Data Analysis Report

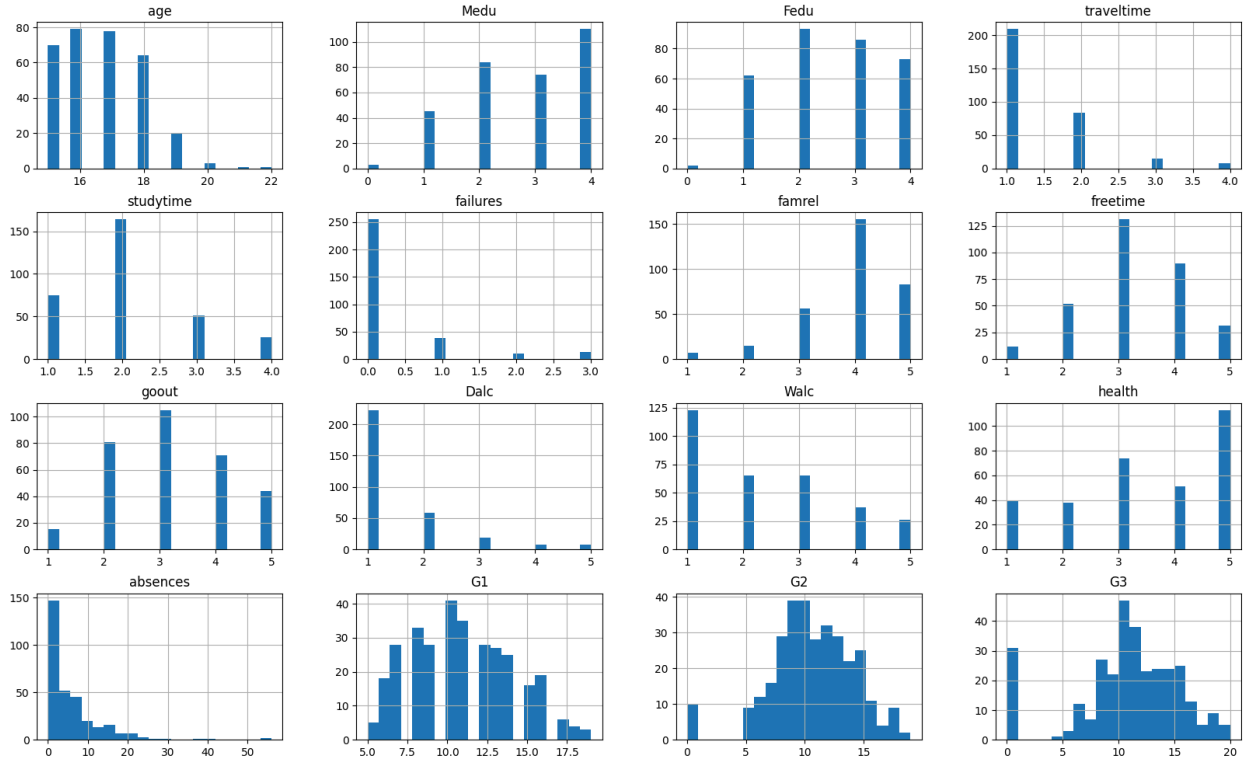
Dataset Summary:

This dataset is a collection of students' personal and socio-economic factors from two schools. These variables include students' age, freetime, failures, studytime, etc.

Objective:

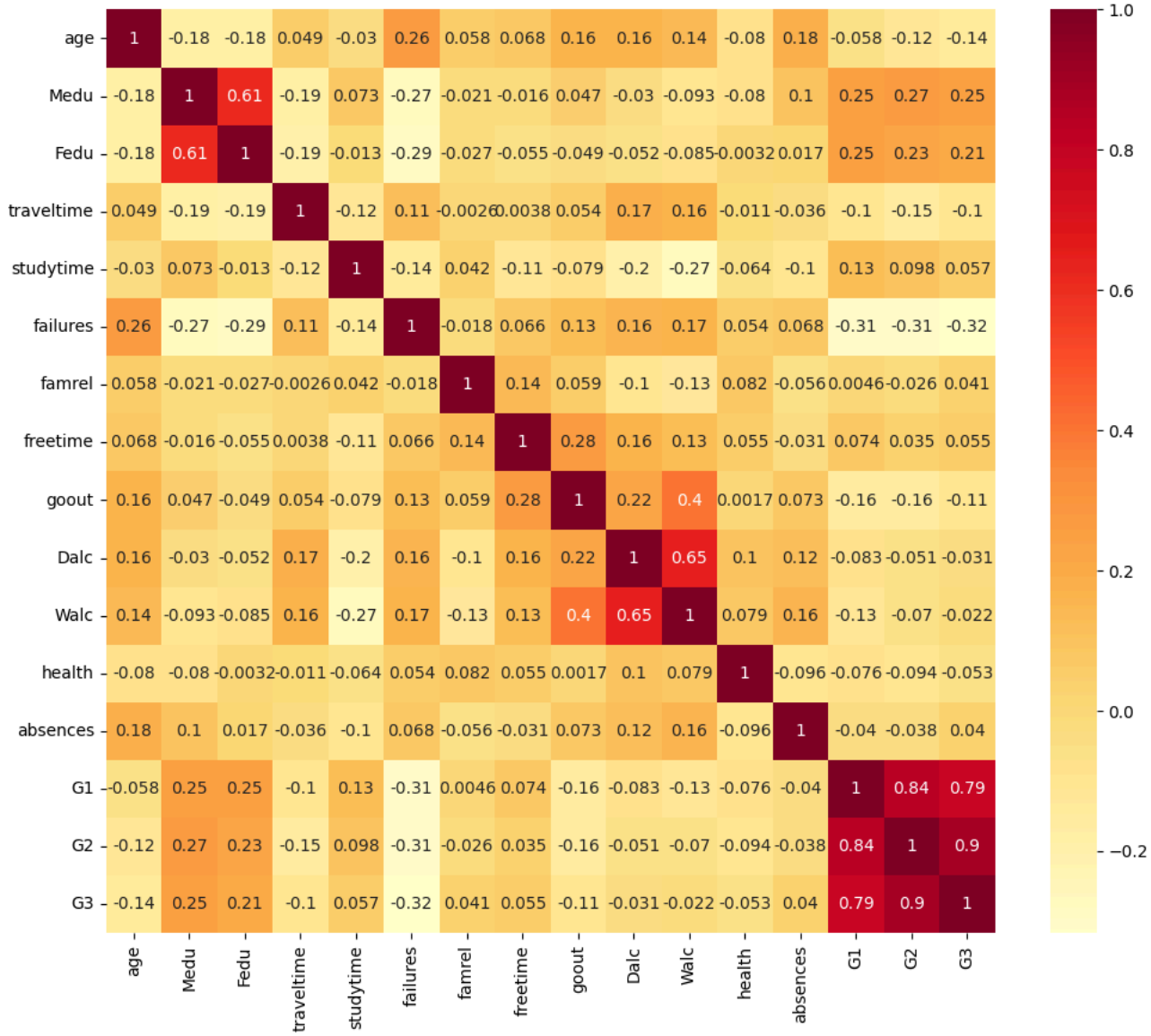
Explore the dataset, select features/factors, and train a regression model to predict final grades (G3).

Distribution of data (numericals)



These variables are numerical values. Most of the data falls into certain amounts except the grades which are G1, G2, and G3. These grades are evenly spread across the range of 0 - 20.

Data Correlation:



This graph shows the correlation between each feature in the dataset, higher number means higher correlation. All the grades are highly correlated with a score .79 or higher. Fedu (father's education) and Medu (mother's education) have median correlation as well as Dalc (workday alcohol consumption) and Walc (weekend alcohol consumption).

Mutual Information:

Top five features with mutual information

G2	1.374960
G1	0.760152
absences	0.177487
Mjob	0.115033
romantic	0.071359

Models:
6 models were tested with training data. The top three are Random Forest, K_Nearest Neighbor, and Bayesian. Random forest was selected based on the metrics below:

R2 score : 0.82
Mean absolute error: 1.25
Mean squared error: 1.91
Root mean squared error: 3.83

Features selection:
A total of five 8 features selection was run and tested with cross validation score on the Random Forest model. The most successful features were age, studytime, famrel, freetime, goout, Wacl, health, absences, G1, and G2. The metrics for these features are below:

R2 score : 0.86
Mean absolute error: 1.10
Mean squared error: 1.65
Root mean squared error: 3.20

Hyperparameters Tuning:
Random Forest was test on selected features with a range of hyperparameters including n_estimators, criterion, max_depth, min_samples_split, min_samples_leaf, max_features. The best parameters are:

N_estimators: 16
Criterion: squared_error
Max_depth: 155
Min_samples_split: 4
Min_samples_leaf: 3
Max_features: None

Validation Set Evaluation:
Final model with selected hyperparameters are tested and evaluated on validation data. The metrics are:

R2 score : 0.78
Mean absolute error: 1.23
Mean squared error: 3.58
Root mean squared error: 1.89

Validation Set Evaluation:
Final model with selected hyperparameters are tested and evaluated on test data. The metrics are:

R2 score : 0.92
Mean absolute error: 0.85
Mean squared error: 1.30
Root mean squared error: 1.14