**Project\_Submission\_Part 1**

**Group 6**

**By Naina Grover, Yifan Wang , Weiwen Qi, Xiaochen Zhang**

**Overview of Dataset**

Our data set portrays students' performance in the exams and marks secured by them in various subjects. This data set includes scores from three exams and a variety of personal, social, and economic factors that have interaction effects upon them. The purpose is to figure out which kind of factors that may impact the final performance of students.

**Number of observations and Variables**

The dataset contains 1,000 observations and 8 variables, including gender, race, lunch (standard or free/reduced), test preparation course, parental level of education and the test score of different exams. Within these variables, the former five ones are categorical variables and the last three are numerical variables.

**Techniques to be used for data analysis**

The methodology for this study includes three key points we have learnt in class: model selection by using “regsubsets()” function for picking better fitted variables. After that, we plan to take linear regression and tree decision as our analysis methods.

The first method is Model Selection. The model selection can help us to decide which variable is statistically significant. For example, maybe the gender is significant at 5% level and the lunch is not significant. After we get the result, we can optimize the model by using fewer variables and get more precise results.

The second method is Linear Regression. This is the most common model and based on initial scatter plots, there are some correlations between scores and other variables. We could further explore the correlations and predict the scores.

The third method is the Decision Tree (classification & regression). For classification, the categorical variables like gender, race and other variables can be treated as a cutoff point according to their own characters, and for numerical data, for instance, the writing or reading scores, we can divide them into several intervals (70-80, 80-90) and then plot them into a classification pattern. For the regression pattern, we may take the result of linear regression and use tree() function to plot the result.

We are inclined in using the above methods and will choose the two most interesting methods out of the above three.

**Link of the Dataset:**

Source:Kaggle <https://www.kaggle.com/spscientist/students-performance-in-exams?select=StudentsPerformance.csv>