Assignment 2: Predicting Wine Quality with Linear Regression

1. What is the distribution of the wine quality scores?

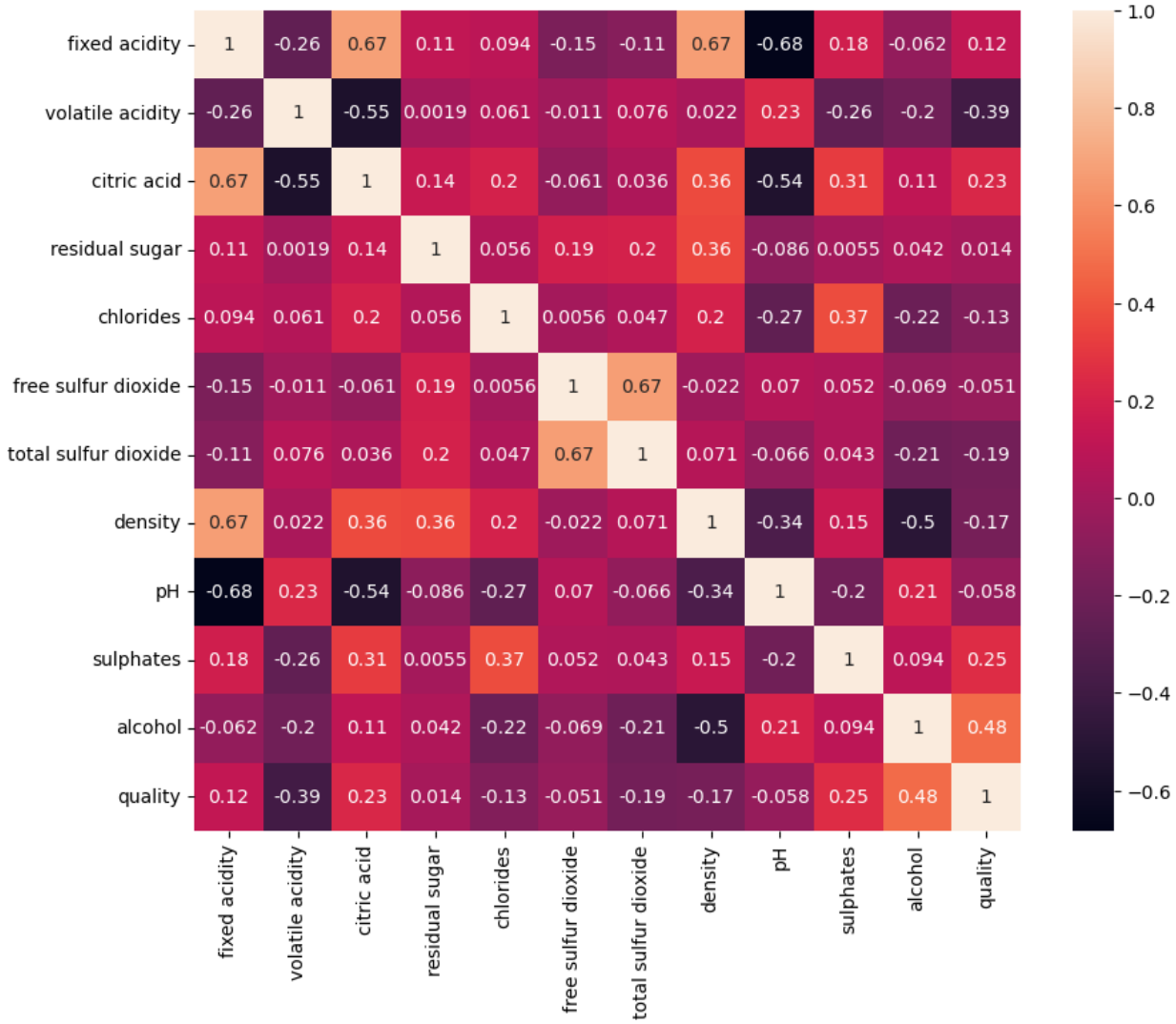
   The values representing the quality of wine are in the range of 3 to 7. The quality of the wine is considered good if the value is above 5 and bad if the value is below 5

   Code : df['quality'].unique()

   Output : array([5, 6, 7, 4, 8, 3], dtype=int64)

2. What are the relationships between the different features?

   Some of the features are moderately correlated, some are decently correlated and some are poorly correlated. Below is a heatmap showing the correlation between various features:



3. Are there any outliers in the data?

Yes. There are outliers in the dataset. I have used boxplots and histograms to identify the outliers and handled them by using both removal and winsorization methods. These methods helped the data with both removal of outliers and to get normally skewed data.

4. What is the accuracy of the linear regression model?

The accuracy of my model for both training and test data is 99.3.

5. What are the most important features for the linear regression model?

All the features - fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, quality together contribute to the prediction of the wine quality. But the important feature which has higher correlation with the target variable is alcohol content.

6. What is the MSE of the linear regression model?

The MSE value by my model for train data prediction is 1.24 and test data prediction is 0.75.

7. What is the R-squared of the linear regression model?

R-squared value of the linear regression model is 0.378

8. How can you improve the performance of the linear regression model?

The performance of the linear regression model can be improved by a more detailed analysis of the features and it's correlation with the target variable. Feature selection and PCA would improve the performance of the model.

9. What are the limitations of the linear regression model?

Linear regression model assumes a linearity between the features which is not possible in the real-world and this would be the drawback of this model.

10. What are the implications of your findings for the real-world problem?

By using this model the quality of the wine can be assessed. This can be used by the food quality check organizations to check the wine which will ultimately help find the adulterated wine.