1
Introduction

This book is based on class notes used to teach undergraduate and graduate students in political science and public policy how to prepare their data to conduct further analysis and provide recommendations to inform decision making. At both levels of education, I found students with different backgrounds in quantitative tools. That required me to develop a teaching approach to address the needs of students new to data analysis, so you should expect detailed explanations in this book.

My courses address different skills: from collecting, organizing, modeling, and analyzing the data, to visualizing and publishing it. This book focuses on data organization and collection and all their related procedures. This stage is arguably the most time-consuming and has been my main concern over the years. In general, books for quantitative analysis devote most of their pages to teaching quantitative concepts while sharing data that has already been well organized into a table readable by most software (Stata, SPSS, SAS, etc), or into a simple comma-separated values (CSV) file to be used by any software. However, every scholar (student, researcher, or professor) or professional will need to prepare a particular data set for his or her current needs. It is then that they may experience a hard time. The literature lacks material on preparing data sets for students in the social sciences, so it has been my goal to provide those competences in basic data organization.

When I embarked on teaching this stage to students, I discovered there is no simple, right, or unique way of doing it. In this book, I simply share my way of teaching it, not how it should necessarily be done. This is a potential challenge or weakness of this book because I do not present a formal paradigm for data organization. Nevertheless, I share strategies that have proven to be helpful for students who lack a quantitative or computational background.

My target audience might be familiar with particular software, but may never have used it to collect or clean data or to automate those processes. There is

no function in an SPSS/Stata/Excel window menu that says either *collecting* or *cleaning* the data. Collecting, cleaning, and formatting data require the flexibility that programming languages provide, which motivated me to teach some coding skills to my students.[1]

1.1 Road Map for the Reader

The use of data has become so pervasive that there is now a field of *data science*. It seems that it is not enough that we already have qualitative disciplines rich in empirical data collection techniques (anthropology, sociology, psychology, history, etc.) and quantitative fields that have dealt with data modeling (statistics, operations research, economics, and so on); now, we feel the need to create a science of data. Rather than being a completely new field, data science is emerging from ad-hoc interdisciplinary combinations that nevertheless use a common tool: the computer (hardware and software).

Advances in computing technology have not simply led to "cool" personal gadgets but also, most importantly, to a different strategy for doing science by means of algorithmic thinking. An algorithm is a finite set of instructions to process some input and effectively reach an output; algorithms enable you not only to create computer programs but also to design a teaching lesson, devise a strategy, establish organizational norms, and so on. Because algorithmic thinking is so closely related to design,[2] it is a key approach to understanding institutions and designing policies. Although I recognize it may be too early for most people to make the connection between binary codes and policy making, I am consciously in favor of calling your attention to this relationship.

This book will neither make a case for data science nor turn you into a data scientist, but it will teach you the first steps to becoming a user of tools to effectively deal with data. And, in my experience, that instruction has been enough to motivate policy scholars to become computational policy analysts.

There are several stages for doing data science, starting with deciding what to study and ending by proposing what actions to take. As you see, the process is no different from a traditional research design where research questions will

[1] If you do not want to learn R or Python while doing the tasks of this book, you may want to take a look at some programs that offer cleaning and formatting capabilities like *Wrangler* (http://idl.cs.washington.edu/papers/wrangler), *OpenRefine* (http://openrefine.org), or *Trifacta* (https://www.trifacta.com).

[2] You can detect my closeness and empathy to Simon (1996) in these lines.

drive much of the project. However, the source of the question is not mainly theoretical as in a dissertation, but is also intuitive, as in any decision-making situation. Once you have identified what matters for your research, you will first get the data, which will be analyzed and interpreted into findings and then shared and stored. This book deals with the first and third steps: getting the data you need, cleaning and formatting it, and sharing and storing, all with the goal of increasing its usefulness.

But, what is the meaning of "getting the data you need'? This simple road map should give you a better idea:

1. **Relevant sources have been *identified* and confirmed to be *trustworthy***. To realize your research goals, you must know what data you need and where you can get it. Most of the time, you also need to know if the source is trustworthy: Do you believe those police reports? Would people be likely to answer this census question honestly? Is this a good proxy for this unobserved social feature? Would this variable be a good instrument for your posterior regression analysis? A mistake here may cause your data to work poorly in the next stages. Knowledge and experience with your subject is more helpful than computer knowledge at this stage.
2. **Data sources have been obtained**. This is the data collection stage. If you requested a survey, you would get a file with the results; if the data you need is in a web repository, you can download it; you may even need to get the data from websites or Twitter messages. Most of the time, you are collecting data from different sources, and this will take a while because some data, by its nature, has some legal constraints and/or technical limitations. You may even wonder how to get and use the data in an ethical manner. Although you may want to be as exhaustive as possible, remember that you need to finish the research in your lifetime. Two key consideration here are the unit of analysis and the time frame. The data collected should share the same units of analysis; that is, if you need information at the city level, you may welcome data at the neighborhood level, but you do not want data at the state level or country level (too aggregated). Similarly make sure that the data is for the same period of time; that is, you do not want one variable from the year 2000 and another from the year 2010; if you are collecting one variable for different years, you will want other variables for those same years. Respect your research goals, and work hard at this stage to keep your data collection coherent.
3. **The values of the data obtained have been cleaned**. Do not expect that all the files you obtained are ready to be compared, integrated, or analyzed. At these later stages the computer needs to read each value you see as you do.

Imagine a spreadsheet with symbols in each cell that may not be correctly interpreted by other software (i.e. ≈, $, €, etc.), or cells with missing values typed in strange ways (`n/a`, for instance); a cell can have footnotes that inform you of something important, but will add an irrelevant value to your data. The problem you have in one file may be different from the one you have in another one. You will be surprised by the different kinds of dirtiness you will encounter. This a very time-consuming stage; just be patient and do your best.

4. **The values have been *formatted***. Your value is now well read by your computer, and it is syntactically valid so far. But now comes a problem of semantics, which again needs your intervention. You may need to tell the computer that the number being read is not an integer but a date or a category, and you need to be careful to apply the right kind of function to those values. At this time, it is even necessary to organize the clean data source into another data structure, particularly if you are planing to analyze longitudinal or relational (network) data. Some people may still call this organizational stage "cleaning"; however, formatting, in this book, is considered a different process than cleaning.
5. **The multiples sources of data have been *integrated and saved***. The last stage combines all your data sources and stores them. You cannot expect that all the variables you need will come from the same source, so you need to find a mechanism to merge all your files. Even though each file is clean and well formatted, you may still find some issues when trying to combine them into one. Once you know they can be well combined, you can safely store what you have, so that the next steps in data science work smoothly.

If you have been used to receiving files ready for analysis and have worked with only one kind of data (censuses, surveys) in your organization, these steps may not be familiar to you. If you need to get and analyze data from different sources, this book will be helpful by sharing the tools you need to do this work; it will give you the flexibility of writing code to automate these processes.

1.2 Main Tools

This book introduces readers to *programming languages*. Readers should not expect to become programmers; instead, they should expect to learn the bright side of programming to make their job easier. There are hundreds of programming languages. From the set of programming languages, I offer you two very popular high-level languages: Python and R.

Why R or Python?

I chose these languages not because they are necessarily the best ones, but because their use is spreading beyond the computer science domain. That has not happened before with JAVA or C, and as powerful as those languages are, policy analysts may find no reason to learn them to carry out their customary tasks. But, as data becomes available in huge amounts and varying complexity, many analysts are feeling the need for a more flexible way to deal with data organization, analysis, and visualization. This book will use R and Python to deal with the organization stage, showing you how some coding can put more power in your hands.

Python and R are also attractive because they are well documented, with lots of applications in different areas of knowledge and active communities in the web sharing code and examples; and, of course, you can use R and Python free of charge. Some other details follow:

- **R** is a high-level programming language. That is, its creators have tried to abstract it, so that some commands are in English. In contrast, low-level languages *talk* to the computer in a language closer to what a computer understands (but far from direct human understanding). Because it is free of charge and open source, R has allowed many scholars not only to carry out data analysis but also to contribute to R itself by adding very specific functionalities, so that R now has support for almost any kind of quantitative technique. It has been said that R poses a slow learning curve at the beginning, but I find this statement to be quite inaccurate. Working with R is very easy. However, that depends in the coding style the user develops. This book emphasizes a good coding style and habits to make R usable.
- **Python** is an all-purpose programming language with many more features for computational work than R has, but for the goals of this book, these differences are not spelled out in the book. I can only say that you cannot build sophisticated information systems with R, but certainly can with Python. Python, as well as R, has a very active community of users and developers, and you can practically write a question about Python in your browser and find an answer. I would suggest reading this book first, before posing any questions, because the repliers are often advanced experts using jargon that a novice user may not understand.

Both R and Python have something in common that you will soon realize: They both need to install and use external programs called *packages / modules / libraries* (I will use these terms interchangeably). These packages are very useful when you need to carry out complicated tasks with a large amount of code, because it is likely that someone has created a package that does what

you need. You do not need to know every package available, but progressively you will become familiar with ones that will save you lots of coding time. As for speed, regular users may not find a difference between Python and R. If you find that computing results is taking too much time, writing a better code can increase the speed, but advanced programming techniques require more preparation, and the code may turn out to be more difficult to read. This book will not turn you into an advanced programmer, but into an effective user of both languages to deal with situations similar to the examples presented here. Your code may be very good, so take into consideration that other factors also affect speed, such as the size and contents of the file, your hardware (laptop, tablet, etc.), your internet provider, your operating system version, the memory available, and so on. For the examples in the book, speed will not be an issue.

I will show you to use Python and R in particular environments:

- **Anaconda**
- **RStudio**

These two programs will make the use of Python and R easier. Anaconda is a free Python distribution that includes the most popular Python packages needed in this book for data analysis: It is almost ready to be used without much downloading. Nevertheless, Anaconda has a simple way to add packages not already included, and I will highlight that when needed. RStudio is an environment that makes the R experience easier and more versatile. However, it does not include R, so you will need to install R first. RStudio is also free.

Both RStudio and Anaconda offer business options, which are not free. These options are needed to deploy large-scale applications. Not everything is free in the data analytics business, but the free versions are enough for all applications you need for academic work.

1.3 Additional Tools

In the world of data science, programming languages are used in combination with lots of other tools to organize the workflow, improve performance, allow for collaboration, and so on. In this book, I will talk about some additional tools for file management. I do not want this to be a burden and you can feel free to not use them; however, I recommend paying attention to the sections where I describe them, in case you believe they are worth learning to help your work.

These are:

- **Google Drive**
- **Dropbox**
- **Github**

All are free. However, you may need to purchase space if you want more than what is offered freely. Google Drive will give you 15 Gigabytes (Gb);[3] Dropbox gives you 2 Gb;[4] and GitHub has no established limit for an account.[5]

Both Google Drive and Dropbox are used very widely. They offer an easy way to create a folder in your computer to store your files and have them available anywhere. I will guide you how to set them up. GitHub is not as common among social scientists as the others, but it will be helpful to become familiar with it. GitHub is very convenient to use when you are programming (developing software) because it is conceived for *version control*.[6]

1.4 The Rest of the Book

In the next chapter, I will guide you to get your computer ready for Python, R, and the associated tools. The installation chapter is very important, so please follow the directions before going on.

Part Two deals with data collection and cleaning, including reading different formats and collecting data from online sites. Because data from different sources may not be ready to be used immediately, the data cleaning process is essential but very time consuming. The chapters in this part seek to make you a more efficient data collector and cleaner. Even if you think that you will only be working with well-formatted data, I recommend paying close attention to these chapters in case you ever need to do data collection and cleaning. In my early years of teaching I always used Excel to do cleaning, and in fact, I still do so for small data sets. Most people in social and policy sciences may also use it for the same purpose. I will show how R and Python can help us in these tasks and be as least as efficient as Excel or any other program in doing these.

Part Three deals with data organization. After reading these chapters, you will be able to make the data ready for modeling, analysis, and visualization (not covered in the book). There are several issues to deal with when preparing your data, particularly deciding if your data will go through posterior

[3] This includes the space for your email account
[4] Free space can increase up to 16 Gb if you invite friends to become Dropbox users, because you get 500 Megabytes (Mb) for each referral.
[5] Please read these restrictions: https://help.github.com/articles/what-is-my-disk-quota/.
[6] This concept is further developed here: http://oss-watch.ac.uk/resources/versioncontrol.

cross-sectional or longitudinal techniques. I will cover these issues including the format for networks and maps.

1.5 For the Reader

In this book, each section depends on something previously presented. It is not a long book, so I would recommend you read it from beginning to end. Of course, if you know most of the material covered, you are most welcome to visit the chapter or section you believe you need, while skipping previous material. I tried to include references to previous sections as needed.

This book has been written for students, professors, and professionals in the social and policy sciences at all levels. It can be used for self-learning and to complement any quantitative analysis course.

I will request that you collect some data using some links, but in case those links are not working when you read this book I will maintain a Dropbox with a copy of those files. I hope you have fun while learning.

1.6 Acknowledgments

I am very grateful to Cambridge University Press for giving me the opportunity to share my work. I am also very grateful to the reviewers who took the time to comment on and suggest improvements for my initial drafts.

This book was made possible by the research appointment I was granted by the eScience Institute at the University of Washington (UW) and my teaching experience while a Visiting Professor at the UW's Evans School of Public Policy and Governance. I have to express my particular gratitude to Ed Lazowska, Bill Howe, Magdalena Balazinska, Tyler McCormick, Bernease Herman, Valentina Staneve, Anthony Arendt, Micaela Parker, and Sarah Stone from the eScience Institute; and to Sandra Archibald, Craig Thomas, Mark Long, and Greg Traxler from the Evans School. They were very supportive during my stay at UW and helped me finish this book in different ways. I am also thankful to the graduate students at the Evans School, whose feedback about my teaching helped me better organize these contents. I also very grateful to my learning, researching, and teaching experience I had at George Mason University. My interactions with Claudio Cioffi-Revilla, Robert Axtell, William Kennedy, Qing Tian, and Andrew Crooks were critical to fine-tuning my exploration of computational social science. During my stay as a Visiting Scholar at Duke University's Social Science Research Institute, I was able to make the

final revisions to this work. I am very thankful to Scott De Marchi for that opportunity and to the EITM class of 2016 for sharing some relevant ideas on what a book like this should contain.

I also express my deepest thanks to my home institution, the Department of Social Sciences of the Pontificia Universidad Católica del Perú, which has made every effort to support my stay in the United States. I particularly appreciate the support of Alejandro Diez, Alan Fairlie, Aldo Panfichi, Edmundo Beteta, Francisco Durand, Jorge Aragon, Ismael Muñoz, Carlos Alza, David Sulmont, Eduardo Dargent, Sinesio Lopez, and Catalina Romero. I would also like to express special thanks to my former teaching assistants who were always there to support and improve my work, especially Jose Luis Incio, Noam Lopez, Maria Paula Brito, Samuel Sanchez, Kely Alfaro, Fernando Contreras, Jorge Abanto, Marylia Cruz, Jorge Vela, Luis Mas, Alejandra Ocaña, Mariana Ramirez, Valerie Tarazona, Rodolfo Benites, Silvana Rebaza, Yamile Guibert, Manuel Figueroa, Rafael Arias, Roberto Diaz, Rosa Arevalo, Veronica Hurtado, Juan Carlos Gonzales, Paolo Rivas, Mariale Campos, Maria Alejandra Guzman, Angela Bravo, Andrea Moncada, Jeniffer Perez, Daniela Lopez, Lorena Levano, Sakimi Leon, and Mariela Mosqueira.

And of course, my eternal thanks to my partners in all my adventures, Diana, my wife, and our son, Rafael.