CS Principles - Data Analysis Project Requirements

As it becomes ever easier and cheaper to collect and store vast amounts of data, all fields of human endeavor are transforming into data-intensive fields. No matter what career path you seek, you will likely need to be adept at manipulating, analyzing, and interpreting data at scales never required previously. The following quote from Hal Varian (chief economist at Google) sums up the importance of learning data analysis skills

> *"The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that's going to be a hugely important skill in the next decades… (Varian, 2009).*

Overview

Given the abundance and variety of publicly available data sets, This project will give you considerable freedom in the selection of your research question. In any data science project, one needs to accomplish the following five steps: 1) to formulate a research question, 2) find data likely to provide insights into the question of interest, 3) clean the data, 4) perform exploratory analyses of the data, 5) interpret the results obtained and write up a report on their findings.

In Lab #10 you learned some basic Google sheets data analysis skills. In Lab #11 you will have some time to explore a variety of different datasets to use for your project. In Labs #12 and #13 you will learn some additional data analysis techniques. Your final report for this project will need to follow a template that explains what things to include in your report - link provided in the last section of this guide. Essentially you will be reporting on the process you used to complete the process, including any insights you gained from your work with the dataset(s) you choose to work with.

As for the timing of this project we expect the different stages of the project to take various amounts of time as explained in the next section. Please note that those section noted as "in Lab" may require some additional time outside of lab.

Experimental Design / Proposed Schedule

1. Identify a research problem (In lab #11): Explore areas of interest and to determine what background reading should be done in preparation for starting the research project

2. Identify available data sources (In lab #11): The student will begin to identify possible data sources and describe how the proposed data may assist in answering the research question

3. Formulate hypotheses to test (1-2 hours): Based on an initial overview of the data, the

student will identify one or more hypotheses to test using the selected dataset(s). This requires you to understand what kind of data is available in your chosen dataset(s), how was the data gathered, the format and layout of the data, and what limitations, if any, there are concerning use of the data.

4. Examine the data sources and clean data as needed (In lab #12): The student can use a variety of tools (text editors, Excel, Google Sheets, on-line tools) to perform exploratory data analysis and data cleaning to determine the adequacy of the data for the research question being investigated and to get the data in a usable format. (this step will also require additional time outside of lab)

5. Perform analyses of the data (4-6 hours): The student will use analysis tools presented in Labs #10 - #12, and/or any other tools they are comfortable with, to perform analyses of the data.

6. Interpret and write up results for report (3-4 hours): Since data visualization is a critical subfield of data science students will use the graphics capabilities of their choice of data visualization tools to create graphics to highlight their research results.

Total Expected Time outside of lab: 8-12. Note this timetable is an estimate of the time it will take to accomplish each step of this process. The actual amount of time that you may require to complete each task may be different.

Report Specifications

Your report should document the process you went through to complete your data analysis together with background on the topic you are working on and finally what results you discovered in the data. The report template is available as a Google doc at:

https://docs.google.com/document/d/1HSx-mxkmN4pmvmGGeD0TNnU0IQTEfFMeKjI82Uu1jT4/edit?usp=sharing

This Google doc is provided for you to work from to help you organize the different parts of your report.

Do not wait to start working on this project: doing so will certainly affect your grade. If you work on each of the listed activities in the order presented above you should have no difficulty in completing this project by the deadline.