Chapter 8
Data Preparation

Torture the data, and it will confess to anything.
Ronald Coase

Data preparation includes data collection and integration, visual data exploration, and data preprocessing by imputing missing data, handling outliers, data transformation, and data balancing. Some of these actions are not trivial and require domain knowledge. Very often they need several iterations until data of acceptable quality is achieved.

Data quality is a decisive factor for the success of data analysis and model building. That makes data preparation the key productivity element in the Data Science workflow. If it is performed inappropriately, it can significantly reduce model development efficiency. In fact, data preparation solves the following dilemma: models or Garbage-In-Garbage-Out. Either we have low-quality data that is useless for solving the defined business problem and can be qualitatively characterized as garbage, or, in the case of data of acceptable quality, we can develop a variety of modeling solutions.

With its high probability of bad results and nasty scenarios, data preparation can be defined as a kingdom of Murphy’s laws. A list of the effects of Murphy’s law on data that presents these scenarios is given in Sect. 17.6.

The first objective of this chapter is to clarify the issues of data collection, such as identifying internal and external data sources, defining metadata, integrating the data, and performing a data sufficiency check. The second is to focus on several techniques for visual data exploration, such as identifying strange patterns, analyzing data distributions, and extracting insight from univariate, bivariate, and multivariate plots. The third is to discuss key data preprocessing capabilities such as handling missing data and outliers, important data transformations, and balancing the data.
8.1 Data Collection

Data is collected in many different ways. The lifecycle of usable data usually involves capture, preprocessing, storage, retrieval, postprocessing, analysis, visualization, modeling, and so on.

Once captured, data is usually referred to as being structured, semi-structured, or unstructured. These distinctions are important because they are directly related to the type of database technologies and storage required, the software and methods by which the data is queried and processed, and the complexity of dealing with the data.

Structured data refers to data that is stored as a structure in a relational database or spreadsheet. Often it is easily searchable using SQL (structured query language), since the structure of the data is known. A record of manufacturing process variables is a good example. Each process variable has a data collection time, variable name, description, units, frequency of data collection, etc.

Unstructured data is data that is not defined by any schema, model, or structure, and is not organized in a specific way. A typical case is most social media data.

Semistructured data is a combination of the two. It is basically unstructured data that also has structured data appended to it. For example, smartphone photos in addition to the picture (unstructured data) have the date and time the photo was taken, the image size, etc. That is the structured part.

A popular approach to data collection is ETL (Extract, Transform, and Load). These three conceptual steps are how most data pipelines are designed and structured:

- Extract. This is the step where raw data waits for upstream data sources to land (e.g., an upstream source could be a machine- or user-generated log, a copy of a relational database, an external data set, etc.).
- Transform. This is the most important ETL job, where business rules and actions such as filtering, grouping, and aggregation to translate raw data into analysis-ready data sets are applied. This step requires a good understanding of the business and domain knowledge acquisition.
- Load. Finally, the processed data is loaded and transported to its final destination. Often, this data set can either be consumed directly by the final users or be treated as yet another upstream dependency on another ETL job, forming a so-called data lineage.

A short summary of other methods for data collection, such as Hadoop, is given in Chap. 12.

8.1.1 Data Sources

Most of the data sources are within the organization related to the defined business problem. They can be spread across different departments and geographies,
however, and be based on various data collection platforms. Usually the domain SMEs have sufficient knowledge about the availability of data sources and the possible ways to access them.

Recently, there has been a growing trend toward using external data for improved data analysis and model building. Typical cases are macroeconomic and microeconomic data and data from social networks. A summary of both internal and external data sources is given below.

**Internal Data Sources**
In most cases the origin of the targeted problem-related data is inside the business. Examples are process-monitoring historians in manufacturing, and data warehouses in businesses. It is common practice that the internal data is extracted during model development by a person from the business who is familiar with the specific sources. Usually the final tables containing the targeted variables are delivered as Excel spreadsheets. However, these are aggregated data collected by scripts from the transaction tables in the corresponding corporate relational databases.

**External Data Sources**
One of the advantages of the Internet is the availability of many services offering almost any data of interest for building business-related models. Not all of these services, however, are equal in quality and in the number of available economic indicators. Some of the best-known sources are discussed below.

**IHS Markit**
The most well-known external source with almost universal collection of economic and financial data is IHS Markit (http://www.ihsmarkit.com.) It includes the main categories of global economic data, global financial data, and industry and sector data for over 200 countries and more than 170 industries. Some of this data includes forecasts for a defined forecast horizon. Access to the data needs a subscription.

**FRED**
The Federal Reserve Bank of St. Louis Economic Data (FRED) service offers more than 500,000 sets of american and international economic historical data for free. No forecasts are offered, however.

**Indexes**
Special types of external data is economic indexes. In the same way as the most famous stock market indexes, such as the Dow Jones and S&P 500, represent the state of the stock market in the United States, economic indexes capture the state of a national economy. Among the numerous available economic indexes, we will focus our attention on the Chicago Fed National Activity Index (CFNAI), which is the de facto normalized economic index for the state of the US economy. One of the advantages of this index is that it can be used as a leading economic indicator of a recession.

**The Chicago Fed National Activity Index**
The CFNAI is a monthly index based on a weighted average of 85 monthly indicators of US national economic activity. The 85 economic indicators that are
included in the CFNAI are taken from four broad categories of economic sectors: (1) production and income; (2) employment, unemployment, and hours; (3) personal consumption and housing; and (4) sales, orders, and inventories. It is a normalized index, i.e., it has an average value of zero and a standard deviation of one. Since economic activity tends toward trend growth rate over time, a positive index reading corresponds to growth above trend and a negative index reading corresponds to growth below trend.

An example of how the CFNAI relates to the recessions recently seen in the US economy is shown in Fig. 8.1. Shading indicates official periods of recession, as identified by the National Bureau of Economic Research. Following a period of economic expansion, an increasing likelihood of a recession has historically been associated with a CFNAI-MA3 (“MA3” means a moving average of the values from the last 3 months) value below −0.70. Conversely, following a period of economic contraction, an increasing likelihood of an expansion has historically been associated with a CFNAI-MA3 value above −0.70 and a significant likelihood of an expansion has historically been associated with a CFNAI-MA3 value above +0.20.

As shown by the first arrow in Fig. 8.1, the CFNAI rang the bell about the upcoming great recession in January–February 2008 several months before many businesses were heavily hit. It rang the bell for the upcoming recession due to coronavirus 19 in February 2020, as shown by the second arrow.

**Metadata Definition**

The key procedure in data definition is clarifying the structure and content of the data. This data about the data is called metadata and provides generic information about key aspects of the data, such as:

- time and date of creation;
- data source;
- data description;
- purpose of the data.
Above all, metadata is data. It is of critical importance for understanding the context of the data. An example of metadata for the emissions estimation project is shown in Table 8.1. This includes the key structural information about the variables used from the process historian, such as the variable name and description, the physical units, the tag in the system, the frequency of data collection, and type (is it only measured or can it be manipulated in control loops?)

On modeling sites, the metadata includes the types of variables available in the selected modeling tools. The most used data types are as follows:

- Numeric or interval type. Numeric variables.
- Categorical or nominal type. Character variables.
- Ordinal type. Categorical variables with clearly ordered categories.
- Binary type. Categorical variables with only two distinct classes.

### Data Integration

In many real-world problems, integrating and synchronizing the different pieces of internal and external data is a challenging task. Two of the most important techniques for data integration are alignment of business-related data with the corresponding business structure and alignment with time in the case of time series data.

#### Alignment with Business Structure

The first question, after identifying the potential economic drivers, is their alignment with a specific business structure or economic category. This allows one to identify the specific data source on the Internet. In the case of the United States, Canada, and Mexico one can use an industry classification system called the North American Industry Classification System (NAICS).  

---

1. [http://www.census.gov/eos/www/naics/](http://www.census.gov/eos/www/naics/)
NAICS is a six-digit hierarchical numeric classification system, with the first digit or two designating the broad industry sector and subsequent digits each reflecting more specific categories. For example, the first sector with the lowest first digit is agriculture, followed by mining, and then construction; manufacturing begins with the digit 3; wholesale trade, retail trade, and transportation are industry sectors in the middle of the list; codes beginning with 6, 7, or 8 are for services, and those beginning with 9 are for public administration. An example of the NAICS code hierarchy in the case of the food manufacturing sector is shown in Table 8.2.

An example of the alignment of some identified economic drivers taken from a business forecasting project is given in Table 8.3.

An example of how to link a NAICS code to available historical data is shown in Fig. 8.2 for the case of shampoo production in the United States. Usually the economic indicator is the producer price index (PPI) of the industry sector where the driver of interest belongs. In the case of shampoo, the available PPI is for toilet preparation manufacturing, part of which is shampoo manufacturing. The starting time of the index is July 2007, when the index is 100%. The dynamics of the index until April 2018 is shown in Fig. 8.2.

### Table 8.2 Food manufacturing NAICS code hierarchy

| NAICS  | Description                                                      |
|--------|------------------------------------------------------------------|
| 31–33  | Manufacturing                                                    |
| 31     | Food manufacturing                                               |
| 311    | Fruit and vegetable preserving and specialty food manufacturing   |
| 31,141 | Frozen food manufacturing                                        |
| 311,411| Frozen fruit, juice, and vegetable manufacturing                 |
| 311,412| Frozen specialty food manufacturing                               |
| 31,142 | Fruit and vegetable canning, pickling, and drying                |
| 311,421| Fruit and vegetable canning                                      |
| 311,422| Specialty canning                                                |
| 311,423| Dried and dehydrated food manufacturing                          |

### Table 8.3 NAICS codes of some identified economic drivers for business forecasting

| Economic driver | NAICS      | Description                                                                 |
|-----------------|------------|-----------------------------------------------------------------------------|
| Trash bags      | 326,111    | Trash bags, plastics film, single wall or multiwall, manufacturing          |
| Liquid laundry  | 333,312    | Commercial laundry, drycleaning, and pressing machine manufacturing          |
| Shampoo         | 81,211     | Hair, nail, and skin care services                                           |
| Bricks          | 327,331    | Bricks, concrete, manufacturing                                              |
| Cheese          | 311,511    | Cheese, cottage, manufacturing                                               |
| Napkins         | 322,121    | Napkins, table, made in paper mills                                          |

Alignment with Time

Alignment of the internal and external data with a common time stamp is critical for the analysis of time series. The first obvious step in integrating the data from the
different sources is to align it relative to the specific time window and frequency of the target (dependent) variable. This process requires expanding or contracting some of the data.

Often, not all of the data is for the same interval. Some might be annual, other data quarterly, and yet other data monthly. Thus, it all has to be brought into a form with the same interval. Sometimes, the time series data has to be converted from a finer interval to a coarser interval (monthly to quarterly). This is called contracting. Other times, an annual interval has to be converted to a quarterly interval. This is called expanding.

Converting time series data that is at a finer interval, such as monthly or weekly, can be simply a matter of summing or averaging data to get a coarser interval. Time series data such as demand or sales should be summed. Data such as indices should be averaged.

Sometimes, data for the independent variables is available only at a more aggregated level than the target variable. This means that the data has to be converted to a more frequent interval, such as going from annual to quarterly. Knowledge of the nature of the time series being expanded from one interval to another is definitely useful. The simplest approach is to divide by 4 (four quarters in a year), but this is generally not too informative. Expanding using some form of spline interpolation
might be more useful. This approach would accommodate a trend from 1 year to the
next. Combining the spline interpolation approach with seasonal adjustments when
the independent variable is known to follow seasonal patterns similar to other
independent variables might provide the best of both worlds.\textsuperscript{2}

**Data Sufficiency Check**

Before beginning visual inspection and shaping the data into a form more suitable
for extracting insight and modeling, the question “do we have the right data to solve
the business problem?” needs an answer. Just because we have a lot of data, it
doesn’t mean that the right data has been collected. We need to ensure that the data is
representative of the entire domain of interest—that the observations cover the range
of values anticipated when the developed model is used in production. The perils of
extrapolation have to be considered. Ideally, training data should include values with
ranges broad enough to cover the expected changes when the model is deployed.

### 8.2 Visual Data Exploration

One of the first questions before beginning data preprocessing is: is the collected data
visually acceptable? The answer is subjective but a short visual inspection of key
process variables may give important information about their usefulness. As a result,
some variables can be removed before preprocessing. A combination of data pattern
identification, distribution analysis, and gaining insight from univariate, bivariate,
and multivariate plots can help in the visual assessment of data quality. Several
examples from manufacturing process data are given below.

#### 8.2.1 Strange Data Patterns

The first task in visual data exploration is to look for patterns that reduce data quality,
such as large deviations, constant values, or a high noise level. The degree of quality
may decline in different ways, as is shown for three selected process variables in
Fig. 8.3. This figure includes three plots of hourly averaged data from the same
manufacturing process.

The first variable (Var1) has a curve with a relatively low noise level. The only
exception is the big drop shown by an arrow. This significant deviation from the
current trend was probably caused by a short process upset and will be treated as an
outlier, which will probably be removed. This single outlier will not have a signif-
icant negative impact on the rest of the data, which can be reliably used in data
preprocessing, analysis, and model development.

\textsuperscript{2}T. Rey, A. Kordon, and C. Wells, *Applied Data Mining in Forecasting*, SAS Institute, 2012.
A more dangerous pattern is identified by an arrow in the Var2 plot. For almost 180 hours, the recorded value is a constant number. The hypothesis that the sensor was broken during this period of time was confirmed later by the SMEs, i.e., this part of the data cannot be used for analysis. In addition, for some reason, the new sensor shows large fluctuations until sample 700, which also cannot be used in the analysis. One option is to focus on the data after sample 700 and not use almost 25% of the

Fig. 8.3 Different patterns in the data for process variables
records. Even in this case, a pattern of changing noise in Var2 is observed (reduced noise around 1100 and 2100 rows and increased variations around 2400 rows in Fig. 8.3). As a result, the quality of the Var2 data is visually estimated as questionable, i.e., it could be used in data analysis and model building but only under selected conditions.

The Var3 plot shows the worst-case scenario, when the process variable is zero 45% of the time. Unfortunately, these periods happened repeatedly and influence the whole period of collected data. The quality of the Var3 data is visually estimated as poor, i.e., it is recommended that it should be removed from further data analysis.

### 8.2.2 Data Distributions

Exploring data distributions is another valuable source of preliminary insight about data quality. The first important piece of information is the type of distribution. Examples of different types of distribution are shown in Fig. 8.4.
The types of the distributions for all cases in Fig. 8.4 were identified by the modeling tool used. Var7 is representative of the most common situation, of normally distributed data. Var16 is a case of a dual normal distribution, which in manufacturing indicates two operating regimes. Many process variables have a tailed distribution type, as is the case with Var10. Often these variables are transformed to make the distribution closer to normal. Sometimes it is difficult to fit a distribution type to the distribution pattern of the data, as in the example of Var13. This is another indicator of potential issues in analyzing the data.

The other benefit of exploring data distributions is identifying outliers. This will be discussed in Sect. 8.3.2.

8.2.3 Univariate Plots

Another way of extracting insight from collected data is by comparing univariate plots of the targeted and independent variables. Of special value is when the comparison is done with process variables sorted by the dependent variable. An example of such univariate plot, related to the emissions estimation project, with eight independent variables \( x_1 - x_8 \) sorted by the target variable \( y \), the emissions, is shown in Fig. 8.5.

A comparison between the trends of some independent variables, such as \( x_1 \) and \( x_7 \), and the trend of the emissions shows a lack of similarity. This is an indication of potential low correlation and a high probability of rejection of those variables. The opposite observation can be made, of a high similarity between the trends, for \( x_6 \) and the emissions. Possible links between \( x_3 \) and \( x_4 \) and the emissions can be defined as well. All three could be good candidates for potential selection of independent variables.

8.2.4 Bivariate Plots

Bivariate plots give important information about potential relationships between variables of interest. They also define another set of outliers, identified by patterns outside the norm. The norm is usually the 95% confidence ellipse. An example of a bivariate plot of three independent variables from the emissions estimation project is shown in Fig. 8.6.

The insight from the plot can be summarized as follows. (1) Variables \( x_6 \) and \( x_8 \) are correlated to some extent but not highly, i.e., multicollinearity is not expected. (2) \( x_7 \) is not correlated with either \( x_6 \) or \( x_8 \); and (3) there are several outliers outside the 95% confidence ellipse that need to be handled.

\(^3\)SAS JMP 14.
Bivariate plots are not practical for data sets with a large number of variables, however. One potential option for visual exploration in this case is multivariate plots.

### 8.2.5 Multivariate Plots

As we’ll discuss in Chap. 9, principal component analysis (PCA) is the most used method for multivariate exploration. The key result is a reduction of the dimensionality from the high number of original variables (dozens to thousands) to a low number of linear combinations of them, called principal components (usually less than 10). The most used plot is called the score plot, which is a scatter plot of the first two principal components. Examples of such plots from two different manufacturing processes are shown in Figs. 8.7 and 8.8.

The score plot, shown in Fig. 8.7, is based on the first two principal components of 89 original process variables. The first principal component explains 45.7% of the process variability, while the second principal component captures 14.4% of the process variability. This is a 2D view of the state of the process, based on 60% variability of the 89 original process variables, where each dot represents the state of the process at a corresponding period of time. Figure 8.7 shows that the manufacturing process operated most of the time in steady conditions with small variations.
This is a clear example of process conditions with one operating regime. An advantage of this mode is that the insight and model development will be defined by this operating condition. The deviations from this mode, shown by black dots in Fig. 8.7, can be treated as outliers and their fate will be decided by discussions with domain SMEs.

In the case shown in Fig. 8.8, the number of original process variables is 186 and the first two principal components represent 59.8% of the process variability. The two operating regimes defined, represented by two clusters in the score plot, may require different modeling solutions for each operating regime or a robust model that satisfies the performance criteria for both situations. This is a more challenging case for developing and applying data-driven solutions than the single operating regime shown in Fig. 8.7.
Fig. 8.7  Multivariate plot for a process with one operating regime

Fig. 8.8  Multivariate plot for a process with two operating regimes
8.3 Data Preprocessing

Data preprocessing is not trivial and is more an art than a science. It includes several steps that help to increase the informational content of the collected raw data. It is an iterative process, with several iterations until acceptable data quality is achieved. These steps are shown in Fig. 8.9 and discussed in this section.

8.3.1 Handling Missing Data

A missing value is a data point that has not been stored or gathered due to a faulty sensor or sampling process, cost restrictions, or limitations in the acquisition process. The treatment of missing values is difficult and has significant consequences in data analysis and modeling. Inappropriately handling the missing values can easily lead to poor knowledge being extracted, as well as wrong conclusions.

The main dilemma in handling missing data is: removal or imputation? The answer depends on the percentage of missing values in the data set, the variables affected by missing values, whether that missing data is a part of the dependent or independent variables, etc. Treatment of missing data becomes important, since the
insights from the data or the performance of developed predictive model could be impacted if the missing values are not appropriately handled.

**Removal**
The first option is usually to discard those records that may contain a missing value. However, this approach is rarely beneficial, as eliminating records may introduce bias into the model development process, and important information can be discarded. There are two ways to implement the removal: (1) listwise and (2) pairwise.

1. Listwise. In this case, rows containing missing variables are deleted. It is recommended that the fraction of deleted rows does not exceed 30%.
2. Pairwise. In this case, only the missing observations are ignored and analysis is done on the variables present.

Both of the above methods suffer from loss of information. Listwise deletion suffers the maximum information loss compared with pairwise removal. But the problem with pairwise deletion is that even though it makes the data available, one cannot compare analyses because the data samples are different every time.

**Imputation**
Imputation is an attempt to keep the missing records by using replacement from available data. The two most popular techniques are averaging and building predictive models. In both cases, it is recommended to limit the fraction of imputed data to less than 20%.

**Imputation by Averaging Techniques**
The mean, median, and mode are the most popular averaging techniques which are used to infer missing values. Approaches ranging from a global average for the variable to averages based on groups are usually applied. Though we can get a quick estimate of the missing values, we significantly reduce the variation in the data set.

**Predictive Modeling Techniques**
A better approach is to apply imputation methods that use the other variables in the data set to build a model to estimate the missing data. The most used methods are multivariate least squares, maximum likelihood, and singular value decomposition. However, it is important to understand the imputation algorithms before applying them, because the results from different algorithms may be significantly different.

An interesting case is decision trees. A decision tree can be trained using a variable that has missing values as its target and all the other variables in the data set as inputs. In this way, the decision tree can learn acceptable replacement values for the missing values in the temporary target variable. This approach requires one decision tree for every input variable that has missing values, so it can become computationally expensive for large and dirty training sets.

We also have to consider the anonymous opinion that imputation is often thought of as the statistical equivalent of witchcraft.
8.3.2 Handling Outliers

Defining Outliers
Outliers can be defined as data points that violate the general pattern of smooth or otherwise regular variation seen in the data sequence.\(^4\) A clear example of an outlier is the sudden drop in Var1 shown in Fig. 8.3. There is no rigid mathematical definition of what constitutes an outlier. Some statistical rules are available, but determining whether or not an observation is an outlier is ultimately a subjective exercise based on detection methods and domain knowledge.

There are several approaches to detecting outliers, which are classified into the following groups:\(^5\)

- **Extreme value analysis.** This is the most basic form of outlier detection and only valid for univariate data. It is assumed that values which are too large or too small relative to the majority of the data are outliers. Examples of such statistical methods are the Z-test and Student’s t-test. These are good heuristics for initial analysis of data but they cannot detect multivariate outliers. They can be used as final steps in interpreting the outputs of other outlier detection methods.

- **Probabilistic and statistical models.** These models assume specific distributions for the data. Then, using expectation-maximization methods they estimate the parameters of the model. Finally, they calculate the probability of membership of each data point in the calculated distribution. The points with low a probability of membership are marked as outliers.

- **Linear multivariate models.** These methods model the data into lower-dimensional subspaces with the use of linear correlations. Then the distance of each data point to a plane that fits the subspace is calculated. This distance is used to find outliers. PCA is an example of a linear model for anomaly detection. It uses an error metric, called \(T^2\), that defines an upper confidence limit for outlier identification.

- **Proximity-based models.** The idea of these methods is to model outliers as points which are isolated from the rest of the observations. Cluster analysis, density-based analysis, and nearest-neighborhood methods are the main approaches of this kind.

It has to be taken into account that some data analysis algorithms are more sensitive than others when dealing with outliers. Supervised machine learning algorithms that use a squared-loss function to determine the parameters that best fit the training data are heavily influenced by outliers. For example, gradient-boosting algorithms add large weights to observations that are considered to be hard cases, i.e., potential outliers. Some clustering algorithms, such as K-means clustering, can be quite

---

\(^4\) R. Pearson, Mining Imperfect Data: Dealing with Contamination and Incomplete Records, SIAM, 2006.

\(^5\) C. Aggarwal, *Outlier Analysis*, Springer, 2013.
sensitive to outliers because they try to minimize the sum of squared distances from cluster member points to the cluster means; a large deviation caused by an outlier receives a lot of weight.

**Outlier Removal**
Fortunately, outliers can usually be detected by some simple initial data exploration using univariate, bivariate, and multivariate plots. An example of identification of outliers from univariate data distributions is shown in Fig. 8.10.

First, we need to determine whether an outlying value is simply an invalid or erroneous entry that can be disregarded. If we have determined that an outlier provides no valuable information, it is acceptable to simply filter it out. This is the case with the single outlier shown in Fig. 8.10, where a single data point is far outside the normal distribution of the rest of the data. Handling the cluster of outlier, shown in Fig. 8.10, is not so trivial, though. In this case, a cluster of data points is far away from the majority of the data. On the one hand, removing the cluster will significantly improve the data distribution and potential for the model building. On the other hand, however, that cluster may include data for a potential new operating regime. The decision about its removal should be based on consulting the SMEs.

![Fig. 8.10 Identification of outliers from univariate data distributions](image-url)
A faster way to identify outliers is by multivariate methods, such as by using $T^2$ statistics derived from PCA. An example of such a determination is shown in Fig. 8.11.

Each dot in the plot is a value of the $T^2$ statistic, which is an indication of how close the data point is to what is statistically defined as the normal state. The important metric is the upper confidence limit (UCL), which defines the threshold between normal and abnormal behavior for a selected confidence level. Records with $T^2$ values above the threshold are identified as multivariate outliers. In the case shown in Fig. 8.11, 22 outliers were identified based on a UCL of 16.65 and a confidence limit of 95%. Handling these outliers can be done in different ways, however. The individual outliers with very high $T^2$ (for example >30) can be removed, but the fate of the set of outliers in the first 10 samples should be discussed with domain experts.

**Special Treatment of Outliers**

An acceptable reason for the removal of an outlier is when it provides no valuable information. However, if it has been determined that outliers might represent some real but rare relationship or if the information from the other features in those observations is too valuable to discard, then they can be used for special treatment by selected modeling techniques.

One such technique, used in time series modeling, is defining outliers as events. Some of the most common events involve a level shift up or down, a change in the trend, or a single observation point, also called a pulse. The events are integrated into the original data set by adding dummy variables with values of zero (no-related to the specific event) and one (related to the specific event).
Another technique for special treatment is to use outliers as a novelty detector. The assumption is that the deviated data points are the seeds of a new operating condition that can define a cluster of a new form. If this assumption is correct, this cluster will attract subsequent data and will gradually become the dominant operating regime. If this assumption is incorrect, the relative weight of this small cluster will decrease in the future, and then it could be removed.

### 8.3.3 Data Transformation

**Rescaling**

Real-world data includes variables whose values vary significantly in magnitude and range. These big differences can degrade the performance of several Data Science methods. Examples are approaches based on a distance measurement (such as K-means clustering), those that use numerical gradient information in their solution (such as neural networks and support vector machines), and those that depend on a measure of the variance (such as principal component analysis). The objective functions used in many of the relevant algorithms can be dominated by the variables that have a large variance relative to other variables, preventing the model from being able to learn the relationship with the other variables.

Two methods for rescaling data are well known—normalization and standardization. Normalization scales all numeric variables into a range of [0,1]. One possible formula is

\[
x_{\text{normalized}} = \frac{x_{\text{original}} - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}}
\]

where \(x_{\text{min}}\) and \(x_{\text{max}}\) are the minimum and maximum values, respectively, of the selected variable, \(x_{\text{original}}\) is the value of the variable in the original numbers, and \(x_{\text{normalized}}\) is its normalized value in the range [0,1].

The other option for rescaling is standardization of the data set. This transforms the original data to have zero mean and unit variance, for example by using the equation

\[
x_{\text{stdrt}} = \frac{x_{\text{original}} - \mu}{\sigma}
\]

where \(x_{\text{original}}\) is the value of the variable in the original numbers, \(\mu\) is either the mean or the median of the selected variable in the original numbers, \(\sigma\) is the corresponding standard deviation also in the original numbers, and \(x_{\text{stdrt}}\) is the standardized variable.

Both of these techniques have their drawbacks. Normalization is very sensitive to outliers. If they are present in the data, normalizing the data will certainly scale the “normal” data to a very small interval. On the other hand, when standardization is used the new, standardized data is not bounded (unlike the case of normalization).

**Improved Distributions by Power Transforms**

The power transform family of functions are typically used to create monotonic data transformations. Their main significance is that they help in stabilizing the variance, keeping the data close to a normal distribution and making the data independent of the mean based on its distribution.
The most popular transform in this family is the log transform. It can be represented mathematically as $y = \log_b(x)$.

This can then be translated into $b^y = x$.

The natural logarithm uses $b = e$, where $e = 2.71828$ is popularly known as Euler’s number.

Log transforms are useful when applied to skewed distributions as they tend to expand the values which fall in the range of lower magnitudes and tend to compress or reduce the values which fall in the range of higher magnitudes. This tends to make the skewed distribution as normal-like as possible. An example of the benefits of a log transform is shown in Fig. 8.12.

As it is shown in Fig. 8.12, the original tailed distribution is transformed into a dual normal distribution that gives a better chance of discovering features and relationships.

**Time Series Transformations**

- **Stationarity**

A stationary time series is one whose statistical properties such as the mean, variance, and autocorrelation are all constant over time. One of the key reasons for trying to make a time series stationary is to be able to obtain meaningful sample statistics such as the mean, variance, and correlation with other variables. For example, if the series has a consistently increasing trend over time, the sample mean and variance will grow with the size of the sample, and they will always underestimate the mean and variance in future periods. If the mean and variance of a time series are not well defined, then neither are its correlations with other variables, which are the basis of data analysis and model building.

Unfortunately, most real-world economic time series are far from stationary, and they exhibit trends, cycles, seasonality, and other nonstationary behavior. An example of such a time series is shown in Fig. 8.13.

Several data transformations, such as differencing, detrending, and deseasonalizing need to be applied until stationarity is achieved. They are discussed briefly below.
Differencing

The first difference of a time series is the series of changes from one period to the next. For our example in Fig. 8.13, the time series $Y(t)$ is a quarterly time series. If $Y(t)$ denotes the value of the time series $Y$ at period $t$, then the first difference of $Y$ at period $t$ is equal to $Y(t) - Y(t-1)$.

In the case of seasonal data (as is the case in Fig. 8.13), it is recommended that nonstationarity be reduced by using the seasonal difference (that is, the difference between an observation and the corresponding observation a year ago). For example, the first seasonal difference for the quarterly data $Y(t)$ with an annual period is equal to $Y(t) - Y(t-4)$. A plot of seasonal and first differences applied to the original data set $Y(t)$ is shown in Fig. 8.14.

When both seasonal and first differences are applied, it is recommended that the seasonal difference be done first because there is a chance that the resulting time
series will be stationary, and there will then be no need for a first difference. If stationarity cannot be accomplished by a first difference, it is possible to apply a second-order regular and seasonal difference to the data.

**Detrending**
A trend in a time series is a slow, gradual change in a property of the series over the whole interval under consideration. Unfortunately, identifying a trend in a time series is subjective because a trend cannot be unequivocally distinguished from low-frequency fluctuations. What looks like a trend in a short-segment of a time series often proves to be a low-frequency fluctuation (perhaps part of a cycle) in a longer series.

Detrending is the statistical transformation of removing a trend from a time-series. Many alternative methods are available for detrending. A simple linear trend in the mean can be removed by subtracting a least-squares-fit straight line. More complicated trends might require different procedures, such as curve fitting, digital filtering, and piecewise polynomials.

The trend decomposition of the original time series $Y(t)$ is shown in Fig. 8.15a.

**Deseasonalizing**
Deseasonalizing is the statistical or mathematical operation of removing seasonal patterns from a time series. Usually, the removal of the seasonal patterns is combined with decomposition of the time series into cyclical and irregular components. The major distinction between a seasonal and a cyclical pattern is that the former has a constant length and appears at regular intervals, and the latter varies in length. In our
example of the original economic time series $Y(t)$, an annual seasonality is observed in the quarterly data. The deseasonalizing transformation is shown in Fig. 8.15b.

### 8.3.4 Data Balance

In many applications, such as fraud detection, analysis of machine failures, and medical screening, the data is heavily unbalanced, with a crowded majority class and a miniscule minority class. Often the minority class is less than 1% of the data and we face “needle in a haystack” problems, where machine learning classifiers are used to sort through huge populations of negative (uninteresting) cases to find the small number of positive (interesting) cases.

Unfortunately, conventional algorithms are often biased toward the majority class because their loss functions attempt to optimize quantities such as the error rate, not taking the data distribution into consideration. It is also possible that the minority examples may be treated as outliers of the majority class and ignored. The learning algorithm simply generates a trivial classifier that classifies every data record into the majority class.

The recommended solution is to resample the training set and create a new data set where both classes are balanced.

**Resampling Training Data**

Balance can be accomplished by two approaches—undersampling and oversampling. Undersampling randomly downsamples the majority class. Oversampling randomly replicates minority instances to increase their population. Some details are given below.

**Undersampling**

Undersampling balances the data set by reducing the size of the majority class. This method is used when the quantity of data is sufficient. By keeping all samples in the minority class and randomly selecting an equal number of samples in the majority class, a balanced new data set can be retrieved for further analysis and modeling. The principle is illustrated in Fig. 8.16.

**Oversampling**

In contrast, oversampling is used when the quantity of data is insufficient. Balance is accomplished by increasing the number of minority samples. Rather than getting rid of majority samples, new minority samples are generated by using repetition, bootstrapping or other methods. The principle of oversampling is illustrated in Fig. 8.17.

Existing experience from implementing data balancing shows that there is no absolute advantage of one resampling method over the other. Application of these two methods depends on the business application and the dataset itself. A combination of over and undersampling is often recommended for improving the results as well.
Another step in data preparation that is very important for successful data analysis and modeling is achieving the proper balance between training and test data. The expectation for both manufacturing and business data is that the test data will be selected from the most recent data. Usually the balance between the most recent data and the rest of available data history is visualized by comparing the distributions of training and test data. An example of emissions data is shown in Figs. 8.18 and 8.19.
Figure 8.18 shows a plot of the target variable and the periods selected for the training (the initial 170 records of hourly averaged measurements) and test (the last 114 records of hourly averaged measurements) emissions data. The quality of the balance is best demonstrated by the ranges of the two distributions, shown in Fig. 8.19. Both the low range (11.4 for training vs. 15.8 for test) and the high range (162.8 for training and 185.5 for test) are close, which confirms the good balance between the data sets. In many cases, however, the balance is not trivial and some data rearrangement is needed.
8.3.5 Data Quality Assessment

The final step in data preparation is estimating the data quality after applying all the procedures discussed. The key question to be answered is: Is the data ready for productive data analysis and model development? Often the answer is not clear in black or white but has several shades of gray. The most generic division of data quality is into three broad categories: good, acceptable, and poor. It is not trivial to define firm quantitative criteria for this ranking, because these categories are strongly problem dependent. An attempt to describe qualitative measures for data quality of these three categories is given below:

- **Good-quality data.** The available internal and external data has a sufficient size, defined by the numbers of variables and records, to allow effective generation of solutions to the problem. The collected data contains a small amount of missing data and outliers. All of this have been successfully handled by removal or imputation. The final size of the data set after data preparation has not been reduced by more than 10%. Several well-balanced training and test data sets are available for effective model development. The recommended decision is to fully accept the prepared data and move on to the next steps of the Data Science workflow.

- **Acceptable quality data.** Most of the needed internal and external data has been collected but the size of the available data, defined by the numbers of variables and records, is insufficient to allow full-scale generation of solutions to the problem. Some data is costly and a decision to purchase some such data can be made depending on the data analysis. The collected data contains a moderate amount, less than 20%, of missing data and outliers. Most of these have been successfully handled by removal or imputation. The final size of the data set after data preparation has been reduced to 70–80% of the original size. Balancing training and test data is difficult and the generated files are barely balanced. It is recommended to prepare several versions of such data sets for model development with different fractions of training and test data. The recommended decision is to cautiously accept the data and to be very careful in the data analysis and model development steps. One option for improving the data quality is not to select the poor-quality variables during variable selection.

- **Poor-or low-quality data.** There are big gaps between the needed and collected internal and external data due to sensor failures, long shutdowns, or a short history. The size of the available data, defined by the numbers of variables and records, is insufficient to generate solutions to the problem based on data-driven methods. The collected data contains a significant amount, greater than 30%, of missing data, outliers, and data with strange patterns. It is impossible to handle these by removal or imputation because the final size of the data set after data preparation would be be reduced to less than 50% of the original size. This is a clear example of the Garbage-In type of data. The recommended decision is to reject the data and discuss with the project team the potential options to the data quality, such as sensor replacement, improvement of the data collection.
infrastructure, purchasing of needed external data, and longer data collection. A decisive message that continuing with this Garbage-In data will lead to a “spectacular” Garbage-Out fiasco should be sent to all stakeholders.

8.4 Common Mistakes

8.4.1 GIGO 2.0

As we have discussed several times in the book, one of the biggest dangers in applied Data Science is the almost religious belief that garbage in the data can be compensated for by the magic of advanced data analysis methods, especially those based on AI. We defined this mode of operation as Garbage-In-Gold-Out or GIGO 2.0, and gave more details in Chap. 1.

8.4.2 Problem Solving with Insufficient Data

Another typical mistake is to try to deliver a solution to the problem with insufficient data in terms of number of variables and with a limited number of records. One of the causes of this mode of operation is a lack of knowledge about the required variables and the available records. A good knowledge acquisition step should give an answer to this key question. An inevitable result of using such an informationally incomplete data set is deriving an imperfect data analysis and models with reduced performance. The issue of insufficient data can be resolved by adding new variables and collecting or purchasing more data records.

8.4.3 Problem Solving with Low-Quality Data

The most common mistake, typical of inexperienced data scientists, is to try to solve the problem at any cost even with low-quality data. They don’t have the courage to say a clear NO and to stop a waste of time and effort after identifying the data quality as poor. They are afraid that killing the project at an early phase, such as data preparation, will be accepted very badly by the project sponsors and their reputation will suffer. The strategy preferred by them is to continue with data analysis and model development and only then to deliver the bad news. The argument is that everything possible has been done but, due to low-quality data, the solutions derived have poor performance. They believe that in this case the negative message is less damaging. The problem is that a lot of time and effort has been lost in fruitless work.
8.4.4 **Low-Quality Data Preparation**

A known habit of less experienced data scientists is ignoring the importance of data preparation. They are obsessed with the significance of modeling and prefer to reach that step of the workflow with minimal effort. The result is low-quality imputation, ineffective handling of outliers, and unbalanced training and test data sets. Often the impact of this fast-track data preparation on the quality of model generation is disastrous.

8.5 **Suggested Reading**

- C. Aggarwal, *Outlier Analysis*, Springer, 2013.
- R. Pearson, *Mining Imperfect Data: Dealing with Contamination and Incomplete Records*, SIAM, 2006.
- D. Pyle, *Data Preparation for Data Mining*, Morgan Kaufmann, 1999.

8.6 **Questions**

**Question 1**
What are the key internal data sources in your organization?

**Question 2**
Why does economic data need to be aligned with the business structure?

**Question 3**
What are the benefits of visual data exploration?

**Question 4**
Can outliers be removed automatically?

**Question 5**
How can data be balanced?