TO UNDERSTAND STUDENTS PERFORMANCE USING DESCRIPTIVE STATISTICS

STUTERN DATA SCIENCE TASK

Fagbolagun Kehinde

Table of Contents

Abstract

This work aims to show how the economic, social and personal status of students influences their performance in school. Data containing/representing the performance of 1,000 students in three subject areas: Maths, Reading and Writing with other attributes such as their gender, race, test preparation, parental level of education etc. was analysed. After statistical analysis, it was shown that the above attributes bordering on social, economic and personal status of the students directly influenced their performance in school.

Introduction

There are several ways of measuring students' ability, one method is the use of test scores which is a direct measure of their academic ability. How students perform academically depends on several factors of which their level of preparation is a huge factor.

This work explores several factors like the students' race/ethnicity, their level of preparation, parental level of education, the category of lunch taken and their gender to find out how these factors impact/influence how students perform.

Descriptive statistics involves summarizing and organizing data so they can be easily understood. Univariate analysis is the simplest form of data analysis where the data being analyzed contains only one variable, it is used to describe data and find patterns that exist within it.

Multivariate analysis is the analysis of three or more variables, it shows the relationship or correlation between variables.

Data Preparation

The data set for this task is sourced from a Kaggle challenge and is contained in a csv file. The data set contains about 1,000 observations (students' results) and 8 attributes which are gender, race/ethnicity, parental level of education, lunch test, test preparation course, math, reading and writing scores respectively.

Analytical Tool

Python programming language was used in combination with some python libraries like Pandas, Seaborn and Matplotlib to prepare and explore the data. The computational environment used was Jupyter Notebook.

Data Preprocessing

The data was read into the notebook and the first five rows were visualized and checked for missing values and errors with none found.

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
0	female	group B	bachelor's degree	standard	none	72	72	74
1	female	group C	some college	standard	completed	69	90	88
2	female	group B	master's degree	standard	none	90	95	93
3	male	group A	associate's degree	free/reduced	none	47	57	44
4	male	group C	some college	standard	none	76	78	75

```
[ ]  df.isnull().values.any()
```

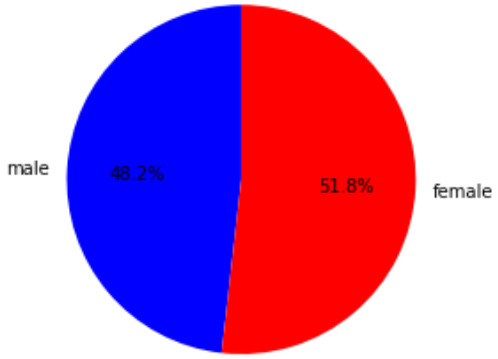
```
False
```

Univariate Analysis

This was carried out using pie charts to visualize the data.

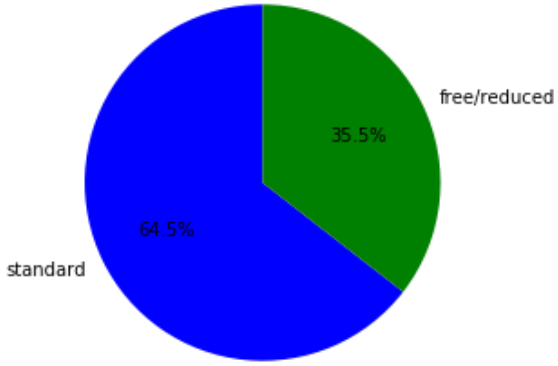
Gender

Out of a total student population of 1,000, 518 were female while 482 were male representing 48.2% and 51.8% respectively.



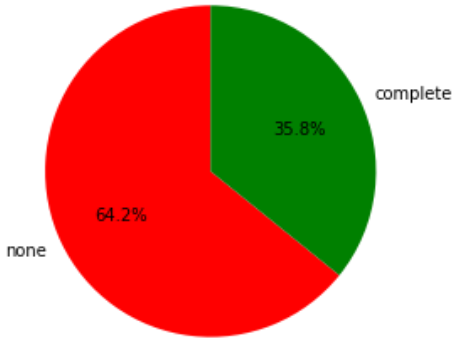
Lunch

Of a total of 1,000 students, 645 had the standard lunch while 355 had the free/reduced lunch.



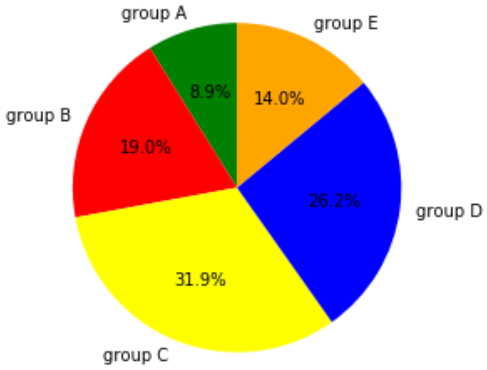
Test preparation course

358 students completed the course while 642 did not.



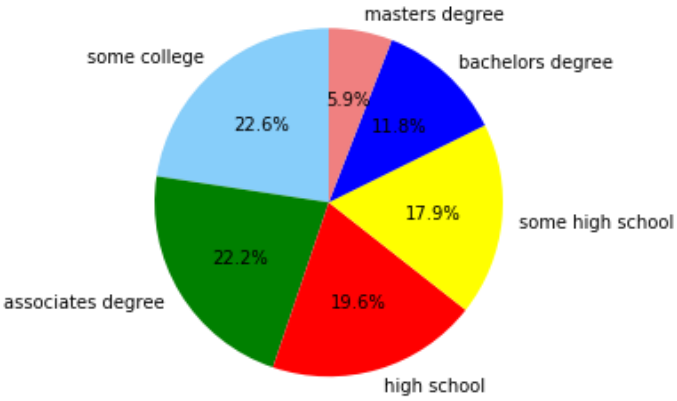
Race/Ethnicity

Of a total of 1,000 students, group A had 89 students, group B had 190 students, group C had 319 students, group D had 262 students while group E had140 students.



Parental level of education

59 students have a parent with masters' degree, 118 have parents with bachelors' degree, 222 have parents with associates' degree etc.



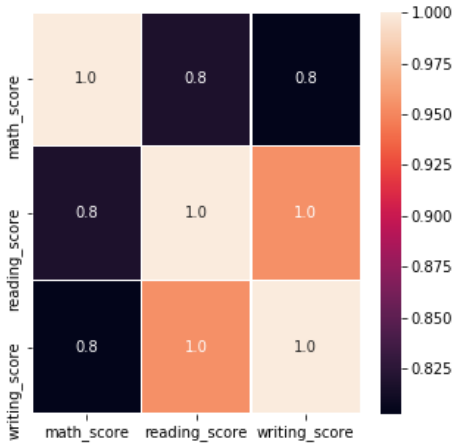
Measure of Central Tendency

The Mean, Median, Mode and Standard Deviation were analysed for the maths, reading and writing scores.

	math score	reading score	writing score
count	1000.00000	1000.000000	1000.000000
mean	66.08900	69.169000	68.054000
std	15.16308	14.600192	15.195657
min	0.00000	17.000000	10.000000
25%	57.00000	59.000000	57.750000
50%	66.00000	70.000000	69.000000
75%	77.00000	79.000000	79.000000
max	100.00000	100.000000	100.000000

Multivariate Analysis

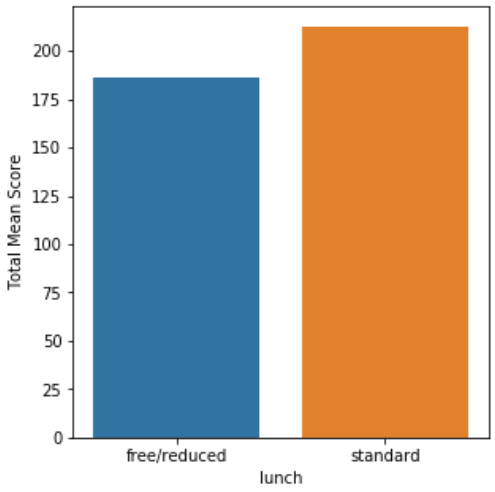
Correlation Matrix is a table that shows the degree of relationship between sets of variables, with a range between 0 and 1, of which 1 is the value for the most correlated variable. This was used to analyse the maths, reading and writing scores.



To check the influence of other attributes on students' performance, the mean of their scores for the three subjects was summed to get a total mean score which was the compared with other attributes.

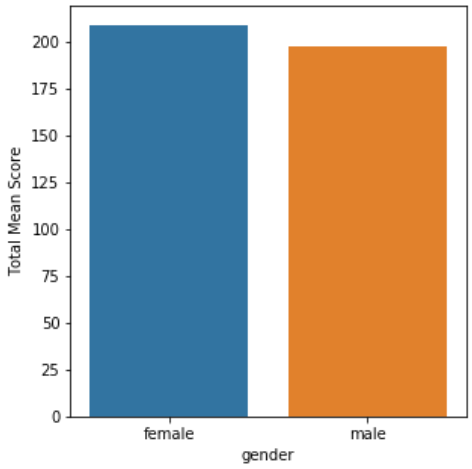
Total mean score against the Type of Lunch taken

Those with standard lunch performed better than those with free or reduced lunch.



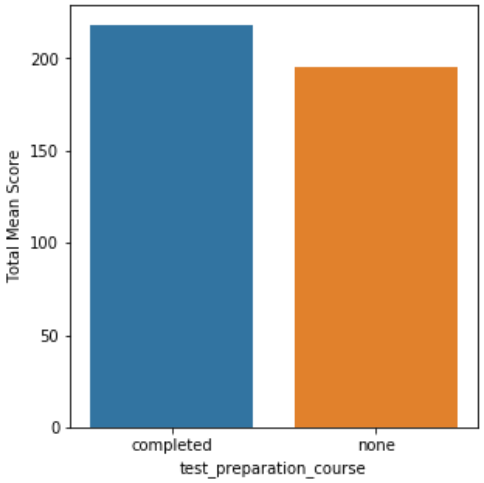
Total mean score against Gender

This shows that females performed better than males.



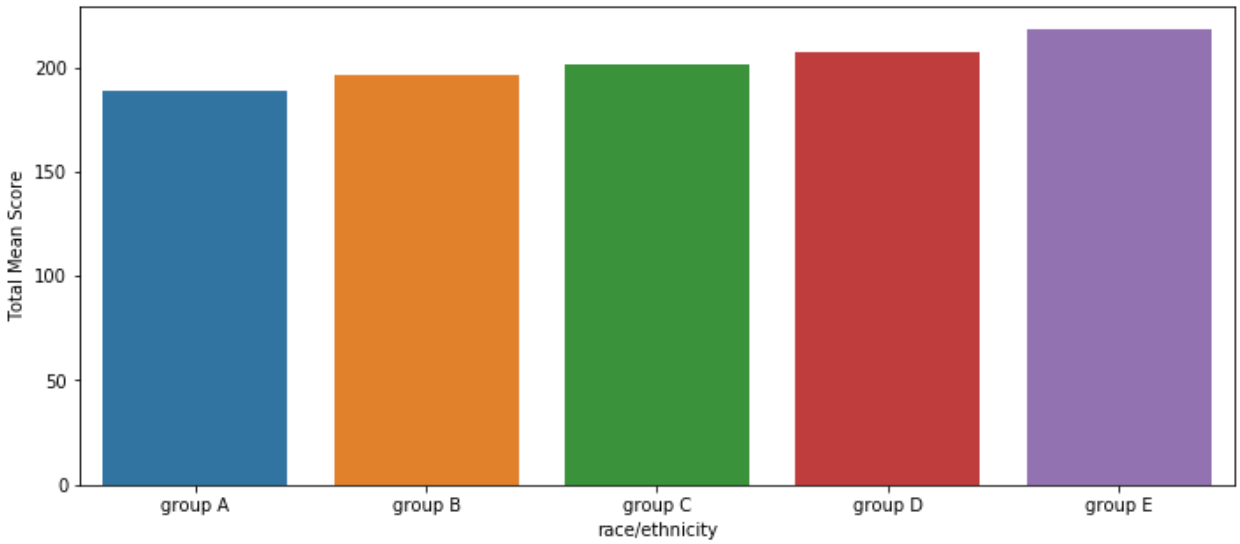
Total mean score against Test preparation course

Those that competed the course did better than those who didn't take the course.



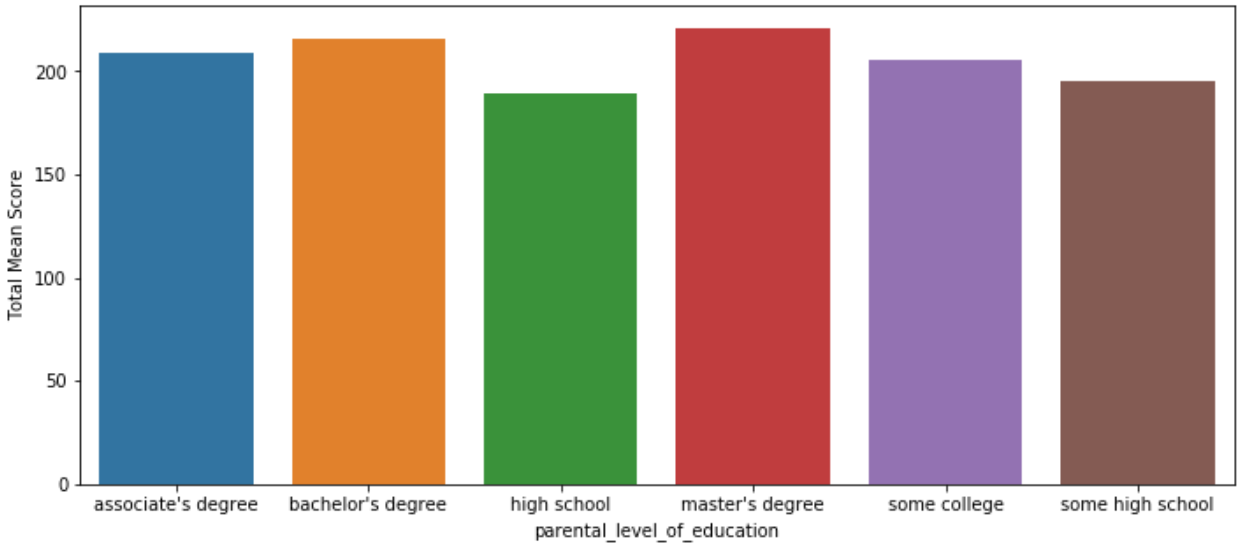
Total mean score against Race/Ethnicity

This shows that group E were the best performers with decreasing order of performance from group D down to group A being the worst performers.



Total mean score against Parental level of education

This shows that students whose parents have a Masters degree were the best performers while those with parents who have only high school education were the worst performers. The general trend shows that the more educated a parent is, the better their children perform.



Conclusion and Recommendation

At the end of the analysis, it can be concluded that several factors influence how students perform. This factors should considered with importance as they can be exploited to the benefits of the students to help increase their performance in school and also mitigate against poor performance.

Link to colab notebook: https://drive.google.com/open?id=1GnEiEEpL0Afp7ofxUWqF1Da6YF8sKA9q