**Project Proposal & EDA**

**Due: 11:59pm, April 16, 2021 (Friday)**

**20 points**

In the 2nd part of the project, you will write a project proposal, do data cleaning and preprocessing and exploratory data analysis.

1. Title: Have a proper title for your project
2. Purpose & Motivation:

In this part, describe the general topic/phenomenon you want to study. State the research questions that you hope to answer about the dataset. You have done this in the Data Appendix. You may use the same questions or modify them.

Build a case for why your research question is important and explain what motivated your question. Don’t assume that “it’s obvious”. Explain why your question is relevant and what the implications might be if you find an association. You may use information from journal articles or reputable websites (e.g. The New York Times, Time magazine, The Wall Street Journal, etc.) .

1. Data cleaning & preprocessing:
2. Deal with N/A values if your dataset has them. The easiest way is to drop the observations that contain N/A values. N/A values might be represented as NaN, None, or other strings.
3. Recode the variables if necessary. For example, if you use a categorical variable as a predictor and it is coded as 0/1, you have to let Python know that it’s categorical. If you use a categorical variable as a response and it contains strings (e.g. Yes/No), you might need to map it to 0/1 so that you know what category is success. If you have a categorical variable with many categories, it might not work well in the model. It’s better to recode it to fewer categories.
4. Make new features if necessary. For example, transform a variable by taking the logarithm or square root. Or, create the product/ratio of two variables.
5. If you have multiple data files, you will need to merge them depending on the research questions that you ask.
6. Exploratory data analysis:
7. Look at the distribution/summary statistics of the variables. You don’t need to show all of them. Just do it for the most relevant ones.
8. Look at some bivariate graphs relevant to your questions (i.e. graphs involving the response variable and one explanatory variable). Show at least two of them. Make sure the graph has labels for x-axis, y-axis and a title.
9. Describe what you see from the graphs.
10. Submit your .ipynb file and the .html file on Moodle. Please change the name of your file(s) to something like “group-X-proposal” + file extension. Each group only needs to submit one copy.