Name: Muneeza Syed   **Student Id:** 23031006

Link to Git Hub Repository: https://github.com/amazingcodemuneeza/Clustering-and-Fitting-Assignment-02

Introduction

This report analyzes a dataset containing information about student academic performance. Attributes such as age, parental education levels, study time, past failures, and final grades are explored to understand the factors influencing student success. The analysis focuses on three key areas:

1. **Identifying Relationships:** Exploring relationships between student characteristics and academic performance.
2. **Grouping Students:** Grouping students with similar profiles using clustering techniques.
3. **Predicting Grades:** Predicting final grades based on past academic performance using linear regression.

Data Exploration

➔ **Relational Graph:** A line plot is generated to visualize the relationship between age and final grade (G3). While a definitive trend might not be immediately evident, further analysis could reveal patterns within specific age groups. For example, older students might have more experience or responsibilities that impact their study time.
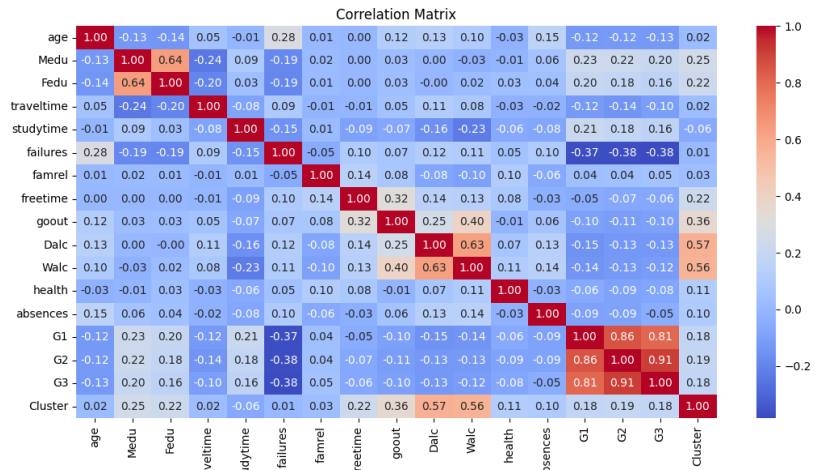

Age vs. Final Grade (G3)

Distribution of Study Time
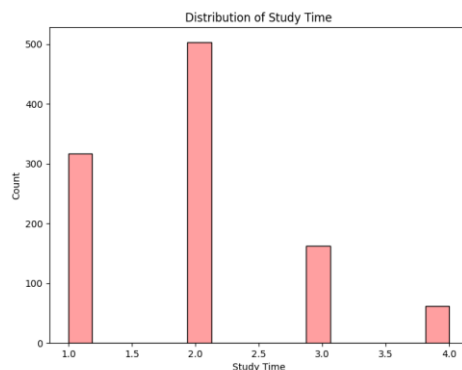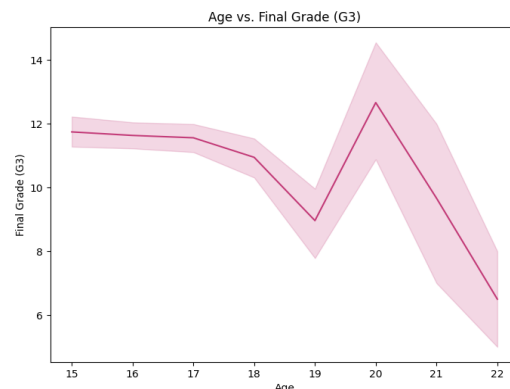
Categorical Graph: A histogram is created to explore the distribution of weekly study time across the student population. This helps identify whether study time is concentrated in a particular range or spread evenly.

Statistical Graph:

The **Correlation Matrix** heatmap (generated by plot_statistical_graph) shows the correlation coefficients between different attributes. Positive values indicate a positive relationship (e.g., higher study time might correlate with higher grades), while negative values suggest an inverse relationship. The heatmap provides a comprehensive overview of potential associations between various factors influencing student performance.


Clustering Analysis:

In our clustering analysis, we applied K-Means clustering to categorize students based on their age and final grade (G3). The algorithm identified three distinct clusters within the student population.

K-Means clustering categorized students into three distinct clusters:

- Cluster 1: 511 students (largest group)

- Cluster 0: 453 students

- Cluster 2: 80 students (smallest group)

The silhouette score of 0.417 indicates a weak cluster separation. This suggests that the clusters might not be very distinct, and exploring different numbers of clusters (k) or alternative clustering algorithms might improve the grouping.
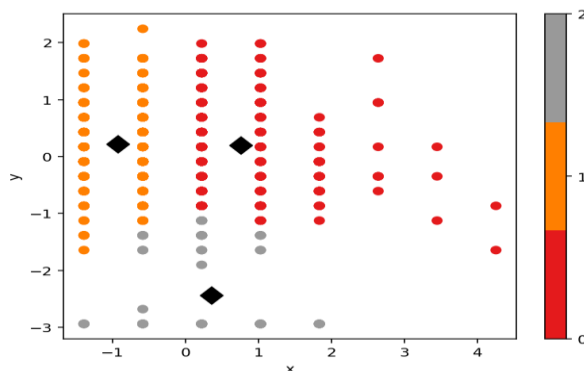
Linear Regression

Linear regression was used to model the relationship between study time and final grade (G3). The results reveal:

- **Positive Relationship:** The model coefficient (0.755) indicates a positive association between study time and final grade. For every unit increase in study time, the model predicts a 0.755 increase in final grade on average.
- **Model Intercept:** The intercept (9.929) represents the predicted final grade when the study time is zero (an unlikely scenario but serves as a reference point).
- **Moderate Prediction Error:** The mean squared error (MSE) of 15.258 signifies a moderate level of error in the model's predictions. On average, the predictions deviate from the actual final grades by around 3.9 units (square root of MSE).
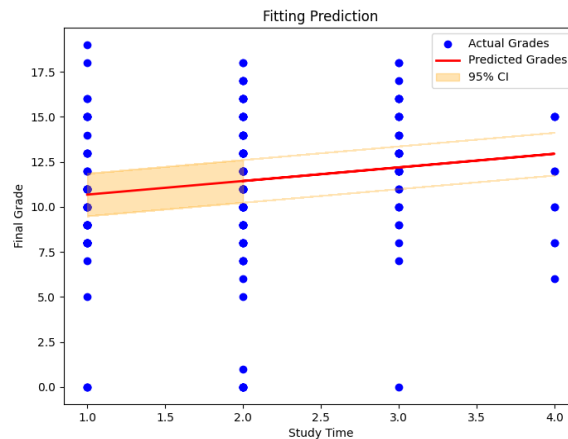
Statistical Depth of Dataset used:

	age	Medu	Fedu	Dalc	Walc	health	absences	freetime
mean	16.72605	2.603448	2.387931	1.494253	2.284483	3.543103	4.4348659	3.2011494
median	17	3	2	1	2	4	2	3
mode	16	4	2	1	1	5	0	3
kurtosis	0.030857	-1.22782	-1.1674	4.449413	-0.7825	-1.08212	26.463253	-0.364357
skewness	0.433405	-0.13933	0.119275	2.154871	0.625023	-0.49808	3.735969	-0.178451

	traveltime	studytime	failures	famrel	goout	G1	G2	G3
mean	1.52298851	1.97030651	0.264368	3.935824	3.15613	11.2136	11.24617	11.34195
median	1	2	0	4	3	11	11	11
mode	1	2	0	4	3	10	11	10
kurtosis	1.46278006	0.00084664	7.45376	1.27986	-0.83724	-0.33318	1.323489	1.730235
skewness	1.36734623	0.67001774	2.779659	-1.05426	0.038872	0.07781	-0.49664	-0.98455

Conclusion:

By analysis of the student performance dataset we found valuable information about student success factors. Clustering helped identify distinct student profiles based on their attributes, while fitting provided a predictive model for final grades. The generated visualizations aided in understanding the relationships between student characteristics, academic performance, and study habits. Overall, this analysis offers valuable insights for educators and policymakers to tailor strategies for supporting student success.