



INTRODUCTION TO DATA SCIENCE FOR SOCIAL AND POLICY RESEARCH

Collecting and Organizing Data
with R and Python

JOSE MANUEL MAGALLANES REYES

Data science has now firmly moved from computer science and engineering to the disciplines of the social sciences, where scholars are harnessing the insightful power of ever larger and more complex data sets. This volume provides a clear introduction for social scientists and policy researchers into the use of R and Python, including best practice of working with data files, command files, and outputs. The step-by-step approach with real-world examples will be of great value to students, scholars, and practitioners engaged in data analytic approaches to social problems."

– Todd Landman, Professor of Political Science and Pro Vice Chancellor,
Faculty of Social Sciences, University of Nottingham

"The irruption of big data and the need to comply with high standards of research reproducibility require social scientists and policy analysts to be conversant in data collection and management techniques. Unfortunately, even those with sophisticated methodological training often lack the necessary tools to take on these requirements. Magallanes Reyes's book at long last collects and organizes a large amount of information and useful advice on how to curate data for scientific analysis. Through agile narrative and compelling examples, he walks the reader through the use of open-source tools of data science such as R, Python, and Github. The book is an invaluable resource for students and scholars at different levels of proficiency, from neophytes to advanced users."

– Guillermo Rosas, Washington University in St. Louis

"This new, practical, reader-friendly how-to manual on computational social data analysis is both long overdue and a must-have for analysts and researchers. The range of problem-solving strategies and demonstrations is impressive. While eminently practical, Magallanes Reyes's contribution is also rigorous and true to its scientific aims, which will please both basic and applied scientists and practitioners."

– Claudio Cioffi-Revilla, Professor of Computational Social Science and
Director, Center for Social Complexity, George Mason University,
and founding President, Computational Social Science Society
of the Americas

"Magallanes Reyes's excellent book on data science for researchers and policy analysts is an accessible yet thorough introduction to data management and analyses in R and Python. It has a broad coverage of the techniques required to capture, clean, and process complex information. It is the perfect companion for sophisticated policy analysts and researchers that are ready to take advantage of the wealth of data that is available to skilled computer scientists."

– Ernesto Calvo, University of Maryland

Cover image courtesy of © Westhoff / E+ / Getty Images

Cover design by Holly Johnson

CAMBRIDGE
UNIVERSITY PRESS
www.cambridge.org

ISBN 978-1-107-54025-5



9 781107 540255 >

INTRODUCTION TO DATA SCIENCE
MAGALANES REYES
FOR SOCIAL AND POLICY RESEARCH

CAMBRIDGE

Introduction to Data Science for
Social and Policy Research
Collecting and Organizing Data
with R and Python

JOSÉ MANUEL MAGALLANES REYES
*University of Washington and Pontificia
Universidad Católica del Perú*



CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

4843/24, 2nd Floor, Ansari Road, Daryaganj, Delhi - 110002, India

79 Anson Road, #06-04/06, Singapore 079906

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org

Information on this title: www.cambridge.org/9781107117419

DOI: 10.1017/9781316338599

© Jose Manuel Magallanes Reyes 2017

This publication is in copyright. Subject to statutory exception and to the provisions of relevant collective licensing agreements, no reproduction of any part may take place without the written permission of Cambridge University Press.

First published 2017

Printed in the United Kingdom by Clays, St Ives plc

A catalogue record for this publication is available from the British Library.

Library of Congress Cataloging-in-Publication Data

Names: Magallanes Reyes, Jose Manuel, author.

Title: Introducing data science to social and policy analysts : from collecting to organizing data with R and Python / Jose Manuel Magallanes Reyes, Pontificia Universidad Católica del Perú.

Description: Cambridge, United Kingdom ; New York, NY : Cambridge University Press, 2017. | Includes bibliographical references and index.

Identifiers: LCCN 2017020965 | ISBN 9781107117419 (alk. paper)

Subjects: LCSH: Policy sciences – Statistical methods. |

Policy sciences – Data processing. | Python (Computer program language) | R (Computer program language)

Classification: LCC H97.M336 2017 | DDC 300.72/7 – dc23

LC record available at <https://lcn.loc.gov/2017020965>

ISBN 978-1-107-11741-9 Hardback

ISBN 978-1-107-54025-5 Paperback

Cambridge University Press has no responsibility for the persistence or accuracy of URLs for external or third-party Internet Web sites referred to in this publication and does not guarantee that any content on such Web sites is, or will remain, accurate or appropriate.

Introduction to Data Science for Social and Policy Research

Real-world data sets are messy and complicated. Written for students in social science and public management, this authoritative but approachable guide describes all the tools needed to collect data and prepare it for analysis. Offering detailed, step-by-step instructions, it covers collection of many different types of data including web files, APIs, and maps; data cleaning; data formatting; the integration of different sources into a comprehensive data set; and storage using third-party tools to facilitate access and shareability (from Google Docs to GitHub).

Assuming no prior knowledge of R and Python, the author introduces programming concepts gradually, using real data sets that provide the reader with practical, functional experience.

JOSE MANUEL MAGALLANES REYES is an associate professor of political science and public policy at Pontificia Universidad Católica del Perú, senior data science fellow at the eScience Institute, and visiting professor at the Evans School of Public Policy and Governance at the University of Washington, Seattle. His research focuses on social complexity, applying computational thinking to governance issues to inform public policy. Over the past 15 years he has served in government and has been involved in several initiatives in Peru's public sector to make better use of data for policy, political research, and decision making.

“It is rare indeed to pick up a new manuscript and immediately think how much you wish it had been written five years earlier, but I suspect many people will have that reaction to this book. This timely, thorough, and remarkably clear tutorial to both R and Python serves as a much needed on-ramp to the data part of data science, and will undoubtedly soon grace the bookshelves of many social scientists – both students and their instructors. If you are intrigued by the possibilities of data science but concerned about the start-up costs, look no further: help has arrived.”

Joshua Tucker, New York University

“If you need to develop new skills in R and Python but you don’t know where to start, this is the book for you. With simple language, Magallanes Reyes shows you how to install the programs, retrieve data using APIs and scrape Internet sources, and how to get the data ready for modeling. This book is a gem.”

Aníbal Pérez-Liñán, University of Pittsburgh

Contents

<i>Illustrations</i>	page vii
<i>Tables</i>	xi

PART ONE GETTING STARTED

1	Introduction	3
	1.1 Road Map for the Reader	4
	1.2 Main Tools	6
	1.3 Additional Tools	8
	1.4 The Rest of the Book	9
	1.5 For the Reader	10
	1.6 Acknowledgments	10
2	Setting up the Tools	12
	2.1 Installing R	12
	2.2 Installing Python	14
	2.3 Setting up Additional Tools	22
3	Basics of R and Python	24
	3.1 First Contact with R and Python	24
	3.2 Introducing Data Structures	32
	3.3 Functions and Control of Execution	69

PART TWO COLLECTING AND CLEANING DATA

4	Collecting Data	85
	4.1 Knowing Where Your Files Are	85
	4.2 Importing Data Sets	90

4.3	Importing Maps from Shapefiles	102
4.4	Collecting via APIs	105
4.5	Collecting Tabular Data by Scraping	115
5	Cleaning Data	126
5.1	Dealing with Missing Values	126
5.2	Dirty Values	152
PART THREE FORMATTING AND STORING DATA		
6	Formatting the “Clean” Data	163
6.1	Formatting Dates	164
6.2	Focusing on Categorical Data	185
6.3	Data Transformation	191
6.4	Transformation for Comparability/Integration	204
6.5	Formatting Longitudinal Data	208
6.6	Formatting Network Data Sets	217
6.7	A Comment on Complex Survey Design Data	224
7	Integrating and Storing Data	226
7.1	Integrating Data	226
7.2	Integrating Network Data	256
7.3	Storing Your Work	261
7.4	Storing and Google Drive	264
7.5	Storing and Dropbox	271
7.6	Storing and GitHub	275
	<i>References</i>	299
	<i>Index of R and Python Commands Used</i>	301

Illustrations

2.1	The RStudio download site	<i>page</i> 13
2.2	The RStudio GUI	14
2.3	The RStudio GUI options	15
2.4	The Anaconda download page	16
2.5	Anaconda Navigator home	17
2.6	Creating an environment in Anaconda	18
2.7	Finding out the packages available in an environment	19
2.8	Installing a package in Anaconda	20
2.9	The Mac terminal	20
2.10	Environment activated in the terminal	21
2.11	Adding a channel to Anaconda	21
3.1	Finding the icons to run Python and R	25
3.2	R as a calculator	26
3.3	Running the calculator	27
3.4	The Anaconda Navigator	28
3.5	The calculator in Python	29
3.6	First interaction with Python	30
3.7	Creating the basic structures	33
3.8	Results for basic structures	34
3.9	Code for lists and vectors in R	35
3.10	Lists and vectors in R	36
3.11	Creating data frames in R	37
3.12	Differences among data frames in R	38
3.13	Creating data frames in Python	39
3.14	Displaying the data frames in Python	40
3.15	Python data frame and incomplete values	42
3.16	Creating data to be manipulated in R	43
3.17	Installing a package in RStudio	43

3.18	Creating data to be manipulated in Python	44
3.19	Installing a new package in Anaconda	45
3.20	Errors and warnings in R	74
3.21	Errors in Python	76
4.1	Data folder in DropBox	86
4.2	Configuring RStudio to access the data (Mac version)	87
4.3	Creating a new script in RStudio	88
4.4	Configuring RStudio to access the data (Windows version)	89
4.5	SPSS data view	90
4.6	SPSS variable view	91
4.7	SPSS data in R	92
4.8	ANES data center webpage	94
4.9	Fixed-width file folder	95
4.10	File in STATA	97
4.11	Python output for STATA import	99
4.12	Updating from Navigator	99
4.13	Spreadsheet data in Excel	100
4.14	Plotting a map in R	104
4.15	Result of an API search in XML	106
4.16	A record in parsed XML into R lists	109
4.17	Structure of a country when recovered using the API	112
4.18	Wikipedia table to be scraped	116
4.19	HMTL behind a webpage	117
4.20	Exploring results from BeautifulSoup (I)	121
4.21	Exploring results from BeautifulSoup (II)	121
4.22	Understanding tags in HTML	122
4.23	Data frame from scraping in Python	123
4.24	First lines of Pandas data frame built from an ordered dict	124
5.1	Looking for missing data in an SPSS file	130
5.2	Missing data in Python (I)	132
5.3	Missing data in Python (II)	133
5.4	Missing data in Python (III)	133
5.5	Missing data in Python (IV)	134
5.6	Plotting a map with NAs in R	137
5.7	Some dirtiness in a spreadsheet	139
5.8	Clean summary with NaNs in Python	147
5.9	NAs obtained in a scraped table	149
6.1	Defective scraping output	169
6.2	Understanding rowspan issues when scraping (I)	179
6.3	Understanding rowspan issues when scraping (II)	180

6.4	Using <code>enumerate</code> in Python	182
6.5	Scraped table repaired in Python	183
6.6	Comparison of different interval building techniques	199
6.7	Comparing continuous index with k-means clustering	201
6.8	Comparing transformations	207
6.9	Simple forecast plot using time-series data	210
6.10	Panel data in a wide format	211
6.11	Some issues in wide-format output in Python	214
6.12	Basic list of edges	217
6.13	A network plot from an edge list	218
6.14	Adjacency matrix	219
6.15	Zippping lists in Python	222
6.16	An adjacency list	223
7.1	Unmatched values from merging HDI and CODES	232
7.2	Unmatched values for Index of Economic Freedom	241
7.3	Unmatched values from merge of WORLD and CODES	244
7.4	Structure of the merge result in Python	249
7.5	A network plot from an adjacency matrix	258
7.6	Contents of nodes in Python	258
7.7	Plotting a network with color by attribute in Python	259
7.8	Finding Google Drive	264
7.9	The Google Drive environment	265
7.10	Google Drive settings	265
7.11	Google Drive's convert files option	266
7.12	Converting data frames into Google Sheets	267
7.13	Files converted into Google Sheets	267
7.14	Info for a Google Sheet	268
7.15	Making a Google spreadsheet public	270
7.16	Making a Google spreadsheet public as a CSV	271
7.17	Dropbox contents before interacting with them	272
7.18	Dropbox contents after interacting with R	274
7.19	Creating a git repository (I)	276
7.20	Creating a git repository (II)	277
7.21	Creating a branch in GitHub	278
7.22	Changing the default branch	278
7.23	Cloning a repo	280
7.24	Repo in the GitHub desktop	280
7.25	Repo folder in GitHub	281
7.26	Detecting local changes in the GitHub desktop	282
7.27	Setting up the R project	283

7.28	Creating a markdown file in RStudio (I)	284
7.29	Creating a markdown file in RStudio (II)	284
7.30	Saving an Rmd file as a webpage	285
7.31	Structure of an Rmd code	286
7.32	Basic output in HTML	287
7.33	Committing the first codes	288
7.34	Finding the URL of your project	289
7.35	More complex code in R Markdown (I)	290
7.36	More complex code in R Markdown (II)	291
7.37	Getting the link to a data file in GitHub	292
7.38	New GitHub repo for Python	293
7.39	Jupyter icon	294
7.40	Jupyter and console events	294
7.41	Creating a new Jupyter notebook	295
7.42	Elements of the new notebook	295
7.43	First step in running notebooks	296
7.44	Coding in notebooks (I)	297
7.45	Coding in notebooks (II)	297

Tables

5.1	Example of missing values	<i>page</i> 127
6.1	Data types in Python and R	163
6.2	Star dates of calendars in different programs	164

PART ONE

GETTING STARTED

