Contents

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.

Questions which you will have to anticipate and you will have to answer are - How do you know that your findings are not random? - And fundamental of all questions: - **So what?**

Even the most impressing looking results may come up randomly. And you will be asked this question along with the question *"what was your p-value and how did you compute it"*

And even if you convince your audience that your results are not random, you will have to be ready to explain why your audience should care about the results you reported. In other words, is there any actionable value in your results? Or they are just simply interesting, good to know, but no one really needs to care much about them otherwise? Hopefully it is the former not the latter.

In the following sections we will address these questions and go through the process of data exploration, validation, and presentation.

- We will start with making plots, follow with free style data exploration – which allows us to form the leads, that is hypotheses. Then we will follow with simple statistical tests which will allow us to validate these hypothesis and defend our findings against randomness claims. - We will learn how to calculate p-values and how to use them to defend our findings.
- We will use as few R commands as possible and reach our goal in the shortest possible path. In fact we will demonstrate how using just 7 R commands 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. 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 (we show them in appendix). 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.

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

2 Best Works of 2022

2.1 DataBlog

Ella Walmsley

2.2 Prediction Challenge 1

Upsham Naik

Jeevanandan Ramasamy

Table 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

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

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

4 Data puzzles secrets

- **Lecture slides:** Data Exploration

Table 2: 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

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.

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.

Table 3: Snippet of Movies Dataset

	country	content	imdb_score	Gross	Budget	genre
12257	USA	R	5.40	Low	Low	Drama
5054	USA	R	6.50	Medium	Low	History
4680	USA	R	7.49	High	Medium	Drama
11958	USA	PG-13	7.00	Medium	Low	Drama
4504	USA	PG	5.60	Medium	Medium	Comedy

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)

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

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.