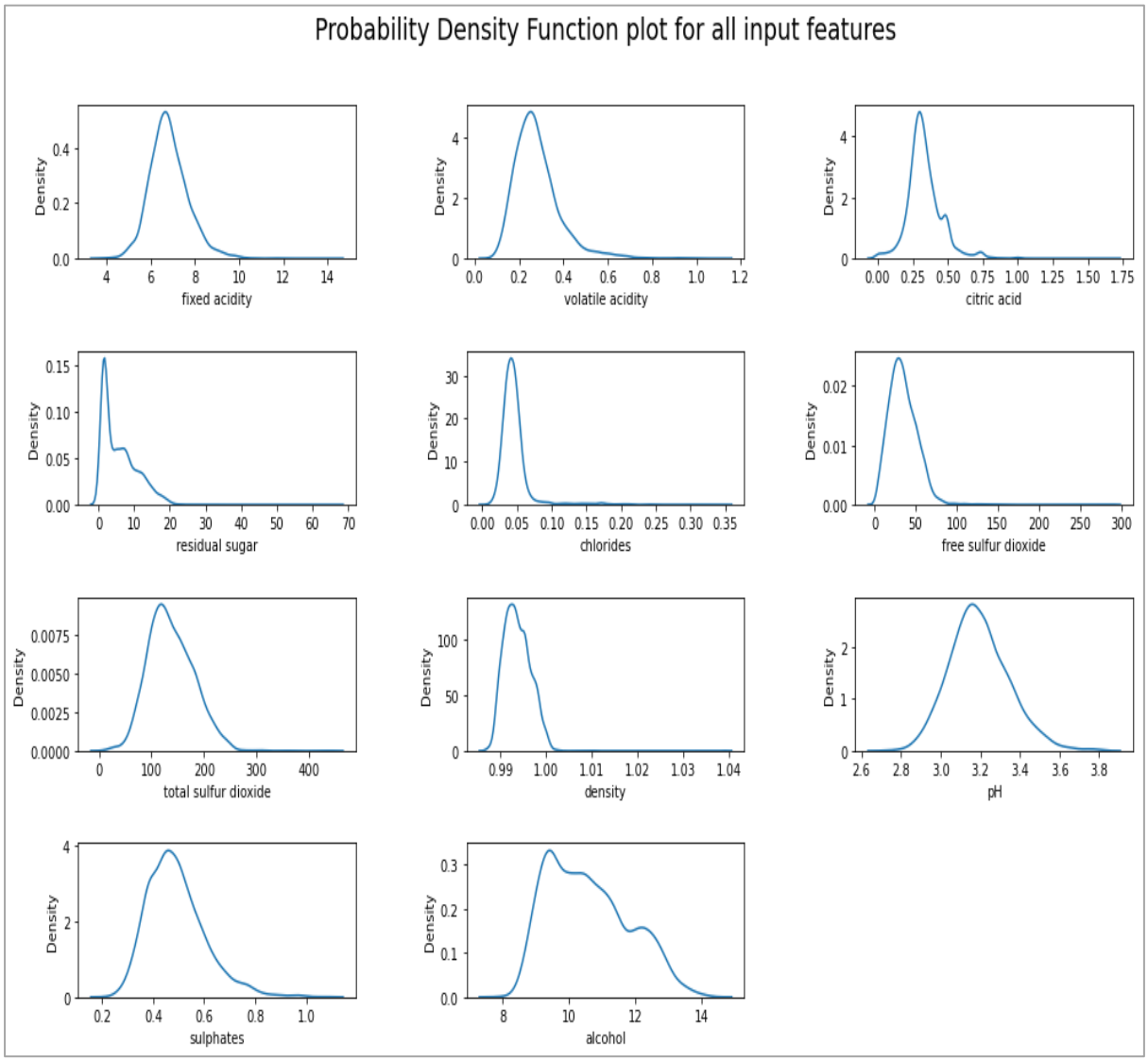
Introduction

This dataset was created using white wine sample. The input includes objective tests (e.g. pH values, acidity etc.) and the output is based on sensory data evaluation given by wine experts with grading from 0(very bad) to 10(very excellent).

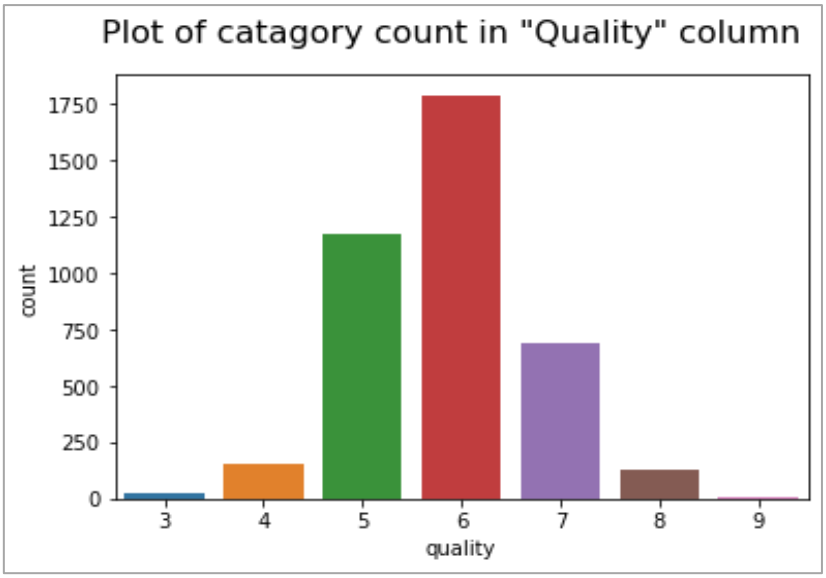
The objective of this report is to summarize the various processes involved in the model making process and also to interpret the data and results.

Interpretation of the data

The original dataset has 4898 entries with 12 columns. But upon examining we can find that although the data had zero null entries, it had 937 rows of duplicate entry. So, after removing the duplicated entries we are left with 3961 rows of data to work with. All the columns have numerical data so it would not require any encoder techniques to fit into a model.

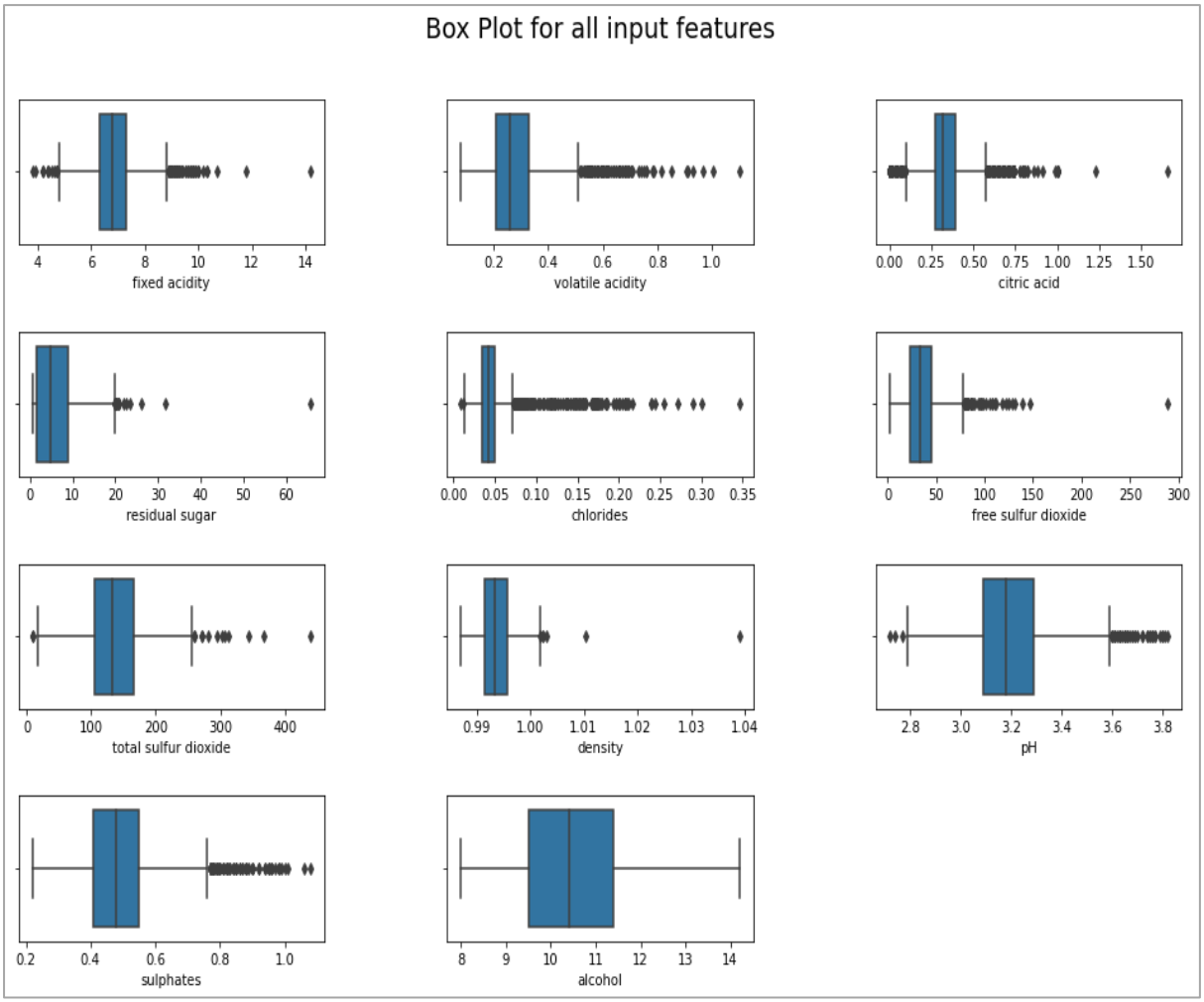


After observing the probability density plot for the features, we can conclude that all the features excluding ‘pH’ column have highly right skewed data which is not good for our machine learning model



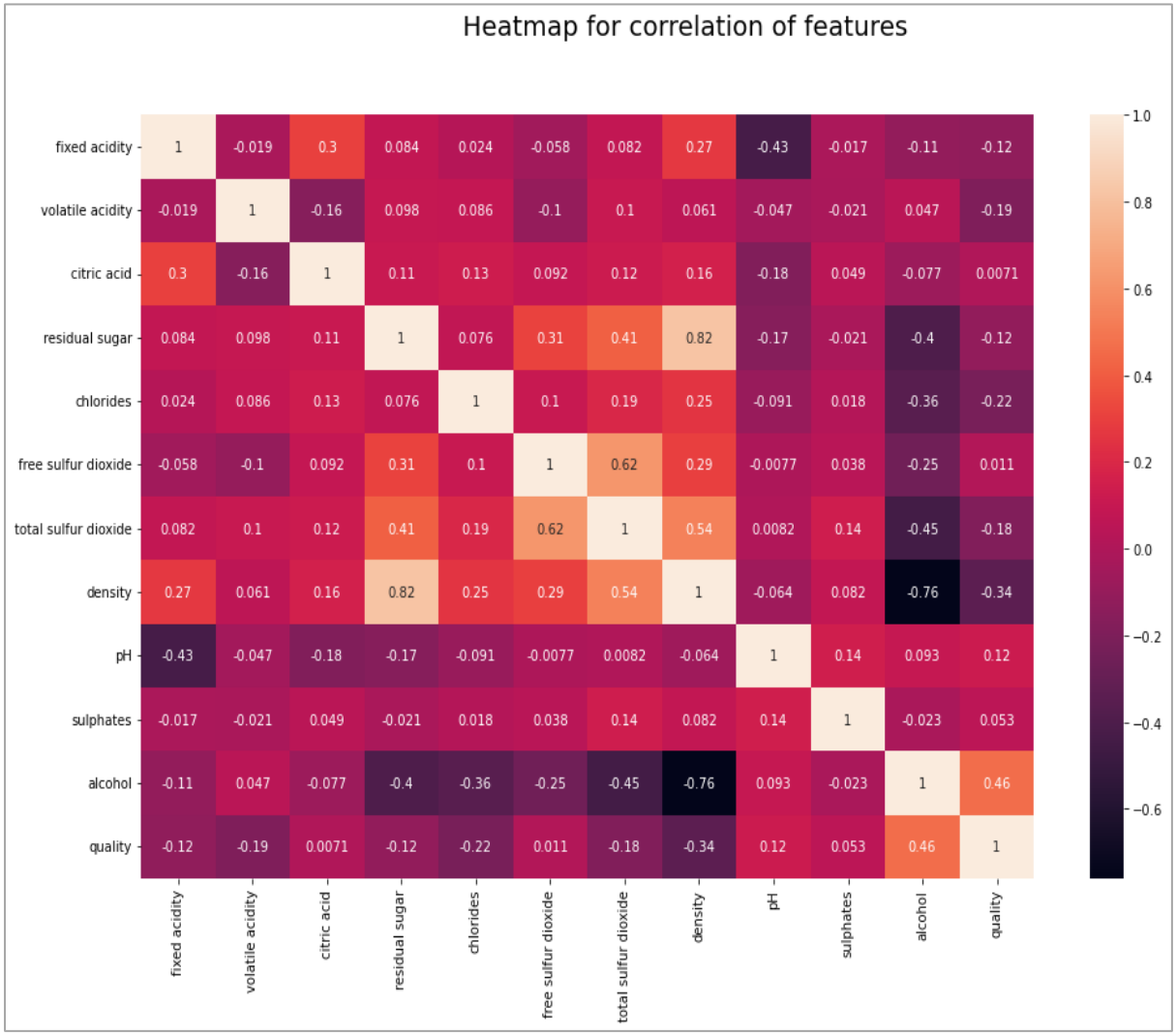
Again, from the count plot in “quality” column we can confirm that the dataset is highly imbalanced having majority of 5, 6 and 7 grade. This will also affect our Regression model.

Now we also have to check for outlier datapoints in each column. For this we used boxplot method. We obtained the following graphs.



The boxplots indicate that all the features have outlier datapoints except “alcohol” column. This will highly impact our Regression model but due to less knowledge of the domain, we can’t make decisions about these outliers.

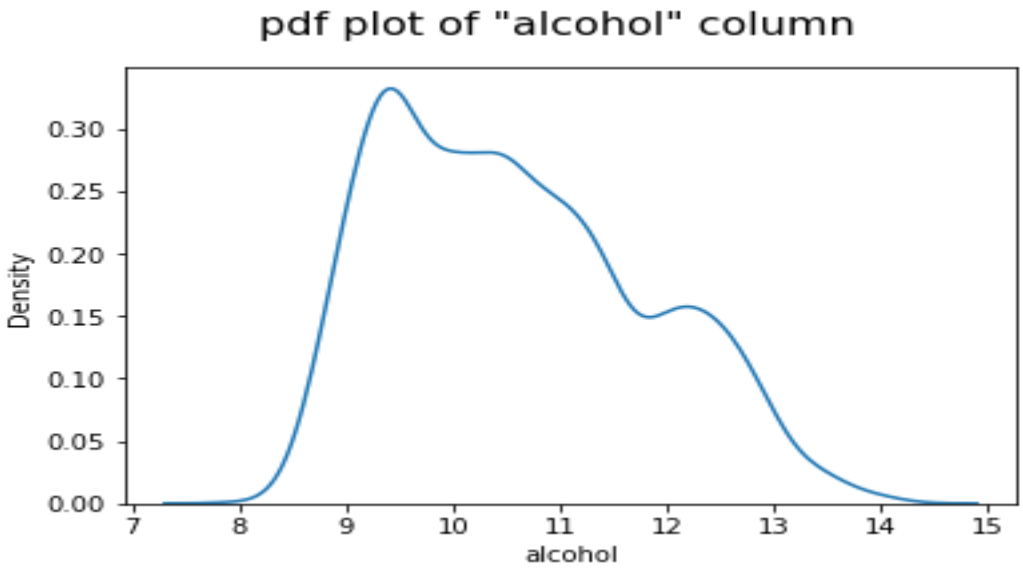
To check for multi-collinearity in the data we plotted a heatmap based on correlation matrix.



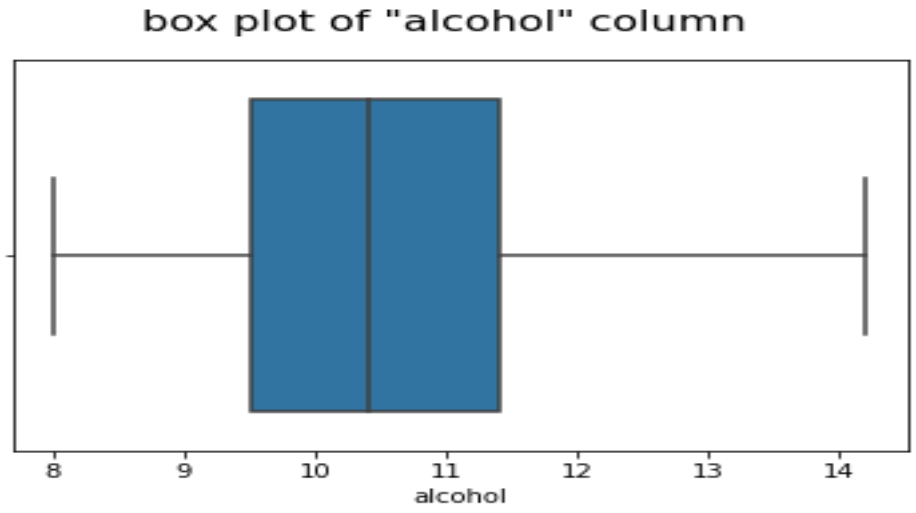
Observing the heatmap, we can easily conclude that some of the features are highly corelated like “residual sugar” & “density” and “total sulfur oxide” & “free sulfur oxide”.

Simple Linear Regression

For making the model we have “quality” as output column and we need to choose a single input feature. So, we chose “alcohol” column as it has highest correlation with the output column. We checked the data distribution using probability density function and found it almost normally distributed.



Again for checking outliers we used box-plot and found that no outliers were present in the data.



Then we divided the data into train and test with 20% data leaving for test purpose. After the split we get 3168 rows in train set and 793 rows in test set.

Finally, we used the Linear Regression Model of *scikit learn* library to train with this data with one input feature and one output feature.

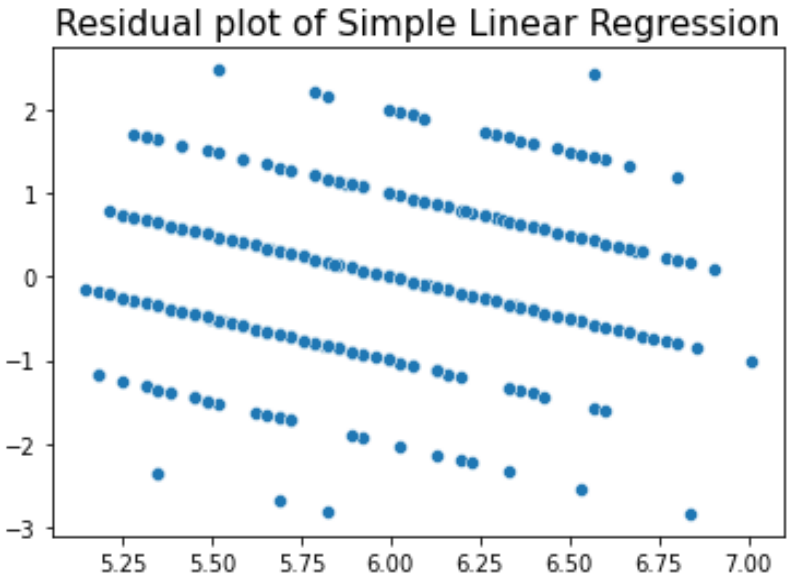
Multiple Linear Regression

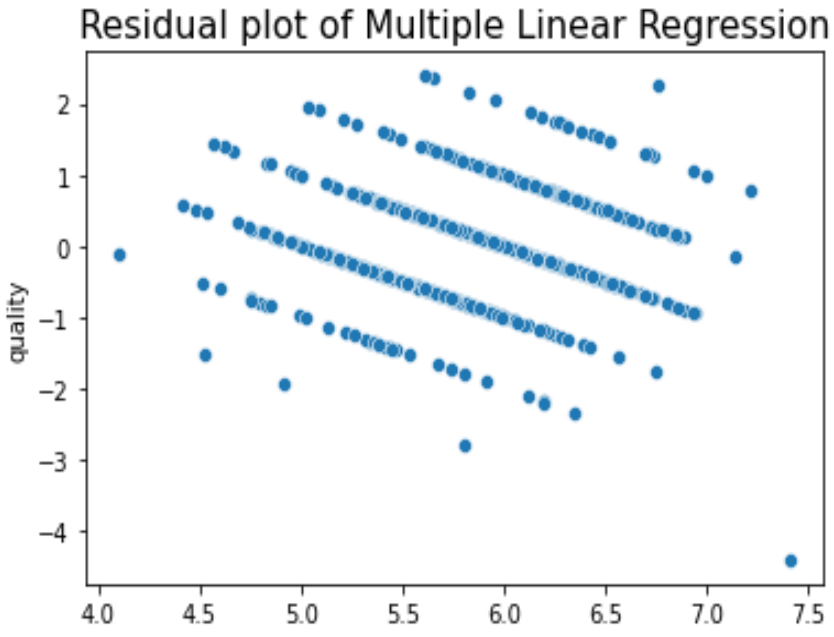
For this model we used all the 11 columns as input and “quality” column as output. After selecting the required column, we divided the data into train and test set with 20% of total data points in the testing set. We now have 3168 rows of training data and 793 rows of test data.

As all of the data were explored before we go directly for model training. We again use Linear Regression model of *scikit learn* for the training purpose.

Exploring the diagnostic plots

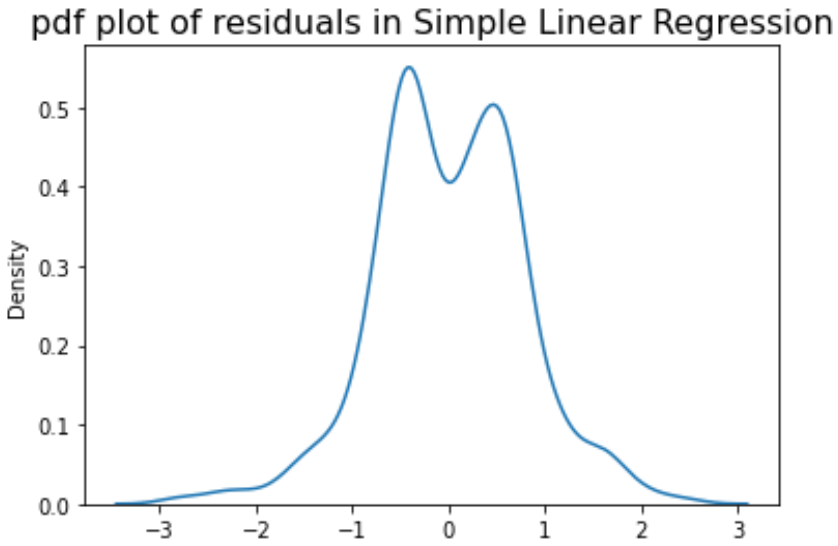
After successfully training both the models, we can now go for checking the assumptions of linear regression. At first, we checked homoscedasticity of residuals using Residual Plot which are presented below –

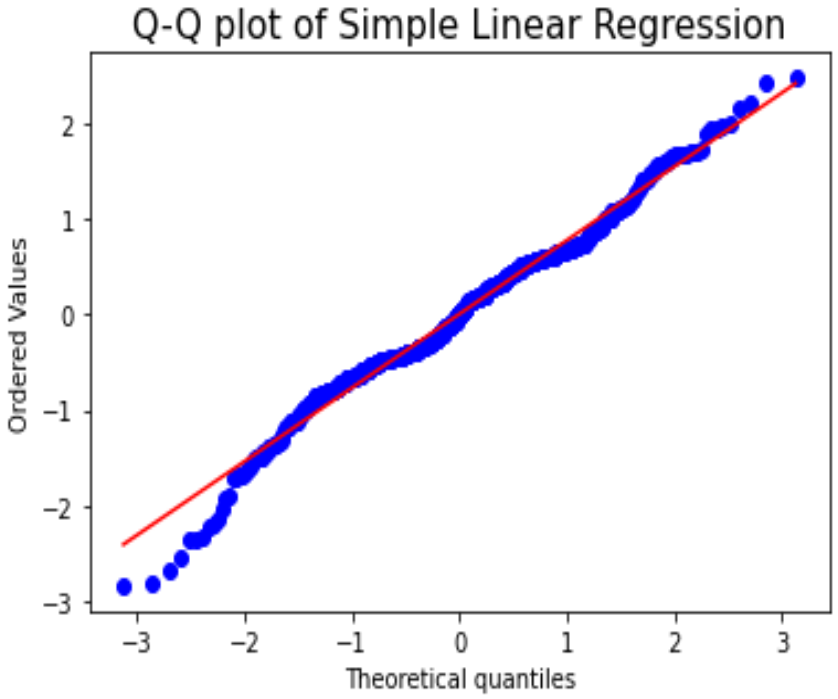


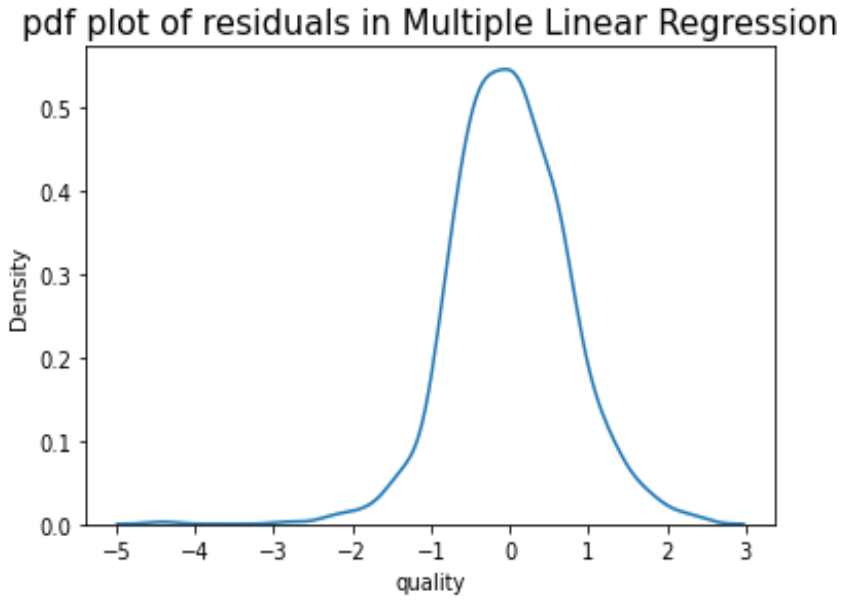


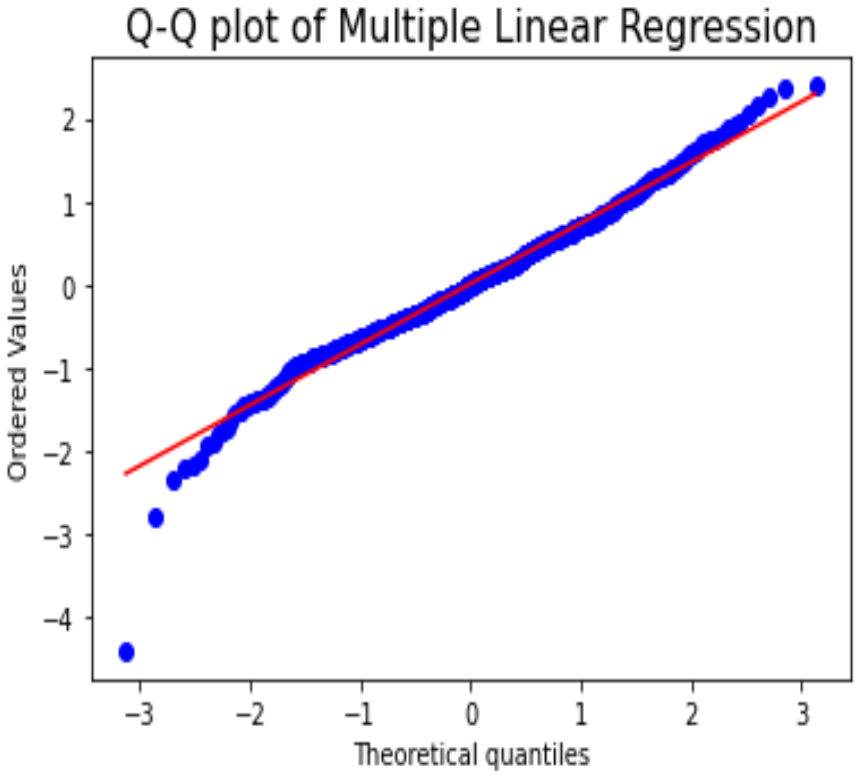
Observing the residual plots, we can assume that homoscedasticity of residuals is maintained.

Again, for checking normality of residuals, we used pdf plot of residuals and Q-Q plot.









After observing the diagnostic plots, we can confirm that normality of residuals assumption is also valid for both the models.

Interpretation of the results

For evaluation purpose of the models, we used three evaluation matrices which are R2 score, adjusted R2 score and Mean Absolute Error. And we find the following performance,

Model	R2 Score	Adjusted R2 Score	MAE
Simple Linear Regression	0.225	0.224	0.617
Multiple Linear Regression	0.304	0.303	0.566

Comparing score in different evaluation matrices, it can be concluded that performance in the Multiple Linear Regression increases from the Simple Linear Regression model.

Again, coefficients and intercept of the hyperplane equation for Multiple Linear Regression Model are listed below –

Intercept of the hyperplane = 132.52

Input feature	Coefficient value
fixed acidity	0.05
volatile acidity	-1.57
citric acid	0.18
residual sugar	0.07
chlorides	-0.27
free sulfur dioxide	0.01
total sulfur dioxide	0.00
Density	-132.32
pH	0.84
sulphates	0.67
alcohol	0.22

From the coefficients, it can be observed the resulting regression equation have best direct relationship with “ph” column and very high indirect relation with “density” column.