STAT 209: Project 3 Description

Deadline

- The final writeup is due by Fri. 12/20 at 9pm, which is the end of the designated “final exam” time for this class.

Goals

This project builds on the previous one in scope. Unlike the first two projects, this one is meant to be done individually. I may make exceptions for unusually ambitious proposals; talk to me if you have an idea that you think might lend itself better to working with a partner.

There are four basic options:

- Option 1: Do a new analysis using a dataset that interests you, along the same lines as project 2, but using a large dataset, accessed from a database via SQL.
- Option 2: Do a new analysis involving one or more of the advanced or specialized techniques that haven’t been part of a project yet. These include: interactive graphics, clustering, dimensionality reduction, geographic data, text data.
- Option 3: Some mix of 1 and 2
- Option 4: “Improve and extend” one of your first two projects. The “improve” part means revising elements of the original project that you weren’t entirely satisfied with the first time. The “extend” part means taking advantage of tools that you learned after you did the original project to look at the data from a different perspective, complement the analysis with additional data, apply some of the machine learning algorithms you will have learned by that point, etc. The degree of extension expected is inversely related to the scope of the original project: a revision of project 1

should be more different from the original than a revision of project 2. If more than one person from the same group decides to extend a project that they worked on together, they should take it in distinct directions.

Note: since revising an existing project means starting with a base of work that was already evaluated (and which was not entirely your own), the expectations for quality are higher in this case than for a new project starting from scratch.

The final writeup

The writeup should be structured similarly to previous projects (see the project 2 description for guidelines)

Data Visualization Requirements

As always, prioritize quality over quantity. A few really well thought out graphs that take a fair amount of work is much better than several hasty ones.

As always, the graphs must be created within your `.Rmd` using `ggplot2`, and you should endeavor to do whatever wrangling is necessary in a separate pipeline so that the visualization code is as well structured and concise as possible.

Wrangling Requirements

As with Project 2, you should employ a number of the five basic verbs (or their SQL equivalents), along with joins, gather, spread, and custom functions/iteration as needed to get your data into a form conducive to the visualizations you want to construct.

You will likely not use every one of these elements, but your wrangling should involve non-trivial manipulation of datasets.

Whether you doing your wrangling entirely within `R` or through a mix of `R` and `SQL`, you should put your code in RMarkdown code chunks, and keep your wrangling pipeline(s) separate from your visualization pipeline(s), for readability's sake.

Extra Elements

You should employ one or more of the following elements in your project:

- SQL
- Substantial text manipulation
- Clustering
- Dimensionality Reduction
- Geographic data and spatial layering/projection
- Interactive Graphics

Data

Any data source that was fair game for Project 2 is fair game here: you can't use a dataset we've used in class or a lab, nor data built into an R package (at least not without running your idea by me).

The GitHub workflow

- See the Project 2 description for an outline of the recommended GitHub workflow.
- The GitHub clasroom link to create your project repo is here (https://classroom.github.com/a/vIzjtPzr)

Turning in your project

Even though you are working by yourself, you should still record the history of edits to your project via GitHub commits.

As always your final submission will consist of the `.Rmd` source, compiled `.html` (or `.pdf` if you prefer), and any other files needed for the `.Rmd` to compile successfully. Whatever state those files are in at the deadline is what I will grade.

Grading

This project makes up 25% of the final grade for the course. In addition, your overall project grade (for projects 1-3, totalling 45% of the course grade) cannot be below your grade for this project.

Rubric

Since the specs for this project are more open, the grading rubric is not as concrete this time, but here is a rough rubric:

Technical baseline (60% of total)

Markdown and Version Control (10%):

- The `.Rmd` compiles successfully with no error messages
- Code, unnecessary messages, and raw R output (other than the plots) are suppressed from the `.html` output
- The GitHub workflow is followed, and *informative* commit messages (that describe what changes were made in that commit) are included.

Wrangling (20%):

- Solid command of core wrangling techniques (five verbs, joins, gather, spread, or their SQL equivalents) is demonstrated
- The visualization pipeline is kept as "clean" as possible by wrangling your data into a form that is conducive to simple visualization commands
- The technical content of the project lines up with one of the options given

Visualization (20%):

- At least one visualization is included, as generated by embedded code
- The choices made are effective and allow information to be conveyed clearly and efficiently

Coding Style (10%):

- Code is well documented with inline comments
- Code follows a consistent style, making use of line breaks and indentation for readability
- Variable names are informative, and code is overall easy to follow

Writeup baseline (25%)

Basic Structure (15%):

- The final post is the right length (800-1000ish words)
- The writing consists of fluent, grammatical sentences of English.
- An engaging introduction is included, which sets up the context of the analysis and describes the data and its source
- Each graphic is interpreted clearly and concisely in the text, including a “take-home” message in no more than two sentences
- A compelling discussion ties together the threads in the writeup
- Data-wrangling methodology is included in the appendix, and is clear
- Data-visualization methodology is included in the appendix, and is clear

Polish (10%):

- The writeup has a coherent narrative throughout, with smooth transitions
- Interesting insights about the data are included in the text and are well supported by the visualizations
- The overall look and format of the writeup is professional in quality

Ambitiousness and Technical Accuracy of the New Elements (15%)

Because there are many different options for the “extra element(s)” in the project, with varying degrees of technical difficulty associated with them, I cannot list detailed criteria for each possibility.

Therefore, I will rate the ambitiousness (A) of the project on a 0-4 scale, and the overall technical accuracy of the project (T), on a 0-6 scale.
The score for this component will be calculated as `min(15, A*T)`.

In other words, to get full credit, the product of your ambitiousness and technical accuracy should be at least 15 (for example, your project could be 3/4 on the ambitiousness scale and 5/6 on the correctness scale, or 4/4 on the ambitiousness scale and 4/6 on the correctness scale).

I will be impressed if your project is both extremely ambitious and flawless, but your grade won’t exceed 15; that is, technical flourish isn’t a substitute for a good writeup, good coding practice, etc.