STAT 220 Final Project

Assignment

The goal of the project is to go through the complete data-scientific process to answer questions you have about some topic of your choosing. You will need to acquire and wrangle your data, design your visualizations, conduct analyses (which could be visualizations), and communicate the results. The project is an opportunity to show off what you have learned about data science! Your task is to use data to tell me something interesting. This project is deliberately open-ended to allow you to fully explore your creativity.

For the final project, you must...

Create a website or Shiny app. The website/app will be your platform to communicate what you have learned.

Create a brief technical report or screencast. Create a brief write-up or screencast/recording for me (not on the website) that explains your methodology for your website/app creation. Include a brief explanation of the class tools/methods you used and any new tools/methods used. This does not need to be a step-by-step explanation, rather it should give me a roadmap for your analysis and the tools from class that you used (as well as major tools not covered in class). This should be less than 750 words if you pursue a writeup or less than 15 minutes if you pursue a screencast.

Use the materials learned in this class. You could include tools on methods such as visualization, joining/cleaning data, processing text/dates, plotting maps, introductory clustering, etc. If you could conduct your project before the term started, it is likely not a good fit.

Center your project around real data. You will likely work with large, complex, and/or messy data. While this is not an explicit requirement of the project, the more challenging your data set is, the more you will have to use the tools learned in this class. For example, one thing that will make your data science project more ambitious is combining two or more data sets that are not directly related. *You need to use data that was not provided to you for class as a STAT 120, 220, 230 or 250 student, and you cannot reuse data from a previous project for a different class, without approval.* (You may want to rethink your project if you could have used your data for a final project in one of those courses.)

Tell me something. An example of a project that doesn't tell me anything would be something that downloads a single data source and summarizes it, with some perfunctory visualizations. Make sure that your project is thought-provoking and has some underlying meaning! This is where the depth of the project comes into play.

Products

1. Web-based presentation of results: Place the URL of your site in your GitHub repo's README file

- You could create a Github page, RPubs page or a **free** Shiny web app page.
- Here is as R Markdown reference for formatting .Rmd as HTML pages and they have info for creating an .Rmd for Github Pages
- Here are some examples of how you could present your work on a (Github built) website:
  - http://cs109hubway.github.io/classp/
  - http://hamelsmu.github.io/AirbnbScrape/

  - https://claralivingston.shinyapps.io/Bachelor/
  - https://yicheng-shen.shinyapps.io/DS_Final_Project/

2. Code and data on GitHub

- Data: make sure to provide your data, if it is too large for GitHub, then store it in Dropbox or Google drive and put the share link in your repo's README file
- README: Provide a list of file names in your repo along with a brief description of their content or use in your project. The README should be a road map to your repo. Be sure to include directions to find your brief technical report or a link to your screencast here (and a link to your data if needed).
- .Rmd/.R files: These are the analysis files that need to be organized and readable. Keep your code organized, well-commented, and don't submit code not used in the analysis.

3. Peer assessment Google form linked here

Data Ideas

On the Moodle page, I will post ideas for where you might find initial dataset ideas. It is definitely not required that you find your data here.

Working in groups

The expectation is that each student shares equal responsibility for the analysis and writing. I will look both at GitHub commits and the group feedback form to help judge whether you made a substantial contribution to your group's project.

Timeline

There are two deadlines associated with the final project:

1. 10pm, Wednesday, November 8 (no late submission). Project proposal. To propose your project, a delegate of your group (only 1 person) will submit a brief proposal to this Google form describing your idea. In this proposal, discuss the datasets you want to use (including the sources), discuss what research question(s) you are trying to answer, and discuss what methods you might use to answer your question. **If you submit this early, I will try to provide feedback early!**
2. 12pm, November 20. Final submission pushed to GitHub/Gradescope and group assessment due. Everyone must submit the group assessment, but only **one person** per group should submit the project. You can indicate your team members when you submit on Gradescope.

Guidelines for professional work

These are some guidelines for what I will be looking for when I grade your final project for this class.

Professionalism - All products were submitted by the deadline.

- The project contains very few grammatical mistakes
- The project contains very few spelling mistakes
- All references are appropriately cited (use any citation style, but your data **must** be cited!)
- The README file was updated to contain the required information about each file in the GitHub repository
- I can easily navigate to your webpage based on a URL in the README file
- If you worked in a group, you submitted the group assessment form

- If you worked in a group, you were a valuable contributor to your group on all facets of the project

Project scope

- Techniques and tools from the course were appropriately applied
- The data was large, complex, and/or messy—that is, the data would not be accessible to a STAT 120, 230 or 250 student
- **Depth**: You've pursued an interesting result beyond simple exploratory data analysis.

Communication

Your website/app...

- clearly describes what information your data set contains and where you obtained the data.
- provides appropriate background information to help someone unfamiliar with the topic understand the context of the research question
- clearly states your research question(s) in an engaging way
- summarizes the major takeaways from your analysis (be sure that you have clear takeaways, not just questions remaining)
- communicates any necessary caveats to your analysis
- does not rely on "raw output" to communicate results

Data visualization and Methods

- Visualizations are appropriate for the task at hand
- Visualizations are properly polished and sized
- Appropriate methods are chosen for the task(s) (e.g. wrangling, modeling, etc)
- Methods are correctly implemented in R

Website/App functionality and Reproducibility

- A website or Shiny app was successfully created
- All components of the website/app function at the URL provided
- The website/app is clearly organized, allowing a user to easily navigate your analysis/results
- GitHub was used for project management and submission
- Commits were used to save progress periodically (I will count commits to make sure there isn't just a final "file dump")
- Informative commit messages were used
- I can pull your repo from GitHub and rerun your code without alteration to obtain your results

Coding style and quality

- The tidyverse style guide is generally followed (new lines after +'s, appropriate spacing, etc.)
- The code is readable
- The code is well organized and commented
- The code is not overly repetitive (i.e., functions are written to reduce repetition and copy/paste errors)
- A randomly selected STAT 220 student could read the code and understand the implementation

Technical report/screen cast

- Your methodology for your website/app creation was explained clearly
- You broadly explain the tools/methods used, pointing out any new tools/methods you needed to learn (and provide references for how you learned them).
- The roadmap for your analysis is clear—a randomly selected STAT 220 student would understand the scope and general tools used.