DATA 115: Personal Dataset Curation

This document provides details and guidelines for the personal dataset project, which accounts for 5% of your final grade. The purpose of this project is both to have an example of completed work for internships and other professional applications as well as to allow you to have access to a meaningful starting point to apply techniques that will be encountered later in the major. Hopefully, you will view this set of assignments as an opportunity to take a deeper dive into a topic that is meaningful to you.

1 Intermediate Deadlines

Throughout the semester you will be required to submit preliminary project reports with more details provided as a part of the weekly homework assignments. Each of these components will become a part of your final product and they will assigned according to this schedule:

Week 5: Make a repository

Week 7: Example questions to answer and potential data source

Week 9: Access, process, and upload

Week 11: Summary visualization and analytical technique

Week 13/14: Final repositories

Week 15: Three minute presentations

2 Final Product

Your final submission will consist of a polished GitHub[1] page, documenting and presenting your dataset. While the data is the central feature of the project, your descriptions of the motivation, cleaning process, and analysis will also be critically important. You will also be graded on the following items, the first four of which should be presented in the readme portion of the associated repository:

- **Motivating Question:** What question were you setting out to answer at the start of this project?
- **Data Process:** The centerpiece of your writeup should be a comprehensive discussion of your process for sourcing, collecting, cleaning, and transforming the raw data into your final dataset. The instructions should be clear enough that someone else could reproduce your process based on just the readme file to obtain the same final results.
- **Visualization:** Choose and construct a single visualization to summarize and represent your data.
- **Analytical Technique:** Apply one of the techniques that we have covered in the course to your dataset and provide an analysis of the results. You can choose from either among the exploratory data analysis topics or from the techniques we cover later in the course[2].
- **Presentation:** In the final week of class, you will give a three minute presentation[3] about your dataset.

[1] Other hosting platforms may be allowed with instructor permission.
[2] You may choose to do some analysis that goes beyond the material presented in the course with permission of the instructor.
[3] Due to technical complications, it is possible this will need to be pre-recorded.