Data Analyst Exam Project

Report on Red Wine Quality Analysis & Prediction

BY: Shloka Sardal

DATE: 10/03/2024

Contents

Introduction

- The dataset consists of physicochemical properties on which the quality of red variants of Portuguese Vinho Verde wine is based. It contains 1599 samples and 12 variables, out of which 11 are input variables and 1 is output/target variable. All variables are numerical. A report has been prepared to predict the quality of red wine. It is important to identify and analyze the main factors that positively and negatively affect the quality of wine so that more emphasis is placed on the positive factors and negative factors are minimized.
- "Red wine quality and style are highly influenced by the qualitative and quantitative composition of aromatic compounds having various chemical structures and properties and their interaction within different red wine matrices. The understanding of interactions between the wine matrix and volatile compounds and the impact on the overall flavor as well as on typical or specific aromas is getting more and more important for the creation of certain wine styles." (Doris Rauhut, 2019)
- The following data processes and techniques are carried out to predict the quality of red wine: -

➢ Exploratory Data Analysis (EDA) allows us to analyze trends and patterns, and spot anomalies in the dataset.

➢ As a part of EDA descriptive statistics of all the numerical variables viz. fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol are performed to get an insight about whether the distribution is normal or skewed and other central tendency measures.

➢ Pivot tables and pivot charts are prepared to analyze how the input variables influence the target variable.

➢ The correlation of all the variables is calculated to analyze which variables strongly correlate with the output variable i.e. quality.

- Multiple regression is run to find out the intercept and parameter estimates of the slope coefficient.
- The statistical significance of parameter estimates is analyzed by checking the p-value and then concluding it with null hypothesis testing.
- The intercept and parameter estimates of the slope coefficient are used to predict the quality of wine.

Body

Initial Assumptions

- Data values have a normal distribution.
- Data values have a linear relationship.

Hypotheses

- Fixed and volatile acids would be an influential(positive) factor in improving the quality of red wine.
- Citric acid would improve the quality of red wine, as citric acid can add freshness and flavor to the wine.
- Residual sugar would improve the quality of red wine.
- A high number of chlorides would not improve the quality of red wine.
- Free sulfur dioxide and total sulfur dioxide improve the quality of red wine.
- The lower the density of red wine the better the quality of red wine.
- High pH values would not improve the quality of red wine.
- Sulphates would improve the quality of red wine.
- Alcohol would improve the quality of red wine.

○ Exploratory Data Analysis

❖ Descriptive Statistics

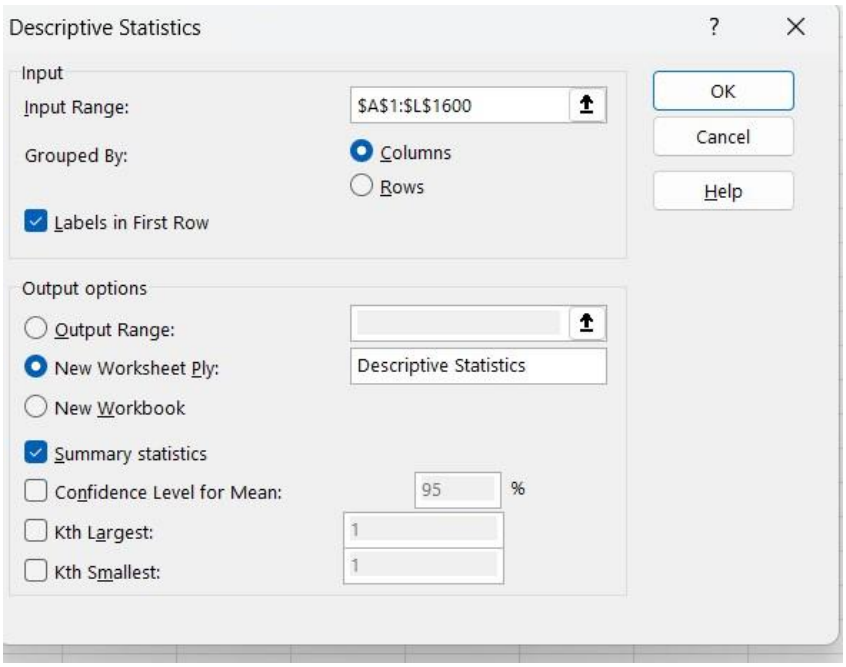


Figure No. 1

- Configuration of Descriptive Statistics is seen in Figure No.1.

1) Fixed acidity

fixed acidity	
Mean	8.319637273
Standard Error	0.043541017
Median	7.9
Mode	7.2
Standard Deviation	1.741096318
Sample Variance	3.031416389
Kurtosis	1.132143398
Skewness	0.982751441
Range	11.3
Minimum	4.6
Maximum	15.9
Sum	13303.1
Count	1599

Figure No. 2

The table in Figure No.2 states descriptive statistics for the variable fixed acidity.

- Average fixed acidity in wine is 8.32 g/dm^3
- Median is 7.9 g/dm^3
- Mode is 7.2 g/dm^3
- Standard deviation is 1.74
- Kurtosis is 1.13, since kurtosis is less than 3 it means the distribution corresponds to a flatter peak and a shorter tail.
- Skewness is 0.98, which means the distribution is positive and moderately skewed.
- Range is 11.3
- Minimum fixed acidity is 4.6 g/dm^3
- Maximum fixed acidity is 15.9 g/dm^3

2) Volatile acidity

volatile acidity	
Mean	0.527820513
Standard Error	0.004477892
Median	0.52
Mode	0.6
Standard Deviatior	0.179059704
Sample Variance	0.032062378
Kurtosis	1.22554225
Skewness	0.671592572
Range	1.46
Minimum	0.12
Maximum	1.58
Sum	843.985
Count	1599

Figure No.3

The table in Figure No.3 states descriptive statistics for the variable volatile acidity.

- The average volatile acidity in wine is 0.53 g/dm^3
- The median is 0.52 g/dm^3
- Mode is 0.6 g/dm^3
- Standard Deviation is 0.18
- Kurtosis is 1.23 since kurtosis is less than 3 it means the distribution corresponds to a flatter peak and a shorter tail.
- Skewness is 0.67, which means the distribution is positive and moderately skewed.
- Range is 1.46
- Minimum volatile acidity is 0.12 g/dm^3
- Maximum volatile acidity is 1.58 g/dm^3

3) Citric acid

citric acid	
Mean	0.27097561
Standard Error	0.004871551
Median	0.26
Mode	0
Standard Deviation	0.194801137
Sample Variance	0.037947483
Kurtosis	-0.788997515
Skewness	0.318337295
Range	1
Minimum	0
Maximum	1
Sum	433.29

Figure No. 4

The table in Figure No.4 states descriptive statistics for the variable citric acid.

- The average citric acid in wine is 0.27 g/dm^3
- The median is 0.26 g/dm^3
- Mode is 0
- Standard Deviation is 0.19
- Kurtosis is -0.79, since kurtosis is less than 3 it means the distribution corresponds to a flatter peak and a shorter tail.
- Skewness is 0.32, which means the distribution is positive and symmetric.
- Range is 1
- Minimum citric acid is 0 g/dm^3
- Maximum citric acid is 1 g/dm^3

4) Residual sugar

residual sugar	
Mean	2.5388055
Standard Error	0.03525922
Median	2.2
Mode	2
Standard Deviation	1.40992806
Sample Variance	1.98789713
Kurtosis	28.6175954
Skewness	4.54065543
Range	14.6
Minimum	0.9
Maximum	15.5
Sum	4059.55

Figure No. 5

The table in Figure No.5 states descriptive statistics for the variable residual sugar.

- The average residual sugar in wine is 2.54 g/dm^3
- Median is 2.2 g/dm^3
- Mode is 2 g/dm^3
- Standard Deviation is 1.41
- Kurtosis is 28.61,since kurtosis is greater than 3, which means the distribution corresponds to a higher peak than normal distribution and a taller tail.
- Skewness is 4.54, which means the distribution is positive and heavily skewed.
- Range is 14.6
- Minimum residual sugar is 0.9 g/dm^3
- Maximum residual sugar is 15.5 g/dm^3

5) Chlorides

chlorides	
Mean	0.087466542
Standard Error	0.001177
Median	0.079
Mode	0.08
Standard Deviation	0.047065302
Sample Variance	0.002215143
Kurtosis	41.71578725
Skewness	5.680346572
Range	0.599
Minimum	0.012
Maximum	0.611
Sum	139.859

Figure No. 6

The table in Figure No.6 states descriptive statistics for the variable chlorides.

- The average of chlorides in a wine is 0.09 g/dm^3
- Median is 0.08 g/dm^3
- Mode is 0.08 g/dm^3
- Standard Deviation is 0.05
- Kurtosis is 41.72, since kurtosis is greater than 3, which means the distribution corresponds to a higher peak than normal distribution and a taller tail.
- Skewness is 5.68, which means the distribution is positive and heavily skewed.
- Range is 0.59
- Minimum number of chlorides in wine is 0.01 g/dm^3
- Maximum number of chlorides in wine is 0.61 g/dm^3

6) Free sulfur dioxide

free sulfur dioxide	
Mean	15.87492183
Standard Error	0.261585683
Median	14
Mode	6
Standard Deviation	10.46015697
Sample Variance	109.4148838
Kurtosis	2.023562046
Skewness	1.250567293
Range	71
Minimum	1
Maximum	72
Sum	25384

Figure No. 7

The table in Figure No.7 states descriptive statistics for the variable free sulfur dioxide.

- The average of free sulfur dioxide in wine is 15.87 mg/dm^3
- Median is 14 mg/dm^3
- Mode is 6 mg/dm^3
- Standard Deviation is 10.46, which means the data points are dispersed.
- Kurtosis is 2.02 since kurtosis is less than 3 it means the distribution corresponds to a flatter peak and a shorter tail.
- Skewness is 1.25 which means the distribution is positive and heavily skewed
- Range is 71
- Minimum free sulfur dioxide in wine is 1 mg/dm^3
- Maximum free sulfur dioxide in wine is 72 mg/dm^3

7) Total sulfur dioxide 2

total sulfur dioxide 2	
Mean	43.1167783
Standard Error	0.66895505
Median	38
Mode	46.4677924
Standard Deviatior	26.7498389
Sample Variance	715.553881
Kurtosis	0.14804131
Skewness	0.9173548
Range	116
Minimum	6
Maximum	122
Sum	68943.7286

Figure No. 8

The table in Figure No.8 states descriptive statistics for the variable total sulfur dioxide2.

- The average total sulfur dioxide 43.12mg/dm^3
- Median is 38 mg/dm^3
- Mode is 46.46 mg/dm^3
- Standard Deviation is 26.74, which means data points are highly dispersed.
- Kurtosis is 0.14, since kurtosis is less than 3 it means the distribution corresponds to a flatter peak and a shorter tail.
- Skewness is 0.91, which means the distribution is positive and moderately skewed.
- Range is 116
- Minimum total sulfur dioxide is 6 mg/dm^3
- Maximum total sulfur dioxide is 122 mg/dm^3

8) density

density	
Mean	0.996746679
Standard Error	4.71981E-05
Median	0.99675
Mode	0.9972
Standard Deviation	0.001887334
Sample Variance	3.56203E-06
Kurtosis	0.934079065
Skewness	0.071287663
Range	0.01362
Minimum	0.99007
Maximum	1.00369
Sum	1593.79794

Figure No. 9

The table in Figure No.9 states descriptive statistics for the variable density.

- The average density of the wine is 0.99 g/cm^3
- Median is 0.99 g/cm^3
- Mode is 0.99 g/cm^3
- Standard Deviation is 0.001
- Kurtosis is 0.93, since kurtosis is less than 3 it means the distribution corresponds to a flatter peak and a shorter tail.
- Skewness is 0.07 which means the distribution is positive and symmetric.
- Range is 0.01
- Minimum density of wine is 0.99
- Maximum density of wine is 1

9) pH

pH	
Mean	3.311
Standard Error	0.004
Median	3.31
Mode	3.3
Standard Deviation	0.154
Sample Variance	0.024
Kurtosis	0.807
Skewness	0.194
Range	1.27
Minimum	2.74
Maximum	4.01
Sum	5294

Figure No. 10

The table in Figure No.10 states descriptive statistics for the variable pH.

- The average pH level is 3.31 units
- Median is 3.31 units
- Mode is 3.3 units
- Standard Deviation is 0.15
- Kurtosis is 0.81, since kurtosis is less than 3 it means the distribution corresponds to a flatter peak and a shorter tail.
- Skewness is 0.19, which means that the distribution is positive and symmetric.
- Range is 1.27
- Minimum pH level is 2.74 units.
- Maximum pH level is 4.01 units.

10) sulphates

sulphates	
Mean	0.658
Standard Error	0.004
Median	0.62
Mode	0.6
Standard Deviation	0.17
Sample Variance	0.029
Kurtosis	11.72
Skewness	2.429
Range	1.67
Minimum	0.33
Maximum	2
Sum	1052

Figure No. 11

The table in Figure No.11 states descriptive statistics for the variable sulphates.

- The average sulphates in wine is 0.66 g /dm^3
- The median is 0.62 g /dm^3
- Mode is 0.6 g /dm^3
- The standard Deviation is 0.17
- Kurtosis is 11.72, kurtosis is greater than 3, which means the distribution corresponds to a higher peak than normal distribution and a taller tail.
- Skewness is 2.43, which means that the distribution is positive and heavily skewed.
- Range is 1.67
- Minimum sulphates in wine is 0.33 g /dm^3
- Maximum sulphates in wine is 2 g /dm^3

11) alcohol

alcohol	
Mean	10.42
Standard Error	0.027
Median	10.2
Mode	9.5
Standard Deviation	1.066
Sample Variance	1.136
Kurtosis	0.2
Skewness	0.861
Range	6.5
Minimum	8.4
Maximum	14.9
Sum	16666

Figure No. 12

The table in Figure No.12 states descriptive statistics for the variable alcohol.

- The average percentage of alcohol in wine is 10.42%
- Median is 10.2%
- Mode is 9.5%
- Standard Deviation is 1.06
- Kurtosis is 0.20, since kurtosis is less than 3 it means the distribution corresponds to a flatter peak and a shorter tail.
- Skewness is 0.86, and the distribution is positive and moderately skewed.
- Range is 6.5
- Minimum percentage of alcohol in wine is 8.4%
- Maximum percentage of alcohol in wine is 14.9%

❖ Pivot tables and Pivot Charts

- Quality count based on fixed acidity.

fixed acidity	Count of quality
4.6	1
4.7	1
4.9	1
5	6
5.1	4
5.2	6
5.3	4
5.4	5
5.5	1
5.6	14
5.7	2
5.8	4
5.9	9
6	13
6.1	16
6.2	20
6.3	14
6.4	25
6.5	17
6.6	37
6.7	28
6.8	46

Figure No.13

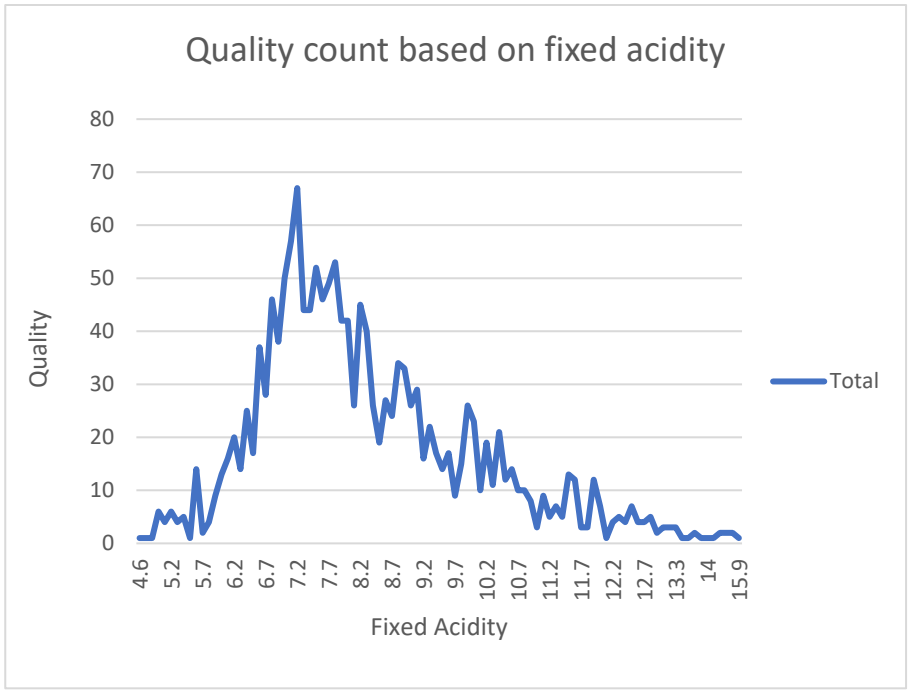


Figure No.14

- It is observed in Figure No. 13 and 14, that fixed acidity at the level of 7.2g/dm^3 has the highest quality count.

- Quality count based on volatile acidity

volatile acidity	Count of quality
0.12	3
0.16	2
0.18	10
0.19	2
0.2	3
0.21	6
0.22	6
0.23	5
0.24	13
0.25	7
0.26	16
0.27	14
0.28	23
0.29	16
0.295	1
0.3	16
0.305	2
0.31	30
0.315	2
0.32	23
0.33	20
0.34	30
0.35	22

Figure No.15

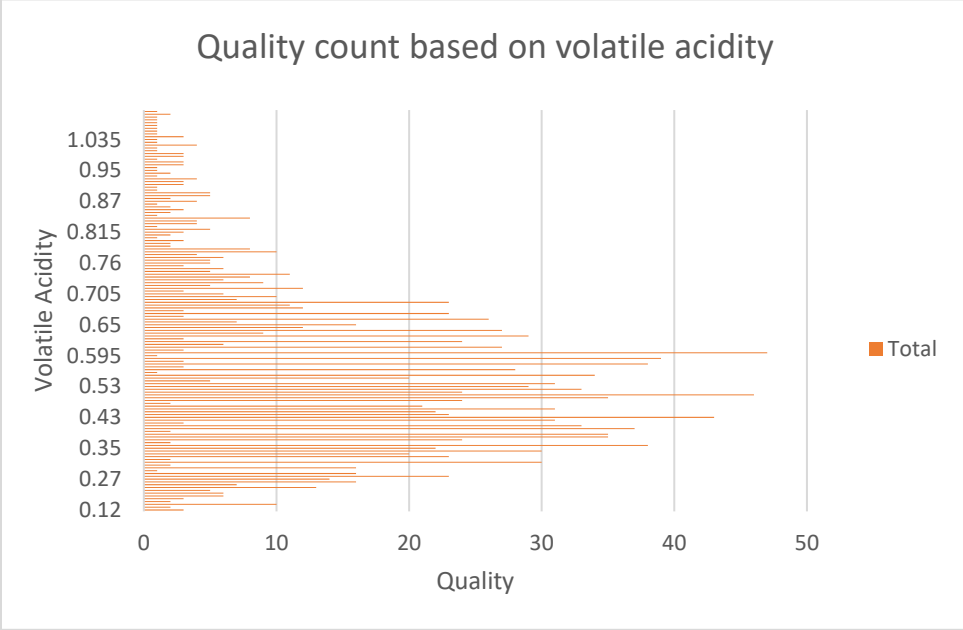


Figure No.16

- It is observed in Figure No. 15 and 16 that volatile acidity at the level of 0.6 g/dm^3 has the highest quality count.

- Quality count based on citric acid

citric acid	Count of quality
0	132
0.01	33
0.02	50
0.03	30
0.04	29
0.05	20
0.06	24
0.07	22
0.08	33
0.09	30
0.1	35
0.11	15
0.12	27
0.13	18
0.14	21

Figure No.17

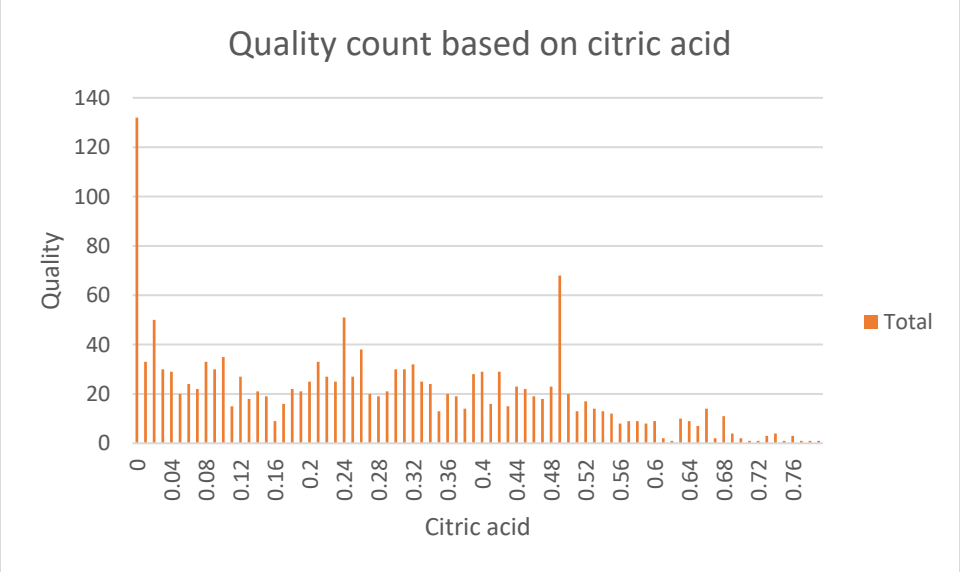


Figure No.18

- It is observed in Figure No. 17 and 18 that citric acid at a level of 0 g/dm^3 has the highest quality count.

- Quality count based on residual sugar

residual sugar	Count of quality
0.9	2
1.2	8
1.3	5
1.4	35
1.5	30
1.6	58
1.65	2
1.7	76
1.75	2
1.8	129
1.9	117
2	156
2.05	2
2.1	128

Figure No.19

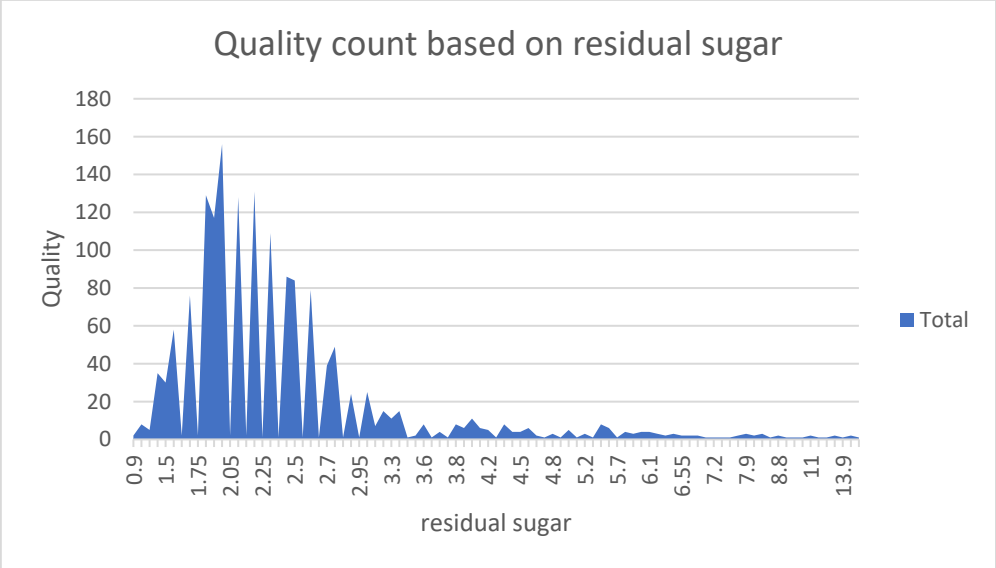


Figure No.20

- It is observed in Figure No. 19 and 20 that residual sugar at a level of 2 g/dm^3 has the highest quality count.

❖ Dashboard No. 1

Figure No. 21

- Four charts are presented in Dashboard 1 in Figure No. 21 are as follows:-

➢ Quality count based on fixed acidity

➢ Quality count based on citric acid

➢ Quality count based on volatile acidity

➢ Quality count based on residual sugar

- Quality count based on chlorides

Chlorides	Count of quality
0.012	2
0.034	1
0.038	2
0.039	4
0.041	4
0.042	3
0.043	1
0.044	5
0.045	4
0.046	4
0.047	4
0.048	8
0.049	8
0.05	12

Figure No. 22

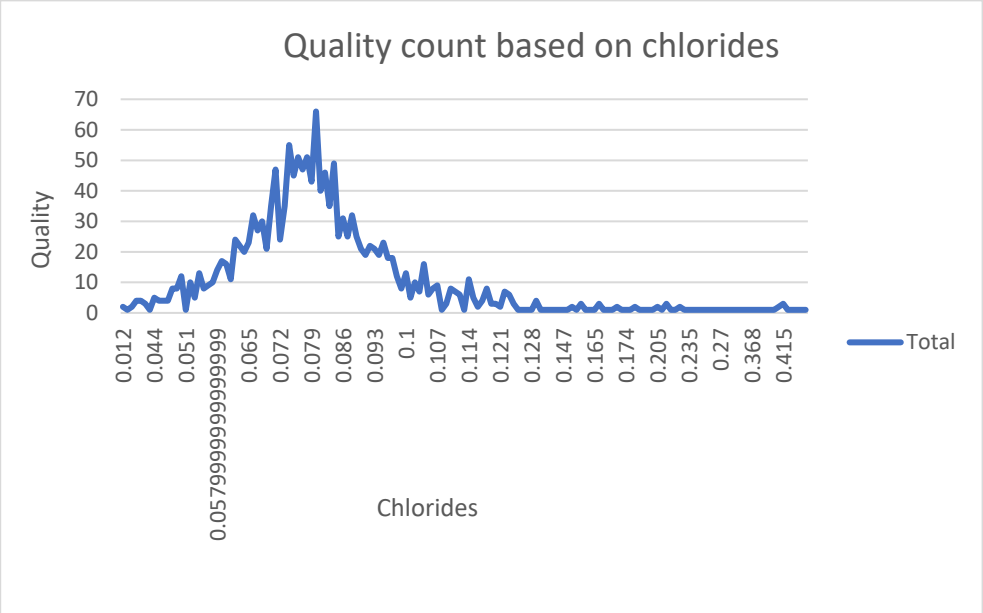


Figure No. 23

- It is observed in Figure No. 22 and 23 that chlorides at a level of 0.08 g/dm^3 have the highest quality count.

- Quality based on free sulfur dioxide

free sulfur dioxide	Count of quality
1	3
2	1
3	49
4	41
5	104
5.5	1
6	138
7	71
8	56
9	62
10	79
11	59
12	75
13	57
14	50
15	78
16	61
17	60
18	46
19	39
20	30
21	41

Figure No. 24

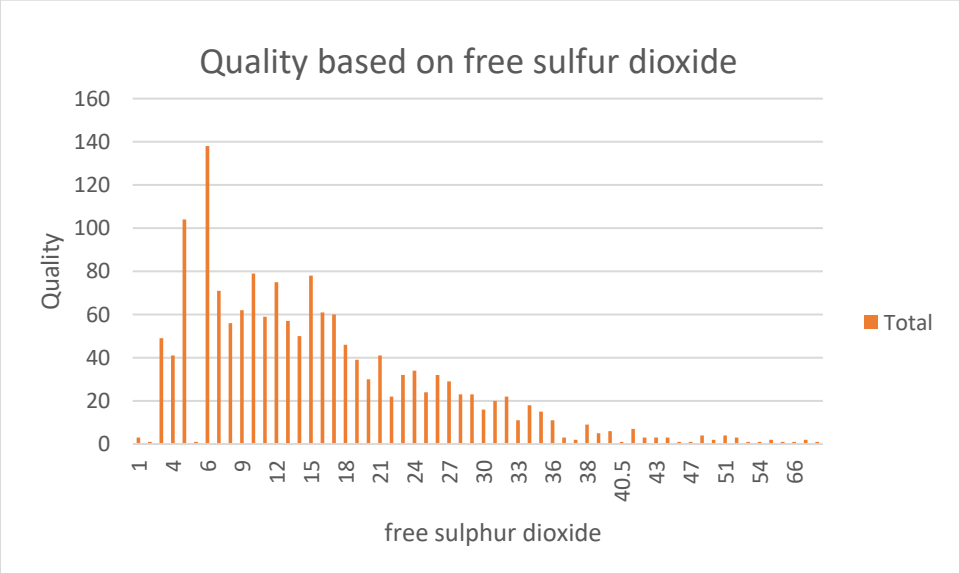


Figure No. 25

- It is observed in Figure No. 24 and 25 that free sulfur dioxide at a level of 6 mg/dm^3 has the highest quality count.

- Quality based on total sulfur dioxide 2

Total Sulfur dioxide 2	Count of quality
6	3
7	4
8	14
9	14
10	27
11	26
12	29
13	28
14	33
15	35
16	26
17	27
18	35
19	29
20	33
21	25

Figure No. 26

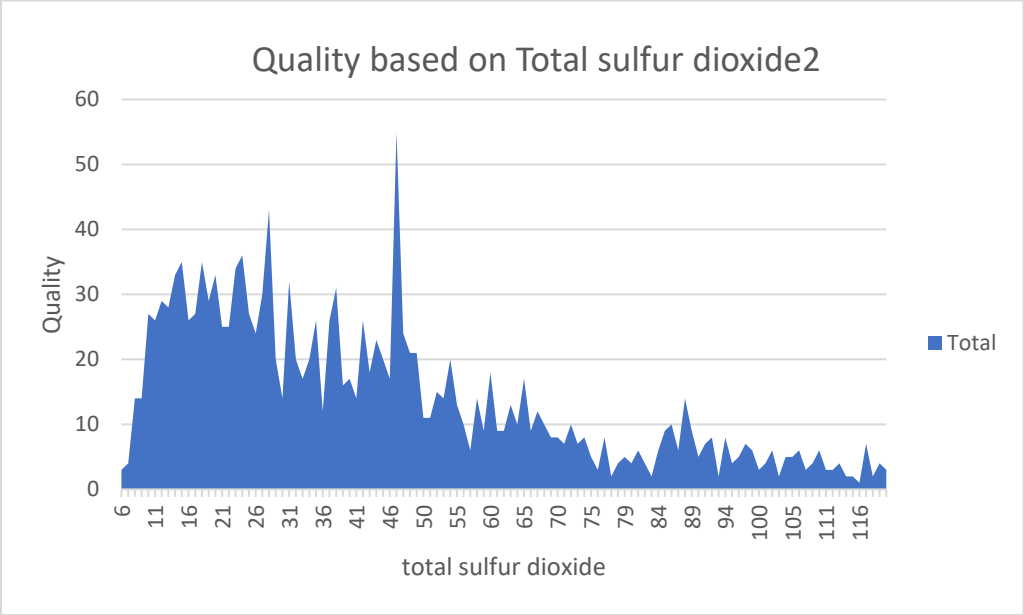


Figure No. 27

- It is observed in Figure No. 26 and 27 that total sulfur dioxide2 at a level of 46.46 mg/dm^3 has the highest quality count.

- Quality count based on density

density	Count of quality
0.99007	2
0.9902	1
0.99064	2
0.9908	1
0.99084	1
0.9912	1
0.9915	1
0.99154	1
0.99157	1
0.9916	2
0.99162	1
0.9917	1
0.99182	2
0.99191	1
0.9921	1
0.9922	2
0.99235	1
0.99236	1

Figure No. 28

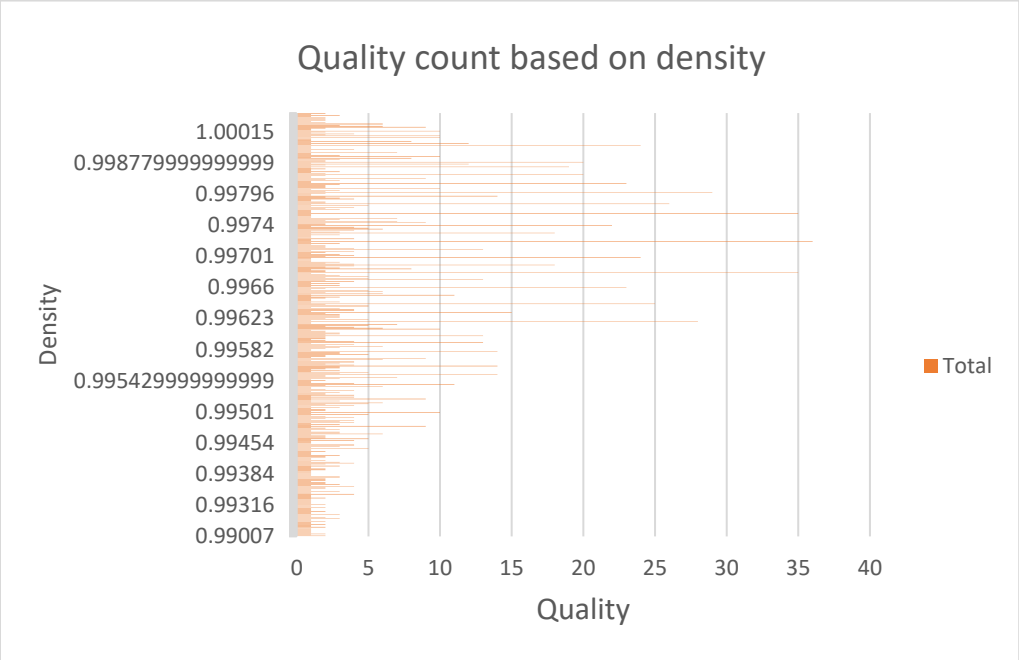


Figure No. 29

- It is observed in Figure No. 28 and 29 that the density at a level of 0.9972 g/cm3 has the highest quality count.

❖ Dashboard No.2

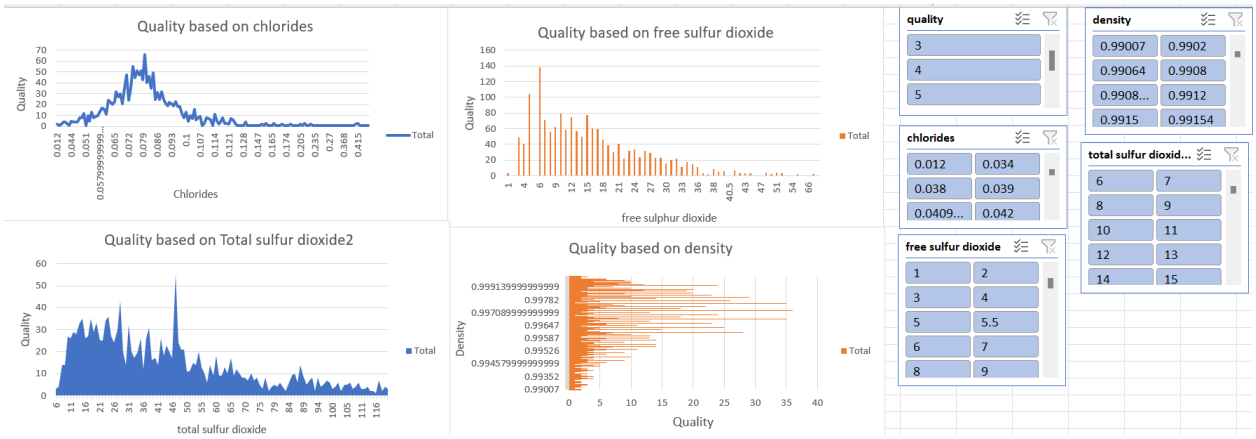


Figure No. 30

- Four charts are presented in Dashboard 2 in Figure No.30 as follows:-

➢ Quality count based on chlorides

➢ Quality count based on free sulfur dioxide

➢ Quality count based on total sulfur dioxide2

➢ Quality count based on density

- Quality count based on pH

Ph	Count of quality
2.74	1
2.86	1
2.87	1
2.88	2
2.89	4
2.9	1
2.92	4
2.93	3
2.94	4
2.95	1
2.98	5
2.99	2
3	6

Figure No. 31

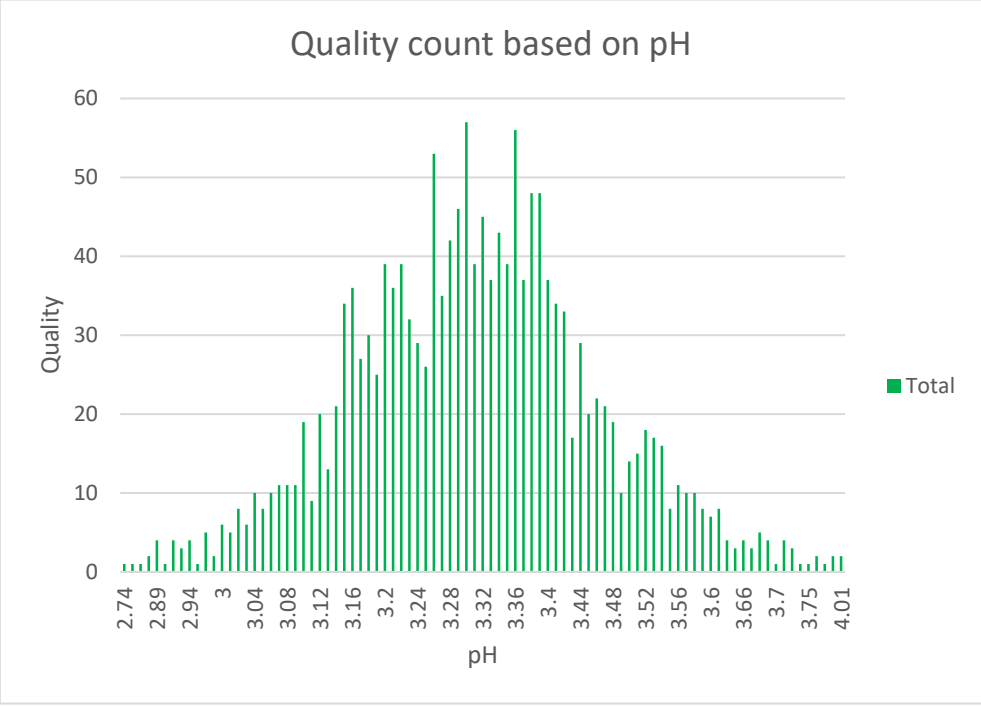


Figure No. 32

- It is observed in Figure No. 31 and 32 that the pH level of 3.3 units has the highest quality count.

- Quality count based on sulphates

sulphates	Count of quality
0.33	1
0.37	2
0.39	6
0.4	4
0.42	5
0.43	8
0.44	16
0.45	12
0.46	18
0.47	19
0.48	29
0.49	31
0.5	27
0.51	26
0.52	47
0.53	51
0.54	68
0.55	50
0.56	60
0.57	55
0.58	68
0.59	51
0.6	69

Figure No. 33

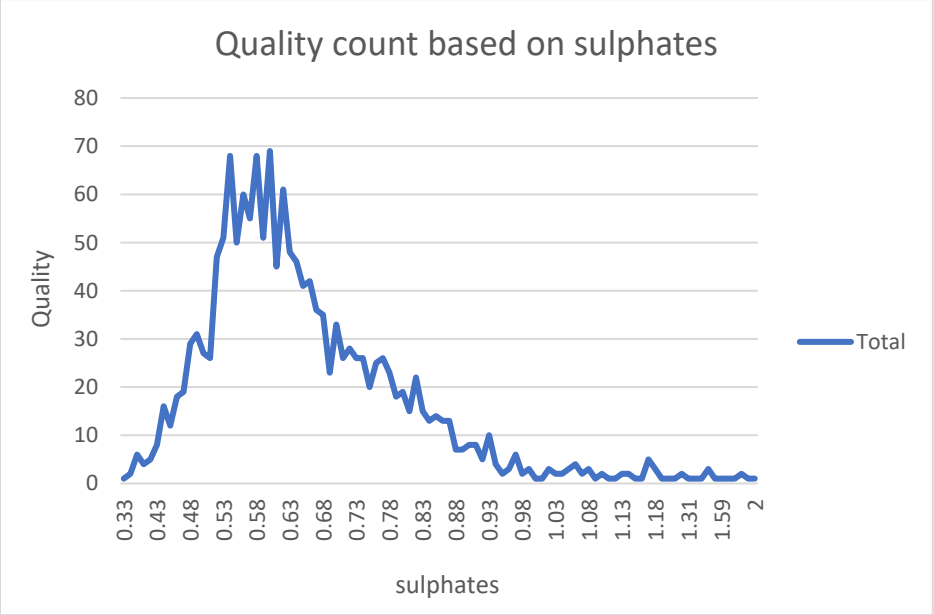


Figure No. 34

- It is observed in Figure No. 33 and 34 that sulphates at the level of 0.6 g/dm^3 have the highest quality count.

- Quality count of wine based on alcohol

alcohol	Count of quality
8.4	2
8.5	1
8.7	2
8.8	2
9	30
9.05	1
9.1	23
9.2	72
9.233333333	1
9.25	1
9.3	59
9.4	103
9.5	139
9.55	2

Figure No. 35

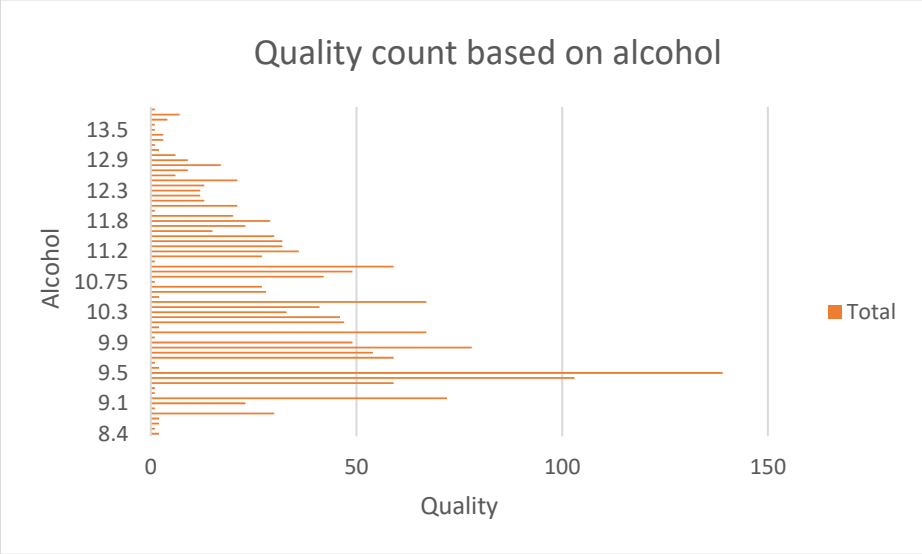


Figure No. 36

- It is observed in Figure No. 35 and 36 that alcohol at the level of 9.5% has the highest quality count.

❖ Dashboard No.3

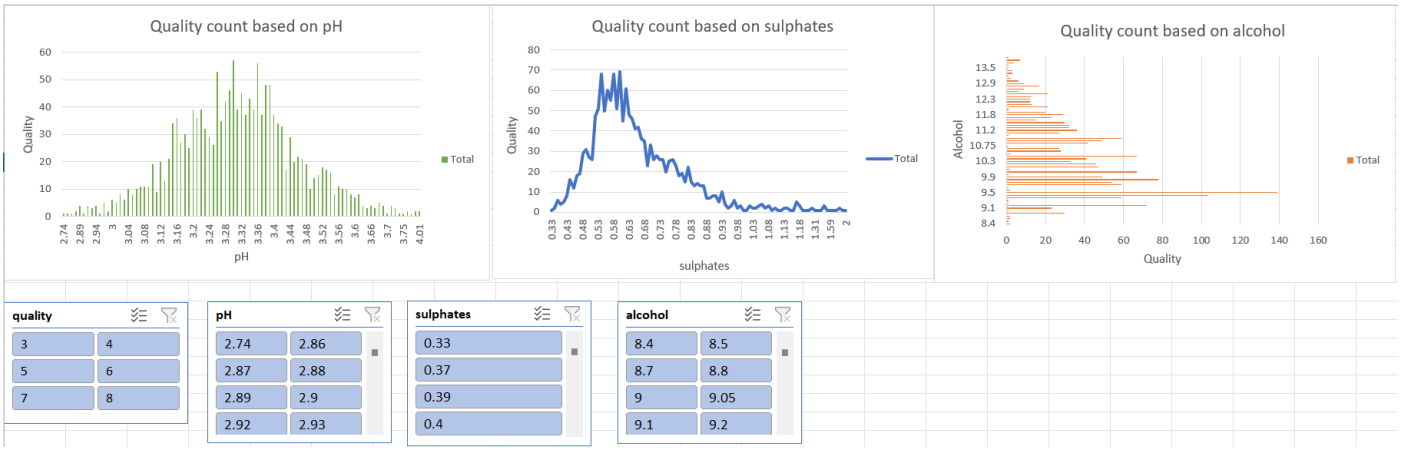


Figure No. 37

- Four charts are presented in Dashboard 3 in Figure No.37 as follows:-
  - ➢ Quality count based on pH
  - ➢ Quality count based on sulphates
  - ➢ Quality count based on alcohol

- Regression Analysis

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.59643236
R Square	0.35573156
Adjusted R Square	0.35126594
Standard Error	0.65044897
Observations	1599

Figure No. 38

- It is observed in Figure No. 38 that, R square is 35.57%, meaning variability of the explanatory variable explains about 35.57% variability of the explained variable.

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	21.8853145	21.35750367	1.024713	0.305655	-20.00657299	63.7772	-20.00657299	63.777202
fixed acidity	0.03773893	0.02582622	1.461264	0.144141	-0.012918162	0.088396	-0.012918162	0.088396
volatile acidity	-1.1556662	0.119750628	-9.65061	1.87E-21	-1.390552304	-0.92078	-1.390552304	-0.9207802
citric acid	-0.2844662	0.144743134	-1.96532	0.049552	-0.568374049	-0.00056	-0.568374049	-0.0005583
residual sugar	0.01238221	0.015058013	0.822301	0.411029	-0.017153475	0.041918	-0.017153475	0.0419179
chlorides	-1.6803987	0.41742422	-4.02564	5.95E-05	-2.499159529	-0.86164	-2.499159529	-0.8616378
free sulfur dioxide	0.00136827	0.002036283	0.671946	0.501716	-0.002625815	0.005362	-0.002625815	0.0053624
total sulfur dioxide	-0.0023042	0.000812351	-2.83646	0.00462	-0.003897591	-0.00071	-0.003897591	-0.0007108
density	-18.251027	21.7979966	-0.83728	0.402561	-61.00692402	24.50487	-61.00692402	24.504869
pH	-0.3103942	0.189978076	-1.63384	0.10249	-0.683028566	0.06224	-0.683028566	0.0622402
sulphates	0.89368793	0.114722524	7.789995	1.2E-14	0.668664294	1.118712	0.668664294	1.1187116
alcohol	0.28285511	0.026515164	10.66767	1.05E-25	0.23084668	0.334864	0.23084668	0.3348635

Figure No. 39

- It is observed in Figure No.39 that, multiple regression is run for the target variable quality, with input variables namely, fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide 2, density, pH, sulphates, and alcohol.

- quality = α + β1*fixed acidity + β2* volatile acidity + β3*citric acid + β4*residual sugar + β5*chlorides + β6* free sulfur dioxide + β7* total sulfur dioxide + β8* density + β9* pH+ β10 * sulphates + β11*alcohol + ξ

  Where

- 21.88 for Intercept α
- 0.03 for slope β1
- -1.15 for slope β2
- -0.28 for slope β3
- 0.01 for slope β4
- -1.68 for slope β5
- 0.001 for slope β6
- -0.002 for slope β7
- -18.25 for slope β8
- -0.31 for slope β9
- 0.89 for slope β10
- 0.28 for slope β11

- In the multiple regression model, where the input variable is fixed acidity, the estimate of the slope coefficient is 0.03, which means:-

  With an increase in every unit (g/dm^3) of fixed acidity, there will be an increase in the quality of red wine on an average of 0.03 units.
- In the multiple regression model, where the input variable is volatile acidity, the estimate of the slope coefficient is -1.15, which means:

  With an increase in every unit (g/dm^3) of volatile acidity, there will be a decrease in the quality of red wine on an average of 1.15 units.
- In the multiple regression model, where the input variable is citric acid, the estimate of the slope coefficient is -0.28, which means:

  With an increase in every unit(g/dm^3) of citric acid, there will be a decrease in the quality of red wine on an average of 0.28 units.
- In the multiple regression model, where the input variable is residual sugar, the estimate of the slope coefficient is 0.01, which means:

  With an increase in every unit(g/dm^3) of residual sugar, there will be an increase in the quality of red wine on an average of 0.01 units.
- In the multiple regression model, where the input variable is chlorides, the estimate of the slope coefficient is -1.68, which means:

  With an increase in every unit (g /dm^3) of chlorides, there will be a decrease in the quality of red wine on an average of 1.68 units.
- In the multiple regression model, where the input variable is free sulfur dioxide, the estimate of the slope coefficient is 0.001, which means:

  With an increase in every unit(mg/dm^3) of free sulfur dioxide, there will be an increase in the quality of red wine on an average of 0.001 units.
- In the multiple regression model, where the input variable is total sulfur dioxide 2, the estimate of the slope coefficient is -0.002, which means:

  With an increase in every unit(mg/dm^3) of total sulfur dioxide 2, there will be a decrease in the quality of red wine on an average of 0.002 units.

- In the multiple regression model, where the input variable is density, the estimate of the slope coefficient is -18.25, which means:

  With an increase in every unit(g/cm^3) of density, there will be a decrease in the quality of red wine by an average of 18.25 units.
- In the multiple regression model, where the input variable is pH, the estimate of the slope coefficient is -0.31, which means:

  With an increase in the level of pH, there will be a decrease in the quality of red wine at an average of 0.31 units.
- In the multiple regression model, where the input variable is sulphates, the estimate of the slope coefficient is 0.89, which means:

  With an increase in every unit (g/dm^3) of sulphates, there will be an increase in the quality of red wine on an average of 0.89 units.
- In the multiple regression model, where the input variable is alcohol, the estimate of the slope coefficient is 0.28, which means:

  With an increase in every percentage of alcohol, there will be an increase in the quality of red wine on an average of 0.28 units.

- Significance of parameter estimates and P-value.

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	21.8853145	21.35750367	1.024713	0.305655	-20.00657299	63.7772	-20.00657299	63.777202
fixed acidity	0.03773893	0.02582622	1.461264	0.144141	-0.012918162	0.088396	-0.012918162	0.088396
volatile acidity	-1.1556662	0.119750628	-9.65061	1.87E-21	-1.390552304	-0.92078	-1.390552304	-0.9207802
citric acid	-0.2844662	0.144743134	-1.96532	0.049552	-0.568374049	-0.00056	-0.568374049	-0.0005583
residual sugar	0.01238221	0.015058013	0.822301	0.411029	-0.017153475	0.041918	-0.017153475	0.0419179
chlorides	-1.6803987	0.41742422	-4.02564	5.95E-05	-2.499159529	-0.86164	-2.499159529	-0.8616378
free sulfur dioxide	0.00136827	0.002036283	0.671946	0.501716	-0.002625815	0.005362	-0.002625815	0.0053624
total sulfur dioxide	-0.0023042	0.000812351	-2.83646	0.00462	-0.003897591	-0.00071	-0.003897591	-0.0007108
density	-18.251027	21.7979966	-0.83728	0.402561	-61.00692402	24.50487	-61.00692402	24.504869
pH	-0.3103942	0.189978076	-1.63384	0.10249	-0.683028566	0.06224	-0.683028566	0.0622402
sulphates	0.89368793	0.114722524	7.789995	1.2E-14	0.668664294	1.118712	0.668664294	1.1187116
alcohol	0.28285511	0.026515164	10.66767	1.05E-25	0.23084668	0.334864	0.23084668	0.3348635

Figure No.40

- It is observed in Figure 40, that the P-value of parameter estimates of variables volatile acidity, citric acid, chlorides, total sulfur dioxide 2, sulphates, and alcohol are statistically significant since the P-value of the respective variables is less than the threshold value of 5% (0.05). This means for these variables we reject the null hypothesis which states that the parameter estimate associated with the variable equals zero, in favor of the alternative one which states that the parameter estimate associated with the variable does not equal zero.

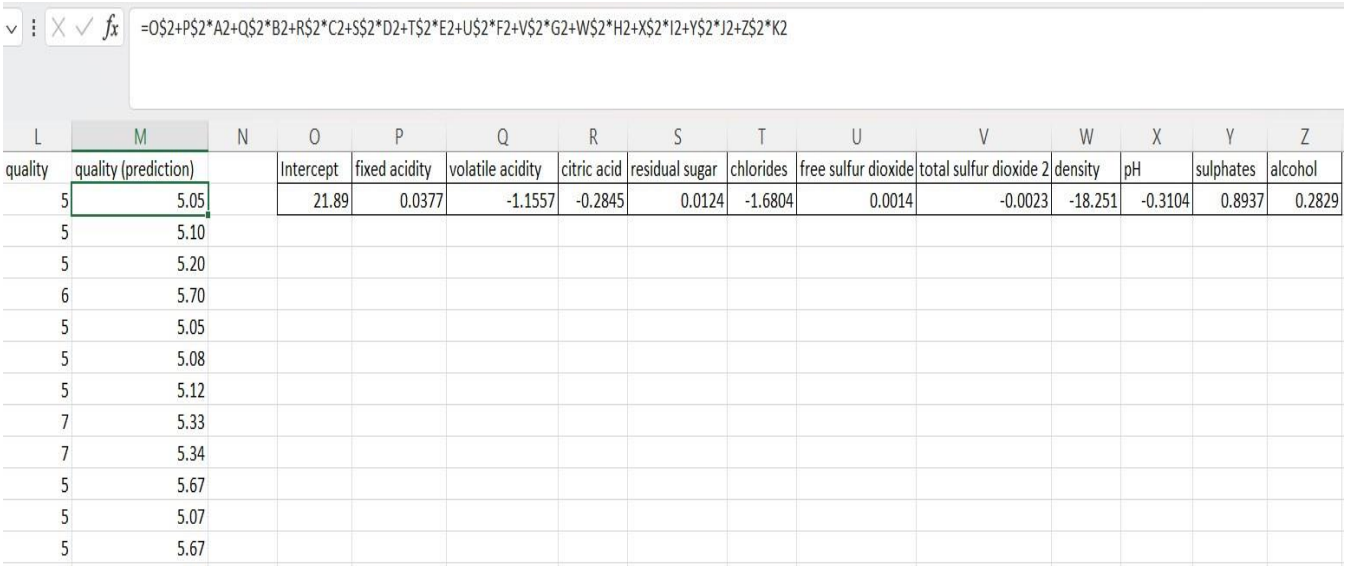
- It is further observed that the P-value of parameter estimates of variables fixed acidity, residual sugar, free sulfur dioxide, density, and pH are not statistically significant since the P-value of the respective variables is greater than the threshold value of 5% (0.05). Therefore, the null hypothesis is accepted for these variables.

- Prediction of quality of red wine

Intercept	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide 2	density	pH	sulphates	alcohol
21.89	0.0377	-1.1557	-0.2845	0.0124	-1.6804	0.0014	-0.0023	-18.251	-0.3104	0.8937	0.2829

Figure No. 41

- It is observed in Figure No. 41 that, the intercept and parameter estimate of variables in the multiple regression model is used to predict the quality of red wine.



L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
quality	quality (prediction)		Intercept	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide 2	density	pH	sulphates	alcohol
5	5.05		21.89	0.0377	-1.1557	-0.2845	0.0124	-1.6804	0.0014	-0.0023	-18.251	-0.3104	0.8937	0.2829
5	5.10													
5	5.20													
6	5.70													
5	5.05													
5	5.08													
5	5.12													
7	5.33													
7	5.34													
5	5.67													
5	5.07													
5	5.67													

Figure No. 42

- quality prediction = α + β1*fixed acidity + β2* volatile acidity + β3*citric acid + β4*residual sugar + β5*chlorides + β6* free sulfur dioxide + β7* total sulfur dioxide 2 + β8* density + β9* pH+ β10*sulphates + β11*alcohol + ξ
- The formula shown above is used in Figure No. 42 to predict the quality of red wine.

- Descriptive statistics of quality and quality(prediction)

quality		*quality (prediction)*	
Mean	5.636023	Mean	5.636023
Standard Error	0.020196	Standard Error	0.012045
Median	6	Median	5.583168
Mode	5	Mode	6.092971
Standard Deviation	0.807569	Standard Deviation	0.481661
Sample Variance	0.652168	Sample Variance	0.231997
Kurtosis	0.296708	Kurtosis	-0.4366
Skewness	0.217802	Skewness	0.394644
Range	5	Range	3.237197
Minimum	3	Minimum	4.278316
Maximum	8	Maximum	7.515512
Sum	9012	Sum	9012

Figure No. 43

It is observed in Figure No. 43 that :-

- The average of both quality and quality prediction is the same i.e 5.63.
- Median of quality is 6 whereas the median of quality prediction is 5.58.
- Mode of quality is 5 whereas the mode of quality prediction is 6.09.
- The standard deviation of quality is 0.80 whereas the standard deviation of quality prediction is 0.48, it is observed that data points of the variable quality are more dispersed than quality prediction.
- Kurtosis of quality is 0.29 meaning the distribution is positive, since kurtosis is less than 3 it means the distribution corresponds to a flatter peak and a shorter tail, whereas kurtosis of quality prediction is -0.43 meaning the distribution is negative, since kurtosis is less than 3 it means the distribution corresponds to a flatter peak and a shorter tail.
- The skewness of quality is 0.21 and the skewness of quality prediction is 0.39, meaning the distribution is positive and symmetric for both the variables.

- The range of quality is 5 whereas the range of quality prediction is 3.23.
- The minimum unit of quality is 3, whereas the minimum unit of quality prediction is 4.28.
- Maximum unit of quality is 8, whereas maximum unit of quality prediction is 7.52.

- Correlation between quality and quality(prediction)

	quality	*quality (prediction)*
quality	1	
quality (pr	0.596432363	1

Figure No. 44

- It is observed in Figure No. 44 that the correlation between quality and quality (prediction) is about 60% which is a positive moderate correlation.

○ Trends, Patterns and Anomalies

❖ Correlation

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide 2	density	pH	sulphates	alcohol	quality
fixed acidi	1											
volatile ac	-0.25613089	1										
citric acid	0.67170343	-0.552495685	1									
residual su	0.11477672	0.001917882	0.143577162	1								
chlorides	0.09370519	0.061297772	0.203822914	0.055609535	1							
free sulfur	-0.15379419	-0.010503827	-0.060978129	0.187048995	0.005562	1						
total sulfu	-0.12086928	0.07726653	-0.005526831	0.149633156	0.058613	0.612954232	1					
density	0.66804729	0.022026232	0.364947175	0.355283371	0.200632	-0.021945831	0.085281739	1				
pH	-0.68297819	0.234937294	-0.541904145	-0.085652422	-0.26503	0.070377499	-0.00527804	-0.3417	1			
sulphates	0.18300566	-0.260986685	0.312770044	0.005527121	0.37126	0.051657572	0.018805148	0.148506	-0.19665	1		
alcohol	-0.06166827	-0.202288027	0.109903247	0.042075437	-0.22114	-0.069408354	-0.206752449	-0.49618	0.205633	0.093595	1	
quality	0.12405165	-0.39055778	0.226372514	0.013731637	-0.12891	-0.050656057	-0.174188536	-0.17492	-0.05773	0.251397	0.476166	1

Figure No. 45

- The correlation between the variable fixed acidity and variable quality is about 12%.
- The correlation between the variable volatile acidity and variable quality is about -39%.
- The correlation between the variable citric acid and variable quality is about 22%.
- The correlation between the variable residual sugar and variable quality is about 1.3%.
- The correlation between the variable chlorides and variable quality is about -12%.
- The correlation between the variable free sulfur dioxide and variable quality is about -5%.
- The correlation between the variable total sulfur dioxide 2 and variable quality is about -17%.
- The correlation between the variable density and variable quality is about -17%.
- The correlation between the variable pH and variable quality is about -5%.
- The correlation between the variable sulphates and variable quality is about 25%.
- The correlation between the variable alcohol and variable quality is about 47%.
- The variable quality has a positive and moderate correlation with the variables citric acid, sulphates, and alcohol.
- The variable quality has a positive and weak correlation with the variables fixed acidity and residual sugar.
- The variable quality has a negative and moderate correlation with the variables volatile acidity, total sulfur dioxide 2, and density.

- The variable quality has a negative and weak correlation with the variables chlorides, free sulfur dioxide, and pH.

❖ Skewness of the variable quality

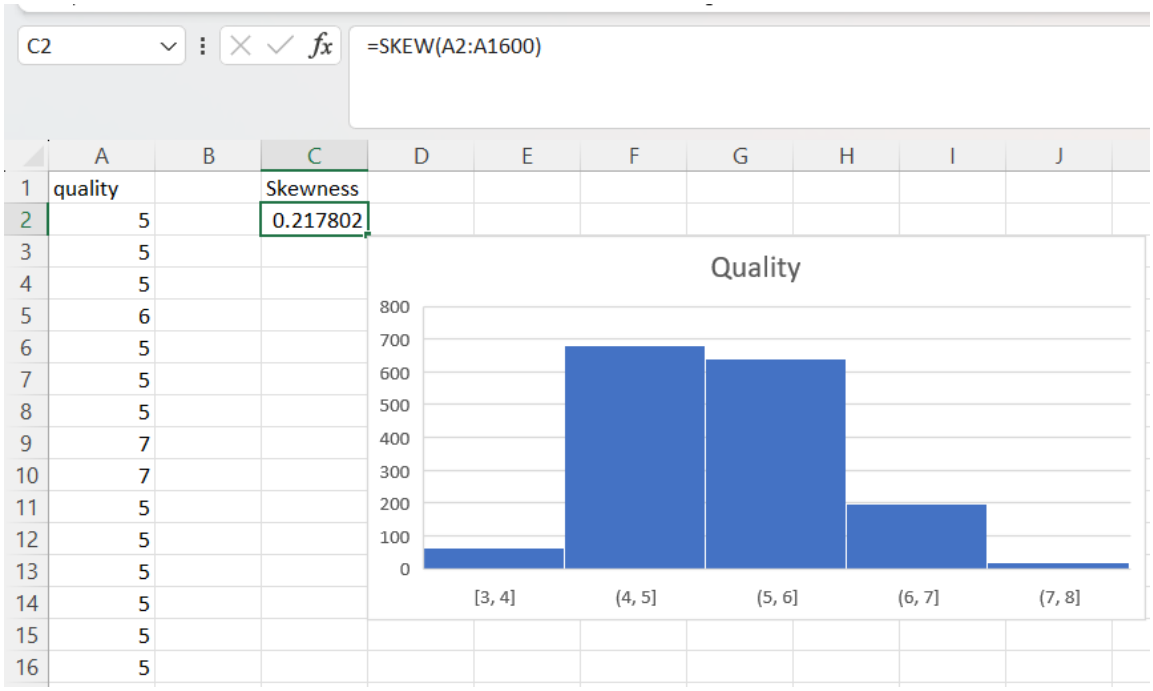


Figure No. 46

- It is observed in Figure No. 46 that the skewness of the variable quality is 0.21, meaning the distribution is positive and symmetric.

❖ Trends and Patterns

- As seen earlier the variable quality has a positive correlation with the variables namely fixed acidity, residual sugar, citric acid, sulphates, and alcohol.

- Fixed acidity and quality

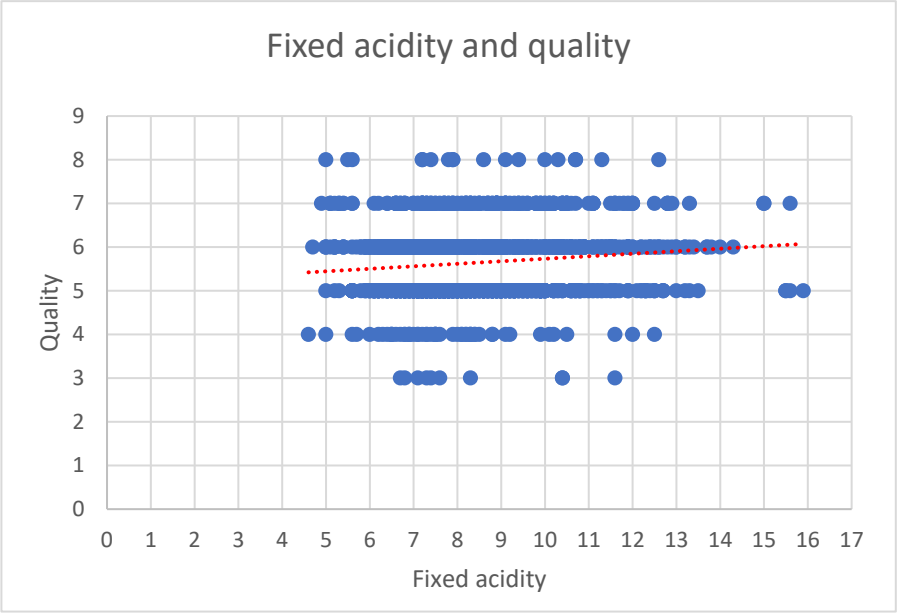


Figure No. 47

- It is observed in Figure No. 47, that there is a gradual increase in the quality of red wine as the fixed acidity level rises.
- It is further observed that the trendline presented in Figure No. 47 is slightly upward.

- Citric acid and quality

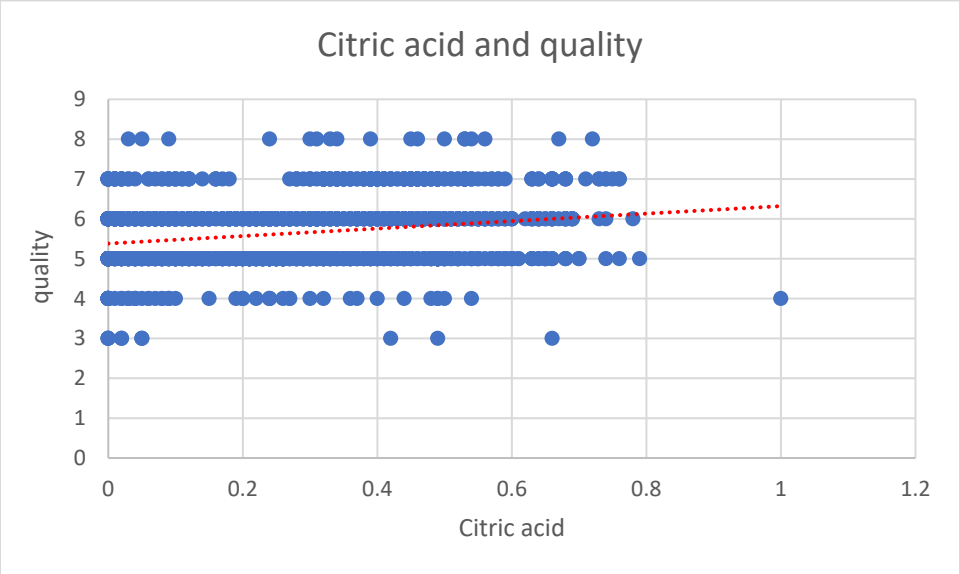


Figure No. 48

- It is observed in Figure No. 48 that there is a gradual increase in the level of quality of red wine when the level of citric acid increases.
- It is further observed that the upward trendline is presented in Figure No. 48

- Residual sugar and quality

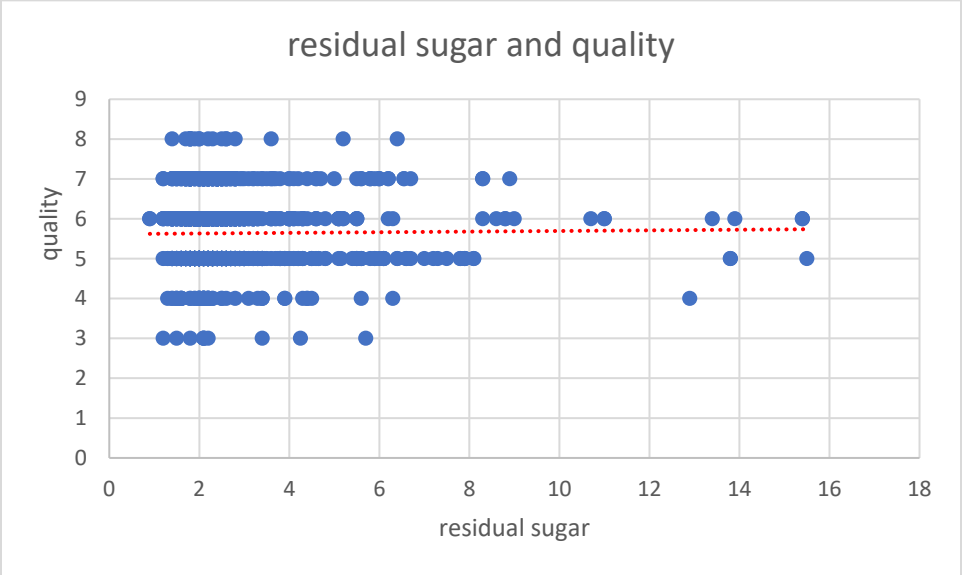


Figure No. 49

- It is observed that overall, there is no trend. The flat trendline in Figure No. 49 confirms the same.

- Sulphates and quality

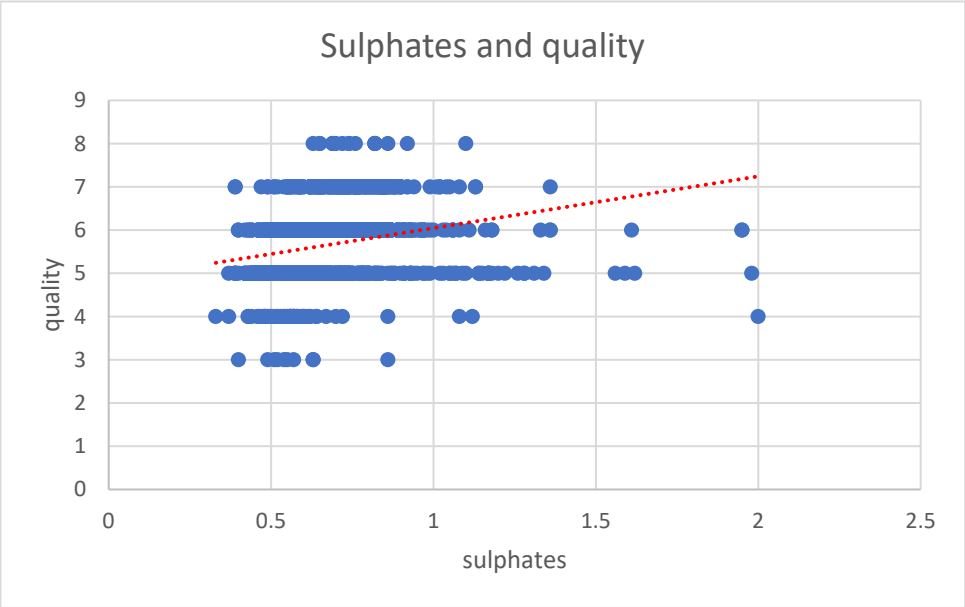


Figure No. 50

- It is observed in Figure No. 50 that there is an overall increase in the level of quality of red wine as the level of sulphates increases.
- An upward trendline is presented in Figure No. 50.

- Alcohol and quality

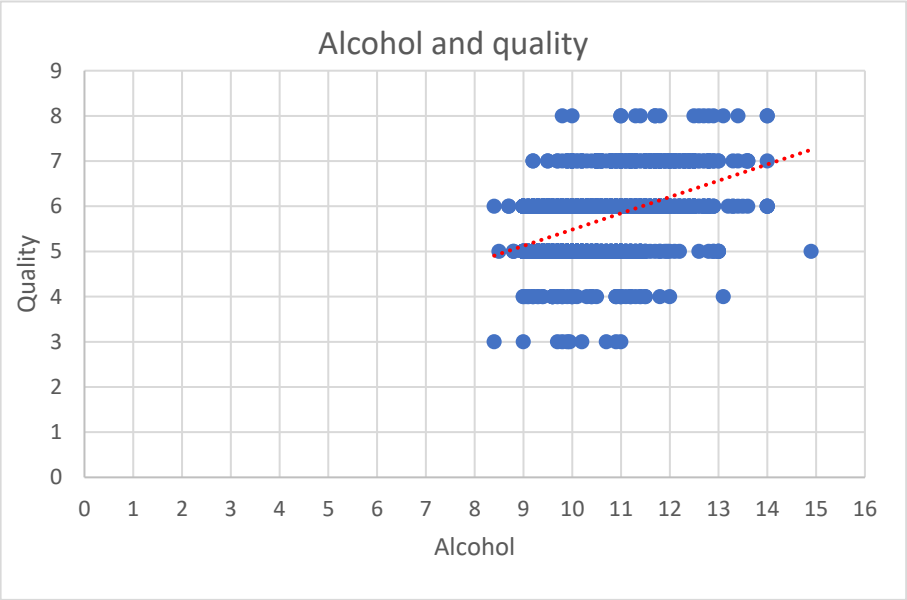


Figure No. 51

- It is observed in Figure No. 51, that there is an overall increase in the level of quality of red wine after the percentage of alcohol increases.
- A sharp upward trendline is observed in Figure No. 51.

❖ Anomalies

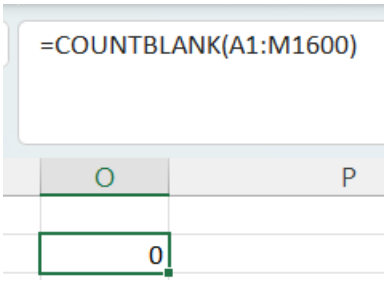


Figure No.52

- As seen in Figure No.52 there are no missing values in the dataset.

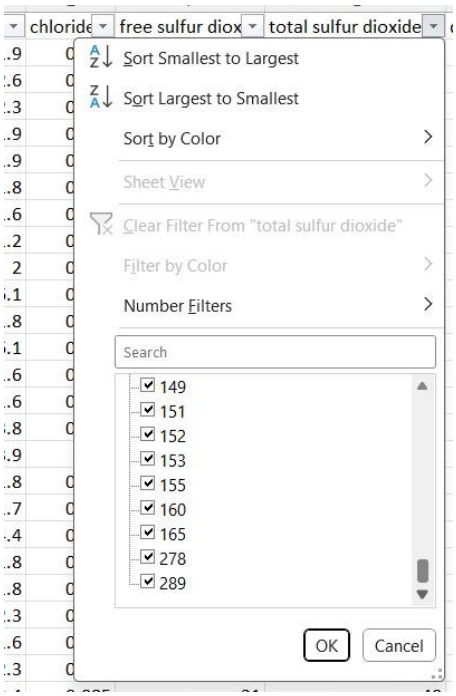


Figure No.53

- It is observed in Figure No.53, that two values i.e. 278 and 289 in the variable total sulfur dioxide may be outliers. Therefore, we need to find out the total number of outliers.

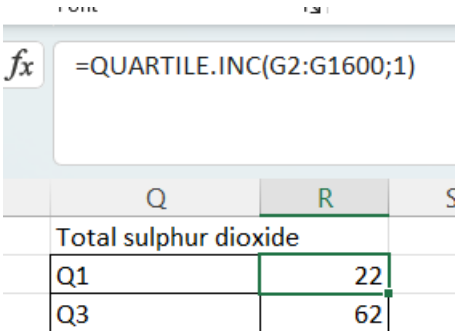


Figure No.54 Calculation of Quartile

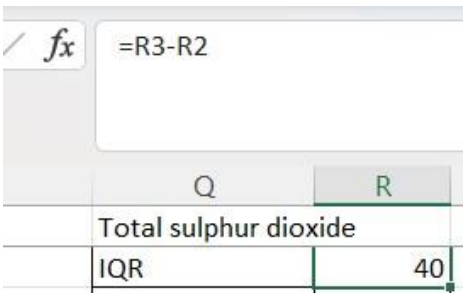


Figure No.55 Calculation of IQR

- In Figure No.54, Quartile 1 and 3 for the variable total sulfur dioxide is calculated.
- In Figure No. 55, Inter Quartile Range is calculated for the variable total sulfur dioxide.

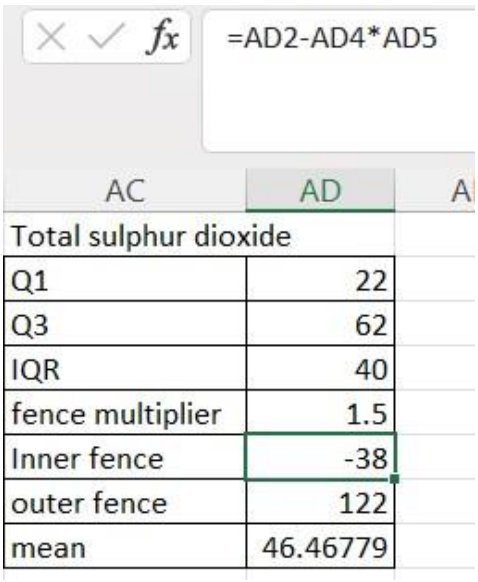


Figure No. 56 Calculation of Inner fence

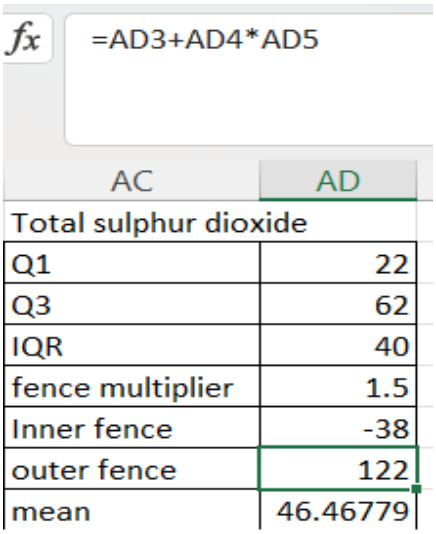


Figure No.57 Calculation of Outer fence

- In Figure No.56 and 57, Inner fence and Outer fence values for the variable total sulfur dioxide are calculated.

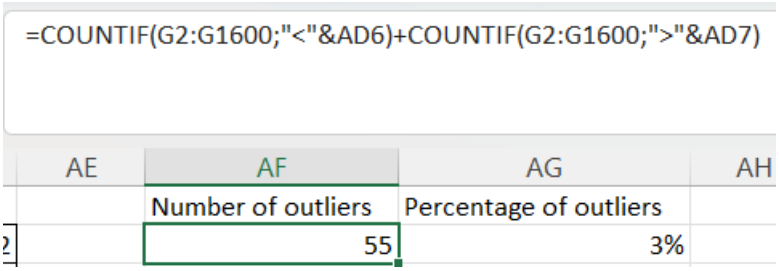


Figure No. 58

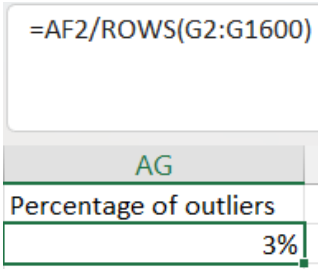


Figure No. 59

- It is observed in Figure No.58 that the number of outliers is 55.
- It is observed in Figure No.59 that the percentage of outliers is 3%.

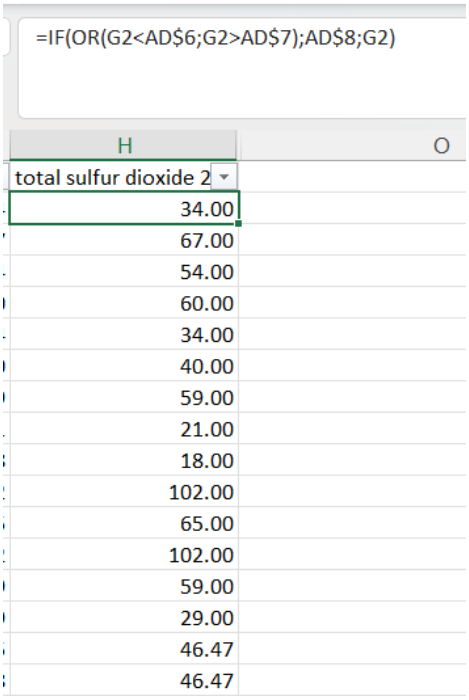


Figure No.60

- It is observed in Figure No.60, that imputation with the mean value approach is adopted for handling the outliers. A new column named total sulfur dioxide 2 is created to store the modified values that overwrite the potential outlying observations with a mean value of the variable.

○ Discussion

❖ Initial Assumptions

- Data values have a normal distribution.

➢ It is observed that all data values do not have a normal distribution.

➢ The variables like citric acid, density, and pH have normal distribution since their skewness is closer to zero.

➢ The variables like fixed acidity, volatile acidity, total sulfur dioxide 2, and alcohol have asymmetric distribution since their data values are moderately skewed.

➢ The variables like residual sugar, chlorides, free sulfur dioxide, and sulphates have asymmetric distribution since their data values are heavily skewed.

- Data values have a linear relationship.

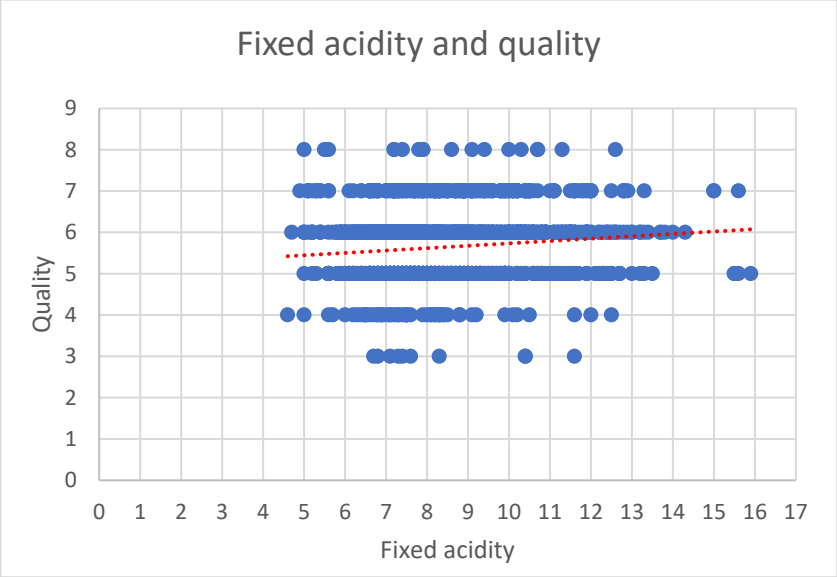


Figure No. 61

- A line of best fit is used in a scatter plot to accurately check the linear relationship between two variables.
- The greater the dispersion of data points around the line of best fit, the weaker the correlation between the variables.
- It is observed in Figure No. 61, that the trendline is slightly upward therefore there is a positive linear relation between the variables fixed acidity and quality.

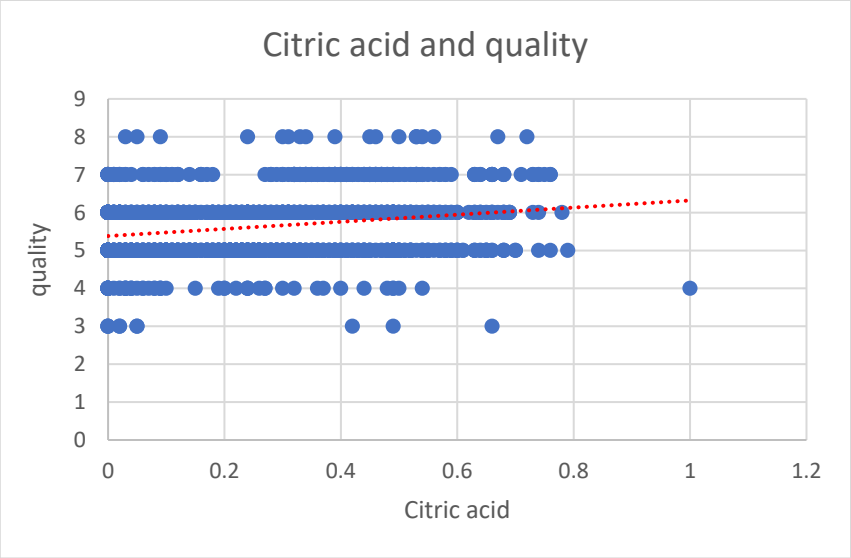


Figure No. 62

- It is observed in Figure No. 62, that there is a lot of dispersion of data points around the line of best fit.
- It is further observed that the line of best fit isn't completely horizontal and is slightly upwards. Therefore, a positive linear relationship exists between the variables citric acid and quality.

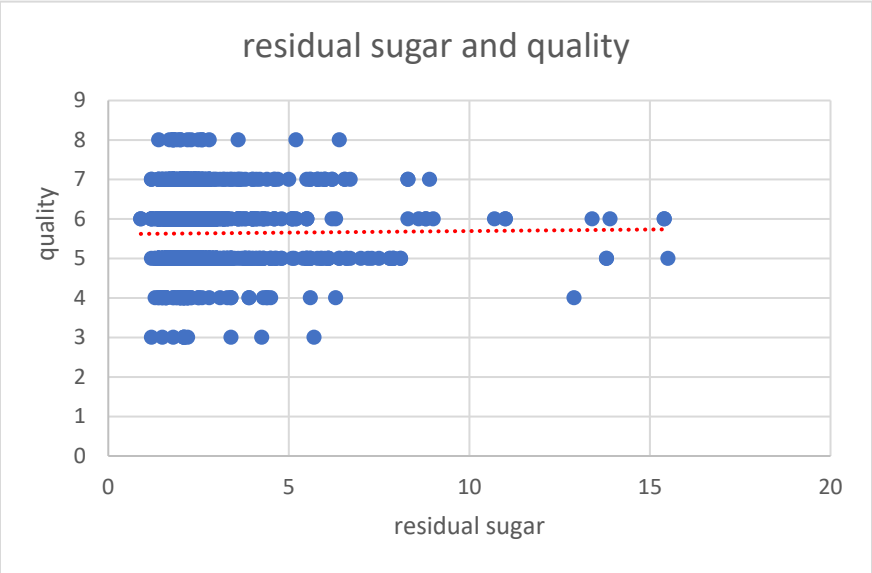


Figure No. 63

- It is observed in Figure No. 63, that the line best fit is horizontal, indicating that there is no linear relationship between the two variables residual sugar and quality.

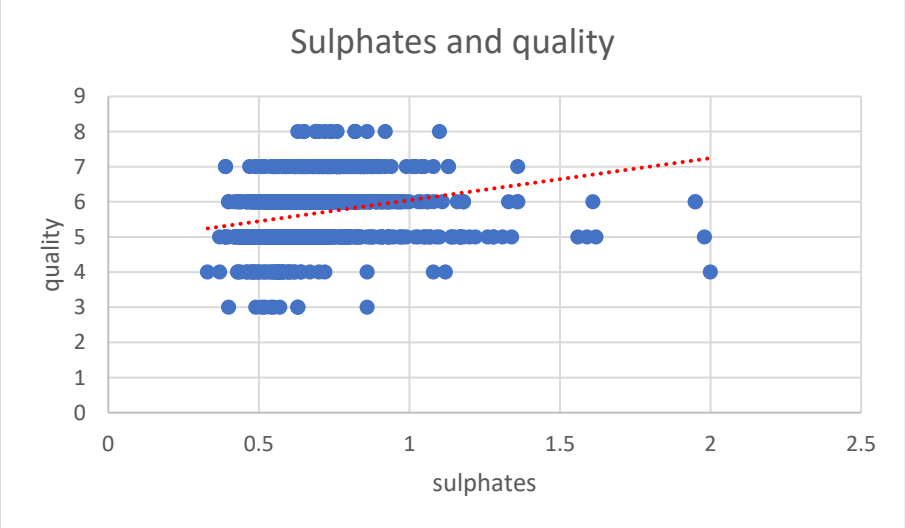


Figure No. 64

- It is observed in Figure No. 64 that there is a positive linear relationship between variables sulphates and quality.

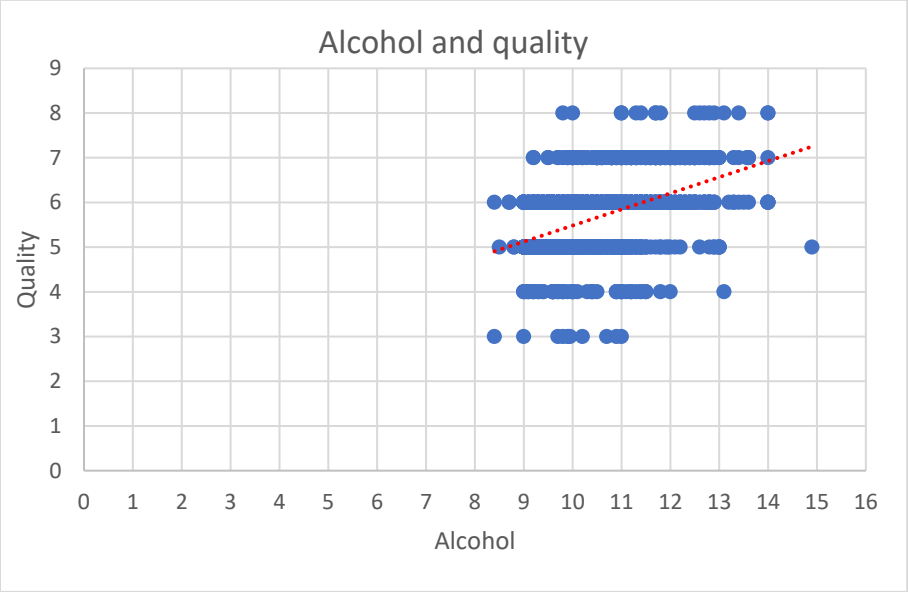


Figure No. 65

- It is observed in Figure No. 65, that there is an upward trendline in the scatter plot. Therefore, the two variables alcohol and quality have a positive linear relationship.

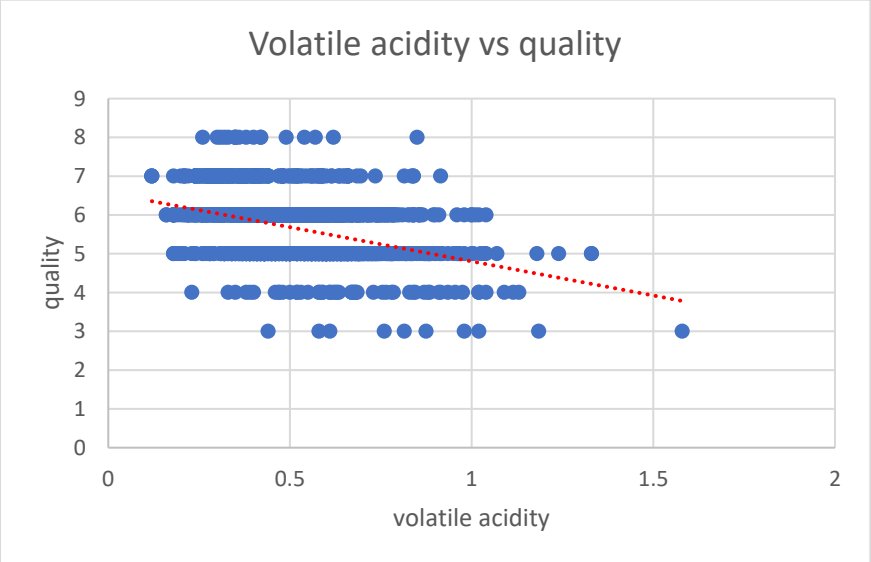


Figure No. 66

- It is observed in Figure No. 66, that there is a downward trendline in the scatter plot. Therefore, the two variables volatile acidity and quality have a negative linear relationship.

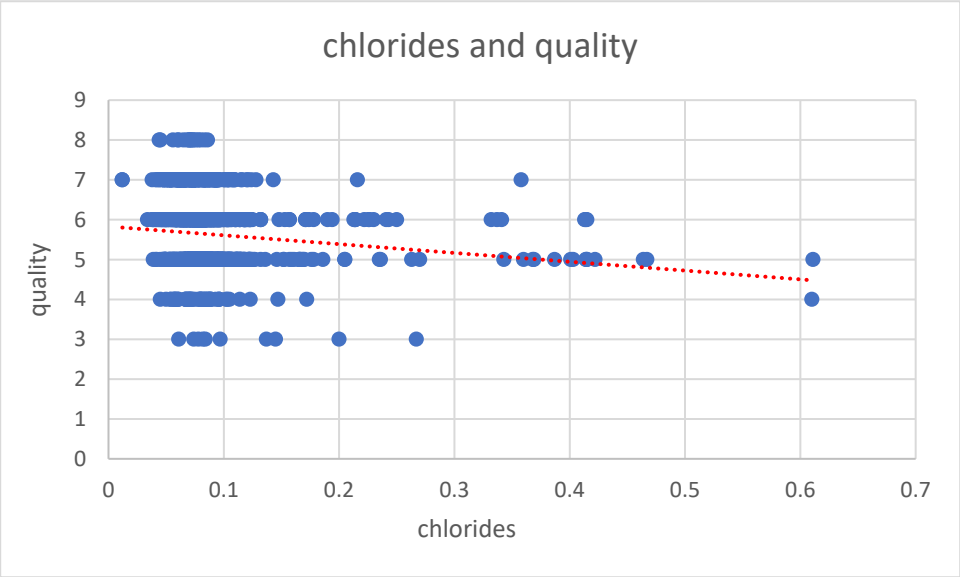


Figure No. 67

- It is observed in Figure No. 67, that there is a downward trendline in the scatter plot. Therefore, the two variables chlorides and quality have a negative linear relationship.

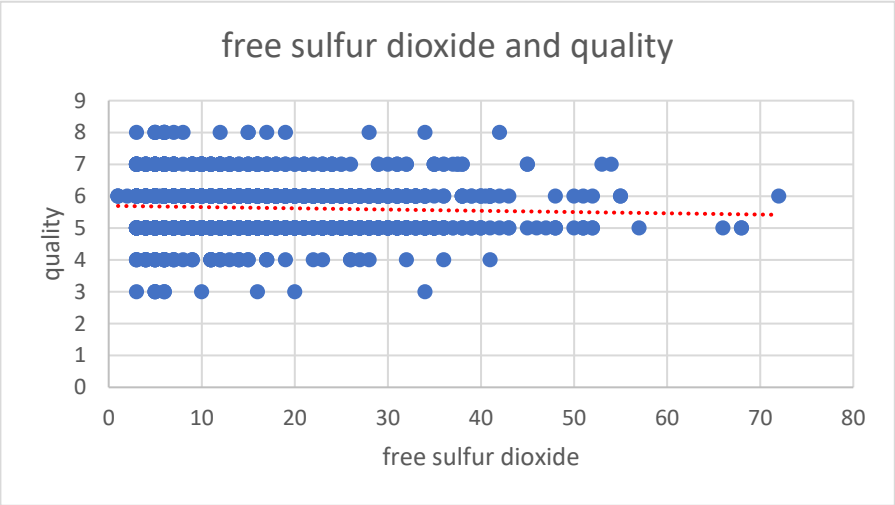


Figure No. 68

- It is observed in Figure No. 68, that the line best fit is horizontal, indicating that there is no linear relationship between the two variables free sulfur dioxide and quality.

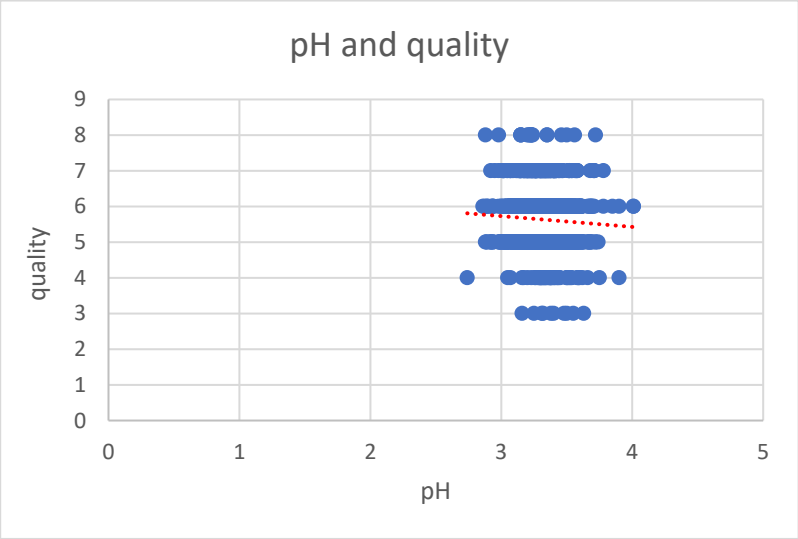


Figure No. 69

- It is observed in Figure No. 69, that the line best fit is slightly downward, indicating that there is a negative linear relationship between the two variables pH and quality.

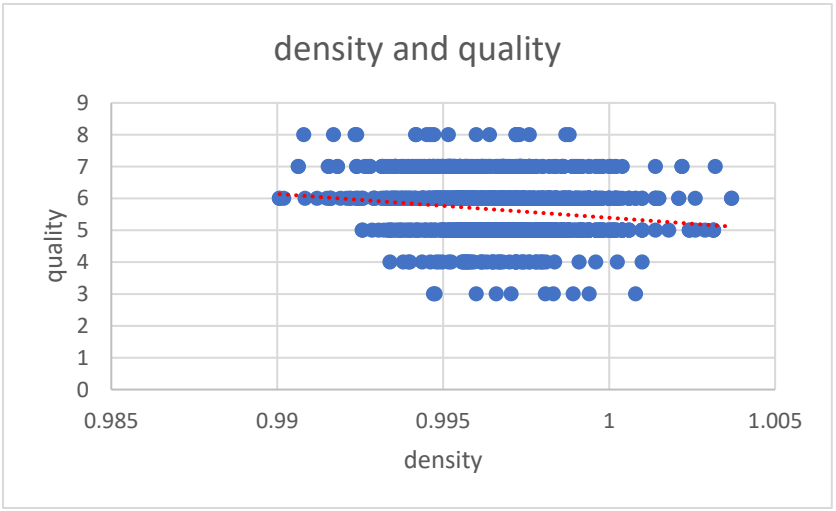


Figure No. 70

- It is observed in Figure No. 70, that the line best fit is slightly downward, indicating that there is a negative linear relationship between the two variables density and quality.

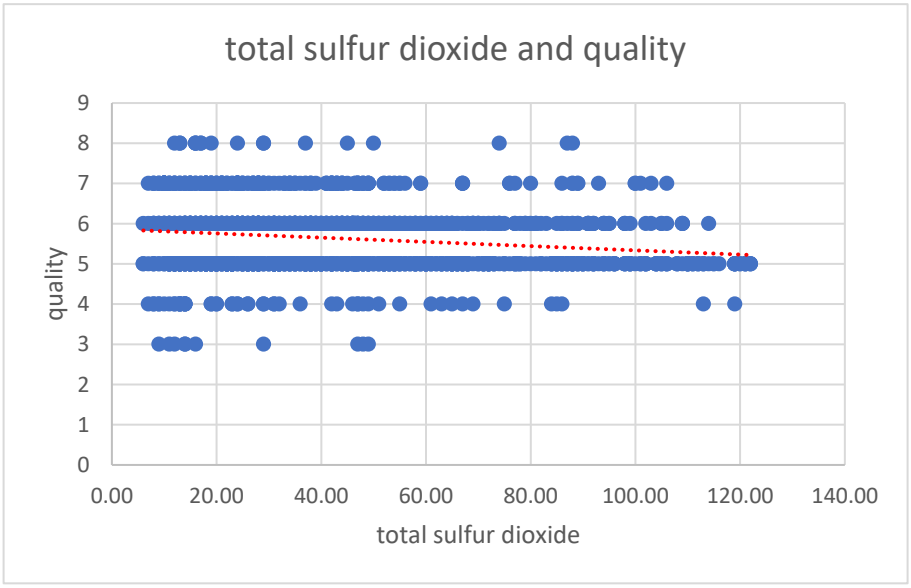


Figure No. 71

- It is observed in Figure No. 71, that the line best fit is slightly downward, indicating that there is a negative linear relationship between the two variables total sulfur dioxide 2 and quality.

❖ Hypotheses

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	21.8853145	21.35750367	1.024713	0.305655	-20.00657299	63.7772	-20.00657299	63.777202
fixed acidity	0.03773893	0.02582622	1.461264	0.144141	-0.012918162	0.088396	-0.012918162	0.088396
volatile acidity	-1.1556662	0.119750628	-9.65061	1.87E-21	-1.390552304	-0.92078	-1.390552304	-0.9207802
citric acid	-0.2844662	0.144743134	-1.96532	0.049552	-0.568374049	-0.00056	-0.568374049	-0.0005583
residual sugar	0.01238221	0.015058013	0.822301	0.411029	-0.017153475	0.041918	-0.017153475	0.0419179
chlorides	-1.6803987	0.41742422	-4.02564	5.95E-05	-2.499159529	-0.86164	-2.499159529	-0.8616378
free sulfur dioxide	0.00136827	0.002036283	0.671946	0.501716	-0.002625815	0.005362	-0.002625815	0.0053624
total sulfur dioxide	-0.0023042	0.000812351	-2.83646	0.00462	-0.003897591	-0.00071	-0.003897591	-0.0007108
density	-18.251027	21.7979966	-0.83728	0.402561	-61.00692402	24.50487	-61.00692402	24.504869
pH	-0.3103942	0.189978076	-1.63384	0.10249	-0.683028566	0.06224	-0.683028566	0.0622402
sulphates	0.89368793	0.114722524	7.789995	1.2E-14	0.668664294	1.118712	0.668664294	1.1187116
alcohol	0.28285511	0.026515164	10.66767	1.05E-25	0.23084668	0.334864	0.23084668	0.3348635

Figure No. 72

- Fixed and volatile acidity would be an influential(positive) factor in the quality of the red wine.

  The hypothesis mentioned above holds in the case of fixed acidity. Fixed acidity positively influences the quality of red wine. The hypothesis does not hold in the case of volatile acidity. Volatile acidity negatively influences the quality of red wine.
- Citric acid would improve the quality of red wine, as citric acid can add freshness and flavor to wines.

  The hypothesis as mentioned above does not hold in the case of citric acid. Citric acid negatively influences the quality of red wine.
- Residual sugar would improve the quality of red wine.

  The hypothesis mentioned above holds in the case of residual sugar. Residual sugar positively influences the quality of red wine.
- The high number of chlorides would not improve the quality of red wine.

  The hypothesis mentioned above holds in the case of chlorides. Chlorides negatively influence the quality of wine.

- Free sulfur dioxide and total sulfur dioxide improve the quality of red wine.

  The hypothesis mentioned above holds in the case of free sulfur dioxide. Free sulfur dioxide positively influences the quality of red wine. In the case of total sulfur dioxide2 the hypothesis does not hold, total sulfur dioxide, negatively influences the quality of red wine.

- The lower the density of wine the better the quality of red wine.

  The hypothesis mentioned above holds in the case of density. Density negatively influences the quality of red wine.

- High pH values would not improve the quality of red wine.

  The hypothesis mentioned above holds in the case of pH. The variable pH negatively influences the quality of red wine.

- Sulphates would improve the quality of red wine.

  The hypothesis mentioned above holds in the case of sulphates. Sulphates positively influence the quality of red wine.

- Alcohol would improve the quality of red wine.

  The hypothesis mentioned above holds in the case of alcohol. Alcohol positively influences the quality of red wine.

○ Conclusion

- After doing EDA (Exploratory Data Analysis), preparing pivot charts, pivot tables, dashboards, running multiple regression, and doing a null hypothesis test on the dataset it is observed that:-
  - Fixed acidity at the level of 7.2g/dm^3 has the highest quality count.
  - Volatile acidity at the level of 0.6 g/dm^3 has the highest quality count.
  - Citric acid at a level of 0 g/dm^3 has the highest quality count
  - Residual sugar at a level of 2 g/dm^3 has the highest quality count.
  - Chlorides at a level of 0.08 g/dm^3 have the highest quality count.
  - Free sulfur dioxide at a level of 6 mg/dm^3 has the highest quality count.
  - Total sulfur dioxide 2 at a level of 46.46 mg/dm^3 has the highest quality count.
  - Density at a level of 0.9972 g/cm^3 has the highest quality count.
  - pH level of 3.3 units has the highest quality count.
  - Sulphates at the level of 0.6 g/dm^3 have the highest quality count.
  - Alcohol at the level of 9.5% has the highest quality count.
  - The output variable quality positively correlates with the variables fixed acidity, citric acid, residual sugar, sulphates, and alcohol.
  - Most of the initial assumptions and hypotheses hold.
  - In the multiple regression model, the variables fixed acidity, residual sugar, free sulfur dioxide, sulphates, and alcohol positively influence the target variable quality of red wine.
- While there is a wide range of techniques to analyze a given data set and predict the results of the final quality. A multiple regression model is used to predict the quality of red wine on the given dataset.
- "Wine quality strongly depends on the grape quality. To obtain high-quality wines, it is necessary to process healthy grapes at the correct ripeness stage, and for this reason, the farmer has to be especially careful in the prevention of parasite attacks on the grapevine." (Pierluigi Caboni, 2010)

- In conducting data analysis on the quality of red wine I aimed to present the data wherever possible in the form of graphs and charts to uncover the trends and patterns of the variables influencing the quality of red wine. Using the imputation with mean value approach I cleaned the data in the variable total sulfur dioxide and removed the outliers. There were no missing values in the dataset. Then I started EDA (Exploratory Data Analysis) which includes descriptive statistics, pivot tables, pivot charts, correlation, and multiple regression analysis. The most interesting part of multiple regression analysis is after performing the analysis, from the intercept and slope coefficient of parameter estimate one gets to know whether the initial hypothesis made at the introductory stage of the analysis holds or not. Also, the preparation of dashboards gives an overview of the data visualization made throughout the project.

Bibliography

Doris Rauhut, F. K., 2019. *Science Direct.* [Online]
Available at: https://www.sciencedirect.com/topics/food-science/wine-quality
[Accessed February 2024].

Pierluigi Caboni, P. C., 2010. *Science Direct.* [Online]
Available at: https://www.sciencedirect.com/topics/food-science/wine-quality
[Accessed 21 February 2024].