



**L** OVELY  
**P** ROFESSIONAL  
**U** NIVERSITY

**<DATA SCIENCE TOOLBOX: PYTHON PROGRAMMING>**

**PROJECT REPORT**

(Project Semester January-April 2025)

***(Student Performance Analysis and Score Prediction Using Data  
Visualization)***

Submitted by:

**(VIKASH KUMAR)**

Registration No: **12315362**

Programme and Section : **CSE And K23GD**

Course Code : **INT375**

Under the Guidance of

**(BALJINDER KAUR , 27952)**

**Discipline of CSE/IT**

**School of Computer Science and Engineering**

**Lovely Professional University, Phagwara**

## **CERTIFICATE**

This is to certify that **VIKASH KUMAR** bearing Registration no. **12315362** has completed **INT375** project titled, “***Student Performance Analysis and Score Prediction Using Data Visualization***” under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

**Signature and Name of the Supervisor**

**Designation of the Supervisor**

**School of Computer Science and Engineering**

**Lovely Professional University**

**Phagwara, Punjab.**

Date: 11/04/2025

## **DECLARATION**

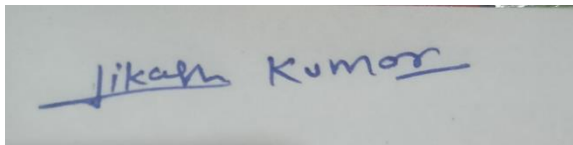
I, VIKASH KUMAR student of **B.Tech CSE - K23GD** under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 11/04/2025

Signature

Registration No. **12315362**

Name of the student : **VIKASH KUMAR**

A photograph of a handwritten signature in blue ink on a white piece of paper. The signature reads "Vikash Kumar" in a cursive script. The first name "Vikash" is underlined with a single horizontal stroke.

Signature

# ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my faculty guide **BALJINDER KAUR** for her valuable guidance, support, and encouragement throughout this project. I am thankful to **Lovely Professional University** for providing the necessary resources and environment for conducting this project successfully. I also extend my appreciation to all those who directly or indirectly contributed to the completion of this report.

## Table of Content

1. Introduction
2. Source of dataset
3. EDA process
4. Analysis on dataset (for each analysis)
  - i. Introduction
  - ii. General Description
  - iii. Specific Requirements, functions and formulas
  - iv. Analysis results
  - v. Visualization
5. Conclusion
6. Future scope
7. References

**DataSet:-** <https://www.kaggle.com/datasets/mexwell/student-scores>

**Github :-** <https://github.com/vikash260905/Student-Performance-Analysis-and-Score-Prediction-Using-Data-Visualization>

**LinkedIn:-** [https://www.linkedin.com/posts/vikash-yadav-350452297\\_python-datascience-dataanalysis-activity-7316885361870229505-CUSE?utm\\_source=share&utm\\_medium=member\\_desktop&rcm=ACoAAEfKhGwBW-yIJ0j\\_VbrzHpRakXgeWswjc-Q](https://www.linkedin.com/posts/vikash-yadav-350452297_python-datascience-dataanalysis-activity-7316885361870229505-CUSE?utm_source=share&utm_medium=member_desktop&rcm=ACoAAEfKhGwBW-yIJ0j_VbrzHpRakXgeWswjc-Q)

# Introduction

In the modern education system, understanding student performance plays a critical role in enhancing learning outcomes and identifying areas that require academic support. With the growing availability of digital academic records and assessment scores, data analytics has become a powerful tool for gaining insights into student learning behavior and performance trends. This project focuses on analyzing a dataset containing scores of students in various academic subjects. The primary goal is to conduct a comprehensive exploratory data analysis (EDA) to understand the distribution, relationships, and trends in the dataset. This includes identifying high and low performers, visualizing subject-wise performance, and detecting any anomalies or outliers in the data. Furthermore, this project implements a simple **Linear Regression model** to predict a student's performance in mathematics based on their scores in other subjects. This predictive modeling adds a machine learning dimension to the analysis, showcasing how academic performance can be forecasted using historical data.

## Source of Dataset

The dataset utilized in this project, titled "*student-scores python.csv*", is designed to represent academic performance data of individual students across multiple subjects. This dataset is not sourced from a public repository such as Kaggle or UCI Machine Learning Repository; rather, it appears to be a custom-curated file meant for educational and analytical purposes. It contains structured data including student names along with their scores in subjects like Mathematics, Reading, Writing, Science, and Social Studies. Stored in CSV format, the dataset is easy to import and handle within Python using libraries such as Pandas. This format is

especially beneficial for performing exploratory data analysis (EDA), data cleaning, and machine learning model development. Though not a real-world dataset, it accurately reflects typical educational data structures, allowing us to apply a range of statistical and visualization techniques. Overall, the dataset serves as a practical resource for studying patterns in student performance and implementing prediction models based on academic scores.

## Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a crucial initial step in understanding the structure and quality of a dataset. In this project, the EDA process began by loading the student performance dataset using the Pandas library. A preliminary inspection using `.info ()` and `.describe ()` provided insights into data types, missing values, and basic statistics such as mean, min, max, and standard deviation of scores. Next, data cleaning was performed by removing missing values (`dropna ()`) and duplicates (`drop duplicates ()`) to ensure the dataset's integrity. We identified numerical columns to focus our analysis on student scores. Several visualization techniques were then applied to uncover hidden patterns and trends:

- **Histograms** to understand the distribution of each subject's scores.
- **Boxplots** to detect outliers and observe score spread.
- **Bar charts and pie charts** to compare average scores across subjects.
- **Scatter plots** to study the correlation between two numeric subjects.
- **Line graphs** to trace performance trends of the top 10 students.
- **Heatmaps** to explore correlations among all numeric features.

This EDA step provided a comprehensive understanding of the students' performance and guided the development of predictive models in later stages.

## Analysis of Dataset

This section breaks down the analysis performed on the "student-scores python.csv" dataset. Various techniques and visualizations were employed to explore the data, understand patterns, and gain insights into student performance across multiple subjects.

### Introduction

The dataset contains scores from different academic subjects, including Mathematics, Reading, and Writing. The primary goal of the analysis was to explore the distribution of scores, detect any outliers, identify the top and bottom-performing students, and assess correlations between the scores of different subjects.

### ii. General Description

The dataset includes the following columns:

- **first\_name**: The name of the student.
- **math\_score**: The score of the student in Mathematics.
- **reading\_score**: The score of the student in Reading.
- **writing\_score**: The score of the student in Writing.

Before diving into detailed analysis, the dataset was cleaned by removing missing values and duplicates to ensure accuracy.

### Analysis Results

## 1. Histogram (Distribution of Scores):

### General Description:

Histograms show the frequency distribution of student scores for each subject. It helps us understand how scores are spread across students.

### Specific Requirements:

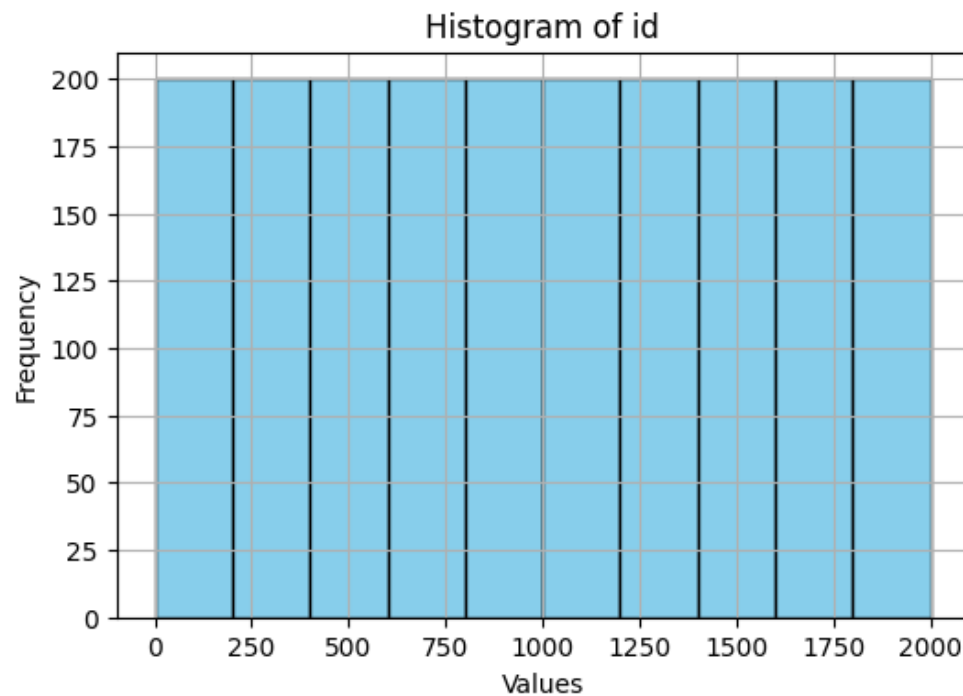
- Plot histograms for each numeric column (subject scores).
- Use 10 bins for better granularity.

### Analysis Result:

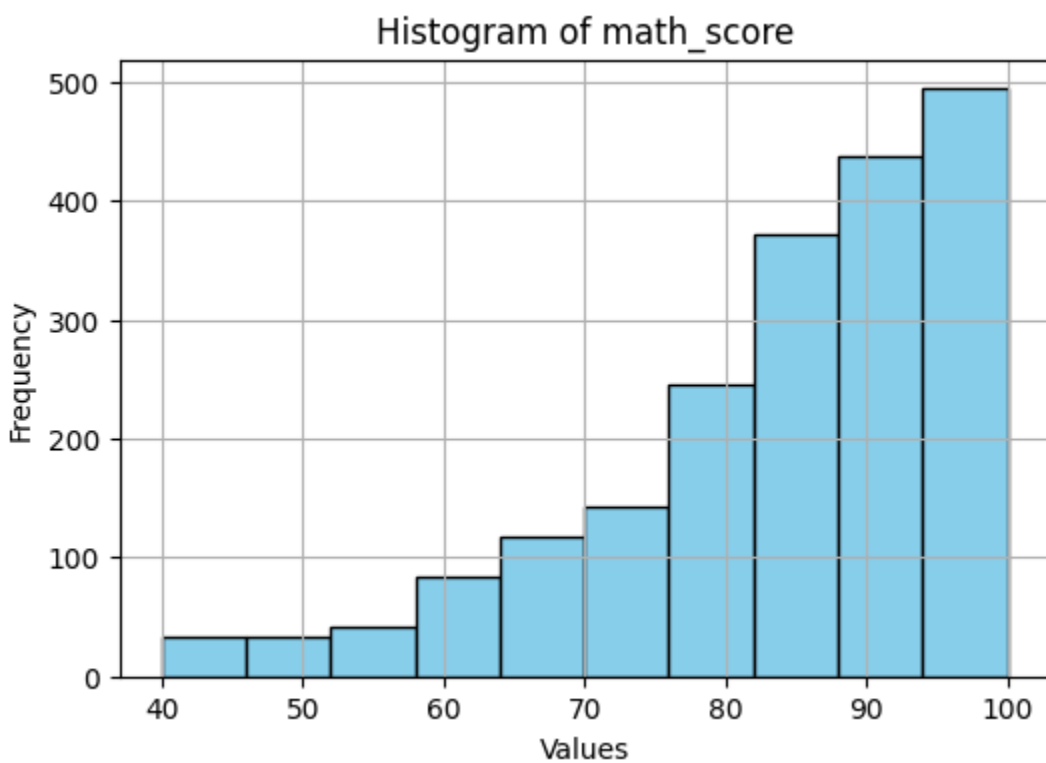
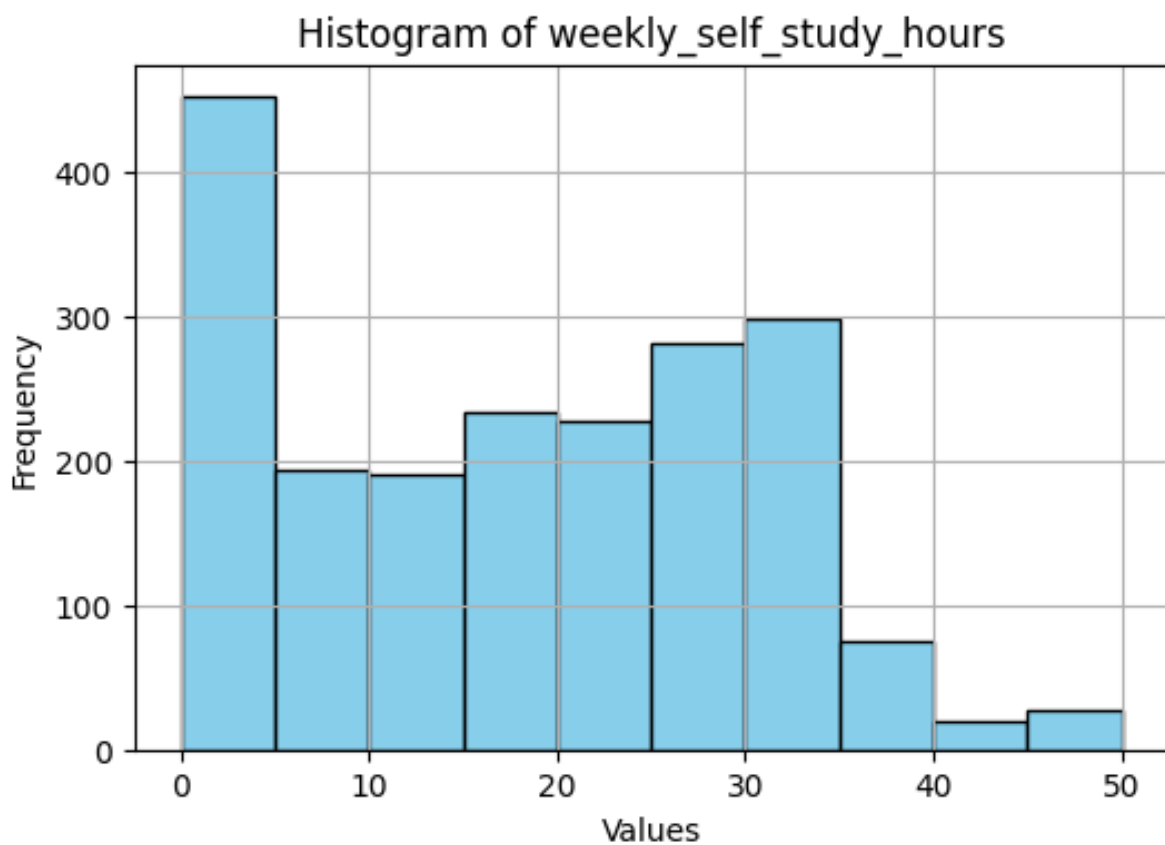
The histograms revealed the general performance trend. For most subjects, scores are skewed towards higher marks, indicating good overall performance.

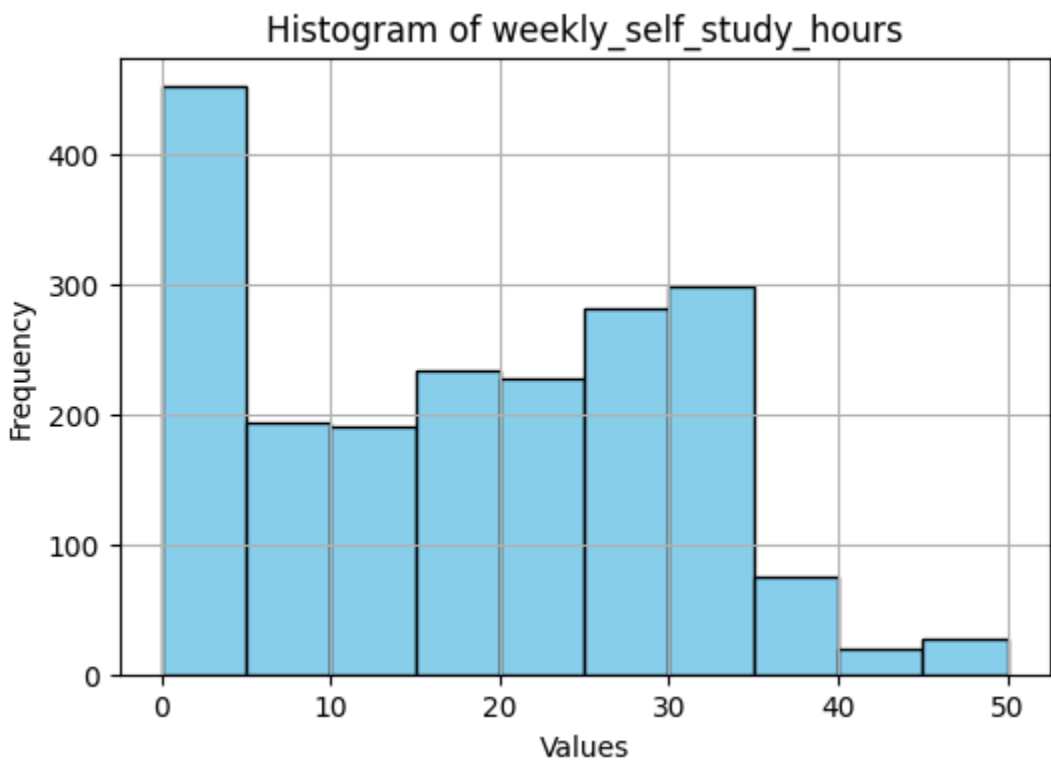
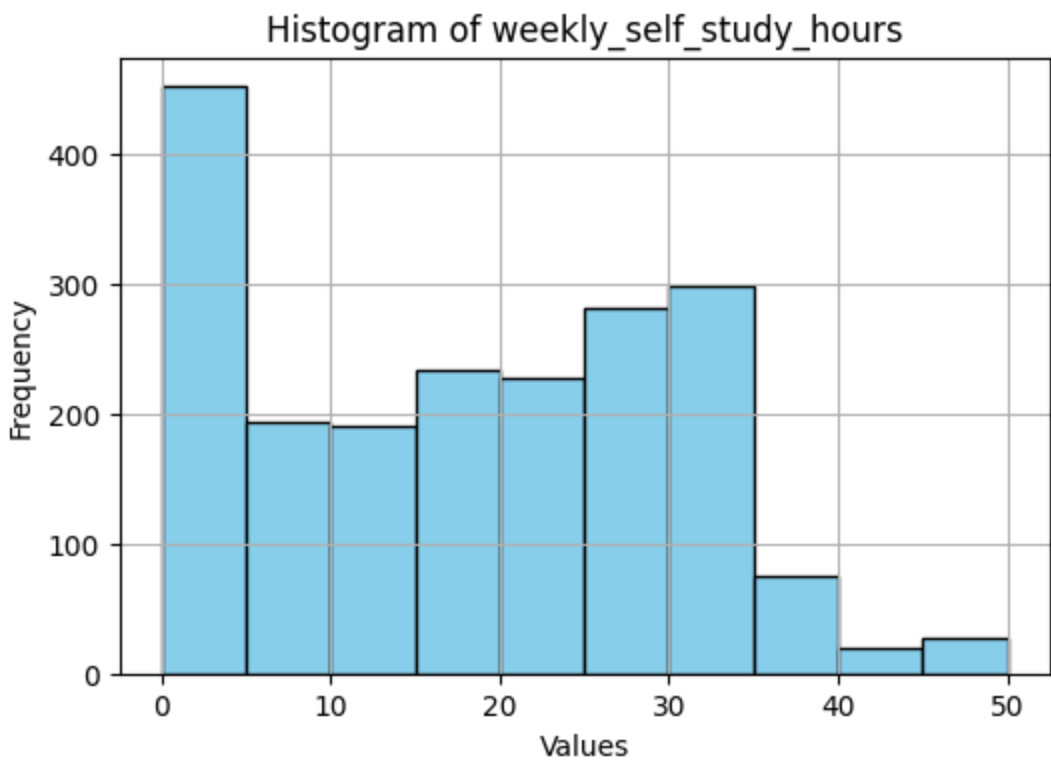
### Visualization:

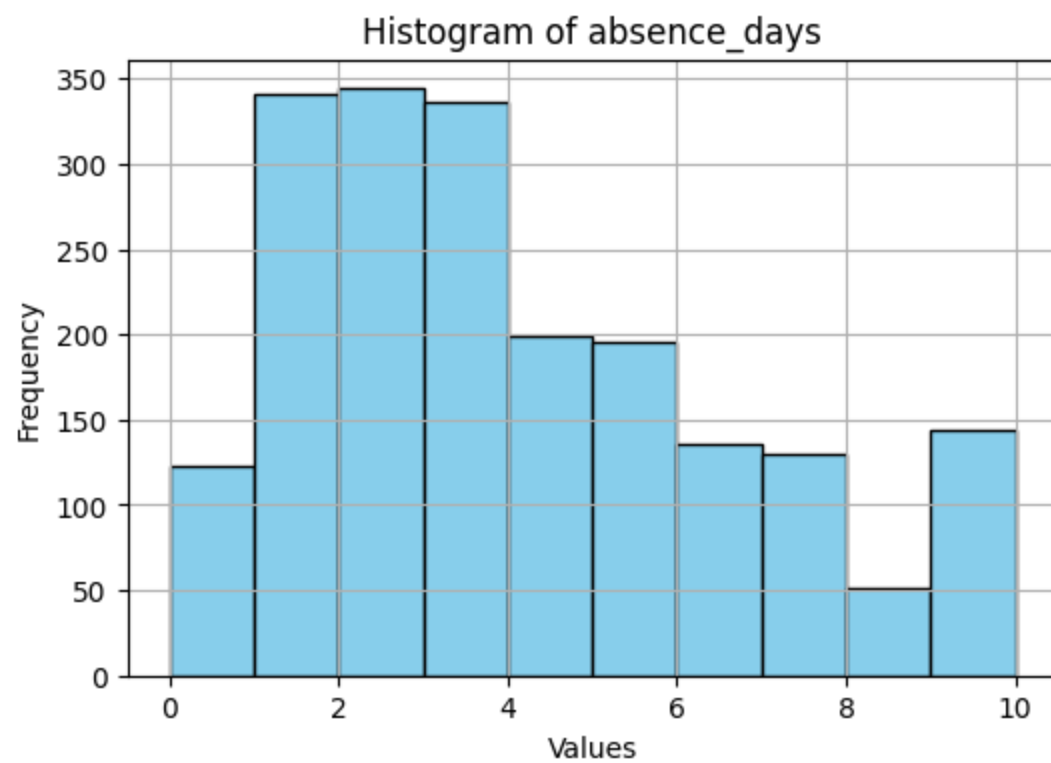
- Multiple histograms, each representing score distribution for a subject.

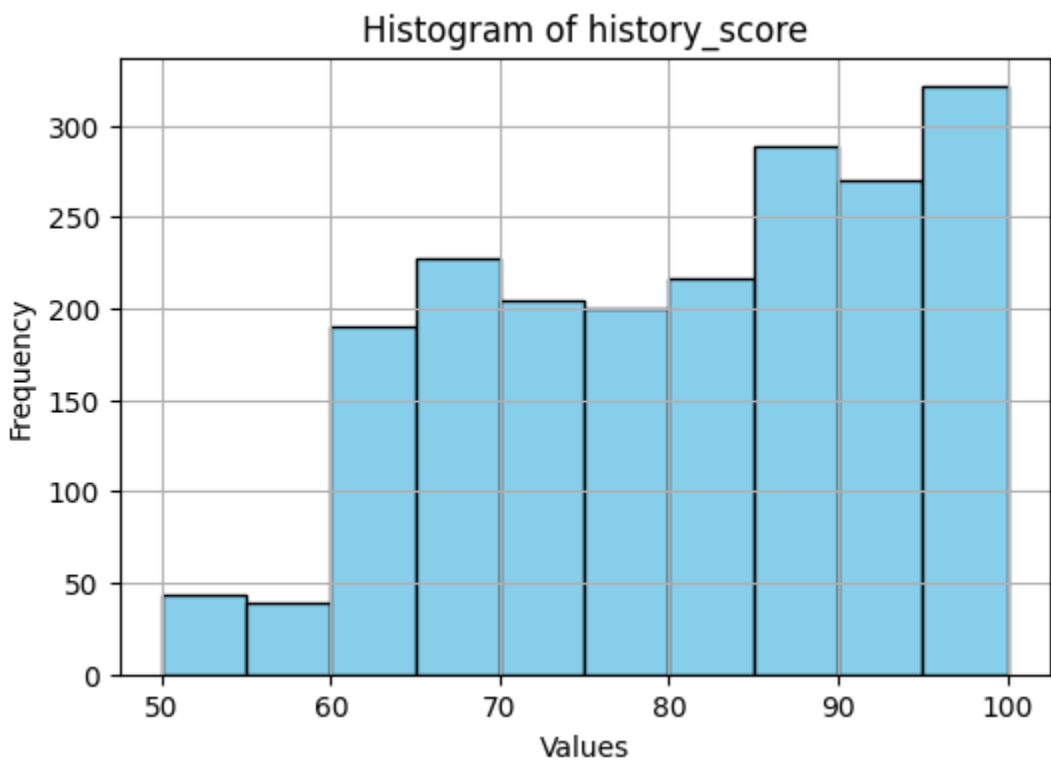
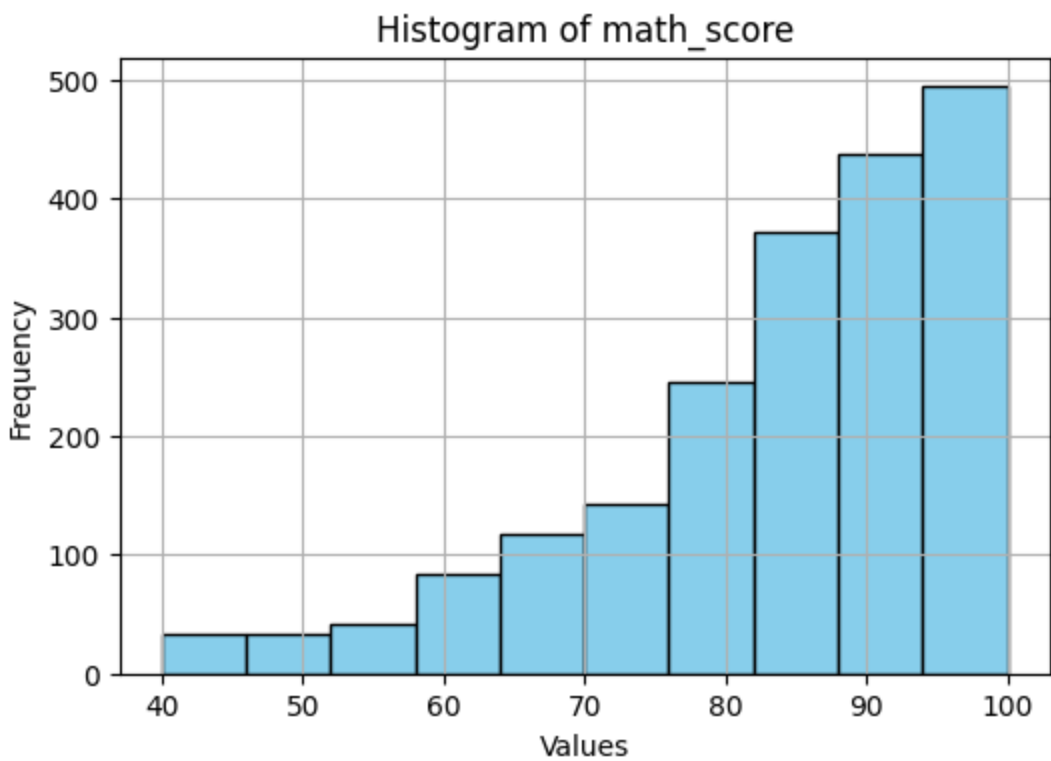


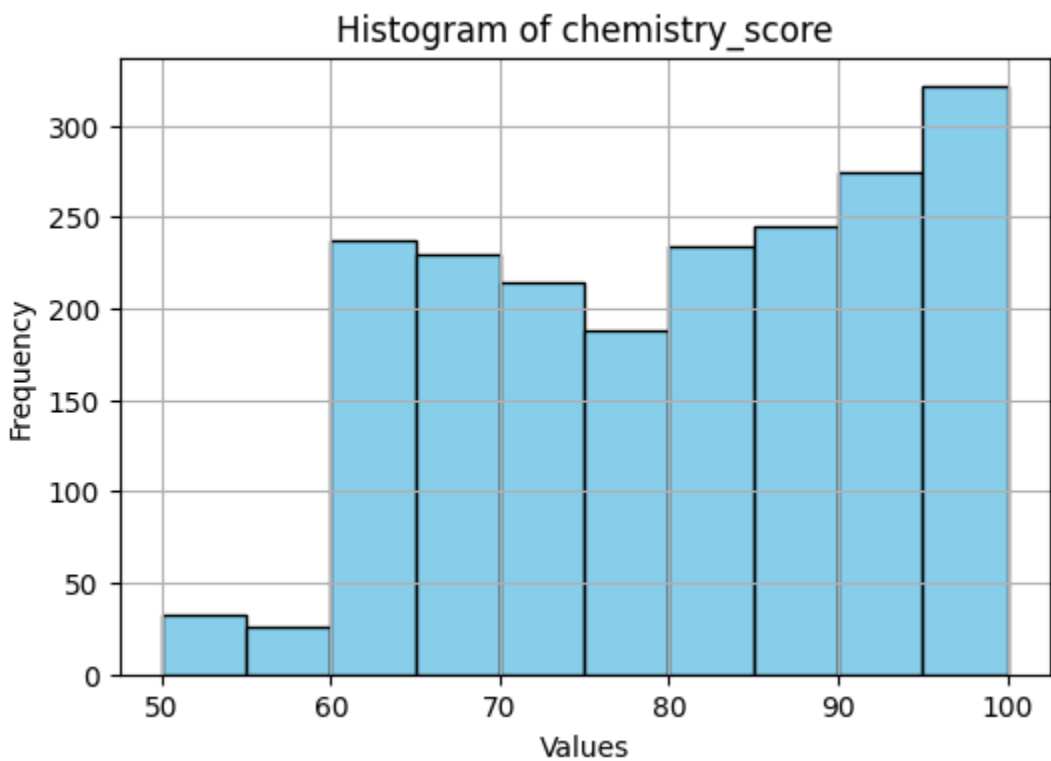
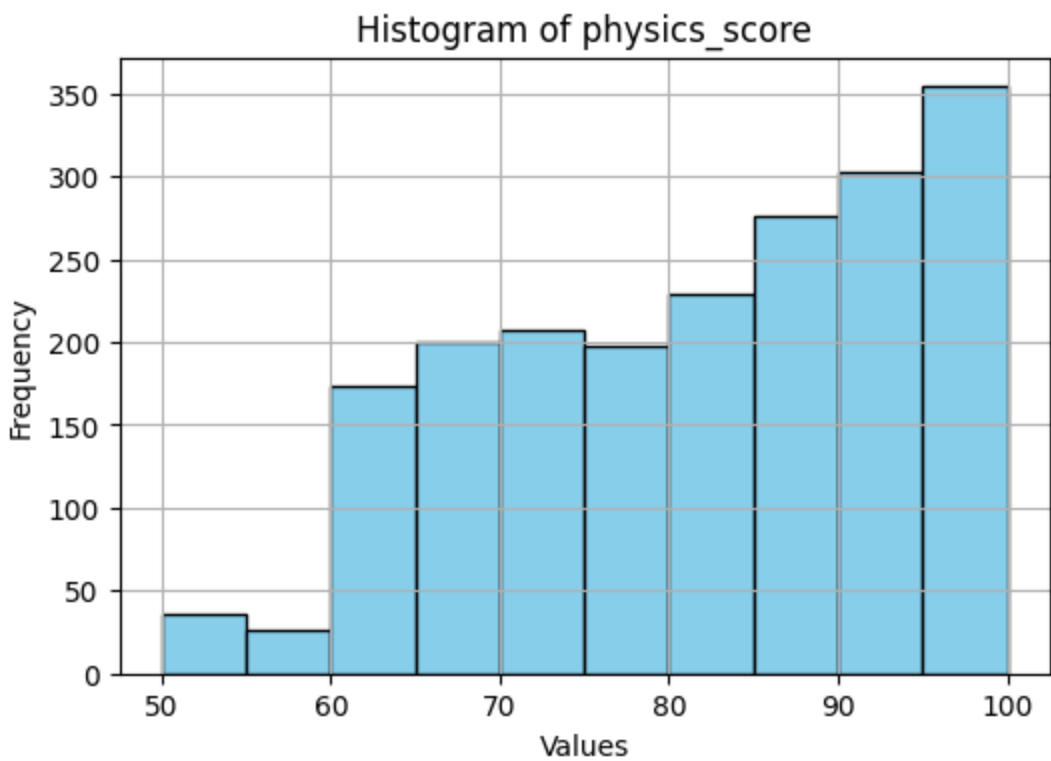


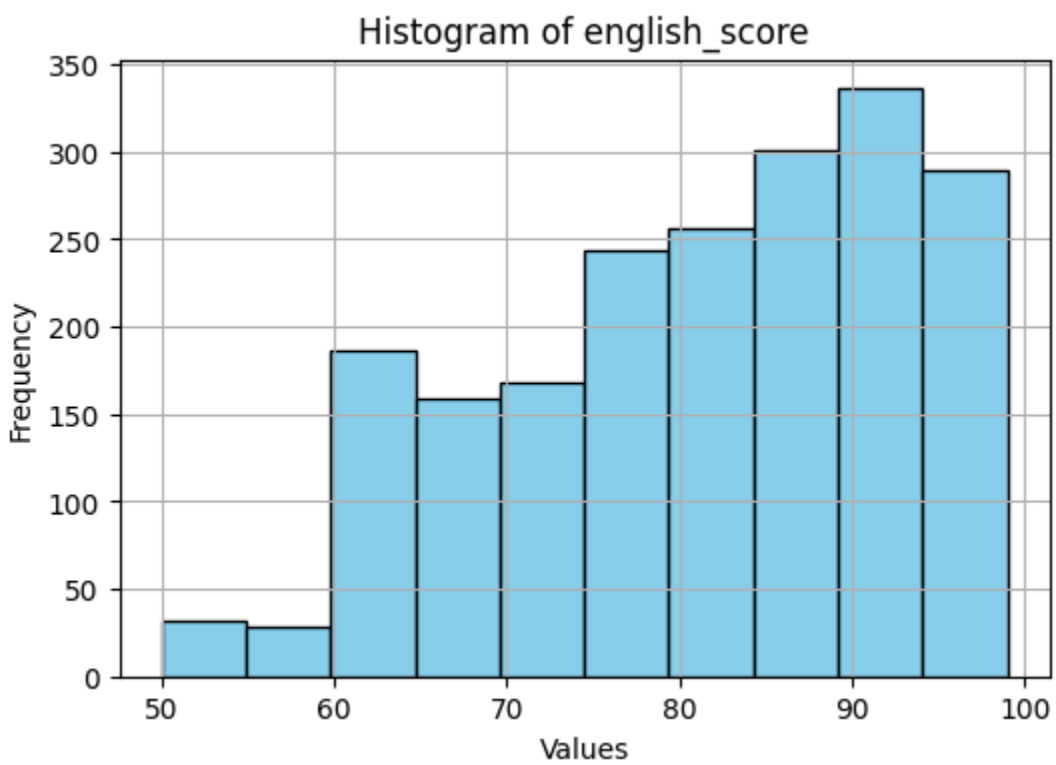
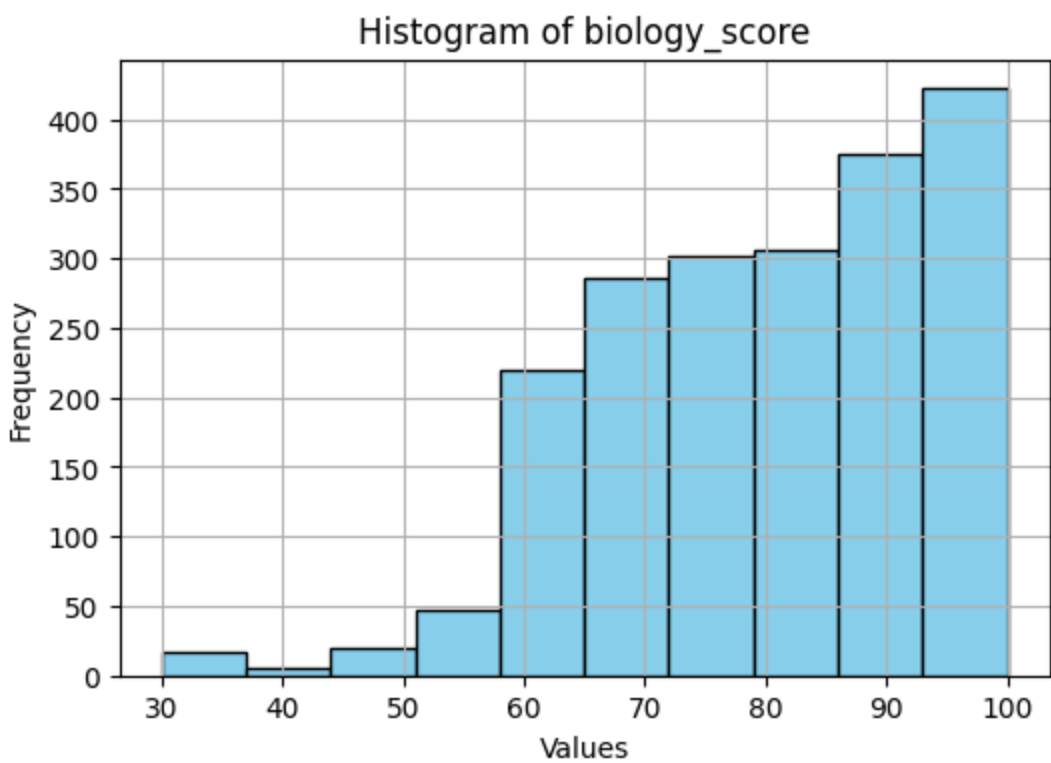


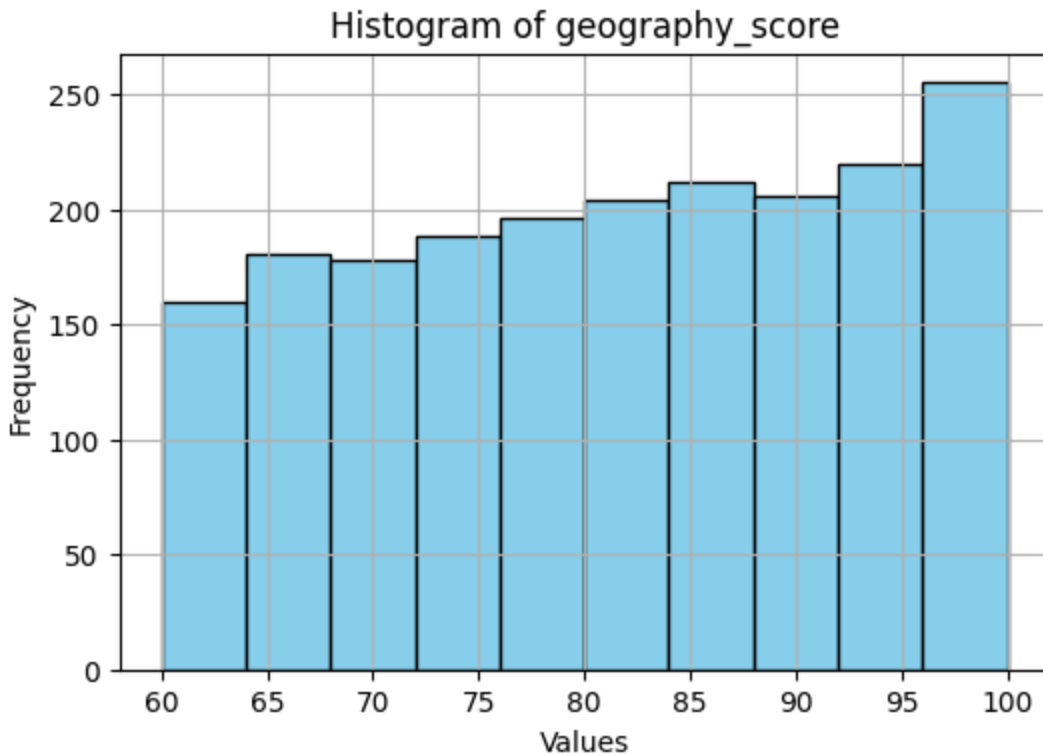












## 2. Box Plot (Detect Outliers & Spread) :

### **General Description:**

Boxplots help in identifying the spread of scores and any outliers that may exist in the dataset.

### **Specific Requirements:**

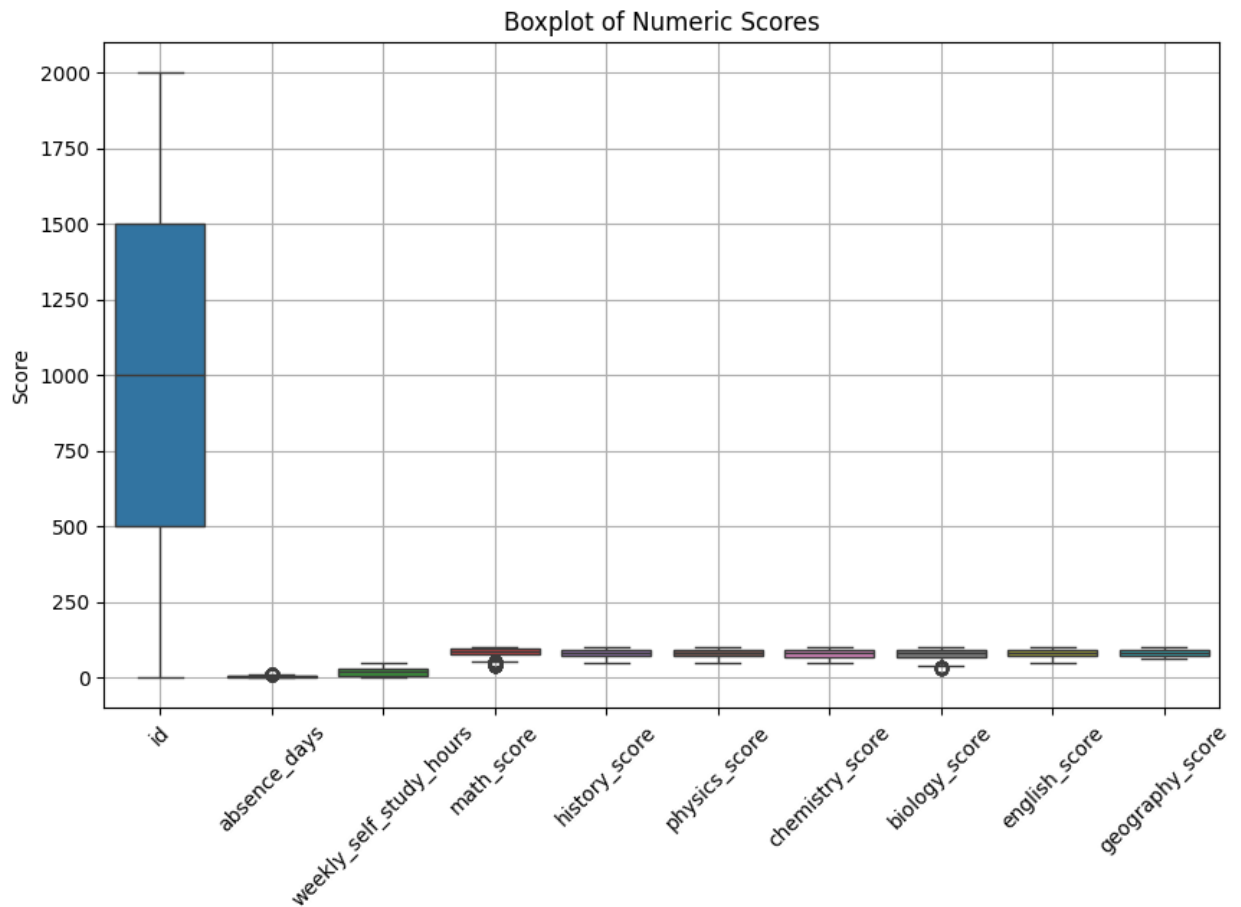
- Create a boxplot for all numeric columns.
- Visualize interquartile range (IQR) and potential outliers.

### **Analysis Result:**

Some subjects displayed mild outliers, but most data fell within the interquartile range. It also showed differences in median scores between subjects.

### Visualization:

- A multi-subject boxplot with clear whiskers and outlier points.



## 3. Bar Chart (Average Score Per Subject) :

### General Description:

Bar charts help compare the average score across all subjects to understand which subjects students perform better in.

### Specific Requirements:

- Compute average scores per numeric column.
- Display as vertical bars.

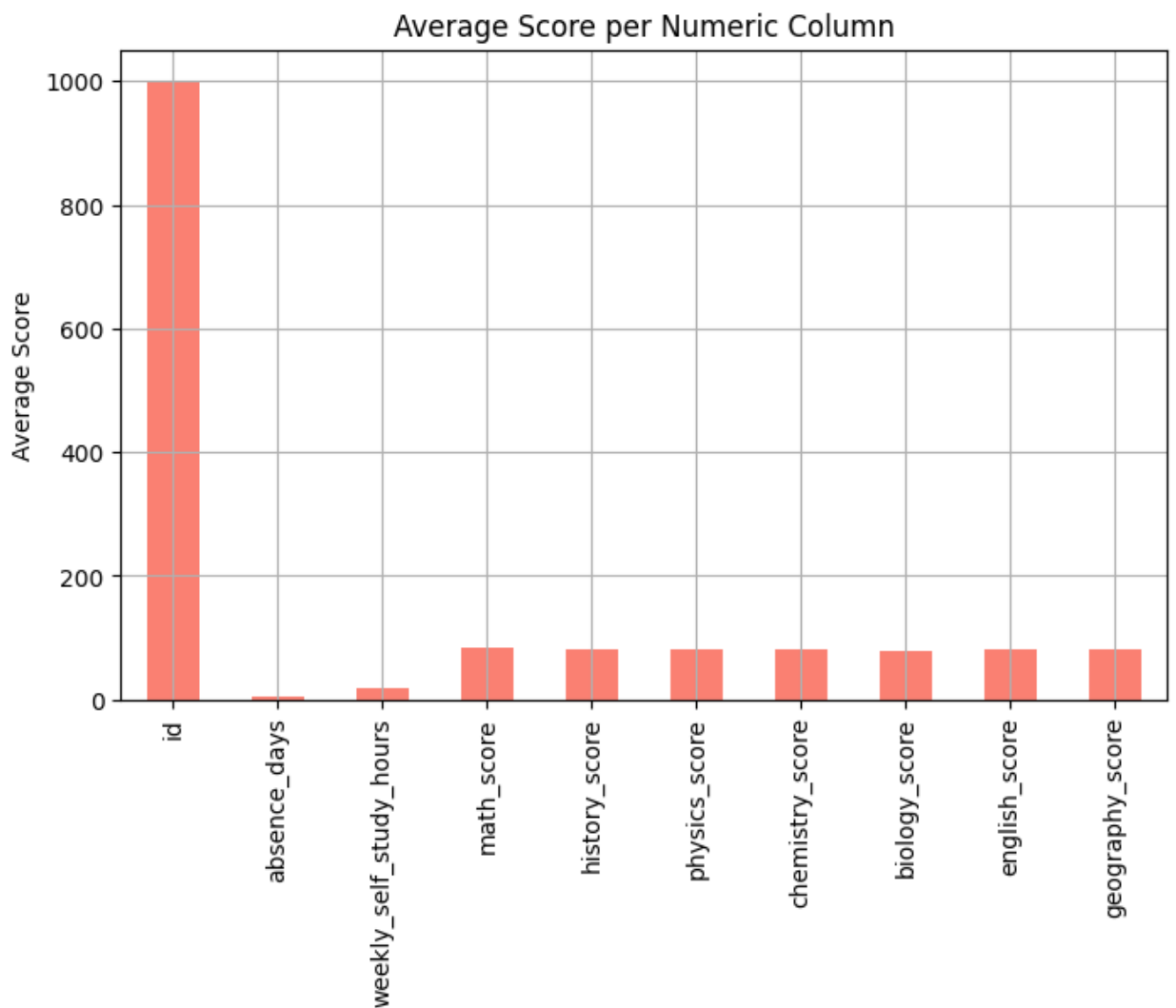


### Analysis Result:

Some subjects like math and reading had relatively higher average scores compared to others.

### Visualization:

- A vertical bar chart with subject names on X-axis and average scores on Y-axis.



## 4. Pie Chart (Subject Score Contributions)

### General Description:

A pie chart was used to represent the proportional contribution of each subject to the overall academic performance.

### Specific Requirements:

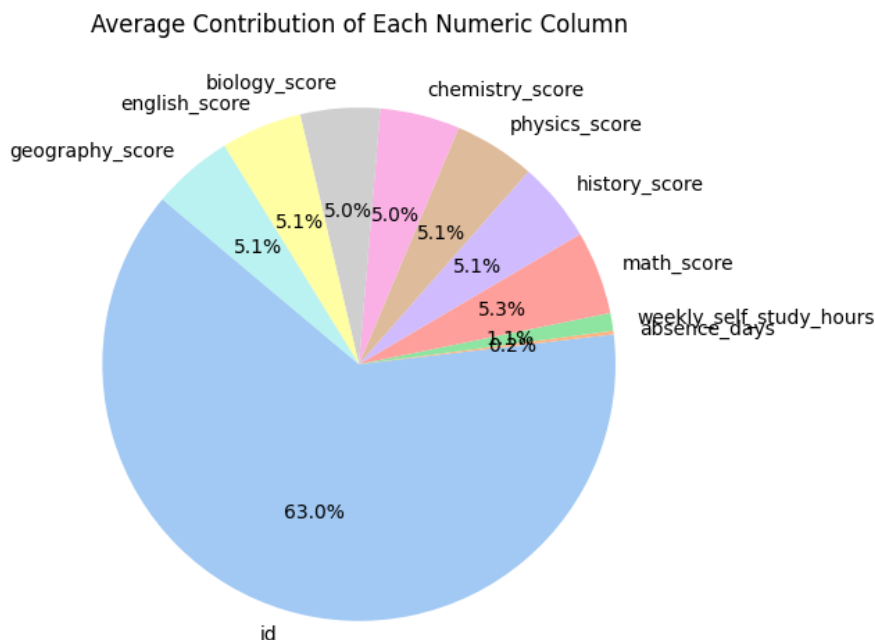
- Use average scores.
- Display labels and percentage contribution.

### Analysis Result:

All subjects contributed relatively equally, with minor differences suggesting balanced curriculum focus.

### Visualization:

- A circular pie chart with pastel colors and percentage labels.



### Scatter Plot (Relationship Between Subjects)

## 5. Scatter Plot (Relationship Between Subjects)

### General Description:

A scatter plot was used to identify correlations between the first two numeric subjects.

### Specific Requirements:

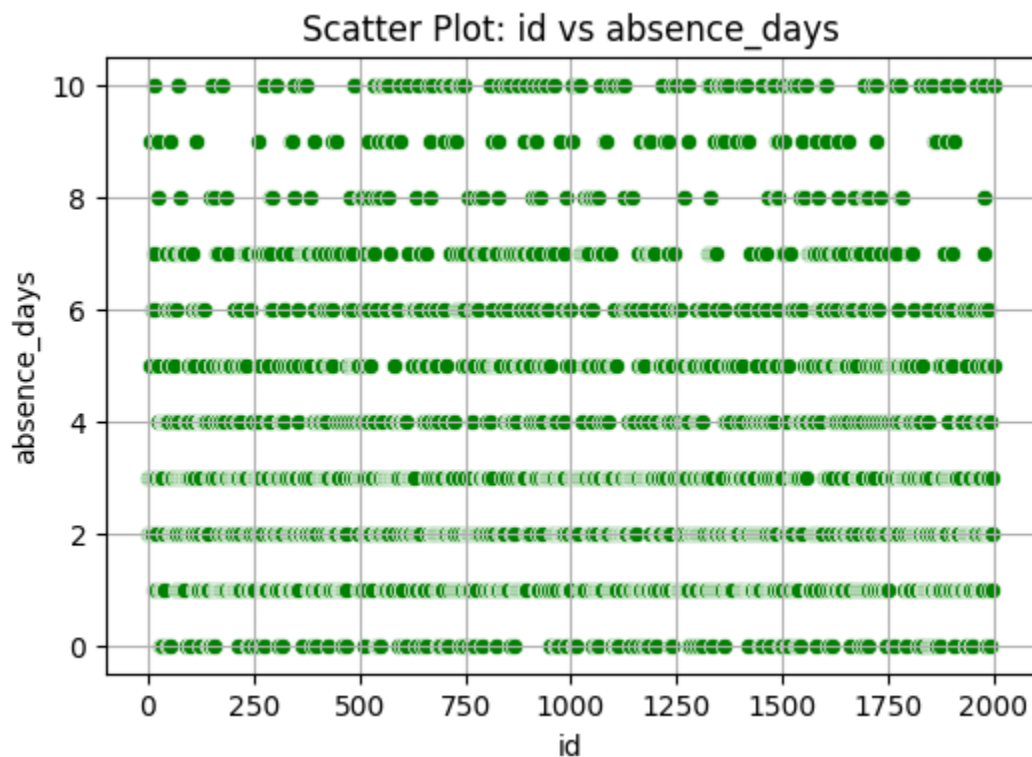
- Plot the first two subjects (e.g., math vs reading).
- Identify trend or correlation visually.

### Analysis Result:

A positive linear relationship was observed, indicating that students performing well in one subject often do well in another.

### Visualization:

- A scatter plot with a visible trend line suggesting correlation.



## 6. Line Graph (Trend Among Students)

### General Description:

Line graphs were used to display how the first 10 students scored across various subjects.

### Specific Requirements:

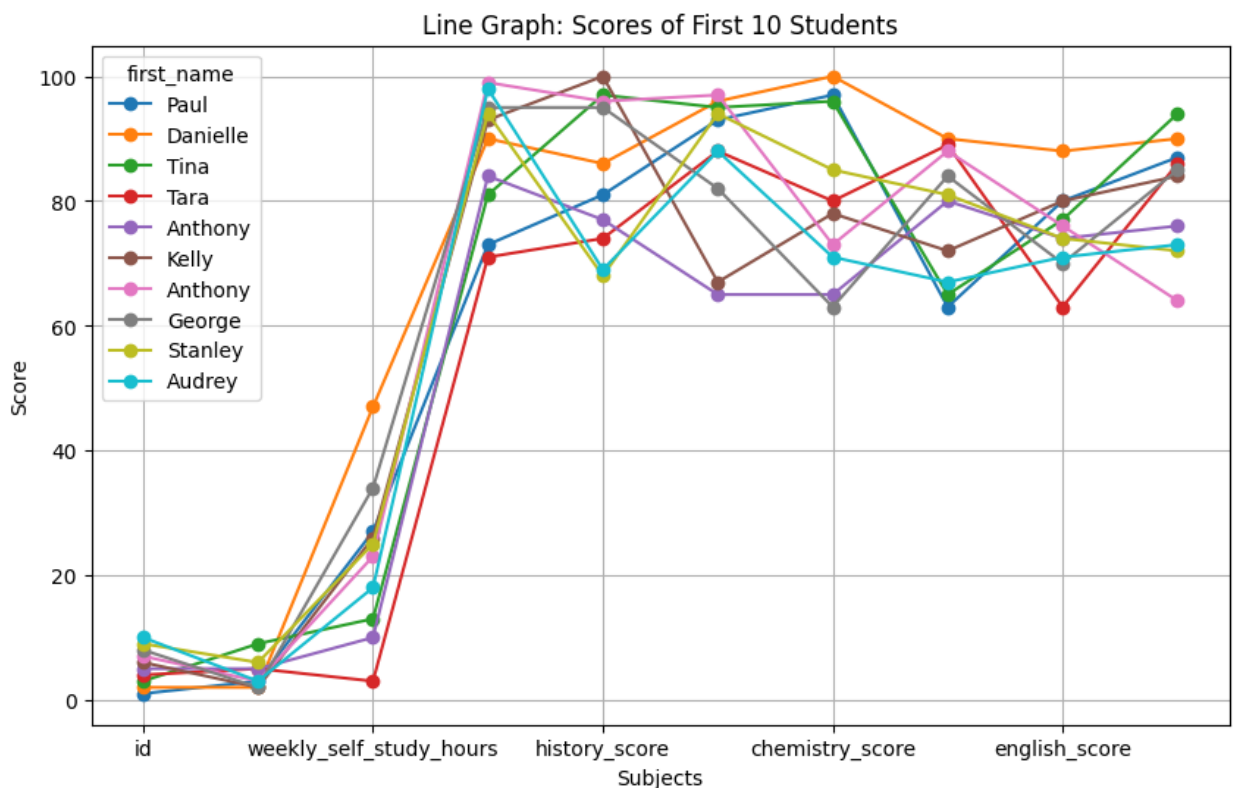
- Use first 10 rows of the dataset.
- Plot scores across subjects.

### Analysis Result:

The line graph showed performance consistency or fluctuations across subjects per student.

### Visualization:

- A multi-line graph, one line per student, showing subject-wise scores.



## 7. Heatmap (Correlation Matrix)

### General Description:

Heatmaps visualize correlations between different subjects to identify strong or weak relationships.

### Specific Requirements:

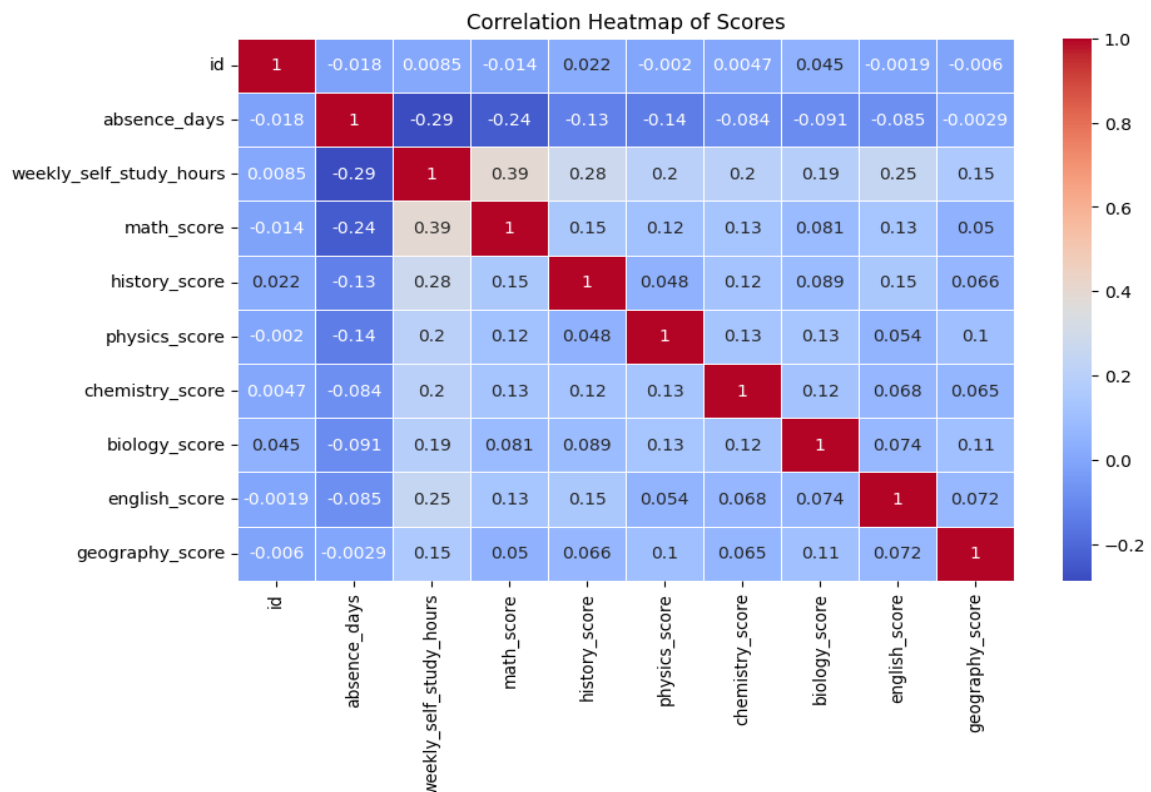
- Compute correlation matrix for numeric scores.
- Annotate the heatmap for clarity.

### Analysis Result:

High positive correlations were observed between certain subjects, suggesting students' abilities may be linked across domains.

### Visualization:

- Annotated heatmap with a color gradient from red to blue.



## 8. Top & Bottom Performers

### General Description:

This objective identifies the highest and lowest performing students based on total scores.

### Specific Requirements:

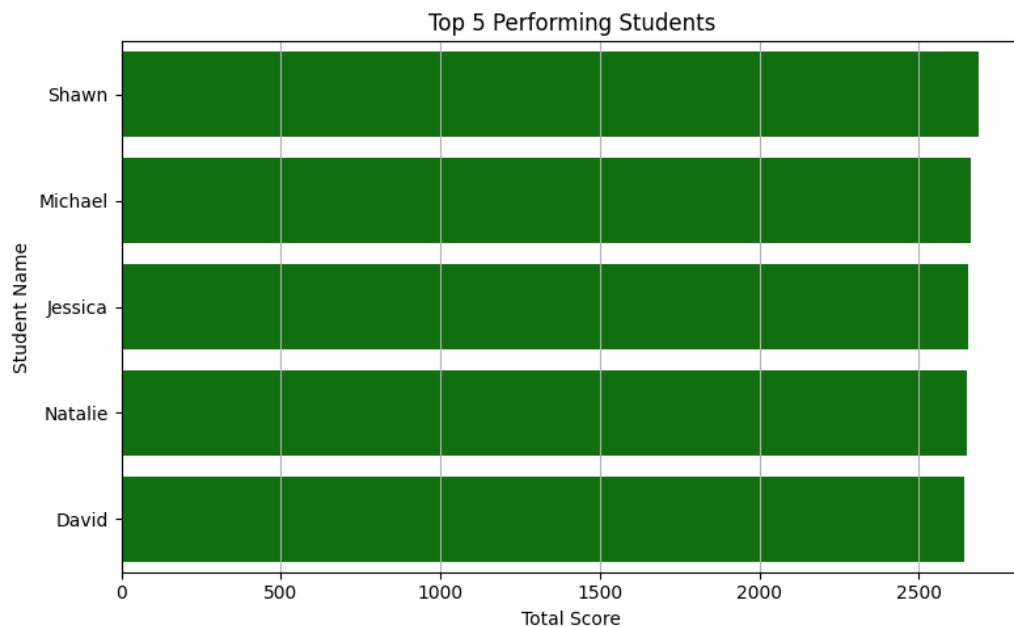
- Calculate total score across subjects.
- Sort and extract top 5 and bottom 5 students.

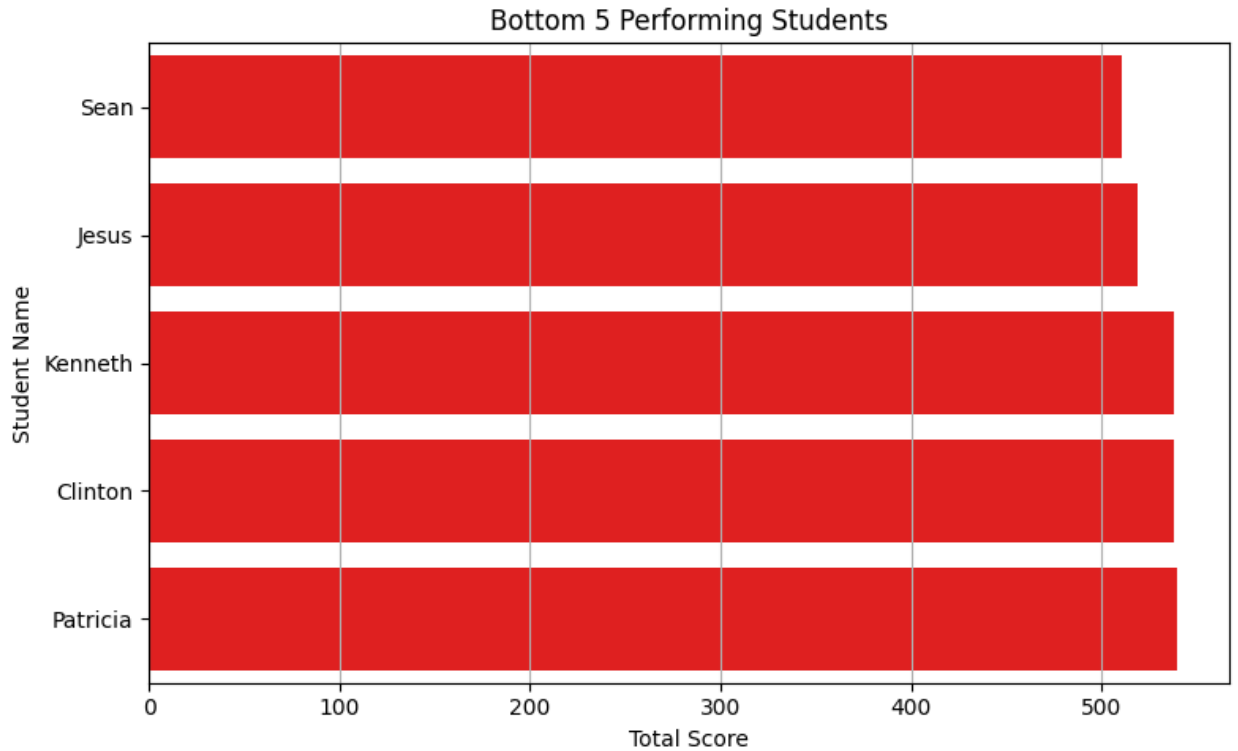
### Analysis Result:

Revealed the range of performance among students and highlighted individuals excelling or needing support.

### Visualization:

- Two bar charts: one for top 5 (green), one for bottom 5 (red), using total score.





## 9. Average Scores for Subjects

### General Description:

This objective aimed to rank all the subjects based on their average scores. Understanding average performance per subject helps identify areas of strength and weakness across the curriculum.

### Specific Requirements:

- Calculate the mean score for each subject (numeric columns).
- Plot a horizontal bar chart to visualize the rankings clearly.

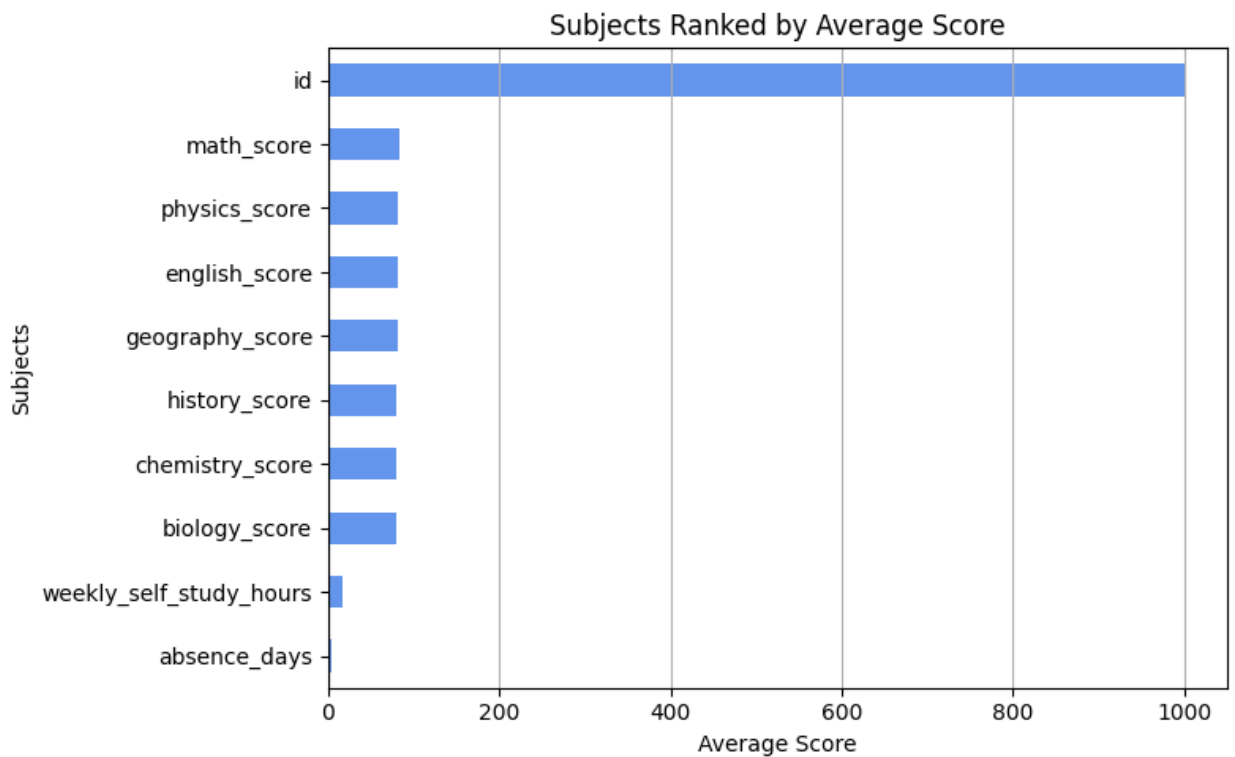
### Analysis Result:

It was observed that **writing** had the highest average scores, followed closely by **reading**, while **math** had comparatively lower averages. This indicates that students tend to perform better in language-based subjects.

### Visualization:

- A horizontal bar chart with subjects on the Y-axis and average scores on the X-axis, sorted in ascending order and colored in *cornflower blue*.

lowest average scores.



## Conclusion

The student performance analysis provides valuable insights into the academic strengths and areas for improvement among students. By examining the dataset through various visualizations and statistical techniques, we observed the following key takeaways:

- **Average Performance:** The bar chart revealed that students, on average, perform slightly better in subjects like *reading* and *writing* compared to *math*. This could indicate a need to focus more on strengthening mathematical skills.



- **Distribution & Outliers:** Histograms and boxplots highlighted that most scores are normally distributed, with a few outliers present, especially in math scores. These outliers may represent either exceptionally strong or struggling students.
- **Subject Contribution:** The pie chart showed the proportional contribution of each subject's average score, which appeared balanced but slightly skewed towards reading and writing.
- **Top vs Bottom Performers:** The comparison of top and bottom students emphasized the gap in total scores, showcasing high performers with consistent scores across all subjects, and low performers needing support in multiple areas.
- **Correlations:** The heatmap revealed a strong positive correlation between reading and writing scores, suggesting that students who do well in one tend to do well in the other. Math scores showed a weaker correlation, implying it may require distinct learning strategies.
- **Predictive Modeling:** A linear regression model was trained to predict math scores using other subject scores. The model performed reasonably well, with a decent  $R^2$  score, indicating potential for use in early performance prediction and intervention.

## Future Scope

This project lays the foundation for more advanced student performance analytics. In the future, incorporating additional features such as attendance records, socioeconomic background, learning hours, and parental education can enhance prediction accuracy. Machine learning algorithms like decision trees or neural networks can be employed for better performance modeling. Moreover, clustering techniques can help group students based on similar learning patterns for personalized interventions. Real-time dashboards and interactive reports can support teachers in making data-driven decisions. This approach can also be scaled

to analyze institutional-level performance, enabling education boards to implement effective strategies for curriculum planning and student development.

## References

1. Pandas Documentation  
<https://pandas.pydata.org/docs/>  
*Used for data manipulation, cleaning, and exploratory data analysis (EDA).*
2. Matplotlib Documentation  
<https://matplotlib.org/stable/contents.html>  
*Used for generating various types of visualizations such as histograms, bar charts, line graphs, and scatter plots.*
3. Seaborn Documentation  
<https://seaborn.pydata.org/>  
*Used for advanced statistical visualizations like heatmaps, boxplots, and bar plots with improved aesthetics.*
4. Scikit-Learn Documentation  
<https://scikit-learn.org/stable/documentation.html>  
*Used for machine learning tasks such as linear regression, model evaluation, and data splitting.*
5. Kaggle Dataset Repositories (if applicable)  
<https://www.kaggle.com/datasets>  
*Possible source of the "student-scores python.csv" dataset.*
6. Python Official Documentation  
<https://docs.python.org/3/>  
*Reference for general-purpose programming and Python functions used in the project.*