10 McDonald quotations from Davenport TH, Iansiti M, Serels A (2013) *Managing with Analytics at Procter & Gamble*, Harvard Business School case study, April.

11 The three core skills needed for analytical managers is adapted research originally published in Harris J (2012) Data is useless without the skills to analyze it. *Harvard Business Review*, September 13. Available at https://hbr.org/2012/09/data-is-useless-without-the-skills.

12 Davenport TH, Harris JG (2017) *Competing on Analytics* (Harvard Business Review Press, revised edition).

13 This section draws on content from a research brief by Harrington E (2014) *Building an Analytics Team for Your Organization*, International Institute for Analytics, September. Available at http://iianalytics.com/research/building-an-analytics-team-for-your-organization-part-i.

14 This section is a revised and updated version of a chapter by Morison R, Davenport TH (2012) Organizing analysts, in *Enterprise Analytics*, Davenport TH, ed. (Prentice Hall).

15 Accenture Institute for High Performance (2010) *Counting on Analytical Talent.*

16 Davenport TH, Harris JG, and Morison R (2010) *Analytics at Work: Smarter Decisions, Better Results* (Harvard Business Press), pp. 104–109.

17 Framework is based on *Building an Analytical Organization*, Business Analytics Concours and nGenera Corporation, 2008.

18 Framework is based on Morison R, Davenport TH (2008) *Mastering the Technologies of Business Analytics*, Business Analytics Concours and nGenera Corporation.

4

The Data

Brian T. Downs

Accenture Digital, Data Science Center of Excellence, Dallas, TX, USA

4.1 Introduction

Regardless of one's area of specialization or interest, it is true that most analytics students and professionals devote most of the effort and energy that goes into training to learning analytics methods and algorithms. A review of a typical curriculum in business analytics will reveal a sharp focus on the tools and techniques required, often in a specific context such as marketing or operations, to be a successful analytics practitioner. Therefore, it is often a surprise for people starting out in the field to discover that on most analytics projects, most of one's time is not spent on using the algorithms recently mastered with such great effort and determination. Rather, it is the lot of an analytics professional to spend most of their time messing with data. This chapter provides a practitioner's view of the different types of data, and some of the challenges in identifying, collecting, and preparing data for analysis.

4.2 Data Collection

4.2.1 Data Types

Before exploring data collection, a review of the various types of data will be useful. Figure 4.1 shows a useful hierarchy for describing these.

Qualitative data result from classifying something or labeling its attributes. There are three main types of qualitative data. *Nominal* data results when we identify things with named categories that do not have any natural or intrinsic value associated with them. For example, the wooden poles a utility company uses to transmit power to its customers can be classified by the species of tree from which they are made. Pine, fir, and cedar are meaningful categories in that each

INFORMS Analytics Body of Knowledge, First Edition. Edited by James J. Cochran.

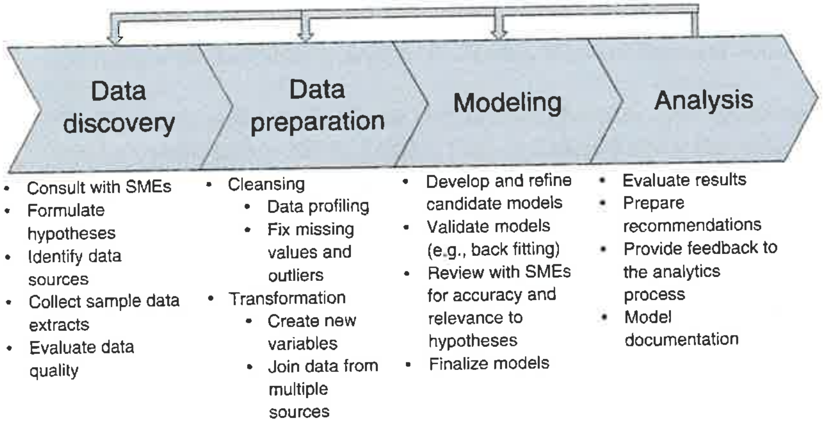

Data discovery
Data preparation
Modeling
Analysis
• Consult with SMEs
• Formulate hypotheses
• Identify data sources
• Collect sample data extracts
• Evaluate data quality
• Cleansing
• Data profiling
• Fix missing values and outliers
• Transformation
• Create new variables
• Join data from multiple sources
• Develop and refine candidate models
• Validate models (e.g., back fitting)
• Review with SMEs for accuracy and relevance to hypotheses
• Finalize models
• Evaluate results
• Prepare recommendations
• Provide feedback to the analytics process
• Model documentation

Figure 4.1 The analytics process.

has intrinsic properties that affect their performance in this application, but there is no obvious way to rank them based on the nominal classification alone. One could use this classification to perform an analysis to see if there are statistically significant differences in the lifespan of wood poles made from each species.

An important special case of nominal data is *binary* data. This type of data places something into one of two mutually exclusive and collectively exhaustive categories, often implying opposite states. A quality inspection of an item on a production line can result in a pass/fail. A production process can be in control/out of control. A magazine subscriber can renew/not renew. This type of data has become increasingly important as methods for predicting how likely an event of interest is to occur have seen widespread use in a variety of contexts. A manufacturing company may wish to predict how likely a machine is to fail given current operating conditions using data that can be collected from the production process. A cable television provider may wish to predict how likely a customer is to drop their cable service given demographic information and their history of problems and complaints. There are many number of classification methods that can predict events with binary outcomes effectively. The challenge in many cases is that historical data that contain multiple instances of each outcome may not be available, or will require some time to collect.

Ordinal data are created when one classifies things into categories where there is an implicit relationship between categories. The use of small, medium, and large to describe the size of things has an implicit meaning in many contexts. We expect a medium drink to contain more than a small drink, and a large drink to contain more than a medium drink. In the context of completing a survey, one might be asked to rank something from worst to best on a scale of 1–10, with the expectation that 5 is better than 2, 10 is better than 6, and so on. The problem with both of these examples is that the rank ordering does not tell us the

magnitude of the difference between each category. There is no way to know from the classification that a medium drink is 33% bigger than a small drink, or how much better in absolute terms a rank of 10 is than a rank of 5.

Quantitative data are created when things are counted or measured. *Discrete* data result from counting things, and therefore is typically expressed as an integer value. The number of nights one has stayed with their preferred hotel chain is an example of discrete data. The number of warranty claims received on a model of smart phone is another. Neither of these things are recorded as fractional values as they refer to discrete events that have occurred an integer number of times.

Continuous data are generally anything that can be measured, and as such may have fractional values depending on how fine of a measurement is made. The flow rate of crude oil through a pipeline, the exhaust temperature of a diesel engine, and the daily output of a chemical process are all things that can be measured and the result will generally be a real number. One thing to be cautious about when using continuous data is that the quality and reliability if the data can be affected by the method of collection. Devices such as electronic sensors can be unreliable or influenced by the surrounding environment. Data recorded by human interaction are naturally prone to errors.

Time is also a potential consideration. Data collected from several subjects at approximately the same point in time are referred to as *cross-sectional* data. Examples of cross-sectional data are candidate preferences of voters immediately prior to an election, the high temperature on a given date in the 100 most populated U.S. cities, or the sizes of donations given to a charity during a fund raising drive. The most common purpose for collecting cross-sectional data is to develop an understanding of characteristics of a population at a particular point in time. Data collected from a single subject at several approximately equally spaced points in time are referred to as *time series* data. Examples of time series data are weekly sales for an item, the daily number of visitors to a museum, or the monthly rainfall measured at a weather station. The most common purpose for collecting time series data is to predict future values of the time series, such as when a sales history is used to predict future sales for planning purposes. In many cases, a time series will have measurable components that can be estimated using appropriate analytical methods. A trend indicates a long-term shift in the overall level of the time series, while seasonality is a cyclic pattern that repeats within a specific time interval such as a day or a year. If a single subject is observed over several points in space, the data are referred to as *spatial data*. Spatial data are similar to time series data and are often analyzed with methods designed for time series data. Finally, data are increasingly collected from several subjects at several approximately equally spaced points. These data, which have characteristics of both cross-sectional and time series data, are referred to as *panel data* (or *longitudinal data* or *cross-sectional time series data*). Time series and cross-sectional data are each a special case of panel data in which either the number of periods of time or the number of subjects is one.

Another type of data that has proliferated in recent years is unstructured *text data*. New technologies have been developed to collect and store this type of data, which can be collected from Web sites, social media, and discussion groups in the form of comments, reviews, and opinions. This type of data is stored as documents and may be used for text mining and sentiment analysis. As will be discussed later, the lack of structure in this type of data makes it difficult to store in a traditional database. Therefore, it has been a driving force in the evolution of nonrelational database technologies in recent years.

INTERVIEW WITH HARRISON SCHRAMM

Sometimes clients have data that could be useful to an analytics project. Harrison Schramm, who recently retired from a 20-year career as a helicopter pilot and an operations research analyst in the U.S. Navy, shares his thoughts on obtaining data for an analytics project from a client:

There are two ways to approach this, and the choice depends on the stakeholder. One way to go about it is to get stakeholders excited about what you are doing and make them want to help you by giving you their data. This is the preferred route.

The other route is to make stakeholders utterly terrified of what you are going to do if they don't give you data. This is a horrible route to take, but sometimes you have to go down this path. If you are working with a large organization, you cannot expect every segment of that organization to be excited about what you're going to do. So if one department is recalcitrant, you just end up having to say, "If you don't give us the data, then we're going to assume this, this, and this . . ." and you pathologically craft those assumptions so all of a sudden that giving you their data looks a lot better to them than those assumptions you're threatening to make. It's a varsity move–it's not for freshman.

This is an excerpt from one of a series of interviews with analytics professionals and educators commissioned by the *INFORMS Analytics Body of Knowledge* Committee.

4.2.2 Data Discovery

There are two types of analytics projects that are often encountered in practice. *Management Consulting*-type projects involve the use of analytics to solve a problem or answer a particular set of questions. These types of projects deal with one-time decisions and the "leave behind" from the effort is a report that contains analysis and recommendations. The questions addressed can be relatively simple, such as "Should I add storage capacity at a facility?", or they can be complex such as "How can I reduce my variable conversion cost while pursuing a high variety, highly customized make-to-order production strategy?" The analytics practitioner may use a single technique, or a combination of predictive analytics, simulation, and optimization. The data may come from a variety of internal and external sources, but are generally discarded after the project is complete. As such, there may not be a need to collect and merge the data into a permanent and sustainable environment. Data often will be collected in spreadsheet format, from a variety of sources, and will require considerable manual effort to prepare. Detective work may be required to locate some data elements as there may not be a system of record that contains what is needed, or even worse the data that are in the system of record may not be accurate. These types of projects can often be completed by people with analytics as their primary skill set as such people usually have some basic data management skills as well. Larger projects with high volumes of data may require data integration specialists to assist with data preparation.

The other common type of analytics project is *Application Development*. In these projects, analytics tools and algorithms are imbedded in an information technology system to support a set of business processes. Examples of such processes include forecasting and demand planning, sales and operations planning, and production process monitoring and control. In these applications, the analytics component is executed on a periodic basis. This can be anywhere from fractions of a second to monthly, depending on the type of process. Supporting a recurring process requires that the data needed by the analytics models be current and complete. This generally requires the development of a data warehouse, or at least an operational data store (ODS), which may involve combining data from a variety of sources in a single environment, and developing extract, transform, and load (ETL) procedures that capture data from source systems and move it into the analytics environment. ETL processes will also clean and transform data, creating tables and views that can be loaded directly into analytics applications. For these types of projects, the "leave behind" will be a functioning application, as well as the data infrastructure necessary to support it. In addition, there will be documented procedures in place to maintain the integrity of both the analytics models and the data. Application Development projects usually require skill sets beyond analytics, including data integration, system architecture, and data visualization.

At a high level, most analytics projects follow a similar process flow, although the amount of effort and complexity can vary widely depending on the scale of the initiative. Figure 4.2 shows a high-level view of typical activities that an analytics practitioner will undertake to complete a project. Note that a sizable proportion of the activities are data related. Data discovery is a critical first step as it is necessary to define the objectives of the work, as well as to determine the likelihood of a successful outcome.

Data discovery must begin with discussions with subject matter experts (SME) to understand the research question to be addressed with analytics. In a business

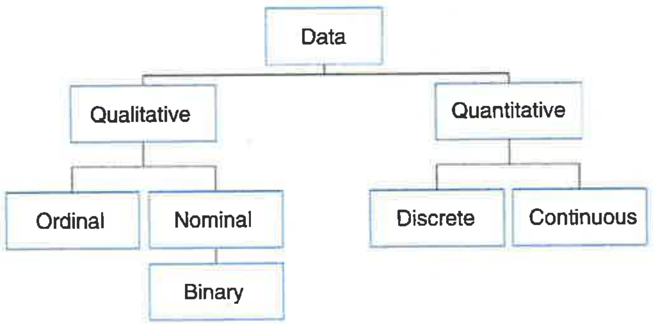
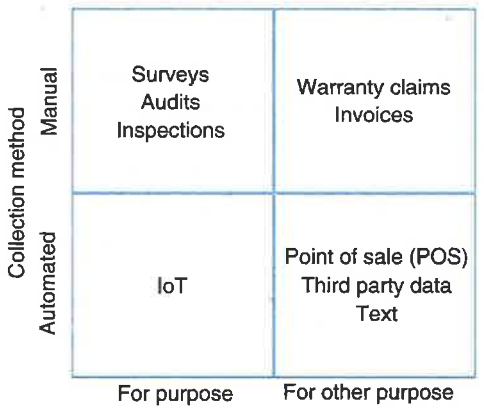


Figure 4.2 Types of data.

context, these are typically people who work in a client business and who are familiar with the attendant operational and business processes. It is the job of the analytics practitioner to understand the business issues at stake, and to frame those issues as testable hypotheses or to propose an approach for addressing a business need. An example of the former is "The results of our supplier audits can be used to predict which of them are most likely to be noncompliant in the next audit cycle." An example of the latter could be "We can use simulation modeling to understand the effect of different lot size policies on our manufacturing conversion costs."

Once the problem and analytics approach have been identified, it is necessary to determine whether the proposed approach is feasible. Critical elements for answering this question are the availability and quality of the necessary data. This requires carefully listing all of the required data elements for the analytics work and identifying possible sources for each. A data source will have an owner, whether it resides in a corporate information system or in a spreadsheet on a personal computer. Enlisting the cooperation of a data source owner is a key to success in analytics projects, as access to data and help understanding its format and structure are essential.

Figure 4.3 provides a way to categorize potential data sources. Along one dimension, one can think of data that are collected manually versus data that are collected by an automated process. Any process that involves a human being recording data on paper or through a form, electronic or paper, is manual data entry. Automated data collection does not require human intervention. Along the other dimension, there are data that are collected specifically in support of the analytics project at hand versus data that have been collected to support another business process but which can be used for the analytics project at hand. This last can be problematic since the specifications of the data being collected were designed to support a different objective, and there is a good chance that it will not be an exact fit for the needs of the current effort. It will likely require additional effort to augment and transform such data into a form fitting the current objective.



Collection method	For purpose	For other purpose
Manual	Surveys Audits Inspections	Warranty claims Invoices
Automated	IoT	Point of sale (POS) Third party data Text

Figure 4.3 Data sources.

Surveys, audits, and inspections are all examples of methods requiring manual data collection that are designed to investigate specific questions using analytics. Surveys use statistical methods to identify a representative sample of a target population to measure their response to a set of research questions. Audits measure compliance to a set of standards, usually along several dimensions. Inspections entail a point-by-point examination of specific operating criteria, usually with a binary (pass/fail) outcome. For each of these, people are directly involved in the gathering and recording of the data, and thus there are opportunities for errors to occur in the process. These can take the form of simple keystroke errors, known as "fat fingering," or the failure to correctly record a response or observation. Other concerns relate to the effect of human judgment on the data collection process. Survey respondents may not answer truthfully because of a reluctance to express an unpopular viewpoint. Different auditors may evaluate the same situation as having different levels of compliance. Different inspectors may employ different thresholds as the standard for a pass/fail recommendation, or even fail to complete the entire inspection process. It is essential that an analytics practitioner be mindful of these potential sources of trouble when using such data to analyze the populations on which these tools are used.

There are transaction-oriented systems that rely upon manual data entry as the means by which data are digitized. Manual processes are often used to create and process invoices, creating records of customer, product, pricing, and shipment information. Such records are initially created for accounting purposes as a financial record of the transaction, but the same data can be used for other analysis such as sales forecasting and production planning. An industrial equipment manufacturer had a system that relied on data entered manually by technicians at their dealer sites to create warranty claims. The system had an electronic form that needed to be completed with data such as the product serial number, time of failure, the parts replaced, and a failure code to classify the nature of the claim. While initially used as a way to receive and process warranty claims so that dealers can be reimbursed for warranty service, over time this system creates a history of warranty claims that can be used by other analyses such as predictive maintenance, root cause analysis, and fraud detection.

An anecdote from the last example highlights the importance of a thoughtful design when creating tools for manual data entry. The field in the form that requested a code to classify the specific failure mode was free text, rather than a pull-down list that provided specific choices. Often the technicians would be in a hurry, or would not have the right code handy, and would enter an invalid code "99" just to complete the form and get the claim submitted. They would write short details in a free text field to describe the work performed. While in most cases this was enough information to get the claim reimbursement, it created a history of warranty claims that did not have the correct failure mode associated with many of the records. This made the data almost unusable for deeper analysis without someone trying to manually review the free text fields to recode the problem records, a task that proved impractical from both time and accuracy perspectives.

Some basic guidelines for design of forms for manual data collection can alleviate some of the data quality risks inherent in this method of data collection:

- Automate the workflow as much as possible, eliminating intermediate steps that use paper or spreadsheets. The use of tablets or other mobile devices for data collection in the field will improve both the accuracy and completeness of data.
- Limit the use of text fields.
- Use pull-down lists, radio boxes, and check boxes wherever possible instead of text fields to limit the potential for errors.
- Make clear which fields are required for the form to be submitted.
- Be sure that required data formats (e.g., dates, currency) are clearly indicated on the form.
- Validate the data and correct errors before allowing the form to be submitted.

This list is not exhaustive and there are many resources available on the Internet to assist in designing forms for data collection. It is important that an analytics practitioner be mindful of these issues when designing tools or applications for these types of applications.

A common lament from many companies is that they are drowning in data, but do not know how to extract value from all the data they possess. This can be attributed in part to factors such as the low cost of data storage, and the proliferation of inexpensive sensors that can be used to collect data from equipment at intervals as small as a fraction of a second. In many industries, the collection of process data from sensors and other systems such as SCADA (Supervisory Control and Data Acquisition) is a prevalent form of automated data collection that is used to monitor and control processes, as well for other

analytics-driven applications such as predictive maintenance. This type of data is the life blood of the Internet of Things (IoT), as the automation of the data collection process enables the automation of monitoring and control processes. This type of data consists of high-frequency time series, typically measurements of process parameters such as pressure, temperature, or velocity. It is usually captured and stored in a *data historian*, such as AspenTech's InfoPlus.21 or OSISoft's PI, that is designed for the efficient storage and retrieval of time series data. Examples of this type of automated data collection can be found in a variety of industries. According to *Aviation Week* (http://aviationweek.com/connected-aerospace/internet-aircraft-things-industry-set-be-transformed), there are now jet engines that have over 5000 sensors, and produce over10 GB of data a second. Industrial equipment manufacturers that serve the mining industry have developed sensors and software that collect operational data from shovels and haul trucks. Utilities collect process data from power generation equipment such as turbines. And processes throughout the oil and gas industry are closely monitored using automated data collection, from upstream drilling and extraction through refining and downstream chemical production. Value can be extracted from this enormous volume of data, but not without considerable effort in the data preparation step of the analytics process.

Other types of automated data collection occur in systems that are used for transaction management purposes such as point of sale (POS) systems. Such systems are used to process transactions in retail operations and perform critical tasks such as invoice preparation, payment and membership discount processing, inventory management, and promotion processing. Since tools such as bar code and credit card readers are a part of the system, the need for manual data entry is nearly eliminated and errors are minimized. The resulting sales records, which contain information about both customers and products, are captured in a data base that can be used for a variety of analytics applications. These include customer segmentation to allow targeted promotions, supply chain segmentation to enable segment-specific strategies such as make to order (MTO) and make to stock (MTS), and forecasting and demand sensing to allow in-season adjustments to production quantities and inventory placement.

Another important data source to be considered is called third party data. These are data provided by an external source that will collect, cleanse, and transform data into usable form, typically on a subscription basis. This includes industry-focused companies such as S&P Global Platts, which provides energy and commodities information, including pricing. Experian is known as a credit reporting corporation with a global footprint. The U.S. Bureau of Labor Statistics compiles a variety of price and production indices, which range from the aggregate level to the industry or commodity specific such as construction, natural gas, and electric utilities. These indices are time series, and are based on good and services specific to the sector that they measure. They are often available in a seasonally adjusted form, with seasonal variation removed, making it easier for analytics models to identify correlations between different time series. This is especially useful when an index is a leading indicator, giving it predictive power for other related time series.

4.3 Data Preparation

Referring to Figure 4.1, one can see that the next step in the analytics process is data preparation. This step can be divided into two parts: data cleansing and data transformation. The objective of the data preparation step is to collect the data that have been gathered from various sources into a single location, and transform it into a form that can be consumed by analytical tools and software. Figure 4.4 shows the flow of data through the process, and the important activities conducted at each step.

Data profiling involves a univariate analysis of each of the variables in a data source, as well as a record-by-record evaluation of the completeness of the data. This is to allow the analyst to evaluate the suitability of the data for the project at hand. For quantitative data, this analysis will involve plotting the distribution of the variable, and identifying measures of central tendency such as the mean and median, as well as measure of dispersion including maximum, minimum, range, variance, and skewness. In addition to providing a sense of the overall shape of the data, this analysis provides insight about the possible probability distributions that may apply to the data, including whether the assumption of normality is warranted. In addition, profiling can help with the identification of missing values, extreme values, or problems with scaling.

Data profiling of qualitative data will involve the creation of frequency histograms to confirm that the data values are valid and complete. This aids in the identification of common data errors, such as missing or inconsistent values, or problems such as high cardinality of values in a categorical variable.

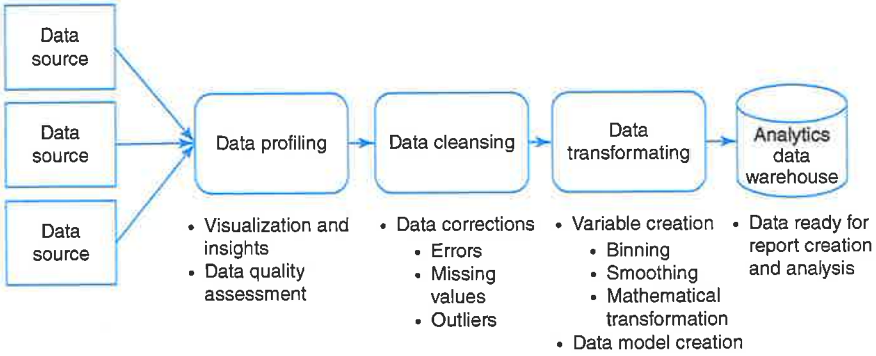


Figure 4.4 Data preparation.

Most commercially available analytics software has standard routines that are available for data profiling. These can greatly reduce the amount of effort involved in the process. However, this task can be completed using the tools that are available in a typical spreadsheet program, although the time and effort required to do so will be much greater than using a tool designed specifically for that purpose. No matter what the tool employed, a key output of data profiling is a list of data issues that need to be remediated to proceed with the analysis. What follows is a discussion of some common data problems that need to be addressed at this stage of the process.

Missing values are endemic to many data sources, and they can occur in a variety of ways often as the result of human involvement in the data collection process. Operational data collected in the field are particularly prone to missing values. Technicians may neglect to enter key information in a form such as identification codes for assets that they are inspecting simply because they cannot see through obstructions such as vegetation. Survey data may have missing values. It is typical that high-income respondents are reluctant to answer questions about their income level and may not respond to them. Whatever the source of the missing values, the critical question to answer is whether the missing values affect the representativeness of the data relative to the population from which it comes. If the sample size is large and the number of missing values is few, the missing values can be discarded without altering the results of the analysis. However, if the number of missing values is large, or if the incidence of the missing values is due to some systemic cause as described in the second example above, it will be necessary to attempt the estimation of the missing values.

In cases where data are gathered from operational systems, it may be necessary to pull data together from multiple sources to create records for analysis. For example, in the case described above, suppose there are technicians performing inspections in the field, and some of the data elements in the inspection form are missing. If there is a master list of assets, it may be possible to fill gaps such as missing identification codes by comparing timestamps from inspection records, and ascertaining the location of the crew at the time the inspection was performed. Comparing this type of data with the geographic coordinate data contained in the master list of assets may make it possible to identify the assets for which the identification codes are missing. In practice, this type of forensic approach to missing value correction is quite common, although it is labor-intensive and usually requires the assistance of someone who has a profound understanding of the data.

When there is a sensitivity to responding to certain questions, perhaps about topics such as income or politics, survey data may contain missing values that indicate response bias, called not missing at random (NMAR). One way to identify this is to create a new binary variable coded as response/no response, and to compare mean values of response variables between the two groups. If there are significant differences, this indicates a nonresponse bias and one should be careful about making inferences using such data.

In other situations, there will not be significant differences between the response and no response group for variables of interest. In this case, the data are missing at random (MAR). One can proceed with the analysis without fear of nonresponse bias, although there will be smaller sample sizes for the questions where there are no responses. If we repeat the comparison of means for all response variables and find no significant differences between the response and no response groups, then we have the best outcome and the data are said to be missing completely at random (MCAR).

There are two approaches often employed in situations where there are missing data due to no response on survey instruments. Pairwise deletion occurs when the responses to each question are summarized individually and the missing value is just excluded from the analysis. Similarly, one can perform correlation analysis on such data, but with a smaller sample size due to the missing values. List-wise deletion is used when using tools such as multiple regression or classification models. Since these methods seek to determine the relative influence of each of a group of predictor variables, any missing value requires that the entire record be excluded from the analysis.

Other techniques for handing missing values fall into the category of imputation. This is the substitution of some value for the missing values using mathematical methods of estimation. The simplest is to use the mean value of all observations for a missing value. This has the desirable property of not changing the mean of the variable, although it will dilute the correlation between that variable and any other. There are many other methods available for imputation of missing values. The reader is advised that many of them are quite advanced and will require some experience and skill to properly implement.

Another common problem with data comes in the form of *nonstandard values*. This happens when the same categorical data value is represented in the data with more than one set of characters. For example, the category "not applicable" may be represented as NA, N/A, n/a, N_A, and so on. This often results where there is manual data entry and nothing in the data entry process enforces the standardization of the response. This is common with abbreviation as well, including those for state and country names. This can also happen when sharing data between countries. A recent example involved data being shared between different groups within a multinational company. Certain special characters such as the @ and % that appear to be the same character have different values in Chinese and English character sets, creating instances of variables that looked the same but were different. While nonstandard values appearing in data sets are very common, it is a straightforward issue to fix. A frequency histogram of all observed values makes it easy to identify such cases, and a best practice is to automate the replacement of nonstandard values as part of an ongoing ETL process.

A related issue with categorical data can occur when a variable has a high cardinality. This occurs with data such as zip codes where the number of unique values is very high. These could also be variables such as e-mail addresses, user names, and social security numbers. The high number of unique values makes these variables impossible to use in tools such as linear models as such variables will not have enough observations per level of the variable to create a model. In practice, this type of data will either be discarded or transformed into a new variable using a technique called binning. For example, a long list of phone numbers may be mapped to a new variable made of just the area code. Zip codes could be mapped to a new variable called region, made up of a small number of geographical areas in the country. Binning is one example of data transformations that will be discussed in a later section of the chapter.

When dealing with quantitative data, the problem of outliers will often occur. An outlier is said to occur when a value is observed for a quantitative variable that is more than three standard deviations away from its mean. Outliers happen for many reasons, and understanding the reason for their occurrence is essential to knowing the proper remedy. Often an outlier is due to a mistake or malfunction. For example, heavy mining trucks have many sensors that monitor critical systems. These track important operating parameters such as the temperature and pressure of oil, fuel, and cooling systems, as well as tire pressure and payload, at intervals of a fraction of a second. However, the normal operating condition of a mining truck is to be carrying hundreds of tons of payload over rough terrain, or to have 40 ton loads dropped into the bed of the truck while loading, often in extremes of altitude and temperature. Under such harsh conditions, sensors can malfunction, as can the software used to collect the data. Data collected from operating assets will often have extreme values that are known to be erroneous. In such environments, observations with extreme values are discarded.

If the outliers in a data set are recurring and predictable, discarding those observations can mean a loss of valuable information that can cause bias in statistical models. However, many statistical modeling techniques are sensitive to extreme values. In such cases, an appropriate variable translation may be necessary to keep the predictive power of the variable reducing the influence of the outliers. An example of such a transformation is the creation of a new variable that is the base 10 logarithm of the observed variable, often after adding a constant to eliminate zeros and negative values.

Another problematic characteristic of some data is that it can be very *noisy*. This occurs when there is a high level of random variability in the data. The data collected from mining trucks described above is an example of data that are noisy. The shocks and vibration endured by the equipment result in data that have a high variance, obscuring the directional changes in the key operating parameters that may be occurring. Another example of data that is noisy is the intra-day price for a stock. Over the course of the day, there may be substantial variation in the price of individual transactions. This obscures directional changes that are of interest. High variability in time series data can be handled by transformations such as smoothing. One approach is to create a new variable that is based on a rolling moving average of a fixed number of the observed values. Since the high and low extreme values are effectively negated by the averaging, the result will be a new time series that has lower variability that will more transparently reveal any trends in the overall level of the time series. Another approach is to reduce the frequency of the time series by taking the minimum, maximum, and average of the time series over fixed intervals such as an hour. This creates three new time series that can be used to monitor not only the average value but also the range and variability.

Another issue with data that may require transformation is *skewness*. This happens when the distribution of continuous data has a long tail on one side, often because the values of the variable are bounded by zero on one side, leading to a long upper tail in the distribution. Such distributions are not consistent with the assumption of normality that is required for many parametric methods, and a transformation is used to mitigate this difficulty. Such transformations involve using a function that will impact the long upper tail the most. Examples of these transformations include logarithmic (ln(x), log10(x)), inverse ($1/x$), and square root (sqr(x)).

Data sets with multiple observations taken on the same population may experience high correlation between variables. While this correlation can be informative and a useful output of the data profiling process, one is advised to be cautious about using correlated variables in predictive models. These collinear variables contain redundant information, and can make it difficult to estimate the parameters of the models. Often this is due to a hierarchical relationship among variables, such as supplier name and source country. The analyst is encouraged to limit the inclusion of all but the most descriptive variables when modeling.

In the previous discussion, the methods for data transformation that have been discussed included *binning*, *smoothing*, and *fitting*. Binning divides the values of a continuous variable into intervals. Binnig discretizes the data, turning quantitative data into categorical data. Most statistical software will have a capability to create new categorical variable from bins using any number of methods such as assigning an equal number of observations to each bin, or creating bins of equal width and assigning a record to a bin if its value falls within the defined interval. Binnig can also be done based upon prior knowledge of the data.

There are mixed views about the use of binning. It does involve the loss of information, and the use of too few bins can hide information such as a multiple modality in the continuous data. However, it does have advantages. Binning reduces the influence of outliers on the model by converting them to a level of a categorical variable. It can also help with the interpretation of the coefficients of

the final predictive models as it has the effect of scaling variables that are of different magnitudes. It can also increase the number of degrees of freedom of a model.

Smoothing is a technique most often used to reduce the volatility of a time series. As already mentioned, simple multi-period moving averages can be effective for reducing volatility. In this approach, a new time series is created from the old one by taking successive averages of a fixed number of periods. For each new point in the new time series, the oldest observation in the previous average is dropped and the next one in the series is added to the calculation. Another popular method for smoothing a time series is called Loess Regression. This is a nonparametric method that performs least-squares regression on a local neighborhood of the time series. The new time series is predicted within a specified range, or span, and may include other predictor variables. The result is a new time series with a smoothness that increases with the width of the span, although this does not minimize the sum of squared errors of the Loess Regression. This functionality is available in commercial and open-source tools.

Several approaches for transforming variables by fitting functions are mentioned above. Another technique of interest for transforming data is *normalization*. Not to be confused with normalization in a database context, this refers to scaling data to eliminate differences of magnitude between continuous variables that can create numerical issues solving for model coefficients, as well as difficulties in comparing and interpreting the estimated coefficients. Typical methods include the following:

- Min–max: The value is scaled by subtracting the minimum from the value, and dividing by the range (max–min) of the observed values.
- *Z*-score: The value is scaled by subtracting the mean from the value, and dividing by the standard deviation of the observed values.
- Decimal scaling: The value is divided by some power of 10, to adjust the range of the observed values.

Another transformation that can be used to discretize time series data is to count the number of events that occur within a specific time interval. For example, suppose data are collected from a diesel engine. A sensor collects data for the temperature of the engine coolant at regular intervals. The engine has a protection system that causes the engine to be derated (the power reduced) when the coolant temperature exceeds 225°F. The raw time series will be noisy and difficult to use. One approach is to consider an event of interest to occur whenever the temperature exceeds this threshold. A new variable can be created that will count the number of these events that occur within a specified interval. The transformed variable can now be used to examine the relationship between these events, which may be transitory in nature, and the occurrence of other events such as unplanned maintenance. This is a valuable transformation as the collection of events and alarms is quite common in asset monitoring systems.

A final consideration on the topic of data transformation is *data reduction.* With the advent of inexpensive data storage and inexpensive devices that can collect data at high frequencies, it is not uncommon for data warehouses to become quite large. Analyses that run using data sets with terabytes of data can become impractical due to the processing time required. Data reduction seeks to reduce the size of the data warehouse while preserving the information contained in the data. There are many techniques used to perform data reduction. This discussion will focus on two examples.

The first is called *principal components analysis* (PCA). PCA finds new variables, called components, that represent the data in a lower dimensional space. PCA reduces the dimensions by an orthogonal transformation of the data that is achieved through the following process:

- Start with a data matrix of m observations of n variables.
- Subtract the mean of each variable from each observation.
- Calculate the $n \times n$ covariance matrix.
- Calculate the eigenvalues and eigenvectors of the covariance matrix.
- The principal component is the eigenvector with the largest eigenvalue.
- Select some subset of p eigenvectors with the p largest eigenvalues.
- Derive the new data by creating a matrix of p eigenvectors and transposing it. Multiply this by the mean adjusted to complete the transformation.

If the correlation between the original n variables is high, the difference between n and p will be significant and there will be substantial reduction in the size of the data. These new data can be used for model development in a fashion like the original data. However, much of the redundancy and unimportant information is removed by the projection into the lower dimensional space.

Another commonly used technique for data reduction is data sampling. There are a variety of sampling methods that can be used to reduce the number of instances submitted to an algorithm while retaining the original characteristics of the data. Simple random sampling without replacement (SRSWOR) is used to select n records from a set of m records, where $n < m$ and every record has an equal probability of being selected. Simple random sampling with replacement (SRSWR) is similar, except that each record that is selected is replaced and may be selected again on the next draw. If the population from which the sample is not homogeneous, then a stratified sample may be taken. Suppose that a sample of individuals consists of three groups or strata: youth, adults, and seniors. A simple random sample (SRS) may be taken from each stratum to accurately reflect the data of the entire population. One challenge with using SRS methods for data reduction is that while they do reduce the size of the data, which will improve computational performance and memory usage, they also increase the sample variance. This will make it more difficult to detect small differences between groups, and will generally reduce the effectiveness of statistical algorithms. More complex algorithms are available for data reduction, also sometimes called *data squashing.* These methods select n records from a set of m records, where n is much smaller than m, and add an additional column that contains a weight that is representative of the frequency of occurrence of that record in the original population. Numerous references to data squashing methods and their application and effectiveness can be found in the statistical literature.

4.4 Data Modeling

4.4.1 Relational Databases

After the data have been cleaned and transformed, the ETL process will deposit them into a data warehouse. The most common type of data warehouse is built using a *relational database.* The software underlying the structure of relational databases is called a relational database management system (RDBMS). Originally proposed by an IBM researcher named E.F. Codd in 1970, a relational database stores data in tables. Each table consists of rows called records that usually represent one entry of the content of the table. For example, each record can contain information about a customer, an asset, or a purchase order. Each record consists of columns or fields that contain data related to that instance. So, a customer record might contain account number, first name, last name, phone number, and e-mail address.

Figure 4.5 illustrates the critical concept in a relational database. Each table will have a *primary key* that serves as a unique identifier for that record in the table. In this example, Asset Type, Asset ID, Work Order Number, and Task Code all serve as a primary key in a table. When they appear in other tables, they are called *foreign keys.* The relationships between the tables are highlighted by the connections. For example, in the Assets table, Asset ID is the primary key as each asset will have a unique Asset ID. In the Work Orders table, Asset ID is a foreign key and the relationship between the two keys is said to be *one to many.* The Asset ID may appear many times in the Work Orders table as the asset may have been serviced many times. However, the information uniquely related to the asset appears just once in the Asset table. This allows the relational database to store the data in a more compact form, and master data such as we see here regarding Assets, Asset Types, and Tasks needs to be retrieved only if we have a need to associate it with transactional data such as we see in the Work Orders table.

Essentially all relational databases use structured query language (SQL) to write queries and maintain the database. A query allows the user to extract data from several different tables to create a new record format specific to a required