Learn Business Analytics in Six Steps Using SAS and R

A Practical, Step-by-Step Guide to Learning Business Analytics

Subhashini Sharma Tripathi

Apress®

Learn Business Analytics in Six Steps Using SAS and R

A Practical, Step-by-Step Guide to Learning Business Analytics

Subhashini Sharma Tripathi

Apress®

Learn Business Analytics in Six Steps Using SAS and R

Subhashini Sharma Tripathi
Bangalore, Karnataka
India

ISBN-13 (pbk): 978-1-4842-1002-4 ISBN-13 (electronic): 978-1-4842-1001-7
DOI 10.1007/978-1-4842-1001-7

Library of Congress Control Number: 2016961720

Copyright © 2016 by Subhashini Sharma Tripathi

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

Trademarked names, logos, and images may appear in this book. Rather than use a trademark symbol with every occurrence of a trademarked name, logo, or image we use the names, logos, and images only in an editorial fashion and to the benefit of the trademark owner, with no intention of infringement of the trademark.

The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Managing Director: Welmoed Spahr
Lead Editor:Celestin Suresh John
Technical Reviewer: Ujjwal Dalmia
Editorial Board: Steve Anglin, Pramila Balan, Laura Berendson, Aaron Black, Louise Corrigan, Jonathan Gennick, Robert Hutchinson, Celestin Suresh John, Nikhil Karkal, James Markham, Susan McDermott, Matthew Moodie, Natalie Pao, Gwenan Spearing
Coordinating Editor: Prachi Mehta
Copy Editor: Kim Wimpsett
Compositor: SPi Global
Indexer: SPi Global
Artist: SPi Global

Distributed to the book trade worldwide by Springer Science+Business Media New York, 233 Spring Street, 6th Floor, New York, NY 10013. Phone 1-800-SPRINGER, fax (201) 348-4505, e-mail orders-ny@springer-sbm.com, or visit www.springeronline.com. Apress Media, LLC is a California LLC and the sole member (owner) is Springer Science + Business Media Finance Inc (SSBM Finance Inc). SSBM Finance Inc is a **Delaware** corporation.

For information on translations, please e-mail rights@apress.com, or visit www.apress.com.

Apress and friends of ED books may be purchased in bulk for academic, corporate, or promotional use. eBook versions and licenses are also available for most titles. For more information, reference our Special Bulk Sales–eBook Licensing web page at www.apress.com/bulk-sales.

Any source code or other supplementary materials referenced by the author in this text are available to readers at www.apress.com. For detailed information about how to locate your book's source code, go to www.apress.com/source-code/. Readers can also access source code at SpringerLink in the Supplementary Material section for each chapter.

Printed on acid-free paper

Contents at a Glance

Contents

About the Author

Subhashini Sharma Tripathi is an analytics enthusiast. After working for a decade with GE Money, Standard Chartered Bank, Tata Motors Finance, and Citi GDM, she started teaching, blogging, and consulting in 2012. As she worked, she became convinced that analytics and data science help reduce dependency on experience. Further, she believes it gives modern managers a conclusive way to solve many real-world problems faster and more accurately. In this evolving business landscape, it also helps define longer-term strategies and makes better choices available. In other words, you can get "more bang for your buck" with analytics.

Subhashini is the founder of pexitics.com, and her first product is the Pexitics Talent Score, a pre-interview score. The company makes tools for effective human resource management and consults in analytics.

You can connect with her via LinkedIn at `https://in.linkedin.com/in/subhashinitripathi` or via e-mail with `subhashini@pexitics.com`.

Acknowledgments

This book is my first, and the experience of writing it has been an exciting and bumpy journey. This book and its writing coincided with the creation and launch of pexitics.com.

The journey would not have been possible without a lot of support and encouragement from my family and the editorial team at Apress, especially Celestin Suresh John, for ensuring that my morale did not flag on the way. I express my heartfelt gratitude to my mother, Dr. M. Tripathi (PhD), for her support and help in words, deeds and prayers.

My thought process has been significantly influenced by the book *Basic Business Statistics* (12th edition) by Mark L. Berenson, David M. Levine, and Timothy C. Krehbiel. I read about the DCOVA process in that book. As I worked with that process, I added another stage, called Insight Generation, and now use the process of DCOVA and I.

When I started my journey into number-based decision-making in 2002, there was a dearth of structured mentoring, and a lot of things were self-discovered and self-taught. I have written this book so that analytics and data science aspirants can start on the journey in a structured way and with a lot of confidence to solve real business problems.

The next edition will cover predictive models.

Introduction

In the last decade, analytics and data science have come into the forefront as support functions for business decisions. A decade ago, business analytics was a little-known career choice. With the drastic dip in data storage costs and the huge increase in data volumes (projected to hit 40 zettabytes in 2020), chief experience officers (CXOs) and modern managers now need analytics and data science to make informed decisions at every point.

Have you wondered how to get started on a career in analytics and data science?

This book teaches you how to solve problems and execute projects in analytics through the Define, Collect, Organize, Visualize, Analyze, and Insights (DCOVA and I) process. Thus, even when the data is very new or the problem is not familiar, you can solve it by using a step-by-step checklist for deduction and inferencing. Finally, for implementing analytics output, the conclusion or insight needs to be understood in plain business terms.

This book teaches you how to do analytics on business data using two popular software tools, SAS and R. SAS is licensed software that is the leader in the sectors that have regulatory supervision (banking, clinical research, insurance, and so on). R is open source software that is popular in sectors without regulators such as retail, technology (including ITES), BPOs, and so on. So, irrespective of the industry in which you work, this book will provide you with the knowledge and skills you and your managers need to make better decisions faster.

You no longer need to choose between the two most popular software tools.

How can business turn this data into useful information in a reasonably fast turnaround time? This question becomes important for running a successful business. Only if the information is available to management at the correct time will the business be able to make the correct decisions. For this, you need business analytics, loosely described as doing statistics on large volumes of data, to arrive at conclusions and models that will aid business decision-making.

The statistical techniques can be divided into the five broad segments of descriptive statistics, inferential statistics, differences statistics, associative statistics, and predictive statistics. I will cover models related to associative and predictive stats in the next edition. In this book, I will focus on developing your understanding of the process of problem-solving and the statistics related to the descriptive, differences, and associative statistical techniques.

Do connect with me via LinkedIn at https://in.linkedin.com/in/subhashinitripathi or via e-mail with subhashini@pexitics.com.

CHAPTER 1

The Process of Analytics

In this chapter, you will look at the process and evolution of analytics. These are some of the topics covered:

- The process of analytics
- What analytics is
- The evolution of analytics
- The dawn of business intelligence

What Is Analytics? What Does a Data Analyst Do?

A casual search on the Internet for *data scientist* offers up the fact that there is a substantial shortage of manpower for this job. In addition, Harvard Business Review has published an article called "Data Scientist: The Sexiest Job of the 21st Century" (http://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/ar/1). So, what does a data analyst actually do?

To put it simply, *analytics* is the use of numbers or business data to find solutions for business problems. Thus, a *data analyst* looks at the data that has been collected across huge enterprise resource planning (ERP) systems, Internet sites, and mobile applications.

In the "old days," we just called upon an expert, who was someone with a lot of experience. We would then take that person's advice and decide on the solution. It's much like we visit the doctor today, who is a subject-matter expert.

As the complexity of business systems went up and we entered an era of continuous change, people found it hard to deal with such complex systems that had never existed before. The human brain is much better at working with fewer variables than many. Also, people started using computers, which are relatively better and unbiased when it comes to new forms and large volumes of data.

An Example

The next question often is, what do I mean by "use of numbers"? Will you have do math again?

The last decade has seen the advent of software as a service (SaaS) in all walks of information gathering and manipulation. Thus, analytics systems now are button-driven systems that do the calculations and provide the results. An analyst or data scientist has to look at these results and make recommendations for the business to implement. For example, say a bank wants to sell loans in the market. It has data of all the customers who have taken loans from the bank over the last 20 years. The portfolio is of, say, 1 million loans. Using this data, the bank wants to understand which customers it should give pre-approved loan offers to.

Electronic supplementary material The online version of this chapter (doi: 10.1007/978-1-4842-1001-7_1) contains supplementary material, which is available to authorized users.

© Subhashini Sharma Tripathi 2016

S. S. Tripathi, *Learn Business Analytics in Six Steps Using SAS and R*, DOI 10.1007/978-1-4842-1001-7_1

The simplest answer may be as follows: all the customers who paid on time every time in their earlier loans should get a pre-approved loan offer. Let's call this set of customers Segment A. But on analysis, you may find that customers who defaulted but paid the loan after the default actually made more money for the bank because they paid interest plus the late payment charges. Let's call this set Segment B.

Hence, you can now say that you want to send out an offer letter to Segment A + Segment B.

However, within Segment B there was a set of customers who you had to send collections teams to their homes to collect the money. So, they paid interest plus the late payment charges minus the collection cost. This set is Segment C.

So, you may then decide to target Segment A + Segment B – Segment C.

You could do this exercise using the decision tree technique that cuts your data into segments (Figure 1-1).

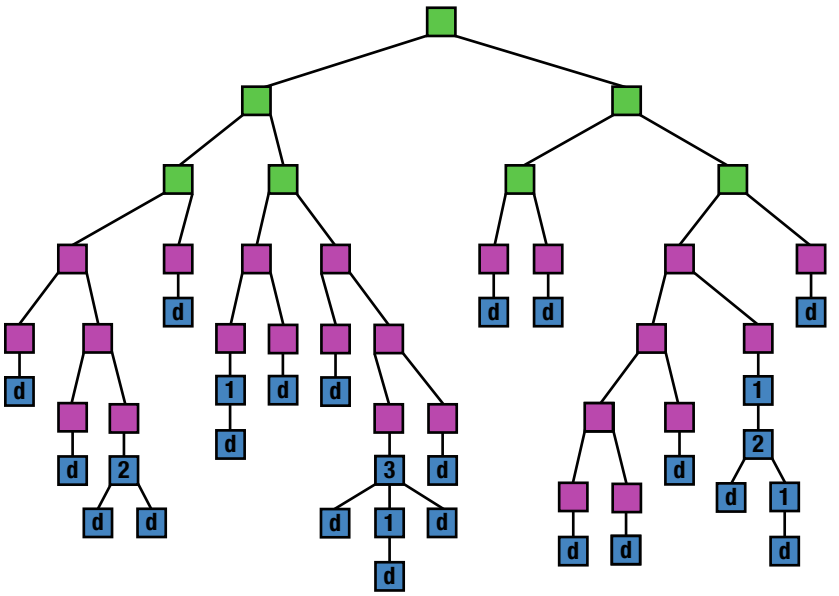


Figure 1-1. *Decision tree*

A Typical Day

The last question to tackle is, what does the workday of an analytics professional look like? It probably encompasses the following:

- The data analyst will walk into the office and be told about the problem that the business needs input on.
- The data analyst will determine the best way to solve the problem.
- The data analyst will then gather the relevant data from the large data sets stored in the server.
- Next, the data analyst will import the data into the analytics software.
- The data analyst will run the technique through the software (SAS, R, SPSS, XLSTAT, and so on).
- The software will produce the relevant output.

- The data analyst will study the output and prepare a report with recommendations.
- The report will be discussed with the business.

Is Analytics for You?

So, is analytics the right career for you? Here are some points that will help you decide:

- *Do you believe that data should be the basis of all decisions?* Take up analytics only if your answer to this question is an unequivocal yes. Analytics is the process of using and analyzing a large quantum of data (numbers, text, images, and so on) by aggregating, visualizing/creating dashboards, checking repetitive trends, and creating models on which decisions can be made. Only people who innately believe in the power of data will excel in this field. If some prediction/analysis is wrong, the attitude of a good analyst is that it is because the data was not appropriate for the analysis or the technique used was incorrect. You will never doubt that a correct decision will be made if the relevant data and appropriate techniques are used.
- *Do you like to constantly learn new stuff?* Take up analytics only if your answer to this question is an unequivocal yes. Analytics is a new field. There is a constant increase in the avenues of data currently regarding Internet data, social networking information, mobile transaction data, and near field communication devices. There are constant changes in technology to store, process, and analyze this data. Hadoop, Google updates, and so on, have become increasingly important. Cloud computing and data management are common now. Economic cycles have shortened, and model building has become more frequent as older models get redundant. Even the humble Excel has an Analysis ToolPak in Excel 2010 with statistical functions. In other words, be ready for change.
- *Do you like to interpret outcomes and then track them to see whether your recommendations were right?* Take up analytics only if your answer to this question is an unequivocal yes. A data analyst will work on a project, and the implementation of the recommendations will generally be valid for a reasonably long period of time, perhaps a year or even three to five years. A good analyst should be interested to know how accurate the recommendations have been and should want to track the performance periodically. You should ideally also be the first person to be able to say when the analysis is not working and needs to be reworked.
- *Are you ready to go back to a text book and brush up on the concepts of math and statistics?* Take up analytics only if your answer to this question is an unequivocal yes. To accurately handle data and interpret results, you will need to brush up on the concepts of math and statistics. It becomes important to justify why you chose a particular path during analysis versus others. Business users will not accept your word blindly.
- *Do you like debating and logical thinking?* Take up analytics only if your answer to this question is an unequivocal yes. As there is no one solution to all problems, an analyst has to choose the best way to handle the project/problem at hand. The analyst has to be able to not only know the best way to analyze the data but also give the best recommendation in the given time constraints and budget constraints. This sector generally has a very open culture where the analyst working on a project/problem will be required to give input irrespective of the analyst's position in the hierarchy.

Do check your answers to the previous questions. If you said yes for three out of these five questions and an OK for two, then analytics is a viable career option for you. Welcome to the world of analytics!

Evolution of Analytics: How Did Analytics Start?

As per the Oxford Dictionary, the definition of *statistics* is as follows:

> *The practice or science of collecting and analyzing numerical data in large quantities, especially for the purpose of inferring proportions in a whole from those in a representative sample.*[1]

Most people start working with numbers, counting, and math by the time we are five years old. Math includes addition, subtraction, theorems, rules, and so on. Statistics is when we start using math concepts to work on real-life data.

Statistics is derived from the Latin word *status*, the Italian word *statista*, or the German word *statistik*, each of which means a political state. This word came into being somewhere around 1780 to 1790.

In ancient times, the government collected the information regarding the population, property, and wealth of the country. This enabled the government to get an idea of the manpower of the country and became the basis for introducing taxes and levies. Statistics are the practical part of math.

The implementation of standards in industry and commerce became important with the onset of the Industrial Revolution, where there arose a need for high-precision machine tools and interchangeable parts. *Standardization* is the process of developing and implementing technical standards. It helps in maximizing compatibility, interoperability, safety, repeatability, and quality.

Nuts and bolts held the industrialization process together; in 1800, Henry Maudslay developed the first practical screw-cutting lathe. This allowed for the standardization of screw thread sizes and paved the way for the practical application of interchangeability for nuts and bolts. Before this, screw threads were usually made by chipping and filing manually.

Maudslay standardized the screw threads used in his workshop and produced sets of nuts and bolts to those standards so that any bolt of the appropriate size would fit any nut of the same size.

Joseph Whitworth's screw thread measurements were adopted as the first unofficial national standard by companies in Britain in 1841 and came to be known as the British standard Whitworth.

By the end of the 19th century, differences and standards between companies were making trading increasingly difficult. The Engineering Standards Committee was established in London in 1901 and by the mid-to-late 19th century, efforts were being made to standardize electrical measurements. Many companies had entered the market in the 1890s, and all chose their own settings for voltage, frequency, current, and even the symbols used in circuit diagrams, making standardization necessary for electrical measurements.

The International Federation of the National Standardizing Associations was founded in 1926 to enhance international cooperation for all technical standards and certifications.

The Quality Movement

Once manufacturing became an established industry, the emphasis shifted to minimizing waste and therefore cost. This movement was led by engineers who were, by training, adept at using math. This movement was called the *quality movement*. Some practices that came from this movement are Six Sigma and just-in-time manufacturing in supply chain management. The point is that all this started in the Industrial Revolution in 1800s.

This was followed with the factory system with its emphasis on product inspection.

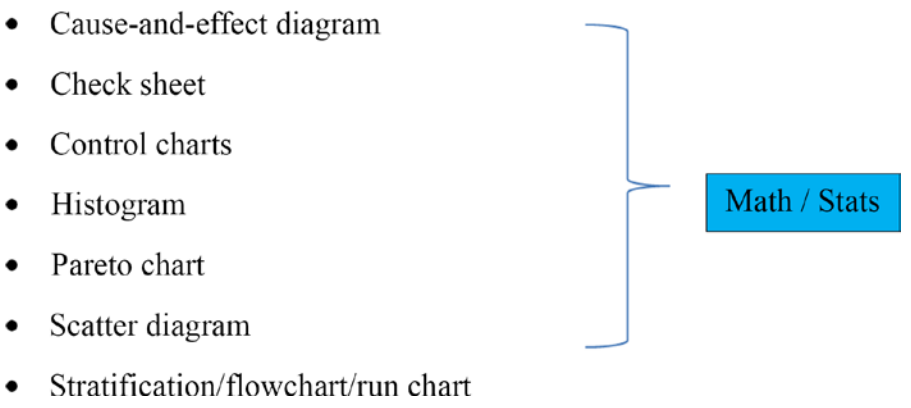
[1] www.oxforddictionaries.com/definition/english/statistics

After the United States entered World War II, the quality became a critical component since bullets from one state had to work with guns manufactured in another state. For example, the U.S. Army had to inspect manually every piece of machinery, but this was very time-consuming. Statistical techniques such as sampling started being used to speed up the processes.

Japan around this time was also becoming conscious of quality.

The quality initiative started with a focus on defects and products and then moved on to look at the process used for creating these products. Companies invested in training their workforce on Total Quality Management (TQM) and statistical techniques.

This phase saw the emergence of seven "basic tools" of quality.



Statistical Process Control from the early 1920s is a method of quality control using statistical methods, where monitoring and controlling the process ensures that it operates at its full potential. At its full potential, a process can churn out as much conforming product or standardize a product as much as possible with a minimum of waste.

This is used extensively in manufacturing lines with a focus on continuous improvement and is practiced in these two phases:

- Initial establishment of the process
- Regular production use of the process

The advantage of Statistical Process Control (SPC) over the methods of quality control such as inspection is that it emphasizes early detection and prevention of problems rather than correcting problems after they occur.

The following were the next steps:

- *Six Sigma*: A process of measurement and improvement perfected by GE and adopted by the world
- *Kaizen*: A Japanese term for continuous improvement; a step-by-step improvement of business processes
- *PDCA*: Plan-Do-Check-Act, as defined by Deming

What was happening on the government front? The maximum data was being captured and used by the military. A lot of the business terminologies and processes used today have been copied from the military: sales *campaigns*, marketing *strategy*, business *tactics*, business *intelligence*, and so on.

The Second World War

As mentioned, statistics made a big difference during World War II. For instance, the Allied forces accurately estimated the production of German tanks using statistical methods. They also used statistics and logical rules to decode German messages.

The Kerrison Predictor was one of the fully automated anti-aircraft fire control systems that could gun an aircraft based on simple inputs such as the angle to the target and the observed speed. The British Army used this effectively in the early 1940s.

The Manhattan Project was a U.S. government research project in 1942–1945 that produced the first atomic bomb. Under this, the first atomic bomb was exploded in July 1945 at a site in New Mexico. The following month, the other atomic bombs that were produced by the project were dropped on Hiroshima and Nagasaki, Japan. This project used statistics to run simulations and predict the behavior of nuclear chain reactions.

Where Else Was Statistics Involved?

Weather predictions, especially rain, affected the world economy the most since weather affected the agriculture industry. The first attempt was made to forecast the weather numerically in 1922 by Lewis Fry Richardson.

The first successful numerical prediction was performed using the ENIAC digital computer in 1950 by a team of American meteorologists and mathematicians.[2]

Then, 1956 saw analytics solve the shortest-path problem in travel and logistics, radically changing these industries.

In 1956 FICO was founded by engineer Bill Fair and mathematician Earl Isaac on the principle that data used intelligently can improve business decisions. In 1958 FICO built its first credit scoring system for American investments, and in 1981 the FICO credit bureau risk score was introduced.[3]

Historically, by the 1960s, most organizations had designed, developed, and implemented centralized computing systems for inventory control. Material requirements planning (MRP) systems were developed in the 1970s.

In 1973, the Black-Scholes model (or Black–Scholes–Merton model) was perfected. It is a mathematical model of a financial market containing certain derivative investment instruments. This model estimates the price of the option/stock overtime. The key idea behind the model is to hedge the option by buying and selling the asset in just the right way and thereby eliminate risk. It is used by investment banks and hedge funds.

By the 1980s, manufacturing resource planning systems were introduced with the emphasis on optimizing manufacturing processes by synchronizing materials with production requirements. Starting in the late 1980s, software systems known as enterprise resource planning systems became the drivers of data accumulation in business. ERP systems are software systems for business management including models supporting functional areas such as planning, manufacturing, sales, marketing, distribution, accounting, and so on. ERP systems were a leg up over MRP systems. They include modules not only related to manufacturing but also to services and maintenance.

[2] http://journals.ametsoc.org/doi/pdf/10.1175/BAMS-89-1-45
[3] www.fico.com/en/about-us#our_history

The Dawn of Business Intelligence

Typically, early business applications and ERP systems had their own databases that supported their functions. This meant that data was in silos because no other system had access to it. Businesses soon realized that the value of data can increase manyfold if all the data is in one system together. This led to the concept of a data warehouse and then an enterprise data warehouse (EDW) as a single system for the repository of all the organization's data. Thus, data could be acquired from a variety of incompatible systems and brought together using extract, transform, load (ETL) processes. Once the data is collected from the many diverse systems, the captured data needs to be converted into information and knowledge in order to be useful. The business intelligence (BI) systems could therefore give much more coherent intelligence to businesses and introduce the concepts of one view of customers and customer lifetime value.

One advantage of an EDW is that business intelligence is now much more exhaustive. Though business intelligence is a good way to use graphs and charts to get a view of business progress, it does not use high-end statistical processes to derive greater value from the data.

The next question that business wanted to answer by the 1990s–2000 was how the data can be used more effectively to understand embedded trends and predict future trends. The business world was waking up to *predictive analytics*.

What are the types of analytics that exist now? The analytics journey generally starts off with the following:

- *Descriptive statistics*: This enables businesses to understand summaries generally about numbers that the management views as part of the business intelligence process.
- *Inferential statistics*: This enables businesses to understand distributions and variations and shapes in which the data occurs.
- *Differences statistics*: This enables businesses to know how the data is changing or if it's the same.
- *Associative statistics*: This enables businesses to know the strength and direction of associations within data.
- *Predictive analytics*: This enables businesses to make predictions related to trends and probabilities.

Fortunately, we live in an era of software, which can help us do the math, which means analysts can focus on the following:

- Understanding the business process
- Understanding the deliverable or business problem that needs to be solved
- Pinpointing the technique in statistics that will be used to reach the solution
- Running the SaaS to implement the technique
- Generating insights or conclusions to help the business

CHAPTER 2

Accessing SAS and R

This chapter gives you an introduction to the popular software called SAS and R. It will cover how to install them and get started using them.

Why SAS and R?

Let's first look at the market reality, as mentioned by Gartner in its 2015 report called "Magic Quadrant for Advanced Analytics Platforms." You can find a copy of this report on the Gartner web site at www.gartner.com/technology/research.jsp.

Market Overview

Gartner estimates that the advanced analytics market amounts to more than $1 billion across a wide variety of industries and geographies. Financial services, retail/e-commerce, and communications are probably the largest industries, although use cases exist in almost every industry. North America and Europe are the largest geographical markets, although Asia/Pacific is also growing rapidly.

This market has existed for more than 20 years. The concept of big data not only has increased interest in this market but has significantly disrupted it. The following are key disruptive trends cited by Gartner:

- The growing interest in applying the results of advanced analytics to improve business performance is rapidly expanding the number of potential applications of this technology and its audience across organizations. Rather than being the domain of a few select groups (for example, those responsible for marketing and risk management), every business function now has a legitimate interest in this capability.
- The rapid growth in the amount of available data, particularly new varieties of data (such as unstructured data from customer interactions and streamed machine-generated data), requires greater levels of sophistication from users and systems, as well as the ability to rapidly interpret and respond to data to realize its full potential.
- The growing demand for these types of capabilities is outpacing the supply of expert users, which necessitates higher levels of automation and increases demand for self-service and citizen data scientist tools.

© Subhashini Sharma Tripathi 2016
S. S. Tripathi, *Learn Business Analytics in Six Steps Using SAS and R*, DOI 10.1007/978-1-4842-1001-7_2

What Is Advanced Analytics?

Gartner defines *advanced analytics* as the analysis of all kinds of data using sophisticated quantitative methods (for example, statistics, descriptive and predictive data mining, simulation, and optimization) to produce insights that traditional approaches to business intelligence (BI)—such as query and reporting—are unlikely to discover.

I find this last part to be significant. Advanced analytics is about using methods beyond BI that involve statistics and data mining.

As mentioned, SAS and R are the leaders in the categories of licensed software and free, open source languages, respectively. Thus, if we as analysts can work with both of these languages, we can be assured of being employable for a large set of projects and companies in analytics.

Here are some other points worth noting:

- *The habit of SAS is hard to break*: Traditionally SAS has been the language of analytics, and years of code has been written and perfected in SAS. For an industry to overthrow all of these established processes and start off with R is difficult.
- *Distrust of freeware is high*: Businesses feel comfortable working on products that they pay for and have customer support for. R is free software (though there are many web forums to focus on it). Tech support is available for paid versions such as Revolution Analytics. (Refer to `www.revolutionanalytics.com/why-revolution-analytics` for more information on the consulting and tech support services from Revolution Analytics.)
- *R has in-memory processing*: Since R works on in-memory processing, there are several issues related to big data processing. However, enterprise versions and RHadoop have offset these limitations. (Enterprise versions of R are not free.)
- *Coding intensity is higher in R while SAS has invested in a lot of point-and-click interfaces such as E Miner and EG*. SAS also has many customized suits for specific business requirements and functions, making it easier to deploy.

Industry-specific solutions exist for the following industries in SAS:

- Automotive
- Banking
- Capital markets
- Casinos
- Communications
- Consumer goods
- Defense and security
- Government
- Healthcare providers
- Health insurance
- High-tech manufacturing
- Higher education
- Hotels

- Insurance
- K-12 education
- Life sciences
- Manufacturing
- Media
- Oil and gas
- Retail
- Small and midsize business
- Sports
- Travel and transportation
- Utilities

■ **Tip** Read more at the SAS web site at `www.sas.com/en_us/industry.html`.

History of SAS and R

I am sure you are curious to understand how SAS and R evolved. Let's look at their histories.

History of SAS

SAS is definitely the tried-and-tested superstar of the analytics industry. In 1966 there was a need for a computerized statistics program to analyze agricultural data collected by the U.S. Department of Agriculture. The U.S. Department of Agriculture was funding the research for a consortium of eight land-grant universities, and these schools came together under a grant from the National Institute of Health to develop a general-purpose statistical software package for the analysis of agricultural data to improve crop yield. The resulting program was called the *statistical analysis system*, and the acronym SAS arose from the name.

Out of the eight universities, North Carolina State University became the leader of the consortium because it had access to a more powerful mainframe computer compared to other universities.

North Carolina State University faculty members Jim Goodnight and Jim Barr were the project leaders. When the National Institute of Health discontinued funding in 1972, members of the consortium agreed to chip in money each year to allow North Carolina State University to continue developing and maintaining the system and supporting the statistical analysis needs. In 1976, the team working on SAS took the project out of the university and incorporated the SAS Institute. In 1985, SAS was rewritten in the C programming language, and the science enterprise Miner was released in 1999. As the name suggests, it was the start of SAS creating suites of products for solving specific business problems, whereas Enterprise Miner was aimed at mining large data sets. In 2002, the Text Minor software was introduced. Today SAS products include the following:

- SAS 9.4 (base SAS)
- SAS/STAT
- SAS Analytics Pro

- SAS Curriculum Pathways
- SAS Data Management
- SAS Enterprise Miner
- SAS Marketing Optimization
- SAS University Edition
- SAS Visual Analytics
- SAS Visual Statistics

What I will cover here is base SAS (which will enable you to write code in SAS), and I will use SAS Enterprise Guide (EG) as the platform so that you also get exposure to the point-and-click functionalities.

What Is EG?

SAS Enterprise Guide provides an intuitive project-based programming and point-and-click interface to SAS. It includes an intelligent program editor, querying capabilities, repeatable process flows, stored process creation and consumption, and a multitude of other features. It allows for point-and-click tasks and the editing of the code for these tasks. Thus, it allows for much less code writing. As an analyst, if you understand the construct of the code and can edit the code to create customized outputs, the time savings is huge. Also, you break up the repetition and monotony.

The other benefit is that noncoders can work more efficiently. Thus, people like me are much more comfortable using it.

How Can You Access SAS Enterprise Guide Software?

SAS has created a SAS on-demand facility for academic users and students. View it and install SAS on your system by visiting `http://support.sas.com/software/products/ondemand-academics/#s1=2`.

Otherwise, just Google *SAS on Demand for Academics* and you will be able to see all the relevant links.

■ **Tip** SAS continues to update versions of the software and sometimes the look and feel of the SAS on-demand site. Don't be surprised to see changes. It's a good way to get an idea of new products or new versions of products that SAS releases.

History of R

In 1975–76, Bell Laboratories designed S, which is a statistical computing language, as an alternative to the common statistical computing that was done by directly calling Fortran subroutines.

In 1995, Ross Ihaka and Robert Gentleman wrote an experimental R, which was "not unlike S." In the last two decades, R has emerged as a software application for statistics, data management, programming, and so on, which exists in a unique quantity and variety. The quality varies, but on average the output is impressive. Most of this is in an open environment that encourages improvements and has wide participation from the statistics profession.

R is freely available under the GNU General Public License; check out `www.r-project.org/`.