ANALYSIS OF STUDENT SCORE

Introduction to analysis

In this project, we conducted a comprehensive data analysis of student scores to uncover key insights and patterns. The analysis began with an exploratory data analysis (EDA) phase, where we examined the dataset’s structure, identified missing values, and visualized distributions to understand trends and potential anomalies. After cleaning and preparing the data, we applied statistical and analytical techniques to extract meaningful insights. The findings from this analysis provide valuable information about student performance, potential influencing factors, and areas for improvement.

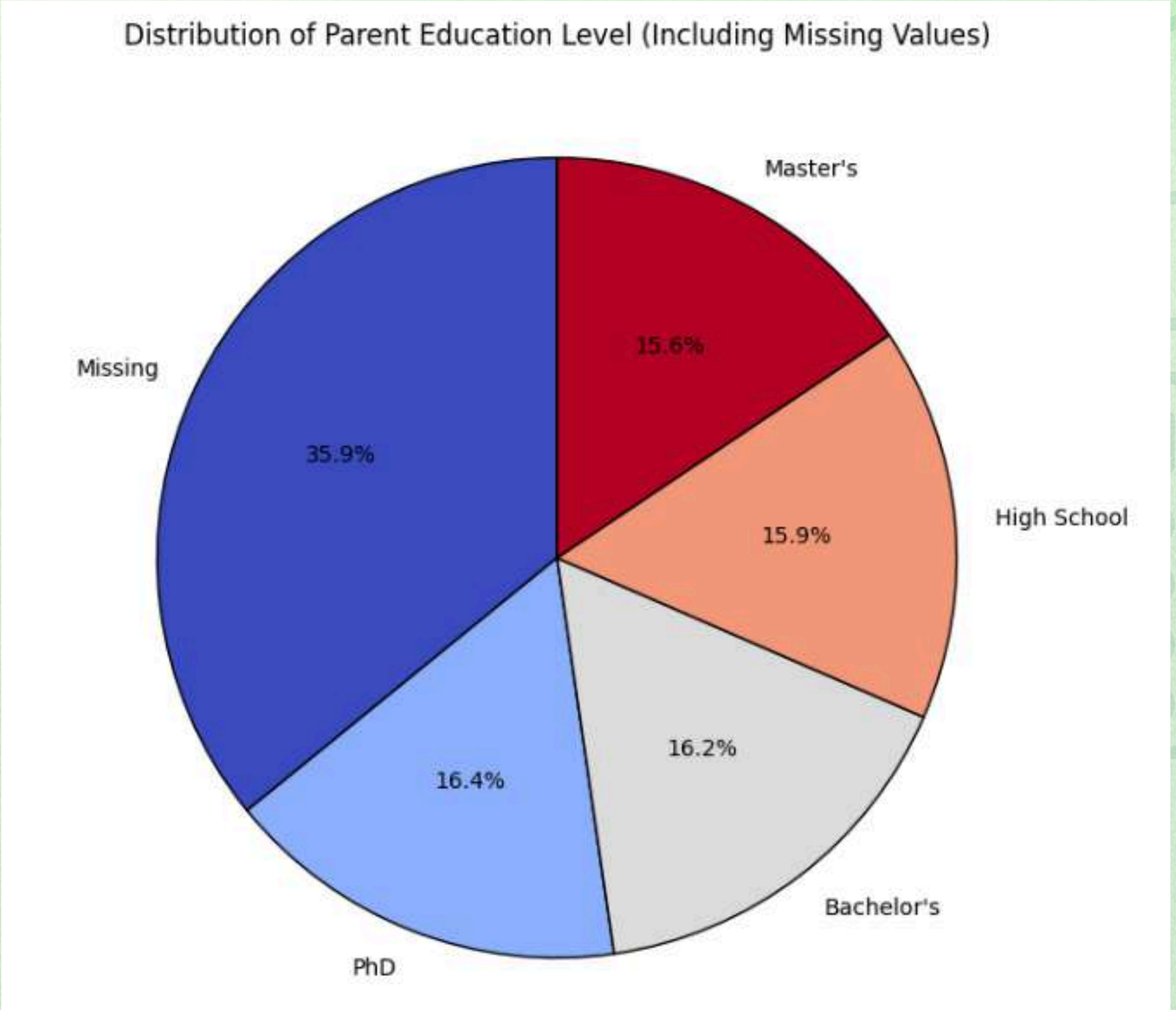
Methodology used in the analysis

The analysis began with an initial exploration of the dataset to understand its structure and characteristics. We first verified the accuracy and consistency of data types to ensure proper processing. A correlation matrix was then generated to identify relationships between different variables. Additionally, we examined the data distribution to assess its normality, helping in selecting appropriate statistical methods. Missing values and null entries were also detected and addressed to maintain data integrity and avoid potential biases in the analysis.

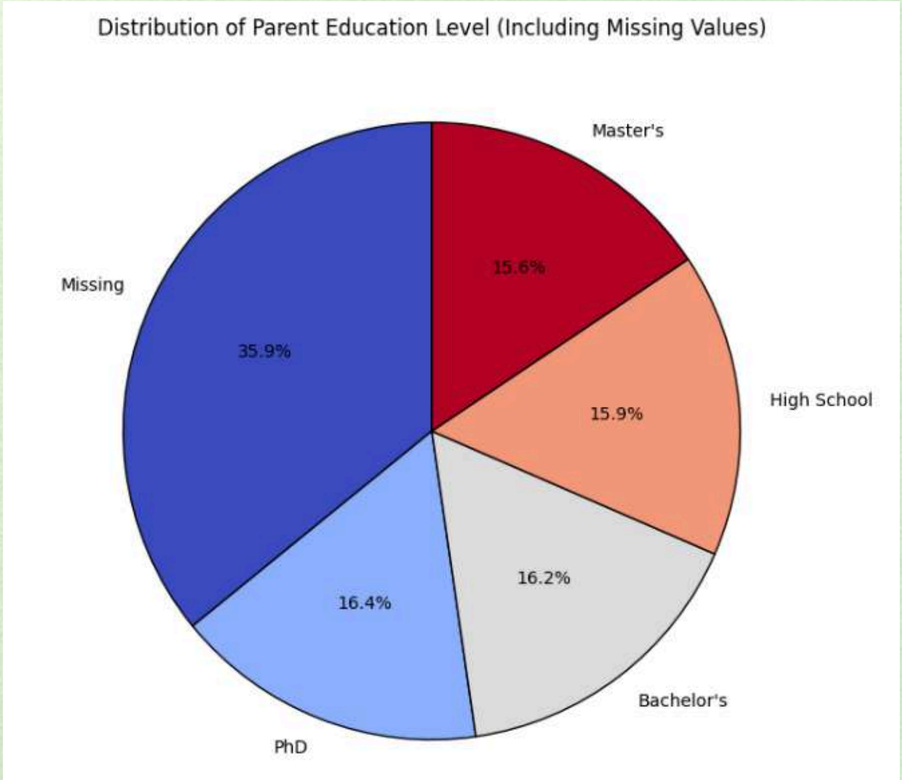
During the data exploration phase, we identified three key attributes with missing values: attendance rate, average assignment score, and parental education level. Addressing these missing values was essential to ensure the reliability of the analysis and prevent biases in the results.

Student_ID	0
First_Name	0
Last_Name	0
Email	0
Gender	0
Age	0
Department	0
Attendance (%)	516
Midterm_Score	0
Final_Score	0
Assignments_Avg	517
Quizzes_Avg	0
Participation_Score	0
Projects_Score	0
Total_Score	0
Grade	0
Study_Hours_per_Week	0
Extracurricular_Activities	0
Internet_Access_at_Home	0
Parent_Education_Level	1794
Family_Income_Level	0
Stress_Level (1-10)	0
Sleep_Hours_per_Night	0

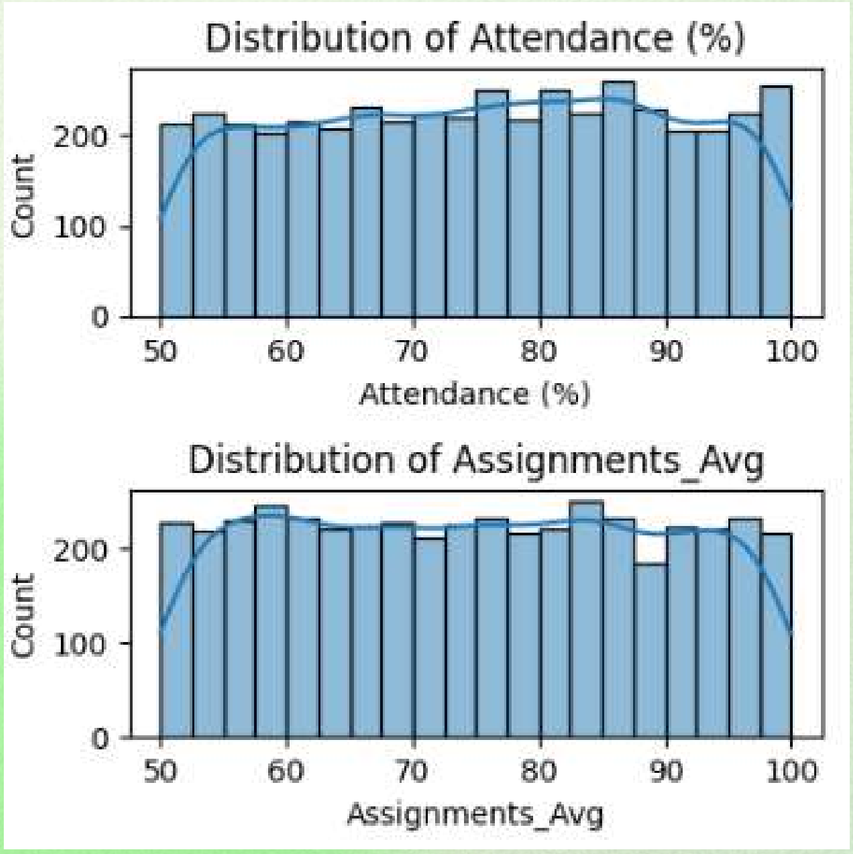
To determine the best approach for handling missing values, we first needed to explore the dataset thoroughly. For filling in the missing values in parental education level, we considered three possible methods: imputing with the most frequent value, filling based on the data's natural distribution, or using K-Nearest Neighbors (KNN) imputation. Each method has its advantages, and the choice depends on the overall data structure and distribution.

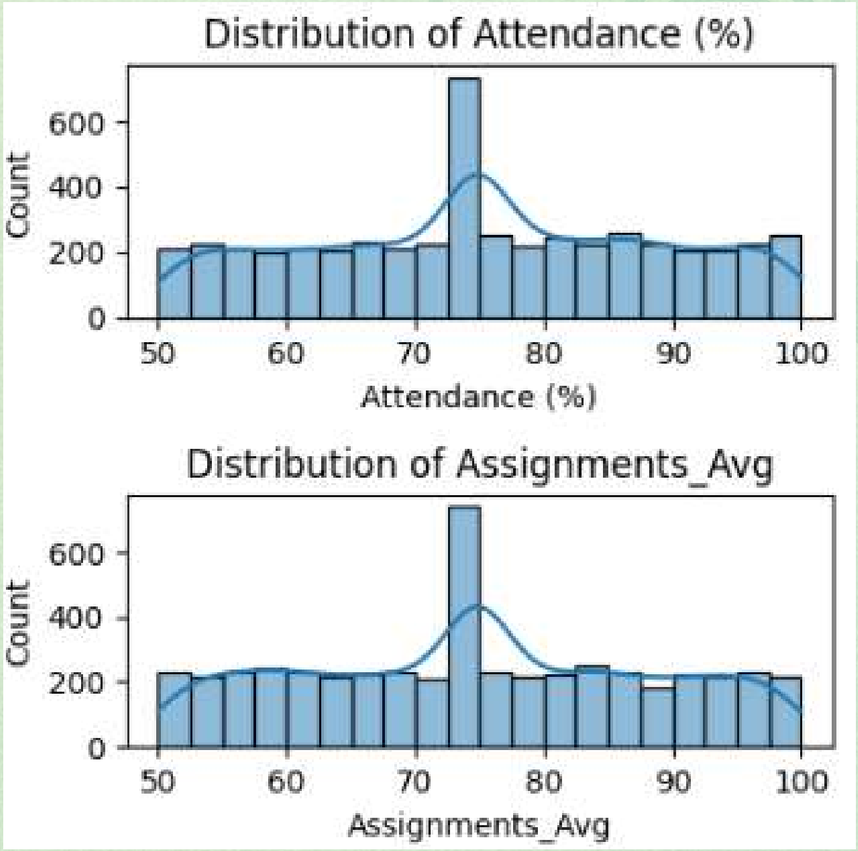


I chose to impute the missing values in parental education level based on the natural distribution of the data. This decision was made because the category percentages were very close to each other. Using the most frequent value would have caused the highest category to nearly double in frequency due to just a 1% difference, which would not be a logical representation of the data.
Additionally, KNN imputation was not suitable in this case due to the weak correlation between features, as we will see later in the analysis. This lack of strong relationships made it difficult for KNN to accurately predict the missing values.

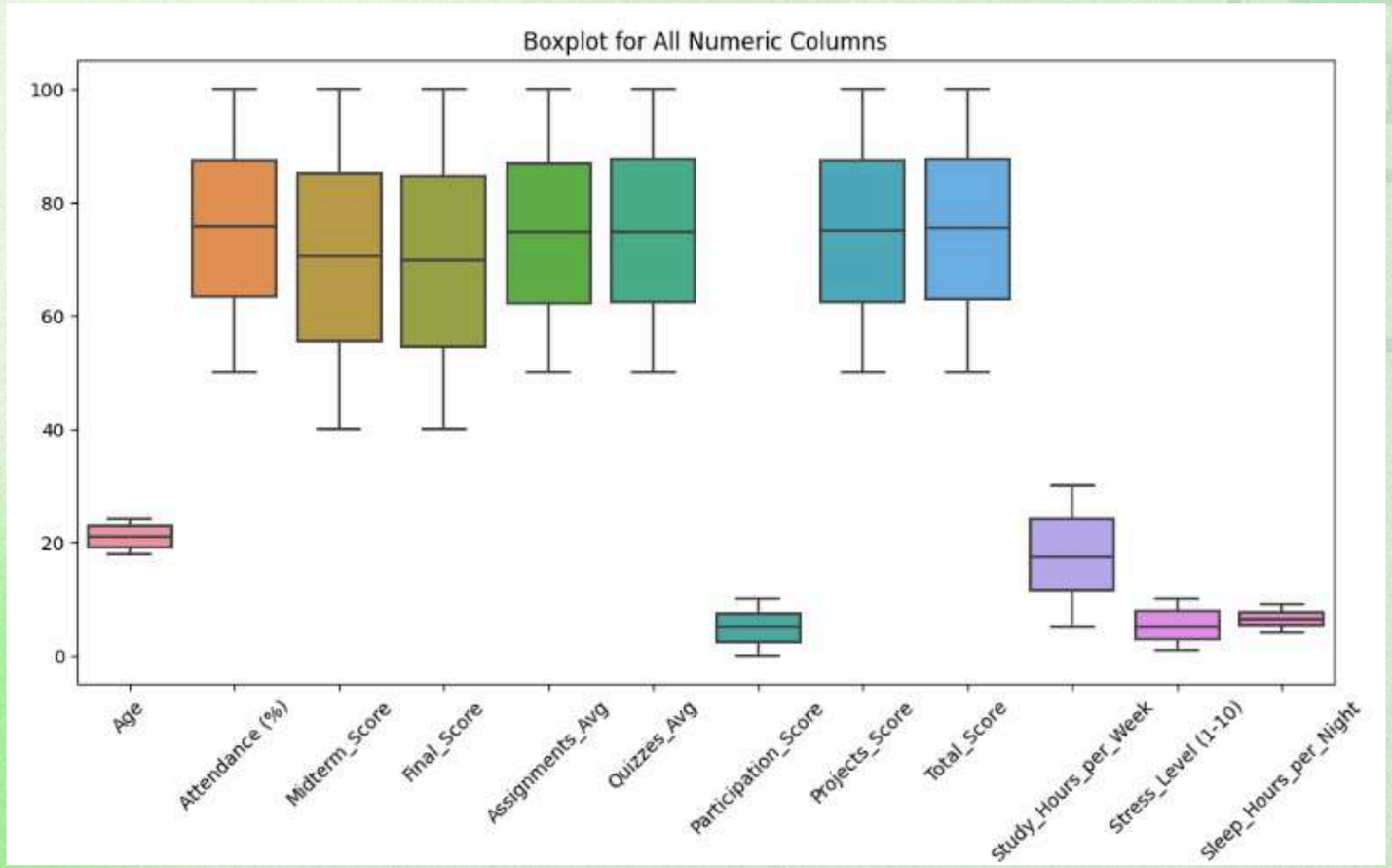


Similarly, for attendance rate and average assignment score, filling the missing values using the mean or median would have altered the natural distribution of the data. This approach would have created an artificial peak at a specific value, which does not accurately represent the real distribution of student performance. To preserve the original data distribution and maintain its statistical integrity, we applied the same method—imputing values based on the natural distribution of the data.

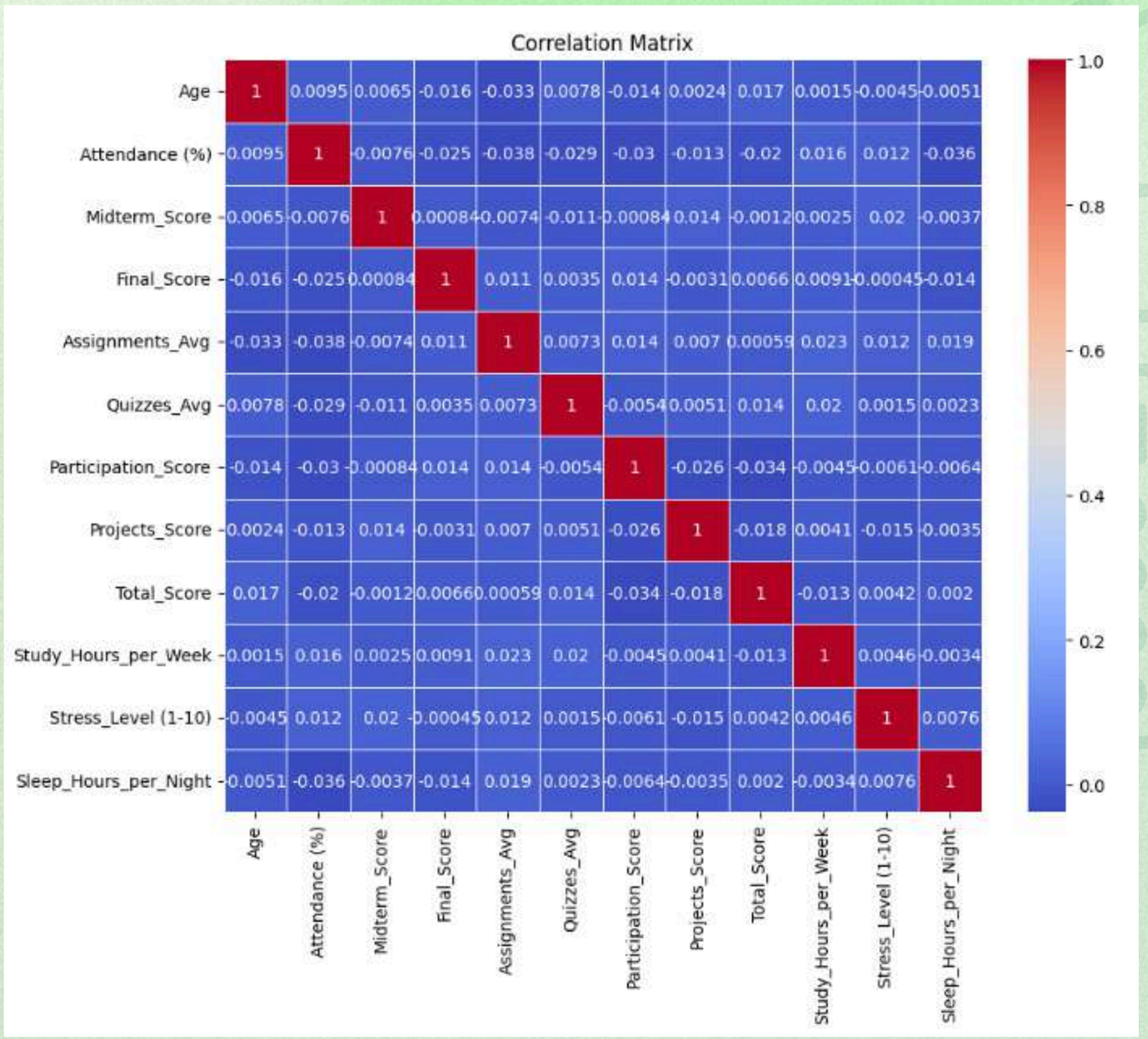




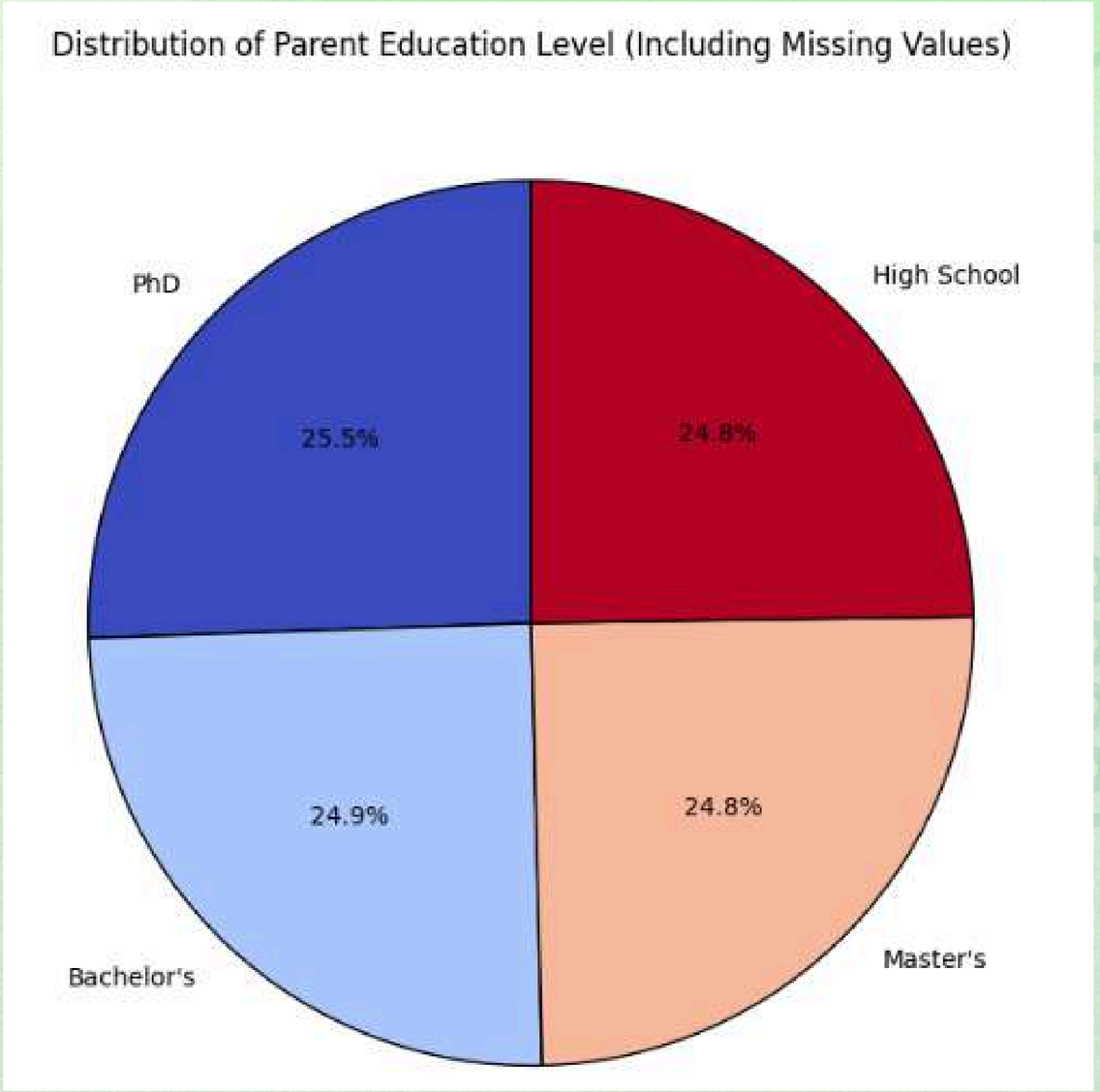
Outliers and data skewness were examined using box plots. The analysis revealed that the dataset did not contain any significant outliers, and there was no noticeable skewness or spreading in any direction. This indicated that the data was well-balanced and did not require transformations to correct distributional imbalances.



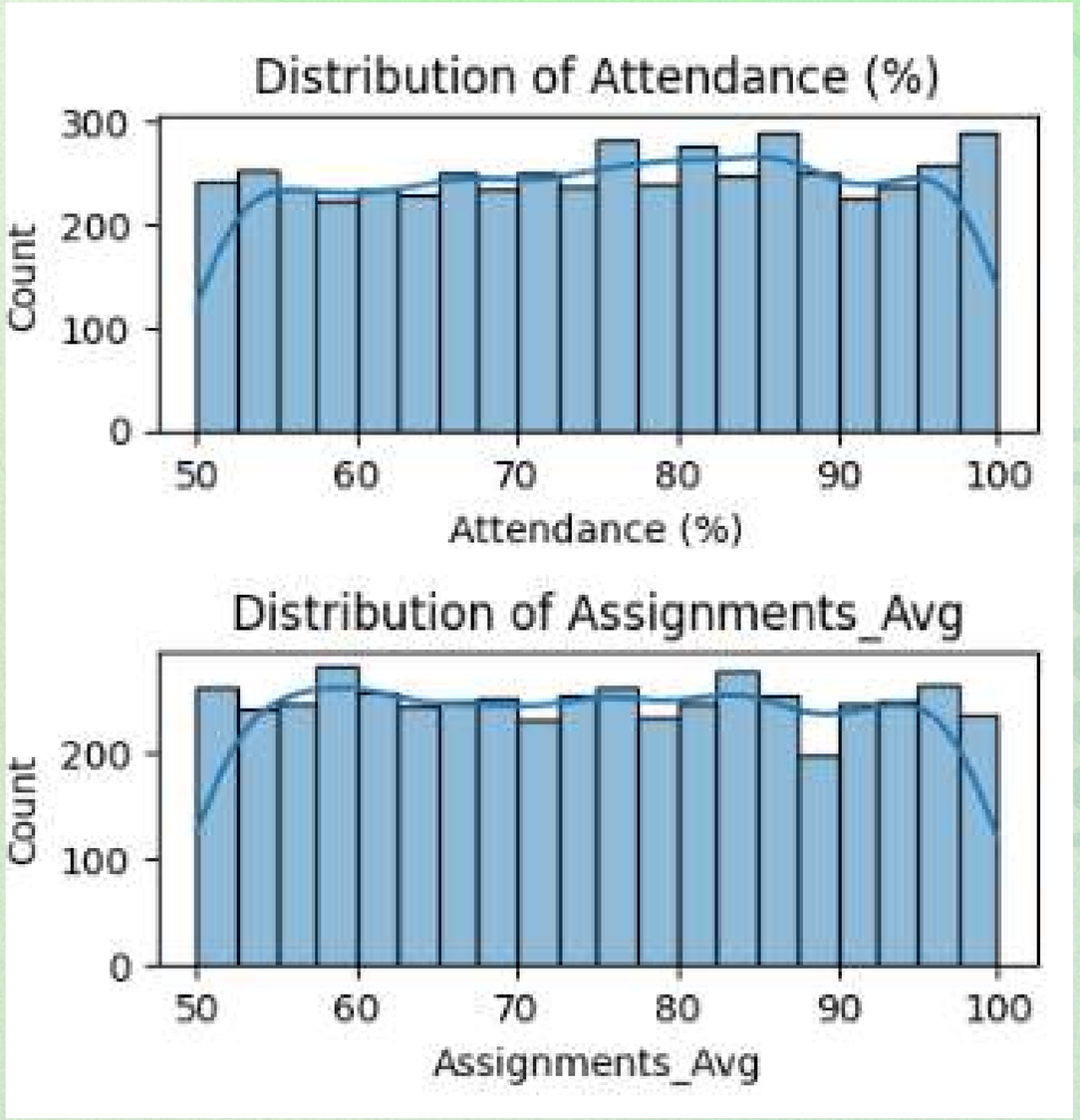
A correlation matrix was generated to analyze relationships between variables. However, the results were disappointing, as the correlations were extremely weak across all features. This indicated that no strong linear relationships existed between the variables, limiting the potential for predictive modeling based on direct correlations.



Missing values were filled based on the original data distribution to ensure that the natural patterns and proportions in the dataset remained intact. This approach helped maintain the statistical integrity of the data without introducing biases or artificial peaks.



Missing values were filled based on the original data distribution to ensure that the natural patterns and proportions in the dataset remained intact. This approach helped maintain the statistical integrity of the data without introducing biases or artificial peaks.



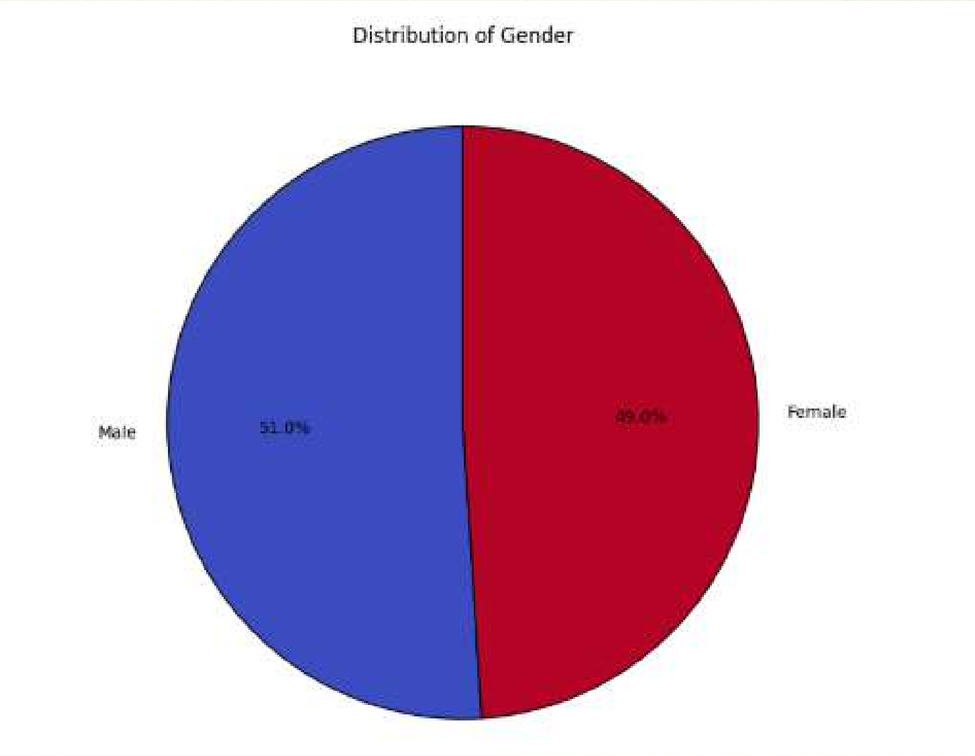
Some aggregations were performed to calculate the total number of students and the number of sections, helping to determine the sample size and understand the dataset's overall structure.

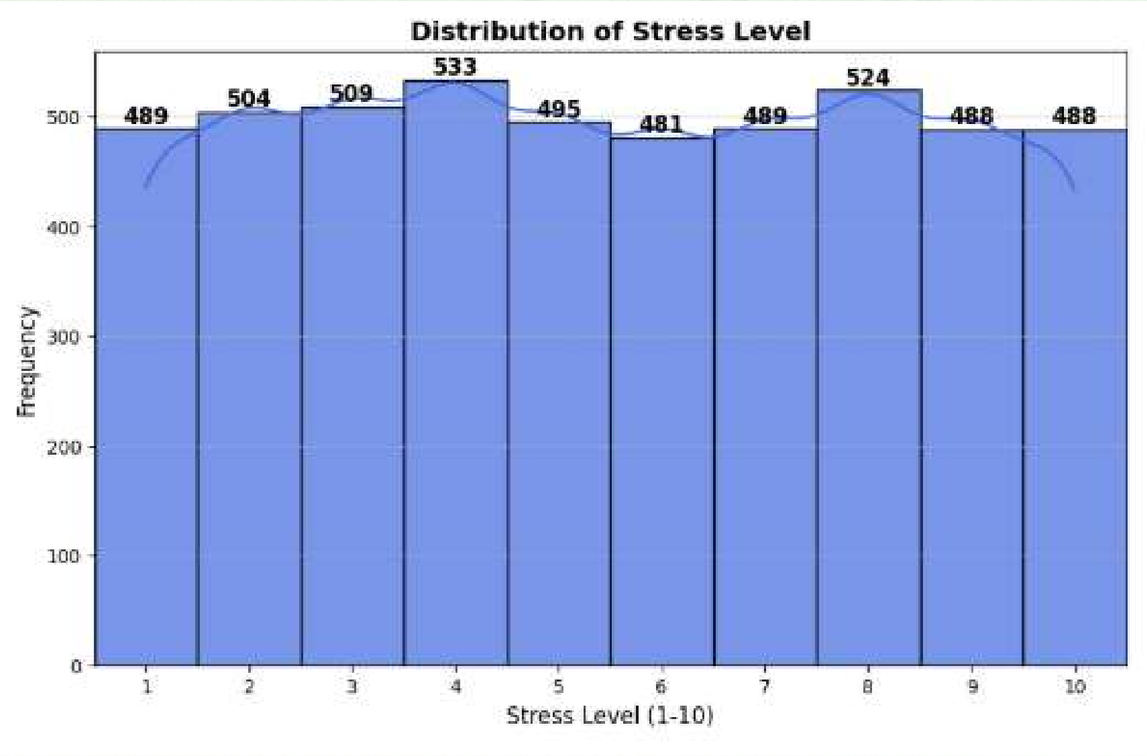
Some aggregations were performed to calculate the total number of students and the number of sections, helping to determine the sample size and understand the dataset's overall structure.

Total Students: 5000
Total Departments: 4

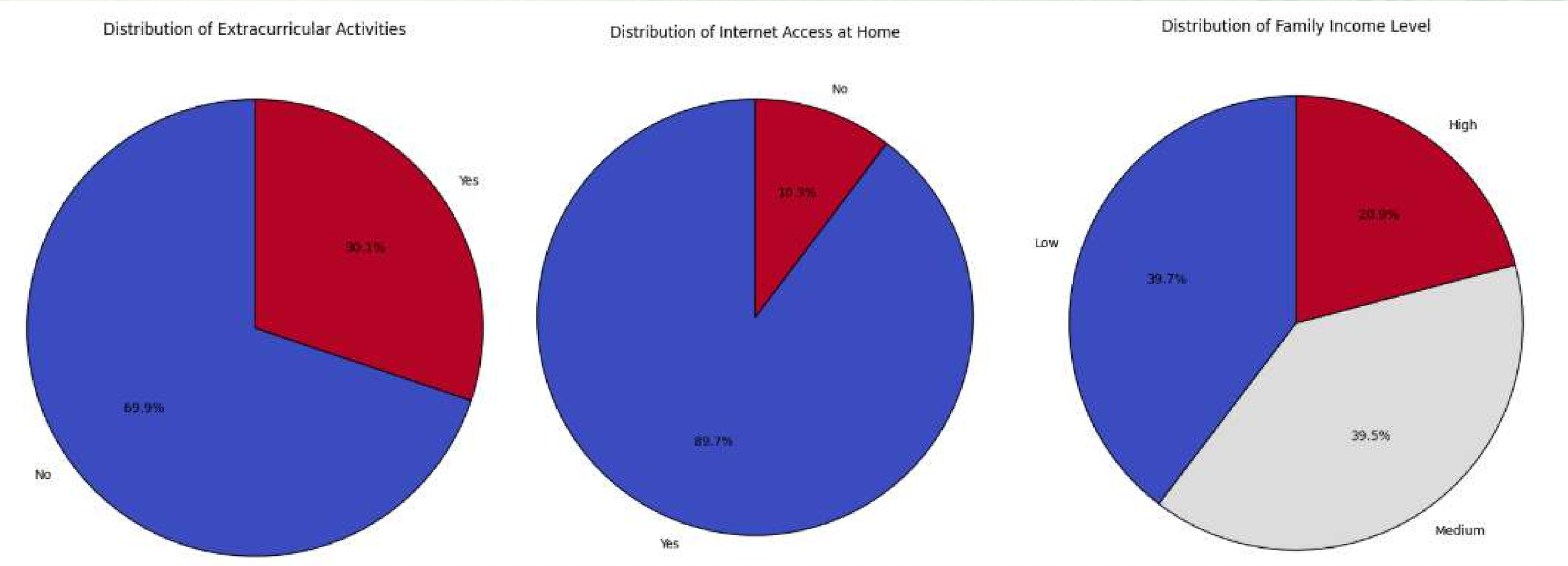
CS	Engineering	Business	Mathematics
2022 Students	1469 Students	1006 Students	503 Students

On the left, a chart displays the male-to-female ratio in the dataset, providing insights into gender distribution. On the right, another chart illustrates the number of students in each category representing different levels of academic pressure, helping to analyze how stress levels vary across the sample.

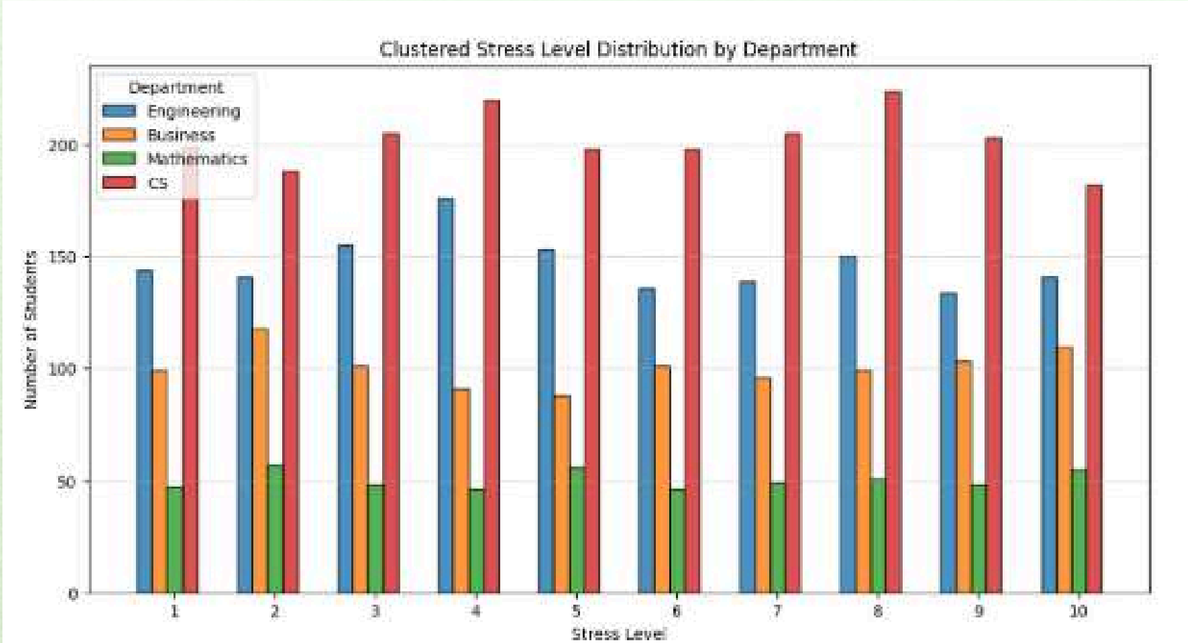


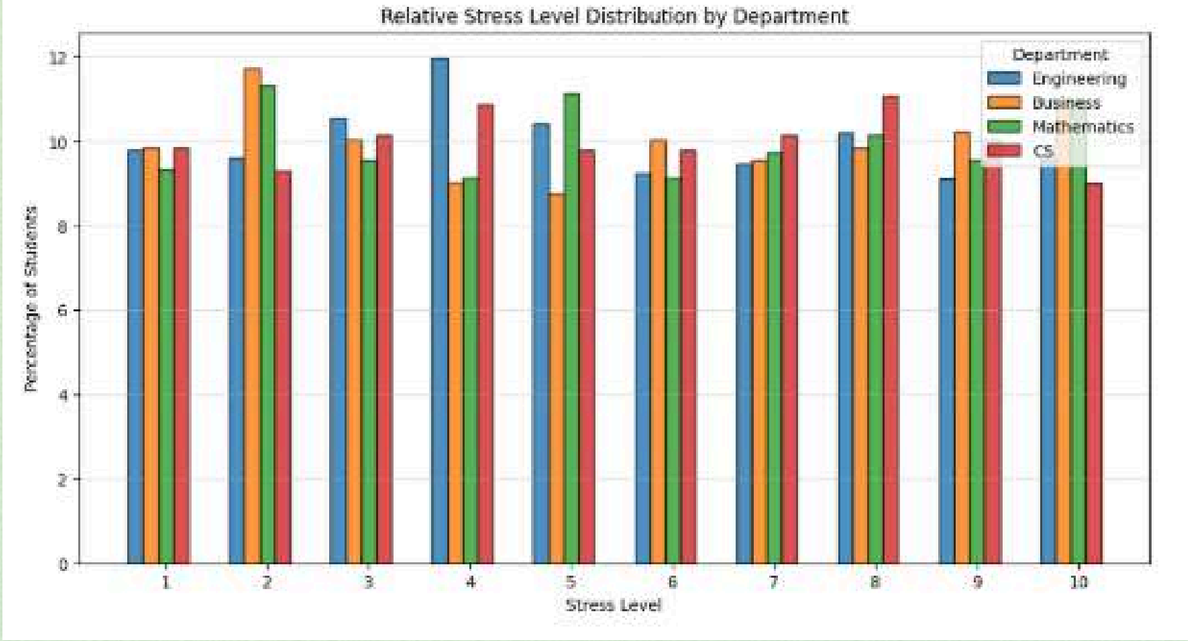


Here, we present the distribution of income levels to understand the financial background of students. Additionally, we compare the percentage of students with home internet access versus those without, highlighting digital accessibility. Lastly, we illustrate the proportion of students engaged in extracurricular activities compared to those who focus solely on academics.



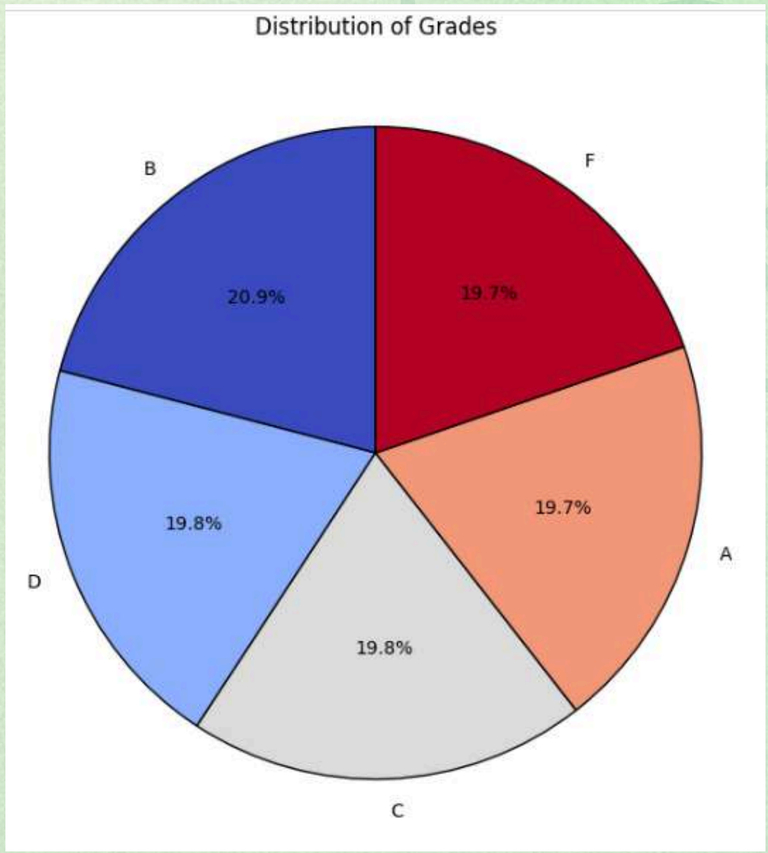
We explored whether certain departments cause more stress by analyzing student distribution across stress levels. Initially, some departments had higher counts, but normalizing by percentage within each department showed no significant differences. Stress levels were evenly distributed, indicating no clear pattern. This aligns with other findings, as the dataset's synthetic nature weakens insights and patterns.





We analyzed the grade distribution using a pie chart and found a highly uniform spread across grades. However, a strange pattern emerged—there was no logical order in the distribution of A, B, C, D, and F grades as typically expected.
To correct this, we redefined the grading system, mapping each 10% range of the upper half of the total score to a letter grade. A represented 90-100, F covered 50-60, and scores below 60 were classified as failing.

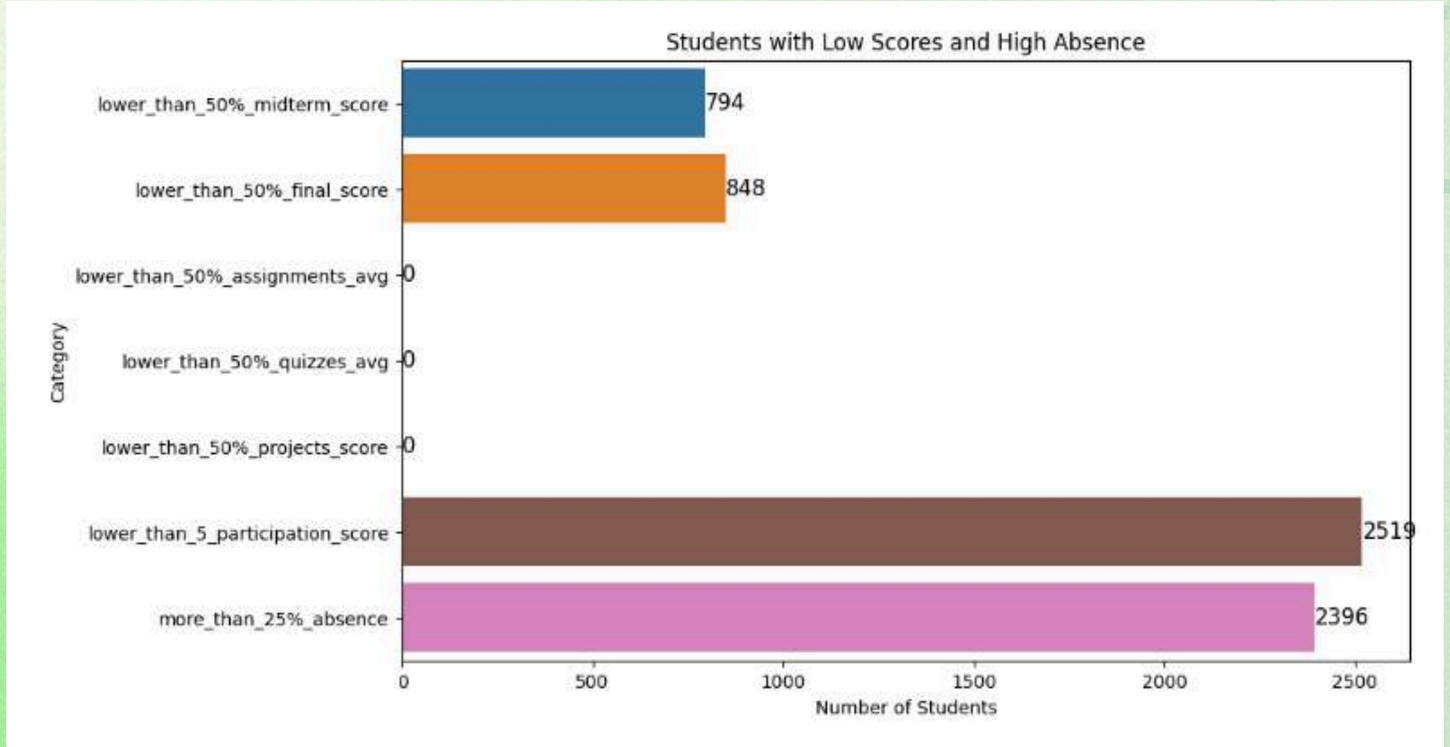
F	83.21
F	81.93
A	95.62
A	84.99
A	58.25
C	87.58
F	59.06
A	77.93
B	53.85
A	52.37
F	73.21
F	55.25
B	65.33
D	66.94
D	88.47



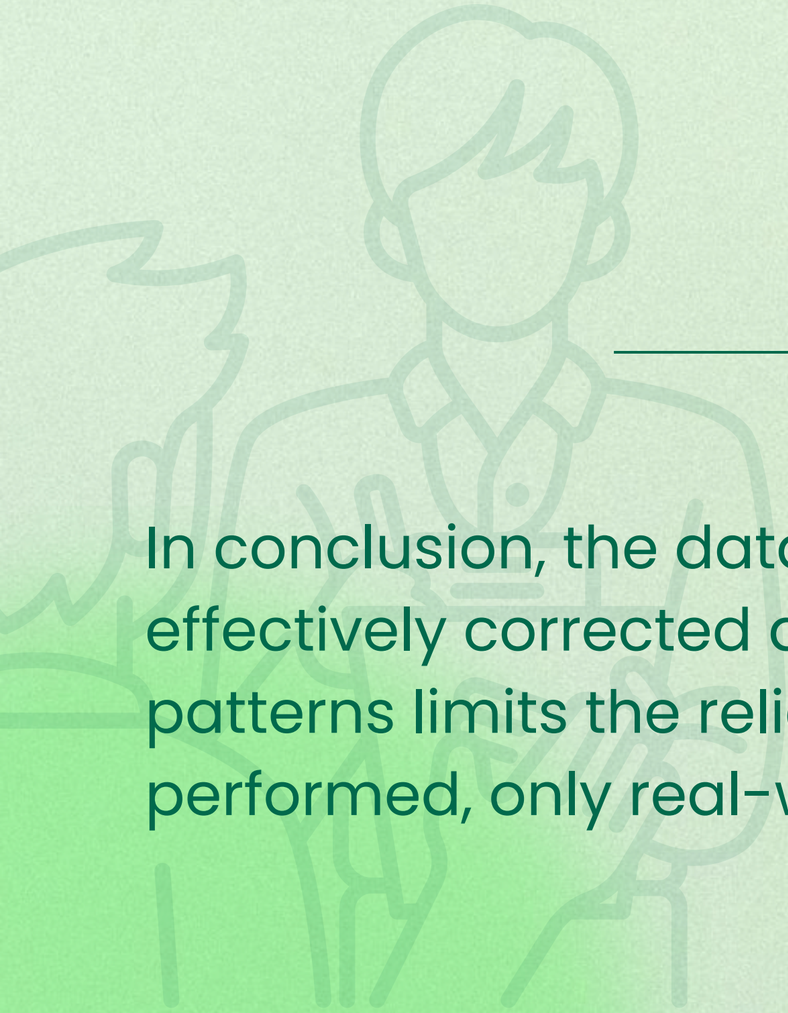
We simulated responses to key business questions to identify the most influential factors affecting the total score and determine how many students struggle with each factor.

We counted students who scored 50% or less in each contributing factor and found that:

- 46% of students are absent for more than 25% of the time.
- 50% participate at only 50% or lower.
- 17% scored below 50% in the final exam.
- 16% scored below 50% in the midterm exam.



Conclusion



In conclusion, the data quality is highly questionable and cannot be effectively corrected due to its synthetic nature. The lack of meaningful patterns limits the reliability of insights. While many analyses could be performed, only real-world data would yield valuable and actionable findings.

THANK YOU SO MUCH!

Notebook Link