CSCI 385 - Introduction to Data Science
Project #1

August 17, 2021

Description

This project will concentrate on the **Discovery** and **Data Preparation** phases of the Data Analytics Life Cycle. In an R Markdown file, document your process and provide R code for importing and cleaning your data and for performing some initial investigation and demonstrating familiarity with the data.

- In an introduction, briefly describe your motivation for choosing the domain of your data set. What interests you about this area?
- Identify the data set that you will be using. The data set should have at least 1000 observations and should have enough variables to explore nontrivial questions ($\sim$ 10 with a combination of continuous and categorical variables).
- Describe your source. Use critical analysis to assess its quality and possible limitations.
- Document your variables and describe in plain language what they mean and how they are represented.
- Document any manipulations you make to clean your data. Make it as "Tidy" as possible.
- Summarize your data using descriptive statistics along with visualizations and accompanying interpretations that help explore what possible insights might be investigated.
- Clearly state **three** data science questions that could be addressed using this data. Discuss how they could be addressed, any information not in the data set that might be useful in addressing them, and any potential social or ethical implications of working with this data.

We will spend time in class on **October 12th** doing peer reviews for this project. This peer review has three main goals:

1. Practice communicating progress to a **technical** audience.
2. Practice giving constructive feedback regarding an initially unfamiliar project.
3. Practice responding to and incorporating feedback.

You do not need to have completed your R Markdown file before this day. However, you will be expected to give a short, informal presentation on your progress. In particular, you will give a 5-10 minute presentation to a small group of your peers on the work that you have completed so far. This presentation should focus on important aspects of the Discovery and Data Preparation phases of the Data Analytics Life Cycle.

There will be time for the audience to ask questions and to provide feedback, both verbal and written. The written feedback will be given to the presenter and **will be submitted by both the reviewer and the presenter** (see below). For each review, make sure it is clear who the feedback is for and who provided the feedback. You may use the following questions to guide the feedback you give:

- What parts of the presentation were most effective?
- Did you find any aspects of the presentation particularly confusing or unclear? If so, what might help to clarify those parts?
- Are the variables in the data set clear?
- Are the findings (descriptive statistics, visualizations, and interpretations) compelling? Were there any potentially unjustified conclusions?
- Are data science questions clear and forward looking? Could they form the basis of more formal falsifiable hypotheses?
- Can you think of any other interesting questions or potentially useful sources of data?
- Can you think of any potential social or ethical implications of the project?

As part of this deliverable, you must respond to the feedback you receive. This response describes, very succinctly, how each suggestion affected or changed the project. For example, if you receive a comment like "It's not clear to me what the variable X in this data set represents", an appropriate response might be "Clarified the meaning of X with an example". It is important to acknowledge suggestions even when they do not result in changes to the project. For instance, you might respond to "Have you considered the question X in the context of this data?" with something like "I had not considered X. That is a very interesting question but, since it might be difficult to address because of Y and Z, I will leave it for a future project".

Always be kind and considerate with the feedback you give and remember that comments you receive are meant to help improve your project.

Submission

You must submit three files for this project, two on GitHub and one on Blackboard. Commit your R Markdown file and a knitted PDF to your GitHub repository. Peer review comments of your work, your responses, and your comments regarding the work of your peers should be submitted on Blackboard as a single PDF. These submissions are due on **October 20th by 11:59pm**.

Grading

Your work will be evaluated on an 100 point scale, based on the following rubric:

- 20% Discovery - the project shows a strong grasp of the activities and goals of the Discovery phase with explanations that can be understood by the general public.
- 20% Data Preparation - the data has been cleaned appropriately and organized according to "Tidy Data" principles; the process has been explained in a clear and accurate manner.
- 20% Information Visualization - visualizations have been designed and interpreted well to communicate an accurate depiction of the data summary.
- 10% R Proficiency - the code blocks demonstrate a strong grasp of R for Data Science along with following best practices to make the code easy to understand and reproducible.
- 10% Peer Review - comments given to peers are thoughtful and constructive. Responses to peer comments are considered and implemented when appropriate.
- 10% Critical Thinking - the documentation and descriptions demonstrate critical thinking about the subject matter, including considerations for ethical and social implications as well as identifying opportunities for data-driven insights.
- 10% Communication - the combination of writing and visualizations help communicate a cohesive narrative that the general public could follow and find interesting.