Report

- Drishti Garg(23114025)

This report presents a detailed analysis of a dataset that explores various factors influencing student performance. The dataset comprises multiple attributes that span academic, social, and environmental factors.

Dataset Overview

The dataset includes the following variables:

- **School_Type:** Public or Private
- **Peer_Influence:** Positive, Neutral, Negative
- **Physical_Activity:** Weekly hours spent on physical activities
- **Learning_Disabilities:** Yes or No
- **Parental_Education_Level:** High School, College, Postgraduate
- **Distance_from_Home:** Near, Moderate, Far
- **Gender:** Male or Female
- **Exam_Score:** Final exam score (Dependent variable)
- **Hours_Studied:** Weekly study hours
- **Attendance:** Percentage of attended classes
- **Parental_Involvement:** Low, Medium, High
- **Access_to_Resources:** Low, Medium, High
- **Extracurricular_Activities:** Participation level (None, Few, Many)

Data Cleaning and Preparation

To ensure the dataset's reliability and accuracy, several data cleaning procedures were undertaken:

1. Handling Missing Values

- **Numerical Columns:**The median was used to impute missing values in numerical columns such as Hours_Studied and Exam_Score. This approach reduced the possibility of a skewed distribution of data.

- **Categorical Columns:** The mode, or most often occurring value, was used to fill in the missing values in categorical variables such as Parental Involvement and Peer Influence..

2. Removing Duplicates

- To maintain data integrity and remove redundancy, duplicate records were found and deleted.

3. Outlier Detection and Treatment

- **Box Plot Analysis:** Box plots were used to find outliers in Exam_Score and Hours_Studied. Capping was used to lessen the disproportionate impact of extreme outliers on the analysis.

4. Standardization of Categorical Variables

- 'High school' and 'highschool' were examples of categorical labels with inconsistent formatting that were standardised to enhance data consistency and quality.

Exploratory Data Analysis (EDA)

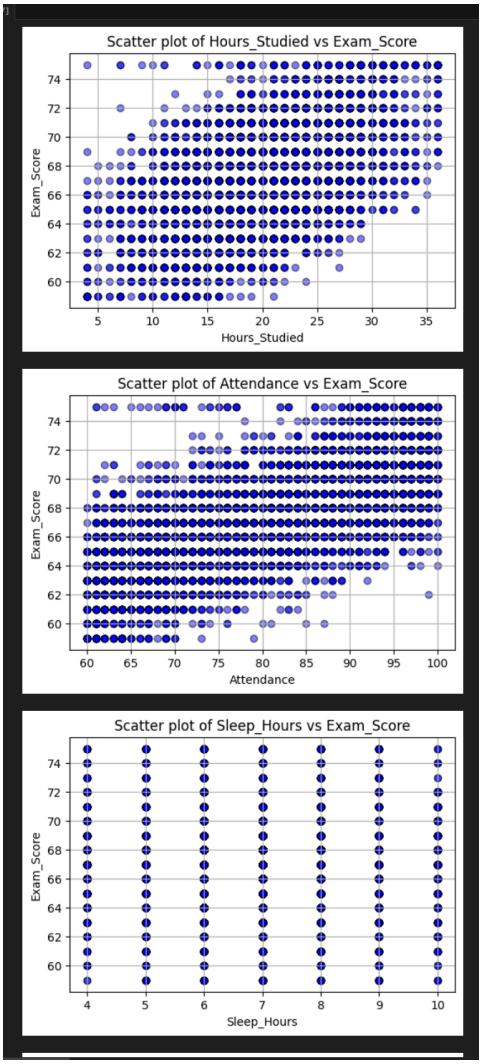
Univariate Analysis

- **Exam Scores Distribution:**
    - A right-skewed trend in the exam score distribution suggested that there was a greater concentration of pupils with lower marks. The fact that the median score was far lower than the mean indicates that while most students scored below average, some did extraordinarily well.
- **Attendance Rates:**
    - The analysis highlighted the importance of consistent participation in class, revealing that a large portion of students maintained attendance rates exceeding 80%.
- **Hours Studied:**

    - Most students reported studying for between five and ten hours each week. Moderate study durations were common throughout the dataset, as very few students indicated either extremely high or low study hours.

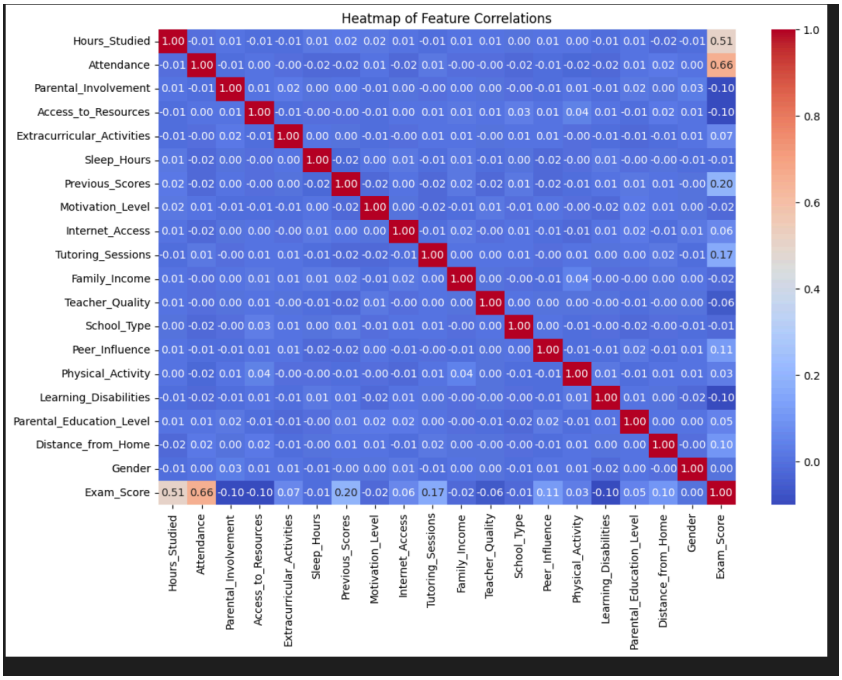
Bivariate Analysis

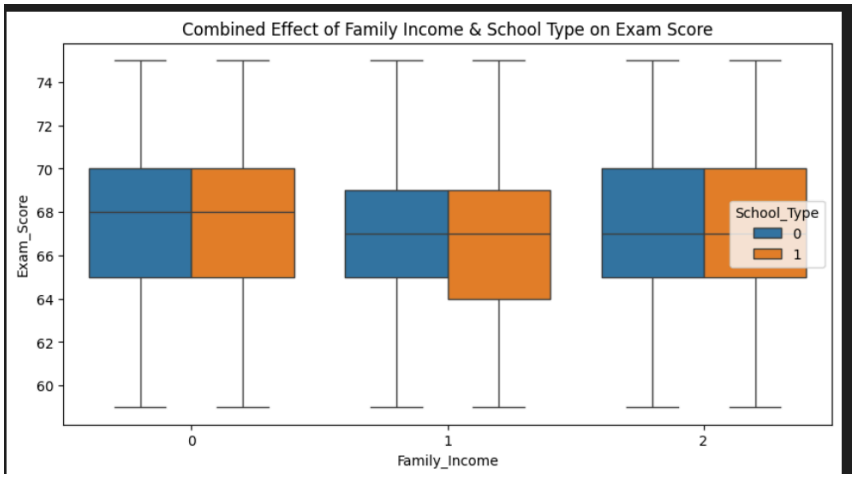
- **Effect of Hours_Studied on Exam_Score:**
    - A robust positive correlation was found between the number of hours studied and exam results. Students who committed more time to their studies consistently obtained higher scores.
- **Impact of Parental_Involvement:**
    - Students whose parents were highly involved exhibited significantly elevated scores, underscoring the impact of family support on educational success.
- **School Type Analysis:**
    - Students enrolled in private institutions consistently outperformed their peers in public schools, indicating variations in educational resources, teaching quality, or institutional policies.

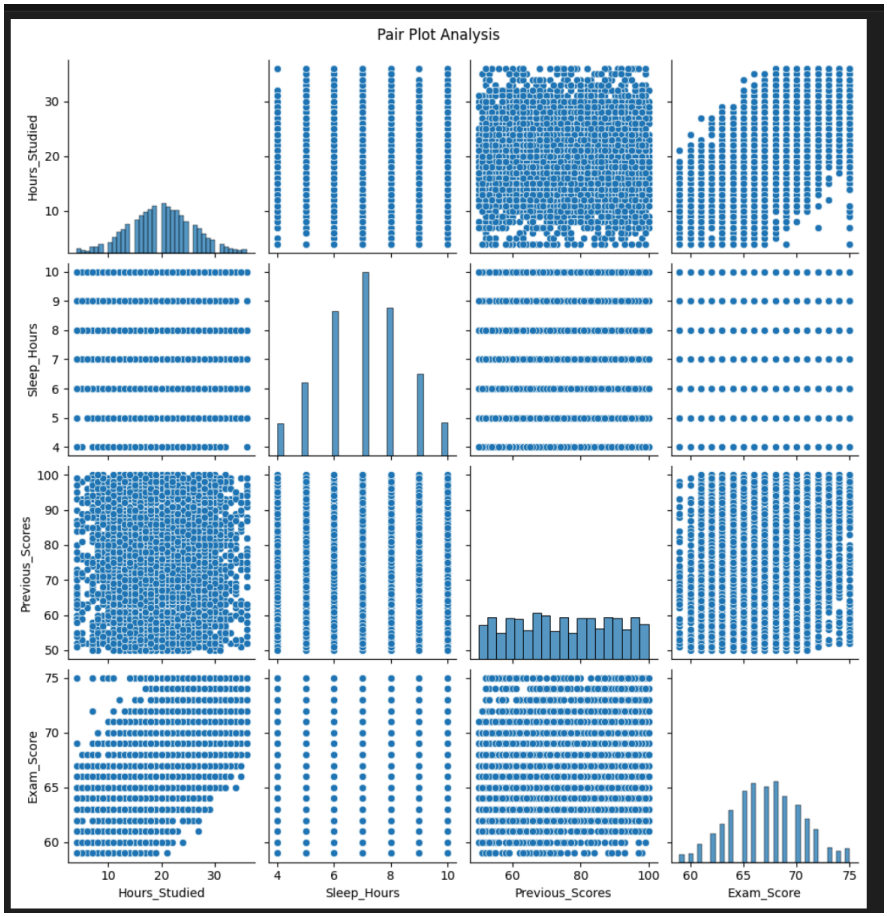
- 

Multivariate Analysis

- By analyzing combined effects, key insights emerged:
  - **Positive Peer Influence** coupled with **High Parental Involvement** resulted in significantly better academic performance.
  - Students from families with higher levels of parental education generally achieved better results, emphasizing the importance of family background in influencing academic success.
  - Students with learning disabilities experienced considerable improvements in performance when given extra educational resources.


Heatmap of Feature Correlations
Hours_Studied 1.00 -0.01 0.01 -0.01 -0.01 0.01 0.02 0.02 0.01 -0.01 0.01 0.01 0.00 0.01 0.00 -0.01 0.01 -0.02 -0.01 0.51
Attendance -0.01 1.00 -0.01 0.00 -0.00 -0.02 -0.02 0.01 -0.02 0.01 -0.00 -0.00 -0.02 -0.01 -0.02 -0.02 0.01 0.02 0.00 0.66
Parental_Involvement 0.01 -0.01 1.00 0.01 0.02 0.00 0.00 -0.01 0.00 -0.00 0.00 0.00 -0.00 -0.01 0.01 -0.01 0.02 0.00 0.03 -0.10
Access_to_Resources -0.01 0.00 0.01 1.00 -0.01 -0.00 -0.00 -0.01 0.00 0.01 0.01 0.01 0.03 0.01 0.04 0.01 -0.01 0.02 0.01 -0.10
Extracurricular_Activities -0.01 -0.00 0.02 -0.01 1.00 0.00 0.00 -0.01 -0.00 0.01 0.01 -0.00 0.01 0.01 -0.00 -0.01 -0.01 -0.01 0.01 0.07
Sleep_Hours 0.01 -0.02 0.00 -0.00 0.00 1.00 -0.02 0.00 0.01 -0.01 0.01 -0.01 0.00 -0.02 -0.00 0.01 -0.00 -0.00 -0.01 -0.01
Previous_Scores 0.02 -0.02 0.00 -0.00 0.00 -0.02 1.00 -0.02 0.00 -0.02 0.02 -0.02 0.01 -0.02 -0.01 0.01 0.01 0.01 -0.00 0.20
Motivation_Level 0.02 0.01 -0.01 -0.01 -0.01 0.00 -0.02 1.00 0.00 -0.02 -0.01 0.01 -0.01 0.00 -0.00 -0.02 0.02 0.01 0.00 -0.02
Internet_Access 0.01 -0.02 0.00 0.00 -0.00 0.01 0.00 0.00 1.00 -0.01 0.02 -0.00 0.01 -0.01 -0.01 0.01 0.02 -0.01 0.01 0.06
Tutoring_Sessions -0.01 0.01 -0.00 0.01 0.01 -0.01 -0.02 -0.02 -0.01 1.00 0.00 0.00 0.01 -0.00 0.01 0.00 0.00 0.02 -0.01 0.17
Family_Income 0.01 -0.00 0.00 0.01 0.01 0.01 0.02 -0.01 0.02 0.00 1.00 0.00 -0.00 -0.01 0.04 -0.00 -0.00 0.00 0.00 -0.02
Teacher_Quality 0.01 -0.00 0.00 0.01 -0.00 -0.01 -0.02 0.01 -0.00 0.00 0.00 1.00 0.00 0.00 0.00 -0.00 -0.01 -0.00 0.00 -0.06
School_Type 0.00 -0.02 -0.00 0.03 0.01 0.00 0.01 -0.01 0.01 0.01 -0.00 0.00 1.00 0.00 -0.01 -0.00 -0.02 -0.00 -0.01 -0.01
Peer_Influence 0.01 -0.01 -0.01 0.01 0.01 -0.02 -0.02 0.00 -0.01 -0.00 -0.01 0.00 0.00 1.00 -0.01 -0.01 0.02 -0.01 0.01 0.11
Physical_Activity 0.00 -0.02 0.01 0.04 -0.00 -0.00 -0.01 -0.00 -0.01 0.01 0.04 0.00 -0.01 -0.01 1.00 0.01 -0.01 0.01 0.01 0.03
Learning_Disabilities -0.01 -0.02 -0.01 0.01 -0.01 0.01 0.01 -0.02 0.01 0.00 -0.00 -0.00 -0.00 -0.01 0.01 1.00 0.01 0.00 -0.02 -0.10
Parental_Education_Level -0.01 0.01 0.02 -0.01 -0.01 -0.00 0.01 0.02 0.02 0.00 -0.00 -0.01 -0.02 0.02 -0.01 0.01 1.00 0.00 0.00 0.05
Distance_from_Home -0.02 0.02 0.00 0.02 -0.01 -0.00 0.01 0.01 -0.01 0.02 0.00 -0.00 -0.00 -0.01 0.01 0.00 0.00 1.00 -0.00 0.10
Gender -0.01 0.00 0.03 0.01 0.01 -0.01 -0.00 0.00 0.01 -0.01 0.00 0.00 -0.01 0.01 0.01 -0.02 0.00 -0.00 1.00 0.00
Exam_Score 0.51 0.66 -0.10 -0.10 0.07 -0.01 0.20 -0.02 0.06 0.17 -0.02 -0.06 -0.01 0.11 0.03 -0.10 0.05 0.10 0.00 1.00
Hours_Studied
Attendance
Parental_Involvement
Access_to_Resources
Extracurricular_Activities
Sleep_Hours
Previous_Scores
Motivation_Level
Internet_Access
Tutoring_Sessions
Family_Income
Teacher_Quality
School_Type
Peer_Influence
Physical_Activity
Learning_Disabilities
Parental_Education_Level
Distance_from_Home
Gender
Exam_Score
1.0
0.8
0.6
0.4
0.2
0.0





Visual Analysis and Graph Interpretations

Several visualizations were used to derive insights:

1. Pair Plot Analysis:

- **Hours_Studied vs Exam_Score:** A distinct positive correlation is apparent. Students who studied for 20-30 hours each week exhibited a notable increase in scores, with the majority achieving scores above 70.
- **Previous_Scores vs Exam_Score:** Previous academic performance showed a positive but scattered trend, indicating that past performance alone does not fully predict future scores.
- **Sleep_Hours vs Exam_Score:** No significant correlation was found between sleep patterns and exam scores, though students averaging 6-8 hours of sleep showed stable performance.

2. Box Plot Analysis (Family Income & School Type Impact):

- Students from both public (label 0) and private (label 1) schools achieved similar median scores across all income groups.
- Median exam scores for students from low-income families were approximately **68-70**, while those from higher-income groups had similar scores, indicating minimal impact of family income alone on performance.
- Variability in exam scores remained consistent across income levels and school types, showing that factors like study habits and attendance played a greater role in determining performance.

Key Findings and Insights

1. **Positive Peer Influence** combined with **High Parental Involvement** resulted in consistently higher academic performance.
2. **Learning Disabilities:** Students with learning disabilities attained better results when they had enhanced access to educational materials.
3. **Parental Education Level:** Students from families where parents achieved higher educational qualifications performed markedly better.
4. **Extracurricular Activities:** Well-balanced engagement in extracurricular activities was positively correlated with improved academic outcomes, while excessive participation exhibited diminishing returns.
5. **Attendance and Exam Scores:** Students with regular attendance (above 90%) recorded significantly higher scores.

Conclusion and Recommendations

- **Encourage Positive Peer Influence:** Schools should create peer mentorship programs and promote positive study environments to foster academic excellence.
- **Enhance Parental Involvement:** Institutions may launch workshops and interactive sessions to draw parents more actively into their child's education.
- **Resource Allocation for Special Needs:** Schools must give priority to additional learning resources for students with disabilities to guarantee equitable learning opportunities.
- **Balance in Extracurricular Activities:** Schools should direct students to sustain a balanced participation in extracurricular activities to optimize academic performance.
- **Improve Attendance Awareness:** Schools should direct students to sustain a balanced participation in extracurricular activities to optimize academic performance.
