Project Report

1. Dataset Description

This project uses a **Student Performance Dataset** that represents a real-world educational scenario. The dataset contains academic and behavioral information of students and is designed to analyze factors that influence students' final academic performance.

Each observation in the dataset corresponds to an individual student, while the variables describe different academic scores, study habits, and attendance-related attributes. The dataset contains **more than 100 observations** and includes **numerical variables**, making it suitable for probability and statistical analysis.

The data was obtained in CSV format and includes students' scores in core academic subjects such as mathematics, reading, and writing, along with supporting variables like study time, number of failures, absences, and age. These variables help in understanding how academic performance and student behavior contribute to the overall final grade.

To ensure data quality and relevance, unnecessary categorical variables (such as student ID, gender, school type, and lifestyle-related factors) were removed. Only numerical variables relevant to academic performance were retained. Missing values were identified and removed to avoid biased or misleading results.

The dataset is realistic and education-focused, making it appropriate for applying statistical measures, exploratory data analysis, correlation, and regression techniques to study student performance patterns.

2. Variables Description

Here is the table formatted based on your description:

Variable Name	Type	Description
age	Discrete	Age of the student in years
study_time_hours_per_week	Continuous	Weekly study time in hours
absences	Discrete	Number of classes missed
failures	Discrete	Number of failed subjects
math_score_0_100	Continuous	Mathematics score (0–100)
reading_score_0_100	Continuous	Reading score (0–100)
writing_score_0_100	Continuous	Writing score (0–100)
final_grade_0_100	Continuous	Final academic grade 0–100

Dependent Variable: Final Grade (final_grade_0_100)
Independent Variables: Age, Study Time, Absences, Failures, Math Score, Reading Score, Writing Score

3. Objective of the Study

The objective of this study is to analyze the factors that influence students' final academic performance using statistical and exploratory data analysis techniques. The dataset is used to examine how academic subject scores (mathematics, reading, and writing) along with student-related factors such as study time, number of failures, absences, and age affect the final grade of a student.

This study aims to identify relationships and patterns between independent variables and the dependent variable (final grade) through descriptive statistics, graphical analysis, correlation, and simple linear regression. Understanding these relationships helps determine which factors have the strongest impact on students' academic outcomes.

The dataset is relevant because student performance is a realistic and practical educational problem, and analyzing it provides meaningful insights into how academic and behavioral variables contribute to overall success. The findings of this study can help in understanding performance trends and in making data-driven academic decisions.

4. Data Cleaning

Data cleaning is a critical step to ensure the accuracy, reliability, and validity of statistical analysis. Before performing any exploratory data analysis, correlation, or regression, the dataset was carefully inspected and cleaned to eliminate inconsistencies and potential sources of bias.

Initially, the dataset was examined to understand its structure, data types, and completeness. Missing values were identified across several variables. As per the project requirements, all rows containing missing data were **removed** rather than imputed. This approach was chosen to prevent introducing artificial values that could distort statistical measures such as mean, variance, correlation, and regression results.

The number of observations before and after cleaning was recorded to quantify the impact of missing data removal. Only complete records were retained, ensuring that all variables used in the analysis had valid numerical values.

In addition to handling missing values, irrelevant and non-numerical variables such as student identification details, demographic attributes, and lifestyle-related factors were removed. This step helped reduce noise in the dataset and allowed the analysis to focus solely on variables directly related to academic performance.

The dataset was also checked for duplicate records, and no duplicate entries were found. Data types of all remaining variables were verified to confirm that they were appropriate for statistical analysis. Numerical consistency was ensured across all columns.

After completing the data cleaning process, the final dataset consisted of clean, complete, and reliable observations. This prepared dataset was suitable for performing descriptive statistics, outlier detection, graphical analysis, correlation analysis, and simple linear regression without risking biased or misleading conclusions.

5. Exploratory Data Analysis (EDA)

Statistical Analysis Results

Descriptive statistical measures were calculated for all numerical variables to understand the central tendency and variability within the dataset.

The **age** variable has a **mean of 16.32 years**, with a **median and mode of 16**, indicating that most students are concentrated around the age of 16. The **low variance (1.27)** and **standard deviation (1.13)** suggest minimal age dispersion, confirming a consistent age group across the dataset.

The **study time per week** shows a **mean of 10.03 hours** and a **median of 10 hours**, indicating that students generally study around 10 hours weekly. The **standard deviation of 3.87** and **variance of 14.94** reflect moderate variability, suggesting differences in individual study habits among students.

The **failures** variable has a **mean of 0.34**, while both the **median and mode are 0**, showing that the majority of students have not experienced academic failure. However, the **variance (0.53)** and **standard deviation (0.73)** indicate the presence of a smaller group of students with higher failure counts, contributing to variability in this variable.

The **absences** variable has a **mean of 2.89**, with a **median of 3** and **mode of 2**, suggesting generally good attendance among students. The relatively low **standard deviation (1.59)** and **variance (2.51)** indicate limited dispersion, meaning most students have a similar number of absences.

Academic performance variables show consistent patterns. The **math score** has a **mean of 56.42**, **median of 56.40**, and **mode of 56.80**, indicating average performance with a fairly symmetric distribution. The **standard deviation of 11.70** reflects moderate variability in math performance.

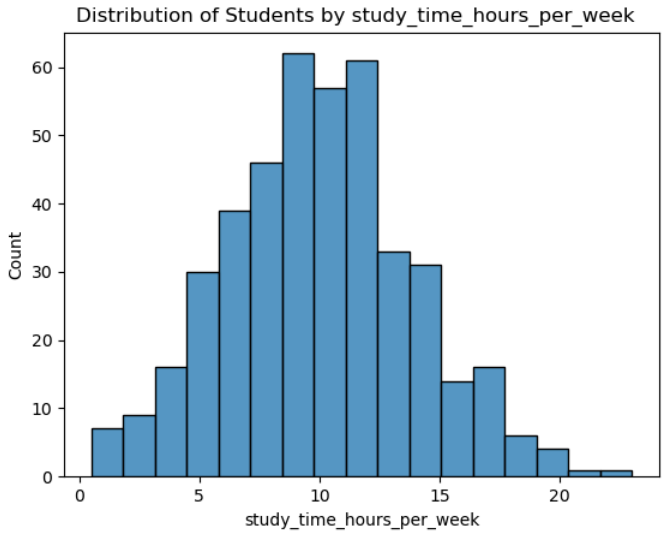
Similarly, the **reading score** has a **mean of 55.10**, **median of 55.40**, and **mode of 54.60**, with a **standard deviation of 11.69**, showing comparable variability and performance trends to math scores.

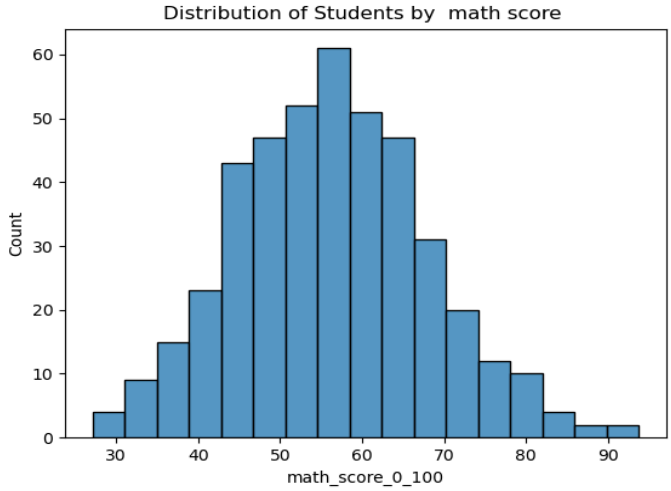
The **writing score** has a **mean of 55.56** and **median of 55.80**, indicating average writing performance. Although the **mode (39.50)** suggests a cluster of lower scores, the **standard deviation (12.05)** and **variance (145.19)** indicate slightly higher variability compared to math and reading scores.

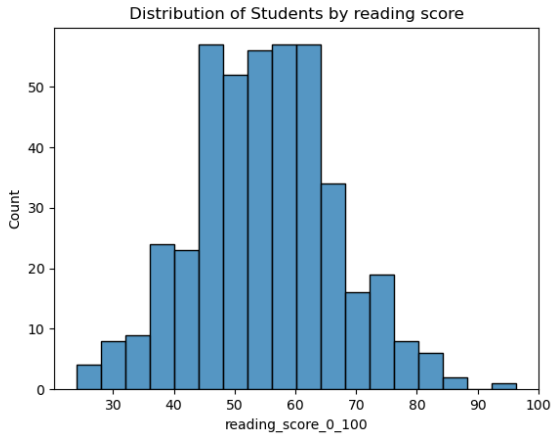
The **final grade** has a **mean of 56.01**, with a **median of 55.40** and **mode of 54.40**, showing that most students achieve an average final result. The **standard deviation of 11.37** indicates moderate dispersion, reflecting differences in overall academic achievement.

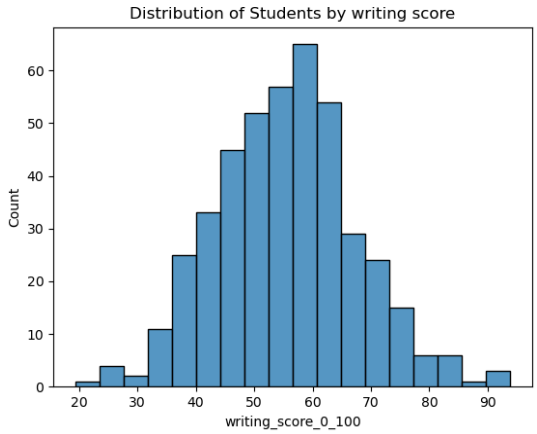
Overall, the close alignment between mean and median values across most variables suggests approximately symmetric distributions, while moderate standard deviations indicate reasonable variability without extreme dispersion. These results confirm that the dataset is well-balanced and suitable for further correlation and regression analysis.

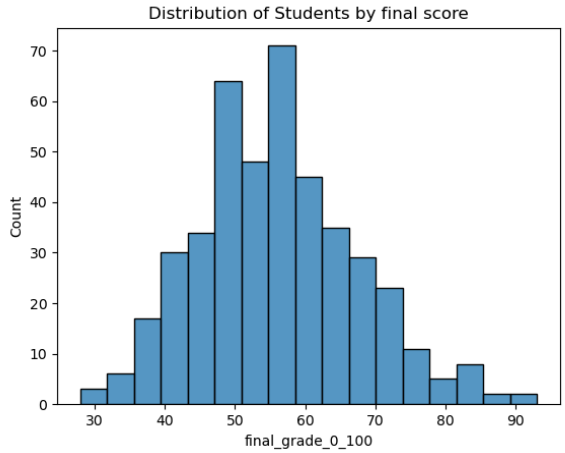
Graphical and Distribution Analysis











Exploratory Data Analysis was performed using histograms, bar charts, box plots, and scatter plots to understand the distribution, spread, and behavior of variables in the dataset.

The distribution of **study time per week** shows that most students study between **10 and 11 hours**, which represents the most frequent study duration. This indicates a moderate and consistent study pattern among students, with very few students studying extremely low or high hours.

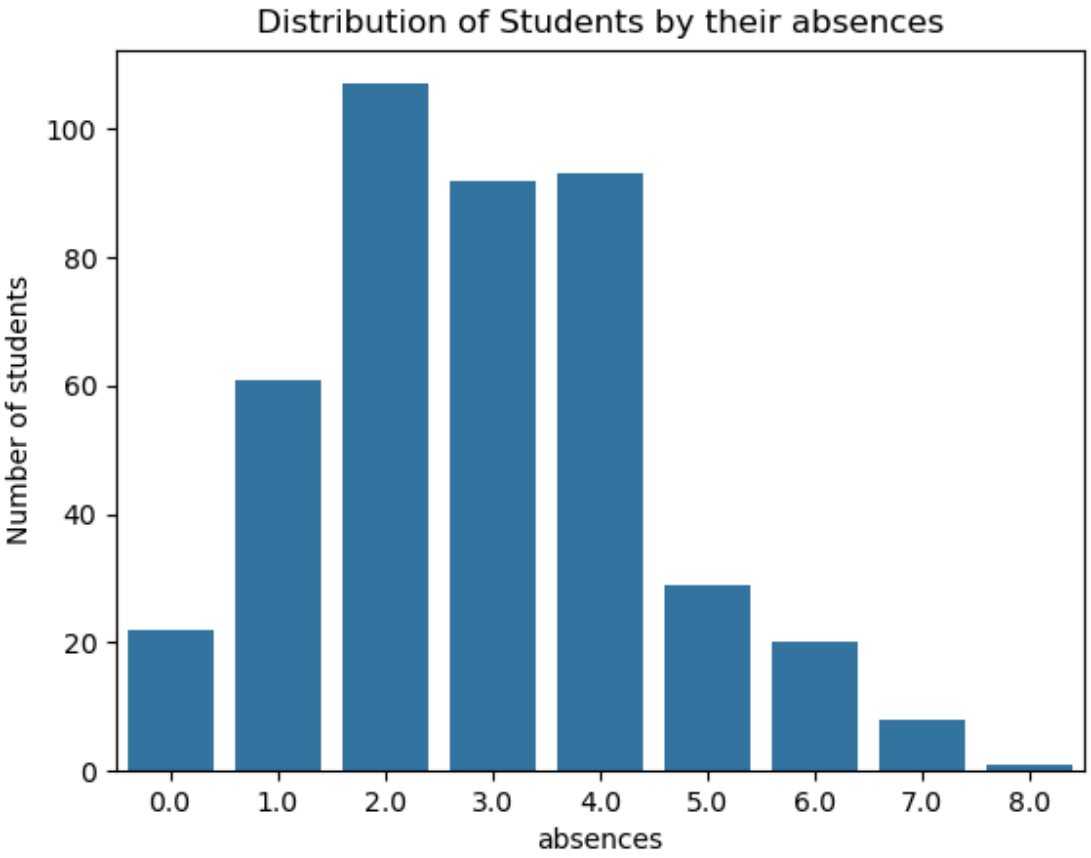
The **math score distribution** reveals that the majority of students scored between **50 and 60**, suggesting average performance in mathematics. The distribution is slightly symmetric, indicating balanced performance across the dataset. Similarly, **reading scores** are mostly concentrated between **45 and 65**, showing moderate variability, while **writing scores** peak at **60 marks**, making it the most frequent score in writing.

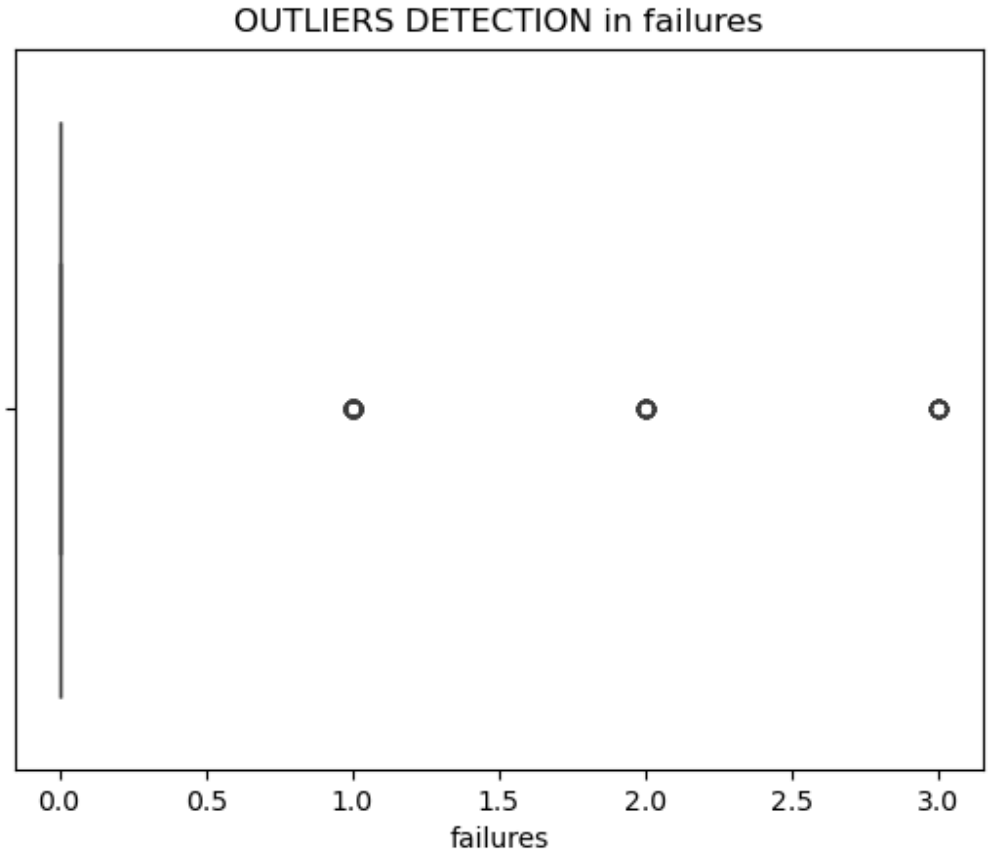
The **final grade distribution** also peaks at **60 marks**, indicating that most students achieve an average overall academic result. The similarity between subject score distributions and final grade distribution suggests that individual subject performance strongly contributes to the final grade.

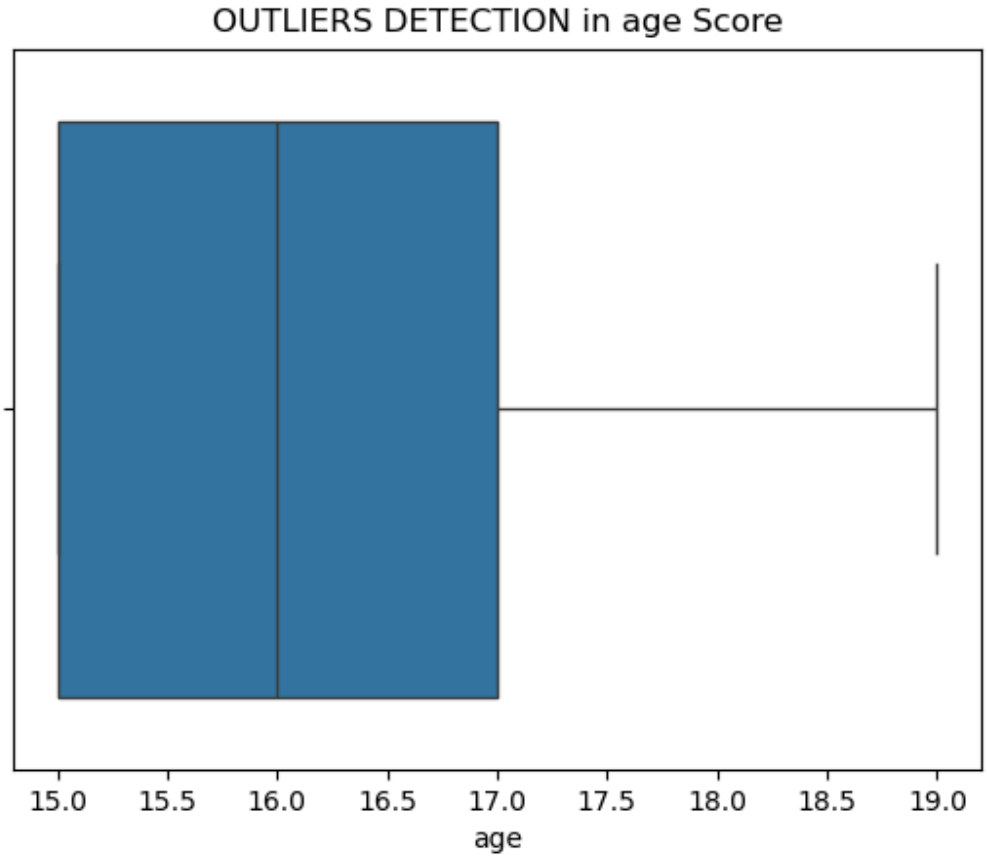
The **age distribution** shows that the majority of students are **16 years old**, with more than **140 observations**, making it the dominant age group in the dataset. This indicates a fairly uniform age range, which reduces age-related bias in performance analysis.

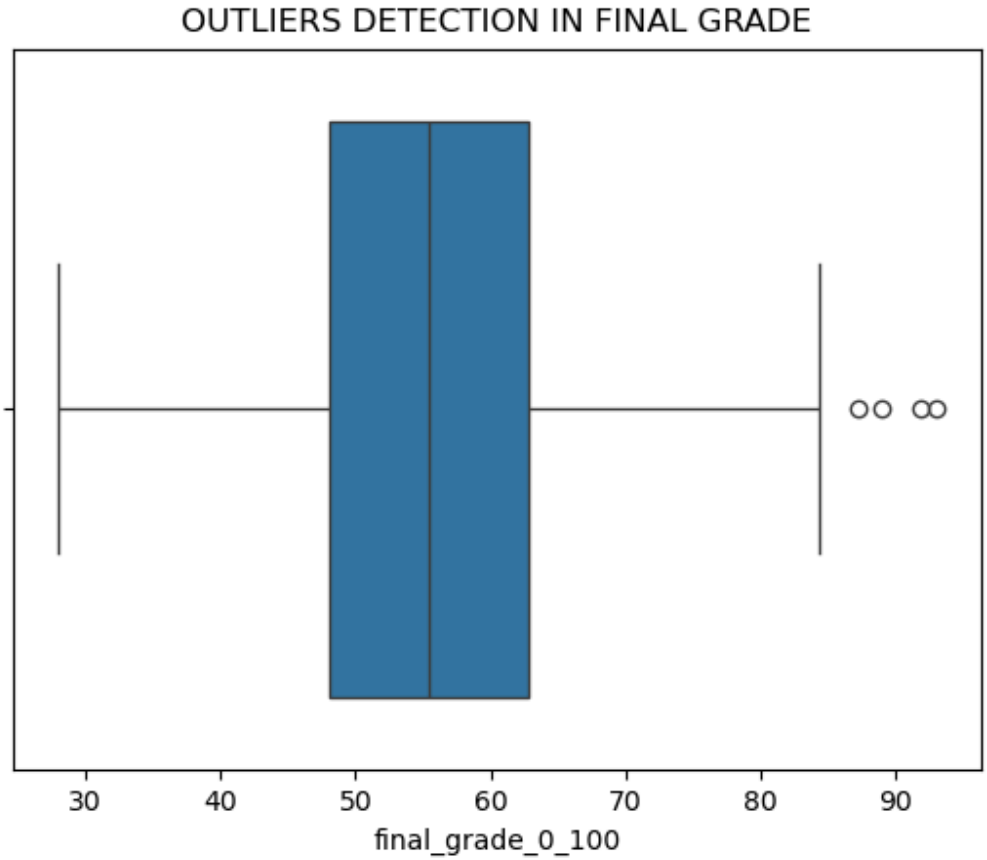
Analysis of **academic failures** shows that over **300 students have zero failures**, indicating that most students successfully pass their subjects. This reflects generally satisfactory academic outcomes. The **absences distribution** shows that more than **100 students have only two absences**, suggesting that attendance behavior is generally good and consistent among students

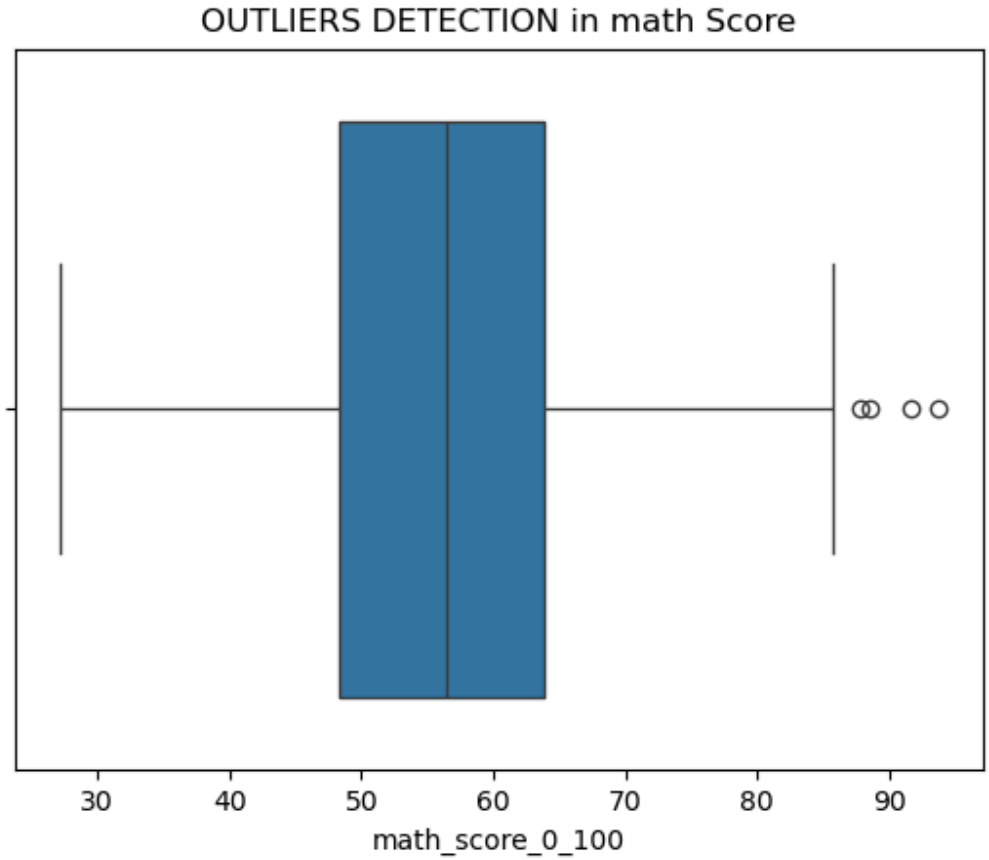
Outlier Analysis

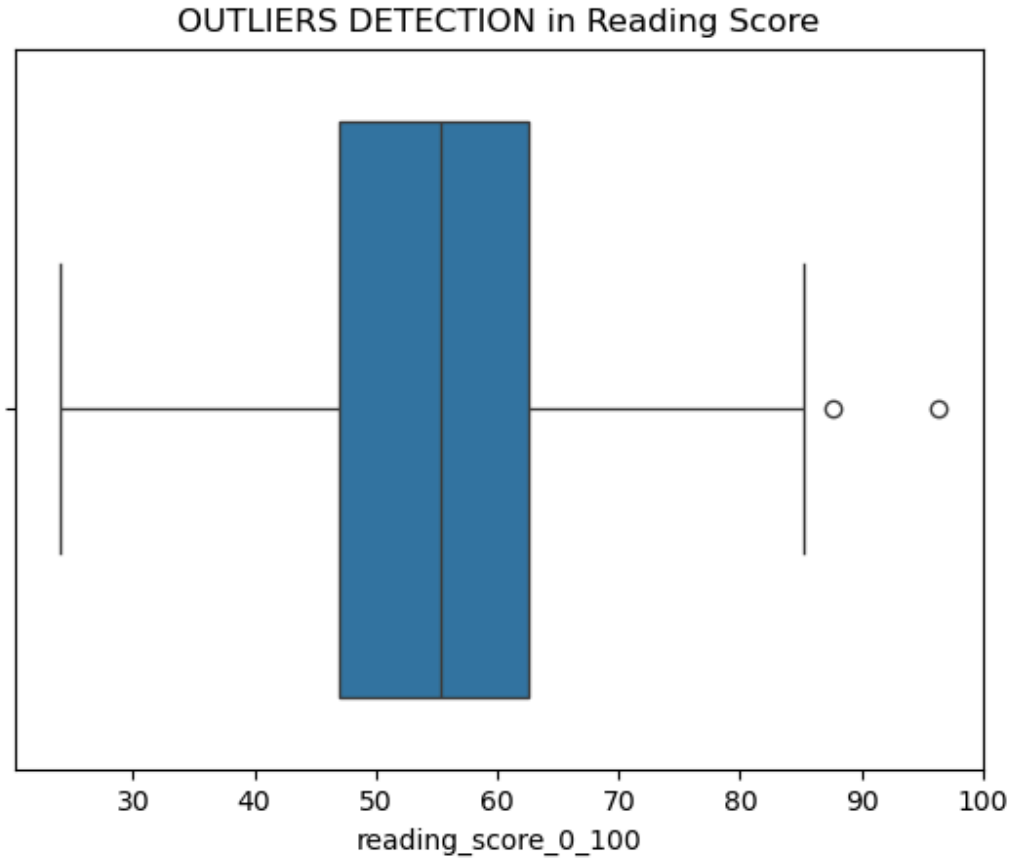


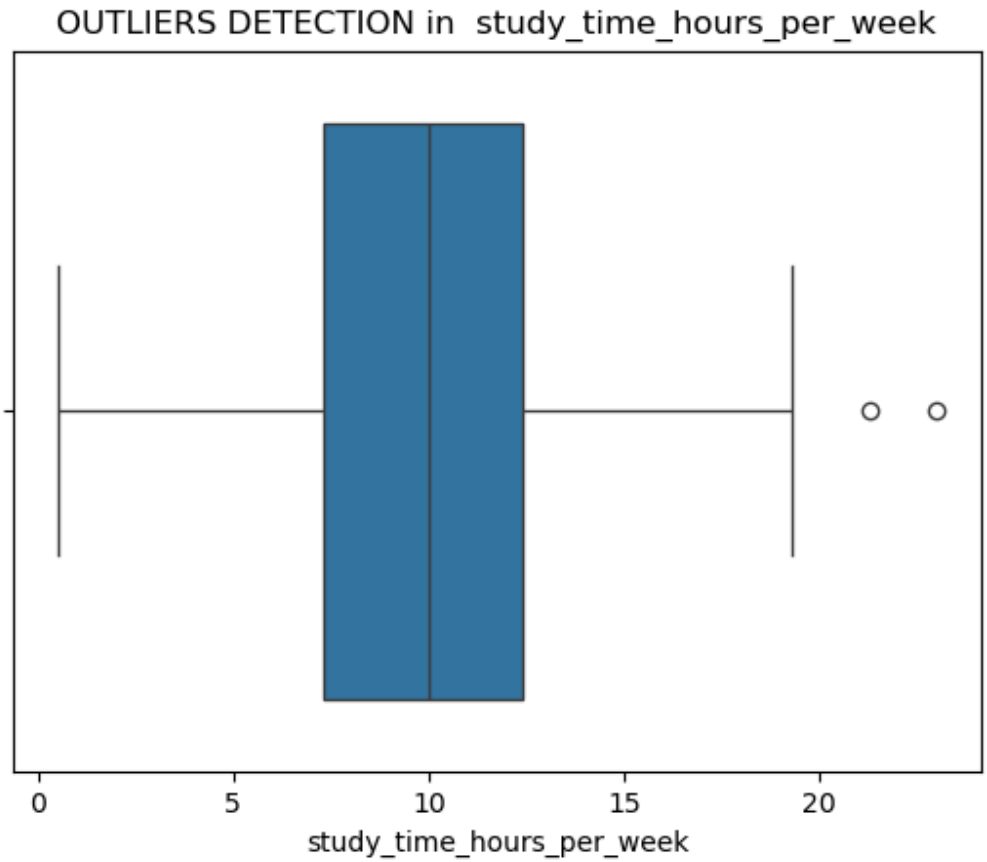

OUTLIERS DETECTION in failures
0.0
0.5
1.0
1.5
2.0
2.5
3.0
failures


OUTLIERS DETECTION in age Score
15.0
15.5
16.0
16.5
17.0
17.5
18.0
18.5
19.0
age


OUTLIERS DETECTION IN FINAL GRADE
30
40
50
60
70
80
90
final_grade_0_100


OUTLIERS DETECTION in math Score
30
40
50
60
70
80
90
math_score_0_100


OUTLIERS DETECTION in Reading Score
30
40
50
60
70
80
90
100
reading_score_0_100


OUTLIERS DETECTION in study_time_hours_per_week
0
5
10
15
20
study_time_hours_per_week


OUTLIERS DETECTION in Writing Score
20
30
40
50
60
70
80
90
writing_score_0_100

Outliers were identified using the **Interquartile Range (IQR) method** and visualized using box plots. No outliers were detected in the **age** variable, indicating a consistent and realistic age range.

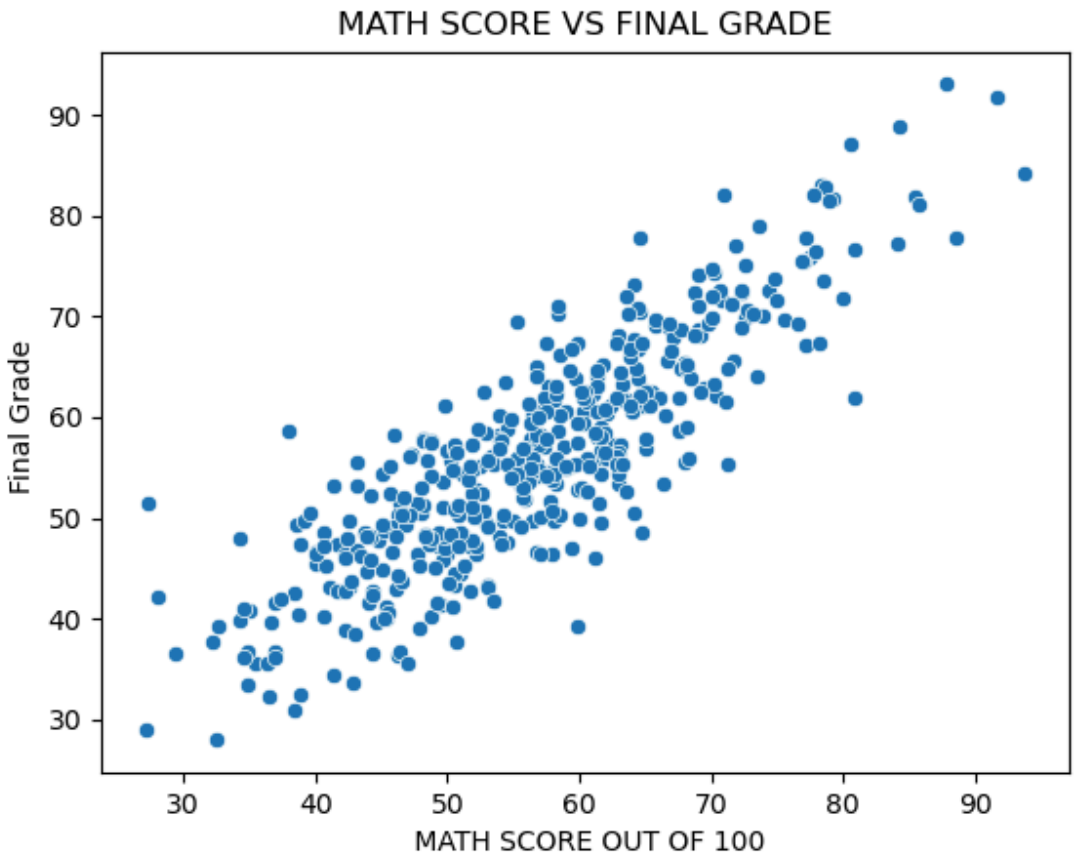
The **study time** variable contains **2 outliers**, suggesting that a small number of students study unusually high or low hours compared to the majority. The **absences** variable contains **1 outlier**, indicating a student with unusually high absenteeism.

The **failures** variable shows **93 outliers**, reflecting high variability in the number of failures. This suggests that although most students have no failures, a subset of students experiences repeated academic difficulties.

Academic performance variables show limited outliers: **math (4)**, **reading (2)**, **writing (4)**, and **final grade (4)**. These limited outliers indicate that extreme academic scores are rare and do not significantly distort overall analysis.

Overall, the dataset remains stable and reliable, with minimal extreme values, except for the failures variable, which naturally exhibits higher variability.

6. Regression and Correlation Analysis



Correlation Analysis

Correlation analysis was applied to measure the **strength and direction** of the relationship between the independent variables and the dependent variable (**final_grade_0_100**).

The results show that academic subject scores have a **strong positive relationship** with the final grade. The **math score** shows the strongest correlation with final grade (**r = 0.860**), indicating a very strong positive relationship. This means that as math scores increase, final grades also increase significantly. Similarly, **writing scores (r = 0.845)** and **reading scores (r = 0.843)** also exhibit strong positive correlations with the final grade.

In contrast, **study time per week** has a **weak positive correlation (r = 0.201)** with the final grade, suggesting that increased study time has only a small effect on final performance. The **absences** variable shows a **weak negative correlation (r = –0.078)**, indicating that higher absences slightly reduce final grades, though the relationship is weak.

Overall, the correlation analysis indicates that **subject-wise academic performance is a much stronger indicator of final grades than behavioral factors such as study time and attendance**

Regression Analysis

A **simple linear regression model** was applied using:

- **Dependent Variable:** Final Grade (final_grade_0_100)
- **Independent Variable:** Math Score (math_score_0_100)

Math score was selected as the independent variable because it showed the **highest correlation** with the final grade.

The regression model establishes a linear relationship between math score and final grade and can be represented by the following equation:

Final Grade = $\beta 0 + \beta 1 \times$ Math Score

Where:

- **Final Grade** is the dependent variable
- **Math Score** is the independent variable
- $\boldsymbol{\beta_0}$ is the intercept
- $\boldsymbol{\beta_1}$ is the regression coefficient representing the change in final grade for each one-unit increase in math score

The regression results indicate a strong positive linear relationship, meaning that students with higher math scores are likely to achieve higher final grades. The model explains a large proportion of the variation in final grades, supporting the correlation findings.

This analysis confirms that **math performance is a significant predictor of overall academic success** and plays a crucial role in determining students' final grades.

7. Interpretation of Results

The analysis shows that most students demonstrate average academic performance, with the highest concentration of final grades around **60 marks**. Study behavior is generally consistent, as the majority of students study **10–11 hours per week** and maintain good attendance, with most students having **zero failures** and only **two absences**.

Outlier analysis indicates that the dataset is stable, with **no outliers in age** and only a few extreme values in study time, absences, and academic scores. A higher number of outliers in the failures variable suggests that while most students do not fail, a small group experiences repeated academic difficulties.

Correlation and regression results clearly indicate that **subject-wise academic performance is the strongest determinant of final grades**. Math, reading, and writing scores show strong positive relationships with the final grade, while study time has a weak positive effect and absences have a weak negative effect.

Overall, the findings confirm the study objective by showing that **academic scores—particularly mathematics have the greatest impact on student performance**, whereas behavioral factors play a secondary role.

8. Conclusion

This study analyzed the factors affecting students' final academic performance using statistical analysis, graphical methods, correlation, and simple linear regression. The results show that subject-wise academic scores, particularly mathematics, have the strongest influence on the final grade. Reading and writing scores also contribute significantly, while behavioral factors such as study time and attendance have a comparatively weaker impact.

The analysis confirms that actual academic performance is a more reliable indicator of student success than study habits alone. Overall, the findings support the study objective and demonstrate the effective application of probability and statistical techniques to a real-world educational dataset.