STAT 184 Final Project

Due: TBD

Final Project Report (& Checklist)

Purpose:

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.

Instructions

Final Project Report

- Polished & professional written investigation of your final project. The Final Project Report must be submitted to Canvas an R Notebook before the deadline. Grading details are provided below.

Final Project Checklist

- This is a secondary document with a separate purpose from the Final Project Report–please do **not** try to combine Final Project Report & Final Project Checklist into a single document.

- Just take 5 minutes to look at your Rmd and complete this using the template in the provided GitHub Repo right before you submit the **Final Project Report**
- Do this last to make sure the Rmd line numbers are correct
- It requires very little effort, but serves two important purposes:
  1. you can make sure you've done everything that's required for the project

  2. it helps the grader find where you completed each of the requirements so they are scored accurately.

Submit–TWO (2) different R Notebooks required

- Submit your Final Project Report R Notebook to Canvas before deadline.

- Submit your Final Project Checklist R Notebook to Canvas before deadline.

Note: If your project does not render properly as an (HTML) R Notebook, it will be graded from the Rmd in your provided GitHub Repo and assessed a 20% penalty.

Tips

- GitHub will not allow any single file larger than 100 MB in the repo
  - this will not be a problem at all for most projects
  - if it's close, you might try compressing the data. It's pretty easy, Google it or visit office hours.
  - (optional) if the compressed file is still too big, you might try see if you can configure Git LFS (Large File Storage) https://git-lfs.github.com/
  - you can also just use a subset of the data if you have a problem with file size–ideally, by dropping variables you were not planning to use anyway, but you can also subset the rows if you must.
- It's okay if you use some things from your EDA but this stage of the project should be much more polished and professional. Not everything you have done in the EDA will be appropriate at this stage.

Grading

See official scoring rubric accompanying "Submit Final Project Report (& Checklist)" available in Canvas for complete detail. The following is intended to summarize key tasks, but scoring is subject to change (Canvas rubric is the authority):

Data Access (15 pts)

1. Analysis includes at least TWO (2) different data sources.
2. Primary data source may NOT be loaded from an R package–though supporting data may.
3. Access to all data sources is contained within the analysis.
4. Data intake is inspected at beginning of analysis (e.g., using one or more R functions like `glimpse`, `head`, `tail`, and more)

Data Wrangling / Processing / etc (40 pts) Students need not use every function and method introduced in STAT 184, but clear demonstration of proficiency should include proper use of at least 4 of these 7 topics from class:

(1) general data wrangling using various data verbs like filter, mutate, summarise, arrange, group_by, etc.
(2) joins for multiple data tables
(3) reduction and/or transformation functions like mean, sum, max, min, n(), rank, pmin, etc.
(4) pivot_wider & pivot_longer –or similar function to stack/unstack variables
(5) regular expressions
(6) user-defined functions
(7) loops/control flow

Note: use of techniques must actually be relevant to the progress of your investigation in the Final Project. For example, use of `pivot_wider` just to show the code will not earn credit if the resulting object has not been useful to the project.

Data Visualization (40 pts) Students need not use every function and method introduced in STAT 184, but clear demonstration of proficiency should include a range of useful of data visualizations that satisfy the following criteria:

1. relevant to stated research question for the analysis
2. are neat with professional appearance including titles, axis labels, guides, etc
3. include at least one effective display of many–at least 2–variables (cannot use a map or decision tree for this requirement)

4. use of multiple geoms (e.g., points, density, lines, boxplots)
5. use of multiple aesthetics–not necessarily all in the same graph (e.g., color, size, shape, facets, alpha) Optional add-ons:
    - layered graphics (e.g., points and accompanying smoother, jittered points and accompanying boxplots, overlaid density distributions, etc
    - maps (e.g., leaflet or choropleth)
    - decision tree and/or dendogram displaying exploratory machine learning results

Code Quality (15 pts) Code formatting is consistent with Style Guide Appendix of DataComputing eBook. Specifically, all code chunks demonstrate proficiency with
(1) meaningful object names
(2) proper use of spacing and new lines
(3) use of meaningful comments

Narrative Quality (25 pts) The narrative text (e.g., sentences and paragraphs describing the progression of your analysis) must satisfy the following:

1. One guiding question or research question for your investigation is clearly stated *with a question mark*
2. Explains why this topic is important and/or interesting to investigate
3. Explains one or more significant findings or conclusions of your investigation that is **clearly related to the guiding/research question**

Overall Quality (15 pts)

- Submitted project shows significant effort to produce a high-quality and thoughtful analysis that showcases STAT 184 skills
- Analysis is well-organized and easy to follow
- Free of extraneous content such as data dumps, unrelated graphs, and other content with unclear purpose or unrelated to advancing an investigation of your research question.

Note: General R Markdown skills will be assessed here. For example, disorganized/missing/poor headers or writing standard narrative as if it's a header/quote/etc would degrade overall quality of the report.

GitHub Repo (15 pts) Requirements for *this* GitHub Repo to you.

- At least 3 commits pushed to provided GitHub repo PER person.
- all commits must make a substantive contribution to the project & have an informative commit message that summarizes the contribution
- Final Project is entirely contained in the provided GitHub Repo, including data sources, documents, and anything else that another analyst would need to reproduce your entire analysis.

Final Project Checklist (5 pts) This is a separate R Notebook to be submitted to Canvas along with the R Notebook for your Final Project Report. Immediately before or after submitting your Final Project Report to Canvas, just take 5 minutes to review the .Rmd that produced your Final Project Report. Complete the Checklist using the template provided in the GitHub Repo (`FinalProject_Checklist.Rmd`). Submit the R Notebook of your completed Checklist to Canvas before the deadline.

- 10 point *penalty* if not submitted to Canvas before the deadline. In other words, you can earn 5 easy points if this is done properly, or you can lose 10 points if this isn't done at all (or it's completely inaccurate).

- The submitted Checklist document must cite accurate line numbers in the `.Rmd` that produced your Final Project Report.
- A penalty (1 pt) will be assessed for each reference that is not accurate (i.e., within a tolerance of 5 lines)

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