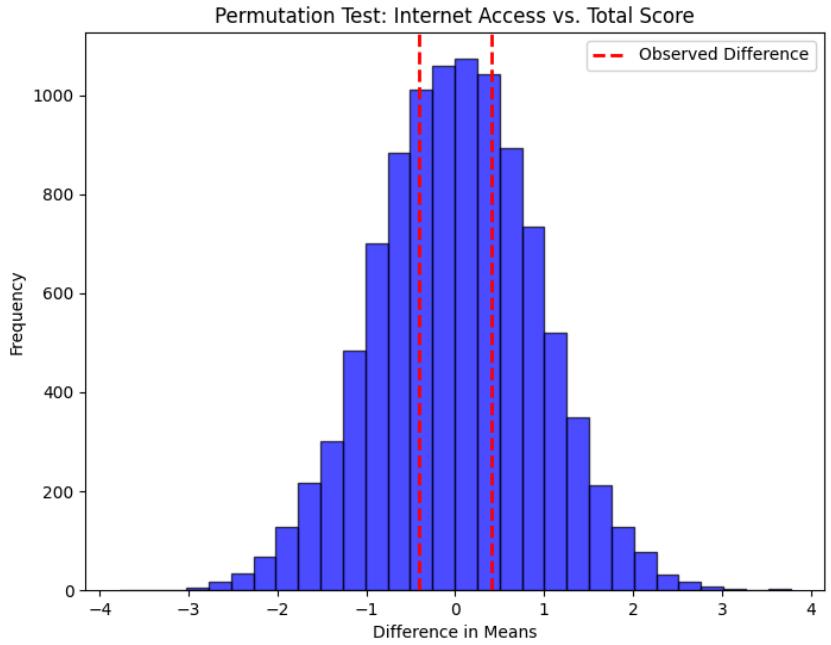
- The Kaggle link for the dataset
  - https://www.kaggle.com/datasets/mahmoudelhemaly/students-grading-dataset
  - Title: Student Performance & Behavior Dataset
  - Attributes:
    - Student_ID: Unique identifier for each student.
    - First_Name: Student's first name.
    - Last_Name: Student's last name.
    - Email: Contact email (can be anonymized).
    - Gender: Male, Female, Other.
    - Age: The age of the student.
    - Department: Student's department (e.g., CS, Engineering, Business).
    - Attendance (%): Attendance percentage (0-100%).
    - Midterm_Score: Midterm exam score (out of 100).
    - Final_Score: Final exam score (out of 100).
    - Assignments_Avg: Average score of all assignments (out of 100).
    - Quizzes_Avg: Average quiz scores (out of 100).
    - Participation_Score: Score based on class participation (0-10).
    - Projects_Score: Project evaluation score (out of 100).
    - Total_Score: Weighted sum of all grades.
    - Grade: Letter grade (A, B, C, D, F).
    - Study_Hours_per_Week: Average study hours per week.
    - Extracurricular_Activities: Whether the student participates in extracurriculars (Yes/No).
    - Internet_Access_at_Home: Does the student have access to the internet at home? (Yes/No).
    - Parent_Education_Level: Highest education level of parents (None, High School, Bachelor's, Master's, PhD).
    - Family_Income_Level: Low, Medium, High.
    - Stress_Level (1-10): Self-reported stress level (1: Low, 10: High).

        - Sleep_Hours_per_Night: Average hours of sleep per night.
- Establish a null and alternative hypothesis, and conduct either a permutation test or z-test
    - Null Hypothesis: There is no significant difference in total scores between students who have internet access at home and those who do not.
    - Alternative Hypothesis: There is a significant difference in total scores between students who have internet access at home and those who do not.
    - ***Conduct a two-proportion z-test to compare the approval rates between applicants with and without prior defaults***

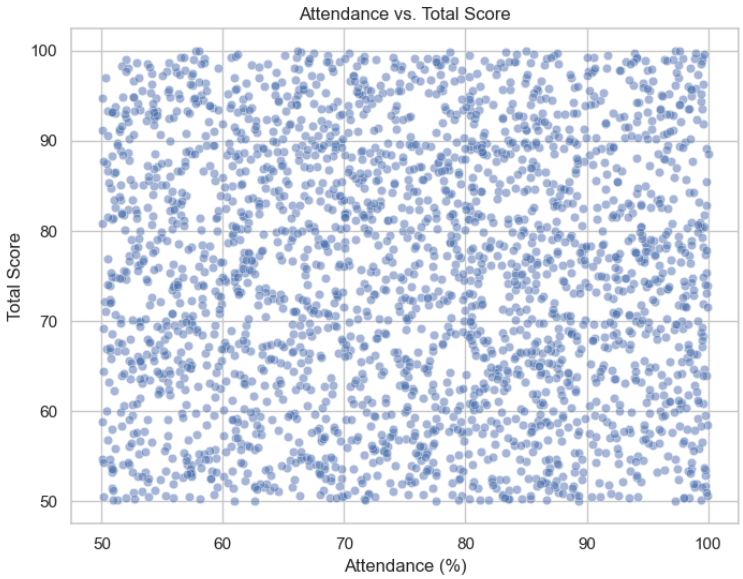



```
C:\Users\msraa\PycharmProjects\pythonProject\venv\Scripts\python.exe C:\Users\msraa\PycharmProjects\pythonProject\capstone2.py
Observed Difference in Means: -0.4013749904465982
P-value: 0.6615
Result: Fail to reject the null hypothesis (no significant difference).

Process finished with exit code 0
```

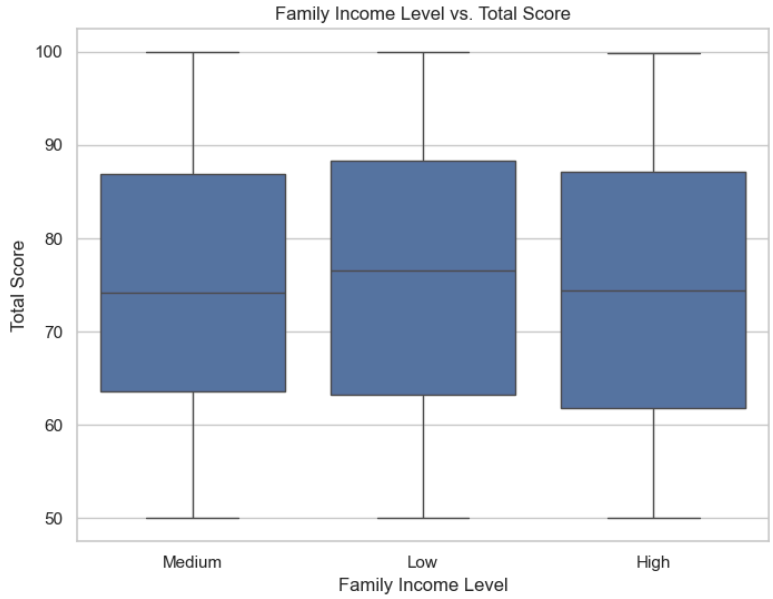
    - Observed Difference in Means: -0.4013749904465982
    - P-value: 0.663
    - Result: Fail to reject the null hypothesis (no significant difference).

    - We can conclude that there is NO significant difference between total scores of students who have internet access at home or do not have internet access at home
    - The image shows a permutation test comparing total scores between students with and without internet access at home. The histogram displays the distribution of differences in means that could occur by random chance, with the observed difference marked by a vertical dashed line. The observed difference in means is -0.4013749004465982, which falls well within the randomized distribution, resulting in a high p-value of 0.6615. This p-value far exceeds the conventional significance threshold of 0.05, leading to the conclusion that we fail to reject the null hypothesis. The result builds a compelling argument that having internet access at home does not significantly impact student academic performance in this dataset, which is somewhat surprising given typical assumptions about digital access and educational advantages. This finding adds to the pattern seen in the following visualizations where expected factors of academic success such as study hours, attendance, and family income showed minimal correlation with performance. This shows that what may have previously been deciding factors in educational success may be less influential than previously believed to be.
- Provide three visualizations that each can make an argument/claim on their own, or collectively make an argument/claim together
    - Scatter Plot: Visually displays the correlation between attendance and total score



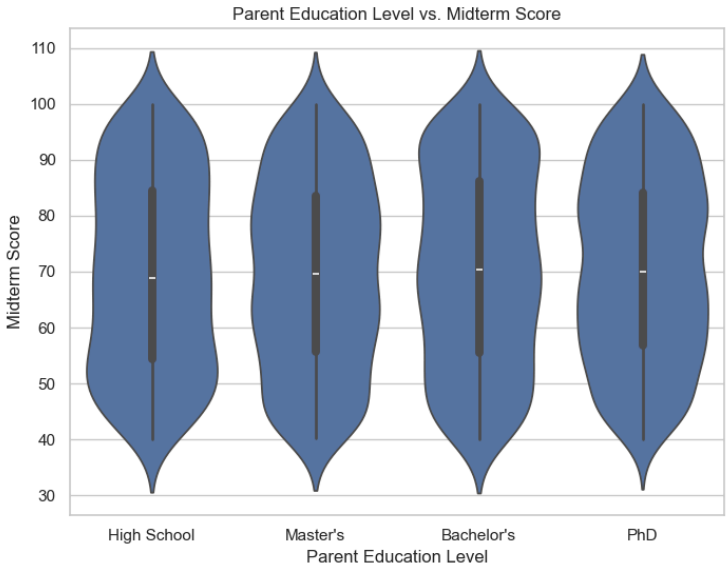
- This is a scatter plot titled "Attendance vs. Total Score" that plots attendance percentage (x-axis, ranging from 50% to 100%) against total score (y-axis, ranging from 50 to 100). The data points appear randomly distributed across the plot with no visible pattern. This shows there is little to no correlation between student attendance rates and their total scores. This goes against common assumption that higher attendance would lead to better academic performance.

- Box Plot: Visually displays the correlation between family income level and total score



- This is a box plot showing "Family Income Level vs. Total Score" with three categories: Medium, Low, and High income levels. The box plots display remarkably similar distributions across all three income groups, with median scores (horizontal lines in the boxes) being very close to each other at around 75 - 78 points. The interquartile ranges (boxes) and overall ranges (whiskers) also look very similar, extending from about 50 to 100 points for all income groups. This visualization strongly suggests that family income level has minimal impact on student academic performance in this dataset. This is against the common assumption about socioeconomic status and educational outcomes.

- Violin Plot: Visually displays the correlation between parent education level and midterm score

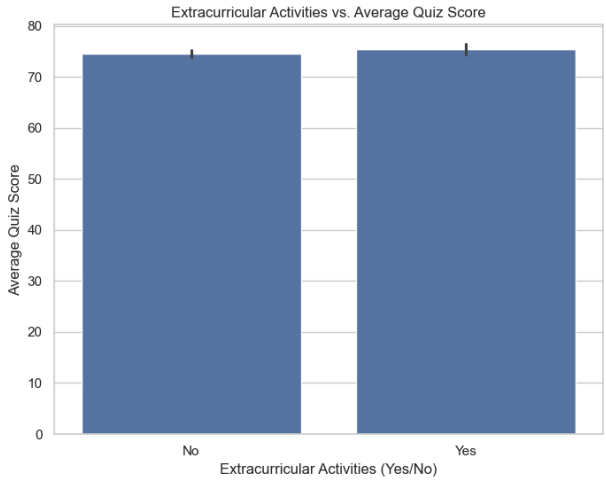


- This is a violin plot titled "Parent Education Level vs. Midterm Score" which displays the distribution of midterm scores across four different parent education levels: High School, Master's, Bachelor's, and PhD. The violin shapes show that scores generally range from about 30 to 110 across all education levels. The distributions appear fairly similar across the four categories, with slightly more density in the middle range (around 70-80) for all groups, suggesting that parent education level might not have a strong effect on midterm scores.

- Regression plot: Visually displays the correlation between study hours per week and total score

- This is a regression plot showing the relationship between "Study Hours per Week" (x-axis, ranging from about 5 to 30 hours) and "Total Score" (y-axis, ranging from about 50 to 100). The plot displays a dense cloud of data points with no clear linear pattern, suggesting a weak/non-existent correlation between study hours and total scores. The points are fairly evenly distributed across the graph, indicating that students who study more hours do not necessarily achieve higher scores, contrary to what might be expected.

- Bar Chart: Visually displays the correlation between study hours per week and total score



- The bar graph titled "Extracurricular Activities vs. Average Quiz Score" compares the academic performance of students based on their participation in extracurricular activities. The x-axis divides students into two categories: those who do not participate in extracurricular activities ("No") and those who do ("Yes"). The y-axis represents the average quiz scores, ranging from 0 to 80. Both quiz averages appear to score approximately 74-75 points on average. This suggests that participation in extracurricular activities has negligible impact on students' quiz performance.

Conclusion

Based on the comprehensive analysis of this educational dataset, a striking pattern emerges across all visualizations and statistical tests: traditional predictors of academic success appear to have remarkably little influence on student performance. The permutation test comparing internet access groups yielded a high p-value (0.6615), demonstrating no significant difference in academic outcomes. Similarly, the visualizations revealed that parent education level, study hours per week, attendance rates, family income levels, and participation in extracurricular activities all showed minimal to no correlation with student scores. This consistent pattern challenges conventional educational wisdom that often emphasizes socioeconomic factors, study habits, and engagement metrics as key determinants of academic success. The data

suggests that either other unmeasured variables are driving academic performance in this population, or that the relationship between traditional predictors and outcomes is more complex than typically assumed. The graphs reinforce the emerging narrative that traditional assumptions about factors influencing academic achievement may need reconsideration. These findings invite educators and researchers to reconsider fundamental assumptions about what truly influences student achievement and highlight the importance of looking beyond conventional factors when developing interventions to improve educational outcomes.