CHAPTER

3

Exploratory Data Analysis

What Data Do I Have?

The first thing to do with any data set is to get to know it. This is done not only to familiarize yourself with all the data you have collected, but also to reduce the workload during analysis. The initial data investigation has been termed *exploratory data analysis* or EDA and it primarily focuses on visually inspecting the data. The main aim of EDA is to understand what data you have, what possible trends there are, and therefore which statistical tests will be appropriate to use.

Figure 3-1 shows the suggested process to follow when conducting EDA.

V. Cox, *Translating Statistics to Make Decisions*, DOI 10.1007/978-1-4842-2256-0_3

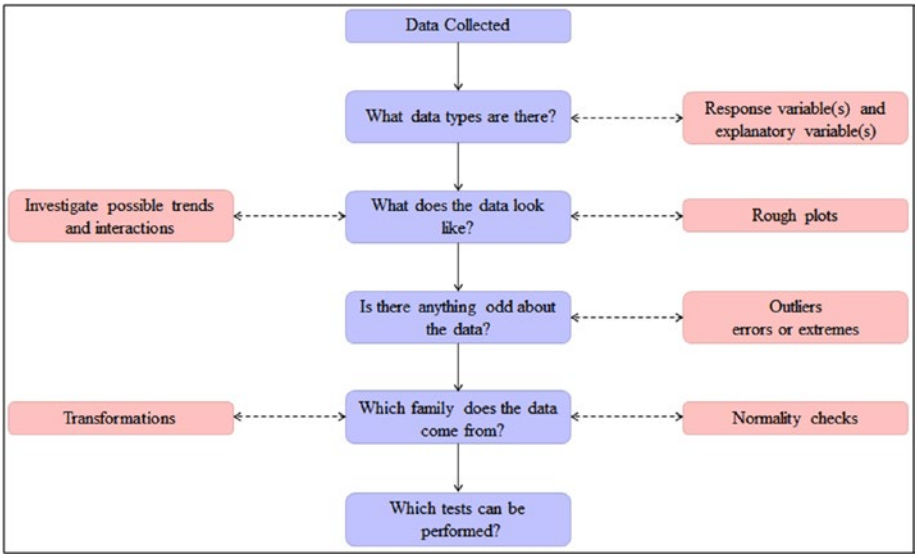

Figure 3-1. *Exploratory data analysis (EDA) process*

The following sections delve into more detail for each of the steps shown in Figure 3-1. However the general idea is to identify the data types you have for each variable, for example, whether the data is continuous or discrete will lead to which plots can be created. These plots are initial investigations and Chapter 9 goes into more explanations about how to make clear, concise graphs to present to a customer as opposed to these quick and dirty inspections.

The next step is to identify any unusual data points and establish whether they are real outliers by using the plots. Once these have been dealt with the family of the data needs to be classified (e.g., normal distribution). By carrying out all these steps you can then move on to the final step that determines which tests can be performed on the data, but this is covered in other chapters.

Data Types

The type of data being collected should have been considered during both the design of experiments and data collection phase, however it is good practice to verify and if you haven't been involved from the beginning of the study it's a good place to start.

The main way to classify data types is into quantitative data or qualitative data; however the data also can be classed as univariate, bivariate, or multivariate. The latter three terms simply refer to the number of variables being recorded: uni (one), bi (two), multi (three or more).

In addition there are also the classifiers of objective or subjective data, which were mentioned in the previous chapters. The data being objective or subjective won't affect the tests used during analysis, however it will affect the assumptions and statements made in the conclusions.

Quantitative Data

Quantitative data is the term given to any data that is recorded as a numerical value. The subcategories within this are continuous data and discrete data.

Continuous Data

Continuous data is data that can be recorded as any value between an interval and as such it can be recorded with decimal points and still make sense (e.g., the strength of a signal in decibels or accuracy of a projectile in meters from a target).

Although age is usually recorded in integer form, it is generally considered to be continuous data due to the fact that you can be 21.5 years, but you just wouldn't record it as such.

Discrete Data

Discrete data is data that can only be recorded as an integer; it would not make sense to have 2.5 people for instance. Other examples include the number of canine detections or the number of survey responses.

Qualitative Data

Qualitative data is the term given to any data that is non-numerical and generally subjective. Qualitative data can be assigned a number to aid with analysis. However the precise value of the number itself is meaningless. The subcategories within this are binary data, nominal data, and ordinal data.

Binary Data

Binary data has two responses such as yes/no, heads/tails, and so forth (e.g., a detector detecting a target or the outcome of flipping a coin). When using binary data in analysis it is generally coded to 0 and 1 with 1 being the measurement of interest.

Binary data is actually a special type of nominal data, one with only two categories. When discussing the graph types later, any reference to nominal data will also include the case of binary data.

Nominal Data

Nominal data also is commonly referred to as categorical data, this is data that contains multiple groups that could be given a numerical value but have no natural ordering. For example, different types of vehicles such as bike, car, truck, boat, plane, could be assigned the numbers 1 to 5 for ease of analysis. However the values themselves are meaningless and have no ordering, car has the value 1 greater than bike by assignment only and not because it's "better."

Ordinal Data

On the other hand, ordinal data also can be given a numerical value but it does have a natural ordering. For example, reactions to a chemical could be none, rash, or blistering and they could be assigned the values 1 to 3 with 3 clearly being more severe than 1, but not necessarily 3 times more severe.

Another example of ordinal data is Likert responses from questionnaires, there is a clear progression from strongly disagree to strongly agree, or the equivalent. This data is not always treated as ordinal data as it should be, and as such the results can be misleading, see more in Chapter 7.

Viewing the Data

As mentioned earlier, in this section I discuss some of the different plot types that can be created given the types of data you have, although this list is by no mean exhaustive.

In EDA graphs are drawn to "get to know the data," so it's about noticing trends and structure for testing as opposed to drawing "pretty" plots. Chapter 9 is more focused on the effective presentation of graphs to highlight messages to the customer, including R code examples for amending color, labels, and so forth.

When plotting data regardless of the type of plot, the response variable should be on the *y*-axis, the vertical axis, and the explanatory variable should be on the *x*-axis, the horizontal axis. If there are multiple explanatory variables then this can be covered by different colors, shapes, or facets on the plot, for more information on graphs see Chapter 9.

Bar Charts

Bar charts are suitable for discrete data and counts of nominal data. There are many different variations such as stacked, percentage, and so forth, but they are very good to compare the frequency of different groups.

Figure 3-2 shows an example of the counts of different car makes bought at a local garage within one month. It clearly shows the order of car type sales: Ford, Vauxhall, Audi, Nissan, and then BMW.

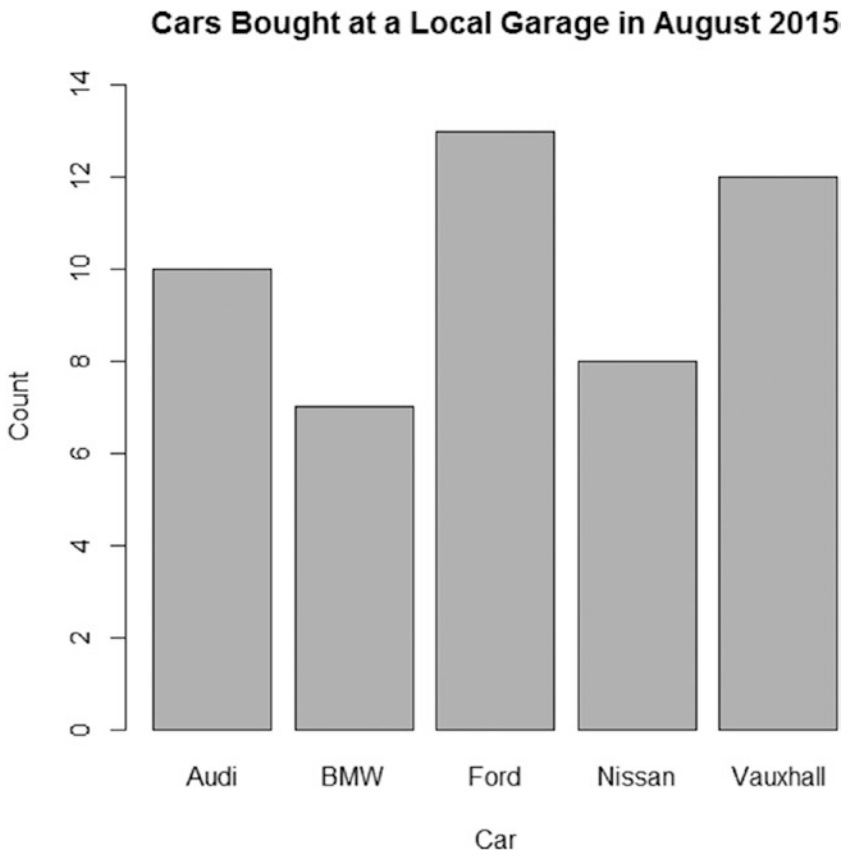

Figure 3-2. Bar chart of cars bought at a local garage

Dot Plots

Dot plots can be used in a similar manner to bar charts and are less cluttered as they show a single point as opposed to a bar. They can be used to show a single statistic, such as a mean, more clearly than a bar chart in which the filled bar below the top line of the bar would be redundant.

Figure 3-3 shows a dot plot of the average waiting times for baggage at selected United Kingdom airports, confidence intervals around the mean could be added to this plot for more information. At first glance the general trend is that the waiting times are shorter for Birmingham, East Midlands, and Manchester and are longer for Edinburgh, Gatwick, and Heathrow; which may be expected due to their size and popularity.

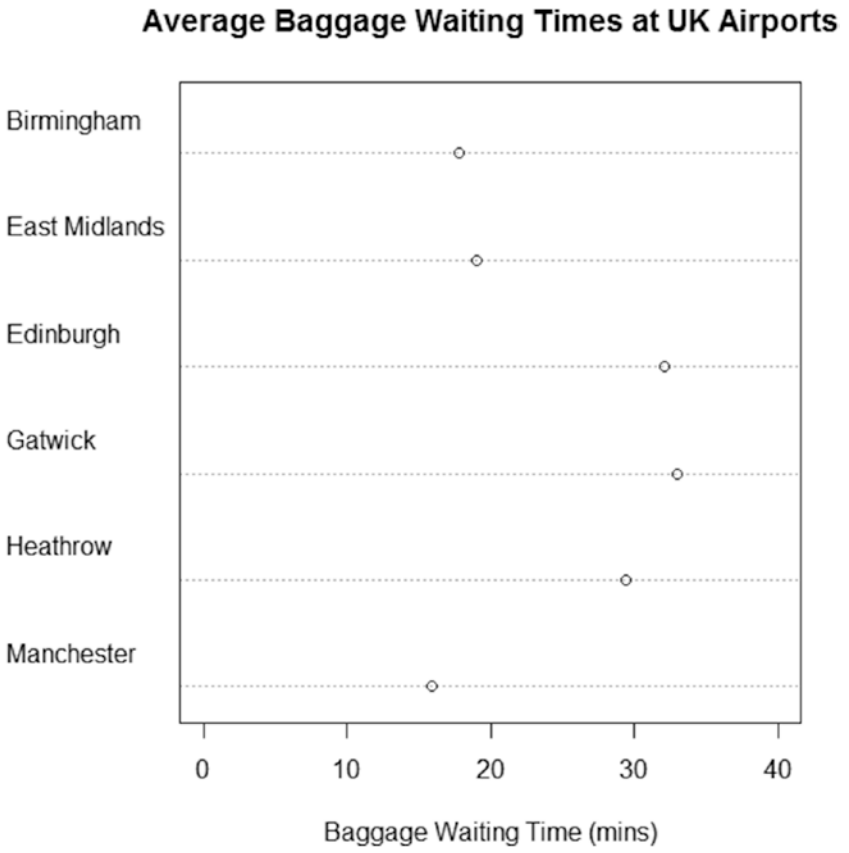

Figure 3-3. Dot plot of average waiting times for baggage at UK airports

Parallel Lines Plots

Parallel lines plots can be used to show paired data as the emphasis should be on the relative change for each subject. This information would be lost using any other plot listed. When creating a parallel lines plot the data shouldn't be stacked; there should be a separate column for subject, then two more columns for the items each person will be doing.

Figure 3-4 shows a parallel lines plot of the time to complete a task both before and after training for each subject. You can see that in all but one case the subjects completed the task quicker after receiving training.

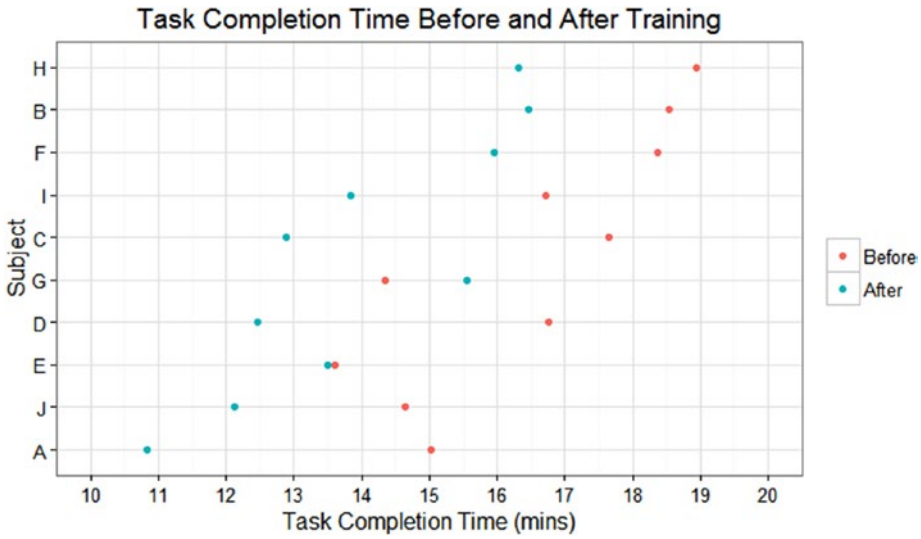

Figure 3-4. Parellel lines plot of training effect on task completion time

Histograms

Histograms are appropriate for continuous data only, they show the frequency counts given in set "bin" sizes. The software being used will choose appropriate bin sizes automatically. Histograms can be useful for highlighting distributions, such as a bell curve to suggest normality, however there is a better plot for investigating this assumption that is shown later in the chapter.

Figure 3-5 shows a histogram of heights from a sample of 100 people; in this case there is a bell curve that would suggest we could assume the data follows a normal distribution.

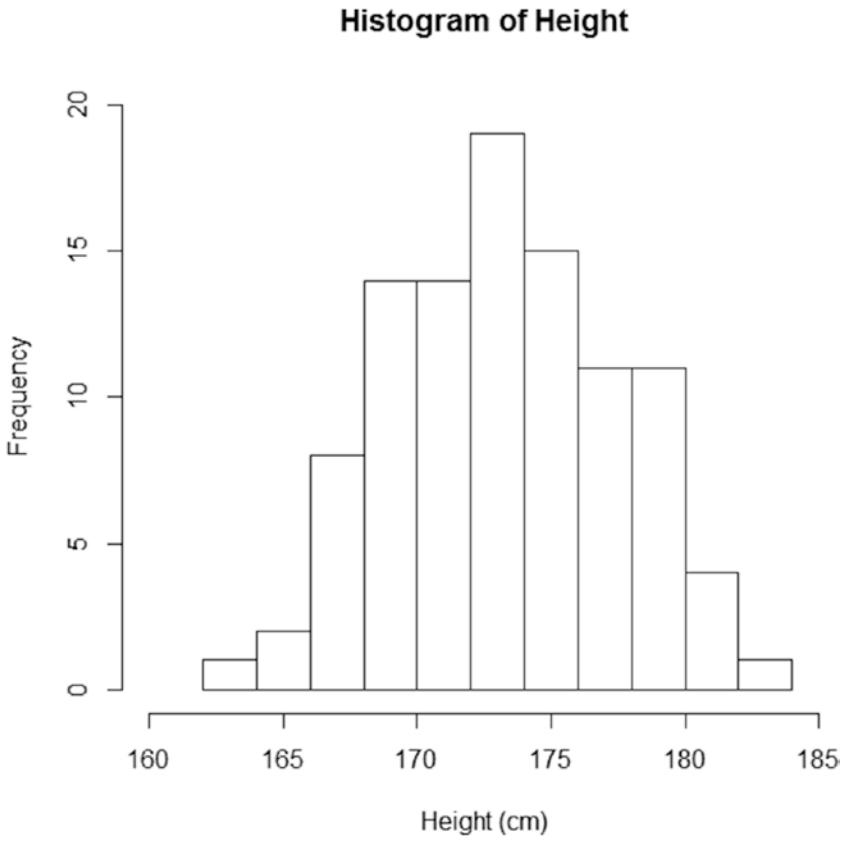

Figure 3-5. Histogram of heights sampled from 100 people

Scatter Plots

Scatter plots are useful for two continuous variables with or without a nominal data variable. These plots are handy for highlighting trends in the data as well as possible differences between any groups.

Figure 3-6 shows a scatter plot of yield by log concentration with a line of best fit. There is undoubtedly a positive trend between log concentration and yield and the points are quite close to the line of best fit. The line is plotted from a linear model, which is explained in Chapter 7.

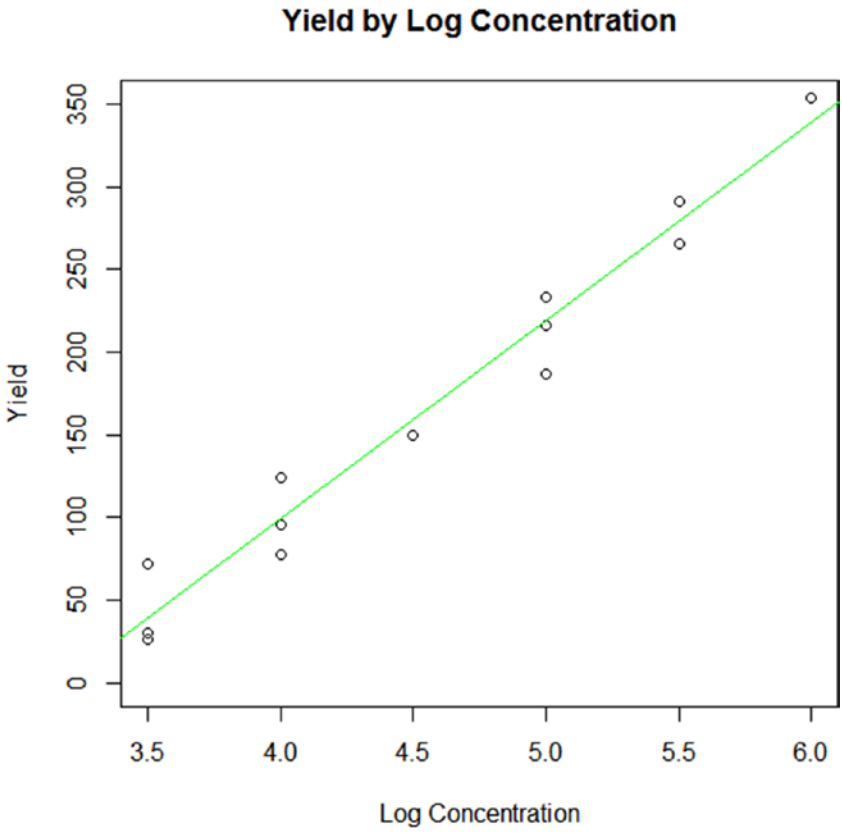

Figure 3-6. Scatter plot of yield by log concentration with line of best fit

Line Graphs

Line graphs are very similar in style to scatter plots. They are used for the same data types, but will generally have a time element across the *x*-axis. These points will be connected by a line to each individual point instead of a smooth trend.

Figure 3-7 shows a line graph of survey response rates by year; there was a sharp drop in 2011, but since then the trend seems to be picking up again.

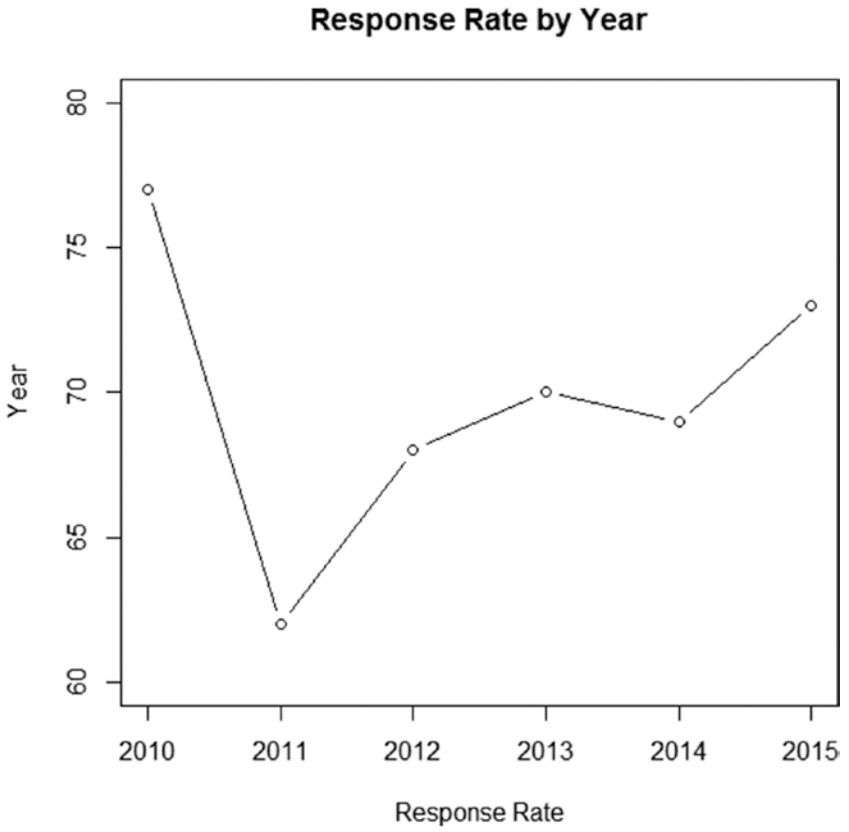

Figure 3-7. Line graph of survey response rates by year

Box Plots

Box plots are extremely useful as they can show a lot of information in a condensed manner. These plots are suitable for a continuous variable and a nominal variable. They can be used to investigate the distribution, equal sections suggest normality, however as earlier there is a better plot for this, and more important they can be very useful for highlighting differences between nominal groups.

A box plot also can be termed a box and whiskers plot as it principally concerns a box with some lines coming from each end. It contains the following information (descriptions of all of the summary statistics listed below are contained in Chapter 4):

- Median: the line within the box.
- Q1 and Q3[1]: the bottom line and top line of the box, respectively.
- IQR[2]: the length of the box itself.
- Range (minus statistical outliers): the length of the whiskers.
- Statistical outliers: any points outside the whiskers. The limits outside which a value is classed as an outlier are usually calculated as 1.5 times the IQR added to/subtracted from the quartiles.[3]

[1]Quartile 1 and Quartile 3
[2]Interquartile range
[3]Lower limit: Q1 – 1.5*IQR and upper limit: Q3 + 1.5*IQR

Figure 3-8 shows a boxplot of summer temperatures across the months June to August for popular holiday destinations. The larger the box and whiskers, the more variable the temperature, as for example Las Vegas; whereas the smaller the box and whiskers, the less variable the temperature, as for example London. In addition we also can start to see some possible differences we may find during testing, such as London and Paris having much colder temperatures than the other destinations and Florida and Las Vegas having much warmer temperatures.

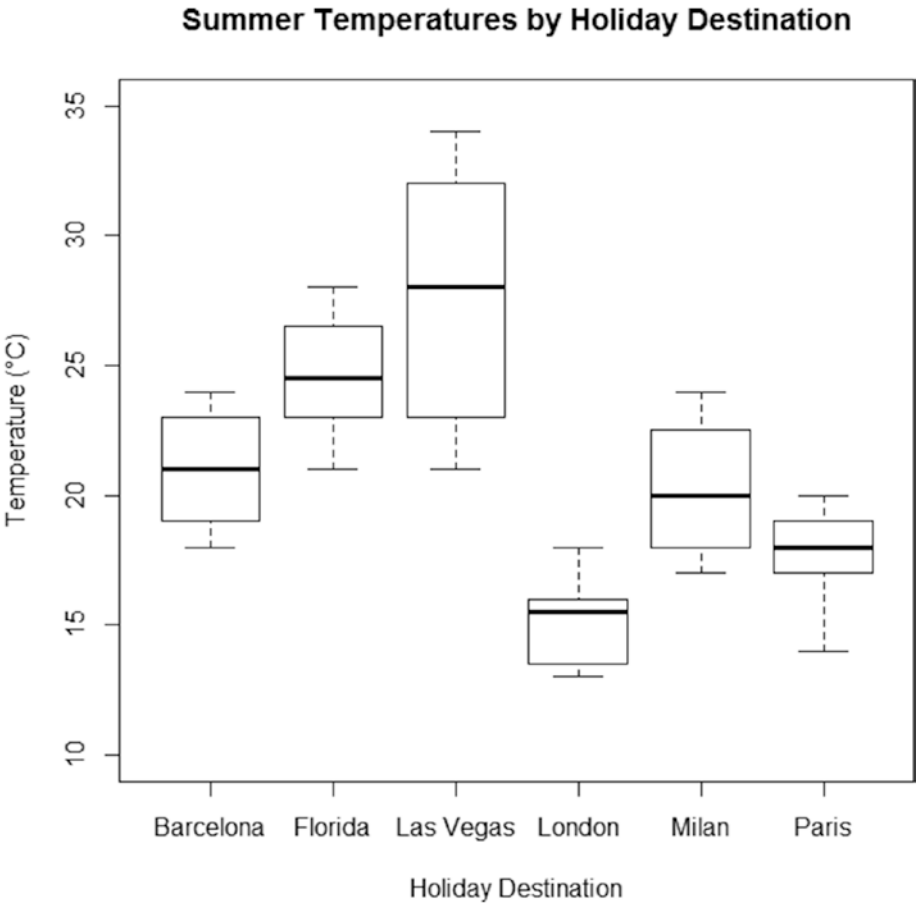

Figure 3-8. Box plot of summer temperatures by popular holiday destinations

Likert Plots

Likert plots are a clearer, more preferable way to view ordinal data rather than using bar charts. These plots are used to highlight the spread of data over the ordinal levels for different nominal groups. They can show the raw values or percentage values, which are preferable with unequal groups, along with the numerical count to the side.

Figure 3-9 shows a Likert plot of responses to three statements, "the equipment was easy to use," "the equipment was comfortable," and "the equipment was reliable over the week," for 15 participants.

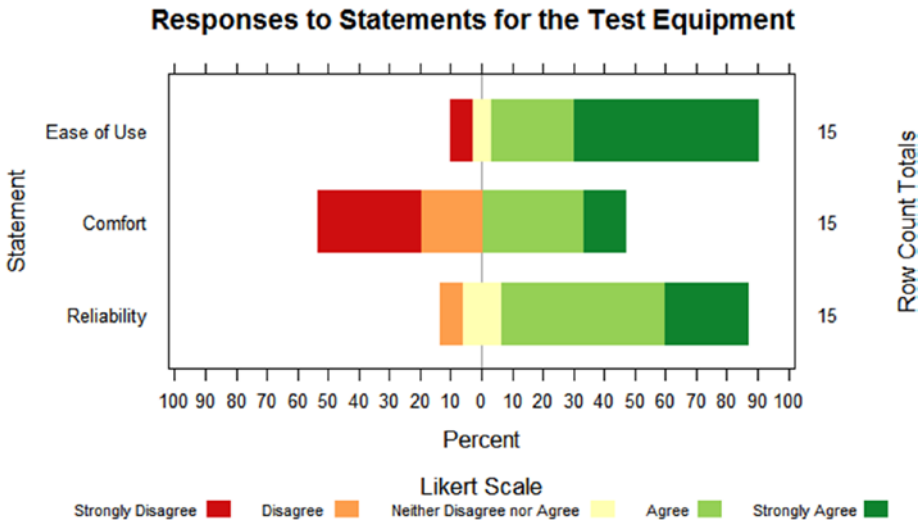

Figure 3-9. Likert plot of statement responses for the test equipment

The red and orange on the left represent the negative views, the yellow in the middle represents the neutral views, and the greens on the right represent the positive views. It can be seen that ease of use and reliability was rated more positive than negative. However, comfort was divided fairly equally between positive and negative.

Trellis Graphs

Trellis graphs are very useful for viewing multivariate data in a clear manner. This plot is limited to a maximum of two continuous variables, but can have multiple nominal variables. The limitation is dependent on the number of nominal variables given the number of levels within each variable. For example, if you had five nominal variables each with ten levels, it would not be sensible to plot them all on the trellis graph. A trellis graph creates multiple panels for nominal variables and also has the option to use shapes and colors for additional nominal variables.

Figure 3-10 shows a trellis plot of accuracy of a projectile by four explanatory variables: distance, ammunition type, operator, and target size. These four explanatory variables have 6, 2, 3, and 3 levels, respectively.

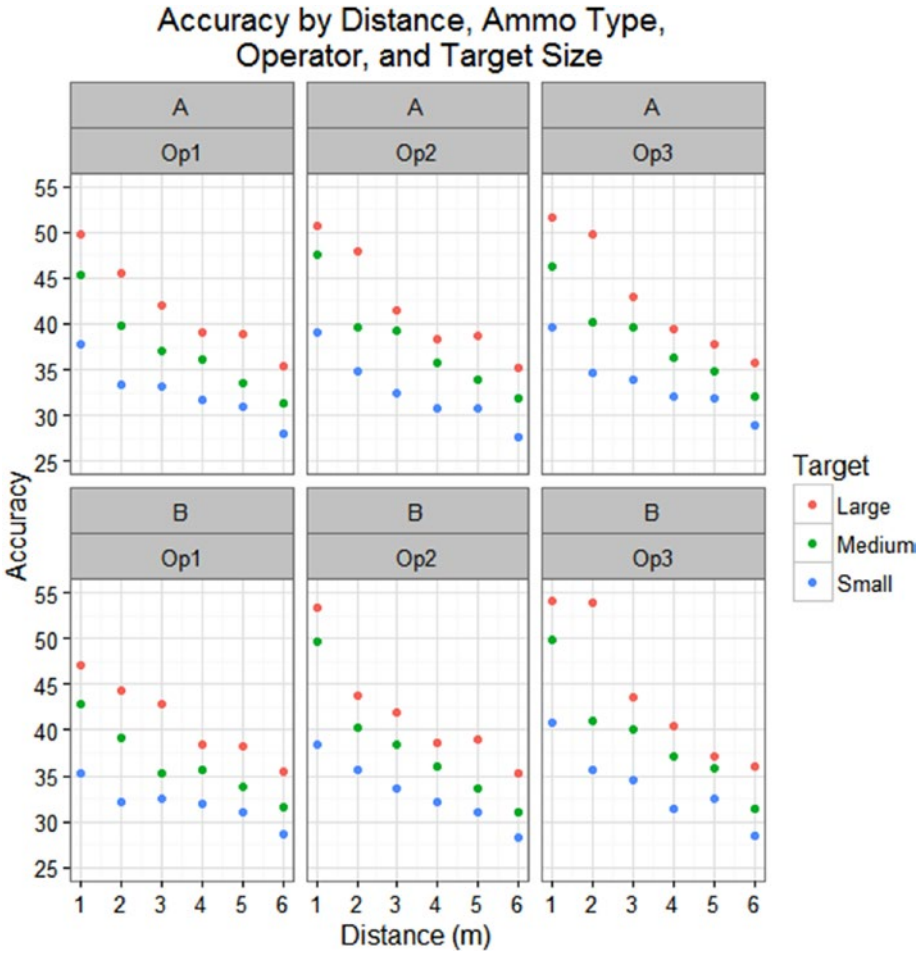

Figure 3-10. *Trellis graph of projectile accuracy by distance, ammo type, operator, and target size*

In all cases it seems that the accuracy decreased the further away from the target you got, which is probably to be expected. It also shows that the accuracy decreases as the size of the target decreases, again expected. There doesn't seem to be any major differences between the operators, comparing the left two boxes, to the middle two, to the right two boxes. There also doesn't seem to be any differences between the ammunition types, comparing the top three boxes to the bottom three boxes.

At initial glance it appears that there are no interactions of interest as there is no change of direction, such as the accuracy increasing as distance increases, in any of the boxes. There also is no change in gradient, such as the accuracy decreasing as distance decreases more rapidly, in any of the boxes; however, this should still be tested in the model.

These plots should be drawn to start to identify trends and group differences; however no conclusions should be drawn at this stage. Statistical testing needs to be undertaken to confirm the trends and add levels of uncertainty to the results due to the sample size and the data variation.

Outliers

There are two types of outliers, and each should be treated very differently otherwise any conclusions drawn from the data may be misleading. These can be classed as data entry errors/technical errors, or statistical outliers.

Data entry errors or technical errors are clearly incorrect data, such as 250% or someone who is 143. In both example cases the raw data would need to be checked to see if the true value could be obtained. However if there was no way of verifying this, then the values need to be removed from the dataset as they will heavily skew the data and influence the results.

Statistical outliers are those values that are highlighted as outliers through box plots and other similar graphs or through statistical testing for outliers such as using Grubbs' test. These outliers are possible, but are either at the extreme ends of the possible values or are just disjointed from the trend of the rest of the data. For example, someone with a height of 6'9" is a realistic value; however it will appear an outlier next to general height recordings. This should be double checked, and if the value was recorded correctly it should not be removed from the data set.

Drawing graphs can help highlight suspect data points that can then be investigated further to determine which type of outlier the points are.

For example, Figure 3-11 shows a scatter plot of the time taken to complete a race by participants of varying ages. It contains a data entry error that, if left in, would completely skew the results.

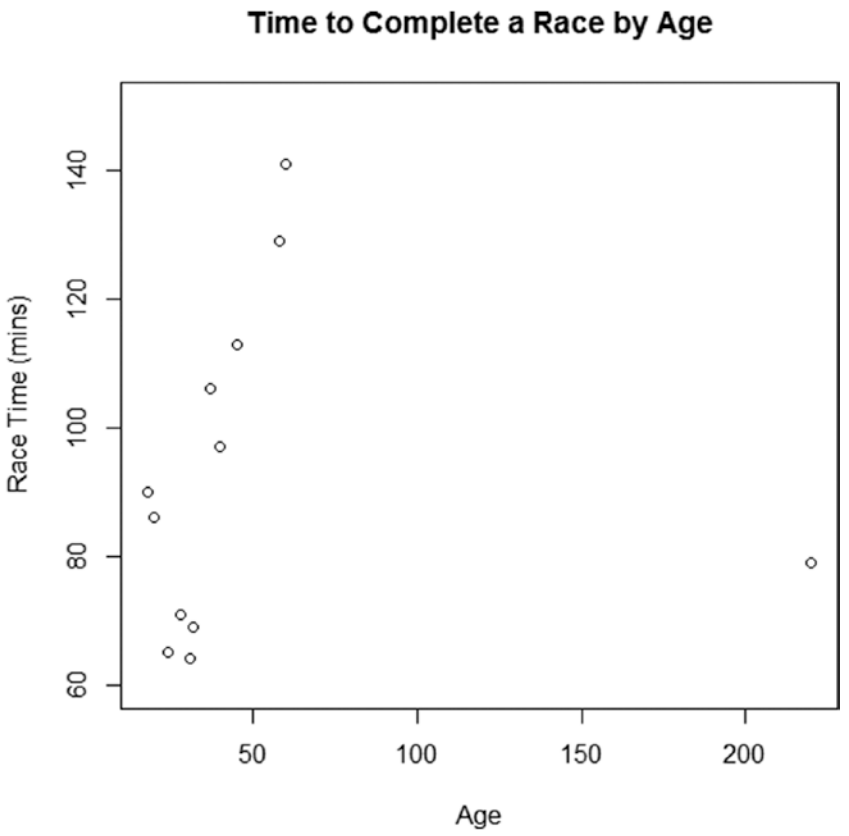

Figure 3-11. Scatter plot of time to complete a race by age, including a data entry error

Figure 3-12 shows the same dataset, but this time with the data entry error removed to highlight the difference that one data point would have made on the results. Now we can start to see a trend of race time decreasing with age until the "sweet spot" around the late twenties, and then see the race time increasing with age.

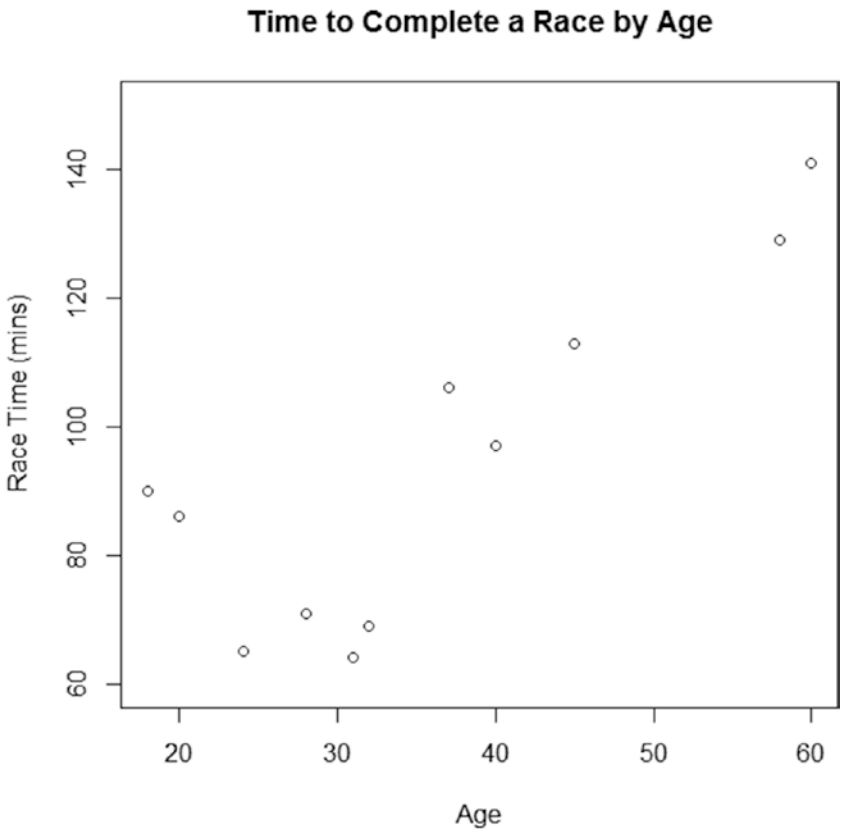

Figure 3-12. Scatter plot of time to complete a race by age, excluding the data entry error

Figure 3-13 shows a box plot of the time taken to read a 300 page book by different school year groups. This highlights a few outliers, which can occur quite often, that are shown as circles; however these are clearly statistical outliers only due to human variation and should therefore be left in and used in the analysis.

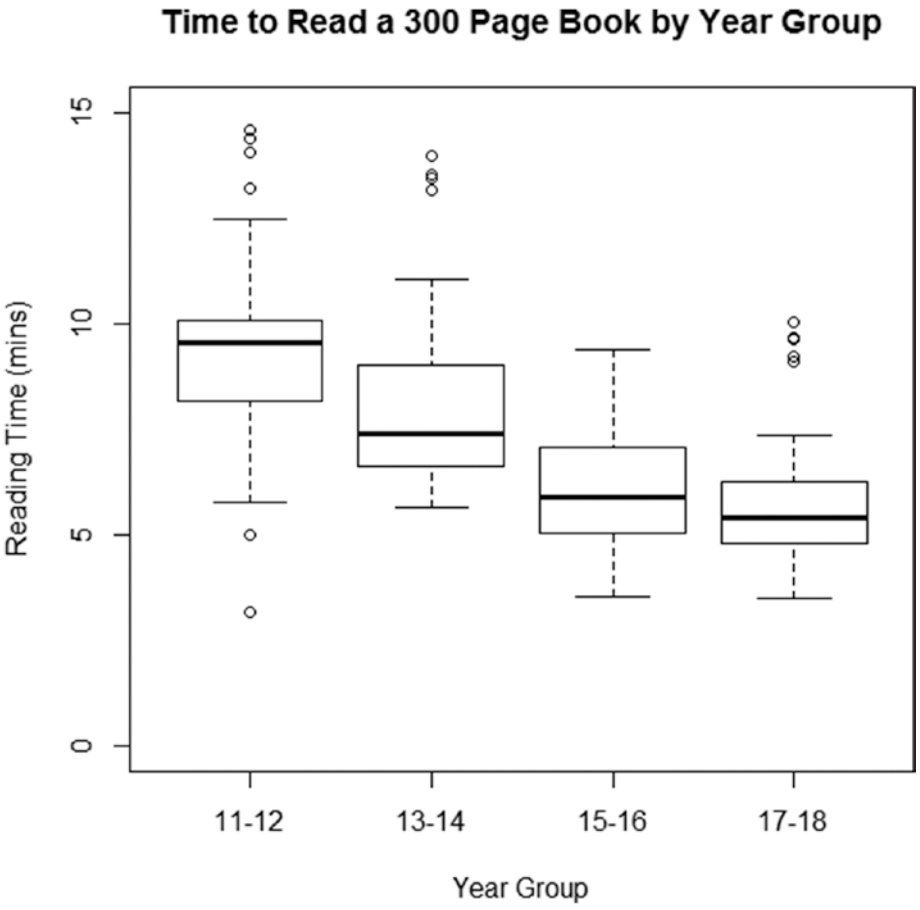

Figure 3-13. Box plot of reading time of a 300 page book by year group

It is important to remove data entry errors and technical errors so that the conclusions from the analysis have not been skewed by these points.

It is equally important to not remove genuine data, those points that are statistical outliers but are still within realistic bounds, as these need to be included to present the whole picture during the conclusions section.

Distribution

There are many different distributions, or families, that data sets can take with arguably the most well-known, for continuous data at least, being the normal distribution, also called the Gaussian distribution.

Figure 3-14 shows an example of a normal distribution that has the nice bell curve shape. You should be able to see how a histogram can aid in determining whether your data can be classed as approximately normal. The "tails" on this plot actually extend from minus infinity to plus infinity. This will not be the same for other distributions.

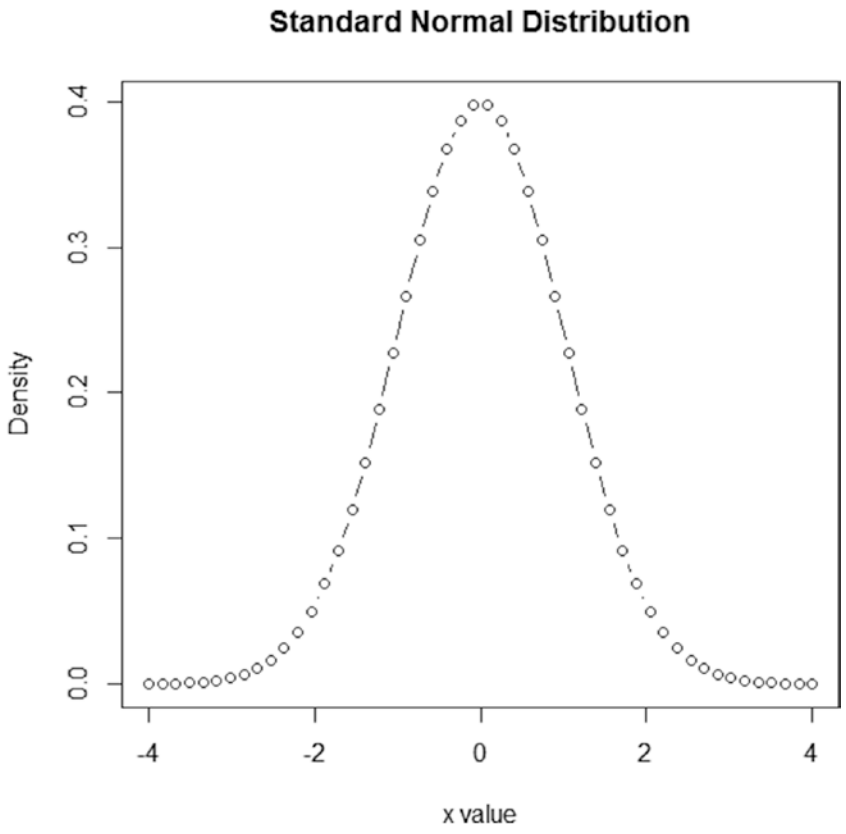

Figure 3-14. Standard normal distribution

When determining if your data follows a normal distribution the best method to use is drawing a graph rather than using formal normality tests, such as the Anderson–Darling or Shapiro–Wilk test. The most useful plot is a quantile-comparison plot, which also can be called a quantile-quantile plot or a Q–Q plot. This plot will draw the ordered values you have observed against theoretical expected values from a normal distribution.

Figure 3-15 shows an example of a Q–Q plot where we could assume normality in our data. Generally you want the points to line up nicely from the bottom left of the plot to the top right of the plot, along the $y = x$ line.

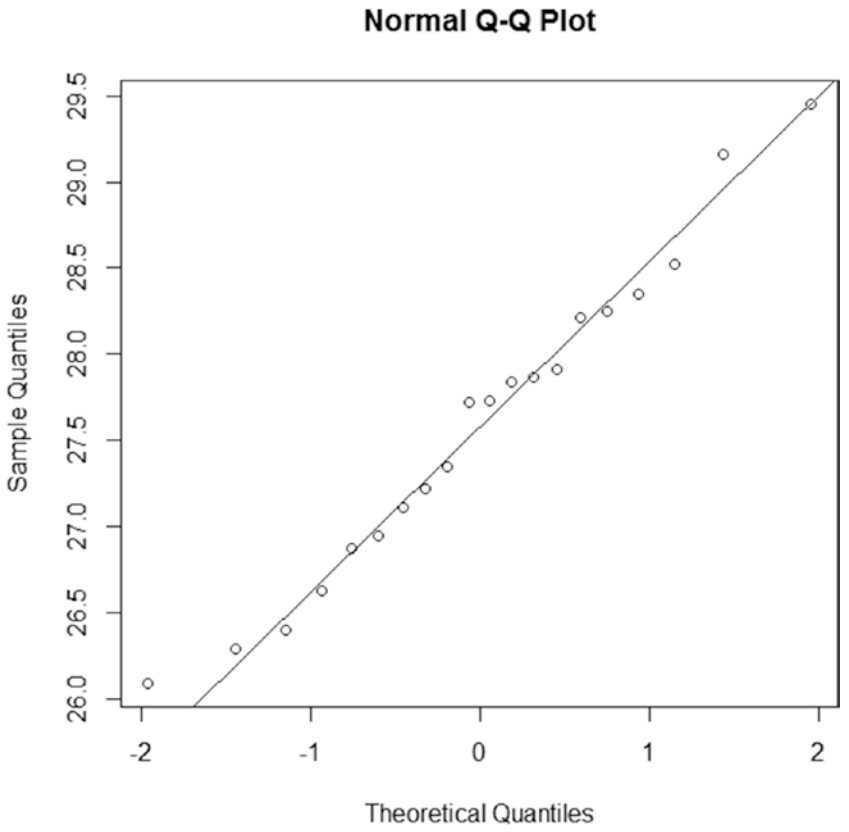

Figure 3-15. Quantile–Quantile (Q–Q) plot

You soon will notice an issue if there is curvature in the line or if the line of points is on a more horizontal line, in which case a transformation may help, see more in Chapter 4.

The reason formal tests are not recommended is that they can be misleading, especially with a large sample size. For example, in Figure 3-16 using the Q–Q plot you would be able to assume normality on that data, however the formal tests suggest strong non-normality. This is purely down to the fact that as there is such a large sample size, those slight deviations at the tails are enough for the formal test to say "non-normality."

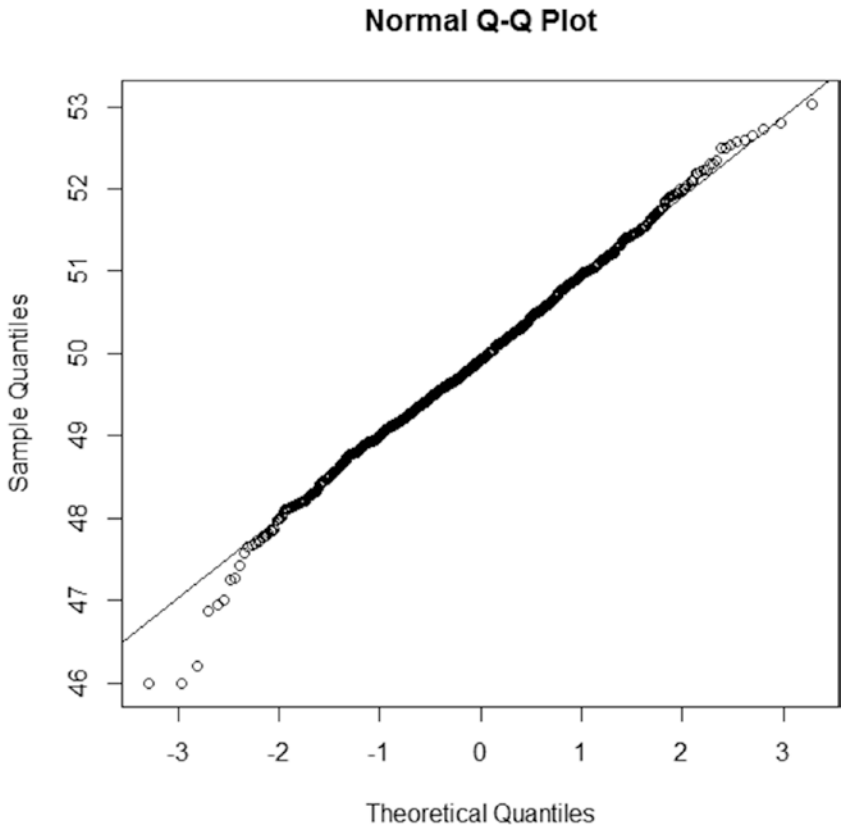

Figure 3-16. Q–Q plot of a large sample size

Other common continuous data distributions include the exponential and the gamma distributions. The exponential distribution is a special case of the gamma distribution with a fixed shape parameter of 1. A good way to think about an exponential distribution is modeling the time until an event occurs. The "tails" on this plot extend from zero to infinity.

Figure 3-17 shows an example exponential distribution with a rate of 1, the higher the rate the more the line will curve into an "L" shape.

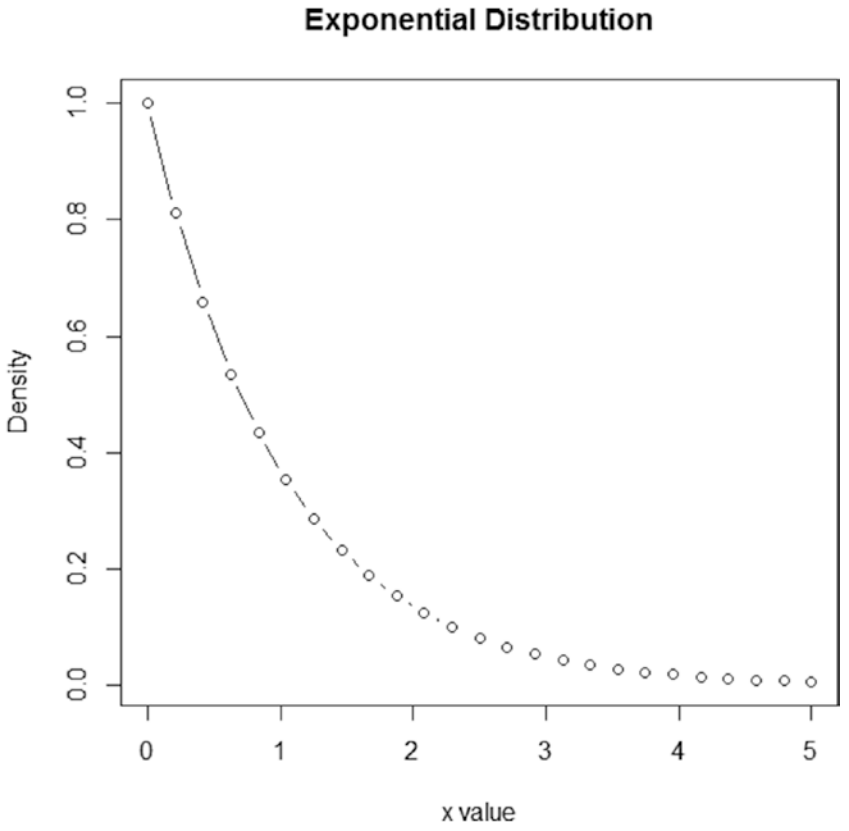

Figure 3-17. Exponential distribution with a rate of 1

The gamma distribution is a generalization of the exponential distribution and has a distinct shape, it is frequently used to model general waiting times, so modeling the time until the next *n* events.

Figure 3-18 shows an example gamma distribution with a rate of 1 and a shape of 2. Generally speaking, though it's not quite this simple, the higher the rate the more "squashed" the curve will become, and the higher the shape the further to the right the curve will move. The "tails" again extend from zero to infinity.

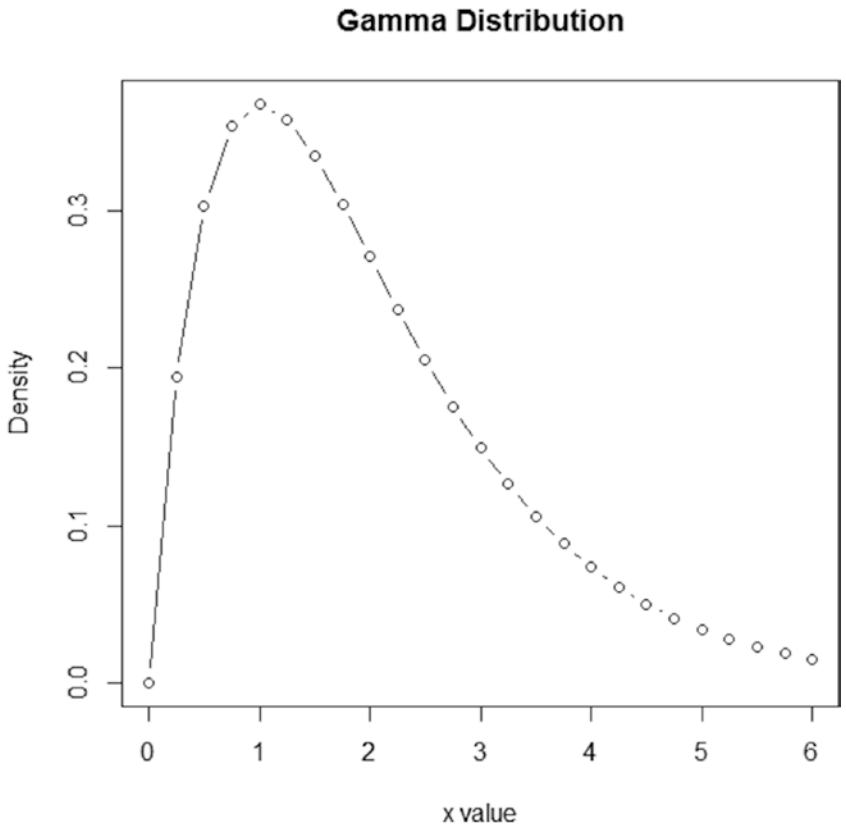

Figure 3-18. Gamma distribution with a rate of 1 and a shape of 2

The most commonly known discrete data distribution is the binomial distribution. This distribution is the probability of "success", which will be between 0 and 1 plotted against the number of trials.

Figure 3-19 shows an example binomial distribution with a size of 50 and a probability of 0.1, the size is just the number of trials, and the probability is the probability of success. The "tails" for a binomial distribution extend from zero to *n*, which in this case is 50.

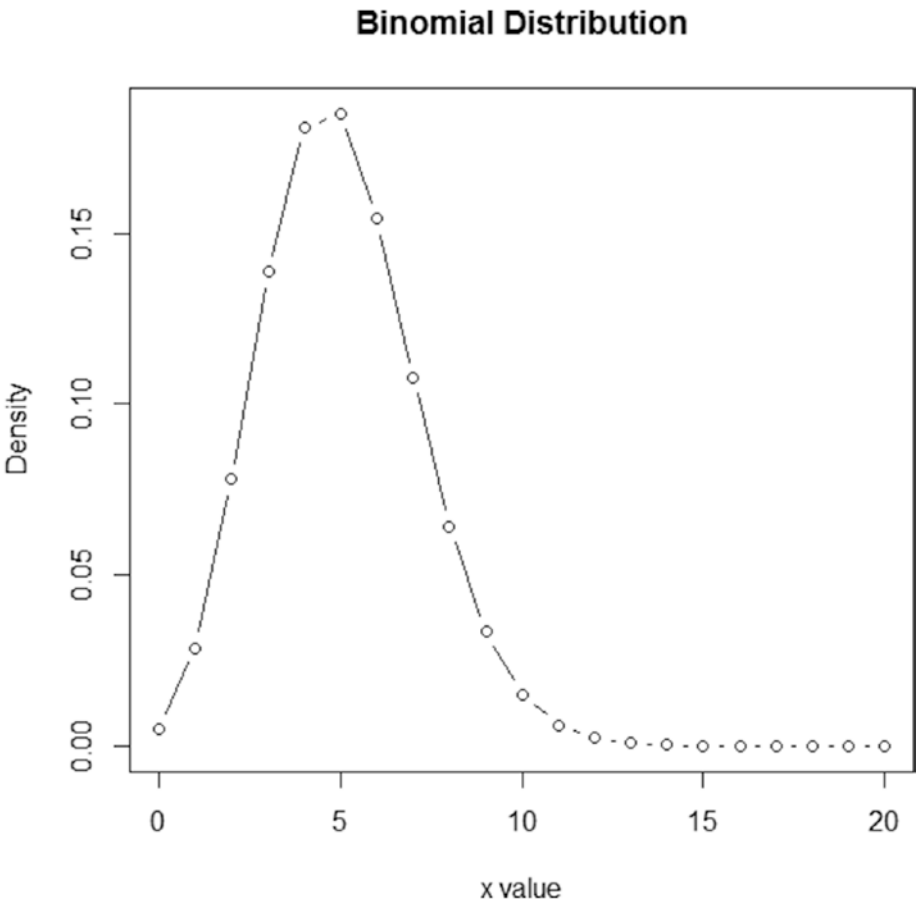

Figure 3-19. Binomial distribution with a size of 50 and a probability of 0.1

Another common distribution for discrete data is the Poisson distribution. This distribution is used for count data plotted against time, generally with there being a much larger count at the lower end of the plot.

Figure 3-20 shows an example Poisson distribution with a lambda of 3, the lambda dictates where the central peak of the curve will be. The "tails" extend from zero to infinity.

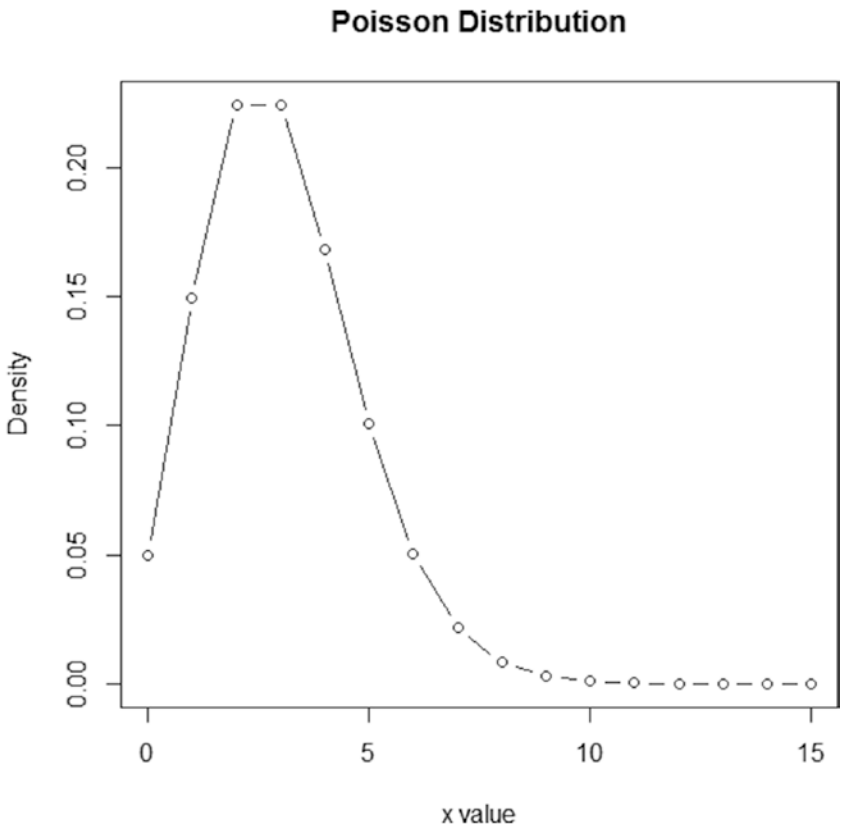

Figure 3-20. Poisson distribution with a lambda of 3

The list is by no means exhaustive, but those are the most commonly used distributions. Ideally if the data can be assumed to be normally distributed then that is the easiest way forward, however that isn't always the case and as such the relevant distribution needs to be identified.

Tests

Once you have progressed through each step of the EDA process and have learned more about the data, then you can move to deciding which tests, and descriptive statistics, would be appropriate. The key is not to jump straight to this step as valuable time and money may be wasted if the analysis is completed including a data entry error, or if the wrong analysis is performed due to not thinking about which data types are being used.

Continuous Data

If your continuous data approximately follows a normal distribution the types of tests used are called parametric tests. If the distribution is not normal and a transformation either doesn't help or is not appropriate then nonparametric tests may be more appropriate.

Generally speaking, nonparametric tests have less power than parametric tests, that is, there's a higher risk of missing a real effect by using nonparametric tests. This is due to the fact that nonparametric tests are distribution free and have to be conservative to account for this fact.

There are nonparametric equivalents for each parametric test and these will be shown alongside each other in the later chapters. The key difference between the tests, in addition to the power mentioned above, is that parametric tests use the means and nonparametric tests use the medians of the data.

In terms of satisfying the normality assumptions, data can actually be non-normal while still being applicable to parametric tests. There are some general rules of thumb to note: each case should be visually investigated to confirm whether appropriate to use parametric tests. As long as the sample size is above 15 for each group and the data is only slightly skewed, parametric tests can be used; or if the sample size is very large and the data clearly doesn't follow another distribution, parametric tests can be used.

Nonparametric tests are commonly used when there is a very small sample size due to the fact that the distribution won't be able to be identified. They also can handle statistical outliers and ranked data quite well. The downside of some nonparametric tests is that they assume that the groups have equal variation, which may not always be an appropriate assumption.

Discrete Data

To add slightly more confusion to the terms, if you had binary data that is clearly not continuous, then the type of tests you would use are parametric tests based on the binomial distribution. There are also parametric tests for Poisson and other distributions.

Basically parametric tests refer to traditional tests relevant to the data type, and nonparametric tests refer to tests used when the data violates the assumptions to be able to use the traditional tests. Most of the time parametric and nonparametric are associated with continuous data (and normality) only, but it's worth noting that these terms apply to all data types.

Thought needs to be given as to which type of tests are most appropriate to the data set you have collected, thinking about things such as data types, sample size, skewness, and variation. You also need to remember that all tests have assumptions, even nonparametric tests, and they need to be satisfied.

Summary

This chapter investigated the sections that make up exploratory data analysis (EDA), which should be performed before undertaking any type of statistical analysis. It was split into five sections corresponding to the five steps shown in the EDA process in Figure 3-1.

The first was identifying the data types for each variable in the data set as this will lead to the second section of plotting the data.

Plotting the data is dependent on which data types you have, as this dictates which plots can be drawn. This section showed a selection of commonly used plots and described the information to be gained from each. One main point in this section was that the plots drawn in EDA are for familiarity with the data only; Chapter 9 focuses on creating clear plots to deliver the main messages to the customer.

The third section discussed outliers in the data and that there are two types of outliers; data entry errors, which should be removed and statistical outliers, which should remain in the data during analysis.

The next section related to the family of the data, or the distribution. It listed the most common distributions used for continuous data and discrete data with an emphasis on checking for normality. It also highlighted that when checking for normality, plots such as the Q–Q plot should be used instead of formal tests.

The final section discussed the use of parametric and nonparametric tests and the benefits and downfalls of each of these along with the fact that these terms do not just apply to continuous data.

Chapter 4 looks at the descriptive statistics that can be gleaned from the data, again before testing is undertaken. It defines the difference between samples and populations, explains the different measures of shape, location, and spread; shows how to transform non-normal continuous data; highlights descriptive statistics for binary data; and also looks at what correlation can and can't tell you.