

**GENERAL INFORMATION****GOAL**

The goal of this project is to perform an exploratory data analysis / create visualizations of a data set of your choosing, in order to gain preliminary insights on questions of interest to you.

**TEAMS**

You may work alone or in teams of up to 4 people. Grading will be by team; more is expected of larger teams. Information will be provided on CourseWorks on how to sign up as a team.

**DATA**

Choose a data set that is not on the beaten track, that is, one that is not included in R (or similar), nor used in Kaggle (or similar) competitions, nor relatively well-known through some other forum. You will begin working with the dataset in Homework #4, to be assigned soon, due Tues, March 28.

**ANALYSIS**

You have a lot of freedom to choose what to do, as long as you restrict yourselves to *exploratory* techniques (rather than modeling / prediction approaches). In addition, your analysis must be clearly *documented* and *reproducible* (more on that below).

**FEEDBACK**

At any point, you may ask the TAs--Ian (iak2119) and Bridget (blr2147)--or me (jtr13) for advice. Our primary role in this regard will be to provide general guidance on your choice of dataset / topic / direction. As always, you are encouraged to post specific questions to Piazza, particularly coding questions and issues. You may also volunteer to discuss your project with the class in order to get feedback--if you'd like to do this, email me to schedule a date.

**PEER REVIEW**

A portion of your grade is based on the feedback you give to other groups. After the due date, you will be assigned projects to review, which you need complete by **Thursday, April 20, 11:59pm**. More specific details on what you need to do will be provided at that time. Your grade is *not* directly based on the feedback you receive. It will be determined by the instructor and TAs.

**REPORT FORMAT**

Your project should be submitted to CourseWorks as a **nb.html** or **.ipynb** file, with graphs / output rendered. Any material that cannot be included in the notebook format, such as certain interactive visualizations, should be clearly referenced, ideally by providing a link in your notebook to an online visualization. You will lose points if we have trouble reading your file, need to ask you to resubmit with graphs visible, if links are broken, or if we have other difficulties accessing your materials due to factors that are in your control.

**REPORT OUTLINE**

Your report should include the following sections, with subtitles ("Introduction", etc.) as indicated:

**1. Introduction**

In this section, explain why you chose this topic, and the questions you are interested in studying. Include a brief description of how you found the data, and clear instructions on where the reader can find the data.

**2. Team**

List team members and a description of how each contributed to the project.

**3. Analysis of Data Quality**

Provide a detailed, well-organized description of data quality, including textual description, graphs, and code.

**4. Executive Summary**

Provide a short **nontechnical** summary of the most revealing findings of your analysis with no more than 3 static graphs or one interactive graph (or link), written for a nontechnical audience. The length should be approximately 2 pages (if we were using pages...) Do not show code, and take extra care to clean up your graphs, ensuring that best practices for presentation are followed.

**5. Main Analysis**

Provide a detailed, well-organized description of your findings, including textual description, graphs, and code. Your focus should be on both the results and the process. Include, as reasonable and relevant, approaches that didn't work, challenges, the data cleaning process, etc.

**6. Conclusion**

Discuss limitations and future directions, lessons learned.

*A note on style:*

You are encouraged to be as intellectually honest as possible. That means pointing out flaws in your work, detailing obstacles, disagreements, decision points, etc. -- the kinds of "behind-the-scene" things that are important but often left out of reports. You may use the first person ("I"/"We") or specific team members' names, as relevant.

**Grading Rubric**

Section	Points
Introduction (including choice of data set, questions), Team description	10
Analysis of Data Quality	10
Executive Summary (focus on quality of presentation choices / techniques)	20
Main Analysis (focus on quality of EDA choices / techniques)	35
Conclusion	5
<b>General</b>	
Reproducibility (will be discussed more in class), sources cited	10
Technical flawlessness (files open, links work, code runs, etc.)	10
<b>TOTAL*</b>	<b>100</b>

**\* If late, 10 points will be deducted per day. Plagiarism of any kind will not be tolerated and will result in a grade of 0 for the project.**