SASURIE COLLEGE OF ENGINEERING

DEPARTMENT OF MASTER OF BUSINESS ADMINISTRATION

REGULATION 2021

I YEAR – II SEMESTER

– BA4206

BUSINESS ANALYTICS

BA4206

BUSINESS ANALYTICS

UNIT I INTRODUCTION TO BUSINESS ANALYTICS

Business Analytics - Terminologies, Process, Importance, Relationship with Organisational Decision Making, BA for CompetitiveAdvantage.

UNIT II MANAGING RESOURCES FOR BUSINESS ANALYTICS

Managing BA Personnel, Data and Technology. Organisational Structures aligning BA. Managing Information policy, data quality and change in BA.

UNIT III DESCRIPTIVE ANALYTICS

Introduction to Descriptive analytics - Visualising and Exploring Data - Descriptive Statistics - Sampling and Estimation - Probability Distribution for Descriptive Analytics - Analysis of Descriptive analytics

UNIT IV PREDICTIVE ANALYTICS

Introduction to Predictive analytics - Logic and Data Driven Models - Predictive Analysis Modeling and procedure - Data Mining for Predictive analytics. Analysis of Predictive analytics

UNIT V PRESCRITIVE ANALYTICS

Introduction to Prescriptive analytics - Prescriptive Modeling - Non Linear Optimisation - Demonstrating Business Performance Improvement.

Unit1–Introduction to Business Analytics

Whatisbusiness analytics?

Businessanalytics(BA)isasetofdisciplinesandtechnologiesforsolvingbusinessproblems using data analysis,statistical models andother quantitative methods. Itinvolves aniterative, methodical exploration of an organization's data, with an emphasis on statistical analysis, to drive decision-making.

Data-drivencompaniestreattheirdataasabusinessassetandactivelylookforwaystoturnit 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.

Howbusinessanalyticsworks

Beforeanydataanalysistakesplace,BAstartswithseveralfoundationalprocesses:

- Determinethebusinessgoalofthe analysis.
- Selectananalysismethodology.
- Getbusinessdatatosupporttheanalysis,oftenfromvarious systemsand sources.
- Cleanseandintegratedataintoasinglerepository,suchasadatawarehouseordata mart.

Initialanalysisistypicallyperformedonasmallersampledatasetofdata.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.

Deploymentofpredictivemodelsinvolvesastatisticalprocessknownasscoringanduses recordstypicallylocatedinadatabase.Scoreshelpenterprisesmakemoreinformed,real-   time decisions within applications and business processes.

BA also supports tactical decision-making in response to unforeseen events. Often the decision-makingisautomatedusingartificialintelligencetosupportreal-time responses.

Types of business analytics

Differenttypesofbusinessanalyticsincludethefollowing:

Descriptive nalytics,whichtrackskeyperformanceindicators(KPIs)tounderstandthe present state of a business;

**predictive analytics,**whichanalyzestrenddatatoassessthelikelihoodoffutureoutcomes; and

prescriptive analytics,whichusespastperformancetogeneraterecommendationsfor handlings
predictiveanalytics,whichanalyzestrenddatatoassessthelikelihoodoffutureoutcomes; and

prescriptive analytics,whichusespastperformancetogeneraterecommendationsfor handling similar situations in the future

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

3 types of data analytics
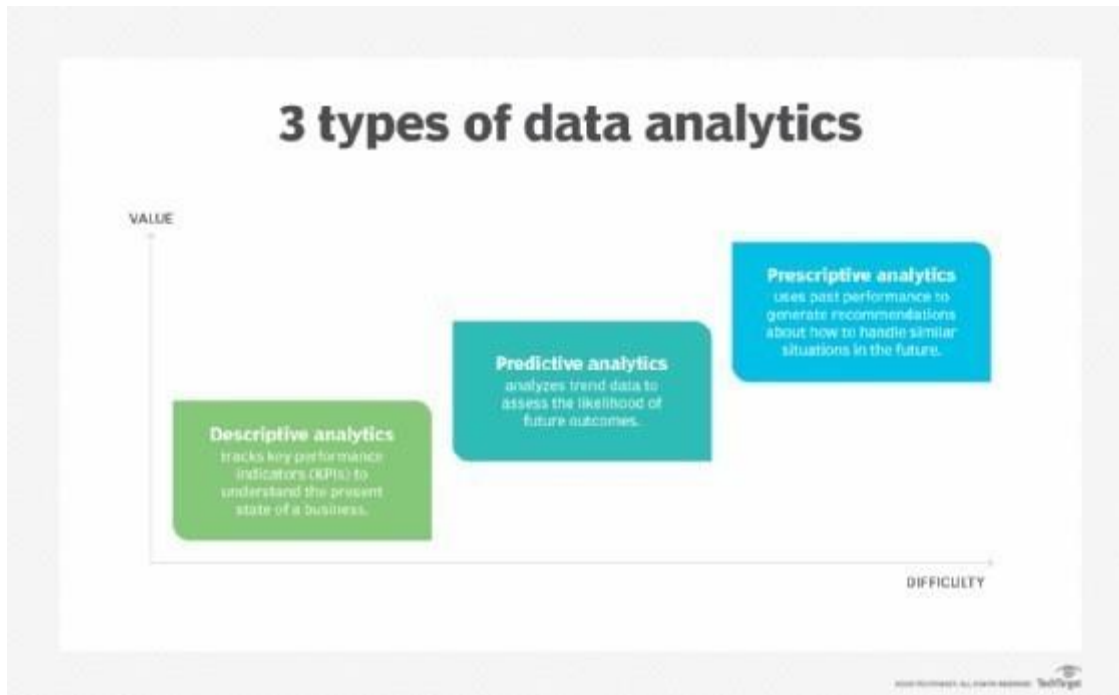
Businessanalyticsvs.business intelligence

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. BIanalyzes 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.

Businessanalyticsvs.data analytics

Data analytics is the analysis of data sets to draw conclusions about the information they contain.Dataanalyticsdoesnothavetobeusedinpursuitofbusinessgoalsorinsights.Itisa   broader practice that includes business analytics.

BAinvolves using dataanalytics tools in pursuitofbusinessinsights. However, becauseit'sa general term, data analytics is sometimes used interchangeably with business analytics.

Businessanalyticsvs.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.

Evenwhenadvancedstatisticalalgorithmsareappliedtodatasets,itdoesn'tnecessarilymean 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.

Commonchallengesofbusinessanalytics

Businesses might encounter both business analytics and business intelligencechallenges 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 processBAdatahasgrown.Somebusinesses,particularlysmallandmedium-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 largevolumesofunstructureddata.

Rolesandresponsibilitiesinbusiness analytics

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

- identifyingstrategicopportunities fromdata patterns;

- identifyingpotentialproblems facingthebusinessandsolutions;
- creatingabudgetandbusiness forecast;
- monitoringprogresswithbusiness initiatives;
- reportingprogressonbusinessobjectivesbackto stakeholders;
- understandingKPIs;and
- understandingregulatoryandreportingrequirements

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

TerminologiesinBusinessAnalytics

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 ora 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.

Businessanalytics–Thistermreferstotheapplication ofanalytics specifically in thesphere of business. It includes subsets like –

- Marketing analytics

- Risk analytics

- Fraudanalytics

- CRManalytics

- Loyalty analytics

- Operationsanalytics

- HRanalytics

Industrieswhichrelyextensively onanalyticsinclude–

- FinancialServices(Banks,CreditCards,Loans,Insuranceetc.)

- Retail

- Telecom

- Healthcare

- Consumergoods

- 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 lotof 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 –Theterm "Reporting" isperhaps themostunglamorous ofall terms in theworld of analytics. Yet it is also one of the most widely used practices within the field. All businessesusereporting to aiddecisionmaking.Whileitisnot"Advancedanalytics" oreven

"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 andinvolves 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

moststatisticaltechniqueswas,untilrecently,basedoneleganttheoryandanalyticalmethods 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,Hadoopetc.

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

Ex:Numberofclicks,probability tobuybasedonsearchtimesofaparticularwordetc.

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

Ex:Facebookdataanalysis,twitteranalysisetc.

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

Ex:Offersbasedoncustomerlocation,insuranceriskcalculationsbasedonproximityto 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 wouldliketoimproveonortheproblem itwantssolved.Sometimes, thegoalisbrokendown

into smaller goals. Relevant data needed to solve these business goals are decidedupon by the business stakeholders, business users with the domain knowledge and thebusiness 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.

Step2.Explore thedata

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 areplotted as they areableto indicate any patterns or outliers. Theremoval 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.

Step3.Analysethe 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 testedusing 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.

Step4.Predictwhat islikelytohappen

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(findthebest 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 givenconstraintsandlimitations.Theanalystwillselecttheoptimalsolutionandmodel

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.

Step6.Makea decision andmeasurethe 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.

Step7.Updatethesystem withtheresultsofthe 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?','howdidthetreatmentgroupcomparewiththecontrolgroup?'and'whatwasthe 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.

Importanceofbusinessanalytics

- Organizations employ Business analytics so they can make data-driven decisions. Business analytics gives business an excellent overview and insight onhow 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 flowof 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.

EvolutionofBusinessAnalytics

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, whichwas 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, managementscienceevolvedintobusinessintelligence,decisionsupportsystemsandintoPC software.

ScopeofBusinessAnalytics

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 companyand 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.

RelationshipofBAProcessandOrganizationDecision-MakingProcess

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.Asdepicted in below shownFigure, thebusinessanalytic process has an inherent relationship to the steps in typical organization decision-making processes.

Figure:Comparisonofbusinessanalyticsandorganizationdecision-making processes
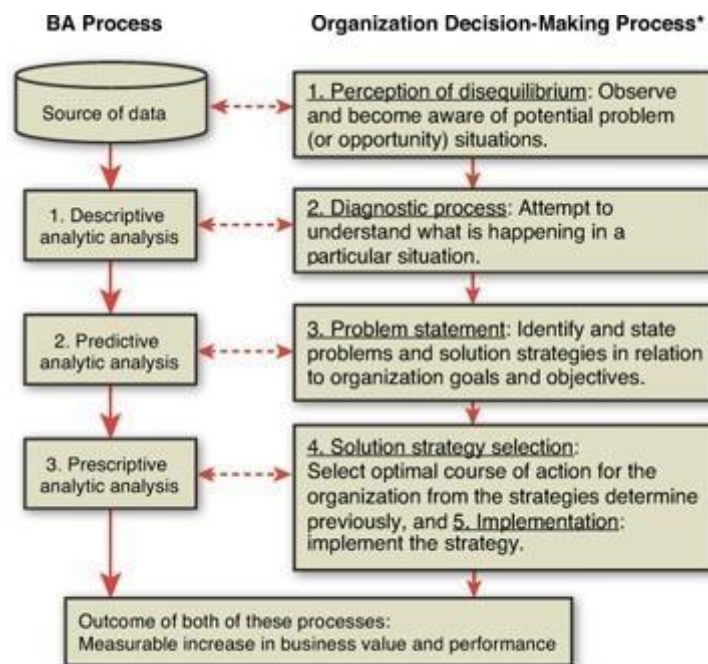
Source: Adapted from Figure 1 in Elbing (1970), pp. 12–13.

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 stepistorecognizethatdatabasesmaycontaininformationthatcouldbothsolveproblems

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.

TheCompetitiveAdvantageofBusiness 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 tothe increasingly thinnermarginfordecisionerror.Itistheretoprovideinsights,predictthefuture of the business and inferences from the treasure chest of raw transactional data, that isinternal 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, economiesof 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 contributesto 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 acompetitive advantage willbe 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 thenextwavefororganizations to successfully compete. This willresult 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.

Analyticsgivescompaniesaninsightintotheircustomers'behaviourandneeds.Italsomakes it possible for a company to understand the public opinion of its brand, to follow the resultsof 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 ora 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 that 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

Unit2 –Managing Resources for Business Analytics

BusinessAnalyticsPersonnel

One way to identify personnel needed for BA staff is to examine what is required for certification in BA by organizations that provide BA services. INFORMS (www.informs.org/Certification-Continuing-Ed/Analytics-Certification), a major academic and professional organization, announced the startup of a Certified Analytic Professional (CAP) program in 2013. Another more established organization, Cognizure (www.cognizure.com/index.aspx), offers a variety of service products, including business analytics services. It offers a general certification Business Analytics Professional (BAP) exam that measures existing skill sets in BA staff and identifies areas needing improvement (www.cognizure.com/cert/bap.aspx). This is a tool to validate technical proficiency,expertise, and professional standards in BA. The certification consists of three examscovering the content areas listed in Table 3.1.

Regardless of the source of data, it has to be put into a structure that makes it usable by BA personnel. We will discuss data warehousing in the next section, but here we focus on a couple ofdata issues that are critical tothe usability of anydatabase ordata file. Those issues are data quality and data privacy. Data quality can be defined as data that serves the purpose for which it is collected. It means different thingsfor different applications, but there are some commonalities of high-quality data. These qualities usually include accurately representing reality, measuring what it is supposed to measure, being timeless, and having completeness. When data is of high quality, it helps ensure competitiveness, aids customer service, and improves profitability. When data is of poor quality, it can provide information that is contradictory, leading to misguided decision-making. For example, having missing data in files can prohibit some forms' statistical modeling, and incorrect codingof information can completely render databases useless. Data quality requires effort on the part

Exam	Topic	Specific Content Areas Covered	Examples
I	Statistical Methods	1. Visualizing and Exploring Data 2. Descriptive Statistics 3. Probability Distributions 4. Sampling and Estimation 5. Statistical Inference 6. Regression Analysis 7. Predictive Modeling and Analysis	1. Graphs and charts 2. Mean, median, mode 3. Normal distribution 4. Confidence intervals 5. Hypothesis testing 6. Multiple regression 7. Curve fitting of models and functions to raw data
II	Operations Research Methods	1. Linear Optimization 2. Integer Optimization 3. Nonlinear Optimization 4. Simulation 5. Decision Analysis 6. Forecasting	1. Linear programming 2. Integer programming 3. Quadratic programming 4. Monte Carlo method 5. Expected value analysis 6. Exponential smoothing
III	Case Studies	Practical knowledge of real-world situations	Application of the preceding methods to solve a real-world problem

Table3.1Cognizure Organization Certification Exam Content Areas*

Mostofthe content areas inTable 3.1willbe discussedand illustrated in subsequent chapters and appendixes. The three exams required in the Cognizure certification program can easily be understood in the context of the three steps of the BA process (descriptive, predictive, and prescriptive) discussed in previous chapters. The topics in Figure 3.1 of the certification program are applicable to the three major steps in the BA process. The basic statistical tools apply to the descriptive analytics step, the more advanced statistical tools apply to the predictive analytics step, and the operations research tools apply to the prescriptive analytics step. Some of the tools can be applied to both the descriptive and the predictive steps. Likewise, tools like simulation can be applied to answer questions in both the predictive and the prescriptive steps, depending on how they're used. At the conjunction of all the tools isthe reality of case studies. The use of casestudies is designed to provide practical experience, whereby all tools are employed to answer important questions or seek opportunities.
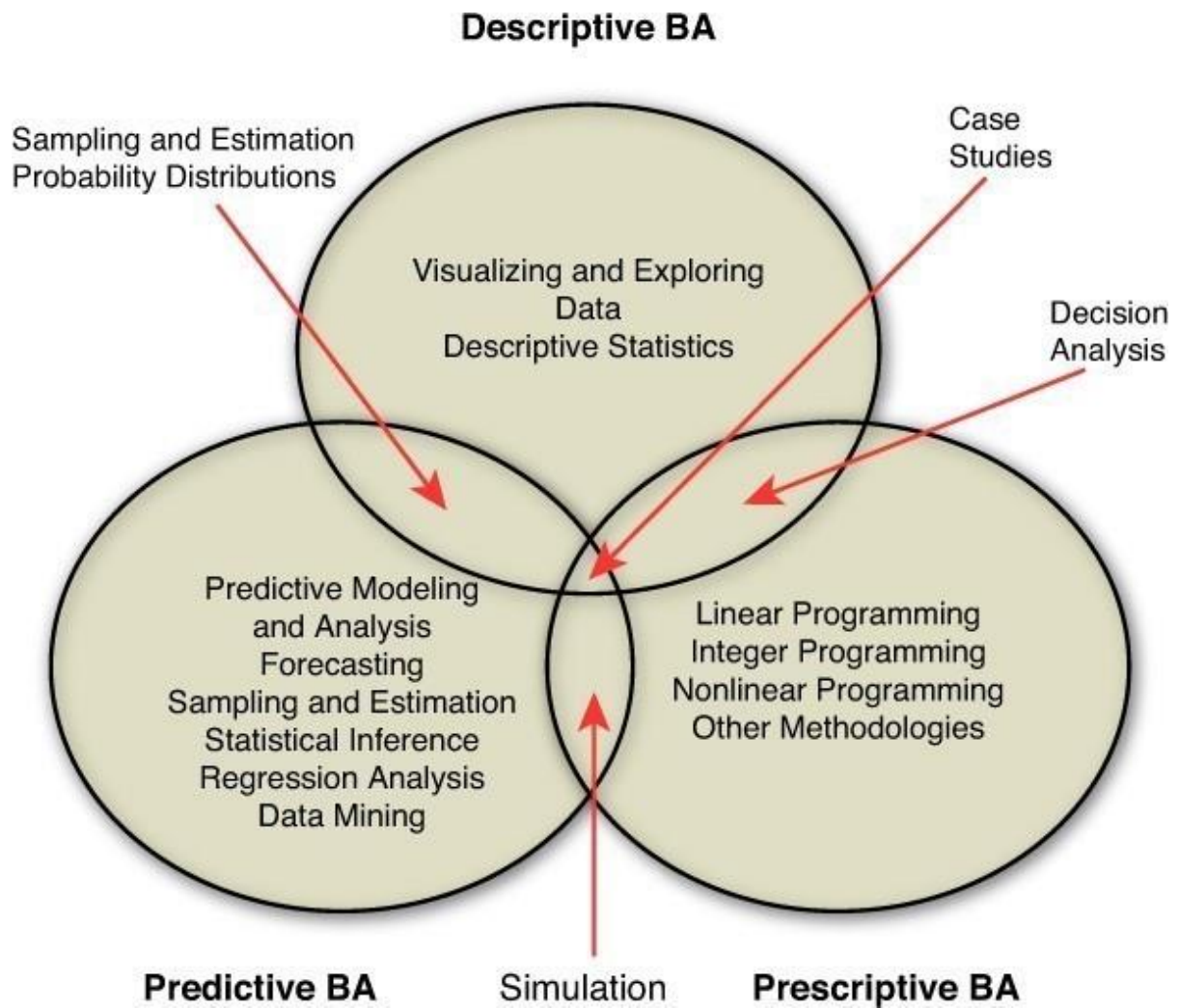
Figure3.1Certificationcontentareasandtheirrelationship tothestepsinBA

Other organizations also offer specialized certification programs. These certifications include other areas of knowledge and skills beyond just analytic tools. IBM, for example, offers a variety of specialized BA certifications (www-03.ibm.com/certify/certs/ba_index.shtml). Although these include certifications in several dozen statistical, information systems, and analytic methodologies related to BA, they also include specialized skill sets related to BA personnel (administrators, designers, developers, solution experts, and specialists), as presented in Table 3.2.

BA Personnel Specialty	Description
Administrators	Within the context of the IBM BA and business intelligence (BI) software platforms, administrators manage servers (their load balancing, installation, and configurations). They manage reports from computer portals, manage dispatchers, and perform troubleshooting for technology. They are also in charge of user authorization and authentication for security.
Designers	As members of a team, designers are responsible for building reports using relational data models, as well as enhancing, customizing, and managing professional reports.
Developers	As members of a team, developers require skills that are closely tied to the BA process and involve the application of analytics, data warehousing, model building, use of operations research and statistical methodologies, and real-time monitoring of data flow to users.
Solution Experts	As members of a team, solution experts analyze, plan, design, deploy, and operate BA applications using an appropriate methodology and development approach. This requires knowledge in many differing BA software applications, including statistical, information systems, and operations research methods.
Technical Specialists	As members of a team, they are responsible for the installation and configuration of BA and BI applications.

Table3.2Types of BA Personnel*

With the variety of positions and roles participants play in the BA process, this leads to the question of what skill sets or competencies are needed to function in BA. In a general sense, BA positions require competencies in business, analytic, and information systems skills. As listed in Table 3.3, business skills involve basic management of people and processes. BA personnel must communicate with BA staffers within the organization (the BA team members) and the other functional areas within a firm (BA customers and users) to be useful. Because they serve a variety of functional areas within a firm, BA personnel need to possess customer service skills so they can interact with the firm's personnel and understand the nature of the problems they seek to solve. BA personnel also need to sell their services to users inside the firm. In addition, some must lead a BA team or department, which requires considerable interpersonal management leadership skills and abilities.

Type of Skill or Competency	Description of Possible Roles
Business	• Leadership • People-related management and communication skills • Manage BA projects (prioritize, schedule, and so on) • Manage BA processes (rules, procedures governing acceptance and use) • Determine project requirements • Train BA personnel to develop competencies
Analytic	• Know how to use statistical methodologies • Know how to use operations research methodologies • Know how to use data mining for quantitative data and text analytics for unstructured data
Information system	• Maintain and use computer portals • Identify and extract data • Maintain quality data

Table3.3 Select Typesof BA Personnel Skills or Competency Requirements

Fundamental to BA is an understanding of analytic methodologies listed in Table 3.1 and others not listed. In addition to any tool sets, there is a need to know how they are integrated into the BA process to leverage data (structured or unstructured) and obtain information that customers who will be guided by the analytics desire.

In summary, people who undertake a career in BA are expected to know how to interact with people and utilize the necessary analytic tools to leverage data into useful information that can be processed, stored, and shared in information systems in a way that guides a firm to higher levels of business performance.

Business Analytics Data

Structured and unstructured data (introduced in Chapter 2, "Why Is Business Analytics Important?") is needed to generate analytics. As a beginning for organizing data into an understandable framework, statisticians usually categorize data into meaning groups.

Categorizing Data

There are many ways to categorize business analytics data. Data is commonly categorized by either internal or external sources (Bartlett, 2013, pp. 238–239). Typical examples of internal datasourcesincludethosepresentedinTable3.4.Whenfirmstrytosolveinternal production

or service operations problems, internally sourced data may be all that is needed. Typical external sources of data (see Table 3.5) are numerous and provide great diversity and unique challenges for BA to process. Data can be measured quantitatively (for example, sales dollars) or qualitatively by preference surveys (for example, products compared based on consumers preferring one product over another) or by the amount of consumer discussion (chatter) on the Web regarding the pluses and minuses of competing products.

Type of Internal Data	Description
Billing and Reminder Systems	Billing systems and reminder systems print bills and monitor customer payment information on valued-based customer segments.
Business	Industry codes, accounting information, personnel information, and so on are routinely collected in the course of business.
Customer	Names, addresses, returns, special contracts, segmentations, and so on are obtained when customers sign for or pay for products or services.
Customer Relationship Management Systems	*Customer relationship management* (CRM) systems collect and provide data on customer history, behavior on matters like complaints, the end of a relationship with a firm, and so on.
Human Resources	Information about employees, salaries, competencies, and so on is recorded by routine efforts over the history of employment.
Information from Enterprise Resource Planning Systems	*Enterprise resource planning* (ERP) systems are used to communicate internal business transactions to provide a direct feed of information on management issues and concerns, as well as other operations activities required to produce and sell products.
Product	Information is collected from procurement through post sales to monitor profitability, durability, and quality.
Production	Information that can be used to optimize production, inventory control, and supply chain delivery of the product to the customers is collected during the production processes.
Questionnaires	Information on customer behavior is obtained by customer questionnaires to measure customer service and product quality, among other things.
Web Logs	Information is collected on the firm's website usage via cookies and other means to learn customer navigation behavior and product interests.

Table 3.4 Typical Internal Sources of Data on Which Business Analytics Can Be Based

Type of External Data	Measured By
Customer Satisfaction	• Revenue, profit • Market share, sales • Product/service survey data • Loyalty • Brand awareness • Average spend per customer
Customer Demographics	• Geographic location (distance from market) • Income level • Market size
Competition	• Market share • Competitor profitability • Advertising/promotion efforts • Preference surveys • Web chatter on products
Economic	• Population statistics • Income distribution statistics

Table 3.5 Typical External Sources of Data on Which Business Analytics Can Be Based

A major portion of the external data sources are found in the literature. For example, the US Census and the International Monetary Fund (IMF) are useful data sources at the macroeconomic level for model building. Likewise, audience and survey data sources might include Nielsen (www.nielsen.com/us/en.html) for psychographic or demographic data, financial data from Equifax (www.equifax.com), Dun & Bradstreet (www.dnb.com), and so forth.

Data Issues

Regardless of the source of data, it has to be put into a structure that makes it usable by BA personnel. We will discuss data warehousing in the next section, but here we focus on a couple of data issues that are critical to the usability of any database or data file. Those issues are data quality and data privacy. Data quality can be defined as data that serves the purpose for which it is collected. It means different things for different applications, but there are some commonalities of high-quality data. These qualities usually include accurately representing reality, measuring what it is supposed to measure, being timeless, and having completeness. When data is of high quality, it helps ensure competitiveness, aids customer service, and improves profitability. When data is of poor quality, it can provide information that is contradictory, leading to misguided decision-making. For example, having missing data in files can prohibit some forms' statistical modeling, and incorrect coding of information can completely render databases useless. Data quality requires effort on the part

of data managers to cleanse data of erroneous information and repair or replace missing data. We will discuss some of these quality data measures in later chapters.

Data privacy refers to the protection of shared data such that access is permitted only to those users for whom it is intended. It is a security issue that requires balancing the need to know with the risks of sharing too much. There are many risks in leaving unrestricted access to a company's database. For example, competitors can steal a firm's customers by accessing addresses. Data leaks on product quality failures can damage brand image, and customers can become distrustful of a firm that shares information given in confidence. To avoid these issues, a firm needs to abide by the current legislation regarding customer privacy and develop a program devoted to data privacy.

Collecting and retrieving data and computing analytics requires the use of computers and information technology. A large part of what BA personnel do is related to managing information systems to collect, process, store, and retrieve data from various sources.

BusinessAnalyticsTechnology

Firms need an information technology (IT) infrastructure that supports personnel in the conduct of their daily business operations. The general requirements for such a system are stated in Table 3.6. These types of technology are elemental needs for business analytics operations.

Type of Technology	Description
Computer Hardware	This is physical equipment used for input, processing, and output activities in an information system. Hardware can include computers of various sizes; various input, output, and storage devices; and telecommunications devices that link computers, including mobile handheld devices.
Computer Software	These are the preprogrammed instructions that control and coordinate the computer hardware components in the information system. They include system-wide software like ERP and smaller *apps* (computer software applications) for mobile devices.
Networking and Telecommunications Technology	Physical devices and software link the various pieces of hardware and transfer data from one physical location to another. They include the computers and communications equipment connected in networks for sharing voice, data, images, sound, and video. They also include the Internet, *intranets* (internal corporate networks based on Internet technology with limited access to employees within the firm), and *extranets* (private intranets extended to authorized users outside the organization).
Data Management Technology	Software governs the organization of data on physical storage media. It includes database management systems, data warehouses, data marts, and online analytical processing, as well as data, text, and Web mining technologies.

Table3.6GeneralInformation Technology(IT) Infrastructure

Of particular importance for BA is the data management technologies listed in Table 3.6. Database management system (DBMS) is a data management technology software that permits firms to centralize data, manage it efficiently, and provide access to stored data by application programs. DBMS usually serves as an interface between applicationprograms and the physical data files of structured data. DBMS makes the task of understanding where and how the data is actually stored more efficient. In addition, other DBMS systems can handle unstructured data. For example, object-oriented DBMS systems are able to store and retrieve unstructured data,likedrawings,images,photographs,andvoicedata.Thesetypesof technology are necessary to handle the load of big data that most firms currently collect.

DBMS includes capabilities and tools for organizing, managing, and accessing data in databases. Four of the more important capabilities are its data definition language, data dictionary, database encyclopedia, and data manipulation language. DBMS has a data definition capability to specify the structure of content in a database. This is used to create database tables and characteristics used in fields to identify content. These tables and characteristics are critical success factors for search efforts as the database grows. These characteristics are documented in the data dictionary (an automated or manual file that stores the size, descriptions, format, and other properties needed to characterize data). The database encyclopedia is a table of contents listing a firm's current data inventory and the data filesthat can be built or purchased. The typical content of the database encyclopedia is presented

in Table 3.7. Of particular importance for BA are the data manipulation language tools included in DMBS. These tools are used to search databases for specific information. An exampleisstructurequerylanguage(SQL),whichallowsuserstofindspecificdatathrougha session of queries and responses in a database.

Database Content Item	Description
Purpose	Why the database exists, including any additional reports or analyses used in leveraging the data.
Time	Window of time when the data is collected or will be useful.
Source	Internal (auditing, accounting, and so on) and external (customers, and so on) sources.
Schematics	Diagrams illustrating the connections between tables and other data files.
Cost	Expense of collecting data, including purchasing prices.
Availability of Data	Window of time when the data may be available.
Collection Techniques	Methods of collection, including observation, data mining, census, and focus groups.
Collection Tools	Web, customer generated, e-survey, and so on.

Table3.7DatabaseEncyclopedia Content

Data warehouses are databases that store current and historical data of potential interest to decision makers. What a data warehouse does is make data available to anyone who needs access to it. In a data warehouse, the data is prohibited from being altered. Data warehouses also provide a set of query tools, analytical tools, and graphical reporting facilities. Some firms use intranet portals to make data warehouse information widely available throughout a firm.

Data marts are focused subsets or smaller groupings within a data warehouse. Firms often build enterprise-wide data warehouses in which a central data warehouse serves the entire organization and smaller, decentralized data warehouses (called data marts) are focused on a limited portion of the organization's data that is placed in a separate database for a specific population of users. For example, a firm might develop a smaller database on just product quality to focus efforts on quality customer and product issues. A data mart canbe constructed more quickly and at lower cost than enterprise-wide data warehouses to concentrate effort in areas of greatest concern.

Oncedatahasbeencapturedandplacedintodatabasemanagementsystems,itisavailablefor analysis with BAtools, including online analytical processing, as well as data, text, and Web

mining technologies. Online analytical processing (OLAP) is software that allows users to view data in multiple dimensions. For example, employees can be viewed in terms of their age, sex, geographic location, and so on. OLAP would allow identification of the number of employees who are age 35, male, and in the western region of a country. OLAP allows users to obtain online answers to ad hoc questions quickly, even when the data is stored in very large databases.

Data mining is the application of a software discovery-driven process that provides insights into business data by finding hidden patterns and relationships in big data or large databases and inferring rules from them to predict future behavior. The observed patterns and rules guide decision-making. They can also act to forecast the impact of those decisions. It is an ideal predictive analytics tool used in the BA process mentioned in Chapter 1, "What Is Business Analytics?" The kinds of information obtained by data mining include those in Table 3.8.

Types of Information	Description	Example
Associations	Occurrences linked to a single event.	An ad in a newspaper is associated with greater sales.
Classification	Recognizes patterns that describe the group an item belongs to by examining previous classified existing items and by inferring a set of rules that guide the classification process.	Identify customers who are likely to need more customer service than those who need less.
Clustering	Similar to classification when no groups have yet been defined, helps to discover different groupings within data.	Identify groups that can be differentiated within a single, large group of customers. An example would be distinguishing tea drinkers from those who would drink other beverages offered in flight on an airline.
Forecasting	Predicts values that can identify patterns in customer behavior.	Estimate the value of a future stream of dollar sales from a typical customer.
Sequence	Links events over time.	Identify a link between a person who buys a new house and subsequently will buy a new car within 90 days.

Table 3.8 Types of Information Obtainable with Data Mining Technology

Text mining (mentioned in Chapter 2) is a software application used to extract key elements from unstructured data sets, discover patterns and relationships in the text materials, and summarize the information. Given that the majority of the information stored in businesses is in the form of unstructured data (e-mails, pictures, memos, transcripts, survey responses, business receipts, and so on), the need to explore and find useful information will require increased use of text mining tools in the future.

Web mining seeks to find patterns, trends, and insights into customer behavior from users of the Web. Marketers, for example, use BA services like Google Trends (www.google.com/trends/) and Google Insights for Search (http://google.about.com/od/i/g/google-insights-for-search.htm) to track the popularity of various words and phrases to learn what consumersare interested in and what they are buying.

In addition to the general software applications discussed earlier, there are focused software applications used every day by BA analysts in conducting the three steps of the BA process (see Chapter 1). These include Microsoft Excel® spreadsheet applications, SAS applications, and SPSS applications. Microsoft Excel (www.microsoft.com/) spreadsheet systems have add-in applications specifically used for BA analysis. These add-in applications broaden the use of Excel into areas of BA. Analysis ToolPak is an Excel add-in that contains a variety of statistical tools (for example, graphics and multiple regression) for the descriptive and predictive BA process steps. Another Excel add-in, Solver, contains operations research optimization tools (for example, linear programming) used in the prescriptive step of the BA process.

SAS® Analytics Pro (www.sas.com/) software provides a desktop statistical toolset allowing users to access, manipulate, analyze, and present information in visual formats. It permits users to access data from nearly any source and transform it into meaningful, usable information presented in visuals that allow decision makers to gain quick understanding of critical issues within the data. It is designed for use by analysts, researchers, statisticians, engineers, and scientists who need to explore, examine, and present data in an easily understandable way and distribute findings in a variety of formats. It is a statistical package chiefly useful in the descriptive and predictive steps of the BA process.

IBM's SPSS software (www-01.ibm.com/software/analytics/spss/) offers users a wide range of statistical and decision-making tools. These tools include methodologies for data collection, statistical manipulation, modeling trends in structured and unstructured data, and optimizing analytics. Depending on the statistical packages acquired, the software can cover all three steps in the BA process.

Other software applications exist to cover the prescriptive step of the BA process. One that will be used in this book is LINGO® by Lindo Systems (www.lindo.com). LINGO is a comprehensivetooldesignedtomakebuildingandsolvingoptimizationmodelsfaster,easier, and more efficient. LINGO provides a completely integrated package that includes an

understandable language for expressing optimization models, a full-featured environment for building and editing problems, and a set of built-in solvers to handle optimization modelingin linear, nonlinear, quadratic, stochastic, and integer programming models.

Insummary, thetechnology needed tosupportaBAprograminanyorganization willentail a general information system architecture, including database management systems and progress in greater specificity down to the software that BA analysts need to compute their unique contributions to the organization. Organizations with greater BA requirements will havesubstantially moretechnology tosupportBA efforts,butall firms that seekto use BA as a strategy for competitive advantage will need a substantial investment intechnology, because BA is a technology-dependent undertaking.

HowDoWeAlignResourcestoSupportBusinessAnalyticswithinanOrganization?

Chapter objectives:

• Explainwhy a centralizedbusiness analytics (BA) organizationstructurehas advantages over other structures.

• Describethedifferences between BAprograms, projects, andteams andhow they alignBA resources in firms.

• DescribereasonswhyBAinitiatives fail.

• Describetypical BAteamrolesandreasonsfortheir failures.

• Explainwhyestablishinganinformationpolicyis important.

• Explaintheadvantagesanddisadvantagesofoutsourcing BA.

• Describehowdatacanbe scrubbed.

• Explainwhatchangemanagementinvolvesandwhatits relationshipisto BA.

OrganizationStructuresAligningBusinessAnalytics

According to Isson and Harriott (2013, p. 124), to successfully implement business analytics (BA) within organizations, the BA in whatever organizational form it takes must be fully integrated throughoutafirm. Thisrequires BAresourcestobealigned in awaythatpermits a view of customer information within and across all departments, access to customer information from multiple sources (internal and external to the organization), access to historical analytics from a central repository, and alignment of technology resources so they're accountable for analytic success. The commonality of theserequirements is thedesire for an alignment that maximizes the flow of information into and through the BAoperation,

which in turn processes and shares information to desired users throughout the organization. Accomplishing this information flow objective requires consideration of differing organizational structures and managerial issues that help align BA resources to best serve an organization.

Organization Structures

As mentioned in Chapter 2, "Why Is Business Analytics Important?" most organizations are hierarchical, with senior managers making the strategic planning decisions, middle-level managers making tactical planning decisions, and lower-level managers making operational planning decisions. Within the hierarchy, other organizational structures exist to support the development and existence of groupings of resources like those needed for BA. These additional structures include programs, projects, and teams. A program in this context is the process that seeks to create an outcome and usually involves managing several related projects withtheintention ofimproving organizational performance. Aprogram can alsobea large project. A project tends to deliver outcomes and can be defined as having temporary rather than permanent social systems within or across organizations to accomplish particular and clearly defined tasks, usually under time constraints. Projects are often composed of teams. A team consists of a group of people with skills to achieve a common purpose. Teams are especially appropriate for conducting complex tasks that have many interdependent subtasks.

The relationship of programs, projects, and teams with a business hierarchy is presented in Figure 4.1. Within this hierarchy, the organization's senior managers establish a BA program initiative to mandate the creation of a BA grouping within the firm as a strategic goal. A BA program does not always have an end-time limit. Middle-level managers reorganize or break down the strategic BA program goals into doable BA project initiatives to be undertaken in a fixed period of time. Some firms have only one project (establish a BA grouping) and others, depending on the organization structure, have multiple BA projects requiring the creation of multiple BA groupings. Projects usually have an end-time date in which to judge the successfulness of the project. The projects in some cases are further reorganized into smaller assignments, called BA team initiatives, to operationalize the broader strategy of the BA program. BA teams may have a long-standing time limit (for example, to exist as the main source ofanalytics foran entire organization) orhaveafixed period(forexample, toworkon a specific product quality problem and then end).
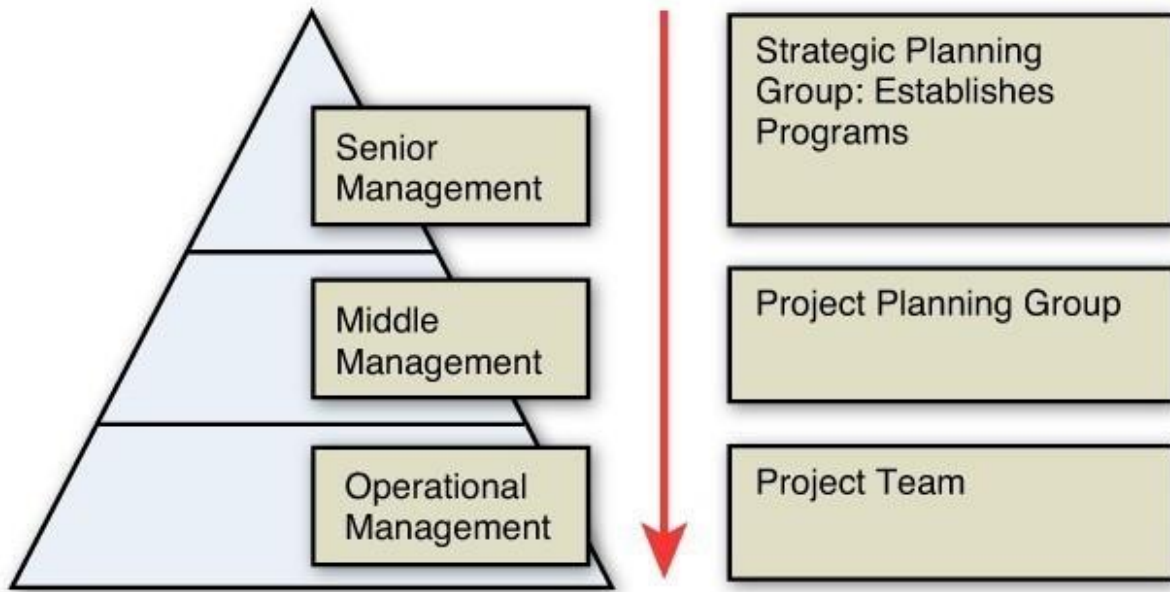
Figure4.1Hierarchal relationshipsprogram,project,andteamplanning

In summary, one way to look at the alignment of BA resources is to view it as a progression of assigned planning tasks from a BA program, to BA projects, and eventually to BA teams for implementation. As shown in Figure 4.1, this hierarchical relationship is a way toexamine how firms align planning and decision-making workload to fit strategic needs and requirements.

BAorganizationstructuresusuallybeginwithaninitiativethatrecognizestheneedtouseand develop some kind of program in analytics. Fortunately, most firms today recognizethis need. The question then becomes how to match the firm's needs within the organization to achieve its strategic, tactical, and operations objectives within resource limitations. Planning the BA resource allocation within the organizational structure of a firm is a starting place for the alignment of BA to best serve a firm's needs.

Aligning the BA resources requires a determination of the amount of resources a firm wants to invest. The outcome of the resource investment might identify only one individual to compute analytics for a firm. Because of the varied skill sets in informationsystems, statistics, and operations research methods, a more common beginning for a BA initiative is the creation of a BA team organization structure possessing a variety of analytical and management skills. (We will discuss BA teams in Section 4.1.2.) Another way of aligningBA resources within an organization is using a project structure. Most firms undertake projects, and some firms actually use a project structure for their entire organization. For example, consulting firms might view each client as a project (or product) and align their resources

around the particular needs of that client. A project structure often necessitates multiple BA teams to dealwith awider variety of analytic needs. Even larger investments in BA resources might be required by firms that decide to establish a whole BA department containing all the BA resources for a particular organization. Although some firms create BA departments, the departments don't have to be large. Whatever the organization structure that is used, the role of BA is a staff (not line management) role in their advisory and consulting mission for the firm.

In general, there are different ways to structure an organization to align its BA resources to serve strategic plans. In organizations in which functional departments are structured on a strict hierarchy, separate BA departments or teams have to be allocated to each functional area, as presented in Figure 4.2. This functional organization structure may have the benefitof stricter functional control by the VPs of an organization and greater efficiency in focusing on just the analytics within each specialized area. On the other hand, this structure does not promote the cross-department access that is suggested as a critical success factor for the implementation of a BA program.
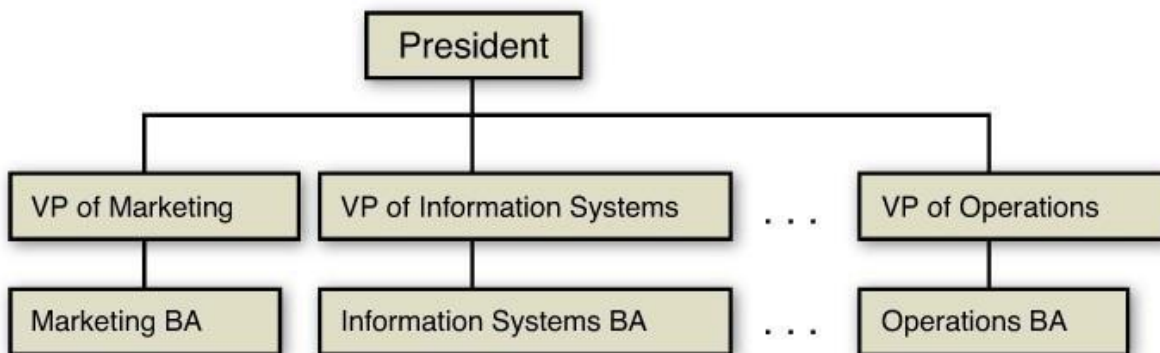
Figure4.2Functionalorganizationstructurewith BA

TheneedsofeachfirmforBAsometimesdictatepositioningBAwithinexistingorganization functional areas. Clearly, many alternative structures can house a BA grouping. For example, because BA provides information to users, BA could be included in the functional area of management information systems, with the chief information officer (CIO) acting as both the director of information systems (which includes database management) and the leader of the BA grouping.

An alternative organizational structure commonly found in large organizations aligns resources by project or product and is called a matrix organization. As illustrated in Figure 4.3,thisstructureallowsthe VPssomeindirectcontrolovertheirrelatedspecialists,which

would include the BA specialists but also allows direct control by the project or product manager. This, similar to the functional organizational structure, does not promote the cross-department access suggested for a successful implementation of a BA program.
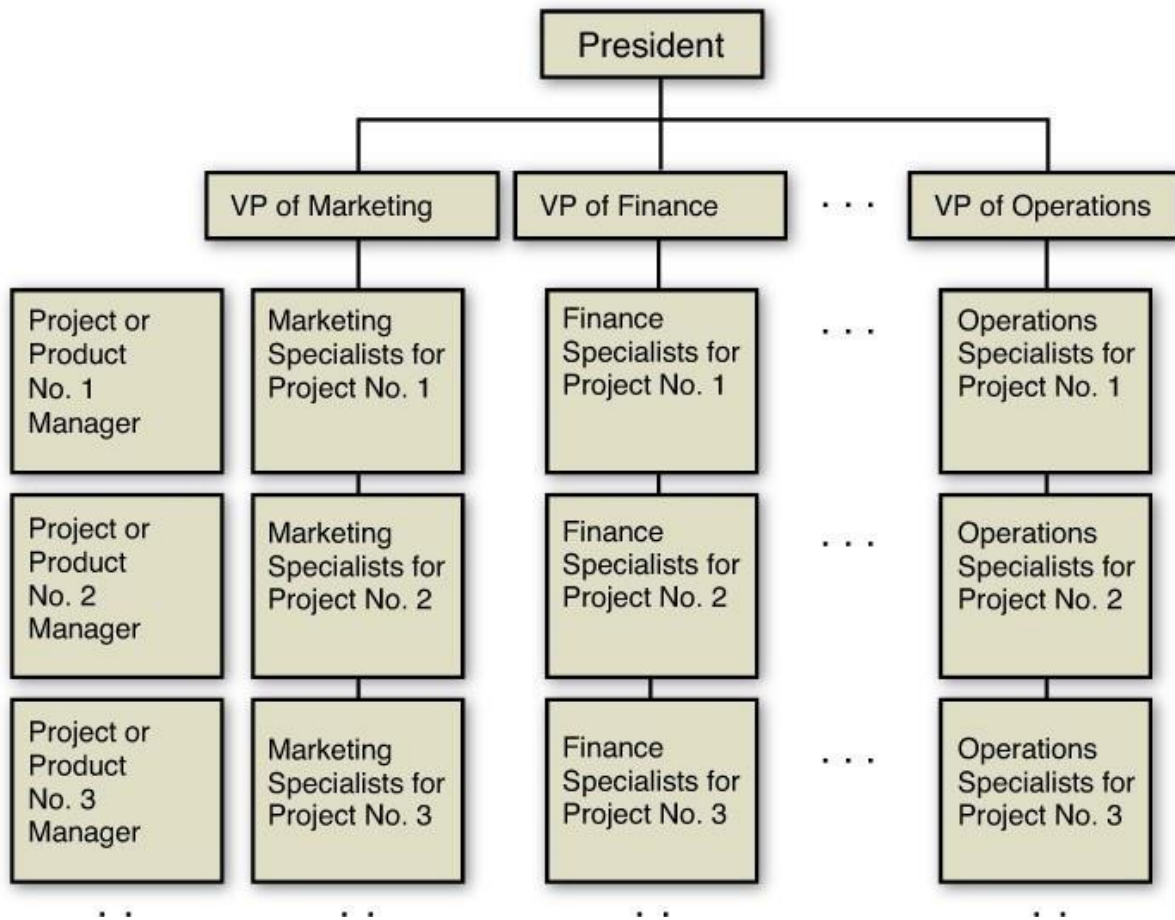
Figure4.3Matrixorganizationstructure

Theliteraturesuggeststhattheorganizational structurethatbestalignsBAresourcesisonein which a department, project, or team is formed in a staff structure where access to and from theBAgrouping ofresources permits access to allareaswithin afirm, as illustrated in Figure 4.4 (Laursen and Thorlund, 2010, pp. 191–192; Bartlett, 2013, pp. 109–111; Stubbs, 2011, p. 68).Thedashed lineindicates astaff(notlinemanagement) relationship. Thiscentralized BA organization structure minimizes investment costs by avoiding duplications found in both the functional and the matrix styles of organization structures. At the same time, it maximizes information flow between and across functional areas in the organization. This is a logical structure for a BA group in its advisory role to the organization. Bartlett (2013, pp. 109–110) suggestsotheradvantagesofacentralizedstructureliketheoneinFigure4.4.Theseincludea reduction in the filtering of information traveling upward through the organization, insulation from political interests, breakdown of the siloed functional area communication barriers, a morecentralplatformforreviewingimportantanalysesthatrequireabroaderfieldof

specialists, analytics-based groupdecision-making efforts,separation of the linemanagement leadership from potential clients (for example, the VP of marketing would not necessarily come between the BA group working on customer service issues for a department within marketing), and better connectivity between BA and all personnel within the area of problem solving.
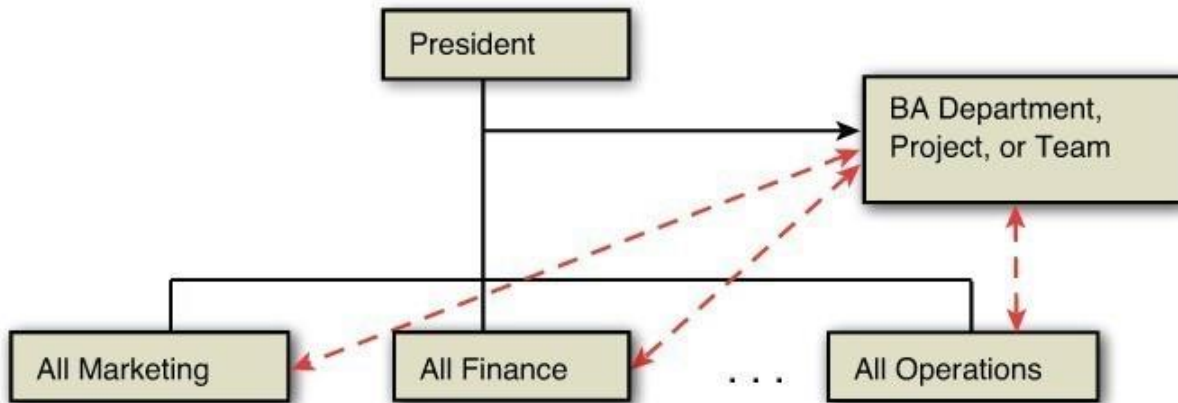
Figure4.4CentralizedBAdepartment,project,orteamorganizationstructure

Giventheadvocacyandlogicrecommendingacentralized   BAgrouping,therearereasonsfor allBAgroupingstobecentralized.   ThesereasonshelpexplainwhyBAinitiativesthatseekto integrate and align BA resources into any type of BA group within the organization sometimes fail. The listing in Table 4.1 is not exhaustive, but it provides some of the important issues to consider in the process of structuring a BA group.

Reason	Description
Lack of Executive Sponsorship	Senior executive failure to recognize the value of BA eventually leads to a reduction in resources and eventual failure.
Limited Context Perception	There is an incorrect perception that analytics must be applied within a particular functional area in order to have the necessary validity to be applied to that area. Example: Financial regression analysis can only be applied correctly in the context of the finance area.
Belief of Physical Proximity	There is misperception that it takes physical proximity of the BA grouping in the business application area to be valid.
Lack of Leadership in BA Groupings	Without an advocate leader in the organization, as well as leaders in BA projects and teams to move the analysis to achieve desired goals, the entire BA effort will lead to eventual failure.
Lack of Support	Without support for needed personnel, collecting data and technology to process the data will lead to failure.
Lack of Collaboration Across All Organizational Groups	Analytics that solve problems across multiple, functional areas are more likely to be accepted and successful than those that lack the cross-over into multiple organizational groups.
Lack of Skilled and Human Resources	BA departments, projects, or teams that don't have the skilled personnel to deal with the execution of analysis will eventually cause the failure of BA.
Inability to Delegate Responsibility	There is a desire to delegate responsibility to solve problems locally (a matter of trusting your own) rather than seeking help throughout the organization. This impedes the flow of problem-solving efforts by an external BA department and impedes communication of information needed to successfully apply BA.
Lack of Integrated Processes	Information that is stored in silos and not shared makes it more difficult for BA analysis to succeed.

Table 4.1 Reasons for BA Initiative and Organization Failure

In summary, the organizational structure that a firm may select for the positioning of its BA grouping can either be aligned within an existing organizational structure, or the BA grouping can be separate, requiring full integration within all areas of an organization. While some firms may start with a number of small teams to begin their BA program, other firms may choose to start with a full-sized BA department. Regardless of the size of the investment in BA resources, it must be aligned to allow maximum information flow between and across functional areas to achieve the most benefits BA can deliver.

Teams

When it comes to getting the BA job done, it tends to fall to a BA team. For firms that employ BA teams, the participants can be defined by the roles they play in the team effort. Some of the roles BA team participants undertake and their typical background are presented in Table 4.2.

Title or Function	Role Description	Background or Skills of Participant
Analytics Modeler	Develop and maintain predictive and forecasting models to provide insight.	Statistics, operation research, analytic modeling.
Analytics Process Designer	Develop and enforce reusable processes to reduce BA execution time.	Management consultant, process mapping, systems design.
Analytics Analyst	Respond to BA inquiries from functional areas within the firm to gain insight.	Reporting, problem solving, communicating, and providing customer service.
BA Team Head	Provide leadership to BA team, define strategies and tactics to ensure improved business performance, and interface with management.	BA manager or administrator.
Business Domain Expert	Provide business experience to ensure relevance of insight, help interpret business measures and the meaning of data.	Business experience in the area of the problem or opportunity.
Data Manager	Ensure data availability and access while minimizing costs.	Data modeling or warehousing, experience in data quality processes.
Implementation Specialist	Ensure rapid and robust model deployment to reduce time in interface.	Information system and data warehousing expertise, enterprise architecture experience.
Monitoring Analyst	Identify, establish, and enforce common analytics to be used to measure value and optimize effort.	Management and BA expert, predictive and financial modeling, process design, and team mentoring.

Table4.2BATeamParticipantRoles*

Aligning BA teams to achieve their tasks requires collaboration efforts from team members and from their organizations. Like BA teams, collaboration involves working with people to achieve a shared and explicit set of goals consistent with their mission. BA teams also have a specific mission to complete. Collaboration through teamwork is the means to accomplish their mission.

Team members' need for collaboration is motivated by changes in the nature of work (no more silos to hide behind, much more open environment, and so on), growth in professions (for example, interactive jobs tend to be more professional, requiring greater variety in expertise sharing), and the need to nurture innovation (creativity and innovation are fostered

by collaboration with a variety of people sharing ideas). To keep their job and to progress in any business career, particularly in BA, team members must encourage working with other membersinsideateamandout.Fororganizations,collaboration ismotivated bythechanging nature of information flow (that is, hierarchical flows tend to be downward, whereas in modern organizations, flow is in all directions) and changes in the scope of business operations (that is, going from domestic to global allows for a greater flow of ideas and information from multiple sources in multiple locations).

How does a firm change its culture of work and business operations to encourage collaboration? One way to affect the culture is to provide the technology to support a more open, cross-departmental information flow. This includes e-mail, instant messaging, wikis (collaboratively edited works, like Wikipedia), use of social media and networking through Facebook and Twitter, and encouragement of activities like collaborative writing, reviewing, and editing efforts. Other technology supporting collaboration includes webinars, audio and video conferencing, and even the use of iPads to enhance face-to-face communication. These can be tools that change the culture of a firm to be more open and communicative.

Reward systems should be put into place to acknowledge team effort. Teams should be commended for their performance, and individuals should be praised for performance in a team. While middle-level managers build teams, coordinate their work, and monitor their performance, senior management should establish collaboration and teamwork as a vital function.

Despite the collaboration and best of intentions, BA teams sometimes fail. There are many reasons for this, butknowing some of themore common ones can help managers avoid them. Some of the more common reasons for team failure are presented in Table 4.3. They also represent issues that can cause a BA program to become unaligned and unproductive.

Reason for Failure	Descriptions
Lack of Communication	It is not enough to come up with valuable information for decision-making and to find business opportunities in data. That information must be shared with users, clients, and everyone within a firm for benefit to come from it. It is only when analytics show a tangible and beneficial outcome that they are considered BA. If those results are not communicated on a continual basis, BA teams can be perceived to provide less value to the organization.
Failure to Deliver	Not every BA team will be able to deliver valued information if the team lacks the ability or resources to deliver needed answers and information. The greater the number of BA team failures, the greater are the chances that the team will be eliminated.
Lack of Justification	BA teams require resource allocations. Those allocations come from other departments that supposedly benefit from BA contributions. Without the role of BA and its potential contributions to a firm being clearly spelled out, users might not associate the ongoing efforts of a BA team as being worth the money spent on them.
Fail to Provide Value	BA teams have to sell their roles and suggested solutions or ideas. Without a clear understanding of value for potential users, the team faces a hard sell.
Inability to Prove Success	BA teams need to document and measure the impact of their ideas and suggestions. Without that proof, potential users might not support future BA efforts.

Table4.3ReasonsforBATeamFailures*

Management Issues

Aligning organizational resources is a management function. There are general management issues that are related to a BA program, and some are specifically important to operating aBA department, project, or team. The ones covered in this section include establishing an information policy, outsourcing business analytics, ensuring data quality, measuring business analytics contribution, and managing change.

Establishing anInformationPolicy

There is a need to manage information. This is accomplished by establishing an information policy to structure rules on how information and data are to be organized and maintained and who is allowed to view the data or change it. The information policy specifies organizational rules for sharing, disseminating, acquiring, standardizing, classifying, and inventorying all types of information and data. It defines the specific procedures and accountabilities that identify which users and organizational units can share information, where the information can be distributed, and who is responsible for updating and maintaining the information.

In small firms, business owners might establish the information policy. For larger firms, data administration may be responsible for the specific policies and procedures for data management (Siegel and Shim, 2003, p. 280). Responsibilities could include developing the information policy, planning data collection and storage, overseeing database design, developingthedatadictionary,aswellasmonitoringhowinformationsystemsspecialistsand end user groups use data.

A more popular term for many of the activities of data administration is data governance, which includes establishing policies and processes for managing the availability, usability, integrity, and security of the data employed in businesses. It is specifically focused on promoting data privacy, data security, data quality, and compliance with government regulations.

Such information policy, data administration, and data governance must be in place to guard and ensure data is managed for the betterment of the entire organization. These steps are also important in the creation of database management systems (see Chapter 3, "What Resource Considerations Are Important to Support Business Analytics?") and their support of BAtasks.

OutsourcingBusinessAnalytics

Outsourcing can be defined as a strategy by which an organization chooses to allocate some business activities and responsibilities from an internal source to an external source (Schniederjans et al., 2005, pp. 3–4). Outsourcing business operations is a strategy that an organization can use to implement a BA program, run BA projects, and operate BA teams. Any business activity can be outsourced, including BA. Outsourcing is an important BA management activity that should be considered as a viable alternative in planning an investment in any BA program.

BA is a staff function that is easier to outsource than other line management tasks, such as running a warehouse. To determine if outsourcing is a useful option in BA programs, management needs to balance the advantages of outsourcing with its disadvantages. Some of the advantages of outsourcing BA include those listed in Table 4.4.

Advantage of Outsourcing BA	Description
Less Expensive	Maintaining a fully functioning BA department when analytics might only be useful periodically may be more expensive than occasionally hiring an outside consulting BA firm to solve a problem.
Superior Analytics	The pool of analytic capabilities is always going to be greater outside a firm.
More Staffing Flexibility	Staff positions are often the first cut in economy downturns. Using consultants is easier and less expensive than hiring full-time BA staff. Outsourcing permits more flexibility to add and reduce BA services as needed.
New Knowledge	Experienced BA consultants bring a variety of knowledge and experience from having worked with many other firms. That type of experience may be of great competitive advantage.

Table 4.4 Advantages of Outsourcing BA

Nevertheless, there are disadvantages of outsourcing BA. Some of the disadvantages to outsourcing are presented in Table 4.5.

Disadvantages of Outsourcing BA	Description
Loss of Control	Once outsourced, most of the control of a BA project is under the control of the outsourcing firm. The client firm might lose not only control, but also opportunities for new and unique information, which the outsourcing firm may not share with the client.
Difficulties in Managing the Relationship	Client firms may find it difficult to communicate with the outsourcing firm because of distance, differing culture, language issues, and more. The lack of management could cause substantial problems with customer service and product quality.
Weakens Innovation	Having outsourced a client firm's internal experts, the remaining collaboration within the firm's personnel is reduced, and that reduces the opportunity for innovation efforts through shared collaboration.
Risk of Information	Outsourcing staff are exposed to client proprietary information, including innovations in analytics. This information could be shared with other competing firms, placing the client firm at risk.
Worthless Analytics	Sometimes outsourcing partners are less capable than internal analysts, wasting time and money.

Table 4.5 Disadvantages of Outsourcing BA

Managing outsourcing of BA does not have to involve the entire department. Most firms outsource projects or tasks found to be too costly to assign internally. For example, firms outsource cloud computing services to outside vendors (Laudon and Laudon, 2012, p. 511), and other firms outsource software development or maintenance of legacy programs to offshore firms in low-wage areas of the world to cut costs (Laudon and Laudon, 2012, p.192).

Outsourcing BA can also be used as a strategy to bring BA into an organization (Schniederjans et al., 2005, pp. 24–27). Initially, to learn how to operate a BA program, project, or team, an outsource firm can be hired for a limited, contracted period. The client firm can then learn from the outsourcing firm's experience and instruction. Once the outsourcing contract is over, the client firm can form itsown BA department, project, or team.

EnsuringData Quality

Business analytics, if relevant, is based on data assumed to be of high quality. Data quality refers to accuracy, precision, and completeness of data. High-quality data is considered to correctly reflect thereal world in which itis extracted. Poor-quality data caused by data entry errors, poorly maintained databases, out-of-date data, and incomplete data usually leadsto bad decisions and undermines BA within a firm. Organizationally, the database management systems (DBMS, mentioned in Chapter 3) personnel are managerially responsiblefor ensuring data quality. Because of its importance and the possible location of the BA departmentoutsidethemanagementinformationsystemsdepartment(whichusuallyhoststhe DBMS), it is imperative that whoever leads the BA program should seek to ensure data quality efforts are undertaken.

Ideally, a properly designed database with organization-wide data standards and efforts taken to avoid duplication or inconsistent date elements should have high-quality data. Unfortunately, times are changing, and more organizations allow customers and suppliers to enter data into databases directly via the Web. As a result, most of the quality problems originate from data input such as misspelled names, transposed numbers, or incorrect or missing codes.

An organization needs to identify and correct faulty data and establish routines and procedures for editing data in the database. The analysis of data quality can begin with a data quality audit, in which a structured survey or inspection of accuracy and levelof completeness of data is undertaken. This audit may be of the entire database, just a sample of files,orasurveyofendusersforperceptionsofthedataquality.Ifduringthedataquality

audit files are found that have errors, a process called data cleansing or data scrubbing is undertaken to eliminate or repair data. Some of the areas in a datafile that should be inspected in the audit and suggestions on how to correct them are presented in Table 4.6.

Data Inspection Items	Description and Cleansing/Scrubbing Recommendation
Current Data	Check to make sure the data is current. If it is out of date, remove it.
Completeness	Check to see if there is missing data. If more than 50 percent is missing, remove the entire file from the database.
Relevance	Check to see if the data is no longer relevant for the purpose for which it was collected. If it's no longer relevant, consider removing it from the database.
Duplication	Check to see if duplicate data files exist in the database. Remove duplicate data.
Outliers	Check for extreme values (outliers) in quantitative data files for possible errors in data coding. Remove from the data file any suspected of being in error, or repair the data.
Inconsistent Values	If data fields contain both characters and real numbers data where only characters or numbers should be, explore repairing the data.
Coding	If suspicious or unknown coding of data exists in data files, remove from the database or repair the coding of data.

Table 4.6 Quality Data Inspection Items and Recommendations

Measuring Business Analytics Contribution

The investment in BA must continually be justified by communicating the BA contribution to the organization for ongoing projects. This means that performance analytics should be computed for every BA project and BA team initiative. These analytics should provide an estimate of the tangible and intangible values being delivered to the organization. This should also involve establishing a communication strategy to promote the value being estimated.

Measuring the value and contributions that BA brings to an organization is essential to helping the firm understand why the application of BA is worth the investment. Some BA contribution estimates can be computed using standard financial methods, such as payback period (how long it takes for the initial costs to be returned by profit) or return on investment (ROI) (see Schniederjans et al., 2010, pp. 90–132), where dollar values or quantitative analysis is possible. When intangible contributions are a major part of the contribution being delivered to the firm, other methods like cost/benefit analysis (see Schniederjans et al., 2010, pp. 143–158), which include intangible benefits, should be used.

The continued measurement of value that BA brings to a firm is not meant to be self-serving, but it aids the organization in aligning efforts to solve problems and find new business opportunities. By continually running BA initiatives, a firm is more likely to identify internal activities that should and can be enhanced by employing optimization methodologies during the Prescriptive step of the BA process introduced in Chapter 1, "What Is Business Analytics?" It can also help identify underperforming assets. In addition, keeping track of investment payoffsforBA initiatives canidentifyareas intheorganization that shouldhavea higher priority for analysis. Indeed, past applications and allocations of BA resources that have shown significant contributions can justify priorities established by the BA leadership about where there should be allocated analysis efforts within the firm. They can also help acquire increases in data support, staff hiring, and further investments in BA technology.

ManagingChange

Wells (2000) found that what is critical in changing organizations is organizational culture and the use of change management. Organizational culture is how an organization supports cooperation, coordination, and empowerment of employees (Schermerhorn, 2001, p. 38). Changemanagement is definedas anapproach fortransitioning theorganization (individuals, teams, projects, departments) to a changed and desired future state (Laudon and Laudon, 2012, pp. 540–542). Change management is a means of implementing change in an organization, such as adding a BA department (Schermerhorn, 2001, pp. 382–390). Changes in an organization can be either planned (a result of specific and planned efforts at change with direction by a change leader) or unplanned (spontaneous changes without direction of a change leader). The application of BA invariably will result in both types of changes because of BA's specific problem-solving role (a desired, planned change to solve a problem) and opportunity-finding exploratory nature (unplanned new knowledge opportunity changes) of BA. Change management can also target almost everything that makes up an organization (see Table 4.7).

Data Inspection Items	Description and Cleansing/Scrubbing Recommendation
Current Data	Check to make sure the data is current. If it is out of date, remove it.
Completeness	Check to see if there is missing data. If more than 50 percent is missing, remove the entire file from the database.
Relevance	Check to see if the data is no longer relevant for the purpose for which it was collected. If it's no longer relevant, consider removing it from the database.
Duplication	Check to see if duplicate data files exist in the database. Remove duplicate data.
Outliers	Check for extreme values (outliers) in quantitative data files for possible errors in data coding. Remove from the data file any suspected of being in error, or repair the data.
Inconsistent Values	If data fields contain both characters and real numbers data where only characters or numbers should be, explore repairing the data.
Coding	If suspicious or unknown coding of data exists in data files, remove from the database or repair the coding of data.

Table4.7Change ManagementTargets*

Itisnotpossibletogainthebenefits ofBAwithoutchange. Theintentischangethatinvolves findingnewanduniqueinformation onwhichchangeshouldtakeplaceinpeople,technology systems, or business conduct. By instituting the concept of change management within an organization,afirmcanalignresourcesandprocessestomorereadilyaccept changesthatBA may suggest. Instituting the concept of change management in any firm depends on the unique characteristics of that firm. There are, though, a number of activities in common with successful change management programs, and they apply equally to changes in BA departments, projects, or teams. Some of these activities that lead to change management success are presented as best practices in Table 4.8.

Best Practice	Description
Champion	Change is scary business for some, and a strong leader for change can champion the change effort, calming fears and explaining the need for change. The champion also helps direct efforts, motivate change, and keep the change activities on track.
Clearly Stated Goals	Any type of change should be clearly defined, including what the changes are, which personnel have to change, and what the processes involve and how they affect technology. This would also include deadlines needed to keep the change effort on track.
Good Communication	To avoid resistance to change (a natural norm to anything that is new), it is useful to help those facing the change understand its value through effective and repeated communications, keeping them informed on progress and easing fears.
Measured Performance	Any goals stated prior to the launch of change can be used to measure performance during the changeover period. Seeing business performance improve with changes can motivate further change and support by those impacted.
Senior Management Support	Critical to all BA departments, projects, or teams is the need for senior management to support change efforts. Sometimes that support is in direct dollars, and sometimes it's in lending authority to get resources needed for BA work.

Table4.8Change ManagementBestPractices

Unit-3 DescriptiveAnalytics

Unit objectives:

• Explain whyweneedto visualizeandexplore data.

• Describestatisticalcharts andhowtoapplythem.

• Describedescriptivestatisticsusefulinthedescriptivebusinessanalytics(BA)process.

• DescribesamplingmethodsusefulinBAandwheretoapply them.

• DescribewhatsamplingestimationisandhowitcanaidintheBA process.

• Describetheuseofconfidenceintervalsandprobability distributions.

• ExplainhowtoundertakethedescriptiveanalyticsstepintheﬁBAprocess.

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 supportofBA.Forpurposes ofillustration, wewill use thedata setin Figure5.1representing four different types of product sales (Sales 1, Sales 2, Sales 3, and Sales 4).

Figure 5.1 Illustrative sales datasets

When using SAS, datasets are placed into files like that in Figure 5.2.
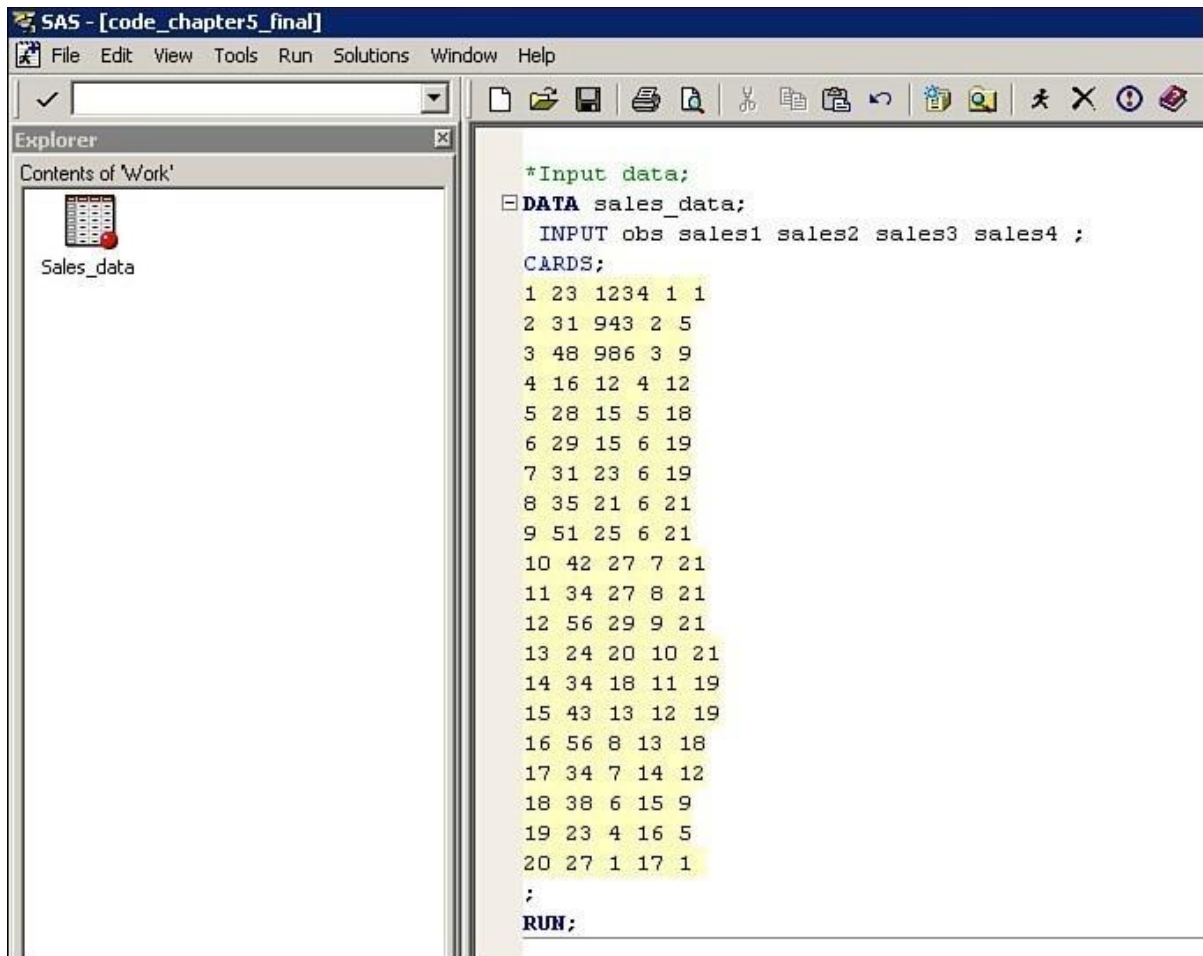
Figure5.2SAScodingofsalesdataset

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 shouldbefamiliar to experienced SASusers.Depending ontheSASversion andintent ofthe data file, the images in these figures may be slightly different.
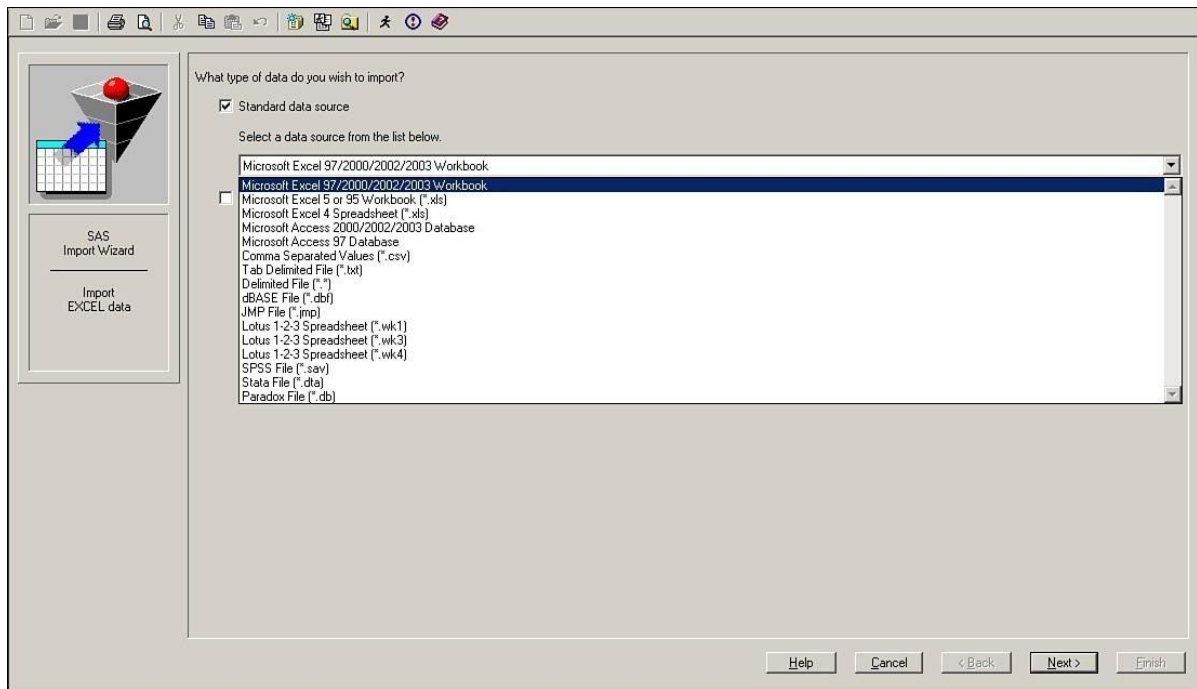
Figure5.3Creatingaworkbookfile for thedataset

Figure5.4Exceldatafile usedinthecreationoftheSASdatafile
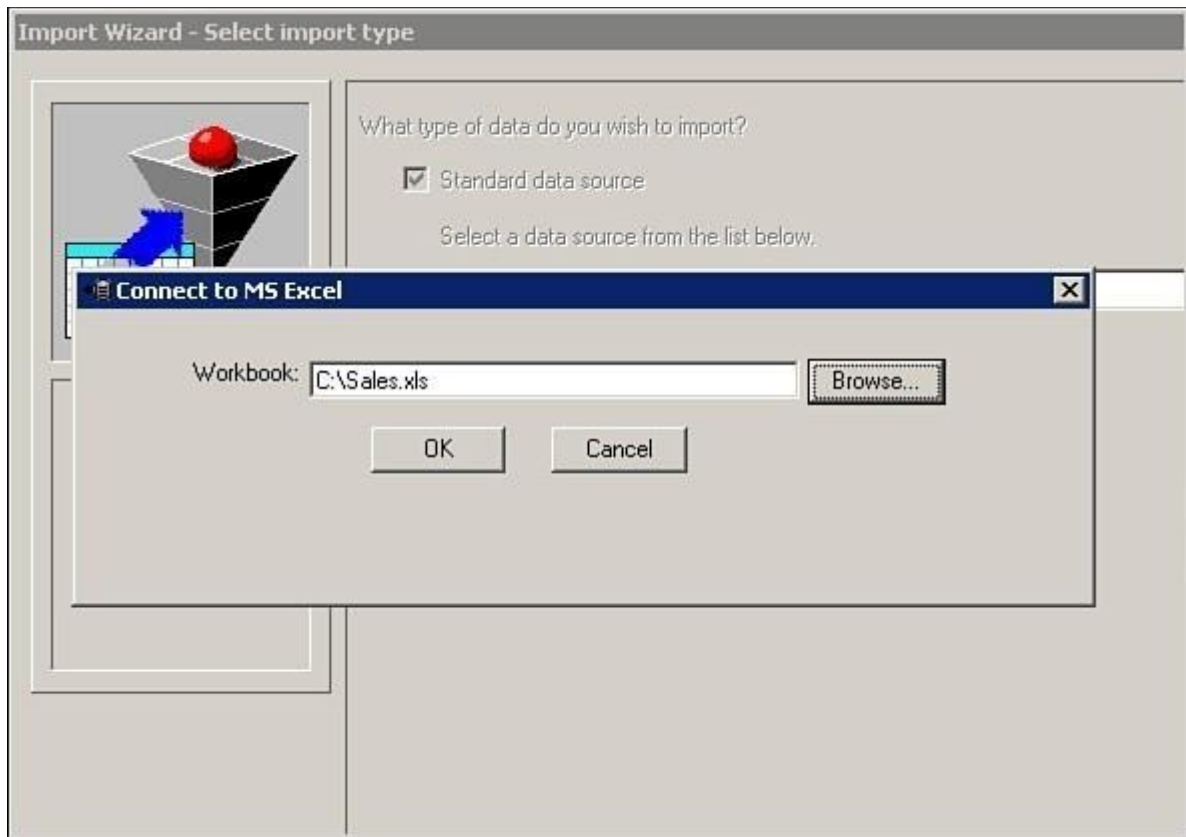
Figure5.5Steptopullthedataset fromanExcel (orany)file using SAS
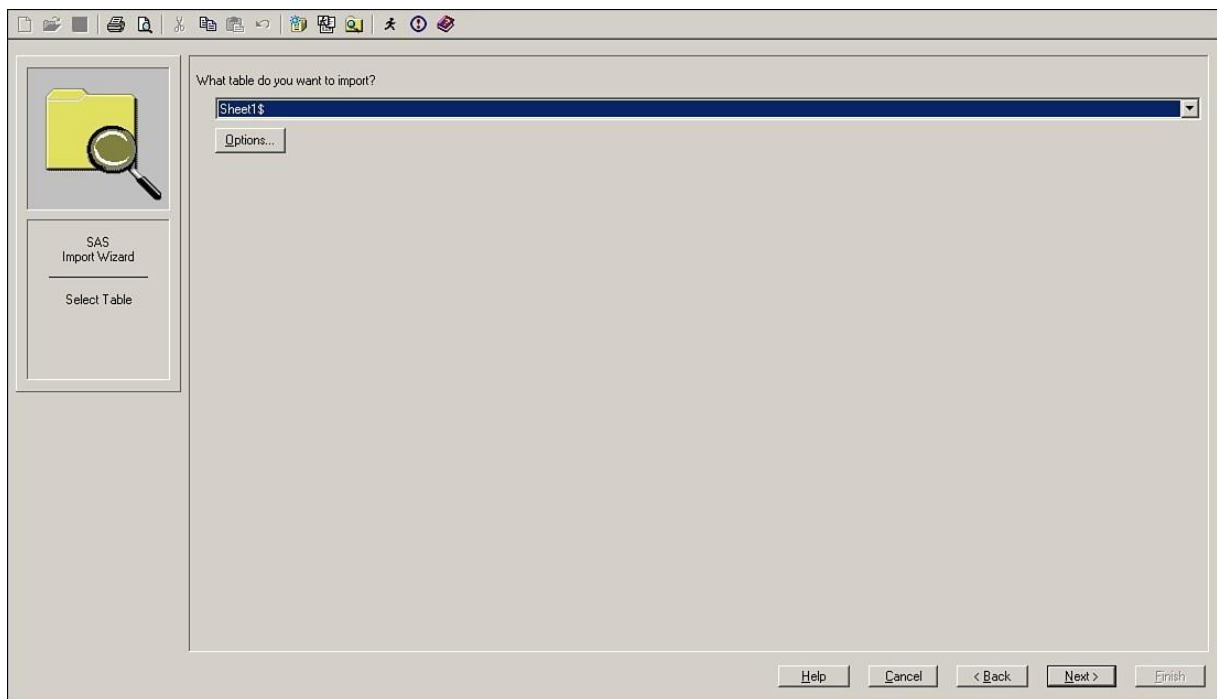
Figure5.6IdentifyfileusingSAS
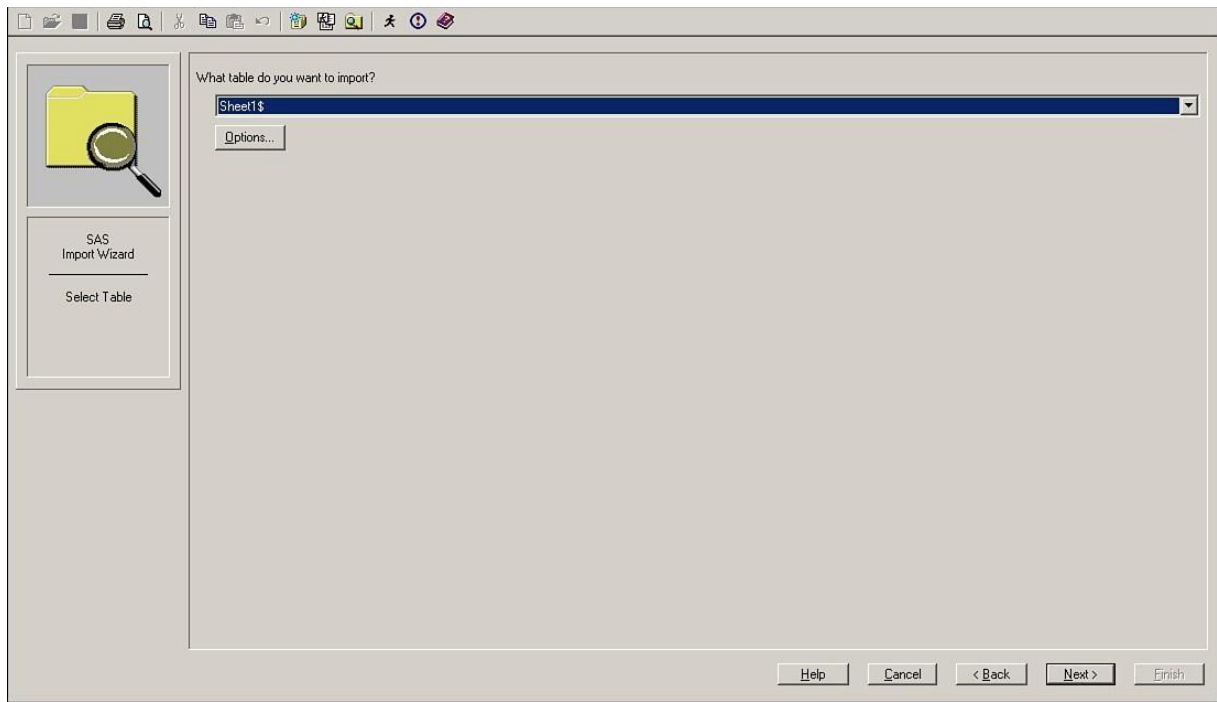
Figure5.7LabelSASfileasSALES_DATA
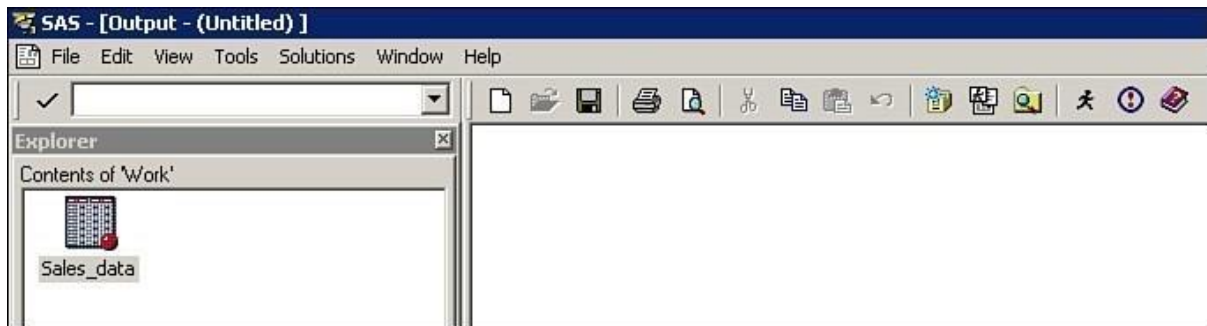
Figure5.8ShowstheSALES_DATASASfileiscreated
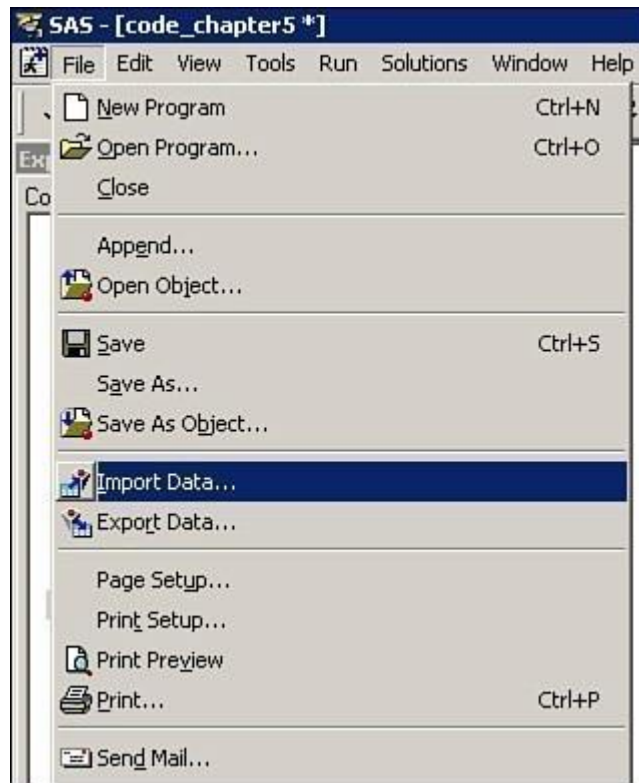
Figure5.9Steptoimport thecreatedfile
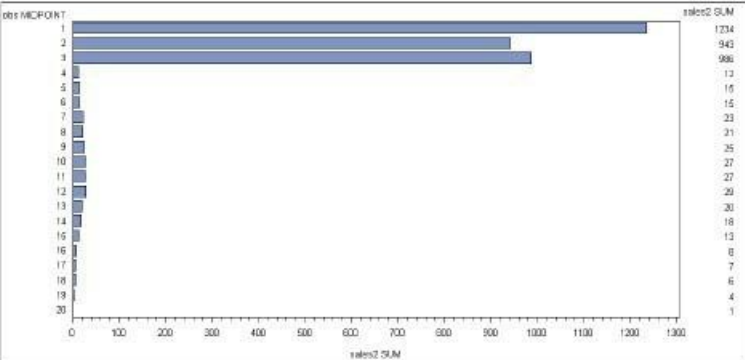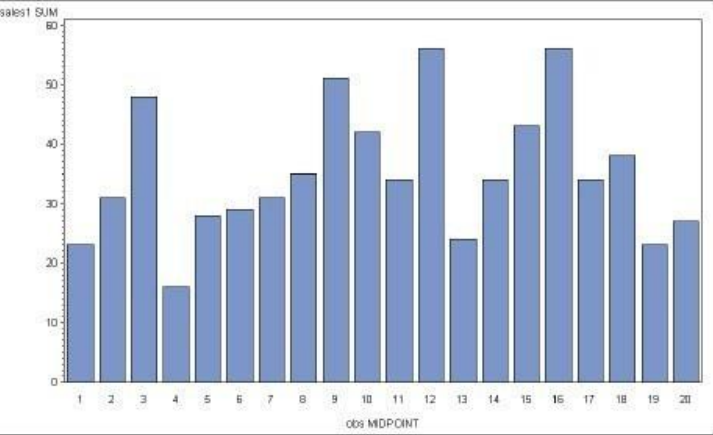
SAS permits the use of many different sources for data sets or data files to be entered into an SAS program. Big data files in Excel, SPSS, or other software applications can be brought intoSASprogramsusingasimilarsetofstepspresentedinthissection.Oncethedatasetsare structured for use with SAS, there 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.

VisualizingandExploringData

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 candraftavariety ofchartsfortheselected variablesinthedatasets.UsingthedatainFigure 5.1, charts can be created for the illustrative sales data sets. Some of these chartsare discussedinTable5.1asasetofexploratorytoolsthatarehelpfulinunderstandingthe

informationalvalueofdatasets.Thecharttoselectdependsontheobjectivessetforthe chart. The SAS program statements used to create each chart in Table 5.1 are provided.

Type of Chart	Application Notes	Chart Example
| Bar | • Can be horizontal, vertical, cone, or cyclically shaped and multi-dimensional with overlaying variables.<br><br>• Ideal for showing comparative improvement over time.<br><br>• Example: Bars showing productivity of one person versus another. | ```
proc gchart data=sales_data;
hbar obs /sumvar=sales2
midpoints=1 to 20 by 1;
run;
quit;
```<br><br> |
| Column | • Same as a bar chart. | ```
proc gchart data=sales_data;
vbar obs /sumvar=sales1
midpoints=1 to 20 by 1;
run;
quit;
```<br><br> |
| Line | • Ideal for showing linear trend and other linear or nonlinear | ```
symbol value=dot interpol=sms line=1 width=2;
proc gplot data=sales_data;
plot sales1*obs;
run;
quit;
``` |

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;
```


SUM of sales1 by obs

Scatter	- Useful when patterns are observed in the data sets.

```
proc sgplot 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.

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

- Reduces the size of data by grouping data points into frequencies.

Table5.1StatisticalChartsUsefulinBA

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 appearanceof 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 structureor 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.

DescriptiveStatistics

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												

Table5.2SASDescriptiveStatistics

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.

(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.

Table5.3DescriptiveStatisticsUsefulinBA

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,Sales2ishighly, 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 predictedthatthedatasetispositivelyskewedandpeaked.NoteinFigure5.10thatthe

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.

Figure5.10SAShistogramofSales2data

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 behavioror 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.

SamplingandEstimation

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.

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 fromwhich it comes. In sampling data, there are three components to recognize: a population, a sample, and a sample element (the items thatmakeupthesample).Afirm'scollectionofcustomerserviceperformancedocuments 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.	• 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 ensurerepresentativesamplesbeingdrawnfromthepopulation,sometimes becauseofcostor 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 natureofthesample.AsmentionedintheapplicationnotesinTable5.4,thesizeofthe

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 thechoice to something like aquota method. Another best practices recommendation is to startwith theobjective(s) oftheBA project andusethem 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 resultsof 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 observedby 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           SAMP1
```

	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

Figure5.11SASprogramstatements forstratification/randomsampling forSales4 variable

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) ± [(confidence coefficient) × (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 likeamean.Theconfidence coefficientissetasapercentage todefinethedegreeof 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:

Confidenceinterval=mean±[(95percent)×(standard errorofthe 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 bythe sample has a 95percent chance of falling in the interval. Inthis 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.

SAScomputesconfidenceintervalswhenanalyzingvariousstatisticalmeasuresandtests.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.

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

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.

Figure5.12Standardnormalprobability 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.

Figure5.13Probabilityfunctionexample

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.

Marketing/PlanningCaseStudyExample:DescriptiveAnalyticsStepintheBAProcess

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 strategiesusedinaBAproblemanalysis.Thisisthefirstinstallmentofthecasestudydealing 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.

CaseStudyBackground

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 onlyduringthemonththatthepromotioneffortsareallocated.Basically,promotion fundsare

allocated atthebeginning or during themonth, and whatever sales occur arerecorded forthat 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, newspaperads, 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.

Thesefivedatasets,inthousandsofdollars,arepresentintheSASprintoutshowninFigure 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 firmallocates 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 aboutthebehaviorofproductsalestopromotionefforts,webeginwiththefirststep   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

Figure5.14Dataformarketing/planningcase study

DescriptiveAnalyticsAnalysis

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
```

Table5.7SASDescriptiveStatisticsfortheMarketing/PlanningCaseStudy

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).Thesechartsareinteresting,butthey'renotveryrevealingofbehaviorthathelps   in understanding future sales trends that may be hiding in this data.

Figure5.15Preliminary SASchartsforthemarketing/planningcasestudy

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 lookingfora graphto showatrendline.Onetypeofsimple graph that is

related to trend analysis is a line chart. Using SAS again, one can compute line charts foreach ofthe five data sets. These charts are presented in Figure 5.16.The vertical axis consists ofthedollarvalues,andthehorizontal axisisthenumber orderingofobservationsaslistedin the data sets.

Figure5.16Preliminary SASlinechartsforthemarketing/planningcasestudy

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 20months ofdata are not inany particular order and are not related totime, they areindependent valuesthatcanbereordered. Reorderingdataorsorting itcanbeapartofthe 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

Figure5.17Reordereddatainline chartsforthemarketing/planningcase study

Figure5.18SASlinechartsbasedonreordereddataforthemarketing/planning case study

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 explorein more detail and sophistication the best set of predictive variables to predictthe 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 ofnewspaperandPOSads tosales.Themanagerial ramifications oftheseresults 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.

Unit4–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 thedescriptiveanalysis 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 todevelopneededforecastsofbusinessactivity.(SeeAppendixE.)Whatevermethodologyis used, the identification of future trends or forecasts is the principal output of the predictive analytics step in the BA process.

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.

Logic-DrivenModels

A logic-driven model is one based on experience, knowledge, and logical relationships of variablesandconstantsconnectedtothedesiredbusinessperformanceoutcomesituation.The question here is how toput variables and constants together to create a model that can predict thefuture.Doingthisrequires businessexperience. Modelbuildingrequires anunderstanding 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 ina 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 termsof 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.

Figure6.1Cause-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 ofprofit. 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:

$$\text{Profit} = \text{Revenue} - \text{Cost, or}$$

$$\text{Profit} = (\text{UnitPrice} \times \text{QuantitySold}) - [(\text{FixedCost}) + (\text{VariableCost} \times \text{Quantity Sold})], \text{ or}$$

$$P = (UP \times QS) - [FC + (VC \times QS)]$$

Figure6.2Aninfluencediagram

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.

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 theAdditional Information column in Table 6.1 toobtainfurther information ontheuseandapplicationofthedata-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.

Table6.1Data-Driven Models

DataMining

As mentioned in Chapter 3, data mining is a discovery-driven software application process that provides insights into business databy 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 andare 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.

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.

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 datamining 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.

Table6.2Types ofInformationandDataMiningMethodologies

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 todevelop, train, andusethesoftwaretoidentify particular traits (suchasbad 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. Theanalysis 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 anoutsource 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 inthe data andbegins togroupthem. The divisive strategy is atop-downapproach, 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 Euclideandistance 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 Kcandidates selected inthe initial runthrough thelist. Each runor iteration through thedatasetallows 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 classifythe 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

Figure6.3Salesdataforclusterclassification 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 thisproblem 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

                                Maximum Distance
                      RMS Std         from Seed     Radius     Nearest     Distance Between
Cluster   Frequency  Deviation    to Observation   Exceeded    Cluster    Cluster Centroids

   1           9       789.8           1857.9                     2             3806.2
   2          11       879.3           2133.9                     1             3806.2
```

Table6.3 SASK- 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.

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.

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

Figure6.4Dataformarketing/planningcase 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 adsto productsales. Soalthough radio andTV commercials aremostpromising, amore 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.

PredictiveAnalyticsAnalysis

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 correlationcoefficientsand theirlevelsof significancearepresentedin 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 r 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      0.97714     -0.28307      0.95797      0.01265
                         <.0001        0.2265      <.0001        0.9578

radio       0.97714      1.00000     -0.23836      0.96610      0.06040
            <.0001                     0.3115      <.0001        0.8003

paper      -0.28307     -0.23836      1.00000     -0.24588     -0.09006
             0.2265       0.3115                    0.2960       0.7057

tv          0.95797      0.96610     -0.24588      1.00000     -0.03602
            <.0001       <.0001        0.2960                    0.8802

pos         0.01265      0.06040     -0.09006     -0.03602      1.00000
             0.9578       0.8003        0.7057       0.8802
```

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

Table6.4SASPearsonCorrelationCoefficients:Marketing/PlanningCaseStudy

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 thedata 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 ita 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 correction 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

                                Sum of          Mean
       Source            DF    Squares        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

                          Parameter      Standard
       Variable     DF     Estimate         Error    t Value   Pr > |t|

       Intercept     1        16649    1398.32203      11.91     <.0001
       pos           1     44.14019     822.47123       0.05     0.9578
```

Table6.5SASPOSRegressionModel:Marketing/Planning CaseStudy

```
                          The REG Procedure
                            Model: MODEL1
                       Dependent Variable: sales

                Number of Observations Read          20
                Number of Observations Used          20


                          Analysis of Variance

                                Sum of          Mean
       Source            DF    Squares        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

                          Parameter      Standard
       Variable     DF     Estimate         Error    t Value   Pr > |t|

       Intercept     1        16649    1398.32203      11.91     <.0001
       pos           1     44.14019     822.47123       0.05     0.9578
```

Table6.6SASRadioRegressionModel:Marketing/Planning CaseStudy

```
                        The REG Procedure
                         Model: MODEL1
                     Dependent Variable: sales

                Number of Observations Read        20
                Number of Observations Used        20


                        Analysis of Variance

                              Sum of          Mean
    Source           DF      Squares        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

                         Parameter      Standard
    Variable     DF       Estimate         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 combinationsforthevariablesandcomputetheirregressionmodels.Thecombinationsare (1)POSand radio; (2)POSandTV; (3)POS,radio,and TV;and (4)radioand 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 decisionis 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

Table6.8SASVariableCombinationsandRegression Model Statistics:Marketing/PlanningCaseStudy

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 theR-Squared (Adjusted) statistic, which is commonly used in multiple regression models.

TheR-Square Adjusted statistic does not havethe sameinterpretation as R-Square (aprecise, 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 accountthe phenomenon of the R-Square automatically increasing when additional independent variablesareaddedtothemodel. Thisphenomenonislikeapainterputting paintonacanvas, wheremorepaintadditivelyincreasesthevalueofthepainting.Yetbycontinuallyadding

paint, there comes a point at which some paint coversother 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 ofany independent value being added. If a set of independent variables is introduced into a regression model one at a time inforward 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.BysystematicexperimentationwiththeR-Squareadjustedrecomputedforeachadded 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

Table6.9SASBestVariableCombinationRegressionModeland Statistics:Marketing/PlanningCaseStudy

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:

$$Yp = -17150 + 275.69065 \, X1 + 48.34057 \, X2$$

where:

Yp= theestimatednumberofdollars ofproductsales

X1= thenumber ofdollarstoinvestinradiocommercials

X2=thenumberofdollarstoinvestinTVcommercials

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 X1), the firm will receive \$275.69 in product sales (represented by Yp in the model). Likewise, for everydollarallocatedtoTVcommercials (representedbyX2),thefirmwillreceive\$48.34in product sales.

Acautionshouldbementionedontheresultsofthiscase study.Manyfactorsmightchallenge a result, particularly those derived from using powerful and complex methodologies like multiple regression.Assuch, theresultsmay notoccur as estimated, becausethemodel isnot reflecting pastperformance. Whatisbeingsuggestedhereisthatmoreanalysiscanalwaysbe 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 moredetailed, quantifiable relationship between the generation of product sales and the sourcesof 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 typesof 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 5:Prescriptive Analytics

Unitobjectives:

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

Explaintheroleofcasestudiesin prescriptiveanalytics.

Explainhowfittingacurvecanbeusedinprescriptiveanalytics. Explain how

to formulate a linear programming model.

Explainthevalueoflinearprogramming intheprescriptive analyticsstep ofBA.

Introduction

Afterundertakingthedescriptiveandpredictiveanalytics stepsintheBAprocess,oneshould 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 resourcesto 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.

InChapter 3,"WhatResourceConsiderations AreImportant toSupportBusinessAnalytics?" 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 BAprocess is again presented in Figure 7.1 to form the basis of this chapter's content.

Figure7.1Prescriptiveanalyticmethodologies

PrescriptiveModeling

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

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 functionto optimize a decision can fall in the prescriptive analytics step.

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

ofthemrequirecarefulestimation ofparameters thatmayormaynotbeaccurate, 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 servicefacilitiesand hascollectedthe 20datapointsinFigure7.2. Whetherthere 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 usefulfor 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

Figure7.2DataforSAScurve 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

adjustedR-Squarestatisticcanrevealthebestestimatedrelationshipbetweentheindependent
(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

Quadratic-Full

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

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

Figure7.4GraphofallSAScurve-fittingmodels

FromTable7.3,theresultingtwostatisticallysignificantcurve-fittedmodelsfollow:

$Yp = 35418 - 5589.432\ X + 268.445 X2$ [Quadraticmodel]

$Yp = 36134 - 5954.738\ X + 310.895\ X2 - 1.347\ X3$ [Cubicmodel]

where:

Yp=theforecastedorpredictedtotalcost

X=thenumber ofcomputer 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 (Yp). 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 costsofserverscanbecomputed usingthe quadratic model, aspresented inFigure7.5.These values were computed by plugging the number of server values (1 to 20) intothe Yp 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, whereasthe 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.

Figure7.5Predicted totalcostinserverproblemforeachserveralternative

Like all regression models, curve fitting is an estimation process with risks, butthe 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 (likeLP). 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).

Continuationof Marketing/PlanningCase Study Example: Prescriptive Step in the BA Analysis

In Chapter 5, "What Is Descriptive Analytics?" and Chapter 6, "What Is Predictive Analytics?" anongoingmarketing/planning casestudywaspresented toillustrate someofthe 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.

CaseBackground 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.

PrescriptiveAnalysis

The allocation problem of the budget to purchase radio and TV commercials isa 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.

Formulation ofLPMarketing/PlanningModel

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 eachdollar

ofradiocommercials estimated withtheregressionmodel,$275.691willbereceived,andfor 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:

$X1$=thenumberofdollarsto investinradiocommercials forthenextmonthlycampaign

$X2$=thenumber ofdollarstoinvestin TVcommercials forthenextmonthlycampaign

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:$Z = 275.691 X1 + 48.341 X2$

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

Constraint 1—Nomorethan $350,000is allowed forthetotal budgetto allocate toboth radio ($X1$) and TV ($X2$) commercials. So add $X1 + X2$ and set it less than or equal to 350,000 to formulate the first constraint as follows:

$X1 + X2 \leq 350000$

Constraint 2—To get a discount on radio ($X1$) commercials, the firm must allocate a minimum of $15,000 to radio. The constraint for this limitation follows:

X1≥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 ≥ 75000

Constraint 4—This is a blending problem constraint (see Appendix B, Section B.6.3). Whatis 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:

2X1 −X2 = 0

5. Statethenonnegativityandgivenrequirements—Withonlytwovariables,thisformal requirement in the formulation of an LP model is expressed as follows:

X1, X2 ≥0

Becausethesevariables areindollars,theydonothavetobeintegervalues.(Theycanbeany real or cardinal number.) The complete LP model formulation is given here:

SolutionfortheLPMarketing/PlanningModel

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.

Figure7.6LINGOLPmodelentryrequirements:marketing/planning casestudy

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 unknownsand 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.

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.


```
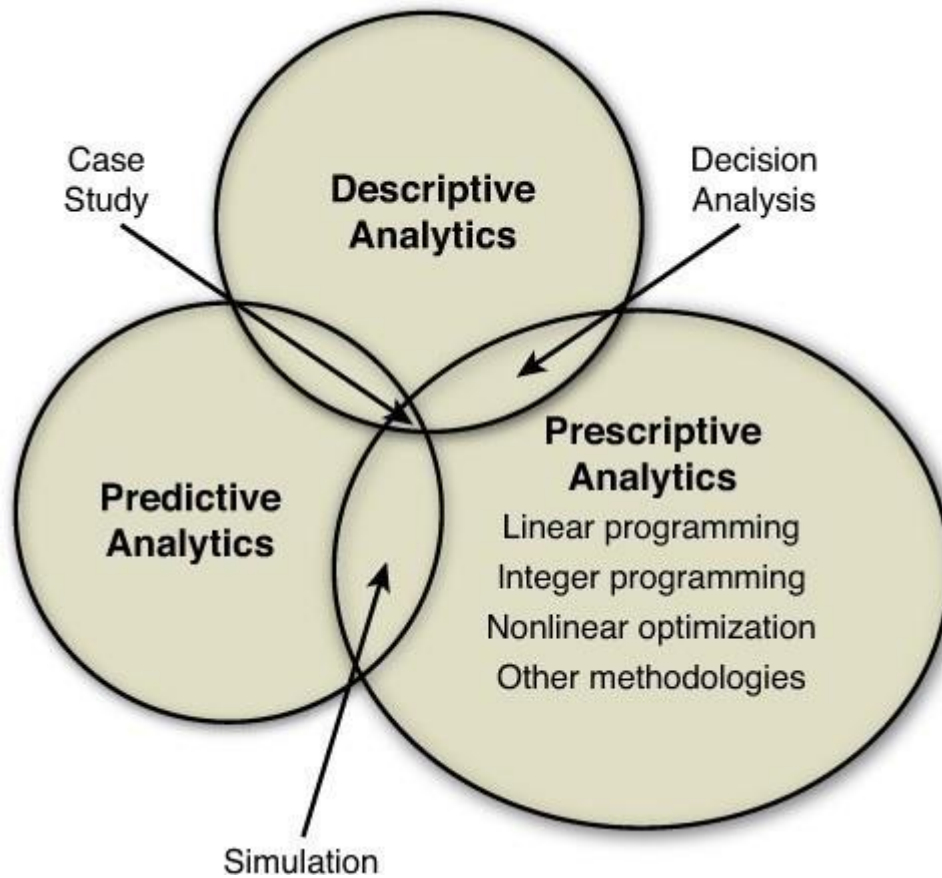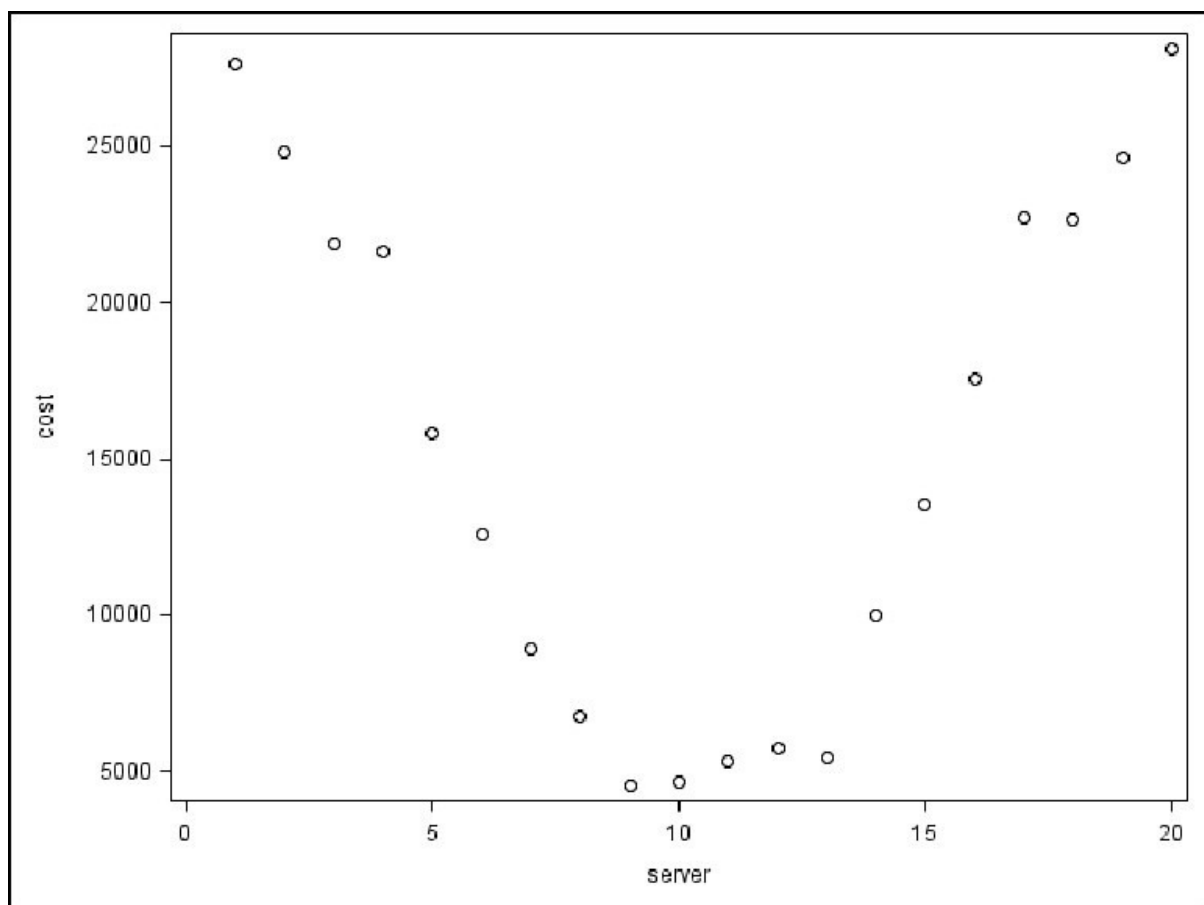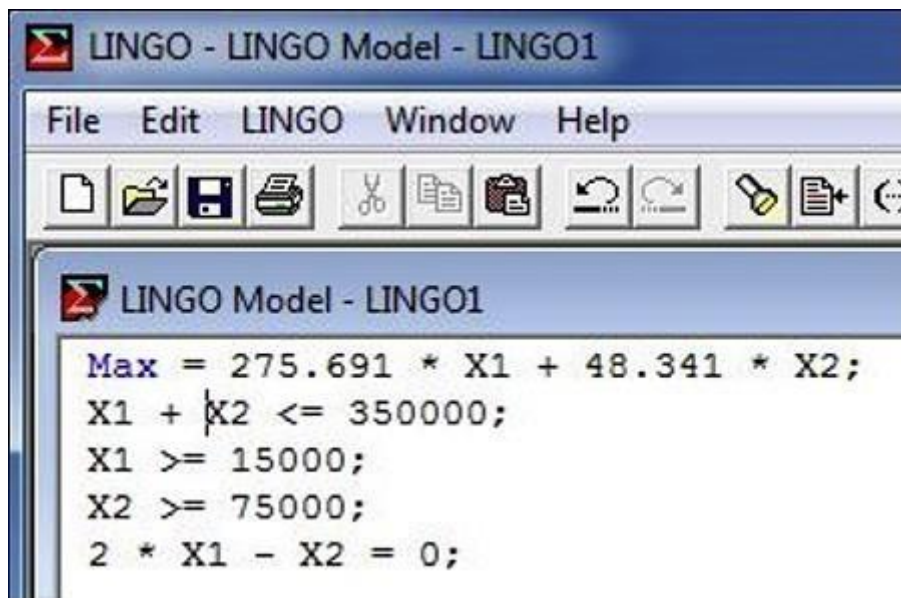Solution Report - LINGO1

   Global optimal solution found.
   Objective value:                        0.4344352E+08
   Total solver iterations:                      0


                  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
```

Figure7.7LINGOLPmodelsolution: marketing/planningcasestudy

Asitturnsout,theoptimaldistributionofthe$350,000promotionbudgetistoallocate $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 businessperformance improvement is shown to better this organization's ability to use its resources more effectively. The management job of monitoring performance and checking to seethat business performance is in fact improved is a needed final step in the BA analysis. Without proofthatbusinessperformanceisimproved,it'sunlikelythatBAwouldcontinuetobeused.

FinalCommentontheMarketing/PlanningModel

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 assuranceof predictive validity. There are many other methods and approaches that could have been used inthiscasestudy.Learninghowtousemorestatisticalanddecisionsciencetools 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 thequadratic 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:

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

Step1.Giventhequadraticfunction above,takeitsfirst derivative.

d(Yp)= –5589.432+536.89 X

Step2. Set the derivative function equal to zero and solve for X.

0 =–5589.432 + 536.89 X

X= 10.4107