Week 3: Exploratory Data Analysis (EDA)

Dataset Overview

There are 2,000 records in the dataset of senior high school students. It covers age, gender and test scores from Math, History, Physics, Chemistry, Biology, English, and Geography subjects. Part-time job, days off from school, involvement in extracurriculars, regular study time, and what a student wants to do after education are considered here.

Data Preprocessing

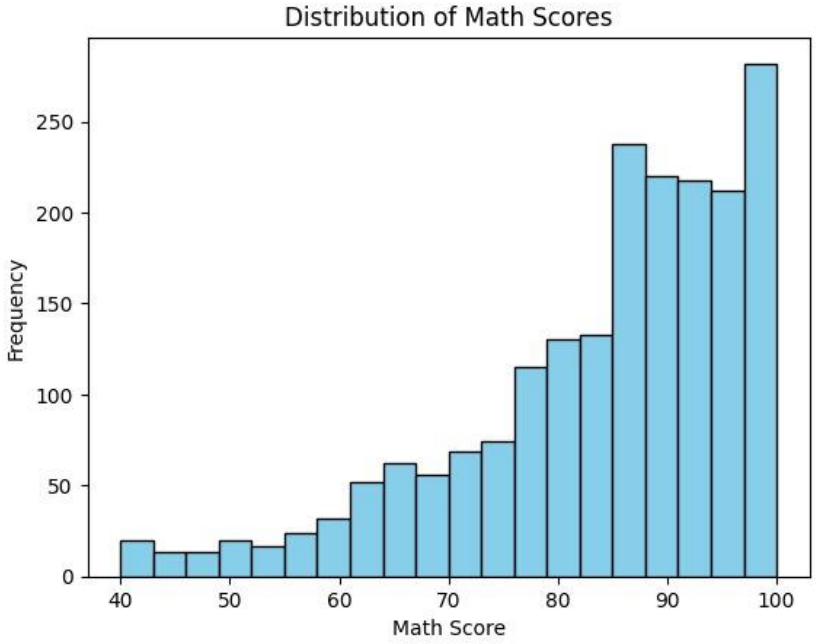
- Missing Values: The presence of missing values was checked in the dataset; none were found, so no imputation was required.
- Duplicates: No records exist that match two or more people.
- Outliers: Outliers were found in absence_days and weekly_self_study_hours by plotting box plots. A larger number of absence days (over 30) was noticed, but the research kept them as they represent the real habits of students.

Descriptive Statistics

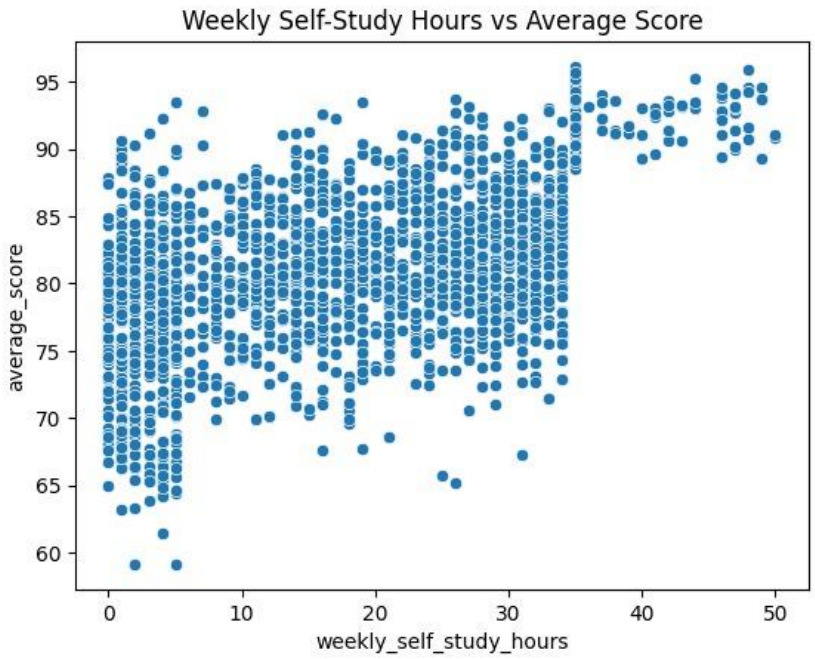
- Scores on subjects were mostly in the area of 65-75 out of 100.
- The number of absences varied from 0 to more than 40, and the average was about 5 days. People spent about 10 hours a week studying on average, but the amount of study time was highly variable.
- Almost 35% of students indicated that they have part-time jobs.
- Very few students did not take part in extra-curricular opportunities (only 30%).

Diagrams and Important Patterns

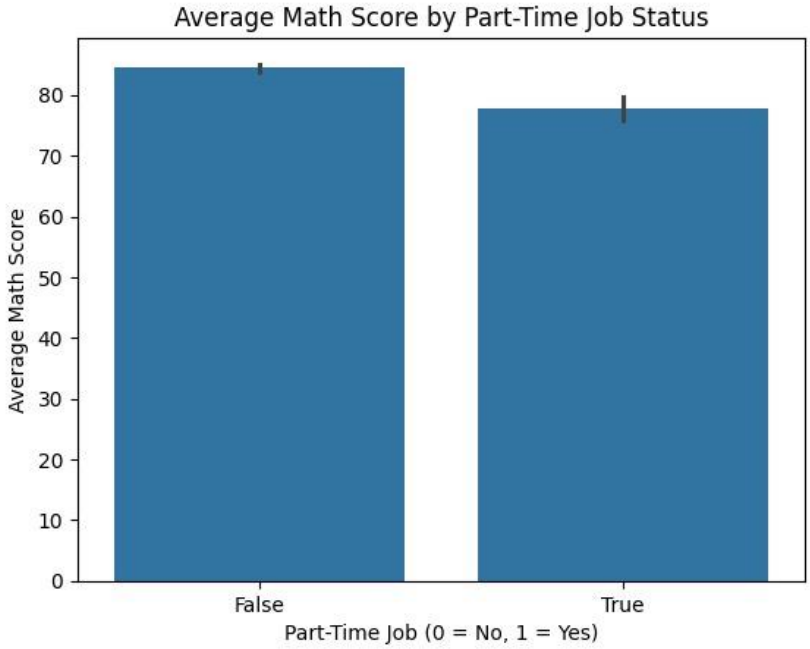
- Histogram: The histogram for math scores shows that the distribution is roughly normal and a bit skewed towards high scores, which suggests most students did moderately well.



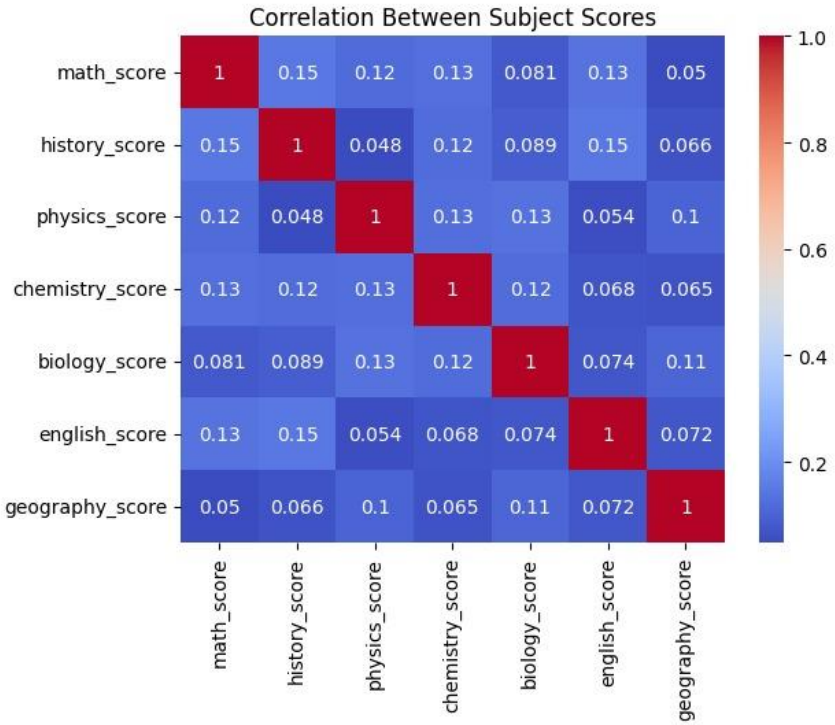
- Scatter Plot: Students who miss more days have lower math scores, which means missing school negatively influences academic achievement.



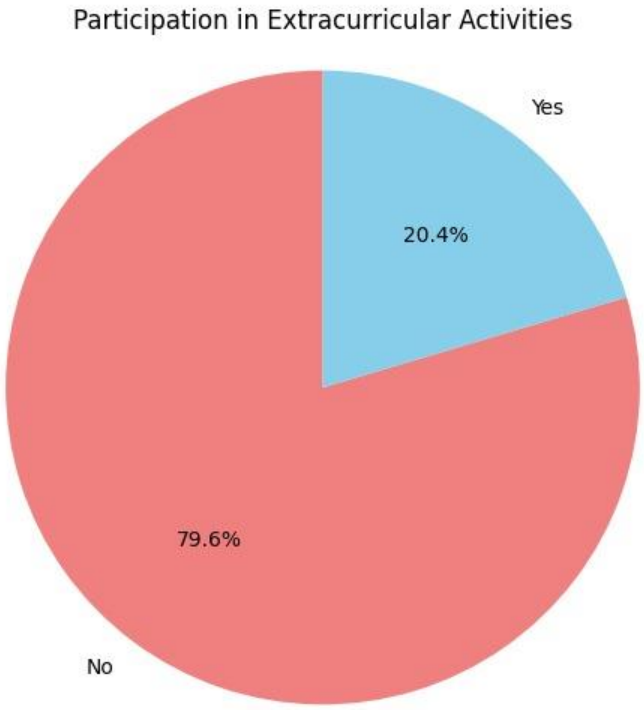
- Bar Plot: The amount of time spent studying each week was connected to better overall scores.



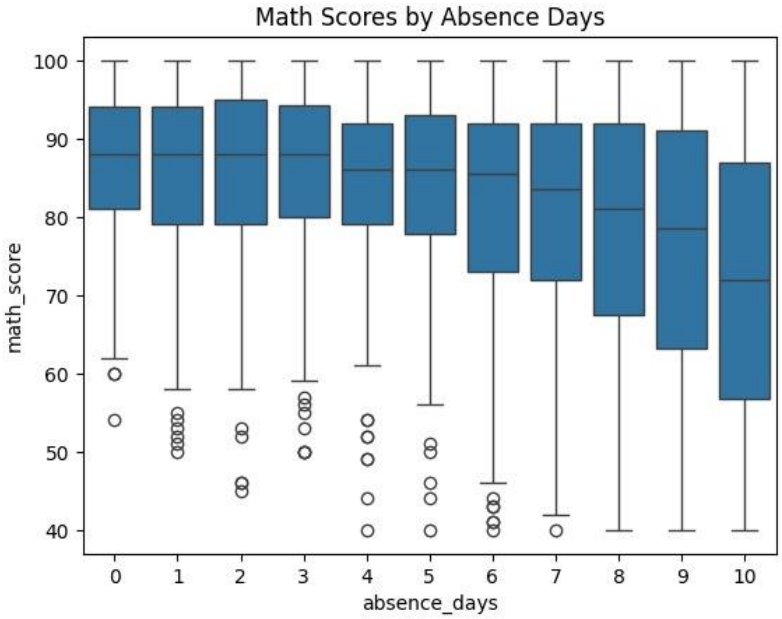
- Heatmap (Matrix Showing Correlation): Shows relationships between different numbers such as scores in several subjects.



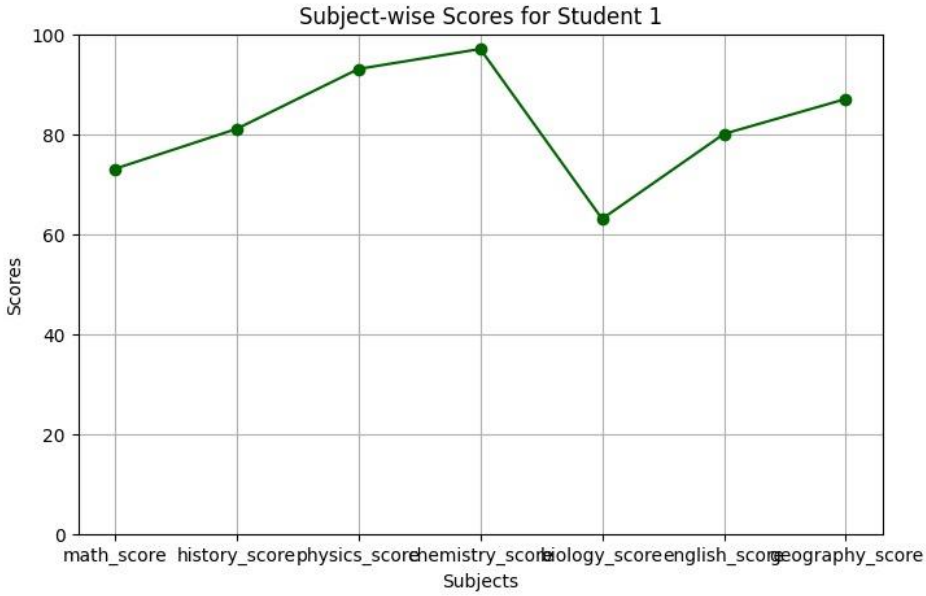
- Pie Chart: Good for illustrating comparisons, for example showing how many students join extracurricular activities.



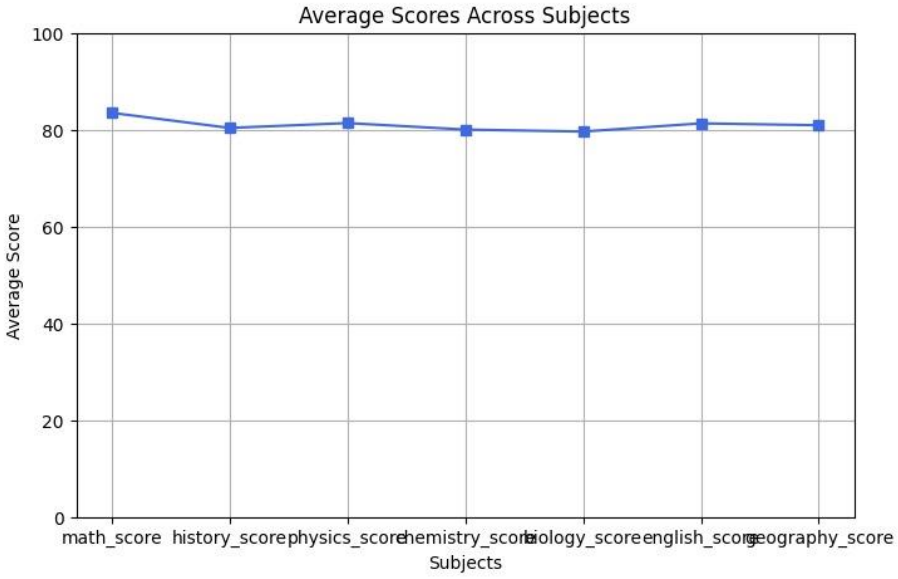
- **Box Plot:** Missing school more often seems to hurt math scores, proving that being absent at school can lower student achievement.



- **Line Plot / Time Series Plot:** If information is given over time (for example, scores across quarters), this is very helpful in tracking progress.



for single student



for all students

Working with Data

- Participation in part-time work reduced students' grades slightly in school which might be because their engagement in study time is lower.
- Being involved in extracurriculars did not directly affect academic performance, yet it probably helps students grow in other ways.
- Missing classes is strongly linked to having bad grades.
- Spending extra self-study time on a regular basis usually improves students' academic results.

Conclusion

Reviewing the data suggests that both attending class and regular study routines impact student achievement a lot. There is a strong link between better scores and missing less school and studying more on your own. The insights will help decide which features and methods to use in further analysis.