**PROJECT MILESTONE 2: INITIAL EXPLORATION**

**Instructions**

1. Have one member of your team submit a write-up on Canvas.

2. Write-ups must be no longer than 750 words (not including references or figure captions).

3. Submit your code, either as a supplementary file in the Canvas submission, or as a public URL (github / box / google drive).

**Description**

After identifying a set of interesting questions and relevant datasets, the next step in a careful data analysis is to conduct preliminary, high-level exploration. The goals of this exploration may include,

* **Checking for potential data quality issues**. Is there any missingness? Are outliers present? Are all the column names and variable levels interpretable? Are some factor levels extremely rare? If these are present, what should be done about them?
* **Build intuition about individual variables**. For example, it can be helpful to save a histogram for every continuous variable and a bar plot for every categorical one. Even though these figures may never appear in any final reports, it can be helpful to skim through these to build intuition (I’ve known people who quickly read through hundreds of histograms when they are acquainting themselves with a dataset).
* **Prepare processed data**. If you find the data un-tidy for the purpose of your study, save a tidied version in a derived\_data folder. If you anticipate needing to derive variables or join tables, computing them now will save effort during downstream analysis. This initial step can help compartmentalize your thinking about data preparation and visual design — a structured, curated dataset will let you focus on visualization without having to continually go back and modify your input.
* **Evaluate initial hypotheses.** Based on the questions you have identified in Milestone 1, there may some natural first figures or interactive views to implement. These can become the starting point for more refined versions that can be shared more widely.
* **Sketch interactive visualizations**. If you are planning to submit a more complex interactive visualization in later milestones, then it will help to sketch the reactive graphs for several competing designs. Which visual comparisons are most easily supported by which designs?

We expect that, depending on the nature of the problem or data studied, 2 – 3 of these goals will be a higher priority than the rest. Your Milestone 2 report should document the steps your team has taken towards reaching this subset of goals.

In the report’s text, it is not necessary to explain the details of every step – think of this instead as a research journal or an executive report. These documents summarize the judgements that guided the tasks that were completed. They also describe the state of the study, highlighting any surprises that were encountered and providing a launching point for further development. Finally, the report lets readers know where they can find more details.

We expect you to submit your code, and it should be reasonably documented and readable. The reason for requiring this is that, the better organized your initial efforts are, the easier they will be to refine and extend.

A rough outline of the tasks involved in successful completion of this milestone is,

1. [30 minute meeting] Decide as a team which of the above goals are highest priority.
2. [2 x 30 minute sessions] Brainstorm individually or in groups some specific tasks that are needed to help reach these goals.
3. [4 x 60 minute sessions] Pursue the tasks brainstormed in the previous sessions.
4. [45 minute meeting] Sync with team about progress and outline the report.
5. [2 x 45 minute sessions] Prepare and revise your section(s) of the milestone report.

**Rubric**

* Depth of exploration (7 points): The team has made substantive progress towards understanding their problem and data. Artifacts generated by this exploration (processed datasets, initial figures, design sketches) facilitate success for the overall project.
* Distillation of takeaways (7 points): The report effectively summarizes the state of the project, including challenges met, surprising findings, and potential next steps.
* Clarity of exposition (6 points): The writing is compact and not verbose, free of technical   
  errors, and divided into clearly marked sections. All supporting figures or tables are properly annotated and referenced.
* Code quality (5 points): Code is organized, readable, and documented. Even if not yet fully reproducible, there is sufficient documentation that there is no doubt about being able to rerun any parts, if the need arose.