Descriptive Statistics

Electronic supplementary material
The online version of this chapter (https://doi.org/10.1007/978-3-662-56707-4_5) contains additional material that is available to authorized users. You can also download the "Springer Nature More Media App" from the iOS or Android App Store to stream the videos and scan the image containing the "Play button".

M. Sarstedt, E. Mooi, *A Concise Guide to Market Research*, Springer Texts in Business and Economics, https://doi.org/10.1007/978-3-662-56707-4_5

Learning Objectives
After reading this chapter, you should understand:
- The workflow involved in a market research study.
- Univariate and bivariate descriptive graphs and statistics.
- How to deal with missing values.
- How to transform data (*z*-transformation, log transformation, creating dummies, aggregating variables).
- How to identify and deal with outliers.
- What a codebook is.
- The basics of using SPSS.

Keywords
Acquiescence • Aggregation • Bar chart • Bivariate statistics • Box plot • Codebook • Construct score • Correlation • Covariance • Crosstabs • Data entry errors • Dummy variables • Extreme response styles • Frequency table • Histogram • Inconsistent answers • Index • Interquartile range • Interviewer fraud • Item non-response • Line chart • Listwise deletion • Little's MCAR test • Log transformation • Mean • Measures of centrality • Measures of dispersion • Median • Middle response styles • Missing (completely or not) at random • Missing data • Multiple imputation • Outliers • Pie chart • Range • Range standardization • Scale transformation • Scatter plot • Skewed data • SPSS • Standard deviation • Standardizing variables • Straight-lining • Survey non-response • Suspicious response patterns • Transforming data • Univariate statistics • Variable respecification • Variance • Workflow • *z*-standardization.

5.1 The Workflow of Data

Market research projects involving data become more efficient and effective if they have a proper **workflow**, which is a strategy to keep track of entering, cleaning, describing, and transforming data. These data may have been collected through surveys or may be secondary data (► Chap. 3). Entering, cleaning, and analyzing bits of data haphazardly is not a (good) strategy, since it increases the likelihood of making mistakes and makes it hard to replicate results. Moreover, without a good data workflow, it becomes hard to document the research process and to cooperate on projects. For example, how can you outsource the data analysis if you cannot indicate what the data are about or what specific values mean? Finally, a lack of a good workflow increases the risk of duplicating work or even losing data. In ◘ Fig. 5.1, we show the steps required to create and describe a dataset after the data have been collected. We subsequently discuss each step in greater detail.

5.2 Create Structure

The basic idea of setting up a good workflow is that good planning saves the researcher time, allows researchers to share the analysis, and/or allows replicating the research. After the data collection phase, the first step is to save the available data. We recommend keeping track of the dataset by providing data and data-related files in separate directories.

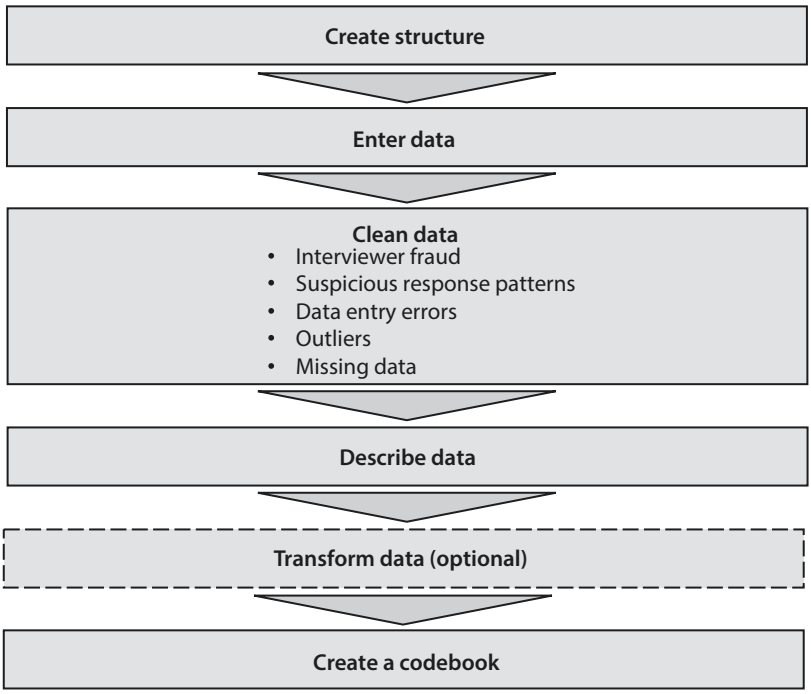


Fig. 5.1 The workflow of data

This directory should have subdirectories for at least: (1) the data files, (2) commands, (3) a temporary directory, and (4) related files; that is, a directory with files that are directly related to a project, such as the survey used to collect the data.[1]

In Table 5.1, we show an example of a directory structure. Within the main directory, there are four subdirectories, each with distinct files. Notice that in the **Data files** subdirectory, we have the original dataset, two modified datasets (one without missing data and one which includes several transformations of the data), as well as a zip file that contains the original dataset. If the data file is contained in a zip or other archive file, it is stored and unlikely to be modified, but can be easily opened if the working file is accidentally overwritten or deleted. In the **Data files** subdirectories, we distinguish between two files with the suffix **rev1** and **rev2**. We use **rev** (abbreviation of revision), but you however, choose another file name as long as it clearly indicates the revision on which you are working. In the **Syntax files** subdirectory, we store all syntax code that were used to manage and analyze our data. These commands may relate to different project phases, including a missing data analysis, descriptives, factor analysis, and other methods used over the course of the project. The **Output files** subdirectory includes a series of files with results from different statistical

1 Alternatively, you could also choose one of the many control system versions, including Subversion, Git, and Mecurial, which enable simple branching and project management. These systems work well with version control in centralized and in distributed environments.

Table 5.1 Example of a directory structure for saving market-research-related files

Directory name	Subdirectory name	Example file names
Oddjob	Data files	oddjob.sav
		oddjob.zip
		oddjob rev1.sav
		oddjob rev2.sav
	Syntax files	Missing data analysis.sps
		Descriptives.sps
		Factor analysis.sps
		Regression analysis.sps
	Output files	Missing data analysis.spv
		Descriptives.spv
		Factor analysis.spv
		Regression analysis.spv
	Temporary files	Missing data analysis rev1.spv
		Descriptives rev1.spv
		Factor analysis rev1.spv
		Regression analysis rev1.spv
	Related files	Codebook.docx
		Survey.pdf
		Initial findings–presentation to client.pptx
		Findings–presentation to client.pptx
		Recommendations rev1.docx
		Recommendations rev2.docx

analyses. As the name indicates, **Temporary files** serve as intermediary files that are kept until the final data or command files are established, after which they are removed. Finally, in the **Related Files** subdirectory, we have a codebook (more on this later), the survey, two presentations, and two documents containing recommendations.

Another aspect of creating a structure is setting up the variables for your study properly. This involves making decisions on the following elements:

- Variable names,
- variable labels,
- data type, and
- coding of variables.

The *variable names* should be clear and short so that they can be read in the dialog boxes. For example, if you have three questions on product satisfaction, three on loyalty, and several descriptors (age and gender), you could code these variables as *satisfaction1, satisfaction2, satisfaction3, loyalty1, loyalty2, loyalty3, age*, and *gender*.

In SPSS, and most other statistical software programs, you can include variable labels that describe what each variable denotes. The description generally includes the original question if the data were collected using surveys. Another point to consider is *variable coding*. Coding means assigning values to a variable. When collecting quantitative data, the task is relatively easy; we use values that correspond with the answers for Likert and semantic differential scales. For example, when using a 7-point Likert scale, responses can be coded as 1–7 or as 0–6 (with 0 being the most negative and 6 being the most positive response). Open-ended questions (*qualitative data*) require more effort, usually involving a three-step process. First, we collect all the responses. In the second step, we group these responses. Determining the number of groups and the group to which a response belongs is the major challenge in this step. Two or three market researchers usually code the responses independently to prevent the process from becoming too subjective and thereafter discuss the differences that may arise. The third step is providing a value for each group. Please see Krippendorff (2012) for more details about coding qualitative variables.

The following video gives a brief introduction into qualitative coding.

© vgajic/Getty Images/iStock
https://www.youtube.com/watch?v=DRL4PF2u9XA

Once a system has been set up to keep track of your progress, you need to consider safeguarding your files. Large companies usually have systems for creating backups (extra copies as a safeguard). If you are working alone or for a small company, you are probably responsible for this. You should save your most recent and second most recent version of your file on a separate drive and have multiple copies of your entire drive! Always keep at least two copies and never keep both backups in the same place, because you could still

lose all your work through theft, fire, or an accident! You can use cloud storage services, such as Amazon's storage services, or Dropbox, Google Drive, or Microsoft's OneDrive for small projects to prevent loss. Always read the terms of the cloud storage services carefully to determine whether your data's privacy is guaranteed.

5.3 Enter Data

How do we enter survey or experimental data into a dataset? Specialized software is often used for large datasets, or datasets created by professional firms. For example, Epidata (http://www.epidata.dk, freely downloadable) is frequently used to enter data from paper-based surveys, Confirmit's mobile survey (http://www.confirmit.com) to enter data from personal intercepts or face-to-face interviewing, and Voxco's Interviewer CATI for telephone interviewing. The SPSS Data Collection Family (http://www.spss.com.hk/software/data-collection/) is a suite of different software packages specifically designed to collect and (automatically) enter data collected from online, telephone, and paper-based surveys.

Such software may not be available for smaller projects, in which case data should be entered directly into SPSS. A significant drawback of direct data entry is the risk of typing errors, for which SPSS cannot check. Professional software, such as Epidata, can directly check if values are admissible. For example, if a survey question has only two answer categories, such as gender (coded 0/1), Epidata (and other packages) can directly check if the value entered is 0 or 1, and not any other value. The software also allows for using multiple typists when very large amounts of data need to be entered simultaneously.

5.4 Clean Data

Cleaning data is the next step in the workflow. It requires checking for:

- Interviewer fraud,
- suspicious response patterns,
- data entry errors,
- outliers, and
- missing data.

These issues require researchers to make decisions very carefully. In the following, we discuss each issue in greater detail. Table 5.3 summarizes the key recommendations discussed in the following sections.

5.4.1 Interviewer Fraud

Interviewer fraud is a difficult and serious issue. It ranges from interviewers "helping" respondents provide answers to entire surveys being falsified. Interviewer fraud often leads to incorrect results. Fortunately, we can avoid and detect interviewer fraud in various ways. First, never base interviewers' compensation on the number of completed responses they submit. Second, check and control for discrepancies in respondent selection and

responses. If multiple interviewers were used, each of whom collected a reasonably large number of responses ($n > 100$), a selection of the respondents should be similar. This means that the average responses obtained should also be similar. In ► Chap. 6 we will discuss techniques to test this. Third, if possible, contact a random number of respondents afterwards for their feedback on the survey. If a substantial number of people claim they were not interviewed, interviewer fraud is likely. Furthermore, if people were previously interviewed on a similar subject, the factual variables collected, such as their gender, should not change (or no more than a trivial percentage), while variables such as a respondent's age and highest education level should only move up. We can check this using descriptive statistics. If substantial interviewer fraud is suspected, the data should be discarded. You should check for interviewer fraud during the data collection process to safeguard the quality of data collection and minimize the risk of having to discard the data in the end.

5.4.2 Suspicious Response Patterns

Before analyzing data, we need to identify **suspicious response patterns**. There are two types of response patterns we need to look for:

- Straight-lining, and
- inconsistent answers.

Straight-lining occurs when a respondent marks the same response in almost all the items. For example, if a 7-point scale is used to obtain answers and the response pattern is 4 (the middle response), or if the respondent selects only 1s, or only 7s in all the items. A common way of identifying straight-lining is by including one or more *reverse-scaled items* in a survey (see ► Chap. 4). By evaluating the response patterns, we can differentiate between those respondents who are not consistent for the sake of consistency and those who are merely mindlessly consistent. Note, however, that this only applies if respondents do not tick the middle option. Straight-lining is very common, especially in web surveys where respondents generally pay less attention to the answers. Likewise, long surveys and those with many similarly worded items trigger straight-lining (Drolet and Morrison 2001). An alternative is to note potential straight-lined responses and include this as a separate category in the subsequent statistical analyses. This step avoids the need to reduce the sample and indicates the size and direction of any bias.

However, straight-lining can also be the result of *culture-specific response styles*. For example, respondents from different cultures have different tendencies regarding selecting the mid points (**middle response styles**) or the end points of a response scale (**extreme response styles**). Similarly, respondents from different cultures have different tendencies regarding agreeing with statements, regardless of the item content; this tendency is also referred to as **acquiescence** (Baumgartner and Steenkamp 2001). For example, respondents from Spanish-speaking countries tend to show higher extreme response styles and high acquiescence, while East Asian (Japanese and Chinese) respondents show a relatively high level of middle response style. Within Europe, the Greeks stand out as having the highest level of acquiescence and a tendency towards an extreme response

style. Harzing (2005) and Johnson et al. (2005) provide reviews of culture effects on response behavior.

Inconsistent answers also need to be addressed before analyzing your data. Many surveys start with one or more screening questions. The purpose of a screening question is to ensure that only individuals who meet the prescribed criteria complete the survey. For example, a survey of mobile phone users may screen for individuals who own an iPhone. If an individual indicates that he/she does not have an iPhone, this respondent should be removed from the dataset.

Surveys often ask the same question with slight variations, especially when reflective measures (see Box 3.1 in ► Chap. 3) are used. If a respondent gives a different answer to very similar questions, this may raise a red flag and could suggest that the respondent did not read the questions closely, or simply marked answers randomly to complete the survey as quickly as possible.

5.4.3 Data Entry Errors

When data are entered manually, **data entry errors** occur routinely. Fortunately, such errors are easy to spot if they happen outside the variable's range. That is, if an item is measured using a 7-point scale, the lowest value should be 1 (or 0) and the highest 7 (or 6). We can check if this is true by using descriptive statistics (minimum, maximum, and range; see next section). Data entry errors should always be corrected by going back to the original survey. If we cannot go back (e.g., because the data were collected using face-to-face interviews), we need to delete this specific observation for this specific variable.

More subtle errors—for example, incorrectly entering a score of 4 as 3—are difficult to detect using statistics. One way to check for these data entry errors is to randomly select observations and compare the entered responses with the original survey. We do, of course, expect a small number of errors (below 1 %). If many data entry errors occur, the dataset should be entered again.

Manual double data entry is another method to detect data entry errors. That is, once the data has been entered manually, a second data checker enters the same data a second time and the two separate entries are compared to ensure they match. Entries that deviate from one another or values that fall outside the expected range of the scales (e.g., 7-point Likert scales should have values that fall within this range) are then indicative of data entry errors (Barchard and Verenikina 2013). Various studies reveal that—although double data entry is more laborious and expensive—it detects errors better than single data entry (Barchard and Pace 2011; Paulsen et al. 2012).

5.4.4 Outliers

Data often contain **outliers**, which are values situated far from all the other observations that may influence results substantially. For example, if we compare the average income of 20 households, we may find that the incomes range between $20,000 and $100,000, with the average being $45,000. If we considered an additional household with an income of, say, $1 million, this would increase the average substantially.

Tip

Malcolm Gladwell's (2008) book "Outliers: The Story of Success" is an entertaining study of how some people became exceptionally successful (outliers).

5.4.4.1 Types of Outliers

Outliers must be interpreted in the context of the study and this interpretation should be based on the types of information they provide. Depending on the source of their uniqueness, outliers can be classified into three categories:

- The first type of outlier is produced by data collection or entry errors. For example, if we ask people to indicate their household income in thousands of US dollars, some respondents may just indicate theirs in US dollars (not thousands). Obviously, there is a substantial difference between $30 and $30,000! Moreover, (as discussed before) data entry errors occur frequently. Outliers produced by data collection or entry errors should be deleted, or we need to determine the correct values by, for example, returning to the respondents.
- A second type of outlier occurs because exceptionally high or low values are a part of reality. While such observations can influence results significantly, they are sometimes highly important for researchers, because the characteristics of outliers can be insightful. Think, for example, of extremely successful companies, or users with specific needs long before most of the relevant marketplace also needs them (i.e., lead users). Deleting such outliers is not appropriate, but the impact that they have on the results must be discussed.
- A third type of outlier occurs when *combinations* of values are exceptionally rare. For example, if we look at income and expenditure on holidays, we may find someone who earns $1,000,000 and spends $500,000 of his/her income on holidays. Such combinations are unique and have a very strong impact on the results (particularly the correlations that we discuss later in this chapter). In such situations, the outlier should be retained, unless specific evidence suggests that it is not a valid member of the population under study. It is very useful to flag such outliers and discuss their impact on the results.

5.4.4.2 Detecting Outliers

In a simple form, outliers can be detected using univariate or bivariate graphs and statistics.[2] When searching for outliers, we need to use multiple approaches to ensure that we detect all the observations that can be classified as outliers. In the following, we discuss both routes to outlier detection:

2 There are multivariate techniques that consider three, or more, variables simultaneously in order to detect outliers. See Hair et al. (2019) for an introduction, and Agarwal (2013) for a more detailed methodological discussion.

- **Univariate Detection**

The univariate detection of outliers examines the distribution of observations of each variable with the aim of identifying those cases falling outside the range of the "usual" values. In other words, finding outliers means finding observations with very low or very high variable values. This can be achieved by calculating the minimum and maximum value of each variable, as well as the range. Another useful option for detecting outliers is through box plots, which visualize the distribution of a variable and pinpoint those observations that fall outside the range of the "usual" values. We introduce the above statistics and box plots in greater detail in the *Describe Data* section.

- **Bivariate Detection**

We can also examine pairs of variables to identify observations whose combinations of variables are exceptionally rare. This is done by using a *scatter plot*, which plots all observations in a graph where the *x*-axis represents the first variable and the *y*-axis the second (usually *dependent*) variable (see ► Sec. 5.5). Observations that fall markedly outside the range of the other observations will show as isolated points in the scatter plot.

A drawback of this approach is the number of scatter plots that we need to draw. For example, with 10 variables, we need to draw 45 scatter plots to map all possible combinations of variables! Consequently, we should limit the analysis to only a few relationships, such as those between a dependent and independent variable in a regression. Scatterplots with large numbers of observations are often problematic when we wish to detect outliers, as there is usually not just one dot, or a few isolated dots, just a cloud of observations where it is difficult to determine a cutoff point.

5.4.4.3 Dealing with Outliers

In a final step, we need to decide whether to delete or retain outliers, which should be based on whether we have an explanation for their occurrence. If there is an explanation (e.g., because some exceptionally wealthy people were included in the sample), outliers are typically retained, because they are part of the population. However, their impact on the analysis results should be carefully evaluated. That is, one should run an analysis with and without the outliers to assess if they influence the results. If the outliers are due to a data collection or entry error, they should be deleted. If there is no clear explanation, outliers should be retained.

5.4.5 Missing Data

Market researchers often have to deal with **missing data**. There are two levels at which missing data occur:

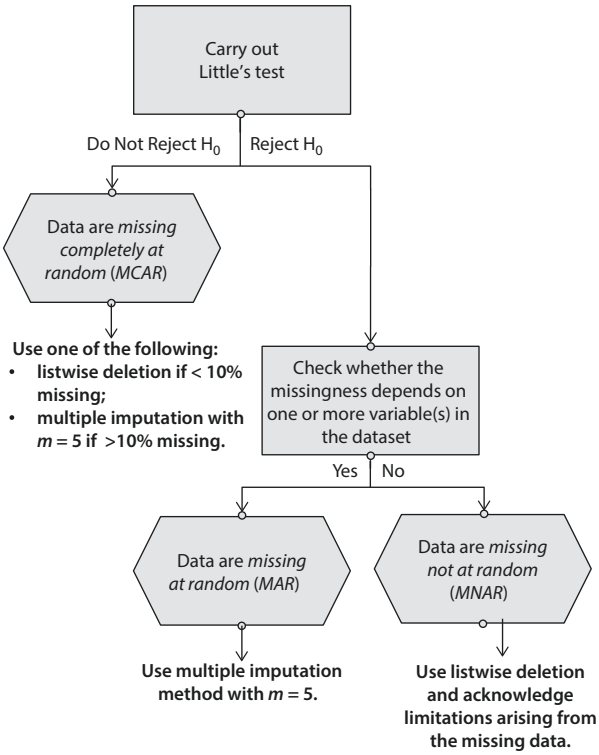
- Entire surveys are missing (survey non-response), and
- Respondents have not answered all the items (item non-response)

Survey non-response (also referred to as *unit non-response*) occurs when entire surveys are missing. Survey non-response is very common and regularly 75–95%, suggesting only 5%–25% of surveys are filled out. Issues such as inaccurate address lists, a lack of

interest and time, people confusing market research with selling, privacy issues, and respondent fatigue also lead to dropping response rates. The issue of survey response is best solved by designing proper surveys and survey procedures (see Box 4.7 in ▶ Chap. 4 for suggestions).

Item non-response occurs when respondents do not provide answers to certain questions. There are different forms of missingness, including people not filling out or refusing to answer questions. Item non-response is common and 2–10 % of questions usually remain unanswered. However, this number greatly depends on factors, such as the subject matter, the length of the questionnaire, and the method of administration. Non-response can be much higher in respect of questions that people consider sensitive and varies from country to country. In some countries, for instance, reporting incomes is a sensitive issue.

The key issue with item non-response is the type of pattern that the missing data follow. Do the missing values occur randomly, or is there some type of underlying system?[3] ◘ Fig. 5.2 illustrates the process of missing data treatment, which we will discuss next.



◘ **Fig. 5.2** Guideline for treating missing data

3 For more information on missing data, see https://www.iriseekhout.com

5.4.5.1 The Three Types of Missing Data: Paradise, Purgatory, and Hell

We generally distinguish between three types of missing data:
- Missing completely at random ("paradise"),
- missing at random ("purgatory"), and
- non-random missing ("hell").

Data are **missing completely at random** (**MCAR**) when the probability of data being missing is unrelated to any other measured variable and is unrelated to the variable with missing values. MCAR data occur when there is no systematic reason for certain data points being missing. For example, MCAR may happen if the Internet server hosting the web survey broke down temporarily. Why is MCAR paradise? When data are MCAR, observations with missing data are indistinguishable from those with complete data. If this is the case and little data are missing (typically less than 10 % in each variable) listwise deletion can be used. Listwise deletion means that we only analyze complete cases; in most statistical software, such as SPSS, this is a default option. Note that this default option in SPSS only works when estimating models and only applies to the variables included in the model. When more than 10 % of the data are missing, we can use multiple imputation (Eekhout et al. 2014), a more complex approach to missing data treatment that we discuss in the section *Dealing with Missing Data*.

Unfortunately, data are rarely MCAR. If the missingness of a variable's observation depends on one or more other variable(s) for which we have complete information (i.e., there are no missing observations for these variables), we consider the data **missing at random** (**MAR**). In this case, the probability that a data point is missing depends on the respondents' traits (e.g., age, gender, or income) or their answering behavior regarding other variables in the dataset. An example of MAR is when women are less likely to reveal their income. That is, the probability of missing data depends on the gender and not on the actual level of the respondent's income.

The term MAR is quite confusing. Many researchers confuse MAR with MCAR but the label has stuck.

Why is MAR purgatory? When data are MAR, the missing value pattern is not random, but this can be handled by more sophisticated missing data techniques such as multiple imputation, which use information in other variables in the dataset to impute the missing data points. On the downside, however, the missingness can typically not be fully accounted for by other variables in the dataset. Hence, we must also rely on its substantive reasonableness to verify whether or not the data are MAR.

Finally, data are **missing not at random** (**MNAR**) when the probability that a data point (e.g., x_i) is missing depends on the variable x. For example, very affluent and poor people are generally less likely to indicate their income even when having the exact same observed values of race, education, and other observed background variables. That is, the missing income values depend on the income variable itself! This is the most severe type of missing data ("hell"). Even sophisticated missing data techniques do not provide satisfactory

solutions as it is impossible to estimate the missing observations from other known observations in the dataset. Thus, any result based on MNAR data should be considered with caution. MNAR data can best be prevented by extensive pretesting and consultations with experts to avoid surveys that cause problematic response behavior. For example, we could use income categories instead of querying the respondents' income directly, or we could simply omit the income variable.

The following website offers a nice visualization of these three typed of missing data.

© 1001Love/Getty Images/iStock
https://iriseekhout.shinyapps.io/missingmechanisms/

5.4.5.2 Testing for the Type of Missing Data

When dealing with missing data, we must ascertain the missing data's type. If the dataset is small, we can browse through the data for obvious nonresponse patterns. However, missing data patterns become more difficult to identify with an increasing sample size and number of variables. Similarly, when we have few observations, patterns should be difficult to spot. In these cases, we should use one (or both) of the following diagnostic tests to identify missing data patterns:

- Little's MCAR test, and
- mean difference tests.

Little's MCAR test (Little 1998) analyzes the pattern of the missing data by comparing the observed data with the pattern expected if the data were randomly missing. If the test indicates no significant differences between the two patterns, the missing data can be classified as MCAR. Put differently, the null hypothesis is that the data are MCAR. Thus,

- if we do **not** reject the null hypothesis, we assume that the data are MCAR, and
- if we reject the null hypothesis, the data are either MAR or MNAR.

If the data cannot be assumed to be MCAR, we need to test whether the missing pattern is caused by another variable in the dataset by using the procedures discussed in ► Chap. 6. For example, we can run a two independent samples *t*-test to explore whether there is a significant difference in the mean of a continuous variable (e.g., income) between the group with missing values and the group without missing values. If we find a significant difference between these two groups, we would conclude that the data are MAR. For nominal or ordinal variables, we would tabulate the occurrence of non-responses against different groups' responses. If we put the (categorical) variable about which we have concerns in one column of a table (e.g., income category), and the number of (non-)responses in another, we obtain a table similar to ◘ Table 5.2.

We can use the χ^2-*test* (pronounced as *chi-square*), discussed in the ↓ Web Appendix (→ Downloads → ► Chap. 6), to test if there is a significant relationship between the respondents' (non-)responses of a certain variable and their income. In this example, the test yields a χ^2 value of 28.88, indicating that there is a significant relationship between the respondents' income and the (non-)response behavior, supporting the assumption that the data are MAR.

In the ↓ Web Appendix (→ Downloads), we illustrate details of Little's MCAR test, together with missing data analysis and imputation procedures.

© MicroStockHub/Getty Images/iStock
https://www.guide-market-research.com/app/download/13488666227/SPSS+3rd_Chapter+5_Multiple+Imputation.pdf?t=1516712968

◘ **Table 5.2** Example of response issues

	Low income	Medium income	High income
Response	65	95	70
Non-response	35	5	30

N = 300

5.4.5.3 Dealing with Missing Data

Research has suggested a broad range of approaches for dealing with missing data. We discuss the listwise deletion and the multiple imputation method.

Listwise deletion uses only those cases with complete responses in respect of all the variables considered in the analysis. If any of the variables used have missing values, the observation is omitted from the computation. If many observations have some missing responses, this decreases the usable sample size substantially and hypotheses are tested with less power (the power of a statistical test is discussed in ▶ Chap. 6).

Multiple imputation is a more complex approach to missing data treatment (Rubin 1987; Carpenter and Kenward 2013). It is a simulation-based statistical technique that facilitates inference by replacing missing observations with a set of possible values (as opposed to a single value) representing the uncertainty about the missing data's true value (Schafer 1997). The missing values are replaced by a set of plausible values not once, but *m* times (e.g., 5 times). This procedure yields *m* imputed datasets, each of which reflects the uncertainty about the missing data's correct value (Schafer 1997). Using these *m* datasets as input, SPSS then analyzes each dataset separately. Depending on the type of statistical analysis, the program additionally analyzes a pooled dataset that combines the *m* datasets into one. According to the literature, deciding on the number of imputations, *m*, can be very challenging, especially when the patterns of the missing data are unclear. As a rule of thumb, an *m* of at least 5 should be sufficient to obtain valid inferences (Rubin, 1987; White et al. 2011).

Now that we have briefly reviewed the most common approaches for handling missing data, there is still one unanswered question: Which one should you use? As shown in ◘ Fig. 5.2, if the data are MCAR, listwise deletion is recommended (Graham 2012) when the missingness is less than 10 % and multiple imputation when this is greater than 10 %. When the data are not MCAR but MAR, listwise deletion yields biased results. You should therefore use the multiple imputation method with an *m* of 5 (White et al. 2011). Finally, when the data are MNAR, the multiple imputation method provides inaccurate results. Consequently, you should choose listwise deletion and acknowledge the limitations arising from the missing data. ◘ Table 5.3 summarizes the data cleaning issues discussed in this section.

5.5 Describe Data

Once we have performed the previous steps, we can turn to the task of describing the data. Data can be described one variable at a time (*univariate descriptives*) or in terms of the relationship between two variables (*bivariate descriptives*). We further divide univariate and bivariate descriptives into graphs and tables, as well as statistics.

The choice between the two depends on the information we want to convey. Graphs and tables can often tell a non-technical person a great deal. On the other hand, statistics require some background knowledge, but have the advantage that they take up little space and are exact. We summarize the different types of descriptive statistics in ◘ Fig. 5.3.

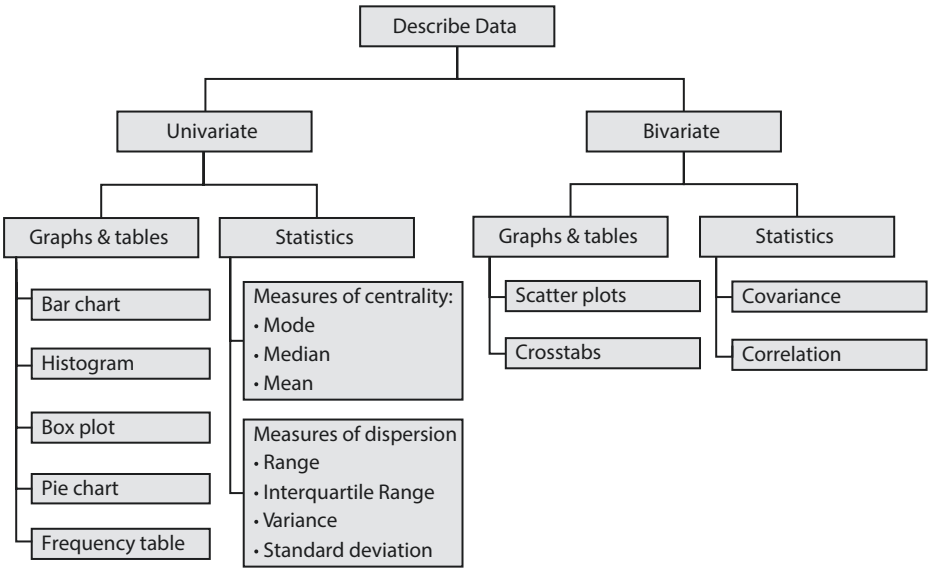


Fig. 5.3 The different types of descriptive statistics

Table 5.3 Data cleaning issues and how to deal with them

Problem	Action
Interviewer fraud	– Check with respondents whether they were interviewed and correlate with previous data if available.
Suspicious response patterns	– Check for straight lining. – Include reverse-scaled items. – Consider removing the cases with straight-lined responses. – Consider cultural differences in response behavior (middle and extreme response styles, acquiescence). – Check for inconsistencies in response behavior.
Data entry errors	– Use descriptive statistics (minimum, maximum, range) to check for obvious data entry errors. – Compare a subset of surveys to the dataset to check for inconsistencies.
Outliers	– Identify outliers using univariate descriptive statistics (minimum, maximum, range), box plots, and scatter plots. – Outliers are usually retained unless they … - … Are a result of data entry errors, - … do not fit the objective of the research, or - … influence the results severely (but report results with and without outliers for transparency).
Missing data	– Check the type of missing data by running Little's MCAR test and, if necessary, mean differences tests. – When the data are MCAR, use either listwise deletion or the multiple imputation method with an *m* of 5. – When the data are MAR, use the multiple imputation method with an *m* of 5. – When the data are MNAR, use listwise deletion and acknowledge the limitations arising from the missing data.

5.5.1 Univariate Graphs and Tables

In this section, we discuss the most common *univariate graphs* and *univariate tables*:

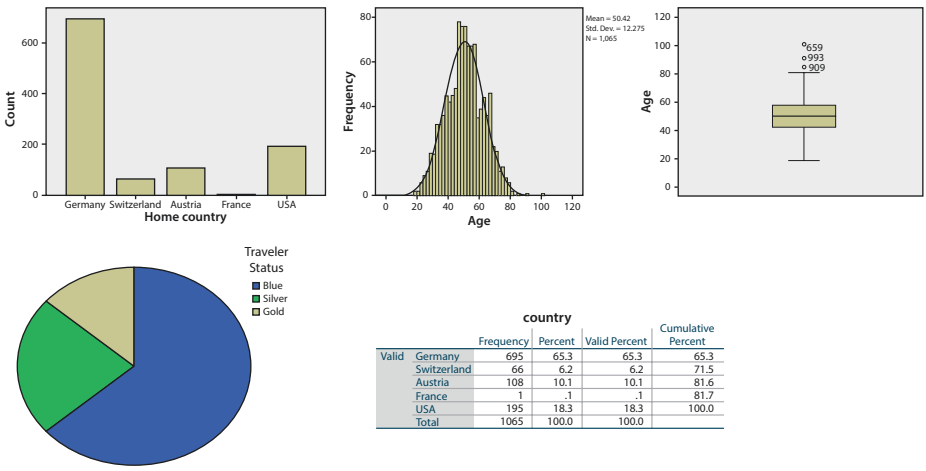
- Bar chart,
- histogram,
- box plot,
- pie chart, and the
- frequency table.

5

Figure 5.4 draws on these different types of charts and tables to provide information on the characteristics of travelers taken from the Oddjob Airways dataset that we use throughout this book.

A **bar chart** (Fig. 5.4 top left) is a graphical representation of a single categorical variable indicating each category's frequency of occurrence. However, each bar's height can also represent other indices, such as centrality measures or the dispersion of different data groups (see next section). Bar charts are primarily useful for describing nominal or ordinal variables. Histograms should be used for interval or ratio-scaled variables.

A **histogram** (Fig. 5.4 top middle) is a graph that shows how frequently categories made from a continuous variable occur. Differing from the bar chart, the variable categories on the *x*-axis are divided into (non-overlapping) classes of equal width. For example, if you create a histogram for the variable *age*, you can use classes of 21–30, 31–40, etc. A histogram is commonly used to examine the distribution of a variable. For this purpose, a curve following a specific distribution (e.g., normal) is typically superimposed on the bars to assess the correspondence of the actual distribution to the



country

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Germany	695	65.3	65.3	65.3
	Switzerland	66	6.2	6.2	71.5
	Austria	108	10.1	10.1	81.6
	France	1	.1	.1	81.7
	USA	195	18.3	18.3	100.0
	Total	1065	100.0	100.0	

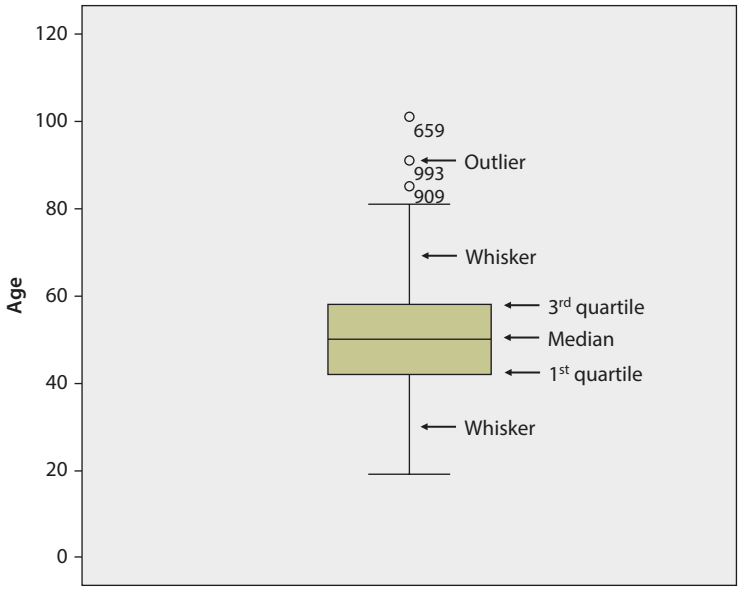
Fig. 5.4 From top left to bottom right; the bar chart, histogram, box plot, pie chart, and frequency table

desired (e.g., normal) distribution. Given that overlaying a normal curve makes most symmetric distributions look more normal then they are, you should be cautious when assessing normality using histograms. In ► Chap. 6 we will discuss several options for checking the normality of data.

Histograms plot continuous variables with ranges of the variables grouped into intervals (bins), while bar charts plot nominal and ordinal variables.

Another way of displaying the distribution of a (continuous) variable is the **box plot** (◘ Fig. 5.4 top right) (also referred to as a **box-and-whisker plot**). The box plot is a graph representing a variable's distribution and consists of elements expressing the dispersion of the data. Note that several elements refer to terminologies discussed in the *Univariate Statistics* section (► Sect. 5.5.2). ◘ Fig. 5.5 shows a box plot for the variable age based on the Oddjob Airways dataset.

- The bottom and top of the box describe the first and third quartiles. That is, the box contains the middle 50 % of the data, which is equivalent to the *interquartile range* (see ► Sect. 5.5.2).
- The solid line inside the box represents the *median*.
- The upper line extending the box (*whisker*) represents the distance to the largest observation that is within the following range: 3rd quartile + interquartile range. If there are no observations within this range, the line is equal to the maximum value.



◘ **Fig. 5.5** Elements of the box plot

- The lower line extending the box (whisker) represents the distance to the smallest observation that is within the following range: 1st quartile—interquartile range. If there are no obervations within this range, the line is equal to the minimum value.
- *Outliers* (observations that range between 1.0 and 3.0 interquartile ranges away from the box) and *extreme values* (observations that range more than 3.0 interquartile ranges away from the box) are depicted by symbols outside the whiskers.

We can make statements about the dispersion of the data with a box plot. The larger the box, the greater the observations' variability. Furthermore, the box plot helps us identify outliers in the data.

The **pie chart** (i.e., ◘ Fig. 5.4 bottom left) visualizes how a variable's different values are distributed. Pie charts are particularly useful for displaying percentages of variables, because people interpret the entire pie as being 100 %, and can easily see how often values occur. The limitation of the pie chart is, however, that it is difficult to determine the size of segments that are very similar.

A **frequency table** (i.e., ◘ Fig. 5.4 bottom right) is a table that includes all possible values of a variable in absolute terms (i.e., frequency), how often they occur relatively (i.e., percentage), and the percentage of the cumulative frequency, which is the sum of all the frequencies from the minimum value to the category's upper bound (i.e., cumulative frequency). It is similar to the histogram and pie chart in that it shows the distribution of a variable's possible values. However, in a frequency table, all values are indicated exactly. Like pie charts, frequency tables are primarily useful if variables are measured on a nominal or ordinal scale.

5.5.2 Univariate Statistics

Univariate statistics fall into two groups: those describing centrality and those describing the dispersion of variables. Box 5.2 at the end of this section shows sample calculation of the statistics used on a small set of values.

5.5.2.1 Measures of Centrality

Measures of centrality (also referred to as *measures of central tendency*) are statistical indices of a "typical" or "average" score. There are two main types of measures of centrality, the median and the mean.[4]

The **median** is the value that occurs in the middle of the set of scores if they are ranked from the smallest to the largest, and it therefore separates the lowest 50 % of cases from the highest 50 % of cases. For example, if 50 % of the products in a market cost less than $1000, then this is the median price. Identifying the median requires at least ordinal data (i.e., it cannot be used with nominal data).

4 The mode is another measure. However, unlike the median and mean, it is ill-defined, because it can take on multiple values. Consequently, we do not discuss the mode.

The most commonly used measure of centrality is the **mean** (also called the *arithmetic mean* or, simply, the *average*). The mean (abbreviated as $\bar{x}$) is the sum of each observation's value divided by the number of observations:

$$\bar{x} = \frac{Sum(x)}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

In the above formula, x_i refers to the value of observation i of variable x and n refers to the total number of observations. The mean is only useful for interval or ratio-scaled variables. SPSS also allows computing the *5 % trimmed mean*, which is the mean that would be obtained if the lower and upper 5 % of values of the variable were deleted.

Tip

Each measure of centrality has its own use. The mean is most frequently used, but is sensitive to very small or large values. Conversely, the median is not sensitive to outliers. Consequently, the relationship between the mean and the median provides us with valuable information about a variable's distribution. If the mean and the median are about the same, the variable is likely to be symmetrically distributed (i.e., the left side of the distribution mirrors the right side). If the mean differs from the median, this suggests that the variable is asymmetrically distributed and/or contains outliers. This is the case when we examine the prices of a set of products valued $500, $530, $530, and $10,000; the median is $530, while the mean is $2890. This example illustrates why a single measure of centrality can be misleading. We also need to consider the variable's dispersion to gain a more complete picture.

5.5.2.2 Measures of Dispersion

Measures of dispersion provide researchers with information about the variability of the data; that is, how far the values are spread out. We differentiate between four types of measures of dispersion:

- Range,
- interquartile range,
- variance, and
- standard deviation.

The **range** is the simplest measure of dispersion. It is the difference between the highest and the lowest value in a dataset and can be used on data measured at least on an ordinal scale. The range is of limited use as a measure of dispersion, because it provides information about extreme values and not necessarily about "typical" values. However, the range is valuable when screening data, as it allows for identifying data entry errors. For example, a range of more than 6 on a 7-point Likert scale would indicate an incorrect data entry.

The **interquartile range** is the difference between the 3rd and 1st quartile. The 1st *quartile* corresponds to the value separating the 25 % lowest values from the 75 % largest values if the values are ordered sequentially. Correspondingly, the 3rd quartile separates the 75 %

lowest from the 25 % highest values. The interquartile range is particularly important for drawing box plots.

The **variance** (generally abbreviated as s^2) is a common measure of dispersion. The variance is the sum of the squared differences of each value and a variable's mean, divided by the sample size minus 1. The variance is only useful if the data are interval or ratio-scaled:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

The variance tells us how strongly observations vary around the mean. A low variance indicates that the observations tend to be very close to the mean; a high variance indicates that the observations are spread out. Values far from the mean increase the variance more than those close to the mean.

The most commonly used measure of dispersion is the **standard deviation** (usually abbreviated as s). It is the square root of—and, therefore, a variant of—the variance:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$$

The variance and standard deviation provide similar information, but while the variance is expressed on the same scale as the original variable, the standard deviation is standardized. Consequently, the following holds for normally distributed variables (this will be discussed in the following chapters in more detail):

- 66 % of all observations are between plus and minus one standard deviation units from the mean,
- 95 % of all observations are between plus and minus two standard deviation units from the mean, and
- 99 % of all observations are between plus and minus three standard deviation units from the mean.

Thus, if the mean price is $1000 and the standard deviation is $150, then 66 % of all the prices fall between $850 and $1150, 95 % fall between $700 and $1300, and 99 % of all the observations fall between $550 and $1450.

5.5.3 Bivariate Graphs and Tables

There are several *bivariate graphs* and *bivariate tables*, of which the scatter plot and the crosstab are the most important. Furthermore, several of the graphs, charts, and tables discussed in the context of univariate analysis can be used for bivariate analysis. For example, box plots can be used to display the distribution of a variable in each group (category) of nominal variables.

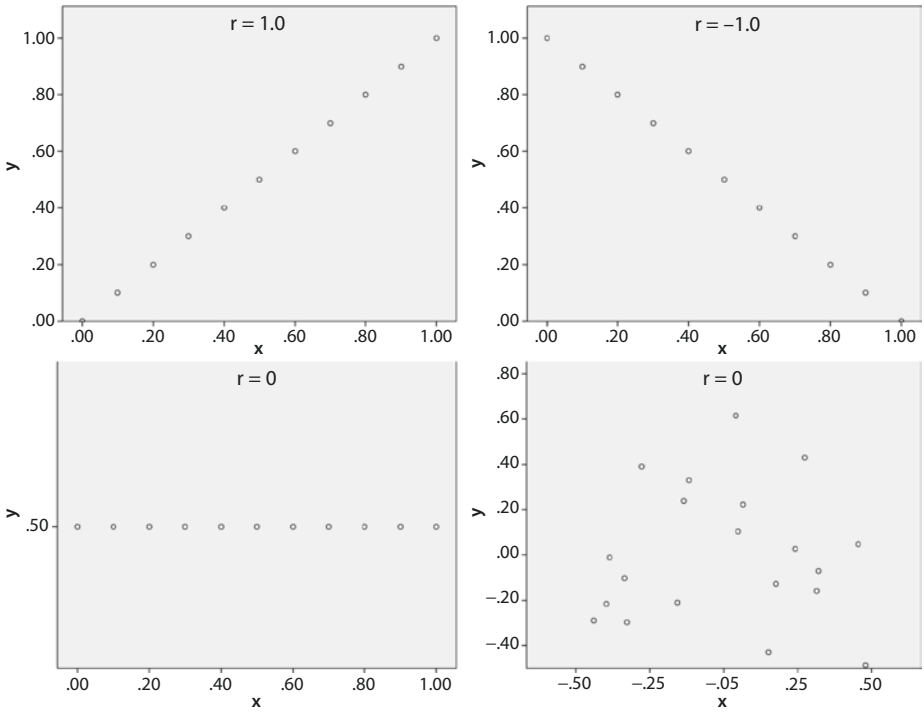


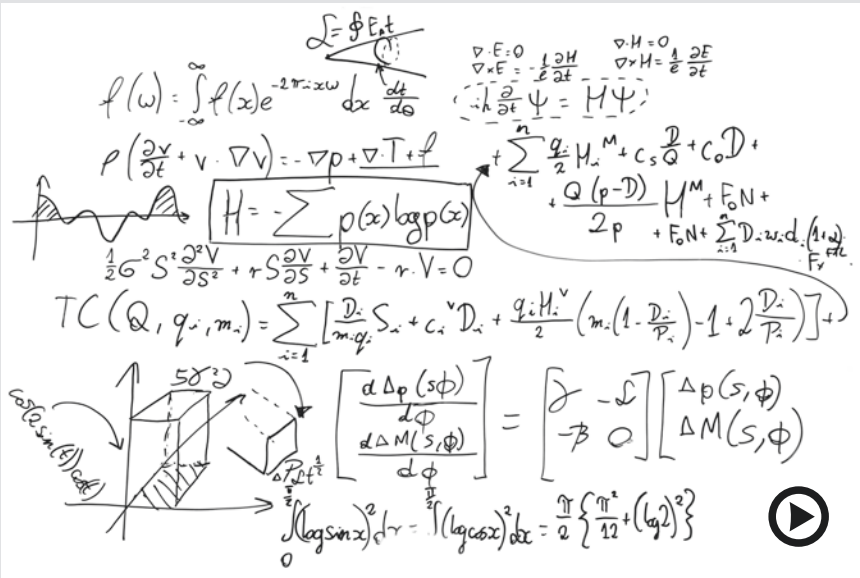
Fig. 5.6 Scatter plots and correlations

A **scatter plot** (see Fig. 5.6) uses both the *y* and *x*-axis to show how two variables relate to each other. If the observations almost form a straight diagonal line in a scatter plot, the two variables are strongly related.[5] Sometimes, a third variable is included, adding another dimension to the plot. Corresponding variants of the scatter plot include the *bubble plot* or *3-D scatter plot*.

Crosstabs (also referred to as *contingency tables*) are tables in a matrix format that show the frequency distribution of nominal or ordinal variables. They are the equivalent of a scatter plot used to analyze the relationship between two variables. While crosstabs are generally used to show the relationship between two variables, they can also be used for three or more variables, which, however, makes them difficult to grasp.

5 A similar type of chart is the **line chart**. In a line chart, measurement points are ordered (typically by their *x*-axis value) and joined with straight line segments.

Crosstabs are also part of the χ^2-*test*, which we discuss in the ↓ Web Appendix (→ Downloads → ► Chap. 6).



https://www.guide-market-research.com/app/download/13488667027/SPSS+3rd_Chapter+6_Chi-square+test.pdf?t=1516713011

5.5.4 Bivariate Statistics

Bivariate statistics involve the analysis of two variables to determine the empirical relationship between them. There are two key measures that indicate (linear) associations between two variables; we illustrate their computation in Box 5.2:

- covariance, and
- correlation.

The **covariance** is the degree to which two variables vary together. If the covariance is zero, then two variables do not vary together. The covariance is the sum of the multiplication of the differences between each value of the x_i and y_i variables and their means, divided by the sample size minus 1:

$$Cov(x_i, y_i) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})\cdot(y_i - \bar{y})$$

Box 5.2 Sample calculation of univariate and bivariate statistics

Consider the following list of values for variables *x* and *y*, which we treat as ratio-scaled:

x	6	6	7	8	8	8	12	14	14
y	7	6	6	9	8	5	10	9	9

Measures of centrality for x:
Median = 8

Mean $\bar{x} = \frac{1}{9}(6+6+\ldots+14+14) = \frac{83}{9} \approx 9.22$

Measures of dispersion for x:
Minimum = 6
Maximum = 14
Range = 14 – 6 = 8
Interquartile range = 6.5

Variance $(s^2) = \frac{\left[(6-9.22)^2+\ldots+(14-9.22)^2\right]}{9-1} = \frac{83.56}{8} \approx 10.44$

Standard deviation $(s) = \sqrt{s^2} = \sqrt{10.44} \approx 3.23$

Measures of association between x and y:

Covariance $(cov(x, y)) = \frac{1}{9-1}[(6-9.22)\cdot(7-7.67)+\ldots+$

$(14-9.22)\cdot(9-7.67)] = \frac{31.67}{8} \approx 3.9$

Correlation $(r) = \frac{3.96}{3.23 \cdot 1.73} \approx 0.71$

The **correlation** (typically abbreviated as *r*) is a common measure of how strongly two variables relate to each other. The most common type of correlation, the *Pearson's correlation coefficient*, is calculated as follows:

$$r = \frac{Cov(x_i, y_i)}{s_x \cdot s_y} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})\cdot(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\cdot\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

The numerator contains the covariance of x_i and y_i $\left(Cov(x_i, y_i)\right)$, while the denominator contains the product of the standard deviations of x_i and y_i.[6] Thus, the correlation is the covariance divided by the product of the standard deviations. As a result, the correlation is standardized and, unlike the covariance, is no longer dependent on the variables' original measurement. More precisely, the correlation coefficient ranges from −1 to 1, where −1 indicates a perfect negative relationship and 1 indicates the contrary. A correlation coefficient of 0 indicates that there is no relationship, also implying that their covariance is zero.

6 Note that the terms *n−1* in the numerator and denominator cancel each other and are therefore not shown here.

Tip

As a rule of thumb (Cohen 1988), an absolute correlation …
- … below 0.30 indicates a weak relationship,
- … between 0.30 and 0.49 indicates a moderate relationship, and
- … above 0.49 indicates a strong relationship.

Strong relationships are not necessarily better than weak relationships. Strong relationships are typically well-established which means they lack novelty. Furthermore, strong relationships may not be actionable. For example, knowing that high net worth is a predictor of sports car sales does not help if we do not have good information on high net worth for targeting purposes.

The scatter plots in ◘ Fig. 5.6 illustrate several correlations between two variables *x* and *y*. If the observations almost form a straight diagonal line in the scatter plot (upper left and right in ◘ Fig. 5.6), the two variables have a high (absolute) correlation. If the observations are uniformly distributed in the scatter plot (lower right in ◘ Fig. 5.6), or one variable is a constant (lower left in ◘ Fig. 5.6), the correlation is zero.

Pearson's correlation coefficient is the most common and is generally simply referred to as the correlation (Agresti and Finlay 2014). Pearson's correlation is appropriate for calculating correlations between two variables that are both interval or ratio-scaled. However, it can also be used when one variable is interval or ratio-scale and the other is, for example, binary. There are other correlation coefficients for variables measured on lower scale levels. Some examples are:
- *Spearman's correlation coefficient* and *Kendall's tau* when at least one variable for determining the correlation is measured on an ordinal scale.
- *Contingency coefficient*, *Cramer's V*, and *Phi* for variables measured on a nominal scale. These statistical measures are used with crosstabs; we discuss these in the context of the χ^2*-test* in the ↓ Web Appendix (→ Downloads → ▶ Chap. 6).

In ◘ Table 5.4, we indicate which descriptive statistics are useful for differently scaled variables. The brackets X indicate that the use of a graph, table, or statistic is potentially useful while (X) indicates that use is possible, but less likely useful, because this typically requires collapsing data into categories, resulting in a loss of information.

5.6 Transform Data (Optional)

Transforming data is an optional step in the workflow. Researchers transform data as certain analysis techniques require this: it might help interpretation or might help meet the assumptions of techniques that will be discussed in subsequent chapters. We distinguish two types of data transformation:
- variable respecification, and
- scale transformation.

Table 5.4 Types of descriptive statistics for differently scaled variables

	Nominal	Ordinal	Interval & ratio
Univariate graphs & tables			
Bar chart	X	X	
Histogram			X
Box plot			X
Pie chart	X	X	(X)
Frequency table	X	X	(X)
Univariate statistics: Measures of centrality			
Median		X	X
Mean			X
Univariate statistics: Measures of dispersion			
Range		(X)	X
Interquartile range		(X)	X
Variance			X
Standard deviation			X
Bivariate graphs/tables			
Scatter plot			X
Crosstab	X	X	(X)
Bivariate statistics			
Contingency coefficient	X		
Cramer's V	X		
Phi	X		
Spearman's correlation		X	
Kendall's tau		X	
Pearson's correlation			X

5.6.1 Variable Respecification

Variable respecification involves transforming data to create new variables or to modify existing ones. The purpose of respecification is to create variables that are consistent with the study's objective. *Recoding* a continuous variable into a categorical variable is an example of a simple respecification. For example, if we have a variable that measures a respondent's number of flights per year, we could code those flights below 5 as low (=1), between 5 and 10 flights as medium (=2), and everything above 11 as high (=3). Recoding

a variable in such a way always results in a loss of information, since the newly created variable contains less detail than the original. While there are situations that require recoding (e.g., we might be asked to give advice based on different income groups where income is continuous), we should generally avoid recoding.

Another example of respecification is swapping the polarity of a question. If you have a variable measured on a 5-point Likert-scale where: 1 = strongly agree; 2 = agree; 3 = undecided; 4 = disagree; 5 = strongly disagree and you wish to switch the polarity of the values so that value 1 reverses to 5, value 2 becomes 4, and so on.

Creating a dummy variable is a special way of recoding data. **Dummy variables** (or simply *dummies*) are binary variables that indicate if a certain trait is present or not. For example, we can use a dummy variable to indicate that advertising was used during a period (value of the dummy is 1) or not (value of the dummy is 0). We can also use multiple dummies to capture categorical variables' effects. For example, three levels of *flight intensity* (low, medium, and high) can be represented by two dummy variables: The first takes a value of 1 if the intensity is high (0 else), the second also takes a value of 1 if the intensity is medium (0 else). If both dummies take the value 0, this indicates low flight intensity. We always construct one dummy less than the number of categories. We show how to create dummy variables in SPSS in the example illustrating the use of regression analysis in ► Sect. 7.4.

The creation of *constructs* is a frequently used type of variable respecification. As described in ► Chap. 3, a construct is a concept that cannot be observed, but can be measured by using multiple items, none of which relate perfectly to the construct. To compute a construct measure, we need to calculate the average (or the sum) of several related items. For example, a traveler's *commitment* to fly with an airline can be measured by using the following three items:

- I am very committed to Oddjob Airways.
- My relationship with Oddjob Airways means a lot to me.
- If Oddjob Airways would no longer exist, it would be a true loss for me.

By calculating the average of these three items, we can form a composite of *commitment*. If one respondent indicated 4, 3, and 4 on the three items' scale, we calculate a **construct score** (also referred to as a composite score) for this person as follows: $(4 + 3 + 4)/3 = 3.67$. Note that we should take the average over the number of nonmissing responses. In ► Chap. 8 we discuss more advanced methods of doing this by, for example, creating factor scores.

Similar to creating constructs, we can create an **index** of sets of variables. For example, we can create an index of information search activities, which is the sum of the information that customers require from promotional materials, the Internet, and other sources. This measure of information search activities is also referred to as a composite measure, but, unlike a construct, the items in an index define the trait to be measured.

5.6.2 Scale Transformation

Scale transformation involves changing the variable values to ensure comparability with other variables or to make the data suitable for analysis. Different scales are often used to measure different variables. For example, we may use a 5-point Likert scale for one set of

variables and a 7-point Likert scale for a different set of variables in our survey. Owing to the differences in scaling, it would not be meaningful to make comparisons across any respondent's measurement scales. These differences can be corrected by **standardizing variables**.

A popular way of standardizing data is by rescaling these to have a mean of 0 and a variance of 1. This type of standardization is called the ***z*-standardization**. Mathematically, standardized scores z_i (also called *z-scores*) can be obtained by subtracting the mean $\bar{x}$ of every observation x_i and dividing it by the standard deviation s. That is:

$$z_i = \frac{\left(x_i - \bar{x}\right)}{s}$$

Range standardization $\left(r_i\right)$ is another standardization technique which scales the data in a specific range. For example, standardizing a set of values to a range of 0 to 1 requires subtracting the minimum value of every observation x_i and then dividing it by the range (i.e., the difference between the maximum and minimum value).

$$r_i = \frac{\left(x_i - x_{min}\right)}{\left(x_{max} - x_{min}\right)}$$

The range standardization is particularly useful if the mean, variance, and ranges of different variables vary strongly and are used for some forms of cluster analysis (see ► Chap. 9.)

A **log transformation**—another type of transformation—is commonly used if we have skewed data. **Skewed data** occur if we have a variable that is asymmetrically distributed and can be positive or negative. A *positive skew* (also called *right-skewed data* or skewed to the right) occurs when many observations are concentrated on the left side of the distribution, producing a long right tail. When data are right-skewed, the mean will be higher than the median. A *negative skew* (also called *left-skewed data* or skewed to the left) is the opposite, meaning that many observations are concentrated on the right of the distribution, producing a long left tail. When data are negatively skewed, the mean will be lower than the median. A histogram will quickly show whether data are skewed. Skewed data can be undesirable in analyses. Log transformations are commonly used to transform data closer to a normal distribution when the data are right-skewed (i.e., the data are non-negative). Taking a natural logarithm will influence the size of the coefficient related to the transformed variable, but will not influence the value of its outcome.[7]

Finally, **aggregation** is a special type of transformation. Aggregation means that we take variables measured at a lower level to a higher level. For example, if we know the average customer's satisfaction with an airline and the distribution channels from which they buy (i.e., the Internet or a travel agent), we can calculate the average satisfaction at the channel level. Aggregation only works from lower to higher levels and is useful if we want to compare groups at a higher level.

7 The logarithm is calculated as follows: If $x = y^b$, then $y = log_b(x)$ where x is the original variable, b the logarithm's base, and y the exponent. For example, log 10 of 100 is 2. Logarithms cannot be calculated for negative values (such as household debt) and for the value of zero.

While transforming data is often necessary to ensure comparability between variables or to make the data suitable for analysis, there are also drawbacks to this procedure. Most notably, we may lose information during most transformations. For example, recoding the *ticket price* (measured at the ratio scale) as a "low," "medium," and "high" ticket price will result in an ordinal variable. In the transformation process, we have therefore lost information by going from a ratio to an ordinal scale. Another drawback is that transformed data are often more difficult to interpret. For example, the log *(ticket price)* is far more difficult to interpret and less intuitive than simply using the ticket price.

5.7 Create a Codebook

After all the variables have been organized and cleaned, and some initial descriptive statistics have been calculated, we can create a **codebook**, containing essential details of the data collection and data files, to facilitate sharing. Codebooks usually have the following structure:

- *Introduction*: The introduction discusses the goal of the data collection, why the data are useful, who participated, and how the data collection effort was conducted (mail, Internet, etc.).
- *Questionnaire(s)*: It is common practice to include copies of all the types of questionnaires used. Thus, if different questionnaires were used for different respondents (e.g., for French and Chinese respondents), a copy of each original questionnaire should be included. Differences in wording may afterwards explain the results of the study, particularly those of cross-national studies, even if a back-translation was used (see ► Chap. 4). These are not the questionnaires received from the respondents themselves, but blank copies of each type of questionnaire used. Most codebooks include details of each variable as comments close to the actual items used. If a dataset was compiled using secondary measures (or a combination of primary and secondary data), the secondary datasets are often briefly discussed (the version that was used, when it was accessed, etc.).
- *Description of the variables*: This section includes a verbal description of each variable used. It is useful to provide the variable name as used in the data file, a description of what the variable is supposed to measure, and whether the measure has previously been used. You should also describe the measurement level (see ► Chap. 3).
- *Summary statistics*: This section includes descriptive statistics of each variable. The average (only for interval and ratio-scaled data), minimum, and maximum are often shown. In addition, the number of observations and usable observations (excluding observations with missing values) are included, just like a histogram (if applicable).
- *Datasets*: This last section includes the names of the datasets and sometimes the names of all the revisions of the used datasets. Codebooks sometimes include the file date or a data signature, to ensure that the right files are used.

In the Web Appendix (↓ Web Appendix → Downloads), we briefly discuss how to create a codebook using SPSS.

© Sezeryadigar/Getty Images/iStock
https://www.guide-market-research.com/app/download/13488664927/SPSS+3rd_Chapter+5_Codebook.pdf?t=1516712939

5.8 The Oddjob Airways Case Study

The most effective way of learning statistical methods is to apply them to a set of data. Before introducing SPSS and how to use it, we present the dataset from a fictitious company called *Oddjob Airways* (but with a real website, http://www.oddjobairways.com) that will guide the examples throughout this book. The dataset *Oddjob.sav* (↓ Web Appendix → Downloads) stems from a customer survey of Oddjob Airways. Founded in 1962 by the Korean businessman Toshiyuki Sakata, Oddjob Airways positions itself as a low-cost carrier, targeting young customers who prefer late or overnight flights. However, the actual customer base is very diverse with many older customers appreciating the offbeat onboard services, high-speed Wi-Fi, and tablet computers on every seat. In an effort to improve its customers' satisfaction, the company's marketing department contacted all the customers who had flown with the airline during the last 12 months and were registered on the company website. A total of 1065 customers who had received an email with an invitation letter completed the survey online.

Learn more about Oddjob Airways at www.oddjobairways.com. The site not only offers background on the airline but also lots of learning material relevant to this book.

Oddjob Airways

Table 5.5 Variable description and label names of the Oddjob Dataset

Variables	Variable description	Variable name in the dataset
Demographic measures		
Age of the customer	Numerical variable ranging between the ages of 19 and 101.	*age*
Customer's gender	Dichotomous variable, where 1 = female; 2 = male.	*gender*
Language of customer	Categorical variable, where 1 = Germany; 2 = English; 3 = French.	*language*
Home country	Categorical variable, whereby: 1 = Germany (de), 2 = Switzerland (ch); 3 = Austria (at); 4 = France (fr), 5 = the United States (us).	*country*
Flight behavior measures		
Flight class	Categorical variable distinguishing between the following categories: 1 = First; 2 = Business; 3 = Economy.	*flight_class*
Latest flight	Categorical variable querying when the customer last flew with Oddjob Airways. Categories are: 1 = Within the last 2 days; 2 = Within the last week; 3 = Within the last month; 4 = Within the last 3 months; 5 = Within the last 6 months; 6 = Within the last 12 months.	*flight_latest*
Flight purpose	Dichotomous variable distinguishing between: 1 = Business; 2 = Leisure.	*flight_purpose*
Flight type	Dummy variable, where: 1 = Domestic; 2 = International.	*flight_type*
Number of flights	Numeric variable ranging between 1 and 457 flights per year.	*nflights*
Traveler's status	Categorical variable, where membership status is defined in terms of: 1 = Blue; 2 = Silver; 3 = Gold.	*status*
Perception and satisfaction measures		
Recommendation	Item on whether a customer is likely to recommend the airline to a friend or colleague. This item is measured on an 11-point Likert-scale ranging from 1 very unlikely to 11 very likely.	*nps*
Reputation	One item stating "Oddjob Airways is a reputable airline." This item is measured on a 7-point Likert-scale ranging from 1 fully disagree to 7 fully agree.	*reputation*
General satisfaction	3 items reflecting a customer's overall satisfaction with the airline. All items are measured on a 7-point Likert scale ranging from 1 fully disagree to 7 fully agree .	*sat1* to *sat3*
Overall price/ performance satisfaction	One item stating "Overall I am satisfied with the price performance ratio of Oddjob Airways." This item is measured on a 7-point Likert-scale ranging from 1 fully disagree to 7 fully agree.	*overall_sat*

Table 5.5 (Continued)

Variables	Variable description	Variable name in the dataset
Loyalty	5 items reflecting a customer's loyalty to the airline. All items are measured on a 7-point Likert-scale ranging from 1 fully disagree to 7 fully agree.	*loy1* to *loy5*
Commitment	3 items reflecting a customer's commitment to fly with the airline. All items are measured on a 7-point Likert-scale ranging from 1 fully disagree to 7 fully agree. The dataset also contains the variable *commitment*, which is the mean of *com1* to *com3*.	*com1* to *com3* and *commitment*
Traveler's expectations	23 items reflecting a customer's expectations with the airline: "How high are your expectations that … " All items are measured on a continuous scale ranging from 1 very low to 100 very high.	*e1* to *e23*
Traveler's satisfaction	23 items reflecting a customer's satisfaction with Oddjob Airways regarding the features asked in the expectation items (*e1-e23*) on a continuous scale ranging from 1 = very unsatisfied to 100 = very satisfied.	*s1* to *s23*

The survey resulted in a rich dataset with information about travelers' demographic characteristics, flight behavior, as well as their price/product satisfaction with and expectations in respect of Oddjob Airways. Table 5.5 describes the variables in detail.

5.8.1 Introduction to SPSS

SPSS is a computer package specializing in quantitative data analysis. It is widely used by market researchers. It is powerful, able to deal with large datasets, and relatively easy to use.

In this book, we use version 25 of IBM SPSS Statistics for Mac (which we simply refer to as SPSS). Prior versions (21 or higher) for Microsoft Windows, Mac, or Linux can also be used for almost all examples throughout the book. The differences between the versions are small enough so that all examples in the book work with all versions.

The regular SPSS package is available at a substantial fee for commercial use. However, large discounts are available for educational use. To obtain these discounts, it is best to go to your university's IT department and enquire if you can purchase a special student license. You can also download a trial version from www.spss.com.

In the next sections, we will use the ► sign to indicate that you should click on something. Options, menu items or drop-down lists that you should look up in dialog boxes are printed in bold. Variable names, data files or data formats are printed in *italics* to differentiate them from the rest of the text.

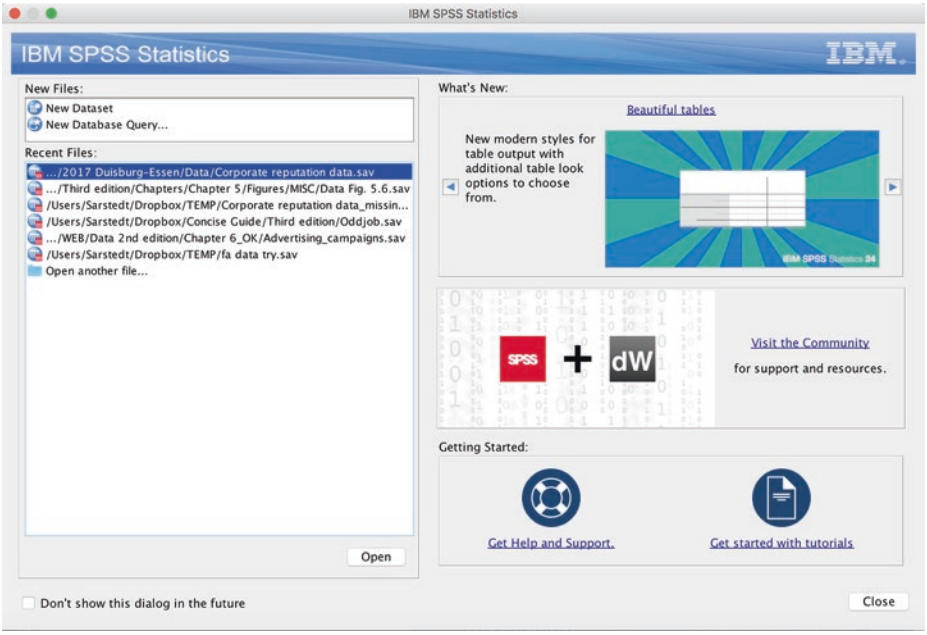


Fig. 5.7 The start-up screen of SPSS

5.8.2 Finding Your Way in SPSS

If you start up SPSS for the first time, it presents a screen similar to Fig. 5.7, unless a previous user has ticked the **Don't show this dialog in the future** box. In that case, you will see a screen similar to Fig. 5.8, but without an active dataset.

In the startup screen, SPSS indicates several options to create or open datasets. The options that you should use are either **Recent Files**, under which you can find a list with recently opened data files, or **New Files**. To open an unlisted file, simply choose **Open another file…** and click **OK** (alternatively, you can click **Close**, and then go to ► File ► Open ► Data). Then search for the directory in which you keep your files, click on *Oddjob.sav*, followed by **Open**.

Tip

SPSS uses multiple file formats. The *.sav* and *.zsav* file format contains data only. SPSS also allows you to open other file formats such as Excel (*.xls* and *.xlsx*), Stata files (*.dta*), and text files (such as *.txt* and *.dat*). Once these files are open, they can be conveniently saved into SPSS's own *.sav* file format. If you are on SPSS's main screen (see Fig. 5.8) simply go to File ► Open ► Data ► Files of type (select the file format) and double click on the file you wish to open.

SPSS uses two windows. The **SPSS Statistics Data Editor** contains the data and information on the variables in the dataset and the **SPSS Statistics Viewer**, which contains the output produced by the analyses. In the following, we will discuss these two windows separately.

Tip

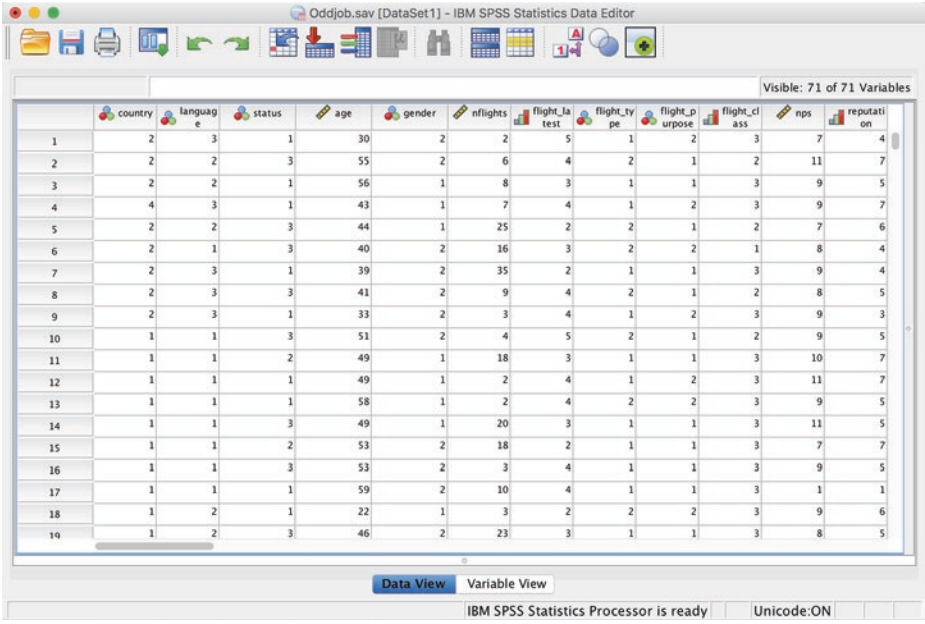
The notation you will see in this book follows the US style. That is, commas are used to separate ten thousands (e.g., 10,000) while decimal points are used to separate whole values from fractions. If you want to change the notation to US style, go to SPSS, then ► File ► New ► Syntax and type in *SET LOCALE = 'English'*. You also need to type in the last point. Now press enter and type in *EXECUTE.* (again, including the point) in the next line. Now run the syntax by choosing Run ► All in the syntax window. SPSS will then permanently apply the US style notation the next time you use SPSS.

5.8.3 SPSS Statistics Data Editor

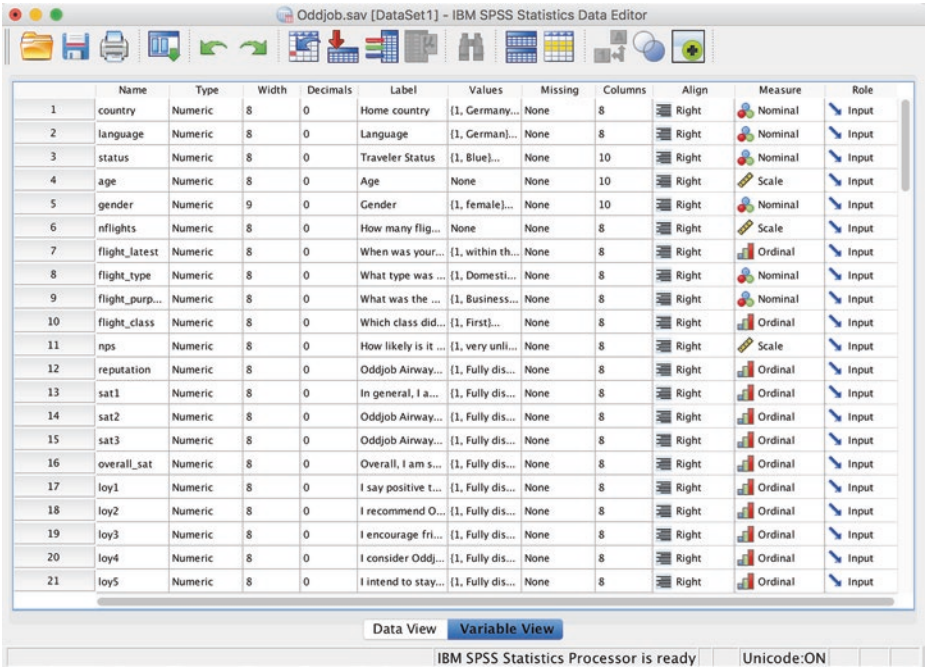
In the **SPSS Statistics Data Editor** (◘ Fig. 5.8), you will find the dataset *Oddjob. sav*. This dataset's variables are included in the columns and their names are indicated at the top of each column. The cases are in the rows, which are numbered from 1 onwards.

If you click on the **Variable View** tab at the bottom of the screen, SPSS will show you a screen similar to ◘ Fig. 5.9. In the **Variable View**, SPSS provides information on the variables included in your dataset:

- **Name**: Here you can indicate the name of the variable. It is best to provide very short names. Variable names must begin with letters (A to Z) or one of the following special characters (@, # or $). Subsequent characters can include letters (A to Z), numbers (0–9), a dot (.), and _, @, #, or $. Note that neither spaces nor other special characters (e.g., %, &,/) are allowed.



◘ **Fig. 5.8** The SPSS data editor



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	country	Numeric	8	0	Home country	{1, Germany...	None	8	Right	Nominal	Input
2	language	Numeric	8	0	Language	{1, German}...	None	8	Right	Nominal	Input
3	status	Numeric	8	0	Traveler Status	{1, Blue}...	None	10	Right	Nominal	Input
4	age	Numeric	8	0	Age	None	None	10	Right	Scale	Input
5	gender	Numeric	9	0	Gender	{1, female}...	None	10	Right	Nominal	Input
6	nflights	Numeric	8	0	How many flig...	None	None	8	Right	Scale	Input
7	flight_latest	Numeric	8	0	When was your...	{1, within th...	None	8	Right	Ordinal	Input
8	flight_type	Numeric	8	0	What type was ...	{1, Domesti...	None	8	Right	Nominal	Input
9	flight_purp...	Numeric	8	0	What was the ...	{1, Business...	None	8	Right	Nominal	Input
10	flight_class	Numeric	8	0	Which class did...	{1, First}...	None	8	Right	Ordinal	Input
11	nps	Numeric	8	0	How likely is it ...	{1, very unli...	None	8	Right	Scale	Input
12	reputation	Numeric	8	0	Oddjob Airway...	{1, Fully dis...	None	8	Right	Ordinal	Input
13	sat1	Numeric	8	0	In general, I a...	{1, Fully dis...	None	8	Right	Ordinal	Input
14	sat2	Numeric	8	0	Oddjob Airway...	{1, Fully dis...	None	8	Right	Ordinal	Input
15	sat3	Numeric	8	0	Oddjob Airway...	{1, Fully dis...	None	8	Right	Ordinal	Input
16	overall_sat	Numeric	8	0	Overall, I am s...	{1, Fully dis...	None	8	Right	Ordinal	Input
17	loy1	Numeric	8	0	I say positive t...	{1, Fully dis...	None	8	Right	Ordinal	Input
18	loy2	Numeric	8	0	I recommend O...	{1, Fully dis...	None	8	Right	Ordinal	Input
19	loy3	Numeric	8	0	I encourage fri...	{1, Fully dis...	None	8	Right	Ordinal	Input
20	loy4	Numeric	8	0	I consider Oddj...	{1, Fully dis...	None	8	Right	Ordinal	Input
21	loy5	Numeric	8	0	I intend to stay...	{1, Fully dis...	None	8	Right	Ordinal	Input

Fig. 5.9 Variable view

- **Type**: Here you can specify what your variable represents. **Numeric** refers to values and **String** refers to words. String is useful if you want to include open- ended answers, email addresses or any other type of information that cannot be adequately captured by numbers. With **Dollar** or **Custom Currency**, you can indicate that your variable represents money.
- **Width** and **Decimals**: These elements indicate the amount of space available for your variables values.
- **Labels**: Here you can provide a longer description of your variables (called variable labels). This can either be the definition of the variables or the original survey question.
- **Values**: Here you can indicate what a certain value represents (called value labels). For example, for the nominal variable *gender*, 1 represents females and 2 males.
- **Missing**: Here you can indicate one or more missing value(s). Generally, SPSS deals with missing values in two ways. If you have blanks in your variables (i.e., you haven't entered any data for a specific observation), SPSS treats these as *system-missing values*. These are indicated in SPSS as a dot (•). Alternatively, you can specify *user-defined missing values* that are meant to signify a missing observation. By explicitly defining missing values, we can indicate why specific scores are missing (e.g., the question didn't apply to the respondent or the respondent refused to answer). In SPSS, you should preferably specify user-defined missing values as this at least allows the true missing values to be separated from data that were not recorded (i.e., no value was entered). Under **Missing**, we can select among three

options to indicate user-defined missing values. The first option, **No missing values**, is the default setting and indicates that no values are user-defined missing. The other two options, **Discrete missing values** and **Range plus one optional discrete missing value**, provide a means to express user-defined missing values. To do so, simply enter values that record that an observation is missing. Each separate value should indicate separate reasons for missing values. For example –99 may record that a respondent could not answer, and –98 that the respondent was unwilling to answer, and –97 might record "other" reasons. Thus, missing values can provide us with valuable information. More importantly, observations with user-defined missing values (just like with system-missing values) are excluded from data transformations and analyses. This is of course essential as, for example, including a user-missing value of –99 in descriptive analyses would greatly distort the results.

Tip

When picking user-defined missing values, take those that would not otherwise occur in the data. For example, for a variable *age*, 1000 might be an acceptable value, as that response cannot occur. However, the same missing value for a variable *income* might lead to problems, as a respondent might have an income of 1000. If 1000 is indicated as a (user-defined) missing value, this observation would be excluded from further analysis. By convention, researchers usually choose (high) negative values to designate missing such as –99 or –999.

- **Columns** and **Align**: These are rarely necessary, so we will skip these.
- **Measure**: Here you can specify the measurement level of your variable. SPSS provides you with the option to indicate whether your variable is nominal, ordinal, or whether it is interval or ratio-scaled. The combination of the last two categories is called **Scale** in SPSS. Note that several procedures such as creating graphs require that all measurement levels are specified correctly.
- The last **Role** option is not necessary for basic analysis.

5.8.4 SPSS Statistics Viewer

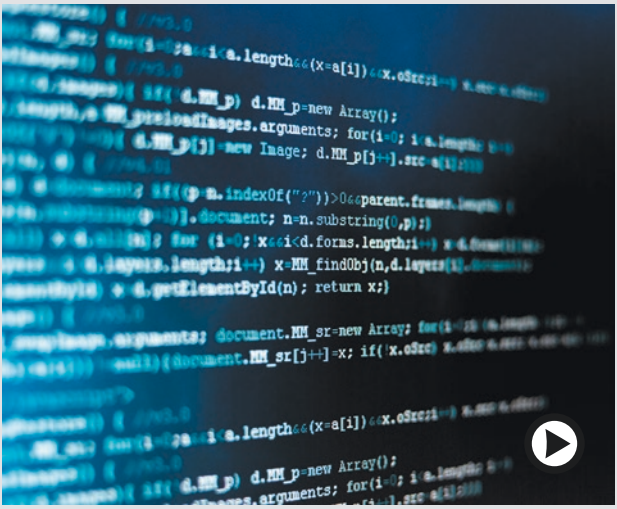
The **SPSS Statistics Viewer** is a separate window, which opens after you carry out an action in SPSS. The viewer contains the output that you may have produced. If you are used to working with software such as Microsoft Excel, where the data and output are included in a single screen, this may be a little confusing at first. Another aspect of the viewer screen is that it does not change your output once made. Unlike, for example, Microsoft Excel, changing the data after an analysis does not dynamically update the results.

The output produced in SPSS can be saved using the *.spv* file format that is particular to SPSS. To partially remedy this, SPSS provides the option to export the output to Microsoft Word, Excel, or PowerPoint, PDF, HTML, or text. It is also possible to export output as a picture. You can find these export options in the Statistics Viewer under File ► Export.

5.8.5 SPSS Menu Functions

In SPSS, you find a number of commands in the menu bar. These include File, ► Edit, ► View, ► Data, ► Transform, ► Analyze, and ► Graphs. In this section, we will briefly discuss these commands. The commands ► Analyze and ► Graphs will be discussed in greater detail in the example, later in this chapter. The last four commands are ► Utilities, ► Add-ons, ► Windows, and ► Help. You are unlikely to need the first three functions but the help function may come in handy if you need further guidance. Under help, you also find a set of tutorials that can show you how to use most of the commands included in SPSS.

In addition to the menu functionalities, we can run SPSS by using its command language, called **SPSS syntax**. You can think of it as a programming language that SPSS uses to translate those elements on which you have clicked in the menus into commands that SPSS can understand. Syntax can be saved (as a *.sps* file) for later use. This is particularly useful if you conduct the same analyses over different datasets. Think, for example, of standardized marketing reports on daily, weekly, or monthly sales. Discussing the syntax in great detail is beyond the scope of this book but we offer a brief introduction in the Web appendix (⇓ Web Appendix → Downloads). Grotenhuis and Visscher (2014) provide a thorough introduction into this subject.



https://www.guide-market-research.com/app/download/13488667727/SPSS+3rd_Chapter+5_SPSS+Syntax.pdf?t=1516713038

Under ► File, you find all the commands that deal with the opening and closing of files. Under this command, you will find subcommands that you can use to open different types of files, save files, and create files. If you open a dataset, you will notice that SPSS also opens a new screen. Note that SPSS can open several datasets simultaneously. You can easily switch from dataset to dataset by just clicking on the one which you would like to activate.

Under ► Edit, you will find subcommands to copy and paste data. Moreover, you will find two options to insert cases or variables. If you have constructed a dataset but need to enter additional data, you can use this to add an additional variable and subsequently add data. **Edit** also contains the **Find** subcommand with which you can look for specific cases or observations. Finally, under ► Edit ► Options, you find a large number of options, including how SPSS formats tables, and where the default file directories are located.

Under ► View, you find several options, of which the **Value Labels** option is the most useful. Value labels are words or short sentences used to indicate what each value represents. For example, value lablels for *status* include *blue, silver*, and *gold*. SPSS shows value labels in the **SPSS Data Editor** window if you click on ► View, and then on **Value Labels**.

Under the ► Data command, you will find many subcommands to change or restructure your dataset. The most prominent option is the ► Sort Cases subcommand with which you can sort your data based on the values of a variable. You could, for example, sort the data based on the respondents' age. The ► Split File subcommand is useful if we want to compare output across different groups such as flight class, purpose, or type. In addition, we can carry out separate analyses over different groups using the **Split File** command. Another very useful command is ► Data ► Select Cases, which allows you to restruct the observations that you want to analyze. Under ► Transform, we find several options to create new variables from existing variables. For example, the first subcommand is **Compute Variable**. This command allows you to create a new variable from one (or more) existing variables. Also included under the ► Transform command are two subcommands to recode variables (i.e., ► Recode into Same Variables and ► Recode into Different Variables). These commands allow you to recode variable values or to summarize sets of values into one value. For example, using the **Recode into Different Variables** command, you could generate a new variable that takes the value 1 if *age* is higher than 40, and 0 else.

Under ► Analyze, you find numerous analysis procedures, several of which we will discuss in the remainder of the book. For example, under **Descriptive Statistics**, you can request univariate and bivariate statistics. Under **Regression**, SPSS provides numerous types of regression techniques.

Finally, under ► Graphs, you can chose among different options to create graphics. For example, the **Chart Builder** is an interactive tool that allows you to design basic and more complex charts. Note that in order to use the **Chart Builder**, all variables' measurement levels must be correctly specified. The same holds for the **Graphboard Template Chooser**, which lets you select the variables to be visualized and suggests different charts that conform to the measurement levels of the variable(s). Finally, the **Legacy Dialogs** command allows for a menu-driven selection of different chart types.

Some of the previous commands are also accessible as shortcut symbols in **Data View** screen's menu bar. As these are very convenient, we present the most frequently used shortcuts in ◘ Table 5.6.

Table 5.6 Toolbar icons

Symbol	Action
	Open dataset
	Save the active dataset
	Recall recently used dialogs
	Undo a user action
	Find
	Split file
	Select cases
	Show value labels

5.9 Data Management in SPSS

In this section, we will illustrate the use of some of the most commonly used commands for managing data in SPSS using the *Oddjob.sav* dataset. These include the following:

- Split file,
- select cases,
- compute variables, and
- recode into same/different variables.

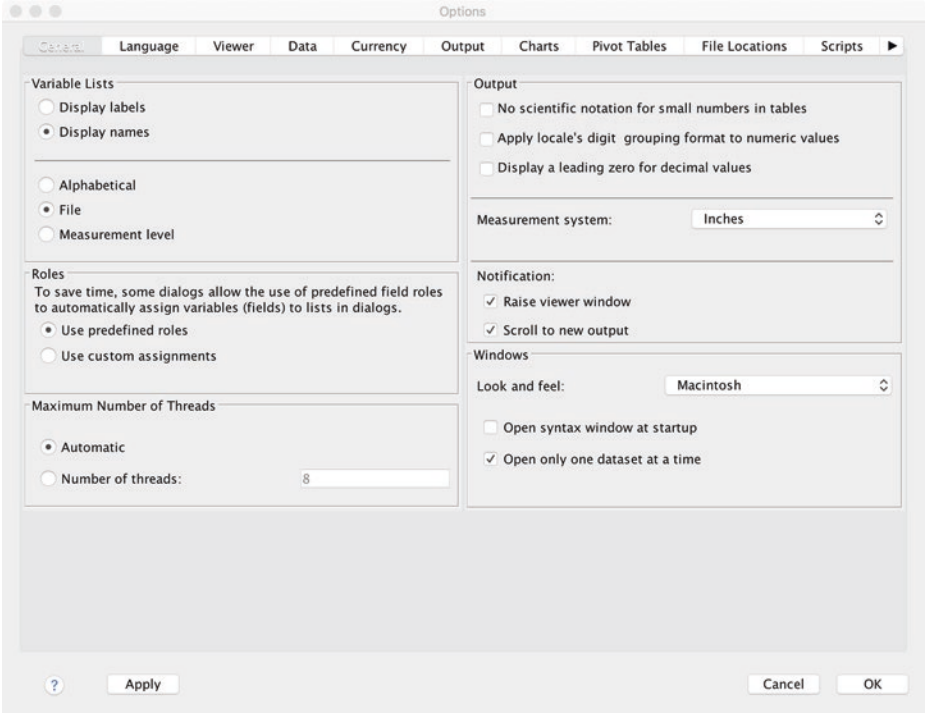
5.9.1 Split File

The **split file** command allows you to split the dataset on the basis of grouping variables. If the split file function is activated, all subsequent analyses will be done separately for each group of data, as defined by the grouping variable.

Tip

A very useful option, which we have adopted in this book throughout, is to show the variable names (instead of the variable labels). The variable names are typically short and the variable labels long. To make it easier to see the variables used, go to SPSS Statistics ► Preferences ► General (for Apple) or Edit ► Options ► General (for PC) and tick **Display names**. We show this in ◘ Fig. 5.10. Under SPSS Statistics ► Preferences ► Language (Apple) or Edit ► Options ► Language (PC) you can change the output and user interface languages.

By clicking on ► Data ► Split File, a dialog box similar to ◘ Fig. 5.11 will open. All you need to do is to enter the variable that indicates the grouping (select **Compare groups**) and move the grouping variable (e.g., *status*) into the **Groups Based on** box. Note that we chose to display the variable names in the variables box at the left side of the dialog box. By right-clicking on a variable name you can switch between **Display Variable Names** and **Display Variable Labels**. Once you have specified the grouping variable and clicked on **OK**, SPSS will automatically carry out all subsequent analyses of each status group (i.e., *blue, silver*, and *gold*) separately. If you want to revert to analyzing the whole dataset, you need to go to ► Data ► Split File, then click on **Analyze all cases, do not create groups**.



◘ **Fig. 5.10** How to display names

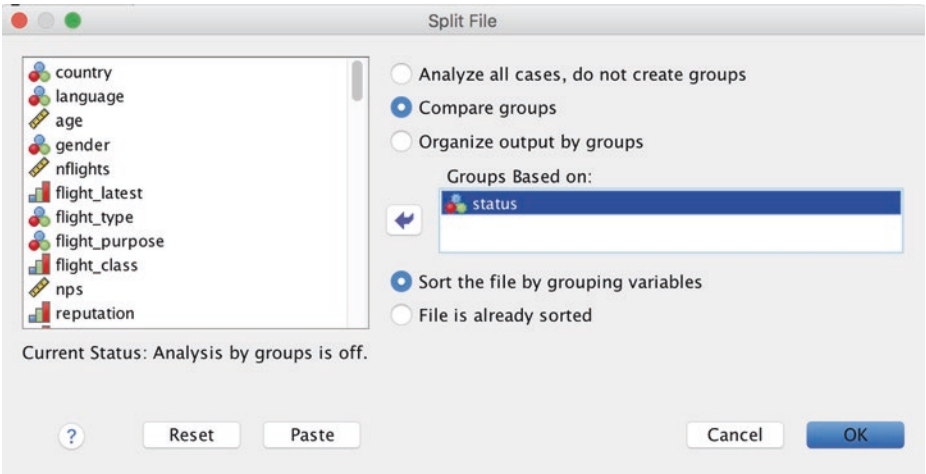


Fig. 5.11 Split file dialog box

It's a common mistake to forget to turn off the split file command. Failing to do so results in all subsequent analyses being carried out for each group separately!

5.9.2 Select Cases

The **select cases** command allows us to restrict certain observations from the analyses through a pre-set condition (e.g., display the summary statistics only for those older than 40). To run the command, go to ► Data ► Select Cases and click on **If condition is satisfied**, followed by **If**. SPSS shows a screen similar to Fig. 5.12 in which you can set the conditions for restricting observations. For example, to select respondents who are older than 40, enter *age>40* in the corresponding box. Next, click on **Continue** and **OK**.

SPSS will only use the selected observations in subsequent analyses, and will omit the others (these are crossed out in the **Data View** screen). Remember to turn the selection off if you do not need it by going back to ► Data ► Select Cases and then click on **All cases**. Since we use the full dataset, please turn this selection off so that you can follow the remainder of the example.

5.9.3 Compute Variables

The **compute variable** command allows you to create a new variable from one (or more) existing variables. For example, if you want to create a composite score of some type (see ► Chap. 3), you need to calculate the average of several variables. In the following, we will use the command to compute a new index variable from the mean of the following three

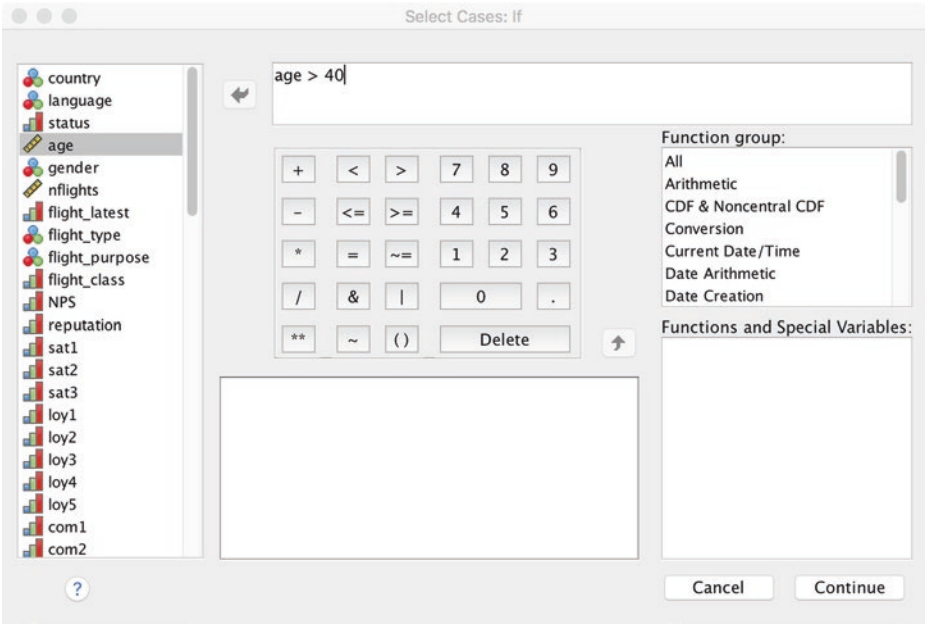


Fig. 5.12 Select cases dialog box

items related to travelers' satisfaction: *sat1, sat2*, and *sat3*. Go to ► Transform ► Compute Variable, which will open a dialog box similar to Fig. 5.13.

Next, enter the name of the new variable (e.g., *rating_index*) in the **Target Variable** box on the upper left part of the screen and select **Statistical** from the **Function group** menu, followed by **Mean** from the **Functions and Special Variables** menu. SPSS will now show the corresponding function in the **Numeric Expression** box as *MEAN(?,?)*. Enter *sat1, sat2, sat3* in the round brackets as in Fig. 5.13 and click on **OK**. You have now created a new variable called *rating_index* that appears at the bottom of the **Variable View**.

5.9.4 Recode Variables

Changing or transforming the values of an existing variable according to a number of rules is a key data management activity. Numeric variables can be changed using the **recode** command, which comes in two forms. You can either overwrite an existing variable or generate a new variable, which includes the new variable coding. We recommend using the recode into different variables option. If you were to use recode into the same variables, any changes you make to the variable will result in the deletion of the original variable. Thus, if you ever want to go back to the original data, you either need to have saved a previous version, or have to enter all the data again as SPSS cannot undo these actions! The recode subcommands allow you to change the scaling of a variable. This is useful if you want to create dummies or create categories of existing variables.

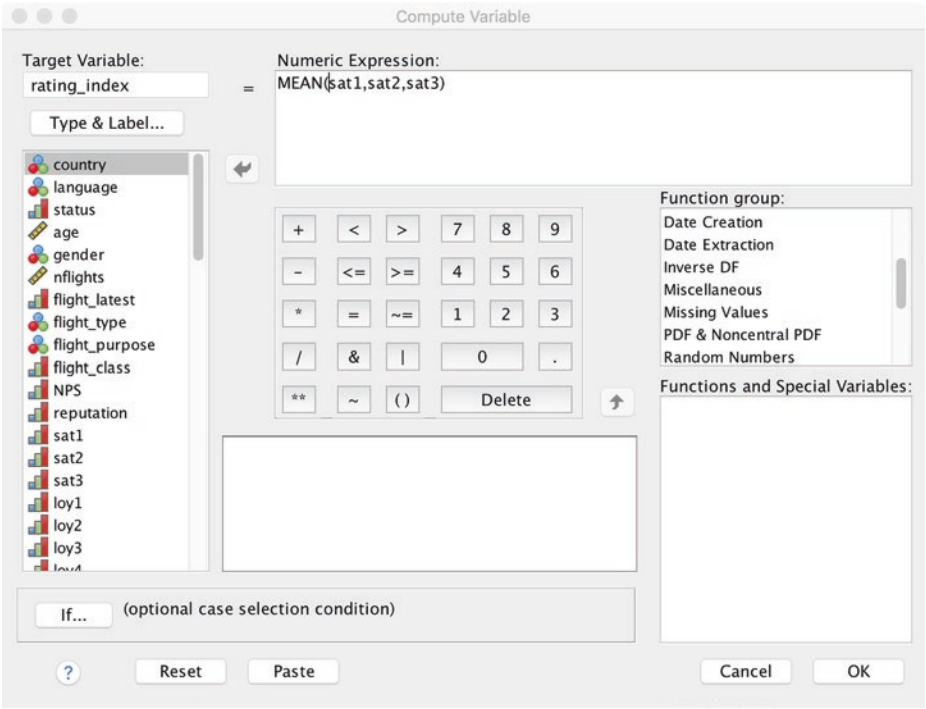


Fig. 5.13 Compute variables dialog box

In the following, we want to use the recode option to generate a new variable *age_dummy*, which should be 1 if a respondent is 40 or younger, and 0 else. To do so, go to ► Transform ► Recode into Different Variables, which will open a dialog box similar to Fig. 5.14.

First, move age into the **Numeric Variable → Output Variable** box. Next, specify the name of the variable that you want to recode (i.e., *age_dummy*) under **Output Variable** and add a label such as *Dummy variable age*. After clicking on **Change**, SPSS will draw an arrow between the original and new variables (Fig. 5.14). The next step is to tell SPSS what values need to be recoded. After clicking on **Old and New Values**, SPSS will open a dialog box similar to Fig. 5.15. On the left of this dialog box, you should indicate the values of the original variable that you want to recode. In our example, we used the option **Range, LOWEST through value** to specify that we want to recode all values of 40 or less. Select this option and enter *40* into the box below. Next, go to the right side of the menu and enter *1* under **New Value** and click on **Add**. To indicate that *age_dummy* should take the value 0 if age is greater than 40, select **All other values** and enter *0* under **New Value**, followed by **Add**. Finally, confirm your changes by clicking on **Continue** and **OK**. You have now created a new dichotomous variable (i.e., *age_dummy*) located at the bottom of the **Variable View**.

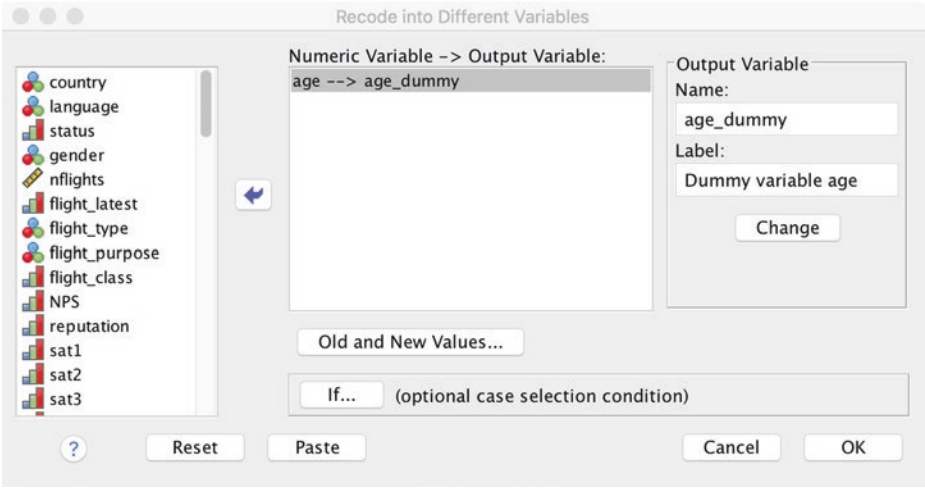


Fig. 5.14 Recode into different variables dialog box

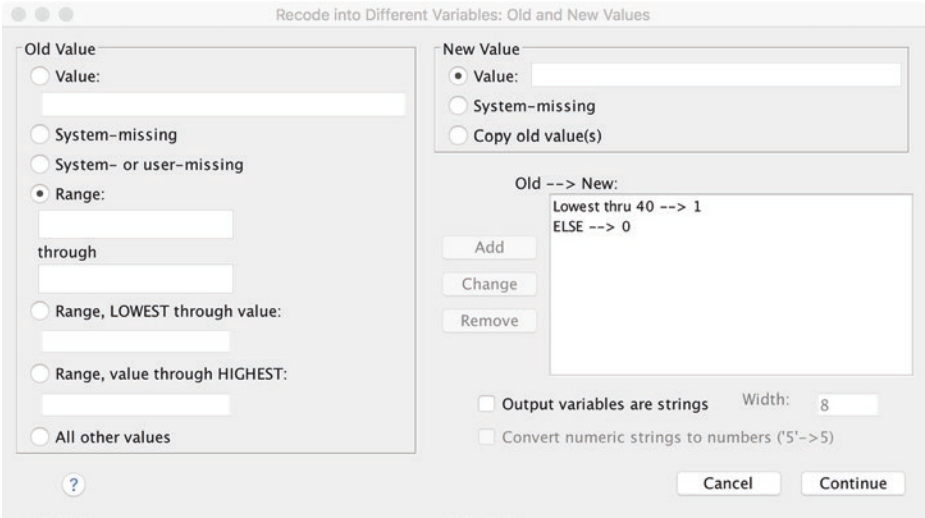


Fig. 5.15 Recode options dialog box

5.10 Example

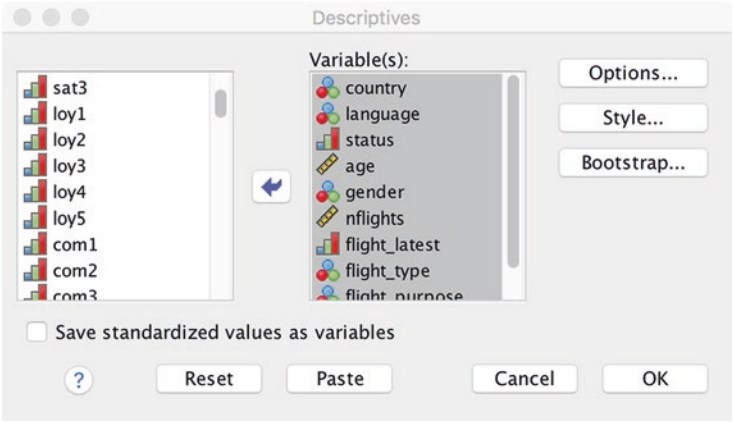
We will now examine the dataset Oddjob.sav in closer detail by following all the steps in Fig. 5.1. Cleaning the data generally requires checking for interviewer fraud, suspicious response patterns, data entry errors, outliers, and missing data. Several of these steps rely on statistics and graphs, which we discussed in the context of descriptive statistics (e.g., box plots and scatter plots). We illustrate the use of Little's MCAR test and multiple imputation of Oddjob.sav in the Web Appendix (→ Downloads).

5.10.1 Clean Data

Since the data were cleaned earlier, we need not check for interviewer fraud or suspicious response patterns. Beside double data entries to detect and minimize errors in the process of data entry, exploratory data analysis is required to spot data entry errors that have been overlooked.

A first step in this procedure is to look at the minimum and maximum values of the relevant variables to detect values that are not plausible (i.e., fall outside the expected range of scale categories). To do so, go to ► Analyze ► Descriptive Statistics ► Descriptives, which will open a dialog box similar to ◘ Fig. 5.16. We will focus our analyses on the first ten variables (i.e., *country, language, status, age, gender, nflights, flight_latest, flight_type, flight_purpose*, and *flight_class*). Enter these variables into the **Variable(s)** box and click on **OK**. ◘ Table 5.7 shows the resulting output.

◘ Table 5.7 lists the minimum, maximum, mean, and standard deviation values of all the variables. Under **N**, we can see that all the listed variables are observed across all 1065 respondents, meaning that none of the selected variables suffer from missing observations. Among others, it appears that the age of the travelers varies between 19 and 101, while the number of flights (variable *nflights*) varies between 1 and 457 flights over the past 12 months. Particularly the maximum value in number of flights appears to be implausible. While this observation could represent a flight attendant, it appears more reasonable to consider this observation an outlier and eliminate it from further analyses that draw on the *nflights* variable.



◘ **Fig. 5.16** Descriptives dialog box

Table 5.7 Descriptives statistics

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
country	1065	1	5	2.00	1.552
language	1065	1	3	1.24	.447
status	1065	1	3	1.50	.720
age	1065	19	101	50.42	12.275
gender	1065	1	2	1.74	.440
nflights	1065	1	457	13.42	20.226
flight_latest	1065	1	6	3.79	1.369
flight_type	1065	1	2	1.48	.500
flight_purpose	1065	1	2	1.51	.500
flight_class	1065	1	3	2.80	.435
Valid N (listwise)	1065				

5.10.2 Describe Data

In the next step, we describe the data in more detail, focusing on those statistics and graphs that were not part of the previous step. To do so, we make use of graphs, tables, and descriptive statistics. In Fig. 5.17, we show how you can request each of the previously discussed graphs and tables using the ► Graphs ► Legacy Dialogs menu. The figure also shows, which menu options to use to request univariate and bivariate statistics. We will not use the **Legacy Dialogs** option as the **SPSS Chart Builder** and **Graphboard Template Chooser** offer users more options when selecting a graph.

5.10.2.1 Univariate Graphs and Tables

- **Bar Charts**

To produce a bar chart that plots the *age* of respondents against their *country* of residence, go to ► Graphs ► Chart Builder. In the dialog box that opens (Fig. 5.18), select **Bar** under **Choose from** and drag the first element into the chart builder box. Next, drag and drop *age* from the **Variables** menu to the *y*-axis and *country* to the *x*-axis. Clicking on **OK**, SPSS will produce a chart similar to Fig. 5.19.

- **Histograms**

Histograms are useful for summarizing numerical variables. We will use the **Graphboard Template Chooser** to generate a histogram of the *age* variable. By going to ► Graphs ► Graphboard Template Chooser, SPSS will open a dialog box as shown in Fig. 5.20.

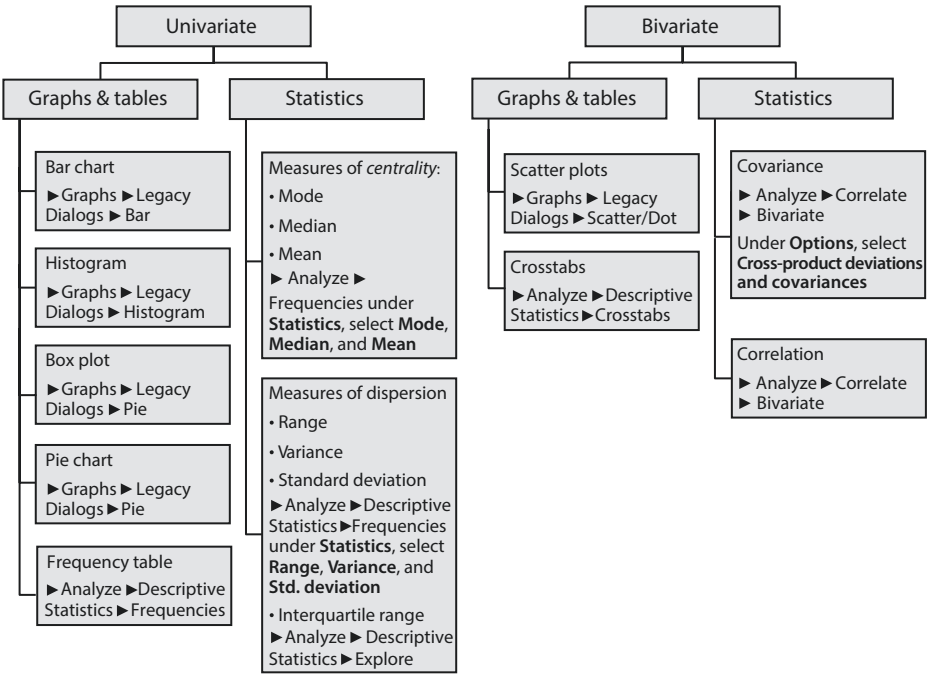


Fig. 5.17 How to request graphs, tables, and statistics in SPSS

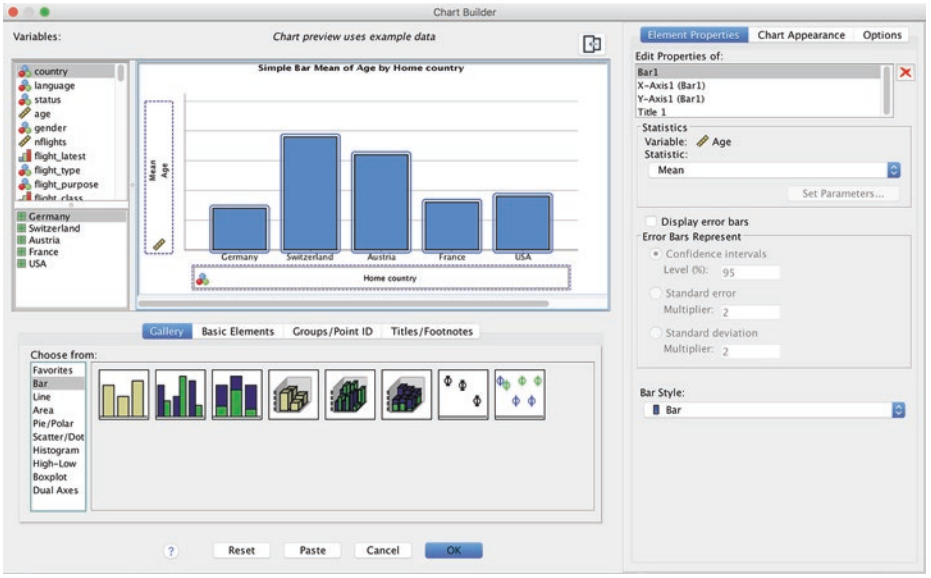


Fig. 5.18 Chart builder (bar charts)

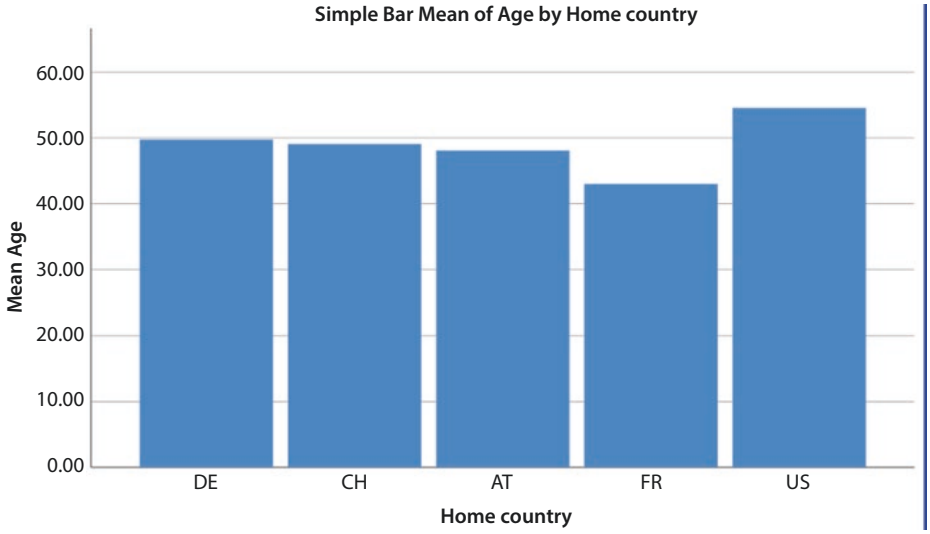


Fig. 5.19 A bar chart

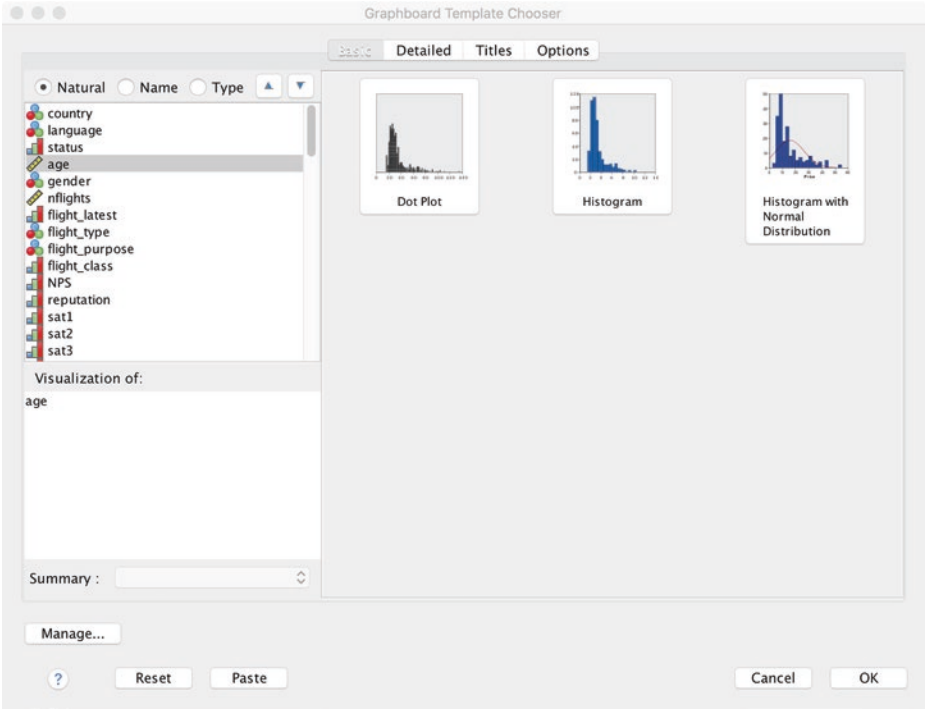


Fig. 5.20 Graphboard template chooser (histogram)

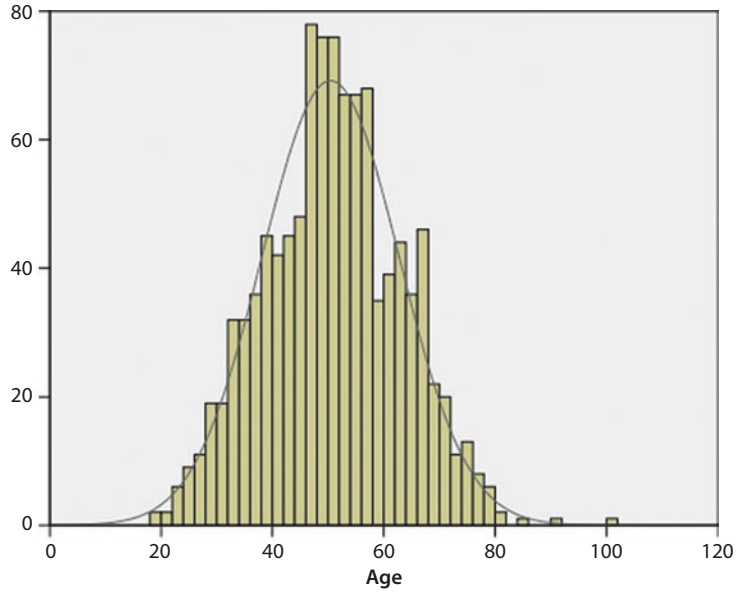


Fig. 5.21 A histogram

When selecting the relevant variable *age* on the left side of the dialog box, SPSS will show different graph options to choose from. Select **Histogram with Normal Distribution** and click on **OK**. SPSS will produce a histogram as shown in Fig. 5.21.

- **Box Plot**

To ask for a box plot, go to ► Graphs ► Chart Builder and select **Boxplot** from the **Choose from** list. SPSS allows you to choose among three boxplot types; select the one to the very right (**1-D Boxplot**) and drag this to the chart builder. Then drop *age* to the *y*-axis. When clicking on **OK**, SPSS will produce output similar to Fig. 5.22.

- **Pie Charts**

Pie charts are useful for displaying categorical or binary variables. We will create a pie chart for the *status* variable by going to ► Graphs ► Chart Builder. Select **Pie/Polar** from the **Choose from** list and drag this to the chart builder. Then drag and drop *status* to the *x*-axis. When clicking on **OK**, SPSS will show a pie chart. By double-clicking on the pie chart, we can open the **Chart Editor**, which allows changing the format and adding further statistics. For example, by selecting the target symbol and clicking on a slice of the chart, SPSS will display the relative frequencies (Fig. 5.23).

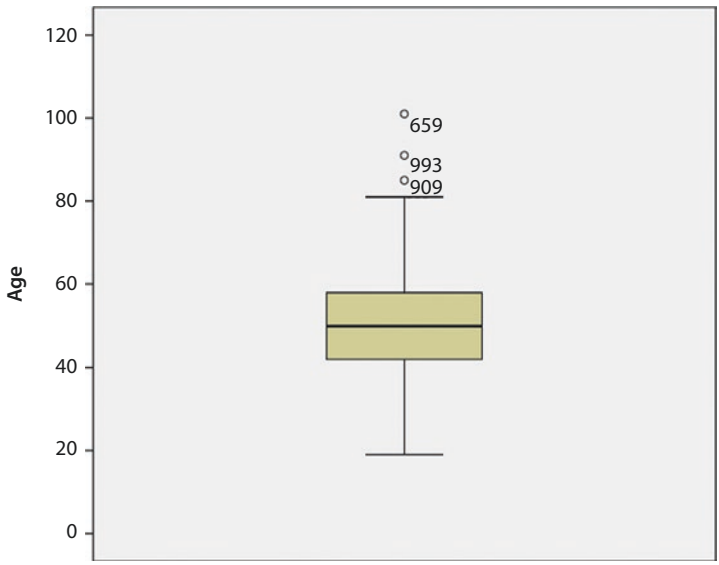


Fig. 5.22 A box plot

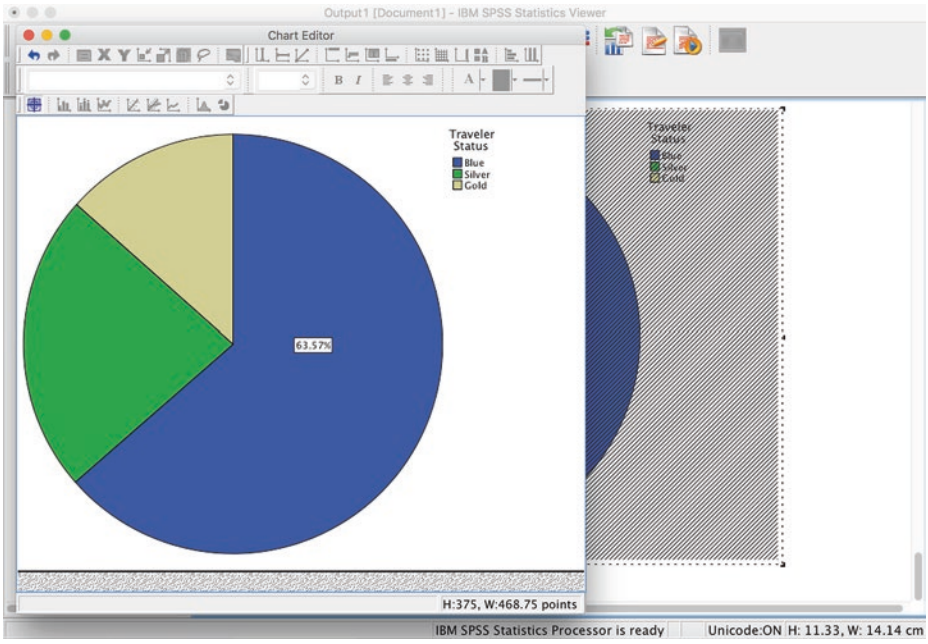


Fig. 5.23 Pie chart in the chart editor

Table 5.8 Example of a frequency table in SPSS

		country			
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Germany	695	65.3	65.3	65.3
	Switzerland	66	6.2	6.2	71.5
	Austria	108	10.1	10.1	81.6
	France	1	.1	.1	81.7
	USA	195	18.3	18.3	100.0
	Total	1065	100.0	100.0	

- **Frequency Tables**

We can produce a frequency table by clicking on ► Analyze ► Descriptive Statistics ► Frequencies. Move the variable *country* into the **Variable(s)** box and then click on **OK**. This operation will produce Table 5.8, which displays the value of each country with the corresponding absolute number of observations (i.e., **Frequency**), the relative values, including and excluding missing values (i.e., **Percent** and **Valid Percent**), as well as the cumulative relative values (i.e., **Cumulative Percent**). It shows that **65.3 %** of our sample consists of travelers who reside in Germany, followed by travelers from the United States (**18.3 %**), Austria (**10.1 %**), Switzerland (**6.2 %**), and, finally, France (**0.1 %**).

5.10.2.2 Univariate Statistics

Another useful way of examining your data is through the **Explore** option, which you can find under ► Analyze ► Descriptive Statistics ► Explore. Selecting this menu option opens a dialog box similar to that in Fig. 5.24. We will use this option to request a series of statistics for *age*, differentiated by *gender*. To do so, move *age* into the **Dependent List** box and *gender* into the **Factor List** box, which indicates the grouping variable. Under **Display**, select **Statistics** and click on **OK**.

SPSS will produce an output similar to Table 5.9, which includes a variety of measures of centrality and dispersion, such as the 95 % confidence interval, the 5 % trimmed mean, the variance, or the interquartile range.

5.10.2.3 Bivariate Graphs and Tables

- **Scatter Plots**

Scatter plots can be easily displayed in SPSS using the **Chart Builder** or the **Graphboard Template Chooser**. For example, to generate a scatter plot, go to ► Graphs ► Chart Builder and select **Scatter/Dot** from the **Choose from** list and drag the upper leftmost (**simple scatter**) into the chart builder box. Next, drag and drop *nps* to the *y*-axis and *age* to the *x*-axis. Clicking on **OK** will produce a scatter plot like in Fig. 5.25.

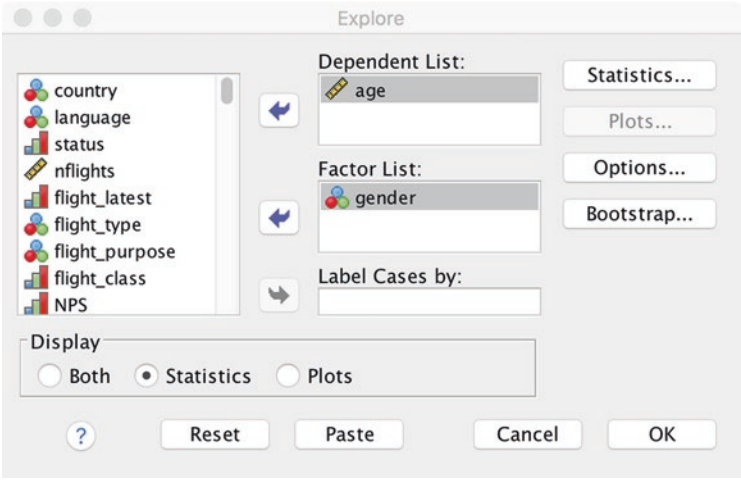


Fig. 5.24 Explore dialog box

Table 5.9 Example of a summary table using the explore option

Descriptives

	gender			Statistic	Std. Error
age	female	Mean		50.81	.761
		95 % Confidence Interval for Mean	Lower Bound	49.32	
			Upper Bound	52.31	
		5 % Trimmed Mean		50.88	
		Median		51.00	
		Variance		162.188	
		Std. Deviation		12.735	
		Minimum		22	
		Maximum		78	
		Range		56	
		Interquartile Range		18	
		Skewness		–.133	.146
		Kurtosis		–.591	.290

Table 5.9 (Continued)

male	Mean		50.28	.432
	95 % Confidence Interval for Mean	Lower Bound	49.43	
		Upper Bound	51.13	
	5 % Trimmed Mean		50.19	
	Median		50.00	
	Variance		146.684	
	Std. Deviation		12.111	
	Minimum		19	
	Maximum		101	
	Range		82	
	Interquartile Range		16	
	Skewness		.151	.087
	Kurtosis		.065	.174

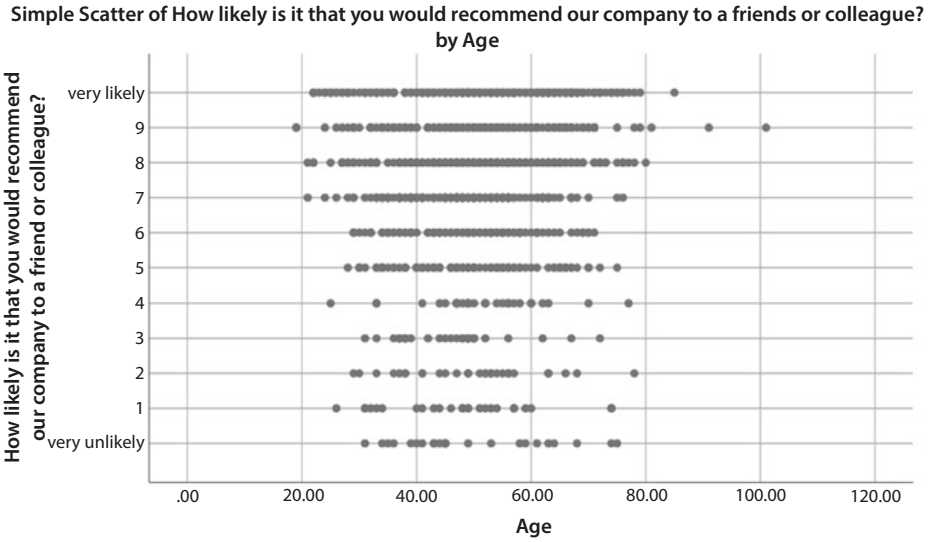


Fig. 5.25 A scatter plot

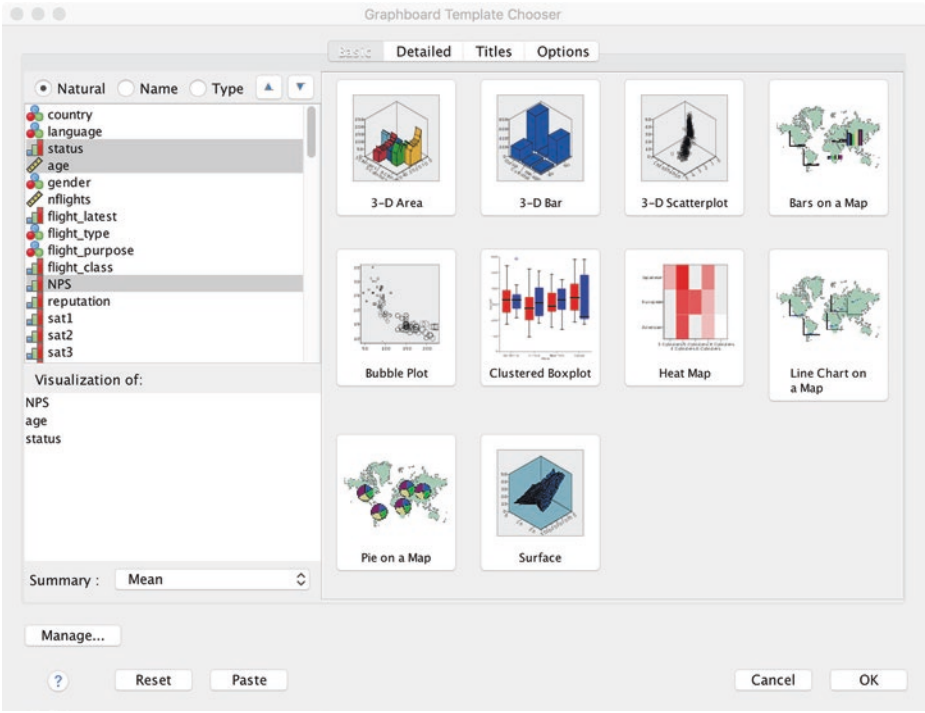
Similarly, we can request a bubble plot that considers a third variable in the scatter plot. Go to Graphs ► Graphboard Template Chooser and select the variables *age*, *nps*, and *status* from the variable list.

Tip

To select multiple variables, simply hold the STRG key (PC) or cmd (or Command) key (Mac) and left-click on the variables you want to select..

SPSS shows several graph types to choose from, including a **3-D Area** plot, **3-D Scatterplot**, and a **Bubble Plot** (◘ Fig. 5.26). Choose the latter and click on **OK**.

The design of the bubble plot depends on the order in which you select the variables. The first variable defines the *x*-axis, the second variable defines the *y*-axis, and the third variable defines the size of the bubbles.



◘ **Fig. 5.26** Graphboard template chooser (bubble plot)

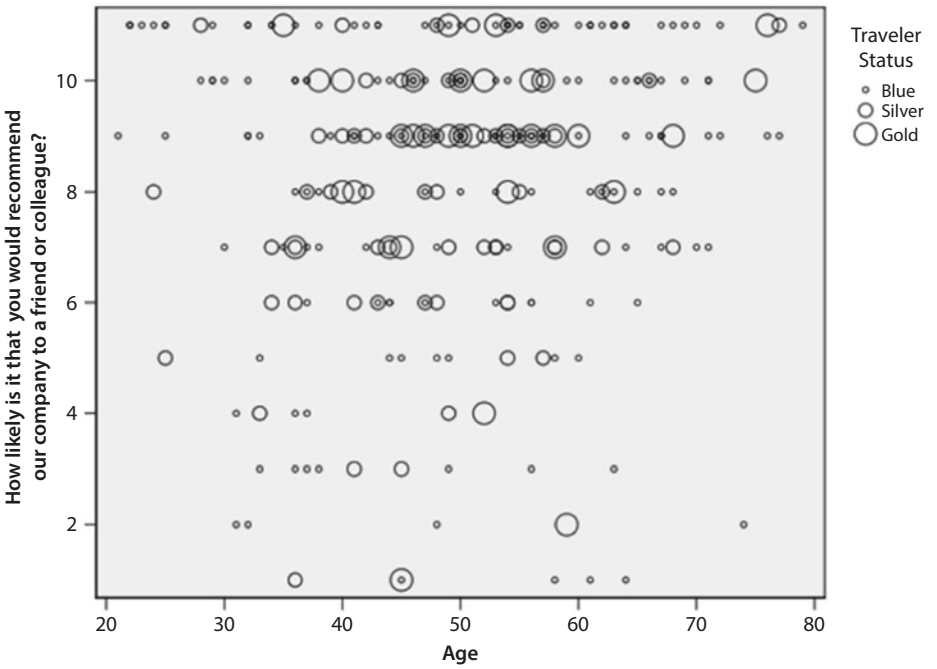


Fig. 5.27 Bubble plot for a subset of data

This will produce the bubble plot shown in Fig. 5.27. The graph shows the relationship between the respondents' age (*x*-axis) and the net promoter score (*y*-axis) differentiated by the travelers' status (size of the dots). Note that for illustrative purposes, the plot shown in Fig. 5.27 does not show the entire data but only a subset of 25 %, which we drew randomly using the select cases command (► Data ► Select Cases).

- **Cross tabulation**

Cross tabulation is useful for understanding the relationship between two variables scaled on a nominal or ordinal scale. To create a crosstab, go to ► Analyze ► Descriptive Statistics ► Crosstabs. It is important that you specify which variable goes in the column and which in the rows. Choose *country* under **Row(s)** and *gender* under **Column(s)**. Next, click on **Cells** and choose **Row** and **Column** under **Percentages**. Clicking on **Continue** and **OK** will produce a table similar to the one in Table 5.10.

5.10.2.4 Bivariate Statistics: Correlations and Covariances

In SPSS, we can calculate bivariate correlations by going to ► Analyze ► Correlate ► Bivariate. In the dialog box that opens (Fig. 5.28), select the variables to be considered in the analysis. For example, enter *s1, s2, s3*, and *s4* in the **Variables** box. When clicking on **OK**, SPSS will produce a correlation matrix like the one in Table 5.11.

Table 5.10 Example of a crosstab

country * gender Crosstabulation

			gender		
			female	male	Total
country	Germany	Count	180	515	695
		% within country	25.9 %	74.1 %	100.0 %
		% within gender	64.3 %	65.6 %	65.3 %
	Switzerland	Count	17	49	66
		% within country	25.8 %	74.2 %	100.0 %
		% within gender	6.1 %	6.2 %	6.2 %
	Austria	Count	25	83	108
		% within country	23.1 %	76.9 %	100.0 %
		% within gender	8.9 %	10.6 %	10.1 %
	France	Count	1	0	1
		% within country	100.0 %	0.0 %	100.0 %
		% within gender	0.4 %	0.0 %	0.1 %
	USA	Count	57	138	195
		% within country	29.2 %	70.8 %	100.0 %
		% within gender	20.4 %	17.6 %	18.3 %
Total		Count	280	785	1065
		% within country	26.3 %	73.7 %	100.0 %
		% within gender	100.0 %	100.0 %	100.0 %

The correlation matrix shows the correlation between each pairwise combination of three variables. For example, the correlation between *s1* (" … with Oddjob Airways you will arrive on time.") and *s2* (" … the entire journey with Oddjob Airways will occur as booked.") is **0.739**, which indicates a strong relationship according to Cohen (1988). As indicated by the two asterisks, the correlation is significant at a 1 % level.

Alternatively, you can let SPSS display a covariance matrix. To do so, go to ► Analyze ► Correlate ► Bivariate and click on **Options**. Under Statistics, tick the box **Cross-product deviations and covariances** and run the analysis.

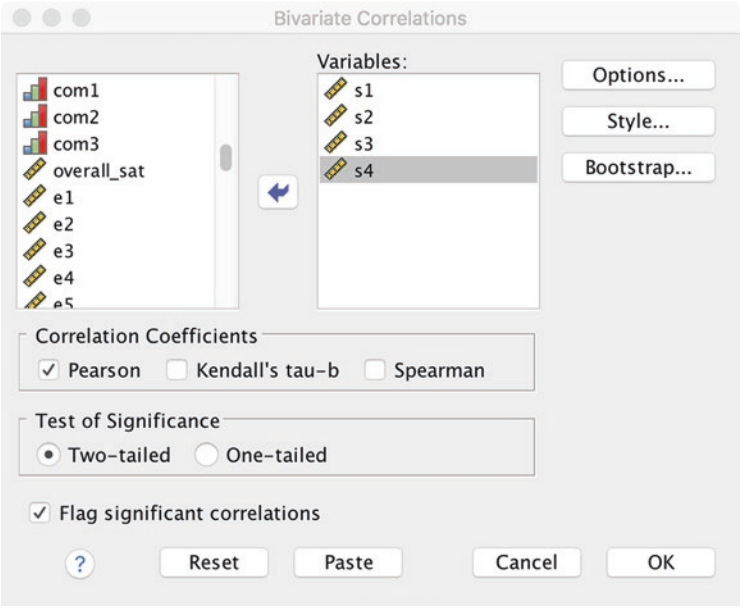


Fig. 5.28 Correlate dialog box

Table 5.11 Correlation matrix

Correlations

		s1	s2	s3	s4
s1	Pearson Correlation	1	.739**	.619**	.717**
	Sig. (2-tailed)		.000	.000	.000
	N	1038	1037	952	1033
s2	Pearson Correlation	.739**	1	.694**	.766**
	Sig. (2-tailed)	.000		.000	.000
	N	1037	1040	952	1034
s3	Pearson Correlation	.619**	.694**	1	.645**
	Sig. (2-tailed)	.000	.000		.000
	N	952	952	954	951
s4	Pearson Correlation	.717**	.766**	.645**	1
	Sig. (2-tailed)	.000	.000	.000	
	N	1033	1034	951	1035

** Correlation is significant at the 0.01 level (2-tailed).

5.11 Cadbury and the UK Chocolate Market (Case Study)

Case Study

The UK chocolate market is expected to be £6.46 billion in 2019. Six subcategories of chocolates are used to identify the different chocolate segments: boxed chocolate, molded bars, seasonal chocolate, count lines, straight lines, and "other."

To understand the UK chocolate market for molded chocolate bars, we have a dataset (chocolate.sav) that includes a large supermarket's weekly sales of 100 g molded chocolate bars from January 2016 onwards. This data file can be downloaded from the book's ⭳ Web Appendix (→ Downloads). This file contains a set of variables. Once you have opened the dataset, you will see the set of variables. The first variable is *week*, indicating the week of the year and starts with Week 1 of January 2016. The last observation for 2016 ends with observation 52, but the variable continues to count onwards for 16 weeks in 2017.

The next variable is *sales*, which indicates the weekly sales of 100 g Cadbury bars in £. Next, four price variables are included, *price1-price4*, which indicate the price of Cadbury, Nestlé, Guylian, and Milka in £. Next, *advertising1-advertising4* indicate the amount of £ the supermarket spent on advertising each product during that week.

A subsequent block of variables, *pop1-pop4*, indicate whether the products were promoted in the supermarket by means of point of purchase advertising. This variable is measured as yes/no. Variables *promo1-promo4* indicate whether the product was put at the end of the supermarket aisle, where it is more noticeable. Lastly, *temperature* indicates the weekly average temperature in degrees Celsius.

You have been tasked with providing descriptive statistics for a client by means of this dataset. To help you with this task, the client has prepared a number of questions:

1. Do Cadbury's chocolate sales vary substantially across different weeks? When are Cadbury's sales at their highest? Please create an appropriate graph to illustrate any patterns.
2. Please tabulate point-of-purchase advertising for Cadbury against point-of-purchase advertising for Nestlé. In addition, create a few more crosstabs. What are the implications of these crosstabs?
3. How do Cadbury's sales relate to the price of Cadbury? What is the strength of the relationship?
4. Which descriptive statistics are appropriate for describing the usage of advertising? Which statistics are appropriate for describing point-of-purchase advertising?

5.12 Review Questions

1. Imagine you are given a dataset on car sales in different regions and are asked to calculate descriptive statistics. How would you set up the analysis procedure?
2. What summary statistics could best be used to describe the change in profits over the last five years? What types of descriptive statistics work best to determine the market shares of five different types of insurance providers? Should we use just one or multiple descriptive statistics?
3. What information do we need to determine if a case is an outlier? What are the benefits and drawbacks of deleting outliers?
4. Download a codebook of the Household Income and Labour Dynamics in Australia (HILDA) Survey at: http://melbourneinstitute.unimelb.edu.au/hilda/for-data-users/user-manuals. Is this codebook clear? What do you think of its structure?

References

Agarwal, C. C. (2013). *Outlier analysis*. New York, NY: Springer.
Agresti, A., & Finlay, B. (2014). *Statistical methods for the social sciences* (4th ed). Upper Saddle River, NJ: Pearson.
Barchard, K. A., & Verenikina, Y. (2013). Improving data accuracy: Electing the best data checking technique. *Computers in Human Behavior*, *29*(50), 1917–1912.
Barchard, K. A., & Pace, L. A. (2011). Preventing human error: The impact of data entry methods on data accuracy and statistical results. *Computers in Human Behavior*, *27*(5), 1834–1839.
Baumgartner, H., & Steenkamp, J.-B. E. M. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, *38*(2), 143–156.
Carpenter, J. & Kenward, M. (2013). *Multiple imputation and its application*. New York, NJ: John Wiley.
Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
Drolet, A. L., & Morrison, D. G. (2001). Do we really need multiple-item measures in service research? *Journal of Service Research*, *3*(3), 196–204.
Eekhout, I., de Vet, H. C. W., Twisk, J. W. R., Brand, J. P. L., de Boer, M. R., & Heymans, M. W. (2014). Missing data in a multi-item instrument were best handled by multiple imputation at the item score level. *Journal of Clinical Epidemiology*, *67*(3), 335–342.
Gladwell, M. (2008). *Outliers: The story of success*. New York, NY: Little, Brown, and Company.
Graham, J. W. (2012). *Missing data: Analysis and design*. Berlin et al.: Springer.
Grotenhuis, M., & Visscher, C. (2014). *How to use SPSS syntax: An overview of common commands*. Thousand Oaks, CA: Sage.
Hair, J. F., Jr., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis* (8th ed.). Boston, MA: Cengage.
Harzing, A. W. (2005). Response styles in cross-national survey research: A 26-country study. *International Journal of Cross Cultural Management*, *6*(2), 243–266.
Johnson, T., Kulesa, P., Lic, I., Cho, Y. I., & Shavitt, S. (2005). The relation between culture and response styles. Evidence from 19 countries. *Journal of Cross-Cultural Psychology*, *36*(2), 264–277.
Krippendorff, K. (2012). *Content analysis: An introduction to its methodology*. Thousand Oaks, CA: Sage.
Little, R. J. A. (1998). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, *83*(404), 1198–1202.
Paulsen, A., Overgaard, S., & Lauritsen, J. M. (2012). Quality of data entry using single entry, double entry and automated forms processing—An example based on a study of patient-reported outcomes. *PloS ONE*, *7*(4), e35087.
Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NJ: Wiley.
Sarstedt, M., Diamantopoulos, A., Salzberger, T., & Baumgartner, P. (2016). Selecting single items to measure doubly-concrete constructs: A cautionary tale. *Journal of Business Research*, *69*(8), 3159–3167.
Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London, UK: Chapman & Hall.
White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, *30*(4), 377–399.

Further Reading

Huck, S. W. (2014). *Reading statistics and research* (6th ed.). Harlow: Pearson Education.
Levesque, R., Programming and data management for IBM SPSS Statistics 20. Chicago, SPSS, Inc. Available at http://www.spsstools.net/en/resources/spss-programming-book/
SticiGui at http://www.stat.berkeley.edu/~stark/SticiGui/Text/correlation.htm