Student Performance Factors Project Report

Sara Akbarzadeh

March 2025

Abstract

In this project, we analyze the Student Performance Factors dataset from Kaggle to identify key elements influencing academic success. Using Exploratory Data Analysis (EDA), we examine patterns and correlations within the data, focusing on study habits, sleep duration, parental involvement, extracurricular activities, internet access, gender, and more. To assess the significance of these factors, we perform statistical tests such as t-tests, ANOVA, and correlation analyses. The results reveal critical aspects that contribute to students' exam scores, providing valuable insights for educators and policymakers. Visualizations further aid in interpreting the findings, demonstrating the importance of data-driven decision-making to improve educational outcomes.

1 Introduction

Academic performance plays a vital role in shaping students' futures by providing access to opportunities and fostering personal and professional growth. In this project, we analyze the Student Performance Factors dataset from Kaggle to identify key factors influencing student performance. Our primary objective is to understand how various attributes—such as study habits, parental involvement, extracurricular activities, sleep patterns, and socio-economic factors—affect exam scores.

We begin by performing Exploratory Data Analysis (EDA) to uncover patterns and trends within the data. Through visualizations and statistical analyses, we explore the relationships between different variables and their impact on academic success. Finally, we conduct hypothesis testing to validate assumptions and quantify the significance of various factors. Our findings provide valuable insights that can help educators and policymakers make data-driven decisions to enhance educational outcomes.

2 Dataset

2.1 Data Description

The dataset used in this project is the Student Performance Factors dataset from Kaggle. It contains a total of 6,607 rows, each representing an individual student, and 20 columns, which correspond to various features related to academic performance.

The features include both numerical and categorical data, encompassing factors such as hours studied, attendance rate, parental involvement, extracurricular activities, sleep hours, motivation level, and more. The target variable is the *Exam Score*, which indicates the academic performance of each student.

The dataset also includes socio-economic factors such as family income, parental education level, and access to resources, along with personal attributes like gender and physical activity levels. This comprehensive data enables us to explore a wide range of factors that may influence student performance and academic success.

2.2 Feature Description

The dataset contains 20 features, each representing a specific aspect of student performance. Below is a detailed description of each feature:

- **Hours_Studied (int)** – Number of hours a student spends studying.

- **Attendance (int)** – Attendance percentage or number of days attended.
- **Parental_Involvement (object)** – Level of parental engagement in academics (Low, Medium, High).
- **Access_to_Resources (object)** – Availability of academic resources (Low, Medium, High).
- **Extracurricular_Activities (object)** – Participation in extracurricular activities (Yes, No).
- **Sleep_Hours (int)** – Average number of hours a student sleeps per night.
- **Previous_Scores (int)** – Previous exam or academic scores.
- **Motivation_Level (object)** – Self-reported motivation level (Low, Medium, High).
- **Internet_Access (object)** – Whether the student has internet access at home (Yes, No).
- **Tutoring_Sessions (int)** – Number of tutoring sessions attended.
- **Family_Income (object)** – Family's financial status (Low, Medium, High).
- **Teacher_Quality (object)** – Perceived quality of teaching (Low, Medium, High).
- **School_Type (object)** – Type of school attended (Public, Private).
- **Peer_Influence (object)** – Influence of peers on academics (Positive, Negative, Neutral).
- **Physical_Activity (int)** – Hours spent on physical activity per week.
- **Learning_Disabilities (object)** – Presence of learning disabilities (Yes, No).
- **Parental_Education_Level (object)** – Highest education level attained by parents (High School, College, Postgraduate).
- **Distance_from_Home (object)** – Distance of the school from home (Near, Moderate, Far).
- **Gender (object)** – Student's gender (Male, Female).
- **Exam_Score (int)** – Final exam score (target variable).

These features encompass a wide range of factors that may influence student performance, including academic habits, socio-economic background, personal attributes, and school-related characteristics. Understanding these factors can help identify key predictors of academic success.

3 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a crucial step in understanding the underlying patterns and trends in the dataset. Before conducting deeper analysis, we first explore the dataset using visualization techniques and check for missing values. Visualizing numerical features helps us understand their distribution, detect potential outliers, and identify patterns that may influence student performance. Additionally, checking for missing values ensures data completeness and helps us decide on appropriate preprocessing steps, such as imputing missing values or handling inconsistencies.

3.1 Visualizations

The following visualizations were instrumental in gaining insights during the EDA phase. These plots helped uncover important patterns and relationships between various factors that could affect student performance. We have visualized the distribution of both categorical and numerical features, and also explored the correlations among numerical features. The visualizations include:

- Distribution of Numerical Features: Histograms for numerical features were created to illustrate the spread and central tendency of variables such as 'Hours Studied', 'Sleep Hours', and 'Previous Scores'. See Figure 1

- Distribution of Categorical Features: This shows the frequency of different categories in each categorical variable, helping us understand how these factors are distributed across the dataset. See Figure 2

- Correlation Heatmap of Numerical Features: A heatmap was used to visualize the correlation between various numerical features, providing insights into which variables are closely related and may influence each other.

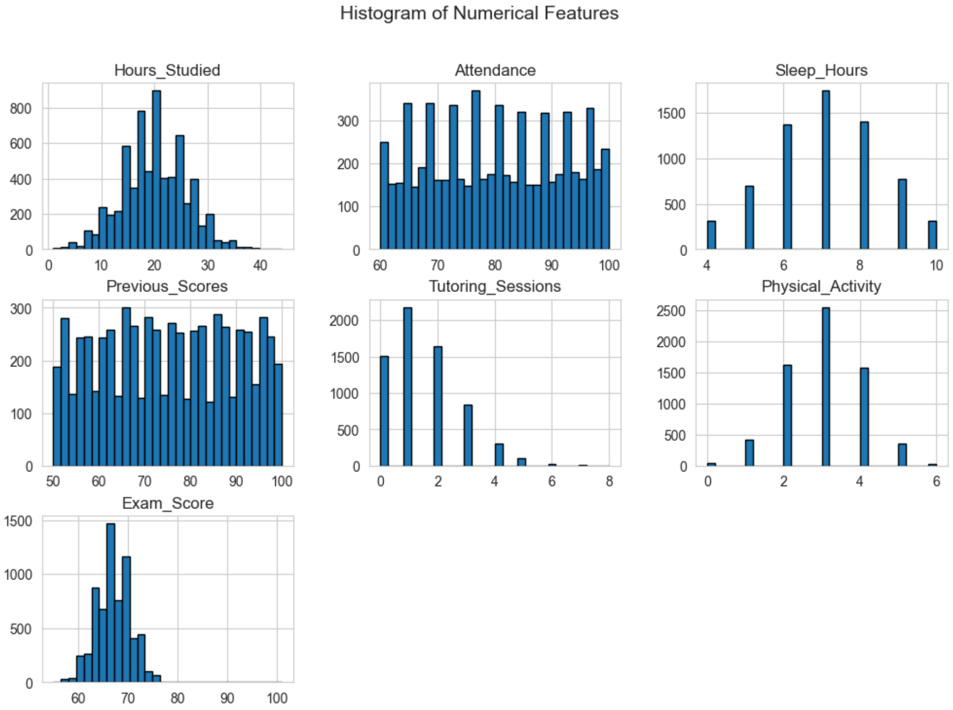

Figure 1: Distribution of Numerical Features

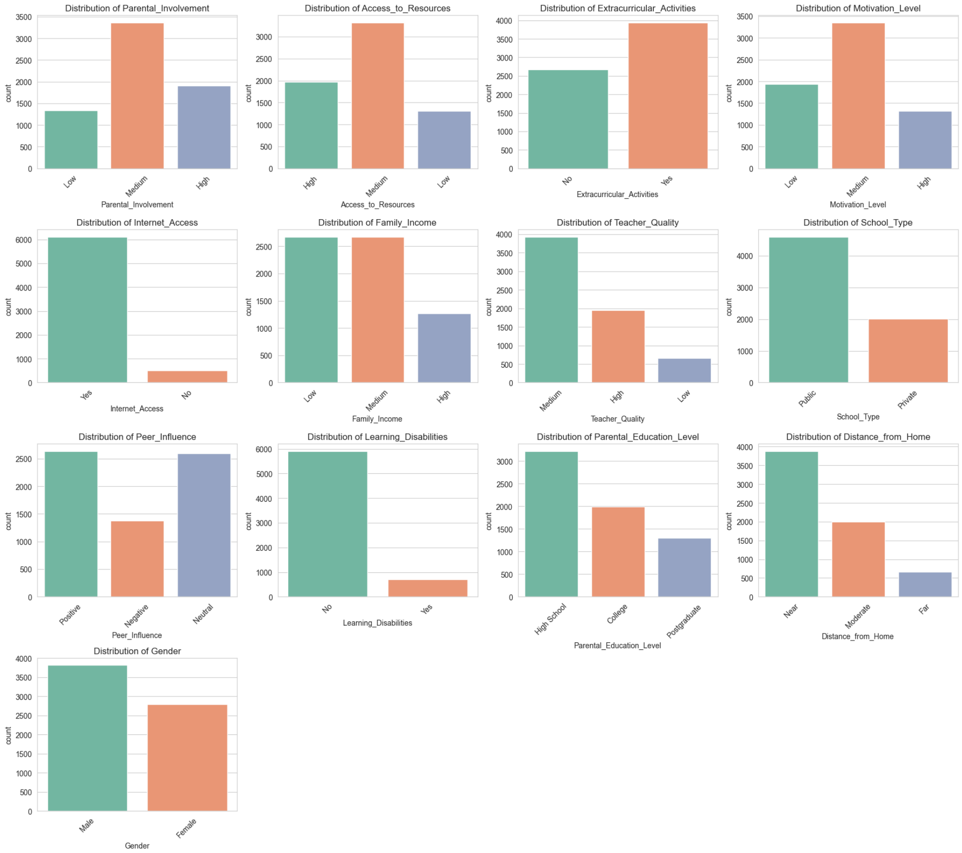

Figure 2: Distribution of Categorical Features

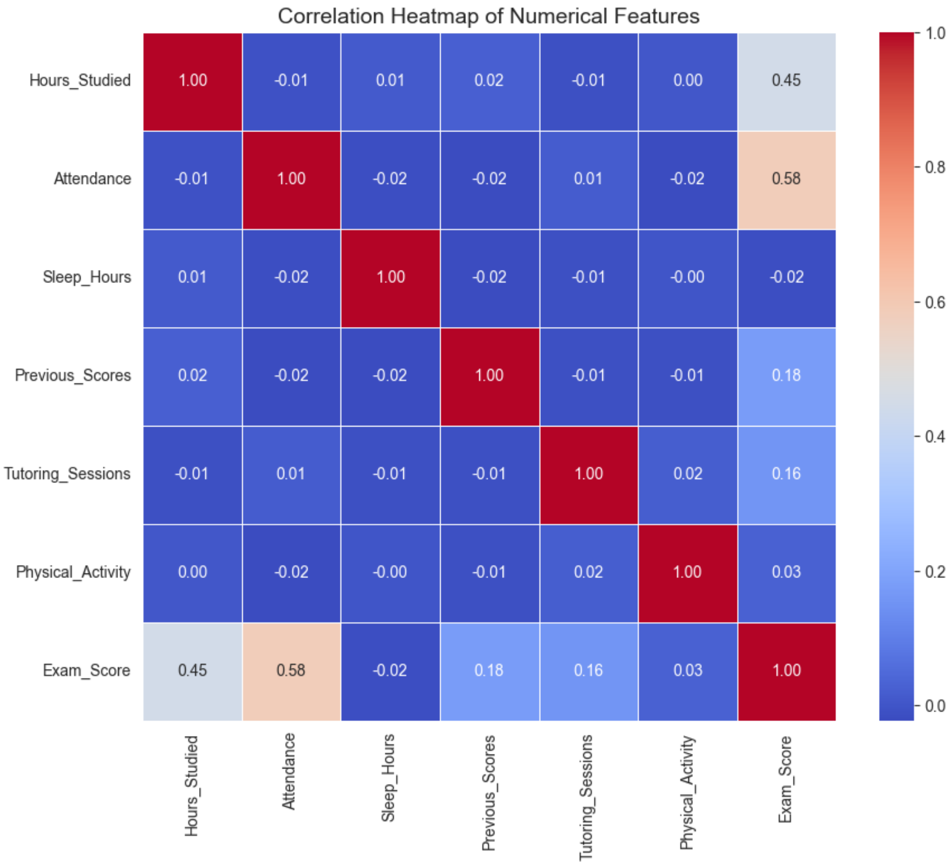

Figure 3: Correlation Heatmap of Numerical Features

These visualizations helped us understand the underlying distribution of the numerical features, the frequency of categories in the categorical features, and the relationships between numerical features. By examining these plots, we were able to identify patterns that guided the subsequent steps in hypothesis testing and modeling.

4 Hypothesis Testing

In this section, we perform hypothesis testing to validate assumptions regarding factors influencing student performance. For each hypothesis, we used appropriate statistical tests and visualizations to draw meaningful conclusions. Below, we present each hypothesis, the statistical test used, and the relevant plot.

4.1 Hypothesis 1: Students with more hours of study tend to have higher exam scores.

- **Null Hypothesis (H_0)**: There is no significant relationship between the number of hours studied and exam scores.
- **Alternative Hypothesis (H_1)**: There is a significant positive relationship between the number of hours studied and exam scores.

We used Pearson's correlation test to analyze the relationship between them. The correlation coefficient and p-value were computed to assess the strength and significance of the relationship.

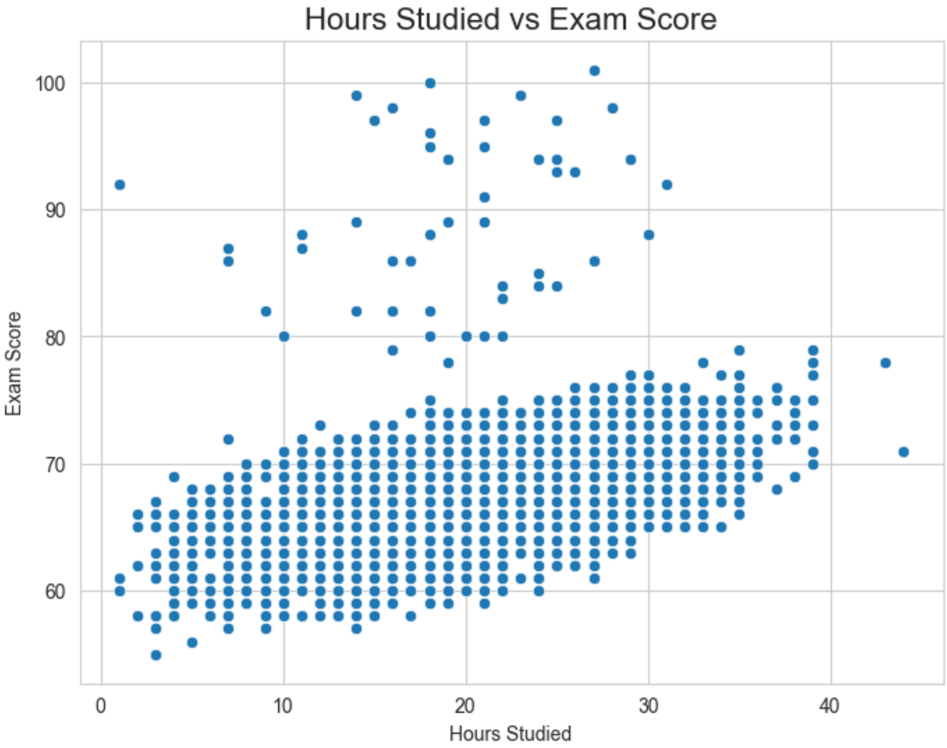

Figure 4: Correlation between Hours Studied and Exam Scores

Interpretation:

- If the p-value is less than 0.05, we reject the null hypothesis and conclude that there is a significant positive correlation between hours studied and exam scores.
- If the p-value is greater than 0.05, there is no significant relationship.

Result:

- Pearson Correlation: 0.44545495407528135, P-Value: 1.28635e-319
- Reject the null hypothesis: There is a significant positive relationship between hours studied and exam score.

4.2 Hypothesis 2: Students with high parental involvement perform better in exams.

- **Null Hypothesis ($\mathbf{H_0}$)**: Parental involvement has no effect on student exam scores.
- **Alternative Hypothesis ($\mathbf{H_1}$)**: Students with high parental involvement tend to have higher exam scores.

We used ANOVA to compare the mean exam scores across different levels of parental involvement.

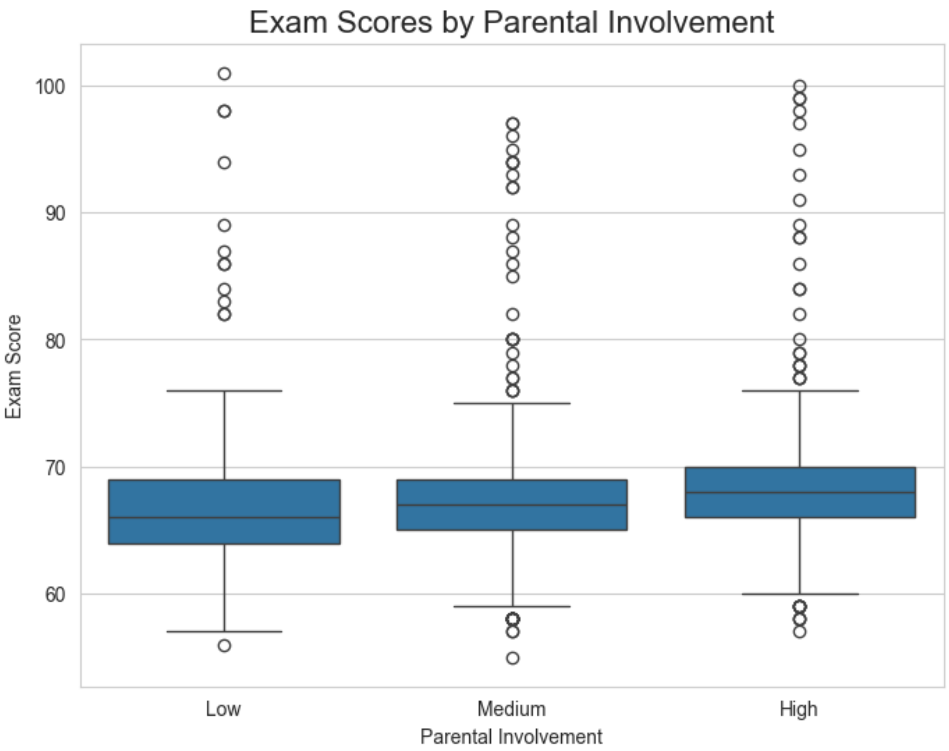

Figure 5: Exam Scores by Parental Involvement

Interpretation:

- If the p-value is less than 0.05, we reject the null hypothesis and conclude that there is a significant difference in exam scores between students with different levels of parental involvement.
- If the p-value is greater than 0.05, the hypothesis that parental involvement affects exam scores cannot be accepted.

Result:

- ANOVA result: F-statistic = 84.48765484606227, P-value = 5.8754791533325443e-37
- Reject the null hypothesis: Parental involvement significantly affects exam scores.

4.3 Hypothesis 3: Students who participate in extracurricular activities perform better in exams.

- **Null Hypothesis (H_0)**: There is no significant difference in exam scores between students who participate in extracurricular activities and those who do not.
- **Alternative Hypothesis (H_1)**: Students who participate in extracurricular activities tend to have higher exam scores.

We used a t-test to compare the exam scores between students who participate and those who do not.

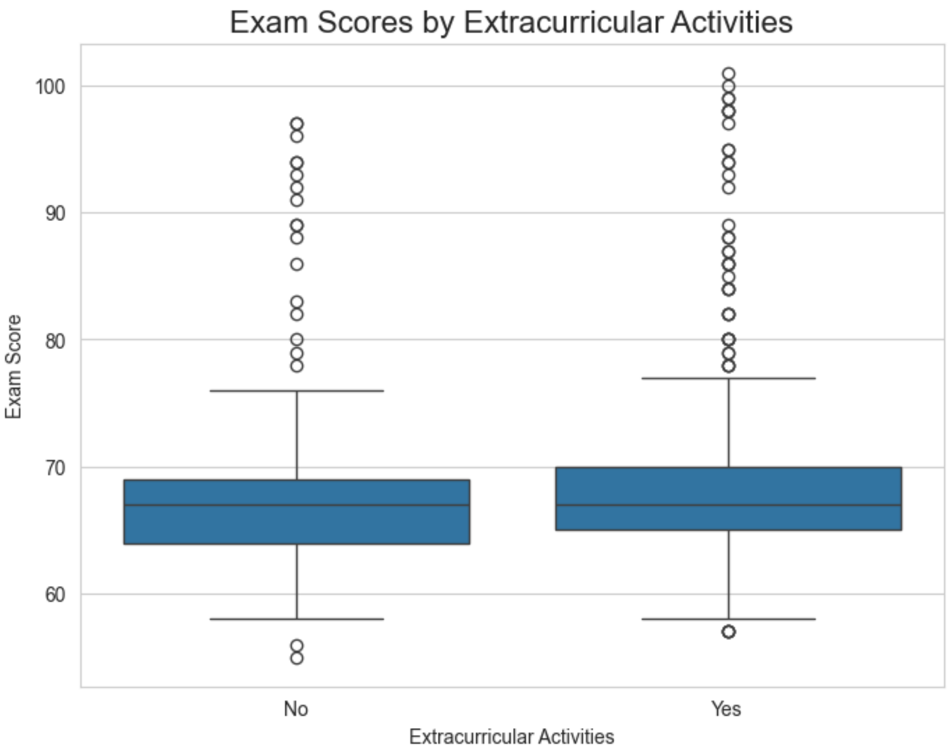

Figure 6: extracurricular Activities

Interpretation:

- If the p-value is less than 0.05, you can reject the null hypothesis and conclude that students who participate in extracurricular activities have significantly different exam scores than those who do not.
- If the p-value is greater than 0.05, there is no significant difference.

Result:

- T-statistic: 5.2432536469508255, P-value: 1.6266777077313432e-07
- Reject the null hypothesis: Students with extracurricular activities perform better in exams.

4.4 Hypothesis 4: Students with internet access perform better in exams.

- **Null Hypothesis ($\mathbf{H_0}$)**: Internet access does not significantly affect exam scores.
- **Alternative Hypothesis ($\mathbf{H_1}$)**: Students with internet access tend to perform better in exams.

We used a t-test to compare the exam scores between students with and without internet access.

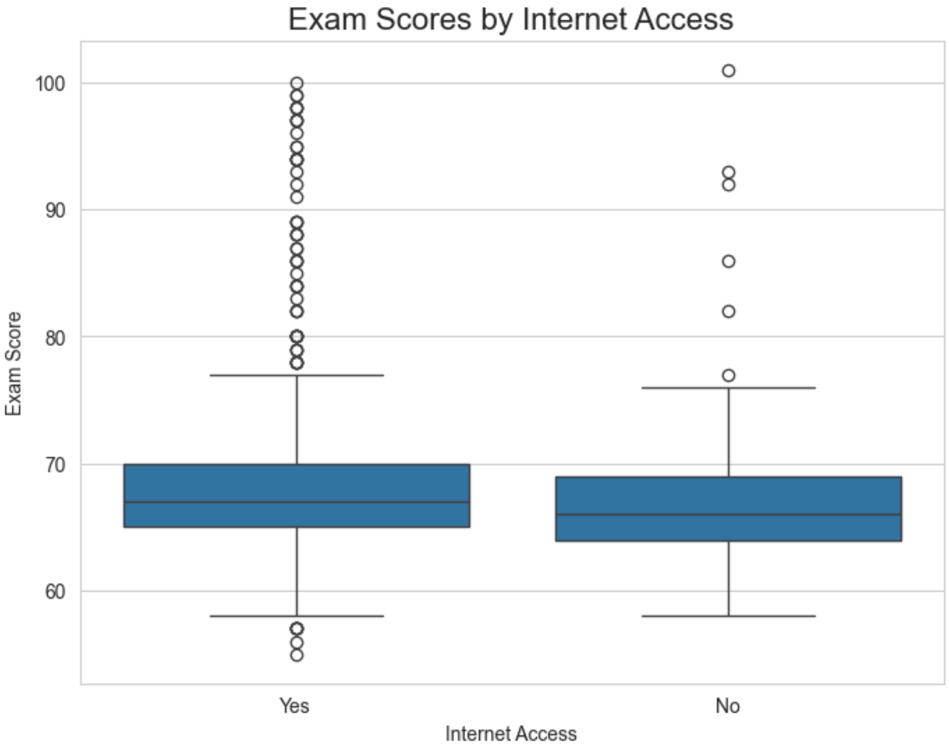

Figure 7: Exam Scores by Internet Acess

Interpretation:

- If the p-value is less than 0.05, you can reject the null hypothesis and conclude that internet access does significantly affect exam performance.
- If the p-value is greater than 0.05, internet access does not significantly affect exam scores.

Result:

- T-statistic: 4.188986958317149, P-value: 2.8385046310278915e-05
- Reject the null hypothesis: Students with internet access perform better in exams.

4.5 Hypothesis 5: Gender has no significant effect on exam performance.

- **Null Hypothesis (H_0)**: There is no significant difference in exam scores between male and female students.
- **Alternative Hypothesis (H_1)**: There is a significant difference in exam scores between male and female students.

A t-test compared exam scores between male and female students. A boxplot helped to visualize the comparison.

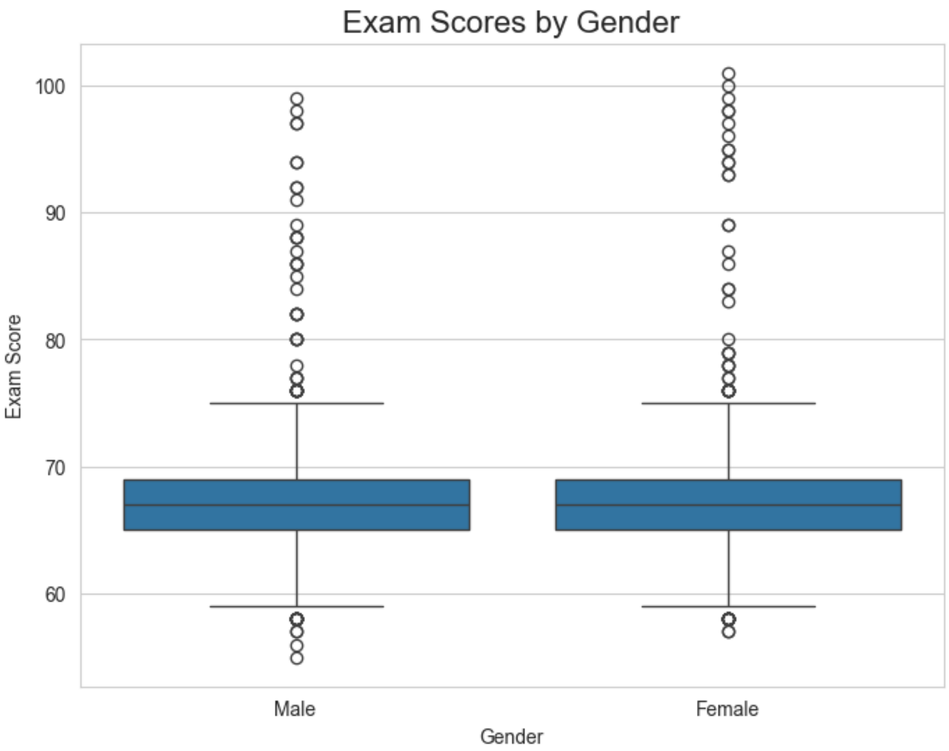

Figure 8: Exam Scores by Gender

Interpretation:

- If p-value is more than 0.05: Fail to reject null hypothesis → No significant difference in exam scores by gender.
- If p-value is less than 0.05: Reject null hypothesis → Gender significantly affects exam scores.

Result:

- T-statistic: -0.16516987601406408, P-value: 0.8688153297340319
- Fail to reject the null hypothesis: There is no significant difference in exam scores between male and female students.

4.6 Hypothesis 6: Sleep duration has a significant effect on exam scores.

- **Null Hypothesis (H_0)**: Sleep duration does not significantly impact exam scores..
- **Alternative Hypothesis (H_1)**: Sleep duration significantly impacts exam scores.

A Pearson correlation test analyzed the relationship between sleep hours and exam scores. A scatter plot helped to visualize this correlation.

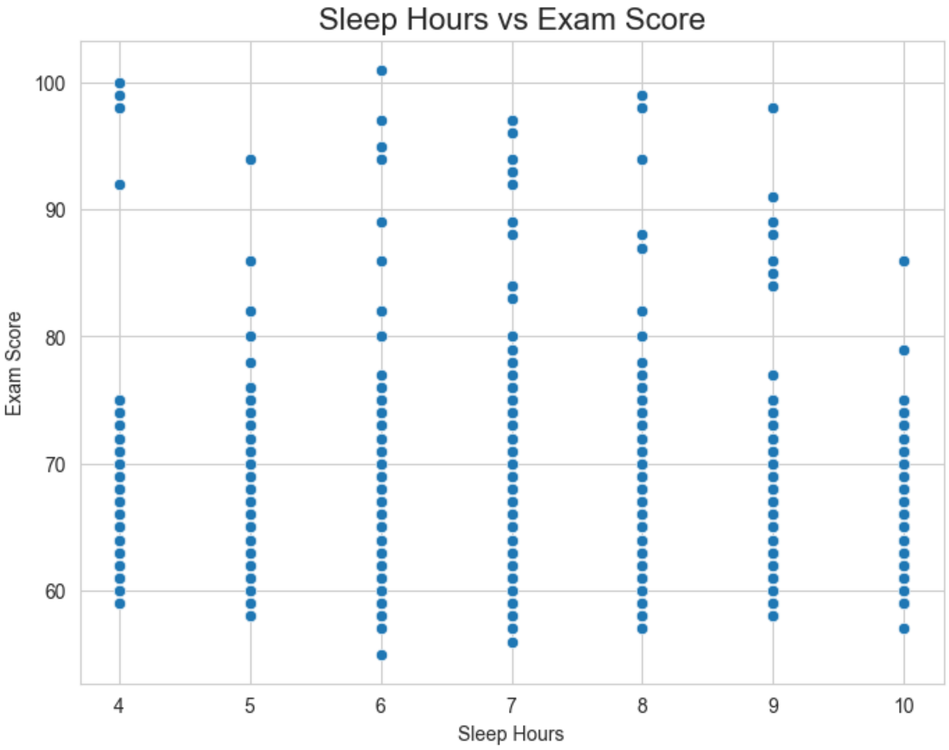

Figure 9: Exam Scores and Sleep Hours

Interpretation:

- If p-value is more than 0.05: Fail to reject null hypothesis → Sleep duration does not significantly affect exam scores.
- If p-value is less than 0.05: Sleep duration significantly affects exam scores.

Result:

- Pearson Correlation: -0.01702162857150259, P-Value: 0.16653759133789678
- Fail to reject the null hypothesis: Sleep duration does not significantly impact exam scores.

4.7 Hypothesis 7: Family income has no significant effect on exam scores.

- **Null Hypothesis (H_0)**: Family income does not significantly impact exam performance.
- **Alternative Hypothesis (H_1)**: Family income significantly impacts exam performance.

An ANOVA test ccompares exam scores across different family income groups (low, middle, high). A boxplot helped to visualize the difference.

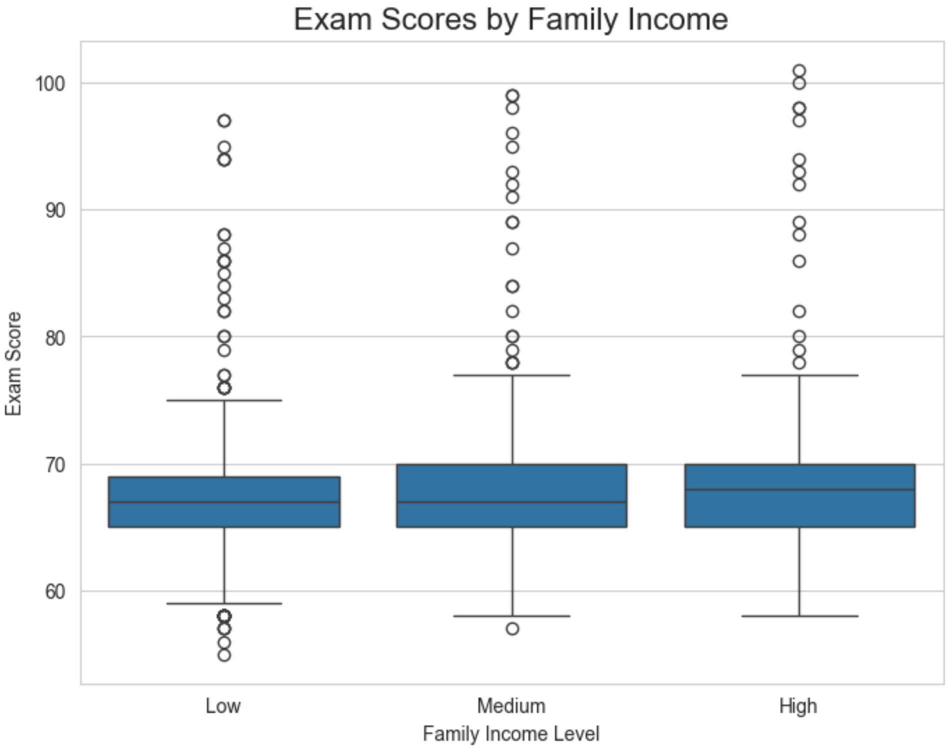

Figure 10: Exam Scores and Family Income

Interpretation:

- If p-value is more than 0.05: Fail to reject null hypothesis → Family income does not significantly affect exam scores.
- If p-value is less than 0.05: Family income significantly affects exam scores.

Result:

- ANOVA result: F-statistic = 29.79386131613895, P-value = 1.3143686049770217e-13
- Reject the null hypothesis: Family income significantly affects exam scores.

4.8 Hypothesis 8: Students attending private schools perform better than those in public schools.

- **Null Hypothesis (H_0)**: There is no significant difference in exam scores between private and public school students.
- **Alternative Hypothesis (H_1)**: There is a significant difference in exam scores between private and public school students.

A t-test compared exam scores between students in public and private schools. A boxplot helped to visualize the comparison.

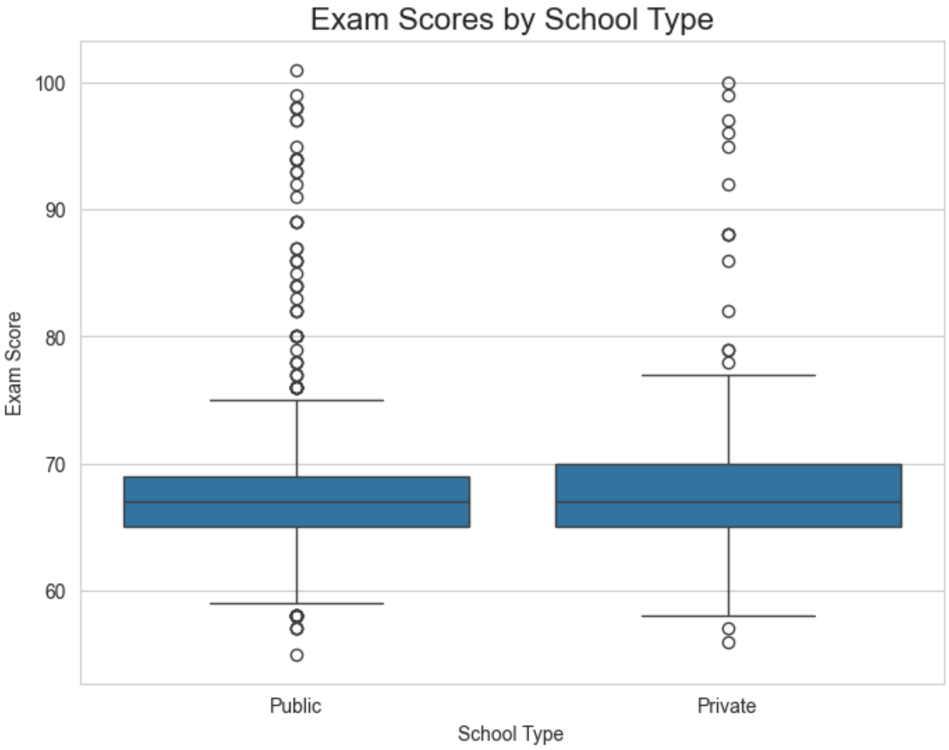

Figure 11: Exam Scores by School Type

Interpretation:

- If p-value is more than 0.05: Fail to reject null hypothesis → School type does not significantly affect exam scores.
- If p-value is less than 0.05:School type significantly affects exam scores.

Result:

- T-statistic: 0.7187537041931684, P-value: 0.47231811262174417
- Fail to reject the null hypothesis: School type does not significantly affect exam scores.

4.9 Hypothesis 9: Tutoring sessions have no significant effect on exam scores.

- **Null Hypothesis (H_0)**: Attending tutoring sessions does not significantly affect exam scores.
- **Alternative Hypothesis (H_1)**: Attending tutoring sessions significantly affects exam scores.

A Pearson correlation test analyzed the relationship between tutoring sessions and exam scores. A scatter plot helped to visualize this correlation.

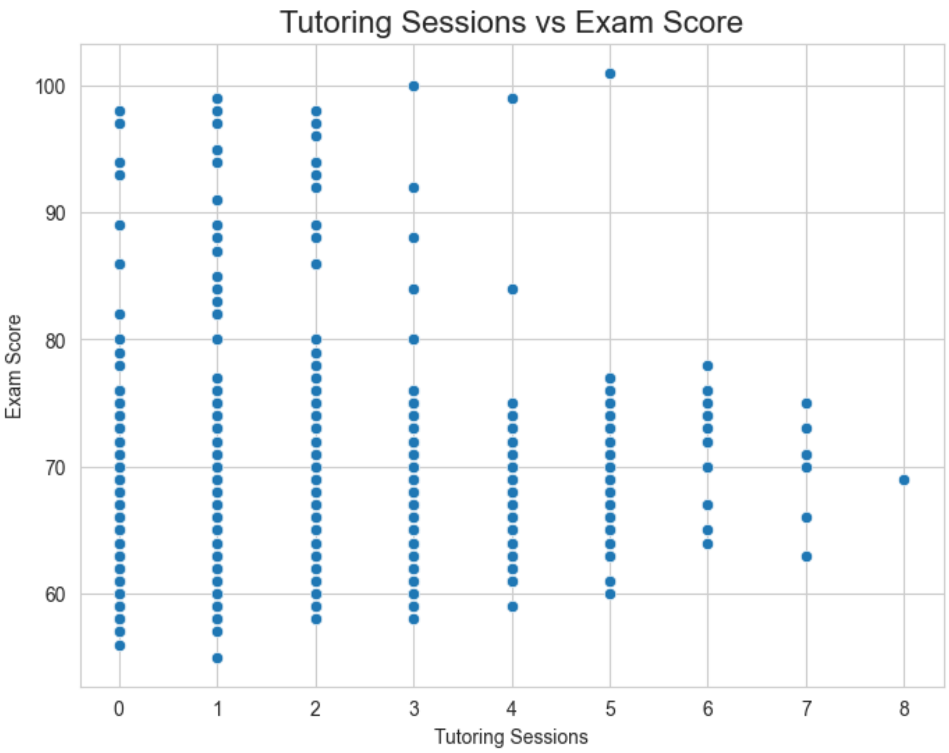

Figure 12: Exam Scores and Tutoring Sessions

Interpretation:

- If p-value is more than 0.05: Fail to reject null hypothesis → Tutoring does not significantly affect exam scores.
- If p-value is less than 0.05: Tutoring significantly affects exam scores.

Result:

- Correlation: 0.1565251853922532, P-value: 1.6508174156274133e-37
- Reject the null hypothesis: Tutoring significantly affects exam scores.

5 Conclusion

Through this comprehensive analysis, we explored various factors influencing student exam performance using statistical hypothesis testing and visualizations. We carefully designed and conducted nine tests, ensuring that all analyses were performed correctly and systematically. Our study examined the impact of critical variables such as study habits, sleep duration, parental involvement, extracurricular activities, internet access, gender, family income, school type, and tutoring sessions on exam scores.

By employing statistical methods including t-tests, ANOVA, and correlation tests, we accurately determined the significance of each variable's effect on performance. The p-value interpretations guided our decision to accept or reject hypotheses, yielding meaningful insights into the factors contributing to student success.

Our findings emphasize the crucial role of study time, parental support, and access to essential resources in achieving academic success. This research provides valuable evidence to inform educational practices and highlights areas where targeted interventions can positively impact student performance.

6 Future Work

Future research could include more features, larger datasets, or machine learning models for deeper insights into the factors affecting student performance.

7 References

- Kaggle. "Student Performance Factors Data Set." Available at: `https://www.kaggle.com/datasets/lainguyn123/student-performance-factors`.
- NumPy Documentation. "NumPy: The Fundamental Package for Scientific Computing with Python." Available at: `https://numpy.org/doc/`.
- Matplotlib Documentation. "Matplotlib: Python Plotting." Available at: `https://matplotlib.org/stable/contents.html`.
- Seaborn Documentation. "Seaborn: Statistical Data Visualization." Available at: `https://seaborn.pydata.org/`.