WHAT WE CAN LEARN FROM THIS VIGNETTE

Striving to thrive in a fast-changing competitive industry, SiriusXM realized the need for a new and improved marketing infrastructure (one that relies on data and analytics) to effectively communicate its value proposition to its existing and potential customers. As is the case in any industry, success or mere survival in entertainment depends on intelligently sensing the changing trends (likes and dislikes) and putting together the right messages and policies to win new customers while retaining the existing ones. The key is to create and manage successful marketing campaigns that resonate with the target population of customers and have a close feedback loop to adjust and modify the message to optimize the outcome. At the end, it was all about the precision in the way that SiriusXM conducted business: being proactive about the changing nature of the clientele and creating and transmitting the right products and services in a timely manner using a fact-based/data-driven holistic marketing strategy. Source identification, source creation, access and collection, integration, cleaning, transformation, storage, and processing of relevant data played a critical role in SiriusXM's success in designing and implementing a marketing analytics strategy as is the case in any analytically savvy successful company today, regardless of the industry in which they are participating.

Sources: C. Quinn, "Data-Driven Marketing at SiriusXM," Teradata Articles & News, 2016. **http://bigdata.teradata.com/US/Articles-News/Data-Driven-Marketing-At-SiriusXM/** (accessed August 2016); "SiriusXM Attracts and Engages a New Generation of Radio Consumers." **http://assets.teradata.com/resourceCenter/downloads/CaseStudies/EB8597.pdf?processed=1**.

3.2 NATURE OF DATA

Data are the main ingredient for any BI, data science, and business analytics initiative. In fact, they can be viewed as the raw material for what popular decision technologies produce—information, insight, and **knowledge**. Without data, none of these technologies could exist and be popularized—although traditionally we have built analytics models using expert knowledge and experience coupled with very little or no data at all; however, those were the old days, and now data are of the essence. Once perceived as a big challenge to collect, store, and manage, data today are widely considered among the most valuable assets of an organization with the potential to create invaluable insight to better understand customers, competitors, and the business processes.

Data can be small or very large. They can be structured (nicely organized for computers to process), or they can be unstructured (e.g., text that is created for humans and hence not readily understandable/consumable by computers). Data can come in small batches continuously or can pour in all at once as a large batch. These are some of the characteristics that define the inherent nature of today's data, which we often call Big Data. Even though these characteristics of data make them more challenging to process and consume, they also make the data more valuable because the characteristics enrich them beyond their conventional limits, allowing for the discovery of new and novel knowledge. Traditional ways to manually collect data (via either surveys or human-entered business transactions) mostly left their places to modern-day data collection mechanisms that use Internet and/or sensor/radio frequency identification (RFID)–based computerized networks. These automated data collection systems are not only enabling us to collect more volumes of data but also enhancing the **data quality** and integrity. Figure 3.1 illustrates a typical analytics continuum—data to analytics to actionable information.

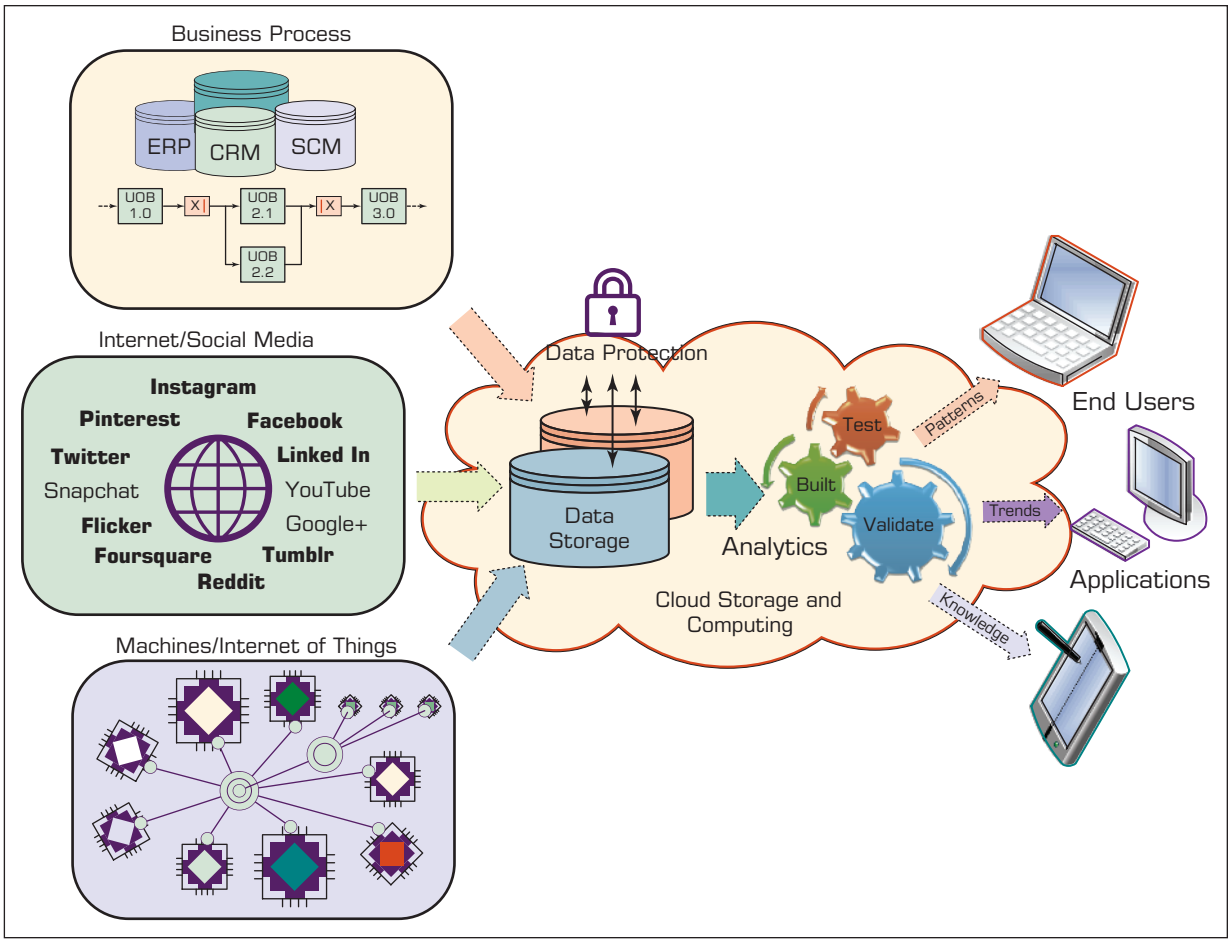


FIGURE 3.1 **A Data to Knowledge Continuum.**

Although their value proposition is undeniable, to live up their promise, data must comply with some basic usability and quality metrics. Not all data are useful for all tasks, obviously. That is, data must match with (have the coverage of the specifics for) the task for which they are intended to be used. Even for a specific task, the relevant data on hand need to comply with the quality and quantity requirements. Essentially, data have to be analytics ready. So what does it mean to make data analytics ready? In addition to its relevancy to the problem at hand and the quality/quantity requirements, it also has to have a certain structure in place with key fields/variables with properly normalized values. Furthermore, there must be an organization-wide agreed-on definition for common variables and subject matters (sometimes also called *master data management*), such as how to define a customer (what characteristics of customers are used to produce a holistic enough representation to analytics) and where in the business process the customer-related information is captured, validated, stored, and updated.

Sometimes the representation of the data depends on the type of analytics being employed. Predictive algorithms generally require a flat file with a target variable, so making data **analytics ready** for prediction means that data sets must be transformed into a flat-file format and made ready for ingestion into those predictive algorithms. It is also imperative to match the data to the needs and wants of a specific predictive algorithm and/or a software tool. For instance, neural network algorithms require all input variables

to be numerically represented (even the nominal variables need to be converted into pseudo binary numeric variables), whereas decision tree algorithms do not require such numerical transformation—they can easily and natively handle a mix of nominal and numeric variables.

Analytics projects that overlook data-related tasks (some of the most critical steps) often end up with the wrong answer for the right problem, and these unintentionally created, seemingly good answers could lead to inaccurate and untimely decisions. Following are some of the characteristics (metrics) that define the readiness level of data for an analytics study (Delen, 2015; Kock, McQueen, & Corner, 1997).

- **Data source reliability.** This term refers to the originality and appropriateness of the storage medium where the data are obtained—answering the question of "Do we have the right confidence and belief in this data source?" If at all possible, one should always look for the original source/creator of the data to eliminate/mitigate the possibilities of data misrepresentation and data transformation caused by the mishandling of the data as they moved from the source to destination through one or more steps and stops along the way. Every move of the data creates a chance to unintentionally drop or reformat data items, which limits the integrity and perhaps true accuracy of the data set.
- **Data content accuracy.** This means that data are correct and are a good match for the analytics problem—answering the question of "Do we have the right data for the job?" The data should represent what was intended or defined by the original source of the data. For example, the customer's contact information recorded within a database should be the same as what the customer said it was. Data accuracy will be covered in more detail in the following subsection.
- **Data accessibility.** This term means that the data are easily and readily obtainable—answering the question of "Can we easily get to the data when we need to?" Access to data can be tricky, especially if they are stored in more than one location and storage medium and need to be merged/transformed while accessing and obtaining them. As the traditional relational database management systems leave their place (or coexist with a new generation of data storage mediums such as data lakes and Hadoop infrastructure), the importance/criticality of data accessibility is also increasing.
- **Data security and data privacy. Data security** means that the data are secured to allow only those people who have the authority and the need to access them and to prevent anyone else from reaching them. Increasing popularity in educational degrees and certificate programs for Information Assurance is evidence of the criticality and the increasing urgency of this data quality metric. Any organization that maintains health records for individual patients must have systems in place that not only safeguard the data from unauthorized access (which is mandated by federal laws such as the Health Insurance Portability and Accountability Act [HIPAA]) but also accurately identify each patient to allow proper and timely access to records by authorized users (Annas, 2003).
- **Data richness.** This means that all required data elements are included in the data set. In essence, richness (or comprehensiveness) means that the available variables portray a rich enough dimensionality of the underlying subject matter for an accurate and worthy analytics study. It also means that the information content is complete (or near complete) to build a predictive and/or prescriptive analytics model.
- **Data consistency.** This means that the data are accurately collected and combined/merged. Consistent data represent the dimensional information (variables of interest) coming from potentially disparate sources but pertaining to the same subject. If the data integration/merging is not done properly, some of the variables of different subjects could appear in the same record—having two different patient

records mixed up; for instance, this could happen while merging the demographic and clinical test result data records.

- **Data currency/data timeliness.** This means that the data should be up-to-date (or as recent/new as they need to be) for a given analytics model. It also means that the data are recorded at or near the time of the event or observation so that the time delay–related misrepresentation (incorrectly remembering and encoding) of the data is prevented. Because accurate analytics relies on accurate and timely data, an essential characteristic of analytics-ready data is the timeliness of the creation and access to data elements.
- **Data granularity.** This requires that the variables and data values be defined at the lowest (or as low as required) level of detail for the intended use of the data. If the data are aggregated, they might not contain the level of detail needed for an analytics algorithm to learn how to discern different records/cases from one another. For example, in a medical setting, numerical values for laboratory results should be recorded to the appropriate decimal place as required for the meaningful interpretation of test results and proper use of those values within an analytics algorithm. Similarly, in the collection of demographic data, data elements should be defined at a granular level to determine the differences in outcomes of care among various subpopulations. One thing to remember is that the data that are aggregated cannot be disaggregated (without access to the original source), but they can easily be aggregated from its granular representation.
- **Data validity.** This is the term used to describe a match/mismatch between the actual and expected data values of a given variable. As part of data definition, the acceptable values or value ranges for each data element must be defined. For example, a valid data definition related to gender would include three values: male, female, and unknown.
- **Data relevancy.** This means that the variables in the data set are all relevant to the study being conducted. Relevancy is not a dichotomous measure (whether a variable is relevant or not); rather, it has a spectrum of relevancy from least relevant to most relevant. Based on the analytics algorithms being used, one can choose to include only the most relevant information (i.e., variables) or, if the algorithm is capable enough to sort them out, can choose to include all the relevant ones regardless of their levels. One thing that analytics studies should avoid is including totally irrelevant data into the model building because this could contaminate the information for the algorithm, resulting in inaccurate and misleading results.

The above-listed characteristics are perhaps the most prevailing metrics to keep up with; the true data quality and excellent analytics readiness for a specific application domain would require different levels of emphasis to be placed on these metric dimensions and perhaps add more specific ones to this collection. The following section will delve into the nature of data from a taxonomical perspective to list and define different data types as they relate to different analytics projects.

SECTION 3.2 REVIEW QUESTIONS

1. How do you describe the importance of data in analytics? Can we think of analytics without data?
2. Considering the new and broad definition of business analytics, what are the main inputs and outputs to the analytics continuum?
3. Where do the data for business analytics come from?
4. In your opinion, what are the top three data-related challenges for better analytics?
5. What are the most common metrics that make for analytics-ready data?

3.3 SIMPLE TAXONOMY OF DATA

The term *data* (**datum** in singular form) refers to a collection of facts usually obtained as the result of experiments, observations, transactions, or experiences. Data can consist of numbers, letters, words, images, voice recordings, and so on, as measurements of a set of variables (characteristics of the subject or event that we are interested in studying). Data are often viewed as the lowest level of abstraction from which information and then knowledge is derived.

At the highest level of abstraction, one can classify data as structured and unstructured (or semistructured). **Unstructured data**/semistructured data are composed of any combination of textual, imagery, voice, and Web content. Unstructured/semistructured data will be covered in more detail in the text mining and Web mining chapter. **Structured data** are what data mining algorithms use and can be classified as categorical or numeric. The **categorical data** can be subdivided into nominal or **ordinal data**, whereas numeric data can be subdivided into intervals or ratios. Figure 3.2 shows a simple **data taxonomy**.

- **Categorical data.** These represent the labels of multiple classes used to divide a variable into specific groups. Examples of categorical variables include race, sex, age group, and educational level. Although the latter two variables can also be considered in a numerical manner by using exact values for age and highest grade completed, for example, it is often more informative to categorize such variables into a relatively small number of ordered classes. The categorical data can also be called *discrete data*, implying that they represent a finite number of values with no continuum between them. Even if the values used for the categorical (or discrete) variables are numeric, these numbers are nothing more than symbols and do not imply the possibility of calculating fractional values.
- **Nominal data.** These contain measurements of simple codes assigned to objects as labels, which are not measurements. For example, the variable *marital status* can be generally categorized as (1) single, (2) married, and (3) divorced. **Nominal**

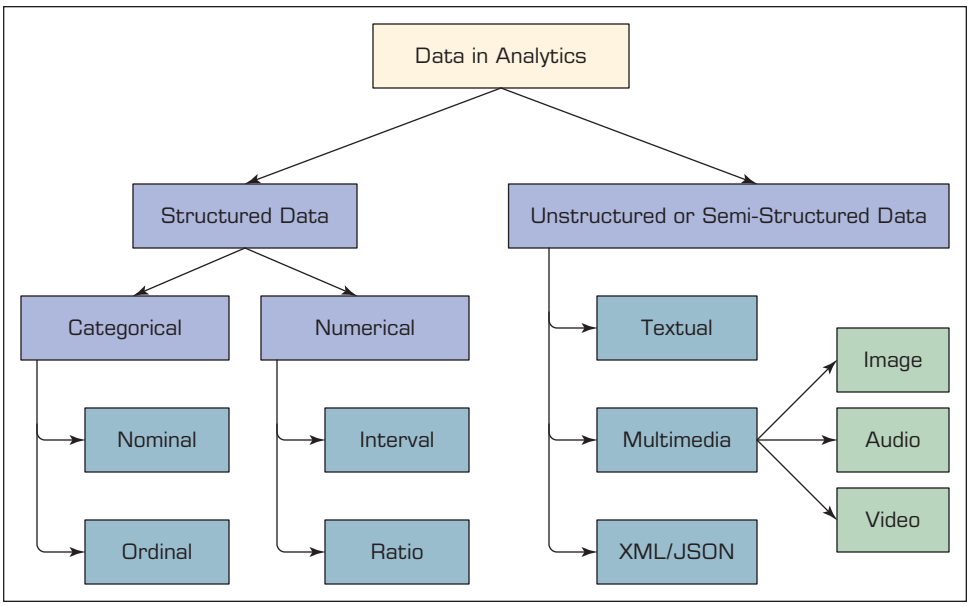


FIGURE 3.2 **A Simple Taxonomy of Data.**

data can be represented with binomial values having two possible values (e.g., yes/no, true/false, good/bad) or multinomial values having three or more possible values (e.g., brown/green/blue, white/black/Latino/Asian, single/married/divorced).

- **Ordinal data.** These contain codes assigned to objects or events as labels that also represent the rank order among them. For example, the variable *credit score* can be generally categorized as (1) low, (2) medium, or (3) high. Similar ordered relationships can be seen in variables such as age group (i.e., child, young, middle-aged, elderly) and educational level (i.e., high school, college, graduate school). Some predictive analytic algorithms, such as *ordinal multiple logistic regression*, take into account this additional rank-order information to build a better classification model.
- **Numeric data.** These represent the numeric values of specific variables. Examples of numerically valued variables include age, number of children, total household income (in U.S. dollars), travel distance (in miles), and temperature (in Fahrenheit degrees). Numeric values representing a variable can be integers (only whole numbers) or real (also fractional numbers). The numeric data can also be called *continuous data*, implying that the variable contains continuous measures on a specific scale that allows insertion of interim values. Unlike a discrete variable, which represents finite, countable data, a continuous variable represents scalable measurements, and it is possible for the data to contain an infinite number of fractional values.
- **Interval data.** These are variables that can be measured on interval scales. A common example of interval scale measurement is temperature on the Celsius scale. In this particular scale, the unit of measurement is 1/100 of the difference between the melting temperature and the boiling temperature of water in atmospheric pressure; that is, there is not an absolute zero value.
- **Ratio data.** These include measurement variables commonly found in the physical sciences and engineering. Mass, length, time, plane angle, energy, and electric charge are examples of physical measures that are ratio scales. The scale type takes its name from the fact that measurement is the estimation of the ratio between a magnitude of a continuous quantity and a unit magnitude of the same kind. Informally, the distinguishing feature of a ratio scale is the possession of a nonarbitrary zero value. For example, the Kelvin temperature scale has a nonarbitrary zero point of absolute zero, which is equal to –273.15 degrees Celsius. This zero point is nonarbitrary because the particles that comprise matter at this temperature have zero kinetic energy.

Other data types, including textual, spatial, imagery, video, and voice, need to be converted into some form of categorical or numeric representation before they can be processed by analytics methods (data mining algorithms; Delen, 2015). Data can also be classified as static or dynamic (i.e., temporal or time series).

Some predictive analytics (i.e., data mining) methods and machine-learning algorithms are very selective about the type of data that they can handle. Providing them with incompatible data types can lead to incorrect models or (more often) halt the model development process. For example, some data mining methods need all the variables (both input and output) represented as numerically valued variables (e.g., neural networks, support vector machines, logistic regression). The nominal or ordinal variables are converted into numeric representations using some type of *1-of-N* pseudo variables (e.g., a categorical variable with three unique values can be transformed into three pseudo variables with binary values—1 or 0). Because this process could increase the number of variables, one should be cautious about the effect of such representations, especially for the categorical variables that have large numbers of unique values.

Similarly, some predictive analytics methods, such as ID3 (a classic decision tree algorithm) and rough sets (a relatively new rule induction algorithm), need all the variables represented as categorically valued variables. Early versions of these methods required the user to discretize numeric variables into categorical representations before they could be processed by the algorithm. The good news is that most implementations of these algorithms in widely available software tools accept a mix of numeric and nominal variables and internally make the necessary conversions before processing the data.

Data come in many different variable types and representation schemas. Business analytics tools are continuously improving in their ability to help data scientists in the daunting task of data transformation and data representation so that the data requirements of specific predictive models and algorithms can be properly executed. Application Case 3.1 illustrates a business scenario in which one of the largest telecommunication companies streamlined and used a wide variety of rich data sources to generate customers insight to prevent churn and to create new revenue sources.

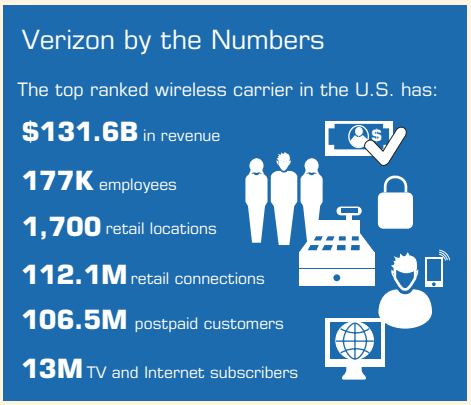
Application Case 3.1

Verizon Answers the Call for Innovation: The Nation's Largest Network Provider Uses Advanced Analytics to Bring the Future to Its Customers

The Problem

In the ultra-competitive telecommunications industry, staying relevant to consumers while finding new sources of revenue is critical, especially since current revenue sources are in decline.

For Fortune 13 powerhouse Verizon, the secret weapon that catapulted the company into the nation's largest and most reliable network provider is also guiding the business toward future success (see the following figure for some numbers about Verizon). The secret weapon? Data and analytics. Because telecommunication companies are typically rich in data, having the right analytics solution and personnel in place can uncover critical insights that benefit every area of the business.



The Backbone of the Company

Since its inception in 2000, Verizon has partnered with Teradata to create a data and analytics architecture that drives innovation and science-based decision making. The goal is to stay relevant to customers while also identifying new business opportunities and making adjustments that result in more cost-effective operations.

"With business intelligence, we help the business identify new business opportunities or make course corrections to operate the business in a more cost-effective way," said Grace Hwang, executive director of Financial Performance & Analytics, BI, for Verizon. "We support decision makers with the most relevant information to improve the competitive advantage of Verizon."

By leveraging data and analytics, Verizon is able to offer a reliable network, ensure customer satisfaction, and develop products and services that consumers want to buy.

"Our incubator of new products and services will help bring the future to our customers," Hwang said. "We're using our network to make breakthroughs in

(*Continued*)

Application Case 3.1 (Continued)

interactive entertainment, digital media, the Internet of Things, and broadband services."

Data Insights across Three Business Units

Verizon relies on advanced analytics that are executed on the Teradata® Unified Data Architecture™ to support its business units. The analytics enable Verizon to deliver on its promise to help customers innovate their lifestyles and provide key insights to support these three areas:

- Identify new revenue sources. Research and development teams use data, analytics, and strategic partnerships to test and develop with the Internet of Things (IoT). The new frontier in data is IoT, which will lead to new revenues that in turn generate opportunities for top-line growth. Smart cars, smart agriculture, and smart IoT will all be part of this new growth.
- Predict churn in the core mobile business. Verizon has multiple use cases that demonstrate how its advanced analytics enable laser-accurate churn prediction—within a one to two percent margin—in the mobile space. For a $131 billion company, predicting churn with such precision is significant. By recognizing specific patterns in tablet data usage, Verizon can identify which customers most often access their tablets, then engage those who do not.
- Forecast mobile phone plans. Customer behavioral analytics allow finance to better predict earnings in fast-changing market conditions. The U.S. wireless industry is moving from monthly payments for both the phone and the service to paying for the phone independently. This opens up a new opportunity for Verizon to gain business. The analytic environment helps Verizon better predict churn with new plans and forecast the impact of changes to pricing plans.

The analytics deliver what Verizon refers to as "honest data" that inform various business units. "Our mission is to be the honest voice and the independent third-party opinion on the success or opportunities for improvement to the business," Hwang explains. "So my unit is viewed as the golden source of information, and we come across with the honest voice, and a lot of the business decisions are through various rungs of course correction."

Hwang adds that oftentimes, what forces a company to react is competitors affecting change in the marketplace, rather than the company making the wrong decisions. "So we try to guide the business through the best course of correction, wherever applicable, timely, so that we can continue to deliver record-breaking results year after year," she said. "I have no doubt that the business intelligence had led to such success in the past."

Disrupt and Innovate

Verizon leverages advanced analytics to optimize marketing by sending the most relevant offers to customers. At the same time, the company relies on analytics to ensure they have the financial acumen to stay number one in the U.S. mobile market. By continuing to disrupt the industry with innovative products and solutions, Verizon is positioned to remain the wireless standard for the industry.

"We need the marketing vision and the sales rigor to produce the most relevant offer to our customers, and then at the same time we need to have the finance rigor to ensure that whatever we offer to the customer is also profitable to the business so that we're responsible to our shareholders," Hwang says.

In Summary—Executing the Seven Ps of Modern Marketing

Telecommunications giant Verizon uses seven Ps to drive its modern-day marketing efforts. The Ps, when used in unison, help Verizon penetrate the market in the way it predicted.

1. **People**: Understanding customers and their needs to create the product.
2. **Place**: Where customers shop.
3. **Product**: The item that's been manufactured and is for sale.

4. **Process**: How customers get to the shop or place to buy the product.
5. **Pricing**: Working with promotions to get customers' attention.
6. **Promo**: Working with pricing to get customers' attention.
7. **Physical evidence**: The business intelligence that gives insights.

"The Aster and Hadoop environment allows us to explore things we suspect could be the reasons for breakdown in the seven Ps," says Grace Hwang, executive director of Financial Performance & Analytics, BI, for Verizon. "This goes back to providing the business value to our decision-makers. With each step in the seven Ps, we ought to be able to tell them where there are opportunities for improvement."

Questions for Case 3.1

1. What was the challenge Verizon was facing?
2. What was the data-driven solution proposed for Verizon's business units?
3. What were the results?

Source: Teradata Case Study "Verizon Answers the Call for Innovation" **https://www.teradata.com/Resources/Case-Studies/Verizon-answers-the-call-for-innovation** (accessed July 2018).

SECTION 3.3 REVIEW QUESTIONS

1. What are data? How do data differ from information and knowledge?
2. What are the main categories of data? What types of data can we use for BI and analytics?
3. Can we use the same data representation for all analytics models? Why, or why not?
4. What is a *1-of-N* data representation? Why and where is it used in analytics?

3.4 ART AND SCIENCE OF DATA PREPROCESSING

Data in their original form (i.e., the real-world data) are not usually ready to be used in analytics tasks. They are often dirty, misaligned, overly complex, and inaccurate. A tedious and time-demanding process (so-called **data preprocessing**) is necessary to convert the raw real-world data into a well-refined form for analytics algorithms (Kotsiantis, Kanellopoulos, & Pintelas, 2006). Many analytics professionals would testify that the time spent on data preprocessing (which is perhaps the least enjoyable phase in the whole process) is significantly longer than the time spent on the rest of the analytics tasks (the fun of analytics model building and assessment). Figure 3.3 shows the main steps in the data preprocessing endeavor.

In the first step of data preprocessing, the relevant data are collected from the identified sources, the necessary records and variables are selected (based on an intimate understanding of the data, the unnecessary information is filtered out), and the records coming from multiple data sources are integrated/merged (again, using the intimate understanding of the data, the synonyms and homonyms are able to be handled properly).

In the second step of data preprocessing, the data are cleaned (this step is also known as *data scrubbing*). Data in their original/raw/real-world form are usually dirty (Hernández & Stolfo, 1998; Kim et al., 2003). In this phase, the values in the data set are identified and dealt with. In some cases, missing values are an anomaly in the data set, in which case they need to be imputed (filled with a most probable value) or ignored; in other cases, the missing values are a natural part of the data set

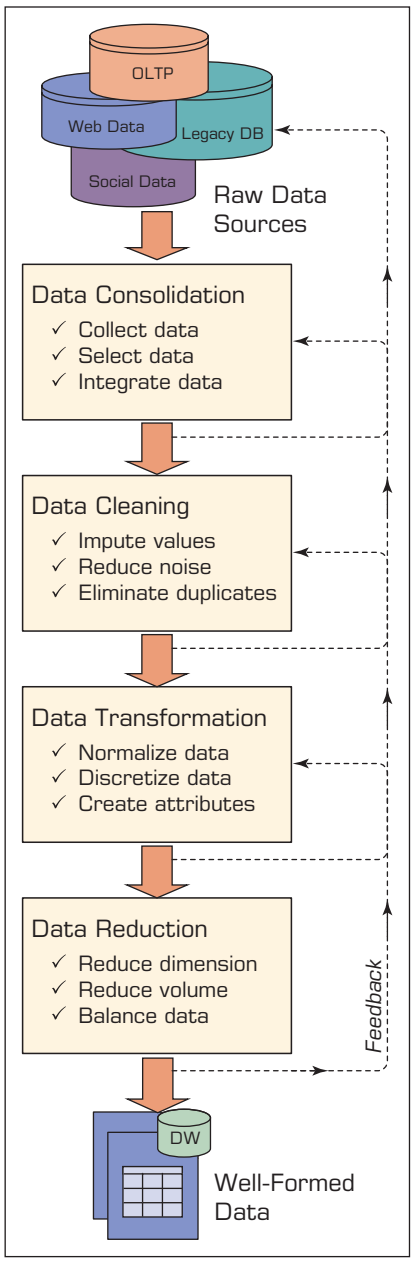


FIGURE 3.3 **Data Preprocessing Steps.**

(e.g., the *household income* field is often left unanswered by people who are in the top income tier). In this step, the analyst should also identify noisy values in the data (i.e., the outliers) and smooth them out. In addition, inconsistencies (unusual values within a variable) in the data should be handled using domain knowledge and/or expert opinion.

In the third step of data preprocessing, the data are transformed for better processing. For instance, in many cases, the data are normalized between a certain minimum and maximum for all variables to mitigate the potential bias of one variable having

large numeric values (such as household income) dominating other variables (such as *number of dependents* or *years in service,* which could be more important) having smaller values. Another transformation that takes place is discretization and/or aggregation. In some cases, the numeric variables are converted to categorical values (e.g., low, medium, high); in other cases, a nominal variable's unique value range is reduced to a smaller set using concept hierarchies (e.g., as opposed to using the individual states with 50 different values, one could choose to use several regions for a variable that shows location) to have a data set that is more amenable to computer processing. Still, in other cases, one might choose to create new variables based on the existing ones to magnify the information found in a collection of variables in the data set. For instance, in an organ transplantation data set, one might choose to use a single variable showing the blood-type match (1: match, 0: no match) as opposed to separate multinominal values for the blood type of both the donor and the recipient. Such simplification could increase the information content while reducing the complexity of the relationships in the data.

The final phase of data preprocessing is data reduction. Even though data scientists (i.e., analytics professionals) like to have large data sets, too much data can also be a problem. In the simplest sense, one can visualize the data commonly used in predictive analytics projects as a flat file consisting of two dimensions: variables (the number of columns) and cases/records (the number of rows). In some cases (e.g., image processing and genome projects with complex microarray data), the number of variables can be rather large, and the analyst must reduce the number to a manageable size. Because the variables are treated as different dimensions that describe the phenomenon from different perspectives, in predictive analytics and data mining, this process is commonly called **dimensional reduction** (or **variable selection**). Even though there is not a single best way to accomplish this task, one can use the findings from previously published literature; consult domain experts; run appropriate statistical tests (e.g., principal component analysis or independent component analysis); and, more preferably, use a combination of these techniques to successfully reduce the dimensions in the data into a more manageable and most relevant subset.

With respect to the other dimension (i.e., the number of cases), some data sets can include millions or billions of records. Even though computing power is increasing exponentially, processing such a large number of records cannot be practical or feasible. In such cases, one might need to sample a subset of the data for analysis. The underlying assumption of sampling is that the subset of the data will contain all relevant patterns of the complete data set. In a homogeneous data set, such an assumption could hold well, but real-world data are hardly ever homogeneous. The analyst should be extremely careful in selecting a subset of the data that reflects the essence of the complete data set and is not specific to a subgroup or subcategory. The data are usually sorted on some variable, and taking a section of the data from the top or bottom could lead to a biased data set on specific values of the indexed variable; therefore, always try to randomly select the records on the sample set. For skewed data, straightforward random sampling might not be sufficient, and stratified sampling (a proportional representation of different subgroups in the data is represented in the sample data set) might be required. Speaking of skewed data, it is a good practice to balance the highly skewed data by either oversampling the less represented or undersampling the more represented classes. Research has shown that balanced data sets tend to produce better prediction models than unbalanced ones (Thammasiri et al., 2014).

The essence of data preprocessing is summarized in Table 3.1, which maps the main phases (along with their problem descriptions) to a representative list of tasks and algorithms.

TABLE 3.1 A Summary of Data Preprocessing Tasks and Potential Methods

Main Task	Subtasks	Popular Methods
Data consolidation	Access and collect the data	SQL queries, software agents, Web services.
	Select and filter the data	Domain expertise, SQL queries, statistical tests.
	Integrate and unify the data	SQL queries, domain expertise, ontology-driven data mapping.
Data cleaning	Handle missing values in the data	Fill in missing values (imputations) with most appropriate values (mean, median, min/max, mode, etc.); recode the missing values with a constant such as "ML"; remove the record of the missing value; do nothing.
	Identify and reduce noise in the data	Identify the outliers in data with simple statistical techniques (such as averages and standard deviations) or with cluster analysis; once identified, either remove the outliers or smooth them by using binning, regression, or simple averages.
	Find and eliminate erroneous data	Identify the erroneous values in data (other than outliers), such as odd values, inconsistent class labels, odd distributions; once identified, use domain expertise to correct the values or remove the records holding the erroneous values.
Data transformation	Normalize the data	Reduce the range of values in each numerically valued variable to a standard range (e.g., 0 to 1 or -1 to $+1$) by using a variety of normalization or scaling techniques.
	Discretize or aggregate the data	If needed, convert the numeric variables into discrete representations using range- or frequency-based binning techniques; for categorical variables, reduce the number of values by applying proper concept hierarchies.
	Construct new attributes	Derive new and more informative variables from the existing ones using a wide range of mathematical functions (as simple as addition and multiplication or as complex as a hybrid combination of log transformations).
Data reduction	Reduce number of attributes	Use principal component analysis, independent component analysis, chi-square testing, correlation analysis, and decision tree induction.
	Reduce number of records	Perform random sampling, stratified sampling, expert-knowledge-driven purposeful sampling.
	Balance skewed data	Oversample the less represented or undersample the more represented classes.

It is almost impossible to underestimate the value proposition of data preprocessing. It is one of those time-demanding activities in which investment of time and effort pays off without a perceivable limit for diminishing returns. That is, the more resources you invest in it, the more you will gain at the end. Application Case 3.2 illustrates an interesting study that used raw, readily available academic data within an educational organization to develop predictive models to better understand attrition and improve freshman student retention in a large higher education institution. As the application case clearly states, each and every data preprocessing task described in Table 3.1 was critical to a successful execution of the underlying analytics project, especially the task that related to the balancing of the data set.

Application Case 3.2 Improving Student Retention with Data-Driven Analytics

Student attrition has become one of the most challenging problems for decision makers in academic institutions. Despite all the programs and services that are put in place to help retain students, according to the U.S. Department of Education's Center for Educational Statistics (**nces.ed.gov**), only about half of those who enter higher education actually earn a bachelor's degree. Enrollment management and the retention of students have become a top priority for administrators of colleges and universities in the United States and other countries around the world. High dropout of students usually results in overall financial loss, lower graduation rates, and an inferior school reputation in the eyes of all stakeholders. The legislators and policy makers who oversee higher education and allocate funds, the parents who pay for their children's education to prepare them for a better future, and the students who make college choices look for evidence of institutional quality and reputation to guide their decision-making processes.

The Proposed Solution

To improve student retention, one should try to understand the nontrivial reasons behind the attrition. To be successful, one should also be able to accurately identify those students who are at risk of dropping out. So far, the vast majority of student attrition research has been devoted to understanding this complex, yet crucial, social phenomenon. Even though these qualitative, behavioral, and survey-based studies revealed invaluable insight by developing and testing a wide range of theories, they do not provide the much-needed instruments to accurately predict (and potentially improve) student attrition. The project summarized in this case study proposed a quantitative research approach in which the historical institutional data from student databases could be used to develop models that are capable of predicting as well as explaining the institution-specific nature of the attrition problem. The proposed analytics approach is shown in Figure 3.4.

Although the concept is relatively new to higher education, for more than a decade now, similar problems in the field of marketing management have been studied using predictive data analytics techniques under the name of "churn analysis" where the purpose has been to identify a sample among current customers to answer the question, "Who among our current customers are more likely to stop buying our products or services?" so that some kind of mediation or intervention process can be executed to retain them. Retaining existing customers is crucial because, as we all know and as the related research has shown time and time again, acquiring a new customer costs on an order of magnitude more effort, time, and money than trying to keep the one that you already have.

Data Are of the Essence

The data for this research project came from a single institution (a comprehensive public university located in the Midwest region of the United States) with an average enrollment of 23,000 students, of which roughly 80 percent are the residents of the same state and roughly 19 percent of the students are listed under some minority classification. There is no significant difference between the two genders in the enrollment numbers. The average freshman student retention rate for the institution was about 80 percent, and the average six-year graduation rate was about 60 percent.

The study used five years of institutional data, which entailed 16,000+ students enrolled as freshmen, consolidated from various and diverse university student databases. The data contained variables related to students' academic, financial, and demographic characteristics. After merging and converting the multidimensional student data into a single flat file (a file with columns representing the variables and rows representing the student records), the resultant file was assessed and preprocessed to identify and remedy anomalies and unusable values. As an example, the study removed all international student records from the data set because they did not contain information about some of the most reputed predictors (e.g., high school GPA, SAT scores). In the data transformation phase, some of the variables were aggregated (e.g., "Major" and "Concentration" variables aggregated to binary variables MajorDeclared and ConcentrationSpecified)

(*Continued*)

Application Case 3.2 (Continued)

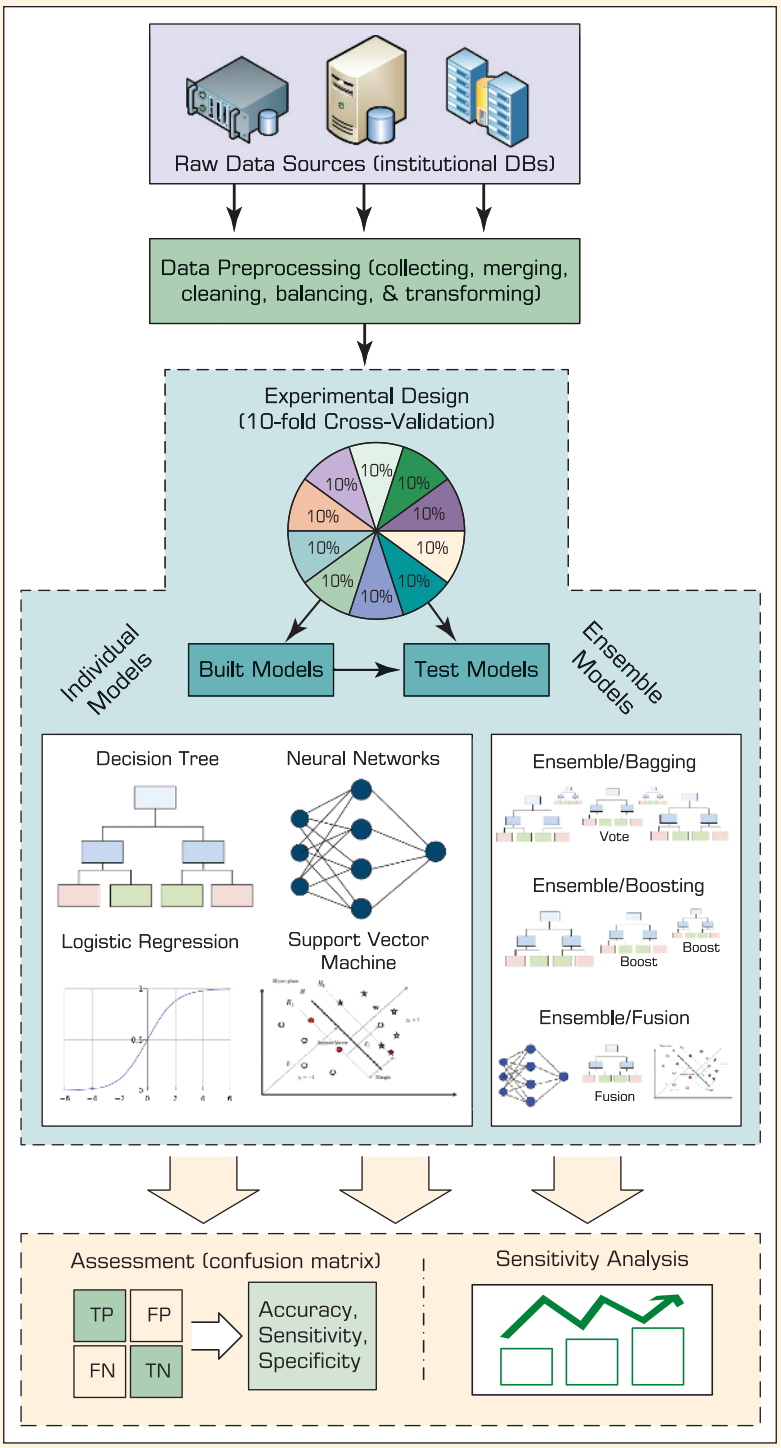


FIGURE 3.4 An Analytics Approach to Predicting Student Attrition.

for better interpretation for the predictive modeling. In addition, some of the variables were used to derive new variables (e.g., Earned/Registered ratio and YearsAfterHighSchool).

$$\text{Earned/Registered} = \frac{\text{EarnedHours}}{\text{RegisteredHours}}$$

$$\text{YearsAfterHighSchool} = \text{FreshmenEnrollmentYear} - \text{HighSchoolGraduationYear}$$

The *Earned/Registered* ratio was created to have a better representation of the students' resiliency and determination in their first semester of the freshman year. Intuitively, one would expect greater values for this variable to have a positive impact on retention/persistence. The *YearsAfterHighSchool* was created to measure the impact of the time taken between high school graduation and initial college enrollment. Intuitively, one would expect this variable to be a contributor to the prediction of attrition. These aggregations and derived variables are determined based on a number of experiments conducted for a number of logical hypotheses. The ones that made more common sense and the ones that led to better prediction accuracy were kept in the final variable set. Reflecting the true nature of the subpopulation (i.e., the freshmen students), the dependent variable (i.e., "Second Fall Registered") contained many more *yes* records (~80 percent) than *no* records (~20 percent; see Figure 3.5).

Research shows that having such imbalanced data has a negative impact on model performance. Therefore, the study experimented with the options of using and comparing the results of the same type of models built with the original imbalanced data (biased for the *yes* records) and the well-balanced data.

Modeling and Assessment

The study employed four popular classification methods (i.e., artificial neural networks, decision trees, support vector machines, and logistic regression) along with three model ensemble techniques (i.e., bagging, busting, and information fusion). The results obtained from all model types were then compared to each other using regular classification model assessment methods (e.g., overall predictive accuracy, sensitivity, specificity) on the holdout samples.

In machine-learning algorithms (some of which will be covered in Chapter 4), sensitivity analysis is a method for identifying the "cause-and-effect" relationship between the inputs and outputs of a given prediction model. The fundamental idea behind sensitivity analysis is that it measures the importance of predictor variables based on the change in modeling performance that occurs if a predictor variable is not included in the model. This modeling and experimentation practice is also called a leave-one-out assessment. Hence, the measure of sensitivity of a specific predictor variable is the ratio of the error of the trained model without the predictor variable to the error of the model that includes this predictor variable. The more sensitive

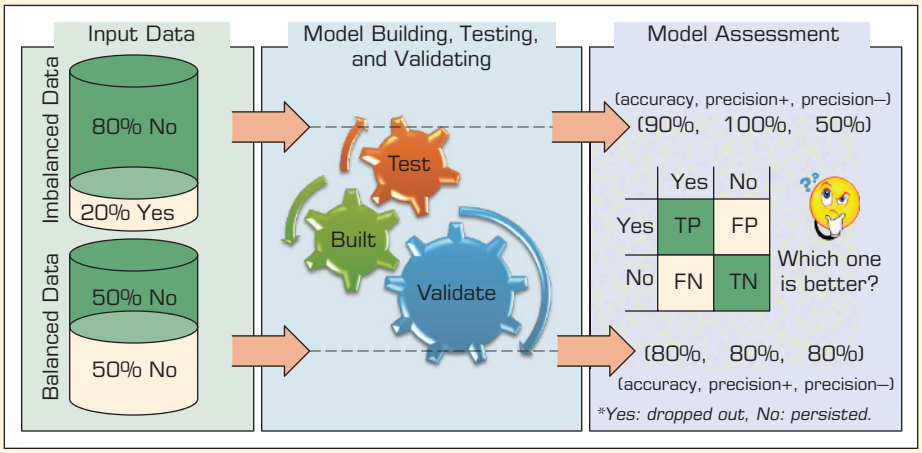


FIGURE 3.5 **A Graphical Depiction of the Class Imbalance Problem.**

(*Continued*)

Application Case 3.2 (Continued)

the network is to a particular variable, the greater the performance decrease would be in the absence of that variable and therefore the greater the ratio of importance. In addition to the predictive power of the models, the study also conducted sensitivity analyses to determine the relative importance of the input variables.

The Results

In the first set of experiments, the study used the original imbalanced data set. Based on the 10-fold cross-validation assessment results, the support vector machines produced the best accuracy with an overall prediction rate of 87.23 percent, and the decision tree was the runner-up with an overall prediction rate of 87.16 percent, followed by artificial neural networks and logistic regression with overall prediction rates of 86.45 percent and 86.12 percent, respectively (see Table 3.2). A careful examination of these results reveals that the prediction accuracy for the "Yes" class is significantly higher than the prediction accuracy of the "No" class. In fact, all four model types predicted the students who are likely to return for the second year with better than 90 percent accuracy, but the types did poorly on predicting the students who are likely to drop out after the freshman year with less than 50 percent accuracy. Because the prediction of the "No" class is the main purpose of this study, less than 50 percent accuracy for this class was deemed not acceptable. Such a difference in prediction accuracy of the two classes can (and should) be attributed to the imbalanced nature of the training data set (i.e., ~80 percent "Yes" and ~20 percent "No" samples).

The next round of experiments used a well-balanced data set in which the two classes are represented nearly equally in counts. In realizing this approach, the study took all samples from the minority class (i.e., the "No" class herein), randomly selected an equal number of samples from the majority class (i.e., the "Yes" class herein), and repeated this process 10 times to reduce potential bias of random sampling. Each of these sampling processes resulted in a data set of 7,000+ records, of which both class labels ("Yes" and "No") were equally represented. Again, using a 10-fold cross-validation methodology, the study developed and tested prediction models for all four model types. The results of these experiments are shown in Table 3.3. Based on the hold-out sample results, support vector machines once again generated the best overall prediction accuracy with 81.18 percent followed by decision trees, artificial neural networks, and logistic regression with an overall prediction accuracy of 80.65 percent, 79.85 percent, and 74.26 percent, respectively. As can be seen in the per-class accuracy figures, the prediction models did significantly better on predicting the "No" class with the well-balanced data than they did with the unbalanced data. Overall, the three machine-learning techniques performed significantly better than their statistical counterpart, logistic regression.

Next, another set of experiments was conducted to assess the predictive ability of the three ensemble models. Based on the 10-fold cross-validation methodology, the information fusion–type ensemble model produced the best results with an overall prediction rate of 82.10 percent, followed by the bagging-type ensembles and boosting-type

TABLE 3.2 Prediction Results for the Original/Unbalanced Data Set

	ANN(MLP)		DT(C5)		SVM		LR	
	No	Yes	No	Yes	No	Yes	No	Yes
No	1,494	384	1,518	304	1,478	255	1,438	376
Yes	1,596	11,142	1,572	11,222	1,612	11,271	1,652	11,150
SUM	3,090	11,526	3,090	11,526	3,090	11,526	3,090	11,526
Per-class accuracy	48.35%	96.67%	49.13%	97.36%	47.83%	97.79%	46.54%	96.74%
Overall accuracy	86.45%		87.16%		87.23%		86.12%	

*ANN: Artificial Neural Network; MLP: Multi-Layer Perceptron; DT: Decision Tree; SVM: Support Vector Machine; LR: Logistic Regression

TABLE 3.3 Prediction Results for the Balanced Data Set

Confusion Matrix	ANN(MLP)		DT(C5)		SVM		LR	
	No	Yes	No	Yes	No	Yes	No	Yes
No	2,309	464	2311	417	2,313	386	2,125	626
Yes	781	2,626	779	2,673	777	2,704	965	2,464
SUM	3,090	3,090	3,090	3,090	3,090	3,090	3,090	3,090
Per-class accuracy	74.72%	84.98%	74.79%	86.50%	74.85%	87.51%	68.77%	79.74%
Overall accuracy	79.85%		80.65%		81.18%		74.26%	

ensembles with overall prediction rates of 81.80 percent and 80.21 percent, respectively (see Table 3.4). Even though the prediction results are slightly better than those of the individual models, ensembles are known to produce more robust prediction systems compared to a single-best prediction model (more on this can be found in Chapter 4).

In addition to assessing the prediction accuracy for each model type, a sensitivity analysis was also conducted using the developed prediction models to identify the relative importance of the independent variables (i.e., the predictors). In realizing the overall sensitivity analysis results, each of the four individual model types generated its own sensitivity measures, ranking all independent variables in a prioritized list. As expected, each model type generated slightly different sensitivity rankings of the independent variables. After collecting all four sets of sensitivity numbers, the sensitivity numbers are normalized and aggregated and plotted in a horizontal bar chart (see Figure 3.6).

The Conclusions

The study showed that, given sufficient data with the proper variables, data mining methods are capable of predicting freshmen student attrition with approximately 80 percent accuracy. Results also showed that, regardless of the prediction model employed, the balanced data set (compared to unbalanced/original data set) produced better prediction models for identifying the students who are likely to drop out of the college prior to their sophomore year. Among the four individual prediction models used in this study, support vector machines performed the best, followed by decision trees, neural networks, and logistic regression. From the usability standpoint, despite the fact that support vector machines showed better prediction results, one might choose to use decision trees, because compared to support vector machines and neural networks, they portray a more transparent model structure. Decision trees

TABLE 3.4 Prediction Results for the Three Ensemble Models

	Boosting (boosted trees)		Bagging (random forest)		Information Fusion (weighted average)	
	No	Yes	No	Yes	No	Yes
No	2,242	375	2,327	362	2,335	351
Yes	848	2,715	763	2,728	755	2,739
SUM	3,090	3,090	3,090	3,090	3,090	3,090
Per-class accuracy	72.56%	87.86%	75.31%	88.28%	75.57%	88.64%
Overall accuracy	80.21%		81.80%		82.10%	

(*Continued*)

Application Case 3.2 (Continued)

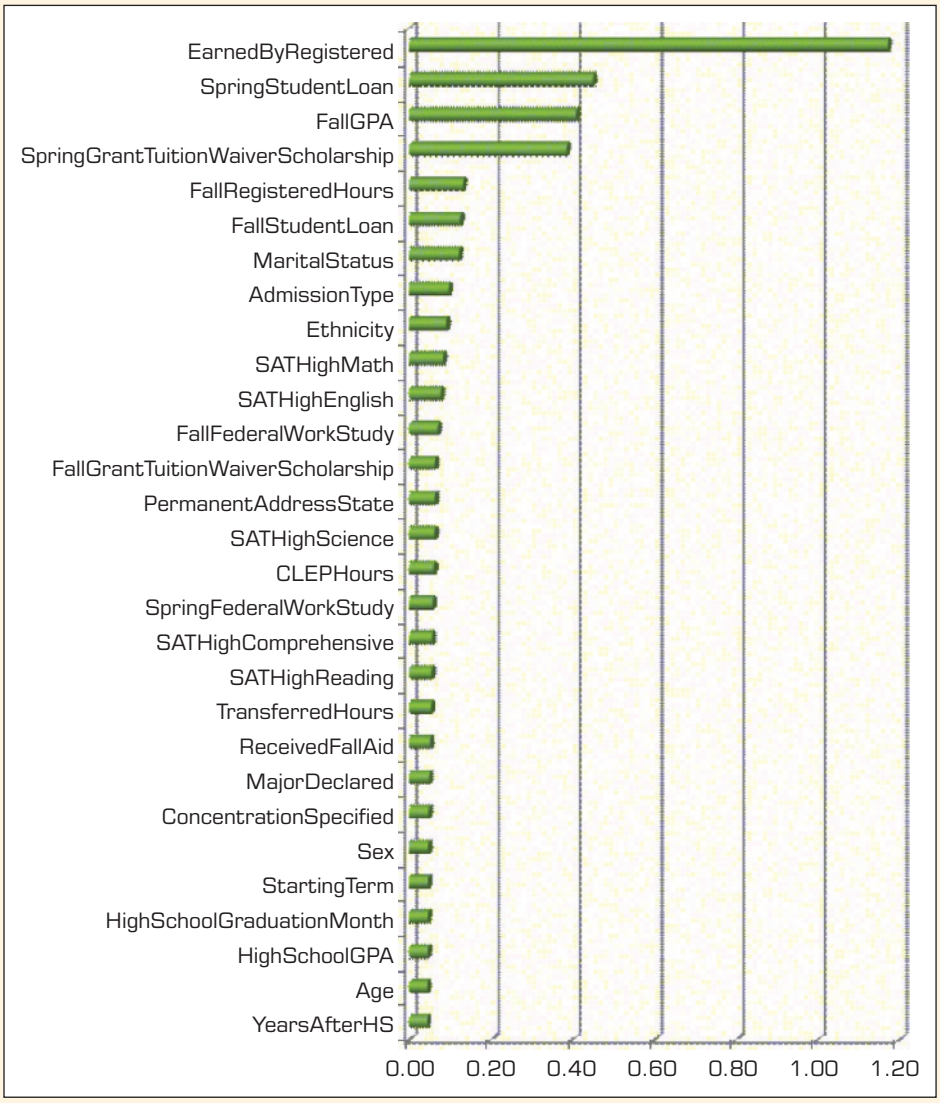


FIGURE 3.6 **Sensitivity-Analysis-Based Variable Importance Results.**

explicitly show the reasoning process of different predictions, providing a justification for a specific outcome, whereas support vector machines and artificial neural networks are mathematical models that do not provide such a transparent view of "how they do what they do."

Questions for Case 3.2

1. What is student attrition, and why is it an important problem in higher education?
2. What were the traditional methods to deal with the attrition problem?
3. List and discuss the data-related challenges within the context of this case study.
4. What was the proposed solution? What were the results?

Sources: D. Thammasiri, D. Delen, P. Meesad, & N. Kasap, "A Critical Assessment of Imbalanced Class Distribution Problem: The Case of Predicting Freshmen Student Attrition," *Expert Systems with Applications, 41*(2), 2014, pp. 321–330; D. Delen, "A Comparative Analysis of Machine Learning Techniques for Student Retention Management," *Decision Support Systems, 49*(4), 2010, pp. 498–506, and "Predicting Student Attrition with Data Mining Methods," *Journal of College Student Retention 13*(1), 2011, pp. 17–35.

SECTION 3.4 REVIEW QUESTIONS

1. Why are the original/raw data not readily usable by analytics tasks?
2. What are the main data preprocessing steps?
3. What does it mean to clean/scrub the data? What activities are performed in this phase?
4. Why do we need data transformation? What are the commonly used data transformation tasks?
5. Data reduction can be applied to rows (sampling) and/or columns (variable selection). Which is more challenging?

3.5 STATISTICAL MODELING FOR BUSINESS ANALYTICS

Because of the increasing popularity of business analytics, the traditional statistical methods and underlying techniques are also regaining their attractiveness as enabling tools to support evidence-based managerial decision making. Not only are they regaining attention and admiration, but this time, they are attracting business users in addition to statisticians and analytics professionals.

Statistics (statistical methods and underlying techniques) is usually considered as part of descriptive analytics (see Figure 3.7). Some of the statistical methods can also be considered as part of predictive analytics, such as discriminant analysis, multiple regression, logistic regression, and k-means clustering. As shown in Figure 3.7, descriptive analytics has two main branches: statistics and **online analytics processing (OLAP)**. OLAP is the term used for analyzing, characterizing, and summarizing structured data stored in organizational databases (often stored in a data warehouse or in a data mart) using cubes (i.e., multidimensional data structures that are created to extract a subset of data values to answer a specific business question). The OLAP branch of descriptive analytics has also been called *business intelligence*. Statistics, on the other hand, helps to characterize the data, either one variable at a time or multivariable, all together using either descriptive or inferential methods.

Statistics—a collection of mathematical techniques to characterize and interpret data—has been around for a very long time. Many methods and techniques have been developed to address the needs of the end users and the unique characteristics of the data being analyzed. Generally speaking, at the highest level, statistical methods can be

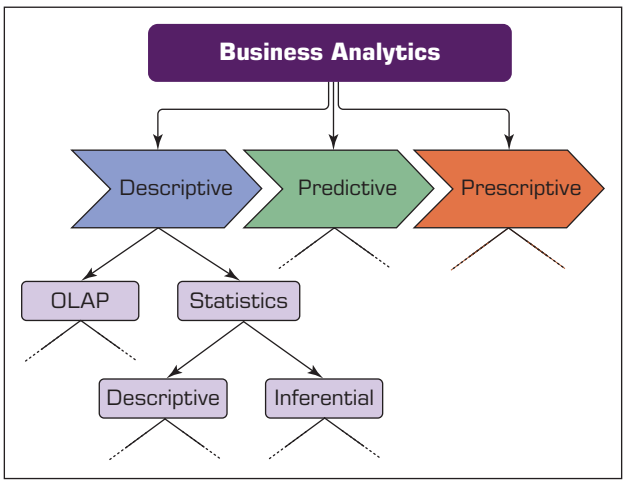


FIGURE 3.7 Relationship between Statistics and Descriptive Analytics

classified as either descriptive or inferential. The main difference between descriptive and inferential statistics is the data used in these methods—whereas **descriptive statistics** is all about describing the sample data on hand, **inferential statistics** is about drawing inferences or conclusions about the characteristics of the population. In this section, we briefly describe descriptive statistics (because of the fact that it lays the foundation for, and is the integral part of, descriptive analytics), and in the following section we cover regression (both linear and logistic regression) as part of inferential statistics.

Descriptive Statistics for Descriptive Analytics

Descriptive statistics, as the name implies, describes the basic characteristics of the data at hand, often one variable at a time. Using formulas and numerical aggregations, descriptive statistics summarizes the data in such a way that often meaningful and easily understandable patterns emerge from the study. Although it is very useful in data analytics and very popular among the statistical methods, descriptive statistics does not allow making conclusions (or inferences) beyond the sample of the data being analyzed. That is, it is simply a nice way to characterize and describe the data on hand without making conclusions (inferences or extrapolations) regarding the population of related hypotheses we might have in mind.

In business analytics, descriptive statistics plays a critical role—it allows us to understand and explain/present our data in a meaningful manner using aggregated numbers, data tables, or charts/graphs. In essence, descriptive statistics helps us convert our numbers and symbols into meaningful representations for anyone to understand and use. Such an understanding helps not only business users in their decision-making processes but also analytics professionals and data scientists to characterize and validate the data for other more sophisticated analytics tasks. Descriptive statistics allows analysts to identify data concertation, unusually large or small values (i.e., outliers), and unexpectedly distributed data values for numeric variables. Therefore, the methods in descriptive statistics can be classified as either measures for central tendency or measures of dispersion. In the following section, we use a simple description and mathematical formulation/representation of these measures. In mathematical representation, we will use $x_1, x_2, \ldots, x_n$ to represent individual values (observations) of the variable (measure) that we are interested in characterizing.

Measures of Centrality Tendency (Also Called *Measures of Location* or *Centrality*)

Measures of centrality are the mathematical methods by which we estimate or describe central positioning of a given variable of interest. A measure of central tendency is a single numerical value that aims to describe a set of data by simply identifying or estimating the central position within the data. The mean (often called the *arithmetic mean* or the *simple average*) is the most commonly used measure of central tendency. In addition to mean, you could also see median or mode being used to describe the centrality of a given variable. Although, the mean, median, and mode are all valid measures of central tendency, under different circumstances, one of these measures of centrality becomes more appropriate than the others. What follows are short descriptions of these measures, including how to calculate them mathematically and pointers on the circumstances in which they are the most appropriate measure to use.

Arithmetic Mean

The **arithmetic mean** (or simply *mean* or *average*) is the sum of all the values/observations divided by the number of observations in the data set. It is by far the most popular

and most commonly used measure of central tendency. It is used with continuous or discrete numeric data. For a given variable x, if we happen to have n values/observations $(x_1, x_2, \ldots, x_n)$, we can write the arithmetic mean of the data sample ($\bar{x}$, pronounced as x-bar) as follows:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

or

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

The mean has several unique characteristics. For instance, the sum of the absolute deviations (differences between the mean and the observations) above the mean is the same as the sum of the deviations below the mean, balancing the values on either side of it. That said, it does not suggest, however, that half the observations are above and the other half are below the mean (a common misconception among those who do not know basic statistics). Also, the mean is unique for every data set and is meaningful and calculable for both interval- and ratio-type numeric data. One major downside is that the mean can be affected by outliers (observations that are considerably larger or smaller than the rest of the data points). Outliers can pull the mean toward their direction and, hence, bias the centrality representation. Therefore, if there are outliers or if the data are erratically dispersed and skewed, one should either avoid using the mean as the measure of centrality or augment it with other central tendency measures, such as median and mode.

Median

The **median** is the measure of center value in a given data set. It is the number in the middle of a given set of data that has been arranged/sorted in order of magnitude (either ascending or descending). If the number of observations is an odd number, identifying the median is very easy—just sort the observations based on their values and pick the value right in the middle. If the number of observations is an even number, identify the two middle values, and then take the simple average of these two values. The median is meaningful and calculable for ratio, interval, and ordinal data types. Once determined, one-half of the data points in the data is above and the other half is below the median. In contrary to the mean, the median is not affected by outliers or skewed data.

Mode

The **mode** is the observation that occurs most frequently (the most frequent value in our data set). On a histogram, it represents the highest bar in a bar chart, and, hence, it can be considered as the most popular option/value. The mode is most useful for data sets that contain a relatively small number of unique values. That is, it could be useless if the data have too many unique values (as is the case in many engineering measurements that capture high precision with a large number of decimal places), rendering each value having either one or a very small number representing its frequency. Although it is a useful measure (especially for nominal data), mode is not a very good representation of centrality, and therefore, it should not be used as the only measure of central tendency for a given data set.

In summary, which central tendency measure is the best? Although there is not a clear answer to this question, here are a few hints—use the mean when the data are not prone to outliers and there is no significant level of skewness; use the median when the data have outliers and/or it is ordinal in nature; use the mode when the data are nominal.

Perhaps the best practice is to use all three together so that the central tendency of the data set can be captured and represented from three perspectives. Mostly because "average" is a very familiar and highly used concept to everyone in regular daily activities, managers (as well as some scientists and journalists) often use the centrality measures (especially mean) inappropriately when other statistical information should be considered along with the centrality. It is a better practice to present descriptive statistics as a package—a combination of centrality and dispersion measures—as opposed to a single measure such as mean.

Measures of Dispersion (Also Called *Measures of Spread* or *Decentrality*)

Measures of **dispersion** are the mathematical methods used to estimate or describe the degree of variation in a given variable of interest. They represent the numerical spread (compactness or lack thereof) of a given data set. To describe this dispersion, a number of statistical measures are developed; the most notable ones are range, variance, and standard deviation (and also quartiles and absolute deviation). One of the main reasons why the measures of dispersion/spread of data values are important is the fact that they give us a framework within which we can judge the central tendency—give us the indication of how well the mean (or other centrality measures) represents the sample data. If the dispersion of values in the data set is large, the mean is not deemed to be a very good representation of the data. This is because a large dispersion measure indicates large differences between individual scores. Also, in research, it is often perceived as a positive sign to see a small variation within each data sample, as it may indicate homogeneity, similarity, and robustness within the collected data.

Range

The **range** is perhaps the simplest measure of dispersion. It is the difference between the largest and the smallest values in a given data set (i.e., variables). So we calculate range by simply identifying the smallest value in the data set (minimum), identifying the largest value in the data set (maximum), and calculating the difference between them (range = maximum − minimum).

Variance

A more comprehensive and sophisticated measure of dispersion is the **variance**. It is a method used to calculate the deviation of all data points in a given data set from the mean. The larger the variance, the more the data are spread out from the mean and the more variability one can observe in the data sample. To prevent the offsetting of negative and positive differences, the variance takes into account the square of the distances from the mean. The formula for a data sample can be written as

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

where n is the number of samples, $\bar{x}$ is the mean of the sample, and x_i is the i^{th} value in the data set. The larger values of variance indicate more dispersion, whereas smaller values indicate compression in the overall data set. Because the differences are squared, larger deviations from the mean contribute significantly to the value of variance. Again, because the differences are squared, the numbers that represent deviation/variance become somewhat meaningless (as opposed to a dollar difference, here you are given a squared dollar difference). Therefore, instead of variance, in

many business applications, we use a more meaningful dispersion measure, called *standard deviation*.

Standard Deviation

The **standard deviation** is also a measure of the spread of values within a set of data. The standard deviation is calculated by simply taking the square root of the variations. The following formula shows the calculation of standard deviation from a given sample of data points.

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

Mean Absolute Deviation

In addition to variance and standard deviation, sometimes we also use **mean absolute deviation** to measure dispersion in a data set. It is a simpler way to calculate the overall deviation from the mean. Specifically, the mean absolute deviation is calculated by measuring the absolute values of the differences between each data point and the mean and then summing them. This process provides a measure of spread without being specific about the data point being lower or higher than the mean. The following formula shows the calculation of the mean absolute deviation:

$$MAD = \frac{\sum_{i=1}^{n}|x_i - \bar{x}|}{n}$$

Quartiles and Interquartile Range

Quartiles help us identify spread within a subset of the data. A **quartile** is a quarter of the number of data points given in a data set. Quartiles are determined by first sorting the data and then splitting the sorted data into four disjoint smaller data sets. Quartiles are a useful measure of dispersion because they are much less affected by outliers or a skewness in the data set than the equivalent measures in the whole data set. Quartiles are often reported along with the median as the best choice of measure of dispersion and central tendency, respectively, when dealing with skewed and/or data with outliers. A common way of expressing quartiles is as an interquartile range, which describes the difference between the third quartile (Q3) and the first quartile (Q1), telling us about the range of the middle half of the scores in the distribution. The quartile-driven descriptive measures (both centrality and dispersion) are best explained with a popular plot called a *box-and-whiskers plot* (or *box plot*).

Box-and-Whiskers Plot

The **box-and-whiskers plot** (or simply a **box plot**) is a graphical illustration of several descriptive statistics about a given data set. They can be either horizontal or vertical, but vertical is the most common representation, especially in modern-day analytics software products. It is known to be first created and presented by John W. Tukey in 1969. Box plot is often used to illustrate both centrality and dispersion of a given data set (i.e., the distribution of the sample data) in an easy-to-understand graphical notation. Figure 3.8 shows two box plots side by side, sharing the same *y*-axis. As shown therein, a single chart can have one or more box plots for visual comparison purposes. In such cases, the *y*-axis would be the common measure of magnitude (the numerical value of the

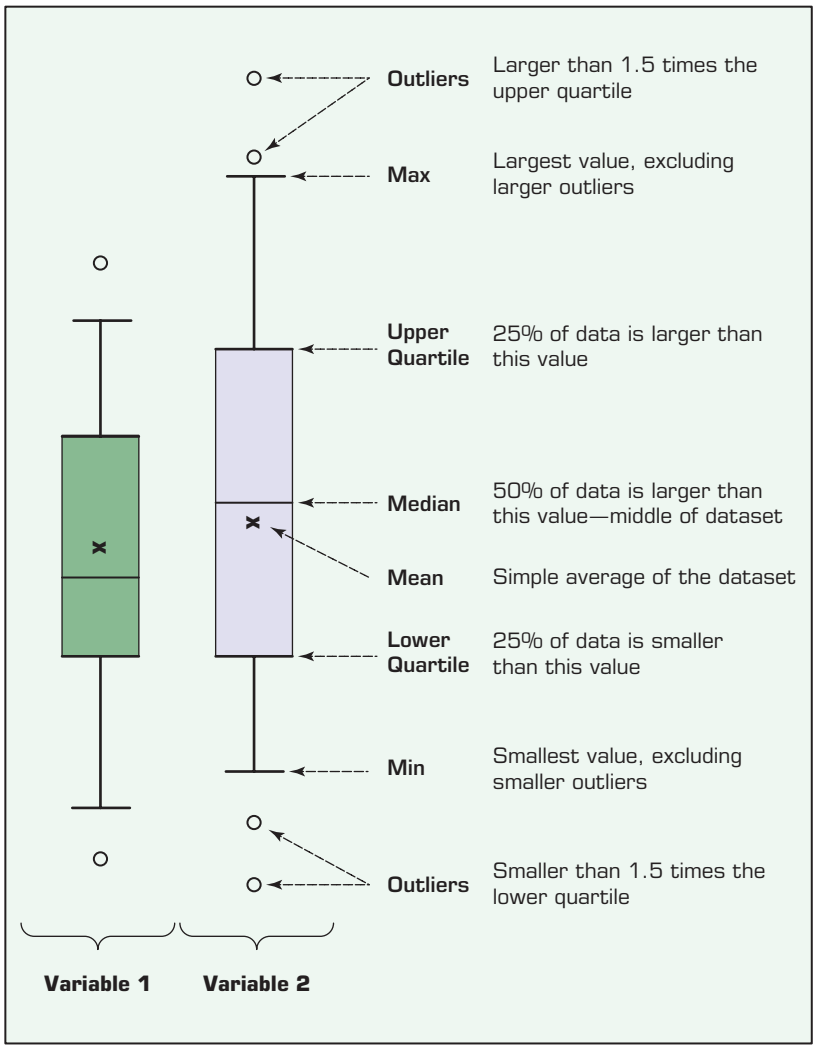


FIGURE 3.8 **Understanding the Specifics about Box-and-Whiskers Plots.**

variable), with the *x*-axis showing different classes/subsets such as different time dimensions (e.g., descriptive statistics for annual Medicare expenses in 2015 versus 2016) or different categories (e.g., descriptive statistics for marketing expenses versus total sales).

Although historically speaking, the box plot has not been used widely and often enough (especially in areas outside of statistics), with the emerging popularity of business analytics, it is gaining fame in less technical areas of the business world. Its information richness and ease of understanding are largely to credit for its recent popularity.

The box plot shows the **centrality** (median and sometimes also mean) as well as the dispersion (the density of the data within the middle half—drawn as a box between the first and third quartiles), the minimum and maximum ranges (shown as extended lines from the box, looking like whiskers, that are calculated as 1.5 times the upper or lower end of the quartile box), and the outliers that are larger than the limits of the whiskers. A box plot also shows whether the data are symmetrically distributed with respect to the mean or sway one way or another. The relative position of the median versus mean and the lengths of the whiskers on both side of the box give a good indication of the potential skewness in the data.

Shape of a Distribution

Although not as common as the centrality and dispersion, the shape of the data distribution is also a useful measure for the descriptive statistics. Before delving into the shape of the distribution, we first need to define the distribution itself. Simply put, *distribution* is the frequency of data points counted and plotted over a small number of class labels or numerical ranges (i.e., bins). In a graphical illustration of distribution, the *y*-axis shows the frequency (count or percentage), and the *x*-axis shows the individual classes or bins in a rank-ordered fashion. A very well-known distribution is called *normal distribution*, which is perfectly symmetric on both sides of the mean and has numerous well-founded mathematical properties that make it a very useful tool for research and practice. As the dispersion of a data set increases, so does the standard deviation, and the shape of the distribution looks wider. A graphic illustration of the relationship between dispersion and distribution shape (in the context of normal distribution) is shown in Figure 3.9.

There are two commonly used measures to calculate the shape characteristics of a distribution: skewness and kurtosis. A histogram (frequency plot) is often used to visually illustrate both skewness and kurtosis.

Skewness is a measure of asymmetry (sway) in a distribution of the data that portrays a unimodal structure—only one peak exists in the distribution of the data. Because normal distribution is a perfectly symmetric unimodal distribution, it does not have

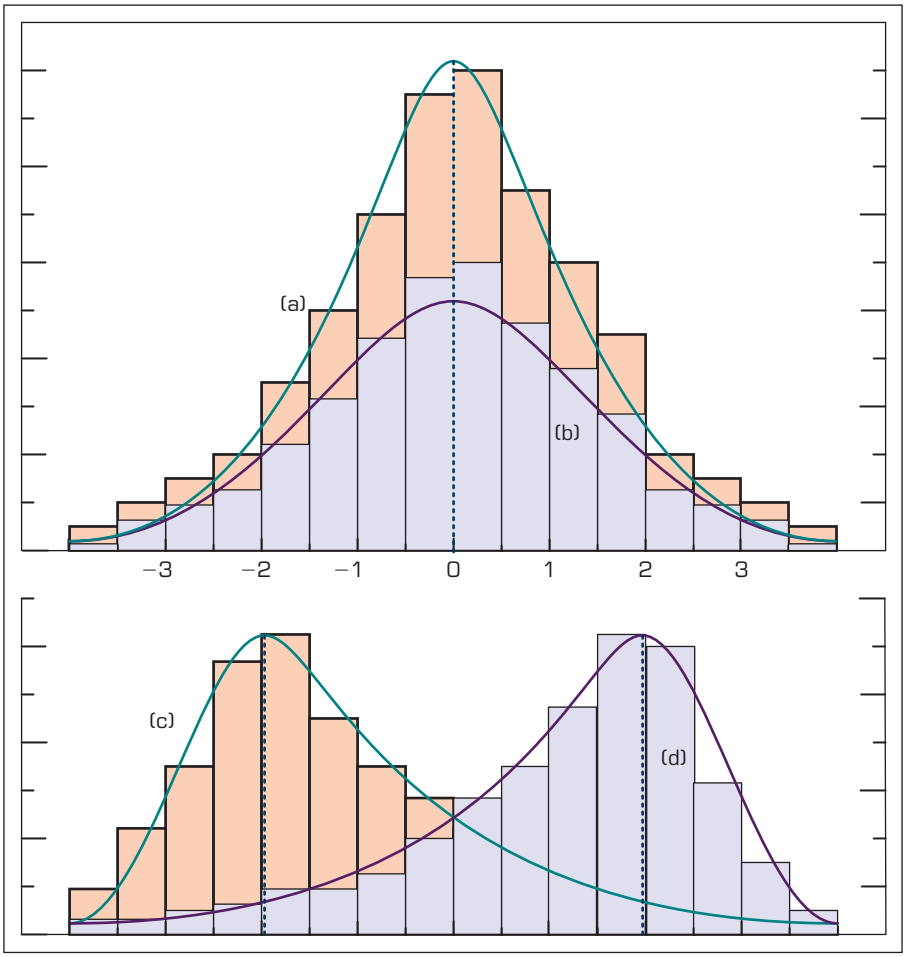


FIGURE 3.9 **Relationship between Dispersion and Distribution Shape Properties.**

skewness; that is, its skewness measure (i.e., the value of the coefficient of skewness) is equal to zero. The skewness measure/value can be either positive or negative. If the distribution sways left (i.e., the tail is on the right side and the mean is smaller than median), then it produces a positive skewness measure; if the distribution sways right (i.e., the tail is on the left side and the mean is larger than median), then it produces a negative skewness measure. In Figure 3.9, (c) represents a positively skewed distribution whereas (d) represents a negatively skewed distribution. In the same figure, both (a) and (b) represent perfect symmetry and hence zero measure for skewness.

$$Skewness = S = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{(n-1)s^3}$$

where s is the standard deviation and n is the number of samples.

Kurtosis is another measure to use in characterizing the shape of a unimodal distribution. As opposed to the sway in shape, kurtosis focuses more on characterizing the peak/tall/skinny nature of the distribution. Specifically, kurtosis measures the degree to which a distribution is more or less peaked than a normal distribution. Whereas a positive kurtosis indicates a relatively peaked/tall distribution, a negative kurtosis indicates a relatively flat/short distribution. As a reference point, a normal distribution has a kurtosis of 3. The formula for kurtosis can be written as

$$Kurtosis = K = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{ns^4} - 3$$

Descriptive statistics (as well as inferential statistics) can easily be calculated using commercially viable statistical software packages (e.g., SAS, SPSS, Minitab, JMP, Statistica) or free/open source tools (e.g., R). Perhaps the most convenient way to calculate descriptive and some of the inferential statistics is to use Excel. Technology Insights 3.1 describes in detail how to use Microsoft Excel to calculate descriptive statistics.

TECHNOLOGY INSIGHTS 3.1 How to Calculate Descriptive Statistics in Microsoft Excel

Excel, arguably the most popular data analysis tool in the world, can easily be used for descriptive statistics. Although the base configuration of Excel does not seem to have the statistics function readily available for end users, those functions come with the Excel installation and can be activated (turned on) with only a few mouse clicks. Figure 3.10 shows how these statistics functions (as part of the Analysis ToolPak) can be activated in Microsoft Excel 2016.

Once activated, the *Analysis ToolPak* will appear in the *Data* menu option under the name of *Data Analysis*. When you click on Data Analysis in the Analysis group under the Data tab in the Excel menu bar, you will see Descriptive Statistics as one of the options within the list of data analysis tools (see Figure 3.11, steps 1, 2); click on OK, and the Descriptive Statistics dialog box will appear (see the middle of Figure 3.11). In this dialog box, you need to enter the range of the data, which can be one or more numerical columns, along with the preference check boxes, and click OK (see Figure 3.11, steps 3, 4). If the selection includes more than one numeric column, the tool treats each column as a separate data set and provides descriptive statistics for each column separately.

As a simple example, we selected two columns (labeled as Expense and Demand) and executed the Descriptive Statistics option. The bottom section of Figure 3.11 shows the output created by Excel. As can be seen, Excel produced all descriptive statistics that are covered in the previous section and added a few more to the list. In Excel 2016, it is also very easy (a few

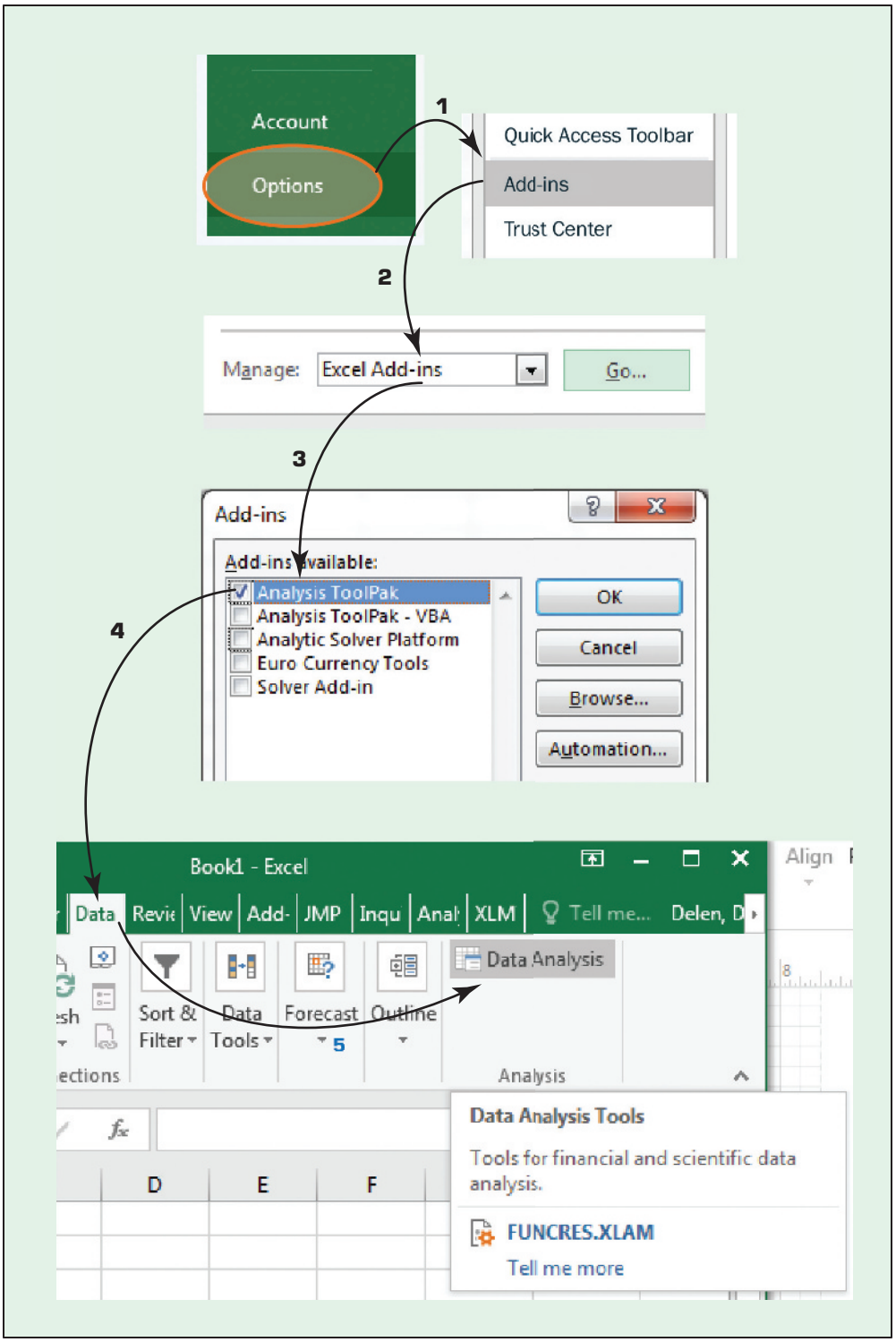


FIGURE 3.10 **Activating Statistics Function in Excel 2016.**

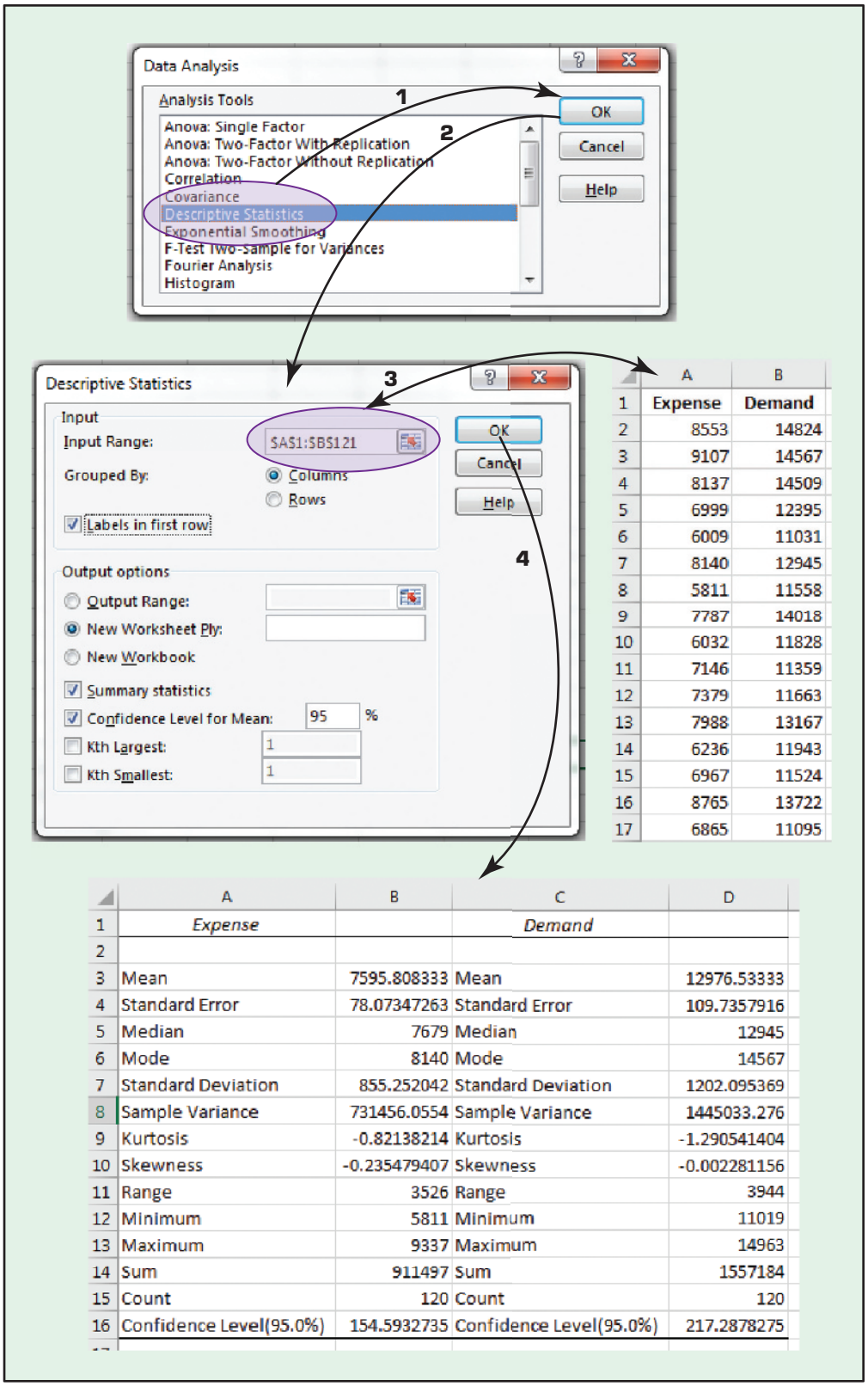


FIGURE 3.11 **Obtaining Descriptive Statistics in Excel.**

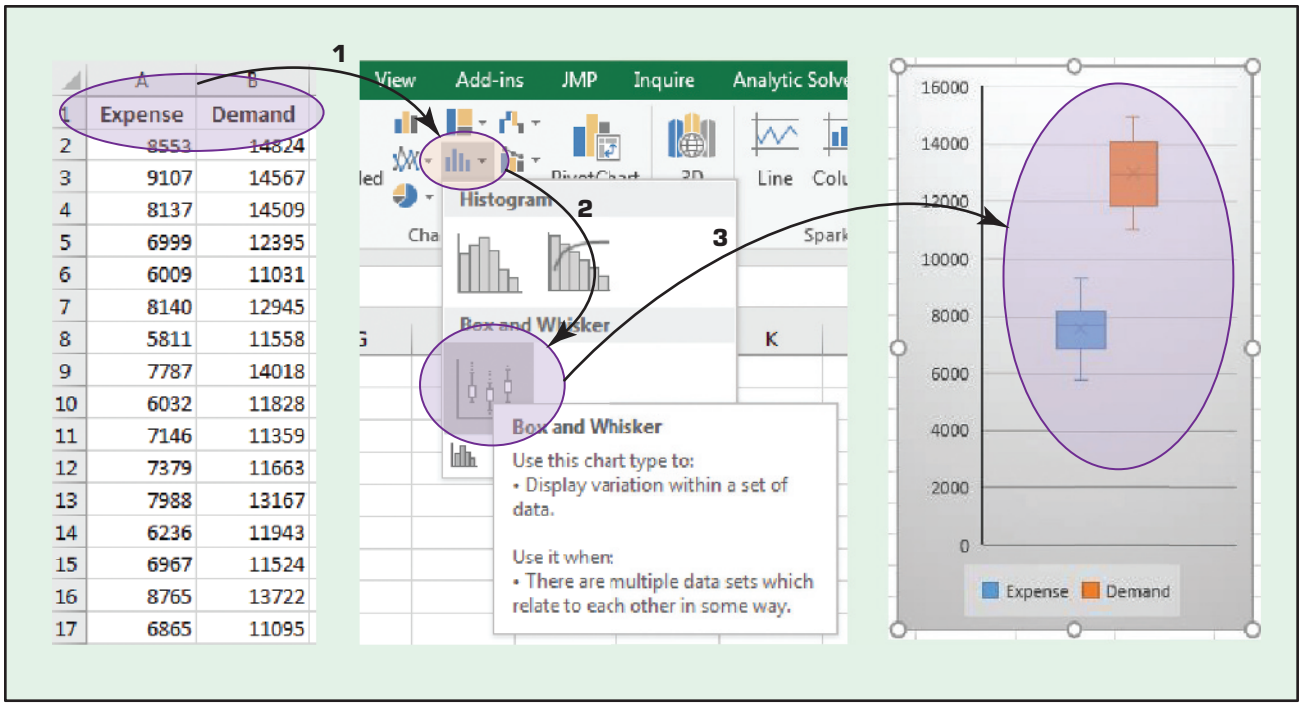


FIGURE 3.12 **Creating a Box-and-Whiskers Plot in Excel 2016.**

mouse clicks) to create a box-and-whiskers plot. Figure 3.12 shows the simple three-step process of creating a box-and-whiskers plot in Excel.

Although Analysis ToolPak is a very useful tool in Excel, one should be aware of an important point related to the results that it generates, which have a different behavior than other ordinary Excel functions: Although Excel functions dynamically change as the underlying data in the spreadsheet are changed, the results generated by the Analysis ToolPak do not. For example, if you change the values in either or both of these columns, the Descriptive Statistics results produced by the Analysis ToolPak will stay the same. However, the same is not true for ordinary Excel functions. If you were to calculate the mean value of a given column (using "=AVERAGE(A1:A121)") and then change the values within the data range, the mean value would automatically change. In summary, the results produced by Analysis ToolPak do not have a dynamic link to the underlying data, and if the data change, the analysis needs to be redone using the dialog box.

Successful applications of data analytics cover a wide range of business and organizational settings, addressing problems once thought unsolvable. Application Case 3.3 is an excellent illustration of those success stories in which a small municipality administration adopted a data analytics approach to intelligently detect and solve problems by continuously analyzing demand and consumption patterns.

SECTION 3.5 REVIEW QUESTIONS

1. What is the relationship between statistics and business analytics?
2. What are the main differences between descriptive and inferential statistics?
3. List and briefly define the central tendency measures of descriptive statistics.
4. List and briefly define the dispersion measures of descriptive statistics.
5. What is a box-and-whiskers plot? What types of statistical information does it represent?
6. What are the two most commonly used shape characteristics to describe a data distribution?

Application Case 3.3

Town of Cary Uses Analytics to Analyze Data from Sensors, Assess Demand, and Detect Problems

A leaky faucet. A malfunctioning dishwasher. A cracked sprinkler head. These are more than just a headache for a home owner or business to fix. They can be costly, unpredictable, and, unfortunately, hard to pinpoint. Through a combination of wireless water meters and a data-analytics-driven, customer-accessible portal, the Town of Cary, North Carolina, is making it much easier to find and fix water loss issues. In the process, the town has gained a big-picture view of water usage critical to planning future water plant expansions and promoting targeted conservation efforts.

When the town of Cary installed wireless meters for 60,000 customers in 2010, it knew the new technology wouldn't just save money by eliminating manual monthly readings; the town also realized it would get more accurate and timely information about water consumption. The Aquastar wireless system reads meters once an hour—that is 8,760 data points per customer each year instead of 12 monthly readings. The data had tremendous potential if they could be easily consumed.

"Monthly readings are like having a gallon of water's worth of data. Hourly meter readings are more like an Olympic-size pool of data," says Karen Mills, finance director for Cary. "SAS helps us manage the volume of that data nicely." In fact, the solution enables the town to analyze half a billion data points on water usage and make them available to and easily consumable by all customers.

The ability to visually look at data by household or commercial customer by the hour has led to some very practical applications:

- The town can notify customers of potential leaks within days.
- Customers can set alerts that notify them within hours if there is a spike in water usage.
- Customers can track their water usage online, helping them to be more proactive in conserving water.

Through the online portal, one business in the town saw a spike in water consumption on weekends when employees are away. This seemed odd, and the unusual reading helped the company learn that a commercial dishwasher was malfunctioning, running continuously over weekends. Without the wireless water-meter data and the customer-accessible portal, this problem could have gone unnoticed, continuing to waste water and money.

The town has a much more accurate picture of daily water usage per person, critical for planning future water plant expansions. Perhaps the most interesting perk is that the town was able to verify a hunch that has far-reaching cost ramifications: Cary residents are very economical in their use of water. "We calculate that with modern high-efficiency appliances, indoor water use could be as low as 35 gallons per person per day. Cary residents average 45 gallons, which is still phenomenally low," explains town Water Resource Manager Leila Goodwin. Why is this important? The town was spending money to encourage water efficiency—rebates on low-flow toilets or discounts on rain barrels. Now it can take a more targeted approach, helping specific consumers understand and manage both their indoor and outdoor water use.

SAS was critical not just for enabling residents to understand their water use but also working behind the scenes to link two disparate databases. "We have a billing database and the meter-reading database. We needed to bring that together and make it presentable," Mills says.

The town estimates that by just removing the need for manual readings, the Aquastar system will save more than $10 million above the cost of the project. But the analytics component could provide even bigger savings. Already, both the town and individual citizens have saved money by catching water leaks early. As Cary continues to plan its future infrastructure needs, having accurate information on water usage will help it invest in the right amount of infrastructure at the right time. In addition, understanding water usage will help the town if it experiences something detrimental like a drought.

"We went through a drought in 2007," says Goodwin. "If we go through another, we have a plan in place to use Aquastar data to see exactly how much water we are using on a day-by-day basis and communicate with customers. We can show 'here's what's happening, and here is how much you can use because our supply is low.' Hopefully, we'll never have to use it, but we're prepared."

Questions for Case 3.3

1. What were the challenges the Town of Cary was facing?
2. What was the proposed solution?
3. What were the results?
4. What other problems and data analytics solutions do you foresee for towns like Cary?

Source: "Municipality Puts Wireless Water Meter-Reading Data To Work (SAS® Analytics)—The Town of Cary, North Carolina Uses SAS Analytics to Analyze Data from Wireless Water Meters, Assess Demand, Detect Problems and Engage Customers." Copyright © 2016 SAS Institute Inc., Cary, NC, USA. Reprinted with permission. All rights reserved.

3.6 REGRESSION MODELING FOR INFERENTIAL STATISTICS

Regression, especially linear regression, is perhaps the most widely known and used analytics technique in statistics. Historically speaking, the roots of regression date back to the 1920s and 1930s, to the earlier work on inherited characteristics of sweet peas by Sir Francis Galton and subsequently by Karl Pearson. Since then, regression has become the statistical technique for characterization of relationships between explanatory (input) variable(s) and response (output) variable(s).

As popular as it is, regression essentially is a relatively simple statistical technique to model the dependence of a variable (response or output variable) on one (or more) explanatory (input) variables. Once identified, this relationship between the variables can be formally represented as a linear/additive function/equation. As is the case with many other modeling techniques, regression aims to capture the functional relationship between and among the characteristics of the real world and describe this relationship with a mathematical model, which can then be used to discover and understand the complexities of reality—explore and explain relationships or forecast future occurrences.

Regression can be used for one of two purposes: hypothesis testing—investigating potential relationships between different variables—and prediction/forecasting—estimating values of a response variable based on one or more explanatory variables. These two uses are not mutually exclusive. The explanatory power of regression is also the foundation of its predictive ability. In hypothesis testing (theory building), regression analysis can reveal the existence/strength and the directions of relationships between a number of explanatory variables (often represented with x_i) and the response variable (often represented with y). In prediction, regression identifies additive mathematical relationships (in the form of an equation) between one or more explanatory variables and a response variable. Once determined, this equation can be used to forecast the values of the response variable for a given set of values of the explanatory variables.

CORRELATION VERSUS REGRESSION Because regression analysis originated from correlation studies, and because both methods attempt to describe the association between two (or more) variables, these two terms are often confused by professionals and even by scientists. **Correlation** makes no a priori assumption of whether one variable is dependent on the other(s) and is not concerned with the relationship between variables; instead it gives an estimate on the degree of association between the variables. On the other hand, regression attempts to describe the dependence of a response variable on one (or more) explanatory variables where it implicitly assumes that there is a one-way causal effect from the explanatory variable(s) to the response variable, regardless of whether the path of effect is direct or indirect. Also, although correlation is interested in the low-level relationships between two variables, regression is concerned with the relationships between all explanatory variables and the response variable.