10/12/2023

Exam Project

Red Wine Quality Data Analysis

Submitted By
Iftekhar Hossain

Table of Contents

1. Introduction

Red wine is an expensive beverage. It is essential to know the quality of the wine before ordering it, especially before ordering a large quantity of red wine. Otherwise, it can cause a considerable loss of money and customers. A business will be far more likely to purchase high-quality wine and continue operating successfully if it can develop a scientific approach for assessing the quality of red wine based on its physicochemical features. The certification procedure frequently includes quality evaluation, which can be used to stratify wines like premium brands (helpful for pricing) and enhance winemaking by identifying the most important features.

We need to do a data analysis to know the effect of different features (physicochemical properties) on red wine quality, for example, which features are more important than others for sensual quality rating, exploring different trends, patterns, and correlations in the wine's physicochemical characteristics, and finding the perfect proportion of the materials to maximize the quality.

2. Initial assumptions and hypotheses

The dataset includes a variety of data about red wine. The following table (Table 2.1) shows a list of variables and their description used in the dataset and for the analysis.
All the twelve variables/features are presented in numeric variables.

Features	**Role**	**Description of the features. (units)**
Fixed acidity	Feature	Mainly tartaric, citric, malic and succinic. They do not evaporate immediately. (g/dm^3)
Volatile acidity	Feature	The wine's acetic acid amount which high amount produces an unpleasant vinegar flavor. (g/dm^3)
Citric acid	Feature	Which gives flavor and freshness in modest amounts. (g/dm^3)
Residual sugar	Feature	The quantity of sugar that is left over once fermentation has ended. (g/dm^3)
Chlorides	Feature	how much salt (sodium chloride) is in the wine. (g/dm^3)
Free sulfur dioxide	Feature	It prevents oxidation of wine and microbial growth. (mg/dm^3)
Total sulfur dioxide	Feature	The total amount of free + bound forms of SO2. In low dilutions, it is generally unnoticeable in wine. However, at free SO2 absorptions over 50 ppm, SO2 becomes apparent in the nose and in taste. (mg/dm^3)
Density	Feature	varying on the proportion of alcohol and sugar substance. sweeter wines are generally higher in density. (g/cm^3)
pH	Feature	measure of how acidic wine is (pH scale value)

Sulphates	Feature	Additive (potassium sulphate) that can aid to (SO^2) sulfur dioxide gas amounts. (g/dm^3)
Alcohol	Feature	percentage of alcohol substance. (% vol.)
Quality	Target	A score between 0 and 10 which was assessed on human sensory data.

Table 2.1: List of all the variables and their description

2.1 Data quality and data preprocessing

Missing Values: No missing values found.

Duplicate values: 240 duplicate values/observation (rows) have been detected and removed. After that 1359 values are remaining.

Outliers: Outlying observation has been removed by both Z-score methods (for features' citric acid, density, and pH since their distribution was approximately normally distributed). The rest of the features outliers have been removed by the IQR method since the rest of the features' values are skewed either more than +0.5 or less than -0.5 of skewness value. If the skewness value is between -0.5 and 0.5, then the feature distribution is approximately symmetrical. One thousand thirty-one observations (rows) are remaining after removing the outliers.

Feature scaling: I have done feature scaling before the regression. Still, in the regression with or without feature scaling, the performance is almost the same as the MS Excel Data Analysis Tools Regression.

2.2 Statistical summary of the data

With Outliers	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
Count	1359	1359	1359	1359	1359	1359	1359	1359	1359	1359	1359	1359
Min	4.6	0.12	0	0.9	0.012	1	6	0.99	2.74	0.33	8.4	3
Max	15.9	1.58	1	15.5	0.611	72	289	1.004	4.01	2	14.9	8
Range	11.3	1.46	1	14.6	0.599	71	283	0.014	1.27	1.67	6.5	5
Mean	8.3	0.5	0.3	2.5	0.1	15.9	46.8	1.0	3.3	0.7	10.4	5.6
Median	7.9	0.52	0.26	2.2	0.079	14	38	0.997	3.31	0.62	10.2	6
Mode	7.2	0.5	0	2	0.08	6	28	0.997	3.3	0.54	9.5	5
stdev (σ)	1.7	0.2	0.2	1.4	0.0	10.4	33.4	0.0	0.2	0.2	1.1	0.8

Q1	7.1	0.39	0.09	1.9	0.07	7	22	0.996	3.21	0.55	9.5	5
Q3	9.2	0.64	0.43	2.6	0.091	21	63	0.998	3.4	0.73	11.1	6

Table 2.2.1: presents a statistical summary of the data before removing the outliers.

After removing Outliers	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
Count	1031	1031	1031	1031	1031	1031	1031	1031	1031	1031	1031	1031
Min	5.1	0.12	0	1.2	0.039	1	6	0.9915	2.88	0.33	8.7	3
Max	12.3	1.01	0.73	3.65	0.122	42	124	1.001	3.72	0.98	13.5	8
Range	7.2	0.89	0.73	2.45	0.083	41	118	0.0095	0.84	0.65	4.8	5
Mean	8.158972	0.52274	0.249534	2.194811	0.078395	14.91465	42.20757	0.996528	3.323453	0.630902	10.39796	5.637245
Median	7.8	0.52	0.24	2.1	0.078	13	36	0.99655	3.32	0.61	10.1	6
Mode	7.2	0.58	0	2	0.08	6	28	0.998	3.36	0.58	9.5	5
stdev (σ)	1.488317	0.16676	0.18263	0.453732	0.014991	8.835082	26.5743	0.001641	0.136563	0.114906	1.003188	0.776928
Q1	7.1	0.39	0.08	1.9	0.069	8	22	0.9955	3.23	0.55	9.5	5
Q3	9	0.63	0.4	2.5	0.087	20	56	0.9976	3.41	0.7	11.08333	6
Skew	0.686652	0.311858	0.3039	0.599249	0.226293	0.842348	0.987611	-0.00784	0.077565	0.628241	0.760384	0.307078

Table 2.2.2: Statistical summary of the data after removing the outliers.

Statistical summary of the best quality wine (7 & 8):

Let’s extract only the wine observations, rated as 7 and 8, and see their statistical summary.

Best Quality Wine	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
No. of observations	128	128	128	128	128	128	128	128	128	128	128	128

Min	5.1	0.12	0	1.2	0.041	3	7	0.99235	2.92	0.47	9.5	7
Max	12	0.85	0.66	3.65	0.121	42	106	1.0002	3.71	0.94	13.4	8
Range	6.9	0.73	0.66	2.45	0.08	39	99	0.00785	0.79	0.47	3.9	1
Mean	8.61	0.41	0.35	2.22	0.07	13.27	30.55	1.00	3.29	0.73	11.49	7.09
Median	8.5	0.375	0.38	2.2	0.07	12	25.5	0.99535	3.3	0.73	11.5	7
Mode	9.1	0.31	0.39	2.4	0.066	6	10	0.9968	3.23	0.76	11.7	7
stdev (σ)	1.66	0.14	0.18	0.50	0.02	8.43	20.45	0.00	0.14	0.10	0.91	0.28
Q1	7.38	0.31	0.30	1.80	0.06	6.00	16.00	0.99	3.20	0.65	10.90	7.0
Q3	9.90	0.51	0.47	2.50	0.08	16.25	37.25	1.00	3.37	0.80	12.13	7.00

Table 2.2.3: Statistical summary of the best quality wine (which were rated as 7 and 8)

2.3 Initial assumptions and hypotheses

Based on the initial insights from the study by (Cortez et al., 2009) in the Red Wine Quality dataset, I have the following assumptions and hypotheses:

1. Since the amount of fixed acidity is the highest among all other types of acid (volatile acidity and citric acid (Table 2.2.2)) and since it does not evaporate readily (non-volatile). It will dominate the pH level. An increase in fixed acidity will make the wine sourer. Also, it could decrease the wine quality.
2. Since the high amount of volatile acidity in wine (acetic acid) produces an unpleasant vinegar flavor. The increase of this will decrease the wine quality.
3. A modest amount of citric acid in wine gives flavor and freshness. Less or too high will decrease the wine quality. Citric acid presence in wine is minimal (Table 2.2.2: Statistical summary). An increase of citric acid will increase the wine quality.
4. pH value of any acidic solutions is less than 7. The more it is acidic, the lesser the pH score. An increase in fixed acidity, volatile acidity, and citric acid will decrease the pH score/value. A low score of pH value is very acidic and sour, thus poor in quality.
5. Residual sugar is the amount of remaining sugar after the fermentation has stopped. Alcohol is less dense than water. An increase in residual sugar will increase the density of the wine. The more residual sugar

will be more watery wine with a lesser alcohol percentage. An increase in the amount of residual sugar will decrease the wine quality.

6. Chlorides is the amount of salt in the wine. An increase in chlorides means saltier. So, an increase in chloride could be a decrease in quality.
7. The density of alcohol is less than water. So, an increase in density means more watery wine. The increase in density will decrease the quality. At room temperature, the alcohol density is around 0.79 g/cm3, and water density is around 0.99 g/cm3.
8. Alcohol is the main ingredient of wine. So, an increase in alcohol percentage will increase the quality. Customers also do not expect a very high percentage of alcohol in the red wine. We should look forward to determining which alcohol percentage yields the best result.

3. Exploratory Data analysis (EDA)

This section presents a detailed analysis of the exploratory data. Here, I show all the different analyses of the variables, their correlations, and finally, the regression analysis to predict the quality.

3.1 Different Features/Variables Distribution

Fixed acidity:

Mean	Median	Mode	Standard Deviation	Kurtosis	Skewness	Range	Minimum	Maximum
8.16	7.80	7.20	1.49	-0.04	0.69	7.20	5.10	12.30

Table 3.1.1: Statistical summary of fixed acidity variable.

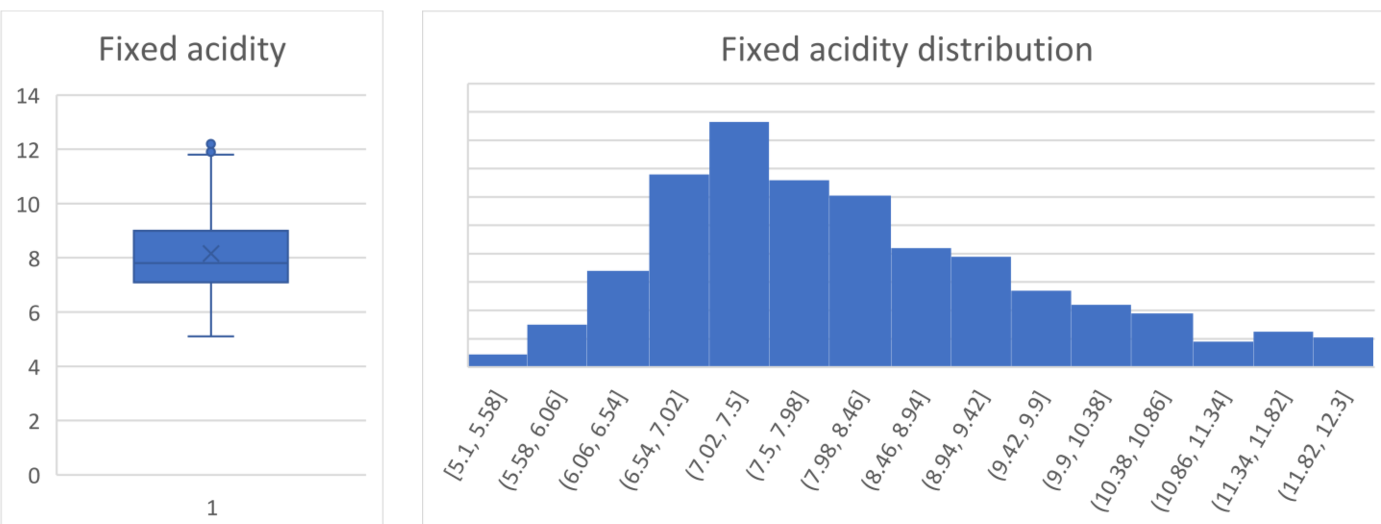


Figure 3.1.1: Fixed acidity distribution

- The fixed acidity distribution is slightly right-skewed and centered around 8 (g/dm3).
- In a fixed acidity variable, the mean is greater than the median.

Volatile acidity:

Mean	Median	Mode	SD (σ)	Kurtosis	Skewness	Range	Minimum	Maximum
0.52	0.52	0.58	0.17	-0.27	0.31	0.89	0.12	1.01

Table 3.1.2: Statistical summary of volatile acidity.

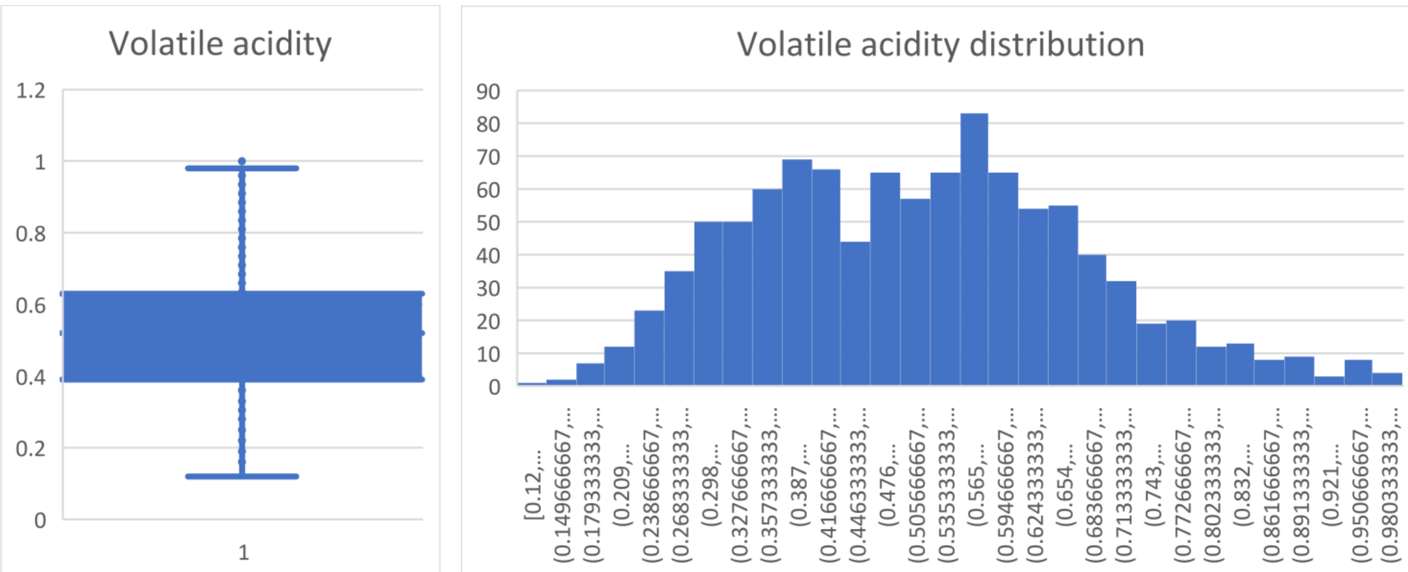


Figure 3.1.2: Volatile acidity distribution

- volatile acidity exists in red wines with an amount mean of 0.52 (g/dm^3).
- The distribution looks bimodal at (around 0.39 and 0.57).

Citric acid:

Mean	Median	Mode	SD (σ)	Kurtosis	Skewness	Range	Minimum	Maximum
0.25	0.24	0	0.18	-0.95	0.30	0.73	0	0.73

Table 3.1.3: Statistical summary of citric acid

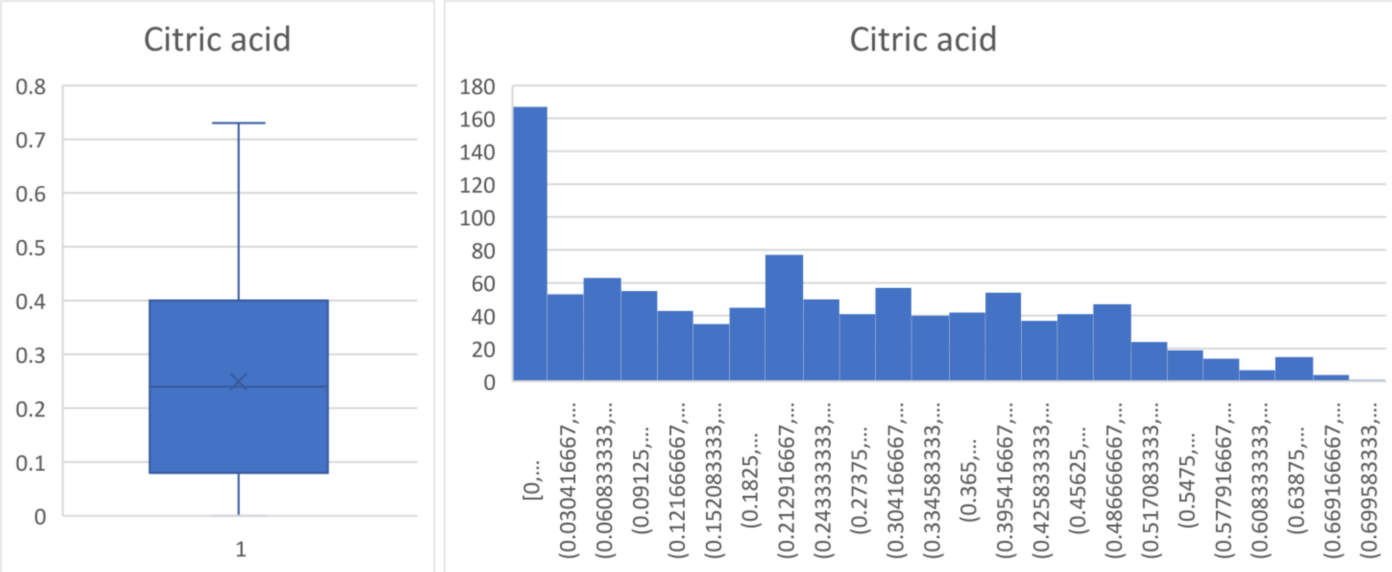


Figure 3.1.3: Citric acid distribution.

- The mode value of the citric acid distribution is 0.
- In red wine, citric acid is present in a very small amount.
- Around 9% of red wine has no citric acid in it.

Residual sugar:

Mean	Median	Mode	SD(σ)	Kurtosis	Skewness	Range	Minimum	Maximum
2.19	2.1	2	0.45	0.30	0.60	2.45	1.2	3.65

Table 3.1.4: Statistical summary of residual sugar.

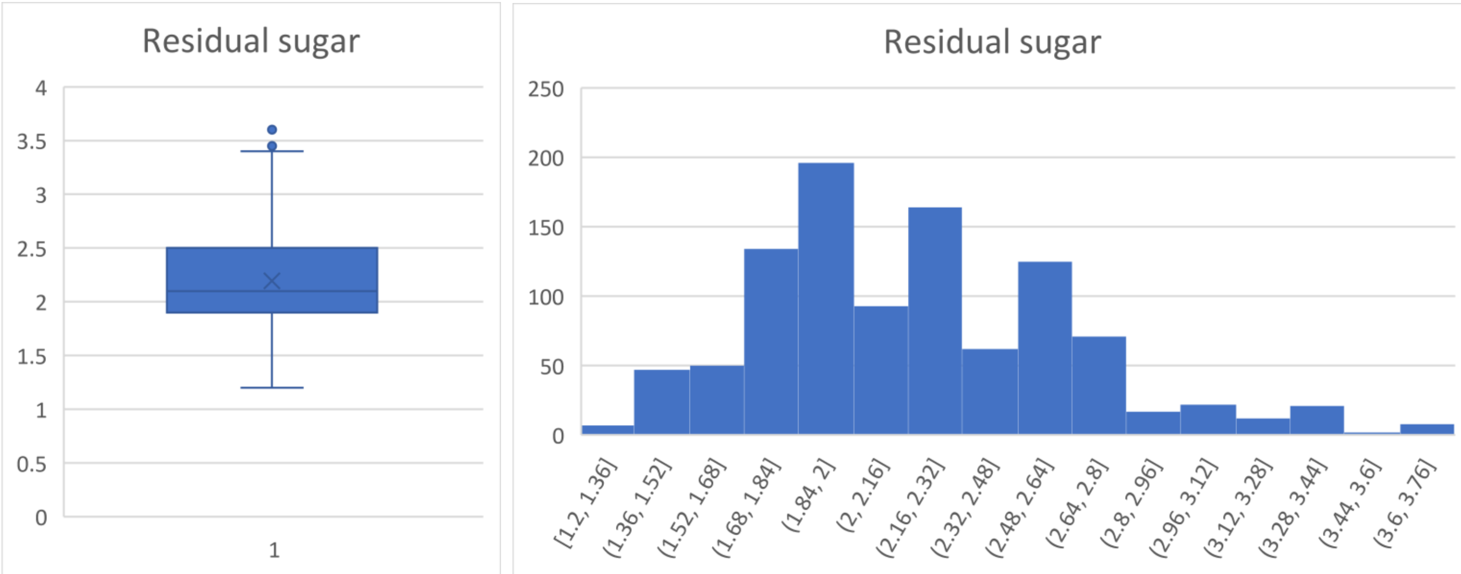


Figure 3.1.4: Residual sugar distribution.

- Residual sugar's Mean 2.2, Medium 2.1, Mode 2 and 75% data is within the range of 2.5 (g/dm3).

Chlorides:

Mean	Median	Mode	SD(σ)	Kurtosis	Skewness	Range	Minimum	Maximum
0.078	0.078	0.080	0.015	0.263	0.227	0.083	0.039	0.122

Table 3.1.5: Statistical summary of chlorides.

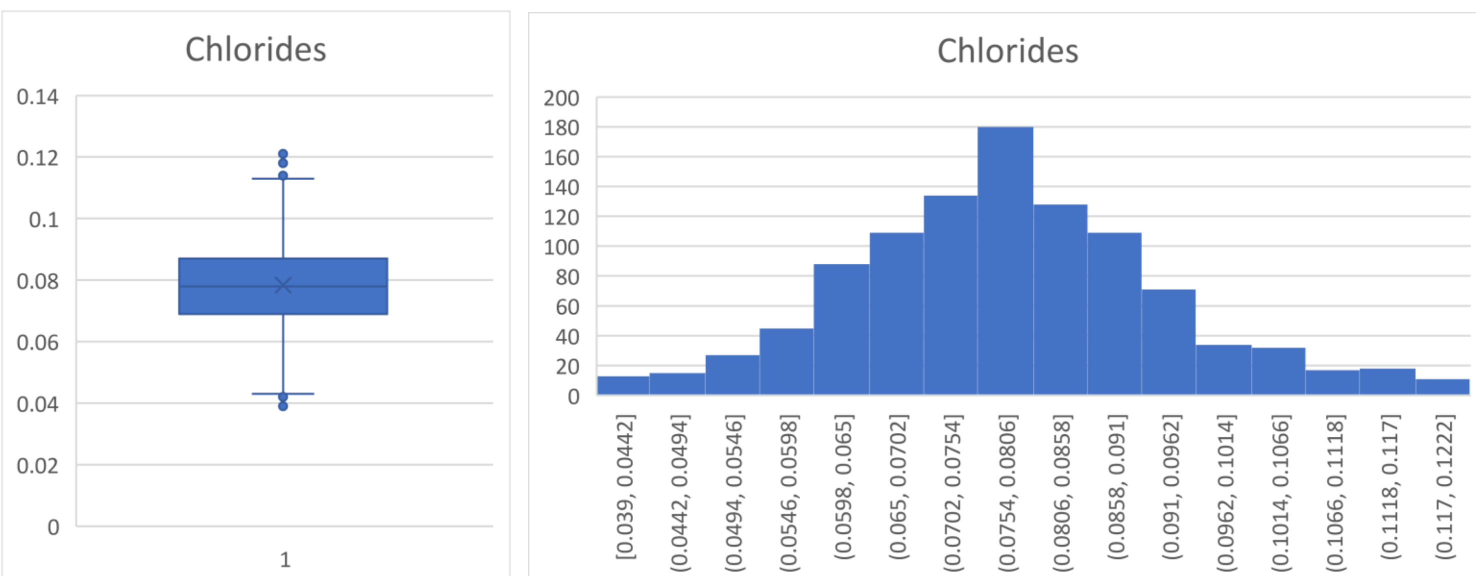


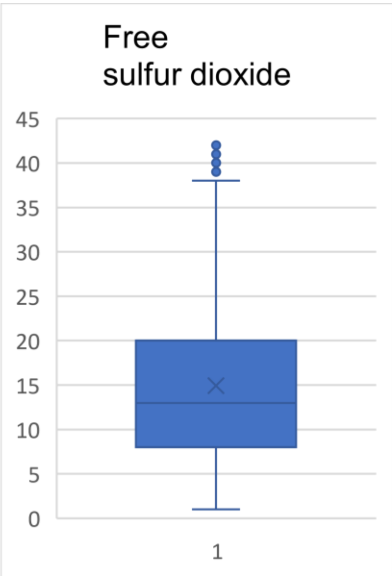
Figure 3.1.5: Chlorides distribution

- Chlorids distribution is normally distributed.
- Around 75% of wine has salt (sodium chloride) less than 0.09 (g/dm3).

Free sulfur dioxide:

Mean	Median	Mode	SD (σ)	Kurtosis	Skewness	Range	Minimum	Maximum
14.91	13	6	8.84	0.05	0.84	41	1	42

Table 3.1.6: Statistical summary of free sulfur dioxide



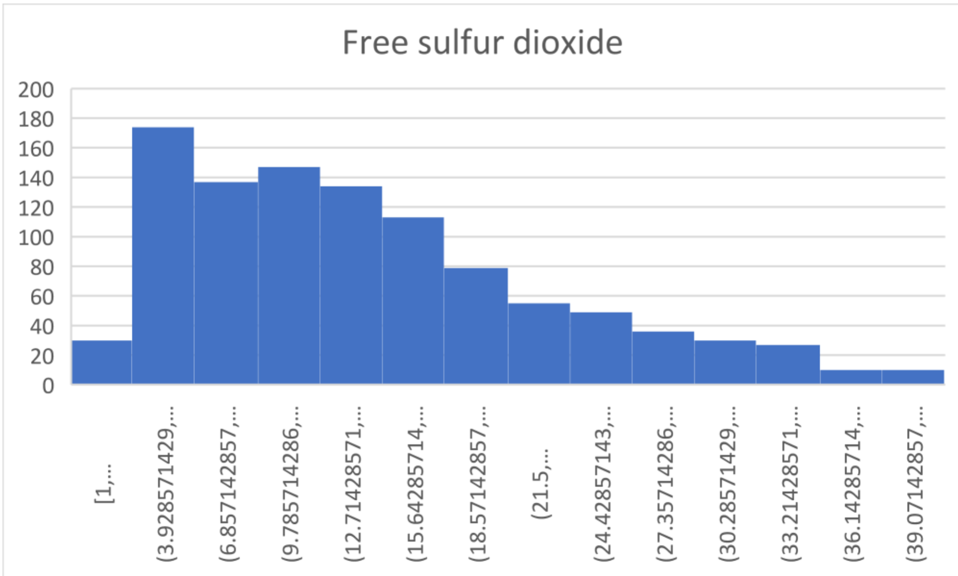


Figure 3.1.6: Free sulfur dioxide distribution

- The distribution of free sulfur dioxide is moderately right skewed.
- 75% of data is within the range of 20 (mg/dm^3).

Total sulfur dioxide:

Mean	Median	Mode	SD(σ)	Kurtosis	Skewness	Range	Minimum	Maximum
42.21	36	28	26.59	0.33	0.99	118	6	124

Table 3.1.7: Statistical summary of total sulfur dioxide.

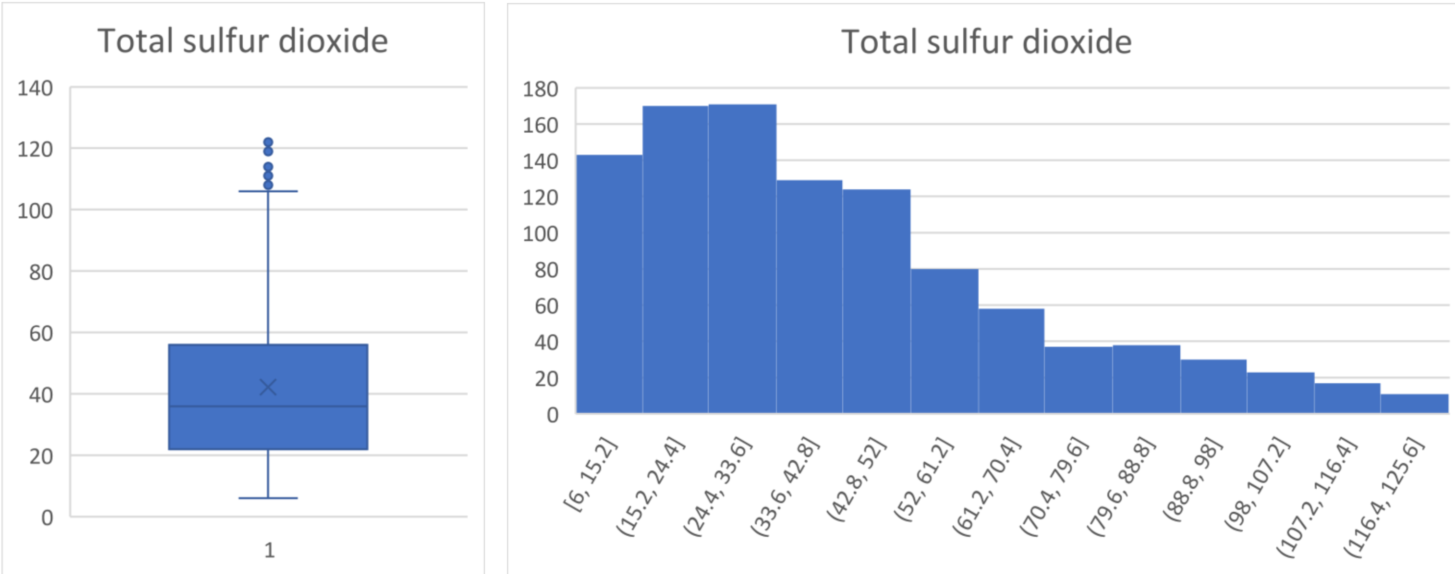


Figure 3.1.7: Total sulfur dioxide distribution

- Total sulfur dioxide's distribution is right skewed.
- 75% of data is within the range of 56 (mg/dm^3).

Density:

Mean	Median	Mode	SD (σ)	Kurtosis	Skewness	Range	Minimum	Maximum
0.997	0.997	0.998	0.002	0.076	-0.008	0.009	0.992	1.001

Table 3.1.8: Statistical summary of density.

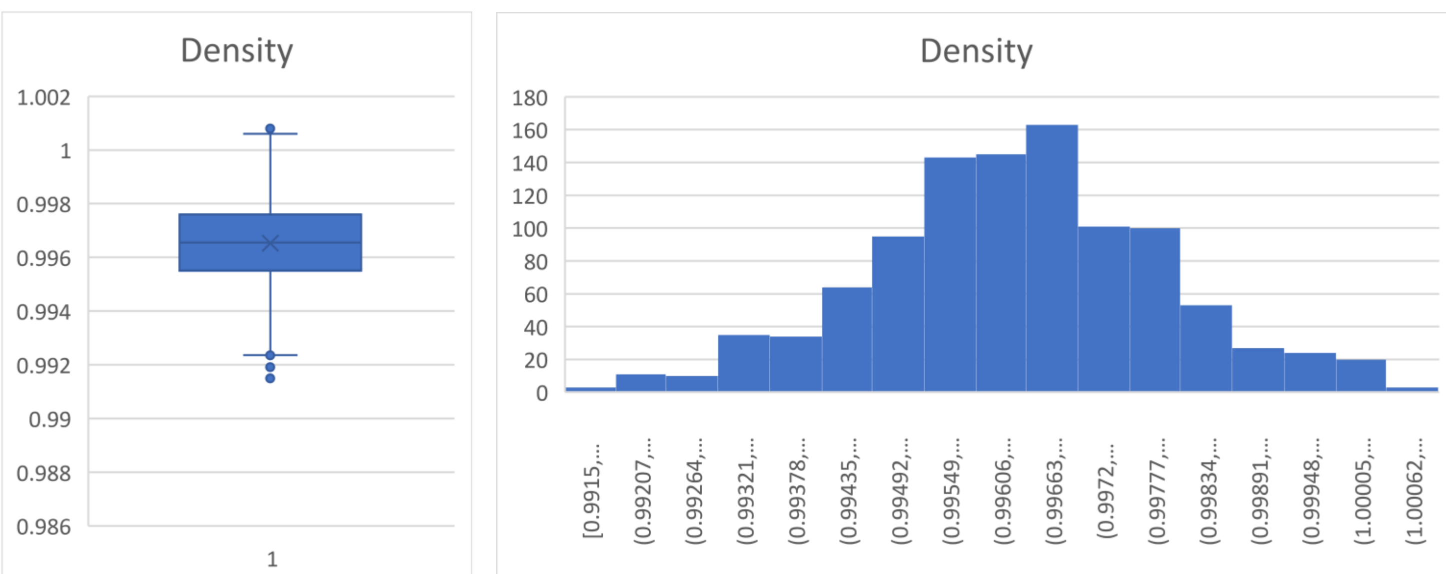


Figure 3.1.8: Density distribution.

- Density feature is normally distributed.

pH:

Mean	Median	Mode	SD(σ)	Kurtosis	Skewness	Range	Minimum	Maximum
3.32	3.32	3.36	0.137	0.020	0.078	0.84	2.88	3.72

Table 3.1.9: Statistical summary of pH.

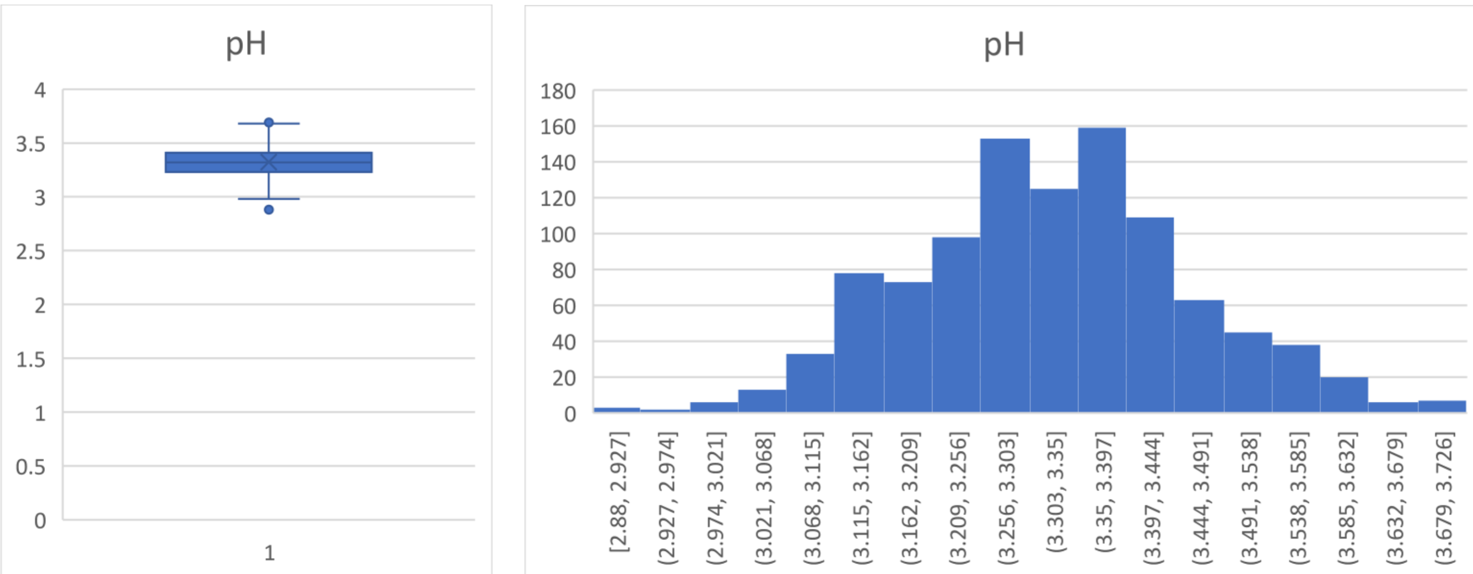


Figure 3.1.9: pH distribution

- pH feature is normally distributed.

Sulphate:

Mean	Median	Mode	SD (σ)	Kurtosis	Skewness	Range	Minimum	Maximum
0.631	0.610	0.580	0.115	0.074	0.629	0.650	0.330	0.980

Table 3.1.10: Statistical summary of sulphates.

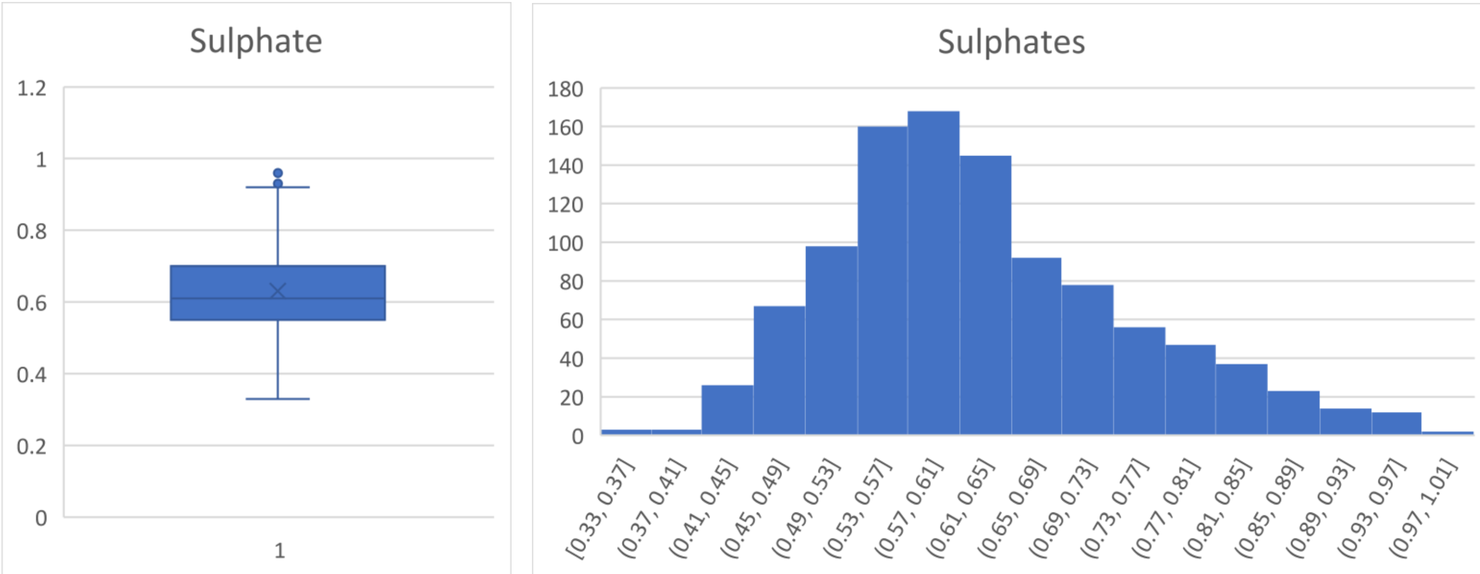


Figure 3.1.10: Sulphate distribution.

- Sulphate is slightly right skewed.
- The mean, median and mode is around 0.6. and 75% of data is within the value of 7 (g/dm^3).

Alcohol:

Mean	Median	Mode	SD	Kurtosis	Skewness	Range	Minimum	Maximum
10.40	10.1	9.5	1.00	-0.26	0.76	4.8	8.7	13.5

Table 3.1.11: Statistical summary of alcohol

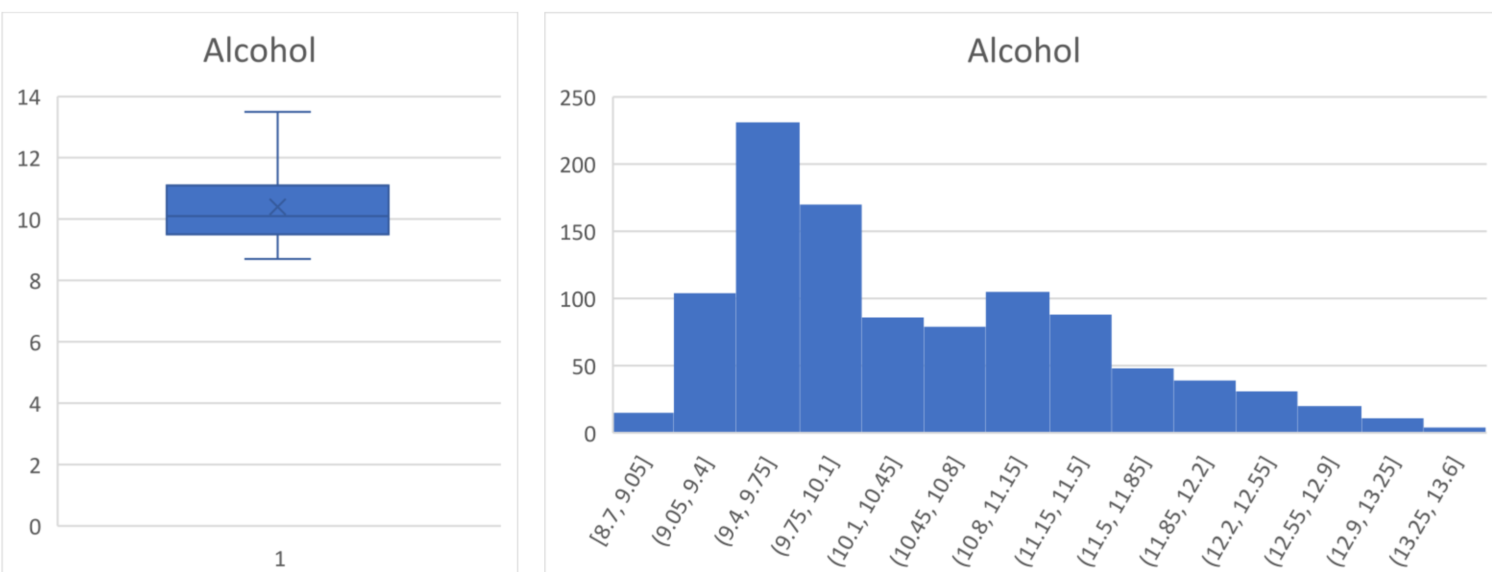


Figure 3.1.11: Alcohol distribution.

- Alcohol features distribution is slightly right skewed.

Quality:

Mean	Median	Mode	SD	Kurtosis	Skewness	Range	Minimum	Maximum
5.64	6	5	0.78	0.22	0.31	5	3	8

Table 3.1.12: Statistical summary of quality.

Figure 3.1.12: Quality distribution

- Most of the wine quality scored either 5 or 6.

3.2 Quality and Different variables

Observations after removing outliers	
Quality	Number of Observations
3	3
4	33
5	438
6	429
7	117
8	11

Table 3.2.1: Number of Observations of different quality's of wine.

Volatile acidity and quality:

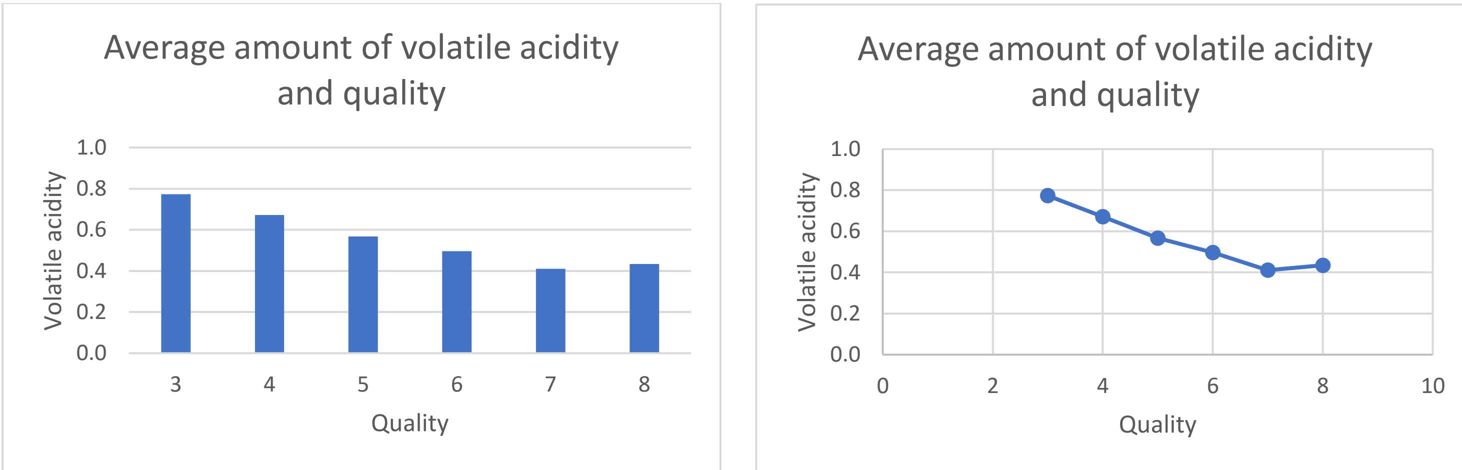


Figure 3.2.1: Average amount of volatile acidity in different quality's of wine

- The lesser the amount of volatile acidity greater the quality.

Citric acid and quality:

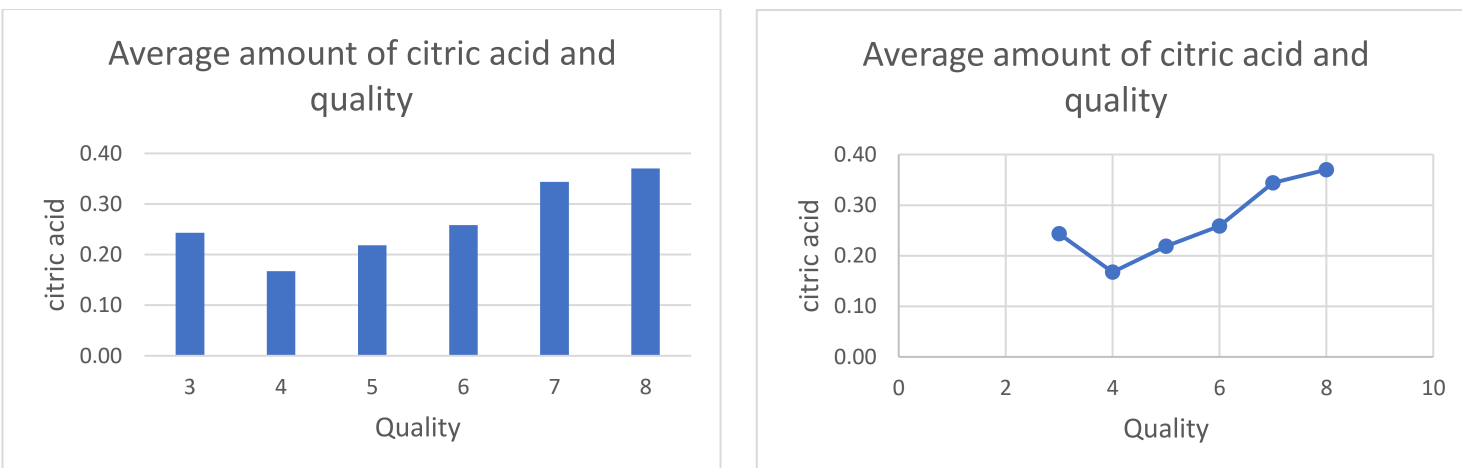


Figure 3.2.2: Average amount of citric acid in different qualities of wine

- After removing the outliers, there are only 03 observations of wine quality of level 3. This is not enough observation to come to any conclusion for level 3. But from the rest of the data, we can see that the greater the amount of citric acid, the greater the quality.
- Citric acid is present in small amounts in the wine. The highest amount of citric acid is 0.73, which was found in a level 6-rated wine.

Chlorides and quality:

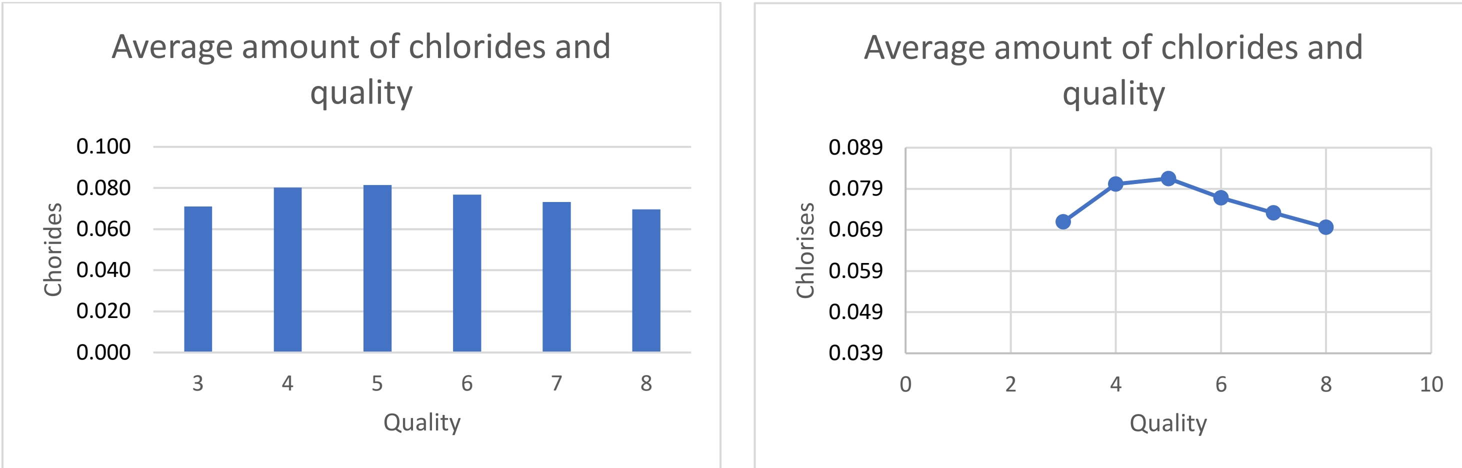


Figure 3.2.3: Average amount of chlorides in different wine qualities.

Note: The minimum value chlorides/salt is 0.039 (g/dm^3) and maximum value is 0.122 (g/dm^3) (Table 2.2.2). So, I have also started the y-axis from 0.039 in figure 3.2.3.

- More salt is lesser in the test (quality).
- From the observation, we can say that we should look for a wine where the chloride value is lesser than 0.07 (if we ignore the observation of level 3 wine since there are only 03 observations of wine quality of level 3)

Sulphates and quality:

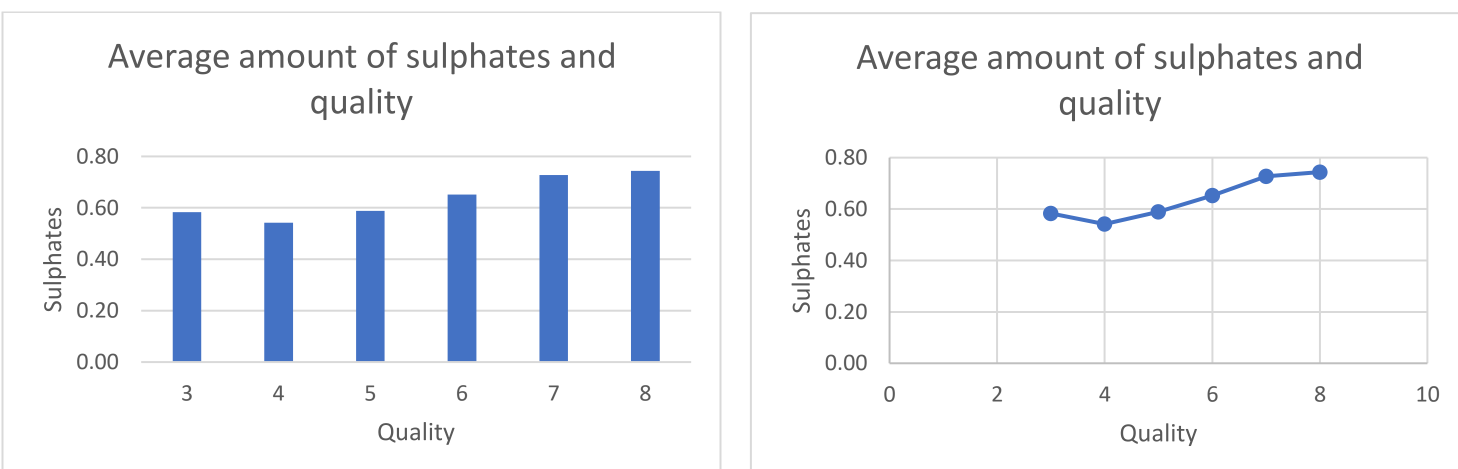
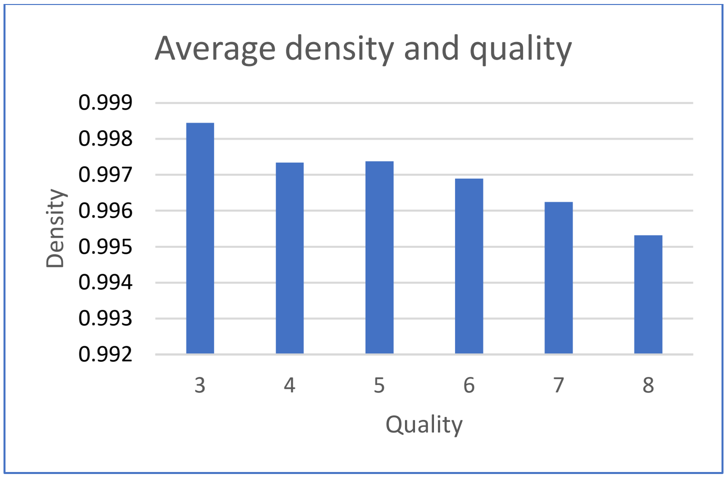


Figure 3.2.4: Average amount of sulphates in different qualities

- Increase in sulphates additive increases the quality.

Density and quality:



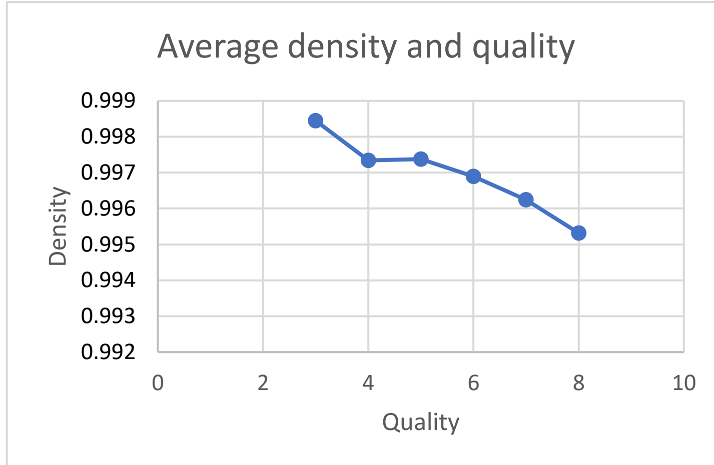
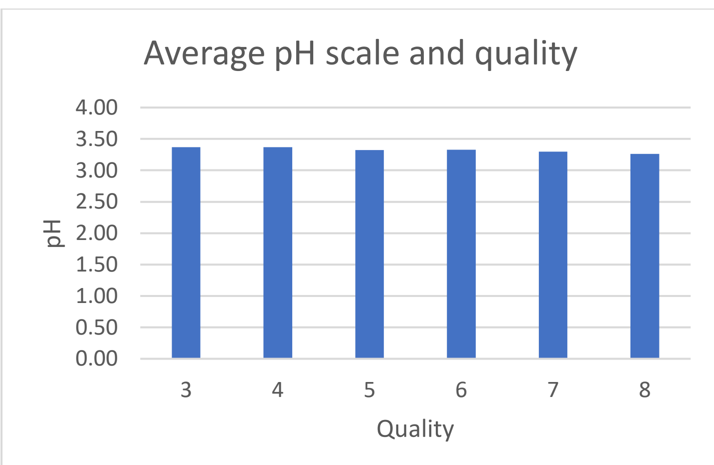


Figure 3.2.5: Average density in different qualities

Note: No liquid is 0 in (g/cm^3) density. In our data set the lowest value of density is 0.9915. So, we also start the y-axis from 0.9915 (Table 2.2.2).

- Increase in density decreases the quality.
- The density of water is higher than the density of alcohol. More dense wine means more watery wine.

pH scale and quality:



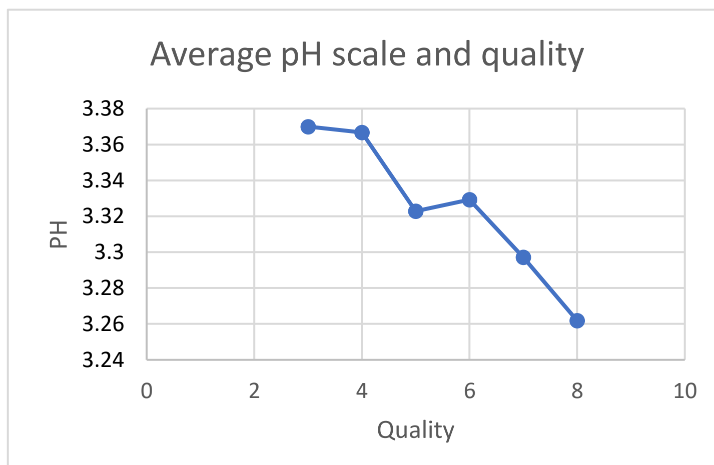


Figure 3.2.6: Average pH scale value in different qualities

Note: In red wine, pH scale value varies in decimal point (Table 2.2.2). So, I have also started the y-axis from 3.24 in figure 3.2.6.

- In the best quality wines rated as 7 or 8, the average pH value is between 3.2 and 3.3.
- In wine quality, the pH value varies in decimal points/fractions. The decimal fraction variation should not be that noticeable in the tongue.

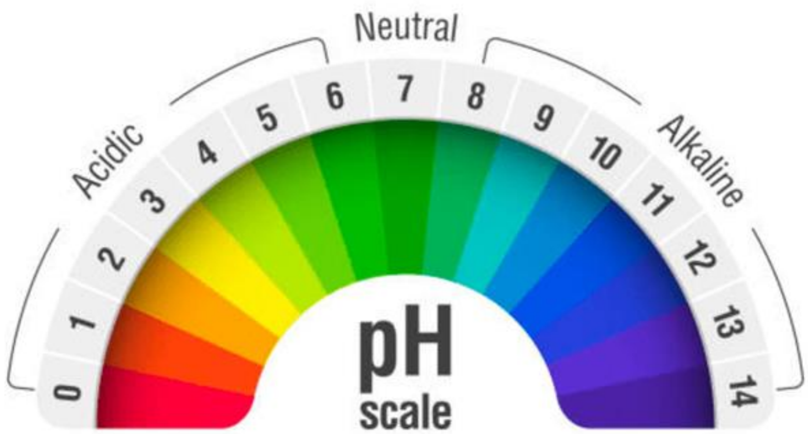


Figure 3.2.7: pH scale

- Red wine is acidic in mild test.
- Decrease in pH value makes little increase in the wine quality.

The pH is a scale measuring how acidic or basic the solution is, which varies from 0 to 14, and where seven is considered neutral. A pH scale value less than seven is considered acidic, and a value greater than seven is considered base.

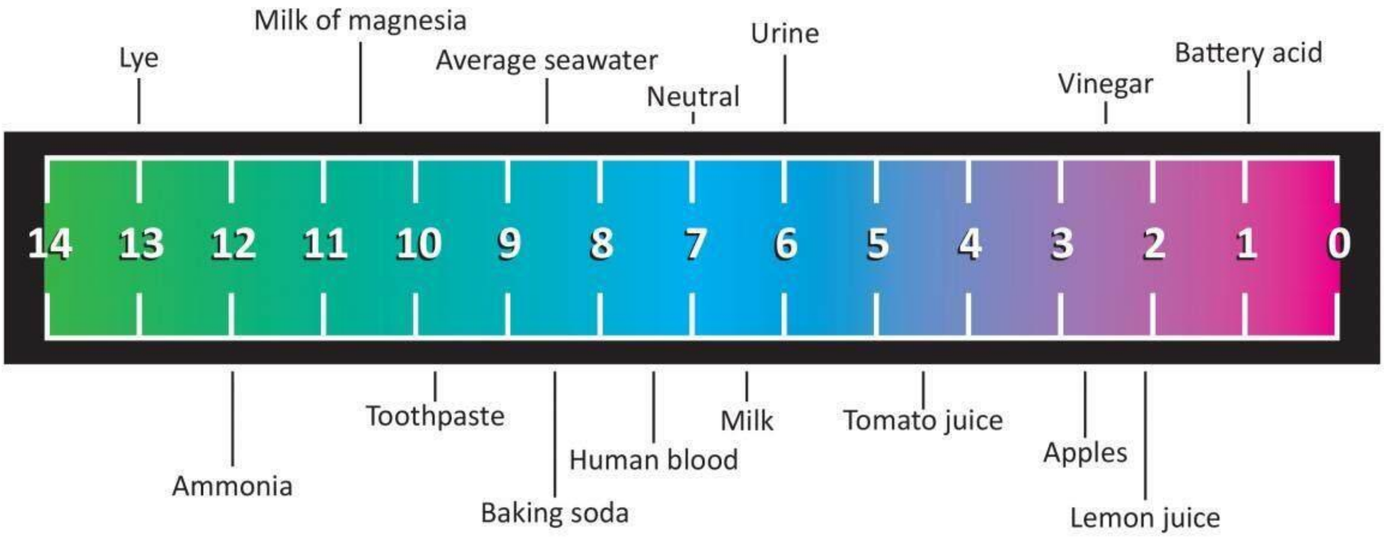


Figure 3.2.8: pH scale value in different item.

Sulphates and wine quality:

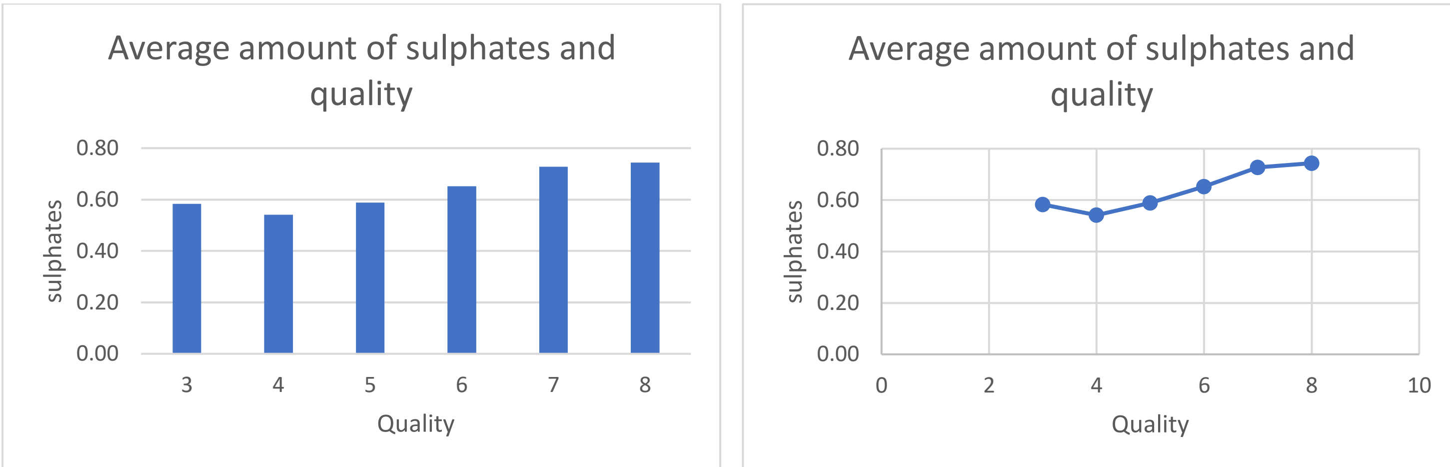


Figure 3.2.9: Average sulphates in different wine quality

- From the cart we can see that increase in sulphates is increase in quality.

Alcohol and wine quality:

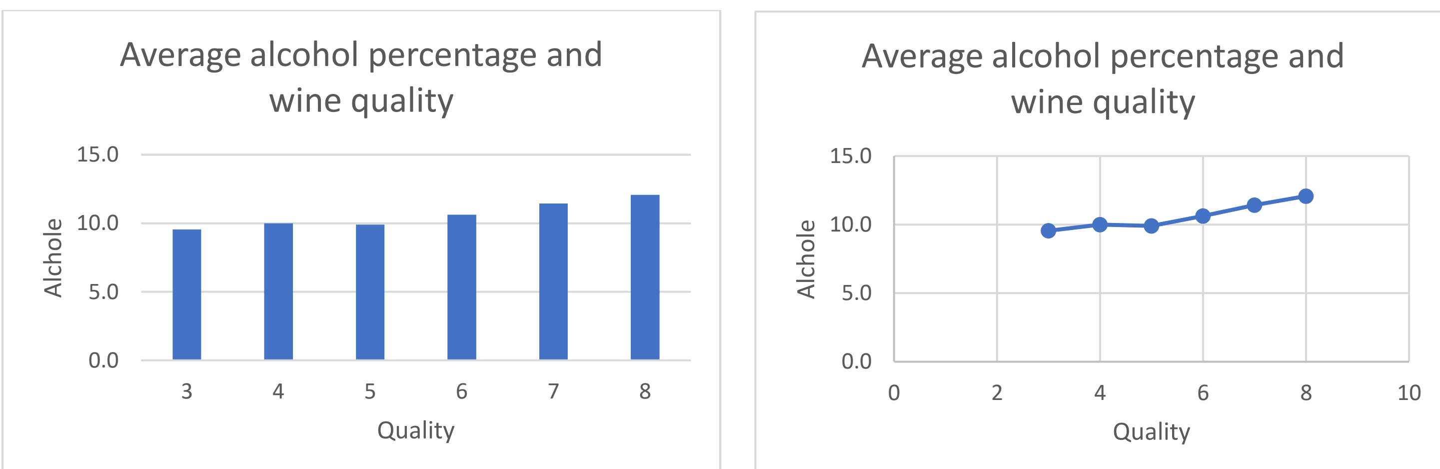


Figure 3.2.10: Average alcohol percentage in different wine quality

- From the cart, we can see an increase in alcohol is an increase in quality.
- In best-quality wines rated as 7 or 8, the average alcohol percentage is between 11 to 13. We should order a wine where the alcohol percentage is within this range. The best choice could be around 12% alcohol.

3.3 Relationship among the variables

Fixed acidity:

Correlation of different features with fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
fixed acidity	-0.28	0.66	0.24	0.21	-0.13	-0.08	0.61	-0.70	0.18	-0.05	0.11

Table 3.3.1: Correlation of different features with fixed acidity.

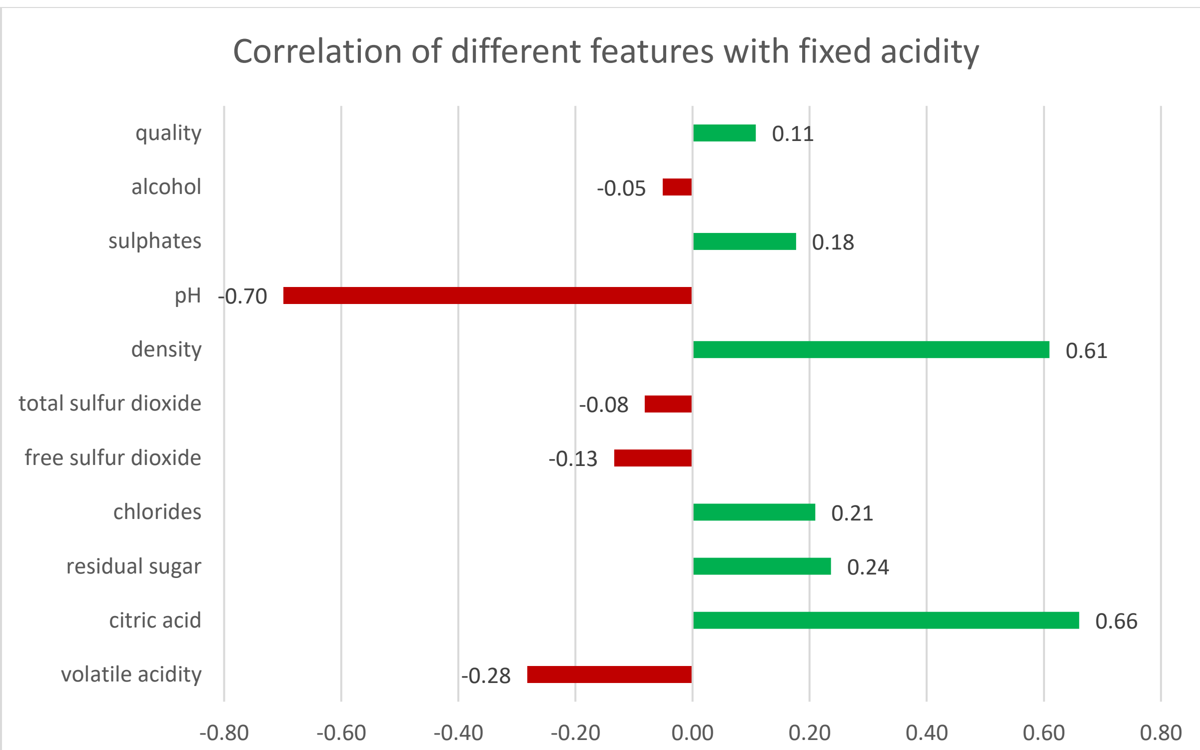


Figure 3.3.1: Correlation of different features with fixed acidity.

- Fixed acidity has a very significant positive correlation with citric acid (with a value of 0.66) and density (with a value of 0.61).
- There is a very significant negative correlation with pH level, with a value of -0.70. It makes sense because the pH value of any acidic solution is less than 7. The more it is acidic, the lesser the pH level.

- Fixed acidity negatively correlates with volatile acidity, with a value of -0.28.
- Fixed acidity hardly correlates negatively with residual sugar, with a value of -0.24.

Volatile acidity:

Correlation of different features with volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
volatile acidity	-0.62	0.01	0.12	-0.02	0.10	0.04	0.24	-0.31	-0.22	-0.35

Table 3.3.2: Correlation of different features with volatile acidity

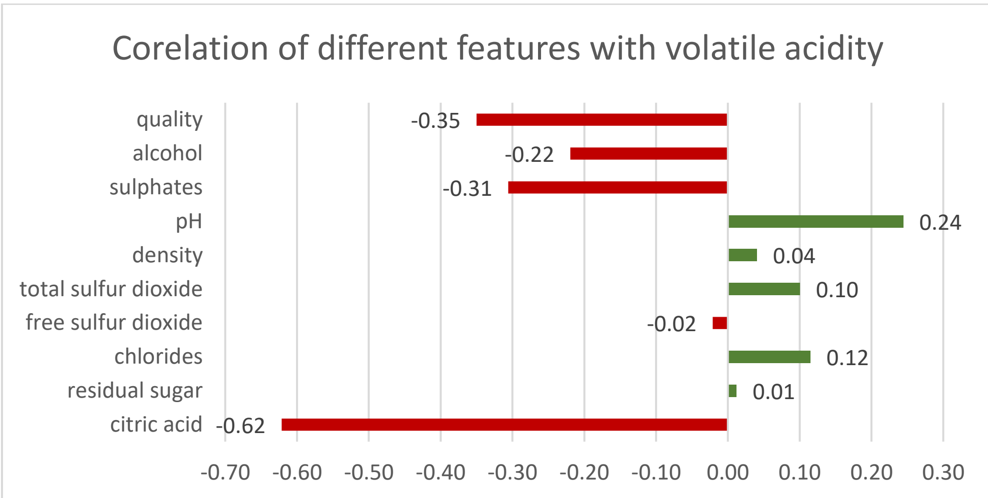


Figure 3.3.2: Correlation of different features with volatile acidity

- There is a very strong negative correlation between volatile acidity and citric acid, with a value of -0.62. Citric acid is able to add freshness and flavor to red wine, where a very high amount of volatile acidity can start a vinegar taste, which is unpleasant. It has a *strong* negative correlation with quality -0.35, which seems very logical.
- Volatile acidity has a strong negative correlation with sulphate -0.31.
- It has a mild positive correlation with pH 0.24. it’s unusual.

Citric acid:

Correlation of different features with citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
citric acid	0.16	0.07	-0.06	0.01	0.29	-0.49	0.27	0.14	0.22

Table 3.3.3: Correlation of different features with citric acid

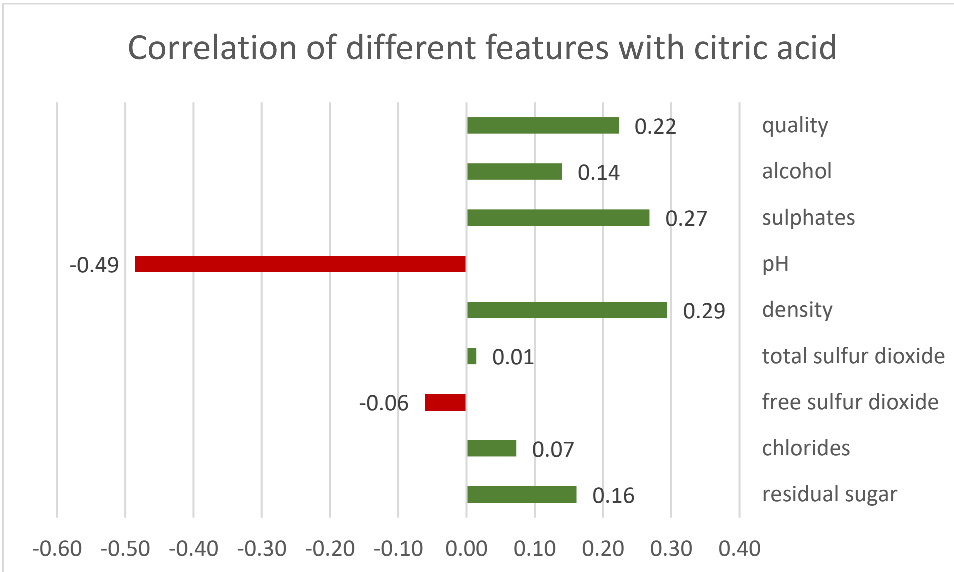


Figure 3.3.3: Correlation of different features with citric acid

- Citric acid has a very strong negative correlation with pH -0.49. The more it increases the more the wine will become acidic. Which makes sense.
- Citric acid has a positive correlation with density 0.29 and with sulphates 0.27.

Residual sugar:

Correlation of different features with residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
residual sugar	0.27	0.08	0.18	0.40	-0.07	0.07	0.08	0.02

Table 3.3.4: Correlation of different features with residual sugar.

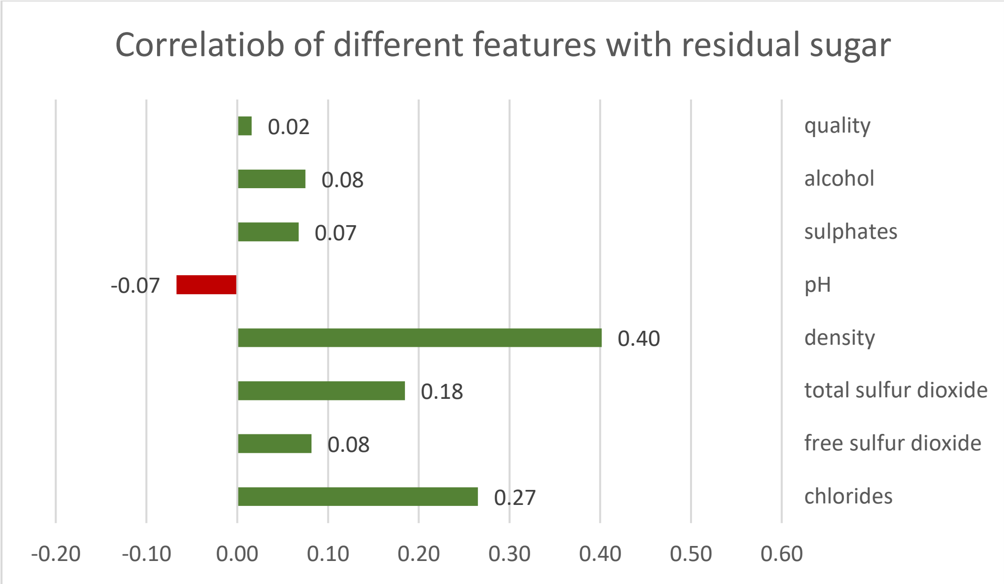


Figure 3.3.4: Correlation of different features with residual sugar.

- Residual sugar has a very strong positive correlation with density 0.40. That makes sense because residual sugar is the amount of remaining sugar after the fermentation has stopped. Alcohol is less dense than water. More residual sugar is more watery wine.
- Residual sugar has a strong positive correlation with chlorides/salt.

Chlorides:

Correlation of different features with chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
chlorides	0.03	0.18	0.43	-0.19	-0.07	-0.30	-0.18

Table 3.3.5: Correlation of different features with chlorides

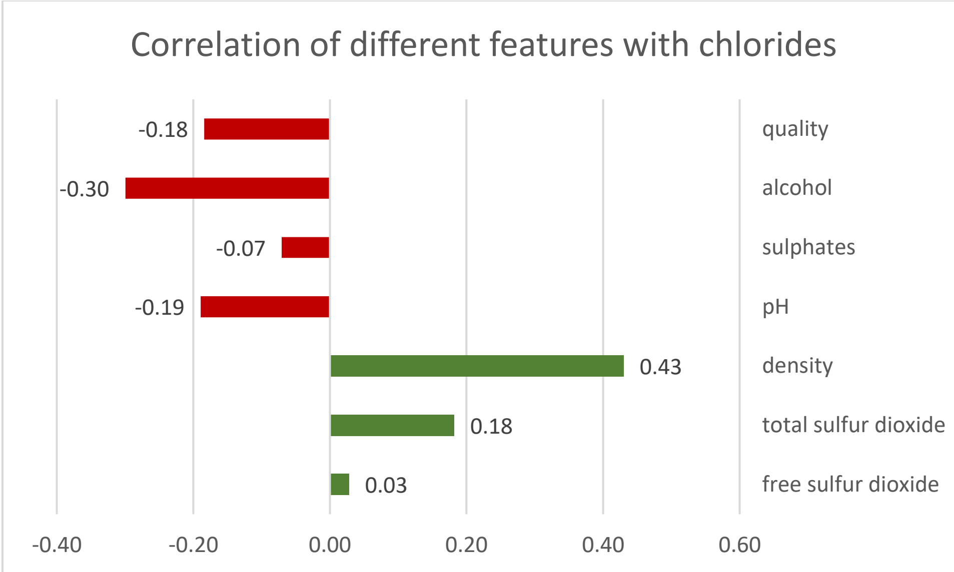


Figure 3.3.5: Correlation of different features with chlorides

- Chlorides has a strong positive correlation with density 0.43. So, more salt means higher density.
- It has a strong negative correlation with alcohol with value -0.30.
- More salty wine means higher in density and lower in alcohol percentage.

Free sulfur dioxide:

Correlation of different features with free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
free sulfur dioxide	0.62	-0.01	0.10	0.08	-0.05	-0.01

Table 3.3.6: Correlation of different features with free sulfur dioxide.

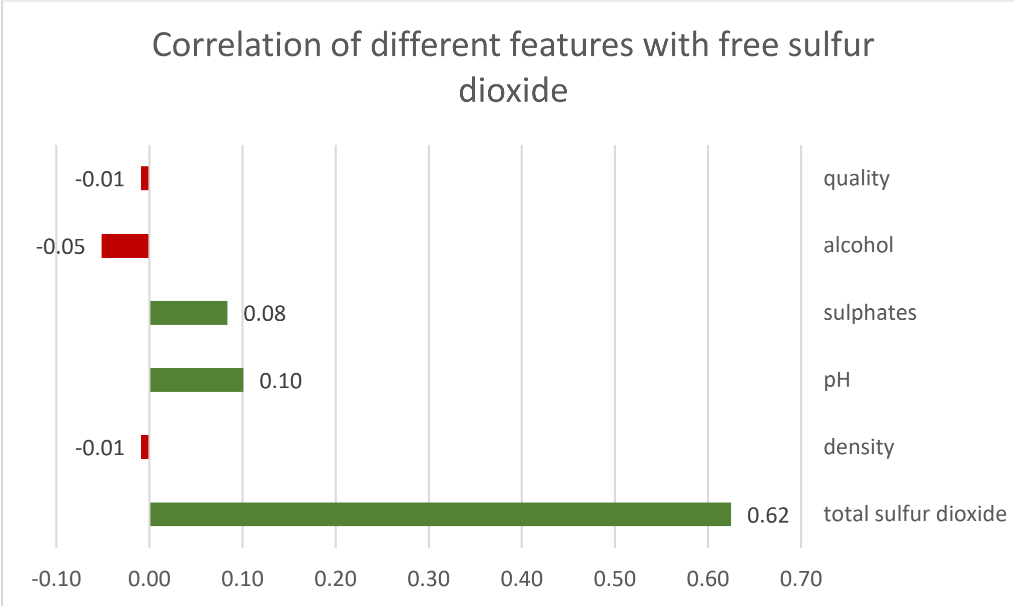


Figure 3.3.6: Correlation of different features with free sulfur dioxide.

- Free sulfur dioxide has a very strong positive correlation with total sulfur dioxide of 0.62. Total Sulfur Dioxide = amount of Free sulfur dioxide + bound forms of sulfur dioxide, which was expected.

Total sulfur dioxide:

Correlation of different features with total sulfur dioxide	density	pH	sulphates	Alcohol	quality
total sulfur dioxide	0.16	-0.01	-0.06	-0.27	-0.20

Table 3.3.7: Correlation of different features with total sulfur dioxide.

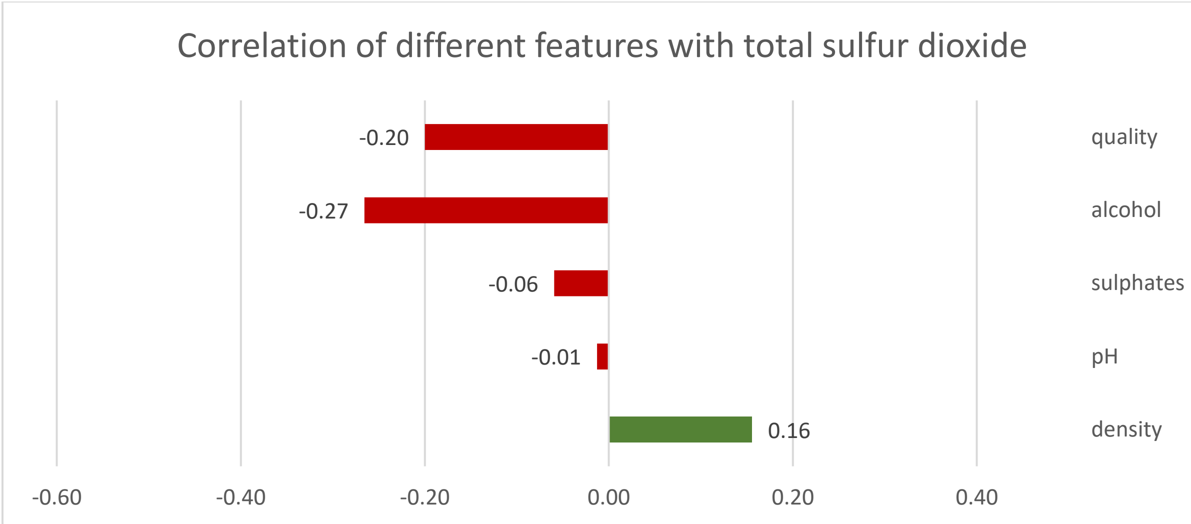


Figure 3.3.7: Correlation of different features with total sulfur dioxide.

- Total sulfur dioxide has a strong negative correlation with an alcohol percentage of -0.27.
- Total sulfur dioxide has a negative correlation with quality-0.20. “at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine” [from the description]

Density:

Correlation of different features with density	pH	sulphates	alcohol	quality
density	-0.24	0.09	-0.56	-0.24

Table 3.3.8: Correlation of different features with density.

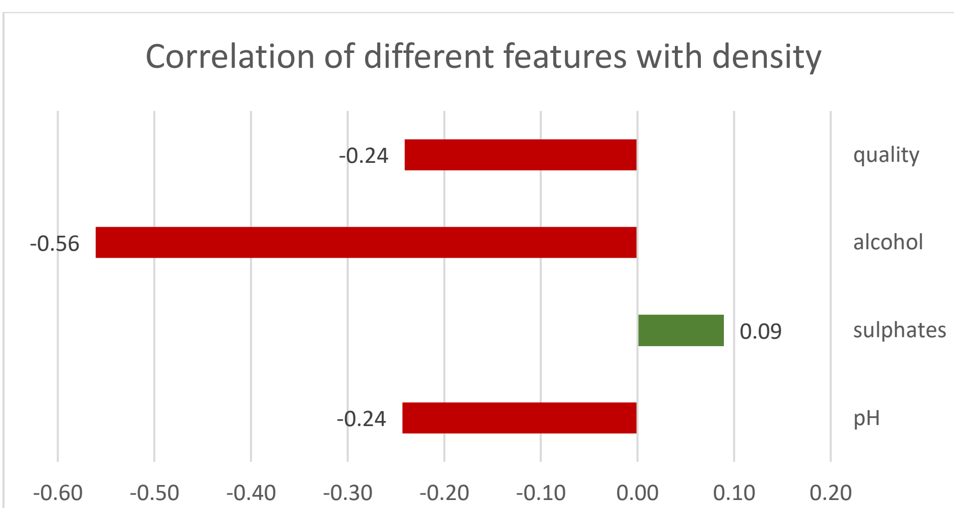


Figure 3.3.8: Correlation of different features with density.

- Density has a very strong negative correlation with alcohol percentage of -0.56. High density wine is more watery wine.
- Density has a mild negative correlation with pH and quality (with a value of -0.24).

pH:

Correlation of different features with pH	sulphates	alcohol	quality
pH	0.01	0.13	-0.07

Table 3.3.9: Correlation of different features with pH.

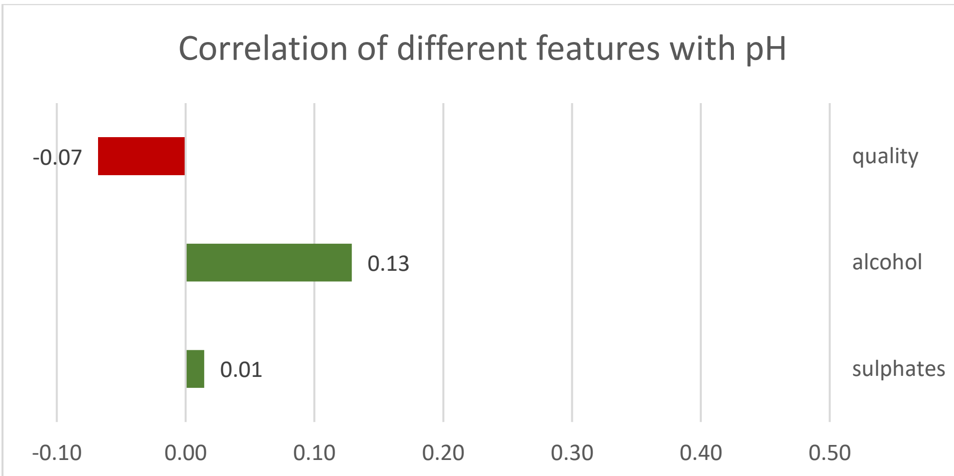


Figure 3.3.9: Correlation of different features with pH.

Before we saw that pH has a very strong negative correlations with fixed acidity -0.70 and with citric acid -0.49. But pH has a mild positive correlation with volatile acidity with value 0.24 (which seems unusual).

Sulphates:

Correlation sulphates with different variables	alcohol	quality	volatile acidity
sulphates	0.27	0.42	-0.31

Table 3.3.10: Correlation of different features with sulphates.

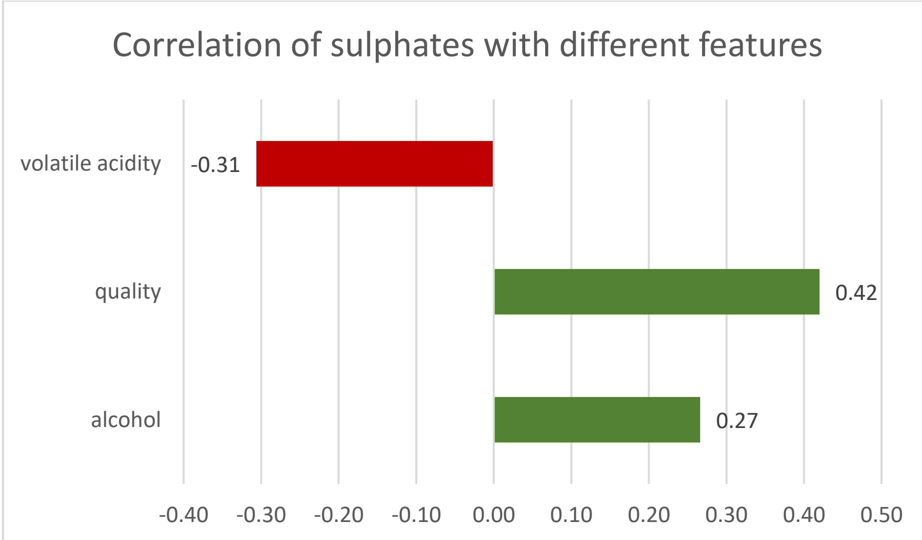


Figure 3.3.10: Correlation of different features with sulphates.

- Sulphate has a very strong positive c*orrelation* with quality 0.42. “Sulphites in wine are used to stop fermentation at a specific point in the winemaking process.” [1]
- Sulphate has a mild positive c*orrelation* with an alcohol percentage of 0.27.
- Sulphate has a high negative correlation with volatile acidity -0.31 (Table 3.3.1) (Figure 3.3.2).

Quality:

Correlation of different features with quality	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
Quality	0.11	-0.35	0.22	0.02	-0.18	-0.01	-0.20	-0.24	-0.07	0.42	0.51	1

Table 3.3.11: Correlation of different features with quality.

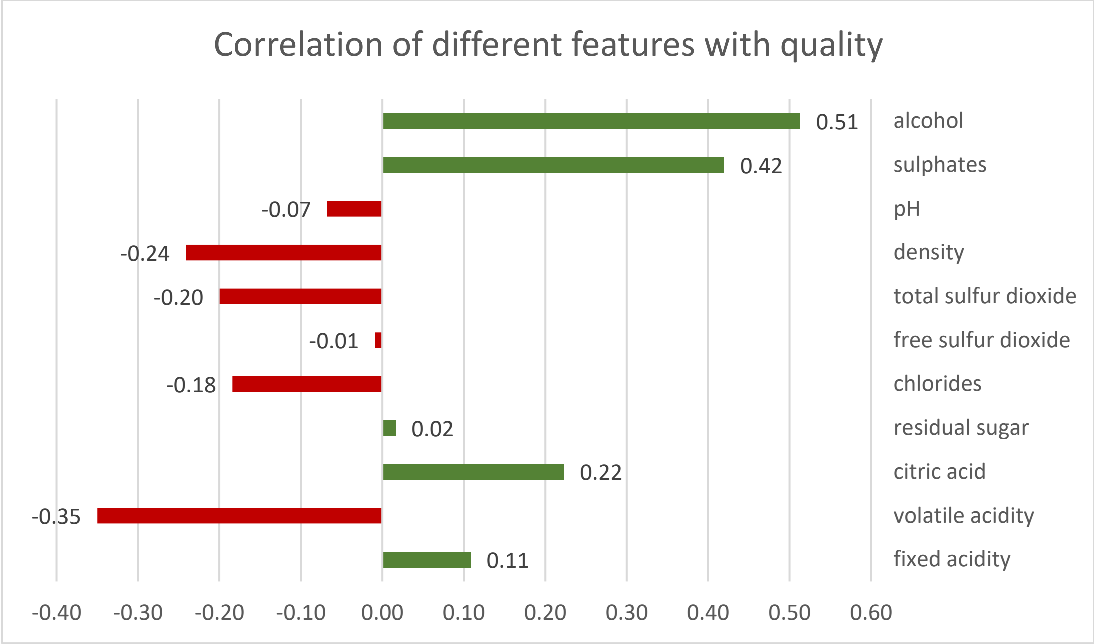


Figure 3.3.11: Correlation of different features with quality.

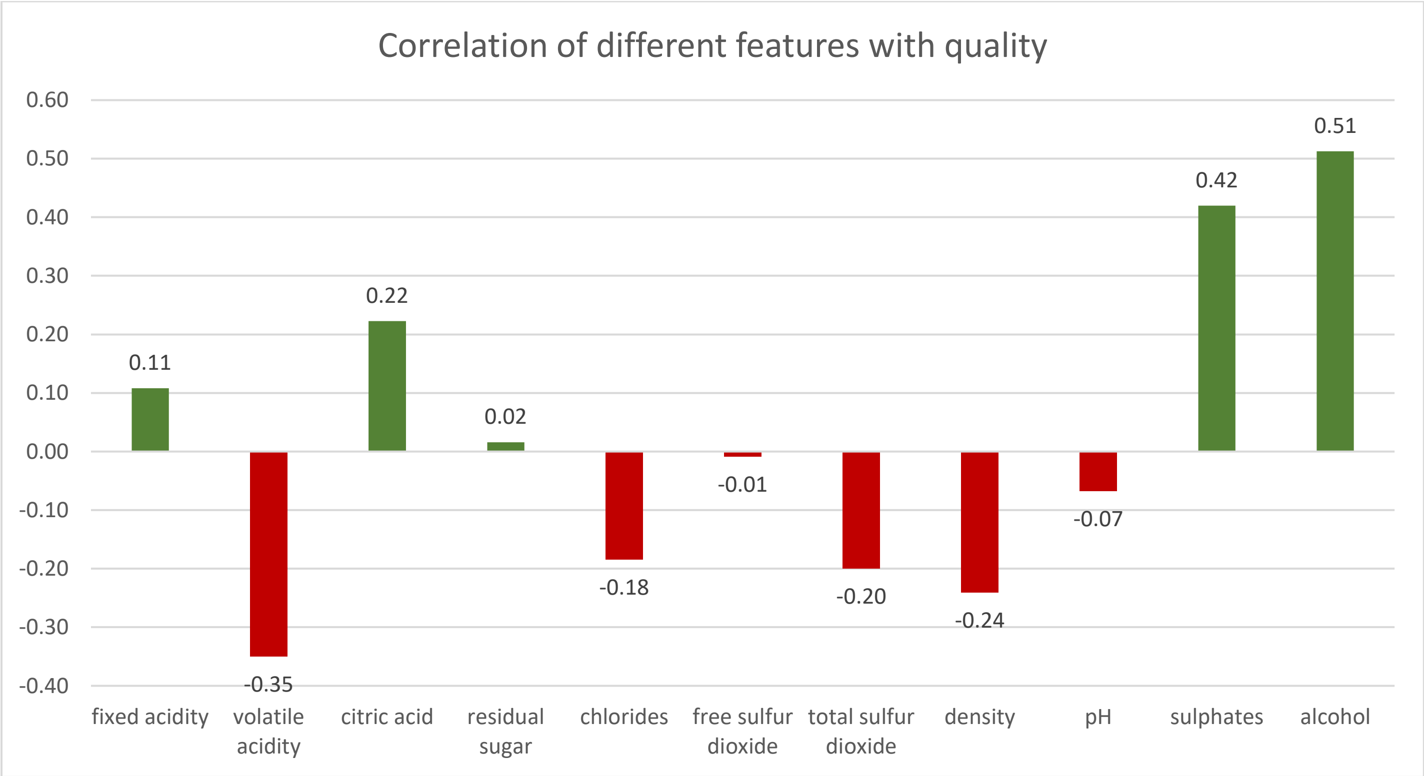


Figure 3.3.12: Correlation of different features with quality.

We can see that wine quality has a very significant positive correlation with alcohol. In fact, alcohol is the most correlated feature with quality in our data set, with a value of 0.51.

The alcohol feature is the amount of alcohol percent content in the wine. A higher percentage of alcohol content would yield better satisfaction for a customer purchasing red wine; it seems like the most important ingredient in wine.

Next, we can see the second strongest positive correlation, 0.41, between sulphates & our quality predictor. It seems that people rate the quality higher when an additive is contributed to the drink. Sulphate acts as an antimicrobial. "Sulphites in wine are used to stop fermentation at a specific point in the winemaking process. Besides, they function as preservatives to prevent spoiling and oxidation and as protection from bacteria. All in all, sulphites help to maintain the freshness and flavour of wine and prolong its shelf life." [1]

Wine quality has the strongest negative correlation with volatile acidity, with a correlation of -0.35. This is expected because too high acetic acid levels can lead to an unpleasant vinegar taste.

Wine quality has a mild positive correlation with citric acid 0.2, even though it adds freshness and flavor to wines. It seems to be present in a very small quantity in the wine, and around 9% of red wine does not have any citric acid (Mode =0) (Table 3.1.3).

Wine quality has a mild negative correlation with density with a value of -0.24.

3.4 Correlation matrix

Correlation matrix	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur	density	pH	sulphates	alcohol	quality
fixed acidity	1											
volatile acidity	-0.28	1										
citric acid	0.66	-0.62	1									
residual sugar	0.24	0.01	0.16	1								
Chlorides	0.21	0.12	0.07	0.27	1							
free sulfur dioxide	-0.13	-0.02	-0.06	0.08	0.03	1						
total sulfur dioxide	-0.08	0.10	0.01	0.18	0.18	0.62	1					
Density	0.61	0.04	0.29	0.40	0.43	-0.01	0.16	1				
pH	-0.70	0.24	-0.49	-0.07	-0.19	0.10	-0.01	-0.24	1			
Sulphates	0.18	-0.31	0.27	0.07	-0.07	0.08	-0.06	0.09	0.01	1		
Alcohol	-0.05	-0.22	0.14	0.08	-0.30	-0.05	-0.27	-0.56	0.13	0.27	1	
Quality	0.11	-0.35	0.22	0.02	-0.18	-0.01	-0.20	-0.24	-0.07	0.42	0.51	1

Table 3.4.1: Correlation matrix of the variables.

3.5 Regression

The target /dependent variable (Y) is wine quality, which is determined by the rest of the 11 features/independent variables (X).

Simple Linear Regression

Fixed acidity and wine quality:

Simple regression between the independent variable/feature fixed acidity and the dependent variable quality:

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.108253376							
R Square	0.011718794							
Adjusted R Square	0.010758365							
Standard Error	0.773112814							
Observations	1031							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	7.292954157	7.292954157	12.20162688	0.00049778			
Residual	1029	615.0368228	0.597703423					
Total	1030	622.3297769						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	5.17618006	0.134171906	38.57871757	4.1169E-202	4.912898278	5.439461843	4.912898278	5.439461843

fixed acidity	0.056510225	0.016177752	3.493082718	0.00049778	0.024765074	0.088255375	0.024765074	0.088255375

Table 3.5.1: a simple regression model with fixed acidity and wine quality.

P-value of the fixed acidity variable is 0.00049778, which is smaller than the threshold value (0.05). So, it is statistically significant.

The confidence interval's lower bound (Lower 95%) for the fixed acidity variable is 0.024765074, and the upper bound (Upper 95%) is 0.088255375. Zero does not fall within this confidence interval. So, the fixed acidity variable is statistically significant.

By the R-squared value, this simple regression model with a fixed acidity variable explains about 1% of the data..

Volatile acidity and Quality:

Simple regression between the independent variable/feature volatile acidity and the dependent variable quality:

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.349875121							
R Square	0.1224126							
Adjusted R Square	0.121559745							
Standard Error	0.728530592							
Observations	1031							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	76.1810061	76.1810061	143.5327872	4.6951E-31			
Residual	1029	546.1487708	0.530756823					
Total	1030	622.3297769						

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	6.489340773	0.074654802	86.92462672	0	6.342847741	6.635833805	6.342847741	6.635833805
volatile acidity	-1.63005564	0.136058873	-11.980517	4.6951E-31	-1.89704017	-1.36307111	-1.89704017	-1.36307111

Table 3.5.2: a simple regression model with volatile acidity and wine quality.

The p-value of the volatile acidity variable is 4.6951E-31, which is smaller than the threshold value (0.05). So, it is statistically significant.

The confidence interval's lower bound (Lower 95%) for the volatile acidity variable is -1.897040167 and the upper bound (Upper 95%) is -1.363071114. Zero does not fall within this confidence interval. So, the volatile acidity variable is statistically significant.

By the R-squared value, this simple regression model with volatile acidity variable explains about 12% of the data.

Citric acid and Quality:

Simple regression between the independent variable/feature citric acid and the dependent variable quality:

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.222887042							
R Square	0.049678634							
Adjusted R Square	0.048755095							
Standard Error	0.758119819							
Observations	1031							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	30.91649292	30.91649292	53.79160746	4.5102E-13			

Residual	1029	591.413284	0.57474566					
Total	1030	622.3297769						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	5.400639908	0.039977333	135.0925525	0	5.322193505	5.479086311	5.322193505	5.479086311
citric acid	0.94818772	0.129281703	7.334276205	4.5102E-13	0.694501845	1.201873595	0.694501845	1.201873595

Table 3.5.3: a simple regression model with citric acid and wine quality.

The p-value of the citric acid variable is 4.5102E-13, which is smaller than the threshold value (0.05). So, it is statistically significant.

The confidence interval's lower bound (Lower 95%) for the citric acid variable is 0.694501845, and the upper bound (Upper 95%) is 1.201873595. Zero does not fall within this confidence interval.

So, the citric acid variable is statistically significant.

By the R-squared value, this simple regression model with citric acid variable explains about 5% of the data.

Residual sugar and Wine quality:

Regression between the independent variable residual sugar and the dependent variable quality:

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.015983874							
R Square	0.000255484							
Adjusted R Square	-0.00071608							
Standard Error	0.777583647							
Observations	1031							
ANOVA								

	df	SS	MS	F	Significance F			
Regression	1	0.158995437	0.158995437	0.262960444	0.608203607			
Residual	1029	622.1707815	0.604636328					
Total	1030	622.3297769						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	5.577174964	0.11961976	46.62419475	5.9036E-256	5.342448451	5.811901476	5.342448451	5.811901476
residual sugar	0.027369296	0.053372608	0.512796689	0.608203607	-0.07736228	0.132100874	-0.07736228	0.132100874

Table 3.5.4: a simple regression model with residual sugar and wine quality.

The p-value of the residual sugar variable is 0.608203607, which is greater than the threshold value. So, it is not statistically significant.

The confidence interval's lower bound (Lower 95%) for the residual sugar is -0.077362281 and the upper bound (Upper 95%) is 0.132100874. Zero falls within this confidence interval. So, the residual sugar variable is not statistically significant.

Chlorides and Wine quality:

Regression between the independent variable chlorides and the dependent variable quality:

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.184235985							
R Square	0.033942898							
Adjusted R Square	0.033004067							

Standard Error	0.764370649							
Observations	1031							
ANOVA								
	Df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	21.12367628	21.12367628	36.15442836	2.52891E-09			
Residual	1029	601.2061006	0.584262488					
Total	1030	622.3297769						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	6.385761197	0.126741622	50.38408946	4.5947E-280	6.137059653	6.634462741	6.137059653	6.634462741
chlorides	-9.54803333	1.587936666	-6.01285526	2.52891E-09	-12.6639971	-6.43206957	-12.6639971	-6.43206957

Table 3.5.5: a simple regression model with chlorides and wine quality.

The p-value of the chloride variable is 2.52891E-09, which is smaller than the threshold value (0.05). So, it is statistically significant.

The confidence interval's lower bound (Lower 95%) for the chloride's variable is -12.6639971 and the upper bound (Upper 95%) is -6.43206957. Zero does not fall within this confidence interval. So, the chlorides variable is statistically significant.

By the R-squared value, this simple regression model with chlorides variable explains about 3% of the data.

Free sulfur dioxide and Wine quality:

Regression between the independent variable free sulfur dioxide and the dependent variable quality:

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.009032395							

R Square	8.15842 E-05							
Adjusted R Square	-0.00089015							
Standard Error	0.777651272							
Observations	1031							
ANOVA								
	Df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	0.05077225	0.05077225	0.083956946	0.772063389			
Residual	1029	622.2790047	0.604741501					
Total	1030	622.3297769						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	5.649091789	0.047519414	118.879661	0	5.555845771	5.742337807	5.555845771	5.742337807
free sulfur dioxide	-0.00079428	0.002741227	-0.28975325	0.772063389	-0.00617331	0.004584754	-0.00617331	0.004584754

Table 3.5.6: a simple regression model with free sulfur dioxide and wine quality.

The p-value of the free sulfur dioxide variable is 0.772063389, which is greater than the threshold value. So, it is not statistically significant.

The confidence interval’s lower bound (Lower 95%) for the free sulfur dioxide is -0.00617331 and the upper bound (Upper 95%) is 0.004584754. Zero falls within this confidence interval. So, the free sulfur dioxide variable is not statistically significant.

Total sulfur dioxide and Wine quality:

Regression between the independent variable total sulfur dioxide and the dependent variable quality:

SUMMARY OUTPUT								
Regression Statistics								

Multiple R	0.199957602							
R Square	0.039983043							
Adjusted R Square	0.039050082							
Standard Error	0.761977339							
Observations	1031							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	24.88263801	24.88263801	42.85606683	9.28112E-11			
Residual	1029	597.4471389	0.580609464					
Total	1030	622.3297769						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	5.883989789	0.044539713	132.1065944	0	5.796590755	5.971388824	5.796590755	5.971388824
total sulfur dioxide	-0.00584598	0.000892999	-6.54645452	9.28112E-11	-0.00759828	-0.00409367	-0.00759828	-0.00409367

Table 3.5.7: a simple regression model with total sulfur dioxide and wine quality.

The p-value of the total sulfur dioxide variable is 9.28112E-11, which is smaller than the threshold value (0.05). So, it is statistically significant.

The confidence interval’s lower bound (Lower 95%) for the total sulfur dioxide variable is -0.00759828 and the upper bound (Upper 95%) is -0.00409367. Zero does not fall within this confidence interval.

So, the total sulfur dioxide variable is statistically significant.

By the R-squared value, this simple regression model with total sulfur dioxide variable explains about 4% of the data.

Density and Wine quality:

Regression between the independent variable Density and the dependent variable quality:

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.241047313							
R Square	0.058103807							
Adjusted R Square	0.057188456							
Standard Error	0.754751742							
Observations	1031							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	36.15972929	36.15972929	63.47707733	4.28065E-15			
Residual	1029	586.1700476	0.569650192					
Total	1030	622.3297769						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	119.3615666	14.27399305	8.362170709	1.98273E-16	91.35210877	147.3710245	91.35210877	147.3710245
density	-114.120529	14.32370321	-7.9672503	4.28065E-15	-142.227531	-86.013526	-142.227531	-86.013526

Table 3.5.8: a simple regression model with density and wine quality.

The p-value of the density variable is 4.28065E-15 which is smaller than the threshold value (0.05). So, it is statistically significant.

The confidence interval's lower bound (Lower 95%) for the density variable is -142.227531 and the upper bound (Upper 95%) is -86.013526. Zero does not fall within this confidence interval.

So, the density variable is statistically significant.

By the R-squared value, this simple regression model with density variable explains about 6% of the data.

pH and Wine quality:

Regression between the independent variable pH and the dependent variable quality:

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.067635918							
R Square	0.004574617							
Adjusted R Square	0.003607246							
Standard Error	0.775902156							
Observations	1031							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	2.846920607	2.846920607	4.728914247	0.029886828			
Residual	1029	619.4828563	0.602024156					
Total	1030	622.3297769						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	6.916083943	0.588574349	11.75056976	5.2633E-30	5.761140939	8.071026947	5.761140939	8.071026947
pH	-0.38479213	0.176947921	-2.17460669	0.029886828	-0.73201209	-0.03757217	-0.73201209	-0.03757217

Table 3.5.9: a simple regression model with pH and wine quality.

The p-value of the pH variable is 0.029886828 which is smaller than the threshold value (0.05). So, it is statistically significant.
The confidence interval's lower bound (Lower 95%) for the pH variable is -0.73201209 and the upper bound (Upper 95%) is -0.03757217. Zero does not fall within this confidence interval.
So, the pH variable is statistically significant.
By the R-squared value, this simple regression model with pH variable explains about 0.5% of the data.

Sulphates and Wine quality:

Regression between the independent variable sulphates and the dependent variable quality:

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.419675136							
R Square	0.17612722							
Adjusted R Square	0.175326566							
Standard Error	0.705882931							
Observations	1031							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	109.6092136	109.6092136	219.9792418	3.02562E-45			
Residual	1029	512.7205633	0.498270713					
Total	1030	622.3297769						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	3.846994314	0.122690018	31.35539786	5.0576E-152	3.60624312	4.087745508	3.60624312	4.087745508
sulphates	2.837605482	0.191320349	14.8316972	3.02562E-45	2.462182906	3.213028059	2.462182906	3.213028059

Table 3.5.10: a simple regression model with sulphates and wine quality.

The p-value of the sulphate variable is 3.02562E-45, which is smaller than the threshold value (0.05). So, it is statistically significant.

The confidence interval's lower bound (Lower 95%) for the sulphate variable is 2.462182906 and the upper bound (Upper 95%) is 3.213028059. Zero does not fall within this confidence interval.

So, the sulphates variable is statistically significant.

By the R-squared value, this simple regression model with sulphates variable explains about 18% of the data.

Alcohol and Wine quality:

Regression between the independent variable alcohol and the dependent variable quality:

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.512761543							
R Square	0.2629244							
Adjusted R Square	0.262208097							
Standard Error	0.667665016							
Observations	1031							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	163.6256832	163.6256832	367.0576092	3.30735E-70			
Residual	1029	458.7040938	0.445776573					
Total	1030	622.3297769						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*

Intercept	1.508080283	0.216524501	6.964940582	5.8515E-12	1.083200303	1.932960263	1.083200303	1.932960263
alcohol	0.397112882	0.020727497	19.15874759	3.30735E-70	0.356439894	0.43778587	0.356439894	0.43778587

Table 3.5.11: a simple regression model with alcohol and wine quality.

The p-value of the alcohol variable is 3.30735E-70 which is smaller than the threshold value (0.05). So, it is statistically significant.

The confidence interval's lower bound (Lower 95%) for the alcohol variable is 0.356439894 and the upper bound (Upper 95%) is 0.43778587. Zero does not fall within this confidence interval.

So, the alcohol variable is statistically significant.

By the R-squared value, this simple regression model with alcohol variable explains about 26% of the data.

Multiple Linear Regression:

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.629295839							
R Square	0.396013253							
Adjusted R Square	0.389493278							
Standard Error	0.607347018							
Observations	1031							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	11	246.4508394	22.40462176	60.73846469	1.2671E-103			
Residual	1019	375.8789375	0.3688704					
Total	1030	622.3297769						

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	42.69182993	29.65096143	1.43981267	0.150227539	-15.4920958	100.8757557	-15.4920958	100.8757557
fixed acidity	0.026885679	0.033811958	0.795152965	0.426709695	-0.03946335	0.093234707	-0.03946335	0.093234707
volatile acidity	-0.808412	0.159201563	-5.07791495	4.53547E-07	-1.12081238	-0.49601161	-1.12081238	-0.49601161
citric acid	-0.26664315	0.181262516	-1.47103301	0.141590915	-0.62233363	0.089047338	-0.62233363	0.089047338
residual sugar	0.023640343	0.054130586	0.436728012	0.662401135	-0.08257982	0.129860508	-0.08257982	0.129860508
chlorides	-0.93491033	1.456921257	-0.64170272	0.521210539	-3.79381926	1.923998591	-3.79381926	1.923998591
free sulfur dioxide	0.004017187	0.002881405	1.394176266	0.163568292	-0.00163698	0.009671353	-0.00163698	0.009671353
total sulfur dioxide	-0.00244519	0.001029016	-2.37623962	0.017674104	-0.00446442	-0.00042596	-0.00446442	-0.00042596
density	-39.0383618	30.25139242	-1.29046496	0.197182003	-98.4005102	20.3237865	-98.4005102	20.3237865
pH	-0.55369758	0.248087134	-2.23186739	0.025840501	-1.04051766	-0.06687751	-1.04051766	-0.06687751
sulphates	1.882830432	0.189410407	9.940480392	2.75978E-22	1.511151387	2.254509477	1.511151387	2.254509477
alcohol	0.272650702	0.036160574	7.539999277	1.03717E-13	0.201692998	0.343608406	0.201692998	0.343608406

Table 3.5.12: Multiple linear Regression to wine quality based on the different explanatory variables of the wine.

Interpret R-squared/adjusted R-squared:

Let's see how well our model fits the data. For this reason, we need to interpret R-squared and adjusted R-squared. Generally, in Multiple Linear Regression, when we have more than one explanatory/Independent variable, it is preferable to use the adjusted R-square than the R-squared. Adjusted R-square considers the number of independent variables in the regression model. It can deliver a more accurate view of the correlation.
By the adjusted R-square, that means controlling the number of independent variables in the regression model, the variability of the features of the Red wine explains almost 0.389 (39%) of the variability of the wine quality.

By the R-squared value, our regression model explains about 0.396 (40%) of the data. Alcohol, sulphate, and volatile acidity have a high influence on the model.

In the simple regression models, the highest R-squared value was 26% with the alcohol. However, in the Multiple Linear Regression Model, the R-squared value is around almost 39%. This means the multiple regression model adds about 13% points of explanatory power compared to the simple regression model. Multiple Linear regression has more/better explanatory power.

4. Trends, Patterns and Anomalies

This section presents the patterns and trends that I have identified from the EDA presented in Section 3. Below, I summarize the identified patterns and trends.

Patterns in fixed acidity:

The fixed acidity distribution is slightly right-skewed and centered around 8 g/dm3 (Figure 3.1.1). Fixed acidity has a very significant positive correlation with citric acid (with a value of 0.66) and with density (with a value of 0.61) (Figure 3.3.1). It has a very significant negative correlation with pH level, with a value of -0.70 (the highest negative correlation among the features). It makes sense because the pH value of any acidic solution is less than 7. The more it is acidic, the lesser the pH level. Fixed acidity negatively correlates with volatile acidity, with a value of -0.28. Fixed acidity hardly correlates negatively with residual sugar, with a value of -0.24 (Figure 3.3.1). Fixed acidity has very weak positive correlation with the wine quality, with a value of 0.11 (Figure 3.3.1).

Patterns in volatile acidity:

Volatile acidity exists in small amounts in wines with a mean of 0.52 (g/dm^3). The distribution looks bimodal (at around 0.39 and 0.57) (Figure 3.1.2). There is a very strong negative correlation between volatile acidity and citric acid, with value of -0.62 (Figure 3.3.2). Citric acid is able to add freshness and flavor to the red wine. Where a very high amount of volatile acidity can start a vinegar taste that is unpleasant, it has a strong negative correlation with quality -0.35 (Figure 3.3.2), which seems very logical. Volatile acidity has a strong negative correlation with sulphate -0.31. It has a mild positive correlation with pH 0.24, which is unusual. The smaller the amount of volatile acidity, the greater the quality (Figure 3.2.1).

Patterns in citric acid:

The mode value of citric acid distribution is 0 (Table 3.1.3). In red wine, citric acid is present in a very small amount. Around 9% of red wine has no citric acid in them. The highest amount of citric acid is 0.73 (g/dm^3), which was found in a level 6-rated wine. It seems like there is a very little possibility that a higher value can reduce the wine quality (because it is present in wine in a very small quantity). Citric acid has a strong negative correlation with pH score, with a value of -0.49 (Figure 3.3.3). The more it increases, the more the wine will become acidic, which makes sense. Citric acid has a positive correlation with a density of 0.29 and with sulphates of 0.27 (Figure 3.3.3). Citric acid has a mild positive correlation with the wine quality, with a value of 0.22 (Figure 3.3.3). Citric acid adds freshness and flavor to the wine. The greater the amount of citric acid, the greater the wine quality (Figure 3.2.2).

Patterns in residual sugar:

Residual sugar's Mean 2.2, Medium 2.1, Mode 2, and 75% of data is within the range of 2.5 (g/dm3) (Table 3.1.4). Residual sugar has a strong positive correlation with density with a value of 0.40 (Figure 3.3.4). That makes sense because residual sugar is the amount of remaining sugar after the fermentation has stopped. Alcohol is less dense than water. More residual sugar is more watery wine. Residual sugar significantly correlates with chlorides/salt (with a value of 0.27) (Table 3.4.1). Residual sugar has a negligible positive correlation with wine quality, with a value of 0.02 (Table 3.3.11).

Patterns in Chlorides:

Chloride distribution is normally distributed (Figure 3.1.5). Around 75% of wine has a salt of less than 0.09 g/dm^3 (Figure 3.1.5). Chloride has a strong positive correlation with density, with a value of 0.43 (Figure 3.3.5). So, more salt means higher density. It has a strong negative correlation with alcohol -0.30 (Figure 3.3.5). More salty wine means higher in density and lower in alcohol percentage. Chloride has a mild positive correlation with residual sugar, with a value of 0.27 (Table 3.4.1). Chlorides have a very weak negative correlation with wine quality, with a value of -0.18 (Table 3.3.11).

Patterns in free sulfur dioxide:

The distribution of free sulfur dioxide is moderately right-skewed (Figure 3.1.6). 75% of the data is within the range of 20 mg/dm3 (Figure 3.1.6). Free sulfur dioxide has a very strong positive correlation with total sulfur dioxide 0.62 (Figure 3.3.6). Which was expected. Since, Total Sulfur Dioxide = amount of Free sulfur dioxide + bound forms of sulfur dioxide. Free sulfur dioxide has a very weak negative correlation with wine quality, with a value of -0.01, which is negligible.

Patterns in total sulfur dioxide:

The total sulfur dioxide distribution is right-skewed (Figure 3.1.7). 75% of the data is within the range of 56 mg/dm3 (Figure 3.1.7). Total sulfur dioxide has a very strong positive correlation with total sulfur dioxide 0.62 (Figure 3.3.6), which was anticipated. Because the amount of total sulfur dioxide = amount of Free sulfur dioxide + bound forms of sulfur dioxide. Total sulfur dioxide has a negative correlation with an alcohol percentage of -0.27 (Figure 3.3.7). So, an increase in total sulfur dioxide means a mild decrease in alcohol (% vol.). Total sulfur dioxide has a negative correlation with quality, with a value of -0.20.

Patterns in density:

The density feature is normally distributed (Figure 3.1.8). Density has a very strong positive correlation with fixed acidity with a value of 0.61, with residual sugar 0.40, with chlorides 0.43 (Table 3.4.1). Residual sugar is the amount of remaining sugar after fermentation has been stopped. More dense wine seems sweeter, more salty, and more acidic wine. Density has a very strong negative correlation with alcohol (% vol.) -0.56 (Figure 3.3.8). High-density wine is more watery wine and has less alcohol. The density of water is higher than the density of alcohol. Density has a negative correlation with pH and quality, with a value of -0.24 (Figure 3.3.8). An increase in density mildly decreases the quality (Figure 3.2.5). More dense wine means more watery wine.

Patterns in pH:

The pH feature is normally distributed (Figure 3.1.9). In the best quality wines rated as 7 or 8, the average pH value is between 3.2 and 3.3 (Figure 3.2.6 and Table 2.2.3). In wine quality, the pH value varies in decimal points. This decimal place variation should not be that much fillable in the tongue. pH has a very strong negative correlation with fixed acidity -0.70 and with citric acid -0.49. This is obvious because a low pH value is more acidic. pH has a mild positive correlation with volatile acidity 0.24 (which is odd) and a mild negative correlation with density -0.24 (Table 3.4.1). pH has a very weak negative correlation with wine quality, with a value of -0.07 (Table 3.4.1), which is negligible.

Patterns in Sulphates:

Sulphites in wine are used to stop fermentation at a specific point in the winemaking process [1]. It works as an antioxidant and antimicrobial in the wine. Sulphates distribution is slightly right skewed (Figure 3.1.10). The mean, median, and mode are around 0.6 g/dm^3. and 75% of the data is within the value of 7 g/dm3 (Figure 3.1.10). Sulphate has a strong negative correlation with volatile acidity -0.31 (Table 3.4.1). Sulphate has a very strong positive correlation with wine quality, with a value of 0.42 (Table 3.4.1). We can say that the higher the amount of sulphate, the higher the wine quality. Which is also supported by other studies in [2][3].

Patterns in alcohol:

Alcohol features distribution is slightly right-skewed (Figure 3.1.11). Alcohol has a very strong negative correlation with density -0.56 (Figure 3.3.8) (Table 3.4.1). Water density is higher than the density of alcohol. An increase in alcohol (% vol.) is a decrease in wine density. So, we can say that high-density wine is more watery wine with less alcohol. Alcohol has a strong negative correlation with chlorides -0.30 (Figure 3.3.5). So, wine with a high percentage of alcohol is less salty and also less in density. As we also observed before in the chart, that increase in alcohol percentage is an increase in the wine quality (Figure 3.2.10). There is a strong positive correlation between alcohol percentage and wine quality, with a value of 0.51 (Table 3.4.1), which is the highest positive correlation of wine quality with other wine features, which is also supported by another analysis [2][3]. In best quality wine, which was rated as 7 or 8, the average alcohol percentage is between 11% and 13% (Figure 3.2.9). We should order a wine where the alcohol percentage is within this range. The best choice could be around 12 (% vol.) alcohol. Alcohol also has a mild positive correlation with sulphates, with a value of 0.27 (Table 3.4.1), and a mild negative correlation with total sulfur dioxide, with a value of -0.27 (Table 3.4.1). It also has a small negative correlation with volatile acidity with a value of -0.22 (Table 3.4.1).

Summary of the patterns and impact on quality:

Most of the wine's quality is either 5 (with 438 observations) or 6 (with 429 observations) (Table 3.2.1). There are only 03 observations of wine quality of level 3, which is not enough to come up with any decision for this type (level 3) of wine. We have 117 observations of level-7 wine and 11 observations of level-8-rated wine (Table 3.2.1).

Red wine quality has the strongest positive correlation with alcohol (% vol.), with a value of 0.51 (Figure 3.3.11). It makes sense that a higher amount of alcohol (% vol.) would bring better customer satisfaction. That is also supported by other analyses [2][3]. Additionally, a simple regression model with alcohol (independent variable) and wine quality (dependent variable) showed that the alcohol variable is statistically significant (Table 3.5.11), and by the R-squared value, this simple regression model with alcohol variable explains about 26% of the data (which is the highest compared to the other independent wine features). So, alcohol (% vol.) is the most important physicochemical characteristic of red wine quality.

Subsequently, the second strongest correlation of wine quality is with the sulphates additive, with a positive value of 0.41. It seems like people rate the wine higher quality when more sulphate additives are added to the wine-making process. Moreover, a simple regression model with sulphates (independent variable) and wine quality (dependent variable) exhibited that the sulphates variable is statistically significant (Table 3.5.10), and by the R-squared value, this simple regression model with sulphates variable explains about 18% of the data (which is the second highest value among the independent features). Therefore, sulphate is the second most important physicochemical characteristic of red wine.

Next, wine quality has the third strongest correlation with the volatile acidity with a negative value of -0.35. This was expected because too high acetic acid (g/dm3) can lead to an unpleasant vinegar taste. The lesser the amount of volatile acidity, the greater the wine quality. which is also supported by other studies [2][3]. Additionally, a simple regression model with volatile acidity (independent variable) and wine quality (dependent variable) showed that the volatile acidity variable is statistically significant (Table 3.5.2), and by the R-squared value, this simple regression model with volatile acidity variable explains about 12% of the data (which is the third highest value among the independent features). Therefore, volatile acidity negatively plays the third most important role in wine quality.

The fourth highest correlated value with quality is with the density, which is a mild negative correlation, with a value of -0.24. Moreover, a simple regression model with density (independent variable) and quality (dependent variable) showed that the density variable is statistically significant (Table 3.5.8), and by the R-squared value, the regression model with density variable explains only 6% of the data (which is the fourth highest value among the independent features). Therefore, Wine density is the fourth most important factor for the red wine quality.

After that, wine quality has a mild positive correlation with citric acid 0.22. Even though it adds freshness and flavor to wines, in a simple regression model with quality (dependent variable) and citric acid (independent variable), by the R-squared value, the citric acid variable explains only 5% of the data, which is the fifth-highest value among the independent variables. The possible reason is that it exists in a very small amount in the wine (Table 3.1.3).

Wine quality also has a very weak negative correlation with the pH score with a value of -0.07, and in the simple regression model with quality (dependent variable) and pH (independent variable), by the R-squared value, the pH variable explains about only 0.5% of the data, which is very small.

5. Discussion

Assumption 01:

As presented in Section 2.3, I initially assumed that since the amount of fixed acidity is the highest quantity among all other types of acid (volatile acidity and citric acid (Table 2.2.2)) and since it does not evaporate readily (non-volatile). It will dominate the pH level. An increase in fixed acidity will make the wine sourer. Also, it could decrease the wine quality. After analyzing the data, I have found that fixed acidity dominates the pH level significantly. Fixed acidity has a very strong negative correlation value with a pH score of -0.70. In fact, it is the highest correlated feature among the features. This means that I was also correct that an increase in fixed acidity will make the wine sourer. However, I was wrong about an increase in fixed acidity, which will decrease the wine quality. However, fixed acidity has a very little positive correlation with the wine quality, with a value of 0.11 (Figure 3.3.1), which means the increase of fixed acidity will increase the wine quality slightly. The correlation is very weak even though the assumption was quite the opposite.

Moreover, as presented in Table 3.5.1, a simple regression model with fixed acidity (independent variable) and wine quality (dependent variable) showed that the fixed acidity variable is statistically significant. By the R-squared value, the regression model with fixed acidity variable explains about 01% of the data. So, I was partly right about my 1st assumption.

Assumption 02:

My second assumption was that since the high amount of volatile acidity in wine (acetic acid) produces an unpleasant vinegar flavor (Cortez et al. (2009).) so, the increase of volatile acidity will decrease the wine quality. After analyzing the data, I have found that there is a strong negative correlation between volatile acidity and wine quality, with a value of -0.35 (Table 3.3.2). As presented in Table 3.5.2, a simple regression model with volatile acidity (independent variable) and wine quality (dependent variable) showed that the volatile acidity variable is statistically significant. By the R-squared value, this simple regression model with volatile acidity variable explains about 12% of the data. Thus, it proves my assumption to be true.

Assumption 03:

My third assumption was also proved true: increased citric acid will increase the wine quality.

In red wine, citric acid is present in a very small amount in the wine (Table 2.2.2). Around 9% of red wine has no citric acid in it. After removing the outliers, there are only 03 observations of wine quality of level 3. This is not enough observation to come to any conclusion for level 3 wine. In level 3 wine, we saw that an increase in it decreased the quality. But from the rest of the data, we can see that the greater the amount of citric acid, the greater the quality (Figure 3.2.3). Also, the highest amount of citric acid is 0.73, which was found in a level 6-rated wine. There is very little chance that a higher value can reduce the wine quality heavily. Since, it is generally present in wine with a very small amount. As per the analysis, there is a mild positive correlation between citric acid and wine quality, with a value of 0.22 (Table 3.4.1). Moreover, as presented in Table 3.5.3, a simple regression model with citric acid (independent variable) and wine quality (dependent variable) showed that the citric acid variable is statistically significant. By the R-squared value, this simple regression model with citric acid variable explains about 5% of the data.

Assumption 04:

My fourth assumption was proved as partly true that an increase of different acid substances (fixed acidity, volatile acidity, and citric acid) will decrease the wine's pH score/level, and since a low score of pH value is very acidic and very sour thus poor in wine quality. pH has a very strong negative correlation with fixed acidity with a value of -0.70 and with citric acid -0.49 (Table 3.4.1). But pH has a mild positive correlation with volatile acidity with a value of 0.24 (Table 3.4.1), which seems odd. In wine quality, the pH value varies in decimal point (Figure 3.2.6) (Table 2.2.2). The decimal fraction variation should not be that noticeable in the tongue. Red wine is mild in the sour/acidic test (Figure 3.2.8). An increase by decimal point, the pH value will cause very little decrease in the wine quality (Figure 3.2.6).

The pH has a minimal negative correlation with the wine quality, with a value of -0.07 (Table 3.4.1), which is a very weak correlation. Additionally, as presented in Table 3.5.9, a simple regression model with pH (independent variable) and wine quality (dependent variable) showed that the pH variable is statistically significant. By the R-squared value, this simple regression model with pH variable explains only about 0.5% of the data.

Assumption 05:

My fifth assumption was proven partly true. I assumed that an increase in residual sugar would increase the wine density. More residual sugar will be more watery wine with a lesser alcohol percentage. An increase in the amount of residual sugar will decrease the wine quality. Residual sugar has a very strong positive correlation with density with a value of 0.40 (Figure 3.3.4). So, I was right. Alcohol is less dense than water. So, more residual sugar is more watery wine. Residual sugar has a very weak positive correlation with alcohol, with a value of 0.08 (Table 3.4.1). This is the opposite of my initial assumption that more residual sugar will be with a lesser alcohol percentage. Residual sugar has a negligible positive correlation with the wine quality, with value of 0.02 (Table 3.3.11). This is a very weak correlation. Additionally, as presented in Table 3.5.4, a simple regression model with residual sugar (independent variable) and wine quality (dependent variable) showed that the residual sugar variable is not statistically significant.

Assumption 06:

My sixth assumption was true. An increase in chlorides can lead to a decrease in quality. More chlorides mean more saltier wine. Chlorides have a strong positive correlation with density, with a value of 0.43 (Figure 3.3.5). It has a strong negative correlation with alcohol -0.30 (Figure 3.3.5). More salty wine means higher in density and lower in alcohol percentage. Chloride has a small negative correlation with wine quality, with a value of -0.18 (Table 3.3.11). Furthermore, as presented in Table 3.5.5, a simple regression model with chlorides (independent variable) and wine quality (dependent variable) showed that the chlorides variable is statistically significant. By the R-squared value, this simple regression model with chloride variable explains about 3% of the data.

Assumption 07:

My seventh assumption was true. An increase in density will decrease the wine quality. At room temperature, the alcohol density is around 0.79 g/cm3, and water density is around 0.99 g/cm3. An increase in density means more watery wine. Density has a very high positive correlation with fixed acidity with a value of 0.61, residual sugar 0.40, and chlorides 0.43 (Table 3.4.1). Density also has a negative correlation with pH, with a value of -0.24 (Figure 3.4.1). More dense wine seems sweeter, saltier, more watery, and more acidic/sour wine. Density strongly correlates negatively with an alcohol percentage of -0.56 (Figure 3.3.8). High-density wine is more watery wine and has less alcohol. In Figure 3.2.5, the increase in density decreases in quality. Density negatively correlates with wine quality, with a value of -0.24 (Figure 3.3.8). In addition, as presented in Table 3.5.8, a simple regression model with density (independent variable) and wine quality (dependent variable) showed that the density variable is statistically significant. By the R-squared value, this simple regression model with density variable explains about 6% of the data.

Assumption 08:

Finally, my eighth assumption was true that an increase in alcohol percentage would increase the wine quality. Alcohol has a very strong negative correlation with density -0.56 (Figure 3.3.8) (Table 3.4.1). An increase in alcohol percentage is a decrease in wine density. High-density wine is more watery wine. Alcohol has a strong negative correlation with chlorides -0.30 (Figure 3.3.5). A high percentage of alcohol in wine is less salty. According to Figure 3.2.10, an increase in average alcohol percentage is an increase in the wine quality. There is a very high positive correlation between alcohol percentage and wine quality, with a value of 0.51 (Table 3.4.1), which is the highest positive correlation of wine quality among all the wine features. Besides, as presented in Table 3.5.11, a simple regression model with alcohol (independent variable) and wine quality (dependent variable) showed that the alcohol variable is statistically significant. By the R-squared value, this simple regression model with the alcohol variable explains about 26% of the data.

6. Conclusion

In 2023, the wine market revenue amounted to USD 333.00 billion. The wine market is expected to grow annually by 5.52% (CAGR 2023-2027) [4]. As a result, companies are trying to acquire new technologies to progress wine production, quality, and sales. Quality certification is a vital phase in all the processes, which now relies heavily on experienced human wine tasting. I looked for a method that would allow me to evaluate red wines scientifically using their physicochemical properties. A sizable dataset (1599 red observations) was considered, including Vinho Verde samples from Portugal's northwest through Kaggle [1]. Firstly, data quality has been checked, and some data preprocessed. After that, I have made some assumptions and hypotheses. To check assumptions and hypotheses, I have done Exploratory Data analysis (EDA) on the data and look for different Trends, Patterns, and Anomalies in the data. I have also done different Simple and Multiple Linear Regressions where quality was always the dependent variable. The multiple linear regression model explains almost 0.389 (39%) of the variability of the wine quality.

Businesses can use this analysis report before ordering red wine in large quantities as this analysis will help to evaluate the red wine quality based on its physicochemical properties and could be a great help for the businesses to ensure choosing red wine of good quality and thus could maintain a good business. For instance, as identified in this analysis, alcohol quantity is the most important substance for red wine quality. The business should give more emphasis on the alcohol quantity. For the best rating, they should collect wine, where the alcohol percentage is around 12%. Sulphate additives play the second most important role in wine quality. Both in the production level and collection phase, this should be the second most considerable factor. Since volatile acidity negatively plays the third most important role in wine quality, it could lead to an unpleasant vinegar taste. Businesses keep an eye on it's (acetic acid)/dm3) quantity in the wine. Wine density is the fourth most important factor for red wine quality.

Wine manufacturers can also use the results of this analysis to enhance the wine's quality because some production-related elements are controllable. For instance, by keeping an eye on the sugar content of the grapes before harvest, the alcohol percentage can be raised or lowered. Additionally, the residual sugar in wine may be controlled by stopping the yeast's process of fermenting sugar by sulphates.

7. References

1. https://www.eufic.org/en/whats-in-food/article/what-are-sulphites-in-wine-and-are-they-bad-for-you, Accessed: 14 October 2023
2. P. Cortez, A. Cerdeira. F. Almeida, T. Matos, J. Reis, Modeling wine preferences by data mining from physicochemical properties, Retrieved from: https://repositorium.sdum.uminho.pt/bitstream/1822/10029/1/wine5.pdf, Accessed: 14 October 2023
3. Paulo Cortez, Juliana Teixeira, Antonio Cerdeira, Fernando Almeida, Elmo Matos, José Reis, Using Data Mining for Wine Quality Assessment, Retrieved from: https://www.researchgate.net/publication/221612614_Using_Data_Mining_for_Wine_Quality_Assessment, Accessed: 14 October 2023
4. https://www.statista.com/outlook/cmo/alcoholic-drinks/wine/worldwide, Accessed: 14 October 2023
5. Image 1: Figure 3.2.7, pH scale, https://www.careerpower.in/blog/ph-values-list, Accessed: 14 October 2023
6. Image 2: Figure 3.2.8, pH scale value in different item, https://www.physics.uoguelph.ca/acids-bases-and-ph, Accessed: 14 October 2023
7. UCI machine learning, Red Wine Quality, https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009, Accessed: 04 October 2023