CHAPTER

2

Descriptive Analytics I: Nature of Data, Statistical Modeling, and Visualization

LEARNING OBJECTIVES

- Understand the nature of data as it relates to business intelligence (BI) and analytics
- Learn the methods used to make real-world data analytics ready
- Describe statistical modeling and its relationship to business analytics
- Learn about descriptive and inferential statistics
- Define business reporting, and understand its historical evolution
- Understand the importance of data/information visualization
- Learn different types of visualization techniques
- Appreciate the value that visual analytics brings to business analytics
- Know the capabilities and limitations of dashboards

In the age of Big Data and business analytics in which we are living, the importance of data is undeniable. The newly coined phrases like "data is the oil," "data is the new bacon," "data is the new currency," and "data is the king" are further stressing the renewed importance of data. But what type of data are we talking about? Obviously, not just any data. The "garbage in garbage out—GIGO" concept/principle applies to today's "Big Data" phenomenon more so than any data definition that we have had in the past. To live up to its promise, its value proposition, and its ability to turn into insight, data has to be carefully created/identified, collected, integrated, cleaned, transformed, and properly contextualized for use in accurate and timely decision making.

Data is the main theme of this chapter. Accordingly, the chapter starts with a description of the nature of data: what it is, what different types and forms it can come in, and how it can be preprocessed and made ready for analytics. The first few sections of the chapter are dedicated to a deep yet necessary understanding and processing of data. The next few sections describe the statistical methods used to prepare data as input to produce both descriptive and inferential measures. Following the statistics sections are sections on reporting and visualization. A report is a communication artifact prepared with the specific intention of converting data into information and knowledge and relaying that

information in an easily understandable/digestible format. Nowadays, these reports are more visually oriented, often using colors and graphical icons that collectively look like a dashboard to enhance the information content. Therefore, the latter part of the chapter is dedicated to subsections that present the design, implementation, and best practices for information visualization, storytelling, and information dashboards.

2.1 OPENING VIGNETTE: SiriusXM Attracts and Engages a New Generation of Radio Consumers with Data-Driven Marketing

SiriusXM Radio is a satellite radio powerhouse, the largest radio company in the world with $3.8 billion in annual revenues and a wide range of hugely popular music, sports, news, talk, and entertainment stations. The company, which began broadcasting in 2001 with 50,000 subscribers, grew to 18.8 million subscribers in 2009, and today has nearly 29 million.

Much of SiriusXM's growth to date is rooted in creative arrangements with automobile manufacturers; today, nearly 70% of new cars are SiriusXM-enabled. Yet the company's reach has extended far beyond car radios in the United States to a worldwide presence on the Internet, on smartphones and through other services and distribution channels, including SONOS, JetBlue, and Dish.

Business Challenge

Despite these remarkable successes, over the past few years changing customer demographics, changing technology, and a changing competitive landscape have posed a new series of business challenges and opportunities for SiriusXM. Here are some notable ones:

- As its market penetration among new cars increased, the demographics of the buyers changed, skewing younger, with less discretionary income. How could SiriusXM reach this new demographic?
- As new cars became used cars and changed hands, how could SiriusXM identify, engage, and convert second owners to paying customers?
- With its acquisition of the connected vehicle business from Agero—the leading provider of telematics in the U.S. car market—SiriusXM gained the ability to deliver its service via both satellite and wireless networks. How could it successfully use this acquisition to capture new revenue streams?

Proposed Solution: Shifting the Vision toward Data-Driven Marketing

SiriusXM recognized that to address these challenges it would need to become a high-performance, data-driven marketing organization. The company began making that shift by establishing three fundamental tenets.

First, personalized interactions—not mass marketing—would rule the day. The company quickly understood that to conduct more personalized marketing, it would have to draw on past history and interactions, as well as on a keen understanding of the consumer's place in the subscription life cycle.

Second, to gain that understanding, information technology (IT) and its external technology partners would need the ability to deliver integrated data, advanced analytics, integrated marketing platforms, and multichannel delivery systems.

And third, the company could not achieve its business goals without an integrated and consistent point of view across the company. Most important, the technology and business sides of SiriusXM would have to become true partners to best address the challenges involved in becoming a high-performance marketing organization that draws on data-driven insights to speak directly with consumers in strikingly relevant ways.

Those data-driven insights, for example, would enable the company to differentiate between consumers, owners, drivers, listeners, and account holders. The insights would help SiriusXM understand what other vehicles and services are part of each household—and to create new opportunities for engagement. In addition, by constructing a coherent and reliable 360-degree view of all its consumers, SiriusXM could ensure that all messaging in all campaigns and interactions would be tailored, relevant, and consistent across all channels. The important bonus is that more tailored and effective marketing is typically more cost-efficient.

Implementation: Creating and Following the Path to High-Performance Marketing

At the time of its decision to become a high-performance marketing company, SiriusXM was working with a third-party marketing platform that did not have the capacity to support SiriusXM's ambitions. The company then made an important, forward-thinking decision to bring its marketing capabilities in-house—and then carefully plotted out what it would need to do to make the transition successfully.

1. Improve data cleanliness through improved master data management and governance. Although the company was understandably impatient to put ideas into action, data hygiene was a necessary first step to creating a reliable window into consumer behavior.
2. Bring marketing analytics in-house and expand the data warehouse to enable scale and fully support integrated marketing analytics.
3. Develop new segmentation and scoring models to run in-database, eliminating latency and data duplication.
4. Extend the integrated data warehouse to include marketing data and scoring, leveraging in-database analytics.
5. Adopt a marketing platform for campaign development.
6. Bring all that capability together to deliver real-time offer management across all marketing channels: call center, mobile, Web, and in-app.

Completing those steps meant finding the right technology partner. SiriusXM chose Teradata because its strengths were a strong match for the project and company. Teradata offered the ability to:

- Consolidate data sources with an integrated data warehouse (IDW), advanced analytics, and powerful marketing applications.

- Solve data latency issues.
- Significantly reduce data movement across multiple databases and applications.
- Seamlessly interact with applications and modules for all of the marketing areas.
- Scale and perform at very high levels for running campaigns and analytics in-database.
- Conduct real-time communications with customers.
- Provide operational support, either via the cloud or on-premise.

This partnership has enabled SiriusXM to move smoothly and swiftly along its road map, and the company is now in the midst of a transformational, 5-year process. After establishing its strong data governance process, SiriusXM began by implementing its Integrated Data Warehouse, which allowed the company to quickly and reliably operationalize insights throughout the organization.

Next, the company implemented Customer Interaction Manager—part of the Teradata Integrated Marketing Cloud, which enables real-time, dialog-based customer interaction across the full spectrum of digital and traditional communication channels. And, SiriusXM will incorporate the Teradata Digital Messaging Center.

Together, the suite of capabilities will allow SiriusXM to handle direct communications across multiple channels. This evolution will enable real-time offers, marketing messages and recommendations based on previous behavior.

In addition to streamlining how they execute and optimize outbound marketing activities, SiriusXM is also taking control of their internal marketing operations with the implementation of Marketing Resource Management, also part of the Teradata Integrated Marketing Cloud. The solution will allow SiriusXM to streamline workflow, optimize marketing resources, and drive efficiency through every penny of their marketing budget.

Results: Reaping the Benefits

As the company continues its evolution into a high-performance marketing organization, already SiriusXM is benefiting from its thoughtfully executed strategy. Household-level consumer insights and a complete view of marketing touch strategy with each consumer enable SiriusXM to create more targeted offers at the household, consumer, and device levels. By bringing the data and marketing analytics capabilities in-house, SiriusXM achieved the following:

- Campaign results in near real-time rather than 4 days, resulting in massive reductions in cycle times for campaigns and the analysts that support them.
- Closed-loop visibility allowing the analysts to support multistage dialogs and in-campaign modifications to increase campaign effectiveness.
- Real-time modeling and scoring to increase marketing intelligence and sharpen campaign offers and responses at the speed of their business.

Finally, SiriusXM's experience has reinforced the idea that high-performance marketing is a constantly evolving concept. The company has implemented both processes and the technology that give it the capacity for continued and flexible growth.

QUESTIONS FOR THE OPENING VIGNETTE

1. What does SiriusXM do? In what type of market does it conduct its business?
2. What were the challenges? Comment on both technology and data-related challenges.
3. What were the proposed solutions?
4. How did they implement the proposed solutions? Did they face any implementation challenges?

5. What were the results and benefits? Were they worth the effort/investment?
6. Can you think of other companies facing similar challenges that can potentially benefit from similar data-driven marketing solutions?

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 the value proposition to its existing and potential customers. As is the case in any industry, in entertainment, success or mere survival 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 they conducted business: being proactive about the changing nature of the clientele, 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 nowadays, regardless of the industry in which they are participating.

Sources: Quinn, C. (2016). Data-driven marketing at SiriusXM. Teradata Articles & News. at http://bigdata.teradata.com/US/Articles-News/Data-Driven-Marketing-At-SiriusXM/ (accessed August 2016); Teradata customer success story. SiriusXM attracts and engages a new generation of radio consumers. http://assets.teradata.com/resourceCenter/downloads/CaseStudies/EB8597.pdf?processed=1.

2.2 The Nature of Data

Data is the main ingredient for any BI, data science, and business analytics initiative. In fact, it can be viewed as the raw material for what these 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 is of the essence. Once perceived as a big challenge to collect, store, and manage, data nowadays is 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 it can be very large. It can be structured (nicely organized for computers to process), or it can be unstructured (e.g., text that is created for humans and hence not readily understandable/consumable by computers). It can come in smaller batches continuously or it 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 it more challenging to process and consume, it also makes it more valuable because it enriches the data beyond its conventional limits, allowing for the discovery of new and novel knowledge. Traditional ways to manually collect data (either via surveys or via human-entered business transactions) mostly left their places to modern-day data collection mechanisms that use Internet and/or sensor/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 2.1 illustrates a typical analytics continuum—data to analytics to actionable information.

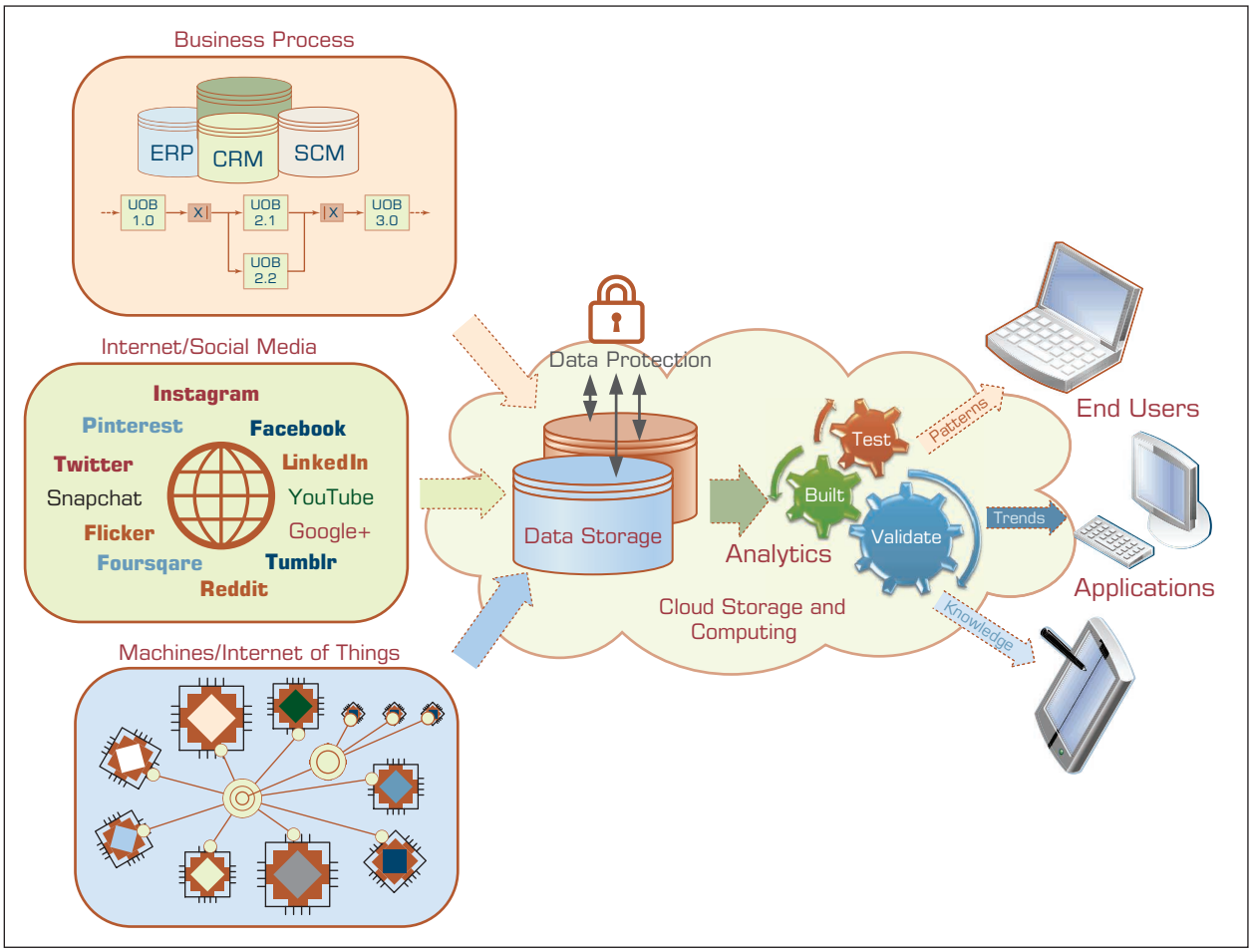


FIGURE 2.1 A Data to Knowledge Continuum.

Although its value proposition is undeniable, to live up its promise, the data has to comply with some basic usability and quality metrics. Not all data is useful for all tasks, obviously. That is, data has to match with (have the coverage of the specifics for) the task for which it is intended to be used. Even for a specific task, the relevant data on hand needs to comply with the quality and quantity requirements. Essentially, data has 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 data 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 you 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 may depend 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) and decision tree algorithms do not require such numerical transformation, easily and natively handling 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 may 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** refers to the originality and appropriateness of the storage medium where the data is obtained—answering the question of "Do we have the right confidence and belief in this data source?" If it 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 it 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** 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 in a record within a database should be the same as what the patient said it was. Data accuracy will be covered in more detail in the following subsection.
- **Data accessibility** 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 may be tricky, especially if the data is stored in more than one location and storage medium and need to be merged/transformed while accessing and obtaining it. As the traditional relational database management systems leave their place (or coexist with) a new generation of data storage mediums like data lakes and Hadoop infrastructure, the importance/criticality of data accessibility is also increasing.
- **Data security and data privacy** means that the data is secured to only allow those people who have the authority and the need to access it and to prevent anyone else from reaching it. Increasing popularity in educational degrees and certificate programs for Information Assurance is an evidence to 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 like Health Insurance Portability and Accountability Act [HIPPA]) but also accurately identifies each patient to allow proper and timely access to records by authorized users (Annas, 2003).
- **Data richness** means that all the 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** 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 may find themselves in the same record—having two different patient records mixed up—for instance, it may happen while merging the demographic and clinical test result data records.

- **Data currency/data timeliness** means that the data should be up-to-date (or as recent/new as it needs to be) for a given analytics model. It also means that the data is 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 rely 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** 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 is aggregated, it may 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 is aggregated cannot be disaggregated (without access to the original source), but it can easily be aggregated from its granular representation.
- **Data validity** 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** 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 may choose to include only the most relevant information (i.e., variables) or if the algorithm is capable enough to sort them out, may choose to include all the relevant ones, regardless of their relevancy level. One thing that analytics studies should avoid is to include totally irrelevant data into the model building, as this may contaminate the information for the algorithm, resulting in inaccurate and misleading results.

Although these 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 paid on these metric dimensions and perhaps add more specific ones to this collection. The following section will dive into the nature of data from a taxonomical perspective to list and define different data types as they relate to different analytics projects.

SECTION 2.2 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 does 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?

2.3 A Simple Taxonomy of Data

Data (**datum** in singular form) refers to a collection of facts usually obtained as the result of experiments, observations, transactions, or experiences. Data may 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 is 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** is 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 2.2 shows a simple **data taxonomy**.

- **Categorical data** 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 may also be considered in a numerical manner by using exact values for age and highest grade completed, it is often more informative to categorize such variables into a relatively small number of ordered classes. The categorical data may also be called discrete data, implying that it represents 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** 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 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).

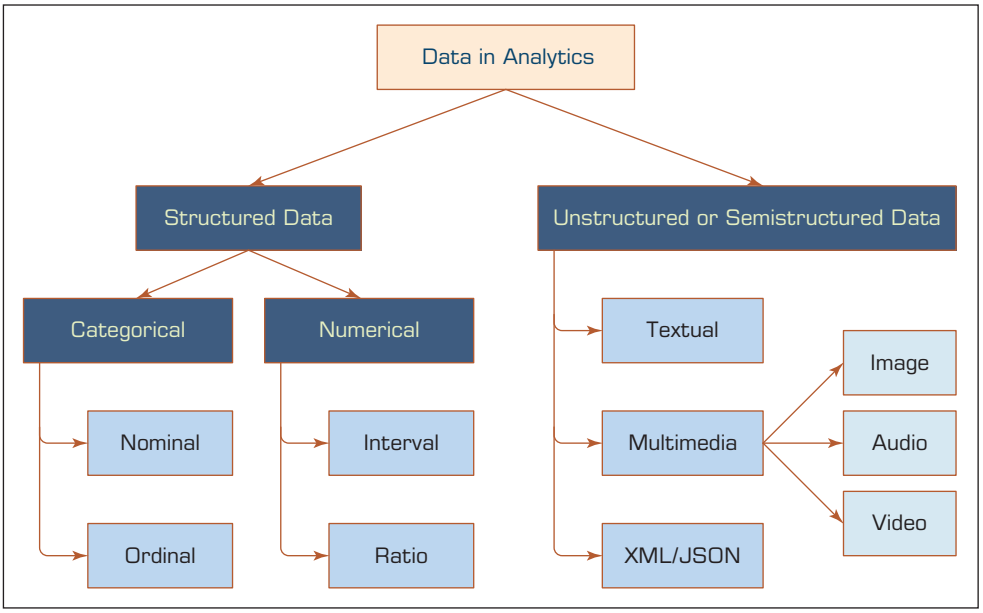


FIGURE 2.2 A Simple Taxonomy of Data.

- **Ordinal data** 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** 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 integer (taking only whole numbers) or real (taking also the fractional number). The numeric data may 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** 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** 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 may lead to incorrect models or (more often) halt the model development process. For example, some data mining methods need all the variables (both input as well as 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 may 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 comes 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 2.1 shows a business scenario in which a data-rich medical device research and development company streamlined their analytics practices to have easy access to both the data and the analyses they need to continue the traditions of innovation and quality at the highest levels.

Application Case 2.1

Medical Device Company Ensures Product Quality While Saving Money

Few technologies are advancing faster than those in the medical field—so having the right advanced analytics software can be a game changer. Instrumentation Laboratory is a leader in the development, manufacturing, and distribution of medical devices and related technologies, including technology that is revolutionizing whole blood and hemostasis testing. To help ensure its continued growth and success, the company relies on data analytics and Dell Statistica.

Problem

As a market leader in diagnostic instruments for critical care and hemostasis, Instrumentation Laboratory must take advantage of rapidly evolving technologies while maintaining both quality and efficiency in its product development, manufacturing, and distribution processes. In particular, the company needed to enable its research and development (R&D) scientists and engineers to easily access and analyze the wealth of test data it collects, as well as efficiently monitor its manufacturing processes and supply chains.

"Like many companies, we were data-rich but analysis-poor," explains John Young, Business Analyst for Instrumentation Laboratory. "It's no longer viable to have R&D analysts running off to IT every time they need access to test data and then doing one-off analyses in Minitab. They need to be able to access data quickly and perform complex analyses consistently and accurately."

Implementing sophisticated analytics was especially critical for Instrumentation Laboratory because of the volume and complexity of its products. For example, every year, the company manufactures hundreds of thousands of cartridges containing a card with a variety of sensors that measure the electrical signals of the blood being tested.

"Those sensors are affected by a wide range of factors, from environmental changes like heat and humidity to inconsistencies in materials from suppliers, so we're constantly monitoring their performance," says Young. "We collect millions of records of data, most of which is stored in SQL Server databases. We needed an analytics platform that would enable our R&D teams to quickly access that data and troubleshoot any problems. Plus, because there are so many factors in play, we also needed a platform that could intelligently monitor the test data and alert us to emerging issues automatically."

Solution

Instrumentation Laboratory began looking for an analytics solution to meet its needs. The company quickly eliminated most tools on the market because they failed to deliver the statistical functionality and level of trust required for the healthcare environment. That left only two contenders: another analytics solution and Dell Statistica. For Instrumentation Laboratory, the clear winner was Statistica.

"Choosing Statistica was an easy decision," recalls Young. "With Statistica, I was able to quickly build a wide range of analysis configurations on top of our data for use by analysts enterprise-wide. Now, when they want to understand specific things, they can simply run a canned analysis from that central store instead of having to ask IT for access to the data or remember how to do a particular test."

(*Continued*)

Application Case 2.1 (Continued)

Moreover, Statistica was far easier to deploy and use than legacy analytics solutions. "To implement and maintain other analytics solutions, you need to know analytics solutions programming," Young notes. "But with Statistica, I can connect to our data, create an analysis and publish it within an hour—even though I'm not a great programmer."

Finally, in addition to its advanced functionality and ease of use, Statistica delivered world-class support and an attractive price point. "The people who helped us implement Statistica were simply awesome," says Young. "And the price was far below what another analytics solution was quoting."

Results

With Statistica in place, analysts across the enterprise now have easy access to both the data and the analyses they need to continue the twin traditions of innovation and quality at Instrumentation Laboratory. In fact, Statistica's quick, effective analysis and automated alerting is saving the company hundreds of thousands of dollars.

"During cartridge manufacturing, we occasionally experience problems, such as an inaccuracy in a chemical formulation that goes on one of the sensors," Young notes. "Scrapping a single batch of cards would cost us hundreds of thousands of dollars. Statistica helps us quickly figure out what went wrong and fix it so we can avoid those costs. For example, we can marry the test data with electronic device history record data from our SAP environment and perform all sorts of correlations to determine which particular changes—such as changes in temperature and humidity—might be driving a particular issue."

Manual quality checks are, of course, valuable, but Statistica runs a variety of analyses automatically for the company as well, helping to ensure that nothing is missed and issues are identified quickly. "Many analysis configurations are scheduled to run periodically to check different things," Young says. "If there is an issue, the system automatically emails the appropriate people or logs the violations to a database."

Some of the major benefits of advanced data analytics with Dell Statistica included the following:

- *Regulatory compliance.* In addition to saving Instrumentation Laboratory money, Statistica also helps ensure the company's processes comply with Food and Drug Administration (FDA) regulations for quality and consistency. "Because we manufacture medical devices, we're regulated by the FDA," explains Young. "Statistica helps us perform the statistical validations required by the FDA—for example, we can easily demonstrate that two batches of product made using different chemicals are statistically the same."
- *Ensuring consistency.* Creating standardized analysis configurations in Statistica that can be used across the enterprise helps ensure consistency and quality at Instrumentation Laboratory. "You get different results depending on the way you go about analyzing your data. For example, different scientists might use different trims on the data, or not trim it at all—so they would all get different results," explains Young. "With Statistica, we can ensure that all the scientists across the enterprise are performing the analyses in the same way, so we get consistent results."
- *Supply chain monitoring.* Instrumentation Laboratory manufactures not just the card with the sensors but the whole medical instrument, and therefore it relies on suppliers to provide parts. To further ensure quality, the company is planning to extend its use of Statistica to supply chain monitoring.
- *Saving time.* In addition to saving money and improving regulatory compliance for Instrumentation Laboratory, Statistica is also saving the company's engineers and scientists valuable time, enabling them to focus more on innovation and less on routine matters. "Statistica's proactive alerting saves engineers a lot of time because they don't have to remember to check various factors, such as glucose slope, all the time. Just that one test would take half a day," notes Young. "With Statistica monitoring our test data, our engineers can focus on other matters, knowing they will get an email if and when a factor like glucose slope becomes an issue."

Future Possibilities

Instrumentation Laboratory is excited about the opportunities made possible by the visibility Statistica advanced analytics software has provided into its

data stores. "Using Statistica, you can discover all sorts of insights about your data that you might not otherwise be able to find," says Young. "There might be hidden pockets of money out there that you're just not seeing because you're not analyzing your data to the extent you could. Using the tool, we've discovered some interesting things in our data that have saved us a tremendous amount of money, and we look forward to finding even more."

Questions for Discussion

1. What were the main challenges for the medical device company? Were they market or technology driven? Explain.
2. What was the proposed solution?
3. What were the results? What do you think was the real return on investment (ROI)?

Source: Dell customer case study. Medical device company ensures product quality while saving hundreds of thousands of dollars. https://software.dell.com/documents/instrumentation-laboratory-medical-device-companyensures-product-quality-while-saving-hundreds-ofthousands-of-dollars-case-study-80048.pdf (accessed August 2016). Used by Permission from Dell.

SECTION 2.3 REVIEW QUESTIONS

1. What is data? How does 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?

2.4 The Art and Science of Data Preprocessing

Data in its original form (i.e., the real-world data) is not usually ready to be used in analytics tasks. It is 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 2.3 shows the main steps in the data preprocessing endeavor.

In the first phase of data preprocessing, the relevant data is 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 phase of data preprocessing, the data is cleaned (this step is also known as data scrubbing). Data in its original/raw/real-world form is usually dirty (Hernández & Stolfo, 1998; Kim et al., 2003). In this step, 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 (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.

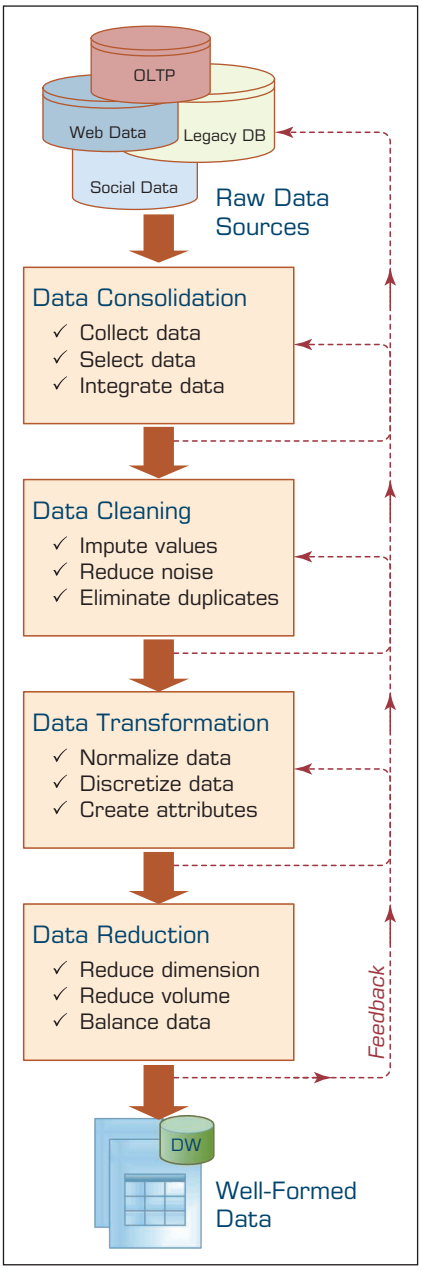


FIGURE 2.3 Data Preprocessing Steps.

In addition, inconsistencies (unusual values within a variable) in the data should be handled using domain knowledge and/or expert opinion.

In the third phase of data preprocessing, the data is transformed for better processing. For instance, in many cases the data is normalized between a certain minimum and maximum for all variables to mitigate the potential bias of one variable (having large numeric values, such as for household income) dominating other variables (such as *number of dependents* or *years in service,* which may potentially 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 may 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 may 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 may 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 down 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 may include millions or billions of records. Even though computing power is increasing exponentially, processing such a large number of records may not be practical or feasible. In such cases, one may 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 may hold well, but real-world data is 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 is usually sorted on some variable, and taking a section of the data from the top or bottom may 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 may not be sufficient, and stratified sampling (a proportional representation of different subgroups in the data is represented in the sample data set) may 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 2.1, which maps the main phases (along with their problem descriptions) to a representative list of tasks and algorithms.

It is almost impossible to underestimate the value proposition of data preprocessing. It is one of those time-demanding activities where 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 2.2 illustrates an interesting study where raw, readily available academic data within an educational organization is used to develop predictive models to better understand attrition and improve freshmen

TABLE 2.1 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	Principal component analysis, independent component analysis, chi-square testing, correlation analysis, and decision tree induction.
	Reduce number of records	Random sampling, stratified sampling, expert-knowledge-driven purposeful sampling.
	Balance skewed data	Oversample the less represented or undersample the more represented classes.

Application Case 2.2

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, 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 has 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 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.

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 that 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 where 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 2.4.

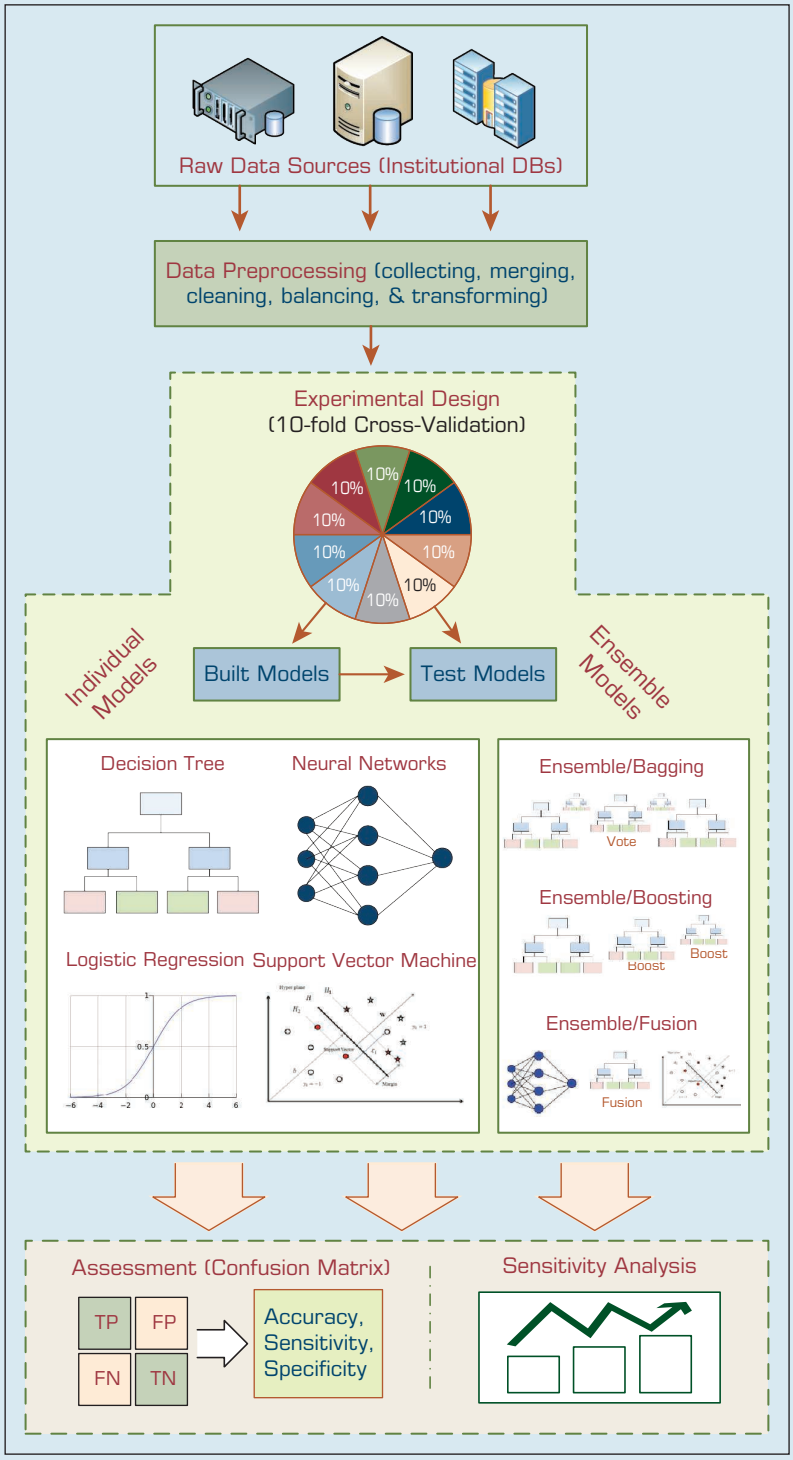


FIGURE 2.4 An Analytics Approach to Predicting Student Attrition.

(Continued)

Application Case 2.2 (Continued)

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 among the 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 Is 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% are the residents of the same state and roughly 19% 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%, and the average 6-year graduation rate was about 60%.

The study used 5 years of institutional data, which entailed to 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) 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).

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

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

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%) than *no* records (~20%; see Figure 2.5).

Research shows that having such an 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.

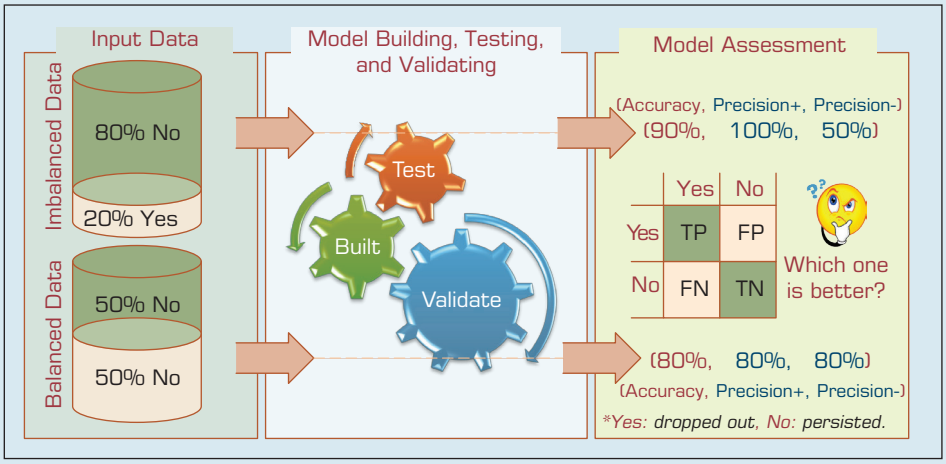


FIGURE 2.5 A Graphical Depiction of the Class Imbalance Problem.

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 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.

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%, the decision tree came out as the runner-up with an overall prediction rate of 87.16%, followed by artificial neural networks and logistic regression with overall prediction rates of 86.45% and 86.12%, respectively (see Table 2.2). A careful examination of these results reveals that the predictions 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% accuracy, but they did poorly on predicting the students who are likely to drop out after the freshman year with less than 50% accuracy. Because the prediction of the "No" class is the main purpose of this study, less than 50% accuracy for this class was deemed not acceptable. Such a difference

TABLE 2.2 Prediction Results for the Original/Unbalanced Dataset

	ANN(MLP)		DT(C5)		SVM		LR	
	No	Yes	No	Yes	No	Yes	No	Yes
No	1494	384	1518	304	1478	255	1438	376
Yes	1596	11142	1572	11222	1612	11271	1652	11150
SUM	3090	11526	3090	11526	3090	11526	3090	11526
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%	

(*Continued*)

Application Case 2.2 (Continued)

TABLE 2.3 Prediction Results for the Balanced Data Set

Confusion Matrix	ANN(MLP)		DT(C5)		SVM		LR	
	No	Yes	No	Yes	No	Yes	No	Yes
No	2309	464	2311	417	2313	386	2125	626
Yes	781	2626	779	2673	777	2704	965	2464
SUM	3090	3090	3090	3090	3090	3090	3090	3090
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%	

in prediction accuracy of the two classes can (and should) be attributed to the imbalanced nature of the training data set (i.e., ~80% "Yes" and ~20% "No" samples).

The next round of experiments used a well-balanced data set where the two classes are represented nearly equally in counts. In realizing this approach, the study took all the samples from the minority class (i.e., the "No" class herein) and randomly selected an equal number of samples from the majority class (i.e., the "Yes" class herein) and repeated this process for 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 2.3. Based on the holdout sample results, support vector machines once again generated the best overall prediction accuracy with 81.18%, followed by decision trees, artificial neural networks, and logistic regression with an overall prediction accuracy of 80.65%, 79.85%, and 74.26%. 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 were 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%, followed by the bagging-type ensembles and boosting-type ensembles with overall prediction rates of 81.80% and 80.21%, respectively (see Table 2.4). Even though the prediction results are slightly better than 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

TABLE 2.4 Prediction Results for the Three Ensemble Models

	Boosting (Boosted Trees)		Bagging (Random Forest)		Information Fusion (Weighted Average)	
	No	Yes	No	Yes	No	Yes
No	2242	375	2327	362	2335	351
Yes	848	2715	763	2728	755	2739
SUM	3090	3090	3090	3090	3090	3090
Per-Class Accuracy	72.56%	87.86%	75.31%	88.28%	75.57%	88.64%
Overall Accuracy	80.21%		81.80%		82.10%	

overall sensitivity analysis results, each of the four individual model types generated its own sensitivity measures ranking all the 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 2.6).

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% accuracy. Results also showed that, regardless of the prediction model employed, the balanced data set (compared to unbalanced/original data set) produced better prediction

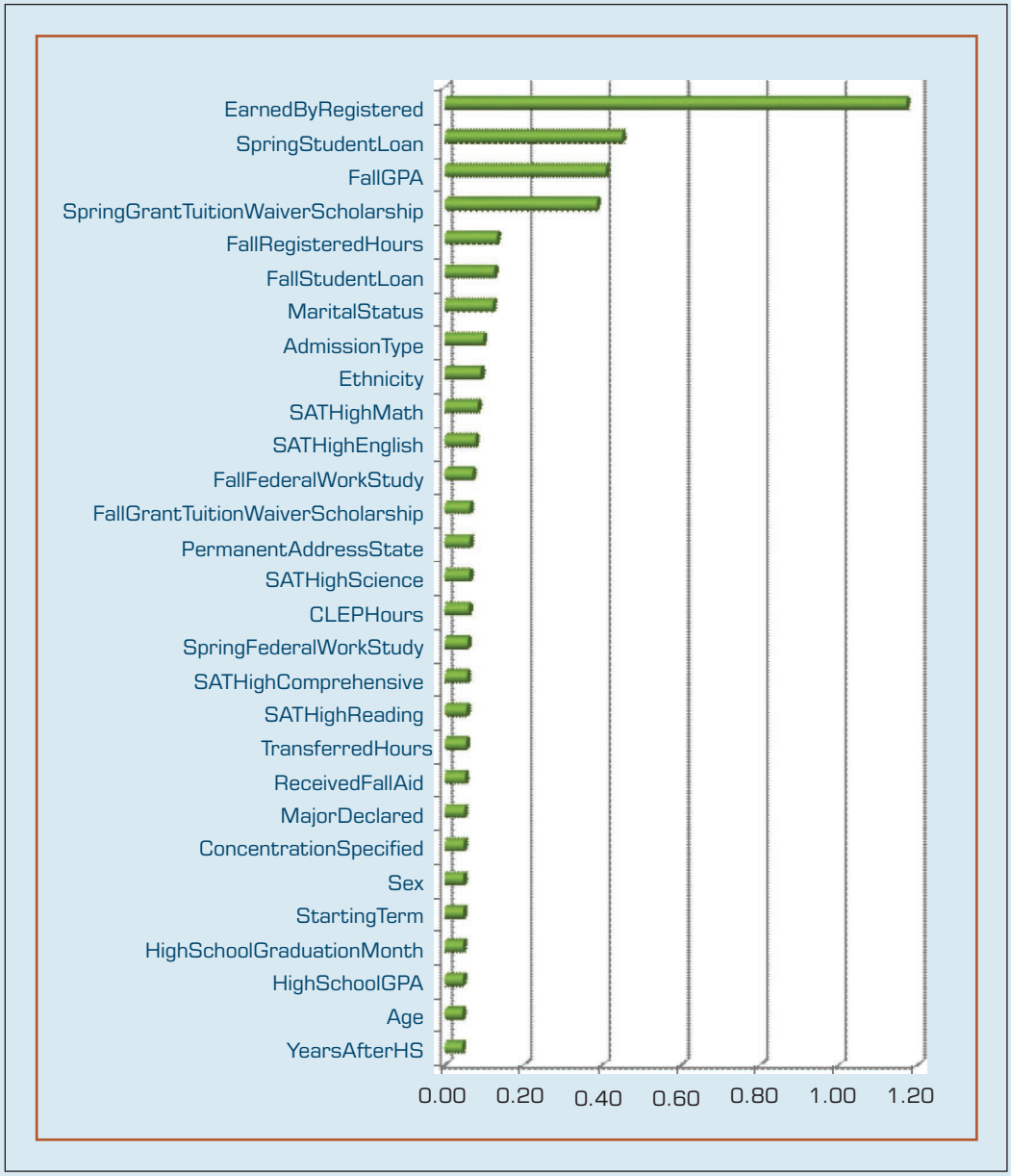


FIGURE 2.6 Sensitivity-Analysis-Based Variable Importance Results.

(*Continued*)

Application Case 2.2 (Continued)

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 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 Discussion

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 context of this case study.
4. What was the proposed solution? And, what were the results?

Sources: Thammasiri, D., Delen, D., Meesad, P., & Kasap N. (2014). A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. *Expert Systems with Applications, 41*(2), 321–330; Delen, D. (2011). Predicting student attrition with data mining methods. *Journal of College Student Retention, 13*(1), 17–35; Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems, 49*(4), 498–506.

student retention in a large higher education institution. As the application case clearly states, each and every data preprocessing task described in Table 2.1 was critical to a successful execution of the underlying analytics project, especially the task that related to the balancing of the data set.

SECTION 2.4 REVIEW QUESTIONS

1. Why is 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?

2.5 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 around, 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 2.7). 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 2.7, descriptive

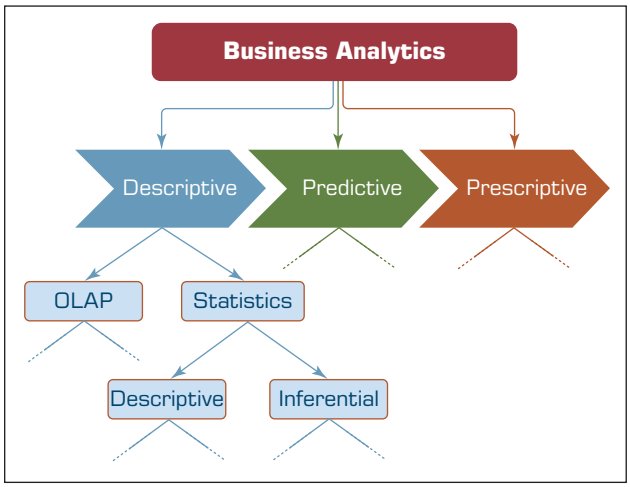


FIGURE 2.7 Relationship between Statistics and Descriptive Analytics.

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—details of data warehousing will be covered in Chapter 3) 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 multivariables 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 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, and **inferential statistics** is about drawing inferences or conclusions about the characteristics of the population. In this section we will 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 will 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 not only helps business users in their decision-making processes, but also helps 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 will 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 (May Also Be 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 are 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 is erratically

dispersed and skewed, one should either avoid using 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 observation 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, then 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, half the data points in the data are above and the other half are 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 may be considered as being 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 may 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 is not prone to outliers and there is no significant level of skewness; use the median when the data has outliers and/or it is ordinal in nature; use the mode when the data is 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 like mean.

Measures of Dispersion (May Also Be 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 are a representation of 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 it gives us a framework within which we can judge the central tendency—gives 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 ith 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, herein 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, it is calculated by measuring the absolute values of the differences between each data point and the mean and summing them. It 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 plot (or box-and-whiskers 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 2.8 shows a couple of box plots side by side, sharing the same *y*-axis. As shown therein, a single

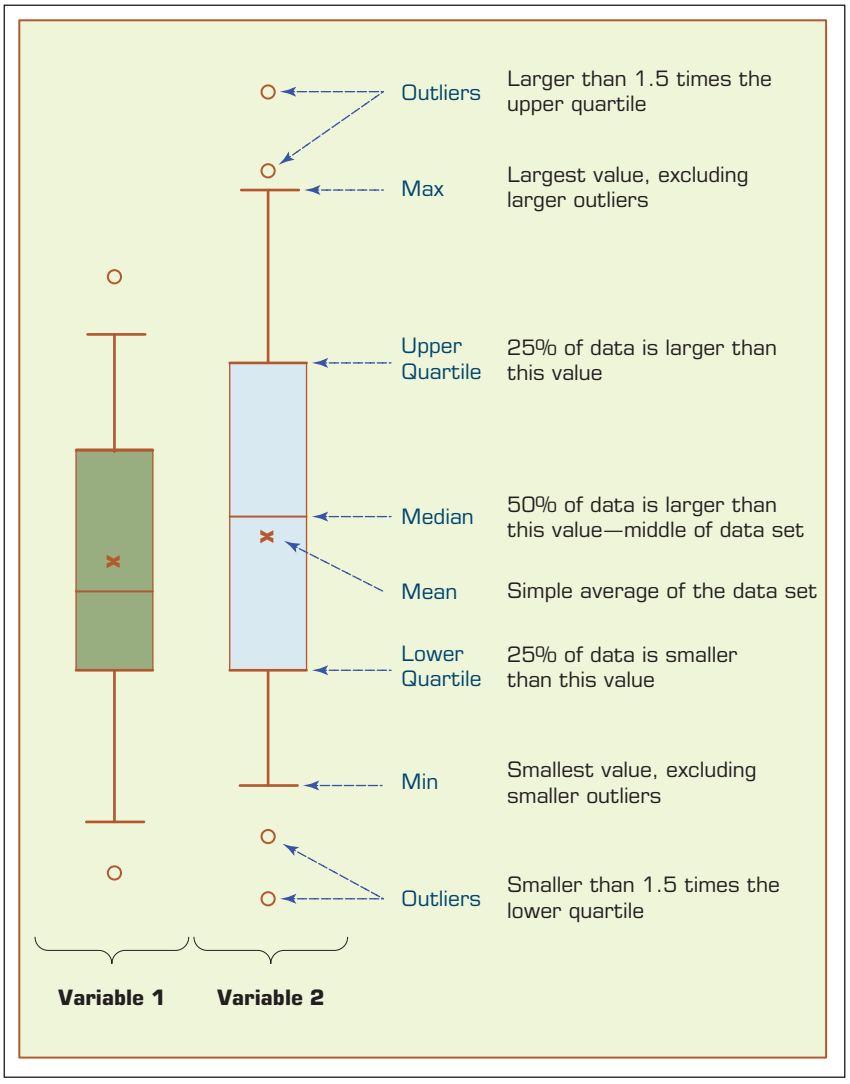


FIGURE 2.8 Understanding the Specifics about Box-and-Whiskers Plots.

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 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 was not 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 quartile), 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) along with the outliers that are larger than the limits of the whiskers. A box plot also shows whether the data is symmetrically distributed with respect to the mean or it sways 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.

The 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 %), 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 2.9.

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 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., tail is on the right side and the mean is smaller than median), then it produces a positive skewness measure, and 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 2.9, (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.

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

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

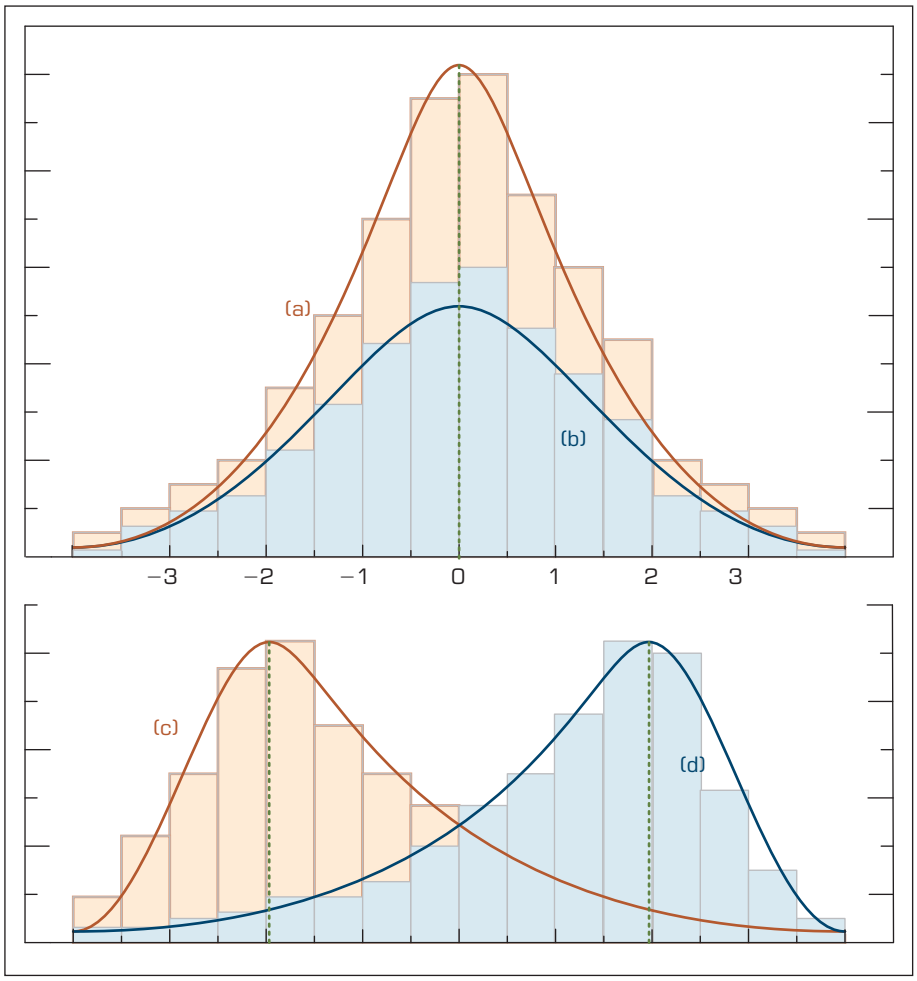


FIGURE 2.9 Relationship between Dispersion and Shape Properties.

Kurtosis is another measure to use in characterizing the shape of a unimodal distribution. As opposed to the sway in shape, kurtosis is more interested in 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 2.1 describes in detail how to use Microsoft Excel to calculate descriptive statistics.

TECHNOLOGY INSIGHTS 2.1
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 installation and can be activated (turned on) with only a few mouse clicks. Figure 2.10 shows how these statistics functions (as part of the Analysis ToolPak) can be activated in Microsoft Excel 2016.

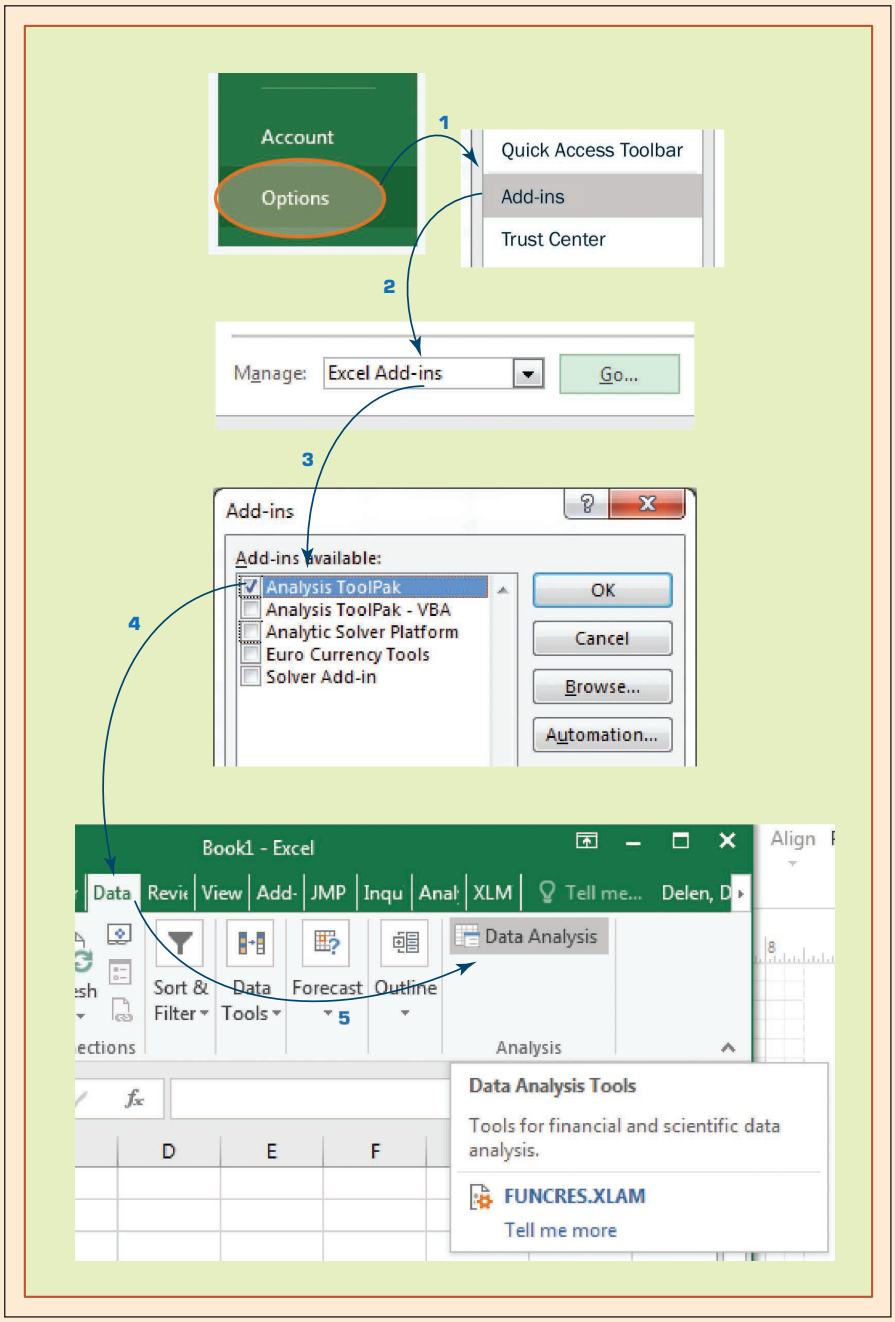
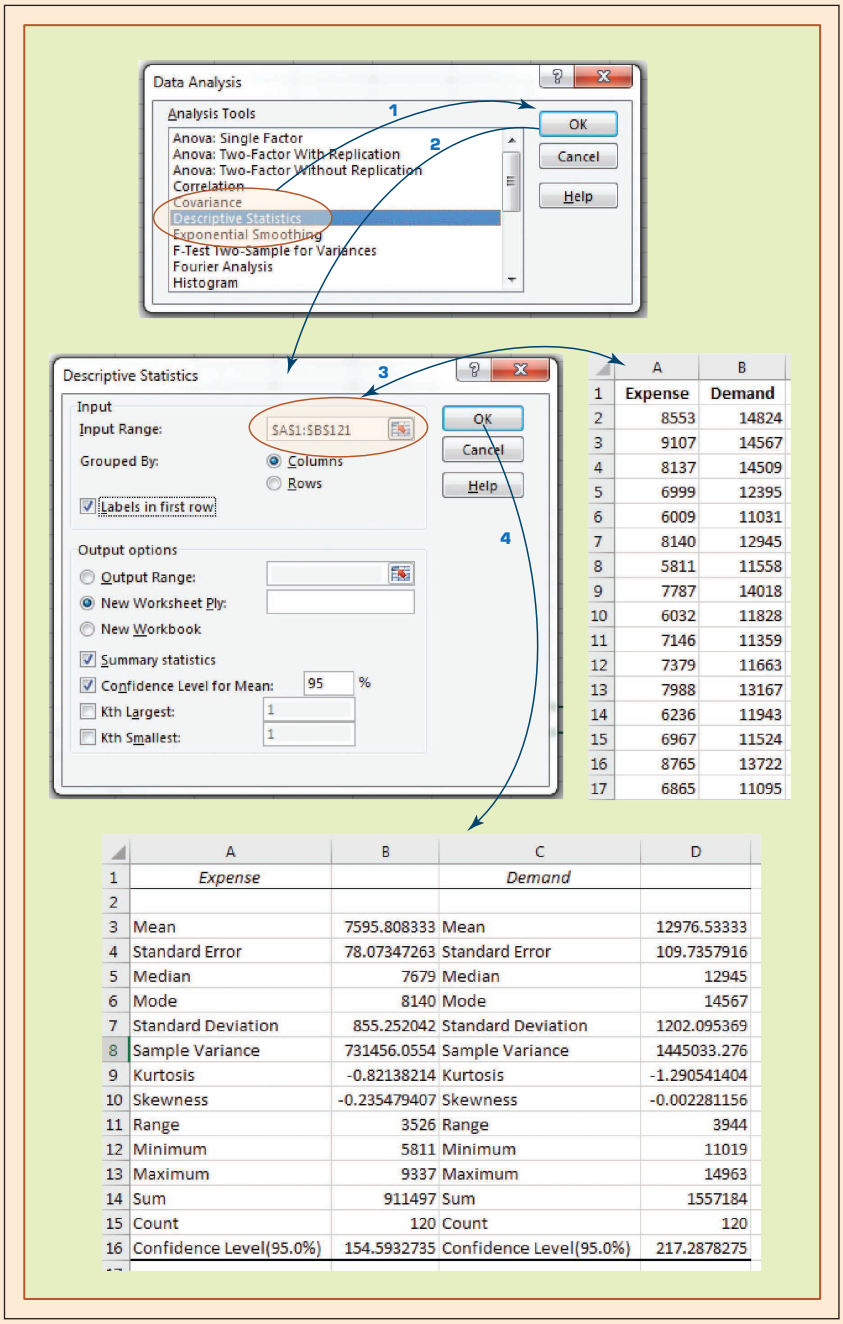


FIGURE 2.10 Activating Statistics Function in Excel 2016.

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 2.11, steps [1, 2]); click on OK, and the Descriptive Statistics dialog box will appear (see middle of Figure 2.11). 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 2.11, 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.

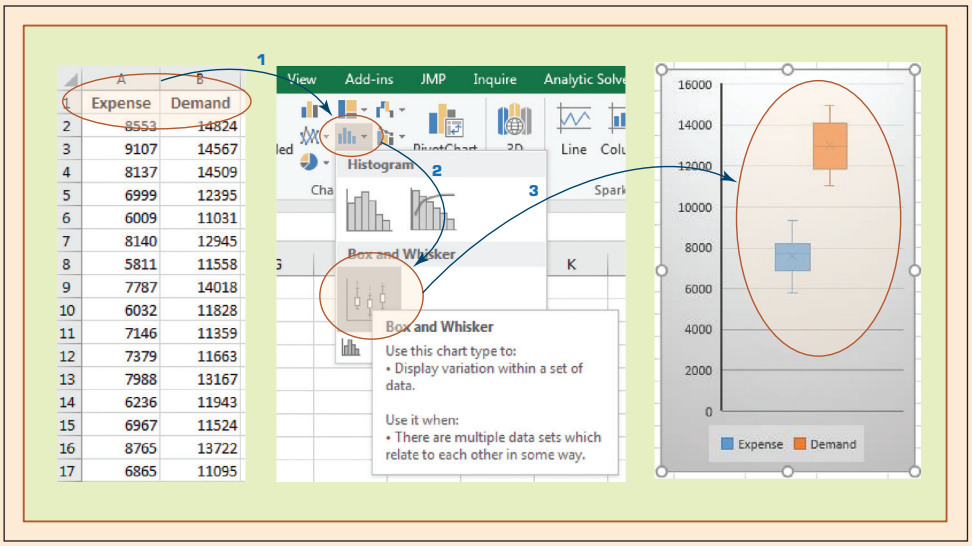


	A	B
1	Expense	Demand
2	8553	14824
3	9107	14567
4	8137	14509
5	6999	12395
6	6009	11031
7	8140	12945
8	5811	11558
9	7787	14018
10	6032	11828
11	7146	11359
12	7379	11663
13	7988	13167
14	6236	11943
15	6967	11524
16	8765	13722
17	6865	11095

	A	B	C	D
1	*Expense*		*Demand*	
2				
3	Mean	7595.808333	Mean	12976.53333
4	Standard Error	78.07347263	Standard Error	109.7357916
5	Median	7679	Median	12945
6	Mode	8140	Mode	14567
7	Standard Deviation	855.252042	Standard Deviation	1202.095369
8	Sample Variance	731456.0554	Sample Variance	1445033.276
9	Kurtosis	-0.82138214	Kurtosis	-1.290541404
10	Skewness	-0.235479407	Skewness	-0.002281156
11	Range	3526	Range	3944
12	Minimum	5811	Minimum	11019
13	Maximum	9337	Maximum	14963
14	Sum	911497	Sum	1557184
15	Count	120	Count	120
16	Confidence Level(95.0%)	154.5932735	Confidence Level(95.0%)	217.2878275

FIGURE 2.11 Obtaining Descriptive Statistics in Excel.

(Continued)



	A	B
1	Expense	Demand
2	8553	14824
3	9107	14567
4	8137	14509
5	6999	12395
6	6009	11031
7	8140	12945
8	5811	11558
9	7787	14018
10	6032	11828
11	7146	11359
12	7379	11663
13	7988	13167
14	6236	11943
15	6967	11524
16	8765	13722
17	6865	11095

FIGURE 2.12 Creating a Box-and-Whiskers Plot in Excel 2016.

As a simple example, we selected two columns (labeled as Expense and Demand) and executed the Descriptive Statistics option. The bottom section of Figure 2.11 shows the output created by Excel. As can be seen, Excel produced all the 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 mouse clicks) to create a box-and-whiskers plot. Figure 2.12 shows the simple three-step process of creating a box-and-whiskers plot in Excel.

Although this is a very useful tool in Excel, one should be aware of an important point related to the results generated by Analysis ToolPak, 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 changes, 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 the problems once thought unsolvable. Application Case 2.3 is an excellent illustration of those success stories where a small municipality administration adopts a data analytics approach to intelligently detect and solve problems by continuously analyzing demand and consumption patterns.

Application Case 2.3

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 the 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's 8,760 data points per customer each year instead of 12 monthly readings. The data had tremendous potential, if it 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 the Town of Cary. "SAS helps us manage the volume of that data nicely." In fact, the solution enables the town to analyze a half-billion data points on water usage and make them available, and easily consumable, to 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 of Cary 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 in 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 the Town of 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 Discussion

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.

SECTION 2.5 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?

2.6 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, essentially, regression 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 may 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 variables 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 prediction 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.

SIMPLE VERSUS MULTIPLE REGRESSION If the regression equation is built between one response variable and one explanatory variable, then it is called simple regression. For instance, the regression equation built to predict/explain the relationship between a height of a person (explanatory variable) and the weight of a person (response variable) is a good example of simple regression. Multiple regression is the extension of simple regression where the explanatory variables are more than one. For instance, in the previous example, if we were to include not only the height of the person but also other personal characteristics (e.g., BMI, gender, ethnicity) to predict the weight of a person, then we would be performing multiple regression analysis. In both cases, the relationship between the response variable and the explanatory variable(s) are linear and additive in nature. If the relationships are not linear, then we may want to use one of many other nonlinear regression methods to better capture the relationships between the input and output variables.

How Do We Develop the Linear Regression Model?

To understand the relationship between two variables, the simplest thing that one can do is to draw a scatter plot, where the *y*-axis represents the values of the response variable and the *x*-axis represents the values of the explanatory variable (see Figure 2.13). A scatter catter plot would show the changes in the response variable as a function of the changes in the explanatory variable. In the case shown in Figure 2.13, there seems to be a positive relationship between the two; as the explanatory variable values increase, so does the response variable.

Simple regression analysis aims to find a mathematical representation of this relationship. In reality, it tries to find the signature of a straight line passing through right between the plotted dots (representing the observation/historical data) in such a way that it minimizes the distance between the dots and the line (the predicted values on the

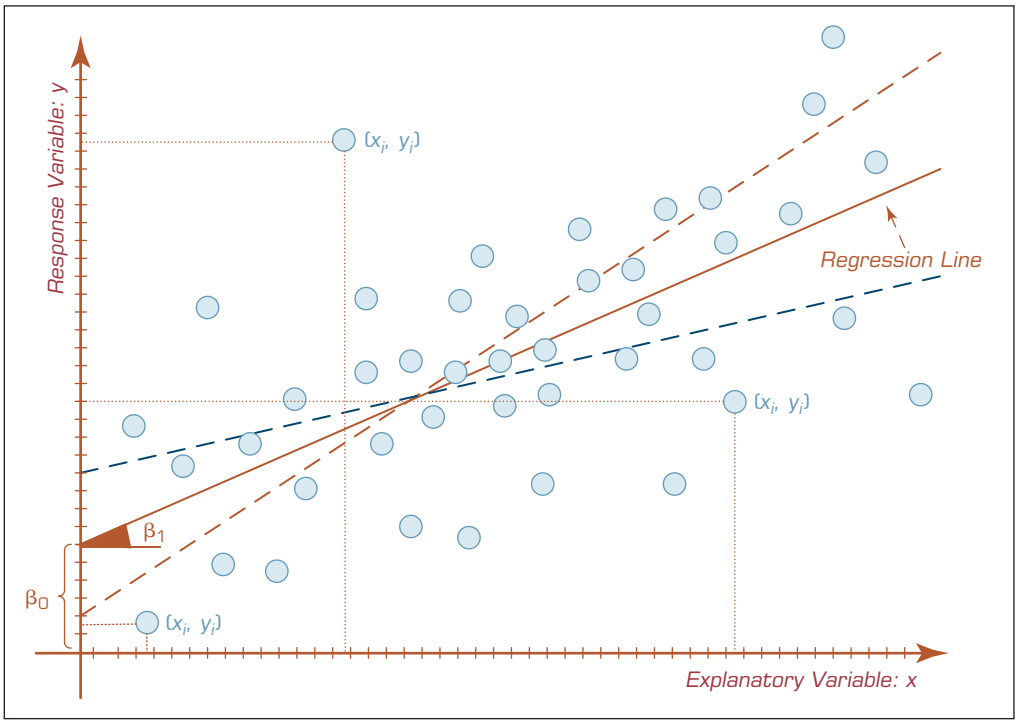


FIGURE 2.13 A Scatter Plot and a Linear Regression Line.

theoretical regression line). Even though there are several methods/algorithms proposed to identify the regression line, the one that is most commonly used is called the **ordinary least squares** (**OLS**) method. The OLS method aims to minimize the sum of squared residuals (squared vertical distances between the observation and the regression point) and leads to a mathematical expression for the estimated value of the regression line (which are known as β parameters). For simple **linear regression**, the aforementioned relationship between the response variable (y) and the explanatory variable(s) (x) can be shown as a simple equation as follows:

$$y = \beta_0 + \beta_1 x$$

In this equation, β_0 is called the intercept, and β_1 is called the slope. Once OLS determines the values of these two coefficients, the simple equation can be used to forecast the values of y for given values of x. The sign and the value of β_1 also reveal the direction and the strengths of relationship between the two variables.

If the model is of a multiple linear regression type, then there would be more coefficients to be determined, one for each additional explanatory variable. As the following formula shows, the additional explanatory variable would be multiplied with the new β_i coefficients and summed together to establish a linear additive representation of the response variable.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_n x_n$$

How Do We Know If the Model Is Good Enough?

Because of a variety of reasons, sometimes models as representations of the reality do not prove to be good. Regardless of the number of explanatory variables included, there is always a possibility of not having a good model, and therefore the linear regression model needs to be assessed for its fit (the degree at which it represents the response variable). In the simplest sense, a well-fitting regression model results in predicted values close to the observed data values. For the numerical assessment, three statistical measures are often used in evaluating the fit of a regression model. R^2 (R-squared), the overall F-test, and the root mean square error (RMSE). All three of these measures are based on the sums of the square errors (how far the data are from the mean and how far the data are from the model's predicted values). Different combinations of these two values provide different information about how the regression model compares to the mean model.

Of the three, R^2 has the most useful and understandable meaning because of its intuitive scale. The value of R^2 ranges from zero to one (corresponding to the amount of variability explained in percentage) with zero indicating that the relationship and the prediction power of the proposed model is not good, and one indicating that the proposed model is a perfect fit that produces exact predictions (which is almost never the case). The good R^2 values would usually come close to one, and the closeness is a matter of the phenomenon being modeled—whereas an R^2 value of 0.3 for a linear regression model in social sciences can be considered good enough, an R^2 value of 0.7 in engineering may be considered as not a good-enough fit. The improvement in the regression model can be achieved by adding mode explanatory variables, taking some of the variables out of the model, or using different data transformation techniques, which would result in comparative increases in an R^2 value. Figure 2.14 shows the process flow of developing regression models. As can be seen in the process flow, the model development task is followed by the model assessment task, where not only is the fit of the model assessed, but because of restrictive assumptions with which the linear models have to comply, also the validity of the model needs to be put under the microscope.

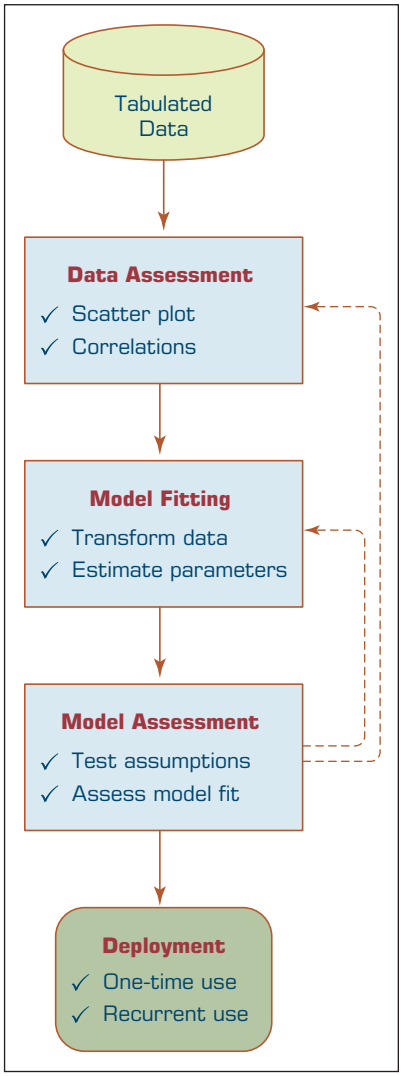


FIGURE 2.14 A Process Flow for Developing Regression Models.

What Are the Most Important Assumptions in Linear Regression?

Even though they are still the choice of many for data analyses (both for explanatory as well as for predictive modeling purposes), linear regression models suffer from several highly restrictive assumptions. The validity of the linear model built depends on its ability to comply with these assumptions. Here are the most commonly pronounced assumptions:

1. **Linearity.** This assumption states that the relationship between the response variable and the explanatory variables are linear. That is, the expected value of the response variable is a straight-line function of each explanatory variable, while holding all other explanatory variables fixed. Also, the slope of the line does not depend on the values of the other variables. It also implies that the effects of different explanatory variables on the expected value of the response variable are additive in nature.
2. **Independence** (of errors). This assumption states that the errors of the response variable are uncorrelated with each other. This independence of the errors is weaker than actual statistical independence, which is a stronger condition and is often not needed for linear regression analysis.

3. **Normality** (of errors). This assumption states that the errors of the response variable are normally distributed. That is, they are supposed to be totally random and should not represent any nonrandom patterns.
4. **Constant variance** (of errors). This assumption, also called homoscedasticity, states that the response variables have the same variance in their error, regardless of the values of the explanatory variables. In practice this assumption is invalid if the response variable varies over a wide enough range/scale.
5. **Multicollinearity.** This assumption states that the explanatory variables are not correlated (i.e., do not replicate the same but provide a different perspective of the information needed for the model). Multicollinearity can be triggered by having two or more perfectly correlated explanatory variables presented to the model (e.g., if the same explanatory variable is mistakenly included in the model twice, one with a slight transformation of the same variable). A correlation-based data assessment usually catches this error.

There are statistical techniques developed to identify the violation of these assumptions and techniques to mitigate them. The most important part for a modeler is to be aware of their existence and to put in place the means to assess the models to make sure that the models are compliant with the assumptions they are built on.

Logistic Regression

Logistic regression is a very popular, statistically sound, probability-based classification algorithm that employs supervised **learning**. It was developed in the 1940s as a complement to linear regression and linear discriminant analysis methods. It has been used extensively in numerous disciplines, including the medical and social sciences fields. Logistic regression is similar to linear regression in that it also aims to regress to a mathematical function that explains the relationship between the response variable and the explanatory variables using a sample of past observations (training data). It differs from linear regression with one major point: its output (response variable) is a class as opposed to a numerical variable. That is, whereas linear regression is used to estimate a continuous numerical variable, logistic regression is used to classify a categorical variable. Even though the original form of logistic regression was developed for a binary output variable (e.g., 1/0, yes/no, pass/fail, accept/reject), the present-day modified version is capable of predicting multiclass output variables (i.e., multinomial logistic regression). If there is only one predictor variable and one predicted variable, the method is called simple logistic regression (similar to calling linear regression models with only one independent variable as simple linear regression).

In predictive analytics, logistic regression models are used to develop probabilistic models between one or more explanatory/predictor variables (which may be a mix of both continuous and categorical in nature) and a class/response variable (which may be binomial/binary or multinomial/multiclass). Unlike ordinary linear regression, logistic regression is used for predicting categorical (often binary) outcomes of the response variable—treating the response variable as the outcome of a Bernoulli trial. Therefore, logistic regression takes the natural logarithm of the odds of the response variable to create a continuous criterion as a transformed version of the response variable. Thus the logit transformation is referred to as the link function in logistic regression—even though the response variable in logistic regression is categorical or binomial, the logit is the continuous criterion on which linear regression is conducted. Figure 2.15 shows a logistic regression function where the odds are represented in the *x*-axis (a linear function of the independent variables), whereas the probabilistic outcome is shown in the *y*-axis (i.e., response variable values change between 0 and 1).

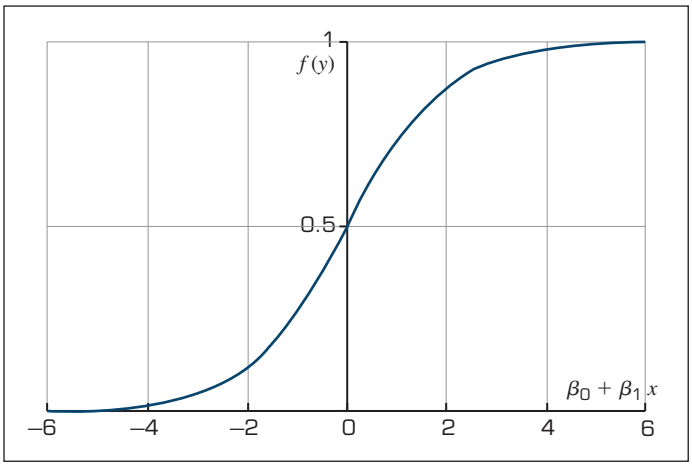


FIGURE 2.15 The Logistic Function.

The logistic function, $f(y)$ in Figure 2.15, is the core of logistic regression, which can only take values between 0 and 1. The following equation is a simple mathematical representation of this function:

$$f(y) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

The logistic regression coefficients (the βs) are usually estimated using the maximum likelihood estimation method. Unlike linear regression with normally distributed residuals, it is not possible to find a closed-form expression for the coefficient values that maximizes the likelihood function, so an iterative process must be used instead. This process begins with a tentative starting solution, then revises the parameters slightly to see if the solution can be improved and repeats this iterative revision until no improvement can be achieved or are very minimal, at which point the process is said to have completed/converged.

Sports analytics—use of data and statistical/analytics techniques to better manage sports teams/organizations—has been gaining tremendous popularity. Use of data-driven analytics techniques have become mainstream for not only professional teams but also college and amateur sports. Application Case 2.4 is an example of how existing and readily available public data sources can be used to predict college football bowl game outcomes using both classification and regression-type prediction models.

Application Case 2.4

Predicting NCAA Bowl Game Outcomes

Predicting the outcome of a college football game (or any sports game, for that matter) is an interesting and challenging problem. Therefore, challenge-seeking researchers from both academics and industry have spent a great deal of effort on forecasting the outcome of sporting events. Large quantities of historic data exist in different media outlets (often publicly available) regarding the structure and outcomes of sporting events in the form of a variety of numerically or symbolically represented factors that are assumed to contribute to those outcomes.

The end-of-season bowl games are very important to colleges both financially (bringing in millions of dollars of additional revenue) as well

(*Continued*)

Application Case 2.4 (Continued)

as reputational—for recruiting quality students and highly regarded high school athletes for their athletic programs (Freeman & Brewer, 2016). Teams that are selected to compete in a given bowl game split a purse, the size of which depends on the specific bowl (some bowls are more prestigious and have higher payouts for the two teams), and therefore securing an invitation to a bowl game is the main goal of any division I-A college football program. The decision makers of the bowl games are given the authority to select and invite bowl-eligible (a team that has six wins against its Division I-A opponents in that season) successful teams (as per the ratings and rankings) that will play in an exciting and competitive game, attract fans of both schools, and keep the remaining fans tuned in via a variety of media outlets for advertising.

In a recent data mining study, Delen, Cogdell, and Kasap (2012) used 8 years of bowl game data along with three popular data mining techniques (decision trees, neural networks, and support vector machines) to predict both the classification-type outcome of a game (win versus loss) as well as the regression-type outcome (projected point difference between the scores of the two opponents). What follows is a shorthand description of their study.

Methodology

In this research, Delen and his colleagues followed a popular data mining methodology called CRISP-DM (Cross-Industry Standard Process for Data Mining), which is a six-step process. This popular methodology, which is covered in detail in Chapter 4, provided them with a systematic and structured way to conduct the underlying data mining study and hence improved the likelihood of obtaining accurate and reliable results. To objectively assess the prediction power of the different model types, they used a cross-validation methodology, called *k*-fold cross-validation. Details on *k*-fold cross-validation can be found in Chapter 4. Figure 2.16 graphically illustrates the methodology employed by the researchers.

Data Acquisition and Data Preprocessing

The sample data for this study is collected from a variety of sports databases available on the Web, including jhowel.net, ESPN.com, Covers.com, ncaa.org, and rauzulusstreet.com. The data set included 244 bowl games, representing a complete set of eight seasons of college football bowl games played between 2002 and 2009. We also included an out-of-sample data set (2010–2011 bowl games) for

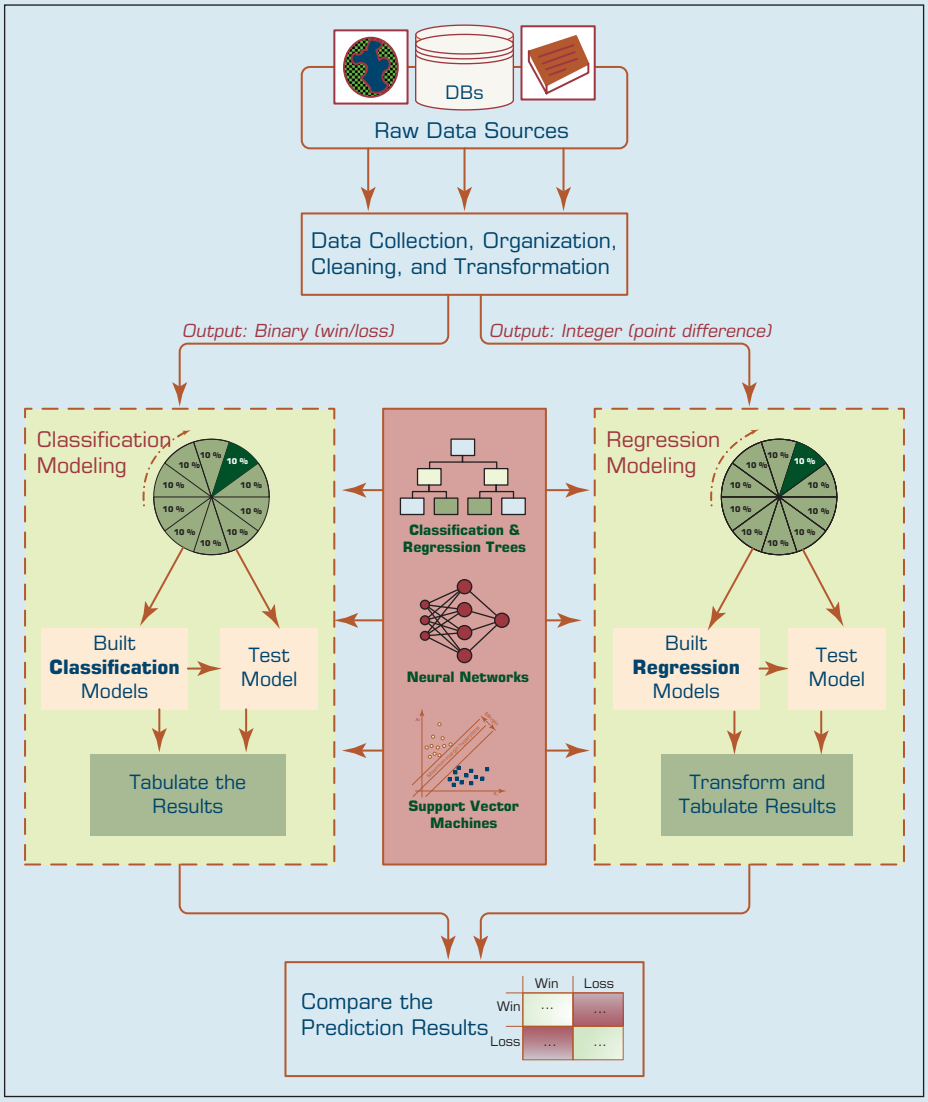


FIGURE 2.16 The Graphical Illustration of the Methodology Employed in the Study.

additional validation purposes. Exercising one of the popular data mining rules-of-thumb, they included as much relevant information into the model as possible. Therefore, after an in-depth variable identification and collection process, they ended up with a data set that included 36 variables, of which the first 6 were the identifying variables (i.e., name and the year of the bowl game, home and away team names and their athletic conferences—see variables 1–6 in Table 2.5), followed by 28 input variables (which included variables delineating a team's seasonal statistics on offense and defense, game outcomes, team composition characteristics, athletic conference characteristics, and how they fared against the odds—see variables 7–34 in Table 2.5), and finally the last two were the output variables (i.e., ScoreDiff—the score difference between the home team and the away team represented with an integer number, and WinLoss—whether the home team won or lost the bowl game represented with a nominal label).

In the formulation of the data set, each row (a.k.a. tuple, case, sample, example, etc.) represented a bowl game, and each column stood for a variable (i.e., identifier/input or output type). To represent the game-related comparative characteristics of the two opponent teams, in the input variables,

(*Continued*)

Application Case 2.4 (Continued)

TABLE 2.5 Description of the Variables Used in the Study

No	Cat	Variable Name	Description
1	ID	YEAR	Year of the bowl game
2	ID	BOWLGAME	Name of the bowl game
3	ID	HOMETEAM	Home team (as listed by the bowl organizers)
4	ID	AWAYTEAM	Away team (as listed by the bowl organizers)
5	ID	HOMECONFERENCE	Conference of the home team
6	ID	AWAYCONFERENCE	Conference of the away team
7	I1	DEFPTPGM	Defensive points per game
8	I1	DEFRYDPGM	Defensive rush yards per game
9	I1	DEFYDPGM	Defensive yards per game
10	I1	PPG	Average number of points a given team scored per game
11	I1	PYDPGM	Average total pass yards per game
12	I1	RYDPGM	Team's average total rush yards per game
13	I1	YRDPGM	Average total offensive yards per game
14	I2	HMWIN%	Home winning percentage
15	I2	LAST7	How many games the team won out of their last 7 games
16	I2	MARGOVIC	Average margin of victory
17	I2	NCTW	Nonconference team winning percentage
18	I2	PREVAPP	Did the team appeared in a bowl game previous year
19	I2	RDWIN%	Road winning percentage
20	I2	SEASTW	Winning percentage for the year
21	I2	TOP25	Winning percentage against AP top 25 teams for the year
22	I3	TSOS	Strength of schedule for the year
23	I3	FR%	Percentage of games played by freshmen class players for the year
24	I3	SO%	Percentage of games played by sophomore class players for the year
25	I3	JR%	Percentage of games played by junior class players for the year
26	I3	SR%	Percentage of games played by senior class players for the year
27	I4	SEASOvUn%	Percentage of times a team went over the O/U* in the current season
28	I4	ATSCOV%	Against the spread cover percentage of the team in previous bowl games
39	I4	UNDER%	Percentage of times a team went under in previous bowl games
30	I4	OVER%	Percentage of times a team went over in previous bowl games
31	I4	SEASATS%	Percentage of covering against the spread for the current season
32	I5	CONCH	Did the team win their respective conference championship game
33	I5	CONFSOS	Conference strength of schedule
34	I5	CONFWIN%	Conference winning percentage
35	O1	ScoreDiff°	Score difference (HomeTeamScore – AwayTeamScore)
36	O2	WinLoss°	Whether the home team wins or loses the game

** Over/Under—Whether or not a team will go over or under of the expected score difference.*

° Output variables—ScoreDiff for regression models and WinLoss for binary classification models.

I1: Offense/defense; I2: game outcome; I3: team configuration; I4: against the odds; I5: conference stats.

ID: Identifier variables; O1: output variable for regression models; O2: output variable for classification models.

we calculated and used the differences between the measures of the home and away teams. All these variable values are calculated from the home team's perspective. For instance, the variable PPG (average number of points a team scored per game) represents the difference between the home team's PPG and away team's PPG. The output variables represent whether the home team wins or loses the bowl game. That is, if the ScoreDiff variable takes a positive integer number, then the home team is expected to win the game by that margin, otherwise (if the ScoreDiff variable takes a negative integer number) then the home team is expected to lose the game by that margin. In the case of WinLoss, the value of the output variable is a binary label, "Win" or "Loss" indicating the outcome of the game for the home team.

Results and Evaluation

In this study, three popular prediction techniques are used to build models (and to compare them to each other): artificial neural networks, decision trees, and support vector machines. These prediction techniques are selected based on their capability of modeling both classification as well as regression-type prediction problems and their popularity in recently published data mining literature. More details about these popular data mining methods can be found in Chapter 4.

To compare predictive accuracy of all models to one another, the researchers used a stratified *k*-fold cross-validation methodology. In a stratified version of *k*-fold cross-validation, the folds are created in a way that they contain approximately the same proportion of predictor labels (i.e., classes) as the original data set. In this study, the value of *k* is set to 10 (i.e., the complete set of 244 samples are split into 10 subsets, each having about 25 samples), which is a common practice in predictive data mining applications. A graphical depiction of the 10-fold cross-validations was shown earlier in this chapter. To compare the prediction models that were developed using the aforementioned three data mining techniques, the researchers chose to use three common performance criteria: accuracy, sensitivity, and specificity. The simple formulas for these metrics were also explained earlier in this chapter.

The prediction results of the three modeling techniques are presented in Table 2.6 and Table 2.7. Table 2.6 presents the 10-fold cross-validation results of the classification methodology where the three data mining techniques are formulated to have a binary-nominal output variable (i.e., *WinLoss*). Table 2.7 presents the 10-fold cross-validation results of the regression-based classification methodology, where the three data mining techniques are formulated to have a numerical output variable (i.e., *ScoreDiff*). In the regression-based classification prediction, the numerical output of the models is converted to a classification type by labeling the positive *WinLoss* numbers with a "Win" and negative *WinLoss* numbers with a "Loss," and then tabulating them in the confusion matrixes. Using the confusion matrices, the overall prediction accuracy, sensitivity, and specificity of each model type are calculated and presented in these two tables. As the results indicate, the classification-type prediction methods performed better than regression-based classification-type prediction methodology. Among the three data mining

TABLE 2.6 Prediction Results for the Direct Classification Methodology

Prediction Method (Classification*)		Confusion Matrix		Accuracy** (in %)	Sensitivity (in %)	Specificity (in %)
		Win	Loss			
ANN (MLP)	Win	92	42	75.00	68.66	82.73
	Loss	19	91			
SVM (RBF)	Win	105	29	79.51	78.36	80.91
	Loss	21	89			
DT (C&RT)	Win	113	21	**86.48**	84.33	89.09
	Loss	12	98			

The output variable is a binary categorical variable (Win or Loss); differences were sig ($p < 0.01$).*

(Continued)

Application Case 2.4 (Continued)

TABLE 2.7 Prediction Results for the Regression-Based Classification Methodology

Prediction Method (Regression-Based*)		Confusion Matrix		Accuracy**	Sensitivity	Specificity
		Win	Loss			
ANN (MLP)	Win	94	40	72.54	70.15	75.45
	Loss	27	83			
SVM (RBF)	Win	100	34	74.59	74.63	74.55
	Loss	28	82			
DT (C&RT)	Win	106	28	77.87	76.36	79.10
	Loss	26	84			

The output variable is a numerical/integer variable (point-diff); differences were sig ($p < 0.01$).*

technologies, classification and regression trees produced better prediction accuracy in both prediction methodologies. Overall, classification and regression tree classification models produced a 10-fold cross-validation accuracy of 86.48%, followed by support vector machines (with a 10-fold cross-validation accuracy of 79.51%) and neural networks (with a 10-fold cross-validation accuracy of 75.00%). Using a t-test, researchers found that these accuracy values were significantly different at 0.05 alpha level, that is, the decision tree is a significantly better predictor of this domain than the neural network and support vector machine, and the support vector machine is a significantly better predictor than neural networks.

The results of the study showed that the classification-type models predict the game outcomes better than regression-based classification models. Even though these results are specific to the application domain and the data used in this study, and therefore should not be generalized beyond the scope of the study, they are exciting because decision trees are not only the best predictors but also the best in understanding and deployment, compared to the other two machine-learning techniques employed in this study. More details about this study can be found in Delen et al. (2012).

Questions for Discussion

1. What are the foreseeable challenges in predicting sporting event outcomes (e.g., college bowl games)?
2. How did the researchers formulate/design the prediction problem (i.e., what were the inputs and output, and what was the representation of a single sample—row of data)?
3. How successful were the prediction results? What else can they do to improve the accuracy?

Sources: Delen, D., Cogdell, D., & Kasap, N. (2012). A comparative analysis of data mining methods in predicting NCAA bowl outcomes. *International Journal of Forecasting, 28*, 543–552; Freeman, K. M., & Brewer, R. M. (2016). The politics of American college football. *Journal of Applied Business and Economics, 18*(2), 97–101.

Time Series Forecasting

Sometimes the variable that we are interested in (i.e., the response variable) may not have distinctly identifiable explanatory variables, or there may be too many of them in a highly complex relationship. In such cases, if the data is available in a desired format, a prediction model, the so-called time series, can be developed. A time series is a sequence of data points of the variable of interest, measured and represented at successive points in time spaced at uniform time intervals. Examples of time series include monthly rain volumes in a geographic area, the daily closing value of the stock market indexes, daily sales totals for

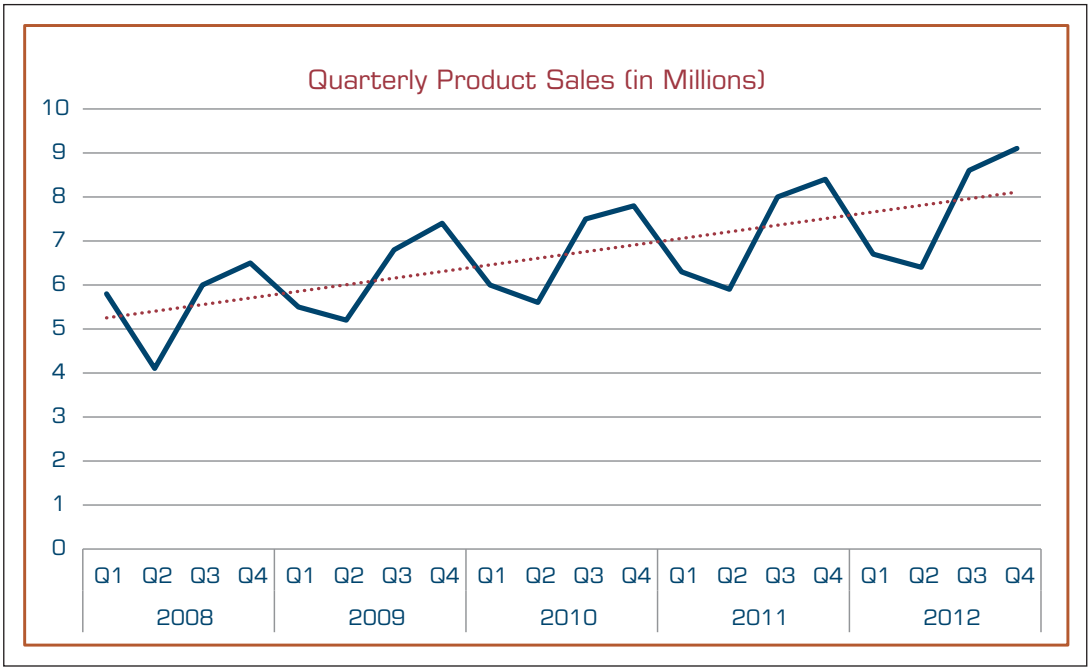


FIGURE 2.17 A Sample Time Series of Data on Quarterly Sales Volumes.

a grocery store. Often, time series are visualized using a line chart. Figure 2.17 shows an example time series of sales volumes for the years 2008 through 2012 in a quarterly basis.

Time series forecasting is the use of mathematical modeling to predict future values of the variable of interest based on previously observed values. The time series plots/charts look and feel very similar to simple linear regression in that as was the case in simple linear regression, in time series there are two variables: the response variable and the time variable presented in a scatter plot. Beyond this look similarity, there is hardly any other commonality between the two. Although regression analysis is often employed in testing theories to see if current values of one or more explanatory variables explain (and hence predict) the response variable, the time series models are focused on extrapolating on their time-varying behavior to estimate the future values.

Time series forecasting assumes all the explanatory variables are aggregated and consumed in the response variable's time-variant behavior. Therefore, capturing of the time-variant behavior is the way to predict the future values of the response variable. To do that the pattern is analyzed and decomposed into its main components: random variations, time trends, and seasonal cycles. The time series example shown in Figure 2.17 illustrates all these distinct patterns.

The techniques used to develop time series forecasts range from very simple (the naïve forecast that suggests today's forecast is the same as yesterday's actual) to very complex like ARIMA (a method that combines autoregressive and moving average patterns in data). Most popular techniques are perhaps the averaging methods that include simple average, moving average, weighted moving average, and exponential smoothing. Many of these techniques also have advanced versions where seasonality and trend can also be taken into account for better and more accurate forecasting. The accuracy of a method is usually assessed by computing its error (calculated deviation between actuals and forecasts for the past observations) via mean absolute error (MAE), mean squared error (MSE), or mean absolute percent error (MAPE). Even though they all use the same core error measure, these three assessment methods emphasize different aspects of the error, some panelizing larger errors more so than the others.

SECTION 2.6 REVIEW QUESTIONS

1. What is regression, and what statistical purpose does it serve?
2. What are the commonalities and differences between regression and correlation?
3. What is OLS? How does OLS determine the linear regression line?
4. List and describe the main steps to follow in developing a linear repression model.
5. What are the most commonly pronounced assumptions for linear regression?
6. What is logistics regression? How does it differ from linear regression?
7. What is time series? What are the main forecasting techniques for time series data?

2.7 Business Reporting

Decision makers are in need of information to make accurate and timely decisions. Information is essentially the contextualization of data. In addition to statistical means that were explained in the previous section, information (descriptive analytics) can also be obtained using online analytics processing [OLTP] systems (see the simple taxonomy of descriptive analytics in Figure 2.7). The information is usually provided to the decision makers in the form of a written report (digital or on paper), although it can also be provided orally. Simply put, a **report** is any communication artifact prepared with the specific intention of conveying information in a digestible form to whoever needs it, whenever and wherever they may need it. It is usually a document that contains information (usually driven from data) organized in a narrative, graphic, and/or tabular form, prepared periodically (recurring) or on an as-needed (ad hoc) basis, referring to specific time periods, events, occurrences, or subjects. Business reports can fulfill many different (but often related) functions. Here are a few of the most prevailing ones:

- To ensure that all departments are functioning properly
- To provide information
- To provide the results of an analysis
- To persuade others to act
- To create an organizational memory (as part of a knowledge management system)

Business reporting (also called OLAP or BI) is an essential part of the larger drive toward improved, evidence-based, optimal managerial decision making. The foundation of these **business reports** is various sources of data coming from both inside and outside the organization (online transaction processing [OLTP] systems). Creation of these reports involves ETL (extract, transform, and load) procedures in coordination with a data warehouse and then using one or more reporting tools (see Chapter 3 for a detailed description of these concepts).

Due to the rapid expansion of information technology coupled with the need for improved competitiveness in business, there has been an increase in the use of computing power to produce unified reports that join different views of the enterprise in one place. Usually, this reporting process involves querying structured data sources, most of which were created using different logical data models and data dictionaries, to produce a human-readable, easily digestible report. These types of business reports allow managers and coworkers to stay informed and involved, review options and alternatives, and make informed decisions. Figure 2.18 shows the continuous cycle of data acquisition → information generation → decision making → business process management. Perhaps the most critical task in this cyclical process is the reporting (i.e., information generation)—converting data from different sources into actionable information.

Key to any successful report are clarity, brevity, completeness, and correctness. The nature of the report and the level of importance of these success factors change significantly

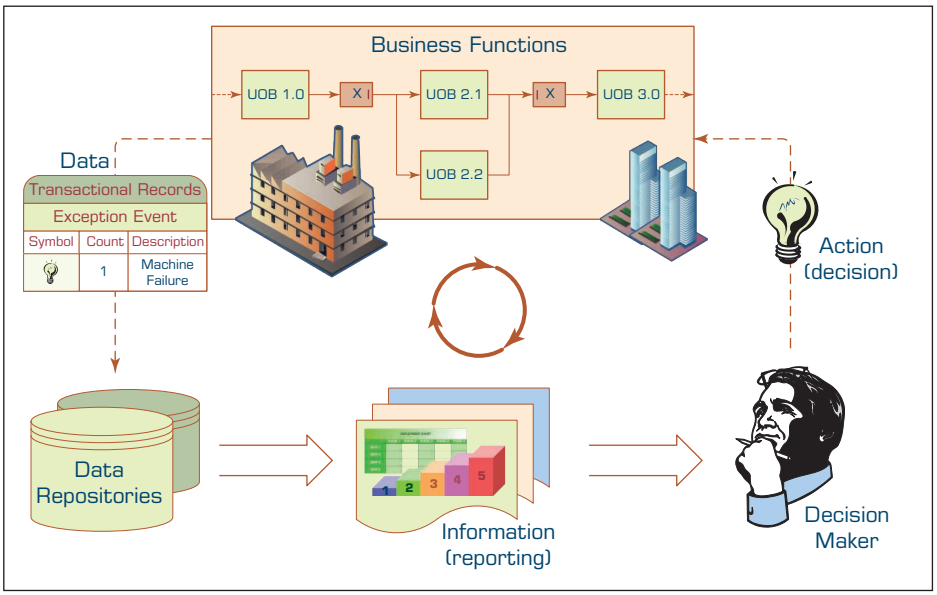


FIGURE 2.18 The Role of Information Reporting in Managerial Decision Making.

based on for whom the report is created. Most of the research in effective reporting is dedicated to internal reports that inform stakeholders and decision makers within the organization. There are also external reports between businesses and the government (e.g., for tax purposes or for regular filings to the Securities and Exchange Commission). Even though there are a wide variety of business reports, the ones that are often used for managerial purposes can be grouped into three major categories (Hill, 2016).

METRIC MANAGEMENT REPORTS In many organizations, business performance is managed through outcome-oriented metrics. For external groups, these are service-level agreements. For internal management, they are **key performance indicators** (**KPIs**). Typically, there are enterprise-wide agreed targets to be tracked against over a period of time. They may be used as part of other management strategies such as Six Sigma or Total Quality Management.

DASHBOARD-TYPE REPORTS A popular idea in business reporting in recent years has been to present a range of different performance indicators on one page, like a dashboard in a car. Typically, dashboard vendors would provide a set of predefined reports with static elements and fixed structure, but also allow for customization of the dashboard widgets, views, and set targets for various metrics. It's common to have color-coded traffic lights defined for performance (red, orange, green) to draw management's attention to particular areas. A more detailed description of dashboards can be found in later part of this chapter.

BALANCED SCORECARD–TYPE REPORTS This is a method developed by Kaplan and Norton that attempts to present an integrated view of success in an organization. In addition to financial performance, balanced scorecard–type reports also include customer, business process, and learning and growth perspectives. More details on balanced scorecards are provided later in this chapter.

Application Case 2.5 is an example to illustrate the power and the utility of automated report generation for a large (and, at a time of natural crisis, somewhat chaotic) organization like FEMA.

Application Case 2.5

Flood of Paper Ends at FEMA

Staff at the Federal Emergency Management Agency (FEMA), a U.S. federal agency that coordinates disaster response when the president declares a national disaster, always got two floods at once. First, water covered the land. Next, a flood of paper, required to administer the National Flood Insurance Program (NFIP) covered their desks—pallets and pallets of green-striped reports poured off a mainframe printer and into their offices. Individual reports were sometimes 18 inches thick, with a nugget of information about insurance claims, premiums, or payments buried in them somewhere.

Bill Barton and Mike Miles don't claim to be able to do anything about the weather, but the project manager and computer scientist, respectively, from Computer Sciences Corporation (CSC) have used WebFOCUS software from Information Builders to turn back the flood of paper generated by the NFIP. The program allows the government to work together with national insurance companies to collect flood insurance premiums and pay claims for flooding in communities that adopt flood control measures. As a result of CSC's work, FEMA staff no longer leaf through paper reports to find the data they need. Instead, they browse insurance data posted on NFIP's BureauNet intranet site, select just the information they want to see, and get an on-screen report or download the data as a spreadsheet. And that is only the start of the savings that WebFOCUS has provided. The number of times that NFIP staff asks CSC for special reports has dropped in half because NFIP staff can generate many of the special reports they need without calling on a programmer to develop them. Then there is the cost of creating BureauNet in the first place. Barton estimates that using conventional Web and database software to export data from FEMA's mainframe, store it in a new database, and link that to a Web server would have cost about 100 times as much—more than $500,000—and taken about 2 years to complete, compared with the few months Miles spent on the WebFOCUS solution.

When Tropical Storm Allison, a huge slug of sodden, swirling cloud, moved out of the Gulf of Mexico onto the Texas and Louisiana coastline in June 2001, it killed 34 people, most from drowning; damaged or destroyed 16,000 homes and businesses; and displaced more than 10,000 families. President George W. Bush declared 28 Texas counties disaster areas, and FEMA moved in to help. This was the first serious test for BureauNet, and it delivered. This first comprehensive use of BureauNet resulted in FEMA field staff readily accessing what they needed when they needed it and asking for many new types of reports. Fortunately, Miles and WebFOCUS were up to the task. In some cases, Barton says, "FEMA would ask for a new type of report one day, and Miles would have it on BureauNet the next day, thanks to the speed with which he could create new reports in WebFOCUS."

The sudden demand on the system had little impact on its performance, noted Barton. "It handled the demand just fine," he says. "We had no problems with it at all. And it made a huge difference to FEMA and the job they had to do. They had never had that level of access before, never had been able to just click on their desktop and generate such detailed and specific reports."

Questions for Discussion

1. What is FEMA, and what does it do?
2. What are the main challenges that FEMA faces?
3. How did FEMA improve its inefficient reporting practices?

Source: Information Builders success story. Useful information flows at disaster response agency. informationbuilders.com/applications/fema (accessed May 2016); and fema.gov.

SECTION 2.7 REVIEW QUESTIONS

1. What is a report? What are reports used for?
2. What is a business report? What are the main characteristics of a good business report?
3. Describe the cyclic process of management, and comment on the role of business reports.
4. List and describe the three major categories of business reports.
5. What are the main components of a business reporting system?

2.8 Data Visualization

Data visualization (or more appropriately, information visualization) has been defined as "the use of visual representations to explore, make sense of, and communicate data" (Few, 2007). Although the name that is commonly used is *data visualization*, usually what is meant by this is information visualization. Because information is the aggregation, summarization, and contextualization of data (raw facts), what is portrayed in visualizations is the information and not the data. However, because the two terms *data visualization* and *information visualization* are used interchangeably and synonymously, in this chapter we will follow suit.

Data visualization is closely related to the fields of information graphics, information visualization, scientific visualization, and statistical graphics. Until recently, the major forms of data visualization available in both BI applications have included charts and graphs, as well as the other types of visual elements used to create scorecards and dashboards.

To better understand the current and future trends in the field of data visualization, it helps to begin with some historical context.

A Brief History of Data Visualization

Despite the fact that predecessors to data visualization date back to the second century AD, most developments have occurred in the last two and a half centuries, predominantly during the last 30 years (Few, 2007). Although visualization has not been widely recognized as a discipline until fairly recently, today's most popular visual forms date back a few centuries. Geographical exploration, mathematics, and popularized history spurred the creation of early maps, graphs, and timelines as far back as the 1600s, but William Playfair is widely credited as the inventor of the modern chart, having created the first widely distributed line and bar charts in his *Commercial and Political Atlas of 1786* and what is generally considered to be the first time series portraying line charts in his *Statistical Breviary*, published in 1801 (see Figure 2.19).

Perhaps the most notable innovator of information graphics during this period was Charles Joseph Minard, who graphically portrayed the losses suffered by Napoleon's army in the Russian campaign of 1812 (see Figure 2.20). Beginning at the Polish–Russian border, the thick band shows the size of the army at each position. The path of Napoleon's retreat from Moscow in the bitterly cold winter is depicted by the dark lower band, which is tied to temperature and time scales. Popular visualization expert, author, and critic Edward Tufte says that this "may well be the best statistical graphic ever drawn." In this graphic Minard managed to simultaneously represent several data dimensions (the size of the army, direction of movement, geographic locations, outside temperature, etc.) in an artistic and informative manner. Many more excellent visualizations were created in the 1800s, and most of them are chronicled on Tufte's Web site (edwardtufte.com) and his visualization books.

The 1900s saw the rise of a more formal, empirical attitude toward visualization, which tended to focus on aspects such as color, value scales, and labeling. In the mid-1900s, cartographer and theorist Jacques Bertin published his *Semiologie Graphique*, which some say serves as the theoretical foundation of modern information visualization. Although most of his patterns are either outdated by more recent research or completely inapplicable to digital media, many are still very relevant.

In the 2000s, the Internet emerged as a new medium for visualization and brought with it a whole lot of new tricks and capabilities. Not only has the worldwide, digital distribution of both data and visualization made them more accessible to a broader audience (raising visual literacy along the way), but it has also spurred the design of new forms that incorporate interaction, animation, and graphics-rendering technology unique to screen

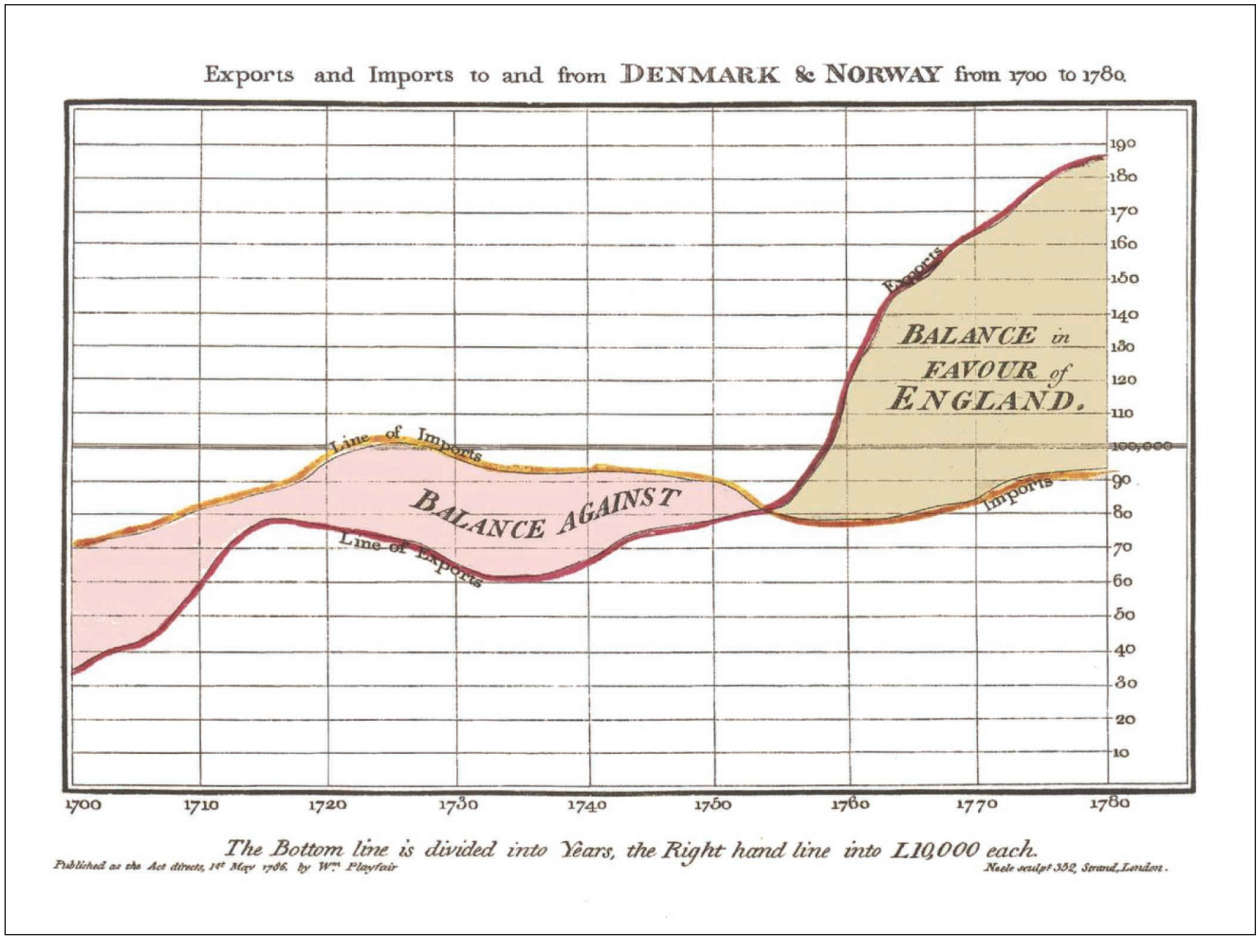


FIGURE 2.19 The First Time Series Line Chart Created by William Playfair in 1801.

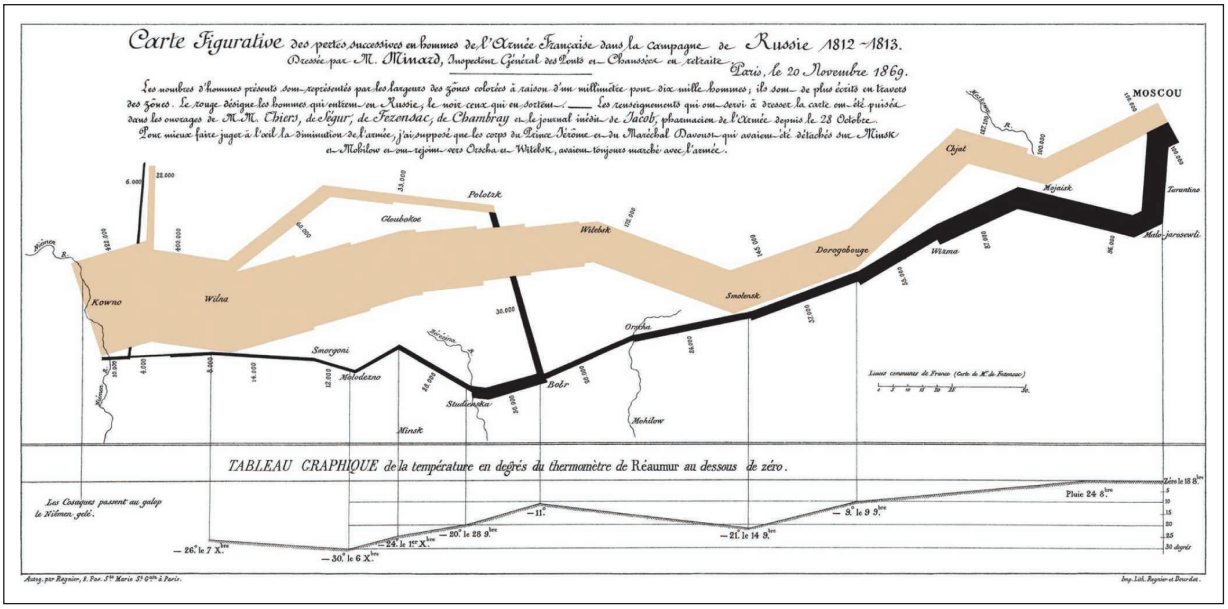


FIGURE 2.20 Decimation of Napoleon's Army during the 1812 Russian Campaign.

media, and real-time data feeds to create immersive environments for communicating and consuming data.

Companies and individuals are, seemingly all of a sudden, interested in data; that interest has in turn sparked a need for visual tools that help them understand it. Cheap hardware sensors and do-it-yourself frameworks for building your own system are driving down the costs of collecting and processing data. Countless other applications, software tools, and low-level code libraries are springing up to help people collect, organize, manipulate, visualize, and understand data from practically any source. The Internet has also served as a fantastic distribution channel for visualizations; a diverse community of designers, programmers, cartographers, tinkerers, and data wonks has assembled to disseminate all sorts of new ideas and tools for working with data in both visual and nonvisual forms.

Google Maps has also single-handedly democratized both the interface conventions (click to pan, double-click to zoom) and the technology (256-pixel square map tiles with predictable file names) for displaying interactive geography online, to the extent that most people just know what to do when they're presented with a map online. Flash has served well as a cross-browser platform on which to design and develop rich, beautiful Internet applications incorporating interactive data visualization and maps; now, new browser-native technologies such as canvas and SVG (sometimes collectively included under the umbrella of HTML5) are emerging to challenge Flash's supremacy and extend the reach of dynamic visualization interfaces to mobile devices.

The future of data/information visualization is very hard to predict. We can only extrapolate from what has already been invented: more three-dimensional visualization, more immersive experience with multidimensional data in a virtual reality environment, and holographic visualization of information. There is a pretty good chance that we will see something that we have never seen in the information visualization realm invented before the end of this decade. Application Case 2.6 shows how visual analytics/reporting tools like Tableau can help facilitate effective and efficient decision making through information/insight creation and sharing.

Application Case 2.6

Macfarlan Smith Improves Operational Performance Insight with Tableau Online

Background

Macfarlan Smith has earned its place in medical history. The company held a royal appointment to provide medicine to Her Majesty Queen Victoria and supplied groundbreaking obstetrician Sir James Simpson with chloroform for his experiments in pain relief during labor and delivery. Today, Macfarlan Smith is a subsidiary of the Fine Chemical and Catalysts division of Johnson Matthey plc. The pharmaceutical manufacturer is the world's leading manufacturer of opiate narcotics such as codeine and morphine.

Every day, Macfarlan Smith is making decisions based on its data. They collect and analyze manufacturing operational data, for example, to allow them to meet continuous improvement goals. Sales, marketing and finance rely on data to identify new pharmaceutical business opportunities, grow revenues and satisfy customer needs. Additionally, the company's manufacturing facility in Edinburgh needs to monitor, trend and report quality data to assure the identity, quality, and purity of its pharmaceutical ingredients for customers and regulatory authorities

(*Continued*)

Application Case 2.6 (Continued)

such as the U.S. Food and Drug Administration (FDA) and others as part of Good Manufacturing Practice (cGMP).

Challenges: Multiple Sources of Truth and Slow, Onerous Reporting Processes

The process of gathering that data, making decisions and reporting was not easy though. The data was scattered across the business: including in the company's bespoke enterprise resource planning (ERP) platform, inside legacy departmental databases such as SQL, Access databases, and standalone spreadsheets. When that data was needed for decision making, excessive time and resources were devoted to extracting the data, integrating it and presenting it in a spreadsheet or other presentation outlet.

Data quality was another concern. Because teams relied on their own individual sources of data, there were multiple versions of the truth and conflicts between the data. And it was sometimes hard to tell which version of the data was correct and which wasn't.

It didn't stop there. Even once the data had been gathered and presented, it was slow and difficult to make changes 'on the fly.' In fact, whenever a member of the Macfarlan Smith team wanted to perform trend or other analysis, the changes to the data needed to be approved. The end result being that the data was frequently out of date by the time it was used for decision making.

Liam Mills, Head of Continuous Improvement at Macfarlan Smith highlights a typical reporting scenario:

"One of our main reporting processes is the 'Corrective Action and Preventive Action', or CAPA, which is an analysis of Macfarlan Smith's manufacturing processes taken to eliminate causes of nonconformities or other undesirable situations. Hundreds of hours every month were devoted to pulling data together for CAPA—and it took days to produce each report. Trend analysis was tricky too, because the data was static. In other reporting scenarios, we often had to wait for spreadsheet pivot table analysis; which was then presented on a graph, printed out, and pinned to a wall for everyone to review."

Slow, labor-intensive reporting processes, different versions of the truth, and static data were all catalysts for change. "Many people were frustrated because they believed they didn't have a complete picture of the business," says Mills. "We were having more and more discussions about issues we faced—when we should have been talking about business intelligence reporting."

Solution: Interactive Data Visualizations

One of the Macfarlan Smith team had previous experience of using Tableau and recommended Mills explore the solution further. A free trial of Tableau Online quickly convinced Mills that the hosted interactive data visualization solution could conquer the data battles they were facing.

"I was won over almost immediately," he says. "The ease of use, the functionality and the breadth of data visualizations are all very impressive. And of course being a software-as-a-service (SaaS)-based solution, there's no technology infrastructure investment, we can be live almost immediately, and we have the flexibility to add users whenever we need."

One of the key questions that needed to be answered concerned the security of the online data. "Our parent company Johnson Matthey has a cloud-first strategy, but has to be certain that any hosted solution is completely secure. Tableau Online features like single sign-on and allowing only authorized users to interact with the data provide that watertight security and confidence."

The other security question that Macfarlan Smith and Johnson Matthey wanted answered was: Where is the data physically stored? Mills again: "We are satisfied Tableau Online meets our criteria for data security and privacy. The data and workbooks are all hosted in Tableau's new Dublin data center, so it never leaves Europe."

Following a six-week trial, the Tableau sales manager worked with Mills and his team to build a business case for Tableau Online. The management team approved it almost straight away and a pilot program involving 10 users began. The pilot involved a manufacturing quality improvement initiative: looking at deviations from the norm, such as when a heating device used in the opiate narcotics manufacturing process exceeds a temperature threshold. From this, a 'quality operations' dashboard was created to track and measure deviations

and put in place measures to improve operational quality and performance.

"That dashboard immediately signaled where deviations might be. We weren't ploughing through rows of data—we reached answers straight away," says Mills.

Throughout this initial trial and pilot, the team used Tableau training aids, such as the free training videos, product walkthroughs and live online training. They also participated in a two-day 'fundamentals training' event in London. According to Mills, "The training was expert, precise and pitched just at the right level. It demonstrated to everyone just how intuitive Tableau Online is. We can visualize 10 years' worth of data in just a few clicks." The company now has five Tableau Desktop users, and up to 200 Tableau Online licensed users.

Mills and his team particularly like the Tableau Union feature in Version 9.3, which allows them to piece together data that's been split into little files. "It's sometimes hard to bring together the data we use for analysis. The Union feature lets us work with data spread across multiple tabs or files, reducing the time we spend on prepping the data," he says.

Results: Cloud Analytics Transform Decision Making and Reporting

By standardizing on Tableau Online, Macfarlan Smith has transformed the speed and accuracy of its decision making and business reporting. This includes:

- New interactive dashboards can be produced within one hour. Previously, it used to take days to integrate and present data in a static spreadsheet.
- The CAPA manufacturing process report, which used to absorb hundreds of man-hours every month and days to produce, can now be produced in minutes—with insights shared in the cloud.
- Reports can be changed and interrogated 'on the fly' quickly and easily, without technical intervention. Macfarlan Smith has the flexibility to publish dashboards with Tableau Desktop and share them with colleagues, partners or customers.
- The company has one, single, trusted version of the truth.
- Macfarlan Smith is now having discussions about its data—not about the issues surrounding data integration and data quality.
- New users can be brought online almost instantly—and there's no technical infrastructure to manage.

Following this initial success, Macfarlan Smith is now extending Tableau Online out to financial reporting, supply chain analytics and sales forecasting. Mills concludes, "Our business strategy is now based on data-driven decisions, not opinions. The interactive visualizations enable us to spot trends instantly, identify process improvements and take business intelligence to the next level. I'll define my career by Tableau."

Questions for Discussion

1. What were the data and reporting related challenges Macfarlan Smith facing?
2. What was the solution and the obtained results/benefits?

Source: Tableau Customer Case Study, "Macfarlan Smith improves operational performance insight with Tableau Online," http://www.tableau.com/stories/customer/macfarlan-smith-improves-operational-performance-insight-tableau-online (accessed October 2016).

SECTION 2.8 REVIEW QUESTIONS

1. What is data visualization? Why is it needed?
2. What are the historical roots of data visualization?
3. Carefully analyze Charles Joseph Minard's graphical portrayal of Napoleon's march. Identify and comment on all the information dimensions captured in this ancient diagram.
4. Who is Edward Tufte? Why do you think we should know about his work?
5. What do you think is the "next big thing" in data visualization?

2.9 Different Types of Charts and Graphs

Often end users of business analytics systems are not sure what type of chart or graph to use for a specific purpose. Some charts or graphs are better at answering certain types of questions. Some look better than others. Some are simple; some are rather complex and crowded. What follows is a short description of the types of charts and/or graphs commonly found in most business analytics tools and what types of questions they are better at answering/analyzing. This material is compiled from several published articles and other literature (Abela, 2008; Hardin et al., 2012; SAS, 2014).

Basic Charts and Graphs

What follows are the basic charts and graphs that are commonly used for information visualization.

LINE CHART Line charts are perhaps the most frequently used graphical visuals for time series data. Line charts (or a line graphs) show the relationship between two variables; they are most often used to track changes or trends over time (having one of the variables set to time on the *x*-axis). Line charts sequentially connect individual data points to help infer changing trends over a period of time. Line charts are often used to show time-dependent changes in the values of some measure, such as changes on a specific stock price over a 5-year period or changes on the number of daily customer service calls over a month.

BAR CHART Bar charts are among the most basic visuals used for data representation. Bar charts are effective when you have nominal data or numerical data that splits nicely into different categories so you can quickly see comparative results and trends within your data. Bar charts are often used to compare data across multiple categories such as percent of advertising spending by departments or by product categories. Bar charts can be vertically or horizontally oriented. They can also be stacked on top of each other to show multiple dimensions in a single chart.

PIE CHART **Pie charts** are visually appealing, as the name implies, pie-looking charts. Because they are so visually attractive, they are often incorrectly used. Pie charts should only be used to illustrate relative proportions of a specific measure. For instance, they can be used to show the relative percentage of an advertising budget spent on different product lines, or they can show relative proportions of majors declared by college students in their sophomore year. If the number of categories to show is more than just a few (say more than four), one should seriously consider using a bar chart instead of a pie chart.

SCATTER PLOT **Scatter plots** are often used to explore the relationship between two or three variables (in 2-D or 2-D visuals). Because they are visual exploration tools, having more than three variables, translating them into more than three dimensions is not easily achievable. Scatter plots are an effective way to explore the existence of trends, concentrations, and outliers. For instance, in a two-variable (two-axis) graph, a scatter plot can be used to illustrate the corelationship between age and weight of heart disease patients or it can illustrate the relationship between the number of customer care representatives and the number of open customer service claims. Often, a trend line is superimposed on a two-dimensional scatter plot to illustrate the nature of the relationship.

BUBBLE CHART **Bubble charts** are often enhanced versions of scatter plots. Bubble charts, though, are not a new visualization type; instead, they should be viewed as a technique to enrich data illustrated in scatter plots (or even geographic maps). By varying the size and/or color of the circles, one can add additional data dimensions, offering more enriched meaning about the data. For instance, a bubble chart can be used to show a competitive view of college-level class attendance by major and by time of the day, or it can be used to show profit margin by product type and by geographic region.

Specialized Charts and Graphs

The graphs and charts that we review in this section are either derived from the basic charts as special cases or they are relatively new and are specific to a problem type and/or an application area.

HISTOGRAM Graphically speaking, a **histogram** looks just like a bar chart. The difference between histograms and generic bar charts is the information that is portrayed. Histograms are used to show the frequency distribution of a variable or several variables. In a histogram, the *x*-axis is often used to show the categories or ranges, and the *y*-axis is used to show the measures/values/frequencies. Histograms show the distributional shape of the data. That way, one can visually examine if the data is normally or exponentially distributed. For instance, one can use a histogram to illustrate the exam performance of a class, where distribution of the grades as well as comparative analysis of individual results can be shown, or one can use a histogram to show age distribution of the customer base.

GANTT CHART Gantt charts are a special case of horizontal bar charts that are used to portray project timelines, project tasks/activity durations, and overlap among the tasks/activities. By showing start and end dates/times of tasks/activities and the overlapping relationships, Gantt charts provide an invaluable aid for management and control of projects. For instance, Gantt charts are often used to show project timelines, task overlaps, relative task completions (a partial bar illustrating the completion percentage inside a bar that shows the actual task duration), resources assigned to each task, milestones, and deliverables.

PERT CHART PERT charts (also called network diagrams) are developed primarily to simplify the planning and scheduling of large and complex projects. They show precedence relationships among the project activities/tasks. A PERT chart is composed of nodes (represented as circles or rectangles) and edges (represented with directed arrows). Based on the selected PERT chart convention, either nodes or the edges may be used to represent the project activities/tasks (activity-on-node versus activity-on-arrow representation schema).

GEOGRAPHIC MAP When the data set includes any kind of location data (e.g., physical addresses, postal codes, state names or abbreviations, country names, latitude/longitude, or some type of custom geographic encoding), it is better and more informative to see the data on a map. Maps usually are used in conjunction with other charts and graphs, as opposed to by themselves. For instance, one can use maps to show distribution of customer service requests by product type (depicted in pie charts) by geographic locations. Often a large variety of information (e.g., age distribution, income distribution, education, economic growth, or population changes) can be portrayed in a geographic map to help decide where to open a new restaurant or a new service station. These types of systems are often called geographic information systems (GIS).

BULLET Bullet graphs are often used to show progress toward a goal. A bullet graph is essentially a variation of a bar chart. Often they are used in place of gauges, meters, and thermometers in a dashboard to more intuitively convey the meaning within a much smaller space. Bullet graphs compare a primary measure (e.g., year-to-date revenue) to one or more other measures (e.g., annual revenue target) and present this in the context of defined performance metrics (e.g., sales quotas). A bullet graph can intuitively illustrate how the primary measure is performing against overall goals (e.g., how close a sales representative is to achieving his/her annual quota).

HEAT MAP Heat maps are great visuals to illustrate the comparison of continuous values across two categories using color. The goal is to help the user quickly see where the intersection of the categories is strongest and weakest in terms of numerical values of the measure being analyzed. For instance, one can use heat maps to show segmentation analysis of target markets where the measure (color gradient would be the purchase amount) and the dimensions would be age and income distribution.

HIGHLIGHT TABLE Highlight tables are intended to take heat maps one step further. In addition to showing how data intersects by using color, highlight tables add a number on top to provide additional detail. That is, they are two-dimensional tables with cells populated with numerical values and gradients of colors. For instance, one can show sales representatives' performance by product type and by sales volume.

TREE MAP Tree maps display hierarchical (tree-structured) data as a set of nested rectangles. Each branch of the tree is given a rectangle, which is then tiled with smaller rectangles representing subbranches. A leaf node's rectangle has an area proportional to a specified dimension on the data. Often the leaf nodes are colored to show a separate dimension of the data. When the color and size dimensions are correlated in some way with the tree structure, one can often easily see patterns that would be difficult to spot in other ways, such as if a certain color is particularly relevant. A second advantage of tree maps is that, by construction, they make efficient use of space. As a result, they can legibly display thousands of items on the screen simultaneously.

Which Chart or Graph Should You Use?

Which chart or graph that we explained in the previous section is the best? The answer is rather easy: there is not one best chart or graph, because if there was we would not have these many chart and graph types. They all have somewhat different data representation "skills." Therefore, the right question should be, "Which chart or graph is the best for a given task?" The capabilities of the charts given in the previous section can help in selecting and using the right chart/graph for a specific task, but it still is not easy to sort out. Several different chart/graph types can be used for the same visualization task. One rule of thumb is to select and use the simplest one from the alternatives to make it easy for the intended audience to understand and digest.

Although there is not a widely accepted, all-encompassing chart selection algorithm or chart/graph taxonomy, Figure 2.21 presents a rather comprehensive and highly logical organization of chart/graph types in a taxonomy-like structure (the original version was published in Abela 2008). The taxonomic structure is organized around the questions of "What would you like to show in your chart or graph?" That is, what the purpose of the chart or graph will be. At that level, the taxonomy divides the purpose into four different types—relationship, comparison, distribution, and composition—and further divides the branches into subcategories based on the number of variables involved and time dependency of the visualization.

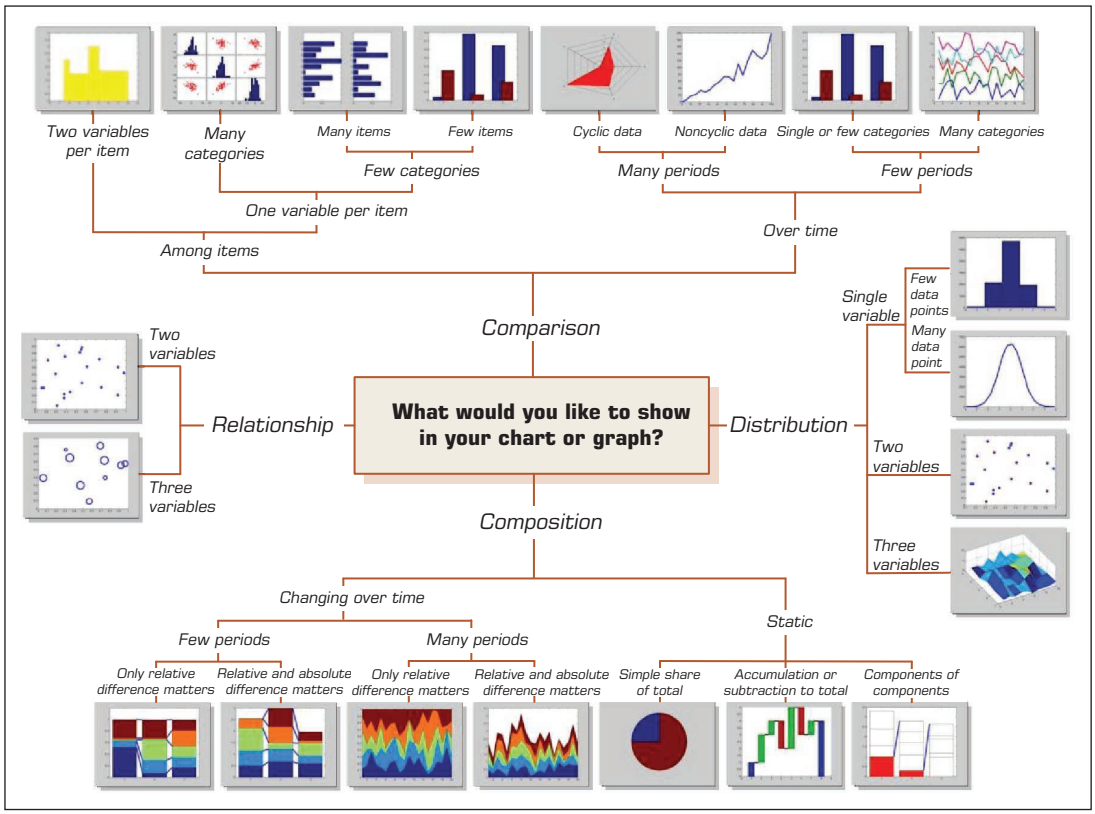


FIGURE 2.21 A Taxonomy of Charts and Graphs. *Source:* Adapted from Abela, A. (2008). Advanced presentations by design: Creating communication that drives action. New York: Wiley.

Even though these charts and graphs cover a major part of what is commonly used in information visualization, they by no means cover it all. Nowadays, one can find many other specialized graphs and charts that serve a specific purpose. Furthermore, the current trend is to combine/hybridize and animate these charts for better-looking and more intuitive visualization of today's complex and volatile data sources. For instance, the interactive, animated, bubble charts available at the Gapminder Web site (gapminder.org) provide an intriguing way of exploring world health, wealth, and population data from a multidimensional perspective. Figure 2.22 depicts the sorts of displays available at the site. In this graph, population size, life expectancy, and per capita income at the continent level are shown; also given is a time-varying animation that shows how these variables change over time.

SECTION 2.9 REVIEW QUESTIONS

1. Why do you think there are many different types of charts and graphs?
2. What are the main differences among line, bar, and pie charts? When should you use one over the others?
3. Why would you use a geographic map? What other types of charts can be combined with a geographic map?
4. Find and explain the role of two types of charts that are not covered in this section.

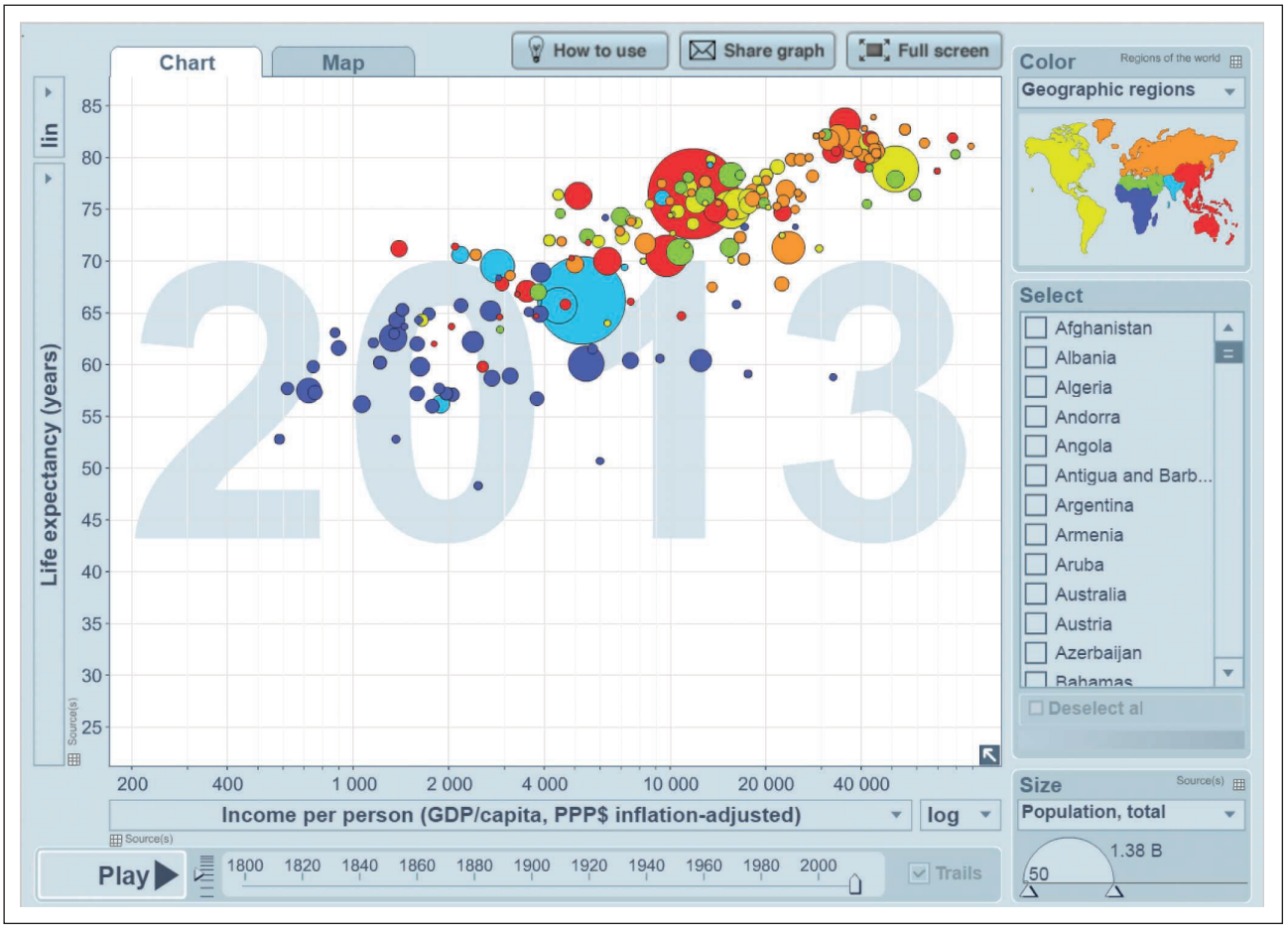


FIGURE 2.22 A Gapminder Chart That Shows the Wealth and Health of Nations. *Source:* gapminder.org.

2.10 The Emergence of Visual Analytics

As Seth Grimes (2009a,b) has noted, there is a "growing palate" of data visualization techniques and tools that enable the users of business analytics and BI systems to better "communicate relationships, add historical context, uncover hidden correlations, and tell persuasive stories that clarify and call to action." The latest Magic Quadrant on Business Intelligence and Analytics Platforms released by Gartner in February 2016 further emphasizes the importance of data visualization in BI and analytics. As the chart shows, all the solution providers in the *Leaders* and *Visionary* quadrants are either relatively recently founded information visualization companies (e.g., Tableau Software, QlikTech) or well-established large analytics companies (e.g., Microsoft, SAS, IBM, SAP, MicroStrategy, Alteryx) that are increasingly focusing their efforts on information visualization and visual analytics. More details on Gartner's latest Magic Quadrant are given in Technology Insights 2.2.

In BI and analytics, the key challenges for visualization have revolved around the intuitive representation of large, complex data sets with multiple dimensions and measures. For the most part, the typical charts, graphs, and other visual elements used in these applications usually involve two dimensions, sometimes three, and fairly small subsets of data sets. In contrast, the data in these systems reside in a data warehouse. At a minimum,

TECHNOLOGY INSIGHTS 2.2
Gartner Magic Quadrant for Business Intelligence and Analytics Platforms

Gartner, Inc., the creator of Magic Quadrants, is the leading information technology research and advisory company publically traded in the United States with over $2 billion annual revenues in 2015. Founded in 1979, Gartner has 7,600 associates, including 1,600 research analysts and consultants, and numerous clients in 90 countries.

Magic Quadrant is a research method designed and implemented by Gartner to monitor and evaluate the progress and positions of companies in a specific, technology-based market. By applying a graphical treatment and a uniform set of evaluation criteria, Magic Quadrant helps users to understand how technology providers are positioned within a market.

Gartner changed the name of this Magic Quadrant from "Business Intelligence Platforms" to "Business Intelligence and Analytics Platforms" to emphasize the growing importance of analytics capabilities to the information systems that organizations are now building. Gartner defines the BI and analytics platform market as a software platform that delivers 15 capabilities across three categories: integration, information delivery, and analysis. These capabilities enable organizations to build precise systems of classification and measurement to support decision making and improve performance.

Figure 2.23 illustrates the latest Magic Quadrant for Business Intelligence and Analytics Platforms. Magic Quadrant places providers in four groups (niche players, challengers, visionaries, and leaders) along two dimensions: completeness of vision (*x*-axis) and ability to execute (*y*-axis). As the quadrant clearly shows, most of the well-known BI/BA providers are positioned in the "leaders" category while many of the lesser known, relatively new, emerging providers are positioned in the "niche players" category.

FIGURE 2.23 Magic Quadrant for Business Intelligence and Analytics Platforms. *Source:* gartner.com.

(Continued)

The BI and analytics platform market's multiyear shift from IT-led enterprise reporting to business-led self-service analytics seem to have passed the tipping point. Most new buying is of modern, business-user-centric visual analytics platforms forcing a new market perspective, significantly reordering the vendor landscape. Most of the activity in the BI and analytics platform market is from organizations that are trying to mature their visualization capabilities and to move from descriptive to predictive and prescriptive analytics echelons. The vendors in the market have overwhelmingly concentrated on meeting this user demand. If there were a single market theme in 2015, it would be that data discovery/visualization became a mainstream architecture. While data discovery/visualization vendors such as Tableau, Qlik, and Microsoft are solidifying their position in the *Leaders* quadrant, others (both emerging and large, well-established tool/solution providers) are trying to move out of *Visionaries* into the *Leaders* quadrant.

This emphasis on data discovery/visualization from most of the leaders and visionaries in the market—which are now promoting tools with business-user-friendly data integration, coupled with embedded storage and computing layers and unfettered drilling—continue to accelerate the trend toward decentralization and user empowerment of BI and analytics and greatly enables organizations' ability to perform diagnostic analytics.

Source: Gartner Magic Quadrant, released on February 4, 2016, gartner.com (accessed August 2016).

these warehouses involve a range of dimensions (e.g., product, location, organizational structure, time), a range of measures, and millions of cells of data. In an effort to address these challenges, a number of researchers have developed a variety of new visualization techniques.

Visual Analytics

Visual analytics is a recently coined term that is often used loosely to mean nothing more than information visualization. What is meant by **visual analytics** is the combination of visualization and predictive analytics. Whereas information visualization is aimed at answering, "What happened?" and "What is happening?" and is closely associated with BI (routine reports, scorecards, and dashboards), visual analytics is aimed at answering, "Why is it happening?" "What is more likely to happen?" and is usually associated with business analytics (forecasting, segmentation, correlation analysis). Many of the information visualization vendors are adding the capabilities to call themselves visual analytics solution providers. One of the top, long-time analytics solution providers, SAS Institute, is approaching it from another direction. They are embedding their analytics capabilities into a high-performance data visualization environment that they call visual analytics.

Visual or not visual, automated or manual, online or paper based, business reporting is not much different than telling a story. Technology Insights 2.3 provides a different, unorthodox viewpoint to better business reporting.

High-Powered Visual Analytics Environments

Due to the increasing demand for visual analytics coupled with fast-growing data volumes, there is an exponential movement toward investing in highly efficient visualization systems. With their latest move into visual analytics, the statistical software giant SAS Institute is now among those who are leading this wave. Their new product, SAS Visual Analytics, is a very **high-performance computing**, in-memory solution for exploring massive amounts of data in a very short time (almost instantaneously). It empowers users to spot patterns, identify opportunities for further analysis, and convey visual results via Web reports or a mobile platform such as tablets and smartphones. Figure 2.25 shows the high-level architecture of the SAS Visual Analytics platform. On one end of the architecture, there is a universal data builder and administrator capabilities, leading into explorer, report designer, and mobile BI modules, collectively providing an end-to-end visual analytics solution.

TECHNOLOGY INSIGHTS 2.3
Telling Great Stories with Data and Visualization

Everyone who has data to analyze has stories to tell, whether it's diagnosing the reasons for manufacturing defects, selling a new idea in a way that captures the imagination of your target audience, or informing colleagues about a particular customer service improvement program. And when it's telling the story behind a big strategic choice so that you and your senior management team can make a solid decision, providing a fact-based story can be especially challenging. In all cases, it's a big job. You want to be interesting and memorable; you know you need to keep it simple for your busy executives and colleagues. Yet you also know you have to be factual, detail oriented, and data driven, especially in today's metric-centric world.

It's tempting to present just the data and facts, but when colleagues and senior management are overwhelmed by data and facts without context, you lose. We have all experienced presentations with large slide decks, only to find that the audience is so overwhelmed with data that they don't know what to think, or they are so completely tuned out that they take away only a fraction of the key points.

Start engaging your executive team and explaining your strategies and results more powerfully by approaching your assignment as a story. You will need the "what" of your story (the facts and data) but you also need the "Who?" "How?" "Why?" and the often-missed "So what?" It's these story elements that will make your data relevant and tangible for your audience. Creating a good story can aid you and senior management in focusing on what is important.

Why Story?

Stories bring life to data and facts. They can help you make sense and order out of a disparate collection of facts. They make it easier to remember key points and can paint a vivid picture of what the future can look like. Stories also create interactivity—people put themselves into stories and can relate to the situation.

Cultures have long used **storytelling** to pass on knowledge and content. In some cultures, storytelling is critical to their identity. For example, in New Zealand, some of the Maori people tattoo their faces with *mokus*. A *moku* is a facial tattoo containing a story about ancestors—the family tribe. A man may have a tattoo design on his face that shows features of a hammerhead to highlight unique qualities about his lineage. The design he chooses signifies what is part of his "true self" and his ancestral home.

Likewise, when we are trying to understand a story, the storyteller navigates to finding the "true north." If senior management is looking to discuss how they will respond to a competitive change, a good story can make sense and order out of a lot of noise. For example, you may have facts and data from two studies, one including results from an advertising study and one from a product satisfaction study. Developing a story for what you measured across both studies can help people see the whole where there were disparate parts. For rallying your distributors around a new product, you can employ a story to give vision to what the future can look like. Most important, storytelling is interactive—typically the presenter uses words and pictures that audience members can put themselves into. As a result, they become more engaged and better understand the information.

So What Is a Good Story?

Most people can easily rattle off their favorite film or book. Or they remember a funny story that a colleague recently shared. Why do people remember these stories? Because they contain certain characteristics. First, a good story has great characters. In some cases, the reader or viewer has a vicarious experience where they become involved with the character. The character then has to be faced with a challenge that is difficult but believable. There must be hurdles that the character overcomes. And finally, the outcome or prognosis is clear by the end of the story. The situation may not be resolved—but the story has a clear endpoint.

Think of Your Analysis as a Story—Use a Story Structure

When crafting a data-rich story, the first objective is to find the story. Who are the characters? What is the drama or challenge? What hurdles have to be overcome? And at the end of your story, what do you want your audience to do as a result?

Once you know the core story, craft your other story elements: define your characters, understand the challenge, identify the hurdles, and crystallize the outcome or decision question. Make sure you are clear

(Continued)

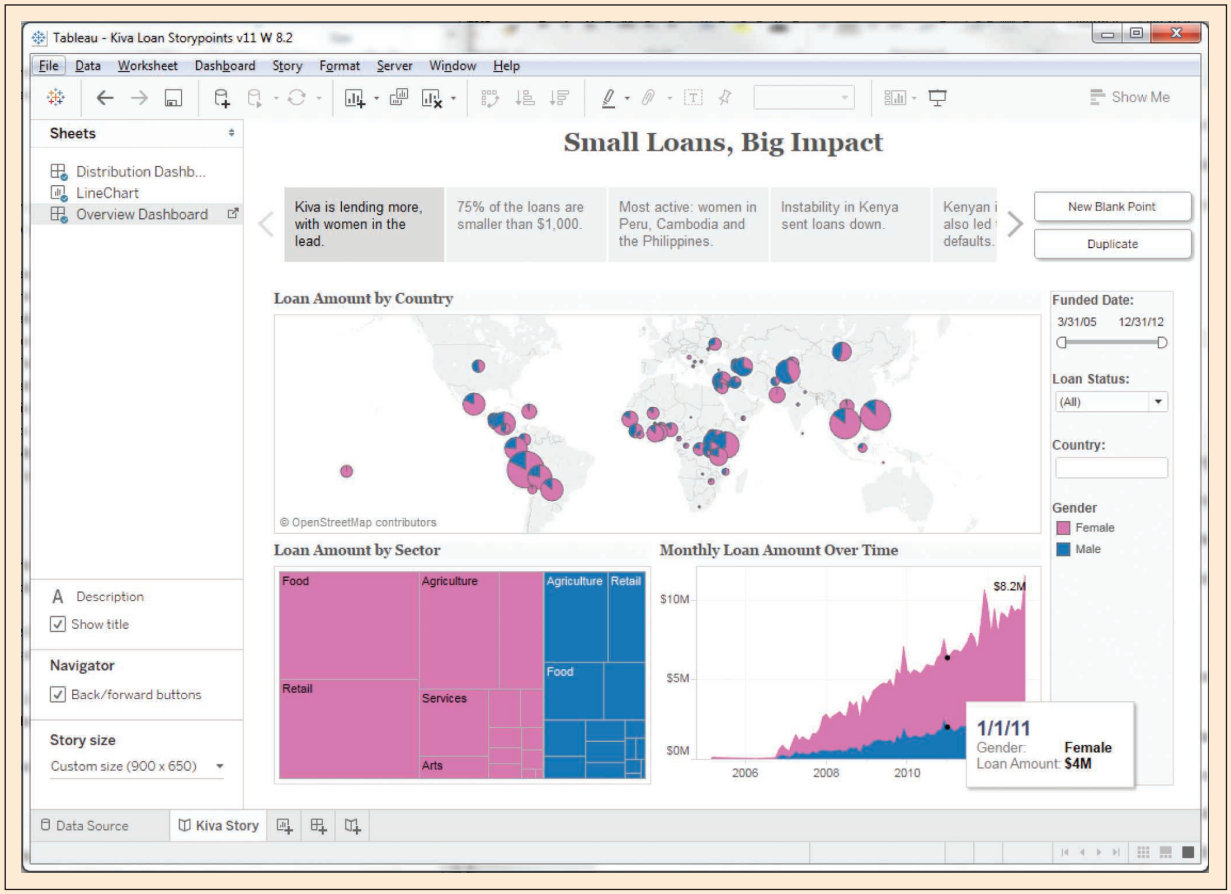


FIGURE 2.24 A Storyline Visualization in Tableau Software.

with what you want people to do as a result. This will shape how your audience will recall your story. With the story elements in place, write out the storyboard, which represents the structure and form of your story. Although it's tempting to skip this step, it is better first to understand the story you are telling and then to focus on the presentation structure and form. Once the storyboard is in place, the other elements will fall into place. The storyboard will help you to think about the best analogies or metaphors, to clearly set up challenge or opportunity, and to finally see the flow and transitions needed. The storyboard also helps you focus on key visuals (graphs, charts, and graphics) that you need your executives to recall. Figure 2.24 shows a storyline for the impact of small loans in a worldwide view within the Tableau visual analytics environment.

In summary, don't be afraid to use data to tell great stories. Being factual, detail oriented, and data driven is critical in today's metric-centric world, but it does not have to mean being boring and lengthy. In fact, by finding the real stories in your data and following the best practices, you can get people to focus on your message—and thus on what's important. Here are those best practices:

1. Think of your analysis as a story—use a story structure.
2. Be authentic—your story will flow.
3. Be visual—think of yourself as a film editor.
4. Make it easy for your audience and you.
5. Invite and direct discussion.

Source: Fink, E., & Moore, S. J. (2012). Five best practices for telling great stories with data. White paper by Tableau Software, Inc., www.tableau.com/whitepapers/telling-data-stories (accessed May 2016).

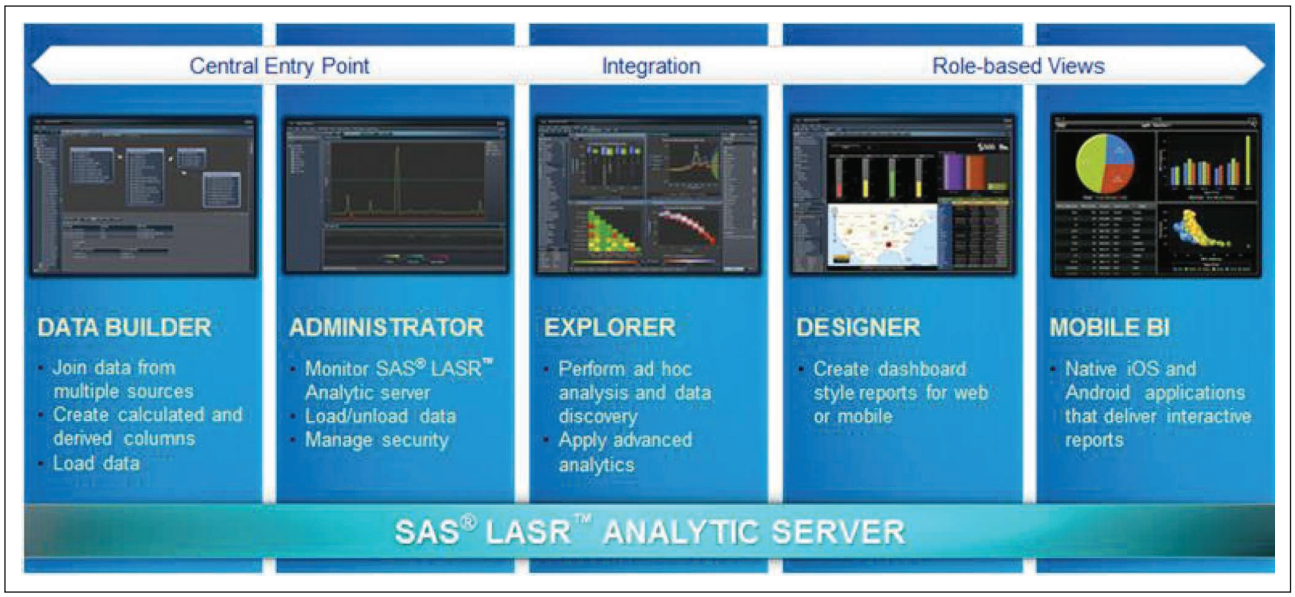


FIGURE 2.25 An Overview of SAS Visual Analytics Architecture. *Source:* SAS.com.

Some of the key benefits proposed by SAS analytics are the following:

- Empowers all users with data exploration techniques and approachable analytics to drive improved decision making. SAS Visual Analytics enables different types of users to conduct fast, thorough explorations on all available data. Sampling to reduce the data is not required and not preferred.
- Easy-to-use, interactive Web interfaces broaden the audience for analytics, enabling everyone to glean new insights. Users can look at more options, make more precise decisions, and drive success even faster than before.
- Answer complex questions faster, enhancing the contributions from your analytic talent. SAS Visual Analytics augments the data discovery and exploration process by providing extremely fast results to enable better, more focused analysis. Analytically savvy users can identify areas of opportunity or concern from vast amounts of data so further investigation can take place quickly.
- Improves information sharing and collaboration. Large numbers of users, including those with limited analytical skills, can quickly view and interact with reports and charts via the Web, Adobe PDF files, and iPad mobile devices, while IT maintains control of the underlying data and security. SAS Visual Analytics provides the right information to the right person at the right time to improve productivity and organizational knowledge.
- Liberates IT by giving users a new way to access the information they need. Frees IT from the constant barrage of demands from users who need access to different amounts of data, different data views, ad hoc reports, and one-off requests for information. SAS Visual Analytics enables IT to easily load and prepare data for multiple users. Once data is loaded and available, users can dynamically explore data, create reports, and share information on their own.
- Provides room to grow at a self-determined pace. SAS Visual Analytics provides the option of using commodity hardware or database appliances from EMC Greenplum and Teradata. It is designed from the ground up for performance optimization and scalability to meet the needs of any size organization.

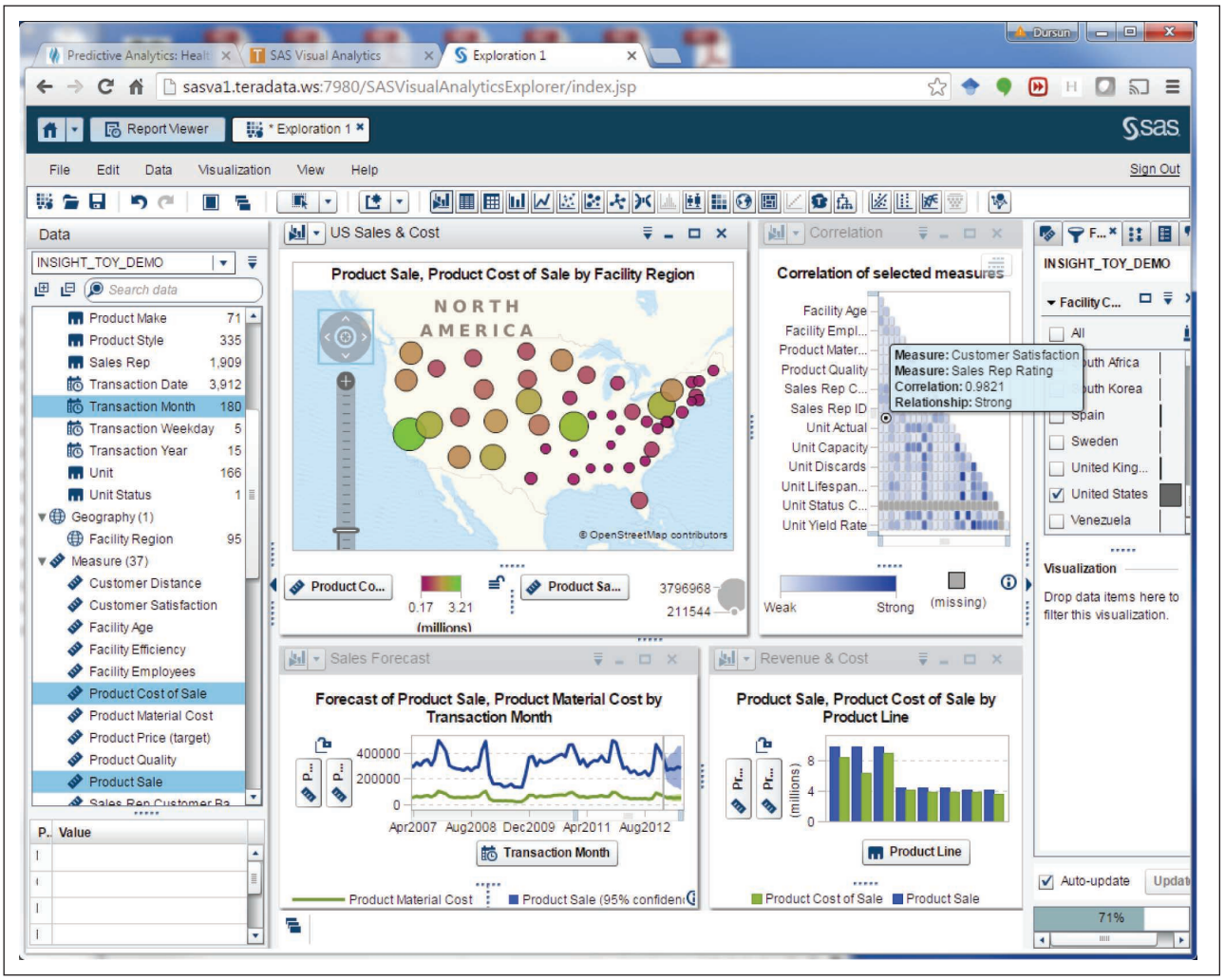


FIGURE 2.26 A Screenshot from SAS Visual Analytics. *Source:* SAS.com.

Figure 2.26 shows a screenshot of an SAS Analytics platform where time series forecasting and confidence interval around the forecast are depicted.

SECTION 2.10 REVIEW QUESTIONS

1. What are the main reasons for the recent emergence of visual analytics?
2. Look at Gartner's Magic Quadrant for Business Intelligence and Analytics Platforms. What do you see? Discuss and justify your observations.
3. What is the difference between information visualization and visual analytics?
4. Why should storytelling be a part of your reporting and data visualization?
5. What is a high-powered visual analytics environment? Why do we need it?

2.11 Information Dashboards

Information dashboards are common components of most, if not all, BI or business analytics platforms, business performance management systems, and performance measurement software suites. **Dashboards** provide visual displays of important information that is consolidated and arranged on a single screen so that information can be digested at a single glance and easily drilled in and further explored. A typical dashboard is shown in Figure 2.27. This particular executive dashboard displays a variety of KPIs for a hypothetical software company called Sonatica (selling audio tools). This executive dashboard shows a high-level view of the different functional groups surrounding the products, starting from a general overview to the marketing efforts, sales, finance, and support departments. All of this is intended to give executive decision makers a quick and accurate idea of what is going on within the organization. On the left side of the dashboard, we can see (in a time series fashion) the quarterly changes in revenues, expenses, and margins, as well as the comparison of those figures to previous years' monthly numbers. On the upper-right side we see two dials with color-coded regions showing the amount of

FIGURE 2.27 A Sample Executive Dashboard. *Source:* dundas.com.

monthly expenses for support services (dial on the left) and the amount of other expenses (dial on the right). As the color coding indicates, although the monthly support expenses are well within the normal ranges, the other expenses are in the red region, indicating excessive values. The geographic map on the bottom right shows the distribution of sales at the country level throughout the world. Behind these graphical icons there are variety of mathematical functions aggregating numerous data points to their highest level of meaningful figures. By clicking on these graphical icons, the consumer of this information can drill down to more granular levels of information and data.

Dashboards are used in a wide variety of businesses for a wide variety of reasons. For instance, in Application Case 2.7, you will find the summary of a successful implementation of information dashboards by the Dallas Cowboys football team.

Application Case 2.7

Dallas Cowboys Score Big with Tableau and Teknion

Founded in 1960, the Dallas Cowboys are a professional American football team headquartered in Irving, Texas. The team has a large national following, which is perhaps best represented by their NFL record for number of consecutive games at sold-out stadiums.

Challenge

Bill Priakos, Chief Operating Officer (COO) of the Dallas Cowboys Merchandising Division, and his team needed more visibility into their data so they could run it more profitably. Microsoft was selected as the baseline platform for this upgrade as well as a number of other sales, logistics, and e-commerce (per MW) applications. The Cowboys expected that this new information architecture would provide the needed analytics and reporting. Unfortunately, this was not the case, and the search began for a robust dashboarding, analytics, and reporting tool to fill this gap.

Solution and Results

Tableau and Teknion together provided real-time reporting and dashboard capabilities that exceeded the Cowboys' requirements. Systematically and methodically the Teknion team worked side by side with data owners and data users within the Dallas Cowboys to deliver all required functionality, on time and under budget. "Early in the process, we were able to get a clear understanding of what it would take to run a more profitable operation for the Cowboys," said Teknion Vice President Bill Luisi. "This process step is a key step in Teknion's approach with any client, and it always pays huge dividends as the implementation plan progresses." Added Luisi, "Of course, Tableau worked very closely with us and the Cowboys during the entire project. Together, we made sure that the Cowboys could achieve their reporting and analytical goals in record time."

Now, for the first time, the Dallas Cowboys are able to monitor their complete merchandising activities from manufacture to end customer and not only see what is happening across the life cycle, but also drill down even further into why it is happening.

Today, this BI solution is used to report and analyze the business activities of the Merchandising Division, which is responsible for all of the Dallas Cowboys' brand sales. Industry estimates say that the Cowboys generate 20% of all NFL merchandise sales, which reflects the fact that they are the most recognized sports franchise in the world.

According to Eric Lai, a *ComputerWorld* reporter, Tony Romo and the rest of the Dallas Cowboys may have been only average on the football field in the last few years, but off the field, especially in the merchandising arena, they remain America's team.

Questions for Discussion

1. How did the Dallas Cowboys use information visualization?
2. What were the challenge, the proposed solution, and the obtained results?

Sources: Lai, E. (2009, October 8). BI visualization tool helps Dallas Cowboys sell more Tony Romo jerseys. *ComputerWorld*; Tableau case study. tableausoftware.com/learn/stories/ tableau-and-teknion-exceed-cowboys-requirements (accessed July 2016).

Dashboard Design

Dashboards are not a new concept. Their roots can be traced at least to the executive information system of the 1980s. Today, dashboards are ubiquitous. For example, a few years back, Forrester Research estimated that over 40% of the largest 2,000 companies in the world used the technology (Ante & McGregor, 2006). Since then, one can safely assume that this number has gone up quite significantly. In fact, nowadays it would be rather unusual to see a large company using a BI system that does not employ some sort of performance dashboards. The Dashboard Spy Web site (dashboardspy.com/about) provides further evidence of their ubiquity. The site contains descriptions and screenshots of thousands of BI dashboards, scorecards, and BI interfaces used by businesses of all sizes and industries, nonprofits, and government agencies.

According to Eckerson (2006), a well-known expert on BI in general and dashboards in particular, the most distinctive feature of a dashboard is its three layers of information:

1. ***Monitoring:*** Graphical, abstracted data to monitor key performance metrics.
2. ***Analysis:*** Summarized dimensional data to analyze the root cause of problems.
3. ***Management:*** Detailed operational data that identify what actions to take to resolve a problem.

Because of these layers, dashboards pack a lot of information into a single screen. According to Few (2005), "The fundamental challenge of dashboard design is to display all the required information on a single screen, clearly and without distraction, in a manner that can be assimilated quickly." To speed assimilation of the numbers, the numbers need to be placed in context. This can be done by comparing the numbers of interest to other baseline or target numbers, by indicating whether the numbers are good or bad, by denoting whether a trend is better or worse, and by using specialized display widgets or components to set the comparative and evaluative context. Some of the common comparisons that are typically made in BI systems include comparisons against past values, forecasted values, targeted values, benchmark or average values, multiple instances of the same measure, and the values of other measures (e.g., revenues versus costs).

Even with comparative measures, it is important to specifically point out whether a particular number is good or bad and whether it is trending in the right direction. Without these types of evaluative designations, it can be time consuming to determine the status of a particular number or result. Typically, either specialized visual objects (e.g., traffic lights, dials, and gauges) or visual attributes (e.g., color coding) are used to set the evaluative context. An interactive dashboard-driven reporting data exploration solution built by an energy company is featured in Application Case 2.8.

Application Case 2.8

Visual Analytics Helps Energy Supplier Make Better Connections

Energy markets all around the world are going through a significant change and transformation, creating ample opportunities along with significant challenges. As is the case in any industry, opportunities are attracting more players in the marketplace, increasing the competition, and reducing the tolerances for less-than-optimal business decision making. Success requires creating and disseminating accurate and timely information to whomever and whenever it is needed. For instance, if you need to easily track marketing budgets, balance employee workloads, and target customers with tailored marketing messages, you would need three different reporting solutions. Electrabel GDF SUEZ is doing all of that for its marketing and sales business unit with SAS® Visual Analytics platform.

(*Continued*)

Application Case 2.8 (Continued)

The one-solution approach is a great time-saver for marketing professionals in an industry that is undergoing tremendous change. "It is a huge challenge to stabilize our market position in the energy market. That includes volume, prices, and margins for both retail and business customers," notes Danny Noppe, Reporting Architecture and Development Manager in the Electrabel Marketing and Sales business unit. The company is the largest supplier of electricity in Belgium and the largest producer of electricity for Belgium and the Netherlands. Noppe says it is critical that Electrabel increase the efficiency of its customer communications as it explores new digital channels and develops new energy-related services.

"The better we know the customer, the better our likelihood of success," he says. "That is why we combine information from various sources—phone traffic with the customer, online questions, text messages, and mail campaigns. This enhanced knowledge of our customer and prospect base will be an additional advantage within our competitive market."

One Version of the Truth

Electrabel was using various platforms and tools for reporting purposes. This sometimes led to ambiguity in the reported figures. The utility also had performance issues in processing large data volumes. SAS Visual Analytics with in-memory technology removes the ambiguity and the performance issues. "We have the autonomy and flexibility to respond to the need for customer insight and data visualization internally," Noppe says. "After all, fast reporting is an essential requirement for action-oriented departments such as sales and marketing."

Working More Efficiently at a Lower Cost

SAS Visual Analytics automates the process of updating information in reports. Instead of building a report that is out of date by the time it is completed, the data is refreshed for all the reports once a week and is available on dashboards. In deploying the solution, Electrabel chose a phased approach starting with simple reports and moving on to more complex ones. The first report took a few weeks to build, and the rest came quickly. The successes include the following:

- Data that took 2 days to prepare now takes only 2 hours.
- Clear graphic insight into the invoicing and composition of invoices for B2B customers.
- A workload management report by the operational teams. Managers can evaluate team workloads on a weekly or long-term basis and can make adjustments accordingly.

"We have significantly improved our efficiency and can deliver quality data and reports more frequently, and at a significantly lower cost," says Noppe. And if the company needs to combine data from multiple sources, the process is equally easy. "Building visual reports, based on these data marts, can be achieved in a few days, or even a few hours."

Noppe says the company plans to continue broadening its insight into the digital behavior of its customers, combining data from Web analytics, e-mail, and social media with data from back-end systems. "Eventually, we want to replace all labor-intensive reporting with SAS Visual Analytics," he says, adding that the flexibility of SAS Visual Analytics is critical for his department. "This will give us more time to tackle other challenges. We also want to make this tool available on our mobile devices. This will allow our account managers to use up-to-date, insightful, and adaptable reports when visiting customers. "We've got a future-oriented reporting platform to do all we need."

Questions for Discussion

1. Why do you think energy supply companies are among the prime users of information visualization tools?
2. How did Electrabel use information visualization for the single version of the truth?
3. What were their challenges, the proposed solution, and the obtained results?

Source: SAS Customer Story, "Visual analytics helps energy supplier make better connections" at http://www.sas.com/en_us/customers/electrabel-be.html (accessed July 2016). Copyright © 2016 SAS Institute Inc., Cary, NC, USA. Reprinted with permission. All rights reserved.

What to Look for in a Dashboard

Although performance dashboards and other information visualization frameworks differ, they all do share some common design characteristics. First, they all fit within the larger BI and/or performance measurement system. This means that their underlying architecture is the BI or performance management architecture of the larger system. Second, all well-designed dashboard and other information visualizations possess the following characteristics (Novell, 2009):

- They use visual components (e.g., charts, performance bars, sparklines, gauges, meters, stoplights) to highlight, at a glance, the data and exceptions that require action.
- They are transparent to the user, meaning that they require minimal training and are extremely easy to use.
- They combine data from a variety of systems into a single, summarized, unified view of the business.
- They enable drill-down or drill-through to underlying data sources or reports, providing more detail about the underlying comparative and evaluative context.
- They present a dynamic, real-world view with timely data refreshes, enabling the end user to stay up to date with any recent changes in the business.
- They require little, if any, customized coding to implement, deploy, and maintain.

Best Practices in Dashboard Design

The real estate saying "location, location, location" makes it obvious that the most important attribute for a piece of real estate property is where it is located. For dashboards, it is "data, data, data." An often overlooked aspect, data is one of the most important things to consider in designing dashboards (Carotenuto, 2007). Even if a dashboard's appearance looks professional, is aesthetically pleasing, and includes graphs and tables created according to accepted visual design standards, it is also important to ask about the data: Is it reliable? Is it timely? Is any data missing? Is it consistent across all dashboards? Here are some of the experience-driven best practices in dashboard design (Radha, 2008).

Benchmark Key Performance Indicators with Industry Standards

Many customers, at some point in time, want to know if the metrics they are measuring are the right metrics to monitor. Sometimes customers have found that the metrics they are tracking are not the right ones to track. Doing a gap assessment with industry benchmarks aligns you with industry best practices.

Wrap the Dashboard Metrics with Contextual Metadata

Often when a report or a visual dashboard/scorecard is presented to business users, questions remain unanswered. The following are some examples:

- Where did you source this data from?
- While loading the data warehouse, what percentage of the data got rejected/encountered data quality problems?
- Is the dashboard presenting "fresh" information or "stale" information?
- When was the data warehouse last refreshed?
- When is it going to be refreshed next?
- Were any high-value transactions that would skew the overall trends rejected as a part of the loading process?

Validate the Dashboard Design by a Usability Specialist

In most dashboard environments, the dashboard is designed by a tool specialist without giving consideration to usability principles. Even though it's a well-engineered data warehouse that can perform well, many business users do not use the dashboard, as it is perceived as not being user friendly, leading to poor adoption of the infrastructure and change management issues. Up-front validation of the dashboard design by a usability specialist can mitigate this risk.

Prioritize and Rank Alerts/Exceptions Streamed to the Dashboard

Because there are tons of raw data, it is important to have a mechanism by which important exceptions/behaviors are proactively pushed to the information consumers. A business rule can be codified, which detects the alert pattern of interest. It can be coded into a program, using database-stored procedures, which can crawl through the fact tables and detect patterns that need immediate attention. This way, information finds the business user as opposed to the business user polling the fact tables for the occurrence of critical patterns.

Enrich the Dashboard with Business-User Comments

When the same dashboard information is presented to multiple business users, a small text box can be provided that can capture the comments from an end-user's perspective. This can often be tagged to the dashboard to put the information in context, adding perspective to the structured KPIs being rendered.

Present Information in Three Different Levels

Information can be presented in three layers depending on the granularity of the information: the visual dashboard level, the static report level, and the self-service cube level. When a user navigates the dashboard, a simple set of 8 to 12 KPIs can be presented, which would give a sense of what is going well and what is not.

Pick the Right Visual Construct Using Dashboard Design Principles

In presenting information in a dashboard, some information is presented best with bar charts, some with time series line graphs, and when presenting correlations, a scatter plot is useful. Sometimes merely rendering it as simple tables is effective. Once the dashboard design principles are explicitly documented, all the developers working on the front end can adhere to the same principles while rendering the reports and dashboard.

Provide for Guided Analytics

In a typical organization, business users can be at various levels of analytical maturity. The capability of the dashboard can be used to guide the "average" business user to access the same navigational path as that of an analytically savvy business user.

SECTION 2.11 REVIEW QUESTIONS

1. What is an information dashboard? Why are they so popular?
2. What are the graphical widgets commonly used in dashboards? Why?
3. List and describe the three layers of information portrayed on dashboards.
4. What are the common characteristics of dashboards and other information visuals?
5. What are the best practices in dashboard design?

Chapter Highlights

- Data has become one of the most valuable assets of today's organizations.
- Data is the main ingredient for any BI, data science, and business analytics initiative.
- Although its value proposition is undeniable, to live up its promise, the data has to comply with some basic usability and quality metrics.
- *Data* (datum in singular form) refers to a collection of facts usually obtained as the result of experiments, observations, transactions, or experiences.
- At the highest level of abstraction, data can be classified as structured and unstructured.
- Data in its original/raw form is not usually ready to be useful in analytics tasks.
- Data preprocessing is a tedious, time-demanding, yet crucial task in business analytics.
- Statistics is a collection of mathematical techniques to characterize and interpret data.
- Statistical methods can be classified as either descriptive or inferential.
- Statistics in general, and descriptive statistics in particular, is a critical part of BI and business analytics.
- Descriptive statistics methods can be used to measure central tendency, dispersion, or the shape of a given data set.
- Regression, especially linear regression, is perhaps the most widely known and used analytics technique in statistics.
- Linear regression and logistic regression are the two major regression types in statistics.
- Logistics regression is a probability-based classification algorithm.
- Time series is a sequence of data points of a variable, measured and recorded at successive points in time spaced at uniform time intervals.
- A report is any communication artifact prepared with the specific intention of conveying information in a presentable form.
- A business report is a written document that contains information regarding business matters.
- The key to any successful business report is clarity, brevity, completeness, and correctness.
- Data visualization is the use of visual representations to explore, make sense of, and communicate data.
- Perhaps the most notable information graphic of the past was developed by Charles J. Minard, who graphically portrayed the losses suffered by Napoleon's army in the Russian campaign of 1812.
- Basic chart types include line, bar, and pie chart.
- Specialized charts are often derived from the basic charts as exceptional cases.
- Data visualization techniques and tools make the users of business analytics and BI systems better information consumers.
- Visual analytics is the combination of visualization and predictive analytics.
- Increasing demand for visual analytics coupled with fast-growing data volumes led to exponential growth in highly efficient visualization systems investment.
- Dashboards provide visual displays of important information that is consolidated and arranged on a single screen so that information can be digested at a single glance and easily drilled in and further explored.

Key Terms

analytics ready
arithmetic mean
box-and-whiskers plot
box plot
bubble chart
business report
categorical data
centrality
correlation
dashboards
data preprocessing
data quality
data security
data taxonomy
data visualization
datum
descriptive statistics
dimensional reduction
dispersion
high-performance computing
histogram
inferential statistics
key performance indicator (KPI)
knowledge
kurtosis
learning
linear regression
logistic regression
mean absolute deviation
median
mode
nominal data
online analytics processing (OLAP)
ordinal data
ordinary least squares (OLS)
pie chart
quartile
range
ratio data
regression
report
scatter plot
skewness
standard deviation
statistics
storytelling
structured data
time series forecasting
unstructured data
variable selection
variance
visual analytics

Questions for Discussion

1. How do you describe the importance of data in analytics? Can we think of analytics without data? Explain.
2. Considering the new and broad definition of business analytics, what are the main inputs and outputs to the analytics continuum?
3. Where does the data for business analytics come from? What are the sources and the nature of that incoming data?
4. What are the most common metrics that make for analytics-ready data?
5. What are the main categories of data? What types of data can we use for BI and analytics?
6. Can we use the same data representation for all analytics models (i.e., do different analytics models require different data representation schema)? Why, or why not?
7. Why is the original/raw data not readily usable by analytics tasks?
8. What are the main data preprocessing steps? List and explain their importance in analytics.
9. What does it mean to clean/scrub the data? What activities are performed in this phase?
10. Data reduction can be applied to rows (sampling) and/or columns (variable selection). Which is more challenging? Explain.
11. What is the relationship between statistics and business analytics (consider the placement of statistics in a business analytics taxonomy)?
12. What are the main differences between descriptive and inferential statistics?
13. What is a box-and-whiskers plot? What types of statistical information does it represent?
14. What are the two most commonly used shape characteristics to describe a data distribution?
15. List and briefly define the central tendency measures of descriptive statistics?
16. What are the commonalities and differences between regression and correlation?
17. List and describe the main steps to follow in developing a linear regression model.
18. What are the most commonly pronounced assumptions for linear regression? What is crucial to the regression models against these assumptions?
19. What are the commonalities and differences between linear regression and logistic regression?
20. What is time series? What are the main forecasting techniques for time series data?
21. What is a business report? Why is it needed?
22. What are the best practices in business reporting? How can we make our reports stand out?
23. Describe the cyclic process of management, and comment on the role of business reports.
24. List and describe the three major categories of business reports.
25. Why has information visualization become a centerpiece in BI and business analytics? Is there a difference between information visualization and visual analytics?
26. What are the main types of charts/graphs? Why are there so many of them?
27. How do you determine the right chart for the job? Explain and defend your reasoning.
28. What is the difference between information visualization and visual analytics?
29. Why should storytelling be a part of your reporting and data visualization?
30. What is an information dashboard? What do they present?
31. What are the best practices in designing highly informative dashboards?
32. Do you think information/performance dashboards are here to stay? Or are they about to be outdated? What do you think will be the next big wave in BI and business analytics in terms of data/information visualization?

Exercises

Teradata University and Other Hands-on Exercises

1. Download the "Voting Behavior" data and the brief data description from the book's Web site. This is a data set manually compiled from counties all around the United States. The data is partially processed, that is, some derived variables are created. Your task is to thoroughly preprocess the data by identifying the error and anomalies and proposing remedies and solutions. At the end you should have an analytics-ready version of this data. Once the preprocessing is completed, pull this data into Tableau (or into some other data visualization software tool) to extract useful visual information from it. To do so, conceptualize relevant questions and hypotheses (come up with at least three of them) and create proper visualizations that address those questions of "tests" of those hypotheses.
2. Download Tableau (at tableau.com, following academic free software download instructions on their site). Using the Visualization_MFG_Sample data set (available as an Excel file on this book's Web site) answer the following questions:
   a. What is the relationship between gross box office revenue and other movie-related parameters given in the data set?

b. How does this relationship vary across different years? Prepare a professional-looking written report that is enhanced with screenshots of your graphic findings.

3. Go to teradatauniversitynetwork.com. Look for an article that deals with the nature of data, management of data, and/or governance of data as it relates to BI and analytics, and critically analyze the content of the article.
4. Go to UCI data repository (archive.ics.uci.edu/ml/datasets.html), and identify a large data set that contains both numeric and nominal values. Using Microsoft Excel, or any other statistical software:
   a. Calculate and interpret central tendency measures for each and every variable.
   b. Calculate and interpret the dispersion/spread measures for each and every variable.
5. Go to UCI data repository (archive.ics.uci.edu/ml/datasets.html), and identify two data sets, one for estimation/regression and one for classification. Using Microsoft Excel, or any other statistical software:
   a. Develop and interpret a linear regression model.
   b. Develop and interpret a logistic regression model.
6. Go to KDnuggest.com, and become familiar with the range of analytics resources available on this portal. Then, identify an article, a white paper, or an interview script that deals with the nature of data, management of data, and/or governance of data as it relates to BI and business analytics, and critically analyze the content of the article.
7. Go to Stephen Few's blog, "The Perceptual Edge" (perceptualedge.com). Go to the section of "Examples." In this section, he provides critiques of various dashboard examples. Read a handful of these examples. Now go to dundas.com. Select the "Gallery" section of the site. Once there, click the "Digital Dashboard" selection. You will be shown a variety of different dashboard demos. Run a couple of the demos.
   a. What sorts of information and metrics are shown on the demos? What sorts of actions can you take?
   b. Using some of the basic concepts from Few's critiques, describe some of the good design points and bad design points of the demos.
8. Download an information visualization tool, such as Tableau, QlikView, or Spotfire. If your school does not have an educational agreement with these companies, then a trial version would be sufficient for this exercise. Use your own data (if you have any) or use one of the data sets that comes with the tool (they usually have one or more data sets for demonstration purposes). Study the data, come up with a couple of business problems, and use data visualization to analyze, visualize, and potentially solve those problems.
9. Go to teradatauniversitynetwork.com. Find the "Tableau Software Project." Read the description, execute the tasks, and answer the questions.
10. Go to teradatauniversitynetwork.com. Find the assignments for SAS Visual Analytics. Using the information and step-by-step instructions provided in the assignment, execute the analysis on the SAS Visual Analytics tool (which is a Web-enabled system that does not require any local installation). Answer the questions posed in the assignment.
11. Find at least two articles (one journal article and one white paper) that talk about storytelling, especially within the context of analytics (i.e., data-driven storytelling). Read and critically analyze the article and paper, and write a report to reflect your understanding and opinions about the importance of storytelling in BI and business analytics.
12. Go to Data.gov—a U.S. government–sponsored data portal that has a very large number of data sets on a wide variety of topics ranging from healthcare to education, climate to public safety. Pick a topic that you are most passionate about. Go through the topic-specific information and explanation provided on the site. Explore the possibilities of downloading the data, and use your favorite data visualization tool to create your own meaningful information and visualizations.

Team Assignments and Role-Playing Projects

1. Analytics starts with data. Identifying, accessing, obtaining, and processing of relevant data are the most essential tasks in any analytics study. As a team, you are tasked to find a large enough real-world data (either from your own organization, which is the most preferred, or from the Internet that can start with a simple search, or from the data links posted on KDnuggets.com), one that has tens of thousands of rows and more than 20 variables to go through and document a thorough data preprocessing project. In your processing of the data, identify anomalies and discrepancies using descriptive statistics methods and measures, and make the data analytics ready. List and justify your preprocessing steps and decisions in a comprehensive report.
2. Go to a well-known information dashboard provider Web site (dundas.com, idashboards.com, enterprise-dashboard.com). These sites provide a number of examples of executive dashboards. As a team, select a particular industry (e.g., healthcare, banking, airline). Locate a handful of example dashboards for that industry. Describe the types of metrics found on the dashboards. What types of displays are used to provide the information? Using what you know about dashboard design, provide a paper prototype of a dashboard for this information.
3. Go to teradatauniversitynetwork.com. From there, go to University of Arkansas data sources. Choose one of the large data sets, and download a large number of records (this may require you to write an SQL statement that creates the variables that you want to include in the data set). Come up with at least 10 questions that can be addressed with information visualization. Using your favorite data visualization tool (e.g., Tableau), analyze the data, and prepare a detailed report that includes screenshots and other visuals.

References

Abela, A. (2008). *Advanced presentations by design: Creating communication that drives action*. New York: Wiley.

Annas, G. J. (2003). HIPAA regulations—A new era of medical-record privacy? *New England Journal of Medicine*, *348*(15), 1486–1490.

Ante, S. E., & McGregor, J. (2006). Giving the boss the big picture: A dashboard pulls up everything the CEO needs to run the show. *Business Week*, 43–51.

Carotenuto, D. (2007). Business intelligence best practices for dashboard design. WebFOCUS white paper. www.datawarehouse.inf.br/papers/information_builders_dashboard_best_practices.pdf (accessed August 2016).

Dell customer case study. Medical device company ensures product quality while saving hundreds of thousands of dollars. https://software.dell.com/documents/instrumentation-laboratory-medical-device-companyensures-product-quality-while-saving-hundreds-ofthousands-of-dollars-case-study-80048.pdf (accessed August 2016).

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems, 49*(4), 498–506.

Delen, D. (2011). Predicting student attrition with data mining methods. *Journal of College Student Retention 13*(1), 17–35.

Delen, D., Cogdell, D., & Kasap, N. (2012). A comparative analysis of data mining methods in predicting NCAA bowl outcomes. *International Journal of Forecasting, 28*, 543–552.

Delen, D. (2015). *Real-world data mining: Applied business analytics and decision making*. Upper Saddle River, NJ: Financial Times Press (A Pearson Company).

Eckerson, W. (2006). *Performance dashboards*. New York: Wiley.

Few, S. (2005, Winter). Dashboard design: Beyond meters, gauges, and traffic lights. *Business Intelligence Journal*, *10*(1).

Few, S. (2007). Data visualization: Past, present and future. perceptualedge.com/articles/Whitepapers/Data_Visualization.pdf (accessed July 2016).

Fink, E., & Moore, S. J. (2012). Five best practices for telling great stories with data. White paper by Tableau Software, Inc., www.tableau.com/whitepapers/telling-data-stories (accessed May 2016).

Freeman, K. M., & Brewer, R. M. (2016). The politics of American college football. *Journal of Applied Business and Economics, 18*(2), 97–101.

Gartner Magic Quadrant, released on February 4, 2016, gartner.com (accessed August 2016).

Grimes, S. (2009a, May 2). Seeing connections: Visualizations makes sense of data. *Intelligent Enterprise*. i.cmpnet.com/intelligententerprise/next-era-business-intelligence/Intelligent_Enterprise_Next_Era_BI_Visualization.pdf (accessed January 2010).

Grimes, S. (2009b). Text analytics 2009: User perspectives on solutions and providers. Alta Plana. altaplana.com/TextAnalyticsPerspectives2009.pdf (accessed July, 2016).

Hardin, M., Hom, D., Perez, R., & Williams, L. (2012). Which chart or graph is right for you? Tableau Software: Tell Impactful Stories with Data‖. Tableau Software. http://www.tableau.com/sites/default/files/media/which_chart_v6_final_0.pdf (accessed August 2016).

Hernández, M. A., & Stolfo, S. J. (1998, January). Real-world data is dirty: Data cleansing and the merge/purge problem. *Data Mining and Knowledge Discovery, 2*(1), 9–37.

Hill, G. (2016). A Guide to enterprise reporting. ghill.customer.netspace.net.au/reporting/definition.html (accessed July 2016).

Kim, W., Choi, B. J., Hong, E. K., Kim, S. K., & Lee, D. (2003). A taxonomy of dirty data. *Data Mining and Knowledge Discovery, 7*(1), 81–99.

Kock, N. F., McQueen, R. J., & Corner, J. L. (1997). The nature of data, information and knowledge exchanges in business processes: Implications for process improvement and organizational learning. *The Learning Organization, 4*(2), 70–80.

Kotsiantis, S. B., Kanellopoulos, D., & Pintelas, P. E. (2006). Data preprocessing for supervised leaning. *International Journal of Computer Science, 1*(2), 111–117.

Lai, E. (2009, October 8). BI visualization tool helps Dallas Cowboys sell more Tony Romo jerseys. *ComputerWorld*.

Quinn, C. (2016). Data-driven marketing at SiriusXM. Teradata Articles & News. at http://bigdata.teradata.com/US/Articles-News/Data-Driven-Marketing-At-SiriusXM/ (accessed August 2016); Teradata customer success story. SiriusXM attracts and engages a new generation of radio consumers. http://assets.teradata.com/resourceCenter/downloads/CaseStudies/EB8597.pdf?processed=1.

Novell. (2009, April). Executive dashboards elements of success. Novell white paper. www.novell.com/docrep/documents/3rkw3etfc3/Executive%20Dashboards_Elements_of_Success_White_Paper_en.pdf (accessed June 2016).

Radha, R. (2008). Eight best practices in dashboard design. *Information Management*. www.information-management.com/news/columns/-10001129-1.html (accessed July 2016).

SAS. (2014). Data visualization techniques: From basics to Big Data. http://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/data-visualization-techniques-106006.pdf (accessed July 2016).

Thammasiri, D., Delen, D., Meesad, P., & Kasap N. (2014). A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. *Expert Systems with Applications, 41*(2), 321–330.