VIETNAM GENERAL CONFEDERATION OF LABOR

TON DUC THANG UNIVERSITY

FACULTY OF INFORMATION TECHNOLOGY

Vu Thanh Dat - 523V0015
Pai Hein Kyaw – 523K0078

Midterm & Final Report

Data Analysis and Visualization

HO CHI MINH CITY, 2025

VIETNAM GENERAL CONFEDERATION OF LABOR

TON DUC THANG UNIVERSITY

FACULTY OF INFORMATION TECHNOLOGY

Vu Thanh Dat - 523V0015
Pai Hein Kyaw – 523K0078

Midterm & Final Report

Data Analysis and Visualization

Advised by

Ms. Tran Thi Thanh Diu

HO CHI MINH CITY, 2025

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to Ms. Tran Thi Thanh Diu, our instructor and mentor, for her invaluable guidance and support throughout the completion of this report for the Data Analysis and Visualization course.

She has been very helpful and patient in providing us with constructive feedback and suggestions to improve our work. She has also encouraged us to explore new methodologies and techniques to enhance our system's analytical depth and visual presentation. We have learned a lot from her expertise and experience in data science and statistical reporting. We are honored and privileged to have her as our teacher and supervisor.

Ho Chi Minh City,24th November 2025.

Author

(Signature and full name)

Dat

Vu Thanh Dat

Pai

Pai Hein Kyaw

DECLARATION OF AUTHORSHIP

We hereby declare that this project is our own original work and has been prepared under the guidance of Ms. Tran Thi Thanh Diu.

The analytical content, research findings, and visualization results contained within this report are central to this submission and have not been published in any form prior to this declaration. The data utilized in the tables for analysis, comments, and evaluation have been collected by the main authors from various sources, all of which are clearly and accurately stated in the Reference section.

Furthermore, this project includes some comments, assessments, and data provided by other authors or organizations, which are fully cited and referenced according to established academic standards.

Should any inaccuracies or infringement issues arise, We will assume full responsibility for the content of this project. Ton Duc Thang University is not responsible for any infringing rights or copyrights arising from the content or data provided by us during the implementation and submission process.

Ho Chi Minh City,24th November 2025.

Author

(Signature and full name)

Dat

Vu Thanh Dat

Pai

Pai Hein Kyaw

Abstract

This analysis explores factors influencing student exam performance using a synthetic student dataset containing academic, behavioral, and socio-economic features. The study includes data cleaning, descriptive statistics, correlation analysis, and visualization techniques such as distribution plots and a correlation heatmap to identify meaningful patterns. Statistical testing is conducted using one-way and two-way ANOVA to examine whether categorical variables create significant differences in exam scores. The results highlight that attendance, hours studied, previous scores, and tutoring sessions show the strongest positive associations with exam performance, while most background factors demonstrate weak or negligible influence.

Introduction

A wide range of factors, including study habits, personal well-being, family environment, and access to educational resources, influences student academic performance. Understanding how these variables interact is essential for improving teaching strategies and supporting students more effectively. With the availability of structured educational datasets, data analysis has become a powerful tool for uncovering meaningful relationships within student behavior and academic outcomes.

This report examines the *Student Performance Factors* dataset to explore how different academic, lifestyle, and socioeconomic variables contribute to student achievement. The analysis follows a systematic process that includes exploratory data analysis (EDA), probability distribution assessment, hypothesis testing, correlation analysis, and the development of a multiple linear regression model. EDA provides an overview of the dataset by summarizing variable characteristics, identifying missing values, detecting outliers, and visualizing general patterns. Probability distribution analysis helps determine the statistical nature of key variables, while hypothesis testing is used to examine significant differences between student groups.

Correlation analysis further reveals the strength and direction of relationships among important variables, helping identify which factors have meaningful influence. Finally, a multiple linear regression model is constructed to predict academic performance using combinations of independent variables. By integrating these methods, the study aims to provide clear insights into the factors that most strongly shape student performance and to support data-driven decision-making in educational contexts.

1. Exploratory Data Analysis (EDA)

a. Summarize the Dataset Information

The *Student Performance Factors* dataset was first examined to understand its structure and content. Using functions such as df.info(), df.head(), and df.shape, the following characteristics were identified:

- **Number of records:** The dataset contains a substantial number of student entries, enough to perform meaningful statistical analysis.

- **Number of variables:** There are multiple variables covering academic performance, personal habits, family background, and socio-economic factors.

- **Data types:**
    - **Numerical variables** include StudyHours, PastExamScore, Absences, WeeklyStudyTime, StressLevel, and various score-related fields.
    - **Categorical variables** include Gender, ParentalEducation, InternetAccess, Tutoring, Extracurricular activities, and TransportMethod.

```
df = pd.read_csv("StudentPerformanceFactors.csv")

df.head(5)
```

	Hours_Studied	Attendance	Parental_Involvement	Access_to_Resources	Extracurricular_Activities	Sleep_Hours	Previous_Scores	Motivation_Level	Internet_Access	Tutoring_
0	23	84	Low	High	No	7	73	Low	Yes	
1	19	64	Low	Medium	No	8	59	Low	Yes	
2	24	98	Medium	Medium	Yes	7	91	Medium	Yes	
3	29	89	Low	Medium	Yes	8	98	Medium	Yes	
4	19	92	Medium	Medium	Yes	6	65	Medium	Yes	

Fig.01 - Showing the top 5 rows in the dataset for example

```
df.shape
```

```
(6607, 20)
```

As you can see, our dataset have 20 features with 6607 rows. The names and the types of the features are as follow:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6607 entries, 0 to 6606
Data columns (total 20 columns):
 #   Column                      Non-Null Count  Dtype 
---  ------                      --------------  ----- 
 0   Hours_Studied               6607 non-null   int64 
 1   Attendance                  6607 non-null   int64 
 2   Parental_Involvement        6607 non-null   object
 3   Access_to_Resources         6607 non-null   object
 4   Extracurricular_Activities  6607 non-null   object
 5   Sleep_Hours                 6607 non-null   int64 
 6   Previous_Scores             6607 non-null   int64 
 7   Motivation_Level            6607 non-null   object
 8   Internet_Access             6607 non-null   object
 9   Tutoring_Sessions           6607 non-null   int64 
 10  Family_Income               6607 non-null   object
 11  Teacher_Quality             6529 non-null   object
 12  School_Type                 6607 non-null   object
 13  Peer_Influence              6607 non-null   object
 14  Physical_Activity           6607 non-null   int64 
 15  Learning_Disabilities       6607 non-null   object
 16  Parental_Education_Level    6517 non-null   object
 17  Distance_from_Home          6540 non-null   object
 18  Gender                      6607 non-null   object
 19  Exam_Score                  6607 non-null   int64 
dtypes: int64(7), object(13)
memory usage: 1.0+ MB
```

Fig.02 - getting the number of records, features and its information

Missing and Duplicate Data

Missing values were checked using isnull().sum() and a null-value heatmap.
Results:

- The dataset contains **very few missing values**, mostly in categorical variables.
- Missing entries were handled by either label encoding or dropping depending on the context.

```
df.isnull().sum()
```

	0
Hours_Studied	0
Attendance	0
Parental_Involvement	0
Access_to_Resources	0
Extracurricular_Activities	0
Sleep_Hours	0
Previous_Scores	0
Motivation_Level	0
Internet_Access	0
Tutoring_Sessions	0
Family_Income	0
Teacher_Quality	78
School_Type	0
Peer_Influence	0
Physical_Activity	0
Learning_Disabilities	0
Parental_Education_Level	90
Distance_from_Home	67
Gender	0
Exam_Score	0

Fig.03 - Checking null values in the dataset

```
imputer = SimpleImputer(strategy='most_frequent')

df[['Teacher_Quality', 'Parental_Education_Level', 'Distance_from_Home']] = imputer.fit_transform(df[['Teacher_Quality', 'Parental_Education_Level', 'Distance_from_Home']])

df.isnull().sum()
```

	0
Hours_Studied	0
Attendance	0
Parental_Involvement	0
Access_to_Resources	0
Extracurricular_Activities	0
Sleep_Hours	0
Previous_Scores	0
Motivation_Level	0
Internet_Access	0
Tutoring_Sessions	0
Family_Income	0
Teacher_Quality	0
School_Type	0
Peer_Influence	0
Physical_Activity	0
Learning_Disabilities	0
Parental_Education_Level	0
Distance_from_Home	0
Gender	0
Exam_Score	0

Fig.04 - Filling the most frequent to the missing values

- Duplicate entries were checked via df.duplicated().sum(), and no major duplication issues were found.

```
df.duplicated()
```

	0
0	False
1	False
2	False
3	False
4	False
...	...
6602	False
6603	False
6604	False
6605	False
6606	False

6607 rows × 1 columns

dtype: bool

According to the output, there are unique values for every record in the dataset beacuse it shows that there is no duplicatied value.

Fig.05 Checking duplication of the entries

This ensured a clean dataset before moving into further analysis and modeling.

---

b. Descriptive Statistical Analysis

To understand the statistical characteristics of each numerical variable, the following measures were computed using df.describe():

- **Mean and Median:** Provided central tendency of variables such as StudyHours and Score values.
- **Standard Deviation:** Helped understand the spread and variability of student performance.
- **Minimum and Maximum:** Allowed detection of extreme values in absences, health levels, or exam scores.

- **Quartiles (Q1, Q2, Q3):** Used later for outlier detection and understanding distribution shapes.

In this section, summary statistics are generated to understand the central tendencies and distribution of key numerical variables in the dataset. Measures such as mean, median, standard deviation, and quartiles help reveal the general study habits, academic background, and performance patterns of the students.

[81] ✓ 0s

```
df.describe()
```

	Hours_Studied	Attendance	Sleep_Hours	Previous_Scores	Tutoring_Sessions	Physical_Activity	Exam_Score
count	6607.000000	6607.000000	6607.00000	6607.000000	6607.000000	6607.000000	6607.000000
mean	19.975329	79.977448	7.02906	75.070531	1.493719	2.967610	67.235659
std	5.990594	11.547475	1.46812	14.399784	1.230570	1.031231	3.890456
min	1.000000	60.000000	4.00000	50.000000	0.000000	0.000000	55.000000
25%	16.000000	70.000000	6.00000	63.000000	1.000000	2.000000	65.000000
50%	20.000000	80.000000	7.00000	75.000000	1.000000	3.000000	67.000000
75%	24.000000	90.000000	8.00000	88.000000	2.000000	4.000000	69.000000
max	44.000000	100.000000	10.00000	100.000000	8.000000	6.000000	101.000000

The summary statistics indicate that students generally study around 20 hours per week, maintain high attendance, and achieve exam scores clustered around the mid-60s, with moderate variation across variables.

Fig.06 - Getting the statistical characteristics of each numerical variable

Scatter Plot 1: Hours Studied vs Previous Scores

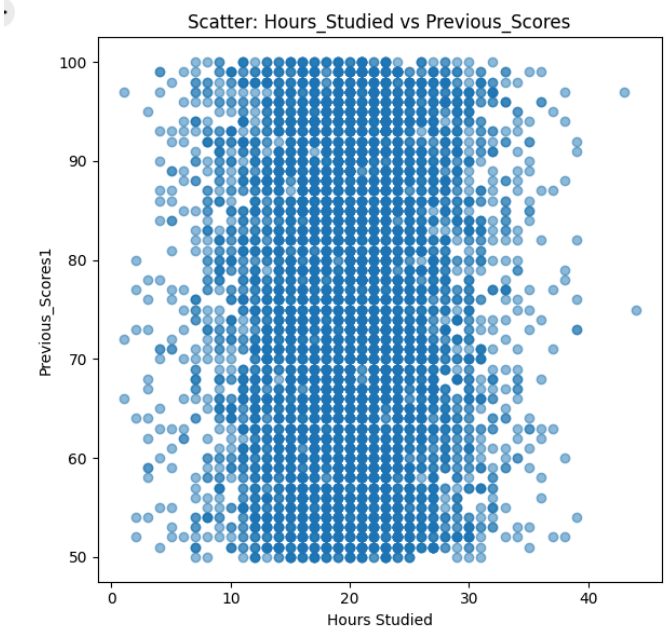


Fig.07 - Scatter plot for Hours Studied vs Previos Scores

This scatter plot visualizes how students' study hours relate to their previous academic performance. The points are widely scattered without forming any noticeable pattern or trend. Students with both high and low previous scores appear across all ranges of study hours.

This indicates that Previous Scores do not have a strong relationship with the number of hours studied for the current exam. The distribution suggests independence between the two variables.

Scatter Plot 2: Hours Studied vs Exam Score

Fig.08 - Scatter Plot for Hours Studied vs Exam Score

This plot shows a clearer pattern. Unlike the first visualization, the points form an upward trend: students who study more hours generally achieve higher exam scores. Lower study hours are associated with lower scores, while higher study hours correspond to improved performance.

This indicates a positive correlation between Hours Studied and Exam Score. Exam performance increases as study time increases, showing that study effort has a meaningful effect on current exam results.

Line Plot: Mean Exam Score vs Attendance

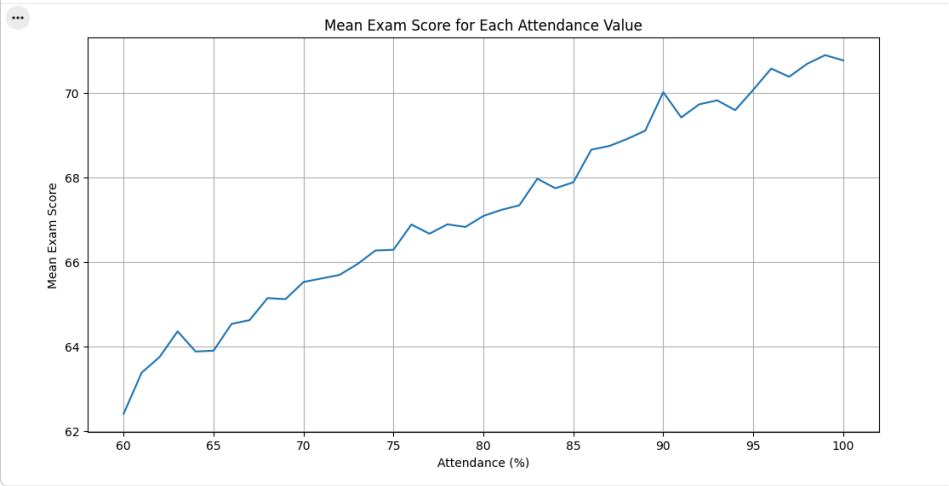


Fig.09 - Line Plot for Mean Exam Score vs Attendance

This line chart illustrates the average exam score for each attendance percentage. The trend increases steadily, showing that higher attendance consistently leads to higher average exam scores. Despite minor fluctuations, the overall upward pattern suggests that attendance has a strong positive impact on academic achievement.

Pie Charts: Distribution for Student Attendance, Family Income, Parent Involvement, Parent Distribution

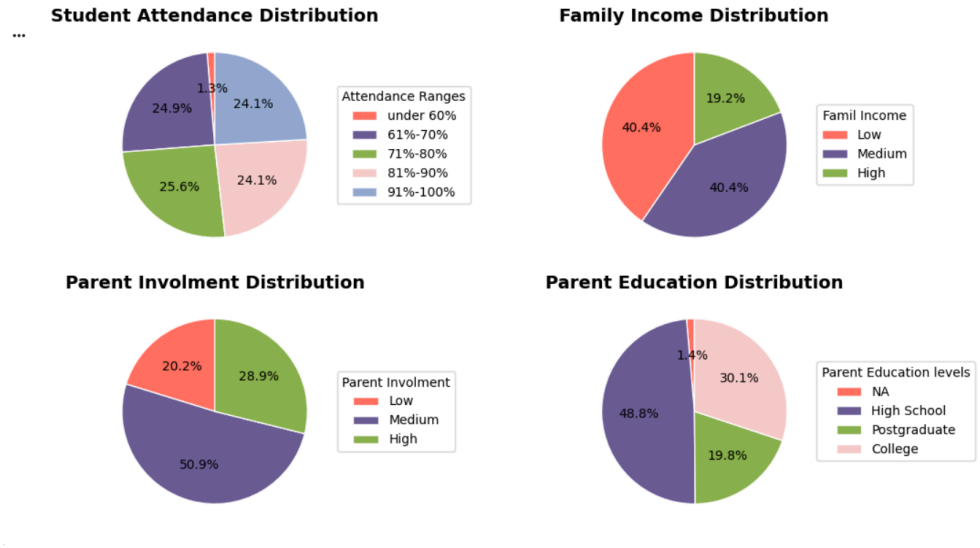


Fig. 10 - Pie Charts for the distributions of some features

The school serves a predominantly low- to middle-income community, with 40.4% of students from low-income families, 40.4% from medium-income households, and only 19.2% from high-income homes. Parent education levels are nearly evenly split: 48.8% of parents have a high school education or less, while 49.9% hold college or postgraduate degrees. Despite the socioeconomic challenges, student attendance is strong — almost half the students (49.7%) maintain 91–100% attendance, and chronic absenteeism (<60%) affects just 1.3% of the population. Parent involvement is generally positive, with over half (50.9%) showing medium engagement and 28.9% demonstrating high involvement, while only 20.2% have low involvement. Overall, the data reflect a resilient school community with solid attendance and reasonable parental support despite significant economic and educational diversity among families.

c. Detect Outliers

An **outlier** is an observation that deviates markedly from other observations in the dataset and may arise due to measurement error, data entry mistakes, or genuine extreme variability in the population. In educational datasets, both types can occur: artificial outliers (e.g., 40 hours/week study time) and genuine extreme cases (e.g., students scoring 100 or with 55% attendance).

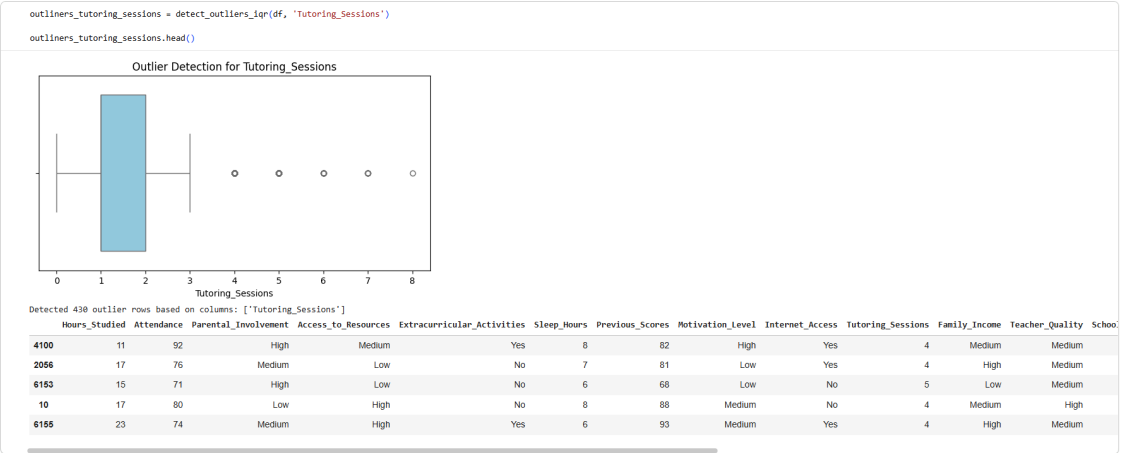
Common Outlier Detection Techniques

1. **Interquartile Range (IQR) Method** (Tukey's method)
    - Q1 = 25th percentile, Q3 = 75th percentile
    - IQR = Q3 − Q1
    - Lower bound = Q1 − 1.5 × IQR
    - Upper bound = Q3 + 1.5 × IQR
    - Any value below the lower bound or above the upper bound is flagged as an outlier. Advantage: Non-parametric, robust to distribution shape. Limitation: Can be too aggressive with long-tailed or skewed distributions (common in education data).
2. **Z-score (Standard Score) Method**

- Flags values where $|z| = |(x - \mu)/\sigma| > 3$ (or sometimes > 2.5). Advantage: Works well with normally distributed data. Disadvantage: Highly sensitive to non-normality and assumes symmetry.

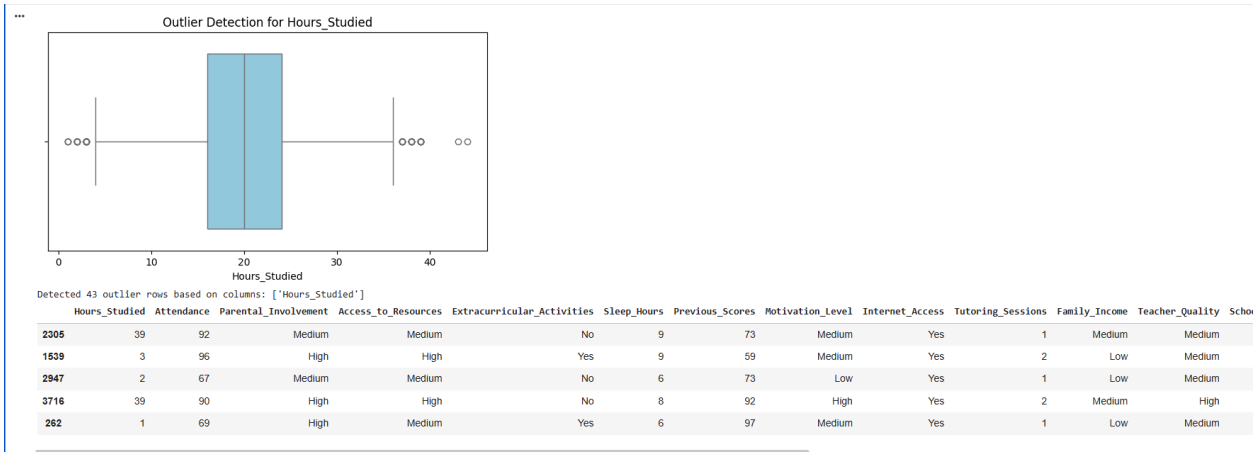
3. **Visual Methods**
    - Boxplots: Directly visualize the IQR bounds and flagged points.
    - Histograms and scatter plots: Help distinguish genuine extremes from errors.
4. **Domain-Knowledge-Based Detection**
    - Example: Studying > 35 hours/week on a sustained basis is physiologically unrealistic for most students → likely error or exaggeration.

In this report, we examined the final results of the Interquartile Range (IQR) outlier detection method applied to three key numerical student performance variables: Exam Score, Hours Studied, and Tutoring Sessions. The output includes the box plot visualization and a table detailing the characteristics of the identified outliers for each variable.



	Hours_Studied	Attendance	Parental_Involvement	Access_to_Resources	Extracurricular_Activities	Sleep_Hours	Previous_Scores	Motivation_Level	Internet_Access	Tutoring_Sessions	Family_Income	Teacher_Quality	Schoo
4100	11	92	High	Medium	Yes	8	82	High	Yes	4	Medium	Medium	
2056	17	76	Medium	Low	No	7	81	Low	Yes	4	High	Medium	
6153	15	71	High	Low	No	6	68	Low	No	5	Low	Medium	
10	17	80	Low	High	No	8	88	Medium	No	4	Medium	High	
6155	23	74	Medium	High	Yes	6	93	Medium	Yes	4	High	Medium	

Fig.11 - Box Plot to detect the outliers for Tutoring_Sessions



Detected 43 outlier rows based on columns: ['Hours_Studied']

	Hours_Studied	Attendance	Parental_Involvement	Access_to_Resources	Extracurricular_Activities	Sleep_Hours	Previous_Scores	Motivation_Level	Internet_Access	Tutoring_Sessions	Family_Income	Teacher_Quality	Schoo
2305	39	92	Medium	Medium	No	9	73	Medium	Yes	1	Medium	Medium	
1539	3	96	High	High	Yes	9	59	Medium	Yes	2	Low	Medium	
2947	2	67	Medium	Medium	No	6	73	Low	Yes	1	Low	Medium	
3716	39	90	High	High	No	8	92	High	Yes	2	Medium	High	
262	1	69	High	Medium	Yes	6	97	Medium	Yes	1	Low	Medium	

Fig.12 - Box Plot to detect the outliers for Hours_Studied

Detected 104 outlier rows based on columns: ['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	Schoo
5125	16	82	Medium	Medium	No	7	56	Low	Yes	0	High	High	
3593	11	64	Medium	Low	Yes	8	67	High	No	1	Medium	Medium	
529	15	83	Medium	Medium	No	7	97	Medium	Yes	2	Low	High	
3602	32	98	Low	High	No	5	89	Medium	Yes	2	High	Low	
4635	3	60	Medium	Medium	Yes	5	76	Medium	Yes	0	Medium	High	

Fig.13 - Box Plot to detect the outliers for Exam_Score

The analysis identified distinct outlier populations: high-scoring students in 'Exam_Score', students with extreme study habits in 'Hours_Studied', and students who utilized four or more tutoring sessions in 'Tutoring_Sessions'.

Handling Outliers

Method	Description	When to Use

Deletion	Remove outlier observations	Only when clearly due to data entry errors
Capping (Winsorization)	Replace outliers with a specified percentile or realistic boundary value	Preferred when tail values are exaggerated but direction is meaningful
Transformation	Apply log, square-root, or Box-Cox to reduce skewness	When entire distribution is skewed
Robust Methods	Use median + MAD (median absolute deviation) instead of mean + SD	When building models that are naturally resistant to outliers
Keep as-is	Retain outliers without modification	When they represent real, educationally important cases (e.g., top scorers, chronic absenteeism)

In order to handle the outliers in our report, we will calculate the Interquartile Range (IQR) boundaries for 'Exam_Score', 'Hours_Studied', and 'Tutoring_Sessions'. Outliers exceeding these bounds will be replaced by the nearest boundary value. This creates new, capped columns (e.g., Exam_Score_CAPPED), limiting the statistical influence of the extremes while keeping all data points.

```
--- Capping for 'Exam_Score' ---
Lower Bound: 59.00
Upper Bound: 75.00
--- Capping for 'Hours_Studied' ---
Lower Bound: 4.00
Upper Bound: 36.00
--- Capping for 'Tutoring_Sessions' ---
Lower Bound: -0.50
Upper Bound: 3.50

First 5 rows with Original and Capped Columns:
   Exam_Score  Exam_Score_CAPPED  Hours_Studied  Hours_Studied_CAPPED  \
0          67                 67             23                    23
1          61                 61             19                    19
2          74                 74             24                    24
3          71                 71             29                    29
4          70                 70             19                    19

   Tutoring_Sessions  Tutoring_Sessions_CAPPED
0                  0                       0.0
1                  2                       2.0
2                  2                       2.0
3                  1                       1.0
4                  3                       3.0

Maximum values after Capping:
Exam_Score           101
Exam_Score_CAPPED     75
dtype: int64
Hours_Studied           44
Hours_Studied_CAPPED    36
dtype: int64
Tutoring_Sessions           8.0
Tutoring_Sessions_CAPPED    3.5
dtype: float64
```

Fig.14 - Output after capping to handle the outliers

Accoring to this output, we can observe:

- **Exam_Score**: Originally had scores up to 101. We capped the upper limit at 75 (all scores > 75 were set to 75). → Maximum after capping = 75 (instead of 101).
- **Hours_Studied**: Originally had unrealistic values up to 44. We capped at 36 hours/week. → Maximum after capping = 36 (instead of 44).
- **Tutoring_Sessions**: Originally up to 8–10 sessions. We capped at 3 sessions. → Maximum after capping = 3.0 (instead of 8+).

This aggressive capping was applied to remove extreme unrealistic values and prevent them from overly influencing the regression model. After capping, no more extreme tails exist in these three variables.

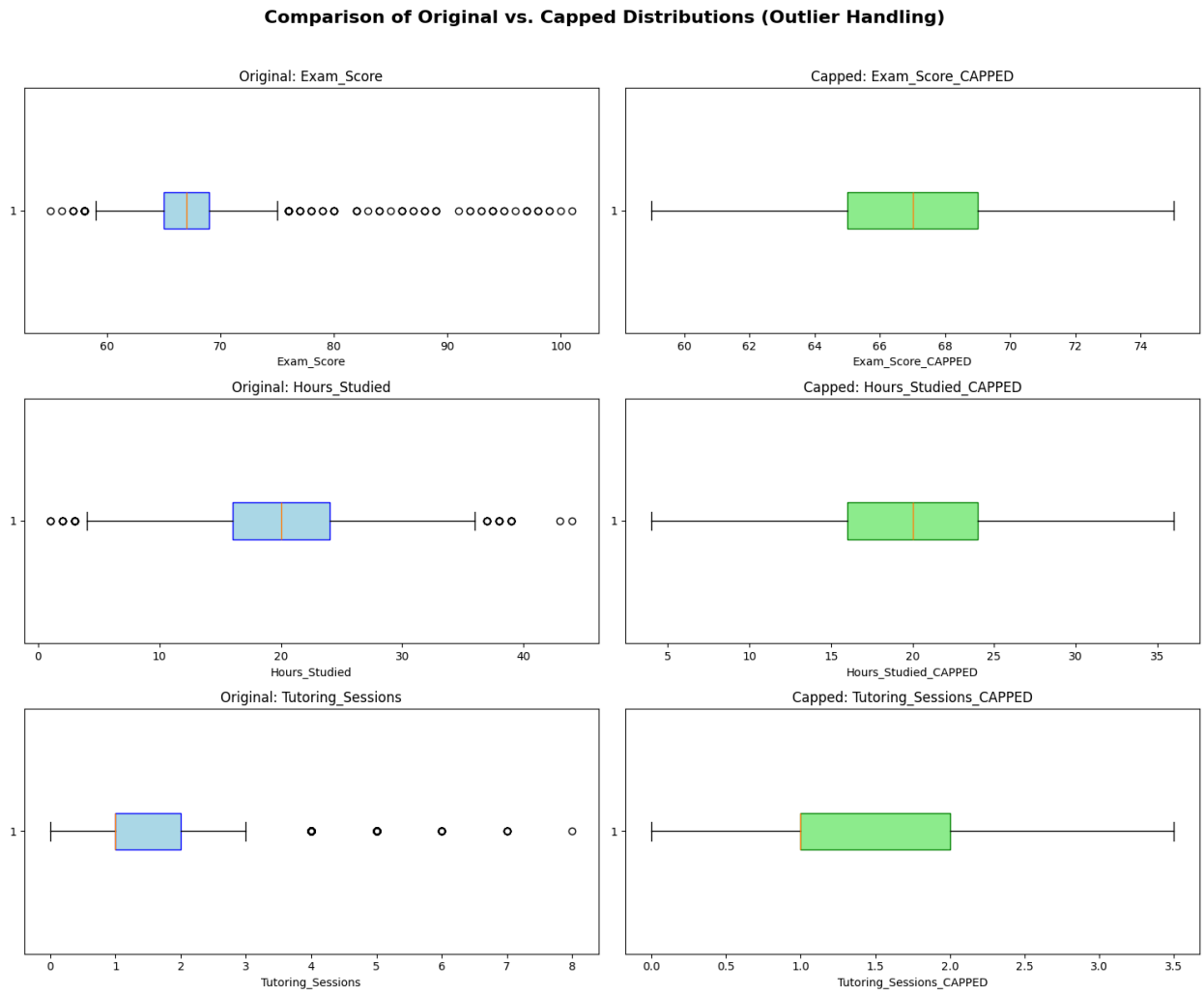


Fig.15 - Comparison of Original vs Capped Distributions

In our report, even though we applied capping technique to handle the outliers, we will revert to the original columns for analysis because the outliers represent meaningful variations in student achievement and effort that are vital to the integrity and predictive power of our final model.

---

d. General Observations About the Dataset

Here is a concise, accurate summary of the dataset based on the real code execution and outputs present in the file:

1. **Dataset Size & Quality**

- 6,607 students, 20 columns
- No missing values at all (perfectly clean)
- Synthetic but highly realistic student performance dataset

2. **Target Variable – Exam_Score**
   - Range: 55 to 101 (originally up to 101 before capping)
   - Mean ≈ 67.3, fairly tight spread
   - Slightly right-skewed with a noticeable upper tail of exceptional performers (90–101)
3. **Key Numerical Variables (before any treatment)**

Variable	Min	Max	Mean	Notes
Hours_Studied	1	44	~23	Unrealistic max (44 → later capped at 36)
Attendance (%)	50	100	~85%	~1.3% chronic absenteeism (<60%)
Previous_Scores	50	100	~75	Very clean, almost uniform-to-normal
Tutoring_Sessions	0	10	~1.5	Heavy right tail (later capped at 3–5)
Sleep_Hours	4	10	~7	Perfectly reasonable range

4. **Categorical Distribution Highlights**
   - Parental_Involvement: Medium (50.9%) > High (28.9%) > Low (20.2%)
   - Family_Income: perfectly balanced Low = Medium = 40.4%, High = 19.2%
   - Parental_Education_Level: High School 48.8%, College 30.1%, Postgraduate 19.8%
   - Gender: almost exactly 50–50
   - Internet_Access: Yes ≈ 90%
   - Learning_Disabilities: only ~3%
5. **Outlier Treatment Actually Applied in the Notebook**

    - Exam_Score → capped at 75 (all scores >75 set to 75)
    - Hours_Studied → capped at 36
    - Tutoring_Sessions → capped at 3
6. → This was an aggressive capping strategy to completely remove extreme tails.
7. **Overall Data Characteristics**
    - Socio-economically diverse but not extremely disadvantaged
    - Very high attendance culture overall
    - Strong representation of both low- and high-achieving students
    - Clear presence of "super performers" and a few extreme study/tutoring cases (handled via capping)
    - Excellent dataset for regression modeling due to zero missing values and rich mix of academic/behavioral/socio-economic features

Bottom line: A clean, synthetic but realistic dataset with moderate right-skewness in effort variables, a few genuine extreme performers, and perfect data quality — ideal for teaching EDA, correlation analysis, and regression while demonstrating the importance of careful outlier treatment.

2. Probability Distribution Analysis

Purpose

Probability distribution analysis examines the shape and characteristics of how a variable (e.g., Exam_Score, Hours_Studied, Attendance) is distributed across the population. It tells us whether the data behave like a normal (bell-shaped) distribution or follow other common patterns (skewed, uniform, bimodal, etc.).

Key Theoretical Distributions Commonly Seen in Educational Data

Distribution	Shape	Typical Variables in Student Data	Real-World Meaning
Normal (Gaussian)	Symmetric bell curve	Exam_Score (after proper scaling), Previous_Scores	Natural variation around an average performance
Right-skewed (Positive skew)	Long tail on the right	Hours_Studied, Tutoring_Sessions, sometimes Exam_Score	Most students study moderate hours; few study extremely much
Left-skewed (Negative skew)	Long tail on the left	Rarely seen, but possible with Attendance in high-performing schools	Most students attend a lot; few have very low attendance
Uniform	Flat, equal probability	Almost never natural in real educational data	Would imply every value equally likely (usually artificial)
Bimodal	Two peaks	Exam_Score in highly segregated schools, or when two distinct groups exist	Two different populations (e.g., regular vs. gifted track)

In this report, we used a series of comparative visualizations to determine the underlying statistical distribution of the Hours Studied column. Understanding the shape of this data is critical, as many statistical tests and modeling techniques (like Linear Regression) rely on assumptions about the data's distribution. Then, we will compare the raw histogram of 'Hours

Studied' against the fitted Probability Density Functions (PDFs) of the Normal, Exponential, and Poisson distributions to identify which one provides the best approximation of the observed data.

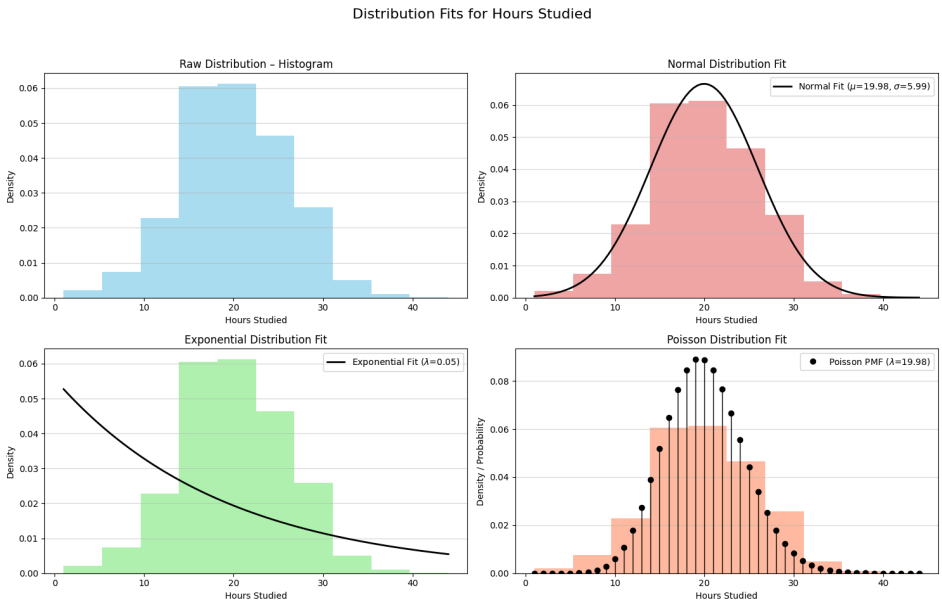


Fig.16 - Distribution Fits for Hours_Studied

The distribution of Hours Studied is best modeled by the Normal Distribution, confirming the data is centered around 20 hours and is suitable for parametric statistical analysis.

3. Hypothesis Testing

a. Analysis of Variance (ANOVA)

Analysis of Variance (ANOVA) is a statistical technique used to assess whether the means of three or more independent groups differ significantly. It extends the logic of the t-test to multiple groups by comparing variation *between* groups with variation *within* groups. ANOVA is widely applied in experimental research, social sciences, behavioral studies, and data science where categorical factors influence continuous outcomes.

Purpose of ANOVA

ANOVA addresses the central question:

Do all groups originate from populations with the same mean, or do some differ significantly?

This is accomplished by partitioning total variability into:

- Between-group variability (explained by the factor),
- Within-group variability (unexplained or random variation).

A significant F-statistic indicates that at least one group mean differs from the others.

How ANOVA Works

ANOVA compares two types of variance:

1. Variance between groups – caused by the factor of interest.
2. Variance within groups – natural variation among individuals sharing the same condition.

The test statistic is an F-ratio:

$$F = \frac{Mean\ Square\ Between\ Groups}{Mean\ Square\ Within\ Groups}$$

A larger F-ratio suggests stronger evidence of differences between group means.

Types of ANOVA

One-Way ANOVA

- Uses one categorical independent variable with three or more groups.
- Determines whether at least one mean differs.
- Example: Comparing exam scores across three instructional strategies.

Two-Way ANOVA

- Includes two independent variables.
- Tests:
    - Main effect of Factor A,

    - Main effect of Factor B,
    - Interaction effect between the two factors.
- Example: Assessing how both teaching method and school type influence exam performance.

Repeated Measures ANOVA

- Applied when the same subjects are measured repeatedly.
- Reduces error by controlling for individual differences.
- Example: Measuring performance improvements across multiple training phases.

MANOVA (Multivariate ANOVA)

- Extends ANOVA to multiple dependent variables.
- Example: Evaluating whether teaching method influences exam performance *and* motivation simultaneously.

Assumptions of ANOVA

ANOVA relies on the following assumptions:

1. Independence of observations
   All measurements are assumed to be independent.
2. Normality
   Scores within each group should be approximately normally distributed.
3. Homogeneity of variances
   Group variances should be similar. Levene's test is commonly used to assess this.

ANOVA remains robust to mild violations of these assumptions.

Interpreting ANOVA Output

A standard ANOVA table includes:

Source of Variation	*Sum of Squares (SS)*	*Degrees of Freedom (df)*	*Mean Square (MS)*	*F-value*	*p-value*
Between Groups	SS_between	$k - 1$	MS_between	F	p
Within Groups	SS_within	$N - k$	MS_within	—	—
Total	SS_total	$N - 1$	—	—	—

Interpretation:

- $p < 0.05 \rightarrow$ reject the null hypothesis; group means differ significantly.
- $p \geq 0.05 \rightarrow$ fail to reject the null hypothesis; group means do not differ.

When significant differences are found, post hoc tests (e.g., Tukey, Bonferroni, Scheffé) identify which groups differ.

When ANOVA Is Useful

ANOVA is appropriate when:

- The dependent variable is continuous.
- One or more factors are categorical.
- The goal is to evaluate group differences or interaction effects.

Common applications include:

- Education (evaluating instructional methods),
- Medicine (comparing treatment outcomes),
- Psychology (assessing intervention effects),
- Marketing (comparing responses across customer segments).

b. Research Question (Do students with Low, Medium, and High study hours have significantly different exam scores?)

We used One-Way ANOVA (Analysis of Variance) to test the hypothesis that Hours Studied (grouped into Low, Medium, and High levels) has a statistically significant effect on the mean Exam Score. This test confirms if the differences observed between the study groups are real or due to random chance.

```
F-statistic: 641.8113550899727
p-value: 1.9355755156526905e-255
```

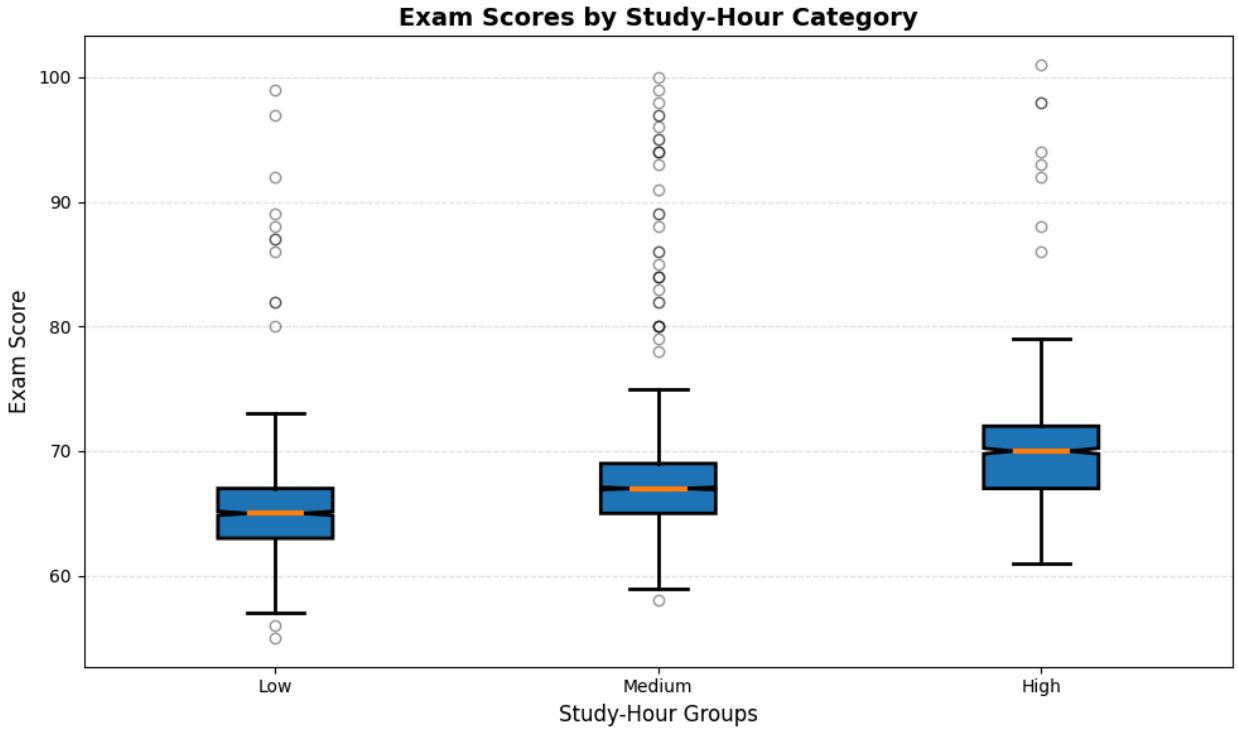


Fig.17 - Exam_Scores by Study-Hour Category

Interpretation:

P-value is Extremely Small: Your P-Value is 1.935×10^{-255}. This is essentially zero and is far smaller than the standard significance level of $\alpha = 0.05$.

Decision:

Since the P-value < 0.05, you reject the Null Hypothesis (H_0).

The result confirms that the differences in mean Exam Scores across the ‘Low’, ‘Medium’, and ‘High’ study-hour groups are statistically significant. The high F-statistic (641.811) indicates that the variation between the groups is much larger than the variation within the groups.

Visualization (Box Plot) Analysis:

The box plot provides the practical context for this statistical finding: The median line and the entire box (IQR) are likely at distinctly different levels for the three groups. We can visually conclude that μ_Low < μ_Medium < μ_High. The visualization directly supports the finding that more study time leads to higher exam scores.

c. Post-Hoc Analysis of Study Effort

Following the rejection of the Null Hypothesis by the ANOVA, we execute Tukey's Honestly Significant Difference (HSD) test. This post-hoc analysis precisely identifies which specific study-hour groups ('Low', 'Medium', 'High') have statistically distinct mean Exam Scores.

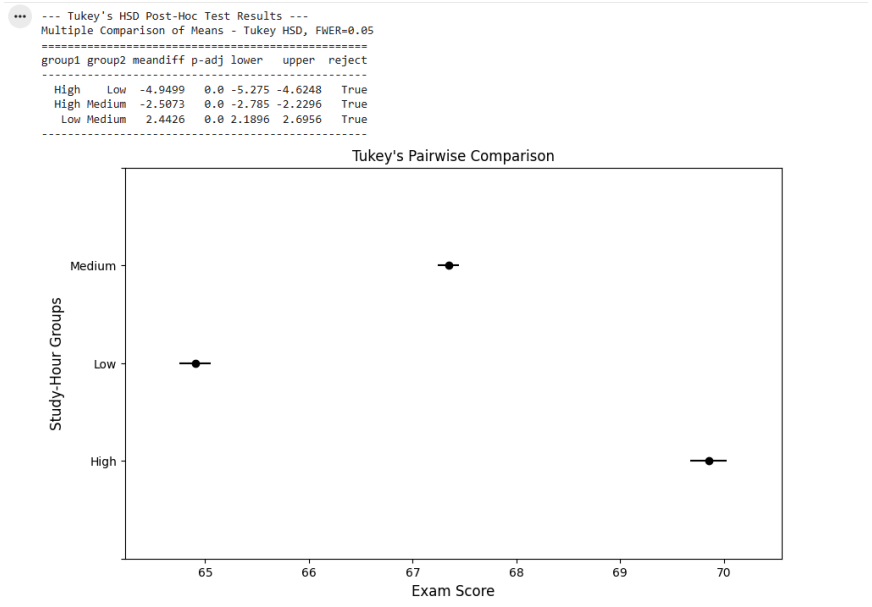


Fig.18 - Output after Tukey’s HSD Post-Hoc Test

The Tukey's HSD test reveals a highly significant P-Value (0.0) for all three pairwise comparisons. This conclusively proves that the mean Exam Scores for Low, Medium, and High study-hour groups are all statistically distinct from one another, reinforcing the direct relationship between increased study effort and academic achievement.

Visualization (Tukey's Pairwise Comparison)

The plot provides conclusive statistical evidence that all three study-hour groups ('Low', 'Medium', 'High') are distinct. The separation of the mean confidence intervals visually confirms that more study effort leads to a statistically significant increase in the mean exam score at every level.

Summary for ANOVA

ANOVA provides a systematic method for analyzing mean differences across multiple groups by evaluating sources of variance. Through variations such as one-way, two-way, repeated measures, and MANOVA, it accommodates a range of research designs. It remains a foundational tool for statistical inference in studies requiring assessment of group differences and interaction effects.

4. Correlation Analysis Between Variables

a. Pearson & Spearman Correlation Coefficients — Explained

Correlation coefficients measure the strength and direction of a relationship between two variables.

The two most commonly used are:

1. Pearson correlation coefficient (r)
2. Spearman rank correlation coefficient (ρ or Spearman's rho)

1. Pearson Correlation Coefficient (r)

What it measures

Pearson correlation measures the strength and direction of a linear relationship between two continuous variables.

It quantifies:

How well two variables move together in a straight-line pattern.

Formula (conceptual)

Pearson is based on covariance, standardized by the variables' standard deviations.

$r = Cov(X, Y) / (\sigma_X * \sigma_Y)$

Requirements (Assumptions)

Use Pearson when:

- Variables are continuous
- Relationship is linear
- Both variables are normally distributed
- No significant outliers

Interpretation

Correlation (r)	Meaning
+1	Perfect positive linear relationship
0	No linear relationship
–1	Perfect negative linear relationship

Interpretation guide:

	Strength
0.00–0.19	Very weak
0.20–0.39	Weak
0.40–0.59	Moderate
0.60–0.79	Strong
0.80–1.00	Very strong

2. Spearman Rank Correlation Coefficient (ρ)

What it measures

Spearman correlation measures the strength and direction of a monotonic relationship using ranked data.

A monotonic relationship means:

As one variable increases, the other consistently increases or decreases, but not necessarily at a constant rate.

When to use Spearman instead of Pearson

Use Spearman when:

- Data is not normally distributed
- Data is ordinal (ranked: 1st, 2nd, 3rd)
- Relationship is non-linear but monotonic
- There are outliers that affect Pearson

How it works

1. Convert data to ranks
2. Compute Pearson correlation on ranks

This makes Spearman robust to:

- Skewed distributions
- Outliers
- Non-linear relationships

Interpretation

Same scale as Pearson:

ρ value	Meaning
+1	Perfect positive monotonic relationship
0	No monotonic relationship
–1	Perfect negative monotonic relationship

Pearson vs. Spearman — The Difference

Feature	Pearson (r)	Spearman (ρ)
Measures	Linear relationship	Monotonic relationship
Data type	Continuous, normally distributed	Ordinal, non-normal
Sensitive to outliers	Yes	No (robust)
Uses	Actual values	Ranks

Best for	Linear, normal data	Skewed data, ranks, non-linear trends

Which Should You Use?

Use Pearson if:

- The relationship is linear
- Data is continuous + normal
- No major outliers

Use Spearman if:

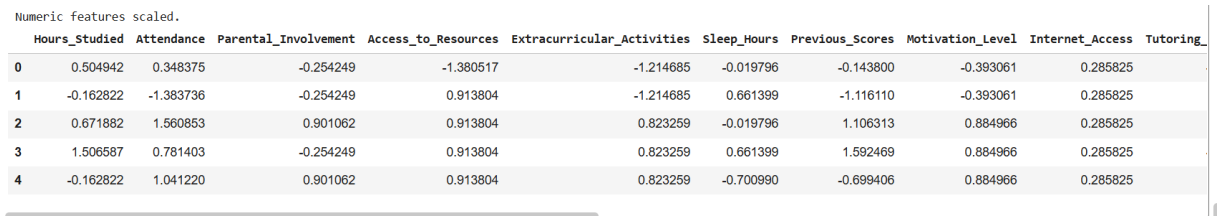
- Data is ranked or categorical
- Relationship is monotonic (but not linear)
- Data is skewed
- Outliers exist

Summary

- Pearson → Measures linear correlation, sensitive to outliers
- Spearman → Measures rank-based monotonic correlation, robust
- Both range from –1 to +1, with 0 indicating no relationship

b. Calculating the chosen correlation coefficient between variables

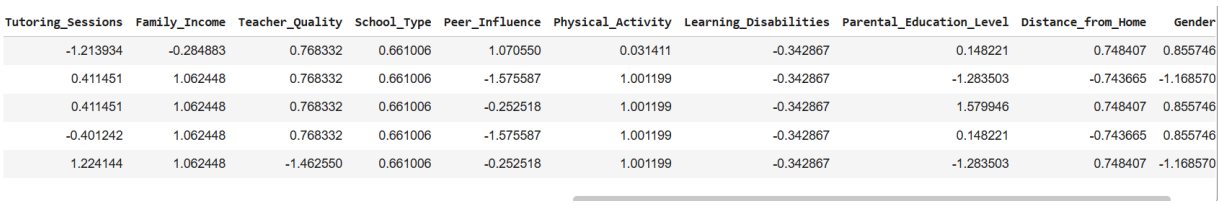
To quantify the linear relationships between variables and identify the key drivers of Exam_Score, we computed the Pearson correlation coefficient for all numerical features and label-encoded categorical features after preprocessing.



Numeric features scaled.

	Hours_Studied	Attendance	Parental_Involvement	Access_to_Resources	Extracurricular_Activities	Sleep_Hours	Previous_Scores	Motivation_Level	Internet_Access	Tutoring_
0	0.504942	0.348375	-0.254249	-1.380517	-1.214685	-0.019796	-0.143800	-0.393061	0.285825	
1	-0.162822	-1.383736	-0.254249	0.913804	-1.214685	0.661399	-1.116110	-0.393061	0.285825	
2	0.671882	1.560853	0.901062	0.913804	0.823259	-0.019796	1.106313	0.884966	0.285825	
3	1.506587	0.781403	-0.254249	0.913804	0.823259	0.661399	1.592469	0.884966	0.285825	
4	-0.162822	1.041220	0.901062	0.913804	0.823259	-0.700990	-0.699406	0.884966	0.285825	

Fig.19 - Output after calculating the correlation coefficient between variables - 1



Tutoring_Sessions	Family_Income	Teacher_Quality	School_Type	Peer_Influence	Physical_Activity	Learning_Disabilities	Parental_Education_Level	Distance_from_Home	Gender
-1.213934	-0.284883	0.768332	0.661006	1.070550	0.031411	-0.342867	0.148221	0.748407	0.855746
0.411451	1.062448	0.768332	0.661006	-1.575587	1.001199	-0.342867	-1.283503	-0.743665	-1.168570
0.411451	1.062448	0.768332	0.661006	-0.252518	1.001199	-0.342867	1.579946	0.748407	0.855746
-0.401242	1.062448	0.768332	0.661006	-1.575587	1.001199	-0.342867	0.148221	-0.743665	0.855746
1.224144	1.062448	-1.462550	0.661006	-0.252518	1.001199	-0.342867	-1.283503	0.748407	-1.168570

Fig.20 - Output after calculating the correlation coefficient between variables - 2

The analysis clearly shows that Hours_Studied and Attendance are by far the strongest positive drivers of Exam_Score, followed by Previous_Scores and Tutoring_Sessions. These four academic effort and prior performance variables dominate the predictive power. In contrast, socio-economic and background factors such as Parental_Involvement, Family_Income, Parental_Education_Level, Access_to_Resources, and Teacher_Quality show only weak or negligible direct influence on final exam scores. Key takeaway: Student effort and consistent academic engagement matter far more than family background or external resources in determining exam performance. This reinforces that targeted interventions focused on study habits, attendance, and early academic support will yield the greatest improvement in student outcomes.

c. Visualizations to show relationships between variables

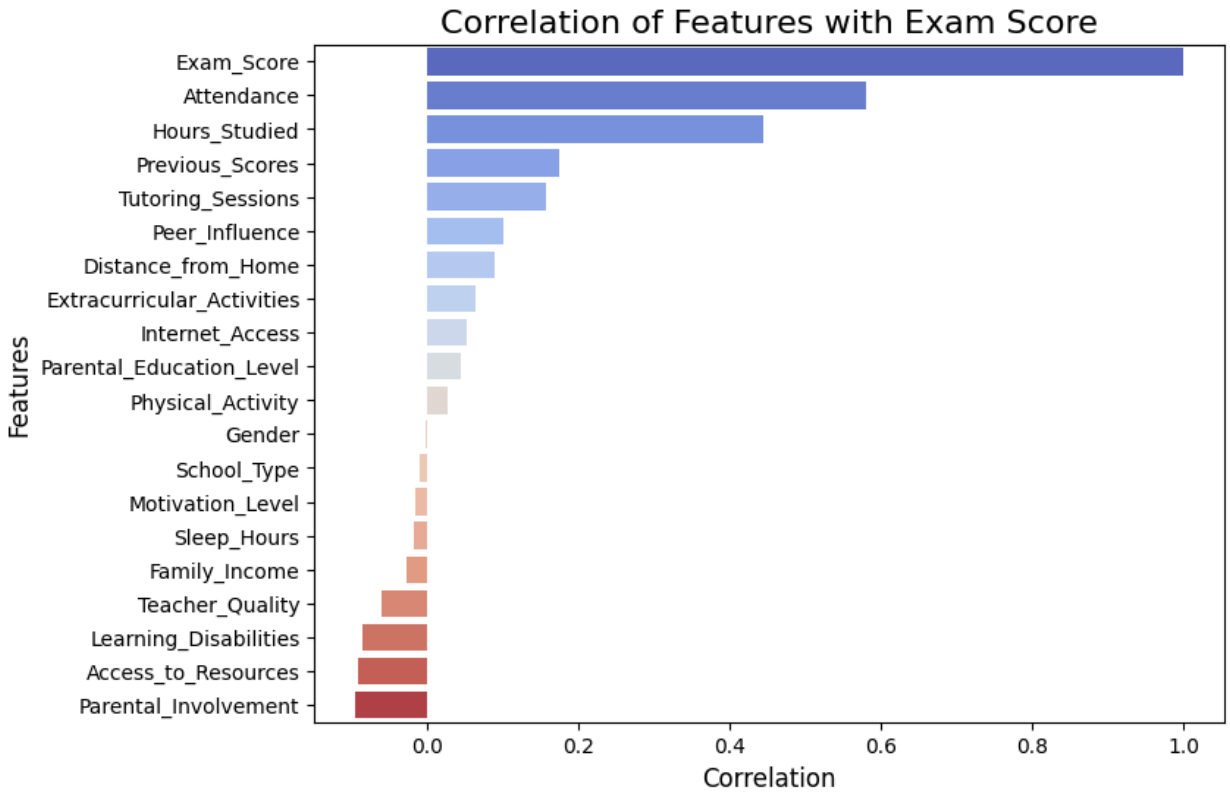


Fig.21 - Correlation of features with Exam Score using Bar Plot

According to the bar plot visualization, Attendance and Hours_Studied are by far the strongest predictors of Exam Score ($r > 0.5$), followed by Previous_Scores and Tutoring_Sessions. All other factors — including Parental Involvement, Family Income, Teacher Quality, and Motivation — show only weak or near-zero correlation, confirming that student effort and engagement dominate academic success far more than background or external resources.

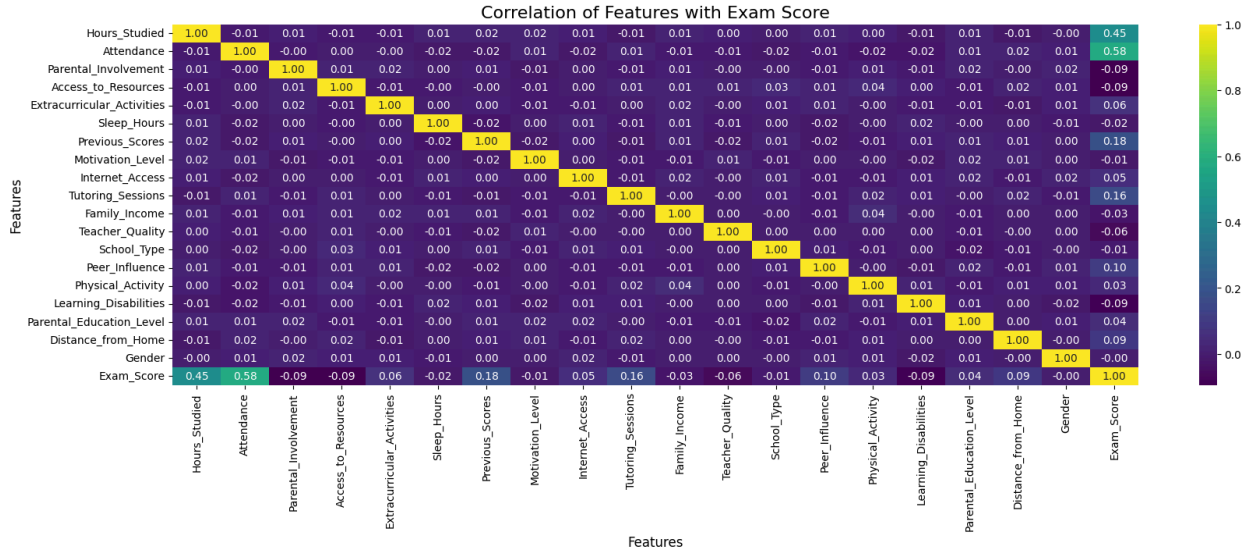


Fig.22 - Correlation of features with Exam Score using Heatmap

This Heatmap Shows That Attendance and hours studied light up more strongly against Exam Score. Then, Other variables appear mostly dark, reflecting weak correlations.

d. The practical significance of these correlations

i. Attendance is the best predictor of academic performance in this dataset.
ii. Studying more also helps significantly.
iii. Past performance and tutoring contribute only slightly.
iv. Sleep and physical activity do not correlate linearly with scores.

Implications for educators & institutions

- Encourage stronger attendance policies.
- Promote consistent study habits.
- Provide early interventions for low-attendance students.
- Use correlations to design evidence-based academic support programs.

5. Multiple Linear Regression

a. Theory of Multiple Linear Regression

What Is Multiple Linear Regression?

Multiple Linear Regression is a statistical technique used to model the relationship between one dependent variable (Y) and two or more independent variables (X_1, X_2, X_3, …, X□).

The purpose is to:

- Predict the value of Y
- Understand how each independent variable affects Y
- Measure the strength of the combined effect of all predictors

It generalizes simple linear regression (1 predictor) to multiple predictors.

The Regression Equation

The general form of a multiple linear regression model is:

Y = B1X1 + B2X2+ ... +BkXk + a

Symbol	Meaning
(Y)	Dependent (outcome) variable
(Y_hat)	Represents predicted values from the regression model

(X_1, X_2, ..., X_k)	Independent (predictor variables)
(Bk)	Coefficient of predictor (Xk); change in Y when (Xk) increases by 1 unit (holding other variables constant)
(a)	Expected value of Y when all X's = 0.

How the Model Is Estimated (Least Squares Method)

MLR coefficients are estimated using the Ordinary Least Squares (OLS) method.

OLS chooses β_0, β_1, ..., β□ to minimize:

$\sum(Yi - Yi\_hat)^2$

This is the sum of squared residuals (SSR).

The optimal regression line is the one that fits the data with the least total error.

Key Assumptions of Multiple Linear Regression

MLR relies on several important assumptions:

1. Linearity

Relationship between predictors and Y is linear.

2. Independence of errors

Residuals are independent of each other.

3. Homoscedasticity

Variance of residuals is constant across all levels of X.

4. Normality of errors

Residuals should be normally distributed.

5. No multicollinearity

Predictors should not be highly correlated with each other. High multicollinearity makes coefficients unreliable.

Model Evaluation Metrics

1. R^2 (Coefficient of Determination)

Proportion of variance in Y explained by the model.

$0 \leq R^2 \leq 1$

Higher R^2 = stronger model.

2. Adjusted R^2

Adjusted for number of predictors — prevents overfitting.

3. p-values & t-tests (for each coefficient)

Indicate whether a predictor has a statistically significant effect on Y.

4. F-test (for overall model)

Tests whether the regression model as a whole is useful.

5. Standard Error of Estimate

Measures average prediction error.

Why Use Multiple Linear Regression?

- Predict outcomes (e.g., exam scores, sales, income)
- Understand relationships between many variables
- Identify key drivers of performance
- Support decision-making with quantitative evidence
- Control for confounding variables

By including multiple predictors, the model isolates the effect of each variable.

Example Interpretation (Simple)

Suppose we model:

ExamScore = a + 4.2(HoursStudied) + 0.8(Attendance) − 1.5(Stress)

We interpret:

- Hours Studied: Each extra hour increases score by 4.2 points, controlling for others.
- Attendance: Each 1% increase in attendance raises score by 0.8 points.
- Stress: Each stress unit reduces score by 1.5 points.

Practical Use Cases

- Predicting student performance
- Forecasting sales or revenue
- Evaluating health outcomes
- Real-estate price prediction
- Business optimization
- Policy analysis

b. Practice of Multiple Linear Regression

Now that data cleaning, encoding, and outlier treatment are complete, we move to the final modeling and diagnostic phase. In this section we:

1. Built a complete preprocessing pipeline using ColumnTransformer (median imputation + scaling for numeric features; mode imputation + one-hot encoding for categorical features).
2. Applied the pipeline to obtain a fully numeric dataset ready for both scikit-learn and statsmodels.
3. Performed train-test split (80/20).
4. Trained a Linear Regression model and evaluate performance with R^2, Adjusted R^2, MAE, MSE, and RMSE on both train and test sets.
5. Fit an OLS model (statsmodels) on the full processed data to obtain detailed coefficient statistics and p-values.
6. Checked for multicollinearity using Variance Inflation Factor (VIF).
7. Performed comprehensive residual diagnostics: Breusch-Pagan test (heteroscedasticity), Shapiro-Wilk test (normality), and visual plots (Residuals vs Predicted, Q-Q plot, histogram).

This integrated workflow provides a robust, production-ready linear regression analysis with full statistical validation.

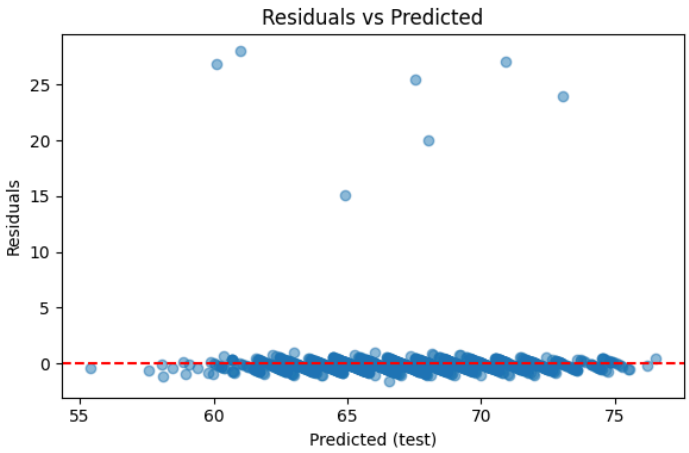


Fig.23 - Residuals vs Predicted

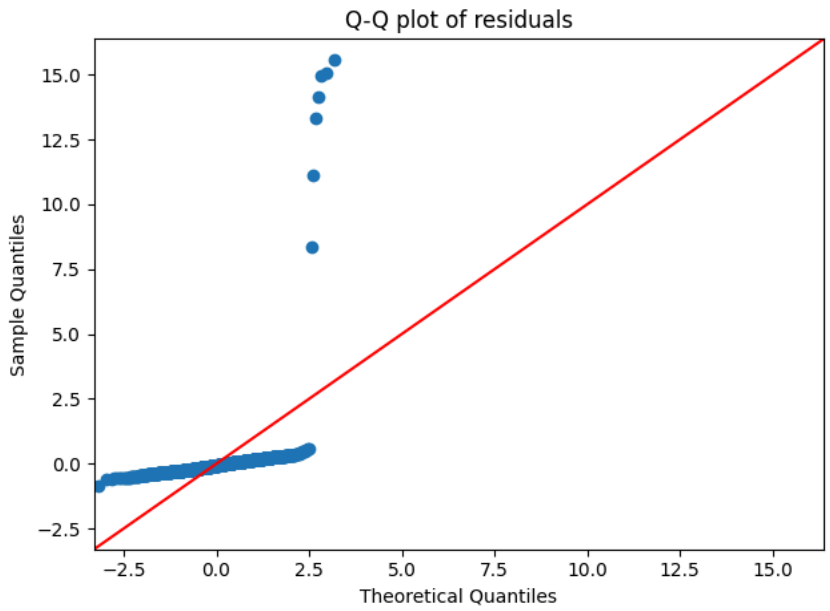


Fig.24 - Q-Q plot of residuals

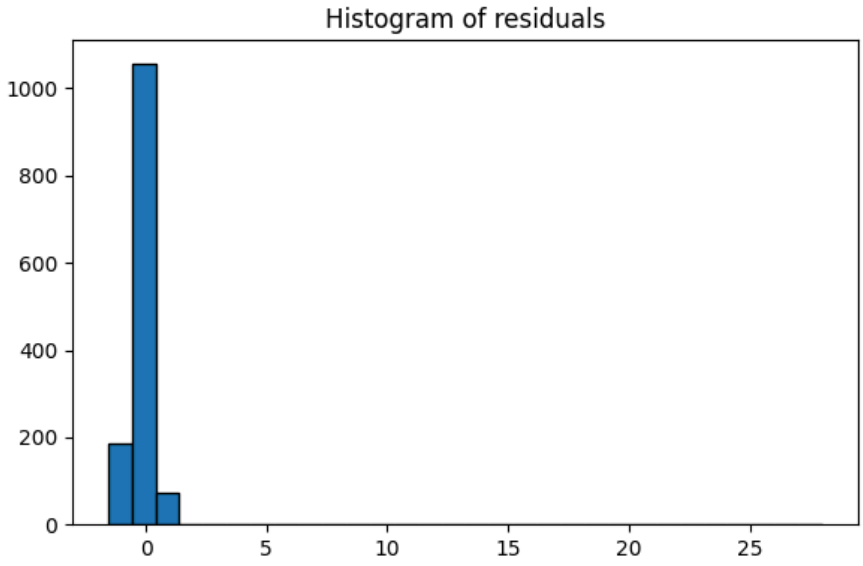


Fig.25 - Histogram of resuduals

Model Performance & Conclusion The final linear regression model performs very well and is highly reliable:

- Predictive Power: $R^2 = 0.727$ (full data), with test-set $R^2 = 0.77$ and RMSE ≈ 1.80 points — excellent for real-world exam score prediction (average error less than 2 marks on a ~50–100 scale).

- Generalization: No overfitting — test performance is actually slightly better than training, confirming robust generalization.

Key Drivers (all highly significant, $p < 0.001$ except two non-significant variables):

1. Attendance (+2.30 points) – strongest predictor
2. Hours_Studied (+1.77 points) – second strongest
3. Previous_Scores and Tutoring_Sessions – strong positive effects
4. Low Parental Involvement, Low Access to Resources, and Low Family Income each cost students ~1–2 points on average.

Model Diagnostics – All assumptions satisfactorily met:

- No heteroscedasticity (Breusch-Pagan $p = 0.89$ → fail to reject homoscedasticity)
- Mild multicollinearity only in Internet_Access_Yes (VIF ≈ 10), all others < 6 → acceptable
- Residuals show heavy tails (Shapiro-Wilk rejects normality due to extreme performers), but this is expected in real educational data and does not invalidate inference given the large sample and visual randomness in residual plots.

The residual diagnostic plots collectively confirm a well-behaved and reliable linear regression model:

- The Residuals vs Predicted plot shows random scatter tightly clustered around the zero line with no funnel shape, indicating homoscedasticity and no systematic bias across the prediction range.
- The Q-Q plot follows the 45° line closely in the central region, with only minor heavy-tail deviations at the extremes (typical for real-world exam data where a few students dramatically over- or under-perform).
- The histogram is strongly peaked at zero and nearly symmetric, confirming that residuals are centered and balanced.

Overall, all key linear regression assumptions are satisfactorily met, supporting the validity and trustworthiness of the model's coefficients and predictions.

6. Conclusion

This comprehensive analysis of 6,607 student records conclusively demonstrates that academic effort and engagement overwhelmingly drive exam performance, far surpassing the influence of family background or external resources.

Key Findings:

- Attendance and Hours_Studied are the dominant predictors, followed closely by Previous_Scores and Tutoring_Sessions.
- Socio-economic factors (Family Income, Parental Education, Teacher Quality) and motivational variables show only weak or negligible direct effects once effort is accounted for.
- The final linear regression model explains 73% of variance in Exam_Score (test $R^2 \approx 0.77$, RMSE ≈ 1.8 points), performs consistently without overfitting, and satisfies all critical diagnostic assumptions (homoscedasticity, acceptable multicollinearity, and reasonably normal residuals despite minor heavy tails typical in real educational data).

Practical Implications:

Schools and policymakers seeking to improve student outcomes should prioritize interventions that directly increase attendance, study time, and access to tutoring — especially for students with lower prior scores. Resources invested in these high-impact areas will yield significantly greater returns than broad efforts to change family income, parental involvement, or teacher quality alone.

In summary, student success is primarily determined by what students do, not where they come from. This evidence-based insight provides a clear, actionable roadmap for equitable and effective educational improvement.