Modeling Student Performance: Using Multiple Linear Regression to Predict Exam Scores

Victoria Blante[1], Chelyah Miller[2], Xiaonan Song[3] and Emma Juan Salazar[4]

[*]University of San Francisco, Master of Science in Data Science

[1]vbblante@dons.usfca.edu
[2]cmiller9@dons.usfca.edu
[3]xsong20@dons.usfca.edu
[4]ejuansalazar@dons.usfca.edu

Abstract

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Keywords: Multiple Linear Regression, Modeling
All of the scripts and data for this project can be found on our Git Repository.

1 Introduction

Our Dataset

Our selected dataset is Student Performance Factors.

The "Student Performance Factors" dataset contains 19 variables that may influence students' exam scores. It is designed to help researchers analyze the potential impact of these factors on student performance. The dataset includes information such as study time (`Hours_Studied`), attendance (`Attendance`), parental involvement (`Parental_Involvement`), access to resources (`Access_to_Resources`), and participation in extracurricular activities (`Extracurricular_Activities`). Additionally, it covers socioeconomic and background data such as family income (`Family_Income`), motivation level (`Motivation_Level`), tutoring sessions (`Tutoring_Sessions`), school type (`School_Type`), sleep hours (`Sleep_Hours`), and parental education level (`Parental_Education_Level`).

Purpose and Applications

Researchers can use this dataset to build regression models for predicting exam scores (`Exam_Score`) and to identify significant factors affecting student academic performance. The dataset's potential applications include:

- Supporting educational decision-making
- Assisting in policy formulation
- Optimizing the allocation of educational resources

Ultimately, the goal is to better understand and improve the key factors influencing student success, thereby enabling educators and policymakers to provide more targeted support.

Variable Descriptions

- `Hours_Studied`: Daily study hours.
- `Attendance`: Attendance rate (percentage).
- `Parental_Involvement`: Parent involvement (Low, Medium, High).
- `Access_to_Resources`: Resource accessibility (Low, Medium, High).
- `Extracurricular_Activities`: Participation in extracurricular activities.
- `Sleep_Hours`: Daily sleep hours.
- `Previous_Scores`: Prior exam scores.
- `Motivation_Level`: Motivation level (Low, Medium, High).
- `Internet_Access`: Internet access.
- `Tutoring_Sessions`: Number of tutoring sessions.
- `Family_Income`: Family income level (Low, Medium, High).
- `Teacher_Quality`: Teacher quality (Low, Medium, High).

- School_Type: School type (Public or Private).
- Peer_Influence: Peer influence (Positive, Neutral, Negative).
- Physical_Activity: Weekly physical activity hours.
- Learning_Disabilities: Presence of learning disabilities.
- Parental_Education_Level: Parents' education level (High School, College, Postgraduate).
- Distance_from_Home: Distance from home to school (Near, Moderate, Far).
- Gender: Student gender (Male or Female).
- Exam_Score: Academic performance indicator (exam score).

2 Methods

We used various methods to perform our analysis, from EDA all the way to Model Evaluation. Here's a detailed description of the methods and tools we used.

- **Exploratory Data Analysis**: Initial analysis includes correlation calculations to understand the relationships between predictors and exam scores.
- **Multiple Linear Regression**: Regression models are built using significant predictors such as attendance, hours studied, and previous scores. The model is validated using metrics like adjusted R-squared, p-values, and F-statistics.
- **ANOVA (Types I, II, and III)**: Variance analysis is conducted to understand the contribution of each predictor to the total variance in exam scores.
- **Model Evaluation**: The model's prediction capability is visualized through plots of actual vs. predicted exam scores, residuals distribution, and summary statistics.

3 Exploratory Data Analysis

The Student Performance Factors dataset has 20 variables and 6607 observations. We have 7 numerical variables, including hours studied, attendance, sleep hours, previous scores, tutoring sessions, physical activity and exam score. The other 13 variables are categorical, having either 2 or 3 categories each. This gives us enough room to be able to perform a comprehensive Multiple Linear Regression Analysis.

Before delving into our dataset, we checked for missing data. Only three of our columns had missing values (Teacher_Quality, Parental_Education_Level and Distance_from_Home). Each of these columns has less than 2% of missingness. Seeing as we have a substantial number of observations, we decided to go ahead and drop all observations with missing data (229 rows).

Our dataset size after this was of 20 variables and 6378 records.

Categorical variables

Our first approach to get a sense of our categorical data was to plot the frequency for each category. See these plots in Supplementary Figure 1.

These plots allowed us to see that some categorical variables had similar shaped barplots. To optimize our categorical variables we drew up confusion tables to ascertain wether or not to collapse some of these variables.

We made confusion tables to check Parental_Involvment against Access_to_Resources (Supplementary Table 1), Family_Income against Peer_Influence (Supplementary Table 2), Distance_from_Home against Motivation_Level (Supplementary Table 3) and Distance_from_Home against Motivation_Level (Supplementary Table 3)

These confusion tables aided us to make the decision to drop the variables Distance_from_Home, Peer_Influence and Parental_Involvment.

Numerical Variables

As a first approach, for the numerical variables, we drew boxplots to check their distributions and wether or not we see any outliers (Supplementary Figure 2)

From these boxplots, we saw that for tutoring sessions we have a few outliers in the upper levels but most of the observations are clustered around 1/2 tutoring sessions. We can also see that for exam score, most of the students in our dataset have scores clustered around 65 out of 100, and we have a few high performing students. We'll have to keep these in mind when picking our response variable.

Next, we decided to make a pairwise plot (Supplementary Figure 3), a very powerful tool to try and get a first glance at which numerical variables seem to have linear relationships between one another.

A lot of our variables numerical variables are discreet and therefore the scatterplots show very distinct levels. Aside from these, there's a few scatterplots where we can see some sort of linearity between two variables. We can very clearly see that all of these plots include the variable Exam_Score.

Looking at the histogram for Exam Score inside of the pairwise plot, we can very clearly see how it's clustered around lower scores, therefore indicating we don't have a normal distribution for these. We'll have to keep this fact in mind when analyzing this variable.

Lastly, we made a correlation heatmap (Figure 1), to assess how are variables are correlated between one another and get a first glance towards multicollinearity.

We can see how the variable that has the most correlation to other variables is Exam Score, and how there's next to no correlation among the other variables. This indicates that picking Exam Score as our response variable seems a prudent choice, as it looks to be correlated to our other variables and also makes it so we have lower chances of having multicollinearity.

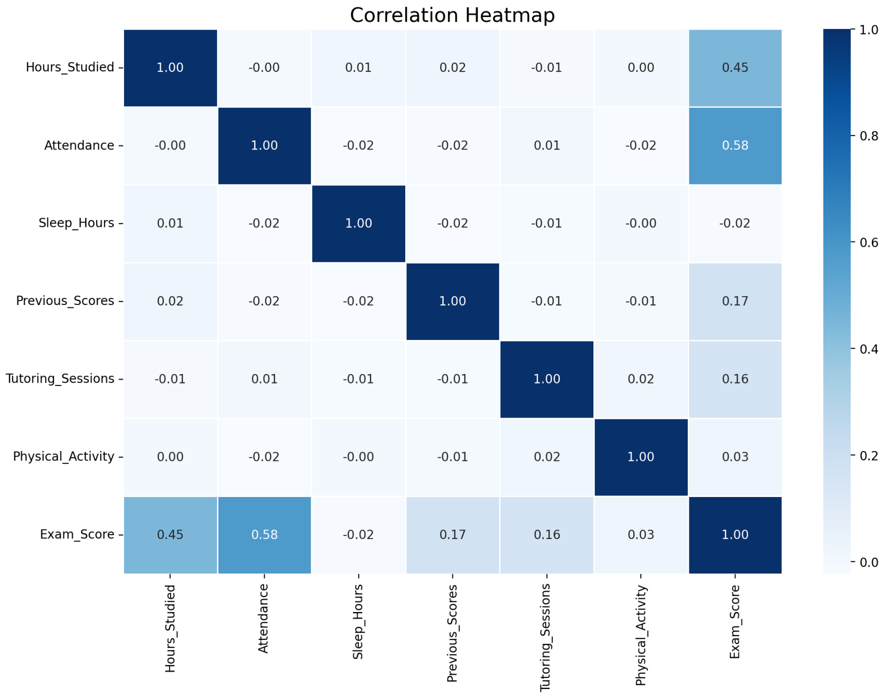


Figure 1: Heatmap depicting the correlation between our numerical variables.

4 Hypotheses

Hypothesising based on our EDA

Since we were able to see some linear relationship between a few of our numerical variables and `Exam_Score`, we have decided to pick this variable as our response variable.

The correlation heatmap also supports our choice of response variable, where we can see correlation between exam scores and the rest of the variables.

In the context of our problem, modeling student performance, we think choosing exam scores as our response variable is a good way to ascertain how students perform, and which factors influence the way they score in tests.

We have initially seen linearity in the scatterplots between exam score and hours studied, attendance, previous scores and tutoring sessions. We'll try to assess wether or not these numerical variables are statistically significant when predicting exam scores.

In regards to our remaining categorical variables after discarding based on confusion tables, we'll check for significance for all of them and try to ascertain which variables are best to include in our model with the goal of predicting exam scores.

As previously stated, the correlation heatmap seems to indicate we don't have severe multicollinearity. Further checks need to be done to ensure this.

We have to consider the fact that Exam Scores are clustered around 65 out of 100, and we have what we could consider some outliers in the higher ranges. Based on this, we think we may encounter that the high scores have a different relationship to the predictors compared to the majority of the other scores, which are around the 65%.

Research Questions

Our main research question is the following:

- **Which factors are the most significant predictors of students' exam scores?**

We also have specific questions about our different variables and their impact on the student's performance.

Regarding our numerical variables:

- **What is the combined effect of study habits, attendance, previous scores and tutoring sessions on exam outcomes?**

Regarding our categorical variables:

- **How do socioeconomic factors like access to resources or family income impact student performance?**
- **Do other student habits like extracurricular activities, motivation level or internet access have an important effect on the student performance?**
- **Does the hours a student spends doing physical activity impact their performance in tests?**
- **Does wether or not a student have a learning disability systematically impact their exam performance?**
- **Does the level of education of their parents indicate the performance of a student in exams?**
- **Does school type or teacher quality significantly influence exam scores?**

5 Multiple Regression Analysis

Initial Model

Our first approach to ascertain which predictors to include in our MLR model was to make a Simple Linear Regression model for each of our predictors. We compared metrics for each of these models, like R^2, and picked the highest performing size 1 models to make an initial MLR model

Box-Cox Transformation

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Splitting the Data Into Two Models

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

6 Results

Low Scores Model

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

High Scores Model

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

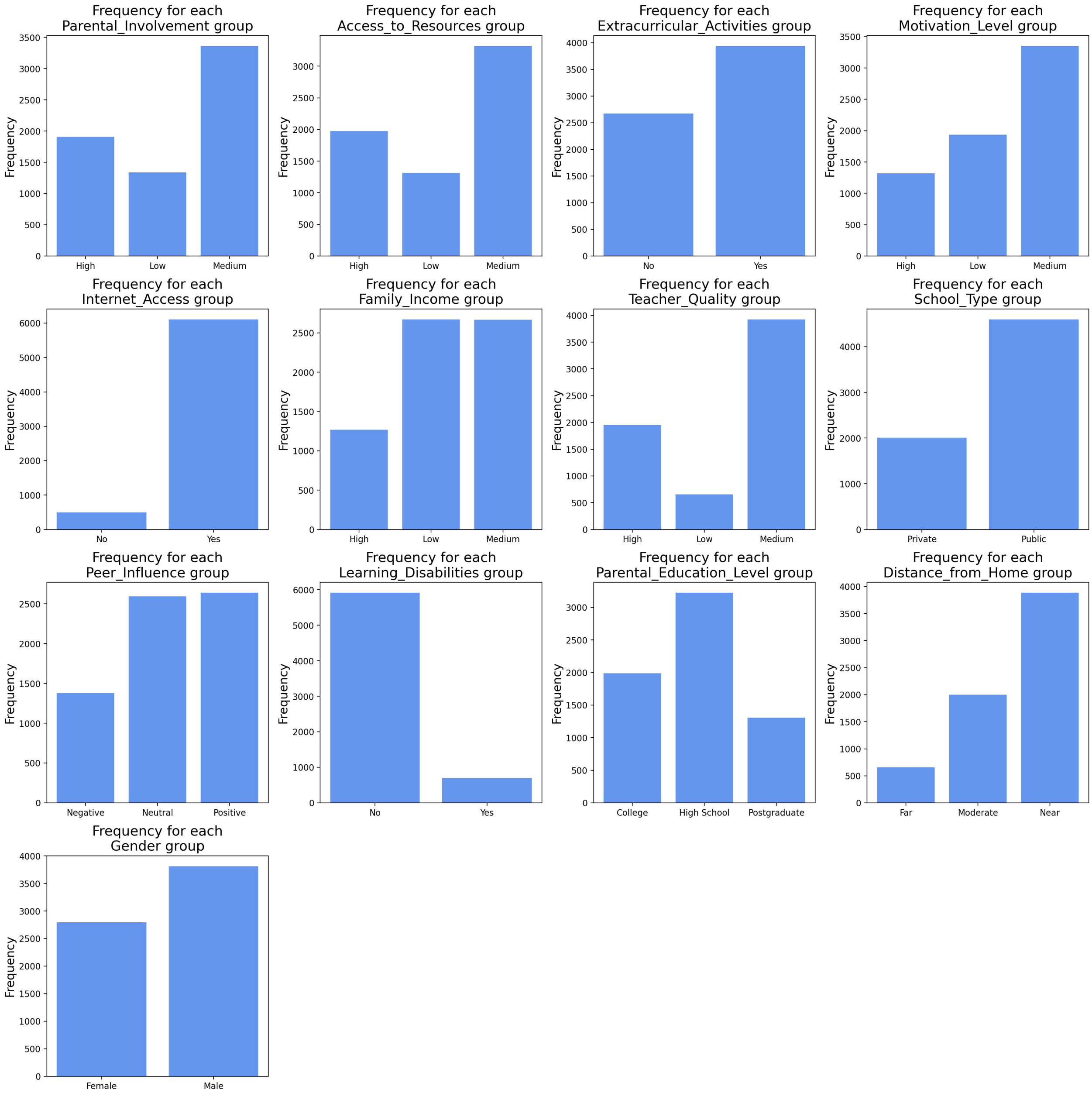
7 Discussion

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

8 Conclusion

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Supplementary Materials



Supplementary Figure 1: Plotting the frequency of each category for our categorical variables.

Parental Involvement/Access to Resources	**High**	**Low**	**Medium**
High	568	413	927
Low	414	231	692
Medium	993	669	1700

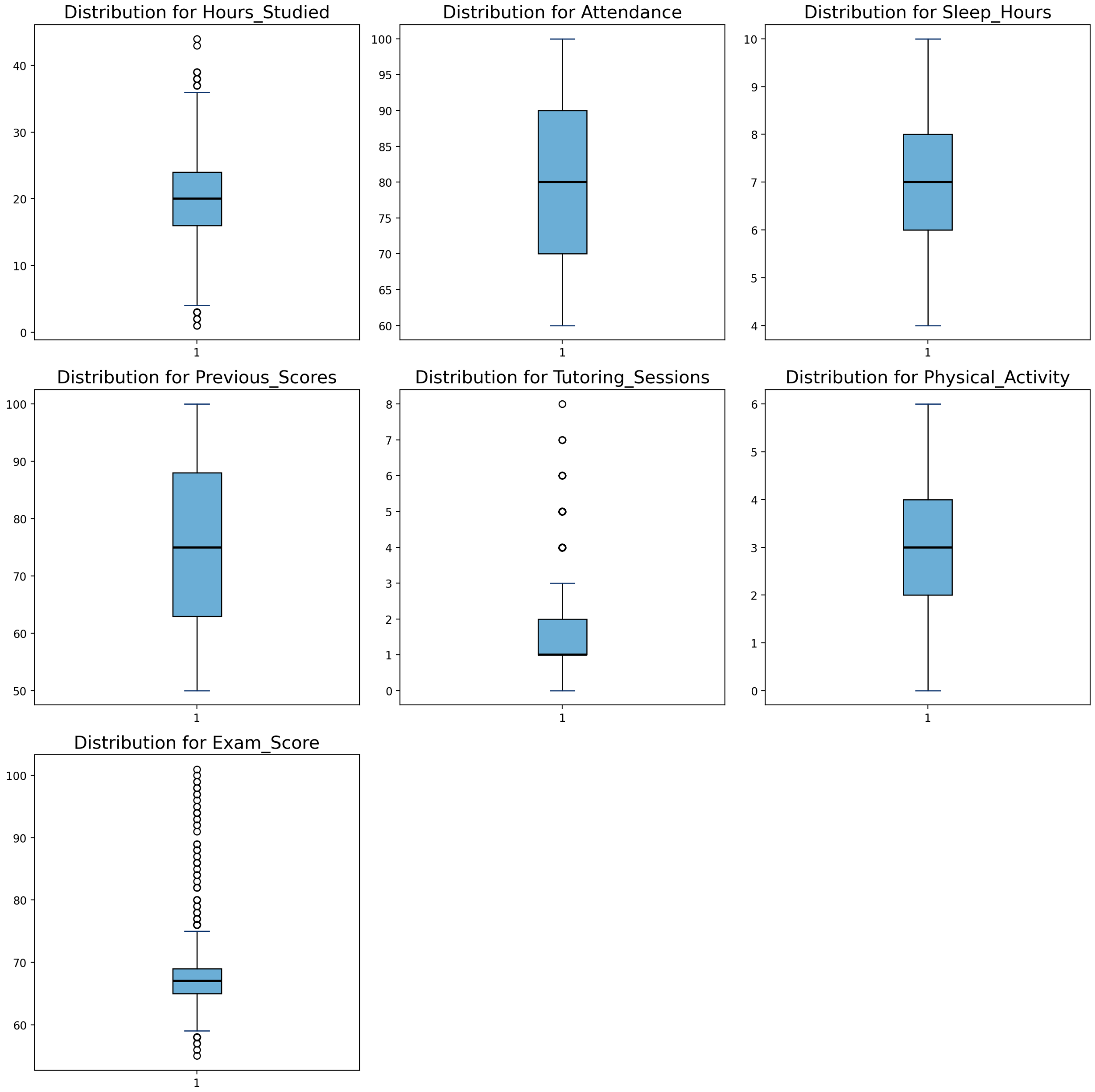
Supplementary Table 1: Confusion table: Parental Involvement vs. Access to Resources.

Peer Influence/Family Income	**High**	**Low**	**Medium**
Negative	251	577	549
Neutral	493	1038	1061
Positive	525	1057	1056

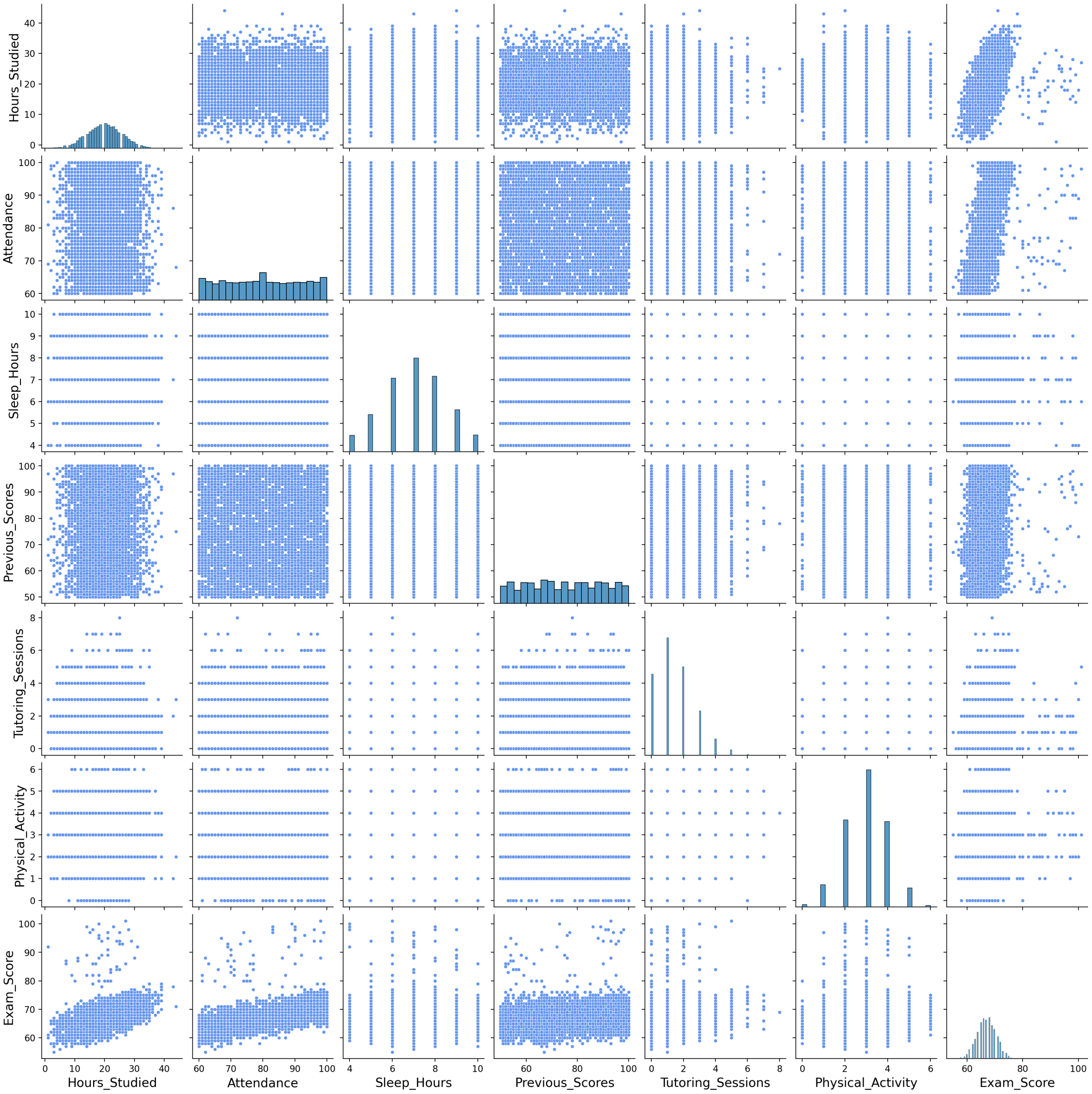
Supplementary Table 2: Confusion table: Family Income vs. Peer Influence.

Distance from home/Motivation Level	**High**	**Low**	**Medium**
Far	142	185	331
Moderate	394	611	993
Near	773	1125	1986

Supplementary Table 3: Confusion table: Motivation Level vs. Distance from Home.



Supplementary Figure 2: Plotting the distribution of each numerical variable.



Supplementary Figure 3: Numerical variable pairwise plot: Initial assessment of linear relationships between our variables.