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

This study explores the factors influencing student performance by applying multiple linear regression models to predict exam scores. Using a dataset of 6,378 students, we analyzed both numerical variables, such as hours studied, attendance, and previous scores, as well as categorical variables, including motivation level, access to resources, and family income. Due to clustering of exam scores around 65%, we split the data into two groups: high-performing students (scores ≥ 80*) and average-performing students (scores* < 80*). The average-performing model achieved an adjusted* R^2 *of 0.898, indicating strong predictive power, while the high-performing model had an adjusted* R^2 *of 0.602 due to limited data. Our results highlight that study habits, attendance, prior performance, and socioeconomic factors significantly impact academic outcomes.*

Keywords: Multiple Linear Regression, Modeling

1 Introduction

Our Dataset

The Student Performance Factors dataset contains 20 variables with information about 6607 students.

This dataset includes information about the different habits a student has, like hours studied, attendance, extracurricular activities, physical activity, or the amount of hours they sleep

We also have information about the student's environment including Parental Involvement, peer influence, school type, or teacher quality.

Moreover, we have information about the student themselves, like their gender, whether or not they have a learning disability, their motivation level, how many tutoring sessions they attend, their previous scores and their exam scores.

Additionally, it covers socioeconomic and background data such as family income, access to resources, internet access, and parental education level.

Purpose and Applications

Our aim is to identify a response variable that will be indicative of student performance, and build a regression model to identify significant factors affecting student academic performance. This analysis' 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 whether 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 whether 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 discrete 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.

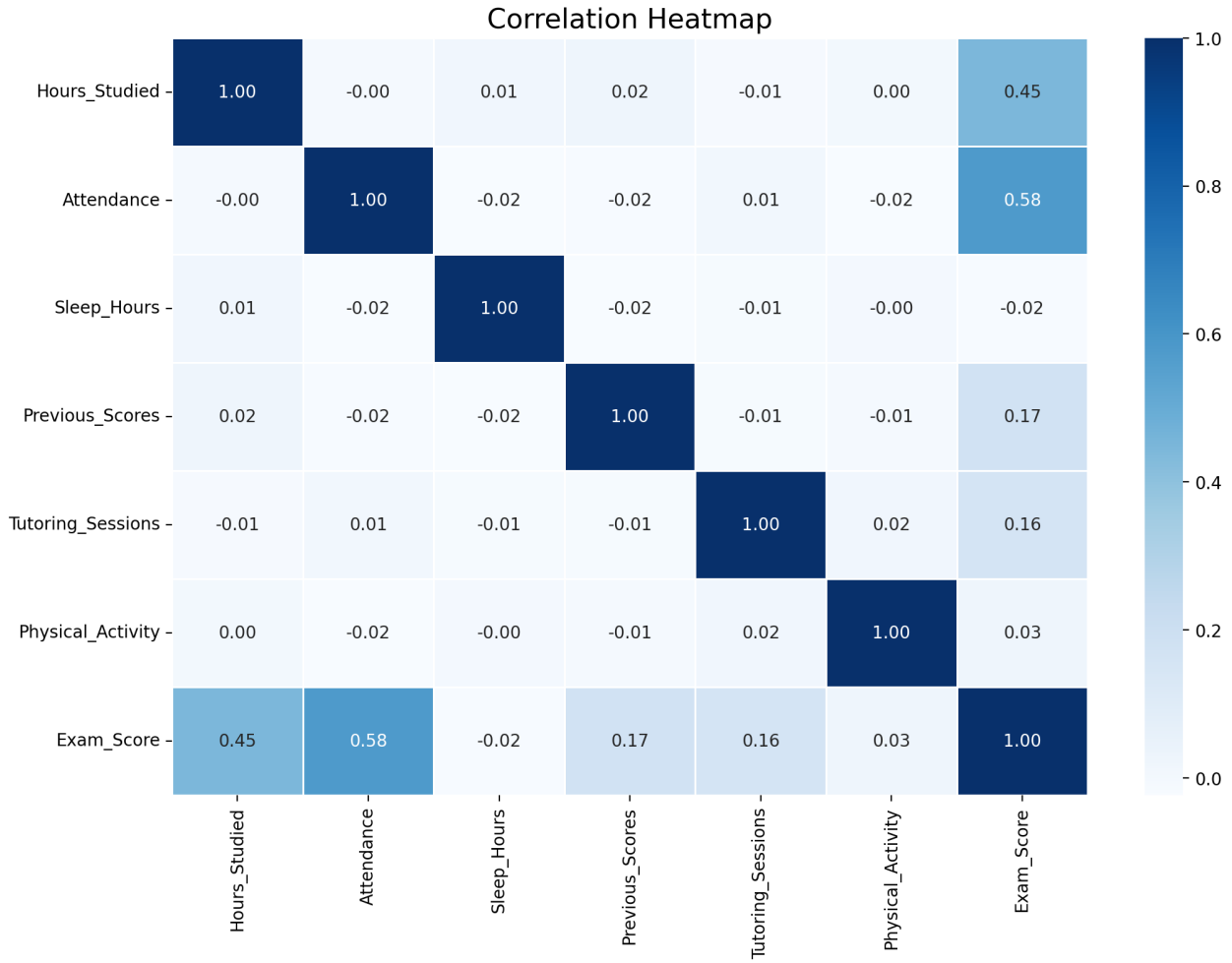


Figure 1: Heatmap depicting the correlation between our numerical variables.

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 whether 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 whether or not a student has 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 Linear 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 predictors based on their SLR models to make an initial MLR model.

After looking at the R_a^2 of exam score regressed onto each individual predictor, we're going to naively choose the 6 predictors with the strongest R_a^2 values to model: attendance, hours studied, previous scores, tutoring sessions, access to resources and learning disabilities.

Categorical variables were encoded using **one-hot encoding** to convert them into numerical format suitable for the models.

An **Ordinary Least Squares (OLS)** regression was performed on standardized predictors, calculating key statistics such as **R-squared**, **adjusted R-squared** and **p-values**.

The summary table for this initial model can be seen at Supplementary Figure 4.

This initial model showed an adjusted R^2 of 0.595, and together with other metrics and diagnostic plots confirmed this was an altogether not very good performing model.

We can see how plotting our predicted scores vs. the observed exam scores (Figure 2) shows observations where the residuals were too high and the predictions were not accurate at all.

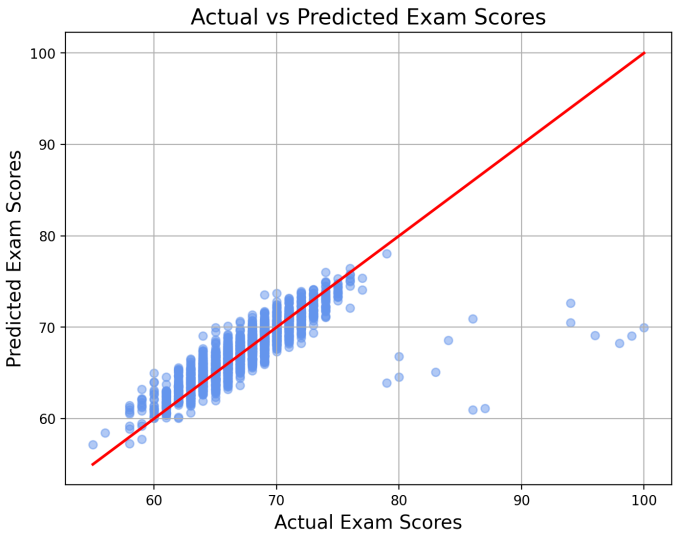


Figure 2: Predicted vs. observed exam scores

At this point, we considered whether the observations that are very far from our line are influential points. An influence plot (Figure 3) shows us how there are just too many of these observations that fit the criteria to be considered an influential point (leverage higher than 3 times the mean leverage, statistically significant externally studentized residuals or Cook's distance higher than $\frac{4}{n}$, n being the sample size).

We cannot simply treat these as outliers and remove them. We seem to not be considering that there are two distinct groups of students that should have different models.

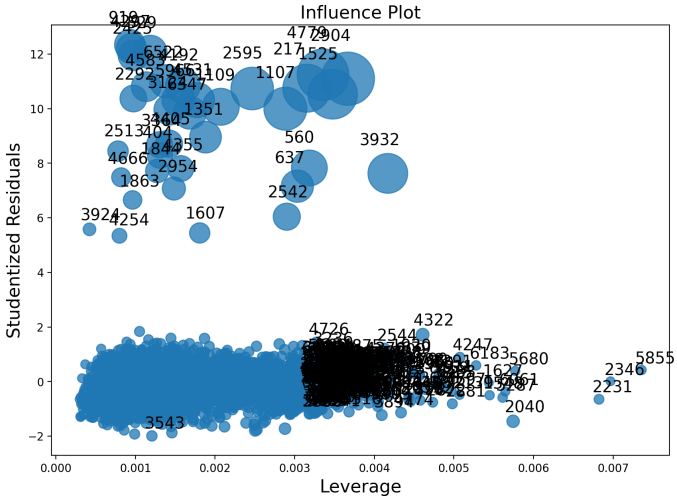


Figure 3: Influence plot for the initial model

To try and improve model fit, we performed a Box-Cox transformation on our model. The summary table for this initial model can be seen at Supplementary Figure 5. After this, even if our metrics had improved ($R_a^2 = 0.719$), our residual plot (Figure 4) showed that our strategy for modeling was not having success.

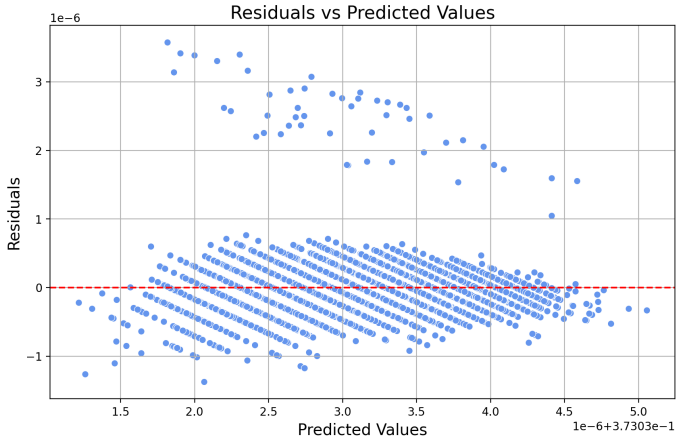


Figure 4: Residual plot for the initial model.

The residual plot showed that there still seemed to be a significant amount of data falling outside the expected bounds. We can clearly see that a chunk of our predictions have very high residuals, following what we could call a pattern.

This residual plot, an influence plot, and our discoveries regarding the Exam Scores variable during EDA, lead us to consider splitting the data.

Splitting the Data Into Two Models

Since most of our observations have their exam scores clustered around 65%, and we find a set of high performing students, we modeled performance under the assumption that these two student groups respond differently to predictors. and the predictors don't impact them in the same way. We decided to divide our dataset into two groups:

- High Performing Students: Students with exam scores of 80 or above. (This will include students that achieve grades of B or higher).
- Average Performing Students: Students with exam scores under 80 (with grades from Fail until C+).

These two datasets were further divided into training and testing datasets.

We ended up with 6330 observations for the average performing students dataset, and 48 observations for the high performing students dataset. We acknowledge our high model will not have a high statistical power, seeing as training a model with only 48 observations limits accuracy.

We will work extensively on our average performing students model, which includes the majority of the students on our dataset. We will also fit a model for the high performing students dataset, but since this is an under represented group in our dataset, we'd need more data to correctly assess student performance for students with high exam scores.

Our new approach to choose predictors involved the Lasso pipeline. LASSO (Least Absolute Shrinkage and Selection Operator) is an effective technique for handling situations with a large number of predictors and potential multicollinearity. It improves model performance by adding an L1 regularization term to the regression model, which penalizes the absolute size of the coefficients. This penalization has two key effects: first, it shrinks some coefficients toward zero, reducing the variance of the model and helping to prevent overfitting. Second, LASSO sets some coefficients exactly to zero, effectively selecting a subset of the most important predictors. By forcing less significant predictors to have a coefficient of zero, LASSO simplifies the model and ensures it retains only the most relevant features, thus improving interpretability and generalization to new data.

6 Results

Low Scores Model

The Lasso pipeline chose 10 predictors for this model:

- Five categorical variables: access to resources, extracurricular activities, motivation level, family income, and parental education level.
- Five numerical variables: hours studied, attendance, previous scores, tutoring sessions, physical activity.

The summary table for the final **Low model** can be seen in Supplementary Table 6.

We can reference different model selection criteria to make sure this model is better and more accurate than our initial model (Table 1)

Model	adj-R^2	AIC	BIC
Initial Model	0.600	2.077e+04	2.081e+04
Box-Cox Model	0.726	-1.718e+05	-1.718e+05
Low Model	0.898	1.511e+04	1.521e+04

Table 1: Model selection criteria: Initial model vs. Low model

For this model, we achieve an adjusted R-squared of 0.898, and the residual plot (Figure 5) looks to be correctly scattered around the x-axis.

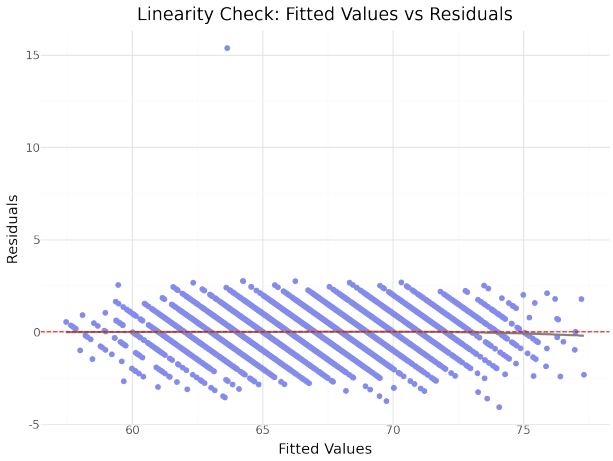


Figure 5: Residual plot for the low model.

The QQ-plot (Figure 6) tells us the same story. We can see how most of our residuals almost perfectly fall on the regression line. This helps us confirm our model isn't violating any model assumptions, like the normality of the residuals.

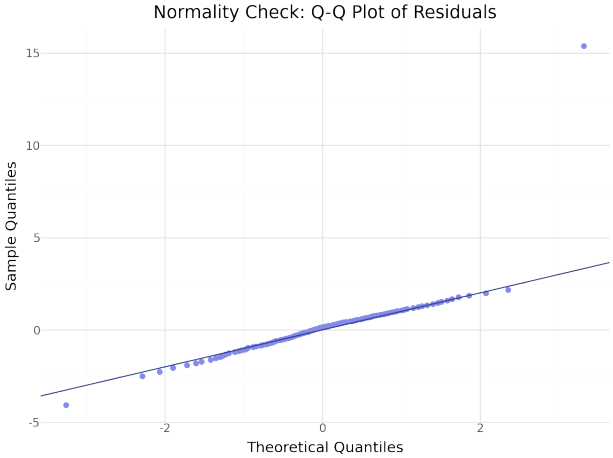


Figure 6: QQ-plot for the low model.

Most of the data falls in the expected region, but we still see a few potential outliers. We decided to explore influential points to get a better look at these potential outliers.

Other diagnostic plots, like scatterplots of individual predictors vs response variable (Supplementary Figure 7) and correlation matrix (Supplementary Figure 8) can be seen on our supplementary material.

Influential Points

To get a sense for the potential problematic observations in our data, we used three metrics: leverage, to get out-

liers, externally studentized residuals, to identify high discrepancy points, and Cook's distance, to definitely identify influence points.

In Table 2 we can see that none of these metrics are high enough to flag the observations as influential points. Since this has no significant effect on our model, we'll leave them.

Index	Leverage	Cook's Distance	Studentized Residual
5043	0.007464	5.039074×10^{-7}	-0.031705
53	0.006856	0.0004598461	-0.999592
1002	0.006809	0.0005532396	1.100212

Table 2: Leverage, Cook's Distance, and Studentized Residual for Selected Observations

Finally, we plotted our line of best fit with the residuals (Figure 7), to get a sense of how our model is performing.

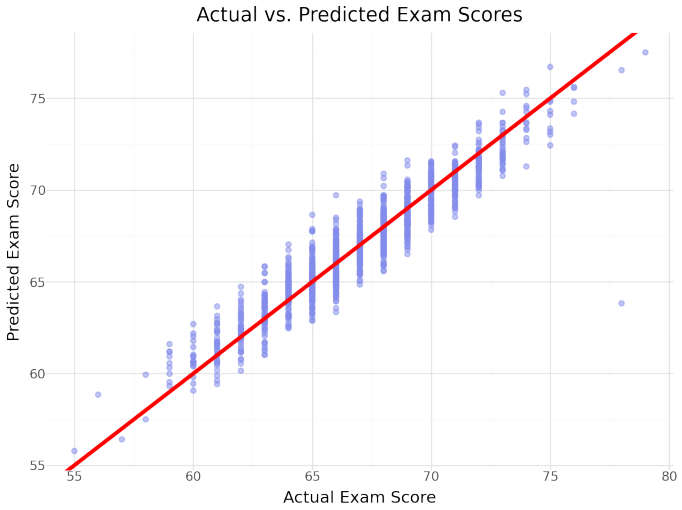


Figure 7: Residual plot for the low model.

We can see how all of the residuals fall in line with the line of best fit.

High Scores Model

The Lasso pipeline chose 11 predictors for this model:

- Six categorical variables: access to resources, internet access, family income, teacher quality, school type and parental education level.
- Five numerical variables: hours studied, attendance, sleep hours, previous scores and tutoring sessions.

The summary table for the final **High model** can be seen in Supplementary Table 9

As expected, the model's performance suggests that the available data may not be sufficient for achieving high accuracy. The **adjusted R-squared** value of $R_a^2 = 0.602$ indicates that approximately 60.2% of the variance in the response variable (exam score) is explained by the predictors in the model. While this suggests some predictive power, it is not particularly high.

The **AIC** of 229.6 and **BIC** of 255.8 are relatively high, suggesting that the model may not fit the data very well and that there could be some overfitting or inefficiency in the model, especially considering the limited number of observations (38).

Our **p-values** for most of the individual predictors are quite large, some being as high as 0.870 (for physical activity). These high p-values suggest that many of the predictors are not statistically significant at common significance levels (such as $\alpha = 0.05$). This implies that, individually, these predictors do not have a strong relationship with the response variable.

However, despite the lack of significance for individual predictors, the **F-test** p-value of 0.000541 suggests that the model as a whole is statistically significant. This means that we can reject the null hypothesis that none of the predictors have any linear relationship with the response variable. In other words, while individual predictors may not show strong significance, the collective set of predictors likely explains some of the variance in exam score.

The diagnostic plots this time have less data to work with, but still don't show violation of our model assumptions. Our residuals seem to be randomly scattered around 0 (Figure 8), but we can also see that the variance it shows is very high, again confirming this model can have fairly high errors.

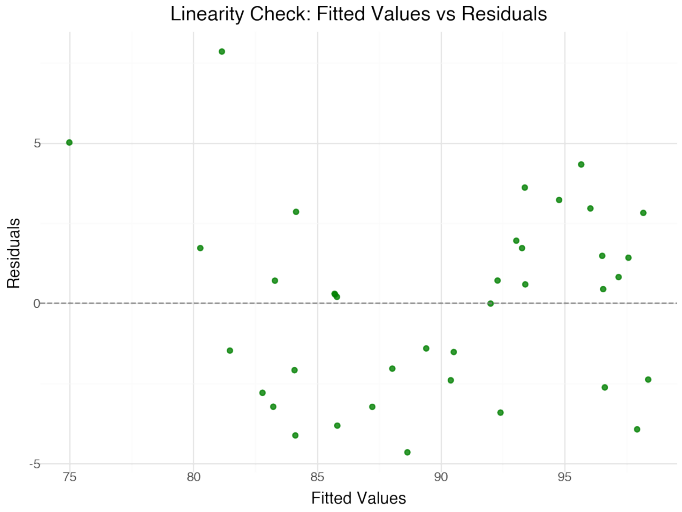


Figure 8: Residual plot for the high model.

The QQ-plot (Figure 9) also shows our theoretical and observed quantiles coinciding for the most part, but the residuals don't fall so much on the line compared to the low model.

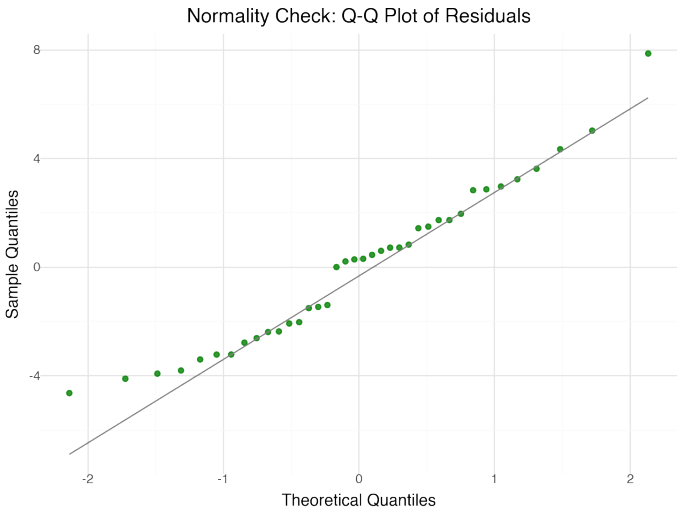


Figure 9: QQ-plot for the high model.

Other diagnostic plots, like scatterplots of individual predictors vs response variable (Supplementary Figure 10) and correlation matrix (Supplementary Figure 11) can be seen on our supplementary material.

In our plot for the actual vs. predicted exam scores for the high model (Figure **??**) we can see that our residuals are very high and therefore, as predicted, our model does not have a lot of predicting power.

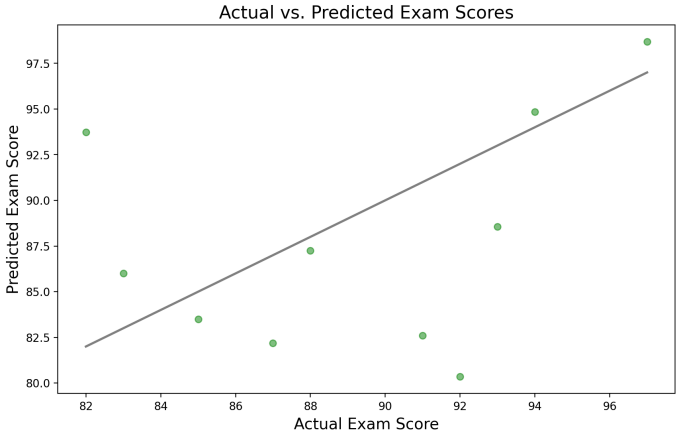


Figure 10: Line of best fit: Actual vs. Predicted Exam Scores

Gathering more data could help improve this particular model's accuracy and the significance of the individual predictors.

7 Discussion

Our final MLR models predict student performace, measured as Exam Scores, based on a set of numerical and categorical predictors. We developed two separate models: one for students with average scores (below 80) and another for high-performing students (80 or above). This split was necessary due to the clustering of most scores around 65%, with only a small subset of students in our dataset achieving high scores. After fitting an initial model with all of the observations, we could clearly see a different pattern between the predictors and the average scores vs. the high scores.

The adjusted R^2 for the average performance model was 0.898, indicating a strong predictive ability. In contrast, the high-performance model achieved a lower adjusted R^2 of 0.602, reflecting the limited predictive power for this group, due to the lack of data.

It is interesting to note which predictors ended up in each model (Table 3).

This selection can already answer some of our research questions. We can see how Learning Disabilities was not picked out for any of the models, indicating that it does not affect a student's performance.

Predictors like hours studied, attendance, previous scores, tutoring sessions, access to resources, family income and parental education level have an impact for both of our models. We can go ahead and call these factors the ones that have the most significant impact in a student's performance.

We see positive coefficients for hours studied and attendance, indicating that the more hours a student studies and the more they attend lessons, the higher they'll score on their exams.

Predictor	Low Model	High Model
Hours Studied	✓	✓
Attendance	✓	✓
Sleep Hours	✗	✓
Previous Scores	✓	✓
Tutoring Sessions	✓	✓
Access to Resources	✓	✓
Extracurricular Activities	✓	✗
Motivation Level	✓	✗
Family Income	✓	✓
Parental Education Level	✓	✓
Physical Activity	✓	✗
Learning Disabilities	✗	✗
School Type	✗	✓
Teacher Quality	✗	✓
Internet Access	✗	✓
Gender	✗	✓

Table 3: Predictor selection for the average and high performance models.

For the access to resources, a lower coefficient for the Low and Medium levels compared to the High level indicates how the less access a student has to resources, the more their performance will suffer. A similar behavior can be concluded for the family income variable. The lower the family income, the lower the student's exam scores. This answers our question about how socioeconomic factors pertaining to a student or their family impacts their performance. The lower the income and access to resources, the more the student's grades decrease.

In regards to a student's parents' education level, we can see how having chosen College as the baseline, our models indicate that for parental education level of High School the student's exam scores are on average 0.47 points lower, and for the Postgraduate level they're 0.53 points higher instead. This shows us how having educated parents is an indicator of a student performing better in exams.

We can see however that different predictors were chosen to model performance for our two groups of students.

Specific predictors for the Average Performing Students Model (Low Model)

For average performing students, we have predictors that don't seem to influence exam scores for high performing students, like extracurricular activities, motivation level, or physical activity.

Looking at our model, we can see how a student participating in extracurricular activities will make them score 0.46 better on average compared to a student who doesn't.

We can also clearly see how motivation level has a clear impact on exam scores, the coefficient for the dummy variable T.Medium indicating a decrease in score of 0.44, and for T.Low of 0.96 compared to students with motivation level high.

8 Conclusion

We were able to identify several factors that significantly impacted student performance, with the most significant predictors being hours studied, attendance, prior academic performance, and motivation. The high-performance

model, despite being less robust, also indicates that these same factors may influence students achieving top scores.

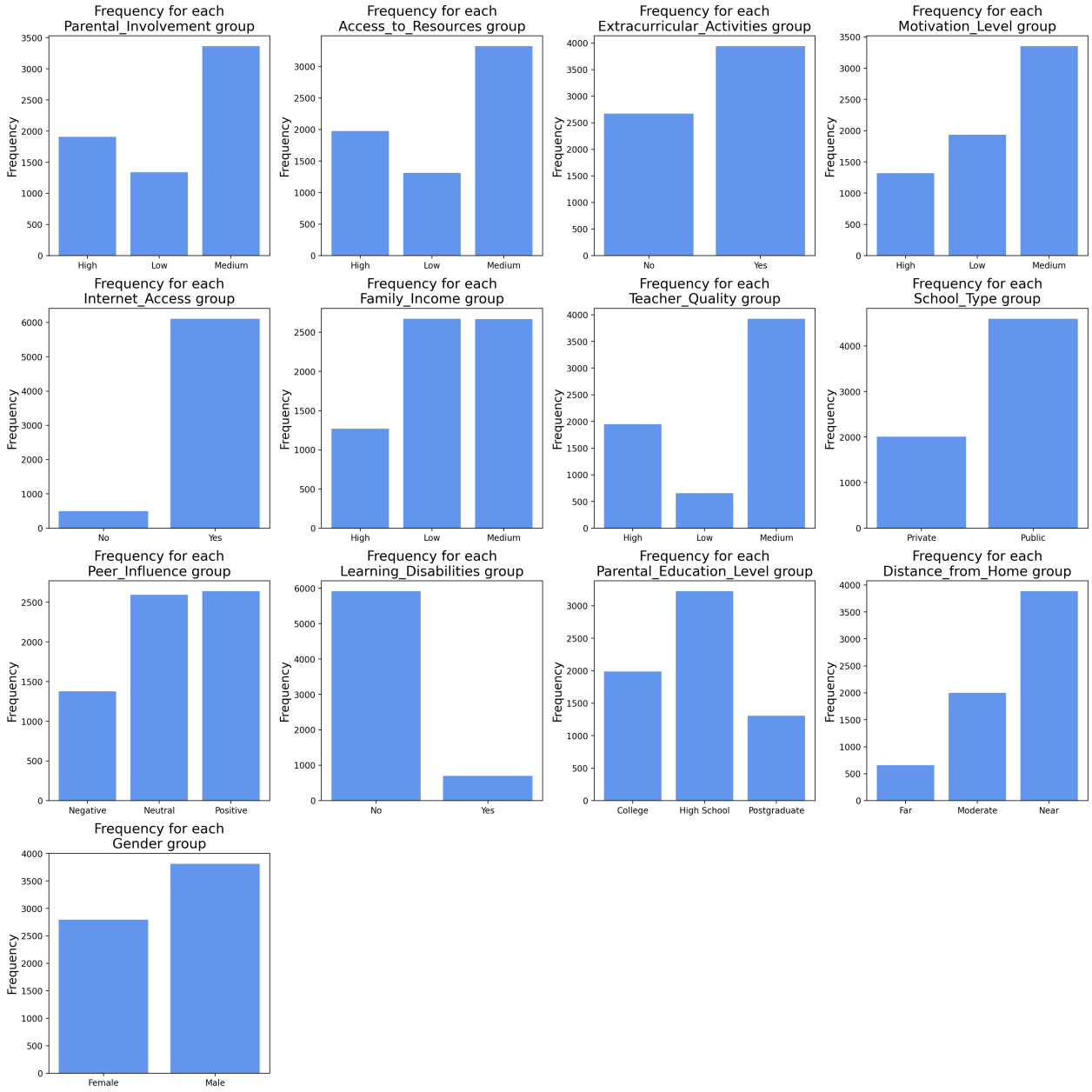
These findings also help to identify areas where educators and policymakers can create significant changes. By increasing study hours, boosting attendance, and providing tutoring sessions, educators will see an improvement in student exam scores. Additionally, factors like access to resources and parental education play a meaningful role, suggesting that there's another avenue for educators to identify students who may require early intervention.

Future research focusing on high-performing students would help to improve model accuracy and predictive power. Including other variables such as mental health or learning environments could also help to capture additional aspects of student performance. Overall, this underscores the value of data-driven strategies in education and provides a foundation into further exploration on how to optimize students' academic success.

9 Data and Code Availability

All of the code and data for this project can be found on our Git Repository.

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

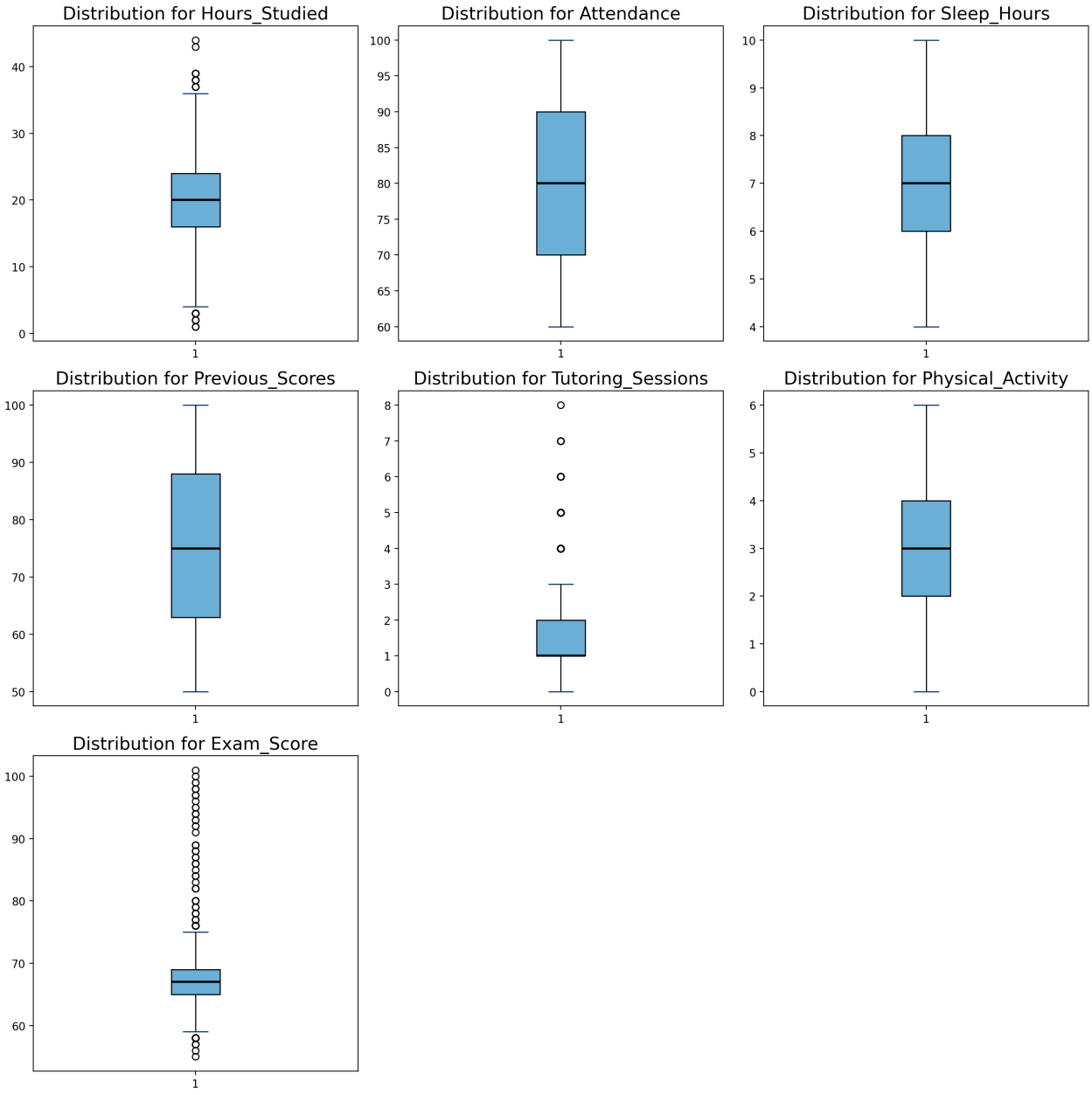
Supplementary Table 1: Confusion table: Parental Involvement vs. Access to Resources.

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

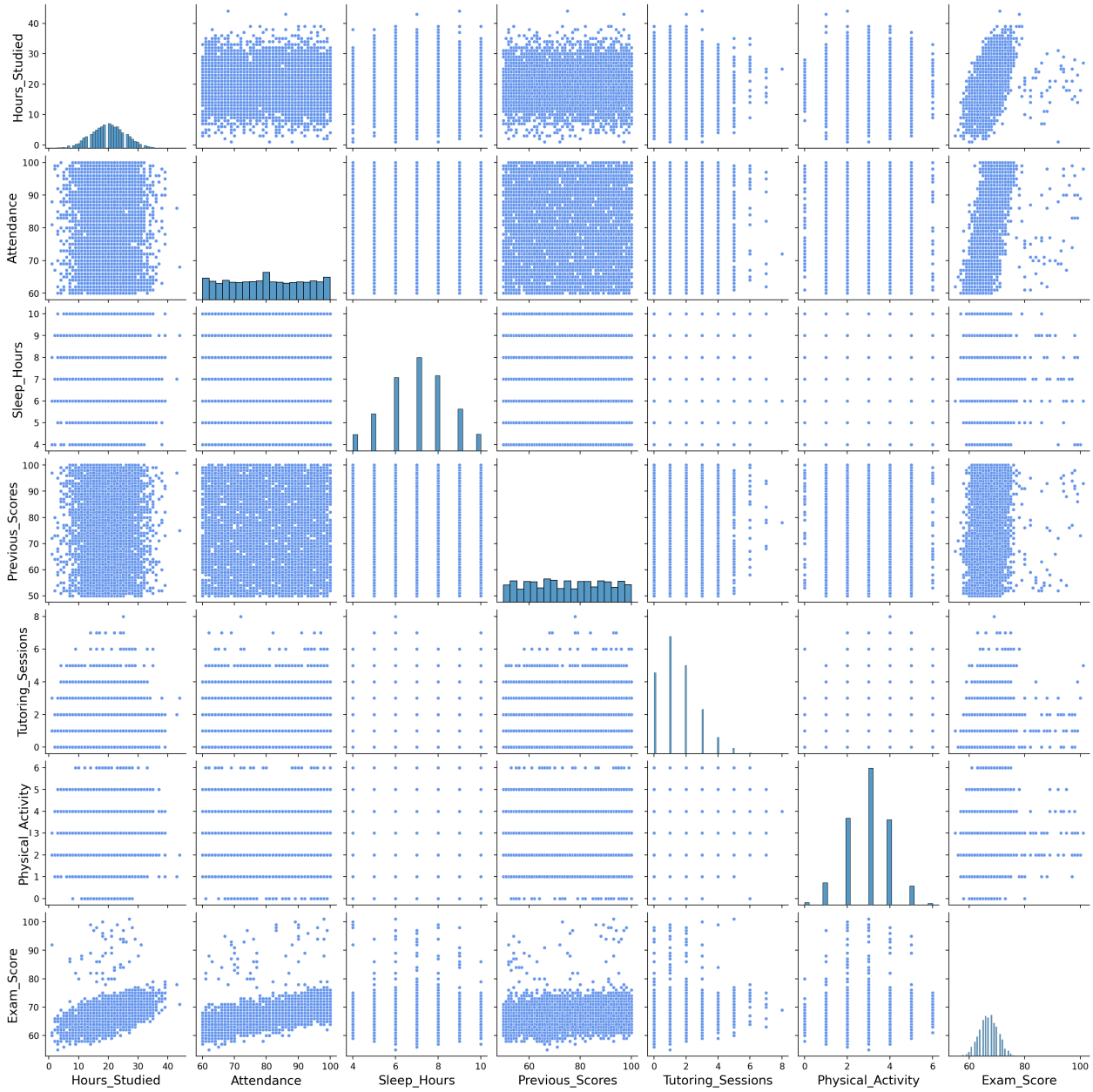
Supplementary Table 2: Confusion table: Family Income vs. Peer Influence.

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.

OLS Regression Results

Dep. Variable:	Exam_Score	R-squared:	0.601
Model:	OLS	Adj. R-squared:	0.600
Method:	Least Squares	F-statistic:	1117.
Date:	Fri, 11 Oct 2024	Prob (F-statistic):	0.00
Time:	19:14:45	Log-Likelihood:	-10378.
No. Observations:	4464	AIC:	2.077e+04
Df Residuals:	4457	BIC:	2.081e+04
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	67.2581	0.037	1815.017	0.000	67.185	67.331
x1	2.2684	0.037	61.190	0.000	2.196	2.341
x2	1.7595	0.037	47.445	0.000	1.687	1.832
x3	0.6932	0.037	18.689	0.000	0.620	0.766
x4	0.5972	0.037	16.108	0.000	0.524	0.670
x5	0.3913	0.037	10.554	0.000	0.319	0.464
x6	-0.3301	0.037	-8.905	0.000	-0.403	-0.257

Omnibus:	5636.291	Durbin-Watson:	1.996
Prob(Omnibus):	0.000	Jarque-Bera (JB):	918821.456
Skew:	6.918	Prob(JB):	0.00
Kurtosis:	71.909	Cond. No.	1.05

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Supplementary Figure 4: Summary of Regression Results for the Initial Model

OLS Regression Results

Dep. Variable:	y	R-squared:	0.726
Model:	OLS	Adj. R-squared:	0.726
Method:	Least Squares	F-statistic:	2814.
Date:	Fri, 11 Oct 2024	Prob (F-statistic):	0.00
Time:	19:15:00	Log-Likelihood:	85921.
No. Observations:	6378	AIC:	-1.718e+05
Df Residuals:	6371	BIC:	-1.718e+05
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.3730	4.28e-09	8.72e+07	0.000	0.373	0.373
x1	4.163e-07	4.28e-09	97.329	0.000	4.08e-07	4.25e-07
x2	3.176e-07	4.28e-09	74.239	0.000	3.09e-07	3.26e-07
x3	1.255e-07	4.28e-09	29.325	0.000	1.17e-07	1.34e-07
x4	1.095e-07	4.28e-09	25.609	0.000	1.01e-07	1.18e-07
x5	7.219e-08	4.28e-09	16.877	0.000	6.38e-08	8.06e-08
x6	-5.812e-08	4.28e-09	-13.590	0.000	-6.65e-08	-4.97e-08

Omnibus:	4705.241	Durbin-Watson:	1.997
Prob(Omnibus):	0.000	Jarque-Bera (JB):	189213.016
Skew:	3.100	Prob(JB):	0.00
Kurtosis:	28.953	Cond. No.	1.05

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Supplementary Figure 5: Summary of Regression Results for the Initial Model after Box Cox transformation

Dep. Variable:	Exam_Score	**R-squared:**	0.898
Model:	OLS	**Adj. R-squared:**	0.898
Method:	Least Squares	**F-statistic:**	3183.
Date:	Sat, 12 Oct 2024	**Prob (F-statistic):**	0.00
Time:	22:32:24	**Log-Likelihood:**	-7540.9
No. Observations:	5064	**AIC:**	1.511e+04
Df Residuals:	5049	**BIC:**	1.521e+04
Df Model:	14		
Covariance Type:	nonrobust		

(a) Model Summary Statistics

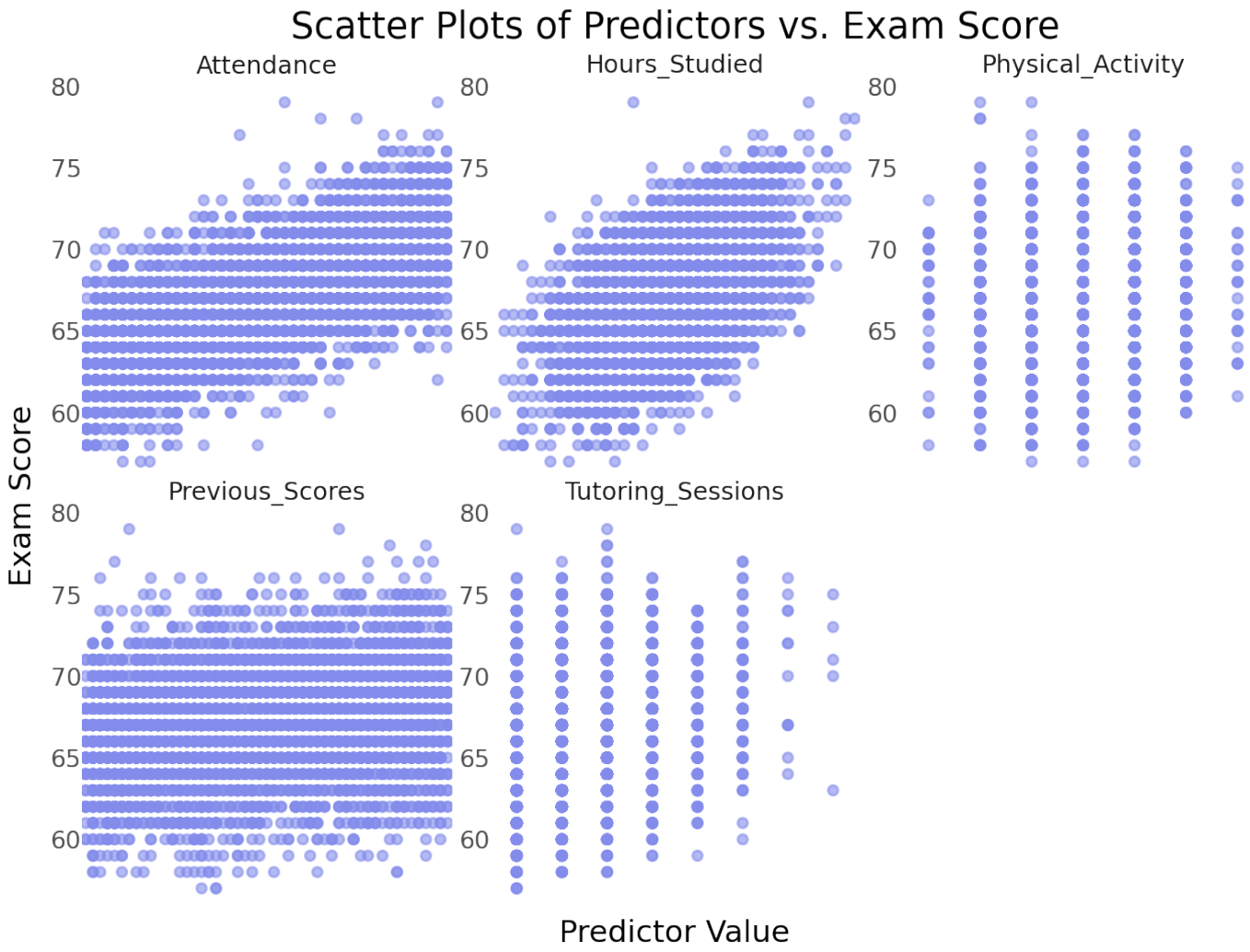
	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
Intercept	42.0979	0.158	266.026	0.000	41.788	42.408
Access_to_Resources[T.Low]	-1.9372	0.044	-44.274	0.000	-2.023	-1.851
Access_to_Resources[T.Medium]	-0.9681	0.035	-27.774	0.000	-1.036	-0.900
Extracurricular_Activities[T.Yes]	0.4688	0.031	15.212	0.000	0.408	0.529
Motivation_Level[T.Low]	-0.9550	0.044	-21.845	0.000	-1.041	-0.869
Motivation_Level[T.Medium]	-0.4403	0.040	-11.040	0.000	-0.519	-0.362
Family_Income[T.Low]	-1.0166	0.042	-24.447	0.000	-1.098	-0.935
Family_Income[T.Medium]	-0.4775	0.042	-11.470	0.000	-0.559	-0.396
Parental_Education_Level[T.High School]	-0.4723	0.035	-13.593	0.000	-0.540	-0.404
Parental_Education_Level[T.Postgraduate]	0.5296	0.043	12.201	0.000	0.444	0.615
Hours_Studied	0.2976	0.003	117.166	0.000	0.293	0.303
Attendance	0.1991	0.001	152.018	0.000	0.196	0.202
Previous_Scores	0.0470	0.001	44.784	0.000	0.045	0.049
Tutoring_Sessions	0.4998	0.012	40.552	0.000	0.476	0.524
Physical_Activity	0.2205	0.015	14.948	0.000	0.192	0.249

(b) Regression Coefficients

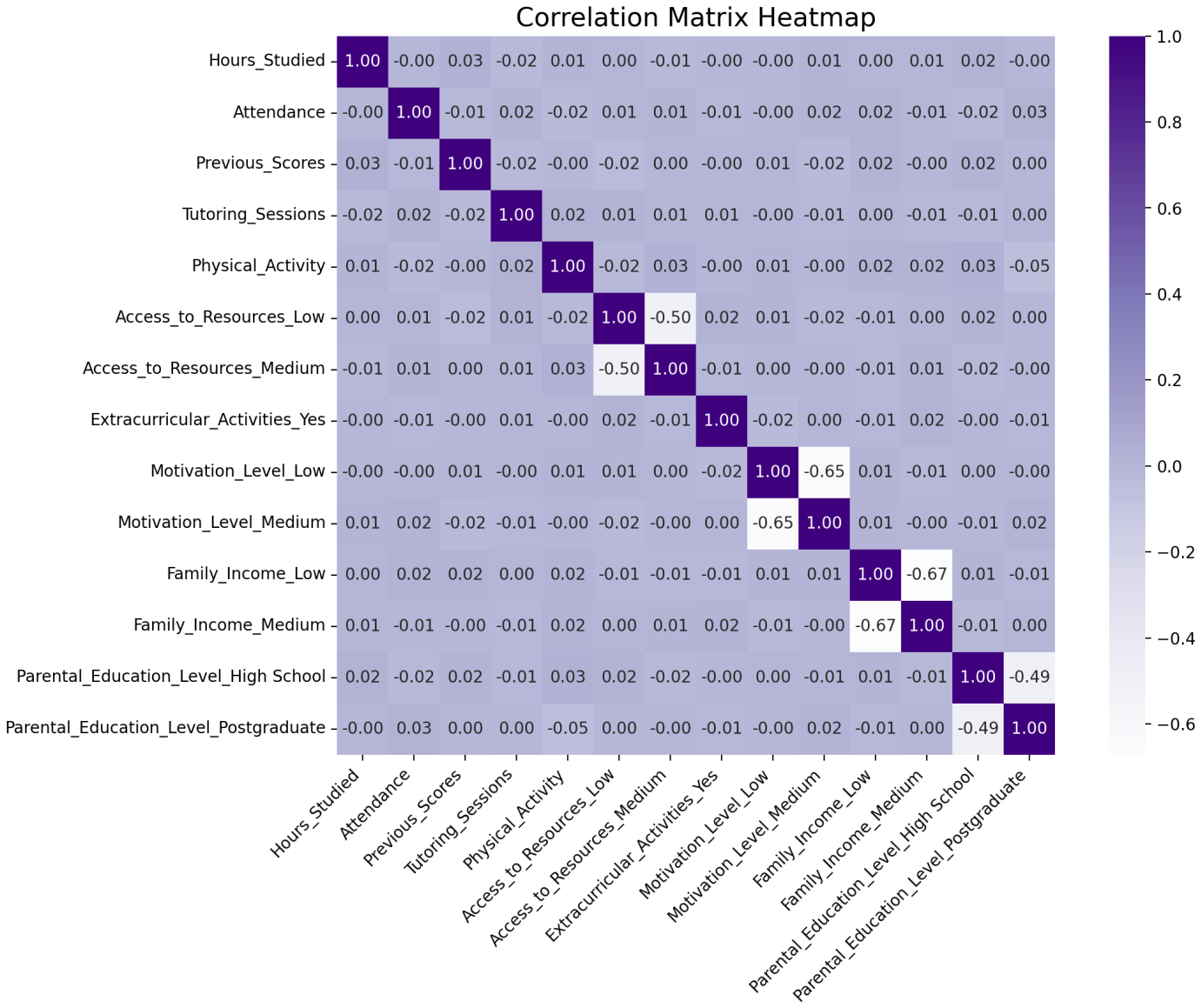
Omnibus:	872.171	**Durbin-Watson:**	1.966
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	13229.088
Skew:	0.352	**Prob(JB):**	0.00
Kurtosis:	10.887	**Cond. No.**	1.19e+03

(c) Model Diagnostics

Supplementary Figure 6: Summary of Regression Results for the Low Model



Supplementary Figure 7: Scatterplot of each predictor vs. Exam Scores for the Low Model



Supplementary Figure 8: Correlation Matrix Heatmap for the Low Model

Dep. Variable:	Exam_Score	**R-squared:**	0.817
Model:	OLS	**Adj. R-squared:**	0.677
Method:	Least Squares	**F-statistic:**	5.845
Date:	Sat, 12 Oct 2024	**Prob (F-statistic):**	0.000128
Time:	22:38:58	**Log-Likelihood:**	-93.933
No. Observations:	38	**AIC:**	221.9
Df Residuals:	21	**BIC:**	249.7
Df Model:	16		
Covariance Type:	nonrobust		

(a) Model Summary Statistics

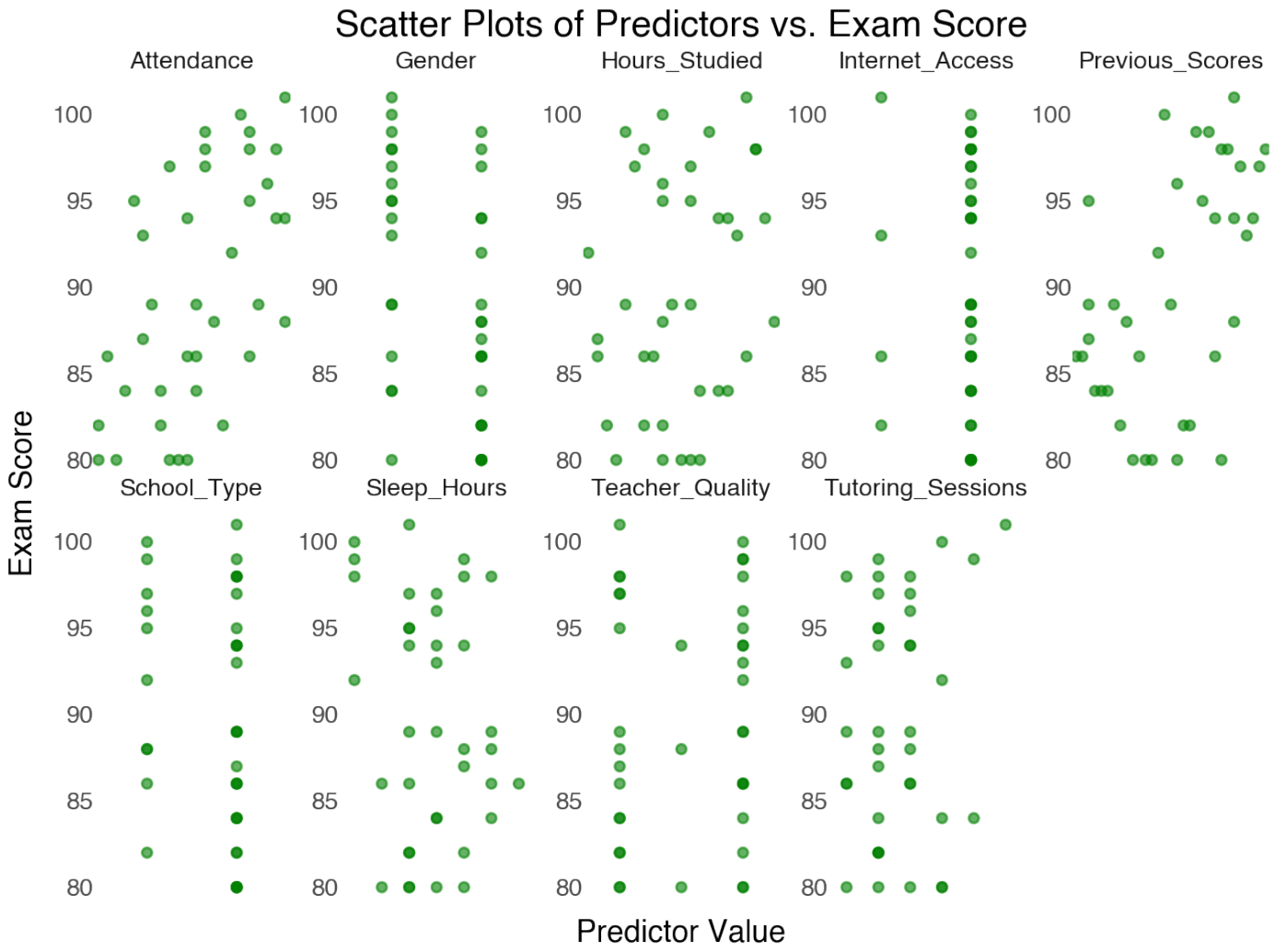
	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
Intercept	61.3570	8.689	7.061	0.000	43.287	79.427
Access_to_Resources[T.Low]	-5.7833	2.147	-2.693	0.014	-10.249	-1.318
Access_to_Resources[T.Medium]	0.9025	2.073	0.435	0.668	-3.408	5.213
Internet_Access[T.Yes]	2.4094	2.284	1.055	0.303	-2.340	7.159
Family_Income[T.Low]	-1.2200	2.065	-0.591	0.561	-5.514	3.074
Family_Income[T.Medium]	1.2700	1.937	0.656	0.519	-2.758	5.298
Teacher_Quality[T.Low]	-3.5358	3.013	-1.174	0.254	-9.802	2.730
Teacher_Quality[T.Medium]	-0.6296	1.633	-0.386	0.704	-4.025	2.766
School_Type[T.Public]	-1.9864	2.269	-0.875	0.391	-6.705	2.732
Parental_Education_Level[T.High School]	1.7369	1.595	1.089	0.289	-1.581	5.055
Parental_Education_Level[T.Postgraduate]	1.9619	2.174	0.902	0.377	-2.560	6.484
Gender[T.Male]	-0.6569	1.600	-0.411	0.686	-3.984	2.671
Hours_Studied	0.1755	0.168	1.046	0.307	-0.173	0.524
Attendance	0.2455	0.078	3.162	0.005	0.084	0.407
Sleep_Hours	-0.6946	0.477	-1.456	0.160	-1.687	0.298
Previous_Scores	0.1439	0.048	3.009	0.007	0.044	0.243
Tutoring_Sessions	-0.3698	0.636	-0.581	0.567	-1.692	0.953

(b) Regression Coefficients

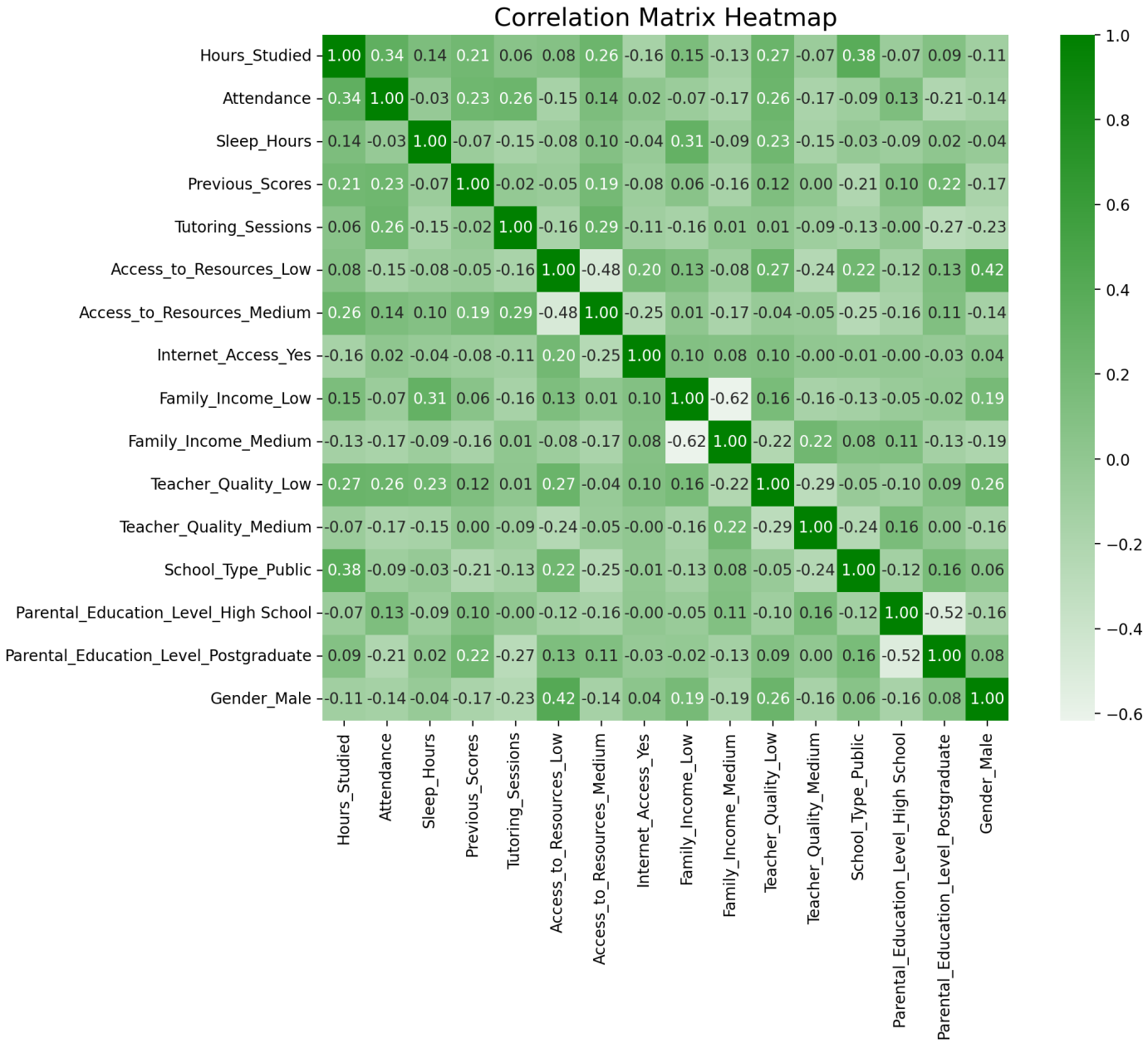
Omnibus:	1.452	**Durbin-Watson:**	1.866
Prob(Omnibus):	0.484	**Jarque-Bera (JB):**	1.252
Skew:	0.429	**Prob(JB):**	0.535
Kurtosis:	2.765	**Cond. No.**	1.60e+03

(c) Model Diagnostics

Supplementary Figure 9: Summary of Regression Results for the High Model



Supplementary Figure 10: Scatterplot of each predictor vs. Exam Scores for the High Model



Supplementary Figure 11: Correlation Matrix Heatmap for the High Model