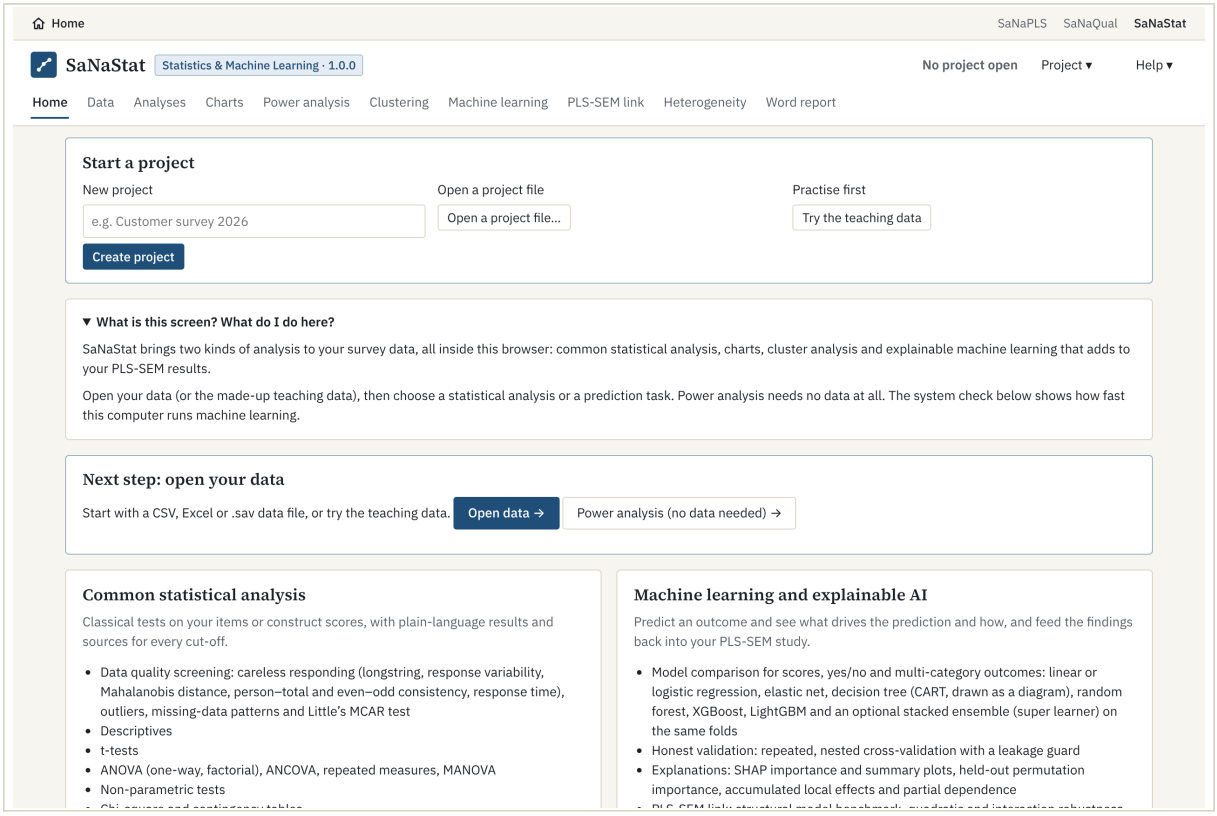
SANASOFT · USER MANUAL

SaNaStat

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

Version 1.0.0 · October 2026



Worked examples: 1,500 simulated retail customers, a 500-respondent brand survey, 360 shoppers for segmentation, three years of online-shop orders and a conjoint study of 150 respondents

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 marketing....6
What you need....8
The practice data....8
How to read the instructions....9
1. Getting started with marketing data....11
1.1 The screens and the list of analyses....11
1.2 Load the customer file....12
1.3 Check the variable types....13
1.4 The variable editor....15
1.5 Missing values....15
1.6 Making new variables....16
1.7 Select cases....17
2. Know your customers....19
2.1 Descriptive statistics....19
2.2 Histograms....22
2.3 Box plots: medians and spread by group....23
2.4 Group means with confidence intervals....23
2.5 Scatter plots and the correlation heat map....25
2.6 Category percentages and the problem of skew....27
2.7 Groups, indicators and merged categories....29
3. Survey scales: trust, value and loyalty....33
3.1 Data quality: straight-lining....33
3.2 Reliability and scale scores....36
3.3 Compare loyalty between groups....38
3.4 What drives loyalty: correlation and regression....41
3.5 Going further with scales....43
4. Testing a campaign....45
4.1 Did more customers respond? The A/B test of a rate....45
4.2 The same table as a chi-square test....48
4.3 Did monthly spend change? The paired test....50
4.4 Did the New offer raise spend more? The A/B test of a mean....51
4.5 Typed-in totals: an A/B test without a data file....53
4.6 How many customers does a test need?....53
5. What drives customer spend....57
5.1 Multiple linear regression....57
5.2 The same model on the logarithm....59
5.3 Visits and basket are parts of spend, not drivers....60
5.4 Compare spend between two groups....61
5.5 Compare spend between three or more groups....63

6. Predicting churn and response....66
6.1 Which variables may be used as predictors?....66
6.2 Logistic regression for churn....67
6.3 Machine learning: set up the prediction....69
6.4 Results · Model comparison....71
6.5 Results · Cut-off: precision and recall....73
6.6 Results · Explanations: what drives churn?....76
6.7 Results · Calibration: can the probabilities be trusted?....79
6.8 Results · Fairness by group....80
6.9 Results · Model card and predictions....81
6.10 Predicting response to the campaign....83
7. Customer segmentation....87
7.1 Segments with clear structure: segments_demo.csv....87
7.2 Segments in the customer file....92
7.3 Keep the clusters and compare them....94
8. Customer value: RFM segments and lifetime value....98
8.1 The data: one row per item bought....98
8.2 Run the RFM analysis....98
8.3 Reading the segments....100
8.4 Cohort retention and the simple customer lifetime value....101
8.5 Choices that change the result....103
9. Market basket analysis: which products are bought together?....105
9.1 Support, confidence and lift....105
9.2 Find the rules....105
9.3 Thresholds and longer rules....110
10. Conjoint analysis: which product features matter?....111
10.1 The data....111
10.2 Estimate the part-worths....111
10.3 The market simulator....114
11. Reporting, exporting and the report builder....117
11.1 Copying and exporting one table or chart....117
11.2 Collecting results in the Word report....117
11.3 A worked mini-report....120
11.4 A checklist before you send the report....121
Appendix A. Codebooks....123
A.1 practice_marketing.csv (1,500 customers)....123
A.2 practice_brand_survey.csv (500 respondents)....124
A.3 segments_demo.csv (360 shoppers)....124
A.4 practice_transactions.csv (21,006 order lines)....125
A.5 practice_conjoint.csv (150 respondents, 1,800 ratings)....125
Appendix B. Which tool answers which marketing question?....127
Appendix C. Glossary....129
Appendix D. Limits and doubtful points to keep in mind....131
References....133

About this manual

This manual shows how to answer common marketing questions with SaNaStat 1.0.0, the statistics and machine learning program of the SaNaSoft suite: who your customers are, how a brand survey scale behaves, whether a campaign worked, what drives spending, who is likely to leave, which kinds of customers you have, which customers are most valuable, which products are bought together and which product features matter most. Every chapter follows one practice file, tells you what to click, shows what you will see, explains how to read the result and gives a sentence for your report. All screenshots and numbers come from real runs of SaNaStat 1.0.0, 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

Marketing students and researchers who have customer, survey or sales data in a table (one row per customer, respondent, order line or rating, one column per variable) and want to analyse it carefully. You do not need earlier experience with any statistics software. Each statistical idea is explained in a "Concept" box when it first appears. Basic knowledge of means and percentages is enough to start. The Getting Started guide describes every screen in detail, and the Common Statistical Tests manual explains the general tests in more depth; here we use the parts that matter for marketing.

Table 1
What each chapter covers

Chapter	Topic	What you learn
1	Getting started with marketing data	Open the file, variable types, the variable editor, missing values, new variables, Select cases
2	Know your customers	Descriptive statistics, charts, skewed spend, groups and indicators
3	Survey scales	Data quality screening, reliability (α and ω), group comparisons, drivers of loyalty

Chapter	Topic	What you learn
4	Testing a campaign	The A/B test of a rate and of a mean, uplift with its interval, before-and-after spend, sample size
5	What drives customer spend	Multiple and hierarchical regression, group comparisons
6	Predicting churn and response	Logistic regression, machine learning on unseen customers, cut-offs, calibration, fairness, the model card
7	Customer segmentation	Cluster analysis, keeping and comparing segments
8	Customer value: RFM and lifetime value	Recency, frequency and money scores, segments, cohort retention, simple CLV
9	Market basket analysis	Association rules: support, confidence and lift
10	Conjoint analysis	Part-worths, attribute importance and the market simulator
11	Reporting and exporting	Export buttons, the report builder, a worked mini-report, a checklist

What SaNaStat can and cannot do for marketing

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. The list on the left of its **Analyses** screen has a group *Marketing analytics* with four tools: **A/B test** (did version B do better?), **RFM and customer value** (segments by recency, frequency and money, and a simple customer lifetime value), **Market basket** (which products are bought together?) and **Conjoint analysis** (which product features matter?). 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 chi-square tests; correlation and regression; power analysis; clustering; and machine learning with explanations, calibration, cut-offs and fairness by group.

Table 2

What the program does and does not offer for marketing analytics

Marketing task	Status in SaNaStat 1.0.0
Describing customers, charts, correlation heat map	Available (chapter 2)
Making new variables (logarithm, difference of two columns, median split, 0/1 indicator, merged categories, reversed items)	Available on the Data screen (sections 1.6 and 2.6 to 2.7)

Marketing task	Status in SaNaStat 1.0.0
Survey scales: data quality screening, reliability (Cronbach's α, McDonald's ω), scale scores	Available (chapter 3); factor analysis and validity in the Common Statistical Tests manual
A/B test of a conversion rate or a mean, with uplift, interval and sample size	Available (chapter 4), also from typed-in totals without a data file
Drivers of spend (linear regression), group comparisons	Available (chapter 5)
Predicting churn or response (logistic regression, machine learning, explanations, calibration, cut-offs, fairness, predictions for each customer)	Available (chapter 6)
Customer segmentation (k-means and hierarchical clustering)	Available (chapter 7)
RFM segments and a simple customer lifetime value	Available (chapter 8)
Association rules (market-basket analysis)	Available (chapter 9)
Ratings-based conjoint analysis with a market simulator	Available (chapter 10)
Mediation and moderation (for example, does satisfaction carry the effect of service quality on loyalty?)	Available; see the Common Statistical Tests manual, chapter 13
Multiple imputation of missing answers	Available; see the Common Statistical Tests manual, chapter 14
Choice-based conjoint (respondents choose one product from a set), hierarchical Bayes part-worths	Not available
Probability models of customer lifetime value (for example models of buying and dropping out for each customer)	Not available
Time-series forecasting of sales or demand	Not available
Uplift modelling (which customers are moved by an offer)	Not available
Marketing-mix models and attribution of sales to channels	Not available

Note: what this means for you

If your project needs one of the unavailable methods, SaNaStat can still help with the surrounding work: describing the data, checking its quality, testing differences between segments or campaign groups, and predicting an outcome. Use other tools for the method itself, and bring the resulting columns back into SaNaStat as ordinary variables.

What you need

- A computer with an up-to-date browser. SaNaStat runs inside the browser and your data stay on your computer; nothing is uploaded.
- Your data as a CSV, Excel (.xlsx) or .sav / .zsav file, with the variable names in the first row.
- The five practice files used here (see "The practice data" below). They ship with SaNaStat.

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

The practice data

Table 3

The practice files used in this manual

File	What it contains	Used in
practice_marketing.csv	1,500 customers of a retailer: profile, shopping behaviour, satisfaction, a campaign test, spend and churn (20 variables)	Chapters 1, 2, 4 to 7 and 11
practice_brand_survey.csv	500 respondents who rated Trust, Value and Loyalty items on a 1 to 7 scale, with age group, gender and customer type (15 variables)	Chapter 3
segments_demo.csv	360 shoppers rated on five attitude scales, with age, income, gender and channel	Chapter 7
practice_transactions.csv	21,006 order lines of an online grocery shop, 2023 to 2025: order number, customer number, date, item and amount (5 variables)	Chapters 8 and 9
practice_conjoint.csv	150 respondents who each rated the same 12 profiles of wireless earbuds (brand, price, battery life, noise cancelling) on a 1 to 10 scale (1,800 rows, 7 variables)	Chapter 10

All five files are simulated. No real customer, respondent, shop or brand appears in them; the brand names in the conjoint study are invented. Effects were planted on purpose, so

that you can check whether your analysis finds what was built in; chapters end with exercises and the planted answer. The codebooks are in appendix A. Because the data are simulated, do not read the results as facts about real customers.

How to read the instructions

Table 4
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.	**Analyses** → **A/B test**
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.	*straight-lining*
Variable names	As they appear in the file.	Annual_spend

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 for the marketing question.

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, with the planted answer where there is one.

1. Getting started with marketing data

1.1 The screens and the list of analyses

Open SaNaStat in your browser (Figure 1). The menu at the top has the screens **Home**, **Data**, **Analyses**, **Charts**, **Power analysis**, **Clustering**, **Machine learning** and **Word report** (in the SaNaSoft suite also **PLS-SEM link** and **Heterogeneity**). Every screen starts with a box *What is this screen? What do I do here?*, which explains the screen in two or three sentences, and a screen that needs data shows "No data yet" until you open a file.

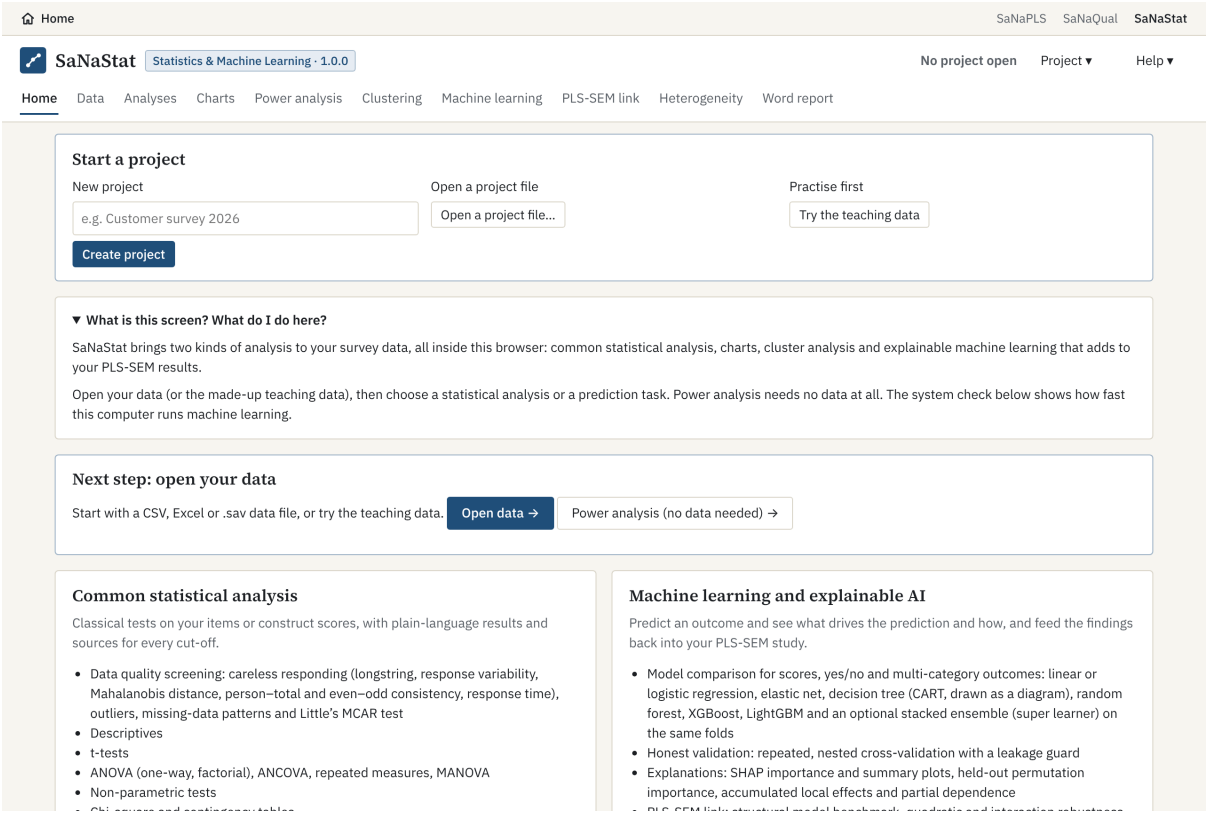


Figure 1. *The Home screen of SaNaStat.*

The list on the left of the **Analyses** screen is grouped (Figure 2): *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* and *Marketing analytics* (**A/B test**, **RFM and customer value**, **Market basket**, **Conjoint analysis**). 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)%.

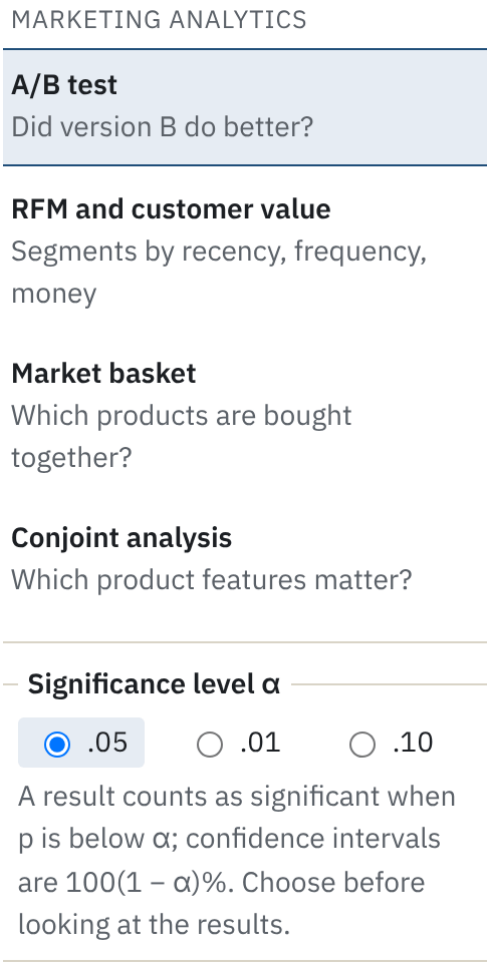


Figure 2. *The Marketing analytics group of the list of analyses, with the significance level α under it.*

SaNaStat keeps your data, results and report in a project in the browser and restores them on your next visit in the same browser. 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 Load the customer file

1. Click **Data** in the menu.
2. Click **Choose a file…** and pick *practice_marketing.csv*.
3. SaNaStat reports *Open practice_marketing.csv: 1500 respondents, 20 variables.* (Figure 3). SaNaStat calls every row a *respondent*; here a respondent is a customer. 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.

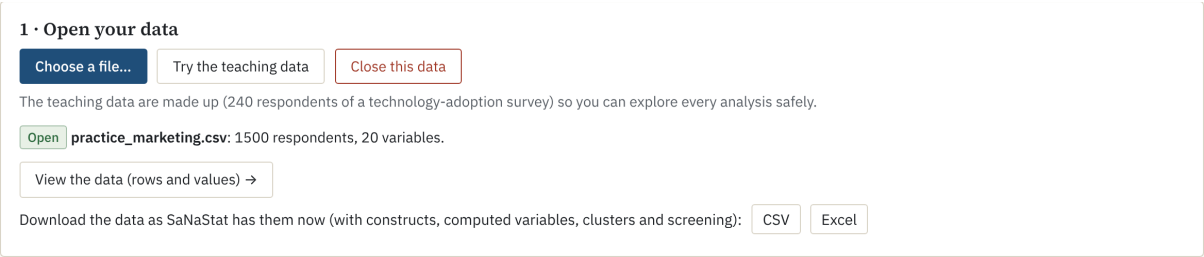


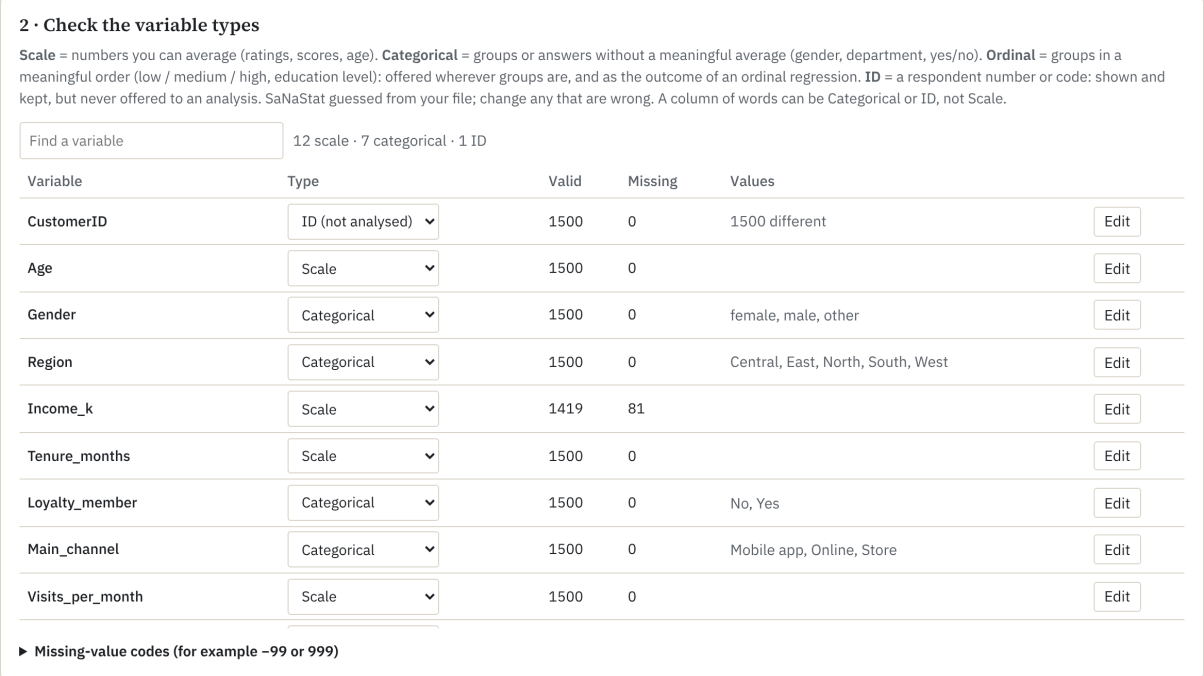
Figure 3. *Step 1 · Open your data, with practice_marketing.csv open.*

Each row is one customer and each column one variable. The file is read inside the browser and is never uploaded. **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.

1.3 Check the variable types

Card *2 · Check the variable types* lists every variable with its type, the number of valid and missing values, and, for categorical variables, the categories (Figure 4). SaNaStat guesses the type from the file and counts the result above the list: *12 scale · 7 categorical · 1 ID*. There are four types, and the type decides which analyses offer the variable:

- **Scale:** numbers that can be averaged, such as Age, Tenure_months, Annual_spend or a 1 to 10 rating.
- **Categorical:** groups or answers without a meaningful average, such as Region, Main_channel or Churned.
- **Ordinal:** groups in a meaningful order (low / medium / high, a rating treated as ordered categories). Offered wherever groups are, and as the outcome of an ordinal regression.
- **ID:** a customer number or code. It is shown and kept, but never offered to an analysis.



2 · Check the variable types

Scale = numbers you can average (ratings, scores, age). **Categorical** = groups or answers without a meaningful average (gender, department, yes/no). **Ordinal** = groups in a meaningful order (low / medium / high, education level): offered wherever groups are, and as the outcome of an ordinal regression. **ID** = a respondent number or code: shown and kept, but never offered to an analysis. SaNaStat guessed from your file; change any that are wrong. A column of words can be Categorical or ID, not Scale.

Find a variable — 12 scale · 7 categorical · 1 ID

Variable	Type	Valid	Missing	Values	
CustomerID	ID (not analysed)	1500	0	1500 different	Edit
Age	Scale	1500	0		Edit
Gender	Categorical	1500	0	female, male, other	Edit
Region	Categorical	1500	0	Central, East, North, South, West	Edit
Income_k	Scale	1419	81		Edit
Tenure_months	Scale	1500	0		Edit
Loyalty_member	Categorical	1500	0	No, Yes	Edit
Main_channel	Categorical	1500	0	Mobile app, Online, Store	Edit
Visits_per_month	Scale	1500	0		Edit

▸ Missing-value codes (for example −99 or 999)

Figure 4. *Card 2 · Check the variable types for practice_marketing.csv (the first rows).*

What it means for our example

SaNaStat guessed correctly. **CustomerID** was given the type ID ("1500 different" values): it will never be offered as a predictor, a grouping variable or a variable to describe, so it cannot leak into a model by mistake. Gender (female, male, other), Region (Central, East, North, South, West), Loyalty_member (No, Yes), Main_channel (Mobile app, Online, Store), Campaign_group (New offer, Standard), Responded (No, Yes) and Churned (No, Yes) are categorical. Age, Income_k, Tenure_months, Visits_per_month, Avg_basket, Discount_share, Complaints_12m, the two ratings, Annual_spend and the two monthly spend variables are scale variables.

Change any type that is wrong with the list in the **Type** column. A common marketing case is a rating such as Recommend_0to10: treat it as scale when you want means and regression, or as categorical or ordinal when you want to compare the share of customers who gave each rating.

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 Central is the reference group for Region, Mobile app for Main_channel and New offer for Campaign_group. The regression screens let you choose another reference group under **Reference groups**; this manual keeps the first one. Remember it when you read coefficients.

1.4 The variable editor

Edit at the end of each row opens the variable editor (Figure 5). 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 Recommend_0to10: SaNaStat reports *Recommend_0to10 is now ordinal.* and lists the eleven answers 0 to 10 with their counts, lowest first, from 2 customers who answered 0 to 116 who answered 10. Click **Close**, and set Recommend_0to10 back to **Scale** in the list for now.

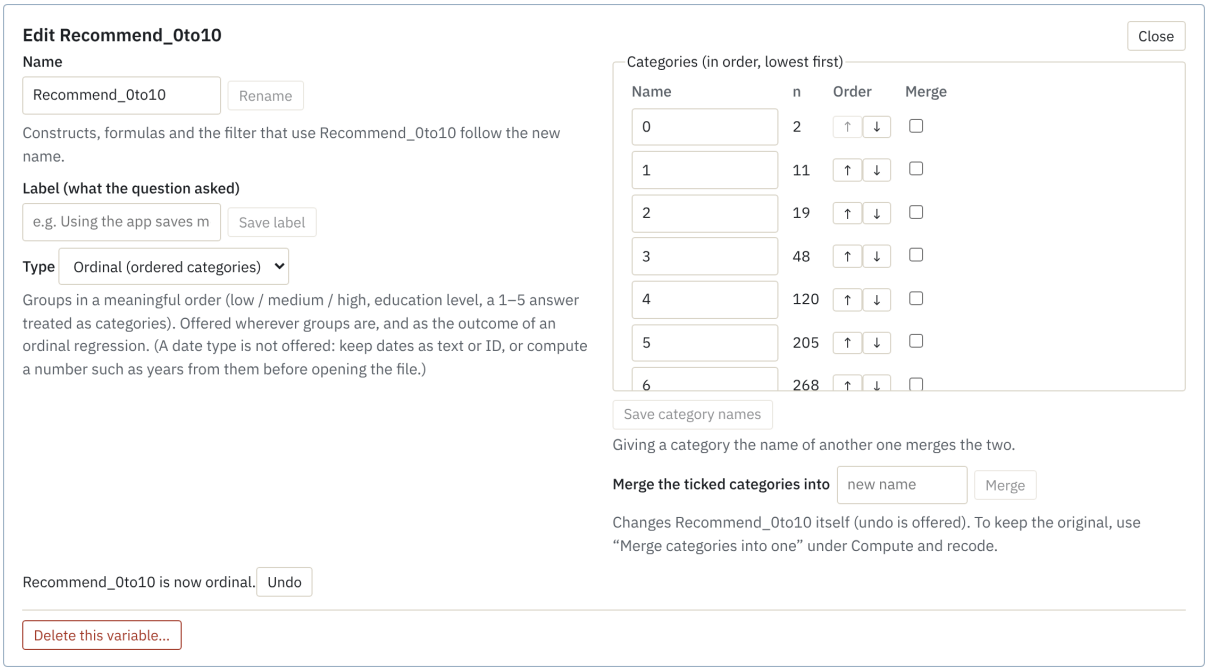


Figure 5. *The variable editor for Recommend_0to10 with the type Ordinal: eleven answers, lowest first.*

1.5 Missing values

The columns Valid and Missing in Figure 4 show the missing data. In practice_marketing.csv three variables have missing values: Income_k (81 missing, so 1419 valid), Satisfaction_1to10 (85 missing, 1415 valid) and Recommend_0to10 (65 missing, 1435 valid). All other variables 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_marketing.csv: SaNaStat reports *Found in 1 variable: Monthly_spend_after (1).* That 99 is a real monthly spend, 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.

Most analyses leave out a customer only for the variables the analysis uses (a correlation uses every pair with both values; a regression uses customers complete on all of its variables). The result always states how many customers were used, for example "n =

1341 (159 respondents with missing values left out)" in the regression of section 5.1. Machine learning fills in missing values using the training customers only (section 6.3).

Note: missing values are information

Customers who skipped a satisfaction question may differ from those who answered it. SaNaStat's data quality screen (shown for the survey in chapter 3) reports whether the missing answers look random (Little's test). Always report how many customers were left out of each analysis.

Tip: multiple imputation

When more than a few per cent of the values are missing, or when the customers 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_marketing.csv the variable with the most missing values is Satisfaction_1to10 (85 of 1,500), so this manual analyses the complete cases. The Common Statistical Tests manual (chapter 14) shows the method step by step, with its diagnostics.

1.6 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.

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, for example *Added Spend_change: 1500 of 1500 respondents have a value.* A new name cannot repeat the name of another variable. This manual makes eight new variables from practice_marketing.csv; sections 2.6, 2.7 and 4.4 show why.

Table 5

The new variables made in this manual

New variable	How it is made	Section
Annual_spend_ln1p	**Transform a scale variable**, Annual_spend, **ln(1 + x)**	2.6
Annual_spend_ln	**Transform a scale variable**, Annual_spend, **Natural logarithm, ln(x)**	2.6
Satisfaction_1to10_group	**Make groups from a scale variable**, Satisfaction_1to10, **Two groups at the median**	2.7
Income_tertile	**Make groups from a scale variable**, Income_k, **Three equal-sized groups (tertiles)**	2.7
Churned_Yes	**Make a 0/1 indicator of a category**, Churned, Yes ticked	2.7
Region_merged	**Merge categories into one**, Region, East and West ticked, *East_West*	2.7
Spend_change	**Combine two scale variables**, Monthly_spend_after **Difference, A − B** Monthly_spend_before	4.4
Spend_change_pct	**Combine two scale variables**, the same two variables, **Percentage change, (A − B) ÷ B × 100**	4.4

After all eight are added, the count above the variable list reads *17 scale · 10 categorical · 1 ID*, and every new variable carries the label *computed* and a short description of how it was made, for example "1 if Churned is Yes".

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

1.7 Select cases

Sometimes an analysis should use only some customers, for example only the members of the loyalty programme. Card *7 · Select cases (optional)* does this without deleting anyone (Figure 6):

1. Under **Analyse only respondents where**, choose **Loyalty_member** in **Variable**, **is** in **Comparison** and **Yes** in **Value**. SaNaStat counts at once: *Loyalty_member = Yes: 646 of 1500 meet this.*
2. Click **Use this filter**.

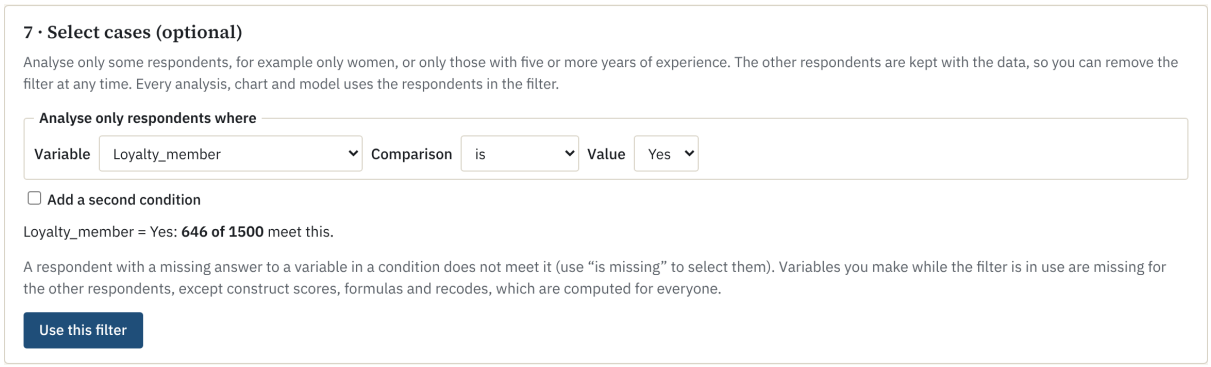


Figure 6. *Select cases: only the loyalty members.*

While the filter is in use, every analysis, chart and model uses only the selected customers, and a banner at the top of the screens says so (Figure 7): *Filter in use: Loyalty_member = Yes (646 of 1500 respondents)*, with **Remove filter** and **Change it**. The descriptive statistics of Annual_spend are then those of the members, M = 4389.38, SD = 2591.22, n = 646; after **Remove filter** they are those of everyone again, M = 3391.70, SD = 2383.64, n = 1500. Remove the filter before you continue.

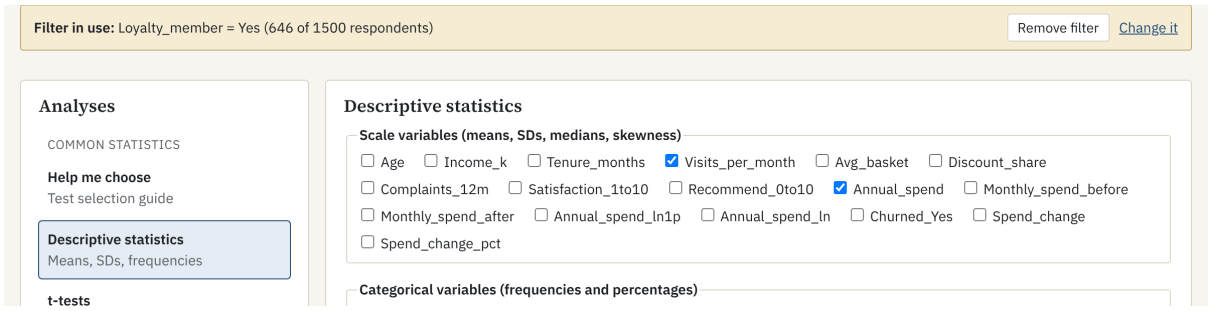


Figure 7. *The banner shows that only the loyalty members are analysed.*

Exercise

Exercise: the file

Open practice_marketing.csv and find: (a) how many variables are categorical; (b) which three variables have missing values and how many; (c) which variable SaNaStat will never offer to an analysis, and why.

Answer: (a) seven (Gender, Region, Loyalty_member, Main_channel, Campaign_group, Responded, Churned); (b) Income_k 81, Satisfaction_1to10 85, Recommend_0to10 65; (c) CustomerID, which has the type ID.

2. Know your customers

Before testing anything, look at the data. This chapter uses the **Analyses** screen (descriptive statistics) and the **Charts** screen.

2.1 Descriptive statistics

1. Click **Analyses** in the menu, then **Descriptive statistics** in the list on the left.
2. Under *Scale variables*, tick Age, Income_k, Tenure_months, Visits_per_month, Avg_basket, Discount_share and Annual_spend (Figure 8).
3. Read the table (Figure 9).

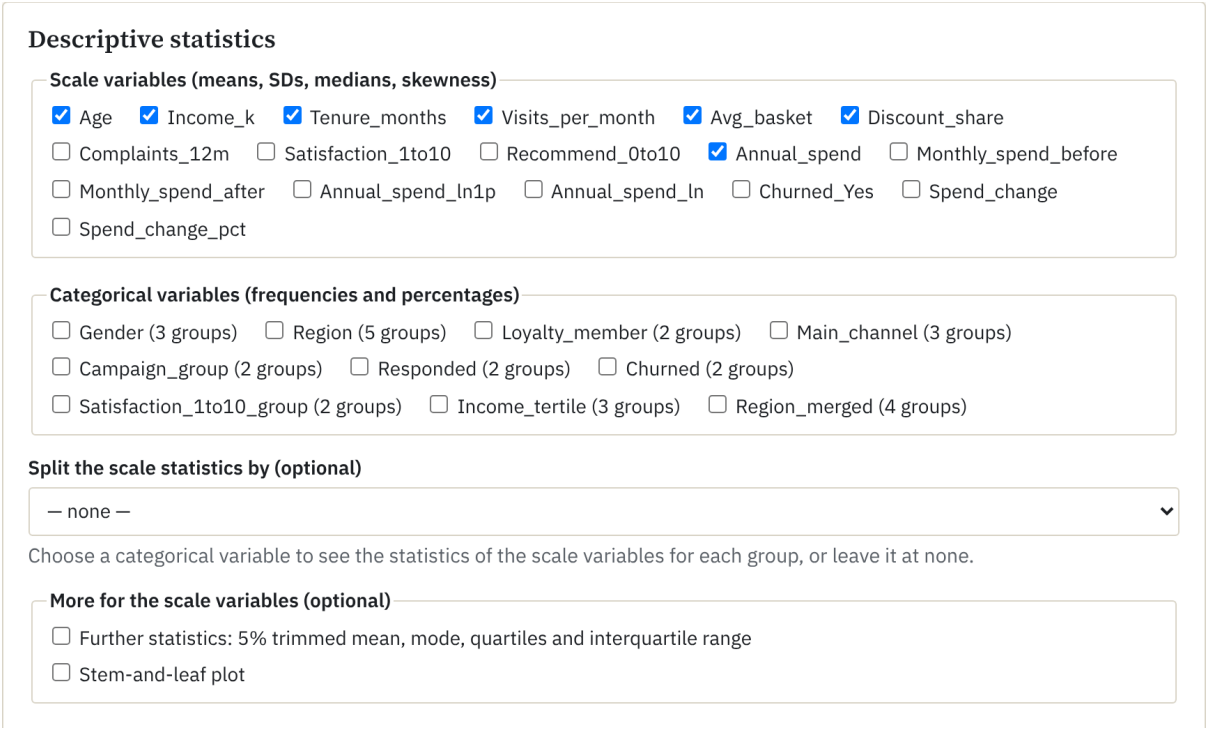


Figure 8. *Descriptive statistics: the variables to describe and the optional split.*

What it means: 3 of the 7 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	1500	0	41.71	12.75	41.00	18.00	80.00	0.22 (0.06)	−0.27 (0.13)
Income_k	1419	81	54.47	25.32	49.00	13.10	190.80	1.38 (0.06)	2.75 (0.13)
Tenure_months	1500	0	30.86	21.36	26.00	1.00	176.00	1.49 (0.06)	3.61 (0.13)
Visits_per_month	1500	0	3.24	1.96	3.00	0.00	10.00	0.59 (0.06)	0.13 (0.13)
Avg_basket	1500	0	93.44	23.99	90.30	39.73	207.58	0.77 (0.06)	1.04 (0.13)
Discount_share	1500	0	0.26	0.14	0.24	0.00	0.80	0.71 (0.06)	0.12 (0.13)
Annual_spend	1500	0	3391.70	2383.64	2861.50	0.00	14085.00	1.07 (0.06)	1.29 (0.13)

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

Figure 9. *Descriptive statistics of seven customer variables.*

Table 6

Descriptive statistics of the customers (n = 1500 unless stated)

Variable	Mean	SD	Median	Min	Max	Skewness
Age	41.71	12.75	41.00	18.00	80.00	0.22
Income_k (n = 1419)	54.47	25.32	49.00	13.10	190.80	1.38
Tenure_months	30.86	21.36	26.00	1.00	176.00	1.49
Visits_per_month	3.24	1.96	3.00	0.00	10.00	0.59
Avg_basket	93.44	23.99	90.30	39.73	207.58	0.77
Discount_share	0.26	0.14	0.24	0.00	0.80	0.71
Annual_spend	3391.70	2383.64	2861.50	0.00	14085.00	1.07

SaNaStat summarises the result in its *What it means* line: "3 of the 7 rows show clear skewness (|skewness| > 1): report medians as well, and read the rank-based tests next to the tests of means." The table also gives the kurtosis and the standard errors of skewness and kurtosis.

Concept: mean, median and skewness

The *mean* is the arithmetic average; the *median* is the middle value. For a symmetric variable they are close. A long tail of large values (*positive skew*) pulls the mean above the median. In the table, the mean of Annual_spend (3391.70) is well above the median (2861.50), and its skewness is 1.07. SaNaStat flags a variable as clearly skewed when the absolute skewness exceeds 1. For skewed money variables, report the median next to the mean.

Choose a categorical variable in **Split the scale statistics by (optional)** to describe each group. Splitting by Loyalty_member (Figure 10) shows what loyalty members do differently.

What it means: 5 of the 14 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	Loyalty_member	n	Missing	Mean	SD	Median	Min	Max	Skewness	Kurtosis
Age	No	854	0	41.64	12.88	41.00	18.00	78.00	0.25 (0.08)	−0.23 (0.17)
Age	Yes	646	0	41.80	12.59	42.00	18.00	80.00	0.18 (0.10)	−0.32 (0.19)
Income_k	No	811	0	54.89	25.49	50.50	13.10	190.80	1.38 (0.09)	2.90 (0.17)
Income_k	Yes	608	0	53.92	25.11	47.90	15.70	182.70	1.38 (0.10)	2.56 (0.20)
Tenure_months	No	854	0	28.50	19.84	23.00	2.00	153.00	1.60 (0.08)	4.20 (0.17)
Tenure_months	Yes	646	0	33.99	22.86	29.00	1.00	176.00	1.34 (0.10)	3.00 (0.19)
Visits_per_month	No	854	0	2.65	1.67	2.00	0.00	8.00	0.60 (0.08)	0.19 (0.17)
Visits_per_month	Yes	646	0	4.01	2.05	4.00	0.00	10.00	0.37 (0.10)	−0.15 (0.19)
Avg_basket	No	854	0	88.77	24.01	86.00	39.73	207.58	0.95 (0.08)	1.58 (0.17)
Avg_basket	Yes	646	0	99.61	22.53	96.15	48.28	204.60	0.73 (0.10)	0.89 (0.19)
Discount_share	No	854	0	0.26	0.14	0.24	0.00	0.80	0.69 (0.08)	0.01 (0.17)
Discount_share	Yes	646	0	0.26	0.15	0.24	0.00	0.76	0.74 (0.10)	0.24 (0.19)
Annual_spend	No	854	0	2637.02	1893.02	2330.50	0.00	12460.00	1.12 (0.08)	1.77 (0.17)
Annual_spend	Yes	646	0	4389.38	2591.22	3949.00	0.00	14085.00	0.79 (0.10)	0.52 (0.19)

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

Figure 10. *Descriptive statistics split by Loyalty_member.*

What it means for our example

Loyalty members visit more often (mean 4.01 visits per month against 2.65 for non-members), fill larger baskets (99.61 against 88.77) and spend more per year (mean 4389.38 against 2637.02). Their age and income are almost the same (mean age 41.80 and 41.64). Note that this is a description of two groups, not proof that the loyalty programme *causes* the extra spending: customers who spend more may be the ones who join.

How to report it

"Customers spent a mean of 3391.70 per year (SD = 2383.64, Mdn = 2861.50, n = 1500); the distribution was right-skewed (skewness = 1.07)."

(The sentence SaNaStat itself writes lists M, SD and n for each variable, and for each group when you split the table; add the median and the skewness from the table when a variable is skewed.) Ticking variables under *Categorical variables* gives their frequencies and percentages in the same result.

Tip: hide small groups

The setting **Hide results for groups smaller than** under the split protects people in small groups: their numbers show as "< k" and those groups are left out of the tests. It applies to the whole project and starts at 5 people. No group in this manual is that small, but a split of customers by region and channel in a small shop can produce cells of two or three people.

2.2 Histograms

A histogram shows how the values are spread. Click **Charts**, choose **Histogram (How are the values spread?)** as the **Chart type**, and choose **Annual_spend** as the **Variable** (Figure 11, Figure 12).



Figure 11. *1 · Choose the chart: a histogram of Annual_spend.*

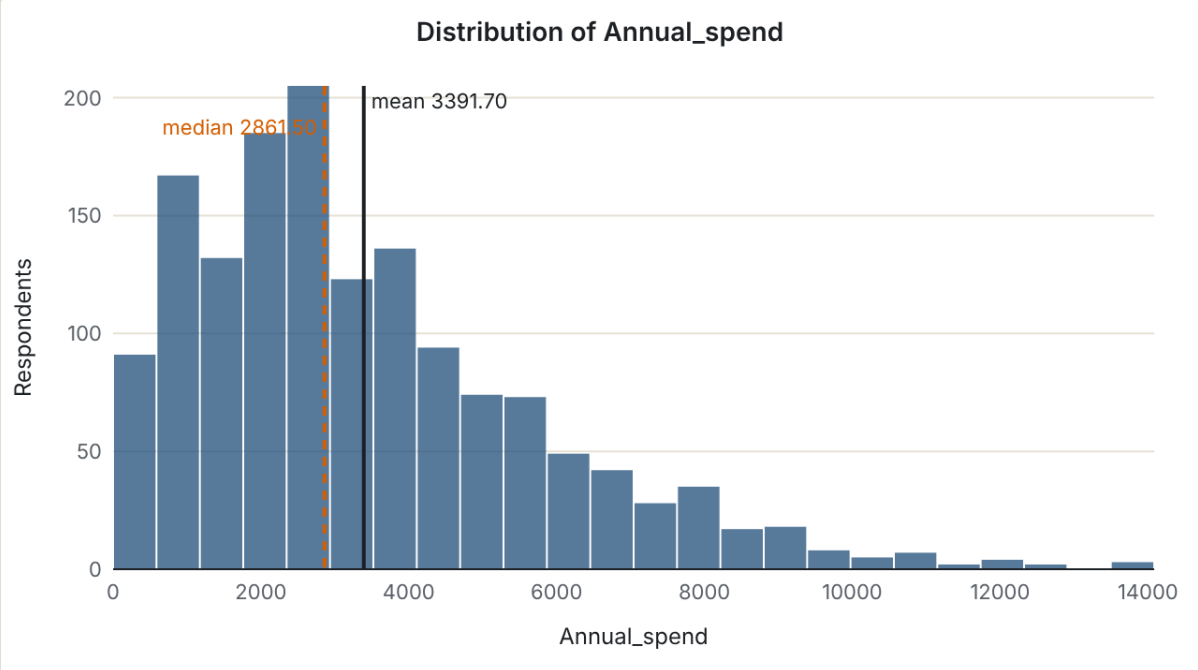


Figure 12. *Histogram of Annual_spend, with the mean and median marked.*

SaNaStat states the finding: "Annual_spend has 1500 values with mean 3391.70, median 2861.50 and SD 2383.64. The distribution is skewed to the right (a long tail of high values)." The histogram of Income_k has the same shape (skewness 1.38).

Every chart has a short caption, a *How to report it* sentence, buttons to copy or save it (chapter 11) and an **Add to Word report** button.

2.3 Box plots: medians and spread by group

Choose **Box plot (Medians and spread, by group)**, **Annual_spend** as the variable and **Loyalty_member** as the groups (Figure 13). The box covers the middle half of the customers, the line is the median, the whiskers reach the values within 1.5 times the interquartile range and the circles are unusual values.

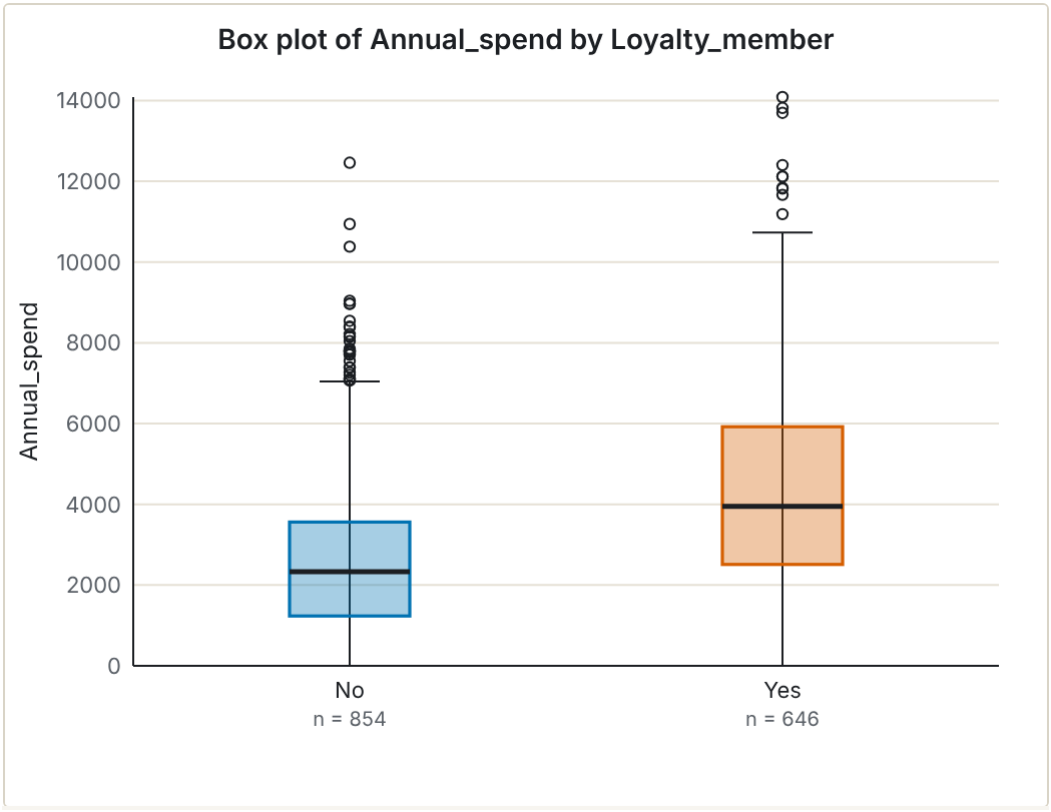


Figure 13. *Box plot of Annual_spend by Loyalty_member.*

Medians of Annual_spend: No 2330.50, Yes 3949.00. The same plot reveals a feature you cannot see in a table: the lowest values of both groups reach 0.00, which means that some customers spent nothing in the year (section 2.6).

2.4 Group means with confidence intervals

Choose **Means with confidence intervals (Group averages)**, the variable **Satisfaction_1to10** and the groups **Main_channel** (Figure 14).

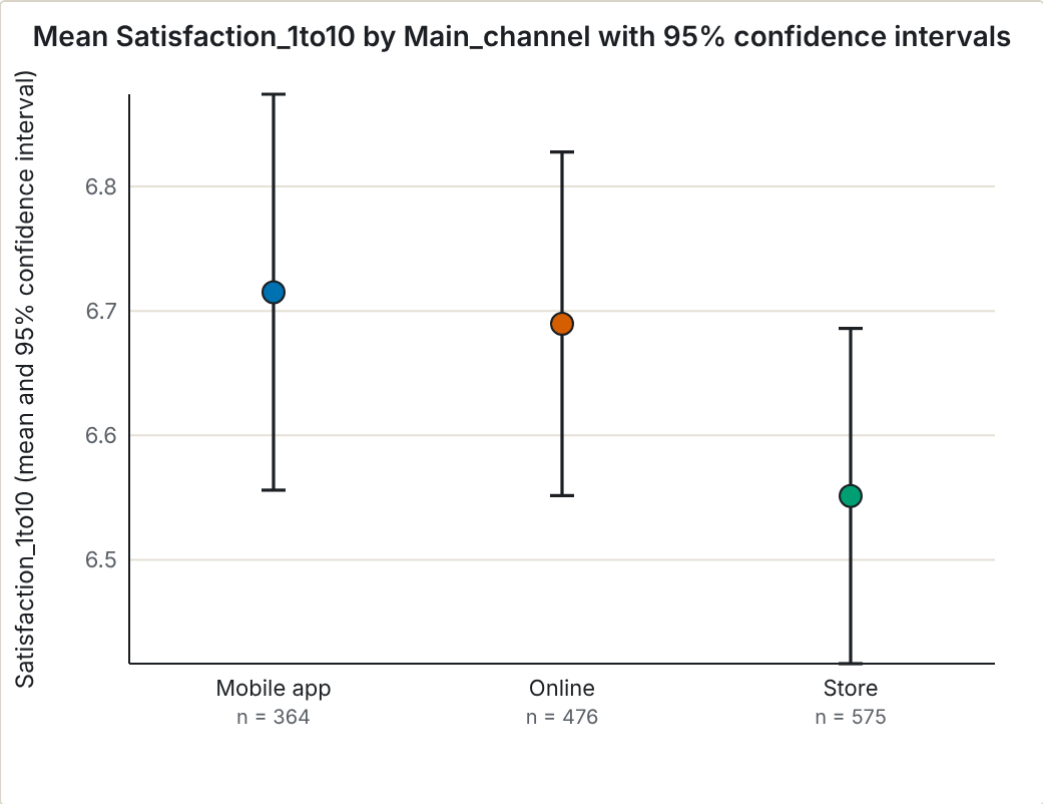


Figure 14. *Mean satisfaction by main channel, with 95% confidence intervals.*

Table 7

Mean Satisfaction_1to10 by Main_channel

Channel	n	Mean	95% CI
Mobile app	364	6.72	[6.56, 6.87]
Online	476	6.69	[6.55, 6.83]
Store	575	6.55	[6.42, 6.69]

Concept: confidence interval

A 95% confidence interval is a range of plausible values for the average of all customers like these, given the sample. Where two intervals do not overlap, the groups very likely differ. Overlapping intervals do not by themselves show that the groups are the same, and the chart is no substitute for a test; section 5.5 tests the same comparison.

What it means for our example

The three intervals overlap heavily: satisfaction is about 6.6 to 6.7 in every channel. Charts show patterns, not proof.

2.5 Scatter plots and the correlation heat map

Scatter plot. Choose **Scatter plot with line (Do two variables go together?)**, **Income_k** on the horizontal axis and **Annual_spend** on the vertical axis (Figure 15). SaNaStat draws the dots and a dashed least-squares line and reports r. Optionally colour the dots by a group, here Region.



Figure 15. *Scatter plot of Annual_spend against Income_k, coloured by region.*

Higher income goes with slightly higher spend: "The relationship is small and positive (Pearson r = .16, n = 1419)", with a slope of 15.368 per extra thousand of income. The dots form a broad cloud around the line.

Correlation heat map. Choose **Correlation heatmap (Many variables at once)**, click **Select none**, and tick the nine variables Income_k, Tenure_months, Visits_per_month, Avg_basket, Discount_share, Complaints_12m, Satisfaction_1to10, Recommend_0to10 and Annual_spend (Figure 16). Each cell shows r; the colour shows its sign and strength.

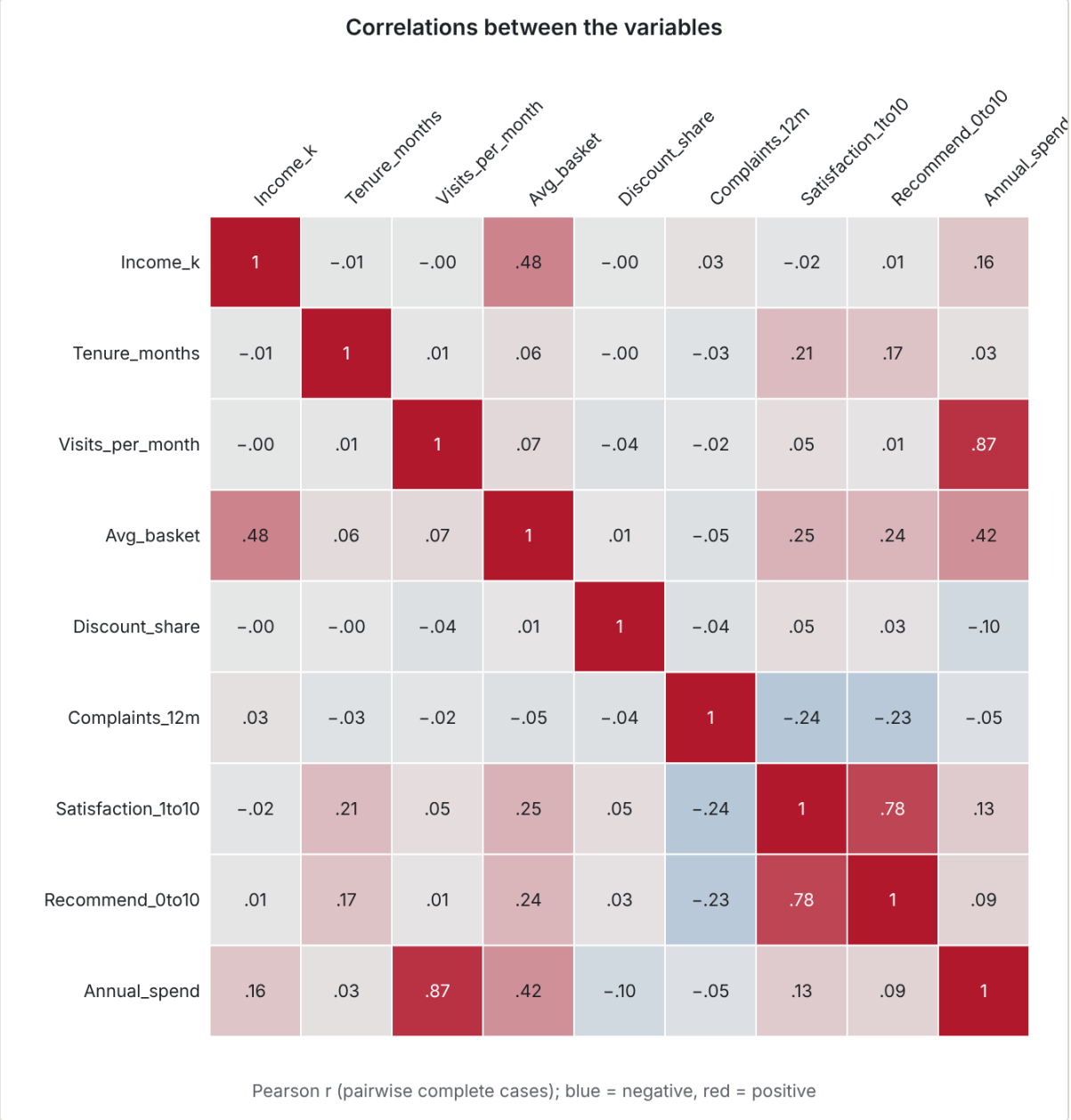


Figure 16. *Correlation heat map of nine customer variables (Pearson r, pairwise complete cases).*

What it means for our example

The strongest pair is Visits_per_month and Annual_spend (r = .87). Next come Satisfaction_1to10 and Recommend_0to10 (r = .78) and Income_k and Avg_basket (r = .48). Annual_spend also goes with Avg_basket (r = .42) and weakly with Income_k (r = .16). Complaints_12m is negatively related to satisfaction (r = −.24) and to willingness to recommend (r = −.23). Tenure_months goes with satisfaction (r = .21).

SaNaStat adds: "Pairs above about .80 (in either direction) measure nearly the same thing, which matters when both are used as predictors." Satisfaction and the recommendation rating are close to that line, so a model should not be expected to separate their effects.

Note: when two variables are parts of a third

Annual spend is, to a good approximation, the number of visits multiplied by the value of the basket. A strong correlation between spend and visits is therefore partly arithmetic, not a discovery about what *drives* spending. Chapter 5 returns to this.

2.6 Category percentages and the problem of skew

Choose **Bar chart of percentages (Categories)** and a categorical variable. For Main_channel: Store 40.7% (611 customers), Online 34.1% (511) and Mobile app 25.2% (378). For Churned: No 83.3% (1250) and Yes 16.7% (250). For Region: South 21.7%, Central 21.5%, North 20.5%, West 19.0% and East 17.3%.

The churn figure matters later: only about one customer in six leaves, so churn is an *imbalanced* outcome (chapter 6).

Annual_spend: skew, zeros and the logarithm

Section 2.2 showed that Annual_spend is right-skewed. A regression of a skewed variable gives residuals (prediction errors) that are not symmetric. The usual remedy for money is the *logarithm*, which turns multiplicative differences into additive ones and pulls in the long right tail. SaNaStat makes the new variable for you in card 4 of the **Data** screen (section 1.6).

1. On the **Data** screen scroll to card *4 · Compute and recode variables (optional)*.
2. In **What do you want to make?** choose **Transform a scale variable (logarithm, standardise, reverse …)**, choose **Annual_spend** as the **Variable** and **ln(1 + x)** as the **Transformation** (Figure 17). The program explains: "Like the logarithm but allows zeros (counts, spend of 0). Needs x > −1."
3. Leave the name as suggested, **Annual_spend_ln1p**, and click **Add the new variable**. SaNaStat confirms: "Added Annual_spend_ln1p: 1500 of 1500 respondents have a value."
4. Repeat with the transformation **Natural logarithm, ln(x)**. The suggested name is **Annual_spend_ln**, and SaNaStat warns: "Added Annual_spend_ln: 1418 of 1500 respondents have a value. Some values could not be computed and are missing (for example a logarithm needs positive values and a ratio needs a denominator other than 0)."

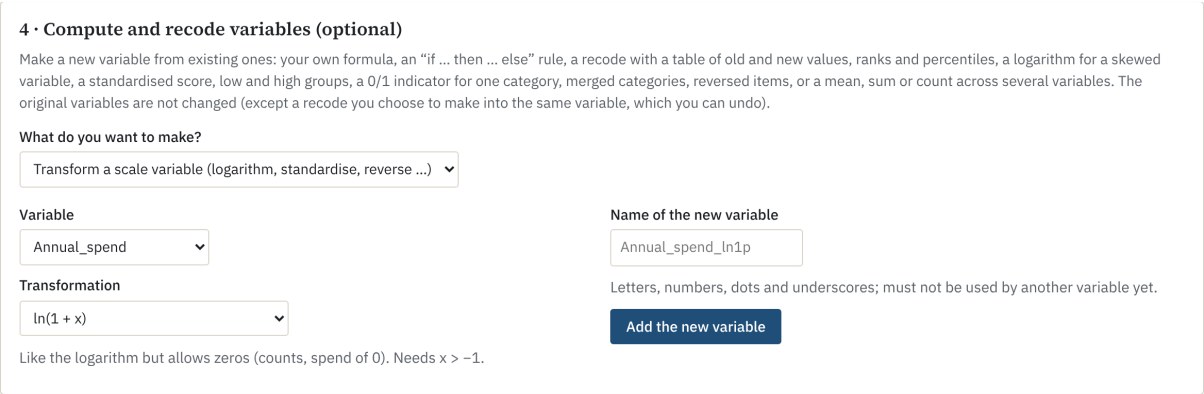


Figure 17. *Card 4 of the Data screen, ready to add ln(1 + Annual_spend).*

The two choices differ in what they do with a spend of 0. The logarithm ln(x) needs a positive value, so the 82 customers who spent nothing get no value (1500 − 82 = 1418 remain). The version ln(1 + x) allows zeros and keeps all 1500 customers. Both new variables are scale variables. Now draw the histograms of the two.

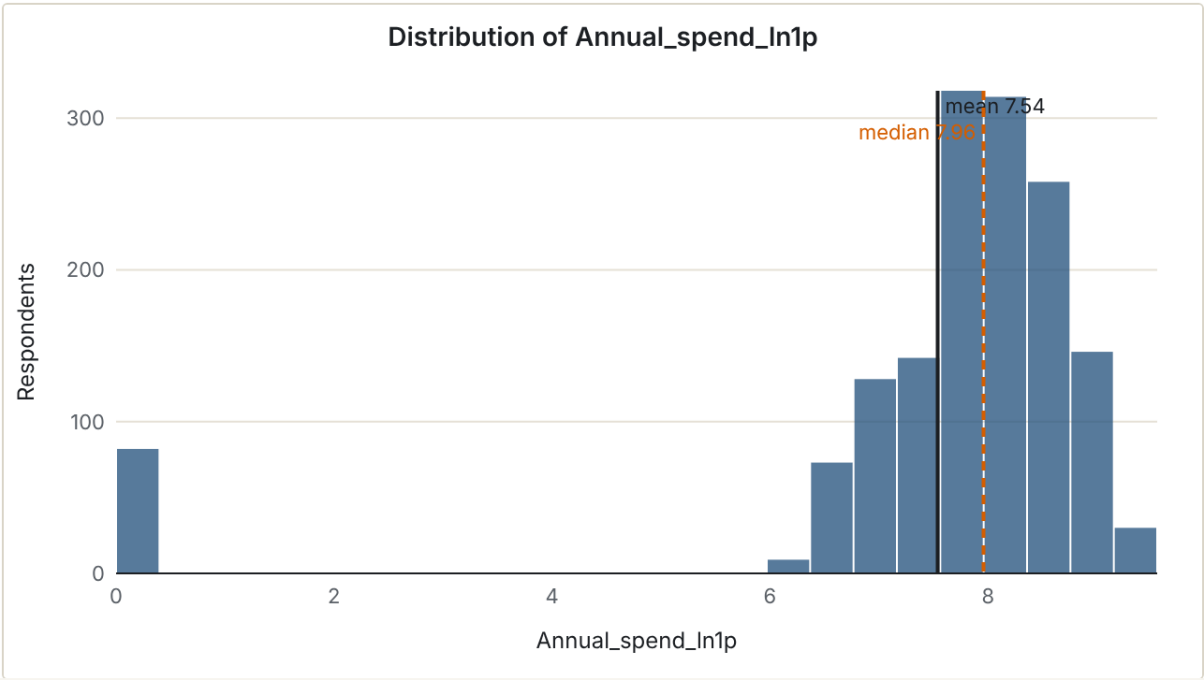


Figure 18. *Histogram of Annual_spend_ln1p (logarithm of one plus Annual_spend).*

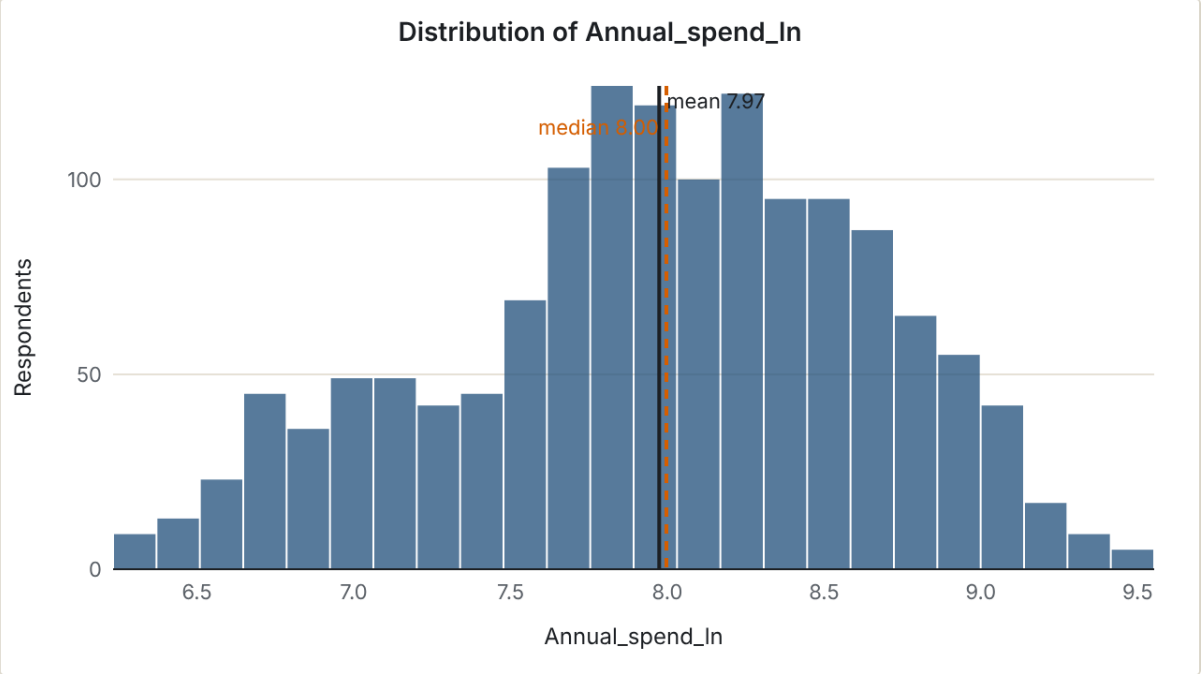


Figure 19. *Histogram of Annual_spend_ln (logarithm of Annual_spend; customers with zero spend have no value).*

What it means for our example

Annual_spend_ln1p is *not* better than the raw variable: its skewness is −3.19, a long tail to the left (mean 7.54, median 7.96, SD 1.93). The reason is that 82 customers spent nothing (Annual_spend = 0). A logarithm of one plus zero is zero, far below the next value (the smallest logarithm of a positive spend is 6.23), and this creates the left tail. Annual_spend_ln leaves those 82 customers out (1418 values remain) and is nearly symmetric: mean 7.97, median 8.00, SD 0.68, skewness −0.30.

Note: choose the transformation with the question in mind

Customers with zero spend are real customers, probably the ones most at risk of leaving. Leaving them out of a logarithmic analysis answers a different question ("among customers who bought something, what drives how much?"). Say so in your report. Section 5.2 compares both versions.

2.7 Groups, indicators and merged categories

The same card 4 makes three other kinds of new variable. All of them leave the original untouched.

Two or more groups from a scale variable. Choose **Make groups from a scale variable (median split, tertiles ...)**, the variable **Satisfaction_1to10** and **Two groups at the median** (Figure 20). SaNaStat shows the cut point before you add the variable: "Cut point: 6.6". A value equal to the cut point goes to the *lower* group. The new categorical variable is

Satisfaction_1to10_group, with the groups *Low (≤ median)* and *High (> median)*; 1415 of the 1500 customers have a value, because 85 did not answer the satisfaction question. The other choices are **Three equal-sized groups (tertiles)**, **Four equal-sized groups (quartiles)** and **My own cut points**: tertiles of Income_k, named Income_tertile, have the cut points 40.4 and 59.733 and the groups *Low third*, *Middle third* and *High third*.

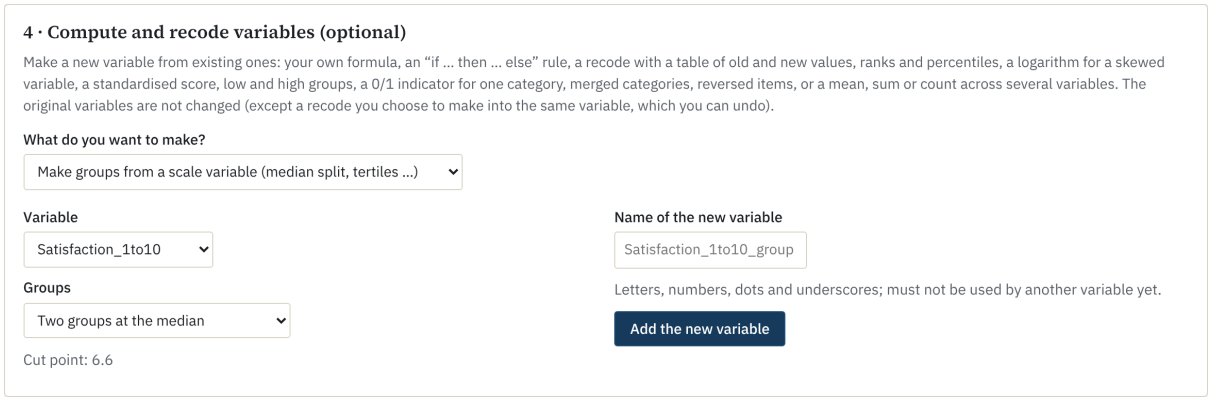


Figure 20. *Making low and high satisfaction groups at the median (cut point 6.6).*

A 0/1 indicator of a category. Choose **Make a 0/1 indicator of a category**, the variable **Churned** and tick **Yes**. The new scale variable **Churned_Yes** is 1 for a customer who churned and 0 for the others (1500 of 1500 customers have a value). Its mean is the share of customers who churned, 16.7% (250 of 1500).

Merging categories. Choose **Merge categories into one**, the variable **Region**, tick **East** and **West** and type *East_West* in **Name of the merged category** (Figure 21). The new variable is **Region_merged**, with the categories Central, East_West, North and South. The merged categories are counted together in every later analysis.

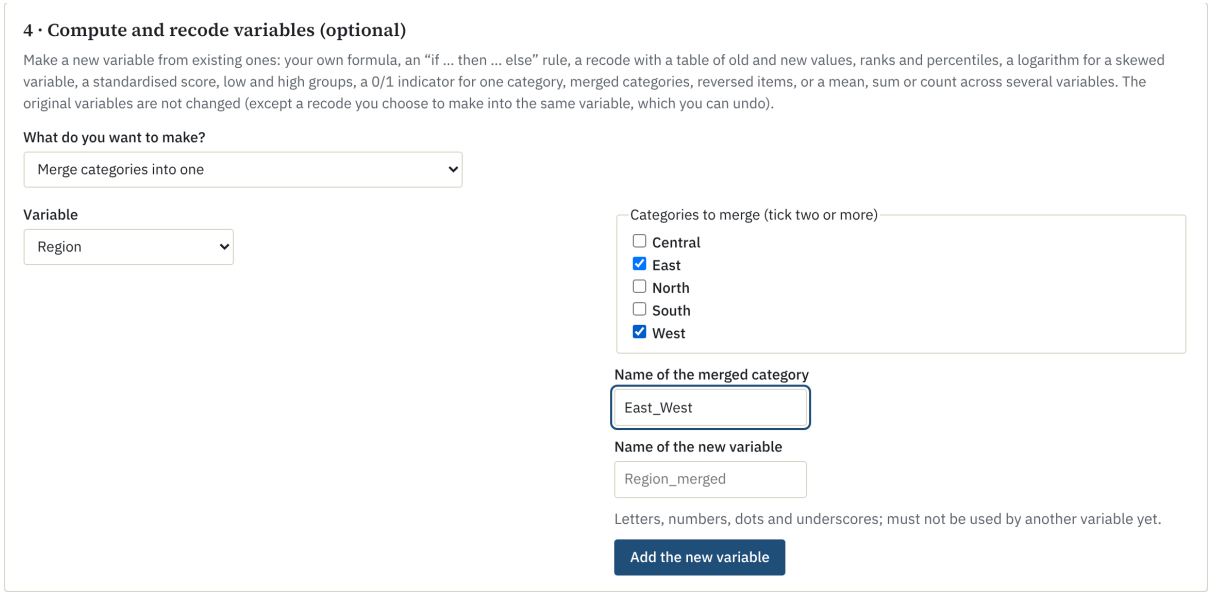


Figure 21. *Merging East and West into one category.*

What it means for our example

Customers with low satisfaction (at or below 6.6) churn far more often than those with high satisfaction: 25% (184 of 726) against 7% (50 of 689). The chi-square test of Satisfaction_1to10_group by Churned gives $\chi^2(1, N = 1415) = 83.79$, $p < .001$, Cramér's $V = .24$ (small). After merging East and West, churn is 23% (75 of 322) in Central, 15% (82 of 544) in East_West, 14% (44 of 308) in North and 15% (49 of 326) in South: $\chi^2(3, N = 1500) = 13.06$, $p = .005$, Cramér's $V = .09$. Only Central has more churners than expected (adjusted residual 3.60).

Note: dichotomising a scale loses information

A median split throws away the differences within each group, and a customer just below the cut point is treated like one far below it. Use groups when the question is about groups, as in a report for managers; for a statistical analysis keep the scale variable where you can (the regression of section 5.1 uses Satisfaction_1to10 itself).

Combining two scale variables. Choose **Combine two scale variables (difference, change, ratio ...)**, the first variable (**Variable A**), the way to combine (**Difference, A − B**, **Sum, A + B**, **Product, A × B**, **Ratio, A ÷ B** or **Percentage change, (A − B) ÷ B × 100**) and the second variable (**Variable B**) (Figure 22). A customer who has no value on A or on B gets no value, and neither does a customer with B = 0 when you ask for a ratio or a percentage change. Section 4.4 uses a difference.

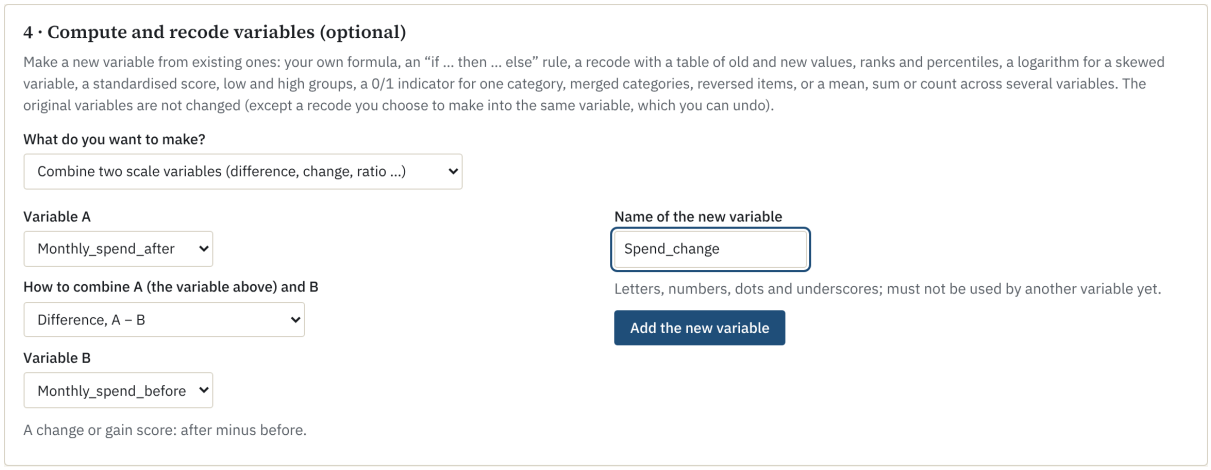


Figure 22. *Combining two variables: Monthly_spend_after − Monthly_spend_before.*

Exercises

Exercise: know your customers

1. Describe Tenure_months and Complaints_12m. Which of the two is more skewed? Answer: Tenure_months (skewness 1.49) against Complaints_12m (1.15), and both are flagged as clearly skewed.

2. Draw the box plot of Visits_per_month by Loyalty_member. Do members visit more often? Answer: yes, the group means are 4.01 and 2.65 (section 2.1).

3. Draw the heat map and find the variables with the weakest correlation with Annual_spend. Answer: Complaints_12m (r = −.05), Tenure_months (r = .03) and Recommend_0to10 (r = .09) have almost no linear relation with spend; Discount_share is weakly negative (r = −.10).

3. Survey scales: trust, value and loyalty

Brand and customer surveys measure ideas such as trust or loyalty with several questions (*items*) and average them into a *scale score*. This chapter uses *practice_brand_survey.csv*: 500 simulated respondents who rated Trust1 to Trust4, Value1 to Value4 and Loyalty1 to Loyalty3 on a 1 to 7 scale, with Age_group (18-24, 25-34, 35-49, 50+), Gender and Customer_type (Long-term, New, Regular).

On the **Data** screen, open the file (start a new project). SaNaStat reports 500 respondents and 15 variables, and counts *11 scale · 3 categorical · 1 ID*: RespondentID has the type ID and is never offered. Value2 and Loyalty2 each have 18 missing answers.

3.1 Data quality: straight-lining

Concept: careless responding and straight-lining

Some respondents answer without reading: the same answer everywhere (*straight-lining*), random answers, or far too fast. Their answers weaken correlations and can create or hide effects. No single index catches every kind of careless answering, so the usual advice is to compute several indices and exclude only respondents flagged by more than one (Curran, 2016; Meade & Craig, 2012).

1. On the **Data** screen click **Check data quality →** in card *6 · Data quality screening*. Check the *Screening settings* (Figure 23): all 11 items are ticked, the rule is **Suggest exclusion when flagged by 2 or more indices (Recommended)**, and no completion time is chosen because the file has none.
2. Click **Screen the data**.

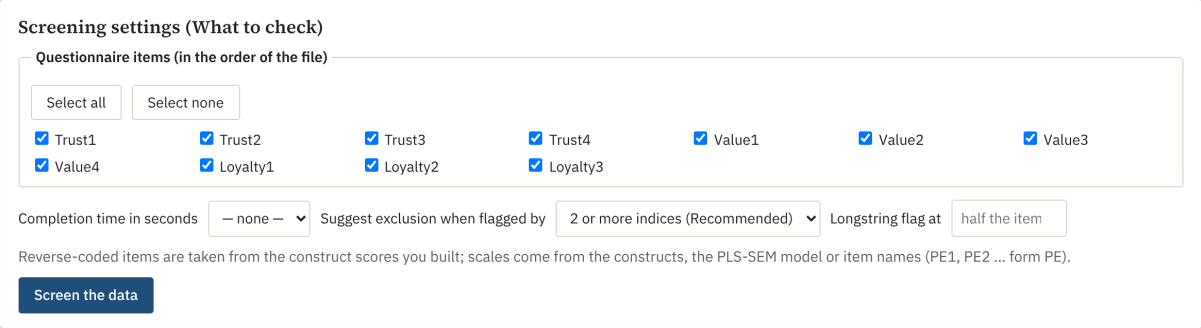


Figure 23. *The screening settings for the brand survey.*

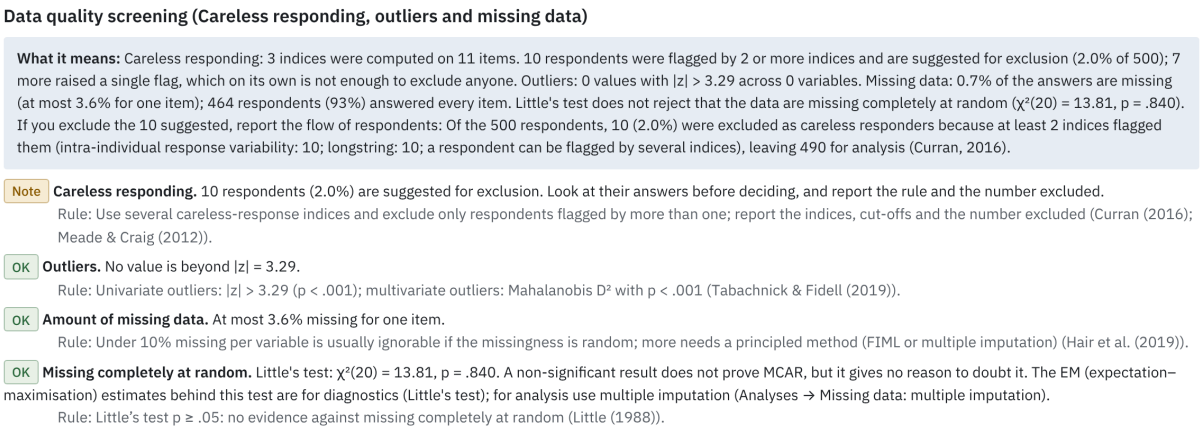


Figure 24. *The start of the screening report: what it means and the checks.*

The report (Figure 24) says: "Careless responding: 3 indices were computed on 11 items. 10 respondents were flagged by 2 or more indices and are suggested for exclusion (2.0% of 500); 7 more raised a single flag, which on its own is not enough to exclude anyone." It adds that there are no outliers, that 0.7% of the answers are missing (at most 3.6% for one item), that 464 respondents (93%) answered every item, and that Little's test does not reject that the answers are missing completely at random, $\chi^2(20) = 13.81$, $p = .840$.

Table 8

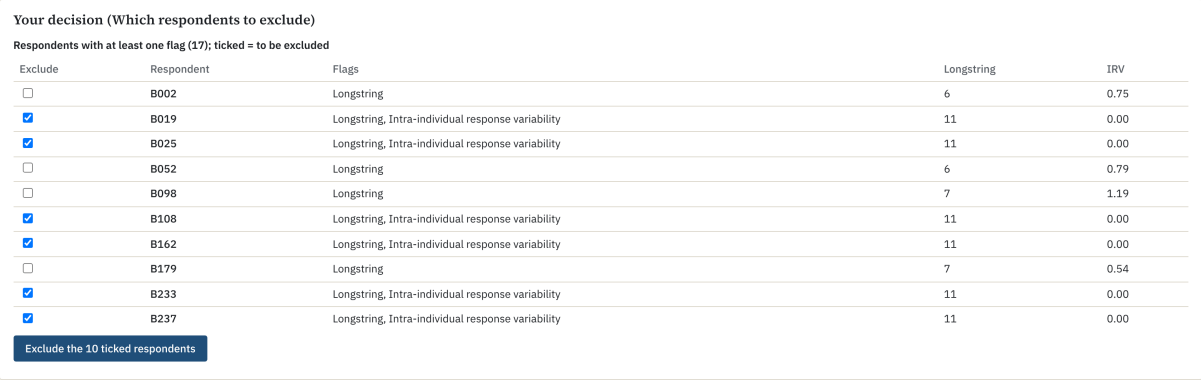
The careless-response indices for the brand survey

Index	Flag when	Respondents flagged
Longstring (same answer many times in a row)	6 or more identical answers in a row (of 11 items)	17 (3.4%)
Intra-individual response variability (no variation at all)	SD of the answers = 0	10 (2.0%)
Mahalanobis distance (unusual answer pattern)	D^2 with $p < .001$	0 (0.0%)
Person–total correlation	Not used: the item means are too similar (SD 0.03 < 0.25)	–
Even–odd consistency	Not used: needs at least 4 scales with two or more items (found 3)	–
Response time	Not used: no completion time chosen	–

What it means for our example

Ten respondents (B019, B025, B108, B162, B233, B237, B352, B374, B424 and B496) gave the same answer to all 11 items (longest run 11, standard deviation 0.00): these are the planted straight-liners, and each is flagged by two indices. Seven more respondents (B002, B052, B098, B179, B244, B431 and B461) have a longest run of 6 to 9 identical answers, so only the longstring index flags them; one flag is not enough, and a run of 6 to 9 in an 11-item questionnaire can occur in an honest answer. This is why no single index is enough, and why the rule needs two flags. The even–odd index is switched off for this file, because it is informative only when a questionnaire has at least four scales and this one has three; SaNaStat says so in the table instead of reporting an index that would be unreliable.

The card *Your decision (Which respondents to exclude)* lists the 17 respondents with at least one flag; the 10 suggested ones are ticked (Figure 25).



Your decision (Which respondents to exclude)

Respondents with at least one flag (17); ticked = to be excluded

Exclude	Respondent	Flags	Longstring	IRV
☐	B002	Longstring	6	0.75
☑	B019	Longstring, Intra-individual response variability	11	0.00
☑	B025	Longstring, Intra-individual response variability	11	0.00
☐	B052	Longstring	6	0.79
☐	B098	Longstring	7	1.19
☑	B108	Longstring, Intra-individual response variability	11	0.00
☑	B162	Longstring, Intra-individual response variability	11	0.00
☐	B179	Longstring	7	0.54
☑	B233	Longstring, Intra-individual response variability	11	0.00
☑	B237	Longstring, Intra-individual response variability	11	0.00

Exclude the 10 ticked respondents

Figure 25. *Your decision: the flagged respondents; the 10 suggested for exclusion are ticked.*

1. Look at the flagged respondents and untick anyone you want to keep.
2. Click **Add to Word report** under the screening report, so that the rule, the cut-offs and the number excluded are documented.
3. Click **Exclude the 10 ticked respondents**. The card *Screened data in use* confirms it: "10 of 500 respondents are excluded: B019, B025, B108, B162, B233, B237, B352, B374, B424, B496. Every analysis now uses the remaining 490."

Nothing is deleted from your file, and **Undo the exclusion (use all 500 respondents)** brings everyone back.

How to report it

"Data were screened for careless responding with 3 indices (longstring, intra-individual response variability, Mahalanobis distance; Curran, 2016; Meade & Craig, 2012); 10 of 500 respondents were flagged by at least 2 indices and excluded; 0.7% of item responses were missing, and Little's (1988) MCAR test was not significant, $\chi^2(20) = 13.81$, $p = .840$."

3.2 Reliability and scale scores

Card *3 · Build construct scores (optional)* of the **Data** screen builds a construct (scale) score as the average of its items and first checks that the items belong together.

1. In **Name of the construct** type **Trust**.
2. In **Quick pick: items starting with** type **Trust** and press Tab. Trust1 to Trust4 are ticked, and **Lowest answer** and **Highest answer** are filled in from the data (1 and 7).
3. Read Cronbach's α, McDonald's ω and the item statistics (Figure 26). No item is reverse-worded here, so leave the **R** boxes empty.
4. Click **Create the construct score**.
5. Repeat for **Value** and **Loyalty**; Table 9 gives the three results.

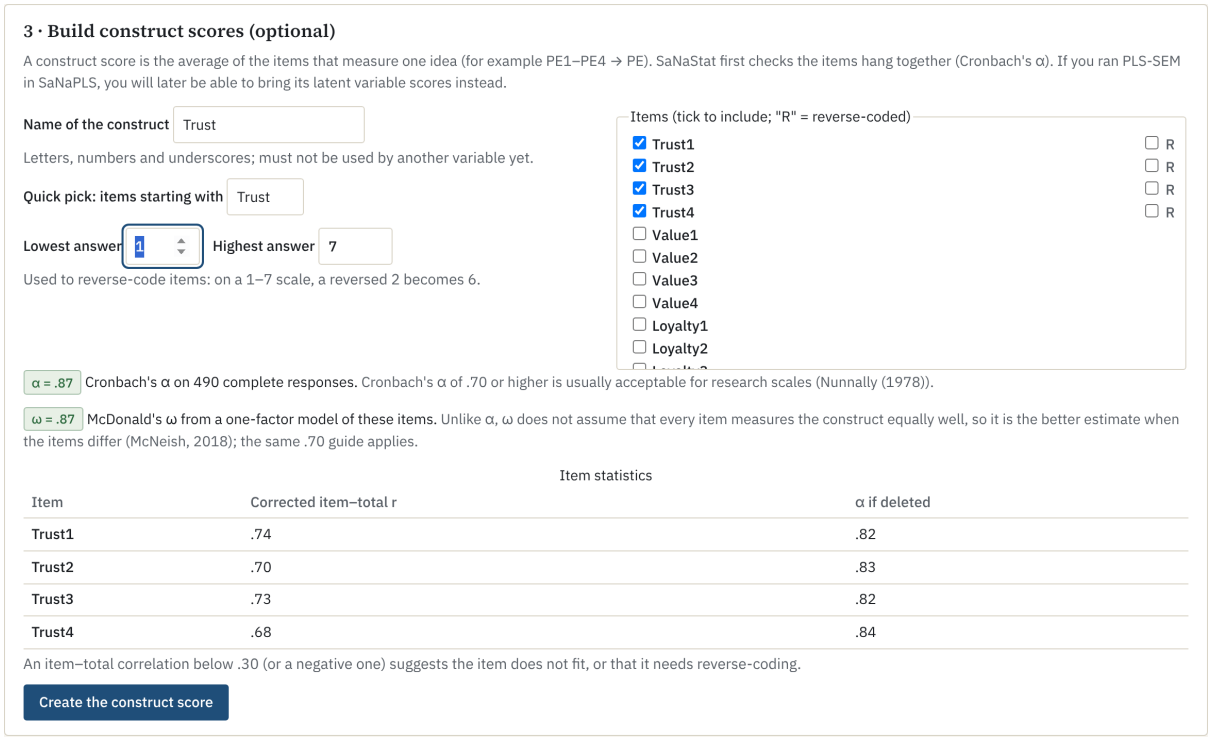


Figure 26. *Building the Trust score: Cronbach's α, McDonald's ω and the item statistics.*

Concept: Cronbach's α and McDonald's ω

Cronbach's α summarises how consistently the items of a scale move together. It rises with the number of items and with their average correlation. McDonald's ω comes from a one-factor model of the items; unlike α 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 is the correlation of an item with the sum of the other items; the column "α if deleted" shows the α the scale would have without that item.

Cut-off values: internal consistency

Cronbach's α and McDonald's ω of .70 or more are usually acceptable for research scales (Nunnally, 1978; McNeish, 2018).

Corrected item–total correlation below .30, or negative, suggests that the item does not fit or needs reverse-coding.

Table 9

Reliability of the three scales (490 respondents after screening)

Scale	α (complete cases)	ω	Corrected item–total r	α if item deleted
Trust (4 items)	.87 (n = 490)	.87	.74, .70, .73, .68	.82, .83, .82, .84
Value (4 items)	.80 (n = 472)	.81	.62, .65, .61, .61	.76, .74, .76, .76
Loyalty (3 items)	.84 (n = 472)	.84	.74, .69, .69	.75, .79, .79

What it means for our example

All three scales are reliable (α and ω between .80 and .87), and deleting an item would lower α every time, so all items stay. α is computed only for respondents who answered all items of the scale: Value2 and Loyalty2 each have 18 missing answers, which is why the Value and Loyalty scores exist for 472 respondents and the Trust score for all 490. If your questionnaire has reverse-worded items, tick **R** for them: otherwise they correlate negatively with the others.

Tip: the construct score is the average of its items

A respondent who skipped any item of a scale gets no score for it. Check the **Valid** and **Missing** columns in card 2 after creating the scores.

Describe the scores with **Analyses** → **Descriptive statistics** (Figure 27).

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

Descriptive statistics

Variable	n	Missing	Mean	SD	Median	Min	Max	Skewness	Kurtosis
Trust	490	0	3.86	1.34	4.00	1.00	7.00	−0.01 (0.11)	−0.52 (0.22)
Value	472	18	3.87	1.17	3.75	1.00	6.75	0.06 (0.11)	−0.29 (0.22)
Loyalty	472	18	3.88	1.33	4.00	1.00	7.00	0.06 (0.11)	−0.49 (0.22)

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

Figure 27. *Descriptive statistics of the three scale scores.*

Table 10

Descriptive statistics of the scale scores (1 to 7 scale)

Score	n	Mean	SD	Median	Min	Max	Skewness
Trust	490	3.86	1.34	4.00	1.00	7.00	−0.01
Value	472	3.87	1.17	3.75	1.00	6.75	0.06
Loyalty	472	3.88	1.33	4.00	1.00	7.00	0.06

No variable is strongly skewed (all |skewness| ≤ 1).

How to report it

"Trust (4 items, α = .87, ω = .87), Value (4 items, α = .80, ω = .81) and Loyalty (3 items, α = .84, ω = .84) were computed as the mean of their items on a 1–7 scale."

3.3 Compare loyalty between groups

Choose **Analyses** → **One-way ANOVA**, **Loyalty** as the score and **Customer_type** as the groups (Figure 28, Figure 29).

One-way ANOVA (Comparing Loyalty across Customer_type)

What it means: Loyalty does not differ significantly across the 3 Customer_type groups (ω² = −.004).

Note **Equal variances.** Levene's test suggests unequal variances: Welch's ANOVA is used with Games–Howell 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: Long-term W = .983, p = .103; New W = .976, p = .034; Regular W = .984, p = .011. 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 = −.004 (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 28. *One-way ANOVA of Loyalty by Customer_type: what it means and the checks.*

Group descriptives

Group	n	Mean	SD	Skewness	Shapiro–Wilk W	p (normality)
Long-term	128	3.86	1.22	0.31	.983	.103
New	116	3.93	1.53	0.09	.976	.034
Regular	228	3.88	1.28	−0.10	.984	.011

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	0.26	2	0.13	0.07	.929
Within groups	830.26	469	1.77		
Total	830.52	471			

Copy table | Copy as picture | PNG | CSV

Robust test and equal variances

Test	Result
Welch's F	F(2, 251.20) = 0.06, p = .940
Levene's test	F(2, 469) = 4.33, p = .014

Copy table | Copy as picture | PNG | CSV

Effect sizes

Measure	Value	90% CI	Size
η²	.000	[.000, .002]	negligible
ω² (less biased)	.000	[.000, .002]	negligible
η² from Welch's F (approximate)	.000	[.000, .002]	negligible
ω² from Welch's F (approximate)	.000	[.000, .002]	negligible

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). Because the variances differ, the Welch-based rows convert Welch's F and its degrees of freedom into η² and ω²: an approximation, as there is no exact effect size for Welch's ANOVA.

Figure 29. *The group descriptives, the ANOVA table, Welch's test and the effect sizes.*

SaNaStat reports the classic ANOVA, Welch's F (robust to unequal variances), the effect sizes η^2 and ω^2 with their intervals, pairwise comparisons, the rank-based Kruskal–Wallis test and normal Q–Q plots of each group.

What it means for our example

Mean loyalty is almost the same for long-term (3.86, n = 128), new (3.93, n = 116) and regular customers (3.88, n = 228). Levene's test suggests unequal variances (F(2, 469) = 4.33, p = .014; new customers vary most, SD = 1.53), so SaNaStat uses Welch's test: F(2, 251.20) = 0.06, p = .940; the effect size $\omega^2 = -.004$ is negligible (the table prints it as .000 with the interval [.000, .002]). The Kruskal–Wallis test agrees, H(2) = 0.17, p = .919. Customer type does not matter for loyalty in this sample.

Now compare Loyalty by **Age_group**.

What it means for our example

The four age groups differ in the sample (18-24: 3.88; 25-34: 3.68; 35-49: 4.05; 50+: 3.91), but not significantly: $F(3, 468) = 1.72$, $p = .161$, $\eta^2 = .011$, $\omega^2 = .005$; Kruskal–Wallis $H(3) = 4.55$, $p = .208$. The largest gap, between 25-34 and 35-49 (0.36 points), has an adjusted p of .109. Do not report this as an age effect.

How to report it

"Loyalty did not differ between age groups, $F(3, 468) = 1.72$, $p = .161$, $\eta^2 = .011$, 90% CI [.000, .026]."

3.4 What drives loyalty: correlation and regression

Choose **Analyses** → **Correlation and regression**, then **Bivariate and partial correlation**, and tick Trust, Value and Loyalty (Figure 30).

Correlation matrix

	1	2	3
1. Trust	—		
2. Value	.37***	—	
3. Loyalty	.63***	.46***	—

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

Figure 30. *Pearson correlations of the three scale scores.*

Trust and Loyalty correlate r = .63 (95% CI [.57, .68]), Value and Loyalty r = .46 ([.38, .53]) and Trust and Value r = .37 ([.29, .44]); all p < .001 after the Holm adjustment for several tests.

Now regress Loyalty on Trust and Value: choose **Multiple linear regression (Which variables predict a score?)**, **Loyalty** as the **Outcome (scale)**, and tick Trust and Value as **Block 1 predictors** (Figure 31, Figure 32).

Multiple linear regression (Predicting Loyalty)

What it means: The predictors together explain 45.4% of the variance in Loyalty (R^2 = .45, f^2 = 0.83, large). Significant predictors: Trust (β = .53), Value (β = .26).

OK **Sample size.** 454 respondents for 2 predictor terms (needed: 66).
Rule: At least 50 + 8k respondents for k predictors to test the whole model (Green (1991)).

OK **Collinearity.** Largest VIF = 1.16.
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 = 0.997, p = .514.
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) $\chi^2(2)$ = 0.27, p = .873.
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 = 0.025; 1 standardised residual beyond ±3.
Rule: No Cook's distance above 1 (Cook & Weisberg (1982)).

Info **Independent errors.** Durbin–Watson = 2.10 (values near 2 mean no order effect; it matters only when the row order means something, such as time).

Info **Effect size.** f^2 = 0.83 (large).
Rule: f^2: .02 small, .15 medium, .35 large (Cohen (1988)).

Figure 31. *The regression result: what it means and the checks.*

Model summary

Model	Predictors	R^2	Adjusted R^2	F	df	p	ΔR^2	p (ΔR^2)
1	Trust, Value	.454	.452	187.86	2, 451	< .001		

Copy table | Copy as picture | PNG | CSV

Coefficients (model 1)

Predictor	B	SE	β	t	p	95% CI of B	VIF
(Intercept)	0.717	0.178		4.03	< .001	[0.367, 1.066]	
Trust	0.519	0.037	.530	14.16	< .001	[0.447, 0.591]	1.16
Value	0.297	0.042	.264	7.06	< .001	[0.215, 0.380]	1.16

n = 454 (36 respondents with missing values left out). β = standardised coefficient.

Figure 32. *The model summary and the coefficients.*

What it means for our example

Trust and Value together explain 45.4% of the variance in Loyalty (R^2 = .45, adjusted R^2 = .452; F(2, 451) = 187.86, p < .001). Both predict loyalty: Trust (B = 0.519, β = .53, t(451) = 14.16, p < .001) more strongly than Value (B = 0.297, β = .26, t(451) = 7.06, p < .001). One more point of trust goes with about half a point more loyalty when value is held constant. All checks are fine: the largest VIF is 1.16, residuals are normal (Shapiro–Wilk p = .514), the error spread is constant (Breusch–Pagan p = .873) and no case is influential (largest Cook's distance 0.025). The regression used 454 respondents; 36 with a missing item were left out.

Cut-off values: regression

Sample size: at least 50 + 8k respondents for k predictors (Green, 1991). **VIF:** below 5 no concern (Hair et al., 2019). **f^2:** .02 small, .15 medium, .35 large (Cohen, 1988).

Note: correlation is not causation

The data are a single survey. A regression shows which scores go together with loyalty, not that raising trust would raise loyalty. For marketing decisions the direction is plausible but needs an experiment or longitudinal data.

How to report it

"A multiple linear regression of Loyalty on Trust and Value was significant, F(2, 451) = 187.86, p < .001, R^2 = .45. Trust predicted Loyalty, β = .53, t(451) = 14.16, p < .001; Value predicted Loyalty, β = .26, t(451) = 7.06, p < .001."

3.5 Going further with scales

Three tools of the *Measurement and models* group take a scale analysis further. They are described step by step in the Common Statistical Tests manual, and they work on the brand survey in the same way:

- **Measurement (chapter 12 of that manual).** **Factor analysis and reliability** checks that the items of Trust, Value and Loyalty load on their own factor (exploratory factor analysis), fits a confirmatory factor analysis with its fit indices, and reports ω, the average variance extracted and the HTMT ratio for discriminant validity. Report this before you interpret a regression of scale scores.
- **Mediation and moderation (chapter 13).** **Mediation and moderation** asks, for example, whether Value carries part of the effect of Trust on Loyalty (an indirect effect, with a bootstrap interval), or whether the link between Trust and Loyalty is stronger for long-term customers (an interaction).

- **Missing answers (chapter 14).** **Missing data: multiple imputation** analyses the 36 respondents who skipped an item instead of leaving them out.

Exercises

Exercise: survey scales

1. Run the screening with **Suggest exclusion when flagged by** set to **any index (Strict)**. How many respondents are suggested now? Compare with the 10 of the recommended rule (the card *Your decision* lists 17 respondents with at least one flag), and decide which rule you would defend in a report.

2. Compare Trust by Gender with an independent-samples t-test. Report what you find, with its effect size. Answer: no difference (d = 0.02, negligible).

3. Planted answer for the chapter: Loyalty depends on both Trust and Value (section 3.4), Customer_type has no effect (section 3.3), and ten respondents straight-lined (section 3.1).

4. Testing a campaign

Practice_marketing.csv contains a randomised campaign test: each customer was randomly given either the *Standard* offer (762 customers) or the *New offer* (738 customers), and the file records whether the customer Responded, and the monthly spend before and after the campaign. Open the file again (**Data** → **Choose a file…**, start a new project) and make Spend_change (section 1.6) if you will follow section 4.4.

4.1 Did more customers respond? The A/B test of a rate

> **Concept: A/B test and uplift**
>
> An A/B test compares two versions of an offer, a page or a message given at random to comparable customers. Version A is the current one (the *control*), version B the new one. Because the groups were formed at random, a difference in the response (or *conversion*) rate can be attributed to the version rather than to who received it. The *uplift* is the change from A to B: in percentage points (B − A) or relative to A ((B ÷ A) − 1). Both need an interval, not only a p-value (Kohavi et al., 2020).

1. Choose **Analyses** → **A/B test** (in the group *Marketing analytics*).
2. In **What did you measure?** choose **A yes/no result per visitor (conversion rate)**; in **Where are the results?** choose **In the data (one row per visitor)**. SaNaStat calls every person a *visitor*; here a visitor is a customer.
3. Choose **Campaign_group** in **Version (which version each visitor saw)**, **Standard** in **Version A (control)** and **New offer** in **Version B (compared with A)**.
4. Choose **Responded** in **Result (two answers, or coded 0/1)** and **Yes** in **Which answer counts as a conversion?** (Figure 33). Keep **Planned share of visitors in version A (%)** at 50.

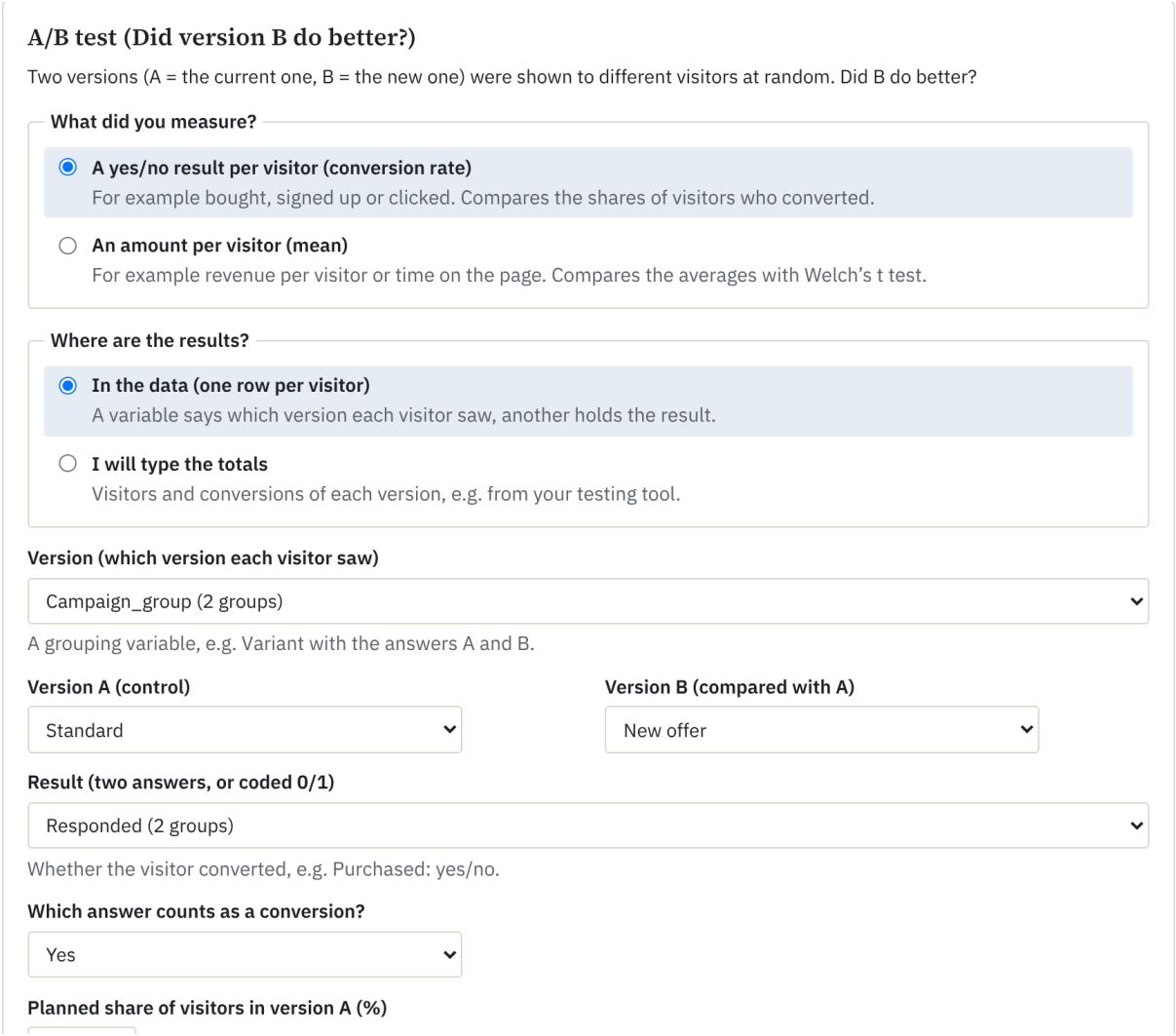


Figure 33. *The A/B test set up for the campaign: Standard (A) against New offer (B), with Responded = Yes as the conversion.*

The result appears at once (Figure 34). It starts with *What it means* and three checks, then gives the rate of each version and the comparison (Figure 35), a chart of the two rates with their intervals (Figure 36) and the sentence for the report (Figure 37).

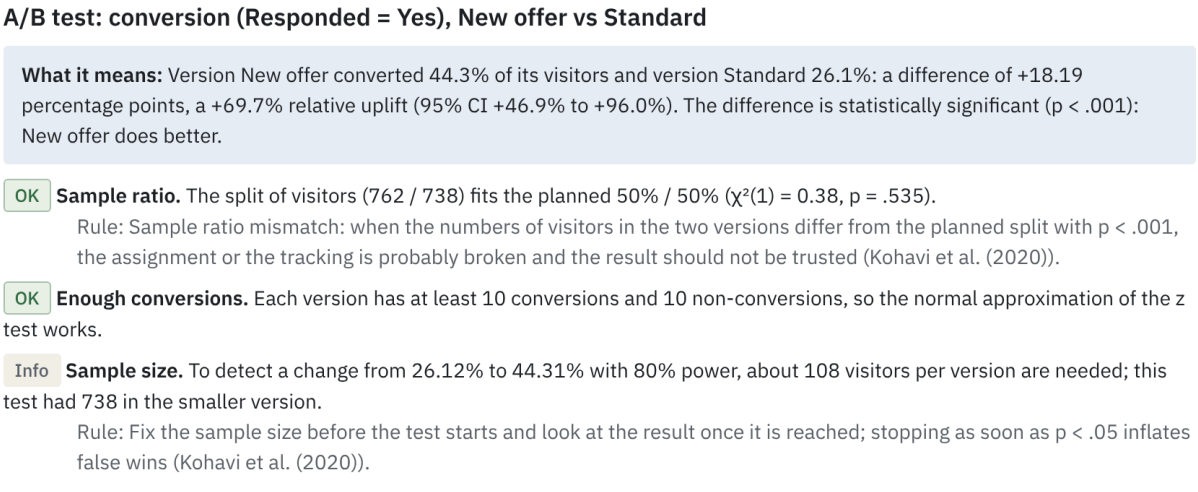


Figure 34. *The A/B test: what it means and the checks.*

Conversion by version

Version	Visitors	Conversions	Rate	95% CI
Standard (A, control)	762	199	26.12%	[23.12%, 29.35%]
New offer (B)	738	327	44.31%	[40.76%, 47.91%]

Interval of each rate: Wilson (1927) score interval.

Figure 35. *Conversion by version: visitors, conversions and the rate with its Wilson interval.*

Table 11

The A/B test of Responded: Standard (A) against New offer (B)

Measure	Estimate	95% CI
Rate of Standard (A, 199 of 762)	26.12%	[23.12%, 29.35%]
Rate of New offer (B, 327 of 738)	44.31%	[40.76%, 47.91%]
Difference (percentage points)	+18.19	[13.40, 22.88]
Relative uplift	+69.7%	[+46.9%, +96.0%]
z test (pooled)	z = 7.38	p < .001
Cohen's h	0.38	

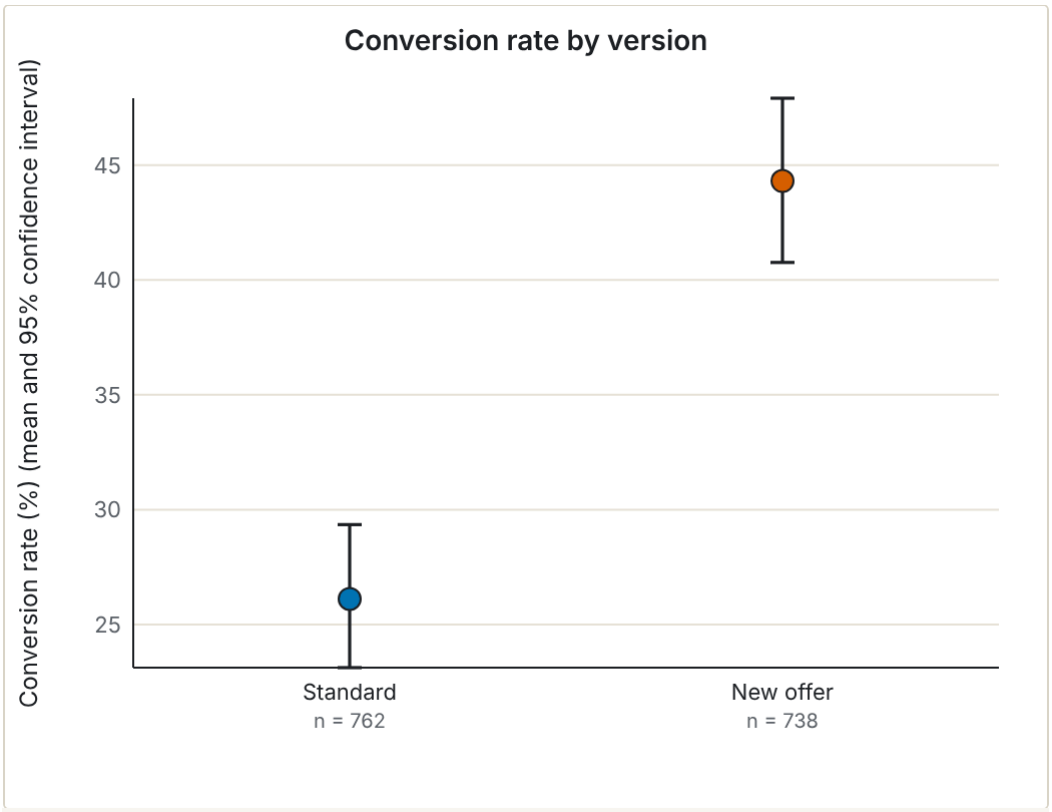


Figure 36. *Conversion rate by version: dot = rate, bars = 95% Wilson interval.*

What it means for our example

SaNaStat writes: "Version New offer converted 44.3% of its visitors and version Standard 26.1%: a difference of +18.19 percentage points, a +69.7% relative uplift (95% CI +46.9% to +96.0%). The difference is statistically significant (p < .001): New offer does better." In business terms, 18 more customers per 100 contacted responded to the new offer, and the 95% interval of the uplift runs from 13.40 to 22.88 percentage points. Cohen's h = 0.38 is the standardised size of the difference between two proportions.

Cut-off values: the checks of an A/B test

Sample ratio: when the numbers of visitors in the two versions differ from the planned split with p < .001, the assignment or the tracking is probably broken and the result should not be trusted (Kohavi et al., 2020). Here the split 762 / 738 fits the planned 50% / 50%, $\chi^2(1) = 0.38$, p = .535. **Enough conversions:** each version needs at least 10 conversions and 10 non-conversions for the z test. **Planned sample size:** fix the sample size before the test starts and look at the result once it is reached; stopping as soon as p < .05 inflates false wins (Kohavi et al., 2020). SaNaStat adds, for information, that about 108 visitors per version would have been enough to detect a change from 26.12% to 44.31% with 80% power; this test had 738 in the smaller version.

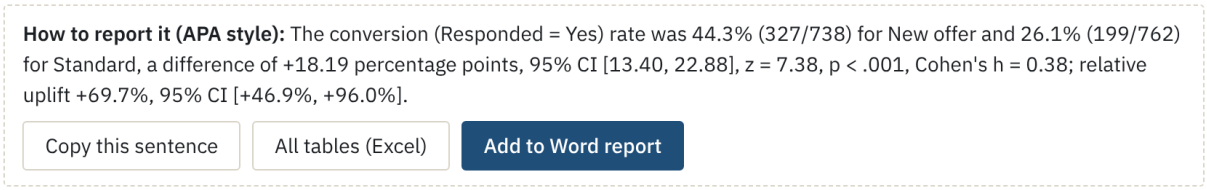


Figure 37. *The sentence for the report, with its buttons.*

How to report it

"The conversion (Responded = Yes) rate was 44.3% (327/738) for New offer and 26.1% (199/762) for Standard, a difference of +18.19 percentage points, 95% CI [13.40, 22.88], z = 7.38, p < .001, Cohen's h = 0.38; relative uplift +69.7%, 95% CI [+46.9%, +96.0%]."

The interval of each rate is the Wilson (1927) score interval, the interval of the difference is Newcombe's (1998) hybrid score interval and the interval of the relative uplift comes from the logarithm of the ratio of the two rates (Katz et al., 1978).

4.2 The same table as a chi-square test

The A/B test of two rates and the chi-square test of a 2 × 2 table answer the same question. Choose **Analyses** → **Chi-square, Fisher, McNemar**, then the test of independence with **Campaign_group** and **Responded** (Figure 38).

Counts (row %)

Campaign_group	No	Yes	Total
New offer	411 (56%)	327 (44%)	738
Standard	563 (74%)	199 (26%)	762

Copy table | Copy as picture | PNG | CSV

Tests

Test	Value	df	p
Pearson chi-square	54.50	1	< .001
Continuity correction (Yates)	53.70	1	< .001
Likelihood ratio	54.90	1	< .001
Fisher's exact test			< .001

Figure 38. *The chi-square test of Campaign_group by Responded: counts and tests.*

Table 12

Responded by Campaign_group (counts and row percentages)

Group	No	Yes	Total
New offer	411 (56%)	327 (44%)	738
Standard	563 (74%)	199 (26%)	762

What it means for our example

$\chi^2(1, N = 1500) = 54.50$, $p < .001$, Cramér's V = .19, 90% CI [.15, .23]. The pooled z of the A/B test is the square root of this χ^2 ($7.38^2 \approx 54.50$), so the two screens agree. The adjusted residuals (±7.38) show that the pattern comes from both cells; all expected counts are adequate (smallest 258.79), and Fisher's exact test ($p < .001$) and the continuity-corrected test ($\chi^2 = 53.70$) agree. Use the A/B screen when you need the uplift and its interval; use the chi-square screen for more than two versions or more than two answers.

Concept: statistical significance and practical importance

Cohen's labels (V of .10 small, .30 medium, .50 large for a 2 × 2 table) describe the size of an association in general, not its commercial value. A "small" V can correspond to a large business difference: here, 18 more responders per 100 customers contacted. Always report the two rates and the uplift, not only the p-value.

4.3 Did monthly spend change? The paired test

Each customer has Monthly_spend_before and Monthly_spend_after. Choose **Analyses** → **t-tests**, then **Paired-samples t-test (The same people measured twice)**, and choose Monthly_spend_before as the first and Monthly_spend_after as the second measurement (Figure 39).

Descriptives (complete pairs)

Measure	n	Mean	SD	Median	Shapiro–Wilk W	p (normality)
Monthly_spend_before	1500	282.64	198.64	238.45		
Monthly_spend_after	1500	304.35	215.75	255.80		
Difference (Monthly_spend_after – Monthly_spend_before)	1500	21.71	29.22	13.50	.737	< .001

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
28.78	1499	< .001	21.71	[20.23, 23.19]

Copy table | Copy as picture | PNG | CSV

Effect size

Measure	Value	95% CI	Size
Cohen's d_z	0.74	[0.69, 0.80]	medium
Hedges' g_z	0.74	[0.69, 0.80]	medium

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	1010454.0	30.61	< .001	.93	[.92, .94]	normal approximation

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

Figure 39. *The paired-samples t-test of monthly spend before and after the campaign.*

Table 13

Monthly spend before and after the campaign (n = 1500)

Measure	Mean	SD	Median
Monthly_spend_before	282.64	198.64	238.45
Monthly_spend_after	304.35	215.75	255.80
Difference (after – before)	21.71	29.22	13.50

What it means for our example

Spend rose by 21.71 per month on average, t(1499) = 28.78, p < .001, d_z = 0.74 (95% CI [0.69, 0.80]), a medium standardised change. The rank-based alternative agrees: the Wilcoxon signed-rank test gives z = 30.61, p < .001, matched rank-biserial r = .93. The mean difference (21.71) is larger than the median difference (13.50), because some customers rose by much more than others.

Note: a before–after change is not an effect of the campaign

The paired test compares everyone's spend before and after, whichever offer they received. Spend may have risen for other reasons (season, prices, other promotions). It cannot tell you whether the *offer* mattered. To judge the offer, compare the *change* between the two campaign groups (section 4.4).

4.4 Did the New offer raise spend more? The A/B test of a mean

A change score is the difference between two variables of the same customer. Make **Spend_change** = Monthly_spend_after − Monthly_spend_before in card 4 of the **Data** screen (section 1.6; Figure 22). SaNaStat confirms: "Added Spend_change: 1500 of 1500 respondents have a value."

Then open **A/B test** again, choose **An amount per visitor (mean)** and **Spend_change** in **Amount per visitor (scale)**; the versions stay as in section 4.1 (Figure 40, Figure 41).

A/B test: mean Spend_change, New offer vs Standard

What it means: The mean Spend_change is 32.13 for version New offer and 11.61 for version Standard: a difference of +20.52 (+176.7% relative uplift, 95% CI +148.4% to +208.3%). The difference is statistically significant ($p < .001$).

OK **Sample ratio.** The split of visitors (762 / 738) fits the planned 50% / 50% ($\chi^2(1) = 0.38$, $p = .535$).
Rule: Sample ratio mismatch: when the numbers of visitors in the two versions differ from the planned split with $p < .001$, the assignment or the tracking is probably broken and the result should not be trusted (Kohavi et al. (2020)).

OK **Equal variances not needed.** Welch's t test is used, so the versions may have different spreads.
Rule: Compare two means with Welch's t-test by default; it stays accurate when the variances differ (Delacre et al. (2017)).

Info **Planned sample size.** Decide the number of visitors before the test and analyse once it is reached.
Rule: Fix the sample size before the test starts and look at the result once it is reached; stopping as soon as $p < .05$ inflates false wins (Kohavi et al. (2020)).

Figure 40. *The A/B test of the mean change in monthly spend: what it means and the checks.*

Spend_change by version

Version	n	Mean	SD
Standard (A, control)	762	11.61	11.12
New offer (B)	738	32.13	37.34

Figure 41. *Spend_change by version: number of customers, mean and SD.*

Table 14

Change in monthly spend by campaign group (A/B test of means)

Measure	Estimate	95% CI
Standard (A): mean (SD), n	11.61 (11.12), 762	
New offer (B): mean (SD), n	32.13 (37.34), 738	
Difference of means	+20.52	[17.71, 23.33]

Measure	Estimate	95% CI
Relative uplift	+176.7%	[+148.4%, +208.3%]
Welch's t test	t(863.00) = 14.33	p < .001
Cohen's d	0.75	

What it means for our example

Spend rose by 32.13 per month after the New offer and by 11.61 after the Standard offer: the New offer added 20.52 per customer and month (95% CI [17.71, 23.33]), almost three times the rise of the Standard offer (relative uplift +176.7%). Welch's t test is used because it does not assume equal variances (Welch, 1947; Delacre et al., 2017), and the variances clearly differ: the New offer produced a few very large increases (SD 37.34 against 11.12). Cohen's d = 0.75 is a medium standardised difference. Both the response rate and the spend change favour the New offer.

Tip: check that randomisation produced comparable groups

Compare a variable that the offer cannot have changed between the groups. Choose **Annual_spend** as the amount: SaNaStat reports "a difference of −21.83 (−0.6% relative uplift, 95% CI −7.5% to +6.7%). The difference is not statistically significant (p = .859)", with Welch's t(1492.71) = −0.18 and d = −0.01 (means 3380.61 and 3402.45). This is what random assignment should give.

How to report it

"Mean Spend_change was 32.13 (SD = 37.34, n = 738) for New offer and 11.61 (SD = 11.12, n = 762) for Standard; Welch's t(863.00) = 14.33, p < .001, difference = 20.52, 95% CI [17.71, 23.33], d = 0.75."

The same comparison can be expressed in per cent: make **Spend_change_pct** with **Percentage change, (A − B) ÷ B × 100** and the same two variables. SaNaStat reports "Added Spend_change_pct: 1418 of 1500 respondents have a value", because the 82 customers whose spend before was 0 have no percentage.

Note: the relative uplift of a mean

The relative uplift of a mean (B ÷ A − 1) needs positive means; its interval uses the delta method on the logarithmic scale. For a change score whose mean is close to 0 in the control version, the relative uplift can be huge and unstable: report the difference of the means as the main result, as above.

4.5 Typed-in totals: an A/B test without a data file

Often you have only the totals from an e-mail or advertising tool: how many people saw each version and how many converted. The A/B test needs no data file for this. Before you open any file, or in any project, choose **Analyses** → **A/B test**, **A yes/no result per visitor (conversion rate)** and **I will type the totals**, and type the visitors and conversions of each version (Figure 42). With the totals of the campaign (A: 762 visitors, 199 conversions; B: 738 visitors, 327 conversions) the result is the same as in section 4.1, with the versions called A and B.

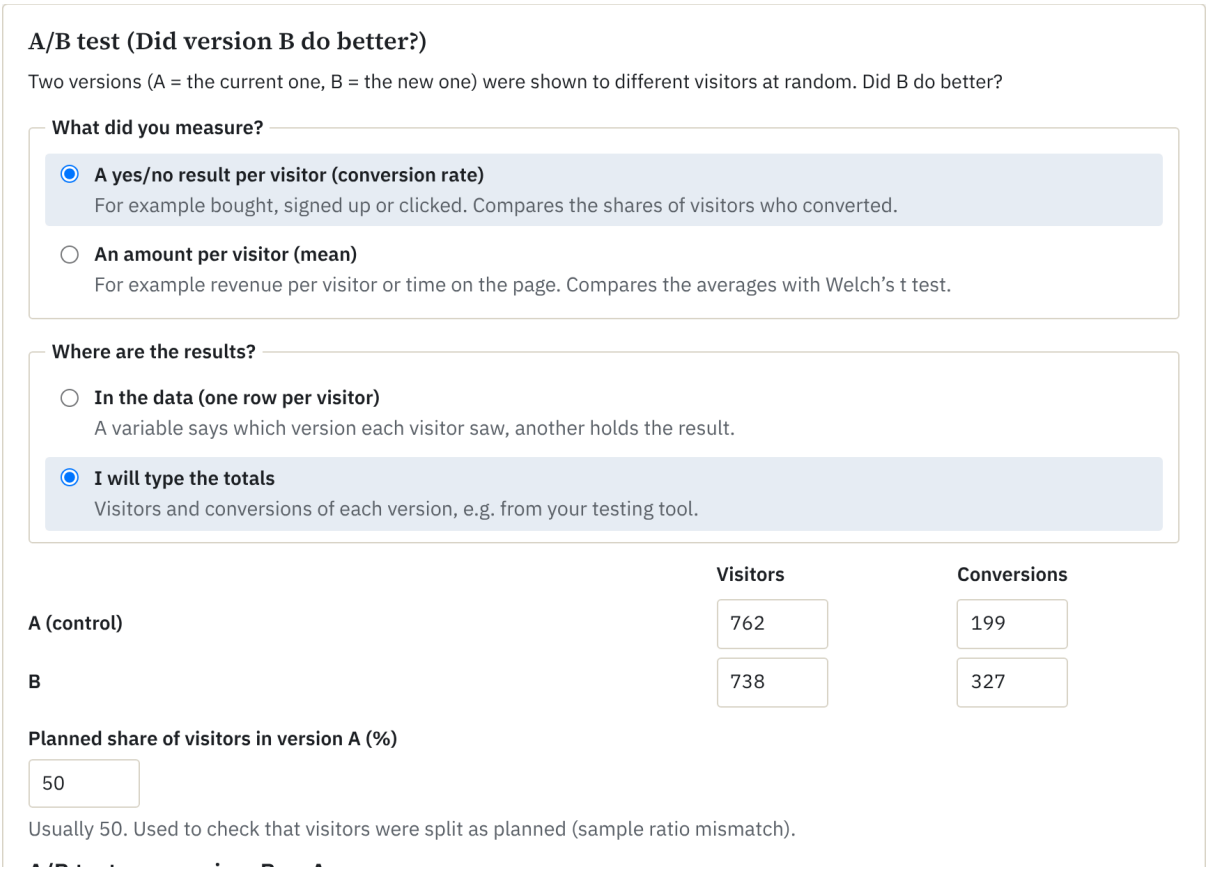


Figure 42. *The A/B test with typed-in totals and no data file: the settings and the totals.*

For a mean, choose **An amount per visitor (mean)** and type the number of visitors, the mean and the standard deviation of each version. The screen starts with an example (A: 200 visitors, mean 50.00, SD 20.00; B: 200 visitors, mean 54.00, SD 22.00), for which SaNaStat reports a +8.0% uplift whose interval (−0.2% to +16.9%) includes 0: "The difference is not statistically significant (p = .058)." Replace the example with your own numbers.

4.6 How many customers does a test need?

A test must be planned before it starts: how many customers are needed to detect the smallest uplift that would matter, with a good chance? The A/B screen has a planner for this. Open **Plan a test: how many visitors do you need?** under the result (Figure 43).

1. For a conversion rate, type the **Current conversion rate (%)**, here 26 (the rate of the Standard offer).
2. Type the **Smallest uplift worth detecting (%)**, as a relative change: 20 means from 26.00% to 31.20%.
3. Keep **Power (%)** at 80 and **Better or worse (two-sided, recommended)**. Optionally type **Visitors per day (optional)**, here 200, for the number of days.

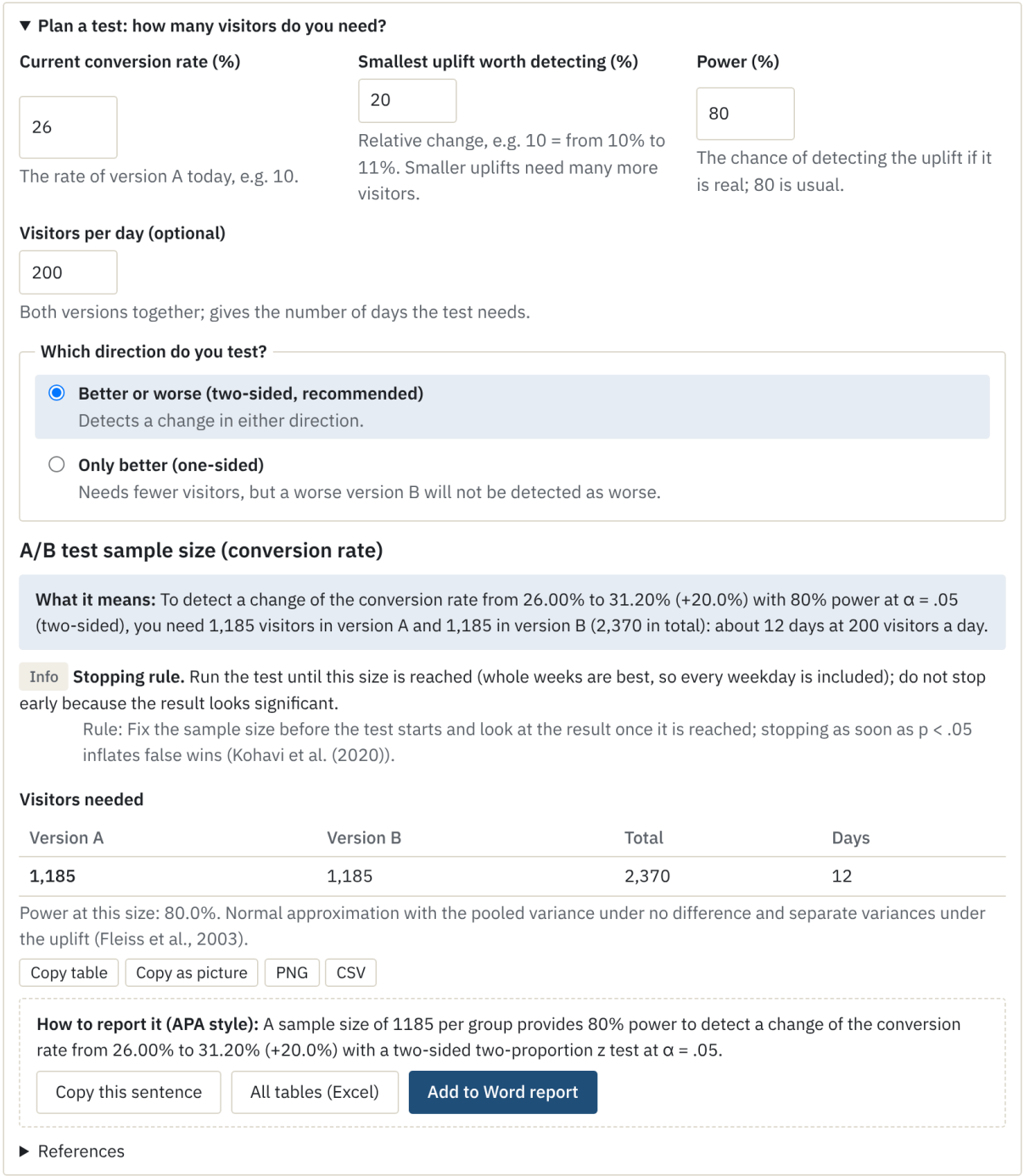


Figure 43. *Planning a test: the sample size for a +20% uplift from a 26% conversion rate.*

Table 15

Visitors needed for a conversion-rate test (from a rate of 26%, power 80%, α = .05)

Smallest uplift worth detecting	Visitors per version (total)	Days at 200 a day
+20% (26.00% to 31.20%), two-sided	1,185 (2,370)	12
+10% (26.00% to 28.60%), two-sided	4,608 (9,216)	47
+10% (26.00% to 28.60%), one-sided	3,630 (7,260)	37

What it means for our example

SaNaStat writes: "To detect a change of the conversion rate from 26.00% to 31.20% (+20.0%) with 80% power at α = .05 (two-sided), you need 1,185 visitors in version A and 1,185 in version B (2,370 in total): about 12 days at 200 visitors a day." Halving the uplift to +10% roughly quadruples the sample (4,608 per version); the calculation uses the normal approximation for two proportions (Fleiss et al., 2003). A one-sided test needs fewer visitors (3,630 per version) but cannot show that version B is worse. For a mean (**An amount per visitor (mean)**), type the **Current mean** and the **Standard deviation** instead: with the change in spend of the Standard offer (mean 11.61, SD 11.12) and a +50% uplift worth detecting (to 17.41, d = 0.52), 59 customers per version are enough.

How to report it

"A sample size of 1185 per group provides 80% power to detect a change of the conversion rate from 26.00% to 31.20% (+20.0%) with a two-sided two-proportion z test at α = .05."

The **Power analysis** screen plans the same kind of study in terms of effect sizes, without data. Choose **Chi-square test (Goodness of fit or independence)** and **A priori power analysis (The sample size I need)**, and enter Cohen's w, α = .05, power = .80 and the degrees of freedom (1 for a 2 × 2 table) (Figure 44).

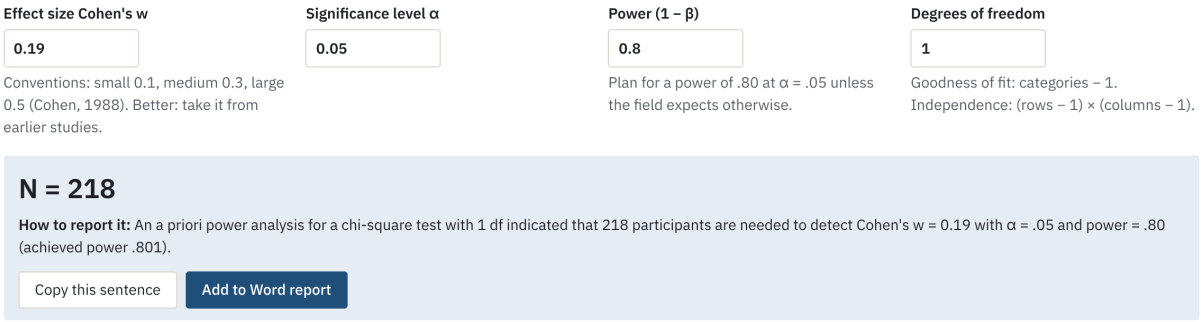


Figure 44. *A priori power analysis for a chi-square test with one degree of freedom (w = 0.19).*

What it means for our example

For a 2 × 2 table, Cohen's w equals φ, which is Cramér's V for a table with one degree of freedom. The A/B test above had V = .19: "An a priori power analysis for a chi-square test with 1 df indicated that 218 participants are needed to detect Cohen's w = 0.19 with α = .05 and power = .80 (achieved power .801)." To detect a small effect (w = 0.10), 785 customers are needed. Choose **Sensitivity power analysis (The smallest effect I can detect)** when the sample size is fixed: with N = 300 customers and power .80, a chi-square test with one degree of freedom can detect effects of w = 0.162 or larger.

Note: where the planning value comes from

Take the uplift worth detecting from the economics of the decision (the smallest uplift that would pay for the change), from a pilot test or from an earlier campaign; conventions such as "a small effect" are a last resort (Cohen, 1988). Plan for missing data, and do not stop the test early because the result looks significant.

Exercises

Exercise: the campaign

1. Repeat the chi-square test for Responded by Loyalty_member. Do members respond more often? (Compare with the logistic regression in section 6.2.)

2. With the sensitivity analysis on the **Power analysis** screen, find the smallest w detectable with N = 1500 and power .80, and compare it with the observed V = .19. Answer: w = 0.072.

3. Planted answer for the chapter: New offer response 44% against 26% for Standard (an uplift of +18.19 percentage points, z = 7.38), and a spend rise after the campaign (paired t(1499) = 28.78) that is larger after the New offer (32.13 against 11.61).

5. What drives customer spend

Which customer characteristics go with higher annual spend, and do groups of customers spend differently? We use practice_marketing.csv and, for section 5.2, the two logarithm variables made in section 2.6 (add them again in card 4 of the **Data** screen if you have reopened the file).

5.1 Multiple linear regression

> **Concept: multiple regression**
>
> A multiple regression predicts a score (here Annual_spend) from several predictors at once. Each coefficient B is the change in the outcome for one more unit of that predictor *with the other predictors held constant*. The standardised coefficient β puts predictors on a common scale, and R^2 is the share of the variance of the outcome that the model explains.

Choose **Analyses** → **Correlation and regression** → **Multiple linear regression (Which variables predict a score?)**, choose **Annual_spend** as the **Outcome (scale)** and tick as **Block 1 predictors** Income_k, Tenure_months, Loyalty_member, Visits_per_month and Satisfaction_1to10 (Figure 45). Categorical predictors, marked "(groups)", are dummy-coded against their first group, so Loyalty_member appears as "Loyalty_member: Yes (vs No)".

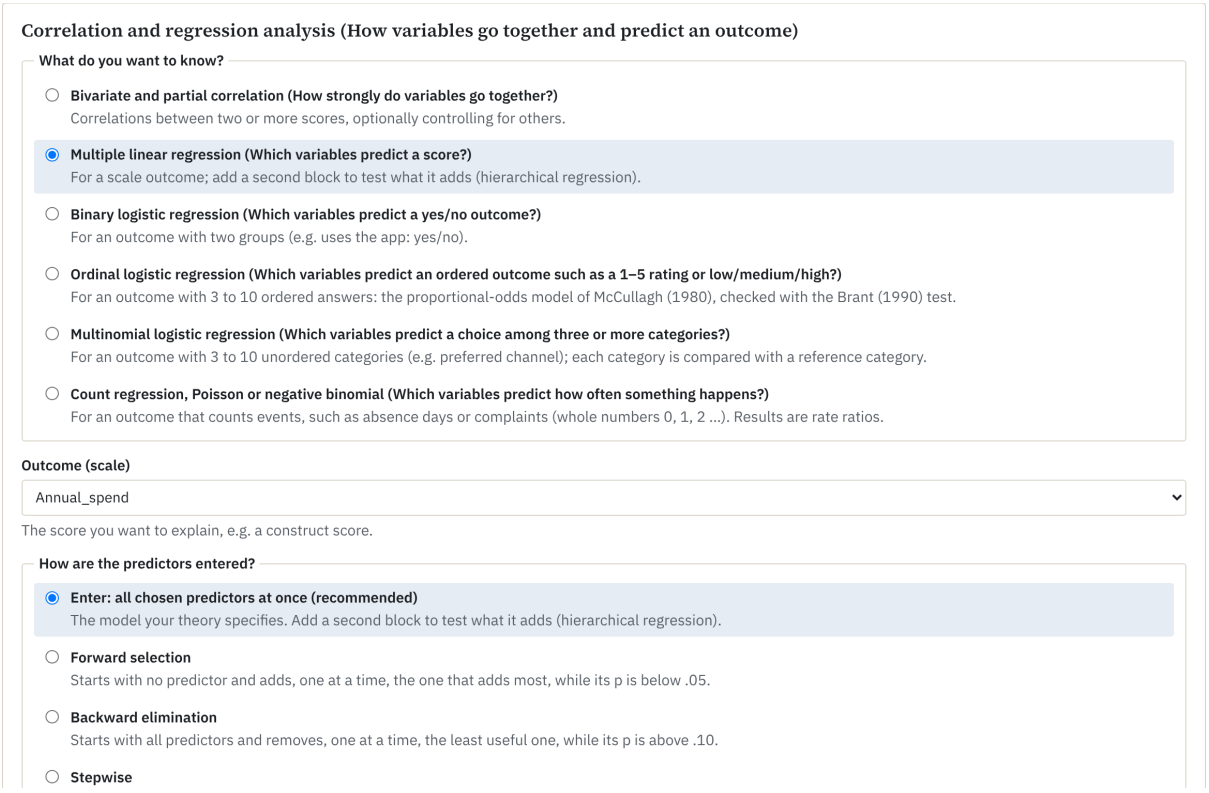


Figure 45. *The regression set-up: the outcome and the predictors in block 1.*

Multiple linear regression (Predicting Annual_spend)

What it means: The predictors together explain 80.6% of the variance in Annual_spend (R^2 = .81, f^2 = 4.15, large). Significant predictors: Income_k (β = .17), Loyalty_member: Yes (vs No) (β = .04), Visits_per_month (β = .86), Satisfaction_1to10 (β = .08).

OK **Sample size.** 1341 respondents for 5 predictor terms (needed: 90).
Rule: At least 50 + 8k respondents for k predictors to test the whole model (Green (1991)).

OK **Collinearity.** Largest VIF = 1.27.
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 = 0.952, p < .001; with this many respondents the coefficient tests are robust.
Rule: Shapiro–Wilk p ≥ .05 for the residuals; with about 300 or more respondents the coefficient tests are robust anyway (Shapiro & Wilk (1965)).

Note **Constant error variance.** Breusch–Pagan (Koenker) $\chi^2(5)$ = 214.22, p < .001.
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 = 0.060; 18 standardised residuals 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 = 4.15 (large).
Rule: f^2: .02 small, .15 medium, .35 large (Cohen (1988)).

Figure 46. *The regression of Annual_spend: what it means and the checks.*

Coefficients (model 1)

Predictor	B	SE	β	t	p	95% CI of B	VIF
(Intercept)	−1738.208	151.929		−11.44	< .001	[−2036.253, −1440.162]	
Income_k	16.071	1.146	.169	14.03	< .001	[13.824, 18.319]	1.00
Tenure_months	−0.456	1.413	−.004	−0.32	.747	[−3.228, 2.316]	1.05
Loyalty_member: Yes (vs No)	217.219	65.995	.045	3.29	.001	[87.754, 346.684]	1.27
Visits_per_month	1057.889	15.835	.860	66.81	< .001	[1026.826, 1088.953]	1.14
Satisfaction_1to10	114.695	19.643	.075	5.84	< .001	[76.160, 153.230]	1.14

n = 1341 (159 respondents with missing values left out). Categorical predictors are dummy-coded against a reference group: Loyalty_member = No. β = standardised coefficient.

Figure 47. *The coefficients of the regression of Annual_spend.*

Table 16

Regression of Annual_spend (n = 1341; 159 customers with a missing value left out)

Predictor	B	β	t	p
Income_k	16.071	.169	14.03	< .001
Tenure_months	−0.456	−.004	−0.32	.747
Loyalty_member: Yes (vs No)	217.219	.045	3.29	.001
Visits_per_month	1057.889	.860	66.81	< .001
Satisfaction_1to10	114.695	.075	5.84	< .001

What it means for our example

The five predictors explain 80.6% of the variance of Annual_spend (R^2 = .81, adjusted R^2 = .805; $F(5, 1335) = 1108.12$, $p < .001$). Visits per month dominate (β = .86: each extra monthly visit goes with 1057.889 more spend per year). Income (β = .17), satisfaction (β = .08) and loyalty membership (β = .04) add a little; tenure adds nothing. Collinearity is not a problem (largest VIF = 1.27). One check needs attention: the Breusch–Pagan test ($\chi^2(5) = 214.22$, $p < .001$) shows that the spread of the errors is not constant, as expected for a skewed money variable, and 18 standardised residuals lie beyond ±3.

Cut-off values: regression checks

Normal residuals: Shapiro–Wilk p ≥ .05; with about 300 or more customers the coefficient tests are robust anyway (Shapiro & Wilk, 1965). **Constant error variance:** Breusch–Pagan (Koenker) p ≥ .05 (Breusch & Pagan, 1979; Koenker, 1981). **Influential cases:** no Cook's distance above 1 (Cook & Weisberg, 1982).

5.2 The same model on the logarithm

Choose **Annual_spend_ln1p** and then **Annual_spend_ln** as the outcome with the same predictors (Figure 48 shows the second).

Coefficients (model 1)

Predictor	B	SE	β	t	p	95% CI of B	VIF
(Intercept)	6.449	0.048		134.09	< .001	[6.354, 6.543]	
Income_k	0.005	0.000	.178	13.28	< .001	[0.004, 0.006]	1.00
Tenure_months	0.000	0.000	.002	0.14	.890	[−0.001, 0.001]	1.05
Loyalty_member: Yes (vs No)	0.074	0.021	.054	3.57	< .001	[0.033, 0.114]	1.25
Visits_per_month	0.309	0.005	.835	58.58	< .001	[0.298, 0.319]	1.13
Satisfaction_1to10	0.026	0.006	.059	4.14	< .001	[0.013, 0.038]	1.14

n = 1268 (232 respondents with missing values left out). Categorical predictors are dummy-coded against a reference group: Loyalty_member = No. β = standardised coefficient.

Figure 48. *The coefficients of the regression of Annual_spend_ln (customers with zero spend left out).*

Table 17

The same five predictors for three versions of the outcome

Outcome	n	R^2	Residual spread (Breusch–Pagan)	Residuals beyond ±3
Annual_spend	1341	.806	$\chi^2(5) = 214.22$, $p < .001$	18

Outcome	n	R^2	Residual spread (Breusch–Pagan)	Residuals beyond ±3
Annual_spend_ln1p (zeros kept)	1341	.415	$\chi^2(5) = 194.05$, $p < .001$	73
Annual_spend_ln (zeros left out)	1268	.773	$\chi^2(5) = 2.97$, $p = .705$	0

What it means for our example

Keeping the zeros (Annual_spend_ln1p) makes the model much worse ($R^2 = .415$; only Visits_per_month is significant, $\beta = .64$), because customers with zero spend are extreme outliers on the logarithmic scale (73 residuals beyond ±3, Shapiro–Wilk W = 0.630). Leaving them out (Annual_spend_ln) gives a well-behaved model: constant error spread ($p = .705$), no extreme residuals and $R^2 = .773$. Its coefficients can be read as percentage changes: each extra monthly visit multiplies spend by about 1.36 (B = 0.309, e^0.309 ≈ 1.36, that is 36% more), loyalty membership by about 1.08 (B = 0.074), and each satisfaction point by about 1.03 (B = 0.026); tenure has no effect (B = 0.000, p = .890).

Note: choose the model that answers your question

The raw model fits well on the whole sample but violates the equal-spread assumption. The logarithmic model on buyers meets the assumptions and describes *how much* buyers spend; it ignores the 82 customers who spent nothing, who need a separate question (who buys at all?), for example with a logistic regression. The low R^2 of the ln(1 + x) version (.415, against .806 for the raw outcome) is caused only by the zero-spend customers.

5.3 Visits and basket are parts of spend, not drivers

In the raw model, visits explain most of the variance. Annual spend is built from visits and basket value, so it is no surprise. Try a hierarchical regression: choose **Annual_spend** as the outcome, put Income_k, Tenure_months, Loyalty_member and Satisfaction_1to10 in **Block 1 predictors** and Visits_per_month and Avg_basket in **Block 2 predictors**.

Table 18

Two-block regression of Annual_spend (n = 1341)

Model	R^2	Adjusted R^2	F (df)	ΔR^2
1: Income_k, Tenure_months, Loyalty_member, Satisfaction_1to10	.157	.154	62.05 (4, 1336)	

Model	R^2	Adjusted R^2	F (df)	ΔR^2
2: add Visits_per_month, Avg_basket	.890	.890	1802.10 (6, 1334)	.734

What it means for our example

Customer characteristics alone explain 15.7% of the variance of annual spend. Adding visits and basket value raises R^2 by .734 ($p < .001$) to .890. In the full model, Visits_per_month ($\beta = .85$, $t(1334) = 87.61$) and Avg_basket ($\beta = .36$, $t(1334) = 32.01$) are the only significant predictors; income ($B = -0.806$, $p = .425$), loyalty membership ($B = -57.331$, $p = .255$) and satisfaction ($p = .924$) no longer matter.

Note: do not adjust away the mechanism

Loyalty members visit more (section 2.1). When visits are in the model, the coefficient of loyalty shows only what loyalty adds *beyond* the extra visits, which is why it shrinks. Whether loyalty *causes* the extra visits cannot be decided from this table. For marketing questions ask first which variables are outcomes of the programme and which are background, and put only background variables in the model when you want the total difference between members and non-members.

5.4 Compare spend between two groups

Choose **Analyses** → **t-tests** → **Independent-samples t-test**, **Annual_spend** as the score and **Loyalty_member** as the groups (Figure 49).

Group descriptives

Group	n	Mean	SD	Median	Skewness	Shapiro–Wilk W	p (normality)
No	854	2637.02	1893.02	2330.50	1.12	.928	< .001
Yes	646	4389.38	2591.22	3949.00	0.79	.958	< .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)	−14.51	1131.47	< .001	−1752.36	[−1989.36, −1515.37]
Student's t (equal variances)	−15.13	1498	< .001	−1752.36	[−1979.50, −1525.22]

Copy table | Copy as picture | PNG | CSV

Effect size

Measure	Value	95% CI	Size
Cohen's d	−0.79	[−0.90, −0.68]	medium
Hedges' g	−0.79	[−0.89, −0.68]	medium

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 49. *The independent-samples t-test of Annual_spend by Loyalty_member: group descriptives, tests and effect sizes.*

Table 19

Annual_spend by Loyalty_member

Group	n	Mean	SD	Median	Skewness
No	854	2637.02	1893.02	2330.50	1.12
Yes	646	4389.38	2591.22	3949.00	0.79

What it means for our example

Members spend 1752.36 more per year on average. Levene's test shows unequal variances ($F(1, 1498) = 70.54$, $p < .001$), so Welch's t-test is read: $t(1131.47) = -14.51$, $p < .001$, 95% CI of the difference [−1989.36, −1515.37], $d = -0.79$ (a medium effect, bordering on large). SaNaStat's effect-size line puts it this way: "$|d| = 0.79$ is medium. The sign of d follows the order of the groups: $d = -0.79$ means Yes scored higher than No." The negative sign therefore says nothing about the direction of the effect by itself; read it from the two means. The rank-based Mann–Whitney test agrees, $z = -14.09$, $p < .001$, rank-biserial $r = -.42$. On the logarithmic scale for buyers (Annual_spend_ln) the result is the same: members 8.24 against 7.76, $t(1371.00) = -14.35$, $p < .001$, $d = -0.76$, and the variances are equal (Levene $p = .104$).

Concept: when to read the rank-based test

SaNaStat shows Welch's t-test and the Mann–Whitney U test side by side. With 30 or more customers per group, tests of means tolerate skewed data (Lumley et al., 2002), so here both agree. With small groups and strong skew, trust the rank-based test, and report medians.

How to report it

"Loyalty members spent more per year (M = 4389.38, SD = 2591.22) than non-members (M = 2637.02, SD = 1893.02), Welch's t(1131.47) = −14.51, p < .001, d = −0.79, 95% CI [−0.90, −0.68]."

5.5 Compare spend between three or more groups

Choose **Analyses** → **One-way ANOVA**, **Annual_spend** as the score and **Main_channel** as the groups (Figure 50, Figure 51).

One-way ANOVA (Comparing Annual_spend across Main_channel)

What it means: Annual_spend differs across the 3 Main_channel groups (small effect, ω² = .039). Pairs that differ: Mobile app vs Online; Mobile app vs Store; Online vs Store.

Note **Equal variances.** Levene's test suggests unequal variances: Welch's ANOVA is used with Games–Howell 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: Mobile app W = .954, p < .001; Online W = .924, p < .001; Store W = .919, 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 = .039 (small).
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 Annual_spend across the three main channels: what it means and the checks.*

Table 20

Annual_spend by Main_channel

Channel	n	Mean	SD
Mobile app	378	4178.99	2521.19
Online	511	2945.69	2130.25
Store	611	3277.66	2382.42

Pairwise comparisons: Games–Howell (chosen automatically: variances differ)

Pair	Mean difference	95% CI	p (adjusted)
Mobile app – Online	1233.30	[856.83, 1609.77]	< .001
Mobile app – Store	901.33	[521.91, 1280.75]	< .001
Online – Store	−331.97	[−648.32, −15.62]	.037

Games–Howell (Games & Howell, 1976) does not assume equal variances; the intervals are simultaneous 95% intervals.

Figure 51. *Games–Howell comparisons of the three channels.*

What it means for our example

Spend differs across channels (small effect, $\omega^2 = .039$). Levene's test shows unequal variances ($F(2, 1497) = 5.20$, $p = .006$), so Welch's ANOVA, $F(2, 899.44) = 29.87$, $p < .001$, and the Games–Howell comparisons are read. All three pairs differ: app customers spend 1233.30 more than online customers (95% CI [856.83, 1609.77]) and 901.33 more than store customers ([521.91, 1280.75]); online customers spend 331.97 less than store customers (adjusted $p = .037$). The Kruskal–Wallis test agrees, $H(2) = 65.95$, $p < .001$, $\varepsilon^2 = .044$. Because Welch's test is used, the sentence for the report gives $\omega^2 = .060$, 90% CI [.038, .088], an approximation computed from Welch's F.

Now compare satisfaction across channels: no difference, $F(2, 1412) = 1.55$, $p = .213$, $\omega^2 = .001$ (Kruskal–Wallis $H(2) = 2.38$, $p = .304$). This matches the overlapping intervals in Figure 14.

Annual spend does not differ between regions ($F(4, 1495) = 0.22$, $p = .926$; Kruskal–Wallis $H(4) = 0.50$, $p = .974$) or genders ($F(2, 1497) = 0.50$, $p = .606$), a useful reminder that not every customer attribute matters. Only 30 customers have Gender "other", so comparisons involving that group are imprecise.

Cut-off values: ANOVA

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

How to report it

"Annual spend differed by main channel, Welch's $F(2, 899.44) = 29.87$, $p < .001$, $\omega^2 = .060$, 90% CI [.038, .088]; customers who mainly use the app spent most ($M = 4178.99$), online customers least ($M = 2945.69$)."

Exercises

Exercise: spend

1. Regress Annual_spend_ln on Income_k and Visits_per_month only. What happens to R^2? (Compare with section 5.2.)

2. Compare Visits_per_month between channels with a one-way ANOVA. Do app customers visit more? Think about what that says about the channel differences in spend.

3. Planted answer for the chapter: Annual_spend rises with income, satisfaction, loyalty membership and above all with visits (section 5.1), and loyalty members spend more (t-test, d = −0.79).

6. Predicting churn and response

Churn means that a customer leaves within the year (Churned = Yes, 250 of 1500 customers, 16.7%). A model that predicts churn lets a firm focus retention effort. SaNaStat offers two routes, and this chapter uses both:

- **Logistic regression** (**Analyses**) explains which characteristics go with a higher chance of churn, with odds ratios and tests.
- **Machine learning** compares several prediction models on customers they have not seen, explains the best one, checks whether its probabilities can be trusted and whether it works equally well for different groups, and gives each customer an honest prediction.

Concept: explanation and prediction

A regression is judged by how well it explains the data it was estimated on. Machine learning asks a different question: how well does a model predict customers it has never seen? If flexible models (random forest, boosting) do not beat a plain logistic regression, the relationships are probably simple, and the regression is the model to report.

6.1 Which variables may be used as predictors?

A prediction is useful only if the predictors are known *before* the outcome. Two kinds of variable must stay out of a churn model:

- **Variables measured over the same year or later than the outcome.** Annual_spend covers the year in which customers left, so a customer who left in March has low annual spend because they left: the predictor partly *contains* the outcome (*leakage*). The two monthly spend variables were measured around the campaign, and the logarithms and Spend_change are made from them.
- **Variables that are results of a campaign you want to evaluate separately**, here Campaign_group and Responded.

The predictors we use for churn are the profile (Age, Gender, Region, Income_k), the relationship (Tenure_months, Loyalty_member, Main_channel), behaviour (Visits_per_month, Avg_basket, Discount_share, Complaints_12m) and the two ratings (Satisfaction_1to10, Recommend_0to10). CustomerID is never offered, because its type is ID (section 1.3).

Note: leakage

A model with leakage looks excellent and fails in use. SaNaStat guards against some forms (a predictor almost identical to the outcome, an ID variable, a construct that contains the outcome), but it cannot know when a variable was measured. That is your responsibility.

6.2 Logistic regression for churn

Choose **Analyses** → **Correlation and regression** → **Binary logistic regression (Which variables predict a yes/no outcome?)**. Choose **Churned** as the outcome, **Yes** as the event (coded 1), and tick as predictors Satisfaction_1to10, Complaints_12m, Tenure_months and Loyalty_member (Figure 52, Figure 53).

Binary logistic regression (Predicting Churned = "Yes")

What it means: The model predicts "Yes" better than chance (Nagelkerke R^2 = .19, AUC = .75, acceptable). Loyalty_member: Yes (vs No): each unit divides the odds of "Yes" by 1.45; Complaints_12m: each unit multiplies the odds of "Yes" by 1.39; Satisfaction_1to10: each unit divides the odds of "Yes" by 1.64.

OK **Events per variable.** EPV = 58.5.
Rule: At least 10 respondents in the smaller outcome group per coefficient (EPV ≥ 10) (Peduzzi et al. (1996)).

OK **Calibration.** Hosmer–Lemeshow $\chi^2(8) = 10.68$, p = .221.
Rule: Hosmer–Lemeshow p ≥ .05: no evidence of poor fit (Hosmer et al. (2013)).

OK **Discrimination.** AUC = .754 (acceptable).
Rule: AUC .7–.8 acceptable, .8–.9 excellent, ≥ .9 outstanding (Hosmer et al. (2013)).

Figure 52. *The logistic regression of Churned: 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
169.84	4	< .001	.191	.134	.125	.754	[.72, .79]

Copy table | Copy as picture | PNG | CSV

Coefficients

Predictor	B	SE	Wald	p	Odds ratio	95% CI of OR
(Intercept)	1.493	0.362	17.03	< .001	4.450	[2.190, 9.043]
Tenure_months	−0.008	0.004	3.72	.054	0.992	[0.983, 1.000]
Loyalty_member: Yes (vs No)	−0.374	0.172	4.73	.030	0.688	[0.491, 0.964]
Complaints_12m	0.329	0.073	20.52	< .001	1.390	[1.205, 1.603]
Satisfaction_1to10	−0.498	0.057	75.33	< .001	0.608	[0.543, 0.680]

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

Figure 53. *The model fit and the coefficients with odds ratios.*

Concept: odds ratio

An odds ratio (OR) is the factor by which the odds of the event are multiplied when the predictor increases by one unit (or when a customer belongs to the group rather than the reference group). An OR of 1 means no effect; above 1 the event becomes more likely; below 1 less likely. The odds of an event are its probability divided by one minus its probability.

Table 21

Logistic regression of Churned (N = 1415 customers with a satisfaction rating)

Predictor	Odds ratio	p	95% CI of OR
Tenure_months (per month)	0.992	.054	[0.983, 1.000]
Loyalty_member: Yes (vs No)	0.688	.030	[0.491, 0.964]
Complaints_12m (per complaint)	1.390	< .001	[1.205, 1.603]
Satisfaction_1to10 (per point)	0.608	< .001	[0.543, 0.680]

What it means for our example

Each additional complaint multiplies the odds of leaving by 1.39, each additional satisfaction point multiplies them by 0.61 (the odds fall by about 39%; SaNaStat says "divides the odds of "Yes" by 1.64"), and loyalty members have 0.69 times the odds of non-members. Longer tenure points the same way but its interval just reaches 1 ($p = .054$). The model beats chance ($\chi^2(4, N = 1415) = 169.84$, $p < .001$; AUC = .754, "acceptable"; Nagelkerke $R^2 = .19$). The checks are fine: 58.5 events per variable (EPV, at least 10 needed; Peduzzi et al., 1996) and the Hosmer–Lemeshow test shows no poor fit, $\chi^2(8) = 10.68$, $p = .221$.

Cut-off values: logistic regression

EPV ≥ 10 (Peduzzi et al., 1996). **AUC** .70–.80 acceptable, .80–.90 excellent, ≥ .90 outstanding (Hosmer et al., 2013). **Hosmer–Lemeshow** p ≥ .05: no evidence of poor fit.

Adding Age, Region, Main_channel, Visits_per_month and Discount_share (nine predictors, 13 coefficients) changes little: Nagelkerke $R^2 = .21$, AUC = .766. Complaints (OR = 1.40) and satisfaction (OR = 0.61) keep their effects; age, channel, visits and discount share do not matter; Region does: relative to Central, the odds of churn are lower in the East (OR = 0.59), North (0.55), South (0.55) and West (0.48), all $p < .05$.

The accuracy trap

The classification table at the bottom of the result shows what the model does at a 0.5 cut-off (a customer is predicted to churn when the predicted probability is at least .50) (Figure 54):

Classification at a 0.5 cut-off

Observed	Predicted No	Predicted Yes	% correct
No	1159	22	98.1
Yes	216	18	7.7
Overall			83.2

A respondent is predicted "Yes" when the predicted probability is at least .50. Guessing the larger group for everyone would be 83.5% correct.

Figure 54. *Classification at a 0.5 cut-off.*

Table 22

Classification of churn at a 0.5 cut-off (model with four predictors)

Observed	Predicted No	Predicted Yes	% correct
No	1159	22	98.1
Yes	216	18	7.7

Overall, 83.2% are classified correctly, and SaNaStat adds: "Guessing the larger group for everyone would be 83.5% correct." **The model is less accurate than a model that predicts "no churn" for everyone.** It finds only 18 of the 234 customers who left.

Concept: precision, recall and why accuracy misleads

For a rare outcome, the share of correct predictions is dominated by the common group. Use instead:

- **Recall (sensitivity):** of the customers who really left, the share the model found.
- **Precision:** of the customers the model flagged, the share who really left.
- **Specificity:** of the customers who stayed, the share the model left alone.

A single cut-off trades recall against precision: lowering the cut-off finds more leavers but also flags more customers who would have stayed. The area under the curve (AUC) summarises ranking quality over all cut-offs and does not depend on the cut-off. Section 6.5 shows the three measures at every cut-off.

6.3 Machine learning: set up the prediction

Click **Machine learning** in the menu. The screen has three set-up cards and, after the run, the results on six tabs.

1. In *1 · Prediction task*, choose **Churned (yes/no: No / Yes)** as the **Outcome to predict**. SaNaStat sets the task (binary classification) and ticks every other variable as a predictor. **Which answer is the event (coded 1)?** is **Yes**.
2. Under **Predictors**, click **Select none** and tick the 13 predictors of section 6.1 (Figure 55). Under the list SaNaStat notes: "Not offered: CustomerID (type ID: a respondent number would let a model memorise respondents)."
3. Read the checks under the predictors: SaNaStat notes "The answers are unbalanced (17% vs 83%). Read AUC and balanced accuracy rather than plain accuracy." and states that missing values, standardising and tuning are learned from the training customers of each fold only (nested cross-validation).
4. In *2 · Validation plan* choose **5-fold cross-validation, no repeats (Quick look)** for exploring (Figure 56); use the recommended repeated cross-validation for results you report. The **Main measure** is AUC for yes/no outcomes, and the **Random seed** (2026) makes every split reproducible.

5. In *3 · Models* keep the default models and click **Run the comparison**. A progress bar shows the steps.

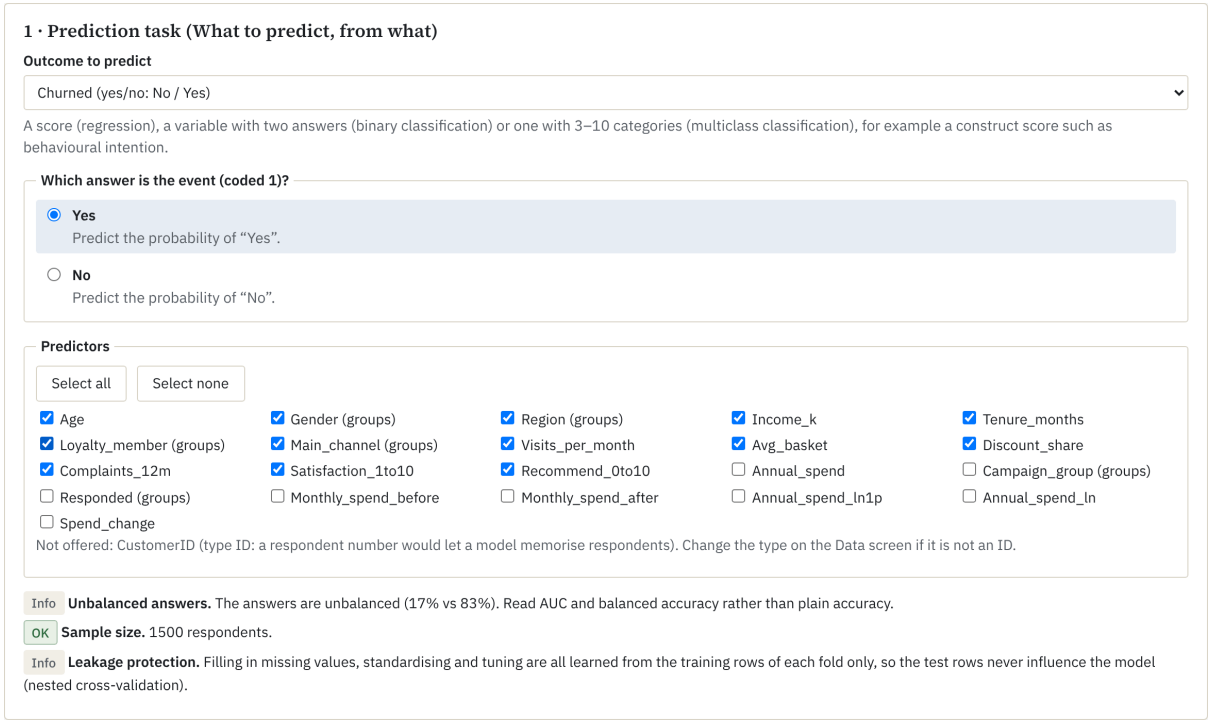


Figure 55. *1 · Prediction task: the outcome, the 13 predictors and the checks.*

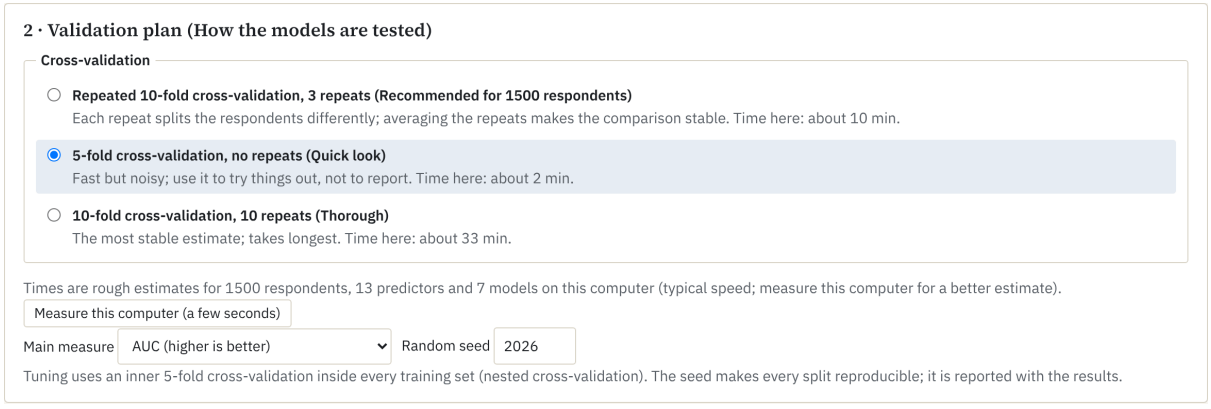


Figure 56. *2 · Validation plan, with a rough time estimate for each choice.*

The plan card estimates the time on your computer. The quick look on 1500 customers took 446 seconds on the computer used for this manual, which was running another study at the same time; speed depends on your computer, the results do not.

Concept: cross-validation

The customers are split into k parts (*folds*). Each model is trained on k − 1 parts and tested on the remaining one, k times over, so every customer is predicted once by a model that did not see them. SaNaStat keeps the share of churners the same in every fold (stratified folds) and tunes each model by an inner cross-validation inside every training set (nested cross-validation), so that the test customers never influence the model (Kuhn & Johnson, 2013).

6.4 Results · Model comparison

The results open on the tab **Comparison** (Figure 57); Table 23 gives the table of out-of-sample performance.

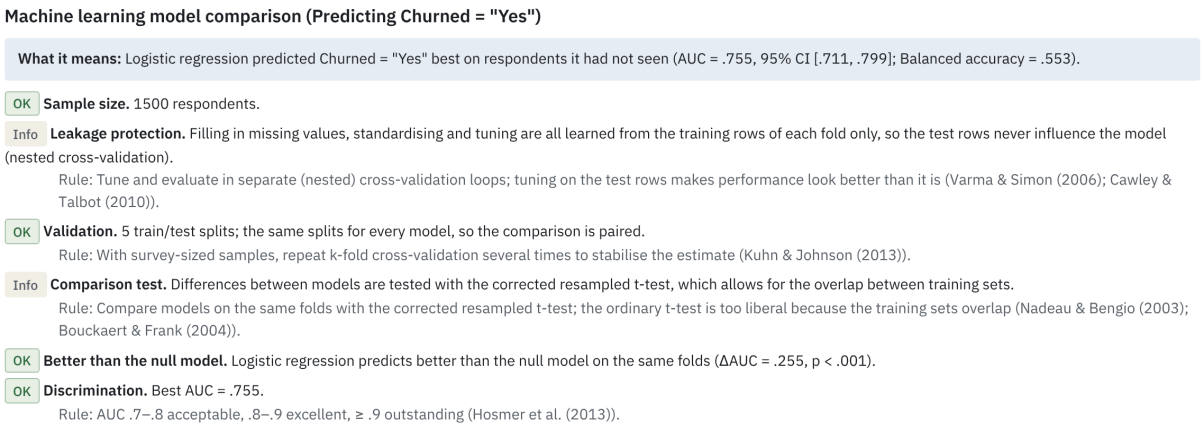


Figure 57. *The model comparison: what it means and the checks.*

Table 23

Out-of-sample prediction of Churned (5-fold cross-validation)

Model	AUC M (SD)	95% CI	Accuracy	Balanced accuracy	F1	MCC	Brier	Δ vs best (p)
Logistic regression	.755 (.024)	[.711, .799]	.833	.553	.207	.188	0.122	best
Elastic net	.754 (.019)	[.719, .789]	.833	.517	.076	.081	0.124	.000 (.959)
XGBoost	.745 (.015)	[.718, .772]	.835	.523	.100	.143	0.123	−.010 (.477)
LightGBM	.743 (.023)	[.701, .786]	.831	.528	.122	.127	0.124	−.011 (.442)
Random forest	.722 (.034)	[.659, .785]	.837	.538	.149	.173	0.125	−.033 (.061)
Decision tree (CART)	.691 (.023)	[.648, .733]	.821	.552	.215	.163	0.134	−.064 (.018)
Null model	.500 (.000)	[.500, .500]	.833	.500	.000	.000	0.139	−.255 (< .001)

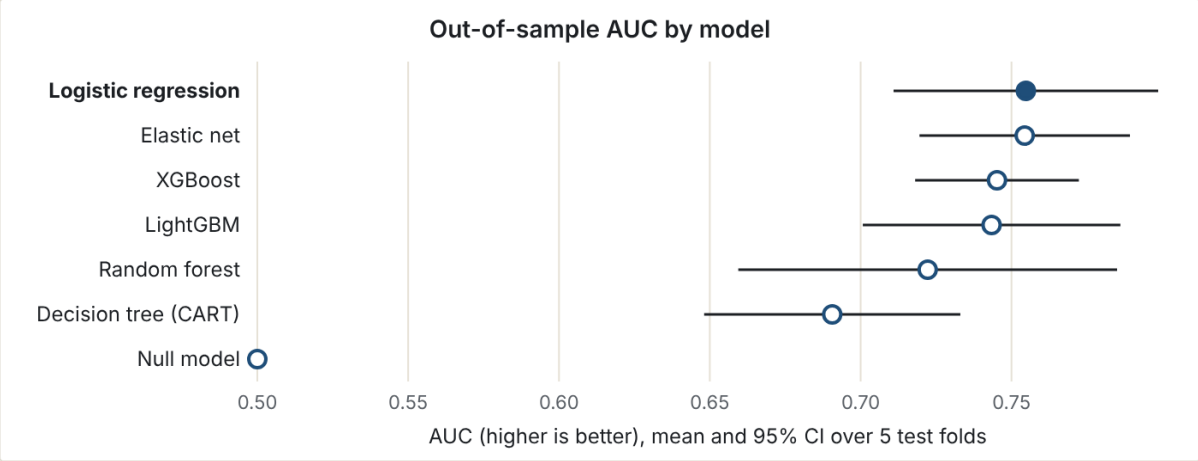


Figure 58. *Out-of-sample AUC of each model (dots: mean over the test folds; lines: 95% confidence intervals).*

What it means for our example

Logistic regression predicts churn best on customers it has not seen (AUC = .755, 95% CI [.711, .799]), and no flexible model beats it: elastic net, XGBoost and LightGBM are statistically indistinguishable (p = .959, .477 and .442), random forest is not significantly worse (p = .061) and the single decision tree is significantly worse (p = .018). Churn in this file is therefore explained by simple, additive relationships, and the logistic regression of section 6.2 is the model to report. Only the null model, which predicts the average for everyone, is clearly worse (ΔAUC = .255, p < .001).

Now look at the accuracy column: every model, including the null model, is about 83% accurate (.821 to .837), because 83% of customers stay. Accuracy does not separate the models at all. The AUC does.

Table 24

Measures reported for a yes/no outcome

Measure	What it tells you
AUC	The chance that a randomly chosen customer who left gets a higher predicted probability than a randomly chosen customer who stayed. 0.5 is chance, 1 is perfect. Independent of the cut-off; the main measure.
Accuracy	The share of customers classified correctly at the 0.5 cut-off. Misleading for rare outcomes.
Balanced accuracy	The average of recall and specificity: .50 for a model that predicts one group for everyone.
F1	The harmonic mean of precision and recall for the churners.
MCC	Matthews correlation coefficient: a correlation between prediction and truth that uses all four cells of the classification table (0 = chance, 1 = perfect).
Brier score	The mean squared error of the predicted probabilities (lower is better); the null model sets the benchmark (0.139).

Differences between models are tested with the corrected resampled t-test, which allows for the overlap between training sets (Nadeau & Bengio, 2003). The table *Models and tuning* under the comparison shows what was tuned and which settings were chosen.

Note: small samples and flexible models

Flexible models need many more customers than linear models before they pay off (van der Ploeg et al., 2014). With a few hundred or a few thousand customers, "no better than logistic regression" is a common and useful result. The Quick look is noisy; use the repeated cross-validation for the results you report.

How to report it

"Seven prediction models were compared with 5-fold cross-validation (5 train/test splits), with every model tuned by an inner 5-fold cross-validation inside each training set (nested cross-validation; seed 2026); missing values were imputed and predictors standardised within each training fold. Logistic regression performed best (AUC = .755, 95% CI [.711, .799]; Balanced accuracy = .553)."

6.5 Results · Cut-off: precision and recall

The prediction models give each customer a *probability* of churn. A decision needs a *cut-off*: contact every customer whose probability is above it. The best cut-off depends on costs, not on statistics (Fawcett, 2006).

1. Estimate the **cost of contacting** a customer who would have stayed (an offer, a call, a discount given away).
2. Estimate the **value of keeping** a customer who would have left, multiplied by the chance that the retention action works.
3. Contact a customer when the probability of churn times the value kept exceeds the cost of contacting. In symbols: contact when p × V > C, that is, when the churn probability p is above C ÷ V. A cut-off of 0.20, for example, is the right one when the value kept is five times the cost of a contact.

For a yes/no outcome the tab **Cut-off** shows *Results · Cut-offs, precision and recall (What happens when probabilities become yes or no?)*. It evaluates the model chosen on the Explanations tab (here the logistic regression) on predictions for customers the model did not see. Type a number between 0 and 1 in **Cut-off (probability, between 0 and 1)**; the box starts at 0.5 (Figure 59). The cut-off is also used on the Fairness tab and in the predictions file.

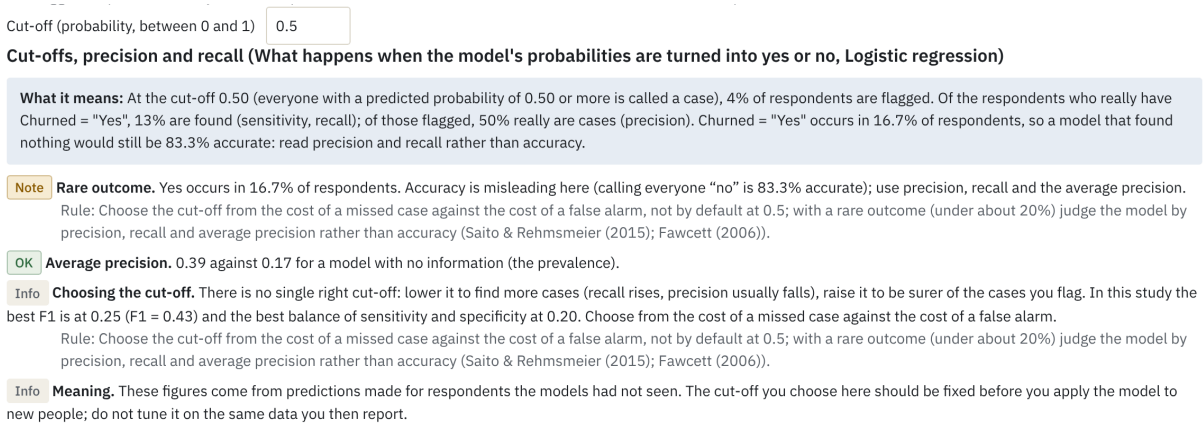


Figure 59. *Cut-offs, precision and recall at the cut-off 0.5: what it means and the checks.*

What it means for our example: the cut-off 0.5

SaNaStat writes: "At the cut-off 0.50 (everyone with a predicted probability of 0.50 or more is called a case), 4% of respondents are flagged. Of the respondents who really have Churned = "Yes", 13% are found (sensitivity, recall); of those flagged, 50% really are cases (precision)." The count table shows the same in customers: of the 250 customers who left, 33 are found and 217 are missed; of the 66 customers flagged, 33 really leave and 33 do not. The checks add that the outcome is rare (Yes occurs in 16.7% of customers, so calling everyone "no" is 83.3% accurate) and that the average precision is 0.39 against 0.17 for a model with no information.

Now type **0.2** (Figure 60). The model flags more customers and finds more of the leavers.

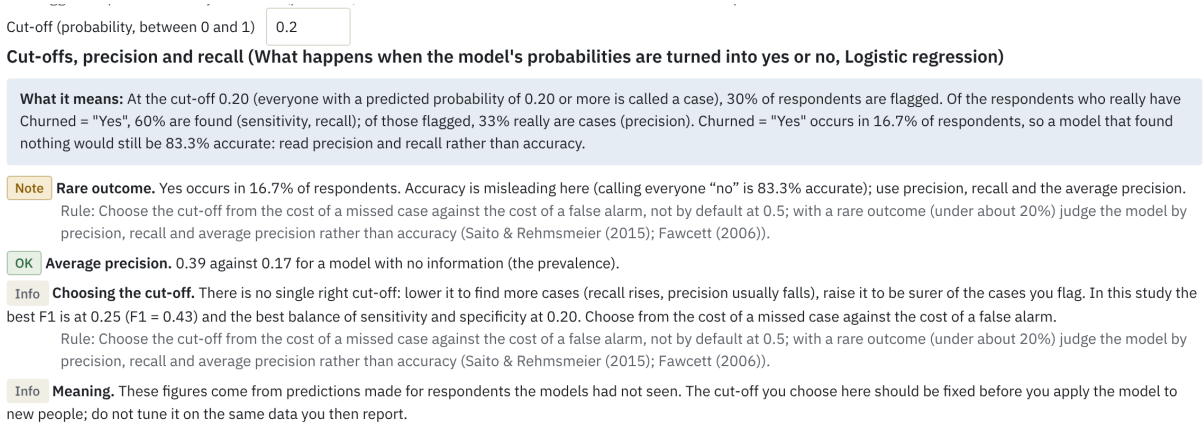


Figure 60. *The same section at the cut-off 0.2.*

Table 25

What the cut-off does (logistic regression, Churned, 5-fold cross-validation)

Cut-off	Flagged	Recall	Precision	Specificity	F1	Accuracy
0.10	58%	0.86	0.25	0.48	0.38	0.54
0.20	30%	0.60	0.33	0.76	0.43	0.73
0.30	15%	0.38	0.41	0.89	0.40	0.81

Cut-off	Flagged	Recall	Precision	Specificity	F1	Accuracy
0.50	4%	0.13	0.50	0.97	0.21	0.83

What it means for our example: the cut-off 0.2

At 0.20, 30% of the customers are flagged (449 of 1500). Of the 250 customers who left, 150 are found (recall 0.60) and 100 are missed; of the 449 flagged customers 150 really leave (precision 0.33) and 299 do not. Compared with the cut-off 0.5, the program finds 150 instead of 33 of the leavers, and the price is 299 instead of 33 false alarms. The accuracy falls from .83 to .73, which shows once more why accuracy is the wrong measure for a rare outcome. SaNaStat's own hint is that the best F1 is at 0.25 (F1 = 0.43) and the best balance of sensitivity and specificity at 0.20, but "There is no single right cut-off".

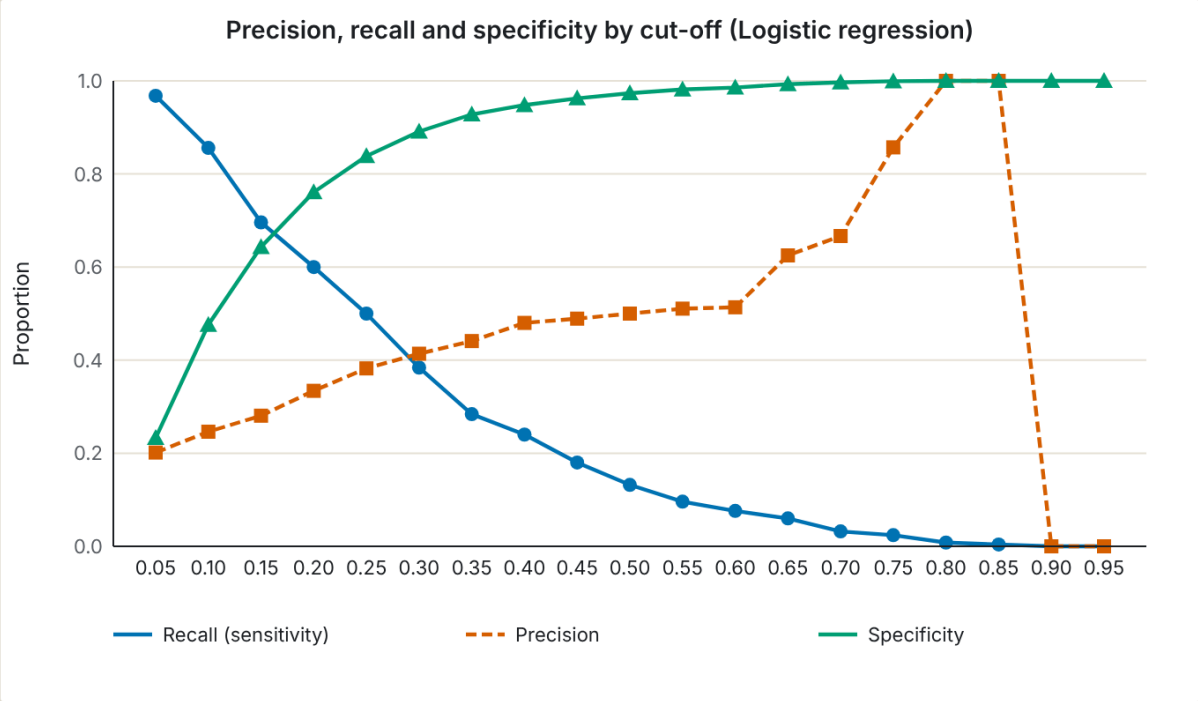


Figure 61. *Precision, recall and specificity of the logistic regression at every cut-off.*

Concept: precision and recall move in opposite directions

A high cut-off flags few customers, mostly true leavers (high precision, low recall); a low cut-off flags many customers and finds most leavers, but most flagged customers would have stayed (low precision, high recall). In Figure 61 the recall curve falls and the specificity curve rises as the cut-off grows, and the precision curve rises with it. Read the far right of the chart with care: at cut-offs of 0.80 and above almost nobody is flagged, so a precision of 1.00 there rests on a handful of customers, and SaNaStat draws an undefined precision (nobody flagged) as 0. A model with no information has a precision equal to the prevalence (0.17); the average precision summarises the whole precision–recall curve (Saito & Rehmsmeier, 2015; He & Garcia, 2009).

Note: choose the cut-off before you look at the result you report

SaNaStat 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". Fix the cut-off from your costs (the formula above). The recall and precision of a rare outcome are also noisy: the Quick look uses one 5-fold cross-validation, and the recommended repeated cross-validation gives more stable figures.

6.6 Results · Explanations: what drives churn?

On the tab **Explanations**, **Explain this model** chooses the model to explain. For churn it starts with the logistic regression, marked "recommended: simplest not significantly worse than the best — also the best".

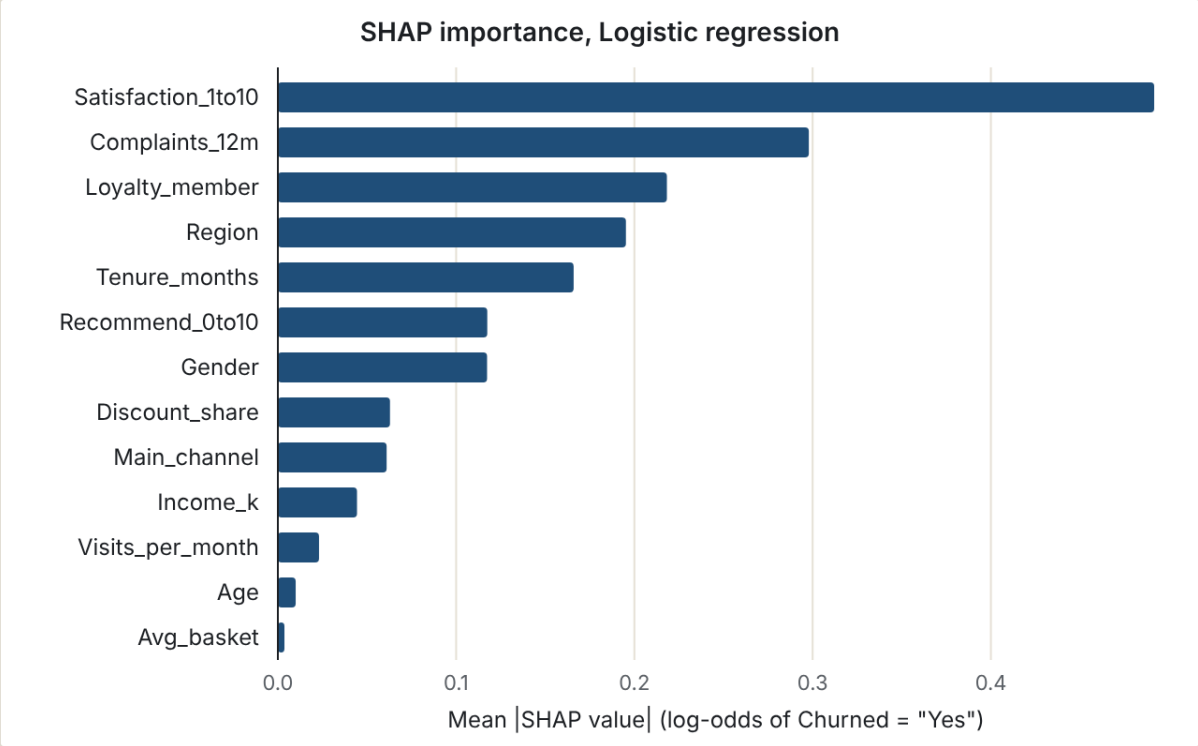


Figure 62. *SHAP importance of each predictor.*

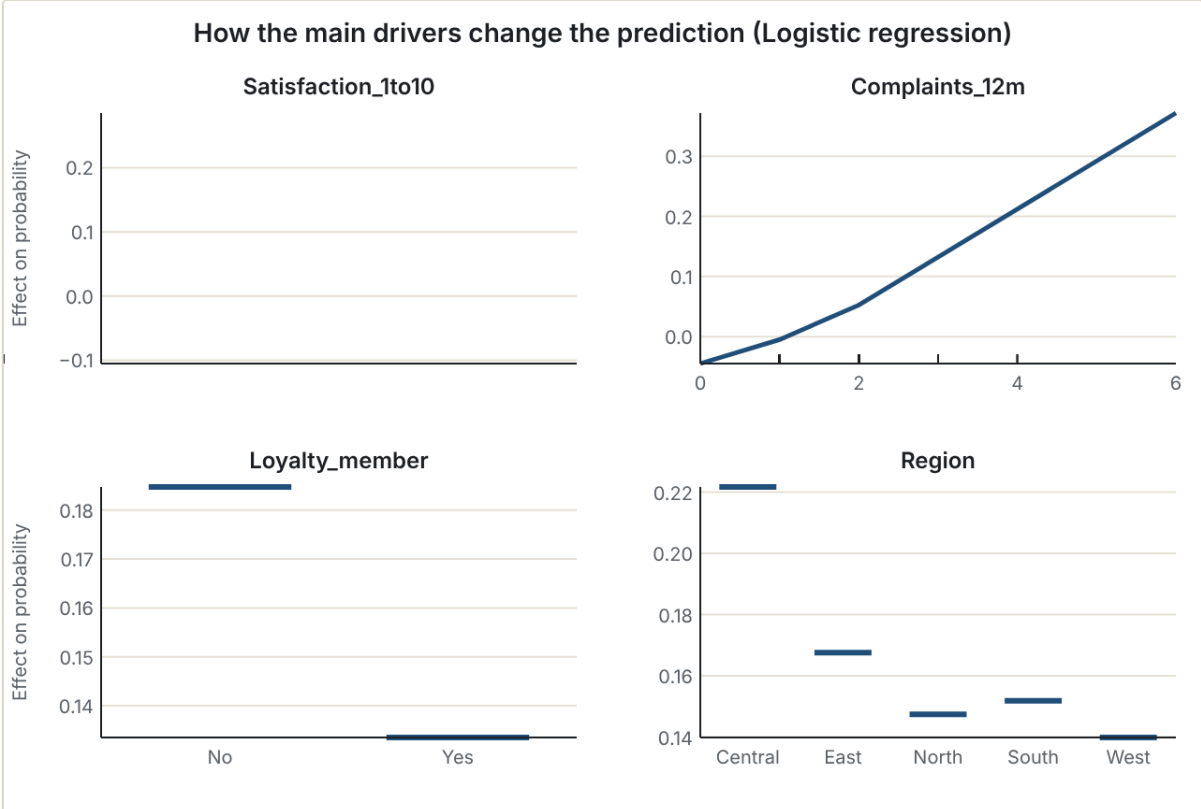


Figure 63. *Accumulated local effects of the main drivers.*

Table 26

The strongest drivers of churn in the logistic regression (importance on held-out customers)

Predictor	Mean abs. SHAP	Direction	Permutation importance M (SD)	Rank (SHAP / permutation)
Satisfaction_1to10	0.492	higher satisfaction, lower chance of churn	0.096 (0.009)	1 / 1
Complaints_12m	0.298	more complaints, higher chance	0.040 (0.018)	2 / 2
Loyalty_member	0.218	depends on the category	0.009 (0.007)	3 / 4
Region	0.195	depends on the category	0.010 (0.009)	4 / 3
Tenure_months	0.166	longer tenure, lower chance	0.008 (0.008)	5 / 5

Concept: SHAP, permutation importance and ALE

SHAP values split each customer's prediction into the contributions of the predictors; the mean absolute value shows how much a predictor moves predictions on average (here in log-odds of churn; Lundberg & Lee, 2017). *Permutation importance* shuffles one predictor in the held-out customers and measures the loss in AUC. *Accumulated local effects* (ALE) show how the predicted probability changes across the range of one predictor while the other predictors stay realistic. These tools describe what the model learned, **not causal effects** (Molnar, 2022). Correlated predictors share credit, and ALE is preferred to partial dependence when predictors are correlated (Apley & Zhu, 2020).

What it means for our example

Satisfaction is the main driver (mean |SHAP| = 0.492), followed by complaints (0.298), loyalty membership (0.218), region (0.195) and tenure (0.166). The two importance methods agree (Spearman's ρ = .97); satisfaction loses 0.096 AUC when shuffled and complaints 0.040. The ALE table shows the shape: the predicted churn probability is 0.299 above average at a satisfaction of 2.30 and 0.117 below average at 10.00, a range of about 0.42, and with complaints it rises from 0.045 below average at 0 complaints to 0.372 above average at 6. Loyalty members have a lower average predicted churn probability (0.13 against 0.18), and the East, North, South and West regions lower than Central (0.14 to 0.17 against 0.22). Main channel, age and average basket add nothing: shuffling them did not hurt the predictions. Satisfaction ranks first in all six models and complaints second or third, which makes these two drivers credible; the rank of loyalty membership and region varies between models.

6.7 Results · Calibration: can the probabilities be trusted?

On the tab **Calibration & intervals**, SaNaStat checks whether the predicted probabilities mean what they say (Figure 64).

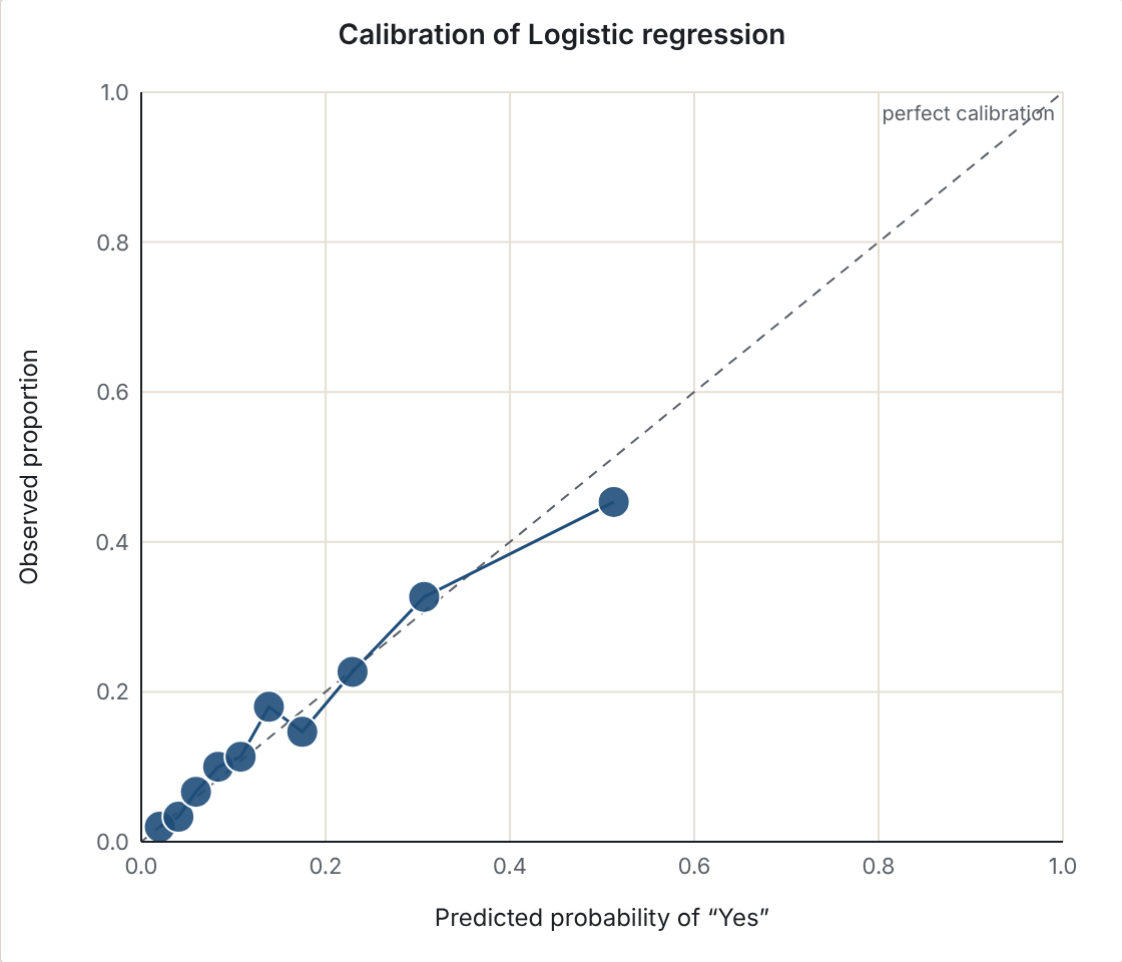


Figure 64. *Reliability diagram: predicted probability against the observed proportion of churners.*

What it means for our example

The calibration slope is 0.89 (95% CI [0.75, 1.04]) and calibration-in-the-large is −0.01 (95% CI [−0.15, 0.14]): the predictions are about as spread out as they should be and there is no systematic over- or under-prediction. In the reliability table, customers in the tenth with the lowest predicted risk (mean .020) churned at .020, and customers in the tenth with the highest predicted risk (mean .513) churned at .453, against the overall rate of 16.7%. A predicted 30% can be read as roughly a 30% chance, which is what a cost-based cut-off needs.

Cut-off values: calibration

Calibration slope close to 1 (below 1: predictions too extreme, a sign of overfitting; above 1: too timid) and calibration-in-the-large close to 0 (Van Calster et al., 2019).

6.8 Results · Fairness by group

A model can work better for some groups of customers than for others. 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 Region in **Group by** (Figure 65).

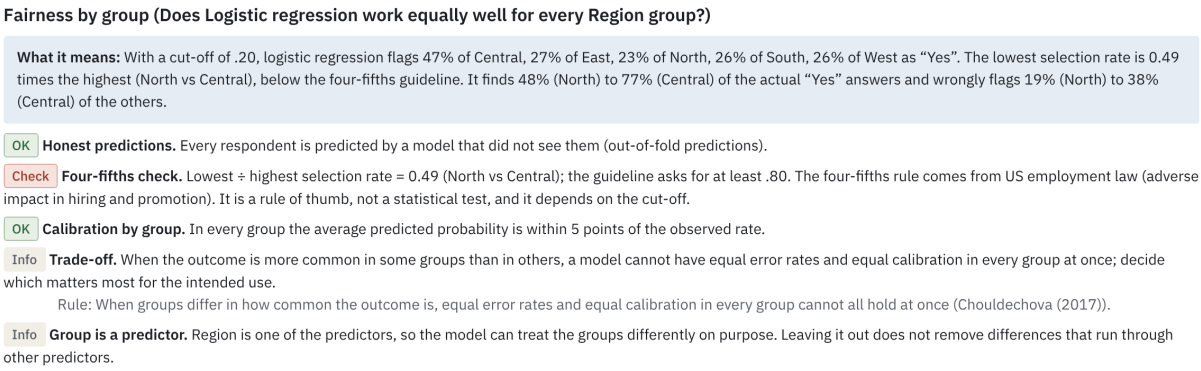


Figure 65. *Fairness of the logistic regression by Region at the cut-off 0.20: what it means and the checks.*

Table 27

Out-of-fold predictions of the logistic regression by Region (cut-off .20; in brackets the ratio to the best group)

Region	n	Observed rate	Selection rate	True-positive rate	False-positive rate	Precision
Central	322	.23	.47 (1.00)	.77 (1.00)	.38 (0.50)	.38 (1.00)
East	259	.15	.27 (0.57)	.55 (0.71)	.21 (0.88)	.32 (0.83)
North	308	.14	.23 (0.49)	.48 (0.62)	.19 (1.00)	.30 (0.77)
South	326	.15	.26 (0.56)	.55 (0.71)	.21 (0.90)	.32 (0.83)
West	285	.15	.26 (0.55)	.52 (0.68)	.21 (0.90)	.30 (0.78)

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 47% of the customers in Central and 23% to 27% elsewhere; the lowest selection rate is 0.49 times the highest (North against Central), and SaNaStat marks the four-fifths check *Check*. This follows the data: churn is more common in Central (an observed rate of .23 against .14 to .15), and Region is one of the predictors. The probabilities are right on average in every region (observed – predicted 0.00), but 38% of the Central customers who stayed are flagged wrongly, against 19% to 21% elsewhere. By Gender the model flags 80% of the 30 customers with Gender "other" (their observed churn rate is .40), a group too small for firm conclusions. By Main_channel the selection rates are close (.26 to .33; ratio 0.79).

Note: fairness in marketing

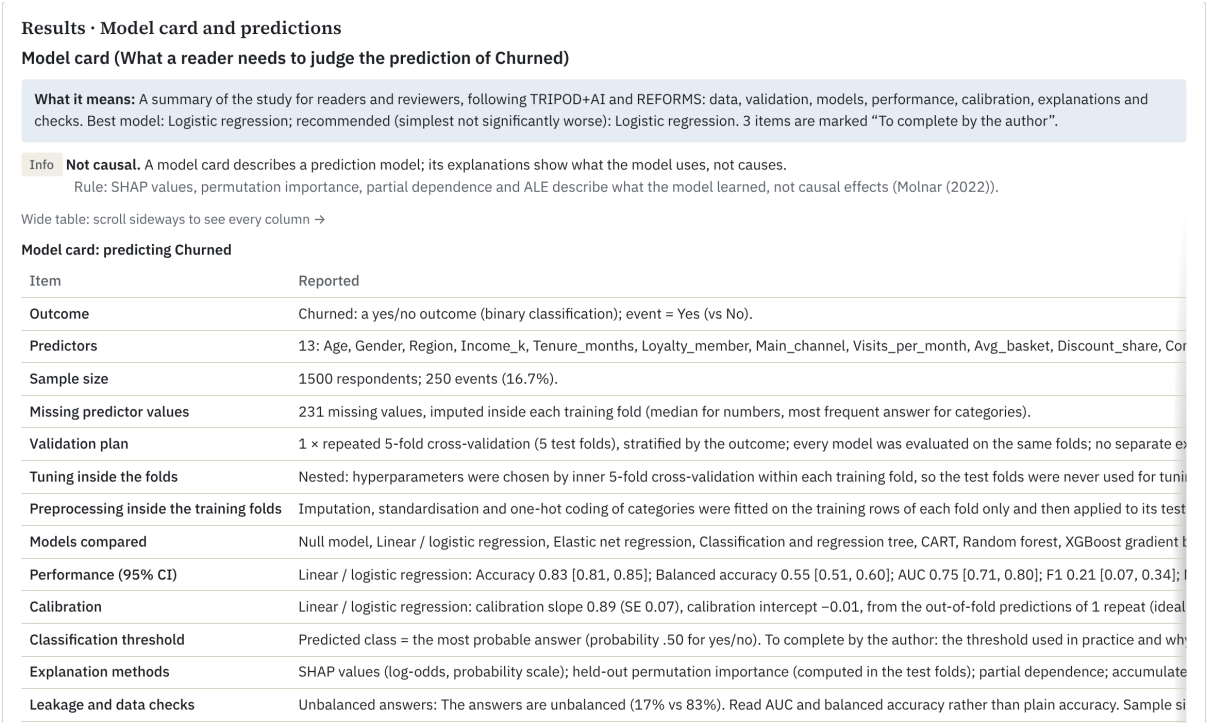
The four-fifths rule comes from employment law and is a rule of thumb, not a test; SaNaStat uses it as a signal. In marketing, a retention offer that reaches one region more often than another may be exactly what the business wants, while an offer that excludes customers by gender or age may be unfair or unlawful. Decide which variables may be used for targeting before you build the model, and check the result by group before you use it.

How to report it

"Group fairness of the logistic regression model was examined with its out-of-fold predictions across the 5 groups of Region (cut-off .20). Selection rates ranged from .23 (North) to .47 (Central; ratio 0.49), true-positive rates from .48 to .77, and false-positive rates from .19 to .38."

6.9 Results · Model card and predictions

The tab **Model card** summarises the study for readers and reviewers (Figure 66), following the TRIPOD+AI and REFORMS recommendations (Collins et al., 2024; Kapoor et al., 2024): the outcome, predictors, sample, missing values, validation, tuning, performance with intervals, calibration, the explanation methods, the checks and the seed. Three items are marked "To complete by the author": an external validation sample, the threshold used in practice and why, and the intended use and population.



Results · Model card and predictions

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

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: Logistic regression; 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)).

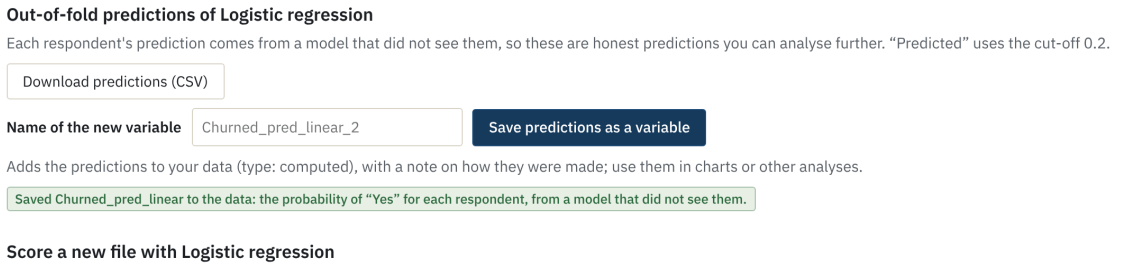
Wide table: scroll sideways to see every column →

Model card: predicting Churned

Item	Reported
Outcome	Churned: a yes/no outcome (binary classification); event = Yes (vs No).
Predictors	13: Age, Gender, Region, Income_k, Tenure_months, Loyalty_member, Main_channel, Visits_per_month, Avg_basket, Discount_share, Cor
Sample size	1500 respondents; 250 events (16.7%).
Missing predictor values	231 missing values, imputed inside each training fold (median for numbers, most frequent answer for categories).
Validation plan	1 × repeated 5-fold cross-validation (5 test folds), stratified by the outcome; every model was evaluated on the same folds; no separate ex
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 tuni
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
Models compared	Null model, Linear / logistic regression, Elastic net regression, Classification and regression tree, CART, Random forest, XGBoost gradient b
Performance (95% CI)	Linear / logistic regression: Accuracy 0.83 [0.81, 0.85]; Balanced accuracy 0.55 [0.51, 0.60]; AUC 0.75 [0.71, 0.80]; F1 0.21 [0.07, 0.34];
Calibration	Linear / logistic regression: calibration slope 0.89 (SE 0.07), calibration intercept −0.01, from the out-of-fold predictions of 1 repeat (ideal
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
Explanation methods	SHAP values (log-odds, probability scale); held-out permutation importance (computed in the test folds); partial dependence; accumulate
Leakage and data checks	Unbalanced answers: The answers are unbalanced (17% vs 83%). Read AUC and balanced accuracy rather than plain accuracy. Sample si

Figure 66. *The start of the model card.*

Under the card (Figure 67), *Out-of-fold predictions of Logistic regression* gives every customer the prediction of a model that did not see them: **Download predictions (CSV)** saves them as a file, and **Save predictions as a variable** adds them to the data. SaNaStat confirms: "Saved Churned_pred_linear to the data: the probability of "Yes" for each respondent, from a model that did not see them." *Score a new file with Logistic regression* predicts the customers of a new file (for example next year's customer list) with the final model, provided that the file has every predictor under the same name and type.



Out-of-fold predictions of Logistic regression

Each respondent's prediction comes from a model that did not see them, so these are honest predictions you can analyse further. "Predicted" uses the cut-off 0.2.

Download predictions (CSV)

Name of the new variable Churned_pred_linear_2 **Save predictions as a variable**

Adds the predictions to your data (type: computed), with a note on how they were made; use them in charts or other analyses.

Saved Churned_pred_linear to the data: the probability of "Yes" for each respondent, from a model that did not see them.

Score a new file with Logistic regression

Figure 67. *Out-of-fold predictions and the scoring of a new file.*

Note: predictions for named customers

A predictions file or a saved prediction variable holds one predicted probability per customer. Use it to target a retention offer within the rules you set in section 6.8, not to treat individual customers worse. Do not base decisions about individual customers on a model without human review.

6.10 Predicting response to the campaign

The same steps predict whether a customer responded to the campaign. On **Machine learning**, choose **Responded (yes/no: No / Yes)** as the outcome and tick the 13 predictors of section 6.1 and **Campaign_group**: customers were assigned to it at random, and it was known before they responded. Use the Quick look and run the comparison.

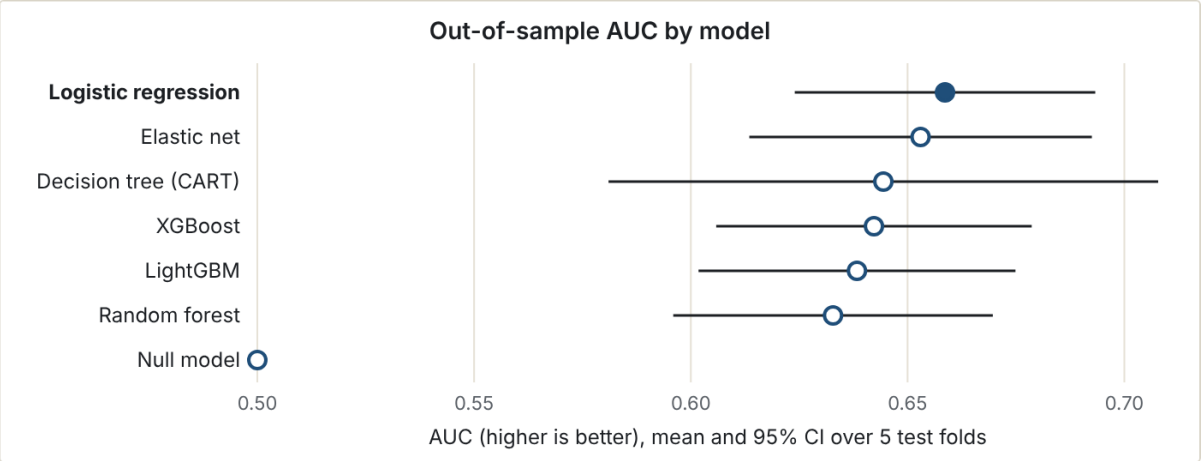


Figure 68. *Out-of-sample AUC of each model when predicting Responded (dots: mean over the test folds; lines: 95% confidence intervals).*

Table 28

Out-of-sample prediction of Responded (5-fold cross-validation)

Model	AUC M (SD)	95% CI	Accuracy	Balanced accuracy	F1	MCC	Brier	Δ vs best (p)
Logistic regression	.659 (.019)	[.624, .693]	.671	.586	.388	.209	0.212	best
Elastic net	.653 (.021)	[.613, .693]	.661	.559	.291	.162	0.214	−.006 (.565)
Decision tree (CART)	.644 (.034)	[.581, .708]	.673	.598	.426	.227	0.213	−.014 (.433)
XGBoost	.642 (.020)	[.606, .679]	.661	.560	.312	.168	0.216	−.016 (.242)
LightGBM	.638 (.020)	[.602, .675]	.659	.561	.319	.165	0.216	−.020 (.121)
Random forest	.633 (.020)	[.596, .670]	.641	.550	.321	.123	0.219	−.026 (.027)
Null model	.500 (.000)	[.500, .500]	.649	.500	.000	.000	0.228	−.159 (< .001)

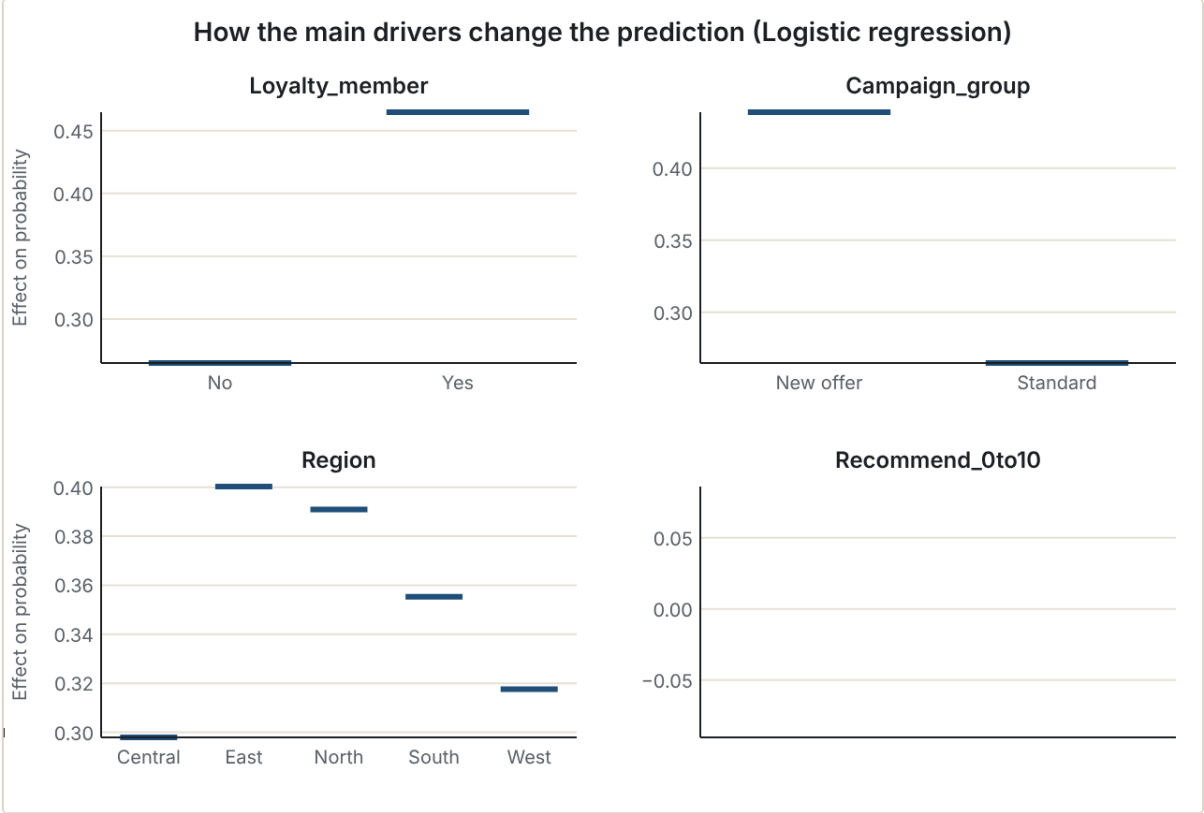


Figure 69. *Accumulated local effects of the main drivers of response.*

What it means for our example

Logistic regression is again the best model (AUC = .659, 95% CI [.624, .693]), no other model is significantly better, and the random forest is significantly worse (p = .027). The AUC is below .70, and SaNaStat marks the discrimination as *Note*: knowing the offer, loyalty status, region and the other variables separates responders from non-responders only modestly. The accuracy of the null model, which predicts "No" for everyone, is .649, close to that of the best model (.671), so accuracy tells us little here too (35% of the customers responded).

The strongest drivers are loyalty membership (mean |SHAP| = 0.460) and the campaign group (0.415); permutation importance puts them first and second as well (loss in AUC 0.087 and 0.070), and all six models rank them in the same two places. The average predicted probability of responding is 0.46 for loyalty members against 0.26 for non-members, and 0.44 after the New offer against 0.26 after the Standard offer. The regions differ somewhat (Central 0.30, East 0.40, North 0.39, South 0.36, West 0.32), and higher scores on Recommend_0to10 go with a higher predicted probability (about 0.18 across its range).

What it means for our example: calibration and cut-offs

The calibration slope is 0.79 (95% CI [0.63, 0.95]) and SaNaStat marks it *Check*: the predictions are too extreme, the usual sign of overfitting. In the reliability table, the tenth of customers with the lowest predicted response (mean .134) responded at .200, and the tenth with the highest predicted response (mean .648) responded at .580. Read the probabilities as a ranking, not as exact chances. Because 35.1% respond, SaNaStat rates the classes as reasonably balanced; the average precision is 0.50 against 0.35 for a model with no information. At the cut-off 0.50, 19% of the customers are flagged; 158 of the 526 responders are found (recall 0.30) and 56% of the flagged customers respond (precision 0.56, 158 of 284). At 0.30, 58% are flagged, 385 of the 526 responders are found (recall 0.73) and the precision falls to 0.44 (385 of 876). At 0.05 everybody is flagged: all responders are found and the precision is 0.35, the response rate. A cut-off is useful only to the extent that precision rises above that baseline.

Note: the offer effect is shown by the A/B test, not by the prediction

Which customers respond is a prediction question; whether the *offer* raised the response is a causal question that the randomised A/B test of section 4.1 answers. The model agrees (the New offer raises the predicted response from 0.26 to 0.44, in line with the observed 26.1% and 44.3%), but SaNaStat does not offer uplift modelling, which asks which customers are moved by the offer.

How to report it

"Seven prediction models were compared with 5-fold cross-validation. Logistic regression performed best (AUC = .659, 95% CI [.624, .693]; Balanced accuracy = .586); discrimination was modest and the predicted probabilities were somewhat too extreme (calibration slope 0.79, 95% CI [0.63, 0.95])."

Exercises

Exercise: churn and response

1. Predict Churned again after adding Annual_spend as a predictor. What happens to the AUC, and why should you not believe it? Think about leakage (section 6.1).

2. In the comparison for Churned, set the cut-off to 0.30 and read the counts. Answer: 96 of the 250 leavers are found (recall 0.38) among 232 flagged customers (precision 0.41); see Table 25 for the other cut-offs.

3. On the Fairness tab, choose Main_channel. Are the selection rates within the four-fifths guideline? Answer: just below it (ratio 0.79, Mobile app against Online).

4. Planted answer for the chapter: churn is driven by satisfaction, complaints, tenure and loyalty (section 6.2); response by the offer, loyalty and channel (logistic regression of Responded on Campaign_group, Loyalty_member and Main_channel: loyalty members OR = 2.42, Standard offer OR = 0.44 relative to the New offer).

7. Customer segmentation

Marketing often asks whether the customers fall into a few groups with different needs, so that each group can be treated differently. *Cluster analysis* looks for groups of customers who are similar to each other and different from the others. SaNaStat's **Clustering** screen offers k-means and hierarchical clustering.

> **Concept: what cluster analysis does, and does not do**
>
> Cluster analysis always returns clusters, even when the data contain none. A solution is only useful if it is (a) reasonably separated (the *silhouette*), (b) stable when the sample changes (*stability*) and (c) interpretable and useful. The clusters describe this sample; they are not proof that customer "types" exist. Naming a cluster is a judgement by the analyst, not a result.

SaNaStat standardises every variable before clustering (mean 0, standard deviation 1), so a variable measured in large units does not dominate. Customers with a missing value on any chosen variable are left out.

7.1 Segments with clear structure: segments_demo.csv

The file *segments_demo.csv* has 360 simulated shoppers rated on five attitude scales (Price_sensitivity, Brand_loyalty, Quality_focus, Online_affinity, Social_influence, each 1 to 7), with Age, Income_k, Gender and Channel (Mixed, Online, Store). Three segments were planted.

1. On **Data**, open segments_demo.csv (360 respondents, 10 variables). Click **Clustering** in the menu.
2. In *1 · Settings (What to cluster)*, click **Select none** and tick the five attitude scales. Leave Age and Income_k unticked: we want to *describe* the clusters with them, not build the clusters from them.
3. Open **Compare the clusters on other variables (optional)** and tick Age, Income_k, Gender and Channel.
4. Keep **Method** k-means, solutions from 2 to 8 clusters, **Show** the best silhouette, and tick **Also compute the gap statistic (slower)** (Figure 70).
5. Click **Find the clusters**.

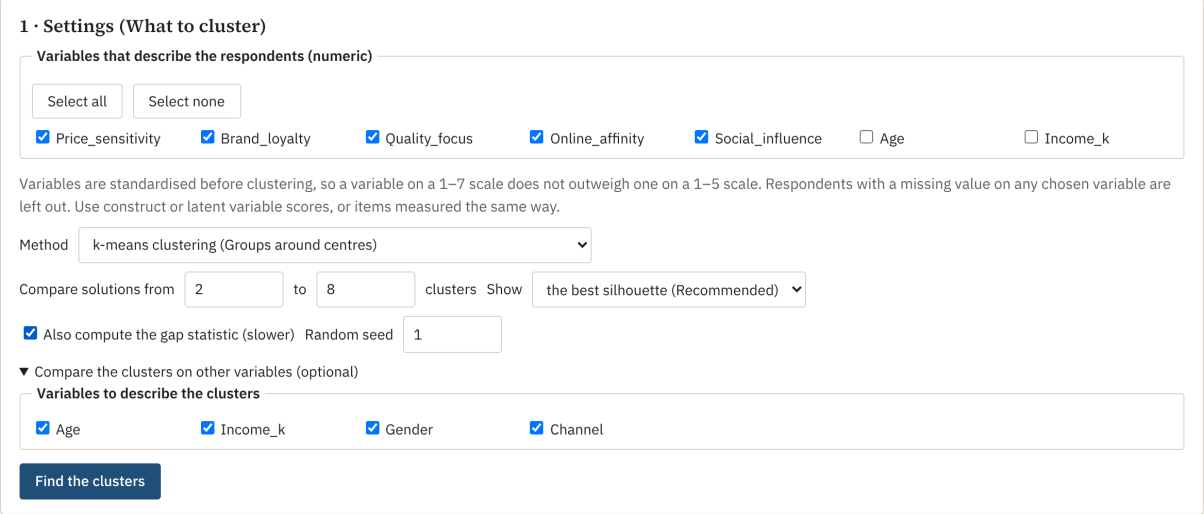


Figure 70. *The clustering settings for segments_demo.csv.*

The result starts with a plain-language summary and the checks (Figure 71), then the table of solutions (Table 29), the cluster means, the comparison with the other variables and the figures.

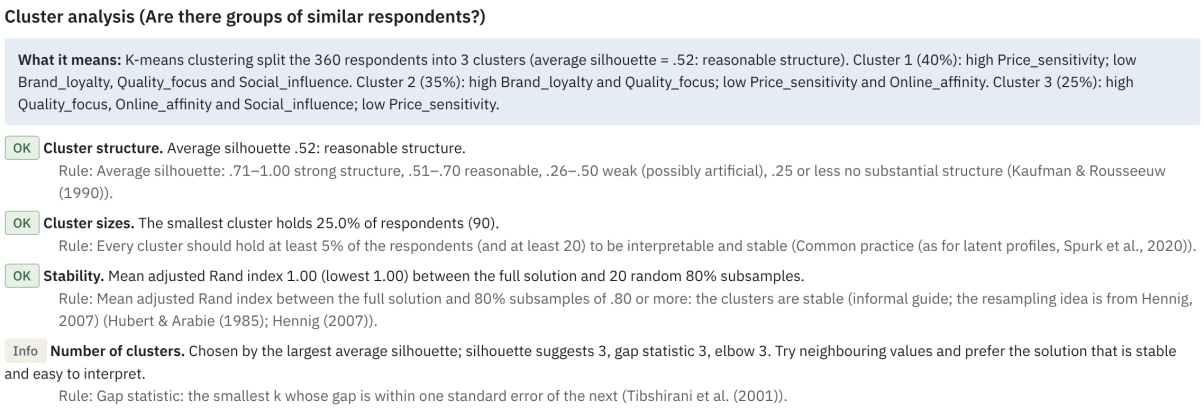


Figure 71. *The start of the cluster report: what it means and the checks.*

Choosing the number of clusters

SaNaStat compares solutions with 2 to 8 clusters.

Table 29

Solutions with 2 to 8 clusters for segments_demo.csv

Clusters	Average silhouette	Calinski–Harabasz	Gap (SE)	Smallest cluster
2	.437	323.7	0.701 (0.020)	40.3%
3 (shown)	.520	466.8	1.194 (0.019)	25.0%
4	.375	349.3	1.157 (0.025)	17.5%
5	.364	283.2	1.129 (0.023)	10.0%
6	.166	243.6	1.101 (0.025)	12.5%
7	.169	220.8	1.099 (0.025)	10.0%
8	.170	200.1	1.057 (0.025)	9.4%

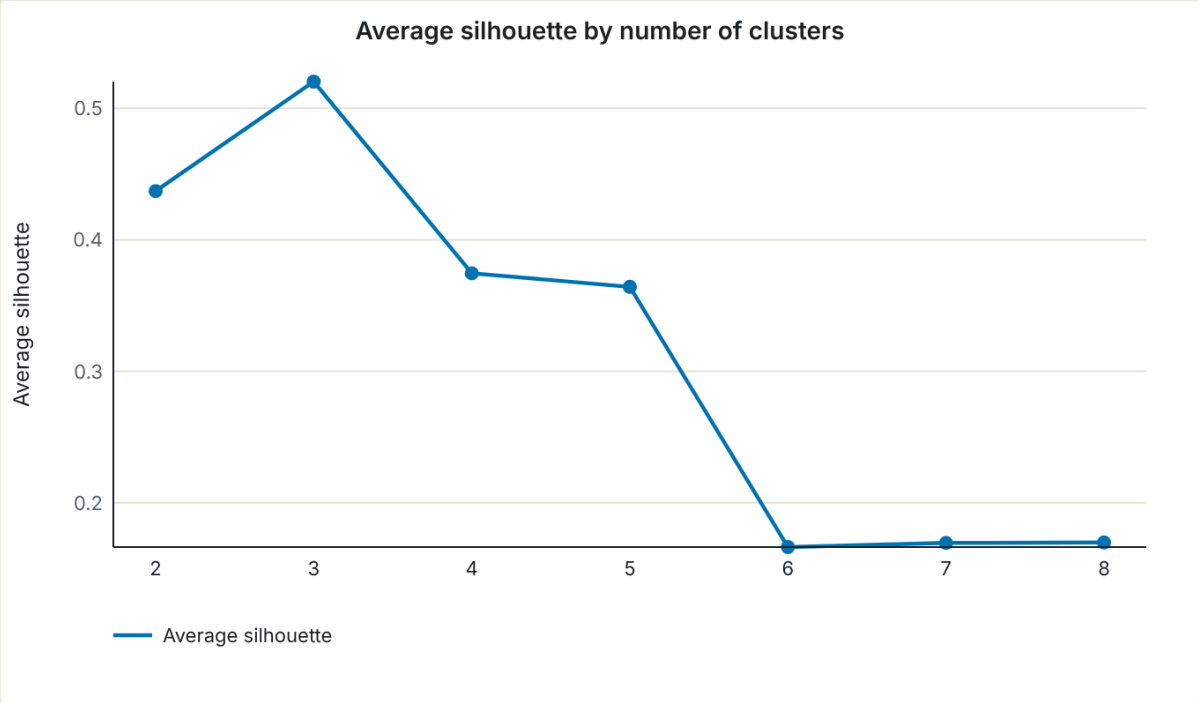


Figure 72. *Average silhouette by number of clusters (higher is better).*

Concept: silhouette, gap statistic and stability

The *silhouette* of a customer compares how close the customer is to the own cluster with how close it is to the nearest other cluster; the average silhouette runs from −1 to 1. The *Calinski–Harabasz* index is the ratio of between-cluster to within-cluster spread. The *gap statistic* compares the within-cluster spread with what random data without groups would give (Tibshirani et al., 2001); choose the smallest k whose gap is within one standard error of the next. *Stability* is the mean adjusted Rand index between the solution on all customers and the solutions on 20 random 80% subsamples (1 = identical groupings). The within-cluster sum of squares always falls as clusters are added, so look for an elbow.

Cut-off values: clusters

Average silhouette: .71–1.00 strong structure, .51–.70 reasonable, .26–.50 weak (possibly artificial), .25 or less no substantial structure (Kaufman & Rousseeuw, 1990). **Cluster size:** at least 5% of the respondents and at least 20. **Stability:** a mean adjusted Rand index of .80 or more (informal guide; Hubert & Arabie, 1985; Hennig, 2007).

What it means for our example

Silhouette (.520), Calinski–Harabasz (466.8), the gap statistic (1.194) and the elbow all point to three clusters; SaNaStat says "silhouette suggests 3, gap statistic 3, elbow 3". The three-cluster solution has reasonable structure (silhouette .52), a smallest cluster of 25.0% (90 shoppers) and perfect stability (mean adjusted Rand index 1.00 over 20 subsamples). With four clusters the silhouette falls to .37 (weak structure), the smallest cluster is 17.5% and stability is .92: the extra cluster splits a group in two similar parts instead of finding a new type.

Profiling and naming the clusters

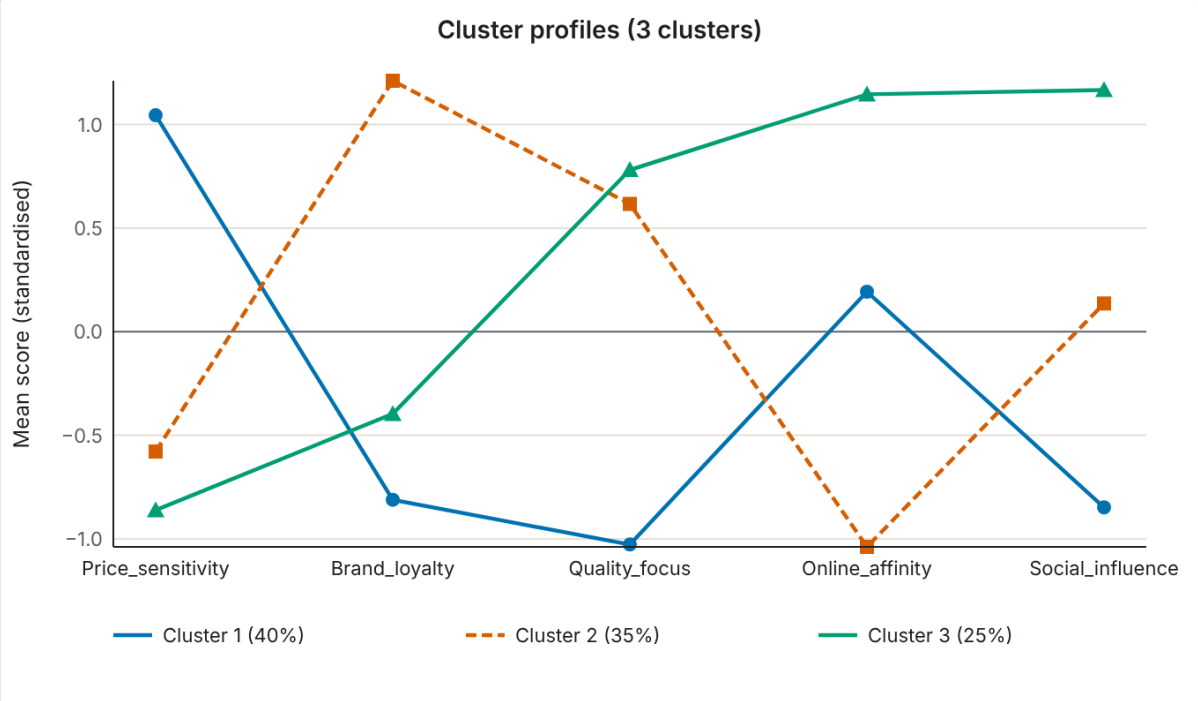


Figure 73. *Cluster profiles: mean standardised scores of the three clusters.*

Table 30

The three clusters of segments_demo.csv (means of the clustering variables; standardised means in brackets)

Variable	Cluster 1 (n = 144)	Cluster 2 (n = 126)	Cluster 3 (n = 90)
Price_sensitivity	5.97 (1.04)	3.42 (−0.58)	2.98 (−0.86)
Brand_loyalty	2.55 (−0.81)	6.12 (1.21)	3.28 (−0.40)
Quality_focus	3.16 (−1.03)	5.56 (0.62)	5.80 (0.78)
Online_affinity	4.74 (0.19)	3.03 (−1.04)	6.07 (1.15)
Social_influence	3.04 (−0.85)	4.37 (0.14)	5.77 (1.17)

Table 31

The clusters on variables that were not used to build them

Variable	Cluster 1	Cluster 2	Cluster 3
Age, mean (SD)	34.29 (8.57)	48.98 (10.22)	27.30 (5.47)
Income_k, mean (SD)	31.88 (8.87)	56.96 (15.62)	46.72 (13.65)
Channel: Store	28.5%	59.5%	8.9%
Channel: Online	20.1%	26.2%	60.0%
Channel: Mixed	51.4%	14.3%	31.1%
Gender: female	60.4%	55.6%	47.8%

What it means for our example

Cluster 1 (40%) is price sensitive and has low brand loyalty, quality focus and social influence; it is younger (mean age 34.29) and has the lowest income (31.88). Cluster 2 (35%) combines high brand loyalty and quality focus with low price sensitivity and low online affinity; it is the oldest group (48.98), has the highest income (56.96) and shops mostly in stores (59.5%). Cluster 3 (25%) is quality focused, online and socially influenced, with low price sensitivity; it is the youngest (27.30) and shops mostly online (60.0%). The differences in age (Welch $F(2, 234.1) = 202.04$, $p < .001$, $\eta^2 = .51$), income ($F(2, 191.2) = 143.69$, $p < .001$, $\eta^2 = .42$) and channel ($\chi^2(4) = 98.99$, $p < .001$, Cramér's $V = .37$) are large, whereas gender does not differ ($\chi^2(2) = 3.58$, $p = .167$, $V = .10$). Possible names, which are our suggestion and not a program result: *price-driven bargain hunters*, *loyal traditionalists* and *connected quality seekers*.

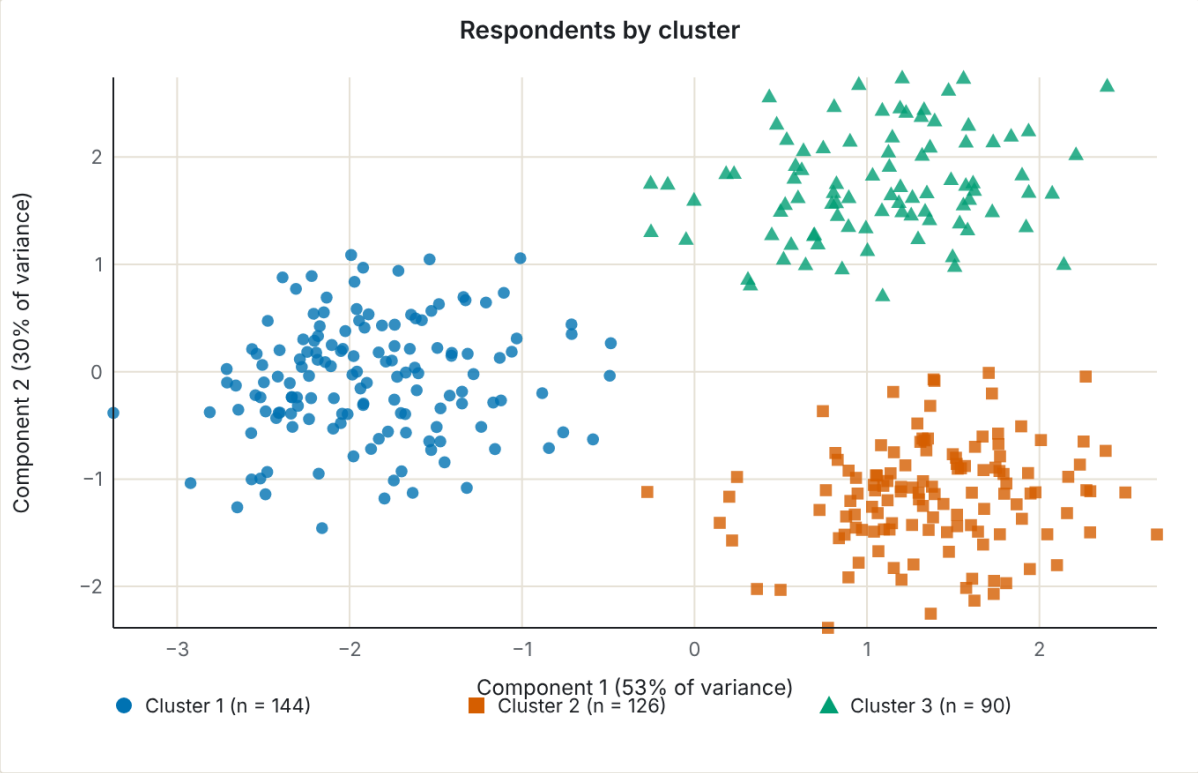


Figure 74. *The shoppers on the first two principal components, coloured by cluster.*

Tip: how to name a cluster

Use the standardised means: values above about +0.5 or below −0.5 characterise the cluster (SaNaStat uses this limit in its summary). Name the cluster by its two or three most distinctive traits and check the names against the variables that were not used to build the clusters (age, income, channel), as above. Cluster numbers are only labels sorted by size; they can change when you change the data or the settings.

Hierarchical clustering with Ward's method (choose it in **Method** and click **Find the clusters** again) gives the same three clusters (sizes 40%, 35% and 25%, silhouette .52, stability .99), which is reassuring: well-separated groups do not depend on the algorithm.

How to report it

"K-means cluster analysis of Price_sensitivity, Brand_loyalty, Quality_focus, Online_affinity and Social_influence (standardised; n = 360) gave 3 clusters of 144, 126 and 90 respondents (k chosen by the largest average silhouette; average silhouette = .52, mean adjusted Rand index over 20 subsamples = 1.00)."

7.2 Segments in the customer file

Now cluster the customers of practice_marketing.csv on five behaviour variables.

1. On **Data**, open practice_marketing.csv. Click **Clustering**.

2. Keep the **Method** k-means, click **Select none**, then tick Visits_per_month, Avg_basket, Discount_share, Satisfaction_1to10 and Tenure_months.
3. Under **Compare the clusters on other variables (optional)**, tick Region, Main_channel, Churned, Loyalty_member, Annual_spend and Complaints_12m.
4. Tick **Also compute the gap statistic (slower)** and click **Find the clusters**.

Cluster analysis (Are there groups of similar respondents?)

What it means: K-means clustering split the 1415 respondents into 5 clusters (average silhouette = .18: no substantial structure). Cluster 1 (33%): low Avg_basket, Discount_share and Satisfaction_1to10. Cluster 2 (19%): high Discount_share. Cluster 3 (18%): high Visits_per_month. Cluster 4 (17%): high Avg_basket and Satisfaction_1to10. Cluster 5 (13%): high Tenure_months and Satisfaction_1to10. The data show no clear groups, so treat the clusters as a convenient way to split the sample, not as types of respondents.

Check **Cluster structure.** Average silhouette .18: no substantial structure. Cluster analysis always returns clusters, even when the data hold none.
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 12.6% of respondents (178).
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 .87 (lowest .81) 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 the largest average silhouette; silhouette suggests 5, elbow 5. Try neighbouring values and prefer the solution that is stable and easy to interpret.

Figure 75. *The report for the customer file: what it means and the checks.*

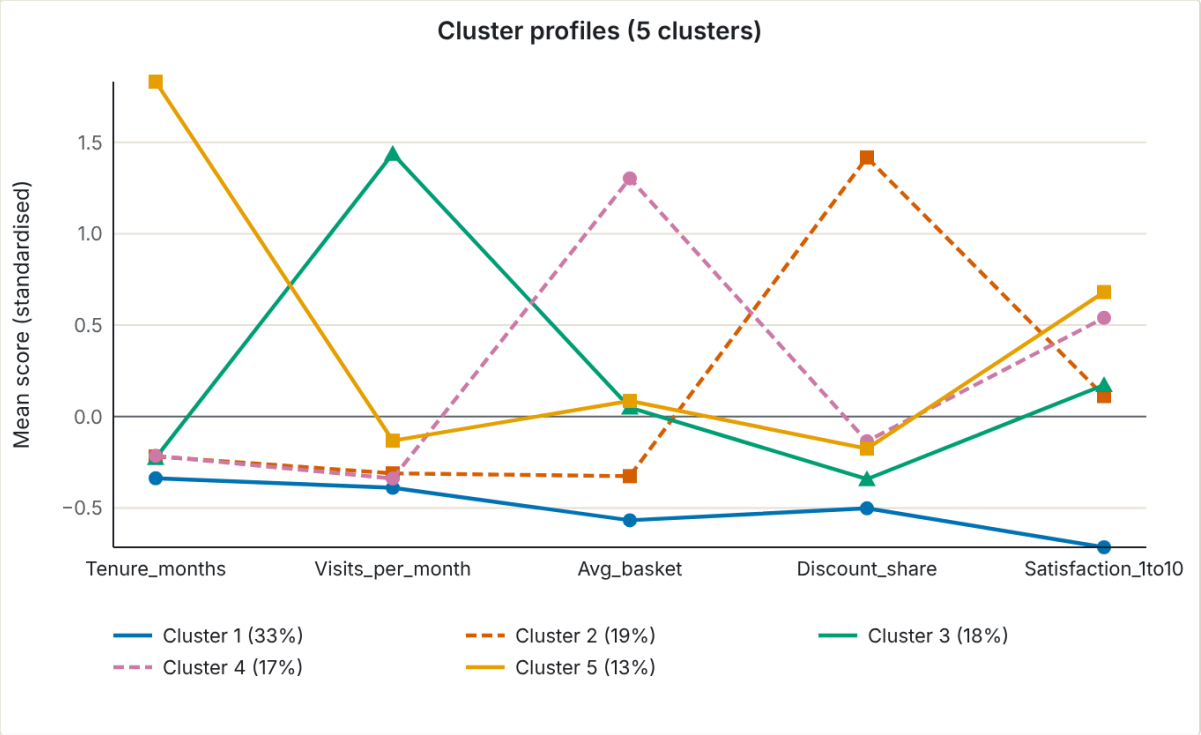


Figure 76. *Cluster profiles of the five customer clusters.*

Table 32

The five customer clusters (means; standardised means in brackets)

Variable	Cluster 1 (n = 461)	Cluster 2 (n = 271)	Cluster 3 (n = 258)	Cluster 4 (n = 247)	Cluster 5 (n = 178)
Tenure_months	23.63 (−0.34)	26.18 (−0.22)	26.00 (−0.23)	26.25 (−0.21)	69.94 (1.83)
Visits_per_month	2.48 (−0.39)	2.64 (−0.31)	6.06 (1.44)	2.58 (−0.34)	2.99 (−0.13)
Avg_basket	79.77 (−0.57)	85.57 (−0.33)	94.58 (0.05)	124.69 (1.30)	95.45 (0.09)

Variable	Cluster 1 (n = 461)	Cluster 2 (n = 271)	Cluster 3 (n = 258)	Cluster 4 (n = 247)	Cluster 5 (n = 178)
Discount_share	0.19 (−0.50)	0.47 (1.42)	0.21 (−0.34)	0.24 (−0.14)	0.24 (−0.18)
Satisfaction_1to10	5.51 (−0.71)	6.82 (0.11)	6.91 (0.17)	7.49 (0.54)	7.72 (0.68)

What it means for our example

SaNaStat chose five clusters (largest silhouette) from 1415 customers (85 with a missing satisfaction rating were left out), but the average silhouette is only .18 and the check reads *Check*: "no substantial structure. Cluster analysis always returns clusters, even when the data hold none." The silhouette hardly changes between 2 and 8 clusters (.150 to .183), and the clusters are stable (mean adjusted Rand index .87). The data therefore form a continuous cloud, not distinct types. The five clusters can still be used as a *descriptive* grouping of customers, as SaNaStat's own summary says: treat the clusters "as a convenient way to split the sample". They are: a large, less satisfied, low-basket group (Cluster 1, 33%); discount shoppers (Cluster 2, 19%, discount share 0.47); frequent visitors (Cluster 3, 18%, 6.06 visits per month); big-basket customers (Cluster 4, 17%, basket 124.69); and long-tenure satisfied customers (Cluster 5, 13%, 69.94 months).

Note: stable does not mean real

High stability (.87) tells you that the same customers fall into the same clusters when the sample is perturbed. It does not tell you that the groups exist in nature: a method that cuts a continuous cloud into pieces can do so repeatedly. Read the silhouette first.

7.3 Keep the clusters and compare them

The second card on the Clustering screen, *2 · Keep the clusters (Use them in other analyses)*, adds each customer's cluster to the data as a grouping variable (Figure 77).

1. In **Name of the new variable** type **Segment**.
2. Click **Add to the data**. SaNaStat confirms: "Added “Segment”."
3. Use Segment in any analysis. Customers who were left out of the clustering have no cluster.

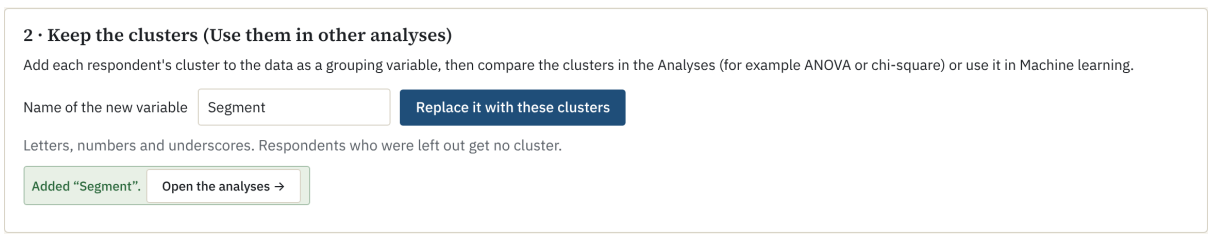


Figure 77. *2 · Keep the clusters: Segment has been added to the data.*

Churn by segment. Choose **Analyses** → **Chi-square, Fisher, McNemar** with Segment and Churned (Figure 78).

Counts (row %)

Segment	No	Yes	Total
Cluster 1	341 (74%)	120 (26%)	461
Cluster 2	222 (82%)	49 (18%)	271
Cluster 3	231 (90%)	27 (10%)	258
Cluster 4	222 (90%)	25 (10%)	247
Cluster 5	165 (93%)	13 (7%)	178

Copy table | Copy as picture | PNG | CSV

Tests

Test	Value	df	p
Pearson chi-square	55.82	4	< .001
Likelihood ratio	56.43	4	< .001

Copy table | Copy as picture | PNG | CSV

Adjusted standardised residuals

Segment	No	Yes
Cluster 1	−6.68 *	6.68 *
Cluster 2	−0.76	0.76
Cluster 3	2.90 *	−2.90 *
Cluster 4	2.99 *	−2.99 *
Cluster 5	3.55 *	−3.55 *

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

Figure 78. *Churned by Segment: counts, tests and adjusted residuals.*

Table 33

Churned by customer segment

Segment	No	Yes	Total
Cluster 1	341 (74%)	120 (26%)	461
Cluster 2	222 (82%)	49 (18%)	271
Cluster 3	231 (90%)	27 (10%)	258
Cluster 4	222 (90%)	25 (10%)	247

Segment	No	Yes	Total
Cluster 5	165 (93%)	13 (7%)	178

What it means for our example

Churn differs between the segments, $\chi^2(4, N = 1415) = 55.82$, $p < .001$, Cramér's V = .20. Cluster 1 (the less satisfied group) churns more than expected (adjusted residual +6.68, 26% against 7% in Cluster 5). Because satisfaction helped to build the clusters, and satisfaction also predicts churn (chapter 6), this result is partly expected. It nevertheless shows the practical value of the segments: one customer in four in the largest segment left, and one in fourteen in the long-tenure segment.

Region by segment. Replace Churned by Region: the segments do not differ by region, $\chi^2(16, N = 1415) = 18.40$, $p = .301$, Cramér's V = .06. A segmentation that does not depend on region can be used in all regions.

Spend by segment. The cluster report itself compares Annual_spend across the clusters: Welch's $F(4, 565.8) = 212.94$, $p < .001$, $\eta^2 = .41$, with the frequent visitors (Cluster 3) spending most (mean 6407.81 against 2240.58 in Cluster 1).

Note: a segment comparison on a variable that built the segment

Cluster 3 was defined by frequent visits, and annual spend is largely visits times basket (chapter 5). The spend difference is therefore built in, not an independent finding. SaNaStat says the same: tests on variables that were *not* used to form the clusters are meaningful, but the clusters were chosen from the same sample, so confirm in new data before calling them types. Region and Churned were not used here, so those comparisons are fair; Annual_spend is a consequence of the clustering variables.

Tip: profile with the right variables

Cluster on the variables that you can act on (behaviour, attitudes) and profile with variables that tell you how to reach the clusters (region, channel, age) and with business outcomes (churn, spend). Keep clustering variables and profiling variables apart. Chapter 8 shows a different kind of segmentation, by recency, frequency and money, which needs no clustering at all.

Exercises

Exercise: segmentation

1. Cluster segments_demo.csv on the five scales and ask for four clusters (**Show 4 clusters**). Which statistic tells you that the fourth cluster is not needed? Answer: the silhouette drops from .520 to .375 and the gap statistic does not rise.

2. Cluster practice_marketing.csv on Visits_per_month and Avg_basket only. Does the silhouette improve? Report what you see.

3. Planted answer for the chapter: three segments in segments_demo.csv (silhouette .52, sizes 40%, 35%, 25%); no clear structure in the customer file (silhouette .18).

8. Customer value: RFM segments and lifetime value

Not every customer is worth the same. A shop that knows which customers buy often and recently, and which have stopped, can reward the first and try to win back the second. *RFM analysis* scores every customer on three simple facts from the sales records, and the *customer lifetime value* (CLV) turns them into money. This chapter uses *practice_transactions.csv*: the order lines of a simulated online grocery shop from 2023 to 2025.

8.1 The data: one row per item bought

On the **Data** screen open practice_transactions.csv (start a new project). SaNaStat reports *Open practice_transactions.csv: 21006 respondents, 5 variables.* Here a "respondent" is one row of the file, that is one item in one order. Card 2 counts *1 scale · 4 categorical*: Order_ID (4961 groups, one per order), Customer_ID (900 groups, one per customer), Order_date (988 different dates, such as 2023-01-04), Item (20 groups, such as Apples, Bread and Butter) and the scale variable Line_amount (the money paid for that item). This *long* layout, with several rows per order and several orders per customer, is how most shop systems export their sales.

Concept: recency, frequency and monetary value

For each customer, *recency* (R) is the number of days since the last purchase, *frequency* (F) the number of purchases and *monetary value* (M) the total spend. Each is turned into a score from 1 to 5 by splitting the customers into five equal-sized groups (quintiles): 5 is the most recent, most frequent or highest-spending fifth *of this customer base*. Customers who bought recently, often and for much money are the most valuable; customers who used to buy often but have not bought for a long time are at risk. Recency, frequency and money together predict future value better than any one of them (Fader et al., 2005).

8.2 Run the RFM analysis

1. Choose **Analyses** → **RFM and customer value**.
2. In **How are the data arranged?** choose **One row per purchase (or per item bought)**.
3. Choose **Customer_ID** in **Customer**, **Order_date** in **Purchase date** and **Line_amount** in **Amount spent**.
4. Choose **Order_ID** in **Order number (optional)**, so that the several rows of one order count as one purchase.
5. Keep **Score groups** at **5 (quintiles, usual)**, the **Gross margin (%)** at 30 and the **Discount rate per year (%)** at 10 (Figure 79).

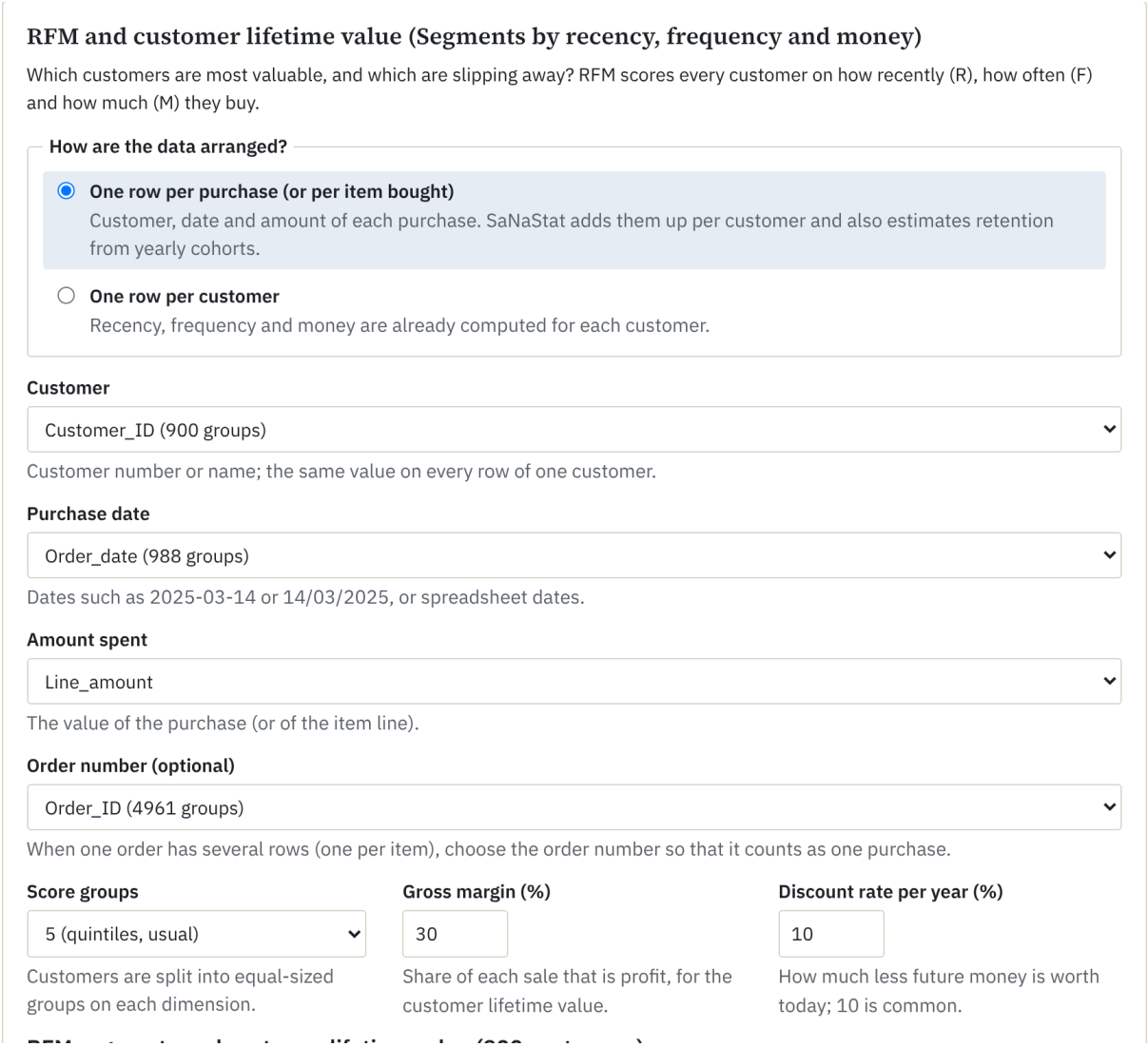


Figure 79. *RFM and customer value: the customer, the date, the amount and the order number of every row.*

The result has a summary with two checks (Figure 80), the table of segments (Table 34), the meaning of the scores (Figure 81), the yearly cohorts (Table 35), two charts and the sentence for the report (Figure 83).

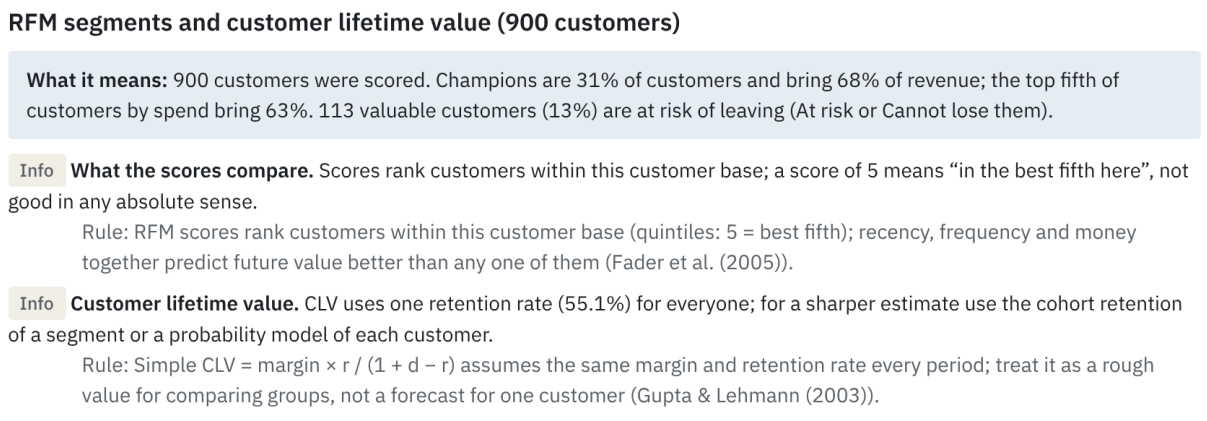


Figure 80. *RFM segments and customer lifetime value: what it means and the checks.*

What it means for our example

SaNaStat writes: "900 customers were scored. Champions are 31% of customers and bring 68% of revenue; the top fifth of customers by spend bring 63%. 113 valuable customers (13%) are at risk of leaving (At risk or Cannot lose them)." The 900 customers made 4,961 purchases up to 2025-12-30, the last date in the file, from which recency is counted.

8.3 Reading the segments

SaNaStat combines the recency score (R) with the mean of the frequency and monetary scores (FM) into nine named segments; the note under the table gives the rule for each, for example Champions R 4–5 and FM 4–5, Cannot lose them R 1–2 and FM 4–5, Lost R 1 and FM 1–2.

Table 34

The RFM segments of the 900 customers (means per customer)

Segment	Customers	Share	Recency (days)	Frequency	Monetary	Share of revenue	CLV (simple)
Champions	277	30.8%	31	11.0	241.57	67.7%	24.37
Loyal customers	170	18.9%	117	5.8	103.18	17.8%	10.41
Potential loyalists	22	2.4%	57	1.4	18.01	0.4%	1.82
New customers	6	0.7%	63	1.0	6.89	0.0%	0.70
Need attention	65	7.2%	178	1.1	15.82	1.0%	1.60
Cannot lose them	49	5.4%	538	6.4	108.32	5.4%	10.93
At risk	64	7.1%	530	3.8	56.82	3.7%	5.73
Hibernating	103	11.4%	445	1.1	16.13	1.7%	1.63
Lost	144	16.0%	885	1.0	15.81	2.3%	1.59

What it means for our example

The 277 Champions bought on average 31 days before the end of the data, 11.0 times, for 241.57 in total, and bring 67.7% of all revenue. The 49 customers in *Cannot lose them* bought almost as often (6.4 purchases, 108.32) but their last purchase was 538 days ago on average: they were valuable and have stopped. The 144 Lost customers bought once (frequency 1.0) and not for 885 days. The last column suggests an action for each segment, for example "Win them back personally; find out why they stopped." for *Cannot lose them*. These are common practice, not results: decide what fits your business.

What the scores mean

Score	Recency (days since last purchase)	Frequency (purchases)	Monetary (total spend)
5	1–25	8–27	137.06–757.56
4	26–94	5–8	85.42–135.59
3	95–270	3–5	41.96–85.39
2	274–634	1–3	14.93–41.48
1	639–1092	1–1	1.64–14.91

5 equal-sized groups of customers per dimension (quintiles); ties are split by the order of the customers. 5 = most recent, most frequent, highest spend.

Figure 81. *What the scores mean: the range of recency, frequency and spend in each score group.*

Note: scores are relative

A score of 5 means "in the best fifth here", not good in any absolute sense, and SaNaStat says so in the first check. In this shop a frequency score of 5 means 8 to 27 purchases in three years and a monetary score of 5 means 137.06 to 757.56; in another shop the same score means something else. Customers with the same value can fall into neighbouring groups, because ties are split by the order of the customers.

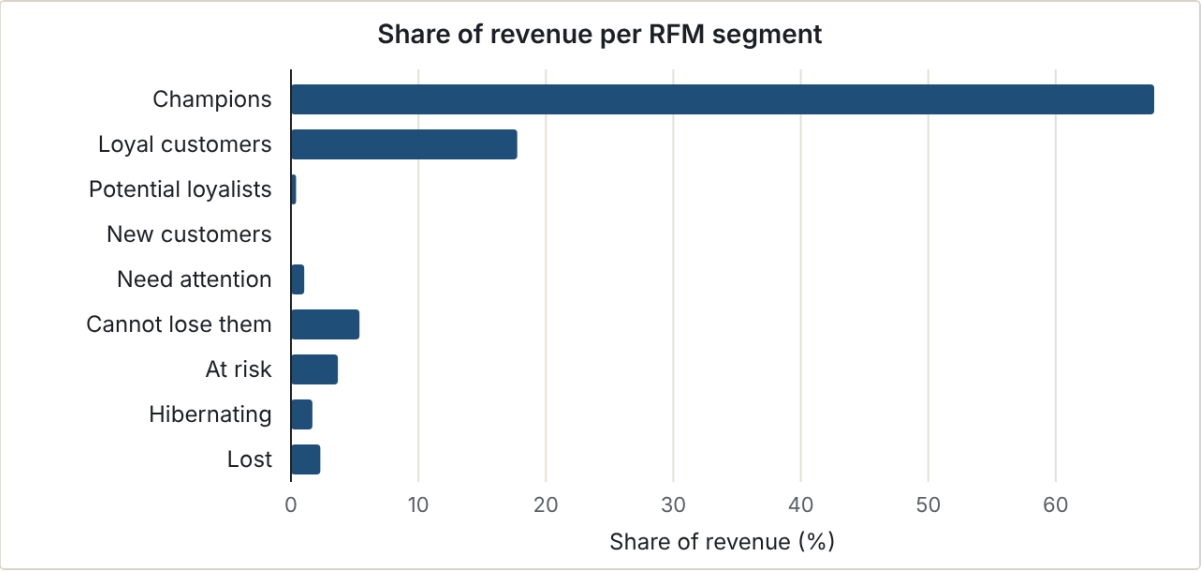


Figure 82. *Share of revenue per RFM segment.*

8.4 Cohort retention and the simple customer lifetime value

Because the data hold dates, SaNaStat groups the customers by the year of their first purchase (the *cohorts*) and follows them (Table 35).

Table 35

Customer cohorts by year of first purchase

First purchase	Customers	Year +1	Year +2	Retention	Spend per active year	CLV (simple)
2023	266	44%	38%	54.7%	66.40	19.71
2024	298	56%		55.7%	74.35	22.88
2025	336			–	95.94	–

Retention is the share of the customers who bought in one year and bought again in the next. Over all cohorts it is 55.1%, and SaNaStat uses this rate for the customer lifetime value.

Concept: a simple customer lifetime value

The customer lifetime value is the money a customer is expected to bring in the future, valued today. SaNaStat uses the simplest form: CLV = margin × r ÷ (1 + d − r), where *margin* is the yearly profit from the customer (here 30% of the yearly spend), *r* the yearly retention rate and *d* the yearly discount rate (Gupta & Lehmann, 2003). The formula assumes that a customer who is still there buys the same every year and leaves with the same chance every year.

Cut-off values: what the CLV is good for

Simple CLV = margin × r ÷ (1 + d − r) assumes the same margin and retention rate every period; treat it as a rough value for comparing groups, not a forecast for one customer (Gupta & Lehmann, 2003). For a sharper estimate use the cohort retention of a segment or a probability model of each customer.

What it means for our example

With a retention rate of 55.1%, a 30% margin and a 10% discount rate, the simple customer lifetime value averaged 11.07 per customer, from 24.37 for Champions to 0.70 for New customers. The 2024 cohort kept more of its customers (56% bought again in 2025) than the 2023 cohort did after one year (44%). The CLV is about one year's margin here, because a customer who returns with a chance of 55.1% a year does not stay long. Change the margin to 20% and the average CLV falls to 7.38 (Champions 16.25): the CLV is proportional to the margin, so use your real margin.

How to report it

"Customers (N = 900) were scored 1–5 on recency, frequency and monetary value (RFM; quintiles) from 4,961 purchases up to 2025-12-30 and grouped into 9 segments; Champions 30.8%, Loyal customers 18.9%, Potential loyalists 2.4%. With a yearly retention rate of 55.1%, a 30% margin and a 10% discount rate, the simple customer lifetime value averaged 11.07."

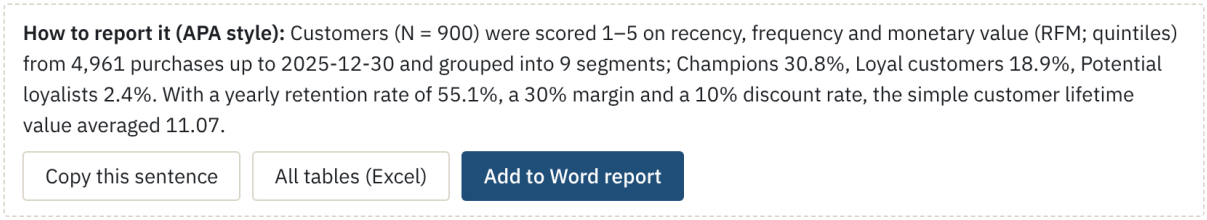


Figure 83. *The sentence for the report, with its buttons.*

8.5 Choices that change the result

- **The order number.** Without **Order number (optional)**, every row counts as a purchase. SaNaStat then reports "21,006 purchases" instead of 4,961, and Champions appear to buy 52.5 times instead of 11.0: every item of an order counts as a purchase. Choose the order number whenever an order has several rows.
- **Score groups.** With **3 (for small customer bases)** there are eight segments, and Champions are 23.2% of the customers. Use quintiles unless the customer base is small.
- **One row per customer.** If your system already gives recency, frequency and money per customer, choose **One row per customer** and the three variables; then type the **Retention per year (%)** from your records and the **Years covered by the data**. In this layout **Save the segment of each customer as a new variable** adds the segment to the data, for use in other analyses (for example, the response to a campaign by segment).

Note: RFM looks back

RFM describes past behaviour. A customer in *Lost* may simply have moved away; a *New customer* may never come back. Treat the segments as a way to prioritise attention, test the actions you take (chapter 4), and recompute the scores regularly.

Exercises

Exercise: customer value

1. Run the RFM analysis without the order number. How many purchases does SaNaStat count, and what happens to the frequency of the Champions? Answer: 21,006 purchases; 52.5 instead of 11.0.

2. Which segment would you contact first in a win-back campaign, and why? Look at *Cannot lose them* and *At risk* (113 customers together).

3. Planted answer for the chapter: the file was built with a group of loyal heavy buyers, many one-off buyers and customers who stop buying. They appear as Champions (30.8%), Lost and Hibernating customers who bought about once, and the *Cannot lose them* and *At risk* segments.

9. Market basket analysis: which products are bought together?

A shop that knows which products end up in the same basket can place them together, bundle them or recommend one to buyers of the other. *Market basket analysis* finds such pairs with *association rules* of the form "if a basket has A, it also has B" (Agrawal et al., 1993). This chapter uses the same practice_transactions.csv as chapter 8: each of its 4,961 orders is a basket.

9.1 Support, confidence and lift

Concept: support, confidence and lift

For a rule A → B:

- **Support** is the share of all baskets that contain both A and B: how common the combination is.
- **Confidence** is the share of the baskets with A that also contain B: how often the rule is right.
- **Lift** is the confidence divided by how common B is in all baskets. A lift of 1 means that A tells you nothing about B; above 1, A and B are bought together more often than if they were unrelated (Brin et al., 1997).
- **Leverage** is the support minus the support expected if the items were unrelated, and **conviction** says how much more often the rule would be wrong if the items were unrelated.

A rule with high confidence but lift near 1 only reflects a popular item: if 43% of all baskets contain milk, "baskets with cheese also contain milk" is right about 43% of the time anyway.

SaNaStat finds the rules with the Apriori algorithm (Agrawal & Srikant, 1994): it first finds the combinations of items that reach the minimum support, then the rules among them that reach the minimum confidence and lift.

9.2 Find the rules

1. On the **Data** screen open practice_transactions.csv, then choose **Analyses** → **Market basket**.
2. In **How are the baskets stored?** choose **One row per item bought (long)**. (A file with one row per basket and one 0/1 column per item is the **One row per basket, one column per item (wide)** layout.)
3. Choose **Order_ID** in **Basket (transaction or order number)** and **Item** in **Item**.
4. Keep the **Minimum support (%)** at 1, the **Minimum confidence (%)** at 30, the **Minimum lift** at 1 and **Items per rule (at most)** at **2 (pairs)** (Figure 84).

5. Click **Find the rules**.

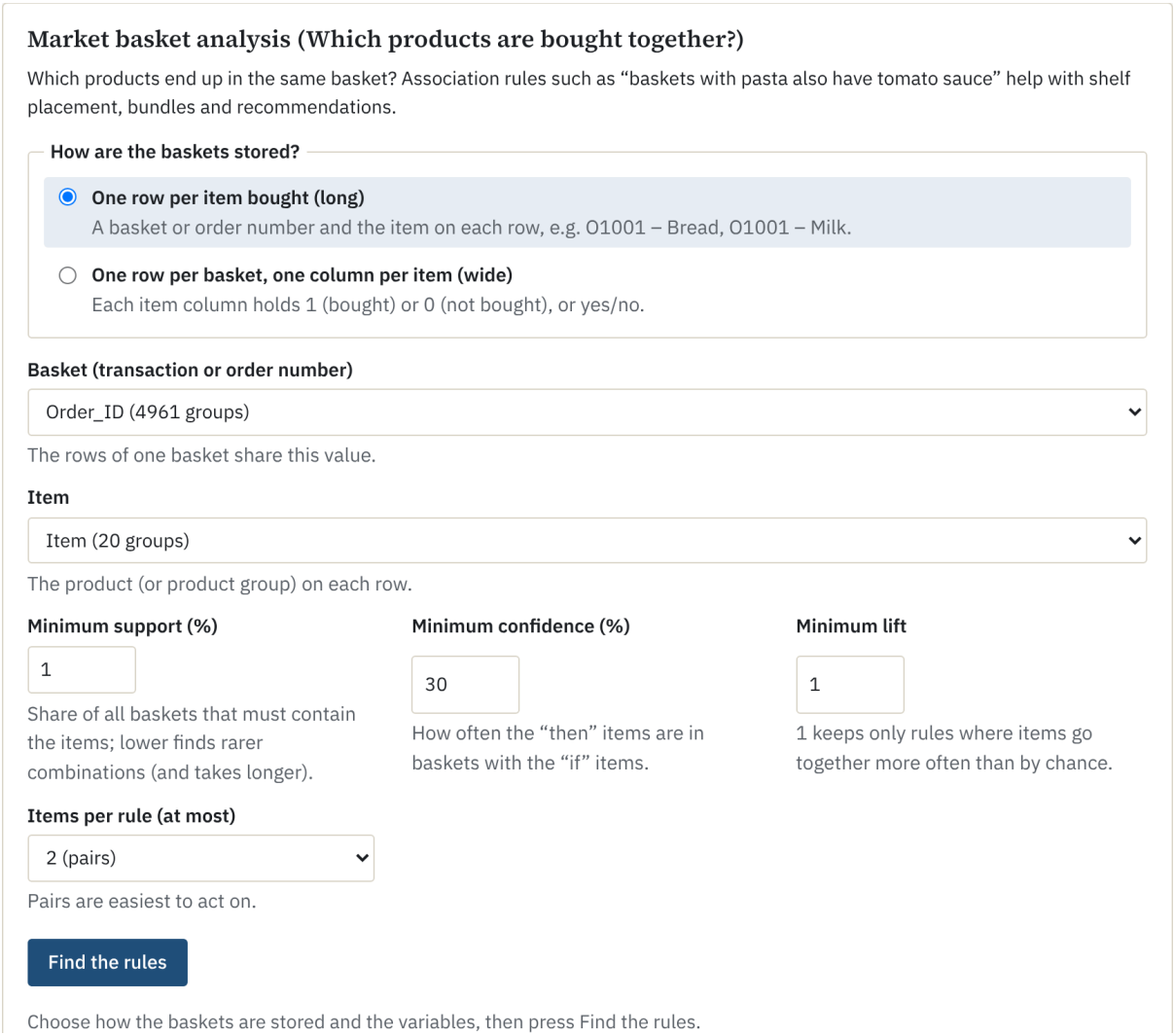


Figure 84. *Market basket analysis: the basket and the item of every row, and the thresholds.*

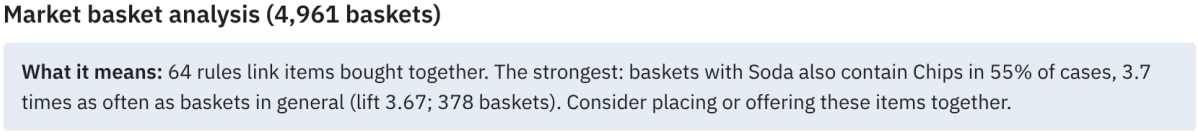


Figure 85. *Market basket analysis: what it means and the checks.*

What it means for our example

SaNaStat writes: "64 rules link items bought together. The strongest: baskets with Soda also contain Chips in 55% of cases, 3.7 times as often as baskets in general (lift 3.67; 378 baskets). Consider placing or offering these items together." The checks confirm that every rule shown rests on at least 20 baskets, and add: "A rule shows items that are bought together; it does not show that offering one makes people buy the other. Test a promotion with an A/B test."

The first table lists the most frequent items (Figure 86): Milk is in 43.2% of the baskets, Bread in 38.1%, Butter in 30.7%. The second lists the rules (Figure 87); **Sort the rules by** and **Rules shown** change the order and the number of rules shown without searching again.

Most frequent items

Item	Baskets	Support
Milk	2,144	43.2%
Bread	1,891	38.1%
Butter	1,523	30.7%
Bananas	1,372	27.7%
Eggs	1,345	27.1%
Apples	1,251	25.2%
Pasta	1,136	22.9%
Coffee	1,112	22.4%
Tomato sauce	1,000	20.2%
Cheese	950	19.1%
Rice	915	18.4%
Chicken	890	17.9%
Sugar	853	17.2%
Yoghurt	852	17.2%
Tea	796	16.0%

Support = share of the 4,961 baskets that contain the item. Items below the minimum support (1.0%) are not shown.

Figure 86. *The most frequent items and their support.*

Association rules (top 20 of 64, sorted by lift)

If a basket has …	… it also has	Baskets	Support	Confidence	Lift	Leverage	Conviction
Soda	Chips	378	7.6%	54.5%	3.67	0.0554	1.87
Chips	Soda	378	7.6%	51.2%	3.67	0.0554	1.76
Tomato sauce	Pasta	715	14.4%	71.5%	3.12	0.0980	2.71
Pasta	Tomato sauce	715	14.4%	62.9%	3.12	0.0980	2.15
Sugar	Coffee	556	11.2%	65.2%	2.91	0.0735	2.23
Coffee	Sugar	556	11.2%	50.0%	2.91	0.0735	1.66
Butter	Bread	1,151	23.2%	75.6%	1.98	0.1150	2.53
Bread	Butter	1,151	23.2%	60.9%	1.98	0.1150	1.77
Cheese	Bananas	312	6.3%	32.8%	1.19	0.0099	1.08
Eggs	Bread	577	11.6%	42.9%	1.13	0.0130	1.08
Bread	Eggs	577	11.6%	30.5%	1.13	0.0130	1.05
Soda	Bread	297	6.0%	42.9%	1.12	0.0066	1.08
Apples	Bananas	388	7.8%	31.0%	1.12	0.0085	1.05
Cheese	Milk	454	9.2%	47.8%	1.11	0.0088	1.09
Chicken	Butter	301	6.1%	33.8%	1.10	0.0056	1.05
Chicken	Bananas	270	5.4%	30.3%	1.10	0.0048	1.04
Bananas	Bread	571	11.5%	41.6%	1.09	0.0097	1.06
Bread	Bananas	571	11.5%	30.2%	1.09	0.0097	1.04
Yoghurt	Bananas	257	5.2%	30.2%	1.09	0.0043	1.04
Tea	Milk	375	7.6%	47.1%	1.09	0.0062	1.07

Support: share of all baskets with every item of the rule. Confidence: share of baskets with the “if” items that also have the “then” items. Lift: confidence divided by how common the “then” items are (above 1 = bought together more often than by chance). Leverage: support minus the support expected if unrelated. Conviction: how much more often the rule would be wrong if the items were unrelated (∞ = never wrong). Minimum support 1.0%, confidence 30%, lift 1.00; at most 2 items per rule.

Figure 87. *The association rules sorted by lift (the top 20 of 64).*

Table 36

The strongest rules by lift (4,961 baskets)

If a basket has …	… it also has	Baskets	Support	Confidence	Lift
Soda	Chips	378	7.6%	54.5%	3.67
Chips	Soda	378	7.6%	51.2%	3.67
Tomato sauce	Pasta	715	14.4%	71.5%	3.12
Pasta	Tomato sauce	715	14.4%	62.9%	3.12
Sugar	Coffee	556	11.2%	65.2%	2.91
Coffee	Sugar	556	11.2%	50.0%	2.91
Butter	Bread	1,151	23.2%	75.6%	1.98
Bread	Butter	1,151	23.2%	60.9%	1.98
Cheese	Bananas	312	6.3%	32.8%	1.19

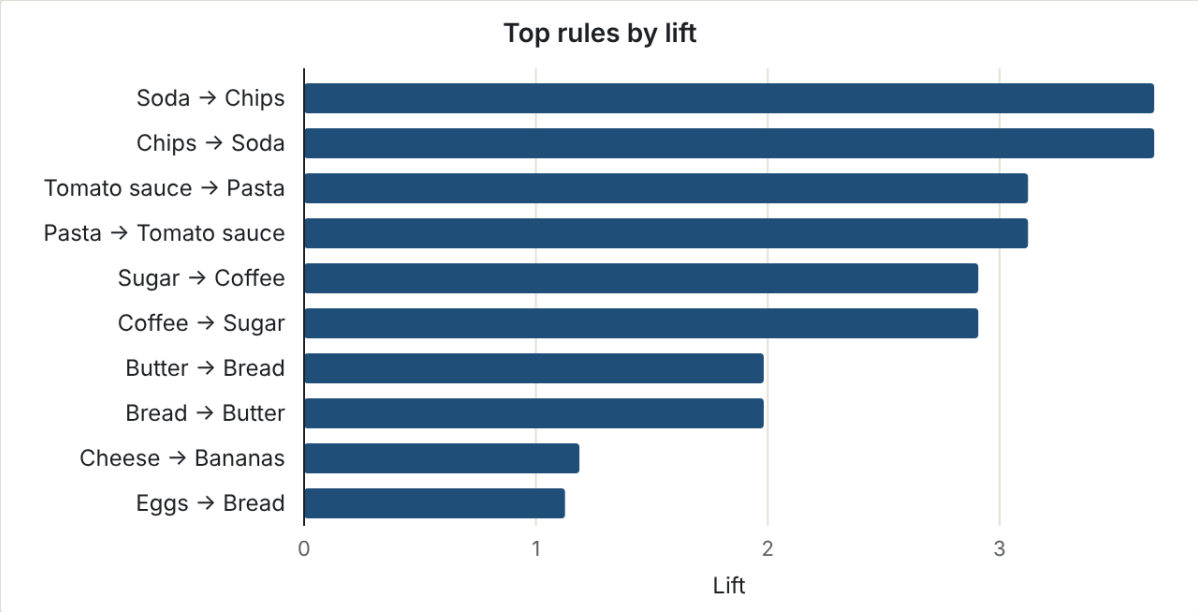


Figure 88. *The top 10 rules by lift. Read "A → B" as: baskets with A also tend to contain B.*

What it means for our example

Four pairs stand out with a lift between 1.98 and 3.67: chips and soda, pasta and tomato sauce, coffee and sugar, bread and butter. After them the lift drops at once to 1.19 and below: the other 56 rules are items that are a little more likely together than by chance. Each pair appears twice, once in each direction, with the same lift (lift is symmetric) but a different confidence: 75.6% of the baskets with butter have bread, but only 60.9% of the baskets with bread have butter, because bread is more common. Bread and butter is the most common pair (support 23.2%, 1,151 baskets); chips and soda the strongest link.

Sort the rules by confidence and a different picture appears: after the four pairs come rules such as Cheese → Milk (confidence 47.8%) and Tea → Milk (47.1%), with a lift of only 1.11 and 1.09. They reflect how common milk is, not a link between the items: this is why the minimum lift is 1 by default and why rules should be read by lift.

How to report it

"An Apriori analysis (Agrawal & Srikant, 1994) of 4,961 transactions (20 items, M = 4.2 items per basket) with minimum support 1.0% and minimum confidence 30% found 209 frequent itemsets (189 with two or more items) and 64 association rules with lift ≥ 1.00. The rule Soda → Chips had support = .076, confidence = .55, lift = 3.67."

How to report it (APA style): An Apriori analysis (Agrawal & Srikant, 1994) of 4,961 transactions (20 items, M = 4.2 items per basket) with minimum support 1.0% and minimum confidence 30% found 209 frequent itemsets (189 with two or more items) and 64 association rules with lift ≥ 1.00. The rule Soda → Chips had support = .076, confidence = .55, lift = 3.67.

Copy this sentence | All tables (Excel) | Add to Word report

Figure 89. *The sentence for the report, with its buttons.*

9.3 Thresholds and longer rules

- **Lower thresholds** find more rules, mostly weak ones. With the minimum confidence and the minimum lift at 0, SaNaStat finds 378 rules from the same 209 frequent itemsets.
- **Longer rules.** With **Items per rule (at most)** at 3, SaNaStat finds 745 frequent itemsets and 755 rules; the strongest is Bread + Chips → Soda (support .031, confidence .57, lift 4.09, 152 baskets). Longer rules rest on fewer baskets and are harder to act on; start with pairs.
- **Lower support** finds rarer combinations and takes longer. A rule backed by a handful of baskets is easily a coincidence; SaNaStat warns when a rule rests on fewer than 20 baskets.

Cut-off values: rules worth acting on

Lift above 1: the items are bought together more often than if they were unrelated; a rule with high confidence but lift near 1 only reflects a popular item. Prefer rules backed by enough baskets (support) (Agrawal et al., 1993; Brin et al., 1997).

Note: what a rule cannot tell you

A rule describes what was bought together, not why. Chips and soda may be bought together because they are placed together already, or because both are bought for parties. Whether a bundle offer or a recommendation raises sales is a causal question: test it with an A/B test (chapter 4).

Exercises

Exercise: baskets

1. Sort the rules by support. Which rule is the most common, and is its lift high? Answer: Butter → Bread and Bread → Butter (support 23.2%), lift 1.98; the next rules, Bread → Milk and Milk → Bread (17.4%), have a lift of only 1.06.

2. Planted answer for the chapter: Bread → Butter, Pasta → Tomato sauce, Coffee → Sugar and Chips → Soda were built into the data, and they are the only rules with a lift above 1.2.

10. Conjoint analysis: which product features matter?

Customers rarely buy a product for one feature. *Conjoint analysis* shows people complete product descriptions (*profiles*) that combine several features (*attributes*, such as brand and price) at different *levels* (such as $49, $79 and $109), asks them to rate each profile, and works out from the ratings how much each level adds to the liking: its *part-worth utility* (Green & Srinivasan, 1978, 1990). This chapter uses *practice_conjoint.csv*: 150 simulated respondents each rated the same 12 profiles of wireless earbuds on a 1 to 10 scale.

10.1 The data

On the **Data** screen open practice_conjoint.csv (start a new project). SaNaStat reports 1800 respondents and 7 variables: here a "respondent" is one rating, and each person has 12 rows. Card 2 counts *2 scale · 5 categorical*: Respondent (R001 to R150), Brand (Auralis, Beatwave, Cirrus), Price ($49, $79, $109), Battery (6 hours, 10 hours) and Noise_cancelling (No, Yes) are categorical; Profile (the number of the profile, 1 to 12) and Rating are scale. The brands are invented.

Concept: part-worths and importance

A *part-worth* is the utility of one level of an attribute. SaNaStat estimates it by regression with *effects coding*: the part-worths of each attribute add up to 0, so a positive value means "liked more than the average level of this attribute". The *importance* of an attribute is the range of its part-worths (best level minus worst) as a share of the sum of the ranges of all attributes. SaNaStat fits one regression for all ratings together (the *aggregate* part-worths) and one for each respondent, and averages the importance over the respondents.

Cut-off values: the design

Each respondent must rate more profiles than there are part-worths to estimate, and the profiles must vary every attribute independently (a balanced, orthogonal design); otherwise individual part-worths cannot be separated (Green & Srinivasan, 1978, 1990). Here there are 7 parameters (the constant, 2 for Brand, 2 for Price, 1 for Battery and 1 for Noise_cancelling) and 12 profiles per respondent.

10.2 Estimate the part-worths

1. Choose **Analyses** → **Conjoint analysis**.
2. Choose **Respondent** in **Respondent**.

3. Under *Attributes (the features that vary between profiles)* tick Brand (3 levels), Price (3 levels), Battery (2 levels) and Noise_cancelling (2 levels).
4. Choose **Rating** in **Rating (scale)** (Figure 90). The result appears at once.

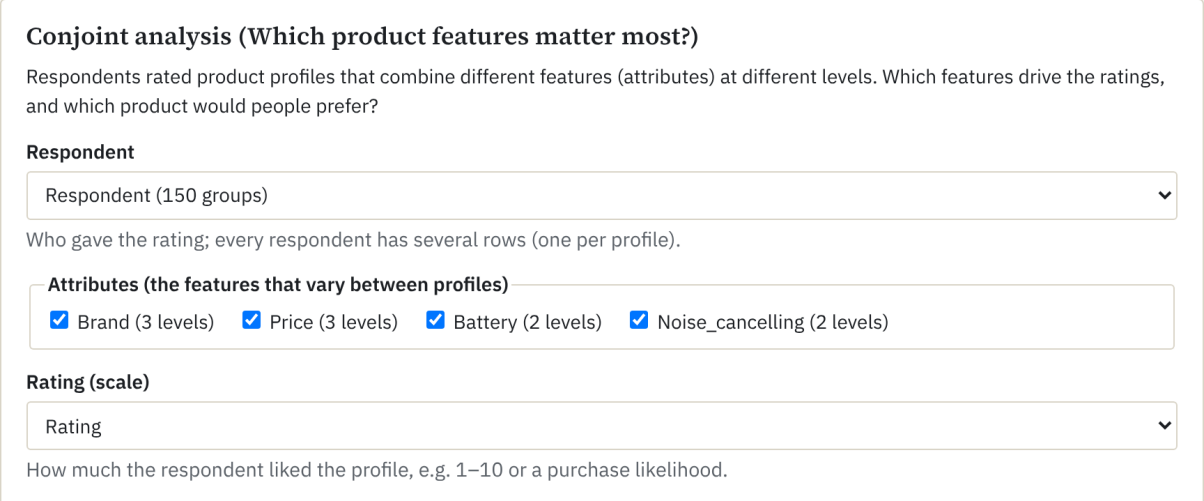


Figure 90. *Conjoint analysis: the respondent, the four attributes and the rating.*

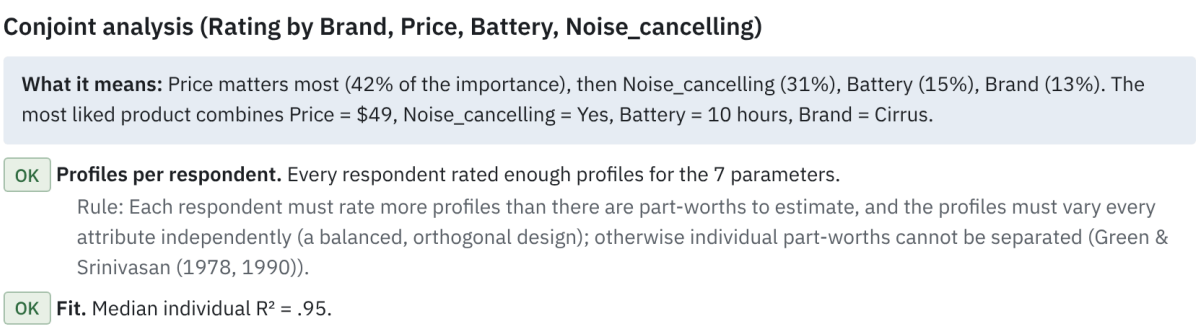


Figure 91. *Conjoint analysis: what it means and the checks.*

What it means for our example

SaNaStat writes: "Price matters most (42% of the importance), then Noise_cancelling (31%), Battery (15%), Brand (13%). The most liked product combines Price = $49, Noise_cancelling = Yes, Battery = 10 hours, Brand = Cirrus." The checks confirm that every respondent rated enough profiles for the 7 parameters, and that the model fits well: median individual R^2 = .95.

Table 37

Attribute importance (150 respondents)

Attribute	Importance	SD across respondents	From the aggregate part-worths	Preferred level
Price	41.9%	16.2%	48.8%	$49
Noise_cancelling	30.5%	13.8%	34.1%	Yes
Battery	14.7%	7.3%	16.7%	10 hours
Brand	12.8%	6.5%	0.5%	Cirrus

Table 38

Part-worth utilities (all respondents together, effects coding)

Attribute	Level	Part-worth	SE	95% CI	p
Brand	Auralis	−0.010	0.042	[−0.093, 0.073]	.813
Brand	Beatwave	−0.008	0.038	[−0.083, 0.066]	.827
Brand	Cirrus	0.018	0.038	[−0.056, 0.093]	.631
Price	$49	1.405	0.038	[1.330, 1.480]	< .001
Price	$79	0.220	0.038	[0.145, 0.295]	< .001
Price	$109	−1.624	0.042	[−1.707, −1.541]	< .001
Battery	6 hours	−0.518	0.026	[−0.569, −0.467]	< .001
Battery	10 hours	0.518	0.026	[0.467, 0.569]	< .001
Noise_cancelling	No	−1.059	0.026	[−1.110, −1.008]	< .001
Noise_cancelling	Yes	1.059	0.026	[1.008, 1.110]	< .001

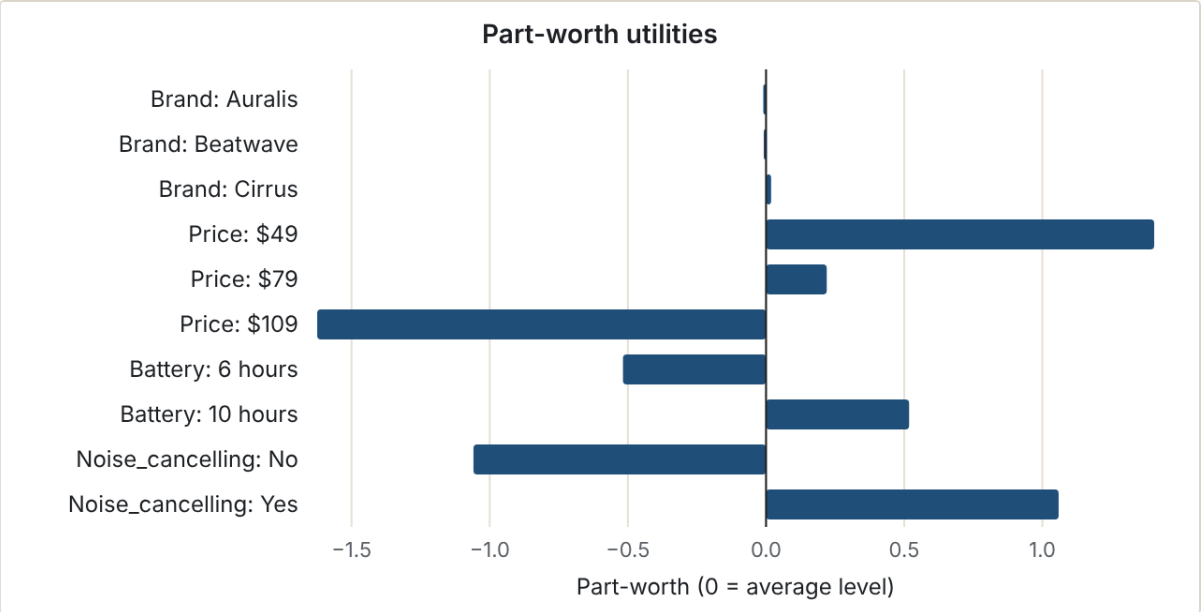


Figure 92. *Part-worth utilities: bars to the right mark a level liked more than the average level of its attribute.*

What it means for our example

The average rating is 5.481 (the constant). A price of $49 adds 1.405 points to the rating and a price of $109 takes 1.624 away, so price spans the widest range of part-worths; noise cancelling moves the rating from −1.059 to 1.059, and 10 hours of battery instead of 6 from −0.518 to 0.518. The three brands hardly differ (all part-worths within ±0.02, p > .6): across all respondents, the brand does not matter. Yet the importance of Brand averaged over the respondents is 12.8%, against 0.5% from the aggregate part-worths. Some respondents prefer one brand and some another, so their preferences cancel out in the aggregate; and each respondent's ratings also hold some noise, which always gives a small range. Read both columns: Price, Noise_cancelling and Battery matter in both, Brand only to individuals.

Note: the p values are optimistic

The regression treats the 1,800 ratings as independent, but the 12 ratings of one person are not, so the standard errors are too small and the p values optimistic. SaNaStat says so under the table. Rely on the size of the part-worths, the importances and the individual results rather than on the p values.

How to report it

"A ratings-based conjoint analysis (N = 150 respondents, 1800 ratings) estimated part-worth utilities by ordinary least squares with effects coding (aggregate R^2 = .71; median individual R^2 = .95). Relative importances were Price 41.9%, Noise_cancelling 30.5%, Battery 14.7%, Brand 12.8%."

10.3 The market simulator

Which of a few products would people prefer? The section *Market simulator* under the result compares 2 to 4 products that you define from the levels of the attributes (Figure 93).

1. Choose 3 in **Number of products**.
2. Name the products and choose a level of every attribute for each: **Budget** (Cirrus, $49, 6 hours, no noise cancelling), **Premium** (Auralis, $109, 10 hours, noise cancelling) and **Mid-range** (Beatwave, $79, 10 hours, noise cancelling).
3. Keep the **Logit scale** at 1.

Market simulator

Define the products to compare: choose a level of every attribute. The shares show which product respondents would prefer among these.

Number of products

3

Logit scale

1

1 is usual; larger values make the shares closer to first choice.

Budget

Name: Budget

Brand: Cirrus

Price: $49

Battery: 6 hours

Noise_cancelling: No

Premium

Name: Premium

Brand: Auralis

Price: $109

Battery: 10 hours

Noise_cancelling: Yes

Mid-range

Name: Mid-range

Brand: Beatwave

Price: $79

Battery: 10 hours

Noise_cancelling: Yes

Market simulator (3 products)

Figure 93. *The market simulator with three products.*

Market simulator (3 products)

What it means: Mid-range would win the largest share of preference: 63% by the logit rule and 76% by first choice.

Info **What the shares mean.** Shares of preference are not market shares: they assume everyone knows and can buy every product. Use them to compare designs.

Simulated shares of preference

Product	Levels	Utility	First choice	Logit share
Budget	Brand Cirrus, Price $49, Battery 6 hours, Noise_cancelling No	5.33	14.3%	19.8%
Premium	Brand Auralis, Price $109, Battery 10 hours, Noise_cancelling Yes	5.42	9.3%	17.2%
Mid-range	Brand Beatwave, Price $79, Battery 10 hours, Noise_cancelling Yes	7.27	76.3%	63.0%

Based on 150 respondents' own part-worths. First choice: each respondent “buys” the product with the highest utility (ties shared). Logit: each respondent spreads the choice by e^(scale × utility); averaged over respondents. Utility shown: aggregate part-worths. Shares are preferences among these products only; they ignore price elasticity outside the tested range, awareness and distribution.

Figure 94. *The simulated shares of preference of the three products.*

Table 39

Simulated shares of preference (150 respondents' own part-worths)

Product	Utility	First choice	Logit share
Budget	5.33	14.3%	19.8%
Premium	5.42	9.3%	17.2%
Mid-range	7.27	76.3%	63.0%

Concept: first choice and the logit rule

The simulator uses each respondent's own part-worths. By the *first-choice* rule, every respondent "buys" the product with the highest utility. By the *logit* rule, every respondent spreads the choice over the products in proportion to e^(scale × utility), so that a product that is nearly as good still gets some share. A larger **Logit scale** brings the logit shares closer to the first-choice shares: at scale 2 Mid-range gets 71.2% instead of 63.0%.

What it means for our example

Mid-range would win the largest share of preference: 63.0% by the logit rule and 76.3% by first choice. It combines the two features respondents value most after price, at a middle price. Premium has the same features but the highest price, and Budget the lowest price without them; both get under 20%. The simulator answers what-if questions: at $49, Mid-range would reach 82.7% (first choice 96.0%); without noise cancelling at $79 it would fall to 24.3%, and Premium would lead with 40.9%.

Note: shares of preference are not market shares

Shares of preference assume that everyone knows every product and can buy it, and they ignore prices outside the tested range, awareness and distribution. Use them to compare designs, not to forecast sales. SaNaStat says the same under the table.

How to report it

"A share-of-preference simulation based on 150 respondents' own part-worths gave Budget 19.8%, Premium 17.2%, Mid-range 63.0% (logit rule, scale = 1.0); first-choice shares were Budget 14.3%, Premium 9.3%, Mid-range 76.3%."

Exercises

Exercise: conjoint analysis

1. Change the price of Mid-range to $49 in the simulator. How much does its logit share rise? Answer: from 63.0% to 82.7%.

2. Why is the importance of Brand 12.8% when the aggregate part-worths of the brands are almost 0? (Section 10.2.)

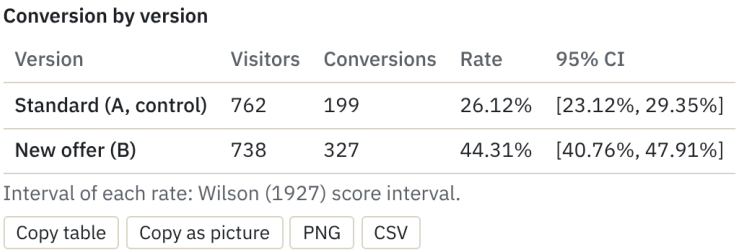
3. Planted answer for the chapter: price matters most, then noise cancelling, then battery, and brand least, with a price-sensitive and a quality-seeking group of respondents (the large SD of the importance of Price, 16.2%, reflects them).

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, how to build the report, and how to turn the results of this manual into a short report for a manager.

11.1 Copying and exporting one table or chart

Every table of a result has a row of buttons under it (Figure 95): **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 96), **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.



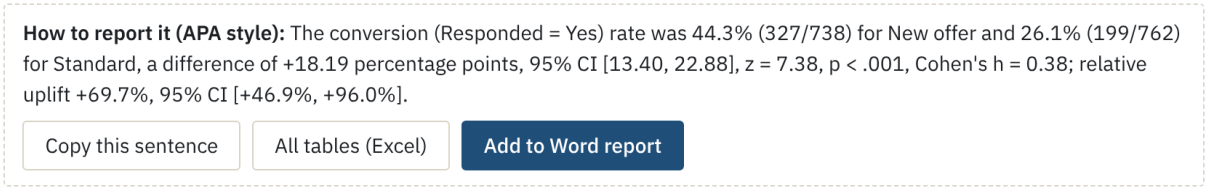
Conversion by version

Version	Visitors	Conversions	Rate	95% CI
Standard (A, control)	762	199	26.12%	[23.12%, 29.35%]
New offer (B)	738	327	44.31%	[40.76%, 47.91%]

Interval of each rate: Wilson (1927) score interval.

Copy table | Copy as picture | PNG | CSV

Figure 95. *The table of the A/B test with its buttons: Copy table, Copy as picture, PNG and CSV.*



How to report it (APA style): The conversion (Responded = Yes) rate was 44.3% (327/738) for New offer and 26.1% (199/762) for Standard, a difference of +18.19 percentage points, 95% CI [13.40, 22.88], z = 7.38, p < .001, Cohen's h = 0.38; relative uplift +69.7%, 95% CI [+46.9%, +96.0%].

Copy this sentence | All tables (Excel) | Add to Word report

Figure 96. *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 four results: the A/B test of Responded (section 4.1), the A/B test of Spend_change (section 4.4), the logistic regression of Churned (section 6.2) and the box plot of Annual_spend by Loyalty_member (section 2.3).

Click **Word report** in the menu (it shows the number of results, here **Word report (4)**) to open the report builder (Figure 97). The first card, *Word report (4 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.

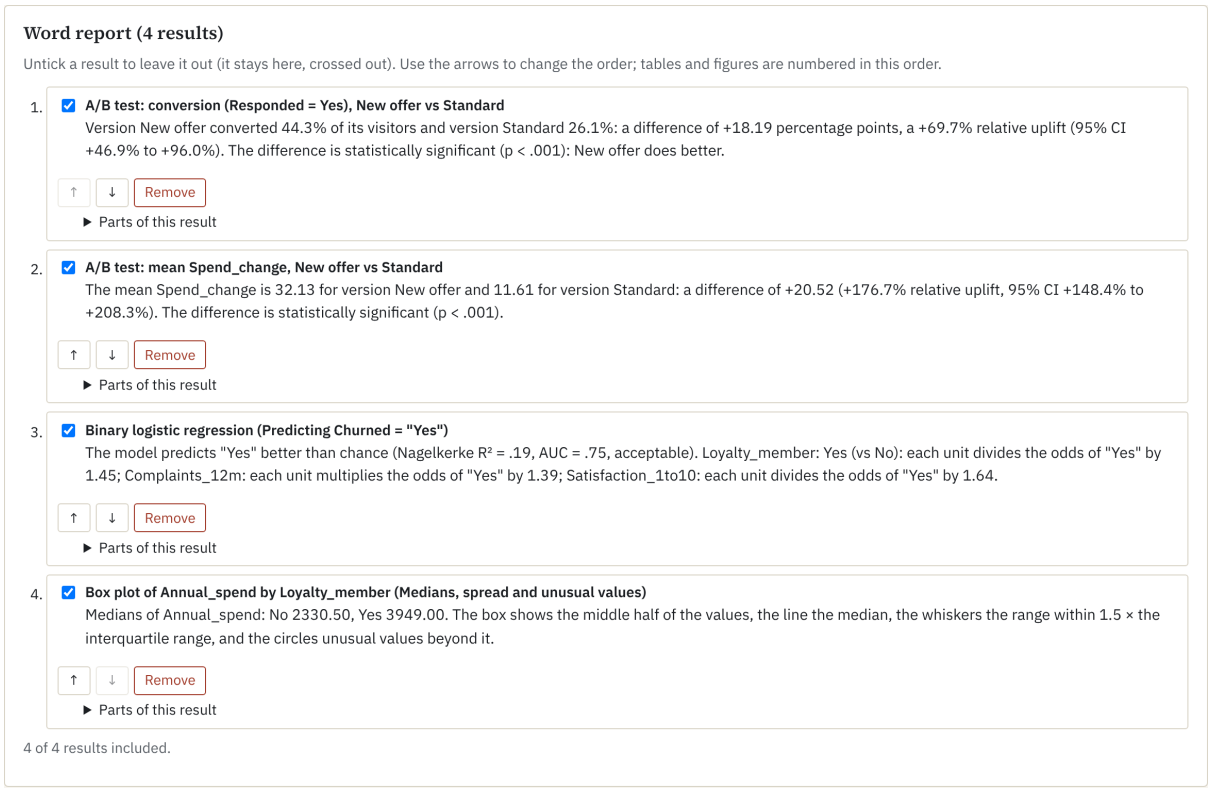


Figure 97. *The report builder with the four results of this chapter.*

The cards below build the report as a whole:

- **What goes in** (Figure 98): 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 "Campaign and churn: a worked mini-report"), the font, the **Line spacing**, the **Paper** (A4 or US Letter) and an optional **Title page**.
- *Preview and download* (Figure 99): **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.

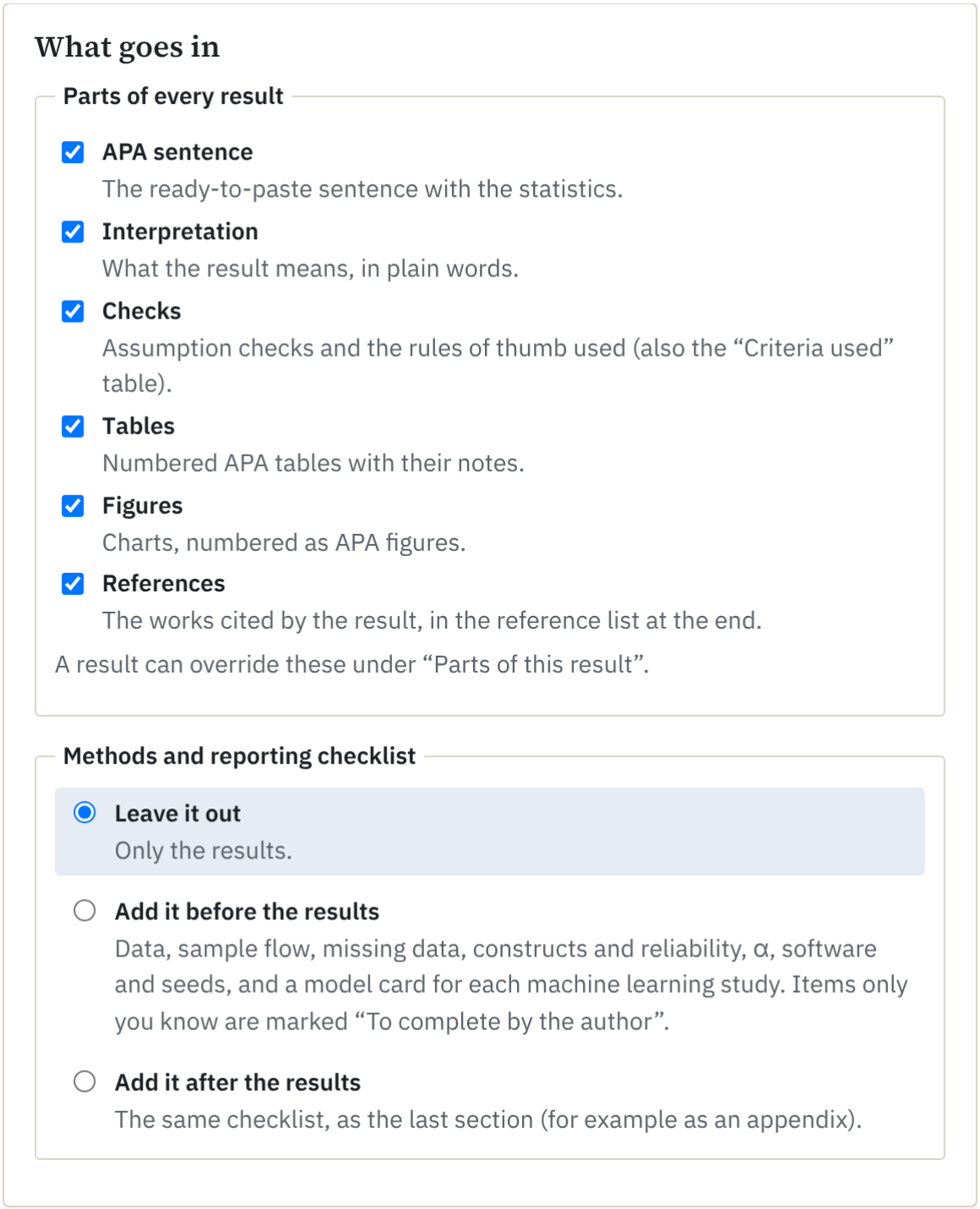


Figure 98. *What goes in: the parts of every result and the methods and reporting checklist.*

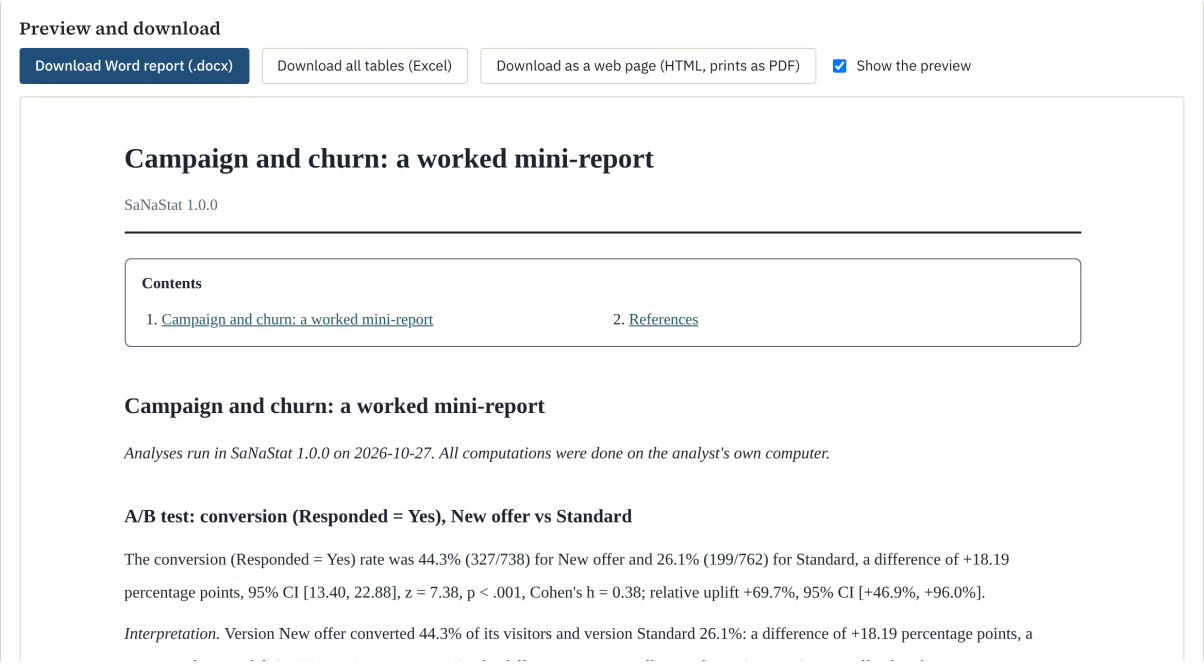


Figure 99. *Preview and download, with the start of the preview.*

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*.

Note: the document is a draft

The report documents the analyses; it does not interpret them for your audience. Add the context, the limitations and the recommendations in Word. In particular state which customers were left out of each analysis, that the practice data are simulated (if you use them), and which variables were chosen on theory and which by the data.

11.3 A worked mini-report

Imagine a retailer that tested a new offer and wants to know four things: did the new offer work, who is at risk of leaving, which kinds of customers it has, and where the value lies. A plan with the tools of this manual:

Table 40

From marketing questions to SaNaStat screens

Question	Analysis	Section
Did the New offer raise the response rate?	A/B test of Responded by Campaign_group	4.1
Did it raise spend more than the Standard offer?	A/B test of the mean of Spend_change	4.4

Question	Analysis	Section
How many customers does the next test need?	The planner of the A/B test	4.6
Who is likely to leave, and why?	Logistic regression of Churned; machine learning comparison; explanations; calibration; cut-offs; fairness	6.2 to 6.9
Which kinds of customers do we have?	Clustering on behaviour; churn and region by segment	7.2 and 7.3
Which customers bring the revenue, and which are slipping away?	RFM segments and customer lifetime value (on the order records)	8.2 to 8.4

The paragraph below is built only from numbers of this manual.

How to report it

"The new offer raised the response rate from 26.1% (199 of 762 customers) to 44.3% (327 of 738), a difference of +18.19 percentage points, 95% CI [13.40, 22.88], z = 7.38, p < .001. Monthly spend rose by 32.13 after the new offer and by 11.61 after the standard offer, a difference of 20.52, 95% CI [17.71, 23.33], Welch's t(863.00) = 14.33, p < .001, d = 0.75. Churn (16.7% of customers) was predicted by low satisfaction (OR = 0.61 per point), more complaints (OR = 1.39 per complaint) and non-membership of the loyalty programme (OR for members = 0.69); the four-predictor model reached AUC = .754 in the sample, and in cross-validation on customers not used to fit it a logistic regression with 13 predictors reached AUC = .755. Flexible machine learning models did not outperform logistic regression. On cross-validated predictions the logistic regression at a cut-off of 0.50 found 13% of the leavers (precision 0.50), whereas at a cut-off of 0.20 it found 60% of them (precision 0.33), so the cut-off for a retention campaign should be chosen from the costs of contact and of losing a customer."

(The AUC of .754 belongs to the logistic regression of section 6.2, fitted to the 1415 customers with a satisfaction rating. The AUC of .755 is the average over the five test folds of the machine learning comparison, which used 13 predictors and all 1500 customers.)

11.4 A checklist before you send the report

- State the **sample**: how many customers, how many were left out of each analysis, and why.
- Report **effect sizes and intervals**, not only p-values, and the rates behind a test (26.1% against 44.3%), with the uplift and its interval.
- For an A/B test, report the **planned sample size** and the sample-ratio check, and do not stop a test early.

- State the **checks** that failed and what you did (for example Welch's test because the variances differed; the logarithm because of skew; zero-spend customers excluded from the logarithmic model).
- For predictions, report **performance on unseen customers** and the **baseline** (the null model), do not report accuracy alone for a rare outcome, state the **cut-off** with the recall and precision it gives, and check **fairness** by group.
- Say that explanations (SHAP, ALE), association rules and conjoint shares describe patterns, **not causes**.
- Describe **segments** (clusters or RFM) as descriptions of this customer base, with the silhouette and the stability for clusters.
- Protect **individual customers**: report groups, not named people, and hide small groups.
- If the data are simulated or a teaching file, say so.

Exercise

Exercise: your own mini-report

Add the results of the A/B tests of sections 4.1 and 4.4, the logistic regression of section 6.2, the cluster analysis of section 7.2 and the RFM analysis of section 8.2 to the Word report, put them in a sensible order, add the methods and reporting checklist after the results, download the report and write two sentences of recommendation under each result in Word.

Appendix A. Codebooks

All five files are simulated; no real person, shop or brand appears in them. Missing values are empty cells.

A.1 practice_marketing.csv (1,500 customers)

Table 41

practice_marketing.csv: variables

Variable	Meaning
CustomerID	Identifier C0001 to C1500 (type ID; never analysed)
Age, Gender, Region	Customer profile; Gender is female, male or other (30 customers); Region is Central, East, North, South or West
Income_k	Household income in thousands (81 missing; right-skewed)
Tenure_months	Months as a customer
Loyalty_member	Yes or No (646 Yes)
Main_channel	Store, Online or Mobile app
Visits_per_month, Avg_basket, Discount_share	Shopping behaviour (visits per month, average basket value, share of spend bought on discount)
Complaints_12m	Count of complaints in the last 12 months
Satisfaction_1to10, Recommend_0to10	Survey ratings (85 and 65 missing)
Annual_spend	Spend in the year (right-skewed; 82 customers spent 0)
Campaign_group	Standard (762) or New offer (738), randomly assigned
Responded	Yes or No: responded to the campaign (526 Yes)
Monthly_spend_before, Monthly_spend_after	Monthly spend around the campaign (paired)
Churned	Yes or No: left within the year (250 Yes, 16.7%)

Table 42

What was planted in practice_marketing.csv, and where this manual finds it

Planted effect	Result in this manual	Section
New offer responds more than Standard (about 44% and 26%)	44.3% against 26.1%; uplift +18.19 points, z = 7.38	4.1
Monthly spend rises after the campaign, more after the New offer	Paired t(1499) = 28.78; difference of the changes 20.52	4.3, 4.4
Annual_spend depends on income, loyalty, satisfaction, visits	R^2 = .806 (raw), .415 (Annual_spend_ln1p), .773 (Annual_spend_ln)	5.1, 5.2

Planted effect	Result in this manual	Section
Loyalty members spend more	d = −0.79, medians 3949.00 against 2330.50	5.4
Churn depends on satisfaction, complaints, tenure and loyalty	OR 0.61, 1.39, 0.992, 0.69; AUC .754	6.2
Response depends on the offer, loyalty and channel	Logistic regression and machine learning of Responded	6.10
Customers can be segmented on behaviour	Silhouette .18: no substantial structure	7.2

A.2 practice_brand_survey.csv (500 respondents)

Table 43
practice_brand_survey.csv: variables

Variable	Meaning
RespondentID	Identifier (type ID)
Age_group, Gender, Customer_type	18-24, 25-34, 35-49 or 50+; female or male; Long-term, New or Regular
Trust1 to Trust4	Trust items, 1 to 7
Value1 to Value4	Perceived value items, 1 to 7 (Value2 has 18 missing)
Loyalty1 to Loyalty3	Loyalty items, 1 to 7 (Loyalty2 has 18 missing)

Ten respondents answered every item identically (straight-lining); see section 3.1.

A.3 segments_demo.csv (360 shoppers)

Table 44
segments_demo.csv: variables

Variable	Meaning
ID	Identifier
Price_sensitivity, Brand_loyalty, Quality_focus, Online_affinity, Social_influence	Attitude scores, 1 to 7
Age, Income_k, Gender, Channel	Profile; Channel is Mixed, Online or Store

Three segments were planted; see section 7.1.

A.4 practice_transactions.csv (21,006 order lines)

One row per item bought in an online grocery shop from 2023-01-04 to 2025-12-30: 4,961 orders by 900 customers, with 20 items.

Table 45
practice_transactions.csv: variables

Variable	Meaning
Order_ID	Order number, O00001 upwards; all rows of one order share it (4,961 orders)
Customer_ID	Customer number, C0001, C0002 and so on (900 customers)
Order_date	Date of the order, written as year-month-day (2023-01-04 to 2025-12-30)
Item	The product: Apples, Bananas, Bread, Butter, Cheese, Chicken, Chips, Chocolate, Coffee, Detergent, Eggs, Milk, Pasta, Rice, Shampoo, Soda, Sugar, Tea, Tomato sauce or Yoghurt
Line_amount	Money paid for this item in this order (price × quantity, with small variations)

Planted on purpose: four pairs of items are bought together (Bread and Butter, Pasta and Tomato sauce, Coffee and Sugar, Chips and Soda; chapter 9), and the customers are a mix of loyal heavy buyers, regular buyers, one-off buyers and customers who stop buying (chapter 8).

A.5 practice_conjoint.csv (150 respondents, 1,800 ratings)

Each respondent rated the same 12 profiles of wireless earbuds, in a random order. The 12 profiles form a balanced design of the four attributes.

Table 46
practice_conjoint.csv: variables

Variable	Meaning
Respondent	Respondent R001 to R150 (12 rows each)
Profile	Number of the profile rated, 1 to 12
Brand	Auralis, Beatwave or Cirrus (invented brands)
Price	$49, $79 or $109
Battery	6 hours or 10 hours
Noise_cancelling	No or Yes
Rating	How much the respondent liked the profile, 1 (not at all) to 10 (very much)

Planted on purpose: price matters most, then noise cancelling, then battery life, and brand least; about half of the respondents are price-sensitive and the others seek quality (noise cancelling); see chapter 10.

Appendix B. Which tool answers which marketing question?

Table 47

Where to find each marketing question

Marketing question	Tool in SaNaStat	Section
Who are my customers? How are spend, tenure and satisfaction distributed?	Analyses → Descriptive statistics; Charts → Histogram, Box plot	2.1 to 2.3
Do groups differ on a measure? (channels, regions, members)	Charts → Means with confidence intervals; Analyses → t-tests, One-way ANOVA	2.4, 5.4, 5.5
Which measures go together?	Charts → Correlation heatmap, Scatter plot; Analyses → Correlation and regression	2.5
How are customers split across categories?	Charts → Bar chart of percentages	2.6
How do I take the logarithm of spend, or make low and high groups, a 0/1 indicator or merged categories?	Data → 4 · Compute and recode variables	1.6, 2.6, 2.7
How do I analyse only some customers?	Data → 7 · Select cases	1.7
Is my survey scale reliable? Is someone straight-lining?	Data → Build construct scores; Data quality screening	3.1, 3.2
Do the items measure what they should? Is there mediation or moderation?	Analyses → Factor analysis and reliability; Mediation and moderation (Common Statistical Tests manual)	3.5
What drives loyalty or satisfaction?	Regression of the scale score	3.4
Did the offer change the conversion rate, and by how much?	Analyses → A/B test (a yes/no result)	4.1, 4.5
Did spend change after a campaign?	Paired t-test; A/B test of the mean change	4.3, 4.4
How many customers does a test need?	A/B test → Plan a test; Power analysis	4.6
What drives spend?	Multiple linear regression; hierarchical regression	5.1 to 5.3
Which customers are likely to churn or respond?	Logistic regression; Machine learning	6.2 to 6.4
What happens when probabilities become yes or no?	Machine learning → Cut-off	6.5

Marketing question	Tool in SaNaStat	Section
What drives the prediction?	Machine learning → Explanations (SHAP, ALE)	6.6
Can the predicted probabilities be trusted?	Machine learning → Calibration & intervals	6.7
Does the model work equally well for every group?	Machine learning → Fairness	6.8
How do I get a prediction for each customer?	Machine learning → Model card (predictions, scoring a new file)	6.9
Which kinds of customers do I have?	Clustering; keep the clusters; compare on churn and region	7.1 to 7.3
Which customers are most valuable, and which are slipping away? What is a customer worth?	Analyses → RFM and customer value	8.2 to 8.4
Which products are bought together?	Analyses → Market basket	9.2, 9.3
Which product features matter, and which product would people prefer?	Analyses → Conjoint analysis; Market simulator	10.2, 10.3
How do I keep and share my results?	Export buttons; Add to Word report → Word report	11.1, 11.2

Appendix C. Glossary

Table 48

Terms used in this manual

Term	Meaning
A/B test	A comparison of two versions of an offer given at random to comparable customers
ALE (accumulated local effects)	How a model's prediction changes across the range of one predictor, keeping the data realistic
Association rule	A rule "if a basket has A, it also has B", with its support, confidence and lift
AUC	Area under the ROC curve: the chance that a customer who left gets a higher predicted risk than one who stayed
Balanced accuracy	The average of recall and specificity
Calibration	Whether predicted probabilities match how often the event happens
Churn	A customer leaving within the observation period
Cohen's d, d_z	Standardised difference between two means (d_z: for paired measurements)
Cohen's h	Standardised difference between two proportions
Cohort	The customers who made their first purchase in the same year
Confidence (of a rule)	The share of baskets with the "if" items that also have the "then" items
Conjoint analysis	A method that estimates how much each level of each product feature adds to the liking of a product
Conversion rate	The share of visitors or customers who did what the version asked (bought, signed up, responded)
Cramér's V, w	Effect size for a table of counts; for a 2 × 2 table equal to φ and to Cohen's w
Cronbach's α, McDonald's ω	Measures of the internal consistency of a scale
Customer lifetime value (CLV)	The money a customer is expected to bring in the future, valued today
Cut-off (threshold)	The predicted probability above which a customer is called a case (for example a leaver)
Cross-validation	Testing a model on customers it was not fitted on, by splitting the sample into folds
Fairness by group	A comparison of a model's flags and errors across groups of customers
ID (type)	A customer or respondent number; kept but never offered to an analysis
Imbalanced outcome	An outcome in which one answer is much rarer than the other (here churn, 16.7%)
Importance (conjoint)	The range of an attribute's part-worths as a share of the sum of the ranges of all attributes
Indicator variable	A 0/1 variable that is 1 for the chosen category (section 2.7)

Term	Meaning
Leakage	A predictor that contains or is a consequence of the outcome
Lift	How much more often the items of a rule are bought together than if they were unrelated (1 = unrelated)
Median split	Two groups made at the median of a scale variable; a value equal to the cut point goes to the lower group
Null model	A model that predicts the average (or the most common answer) for everybody
Odds ratio	The factor by which the odds of an event are multiplied for one more unit of a predictor
Part-worth	The utility of one level of a product attribute in a conjoint analysis
Precision	Of the customers flagged, the share who really have the outcome
Recall (sensitivity)	Of the customers who really have the outcome, the share that was flagged
Retention rate	The share of the customers who bought in one year and bought again in the next
RFM	Recency, frequency and monetary value of a customer's purchases, scored 1 to 5
Sample ratio mismatch	Numbers of visitors in the versions of an A/B test that differ from the planned split
Share of preference	The share of respondents who would prefer a product among a set, from the conjoint part-worths
SHAP	A method that splits a prediction into the contributions of the predictors
Silhouette	A measure of how well a customer fits its own cluster compared with the nearest other cluster
Stability (clusters)	Agreement between the clusters on all customers and on random subsamples (adjusted Rand index)
Straight-lining	Giving the same answer to all items of a questionnaire
Support	The share of all baskets that contain the items of a rule
Uplift	The change from version A to version B, in percentage points or relative to A
Welch's test	A test of means that stays valid when the group variances differ

Appendix D. Limits and doubtful points to keep in mind

Table 49

Things to check when you work with these tools

What you see	What to do
A ratio or percentage change is missing for some customers	A customer whose B value is 0 gets no value (82 customers have a Monthly_spend_before of 0, so the percentage change exists for 1418); say how many customers were left out (section 4.4)
A missing-value code such as 99 is "found"	Check that it is a code and not a real value before **Set as missing** (section 1.5)
A median split of a scale variable	It discards information; keep the scale variable for the analysis and use the groups for reporting (section 2.7)
The even–odd consistency index is listed as "Not used"	It needs at least four scales with two or more items; with fewer, rely on the other indices and the rule of two flags
The sign of d in a t-test depends on the order of the groups	SaNaStat reports |d| and explains the sign under the effect size; read the group means and say which group is higher
An A/B test is significant after a few days	Do not stop early; run the test to the planned size and check the sample ratio (section 4.6)
The relative uplift of a mean is very large	It divides by the mean of version A; for a change score report the difference of the means (section 4.4)
The 0.5 cut-off of the logistic regression's classification table	Look at other cut-offs on the Cut-off tab of Machine learning (section 6.5)
The four-fifths check of the Fairness tab reads Check	It is a rule of thumb from employment law; decide which differences between groups are acceptable for the intended use (section 6.8)
Machine learning is slow	Use the Quick look while exploring, untick slow models and unneeded predictors
RFM counts far too many purchases	An order has several rows: choose the order number (section 8.5)
The CLV looks small or large	It is proportional to the margin and depends strongly on the retention rate; use your own figures and compare groups rather than reading one value (section 8.4)
Many association rules have a lift just above 1	They are weak; read the rules with the highest lift and enough baskets, and test any promotion (section 9.2)
Conjoint p values are all below .001	The ratings of one person are not independent, so the p values are optimistic; rely on the part-worths and importances (section 10.2)

What you see	What to do
Brand matters for individuals but not on average	Preferences that differ between respondents cancel out in the aggregate part-worths; compare the two importance columns (section 10.2)
My results are gone	Results are kept per browser; download the Word report and export the project at the end of each session

References

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. ACM SIGMOD Record, 22(2), 207–216. https://doi.org/10.1145/170036.170072

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. In Proceedings of the 20th International Conference on Very Large Data Bases (pp. 487–499). Morgan Kaufmann.

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

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

Brin, S., Motwani, R., Ullman, J. D., & Tsur, S. (1997). Dynamic itemset counting and implication rules for market basket data. ACM SIGMOD Record, 26(2), 255–264. https://doi.org/10.1145/253262.253325

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.

Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., van Smeden, M., Boulesteix, A.-L., Camaradou, J. C., Celi, L. A., Denaxas, S., Denniston, A. K., Glocker, B., Golub, R. M., Harvey, H., Heinze, G., . . . Logullo, P. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ, 385, e078378. https://doi.org/10.1136/bmj-2023-078378

Cook, R. D., & Weisberg, S. (1982). Residuals and influence in regression. Chapman & Hall.

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. Journal of Experimental Social Psychology, 66, 4–19. https://doi.org/10.1016/j.jesp.2015.07.006

Delacre, M., Lakens, D., & Leys, C. (2017). Why psychologists should by default use Welch's t-test instead of Student's t-test. International Review of Social Psychology, 30(1), 92–101. https://doi.org/10.5334/irsp.82

Fader, P. S., Hardie, B. G. S., & Lee, K. L. (2005). RFM and CLV: Using iso-value curves for customer base analysis. Journal of Marketing Research, 42(4), 415–430. https://doi.org/10.1509/jmkr.2005.42.4.415

Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). Statistical methods for rates and proportions (3rd ed.). Wiley. https://doi.org/10.1002/0471445428

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

Green, P. E., & Srinivasan, V. (1978). Conjoint analysis in consumer research: Issues and outlook. Journal of Consumer Research, 5(2), 103–123. https://doi.org/10.1086/208721

Green, P. E., & Srinivasan, V. (1990). Conjoint analysis in marketing: New developments with implications for research and practice. Journal of Marketing, 54(4), 3–19. https://doi.org/10.1177/002224299005400402

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

Gupta, S., & Lehmann, D. R. (2003). Customers as assets. Journal of Interactive Marketing, 17(1), 9–24. https://doi.org/10.1002/dir.10045

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

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

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

Kapoor, S., Cantrell, E. M., Peng, K., Pham, T. H., Bail, C. A., Gundersen, O. E., Hofman, J. M., Hullman, J., Lones, M. A., Malik, M. M., Nanayakkara, P., Poldrack, R. A., Raji, I. D., Roberts, M., Salganik, M. J., Serra-Garcia, M., Stewart, B. M., Vandewiele, G., & Narayanan, A. (2024). REFORMS: Consensus-based recommendations for machine-learning-based science. Science Advances, 10(18), eadk3452. https://doi.org/10.1126/sciadv.adk3452

Katz, D., Baptista, J., Azen, S. P., & Pike, M. C. (1978). Obtaining confidence intervals for the risk ratio in cohort studies. Biometrics, 34(3), 469–474. https://doi.org/10.2307/2530610

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

Kohavi, R., Tang, D., & Xu, Y. (2020). Trustworthy online controlled experiments: A practical guide to A/B testing. Cambridge University Press. https://doi.org/10.1017/9781108653985

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.

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. Journal of the American Statistical Association, 83(404), 1198–1202. https://doi.org/10.1080/01621459.1988.10478722

Lumley, T., Diehr, P., Emerson, S., & Chen, L. (2002). The importance of the normality assumption in large public health data sets. Annual Review of Public Health, 23, 151–169. https://doi.org/10.1146/annurev.publhealth.23.100901.140546

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems 30 (pp. 4765–4774). Curran Associates.

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

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. Psychological Methods, 17(3), 437–455. https://doi.org/10.1037/a0028085

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

Nadeau, C., & Bengio, Y. (2003). Inference for the generalization error. Machine Learning, 52(3), 239–281. https://doi.org/10.1023/A:1024068626366

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.

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

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE, 10(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). Biometrika, 52(3–4), 591–611. https://doi.org/10.1093/biomet/52.3-4.591

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B, 63(2), 411–423. https://doi.org/10.1111/1467-9868.00293

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

van der Ploeg, T., Austin, P. C., & Steyerberg, E. W. (2014). Modern modelling techniques are data hungry: A simulation study for predicting dichotomous endpoints. BMC Medical Research Methodology, 14, 137. https://doi.org/10.1186/1471-2288-14-137

Welch, B. L. (1947). The generalization of 'Student's' problem when several different population variances are involved. Biometrika, 34(1–2), 28–35. https://doi.org/10.1093/biomet/34.1-2.28

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. Journal of the American Statistical Association, 22(158), 209–212. https://doi.org/10.1080/01621459.1927.10502953