STAT 184 Final Project

Due: 12/9/2016 at Noon

Final Project Description

The final project provides an opportunity to combine content learned throughout the course for use in some realistic application. Each student is required to individually complete and sumbit their own work, but students consult one another with questions (post questions on Piazza, chat with group members, etc). A successful project will find one or more real & interesting data sets, and use their R programming skills to tell a story that reveals insights from the data. The weekly group 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 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.

Scoring

The project is worth 30% of the final course grade and will be evaluated according to the following:

- **Reproducible Research**: The final product is an HTML file with embedded .Rmd that can be run without errors by the instructor or TA without prior exposure to the project or modification of your code.
- **Data Access**: The project uses one or more real data sets. The primary data should not be loaded from an R package, but it may be joined to data available in an R package. For example, the primary data could be joined to a data set like `CountryCentroids` from the `DataComputing` 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 or scrape), or uploaded to Canvas as a CSV so the instructor & TA can load the data when we run your code.
- **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 shows a useful visualization involving 3 or more variables through faceting, coloring, linetype, etc.
- **Code Quality**: Code conforms to syntax and style conventions advocated by the 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

- (recommended) Data set approval
- (optional) Interim feedback: Analysis Plan
- (optional) Interim feedback: Complete Draft Peer Review
- Final Project due at NOON: 12/9/2016

Each student must choose a different primary data set. Data set approval 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. One round of interim feedback on the analysis plan and/or one round of interim feedback on a complete draft of the project is available to each student. Students may also submit a complete draft of their project for Peer Review. All students that submit a draft project for Peer Review are required to participate by reviewing project drafts submitted by their peers, but no student will be asked to participate in the peer review process if they have opted not to submit a project draft of their own. Interim feedback requests must be made & associated materials submitted at least two weeks before the project due date.

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/

http://www.datasciencecentral.com/profiles/blogs/great-github-list-of-public-data-sets

https://github.com/fivethirtyeight/data

DataCamp Alternative

Each student may choose to complete a selection of DataCamp tutorials as an alternative to the Final Project for partial credit. The eligible DataCamp tutorials must be selected from the list "assigned" to STAT 184 on the DataCamp website (www.datacamp.com). Scoring is based on the number of complete DataCamp tutorials as follows:

- 1 course: 10% credit
- 2 courses: 20% credit
- 3 courses: 30% credit
- 4 courses: 50% credit
- 5 courses: 70% credit
- 6 courses: 80% credit
- 7 courses: 85% credit

Incomplete courses or those completed after the due date for the final project will NOT be counted for credit, and the "Introduction to R" course is a homework assignment for the class so it will NOT be counted toward the total of 7 courses. Also, note that the maximum allowable credit earned for the DataCamp Alternative is 85%.