1

Welcome to R

1.1 What Is R?

As a marketing analyst, you have no doubt heard of R. You may have tried R and become frustrated and confused, after which you returned to other tools that are "good enough." You may know that R uses a command line and dislike that. Or you may be convinced of R's advantages for experts but worry that you don't have time to learn or use it.

We are here to help! Our goal is to present *just the essentials*, in the *minimal necessary time*, with *hands-on learning* so you will come up to speed as quickly as possible to be productive in R. In addition, we'll cover a few advanced topics that demonstrate the power of R and might teach advanced users some new skills.

A key thing to realize is that *R is a programming language*. It is *not* a "statistics program" like SPSS, SAS, JMP, or Minitab, and doesn't wish to be one. The official R Project describes R as "a language and environment for statistical computing and graphics." Notice that "language" comes first, and that "statistical" is coequal with "graphics." R is a great programming language for doing statistics. The inventor of the underlying language, John Chambers received the 1998 Association for Computing Machinery (ACM) Software System Award for a system that "will forever alter the way people analyze, visualize, and manipulate data . . . "[6].

R was based on Chambers's preceding S language (S as in "statistics") developed in the 1970s and 1980s at Bell Laboratories, home of the UNIX operating system and the C programming language. S gained traction among analysts and academics in the 1990s as implemented in a commercial software package, S-PLUS. Robert Gentleman and Ross Ihaka wished to make the S approach more widely available and offered R as an open source project starting in 1997.

C. Chapman, E.M. Feit, *R for Marketing Research and Analytics*, Use R!,
DOI 10.1007/978-3-319-14436-8_1

Since then, the popularity of R has grown geometrically. The real magic of R is that its users are able to contribute developments that enhance R with everything from additional core functions to highly specialized methods. And many do contribute! Today there are over 6,000 packages of add-on functionality available for R (see http://cran.r-project.org/web/packages for the latest count).

If you have experience in programming, you will appreciate some of R's key features right away. If you're new to programming, this chapter describes why R is special and Chap. 2 introduces the fundamentals of programming in R.

1.2 Why R?

There are many reasons to learn and use R. It is the platform of choice for the largest number of statisticians who create new analytics methods, so emerging techniques are often available first in R. R is rapidly becoming the default educational platform in university statistics programs and is spreading to other disciplines such as economics and psychology.

For analysts, R offers the largest and most diverse set of analytic tools and statistical methods. It allows you to write analyses that can be reused and that extend the R system itself. It runs on most operating systems and interfaces well with data systems such as online data and SQL databases. R offers beautiful and powerful plotting functions that are able to produce graphics vastly more tailored and informative than typical spreadsheet charts. Putting all of those together, R can vastly improve an analyst's overall productivity. Elea knows an enterprising analyst who used R to automate the process of downloading data and producing a formatted monthly report. The automation saved him almost 40 h of work each month . . . which he didn't tell his manager for a few months!

Then there is the community. Many R users are enthusiasts who love to help others and are rewarded in turn by the simple joy of solving problems and the fact that they often learn something new. R is a dynamic system created by its users, and there is always something new to learn. Knowledge of R is a valuable skill in demand for analytics jobs at a growing number of top companies.

R code is also inspectable; you may choose to trust it, yet you are also free to verify. All of its core code and most packages that people contribute are open source. You can examine the code to see exactly how analyses work and what is happening under the hood.

Finally, R is free. It is a labor of love and professional pride for the R Core Development Team, which includes eminent statisticians and computer scientists. As with all masterpieces, the quality of their devotion is evident in the final work.

1.3 Why Not R?

What's not to love? No doubt you've observed that not everyone in the world uses R. Being R-less is unimaginable to us, yet there are reasons why some analysts might not want to use it.

One reason not to use R is this: until you've mastered the basics of the language, many simple analyses are cumbersome to do in R. If you're new to R and want a table of means, cross-tabs, or a t-test, it may be frustrating to figure out how to get them. R is about power, flexibility, control, iterative analyses, and cutting-edge methods, not point-and-click deliverables.

Another reason is if you do not like programming. If you're new to programming, R is a great place to start. But if you've tried programming before and didn't enjoy it, R will be a challenge as well. Our job is to help you as much as we can, and we will try hard to teach R to you. However, not everyone enjoys programming. On the other hand, if you're an experienced coder, R will seem simple (perhaps deceptively so), and we will help you avoid a few pitfalls.

Some companies and their information technology or legal departments are skeptical of R because it is open source. It is common for managers to ask, "If it's free, how can it be good?" There are many responses to that, including pointing out the hundreds of books on R, its citation in peer-reviewed articles, and the list of eminent contributors (in R, run the `contributors()` command and web search some of them). Or you might try the engineer's adage: "It can be good, fast, or cheap: pick 2." R is good and cheap, but not fast, insofar as it requires time and effort to master.

As for R being free, you should realize that contributors to R actually do derive benefit; it just happens to be non-monetary. They are compensated through respect and reputation, through the power their own work gains, and by the contributions back to the ecosystem from other users. This is a rational economic model even when the monetary price is zero.

A final concern about R is the unpredictability of its ecosystem. With packages contributed by thousands of authors, there are priceless contributions along with others that are mediocre or flawed. The downside of having access to the latest developments is that many will not stand the test of time. It is up to you to determine whether a method meets your needs, and you cannot always rely on curation or authorities to determine it for you (although you will rapidly learn which authors and which experts' recommendations to trust). If you trust your judgment, this situation is no different than with any software. *Caveat emptor*.

We hope to convince you that for many purposes, the benefits of R outweigh the difficulties.

1.4 When R?

There are a few common use cases for R:

- You want access to methods that are newer or more powerful than available elsewhere. Many R users start for exactly that reason; they see a method in a journal article, conference paper, or presentation, and discover that the method is available only in R.
- You need to run an analysis many, many times. This is how Chris started his R journey; for his dissertation, he needed to bootstrap existing methods in order to compare their typical results to those of a new machine learning model. R is perfect for model iteration.
- You need to apply an analysis to multiple data sets. Because everything is scripted, R is great for analyses that are repeated across data sets. It even has tools available for automated reporting.
- You need to develop a new analytic technique or wish to have perfect control and insight into an existing method. For many statistical procedures, R is easier to code than other programming languages.
- Your manager, professor, or coworker is encouraging you to use R. We've influenced students and colleagues in this way and are happy to report that a large number of them are enthusiastic R users today.

By showing you the power of R, we hope to convince you that your current tools are *not* perfectly satisfactory. Even more deviously, we hope to rewrite your expectations about what *is* satisfactory.

1.5 Using This Book

This book is intended to be *didactic* and *hands-on*, meaning that we want to teach you about R and the models we use in plain English, and we expect you to engage with the code interactively in R. It is designed for you to type the commands as you read. (We also provide code files for download from the book's website; see Sect. 1.5.3 below.)

1.5.1 About the Text

R commands for you to run are presented in code blocks like this:

```
> citation()

To cite R in publications use:
```

```
  R Core Team (2014). R: A language and environment for statistical computing.
     R Foundation for Statistical
  Computing, Vienna, Austria. URL http://www.R-project.org/.
...
```

We describe these code blocks and interacting with R in Chap. 2. The code generally follows the Google style guide for R (available at http://google-styleguide.googlecode.com/svn/trunk/Rguide.xml) except when we thought a deviation might make the code or text clearer. (As you learn R, you will wish to make your code readable; the Google guide is very useful for code formatting.)

When we refer to R commands, add-on packages, or data in the text outside of code blocks, we set the names in monospace type like this: `citation()`. We include parentheses on function (command) names to indicate that they are functions, such as the `summary()` function (Sect. 2.4.1), as opposed to an object such as the `Groceries` data set (Sect. 12.2.1).

When we introduce or define significant new concepts, we set them in italic, such as *vectors*. Italic is also used simply for *emphasis*.

We teach the R language progressively throughout the book, and much of our coverage of the language is blended into chapters that cover marketing topics and statistical models. In those cases, we present crucial language topics in *Language Brief* sections (such as Sect. 3.4.5). To learn as much as possible about the R language, you'll need to read the Language Brief sections even if you only skim the surrounding material on statistical models.

Some sections cover deeper details or more advanced topics, and may be skipped. We note those with an asterisk in the section title, such as *Learning More**.

1.5.2 About the Data

Most of the data sets that we analyze in this book are *simulated* data sets. They are created with R code to have a specific structure. This has several advantages:

- It allows us to illustrate analyses where there is no publicly available marketing data. This is valuable because few firms share their proprietary data for analyses such as segmentation.
- It allows the book to be more self-contained and less dependent on data downloads.
- It makes it possible to alter the data and rerun analyses to see how the results change.
- It lets us teach important R skills for handling data, generating random numbers, and looping in code.

- It demonstrates how one can write analysis code while waiting for real data. When the final data arrives, you can run your code on the new data.

An exception to this is the transactional data in Chap. 12; such data is complex to create and appropriate data has been published [20].

We recommend to work through data simulation sections where they appear; they are designed to teach R and to illustrate points that are typical of marketing data. However, when you need data quickly to continue with a chapter, it is available for download as noted in the next section and again in each chapter.

Whenever possible you should also try to perform the analyses here with your own data sets. We work with data in every chapter, but the best way to learn is to adapt the analyses to other data and work through the issues that arise. Because this is an educational text, not a cookbook, and because R can be slow going at first, we recommend to conduct such parallel analyses on tasks where you are not facing urgent deadlines.

At the beginning, it may seem overly simple to repeat analyses with your own data, but when you try to apply an advanced model to another data set, you'll be much better prepared if you've practiced with multiple data sets all along. The sooner you apply R to your own data, the sooner you will be productive in R.

1.5.3 Online Material

This book has a companion website: `http://r-marketing.r-forge.r-project.org`. The website exists primarily to host the R code and data sets for download, although we encourage you to use those sparingly; you'll learn more if you type the code and create the data sets by simulation as we describe.

On the website, you'll find:

- A welcome page for news and updates: `http://r-marketing.r-forge.r-project.org`
- Code files in .R (text) format: `http://r-marketing.r-forge.r-project.org/code`
- Copies of data sets that are simulated in the book: `http://r-marketing.r-forge.r-project.org/data`. These can be downloaded directly into R using the `read.csv()` command (you'll see that command in Sect. 2.6.2, and will find code for an example download in Sect. 3.1)
- A ZIP file containing all of the data and code files: `http://r-marketing.r-forge.r-project.org/data/chapman-feit-rintro.zip`

Links to online data are provided in the form of shortened `goo.gl` links to save typing. More detail on the online materials and ways to access the data are described in Appendix D.

1.5.4 When Things Go Wrong

When you learn something as complex as R or new statistical models, you will encounter many large and small warnings and errors. Also, the R ecosystem is dynamic and things will change after this book is published. We don't wish to scare you with a list of concerns, but we do want you to feel reassured about small discrepancies and to know what to do when larger bugs arise. Here are a few things to know and to try if one of your results doesn't match this book:

- **With R**. The basic error correction process when working with R is to check everything very carefully, especially parentheses, brackets, and upper- or lowercase letters. If a command is lengthy, deconstruct it into pieces and build it up again (we show examples of this along the way).
- **With packages** (add-on libraries). Packages are regularly updated. Sometimes they change how they work, or may not work at all for a while. Some are very stable while others change often. If you have trouble installing one, do a web search for the error message. If output or details are slightly different than we show, don't worry about it. The error `"There is no package called ..."` indicates that you need to install the package (Sect. 2.2). For other problems, see the remaining items here or check the package's help file (Sect. 2.4.2).
- **With R warnings and errors**. An R "warning" is often informational and does not necessarily require correction. We call these out as they occur with our code, although sometimes they come and go as packages are updated. If R gives you an "error," that means something went wrong and needs to be corrected. In that case, try the code again, or search online for the error message.
- **With data**. Our data sets are simulated and are affected by random number sequences. If you generate data and it is slightly different, try it again from the beginning; or load the data from the book's website (Sect. 1.5.3).
- **With models**. There are three things that might cause statistical estimates to vary: slight differences in the data (see the preceding item), changes in a package that lead to slightly different estimates, and statistical models that employ random sampling. If you run a model and the results are very similar but slightly different, you can assume that one of these situations occurred. Just proceed.
- **With output**. Packages sometimes change the information they report. The output in this book was current at the time of writing, but you can expect some packages will report things slightly differently over time.
- **With names that can't be located**. Sometimes packages change the function names they use or the structure of results. If you get a code error when trying to extract something from a statistical model, check the model's help file (Sect. 2.4.2); it may be that something has changed names.

Our overall recommendation is this. If the difference is small—such as the difference between a mean of 2.08 and 2.076, or a p-value of 0.726 vs. 0.758—don't

worry too much about it; you can usually safely ignore these. If you find a large difference—such as a statistical estimate of 0.56 instead of 31.92—try the code block again in the book's code file (Sect. 1.5.3).

1.6 Key Points

At the end of each chapter we summarize crucial lessons. For this chapter, there is only one key point: if you're ready to learn R, let's get started with Chap. 2!