Data Files

We are using the student performance dataset from Kaggle. The dataset contains marks obtained by students in college from the United States.

The goal of our analysis is to predict variables that provides us with the better prediction for gender identification. The focus of our prediction is not based on whether the gender is male or female, instead we are focused on identifying features that may be more relatable to the gender classification. We are taking gender as our target variable(y) and other features as responses(X). There are 8 features in our dataset.

1. Gender – Describes the gender of student
2. Ethnicity- Ethnicity of the student
3. Parental education – Level of parental education
4. Lunch – Type of lunch student took before the exam
5. Test preparation- Whether the student prepared for exam and took the preparation test
6. Math score – Score in math
7. Reading score - Score in reading
8. Writing score- Score in writing

We have performed some data wrangling techniques to perform feature engineering. Below, we are firstly describing the dataset to know its mean, variance, quartile, and min and max deviations to understand the dataset. But only three features have shown the summary of data as data is given in the continuous form in these features.

In Data Cleaning step, we will find if the dataset contains any null values or not. After operating we concluded that there are no null values in the data set. Later we operated to find duplicate records. After handling that, we converted categorical data into numerical data by performing

df = pd.get_dummies(df). The last step, in the data wrangling that we did, was to handle an inappropriatenaming convention. Finally, all the cleaned data is stored in **StudentPerformance_cleaned.csv.**