**09-26-2024**

**Principles of Data Science**

**Assignment -1**

**Achyuth Pothuganti- 16355349**

**Question 2:** I have conducted various analytical operations on the student performance dataset and presented the results in the form of visualizations.

* **Raw Data:**

A screenshot of a computer

Description automatically generated

I performed data cleaning by first removing rows with less than five non-null values and filling missing values in the 'math score,' 'reading score,' and 'writing score' columns with their respective column means. Additionally, I standardized text data in categorical columns such as 'gender,' 'race/ethnicity,' 'parental level of education,' 'lunch,' and 'test preparation course' by converting them to lowercase and trimming whitespace for consistency.

* **Clean Data:**  The table below presents the cleaned dataset.

A screenshot of a computer

Description automatically generated

* **Visualization:**

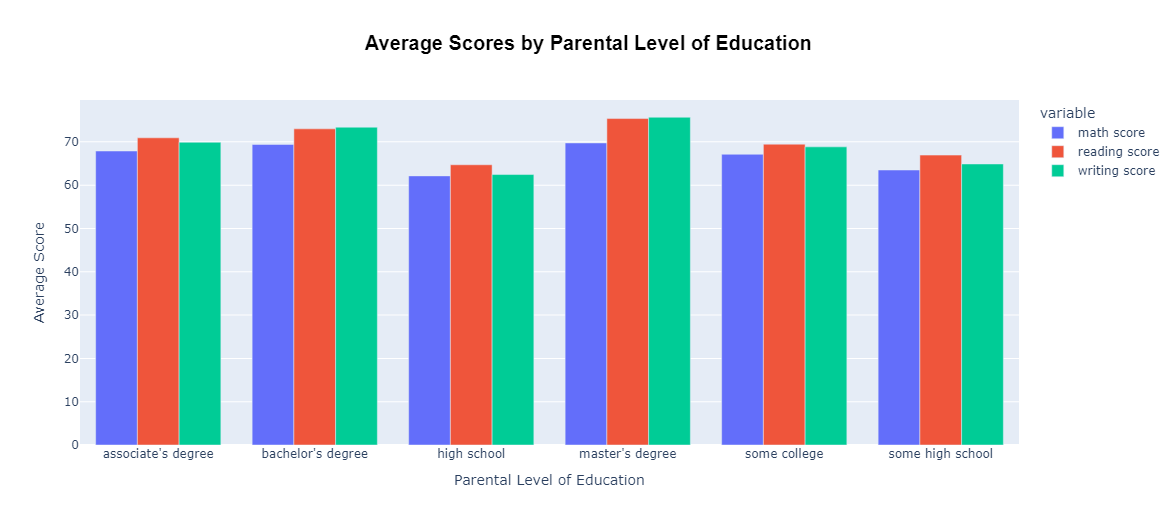
1. **Box plot of Math Score by Gender:**

**A graph showing a box plot

Description automatically generated**

The box plot illustrates that female students generally have higher math scores than male students, with a median of around 66 for females compared to 60 for males. The interquartile range (IQR) for females (55 to 80) is wider than that for males (45 to 75), indicating more variability in male scores. Additionally, males have more outliers, suggesting a broader range of performance. Overall, the data suggests a consistent trend of higher math achievement among females compared to their male counterparts.

1. **Average Scores by Parental Level of Education:**

****

The bar plot illustrates that students' average scores in math, reading, and writing improve with higher parental education levels. Those with parents holding a master's degree achieve the highest scores across all subjects, while scores decline for those whose parents have only completed high school or have some college education. Reading scores generally surpass math and writing scores, suggesting a greater emphasis on reading skills. Overall, the data highlights the significant impact of parental education on student academic performance.

1. **Scatter Plot of Reading vs. Writing Scores:**

**A graph showing a graph of reading and writing

Description automatically generated**

The scatter plot shows a positive correlation between reading and writing scores, indicating that higher reading scores are associated with higher writing scores. Female students who completed a test preparation course tend to achieve higher writing scores, while male students also benefit from the course, though with more variability. Overall, the plot highlights the importance of test preparation in enhancing writing skills and suggests that proficiency in reading contributes to writing ability.

1. **Correlation Matrix Heatmap:**

**A screenshot of a graph

Description automatically generated**

The correlation matrix heatmap illustrates strong positive relationships between math, reading, and writing scores. The highest correlation is between reading and writing scores, almost perfect at **0.9946,** indicating that students who excel in one of these subjects tend to excel in the other. Math scores also have strong but slightly weaker correlations with reading and writing, suggesting that while related, math skills are less closely tied to reading and writing abilities. Overall, the heatmap reveals that students performing well in one subject generally perform well in the others.

1. **Count of Students by Race/Ethnicity and Lunch Type:**

A graph of students by race ethnicity and lunch type

Description automatically generated

The bar chart highlights the distribution of students from different racial/ethnic groups (a to e) receiving either standard or free/reduced lunch. Group **c** has the highest number of students overall, with 110 receiving standard lunch and 90 receiving free/reduced lunch. Group **a** has the fewest students, with most receiving standard lunch. Groups **b, d**, and **e** have similar numbers of students receiving standard lunch, but group **b** leads slightly. The chart reveals varying proportions of students receiving free/reduced lunch across the groups, with group **c** having the highest proportion at 45%.