09-26-2024

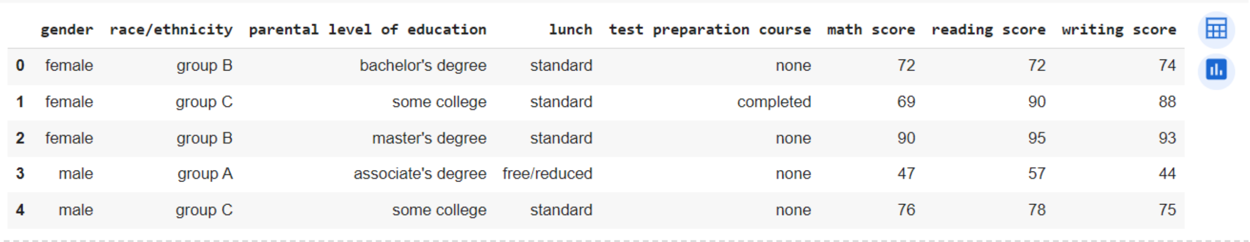
Principles of Data Science

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.

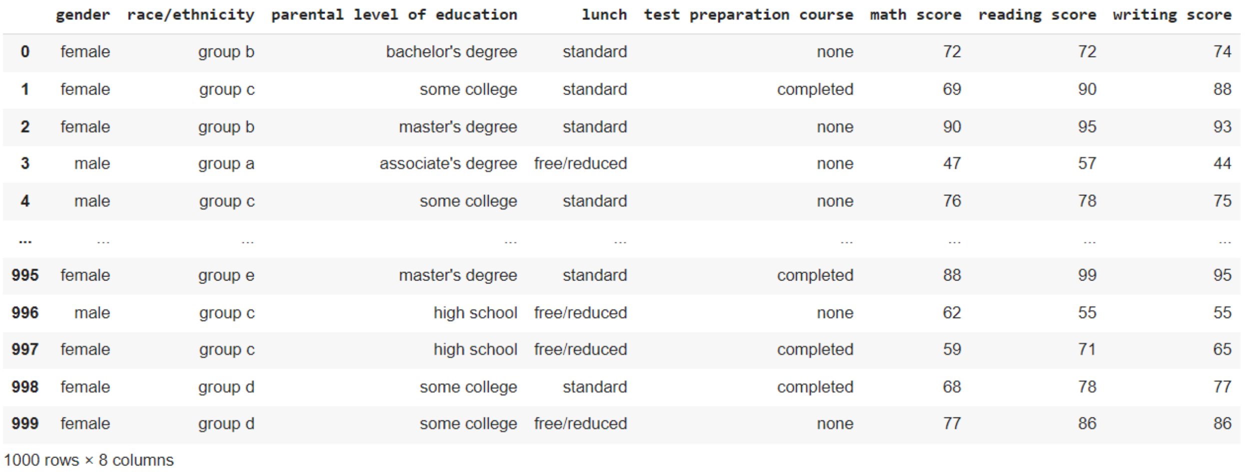
- **Raw Data:**



	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

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.

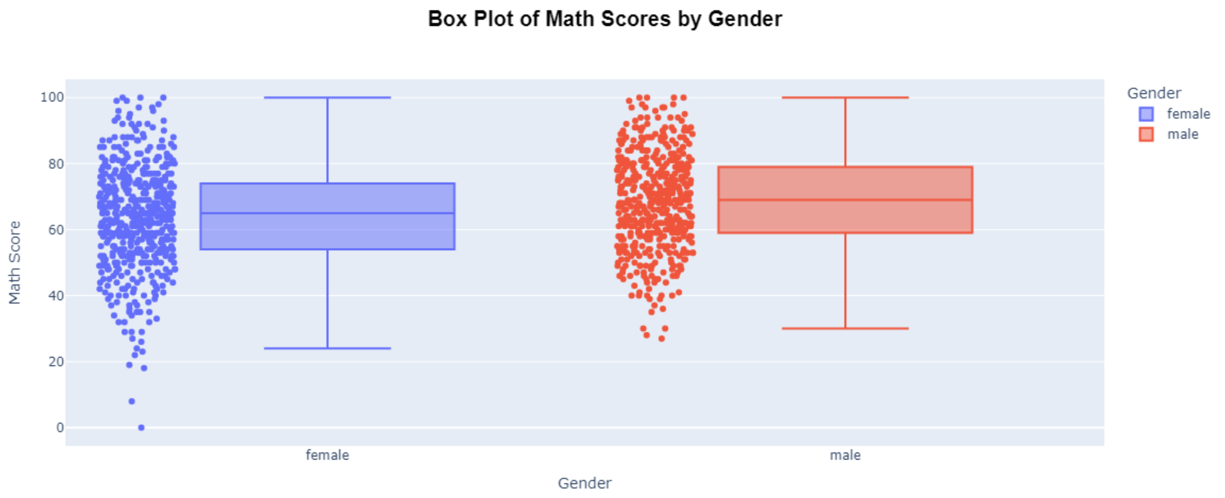


	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
0	female	group b	bachelor's degree	standard	none	72	72	74
1	female	group c	some college	standard	completed	69	90	88
2	female	group b	master's degree	standard	none	90	95	93
3	male	group a	associate's degree	free/reduced	none	47	57	44
4	male	group c	some college	standard	none	76	78	75
...	...	...	...	...	...	...	...	...
995	female	group e	master's degree	standard	completed	88	99	95
996	male	group c	high school	free/reduced	none	62	55	55
997	female	group c	high school	free/reduced	completed	59	71	65
998	female	group d	some college	standard	completed	68	78	77
999	female	group d	some college	free/reduced	none	77	86	86

1000 rows × 8 columns

➢ Visualization:

1. Box plot of Math Score by Gender:



Purpose:

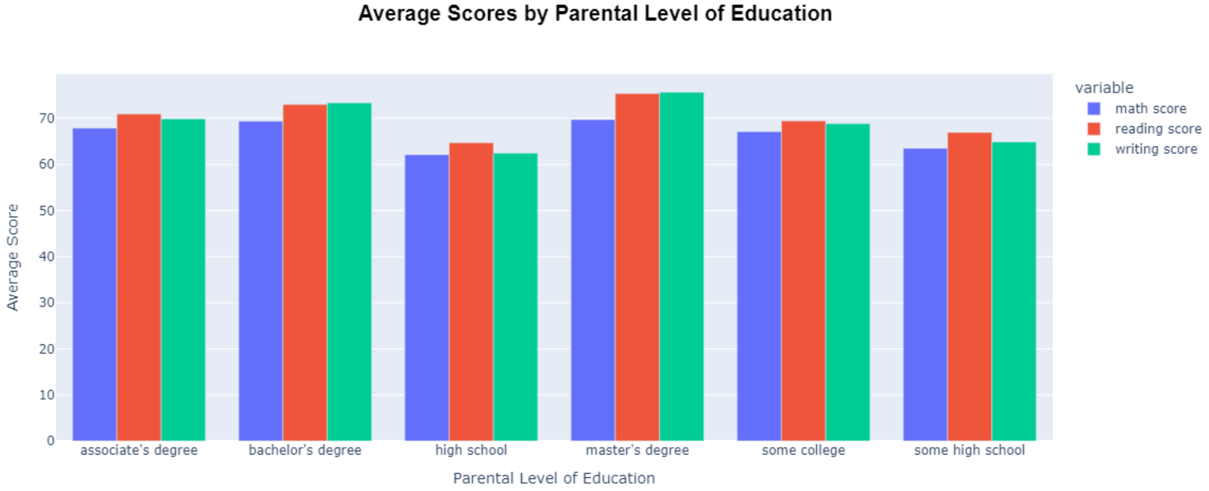
This visualization shows the frequency distribution of scores across three subjects.

Easier Analysis:

Identifies where students perform strongest or weakest. Reveals if certain subjects have a skewed distribution (e.g., students performing better in reading/writing than math). Allows educators to focus on subjects where students tend to struggle.

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.

2. Average Scores by Parental Level of Education:



Purpose:

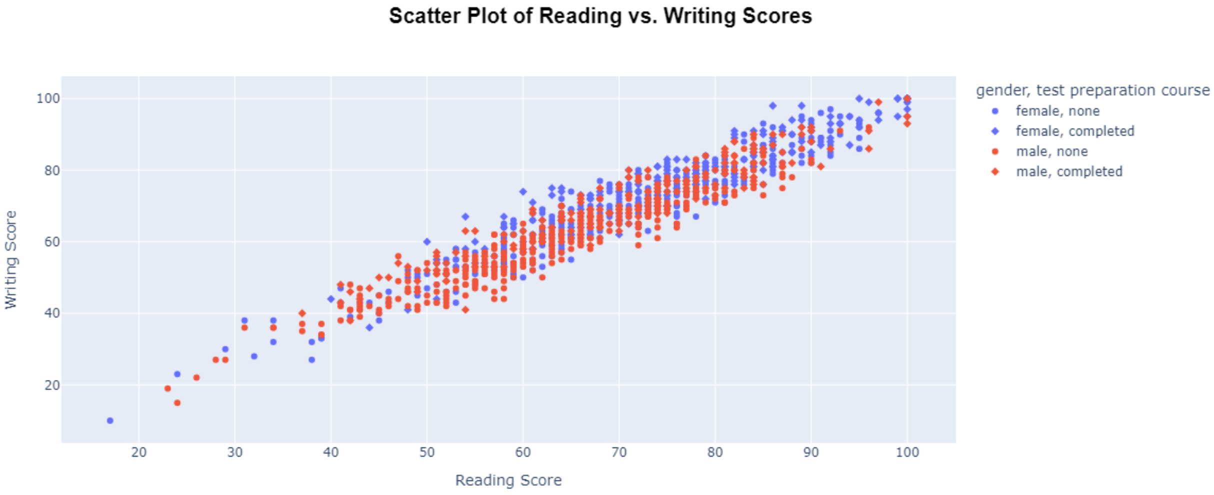
This compares the average scores between male and female students.

Easier Analysis:

- Highlights gender performance differences across all subjects.
- Simplifies understanding whether gender influences student performance.
- Useful for addressing performance gaps between genders if present.

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.

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



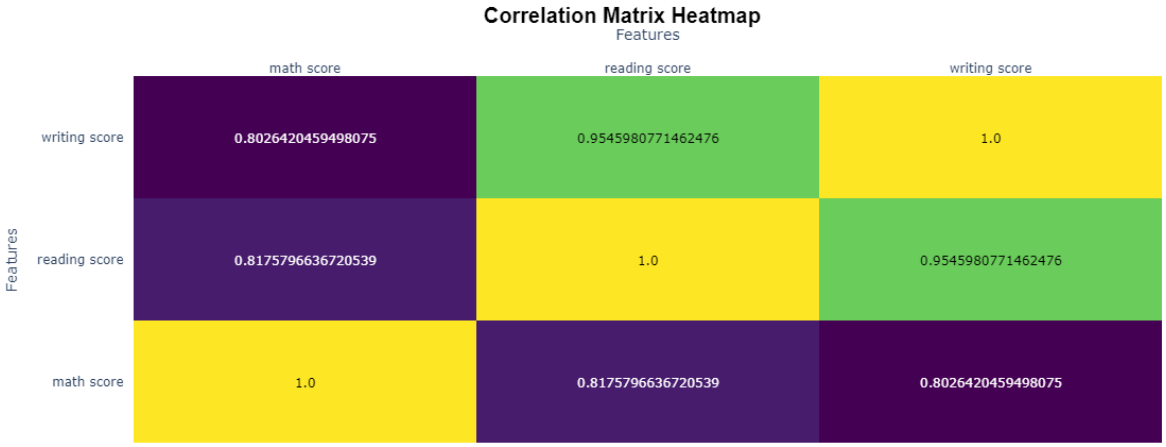
Purpose: Shows how students' scores vary based on their parents' educational background.

Easier Analysis:

- Reveals patterns in student performance linked to parental education.
- Highlights the influence of a parent's education level on student outcomes.
- Useful for identifying groups that may need additional support or resources.

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.

4. Correlation Matrix Heatmap:



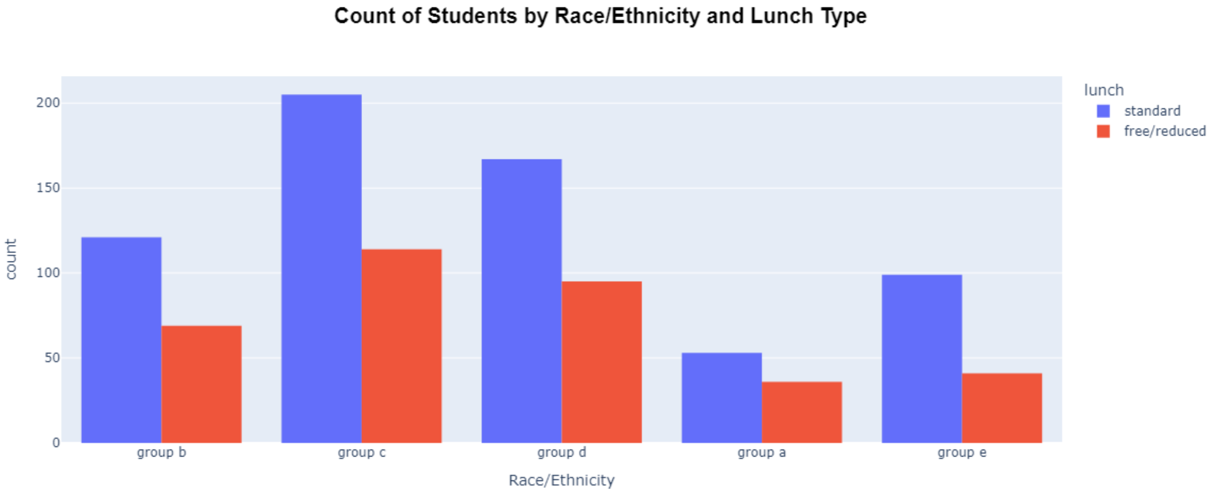
Purpose: Visualizes the distribution of scores based on whether students completed a test preparation course.

Easier Analysis:

- Provides insights into how test preparation affects performance.
- Reveals whether students who completed the course perform significantly better.
- Useful for promoting or redesigning test preparation programs.

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.

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



Purpose: Displays the correlation between the three score variables (math, reading, writing).

Easier Analysis:

- Identifies how scores in one subject relate to another.
- Helps in understanding whether a student's performance in one subject influences others.
- Useful for creating strategies that address subject-specific weaknesses while leveraging strengths in other subjects.

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%.