PSTAT 126: REGRESSION ANALYSIS

FINAL PROJECT

The purpose of this document is to provide guidance for structuring your project. The deliverables will consist of an R markdown in .Rmd to be submitted on Canvas and knitted pdf version to be submitted on Gradescope.

PART - 1
Data Description and Descriptive Statistics

The first objective is to understand the variables in your dataset and their relationships. Your task is to choose a dataset Diamonds from Kaggle: Here's the link, describe it, and perform some descriptive statistical analyses. You should carefully consider the observational units in the dataset to ensure they are independent. The dataset should have at least 100 independent cases/observations (the ideal number of observations is 200-500). Be wary of missing observations! If you have too many, pick 500 observations randomly from the dataset. That way, scatter plots will not seem overwhelmingly charged. You should include atleast 2 categorical variables and 3 independent quantities.

The final report should include the following:

- A description of all relevant variables, and the observational unit (i.e., row).
- Appropriate summary statistics with adequate explanation and interpretation. You should examine binary or other relationships, as well as describing individual distributions. Consider making an appropriate table of some kind. The package skimr has a function called skim, which is great for summarizing data succinctly
- comment on anything of interest that occurred during the project. Were the data approximately what you expected, or did some of the results surprise you? How did the sampling go? Do you think you got a representative sample of your population?