FACTORS AFFECTING STUDENT ACADEMIC PERFORMANCE

TERM PAPER FOR THE COURSE: MSC 609 – QUANTITATIVE DATA ANALYSIS FOR MANAGEMENT SCIENCE

SUBMITTED BY

ANJIYA ADIL NOORUDDIN

ARYAN GANDHI

AVERY FUNSTON

UTKARSH SINGH

Contents

Introduction

Student performance has been a debated topic for years and understanding the factors that contribute to it has been a critical task in the field of education. This report explores a range of quantitative and qualitative variables that affect the achievement of students both negatively and positively. The factors include, but not limited to, study hours, access to resources, prior performance, attendance, motivation level, socio-economic factors and additional variables relevant in this context. Each element plays a crucial role in the performance of students and their educational outcomes which is why it is imperative to understand the effect of each factor for educational advancements to promote academic success.

An extensive amount of research has been conducted previously in similar domains to explore the influence of various variables on the performance of students both negatively and positively and identify the correlation between various variables. Personal attributes, Socio-economic conditions, and academic support facilities have been linked with academic performance in various studies. A thorough understanding of the factors and their affects is crucial for not just students, but for policymakers, educators, and businesses that develop educational products and services to promote enhanced academic resources and environments.

This study sheds light on the variables that highly influence the performance of students by analyzing a synthetic set of data on study habits, socio-economic conditions, teacher quality and various other key factors. An in-depth analysis of the findings of this study will help in creating potential strategies for policymakers, students, educators and businesses to lay strong foundations for interventions in the field of education based on evidence from various models.

Data Description

Data Source

- The dataset for this study is an artificial dataset with various variables that have been generated for educational purposes. The data has not been obtained from any institution, rather it has been synthetically created to model realistic performance scenarios of students.
- License: CC0: Public Domain
- Link: Student Performance Factors (kaggle.com)
- Vetting the dataset for reliability included the following stages:

1. **Data Completeness**: The artificial dataset had three variables namely, 'Parental education Level', 'Teacher Quality', and 'Distance from Home' with missing values which were accounted for during the modelling process. (Figure 2)
2. **Variable Types**: The artificial dataset was created such that it has both quantitative and qualitative variables, all of which are represented appropriately.

   a) <u>Numerical Variables:</u>
   - Hours Studied: ranges from 1 to 44, with a mean of approximately 20 hours. This range is plausible, though hours as high as 44 might represent outliers.

- Attendance: ranges from 60 to 100, with an average attendance of around 80%. This distribution is within realistic bounds.
- Sleep Hours: ranges from 4 to 10, averaging 7 hours, which aligns with typical sleep patterns.
- Previous Scores and Exam Score: range from 50 to 100, with `Exam Score` having a maximum value of 101, suggesting an error or entry exceeding a typical score scale.
- Tutoring Sessions and Physical Activity show logical ranges.

b) Categorical Variables: Most categorical variables have two or three levels

- 'Low, Medium, High' for Parental Involvement, Motivation Level, Family Income, Teacher Quality and Access to Resources).
- 'Yes, No' for Extracurricular Activities, Internet Access, Learning Disabilities
- 'Public, Private' for School Type
- 'Positive, Negative, Neutral' for Peer Influence
- 'Male, Female' for Gender
- 'College, High School, Postgraduate' for Parental Education Level
- 'Far, Moderate, Near' for Distance from Home

3. **Analyzing Statistical Summaries for Numerical Variables**: All numerical variables have values within expected ranges, which aligns with typical educational and student performance metrics with the exception of:
   - Exam Score: The maximum value of the variable is 101, which could be an error or indicate an intended simulation of exceptional performance.

4. **Inspecting Categorical Variables:** The categorical variables appeared to be consistent throughout with either having defined set of values like (Yes or No) etc. or binary labels.

Key Variables

The dataset consists of student data for 6,607 students for the various quantitative and qualitative variables that capture and influence a student's study habits, socio-economic conditions, school environment, and personal characteristics, which can significantly influence student exam performance and help identify the most impactful contributors to academic success.

Variable Name	Description	Information Provided
Hours Studied	Number of hours studied per week.	Indicates students' study habits and dedication.
Attendance	Percentage of classes attended.	Reflects students' discipline and engagement.
Parental Involvement	Level of parental involvement in the student's education (Low, Medium, High).	Represents parental encouragement and support
Access to Resources	Availability of educational resources (Low, Medium, High).	Reflects access to learning tools – linked to family income
Extracurricular Activities	Participation in extracurricular activities (Yes, No).	Suggests involvement in non-academic activities
Sleep Hours	Average number of hours of sleep per night.	Indicates sleep habits and potential effects on concentration
Previous Scores	Scores from previous exams.	Reflects prior academic achievement and baseline performance
Motivation Level	Student's level of motivation (Low, Medium, High).	Captures the student's internal drive
Internet Access	Availability of internet access (Yes, No).	Indicates ability to access online learning resources
Tutoring Sessions	Number of tutoring sessions attended per month.	Shows additional academic support received
Family Income	Family income level (Low, Medium, High).	Reflects economic background
Teacher Quality	Quality of the teachers (Low, Medium, High).	Indicates the influence of teacher effectiveness
School Type	Type of school attended (Public, Private).	Reflects schooling environment
Peer Influence	Influence of peers on academic performance (Positive, Neutral, Negative).	Describes the social environment and peer effect
Physical Activity	Average number of hours of physical activity per week.	Indicates the level of physical health engagement
Learning Disabilities	Presence of learning disabilities (Yes, No).	Identifies special education needs
Parental Education Level	Highest education level of parents (High School, College, Postgraduate).	Provides context on family educational background
Distance from Home	Distance from home to school (Near, Moderate, Far).	Suggests potential commute impact on study time and fatigue

Gender	Gender of the student (Male, Female).	Basic demographic information
Exam Score	Final exam score - the response variable.	Measures academic performance outcome

Descriptive Data Analysis

This section aims to provide a thorough overview of the dataset, unravel patterns and relationships between various quantitative and qualitative variables and extract significant insights from statistical summaries and visualizations.

After observing the summaries of numerical variables, hours studied, and exam scores show moderate skewness. Categorical variables like gender, school type and access to resources display balanced distribution while extracurricular is skewed towards fewer participants. (Figure 1: Summary of Variables)

Histograms of Numerical Variables (Figures 5-11)

Variable	Distribution
Previous Scores	Appears uniform, spread across the range of values (60-100).
Sleep Hours	Bimodal distribution, with a higher frequency at 2.5-7.5 and lower frequency between 7.5-12.5 hours.
Attendance	Appears uniform, spread across the range of values (60-100)
Hours Studied	Normal Distribution with mean 20.
Tutoring Sessions	Binomial distribution with higher frequency from 0 to 2.5.
Physical Activity	Binomial distribution with higher frequency from 2.5 to 7.5.

Bar Charts of Categorical Variables (Figures 12-24)

Variable	Distribution
Extracurricular Activities	Shows high distribution for students performing extra curriculars than those who don’t. Overall structure of distribution is left skewed
Access to Resources	Shows right skewed with majority students reporting from moderate to high access
Parental Involvement	Depicts slightly skewed distribution showing many parents having moderate involvement
Internet Access	Highlights left skewed distribution with many having internet access
Motivation Level	Shows a left skewed distribution that is mostly concentrated around moderate levels.
Teacher Quality	Demonstrates quite uniform distribution
Family Income	Shows left skewed distribution -higher distribution from lower and medium income families
School Type	Shows skewed distribution - public distribution is higher
Peer Influence	Skewed distribution highlighting different levels of influence - mainly neutral & positive influence

Learning Disabilities	Skewed right distribution with most without disabilities
Parental Education Level	Depicts normal distribution having majority of parents with high school or some college education
Distance from Home	Left skewed having majority of students nearer to the schools
Gender	Skewed distribution with high proportion of males

Relationships between the Dependent and Independent Variables

Distribution of Exam Score

Figure 3 shows a histogram that depicts the distribution of Exam Score which is slightly skewed with many students scoring in the mid to high range, showing moderate to high academic performance while also showcasing a small percentage of low performers.

Box Plot of Exam Score

Figure 4 shows a box plot that illustrates that most students score between in the range of 65-68. Also, relatively tight clustering is observed. Additionally, some exceptional performers who scored well are seen above the typical range while few are below the range.

Scatter Plot: Hours Studied vs Exam Score

Figure 27 shows a scatter plot illustrating hours studied by student and their exam score. Basically, it shows a positive correlation between the two as confirmed from the blue trend line that shows a moderate upward trend. The study hour range is between 0 to 40. Most scores are clustered around 60-80. Some students achieve quite high score (90-100) with varying study hours indicating other external factors might also be leading to these outlier points.

Bar Plot for Attendance bins with Exam Score

Figure 28 shows a plot that illustrates a positive relationship between attendance level of students and their average exam score. Students with high attendance are across 70 range, medium attendance around 65 while low attendance direct towards low average score that is around 62 range.

Scatter plot faceted by school type to show teacher quality impact

Figure 29 shows a scatter plot that compares the teacher quality (low, medium, high) across both private and public schools. The exam score distribution is shown by vertical spread of points. Both schools show similar pattern in distribution. Teacher quality is observed to have similar positive impact on both public and private schools. There are also high performing exam scores (outliers) scores near 100. Median scores are seen as relatively consistent across categories.

Box plot of Motivation Level vs. Exam Score

Figure 30 shows box plot distribution of (low, medium, high) motivation level across exam scores. High motivation plot shows positive skewness, and the median line is closer to the bottom. Also, plot has longer upper whisker and multiple high outliers. This shows that while most high motivation students are in the lower range of distribution of box plot but still there's significant tail of high performers. On the other hand, low motivation plot shows high positive skewness, and median line is close to above end of plot. It has some outliers on high end. Lastly, medium plot shows moderate skewness as median line is centered across the plot. Suggest more like a bell-shaped distribution. Surprisingly, it has many high outliers, but overall distribution is balanced.

Contingency Table - Parental Involvement vs Motivation

Figure 31 shows a cross tabulation/contingency table of categorical variables. Here, it is observed that parental involvement and motivation levels depict a strong association where high parental engagement correlates with higher motivation in students.

Proportional table - School Type vs Learning Disability

Figure 32 shows the proportional table that compares school type and learning disabilities showing that resource constrained schools have high proportion of students classified with learning disabilities.

Correlation of Hours Studied, Attendance, Previous Scores and Exam Score

Figure 33 shows a correlation heatmap that shows correlation between some key variables and exam scores:

1. Hours Studied and Exam Scores: 0.45 shows a moderate positive correlation, indicating that more study hours may direct towards better exam score
2. Attendance and Exam Scores: 0.58 it shows strong positive correlation when compared with other variables indicating strong positive relationship between attendance and exam scores.
3. Previous Scores and Exam Scores: 0.18 shows weak positive correlation suggesting that past scores have only a very slight positive effect on exam scores.

Question of Interest

Research Question

"What factors have the most influential contribution to student academic performance?"

Hypothesis 1: Hours studied for an exam will have a significant impact on exam grades. An increase in study time is strongly correlated with higher exam grades.

Hypothesis 2: Attendance will positively impact exam grades, with higher attendance rates strongly correlating with better performance.

Hypothesis 3: Family income affects exam grades, as higher-income families may provide supplementary resources, leading to better exam outcomes.

Hypothesis 4: Sleep hours will significantly affect performance, with longer sleep associated with higher grades.

Hypothesis 5: Teacher quality plays a significant role in exam outcomes, with better teaching quality leading to higher student performance.

Literature Review

This chapter sheds some light on past research and studies that have been carried out to understand which factors have the most influential contribution to student academic performance.

The predictors we have handpicked for our research-based model are sleep hours, hours studied, and attendance. These factors were selected based on strong empirical evidence linked to a student's academic outcomes. Extensive research highlights the critical role these factors play in shaping student performance, particularly in exam grades. Although sleep measures can be inconclusive strong statistical evidence pointed towards its impact on grades. Similarly, the number of hours devoted to studying reflects the effort and time investment necessary for mastery of academic content. Attendance serves as a proxy for engagement and access to instructional content, both of which are essential for understanding material and meeting academic expectations. At the secondary school level, additional factors such as socioeconomic status (SES) and parental education also emerge as influences on student performance Students from higher SES backgrounds often benefit from enriched learning environments, additional academic support, and reduced stressors, which collectively enhance performance. Parental education further contributes to this dynamic, as more educated parent are likely to emphasize academic success, provide better guidance, and create a supportive environment for learning.

Attendance

Attendance has been consistently cited as one of the major factors influencing student academic performance. Immense emphasis has been laid on regular class attendance, which is linked to improved academic results, as it ensures continuous exposure to course material and active engagement in the learning process (Hijazi and Naqvi).

A meta-analytics review on attendance and grade score found that attendance is a strong predictor for both class grades and GPA, included but not limited to standardized testing like SAT. The study found that mandatory attendance in college level classes had a small impact on average grade received. The study concluded that attendance was a strong predictor of academic performance and explained a significant amount of unique variance in college students' grades since students who attend class regularly are more likely to grasp the subject matter, stay engaged, and participate in class discussions, all of which contribute to better academic performance (Crede, 2009).

A relevant study in this regard was conducted by (Malini and Kalpana, 2021) who found that students' engagement with educational materials during classes—facilitated by regular attendance—improves their ability to perform academically. This engagement, coupled with the opportunity for immediate feedback from instructors, enhances students' learning experiences.

According to another research, the importance of attendance was emphasized especially for students with leaning disabilities. The author shed light on how regular attendance can help such students in receiving the structured support that helps them perform well academically (Whitley, 2010).

Socio-Economic Status & Access to Resources

According to research, socio-economic class of students is one of the most substantial variables affecting their academic performance. Higher family incomes and parental education levels have been associated positively with the student's performance leading to improved outcomes. The authors have laid emphasis on the fact that students coming from higher socio-economic backgrounds have access to enhanced educational resources for example, enriching environment at home, private tutors, online learning material and much more which fosters the importance of education in kids and helps them build improved study habits, resulting in academic success (Hijazi and Naqvi, 2006).

Another research builds upon the same notion by highlighting the importance of family income to student success since higher family income means access to advanced study resources like textbooks, tutors, computers, educational tools etc. along with extracurricular activities which help students in building a well-balanced personality to strive in both the academic and professional worlds (Farooq et al., 2011).

Various studies have highlighted the struggles faced by students belonging to the lower socio-economic class. The psychological stress of financial instability affects students adversely, hampering their ability to perform well in academic tasks (Jaggia and Kelly-Hawke, 1999).

In summary, the findings from these studies show that differences in socio-economic statuses of students results in disparities between the access to additional support and resources available to students and their related outcomes.

Study Habits and Effort

Academic success depends on various factors of which study habits and effort have been acknowledged across many studies for better academic performance. According to a study, regular review of academic material, set academic goals, coupled with time management resulted in students getting better grades due to better learning and retention, even in the cases where the study material was not interesting and engaging, Hijazi and Naqvi (2006).

Another study used the technique of data mining to study the correlation between the amount of effort a student puts in and their academic performance. The study revealed a positive relationship since students who actively engaged in class, utilized learning resources and implemented problem solving techniques through practice performed better than those who didn't because of their ability to retain, comprehend, and apply the gained knowledge, (Malini and Kalpana, 2021).

Farooq et al. also added to the studies by corroborating that despite the presence of external factors, students who devote their time and effort to academics have a higher probability of achieving better outcomes further proving that academic success does not only depend on external circumstances or intelligence, but also on the student's dedication to the process.

Sleep and Psychological Well-being

The importance of sleep quality & psychological health has been emphasized in relation to academic outcome by many researchers. According to a study, lack of sleep damages the cognitive functions of the brain which eventually results in problems related to concentration, retention, and problem solving, all of which have been known to be crucial for improving academic performance of students. According to research, inadequate sleep also affects the emotional control of our brain leading to an increase in irritability and stress, that hinders the performance of students (Hershner, 2020).

The relationship between the performance of students and sleep has also been argued by Chow (2010) where he discusses the relationship of these two variables with psychological factors. Anxiety is known to worsen concentration levels, while depression is known to reduce motivation to engage with academic tasks, thus reducing overall performance particularly for undergraduate students, who face the dual stress of handling academic work with personal development.

Summary

While our model focuses primarily on sleep hours, study hours, access to resources, family income, and attendance for simplicity and manageability, it is important to acknowledge the broader context in which these variables operate. Other factors reviewed in the original study often intersect with our selected predictors, amplifying or mitigating their effects. This interconnectedness underscores the complexity of academic performance and the need for a holistic understanding of the factors influencing student success.

Previous Literature	Hypothesis Description	Dependent Variable	Main Independent Variable	Expected Sign on IV
(Hijazi and Naqvi) (Crede, 2009) (Malini and Kalpana, 2021) (Whitley, 2010)	Attendance positively impacts exam grades	Student Performance	Attendance	Positive
(Hijazi and Naqvi, 2006) (Farooq et al., 2011) (Jaggia and Kelly-Hawke, 1999)	Higher SES results in access to supplementary resources leading to better exam performance	Student Performance	SES and Access to Resources	Positive
(Malini and Kalpana, 2021) Hijazi and Naqvi (2006) Farooq et al. (2011)	Increase in study time results in higher grades	Student Performance	Study Habits and Effort	Positive
(Hershner, 2020) (Chow, 2010)	Sufficient sleep and mental well-being result in better performance	Student Performance	Sleep & Psychological Well-being	Positive

Data Relevance

The dataset includes variables explicitly aligned with the research question and hypotheses, enabling a direct investigation of their impacts on exam scores.

Hours Studied (numeric): Aligns directly with the hypothesis on study habits and effort and its influence on exam scores.

Attendance (numeric): Captures a student's attendance percentage, directly relevant to the hypothesis about attendance.

Family Income (categorical): Provides income levels ("Low," "Medium," "High"), which depict the socio-economic situation of the student which can be analyzed in relation to exam performance.

Sleep Hours (numeric): Directly aligns with the hypothesis on the relationship between sleep and academic performance.

Access to Resources (categorical): Reflects the availability of resources (e.g., "High," "Medium," "Low") for students.

The key variable, **Exam Score** (numeric), serves as the dependent variable, measuring academic performance.

Initial Model and Estimation

Our initial model was multiple linear regression model consisting of all the predictor variables from the dataset, while our response was Exam Score which is continuous. The reason we initially chose all variables was to clearly understand the relationship between various factors and exam performance. We had several predictor variables which were both categorical and numerical and our initial model helped us to look out for the impact of each factor while controlling for others.

We then fit a reduced model on our dataset where the Exam Score was dependent on all independent variables except for the ones deemed to be insignificant in the initial model.

We also fit the Least Absolute Shrinkage and Selection Operator (Lasso) model which is a L1 regularized least squares regression technique that applies 10-fold cross validation to find the optimal regularization parameter to enhance both model interpretability and predictive accuracy. Given the high dimensionality of our dataset, feature selection was crucial to identify and exclude irrelevant variables (those with coefficients equal to zero). This process simplifies the model and helps prevent overfitting.

Stepwise selection was employed to refine the model by combining forward selection and backward elimination, ensuring a thorough search for the most influential predictors of Exam Score. In this case, AIC (Akaike Information Criterion) was used to guide the selection process where $AIC = 2K - 2\ln(L)$ where k = number of parameters and L = maximum likelihood, identifying the most relevant variables while balancing model fit against complexity.

For further improvement, we selected a subset of predictors sleep hours, hours studied, access to resources, family income, and attendance. Reason for selecting was based upon strong influence on academic outcome according to previous research which is summarized as follows:

- **Sleep Hours**: Strongly linked to cognitive functioning and academic outcomes (Curcio et al., 2006; Hershner et al., 2020).
- **Hours Studied**: Correlates with commitment and academic success (Zubair et al., 2024).
- **Access to Resources:** Research underscores socio-economic status as a substantial factors influencing academic performance. (Hijazi and Naqvi, 2006).
- **Family Income: P**ositively associated with improved academic outcomes. (Naqvi, 2006)
- **Attendance**: Strong predictor of academic excellence (Crede, 2009).

By focusing on these predictors, the handpicked model's goal is to simplify explanation while ensuring high relevance to the research hypothesis. The model's estimation and selection conditions highlight the interaction of these predictors in effecting student performance, giving impactful insights for enhancing academic results.

We implemented various estimation strategies to optimize the model.

1. **Training Test Split**
   - Estimation method: 80-20 Train Test Split
   - Includes data splitting into 80-20 (80% is training set for model fitting, 20% is test set for validation)

- Random sample with set seed is used for reproducibility.

2. **F-Test**
   - Used to determine whether there is significant difference in explaining variability in the response variable compared to a baseline model.
   - Used to identify whether number of predictors in the model can be reduced without sacrificing significant explanatory power.

Interpretation of Results

Models

Linear Regression

We began by performing a linear regression analysis on our train dataset to assess the relationship between each independent variable and the dependent variable (Exam Score). The results revealed that all independent variables, except for Sleep Hours, School Type, and Gender, demonstrated a statistically significant relationship with Exam Score, as indicated by p-values less than 0.05. In contrast, Sleep Hours, School Type, and Gender had p-values greater than 0.05, suggesting that these variables do not significantly contribute to predicting the exam score. Furthermore, the model's overall significance is confirmed by the high F-statistic of 451.6 and the extremely low p-value (2.2e-16), both of which indicate that the model as a whole is highly significant. The adjusted R-squared value of 0.7046 suggests that approximately 70.46% of the variance in Exam Scores is explained by the independent variables included in the model. The coefficients of each independent variable were further analyzed to understand their individual impact on the exam scores shown in a table in the appendix.

Reduced Linear Model

The full linear model, which included all independent variables, was compared to a reduced model that excluded Sleep Hours, School Type, and Gender. These three variables were removed based on the previous analysis showing their insignificance (p-value > 0.05) in predicting the dependent variable (Exam Score). The full model served as the baseline for an ANOVA test, which was used to evaluate whether removing these variables significantly impacted the model fit.

The p-value from the F-test (0.4787) is greater than the significance level of 0.05, indicating that the reduction in model complexity does not lead to a statistically significant loss of explanatory power. Thus, removing those predictors allows us to maintain explanatory power while simplifying the model.

Thus, the reduced model, which is simpler and more efficient, is just as effective as the full model in explaining Exam Score, and there is no evidence to suggest that retaining Sleep Hours, School Type, and Gender would improve predictive accuracy.

Lasso Model

To determine the optimal penalty parameter, the data was divided into 10 folds (nfolds = 10) for cross-validation. The best lambda value that minimized the cross-validation error was found to be 0.003066671. This value strikes an appropriate balance between bias (underfitting) and variance (overfitting), ensuring a robust model.

The intercept of 40.966 represents the baseline predicted exam score when all predictors are at their reference levels (i.e., categorical predictors at their baseline categories and continuous predictors at zero). The coefficients were categorized into Positive Impact, Negative Impact, and Zero Impact, which are detailed in the tables in the appendix, offering clearer insights into the influence of each predictor on the exam score.

Stepwise Selection Model

Stepwise selection was employed to refine the model by combining forward selection and backward elimination, using AIC (Akaike Information Criterion) to guide the selection process, identifying the most relevant variables while balancing model fit against complexity.

Step 1: Removing Sleep Hours

Excluding Sleep Hours led to a slight reduction in AIC, from 7815.77 to 7814.13, signaling a minor improvement in model fit. However, the Residual Sum of Squares (RSS) increased slightly, from 23349 to 23351, suggesting a negligible loss in explanatory power.

Step 2: Removing Gender

Excluding Gender further reduced AIC to 7813.15. Again, the increase in RSS (23355) was minimal, indicating a minor loss in model accuracy.

Step 3: Removing School Type

Finally, removing School Type reduced the AIC to 7812.27, marking the last improvement. The RSS increased marginally to 23361, further confirming the minor impact of the removed variables.

The minimal changes in RSS between models suggest that the removed variables—Sleep Hours, Gender, and School Type—had negligible effect on the overall explanatory power of the model. Meanwhile, the consistent decrease in AIC reflects a more parsimonious and efficient model.

The final model (AIC = 7812.27) is more efficient than the initial one, aligning with previous analyses where the p-values for these variables were greater than 0.05, indicating their limited contribution. Key variables, such as Hours Studied, Attendance, and Previous Scores, remain central in the final model, underscoring their strong relationship with Exam Score. In conclusion, the final model strikes an optimal balance between model fit and predictive performance, as evidenced by the AIC reduction and minimal change in RSS.

Model with Handpicked Predictors

Based on the rationale discussed earlier, a regression model was developed using carefully selected predictor variables deemed to have a substantial impact on the dependent variable, Exam Score. These predictors—Sleep Hours, Hours Studied, Attendance, Access to Resources, and Family Income—were identified through a review of prior research on factors influencing academic performance.

The model accounts for approximately 56.35% of the variance in Exam Score (adjusted R squared), indicating a moderately strong explanatory power. The overall model is highly significant ($p<2.2\times10^{-16}$), demonstrating that the predictors collectively explain a significant portion of the variance in the outcome variable. However, the F-test with this model and the full model as the baseline, indicate the full model has some additional predictors that are significant to the response with p-value ($p<2.2\times10-16$).

Among the predictors, Hours Studied, Attendance, Access to Resources, and Family Income were found to be highly significant contributors to Exam Score. Notably, Attendance emerged as the most influential predictor, as evidenced by its exceptionally high t-value (62.885). In contrast, Sleep Hours did not exhibit a statistically significant relationship with Exam Score. This finding suggests that the role of sleep may be more complex or mediated by other variables not included in the model. The effect of each variable on the dependent variables is explained in a table in the appendix.

These results underscore several actionable insights: enhancing Hours Studied and Attendance, as well as improving Access to Resources, are likely to yield substantial improvements in student performance. Furthermore, the significant effects of Family Income and Access to Resources highlight persistent socioeconomic disparities. Addressing these inequities through targeted interventions could play a crucial role in supporting students from disadvantaged backgrounds and promoting educational equity.

To further analyze the model, we compared the handpicked model with the full model through ANOVA. A p-value of less than 0.05 (< 2.2e-16) showed that the additional variables in the full model are significant and capture a large proportion of variability in the dependent variable, so it may be necessary to include some other predictors.

Comparing Models using AIC

Akaike Information Criterion (AIC) is a widely used metric for model selection, designed to balance the goodness of fit of the model with the model complexity. By penalizing models with excessive parameters, AIC helps identify the model that most adequately describes the data. In this study, we used AIC to compare the Reduced Linear Model, the Stepwise Model, the Lasso model, and the Handpicked Model to determine which provided the best analysis of our dataset. A lower AIC value suggests a better model fit when comparing models that have been fit on the same dataset.

Based on the AIC for each model, the Reduced, Stepwise, and Lasso Models have the best balance between goodness of fit and model complexity, as they have low negligible AIC values of 7812.267 and 7814.304. This shows that these models maintain a balance between model complexity and goodness of fit. In contrast, the results show us that the custom model has the highest AIC value of all models, indicating that the model is likely underfitting the data due to its lack of complexity which compromised the model's predictive power.

Comparing Models using R-squared

Adjusted R squared is a statistical measure that shows the amount of variance in the dependent variable (Exam Score) that can be explained by the independent variables of the Student Performance dataset, while accounting for the different number of predictors between models. Adjusted R squared is a more accurate measure as compared to R squared when models with different numbers of predictors are being compared.

```
Adjusted R^2 for Reduced Model: 0.7046002
Adjusted R^2 for Stepwise Model: 0.7046002
Adjusted R^2 for Lasso Model: 0.7045392
Adjusted R^2 for Custom Model: 0.5635486
```

Stepwise & Reduced Model: These models explain the highest proportion of variance, with an Adjusted R^2 of 0.7046. This suggests that the models have effectively identified the most relevant predictors for explaining the variance in the dependent variable.

Lasso Model: The Lasso Model shows an Adjusted R^2 of 0.7045, which is slightly lower than the Stepwise Model (by 0.0001). While this is a negligible difference, it suggests that the Lasso model's regularization (through shrinkage and variable selection) does not substantially reduce explanatory power. Nevertheless, Lasso could still offer advantages in terms of model stability and interpretability by reducing overfitting, especially in the presence of highly collinear or irrelevant predictors. Its ability to perform variable selection and shrinkage makes it particularly useful in scenarios where the goal is to prevent overfitting and improve model generalization.

Custom Model: The Custom Model exhibits the lowest Adjusted R^2 value of 0.5635, indicating that it explains significantly less of the variance in the dependent variable compared to the other models. Since the custom model has been fit by using handpicked variables (based on research), this indicates that the chosen variables may not be relevant or may not be explain an extensive amount of variation in the dependent variables due to over simplicity and there must be other significant variables that should be considered. This could also indicate the presence of non-linear relationship between variables.

Choice of Final Model

After a comprehensive comparison of various models, we conclude that the Lasso model emerged as the top choice for predicting the effect of various factors on student performance. Despite having an AIC value slightly higher than the reduced and stepwise model, the negligible difference in value of 2.037 can be ignored since the lasso model can be advantageous considering the dataset has some redundant variables and Lasso regularization improves the interpretability of the model by removing or reducing the coefficients of predictor variables that are insignificant and helps prevent overfitting due to high correlation between variables. Considering the stepwise and reduced models depend on variable selection methods like the stepwise regression or the backward elimination, they might not effectively handle data that is highly dimensional and multicollinear.

Additionally, when diagnosing the Lasso model's assumptions, we find that although there are outliers that could potentially introduce bias to the model, for the most part the linearity, independence, constant variances and normality assumptions hold. Thus, the model results can be trusted with few doubts.

We then test the chosen model on the test set which demonstrated that the Lasso Model portrayed exceptional predictive performance with a root mean squared error equal to 1.7401. This means that on average, the predictions made by the lasso model deviate by 1.74 points. A low value of RMSE indicates small errors and better performance of the model on average making it an effective model to precisely predict the performance of students.

Limitations and Further Research

Model Limitations

One major limitation of our model is that it was trained and tested on synthetically generated data. While synthetic data can provide diverse scenarios and sufficient volume for training, it may not capture the nuanced relationships and variability present in real-world settings. As a result, it is challenging to generalize the findings to real-world scenarios.

Additionally, there were some rows with missing values for ‘Parental education Level’, ‘Teacher Quality’, and ‘Distance from Home’ which were removed with the assumption that the data was missing completely at random. There was an outlier present in the Exam Score with a value of 101. These outliers and missing values may have created bias in model’s predictions.

Future Research

To help address these limitations in future research, we recommend conducting a study using real-world data to better understand the factors affecting student academic performance. Collecting real-world data can be cumbersome, time-consuming and difficult to obtain a sufficiently large sample size for training. Therefore, it may be beneficial to use a combination of real-world and synthetic data. This approach would allow the model to be validated on real-world data while still leveraging synthetic data to achieve sufficient data volume for training. Incorporating authentic datasets will allow for a more accurate evaluation of the model's predictive accuracy and generalizability.

Additionally, including additional factors that influence academic performance would further enhance the model's reliability and applicability. The current dataset captures only a limited set of variables, leaving out other significant contributors to academic performance. Factors such as mental health and sleep disorders are three factors that are associated with impacting academic performance. Mental health plays a crucial role in an individual’s ability to perform effectively. Clinical research highlights that a student’s capacity to manage stress and anxiety, along with the presence of underlying mental health disorders, can significantly influence their academic performance (Chu et al., 2022). Incorporating further research on this variable would provide valuable insights to enhance our model. Similarly, sleep disorders have been strongly associated with an increased risk of poor academic outcomes, underscoring the importance of including this factor in future studies as well (Curcio et al., 2006). Future studies should integrate these variables to provide a more comprehensive understanding of the determinants of student success.

Also, while the data was missing with complete randomness, applying more robust preprocessing techniques, such as advanced imputation methods or outlier handling, could minimize their influence and improve the model’s overall reliability and accuracy. Furthermore, if predictive power of the model is the top priority for other researchers, transformations such as a log transformation to the response may improve model performance and minimize the influence of the outliers on the model.

Another methodology for evaluating student performance prediction could be to break up the percentiles into categories such as “Pass” (70-100%) or “Fail” (0-69%) and use a logistic regression model to predict the category of the student. This approach may be more practical and reliable than predicting the student’s exact exam score percentage as it may allow institutions to be able to identify students that may be at a disadvantage or require special attention.

References

Hijazi, S. T., & Naqvi, S. M. M. (2006). FACTORS AFFECTING STUDENTS'PERFORMANCE. Bangladesh e-journal of Sociology, 3(1).

Malini, J., & Kalpana, Y. (2021). Investigation of factors affecting student performance evaluation using education materials data mining technique. Materials Today: Proceedings, 47, 6105-6110

Farooq, M. S., Chaudhry, A. H., Shafiq, M., & Berhanu, G. (2011). Factors affecting students' quality of academic performance: A case of secondary school level. Journal of quality and technology management, 7(2), 1-14.

Jaggia, S., & Kelly-Hawke, A. (1999). An analysis of the factors that influence student performance: A fresh approach to an old debate. Contemporary Economic Policy, 17(2), 189-198.

Chow, H. P. (2010). Predicting academic success and psychological wellness in a sample of Canadian undergraduate students. Electronic Journal of Research in Educational Psychology, 8(2), 473-496.

Whitley, J. (2010). Modelling the Influence of Teacher Characteristics on Student Achievement for Canadian Students with and without Learning Disabilities. International Journal of Special Education, 25(3), 88-97.

Credé, M., Roch, S. G., & Kieszczynka, U. M. (2010). Class attendance in college: A meta-analytic review of the relationship of class attendance with grades and student characteristics. Review of Educational Research, 80(2), 272-295.

Hershner, S. (2020). Sleep and academic performance: measuring the impact of sleep. Current Opinion in Behavioral Sciences, 33, 51-56.

Zubair, T., Qazi, U., Faisal, S. M., & Khan, A. K. (2024). The impact of study hours on academic performance: A statistical analysis of students' grades.

Curcio, G., Ferrara, M., & De Gennaro, L. (2006). Sleep loss, learning capacity and academic performance. *Sleep medicine reviews*, *10*(5), 323-337.

Chu, T., Liu, X., Takayanagi, S., Matsushita, T., & Kishimoto, H. (2023). Association between mental health and academic performance among university undergraduates: The interacting role of lifestyle behaviors. *International Journal of Methods in Psychiatric Research*, *32*(1), e1938.

Appendix

Descriptive Data Analysis

Summary of Data

```
 Hours_Studied     Attendance     Parental_Involvement Access_to_Resources Extracurricular_Activities  Sleep_Hours      Previous_Scores  Motivation_Level
 Min.   : 1.00   Min.   : 60.00   Length:6607          Length:6607         Length:6607                Min.   : 4.000   Min.   : 50.00   Length:6607
 1st Qu.:16.00   1st Qu.: 70.00   Class :character     Class :character    Class :character           1st Qu.: 6.000   1st Qu.: 63.00   Class :character
 Median :20.00   Median : 80.00   Mode  :character     Mode  :character    Mode  :character           Median : 7.000   Median : 75.00   Mode  :character
 Mean   :19.98   Mean   : 79.98                                                                       Mean   : 7.029   Mean   : 75.07
 3rd Qu.:24.00   3rd Qu.: 90.00                                                                       3rd Qu.: 8.000   3rd Qu.: 88.00
 Max.   :44.00   Max.   :100.00                                                                       Max.   :10.000   Max.   :100.00
 Internet_Access    Tutoring_Sessions Family_Income      Teacher_Quality    School_Type        Peer_Influence     Physical_Activity Learning_Disabilities
 Length:6607        Min.   :0.000     Length:6607        Length:6607        Length:6607        Length:6607        Min.   :0.000     Length:6607
 Class :character   1st Qu.:1.000     Class :character   Class :character   Class :character   Class :character   1st Qu.:2.000     Class :character
 Mode  :character   Median :1.000     Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :3.000     Mode  :character
                    Mean   :1.494                                                                                 Mean   :2.968
                    3rd Qu.:2.000                                                                                 3rd Qu.:4.000
                    Max.   :8.000                                                                                 Max.   :6.000
 Parental_Education_Level Distance_from_Home    Gender            Exam_Score
 Length:6607              Length:6607        Length:6607        Min.   : 55.00
 Class :character         Class :character   Class :character   1st Qu.: 65.00
 Mode  :character         Mode  :character   Mode  :character   Median : 67.00
                                                                Mean   : 67.24
                                                                3rd Qu.: 69.00
                                                                Max.   :101.00
```

Figure 1: Summary of Variables

Missing Data

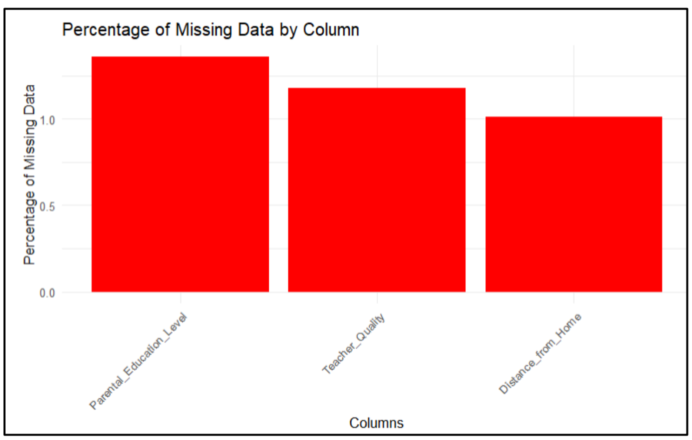


Figure 2: Percentage of Missing Values

Distribution of Dependent Variable

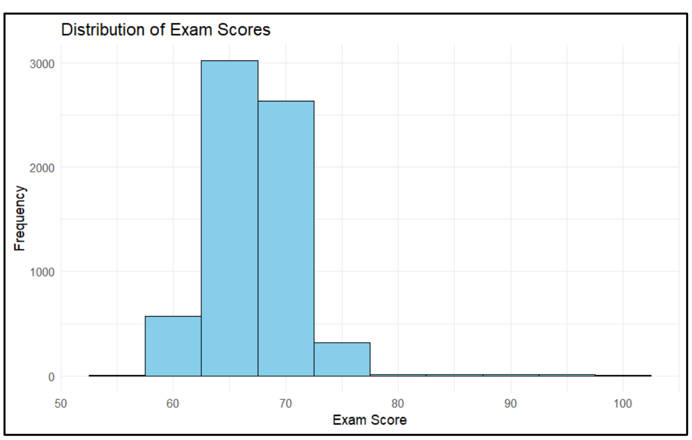


Figure 3: Distribution of Exam Score

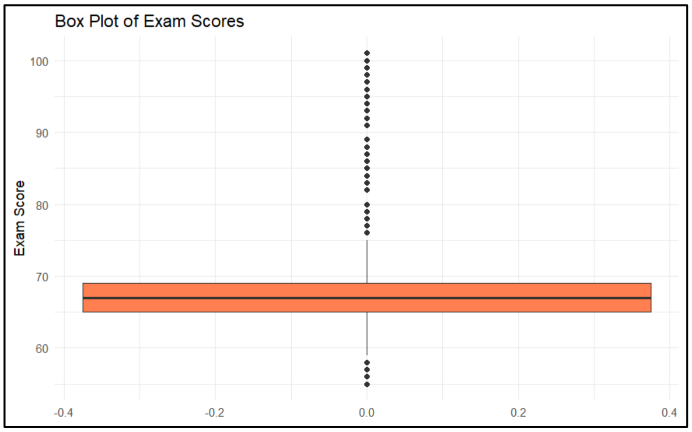


Figure 4: Box Plot of Exam Score

Distributions of Numerical Independent Variables

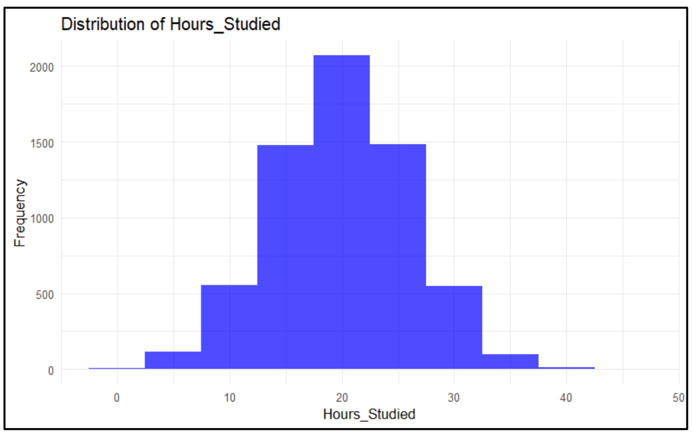


Figure 5: Distribution of Hours Studied

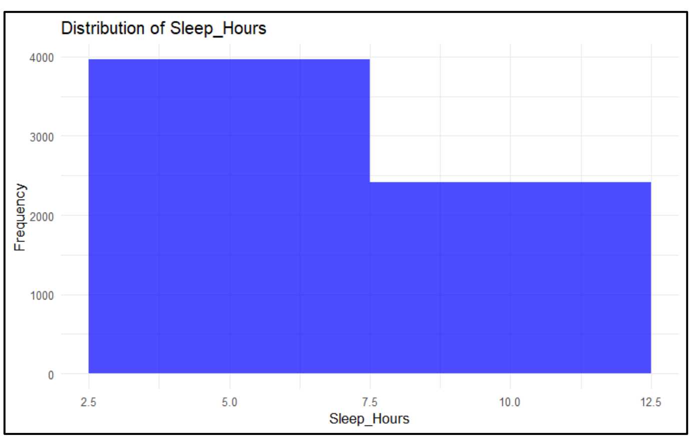


Figure 6: Distribution of Sleep Hours

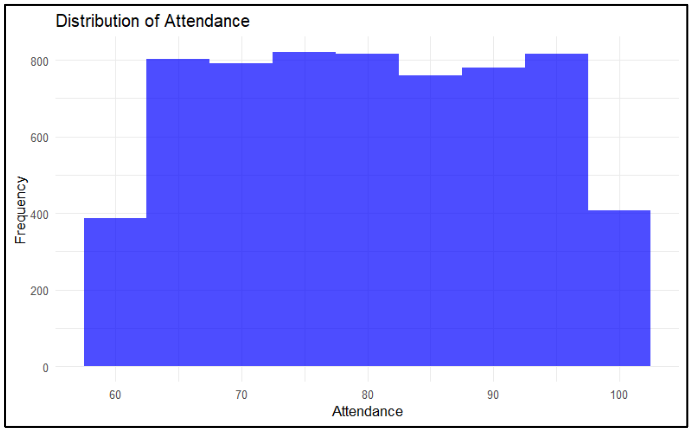


Figure 7: Distribution of Attendance

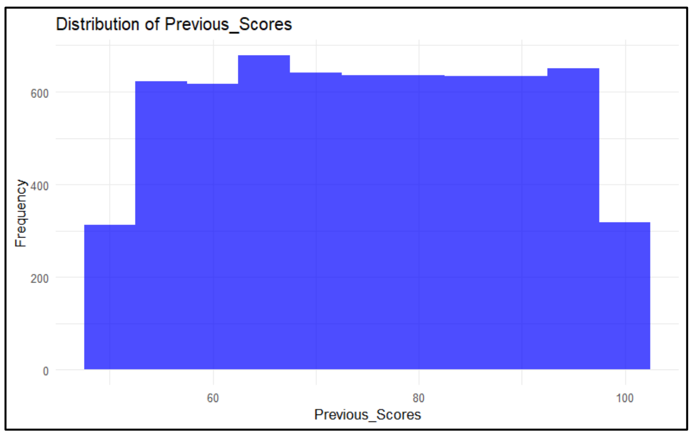


Figure 8: Distribution of Previous Scores

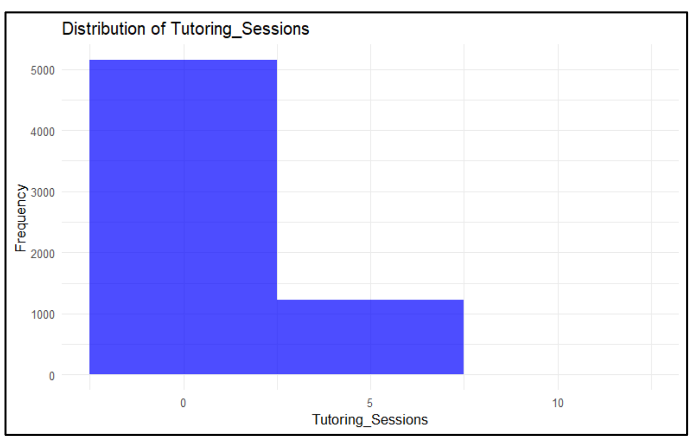


Figure 9: Distribution of Tutoring Sessions

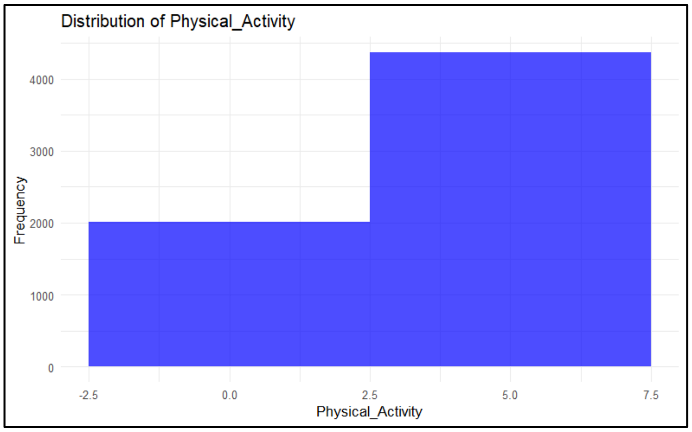


Figure 10: Distribution of Physical Activity

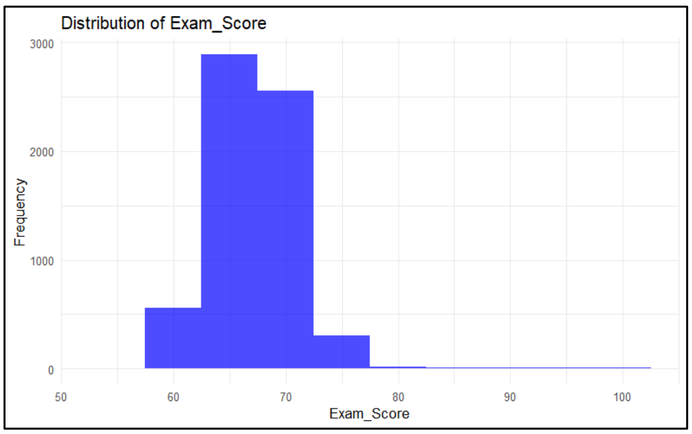


Figure 11: Distribution of Exam Score

Distribution of Categorical Independent Variables

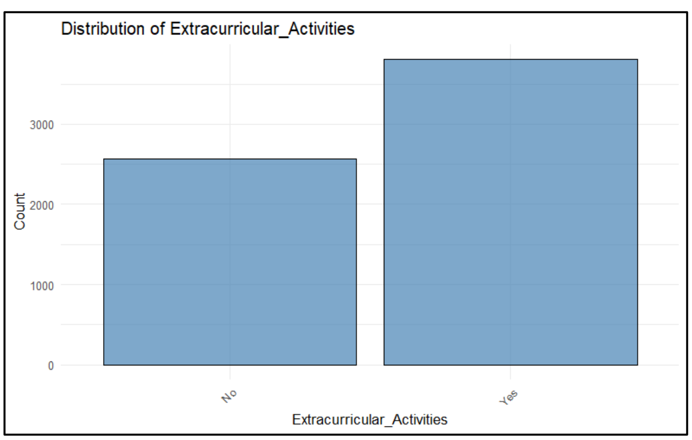


Figure 12: Distribution of Extracurricular Activities

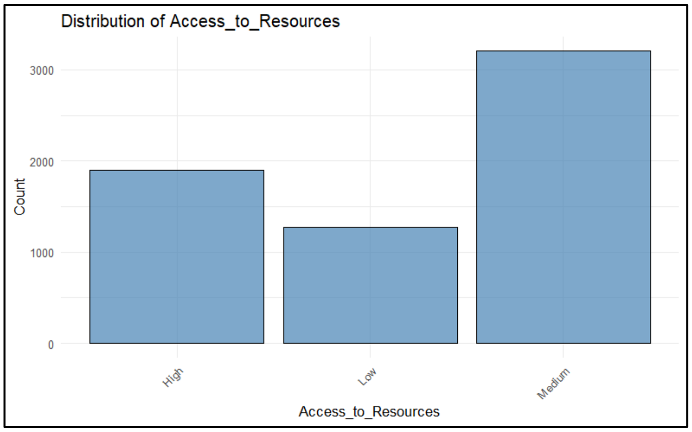


Figure 13: Distribution of Access to Resources

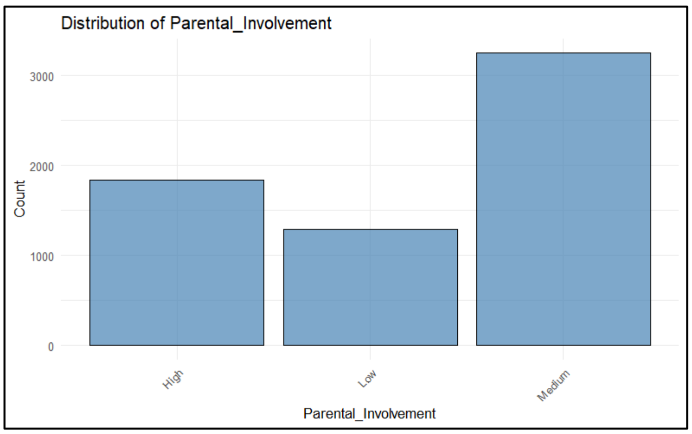


Figure 14: Distribution of Parental Involvement

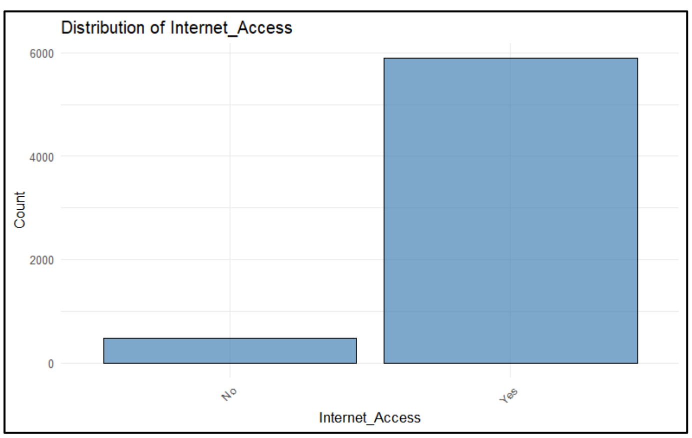


Figure 15: Distribution of Internet Access

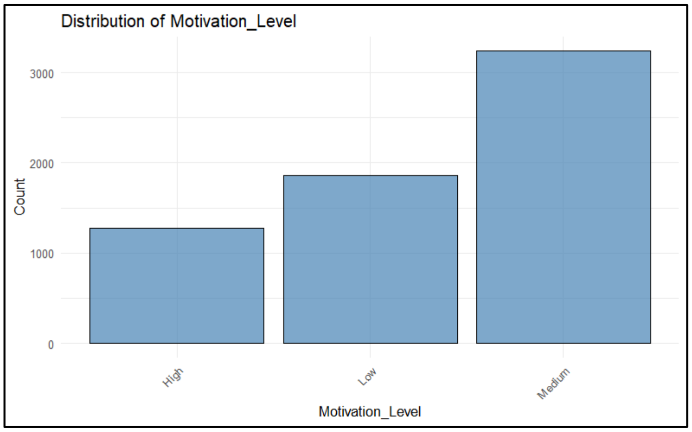


Figure 16: Distribution of Motivation Level

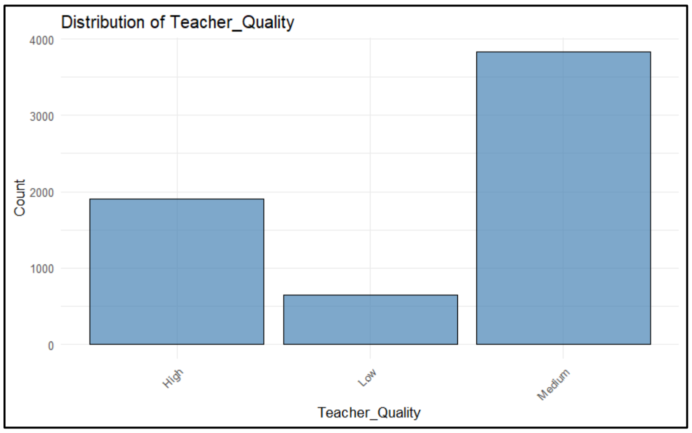


Figure 17: Distribution of Teacher Quality

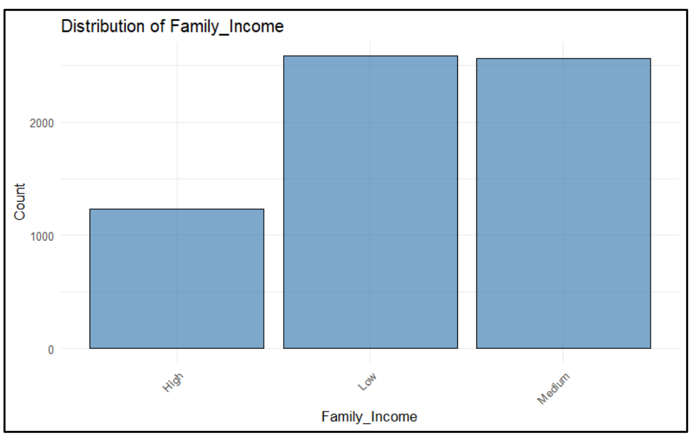


Figure 18: Distribution of Family Income

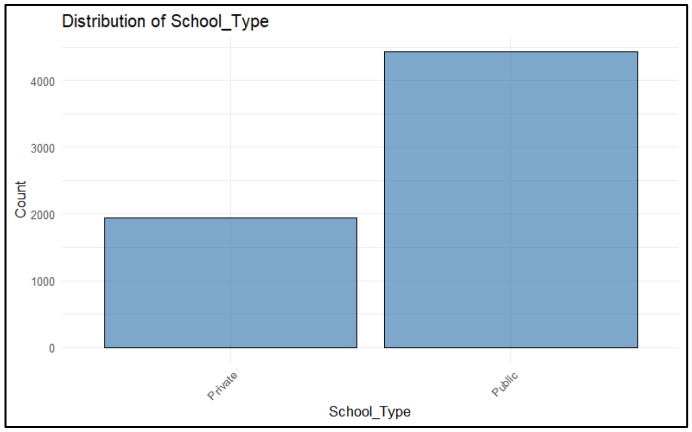


Figure 19: Distribution of School Type

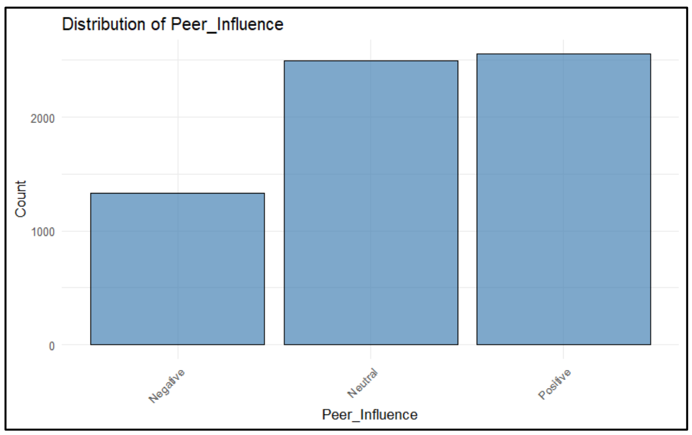


Figure 20: Distribution of Peer Influence

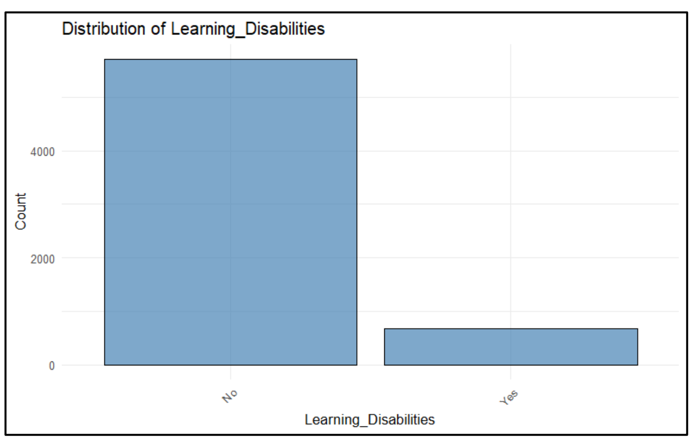


Figure 21: Distribution of Learning Disabilities

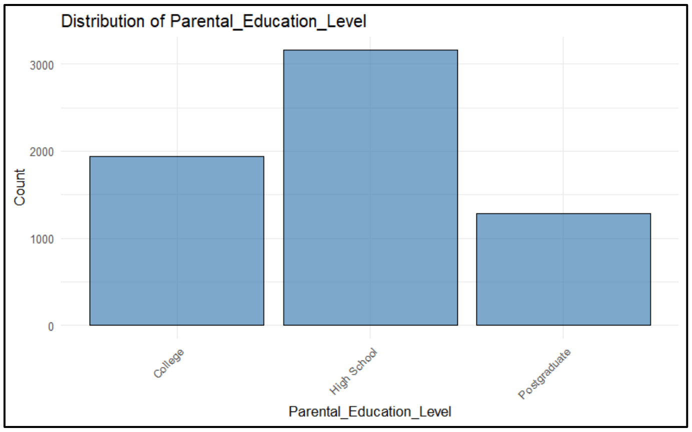


Figure 22: Distribution of Parental Education Level

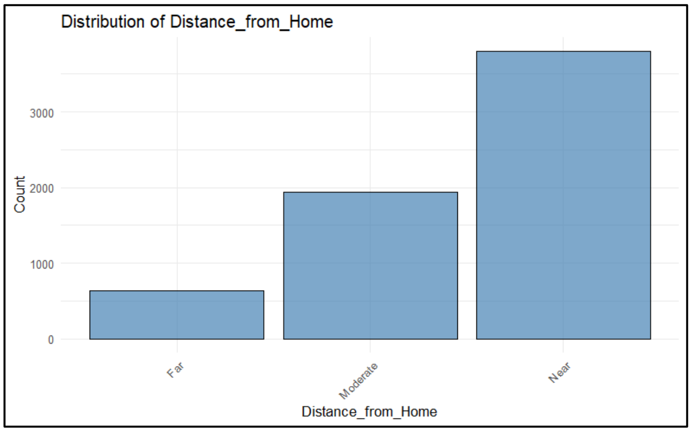


Figure 23: Distribution of Distance from Home

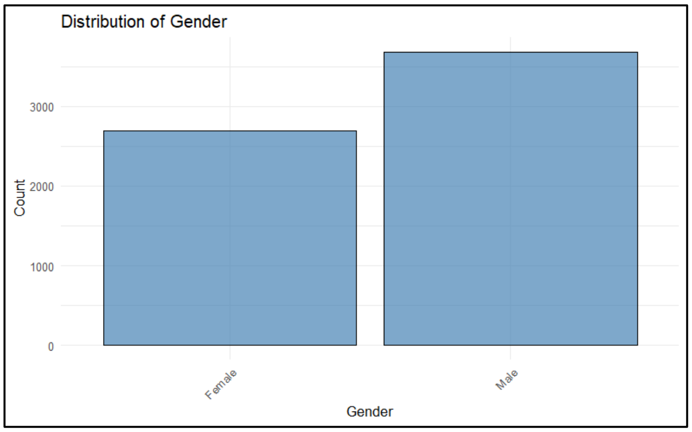


Figure 24: Distribution of Gender

Relationships between Dependent and Independent Variables

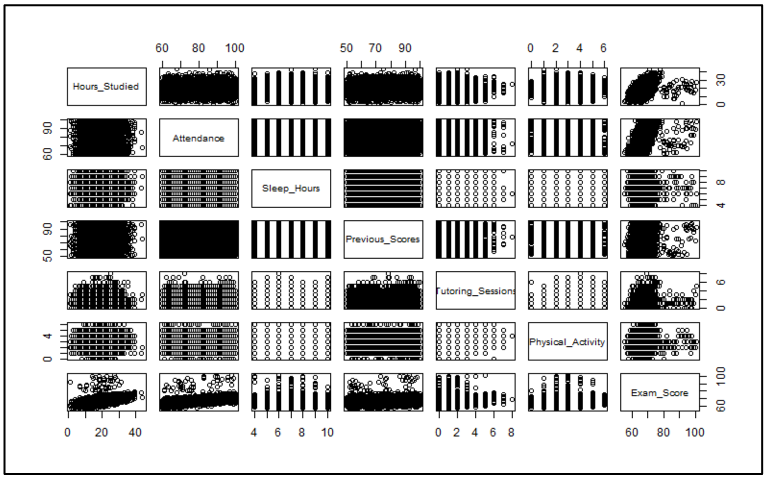


Figure 25: Pair plot for Numeric Variables

	High	Low	Medium
High	347	553	936
Low	271	355	665
Medium	659	956	1636

	No	Yes
Private	0.27249922	0.03229853
Public	0.62276576	0.07243650

Figure 26: Cross-tabulations for categorical variables

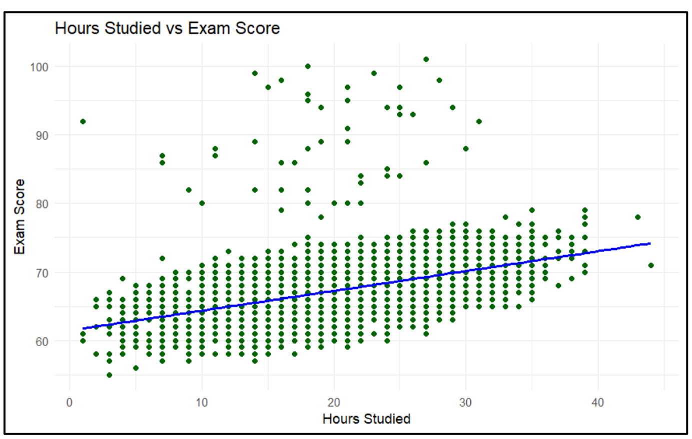


Figure 27: Scatter Plot: Hours Studied vs Exam Score

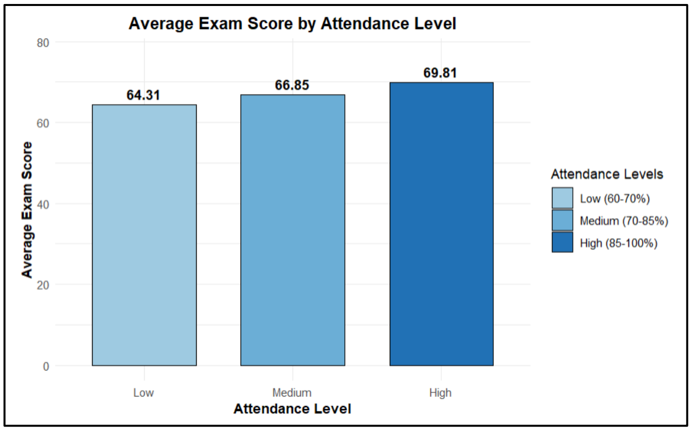


Figure 28: Bar Plot for Attendance bins with Exam Score

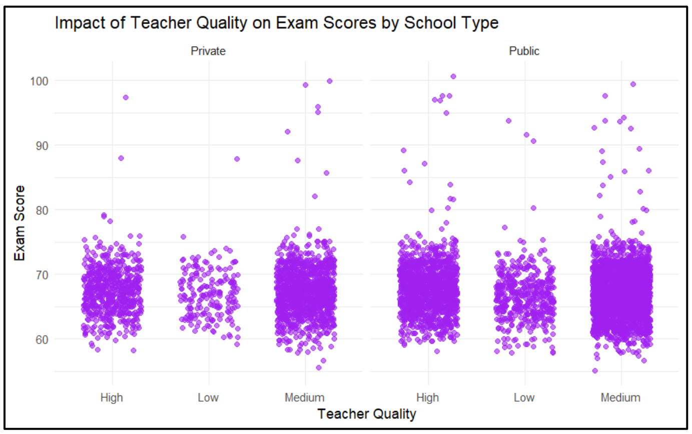


Figure 29: Scatter plot faceted by school type to show teacher quality impact

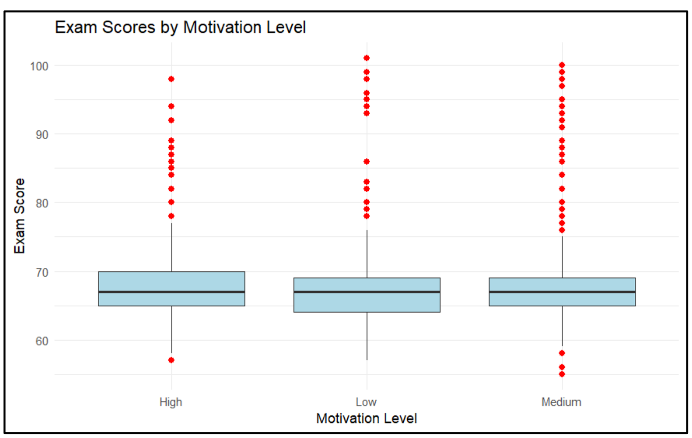


Figure 30: Box plot of Motivation Level vs. Exam Score

Contingency Table: Number of Students by Parental Involvement and Motivation Level

Rows: Parental Involvement Levels
Columns: Motivation Levels

	High	Low	Medium
High	359	574	975
Low	278	368	691
Medium	682	995	1685

Figure 31: Contingency Table - Parental Involvement vs Motivation

Proportional Table: Percentage Distribution of Students
Rows: School Types
Columns: Learning Disabilities Status (Values in %)

	No	Yes
Private	27.21	3.19
Public	62.27	7.33

Figure 32: Proportional table - School Type vs Learning Disability

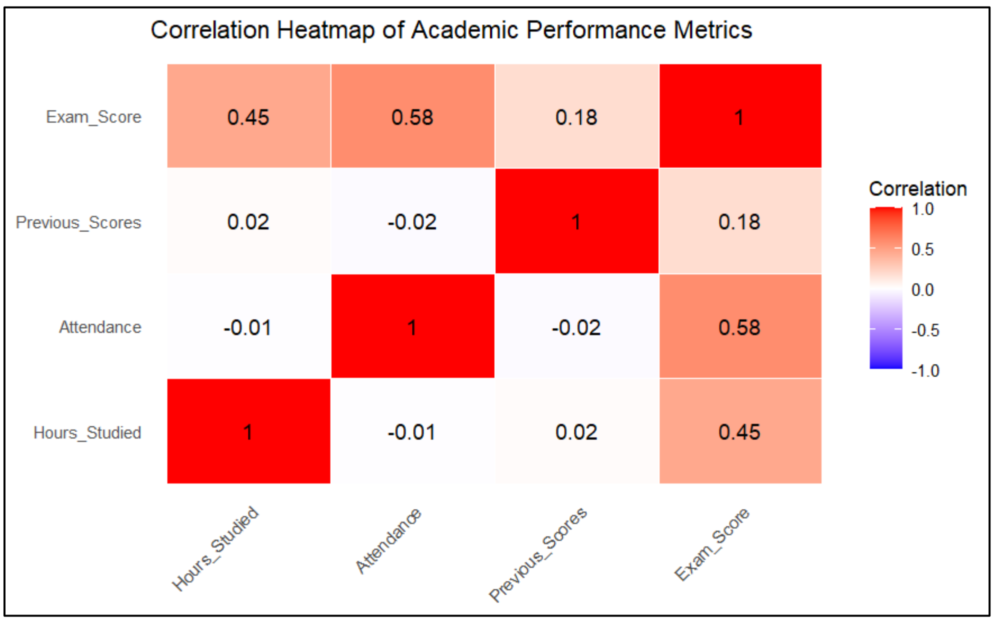


Figure 33: Correlation of Hours Studied, Attendance, Previous Scores and Exam Score

Results

Models

Initial Linear Regression Model

```
Call:
lm(formula = Exam_Score ~ ., data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.2211 -0.4419 -0.1864  0.0714 29.3231 

Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)                          41.980622   0.394170 106.504  < 2e-16 ***
Hours_Studied                         0.294973   0.005070  58.185  < 2e-16 ***
Attendance                            0.199043   0.002602  76.484  < 2e-16 ***
Parental_InvolvementLow              -1.947154   0.087326 -22.298  < 2e-16 ***
Parental_InvolvementMedium           -1.077560   0.070216 -15.346  < 2e-16 ***
Access_to_ResourcesLow               -2.029452   0.087531 -23.185  < 2e-16 ***
Access_to_ResourcesMedium            -0.999113   0.069628 -14.349  < 2e-16 ***
Extracurricular_ActivitiesYes         0.591311   0.061433   9.625  < 2e-16 ***
Sleep_Hours                          -0.012253   0.020445  -0.599 0.549005    
Previous_Scores                       0.047681   0.002101  22.692  < 2e-16 ***
Motivation_LevelLow                  -1.099353   0.087541 -12.558  < 2e-16 ***
Motivation_LevelMedium               -0.562589   0.079313  -7.093 1.49e-12 ***
Internet_AccessYes                    0.884375   0.112809   7.840 5.47e-15 ***
Tutoring_Sessions                     0.502704   0.024230  20.748  < 2e-16 ***
Family_IncomeLow                     -1.137368   0.082652 -13.761  < 2e-16 ***
Family_IncomeMedium                  -0.606801   0.083140  -7.299 3.36e-13 ***
Teacher_QualityLow                   -1.058654   0.108356  -9.770  < 2e-16 ***
Teacher_QualityMedium                -0.564183   0.067673  -8.337  < 2e-16 ***
School_TypePublic                     0.068405   0.065438   1.045 0.295913    
Peer_InfluenceNeutral                 0.556676   0.081689   6.815 1.06e-11 ***
Peer_InfluencePositive                1.041562   0.081188  12.829  < 2e-16 ***
Physical_Activity                     0.196847   0.029484   6.676 2.71e-11 ***
Learning_DisabilitiesYes             -0.886120   0.096644  -9.169  < 2e-16 ***
Parental_Education_LevelHigh School  -0.515858   0.069344  -7.439 1.18e-13 ***
Parental_Education_LevelPostgraduate  0.466063   0.086481   5.389 7.40e-08 ***
Distance_from_HomeModerate            0.394752   0.110302   3.579 0.000348 ***
Distance_from_HomeNear                0.947183   0.103579   9.145  < 2e-16 ***
GenderMale                           -0.061969   0.060907  -1.017 0.308995    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.145 on 5074 degrees of freedom
Multiple R-squared:  0.7061,	Adjusted R-squared:  0.7046 
F-statistic: 451.6 on 27 and 5074 DF,  p-value: < 2.2e-16
```

Independent Variable	Relationship with Dependent Variable	P-Value	Statistical Significance
Hours Studied	Positive	< 2e-16	Significant
Attendance	Positive	< 2e-16	Significant
Parental Involvement	Negative	< 2e-16	Significant
Access to Resources	Negative	< 2e-16	Significant
Extracurricular Activities	Positive	< 2e-16	Significant
Sleep Hours	Negative	0.549005	Insignificant
Previous Scores	Positive	< 2e-16	Significant
Motivation Level	Negative	< 1.49e-12	Significant
Internet Access	Positive	5.47e-15	Significant
Tutoring Sessions	Positive	< 2e-16	Significant

Family Income	Negative	< 3.36e-13	Significant
Teacher Quality	Negative	< 2e-16	Significant
School Type	Positive	0.295913	Insignificant
Peer Influence	Positive	< 1.06e-11	Significant
Physical Activity	Positive	2.71e-11	Significant
Learning Disabilities	Negative	< 2e-16	Significant
Parental Education Level	Negative	< 7.40e-08	Significant
Distance from Home	Positive	< 0.000348	Significant
Gender	Negative	0.308995	Insignificant

Reduced Linear Model

```
Call:
lm(formula = Exam_Score ~ . - Sleep_Hours - School_Type - Gender,
    data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-1.2152 -0.4409 -0.1882  0.0713 29.3273

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      41.909352   0.360145 116.368  < 2e-16 ***
Hours_Studied                     0.294914   0.005069  58.185  < 2e-16 ***
Attendance                        0.199025   0.002601  76.526  < 2e-16 ***
Parental_InvolvementLow          -1.945263   0.087275 -22.289  < 2e-16 ***
Parental_InvolvementMedium       -1.078873   0.070202 -15.368  < 2e-16 ***
Access_to_ResourcesLow           -2.028451   0.087446 -23.197  < 2e-16 ***
Access_to_ResourcesMedium        -0.998001   0.069573 -14.345  < 2e-16 ***
Extracurricular_ActivitiesYes     0.590383   0.061422   9.612  < 2e-16 ***
Previous_Scores                   0.047714   0.002101  22.714  < 2e-16 ***
Motivation_LevelLow              -1.100384   0.087527 -12.572  < 2e-16 ***
Motivation_LevelMedium           -0.563311   0.079302  -7.103 1.39e-12 ***
Internet_AccessYes                0.884165   0.112786   7.839 5.48e-15 ***
Tutoring_Sessions                 0.503511   0.024222  20.787  < 2e-16 ***
Family_IncomeLow                 -1.139715   0.082629 -13.793  < 2e-16 ***
Family_IncomeMedium              -0.608913   0.083120  -7.326 2.75e-13 ***
Teacher_QualityLow               -1.057721   0.108329  -9.764  < 2e-16 ***
Teacher_QualityMedium            -0.564064   0.067668  -8.336  < 2e-16 ***
Peer_InfluenceNeutral             0.554700   0.081667   6.792 1.23e-11 ***
Peer_InfluencePositive            1.042178   0.081155  12.842  < 2e-16 ***
Physical_Activity                 0.196407   0.029479   6.663 2.98e-11 ***
Learning_DisabilitiesYes         -0.884736   0.096614  -9.157  < 2e-16 ***
Parental_Education_LevelHigh School -0.515994   0.069334  -7.442 1.16e-13 ***
Parental_Education_LevelPostgraduate 0.463478   0.086425   5.363 8.56e-08 ***
Distance_from_HomeModerate        0.394432   0.110294   3.576 0.000352 ***
Distance_from_HomeNear            0.947996   0.103567   9.153  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.145 on 5077 degrees of freedom
Multiple R-squared:  0.706,	Adjusted R-squared:  0.7046
F-statistic:   508 on 24 and 5077 DF,  p-value: < 2.2e-16
```

```
Analysis of Variance Table

Model 1: Exam_Score ~ (Hours_Studied + Attendance + Parental_Involvement +
    Access_to_Resources + Extracurricular_Activities + Sleep_Hours +
    Previous_Scores + Motivation_Level + Internet_Access + Tutoring_Sessions +
    Family_Income + Teacher_Quality + School_Type + Peer_Influence +
    Physical_Activity + Learning_Disabilities + Parental_Education_Level +
    Distance_from_Home + Gender) - Sleep_Hours - School_Type -
    Gender
Model 2: Exam_Score ~ Hours_Studied + Attendance + Parental_Involvement +
    Access_to_Resources + Extracurricular_Activities + Sleep_Hours +
    Previous_Scores + Motivation_Level + Internet_Access + Tutoring_Sessions +
    Family_Income + Teacher_Quality + School_Type + Peer_Influence +
    Physical_Activity + Learning_Disabilities + Parental_Education_Level +
    Distance_from_Home + Gender
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1   5077 23361
2   5074 23349  3     11.42 0.8272 0.4787
```

Lasso Model

```
Best lambda selected by cross-validation: 0.003066671

29 x 1 sparse Matrix of class "dgCMatrix"
                                          s0
(Intercept)                      40.96626337
Hours_Studied                     0.29438791
Attendance                        0.19876817
Parental_InvolvementHigh          1.07232187
Parental_InvolvementLow          -0.86317739
Parental_InvolvementMedium         .
Access_to_ResourcesLow           -2.01370531
Access_to_ResourcesMedium        -0.98586768
Extracurricular_ActivitiesYes     0.58451805
Sleep_Hours                      -0.01053661
Previous_Scores                   0.04748149
Motivation_LevelLow              -1.07931574
Motivation_LevelMedium           -0.54473804
Internet_AccessYes                0.87165924
Tutoring_Sessions                 0.50047718
Family_IncomeLow                 -1.11883434
Family_IncomeMedium              -0.58756425
Teacher_QualityLow               -1.03985685
Teacher_QualityMedium            -0.55367719
School_TypePublic                 0.06107669
Peer_InfluenceNeutral             0.53730882
Peer_InfluencePositive            1.02333041
Physical_Activity                 0.19275249
Learning_DisabilitiesYes         -0.87574869
Parental_Education_LevelHigh School -0.51073570
Parental_Education_LevelPostgraduate  0.46114660
Distance_from_HomeModerate        0.36019437
Distance_from_HomeNear            0.91477253
GenderMale                       -0.05545794
```

Positive Impact Variables

Variable	Coefficient	Interpretation
Parental Involvement High	1.072	High parental involvement significantly boosts exam scores by ~1.07 points compared to the baseline category.
Peer Influence Positive	1.023	Positive peer influence adds ~1.02 points to the score, emphasizing the value of supportive peer groups.
Distance from Home Near	0.915	Students living near school score ~0.91 points higher than those farther away, likely due to reduced stress and better access to resources.
Internet Access Yes	0.872	Internet access increases scores by ~0.87 points, highlighting the importance of connectivity in learning.
Hours Studied	0.294	Each additional hour studied increases exam scores by ~0.29 points, reflecting the direct benefit of effort.
Attendance	0.199	Higher attendance adds ~0.20 points to the score, reinforcing the importance of consistent engagement in education.
Physical Activity	0.193	Physical activity positively influences scores by ~0.19 points, potentially through improved cognitive function and focus.
Extracurricular Activities Yes	0.585	Participation in extracurricular activities contributes ~0.59 points to the score, likely due to enhanced soft skills and discipline.
Tutoring Sessions	0.500	Each tutoring session adds ~0.50 points, demonstrating the impact of individualized support.
Parental Education Level Postgraduate	0.461	Students whose parents have postgraduate education score ~0.46 points higher, reflecting educational advantages passed on by parents.
Peer Influence Neutral	0.537	Neutral peer influence adds ~0.54 points, though less impactful than positive peer influence.

Distance from Home Moderate	0.360	Moderate proximity to school increases scores by ~0.36 points compared to far distances.

Negative Impact Variables

Variable	Coefficient	Interpretation
Access to Resources Low	-2.014	Limited access to resources decreases scores by ~2.01 points, emphasizing the critical role of learning materials and facilities.
Motivation Level Low	-1.079	Low motivation reduces scores by ~1.08 points, indicating a strong need to foster intrinsic and extrinsic motivation.
Family Income Low	-1.119	Students from low-income families score ~1.12 points lower than those from high-income families, highlighting economic disparities in education outcomes.
Teacher Quality Low	-1.040	Poor teacher quality reduces scores by ~1.04 points, underlining the importance of qualified educators.
Learning Disabilities Yes	-0.876	The presence of a learning disability reduces scores by ~0.88 points, signaling the need for tailored support mechanisms.
Parental Involvement Low	-0.863	Low parental involvement decreases scores by ~0.86 points, showing the importance of parental engagement in education.
Parental Education Level High School	-0.511	Students whose parents have only high school education score ~0.51 points lower, reflecting limited parental academic influence.
Motivation Level Medium	-0.545	Medium motivation reduces scores slightly (~0.54 points), compared to high motivation levels.
Family Income Medium	-0.588	Medium income families score ~0.59 points lower than high-income families, though less pronounced than low-income families.
Access to Resources Medium	-0.986	Medium access to resources reduces scores by ~0.99 points compared to high access.
Teacher Quality Medium	-0.554	Medium teacher quality lowers scores by ~0.55 points compared to high-quality teaching.

Zero Impact Variables

Variable	Coefficient	Interpretation
Gender Male	-0.055	Males score slightly (~0.06 points) lower than females, though the impact is minimal.
Sleep Hours	-0.011	Negligible impact on scores, suggesting sleep hours in this dataset do not significantly influence academic performance.
School Type Public	0.061	Minimal positive impact, indicating school type (public/private) is not a strong predictor of exam scores.
Parental Involvement Medium	0	Eliminated by LASSO, meaning medium parental involvement does not significantly differ from the baseline in predicting scores.

Stepwise Selection Model

```
Start:  AIC=7815.77
Exam_Score ~ Hours_Studied + Attendance + Parental_Involvement +
    Access_to_Resources + Extracurricular_Activities + Sleep_Hours +
    Previous_Scores + Motivation_Level + Internet_Access + Tutoring_Sessions +
    Family_Income + Teacher_Quality + School_Type + Peer_Influence +
    Physical_Activity + Learning_Disabilities + Parental_Education_Level +
    Distance_from_Home + Gender

                             Df Sum of Sq   RSS     AIC
- Sleep_Hours                 1       1.7 23351  7814.1
- Gender                      1       4.8 23354  7814.8
- School_Type                 1       5.0 23354  7814.9
<none>                                    23349  7815.8
- Physical_Activity           1     205.1 23554  7858.4
- Internet_Access             1     282.8 23632  7875.2
- Learning_Disabilities       1     386.9 23736  7897.6
- Extracurricular_Activities  1     426.3 23775  7906.1
- Teacher_Quality             2     539.7 23889  7928.4
- Distance_from_Home          2     573.8 23923  7935.6
- Motivation_Level            2     734.9 24084  7969.9
- Parental_Education_Level    2     758.0 24107  7974.8
- Peer_Influence              2     777.2 24126  7978.8
- Family_Income               2     903.4 24253  8005.4
- Tutoring_Sessions           1    1980.9 25330  8229.2
- Previous_Scores             1    2369.6 25719  8306.9
- Parental_Involvement        2    2380.9 25730  8307.2
- Access_to_Resources         2    2520.2 25869  8334.7
- Hours_Studied               1   15579.1 38928 10421.7
- Attendance                  1   26919.3 50268 11726.1
```

```
Step:  AIC=7814.13
Exam_Score ~ Hours_Studied + Attendance + Parental_Involvement +
    Access_to_Resources + Extracurricular_Activities + Previous_Scores +
    Motivation_Level + Internet_Access + Tutoring_Sessions +
    Family_Income + Teacher_Quality + School_Type + Peer_Influence +
    Physical_Activity + Learning_Disabilities + Parental_Education_Level +
    Distance_from_Home + Gender

                             Df Sum of Sq   RSS     AIC
- Gender                      1       4.7 23355  7813.2
- School_Type                 1       5.0 23356  7813.2
<none>                                    23351  7814.1
+ Sleep_Hours                 1       1.7 23349  7815.8
- Physical_Activity           1     205.4 23556  7856.8
- Internet_Access             1     282.7 23634  7873.5
- Learning_Disabilities       1     387.3 23738  7896.1
- Extracurricular_Activities  1     426.0 23777  7904.4
- Teacher_Quality             2     539.5 23890  7926.7
- Distance_from_Home          2     573.4 23924  7933.9
- Motivation_Level            2     735.8 24087  7968.4
- Parental_Education_Level    2     757.6 24108  7973.0
- Peer_Influence              2     779.6 24130  7977.7
- Family_Income               2     904.8 24256  8004.1
- Tutoring_Sessions           1    1983.1 25334  8228.0
- Previous_Scores             1    2372.8 25724  8305.9
- Parental_Involvement        2    2383.4 25734  8306.0
- Access_to_Resources         2    2525.9 25877  8334.2
- Hours_Studied               1   15578.2 38929 10419.8
- Attendance                  1   26954.7 50305 11727.8
```

```
Step:  AIC=7813.15
Exam_Score ~ Hours_Studied + Attendance + Parental_Involvement +
    Access_to_Resources + Extracurricular_Activities + Previous_Scores +
    Motivation_Level + Internet_Access + Tutoring_Sessions +
    Family_Income + Teacher_Quality + School_Type + Peer_Influence +
    Physical_Activity + Learning_Disabilities + Parental_Education_Level +
    Distance_from_Home

                             Df Sum of Sq   RSS     AIC
- School_Type                 1       5.1 23361  7812.3
<none>                                    23355  7813.2
+ Gender                      1       4.7 23351  7814.1
+ Sleep_Hours                 1       1.6 23354  7814.8
- Physical_Activity           1     204.6 23560  7855.7
- Internet_Access             1     281.8 23637  7872.4
- Learning_Disabilities       1     385.6 23741  7894.7
- Extracurricular_Activities  1     424.8 23780  7903.1
- Teacher_Quality             2     540.4 23896  7925.9
- Distance_from_Home          2     574.5 23930  7933.1
- Motivation_Level            2     735.4 24091  7967.3
- Parental_Education_Level    2     758.5 24114  7972.2
- Peer_Influence              2     778.7 24134  7976.5
- Family_Income               2     905.6 24261  8003.2
- Tutoring_Sessions           1    1985.2 25341  8227.4
- Previous_Scores             1    2372.1 25728  8304.7
- Parental_Involvement        2    2384.0 25739  8305.0
- Access_to_Resources         2    2527.3 25883  8333.4
- Hours_Studied               1   15577.3 38933 10418.3
- Attendance                  1   26950.9 50306 11725.9
```

```
Step:  AIC=7812.27
Exam_Score ~ Hours_Studied + Attendance + Parental_Involvement +
    Access_to_Resources + Extracurricular_Activities + Previous_Scores +
    Motivation_Level + Internet_Access + Tutoring_Sessions +
    Family_Income + Teacher_Quality + Peer_Influence + Physical_Activity +
    Learning_Disabilities + Parental_Education_Level + Distance_from_Home

                             Df Sum of Sq   RSS     AIC
<none>                                    23361  7812.3
+ School_Type                 1       5.1 23355  7813.2
+ Gender                      1       4.7 23356  7813.2
+ Sleep_Hours                 1       1.6 23359  7813.9
- Physical_Activity           1     204.2 23565  7854.7
- Internet_Access             1     282.8 23643  7871.7
- Learning_Disabilities       1     385.9 23746  7893.8
- Extracurricular_Activities  1     425.1 23786  7902.3
- Teacher_Quality             2     539.2 23900  7924.7
- Distance_from_Home          2     575.5 23936  7932.4
- Motivation_Level            2     736.4 24097  7966.6
- Parental_Education_Level    2     755.6 24116  7970.7
- Peer_Influence              2     779.5 24140  7975.7
- Family_Income               2     907.1 24268  8002.6
- Tutoring_Sessions           1    1988.2 25349  8227.0
- Parental_Involvement        2    2380.0 25741  8303.3
- Previous_Scores             1    2373.9 25734  8304.0
- Access_to_Resources         2    2522.5 25883  8331.4
- Hours_Studied               1   15577.4 38938 10417.0
- Attendance                  1   26945.9 50306 11723.9
```

```
coef(stepwise_model)
```

```
            (Intercept)                    Hours_Studied                          Attendance         Parental_InvolvementLow      Parental_InvolvementMedium
            41.90935216                       0.29491359                          0.19902456                     -1.94526254                     -1.07887333
 Access_to_ResourcesLow           Access_to_ResourcesMedium        Extracurricular_ActivitiesYes                 Previous_Scores             Motivation_LevelLow
            -2.02845066                      -0.99800073                          0.59038319                      0.04771392                     -1.10038424
  Motivation_LevelMedium                 Internet_AccessYes                   Tutoring_Sessions                 Family_IncomeLow              Family_IncomeMedium
            -0.56331139                       0.88416483                          0.50351128                     -1.13971485                     -0.60891281
     Teacher_QualityLow              Teacher_QualityMedium               Peer_InfluenceNeutral          Peer_InfluencePositive               Physical_Activity
            -1.05772148                      -0.56406350                          0.55469981                      1.04217817                      0.19640714
 Learning_DisabilitiesYes Parental_Education_LevelHigh School Parental_Education_LevelPostgraduate  Distance_from_HomeModerate          Distance_from_HomeNear
            -0.88473557                      -0.51599413                          0.46347751                      0.39443192                      0.94799644
```

Handpicked Model

```
Call:
lm(formula = Exam_Score ~ Sleep_Hours + Hours_Studied + Attendance + 
    Access_to_Resources + Family_Income, data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.8160 -1.2181 -0.1812  0.8550 31.1379 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)              47.409762   0.346575 136.795  < 2e-16 ***
Sleep_Hours              -0.040612   0.024820  -1.636    0.102    
Hours_Studied             0.291458   0.006153  47.367  < 2e-16 ***
Attendance                0.198370   0.003154  62.885  < 2e-16 ***
Access_to_ResourcesLow   -1.986387   0.106109 -18.720  < 2e-16 ***
Access_to_ResourcesMedium -0.975440   0.084386 -11.559  < 2e-16 ***
Family_IncomeLow         -1.128528   0.100302 -11.251  < 2e-16 ***
Family_IncomeMedium      -0.534823   0.100898  -5.301  1.2e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.607 on 5094 degrees of freedom
Multiple R-squared:  0.5641,	Adjusted R-squared:  0.5635 
F-statistic: 941.9 on 7 and 5094 DF,  p-value: < 2.2e-16
```

<u>Comparison of Handpicked Model and Full Model</u>

```
Analysis of Variance Table

Model 1: Exam_Score ~ Sleep_Hours + Hours_Studied + Attendance + Access_to_Resources + 
    Family_Income
Model 2: Exam_Score ~ Hours_Studied + Attendance + Parental_Involvement + 
    Access_to_Resources + Extracurricular_Activities + Sleep_Hours + 
    Previous_Scores + Motivation_Level + Internet_Access + Tutoring_Sessions + 
    Family_Income + Teacher_Quality + School_Type + Peer_Influence + 
    Physical_Activity + Learning_Disabilities + Parental_Education_Level + 
    Distance_from_Home + Gender
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)    
1   5094 34631                                  
2   5074 23349 20     11282 122.58 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

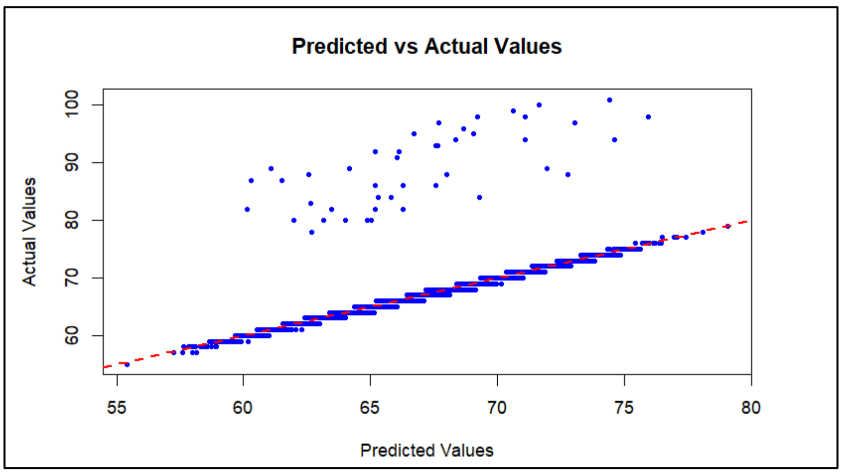
Model Comparison using AIC

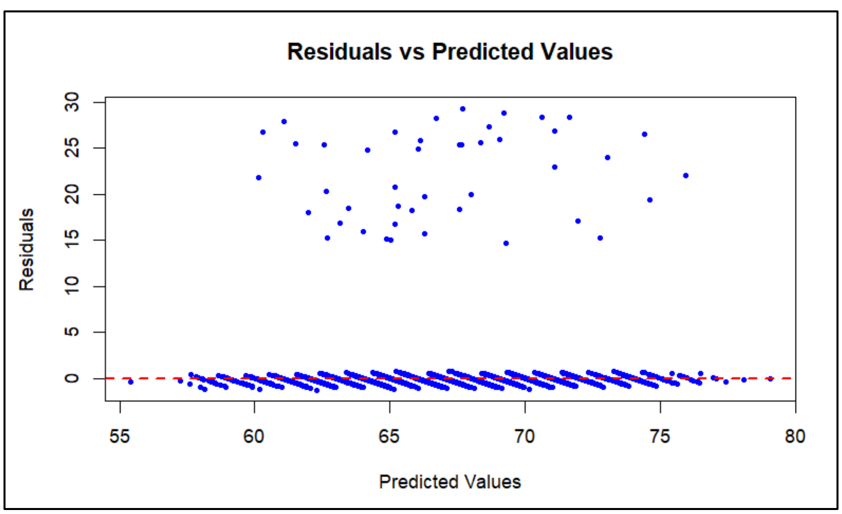
```
AIC for Reduced Model: 7812.267
AIC for Stepwise Model: 7812.267
AIC for Lasso Model: 7814.304
AIC for Custom Model: 9786.873
```

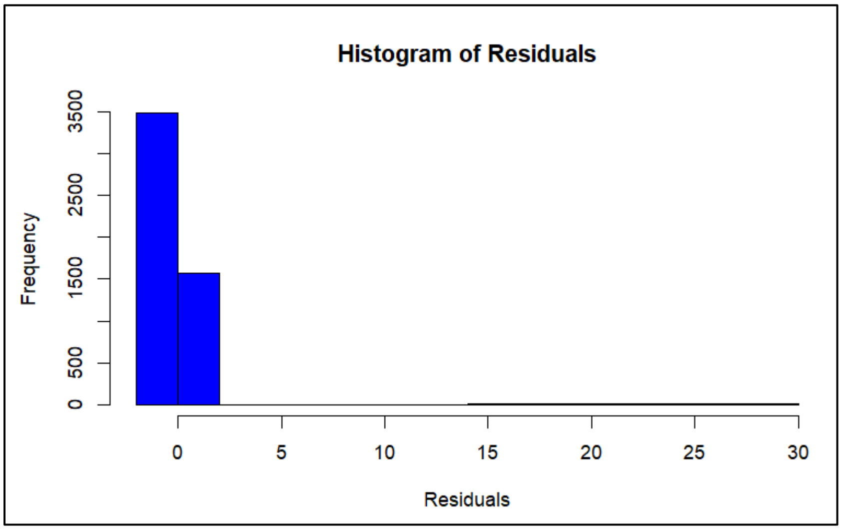
Model Comparison using R squared

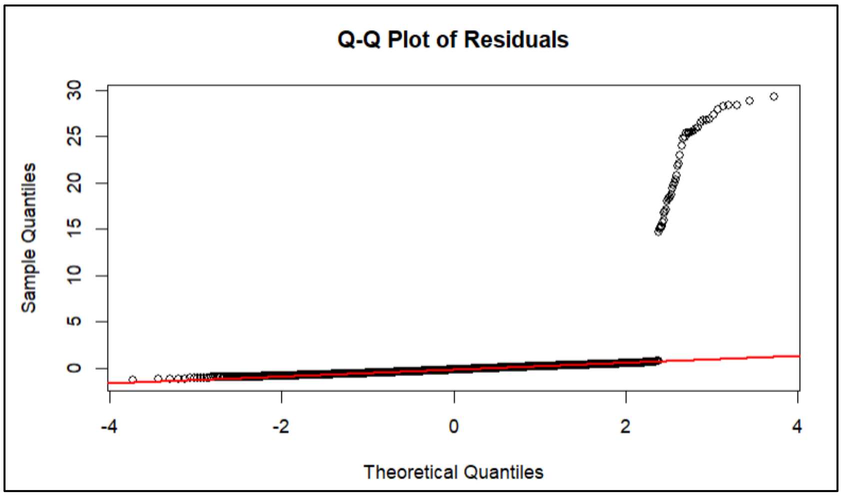
```
Adjusted R^2 for Reduced Model: 0.7046002
Adjusted R^2 for Stepwise Model: 0.7046002
Adjusted R^2 for Lasso Model: 0.7045392
Adjusted R^2 for Custom Model: 0.5635486
```

Plots of Chosen Model (Lasso) on Training Data









Lasso Results on Test Set

```
Mean Squared Error (MSE) for Lasso Model on Test Data: 3.028119
Root Mean Squared Error (RMSE): 1.740149
```

609 Final_Project

Anjiya Adil, Aryan Gandhi, Avery Funston, Utkarsh Singh

2024-11-03

Loading in the data

```
data =
read.csv('C:/Users/anjiy/Downloads/StudentPerformanceFactors.csv')
summary(data)
```

Data Preprocessing

Indentifying and removing rows with missing values

```
cat("Number of rows with missing data: ", length(data[data == ""]))

# Load necessary libraries
library(ggplot2)

data[data == ""] <- NA

# Calculate the percentage of missing data for each column
missing_percentage <- round(colSums(is.na(data)) / nrow(data) * 100,
2)
missing_percentage <- missing_percentage[missing_percentage > 0]  #
Filter columns with missing data

# Convert to a data frame for plotting
missing_df <- data.frame(
  column = names(missing_percentage),
  percentage = missing_percentage
)

missing_df

# Create the bar graph
ggplot(missing_df, aes(x = reorder(column, -percentage), y =
percentage)) +
  geom_bar(stat = "identity", fill = "red") +
  labs(
    title = "Percentage of Missing Data by Column",
    x = "Columns",
    y = "Percentage of Missing Data"
  ) +
  theme_minimal() +
```

```
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Rotate
x-axis labels

# set empty rows to NA values and omit them from the dataset
data <- na.omit(data)
```

Converting categorical variables to factor data type

```
library(dplyr)
library(caret)


# Convert categorical features to factors first
categorical_features <- c("Parental_Involvement",
"Access_to_Resources",
                          "Extracurricular_Activities",
"Motivation_Level",
                          "Internet_Access", "Family_Income",
"Teacher_Quality", "School_Type", "Peer_Influence",
"Learning_Disabilities", "Parental_Education_Level",
"Distance_from_Home", "Gender")

data[categorical_features] <- lapply(data[categorical_features],
as.factor)
```

Exploratory Data Analysis

```
# Cross-tabulations for categorical variables
table(data$Parental_Involvement, data$Motivation_Level)
prop.table(table(data$School_Type, data$Learning_Disabilities))

# Histogram of Exam Scores
ggplot(data, aes(x = Exam_Score)) +
  geom_histogram(binwidth = 5, fill = "skyblue", color = "black") +
  labs(title = "Distribution of Exam Scores",
       x = "Exam Score",
       y = "Frequency") +
  theme_minimal()

# Box plot of Exam Scores
ggplot(data, aes(y = Exam_Score)) +
  geom_boxplot(fill = "coral") +
  labs(title = "Box Plot of Exam Scores",
       y = "Exam Score") +
  theme_minimal()

library(ggplot2)
library(GGally)
library(dplyr)

# List of numeric variables to plot
```

```r
numeric_vars <- data %>% select_if(is.numeric)

# Create histograms
for (var in names(numeric_vars)) {
  print(ggplot(data, aes_string(x = var)) +
          geom_histogram(binwidth = 5, fill = "blue", alpha = 0.7) +
          labs(title = paste("Distribution of", var), x = var, y =
"Frequency") +
          theme_minimal())
}



# Scatter Plot: Hours Studied vs Exam Score
ggplot(data, aes(x = Hours_Studied, y = Exam_Score)) +
  geom_point(color = "darkgreen") +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +  # Optional
linear trend line
  labs(title = "Hours Studied vs Exam Score",
       x = "Hours Studied",
       y = "Exam Score") +
  theme_minimal()

#Here, Low range [0-70), Medium range[70-84) and High range[85-10]
temp_data = data
temp_data$attendance_bin <- cut(data$Attendance,
                           breaks = c(0, 70, 85, 100),
                           labels = c("Low", "Medium", "High"),
                           right = TRUE)

attendance_scores_binned <- temp_data %>%
  group_by(attendance_bin) %>%
  summarize(mean_exam_score = mean(Exam_Score, na.rm = TRUE))

sum(is.na(temp_data))

# Bar plot for attendance bins with legend
ggplot(attendance_scores_binned,
       aes(x = attendance_bin, y = mean_exam_score, fill =
attendance_bin)) +
  geom_bar(stat = "identity",
           color = "black",
           width = 0.7) +
  scale_fill_manual(values = c("Low" = "#9ecae1",
                               "Medium" = "#6baed6",
                               "High" = "#2171b5"),
                    name = "Attendance Levels",
                    labels = c("Low (60-70%)", "Medium (70-85%)",
"High (85-100%)")) +
  labs(title = "Average Exam Score by Attendance Level",
       x = "Attendance Level",
```

```r
       y = "Average Exam Score") +
  geom_text(aes(label = round(mean_exam_score, 2)),
            position = position_dodge(width = 0.9),
            vjust = -0.5,
            fontface = "bold") +
  theme_minimal() +
  theme(legend.position = "right",
        plot.title = element_text(hjust = 0.5, face = "bold"),
        axis.title = element_text(face = "bold")) +
  coord_cartesian(ylim = c(0,
max(attendance_scores_binned$mean_exam_score) * 1.1))

# Scatter plot faceted by school type to show teacher quality impact
ggplot(data = data,
       aes(x = Teacher_Quality, y = Exam_Score)) +
  geom_jitter(width = 0.3, alpha = 0.6, color = "purple") +
  geom_smooth(method = "lm", color = "darkred", se = FALSE) +
  labs(title = "Impact of Teacher Quality on Exam Scores by School
Type",
       x = "Teacher Quality",
       y = "Exam Score") +
  facet_wrap(~ School_Type) +
  theme_minimal()

# Box plot of Motivation Level vs. Exam Score
ggplot(data, aes(x = factor(Motivation_Level), y = Exam_Score)) +
  geom_boxplot(fill = "lightblue", outlier.color = "red",
outlier.shape = 16, outlier.size = 2) +  # Boxplot with outliers
  labs(title = "Exam Scores by Motivation Level",
       x = "Motivation Level",
       y = "Exam Score") +
  theme_minimal()


# List of categorical variables
categorical_vars <- data %>% select_if(is.factor)

# Loop through each categorical variable and create a bar plot
for (var in colnames(categorical_vars)) {
  print(
    ggplot(data, aes_string(x = var)) +
      geom_bar(fill = "steelblue", color = "black", alpha = 0.7) +
      labs(title = paste("Distribution of", var), x = var, y =
"Count") +
      theme_minimal() +
      theme(axis.text.x = element_text(angle = 45, hjust = 1)) #
Rotate x-axis labels for readability
  )
}
```

```
# For the first table (Contingency Table)
cat("\nContingency Table: Number of Students by Parental Involvement
and Motivation Level\n")
cat("\nRows: Parental Involvement Levels")
cat("\nColumns: Motivation Levels\n\n")

parental_motivation_table <- table(data$Parental_Involvement,
data$Motivation_Level)
print(parental_motivation_table)

# For the second table (Proportional Table)
cat("\nProportional Table: Percentage Distribution of Students\n")
cat("Rows: School Types")
cat("\nColumns: Learning Disabilities Status (Values in %)\n\n")

# Calculate proportions and convert to percentages
school_disability_prop <- prop.table(table(data$School_Type,
data$Learning_Disabilities)) * 100
# Round to 2 decimal places
school_disability_prop <- round(school_disability_prop, 2)
print(school_disability_prop)

# Install and load required packages if not already installed
if (!require(ggplot2)) install.packages("ggplot2")
if (!require(reshape2)) install.packages("reshape2")
library(ggplot2)
library(reshape2)

# Select the variables we want to correlate
variables <- c("Hours_Studied", "Attendance", "Previous_Scores",
"Exam_Score")
cor_data <- data[, variables]

# Calculate correlation matrix
cor_matrix <- round(cor(cor_data), 2)

# Convert correlation matrix to long format for ggplot
cor_melted <- melt(cor_matrix)

# Create the heatmap
ggplot(cor_melted, aes(x = Var1, y = Var2, fill = value)) +
  geom_tile(color = "white") +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                       midpoint = 0, limit = c(-1,1), name =
"Correlation") +
  geom_text(aes(label = value), color = "black", size = 4) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        axis.title = element_blank(),
        panel.grid.major = element_blank(),
```

```
        panel.border = element_blank(),
        panel.background = element_blank(),
        axis.ticks = element_blank(),
        legend.position = "right") +
  ggtitle("Correlation Heatmap of Academic Performance Metrics")

# Summarize data by extracurricular and physical activity, taking mean
of exam_score
heatmap_data <- data %>%
  group_by(Extracurricular_Activities, Physical_Activity) %>%
  summarize(mean_exam_score = mean(Exam_Score, na.rm = TRUE))

# Plot heatmap
ggplot(heatmap_data, aes(x = Extracurricular_Activities, y =
Physical_Activity, fill = mean_exam_score)) +
  geom_tile(color = "white") +
  scale_fill_gradient(low = "lightblue", high = "darkblue") +
  labs(title = "Impact of Extracurricular Activities and Physical
Activity on Exam Scores",
       x = "Extracurricular Activities Level",
       y = "Physical Activity Level",
       fill = "Avg Exam Score") +
  theme_minimal()

library(GGally)

# Select only numeric columns from the dataset
numeric_vars <- data[sapply(data, is.numeric)]

# Create pairwise plots for each numeric variable
pairs(numeric_vars)
```

Training Phase

Splitting the data into train and test datasets

```
# Set a seed for reproducibility
set.seed(123)

# Create subset of 80% of the data for training
sample_size <- floor(0.8 * nrow(data))

train_indices <- sample(seq_len(nrow(data)), size = sample_size)

# Subset the data into training and testing sets
train_data <- data[train_indices, ]
test_data <- data[-train_indices, ]

sum(is.na(train_data))
```

```
cat("Size of training data: ", nrow(train_data), "\n")
cat("Size of test data: ", nrow(test_data), "\n")
nrow(data) == nrow(train_data) + nrow(test_data)
```

Full Model with all the predictors

```
lmod <-lm(Exam_Score ~ ., data=train_data)
summary(lmod)
```

Insignifcant Variables to the Response (p-value > 0.05)

Sleep_Hours School_TypePublic GenderMale

Creating a reduced model based on the p-values

```
reduced_model <- lm(Exam_Score ~ . - Sleep_Hours -School_Type -
Gender, data = train_data)

f_test <- anova(reduced_model, lmod)
print(f_test)

summary(reduced_model)
```

Results

Since the p-value is >= 0.05 we conclude that the additional predictors we can reasonably conclude that removing these variables does not significantly reduce the model's predictive power. Thus, the reduced model (without Sleep_Hours, School_Type, and Gender) may be preferred for simplicity.

L1 Regression Model

```
library(glmnet)

x <- model.matrix(Exam_Score ~ . -1, data = train_data)  # `-1`
removes the intercept
y <- train_data$Exam_Score


cv_lasso <- cv.glmnet(x, y, alpha = 1, nfolds = 10)


best_lambda <- cv_lasso$lambda.min
cat("\nBest lambda selected by cross-validation:", best_lambda, "\n\
n")


lasso_model <- glmnet(x, y, alpha = 1, lambda = best_lambda)


print(coef(lasso_model))
```

Stepwise Selection Model

```
library(MASS)

# Fit a full model with all predictors
full_model <- lm(Exam_Score ~ ., data = train_data)

stepwise_model <- stepAIC(full_model, direction = "both")

print(coef(stepwise_model))
```

Model with handpicked predictors

This model is created by hand picking predictors based on factors outlined in previous research.

```
handpick_model <- lm(Exam_Score ~ Sleep_Hours + Hours_Studied +
Attendance + Access_to_Resources + Family_Income, train_data)

summary(handpick_model)
```

Use F-Test to compare model with handpicked predictors with full model as baseline

```
f_test2 <- anova(handpick_model, lmod)
print(f_test2)
```

F-test shows that other predictors in full model are significant to the response. Thus, it may be necessary to include some other predictors.

Model Comparison

Comparing models using AIC

```
reduced_rss <- sum(residuals(reduced_model)^2)
n <- nrow(train_data)
reduced_k <- length(coef(reduced_model))

aic_reduced <- n * log(reduced_rss / n) + 2 * reduced_k

step_rss <- sum(residuals(stepwise_model)^2)
n <- nrow(train_data)
step_k <- length(coef(stepwise_model))

aic_stepwise <- n * log(step_rss / n) + 2 * step_k

hp_rss <- sum(residuals(handpick_model)^2)
n <- nrow(train_data)
hp_k <- length(coef(handpick_model))

aic_handpick <- n * log(hp_rss / n) + 2 * hp_k

lasso_predictions <- predict(lasso_model, newx = x, s = best_lambda)
lasso_predictions <- as.numeric(lasso_predictions)
```

```
rss_lasso <- sum((y - lasso_predictions)^2)

n <- length(y)

p_lasso <- sum(coef(lasso_model, s = best_lambda) != 0) - 1  # Exclude
the intercept

aic_lasso <- n * log(rss_lasso / n) + 2 * p_lasso

# Display AIC values for comparison
cat("AIC for Reduced Model:", aic_reduced, "\n")
cat("AIC for Stepwise Model:", aic_stepwise, "\n")
cat("AIC for Lasso Model:", aic_lasso, "\n")
cat("AIC for Custom Model:", aic_handpick, "\n")
```

Based on the AIC for each model, based on these results the Reduced, Stepwise, and Lasso Model have the best balance between goodness of fit and model complexity, as they have low negligible AIC values of 7812.267 and 7814.304.

```
# Adjusted R^2 for the Reduced Model
adj_r2_reduced <- summary(reduced_model)$adj.r.squared

# Adjusted R^2 for the Stepwise Model
adj_r2_stepwise <- summary(stepwise_model)$adj.r.squared

# Adjusted R^2 for the Custom Model
adj_r2_handpick <- summary(handpick_model)$adj.r.squared

# Adjusted R^2 for the Lasso Model
# Manually calculate adjusted R^2 for Lasso since glmnet does not
provide it directly
lasso_predictions <- as.numeric(predict(lasso_model, newx = x, s =
best_lambda))
rss_lasso <- sum((y - lasso_predictions)^2)
tss <- sum((y - mean(y))^2)
n <- length(y)
p_lasso <- sum(coef(lasso_model, s = best_lambda) != 0) - 1
r2_lasso <- 1 - (rss_lasso / tss)
adj_r2_lasso <- 1 - ((1 - r2_lasso) * (n - 1) / (n - p_lasso - 1))


cat("Adjusted R^2 for Reduced Model:", adj_r2_reduced, "\n")
cat("Adjusted R^2 for Stepwise Model:", adj_r2_stepwise, "\n")
cat("Adjusted R^2 for Lasso Model:", adj_r2_lasso, "\n")
cat("Adjusted R^2 for Custom Model:", adj_r2_handpick, "\n")
```

Based on the adjusted R^2 values the best model is the reduced model or model created using stepwise selection. However, with the Lasso model having a very close second best adjusted R^2 value of 0.7045392.

Final Model

Based on our comparisons the best model with the best balance of predictive power, goodness of fit and interpretability is the Lasso Model

Model Diagonstics

```
plot(lasso_predictions, y,
     main = "Predicted vs Actual Values",
     xlab = "Predicted Values",
     ylab = "Actual Values",
     pch = 20, col = "blue")
abline(0, 1, col = "red", lwd = 2, lty = 2)  # Add y=x line for
perfect predictions
```

```
residuals <- y - lasso_predictions
plot(lasso_predictions, residuals,
     main = "Residuals vs Predicted Values",
     xlab = "Predicted Values",
     ylab = "Residuals",
     pch = 20, col = "blue")
abline(h = 0, col = "red", lwd = 2, lty = 2)  # Horizontal line at
residual = 0
```

While there are a few outliers, it seems that residuals are mostly centered around 0 and there does not seem to be heteroskedasticity. Thus, linearity, constant variance and independence assumptions hold.

```
hist(residuals,
     main = "Histogram of Residuals",
     xlab = "Residuals",
     breaks = 20, col = "blue")
```

```
qqnorm(residuals, main = "Q-Q Plot of Residuals")
qqline(residuals, col = "red", lwd = 2)
```

While there are a few outliers at the tails of the plot, based on the qqplot and line, it seems that for the most part the residuals are normally distributed. Thus, normality assumption holds.

Final Model Results on Test Set

```
x_test <- model.matrix(Exam_Score ~ . - 1, data = test_data)

y_test <- test_data$Exam_Score
```

```
lasso_test_preds <- as.numeric(predict(lasso_model, newx = x_test, s =
best_lambda))


mse_lasso <- mean((y_test - lasso_test_preds)^2)


cat("Mean Squared Error (MSE) for Lasso Model on Test Data:",
mse_lasso, "\n")

rmse_lasso <- sqrt(mse_lasso)
cat("Root Mean Squared Error (RMSE):", rmse_lasso, "\n")
```

This means that, on average, the model's predictions deviate from the actual Exam_Score by approximately 1.74 percent from values in the test data.

Model Diagnostic on test set

```
plot(lasso_test_preds, y_test,
     main = "Predicted vs Actual Values",
     sub = paste("RMSE =", round(sqrt(mse_lasso), 3)),
     xlab = "Predicted Values",
     ylab = "Actual Values",
     pch = 20, col = "blue")
abline(0, 1, col = "red", lwd = 2, lty = 2)  # Add y=x line for
perfect predictions

test_residuals <- y_test - lasso_test_preds
plot(lasso_test_preds, test_residuals,
     main = "Residuals vs Predicted Values",
     xlab = "Predicted Values",
     ylab = "Residuals",
     pch = 20, col = "blue")
abline(h = 0, col = "red", lwd = 2, lty = 2)  # Horizontal line at
residual = 0

hist(test_residuals,
     main = "Histogram of Residuals",
     xlab = "Residuals",
     breaks = 20, col = "blue")

qqnorm(test_residuals, main = "Q-Q Plot of Residuals")
qqline(test_residuals, col = "red", lwd = 2)
```