

Dan Toomey

Jupyter for Data Science

Exploratory analysis, statistical modeling, machine learning, and data visualization with Jupyter



Packt>

Jupyter for Data Science

Jupyter Notebook is a web-based environment that enables interactive computing in notebook documents. It allows you to create documents that contain live code, equations, and visualizations. This book is a comprehensive guide to getting started with data science using the popular Jupyter Notebook.

If you are familiar with Jupyter Notebook and want to learn how to use it to perform various data science tasks, this is the book for you! From data exploration to visualization, this book will take you through every step of implementing an effective data science pipeline using Jupyter. You will also see how to utilize Jupyter's features to share your documents and codes with your colleagues. The book also explains how Python 3, R, and Julia can be integrated with Jupyter for various data science tasks.

By the end of this book, you will comfortably leverage the power of Jupyter to perform various tasks in data science successfully.

Things you will learn:

- Understand why Jupyter Notebooks are a perfect fit for your data science tasks
- Perform scientific computing and data analysis tasks with Jupyter
- Interpret and explore different kinds of data visually with charts, histograms, and more
- Extend SQL's capabilities with Jupyter Notebooks
- Combine the power of R and Python 3 with Jupyter to create dynamic notebooks
- Create interactive dashboards and dynamic presentations
- Master the best coding practices and deploy your Jupyter Notebooks efficiently

Jupyter for Data Science

Dan Toomney



Jupyter for Data Science

Exploratory analysis, statistical modeling, machine learning,
and data visualization with Jupyter

Dan Toomey



BIRMINGHAM - MUMBAI

Jupyter for Data Science

Copyright © 2017 Packt Publishing

All rights reserved. No part of this book may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, without the prior written permission of the publisher, except in the case of brief quotations embedded in critical articles or reviews.

Every effort has been made in the preparation of this book to ensure the accuracy of the information presented. However, the information contained in this book is sold without warranty, either express or implied. Neither the author, nor Packt Publishing, and its dealers and distributors will be held liable for any damages caused or alleged to be caused directly or indirectly by this book.

Packt Publishing has endeavored to provide trademark information about all of the companies and products mentioned in this book by the appropriate use of capitals. However, Packt Publishing cannot guarantee the accuracy of this information.

First published: October 2017

Production reference: 1171017

Published by Packt Publishing Ltd.
Livery Place
35 Livery Street
Birmingham
B3 2PB, UK.

ISBN 978-1-78588-007-0

www.packtpub.com

Credits

Author

Dan Toomey

Reviewers

Jesse Bacon

Ruben Oliva Ramos

Commissioning Editor

Amey Varangaonkar

Acquisition Editor

Amey Varangaonkar

Content Development Editor

Aishwarya Pandere

Technical Editor

Prasad Ramesh

Copy Editor

Safis Editing

Project Coordinator

Nidhi Joshi

Proofreader

Safis Editing

Indexer

Tejal Daruwale Soni

Graphics

Tania Dutta

Production Coordinator

Aparna Bhagat

About the Author

Dan Toomey has been developing applications for over 20 years. He has worked in a variety of industries and companies of all sizes, in roles from sole contributor to VP/CTO level. For the last 10 years or so, he has been contracting companies in the eastern Massachusetts area under Dan Toomey Software Corp. Dan has also written *R for Data Science* and *Learning Jupyter* with Packt Publishing.

About the Reviewers

Jesse Bacon is a hobbyist programmer and technologist in the Washington D.C. metro area. In his free time, he mostly studies a new and interesting technology or visits the gym pursuing fitness goals. Jesse values the opinions of the development community and looks forward to a new generation of programmers with all the gifts of today's computing environments.

Ruben Oliva Ramos is a computer systems engineer from Tecnologico of León Institute, with a master's degree in computer and electronic systems engineering, teleinformatics, and networking specialization from the University of Salle Bajio in Leon, Guanajuato Mexico. He has more than 5 years of experience in developing web applications to control and monitor devices connected with Arduino and Raspberry Pi using web frameworks and cloud services to build Internet of Things applications.

He is a mechatronics teacher at the University of Salle Bajio and teaches students of master's in design and engineering of mechatronics systems. He also works at Centro de Bachillerato Tecnologico Industrial 225 in Leon, teaching electronics, robotics and control, automation, and microcontrollers at Mechatronics Technician Career.

Ruben is a consultant and developer projects in areas such as monitoring systems and datalogger data using technologies such as Android, iOS, Windows Phone, HTML5, PHP, CSS, Ajax, JavaScript, Angular, ASP .NET databases (SQLite, mongoDB, and MySQL), web servers (Node.js and IIS), hardware programming (Arduino, Raspberry pi, Ethernet Shield, GPS and GSM/GPRS, ESP8266), and control and monitor systems for data acquisition and programming.

He is the author of these two books for Packt Publishing:

- Internet of Things Programming with JavaScript
- Advanced Analytics with R and Tableau

Also Monitoring, controlling and acquisition of data with Arduino and Visual Basic .NET for Alfaomega.

I would like to thank my savior and lord, Jesus Christ, for giving me the strength and courage to pursue this project; to my dearest wife, Mayte; our two lovely sons, Ruben and Dario; my dear father, Ruben; my dearest mom Rosalia; my brother, Juan Tomas; and my sister, Rosalia, whom I love. This is for all their support while reviewing this book, for allowing me to pursue my dream and tolerating not being with them after my busy day. I'm very grateful to Packt Publishing for giving the opportunity to collaborate as an author and reviewer and join to this honest and professional team.

www.PacktPub.com

For support files and downloads related to your book, please visit www.PacktPub.com. Did you know that Packt offers eBook versions of every book published, with PDF and ePub files available? You can upgrade to the eBook version at www.PacktPub.com and as a print book customer, you are entitled to a discount on the eBook copy. Get in touch with us at service@packtpub.com for more details. At www.PacktPub.com, you can also read a collection of free technical articles, sign up for a range of free newsletters and receive exclusive discounts and offers on Packt books and eBooks.



<https://www.packtpub.com/mapt>

Get the most in-demand software skills with Mapt. Mapt gives you full access to all Packt books and video courses, as well as industry-leading tools to help you plan your personal development and advance your career.

Why subscribe?

- Fully searchable across every book published by Packt
- Copy and paste, print, and bookmark content
- On demand and accessible via a web browser

Customer Feedback

Thanks for purchasing this Packt book. At Packt, quality is at the heart of our editorial process. To help us improve, please leave us an honest review on this book's Amazon page at <https://www.amazon.com/dp/1785880071>.

If you'd like to join our team of regular reviewers, you can email us at customerreviews@packtpub.com. We award our regular reviewers with free eBooks and videos in exchange for their valuable feedback. Help us be relentless in improving our products!

Table of Contents

Preface	1
Chapter 1: Jupyter and Data Science	7
Jupyter concepts	7
A first look at the Jupyter user interface	8
Detailing the Jupyter tabs	8
What actions can I perform with Jupyter?	9
What objects can Jupyter manipulate?	11
Viewing the Jupyter project display	12
File menu	13
Edit menu	14
View menu	15
Insert menu	16
Cell menu	16
Kernel menu	17
Help menu	18
Icon toolbar	18
How does it look when we execute scripts?	19
Industry data science usage	20
Real life examples	21
Finance, Python - European call option valuation	21
Finance, Python - Monte Carlo pricing	23
Gambling, R - betting analysis	25
Insurance, R - non-life insurance pricing	27
Consumer products, R - marketing effectiveness	31
Using Docker with Jupyter	34
Using a public Docker service	35
Installing Docker on your machine	36
How to share notebooks with others	36
Can you email a notebook?	36
Sharing a notebook on Google Drive	37
Sharing on GitHub	37
Store as HTML on a web server	37
Install Jupyter on a web server	37
How can you secure a notebook?	38
Access control	38
Malicious content	38
Summary	39
Chapter 2: Working with Analytical Data on Jupyter	41

Data scraping with a Python notebook	41
Using heavy-duty data processing functions in Jupyter	45
Using NumPy functions in Jupyter	45
Using pandas in Jupyter	48
Use pandas to read text files in Jupyter	48
Use pandas to read Excel files in Jupyter	49
Using pandas to work with data frames	50
Using the groupby function in a data frame	50
Manipulating columns in a data frame	51
Calculating outliers in a data frame	52
Using SciPy in Jupyter	53
Using SciPy integration in Jupyter	53
Using SciPy optimization in Jupyter	55
Using SciPy interpolation in Jupyter	56
Using SciPy Fourier Transforms in Jupyter	58
Using SciPy linear algebra in Jupyter	59
Expanding on panda data frames in Jupyter	61
Sorting and filtering data frames in Jupyter/IPython	63
Filtering a data frame	63
Sorting a data frame	64
Summary	66
Chapter 3: Data Visualization and Prediction	67
Make a prediction using scikit-learn	67
Make a prediction using R	72
Interactive visualization	77
Plotting using Plotly	79
Creating a human density map	80
Draw a histogram of social data	83
Plotting 3D data	86
Summary	88
Chapter 4: Data Mining and SQL Queries	89
Special note for Windows installation	90
Using Spark to analyze data	90
Another MapReduce example	92
Using SparkSession and SQL	94
Combining datasets	97
Loading JSON into Spark	101
Using Spark pivot	103
Summary	105

Chapter 5: R with Jupyter	107
How to set up R for Jupyter	107
R data analysis of the 2016 US election demographics	108
Analyzing 2016 voter registration and voting	111
Analyzing changes in college admissions	120
Predicting airplane arrival time	126
Summary	130
Chapter 6: Data Wrangling	131
Reading a CSV file	131
Reading another CSV file	135
Manipulating data with dplyr	138
Converting a data frame to a dplyr table	139
Getting a quick overview of the data value ranges	139
Sampling a dataset	140
Filtering rows in a data frame	142
Adding a column to a data frame	143
Obtaining a summary on a calculated field	144
Piping data between functions	144
Obtaining the 99% quantile	145
Obtaining a summary on grouped data	146
Tidying up data with tidyr	147
Summary	152
Chapter 7: Jupyter Dashboards	153
Visualizing glyph ready data	153
Publishing a notebook	159
Font markdown	159
List markdown	160
Heading markdown	160
Table markdown	161
Code markdown	161
More markdown	162
Creating a Shiny dashboard	162
R application coding	163
Publishing your dashboard	167
Building standalone dashboards	168
Summary	170
Chapter 8: Statistical Modeling	171

Converting JSON to CSV	171
Evaluating Yelp reviews	172
Summary data	173
Review spread	174
Finding the top rated firms	175
Finding the most rated firms	176
Finding all ratings for a top rated firm	176
Determining the correlation between ratings and number of reviews	178
Building a model of reviews	179
Using Python to compare ratings	180
Visualizing average ratings by cuisine	182
Arbitrary search of ratings	183
Determining relationships between number of ratings and ratings	184
Summary	188
Chapter 9: Machine Learning Using Jupyter	189
Naive Bayes	189
Naive Bayes using R	189
Naive Bayes using Python	192
Nearest neighbor estimator	193
Nearest neighbor using R	193
Nearest neighbor using Python	195
Decision trees	198
Decision trees in R	199
Decision trees in Python	201
Neural networks	203
Neural networks in R	203
Random forests	205
Random forests in R	206
Summary	207
Chapter 10: Optimizing Jupyter Notebooks	209
Deploying notebooks	209
Deploying to JupyterHub	209
Installing JupyterHub	210
Accessing a JupyterHub Installation	210
Jupyter hosting	211
Optimizing your script	212
Optimizing your Python scripts	212
Determining how long a script takes	212
Using Python regular expressions	213

Using Python string handling	213
Minimizing loop operations	213
Profiling your script	214
Optimizing your R scripts	214
Using microbenchmark to profile R script	214
Modifying provided functionality	215
Optimizing name lookup	215
Optimizing data frame value extraction	215
Changing R Implementation	216
Changing algorithms	216
Monitoring Jupyter	216
Caching your notebook	217
Securing a notebook	217
Managing notebook authorization	217
Securing notebook content	217
Scaling Jupyter Notebooks	218
Sharing Jupyter Notebooks	218
Sharing Jupyter Notebook on a notebook server	218
Sharing encrypted Jupyter Notebook on a notebook server	219
Sharing notebook on a web server	219
Sharing notebook on Docker	219
Converting a notebook	220
Versioning a notebook	220
Summary	221
Index	223
