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
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
Contents

Illustrations

Tables

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


# Contents

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

References 299

Index of R and Python Commands Used 301
Illustrations

2.1 The RStudio download site  
2.2 The RStudio GUI  
2.3 The RStudio GUI options  
2.4 The Anaconda download page  
2.5 Anaconda Navigator home  
2.6 Creating an environment in Anaconda  
2.7 Finding out the packages available in an environment  
2.8 Installing a package in Anaconda  
2.9 The Mac terminal  
2.10 Environment activated in the terminal  
2.11 Adding a channel to Anaconda  
3.1 Finding the icons to run Python and R  
3.2 R as a calculator  
3.3 Running the calculator  
3.4 The Anaconda Navigator  
3.5 The calculator in Python  
3.6 First interaction with Python  
3.7 Creating the basic structures  
3.8 Results for basic structures  
3.9 Code for lists and vectors in R  
3.10 Lists and vectors in R  
3.11 Creating data frames in R  
3.12 Differences among data frames in R  
3.13 Creating data frames in Python  
3.14 Displaying the data frames in Python  
3.15 Python data frame and incomplete values  
3.16 Creating data to be manipulated in R  
3.17 Installing a package in RStudio  

page 13  
14  
15  
16  
17  
18  
19  
20  
20  
21  
21  
25  
26  
27  
28  
29  
30  
33  
34  
35  
36  
37  
38  
39  
40  
42  
43  
43
viii Illustrations

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 HTML 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 Deﬁective scraping output 169
6.2 Understanding rowspan issues when scraping (I) 179
6.3 Understanding rowspan issues when scraping (II) 180
<table>
<thead>
<tr>
<th>Illustrations</th>
<th>ix</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.4 Using <code>enumerate</code> in Python</td>
<td>182</td>
</tr>
<tr>
<td>6.5 Scraped table repaired in Python</td>
<td>183</td>
</tr>
<tr>
<td>6.6 Comparison of different interval building techniques</td>
<td>199</td>
</tr>
<tr>
<td>6.7 Comparing continuous index with k-means clustering</td>
<td>201</td>
</tr>
<tr>
<td>6.8 Comparing transformations</td>
<td>207</td>
</tr>
<tr>
<td>6.9 Simple forecast plot using time-series data</td>
<td>210</td>
</tr>
<tr>
<td>6.10 Panel data in a wide format</td>
<td>211</td>
</tr>
<tr>
<td>6.11 Some issues in wide-format output in Python</td>
<td>214</td>
</tr>
<tr>
<td>6.12 Basic list of edges</td>
<td>217</td>
</tr>
<tr>
<td>6.13 A network plot from an edge list</td>
<td>218</td>
</tr>
<tr>
<td>6.14 Adjacency matrix</td>
<td>219</td>
</tr>
<tr>
<td>6.15 Zipping lists in Python</td>
<td>222</td>
</tr>
<tr>
<td>6.16 An adjacency list</td>
<td>223</td>
</tr>
<tr>
<td>7.1 Unmatched values from merging HDI and CODES</td>
<td>232</td>
</tr>
<tr>
<td>7.2 Unmatched values for Index of Economic Freedom</td>
<td>241</td>
</tr>
<tr>
<td>7.3 Unmatched values from merge of WORLD and CODES</td>
<td>244</td>
</tr>
<tr>
<td>7.4 Structure of the merge result in Python</td>
<td>249</td>
</tr>
<tr>
<td>7.5 A network plot from an adjacency matrix</td>
<td>258</td>
</tr>
<tr>
<td>7.6 Contents of nodes in Python</td>
<td>258</td>
</tr>
<tr>
<td>7.7 Plotting a network with color by attribute in Python</td>
<td>259</td>
</tr>
<tr>
<td>7.8 Finding Google Drive</td>
<td>264</td>
</tr>
<tr>
<td>7.9 The Google Drive environment</td>
<td>265</td>
</tr>
<tr>
<td>7.10 Google Drive settings</td>
<td>265</td>
</tr>
<tr>
<td>7.11 Google Drive’s convert files option</td>
<td>266</td>
</tr>
<tr>
<td>7.12 Converting data frames into Google Sheets</td>
<td>267</td>
</tr>
<tr>
<td>7.13 Files converted into Google Sheets</td>
<td>267</td>
</tr>
<tr>
<td>7.14 Info for a Google Sheet</td>
<td>268</td>
</tr>
<tr>
<td>7.15 Making a Google spreadsheet public</td>
<td>270</td>
</tr>
<tr>
<td>7.16 Making a Google spreadsheet public as a CSV</td>
<td>271</td>
</tr>
<tr>
<td>7.17 Dropbox contents before interacting with them</td>
<td>272</td>
</tr>
<tr>
<td>7.18 Dropbox contents after interacting with R</td>
<td>274</td>
</tr>
<tr>
<td>7.19 Creating a git repository (I)</td>
<td>276</td>
</tr>
<tr>
<td>7.20 Creating a git repository (II)</td>
<td>277</td>
</tr>
<tr>
<td>7.21 Creating a branch in GitHub</td>
<td>278</td>
</tr>
<tr>
<td>7.22 Changing the default branch</td>
<td>278</td>
</tr>
<tr>
<td>7.23 Cloning a repo</td>
<td>280</td>
</tr>
<tr>
<td>7.24 Repo in the GitHub desktop</td>
<td>280</td>
</tr>
<tr>
<td>7.25 Repo folder in GitHub</td>
<td>281</td>
</tr>
<tr>
<td>7.26 Detecting local changes in the GitHub desktop</td>
<td>282</td>
</tr>
<tr>
<td>7.27 Setting up the R project</td>
<td>283</td>
</tr>
</tbody>
</table>
Illustrations

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

6.1 Data types in Python and R

6.2 Star dates of calendars in different programs

page 127

163

164