STAT 408 Project Instructions

Spring 2022

Project Setting

The data visualization and wrangling techniques learned in this course help you tell stories by transforming data into a form easier to understand. This project allows you to explore a data set you are interested in, as we continue to refine your skills as users and storytellers of data. This project contains five components and has three deadlines. It is highly encouraged that you reach out (staceyhancock@montana.edu and elijah.meyer@montana.edu) if you have any questions.

You will be working within Github with your lab teammates. A team project repo will be created for you in the stat408-s22 Github organization titled "project-[YOURTEAMNAME]". All components of the project will be turned in via Github and **must be reproducible**.

Project Learning Outcomes

1. Identify and scrape a data set from the web.
2. Determine research questions that can be addressed using a particular data set.
3. Create interactive visualizations and summaries of a data set that address a specific research question.
4. Write a data exploration summary that addresses a specific research question.
5. Critique data visualizations and summaries created by other teams.

Project Components

Pick a Context

What are you interested in exploring? For this project, your group is challenged to collectively come up with one context that you are excited about. This could sports, education, etc. Then, find a source of data online that is relevant to your context. Make sure that these data:

1. have **at least 5 different variables and at least 40 observations**, ideally with a mixture of categorical and quantitative variables, and
2. are not available in downloadable "tidy" data form (e.g., .csv, .xlsx); that is, you must make use of some of the web scraping tools we learned in class.

START EARLY! Finding a data set often takes longer than people think. Treat this project as opportunity to explore something you care about, practicing using what you've learned in class in a context you find interesting.

1. Data Scraping - 5 points

Once you have found a source of data online you are interested in, you will practice your data scraping skills learned in Lab 6. Create an R script titled "scrape.R" in your Github repo, and save the data set in the "data" folder of your repo.

2. Research Questions - 5 points

Once you have scraped your data, you will come up with **two research questions of interest** *that can be explored using the data you scrape*. At least one of these questions must include **at least three variables**. What are you curious about? What are you interested in? Ask two questions to be answered using your data and the subsequent data visualizations you will be creating.

Deadline 1: Friday, April 8th by 11:00pm

By April 8th at 11:00 p.m., the following must be completed and available in your Github repo:

1. Topic you are interested in researching (in README.md)
2. Data source (in README.md)
3. Two research questions (in README.md)
4. R script that scrapes a data set of interest (scrape.R)
5. Scraped and cleaned data set (in the data folder)

Consider this a *proposal* to obtain feedback on your research questions before diving into data exploration. Your proposal is graded on effort, creativity, reproducibility, and completeness.

3. R Shiny Dashboard

Your final document will be an R Markdown document that embeds Shiny, published on https://www.shinyapps.io/. Shiny is an R package that makes it easy to build interactive web apps straight from R. You will get practice with this, and are encouraged to reference Lab 7. Using these tools, you are tasked to create a combination of tables, visualizations, annotations and discussion that allow the user to answer your research questions. The purpose of this R Shiny dashboard is to tell the *story* of your data; it should be a stand-alone document similar to what one might find from the *New York Times* graphics data journalists.

Data Summary & Visualization - 20 points

Using what you've learned in class, create appropriate and neatly formatted summary statistics tables from your data relevant to your question(s). Additionally, create appropriate and professional figures that help answer your research question. We stress **quality** over quantity. A couple high quality visualizations will receive a much higher grade than a large number of poor quality visualizations. There is no set number of visualizations that you need to create (as long as there is at least one). Create enough to explore your research questions.

Organize this combination of tables and visualizations into your R Shiny dashboard. At least one table OR visualization must be interactive. **Use the interactive functionality purposefully!**

Discussions - 10 points

Embedded in your R Shiny dashboard, write 2–3 paragraphs that:

- give the reader a brief explanation of your context, research questions, & why your research is important,
- cite the source of the data with a brief explanation of the variables,

- using evidence from your generated summary statistics and visualizations, answer each of your research questions, and
- discuss any potential limitations of your investigation.

Deadline 2a: Friday, April 22 by 11:00pm

By April 22th at 11:00 p.m., the following must be completed:

1. "Data Summary & Visualization" portion of your RShiny dashboard R code completed and available in GitHub

2. "Data Summary & Visualization" portion of your RShiny app deployed to https://www.shinyapps.io/
3. Link to RShiny app posted at the bottom of README.md in your Github repo

Deadline 2b: Friday, April 29 by 11:00pm

By April 29th at 11:00 p.m., the following must be completed:

1. Complete RShiny dashboard (including both "Data Summary & Visualization" and "Discussions") R code completed in GitHub
2. Link to RShiny app posted in D2L Project Shiny Apps discussion board

R Shiny Dashboard Grading Criteria

Your R Shiny Dashboard will be assessed for the following:

- **Correctness** - Are the statistical tables / figures produced appropriate for the research question? Are the figures professional? (i.e., is this at the level of quality to be published in a journal article)

- **Discussion** - What is the quality of the writing? Does the writing reference created tables / figures in a meaningful way to help address the research questions? Does the writing set the stage for why these questions are important to investigate? Are limitations carefully considered?

- **Creativity and Critical Thought** - Does it appear that time and effort went into the planning and implementation of the project?

- **Reproducability** - Can another reproduce the scraping of your data set? Can others interact with a figure / table created to help further investigate your research questions?

4. Peer Feedback - 5 points

Deadline 3: Friday, May 6 at 11:00pm

By May 6th at 11:00 p.m., the following must be completed:

1. Post comments on at least **two** RShiny app discussion post including at (a) least one feature you think works well, and (b) at least one feature that you might have done differently, and (c) how you could extend the study (e.g., what other research questions does the analysis inspire?).

5. Teamwork - 5 points

Other than the peer feedback, you are to complete the project as a team. All team members are expected to contribute equally to the completion of this assignment and team evaluations will be given at its completion — anyone judged to not have sufficiently contributed to the final product will have their grade penalized. While different teams members may have different backgrounds and abilities, it is the responsibility of every team member to understand how and why all code and approaches in the assignment work.