*Sanjeev Kumar*

# *Advanced Big Data Analytics for Business with R*

dedicated to the one to whom I owe all

# *Contents*

# *List of Tables*

# *List of Figures*

# *Preface*

We live in a world awash in data. Companies are increasingly turning to data analytics to extract a competitive edge from data, especially large, complex datasets often called *Big Data*. However, companies are increasingly facing the challenge posed by the scarcity of analytical talent - people who can turn data into better decisions, people who can extract insights and information from data. There is growing demand for professionals with strong quantitative skills combined with an understanding of how data analytic techniques can be applied to business contexts and managerial decision making. To help my students succeed in this growing field, I teach classes in **Advanced Big Data Analytics** in Ross School of Business, Univ. of Michigan. In these courses I teache advanced analytical, statistical and data mining tools with an applied focus. This book is developed specifically for these courses.

The main focus of this book (and the associated coures) is to prepare students to model and manage business decisions with data analytics and decision models using real life case contexts and datasets. By the end of this book students will have a better understanding of processes, methodologies and tools used to transform the large amount of business data available into useful information and support business decision making. The book will focus on extracting actionable business intelligence by analyzing traditional business data as well as more recently available datasets such as social media, crowdsourcing and recommendation engines. The book focuses on the powerful, open source (and hence free) data analysis environment R.

As I designed these courses, I realized that while there are a lot of references available for doing Advanced Analytics on Big Data using R, there isn't a good reference that approaches the topic from the perspective of business students and is accessible for students who do not have extensive background in statistics or computer programming. I designed my classes to have an applied nature with significant amount of hands-on work on real business Big Data with significant managerial implications. I emphasized aspects of the analytics process that are important for actual practice of data science but are not typically well covered in traditonal textbooks - like data cleaning, managing large datasets and building data dashboards. I ended up creating a significant amount of material for the classes - much of it collated from already available but widely dispersed sources. This book is a collection of these material.

Business students have a unique mixture of tech savvy, super smarts and learn-

ing ability with a relative lack of computer programming or coding experience. They further have a great applied sense of statistics and data analysis but typically without the theoretical expertise of a Statistics student. This book is directed towards such students. This book assumes that business students are joining an Advanced Analytics course without any knowledge of computer programming and without any background in R. This book assumes that students are familiar with basic probability and statistics but their focus is on applied statistics - not on the theory. The book then builds up student's comfort level with R while at the same time makeing progress on key Advanced Analytics materials.

Direct all feedback on this book to the author at email:`sankum at umich.edu` or twitter: `@a_teachr`

## Structure of the book

Add structure information for the book here.

## Software information and conventions

This book is built using the **knitr** package (Xie, 2015) and the **bookdown** package (Xie, 2017). Following is the R session information that built the current version:

```
xfun::session_info()
```

```
## R version 3.4.2 (2017-09-28)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 17134)
##
## Locale:
##   LC_COLLATE=English_United States.1252
##   LC_CTYPE=English_United States.1252
##   LC_MONETARY=English_United States.1252
##   LC_NUMERIC=C
##   LC_TIME=English_United States.1252
```

```
##
## Package version:
##   base64enc_0.1.3 bookdown_0.8    compiler_3.4.2
##   digest_0.6.12   evaluate_0.10.1 graphics_3.4.2
##   grDevices_3.4.2 highr_0.6       htmltools_0.3.6
##   jsonlite_1.5    knitr_1.21      magrittr_1.5
##   markdown_0.8    methods_3.4.2   mime_0.5
##   Rcpp_0.12.13    rmarkdown_1.11  stats_3.4.2
##   stringi_1.1.5   stringr_1.2.0   tinytex_0.9
##   tools_3.4.2     utils_3.4.2     xfun_0.4
##   yaml_2.2.0
```

Package names are in bold text (e.g., **rmarkdown**), and inline code and filenames are formatted in a typewriter font (e.g., `knitr::knit('foo.Rmd')`). Function names are followed by parentheses (e.g., `bookdown::render_book()`).

## Acknowledgments

Add acknowledgements here.

Sanjeev Kumar
Technology and Operations, Ross School of Business, Univ. of Michigan

# *About the Author*

Sanjeev Kumar is part of the Technology and Operations faculty at the Ross School of Business, Univ. of Michigan.

{-}

test

# 1

# *Introduction*

Welcome to **Advanced Big Data Analytics for Business with R**. Let's get started.

## 1.1 Who Is This Book For?

This book is for business students, practitioners and executives who want to have hands-on experience in working with business related big data and want to extract actionable insights from the data using advanced analytical techniques. This book is application oriented and hence focuses on the application of advanced analytical techniques rather than the theoretical details. Consequently, this book is not for readers solely focused on the theoretical, statistical part of data analytics.

This book assumes that the readers have basic familiarity with introductory probability and statistics. This domain is easier to follow and understand if the reader also has *some* exposure to *some* kind of computer programming although that is not a necessity. This book has been written for a practitioner audience, not an academic one. The book aims to be an exhaustive reference resource for practitioners - that's why it starts with baby steps of introducing R and basic data manipulation and ends at the other end with advanced Machine Learning algorithms and complex model improvement algorithms.

Thank you for placing your trust in this book. This chapter will provider further details about this book and how to best read/use this book.

## 1.2 About This Document

This document, which you are probably seeing as a GitHub eBook or a PDF document or even as printed book, has been written entirely using R and

RStudio. It uses the R packages `bookdown` along with `rmarkdown` to integrated R code inside my favorite document creation system called **LaTeX**. This document has been created entirely using Open Source tools and has been released back into the Open Source ecosystem for free using the GNU General Public License (GPL). You are free to share this document with others as long as you comply with the GPL license. GPL License usually require that the product (in this case, this book) be made available free or charge; and any subsequent product made using the GPL Licensed product must also be made available free of charge (Foundation, 2007).

This is my first attempt at writing a book size document. I have had to develop a bunch of workarounds to make the process work and get a quality output. The source code for this book is available for public use free of charge at: `https://github.com/clarifyR/AABookDown`.

## 1.3 How is This Book Structured

This book covers a wide range of material. The material is organized in 5 parts, 19 chapters and several appendices.

### 1.3.1 Part I: Getting Started

We set ourselves for the book - download and install all needed software; figure our way around R and RStudio and learn the basics of the R language. Essentially lay down enough of a foundation that we can start getting productive. For students without a computer programming background, it is essential that they do not rush this part. Our success in later parts depend upon us getting comfortable with the material here.

### 1.3.2 Part II: Data Exploration and Visualization

Real data analytics begins with us understanding and getting a handle on the data. This typically involves cleaning up the data, generating descriptive statistics to better understand the data and finally creating visualizations that allow us to better understand the underlying complexity of the data. In my opinion, data cleaning is the most under-appreciated part of data analytics. It often takes more time and effort than the actual analysis that follows.

Data visualization has emerged as a key tool in Big Data Analytics. We specifically focus our attention here on a subset of data visualization that is im-

portant in the business context - creating data dashboards that can help managerial decision making.

### 1.3.3 Part III: Traditional Statistical Modeling

Even advanced analytic techniques like machine learning algorithms in the next part have their foundations in traditional statistical methodologies. Classical statistical modeling approaches like Linear Regression are still the benchmark given their immense popularity, flexibility and stability. In this part of the book, we explore traditional statistical analysis tools like Linear Regression, Generalized Linear Models like Logistic and Survival Models, Principal Components and Factor Analysis, Time Series Analysis and so on.

As we assume that you are already familiar with basic probability and statistics, we will focus on application of the methodologies and not the underlying theory. We will also focus on how to tweak these tools for the Big Data world as many of these tools run into trouble when sample sizes are quite large. Bigger is not always better - traditional statisticall methodologies were developed/optimized for smaller sample sizes. Using them for large sample sizes give rise to unique issues and problems - we will discuss how to address them.

### 1.3.4 Part IV: Machine Learning and Predictive Analytics

This is the largest and the most important part of the book. This is why this book was written in the first place - an overview of data analytics tools specific to Big Data - variously known as Machine Learning, Statistical Learning, Predictive Analytics, Data Mining as so on. This book focuses on tools that help us make sense of Big Data and helps us automate the extraction of managerial decision making insights from Big Data.

As with much of the rest of the book, the focus is on applying the tools rather than their theoretical/mathematical underpinnings. We will discuss enough theory to develop an overall understanding and then devote our energies on making these tools work on real datasets.

### 1.3.5 Part V: Putting It All Together

Now that we have gone through all the elements individually, we can move forward to create a combined, integrated approach that puts all these pieces together. How to combine different models so that they result in an output better than sum of their parts? How to ensure that our models improve as more data become available?

We will conclude by running a couple of large integrated data analytics projects that will combine elements discussed in the book.

### 1.3.6 Part V: Appendices, Bibliography and Index

Back matter of the book. The Index in the end has three main components - Key Concepts, R Commands and R Packages. Appendices include syllabus for the two courses this book is primarily used for and other miscellaneous content that did not fit in one of the main matter chapters.

There are a lot of quality, free, online resources available for building up your R and Data Analytics skills before you go through this book or the associated courses. They are listed in Appendix: Key References. A fully fleshed data analysis example has been presented on Appendix: Data Analysis Example to give you an idea of the power and range of what can be accomplished with just a little bit of familiarity with R.

## 1.4 How to Read This Book

You would have noticed by now that this book integrates R code within its text. Most of the time R code takes the form of dedicated code blocks like the one below:

```
#This is a demo code block
print("Hello World")
```

```
## [1] "Hello World"
```

As you can see from above, code blocks are printed in color, in `fixed width font`. There is a color scheme here that you will soon become familiar with - comments are in italics and kind of violet looking, functions are in reddish color, strings appear bluish and so on. The output of the code block appears right after - for example - the output of the `print()` command follows the code above.

You would notice that whenever we refer to a R command in a significant way (like `print()` in paragraph above), the command is printed in `fixed width font`, in red color with a yellow highlight that makes it easy for you to see which commands are discussed significantly on the page. The highlighting also

ensures that an entry for the command is placed in the Index provided at the end of the book. For times when a command is mentioned in text in a minor way not necessating a margin and Index entry, they will be typed in `fixed width font`.

Like R Commands, this book provides special formatting for R Packages used in the book. R Packages are fomatted like R Commands but in blue color - check our reference of the `clarifyR` package earlier - blue fixed width font text with yellow highlighting and finally an entry in the Packages section of the Index.

Lastly, special formatting is provided for key concepts - similar to R Commands and Packages in all respects except that they are in bolded and in black color and their Index entry is in the Key Concepts section. For example: this book focuses on `Open Source` software - that are created by a community of software developers for use by the community, usually made available for free.

The three special formatting elements - commands, packages and key concepts - ensure that you have a summary of significant elements discussed in the page just by looking for highlighted items.

---

## Is This Book Suitable For You?

Well - you wouldn't know until you spend some time with it. Dig in.

A quick word of caution though before you get too deep: this book (and the associated courses) are very hands-on. There is no point in reading this book like a sequence of text. This book should be seen more as an illustrated text - illustrated with relevant R commands and material. You should read this book alongside an RStudio session - trying all the commands there as you read along. You will need to get comfortable with a lot of command line typing, keyboard shortcuts and figuring workarounds to inevitable problems that will arise. We will often get into situations where there will be no tested/optimized/prescribed solution and we will need to figure our way out - often with some trial and error. You should be comfortable with such ambiguity.

## 1.5 Archive - Delete

This section holds the placeholder information that will be deleted after the book is built.

We have a nice figure in Figure 1.1, and also a table in Table 1.1.

```
par(mar = c(4, 4, 1, .1))
plot(cars, pch = 19)
```

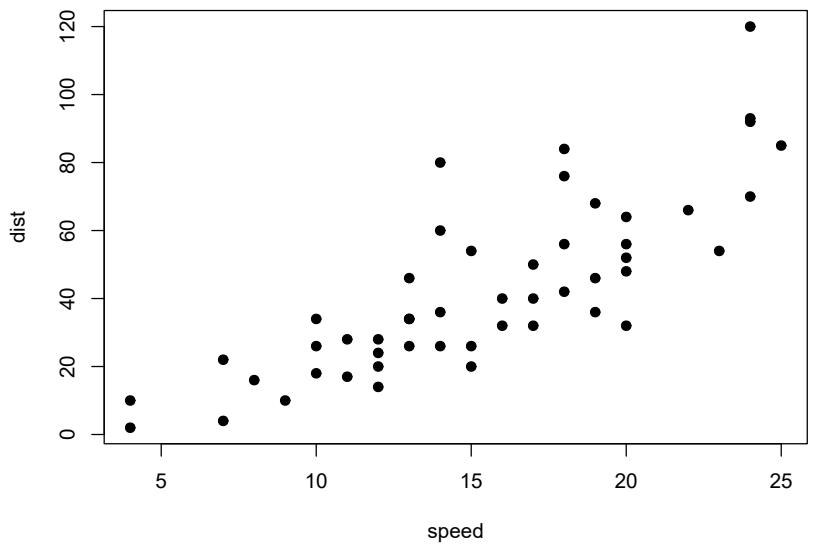


**FIGURE 1.1:** Hello World!

```
knitr::kable(
  head(iris), caption = 'The boring iris data.',
  booktabs = TRUE
)
```

**TABLE 1.1:** The boring iris data.

| Sepal.Length | Sepal.Width | Petal.Length | Petal.Width | Species |
|---|---|---|---|---|
| 5.1 | 3.5 | 1.4 | 0.2 | setosa |
| 4.9 | 3.0 | 1.4 | 0.2 | setosa |
| 4.7 | 3.2 | 1.3 | 0.2 | setosa |
| 4.6 | 3.1 | 1.5 | 0.2 | setosa |
| 5.0 | 3.6 | 1.4 | 0.2 | setosa |
| 5.4 | 3.9 | 1.7 | 0.4 | setosa |

# 2

## *The Second Chapter*

We talk about the *FOO* method in this chapter.

# A

# *Syllabi*

This appendix will have the syllabus for the relevant Ross courses.

## A.1 TO404 Big Data Manipulation and Visualization

## A.2 TO414 Advanced Analytics

## A.3 TO628 Advanced Big Data Analytics

# *Bibliography*

Foundation, F. S. (2007). The gnu general public license v3.0 - gnu project - free software foundation. https://www.gnu.org/licenses/gpl-3.0.en.html. (Accessed on 09/23/2016).

Xie, Y. (2015). *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition. ISBN 978-1498716963.

Xie, Y. (2017). *bookdown: Authoring Books and Technical Documents with R Markdown*. R package version 0.5.

# *Index*