PYTHON® for R USERS

A Data Science Approach

A. Ohri

WILEY

Python® for R Users

Python® for R Users

A Data Science Approach

Ajay Ohri

WILEY

This edition first published 2018

Registered Office
John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA

Editorial Office
111 River Street, Hoboken, NJ 07030, USA

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Library of Congress Cataloguing-in-Publication Data

Name: Ohri, A. (Ajay), author.
Title: Python® for R users : a data science approach / Ajay Ohri.
Description: Hoboken, NJ : John Wiley & Sons, 2018. | Includes bibliographical references and index. |
Identifiers: LCCN 2017022045 (print) | LCCN 2017036415 (ebook) | ISBN 9781119126775 (pdf) | ISBN 9781119126782 (epub) | ISBN 9781119126768 (pbk.)
Subjects: LCSH: Python (Computer program language) | R (Computer program language)
Classification: LCC QA76.73.P98 (ebook) | LCC QA76.73.P98 O37 2017 (print) | DDC 005.13/3–dc23
LC record available at https://lccn.loc.gov/2017022045

Cover design: Wiley
Cover images: (Background) © Duncan Walker/iStockphoto

Set in 10/12pt Warnock by SPi Global, Pondicherry, India

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

Dedicated to my family in Delhi, Mumbai, and the United States

and

Kush Ohri (my son whom I love very much)

and

Jesus Christ (my personal savior)

Contents

Preface

I started my career with selling cars in 2003. That was my first job after 2 years of MBA and 4 years of engineering. In addition, I took off 2 years to enter a military academy as an officer cadet (dropped out in 1 year) and as a physicist (dropped out after 1 year). Much later, I dropped out of my PhD Track (MS Stats) after 1 year in Knoxville. I did not do very well in statistics theory in my engineering, my MBA, or even my grad school. I was only interested in statistical software and fortunately I was not very bad at using it. So in 2004, I dropped out of selling cars and entered into writing statistical software for General Electric's then India-based offshore company.

I used a language called SAS for a software called Base SAS. The help provided by the software company called SAS for this software and language was quite nice, so it was nice to play with data and code all day and be paid to have fun. After a few years of job changes, I came across open-source software when I started building my own start-up. I really like SAS as a language and a company, but as a start-up guy I could not afford it, and the SAS University Edition was not there in 2007. Since I needed money to pay for diapers of my baby Kush, and analysis was the only gift God had given me, I turned to R.

R, Open Office, and Ubuntu Linux were my first introduction to open-source statistical computing, and I persevered in it. In 2007 I started my own start-up in business analytics writing and consulting, Decisionstats.com. In 2009 I entered the University of Tennessee for a funded assistantship, I interned in Silicon Valley for a few weeks in the winter, and I dropped out on medical reasons after taking courses across multiple departments from graphics design and genetic algorithms from Computer Science Department, apart from Statistics Department. Cross-domain training helped me a lot to think in various ways to give simple solutions, and I will always be thankful to the kind folks in Statistics and Computer Science Department of the University of Tennessee.

Once I mastered my brain around the vagaries of troubleshooting in Linux and of object-oriented programming on R, I was good to go to give consulting projects for data analysis. Those days we used to call it business analytics, but today of course we call it data science.

Since I often forget things including where I kept my code, I started blogging on things that I felt were useful and might be useful to others. After a few years I discovered that in the real world it was not what I knew, but who I knew that really helped my career. So I began interviewing people in Analytics and R and my blog viewership took off. My blog philosophy continues to be—a blog post should be useful, it should be unique, and it should be interesting. In 2016, I had amassed 1,000,000 views on DecisionStats.com—again a surprising turn of events for me. I am most grateful to the 100 plus people who agreed to be interviewed by me.

2007 and 2008 were early days for analytics blogging for sure. After a few years I had enough material to put together a book and enough credibility to publish with a publisher. In 2012 I came up with my first book and in 2014 I came up with my second book. In 2016, the Chinese translation of my first book was realized. Surprisingly for me, a review of my second book appeared in the *Journal of Statistical Software*.

After publishing two books on R, mentoring many start-ups by consulting and training, engaging consulting clients in real-world problems, and making an established name in social media, I still felt I needed to learn more.

Data was getting bigger and bigger. It was not enough to know how to write small data analytics using a single machine in serialized code; perhaps it was time to write parallel code in multiple machines on big data analytics. Then there was the divide between statisticians and computer science that fascinated me since I see data as data, a problem to be solved. As Eric S. Raymond wrote in the Hacker's attitude, "The world is full of interesting problems."

Then there was temptation and intellectual appeal of an alternative to R, called Python, which came with batteries attached (allegedly).

Once my scientific curiosity was piqued, I started learning Python. I found Python was both very good and very bad compared with R. Its community has different sets of rules and behavior (which are always turbulent in the passionate world of open-source developer ninjas). But the language itself was very different. I don't care about the language. I love science. But if a person like me who at least knows how to code a wee bit in R found it so tough to redo the same thing in Python, I thought maybe others were facing this transitioning problem too. For big data and for some specific use cases, Python was better in terms of speed. Speed matters, no matter how much Moore's law conspires with the either to make it easier for you to write code. R also seemed to turn into a language where all I did was import a package and run a function with tweaked parameters. As R became the scientific mainstream replacing SAS

language, and SAS remained the enterprise statistical language, Python and how to write code in it became the thing for anonymous red hat hackers like me to venture delve and explore into.

As the Internet of people expands to Internet of things, I feel that budding data scientists should know at least two languages in analytics so they can be secure on career. This also gives enterprises an open choice on which software to prototype models and which software to deploy in production environments.

Acknowledgments

The author is grateful to many people working in both the Python and R community for making this book possible. He would especially like to thank Dr. Eric Siegel of Predictive Analytics Conference and John Sall of JMP. He would like to thank all his students in 2012–2016.

This book would not be done without the support from Madhur Batra for mentoring and logistical support. On a technical side, inputs and hard work from his interns Yashika and Chandan Routray (IIT Kharagpur) and his DecisionStats team helped him. His coresearcher F. Xavier provided invaluable help with case studies.

Scope

The scope of the book is to introduce Python as a platform for data science practitioners including aspiring budding data scientists. The book is aimed at people who know R coding at various levels of expertise, but even those who know no coding in no language may find some value in it. It is not aimed at members of research communities and research departments. The focus is on simple tutorials and actionable analytics, not theory. I have also tried to incorporate R code to give a compare and contrast approach to learners.

Chapter 1

Introduction deals with Python and comparison with R. It also lists the functions and packages used in both languages. It also lists some managerial models that the author feels data scientists should be aware of. It introduces the reader to basics of Python and R language.

Chapter 2

"Data Input" deals with an approach for people to get data of various volume variety and velocity in Python. This includes web scraping, databases, noSQL data, and spreadsheet like data.

Chapter 3

"Data Inspection and Data Quality"—Data Inspection deals with choices in verifying data quality in Python.

Chapter 4

"Exploratory Data Analysis" deals with basic data exploration and data summarization with rolling up data with group by criterion.

Chapter 5

"Statistical Modeling" deals with creating models based on statistical analysis including OLS regression that are useful for industry to build propensity models.

Chapter 6

"Data Visualization" deals with visual methods to inspect raw and rolled-up data.

Chapter 7

"Machine Learning Made Easier" deals with commonly used data mining methods for model building. This is done with an emphasis on both supervised and unsupervised methods and further emphasis on regression and clustering techniques. Time series forecasting helps the user with time series forecasting. Text mining deals with text mining methods and natural language processing. Web analytics looks at using Python for analyzing web data. Advanced data science looks at methods and techniques for newer age use cases including cloud computing-enabled big data analysis, social network analysis, Internet of things, etc.

Chapter 8

Conclusion and Summary—We list down what we learned and tried to achieve in this book, and our perspective for future growth of R and Python as well as statistical computing to grow, and render data science a credible foothold for the future.

Purpose

The book has been written from a practical use case perspective for helping people navigate multiple open-source languages in the pursuit of data science excellence. The author believes that there is no one software or language that can solve all kinds of data problems all the time. An optimized approach to learning is better than an ideological approach to learning statistical software. Past habits of thinking must be confronted to enhance speed of future knowledge enhancement.

Plan

I will continue to use screenshots as a tutorial device and I will draw upon my experience in data science consulting to highlight practical data parsing problems. This is because choosing the right tool and technique and even package is not so time consuming but the sheer variety of data and business problems can suck up the data scientist's time that can later affect quality of his judgment and solution.

Intended Audience

This is a book for budding data scientists and existing data scientists married to other languages like SPSS or R or Julia. I am trying to be practical about solving problems in data. Thus there will be very little theory.

Afterthoughts

I am focused on practical solutions. I will therefore proceed on the assumption that the user wants to do data science or analytics at the lowest cost and greatest accuracy, robustness, and ease possible. A true scientist always keeps his mind open to data and options regardless of who made whom. The author finds that information asymmetry and brand clutter have managed to confuse audiences of the true benefits of R versus Python versus other languages. The instructions and tutorials within this book have no warranty and you are doing so at your own risk.

As a special note on formatting of this manuscript, the author mostly writes on Google Docs, but here he is writing using the GUI LyX for the typesetting software LaTex, and he confesses he is not very good at it. We do hope the book is read by business users, technical users, CTOs keen to know more on R and Python and when to use open-source analytics, and students wishing to enter a very nice career as data scientists. R is well known for excellent graphics but

not so suitable for bigger datasets in its native straight to use open-source version. Python is well known for being great with big datasets and flexibility but has always played catch-up to the number of good statistical libraries as available in R.

The enterprise CTO can reduce costs incredibly by using open-source software and hardware via blended cloud and blended open-source software.

The Zen of Python

Tim Peters

- Beautiful is better than ugly.
- Explicit is better than implicit.
- Simple is better than complex.
- Complex is better than complicated.
- Flat is better than nested.
- Sparse is better than dense.
- Readability counts.
- Special cases aren't special enough to break the rules.
- Although practicality beats purity.
- Errors should never pass silently. Unless explicitly silenced.
- In the face of ambiguity, refuse the temptation to guess.
- There should be one—and preferably only one—obvious way to do it.
- Although that way may not be obvious at first unless you're Dutch.
- Now is better than never. Although never is often better than right now.
- If the implementation is hard to explain, it's a bad idea.
- If the implementation is easy to explain, it may be a good idea.
- Namespaces are one honking great idea—let's do more of those!

Source: https://www.python.org/dev/peps/pep-0020/

1

Introduction to Python R and Data Science

1.1 What Is Python?

Python is a programming language that lets you work more quickly and integrate your systems more effectively. It was created by Guido van Rossum. You can read Guido's history of Python at the History of Python blog at http://python-history.blogspot.in/2009/01/introduction-and-overview.html.

It is worth reading for beginners and even experienced people in Python. The following is just an extract:

> many of Python's keywords (if, else, while, for, etc.) are the same as in C, Python identifiers have the same naming rules as C, and most of the standard operators have the same meaning as C. Of course, Python is obviously not C and one major area where it differs is that instead of using braces for statement grouping, it uses indentation. For example, instead of writing statements in C like this

```
if (a < b) {
   max = b;
} else {
   max = a;
}
```

> Python just dispenses with the braces altogether (along with the trailing semicolons for good measure) and uses the following structure:

```
if a < b:
   max = b
else:
   max = a
```

Python® for R Users: A Data Science Approach, First Edition. Ajay Ohri.

> The other major area where Python differs from C-like languages is in its use of dynamic typing. In C, variables must always be explicitly declared and given a specific type such as int or double. This information is then used to perform static compile-time checks of the program as well as for allocating memory locations used for storing the variable's value. In Python, variables are simply names that refer to objects.

The Python Package Index (PyPI) https://pypi.python.org/pypi hosts third-party modules for Python. There are currently **91 625** packages there. You can browse Python packages by topic at https://pypi.python.org/pypi?%3A action=browse

1.2 What Is R?

The official definition of what is R is given on the main website at http://www.r-project.org/about.html

> R is an integrated suite of software facilities for data manipulation, calculation and graphical display. It includes an effective data handling and storage facility, a suite of operators for calculations on arrays, in particular matrices, a large, coherent, integrated collection of intermediate tools for data analysis, graphical facilities for data analysis and display either on-screen or on hardcopy, and a well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities.
>
> The term 'environment' is intended to characterize it as a fully planned and coherent system, rather than an incremental accretion of very specific and inflexible tools, as is frequently the case with other data analysis software.

The Comprehensive R Archive Network (CRAN) hosts thousands of packages for R at https://cran.r-project.org/web/packages/, so does GitHub (see https://github.com/search?utf8=%E2%9C%93&q=stars%3A%3E1+language%3AR) as well as Bioconductor as package repositories. You can see all the packages from these repositories for R at http://www.rdocumentation.org/ (11 885 packages as of 2016).

As per the author, R is both a language in statistics as well as computer science and an analytics software with great usefulness in analyzing business data and applying data science to it. In particular the appeal of R remains: it is a free open source and has a huge number of packages particularly dealing with analysis of data.

Disadvantages of R remain memory handling in production environments, lack of incentives for R developers, and a sometimes turgid documentation that is mildly academic oriented rather than enterprise user oriented.

1.3 What Is Data Science?

Data science lies at the intersection of programming, statistics, and business analysis. It is the use of programming tools with statistical techniques to analyze data in a systematic and scientific way. A famous diagram by Drew Conway put data science as the intersection of the three. It is given at http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

The author defines a data scientist as follows:

> A data scientist is simply a person who can **write code** (in languages like R, Python, Java, SQL, Hadoop (Pig, HQL, MR) etc.) **for data** (storage, querying, summarization, visualization) **efficiently** and **quickly on hardware** (local machines, on databases, on cloud, on servers) and **understand enough statistics** to derive **insights** from **data** so business can make **decisions.**

1.4 The Future for Data Scientists

The respectable *Harvard Business Review* defines data scientist to be the sexiest job of the twenty-first century (https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/).

Surveys on salaries point out to both rising demand and salaries for data scientists and a big shortage for trained professionals (see http://www.forbes.com/sites/gilpress/2015/10/09/the-hunt-for-unicorn-data-scientists-lifts-salaries-for-all-data-analytics-professionals/). Indeed this has coined a new term unicorn data scientists. A unicorn data scientist is rare to find for he has all the skills in programming, statistics, and business aptitude. A modification of the Data Science Venn Diagram in Figure 1.1 is available at http://www.anlytcs.com/2014/01/data-science-venn-diagram-v20.html, which the author found more updated.

In addition, unicorn is a term in the investment industry, and in particular the venture capital industry, which denotes a start-up company whose valuation has exceeded $1 billion. The term has been popularized by Aileen Lee of Cowboy Ventures. They can be seen at http://graphics.wsj.com/billion-dollar-club/ and http://fortune.com/unicorns/

Not surprisingly data science offers a critical edge to these start-ups as well. So we can have both rising demand and short supply of data scientists, leading to a more secure work environment. A list of start-ups can be seen at Y Combinator at http://yclist.com/ including data science related start-ups. You can see a survey here on data scientist salaries at http://www.burtchworks.com/2015/07/14/compensation-of-data-scientists-insights-from-the-past-year. The annual Rexer Analytics survey helps gauge skills and usage by data miners. You can read an interview at http://decisionstats.com/2013/12/25/

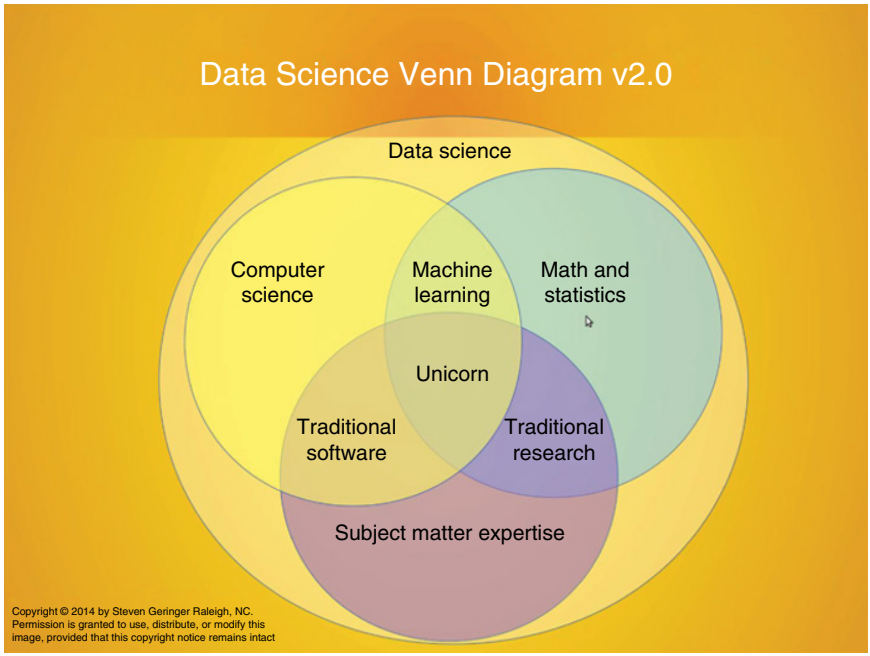

Figure 1.1 Data Science Venn diagram. *Source:* Copyright © 2014 Steven Geringer Raleigh, NC.

karl-rexer-interview-on-the-state-of-analytics/ or read the report at www.rexeranalytics.com. We can thus sum up and say that data scientists who have the right skills have a great future ahead professionally.

A note of caution is that skills need to be updated by data scientists very quickly and they need to be responsive to business needs to frame the data science solutions. So the risk of being obsolete remains an encouragement for data scientists to get multiple skills. An interesting fellowship program for data scientists is run by Insight at http://insightdatascience.com/, and a repository for data science is available for free at https://github.com/okulbilisim/awesome-datascience

Closer home, the NY-based Byte academy offers a Python-based program for data science at http://byteacademy.co/

1.5 What Is Big Data?

Big data is a broad term for datasets so large or complex that traditional data processing applications are inadequate. The 3Vs model helps with understanding big data.