CSDA-1 Final Project-Spring 2021

Prof. Juming Pan

General Information

The goal of this project is to enhance your understanding of the processes involved in statistical analyses and especially in displaying and interpreting data. This project consists of two parts: 1) Demonstration of abilities to use R to perform statistical analyses, 2) A report on the findings of the analysis performed.

- Before performing the analyses, carefully read the description in the file (see below).
- This is an individual project and all work must be your own.
- For all analyses you MUST use R. Analyses using other software packages receive no credit.
- The last day for asking questions on the project is two days before the project is due.
- A 10% penalty for each part will be applied for each calendar day beyond the deadline.

Data File

A list of data sets are available at UCI Machine Learning Repository `https://archive.ics.uci.edu/ml/datasets.php`. A data can only be used by one student; if a data has already been approved for a student, it will not be approved for any other student (first-come, first-serve basis). Please use the Google Doc in below to claim which the data you want to summarize.

`https://docs.google.com/document/d/1DncZgFOKJt5dVgGLp3E7g19_ChjBcDbLKzDcuzZOmzY/edit?usp=sharing`

1 Proposal (25 pts) – due March 15 by 11:59 pm

Goal

Your goal is to submit a proposal giving me enough detail about what you want to do in your project, so that I can give you feedback before you set out to complete it. You should also complete a thorough exploratory data analysis so that you familiarize yourself with your data and decide whether or not the data are appropriate for the project and whether you should be expecting surprising results.

Content

Your proposal should consist of the following sections:

- A list of research goals (at least 3) for the project. For example, what questions do you hope to answer? What story you want to tell? Research question: In one sentence, what is your research question? This should be something you are genuinely interested in! Then provide a sentence or two explaining why you are interested in this question.
- Data
  - Data source: Include the citation for your data, and (if available) link to the source.
  - Data collection: How was the data collected? Who collected the data and when it was collected?
  - Variables: What are the variables you will be studying? Are they numerical, categorical?
  - Scope of inference: Can these data be used to establish causal links and/or can findings be generalized to the population at large?
  - Data clean-up: (Optional) If you had to do any data clean up in R, you can include the code and a very brief description of your steps here.
- Exploratory data analysis: Perform relevant descriptive statistics, including summary statistics and visualization of the data. Think about what conditions you might need to check for your analysis or what summaries of the data might be useful for answering your question. Address what the exploratory data analysis suggests about your research question.

Print out 1 page of your data set and attach it to your proposal. If your data fits in one page, great. If you have too many observations and it won't fit, that's ok too. I just want to get a sense of your data set, I do not need to see all rows. However your print out should contain all relevant columns.

Format&Length

Your proposal should be at most two pages. Keep it brief as the information on this document will eventually make it into your project. If you make it too long you'll end up having to cut it down later. I would like any relevant figures and R code (e.g., from your exploratory data analysis). The R code and the print-out of you data (just one page) can be included at the end of your proposal as an appendix and will not count toward the limit of two pages.

Submission

Proposals are due at 11:59 pm on March 15 (Monday) via Canvas. Late work policy applies (10% off for every 24 hours late).

2 Project Report (75 pts) - due May 4 by 11:59 pm

Goal

Your goal is to submit a cohesive project report that conveys that you have mastered statistical inference techniques that we have learned in class and that helps you answer your research question.

Content

Your project should be a write up of the parts below in the form of a research paper.

- Part 1: Introduction. What is your research question? Why do you care? Why should others care? If you know of any other related work done by others, feel free to include a brief description
- Part 2: Data. This should be a cohesive write-up of the data section from your proposal (not just a list of numbers).
- Part 3: Exploratory data analysis. Perform relevant descriptive statistics, including summary statistics and visualization of the data. Also address what the exploratory data analysis suggests about your research question.
- Part 4: Inference. Perform relevant descriptive statistics, including summary statistics and visualization of the data. Also address what the exploratory data analysis suggests about your research question.
  - Check conditions
  - Theoretical inference (if possible) - hypothesis test and confidence interval
  - Simulation-based inference (if theoretical not possible) - hypothesis test and confidence interval

  If your data fails some conditions and you can't use a theoretical method, then you should use simulation. It is your responsibility to figure out the appropriate methodology, but feel free to ask me. Be sure to interpret your findings in context of the problem.

  Important note: If you're using a categorical variable with more than two levels, for the inference section you may want to either combine or drop some of the levels, so that you only have two levels to work with. This only applies to the inference section, there are no restrictions on the number of levels you can use for the exploratory data analysis.
- Part 5 Regression. Find the correlation coefficient describes the strength and direction of the linear association between two numerical variables, obtain least squares line minimizes squared residuals and interpreting the least squares line. Finally, predict, but don't extrapolate.
- Part 6 Conclusion. Summarize your findings without repeating verbatim your statements from earlier. Also include a discussion of what you have learned about your research question and the data you collected. You may also want to note limitations in your study and include ideas for possible future research.

Format&Length

Your write up should be at most 7 pages (including figures and R code) and submitted as a pdf. This is not very long, so you will need to be concise. Every sentence should add something to your paper. I would like any relevant R code to be given in an appendix. Figures may be inserted into the body of the report (preferred) or in an appendix. All graphs and tables need to be resized so that each covers at most 1/6 of a standard page. I prefer single-spacing, but double-spacing is also acceptable.

Tone

Write as if you are explaining your results to whoever would be interested in your research question, whether this is other scholars in your field or peers sharing your interest in the topic. Keep in mind this audience may or may not have taken statistics. You must be statistically accurate and use correct statistical terminology, but must also explain your conclusions in a way that anyone can understand. Remember to interpret your findings in context of your research question.

Submission

Reports are due at 11:59 pm on May 4 (Tuesday) via Canvas. Late work policy applies (10% off for every 24 hours late).