Student Performance Factors

February 5, 2025

1 Role of Family on Student Performance

1.1 Purpose and Questions

Throughout our society today, exams have become woven into almost every part of our life. Almost every child in the United States spends at least 13 years taking exams and having their performance evaluated to determine their future pursuits. However, performance on exams can vary heavily. Many people blame various factors including gender, race, parental involvement, money, and more. Despite various explanations the debate still continues. Throughout this paper, I hope to provide some understanding to the reasons behind student performance on exams and to discern whether correlation is equal to causation or if there is some deeper reason.

1.1.1 Questions to Answer

- Does family income play a significant role in the performance of students?
- Does tutoring, typically only afforded by the well-off families, increase performance?
- What role does the family of a student have in their performance on exams?

I will be using a dataset from Kaggle with close to 20 features. You can access the dataset with this link https://www.kaggle.com/datasets/lainguyn123/student-performance-factors/data. I will be focusing on the role of family in student performance.

[4]:
```python
#first, let's load the dataset, we'll need pandas
import pandas as pd

df = pd.read_csv('StudentPerformanceFactors.csv')
```

[5]:
```python
df.shape #rows x columns
```

[5]:
```
(6607, 20)
```

1.2 Preprocessing

Before we start visualizing data, we should evaluate the dataset and determine whether the dataset needs any cleaning done.

[7]:
```python
print(df.isnull().sum())    #check for nulls in every column
print(df.isnull().sum().sum())
```

```
Hours_Studied                  0
Attendance                     0
Parental_Involvement           0
Access_to_Resources            0
Extracurricular_Activities     0
Sleep_Hours                    0
Previous_Scores                0
Motivation_Level               0
Internet_Access                0
Tutoring_Sessions              0
Family_Income                  0
Teacher_Quality               78
School_Type                    0
Peer_Influence                 0
Physical_Activity              0
Learning_Disabilities          0
Parental_Education_Level      90
Distance_from_Home            67
Gender                         0
Exam_Score                     0
dtype: int64
235
```

[8]:
```
df = df.copy()

# Fill categorical columns with their most frequent value (mode)
categorical_cols = ['Teacher_Quality', 'Parental_Education_Level',␣
 ↪'Distance_from_Home']
df[categorical_cols] = df[categorical_cols].apply(lambda col: col.fillna(col.
 ↪mode()[0]))

# Check if nulls are removed
print(df.isnull().sum())
```

```
Hours_Studied                 0
Attendance                    0
Parental_Involvement          0
Access_to_Resources           0
Extracurricular_Activities    0
Sleep_Hours                   0
Previous_Scores               0
Motivation_Level              0
Internet_Access               0
Tutoring_Sessions             0
Family_Income                 0
Teacher_Quality               0
School_Type                   0
Peer_Influence                0
```

```
Physical_Activity             0
Learning_Disabilities         0
Parental_Education_Level      0
Distance_from_Home            0
Gender                        0
Exam_Score                    0
dtype: int64
```

[9]:

```
print(df.duplicated().sum())  # Count duplicate rows
```

```
0
```

[10]:

```
bins = [0, 60, 80, 100]  # Chose Ranges: 0-60, 70-80, 81-100
labels = ['Low', 'Average', 'High']

# Used feature engineering to bin exam scores
df['Score_Category'] = pd.cut(df['Exam_Score'], bins=bins, labels=labels,␣
  ↪right=True)
```

So far, we've checked for any null values and checked for duplicates. I chose to replace missing values with the mode because the amount of nulls weren't considerable enough for a dataset with almost 7000 datapoints. At this point, I am comfortable with our preprocessing steps because it looks like this dataset is ready to go. I chose to create a binned feature called Score Category so that it would be easier for me to analyze the overll performance. I personally don't believe the difference between a 65 and a 67 is necessarily significant enough to evaluate performance.

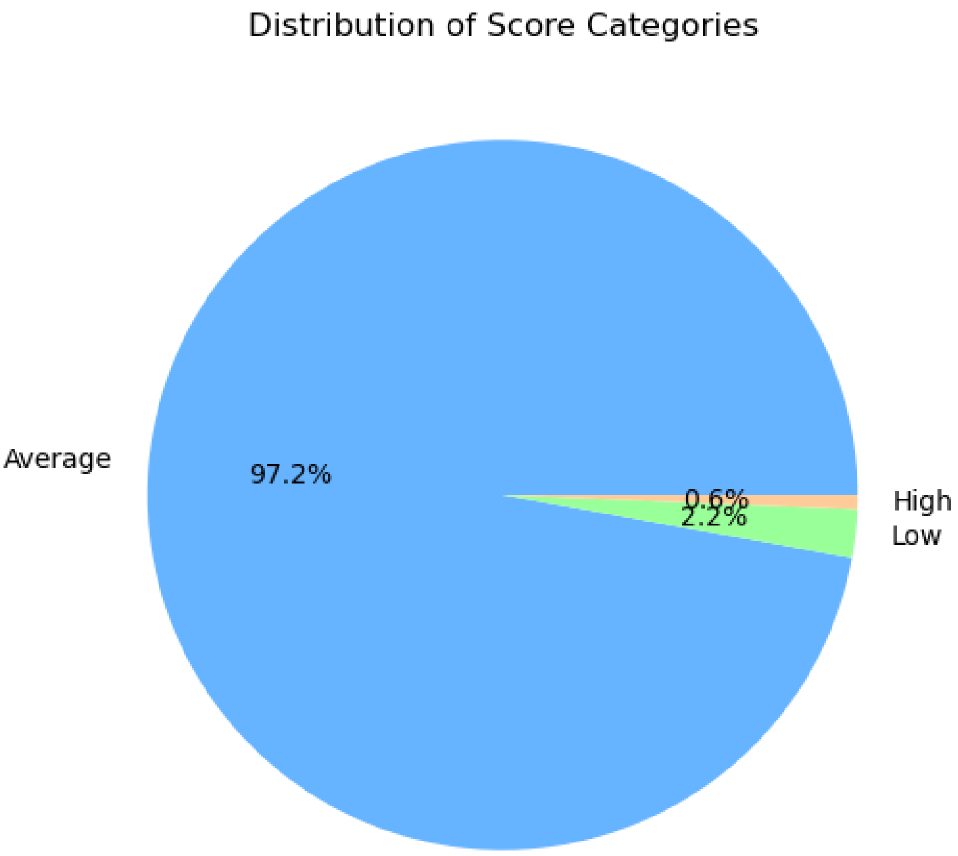
1.3 Visualizations

Before we move on, I believe it is important to understand our data's limitations, specifically when it comes to distribution between category. I believe that understanding any differences in distribution is valuable which is why I will be making a pie chart so we can see if there is an equal distribution between each score category.

[14]:

```
import seaborn as sns
import matplotlib.pyplot as plt
score_category_counts = df["Score_Category"].value_counts()

plt.figure(figsize=(10, 6))
plt.pie(score_category_counts, labels=score_category_counts.index, autopct='%1.
  ↪1f%%',
        startangle=360, colors=["#66b3ff", "#99ff99", "#ffcc99"])
plt.title("Distribution of Score Categories")
plt.show()
```

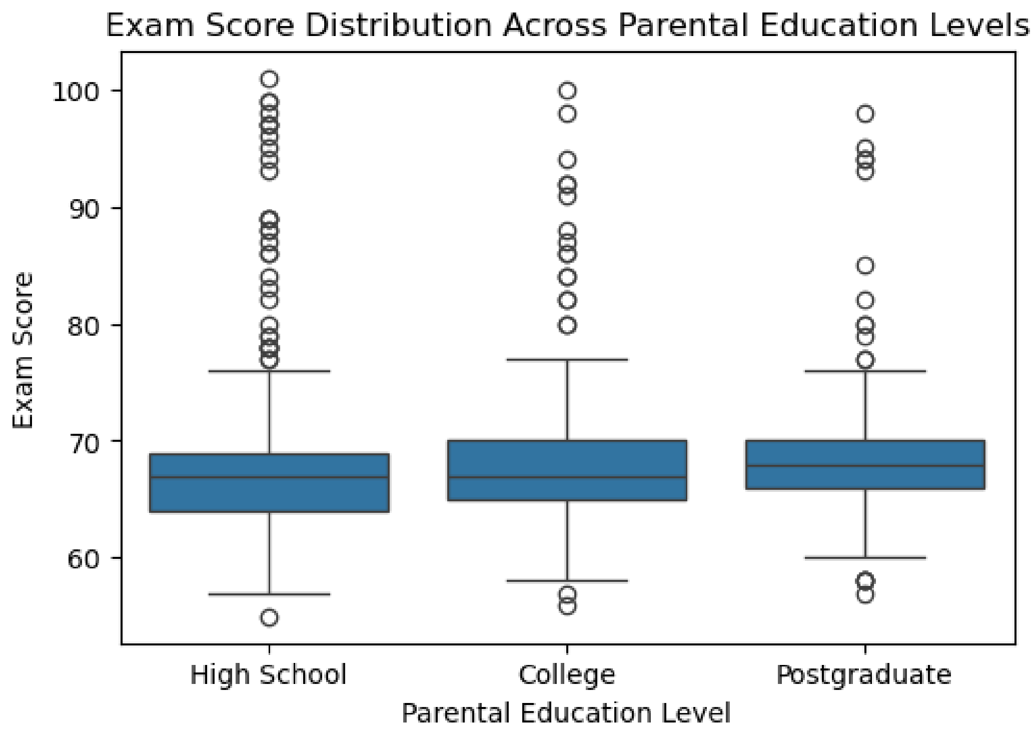


As you may notice, most students in this dataset range in the Average category, or 61-80. Therefore, I will most likely be utilizing Exam_Score more than I previously anticipated because of the overwhelming amount of Average scores in this dataset.

1.3.1 Key Factors v Exam Performance

Educational Level v Exam Score Oftentimes, many students emphasize how their family's background is a big part in how they view school. Students whose families have less formal education may have a different view on the importance of school and that might correlate to the effort students put in towards their exams. Therefore, I would like to create a correlation matrix to analyze the correlation between parental education and exam performance.

[18]:
```
plt.figure(figsize=(6, 4))
sns.boxplot(x=df['Parental_Education_Level'], y=df['Exam_Score'])
plt.xlabel("Parental Education Level")
plt.ylabel("Exam Score")
plt.title("Exam Score Distribution Across Parental Education Levels")
plt.show()
```



As we see in these boxplots, the median score is about the same. Median is a good measure of the central tendency of each group. We notice that each group has a very similar distribution. The only thing that is noticeable is that students whose parents had postgraduate education have a higher floor than the others. I think this goes to show that there is likely more than just your parents' education that affects your exam performance. Maybe what matters more is the individual, not just family life. However, we have a lot more features to evaluate.

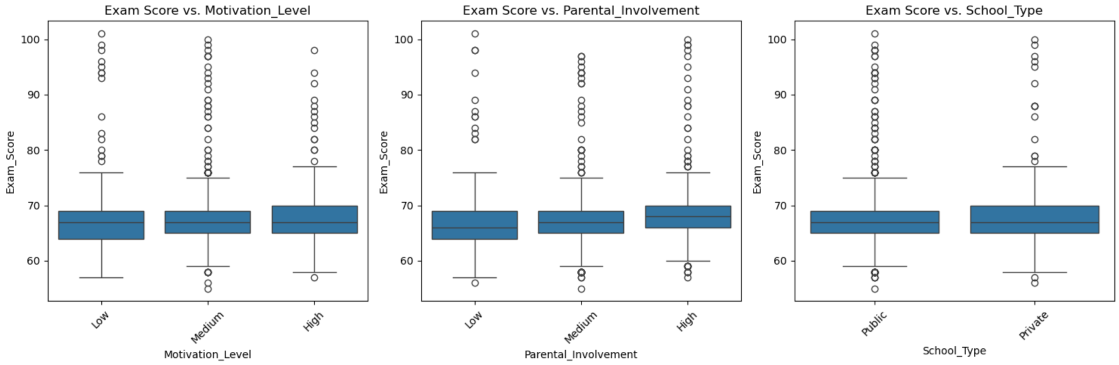
1.3.2 Examining Motivation Level, Parental Involvement, and School Type

Motivation and parental involvement are often cited as key factors influencing student performance, with the assumption that higher levels of both lead to better outcomes. Similarly, the type of school a student attends is frequently debated in terms of its impact on academic success. To better understand these relationships, I chose to visualize exam scores across these three factors using boxplots. This allows for a clear comparison of score distributions and potential trends, helping to determine whether these commonly held beliefs hold true within this dataset.

[21]:
```
categorical_factors = ["Motivation_Level", "Parental_Involvement",␣
 ↪"School_Type"]

plt.figure(figsize=(15, 5))
for i, factor in enumerate(categorical_factors, 1):
    plt.subplot(1, 3, i)
    sns.boxplot(x=df[factor], y=df["Exam_Score"])
    plt.xticks(rotation=45)
    plt.title(f"Exam Score vs. {factor}")

plt.tight_layout()
plt.show()
```



Based on these boxplots, we can see that each of these boxplots have roughly the same median, showing that each of these factors are unlikely to have a significant impact on exam scores alone. One thing we can notice is that some students perform exceptionally despite their motivation level or parental involvement. While the interquartile range for Parental Involvement - High is slightly higher than Low and Medium, it suggests some change based on involvement, but it is not extremely significant.

1.3.3 Score Category Breakdown by Family Income

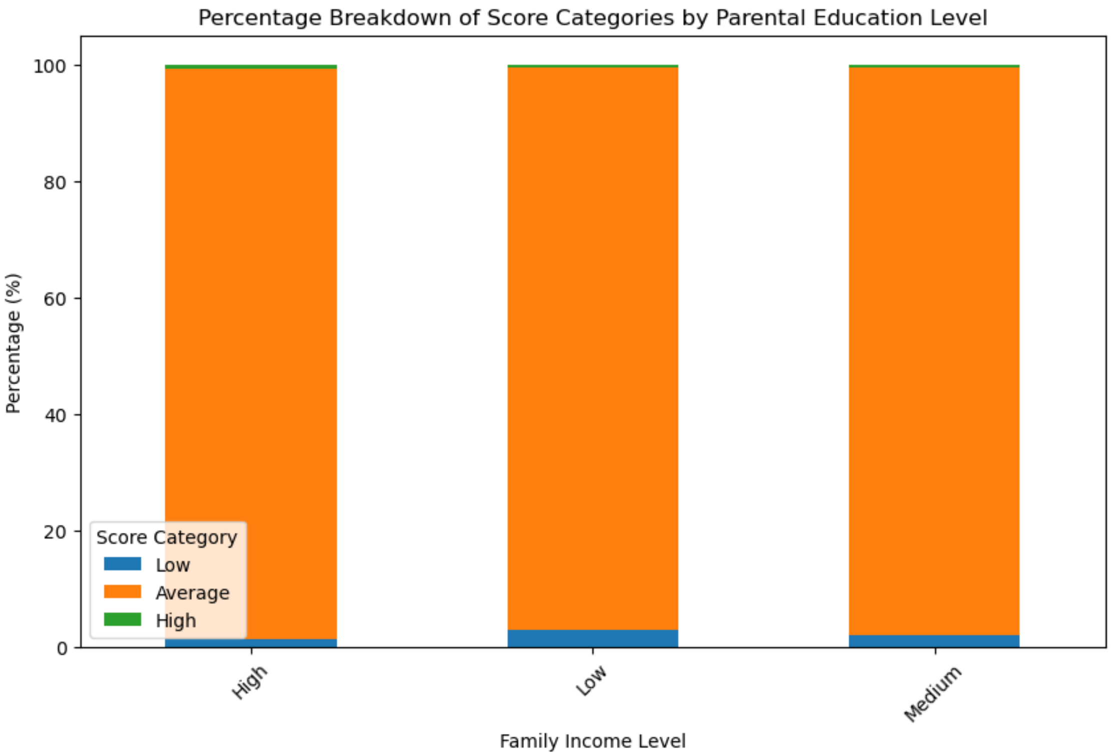
Many studies, groups, and institutions have a tendency to see family income as the "single most significant predictor" of educational success (García & Weiss, 2017). This perspective is something the Economic Policy Institute truly believes in based on their own research. I would like to approach this question based on our dataset and see if there is som

[24]:
```
# I chose to use percentages because there is a count discrepancy between␣
 ↪categories
education_category_counts = df.groupby("Family_Income")["Score_Category"].
 ↪value_counts(normalize=True).unstack() * 100

plt.figure(figsize=(10, 6))
education_category_counts.plot(kind="bar", stacked=True, figsize=(10,6))

plt.title("Percentage Breakdown of Score Categories by Parental Education␣
 ↪Level")
plt.xlabel("Family Income Level")
plt.ylabel("Percentage (%)")
plt.legend(title="Score Category")
plt.xticks(rotation=45)
plt.show()
```

```
<Figure size 1000x600 with 0 Axes>
```



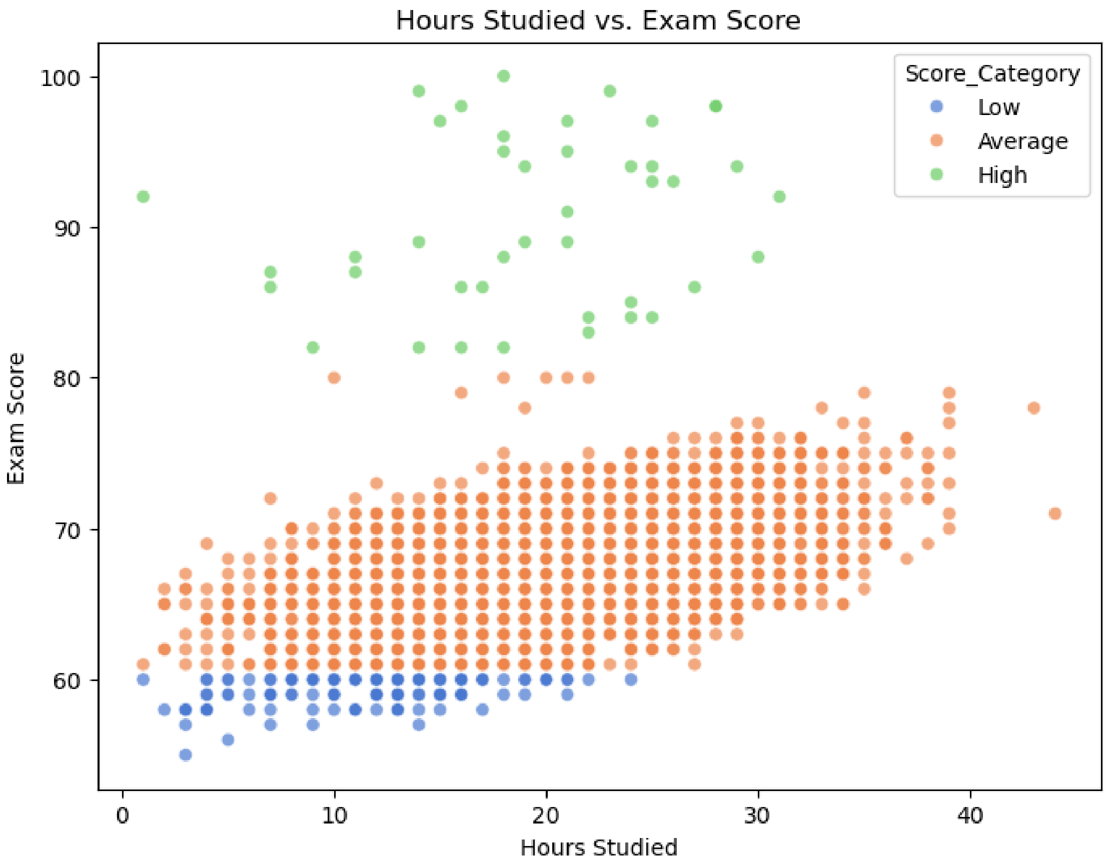
As we can see in this stacked bar chart, students with a higher family income are less likely to be grouped into the Low score category. However, the difference isn't necessarily massive. This might be a possible counterargument to the Economic Policy Institute's conclusion as our dataset seems to point towards income not necessarily being the single most driver of exam performance. However, there are differences in this dataset and the study's assumptions.

1.3.4 Study Habits v Performance

Now that we've evaluated some key performance factors, we can incorporate different features focused more on study habits. No matter the age, place, or situation, one of the most celebrated actions is studying. Studying is glorified as the gateway to better performance on exams and using our dataset, we can analyze this assumption.

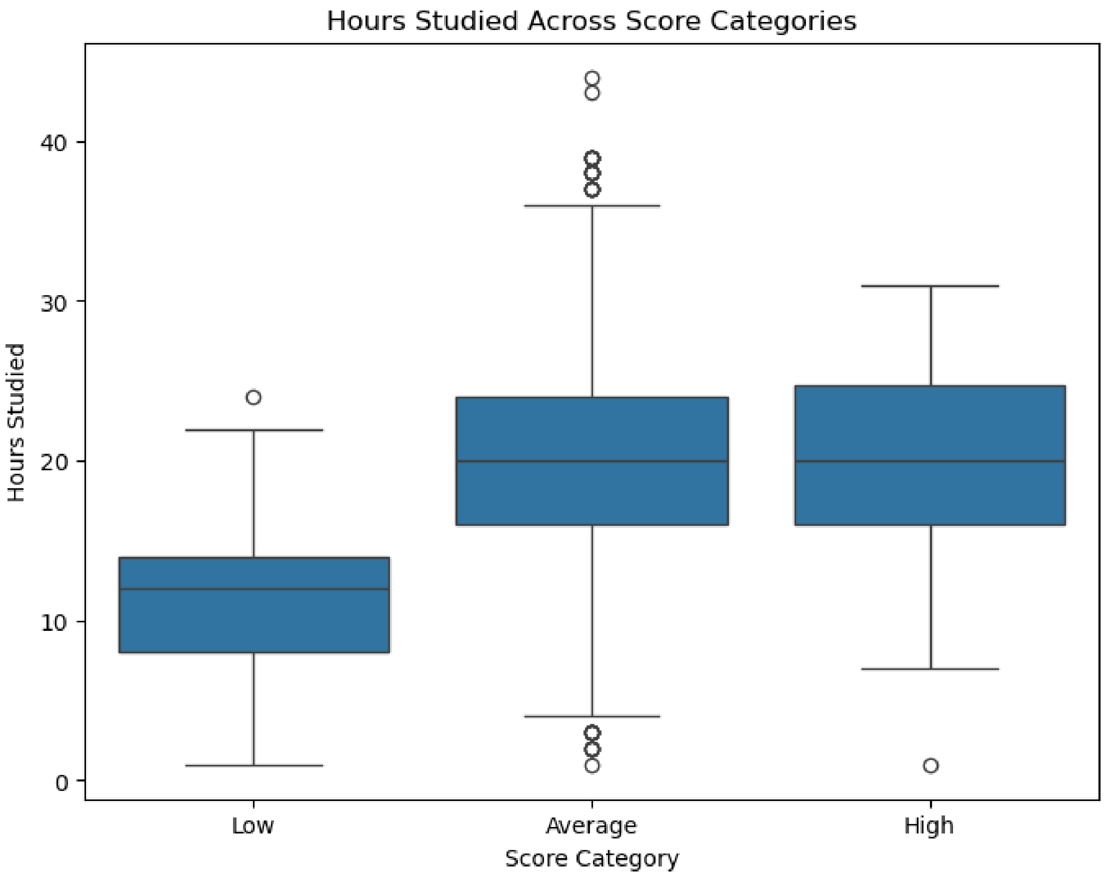
[28]:

```
plt.figure(figsize=(8, 6))
sns.scatterplot(x=df["Hours_Studied"], y=df["Exam_Score"],␣
  ↪hue=df["Score_Category"], palette="muted", alpha=0.7)
plt.title("Hours Studied vs. Exam Score")
plt.xlabel("Hours Studied")
plt.ylabel("Exam Score")
plt.show()
```



This scatter plot illustrates the relationship between Hours Studied and Exam Score, with students categorized into Low (blue), Average (orange), and High (green) score groups. The trend shows a positive correlation—as study hours increase, exam scores tend to rise. However, the majority of students fall within the Average category, suggesting that while more study hours generally lead to better performance, other factors may also influence scores. Additionally, High-scoring students (green) are more spread out, indicating that some achieve top scores with relatively fewer study hours, potentially due to other factors like prior knowledge or study efficiency.

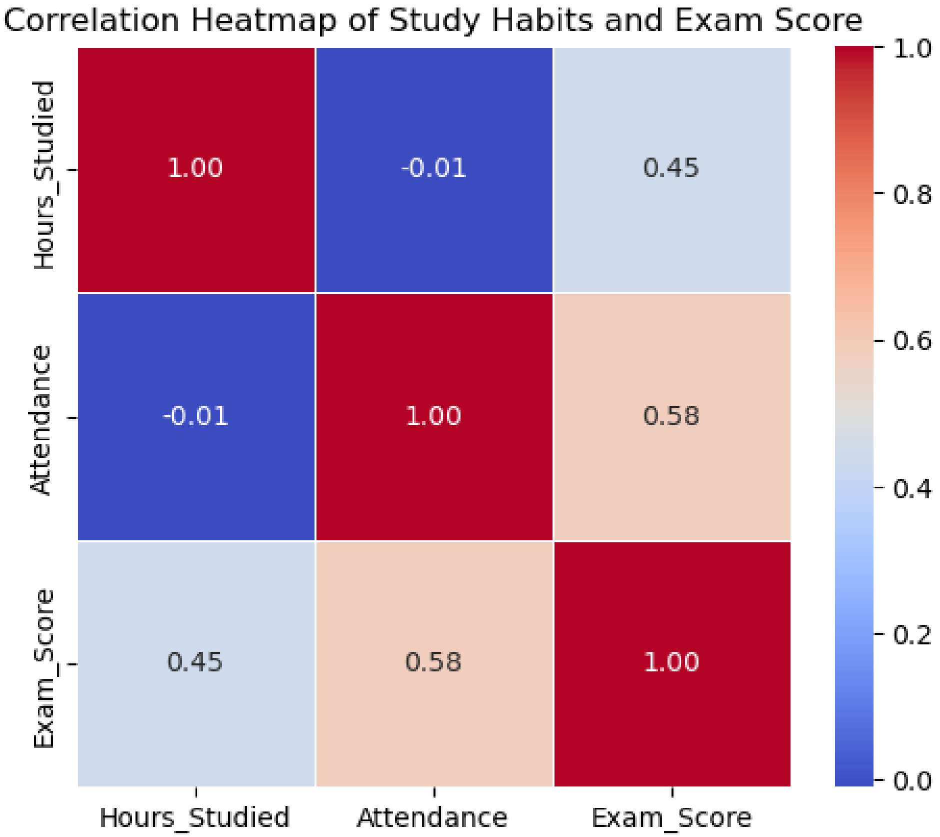
[30]:
```
plt.figure(figsize=(8, 6))
sns.boxplot(x=df["Score_Category"], y=df["Hours_Studied"])
plt.title("Hours Studied Across Score Categories")
plt.xlabel("Score Category")
plt.ylabel("Hours Studied")
plt.show()
```



This boxplot shows the distribution of Hours Studied across different Score Categories (Low, Average, High). As expected, students in the High score category tend to have a higher median study time compared to those in the Low category. However, there is significant overlap between the Average and High categories, suggesting that while studying more generally leads to better performance, other factors likely play a role in determining scores. The presence of outliers (particularly low study hours in the High category) suggests that some students achieve high scores with relatively little studying, possibly due to prior knowledge, efficient study habits, or external support.

[32]:
```
import numpy as np

corr_features = df[["Hours_Studied", "Attendance", "Exam_Score"]].corr()
plt.figure(figsize=(6, 5))
sns.heatmap(corr_features, annot=True, cmap="coolwarm", fmt=".2f", linewidths=0.
↪5)
plt.title("Correlation Heatmap of Study Habits and Exam Score")
plt.show()
```

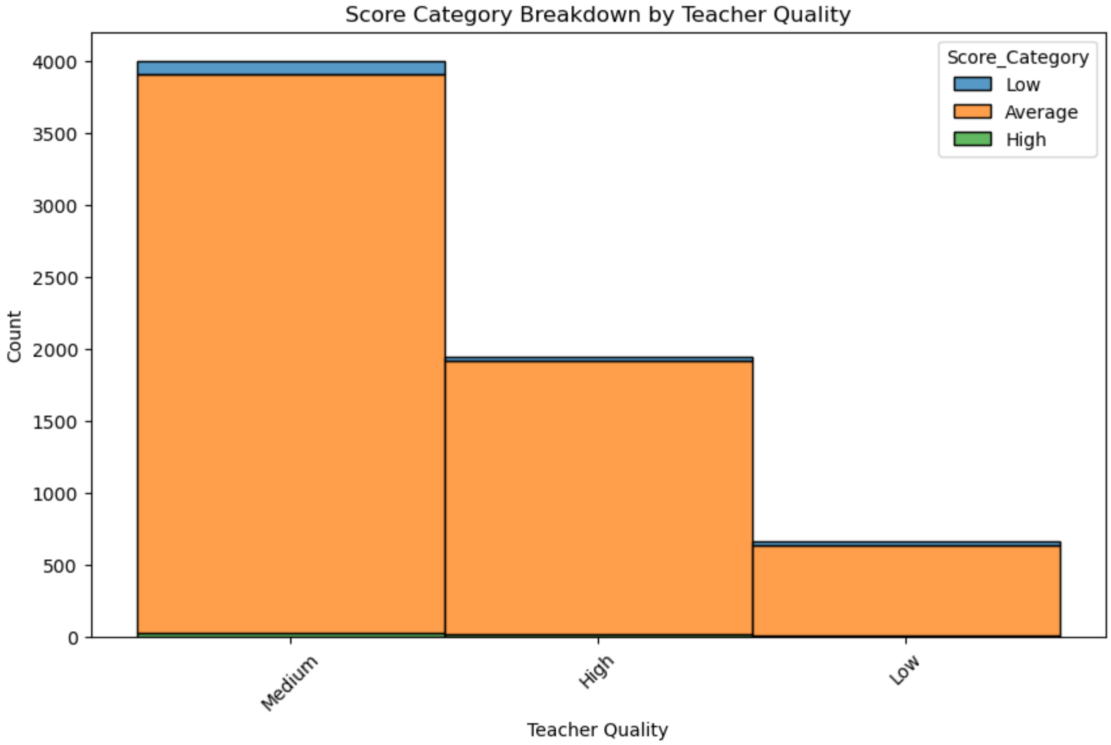


This heatmap shows that attendance (0.58) has a stronger correlation with exam scores than hours studied (0.45), suggesting that class participation plays a key role in performance. The near-zero correlation between hours studied and attendance (-0.01) indicates that studying habits are independent of class attendance. While both studying and attending class contribute to better scores, attendance appears to be the more influential factor.

1.3.5 School and Environmental Impact

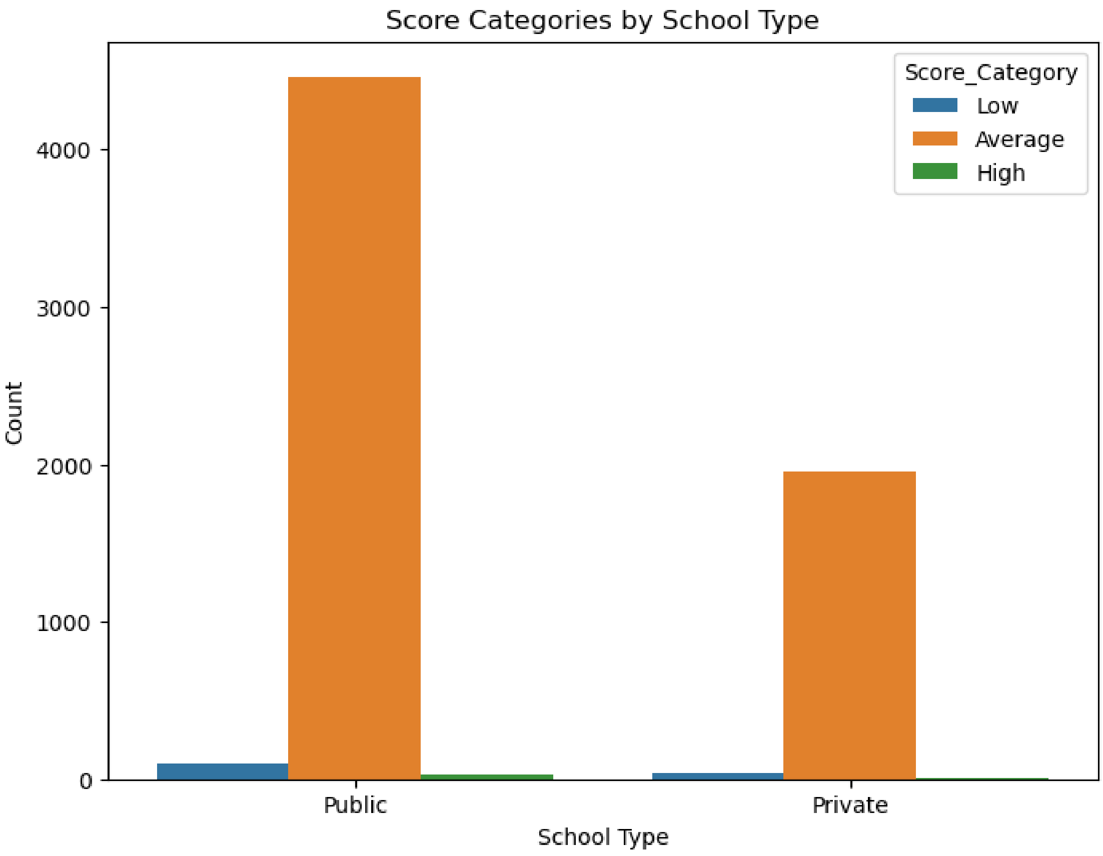
Teacher Quality Teacher quality is often considered a crucial factor in student performance, with the assumption that higher-quality teachers lead to better outcomes. To explore this, I created a count plot comparing score categories across different levels of teacher quality. This visualization helps assess whether students with higher-rated teachers are more likely to achieve High scores or if the impact of teacher quality is less pronounced. By analyzing these distributions, we can better understand whether teacher quality is a key determinant of exam success or if other factors play a larger role.

[35]:
```
plt.figure(figsize=(10, 6))
sns.histplot(data=df, x="Teacher_Quality", hue="Score_Category",␣
 ↪multiple="stack")
plt.xticks(rotation=45)
plt.title("Score Category Breakdown by Teacher Quality")
plt.xlabel("Teacher Quality")
plt.ylabel("Count")
plt.show()
```



School Type School type is often discussed as a factor influencing academic performance, with the assumption that private schools provide better outcomes. To explore this, I created a count plot comparing score categories across public and private schools. This visualization helps identify any disparities in performance distribution and whether one school type has a higher proportion of students in the High or Low score categories. By analyzing these trends, we can assess whether school type plays a significant role in exam outcomes or if other factors might be more influential.

[37]:
```
plt.figure(figsize=(8, 6))
sns.countplot(x=df["School_Type"], hue=df["Score_Category"])
plt.title("Score Categories by School Type")
plt.xlabel("School Type")
plt.ylabel("Count")
plt.show()
```



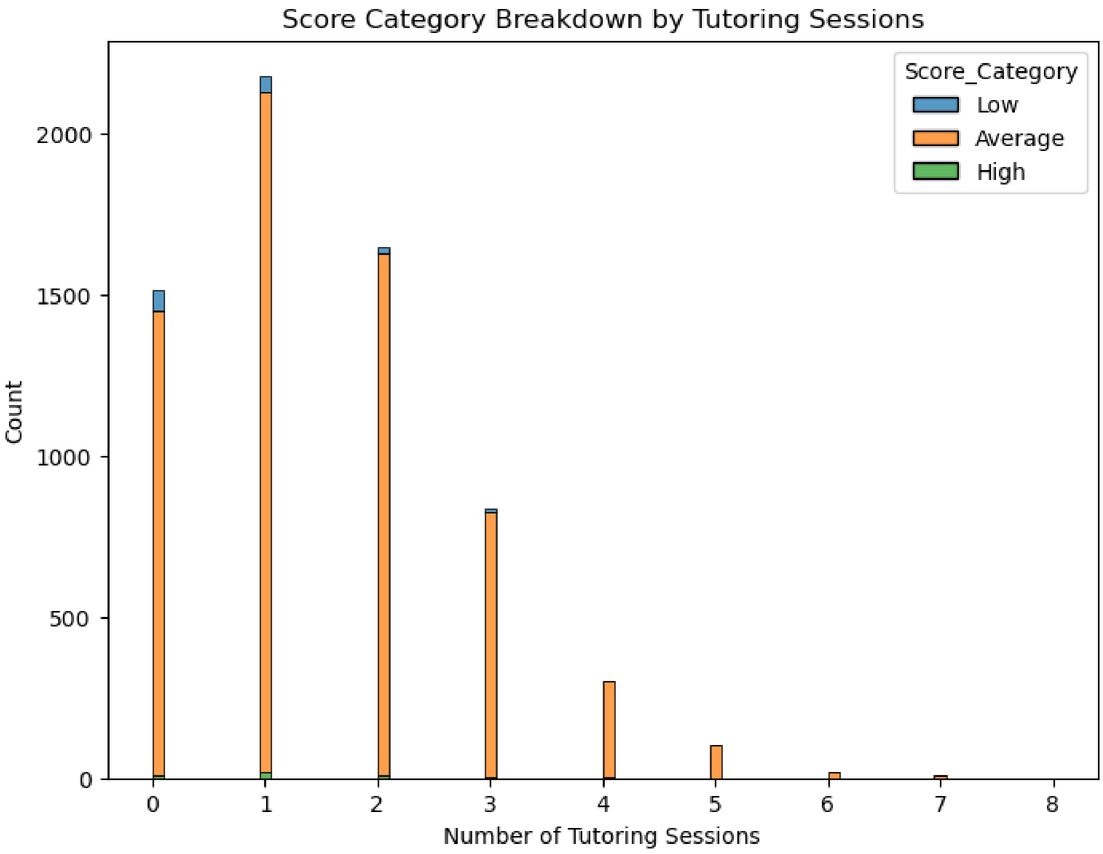
As we can see in this stacked bar chart, the majority of students fall into the Average score category, with Public schools having a significantly larger representation compared to Private schools. While there are students in the Low and High score categories, their proportions are relatively small in both school types. This might suggest that school type alone isn't the primary determinant of exam performance, as both Public and Private schools show similar distributions outside the Average category. However, there may be other underlying factors influencing these results that aren't captured in this dataset.

1.3.6 Effect of Interventions

Tutoring Sessions Tutoring is often seen as a way to improve academic performance, with the expectation that more sessions lead to better outcomes. To explore this, I created a count plot comparing score categories across different numbers of tutoring sessions. This visualization helps assess whether students who attend more tutoring sessions are more likely to achieve High scores or if the distribution remains consistent regardless of tutoring. By analyzing these trends, we can determine whether tutoring plays a significant role in exam success or if its impact is less pronounced than expected.

[40]:
```
plt.figure(figsize=(8, 6))
sns.histplot(data=df, x="Tutoring_Sessions", hue="Score_Category",␣
  ↪multiple="stack")
```

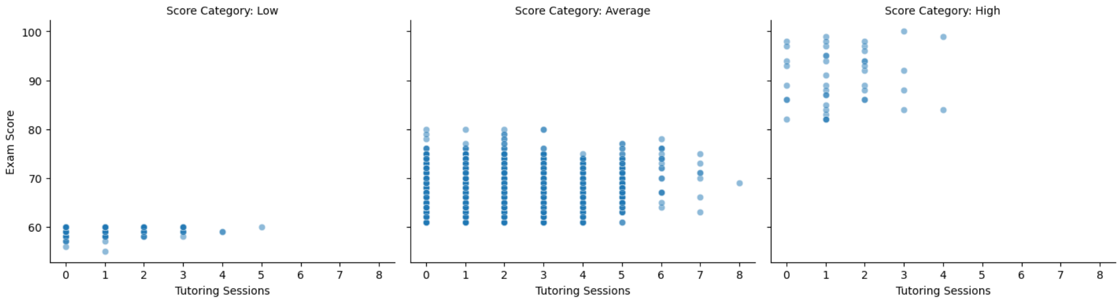
```
plt.title("Score Category Breakdown by Tutoring Sessions")
plt.xlabel("Number of Tutoring Sessions")
plt.ylabel("Count")
plt.show()
```



This stacked bar chart shows that most students fall into the Average score category, regardless of the number of tutoring sessions attended. The majority of students had zero or one tutoring session, with participation dropping significantly as the number of sessions increased. The Low and High score categories appear in very small proportions, suggesting that tutoring may not be the sole driver of improved performance. However, other factors could be influencing these results, and the dataset's limitations should be considered when interpreting the relationship between tutoring and exam scores.

Tutoring Tutoring is often recommended as a way to boost academic performance, with the expectation that more sessions will lead to higher scores. To analyze this, I created scatter plots that break down exam scores by tutoring sessions across three score categories: Low, Average, and High. This visualization helps assess whether students in different score categories benefit differently from tutoring. While the Low category remains clustered around the 60s regardless of tutoring, the Average category shows a wider spread with some increase in variability. The High category consists of students scoring above 90, with some achieving high scores even with minimal tutoring. This suggests that tutoring may not have a uniform effect on all students, and other factors could be influencing performance.

[43]:
```
g = sns.FacetGrid(df, col="Score_Category", height=4, aspect=1.2)
g.map_dataframe(sns.scatterplot, x="Tutoring_Sessions", y="Exam_Score", alpha=0.
  ↪5)
g.set_axis_labels("Tutoring Sessions", "Exam Score")
g.set_titles(col_template="Score Category: {col_name}")
plt.show()
```



This scatter plot breaks down exam scores by tutoring sessions across three score categories: Low, Average, and High. Students in the Low category consistently score around the 60s, regardless of the number of tutoring sessions attended. The Average category shows a wider spread, with most students clustering between 70 and 80, and a slight increase in score variability as tutoring sessions increase. The High category consists of students scoring above 90, with some attending more tutoring sessions but others achieving high scores with minimal tutoring. This suggests that while tutoring may have some influence, other factors likely contribute to exam performance, and its impact varies across different score groups.

While tutoring is often assumed to enhance academic performance, our dataset suggests its impact is limited. One explanation for this could be that students who require more tutoring may already be struggling with foundational concepts, meaning tutoring alone is insufficient to bridge performance gaps. Additionally, the dataset does not specify the quality of tutoring sessions or whether they were effectively structured to improve understanding. Future research could examine the difference between structured tutoring programs and casual one-on-one sessions

1.4 Summary

Many factors are often assumed to influence student performance, including school type, tutoring, teacher quality, and parental involvement. However, the analysis suggests that these factors may not be as impactful as commonly believed. Private and public schools show similar distributions of exam scores, with most students falling into the Average category. Tutoring also appears to have minimal influence, as students across all score categories exhibit similar patterns regardless of the number of sessions attended. Likewise, teacher quality does not seem to strongly correlate with higher performance, as students with highly rated teachers still primarily fall within the Average range.

Motivation and parental involvement also show little variation in exam scores, challenging the notion that increased engagement in these areas leads to significantly better outcomes. Scatter plots further support this by demonstrating that students in the Low category consistently score around the same level, regardless of tutoring, while high-scoring students achieve their results through a variety of different experiences. These findings suggest that external factors like school environment and tutoring may not be the biggest determinants of success—other variables, such as individual study habits or intrinsic ability, could play a much larger role in shaping academic performance.

From my analysis, I found that while parental involvement and income levels do show some correlation with student performance, they are not the sole predictors. Motivation and study habits play a crucial role, with attendance being a particularly strong factor. Additionally, tutoring and teacher quality, which are often assumed to have a significant impact, do not show as strong of a correlation in this dataset. These findings suggest that a student's personal study habits and engagement may outweigh traditional assumptions about academic success.

1.4.1 Bias and Limitations

One limitation of this dataset is that it does not account for intrinsic abilities or learning differences that could play a significant role in student success. Additionally, socioeconomic status is often linked to access to resources beyond just tutoring, such as stable home environments, reduced stress, and better schools, which this dataset may not fully capture. Another potential bias is self-reporting inaccuracies in study habits or parental involvement. Future research could incorporate more qualitative data, such as student interviews, to provide a fuller picture of academic success factors.

1.5 References

García, E., & Weiss, E. (2017, September 27). Education inequalities at the school starting gate: Gaps, trends, and strategies to address them. Retrieved from Economic Policy Institute website: https://www.epi.org/publication/education-inequalities-at-the-school-starting-gate/

Kaggle. (2024). Student Performance Factors. Retrieved from Kaggle.com website: https://www.kaggle.com/datasets/lainguyn123/student-performance-factors/data