Statistical Principles in Data Science Group Project

Winter 2022

1 Background

The goal of the group project is to work in groups and to apply the statistical methods you learned from this course to a data set chosen by your group.

To help you with choosing your data set, here are links to two common sources of data used by students all over the world:

UC Irvine Machine Learning Repository

Kaggle

2 Project Proposal

The very first step of the group project is to form a project proposal of your group. In this step, you and your teammate need to find a dataset and come up with a realistic problem your team try to solve. The structure could be:

- Name and Student ID of all teammates.
- Title of the project.
- Introduce your data (the source of the data, variables description, background information etc)
- The realistic problem your team try to solve from this data. And why it matters (motivations). (around 200 words)
- The statistics method will be used (Regression or Classification). You don't have to be very specific.

- A rough timeline and plan how your team will proceed to accomplish this project. (How often your team meet each other? Who will responsible for Introduction etc)

Note that we ask you to provide a plan, but it does not have to be very specific. Just need to show your team thinks through every component you will need in the final project. The deadline of the submission of this proposal will be posted on the Moodle.

3 Analysis for Final Report

For this group project, you are required to write a complete and readable report that contains the following sections:

- Abstract
- Introduction
- Methods
- Results
- Discussion
- Appendix

Note: Your goal is not to use as many methods as possible. You should use appropriate methods to find a desired model for the statistical analysis task. Hence, please provide explanations why you chose a particular method. Also, discuss why you think that you completed your task and how well the goals were achieved.

The details of what you should do in each section is given below.

3.1 Abstract

The abstract is the first thing to appear in the report. However, it should be the last thing you write. Generally, the abstract should serve as a summary of the entire report. Reading only the abstract, the reader should have a good idea about what to expect from the rest of the document. Abstracts can be extremely variable in length,

but a good heuristic is to use a sentence for each of the main sections of the IMRaD organization structure: 1. Why are you doing this analysis? (Introduction) 2. What did you do? (Methods) 3. What did you find? (Results) 4. What does it mean? Why does it matter? (Discussion)

3.2 Introduction

Essentially, in this part, you need to motivate your work and put it into context. Why does this analysis need to be done? What is the goal of this analysis? The introduction should also provide enough background on the subject area for a reader to understand your analysis. Do not assume your reader knows anything about the subject area that your data comes from.

A complete data dictionary should be provided in the appendix rather than the introduction, but you can include some data, which are helpful to motivate of the analysis.

Also, you can consider including some results of an exploratory data analysis (EDA) here, if you feel it helps to understand the data.

3.3 Methods

For this project you have to use **at least TWO** methods learn from the course to analyze your data. For example, you can fit regression model to make predictions and use PCA to explore the underlying structure of the data. *Note that fitting two regression models does not count as two methods.*

In this section, you should discuss what you have done, including the choice of an appropriate model. Which method did you apply, and why is it appropriate in your context? You are allowed to use any method you learned from this course. This section will contain most of the `R` code that is used to generate the results. Your R code is not expected to be perfect idiomatic `R`, but it is expected to be understood by a reader without too much effort. The majority of your code should be suppressed from the final report, but consider displaying code that helps to illustrate the analysis you performed.

3.4 Results

The results section should contain numerical and graphical summaries of your results. What was the output from your chosen methods? Consider reporting a "final" or "best" model you have chosen. There is not necessarily one, singular correct model, but certainly, some methods and models are better than others in your situation.

3.5 Discussion

In this section, you should discuss of your results. What do your results mean? Usually, results are just numbers, hence you need to explain what they tell you about the analysis you are performing. How good are your results? The results section tells the reader what the results are. The discussion section tells the reader why those results matter.

3.6 Appendix

The appendix section should contain any additional code, tables, and graphics that are not explicitly referenced in the narrative of the report. The appendix must contain a data dictionary.

4 Read the following carefully

- Please submit your report as q `.pdf` file.
- Load all your package at the beginning of the code.
- Only include the most importance plots, outputs, and tables in the main body of the report. Other result goes to appendix.
- You will lose points with unreadable plots. This means you should always use meaningful labels and title with a plot.
- Make sure that your analysis is meaningful. You should choose an appropriate method and follow the guidelines concerning the structure of your report.
- Try to write your report in good English!

- The length of the report should be around 10-20 pages, not counting the appendix.

Peer Evaluation

It is important to evaluate your peers' contribution to this project. We are expecting every group members to equally contribute to the project. It would not be fair, if one person does all the work. You need to state what exactly each team mate contributed to the project. For example, you might write; "My team mate Tim was mainly responsible for finding the data and clean the data for us to use. He participated in every group meeting, and also wrote the "introduction" section of our report." This evaluation should be in a separate document and turned in by each participant separately.