Linear Regression Analysis

Data collection and Cleaning

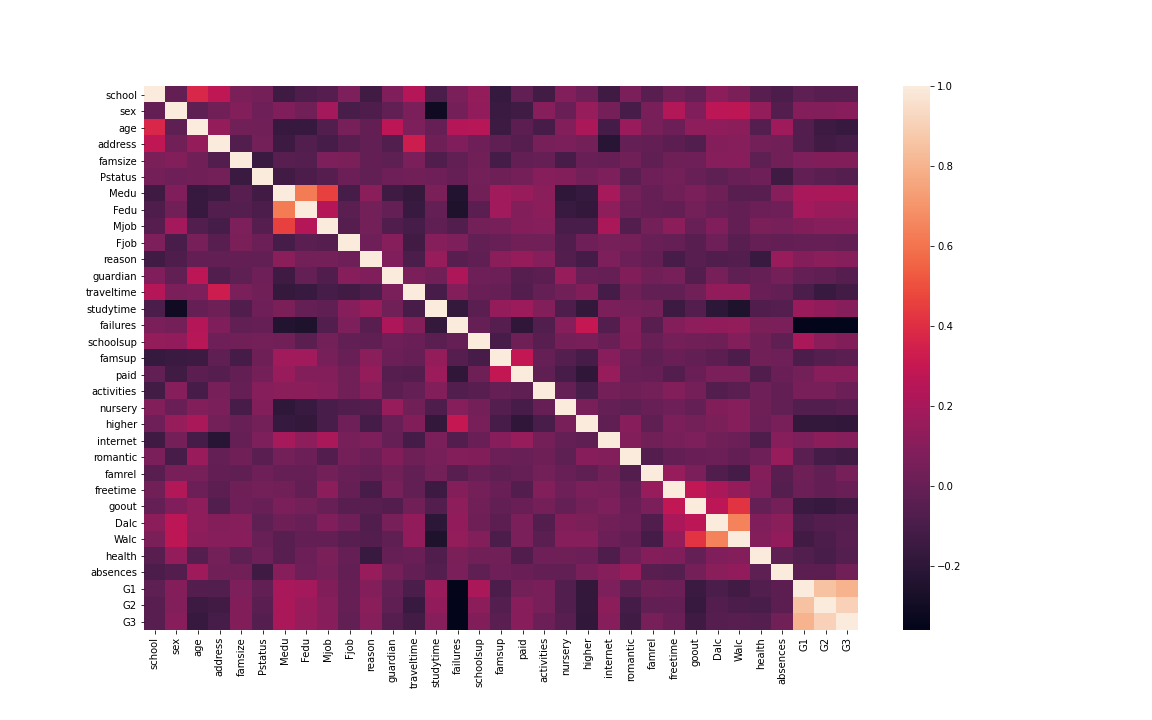
* Read the CSV file and construct a Data Frame
* Check for null values in Data Frame { df.isnull().sum() gives the sum of null values in each column }. No null values in given Data Set

Categorical Variables

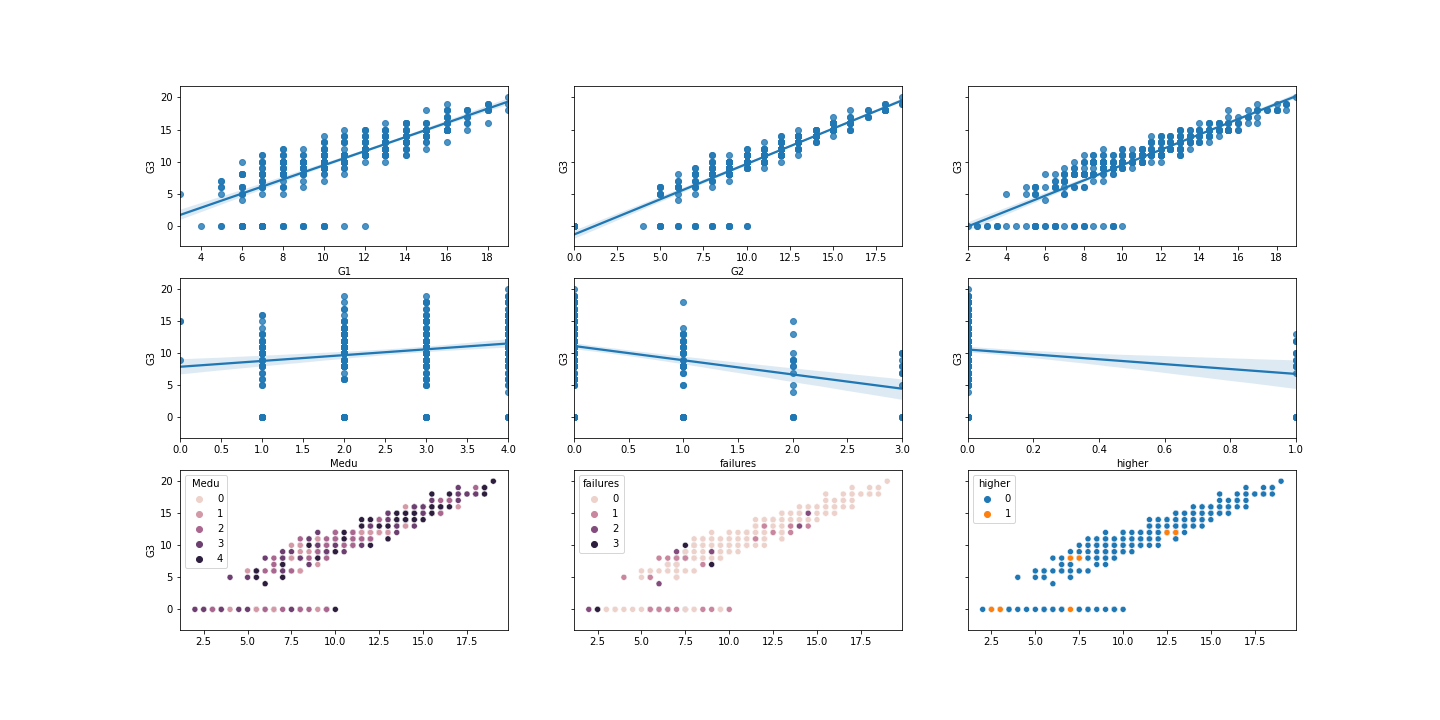
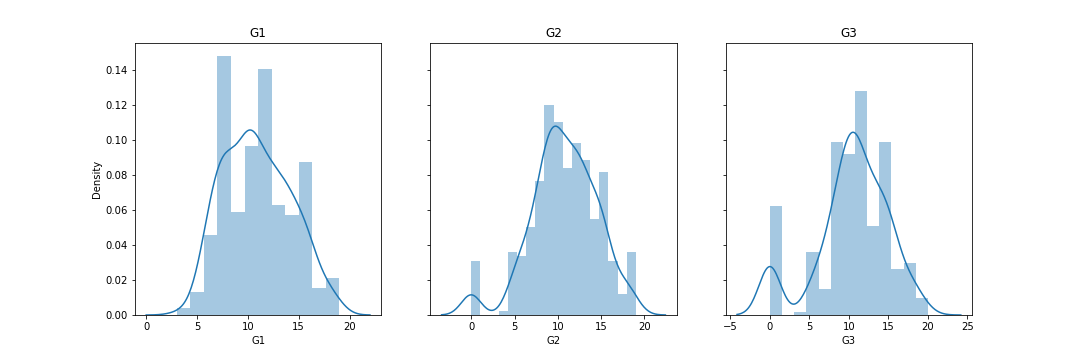
* Select all the columns with categorical variables
* If there are less categorical variables we can use OneHotEncoding but there were 17 columns so I used factorize method to convert it to integer classes (similar to Ordinal encoding)

Co- relation matrix

* Using method .corr() I was able to get the corelation of every column in the dataset
* Using heatmap to plot the data I was able to quickly identify the columns which have strong relation with target variable



Data Visualization

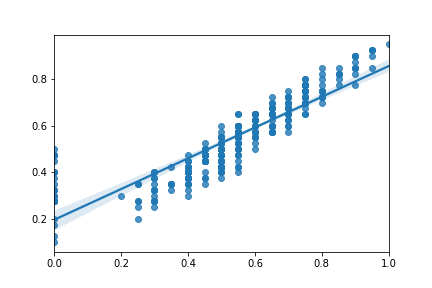
* Plotting an histograph to check the overall data and any hints for outliers.
* Plotting a scatter plot to know the trend of relationship with G2 and G1.
* Formulated an average of G1 and G2 with gave me a better and concise plot.
* Checking if any other combination with other variable could give better results

Feature Selection/Train and Test Split

* From all the above plots the average of G1 and G2 seemed close to the regression line so this became my feature.
* Splitting Data into test(20%) and train (80%)

Model Training

* Used a Linear Regression Model from sklearn module.
* Normalized the data and then feed it to the Linear Regression Model
* X- Coefficient = 1.17275541, X intercept = -0.11151271



Validation/Testing

* MSE = 0.00984681464304563
* RMSE = 0.09923111731229085
* r2 \_score = 0.8394388687317985