STAT 184 Final Project

Soumya Mukherjee

Due: August 11, 2023

Final Project Description

Polished & professional written investigation of your final project topic. The result should be something you would be proud to include in a work portfolio or discuss in an interview or cover letter for an internship, research opportunity, job, etc. Your project GitHub Repo should be "self-contained" meaning that it includes access to the source data such that another person (e.g., STAT 184 grader, future supervisor) can clone your provided GitHub Repo and execute your entire analysis without errors.

The weekly activities in the Data Computing text book are good examples of the type of work expected for a successful project, with the differences that you are expected to do the work independently (or in your group) using your own data (not loaded from an R package), and you are responsible for the narrative explaining your reasoning and conclusions as you work through the analysis.

Students need to work independently on their projects (this is not a group assignment).

Scoring

The project is worth 25% of the final course grade and will be evaluated according to the following:

- **Reproducible Research**: The final product is published to a self-contained GitHub repository and presented as a and RNotebook. Your actual analysis should be an **.nb.html** file with embedded .Rmd (like usual) that can be run without errors by the instructor or graders without prior exposure to the project or modification of your code.
- **Data Access**: The project uses 2 or more real data sets - a primary dataset and a secondary dataset. The primary data should not be loaded from an R package, but it may be joined to secondary data available in an R package. For example, the primary data could be joined to a data set like `CountryCentroids` from the `dcData` package if `iso_a3` country labels or latitude/longitude information are needed. Data access should be reproduced in the .Rmd file (e.g. read in from URL, scrape from web, or in a CSV in the GitHub repo). If you use a live data source (e.g. from the web), it's a good idea to save a static copy of the data in your GitHub repo as a backup just in case the source data fails or changes the day your project is due (!)
- **Data Wrangling**: The project demonstrates proficiency with data wrangling techniques learned in STAT 184 (e.g., `dplyr`, `tidyr`)
- **Visualization**: The project demonstrates proficiency with graphics tools learned in STAT 184 (e.g. `ggplot()`, choropleths, leaflets) to explore several variables of interest. At least one graphic should show a useful visualization involving 3 or more variables through faceting, coloring, linetype, etc.

- **Code Quality**: Code conforms to syntax and style conventions. The easiest one to pick is the on in Data Computing text book (e.g. chain syntax, readability, variable and table naming, commented code)
- **Narrative Quality**: The project is a complete report that describes the background and context of the data set as well as detailed rationale of each decision and explanation of each observation in the analysis.
- **Overall Quality**: Judgment of holistic quality of project. Reports should follow a logical progression and maintain a polished, professional appearance.

Project Milestones

- August 2, 2023 @ 9:59am: Submit project ideas (on canvas only)
- August 11, 2023 @ 11:59pm: Final Project Report Due (Submit the nb.html file along with link to the github repo)

More detailed requirements will be announced on Canvas and in class over the next few lectures. The changes will be updated in this document as well.

Each student must choose at least 2 different data sets. Data set approval from the instructor is strongly recommended in order to be confident that you have chosen a data set likely to align with the goals of the Project. Students may request data set approval as many times as necessary until they have an appropriate data set for the project. If you wish to get feedback from the instructor regarding the final project report, all materials must be submitted to the Github repo within August 4, Monday, 9:59am and let the instructor know.

Getting Started

For some it will seem daunting to start from scratch looking for one or more "interesting" data sets. There are lots of useful repositories out there. Here are a few links to get you started, but please feel free to use any data that interest you!

- https://www.springboard.com/blog/free-public-data-sets-data-science-project/
- https://www.dataquest.io/blog/free-datasets-for-projects/
- https://data.cityofnewyork.us/
- http://www.icpsr.umich.edu/icpsrweb/ICPSR/
- https://github.com/awesomedata/awesome-public-datasets
- https://github.com/fivethirtyeight/data