CAP 5302, Project: Data Description.

Provide a detailed description of the data you've selected for your project, including

1. The source and inspiration for selecting this particular data set. If you don't have any data preferences off the top of your head, some examples of data sources are *Kaggle.com*, *https://datasetsearch.research.google.com/*, UCI Machine Learning repository, *sports-reference.com*, among many others.

   Note: It shouldn't be among data sets that we've already used in class, or that are contained in any of the *R* packages. It should be obtained from external source (most likely a *.csv* file, or via web scraping).

2. The size of the data (# of observations and # of variables).

3. Description of all variables (similarly to descriptions you encounter for *R* data sets, e.g. https://stat.ethz.ch/R-manual/R-devel/library/MASS/html/Boston.html)

MINIMUM REQUIREMENTS: Make sure that your data set

- Contains a total of at least 10 variables and 50 observations.
- Is NOT a time series.
- Has several numerical variables, one of which is expected to act as a response variable in linear regression (your main variable of interest).
- Contains at least two categorical variables, one of which can potentially act as a response, and another one as a predictor. Variables such as unique ID's or unique names don't count as categorical variables.
- **NOTE**: Keep in mind that (eventually) you will be expected to fit **at least one linear regression model**, and **at least one logistic regression model**. For the latter, if you don't have an appropriate categorical variable in your data set, might potentially create a "synthetic" binary response variable via breaking down the numerical response variable (that you used for linear regression) into two categories (e.g. "High"/"Low" depending on whether the value is above or below median).

The report should be

- **no more (!) than 2 pages long**.