Analyzing Student Performance

1. Introduction

The objective of this project is to analyze factors influencing student performance using a dataset containing various attributes related to students' demographics, family background, and academic records. By exploring and visualizing the data, we aim to identify key factors affecting grades and develop a predictive model for student performance.

2. Data Overview

Dataset Description: The dataset consists of information about students, including their school, age, parents' jobs, study habits, and academic grades. The key columns are:

- ID: Student ID
- school: Student's school
- age: Student's age
- Fjob: Father's job
- Mjob: Mother's job
- goout: Frequency of going out
- internet: Access to internet
- romantic: In a romantic relationship
- studytime: Weekly study hours
- failures: Number of past class failures
- health: Health status
- G1, G2, G3: Grades for the first, second, and third terms

3. Data Exploration and Cleaning

Initial Exploration:

- The dataset was explored to understand the distribution and nature of each feature.
- Categorical features were examined for unique values, and numerical features were described using summary statistics.

Data Cleaning:

- Missing values were addressed: internet and G1 columns were filled with the mode and mean, respectively.
- Incorrect values in the romantic column were corrected, and Gender values were standardized.
- Outliers, such as unrealistic ages, were identified and removed.

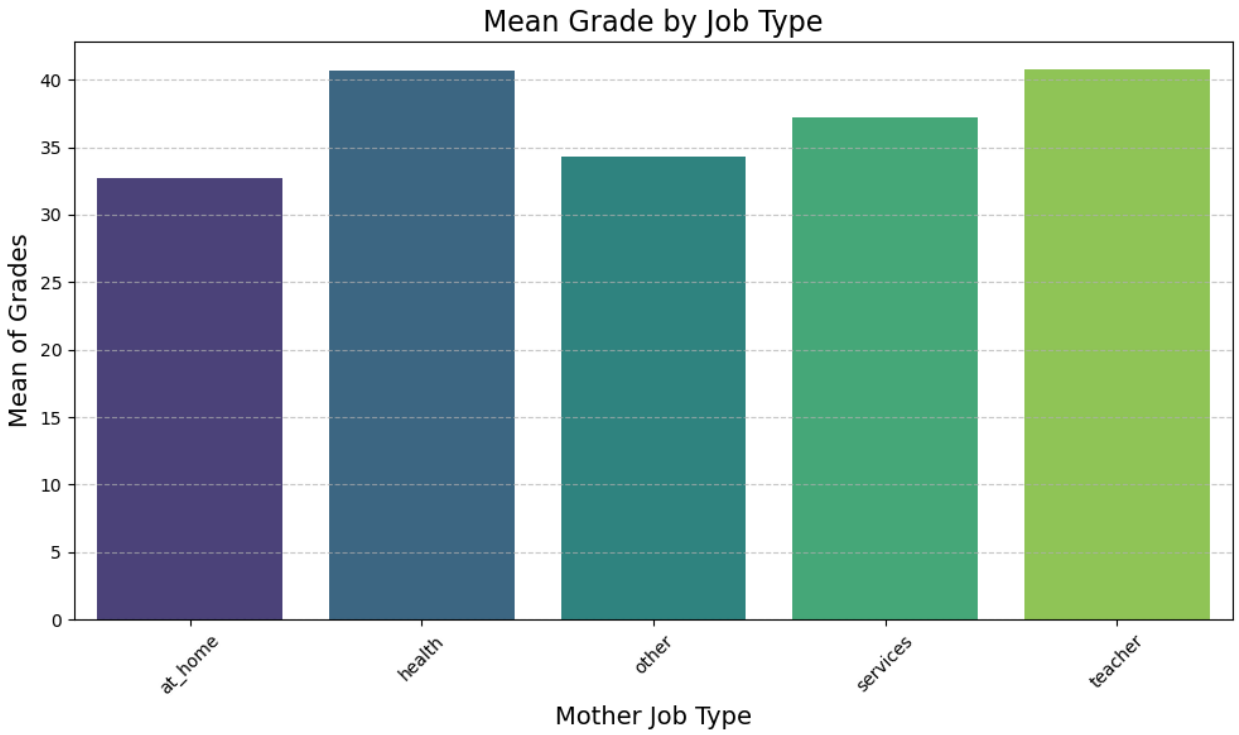
Feature Engineering:

- A new column sum_grades was created by summing G1, G2, and G3 to represent overall academic performance.
- Irrelevant columns, such as ID and individual grade columns, were dropped.

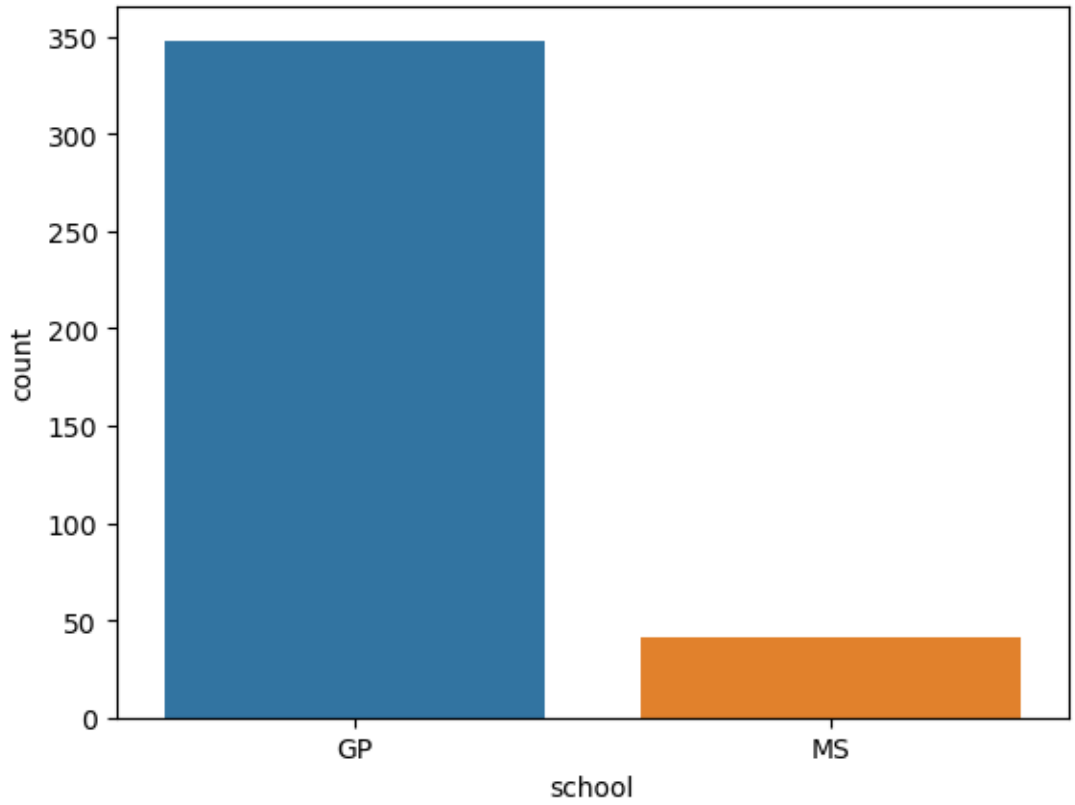
4. Data Visualization

Key Visualizations:

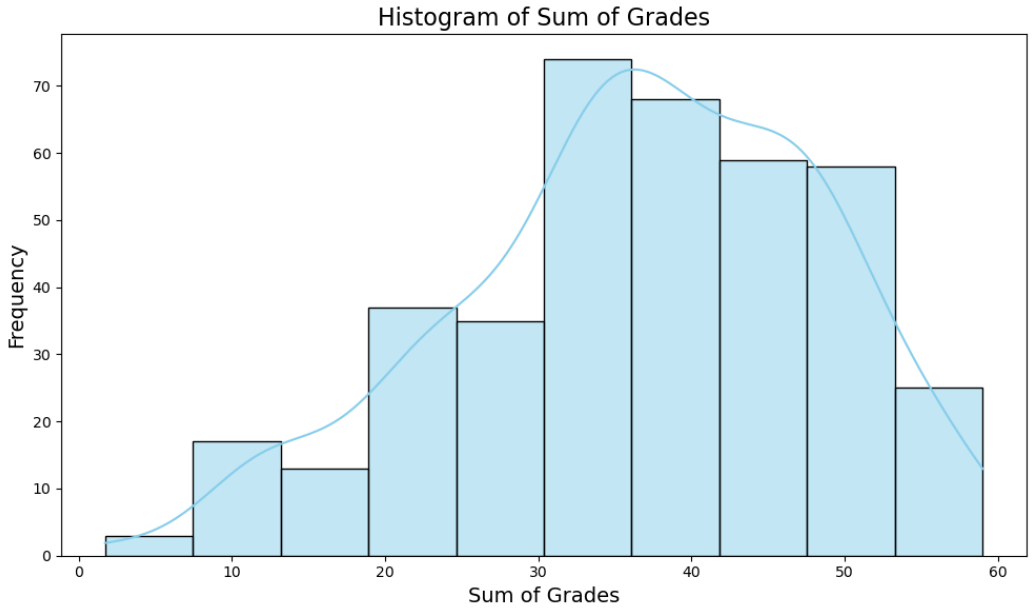
1. **Mean Grade by Mother's Job Type:** A bar plot was created to show the mean grade for each type of mother's job, highlighting differences in student performance.



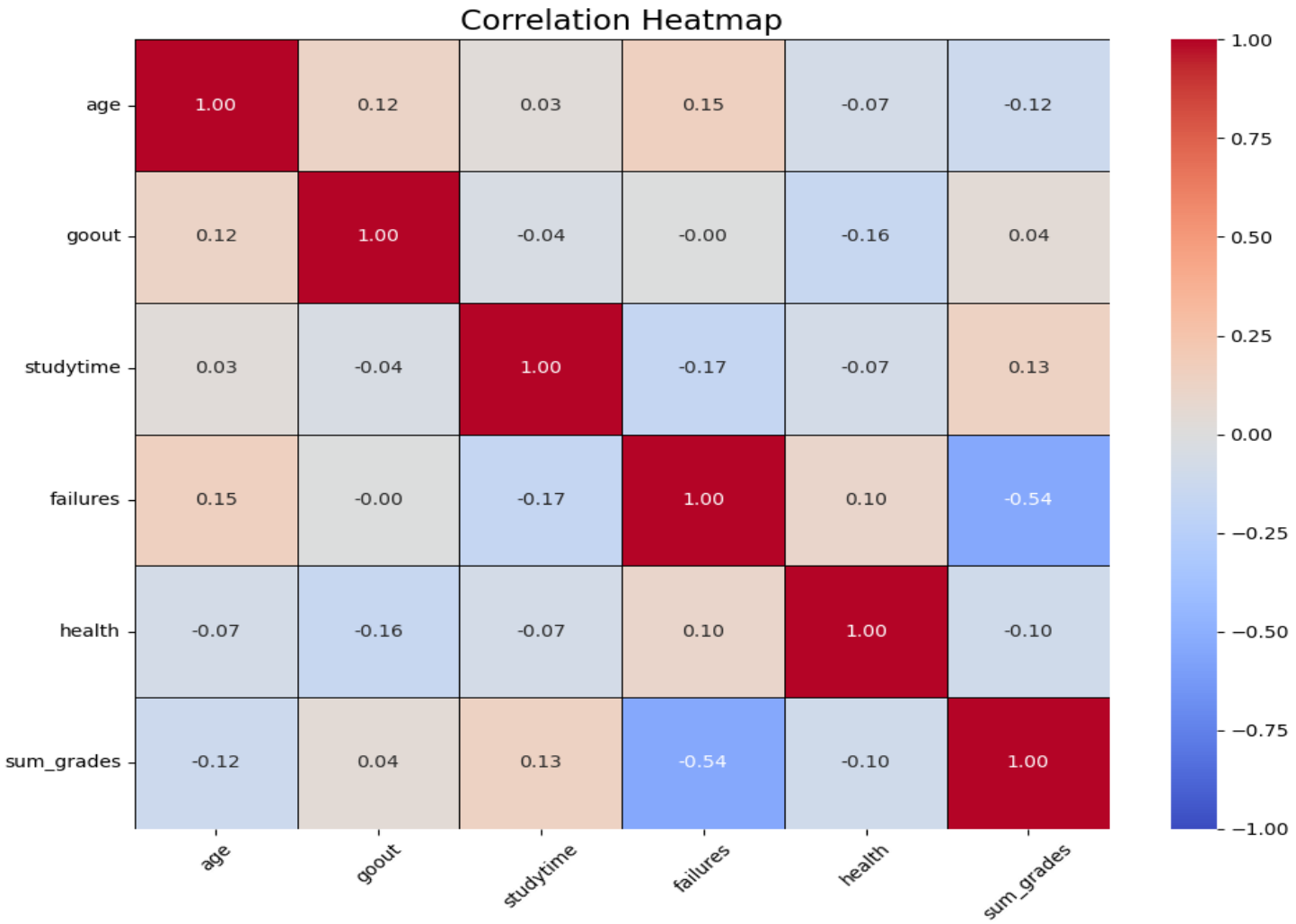
2. **School Distribution:** A count plot illustrated the distribution of students across different schools.



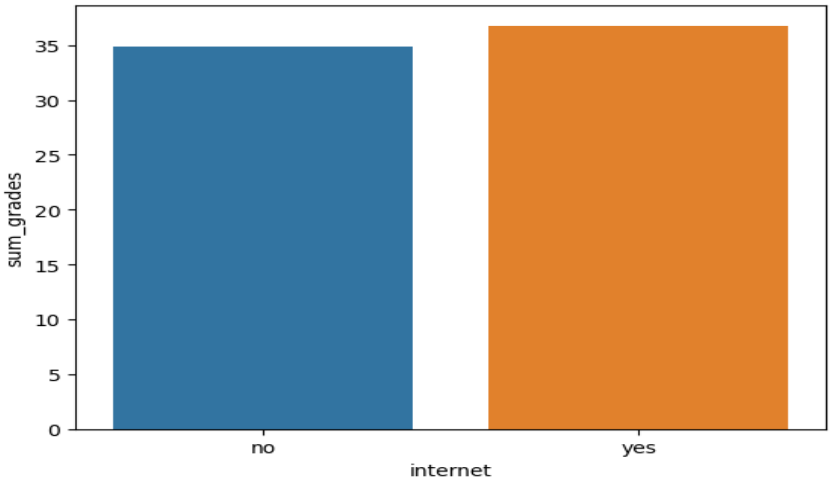
3. **Distribution of Sum of Grades:** A histogram with a KDE plot showed the distribution of the total grades.



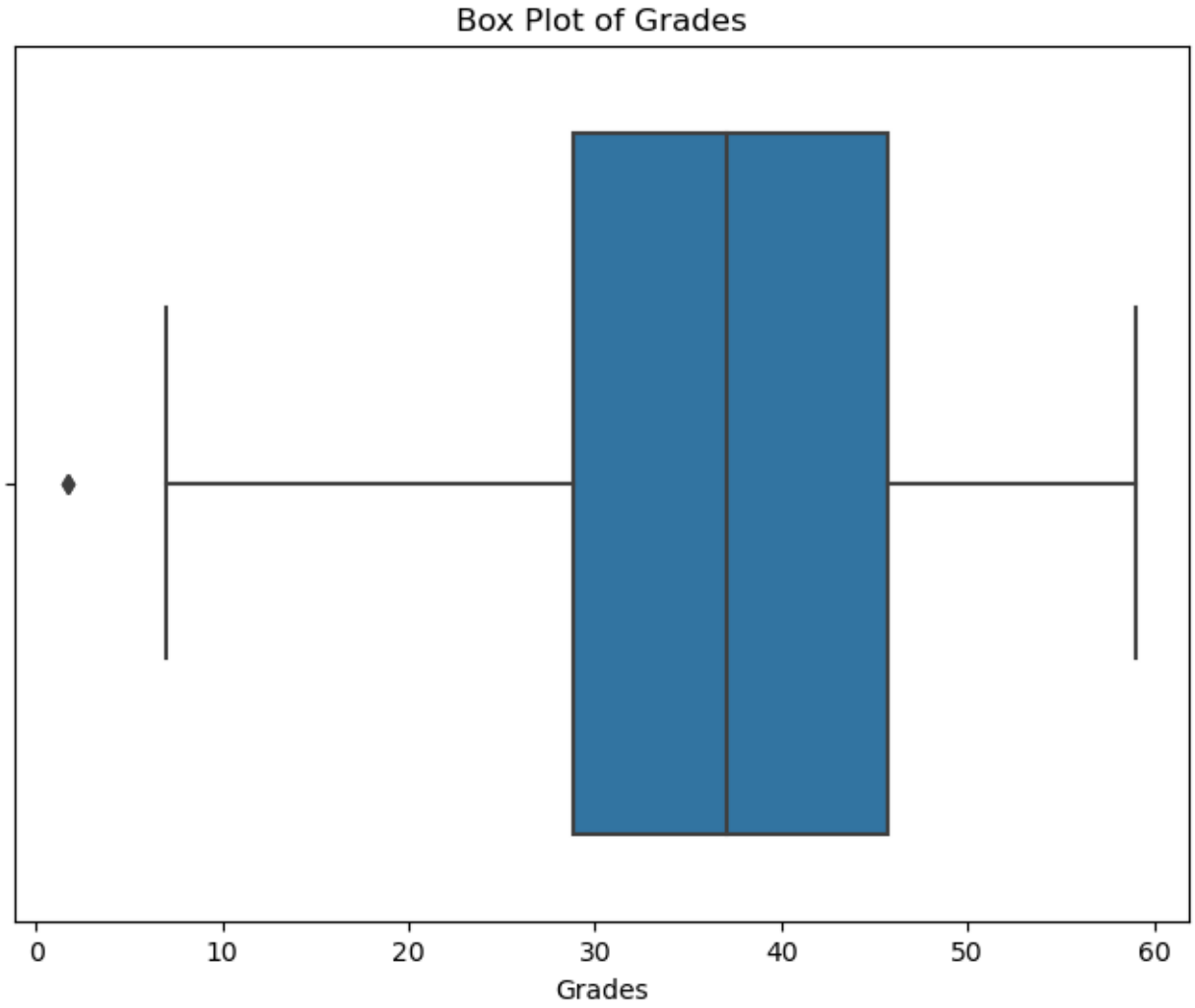
4. **Correlation Heatmap:** A heatmap visualized correlations between numerical features, indicating relationships such as the positive correlation between study time and grades.



5. **Impact of Internet Access:** A bar chart demonstrated the difference in mean grades between students with and without internet access.



6. **Box Plot of Grades:** A box plot identified the distribution and outliers in the sum_grades column.



5. Data Encoding and Scaling

Label Encoding: Categorical variables were encoded into numerical values using LabelEncoder, allowing them to be used in machine learning models.

Feature Scaling: Numerical features, such as age, were scaled using StandardScaler to standardize the data.

6. Feature Selection

Selecting Important Features: Using SelectKBest with f_regression, the top 5 features most relevant to predicting sum_grades were selected. This step helped in reducing the dimensionality of the dataset and focusing on the most influential factors.

7. Model Building and Evaluation

Regression Model: A linear regression model was developed to predict the sum_grades. The data was split into training and test sets to evaluate the model's performance.

Model Performance:

- **Mean Absolute Error (MAE):** The average error in predictions, indicating how far off the predictions are from the actual values.
    - **Mean Absolute Error** = 8.168804140374798

- **R-squared (R^2) Score:** This score indicated the proportion of variance in the dependent variable that is predictable from the independent variables.
    - 0.3322767436224593

8. Conclusion

The R-squared (R^2) score of the linear regression model was low, indicating a weak predictive relationship. However, the analysis revealed that students with internet access generally had better grades than those without, and students whose mothers are teachers performed better than those with mothers in other jobs. Additionally, some students exhibited outlier grades around 1.7, which may require further investigation. These findings suggest that while the model's accuracy was limited, certain factors like internet access and parental involvement play significant roles in student performance.