Project-2 Description

The Project-2 for this class is the analysis of a data of your own choosing. The dataset may already exist, or you may collect your own data using a survey, by conducting an experiment, or by webscraping. You can choose the data based on your interests or based on work in other courses or research projects. The goal of this project is for you to demonstrate proficiency in the techniques we have covered in this class (and beyond, if you like) and apply them to a dataset in a meaningful way.

Data

In order for you to have the greatest chance of success with this project it is important that you choose a manageable dataset. This means that the data should be readily accessible and large enough. As such, your dataset must have at least 50 observations and between 9 to 20 variables.

All analyses must be done in R. If you are using a dataset that comes in a format that we haven't encountered in class, make sure that you are able to load it into RStudio/Jupyter as this can be tricky depending on the source.

Do NOT reuse datasets from textbooks.

Analysis and Write up

Your write up and all typesetting must be done with using R Markdown or Jupyter Notebook.

Your **introduction** should introduce your general research question, your data (where it came from, how it was collected, what are the variables, etc.).

After providing the description of your dataset and research question in the introduction, provide a very preliminary exploratory data analysis, including some summary statistics and visualizations, along with some explanation on how they help you learn more about your data.

Use the remainder of your write up to showcase how you have arrived at an answer / answers to your question using regression analysis and any techniques we have learned in this class (and some beyond, if you're feeling adventerous). The goal is not to do an exhaustive data analysis i.e., do not calculate every statistic and procedure you have learned for every variable, but rather let me know that you are proficient at asking meaningful

questions[1] and answering them with results of data analysis, that you are proficient in using R, and that you are proficient at interpreting and presenting the results. Focus on methods that help you begin to answer your research questions. You do not have to apply every statistical procedure we learned. Also pay attention to your writing. Neatness, coherency, and clarity will count.

Your write up must also include a one to two paragraphs **conclusion** and discussion. This will require a summary of what you have learned about your research question along with statistical arguments supporting your conclusions. Also critique your own methods and provide suggestions for improving your analysis. Issues pertaining to the reliability and validity of your data, and appropriateness of the statistical analysis should be discussed here. A paragraph on what you would do differently if you were able to start over with the project or what you would do next if you were going to continue work on the project should also be included.

The project is very open ended. You should create some kind of compelling visualization(s) of this data in R. There is no limit on what tools or packages you may use. You do not need to visualize all of the data at once. A single high-quality visualization will receive a much higher grade than a large number of poor-quality visualizations.

You can add sections as you see fit. Make sure you have a section called Introduction at the beginning and a section called Conclusion at the end. The rest is up to you!

Deliverables

Your submission should be a ZIP-file `Lastname.zip`, that contains:

- `RMarkdown + knitted HTML files / Jupyter Notebook`
- dataset file

[1] Please watch this to understand what a **good** research question is.