Research Project: Exploratory Data Analysis

Due date: 7 March, 11:59pm

Details

Now that you have chosen a research question and dataset, the next step is to perform some exploratory data analysis to better understand characteristics of your dataset, patterns in the raw data, and to present descriptive statistics related to your data and your research question.

There is no set format or page limit, but here are a few pointers:

- **General characteristics of dataset**: for example, how many observations, how were the data collected (is the dataset representative of the population of interest?)
- **Missing data**: If your dataset does not have any missing observations, then fine to just say this (don't need to do EDA graphs or discuss). If you have missing observations, summarize what is missing, and give a brief discussion about whether or not you think missingness may be a problem (e.g. is there more likely to be missing data for some groups compared to others?)
- **Summary statistics of variables of interest**: for example, you could present a table of summary statistics of main variables, including things like number of observations, mean/median/sd (if a continuous variable), proportions by group, etc…
- **Graphs showing both univariate and bivariate patterns**: Remember back to EDA lecture about appropriate graphs to show patterns in different types of variables. You, me, and Julia are likely to be interested in both univariate patterns (e.g., the distribution of continuous variables, proportions for categorical outcomes…) and bivariate patterns (e.g. scatterplots, proportions/boxplots by group, trends over time…).
- **Pick a few key graphs**: You could keep going ad infinitum. But good EDA reports will just pick a few key graphs to show key relationships/patterns, and have a good discussion of these. If you put in 20,000 graphs things become hard to distill and understand. I know you have probably done 20,000 graphs to pick the 3-5 graphs that go into the report.
- **Discuss what you see**: Good reports will have a good discussion about patterns in the graphs and what they potentially mean. In most cases this is more than just one sentence. If patterns (or absence of patterns) are surprising, you can note that down.
- **How does this inform your model choice**: One of the big reasons why we do EDA is to work out reasonable models to try. Summarizing what you found and how this will affect your decisions about what variables to include is a good thing to include.

What to submit

It is expected that you present and write up your findings in Rmd. You should submit:

- your R Markdown file; and
- the knitted PDF resulting from your R Markdown file.

Please submit both files via Quercus.

If you copy paste graphs/tables into a Word Document then you will lose marks.