Final Project Assignment

**STAT 5702 FINAL PROJECT due Tuesday, Apr 24, 2018, 11:59pm**

**GENERAL INFORMATION**

General note: Any clarifications / additional advice made after the final project was first assigned will be in blue.  Please note however that this assignment sheet cannot possibly cover all the "do's and don'ts" of data analysis and visualization. You are expected to follow all of the best practices discussed in class throughout the semester.

SIGNUP: If you have not done so already, please create a final project group (click on the People tab, and then Student Groups) and add the members of your group to it. Please use the following naming convention: Columbia UNIs in *alphabetical order*, with underscores separating names.  For example: **jtr13\_rar3010\_sw2934**(If you already created a group you do not need to redo it.)

(We will recreate these groups as Final Project Groups since student created groups cannot be used for graded assignments in CourseWorks.)

**GOAL**

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

**TEAMS**

You must work in teams of 2-4 people. Please [fill out this form (Links to an external site.)Links to an external site.](https://goo.gl/forms/PciVMLUmQK8rRAUn2) if you would like to be assigned a partner or partners.  (If you have specific interests you should try to find partners on Piazza first as I will not be able to match on specific criteria -- I will simply assign groups in the order in which responses come in.)

**TOPICS**

Start with a topic / question that interests you and *then* look for data!

**DATA**

Choose data from the original source, 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. If in doubt, *ask!*

A few examples are:

[https://opendata.cityofnewyork.us/ (Links to an external site.)Links to an external site.](https://opendata.cityofnewyork.us/)                    [https://www.bls.gov/ (Links to an external site.)Links to an external site.](https://www.bls.gov/)

**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.*

**FEEDBACK**

At any point, you may ask the head TA-- Jing Wu (jw3233) -- or me (jtr13) for advice. Our primary role in this regard will be to provide general guidance on your choice of data / 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, each *individual* will be assigned two project *groups* to review, and instructions will be provided.  The reviews need to be completed by **Tues, May 1, 11:59pm**.

Note: part of the grade you receive for the class is based on the quality of review that you *write*, not on the feedback that your project *receives.*Your grade for the project (as for all other assignments for the class) will be determined solely by the instructor and TAs.

**REPORT FORMAT**

With the exception of the interactive part, your project should be submitted to CourseWorks in the same manner as homework assignments, as **.Rmd** and **.html** files, with graphs / output rendered. 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. Note: Using Markdown + code chunks is supposed to make combining code, text and graphs easier. If it is making it more difficult, you are probably trying to do something that isn't well suited to the tool set. Focus on the text and graphs, not the formatting. If you're not sure if something is important to focus on or not, please ask.

Advice: don't wait to start writing.  Your overall project will undoubtedly be better if you give up trying to get that last graph perfect or the last bit of analysis done and *get to the writing*!

**OUTLINE**

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

1. **Introduction**

Explain why you chose this topic, and the questions you are interested in studying.

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

1. **Description of Data**

Describe how the data was collected, how you accessed it, and any other noteworthy features.

3. **Analysis of Data Quality**

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

1. **Main Analysis (Exploratory Data 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.

1. **Executive Summary (Presentation-style)**

**Note:** "Presentation" here refers to the *style of graph*, that is, graphs that are cleaned up for presentation, as opposed to the rough ones we often use for exploratory data analysis. *You do not have to present your work to the class!* However, you may choose to present your work as your community contribution, in which case you need to email me to set a date before the community contribution due date (Apr 3). (The presentation itself may be later.)

Provide a short **nontechnical**summary of the most revealing findings of your analysis  written for a nontechnical audience. The length should be approximately two pages (if we were using pages...) Take extra care to clean up your graphs, ensuring that best practices for presentation are followed.

1. **Interactive Component** (details to be provided)

1. **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** |
| 1. Introduction (including choice of data, questions), Team description | 10 |
| 2. Description of Data | 10 |
| 3. Graphical Analysis of Data Quality | 10 |
| 4. Main Analysis (focus on quality of EDA choices / techniques) | 20 |
| 5. Executive Summary (focus on quality of presentation choices / techniques) | 20 |
| 6. Interactive Component (specific instructions to be provided) | 20 |
| 7. Conclusion | 10 |
| **TOTAL** | 100 |

\* **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.**

\* Points will be deducted for technical flaws (problems opening files, following links, etc.), for not citing sources, and for lack of reproducibility.