AN ANALYTICAL STUDY OF STUDENT PERFORMANCE USING STATISTICAL TECHNIQUES

Group: G-04, Section: F

Submitted To
Dr. Ashraf Uddin
Assistant Professor, CS, AIUB

AI Usage Declaration

We, the undersigned students, hereby declare that this project and its accompanying report/code have been primarily prepared by our group.

We acknowledge that the use of Artificial Intelligence (AI) tools such as ChatGPT, GitHub Copilot, Grammarly, or similar systems was permitted only to assist in learning, idea generation, code debugging, or language improvement.

We further declare that:

1. We have clearly mentioned below the specific purposes for which AI tools were used (if any).
2. The core design, implementation, analysis, and conclusions are our own original work.
3. We collectively take full academic responsibility for the content of this submission.

AI Usage Details:

☐ AI tools were used for the following purposes (please specify clearly):

---

1. To refine language, grammar and presentation, ensuring clarity and readability.
2. To provide assistance in code debugging or optimization, where necessary.

	Name	**Student ID**	**Signature with Date**
1.	ABDULLAH AL HASIB	22-47055-1	
2.	ANINDITA BHATTACHARJEE	22-48606-3	
3.	DIBAJIT ROY	22-48569-3	
4.	MD ASHFAK UZZAMAN KHAN LIMON	22-49976-3	

Table of Contents

List of Table and Figures

Data-Driven Exploration of Academic, Behavioral and Environmental Predictors of Achievement

This project focuses on analyzing the Student Performance dataset to understand the academic, behavioral and environmental factors influencing exam outcomes. The dataset contains multiple numerical and categorical variables that represent study habits, attendance, sleep quality, parental involvement and several motivational attributes. The goal of this project is to examine how these factors contribute to student's final exam scores and to uncover meaningful patterns through statistical analysis, feature engineering and visualization. Using R programming, the dataset was explored through descriptive analytics, cleaned to ensure consistency, transformed to improve usability and analyzed through both correlation-based and information-based feature selection. The study provides insights into which factors most strongly predict student performance and prepares the dataset for future predictive modeling. This analysis demonstrates the usefulness of data-driven approaches in understanding academic outcomes and supporting evidence-based educational decisions.

Dataset Description

The dataset used in this project was collected from Kaggle and contains a comprehensive set of student performance related variables. It includes both numerical features-such as Exam_Score, Hours_Studied, Attendance, Sleep_Hours, Previous_Scores, Tutoring_Sessions, Physical_Activity and categorical attributes including Gender, Parental_Involvement, Access_to_Resources, Extracurricular_Activities, Motivation_Level and Peer_Influence. These variables collectively capture academic outcomes as well as behavioral and environmental factors that may influence performance. The dataset's structure, consisting of 6,607 rows and 20 columns, makes it suitable for detailed exploratory analysis, data visualization, data preprocessing, and feature engineering to better understand the key determinants of student achievement.

Dataset Source: https://www.kaggle.com/datasets/lainguyn123/student-performance-factors

A.Dataset Understanding

1.Load Dataset

Import the raw student performance data into R for analysis using functions like read.csv ().

2.Display the first few rows (10 rows)

Use head () to quickly inspect sample records and understand the structure and contents of the dataset.

3.Show Shape (rows x columns)

Retrieve and print the number of observations and variables with nrow() and ncol() to understand dataset size and get 6607 rows and 20 columns.

4.Display data types of each column

Use str () to examine variable types, helping determine how each attribute should be processed.

5.Generate Basic Descriptive Statistics

Calculate summary measures (mean, median, mode, standard deviation, min, max, count) to understand the distribution of numerical variables.

6.Identify Categorical and Numerical Features

Separate variables based on their data types to guide preprocessing, visualization, and feature selection steps.

B.Data Exploration and Visualization

Univariate Analysis

- **1.Histogram plot**

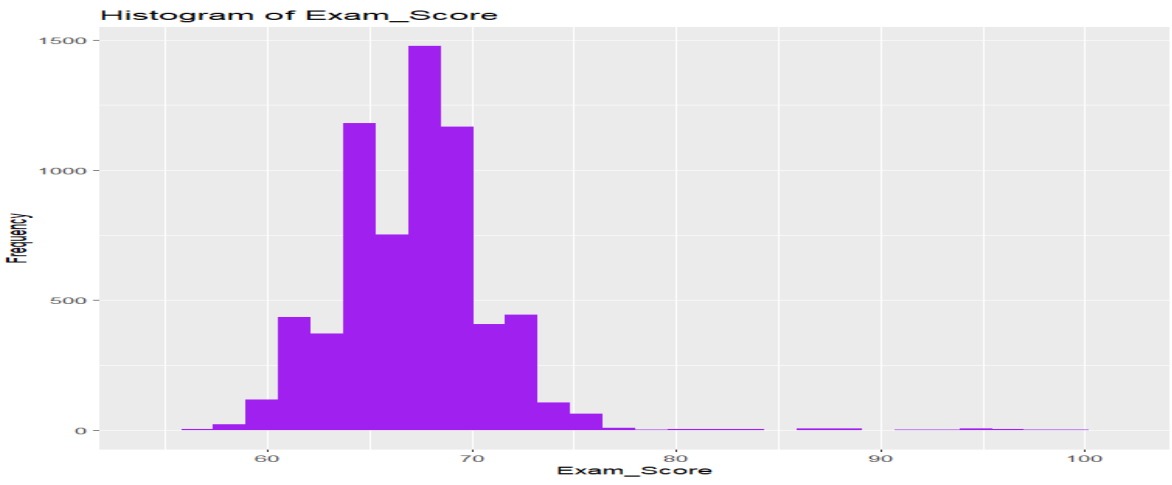

Figure 1. Histogram of Student Exam score

This shows maximum student exam score is under 70. There is very small portion of student getting up to 80 marks.

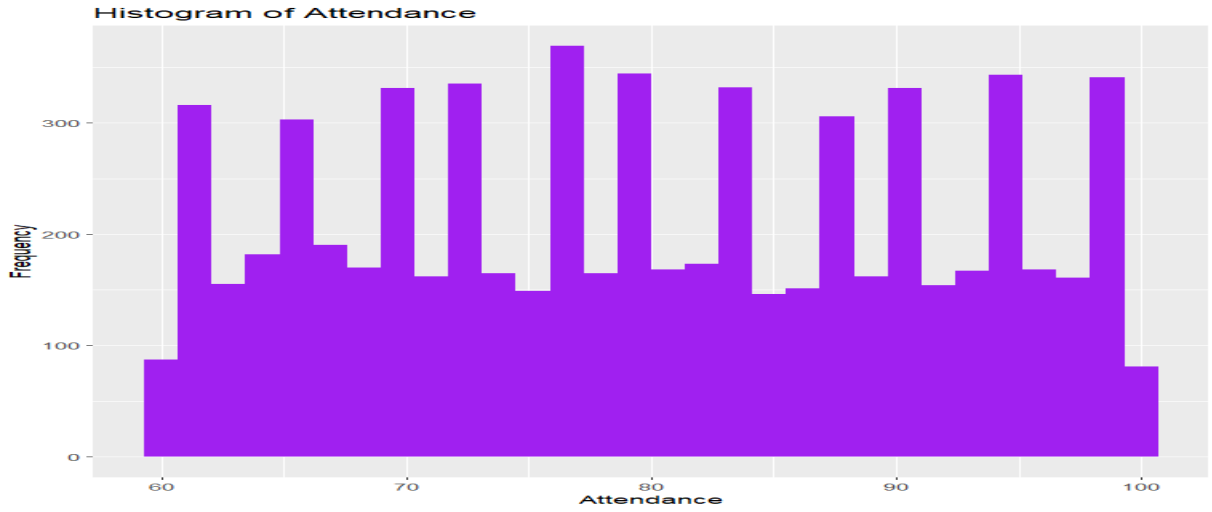

Figure 2. Histogram of student attendance at school

This histogram shows attendance ranging from 60 to 100 with several ups and downs creating a bumpy pattern. People tend to group around certain attendance numbers, likely round numbers like 70, 80, 90, rather than spreading out evenly.

- **2.Bar Chart**

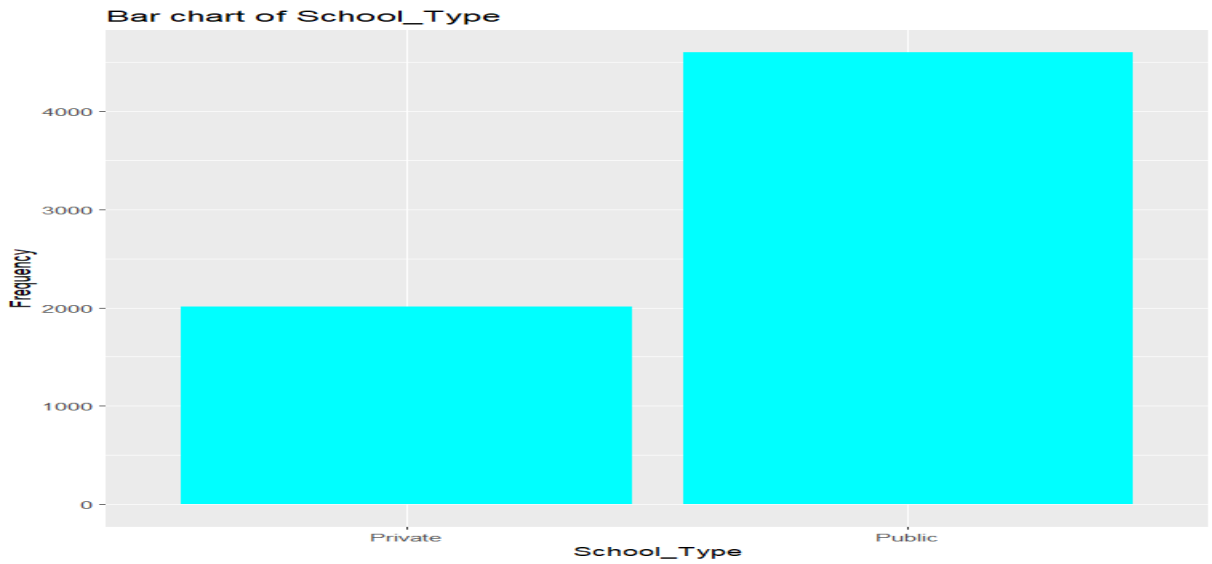

Figure 3. Bar chart of School type where number of public school are more rather than private school

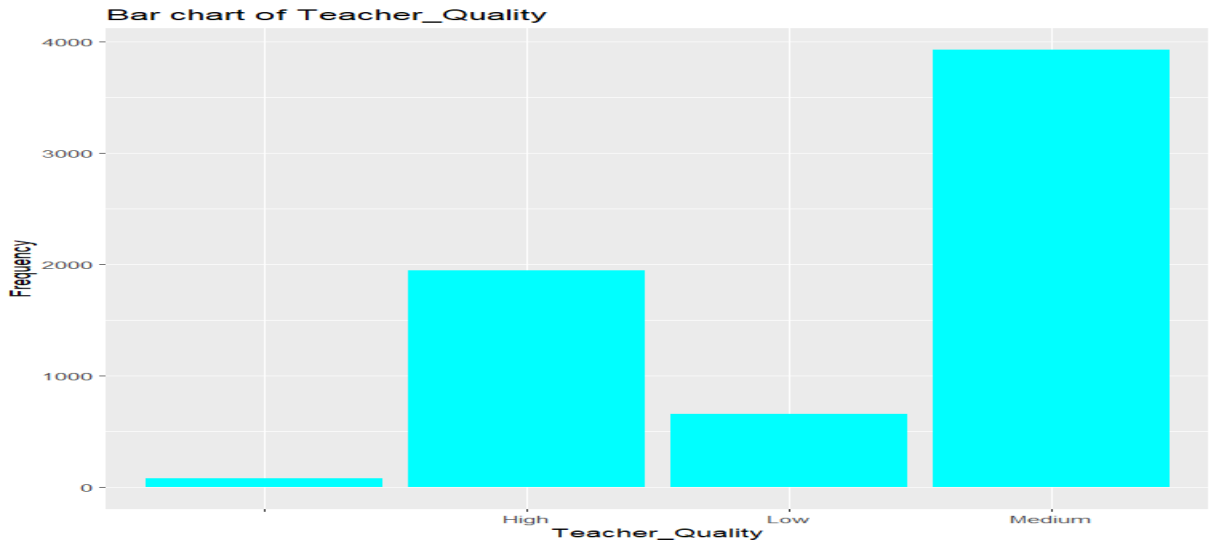

Figure 4. Bar chart of School type where number of public school are more rather than private school

- **3.Box Plot**

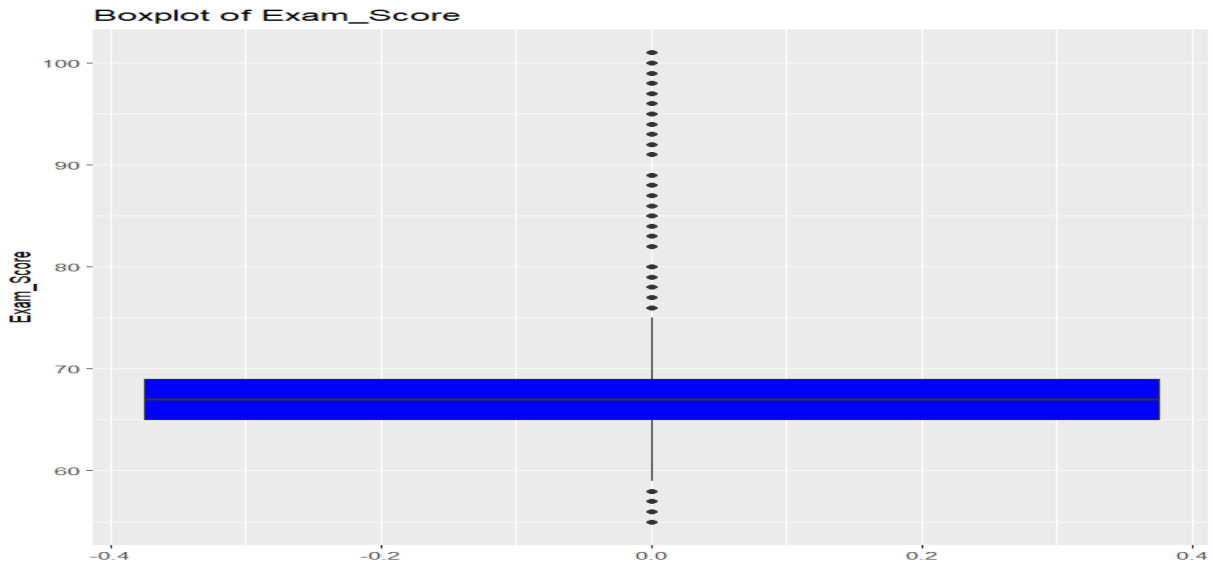

Figure 5: Box plot of student exam score

Most exam scores are clustered between 65-69 with many outliers above 75 reaching up to about 101.

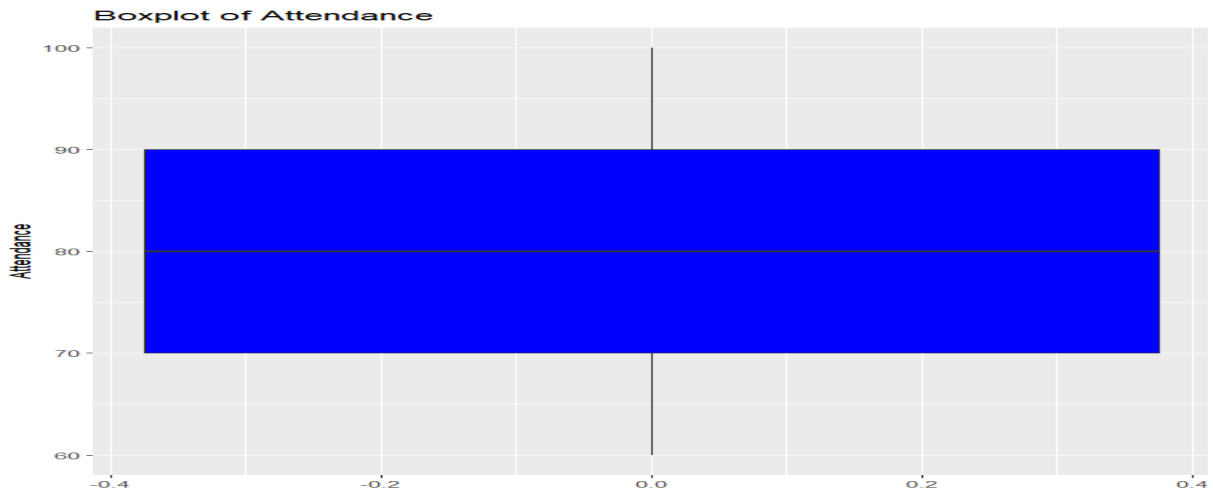

Figure 6. Boxplot of student attendance at school

Attendance is tightly packed between 70-90 with very few outliers, indicating most students have consistent attendance in this range.

- **4.Frequency of categorical features**

Categorical features shows how often each category appears in the dataset, helping identify dominant groups, class imbalance and distribution patterns across qualitative variable.

Bivariate Analysis

- **1.Heatmap**

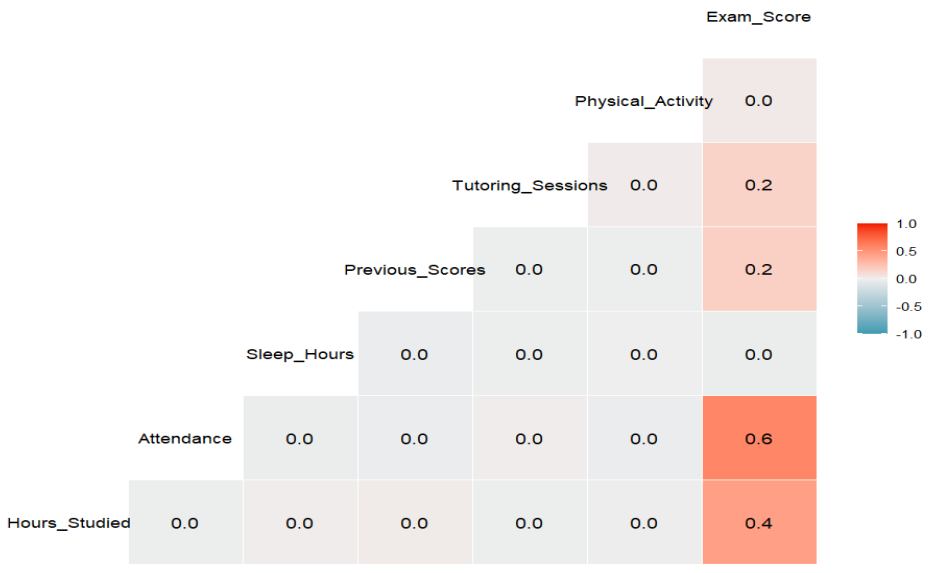

Figure 7: Heatmap

Heatmap shows the correlation between different student factors and exam scores. Attendance (0.6) and Hours Studied (0.4) have moderate positive relationships with exam scores, while Tutoring Sessions (0.2) and Previous Scores (0.2) show weak positive connections, and Physical Activity and Sleep Hours show no relationship (0.0) with exam performance.

- **2.Scatter plot**

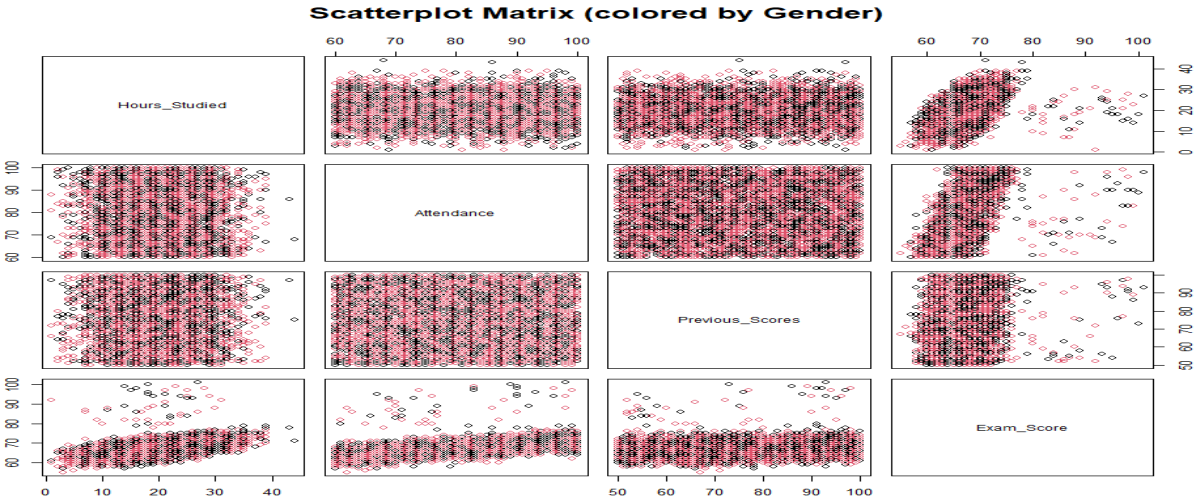

Figure 8: Scatter plot matrix which is colored by Gender

Scatterplot matrix shows the relationships between multiple student variables (Hours Studied, Attendance, Previous Scores, and Exam Score) with data points colored by gender. The diagonal cells display the variable names, revealing patterns like the positive relationship between attendance and exam scores and how gender differences appear across these academic factors.

- **3.Box plot- Categorical vs Numerical Feature**

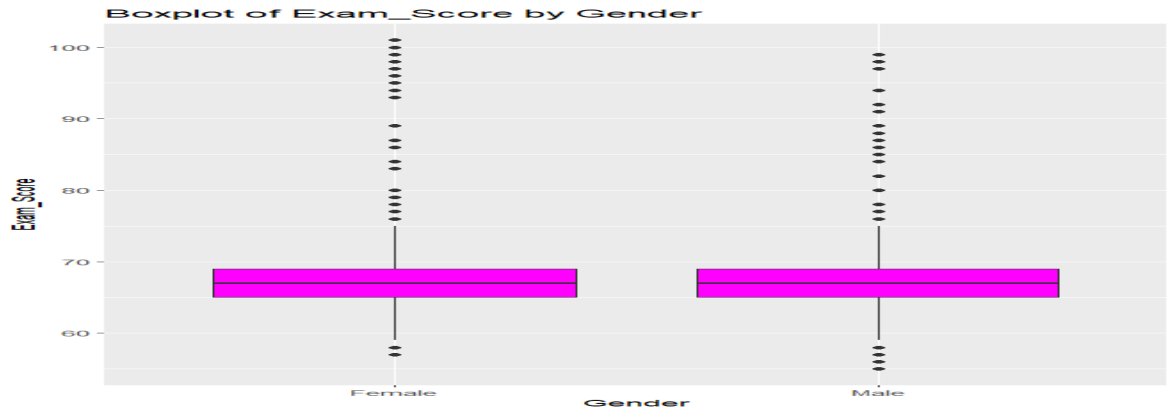

Figure 9. Exam Score between Female and Male Student where both gender marks in 65-70

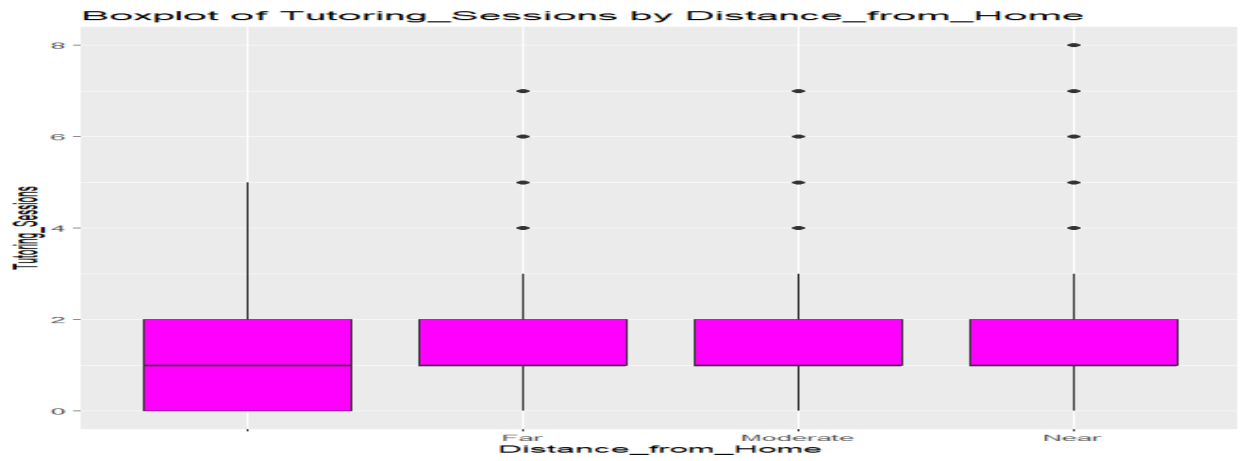

Figure 10. Box plot between Tutoring_Session vs Distance_From_Home

This boxplot shows tutoring sessions grouped by distance from home (Close, Far, Moderate, and Near). All four distance categories have similar patterns with most students attending 0-2 tutoring sessions and a few outliers attending 4-5 sessions, indicating that distance from home doesn't significantly affect tutoring attendance.

- **4.Skewness of column**

Most numeric features are roughly symmetric, except Exam_Score, which is positively skewed, indicating a few students scored much higher than the majority.

C.Data Preprocessing

1. Handling Missing Values

The preprocessing stage began with identifying missing values using colSums(is.na(data)). All rows containing missing entries were removed using na.omit(), resulting in a fully clean dataset with zero remaining NA values. Duplicate records were also eliminated to prevent redundancy and ensure analytical accuracy. After cleaning, the dataset was further filtered to retain only students with an Exam_Score greater than 60 and selected variables were used to create a new feature, Study_Efficiency = Exam_Score / Hours_Studied, providing an additional measure of student productivity.

2. Handling Outliers

Outliers in the numerical variables were first detected visually using boxplots. To formally identify them, the IQR method was applied by calculating lower and upper bounds based on Q1 − 1.5×IQR and Q3 + 1.5×IQR. The number of outliers in each numeric feature was reported. Instead of removing outliers, the dataset was treated using capping (winsorization), where values below the lower bound were replaced with the lower limit and values above the upper bound were replaced with the upper limit. This method preserved the dataset's size while reducing the influence of extreme values.

3. Data Conversion (Encoding Categorical Variables)

Categorical columns were converted into numerical format using one-hot encoding via the fastDummies::dummy_cols() function. To avoid the dummy variable trap, the first dummy of each categorical feature was removed and original categorical columns were dropped. The resulting dataset became entirely numeric, making it compatible with various statistical models and machine learning algorithms.

4. Data Transformation

Data transformation was performed using multiple scaling techniques. First, Z-score standardization was applied using the scale() function, transforming numeric features into values with mean 0 and standard deviation 1. Next, Min-Max normalization was implemented to scale variables between 0 and 1. To address skewness, the dataset was evaluated using the skewness() function. Variables with skewness greater than 1 were log-transformed, while those with skewness above 0.5 received square-root transformation. These adjustments improved the distributional normality of the numeric features.

5. Feature Selection

Feature selection was carried out using three different approaches. First, correlation analysis was performed to identify the most strongly associated features for each numeric variable. Second, variance thresholding was applied, retaining features with variance greater than 0.01 and removing low-variance attributes. Finally, mutual information (information gain) was computed using the information_gain() function from the FSelectorRcpp package for four target variables: Exam_Score, Previous_Scores, Hours_Studied, and Attendance. Only features with positive

mutual information scores were selected as significant predictors, ensuring that the final dataset retained the most meaningful attributes for analysis.

Conclusion

This project analyzed the factors influencing student academic performance using detailed statistical analysis and data preprocessing techniques. The dataset underwent rigorous cleaning, exploration and transformation to ensure accurate insights. Visualizations highlighted meaningful distribution patterns, while correlation and mutual information analyses identified the most influential predictors of exam performance. The results suggest that Hours Studied, Previous Scores and Attendance are the strongest drivers of academic success. Behavioral factors such as Motivation Level and Peer Influence also showed significant secondary effects. Although the study is limited to the features present in the dataset, the methodology provides a strong foundation for future predictive modeling, such as regression or machine learning approaches. Overall, this project demonstrates how data-driven analysis can offer valuable insights into improving educational outcomes.