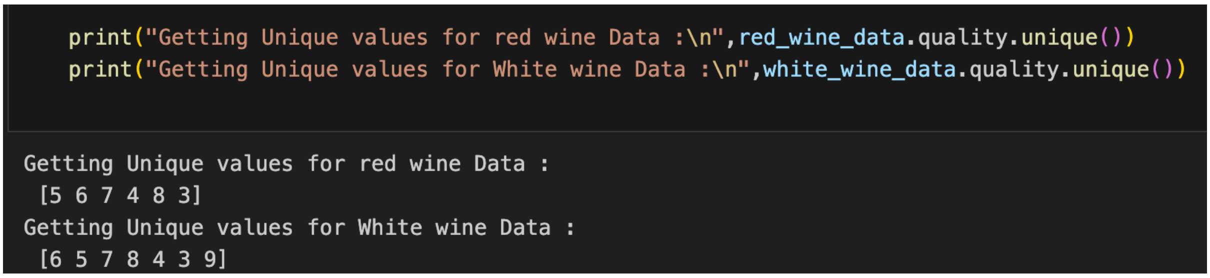
# Assignment 2: Predicting Wine Quality with Linear Regression

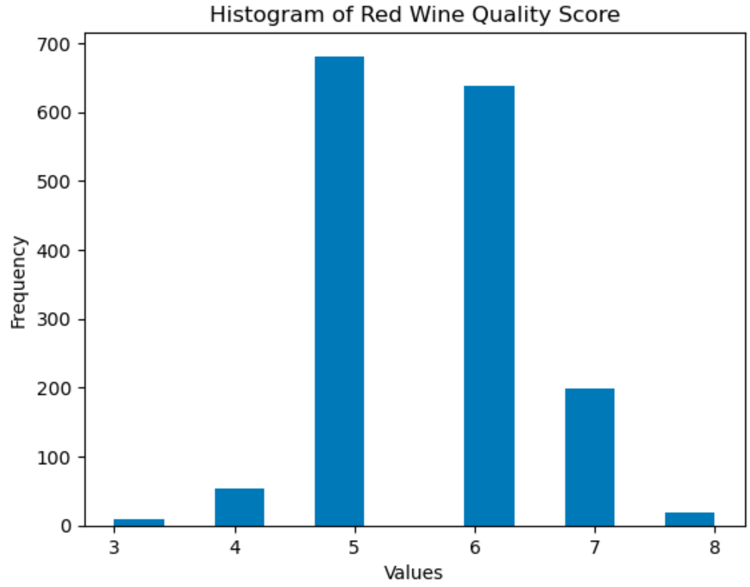
## 1. What is the distribution of the wine quality scores?

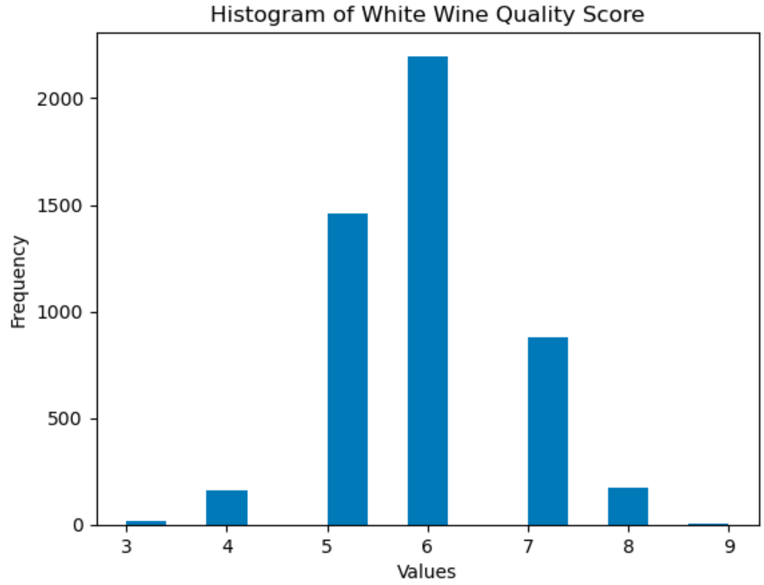
Ans:



```
print("Getting Unique values for red wine Data :\n",red_wine_data.quality.unique())
print("Getting Unique values for White wine Data :\n",white_wine_data.quality.unique())
```

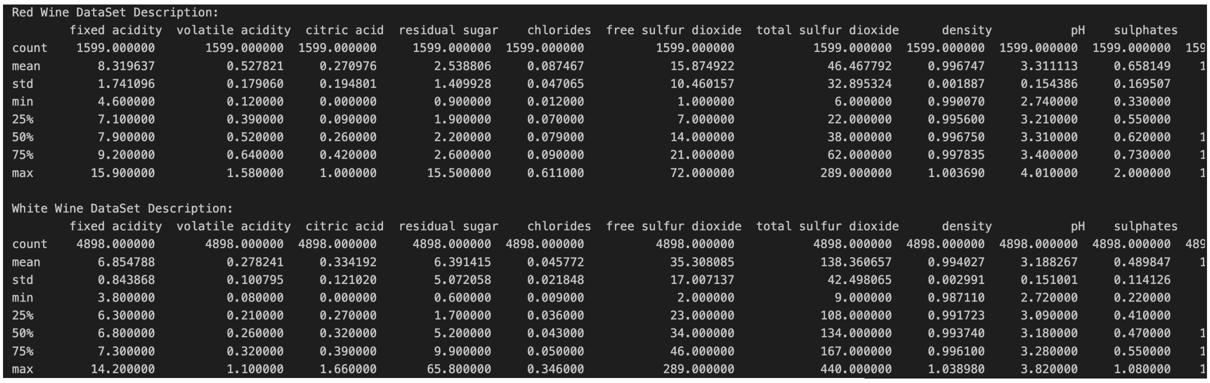
```
Getting Unique values for red wine Data :
 [5 6 7 4 8 3]
Getting Unique values for White wine Data :
 [6 5 7 8 4 3 9]
```





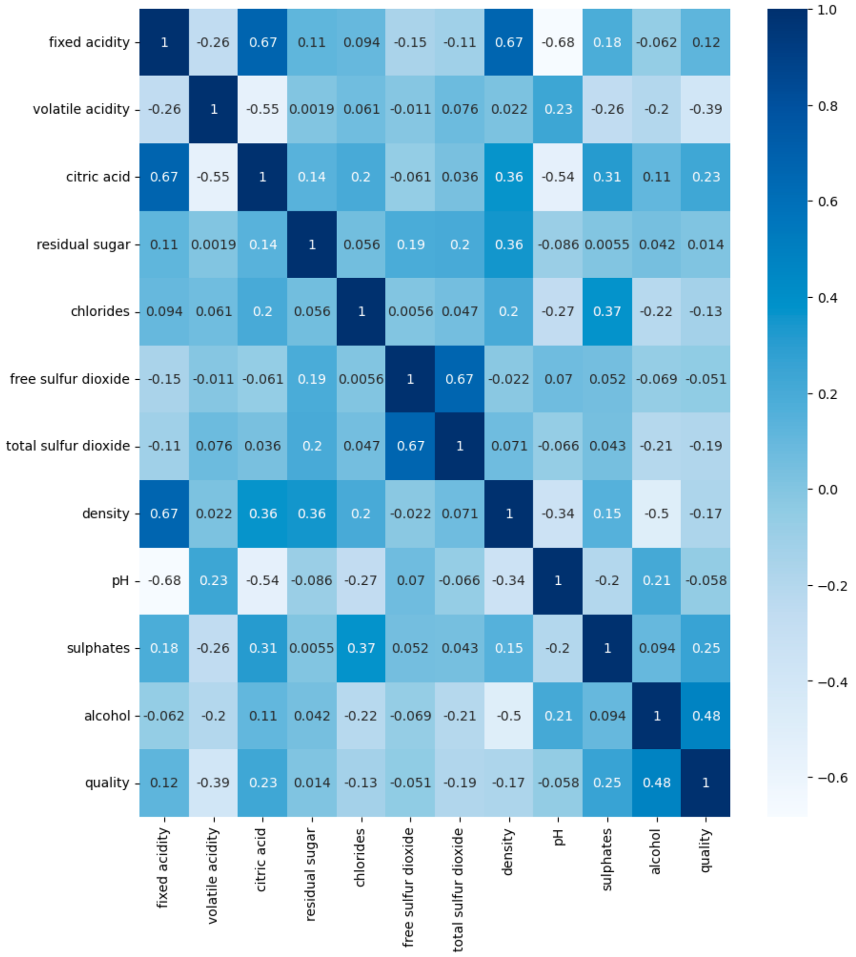
## 2. What are the relationships between the different features?

# Ans:

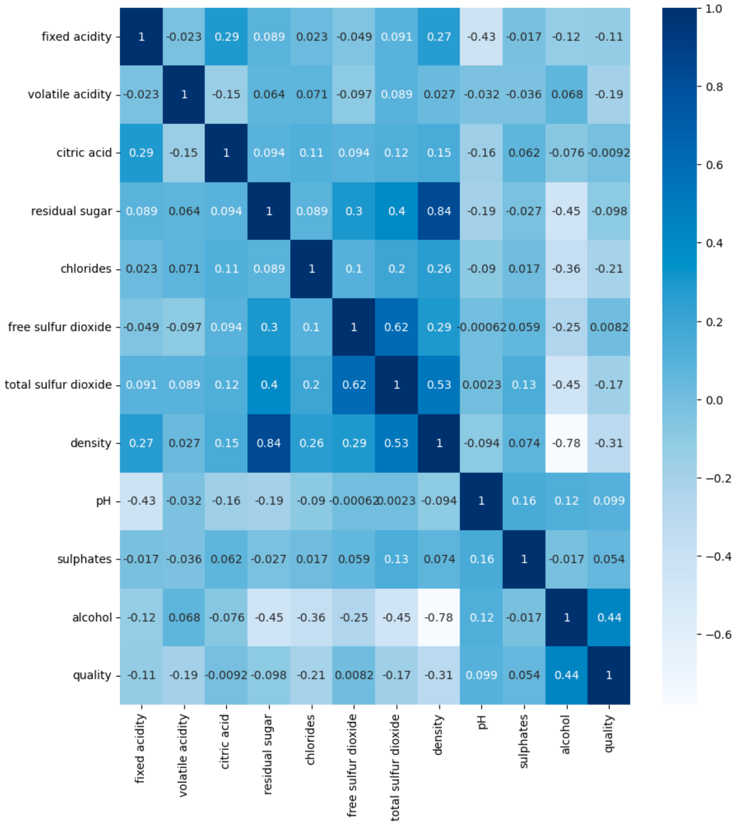


```
Red Wine DataSet Description:
        fixed acidity  volatile acidity  citric acid  residual sugar    chlorides  free sulfur dioxide  total sulfur dioxide      density           pH    sulphates
count     1599.000000       1599.000000  1599.000000     1599.000000  1599.000000          1599.000000           1599.000000  1599.000000  1599.000000  1599.000000  159
mean         8.319637          0.527821     0.270976        2.538806     0.087467            15.874922             46.467792     0.996747     3.311113     0.658149    1
std          1.741096          0.179060     0.194801        1.409928     0.047065            10.460157             32.895324     0.001887     0.154386     0.169507
min          4.600000          0.120000     0.000000        0.900000     0.012000             1.000000              6.000000     0.990070     2.740000     0.330000
25%          7.100000          0.390000     0.090000        1.900000     0.070000             7.000000             22.000000     0.995600     3.210000     0.550000
50%          7.900000          0.520000     0.260000        2.200000     0.079000            14.000000             38.000000     0.996750     3.310000     0.620000    1
75%          9.200000          0.640000     0.420000        2.600000     0.090000            21.000000             62.000000     0.997835     3.400000     0.730000    1
max         15.900000          1.580000     1.000000       15.500000     0.611000            72.000000            289.000000     1.003690     4.010000     2.000000    1

White Wine DataSet Description:
        fixed acidity  volatile acidity  citric acid  residual sugar    chlorides  free sulfur dioxide  total sulfur dioxide      density           pH    sulphates
count     4898.000000       4898.000000  4898.000000     4898.000000  4898.000000          4898.000000           4898.000000  4898.000000  4898.000000  4898.000000  489
mean         6.854788          0.278241     0.334192        6.391415     0.045772            35.308085            138.360657     0.994027     3.188267     0.489847    1
std          0.843868          0.100795     0.121020        5.072058     0.021848            17.007137             42.498065     0.002991     0.151001     0.114126
min          3.800000          0.080000     0.000000        0.600000     0.009000             2.000000              9.000000     0.987110     2.720000     0.220000
25%          6.300000          0.210000     0.270000        1.700000     0.036000            23.000000            108.000000     0.991723     3.090000     0.410000
50%          6.800000          0.260000     0.320000        5.200000     0.043000            34.000000            134.000000     0.993740     3.180000     0.470000    1
75%          7.300000          0.320000     0.390000        9.900000     0.050000            46.000000            167.000000     0.996100     3.280000     0.550000    1
max         14.200000          1.100000     1.660000       65.800000     0.346000           289.000000            440.000000     1.038980     3.820000     1.080000    1
```

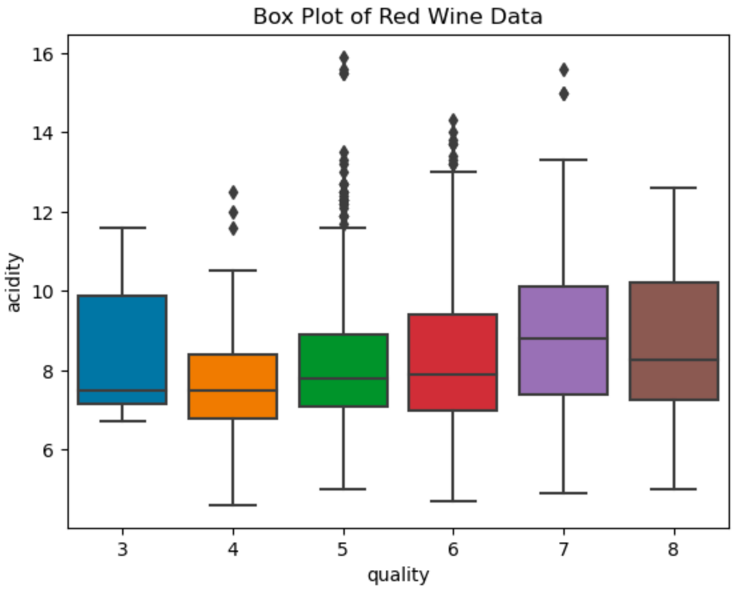


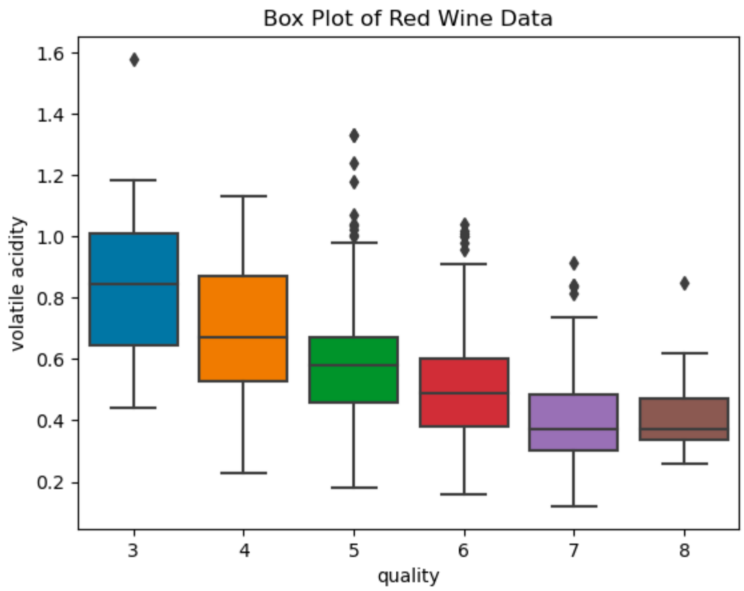
| | fixed acidity | volatile acidity | citric acid | residual sugar | chlorides | free sulfur dioxide | total sulfur dioxide | density | pH | sulphates | alcohol | quality |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| fixed acidity | 1 | -0.26 | 0.67 | 0.11 | 0.094 | -0.15 | -0.11 | 0.67 | -0.68 | 0.18 | -0.062 | 0.12 |
| volatile acidity | -0.26 | 1 | -0.55 | 0.0019 | 0.061 | -0.011 | 0.076 | 0.022 | 0.23 | -0.26 | -0.2 | -0.39 |
| citric acid | 0.67 | -0.55 | 1 | 0.14 | 0.2 | -0.061 | 0.036 | 0.36 | -0.54 | 0.31 | 0.11 | 0.23 |
| residual sugar | 0.11 | 0.0019 | 0.14 | 1 | 0.056 | 0.19 | 0.2 | 0.36 | -0.086 | 0.0055 | 0.042 | 0.014 |
| chlorides | 0.094 | 0.061 | 0.2 | 0.056 | 1 | 0.0056 | 0.047 | 0.2 | -0.27 | 0.37 | -0.22 | -0.13 |
| free sulfur dioxide | -0.15 | -0.011 | -0.061 | 0.19 | 0.0056 | 1 | 0.67 | -0.022 | 0.07 | 0.052 | -0.069 | -0.051 |
| total sulfur dioxide | -0.11 | 0.076 | 0.036 | 0.2 | 0.047 | 0.67 | 1 | 0.071 | -0.066 | 0.043 | -0.21 | -0.19 |
| density | 0.67 | 0.022 | 0.36 | 0.36 | 0.2 | -0.022 | 0.071 | 1 | -0.34 | 0.15 | -0.5 | -0.17 |
| pH | -0.68 | 0.23 | -0.54 | -0.086 | -0.27 | 0.07 | -0.066 | -0.34 | 1 | -0.2 | 0.21 | -0.058 |
| sulphates | 0.18 | -0.26 | 0.31 | 0.0055 | 0.37 | 0.052 | 0.043 | 0.15 | -0.2 | 1 | 0.094 | 0.25 |
| alcohol | -0.062 | -0.2 | 0.11 | 0.042 | -0.22 | -0.069 | -0.21 | -0.5 | 0.21 | 0.094 | 1 | 0.48 |
| quality | 0.12 | -0.39 | 0.23 | 0.014 | -0.13 | -0.051 | -0.19 | -0.17 | -0.058 | 0.25 | 0.48 | 1 |

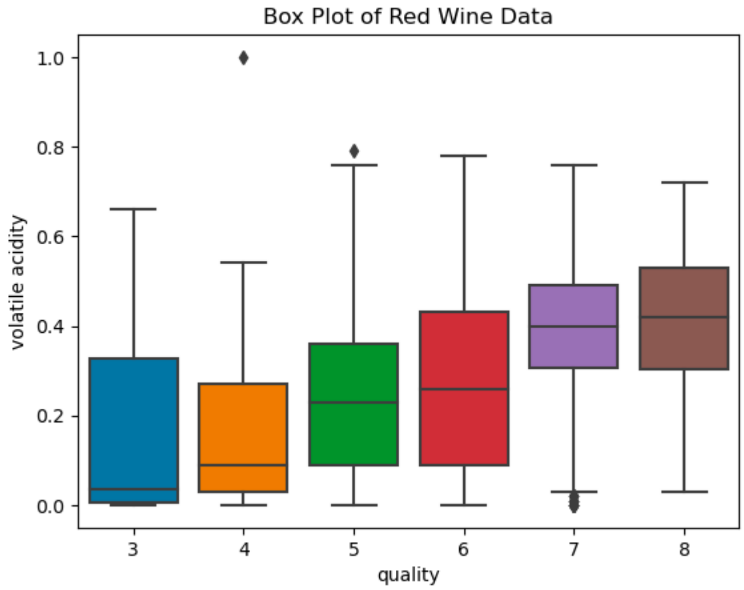


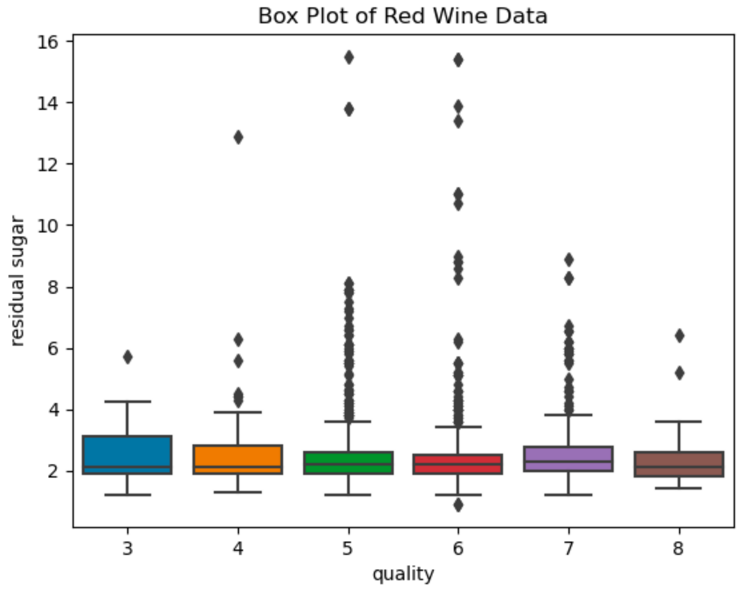
# 3. Are there any outliers in the data?

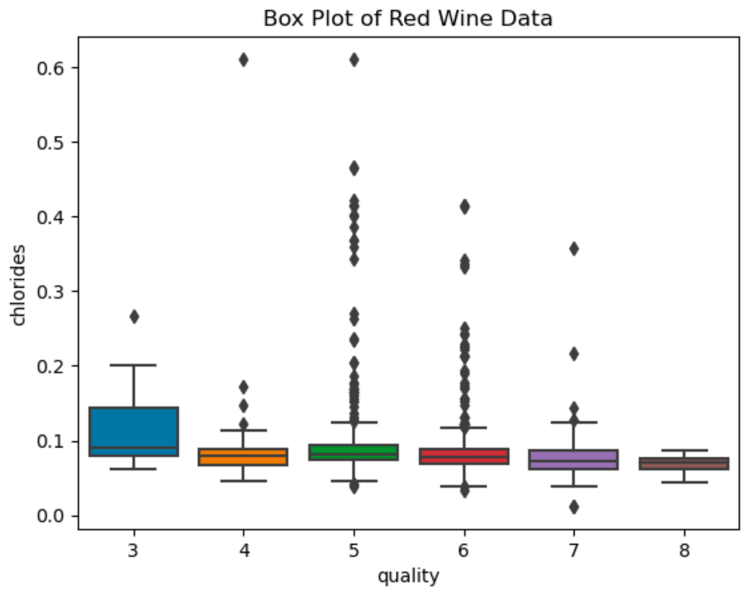
Ans:

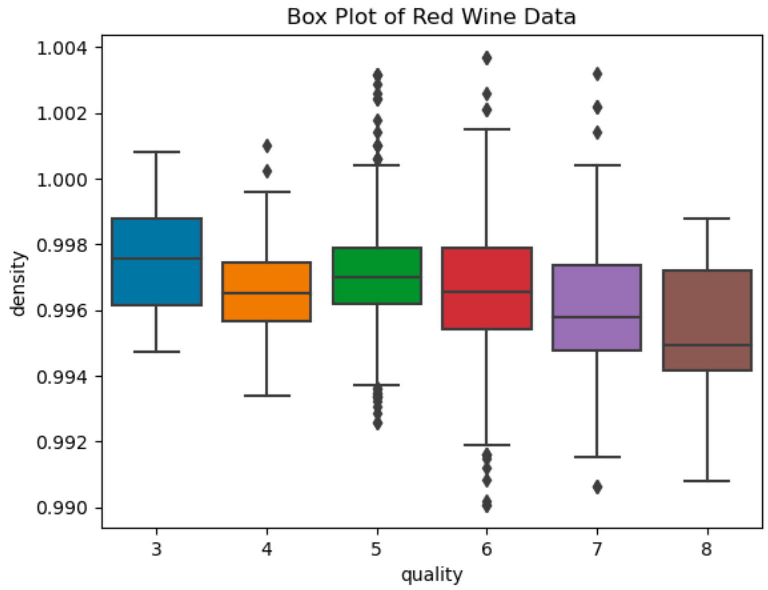



Box Plot of Red Wine Data
volatile acidity
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
3
4
5
6
7
8
quality


Box Plot of Red Wine Data
volatile acidity
1.0
0.8
0.6
0.4
0.2
0.0
3
4
5
6
7
8
quality

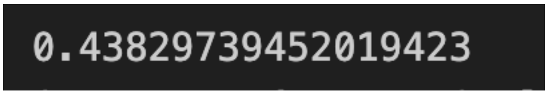

Box Plot of Red Wine Data
residual sugar
16
14
12
10
8
6
4
2
3
4
5
6
7
8
quality





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

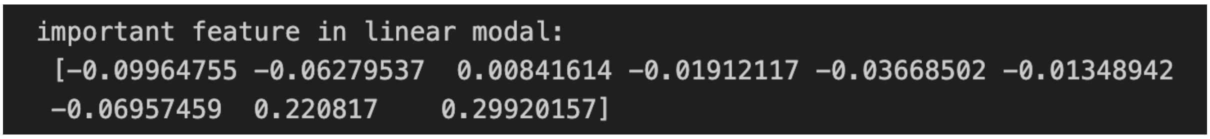
Ans:



```
0.43829739452019423
```

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

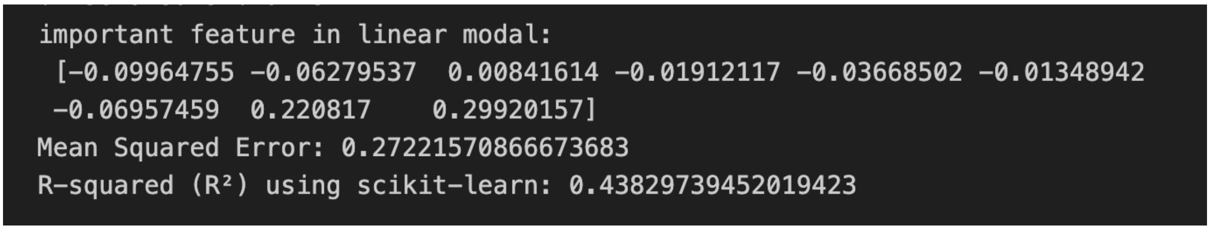
Ans



```
important feature in linear modal:
 [-0.09964755 -0.06279537  0.00841614 -0.01912117 -0.03668502 -0.01348942
 -0.06957459  0.220817    0.29920157]
```

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

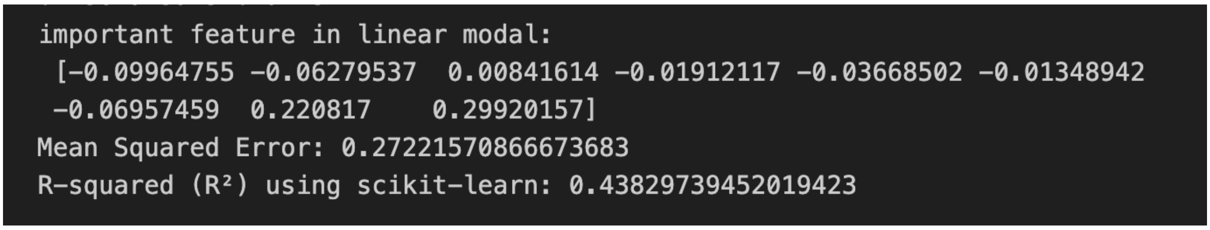
Ans:



```
important feature in linear modal:
 [-0.09964755 -0.06279537  0.00841614 -0.01912117 -0.03668502 -0.01348942
 -0.06957459  0.220817    0.29920157]
Mean Squared Error: 0.27221570866673683
R-squared (R²) using scikit-learn: 0.43829739452019423
```

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

Ans:



```
important feature in linear modal:
 [-0.09964755 -0.06279537  0.00841614 -0.01912117 -0.03668502 -0.01348942
 -0.06957459  0.220817    0.29920157]
Mean Squared Error: 0.27221570866673683
R-squared (R²) using scikit-learn: 0.43829739452019423
```

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

Ans: firstly by normalizing the data

2. by removing the outliers

3. by feature engineering

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

Ans: first and foremost the we need to make an assumption that the relationship between dependent and independent variable is linear in nature.

2.Just like in our case the quality can be divided into 3-4 categories easily but as we are using linear modal it is difficult to handle categorical data

3.it might be possible that the modal will not hold onto the prediction made outside the observed modal

4.In linear modal the outliers and other kind of noise effects the modal adversely

5. Overfitting can be problematic in linear modal as if we try to increase the degree of the features how well it will hold for the future predictions

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

Ans: We can use this modal to standardise the production of the wine based on our finding i.e: what should be the appropriate acidity level or alcohol levels and like wise we can standardise the percentage of the component involved

2. second use case that I can see is how the wine aging is affected based on its component and quality and weather any component is effecting it’s taste or not.

3. Lastly if we are introducing any new product in the future these finding can act as a baseline for the new Product