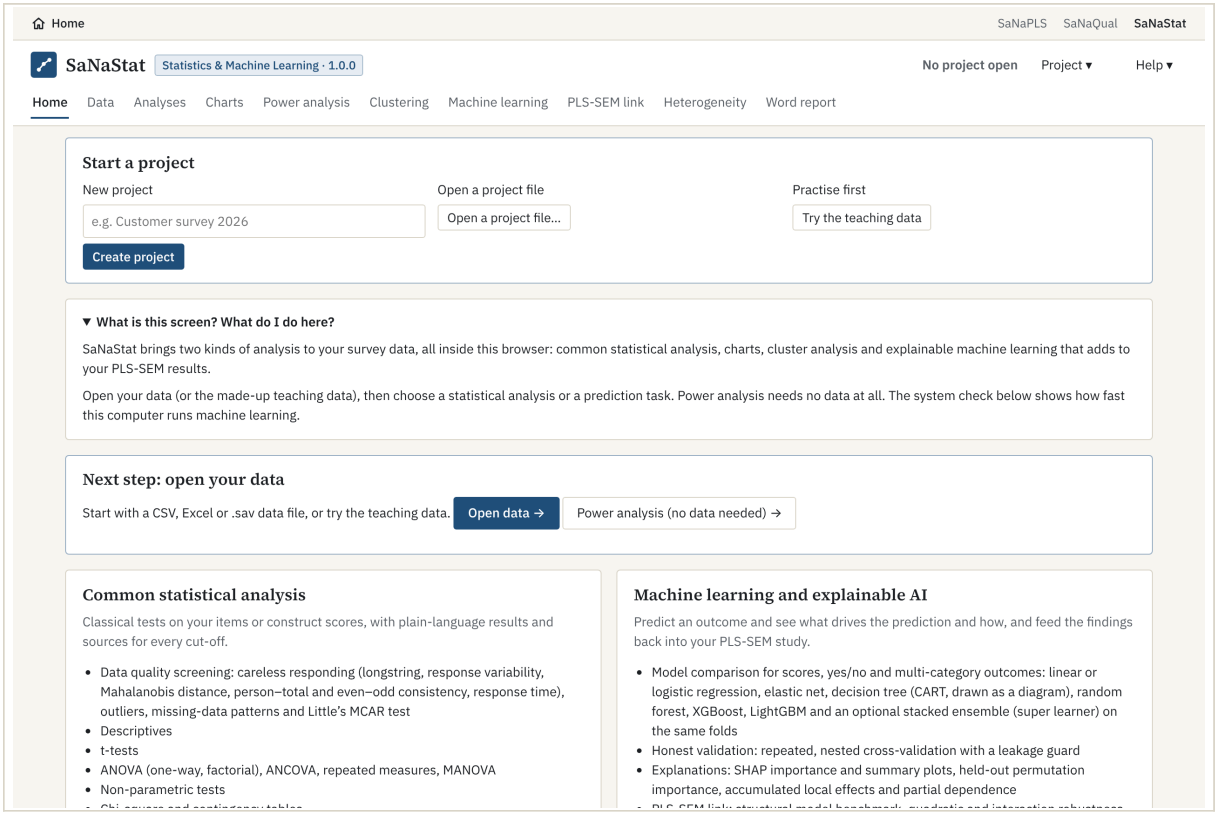
SANASOFT · USER MANUAL

SaNaStat

HR Analytics: a step-by-step manual with practice data

Version 1.0.0 · October 2026



Worked examples: 1,200 simulated employees and a 600-person simulated engagement survey (practice_hr.csv and practice_engagement.csv)

Prof. S. Sabraz Nawaz and Prof. Ghazanfar Ali
SaNaSoft · 2026

Contents

About this manual....5
Who this manual is for....5
What SaNaStat can and cannot do for HR....6
The practice data are simulated....7
What you need....7
How to read the instructions....7
Responsible people analytics....9
The privacy setting: hiding small groups....10
1. Getting started with HR data....12
1.1 What the screens are for....12
1.2 The worked examples....14
1.3 Open the file....14
1.4 Check the variable types....15
1.5 The variable editor....17
1.6 Missing values....17
1.7 Files that are not ready to use....18
1.8 Anonymisation and small groups....18
1.9 Making new variables....19
1.10 Select cases....21
2. Describing the workforce....23
2.1 Descriptive statistics....23
2.2 Describing groups....24
2.3 Charts....26
2.4 Salary is skewed: when to transform....28
3. Survey scales: the engagement survey....31
3.1 Open the file....31
3.2 Data quality screening....31
3.3 Reliability of the scales....33
3.4 Scale scores....35
3.5 Comparing departments: ANOVA and Kruskal–Wallis....36
3.6 Comparing gender and remote work....40
4. Pay equity and adverse impact....42
4.1 Two questions, not one....42
4.2 The raw gap: t-test and Mann–Whitney test....42
4.3 Do women and men do different jobs?....44
4.4 The adjusted gap: regression of log salary....45
4.5 Is the result robust?....48
4.6 Does the gap differ by department?....49
4.7 The pay-gap tool: raw gap, adjusted gap and the Oaxaca–Blinder decomposition..49
4.8 Promotion: selection rates and the four-fifths rule....55
4.9 What these analyses cannot show....57

5. Training evaluation....59
5.1 The data....59
5.2 Did scores change? Paired t-test and Wilcoxon test....59
5.3 Does more training go with more gain?....61
5.4 Regression....62
5.5 What the design cannot show....63
5.6 Planning the next evaluation: power analysis....63
6. Engagement and job satisfaction: what goes with them?....65
6.1 The correlation matrix....65
6.2 Multiple regression of engagement....66
6.3 Hierarchical regression: background first, working conditions second....68
6.4 Job satisfaction....69
6.5 Engagement over the year: repeated-measures ANOVA....69
6.6 Does one factor act through another? Mediation and moderation....71
7. Absence and department differences....72
7.1 Absence by department....72
7.2 What goes with absence?....73
7.3 Counting absence days: Poisson and negative binomial regression....73
7.4 Leaving by department: the chi-square test....77
7.5 Leavers and stayers compared....80
8. Time to leaving: survival analysis....82
8.1 Why time matters, and what censoring is....82
8.2 The survival curve for everyone....82
8.3 Comparing departments: the log-rank test....84
8.4 Cox regression: which factors make leaving come sooner?....88
8.5 A trap: predictors that grow with time....92
9. Predicting attrition....94
9.1 What is being predicted....94
9.2 Logistic regression first....94
9.3 Machine learning on unseen employees....97
9.4 Results · Model comparison....99
9.5 Results · Explanations: what drives leaving?....101
9.6 The imbalance problem....104
9.7 Results · Calibration....105
9.8 Results · Cut-offs, precision and recall....106
9.9 Results · Fairness by group....109
9.10 Ethical use of attrition predictions....110
10. Employee types: clustering....112
10.1 Choosing the variables....112
10.2 The default solution: two clusters....113
10.3 Comparing solutions and choosing k....114
10.4 Reading the three-cluster solution....115
10.5 Describing the clusters with other variables....118
10.6 Keeping the clusters as a variable....119

10.7 Do the employee types differ in leaving?....119
11. Reporting, exporting and the report builder....122
11.1 Copying and exporting one table or chart....122
11.2 Collecting results in the Word report....122
11.3 Using the report well....125
11.4 A checklist for an HR analysis....126
Appendix A. Codebooks....127
A.1 practice_hr.csv (1,200 employees)....127
A.2 practice_engagement.csv (600 respondents)....128
Appendix B. Which tool answers which HR question?....129
Appendix C. Glossary....131
References....133

About this manual

This manual shows how to use SaNaStat 1.0.0 to answer common questions of human resource (HR) management with employee data: How do employees describe their work? Do women and men earn the same after allowing for the job they do, and how much of the gap do job-related factors explain? Are some groups promoted less often? Did a training programme work? What goes with engagement and with absence? How long do employees stay, and who is likely to leave? Every chapter says what to click, what you will see, how to read it, how to report it and what the method cannot tell you. All screenshots and numbers come from real runs of SaNaStat 1.0.0 on the two practice files described below, so you can follow along and obtain the same results.

New in SaNaStat 1.0.0

SaNaML is now called **SaNaStat**, at the address /stat/ (old links lead there). Projects saved in your browser stay, and project files are now saved as .sanastat; older .sanaml files still open. The methods and results are those of SaNaML 0.15.0, which added:

Split file, weight cases and restructuring on the Data screen; the Scheffé, Šidák, LSD, Dunnett and Student–Newman–Keuls post hoc tests; stepwise selection in linear regression; further statistics and stem-and-leaf plots; layered cross-tables with the Mantel–Haenszel test; saved factor scores; and a new Agreement screen (Cohen's kappa and the intraclass correlation). The Common Statistical Tests manual shows each of them step by step.

Who this manual is for

Students of human resource management and organisational behaviour, researchers who analyse employee or survey data, and HR practitioners who want to understand their own data without writing code. You do not need to know any statistics software. Each statistical idea is explained in plain language in a "Concept" box when it first appears.

Table 1

What each chapter covers

Chapter	Topic	What you learn
–	Responsible people analytics	Rules for employee data, and the setting that hides small groups
1	Getting started with HR data	Open the file, variable types, the variable editor, missing values, new variables, Select cases
2	Describing the workforce	Descriptive statistics, frequencies, charts, skewed salaries
3	Survey scales	Data quality screening, reliability (α and ω), group comparisons

Chapter	Topic	What you learn
4	Pay equity and adverse impact	Raw and adjusted pay gaps, the Oaxaca–Blinder decomposition, selection rates and the four-fifths rule
5	Training evaluation	Paired tests, gain scores, partial correlation, power analysis
6	Engagement and job satisfaction	Correlation, multiple and hierarchical regression, repeated measures
7	Absence and department differences	Rank-based tests, chi-square, Poisson and negative binomial regression for counts
8	Time to leaving	Kaplan–Meier curves, the log-rank test, Cox regression with the proportional-hazards check
9	Predicting attrition	Logistic regression, machine learning on unseen employees, cut-offs, fairness by group
10	Employee types	Cluster analysis
11	Reporting and exporting	Export buttons, the report builder, a checklist

What SaNaStat can and cannot do for HR

Please read this section before you start, so that you know what to expect.

SaNaStat is a general program for statistics and machine learning on tables of data. Its **Analyses** screen has a group *HR and people analytics* with two HR tools: **Survival analysis** (time until an event such as leaving) and **Adverse impact and pay gaps** (selection rates and the four-fifths rule; the raw and adjusted pay gap with the Oaxaca–Blinder decomposition). Count regression for outcomes such as absence days is part of **Correlation and regression**. Everything else in this manual uses the general tools: descriptive statistics and charts; making new variables; data quality screening; reliability; t-tests, analysis of variance and their rank-based alternatives; chi-square tests; correlation and regression; power analysis; clustering; and machine learning with explanations, calibration and fairness by group. A privacy setting hides the results of small groups on the screens that report groups. Appendix B shows which tool answers which HR question.

SaNaStat does not know what "turnover" or "span of control" means: you decide what the variables mean and whether an analysis fits the question. The following are **not available in SaNaStat 1.0.0**, and this manual does not pretend otherwise:

- survival analysis with *competing risks* (for example voluntary leaving and dismissal as different events) or with predictors that change over time (chapter 8 explains why this matters);
- models for employees nested in teams or managers (multilevel models);
- imbalance remedies for rare outcomes (class weights or resampling). SaNaStat flags an unbalanced outcome, reports measures suited to it and shows precision and recall at any cut-off you choose, but it does not rebalance the data (chapter 9).

The practice files also have limits: they hold no applicant or hiring records and no dates, so chapter 4 shows adverse impact on promotions and chapter 8 treats tenure as the time to leaving. Both points are explained where they matter.

The practice data are simulated

Both data files were generated by computer for teaching. **No real employee appears in them**, and no company is described. Effects were planted on purpose, so you can check that an analysis finds what was built in. Appendix A gives the codebooks. Real data are messier than these files: the effects will be weaker, the missing data less random and the questions less clean.

What you need

- A computer with an up-to-date browser. SaNaStat runs entirely inside the browser; your data are read on your own computer and are not uploaded.
- The two practice files, practice_hr.csv and practice_engagement.csv, which are shipped with SaNaStat.

Feedback on this manual is welcome at sabraz@seu.ac.lk.

How to read the instructions

Table 2

Conventions used in this manual

Convention	Meaning	Example
Bold coloured text	A button, tab, menu item, box or list on the screen. Click or choose exactly this.	**Run the comparison**
→	A sequence of clicks.	**Data** → **Choose a file…**
Numbered steps	Actions to carry out in this order.	1. Open … 2. Click …
Italics	A technical term explained nearby, or text exactly as it appears in the results.	*nested cross-validation*

Concept: what a term means

Explains a statistical idea in plain language.

Cut-off values: the rule of thumb

The thresholds SaNaStat applies, with their sources. SaNaStat prints the same rule under every check ("Rule: …").

What it means for our example

What the numbers of the worked example mean.

Tip

A shortcut or good practice.

Note

A warning or a point that needs judgement.

How to report it

A model sentence for your report. SaNaStat writes one for every result.

Exercise

A task for you to try, usually with the answer that was planted in the data.

Responsible people analytics

Employee data are about people who depend on the organisation that holds the data. Before any analysis, decide how the results will be used. These principles apply to every chapter.

Note: ten rules for working with employee data

- **Have a purpose and a legal basis.** Collect and analyse only what you need for a stated purpose, and follow the data protection law and the policies that apply to your organisation and country. This manual is not legal advice.
- **Tell people and, where required, ask them.** Employees should know what is collected and why. A survey that promises confidentiality must keep the promise: do not use it to identify individuals.
- **Remove names and numbers that identify people.** Analyse a file in which the employee number is the only identifier, keep the key that links numbers to names somewhere else, and delete the identifier too if you do not need it (section 1.8).
- **Report patterns, not people.** Statistics describe groups. A result such as "attrition is higher among people with low engagement" is a statement about a group; it does not say that this particular employee will leave.
- **Protect small groups.** A table of salary by department and gender may contain a cell with two people, and everybody knows who they are. Decide a minimum group size in advance (many organisations use five to ten people), let SaNaStat hide smaller groups where it can (below), and merge categories elsewhere.
- **Be careful with sensitive variables.** Gender, age, ethnicity, religion, disability and health are protected or sensitive in many places. Use them to check for unfair treatment (chapter 4, and the fairness check of chapter 9), not to predict or to rank individuals. A model that never receives gender can still treat women differently if other variables stand in for it.
- **Do not decide about a person from a model.** A prediction of leaving is a prompt for a supportive conversation, never a reason to withhold a promotion or training. Chapter 9 returns to this.
- **Correlation is not causation.** Engagement and satisfaction go together; that does not show that raising one raises the other. Only design (for example a randomised trial) supports causal claims, and even then with limits.
- **Past data carry past bias.** If past promotions favoured one group, a model trained on them learns that pattern and repeats it. A well-known case outside HR is an algorithm that used health costs as a proxy for health needs and so under-served Black patients (Obermeyer et al., 2019).
- **Keep the humans in charge.** Document what was done, report its limits, and let people who understand the work review conclusions before anyone acts on them.

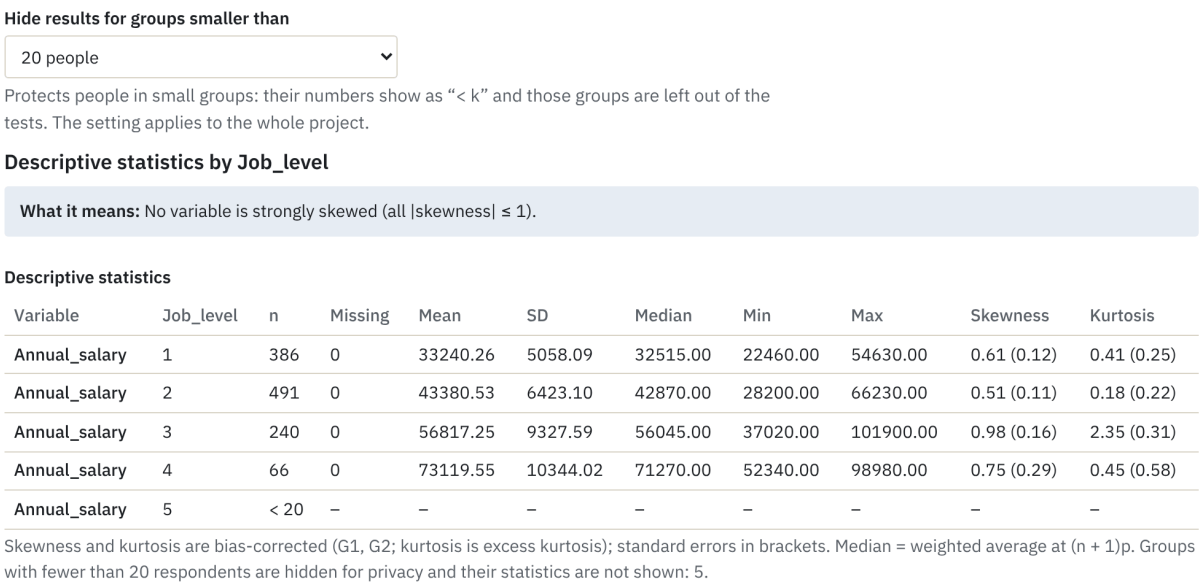
Concept: anonymous, pseudonymous, aggregated

Data are *anonymous* when nobody can work out who a row belongs to, even by combining columns. Data with an employee number are only *pseudonymous*: whoever holds the key can identify people. Combinations of ordinary columns (department, level, age, gender) can single out one person even without a name; the idea of *k-anonymity* asks that every combination occur at least k times (Sweeney, 2002). *Aggregated* results, such as group means, are safest, provided the groups are large enough.

The privacy setting: hiding small groups

SaNaStat can hide the results of small groups for you. The setting **Hide results for groups smaller than** appears on the screens that report groups: **Descriptive statistics** (split by a group), **Chi-square, Fisher, McNemar** (the test of independence and the goodness-of-fit test), **Survival analysis** and **Adverse impact and pay gaps**. Its choices are **Show every group (no hiding)**, 3, 5, 10 and 20 people; the screen explains: *Protects people in small groups: their numbers show as "< k" and those groups are left out of the tests. The setting applies to the whole project.* One choice therefore holds on every screen of the project, and it starts at 5 people. No group in the worked examples of this manual is smaller than 5, so the default hides nothing in them.

To see it work, set Job_level to **Categorical** on the Data screen (section 1.4), open **Descriptive statistics**, tick Annual_salary, split it by Job_level and choose **20 people** (Figure 1). Level 5 has 17 employees, so its row shows *< 20* in place of the count and dashes in place of the statistics, and the note under the table names the hidden group: *Groups with fewer than 20 respondents are hidden for privacy and their statistics are not shown: 5.* The other four levels are shown as usual (for example level 4: n = 66, M = 73119.55).



Hide results for groups smaller than

20 people

Protects people in small groups: their numbers show as "< k" and those groups are left out of the tests. The setting applies to the whole project.

Descriptive statistics by Job_level

What it means: No variable is strongly skewed (all |skewness| ≤ 1).

Descriptive statistics

Variable	Job_level	n	Missing	Mean	SD	Median	Min	Max	Skewness	Kurtosis
Annual_salary	1	386	0	33240.26	5058.09	32515.00	22460.00	54630.00	0.61 (0.12)	0.41 (0.25)
Annual_salary	2	491	0	43380.53	6423.10	42870.00	28200.00	66230.00	0.51 (0.11)	0.18 (0.22)
Annual_salary	3	240	0	56817.25	9327.59	56045.00	37020.00	101900.00	0.98 (0.16)	2.35 (0.31)
Annual_salary	4	66	0	73119.55	10344.02	71270.00	52340.00	98980.00	0.75 (0.29)	0.45 (0.58)
Annual_salary	5	< 20	–	–	–	–	–	–	–	–

Skewness and kurtosis are bias-corrected (G1, G2; kurtosis is excess kurtosis); standard errors in brackets. Median = weighted average at (n + 1)p. Groups with fewer than 20 respondents are hidden for privacy and their statistics are not shown: 5.

Figure 1. *The privacy setting at 20 people: the 17 employees at level 5 are hidden.*

A hidden group is also left out of the tests. The chi-square test of Job_level by Gender (section 4.3) with the setting at 20 people uses four levels and 1,183 employees, $\chi^2(3, N = 1183) = 0.63$, $p = .890$, and the table notes *Groups with fewer than 20 respondents are hidden for privacy and left out of the tests: 5.* At 5 people the same test uses all five levels, $\chi^2(4, N = 1200) = 1.61$, $p = .807$. Set the choice back to **5 people** and Job_level back to **Scale** before you continue.

Note: what the setting does not cover

The setting is used by the four screens named above. The t-tests, the analyses of variance, the regressions, the charts, clustering and machine learning show every group, whatever its size: check the group sizes (the column n, or the labels n = ... under a chart) yourself before you report them. Hiding a group in one table does not stop a reader from working out its numbers from the total and the other groups; merge small groups (section 1.9) when that is a risk.

How to report it

"Results for groups with fewer than [k] employees are not reported, and these groups were left out of the tests."

1. Getting started with HR data

1.1 What the screens are for

SaNaStat has these screens, shown in the menu at the top: **Home**, **Data**, **Analyses**, **Charts**, **Power analysis**, **Clustering**, **Machine learning** and **Word report** (in the SaNaSoft suite also **PLS-SEM link** and **Heterogeneity**). Figure 2 shows the Home screen. Every screen starts with a box *What is this screen? What do I do here?* that explains it in two or three sentences, and a screen that needs data shows "No data yet" until you open a file.

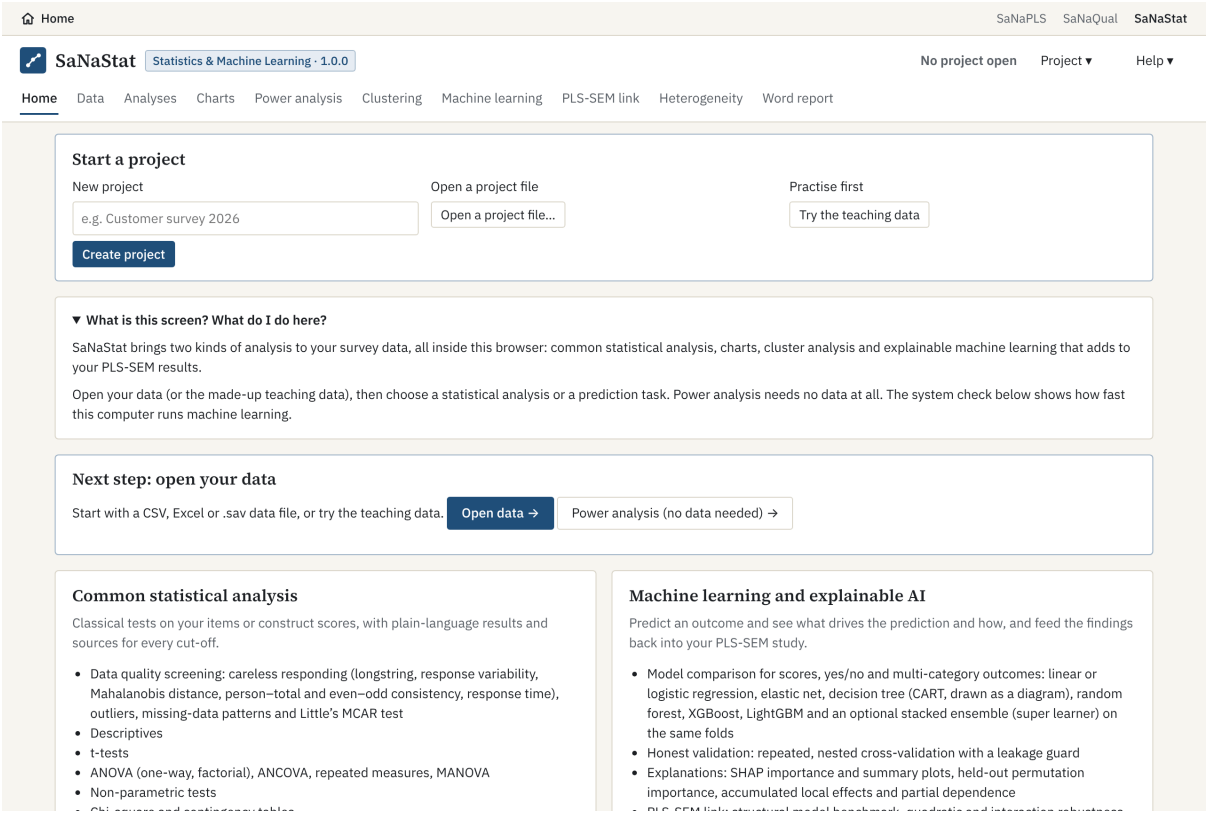


Figure 2. *The Home screen of SaNaStat.*

The list on the left of the **Analyses** screen is grouped (Figure 3): *Common statistics* (from **Help me choose** to **Correlation and regression**), *Measurement and models* (**Factor analysis and reliability**, **Mediation and moderation**, **Missing data: multiple imputation**), *HR and people analytics* (**Survival analysis**, **Adverse impact and pay gaps**) and *Marketing analytics*. Under the list is the *Significance level α* of the project, .05 unless you change it; this manual uses .05 throughout. A result counts as significant when p is below α, and confidence intervals are 100(1 − α)%.

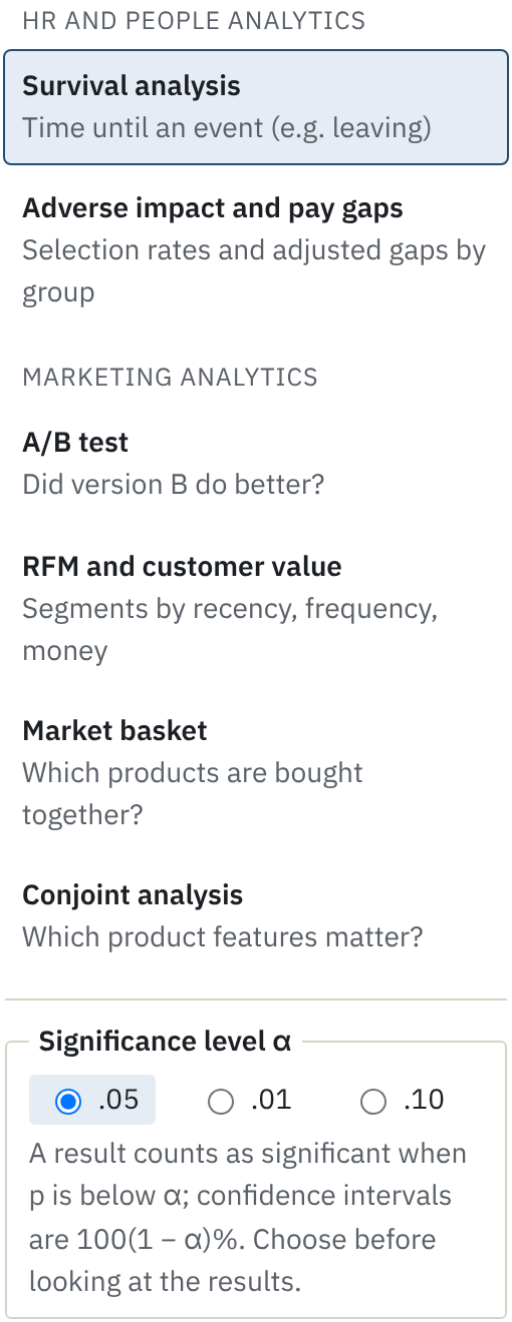


Figure 3. *The list of analyses, grouped, with the significance level α under it.*

Three principles apply throughout:

- **Your data stay on your computer.** Every calculation, machine learning included, runs inside your browser. Nothing is uploaded. This matters for employee data, but it does not replace the rules of the previous chapter: a file on a shared laptop is still a file.
- **Plain language with sources.** Every result starts with *What it means*, lists its checks with a verdict (OK, Note, Check, Info) and the rule behind it, and ends with *How to report it (APA style)*.
- **Honest testing.** Prediction models are judged on employees they have not seen, and every step that learns from the data is repeated inside each training set.

SaNaStat keeps your data, results and report in a project in the browser and restores them on your next visit. They are lost if you clear the browser's site data, so download the Word report (chapter 11) and export the project file (**Project ▾**) to keep your work.

1.2 The worked examples

Table 3
The datasets used in this manual

File	What it holds	Used in
practice_hr.csv	1,200 simulated employees of one company: department, level, gender, age, education, tenure, salary, performance, training, work conditions, survey scores, absence, promotion, a pre/post skills test, engagement in four quarters, and whether the person left within the year	Chapters 1, 2 and 4 to 11
practice_engagement.csv	600 simulated survey respondents: Vigor, Dedication and Absorption items (three each, rated 1 to 7), with department, gender, tenure band and whether the person works remotely	Chapter 3

Tip: follow along

Open SaNaStat next to this manual and repeat each step. Random parts of the analysis (cross-validation splits, bootstrap resamples, clusters) use fixed seeds, so you get the same numbers as the screenshots. Machine learning takes longer on a slow computer, but the results do not change.

1.3 Open the file

1. Click **Data** in the menu.
2. Click **Choose a file…** and pick practice_hr.csv. SaNaStat reads CSV, Excel (.xlsx) and .sav / .zsav files with one row per person and one column per variable, with the column names in the first row.
3. SaNaStat confirms: *Open practice_hr.csv: 1200 respondents, 26 variables.* (Figure 4). SaNaStat calls every row a *respondent*; here a respondent is an employee. If a project already holds data, SaNaStat asks whether to start a new project with the new file or to replace the data; start a new project.

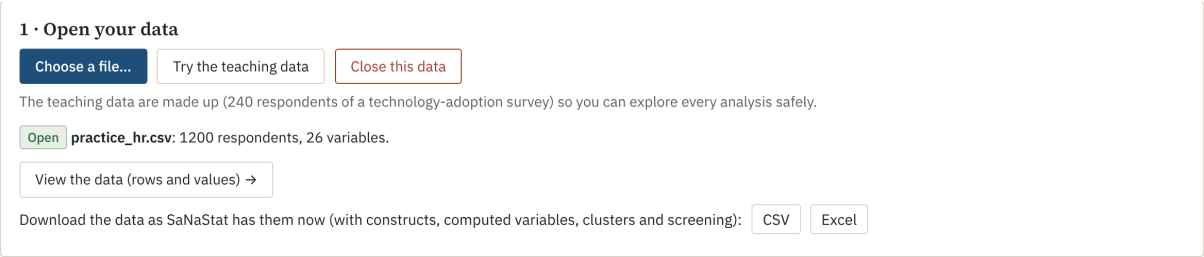


Figure 4. *Step 1 · Open your data, with practice_hr.csv open.*

View the data (rows and values) → shows the rows as SaNaStat analyses them, and the buttons **CSV** and **Excel** download the data as SaNaStat has them now, with the variables you add.

> **Note: the data are simulated**
>
> practice_hr.csv is a simulated company. No real employee appears in it. The effects in it were planted so that you can check your results against the answers in the exercises.

1.4 Check the variable types

Card *2 · Check the variable types* lists every variable with its type, the number of valid values and the number missing (Figure 5). SaNaStat guesses the type from the file and counts the result above the list: *20 scale · 5 categorical · 1 ID*. There are four types:

- **Scale:** numbers you can average (salary, tenure, survey scores).
- **Categorical:** groups without a meaningful order (department, gender, yes/no).
- **Ordinal:** groups in a meaningful order (low / medium / high, a job level treated as groups). Offered wherever groups are, and as the outcome of an ordinal regression.
- **ID:** an employee number or code. It is shown and kept, but never offered to an analysis.

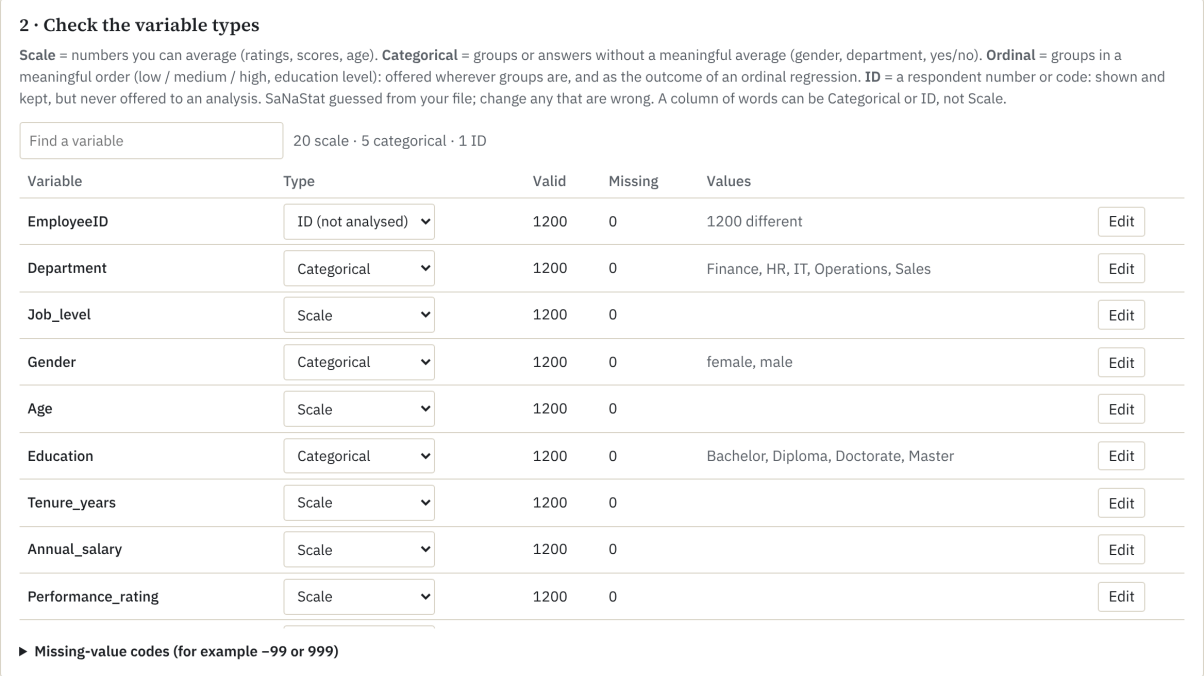


Figure 5. *Card 2 · Check the variable types for practice_hr.csv (the first rows).*

Read the list with HR eyes:

- **EmployeeID** was given the type ID ("1200 different" values). SaNaStat will never offer it as a predictor, a grouping variable or a variable to describe, so it cannot leak into a model by mistake.
- **Department, Gender, Education, Promoted_last_3y and Left_company** are Categorical. Department has the groups Finance, HR, IT, Operations and Sales; Education has Bachelor, Diploma, Doctorate and Master; Gender has the two groups female and male (as coded in the file).
- **Job_level** (1 to 5) was read as Scale because it is a number. This is a judgement call. Treated as Scale, it enters a regression as one straight-line effect per level. Changing its type to **Categorical** or **Ordinal** (the list **Type of Job_level**) turns it into a grouping variable with the groups 1, 2, 3, 4 and 5 that can be used in tables and tests, and gives each level its own coefficient in a regression. Use whichever your question needs; section 4.5 shows both.
- The survey scores (Manager_support, Work_life_balance, Engagement, Job_satisfaction) are averages on a 1 to 5 scale and are correctly Scale.

Note: the order of categories and the reference group

SaNaStat lists categories alphabetically, and a regression compares every other category with the first one, the *reference group*. So female is the reference group for Gender, Finance for Department and Bachelor for Education. The regression screens let you choose another reference group under **Reference groups**; this manual keeps the first one. Remember it when you read coefficients.

1.5 The variable editor

Edit at the end of each row opens the variable editor (Figure 6). In it you can rename the variable (**Rename**), give it a label (**Label (what the question asked)**, then **Save label**), change its **Type**, and, for a categorical or ordinal variable, rename, reorder and merge its categories. **Delete this variable…** removes it. Choose **Ordinal (ordered categories)** for Job_level: SaNaStat reports *Job_level is now ordinal.* and lists the five levels with their counts, lowest first: 386, 491, 240, 66 and 17 employees. Click **Close**, and set Job_level back to **Scale** in the list for now.

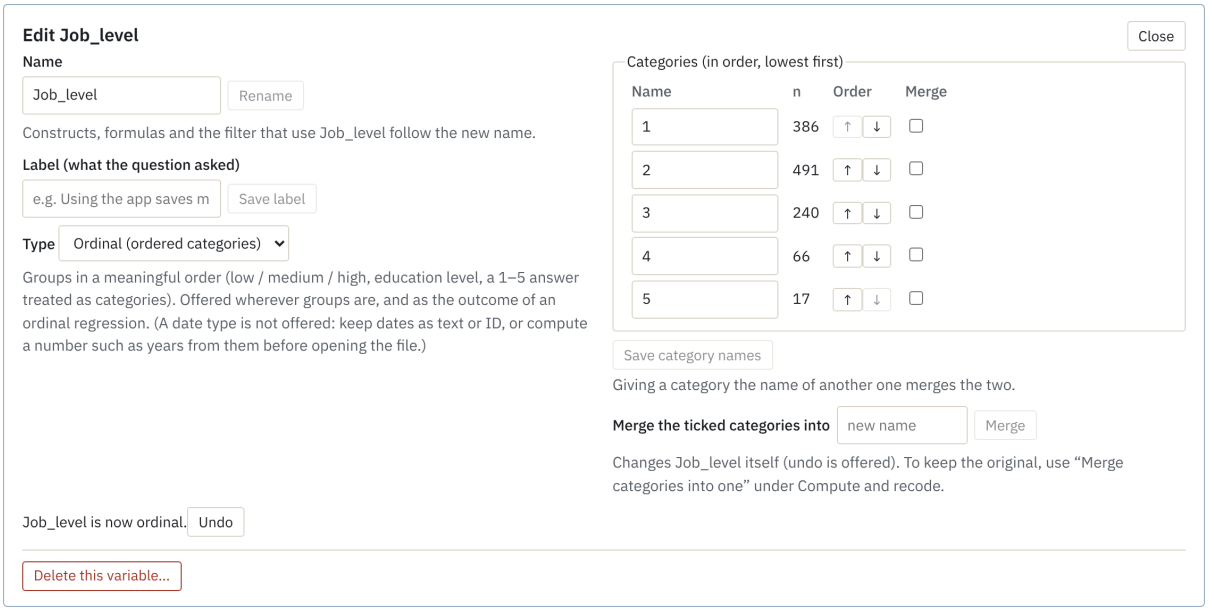


Figure 6. *The variable editor for Job_level with the type Ordinal: five levels, lowest first.*

1.6 Missing values

The columns Valid and Missing in Figure 5 show the missing data: Commute_km has 35 missing values, Manager_support 40 and Job_satisfaction 52. All other columns are complete. In the data file these are empty cells.

Some files use a number such as −99 or 999 for "no answer". Open **Missing-value codes (for example −99 or 999)** under the variable list and type the codes; SaNaStat shows where they occur, and **Set as missing** turns them into missing values. Type **−99, 99, 999** for practice_hr.csv: SaNaStat reports *Found in 1 variable: Skills_test_after (1).* That 99 is a real test score, not a code, so do **not** set it as missing. Codes must be values that cannot occur as answers; check what each one means before you use the button.

How SaNaStat treats missing values: an analysis uses every person who has the values it needs. A t-test of Job_satisfaction uses the 1,148 people who answered it; a regression uses only people with complete data on every variable in the model and says how many were left out (for example "n = 1160"). Machine learning fills in missing values inside each training set (section 9.3).

Concept: why missing data matter

If values are missing at random, dropping those people loses precision but not accuracy. If people are missing *because of* what they would have answered (for example unhappy employees skipping the satisfaction question), the results are biased, and no program can see this from the data alone. Ask why the values are missing before you decide that it does not matter. The data quality screen (section 3.2) tests one simple version, Little's test of missing completely at random.

Tip: multiple imputation

When more than a few per cent of the values are missing, or when the people with missing values differ from the others, **Analyses** → **Missing data: multiple imputation** fills in the missing values several times from the other variables and pools the results with Rubin's rules (Rubin, 1987), so that the uncertainty about the missing values is kept in the standard errors. In practice_hr.csv the variable with the most missing values is Job_satisfaction (52 of 1,200), so this manual analyses the complete cases. The Common Statistical Tests manual (chapter 14) shows the method step by step, with its diagnostics.

1.7 Files that are not ready to use

A few things to fix in your own files before opening them:

- one row per person, no merged cells, no totals row, no notes above the header;
- consistent spelling of categories ("Finance", not "Finance " and "finance");
- numbers stored as numbers, not as text such as "52,000 USD";
- yes/no outcomes coded in the same way throughout;
- dates converted into numbers that mean something, such as tenure in years or the time from hiring to leaving (chapter 8).

1.8 Anonymisation and small groups

Before analysis:

1. **Replace names by numbers.** practice_hr.csv has an employee number and no name. Keep the list that links numbers to names outside the analysis file.
2. **Drop columns you do not need.** Date of birth is not needed when Age is; a home address is not needed when Commute_km is.
3. **Coarsen identifying detail.** Age in five-year bands is safer than the exact age; exact hire dates are safer as tenure bands.
4. **Set the minimum group size.** Choose it in **Hide results for groups smaller than** (Responsible people analytics, above). In practice_hr.csv the smallest education group is Doctorate with 46 people and the smallest department, HR, has 113, but job

level 5 has only 17 people, and the HR department split by gender and education would produce cells with very few people.

5. **Merge or leave out small groups** where the setting does not apply.

> **How to report it**
>
> "Employee numbers were retained for linkage only and were removed before analysis. Groups with fewer than [minimum] employees are not reported."

1.9 Making new variables

Card *4 · Compute and recode variables (optional)* of the Data screen makes a new variable from existing ones. The original is never changed; the new variable is added to the list and can be used on every other screen. **What do you want to make?** offers eleven kinds: **Your own formula (+ − * / ^, ln, mean, if …)**, **If … then … else (groups or scores by rules)**, **Recode with a table of old → new values**, **Ranks, percentiles, or z-scores within groups**, **Transform a scale variable (logarithm, standardise, reverse …)**, **Make groups from a scale variable (median split, tertiles …)**, **Make a 0/1 indicator of a category**, **Combine two scale variables (difference, change, ratio …)**, **Merge categories into one**, **Reverse-code several items at once** and **Statistic across several variables (mean, sum, count …)**. The Common Statistical Tests manual (chapter 2) shows each of them; this section makes the five variables used later in this manual (Figure 7).

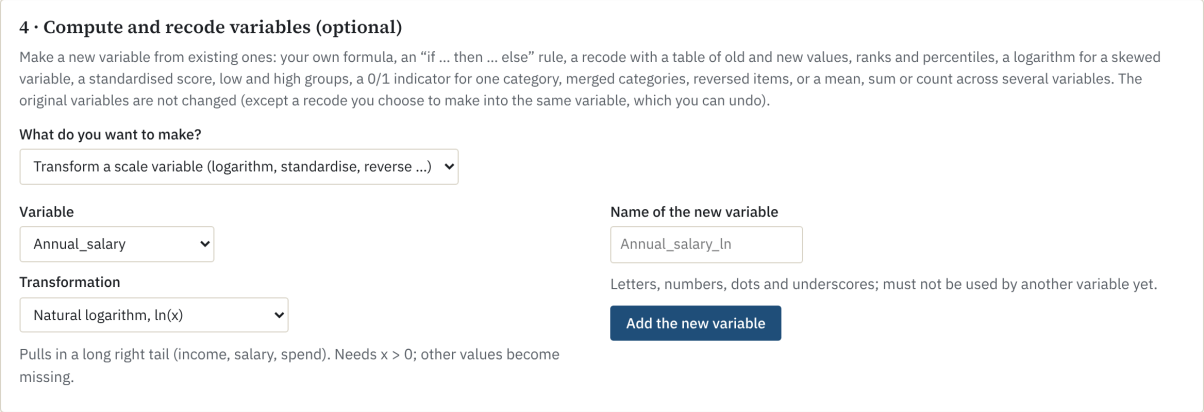


Figure 7. *Card 4: a natural logarithm of Annual_salary is ready to be added.*

Pick the kind, the variable and the operation, and optionally type a name in **Name of the new variable**; if you leave it empty the program supplies one. Select **Add the new variable**. The program confirms with, for example, *Added Annual_salary_ln: 1200 of 1200 respondents have a value.* A new name cannot repeat the name of another variable.

- **Annual_salary_ln: Transform a scale variable**, Annual_salary, **Natural logarithm, ln(x)**. The program notes: "Pulls in a long right tail (income, salary, spend). Needs x > 0; other values become missing." Used for pay equity in chapter 4; section 2.4 shows why.
- **Tenure_group: Make groups from a scale variable**, Tenure_years, **Three equal-sized groups (tertiles)**. The program shows the cut points it found, *Cut points: 3.1,*

6.5, and names the groups Low third, Middle third and High third; a median split would have used the single cut point 4.7 years. The three groups hold 401, 407 and 392 employees.

- **Left_yes:** **Make a 0/1 indicator of a category**, Left_company, with Yes ticked. Its mean is the proportion who left.
- **Skills_gain:** **Combine two scale variables**, Variable A Skills_test_after, **Difference, A − B**, Variable B Skills_test_before (Figure 8). The program describes this choice as "A change or gain score: after minus before." It is used in chapter 5.
- **Dept_merged:** **Merge categories into one**, Department, with Finance and HR ticked and **Corporate** in **Name of the merged category** (Figure 9). The groups of Dept_merged are Corporate (292 employees), IT (195), Operations (363) and Sales (350).

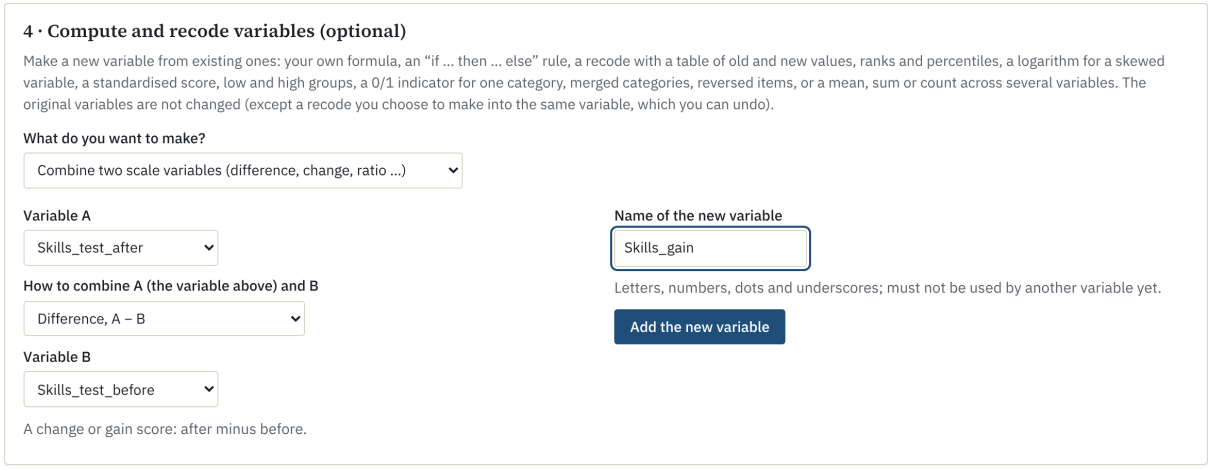


Figure 8. *Combining two variables: Skills_test_after − Skills_test_before.*

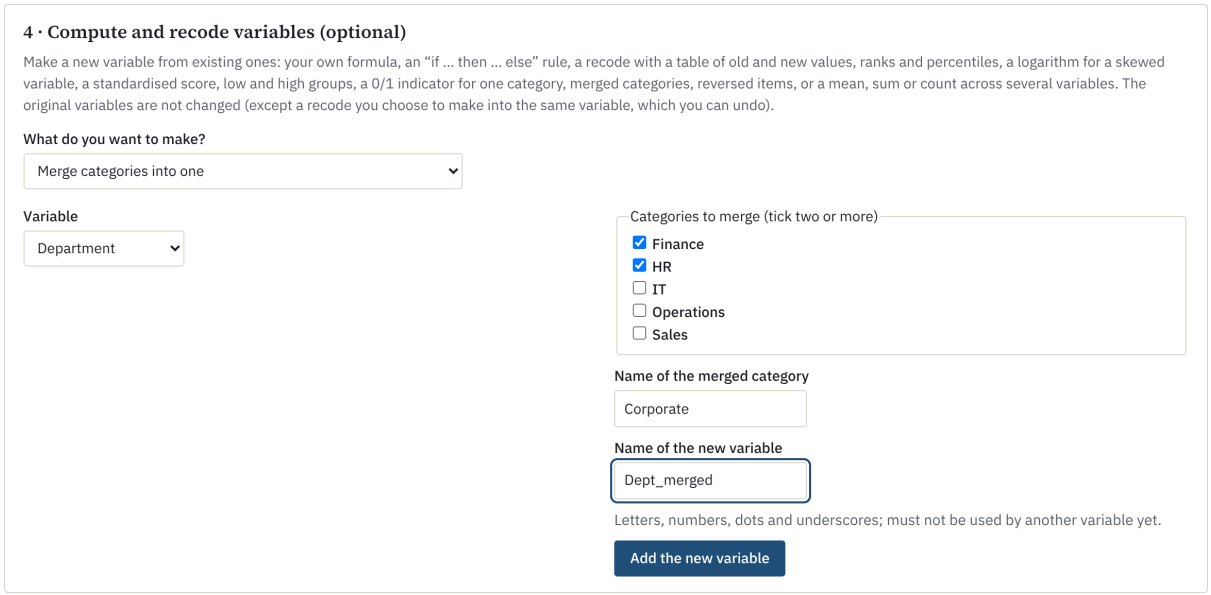


Figure 9. *Merging Finance and HR into Corporate.*

The variable list now shows the new variables with the label *computed* and a short description of how each was made, for example "Tenure_years in groups" and "1 if Left_company is Yes"; the count above the list reads *23 scale · 7 categorical · 1 ID*. When

the items of a questionnaire include statements worded the other way round, **Reverse-code several items at once** makes a reversed copy of each ticked item in one step (the engagement survey of chapter 3 has none).

> **Tip: dividing a scale into groups loses information**
>
> Groups of tertiles or a median split make results easy to show to managers, but they throw away the detail in the scale and can hide or create effects. Use groups for describing and presenting; for testing, prefer the scale itself (for instance, correlate or regress on Tenure_years instead of comparing tenure groups). In this manual Tenure_group is used only to describe and to show a cross-table with leaving (section 7.4).

> **Note: new variables belong to the project**
>
> The new variables are kept with the project in the browser and are included when you download the data from card 1. They are not written into your original file. If you open the original file in a new project, make them again, or keep a note of the recipe (variable, operation, name) with your analysis plan.

1.10 Select cases

Sometimes an analysis should use only some employees, for example one department. Card *7 · Select cases (optional)* does this without deleting anyone (Figure 10):

1. Under **Analyse only respondents where**, choose **Department** in **Variable**, **is** in **Comparison** and **Operations** in **Value**. SaNaStat counts at once: *Department = Operations: 363 of 1200 meet this.*
2. Click **Use this filter**.

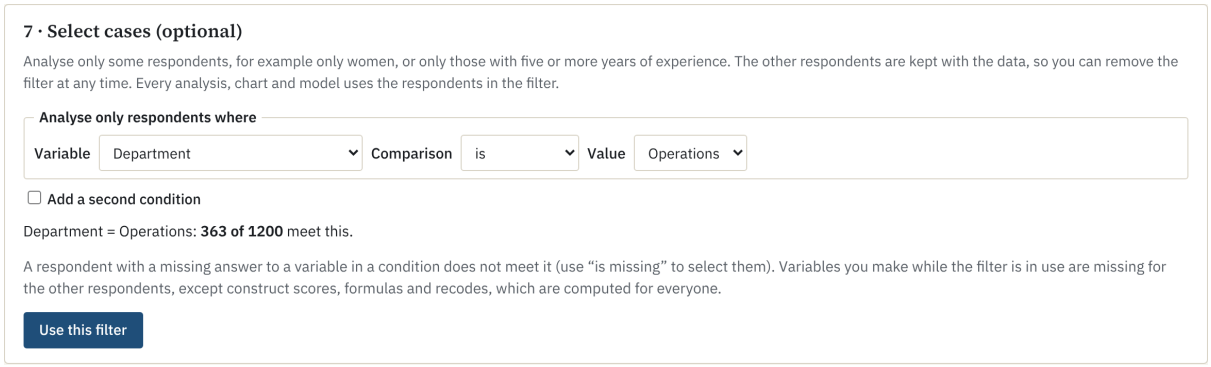


Figure 10. *Select cases: only the Operations department.*

While the filter is in use, every analysis, chart and model uses only the selected employees, and a banner at the top of the screens says so (Figure 11): *Filter in use: Department = Operations (363 of 1200 respondents)*, with **Remove filter** and **Change it**. The descriptive statistics of Engagement are then those of Operations, M = 4.21, SD =

0.61, n = 363; after **Remove filter** they are those of everyone again, M = 4.20, SD = 0.58, n = 1200. Remove the filter before you continue.

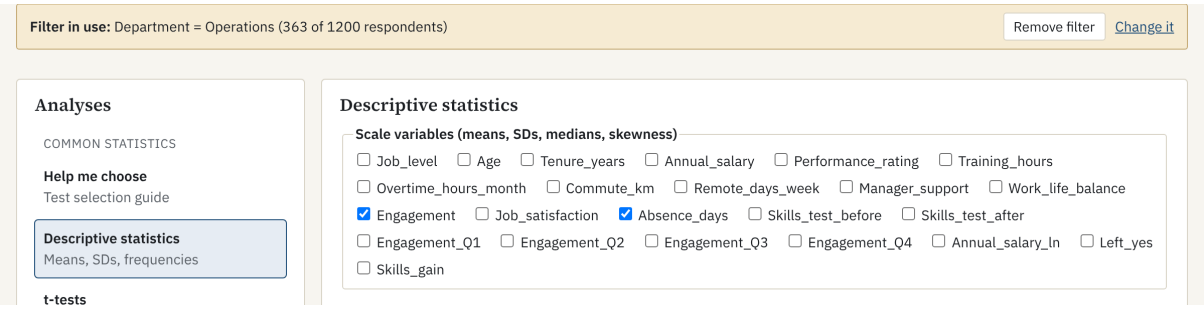


Figure 11. *The banner shows that only Operations is analysed.*

2. Describing the workforce

Every HR analysis should start with a description of who is in the data. In this chapter we describe the 1,200 employees of practice_hr.csv with numbers (**Analyses**) and with pictures (**Charts**), and deal with the first practical problem of pay data: salaries are skewed.

2.1 Descriptive statistics

1. Open practice_hr.csv (section 1.3).
2. Click **Analyses** → **Descriptive statistics**.
3. Under *Scale variables (means, SDs, medians, skewness)*, tick Age, Tenure_years, Annual_salary, Engagement and Absence_days.
4. Leave **Split the scale statistics by (optional)** at "none". The result appears below (Figure 12).

What it means: 2 of the 5 rows show clear skewness (|skewness| > 1): report medians as well, and read the rank-based tests next to the tests of means.

Descriptive statistics

Variable	n	Missing	Mean	SD	Median	Min	Max	Skewness	Kurtosis
Age	1200	0	37.17	8.81	37.00	21.00	63.00	0.24 (0.07)	−0.31 (0.14)
Tenure_years	1200	0	5.53	4.01	4.70	0.10	25.10	1.14 (0.07)	1.62 (0.14)
Annual_salary	1200	0	45242.77	14591.35	42310.00	22460.00	125020.00	1.63 (0.07)	4.38 (0.14)
Engagement	1200	0	4.20	0.58	4.24	2.09	5.00	−0.58 (0.07)	0.01 (0.14)
Absence_days	1200	0	3.88	1.98	4.00	0.00	11.00	0.49 (0.07)	0.24 (0.14)

Skewness and kurtosis are bias-corrected (G1, G2; kurtosis is excess kurtosis); standard errors in brackets. Median = weighted average at (n + 1)p.

Figure 12. *Descriptive statistics of five variables for the whole company.*

Concept: mean, median, skewness

The *mean* is the arithmetic average. The *median* is the middle value: half of the people are above it. For a skewed variable such as salary the two differ, and the median describes the typical person better. *Skewness* measures the lack of symmetry: 0 is symmetric, a positive value means a long tail of high values, and values beyond ±1 are clearly skewed. *Kurtosis* (excess kurtosis) measures how heavy the tails are; 0 is like a normal distribution. SaNaStat gives both with their standard errors in brackets.

For the whole company the table reads: Age M = 37.17 (SD = 8.81), Tenure_years M = 5.53 (SD = 4.01, median 4.70, skewness 1.14), Annual_salary M = 45242.77 (SD = 14591.35, median 42310.00, skewness 1.63, kurtosis 4.38), Engagement M = 4.20 (SD = 0.58) and Absence_days M = 3.88 (SD = 1.98). SaNaStat adds its own reading: *2 of the 5 rows show*

clear skewness (|skewness| > 1): report medians as well, and read the rank-based tests next to the tests of means.

What it means for our example

The average salary (45242.77) is well above the median (42310.00) because a few very high salaries (the maximum is 125020.00) pull the mean up. Tenure is skewed in the same way: many people are new and a few have worked for twenty years or more. Age and engagement are roughly symmetric. The salary unit is not stated in the file; we simply call the numbers salary units.

Ticking variables under *Categorical variables (frequencies and percentages)* gives frequency tables with a bar chart of each. Department: Finance 179 (14.9%), HR 113 (9.4%), IT 195 (16.3%), Operations 363 (30.3%) and Sales 350 (29.2%). Education: Bachelor 595 (49.6%), Diploma 267 (22.3%), Doctorate 46 (3.8%) and Master 292 (24.3%). Left_company: 1,024 stayed (85.3%) and 176 left (14.7%).

2.2 Describing groups

Use **Split the scale statistics by (optional)** to see the statistics for each group (Figure 13). Choose Department.

What it means: 8 of the 25 rows show clear skewness (|skewness| > 1): report medians as well, and read the rank-based tests next to the tests of means.

Descriptive statistics

Variable	Department	n	Missing	Mean	SD	Median	Min	Max	Skewness	Kurtosis
Age	Finance	179	0	37.39	9.11	37.00	21.00	63.00	0.32 (0.18)	−0.30 (0.36)
Age	HR	113	0	36.30	8.79	37.00	21.00	56.00	−0.09 (0.23)	−0.80 (0.45)
Age	IT	195	0	37.36	9.06	37.00	21.00	63.00	0.21 (0.17)	−0.33 (0.35)
Age	Operations	363	0	37.43	8.60	37.00	21.00	63.00	0.27 (0.13)	−0.19 (0.26)
Age	Sales	350	0	36.96	8.75	36.00	21.00	62.00	0.28 (0.13)	−0.28 (0.26)
Tenure_years	Finance	179	0	5.71	4.33	5.00	0.10	25.10	1.44 (0.18)	3.06 (0.36)
Tenure_years	HR	113	0	5.45	4.45	4.40	0.10	24.60	1.32 (0.23)	2.75 (0.45)
Tenure_years	IT	195	0	5.84	4.11	5.00	0.10	19.30	0.97 (0.17)	0.51 (0.35)
Tenure_years	Operations	363	0	5.33	3.78	4.50	0.10	21.10	1.17 (0.13)	1.69 (0.26)
Tenure_years	Sales	350	0	5.51	3.87	4.80	0.10	18.70	0.91 (0.13)	0.48 (0.26)
Annual_salary	Finance	179	0	47920.95	14857.81	44530.00	26850.00	124920.00	2.28 (0.18)	8.30 (0.36)
Annual_salary	HR	113	0	44747.96	13458.11	43350.00	25500.00	102110.00	1.58 (0.23)	4.37 (0.45)
Annual_salary	IT	195	0	52486.00	17533.10	47880.00	28710.00	125020.00	1.62 (0.17)	3.42 (0.35)
Annual_salary	Operations	363	0	41304.52	12491.95	38480.00	22460.00	85620.00	1.14 (0.13)	1.12 (0.26)
Annual_salary	Sales	350	0	44081.83	13320.54	41240.00	25510.00	98000.00	1.32 (0.13)	2.19 (0.26)
Engagement	Finance	179	0	4.20	0.62	4.25	2.49	5.00	−0.59 (0.18)	−0.41 (0.36)
Engagement	HR	113	0	4.24	0.53	4.16	2.89	5.00	−0.26 (0.23)	−0.61 (0.45)
Engagement	IT	195	0	4.19	0.55	4.26	2.39	5.00	−0.61 (0.17)	0.27 (0.35)
Engagement	Operations	363	0	4.21	0.61	4.27	2.09	5.00	−0.72 (0.13)	0.30 (0.26)
Engagement	Sales	350	0	4.18	0.57	4.18	2.17	5.00	−0.47 (0.13)	−0.09 (0.26)
Absence_days	Finance	179	0	4.05	1.98	4.00	0.00	10.00	0.32 (0.18)	−0.37 (0.36)
Absence_days	HR	113	0	3.67	1.93	4.00	0.00	10.00	0.67 (0.23)	0.80 (0.45)
Absence_days	IT	195	0	4.02	1.80	4.00	0.00	10.00	0.36 (0.17)	0.18 (0.35)
Absence_days	Operations	363	0	3.87	2.00	4.00	0.00	10.00	0.46 (0.13)	0.04 (0.26)
Absence_days	Sales	350	0	3.80	2.06	4.00	0.00	11.00	0.62 (0.13)	0.69 (0.26)

Skewness and kurtosis are bias-corrected (G1, G2; kurtosis is excess kurtosis); standard errors in brackets. Median = weighted average at (n + 1)p.

Figure 13. *Descriptive statistics by Department (the first rows of the table are Age and Tenure_years).*

Table 4

Annual salary by department (simulated data)

Department	n	Mean	SD	Median
Finance	179	47920.95	14857.81	44530.00
HR	113	44747.96	13458.11	43350.00
IT	195	52486.00	17533.10	47880.00
Operations	363	41304.52	12491.95	38480.00
Sales	350	44081.83	13320.54	41240.00

Pay differs by department: IT has the highest median (47880.00) and Operations the lowest (38480.00). By education the gap is larger still: Doctorate (n = 46) averages 55714.13 and Master (n = 292) 55121.34, against 41919.87 for Bachelor (n = 595) and 40040.15 for Diploma (n = 267). By gender, women (n = 563) average 44213.37 and men

(n = 637) 46152.57. That difference is not yet a statement about pay equity: women and men may differ in level, tenure, department and education. Chapter 4 deals with that.

Engagement, by contrast, hardly differs between departments (means from 4.18 in Sales to 4.24 in HR) and absence differs very little (means from 3.67 in HR to 4.05 in Finance).

> **Note: small groups**
>
> With a second split, groups quickly become small. The Doctorate row has only 46 people. Before you put a table like this in a report, apply the minimum group size you decided on (section 1.8).

2.3 Charts

The **Charts** screen draws six kinds of picture: a histogram, a box plot, means with confidence intervals, a scatter plot with a fitted line, a correlation heatmap and a bar chart of percentages. Choose the **Chart type**, then the variables. Every chart comes with a short description, a summary table, and a sentence and caption for the report. Charts show patterns, not proof: an apparent difference needs a test.

- **Histogram** (Figure 14) shows how the values are spread; the lines mark the mean and the median.
- **Box plot** (Figure 15) shows the median, the middle half of the values (the box) and unusual values beyond 1.5 times the interquartile range (the circles). Choose **Groups** to split it.
- **Scatter plot** (Figure 16) shows two variables together with the straight-line fit.
- **Bar chart** of percentages and **means with confidence intervals** are used for categories and for group averages.

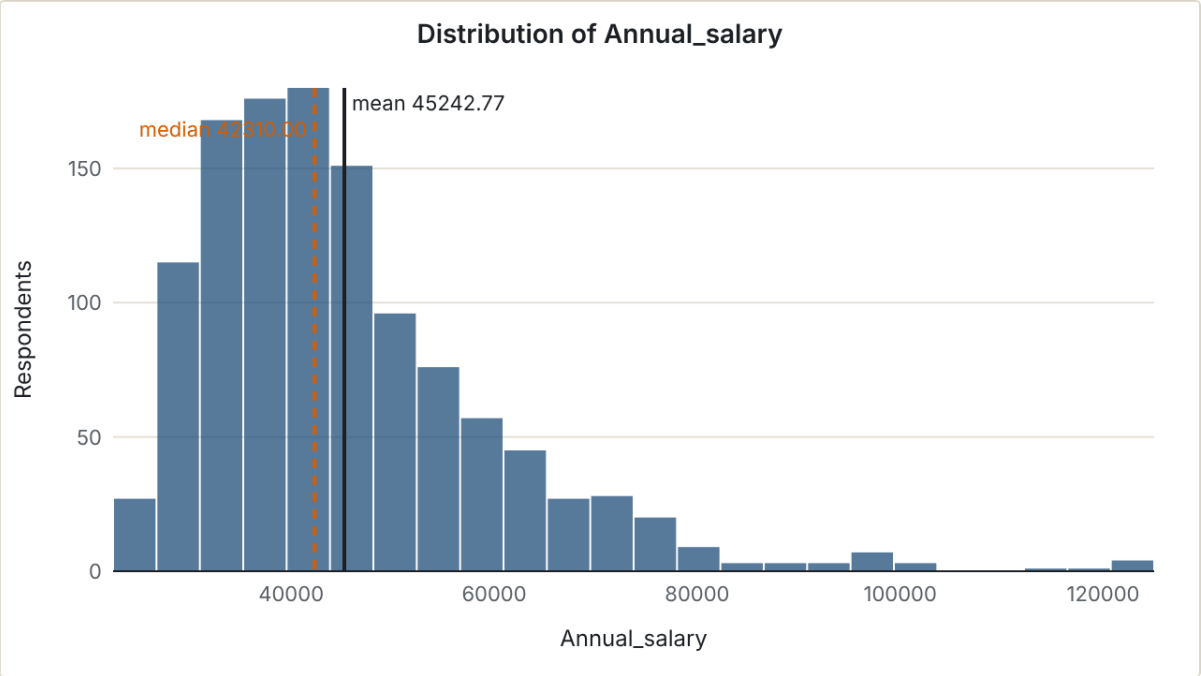


Figure 14. *Histogram of Annual_salary: a long tail to the right.*

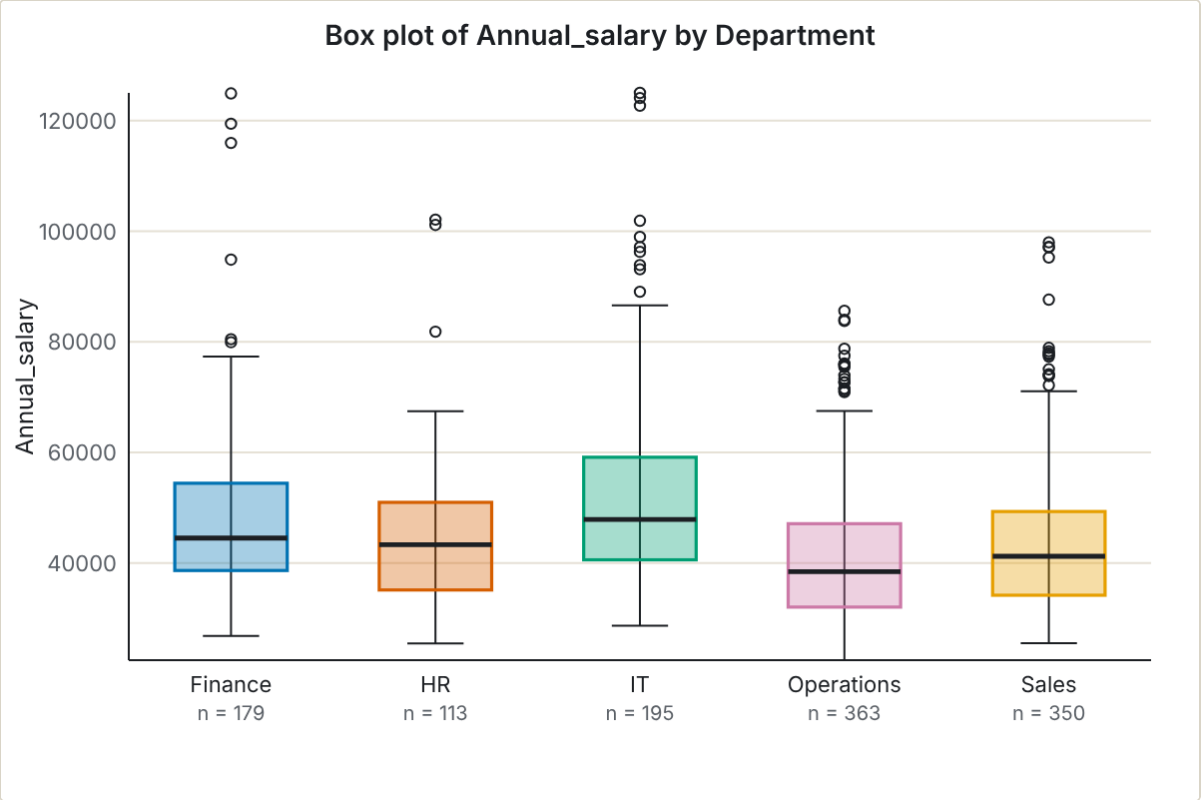


Figure 15. *Box plot of Annual_salary by Department. The circles are employees with unusually high salaries.*

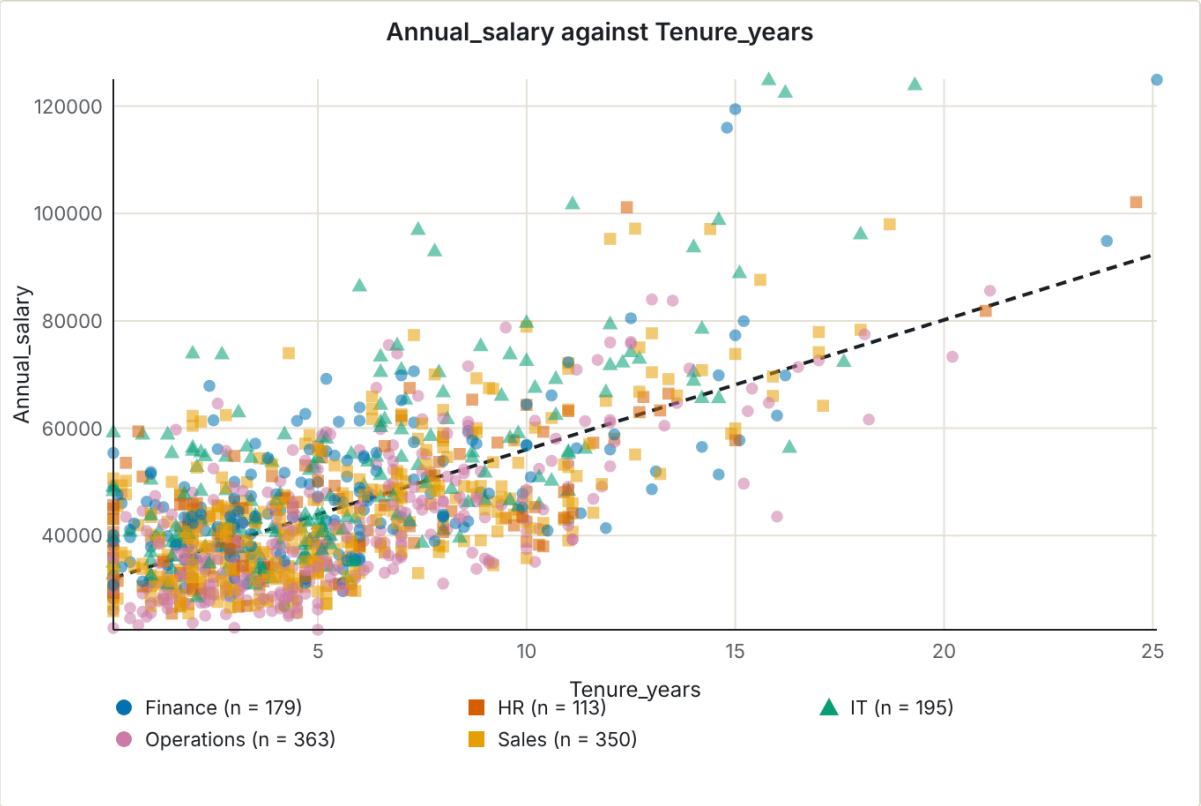


Figure 16. *Annual_salary against Tenure_years, with the fitted line and the departments marked.*

What it means for our example

The histogram shows the right skew in numbers (skewness 1.63). The box plot shows that IT and Finance have the longest upper tails. The scatter plot shows that salary rises with tenure: r = .66, with a fitted slope of 2415.372 salary units per additional year (intercept 31878.313). The fitted line is only a summary; chapter 4 asks how much of this is tenure itself and how much is level and department.

2.4 Salary is skewed: when to transform

Pay is a classic skewed variable: most people earn a modest amount and few earn a lot, and differences between people are better described as percentages than as amounts (a rise of 2,000 means more to someone earning 25,000 than to someone earning 100,000). Taking the natural logarithm of salary makes the distribution more symmetric and turns differences into approximate percentages.

Make the logarithm on the Data screen, card 4, as in section 1.9. Compare the two histograms.

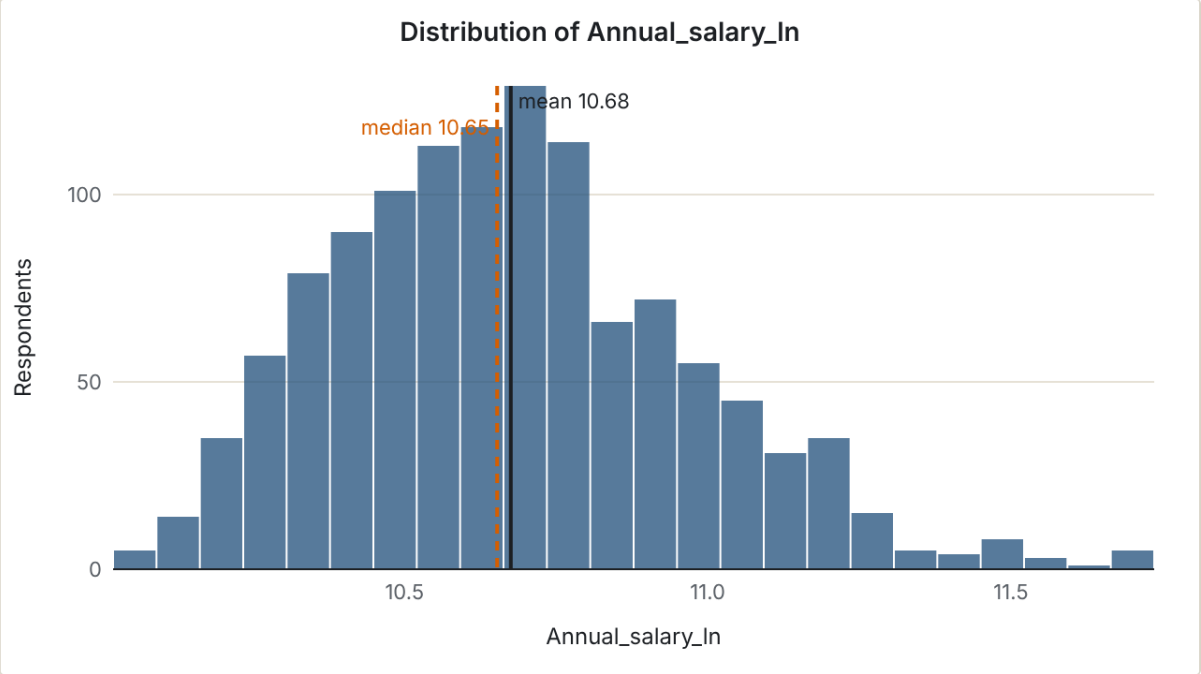


Figure 17. *Histogram of Annual_salary_ln. The tail is much shorter than in the histogram of Annual_salary.*

Table 5

Annual salary and its logarithm

Variable	Mean	SD	Median	Skewness	Kurtosis
Annual_salary	45242.77	14591.35	42310.00	1.63	4.38
Annual_salary_ln	10.68	0.29	10.65	0.55	0.32

The logarithm brings the skewness down from 1.63 to 0.55 and the kurtosis from 4.38 to 0.32, and mean and median now nearly agree (10.68 and 10.65). It does not remove the skew completely.

Concept: differences on the log scale are percentages

If log salary differs by a small amount b between two groups, the ratio of their salaries is about e^b. For b = 0.044 the ratio is e^0.044 = 1.045: the first group earns about 4.5% more. For small b the percentage is close to 100 × b. When this manual quotes a percentage that the program did not print, it says that it was calculated by hand.

Tip: when to transform

Transform a variable when its skew or its spread of errors makes the standard methods unreliable, and when the transformed scale has a sensible meaning. Percentage differences in pay are a sensible meaning. Do not transform a variable only to get a significant result.

Exercise

1. Which department has the highest median tenure, and what is it? (Answer: IT and Finance both have a median tenure of 5.00 years.)

1. Draw the histogram of Tenure_years. How skewed is it? (Answer: skewness 1.14, with mean 5.53 and median 4.70.)

1. Split Annual_salary by Education. Which two education groups have the highest means? (Answer: Doctorate 55714.13 and Master 55121.34.)

3. Survey scales: the engagement survey

Most HR questions about attitudes are answered with a questionnaire. A single question is a noisy measure, so questionnaires use several items for one idea and average them. This chapter uses the engagement survey practice_engagement.csv: 600 respondents answered nine statements on a scale from 1 to 7. Three items measure *Vigor* (energy at work), three measure *Dedication* (involvement and pride) and three measure *Absorption* (being immersed in the work), the three dimensions of work engagement described by Schaufeli et al. (2002). The responses are simulated; no real person answered them.

3.1 Open the file

Click **Data** → **Choose a file…** and open practice_engagement.csv; when SaNaStat asks, start a new project for it. SaNaStat reports 600 respondents and 14 variables, counted as *9 scale · 4 categorical · 1 ID*: RespondentID is the ID; Department, Gender, Tenure_band and Works_remotely are Categorical; the nine items are Scale. Note that Tenure_band lists its groups in alphabetical order ("<2 years, >10 years, 2-5 years, 6-10 years"), not in the natural order; the variable editor (section 1.5) can put them in order.

Vigor2 has 20 missing answers and Absorption3 has 21; the other items are complete.

3.2 Data quality screening

Survey answers can be careless: the same answer to every item, answers at random, answers given too fast to have been read. Screening takes a minute and is good practice before scoring a questionnaire.

1. On the Data screen click **Check data quality →** (card 6).
2. Under *Screening settings (What to check)* SaNaStat has ticked the nine items. The file has no completion time, so leave **Completion time in seconds** at "none". Leave the rule at **2 or more indices (Recommended)**.
3. Click **Screen the data**.

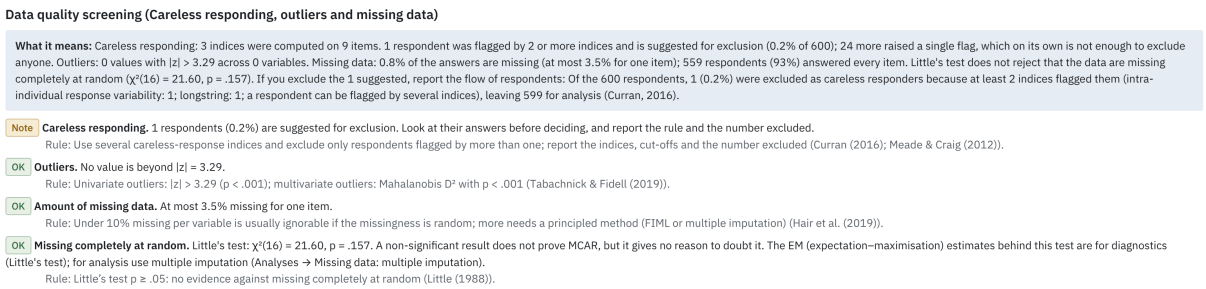
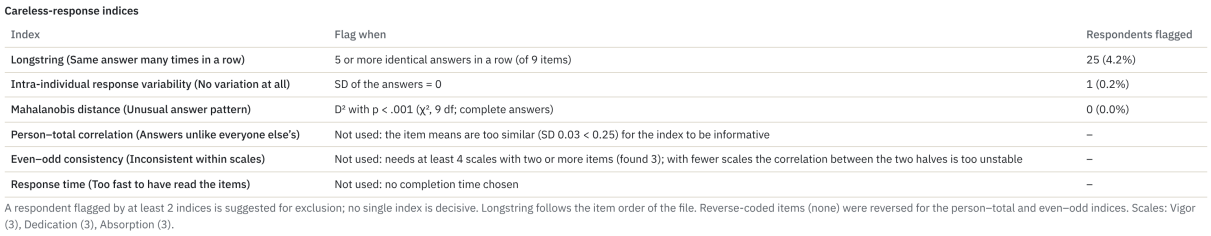


Figure 18. *The start of the data quality report: what it means and the checks.*

Figure 18 shows the result. SaNaStat computed three indices on nine items: 1 respondent (0.2% of 600) was flagged by two or more indices and is suggested for exclusion, and 24 more raised a single flag, which on its own is not enough to exclude anyone. No value is an outlier ($|z| > 3.29$), 0.8% of the answers are missing (at most 3.5% for one item), 559

respondents (93%) answered every item, and Little's test does not reject that the data are missing completely at random, $\chi^2(16) = 21.60$, $p = .157$.



Careless-response indices

Index	Flag when	Respondents flagged
Longstring (Same answer many times in a row)	5 or more identical answers in a row (of 9 items)	25 (4.2%)
Intra-individual response variability (No variation at all)	SD of the answers = 0	1 (0.2%)
Mahalanobis distance (Unusual answer pattern)	D^2 with $p < .001$ (χ^2, 9 df; complete answers)	0 (0.0%)
Person–total correlation (Answers unlike everyone else's)	Not used: the item means are too similar (SD 0.03 < 0.25) for the index to be informative	–
Even–odd consistency (Inconsistent within scales)	Not used: needs at least 4 scales with two or more items (found 3); with fewer scales the correlation between the two halves is too unstable	–
Response time (Too fast to have read the items)	Not used: no completion time chosen	–

A respondent flagged by at least 2 indices is suggested for exclusion; no single index is decisive. Longstring follows the item order of the file. Reverse-coded items (none) were reversed for the person–total and even–odd indices. Scales: Vigor (3), Dedication (3), Absorption (3).

Figure 19. *The careless-response indices, with the reason for each index that was not used.*

Table 6

Careless-response indices for the engagement survey

Index	Flag when	Flagged
Longstring	5 or more identical answers in a row (of 9 items)	25 (4.2%)
Intra-individual response variability	SD of the answers = 0	1 (0.2%)
Mahalanobis distance	D^2 with $p < .001$	0 (0.0%)
Person–total correlation	Not used: the item means are too similar	–
Even–odd consistency	Not used: needs at least 4 scales (found 3)	–
Response time	Not used: no completion time chosen	–

The even–odd index compares the two halves of each scale. The program uses it only when the questionnaire has at least four scales with two or more items, because with fewer scales the correlation between the halves is too unstable; the engagement survey has three scales of three items, so the program states that the index is not used (Figure 19).

Note: read the indices, do not just count them

The suggested respondent is R364, who gave the same answer to all nine items (longest run 9, response variability 0.00; Mahalanobis $D^2 = 7.9$, $p = .540$): a genuine straight-liner. The other 24 respondents with one flag all have a longstring of 5 or 6 identical answers. A run of five identical answers can occur in honest responding, so use the flags to look at the answers, not as a verdict. Respondents in the lists are labelled by the ID column of the file (here RespondentID), not by a row number.

In this manual we **keep all 600 respondents**, so that your numbers match ours. In a real project you would look at R364's answers and those of the 24 single-flag respondents, decide, and report the rule and the number excluded (the report sentence is given under the result). Exclusion is done under *Your decision*, with **Exclude the 1 ticked respondent**; it can be undone.

3.3 Reliability of the scales

A scale is *reliable* if its items hang together. The usual measures are Cronbach's α and McDonald's ω.

Concept: Cronbach's α and McDonald's ω

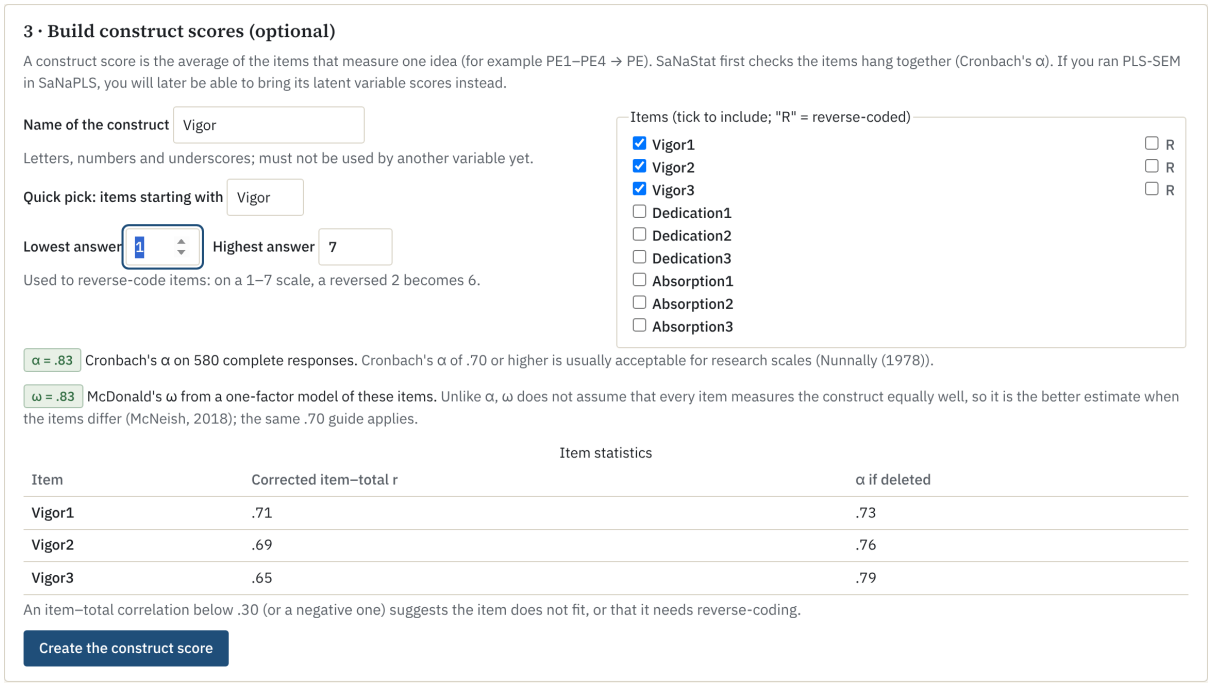
Alpha is the average relationship between the items of a scale, raised by the number of items. It runs from 0 to 1. Because it grows with the number of items, a scale of nine items can have a high alpha even if its items measure different things, so alpha does not prove that a scale measures one idea. Omega comes from a one-factor model of the items; unlike alpha it does not assume that every item measures the construct equally well, so it is the better estimate when the items differ (McNeish, 2018). The *corrected item–total correlation* shows how strongly each item goes with the rest of the scale, and *α if deleted* shows what would happen if the item were dropped.

Cut-off values: reliability

Cronbach's α (and ω) of .70 or higher is usually acceptable for research scales (Nunnally, 1978). A corrected item–total correlation below .30, or a negative one, suggests that the item does not fit or needs reverse-coding.

Card *3 · Build construct scores (optional)* of the Data screen builds construct scores:

1. In **Name of the construct** type Vigor.
2. In **Quick pick: items starting with** type Vigor and press Tab. SaNaStat ticks Vigor1 to Vigor3 and sets **Lowest answer** and **Highest answer** to 1 and 7. (If an item were worded in the opposite direction, you would tick its **R** box to reverse it. None of these items is.)
3. Read α, ω and the item table (Figure 20).
4. Click **Create the construct score**. Repeat for Dedication and Absorption.
5. For the total, name the construct Work_engagement and tick all nine items (Figure 21).



3 · Build construct scores (optional)

A construct score is the average of the items that measure one idea (for example PE1–PE4 → PE). SaNaStat first checks the items hang together (Cronbach's α). If you ran PLS-SEM in SaNaPLS, you will later be able to bring its latent variable scores instead.

Name of the construct: Vigor

Letters, numbers and underscores; must not be used by another variable yet.

Quick pick: items starting with: Vigor

Lowest answer: 1 Highest answer: 7

Used to reverse-code items: on a 1–7 scale, a reversed 2 becomes 6.

Items (tick to include; "R" = reverse-coded)

☑ Vigor1	☐ R
☑ Vigor2	☐ R
☑ Vigor3	☐ R
☐ Dedication1	
☐ Dedication2	
☐ Dedication3	
☐ Absorption1	
☐ Absorption2	
☐ Absorption3	

α = .83 Cronbach's α on 580 complete responses. Cronbach's α of .70 or higher is usually acceptable for research scales (Nunnally (1978)).

ω = .83 McDonald's ω from a one-factor model of these items. Unlike α, ω does not assume that every item measures the construct equally well, so it is the better estimate when the items differ (McNeish, 2018); the same .70 guide applies.

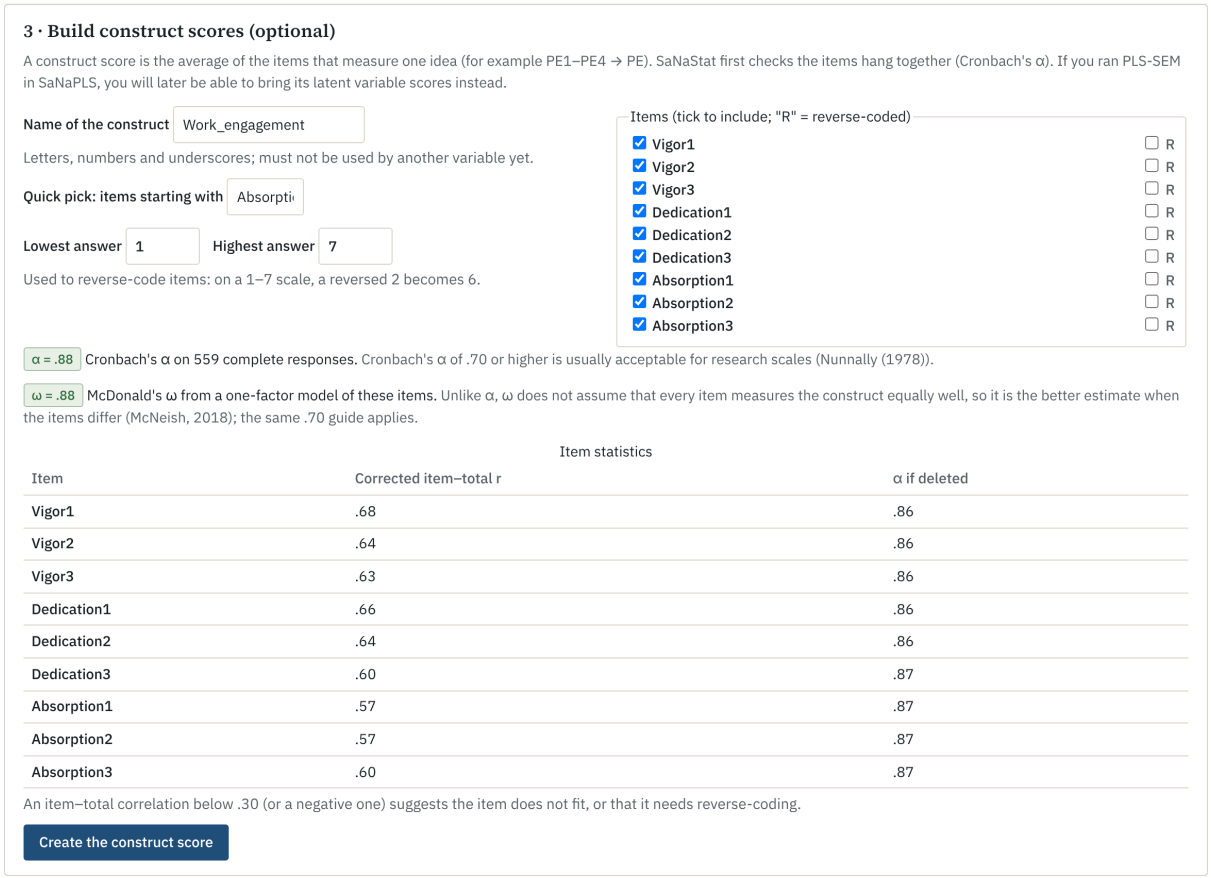
Item statistics

Item	Corrected item–total r	α if deleted
Vigor1	.71	.73
Vigor2	.69	.76
Vigor3	.65	.79

An item–total correlation below .30 (or a negative one) suggests the item does not fit, or that it needs reverse-coding.

Create the construct score

Figure 20. *Reliability of the three Vigor items: α, ω and the item statistics.*



3 · Build construct scores (optional)

A construct score is the average of the items that measure one idea (for example PE1–PE4 → PE). SaNaStat first checks the items hang together (Cronbach's α). If you ran PLS-SEM in SaNaPLS, you will later be able to bring its latent variable scores instead.

Name of the construct: Work_engagement

Letters, numbers and underscores; must not be used by another variable yet.

Quick pick: items starting with: Absorpti

Lowest answer: 1 Highest answer: 7

Used to reverse-code items: on a 1–7 scale, a reversed 2 becomes 6.

Items (tick to include; "R" = reverse-coded)

☑ Vigor1	☐ R
☑ Vigor2	☐ R
☑ Vigor3	☐ R
☑ Dedication1	☐ R
☑ Dedication2	☐ R
☑ Dedication3	☐ R
☑ Absorption1	☐ R
☑ Absorption2	☐ R
☑ Absorption3	☐ R

α = .88 Cronbach's α on 559 complete responses. Cronbach's α of .70 or higher is usually acceptable for research scales (Nunnally (1978)).

ω = .88 McDonald's ω from a one-factor model of these items. Unlike α, ω does not assume that every item measures the construct equally well, so it is the better estimate when the items differ (McNeish, 2018); the same .70 guide applies.

Item statistics

Item	Corrected item–total r	α if deleted
Vigor1	.68	.86
Vigor2	.64	.86
Vigor3	.63	.86
Dedication1	.66	.86
Dedication2	.64	.86
Dedication3	.60	.87
Absorption1	.57	.87
Absorption2	.57	.87
Absorption3	.60	.87

An item–total correlation below .30 (or a negative one) suggests the item does not fit, or that it needs reverse-coding.

Create the construct score

Figure 21. *Reliability of all nine items together.*

Table 7

Reliability of the engagement scales (complete responses only)

Scale	Items	α	ω	Complete responses	Corrected item–total r
Vigor	3	.83	.83	580	.71, .69, .65
Dedication	3	.84	.84	600	.72, .72, .66
Absorption	3	.79	.79	579	.61, .62, .65
All nine items	9	.88	.88	559	.57 to .68

What it means for our example

All three scales are reliable (α and ω from .79 to .84) and no item fits badly (every item–total correlation is above .56). Dropping an item would lower α in every scale, so keep all items. The nine items together give α = .88, but see the concept box: this reflects the number of items, and the three dimensions are related but not identical (section 3.4). Report the reliability of each scale, and use the total only if your theory says engagement is one overall idea.

Tip: does the questionnaire measure three dimensions?

Reliability says that items hang together; it does not say that the three dimensions are distinct. **Analyses** → **Factor analysis and reliability** tests that with an exploratory factor analysis, a confirmatory factor analysis of the three-factor model (fit indices, AVE, the Fornell–Larcker criterion and the HTMT) and a reliability table with α and ω for every scale. The Common Statistical Tests manual (chapter 12) shows these steps; run them before you rely on a new or adapted questionnaire.

3.4 Scale scores

Clicking **Create the construct score** adds the *mean of the items* as a new variable. A person gets a score only if they answered every item of the scale. That is why the scores have fewer valid values than the items: Vigor 580, Absorption 579 and Work_engagement 559 (41 people missing at least one item).

Describe the scores with **Analyses** → **Descriptive statistics**: Vigor M = 3.93 (SD = 1.32), Dedication M = 3.95 (SD = 1.33), Absorption M = 3.90 (SD = 1.30) and Work_engagement M = 3.93 (SD = 1.10); no score is clearly skewed. The scales correlate positively (Figure 22): Vigor with Dedication r = .57, Vigor with Absorption r = .54 and Dedication with Absorption r = .47 (all $p < .001$ after Holm adjustment).

Correlation matrix

	1	2	3
1. Vigor	—		
2. Dedication	.57***	—	
3. Absorption	.54***	.47***	—

* p < .05, ** p < .01, *** p < .001 (Holm-adjusted). Each pair uses every respondent with both values.

Figure 22. *Pearson correlations of the three engagement scales.*

Tip: unreliable scores need care

A score based on three items is a rough measure of a person. Group means (department, gender) are much more trustworthy than a single person's score.

3.5 Comparing departments: ANOVA and Kruskal–Wallis

Does engagement differ between departments? With five groups, use **Analyses** → **One-way ANOVA**. Choose **Score to compare** Work_engagement and **Groups** Department (Figure 23, Figure 24).

One-way ANOVA (Comparing Work_engagement across Department)

What it means: Work_engagement differs across the 5 Department groups (medium effect, ω² = .075). Pairs that differ: Finance vs Operations; HR vs Operations; HR vs Sales; IT vs Operations.

OK **Equal variances.** Levene's test does not suggest unequal variances: the classic ANOVA is used with Tukey–Kramer comparisons.

Rule: Levene's test p < .05 suggests unequal variances: rely on Welch's test and Games–Howell comparisons (Levene (1960); Games & Howell (1976)).

OK **Normality.** Every group has 30 or more cases, so the test of means is robust to non-normal data. Shapiro–Wilk: Finance W = .970, p = .069; HR W = .985, p = .602; IT W = .992, p = .826; Operations W = .988, p = .171; Sales W = .989, p = .279.

Rule: With about 30 or more cases per group, t-tests and ANOVA tolerate non-normal data; with fewer, prefer the rank-based test if the data are clearly skewed (Lumley et al. (2002)).

Info **Effect size.** ω² = .075 (medium).

Rule: η² / ω²: about .01 small, .06 medium, .14 large (Cohen (1988)).

Info **Effect-size intervals.** η² and ω² are given with 90% confidence intervals (100(1 − 2α)%), so the interval excludes 0 exactly when the test is significant at α = .05. Other effect sizes and differences use 95% intervals.

Rule: Report η², ω², partial η², Cramér's V, φ and w with a 100(1 − 2α)% interval (90% at α = .05): its lower limit is above 0 exactly when the F or χ² test is significant (Steiger (2004)).

Figure 23. *One-way ANOVA of Work_engagement by Department: what it means and the checks.*

Group descriptives

Group	n	Mean	SD	Skewness	Shapiro–Wilk W	p (normality)
Finance	75	4.12	1.11	−0.05	.970	.069
HR	68	4.56	0.99	−0.02	.985	.602
IT	97	4.13	1.06	−0.05	.992	.826
Operations	171	3.61	0.99	−0.08	.988	.171
Sales	148	3.79	1.12	0.04	.989	.279

Shapiro–Wilk test of normality per group (Shapiro & Wilk, 1965).

Copy table | Copy as picture | PNG | CSV

One-way ANOVA

Source	SS	df	MS	F	p
Between groups	54.98	4	13.75	12.38	< .001
Within groups	614.90	554	1.11		
Total	669.88	558			

Copy table | Copy as picture | PNG | CSV

Robust test and equal variances

Test	Result
Welch's F	F(4, 226.45) = 13.47, p < .001
Levene's test	F(4, 554) = 0.62, p = .647

Copy table | Copy as picture | PNG | CSV

Effect sizes

Measure	Value	90% CI	Size
η²	.082	[.044, .115]	medium
ω² (less biased)	.075	[.044, .115]	medium

CIs of η², ω² and partial η² are two-sided 90% intervals (100(1 − 2α)% with α = .05), from inverting the noncentral F distribution. This level matches the F test: the lower limit is above 0 exactly when the test is significant (Steiger, 2004). η² and ω² share one interval because both estimate the same population share of variance (Ben-Shachar et al., 2020).

Figure 24. *Group descriptives with Shapiro–Wilk, the ANOVA table, Welch's test with Levene's test and the effect sizes.*

Concept: one-way ANOVA and its rank-based partner

Analysis of variance (ANOVA) asks whether the group means differ by more than chance would produce. A significant F says that at least one group differs, not which one; *post hoc* comparisons (Tukey–Kramer, which corrects for the number of pairs) tell you which pairs differ. The *Kruskal–Wallis test* does the same job on ranks and does not assume a normal distribution; *Dunn's test* is its pairwise follow-up. SaNaStat shows both, reads Levene's test of equal variances for you and chooses the follow-up test. The effect size ω² says how much of the variance the groups explain; it is given with a 90% confidence interval, whose lower limit is above zero exactly when the F test is significant.

Cut-off values: ANOVA

Levene's test $p < .05$: use Welch's test and Games–Howell comparisons (Levene, 1960; Games & Howell, 1976). Effect size η^2 and ω^2: about .01 small, .06 medium, .14 large (Cohen, 1988).

Table 8

Work engagement by department (simulated survey)

Department	n	Mean	SD
Finance	75	4.12	1.11
HR	68	4.56	0.99
IT	97	4.13	1.06
Operations	171	3.61	0.99
Sales	148	3.79	1.12

What it means for our example

Work_engagement differs across the five departments, $F(4, 554) = 12.38$, $p < .001$, $\eta^2 = .082$, 90% CI [.044, .115], $\omega^2 = .075$ (a medium effect). Levene's test does not suggest unequal variances ($p = .647$), and Welch's test agrees, $F(4, 226.45) = 13.47$, $p < .001$. The Kruskal–Wallis test gives the same answer, $H(4) = 42.88$, $p < .001$, $\varepsilon^2 = .077$. HR is highest (4.56) and Operations lowest (3.61). The Tukey–Kramer comparisons (Figure 25) show that HR scores 0.96 [0.54, 1.37] above Operations ($p < .001$); Finance and IT are each about 0.5 above Operations ($p = .004$ and $p < .001$), and HR is 0.78 above Sales ($p < .001$). The rank-based pairwise tests (Holm-adjusted) agree on all of these.

Pairwise comparisons: Tukey–Kramer (chosen automatically: variances similar)

Pair	Mean difference	95% CI	p (adjusted)
Finance – HR	−0.44	[−0.93, 0.04]	.089
Finance – IT	−0.01	[−0.45, 0.43]	1.000
Finance – Operations	0.51	[0.12, 0.91]	.004
Finance – Sales	0.33	[−0.07, 0.74]	.168
HR – IT	0.43	[−0.02, 0.89]	.072
HR – Operations	0.96	[0.54, 1.37]	< .001
HR – Sales	0.78	[0.36, 1.20]	< .001
IT – Operations	0.52	[0.16, 0.89]	< .001
IT – Sales	0.34	[−0.03, 0.72]	.091
Operations – Sales	−0.18	[−0.50, 0.14]	.546

Tukey–Kramer (Kramer, 1956) assumes equal variances and keeps the family-wise error at α; the intervals are simultaneous 95% intervals.

Figure 25. *Tukey–Kramer pairwise comparisons of the five departments.*

The same test on the three scales gives the same picture (Table 9). The department differences are largest for Dedication and smallest for Absorption.

Table 9

ANOVA and Kruskal–Wallis test of the engagement scales by department

Scale	ANOVA	ω^2	Kruskal–Wallis H(4)	p
Vigor	F(4, 575) = 9.18, p < .001	.053	33.46	< .001
Dedication	F(4, 595) = 10.84, p < .001	.062	39.24	< .001
Absorption	F(4, 574) = 6.76, p < .001	.038	25.14	< .001
Work_engagement	F(4, 554) = 12.38, p < .001	.075	42.88	< .001

The **Charts** screen draws the group means with their 95% confidence intervals (Figure 26): Finance 4.12 [3.87, 4.37], HR 4.56 [4.32, 4.80], IT 4.13 [3.92, 4.34], Operations 3.61 [3.46, 3.76] and Sales 3.79 [3.60, 3.97].

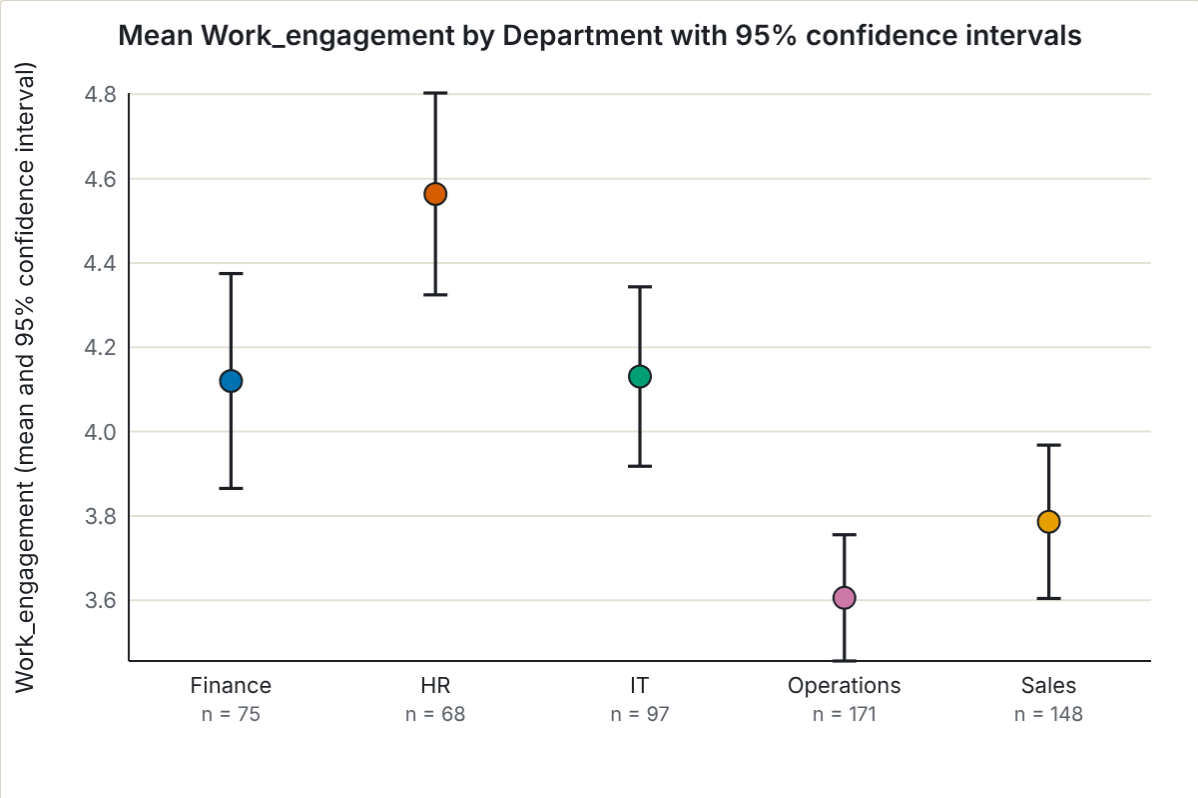


Figure 26. *Mean Work_engagement by Department with 95% confidence intervals.*

How to report it

"A one-way ANOVA showed a significant effect of Department on Work_engagement, $F(4, 554) = 12.38$, $p < .001$, $\eta^2 = .082$, 90% CI [.044, .115], $\omega^2 = .075$. Tukey–Kramer comparisons showed that HR ($M = 4.56$, $SD = 0.99$) scored higher than Operations ($M = 3.61$, $SD = 0.99$) and Sales ($M = 3.79$, $SD = 1.12$). The Kruskal–Wallis test agreed, $H(4) = 42.88$, $p < .001$."

Note: a department difference is not a manager difference

A difference between departments may come from the work, the people, the managers or the way people were chosen to answer. With survey data you cannot tell which. A department with 68 respondents is also a small group: check the minimum group size before you publish department-level results.

3.6 Comparing gender and remote work

For two groups use **Analyses** → **t-tests**, with **Score to compare** Work_engagement and **Groups** Gender. SaNaStat reports Welch's t-test first and the Mann–Whitney U test as the rank-based alternative.

What it means for our example

Women ($M = 3.92$, $SD = 1.09$, $n = 256$) and men ($M = 3.94$, $SD = 1.10$, $n = 303$) do not differ, $t(543.87) = -0.26$, $p = .798$, $d = -0.02$, 95% CI [−0.19, 0.14]; the Mann–Whitney test agrees ($U = 37820.5$, $p = .613$). The difference is female minus male, which is why the sign is negative.

Respondents who work remotely score a little higher (Yes: $M = 4.02$, $SD = 1.08$, $n = 202$; No: $M = 3.88$, $SD = 1.10$, $n = 357$), but not significantly, $t(426.12) = -1.45$, $p = .148$, $d = -0.13$, 95% CI [−0.30, 0.05]; Mann–Whitney $p = .214$. A confidence interval from −0.30 to 0.05 is compatible with no effect and with a small effect: the data are not informative enough to say more.

Tenure_band shows how the two tests can disagree near the boundary: the ANOVA is just significant, $F(3, 555) = 2.80$, $p = .039$, $\omega^2 = .010$ (negligible), while the Kruskal–Wallis test is not, $H(3) = 6.83$, $p = .078$. Do not make much of such a result.

Exercise

1. Compare Dedication by Works_remotely and write the result in one sentence. Then compare your sentence with the one SaNaStat writes under *How to report it*. (Answer: t(449.36) = −1.20, p = .233, d = −0.10, 95% CI [−0.27, 0.07].)

1. Which pair of departments differs most on Vigor, and by how much? (Answer: HR and Operations, 0.97, 95% CI [0.47, 1.48], Tukey–Kramer p < .001.)

1. Why is the Work_engagement score available for only 559 respondents when the file has 600? (Answer: a score needs all nine items; 41 respondents skipped at least one.)

4. Pay equity and adverse impact

Pay equity is one of the most common and most sensitive questions in people analytics. This chapter asks whether women and men in the simulated company earn the same, first in raw terms and then after allowing for the job they do; it splits the gap into a part explained by job-related factors and a part that is not (the Oaxaca–Blinder decomposition); and it compares promotion rates between groups with the four-fifths rule. It also explains, as plainly as we can, what such analyses can and cannot show. We use practice_hr.csv with the variable Annual_salary_ln made in section 1.9; the company, its employees and its pay are simulated.

> **Note: before you start**
> Pay data are among the most sensitive data in an organisation. Work on an anonymised copy, apply the minimum group size (Responsible people analytics), and decide before the analysis who will see the results and what they will be used for.

4.1 Two questions, not one

- **The raw gap.** Do women and men earn different amounts on average? This describes the organisation as it is.
- **The adjusted gap.** Do women and men earn different amounts when they do comparable work, that is, at the same level, with the same tenure, in the same department and with the same education? This asks whether pay differs for people who look alike on the measured characteristics.

Both are legitimate and they can differ. A large raw gap with a small adjusted gap means that women and men are in different jobs; a small raw gap with a large adjusted gap means that people in the same jobs are paid differently. The two point to different remedies. Sections 4.2 to 4.6 answer them with the general tools, step by step; section 4.7 does it in one run with the pay-gap tool.

4.2 The raw gap: t-test and Mann–Whitney test

1. Click **Analyses** → **t-tests**.
2. Choose **Score to compare** Annual_salary and **Groups** Gender. Leave **Two-tailed test**.

Figure 27 and Figure 28 show the result. Welch's t-test is reported first and the Mann–Whitney U test is the rank-based alternative, which does not depend on the skewed shape of salary.

Independent-samples t-test (Comparing Annual_salary between female and male)

What it means: male scored higher than female on Annual_salary (difference 1939.20 points). The difference is statistically significant and its size is negligible (|d| = 0.13).

Info **Which t-test.** Welch's t-test is reported first because it is accurate whether or not the variances are equal.
Rule: Compare two means with Welch's t-test by default; it stays accurate when the variances differ (Delacre et al. (2017)).

OK **Equal variances.** Levene's test does not suggest unequal variances; Student's and Welch's tests agree closely.
Rule: Levene's test p < .05 suggests unequal variances: rely on Welch's test and Games–Howell comparisons (Levene (1960); Games & Howell (1976)).

OK **Normality.** Every group has 30 or more cases, so the test of means is robust to non-normal data. Shapiro–Wilk: female W = .881, p < .001; male W = .889, p < .001. With this many cases the Shapiro–Wilk test flags even small, harmless departures; look at the Q–Q plot.
Rule: With about 30 or more cases per group, t-tests and ANOVA tolerate non-normal data; with fewer, prefer the rank-based test if the data are clearly skewed (Lumley et al. (2002)).

Info **Effect size.** |d| = 0.13 is negligible. The sign of d follows the order of the groups: d = −0.13 means male scored higher than female.
Rule: Cohen's d: about 0.2 small, 0.5 medium, 0.8 large (Cohen (1988)).

Figure 27. *Independent-samples t-test of Annual_salary by Gender: what it means and the checks.*

Group descriptives

Group	n	Mean	SD	Median	Skewness	Shapiro–Wilk W	p (normality)
female	563	44213.37	14237.12	41420.00	1.70	.881	< .001
male	637	46152.57	14849.22	42860.00	1.57	.889	< .001

Shapiro–Wilk test of normality per group (Shapiro & Wilk, 1965): p < α suggests the scores are not normally distributed.

Copy table | Copy as picture | PNG | CSV

t-tests (two-sided)

Test	t	df	p	Mean difference	95% CI
Welch's t (recommended)	−2.31	1190.08	.021	−1939.20	[−3587.92, −290.48]
Student's t (equal variances)	−2.30	1198	.022	−1939.20	[−3592.20, −286.20]

Copy table | Copy as picture | PNG | CSV

Effect size

Measure	Value	95% CI	Size
Cohen's d	−0.13	[−0.25, −0.02]	negligible
Hedges' g	−0.13	[−0.25, −0.02]	negligible

CIs of d and g are 95% intervals from the noncentral t distribution (Steiger & Fouladi, 1997); Hedges' g corrects d for small-sample bias (Hedges, 1981). The CI of the rank-biserial r uses Fisher's z (Ben-Shachar et al., 2020).

Figure 28. *The group descriptives with Shapiro–Wilk, the t-tests and the effect sizes with their confidence intervals.*

Table 10

Raw salary gap between women and men

Group	n	Mean	SD	Median
female	563	44213.37	14237.12	41420.00
male	637	46152.57	14849.22	42860.00

What it means for our example

Men earn more on average: the difference (female minus male) is −1939.20, 95% CI [−3587.92, −290.48], Welch's t(1190.08) = −2.31, p = .021, d = −0.13, 95% CI [−0.25, −0.02]. The Mann–Whitney test agrees, U = 164623.0, z = −2.45, p = .014, with a rank-biserial correlation of −.08. The effect size is negligible by Cohen's convention, but a salary gap of a few percent is not negligible to the person who earns less: statistical "size" and practical importance are different questions.

The effect-size check of the program reads: "|d| = 0.13 is negligible. The sign of d follows the order of the groups: d = −0.13 means male scored higher than female." The size is judged on the absolute value of d; the sign only tells you which group is higher. The normality check notes that Shapiro–Wilk rejects normality in both groups (W = .881 and .889), and adds that with this many cases the test of means is robust; that is why the rank-based test is shown alongside.

On the log scale the picture is the same: Welch's t(1181.44) = −2.54, p = .011, with a difference of −0.04 in log salary, 95% CI [−0.08, −0.01]. The box plot (**Charts**) shows heavily overlapping distributions: knowing that someone is a woman tells you very little about her salary.

How to report it

"An independent-samples Welch's t-test showed that Annual_salary differed between female (M = 44213.37, SD = 14237.12) and male (M = 46152.57, SD = 14849.22), t(1190.08) = −2.31, p = .021, d = −0.13, 95% CI [−0.25, −0.02]."

4.3 Do women and men do different jobs?

If the raw gap came from women and men working in different places, we would see it in the composition. Use **Analyses** → **Chi-square, Fisher, McNemar** → **Chi-square test of independence**.

- **Department by Gender:** $\chi^2(4, N = 1200) = 3.68$, p = .451, Cramér's V = .06, 90% CI [.00, .08]. The share of women ranges from 44% (Finance and Operations) to 53% (HR).
- **Job_level by Gender** (set Job_level to **Categorical** on the Data screen first): $\chi^2(4, N = 1200) = 1.61$, p = .807, Cramér's V = .04 (Figure 29). Women are 46% of level 1, 47% of levels 2 and 3, 42% of level 4 and 10 of the 17 people at level 5.

Counts (row %)

Job_level	female	male	Total
1	179 (46%)	207 (54%)	386
2	233 (47%)	258 (53%)	491
3	113 (47%)	127 (53%)	240
4	28 (42%)	38 (58%)	66
5	10 (59%)	7 (41%)	17

Copy table | Copy as picture | PNG | CSV

Tests

Test	Value	df	p
Pearson chi-square	1.61	4	.807
Likelihood ratio	1.61	4	.807

Figure 29. *Counts with row percentages and the tests for Job_level by Gender.*

What it means for our example

Neither department nor level differs significantly by gender. Women and men are distributed over departments and levels in much the same way, so the raw gap is not explained by women being concentrated in lower levels or lower-paying departments. (Level 5 has only 17 people, so do not read its 59% too closely; with the privacy setting at 20 people the level would be hidden.) Set Job_level back to **Scale** if you want it as a single straight-line predictor.

4.4 The adjusted gap: regression of log salary

A multiple regression estimates the gender difference while holding the other predictors constant. We use log salary as the outcome, so the coefficients are approximate percentage differences (section 2.4).

1. Click **Analyses** → **Correlation and regression** → **Multiple linear regression (Which variables predict a score?)**.
2. **Outcome (scale)**: Annual_salary_ln.
3. **Block 1 predictors (e.g. control variables)**: Job_level, Tenure_years, Department, Education.
4. **Block 2 predictors (optional: what do they add?)**: Gender.

Putting gender alone in the second block is a *hierarchical regression*: SaNaStat reports the model without gender, then the model with gender, and the change in R^2 (ΔR^2) that gender

adds (Figure 30, Figure 31). The reference groups (Finance, Bachelor, female) can be changed under **Reference groups**.

Multiple linear regression (Predicting Annual_salary_ln)

What it means: The predictors together explain 83.8% of the variance in Annual_salary_ln (R^2 = .84, f^2 = 5.19, large). Significant predictors: Department: HR (vs Finance) (β = −.09), Department: IT (vs Finance) (β = .09), Department: Operations (vs Finance) (β = −.24), Department: Sales (vs Finance) (β = −.16), Job_level (β = .68), Education: Diploma (vs Bachelor) (β = −.07), Education: Doctorate (vs Bachelor) (β = .07), Education: Master (vs Bachelor) (β = .08), Tenure_years (β = .18), Gender: male (vs female) (β = .08). Block 2 adds ΔR^2 = .006 (p < .001).

OK **Sample size.** 1200 respondents for 10 predictor terms (needed: 130).
Rule: At least 50 + 8k respondents for k predictors to test the whole model (Green (1991)).

OK **Collinearity.** Largest VIF = 3.18.
Rule: VIF below 5: no concern; 5–10: moderate; above 10: serious collinearity (Hair et al. (2019); Kutner et al. (2005)).

OK **Normal residuals.** Shapiro–Wilk W = 0.998, p = .116.
Rule: Shapiro–Wilk p ≥ .05 for the residuals; with about 300 or more respondents the coefficient tests are robust anyway (Shapiro & Wilk (1965)).

OK **Constant error variance.** Breusch–Pagan (Koenker) $\chi^2(10)$ = 17.57, p = .063.
Rule: Breusch–Pagan (Koenker) p ≥ .05: no evidence that the error spread changes (Breusch & Pagan (1979); Koenker (1981)).

OK **Influential cases.** Largest Cook's distance = 0.020; 1 standardised residual beyond ±3.
Rule: No Cook's distance above 1 (Cook & Weisberg (1982)).

Info **Independent errors.** Durbin–Watson = 1.98 (values near 2 mean no order effect; it matters only when the row order means something, such as time).

Info **Effect size.** f^2 = 5.19 (large).
Rule: f^2: .02 small, .15 medium, .35 large (Cohen (1988)).

Figure 30. *Regression of Annual_salary_ln: what it means and the checks.*

Model summary

Model	Predictors	R^2	Adjusted R^2	F	df	p	ΔR^2	p (ΔR^2)
1	Department, Job_level, Education, Tenure_years	.833	.831	657.72	9, 1190	< .001		
2	Department, Job_level, Education, Tenure_years, Gender	.838	.837	616.53	10, 1189	< .001	.006	< .001

Copy table | Copy as picture | PNG | CSV

Coefficients (model 2)

Predictor	B	SE	β	t	p	95% CI of B	VIF
(Intercept)	10.214	0.012		817.32	< .001	[10.190, 10.239]	
Department: HR (vs Finance)	−0.087	0.014	−.088	−6.18	< .001	[−0.115, −0.060]	1.49
Department: IT (vs Finance)	0.070	0.012	.089	5.75	< .001	[0.046, 0.094]	1.75
Department: Operations (vs Finance)	−0.154	0.011	−.243	−14.31	< .001	[−0.175, −0.133]	2.12
Department: Sales (vs Finance)	−0.100	0.011	−.157	−9.30	< .001	[−0.122, −0.079]	2.10
Job_level	0.212	0.006	.681	32.76	< .001	[0.200, 0.225]	3.18
Education: Diploma (vs Bachelor)	−0.048	0.009	−.069	−5.57	< .001	[−0.066, −0.031]	1.14
Education: Doctorate (vs Bachelor)	0.105	0.019	.069	5.49	< .001	[0.067, 0.142]	1.17
Education: Master (vs Bachelor)	0.052	0.011	.077	4.90	< .001	[0.031, 0.073]	1.80
Tenure_years	0.013	0.001	.182	10.08	< .001	[0.011, 0.016]	2.40
Gender: male (vs female)	0.044	0.007	.076	6.48	< .001	[0.031, 0.058]	1.01

n = 1200. Categorical predictors are dummy-coded against a reference group: Department = Finance, Education = Bachelor, Gender = female. β = standardised coefficient.

Figure 31. *The model summary of both blocks and the coefficients of model 2.*

Concept: what a regression coefficient means

A coefficient B is the average change in the outcome for one unit more of the predictor, holding the other predictors fixed. For a categorical predictor, the coefficient compares a group with the *reference group*: "Gender: male (vs female)" is men minus women. Because the outcome is log salary, B multiplied by 100 is roughly the percentage difference. R^2 is the share of the variance explained.

Table 11

Regression of Annual_salary_ln (n = 1200). The last column was calculated by hand as e^B − 1.

Predictor	B	95% CI of B	p	Approx. %
Job_level (per level)	0.212	[0.200, 0.225]	< .001	+23.6
Tenure_years (per year)	0.013	[0.011, 0.016]	< .001	+1.3
Department: HR (vs Finance)	−0.087	[−0.115, −0.060]	< .001	−8.3
Department: IT (vs Finance)	0.070	[0.046, 0.094]	< .001	+7.3
Department: Operations (vs Finance)	−0.154	[−0.175, −0.133]	< .001	−14.3
Department: Sales (vs Finance)	−0.100	[−0.122, −0.079]	< .001	−9.5
Education: Diploma (vs Bachelor)	−0.048	[−0.066, −0.031]	< .001	−4.7
Education: Master (vs Bachelor)	0.052	[0.031, 0.073]	< .001	+5.3
Education: Doctorate (vs Bachelor)	0.105	[0.067, 0.142]	< .001	+11.1
Gender: male (vs female)	0.044	[0.031, 0.058]	< .001	+4.5

What it means for our example

Job level, tenure, department and education together explain much of the variance in log salary (model 1: R^2 = .833, F(9, 1190) = 657.72, p < .001). Adding gender raises R^2 to .838 (model 2: F(10, 1189) = 616.53, ΔR^2 = .006, p < .001). The coefficient for gender is B = 0.044 (SE = 0.007), t(1189) = 6.48, p < .001, 95% CI [0.031, 0.058]: **men earn about 4.5% more than women at the same level, with the same tenure, department and education**, with a 95% confidence interval from about 3.1% to 6.0% (calculated by hand). This is the gap that was planted in the data (about 4%). Each level adds about 24%, each year of tenure about 1.3%, and Operations pays about 14% less than Finance at the same level.

The checks under the result are all acceptable: no coefficient is hampered by collinearity (largest VIF = 3.18), the residuals are close to normal (Shapiro–Wilk W = 0.998, p = .116), the error variance is constant (Breusch–Pagan p = .063) and no case is influential (largest Cook's distance 0.020). That the log scale works well is part of the reason for using it; the same model on raw salary gives a male coefficient of 1935.101 salary units, 95% CI [1233.356, 2636.846], t = 5.41.

How to report it

"A hierarchical multiple linear regression of log annual salary on job level, tenure, department and education (block 1, R^2 = .83) and gender (block 2) showed that men earned 4.5% more than women with the same level, tenure, department and education, B = 0.044, 95% CI [0.031, 0.058], t(1189) = 6.48, p < .001, ΔR^2 = .006."

4.5 Is the result robust?

A single regression is one set of choices. Try the reasonable alternatives and see whether the gender coefficient moves.

- **Job level as groups.** Set Job_level to **Categorical** on the Data screen and repeat the regression with all predictors in one block. Each level gets its own coefficient: 0.212 (level 2), 0.431 (level 3), 0.626 (level 4) and 0.841 (level 5), all against level 1. The steps are close to equal, so treating level as a straight line was reasonable. The gender coefficient is unchanged, B = 0.044, t(1186) = 6.47, 95% CI [0.031, 0.058]. Set Job_level back to **Scale** afterwards.
- **More controls.** Adding Performance_rating and Age leaves the gender coefficient at B = 0.044, 95% CI [0.031, 0.058]; neither Age (p = .440) nor Performance_rating (p = .615) predicts salary once level is in the model.
- **Fewer controls.** Leaving Job_level out gives a gender coefficient of 0.037, 95% CI [0.019, 0.056], and R^2 falls from .84 to .69. The gap is similar, but the model is much poorer, because level is by far the strongest predictor of pay.

Note: "adjusted" is not the same as "fair"

Whether to control for a variable is a judgement about causes, not a statistical question. If women are held back from higher levels, then level is *part of* the gap, and controlling for it hides that part. Conversely, if level reflects real differences in responsibility, controlling for it is right. Report the gap with and without the contested controls, and say which you consider appropriate and why.

4.6 Does the gap differ by department?

A factorial ANOVA of Annual_salary_ln by Gender and Department (**Analyses** → **Factorial ANOVA, ANCOVA, MANOVA**, with Gender and Department ticked under the grouping factors; Figure 32) asks whether the gap differs between departments. It does not adjust for level or tenure; it asks only whether the raw log-salary gap is the same everywhere.

Factorial ANOVA (Type III sums of squares)

Source	SS	df	MS	F	p	Partial η²	90% CI
Department	7.86	4	1.96	25.22	< .001	.078	[.053, .101]
Gender	0.41	1	0.41	5.25	.022	.004	[.000, .013]
Department × Gender	0.07	4	0.02	0.22	.925	.001	[.000, .001]
Error	92.68	1190	0.08				
Total (corrected)	101.20	1199					

R^2 = .084, adjusted R^2 = .077, n = 1200. CIs of η^2, ω^2 and partial η^2 are two-sided 90% intervals (100(1 − 2α)% with α = .05), from inverting the noncentral F distribution. This level matches the F test: the lower limit is above 0 exactly when the test is significant (Steiger, 2004).

Figure 32. *Factorial ANOVA of Annual_salary_ln by Department and Gender (Type III sums of squares).*

What it means for our example

There is a main effect of Department, $F(4, 1190) = 25.22$, $p < .001$, partial $\eta^2 = .078$, and a small main effect of Gender, $F(1, 1190) = 5.25$, $p = .022$, partial $\eta^2 = .004$, but no interaction, $F(4, 1190) = 0.22$, $p = .925$: the data give no sign that the gap is larger in one department than another. Every cell has at least 53 people. A non-significant interaction does not prove that the gap is the same everywhere; with five departments the test has limited power to find a difference in one of them.

4.7 The pay-gap tool: raw gap, adjusted gap and the Oaxaca–Blinder decomposition

Analyses → **Adverse impact and pay gaps** → **Pay gap between two groups (How much of the gap do job-related factors explain?)** does the work of sections 4.2 and 4.4 in one run and adds a decomposition of the gap (Figure 33).

1. **Pay (scale)**: Annual_salary.

2. **Group (e.g. gender)**: Gender. **Group A (usually the higher-paid group)**: male. **Group B (compared with A)**: female.
3. Under **Job-related factors to control for (optional)** tick Job_level, Tenure_years, Department and Education. The reference groups of the categorical factors can be changed below the list.
4. Leave **Analyse the logarithm of pay (gap in percent)** ticked, the pay structure at **Both groups together (recommended)** and **Bootstrap resamples for the intervals** at 1000.
5. Click **Run the pay-gap analysis**. The bootstrap takes a few seconds.

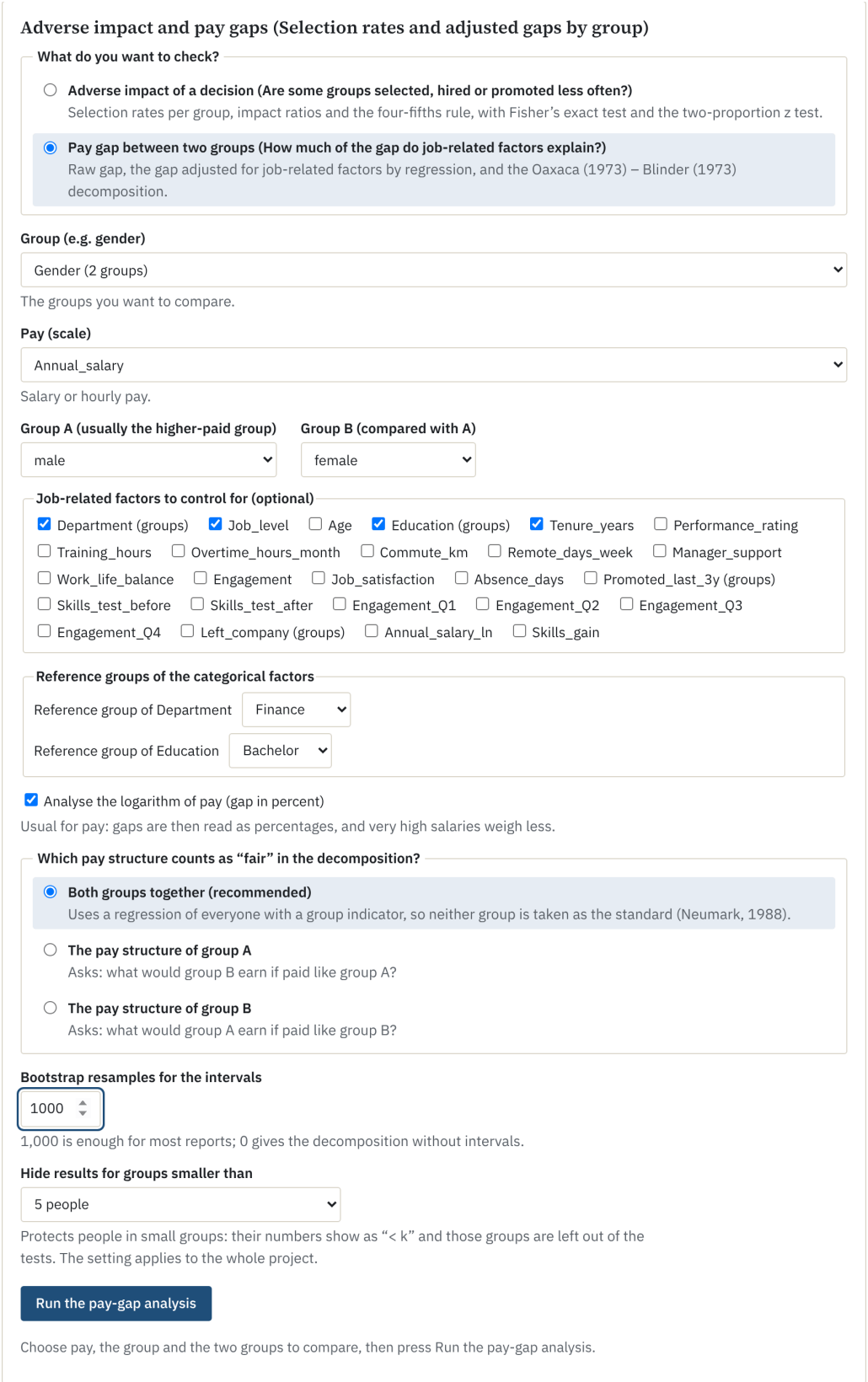


Figure 33. *The pay-gap tool: pay, the two groups, the job-related factors, the logarithm, the pay structure and the bootstrap.*

Concept: the Oaxaca–Blinder decomposition

The decomposition (Oaxaca, 1973; Blinder, 1973) splits the average pay gap between two groups into two parts. The *explained* part is what the gap would be if the two groups differed only in their characteristics (level, tenure, department, education), paid at one common pay structure. The *unexplained* part is the rest: the difference in pay for the same characteristics. The choice of the common pay structure matters; SaNaStat recommends a regression of both groups together with a group indicator, so that neither group is taken as the standard (Neumark, 1988). The table *Contribution of each variable* splits both parts by variable.

Pay gap (Annual_salary: female vs male)

What it means: Before controls, pay in group female is 4.2% lower on average than in group male; with the same Department, Job_level, Education, Tenure_years it is 4.3% lower (adjusted p < .001). Of the gap, −3.7% is explained by differences in Department, Job_level, Education, Tenure_years and 103.7% remains unexplained.

Info **What the gap means.** An adjusted or unexplained gap is not proof of unfair pay: it also contains everything the controls leave out. A control that is itself affected by unfair treatment (e.g. job level) can hide part of the gap.

Rule: The unexplained part of an Oaxaca–Blinder decomposition is not proof of discrimination: it also contains the effect of anything the model leaves out (Oaxaca (1973); Blinder (1973)).

OK **Group sizes.** 637 male and 563 female.

Figure 34. *The result: what it means and the checks.*

Pay of male and female

Gender	n	Mean	Median
male	637	46152.57	42860.00
female	563	44213.37	41420.00

People with complete data on Annual_salary and the control variables.

Copy table | Copy as picture | PNG | CSV

Raw and adjusted gap

Gap	female − male (log)	In words	SE	p	95% CI
Raw (no controls)	−0.043	−4.2%			
Adjusted for Department, Job_level, Education, Tenure_years	−0.044	−4.3%	0.007	< .001	[−0.058, −0.031]

Adjusted gap = coefficient of the group indicator in a linear regression of log Annual_salary on the group and the controls (R^2 = .84, n = 1200). In words: percentage difference of the geometric means, 100 × (e^d − 1).

Figure 35. *The pay of the two groups, and the raw and adjusted gap.*

Oaxaca–Blinder decomposition of the gap (male − female)

Part	Estimate	95% bootstrap CI	Share of the gap
Total gap	0.043	[0.009, 0.075]	100.0%
Explained (differences in characteristics)	−0.002	[−0.033, 0.028]	−3.7%
Unexplained (differences in pay for the same characteristics)	0.044	[0.030, 0.057]	103.7%

On the log scale: a value d means about 100 × (e^d − 1)% difference. Coefficients for the "fair" pay structure: a pooled regression of both groups with a group indicator. Percentile bootstrap with 1000 resamples within groups.

Copy table | Copy as picture | PNG | CSV

Contribution of each variable

Variable	Mean (male)	Mean (female)	Explained	Unexplained
Constant (base pay)			0.000	0.016
Department: HR (vs Finance)	0.08	0.11	0.002	0.000
Department: IT (vs Finance)	0.16	0.17	−0.000	0.006
Department: Operations (vs Finance)	0.32	0.29	−0.005	0.004
Department: Sales (vs Finance)	0.28	0.30	0.002	0.009
Job_level	2.03	2.04	−0.002	0.015
Education: Diploma (vs Bachelor)	0.23	0.21	−0.001	0.006
Education: Doctorate (vs Bachelor)	0.03	0.05	−0.002	−0.001
Education: Master (vs Bachelor)	0.24	0.25	−0.001	−0.002
Tenure_years	5.71	5.33	0.005	−0.008

The detailed split of the unexplained part depends on which category is the reference of a categorical control; read it with care.

Figure 36. *The Oaxaca–Blinder decomposition with bootstrap intervals, and the contribution of each variable.*

Figure 34, Figure 35 and Figure 36 show the result. SaNaStat summarises: *Before controls, pay in group female is 4.2% lower on average than in group male; with the same Department, Job_level, Education, Tenure_years it is 4.3% lower (adjusted p < .001). Of the gap, −3.7% is explained by differences in Department, Job_level, Education, Tenure_years and 103.7% remains unexplained.*

Table 12

The pay gap between women and men (log scale; from the program)

Part	Estimate	95% CI	Share of the gap
Raw gap, female − male	−0.043 (−4.2%)	–	–
Adjusted gap, female − male	−0.044 (−4.3%)	[−0.058, −0.031]	–
Total gap, male − female	0.043	[0.009, 0.075]	100.0%
Explained by the four factors	−0.002	[−0.033, 0.028]	−3.7%
Unexplained	0.044	[0.030, 0.057]	103.7%

What it means for our example

The raw gap is 4.2% and the gap at the same level, tenure, department and education is 4.3%: the adjusted gap is the gender coefficient of section 4.4 (b = −0.044, t(1189) = −6.48), with the sign of female minus male. The decomposition says the same thing in another way. Women and men hardly differ in their characteristics (both average about level 2.03 to 2.04; men have a little more tenure, 5.71 against 5.33 years), so the explained part is close to zero and its interval [−0.033, 0.028] includes zero. Slightly more than the whole gap is unexplained (103.7%; a share above 100% means that the characteristics, taken together, would favour women very slightly). The unexplained part, 0.044 log points, has a bootstrap interval of [0.030, 0.057], well away from zero. In this company the gap is not about who holds which job; it is about pay within comparable jobs.

The check *What the gap means* repeats the most important caution: "An adjusted or unexplained gap is not proof of unfair pay: it also contains everything the controls leave out. A control that is itself affected by unfair treatment (e.g. job level) can hide part of the gap." Two variants are worth a look:

- Untick **Analyse the logarithm of pay (gap in percent)** to see the gap in salary units: raw −1939.20, adjusted −1935.10, 95% CI [−2636.846, −1233.356]; 0.2% explained and 99.8% unexplained.
- Choose **The pay structure of group A** (what would women earn if paid like men?): the explained share is then −3.4% and 103.4% is unexplained. The conclusion does not depend on the choice here; when it does, report both.

How to report it

"The raw annual salary gap between women and men was 4.2% (−0.043 log points). Controlling for job level, tenure, department and education, the gap was b = −0.044, SE = 0.007, t(1189) = −6.48, p < .001, 95% CI [−0.058, −0.031]. An Oaxaca–Blinder twofold decomposition with a pooled pay structure (Neumark, 1988) attributed −0.002 log points to differences in characteristics and 0.044 to unexplained differences (95% bootstrap CI [0.030, 0.057], 1,000 resamples)."

Note: the detailed decomposition

The table *Contribution of each variable* shows how much each control adds to each part. The program warns: "The detailed split of the unexplained part depends on which category is the reference of a categorical control; read it with care." Read the explained column (which does not have this problem) and the total of the unexplained part; do not interpret the unexplained part of a single category.

4.8 Promotion: selection rates and the four-fifths rule

Pay is one outcome of HR decisions; selection is another: who is hired, promoted, given training or a bonus. *Adverse impact* is a substantially lower selection rate for one group. practice_hr.csv contains **no applicant or hiring data**, so we use the one selection decision it records, promotion in the last three years. For hiring, your file would need one row per applicant with the group and whether the person was hired; the steps are the same.

1. **Analyses** → **Adverse impact and pay gaps** → **Adverse impact of a decision (Are some groups selected, hired or promoted less often?)**.
2. **Group (e.g. gender)**: Gender. **Decision (two answers, or coded 0/1)**: Promoted_last_3y. **Which answer means "selected"?**: Yes.
3. Leave **Compare every group with** at **The group with the highest rate (four-fifths rule)** (Figure 37).

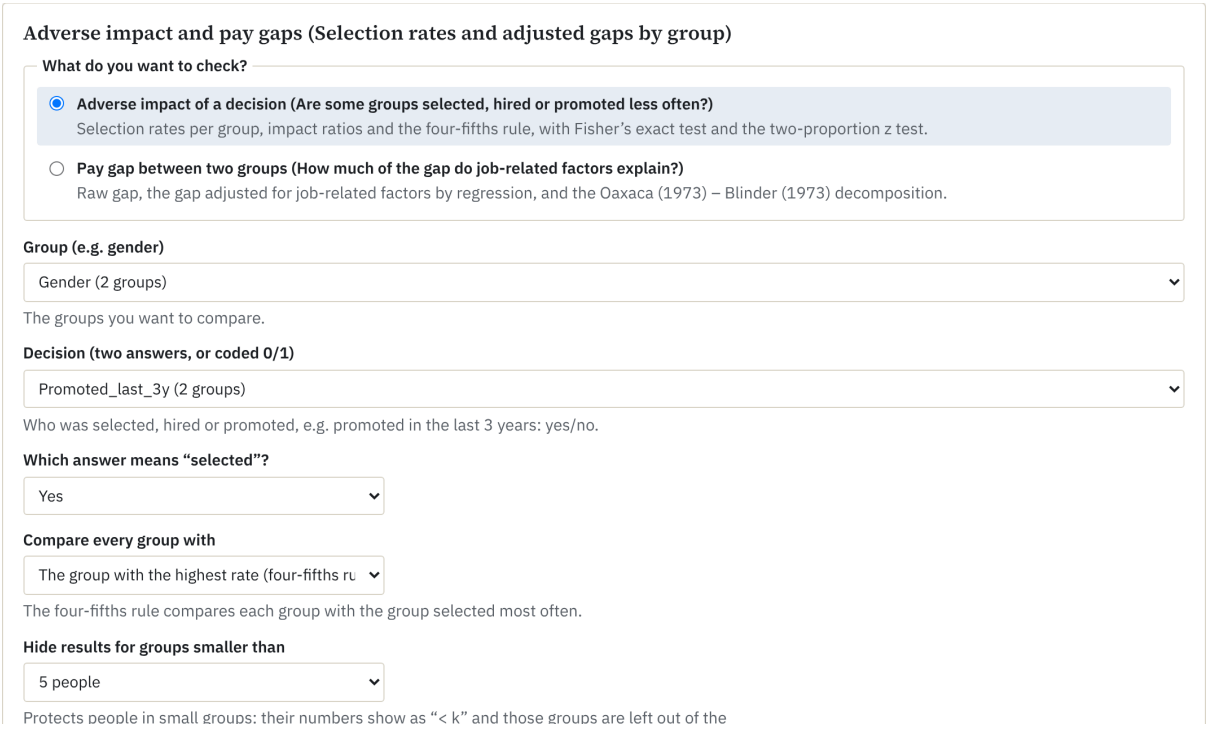


Figure 37. *The adverse-impact check: the group, the decision, the answer that means selected, and the comparison group.*

Concept: selection rate, impact ratio and the four-fifths rule

The *selection rate* of a group is the share of its members who were selected. The *impact ratio* divides a group's rate by the rate of the group selected most often. The *four-fifths rule* of the Uniform Guidelines on Employee Selection Procedures (Equal Employment Opportunity Commission et al., 1978) regards a ratio below .80 as evidence of adverse impact. It is a rule of thumb: with small groups a ratio can fall below .80 by chance, and with large groups a small difference can be significant while the ratio stays above .80. SaNaStat therefore adds the difference in percentage points with a confidence interval (Newcombe, 1998), a two-proportion z test and Fisher's exact test (Morris & Lobsenz, 2000).

Adverse impact (Promoted_last_3y = Yes by Gender)

What it means: Every group's selection rate is at least four-fifths of the rate of female (11.5%): no sign of adverse impact by the four-fifths rule.

OK **Four-fifths rule.** All impact ratios are .80 or higher.
Rule: Four-fifths rule: a selection rate for any group below 80% of the rate of the group with the highest rate is generally regarded as evidence of adverse impact; with small groups, add a significance test (Uniform Guidelines (1978); Morris & Lobsenz (2000)).

OK **Group sizes.** Every group has at least 30 people.

Figure 38. *Adverse impact of promotion by Gender: what it means and the checks.*

Selection rates and impact ratios

Gender	People	Selected	Rate	Impact ratio	Four-fifths rule	Difference (points) [95% CI]	p (z test)	p (Fisher's exact)
female (reference)	563	65	11.5%	1.00				
male	637	69	10.8%	0.94	Meets	−0.7 [−4.3, 2.9]	.695	.714

Impact ratio = the group's selection rate divided by the rate of female (the highest rate). Difference in percentage points with a Newcombe (1998) interval; z test with the pooled proportion; Fisher's exact test two-sided.

Figure 39. *Selection rates and impact ratios by Gender.*

What it means for our example

65 of 563 women (11.5%) and 69 of 637 men (10.8%) were promoted (Figure 38, Figure 39). Women have the higher rate, so they are the comparison group, and the impact ratio for men is 0.94: it *Meets* the four-fifths rule. The difference is −0.7 percentage points, 95% CI [−4.3, 2.9], z test p = .695, Fisher's exact p = .714. There is no sign of adverse impact by gender in promotion. The chi-square test of Gender by Promoted_last_3y gives the same answer, $\chi^2(1, N = 1200) = 0.15$, p = .695, Cramér's V = .01.

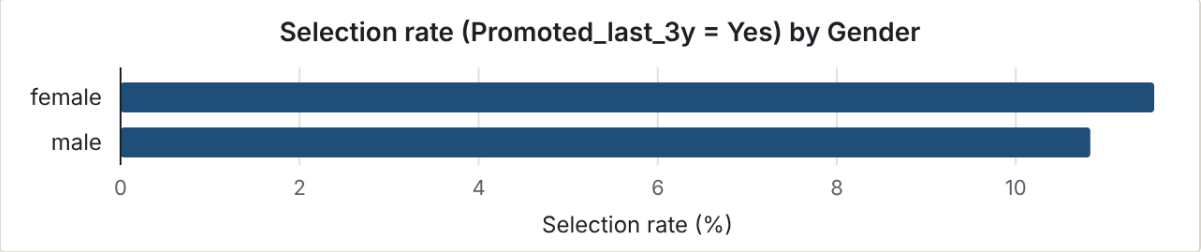


Figure 40. *The selection rates as a chart.*

Change **Group (e.g. gender)** to Department (Figure 41). HR has the highest rate (14.2%); IT (7.7%, impact ratio 0.54) and Operations (9.1%, ratio 0.64) fall *Below .80*. SaNaStat flags them with *Check* and adds: "None of these differences is statistically significant, so with these group sizes it may be chance; watch it over time." (Fisher's exact p = .079 for IT and p = .155 for Operations.) By Education the picture is clearer: Diploma holders have the highest rate (15.4%), and Bachelor holders are promoted at 8.7%, an impact ratio of 0.57, a difference of −6.6 percentage points, 95% CI [−11.8, −2.0], Fisher's exact p = .006.

Selection rates and impact ratios

Department	People	Selected	Rate	Impact ratio	Four-fifths rule	Difference (points) [95% CI]	p (z test)	p (Fisher's exact)
Finance	179	23	12.8%	0.91	Meets	−1.3 [−10.0, 6.4]	.749	.860
HR (reference)	113	16	14.2%	1.00				
IT	195	15	7.7%	0.54	Below .80	−6.5 [−14.6, 0.5]	.069	.079
Operations	363	33	9.1%	0.64	Below .80	−5.1 [−13.1, 1.2]	.122	.155
Sales	350	47	13.4%	0.95	Meets	−0.7 [−9.0, 5.9]	.844	.875

Impact ratio = the group's selection rate divided by the rate of HR (the highest rate). Difference in percentage points with a Newcombe (1998) interval; z test with the pooled proportion; Fisher's exact test two-sided.

Figure 41. *Selection rates and impact ratios by Department: two departments fall below .80.*

Note: adverse impact is a signal, not a verdict

A ratio below .80 asks a question; it does not answer it. Promotion depends on eligibility: tenure, performance and the number of openings at each level. In the simulated company promotion was planted to depend on performance, tenure and training, not on department or education, so the Bachelor result must come from differences in those factors between the groups or from chance, and a real analysis would check the rates within comparable groups (for example a logistic regression of promotion with the eligibility factors and the group). Report the rates, the ratios and the tests, say what eligibility means for the decision, and involve the people responsible for equality before anyone draws conclusions.

How to report it

"Promotion rates over the last three years were 11.5% for women (65/563) and 10.8% for men (69/637). Compared with women, the impact ratio for men was 0.94, above the four-fifths threshold; difference = −0.7 percentage points, 95% CI [−4.3, 2.9], Fisher's exact p = .714."

4.9 What these analyses cannot show

- **Cause.** A gap of 4.5% that remains after the controls does not tell you why. It may reflect discrimination, differences in negotiation, in performance that the file does not capture, in working hours or in the history of each person's pay. The regression and the decomposition show that a difference exists; they cannot name the reason.
- **Unmeasured differences.** Anything not in the model (the quality of work, the starting salary, the labour market at the time of hiring) is invisible, and sits in the unexplained part. Adding all available variables does not make the model complete.
- **Individuals.** The result is an average. It does not say that any particular woman is underpaid, and it must not be used to adjust one person's salary without a case-by-case review.
- **Small groups.** In a company with few women at senior levels, the confidence interval at each level is wide. Report the number of people behind each estimate, and use the privacy setting.

- **Other forms of pay.** Bonuses, allowances and benefits are separate questions; so are part-time work and pay per hour.

Exercise

1. Repeat the regression of section 4.4 with Gender in block 1 alongside the other predictors (one block). What is the male coefficient? (Answer: B = 0.044, the same, because a single block gives the same coefficient as the last block.)

1. What is the planted gap? Does the 95% interval for the coefficient contain it? (Answer: the planted gap is about 4%; the interval 3.1% to 6.0% contains it.)

1. Run the pay-gap tool with Job_level as the only control. Is more or less of the gap explained than with the four controls, and why?

1. Run the adverse-impact check of promotion by Education. Which group is the comparison group, and which group has a significant difference? (Answer: Diploma, 15.4%; Bachelor, 8.7%, Fisher's exact p = .006.)

1. Why is it wrong to conclude from this chapter that "the company discriminates", or that "it does not"?

5. Training evaluation

A company trains its employees, measures a skills test before and after, and wants to know whether the training worked. This chapter shows the usual first answer (a paired test), what else the data allow (does more training go with more gain?) and what the design cannot prove. Data: practice_hr.csv, columns Skills_test_before, Skills_test_after and Training_hours, all simulated.

5.1 The data

Descriptive statistics of the three variables: Skills_test_before M = 58.13 (SD = 11.07), Skills_test_after M = 63.74 (SD = 12.74) and Training_hours M = 23.91 (SD = 13.52, median 21.00, range 1.00 to 96.00, skewness 1.01).

5.2 Did scores change? Paired t-test and Wilcoxon test

The same people were measured twice, so use the paired test, not the independent one.

1. Click **Analyses** → **t-tests** and choose **Paired-samples t-test (The same people measured twice)**.
2. **First measurement**: Skills_test_before. **Second measurement**: Skills_test_after. The change is second minus first.

Descriptives (complete pairs)

Measure	n	Mean	SD	Median	Shapiro–Wilk W	p (normality)
Skills_test_before	1200	58.13	11.07	58.10		
Skills_test_after	1200	63.74	12.74	63.90		
Difference (Skills_test_after – Skills_test_before)	1200	5.61	6.19	5.60	.998	.357

The paired t-test assumes that the differences are roughly normal, so Shapiro–Wilk (Shapiro & Wilk, 1965) is shown for the differences.

Copy table | Copy as picture | PNG | CSV

Paired t-test (two-sided)

t	df	p	Mean difference	95% CI
31.37	1199	< .001	5.61	[5.26, 5.96]

Copy table | Copy as picture | PNG | CSV

Effect size

Measure	Value	95% CI	Size
Cohen's d_z	0.91	[0.84, 0.97]	large
Hedges' g_z	0.91	[0.84, 0.97]	large

d_z = mean change / SD of the changes. CIs of d and g are 95% intervals from the noncentral t distribution (Steiger & Fouladi, 1997); Hedges' g corrects d for small-sample bias (Hedges, 1981). The CI of the rank-biserial r uses Fisher's z (Ben-Shachar et al., 2020).

Copy table | Copy as picture | PNG | CSV

Rank-based alternative

Test	V	z	p	Matched rank-biserial r	95% CI	Method
Wilcoxon signed-rank	638716.0	23.98	< .001	.80	[.78, .82]	normal approximation

10 pairs with no change were left out of the signed-rank test.

Figure 42. *Paired-samples t-test of the skills test before and after training: descriptives with Shapiro–Wilk for the differences, the test, the effect sizes and the Wilcoxon test.*

Concept: paired test and d_z

A paired test works on each person's *difference* (after minus before). It asks whether the mean difference is zero. Because it compares each person with themselves, it removes the differences between people, which are large here (SD = 11.07), and focuses on the change (SD = 6.19). The effect size d_z is the mean change divided by the standard deviation of the changes. The Wilcoxon signed-rank test is the rank-based alternative.

What it means for our example

Scores rose by 5.61 points on average (SD of the change = 6.19), t(1199) = 31.37, p < .001, 95% CI of the mean change [5.26, 5.96], d_z = 0.91, 95% CI [0.84, 0.97] (a large effect). The differences are close to normal (Shapiro–Wilk W = .998, p = .357). The Wilcoxon signed-rank test agrees, V = 638716.0, z = 23.98, p < .001, matched rank-biserial r = .80; 10 pairs with no change were left out of that test.

How to report it

"A paired-samples t-test showed that Skills_test_after (M = 63.74, SD = 12.74) differed from Skills_test_before (M = 58.13, SD = 11.07), t(1199) = 31.37, p < .001, d_z = 0.91, 95% CI [0.84, 0.97]."

5.3 Does more training go with more gain?

If the training works, more hours of training should go with a larger gain. Use the variable Skills_gain made in section 1.9 (Skills_test_after − Skills_test_before). The paired t-test of section 5.2 found an average gain of 5.61 points. Now ask who gains more.

Gain by department. Skills_gain has means of 5.96 (Finance), 6.02 (HR), 5.29 (IT), 5.39 (Operations) and 5.71 (Sales). A one-way ANOVA finds no difference between departments, F(4, 1195) = 0.54, p = .707, ω^2 = −.002 (Welch's F(4, 444.84) = 0.52, p = .723). The training worked about equally everywhere.

Gain and hours. **Correlation and regression** → **Bivariate and partial correlation** with Training_hours and Skills_gain gives r(1198) = .15, 95% CI [.09, .20], p < .001. A regression of Skills_gain on Training_hours gives B = 0.067 (SE = 0.013), t(1198) = 5.12, p < .001, 95% CI [0.041, 0.093], R^2 = .021: every additional ten hours of training goes with 0.67 more points of gain (10 × 0.067, calculated by hand), but hours explain only 2.1% of the variance in the gain.

Concept: gain scores and the starting level

A gain score subtracts the starting level, but people who start high have less room to rise, and a high or low first score partly reflects chance, which tends to move back towards the average on the second test (*regression to the mean*; Barnett et al., 2005). The usual remedy is to include the before score as a predictor of the after score, as in section 5.4. Here the two approaches agree because the before score has a coefficient close to 1 (1.006), so the gain score does not depend on the starting level.

Two other ways to answer the same question use the original columns.

Correlations. With Training_hours, Skills_test_before and Skills_test_after, hours are unrelated to the before score (r = .01, p = .761) and slightly related to the after score (r = .08, 95% CI [.02, .13], Holm-adjusted p = .013); before and after correlate strongly (r = .87, 95% CI [.86, .89]).

Partial correlation. Tick Training_hours and Skills_test_after, and under **Control for (optional: partial correlation)** tick Skills_test_before (Figure 43). The correlation between hours and the after score, controlling for the before score, is the correlation between hours and *the part of the after score not predicted by the before score*, which is the change. It is r(1197) = .15, 95% CI [.09, .20], p < .001.

Partial correlation (Pearson r, controlling for Skills_test_before)

What it means: 1 of 1 partial correlations are significant after Holm adjustment. The strongest is Training_hours with Skills_test_after (r = .15, small, higher goes with higher).

Info **Strength.** Read the size of a correlation from its absolute value.
Rule: Correlation-type effect sizes (r, rank-biserial, φ): about .10 small, .30 medium, .50 large (Cohen (1988)).

Info **Method.** Pearson's r measures straight-line relationships. With strongly skewed data or many tied ratings, Spearman's ρ is a safer choice.

Correlation matrix

	1	2
1. Training_hours	—	
2. Skills_test_after	.15***	—

* p < .05, ** p < .01, *** p < .001 (Holm-adjusted). Complete cases only.

Figure 43. *Partial correlation of Training_hours with Skills_test_after, controlling for Skills_test_before.*

The hours alone hide the effect, because most of the variation in the after score is the person's starting level. Controlling for the starting level reveals it.

5.4 Regression

A hierarchical regression says the same in the units of the test (Figure 44). Choose **Multiple linear regression**, **Outcome (scale)** Skills_test_after, block 1 Skills_test_before, block 2 Training_hours.

Model summary

Model	Predictors	R^2	Adjusted R^2	F	df	p	ΔR^2	p (ΔR^2)
1	Skills_test_before	.764	.764	3872.34	1, 1198	< .001		
2	Skills_test_before, Training_hours	.769	.768	1989.95	2, 1197	< .001	.005	< .001

Copy table | Copy as picture | PNG | CSV

Coefficients (model 2)

Predictor	B	SE	β	t	p	95% CI of B	VIF
(Intercept)	3.681	0.995		3.70	< .001	[1.729, 5.633]	
Skills_test_before	1.006	0.016	.873	62.83	< .001	[0.974, 1.037]	1.00
Training_hours	0.067	0.013	.071	5.12	< .001	[0.041, 0.093]	1.00

n = 1200. β = standardised coefficient.

Figure 44. *Regression of Skills_test_after on Skills_test_before (block 1) and Training_hours (block 2).*

What it means for our example

The before score alone explains most of the variance in the after score (R^2 = .764, F(1, 1198) = 3872.34). Training_hours adds ΔR^2 = .005 ($p < .001$), R^2 = .769. The coefficient for Training_hours is B = 0.067 (SE = 0.013), t(1197) = 5.12, $p < .001$, 95% CI [0.041, 0.093]: with the same before score, every additional ten hours of training goes with 0.67 more points on the after test. The coefficient of the before score is 1.006: someone who scored ten points higher before scores about ten points higher after. The scatter plot of hours against the after score shows how weak the raw relation is (r = .08, slope 0.074); the effect is only visible once the starting level is controlled.

5.5 What the design cannot show

- **No control group.** A rise from before to after may reflect the training, but also practice on the same test, the time of the year, other development at work, or *regression to the mean*: people chosen for training because of low scores tend to score higher on any retest (Barnett et al., 2005). The paired test cannot tell these apart. A comparison group of similar employees who were not trained is needed.
- **Hours are not randomly assigned.** People who chose more training may be more motivated, and motivated people improve more for other reasons. The dose–response result (r = .15) is consistent with an effect of training but does not prove it.
- **Skills are not performance.** A better test score does not show better work. Link training to outcomes that matter (performance, errors, promotion) if you can, and say when you cannot.
- **Statistical and practical significance.** With 1,200 people almost any difference is "significant". Judge the size: 5.61 points on a test with SD 11 to 13 is a large effect, 0.67 points per ten hours is a small one.

5.6 Planning the next evaluation: power analysis

Before the next evaluation, work out how many people you need. Click **Power analysis** and choose **Paired-samples t-test (Same people twice)** (Figure 45). Use a cautious effect size: a new evaluation may find a smaller effect than a first one, so take half of the d_z found here, 0.45. With **A priori power analysis**, α = .05 (two-sided) and power .80, SaNaStat reports N = 41: *An a priori power analysis for a paired-samples t-test indicated that 41 participants are needed to detect d_z = 0.45 with α = .05 (two-sided) and power = .80 (achieved power .803).*

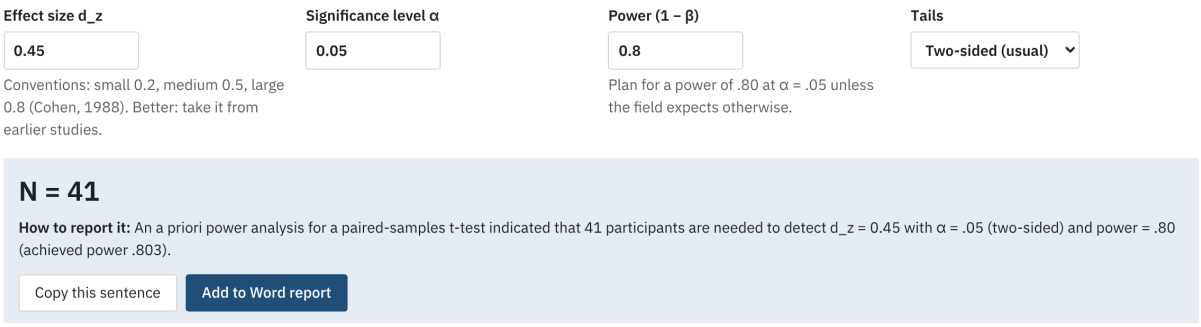


Figure 45. *A priori power analysis for a paired-samples t-test.*

Two other questions can be asked of the same screen. If only one class of 40 people is available, **Sensitivity power analysis** shows the smallest effect it can detect with power .80: d_z = 0.454. And **Post hoc power analysis** with N = 100 shows that a study of 100 people has a power of only .508 to detect d_z = 0.20: it would miss a small effect about half of the time. A power analysis has its own **Add to Word report** button, like every other result.

Exercise

1. What is the mean gain on the skills test, and its 95% interval? (Answer: 5.61, 95% CI [5.26, 5.96].)

1. Run the partial correlation of Training_hours and Skills_test_after controlling for Skills_test_before. (Answer: r = .15.)

1. Change the effect size in the power analysis to d_z = 0.30. How many people are needed? (Answer: N = 90, achieved power .804.)

1. List two reasons why the 5.61-point gain could be smaller than it looks.

6. Engagement and job satisfaction: what goes with them?

Which working conditions go with higher engagement and higher job satisfaction? This chapter uses correlations to see the whole picture, multiple regression to see each condition's share, a hierarchical regression to separate background from working conditions, and a repeated-measures analysis to follow engagement through the year. Data: practice_hr.csv (simulated). Engagement and Job_satisfaction are survey scores from 1 to 5.

6.1 The correlation matrix

Click **Analyses** → **Correlation and regression** → **Bivariate and partial correlation**. Tick Engagement, Job_satisfaction, Manager_support, Work_life_balance, Overtime_hours_month, Performance_rating, Tenure_years and Training_hours. Keep **Pearson product-moment correlation (Pearson r)**.

Concept: correlation

A correlation r measures how closely two variables follow a straight line, from −1 to +1. By convention .10 is small, .30 medium and .50 large (Cohen, 1988). With many pairs, some will look significant by chance; SaNaStat therefore adjusts the p-values for the number of pairs (the Holm method) and uses every person with both values for each pair. *Partial* correlation removes the effect of a third variable.

What it means for our example

8 of the 28 correlations are significant after Holm adjustment. Engagement goes with Job_satisfaction (r = .66, 95% CI [.63, .69]), with Manager_support (r = .40, [.35, .45]) and with Work_life_balance (r = .29, [.24, .34]). Overtime_hours_month goes with lower Work_life_balance (r = −.33, [−.38, −.28]) and, weakly, with lower Engagement (r = −.11, Holm-adjusted p = .003). Job_satisfaction also goes with Manager_support (r = .26), Work_life_balance (r = .25) and Tenure_years (r = .14). Performance_rating and Training_hours are unrelated to everything else (|r| ≤ .06). Manager support and work-life balance are unrelated to each other (r = −.05).

The **Charts** screen draws the same matrix as a heatmap (Figure 46): one colour for positive and another for negative correlations, darker for stronger ones.

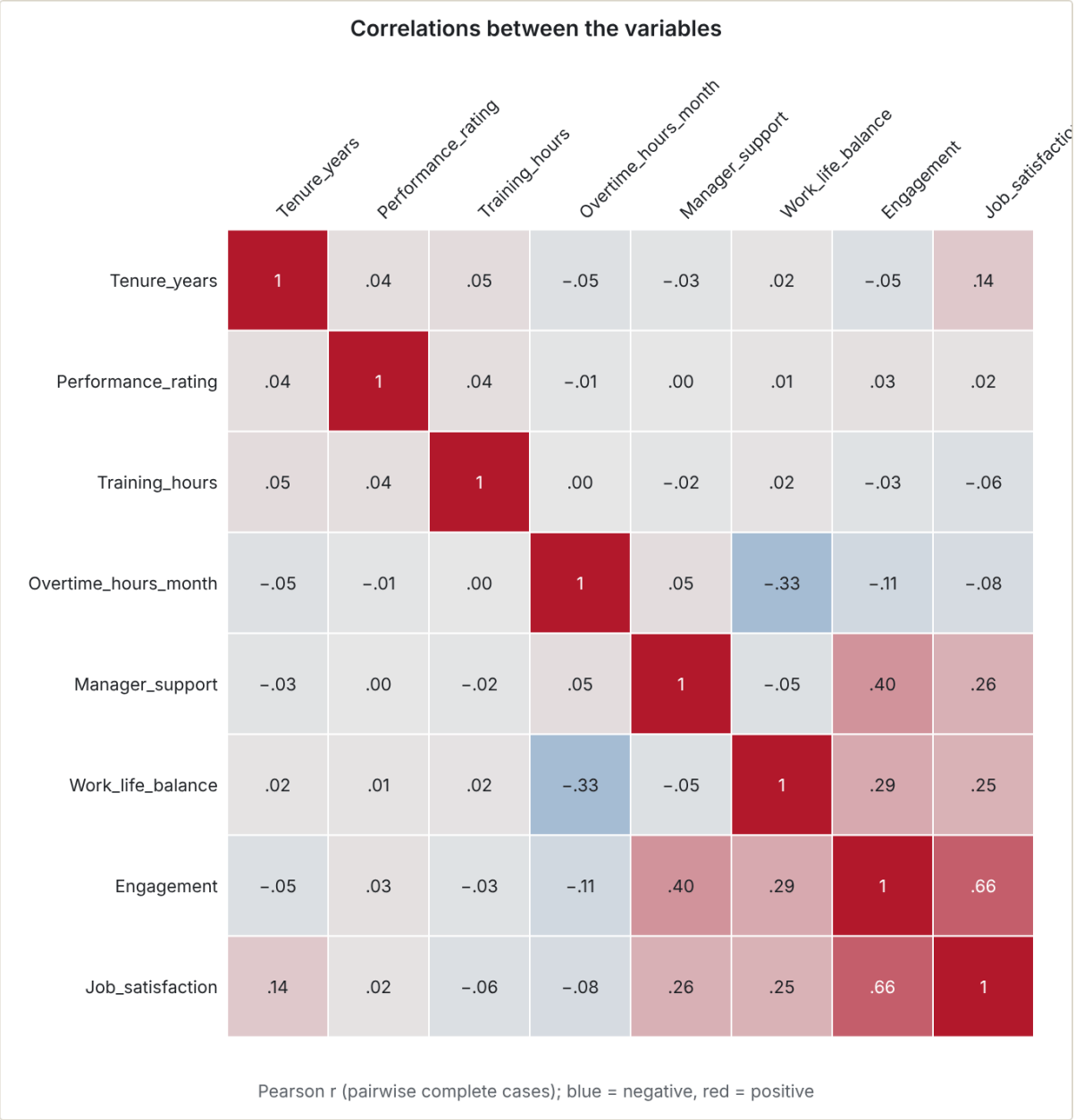


Figure 46. *Correlation heatmap of the same eight variables.*

6.2 Multiple regression of engagement

A multiple regression estimates each condition's association with engagement while holding the others constant. Choose **Multiple linear regression**, **Outcome (scale)** Engagement, and tick Manager_support, Work_life_balance, Overtime_hours_month and Performance_rating as block 1 predictors (Figure 47).

Model summary

Model	Predictors	R^2	Adjusted R^2	F	df	p	ΔR^2	p (ΔR^2)
1	Performance_rating, Overtime_hours_month, Manager_support, Work_life_balance	.263	.260	102.84	4, 1155	< .001		

Copy table | Copy as picture | PNG | CSV

Coefficients (model 1)

Predictor	B	SE	β	t	p	95% CI of B	VIF
(Intercept)	2.440	0.114		21.31	< .001	[2.215, 2.664]	
Performance_rating	0.016	0.017	.023	0.90	.369	[−0.018, 0.050]	1.00
Overtime_hours_month	−0.002	0.002	−.022	−0.81	.418	[−0.007, 0.003]	1.12
Manager_support	0.277	0.017	.418	16.53	< .001	[0.244, 0.310]	1.00
Work_life_balance	0.235	0.020	.310	11.59	< .001	[0.195, 0.275]	1.12

n = 1160 (40 respondents with missing values left out). β = standardised coefficient.

Figure 47. *Multiple regression of Engagement on four working conditions: model summary and coefficients.*

Table 13

Regression of Engagement (n = 1160; people with missing values left out)

Predictor	B	β	t	p
Manager_support	0.277	.42	16.53	< .001
Work_life_balance	0.235	.31	11.59	< .001
Overtime_hours_month	−0.002	−.02	−0.81	.418
Performance_rating	0.016	.02	0.90	.369

What it means for our example

The four predictors explain 26.3% of the variance in Engagement (R^2 = .263, adjusted R^2 = .260, $F(4, 1155) = 102.84$, $p < .001$). A one-point rise in manager support goes with 0.28 points higher engagement, and a one-point rise in work-life balance with 0.24 points, each with the other conditions held constant. Overtime and performance add nothing once the other two are in the model. Overtime was weakly correlated with engagement ($r = -.11$), but it is related to work-life balance ($r = -.33$), which carries the association. Whether overtime lowers engagement *through* work-life balance is a question of mediation (section 6.6); and survey data measured at one time cannot settle the order of cause and effect.

SaNaStat's checks all pass except one, flagged *Note*: the error variance is not constant (Breusch–Pagan $\chi^2(4) = 36.49$, $p < .001$). Engagement cannot exceed 5, so its spread shrinks near the top of the scale. The residuals are not perfectly normal (Shapiro–Wilk $p < .001$), but with 1,160 people the coefficient tests are robust, as SaNaStat says. VIF is at most 1.12.

Cut-off values: regression

Sample size at least 50 + 8k for k predictors (Green, 1991). VIF below 5: no concern (Hair et al., 2019). Breusch–Pagan p ≥ .05: no evidence that the error spread changes (Breusch & Pagan, 1979; Koenker, 1981).

6.3 Hierarchical regression: background first, working conditions second

Does the background (level, gender, age, tenure) explain engagement, and what do the working conditions add? Put Job_level, Gender, Age and Tenure_years in block 1 and the four conditions in block 2 (Figure 48).

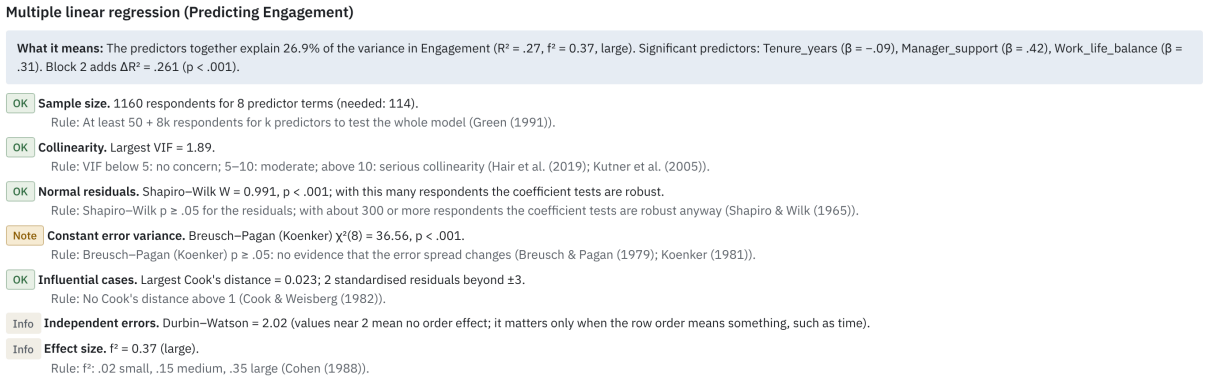


Figure 48. *Hierarchical regression of Engagement: what it means and the checks.*

What it means for our example

The background explains almost nothing (model 1: R^2 = .007, F(4, 1155) = 2.15, p = .073). Adding the working conditions raises R^2 to .269 (ΔR^2 = .261, p < .001). In the full model only Tenure_years (B = −0.014, β = −.09, p = .007), Manager_support (β = .42) and Work_life_balance (β = .31) are significant; gender is not (B = −0.036, p = .222). Engagement depends much more on how people are managed and how they live than on who they are.

How to report it

"A hierarchical multiple linear regression of Engagement on background variables (block 1) and working conditions (block 2) showed that the working conditions added significantly, ΔR^2 = .261, p < .001, final model R^2 = .27, F(8, 1151) = 52.82, p < .001. Manager support (β = .42, p < .001) and work-life balance (β = .31, p < .001) predicted engagement."

6.4 Job satisfaction

Repeat the regression with Job_satisfaction as the outcome and Manager_support, Work_life_balance, Overtime_hours_month, Performance_rating and Engagement as predictors (n = 1111). The model explains 43.5% of the variance (R^2 = .435, $F(5, 1105)$ = 170.08, $p < .001$). Engagement dominates (β = .63, t = 23.70, $p < .001$) and Work_life_balance adds a little (β = .08, p = .001); Manager_support, which correlated with satisfaction at r = .26, has no coefficient of its own once engagement is in the model. The pattern fits the idea that support works through engagement, but this is an interpretation that a regression cannot test.

> **Note: two scores that are nearly one thing**
>
> Engagement and satisfaction correlate at r = .66, and both are answers from the same person at the same time. Predicting one from the other is partly asking the same question twice. Do not read the large coefficient as "engagement drives satisfaction".

6.5 Engagement over the year: repeated-measures ANOVA

The file has engagement measured in four quarters. The same people were measured four times, so use **Analyses** → **Repeated-measures ANOVA**, and tick Engagement_Q1 to Engagement_Q4 in the order they were taken (Figure 49).

Descriptives (complete cases)

Measure	n	Mean	SD
Engagement_Q1	1200	4.19	0.61
Engagement_Q2	1200	4.22	0.61
Engagement_Q3	1200	4.24	0.61
Engagement_Q4	1200	4.25	0.61

Copy table | Copy as picture | PNG | CSV

Repeated-measures ANOVA

Correction	df	F	p	ε	Partial η²	90% CI
None (sphericity assumed)	3, 3597	15.79	< .001	1	.013	[.007, .019]
Greenhouse–Geisser	2.99, 3584.78	15.79	< .001	.997	.013	[.007, .019]
Huynh–Feldt	3.00, 3594.73	15.79	< .001	.999	.013	[.007, .019]

Partial η² = .013; generalised η² = .002 (Bakeman, 2005; no interval is given for it). CIs of η², ω² and partial η² are two-sided 90% intervals (100(1 − 2α)% with α = .05), from inverting the noncentral F distribution. This level matches the F test: the lower limit is above 0 exactly when the test is significant (Steiger, 2004). Each row's interval uses that row's degrees of freedom.

Copy table | Copy as picture | PNG | CSV

Mauchly's test of sphericity

W	χ²	df	p
0.995	6.13	5	.294

Figure 49. *Repeated-measures ANOVA of Engagement_Q1 to Engagement_Q4: descriptives, the test with its corrections and Mauchly's test.*

Cut-off values: sphericity

If Mauchly's test is significant, correct the degrees of freedom: Greenhouse–Geisser when ε < .75, Huynh–Feldt when ε ≥ .75 (Girden, 1992). The ANOVA assumes the differences between all pairs of time points have equal variance (sphericity).

What it means for our example

Mean engagement rose slightly each quarter: 4.19, 4.22, 4.24 and 4.25 (SD = 0.61 in each). Mauchly's test is not significant (W = 0.995, $\chi^2(5) = 6.13$, p = .294), so no correction is needed, F(3, 3597) = 15.79, p < .001, partial $\eta^2 = .013$, 90% CI [.007, .019], generalised $\eta^2 = .002$ (negligible). The Friedman test agrees, $\chi^2(3) = 47.19$, p < .001, Kendall's W = .013. The difference between the first and the last quarter is 0.07 points (Holm-adjusted p < .001, $d_z = -0.19$). The change is statistically clear and practically tiny: with 1,200 people SaNaStat can detect a change of a few hundredths of a point, but whether 0.07 points on a five-point scale matters to the company is a question for the company.

6.6 Does one factor act through another? Mediation and moderation

Two questions of this chapter go beyond regression. Does overtime lower engagement *through* worse work-life balance (mediation)? Does manager support matter more for employees with heavy overtime (moderation)? **Analyses** → **Mediation and moderation** answers such questions: simple and parallel mediation with bootstrap confidence intervals for the indirect effect, moderation with simple slopes and the Johnson–Neyman technique, and moderated mediation (Hayes, 2022). The Common Statistical Tests manual (chapter 13) shows each model step by step. Remember that a mediation model estimated on survey data from one occasion describes a pattern that fits a causal story; it does not prove the story.

Exercise

1. Which working condition has the largest standardised coefficient for Engagement? (Answer: Manager_support, β = .42.)

1. Run the regression of Engagement on Tenure_years alone. Is the coefficient significant, and is its sign the same as in the hierarchical model? (Answer: B = −0.008, p = .059; negative in both, but not significant on its own.)

1. Draw a scatter plot of Engagement against Manager_support. What is r? (Answer: r = .40.)

1. Why is the repeated-measures ANOVA better than four separate tests of the quarters?

7. Absence and department differences

This chapter uses non-parametric tests, which do not assume a bell-shaped distribution, count regression for the number of days absent, and the chi-square test for categories. The questions are whether absence and leaving differ between departments and what goes with absence.

7.1 Absence by department

Absence_days is a count from 0 to 11 with a median of 4.00 in every department. Choose **One-way ANOVA**, **Score to compare** Absence_days and **Groups** Department (Figure 50).

One-way ANOVA (Comparing Absence_days across Department)

What it means: Absence_days does not differ significantly across the 5 Department groups (ω² = .000).

OK **Equal variances.** Levene's test does not suggest unequal variances: the classic ANOVA is used with Tukey–Kramer comparisons.

Rule: Levene's test p < .05 suggests unequal variances: rely on Welch's test and Games–Howell comparisons (Levene (1960); Games & Howell (1976)).

OK **Normality.** Every group has 30 or more cases, so the test of means is robust to non-normal data. Shapiro–Wilk: Finance W = .963, p < .001; HR W = .945, p < .001; IT W = .965, p < .001; Operations W = .962, p < .001; Sales W = .955, p < .001. With this many cases the Shapiro–Wilk test flags even small, harmless departures; look at the Q–Q plot.

Rule: With about 30 or more cases per group, t-tests and ANOVA tolerate non-normal data; with fewer, prefer the rank-based test if the data are clearly skewed (Lumley et al. (2002)).

Info **Effect size.** ω² = .000 (negligible).

Rule: η² / ω²: about .01 small, .06 medium, .14 large (Cohen (1988)).

Info **Effect-size intervals.** η² and ω² are given with 90% confidence intervals (100(1 − 2α)%, so the interval excludes 0 exactly when the test is significant at α = .05. Other effect sizes and differences use 95% intervals.

Rule: Report η², ω², partial η², Cramér's V, φ and w with a 100(1 − 2α)% interval (90% at α = .05): its lower limit is above 0 exactly when the F or χ² test is significant (Steiger (2004)).

Figure 50. *One-way ANOVA of Absence_days by Department: what it means and the checks.*

What it means for our example

Mean absence is 4.05 days in Finance, 3.67 in HR, 4.02 in IT, 3.87 in Operations and 3.80 in Sales. The departments do not differ: $F(4, 1195) = 1.01$, $p = .403$, $\eta^2 = .003$, $\omega^2 = .000$. Levene's test does not suggest unequal variances ($p = .161$), and the Kruskal–Wallis test, which does not depend on the distribution of a count, agrees: $H(4) = 5.43$, $p = .246$, $\varepsilon^2 = .005$. None of the ten pairs differs after correction. The same is true of Engagement in this file: $F(4, 1195) = 0.28$, $p = .891$ and Kruskal–Wallis $p = .838$ (means from 4.18 to 4.24).

Note that a non-significant result is a finding. "No evidence of a difference between departments" is useful to know. Do not describe it as "no difference" without the confidence intervals; the Tukey–Kramer intervals in the report show the range of differences that remain plausible (for example Finance minus HR, 0.38 days, 95% CI [−0.27, 1.03]).

7.2 What goes with absence?

Absence is a count with many ties, so use Spearman's rank correlation: **Bivariate and partial correlation**, choose **Spearman's rank-order correlation (Spearman's ρ)**, and tick Absence_days, Engagement, Job_satisfaction, Commute_km, Overtime_hours_month and Tenure_years.

What it means for our example

Absence is lower among more engaged employees (ρ = −.18, 95% CI [−.24, −.13], Holm-adjusted p < .001) and, less strongly, among more satisfied ones (ρ = −.10, [−.16, −.05], p = .006). It is slightly higher with a longer commute (ρ = .08, p = .044 after adjustment). Overtime (ρ = .07, adjusted p = .114) and tenure (ρ = −.00) are not related to absence. Absence does not differ by gender (3.86 against 3.90 days, Welch p = .729) or between people who stayed and people who left (3.88 days in both groups, p = .957).

These are weak associations. Each correlation looks at one variable at a time; the count regression of the next section looks at them together.

7.3 Counting absence days: Poisson and negative binomial regression

Days of absence, sick-leave spells, complaints or accidents are *counts*: whole numbers 0, 1, 2 … with no upper limit and often a long right tail. A linear regression can predict impossible negative counts and treats an increase from 0 to 1 day like one from 10 to 11. Count regression models the logarithm of the expected count, so its coefficients become *rate ratios*.

1. **Analyses** → **Correlation and regression** → **Count regression, Poisson or negative binomial (Which variables predict how often something happens?)**.
2. **Outcome (a count)**: Absence_days. Only variables of whole numbers from 0 upwards are listed.
3. **Count model**: leave **Negative binomial regression (recommended)**.
4. **Exposure (optional)**: leave it at none. Choose a variable here when people were observed for different lengths of time (for example months employed in the year): the results then become rates per unit of time.
5. **Predictors**: Engagement, Job_satisfaction, Commute_km, Overtime_hours_month, Tenure_years and Department (Figure 51).

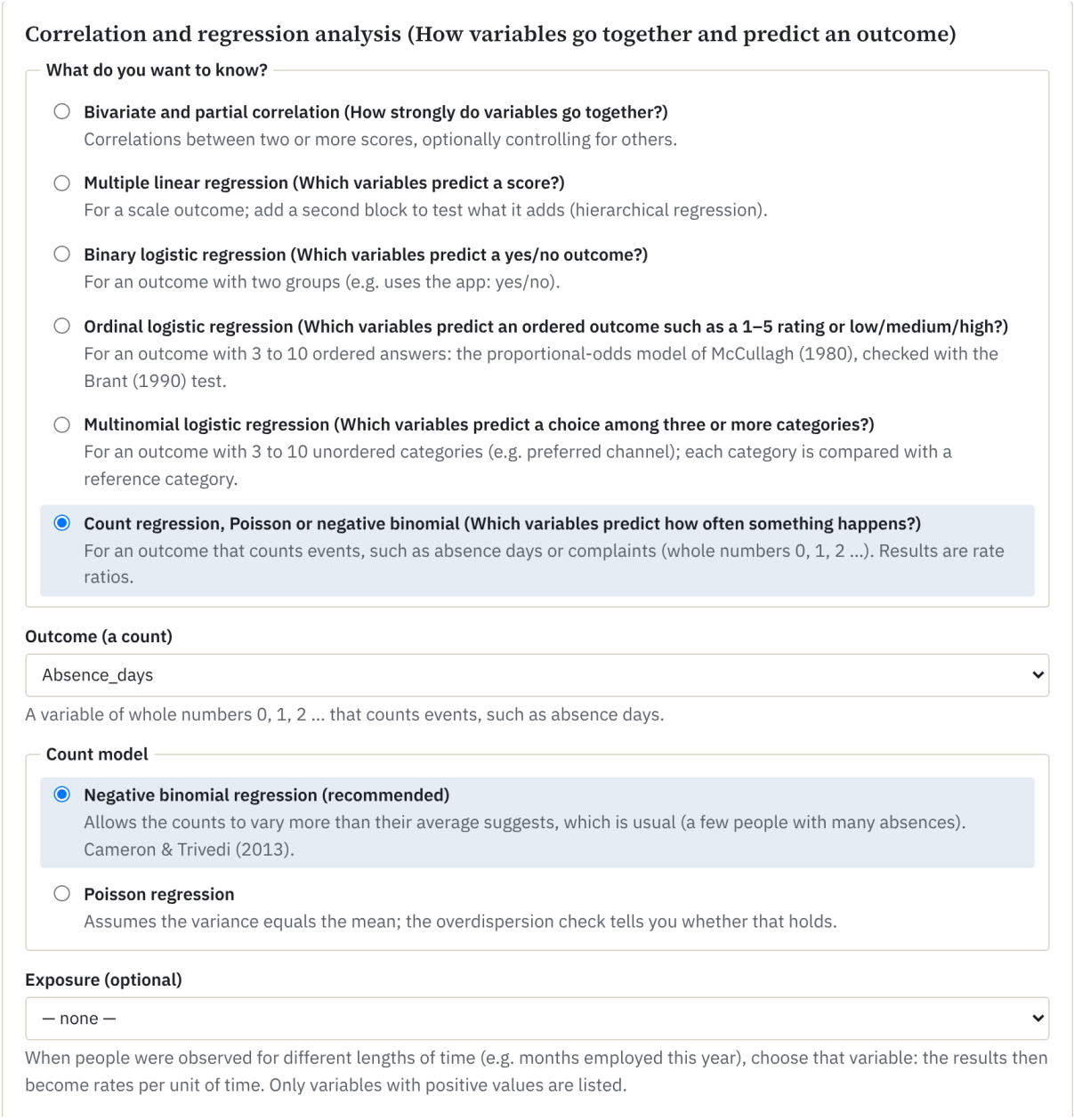


Figure 51. *Count regression: the count model, the exposure and the predictors.*

Concept: rate ratios and overdispersion

The *incidence-rate ratio* (IRR) is the factor by which the expected count changes for one unit more of a predictor: IRR = 0.86 means 14% fewer days per unit, IRR = 1.10 means 10% more. *Poisson regression* assumes that the variance of the counts equals their mean. Real counts often vary more (a few people with many absences): this is *overdispersion*, and it makes the Poisson standard errors too small. *Negative binomial regression* (NB2) adds a dispersion parameter α that allows for it; α = 0 gives the Poisson model back (Cameron & Trivedi, 2013).

Cut-off values: overdispersion

Poisson regression assumes the variance equals the mean; when the Pearson dispersion clearly exceeds 1 or the overdispersion test is significant, use negative binomial regression, otherwise the standard errors are too small (Cameron & Trivedi, 1990, 2013).

Negative binomial regression (Predicting the count Absence_days)

What it means: The predictors explain differences in Absence_days. Commute_km: 0.4% more Absence_days per unit; Engagement: 14% fewer Absence_days per unit.

OK **Overdispersion.** Pearson dispersion of the Poisson model = 0.98 (1 = as Poisson expects); Cameron–Trivedi test t(1112) = −0.75, p = .772; NB2 vs Poisson LR χ² = 0.00, p = .497. Little overdispersion: the Poisson model would give much the same result.
Rule: Poisson regression assumes the variance equals the mean; when the Pearson dispersion clearly exceeds 1 or the overdispersion test is significant, use negative binomial regression, otherwise the standard errors are too small (Cameron & Trivedi (1990, 2013)).

Info **Zeros.** 28 of 1113 respondents (3%) have a count of 0; mean 3.90, variance 3.95.

Figure 52. *Negative binomial regression of Absence_days: what it means and the checks.*

Model fit

N	LR χ²	df	p	Log-likelihood	AIC	BIC	McFadden R²	α (dispersion)	95% CI of α
1113	49.97	9	< .001	−2293.19	4608.39	4663.55	.011	0.000	[0.000, 0.000]

LR χ² compares the model with a model without predictors. Variance = μ + α μ² (NB2); α = 0 would be the Poisson model.

Copy table | Copy as picture | PNG | CSV

Incidence-rate ratios

Predictor	B	SE	z	p	IRR	95% CI of IRR
(Intercept)	1.958	0.125	15.68	< .001	7.084	[5.546, 9.048]
Department: HR (vs Finance)	−0.083	0.064	−1.31	.191	0.920	[0.812, 1.042]
Department: IT (vs Finance)	−0.008	0.053	−0.16	.875	0.992	[0.894, 1.101]
Department: Operations (vs Finance)	−0.048	0.048	−1.01	.312	0.953	[0.868, 1.046]
Department: Sales (vs Finance)	−0.069	0.048	−1.45	.148	0.933	[0.850, 1.025]
Tenure_years	−0.001	0.004	−0.32	.753	0.999	[0.991, 1.006]
Overtime_hours_month	0.001	0.002	0.62	.537	1.001	[0.997, 1.006]
Commute_km	0.004	0.002	2.81	.005	1.004	[1.001, 1.008]
Engagement	−0.154	0.035	−4.43	< .001	0.857	[0.801, 0.918]
Job_satisfaction	0.005	0.034	0.14	.890	1.005	[0.941, 1.073]

IRR = exp(B): how many times the expected count changes per unit of the predictor (or for a group compared with its reference group). 87 respondents with missing values left out. Categorical predictors are dummy-coded against a reference group: Department = Finance.

Figure 53. *The model fit and the incidence-rate ratios.*

Figure 52 and Figure 53 show the result. The model uses 1,113 employees (87 with missing values are left out) and is significant, $\chi^2(9, N = 1113) = 49.97$, $p < .001$, McFadden $R^2 = .011$.

Table 14

Negative binomial regression of Absence_days (from the program)

Predictor	IRR	95% CI of IRR	p
Engagement	0.857	[0.801, 0.918]	< .001
Commute_km	1.004	[1.001, 1.008]	.005
Overtime_hours_month	1.001	[0.997, 1.006]	.537
Tenure_years	0.999	[0.991, 1.006]	.753
Job_satisfaction	1.005	[0.941, 1.073]	.890
Department: HR (vs Finance)	0.920	[0.812, 1.042]	.191
Department: Sales (vs Finance)	0.933	[0.850, 1.025]	.148

What it means for our example

With the other predictors held constant, each extra point of engagement goes with 14% fewer days of absence (IRR = 0.857, 95% CI [0.801, 0.918]), and each extra kilometre of commuting with 0.4% more (IRR = 1.004, [1.001, 1.008]). Job satisfaction, which was weakly related to absence on its own (ρ = −.10), adds nothing once engagement is in the model. Overtime, tenure and department are not related to absence. The checks answer the question of which model to use: the Pearson dispersion of the Poisson model is 0.98 (1 is what Poisson expects), the Cameron–Trivedi test is not significant (p = .772) and the likelihood-ratio test of the negative binomial against the Poisson model gives $\chi^2 = 0.00$, p = .500. The estimated α is 0, so *the negative binomial model reduces to the Poisson model*. In these simulated data the counts behave exactly as Poisson counts; real absence data are usually overdispersed.

Choose **Poisson regression** to compare (Figure 54). The coefficients are the same, the check reads *No sign of overdispersion*, and the AIC of the Poisson model (4606.39) is a little smaller than that of the negative binomial model (4608.39), because the Poisson model has one parameter fewer. Either model may be reported here; with overdispersed data, report the negative binomial model.

Poisson regression (Predicting the count Absence_days)

What it means: The predictors explain differences in Absence_days. Commute_km: 0.4% more Absence_days per unit; Engagement: 14% fewer Absence_days per unit.

OK **Overdispersion.** Pearson dispersion of the Poisson model = 0.98 (1 = as Poisson expects); Cameron–Trivedi test t(1112) = −0.75, p = .772; NB2 vs Poisson LR $\chi^2 = 0.00$, p = .497. No sign of overdispersion.

Rule: Poisson regression assumes the variance equals the mean; when the Pearson dispersion clearly exceeds 1 or the overdispersion test is significant, use negative binomial regression, otherwise the standard errors are too small (Cameron & Trivedi (1990, 2013)).

Info **Zeros.** 28 of 1113 respondents (3%) have a count of 0; mean 3.90, variance 3.95.

Figure 54. *The Poisson model of the same counts: no sign of overdispersion.*

How to report it

"A negative binomial regression of absence days on engagement, job satisfaction, commuting distance, overtime, tenure and department was significant, $\chi^2(9, N = 1113) = 49.97$, $p < .001$. Each additional point of engagement was associated with 14% fewer absence days, IRR = 0.857, 95% CI [0.801, 0.918], $p < .001$, and each additional kilometre of commuting with 0.4% more, IRR = 1.004, 95% CI [1.001, 1.008], $p = .005$. There was no evidence of overdispersion (dispersion = 0.98)."

Note: counts with many zeros, and counts over different periods

Only 28 of the 1,113 employees (3%) had no absence at all, and the check *Zeros* reports it. When a count has far more zeros than the model expects (for example accidents, where most people have none), say so and be cautious: SaNaStat has no zero-inflated model. When people were employed for different parts of the year, use **Exposure (optional)**; otherwise a new employee's few days look like low absence.

7.4 Leaving by department: the chi-square test

Is leaving linked to department? Both variables are categories, so use **Chi-square, Fisher, McNemar** → **Chi-square test of independence**: **Categorical variable** Department, **Second categorical variable** Left_company (Figure 55).

Counts (row %)

Department	No	Yes	Total
Finance	155 (87%)	24 (13%)	179
HR	97 (86%)	16 (14%)	113
IT	173 (89%)	22 (11%)	195
Operations	294 (81%)	69 (19%)	363
Sales	305 (87%)	45 (13%)	350

Copy table | Copy as picture | PNG | CSV

Tests

Test	Value	df	p
Pearson chi-square	8.42	4	.077
Likelihood ratio	8.20	4	.085

Copy table | Copy as picture | PNG | CSV

Adjusted standardised residuals

Department	No	Yes
Finance	0.52	−0.52
HR	0.16	−0.16
IT	1.46	−1.46
Operations	−2.80 *	2.80 *
Sales	1.14	−1.14

* beyond ±1.96: this cell holds clearly more (positive) or fewer (negative) cases than expected.

Figure 55. *Chi-square test of Department by Left_company: counts, tests and adjusted residuals.*

Concept: the chi-square test of independence

The test compares the counts in a table with the counts expected if the two variables were unrelated. The *adjusted standardised residual* of each cell says how far its count is from expectation, in units that behave like z-scores: beyond ±1.96 is notable. Cramér's V is the effect size, from 0 to 1, given with a 90% confidence interval. The chi-square test needs expected counts of at least 5 in nearly all cells; otherwise use Fisher's exact test.

Table 15

Left_company by Department

Department	Stayed	Left	% left
Finance	155	24	13%
HR	97	16	14%
IT	173	22	11%
Operations	294	69	19%
Sales	305	45	13%

What it means for our example

Overall 176 of the 1,200 employees (14.7%) left. The test does not reject independence: $\chi^2(4, N = 1200) = 8.42$, $p = .077$, Cramér's V = .08 (negligible). One cell is flagged: Operations has more leavers than expected (adjusted residual 2.80, 19% left against 11% to 14% elsewhere). Because the overall test is not significant, treat that as a lead to examine, not as a finding: with ten cells, one will fall outside ±1.96 by chance fairly often. If Operations is a concern, test it as a pre-planned question on new data.

The same test on two other variables gives no significant relation either. Promotion and leaving: 15% of those not promoted left against 11% of those promoted, $\chi^2(1, N = 1200) = 1.45$, $p = .228$, Cramér's V = .03 (Fisher's exact test $p = .247$). Gender and leaving: 17% of women and 13% of men left, $\chi^2(1, N = 1200) = 2.91$, $p = .088$, Cramér's V = .05 (Fisher $p = .102$). Chapter 8 returns to both with the timing of leaving.

The variables made in section 1.9 can be used here. Tenure_group (three equal-sized groups) against Left_company (Figure 56): 18% of the low-tenure third left (71 of 401), 17% of the middle third (68 of 407) and 9% of the high-tenure third (37 of 392), $\chi^2(2, N = 1200) = 12.87$, $p = .002$, Cramér's V = .10 (small). The high-tenure third has fewer leavers than expected (adjusted residual −3.57 for leaving) and the low-tenure third more (2.11).

Counts (row %)

Tenure_group	No	Yes	Total
Low third	330 (82%)	71 (18%)	401
Middle third	339 (83%)	68 (17%)	407
High third	355 (91%)	37 (9%)	392

Copy table | Copy as picture | PNG | CSV

Tests

Test	Value	df	p
Pearson chi-square	12.87	2	.002
Likelihood ratio	13.71	2	.001

Copy table | Copy as picture | PNG | CSV

Adjusted standardised residuals

Tenure_group	No	Yes
Low third	−2.11 *	2.11 *
Middle third	−1.43	1.43
High third	3.57 *	−3.57 *

* beyond ±1.96: this cell holds clearly more (positive) or fewer (negative) cases than expected.

Figure 56. *Chi-square test of Tenure_group by Left_company.*

Merging Finance and HR into one category called Corporate (Dept_merged) changes the department test: $\chi^2(3, N = 1200) = 8.39$, $p = .039$, Cramér's V = .08 (negligible), with the same flagged cell, Operations (adjusted residual 2.80).

Note: merging categories after looking at the data

With all five departments the test gave $p = .077$; with Finance and HR merged it gives $p = .039$. No employee changed; only the number of categories did. If you choose a merge, a cut point or a grouping after seeing which version gives a significant result, the p-value is no longer trustworthy. Decide such groupings from the meaning of the groups before you look at the outcome, and report every version you tried.

7.5 Leavers and stayers compared

Descriptive statistics split by Left_company gives a first look at how leavers differ, which chapters 8 and 9 take further.

Table 16

Leavers and stayers (means)

Variable	Stayed (n = 1024)	Left (n = 176)
Tenure_years	5.71	4.48
Annual_salary	45776.58	42136.93
Overtime_hours_month	8.76	11.67
Engagement	4.25	3.91

Leavers had shorter tenure, lower pay, more overtime and lower engagement. Because these variables are correlated with each other and with the outcome, a model is needed to separate them; chapters 8 and 9 do this.

Exercise

1. Run the chi-square test of Gender by Left_company. Is the difference significant? (Answer: no, p = .088; 17% of women and 13% of men left.)

1. Repeat the department test on Job_satisfaction. Compare the Kruskal–Wallis and the ANOVA results. (Answer: F(4, 1143) = 0.31, p = .871; Kruskal–Wallis p = .909.)

1. In the count regression, keep only Engagement and Commute_km as predictors and compare the Poisson and the negative binomial model. Do the conclusions change?

1. In the chi-square table, why should you not conclude that Operations has a leaving problem? (Answer: the overall test is not significant and one flagged cell out of ten is not unusual.)

8. Time to leaving: survival analysis

Chapter 7 asked *whether* employees left. This chapter asks *when*. How long do employees typically stay? Do some departments lose people sooner? Which factors make leaving come earlier? These are questions of *survival analysis* (also called time-to-event or event history analysis). Data: practice_hr.csv (simulated).

8.1 Why time matters, and what censoring is

A yes/no outcome treats someone who left after three months like someone who left after ten years, and it treats an employee who joined last month and has not left (yet) like someone who has stayed for twenty years. Survival analysis uses the time each person was observed and whether the event happened in that time.

Concept: censoring, the survival curve and the hazard

For an employee who has left, the time is the time until leaving. For an employee who is still there, we know only that the time to leaving is longer than the time observed so far: the time is *censored*. Censored people count as "at risk" up to their time and then drop out of the calculation; they are not treated as leavers or as people who will never leave. The *Kaplan–Meier curve* estimates the share of employees still without the event (still employed) at each time (Kaplan & Meier, 1958). The *hazard* is the risk of the event at a given time among those still at risk; the *log-rank test* compares curves between groups (Mantel, 1966), and *Cox regression* relates the hazard to predictors (Cox, 1972).

Note: the practice file has no dates

Real survival analysis needs, for each person, the time from hiring to leaving (for leavers) or to the end of the observation period (for those who stayed), computed from the dates in the HR records. practice_hr.csv has no dates. In this chapter we use **Tenure_years as the time** (the time at which a leaver left, and the time observed so far for a stayer) and **Left_company = Yes as the event**. Because the file records leaving within one year, this is an approximation made for teaching: the curves show the method, not the true timing of leaving in a real company.

8.2 The survival curve for everyone

1. **Analyses** → **Survival analysis**.
2. **Time (scale)**: Tenure_years. **Event (two groups, or coded 0/1)**: Left_company. **Which answer means the event happened?**: SaNaStat proposes Yes; the other answer counts as "not yet" (censored).
3. Leave **Compare groups (optional)** at none and **Show confidence bands around the curves** ticked (Figure 57).

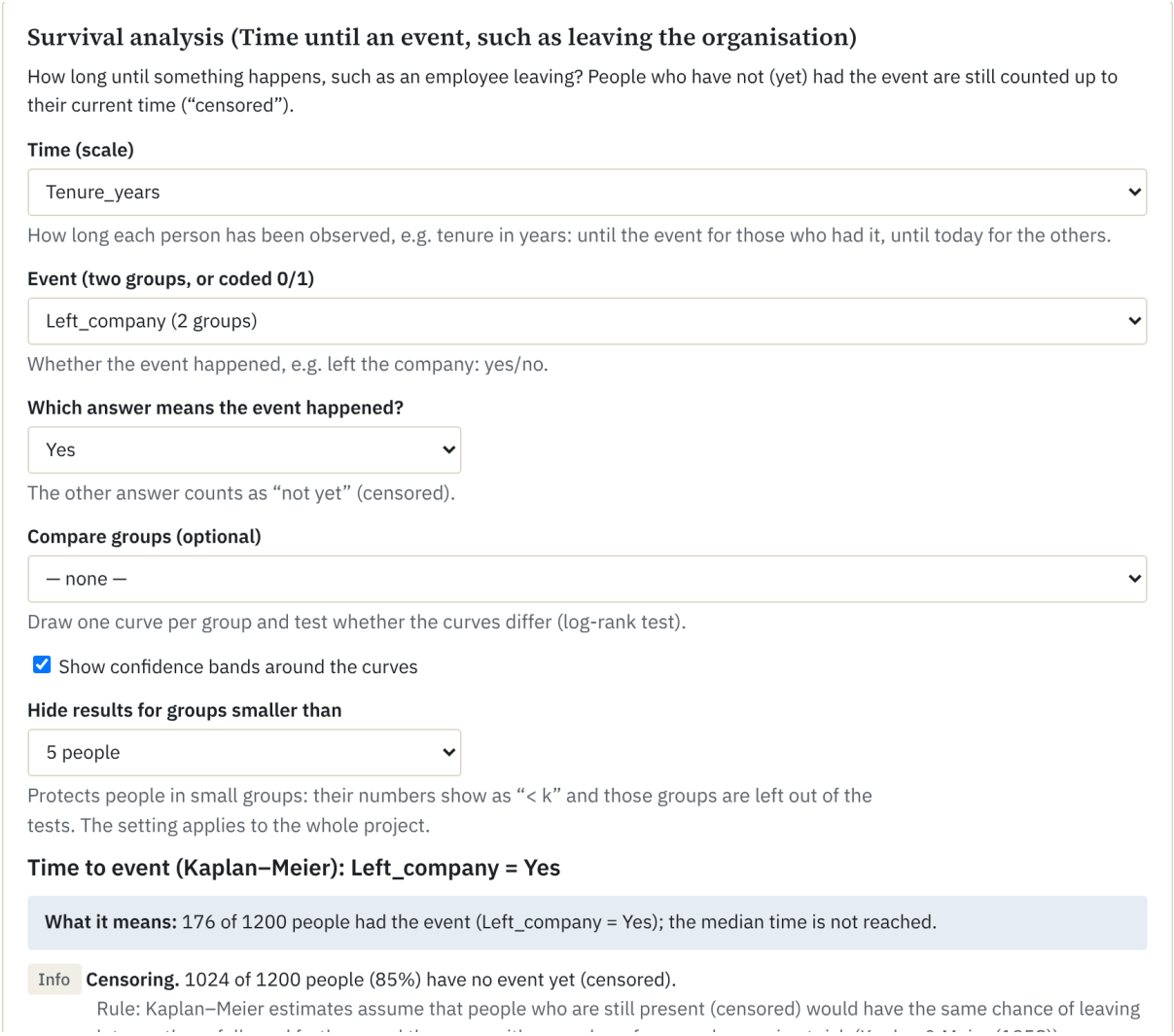


Figure 57. *The survival analysis screen: time, event, the answer that means the event, groups, bands and the privacy setting.*

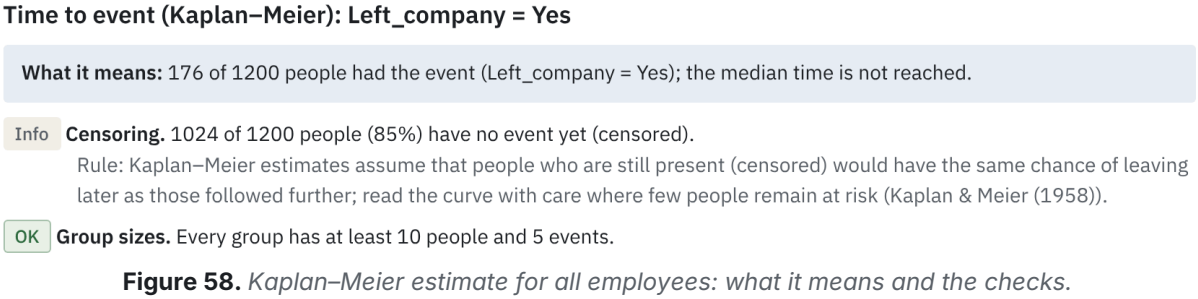
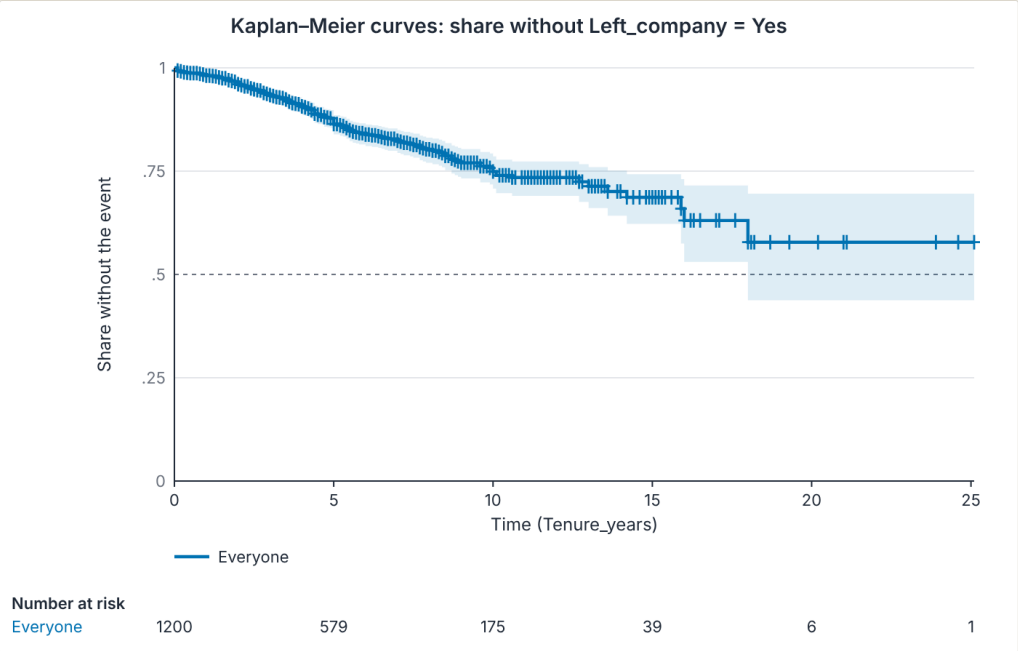


Figure 58. *Kaplan–Meier estimate for all employees: what it means and the checks.*



Kaplan–Meier curves of the share without Left_company = Yes. Shaded: 95% confidence band. + = censored (still present at that time). Dashed line: half of the group.

Figure 59. *The Kaplan–Meier curve with its 95% confidence band and the number at risk.*

Figure 58 and Figure 59 show the result. 176 of 1,200 employees had the event, and 1,024 (85%) are censored. The *Median* is *not reached*: fewer than half of the employees left within the longest tenure observed, so the time by which half have left cannot be estimated. The table *Share without the event at selected times* gives the estimated share still employed at 5, 10, 15 and 20 years of tenure: .86 [.84, .89], .75 [.71, .79], .69 [.62, .74] and .58 [.44, .70].

What it means for our example

About 86% of employees are estimated to stay at least five years and about 75% at least ten. The intervals widen with time because fewer and fewer employees have been there long enough to be at risk: the number-at-risk table under the curve shows how many. Read the right-hand end of a survival curve with care; the check *Censoring* and its rule say the same.

8.3 Comparing departments: the log-rank test

Choose Department in **Compare groups (optional)** (Figure 60, Figure 61, Figure 62).

Time to event (Kaplan–Meier and log-rank test): Left_company = Yes

What it means: Time to Left_company = Yes differs between the Department groups (log-rank p = .048). Operations has the event soonest: 69 events against 51.2 expected if every group had the same curve.

Info **Censoring.** 1024 of 1200 people (85%) have no event yet (censored).
Rule: Kaplan–Meier estimates assume that people who are still present (censored) would have the same chance of leaving later as those followed further; read the curve with care where few people remain at risk (Kaplan & Meier (1958)).

OK **Group sizes.** Every group has at least 10 people and 5 events.

Figure 60. *Kaplan–Meier curves by Department: what it means and the checks.*

Median time to Left_company = Yes

Department	n	Events	Censored	Median	95% CI
Finance	179	24	155	not reached	[16.00, ∞]
HR	113	16	97	not reached	[–, ∞]
IT	195	22	173	not reached	[–, ∞]
Operations	363	69	294	not reached	[–, ∞]
Sales	350	45	305	18.00	[13.60, ∞]

Time in the units of Tenure_years. Censored = still present at the end of follow-up (no event yet). The median is the time by which half of the group has had the event; "not reached" means fewer than half had it. Interval: Brookmeyer–Crowley type, from the 95% band of the curve.

Copy table | Copy as picture | PNG | CSV

Share without the event (Left_company = Yes) at selected times

Department	t = 5	t = 10	t = 15	t = 20
Finance	.88 [.81, .93]	.77 [.65, .86]	.74 [.61, .84]	.56 [.20, .81]
HR	.79 [.68, .87]	.79 [.68, .87]	.79 [.68, .87]	.79 [.68, .87]
IT	.91 [.85, .95]	.78 [.66, .86]	.71 [.51, .83]	.71 [.51, .83]
Operations	.83 [.78, .87]	.65 [.56, .73]	.63 [.54, .71]	.63 [.54, .71]
Sales	.89 [.84, .92]	.80 [.73, .86]	.65 [.48, .78]	.28 [.02, .66]

Kaplan–Meier estimate S(t) with its 95% interval (Greenwood variance, log(–log) scale).

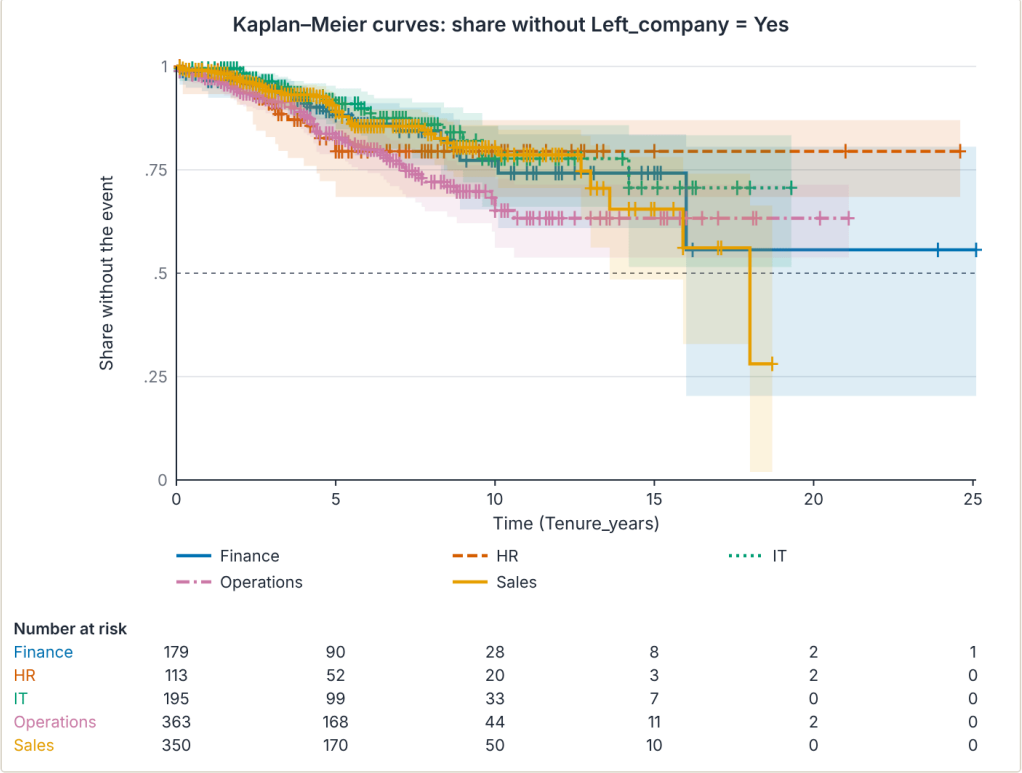
Copy table | Copy as picture | PNG | CSV

Log-rank test

Comparison	χ^2	df	p	p (Holm)
All groups	9.56	4	.048	
Finance vs HR	0.12	1	.725	1.000
Finance vs IT	0.54	1	.464	1.000
Finance vs Operations	2.86	1	.091	.728
Finance vs Sales	0.00	1	.970	1.000
HR vs IT	0.99	1	.321	1.000
HR vs Operations	1.18	1	.278	1.000
HR vs Sales	0.18	1	.672	1.000
IT vs Operations	6.52	1	.011	.107
IT vs Sales	0.67	1	.414	1.000
Operations vs Sales	5.12	1	.024	.212

Observed / expected events: Finance 24 / 27.0; HR 16 / 16.2; IT 22 / 30.3; Operations 69 / 51.2; Sales 45 / 51.3.

Figure 61. *The median time, the share still employed at 5, 10, 15 and 20 years, and the log-rank tests.*



Kaplan–Meier curves of the share without Left_company = Yes by Department. Shaded: 95% confidence band. + = censored (still present at that time). Dashed line: half of the group.

Figure 62. *Kaplan–Meier curves of the five departments with the number at risk.*

Concept: the log-rank test

At every time at which someone leaves, the log-rank test compares the number of leavers in each group with the number expected if all groups had the same curve, and adds up the differences. A significant test says the curves differ somewhere. The note under the table lists the observed and expected events of each group, which show *which* group leaves sooner. With more than two groups SaNaStat also compares every pair, with Holm-adjusted p-values.

What it means for our example

The curves differ, $\chi^2(4, N = 1200) = 9.56$, $p = .048$. Operations had 69 leavers against 51.2 expected if every department had the same curve; IT had 22 against 30.3. At ten years the estimated share still employed is .65 [.56, .73] in Operations against .77 to .80 in the other departments. No pair of departments differs after the Holm adjustment (the smallest adjusted p is .107, IT against Operations). The overall test is just significant at $\alpha = .05$, so treat the Operations pattern as worth watching, not as settled; it is the same lead as in section 7.4, now with the timing included.

The only department with a median is Sales, 18.00 years, 95% CI [13.60, ∞]. This is a warning example: the Sales curve drops to .28 at 20 years because at that point hardly anybody in Sales is still at risk (10 people at 15 years and none at 20 in the number-at-risk

table), so one or two leavers move the curve a long way. The interval of the share at 20 years, [.02, .66], shows how little the data say there.

Two other groupings give significant log-rank tests although the yes/no comparisons of section 7.4 did not. Employees promoted in the last three years leave later: 15 leavers against 24.4 expected, $\chi^2(1, N = 1200) = 4.24$, $p = .040$. Women leave a little sooner than men: 93 leavers against 79.5 expected, $\chi^2(1, N = 1200) = 4.23$, $p = .040$ (the chi-square test of leaving by gender gave $p = .088$). The log-rank test uses when people left, not only whether, and so can detect a difference the yes/no comparison misses. Two tests of the same question that disagree near $p = .05$ are still weak evidence; the Cox model below asks whether gender matters once other factors are held constant.

How to report it

"Kaplan–Meier estimates of time to leaving (N = 1200, 176 leavers) estimated the share still employed at five years of tenure as .86, 95% CI [.84, .89]. A log-rank test showed a significant difference between departments, $\chi^2(4, N = 1200) = 9.56$, $p = .048$; Operations had more leavers than expected (69 observed, 51.2 expected), but no pairwise comparison was significant after Holm adjustment."



Figure 63. *The sentence for the report, with the export and report buttons.*

8.4 Cox regression: which factors make leaving come sooner?

Under the curves, the section *Which factors make the event come sooner? (Cox regression)* fits a Cox proportional-hazards model (Figure 64).

1. Under **Predictors (optional)** tick Department, Overtime_hours_month, Commute_km, Engagement and Promoted_last_3y. The time variable (Tenure_years) and the event are not offered.
2. The reference groups of the categorical predictors are Finance and No; change them in **Reference groups (each other group is compared with this one)** if you wish.
3. Leave **Tied times (several events at the same time)** at **Efron (1977) method (recommended)**. Tenure is recorded to one decimal, so many people share the same time; Efron's method handles such ties more accurately than Breslow's.

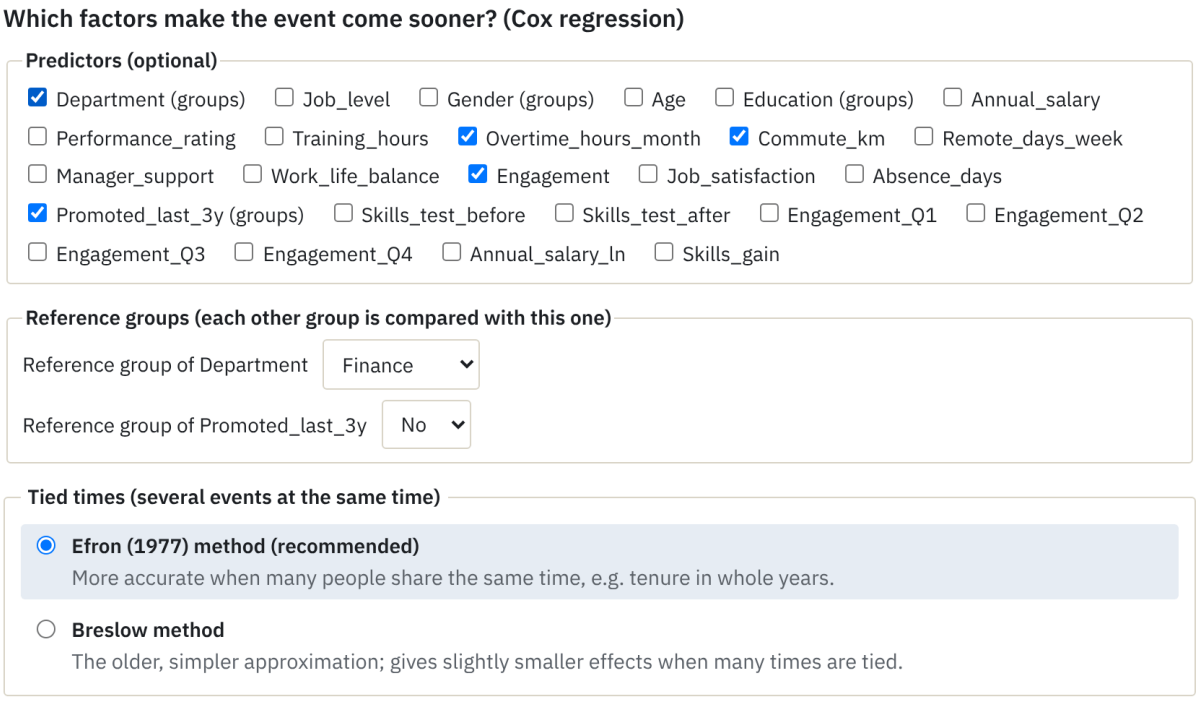


Figure 64. *Cox regression: predictors, reference groups and the method for tied times.*

Concept: the hazard ratio

The *hazard ratio* (HR) is the factor by which the risk of leaving at any time changes for one unit more of the predictor, holding the others constant. HR = 1 means no relationship, HR above 1 means leaving sooner, HR below 1 later. A Cox model makes no assumption about the shape of the survival curve itself; it assumes that each predictor multiplies the hazard by the same factor at every time: the *proportional-hazards* assumption.

Cut-off values: Cox regression

At least 10 events (e.g. leavers) per predictor term in a Cox model; with fewer, hazard ratios and their intervals become unreliable (Peduzzi et al., 1995). Proportional hazards: a predictor's effect should stay the same over time; a significant test of the scaled Schoenfeld residuals against (ranked) time suggests it changes (Grambsch & Therneau, 1994). Concordance: .5 is chance, 1 perfect ordering.

Cox regression (Predictors of time to Left_company = Yes)

What it means: The predictors are related to how soon people have the event (Left_company = Yes). Overtime_hours_month: the hazard is 3.9% higher per unit; Commute_km: the hazard is 2.4% higher per unit; Engagement: the hazard is 46% lower per unit.

OK **Events per predictor.** 172 events for 8 predictor terms: 21.5 per term.
Rule: At least 10 events (e.g. leavers) per predictor term in a Cox model; with fewer, hazard ratios and their intervals become unreliable (Peduzzi et al. (1995)).

OK **Proportional hazards.** No predictor shows a clear change of its effect over time.
Rule: Proportional hazards: a predictor's effect should stay the same over time; a significant test of the scaled Schoenfeld residuals against (ranked) time suggests it changes (Grambsch & Therneau (1994)).

Figure 65. *Cox regression of time to leaving: what it means and the checks.*

Model fit

N	Events	LR χ^2	df	p	Partial AIC	Concordance
1165	172	69.79	8	< .001	2122.37	.687

LR χ^2 compares the model with a model without predictors. Concordance: the share of pairs of people whose order of events the model predicts correctly (.5 = chance).

Copy table | Copy as picture | PNG | CSV

Hazard ratios

Predictor	B	SE	z	p	Hazard ratio	95% CI of HR
Department: HR (vs Finance)	0.191	0.325	0.59	.556	1.211	[0.641, 2.288]
Department: IT (vs Finance)	−0.335	0.296	−1.13	.258	0.715	[0.400, 1.279]
Department: Operations (vs Finance)	0.314	0.242	1.29	.196	1.368	[0.851, 2.201]
Department: Sales (vs Finance)	0.031	0.253	0.12	.901	1.032	[0.628, 1.695]
Overtime_hours_month	0.038	0.010	3.97	< .001	1.039	[1.019, 1.058]
Commute_km	0.024	0.007	3.40	< .001	1.024	[1.010, 1.039]
Engagement	−0.615	0.119	−5.16	< .001	0.541	[0.428, 0.683]
Promoted_last_3y: Yes (vs No)	−0.483	0.273	−1.77	.077	0.617	[0.361, 1.053]

HR > 1: the event happens sooner (higher hazard); HR < 1: later. Categorical predictors are compared with their reference group. 35 people with missing values left out.

Figure 66. *The model fit and the hazard ratios.*

Table 17

Cox regression of time to leaving (N = 1165, 172 events; from the program)

Predictor	HR	95% CI of HR	p
Engagement	0.541	[0.428, 0.683]	< .001
Overtime_hours_month	1.039	[1.019, 1.058]	< .001
Commute_km	1.024	[1.010, 1.039]	< .001
Promoted_last_3y: Yes (vs No)	0.617	[0.361, 1.053]	.077
Department: Operations (vs Finance)	1.368	[0.851, 2.201]	.196
Department: IT (vs Finance)	0.715	[0.400, 1.279]	.258

What it means for our example

The model uses 1,165 employees (35 with a missing commute are left out) and 172 leavers, 21.5 per predictor term, and it is significant, $\chi^2(8) = 69.79$, $p < .001$, concordance = .687. With the other factors held constant, each extra point of engagement roughly halves the hazard of leaving (HR = 0.541, 95% CI [0.428, 0.683]); each extra hour of overtime a month raises it by 3.9% (HR = 1.039) and each extra kilometre of commuting by 2.4% (HR = 1.024). Promotion points towards later leaving (HR = 0.617) but its interval includes 1 ($p = .077$), and no department differs from Finance once the other factors are in the model: the department difference of section 8.3 is largely a matter of who works where and how much overtime they do. These are the factors the logistic regression of chapter 9 will also find.

Proportional-hazards check

Predictor	χ^2	df	p
Department: HR (vs Finance)	2.07	1	.150
Department: IT (vs Finance)	0.23	1	.631
Department: Operations (vs Finance)	0.07	1	.797
Department: Sales (vs Finance)	0.20	1	.658
Overtime_hours_month	0.97	1	.324
Commute_km	0.24	1	.622
Engagement	2.17	1	.141
Promoted_last_3y: Yes (vs No)	0.65	1	.420

Test of the scaled Schoenfeld residuals against the rank of the event times (Grambsch & Therneau, 1994). $p < .05$: the effect of that predictor changes over time.

Figure 67. *The proportional-hazards check: one test per predictor term.*

Check the proportional-hazards assumption. The table *Proportional-hazards check* (Figure 67) tests, for each predictor, whether its effect changes over time. Every p-value is above .05 (the smallest is .141 for Engagement), and the check reads *No predictor shows a clear change of its effect over time.* Had a predictor failed ($p < .05$), its hazard ratio would be an average over time that hides a changing effect; the usual remedies, such as splitting the time axis or stratifying by the predictor, are not available in SaNaStat, so report the failure and interpret that predictor with care.

The choice of method for ties makes little difference here: with **Breslow method** the likelihood-ratio χ^2 is 69.54 instead of 69.79 and the hazard ratios agree to the third decimal (Engagement 0.541).

How to report it

"A Cox proportional-hazards regression (Efron method for ties) of time to leaving on engagement, overtime, commuting distance, promotion and department (N = 1165, 172 events) was significant, $\chi^2(8) = 69.79$, $p < .001$, concordance = .69. Engagement (HR = 0.54, 95% CI [0.43, 0.68]), overtime (HR = 1.04, 95% CI [1.02, 1.06]) and commuting distance (HR = 1.02, 95% CI [1.01, 1.04]) predicted the hazard of leaving. Tests of the scaled Schoenfeld residuals gave no evidence against proportional hazards (all $p > .14$)."

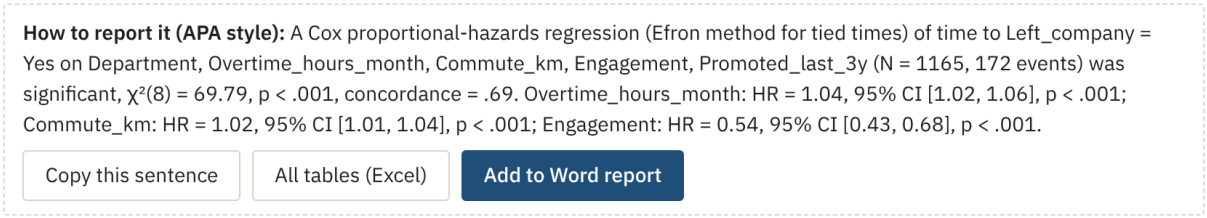


Figure 68. *The sentence for the report of the Cox regression.*

8.5 A trap: predictors that grow with time

It is tempting to add every available variable. Add Gender, Annual_salary_ln and Job_level to the model. Job_level now has a strong protective effect, HR = 0.448, 95% CI [0.299, 0.671], and the concordance rises to .770. Gender is not significant (HR = 0.785, p = .123), so the gender difference of section 8.3 does not hold once the other factors are held constant.

The job level result is mostly an artefact. Job level and salary are measured *now*, at the end of each person's time, and both grow with tenure: in the simulated company, people move up a level roughly every six years. An employee who has stayed twelve years has had time to reach level 3; one who left after one year never had the chance. The model therefore "learns" that high levels go with late leaving, partly because staying long produces high levels. A Cox predictor must be known at the start of the time (or be fixed, such as department at hiring), or be modelled as a predictor that changes over time, which SaNaStat does not offer. Keep such variables out, or use their value at hiring.

Note: what the method cannot show

- **Cause.** Hazard ratios describe associations, like any regression on observational data.
- **Different reasons for leaving.** Resignations, dismissals and retirements are different events with different causes. Analysing them as one event mixes them; a real study would at least analyse voluntary leaving separately, treating other exits as censored, and say so.
- **Informative censoring.** The method assumes that people who are censored (for example because they were transferred) would have left at the same rate as the others. If not, the curves are biased.

Exercise

1. Draw the curves by Promoted_last_3y. What share of the promoted and of the not promoted employees is estimated to be still employed at ten years? (Answer: .85 and .73.)

1. In the Cox model of section 8.4, which predictor has the smallest p-value in the proportional-hazards check, and does it fail the check? (Answer: Engagement, p = .141; no.)

1. Why must a predictor in a Cox model be known at the start of the time? Give an HR example other than job level.

9. Predicting attrition

Which employees are likely to leave in the next year, and why? This chapter first fits a logistic regression, then uses the **Machine learning** screen to compare prediction models on employees they have not seen, explain what drives the predictions, check that the predicted probabilities can be trusted, look at the trade-off between catching leavers and raising false alarms, and check whether the model works equally well for women and men. The outcome is Left_company ("Yes" for 176 of the 1,200 employees, 14.7%) in practice_hr.csv; the data are simulated.

> **Note: read this before you build a model of people**
>
> A prediction of leaving describes a group of employees with similar characteristics. It is not a statement about a particular person, it is wrong for many people, and it can harm the people it is wrong about. Section 9.10 sets out how such predictions may be used. SaNaStat's output is meant for understanding and planning (where is attention needed, and why), not for ranking or selecting individual employees.

9.1 What is being predicted

Left_company records whether the person left within the year. It does not record when, voluntarily or not, or why. Three consequences:

- People who left in January and in December are treated alike. Chapter 8 used tenure as the time to leaving; this chapter predicts the yes/no outcome.
- Dismissals, retirements and resignations are mixed. Real projects usually predict voluntary leaving only, using the reasons held in HR records.
- The predictors must be known **before** the outcome. Engagement_Q1 to Engagement_Q4 measure engagement through the same year; someone who left in the second quarter has no engagement for the third and fourth. We therefore leave them out and use the annual Engagement score, which we treat as measured at the start of the year. In real data, check the dates of every variable.

9.2 Logistic regression first

Logistic regression is the standard model for a yes/no outcome. Choose **Analyses** → **Correlation and regression** → **Binary logistic regression (Which variables predict a yes/no outcome?)**, **Outcome (two groups)** Left_company, the event "Yes", and as predictors the fifteen work-related variables: Department, Job_level, Tenure_years, Annual_salary, Performance_rating, Training_hours, Overtime_hours_month, Commute_km, Remote_days_week, Manager_support, Work_life_balance, Engagement, Job_satisfaction, Absence_days and Promoted_last_3y (Figure 69, Figure 70). We deliberately leave out Gender, Age and Education and the variables that duplicate others (section 9.3 explains why).

Binary logistic regression (Predicting Left_company = "Yes")

What it means: The model predicts "Yes" better than chance (Nagelkerke R^2 = .16, AUC = .75, acceptable). Tenure_years: each unit divides the odds of "Yes" by 1.08; Overtime_hours_month: each unit multiplies the odds of "Yes" by 1.05; Commute_km: each unit multiplies the odds of "Yes" by 1.03; Engagement: each unit divides the odds of "Yes" by 2.61.

Check **Events per variable.** EPV = 8.9.
Rule: At least 10 respondents in the smaller outcome group per coefficient (EPV ≥ 10) (Peduzzi et al. (1996)).

OK **Calibration.** Hosmer–Lemeshow $\chi^2(8) = 12.33$, p = .137.
Rule: Hosmer–Lemeshow p ≥ .05: no evidence of poor fit (Hosmer et al. (2013)).

OK **Discrimination.** AUC = .750 (acceptable).
Rule: AUC .7–.8 acceptable, .8–.9 excellent, ≥ .9 outstanding (Hosmer et al. (2013)).

Figure 69. *Binary logistic regression of Left_company on fifteen predictors: what it means and the checks.*

Model fit

χ^2	df	p	Nagelkerke R^2	McFadden R^2	Tjur R^2	AUC	95% CI of AUC
104.71	18	< .001	.163	.116	.108	.750	[.71, .79]

Copy table | Copy as picture | PNG | CSV

Coefficients

Predictor	B	SE	Wald	p	Odds ratio	95% CI of OR
(Intercept)	1.969	1.016	3.76	.053	7.163	[0.978, 52.445]
Department: HR (vs Finance)	0.198	0.394	0.25	.615	1.219	[0.563, 2.640]
Department: IT (vs Finance)	−0.280	0.361	0.60	.438	0.756	[0.373, 1.533]
Department: Operations (vs Finance)	0.454	0.340	1.79	.181	1.575	[0.809, 3.066]
Department: Sales (vs Finance)	0.142	0.311	0.21	.647	1.153	[0.627, 2.120]
Job_level	−0.187	0.225	0.69	.406	0.829	[0.534, 1.289]
Tenure_years	−0.077	0.034	5.14	.023	0.926	[0.867, 0.990]
Annual_salary	0.000	0.000	0.23	.630	1.000	[1.000, 1.000]
Performance_rating	−0.125	0.109	1.31	.252	0.882	[0.712, 1.093]
Training_hours	0.009	0.007	1.85	.174	1.009	[0.996, 1.022]
Overtime_hours_month	0.046	0.013	12.09	< .001	1.047	[1.020, 1.075]
Commute_km	0.034	0.009	14.83	< .001	1.035	[1.017, 1.053]
Remote_days_week	0.037	0.105	0.12	.727	1.037	[0.844, 1.275]
Manager_support	0.096	0.117	0.67	.411	1.101	[0.876, 1.384]
Work_life_balance	−0.114	0.135	0.71	.398	0.892	[0.685, 1.162]
Engagement	−0.960	0.227	17.95	< .001	0.383	[0.246, 0.597]
Job_satisfaction	0.003	0.207	0.00	.990	1.003	[0.669, 1.503]
Absence_days	−0.075	0.047	2.57	.109	0.928	[0.847, 1.017]
Promoted_last_3y: Yes (vs No)	−0.215	0.332	0.42	.517	0.807	[0.421, 1.545]

"Yes" = 1, "No" = 0. Categorical predictors are dummy-coded against a reference group: Department = Finance, Promoted_last_3y = No.

Figure 70. *The model fit and the coefficients with odds ratios.*

Concept: odds ratio

The *odds* of an event are its probability divided by one minus its probability: a probability of 15% is odds of 0.15 / 0.85, about 0.18. An *odds ratio* (OR) is the factor by which the odds change when the predictor rises by one unit. OR = 1 means no relationship; OR above 1 means higher odds of leaving; OR below 1, lower odds. For small probabilities the odds ratio is close to the ratio of probabilities, but the two are not the same.

Cut-off values: logistic regression

At least 10 people in the smaller outcome group per coefficient (EPV ≥ 10; Peduzzi et al., 1996). AUC .70 acceptable, .80 excellent, .90 outstanding; Hosmer–Lemeshow p ≥ .05: no evidence of poor calibration (Hosmer et al., 2013).

What it means for our example

The model uses 1,077 employees (123 with missing values are left out) and is significant, $\chi^2(18) = 104.71$, $p < .001$, Nagelkerke $R^2 = .16$, AUC = .750, 95% CI [.71, .79] (acceptable); the Hosmer–Lemeshow test does not suggest poor calibration, $\chi^2(8) = 12.33$, $p = .137$. Four predictors stand out:

- Engagement: OR = 0.383, 95% CI [0.246, 0.597], $p < .001$. Each extra point of engagement divides the odds of leaving by 2.61.
- Overtime_hours_month: OR = 1.047, [1.020, 1.075], $p < .001$: each extra hour of overtime per month raises the odds by about 5%.
- Commute_km: OR = 1.035, [1.017, 1.053], $p < .001$: each extra kilometre raises the odds by about 3.5%.
- Tenure_years: OR = 0.926, [0.867, 0.990], $p = .023$: each extra year lowers the odds by about 7%.

Promotion (OR = 0.807, $p = .517$), salary, level, performance, training, manager support, work-life balance, satisfaction and absence are not significant once the others are in the model. The salary coefficient is shown as 1.000 because one salary unit is a very small step; a standardised salary (**Standardise (z-score)** in card 4) would give an odds ratio per standard deviation instead.

The *Check* flag on the sample size is worth noting: EPV = 8.9, a little under 10, because there are 18 coefficients for 160 leavers with complete data. A model with only four predictors (Engagement, Overtime_hours_month, Tenure_years, Promoted_last_3y) uses all 1,200 employees, has EPV = 44.0 and gives nearly the same discrimination (AUC = .707, 95% CI [.67, .75], Nagelkerke $R^2 = .12$; Engagement OR = 0.390, Overtime OR = 1.051, Tenure OR = 0.906).

The table *Classification at a 0.5 cut-off* (Figure 71) classifies each person as a predicted leaver when the predicted probability is at least 0.5. Of the 160 employees with complete

data who left, the model flagged 13 (8.1%); of those who stayed, it flagged 14. Overall it classified 85.1% correctly, which is exactly what you would get by guessing that nobody leaves (85.1%). The model ranks people reasonably (AUC .75) but, with only 15% leavers, almost nobody reaches a predicted probability of 0.5. **Cut-off for the classification table** lets you choose a lower cut-off; sections 9.6 and 9.8 return to this.

Classification at a 0.5 cut-off

Observed	Predicted No	Predicted Yes	% correct
No	903	14	98.5
Yes	147	13	8.1
Overall			85.1

A respondent is predicted "Yes" when the predicted probability is at least .50. Guessing the larger group for everyone would be 85.1% correct.

Figure 71. *The classification table at the cut-off 0.5.*

How to report it

"A binary logistic regression predicting leaving from fifteen work-related variables was significant, $\chi^2(18, N = 1077) = 104.71$, $p < .001$, Nagelkerke $R^2 = .16$, AUC = .75. Engagement (OR = 0.38, 95% CI [0.25, 0.60]), overtime (OR = 1.05, 95% CI [1.02, 1.07]), commuting distance (OR = 1.03, 95% CI [1.02, 1.05]) and tenure (OR = 0.93, 95% CI [0.87, 0.99]) were significant predictors."

9.3 Machine learning on unseen employees

A regression is judged by how well it fits the data it was estimated on, which flatters it. The **Machine learning** screen asks a harder question: how well does a model predict employees it has never seen? It splits the employees into parts, trains on most, tests on the rest, and repeats, so every employee is predicted by a model that did not see them (*cross-validation*). Every step that learns from the data, including filling in missing values and choosing the settings of the model (*tuning*), is repeated inside each training set so that nothing leaks from the test employees (*nested cross-validation*).

Concept: cross-validation

The employees are split into k parts (*folds*). Each model is trained on k − 1 parts and tested on the remaining one, k times over. Repeating the whole split several times averages out the luck of one particular split (Kuhn & Johnson, 2013).

Step 1 · Prediction task. Choose **Outcome to predict** Left_company. SaNaStat recognises a yes/no outcome and asks which answer is the event; keep "Yes". It ticks every variable as a predictor. Click **Select none** and tick the fifteen predictors of section 9.2 (Figure 72). The reasons for what we left out:

- **EmployeeID** is not offered at all. SaNaStat says under the list: *Not offered: EmployeeID (type ID: a respondent number would let a model memorise respondents)*.
- **Gender, Age and Education** are personal characteristics. We do not want a model of leaving to use them (section 9.10). The remaining variables can still stand in for them, which is why section 9.9 checks the model by gender.
- **Skills_test_before and Skills_test_after** belong to the training evaluation, and **Skills_gain** is made from them.
- **Engagement_Q1 to Engagement_Q4** overlap with Engagement and run through the year of the outcome (section 9.1).
- **Annual_salary_ln** is a duplicate of Annual_salary.

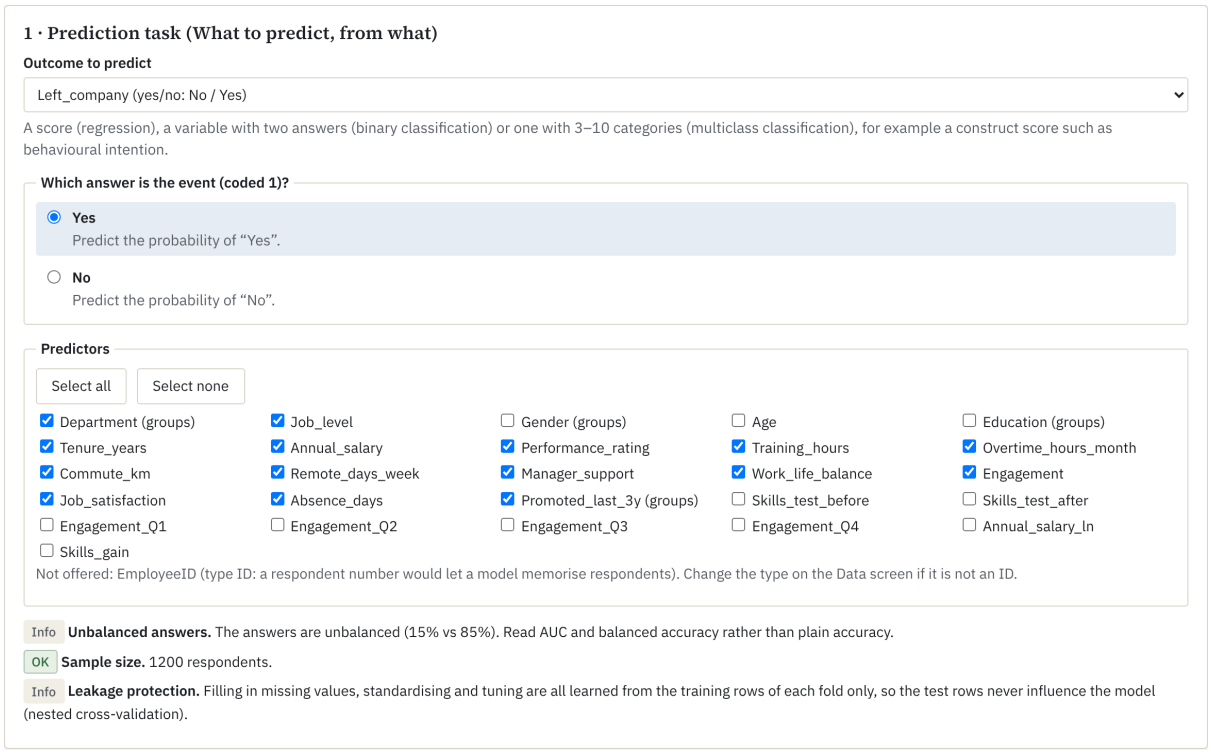


Figure 72. *1 · Prediction task: outcome, event, the fifteen predictors and the checks.*

Under the predictors SaNaStat shows three checks. *Unbalanced answers (Info):* "The answers are unbalanced (15% vs 85%). Read AUC and balanced accuracy rather than plain accuracy." *Sample size (OK):* 1200 respondents. *Leakage protection (Info):* filling in missing values, standardising and tuning are learned from the training rows of each fold only.

Step 2 · Validation plan. The recommended plan for 1,200 employees is *Repeated 10-fold cross-validation, 3 repeats* (30 train/test splits) (Figure 73). The main measure for a yes/no outcome is AUC. The random seed (2026) makes every split reproducible.

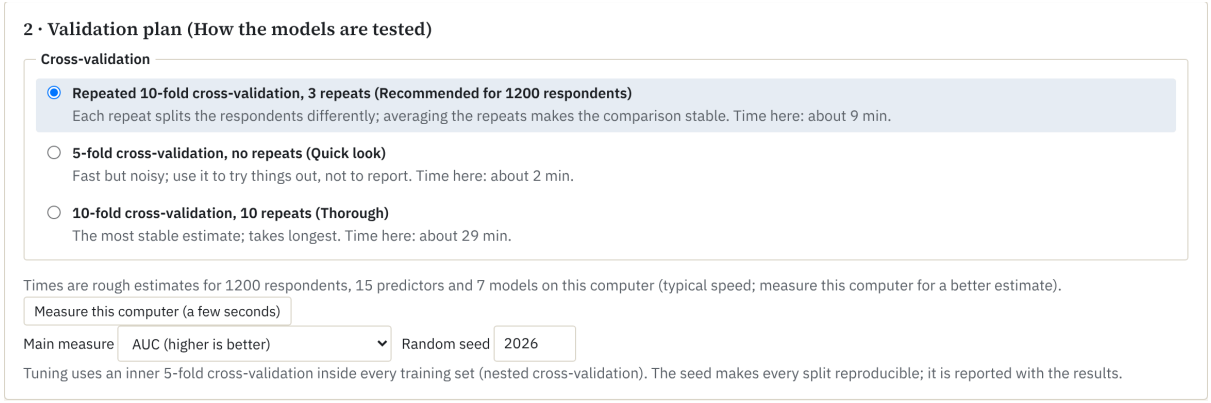


Figure 73. *2 · Validation plan: the recommended repeated 10-fold cross-validation, the main measure and the seed.*

Step 3 · Models. Seven models are compared with the always-included null model: logistic regression, elastic net (a penalised logistic regression), a decision tree (CART), a random forest, XGBoost and LightGBM, and optionally a stacked ensemble (we left it unticked). Click **Run the comparison**. The screen estimates the time before the run (here *216 steps on 1 background worker · about 9 min.*); on our two-core computer the run took 1159 s, about 19 minutes. **Stop** cancels the run.

> **Tip: use the Quick look to try things out**
>
> The **5-fold cross-validation, no repeats (Quick look)** plan is fast but noisy. Use it to try settings; use the recommended plan for the results you report.

9.4 Results · Model comparison

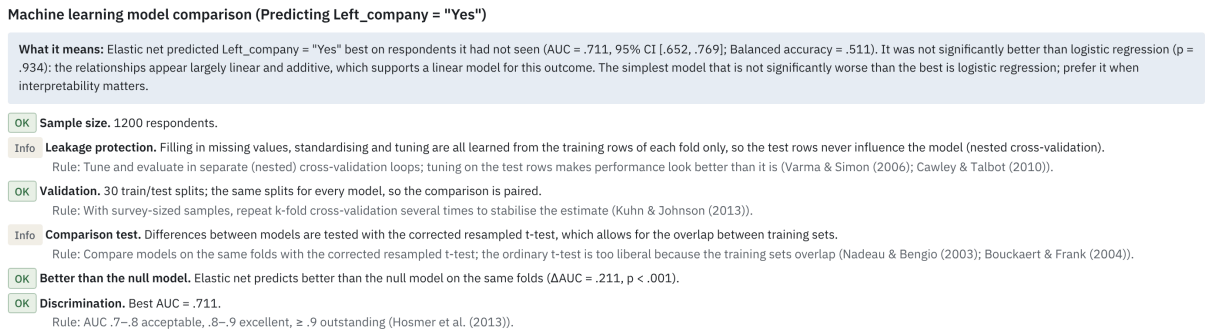


Figure 74. *The model comparison: what it means and the checks.*

Out-of-sample performance (nested cross-validation)

Model	AUC M (SD)	95% CI	Accuracy	Balanced accuracy	F1	MCC	Brier score	Δ vs best (p)
Elastic net	.711 (.075)	[.652, .769]	.850	.511	.053	.056	0.116	best
Logistic regression	.710 (.076)	[.651, .769]	.850	.525	.108	.111	0.117	−.001 (.934)
XGBoost	.693 (.078)	[.632, .754]	.848	.506	.037	.030	0.118	−.018 (.097)
LightGBM	.688 (.074)	[.630, .746]	.847	.502	.024	.009	0.119	−.023 (.045)
Random forest	.678 (.065)	[.627, .728]	.851	.510	.050	.054	0.119	−.033 (.038)
Decision tree (CART)	.645 (.075)	[.587, .704]	.847	.527	.117	.105	0.126	−.065 (.009)
Null model	.500 (.000)	[.500, .500]	.853	.500	.000	.000	0.125	−.211 (< .001)

AUC M (SD) over 30 test folds; 95% CI with the Nadeau–Bengio corrected variance. Δ vs best: mean per-fold difference from the best model, corrected resampled t-test p.

Figure 75. *The out-of-sample performance of the seven models.*

Figure 74 and Figure 75 show the report on the tab **Comparison**. All numbers are measured on employees the models did not see.

Table 18

Out-of-sample prediction of Left_company = "Yes" (10-fold cross-validation, 3 repeats)

Model	AUC M (SD)	95% CI	Accuracy	Balanced accuracy	Δ vs best (p)
Elastic net	.711 (.075)	[.652, .769]	.850	.511	best
Logistic regression	.710 (.076)	[.651, .769]	.850	.525	−.001 (.934)
XGBoost	.693 (.078)	[.632, .754]	.848	.506	−.018 (.097)
LightGBM	.688 (.074)	[.630, .746]	.847	.502	−.023 (.045)
Random forest	.678 (.065)	[.627, .728]	.851	.510	−.033 (.038)
Decision tree (CART)	.645 (.075)	[.587, .704]	.847	.527	−.065 (.009)
Null model	.500 (.000)	[.500, .500]	.853	.500	−.211 (< .001)

Concept: AUC

The *area under the ROC curve* is the probability that a randomly chosen leaver gets a higher predicted risk than a randomly chosen stayer. .50 is guessing and 1.00 is perfect. It measures how well the model *ranks* people, and it does not depend on a cut-off, which makes it suitable for an outcome that is rare.

What it means for our example

The elastic net predicts best (AUC = .711, 95% CI [.652, .769]), but it is not significantly better than plain logistic regression (ΔAUC = .001, corrected resampled t(29) = 0.08, p = .934). The flexible models are no better and some are significantly worse (random forest p = .038, LightGBM p = .045, decision tree p = .009). SaNaStat therefore names logistic regression as the simplest model that is not significantly worse than the best; its summary says that the relationships "appear largely linear and additive, which supports a linear model for this outcome". Only the null model, which has AUC .500 by definition, is clearly worse (ΔAUC = .211, p < .001).

Notice the accuracy column. The null model, which says "nobody leaves" for everyone, is 85.3% accurate, more than any of the real models (.847 to .851). Balanced accuracy is close to .50 for every model. These numbers do not say that the models are useless (the AUCs show that they rank employees much better than chance); they show that accuracy and balanced accuracy at a fixed cut-off of 0.5 are the wrong yardsticks when only 15% of employees leave (section 9.6).

How to report it

"Seven prediction models were compared with 10-fold cross-validation repeated 3 times (30 train/test splits), with every model tuned by an inner 5-fold cross-validation inside each training set (nested cross-validation; seed 2026). Elastic net performed best (AUC = .711, 95% CI [.652, .769]), but not significantly higher than logistic regression, ΔAUC = .001, corrected resampled t(29) = 0.08, p = .934."

9.5 Results · Explanations: what drives leaving?

On the tab **Explanations**, **Explain this model** chooses the model to explain. It opens on the recommended model (logistic regression); choose *Elastic net (best)*. The same model is then used on the other tabs. The explanations describe what the model learned.

- **SHAP feature importance** shares each prediction among the predictors. The mean absolute SHAP value shows how far a predictor moves the predictions, in log-odds of leaving; the report adds the direction and a second opinion, the *permutation importance* (how much the AUC falls on the test employees when the predictor is shuffled).
- The **SHAP importance** chart (Figure 76) draws the mean absolute SHAP value of every predictor; the *SHAP summary* plot under it shows every employee as a dot (right means the predictor raised that person's predicted risk; red means a high value of the predictor).
- **Accumulated local effects (ALE)** (Figure 77) show how the prediction changes across the range of each main driver, with the other predictors kept realistic.

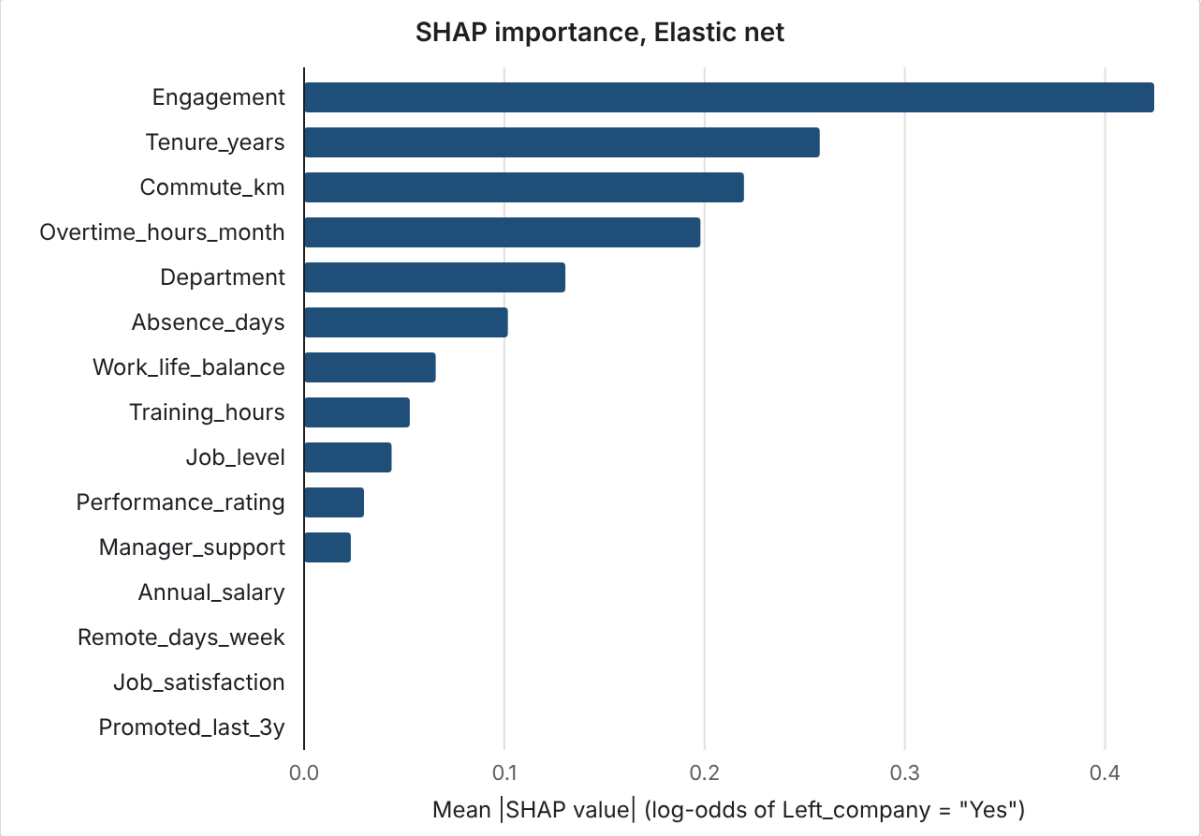


Figure 76. *SHAP importance of each predictor in the elastic net.*

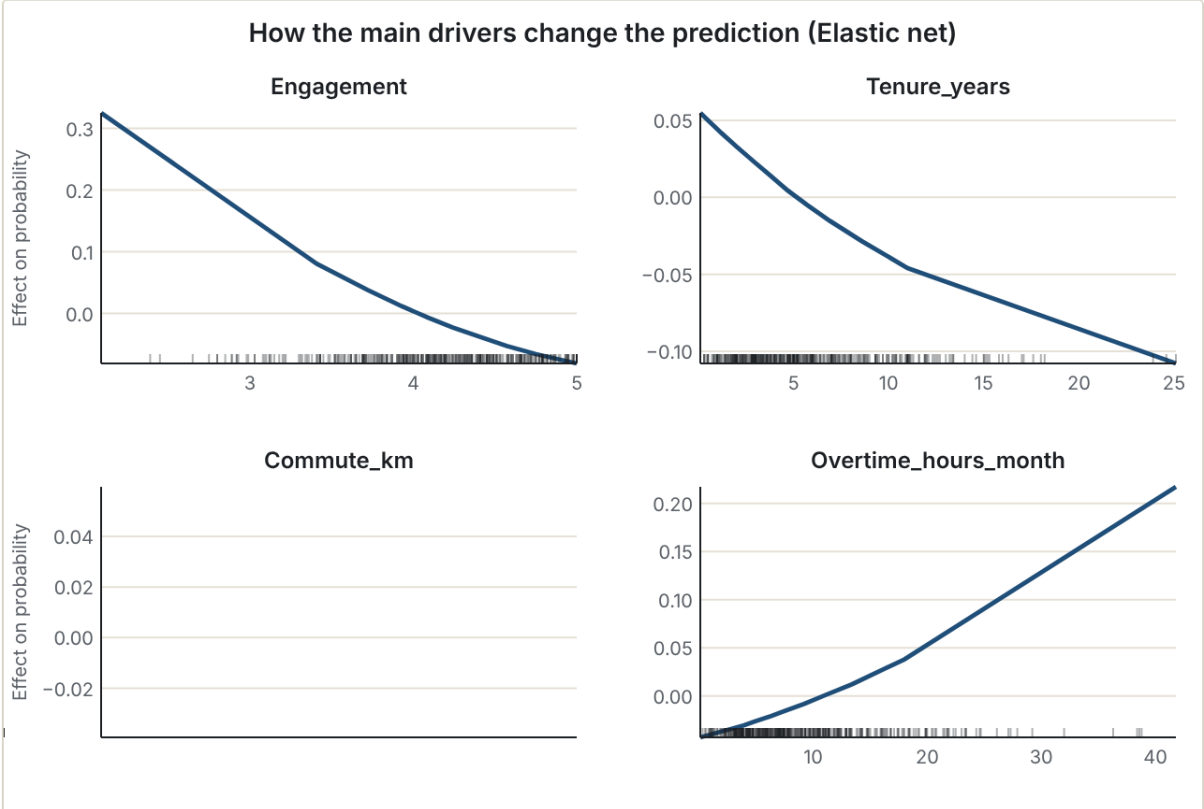


Figure 77. *Accumulated local effects of the four main drivers.*

Table 19

The strongest drivers in the elastic net model (from the program)

Predictor	Mean abs. SHAP	SHAP rank	Permutation importance M (SD)	Permutation rank
Engagement	0.425	1	0.086 (0.050)	1
Tenure_years	0.257	2	0.024 (0.036)	4
Commute_km	0.220	3	0.030 (0.024)	2
Overtime_hours_month	0.198	4	0.026 (0.032)	3
Department	0.130	5	0.007 (0.013)	5
Absence_days	0.102	6	0.003 (0.012)	6

What it means for our example

Engagement is by far the strongest driver (mean |SHAP| = 0.425): higher engagement, lower chance of leaving. It is followed by tenure (0.257, lower chance), commute (0.220, higher chance) and overtime (0.198, higher chance). The two methods agree (Spearman ρ = .88 between the rankings), and Engagement is ranked first by every one of the six models in the stability table; the correlations with the other models' rankings range from .68 (random forest) to .93 (logistic regression). The table across the 30 test folds adds that only Engagement helped the predictions in practically every fold (97%); the importance of the others varies from fold to fold. Annual_salary, Remote_days_week, Job_satisfaction and Promoted_last_3y add nothing once the others are included (SHAP 0.000). Job_satisfaction is almost the same as Engagement (r = .66), so it adds nothing beyond it; that is different from saying that satisfaction does not matter.

The accumulated local effects are nearly straight lines: across its observed range, engagement changes the predicted probability of leaving by about 0.41, tenure by about 0.16, commute by about 0.25 and overtime by about 0.26. At the lowest engagement (2.09) the effect is +0.325; at the highest (5.00) it is −0.081. The effects grow most for the longest commutes (from 35.60 km) and the heaviest overtime (from 18.00 hours a month).

Choose **Decision tree (CART)** under **Explain this model** to see one readable tree (Figure 78). Its first question is *Engagement ≤ 4.255?*; the next questions are about tenure, engagement again, absence, overtime (*Overtime_hours_month ≤ 22?*) and commuting (*Commute_km ≤ 29.2?*). Each box shows the employees who reach it and the predicted share of leavers there. The tree is easy to read but, as the comparison showed, less accurate than the linear models.

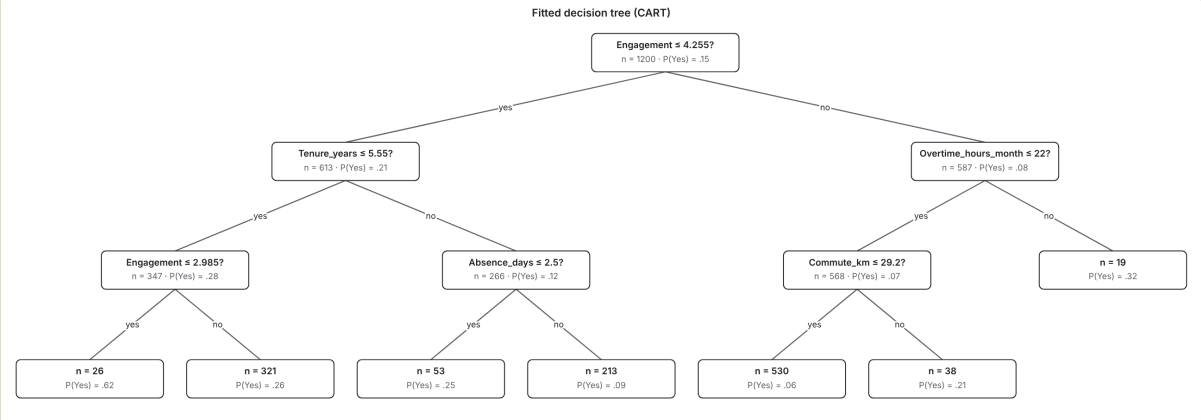


Figure 78. *The fitted decision tree (CART): depth 3, 7 leaves.*

Cut-off values: reading explanations

SHAP values, permutation importance and accumulated local effects describe what the model learned, **not causal effects** (Molnar, 2022). Correlated predictors share credit. Trust a driver that ranks high on both SHAP and permutation importance. ALE is preferred over partial dependence when predictors are correlated (Apley & Zhu, 2020).

Note: drivers are not levers

That engagement is the strongest predictor does not mean that raising someone's engagement will stop them from leaving. Engagement may be a symptom (a person already planning to leave answers a survey differently), and unmeasured causes may sit behind both. Use the drivers to decide where to look and what to ask, and test any intervention.

9.6 The imbalance problem

Only 14.7% of the employees leave. This is typical for attrition and for many other HR outcomes (promotions, misconduct, long absences), and it causes three problems.

1. **Accuracy is misleading.** Predicting "stays" for everybody is right 85.3% of the time. Any model that has to beat this on accuracy must be very good indeed.
2. **A fixed cut-off of 0.5 flags almost nobody.** A probability above 0.5 is rare when the average is 0.15, so sensitivity (the share of leavers flagged) is very low at 0.5 (8.1% for the logistic regression of section 9.2, and 3% for the elastic net in section 9.8), and so are F1 (between .024 and .117 for the real models) and the balanced accuracy (Table 18).
3. **There are few leavers to learn from.** With 176 leavers, a flexible model has little to work on, which is another reason why the simple models do as well.

What SaNaStat does about it, and what it does not:

- It **flags** the imbalance before the run and uses **AUC** as the main measure for a yes/no outcome.
- It reports balanced accuracy, F1, MCC (the Matthews correlation coefficient, which is high only if both groups are predicted well) and the Brier score (the average squared error of the predicted probabilities; lower is better; the null model's is 0.125 and the elastic net's 0.116) next to accuracy.
- It checks **calibration**, which does not depend on a cut-off (section 9.7).
- It lets you **try cut-offs** and shows precision, recall and the average precision against the prevalence (section 9.8).
- It does **not** rebalance the data or weight the classes: the models are fitted to the data as they are, and the remedy is to judge them by AUC, precision and recall.

9.7 Results · Calibration

A model is *calibrated* if its predicted probabilities mean what they say: of the employees given a 20% chance, about 20% should leave. This matters more than the AUC if the numbers will be used for planning. The tab **Calibration & intervals** checks calibration on the out-of-fold predictions (Figure 79).

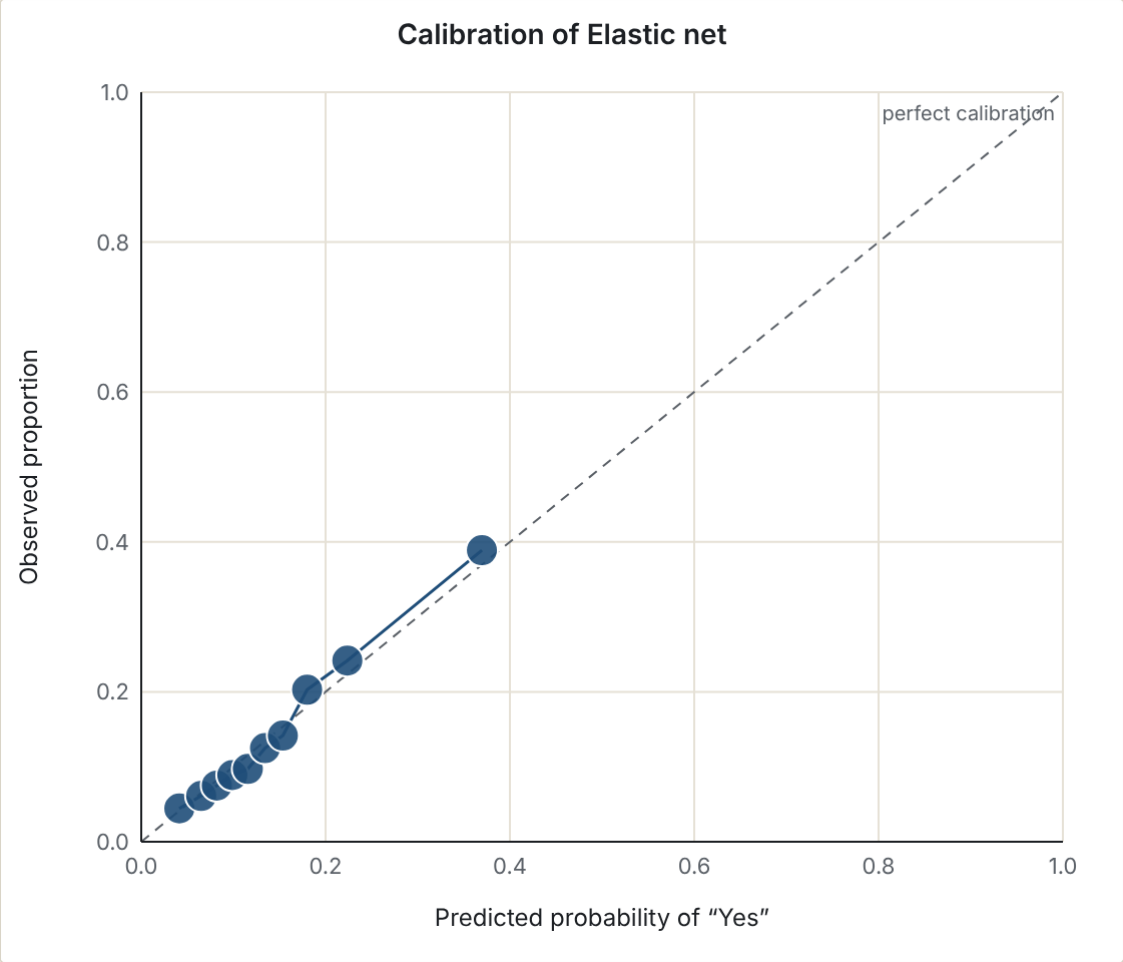


Figure 79. *Reliability diagram of the elastic net: predicted probability against the observed share of leavers.*

What it means for our example

The elastic net is well calibrated: the calibration slope is 1.01, 95% CI [0.78, 1.24], and the calibration-in-the-large is 0.00, 95% CI [−0.16, 0.17] (no systematic over- or under-prediction). A predicted 30% can be read as roughly a 30% chance. The ten equal-sized groups in the reliability table follow the diagonal closely: the group with the lowest predicted risk (mean .041) had an observed share of .044, and the group with the highest (mean .370) had .389.

Cut-off values: calibration

Slope close to 1 and calibration-in-the-large close to 0 (Van Calster et al., 2019). A slope well below 1 means the predictions are too extreme (overfitting); well above 1, too cautious.

Calibration says that the probabilities are honest; it does not say that they are sharp. The highest-risk tenth of employees has an average predicted risk of 37%: it is the best the model can do for them, and it also means that most of them stay.

9.8 Results · Cut-offs, precision and recall

A model gives each employee a probability of leaving. To act on it, someone has to turn the probability into yes or no with a *cut-off*. For a yes/no outcome the tab **Cut-off** has a section for this, *Results · Cut-offs, precision and recall (What happens when probabilities become yes or no?)* (Figure 80). It uses the out-of-sample predictions of the model chosen on the Explanations tab (here the elastic net). Suppose that HR can offer a supportive "stay conversation" to a limited number of employees. Whom to invite depends on a trade-off:

- *Recall* (also called sensitivity) is the share of all leavers who are flagged.
- *Precision* is the share of the flagged employees who really would have left.

Flagging more people raises recall and lowers precision. The box **Cut-off (probability, between 0 and 1)** starts at 0.5.

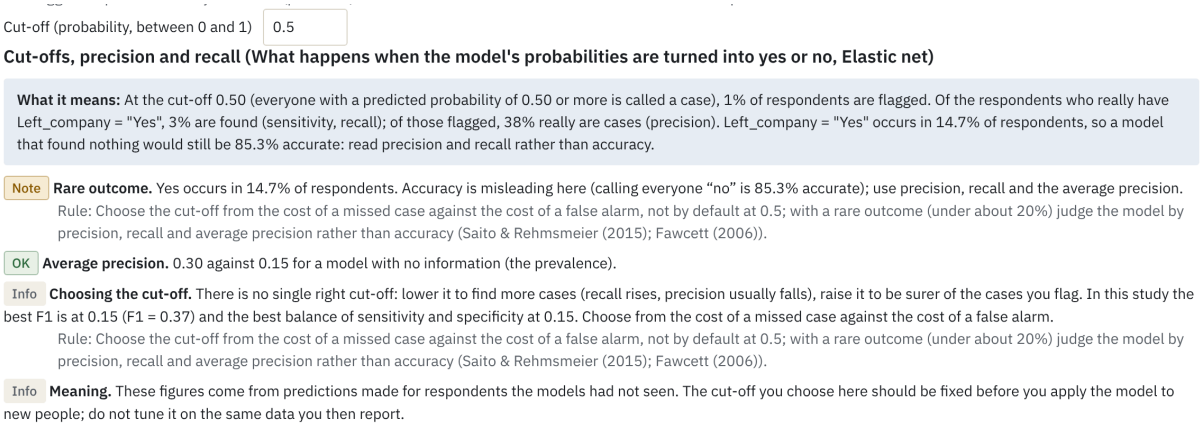


Figure 80. *The cut-off section at the default cut-off of 0.50: what it means and the checks.*

What it means for our example

At 0.50 the program says: "1% of respondents are flagged. Of the respondents who really have Left_company = "Yes", 3% are found (sensitivity, recall); of those flagged, 38% really are cases (precision)." Of the 176 leavers, on average 5.3 are flagged and 170.7 are missed. The default cut-off is useless for this purpose, as section 9.6 predicted. The program also reports the average precision, 0.30, against 0.15 for a model with no information (the prevalence): the model is better than chance, but far from precise.

Type a lower cut-off, for example 0.2 (Figure 81). The program now says: "19% of respondents are flagged. ... 42% are found (sensitivity, recall); of those flagged, 32% really are cases (precision)." The counts table gives the numbers behind these percentages, averaged over the three repeats of the cross-validation (Table 20).

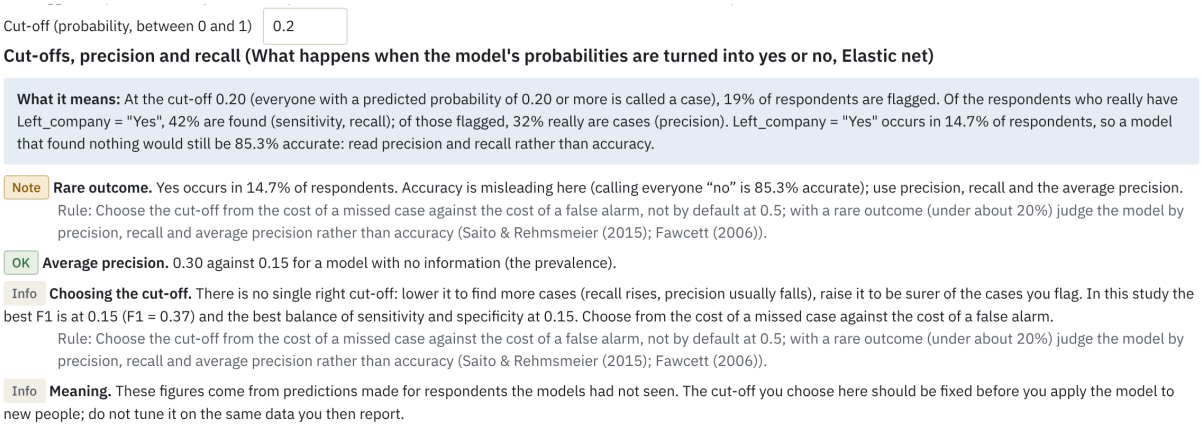


Figure 81. *The same section at a cut-off of 0.20.*

Table 20

Counts at the cut-off 0.20 (average over the repeats of the cross-validation).

Actual outcome	Predicted: case	Predicted: not a case	Total
Actually Yes	74.3	101.7	176
Actually not Yes	158	866	1024
Total	232.3	967.7	1200

Table 21 is part of the program's table *Measures at every cut-off*, and the chart shows the same measures (Figure 82).

Table 21

Measures at some cut-offs (from the program; elastic net, out-of-sample).

Cut-off	Flagged	Sensitivity (recall)	Specificity	Precision	F1	Accuracy
0.10	64%	0.85	0.39	0.19	0.32	0.46
0.15	37%	0.64	0.68	0.26	0.37	0.67

Cut-off	Flagged	Sensitivity (recall)	Specificity	Precision	F1	Accuracy
0.20	19%	0.42	0.85	0.32	0.36	0.78
0.30	7%	0.18	0.95	0.40	0.25	0.84
0.50	1%	0.03	0.99	0.38	0.06	0.85

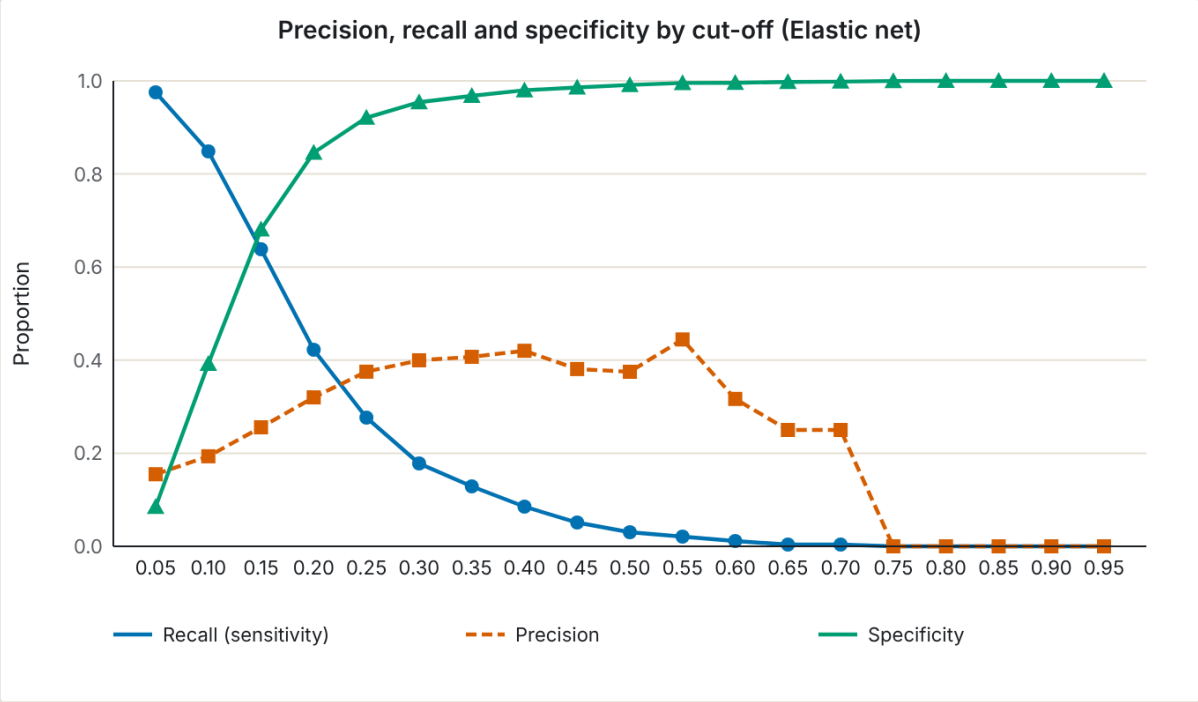


Figure 82. *Precision, recall and specificity of the elastic net at every cut-off.*

What it means for our example

At the cut-off 0.20 about 232 employees (19%) are flagged, which reaches 42% of the leavers (74 of 176), but 158 of the 232 flagged would have stayed anyway. Lowering the cut-off to 0.10 reaches 85% of the leavers but flags 64% of all employees, of whom only 19% would leave: barely more than the base rate of 14.7%, so the flag tells you little. Raising it to 0.30 gives a more accurate list (precision 0.40) that finds only 18% of the leavers. The program notes that the best F1 is at 0.15 (F1 = 0.37), but "There is no single right cut-off": choose it from the cost of a missed case against the cost of a false alarm.

The program also warns that the cut-off "should be fixed before you apply the model to new people; do not tune it on the same data you then report". The numbers in Table 21 are honest estimates for a given cut-off, but picking the best-looking cut-off from the same predictions makes the reported precision and recall slightly optimistic.

There is no single right threshold. It depends on the cost of the intervention, the cost of losing someone and, above all, whether being flagged is a benefit or a burden for the

employee. If the "intervention" could damage careers, the precision of the list must be much higher than it is here.

9.9 Results · Fairness by group

A model that never received Gender can still treat women and men differently, through the predictors that go with gender. The tab **Fairness** compares the out-of-fold predictions of the chosen model across the groups of a grouping variable, at the cut-off chosen on the Cut-off tab (here 0.20). Choose Gender in **Group by** (Figure 83).

Fairness by group (Does Elastic net work equally well for every Gender group?)

What it means: With a cut-off of .20, elastic net flags 21% of female, 19% of male as "Yes". The lowest selection rate is 0.88 times the highest (male vs female), within the four-fifths guideline. It finds 35% (male) to 52% (female) of the actual "Yes" answers and wrongly flags 15% (female) to 16% (male) of the others.

OK **Honest predictions.** Every respondent is predicted by a model that did not see them (out-of-fold predictions, averaged over 3 repeats).

OK **Four-fifths check.** Lowest ÷ highest selection rate = 0.88 (male vs female); the guideline asks for at least .80. The four-fifths rule comes from US employment law (adverse impact in hiring and promotion). It is a rule of thumb, not a statistical test, and it depends on the cut-off.

OK **Calibration by group.** In every group the average predicted probability is within 5 points of the observed rate.

Info **Trade-off.** When the outcome is more common in some groups than in others, a model cannot have equal error rates and equal calibration in every group at once; decide which matters most for the intended use.

Rule: When groups differ in how common the outcome is, equal error rates and equal calibration in every group cannot all hold at once (Chouldechova (2017)).

Out-of-fold predictions of Elastic net by Gender (cut-off .20)

Group	n	Observed rate of "Yes"	Selection rate (ratio)	True-positive rate (ratio)	False-positive rate (ratio)	Precision (ratio)	Observed − predicted
female	563	.17	.21 (1.00)	.52 (1.00)	.15 (1.00)	.40 (1.00)	0.02
male	637	.13	.19 (0.88)	.35 (0.68)	.16 (0.94)	.25 (0.61)	−0.02

Selection rate: share predicted "Yes" (probability ≥ .20). True-positive rate: share of actual "Yes" found (recall). False-positive rate: share of the others wrongly flagged. Precision: share of the flagged who really are "Yes". In brackets: ratio to the best group (1 = best; the best has the highest rate, or for false positives the lowest). Observed − predicted: calibration-in-the-large (0 = the predicted probabilities are right on average; positive = the model under-predicts in that group). Groups under 20 respondents are marked small.

Figure 83. *Fairness of the elastic net by Gender at the cut-off 0.20.*

Concept: what "fair" can mean for a prediction

The table compares, for each group, the *selection rate* (share flagged), the *true-positive rate* (share of the real leavers who are flagged), the *false-positive rate* (share of the stayers wrongly flagged), the *precision* and the calibration (*Observed − predicted*). These are different ideas of fairness, and when the outcome is more common in one group than in another, a model cannot satisfy all of them at once (Chouldechova, 2017). Decide which matters most for the intended use before you look.

What it means for our example

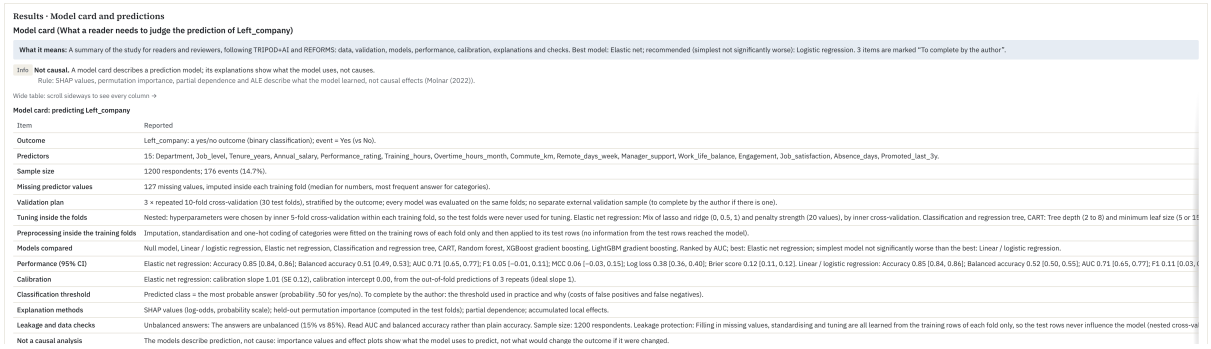
At the cut-off .20 the model flags 21% of women and 19% of men; the ratio of the selection rates is 0.88, within the four-fifths guideline. It finds 52% of the women who left but only 35% of the men who left (ratio 0.68), with similar false-positive rates (.15 and .16), and the precision is .40 for women and .25 for men. The predicted probabilities are right on average in both groups (observed − predicted 0.02 and −0.02). Women left more often in these data (an observed rate of .17 against .13), and the model, without seeing gender, reproduces that difference through the other predictors. Whether the lower recall for men matters depends on the use: for an offer of support, it means that men who are about to leave are offered it less often.

Choose Department to see a stronger difference: Operations employees are flagged at 30%, IT employees at 14% (ratio 0.47, *Check: below the four-fifths guideline*), and 26% of the Operations employees who stayed are flagged wrongly, against 10% in Finance, HR and IT. Department is one of the predictors, and the program says so: "Department is one of the predictors, so the model can treat the groups differently on purpose. Leaving it out does not remove differences that run through other predictors."

How to report it

"Group fairness of the elastic net model was examined with its out-of-fold predictions across the 2 groups of Gender (cut-off .20). Selection rates ranged from .19 (male) to .21 (female; ratio 0.88), true-positive rates from .35 to .52, and false-positive rates from .15 to .16."

The tab **Model card** summarises the study for readers and reviewers (Figure 84): the outcome, predictors, sample, validation, tuning, performance with intervals, calibration, the explanation methods, the checks and the seed, with three items marked "To complete by the author" (an external validation sample, the threshold used in practice and why, and the intended use and population). Under it, **Download predictions (CSV)** and **Save predictions as a variable** give each employee's out-of-fold prediction, and *Score a new file with Elastic net* predicts employees of a new file with the final model.


Results · Model card and predictions

Model card (What a reader needs to judge the prediction of Left_company)

What it means: A summary of the study for readers and reviewers, following TRIPOD+AI and REFORMS: data, validation, models, performance, calibration, explanations and checks. Best model: Elastic net; recommended (simplest not significantly worse): Logistic regression. 3 items are marked "To complete by the author".

Info Not causal. A model card describes a prediction model; its explanations show what the model uses, not causes.
Rule: SHAP values, permutation importance, partial dependence and ALE describe what the model learned, not causal effects (Molnar (2022)).

Wide table: scroll sideways to see every column →

Model card: predicting Left_company

Item	Reported
Outcome	Left_company: a yes/no outcome (binary classification); event = Yes (vs No).
Predictors	15: Department, Job_level, Tenure_years, Annual_salary, Performance_rating, Training_hours, Overtime_hours_month, Commute_km, Remote_days_week, Manager_support, Work_life_balance, Engagement, Job_satisfaction, Absence_days, Promoted_last_3y.
Sample size	1200 respondents; 176 events (14.7%).
Missing predictor values	127 missing values, imputed inside each training fold (median for numbers, most frequent answer for categories).
Validation plan	3 × repeated 10-fold cross-validation (30 test folds), stratified by the outcome; every model was evaluated on the same folds; no separate external validation sample (to complete by the author if there is one).
Tuning inside the folds	Nested: hyperparameters were chosen by inner 5-fold cross-validation within each training fold, so the test folds were never used for tuning. Elastic net regression: Mix of lasso and ridge (0, 0.5, 1) and penalty strength (20 values), by inner cross-validation. Classification and regression tree, CART: Tree depth (2 to 8) and minimum leaf size (5 or 1
Preprocessing inside the training folds	Imputation, standardisation and one-hot coding of categories were fitted on the training rows of each fold only and then applied to its test rows (no information from the test rows reached the model).
Models compared	Null model, Linear / logistic regression, Elastic net regression, Classification and regression tree, CART, Random forest, XGBoost gradient boosting, LightGBM gradient boosting. Ranked by AUC; best: Elastic net regression; simplest model not significantly worse than the best: Linear / logistic regression.
Performance (95% CI)	Elastic net regression: Accuracy 0.85 [0.84, 0.86]; Balanced accuracy 0.51 [0.49, 0.53]; AUC 0.71 [0.65, 0.77]; F1 0.05 [−0.01, 0.11]; MCC 0.06 [−0.03, 0.15]; Log loss 0.38 [0.36, 0.40]; Brier score 0.12 [0.11, 0.12]. Linear / logistic regression: Accuracy 0.85 [0.84, 0.86]; Balanced accuracy 0.52 [0.50, 0.55]; AUC 0.71 [0.65, 0.77]; F1 0.11 [0.03,
Calibration	Elastic net regression: calibration slope 1.01 (SE 0.12), calibration intercept 0.00, from the out-of-fold predictions of 3 repeats (ideal slope 1).
Classification threshold	Predicted class = the most probable answer (probability .50 for yes/no). To complete by the author: the threshold used in practice and why (costs of false positives and false negatives).
Explanation methods	SHAP values (log-odds, probability scale); held-out permutation importance (computed in the test folds); partial dependence; accumulated local effects.
Leakage and data checks	Unbalanced answers: The answers are unbalanced (15% vs 85%). Read AUC and balanced accuracy rather than plain accuracy. Sample size: 1200 respondents. Leakage protection: Filling in missing values, standardising and tuning are all learned from the training rows of each fold only, so the test rows never influence the model (nested cross-val
Not a causal analysis	The models describe prediction, not cause: importance values and effect plots show what the model uses to predict, not what would change the outcome if it were changed.

Figure 84. *The start of the model card.*

Note: predictions of named people

The predictions file and a saved prediction variable hold one predicted probability per employee number. That turns a study of groups into a list of individuals with risk scores. Do not create or share such a list without a clear policy, a legal basis and the involvement of the people responsible for data protection (section 9.10).

9.10 Ethical use of attrition predictions

- **Use it for groups and for support.** Findings such as "low engagement, long commutes and heavy overtime go with leaving" support action on workload,

commuting arrangements and management. They do not justify naming employees as "flight risks".

- **Never use a prediction to withhold something.** A person flagged as likely to leave must not be denied training, promotion or responsibility because "they will leave anyway". That turns a prediction into a cause.
- **Keep it away from line managers' decisions about individuals.** A list of names with risk scores changes how people are treated, and it can become true because of the way people are treated.
- **Check fairness, and do not stop at one number.** We did not give the model Gender, Age or Education, but section 9.9 showed that it still finds leavers among women and men at different rates. Compare the rates of flagging and the error rates across groups before any use, and involve the people responsible for equality.
- **Be open about it.** Employees should know that such analyses exist and how the results are used. In many places, automated decisions about people are regulated: check the law that applies to you.
- **Remember the limits.** The model was tested on the same company in the same year. Next year's conditions will differ; recheck calibration on new data before trusting the numbers, and update the model when conditions change.

Exercise

1. Run the logistic regression with only Engagement, Overtime_hours_month, Tenure_years and Promoted_last_3y. What are the AUC and the EPV? (Answer: AUC = .707, EPV = 44.0.)

1. In the model comparison, which model has the highest accuracy, and why is this not a good sign? (Answer: the null model, .853; it predicts that nobody leaves.)

1. Which four predictors have the highest SHAP importance? (Answer: Engagement 0.425, Tenure_years 0.257, Commute_km 0.220, Overtime_hours_month 0.198.)

1. Using Table 21, what share of the leavers is found at the cut-off 0.30, and what share of the flagged employees really leave? (Answer: 18% of the leavers; precision 0.40.)

1. Type the cut-off 0.15 in the program. What does it say about the share flagged? (Answer: 37% flagged, sensitivity 0.64, precision 0.26.)

1. On the Fairness tab, choose Education. Do the selection rates of the four groups meet the four-fifths guideline at the cut-off 0.20?

1. Describe in three sentences how you would explain the results of this chapter to a works council, without naming individuals.

10. Employee types: clustering

Are there groups of employees who resemble each other, so that one policy does not have to fit everyone? Cluster analysis looks for such groups. It is **exploratory**: it always returns groups, whether or not real groups exist, and the groups depend on which variables and which method you choose. This chapter uses the **Clustering** screen on practice_hr.csv (simulated data; no real employee appears) and is honest about how weak the structure here is.

Concept: what clustering does

K-means clustering places a chosen number of centres in the space of the variables and assigns every employee to the nearest centre, then moves the centres and repeats until nothing changes. Hierarchical clustering instead merges the most similar employees step by step. Both need the variables to be on comparable scales, so SaNaStat standardises each variable before clustering. The result is a *description* of the sample, not proof that employee types exist.

10.1 Choosing the variables

Open **Clustering**. Under *Variables that describe the respondents (numeric)* click **Select none** and tick the variables that define a type. We choose four: Tenure_years, Overtime_hours_month, Engagement and Annual_salary_ln (Figure 85). Annual_salary_ln is used instead of Annual_salary because salary is strongly right-skewed. Leave the method at k-means and the number of clusters on "the best silhouette (Recommended)", with the seed at its default. Under **Compare the clusters on other variables (optional)** tick Manager_support, Work_life_balance, Job_satisfaction, Absence_days, Age, Commute_km and Performance_rating.

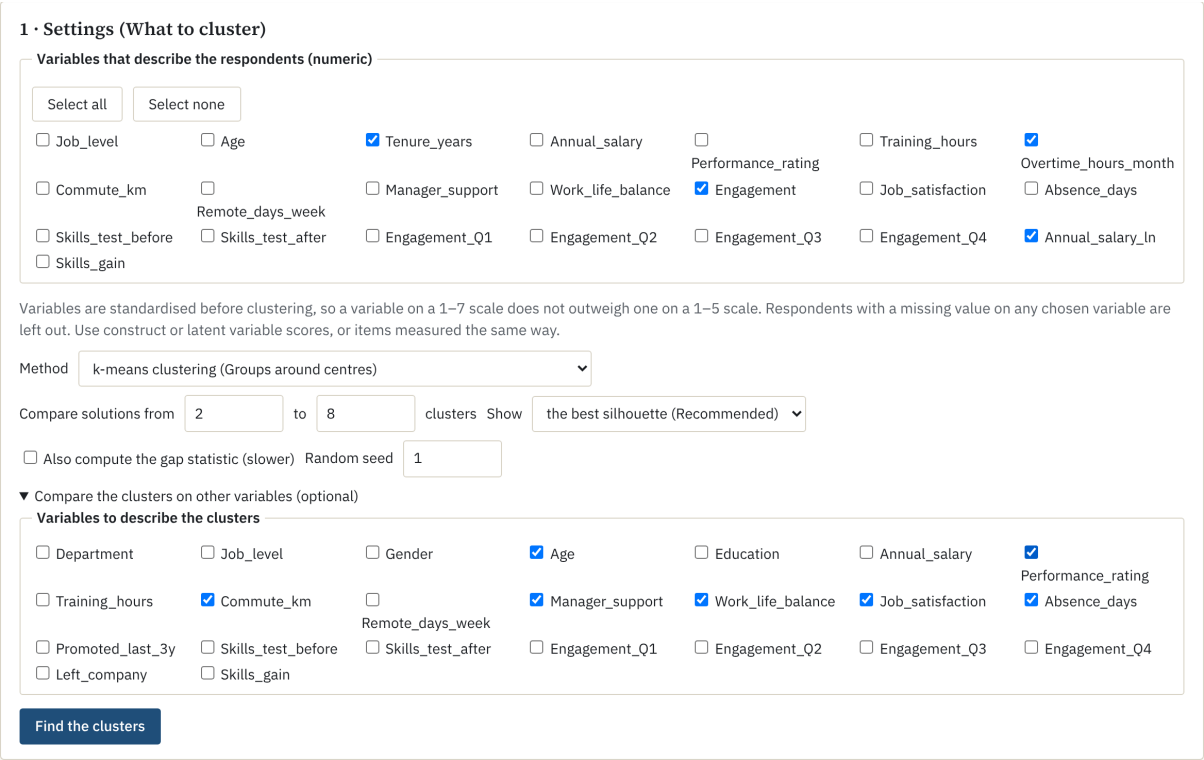


Figure 85. *The Clustering settings: four variables, k-means, the best silhouette, and the variables to describe the clusters.*

Two cautions about the choice of variables:

- **Clusters reflect what you put in.** If you add many pay variables, you will get pay clusters. Choose variables that are meaningful for the question and measured in the same way.
- **Do not include the outcome you want to explain.** We keep Left_company out of the clustering and compare the clusters on it afterwards. If it were inside, the finding that one cluster contains more leavers would be circular.

Respondents with a missing value on any chosen variable are left out; here all 1,200 are used.

10.2 The default solution: two clusters

Select **Find the clusters**. The first result is the solution with the best average silhouette, which is two clusters. The text says: "K-means clustering split the 1200 respondents into 2 clusters (average silhouette = .27: weak structure, possibly artificial). Cluster 1 (64%): low Tenure_years and Annual_salary_ln. Cluster 2 (36%): high Tenure_years and Annual_salary_ln."

Concept: the average silhouette

For each employee, the silhouette compares the average distance to the others in the same cluster with the average distance to the nearest other cluster. It runs from −1 to 1. The average over all employees is judged with the scale of Kaufman and Rousseeuw (1990): .71 to 1.00 strong structure, .51 to .70 reasonable, .26 to .50 weak (possibly artificial), .25 or less no substantial structure.

What it means for our example

The silhouette of .27 is only just above the .25 line. The two clusters split the employees mainly by tenure and pay, which are strongly related: this is a gradient, not two separate groups. The program says so itself ("weak structure, possibly artificial").

10.3 Comparing solutions and choosing k

Below the first result, SaNaStat lists the solutions with two to eight clusters (Table 22).

Table 22

Solutions with 2 to 8 clusters (from the program).

Clusters	Within-cluster sum of squares	Average silhouette	Calinski–Harabasz	Smallest cluster
2	3503.9	.270	443.1	35.6%
3	2864.1	.265	404.5	17.0%
4	2410.6	.228	395.2	13.9%
5	2134.9	.219	372.9	9.8%
6	1933.9	.216	353.9	7.8%
7	1805.4	.203	329.8	7.6%
8	1693.2	.208	312.4	6.4%

The within-cluster sum of squares always falls as clusters are added, so one looks for an elbow; the program notes: "silhouette suggests 2, elbow 3." The silhouette chart is shown in Figure 86; the program also draws the sum-of-squares chart.

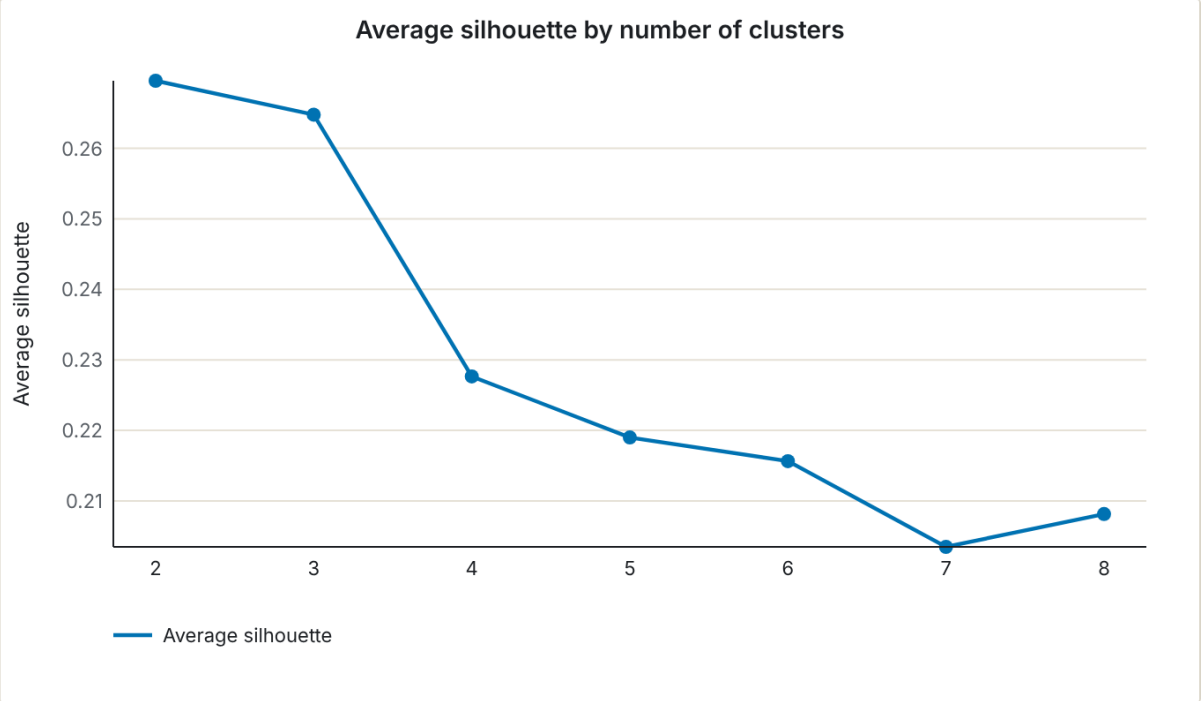


Figure 86. *Average silhouette by number of clusters.*

The criteria disagree, as they often do. Two clusters give the best silhouette, but the pattern is a tenure gradient that the manager already knows. Three clusters have a silhouette that is almost the same (.265 against .270) and an elbow. We therefore choose three, which is a decision of the analyst, not of the program: change **Show** from the best silhouette to "3 clusters" and select **Find the clusters** again (Figure 87).

Cluster analysis (Are there groups of similar respondents?)

What it means: K-means clustering split the 1200 respondents into 3 clusters (average silhouette = .26: weak structure, possibly artificial). Cluster 1 (55%): low Tenure_years. Cluster 2 (28%): high Tenure_years and Annual_salary_ln. Cluster 3 (17%): high Overtime_hours_month; low Engagement.

Note **Cluster structure.** Average silhouette .26: weak structure, possibly artificial.
Rule: Average silhouette: .71–1.00 strong structure, .51–.70 reasonable, .26–.50 weak (possibly artificial), .25 or less no substantial structure (Kaufman & Rousseeuw (1990)).

OK **Cluster sizes.** The smallest cluster holds 17.0% of respondents (204).
Rule: Every cluster should hold at least 5% of the respondents (and at least 20) to be interpretable and stable (Common practice (as for latent profiles, Spurk et al., 2020)).

OK **Stability.** Mean adjusted Rand index .90 (lowest .75) between the full solution and 20 random 80% subsamples.
Rule: Mean adjusted Rand index between the full solution and 80% subsamples of .80 or more: the clusters are stable (informal guide; the resampling idea is from Hennig, 2007) (Hubert & Arabie (1985); Hennig (2007)).

Info **Number of clusters.** Chosen by you; silhouette suggests 2, elbow 3. Try neighbouring values and prefer the solution that is stable and easy to interpret.

Figure 87. *The three-cluster solution: what it means and the checks.*

10.4 Reading the three-cluster solution

The program summarises: "K-means clustering split the 1200 respondents into 3 clusters (average silhouette = .26: weak structure, possibly artificial). Cluster 1 (55%): low Tenure_years. Cluster 2 (28%): high Tenure_years and Annual_salary_ln. Cluster 3 (17%): high Overtime_hours_month; low Engagement." The checks say the smallest cluster holds 17.0% of respondents (204), and that the mean adjusted Rand index over 20 random 80% subsamples is .90 (lowest .75). The program's informal guide is .80 or more for stable clusters; here the mean passes but the lowest value does not, so the solution is reasonably, not perfectly, stable.

Concept: stability

The program repeats the clustering on 20 random 80% subsamples and compares each result with the full solution using the adjusted Rand index (Hubert & Arabie, 1985; the resampling idea is from Hennig, 2007). A value of 1 means identical groupings; 0 means agreement no better than chance. A stable solution is necessary, though not sufficient, for the clusters to be real.

Table 23 gives the size and silhouette of each cluster, and Table 24 the profile.

Table 23

The three clusters (from the program).

Cluster	Respondents	Share	Average silhouette
Cluster 1	657	54.8%	.33
Cluster 2	339	28.2%	.20
Cluster 3	204	17.0%	.16

Table 24

Mean of each clustering variable by cluster, original units (standardised value in brackets).

Variable	Cluster 1 (n = 657)	Cluster 2 (n = 339)	Cluster 3 (n = 204)	η^2 (descriptive)
Tenure_years	3.46 (−0.52)	10.28 (1.19)	4.32 (−0.30)	.56
Overtime_hours_month	6.87 (−0.36)	7.54 (−0.25)	19.41 (1.58)	.51
Engagement	4.34 (0.25)	4.11 (−0.14)	3.86 (−0.57)	.10
Annual_salary_ln	10.54 (−0.45)	10.98 (1.06)	10.58 (−0.32)	.45

The charts on the screen include the cluster profiles (Figure 88) and the respondents on the first two principal components (Figure 89).

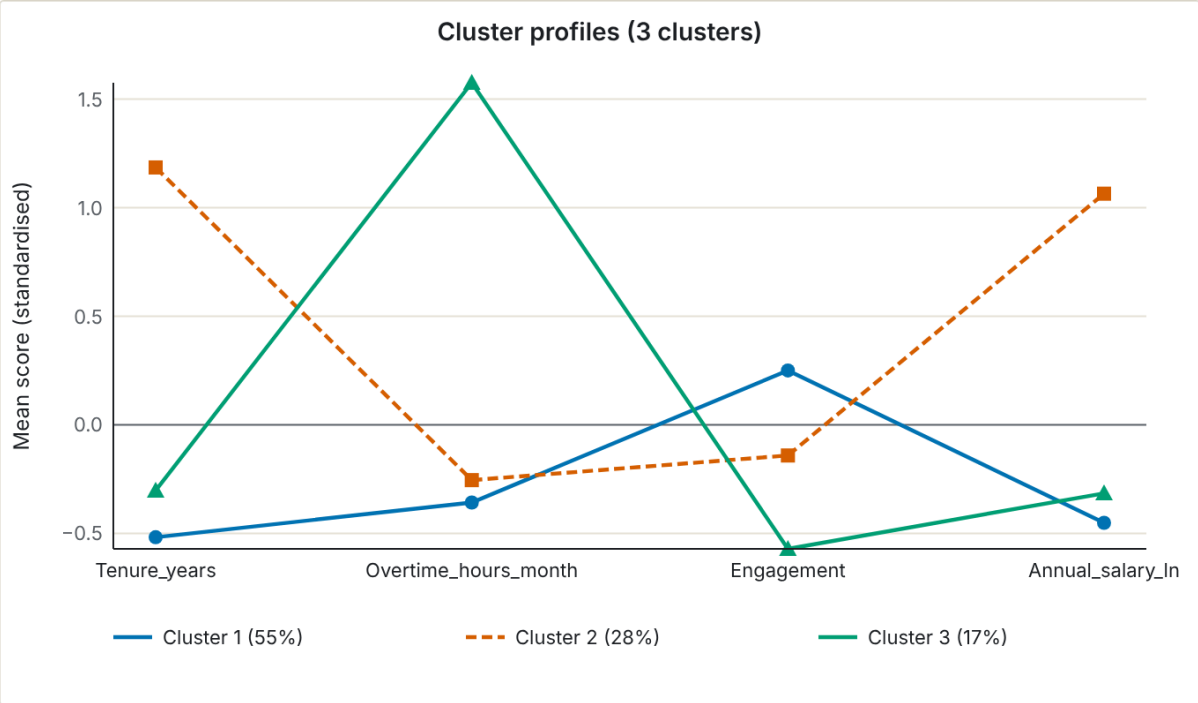


Figure 88. *Cluster profiles: mean standardised scores of the three clusters.*

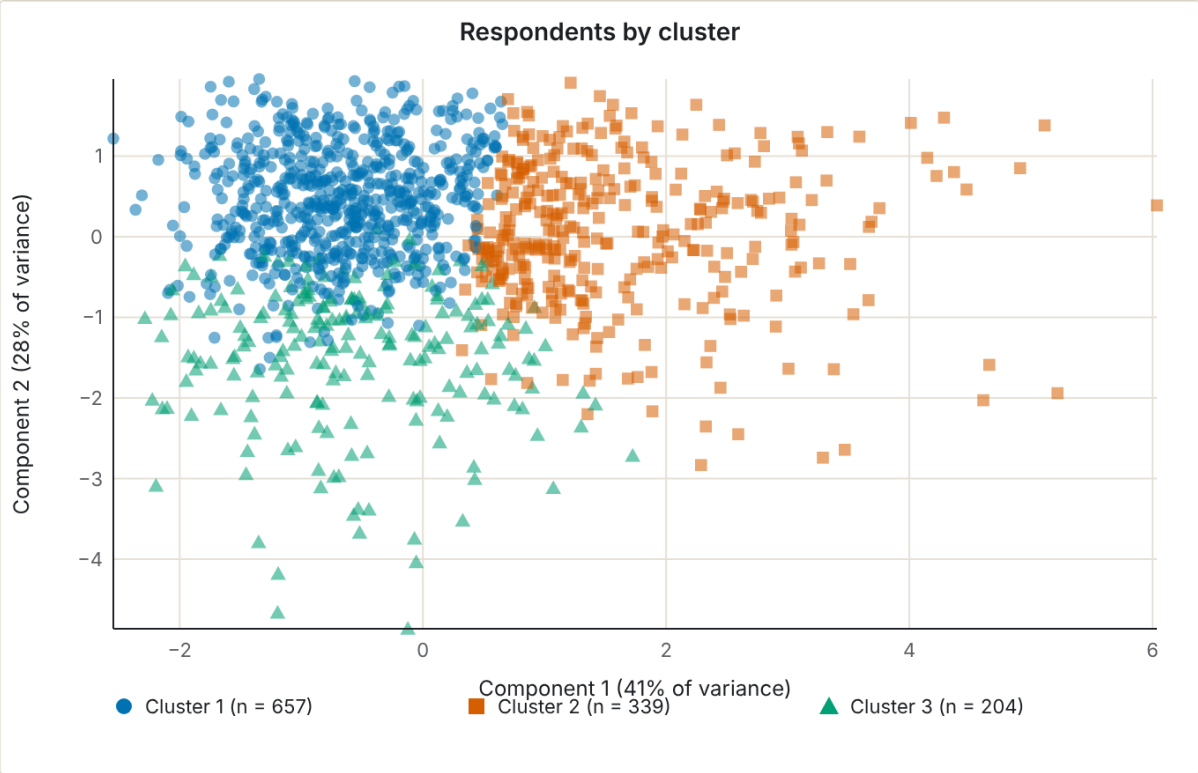


Figure 89. *Respondents on the first two principal components (69% of the variance), coloured by cluster.*

Note: differences are large by construction

The program states: "The clusters were built from these variables, so differences between clusters are large by construction." The η^2 column only shows which variables separate the clusters most; it carries no p-values on purpose. Here tenure and overtime separate them most, engagement least.

Putting the profiles into words, with a descriptive label as a working name only:

- **Cluster 1, "newer employees" (55%).** Short tenure (mean 3.46 years), average overtime (6.87 hours a month), the highest engagement (4.34).
- **Cluster 2, "established and well paid" (28%).** Long tenure (10.28 years) and the highest Annual_salary_ln (10.98).
- **Cluster 3, "overworked" (17%).** Much overtime (19.41 hours a month) and the lowest engagement (3.86).

Tip: be careful with labels

A label such as "overworked" is shorthand for a pattern in the numbers. It must never become a label attached to a person, in a file or in conversation. State the criterion instead: "employees with more than about 19 hours of overtime a month on average in this cluster". Also remember the silhouette of .26: cluster boundaries are fuzzy, and many employees sit near a boundary (the average silhouettes of clusters 2 and 3, .20 and .16, are low).

10.5 Describing the clusters with other variables

The variables ticked under **Compare the clusters on other variables (optional)** were not used to form the clusters. For each, the program gives the mean, SD and Welch's test (Table 25).

Table 25

Other numeric variables by cluster: mean (SD) and Welch's test (from the program).

Variable	Cluster 1	Cluster 2	Cluster 3	Welch F (df)	p	η^2
Age	36.39 (9.37)	39.26 (7.16)	36.20 (8.88)	17.04 (2, 527.2)	< .001	.02
Performance_rating	3.39 (0.84)	3.47 (0.81)	3.35 (0.89)	1.66 (2, 502.4)	.192	.00
Commute_km	14.09 (9.14)	14.21 (9.27)	14.83 (9.40)	0.47 (2, 482.9)	.626	.00
Manager_support	3.42 (0.87)	3.31 (0.89)	3.28 (0.85)	2.93 (2, 485.8)	.055	.00
Work_life_balance	3.51 (0.72)	3.41 (0.72)	2.86 (0.74)	61.37 (2, 504.3)	< .001	.10
Job_satisfaction	3.73 (0.57)	3.80 (0.61)	3.40 (0.68)	24.21 (2, 461.4)	< .001	.05

Variable	Cluster 1	Cluster 2	Cluster 3	Welch F (df)	p	η^2
Absence_days	3.85 (2.00)	3.78 (1.90)	4.15 (2.03)	2.28 (2, 509.9)	.103	.00

The program adds: "These variables were not used to form the clusters, so the tests are meaningful, but the clusters were still chosen from the same sample: confirm in new data before calling them types." Cluster 3 stands out on work-life balance (2.86 against 3.51 and 3.41) and job satisfaction (3.40 against 3.73 and 3.80). Absence does not differ.

10.6 Keeping the clusters as a variable

Under *2 · Keep the clusters (Use them in other analyses)*, type a name for the new variable in **Name of the new variable** (we use Employee_type) and select **Add to the data** (Figure 90). The program confirms: *Added "Employee_type".* Employee_type is now a grouping variable that every other screen can use. If you run the clustering again and keep it under the same name, the button reads **Replace it with these clusters** and replaces the old column; choose a new name to keep both.

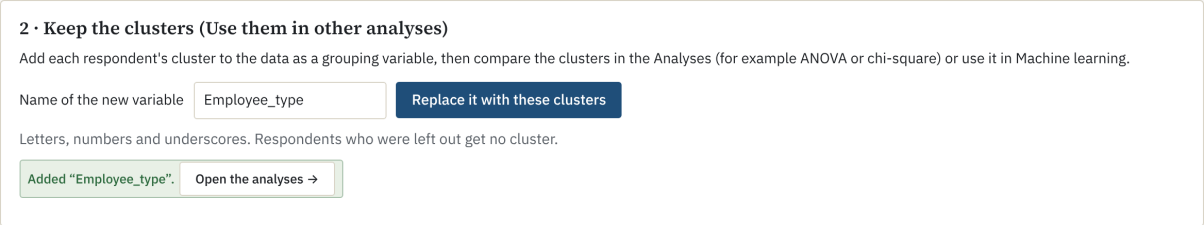


Figure 90. *Keeping the clusters as the new variable Employee_type.*

10.7 Do the employee types differ in leaving?

Choose **Analyses** → **Chi-square, Fisher, McNemar** → **Chi-square test of independence** with Employee_type and Left_company (Figure 91). The program reports $\chi^2(2, N = 1200) = 32.08$, $p < .001$, Cramér's V = .16, 90% CI [.11, .21] (small); all expected counts are adequate (smallest 29.92).

Counts (row %)

Employee_type	No	Yes	Total
Cluster 1	569 (87%)	88 (13%)	657
Cluster 2	306 (90%)	33 (10%)	339
Cluster 3	149 (73%)	55 (27%)	204

Copy table | Copy as picture | PNG | CSV

Tests

Test	Value	df	p
Pearson chi-square	32.08	2	< .001
Likelihood ratio	28.82	2	< .001

Copy table | Copy as picture | PNG | CSV

Adjusted standardised residuals

Employee_type	No	Yes
Cluster 1	1.37	−1.37
Cluster 2	3.03 *	−3.03 *
Cluster 3	−5.45 *	5.45 *

* beyond ±1.96: this cell holds clearly more (positive) or fewer (negative) cases than expected.

Figure 91. *Chi-square test of Employee_type by Left_company.*

Table 26

Employee_type by Left_company: counts (row %) and adjusted residuals (from the program).

Employee_type	No	Yes	Total	Residual (No)	Residual (Yes)
Cluster 1	569 (87%)	88 (13%)	657	1.37	−1.37
Cluster 2	306 (90%)	33 (10%)	339	3.03 *	−3.03 *
Cluster 3	149 (73%)	55 (27%)	204	−5.45 *	5.45 *

27% of cluster 3 left, against 13% of cluster 1 and 10% of cluster 2. The adjusted residuals beyond ±1.96 mark cluster 2 (fewer leavers than expected) and cluster 3 (more).

How to report it

"A chi-square test of independence showed a significant association between Employee_type and Left_company, $\chi^2(2, N = 1200) = 32.08$, $p < .001$, Cramér's V = .16, 90% CI [.11, .21]."

Two ANOVAs complete the picture. Job_satisfaction differs across the types (Welch's $F(2, 461.40) = 24.21$, $p < .001$, $\omega^2 = .048$, small); Games–Howell comparisons show cluster 3 below both others (differences 0.33 and 0.40 points, both $p < .001$) and no difference between clusters 1 and 2 ($p = .125$). Absence_days does not differ ($F(2, 1197) = 2.35$, $p = .096$, $\omega^2 = .002$).

What it means

The group with heavy overtime and low engagement has about twice the leaving rate of the others, and also reports lower work-life balance and satisfaction. This agrees with chapters 8 and 9 (engagement and overtime among the main factors). It is an association in one simulated sample, not evidence that overtime causes leaving, and the clusters were chosen using overtime and engagement themselves, so the finding is partly built in.

Exercise

1. Run the clustering with Engagement, Overtime_hours_month, Work_life_balance and Job_satisfaction (all four are survey or workload measures). Compare the silhouette with the .27 found here. (Check your own result: no answer is given, because it is not part of the planted analyses.)

1. In Table 22, how much does the average silhouette fall between 3 and 4 clusters? (Answer: from .265 to .228, a fall of .037; calculated by hand.)

1. What share of cluster 2 left the company? (Answer: 33 of 339, about 10%.)

1. Why is no p-value for tenure reported in Table 24? (Answer: tenure was used to build the clusters, so a test would be circular.)

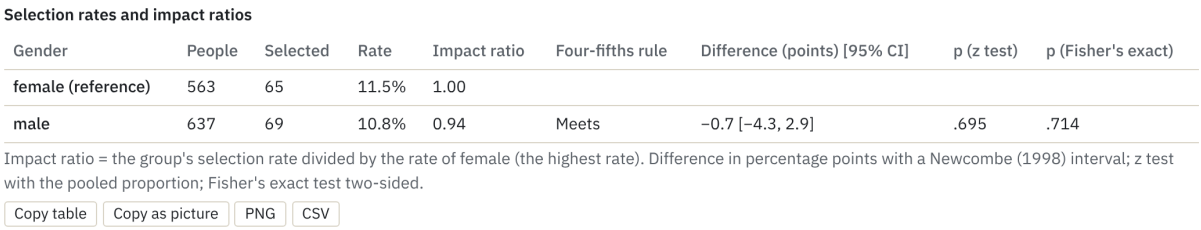
1. Write two sentences that describe cluster 3 for a managers' meeting without using a label for people.

11. Reporting, exporting and the report builder

SaNaStat does not write your report for you, but every result comes with the sentence you would write, tables ready to paste and a report builder that collects the results you choose into one Word document. This chapter shows how to take single tables and charts out of SaNaStat and how to build the report.

11.1 Copying and exporting one table or chart

Every table of a result has a row of buttons under it (Figure 92): **Copy table** copies it as formatted text, so that it arrives in a word processor or a spreadsheet as a table; **Copy as picture** copies an image, and **PNG** saves it as a picture file; **CSV** saves the numbers as a CSV file. Every chart has **Copy picture**, **PNG** and **SVG**; SVG is a vector format that stays sharp at any size. Under the sentence of each result (Figure 93), **Copy this sentence** copies the APA-style sentence, **All tables (Excel)** saves every table of the result in one Excel workbook, one sheet per table, and **Add to Word report** adds the result to the report.



Selection rates and impact ratios

Gender	People	Selected	Rate	Impact ratio	Four-fifths rule	Difference (points) [95% CI]	p (z test)	p (Fisher's exact)
female (reference)	563	65	11.5%	1.00				
male	637	69	10.8%	0.94	Meets	−0.7 [−4.3, 2.9]	.695	.714

Impact ratio = the group's selection rate divided by the rate of female (the highest rate). Difference in percentage points with a Newcombe (1998) interval; z test with the pooled proportion; Fisher's exact test two-sided.

Copy table | Copy as picture | PNG | CSV

Figure 92. *The table of the adverse-impact check with its buttons: Copy table, Copy as picture, PNG and CSV.*

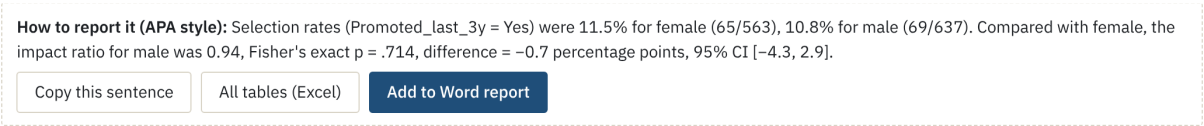


Figure 93. *The sentence for the report, with Copy this sentence, All tables (Excel) and Add to Word report.*

11.2 Collecting results in the Word report

Add the results you may want to keep as you work; they are kept in the order added, and a result that has been added is marked *In the Word report*. In the session of this chapter we added five results: the adverse-impact check of promotion by gender (section 4.8), the t-test of Annual_salary by Gender (section 4.2), the hierarchical regression of Annual_salary_ln (section 4.4), the Kaplan–Meier curves by department (section 8.3) and the negative binomial regression of Absence_days (section 7.3).

Click **Word report** in the menu (it shows the number of results, here **Word report (5)**) to open the report builder (Figure 94). The first card, *Word report (5 results)*, lists the results with their finding:

- untick a result to leave it out (it stays in the list, crossed out);

- the arrows move a result up or down (tables and figures are numbered in this order), and **Remove** drops it;
- **Parts of this result** chooses what goes in for that result alone.

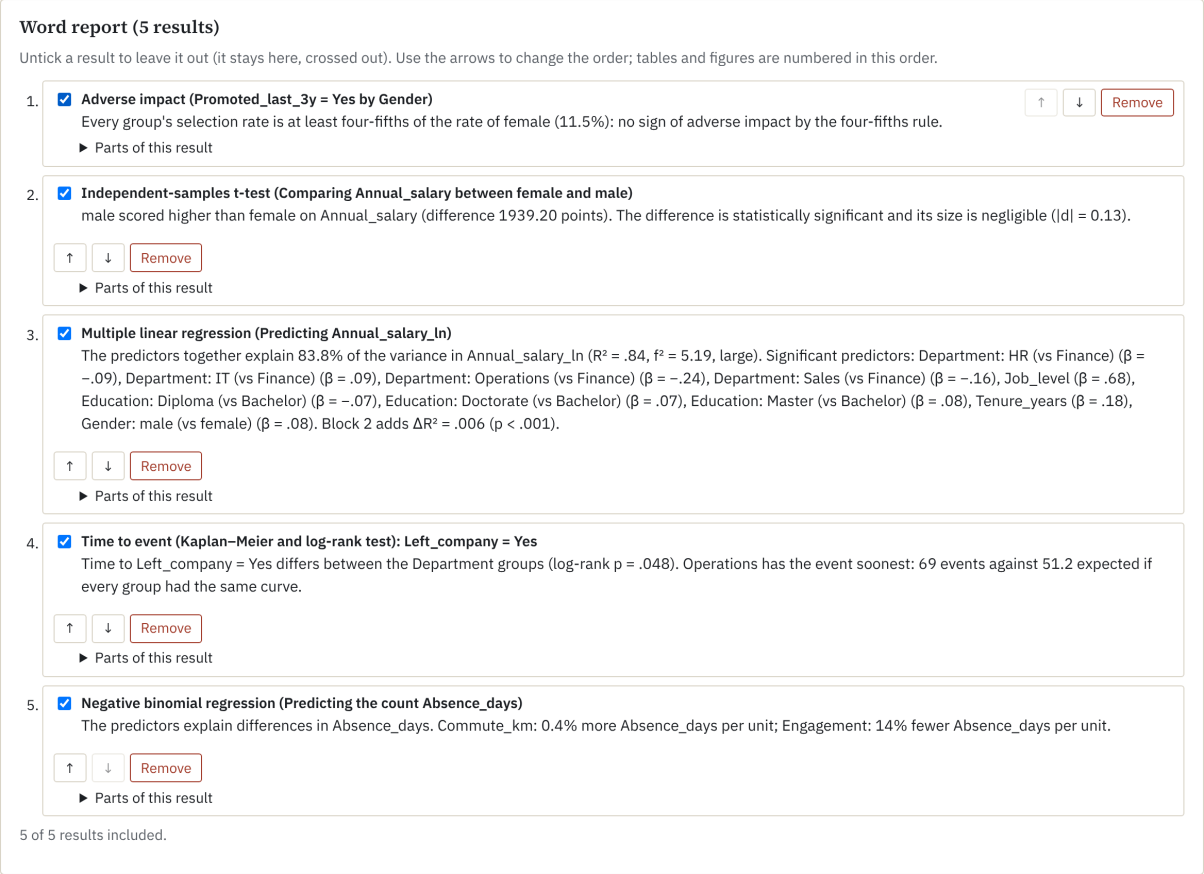


Figure 94. *The report builder with the five results of this chapter.*

The cards below build the report as a whole:

- **What goes in** (Figure 95): the parts of every result (APA sentence, Interpretation, Checks, Tables, Figures, References) and the **Methods and reporting checklist**, left out or added before or after the results. The checklist covers the data, the flow of respondents, missing data, constructs and reliability, α, the software and the seeds, and a model card for each machine learning study; items only you know are marked "To complete by the author".
- **Style**: the **Report title** (we typed "HR analytics: practice data"), the font (Times New Roman 12 pt, Calibri 11 pt or Arial 11 pt), the **Line spacing** (single, 1.5 or double), the **Paper** (A4 or US Letter) and an optional **Title page**.
- *Preview and download* (Figure 96): **Download Word report (.docx)**, **Download all tables (Excel)**, **Download as a web page (HTML, prints as PDF)** and **Show the preview**. The Word file is named after the title, here "HR analytics practice data (SaNaStat).docx".

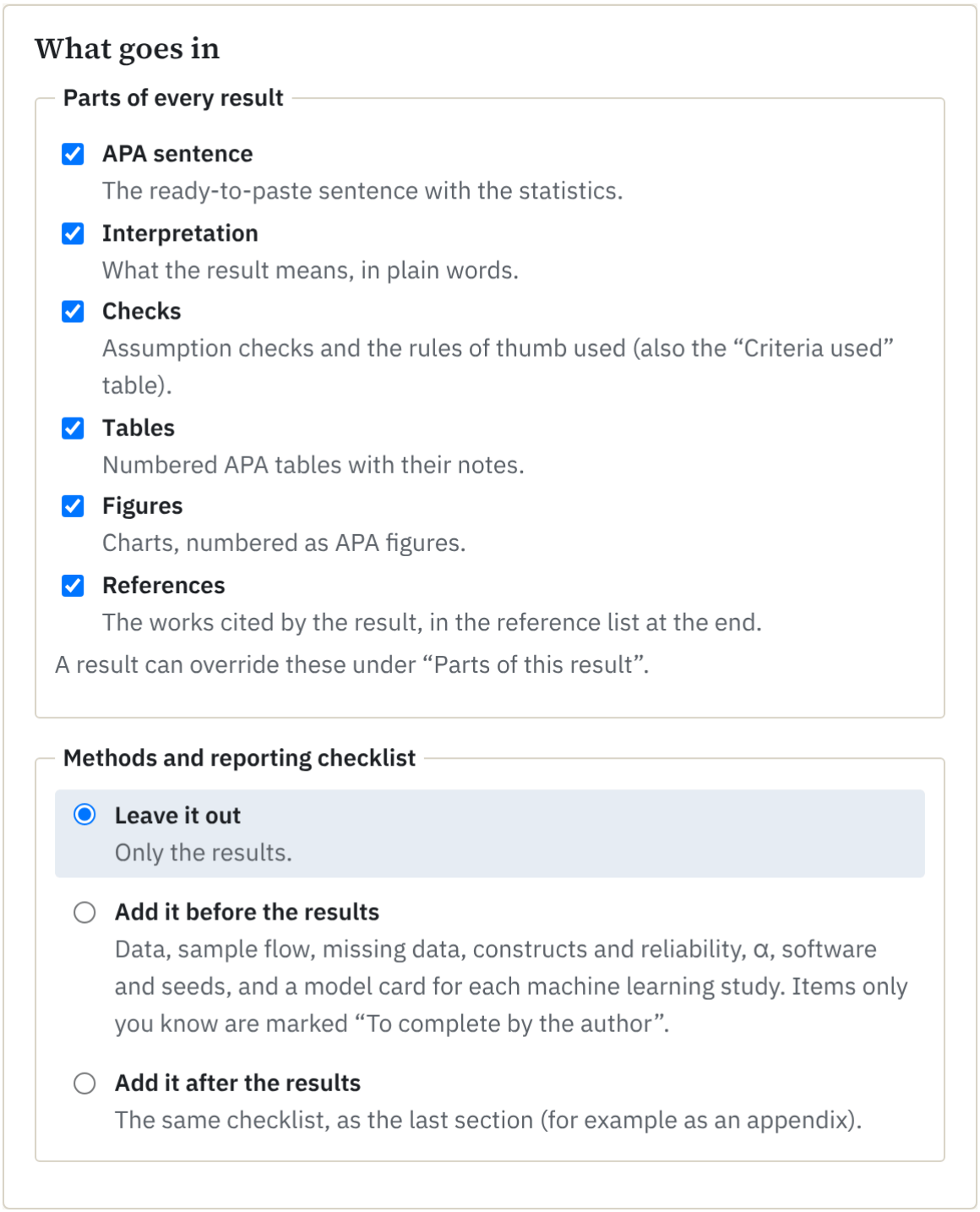


Figure 95. *What goes in: the parts of every result and the methods and reporting checklist.*

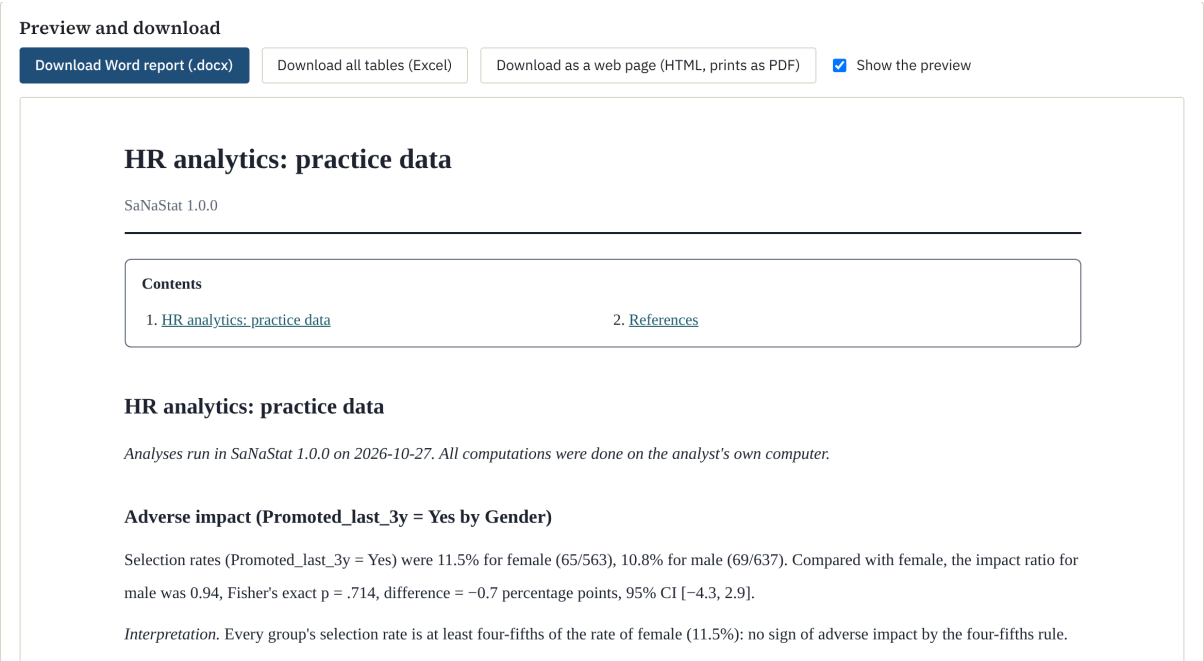


Figure 96. *Preview and download, with the start of the preview.*

The document contains the title, a line stating the SaNaStat version and date and that all computations were done on the analyst's own computer, then each result with its heading, APA sentence, interpretation, checks, numbered tables and figures, then a table of the *Criteria used* (each rule of thumb applied, with its source) and the *References*.

11.3 Using the report well

- **Edit it.** The document is a draft. Replace the variable names by plain words, delete what is not needed, and add the context.
- **Check each sentence against the table.** The APA sentence is generated from the numbers; confirm that it says what you mean. For instance, the t-test result reads "male scored higher than female on Annual_salary (difference 1939.20 points)", while the reported d = −0.13 in the APA sentence has the sign of the group order (female minus male): read the two together.
- **Report the non-significant results too.** For example, the absence by department test (p = .403) belongs in the report as a null result.
- **Say what the data are.** State the source, the date, the number of employees, who was excluded and how missing values were handled.
- **Share with care.** A report on groups of employees can still identify people in small groups. Use the privacy setting and merge or leave out small groups before sharing.
- **Keep the data and the report together.** Download the report and export the project at the end of the session: SaNaStat keeps the project in the browser only until the site data are cleared.

11.4 A checklist for an HR analysis

1. Is the question clear, is there a legal basis and a purpose, and are the data good enough for it (Responsible people analytics, chapter 1)?
2. Have you looked at the data (counts, missing values, outliers) before any test (chapter 2)?
3. Are the scales reliable and valid before you use their scores (chapter 3)?
4. Does the test fit the design and the kind of outcome: independent groups, repeated measures, categories, counts, time to an event (chapters 3 to 8)?
5. Have you looked at the effect size and its interval, not only the p-value?
6. For pay and selection: have you reported the raw and the adjusted gap, the impact ratios with their tests, and what the controls can and cannot explain (chapter 4)?
7. For predictions: was the model tested on unseen employees, is it calibrated, does it work equally well across groups, and who is harmed if it errs (chapter 9)?
8. Are small groups protected, and is the result written in plain words, with its limits, for people who can act on it?

Exercise

1. Add the results of three analyses of this manual to the Word report, give the report a title, add the methods and reporting checklist after the results, and download it. Open the file and find the table of criteria used.

1. Copy the hazard-ratio table of the Cox regression of chapter 8 and paste it into a word processor. Does it arrive as a table?

1. Pick one APA sentence and rewrite it for a non-statistician in two sentences.

Appendix A. Codebooks

Both files are simulated; no real person or company appears in them. Missing values are empty cells. The planted effects are listed so that you can check your results.

A.1 practice_hr.csv (1,200 employees)

Table 27
Variables of practice_hr.csv

Variable	Meaning
EmployeeID	Identifier (type ID; never analysed)
Department	Sales, Operations, Finance, IT, HR
Job_level	1 to 5
Gender	female, male
Age	Years
Education	Diploma, Bachelor, Master, Doctorate
Tenure_years	Years with the company (used as the time to leaving in chapter 8)
Annual_salary	Skewed; after controls, women earn about 4% less (planted)
Performance_rating	1 to 5
Training_hours	Hours of training
Overtime_hours_month	Hours of overtime per month
Commute_km	Commuting distance (35 missing)
Remote_days_week	Days worked remotely per week
Manager_support	Survey score, 1 to 5 (40 missing)
Work_life_balance	Survey score, 1 to 5
Engagement	Survey score, 1 to 5
Job_satisfaction	Survey score, 1 to 5 (52 missing)
Absence_days	Count of days absent (0 to 11)
Promoted_last_3y	Yes/No (about 11%)
Skills_test_before, Skills_test_after	Skills test around a training programme (paired)
Engagement_Q1 to Engagement_Q4	Engagement in four quarters (repeated measures)
Left_company	Yes/No, left within the year (about 15%)

Added by you in section 1.9 (Data screen, card 4): **Annual_salary_ln**, the natural logarithm of Annual_salary; **Tenure_group**, tertiles of Tenure_years; **Left_yes**, a 0/1 indicator of

Left_company = Yes; **Skills_gain**, Skills_test_after − Skills_test_before; **Dept_merged**, Department with Finance and HR merged into Corporate. Chapter 10 adds **Employee_type**.

What is planted: pay gap after controls (chapter 4); promotion driven by performance, tenure and training (chapter 4); gain on the skills test and a dose–response with Training_hours (chapter 5); engagement driven by Manager_support and Work_life_balance (chapter 6); absence days as counts that fall with engagement and rise with commuting distance (chapter 7); leaving linked to engagement, overtime, commuting and tenure (chapters 8 and 9); about 15% of employees leave. The file has no dates and no applicant records.

A.2 practice_engagement.csv (600 respondents)

Table 28
Variables of practice_engagement.csv

Variable	Meaning
RespondentID	Identifier (type ID)
Department	Sales, Operations, Finance, IT, HR
Gender	female, male
Tenure_band	<2 years, 2-5 years, 6-10 years, >10 years
Works_remotely	Yes/No
Vigor1 to Vigor3	Energy at work, rated 1 to 7 (Vigor2: 20 missing)
Dedication1 to Dedication3	Involvement and pride, rated 1 to 7
Absorption1 to Absorption3	Immersion in work, rated 1 to 7 (Absorption3: 21 missing)

What is planted: three correlated scales, and differences between departments with Operations lowest (chapter 3).

Appendix B. Which tool answers which HR question?

Table 29

HR questions and the SaNaStat 1.0.0 tools that answer them (or do not)

HR question	Tool in SaNaStat 1.0.0	Chapter
What does the workforce look like?	Analyses → Descriptive statistics; Charts	2
How do I protect small groups in tables?	Hide results for groups smaller than (privacy setting)	Responsible people analytics
Which variable would be better as a logarithm, as groups or as an indicator?	Data → Compute and recode variables	1, 2
How do I analyse one department only?	Data → Select cases	1
Is a survey scale reliable, and does it measure what it should?	Data → Build construct scores (α, ω); Analyses → Factor analysis and reliability	3
Were survey answers careless?	Data → Check data quality	3
How do I handle missing survey answers?	Analyses → Missing data: multiple imputation	1
Does engagement differ between departments or groups?	One-way ANOVA with Kruskal–Wallis; t-test with Mann–Whitney	3, 7
Do women and men earn the same?	t-test; regression of log salary with gender; Adverse impact and pay gaps → Pay gap (adjusted gap, Oaxaca–Blinder)	4
Are women and men in different jobs?	Chi-square test of Gender by Department or Job_level	4
Are some groups selected, hired or promoted less often?	Adverse impact and pay gaps → Adverse impact (selection rates, four-fifths rule, Fisher's exact test)	4
Did training change test scores?	Paired t-test with Wilcoxon test	5
Does more training go with more gain?	Partial correlation; regression with the baseline score	5
How many people do I need to evaluate a programme?	Power analysis	5
What goes with engagement or satisfaction?	Correlation; multiple and hierarchical regression	6
Does a factor act through another, or depend on another?	Analyses → Mediation and moderation	6

HR question	Tool in SaNaStat 1.0.0	Chapter
Does a score change across several occasions?	Repeated-measures ANOVA with Friedman test	6
Is absence or leaving different between departments?	ANOVA / Kruskal–Wallis; chi-square test	7
What predicts the number of absence days?	Correlation and regression → Count regression (Poisson or negative binomial)	7
How long do employees stay? Do groups differ?	Survival analysis → Kaplan–Meier curves, log-rank test	8
Which factors make leaving come sooner?	Survival analysis → Cox regression with the proportional-hazards check	8
Which employees are at risk of leaving?	Logistic regression; Machine learning	9
What drives the prediction?	Machine learning → SHAP, accumulated local effects	9
Are the predicted probabilities trustworthy?	Machine learning → Calibration	9
Choose a cut-off for a rare outcome	Machine learning → Cut-offs, precision and recall	9
Does a model work equally well for women and men?	Machine learning → Fairness by group	9
Are there types of employee?	Clustering	10
Different reasons for leaving (competing risks); predictors that change over time	Not available	–
Employees nested in teams (multilevel models)	Not available	–
Rebalance a rare outcome (weights, resampling)	Not available (read AUC, precision and recall)	–

Appendix C. Glossary

Table 30

Terms used in this manual

Term	Meaning
Adjusted gap	A difference between groups that remains after other measured variables are held constant in a regression.
Adverse impact	A substantially lower selection rate (hiring, promotion) for one group than for another.
Alpha (Cronbach's)	A measure of how well the items of a scale hang together, from 0 to 1.
ANOVA	Analysis of variance: a test of whether three or more group means differ.
AUC	Area under the ROC curve: the probability that a randomly chosen leaver gets a higher predicted risk than a randomly chosen stayer. .50 is chance, 1.00 perfect.
Balanced accuracy	The average of sensitivity and specificity; it is not inflated by a large majority group.
Calibration	Whether predicted probabilities match the observed frequencies.
Censored	A person whose event has not (yet) happened at the end of observation; counted as at risk up to that time.
Confidence interval	A range of values compatible with the data at a stated level (usually 95%).
Count regression	Poisson or negative binomial regression for outcomes that count events (0, 1, 2 …); the results are rate ratios.
Cox regression	A model of the hazard of an event that gives a hazard ratio for each predictor, without assuming a shape for the survival curve.
Cross-validation	Testing a model on people it was not trained on, by splitting the data into folds.
d, d_z	Standardised mean differences: 0.2 small, 0.5 medium, 0.8 large (Cohen, 1988).
Dummy coding	Turning a categorical predictor into 0/1 comparisons with a reference group.
Effect size	How large an effect is, as opposed to whether it is statistically significant.
Four-fifths rule	A rule of thumb that regards a selection rate below 80% of the highest group's rate as evidence of adverse impact.
Hazard ratio	The factor by which the risk of the event at any time changes for one unit more of a predictor.
Hierarchical regression	A regression entered in blocks, to see what each block adds (ΔR^2).
Impact ratio	A group's selection rate divided by the rate of the group selected most often.
Incidence-rate ratio (IRR)	The factor by which an expected count changes for one unit more of a predictor.

Term	Meaning
Kaplan–Meier curve	An estimate of the share of people still without the event at each time, allowing for censoring.
Kruskal–Wallis test	A rank-based test for three or more groups.
Leakage	Information in a predictor that would not be available when the prediction is made, or that contains the outcome.
Log-rank test	A test of whether the survival curves of two or more groups differ.
Mann–Whitney U test	A rank-based test for two independent groups.
Multiple imputation	Filling in missing values several times from the other variables and pooling the results, so that the uncertainty about them is kept.
Negative binomial regression	Count regression that allows the counts to vary more than their mean (overdispersion).
Oaxaca–Blinder decomposition	A split of an average pay gap into a part explained by differences in characteristics and an unexplained part.
Odds ratio	The factor by which the odds of the outcome change for one unit more of a predictor.
Omega (McDonald's)	A reliability coefficient from a one-factor model of the items; it does not assume that all items measure equally well.
Overdispersion	Counts that vary more than a Poisson model allows; it makes Poisson standard errors too small.
p-value	The probability of a result at least as extreme as the one observed if there were no effect. It is not the probability that the hypothesis is true.
Partial correlation	A correlation after removing the effect of one or more other variables.
Precision (positive predictive value)	Of those predicted to leave, the share who did leave.
Proportional hazards	The assumption of Cox regression that a predictor's effect on the hazard is the same at every time.
R^2	The share of the variance of the outcome explained by the model.
Reference group	The comparison group for a categorical predictor (by default the first category alphabetically).
Sensitivity (recall)	Of those who left, the share the model flagged.
SHAP	A method that shares each prediction among the predictors; mean absolute SHAP shows how much each predictor moves the predictions.
Specificity	Of those who stayed, the share the model did not flag.
Wilcoxon signed-rank test	A rank-based test for paired measurements.

References

Apley, D. W., & Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 82(4), 1059–1086. https://doi.org/10.1111/rssb.12377

Barnett, A. G., van der Pols, J. C., & Dobson, A. J. (2005). Regression to the mean: What it is and how to deal with it. International Journal of Epidemiology, 34(1), 215–220. https://doi.org/10.1093/ije/dyh299

Blinder, A. S. (1973). Wage discrimination: Reduced form and structural estimates. Journal of Human Resources, 8(4), 436–455. https://doi.org/10.2307/144855

Breusch, T. S., & Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. Econometrica, 47(5), 1287–1294. https://doi.org/10.2307/1911963

Cameron, A. C., & Trivedi, P. K. (1990). Regression-based tests for overdispersion in the Poisson model. Journal of Econometrics, 46(3), 347–364. https://doi.org/10.1016/0304-4076(90)90014-K

Cameron, A. C., & Trivedi, P. K. (2013). Regression analysis of count data (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781139013567

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 5(2), 153–163. https://doi.org/10.1089/big.2016.0047

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum.

Cox, D. R. (1972). Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological), 34(2), 187–220. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. Journal of the American Statistical Association, 72(359), 557–565. https://doi.org/10.1080/01621459.1977.10480613

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice. (1978). Uniform guidelines on employee selection procedures. Federal Register, 43(166), 38290–38315.

Games, P. A., & Howell, J. F. (1976). Pairwise multiple comparison procedures with unequal n's and/or variances: A Monte Carlo study. Journal of Educational Statistics, 1(2), 113–125. https://doi.org/10.3102/10769986001002113

Girden, E. R. (1992). ANOVA: Repeated measures. Sage.

Grambsch, P. M., & Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. Biometrika, 81(3), 515–526. https://doi.org/10.1093/biomet/81.3.515

Green, S. B. (1991). How many subjects does it take to do a regression analysis? Multivariate Behavioral Research, 26(3), 499–510. https://doi.org/10.1207/s15327906mbr2603_7

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). Multivariate data analysis (8th ed.). Cengage Learning.

Hayes, A. F. (2022). Introduction to mediation, moderation, and conditional process analysis: A regression-based approach (3rd ed.). Guilford Press.

Hennig, C. (2007). Cluster-wise assessment of cluster stability. Computational Statistics & Data Analysis, 52(1), 258–271. https://doi.org/10.1016/j.csda.2006.11.025

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (3rd ed.). Wiley.

Hubert, L., & Arabie, P. (1985). Comparing partitions. Journal of Classification, 2, 193–218. https://doi.org/10.1007/BF01908075

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. Journal of the American Statistical Association, 53(282), 457–481. https://doi.org/10.1080/01621459.1958.10501452

Kaufman, L., & Rousseeuw, P. J. (1990). Finding groups in data: An introduction to cluster analysis. Wiley.

Koenker, R. (1981). A note on studentizing a test for heteroscedasticity. Journal of Econometrics, 17(1), 107–112. https://doi.org/10.1016/0304-4076(81)90062-2

Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. Springer.

Levene, H. (1960). Robust tests for equality of variances. In I. Olkin (Ed.), Contributions to probability and statistics (pp. 278–292). Stanford University Press.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemotherapy Reports, 50(3), 163–170.

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. Psychological Methods, 23(3), 412–433. https://doi.org/10.1037/met0000144

Molnar, C. (2022). Interpretable machine learning: A guide for making black box models explainable (2nd ed.). Independently published.

Morris, S. B., & Lobsenz, R. E. (2000). Significance tests and confidence intervals for the adverse impact ratio. Personnel Psychology, 53(1), 89–111. https://doi.org/10.1111/j.1744-6570.2000.tb00195.x

Neumark, D. (1988). Employers' discriminatory behavior and the estimation of wage discrimination. Journal of Human Resources, 23(3), 279–295. https://doi.org/10.2307/145830

Newcombe, R. G. (1998). Interval estimation for the difference between independent proportions: Comparison of eleven methods. Statistics in Medicine, 17(8), 873–890. https://doi.org/10.1002/(SICI)1097-0258(19980430)17:8<873::AID-SIM779>3.0.CO;2-I

Nunnally, J. C. (1978). Psychometric theory (2nd ed.). McGraw-Hill.

Oaxaca, R. (1973). Male-female wage differentials in urban labor markets. International Economic Review, 14(3), 693–709. https://doi.org/10.2307/2525981

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453. https://doi.org/10.1126/science.aax2342

Peduzzi, P., Concato, J., Feinstein, A. R., & Holford, T. R. (1995). Importance of events per independent variable in proportional hazards regression analysis II. Accuracy and

precision of regression estimates. Journal of Clinical Epidemiology, 48(12), 1503–1510. https://doi.org/10.1016/0895-4356(95)00048-8

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. Journal of Clinical Epidemiology, 49(12), 1373–1379. https://doi.org/10.1016/S0895-4356(96)00236-3

Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. Wiley. https://doi.org/10.1002/9780470316696

Schaufeli, W. B., Salanova, M., González-Romá, V., & Bakker, A. B. (2002). The measurement of engagement and burnout: A two sample confirmatory factor analytic approach. Journal of Happiness Studies, 3(1), 71–92. https://doi.org/10.1023/A:1015630930326

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5), 557–570. https://doi.org/10.1142/S0218488502001648

Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: The Achilles heel of predictive analytics. BMC Medicine, 17, 230. https://doi.org/10.1186/s12916-019-1466-7