A Ohri

R for Business Analytics

Springer

R for Business Analytics

A. Ohri

R for Business Analytics

Springer

A. Ohri
Founder-Decisionstats.com
Delhi, India

ISBN 978-1-4614-4342-1 ISBN 978-1-4614-4343-8 (eBook)
DOI 10.1007/978-1-4614-4343-8
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2012941362

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Dedicated to Dad (A. K. Ohri)
and Father (Jesus Christ)
and my 5 year old son Kush

Foreword

I basically structured the book according to the tasks that I have been doing most frequently in my decade-long career as a business analyst. The chapters are thus divided into most frequently used tasks, and I have added references to multiple sources to help the reader explore a particular subject in more depth. Again, I emphasize that this is a business analytics book, not for statistics, and my own experience as an MBA and with the literature available for MBAs in business analytics (particularly R) led me to these choices.

This book is thus organized for a business analyst rather than a statistician. It will not help you get a better grade on your graduate school thesis, but it will definitely help you get, or retain, a job in analytics. If you are a student studying R, it should help you do your homework faster.

In the current business environment, I believe that focus will shift back to the analyst rather than the software tool, and having multiple platform skills, especially in both high-end and low-cost analytical platforms, can be of some benefit to the user.

This book will focus explicitly on graphical user interfaces, tricks, tips, techniques, and shortcuts, and focus on case studies from the most commonly used tasks that a business analyst will face on a day-to-day basis. Things will be made as simple as possible but no simpler than that. Each chapter will have a case study, tutorial, or example problem. Functions and packages used in each chapter will be listed at the end to help the reader review. There might be times when some functions appear to have been repeated or stressed again; this has more to do with their analytical use and necessity. The brief interviews with creators, authors, and developers is aimed at making it easy for the business analyst to absorb aspects of R. The use cases of existing R deployments are designed to help decision makers within the analytics community to give R a chance, if they have not done so already.

The book has a pragmatic purpose and is aimed at those using or hoping to use R in a corporate business environment. Adequate references will be provided to help the reader with theoretical aspects or advanced levels and to assist the reader on his or her journey in R with other available resources.

Readers are encouraged to skip chapters that have no immediate relevance to them and go directly to those that are of maximum utility for their purposes. One issue that I faced was that the R project released almost four new versions of the software by the time I wrote the book, so please feel free to let me know about any inaccuracies or errata.

Organization of the Book

Chapter 1: Why R? Reasons for Using R in Business Analytics. In this chapter we discuss reasons for choosing R as an analytical and not just a statistical computing platform, comparisons with other analytical software, and some general costs and benefits in using R in a business environment. It lists the various reasons R should be chosen for learning by business analysts and the additional features that R has over other analytical platforms. The benefits of R are subdivided into three major categories: business analytics, data mining, and business intelligence/data visualization.

Chapter 2: R Infrastructure: Setting up Your R Analytical Infrastructure. In this chapter we discuss the practical realities in setting up an analytical environment based on R, including which hardware, operating system, additional software, requisite budgeting, and the training and software support needs. We discuss the various operating systems, hardware choices, and various providers of R-based solutions. We walk through the basics of installing R, R's library of packages, updating R, and accessing the comprehensive user help and a simple basic tutorial for starting R.

Chapter 3: R Interfaces: Ways to Use Your R Analytics Based on Your Needs. In this chapter we compare and contrast the various ways to interact with the R analytical platform. R can be used in a command line, with graphical user interfaces (GUIs), and via Web interfaces (including cloud computing interfaces). The chapter outlines both the advantages and disadvantages of a GUI-based approach. The chapter compares the features of nine kinds of GUI, including a summary sheet of comparative advantages and disadvantages. It also discusses using R from other software and from Web interfaces. Lastly, the chapter includes some useful tutorials on running R from the Amazon cloud.

Chapter 4: Manipulating Data: Obtaining Your Data Within R. This chapter talks of various ways of obtaining data within R, including the basic syntax. It deals specifically with data in databases as this is often the case for business data. The chapter shows how the user can connect to MySQL and Pentaho databases, which are the two leading open source databases used. Specific sections are devoted to using SQLite with R and to business intelligence practitioners. We take a brief look at Jaspersoft and Pentaho, the leading open source solutions within BI, and how they interact with R. While the chapter briefly mentions additional resources to handle larger datasets, it also gives a series of common analytical tasks that a business analyst is expected to perform on any data to help someone transitioning to R.

Chapter 5: Exploring Data—The Booming Business in Data Visualization. This chapter discusses exploring data in R using visual and graphical means. It talks about the basic graphs and a series of advanced graphs in R that can be easily created by an average programmer learning R in a very short time. It introduces a specialized GUI for data exploration, grapheR and Deducer, and also features ggplot2 creator Hadley Wickham 's interview. Aspects of graphs include code, easy-to-recreate examples, and information on interactive graphs. The chapter is aimed at demystifying the sometimes intimidating art of data visualization for someone who has been creating graphs mostly using spreadsheet programs.

Chapter 6: Building Regression Models. Regression models are the statistical workhorses of the business analytics industry. They are used perhaps to the point of overuse because of the inherent simplicity in communicating them to business audiences internally. We learn how to build linear and logistic regression models, study some risk models and scorecards, and discuss PMML as the way to implement models. We also have a brief case study for simplifying the process of building logistic models in R, and no theory of regression is introduced, in keeping with the business analytics focus of the book.

Chapter 7: Data Mining Using R. Data mining using R employs the Rattle GUI for simplifying and speeding the process of data modeling in R. However, it begins by introducing the concepts of information ladder and various data mining methodologies to the reader including, briefly, CRISP-DM, SEMMA, and KDD. It also features extracts from brief interviews with two authors who have written books on data mining using R. As special cases, text mining, Web mining, and the Google Prediction application interface (API) are featured.

Chapter 8: Data Segmentation. Data segmentation in this book deals mainly with cluster analysis, and we walk through the various types of clustering. Clustering is added here because of the inherent and increasing need for data reduction techniques in business environments of big data and because the size of datasets is increasing rapidly. We refer once again to the Rattle GUI but also briefly touch on other clustering GUIs in R. A small case study is presented regarding the use of Revolution R for clustering large amounts of data.

Chapter 9: Forecasting and Time Series. While business uses business intelligence and reporting for knowing the past and present of operations, their focus is to improve decision making for the future. Powerful but underutilized in many businesses, time series and forecasting are explained here with a case study and an R Commander epack GUI. I have tried to make it a practical chapter to help your business team do more forecasting utilizing the nascent data present in all organizations.

Chapter 10: Data Export and Output. Obtaining results for analytics is only part of the job. Output should be presented in a manner that convinces decision makers to take actionable decisions R provides many flexible ways to generate and embed output, and they are presented here.

Chapter 11: Optimizing R Coding. Now that you have learned how to functionally use R for business analytics, the next step is to understand how to utilize its

powerful flexibility without drowning in the huge library available. This chapter discusses tips, tweaks, and tools including code editors to help you code better and faster.

Chapter 12: Additional Training Literature. This chapter is meant as a follow-up for readers interested in expanding their R knowledge and in a more structured view of the documentation environment of R.

Chapter 13: Case Studies Using R. This chapter presents coding case studies based on various business uses, including Web analytics, and is designed to help readers find a ready reference for using R for business analytics in their operational business context.

Preface

There are 115 books listed on the R project Web site, more books have been published since that list was last updated, and if you are wondering why one more is needed, well, I am here to explain why. I have been working in the field known as business analytics for nearly a decade now. You may the know the field as statistical computing, data mining, business intelligence, or, lately, data science. I generally prefer the term decision sciences, but overall the field refers to using data, statistical techniques, and specialized tools to assist decision makers in government, research, and industry with insights to maximize positive outcomes and minimize costs.

I found the field of business analytics to be both very enjoyable and lucrative from a career point of view. Selling credit cards and loans and getting more revenue from the same people was a lot of fun, at least in the early years for me. The reason I found business analytics interesting was that it combined the disciplines of investigative and diligent thinking with business insights.

I also found the field of business analytics a bit confusing. There were two kinds of people with almost opposing views on how to apply business analytics: technically minded people (like computer science and statistics majors) who aimed for robust results and business-minded people (like MBAs) who aimed for revenue and quicker results. I was caught in the middle of the occasional crossfire.

When I started learning all this—in 2001—the predominant platform was the SPSS and SAS languages. In 2007, following the birth of my first son, I discovered the R programming language almost as a necessity to keep my fledgling consulting shop open as I needed a reliable analytics platform without high annual fees. In launching an analytics startup at the age of 30, I found that I could not afford the tools that I had been using in some of the world's largest corporations (and I discovered that there were no separate discounts for small enterprises in India). So I moved to R because it was free. Or so I thought.

But R took me a lot of time to learn, and time is not free. Something was wrong – either with me, or the language, or the whole universe. In the meantime, I started writing an analytics blog called DecisionStats (http://decisionstats.com) to network with other business analysts, and with almost 20,000 views every month and nearly

100 interviews with leading analytics practitioners, I slowly obtained more practical insights into the field of business analytics.

Over a period of time, as I slowly immersed myself in R, I discovered that it, like all languages, had its own set of tricks and techniques with tradeoffs to get things done faster, and I did need not to memorize huge chunks of code or be overawed by those who could. I was aided in this by a few good books including R for SAS and SPSS users by Bob Muenchen, great blogs like those by David Smith and Tal G. Overall, I learned R at a much faster pace than I had expected to initially, thanks to Prof. John Fox who made R Commander, Prof. Graham Williams who made Rattle, and many others who have helped to make R what it is today. The R community exploded in size, acceptability, and organization, and I was no longer fighting a lonely battle to learn R on my own for analytical purposes.

In 2012, R has grown at a pace I could not have imagined in 2007. I am both humbled and blessed to know that R is the leading statistical computing language and used and completely supported by leading technology companies including Microsoft, SAS Institute, Oracle, Google, and others. *No R&D budget can compete with nearly ALL the statistics departments of the world and their professors working for free on this project.*

If you are a decision maker thinking of using R in combination with your existing analytical infrastructure, you will find the brief interviews of various partners and contributors to R to be very enlightening and helpful. They have specifically been added to increase the book's readability for a business analytics audience that may prefer reading English to code.

I would like to thank Anne Milley of JMP, Jill Dyche of Baseline Consulting, Bob Muenchen of the University of Tennessee, Karl Rexer of Rexer Analytics, and Gregory Shapiro of KD Nuggets specifically for their help in mentoring me through my analytical wanderings. I would also like to thank Marc Strauss and Susan Westendorf of Springer USA and Leo Augustine for their help.

I am humbled by the patient readers of Decisionstats and appreciate the attention that the R and broader analytics community has shown towards the author.

Delhi, India A. Ohri

Contents

Chapter 1
Why R

Chapter summary: In this chapter we introduce the reader to R, discuss reasons for choosing R as an analytical and not just a statistical computing platform, make comparisons with other analytical software, and present some broad costs and benefits in using R in a business environment.

R is also known as GNU S, as it is basically an open source derivative and descendant of the S language. In various forms and avatars, R has been around for almost two decades now, with an ever growing library of specialized data visualization, data analysis, and data manipulation packages. With around two million users, R has one of the largest libraries of statistical algorithms and packages.

While R was initially a statistical computing language, in 2012 you could call it a complete analytical environment.

1.1 Reasons for Classifying R as a Complete Analytical Environment

R may be classified as a complete analytical environment for the following reasons.

- Multiple platforms and interfaces to input commands: R has multiple interfaces ranging from command line to numerous specialized graphical user interfaces (GUIs) (Chap. 2) for working on desktops. For clusters, cloud computing, and remote server environments, R now has extensive packages including SNOW, RApache, RMpi, R Web, and Rserve.
- Software compatibility: Official commercial interfaces to R have been developed by numerous commercial vendors including software makers who had previously thought of R as a challenger in the analytical space (Chap. 4). Oracle, ODBC, Microsoft Excel, PostgreSQL, MySQL, SPSS, Oracle Data Miner, SAS/IML, JMP, Pentaho Kettle, and Jaspersoft BI are just a few examples of commercial

A. Ohri, *R for Business Analytics*, DOI 10.1007/978-1-4614-4343-8_1,

software that are compatible with R usage. In terms of the basic SAS language, a WPS software reseller offers a separate add-on called the Bridge to R. Revolution Analytics offers primarily analytical products licensed in the R language, but other small companies have built successful R packages and applications commercially.
- Interoperability of data: Data from various file formats as well as various databases can be used directly in R, connected via a package, or reduced to an intermediate format for importing into R (Chap. 2).
- Extensive data visualization capabilities: These include much better animation and graphing than other software (Chap. 5).
- Largest and fastest growing open source statistical library: The current number of statistical packages and the rate of growth at which new packages continue to be upgraded ensures the continuity of R as a long-term solution to analytical problems.
- A wide range of solutions from the R package library for statistical, analytical, data mining, dashboard, data visualization, and online applications make it the broadest analytical platform in the field.

1.2 Additional Advantages of R over Other Analytical Packages

So what all is extra in R? The list below shows some of the additional features in R that make it superior to other analytical software.

- R's source code is designed to ensure complete custom solutions and embedding for a particular application. Open source code has the advantage of being extensively peer-reviewed in journals and the scientific literature. This means bugs will found, information about them shared, and solutions delivered transparently.
- A wide range of training material in the form of books is available for the R analytical platform (Chap. 12).
- R offers the best data visualization tools in analytical software (apart from Tableau Software's latest version). The extensive data visualization available in R comprises a wide variety of customizable graphics as well as animation. The principal reason why third-party software initially started creating interfaces to R is because the graphical library of packages in R was more advanced and was acquiring more features by the day.
- An R license is free for academics and thus budget friendly for small and large analytical teams.
- R offers flexible programming for your data environment. This includes packages that ensure compatibility with Java, Python, and C++.
- It is easy to migrate from other analytical platforms to the R platform. It is relatively easy for a non-R platform user to migrate to the R platform, and there is no danger of vendor lock-in due to the GPL nature of the source code and the open community, the GPL can be seen at http://www.gnu.org/copyleft/gpl.html.

- The latest and broadest range of statistical algorithms are available in R. This is due to R's package structure in which it is rather easier for developers to create new packages than in any other comparable analytics platform.

1.3 Differences Between R as a Statistical Language and R as an Analytical Platform

Sometimes the distinction between statistical computing and analytics does come up. While statistics is a tool- and technique-based approach, analytics is more concerned with business objectives. Statistics are basically numbers that inform (descriptive), advise (prescriptive), or forecast (predictive). Analytics is a decision-making-assistance tool. Analytics on which no decision is to be made or is being considered can be classified as purely statistical and nonanalytical. Thus the ease with which a correct decision can be made separates a good analytical platform from a not-so-good one. The distinction is likely to be disputed by people of either background, and business analysis requires more emphasis on how practical or actionable the results are and less emphasis on the statistical metrics in a particular data analysis task. I believe one way in which business analytics differs from statistical analysis is the cost of perfect information (data costs in the real world) and the opportunity cost of delayed and distorted decision making.

1.4 Costs and Benefits of Using R

The only cost of using R is the time spent learning it. The lack of a package or application marketplace in which developers can be rewarded for creating new packages hinders the professional mainstream programmer's interest in R to the degree that several other platforms like iOS and Android and Salesforce offer better commercial opportunities to coding professionals. However, given the existing enthusiasm and engagement of the vast numbers of mostly academia-supported R developers, the number of R packages has grown exponentially over the past several years. The following list enumerates the advantages of R by business analytics, data mining, and business intelligence/data visualization as these have three different domains in the data sciences.

1.4.1 Business Analytics

R is available for free download.

1. R is one of the few analytical platforms that work on Mac OS.

2. Its results have been established in journals like the *Journal of Statistical Software*, in places such as LinkedIn and Google, and by Facebook's analytical teams.
3. It has open source code for customization as per GPL and adequate intellectual protection for developers wanting to create commercial packages.
4. It also has a flexible option for enterprise users from commercial vendors like Revolution Analytics (who support 64-bit Windows and now Linux) as well as big data processing through its RevoScaleR package.
5. It has interfaces from almost all other analytical software including SAS, SPSS, JMP, Oracle Data Mining, and RapidMiner. Exist huge library of packages is available for regression, time series, finance, and modeling.
6. High-quality data visualization packages are available for use with R.

1.4.2 Data Mining

As a computing platform, R is better suited to the needs of data mining for the following reasons.

1. R has a vast array of packages covering standard regression, decision trees, association rules, cluster analysis, machine learning, neural networks, and exotic specialized algorithms like those based on chaos models.
2. R provides flexibility in tweaking a standard algorithm by allowing one to see the source code.
3. The Rattle GUI remains the standard GUI for data miners using R. This GUI offers easy access to a wide variety of data mining techniques. It was created and developed in Australia by Prof. Graham Williams. Rattle offers a very powerful and convenient free and open source alternative to data mining software.

1.4.3 Business Dashboards and Reporting

Business dashboards and reporting are an essential piece of business intelligence and decision making systems in organizations.

1. R offers data visualization through ggplot, and GUIs such as Deducer, GrapheR, and Red-R can help even business analysts who know none or very little of the R language in creating a metrics dashboard.
2. For online dashboards R has packages like RWeb, RServe, and R Apache that, in combination with data visualization packages, offer powerful dashboard capabilities. Well-known examples of these will be shown later.
3. R can also be combined with Microsoft Excel using the R Excel package to enable R capabilities for importing within Excel. Thus an Excel user with no knowledge of R can use the GUI within the R Excel plug-in to take advantage of the powerful graphical and statistical capabilities.

4. R has extensive capabilities to interact with and pull data from databases including those by Oracle, MySQL, PostGresSQL, and Hadoop-based data. This ability to connect to databases enables R to pull data and summarize them for processing in the previsualization stage.

1.5 Using SAS and R Together

What follows is a brief collection of resources that describe how to use SAS Institute products and R: Base SAS, SAS/Stat, SAS/Graph.

- A great blog on using both SAS and R together is http://sas-and-r.blogspot.com/.
- The corresponding book *SAS and R* http://www.amazon.com/gp/product/1420070576.
- Sam Croker's paper on the use of time series analysis with Base SAS and R: http://www.nesug.org/proceedings/nesug08/sa/sa07.pdf.
- Phil Holland's paper "SAS to R to SAS," available at http://www.hollandnumerics.co.uk/pdf/SAS2R2SAS_paper.pdf, describes passing SAS data from SAS to R, using R to produce a graph, then passing that graph back to SAS for inclusion in an ODS document.
- One of the first books on R for SAS and SPSS users was by Bob Muenchen: http://www.amazon.com/SAS-SPSS-Users-Statistics-Computing/dp/0387094172.
- A free online document by Bob Muenchen for SAS users of R is at https://sites.google.com/site/r4statistics/books/free-version.
- A case study, "Experiences with using SAS and R in insurance and banking," can be found at http://files.meetup.com/1685538/R%20ans%20SAS%20in%20Banking.ppt.
- The document "Doing More than Just the Basics with SAS/Graph and R: Tips, Tricks, and Techniques" is available at http://biostat.mc.vanderbilt.edu/wiki/pub/Main/RafeDonahue/doingmore_currentversion.pdf.
- The paper "Multiple Methods in JMP® to Interact with R" can be downloaded at http://www.nesug.org/Proceedings/nesug10/po/po06.pdf.
- Official documentation on using R from within SAS/IML is available at http://support.sas.com/documentation/cdl/en/imlug/63541/HTML/default/viewer.htm#imlug_r_sect010.htm.

1.6 Brief Interview: Using R with JMP

An indicator of the long way R has come from being a niche player to a broadly accepted statistical computing platform is the SAS Institute's acceptance of R as a complementary language. What follows is a brief extract from a February 2012 interview with researcher Kelci Miclaus from the JMP division at SAS Institute that includes a case study on how adding R can help analytics organizations even more.

Ajay: How has JMP been integrating with R? What has been the feedback from customers so far? Is there a single case study you can point to where the combination of JMP and R was better than either one of them alone?

Kelci: Feedback from customers has been very positive. Some customers use JMP to foster collaboration between SAS and R modelers within their organizations. Many use JMP's interactive visualization to complement their use of R. Many SAS and JMP users use JMP's integration with R to experiment with more bleeding-edge methods not yet available in commercial software. It can be used simply to smooth the transition with regard to sending data between the two tools or to build complete custom applications that take advantage of both JMP and R.

One customer has been using JMP and R together for Bayesian analysis. He uses R to create MCMC chains and has found that JMP is a great tool for preparing data for analysis and for displaying the results of the MCMC simulation. For example, the control chart and bubble plot platforms in JMP can be used to quickly verify convergence of an algorithm. The use of both tools together can increase productivity since the results of an analysis can be achieved faster than through scripting and static graphics alone.

I, along with a few other JMP developers, have written applications that use JMP scripting to call out to R packages and perform analysis like multidimensional scaling, bootstrapping, support vector machines, and modern variable selection methods. These really show the benefit of interactive visual analysis coupled with modern statistical algorithms. We've packaged these scripts as JMP add-ins and made them freely available on our JMP User Community file exchange. Customers can download them and employ these methods as they would a regular JMP platform. We hope that our customers familiar with scripting will also begin to contribute their own add-ins so a wider audience can take advantage of these new tools (see http://www.decisionstats.com/jmp-and-r-rstats/).

Ajay: How is R a complementary fit to JMP's technical capabilities?

Kelci: R has an incredible breadth of capabilities. JMP has extensive interactive, dynamic visualization intrinsic to its largely visual analysis paradigm, in addition to a strong core of statistical platforms. Since our brains are designed to visually process pictures and animated graphics more efficiently than numbers and text, this environment is all about supporting faster discovery. Of course, JMP also has a scripting language (JSL) that allows you to incorporate SAS code and R code and to build analytical applications for others to leverage SAS, R, and other applications for users who don't code or who don't want to code. JSL is a powerful scripting language on its own.

It can be used for dialog creation, automation of JMP statistical platforms, and custom graphic scripting. In other ways, JSL is very similar to the R language. It can also be used for data and matrix manipulation and to create new analysis functions. With the scripting capabilities of JMP, you can create custom applications that provide both a user interface and an interactive visual backend to R functionality.

Alternatively, you could create a dashboard using statistical or graphical platforms in JMP to explore the data and, with the click of a button, send a portion of the data to R for further analysis.

Another JMP feature that complements R is the add-in architecture, which is similar to how R packages work. If you've written a cool script or analysis workflow, you can package it into a JMP add-in file and send it to your colleagues so they can easily use it.

Ajay: What is the official view on R at your organization? Do you think it is a threat or a complementary product or statistical platform that coexists with your offerings?

Kelci: Most definitely, we view R as complementary. R contributors provide a tremendous service to practitioners, allowing them to try a wide variety of methods in the pursuit of more insight and better results. The R community as a whole provides a valued role to the greater analytical community by focusing attention on newer methods that hold the most promise in so many application areas. Data analysts should be encouraged to use the tools available to them in order to drive discovery, and JMP can help with that by providing an analytic hub that supports both SAS and R integration.

Ajay: **Since you do use R, are there any plans to give back something to the R community in terms of your involvement and participation (say at useR events) or sponsoring contests?**

Kelci: We are certainly open to participating in useR groups. At Predictive Analytics World in New York last October, they didn't have a local useR group, but they did have a Predictive Analytics meet-up group comprised of many R users. We were happy to sponsor this. Some of us within the JMP division have joined local R user groups, myself included. Given that some local R user groups have entertained topics like Excel and R, Python and R, databases and R, we would be happy to participate more fully here. I also hope to attend the useR annual meeting later this year to gain more insight on how we can continue to provide tools to help both the JMP and R communities with their work. We are also exploring options to sponsor contests and would invite participants to use their favorite tools, languages, etc. in pursuit of the best model. Statistics is about learning from data, and this is how we make the world a better place.

Citations and References

- R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org
- SAS/IML and JMP are analytical software applications that are © 2011 SAS Institute, SAS Campus Drive, Cary, NC 27513, USA

In the next chapter we will discuss setting up the basic R infrastructure.

Chapter 2
R Infrastructure

Chapter summary: In this chapter we discuss the practical realities in setting up an analytical environment based on R, including hardware, software, budgeting, and training needs. We will also walk through the basics of installing R, R's library of packages, updating R, and accessing the comprehensive user help.

Congratulations if you decided to install R! As choices go, this is the best one in open source statistical software that you could make for at least the next decade.

2.1 Choices in Setting up R for Business Analytics

Some options await you now before you set up your new analytical environment:

2.1.1 Licensing Choices: Academic, Free, or Enterprise Version of R

You can choose between two kinds of R installations. One is free and open source and is available at http://r-project.org; the other is commercial and offered by many vendors including Revolution Analytics. However, there are other commercial vendors too.

Commercial Vendors of R Language Products:

- Revolution Analytics: http://www.revolutionanalytics.com/
- XL Solutions: http://www.experience-rplus.com/
- Information Builder: http://www.informationbuilders.com/products/webfocus/PredictiveModeling.html
- Blue Reference (Inference for R): http://inferenceforr.com/default.aspx
- R for RExcel: http://www.statconn.com/

A. Ohri, *R for Business Analytics*, DOI 10.1007/978-1-4614-4343-8_2,

Enterprise R from Revolution Analytics has a complete R Development environment for Windows including the use of code snippets to make programming faster. Revolution is also expected to make a GUI available by 2012. Revolution Analytics claims several enhancements for its version of R including the use of optimized libraries for faster performance and the RevoScaleR package that uses the xdf format to handle large datasets.

2.1.2 Operating System Choices

Which operating system should the business analyst choose, Unix, Windows, or Mac OS? Often the choice is dictated by the information technology group in the business. However, we compare some of the advantages and disadvantages of each.

1. Microsoft Windows: This remains the most widely used operating system on the planet. If you are experienced in Windows-based computing and are active on analytical projects, it would not make sense for you to move to other operating systems unless there are significant cost savings and minimal business disruption as a result of the transition. In addition, compatibility issues are minimal for Microsoft Windows, and extensive help documentation is available. However, there may be some R packages that would not function well under Windows; in that case, a multiple operating system is your next option.
2. MacOS and iOS: The reasons for choosing MacOS remain its considerable appeal in esthetically designed software and performance in art or graphics related work, but MacOS is not a standard operating system for enterprise systems or statistical computing. However, open source R claims to be quite optimized and can be used for existing Mac users.
3. Linux: This is the operating system of choice for many R users due to the fact that it has the same open source credentials and so is a much better fit for all R packages. In addition, it is customizable for large-scale data analytics. The most popular versions of Linux are Ubuntu/Debian, Red Hat Enterprise Linux, OpenSUSE, CentOS, and Linux Mint.
   (a) Ubuntu Linux is recommended for people making the transition to Linux for the first time. Ubuntu Linux had a marketing agreement with Revolution Analytics for an earlier version of Ubuntu, and many R packages can be installed in a straightforward way. Ubuntu/Debian packages are also available.
   (b) Red Hat Enterprise Linux is officially supported by Revolution Analytics for its enterprise module.
4. Multiple operating systems

Virtualization versus dual boot: if you are using more than two operating systems on your PC. You can also choose between having VMware Player from VMware

(http://www.vmware.com/products/player/), if you want a virtual partition on your computer that is dedicated to R-based computing, and having a choice of operating system at startup. In addition, you can dual boot your computer with a USB installer from Ubuntu's Netbook remix (http://www.ubuntu.com/desktop/get-ubuntu/windows-installer).

A software program called wubi helps with the dual installation of Linux and Windows.

2.1.3 Operating System Subchoice: 32- or 64-bit

Given a choice between a 32-bit versus 64-bit version of an operating system like Linux Ubuntu, keep in mind that the 64-bit version would speed up processing by an approximate factor of 2. However, you need to check whether your current hardware can support 64-bit operating systems; if so, you may want to ask your information technology manager to upgrade at least some of the operating systems in your analytics work environment to 64-bit versions. Smaller hardware like netbooks do not support 64-bit Linux, whereas Windows Home Edition computers may have 32-bit version installed on it. There are cost differences due to both hardware and software. One more advantage for 64-bit computing is the support from Revolution Analytics for its version of R Enterprise.

2.1.4 Hardware Choices: Cost-Benefit Tradeoffs for Additional Hardware for R

At the time of writing of this book, the dominant computing paradigm is workstation computing followed by server–client computing. However, with the introduction of cloud computing, netbooks, and tablet PCs, hardware choices are much more flexible in 2011 than just a couple years ago.

Hardware costs represent a significant expense for an analytics environment and are also remarkably depreciated over a short period of time. Thus, it is advisable to examine your legacy hardware and your future analytical computing needs and decide accordingly regarding the various hardware options available for R.

Unlike other analytical software that can charge by the number of processors, or servers, which can be more expensive than workstations, or grid computing, which can be very costly as well if it is even available, R is well suited for all kinds of hardware environments with flexible costs.

Given the fact that R is memory intensive (it limits the size of data analyzed to the RAM size of the machine unless special formats or chunking is used), the speed at which R can process data depends on the size of the datasets used and the number of users analyzing a dataset concurrently. Thus the defining issue is not R but the

size of the data being analyzed and the frequency, repeatability, and level of detail of analysis required.

2.1.4.1 Choices Between Local, Cluster, and Cloud Computing

- Local computing: This denotes when the software is installed locally. For big data, the data to be analyzed are stored in the form of databases. The server version—Revolution Analytics has differential pricing for server–client versions (as is true for all analytical software pricing), but for the open source version it is free, as it is for server or workstation versions. The issue of number of servers versus workstations is best determined by the size of the data. R processes data in RAM, so it needs more RAM than other software of its class.

Cloud computing is defined as the delivery of data, processing, and systems via remote computers. It is similar to server–client computing, but the remote server (also called the cloud) has flexible computing in terms of number of processors, memory, and data storage. Cloud computing in the form of a public cloud enables people to do analytical tasks on massive datasets without investing in permanent hardware or software as most public clouds are priced on pay per usage. The biggest cloud computing provider is Amazon, and many other vendors provide services on top of it. Google also does data storage in the form of clouds (Google Storage) and uses machine learning in the form of an API (Google Prediction API).

1. *Amazon*: We will describe how to set up an R session on an Amazon EC2 machine.
2. *Google*: We will describe how to use Google Cloud Storage as well as Google Prediction API using packages.
3. *Cluster-grid computing/parallel processing*: To build a cluster, you need the RMpi and SNOW packages, plus other packages that help with parallel processing. This will be covered in general detail but detailed instructions for building a big cluster will not be provided as that is more suitable for a high-performance computing environment.

2.1.5 Interface Choices: Command Line Versus GUI. Which GUI Should You Choose as the Default Startup Option?

R can be used in various ways depending on the level of customization. The main GUIs suitable for business analyst audiences are as follows:

1. R Commander
2. Rattle
3. Deducer and JGR
4. GrapheR

5. RKWard
6. Red-R
7. Others including Sciviews-K

The interfaces to R will be covered in detail in Chap. 3, where a detailed description will also be given of how to access R from other mainstream analytical software applications like Oracle Data Miner, JMP, SAS/IML, KNIME, and Microsoft Excel. In addition to the standard desktop GUI, there are Web interfaces that use R and command line for default coding.

2.1.6 Software Component Choice

Which R packages should you install? There are almost 3,000 packages, some of them are complementary, others depend on each other, and almost all are free.

Throughout this book we will describe specialized packages that are best suited for creating the results of certain analytical tasks. In the R Programming language, multiple approaches, code, functions, and packages can be used to achieve the same result. The objective of this book is to focus on analysis rather than the language, and accordingly we will indicate the easiest approach to accomplishing the given business analysis task and mention other options and the advantages and disadvantages of using multiple options and approaches.

2.1.7 Additional Software Choices

What other applications do you need to achieve maximum accuracy, robustness, and speed of computing and how do you make use of existing legacy software and hardware to achieve the best complementary results with R?

Once we have covered the basics, we will describe, in Chap. 11, additional tips, tricks, and tweaks to help you optimize your R code. These include setting up benchmarks to measure and improve code efficiency and using syntax editors and integrated development environments.

2.2 Downloading and Installing R

To download and install the open source version of R, visit R's home page at http://www.r-project.org/.

You will be directed to the CRAN mirror closest to your location. CRAN, which stands for Comprehensive R Archive Network, is a set of online mirror servers that enable you to download R and its various packages. The global network thus ensures a fast, dedicated, reliable network for downloading and accessing software. In this manner, CRAN guarantees the highest likelihood of availability of R as it is very difficult to bring down servers of the entire CRAN, but an isolated server might get

overwhelmed due to traffic (especially at new product release times). It consists of 79 sites in 34 regions. R can be downloaded from http://cran.r-project.org/mirrors.html.

For Windows-R, installers exist in the form of downloadable binaries. Download the Windows.exe file and install the program. In addition, read the Frequently Asked Questions.

On Linux (Ubuntu): To install the complete R system, open a terminal window and use *sudo apt-get update sudo apt-get install r-base*.

Debian packages for R are a bit dated, but this is the easiest way to install. The other way is to modify your source file with a CRAN mirror before running apt-get. Documentation for this is on the Web site given previously.

Mac OS has a separate, downloadable installer. You will need to refer to the main R Web site http://www.r-project.org.

The Australian CRAN Mirror can be accessed at http://cran.ms.unimelb.edu.au/bin/windows/base/README.R-2.15.1 and FAQs at http://cran.ms.unimelb.edu.au/bin/windows/base/rw-FAQ.html#Introduction.

<CRAN MIRROR>/bin/windows/base/release.htm is the generic link for Windows releases. The latest version was 2.14.1 in January 2012, but this will change every 6 months.

2.3 Installing R Packages

Unlike other traditional software applications that come in bundles, R comes in the form of one installer and a large number of small packages. There are an estimated 3,000 packages in R—so if you have a specific analytical need, chances are someone has created a package already for it. To launch R, simply click the icon that was created (for Windows users) or type R (at command terminal for Linux users).