Student Performance Analysis

Student Performance Analysis

Proposed Work & Real-time Usage

To know which factor may affect the student's performance, we classify the score into a couple of ranks and figure out which feature affects the score more significantly. The independent variables follow:

1. Gender: sex of students
2. Race/ethnicity: ethnicity of students
3. Parental level of education: parents' final education
4. Lunch: having lunch before the test (normal or abnormal)
5. Test preparation course: complete or not complete before the test

Problem Statement:

To understand the influence of various factors like economic, personal and social on the student's performance

1. How to improve the students' performance in each test?

2. What are the major factors influencing the test scores?

3. Effectiveness of test preparation course?

4. Other inferences

An Exploratory Data Analysis Certain SocioeconomicIndicators with Classification

Objectives :

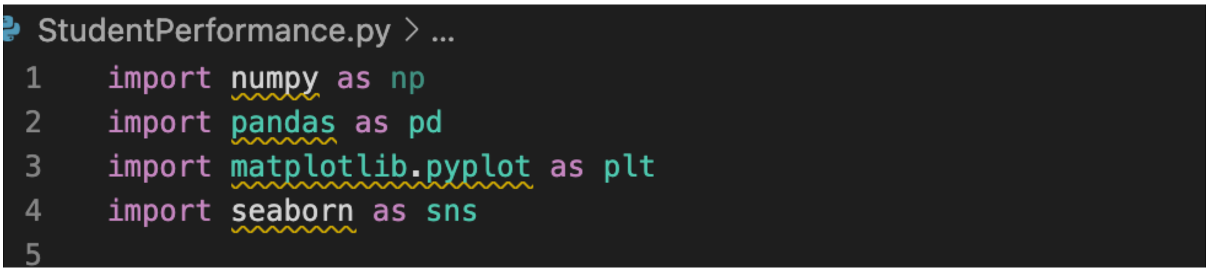
In this project, we look to explore the relationship between students' performance and certain socioeconomic indicators. We will look at how different socioeconomic categories impact students' performance, and the correlation (if any) it has on students' grades. We will perform classification tasks on the cleaned dataset. After using the original clean dataset to train models, we will then balance the dataset using oversampling and under-sampling methods to see how the metric scores change when balancing the dataset.

- [X] Import Data
- [X] Clean Data
- [X] Perform Exploratory Analysis of Data
- [X] Draw Conclusion from Analysed Data
- [X] Train Classification Models on Data
- [X] Create a Confusion Matrix of Predicted Values vs Actual Values (Unbalanced Data)
- [X] Balance Data Using an Oversampling Method
- [X] Train Classification Models on Oversampled Data
- [X] Create a Confusion Matrix of Predicted Values vs Actual Values (oversampled balanced Data)
- [X] Balance Data Using an Under-sampling Method
- [X] Train Classification Models on Under-sampled Data
- [X] Create a Confusion Matrix of Predicted Values vs Actual Values (under-sampled balanced Data)

Modules & Explanation

Module 1: DATA CLEANING

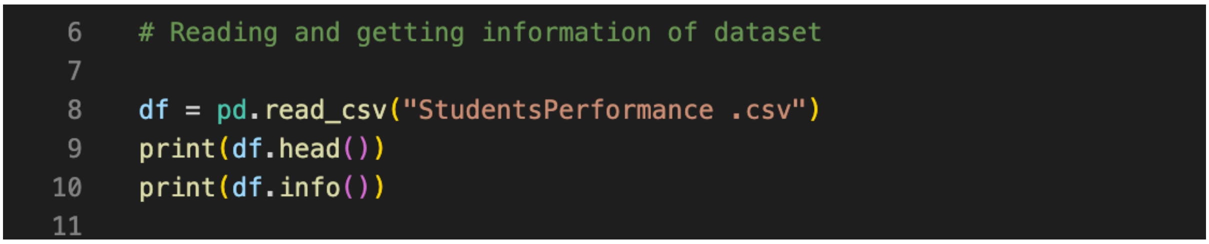
- This module will use various in-built functions from Python libraries NumPy and Pandas to clean the dataset.



StudentPerformance.py > ...

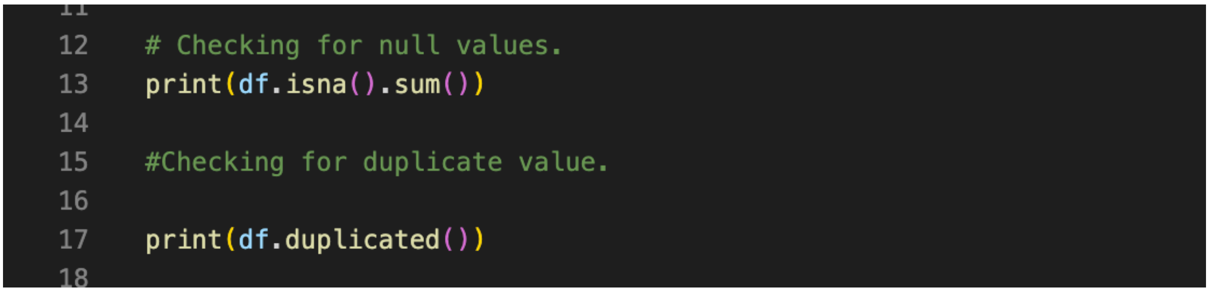
```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

```

- We will begin by loading the CSV file in a data frame, reading the first five rows using the head function, and getting the information about all the columns using the info function.



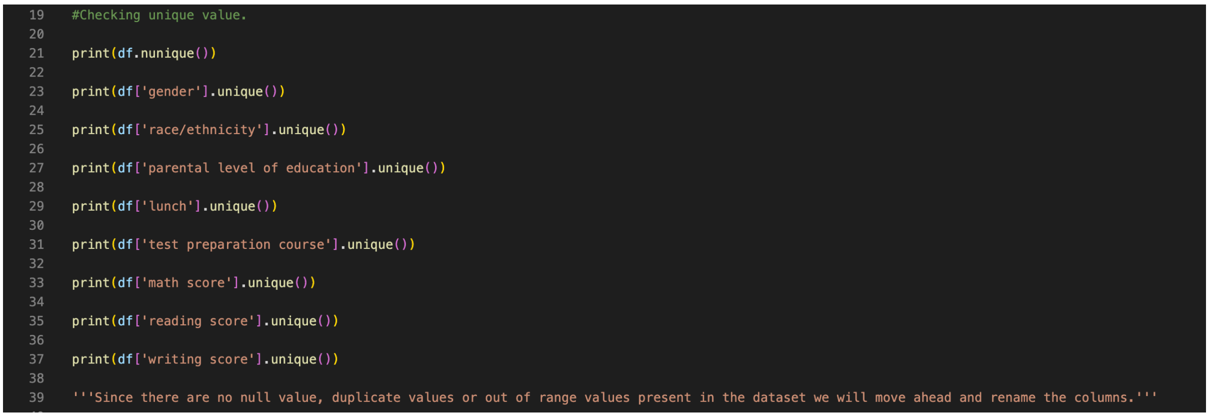
```python
# Reading and getting information of dataset

df = pd.read_csv("StudentsPerformance .csv")
print(df.head())
print(df.info())

```

- After this, we will use the isna().sum() and the duplicated() function to check for the null values and duplicated values.



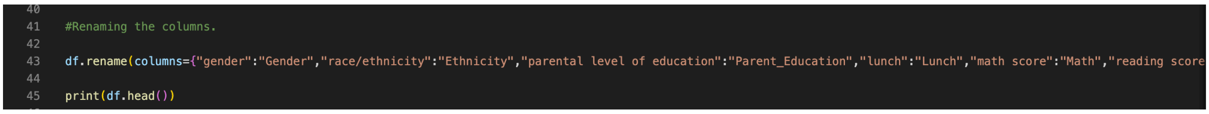
```python
# Checking for null values.
print(df.isna().sum())

#Checking for duplicate value.

print(df.duplicated())

```

- After this, we will examine unique values present in our dataset using the unique function.



```python
#Checking unique value.

print(df.nunique())

print(df['gender'].unique())

print(df['race/ethnicity'].unique())

print(df['parental level of education'].unique())

print(df['lunch'].unique())

print(df['test preparation course'].unique())

print(df['math score'].unique())

print(df['reading score'].unique())

print(df['writing score'].unique())

'''Since there are no null value, duplicate values or out of range values present in the dataset we will move ahead and rename the columns.'''
```

- Since our dataset contains no null, duplicate, or out-of-range values, we will move forward with renaming our columns for easier access to them.
- For renaming our columns we will use rename function.



```
#Renaming the columns.

df.rename(columns={"gender":"Gender","race/ethnicity":"Ethnicity","parental level of education":"Parent_Education","lunch":"Lunch","math score":"Math","reading score

print(df.head())
```

- Since there isn’t a current requirement to change the columns' data types, we will end our data cleaning and move ahead to our next module.

Module 2: DATA EXPLORATION

- This module will use various in-built functions from Python libraries Seaborn and Matplotlib and visualise our data using various types of graphs.
- These graphs will help us to understand our dataset in a much better way and will also help with drawing comparisons.
- All of this together in the end will help us to conclude our analysis.

Module 2: DATA EXPLORATION

- This module will use various in-built functions from Python libraries Seaborn and Matplotlib and visualise our data using various types of graphs.
- These graphs will help us to understand our dataset in a much better way and will also help with drawing comparisons.
- All of this together in the end will help us to conclude our analysis.

Module 3: HYPOTHESIS TESTING

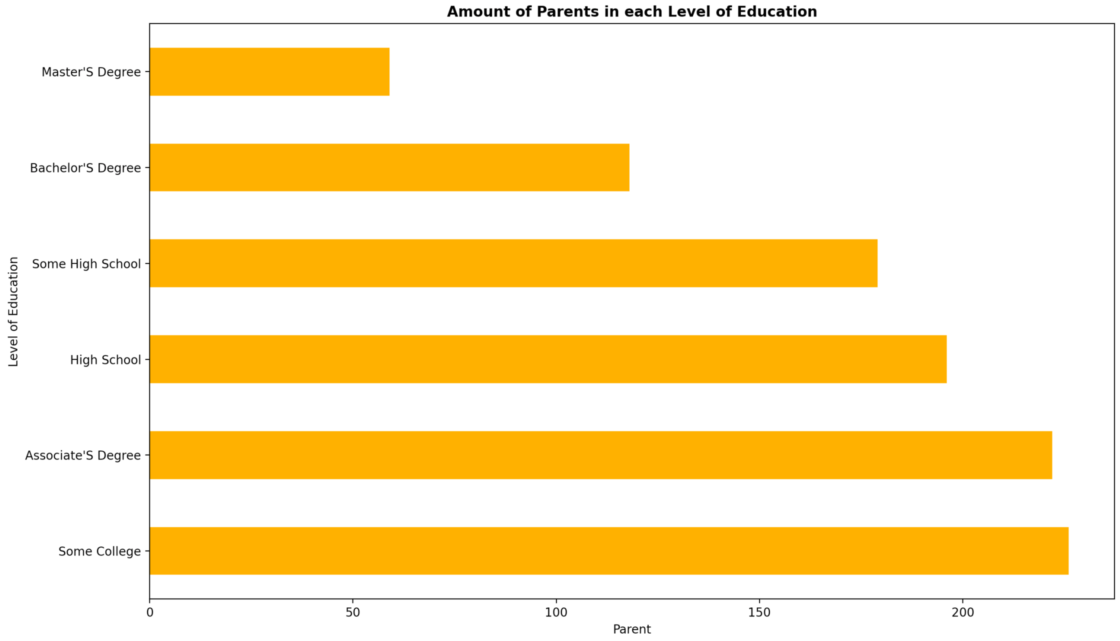
- Hypothesis testing is done to confirm our observation about the population using sample data, within the desired error level.
- Through hypothesis testing, we can determine whether we have enough statistical evidence to conclude if the hypothesis about the population is true or not.
- To trust your model and make predictions, we utilize hypothesis testing. When we will use sample data to train our model, we make assumptions about our population.
- By performing hypothesis testing, we validate these assumptions for a desired significance level.

Module 4: MULTIPLE LINEAR REGRESSION MODEL

- One of the most common types of predictive analysis is multiple linear regression.
- This type of analysis allows you to understand the relationship between a continuous dependent variable and two or more independent variables.
- The independent variables can be either continuous (like age and height) or categorical (like gender and occupation).
- It's important to note that if our dependent variable is categorical, we will dummy code it before running the analysis.

OBSERVATIONS FROM DATA EXPLORATION

Clear Understanding of Education Level



A clear understanding of the differences between degrees will help us analyse the data. Below is a brief summary that can help us make the distinction:

1. *Associate Degree*: Full-time associate's degree will take 2 years to complete
2. *Bachelor's Degree*: Full-time bachelor's degree will take 4 years to complete
3. *Master's Degree*: A master’s degree is an advanced graduate degree that can be pursued after you have completed a bachelor’s degree.

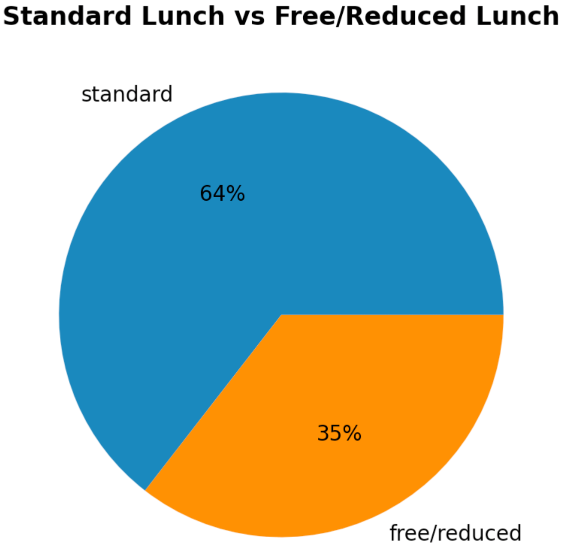
It is also important to understand the differences between College and University. These two words are used interchangeably, but there is a distinction between the two. Here is a short summary of the differences between the two:

Colleges, compared to universities, tend to be a smaller institution that focuses on undergraduate programs. Often, colleges offer liberal arts programs with broad areas of study in subjects like the humanities, science, and creative arts The Universities offer both undergraduate and graduate programs.

If we look at the Parent_Education column, we find 2 categories which are some high schools and some colleges. Since there are no explanations about the definition of these two terms, I am going to infer that 'some high school' means a parent who went to high school but did not finish their education. The same goes as 'some college', meaning that a parent who went to college but did not finish their education

From the graph above, we can see that a lot of the parents attended college but did not finish, Therefore, they can't be categorised as having a degree.

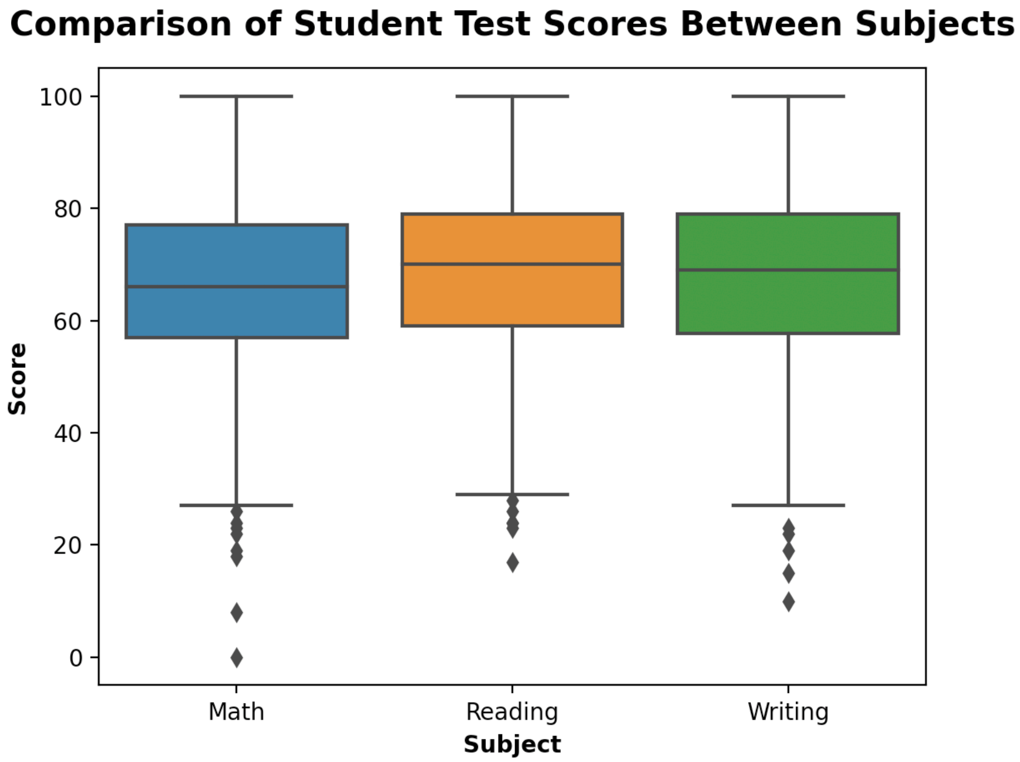
Clear Understanding of Lunch Terminology



1. *Standard*: A student who has to pay full price for their lunch
2. *Free/reduced*: A student who pays a reduced amount for their lunch or receives their lunch for free

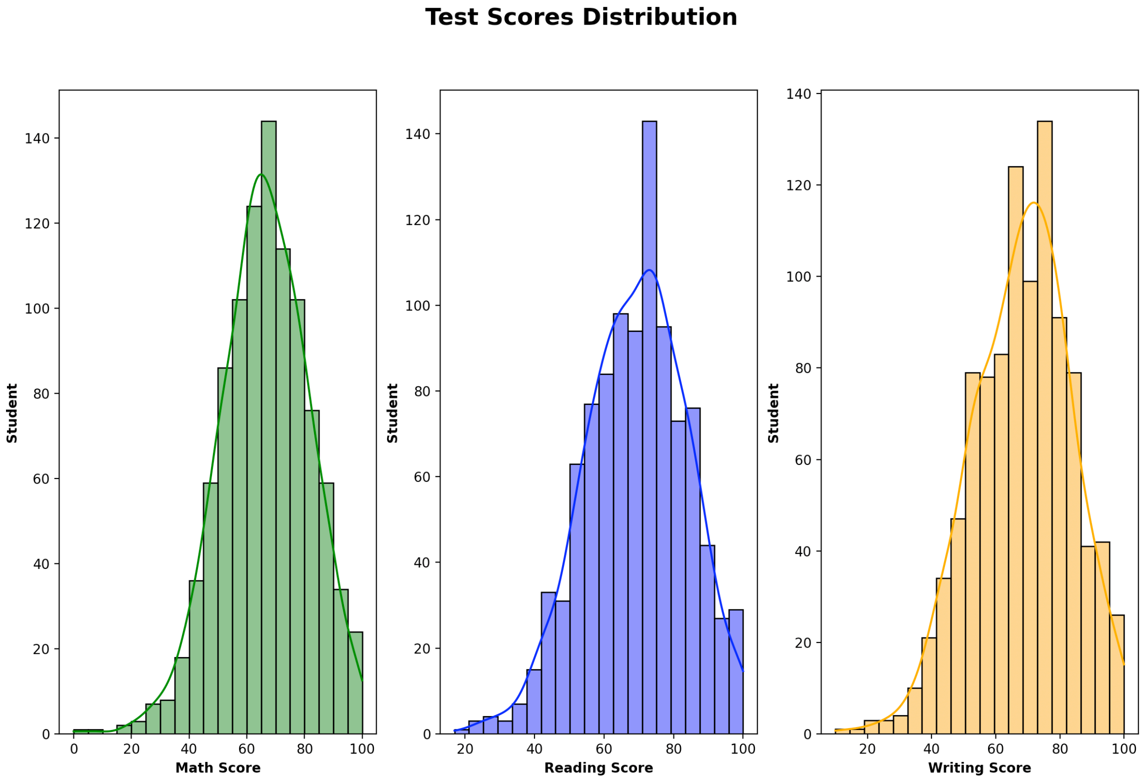
From the graph above, 64% of students need to pay standard pricing for their lunches and only 35% paid a reduced price or gets free lunch.

The Overall Student's Performance In Each Subject



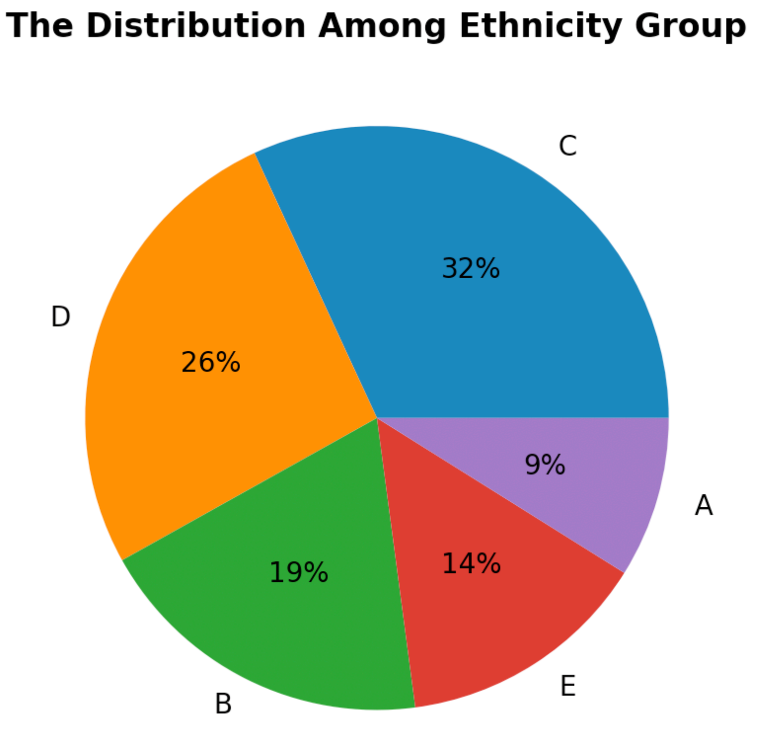
By the box plot above, we can conclude that most of the students achieve a slightly higher score in reading compared to writing and math. There are students who has noticeably low scores on Math reaching below the score of 20.

The Distribution of Student's Test Scores For Each Subject



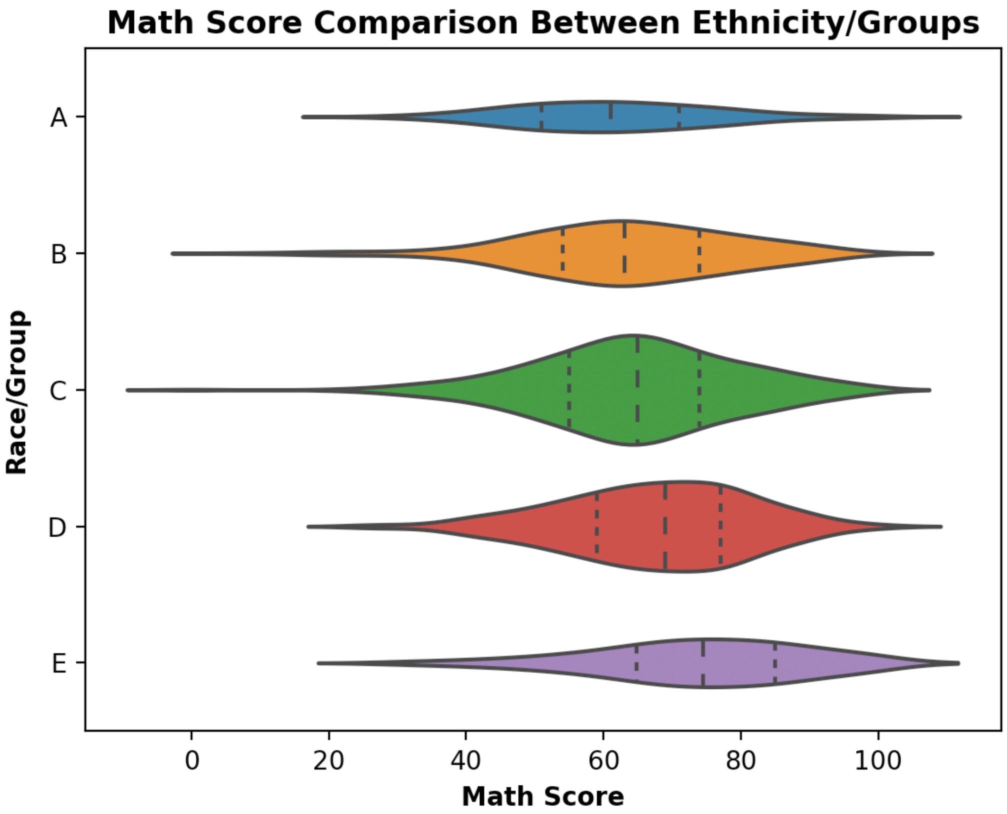
The histograms above show that the data is slightly left-skewed. Therefore the use of the median is better due to the average score being affected by extreme values or outliers.

The Distribution of Ethnicity/Group among Students

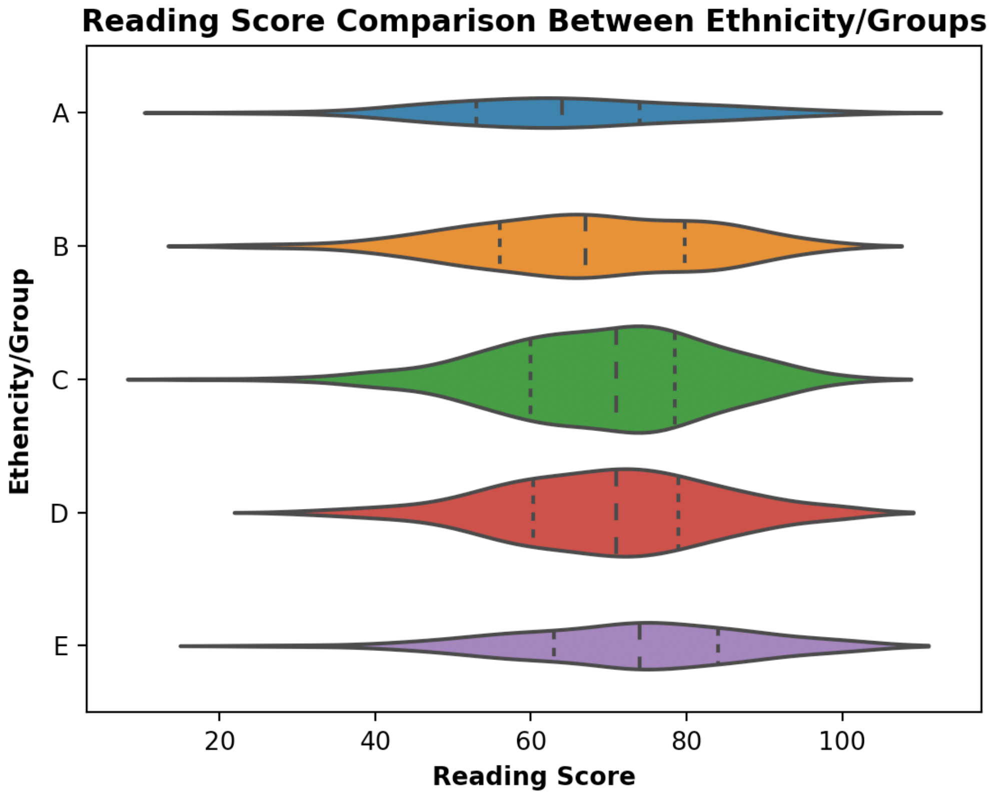


The students in each race/ethnicity are not evenly represented. Group A is very under-represented compared to other race/ethnicity. If for example, a few students produced in group A produce really, high scores and really low scores, the spread of the scores within the group would be quite wide.

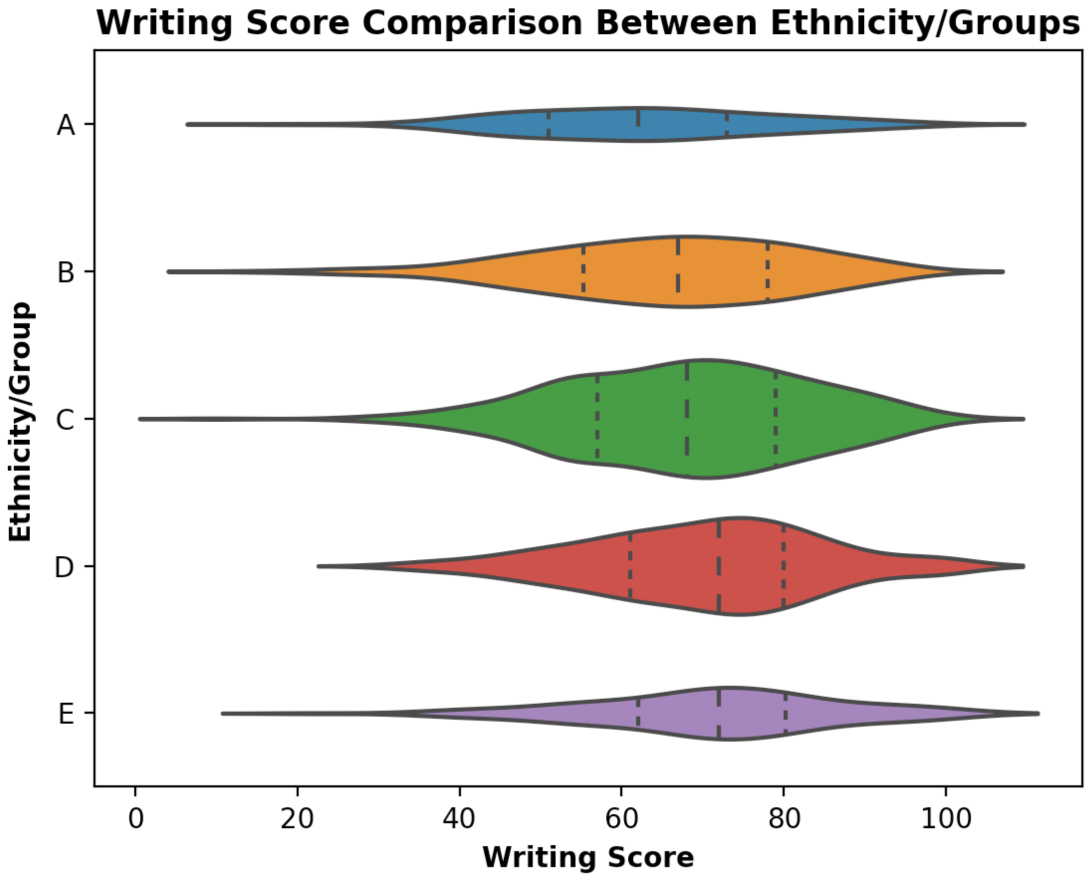
Student's Performance Compared By Each Ethnicity/Group



Group E performs better in Math. Interestingly, **Group C students produce one of the lowest scores in Math**. This means that from the 'Comparison of Student Test Scores Between Subjects' boxplot, we can conclude that the students who achieve low scores in Math are from Group C.



Almost all groups perform the same on the Reading Subject with group A falling a little bit behind the other groups. We can see that **some students from each group scored lower than 20 except for students who are in group D** which not even one student scored below 20 for Reading.



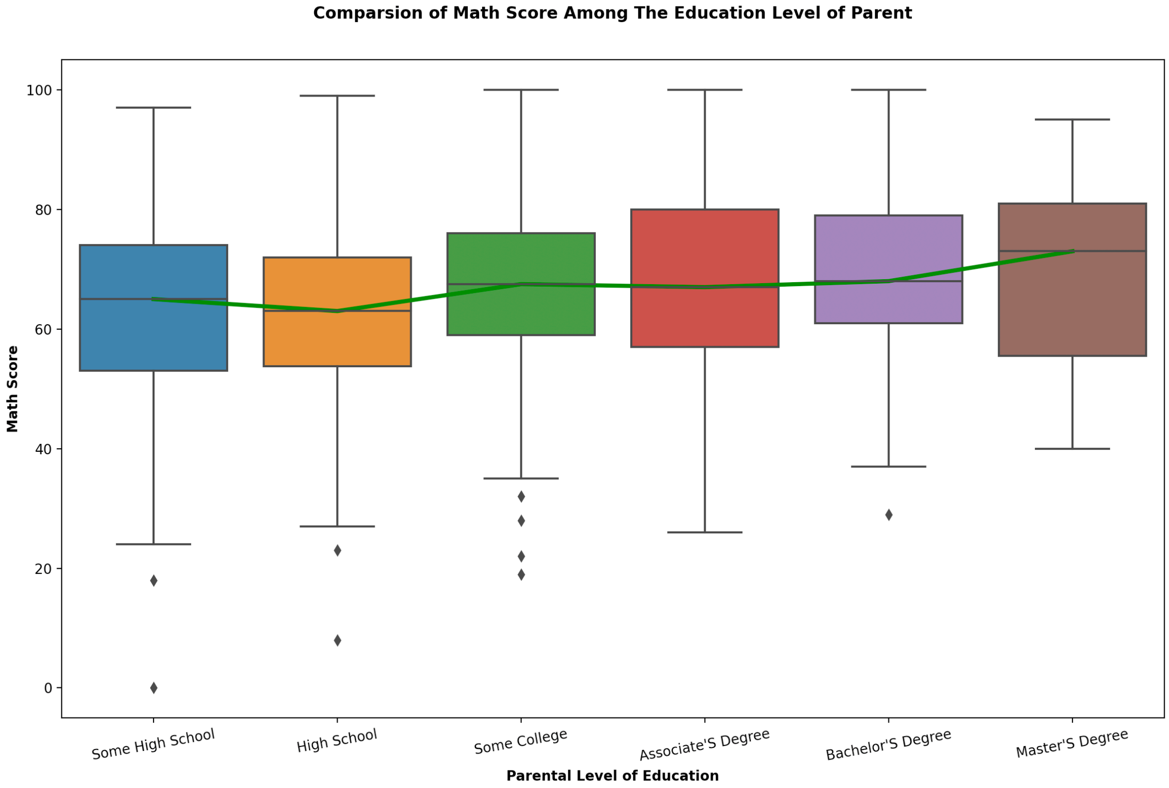
Group E performs the best in writing. However, it is not significantly ahead of other groups.

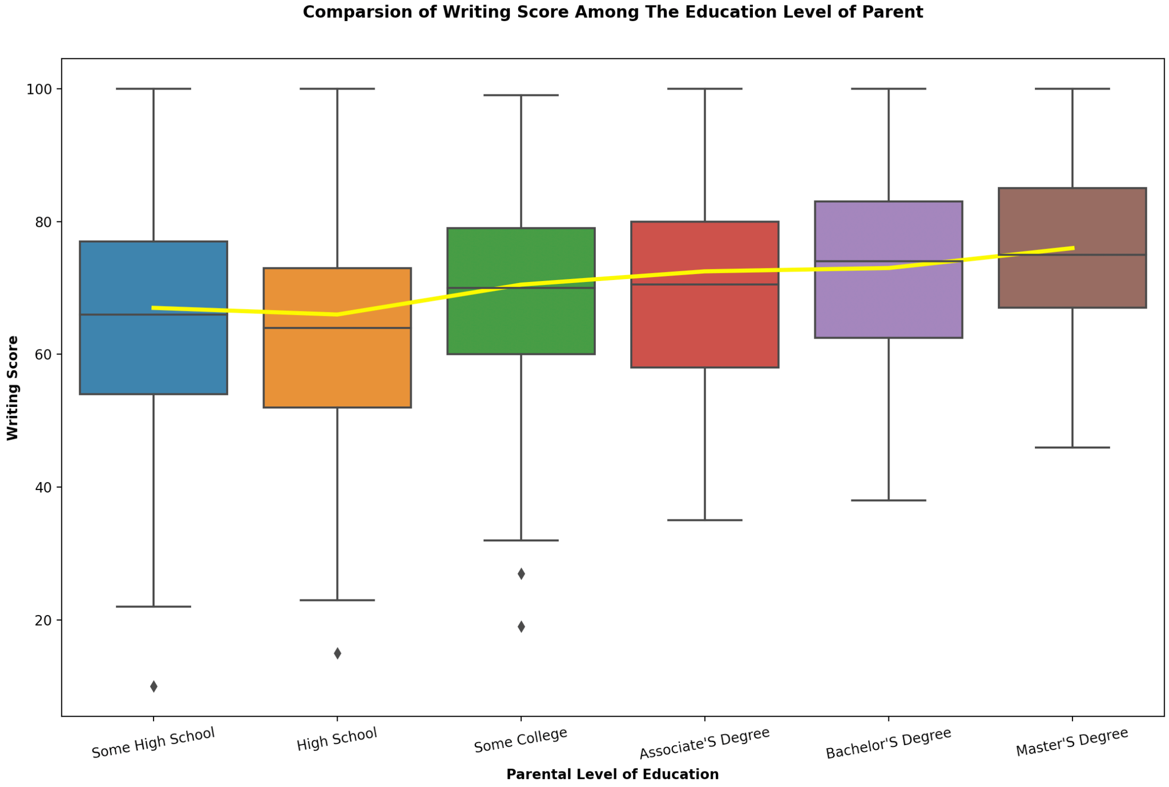
Overall Observations From Violin Plots

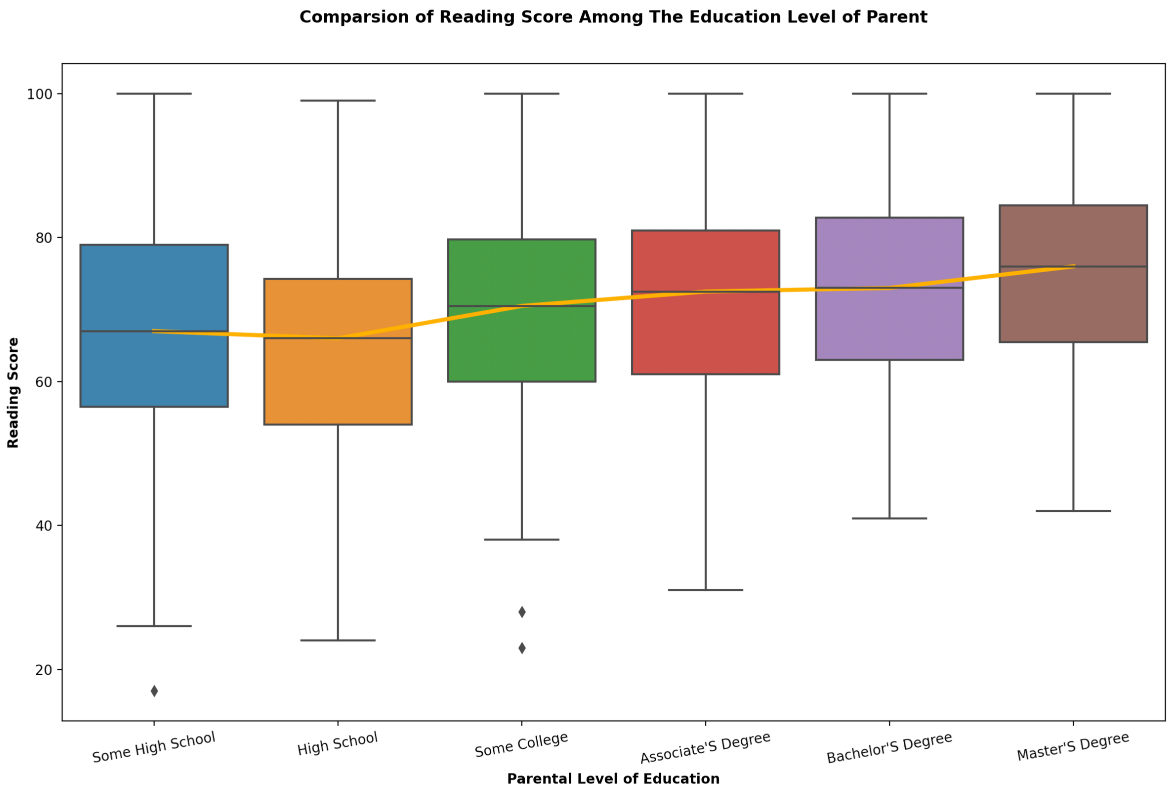
From the 3 violin plots above, we can conclude that **group E performs better than the other groups**. This might be due to the number of students in group E which is only 14 per cent of the overall students. **Group A performs the worst in all subjects even though the number of students in that group compared to overall is only 9 per cent**. We can infer that the amount of students in each ethnicity/group does not really affect the overall group performance.

However interesting to note, since **group C** has the largest amount of students (32%), they are the group **to produce the largest amount of low scores (<40) in all subjects**.

The Effect of Parent Education on Student's Performance

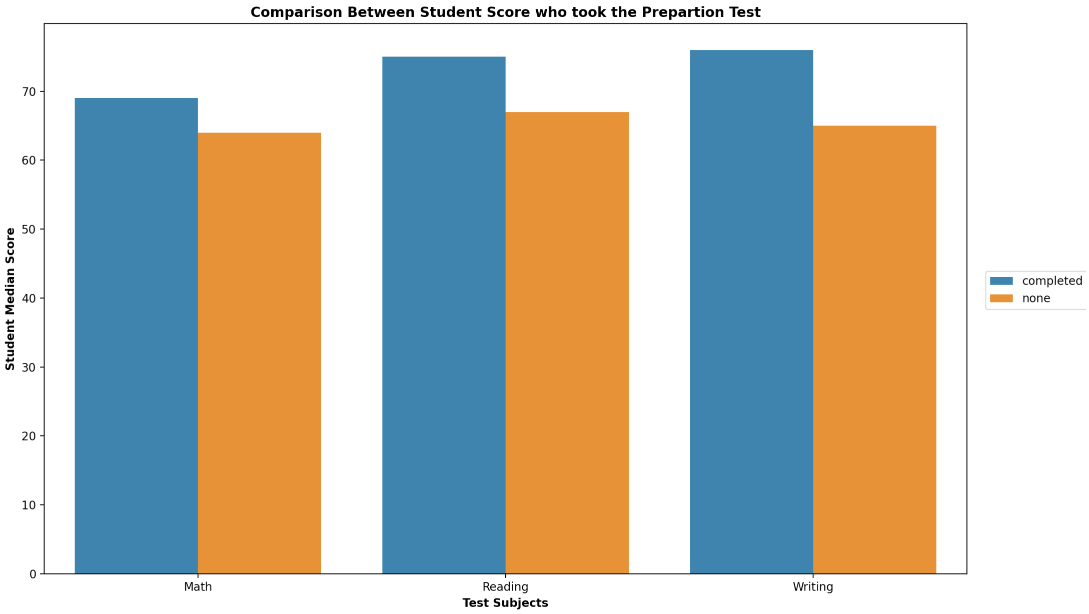






Parent education affects the student's performance in the test. The higher the academic level of the parents, the higher the score. However, there's slight drop from parents who didn't finish high school and parents who did. Surprisingly, **the students whose parents dropped out of High school scored higher than students whose parents finished high school. However, it is not significantly different.**

The Effect of Taking Preparation Test on Student's Performance



-

- To prevent outliers from affecting the mean, I choose to use the median to compare the score.

- We can see from the graph above that **there is a significant difference in scores between the students who completed the preparation test**, and the student who did not.

- The student who did complete the preparation test performed better in all subjects.

OBSERVATIONS FROM HYPOTHESIS TESTING

Note:-

This dataset is a sample of scores from 1000 students and not the population.

Type of Hypothesis Testing Used:-

Since we are looking if there is a difference, we will be using a **Two-Tailed Test**.

First Hypothesis

A. Math Scores Between Students Whose Parental Level of Education is Bachelor's and Master's Degree

Null Hypothesis: The Math scores of the students whose parental level of education is Bachelor's Degree **are the same** as Master's Degree.

Alternate Hypothesis: The Math scores of the students whose parental level of education is a Bachelor's Degree **are different** Master's Degree.

RESULT:

- We accept the Null Hypothesis
- Meaning that The Math scores of the students whose parental level of education is a Bachelor's Degree **are the same as** the students whose parental level of education is a Master's Degree.

Second Hypothesis

B. Math Scores Between Students Whose Parental Level of Education is Some High School and Master's Degree

Null Hypothesis: The Math scores of the students whose parental level of education is Some High School **is the same** as Master's Degree.

Alternate Hypothesis: The Math scores of the students whose parental level of education is Some High School **is different** Master's Degree.

RESULT:

- We Reject the Null Hypothesis
- Meaning that The Math scores of the students whose parental level of education is Some High School **is different** from the students whose parental level of education is a Master's Degree.

OBSERVATIONS FROM Multiple Linear Regression Model

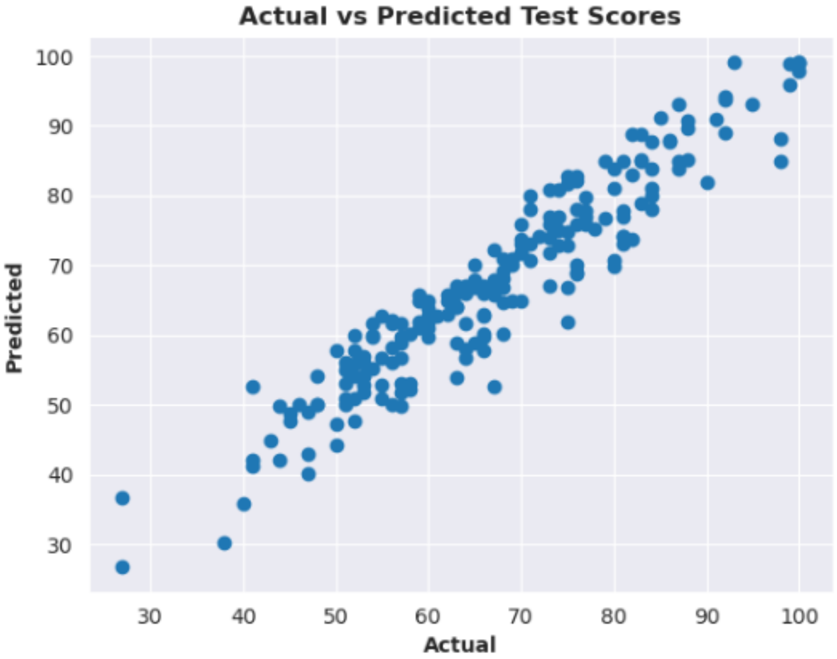
We are going to attempt to create a Multiple Linear Regression Model to predict the **writing score** of students.

The **Reading Score and Parent Education** will be used as a predictor.

From the Analysis above, we found out that the parental level of education has an effect on the student's score.

The Parent_Education column already has a hierarchy starting from "Some High School" to "Master's Degree".

Based on the R-squared score of **0.90** proves that the **linear Regression model is a good model**.


Actual vs Predicted Test Scores
Predicted
Actual