Contents

Chapter 1

Introduction

The objective of this textbook is to provide you with the shortest path to exploring your data, visualizing it, forming hypotheses and validating and defending them.

Given a data set, you want to be able to make any plot you wish, find plots which show something actionable and interesting, explore data by slicing and dicing it and finally present your results in a statistically convincing manner, perhaps in a colorful and visually appealing way. Finally, you will be able to apply some basic machine learning methods to build, train and test prediction models. All of this will be accomplished in a succinct and crisp way using a small subset of R instructions.

It is an active textbook. It assumes no prior programming background. We will teach you as little R as possible to achieve the goals of this book which are quite impressive. In fact you will be able to do a good chunk of work which data scientists do. We will accomplish this goal through active snippets of executable code. These are examples of R code (**around 80 executable snippets of code**) embedded in the textbook itself. More importantly, you will be able to modify the code and execute the modified code without need to install any application on your machine. This will allow you to understand the code in the book through "what if" exploratory process. Thus, every code snippet is just an invitation to endless modifications. This is why we call this textbook - an active textbook.

Another unique aspect of this textbook is its reliance on data puzzles. These are synthetic data sets with embedded patterns and rules generated by our tool called **DataMaker**. We will present our data puzzles (dynamic list, may vary from year to year) in section 4 and follow by showing how to make plots of data. Then we will proceed to freestyle data exploration. This will allow us to learn more about our data, form the leads, and finally state our hypotheses. We will follow up by an elementary introduction to hypothesis testing through a

permutation test. We will learn how to calculate p-values and how to use them to defend our findings against the randomness trap.

We will use as few R functions as possible to achieve our goals. In fact we will demonstrate how using **less than ten R functions** we can perform quite sophisticated data exploration. In the appendix, we show many more useful commands of R which eventually you would have to use. However, our goal in this short textbook, is to present the shortest path to data analysis which will let you import the data, plot it, make some analysis yourself and use R-libraries to build machine learning models. In this textbook and in this class we do not teach how to clean the data (data wrangling) and how to deal with a wide variety of data types. We also do not address complex data transformations such as multi-frame operations like merge function. We also do not explain how different machine learning methods work, we only show you how to use them. It is similar to teaching one how to drive a car without knowing how a car engine works.

Sections **5.6** and **5.7** provide the lists of all concepts which we cover in our active textbook and all R functions which are needed. Notice how small the set of R functions is. It is important for programming novices to start small and also see how far this small set of functions can get you.

Our **question roulette** allows self-testing on nearly 100 questions relevant to the material. Each question is answered, but students are encouraged first to answer questions themselves and only then follow it with checking the correct answer.

Acknowledgement: This textbook would not have been created without extensive help from Devarsh Shah

Chapter 2

Best Works of 2022

2.1 DataBlog

Ella Walmsley

2.2 Prediction Challenge 1

Upsham Naik

Jeevanandan Ramasamy

2.3 Prediction Challenge 2

Jeevanandan Ramasamy

2.4 Prediction Challenge 3

Eva Zhang

2.5 Boundless Analytics

Anastasiya Chuchkova

Shreya Tiwari

George Basta

Paul Kotys

Selin Altimparmak

Chapter 3

Data League Leaderboard

Honourable Mentions: Upsham Naik, Joshua B. Sze, Kirtan Patel, Maria Xu, Devam Patel, Eva Zhang, Toshanraju Vysyaraju, Maanas Pimplikar, Jared Chiou, Nitya Narayanan, Shrish Vellore, Yousra Belgaid, Mitali Shroff, Michael Jucan, Jackie Hong, Arvin Sung, Eric Xuan, Eva Allred, Leah Ranavat, Nami Jain, Gautam Agarwal, Aditya Patil

Table 3.1: Leaderboard 2022

Rank	Participant.Name
1	Jeevanandan Ramasamy
2	George Basta
3	Joyce Huang
4	Jiaxu Hu
5	Dhiren Patel
6	Chicheng Shao
7	Cheyenne Pourkay
8	Christopher Nguyen
9	Aaron Mok
10	Ethan Matta

Chapter 4

Data puzzles secrets

- **Lecture slides:** Data Exploration

4.1 Moody Data Puzzle

Moody Data Puzzle is our first example of a data puzzle. By data puzzles we mean synthetically generated data sets which have some embedded patterns. Your goal is to find the embedded pattern(s). You may also find patterns similar (implied) by patterns embedded in the data puzzle. This is fine too. The goal of data puzzles is to excite you about exploratory data analysis. In many ways it is like a game.

Puzzle description:

Professor Moody has been teaching statistics 101 class for many years. His teaching evaluations went considerably south with the chief complaint: he DOES NOT seem to assign grades fairly. Students compared their scores among themselves and found quite a bit of discrepancies! But their complaints went nowhere since Professor promptly disappeared after posting the final grades and scores.

Table 4.1: Snippet of Moody Dataset

SCORE	GRADE	DOZES_OFF	TEXTING_IN_CLASS	PARTICIPATION
21.33	F	never	never	0.29
71.57	C	always	rarely	0.11
90.11	A	always	never	0.26
31.52	D	sometimes	rarely	0.03
95.94	A	always	rarely	0.21

A new brave TA, managed to get hold of the carefully maintained grading table (spanning multiple years) of professor Moody by ….messing a bit with Moody's computer….well, let's not explain the details because he would get in trouble. What he found out was a remarkably structured account of how professor Moody assigns his grades.

Looks like Professor Moody is in fact very alert in class. He is aware of what students do, detecting texting during class and remembering exactly who was dozed off in class. He also keeps the mysterious "participation index" which is a numerical score from 0 to 1. This is probably related to questions asked and answered by students as well as their general attentiveness in class. Remarkable but a little creepy, isn't it?

What is the best advice the new TA, can give future students how to get a good grade in Professor Moody's class? What factors influence the grade besides the score? Back your recommendation up with plots and evidence from the attached data.

What are examples of patterns we are looking for here?

Here are some:

- "Students who text a lot" have lower chance to get an A in the class"
- "Students whose participation is lower than 0.25 fail the class more often"
- "Dozing off does not matter if your score is more than 90, you still get an A"
- "If you score is less than 30, you fail the class regardless of what your other attributes are"

4.1.1 Secrets Revealed- Patterns in Professor Moody's data?

My Patterns

Many student solutions falsely attribute higher grades to higher values of participation attribute..

This is a classic example of a hidden variable described in the reference attached below.

The truth is that participation attribute value impacts the score attribute value. Generally, the higher the participation in Moody's class, the higher the score. But it is the score attribute which has a direct impact on the grade. Thus, it is the score which is the real "hidden variable" impacting the final grade.

Thus, the score already reflects participation. Professor Moody seemed to look only at texting and dozing off attributes in grade determination (see the power points above with the explanation)

Table 4.2: Snippet of Movies Dataset

	country	content	imdb_score	Gross	Budget	genre
2012	USA	R	5.90	Medium	High	Drama
4180	USA	R	7.07	Medium	Medium	Comedy
21	USA	PG	7.22	High	High	Family
11915	USA	R	5.22	Low	Low	History
4430	France	PG	6.86	Low	Medium	Comedy

Compare with examples of hidden variables in the following reference about correlation and causation.

https://www.stewartmath.com/precalc_7e_dp/precalc_7e_dp6.html

4.1.2 Best Student's Submissions 2022

Lauretta Martin

Sanjaya Budhathoki

Sandhya Senthilkumar

4.2 Movies Data Hunt

Puzzle description:

contains imdb scores of 12,800+ movies along with several attributes including budget, gross genre, content rating etc.

What are the most promising alternative hypotheses about imdb scores to test? Name your three top candidates along with the evidence which backs them up: either in the form of R instruction(s) or plot.

4.2.1 Secrets Revealed- Patterns in Movies data?

Secrets Revealed

4.2.2 Best Student's Submissions 2022

Joshua Sze

Andrew Fasano

Table 4.3: Snippet of Minimarket Dataset

	BREAD	BUTTER	COOKIES	COFFEE	TEA
6585	0	0	0	0	0
1998	0	0	1	1	0
5835	1	1	0	0	0
7207	0	0	0	1	1
7192	1	0	0	0	1

4.3 Minimarket Data Hunt

Puzzle description:

Each row of Minimarket.csv contains one customer transaction which is represented as binary vector (treat this as NUM values). 1 means that customer bought an item, 0 - means that customer did not but that item. For example if customer bought Bread but did not buy Butter you will see 1 in the Bread column and 0 in the Butter column.

Summary

Here's what you'd do: 1. Come up with a null hypothesis: "Bread does not impact the sales of butter" 2. Come up with an alternative hypothesis: "Bread impacts the sale of butter" 3. Compute the mean value of the Butter column for all the rows where Bread value = 0. Let's say this is mean1. 4. Compute the mean value of the Butter column for all the rows where Bread value = 1. Let's say this is mean2.

4.3.1 What were the secret associations between items in the minimarket?

Secrets Revealed

4.4 Predicting grades in Professor Moody's class

4.4.1 How did I cook the Professor Moody Prediction challenge data?

Secrets Revealed