STAT 209: Project 1 Description

NOTE: Please have *exactly one* person in your group create project repo through the GitHub Classroom link here (https://classroom.github.com/g/NESDtsu9), and when asked to create a team, please name your team simply `group1`, `group2`, etc. Then, the other team members should use the same link and join the existing group. This will do a few things: (1) It gives me easy access to your repos without you explicitly having to invite me, as well as the ability to check out your repos via script without much if any hand-tuning,and (2) It makes your repository private, since it is part of my `ocstats` organization which has some extra features.

Goals

The goal of this project is to create a "blog-post" style writeup in the style of FiveThirtyEight or the NYTimes "The Upshot" column, centered on a related set of informative, accurate, and aesthetically pleasing data graphics illustrating something about a topic of your choice.

Required skills

Proficiency with `RMarkdown`, `git` and the `ggplot2` package

Relevant background and resources

- Chs. 2 and 3 of the textbook
- Leada tutorial on R
- RStudio tutorial on RMarkdown
- Slides and labs from 9/3 - 9/24
- `ggplot2` Quick Reference Sheet (http://colindawson.net/stat209/resources/ggplot2-quick-reference.pdf)
- RMarkdown Quick Reference (http://colindawson.net/stat209/resources/rmarkdown-quick-reference.pdf)

Instructions

Your group will work together to write a blog post that contains at least three related data graphics that tell the reader something interesting about the domain that the data comes from. The following are some examples of the kind of structure I have in mind (though most of these are longer than your post will be).

- Typical ages of people with common female names (https://fivethirtyeight.com/features/how-to-tell-someones-age-when-all-you-know-is-her-name/)
- Quality of various flights and airports (https://fivethirtyeight.com/features/fastest-airlines-fastest-airports/)
- Learning about New Yorkers' lives via taxi and Uber records (http://toddwschneider.com/posts/analyzing-1-1-billion-nyc-taxi-and-uber-trips-with-a-vengeance/)
- People admitted to the ER for wall-punching (https://qz.com/582720/americas-most-prolific-wall-punchers-charted/)

Conciseness is of value here: aim for a post in the 500-700 word range (not including the Methodology section at the end). A suggested structure is as follows (though, apart from the inclusion of the Methodology section at the end, you do not have to adhere to this exactly):

- An introductory paragraph that sets up the context and introduces the dataset and its source. Tell the reader what the units of observation ("cases") are, and what the relevant variables are. However, don't just list these things: work them into one or more paragraphs that inform the reader about your data as though you were writing an article for a blog.
- For each graphic you include, a paragraph or two discussing what the graphic shows, including a concise "takehome" message in one or two sentences. Again, don't just show graphic, paragraph, graphic paragraph, …, connect your text and graphics in a coherent narrative.
- A discussion section that ties together the insights from the various views of the data you have created, and suggests open questions that were not possible to answer in the scope of this project (either because the relevant data was not available, or because of a technical hurdle that we have not yet learned enough to overcome)
- A "methodology" section (in an Appendix section at the end, separate from the main narrative) that explains the technical details of your project for a reader interested in data visualization. Explain the choices you made in your graphic: why did you choose the types of graphs (geometries) that you did; why did you choose the aesthetic mappings you did, why did you choose the color schemes you did, etc.

Turning in your project

Collaboration on your project should take place via GitHub commits. Your `.Rmd` source should be part of a GitHub repo from its inception, and changes should be recorded via commits from the account of the person who made the edit. Everyone in the group must make at least one commit.

Your final submission will consist of the `.Rmd` source, compiled `.html`, and any other files needed for the `.Rmd` to compile successfully. For example, if you are reading in the data from a `.csv` file stored locally (that is, in your RStudioPro server account), commit this file, and make sure that you are using a relative path to the file when you read in the data. If you are reading the dataset directly from an R package or from a URL, this is not necessary.

Whatever state those files are in at the deadline is what I will grade.

Data

You can use any data source you want. For this first project, you are not expected to do much data wrangling, so you should spend minimal (if any) time manipulating the dataset. You might want to do some `filter()`ing to select subsets, and/or `mutate()`ing to create new variables, but that's the extent of the wrangling you should do (and only do that if it's appropriate for what you want to show!). If you are finding that creating the graphics you would need to create requires more involved wrangling, you might want to redefine your topic (but keep the original one in mind as a candidate for Project 2!)

Some possible sources for data are:

- The federal government's Data.gov (http://data.gov) site
- The American Psychological Association (http://apa.org/research/responsible/data-links.aspx)
- The data science competition Kaggle (https://www.kaggle.com/datasets)
- The UC Irvine machine learning repository (http://archive.ics.uci.edu/ml/index.php)
- The Economics Network (http://www.economicsnetwork.ac.uk/data_sets)
- Data provided by an R package, such as
    - `nycflights13` : data about flights leaving from the three major NYC airports in 2013
    - `Lahman` : comprehensive historical archive of major league baseball data
    - `fueleconomy` : fuel economy data from the EPA, 1985–2015
    - `fivethirtyeight` : provides access to data sets that drive many articles on FiveThirtyEight

You can find data anywhere else you like. But don't use a dataset we've used in class or homework, and if you are using a dataset from an R package, ensure that you're doing something different from what might be in the examples given in the documentation on the dataset.

To see a list of the datasets provided by a given R package, you can type the following at the console (fill in the package name).

```
data(package = "somePackageName")
```

Grading Rubric

This project is worth a total of 10% of the course grade. There is a group component and an individual component to the grade, each weighted equally (5% each). The typical division of labor is that each group member is individually responsible for at least one graphic, along with the part of the writeup and methodology section directly pertaining to that graph, and the group as a whole works together to write and edit the general introduction and conclusion, along with any components of the Methods section that pertain to the project as a whole. Your group may choose to divide the work differently, but be sure that each person is involved in the topic selection and planning stage, the coding component, the "general audience" writing element, and the "technical writing" element.

Group grade: Baseline Criteria (3/5 credit)

- a suitable dataset is chosen, and the visualizations chosen fit together to illustrate interesting features about the data
- at least two related graphics are included
- the graphics are generated by the code embedded in the `.Rmd` (not included from an external file)
- the `.Rmd` compiles successfully

- the motivation and goals of the blog post are laid out clearly at the beginning of the writeup
- a description of the dataset is provided, along with contextual information
- the post is a reasonable length (500-700 words of text)
- a description of the technical elements of your project is included in a separate methodology section at the end
- there is a description of “big picture” insights gained from considering the visualizations as a set
- there is a GitHub record of commits, with informative commit messages

Group grade: Finishing touches (5/5 credit)

- The above, plus:
- Code, unnecessary messages, and raw R output (other than the plots) are suppressed from the `.html` output
- The choices made are effective and allow information to be conveyed clearly and efficiently
- The writeup is organized well, making it easy to follow your narrative, and looks professional and polished

Group grade: Above and beyond (6/5 credit)

- The writeup could be mistaken for an article in a prominent data journalism outlet
- The graphics used collectively convey particularly surprising or subtle aspects of the data that would have been difficult to notice with any single view

Individual grade: Baseline Criteria (3/5 credit)

- Your group members report that you were a meaningful participant in planning the project, and made meaningful contributions to the “collective” parts of the project.
- You are present for all in-class workshop and presentation days
- You are present for all out-of-class group planning sessions, or if you have to miss one, you make up for it later
- Your graphic includes relevant context (title, axis labels, etc.)
- The graphic you were responsible for is constructed using clean `ggplot2` code
- The interpretation of your individual graphic fits with what is shown in the graphic, and the graphic illustrates what you say it does
- Your contributions to the writing consist of complete, grammatical English sentences
- The visual (aesthetic) mapping is motivated in your piece of the Methods section
- Your individual contributions are clearly documented with commits from your account in GitHub

Individual grade: Full Contributor (5/5 credit)

- Your group members report that you were at least a co-equal participant in all phases of planning and execution of the project
- Your graphic displays aspects of the data that are not easily seen in the other graphics in the writeup
- Your graphic employs thoughtful choices for geometries, aesthetic mappings, color palettes
- Your interpretation exhibits insight into non-trivial aspects of the data
- Your writing is clean, concise, and easy to follow.

Individual grade: Above and beyond (6/5 credit)

- Your individual graphic displays especially original, subtle or surprising insights into the data