3 Data Exploration

Fail to prepare, prepare to fail.

—Roy Keane

In Chapter 2[21] we described the process of moving from a business problem to an analytics solution and, from there, to the design and construction of an **analytics base table** (**ABT**). An ABT for a predictive analytics solution contains a set of instances that are represented by a set of descriptive features and a target feature. Before attempting to build predictive models based on an ABT it is important that we undertake some exploratory analysis, or **data exploration**, of the data contained in the ABT. **Data exploration** is a key part of both the **Data Understanding** and, **Data Preparation** phases of CRISP-DM.

There are two goals in data exploration. The first goal is to fully understand the characteristics of the data in the ABT. It is important that for each feature in the ABT, we understand characteristics such as the types of values a feature can take, the ranges into which the values in a feature fall, and how the values in a dataset for a feature are distributed across the range that they can take. We refer to this as *getting to know* the data. The second goal of data exploration is to determine whether or not the data in an ABT suffer from any **data quality issues** that could adversely affect the models that we build. Examples of typical data quality issues include an instance that is missing values for one or more descriptive features, an instance that has an extremely high value for a feature, or an instance that has an inappropriate level for a feature. Some data quality issues arise due to invalid data and will be corrected as soon as we discover them. Others, however, arise because of perfectly valid data that may cause difficulty to some machine learning techniques. We note these types of data quality issues during exploration for potential handling when we reach the modeling phase of a project.

The most important tool used during data exploration is the **data quality report**. This chapter begins by describing the structure of a data quality report and explaining how it is used to *get to know* the data in an ABT and to identify data quality issues. We then describe a number of strategies for handling data quality issues and when it is appropriate to use them. Throughout the discussion of the data quality report and how we use it, we return to the motor insurance fraud case study from Chapter 2 [21]. Toward the end of the chapter, we introduce some more advanced data exploration techniques that, although not part of the standard data quality report, can be useful at this stage of an analytics project and present some data preparation techniques that can be applied to the data in an ABT prior to modeling.

3.1 The Data Quality Report

The **data quality report** is the most important tool of the data exploration process. A data quality report includes tabular reports (one for continuous features and one for categorical features) that describe the characteristics of each feature in an ABT using standard statistical measures of **central tendency** and **variation**. The tabular reports are accompanied by data visualizations that illustrate the distribution of the values in each feature in an ABT. Readers who are not already familiar with standard measures of central tendency (**mean**, **mode**, and **median**), standard measures of variation (**standard deviation** and **percentiles**), and standard data visualization plots (**bar plots**, **histograms**, and **box plots**) should read Appendix A[525] for the necessary introduction.

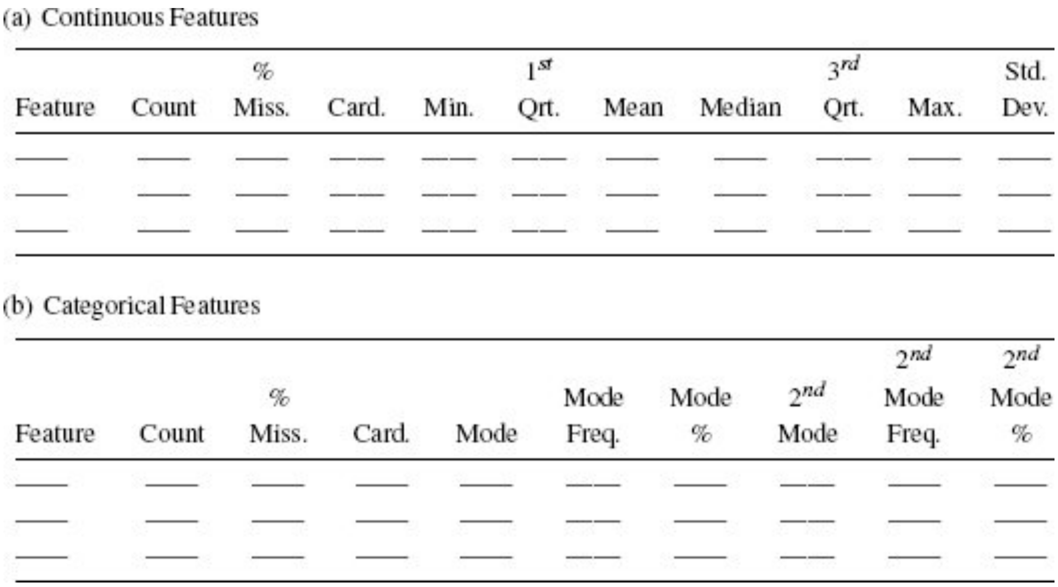
The table in a data quality report that describes continuous features should include a row containing the minimum, 1st quartile, mean, median, 3rd quartile, maximum, and standard deviation statistics for that feature as well as the total number of instances in the ABT, the percentage of instances in the ABT that are missing a value for each feature and the **cardinality** of each feature, (cardinality measures the number of distinct values present in the ABT for a feature). Table 3.1(a)[57] shows the structure of the table in a data quality report that describes continuous features.

The table in the data quality report that describes categorical features should include a row for each feature in the ABT that contains the two most frequent levels for the feature (the mode and 2nd mode) and the frequency with which these appear (both as raw frequencies and as a proportion of the total number of instances in the dataset). Each row should also include the percentage of instances in the ABT that are missing a value for the feature and the cardinality of the feature. Table 3.1(b)[57] shows the structure of the table in a data quality report that describes categorical features.

The data quality report should also include a histogram for each continuous feature in an ABT. For continuous features with cardinality less than 10, we use bar plots instead of histograms as this usually produces more informative data visualization. For each categorical feature in an ABT, a bar plot should be included in the data quality report.

Table 3.1

The structures of the tables included in a data quality report to describe (a) continuous features and (b) categorical features.



(a) Continuous Features

		%			1st			3rd		Std.
Feature	Count	Miss.	Card.	Min.	Qrt.	Mean	Median	Qrt.	Max.	Dev.
____	____	____	____	____	____	____	____	____	____	____
____	____	____	____	____	____	____	____	____	____	____
____	____	____	____	____	____	____	____	____	____	____

(b) Categorical Features

								2nd	2nd
		%			Mode	Mode	2nd	Mode	Mode
Feature	Count	Miss.	Card.	Mode	Freq.	%	Mode	Freq.	%
____	____	____	____	____	____	____	____	____	____
____	____	____	____	____	____	____	____	____	____
____	____	____	____	____	____	____	____	____	____

3.1.1 Case Study: Motor Insurance Fraud

Table 3.2[58] shows a portion of the ABT that has been developed for the motor insurance claims fraud detection solution based on the design described in Section 2.4.6[43].[1] The data quality report for this ABT is shown across Table 3.3[59] (tabular reports for continuous and categorical features) and Figure 3.1[60] (data visualizations for each feature in the dataset).

Table 3.2

Portions of the ABT for the motor insurance claims fraud detection problem discussed in Section 2.4.6 [43].

ID	Type	Inc.	Marital Status	Num. Clmnts.	Injury Type	Hospital Stay	Claim Amt.	Total Claimed	Num Claims	Num. Soft Tiss.	% Soft Tiss.	Claim Amt. Rcvd.	Fraud Flag
1	ci	0		2	soft tissue	no	1,625	3,250	2	2	1.0	0	1
2	ci	0		2	back	yes	15,028	60,112	1		0	15,028	0
3	ci	54,613	married	1	broken limb	no	-99,999	0	0	0	0	572	0
4	ci	0		4	broken limb	yes	5,097	11,661	1	1	1.0	7,864	0
5	ci	0		4	soft tissue	no	8,869	0	0	0	0	0	1
			⋮			⋮					⋮		
300	ci	0		2	broken limb	no	2,244	0	0	0	0	2,244	0
301	ci	0		1	broken limb	no	1,627	92,283	3	0	0	1,627	0
302	ci	0		3	serious	yes	270,200	0	0	0	0	270,200	0
303	ci	0		1	soft tissue	no	7,668	92,806	3	0	0	7,668	0
304	ci	46,365	married	1	back	no	3,217	0	0		0	1,653	0
			⋮			⋮					⋮		
458	ci	48,176	married	3	soft tissue	yes	4,653	8,203	1	0	0	4,653	0
459	ci	0		1	soft tissue	yes	881	51,245	3	0	0	0	1
460	ci	0		3	back	no	8,688	729,792	56	5	0.08	8,688	0
461	ci	47,371	divorced	1	broken limb	yes	3,194	11,668	1	0	0	3,194	0
462	ci	0		1	soft tissue	no	6,821	0	0	0	0	0	1
			⋮			⋮					⋮		
496	ci	0		1	soft tissue	no	2,118	0	0	0	0	0	1
497	ci	29,280	married	4	broken limb	yes	3,199	0	0	0	0	0	1
498	ci	0		1	broken limb	yes	32,469	0	0	0	0	16,763	0
499	ci	46,683	married	1	broken limb	no	179,448	0	0		0	179,448	0
500	ci	0		1	broken limb	no	8,259	0	0	0	0	0	1

Table 3.3

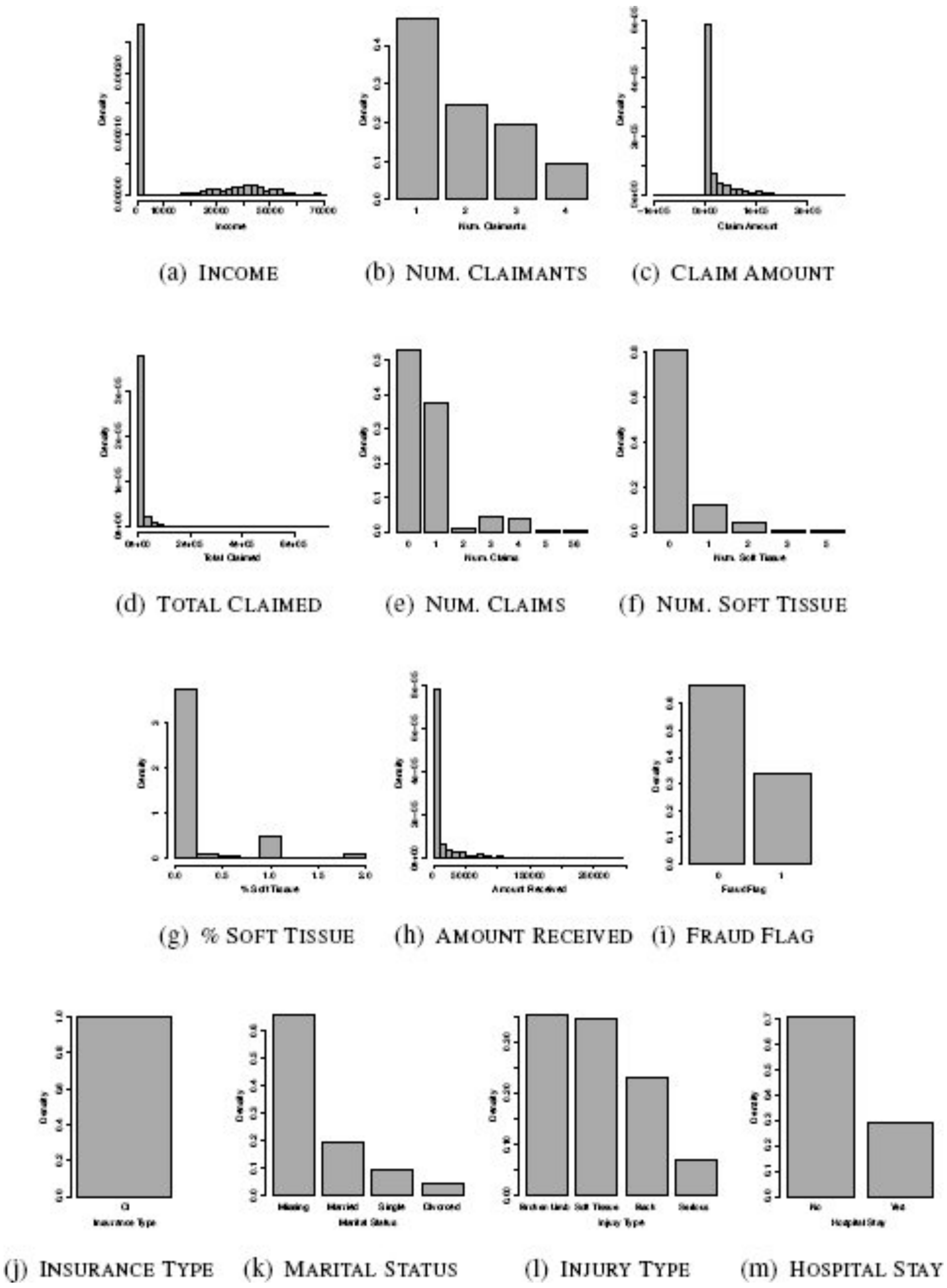
A data quality report for the motor insurance claims fraud detection ABT displayed in Table 3.2[58].

(a) Continuous Features

Feature	Count	% Miss.	Card.	Min	1st Qrt.	Mean	Median	3rd Qrt.	Max	Std. Dev.
Income	500	0.0	171	0.0	0.0	13,740.0	0.0	33,918.5	71,284.0	20,081.5
Num. Claimants	500	0.0	4	1.0	1.0	1.9	2	3.0	4.0	1.0
Claim Amount	500	0.0	493	-99,999	3,322.3	16,373.2	5,663.0	12,245.5	270,200.0	29,426.3
Total Claimed	500	0.0	235	0.0	0.0	9,597.2	0.0	11,282.8	729,792.0	35,655.7
Num. Claims	500	0.0	7	0.0	0.0	0.8	0.0	1.0	56.0	2.7
Num. Soft Tissue	500	2.0	6	0.0	0.0	0.2	0.0	0.0	5.0	0.6
% Soft Tissue	500	0.0	9	0.0	0.0	0.2	0.0	0.0	2.0	0.4
Amount Received	500	0.0	329	0.0	0.0	13,051.9	3,253.5	8,191.8	295,303.0	30,547.2
Fraud Flag	500	0.0	2	0.0	0.0	0.3	0.0	1.0	1.0	0.5

(b) Categorical Features

Feature	Count	% Miss.	Card.	Mode	Mode Freq.	Mode %	2nd Mode	2nd Mode Freq.	2nd Mode %
Insurance Type	500	0.0	1	ci	500	1.0	–	–	–
Marital Status	500	61.2	4	married	99	51.0	single	48	24.7
Injury Type	500	0.0	4	broken limb	177	35.4	soft tissue	172	34.4
Hospital Stay	500	0.0	2	no	354	70.8	yes	146	29.2



(a) INCOME (b) NUM. CLAIMANTS (c) CLAIM AMOUNT

(d) TOTAL CLAIMED (e) NUM. CLAIMS (f) NUM. SOFT TISSUE

(g) % SOFT TISSUE (h) AMOUNT RECEIVED (i) FRAUD FLAG

(j) INSURANCE TYPE (k) MARITAL STATUS (l) INJURY TYPE (m) HOSPITAL STAY

Figure 3.1

Visualizations of the continuous and categorical features in the motor insurance claims fraud detection ABT in Table 3.2[58].

3.2 Getting to Know the Data

The data quality report gives an in-depth picture of the data in an ABT, and we should study it in detail in order to *get to know* the data that we will work with. For each feature, we should examine the central tendency and variation to understand the types of values that each feature can take. For categorical features, we should first examine the mode, 2^{nd} mode, mode %, and 2^{nd} mode % in the categorical features table in the data quality report. These tell us the most common levels within these features and will identify if any levels dominate the dataset (these levels will have a very high mode %). The bar plots shown in the data quality report are also very useful here. They give us a quick overview of all the levels in the domain of each categorical feature and the frequencies of these levels.

For continuous features we should first examine the mean and standard deviation of each feature to get a sense of the central tendency and variation of the values within the dataset for the feature. We should also examine the minimum and maximum values to understand the range that is possible for each feature. The histograms for each continuous feature included in a data quality report are a very easy way for us to understand how the values for a feature are distributed across the range they can take.[2] When we generate histograms of features, there are a number of common, well-understood shapes that we should look out for. These shapes relate to well-known standard **probability distributions**,[3] and recognizing that the distribution of the values in an ABT for a feature closely matches one of these standard distributions can help us when building machine learning models. During data exploration we don’ t need to go any further than simply recognizing that features seem to follow particular distributions, and this can be done from examining the histogram for each feature. Figure 3.2[62] shows a selection of histogram shapes that exhibit characteristics commonly seen when analyzing features and that are indicative of standard, well-known probability distributions.

Figure 3.2(a)[62] shows a histogram exhibiting a **uniform distribution**. A uniform distribution indicates that a feature is equally likely to take a value in any of the ranges present. Sometimes a uniform distribution is indicative of a descriptive feature that contains an ID rather than a measure of something more interesting.

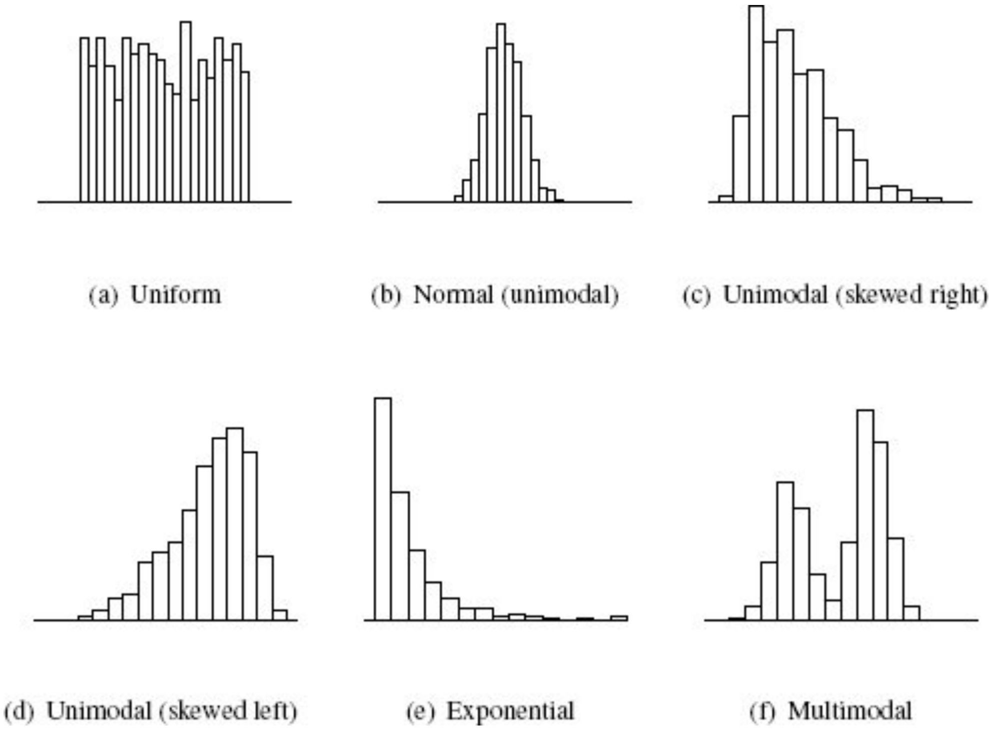


Figure 3.2

Histograms for six different sets of data, each of which exhibit well-known, common characteristics.

Figure 3.2(b)[62] shows a shape indicative of a **normal distribution**. Features following a normal distribution are characterized by a strong tendency toward a central value and symmetrical variation to either side of this central tendency. Naturally occurring phenomena—for example, the heights or weights of a randomly selected group of men or women—tend to follow a normal distribution. Histograms that follow a normal distribution can also be described as **unimodal** because they have a single peak around the central tendency. Finding features that exhibit a normal distribution is a good thing, as many of the modeling techniques we discuss in later chapters work particularly well with normally distributed data.

Figures 3.2(c)[62] and 3.2(d)[62] show unimodal histograms that exhibit **skew**. Skew is simply a tendency toward very high (**right skew** as seen in Figure 3.2(c)[62]) or very low (**left skew** as seen in Figure 3.2(d)[62]) values. Features recording salaries often follow a right skewed, distribution as most people are paid salaries near a well-defined central tendency, but there are usually a small number of people who are paid very large salaries. Skewed distributions are often said to have **long tails** toward these very high or very low values.

In a feature following an **exponential distribution**, as shown in Figure 3.2(e)[62], the likelihood of low values occurring is very high but diminishes rapidly for higher values. Features such as the number of times a person has made an insurance claim or the number of times a person has been married tend to follow an exponential distribution. Recognizing that a feature follows an exponential distribution is another clear warning sign that outliers are likely. As shown in Figure 3.2(e)[62], exponential distributions have a long tail, and so very high values are not uncommon.

Finally, a feature characterized by a **multimodal distribution** has two or more very commonly occurring ranges of values that are clearly separated. Figure 3.2(f)[62] shows a **bi-modal distribution** with two clear peaks—we can think of this as two normal distributions pushed together. Multimodal distributions tend to occur when a feature contains a measurement made across a number of distinct groups. For example, if we were to measure the heights of a randomly selected group of Irish men and women, we would expect a bi-modal distribution with a peak at around 1.635m for women and 1.775m for men.

Observing a multimodal distribution is cause for both caution and optimism. The caution comes from the fact that measures of central tendency and variation tend to break down for multimodal data. For example, consider that the mean value of the distribution shown in Figure 3.2(f)[62] is likely to sit right in the valley between the two peaks, even though very few instances actually have this value. The optimism associated with finding multimodally distributed data stems from the fact that, if we are lucky, the separate peaks in the distribution will be associated with the different target levels we are trying to predict. For example, if we were trying to predict gender from a set of physiological measurements, height would most likely be a very predictive value, as it would separate people into male and female groups.

This stage of data exploration is mostly an information-gathering exercise, the output of which is just a better understanding of the contents of an ABT. It does, however, also present a good opportunity to discuss anything unusual that we notice about the central tendency and variation of features within the ABT. For example, a salary feature with a mean of 40 would seem unlikely (40,000 would seem more reasonable) and should be investigated.

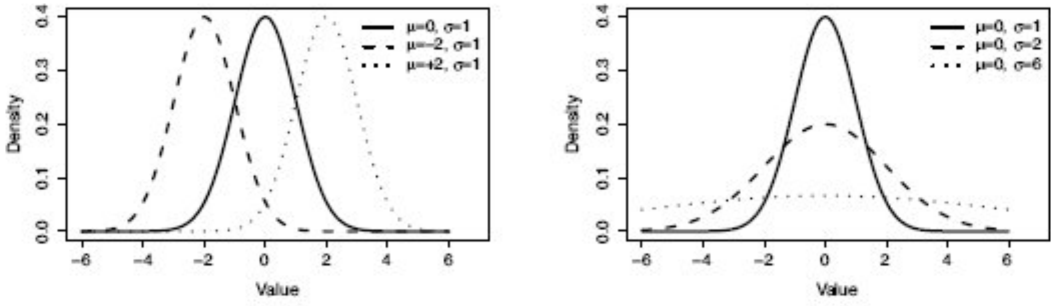
3.2.1 The Normal Distribution

The **normal distribution** (also known as a **Gaussian distribution**) is so important that it is worth spending a little extra time discussing its characteristics. Standard probability distributions have associated **probability density functions**, which define the characteristics of the distribution. The probability density function for the normal distribution is

$$N(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{3.1}$$

where x is any value, and μ and σ are parameters that define the shape of the distribution. Given a probability density function, we can plot the **density curve** associated with a distribution, which gives us a different way to visualize standard distributions like the normal. Figure 3.3[65] shows the density curves for a number of different normal distributions. The higher the curve for a particular value on the horizontal axis, the more likely that value is.

The curve defined by a normal probability distribution is symmetric around a single peak value. The location of the peak value is defined by the parameter μ (pronounced *mu*), which denotes the **population mean** (in other words, the mean value of the feature if we had access to every value that could possibly occur). The height and slope of the curve is dependent on the parameter σ (pronounced *sigma*), which denotes the **population standard deviation**. The larger the value of σ, the lower the maximum height of the curve and the shallower the slope. Figure 3.3(a)[65] illustrates how the location of the peak moves as the value for μ changes, and Figure 3.3(b)[65] illustrates how the shape of the curve changes as we vary the value for σ. Notice that in both figures, the normal distribution plotted with the continuous black line has mean $\mu = 0$ and standard deviation $\sigma = 1$. This normal distribution is known as the **standard normal distribution**. The notation *X is N(*μ*,* σ*)* is often used as a shorthand for *X is a normally distributed feature with mean* μ *and standard deviation* σ*.*[4] One important characteristic of the normal distribution is often described as the 68-95-99.7 **rule**. The rule states that approximately 68% of the values in a sample that follows a normal distribution will be within one σ of μ, 95% of the values will be within two σ of μ, and 99.7% of values will be within three σ of μ. Figure 3.4 [65] illustrates this rule. This rule highlights that in data that follows a normal distribution, there is a very low probability of observations occurring that differ from the mean by more than two standard deviations.



(a) Different means

(b) Different standard deviations

Figure 3.3

(a) Three normal distributions with different means but identical standard deviations; (b) three normal distributions with identical means but different standard deviations.

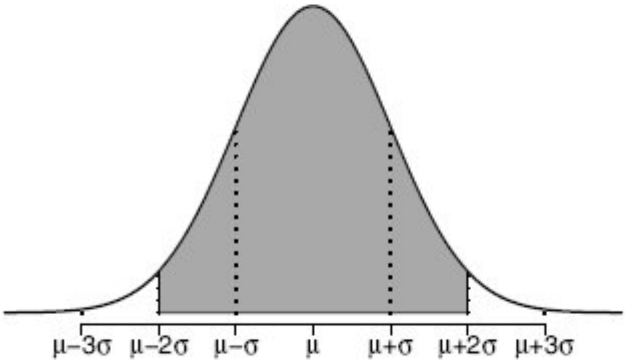


Figure 3.4

An illustration of the 68-95-99.7 rule. The gray region defines the area where 95% of values in a sample are expected.

3.2.2 Case Study: Motor Insurance Fraud

The data quality report in Table 3.3[59] and in Figure 3.1[60] and allows us to very quickly become familiar with the central tendency and variation of each feature in the ABT. These were all broadly as the business expected. In the bar plots in Figure 3.1[60], the different levels in the domain of each categorical feature, and how these levels are distributed, are obvious. For example, INJURY TYPE has four levels. Three of these, *broken limb*, *soft tissue*, and *back*, are quite frequent in the ABT, while *serious* is quite rare. The distribution of INSURANCE TYPE is a little strange, as it displays only one level.

From the histograms in Figure 3.1[60], we see that all the continuous features except for INCOME and FRAUD FLAG seem to follow an exponential distribution pretty closely. INCOME is interesting as it seems to follow what looks like a normal distribution except that there is one large bar at about 0. The distribution of the FRAUD FLAG feature that can be seen in its histogram is not typical of a continuous feature.

By analyzing the data quality report, we are able to understand the characteristics of the data in the ABT. We will return to the features that seemed to have slightly peculiar distributions.

3.3 Identifying Data Quality Issues

After getting to know the data, the second goal of data exploration is to identify any data quality issues in an ABT. A **data quality issue** is loosely defined as anything *unusual* about the data in an ABT. The most common data quality issues, however, are **missing values**, **irregular cardinality** problems, and **outliers**. In this section we describe each of these data quality issues and outline how the data quality report can be used to identify them.

The data quality issues we identify from a data quality report will be of two types: data quality issues due to **invalid data** and data quality issues due to **valid data**. Data quality issues due to invalid data typically arise because of errors in the process used to generate an ABT, usually in relation to calculating derived features. When we identify data quality issues due to invalid data, we should take immediate action to correct them, regenerate the ABT, and recreate the data quality report. Data quality issues due to valid data can arise for a range of domain-specific reasons (we discuss some of these later in this section), and we do not necessarily need to take any corrective action to address these issues. We do not correct data quality issues due to valid data unless the predictive models we will use the data in the ABT to train require that particular data quality issues be corrected. For example, we cannot train error-based models with data that contains missing values, and data that contains outliers significantly damages the performance of similarity-based models. At this stage we simply record any data quality issues due to valid data in a **data quality plan** so that we remain aware of them and can handle them later if required. Table 3.4[67] shows the structure of a data quality plan. For each of the data quality issues found, we include the feature it was found in and the details of the data quality issue. Later we add information on potential handling strategies for each data quality issue.

Table 3.4

The structure of a data quality plan.

Feature	Data Quality Issue	Potential Handling Strategies
——	——————————	————————————————————
——	——————————	————————————————————
——	——————————	————————————————————
——	——————————	————————————————————

3.3.1 Missing Values

Often when an ABT is generated, some instances will be missing values for one or more features. The **% Miss.** columns in the data quality report highlight the percentage of missing values for each feature (both continuous and categorical) in an ABT, and so it is very easy to identify which features suffer from this issue. If features have missing values, we must first determine why the values are missing. Often missing values arise from errors in data integration or in the process of generating values for derived fields. If this is the case, these missing values are due to invalid data, so the data integration errors can be corrected, and the ABT can be regenerated to populate the missing values. Missing values can also arise for legitimate reasons, however. Sometimes in an organization, certain values will only have been collected after a certain date, and the data used to generate an ABT might cover time both before and after this date. In other cases, particularly where data arises from manual entry, certain personally sensitive values (for example, salary, age, or weight) may be entered only for a small number of instances. These missing values are due to valid data, so they do not need to be handled but should instead be recorded in the data quality plan.

There is one case in which we might deal directly with missing values that arise from valid data during data exploration. If the proportion of missing values for a feature is very high, a good rule of thumb is anything in excess of 60%, then the amount of information stored in the feature is so low that it is probably a good idea to simply remove that feature from the ABT.

3.3.2 Irregular Cardinality

The **Card.** column in the data quality report shows the number of distinct values present for a feature within an ABT. A data quality issue arises when the cardinality for a feature does not match what we expect, a mismatch called an **irregular cardinality**. The first things to check the cardinality column for are features with a cardinality of 1. This indicates a feature that has the same value for every instance and contains no information useful for building predictive models. Features with a cardinality of 1 should first be investigated to ensure that the issue is not due to an ABT generation error. If this is the case, then the error should be corrected, and the ABT should be regenerated. If the generation process proves to be error-free, then features with a cardinality of 1, although valid, should be removed from an ABT because they will not be of any value in building predictive models.

The second things to check for in the cardinality column are categorical features incorrectly labeled as continuous. Continuous features will usually have a cardinality value close to the number of instances in the dataset. If the cardinality of a continuous feature is significantly less than the number of instances in the dataset, then it should be investigated. Sometimes a feature is actually continuous but in practice can assume only a small range of values—for example, the number of children a person has. In this case there is nothing wrong, and the feature should be left alone. In other cases, however, a categorical feature will have been developed to use numbers to indicate categories and might be mistakenly identified as a continuous feature in a data quality report. Checking for features with a low cardinality will highlight these features. For example, a feature might record gender using 1 for female and 0 for male. If treated as a continuous feature in a data quality report, this would have a cardinality of 2. Once identified, these features should be recoded as categorical features.

The third way in which a data quality issue can arise due to an irregular cardinality is if a categorical feature has a much higher cardinality than we would expect given the definition of the feature. For example, a categorical feature storing gender with a cardinality of 6 is worthy of further investigation. This issue often arises because multiple levels are used to represent the same thing—for example, in a feature storing gender, we might find levels of *male*, *female*, *m*, *f*, *M*, and *F*, which all represent male and female in slightly different ways. This is another example of a data quality issue due to invalid data. It should be corrected through a mapping to a standard set of levels, and the ABT should be regenerated.

The final example of a data quality issue due to an irregular cardinality is when a categorical feature simply has a very high number of levels—anything over 50 is worth investigation. There are many genuine examples of features that will have such high cardinality, but some of the machine learning algorithms that we will look at will struggle to effectively use features with such high cardinality. This is an example of a data issue due to valid data, so if this occurs for features in an ABT, it should be noted in the data quality plan.

3.3.3 Outliers

Outliers are values that lie far away from the central tendency of a feature. There are two kinds of outliers that might occur in an ABT: **invalid outliers** and **valid outliers**. Invalid outliers are values that have been included in a sample through error and are often referred to as noise in the data. Invalid outliers can arise for all sorts of different reasons. For example, during a manual data entry process, a *fat fingered*[5] analyst may have entered 100,000 instead of 1,000. Valid outliers are correct values that are simply very different from the rest of the values for a feature, for example, a billionaire who has a massive salary compared to everyone else in a sample.

There are two main ways that the data quality report can be used to identify outliers within a dataset. The first is to examine the minimum and maximum values for each feature and use domain knowledge to determine whether these are plausible values. For example, a minimum age value of −12 would jump out as an error. Outliers identified in this way are likely to be invalid outliers and should immediately be either corrected, if data sources allow this, or removed and marked as missing values if correction is not possible. In some cases we might even remove a complete instance from a dataset based on the presence of an outlier.

The second approach to identifying outliers is to compare the gaps between the median, minimum, maximum, 1^{st} quartile, and 3^{rd} quartile values. If the gap between the 3^{rd} quartile and the maximum value is noticeably larger than the gap between the median and the 3^{rd} quartile, this suggests that the maximum value is unusual and is likely to be an outlier. Similarly, if the gap between the 1^{st} quartile and the minimum value is noticeably larger than the gap between the median and the 1^{st} quartile, this suggests that the minimum value is unusual and is likely to be an outlier. The outliers shown in box plots also help to make this comparison. Exponential or skewed distributions in histograms are also good indicators of the presence of outliers.

It is likely that outliers found using the second approach are valid outliers, so they are a data quality issue due to valid data. Some machine learning techniques do not perform well in the presence of outliers, so we should note these in the data quality plan for possible handling later in the project.

3.3.4 Case Study: Motor Insurance Fraud

Using the data quality report in Table 3.3[59] and Figure 3.1[60] together with the ABT extract in Table 3.2[58], we can perform an analysis of this ABT for data quality issues. We do this by describing separately missing values, irregular cardinality, and outliers.

3.3.4.1 Missing Values

The **% Miss.** column of the data quality report in Table 3.3[59] shows that Marital Status and Num. Soft Tissue are the only features with an obvious problem with missing values. Indeed, over 60% of the values for Marital Status are missing, so this feature should almost certainly be removed from the ABT (we return to this feature shortly). Only 2% of the values for the Num. Soft Tissue feature are missing, so removal would be extreme in this case. This issue should be noted in the data quality plan.

An examination of the histogram for the Income feature (shown in Figure 3.1(a)[60]) and the actual data for this feature in Table 3.2[58] reveals an interesting pattern. In the histogram we can see an unusual number of zero values for Income that seems set apart from the central tendency of the data, which appears to be at about 40,000. Examining the Income row in the data quality report also shows a large difference between the mean and median values, which is unusual. Examining the actual raw data in Table 3.2[58] shows that these zeros always co-occur with missing values in the Marital Status feature. This pattern was investigated with the business to understand whether this was an issue due to valid or invalid data. It was confirmed by the business that the zeros in the Income feature actually represent missing values and that Marital Status and Income were collected together, leading to their both being missing for the same instances in the ABT. No other data source existed from which these features could be populated, so it was decided to remove both of them from the ABT.

3.3.4.2 Irregular Cardinality

Reading down the **Card.** column of the data quality report, we can see that the cardinality of the Insurance Type feature is 1, aa obvious data problem that needs investigation. The cardinality value indicates that every instance has the same value for this feature, *ci*. Investigation of this issue with the business revealed that nothing had gone wrong during the ABT generation process, and that *ci* refers to *car insurance*. Every instance in this ABT should have that value, and this feature was removed from the ABT.

Many of the continuous features in the dataset also have very low cardinality values. Num. Claimants, Num. Claims, Num. Soft Tissue, % Soft Tissue, and Fraud Flag all have cardinality less than 10, which is unusual in a dataset of 500 instances. These low cardinalities were investigated with the business. The low cardinality for the Num. Claimants, Num. Claims, and Num. Soft Tissue features was found to be valid, because these are categorical features and can only take values in a small range as people tend not to make very many claims. The % Soft Tissue feature is a ratio of the Num. Claims and Num. Soft Tissue features, and its low cardinality arises from their low cardinality.

The cardinality of 2 for the Fraud Flag feature highlights the fact that this is not really a continuous feature. Rather, Fraud Flag is a categorical feature that just happens to use *0* and *1* as its category labels which has led to its being treated as continuous in the ABT. Fraud Flag was changed to be a categorical feature. This is particularly important in this case because Fraud Flag is the target feature, and as we will see in upcoming chapters, the type of the target feature has a big impact on how we apply machine learning techniques.

3.3.4.3 Outliers

From an examination of the minimum and maximum values for each continuous feature in Table 3.3(a)[59], C LAIM AMOUNT jumps out as having an unusual minimum value of −99,999. A little investigation revealed that this minimum value arises from $\mathbf{d}_3$ in Table 3.2[58]. The absence of a large bar at −99,999 in Figure 3.1(c)[60] confirms that there are not multiple occurrences of this value. The pattern 99,999 also suggests that this is most likely a data entry error or a system default remaining in the ABT. This was confirmed with the business in this case, and this value was treated as a invalid outlier and replaced with a missing value.

Table 3.5

The data quality plan for the motor insurance fraud prediction ABT.

Feature	Data Quality Issue	Potential Handling Strategies
NUM. SOFT TISSUE	Missing values (2%)	
CLAIM AMOUNT	Outliers (high)	
AMOUNT RECEIVED	Outliers (high)	

CLAIM AMOUNT, TOTAL CLAIMED, NUM. CLAIMS and AMOUNT RECEIVED all seem to have unusually high maximum values, especially when compared to their median and 3rd quartile values. To investigate outliers, we should always start by locating the instance in the dataset that contains the strange maximum or minimum values. In this case the maximum values for TOTAL CLAIMED and NUM. CLAIMS both come from $\mathbf{d}_{460}$ in Table 3.2[58]. This policy holder seems to have made many more claims than anyone else, and the total amount claimed reflects this. This deviation from the norm was investigated with the business, and it turned out that although these figures were correct, this policy was actually a company policy rather than an individual policy, which was included in the ABT by mistake. For this reason, instance $\mathbf{d}_{460}$ was removed from the ABT.

The offending large maximums for CLAIM AMOUNT and AMOUNT RECEIVED both come from $\mathbf{d}_{302}$ in Table 3.2[58]. Investigation of this claim with the business revealed that this is in fact a valid outlier and represents an unusually large claim for a very serious injury. Examination of the histograms in Figures 3.1(c)[60] and 3.1(h)[60] show that the CLAIM AMOUNT and AMOUNT RECEIVED features have a number of large values (evidenced by the small bars to the right hand side of these histograms) and that $\mathbf{d}_{302}$ is not unique. These outliers should be noted in the data quality plan for possible handling later in the project.

3.3.4.4 The Data Quality Plan

Based on the analysis described in the preceding sections, the data quality plan shown in Table 3.5[72] was created. This records each of the data quality issues due to valid data that have been identified in the motor insurance fraud ABT. During the Modeling phase of the project, we will use this table as a reminder of data quality issues that could affect model training. At the end of the next section we complete this table by adding potential handling strategies.

3.4 Handling Data Quality Issues

When we find data quality issues due to valid data during data exploration, we should note these issues in a data quality plan for potential handling later in the project. The most common issues in this regard are missing values and outliers, which are both examples of **noise** in the data. Although we usually delay handling noise issues until the modeling phase of a project (different predictive model types require different levels of noise handling, and we should in general do as little noise handling as we can), in this section we describe the most common techniques used to handle missing values and outliers. It is a good idea to add suggestions for the best technique to handle each data quality issue in the data quality plan during data exploration as it will save time during modeling.

3.4.1 Handling Missing Values

The simplest approach to handling missing values is to simply drop from an ABT any features that have them. This, however, can result in massive, and frequently needless, loss of data. For example, if in an ABT containing 1,000 instances, one value is missing for a particular feature, it would be pretty extreme to remove that whole feature. As a general rule of thumb, only features that are missing in excess of 60% of their values should be considered for complete removal, and more subtle handling techniques should be used for features missing less data.

An alternative to entirely deleting features that suffer from large numbers of missing values is to a derive a missing indicator feature from them. This is a binary feature that flags whether the value was present or missing in the original feature. This can be useful if the reason that specific values for a feature are missing might have some relationship to the target feature—for example, if a feature that has missing values represented sensitive personal data, people' s readiness to provide this data (or not) might tell us something about them. When missing indicator features are used, the original feature is usually discarded.

Another simple approach to handling missing values is **complete case analysis**, which deletes from an ABT any instances that are missing one or more feature values. This approach, however, can result in significant amounts of data loss and can introduce a bias into the dataset if the distribution of missing values in the dataset is not completely random. In general, we recommend the use of complete case analysis only to remove instances that are missing the value of the target feature. Indeed, any instances with a missing value for the target feature should always be removed from an ABT.

Imputation replaces missing feature values with a plausible estimated value based on the feature values that are present. The most common approach to imputation is to replace missing values for a feature with a measure of the central tendency of that feature. For continuous features, the mean or median are most commonly used, and for categorical features, the mode is most commonly used.

Imputation, however, should not be used for features that have very large numbers of missing values because imputing a very large number of missing values will change the central tendency of a feature too much. We would be reluctant to use imputation on features missing in excess of 30% of their values and would strongly recommend against the use of imputation on features missing in excess of 50% of their values.

There are other, more complex approaches to imputation. For example, we can actually build a predictive model that estimates a replacement for a missing value based on the feature values that are present in a dataset for a given instance. We recommend, however, using simple approaches first and turning to more complex ones only if required.

Imputation techniques tend to give good results and avoid the data loss associated with deleting features or complete case analysis. It is important to note, however, that all imputation techniques suffer from the fact that they change the underlying data in an ABT and can cause the variation within a feature to be underestimated, which can negatively bias the relationships between a descriptive feature and a target feature.

3.4.2 Handling Outliers

The easiest way to handle outliers is to use a **clamp transformation**. This clamps all values above an upper threshold and below a lower threshold to these threshold values, thus removing the offending outliers:

$$a_i = \begin{cases} lower & \text{if } a_i < lower \\ upper & \text{if } a_i > upper \\ a_i & otherwise \end{cases} \tag{3.2}$$

where a_i is a specific value of feature *a*, and *lower* and *upper* are the lower and upper thresholds.

The upper and lower thresholds can be set manually based on domain knowledge or can be calculated from data. One common way to calculate clamp thresholds is to set the lower threshold to the 1^{st} quartile value minus 1.5 times the **inter-quartile range** and the upper threshold to the 3^{rd} quartile plus 1.5 times the inter-quartile range. This works effectively and takes into account the fact that the variation in a dataset can be different to either side of a central tendency.

If this approach were to be used for the CLAIM AMOUNT feature from the motor claims insurance fraud detection scenario, then the upper and lower thresholds would be defined as follows:

$$lower = 3{,}322.3 - 1.5 \times 8{,}923.2 = -10{,}062.5$$

$$upper = 12{,}245.5 + 1.5 \times 8{,}923.2 = 25{,}630.3$$

where the values used are extracted from Table 3.3[59]. Any values outside these thresholds would be converted to the threshold values. Examining the histogram in Figure 3.1(c)[60] is useful in considering the impact of applying the clamp transformation using these thresholds. Locating 25,630.3

on the horizontal axis shows that this upper threshold would cause a relatively large number of values to be changed. The impact of the clamp transformation can be reduced by changing the multiplier used to calculate the thresholds from 1.5 to a larger value.

Another commonly used approach to setting the upper and lower thresholds is to use the mean value of a feature plus or minus 2 times the standard deviation.[6] Again this works well, but it does assume that the underlying data follows a normal distribution.

If this approach were to be used for the AMOUNT RECEIVED feature from the motor claims insurance fraud detection scenario, then the upper and lower thresholds would be defined as follows:

$$lower = 13{,}051.9 - 2 \times 30{,}547.2 = -48{,}042.5$$

$$upper = 13{,}051.9 + 2 \times 30{,}547.2 = 74{,}146.3$$

where the values used are again extracted from Table 3.3[59]. Examining the histogram in Figure 3.1(h)[60] is again a good indication of the impact of using this transformation. This impact can be reduced by changing the multiplier used to calculate the thresholds from 2 to a larger value.

Opinions vary widely on when transformations like the clamp transformation should be used to handle outliers in data. Many argue that performing this type of transformation may remove the most interesting and, from a predictive modeling point of view, informative instances from a dataset. On the other hand, some of the machine learning techniques that we discuss in upcoming chapters perform poorly in the presence of outliers. We recommend only applying the clamp transformation in cases where it is suspected that a model is performing poorly due to the presence of outliers. The impact of the clamp transformation should then be evaluated by comparing the performance of different models trained on datasets where the transformation has been applied and where it has not.

3.4.3 Case Study: Motor Insurance Fraud

If we needed to do it, the most sensible approach to handling the missing values in the NUM. SOFT TISSUE feature would be to use imputation. There are very few missing values for this feature (2%), so replacing them with an imputed value should not excessively affect the variance of the feature. In this case the median value of 0.0 (shown in Table 3.3(a)[59]) is the most appropriate value to use to replace the missing values; because this feature only actually takes discrete values, the mean value of 0.2 never naturally occurs in the dataset.

The outliers present in the CLAIM AMOUNT and AMOUNT RECEIVED features could be easily handled using a clamp transformation. Both features follow a broadly exponential distribution, however, which means that the methods described for setting the thresholds of the clamp will not work especially well (both methods work best for normally distributed data). Therefore, manually setting upper and lower thresholds based on domain knowledge is most appropriate in this case. The business advised that for both features, a lower threshold of 0 and an upper threshold of 80,000 would make sense.

We completed the data quality plan by including these potential handling strategies. The final data quality plan is shown in Table 3.6[77]. Together with the data quality report, these are the outputs of the data exploration work for the motor insurance fraud detection project.

Table 3.6

The data quality plan with potential handling strategies for the motor insurance fraud prediction ABT.

Feature	Data Quality Issue	Potential Handling Strategies
NUM. SOFT TISSUE	Missing values (2%)	Imputation (median: 0.0)
CLAIM AMOUNT	Outliers (high)	Clamp transformation (manual: 0, 80,000)
AMOUNT RECEIVED	Outliers (high)	Clamp transformation (manual: 0, 80,000)

3.5 Advanced Data Exploration

All the descriptive statistics and data visualization techniques that we have used in the previous sections of this chapter have focused on the characteristics of individual features. This section will introduce techniques that enable us to examine relationships between pairs of features.

3.5.1 Visualizing Relationships Between Features

In preparing to create predictive models, it is always a good idea to investigate the relationships between pairs of features. This can help indicate which descriptive features might be useful for predicting a target feature and help find pairs of descriptive features that are closely