

Unit 1 – Introduction to Business Analytics

What is business analytics?

Business analytics (BA) is a set of disciplines and technologies for solving business problems using data analysis, statistical models and other quantitative methods. It involves an iterative, methodical exploration of an organization's data, with an emphasis on statistical analysis, to drive decision-making.

Data-driven companies treat their data as a business asset and actively look for ways to turn it into a competitive advantage. Success with business analytics depends on data quality, skilled analysts who understand the technologies and the business, and a commitment to using data to gain insights that inform business decisions.

How business analytics works

Before any data analysis takes place, BA starts with several foundational processes:

- Determine the business goal of the analysis.
- Select an analysis methodology.
- Get business data to support the analysis, often from various systems and sources.
- Cleanse and integrate data into a single repository, such as a data warehouse or data mart.

Initial analysis is typically performed on a smaller sample data set of data. Analytics tools range from spreadsheets with statistical functions to complex data mining and predictive modeling applications. Patterns and relationships in the raw data are revealed. Then new questions are asked, and the analytic process iterates until the business goal is met.

Deployment of predictive models involves a statistical process known as scoring and uses records typically located in a database. Scores help enterprises make more informed, real-time decisions within applications and business processes.

BA also supports tactical decision-making in response to unforeseen events. Often the decision-making is automated using artificial intelligence to support real-time responses.

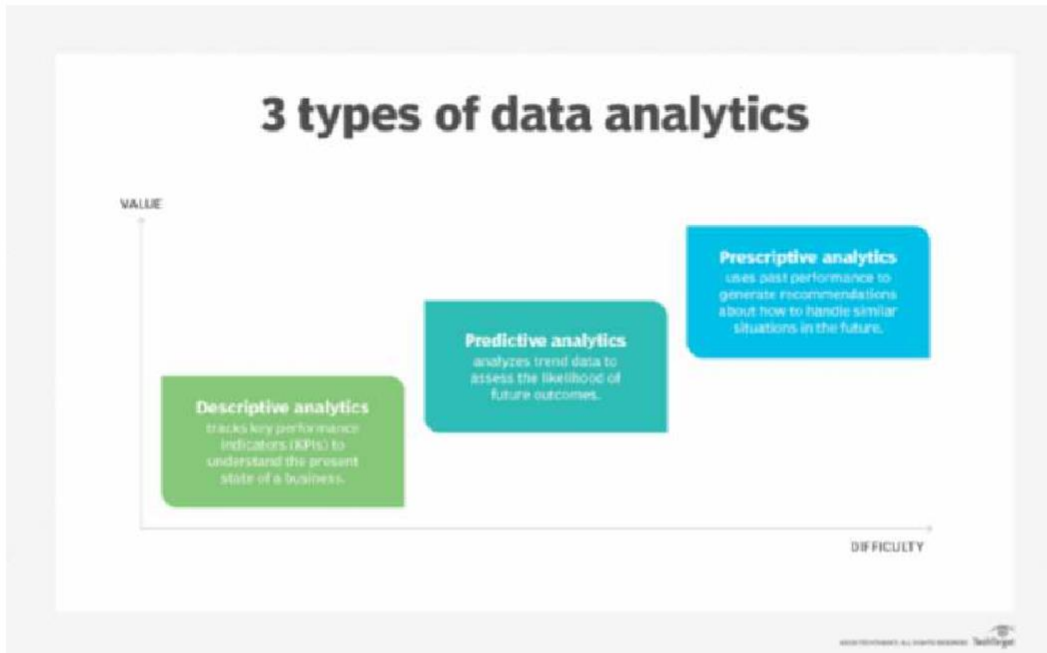
Types of business analytics

Different types of business analytics include the following:

descriptive analytics, which tracks key performance indicators (KPIs) to understand the present state of a business;

predictive analytics, which analyzes trend data to assess the likelihood of future outcomes; and

prescriptive analytics, which uses past performance to generate recommendations for handling similar situations in the future.



Business analytics vs. business intelligence

The terms business intelligence (BI) and business analytics are often used interchangeably. However, there are key differences.

Companies usually start with BI before implementing business analytics. BI analyzes business operations to determine what practices have worked and where opportunities for improvement lie. BI uses descriptive analytics.

In contrast, business analytics focuses on predictive analytics, generating actionable insights for decision-makers. Instead of summarizing past data points, BA aims to predict trends.

The data collected using BI lays the groundwork for BA. From that data, companies can choose specific areas to analyze further using business analytics.

Business analytics vs. data analytics

Data analytics is the analysis of data sets to draw conclusions about the information they contain. Data analytics does not have to be used in pursuit of business goals or insights. It is a broader practice that includes business analytics.

BA involves using data analytics tools in pursuit of business insights. However, because it's a general term, data analytics is sometimes used interchangeably with business analytics.

Business analytics vs. data science

Data science uses analytics to inform decision-making. Data scientists explore data using advanced statistical methods. They allow the features in the data to guide their analysis. The more advanced areas of business analytics resemble data science, but there is a distinction between what data scientists and business analysts do.

Even when advanced statistical algorithms are applied to data sets, it doesn't necessarily mean data science is involved. That's because true data science uses custom coding and explores answers to open-ended questions. In contrast, business analytics aims to solve a specific question or problem.

Common challenges of business analytics

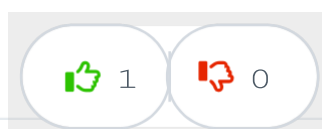
Businesses might encounter both business analytics and business intelligence challenges when trying to implement a business analytics strategy:

- **Too many data sources.** There is an increasingly large spectrum of internet-connected devices generating business data. In many cases, they are generating different types of data that must be integrated into an analytics strategy. However, the more complex a data set becomes, the harder it is to use it as part of an analytics framework.
- **Lack of skills.** The demand for employees with the data analytic skills necessary to process BA data has grown. Some businesses, particularly small and medium-sized businesses (SMBs), may have a hard time hiring people with the BA expertise and skills they need.
- **Data storage limitations.** Before a business can begin to decide how it will process data, it must decide where to store it. For instance, a data lake can be used to capture large volumes of unstructured data.

Roles and responsibilities in business analytics

Business analytics professionals' main responsibility is to collect and analyze data to influence strategic decisions that a business makes. Some initiatives they might provide analysis for include the following:

- identifying strategic opportunities from data patterns;



- identifying potential problems facing the business and solutions;
- creating a budget and business forecast;
- monitoring progress with business initiatives;
- reporting progress on business objectives back to stakeholders;
- understanding KPIs; and
- understanding regulatory and reporting requirements

Terminologies in Business Analytics

Analytics – Analytics can simply be defined as the process of breaking a problem into simpler parts and using inferences based on data to drive decisions. Analytics is not a tool or a technology; rather it is a way of thinking and acting.

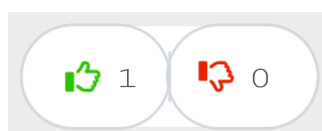
Analytics has widespread applications in spheres as diverse as science, astronomy, genetics, financial services, telecom, retail, marketing, sports, gaming and health care.

Business analytics – This term refers to the application of analytics specifically in the sphere of business. It includes subsets like –

- Marketing analytics
- Risk analytics
- Fraud analytics
- CRM analytics
- Loyalty analytics
- Operations analytics
- HR analytics

Industries which rely extensively on analytics include –

- Financial Services (Banks, Credit Cards, Loans, Insurance etc.)
- Retail
- Telecom
- Health care
- Consumer goods
- Manufacturing
- Sports



- Hotels
- Airlines
- Any industry where large amounts of data is generated

Predictive Analytics – Predictive analytics is one of the most popular analytics terms. Predictive analytics is used to make predictions on the likelihood of occurrence of an event or determine some future patterns based on data. Remember it does not tell whether an event will happen. It only assigns probabilities to the future events or patterns.

Google Trends analysis of “predictive analytics”

The term emphasizes the predictive nature of analytics (as opposed to, say the retrospective nature of tools like OLAP). This is one of those terms that is designed by sales people and marketers to add glamour to any business. “Predictive analytics” sounds fancier than just plain “analytics”. In practise, predictive analytics is rarely used in isolation from descriptive analytics.

Descriptive analytics – Descriptive analytics refers to a set of techniques used to describe or explore or profile any kind of data. Any kind of reporting usually involves descriptive analytics. Data exploration and data preparation are essential ingredients for predictive modelling and these rely heavily on descriptive analytics.

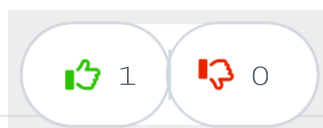
Inquisitive analytics – Whereas descriptive analytics is used for data presentation and exploration, inquisitive analytics answers terms why, what, how and what if. Ex: Why have the sales in the Q4 dropped could be a question based on which inquisitive analysis can be performed on the data

Advanced analytics – Like “Predictive analytics”, “Advanced analytics” too is a marketing driven terminology. “Advanced” adds a little more punch, a little more glamour to “Analytics” and is preferred by marketers.

Big data analytics – When analytics is performed on large data sets with huge volume, variety and velocity of data it can be termed as big data analytics. The annual amount of data we have is expected to grow from 8 zettabytes (trillion gigabytes) in 2015 to 35 zettabytes in 2020.

Growing data sizes would inevitably require advanced technology like Hadoop and Map Reduce to store and map large chunks of data. Also, large variety of data (structured, unstructured) is flowing in at a very rapid pace. This would not only require advance technology but also advanced analytical platforms. So to summarize, large amounts of data together with the technology and the analytics platforms to get insights out of such a data can be called as the Big data analytics.

Data Mining – Data mining is the term that is most interchangeably used with “Analytics”. Data Mining is an older term that was more popular in the nineties and the early 2000s.



However, data mining began to be confused with OLAP and that led to a drive to use more descriptive terms like “Predictive analytics”.

According to Google trends, “Analytics” overtook “Data mining” in popularity at some point in 2005 and is about 5 times more popular now. Incidentally, Coimbatore is one of the only cities in the world where “**Data mining**” is still more popular than “**Analytics**”.

Data Science – Data science and data analytics are mostly used interchangeably. However, sometimes a data scientist is expected to possess higher mathematical and statistical sophistication than a data analyst. A Data scientist is expected to be well versed in linear algebra, calculus, machine learning and should be able to navigate the nitty-gritty details of mathematics and statistics with much ease.

Artificial Intelligence –During the early stages of computing, there were a lot of comparisons between computing and human learning process and this is reflected in the terminology.

The term “Artificial intelligence” was popular in the very early stages of computing and analytics (in the 70s and 80s) but is now almost obsolete.

Machine learning – involves using statistical methods to create algorithms. It replaces explicit programming which can become cumbersome due to the large amounts of data, inflexible to adapt to the solution requirements and also sometimes illegible.

It is mostly concerned with the algorithms which can be a black box to interpret but good models can give highly accurate results compared to conventional statistical methods. Also, visualization, domain knowledge etc. are not inclusive when we speak about machine learning. Neural networks, support vector machines etc. are the terms which are generally associated with the machine learning algorithms

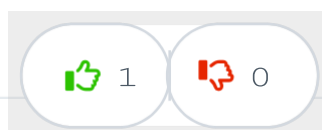
Algorithm – Usually refers to a mathematical formula which is output from the tools. The formula summarizes the model

Ex: Amazon recommendation algorithm gives a formula that can recommend the next best buy

Machine Learning – Similar to “Artificial intelligence” this term too has lost its popularity in the recent past to terms like “Analytics” and its derivatives.

OLAP – Online analytical processing refers to descriptive analytic techniques of slicing and dicing the data to understand it better and discover patterns and insights. The term is derived from another term “OLTP” – online transaction processing which comes from the data warehousing world.

Reporting – The term “Reporting” is perhaps the most unglamorous of all terms in the world of analytics. Yet it is also one of the most widely used practices within the field. All businesses use reporting to aid decision making. While it is not “Advanced analytics” or even



“Predictive analytics”, effective reporting requires a lot of skill and a good understanding of the data as well as the domain.

Data warehousing – Ok, this may actually be considered more unglamorous than even “Reporting”. Data warehousing is the process of managing a database and involves extraction, transformation and loading (ETL) of data. Data warehousing precedes analytics. The data managed in a data warehouse is usually taken out and used for business analytics.

Statistics – Statistics is the study of the collection, organization, and interpretation of data. Data mining does not replace traditional statistical techniques. Rather, it is an extension of statistical

methods that is in part the result of a major change in the statistics community. The development of

most statistical techniques was, until recently, based on elegant theory and analytical methods that

worked quite well on the modest amounts of data being analyzed. The increased power of computers and their lower cost, coupled with the need to analyze enormous data sets with millions of rows, have allowed the development of new techniques based on a brute-force exploration of possible solutions.

Analytics platform – Software that provides for the computation required to carry out the statistical methods, descriptive and inquisitive queries, machine learning, visualization and Big data (which is software plus hardware).

Ex: SAS, R, Tableau, Hadoop etc.

Clickstream analytics/ Web analytics – Analysis on user imprints created on the web

Ex: Number of clicks, probability to buy based on search times of a particular word etc.

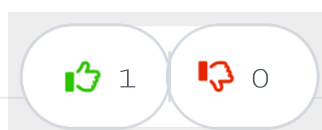
Text analytics – Usually refers to analysing unstructured (not tabulated) data in the form of continuous text.

Ex: Facebook data analysis, twitter analysis etc.

Location analytics – With advanced GPS and location data available location analytics has become quite popular

Ex: Offers based on customer location, insurance risk calculations based on proximity to hazards

Sports analytics – Analysis of sports data using analytical tool and methods. Performance as well as revenue data can be subjected to analytical procedures to achieve better results



The 7-step Business Analytics Process

Real-time analysis is an emerging business tool that is changing the traditional ways enterprises do business. More and more organisations are today exploiting business analytics to enable proactive decision making; in other words, they are switching from reacting to situations to anticipating them.

One of the reasons for the flourishing of business analytics as a tool is that it can be applied in any industry where data is captured and accessible. This data can be used for a variety of reasons, ranging from improving customer service as well improving the organisation's capability to predict fraud to offering valuable insights on online and digital information.

However business analytics is applied, the key outcome is the same: The solving of business problems using the relevant data and turning it into insights, providing the enterprise with the knowledge it needs to proactively make decisions. In this way the enterprise will gain a competitive advantage in the marketplace.

So what is business analytics? Essentially, business analytics is a 7-step process, outlined below.

Step 1. Defining the business needs



The first stage in the business analytics process involves understanding what the business would like to improve on or the problem it wants solved. Sometimes, the goal is broken down



1



0

into smaller goals. Relevant data needed to solve these business goals are decided upon by the business stakeholders, business users with the domain knowledge and the business analyst. At this stage, key questions such as, “what data is available”, “how can we use it”, “do we have sufficient data” must be answered.

Step 2. Explore the data

This stage involves cleaning the data, making computations for missing data, removing outliers, and transforming combinations of variables to form new variables. Time series graphs are plotted as they are able to indicate any patterns or outliers. The removal of outliers from the dataset is a very important task as outliers often affect the accuracy of the model if they are allowed to remain in the data set. As the saying goes: Garbage in, garbage out (GIGO)!

Once the data has been cleaned, the analyst will try to make better sense of the data. The analyst will plot the data using scatter plots (to identify possible correlation or non-linearity). He will visually check all possible slices of data and summarise the data using appropriate visualisation and descriptive statistics (such as mean, standard deviation, range, mode, median) that will help provide a basic understanding of the data. At this stage, the analyst is already looking for general patterns and actionable insights that can be derived to achieve the business goal.

Step 3. Analyse the data

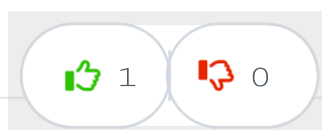
At this stage, using statistical analysis methods such as correlation analysis and hypothesis testing, the analyst will find all factors that are related to the target variable. The analyst will also perform simple regression analysis to see whether simple predictions can be made. In addition, different groups are compared using different assumptions and these are tested using hypothesis testing. Often, it is at this stage that the data is cut, sliced and diced and different comparisons are made while trying to derive actionable insights from the data.

Step 4. Predict what is likely to happen

Business analytics is about being proactive in decision making. At this stage, the analyst will model the data using predictive techniques that include decision trees, neural networks and logistic regression. These techniques will uncover insights and patterns that highlight relationships and ‘hidden evidences’ of the most influential variables. The analyst will then compare the predictive values with the actual values and compute the predictive errors. Usually, several predictive models are ran and the best performing model selected based on model accuracy and outcomes.

Step 5. Optimise (find the best solution)

At this stage the analyst will apply the predictive model coefficients and outcomes to run ‘what-if’ scenarios, using targets set by managers to determine the best solution, with the given constraints and limitations. The analyst will select the optimal solution and model



based on the lowest error, management targets and his intuitive recognition of the model coefficients that are most aligned to the organisation's strategic goal.

Step 6. Make a decision and measure the outcome

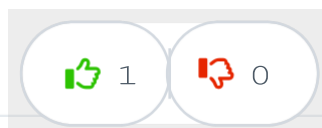
The analyst will then make decisions and take action based on the derived insights from the model and the organisational goals. An appropriate period of time after this action has been taken, the outcome of the action is then measured.

Step 7. Update the system with the results of the decision

Finally the results of the decision and action and the new insights derived from the model are recorded and updated into the database. Information such as, 'was the decision and action effective?', 'how did the treatment group compare with the control group?' and 'what was the return on investment?' are uploaded into the database. The result is an evolving database that is continuously updated as soon as new insights and knowledge are derived.

Importance of business analytics

- Organizations employ Business analytics so they can make data-driven decisions. Business analytics gives business an excellent overview and insight on how companies can become more efficient, and these insights will enable such business optimize and automate their processes. It is no surprise that data-driven companies, and also make use of business analytics usually outperform their contemporaries. The reason for this is that the insights gained via business analytics enable them to; understand why specific results are achieved, explore more effective business processes, and even predict the likelihood of certain results.
- Business analytics also offers adequate support and coverage for businesses who are looking to make the right proactive decisions. Business analytics also allows organizations to automate their entire decision-making process, so as to deliver real-time responses when needed.
- One of the apparent importance of business analytics is the fact that it helps to gain essential business insights. It does this by presenting the right data to work it. This goes a long way in making decision making more efficient, but also easy.
- Efficiency is one area of business analytics helps any organization to achieve immediately. Since its inception, business analytics have played a key role in helping business improve their efficiency. Business analytics collates a considerable volume of data in a timely manner, and also in a way that it can easily be analyzed. This allows businesses to make the right decisions faster.
- Business analytics help organizations to reduce risks. By helping them make the right decisions based on available data such as customer preferences, trends, and so on, it can help businesses to curtail short and long-term risk.



- Business analytics is a methodology or tool to make a sound commercial decision. Hence it impacts functioning of the whole organization. Therefore, business analytics can help improve profitability of the business, increase market share and revenue and provide better return to a shareholder.
- Facilitates better understanding of available primary and secondary data, which again affect operational efficiency of several departments.
- Provides a competitive advantage to companies. In this digital age flow of information is almost equal to all the players. It is how this information is utilized makes the company competitive. Business analytics combines available data with various well thought models to improve business decisions.
- Converts available data into valuable information. This information can be presented in any required format, comfortable to the decision maker.

Evolution of Business Analytics

Business analytics has been existence since very long time and has evolved with availability of newer and better technologies. It has its roots in operations research, which was extensively used during World War II. Operations research was an analytical way to look at data to conduct military operations. Over a period of time, this technique started getting utilized for business. Here operation's research evolved into management science. Again, basis for management science remained same as operation research in data, decision making models, etc.

As the economies started developing and companies became more and more competitive, management science evolved into business intelligence, decision support systems and into PC software.

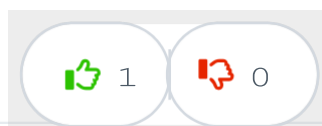
Scope of Business Analytics

Business analytics has a wide range of application and usages. It can be used for descriptive analysis in which data is utilized to understand past and present situation. This kind of descriptive analysis is used to asses' current market position of the company and effectiveness of previous business decision.

It is used for predictive analysis, which is typical used to asses' previous business performance.

Business analytics is also used for prescriptive analysis, which is utilized to formulate optimization techniques for stronger business performance.

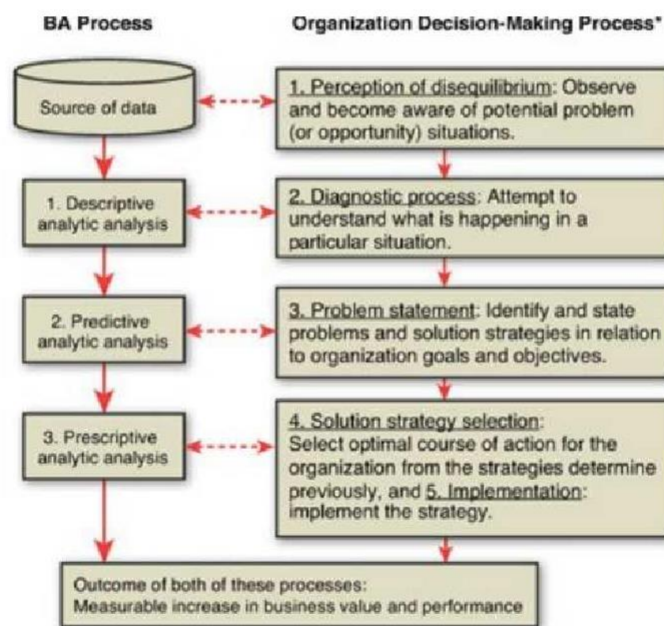
For example, business analytics is used to determine pricing of various products in a departmental store based past and present set of information.



Relationship of BA Process and Organization Decision-Making Process

The BA process can solve problems and identify opportunities to improve business performance. In the process, organizations may also determine strategies to guide operations and help achieve competitive advantages. Typically, solving problems and identifying strategic opportunities to follow are organization decision-making tasks. The latter, identifying opportunities, can be viewed as a problem of strategy choice requiring a solution. It should come as no surprise that the BA process closely parallels classic organization decision-making processes. As depicted in below shown Figure, the business analytic process has an inherent relationship to the steps in typical organization decision-making processes.

Figure: Comparison of business analytics and organization decision-making processes



*Source: Adapted from Figure 1 in Elbing (1970), pp. 12-13.

The organization decision-making process (ODMP) developed by Elbing (1970) and presented in Figure 1.2 is focused on decision making to solve problems but could also be applied to finding opportunities in data and deciding what is the best course of action to take advantage of them. The five-step ODMP begins with the perception of disequilibrium, or the awareness that a problem exists that needs a decision. Similarly, in the BA process, the first step is to recognize that databases may contain information that could both solve problems



and find opportunities to improve business performance. Then in Step 2 of the ODMP, an exploration of the problem to determine its size, impact, and other factors is undertaken to diagnose what the problem is. Likewise, the BA descriptive analytic analysis explores factors that might prove useful in solving problems and offering opportunities. The ODMP problem statement step is similarly structured to the BA predictive analysis to find strategies, paths, or trends that clearly define a problem or opportunity for an organization to solve problems. Finally, the ODMP's last steps of strategy selection and implementation involve the same kinds of tasks that the BA process requires in the final prescriptive step (make an optimal selection of resource allocations that can be implemented for the betterment of the organization).

The decision-making foundation that has served ODMP for many decades parallels the BA process. The same logic serves both processes and supports organization decision-making skills and capacities.

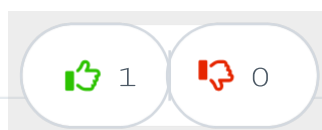
Business analytics is the process of gathering data, measuring business performance, and producing valuable conclusions that can help companies make informed decisions on the future of the business, through the use of various statistical methods and techniques.

Analytics has become one of the most important tools at an organization's disposal. When data and analytics work hand in hand, the benefits become obvious. Companies can leverage data to improve cost savings, redefine processes, drive market strategy, establish competitive differentiators and, perhaps most importantly, build exceptional and truly personalized customer experience.

The Competitive Advantage of Business Analytics

Business analytics for organisations is becoming a competitive advantage and is now necessary to apply business analytics, particularly its subset of predictive business analytics. The use of business analytics is a skill that is gaining mainstream value due to the increasingly thinner margin for decision error. It is there to provide insights, predict the future of the business and inferences from the treasure chest of raw transactional data, that is internal and external data that many organizations now store (and will continue to store) as soft copy.

Business analytics enables differentiation. It is primarily about driving change. Business analytics drives competitive advantage by generating economies of scale, economies of scope, and quality improvement. Taking advantage of the economies of scale is the first way organizations achieve comparative cost efficiencies and drive competitive advantage against their peers. Taking advantage of the economies of scope is the second-way organizations achieve comparative cost efficiencies and drive competitive advantage against their peers.



Business analytics improves the efficiency of business operations. The efficiencies that accumulate when a firm embraces big data technology eventually contribute to a ripple effect of increased production and reduced business costs. In the modern world, the vast quantities of data produced by corporations make their study and management practically impossible.

One can make the case that increasing the primary source of attaining a competitive advantage will be an organization's competence in mastering all flavours of analytics. If your management team is analytics-impaired, then your organization is at risk. Predictive business analytics is arguably the next wave for organizations to successfully compete. This will result not only from being able to predict outcomes but also to reach higher to optimize the use of their resources, assets and trading partners. It may be that the ultimate sustainable business strategy is to foster analytical competency and eventually mastery of analytics among an organization's workforce.

Analytics gives companies an insight into their customers' behaviour and needs. It also makes it possible for a company to understand the public opinion of its brand, to follow the results of various marketing campaigns, and strategize how to create a better marketing strategy to nurture long and fruitful relationships with its customers.

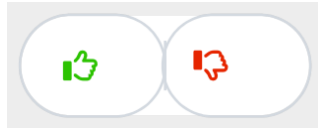
Business analytics helps organisations to know where they stand in the industry or a particular niche provides the company with the needed clarity to develop effective strategies to position itself better in the future.

For a company to remain competitive in the modern marketplace that requires constant change and growth, it must stay informed on the latest industry trends and best practices. Not only does business analytics provide the needed knowledge for companies to survive in today's constantly changing business environment, but it also makes room for growth and improvement, providing a detailed look into various opportunities and challenges that companies face on a day-to-day basis.

The retention of company employees has been a concern for business enterprises although it is taken more seriously in some niches than it is in other industries. A recent study that was conducted by IBM infers that a business enterprise had over 5,000 job applications reviewed but only hired 200 employees monthly. Big data has made it possible for companies to quickly analyse long time worker's histories to identify the job traits for long-term employment prospects.

As a result, corporations and small business enterprises are revamping their recruitment process which reduces employee turnover significantly. Companies can dedicate resources that are newly available to activities that are of more productive value to the business and increase their levels of service delivery. The

retention of an experienced pool of employees can significantly assist a business enterprise to outperform its competitors using their long-term experiences.



Unit - 3 Descriptive Analytics

Unit objectives:

- Explain why we need to visualize and explore data.
- Describe statistical charts and how to apply them.
- Describe descriptive statistics useful in the descriptive business analytics (BA) process.
- Describe sampling methods useful in BA and where to apply them.
- Describe what sampling estimation is and how it can aid in the BA process.



- Describe the use of confidence intervals and probability distributions.
- Explain how to undertake the descriptive analytics step in the BA process.

Introduction

In any BA undertaking, referred to as BA initiatives or projects, a set of objectives is articulated. These objectives are a means to align the BA activities to support strategic goals. The objectives might be to seek out and find new business opportunities, to solve operational problems the firm is experiencing, or to grow the organization. It is from the objectives that exploration via BA originates and is in part guided. The directives that come down, from the strategic planners in an organization to the BA department or analyst, focus the tactical effort of the BA initiative or project. Maybe the assignment will be one of exploring internal marketing data for a new marketing product. Maybe the BA assignment will be focused on enhancing service quality by collecting engineering and customer service information. Regardless of the type of BA assignment, the first step is one of exploring data and revealing new, unique, and relevant information to help the organization advance its goals. Doing this requires an exploration of data.

This chapter focuses on how to undertake the first step in the BA process: descriptive analytics. The focus in this chapter is to acquaint readers with more common descriptive analytic tools used in this step and available in SAS software. The treatment here is not computational but informational regarding the use and meanings of these analytic tools in support of BA. For purposes of illustration, we will use the data set in Figure 5.1 representing four different types of product sales (Sales 1, Sales 2, Sales 3, and Sales 4).



SAS - [VIEWTABLE: Work.Sales_data]

File Edit View Tools Data Solutions Window Help

Explorer

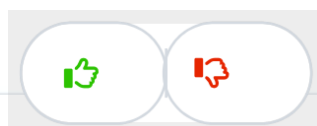
Contents of 'Work'

Sales_data

	obs	sales1	sales2	sales3	sales4
1	1	23	1234	1	1
2	2	31	943	2	5
3	3	48	986	3	9
4	4	16	12	4	12
5	5	28	15	5	18
6	6	29	15	6	19
7	7	31	23	6	19
8	8	35	21	6	21
9	9	51	25	6	21
10	10	42	27	7	21
11	11	34	27	8	21
12	12	56	29	9	21
13	13	24	20	10	21
14	14	34	18	11	19
15	15	43	13	12	19
16	16	56	8	13	18
17	17	34	7	14	12
18	18	38	6	15	9
19	19	23	4	16	5
20	20	27	1	17	1

Figure 5.1 Illustrative sales data sets

When using SAS, data sets are placed into files like that in Figure 5.2.



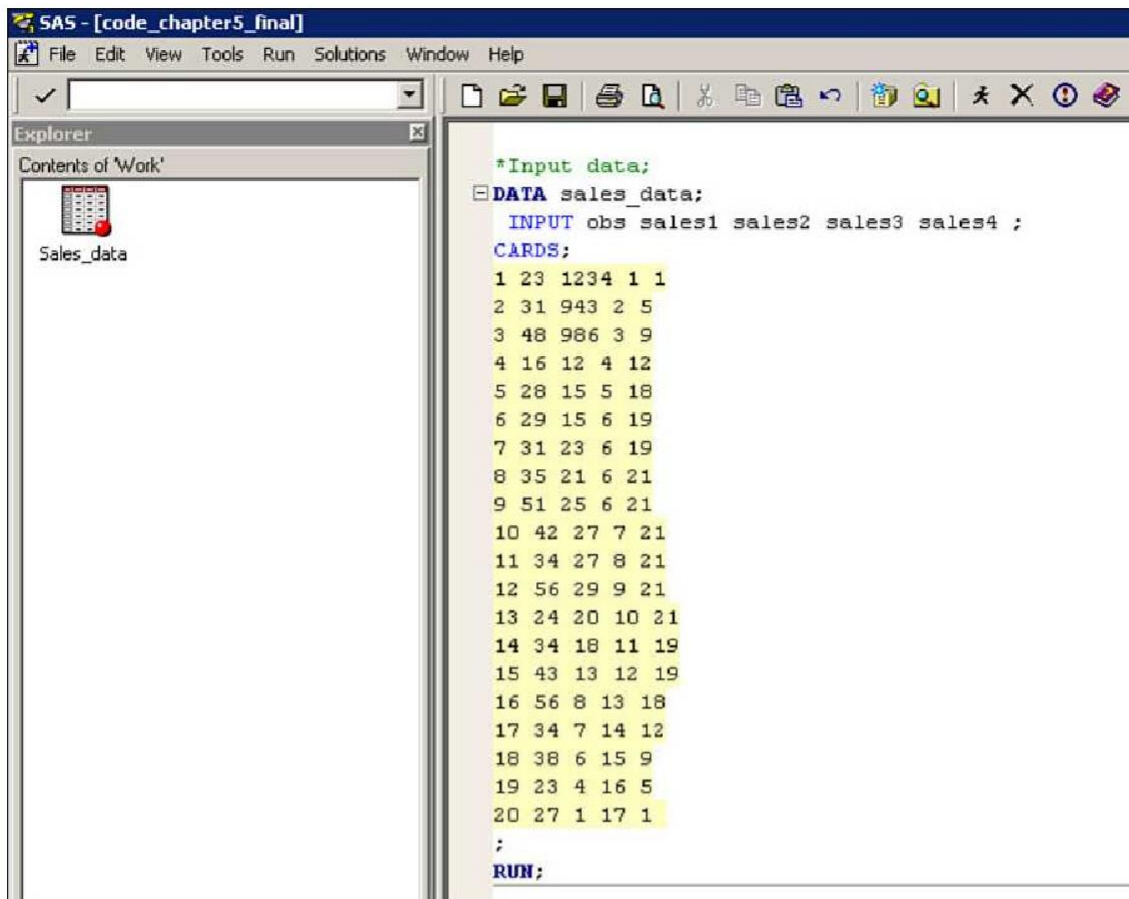
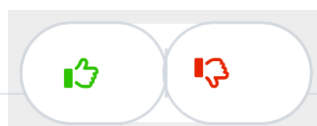


Figure 5.2 SAS coding of sales data set

Creating the data set in Figure 5.2 requires a sequence of SAS steps. Because this is the first use of SAS, the sequence of instructions is illustrated in Figures 5.3 through 5.9. These steps should be familiar to experienced SAS users. Depending on the SAS version and intent of the data file, the images in these figures may be slightly different.



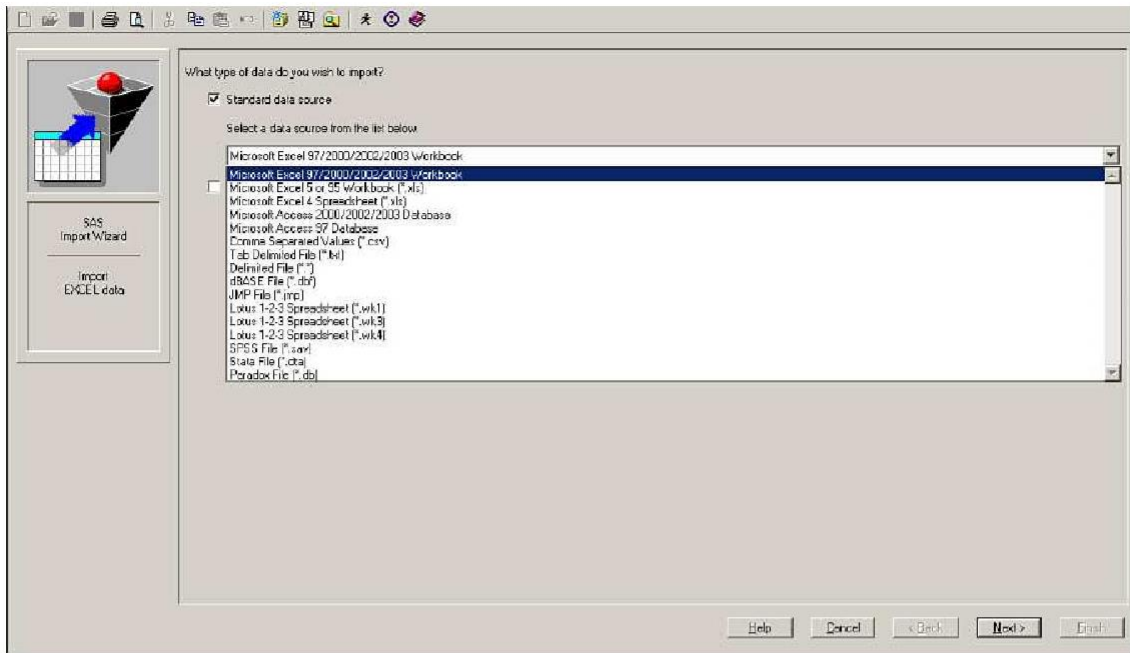
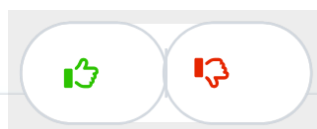


Figure 5.3 Creating a workbook file for the data set



The image shows a screenshot of the Microsoft Excel application. The ribbon at the top includes 'File', 'Home', 'Insert', and 'Page Layout'. The 'Home' ribbon is active, showing the 'Clipboard' group with 'Paste' and 'Clipboard' options, and the 'Font' group with 'Calibri' font, size '11', and options for bold, italic, and underline. The active cell is E22, containing a formula icon. The spreadsheet grid shows columns A through D and rows 1 through 21. The data is as follows:

	A	B	C	D
1	Sales 1	Sales 2	Sales 3	Sales 4
2	23	1234	1	1
3	31	943	2	5
4	48	896	3	9
5	16	12	4	12
6	28	15	5	18
7	29	15	6	19
8	31	23	6	19
9	35	21	6	21
10	51	25	6	21
11	42	27	7	21
12	34	27	8	21
13	56	29	9	21
14	24	20	10	21
15	34	18	11	19
16	43	13	12	19
17	56	8	13	18
18	34	7	14	12
19	38	6	15	9
20	23	4	16	5
21	27	1	17	1

Figure 5.4 Excel data file used in the creation of the SAS data file



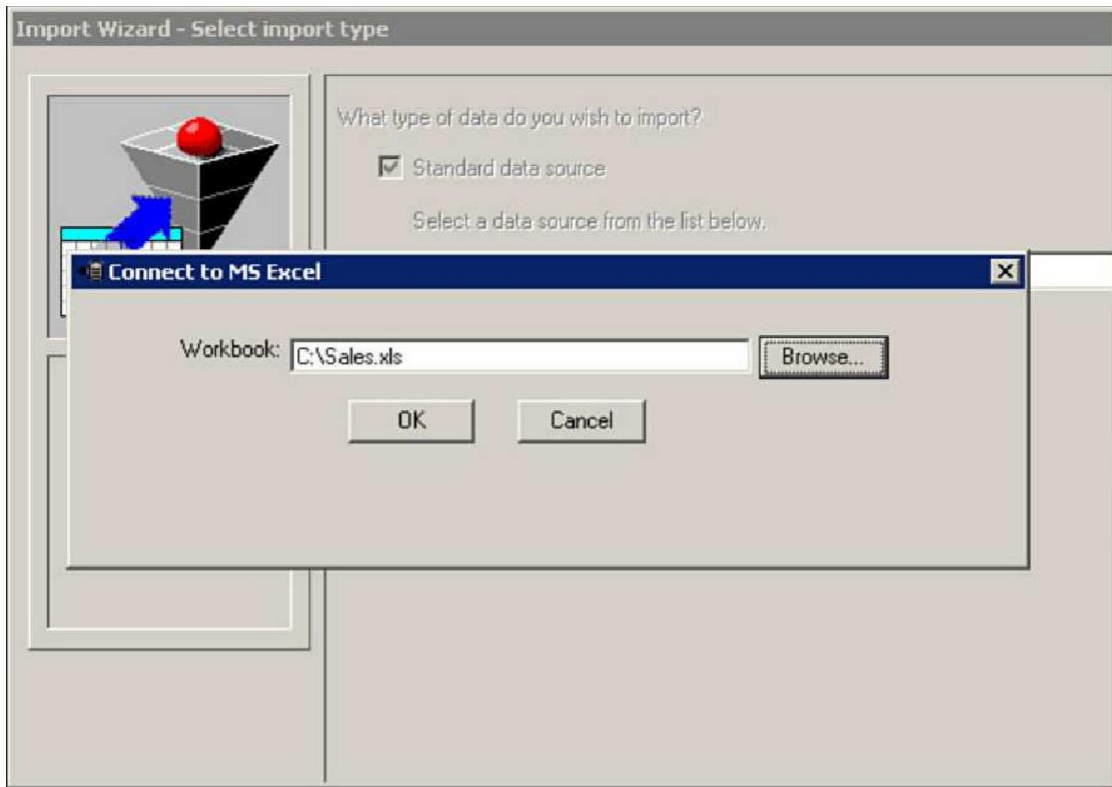


Figure 5.5 Step to pull the data set from an Excel (or any) file using SAS

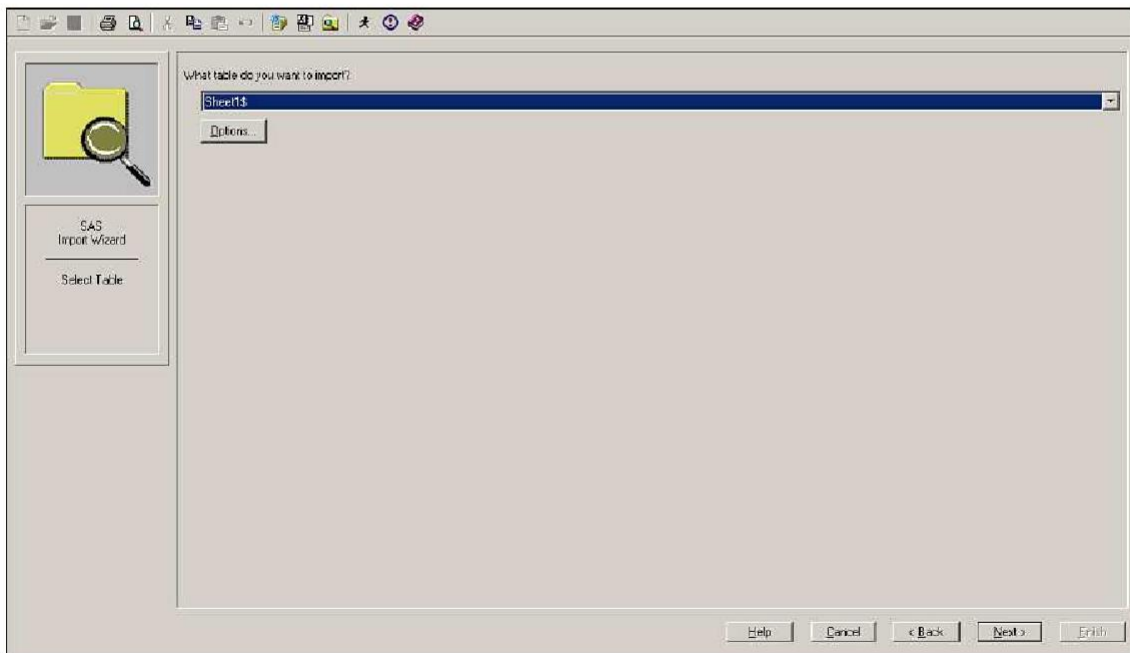
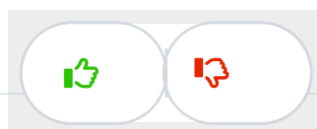


Figure 5.6 Identify file using SAS



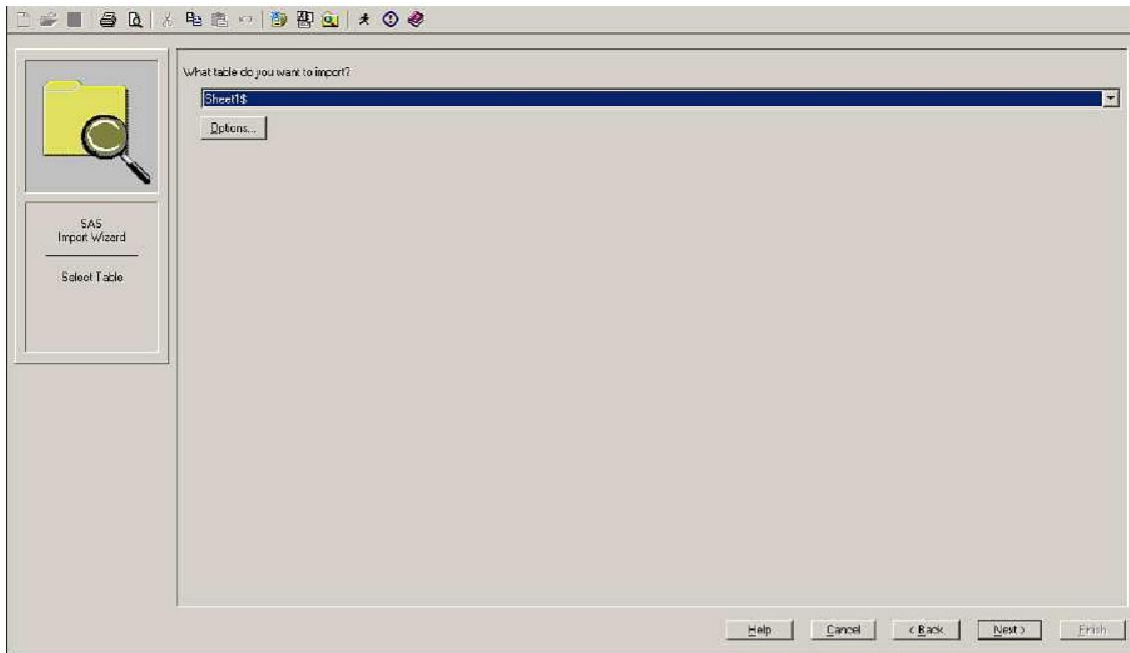


Figure 5.7 Label SAS file as SALES_DATA

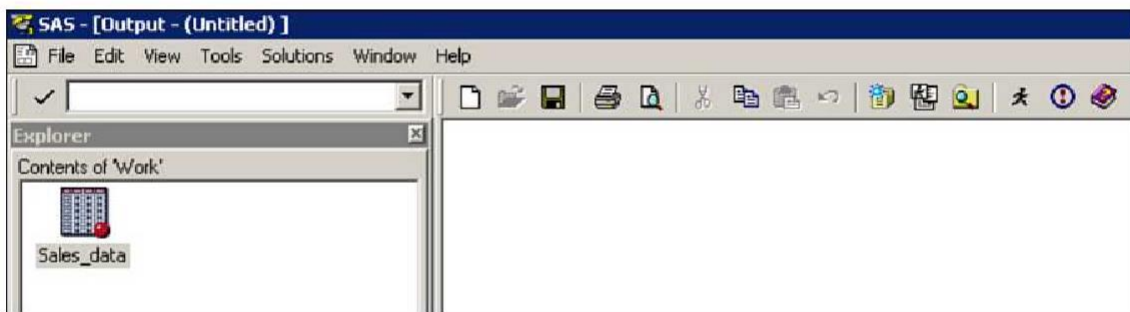
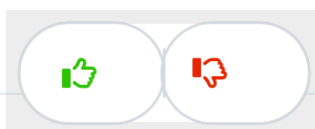


Figure 5.8 Shows the SALES_DATA SAS file is created



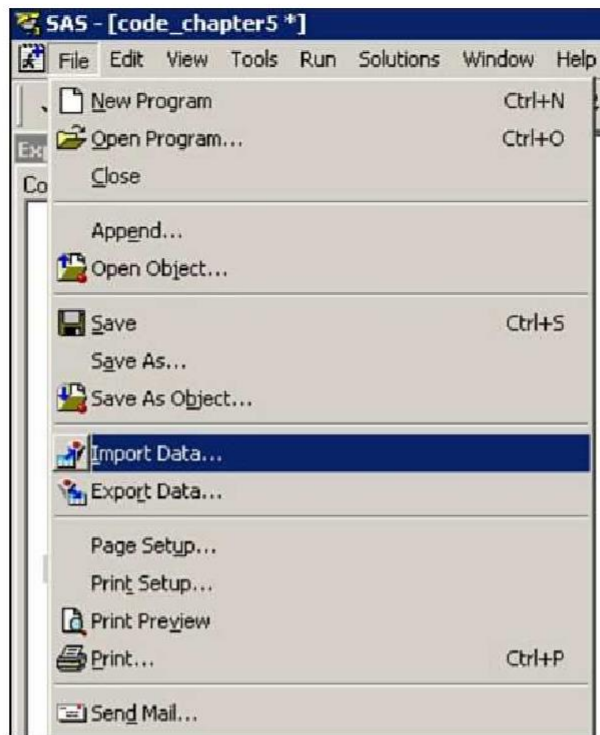


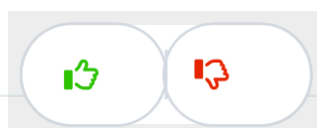
Figure 5.9 Step to import the created file

SAS permits the use of many different sources for data sets or data files to be entered into an SAS program. Big data files in Excel, SPSS, or other software applications can be brought into SAS programs using a similar set of steps presented in this section. Once the data sets are structured for use with SAS, there is still considerable SAS programming effort needed to glean useful information from any big data or small data files. Fortunately, SAS provides the means by which any sized data set can be explored and visualized by BA analysts. Because SAS is a programming language, it permits a higher degree of application customization than most other statistical software.

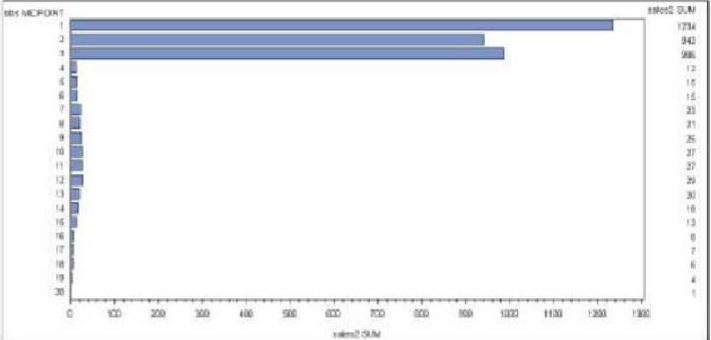
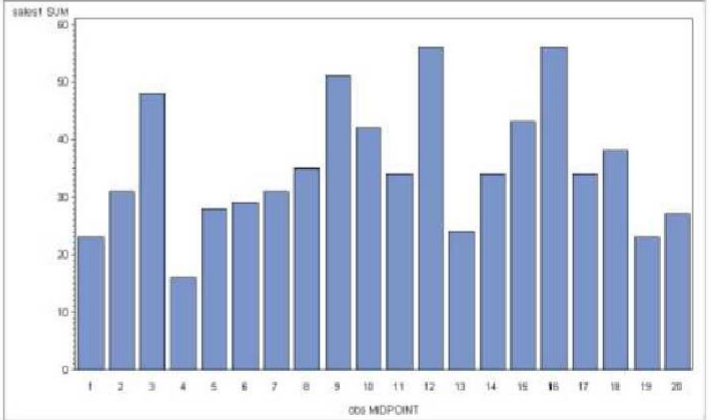
5.2 Visualizing and Exploring Data

There is no single best way to explore a data set, but some way of conceptualizing what the data set looks like is needed for this step of the BA process. Charting is often employed to visualize what the data might reveal.

When determining the software options to generate charts in SAS, consider that the software can draft a variety of charts for the selected variables in the data sets. Using the data in Figure 5.1, charts can be created for the illustrative sales data sets. Some of these charts are discussed in Table 5.1 as a set of exploratory tools that are helpful in understanding the



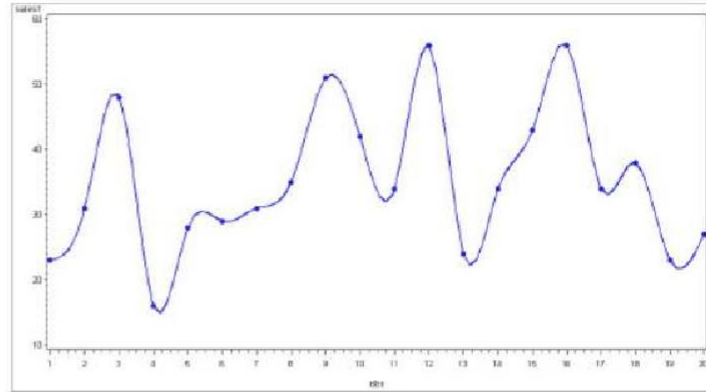
informational value of data sets. The chart to select depends on the objectives set for the chart. The SAS program statements used to create each chart in Table 5.1 are provided.

Type of Chart	Application Notes	Chart Example
Bar	<ul style="list-style-type: none"> • Can be horizontal, vertical, cone, or cyclically shaped and multi-dimensional with overlaying variables. • Ideal for showing comparative improvement over time. • Example: Bars showing productivity of one person versus another. 	<pre data-bbox="660 394 1027 512">proc gchart data=sales_data; hbar obs /sumvar=sales2 midpoints=1 to 20 by 1; run; quit;</pre> 
Column	<ul style="list-style-type: none"> • Same as a bar chart. 	<pre data-bbox="660 1059 1027 1178">proc gchart data=sales_data; vbar obs /sumvar=sales1 midpoints=1 to 20 by 1; run; quit;</pre> 
Line	<ul style="list-style-type: none"> • Ideal for showing linear trend and other linear or nonlinear 	<pre data-bbox="660 1731 1177 1850">symbol value=dot interpol=sms line=1 width=2; proc gplot data=sales_data; plot sales1*obs; run; quit;</pre>



appearance.

- Best applied with time series data with time as the X-axis.



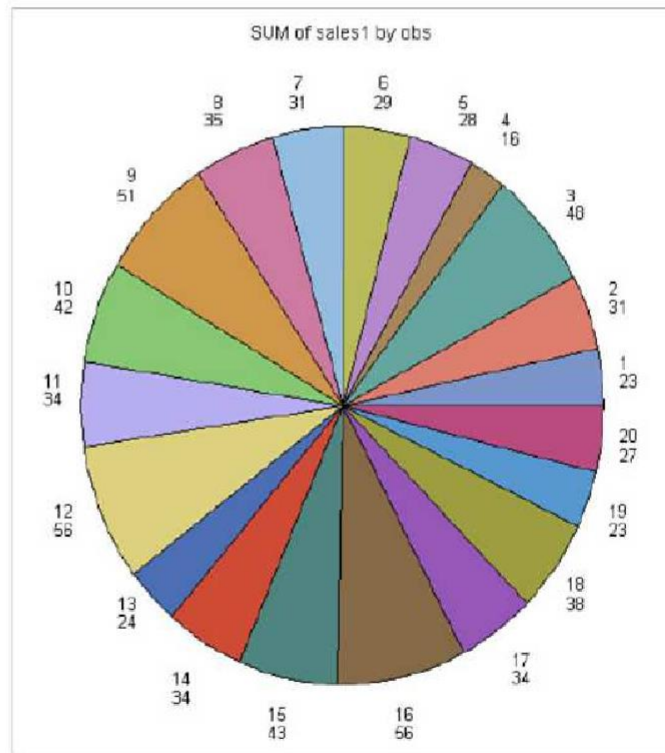
Pie

- Useful in conceptualizing proportions.
- Various other versions, like

the donut chart (with a hollow center), can also be used.

- Useful when the number of variables is limited (not like the illustration to the right).

```
proc gchart data=sales_data;  
  pie obs /sumvar=sales1  
  midpoints=1 to 20 by 1 other=0;  
run;  
quit;
```



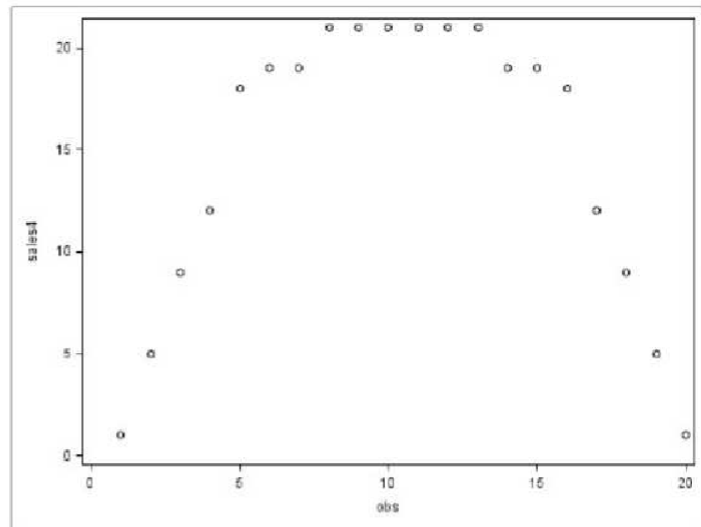
Scatter

- Useful when patterns are observed in the data sets.

```
proc splot data=sales_data;  
  scatter x=obs y=sales4;  
run;
```



- Useful when outliers are observed in the data that may need to be cleaned out.
- Outline trends that a linear chart can augment.



Histogram

- Ideal to help reveal frequency distributions in variable data sets.
- Reduces the size of data by grouping data points into frequencies.

```
proc gchart data=sales_data;
vbar sales3 / midpoints=1 to 20 by 2;
run;
```

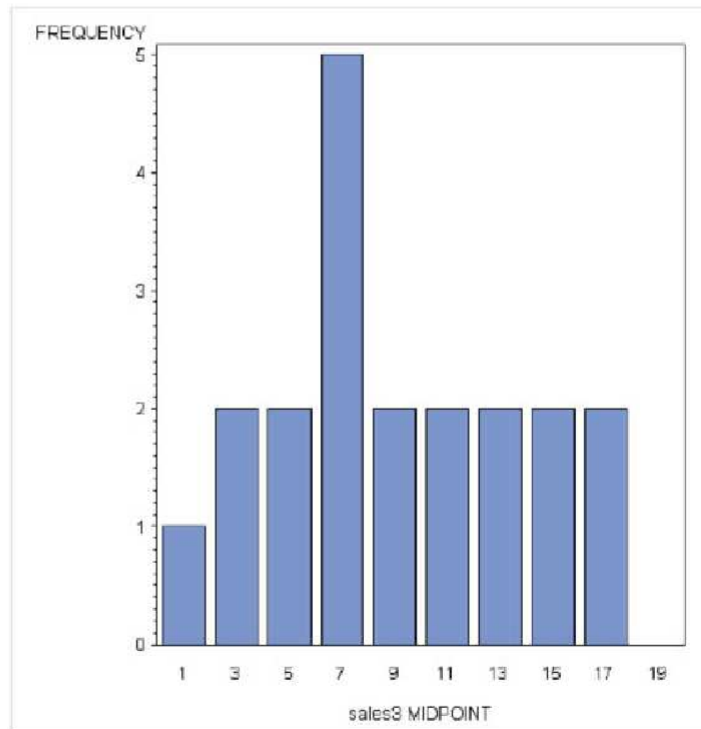


Table 5.1 Statistical Charts Useful in BA



The charts presented in Table 5.1 reveal interesting facts. The column chart is useful in revealing the almost perfect linear trend in the Sales 3 data, whereas the scatter chart reveals an almost perfect nonlinear function in Sales 4 data. Additionally, the cluttered pie chart with 20 different percentages illustrates that all charts can or should be used in some situations. The best practices suggest charting should be viewed as an exploratory activity of BA. BA analysts should run a variety of charts and see which ones reveal interesting and useful information. Those charts can be further refined to drill down to more detailed information and more appropriate charts related to the objectives of the BA initiative.

Of course, a cursory review of the Sales 4 data in Figure 5.1 makes the concave appearance of the data in the scatter chart in Table 5.1 unnecessary. But most BA problems involve big data—so large as to make it impossible to just view it and make judgment calls on structure or appearance. This is why descriptive statistics can be employed to view the data in a parameter-based way in the hopes of better understanding the information that the data has to reveal.

5.3 Descriptive Statistics

SAS has a number of useful statistics that can be automatically computed for the variables in the data sets. The SAS printout of the sales data from Figure 5.1 is summarized in Table 5.2. Some of these descriptive statistics are discussed in Table 5.3 as exploratory tools that are helpful in understanding the informational value of data sets.

	N	Range	Min	Max	Sum	Mean	Std. Dev.	Variance	Skewedness	Kurtosis			
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Sales 1	20	40	16	56	703	35.15	2.504	11.198	125.397	.490	.512	-.429	.992
Sales 2	20	1233	1	1234	3344	167.20	83.686	374.254	140065.853	2.241	.512	3.636	.992
Sales 3	20	16	1	17	171	8.55	1.065	4.763	22.682	.272	.512	-.988	.992
Sales 4	20	20	1	21	292	14.60	1.603	7.170	51.411	-.824	.512	-.825	.992
Valid N	20												

Table 5.2 SAS Descriptive Statistics



Statistics	Computation (in Data Set)	Application Area	Example	Application Notes
N or Count	Number of values.	Any.	Sample size of a company's transactions during a month.	Useful in knowing how many items were used in the statistics computations.
Sum	Total of the values in the entire data set.	Any.	Total sales for a company.	Useful in knowing the total value.
Mean	Average of all values.	Any.	Average sales per month.	Useful in capturing the central tendency of the data set.
Median	Midpoint value in the data set arranged from high to low.	Finding the midpoint in the distribution of data.	Total income for citizens of a country.	Useful in finding the point where 50 percent of the data is above and below.
Mode	Most common value in the data set.	Where values are highly repeated in the data set.	Fixed annual salaries where a limited number of wage levels is used.	Useful in declaring a common value in highly repetitive data sets.
Maximum/ Minimum	Largest and smallest values, respectively.	To conceptualize the spread of the data's distribution.	Largest and smallest sales in a day.	Useful in providing a scope or end points in the data.



Range	Difference between the max and min values.	A crude estimate of the spread of the data's distribution.	Spread of sales in units during a month.	Useful as a simple estimate of dispersion.
Standard deviation	Square root of the average of the differences squared between the mean and all other values in the data set.	A precise estimate of the spread of the data's distribution from a mean value in terms of the units used in its computation.	Standard deviation in dollars from mean sales.	The smaller the value, the less the variation and the more predictable using the data set.
Variance	Average differences squared between the mean and all other values.	A variance estimate of the spread of the data's distribution from a mean value, not in terms of the units used in its computation.	Measure of variance that is best used when compared with another variance computed on the same data set.	The smaller the value, the less the variation and the more predictable the data set.
(Coefficient of) Skewedness	Positive or negative values. If value sign is +, distribution is positively skewed; if -, it is negatively skewed. The larger the value, the greater it is skewed.	Measure of the degree of asymmetry of data about a mean.	As the age of residents in a country becomes older, the population age distribution becomes more negatively skewed.	The closer the value is to 0, the better the symmetry. A positively skewed distribution has its largest allocation to the left, and a negative distribution to the right.



1



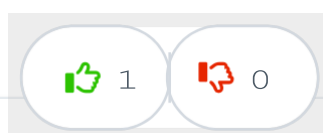
0

(Coefficient of) Kurtosis	Value where less than 3 means a flat distribution and more than 3 means a peaked distribution.	Measure of the degree of spread vertically in a distribution about a mean. Also, it reveals a positive and a negative symmetry depending on its sign.	Distribution of customers at lunch and dinner times peaks and then flattens out.	The closer the value is to 2, the less is the kurtosis (peaking or flattening in the distribution).
Standard Error (of the Mean)	Mean of the sample standard deviation (that is, a standard deviation adjusted to reflect a sample size).	Standard deviation of a sampling distribution.	Standard deviation in dollars from mean sales based on a sample.	The smaller the value, the less the variation and the more predictable the sample data set.
Sample Variance	Same as variance but adjusted for sample sizes.	Variance estimate of the spread of the sampling data distribution.	Measure of variance when sampling is used for collection purposes.	The smaller the value, the less the variation and the more predictable the sample data set.

Table 5.3 Descriptive Statistics Useful in BA

Fortunately, we do not need to compute these statistics to know how to use them. Computer software provides these descriptive statistics when they're needed or requested. When you look at the data sets for the four variables in Figure 5.1 and at the statistics in Table 5.2, there are some obvious conclusions based on the detailed statistics from the data sets. It should be no surprise that Sales 2, with a few of the largest values and mostly smaller ones making up the data set, would have the largest variance statistics (standard deviation, sample variance, range, maximum/minimum). Also, Sales 2 is highly, positively skewed (Skewedness > 1) and highly peaked (Kurtosis > 3). Note the similarity of the mean, median, and mode in Sales 1 and the dissimilarity in Sales 2. These descriptive statistics provide a more precise basis to envision the behavior of the data. Referred to as measures of central tendency, the mean, median, and mode can also be used to clearly define the direction of a skewed distribution. A negatively skewed distribution orders these measures such that mean < median < mode, and a positive skewed distribution orders them such that mode < median < mean.

So what can be learned from these statistics? There are many observations that can be drawn from this data. Keep in mind that, in dealing with the big data sets, one would only have the charts and the statistics to serve as a guide in determining what the data looks like. Yet, from these statistics, one can begin describing the data set. So in the case of Sales 2, it can be predicted that the data set is positively skewed and peaked. Note in Figure 5.10 that the



histogram of Sales 2 is presented. The SAS chart also overlays a normal distribution (a bell-shaped curve) to reflect the positioning of the mean (highest point on the curve, 167.2) and the way the data appears to fit the normal distribution (not very well in this situation). As expected, the distribution is positively distributed with a substantial variance between the large values in the data set and the many more smaller valued data points.

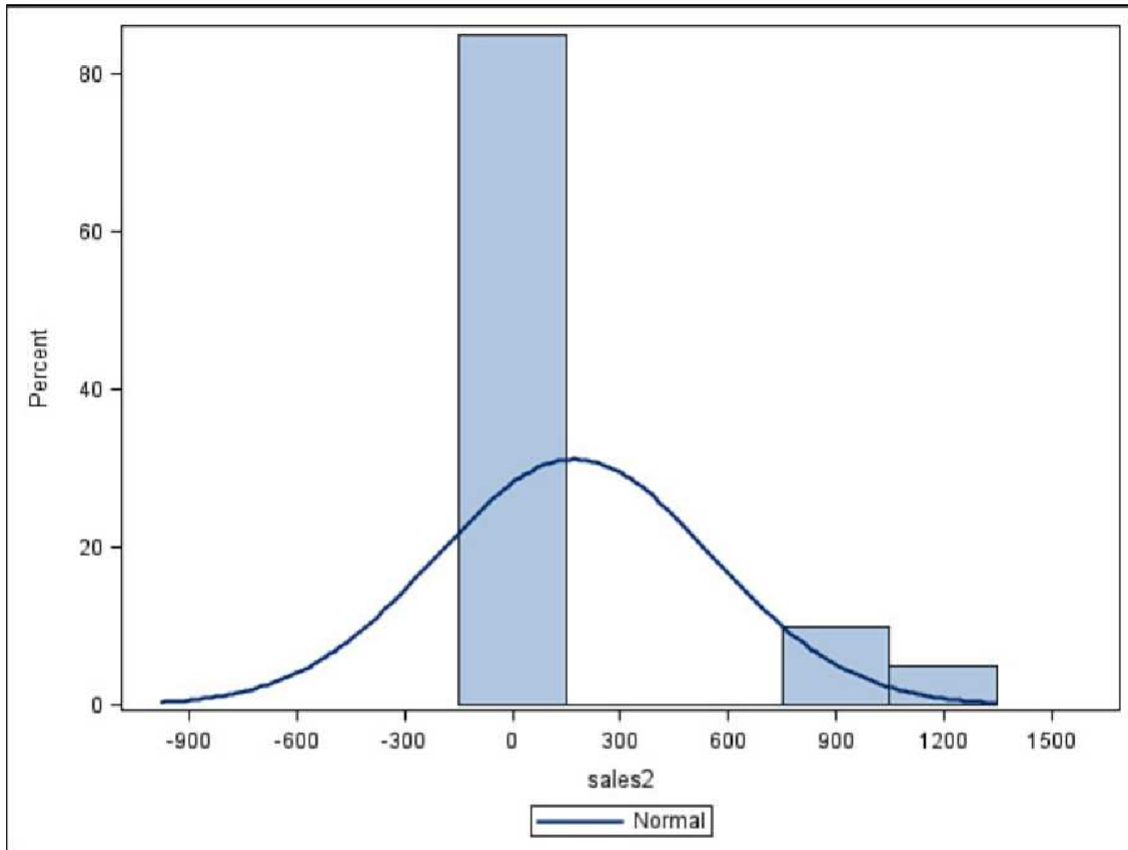
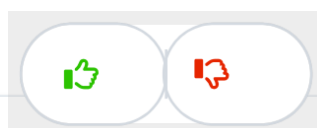


Figure 5.10 SAS histogram of Sales 2 data

We also know that substantial variance in the data points making up the data set is highly diverse—so much so that it would be difficult to use this variable to predict future behavior or trends. This type of information may be useful in the further steps of the BA process as a means of weeding out data that will not help predict anything useful. Therefore, it would not help an organization improve its operations.

Sometimes big data files become so large that certain statistical software systems cannot manipulate them. In these instances, a smaller but representative sample of the data can be obtained if necessary. Obtaining the sample for accurate prediction of business behavior requires understanding the sampling process and estimation from that process.



5.4 Sampling and Estimation

The estimation of most business analytics requires sample data. In this section, we discuss various types of sampling methods and follow that with a discussion of how the samples are used in sampling estimation.

5.4.1 Sampling Methods

Sampling is an important strategy of handling large data. Big data can be cumbersome to work with, but a smaller sample of items from the big data file can provide a new data file that seeks to accurately represent the population from which it comes. In sampling data, there are three components to recognize: a population, a sample, and a sample element (the items that make up the sample). A firm's collection of customer service performance documents for one year could be designated as a population of customer service performance for that year. From that population, a sample of a lesser number of sample elements (the individual customer service documents) can be drawn to reduce the effort of working with the larger data. Several sampling methods can be used to arrive at a representative sample. Some of these sampling methods are presented in Table 5.4.

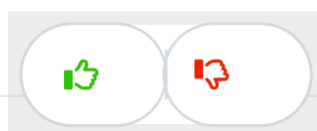


Sampling Method	Description	Application	Application Notes
Simple Random	Allows each sample element in a population to have an equal chance of selection.	Selecting customers based on their percentage of occurrence as a member of a particular race.	Sample size must be sufficient to avoid sampling bias.
Systematic Random (or Period)	Selects sample elements from a population based on a fixed number in an interval.	Selecting every fifth person leaving an airport to interview.	Assumes the sample elements order in the interval is presented in a random fashion; otherwise, it can result in sampling bias.
Stratified Random	Stage 1: Divide a population into groups (called strata); Stage 2: Apply simple random sampling.	Randomly selecting an equal number of people in each of three different economic strata.	Strata must be representative of the population, or it can result in sampling bias.
Cluster Random	Stage 1: Group sample elements geographically (called clusters); Stage 2: Apply simple random sampling.	Randomly selecting an equal number of people from voting districts.	Cluster must be representative of the population, or it can result in sampling bias.
Quota	Based on a fixed quota or number of sample elements.	Selecting the first 200 people who enter a store.	<ul style="list-style-type: none"> • Mainly used to save time and money. • Sample size must be sufficient to avoid sampling bias.
Judgment	Selects sample elements based on expert judgment.	Selecting candidates for an interview with a special offer based on their appearance.	Prone to bias without defined criteria for selection because of dependency on interviewer experience.

Table 5.4 Sampling Methods

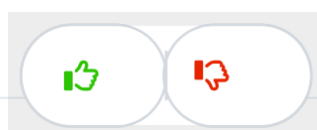
The simple, systematic, stratified, and cluster random methods are based on some kind of probability of their occurrence in a population. The quota and judgment methods are nonprobability-based tools. Although the randomization process in some methods helps ensure representative samples being drawn from the population, sometimes because of cost or time constraints, nonprobability methods are the best choice for sampling.

Which sampling method should be selected for a particular BA analysis? It depends on the nature of the sample. As mentioned in the application notes in Table 5.4, the size of the



population, the size of the sample, the area of application (geography, strata, ordering of the data, and so on), and even the researchers running the data collection effort impact the particular methodology selected. A best practices approach might begin with a determination of any constraints (time allowed and costs) that might limit the selection of a sample collection effort. That may narrow the choice to something like a quota method. Another best practices recommendation is to start with the objective(s) of the BA project and use them as a guide in the selection of the sampling method. For example, suppose the objective of a BA analysis is to increase sales of a particular product. This might lead to random sampling of customers or even a stratified sample by income levels, if income is important to the results of the analysis. Fortunately, there is software to make the data collection process easier and less expensive.

SAS software can be used with the methods mentioned earlier to aid in sampling analysis. For example, SAS permits simple, systematic, stratified, and cluster random methods, among others. Using this software requires a designation of the number of sample elements in each stratum. (For example, we selected 2 for each stratum in this example.) In Figure 5.11, SAS has defined seven strata for the Sales 4 data. The logic of this stratification can be observed by looking at the Sales 4 data in Figure 5.1, where only seven different types of values exist (1, 5, 9, 12, 18, 19, and 20). The additional SAS printout in Figure 5.11 shows the specific sample elements that were randomly selected in each stratum, as well as totals and their percentages in the resulting sample. For example, only 0.33, or 33 percent, of the “21” strata sample elements were randomly selected by the SAS program.



```

proc sort data= sales_data; by sales4;

proc surveyselect data =sales_data out = samp1 method = srs n=2;
strata sales4;
run;

```

The SAS System 17:53 Friday,

The SURVEYSELECT Procedure

Selection Method Simple Random Sampling
Strata Variable sales4

Input Data Set SALES_DATA
Random Number Seed 522937001
Stratum Sample Size 2
Number of Strata 7
Total Sample Size 14
Output Data Set SAMPI

	sales4	Probability of Selection	Sampling Weight
1	1	1	1
2	1	1	1
3	5	1	1
4	5	1	1
5	9	1	1
6	9	1	1
7	12	1	1
8	12	1	1
9	18	1	1
10	18	1	1
11	19	0.5	2
12	19	0.5	2
13	21	0.3333333333	3
14	21	0.3333333333	3

Figure 5.11 SAS program statements for stratification/random sampling for Sales 4 variable

5.4.2 Sampling Estimation

Invariably, using any sampling method can cause errors in the sample results. Most of the statistical methods listed in Table 5.2 are formulated for population statistics. Once sampling is introduced into any statistical analysis, the data must be treated as a sample and not as a population. Many statistical techniques, such as standard error of mean and sample variance,



incorporate mathematical correction factors to adjust descriptive analysis statistical tools to compensate for the possibility of sampling error.

One of the methods of compensating for error is to show some degree of confidence in any sampling statistic. The confidence in the sample statistics used can be expressed in a confidence interval, which is an interval estimate about the sample statistics. In general, we can express this interval estimate as follows:

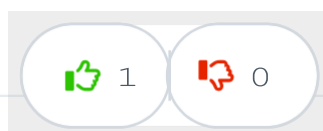
Confidence interval = (sample statistic) \pm [(confidence coefficient) \times (standard error of the estimate)]

The sample statistic in the confidence interval can be any measure or proportion from a sample that is to be used to estimate a population parameter, such as a measure of central tendency like a mean. The confidence coefficient is set as a percentage to define the degree of confidence to accurately identify the correct sample statistic. The larger the confidence coefficient, the more likely the population mean from the sample will fall within the confidence interval. Many software systems set a 95 percent confidence level as the default confidence coefficient, although any percentage can be used. SAS permits the user to enter a desired percentage. The standard error of the estimate in the preceding expression can be any statistical estimate, including proportions used to estimate a population parameter. For example, using a mean as the sample statistic, we have the following interval estimate expression:

Confidence interval = mean \pm [(95 percent) \times (standard error of the mean)]

The output of this expression consists of two values that form high and low values defining the confidence interval. The interpretation of this interval is that the true population mean represented by the sample has a 95 percent chance of falling in the interval. In this way, there is still a 5 percent chance that the true population mean will not fall in the interval due to sampling error. Because the standard error of the mean is based on variation statistics (standard deviation), the larger the variance statistics used in this expression, the wider the confidence interval and the less precise the sample mean value, which results in a good estimate for the true population mean.

SAS computes confidence intervals when analyzing various statistical measures and tests. For example, the SAS summary printout in Table 5.5 is of the 95 percent confidence interval for



the Sales 1 variable. With a sample mean value of 35.15, the confidence interval suggests there is a 95 percent chance that the true population mean falls between 29.91 and 40.39. When trying to ascertain if the sample is of any value, this kind of information can be of great significance. For example, knowing with 95 percent certainty there is at least a mean of 29.91 might make the difference between continuing to sell a product or not because of a needed requirement for a breakeven point in sales.

<i>One-Sample Statistics</i>						
	N	Mean	Std. Deviation	Std. Error Mean	95 Percent Confidence Interval of the Difference	
Sales 1	20	35.15	11.198	2.504	Lower 29.91	Upper 40.39

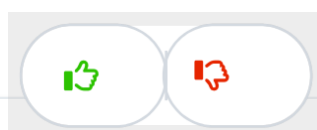
Table 5.5 SAS 95 Percent Confidence Interval Summary for Sales 1 Variable

Confidence intervals are also important for demonstrating the accuracy of some forecasting models. For example, confidence intervals can be created about a regression equation model forecast to see how far off the estimates might be if the model is used to predict future sales. For additional discussion on confidence intervals, see Appendix A, “Statistical Tools.”

5.5 Introduction to Probability Distributions

By taking samples, one seeks to reveal population information. Once a sample is taken on which to base a forecast or a decision, it may not accurately capture the population information. No single sample can assure an analyst that the true population information has been captured. Confidence interval statistics are employed to reflect the possibility of error from the true population information.

To utilize the confidence interval formula expressed in Section 5.4, you set a confidence coefficient percentage (95 percent) as a way to express the possibility that the sample statistics used to represent the population statistics may have a potential for error. The confidence coefficient used in the confidence interval is usually referred to as a Z value. It is spatially related to the area (expressed as a percentage or frequency) representing the probability under the curve of a distribution. The sample standard normal distribution is the bell-shaped curve illustrated in Figure 5.12. This distribution shows the relationship of the Z value to the area under the curve. The Z value is the number of standard error of the means.



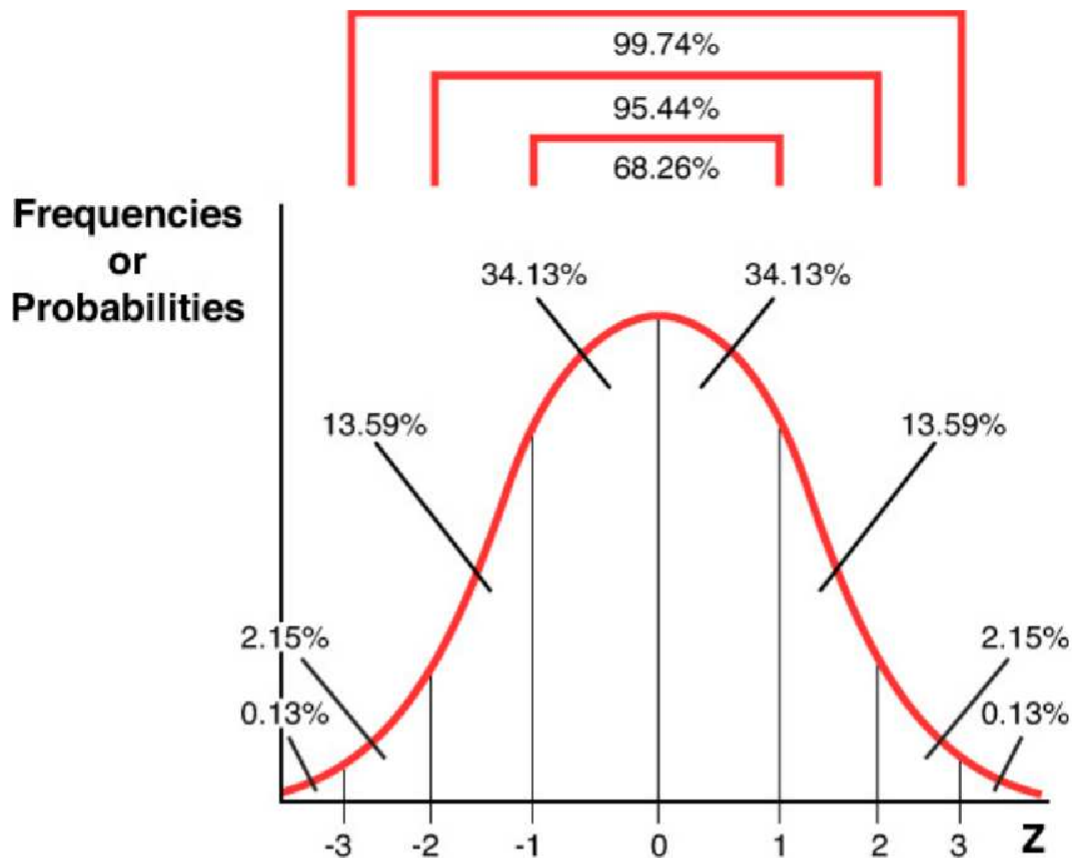
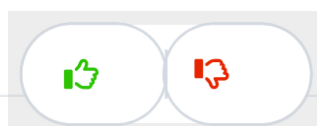


Figure 5.12 Standard normal probability distribution

The confidence coefficient is related to the Z values, which divide the area under a normal curve into probabilities. Based on the central limit theorem, we assume that all sampling distributions of sufficient size are normally distributed with a standard deviation equal to the standard error of the estimate. This means that an interval of plus or minus two standard errors of the estimate (whatever the estimate is) has a 95.44 percent chance of containing the true or actual population parameter. Plus or minus three standard errors of the estimate has a 99.74 percent chance of containing the true or actual population parameter. So the Z value represents the number of standard errors of the estimate. Table 5.6 has selected Z values for specific confidence levels representing the probability that the true population parameter is within the confidence interval and represents the percentage of area under the curve in that interval.

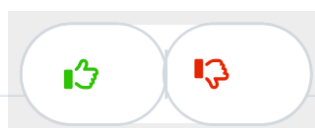


Confidence Level	Related Z Value
0.60	0.253
0.70	0.524
0.80	0.842
0.90	1.282
0.95	1.645
0.99	2.327
0.999	3.080

Table 5.6 Selected Z Values and Confidence Levels

The important BA use of the probability distributions and confidence intervals is that they suggest an assumed parameter based on a sample that has properties that allow analysts to predict or forecast with some assessed degree of statistical accuracy. In other words, BA analysts can, with some designated confidence level, use samples from large databases to accurately predict population parameters.

Another important value to probability distributions is that they can be used to compute probabilities that certain outcomes like success with business performance may occur. In the exploratory descriptive analytics step of the BA process, assessing the probabilities of some events occurring can be a useful strategy to guide subsequent steps in an analysis. Indeed, probability information may be useful in weighing the choices an analyst faces in any of the steps of the BA process. Suppose, for example, the statistics from the Sales 1 variable in Table 5.5 are treated as a sample to discover the probability of sales greater than one standard error of the mean above the current mean of 35.15. In Figure 5.13, the mean (35.15) and standard error of the mean (2.504) statistics are included at the bottom of the standard sampling normal distribution. When one standard error of the mean is added to the sample mean, the resulting value is 37.654. The sum of the area (the shaded region in Figure 5.13) representing the total probability beyond 37.654 is a probability of 15.87 (13.59+2.15+0.13). So there is only a 15.87 percent probability that sales will exceed 37.654 based on the sample information for the Sales 1 variable.



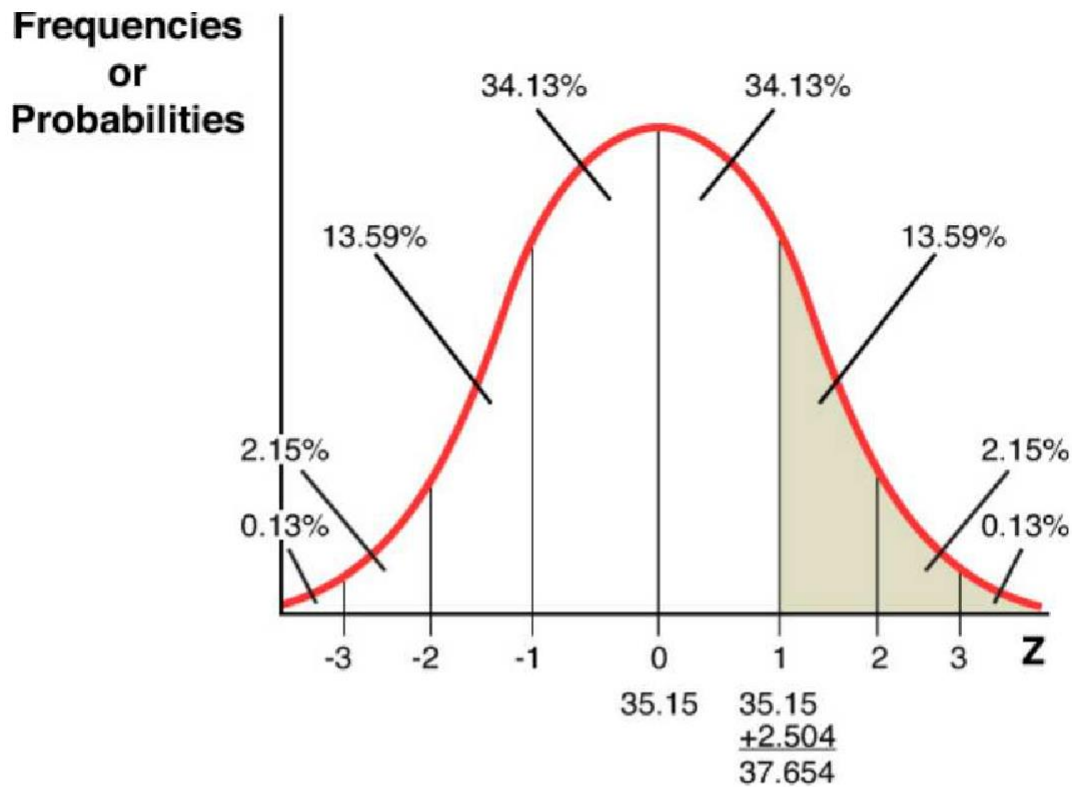


Figure 5.13 Probability function example

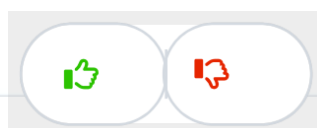
The ability to assess probabilities using this approach is applicable to other types of probability distributions. For a review of probability concepts and distributions, probability terminology, and probability applications, see Appendix A.

5.6 Marketing/Planning Case Study Example: Descriptive Analytics Step in the BA Process

In the last section of this chapter and in Chapters 6, “What Is Predictive Analytics?” and 7, “What Is Prescriptive Analytics?” an ongoing marketing/planning case study of the relevant BA step discussed in those chapters will be presented to illustrate some of the tools and strategies used in a BA problem analysis. This is the first installment of the case study dealing with the descriptive analytics step in BA. The predictive analytics step (in Chapter 6) and prescriptive analytics step (in Chapter 7) will continue with this ongoing case study.

5.6.1 Case Study Background

A firm has collected a random sample of monthly sales information on a service product offered infrequently and only for a month at a time. The sale of this service product occurs only during the month that the promotion efforts are allocated. Basically, promotion funds are



allocated at the beginning or during the month, and whatever sales occur are recorded for that promotion effort. There is no spillover of promotion to another month, because monthly offerings of the service product are independent and happen randomly during any particular year. The nature of the product does not appear to be impacted by seasonal or cyclical variations, which prevents forecasting and makes planning the budget difficult.

The firm promotes this service product by using radio commercials, newspaper ads, television commercials, and point-of-sale (POS) ad cards. The firm has collected the sales information as well as promotion expenses. Because the promotion expenses are put into place before the sales take place and on the assumption that the promotion efforts impact products, the four promotion expenses can be viewed as predictive data sets (or what will be the predictive variables in a forecasting model). Actually, in terms of modeling this problem, product sales is going to be considered the dependent variable, and the other four data sets represent independent or predictive variables.

These five data sets, in thousands of dollars, are present in the SAS printout shown in Figure 5.14. What the firm would like to know is, given a fixed budget of \$350,000 for promoting this service product, when offered again, how best should budget dollars be allocated in the hope of maximizing future estimated months' product sales? This is a typical question asked of any product manager and marketing manager's promotion efforts. Before the firm allocates the budget, there is a need to understand how to estimate future product sales. This requires understanding the behavior of product sales relative to sales promotion. To begin to learn about the behavior of product sales to promotion efforts, we begin with the first step in the BA process: descriptive analytics.



	case_number	sales	radio	paper	tv	pos
1	1	11125	65	89	250	1.3
2	2	16121	73	55	260	1.6
3	3	16440	74	58	270	1.7
4	4	16876	75	82	270	1.3
5	5	13965	69	75	255	1.5
6	6	14999	70	71	255	2.1
7	7	20167	87	59	280	1.2
8	8	20450	89	65	280	3
9	9	15789	72	62	260	1.6
10	10	15991	73	56	260	1.6
11	11	15234	70	66	255	1.5
12	12	17522	78	50	270	0
13	13	17933	79	47	275	0.2
14	14	18390	81	78	275	0.9
15	15	18723	81	41	275	1
16	16	19328	84	63	280	2.6
17	17	19399	84	77	280	1.2
18	18	19641	85	35	280	2.5
19	19	12369	65	37	250	2.5
20	20	13882	68	80	252	1.4

Figure 5.14 Data for marketing/planning case study

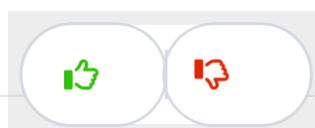
5.6.2 Descriptive Analytics Analysis

To begin conceptualizing possible relationships in the data, one might compute some descriptive statistics and graph charts of data (which will end up being some of the variables in the planned model). SAS can be used to compute these statistics and charts. The SAS printout in Table 5.7 provides a typical set of basic descriptive statistics (means, ranges, standard deviations, and so on) and several charts.

The MEANS Procedure							
Variable	N	Range	Minimum	Maximum	Mean	Std Dev	Variance
radio	20	24.0000000	65.0000000	89.0000000	76.1000000	7.3549124	54.0947368
paper	20	54.0000000	35.0000000	89.0000000	62.3000000	15.3592078	235.9052632
tv	20	30.0000000	250.0000000	280.0000000	266.6000000	11.3388016	128.5684211
pos	20	3.0000000	0	3.0000000	1.5350000	0.7499298	0.5623947
sales	20	9325.00	11125.00	20450.00	16717.20	2617.05	6848960.59

Table 5.7 SAS Descriptive Statistics for the Marketing/Planning Case Study

Remember, this is the beginning of an exploration that seeks to describe the data and get a handle on what it may reveal. This effort may take some exploration to figure out the best way to express data from a file or database, particularly as the size of the data file increases.



In this simple example, the data sets are small but can still reveal valuable information if explored well.

In Figure 5.15, five typical SAS charts are presented. Respectively, these charts include a histogram chart (sales), a block chart (radio), a line chart (TV), a pie chart (paper), and a 3D chart (POS).

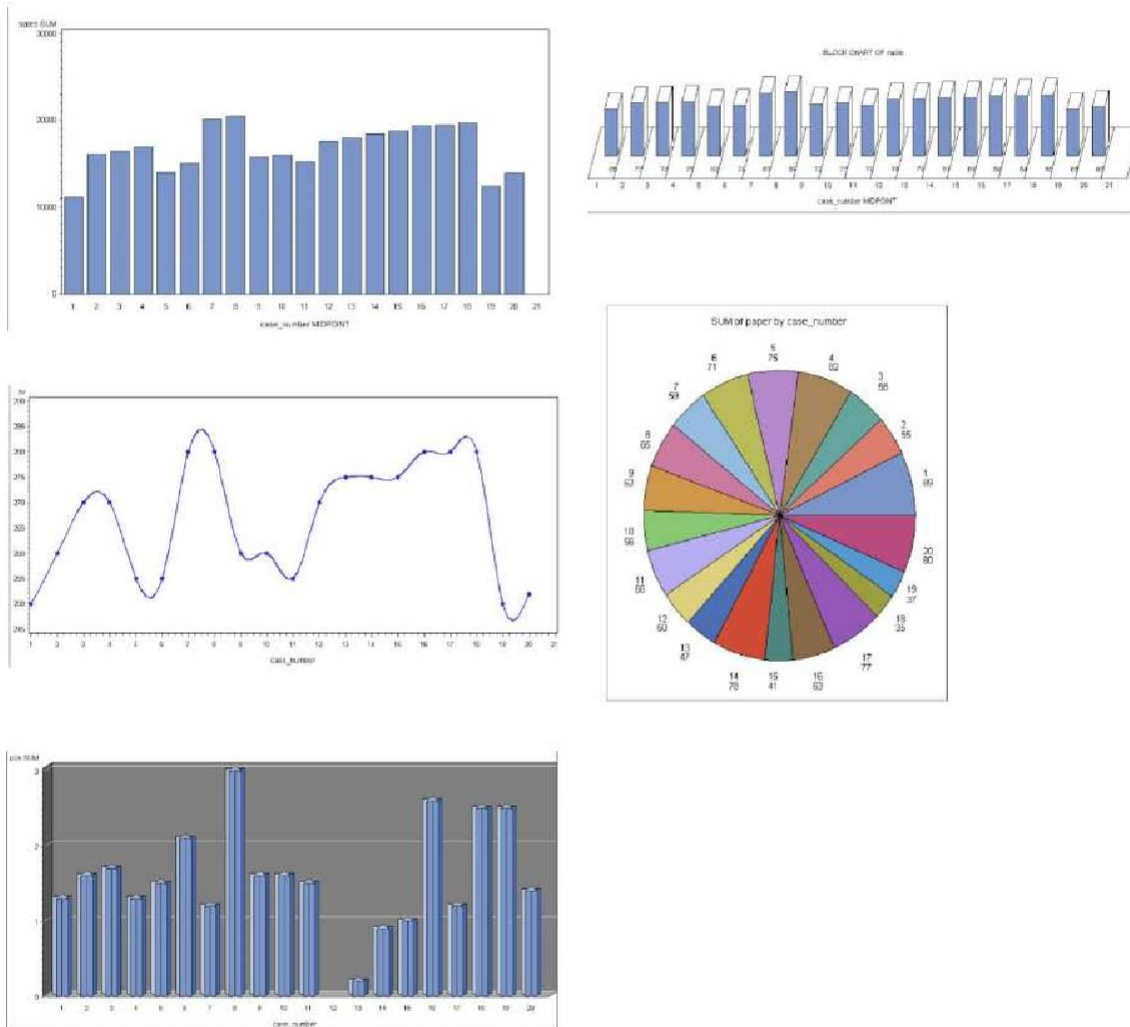


Figure 5.15 Preliminary SAS charts for the marketing/planning case study

To expedite the process of revealing potential relational information, think in terms of what one is specifically seeking. In this instance, it is to predict the future sales of the service product. That means looking for a graph to show a trend line. One type of simple graph that is

related to trend analysis is a line chart. Using SAS again, one can compute line charts for each of the five data sets. These charts are presented in Figure 5.16. The vertical axis consists of the dollar values, and the horizontal axis is the number ordering of observations as listed in the data sets.

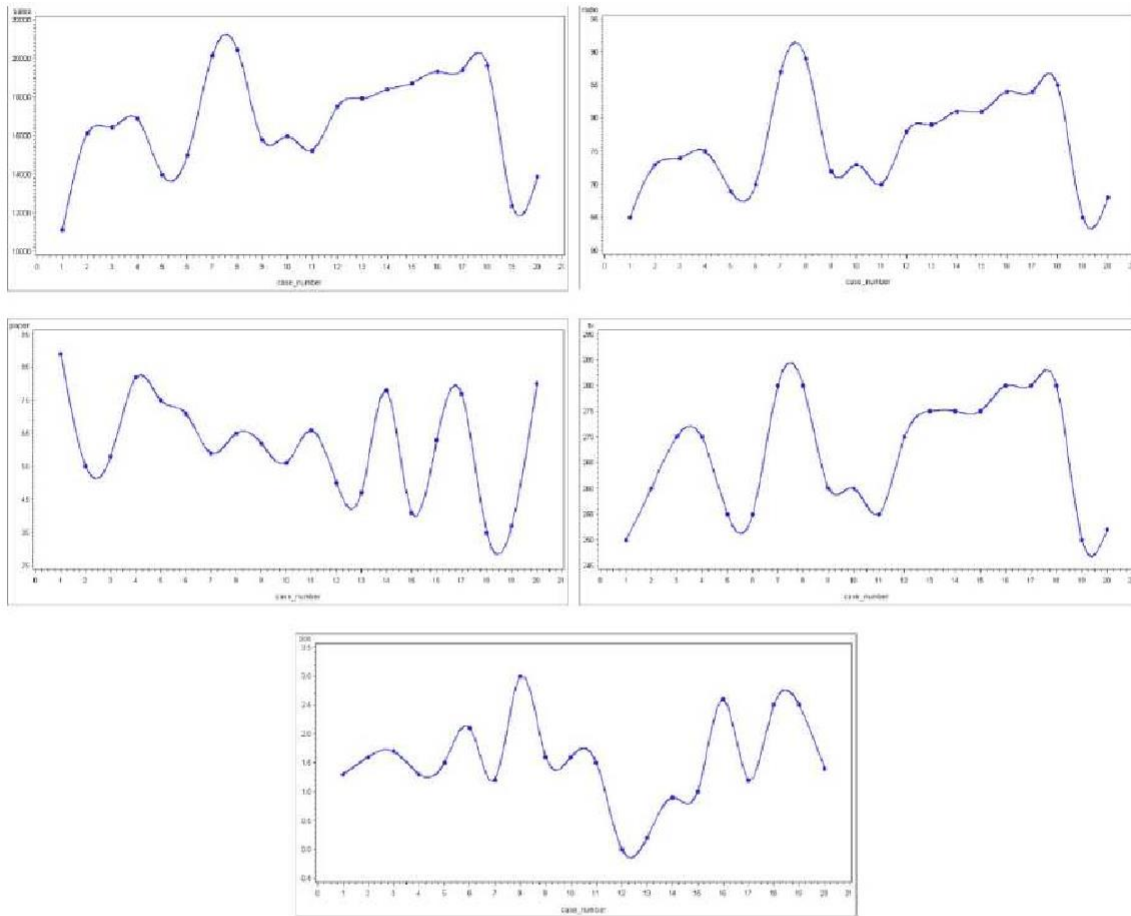
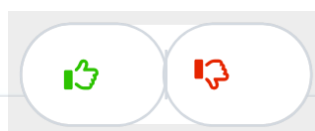


Figure 5.16 Preliminary SAS line charts for the marketing/planning case study

While providing a less confusing graphic presentation of the up-and-down behavior of the data, the charts in these figures still do not clearly reveal any possible trend information. Because the 20 months of data are not in any particular order and are not related to time, they are independent values that can be reordered. Reordering data or sorting it can be a part of the descriptive analytics process. Because trend is usually an upward or downward linear behavior, one might be able to observe a trend in the product sales data set if that data is reordered from low to high (or high to low). Reordering the sales by moving the 20 rows of data around such that sales is arranged from low to high is presented in Figure 5.17. Using this reordered data set, the SAS results are illustrated in the new line charts in Figure 5.18.



	case_number	sales	radio	paper	tv	pos
1	1	11125	65	89	250	1.3
2	19	12369	65	37	250	2.5
3	20	13882	68	80	252	1.4
4	5	13965	69	75	255	1.5
5	6	14999	70	71	255	2.1
6	11	15234	70	66	255	1.5
7	9	15789	72	62	260	1.6
8	10	15991	73	56	260	1.6
9	2	16121	73	55	260	1.6
10	3	16440	74	58	270	1.7
11	4	16876	75	82	270	1.3
12	12	17522	78	50	270	0
13	13	17933	79	47	275	0.2
14	14	18390	81	78	275	0.9
15	15	18723	81	41	275	1
16	16	19328	84	63	280	2.6
17	17	19399	84	77	280	1.2
18	18	19641	85	35	280	2.5
19	7	20167	87	59	280	1.2
20	8	20450	89	65	280	3

Figure 5.17 Reordered data in line charts for the marketing/planning case study

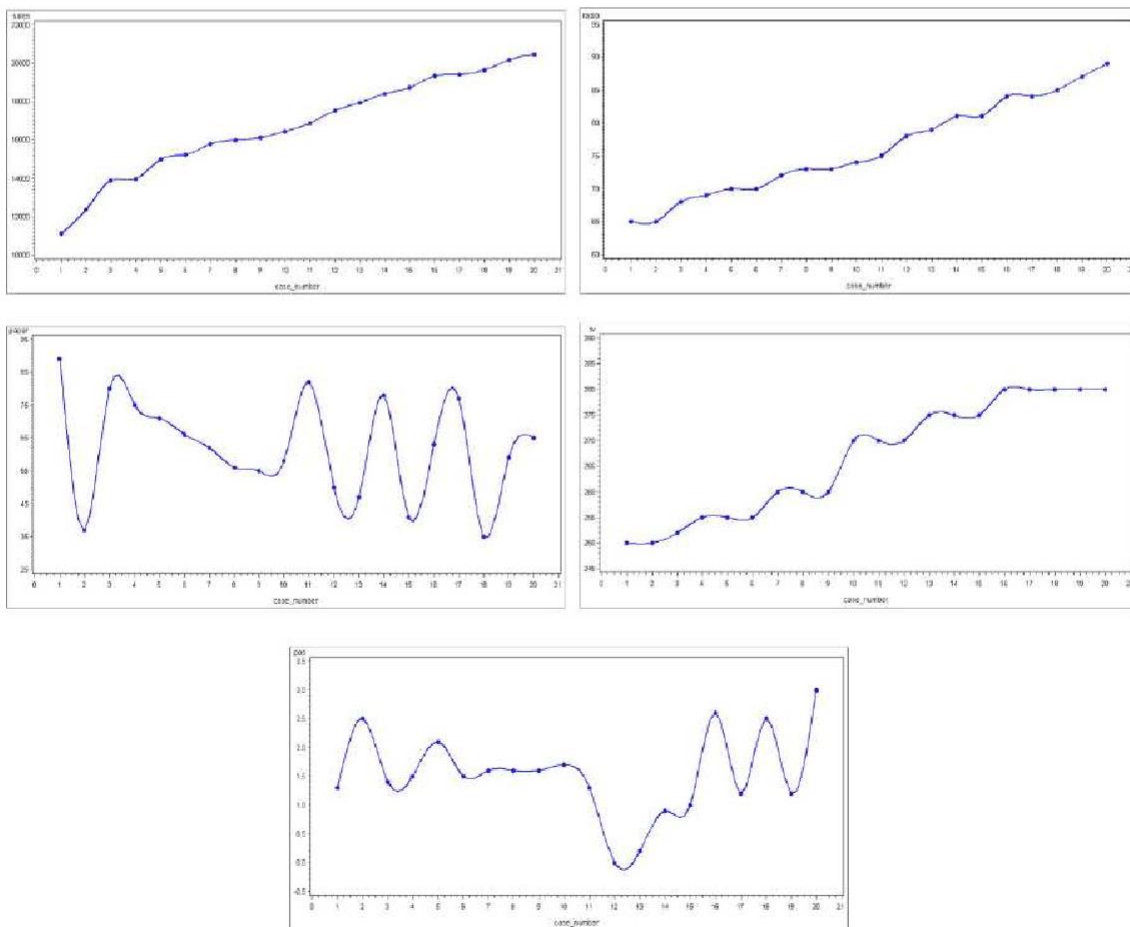
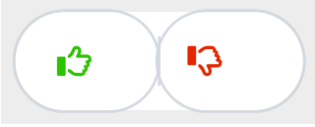


Figure 5.18 SAS line charts based on reordered data for the marketing/planning casestudy

Given the low to high reordering of the product sales as a guide, some of the other four line charts suggest a relationship with product sales. Both radio and TV commercials appear to have a similar low to high trending relationship that matches product sales. This suggests these two will be good predictive variables for product sales, whereas newspaper and POSads are still volatile in their charted relationships with product sales. Therefore, these two latter variables might not be useful in a model seeking to predict product sales. They cannot be ruled out at this point in the analysis, but they are suspected of adding little to a model for accurately forecasting product sales. Put another way, they appear to add unneeded variation that may take away from the accuracy of the model. Further analysis is called for to explore in more detail and sophistication the best set of predictive variables to predict the relationships in product sales.

In summary, for this case study, the descriptive analytics analysis has revealed a potential relationship between radio and TV commercials and future product sales, and it questions the relationship of newspaper and POS ads to sales. The managerial ramifications of these results might suggest discontinuing investing in newspaper and POS ads and more productively allocating funds to radio and TV commercials. Before such a reallocation can be justified, more analysis is needed.



Unit 4 – Predictive Analytics

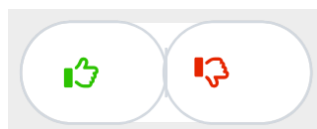
- Explain what logic-driven models are used for in business analytics (BA).
- Describe what a cause-and-effect diagram is used for in BA.
- Explain the difference between logic-driven and data-driven models.
- Explain how data mining can aid in BA.
- Explain why neural networks can be helpful in determining both associations and classification tasks required in some BA analyses.
- Explain how clustering is undertaken in BA.
- Explain how step-wise regression can be useful in BA.
- Explain how to use R-Squared adjusted statistics in BA.

Introduction

In Chapter 1, “What Is Business Analytics?” we defined predictive analytics as an application of advanced statistical, information software, or operations research methods to identify predictive variables and build predictive models to identify trends and relationships not readily observed in the descriptive analytic analysis. Knowing that relationships exist explains why one set of independent variables (predictive variables) influences dependent variables like business performance. Chapter 1 further explained that the purpose of the descriptive analytics step is to position decision makers to build predictive models designed to identify and predict future trends.

Picture a situation in which big data files are available from a firm’s sales and customer information (responses to differing types of advertisements, customer surveys on product quality, customer surveys on supply chain performance, sale prices, and so on). Assume also that a previous descriptive analytic analysis suggests that there is a relationship between certain customer variables, but there is a need to precisely establish a quantitative relationship between sales and customer behavior. Satisfying this need requires exploration into the big data to first establish whether a measurable, quantitative relationship does in fact exist and then develop a statistically valid model in which to predict future events. This is what the predictive analytics step in BA seeks to achieve.

Many methods can be used in this step of the BA process. Some are just to sort or classify big data into manageable files in which to later build a precise quantitative model. As previously mentioned in Chapter 3, “What Resource Considerations Are Important to Support Business



Analytics?” predictive modeling and analysis might consist of the use of methodologies, including those found in forecasting, sampling and estimation, statistical inference, data mining, and regression analysis. A commonly used methodology is multiple regression. (See Appendixes A, “Statistical Tools,” and E, “Forecasting,” for a discussion on multiple regression and ANOVA testing.) This methodology is ideal for establishing whether a statistical relationship exists between the predictive variables found in the descriptive analysis and the dependent variable one seeks to forecast. An example of its use will be presented in the last section of this chapter.

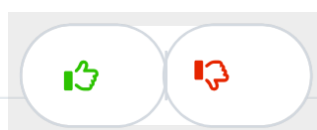
Although single or multiple regression models can often be used to forecast a trend line into the future, sometimes regression is not practical. In such cases, other forecasting methods, such as exponential smoothing or smoothing averages, can be applied as predictive analytics to develop needed forecasts of business activity. (See Appendix E.) Whatever methodology is used, the identification of future trends or forecasts is the principal output of the predictive analytics step in the BA process.

6.2 Predictive Modeling

Predictive modeling means developing models that can be used to forecast or predict future events. In business analytics, models can be developed based on logic or data.

6.2.1 Logic-Driven Models

A logic-driven model is one based on experience, knowledge, and logical relationships of variables and constants connected to the desired business performance outcome situation. The question here is how to put variables and constants together to create a model that can predict the future. Doing this requires business experience. Model building requires an understanding of business systems and the relationships of variables and constants that seek to generate a desirable business performance outcome. To help conceptualize the relationships inherent in a business system, diagramming methods can be helpful. For example, the cause-and-effect diagram is a visual aid diagram that permits a user to hypothesize relationships between potential causes of an outcome (see Figure 6.1). This diagram lists potential causes in terms of human, technology, policy, and process resources in an effort to establish some basic relationships that impact business performance. The diagram is used by tracing contributing and relational factors from the desired business performance goal back to possible causes, thus allowing the user to better picture sources of potential causes that could affect the performance. This diagram is sometimes referred to as a fishbone diagram because of its appearance.



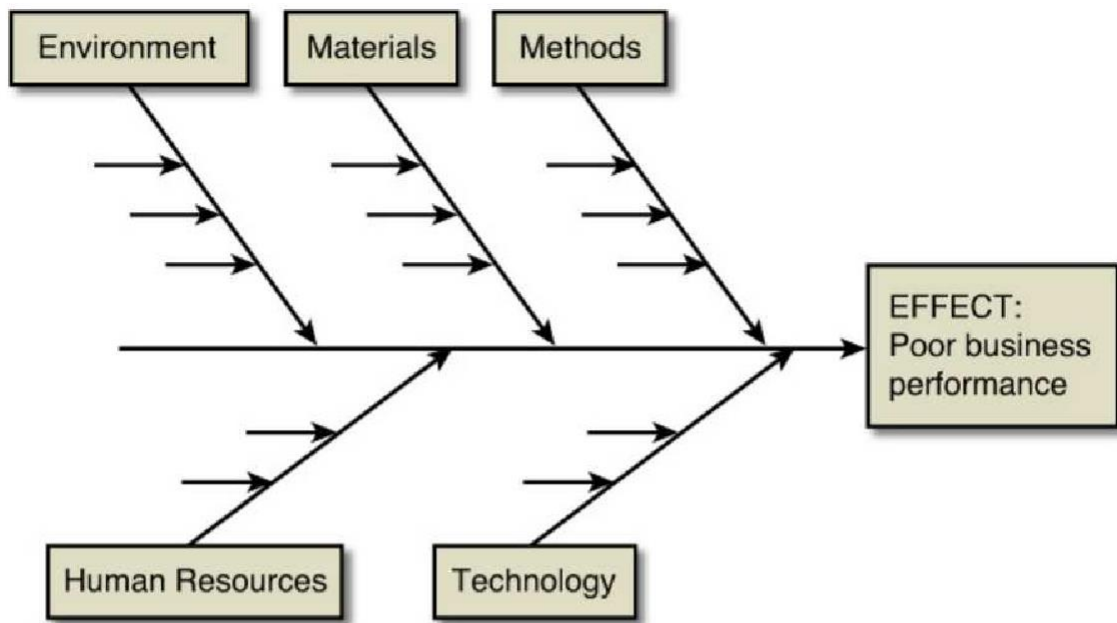


Figure 6.1 Cause-and-effect diagram*

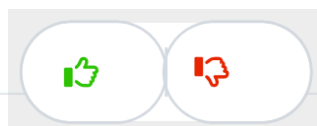
Another useful diagram to conceptualize potential relationships with business performance variables is called the influence diagram. According to Evans (2013, pp. 228–229), influence diagrams can be useful to conceptualize the relationships of variables in the development of models. An example of an influence diagram is presented in Figure 6.2. It maps the relationship of variables and a constant to the desired business performance outcome of profit. From such a diagram, it is easy to convert the information into a quantitative model with constants and variables that define profit in this situation:

Profit = Revenue – Cost, or

Profit = (Unit Price × Quantity Sold) - [(Fixed Cost) + (Variable Cost × Quantity Sold)],

or

P = (UP × QS) - [FC + (VC × QS)]



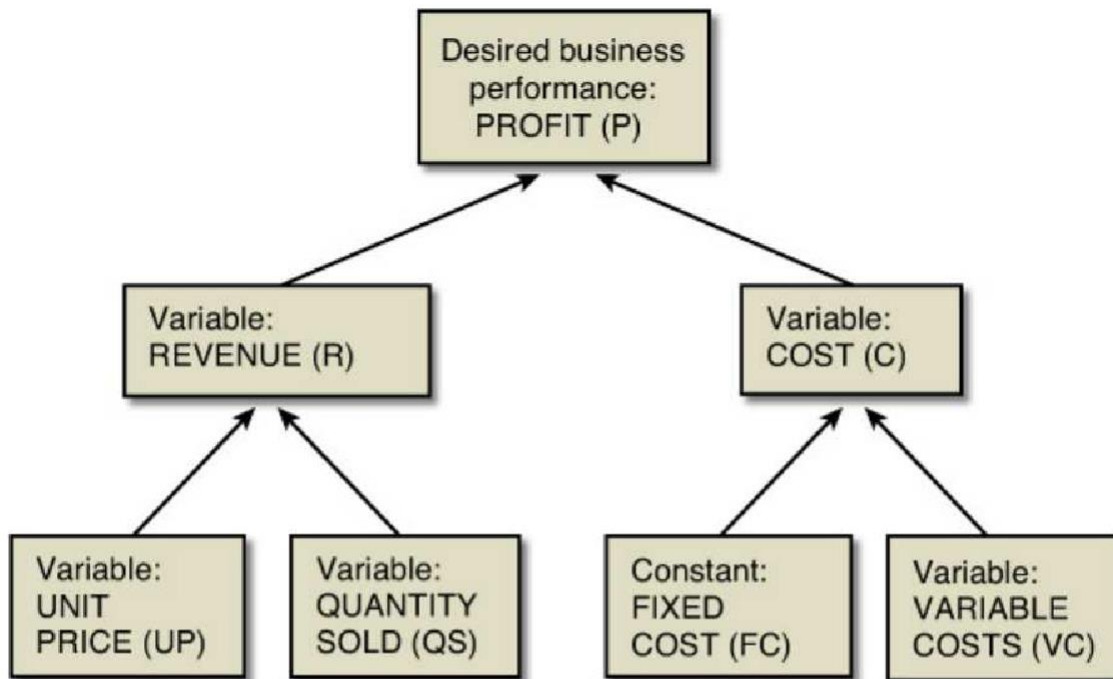


Figure 6.2 An influence diagram

The relationships in this simple example are based on fundamental business knowledge. Consider, however, how complex cost functions might become without some idea of how they are mapped together. It is necessary to be knowledgeable about the business systems being modeled in order to capture the relevant business behavior. Cause-and-effect diagrams and influence diagrams provide tools to conceptualize relationships, variables, and constants, but it often takes many other methodologies to explore and develop predictive models.

6.2.2 Data-Driven Models

Logic-driven modeling is often used as a first step to establish relationships through data-driven models (using data collected from many sources to quantitatively establish model relationships). To avoid duplication of content and focus on conceptual material in the chapters, we have relegated most of the computational aspects and some computer usage content to the appendixes. In addition, some of the methodologies are illustrated in the case problems presented in this book. Please refer to the Additional Information column in Table 6.1 to obtain further information on the use and application of the data-driven models.

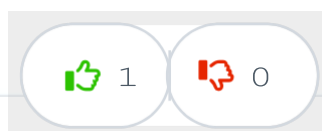


Data-Driven Models	Possible Applications	Additional Information
Sampling and Estimation	Generate statistical confidence intervals to define limitations and boundaries on future forecasts for other forecasting models.	Chapter 5, “What Is Descriptive Analytics?” Appendix A, Appendix E.
Regression Analysis	(1) Create a predictive equation useful for forecasting time series forecasts. (2) Weed out predictive variables in forecasting models that add little to predicting values. (3) Generate a trend line for forecasting.	Chapter 6, “What Is Predictive Analytics?” Chapter 8, “A Final Business Analytics Case Problem,” Appendix E.
Correlation Analysis	(1) Assess variable relationships. (2) Weed out predictive variables in forecasting models that add little to predicting values.	Chapter 6, Appendix E.
Probability Distributions	(1) Estimate trend behavior that follows certain types of probability distributions. (2) Conduct statistical tests to confirm significance of variables.	Chapter 5, Appendix A.

Table 6.1 Data-Driven Models

6.3 Data Mining

As mentioned in Chapter 3, data mining is a discovery-driven software application process that provides insights into business data by finding hidden patterns and relationships in big or small data and inferring rules from them to predict future behavior. These observed patterns and rules guide decision-making. This is not just numbers, but text and social media information from the Web. For example, Abrahams et al. (2013) developed a set of text-mining rules that automobile manufacturers could use to distill or mine specific vehicle component issues that emerge on the Web but take months to show up in complaints or other damaging media. These rules cut through the mountainous data that exists on the Web and are reported to provide marketing and competitive intelligence to manufacturers, distributors, service centers, and suppliers. Identifying a product’s defects and quickly recalling or correcting the problem before customers experience a failure reduce customer dissatisfaction when problems occur.



6.3.1 A Simple Illustration of Data Mining

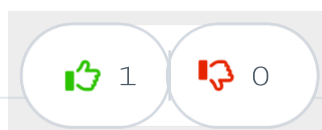
Suppose a grocery store has collected a big data file on what customers put into their baskets at the market (the collection of grocery items a customer purchases at one time). The grocery store would like to know if there are any associated items in a typical market basket. (For example, if a customer purchases product A, she will most often associate it or purchase it with product B.) If the customer generally purchases product A and B together, the store might only need to advertise product A to gain both product A's and B's sales. The value of knowing this association of products can improve the performance of the store by reducing the need to spend money on advertising both products. The benefit is real if the association holds true.

Finding the association and proving it to be valid require some analysis. From the descriptive analytics analysis, some possible associations may have been uncovered, such as product A's and B's association. With any size data file, the normal procedure in data mining would be to divide the file into two parts. One is referred to as a training data set, and the other as a validation data set. The training data set develops the association rules, and the validation data set tests and proves that the rules work. Starting with the training data set, a common data mining methodology is what-if analysis using logic-based software. SAS has a what-if logic-based software application, and so do a number of other software vendors (see Chapter 3). These software applications allow logic expressions. (For example, if product A is present, then is product B present?) The systems can also provide frequency and probability information to show the strength of the association. These software systems have differing capabilities, which permit users to deterministically simulate different scenarios to identify complex combinations of associations between product purchases in a market basket.

Once a collection of possible associations is identified and their probabilities are computed, the same logic associations (now considered association rules) are rerun using the validation data set. A new set of probabilities can be computed, and those can be statistically compared using hypothesis testing methods to determine their similarity. Other software systems compute correlations for testing purposes to judge the strength and the direction of the relationship. In other words, if the consumer buys product A first, it could be referred to as the Head and product B as the Body of the association (Nisbet et al., 2009, p. 128). If the same basic probabilities are statistically significant, it lends validity to the association rules and their use for predicting market basket item purchases based on groupings of products.

6.3.2 Data Mining Methodologies

Data mining is an ideal predictive analytics tool used in the BA process. We mentioned in Chapter 3 different types of information that data mining can glean, and Table 6.2 lists a small sampling of data mining methodologies to acquire different types of information. Some of the same tools used in the descriptive analytics step are used in the predictive step but are

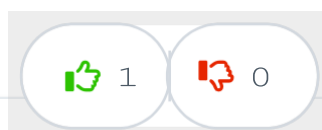


employed to establish a model (either based on logical connections or quantitative formulas) that may be useful in predicting the future.

Types of Information	Description	Sample of Data Mining Methodologies
Association	Occurrence linked to a single event.	Association rules (for example, if-then analysis), correlation analysis, neural networks.
Classification	Pattern that describes the group an item belongs to. Found by examining previous classified existing items and inferring a set of rules that guide the classification process.	Discriminant analysis, logistics regression, neural networks.
Clustering	Similar to classification when no groups have yet been defined. Helps discover different groupings within data.	Hierarchical clustering, K-mean clustering.
Forecasting	Used to predict values that can identify patterns in customer behavior.	Regression analysis, correlation analysis.
Sequence	Event that is linked over time.	Lag correlation analysis, cause-and-effect diagrams.

Table 6.2 Types of Information and Data Mining Methodologies

Several computer-based methodologies listed in Table 6.2 are briefly introduced here. Neural networks are used to find associations where connections between words or numbers can be determined. Specifically, neural networks can take large volumes of data and potential variables and explore variable associations to express a beginning variable (referred to as an input layer), through middle layers of interacting variables, and finally to an ending variable (referred to as an output). More than just identifying simple one-on-one associations, neural networks link multiple association pathways through big data like a collection of nodes in a network. These nodal relationships constitute a form of classifying groupings of variables as related to one another, but even more, related in complex paths with multiple associations (Nisbet et al., 2009, pp. 128–138). Differing software have a variety of association network function capabilities. SAS offers a series of search engines that can identify associations. SPSS has two versions of neural network software functions: Multilayer Perception (MLP) and Radial Basis Function (RBF). Both procedures produce a predictive model for one or more dependent variables based on the values of the predictive variables. Both allow a decision maker to develop, train, and use the software to identify particular traits (such as bad loan risks for a bank) based on characteristics from data collected on past customers.

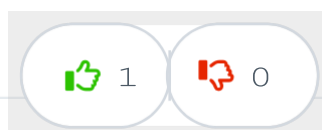


Discriminant analysis is similar to a multiple regression model except that it permits continuous independent variables and a categorical dependent variable. The analysis generates a regression function whereby values of the independent variables can be incorporated to generate a predicted value for the dependent variable. Similarly, logistic regression is like multiple regression. Like discriminant analysis, its dependent variable can be categorical. The independent variables in logistic regression can be either continuous or categorical. For example, in predicting potential outsource providers, a firm might use a logistic regression, in which the dependent variable would be to classify an outsource provider as either rejected (represented by the value of the dependent variable being zero) or acceptable (represented by the value of one for the dependent variable).

Hierarchical clustering is a methodology that establishes a hierarchy of clusters that can be grouped by the hierarchy. Two strategies are suggested for this methodology: agglomerative and divisive. The agglomerative strategy is a bottom-up approach, in which one starts with each item in the data and begins to group them. The divisive strategy is a top-down approach, in which one starts with all the items in one group and divides the group into clusters. How the clustering takes place can involve many different types of algorithms and differing software applications. One method commonly used is to employ a Euclidean distance formula that looks at the square root of the sum of distances between two variables, their differences squared. Basically, the formula seeks to match up variable candidates that have the least squared error differences. (In other words, they're closer together.)

K-mean clustering is a classification methodology that permits a set of data to be reclassified into K groups, where K can be set as the number of groups desired. The algorithmic process identifies initial candidates for the K groups and then interactively searches other candidates in the data set to be averaged into a mean value that represents a particular K group. The process of selection is based on maximizing the distance from the initial K candidates selected in the initial run through the list. Each run or iteration through the data set allows the software to select further candidates for each group.

The K-mean clustering process provides a quick way to classify data into differentiated groups. To illustrate this process, use the sales data in Figure 6.3 and assume these are sales from individual customers. Suppose a company wants to classify the sales customers into high and low sales groups.



	time	sale
1	1	13444
2	2	12369
3	3	15322
4	4	13965
5	5	14999
6	6	15234
7	7	12999
8	8	15991
9	9	16121
10	10	18654
11	11	16876
12	12	17522
13	13	17933
14	14	15233
15	15	18723
16	16	13855
17	17	19399
18	18	16854
19	19	20167
20	20	18654

Figure 6.3 Sales data for cluster classification problem

The SAS K-Mean cluster software can be found in Proc Cluster. Any integer value can designate the K number of clusters desired. In this problem set, K=2. The SAS printout of this classification process is shown in Table 6.3. The Initial Cluster Centers table listed the initial high (20167) and a low (12369) value from the data set as the clustering process begins. As it turns out, the software divided the customers into 9 high sales customers and 11 low sales customers.

The FASTCLUS Procedure						
Replace=FULL Radius=0 Maxclusters=2 Maxiter=1						
Initial Seeds						
Cluster	time	sale				
1	19.00000	20167.00000				
2	2.00000	12369.00000				
Criterion Based on Final Seeds = 797.6						
Cluster Summary						
Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Radius Exceeded	Nearest Cluster	Distance Between Cluster Centroids
1	9	789.8	1857.9		2	3806.2
2	11	879.3	2133.9		1	3806.2

Table 6.3 SAS K-Mean Cluster Solution



Consider how large big data sets can be. Then realize this kind of classification capability can be a useful tool for identifying and predicting sales based on the mean values.

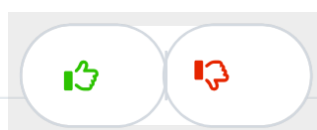
There are so many BA methodologies that no single section, chapter, or even book can explain or contain them all. The analytic treatment and computer usage in this chapter have been focused mainly on conceptual use. For a more applied use of some of these methodologies, note the case study that follows and some of the content in the appendixes.

6.4 Continuation of Marketing/Planning Case Study Example: Prescriptive Analytics Step in the BA Process

In the last sections of Chapters 5, 6, and 7, an ongoing marketing/planning case study of the relevant BA step discussed in those chapters is presented to illustrate some of the tools and strategies used in a BA problem analysis. This is the second installment of the case study dealing with the predictive analytics analysis step in BA. The prescriptive analysis step coming in Chapter 7, “What Is Prescriptive Analytics?” will complete the ongoing case study.

6.4.1 Case Study Background Review

The case study firm had collected a random sample of monthly sales information presented in Figure 6.4 listed in thousands of dollars. What the firm wants to know is, given a fixed budget of \$350,000 for promoting this service product, when it is offered again, how best should the company allocate budget dollars in hopes of maximizing the future estimated month’s product sales? Before the firm makes any allocation of budget, there is a need to understand how to estimate future product sales. This requires understanding the behavior of product sales relative to sales promotion efforts using radio, paper, TV, and point-of-sale (POS) ads.



	case_number	sales	radio	paper	tv	pos
1	1	11125	65	89	250	1.3
2	2	16121	73	55	260	1.6
3	3	16440	74	58	270	1.7
4	4	16876	75	82	270	1.3
5	5	13965	69	75	255	1.5
6	6	14999	70	71	255	2.1
7	7	20167	87	59	280	1.2
8	8	20450	89	65	280	3
9	9	15789	72	62	260	1.6
10	10	15991	73	56	260	1.6
11	11	15234	70	66	255	1.5
12	12	17522	78	50	270	0
13	13	17933	79	47	275	0.2
14	14	18390	81	78	275	0.9
15	15	18723	81	41	275	1
16	16	19328	84	63	280	2.6
17	17	19399	84	77	280	1.2
18	18	19641	85	35	280	2.5
19	19	12369	65	37	250	2.5
20	20	13882	68	80	252	1.4

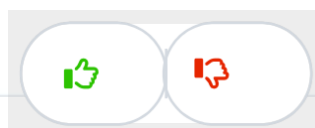
Figure 6.4 Data for marketing/planning case study

The previous descriptive analytics analysis in Chapter 5 revealed a potentially strong relationship between radio and TV commercials that might be useful in predicting future product sales. The analysis also revealed little regarding the relationship of newspaper and POS ads to product sales. So although radio and TV commercials are most promising, a more in-depth predictive analytics analysis is called for to accurately measure and document the degree of relationship that may exist in the variables to determine the best predictors of product sales.

6.4.2 Predictive Analytics Analysis

An ideal multiple variable modeling approach that can be used in this situation to explore variable importance in this case study and eventually lead to the development of a predictive model for product sales is correlation and multiple regression. We will use SAS's statistical package to compute the statistics in this step of the BA process.

First, we must consider the four independent variables—radio, TV, newspaper, POS—before developing the model. One way to see the statistical direction of the relationship (which is better than just comparing graphic charts) is to compute the Pearson correlation coefficients r between each of the independent variables with the dependent variable (product sales). The SAS correlation coefficients and their levels of significance are presented in Table 6.4. The



larger the Pearson correlation (regardless of the sign) and the smaller the Significance test values (these are t-tests measuring the significance of the Pearson r value; see Appendix A), the more significant the relationship. Both radio and TV are statistically significant correlations, whereas at a 0.05 level of significance, paper and POS are not statistically significant.

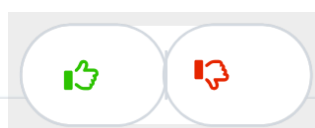
Statistic	Radio	Paper	TV	POS
Pearson Correlation <i>r</i> with Product Sales	.977	-.283	.958	.013
Significance Test (1-Tailed)*	.000	.113	.000	.479

Pearson Correlation Coefficients, N = 20 Prob > r under H0: Rho=0					
	sales	radio	paper	tv	pos
sales	1.00000 <.0001	0.97714 <.0001	-0.28307 0.2265	0.95797 <.0001	0.01265 0.9578
radio	0.97714 <.0001	1.00000	-0.23836 0.3115	0.96610 <.0001	0.06040 0.8003
paper	-0.28307 0.2265	-0.23836 0.3115	1.00000	-0.24588 0.2960	-0.09006 0.7057
tv	0.95797 <.0001	0.96610 <.0001	-0.24588 0.2960	1.00000	-0.03602 0.8802
pos	0.01265 0.9578	0.06040 0.8003	-0.09006 0.7057	-0.03602 0.8802	1.00000

*Values of 0.05 or less would designate a significant relationship with product sales

Table 6.4 SAS Pearson Correlation Coefficients: Marketing/Planning Case Study

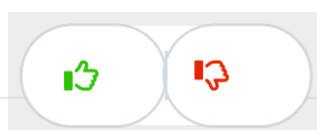
Although it can be argued that the positive or negative correlation coefficients should not automatically discount any variable from what will be a predictive model, the negative correlation of newspapers suggests that as a firm increases investment in newspaper ads, it will decrease product sales. This does not make sense in this case study. Given the illogic of such a relationship, its potential use as an independent variable in a model is questionable. Also, this negative correlation poses several questions that should be considered. Was the data set correctly collected? Is the data set accurate? Was the sample large enough to have included enough data for this variable to show a positive relationship? Should it be included for further analysis? Although it is possible that a negative relationship can statistically show up like this, it does not make sense in this case. Based on this reasoning and the fact that the correlation is not statistically significant, this variable (newspaper ads) will be removed from further consideration in this exploratory analysis to develop a predictive model.



Some researchers might also exclude POS based on the insignificance ($p=0.479$) of its relationship with product sales. However, for purposes of illustration, continue to consider it a candidate for model inclusion. Also, the other two independent variables (radio and TV) were found to be significantly related to product sales, as reflected in the correlation coefficients in the tables.

At this point, there is a dependent variable (product sales) and three candidate independent variables (POS, TV, and Radio) in which to establish a predictive model that can show the relationship between product sales and those independent variables. Just as a line chart was employed to reveal the behavior of product sales and the other variables in the descriptive analytic step, a statistical method can establish a linear model that combines the three predictive variables. We will use multiple regression, which can incorporate any of the multiple independent variables, to establish a relational model for product sales in this case study. Multiple regression also can be used to continue our exploration of the candidacy of the three independent variables.

The procedure by which multiple regression can be used to evaluate which independent variables are best to include or exclude in a linear model is called step-wise multiple regression. It is based on an evaluation of regression models and their validation statistics—specifically, the multiple correlation coefficients and the F-ratio from an ANOVA. SAS software and many other statistical systems build in the step-wise process. Some are called backward selection or step-wise regression, and some are called forward selection or step-wise regression. The backward step-wise regression starts with all the independent variables placed in the model, and the step-wise process removes them one at a time based on worst predictors first until a statistically significant model emerges. The forward step-wise regression starts with the best related variable (using correlation analysis as a guide), and then step-wise adds other variables until adding more will no longer improve the accuracy of the model. The forward step-wise regression process will be illustrated here manually. The first step is to generate individual regression models and statistics for each independent variable with the dependent variable one at a time. These three SAS models are presented in Tables 6.5, 6.6, and 6.7 for the POS, radio, and TV variables, respectively.



The REG Procedure
Model: MODEL1
Dependent Variable: sales

Number of Observations Read 20
Number of Observations Used 20

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	20819	20819	0.00	0.9578
Error	18	130109432	7228302		
Corrected Total	19	130130251			

Root MSE 2688.55013 R-Square 0.0002
Dependent Mean 16717 Adj R-Sq -0.0554
Coeff Var 16.08254

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	16649	1398.32203	11.91	<.0001
pos	1	44.14019	822.47123	0.05	0.9578

Table 6.5 SAS POS Regression Model: Marketing/Planning Case Study

The REG Procedure
Model: MODEL1
Dependent Variable: sales

Number of Observations Read 20
Number of Observations Used 20

Analysis of Variance

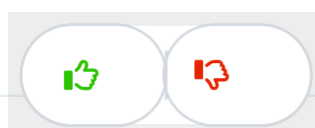
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	20819	20819	0.00	0.9578
Error	18	130109432	7228302		
Corrected Total	19	130130251			

Root MSE 2688.55013 R-Square 0.0002
Dependent Mean 16717 Adj R-Sq -0.0554
Coeff Var 16.08254

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	16649	1398.32203	11.91	<.0001
pos	1	44.14019	822.47123	0.05	0.9578

Table 6.6 SAS Radio Regression Model: Marketing/Planning Case Study

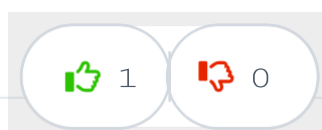


The REG Procedure					
Model: MODEL1					
Dependent Variable: sales					
Number of Observations Read					20
Number of Observations Used					20
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	119421443	119421443	200.73	<.0001
Error	18	10708808	594934		
Corrected Total	19	130130251			
Root MSE		771.31951	R-Square	0.9177	
Dependent Mean		16717	Adj R-Sq	0.9131	
Coeff Var		4.61393			
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-42229	4164.12104	-10.14	<.0001
tv	1	221.10431	15.60596	14.17	<.0001

Table 6.7 SAS TV Regression Model: Marketing/Planning Case Study

The computer printouts in the tables provide a variety of statistics for comparative purposes. Discussion will be limited here to just a few. The R-Square statistics are a precise proportional measure of the variation that is explained by the independent variable's behavior with the dependent variable. The closer the R-Square is to 1.00, the more of the variation is explained, and the better the predictive variable. The three variables' R-Squares are 0.0002 (POS), 0.9548 (radio), and 0.9177 (TV). Clearly, radio is the best predictor variable of the three, followed by TV and, without almost any relationship, POS. This latter result was expected based on the prior Pearson correlation. What it is suggesting is that only 0.0823 percent (1.000–0.9177) of the variation in product sales is explained by TV commercials.

From ANOVA, the F-ratio statistic is useful in actually comparing the regression model's capability to predict the dependent variable. As R-Square increases, so does the F-ratio because of the way in which they are computed and what is measured by both. The larger the F-ratio (like the R-Square statistic), the greater the statistical significance in explaining the variable's relationships. The three variables' F-ratios from the ANOVA tables are 0.00 (POS), 380.22 (radio), and 200.73 (TV). Both radio and TV are statistically significant, but POS has an insignificant relationship. To give some idea of how significant the relationships are, assuming a level of significance where $\alpha=0.01$, one would only need a cut-off value for the F-ratio of 8.10 to designate it as being significant. Not exceeding that F-ratio (as in the case of POS at 0.00) is the same as saying that the coefficient in the regression model for POS is no



different from a value of zero (no contribution to Product Sales). Clearly, the independent variables radio and TV appear to have strong relationships with the dependent variable. The question is whether the two combined or even three variables might provide a more accurate forecasting model than just using the one best variable like radio.

Continuing with the step-wise multiple regression procedure, we next determine the possible combinations of variables to see if a particular combination is better than the single variable models computed previously. To measure this, we have to determine the possible combinations for the variables and compute their regression models. The combinations are (1) POS and radio; (2) POS and TV; (3) POS, radio, and TV; and (4) radio and TV.

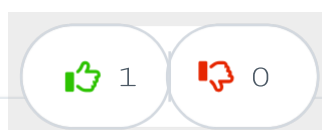
The resulting regression model statistics are summarized and presented in Table 6.8. If one is to base the selection decision solely on the R-Square statistic, there is a tie between the POS/radio/TV and the radio/TV combination (0.979 R-Square values). If the decision is based solely on the F-ratio value from ANOVA, one would select just the radio/TV combination, which one might expect of the two most significantly correlated variables.

Variable Combination	R-Square	R-Square (Adjusted)	F-Ratio
POS/radio	0.957	0.952	188.977
POS/TV	0.920	0.911	97.662
POS/radio/TV	0.979	0.951	123.315
Radio/TV	0.979	0.953	192.555

**Table 6.8 SAS Variable Combinations and Regression Model Statistics:
Marketing/Planning Case Study**

To aid in supporting a final decision and to ensure these analytics are the best possible estimates, we can consider an additional statistic. That tie breaker is the R-Squared (Adjusted) statistic, which is commonly used in multiple regression models.

The R-Square Adjusted statistic does not have the same interpretation as R-Square (a precise, proportional measure of variation in the relationship). It is instead a comparative measure of suitability of alternative independent variables. It is ideal for selection between independent variables in a multiple regression model. The R-Square adjusted seeks to take into account the phenomenon of the R-Square automatically increasing when additional independent variables are added to the model. This phenomenon is like a painter putting paint on a canvas, where more paint additively increases the value of the painting. Yet by continually adding



paint, there comes a point at which some paint covers other paint, diminishing the value of the original. Similarly, statistically adding more variables should increase the ability of the model to capture what it seeks to model. On the other hand, putting in too many variables, some of which may be poor predictors, might bring down the total predictive ability of the model. The R-Square adjusted statistic provides some information to aid in revealing this behavior.

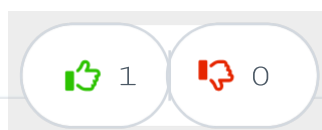
The value of the R-Square adjusted statistic can be negative, but it will always be less than or equal to that of the R-Square in which it is related. Unlike R-Square, the R-Square adjusted increases when a new independent variable is included only if the new variable improves the R-Square more than would be expected in the absence of any independent value being added. If a set of independent variables is introduced into a regression model one at a time in forward step-wise regression using the highest correlations ordered first, the R-Square adjusted statistic will end up being equal to or less than the R-Square value of the original model. By systematic experimentation with the R-Square adjusted recomputed for each added variable or combination, the value of the R-Square adjusted will reach a maximum and then decrease. The multiple regression model with the largest R-Square adjusted statistic will be the most accurate combination of having the best fit without excessive or unnecessary independent variables. Again, just putting all the variables into a model may add unneeded variability, which can decrease its accuracy. Thinning out the variables is important.

Finally, in the step-wise multiple regression procedure, a final decision on the variables to be included in the model is needed. Basing the decision on the R-Square adjusted, the best combination is radio/TV. The SAS multiple regression model and support statistics are presented in Table 6.9.

Variable Combination	R-Square	R-Square (Adjusted)	F-Ratio
POS/radio	0.957	0.952	188.977
POS/TV	0.920	0.911	97.662
POS/radio/TV	0.979	0.951	123.315
Radio/TV	0.979	0.953	192.555

**Table 6.9 SAS Best Variable Combination Regression Model and Statistics:
Marketing/Planning Case Study**

Although there are many other additional analyses that could be performed to validate this model, we will use the SAS multiple regression model in Table 6.9 for the firm in this case study. The forecasting model can be expressed as follows:



$$Y_p = -17150 + 275.69065 X_1 + 48.34057 X_2$$

where:

Y_p = the estimated number of dollars of product sales

X_1 = the number of dollars to invest in radio commercials

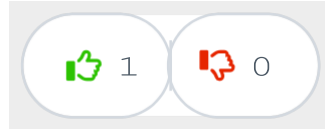
X_2 = the number of dollars to invest in TV commercials

Because all the data used in the model is expressed as dollars, the interpretation of the model is made easier than using more complex data. The interpretation of the multiple regression model suggests that for every dollar allocated to radio commercials (represented by X_1), the firm will receive \$275.69 in product sales (represented by Y_p in the model). Likewise, for every dollar allocated to TV commercials (represented by X_2), the firm will receive \$48.34 in product sales.

A caution should be mentioned on the results of this case study. Many factors might challenge a result, particularly those derived from using powerful and complex methodologies like multiple regression. As such, the results may not occur as estimated, because the model is not reflecting past performance. What is being suggested here is that more analysis can always be performed in questionable situations. Also, additional analysis to confirm a result should be undertaken to strengthen the trust that others must have in the results to achieve the predicted higher levels of business performance.

In summary, for this case study, the predictive analytics analysis has revealed a more detailed, quantifiable relationship between the generation of product sales and the sources of promotion that best predict sales. The best way to allocate the \$350,000 budget to maximize product sales might involve placing the entire budget into radio commercials because they give the best return per dollar of budget. Unfortunately, there are constraints and limitations regarding what can be allocated to the different types of promotional methods. Optimizing the allocation of a resource and maximizing business performance necessitate the use

of special business analytic methods designed to accomplish this task. This requires the



Unit No 5 : Prescriptive Analytics

Unit objectives:

List and describe the commonly used prescriptive analytics in the business analytics (BA) process.

Explain the role of case studies in prescriptive analytics.

Explain how fitting a curve can be used in prescriptive analytics.

Explain how to formulate a linear programming model.

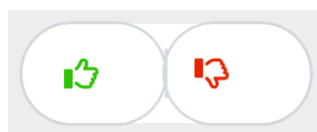
Explain the value of linear programming in the prescriptive analytics step of BA.

Introduction

After undertaking the descriptive and predictive analytics steps in the BA process, one should be positioned to undertake the final step: prescriptive analytics analysis. The prior analysis should provide a forecast or prediction of what future trends in the business may hold. For example, there may be significant statistical measures of increased (or decreased) sales, profitability trends accurately measured in dollars for new market opportunities, or measured cost savings from a future joint venture.

If a firm knows where the future lies by forecasting trends, it can best plan to take advantage of possible opportunities that the trends may offer. Step 3 of the BA process, prescriptive analytics, involves the application of decision science, management science, or operations research methodologies to make best use of allocable resources. These are mathematically based methodologies and algorithms designed to take variables and other parameters into a quantitative framework and generate an optimal or near-optimal solution to complex problems. These methodologies can be used to optimally allocate a firm's limited resources to take best advantage of the opportunities it has found in the predicted future trends. Limits on human, technology, and financial resources prevent any firm from going after all the opportunities. Using prescriptive analytics allows the firm to allocate limited resources to optimally or near-optimally achieve the objectives as fully as possible.

In Chapter 3, "What Resource Considerations Are Important to Support Business Analytics?" the relationships of methodologies to the BA process were expressed as a function of certification exam content. The listing of the prescriptive analytic methodologies as they are



in some cases utilized in the BA process is again presented in Figure 7.1 to form the basis of this chapter's content.

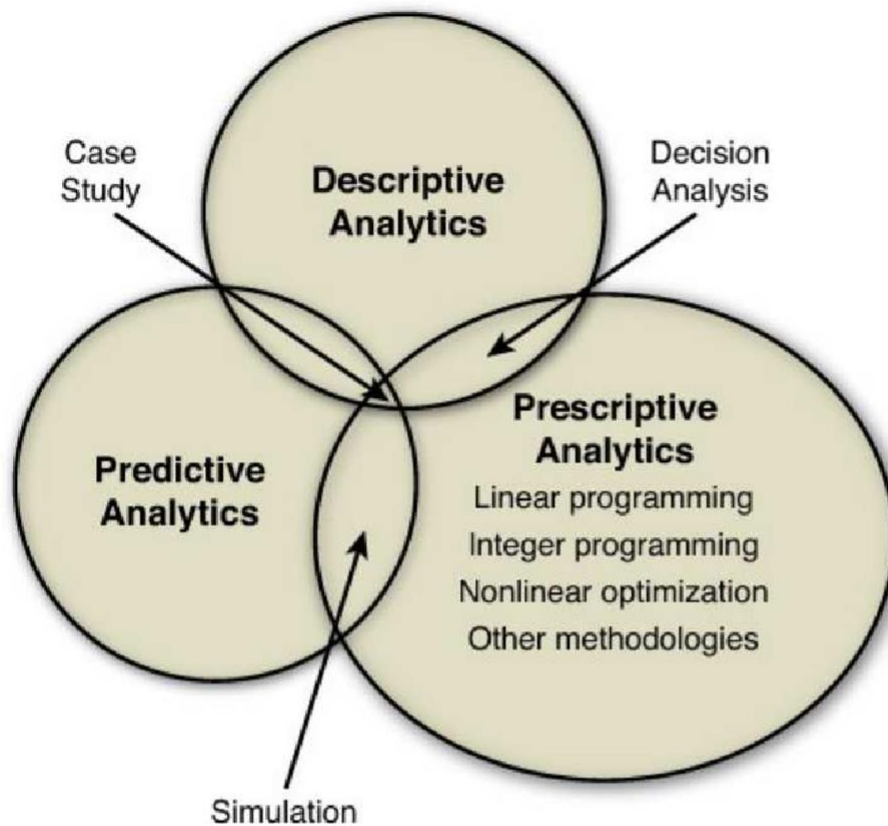
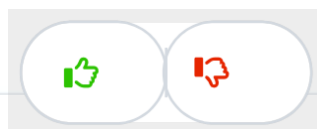


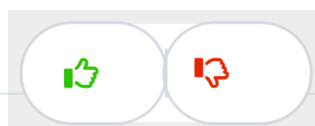
Figure 7.1 Prescriptive analytic methodologies

7.2 Prescriptive Modeling

The listing of prescriptive analytic methods and models in Figure 7.1 is but a small grouping of many operations research, decision science, and management science methodologies that are applied in this step of the BA process. Most of the methodologies in Table 7.1 are explained throughout this book. (See the Additional Information column in Table 7.1.)



Data-Driven Models	Possible Applications	Additional Information
Linear Programming (LP)	A general-purpose modeling methodology is applied to multiconstrained, multivariable problems when an optimal solution is sought. It is ideal for complex and large-scale problems when limited resources are being allocated to multiple uses. Examples include allocating advertising budgets to differing media, allocating human and technology resources to product production, and optimizing blends of mixing ingredients to minimize costs of food products.	Chapters 7 and 8, "A Final Business Analytics Case Problem" Appendix B, "Linear Programming" Appendix C, "Duality and Sensitivity Analysis in Linear Programming"
Integer Programming	This is the same as LP, but it permits decision variables to be integer values. Examples include allocating stocks to portfolios, allocating personnel to jobs, and allocating types of crops to farm lands.	Appendix D, "Integer Programming"
Nonlinear Optimization	A large class of methodologies and algorithms is used to analyze and solve for optimal or near-optimal solutions when the behavior of the data is nonlinear. Examples include solving for optimized allocations of human, technology, and systems whose data appears to form a cost or profit function that is quadratic, cubic, or nonlinear in some way.	Chapters 7 and 8 Appendix E, "Forecasting"



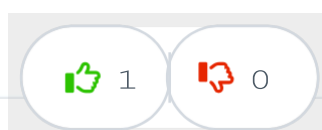
Decision Analysis	A set of methodologies, models, or principles is used to analyze and guide decision-making when multiple choices face the decision maker in differing decision environments (for example, certainty, risk, and uncertainty). Examples include selecting one from a set of computer systems, trucks, or site locations for a service facility.	Appendix G, "Decision Theory"
Case Studies	A learning aid provides practical experience by offering real or hypothetical case studies of real-world applications of BA. For example, case studies can simulate the issues and challenges in an actual problem setting. This kind of simulation can prep decision makers to anticipate and prepare for what has been predicted to occur by the predicted analytics step in the BA process. For example, a case study discussion on how to cope with organization growth might provide a useful decision-making environment for a firm whose analytics have predicted growth in the near future.	This is beyond the scope of this book. See Sekaran and Bougie (2013); Adkins (2006).
Simulation	This methodology can be used in prescriptive analysis in situations where parameters are probabilistic, nonlinear, or just too complex to use with other optimization models that require deterministic or linear behavior. For example, a bank might want to simulate the transactions it currently uses to process a loan application to determine if changes in the process might reduce time and improve performance. The simulation model might be used to test alternative process scenarios.	Appendix F, "Simulation"

Table 7.1 Select Prescriptive Analytic Models

7.3 Nonlinear Optimization

The prescriptive methodologies in Table 7.1 are explained in detail in the referenced chapters and appendixes, but nonlinear optimization will be discussed here. When business performance cost or profit functions become too complex for simple linear models to be useful, exploration of nonlinear functions is a standard practice in BA. Although the predictive nature of exploring for a mathematical expression to denote a trend or establish a forecast falls mainly in the predictive analytics step of BA, the use of the nonlinear function to optimize a decision can fall in the prescriptive analytics step.

As mentioned previously, there are many mathematical programming nonlinear methodologies and solution procedures designed to generate optimal business performance solutions. Most



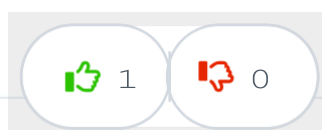
of them require careful estimation of parameters that may or may not be accurate, particularly given the precision required of a solution that can be so precariously dependent upon parameter accuracy. This precision is further complicated in BA by the large data files that should be factored into the model-building effort.

To overcome these limitations and be more inclusive in the use of large data, we can apply regression software. As illustrated in Appendix E, curve-fitting software can be used to generate predictive analytic models that can also be utilized to aid in making prescriptive analytic decisions.

For purposes of illustration, SAS's software will be used to fit data to curves in this chapter. Suppose that a resource allocation decision is being faced whereby one must decide how many computer servers a service facility should purchase to optimize the firm's costs of running the facility. The firm's predictive analytics effort has shown a growth trend. A new facility is called for if costs can be minimized. The firm has a history of setting up large and small service facilities and has collected the 20 data points in Figure 7.2. Whether there are 20 or 20,000 items in the data file, SAS can be used to fit data based on regression mathematics to a nonlinear line that best minimizes the distance from the data items to the line. The software then converts the line into a mathematical expression useful for forecasting.

	server	cost
1	1	27654
2	2	24789
3	3	21890
4	4	21633
5	5	15843
6	6	12567
7	7	8943
8	8	6789
9	9	4533
10	10	4678
11	11	5321
12	12	5765
13	13	5432
14	14	9995
15	15	13522
16	16	17563
17	17	22732
18	18	22643
19	19	24621
20	20	28111

Figure 7.2 Data for SAS curve fitting



In this server problem, the basic data has a u-shaped function, as presented in Figure 7.3. This is a classic shape for most cost functions in business. In this problem, it represents the balancing of having too few servers (resulting in a costly loss of customer business through dissatisfaction and complaints with the service) or too many servers (excessive waste in investment costs because of underutilized servers). Although this is an overly simplified example with little and nicely ordered data for clarity purposes, in big data situations, cost functions are considerably less obvious.

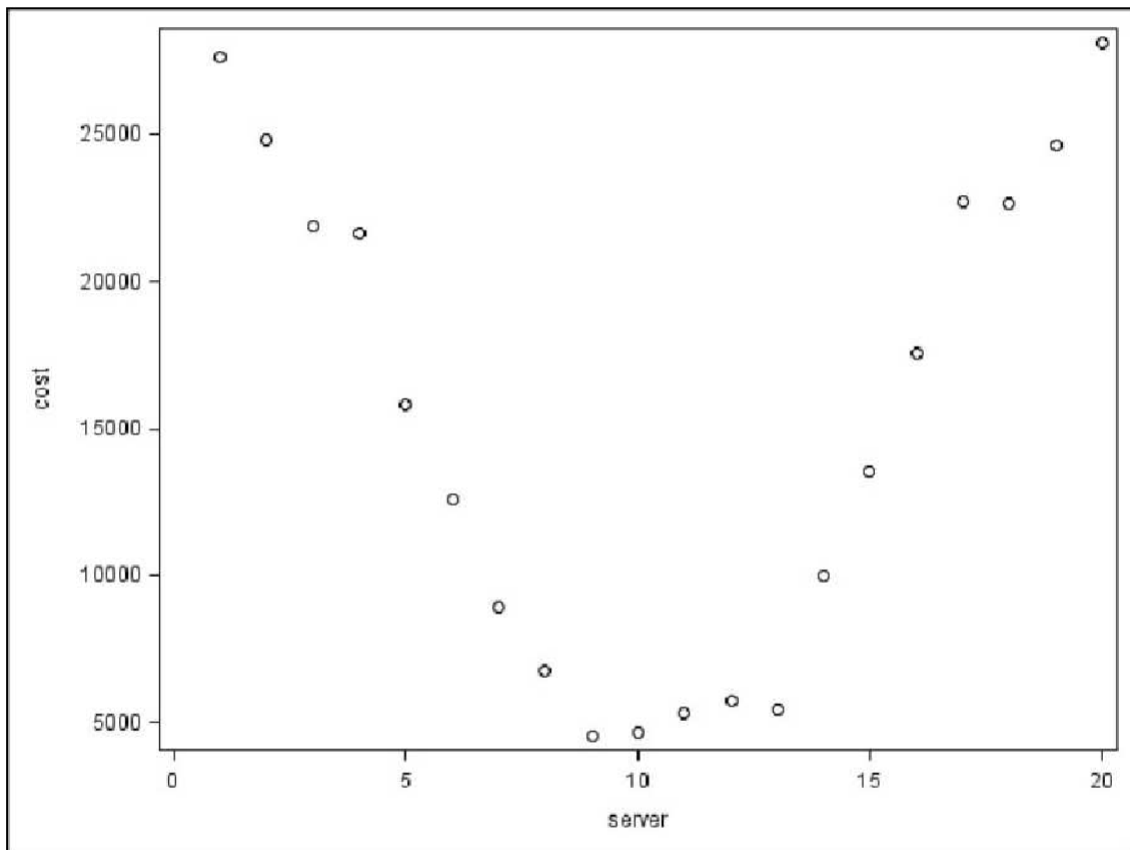
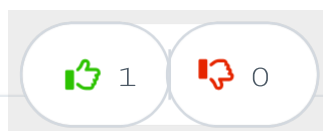


Figure 7.3 Server problem basic data cost function

The first step in curve fitting is to generate the best-fitting curve to the data. Using SAS and the data in Figure 7.2, the regression process seeks to minimize the distance by creating a line in one of the eight regression models in Figure 7.3. Doing this in SAS requires the selection of a set of functions that the analyst might believe is a good fit. The number of regression functions selected can be flexible. SAS offers a wide number of possible regression models to choose from. The result is a series of regression models and statistics, including ANOVA and other testing statistics. It is known from the previous illustration of regression that the



adjusted R-Square statistic can reveal the best estimated relationship between the independent (number of servers) and dependent (total cost) variables. These statistics are presented in Table 7.2. The best adjusted R-Square value (the largest) occurs with the quadratic model, followed by the cubic model. The more detailed supporting statistics for both of these models are presented in Table 7.3. The graph for all the SPSS curve-fitting models appears in Figure 7.4.

Linear			
Root MSE	8687.28965	R-Square	0.0011
Dependent Mean	15251	Adj R-Sq	-0.0544
Coeff Var	56.96135		
Logarithmic			
Root MSE	8376.01994	R-Square	0.0714
Dependent Mean	15251	Adj R-Sq	0.0198
Coeff Var	54.92040		
Inverse			
Root MSE	7825.69613	R-Square	0.1894
Dependent Mean	15251	Adj R-Sq	0.1444
Coeff Var	51.31200		
Quadratic			
Root MSE	2342.31463	R-Square	0.9314
Dependent Mean	15251	Adj R-Sq	0.9233
Coeff Var	15.35823		
Cubic			
Root MSE	2404.00949	R-Square	0.9320
Dependent Mean	15251	Adj R-Sq	0.9193
Coeff Var	15.76276		
S-Curve			
Root MSE	0.62652	R-Square	0.1454
Dependent Mean	9.44856	Adj R-Sq	0.0979
Coeff Var	6.63085		
Logistic			
R-Square	0.0047	Max-rescaled R-Square	0.0047
Growth			
Root MSE	0.67750	R-Square	0.0006
Dependent Mean	9.44856	Adj R-Sq	-0.0549
Coeff Var	7.17041		

Table 7.2 Adjusted R-Square Values of All SAS Models



1



0

Quadratic-Full

The REG Procedure
Model: MODEL1
Dependent Variable: cost

Number of Observations Read	20	
Number of Observations Used	20	

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1266704838	633352419	115.44	<.0001
Error	17	93269443	5486438		
Corrected Total	19	1359974281			

	Root MSE	2342.31463	R-Square	0.9314
	Dependent Mean	15251	Adj R-Sq	0.9233
	Coeff Var	15.35823		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	35418	1742.63922	20.32	<.0001
server	1	-5589.43151	382.18778	-14.62	<.0001
quadratic_server	1	268.44919	17.67797	15.19	<.0001

Cubic-Full

The REG Procedure
Model: MODEL1
Dependent Variable: cost

Number of Observations Read	20	
Number of Observations Used	20	

Analysis of Variance

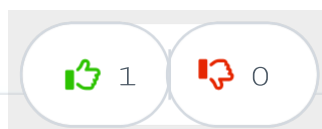
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1267506095	422502032	73.11	<.0001
Error	16	92468186	5779262		
Corrected Total	19	1359974281			

	Root MSE	2404.00949	R-Square	0.9320
	Dependent Mean	15251	Adj R-Sq	0.9193
	Coeff Var	15.76276		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	36134	2625.97612	13.76	<.0001
server	1	-5954.73759	1056.59555	-5.64	<.0001
quadratic_server	1	310.89531	115.43050	2.69	0.0160
cubic_server	1	-1.34750	3.61891	-0.37	0.7145

Table 7.3 Quadratic and Cubic Model SAS Statistics



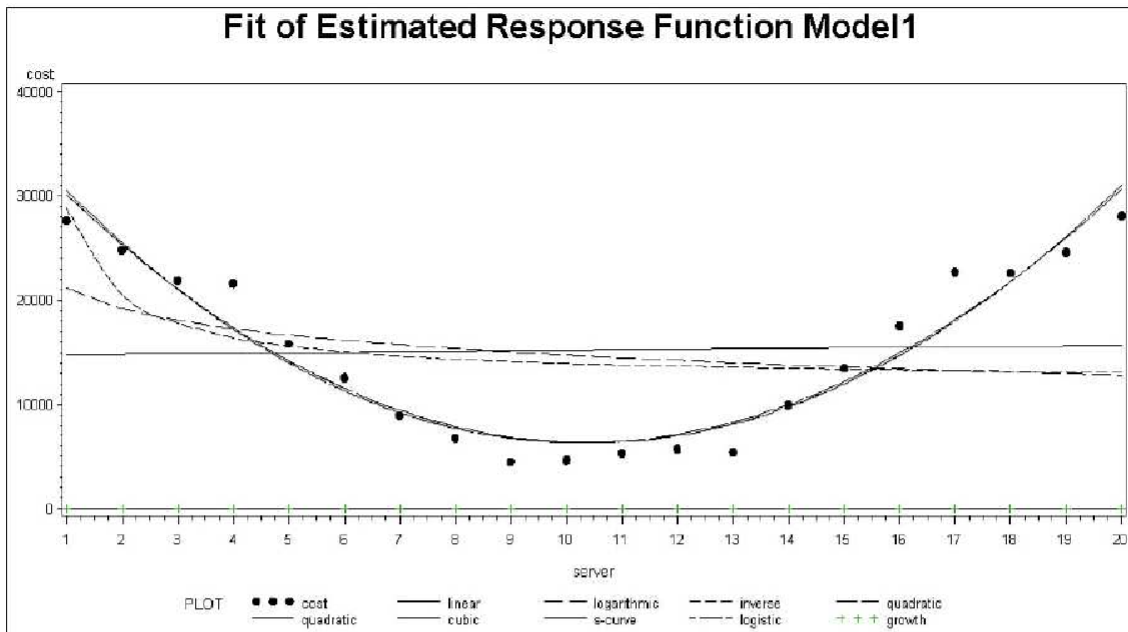


Figure 7.4 Graph of all SAS curve-fitting models

From Table 7.3, the resulting two statistically significant curve-fitted models follow:

$$Y_p = 35418 - 5589.432 X + 268.445 X^2 \text{ [Quadratic model]}$$

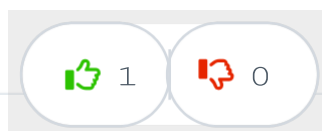
$$Y_p = 36134 - 5954.738 X + 310.895 X^2 - 1.347 X^3 \text{ [Cubic model]}$$

where:

Y_p = the forecasted or predicted total cost

X = the number of computer servers

For purposes of illustration, we will use the quadratic model. In the next step of using the curve-fitted models, one can either use calculus to derive the cost minimizing value for X (number of servers) or perform a deterministic simulation where values of X are substituted



into the model to compute and predict the total cost (Y_p). The calculus-based approach is presented in the “Addendum” section of this chapter.

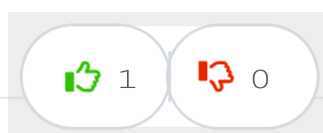
As a simpler solution method to finding the optimal number of servers, simulation can be used. Representing a deterministic simulation (see Appendix F, Section F.2.1), the resulting costs of servers can be computed using the quadratic model, as presented in Figure 7.5. These values were computed by plugging the number of server values (1 to 20) into the Y_p quadratic function one at a time to generate the predicted values for each of the server possibilities. Note that the lowest value in these predicted values occurs with the acquisition of 10 servers at \$6367.952, and the next lowest is at 11 servers at \$6415.865. In the actual data in Figure 7.2, the minimum total cost point occurs at 9 servers at \$4533, whereas the next lowest total cost is \$4678 occurring at 10 servers. The differences are due to the estimation process of curve fitting. Note in Figure 7.3 that the curve that is fitted does not touch the lowest 5 cost values. Like regression in general, it is an estimation process, and although the ANOVA statistics in the quadratic model demonstrate a strong relationship with the actual values, there is some error. This process provides a near-optimal solution but does not guarantee one.

Figure 7.5 Predicted total cost in server problem for each server alternative

Like all regression models, curve fitting is an estimation process with risks, but the supporting statistics, like ANOVA, provide some degree of confidence in the resulting solution.

Finally, it must be mentioned that many other nonlinear optimization methodologies exist. Some, like quadratic programming, are considered constrained optimization models (like LP). These topics are beyond the scope of this book. For additional information on nonlinear programming, see King and Wallace (2013), Betts (2009), and Williams (2013). Other methodologies, like the use of calculus in this chapter, are useful in solving for optimal solutions in unconstrained problem settings. For additional information on calculus methods, see Spillers and MacBain (2009), Luptacik (2010), and Kwak and Schniederjans (1987).

7.4 Continuation of Marketing/Planning Case Study Example: Prescriptive Step in the BA Analysis



In Chapter 5, “What Is Descriptive Analytics?” and Chapter 6, “What Is Predictive Analytics?” an ongoing marketing/planning case study was presented to illustrate some of the tools and strategies used in a BA problem analysis. This is the third and final installment of the case study dealing with the prescriptive analytics step in BA.

7.4.1 Case Background Review

The predictive analytics analysis in Chapter 6 revealed a statistically strong relationship between radio and TV commercials that might be useful in predicting future product sales. The ramifications of these results suggest a better allocation of funds away from paper and POS ads to radio and TV commercials. Determining how much of the \$350,000 budget should be allocated between the two types of commercials requires the application of an optimization decision-making methodology.

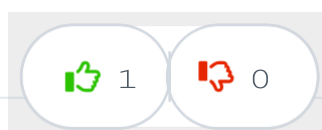
7.4.2 Prescriptive Analysis

The allocation problem of the budget to purchase radio and TV commercials is a multivariable (there are two media to consider), constrained (there are some limitations on how one can allocate the budget funds), optimization problem (BA always seeks to optimize business performance). Many optimization methods could be employed to determine a solution to this problem. Considering the singular objective of maximizing estimated product sales, linear programming (LP) is an ideal methodology to apply in this situation. To employ LP to model this problem, use the six-step LP formulation procedure explained in Appendix B.

7.4.2.1 Formulation of LP Marketing/Planning Model

In the process of exploring the allocation options, a number of limitations or constraints on placing radio and TV commercials were observed. The total budget for all the commercials was set at a maximum of \$350,000 for the next monthly campaign. To receive the radio commercial price discount requires a minimum budget investment of \$15,000. To receive the TV commercials price discount requires a minimum budget investment of \$75,000. Because the radio and TV stations are owned by the same corporation, there is an agreement that for every dollar of radio commercials required, the client firm must purchase \$2 in TV commercials. Given these limitations and the modeled relationship found in the previous predictive analysis, one can formulate the budget allocation decision as an LP model using a five-step LP formulation procedure (see Appendix B, Section B.4.1):

1. Determine the type of problem—This problem seeks to maximize dollar product sales by determining how to allocate budget dollars over radio and TV commercials. For each dollar



of radio commercials estimated with the regression model, \$275.691 will be received, and for each dollar of TV commercials, \$48.341 will be received. Those two parameters are the product sales values to maximize. Therefore, it will be a maximization model.

2. Define the decision variables—The decision variables for the LP model are derived from the multiple regression model's independent variables. The only adjustment is the monthly timeliness of the allocation of the budget:

X_1 = the number of dollars to invest in radio commercials for the next monthly campaign

X_2 = the number of dollars to invest in TV commercials for the next monthly campaign

3. Formulate the objective function—Because the multiple regression model defines the dollar sales as a linear function with the two independent variables, the same dollar coefficients from the regression model can be used as the contribution coefficients in the objective function. This results in the following LP model objective function:

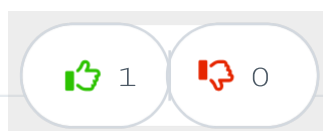
Maximize: $Z = 275.691 X_1 + 48.341 X_2$

4. Formulate the constraints—Given the information on the limitations in this problem, there are four constraints:

Constraint 1—No more than \$350,000 is allowed for the total budget to allocate to both radio (X_1) and TV (X_2) commercials. So add $X_1 + X_2$ and set it less than or equal to 350,000 to formulate the first constraint as follows:

$X_1 + X_2 \leq 350000$

Constraint 2—To get a discount on radio (X_1) commercials, the firm must allocate a minimum of \$15,000 to radio. The constraint for this limitation follows:



$$X1 \geq 15000$$

Constraint 3—Similar to Constraint 2, to get a discount on TV (X2) commercials, the firm must allocate a minimum of \$75,000 to TV. The constraint for this limitation follows:

$$X2 \geq 75000$$

Constraint 4—This is a blending problem constraint (see Appendix B, Section B.6.3). What is needed is to express the relationship as follows:

which is to say, for each one unit of X1, one must acquire two units of X2. Said differently, the ratio of one unit of X1 is equal to two units of X2. Given the expression, use algebra to cross-multiply such that

$$2 X1 = X2$$

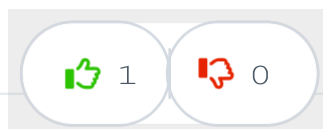
Convert it into an acceptable constraint with a constant on the right side and the variables on the left side as follows:

$$2 X1 - X2 = 0$$

5. State the nonnegativity and given requirements—With only two variables, this formal requirement in the formulation of an LP model is expressed as follows:

$$X1, X2 \geq 0$$

Because these variables are in dollars, they do not have to be integer values. (They can be any real or cardinal number.) The complete LP model formulation is given here:



7.4.2.2 Solution for the LP Marketing/Planning Model

Appendix B explains that both Excel and LINGO software can be used to run the LP model and solve the budget allocation in this marketing/planning case study problem. For purposes of brevity, discussion will be limited to just LINGO. As will be presented in Appendix B, LINGO is a mathematical programming language and software system. It allows the fairly simple statement of the LP model to be entered into a single window and run to generate LP solutions.

LINGO opens with a blank window for entering whatever type of model is desired. After the LP model formulation is entered into the LINGO software, the resulting data entry information is presented in Figure 7.6.

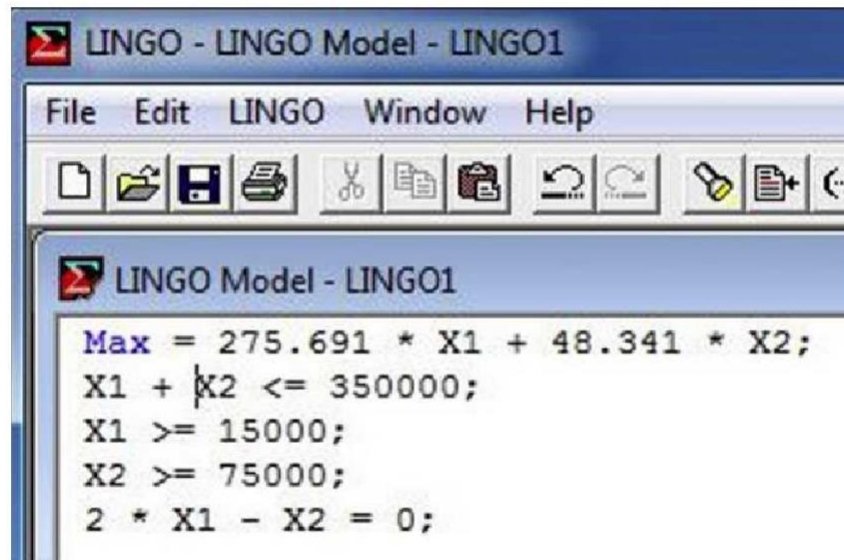


Figure 7.6 LINGO LP model entry requirements: marketing/planning case study

There are several minor differences in the model entry requirements over the usual LP model formulation. These differences are required to run a model in LINGO. These include (1) using the term "Max" instead of "Maximize," (2) dropping off "Subject to" and "and" in the model formulation, (3) placing an asterisk and a space between unknowns and constant values in the objective and constraint functions where multiplication is required, (4) ending each expression with a semicolon, and (5) omitting the nonnegativity requirements, which aren't necessary.



1



0

Now that the model is entered into LINGO, a single click on the SOLVE option in the bar at the top of the window generates a solution. The marketing budget allocation LP model solution is found in Figure 7.7.

The screenshot shows a window titled "Solution Report - LINGO1". The text inside the window reads: "Global optimal solution found. Objective value: 0.4344352E+08 Total solver iterations: 0". Below this, there are two tables. The first table lists variables X1 and X2 with their values and reduced costs. The second table lists rows 1 through 5 with their slack or surplus and dual prices.

Variable	Value	Reduced Cost
X1	116666.7	0.000000
X2	233333.3	0.000000

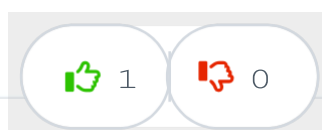
Row	Slack or Surplus	Dual Price
1	0.4344352E+08	1.000000
2	0.000000	124.1243
3	101666.7	0.000000
4	158333.3	0.000000
5	0.000000	75.78333

Figure 7.7 LINGO LP model solution: marketing/planning case study

As it turns out, the optimal distribution of the \$350,000 promotion budget is to allocate \$116,666.70 to radio commercials and \$233,333.30 to TV commercials. The resulting Z value, which in this model is the total predicted product sales in dollars, is 0.4344352E+08, or \$43,443,524. When we compare that future estimated month's product sales with the average current monthly product sales of \$16,717,200 presented in Figure 7.7, it does appear that the firm in this case study will optimally maximize future estimated monthly product sales if it allocates the budget accordingly (that is, if the multiple regression model estimates and the other parameters in the LP model hold accurate and true).

In summary, the prescriptive analytics analysis step brings the prior statistical analytic steps into an applied decision-making process where a potential business performance improvement is shown to better this organization's ability to use its resources more effectively. The management job of monitoring performance and checking to see that business performance is in fact improved is a needed final step in the BA analysis. Without proof that business performance is improved, it's unlikely that BA would continue to be used.

7.4.2.3 Final Comment on the Marketing/Planning Model



Although the LP solution methodology used to generate an allocation solution guarantees an optimal LP solution, it does not guarantee that the firm using this model's solution will achieve the results suggested in the analysis. Like any forecasting estimation process, the numbers are only predictions, not assurances of outcomes. The high levels of significance in the statistical analysis and the added use of other conformational statistics (R-Square, adjusted R-Square, ANOVA, and so on) in the model development provide some assurance of predictive validity. There are many other methods and approaches that could have been used in this case study. Learning how to use more statistical and decision science tools helps ensure a better solution in the final analysis.

Summary

This chapter discussed the prescriptive analytics step in the BA process. Specifically, this chapter revisited and briefly discussed methodologies suggested in BA certification exams. An illustration of nonlinear optimization was presented to demonstrate how the combination of software and mathematics can generate useful decision-making information. Finally, this chapter presented the third installment of a marketing/planning case study illustrating how prescriptive analytics can benefit the BA process.

We end this book with a final application of the BA process. Once again, several of the appendixes are designed to augment this chapter's content by including technical, mathematical, and statistical tools. For both a greater understanding of the methodologies discussed in this chapter and a basic review of statistical and other quantitative methods, a review of the appendixes and chapters is recommended.

Addendum

The differential calculus method for finding the minimum cost point on the quadratic function that follows involves a couple of steps. It finds the zero slope point on the cost function (the point at the bottom of the u-shaped curve where a line could be drawn that would have a zero slope). There are limitations to its use, and qualifying conditions are required to prove minimum or maximum positions on a curve. The quadratic model in the server problem follows:

$$Y_p = 35418 - 5589.432 X + 268.445 X^2 \text{ [Quadratic model]}$$

Step 1. Given the quadratic function above, take its first derivative.