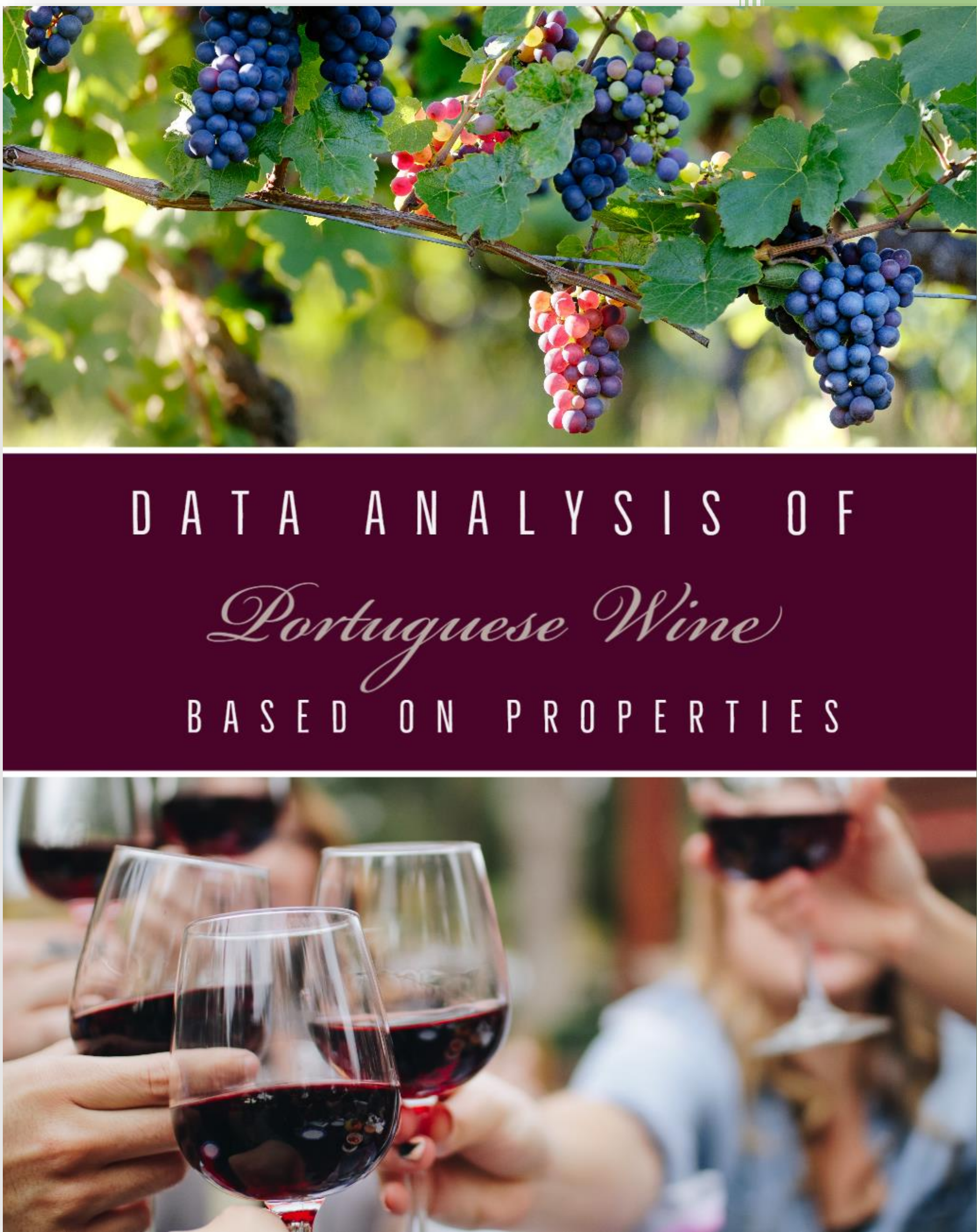


Swati Kohli

Fall 2019

Quality Assessment of Wine
through the physiochemical properties
using Multiple Regression Analysis

TABLE OF CONTENTS

1. Introduction ……………………………………………………………………….pg 02
2. Objective ……………………………………………………………………….pg 02
3. About Wine ……………………………………………………………………….pg 02
4. Wine Making Process ……………………………………………………………….pg 03
5. Wine Properties And Their Impact………………………………………………….pg 03
6. Intuitive Relation …………………………………………………………….....pg 04
7. Analysis Results ……………………………………………………………….….....pg 05
8. Unintuitive Insight & Conclusion ………………………..……………………......pg 06
9. Appendix …………….…………….……………………………………………......pg 06
    a. Dataset
    b. Methodology
    c. Level of Significance
    d. Final Equation
    e. Stepwise Regression
    f. Test of Assumptions through Hypothesis for Normality
    g. Test of Assumptions through Hypothesis for Homoskedasticity
    h. Cross Validation
10. Citation ……….…………………..……………………………………………….....pg 13

1. INTRODUCTION

"A bottle of wine contains more philosophy than all the books in the world"

- Louis Pasteur

As a recent wine enthusiast, I have had some pleasurable experience in tasting wines in the USA and Australia of different origins like Napa, Yara Valley etc. Getting familiar and developing a palatte, I am now more aware of the fact that wine is a luxury beverage appreciated by people all over the world. And this is now increasingly becoming a trend.

PRICING

There are two factors which govern the **price of a wine- quality and vintage**. Nowadays industries invest in new technologies for the growth of manufacturing and selling process. In this context, identifying most influential factors to improve winemaking would help stratify them as premium and other quality. Quality assement for determination and therafter enhancement of the superiority of wine for improving brand quality is a key factor which would increase winemaker's returns. Once it is manufactured, the assessment of the fineness of wine can be done by **two methods**.

1. P**hysiochemical tests** based on chemical characteristics that affect the taste and
2. **Sensory tests** by wine tasting experts called sommeliers.

2. OBJECTIVE

Tasting and enjoying wine is one thing, understand what makes the wine taste good is another. This paper evaluates through a data-driven approach, the effects of chemical attributes towards the quality assessment aspect of wine. These attributes are pH, density, sulphates, alcohol, residual sugar, acidity and chlorides. Do some of these variables have a significant effect on quality? If so, which ones? Can variables be identified for which there is a considerable change between a good wine and a bad one. These variables might be significant predictor of a good wine. Since the study is data-backed, this summary is useful for both manufacturers and sellers of wine to improve their decision process vis a vis product enhancement, revenue generation and marketing strategy. Multiple regression method (Refer Appendix) is used to investigate with a particular level of accuracy, which physiochemical characteristics are related significantly and to what extent they are influential contributors to the valuation of superiority. In other words, the purpose is to present an equation (as best possible) to predict the quality of wine through relevant quantitative and qualitative input variables of physiochemical attributes.

3. ABOUT WINE

Oenology is the science and study of wine and winemaking. The analysis uses Portuguese wine, Vinho Verde (named after the same region) which is considered as a young wine because it is consumed fresh, after harvesting for merely 3-6 months. Broadly speaking, among different variants of wines two are worldwide popular by their names: Red wine and White wine. The 11 physiochemical properties predict the quality based on sensory results given by Sommeliers scored between 0-10.

4. WINE MAKING PROCESS

As sugar levels rise in each grape being used in the manufacturing process, the acid levels drop. Harvesting grapes at just the right balance for sugar and acid is one of the most critical decisions of the winemaker. Also, this decision is often affected by climatic factors like sun, rain etc which are out of one's control. As the harvested grapes go through fermentation, this sugar is principally used up by the yeast to convert into alcohol. Therefore, the higher the sugar level in the grape, the higher the alcohol in the resulting wine. The properties are as follows with a description of how they relate with the taste of the final product.

5. PROPERTIES AND THEIR IMPACT

Wine properties and how they relate with the taste of the final product are described below.

1. **Alcohol content (% volume)**

   Alcohol level in wine is strongly related with the amount of sugar developed in the grapes during the harvest time: higher sugar levels have higher potential of alcohol. However, it does not imply that high alcohol wines are sweeter, though sometimes it could be so. It comes through as heat in the back of the throat. More alcohol will taste warmer and bolder with slight burning sensation. They typically range from 10-14%. Above 14% alcohol content are considered high alcohol wines.

2. **Residual Sugar (g/l)**

   This is the amount of sugar remaining after fermentation stops. Having no perceptible taste of sugar is called dry tasting wine (arounf 10g/l) while wines high in sugar, taste soft (above 35 g/l).

3. **Density (g/cm^3)**

   The density of wine is less to that of water (1 g/cm^3) depending on the percent alcohol and sugar content. The density of ethanol is .789 g/cm^3. Good wines usually have lower density observed at the time they are swirled in the glass before tasting for oxidation.

4. **Acidity and pH**

   Acidity parameters namely citric acid (g/l), fixed acidity (g/l) and volatile acidity (g/l) are used to describe a wine's sour taste. They lend a tart taste or have a sharp edge on the palatte. At optimum levels it gives a zestful tang or freshness and flavor, More acidity can give unpleasantly harsh vinegar taste. Usually red wines are more acidic.

   pH describes how acidic or basic a wine is and most wines are between 2.9 to 4.2 on the pH scale. On a scale from 2.9 (very acidic) to 4.2 (very basic). Lower pH also means the wine retains color better.

5. **Sulfurs**

Sulfates (g/l), free sulfur dioxide (mg/l) and total sulfur dioxide (mg/l) are another set of chemical attributes. SO2 is not present only in free form, therefore, total SO2 is a better factor than other sulfates and its forms to find a good wine. These play the role of are wine additive which act as antimicrobial and antioxidant and usually undetectable in wine. However, in concentrations over 50ppm, SO2 becomes evident in the nose and taste of wine

6. **Chlorides (g/l)**

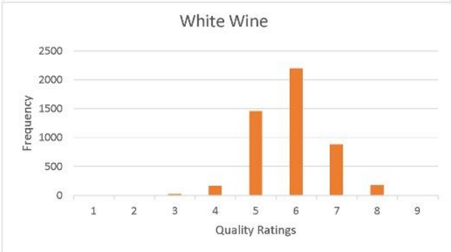
It is the amount of salt in the wine. If exceeding limits, selling of such wines might not be allowed in some countries. The characteristics like flatness or burning persistence are associated with this property. Countries have different threshold values. It is observed that wines from Australia and Argentina have more concentration of chlorides than USA or Europe.

7. **Quality Rating (0-10)**

The assessment associated with wine designation is based on the sensory data. The 11 physiochemical properties predict the quality based on tasting results given by sommeliers who score between 0-10. For the data under consideration, following has been observed. It is observed that maximum data is for average rating 5-7.

rating	*Frequency*
1	0
2	0
3	10
4	53
5	681
6	638
7	199
8	18
9	0
10	0

rating	*Frequency*
1	0
2	0
3	20
4	163
5	1457
6	2198
7	880
8	175
9	5
10	0



6. INTUITIVE RELATIONS

Based on the above properties, following variables have known relations to one another and are expected to reflect in the data

1. **pH and Acidity**

If the data is sound, pH should have an inverse relationship with the acid variables in the data. Either of these factors are expected to have a significance in the quality assessment.

2. **Density and Alcohol**

Alcohol has less density compared to other consumable liquids, specifically water. At equal pressure and temperature, water has 1 g/cc while alcohol has 0.789 gm/cc density. Therefore, more alcohol compared to other liquids, should decrease the overall density of wine. This implies an inverse relationship between alcohol and density. It is expected that alcohol plays a significant role in the quality of wine.

3. **Density and Fixed acidity, Residual Sugar**

Furthermore, generally the acids in wine have a higher density than water like tartaric acid, citric acid and malic acid. Therefore, a positive relationship, i.e., higher the concentration of acidic compounds, wine should become more dense. Same goes for residual sugar and density.

Alcohol < Water < Sugar

In the end, a good bottle of wine should have a density close to that of water. With a density slightly less than water it tends to be a dry wine, and slightly greater tends to be a sweeter wine

7. ANALYSIS RESULTS

Regression analysis is a good method to look at wine tasting variables because it gives an insight on how these multiple factors affect the product. With an aim to determine patterns in the overall designation of wine, the 2 datasets (red and white wine) are combined and added as a categorical variable called type. This would streamline the analysis of relationship between quality and physiochemical properties. The distinct relationship extrapolated for quality of wine are with alcohol, volatile acidity, total sulfur dioxide.

Quality(0-10) = (0.65)*Alcohol % by vol – (2.655)*Volatile Acidity g/l – (0.005)*Sulfur Dioxide mg/l

How the variables are related significantly to the quality metric is explained below.

1. **More Alcohol is desired for a better quality**

   A widely accepted idea is that alcohol reduces the palatte sensitivity and therefore, lower alcohol wines are usually considered balanced which thus pair better with foods. However, it is interesting that when tasting wine, alcohol is well received due to the fact that in wines it tends to draw out more intense flavors. They are called fuller bodied wines. It is the most significant factor to determine the quality of wine. As per the analysis, for a wine sample if volatile acidity and sulfur dioxide which are significant in determining the quality (as per the equation) of wine are kept constant, one percent increase in volume of alcohol would increase the score by 0.65. For example, a quality score of 5 would go to 5.65 if percent by volume of alcohol increases from 9% to 10% . This demonstrates a very substantial raise in the quality.

2. **Result of acid**

   The second important factor for wine quality assessment is **volatile acidity** (acetic acid). Wines are perceived negatively (score reduces) when acid concentration increases because as expected, the acetic acid or vinegar flavour reduces the quality of a wine. This could also be due to the positive relation with pH level because acetic acid is one of the weak acid.

   Delving deeper into this, one expects an inverse relation between pH and acids. Although pH shows to have a negative correlation, i.e. inverse relation with fixed and citric acid as expected, it shows to have a positive one with volatile acidity. This is an unexpected result since acidity means lower pH values on the scale. Further research shows that volatile acidity usually means acetic acid which is a weak acid. Weak acid simultaneously contain their related base in one solution (https://en.wikipedia.org/wiki/Acid_strength#Conjugate_acid/base_pair). This possibly explains the positive correlation.

3. **How clean is tasty?**

   Even though it makes sense that an antimicrobial compound would make the wine cleaner, it is a factor that hampers the quality of wine. This is understandable because this antimicrobial compound, **sulfur dioxide,** has a pungent repelling aroma which in more quantity results in lower quality level.

8. UNINTUITIVE INSIGHT & CONCLUSION

Upon review of some other factors which are not too significant but worth exploring are:

1. Density seems to have a negative effect on taste, while pH has a positive effect.
2. The salts or chlorides also negatively affect the quality.
3. It would have been interesting to see how stronger acidity affects the quality of wine. As a taster, in red wine, more acidity is welcome but undesired in white wine at high levels.

More data and research might give some answers to those relationships. Therefore, data analysis is a productive method to gain insight into factors that are important—even if, like taste preferences, they seem hard to measure.

9. APPENDIX

Following is the information on the dataset:

a. DATASET

The dataset for the study is available on Kaggle at
https://www.kaggle.com/maitree/wine-quality-selection
Red Wine : 1599 Observations; White Wine : 4898 Observation
To limit the scope of work, combine and take a random sample of 50 from the dataset based on average quality rating (4 to 8). Therefore 6 random samples each for range 4 to 8.
Refer data set on pg 12

b. METHODOLOGY

Multiple regression with forward stepwise method is used

c. LEVEL OF SIGNIFICANCE is taken at 5%

d. FINAL EQUATION

Quality(0-10) = (0.65)*Alcohol % by vol – (2.655)*Volatile Acidity g/l – (0.005)*Sulfur Dioxide mg/l

e. STEPWISE REGRESSION

1. Check Correlation matrix with Quality

	Fixed_Acidity	*Volatile_Acidity*	*Citric_Acid*	*Residual_Sugar*	*Chlorides*	*Free_Sulphur*	*Total_Sulfur_Dioxi*	*Density*	*pH*	*Sulphates*	*Alcohol*	*Color_(1=white)*	*Quality*
Fixed_Acidity	1												
Volatile_Acidity	0.07305986	1											
Citric_Acid	0.608272517	-0.3843635	1										
Residual_Sugar	-0.125891886	-0.2125785	0.20405057	1									
Chlorides	0.333858403	0.70971374	-0.110663321	-0.219551289	1								
Free_Sulphur	-0.455611895	-0.4084594	0.036655181	0.444615608	-0.52244372	1							
Total_Sulfur_Dic	-0.487191727	-0.3941495	-0.046524541	0.400942912	-0.55531642	0.76314177	1						
Density	0.703356249	0.35521904	0.328946302	0.274417103	0.578634885	-0.37748427	-0.396813934	1					
pH	-0.465261239	0.31725092	-0.599819449	-0.378417421	0.209114692	-0.10187742	-0.126487176	-0.25224	1				
Sulphates	0.195407805	0.06698744	0.093769473	-0.287093232	0.315897421	-0.19384216	-0.311969761	0.199907	0.09134486	1			
Alcohol	-0.08745087	-0.1829914	0.063230311	-0.339459415	-0.24547195	-0.05102114	-0.128136334	-0.49675	0.23742021	0.129004	1		
Color_(1=white)	-0.46081849	-0.5431342	0.040339215	0.381725663	-0.70045487	0.68334049	0.847399397	-0.53423	-0.2320378	-0.435749	-0.07239333	1	
Quality	-0.006596111	-0.438645	0.1786953	-0.221608132	-0.34670747	0.07568974	-0.131581489	-0.43705	0.04769638	0.138005	0.700320228	0	1

Alcohol has highest correlation with Quality, check if transformation required & Run Quality = f(alcohol).

2. No transformation required as the slope ration falls between 1/3 and 3

Sort on X values, since 50 observations, there should be
50/3 = 16 in middle third and 17 each in top and bottom

Quality	Alcohol	Which Third			
4	8.6	L1	Y-left	X-left	
5	8.7	L2	5	9.3	
4	9	L3			
4	9	L4		Slope-left	
5	9.1	L5		1.176471	
5	9.1	L6			
5	9.2	L7			Slope ratio
7	9.2	L8			0.970588
5	9.3	L9			
4	9.3	L10			Since slope ratio is between
4	9.4	L11			1/3 and 3, no need to transform
5	9.4	L12			Alcohol

5	9.5	L17			
7	9.5	M1	Y-mid	X-mid	
5	9.8	M2	6	10.15	
6	9.8	M3			
6	9.8	M4		Slope-right	
4	9.8	M5		1.212121	
6	9.9	M6			

6	10.9	M16			
6	11	R1	Y-right	X-right	
8	11	R2	8	11.8	

3. Anova table & residuals – quality with alcohol (one variable)
   T * for 50 observations at 5% significance (for 1 variable) = 2.0106. It varies upto 2.0141 if say 4 variables are added. This is considered for comparing t stat to assess significance of a variable to be added or not.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.700320228
R Square	0.490448421
Adjusted R Square	0.479832763
Standard Error	1.030323148
Observations	50

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	49.04484212	49.04484212	46.20047352	1.50544E-08
Residual	48	50.95515788	1.061565789		
Total	49	100			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-1.998199293	1.18569598	-1.685254336	0.098431119	-4.38220084	0.385802258	-4.382200843	0.385802258
Alcohol	0.762750266	0.112217129	6.797093019	1.50544E-08	0.537122606	0.988377927	0.537122606	0.988377927

t stat at 95% significance	2.0106		t* for Alcohol is significant

RESIDUAL OUTPUT

Observation	*Predicted Quality*	*Resid(Y-1var)*
1	4.866553104	-0.866553104
2	4.561452998	-0.561452998

4. Add new column of Residual with one variable

Fixed_Acidity	Volatile_Acidity	Citric_Acid	Residual_Sugar	Chlorides	Free_Sulphur	Total_Sulfur_Dioxide	Density	pH	Sulphates	Alcohol	Color (1=white, 0=red)	Resid(Y-1var)	Quality
6.4	0.595	0.14	5.2	0.058	15	97	0.9951	3.38	0.36	9	1	-0.8666	4
7.2	0.4	0.62	10.8	0.041	70	189	0.9976	3.08	0.49	8.6	1	-0.5615	4
6.1	0.28	0.25	12.9	0.054	34	189	0.9979	3.25	0.43	9	1	-0.8666	4
8.2	0.68	0.3	2.1	0.047	17	138	0.995	3.22	0.71	10.8	1	-2.2395	4

5. Check correlation of Residual (y-1Var) with all other variables

	Fixed_Acidity	*Volatile_Acidity*	*Citric_Acid*	*Residual_Sugar*	*Chlorides*	*Free_Sulphur*	*otal_Sulfur_Dioxid*	*Density*	*pH*	*Sulphates*	*Alcohol*	*Color_(1=white)*	*esid(Y-1var*	*Quality*
Fixed_Acidity	1													
Volatile_Acidity	0.07305986	1												
Alcohol	-0.08745087	-0.182991413	0.063230311	-0.339459415	-0.24547195	-0.05102114	-0.128136334	-0.49675	0.23742021	0.129004	1			
Color_(1=white)	-0.46081849	-0.543134165	0.040339215	0.381725663	-0.70045487	0.68334049	0.847399397	-0.53423	-0.2320378	-0.435749	-0.07239333	1		
Resid(Y-1var)	0.076555472	-0.434967673	0.188299815	0.022585475	-0.24487453	0.15608907	-0.058620529	-0.12492	-0.1661096	0.066768	3.28627E-16	0.071023349	1	

Residual (y-1Var) has highest absolute correlation with Volatile Acidity. Check if it needs transformation.

6. No transformation required for Volatile Acidity since one slope positive and other negative

7. Columns readjusted alcohol, volatile acidity adjacent to Residual (y-1Var)

8. Run Quality = f(Alcohol, Volatile acidity). We observe both variables are significant

Sort on X values, since 50 observations, there should be 50/3 = 16 in middle third and 17 each in top and bottom					
Resid(Y-1var)	Volatile_ Acidity	Which Third			
0.9977	0.17	L1	Y-left	X-left	
-0.0954	0.18	L2	0.302846	0.23	
1.2266	0.18	L3			
-0.2395	0.21	L4		Slope-left	
0.5232	0.21	L5		0.616822	
0.3028	0.21	L6			
-0.6125	0.21	L7			Slope ratio
1.9809	0.21	L8			-0.12224
0.9893	0.23	L9			
1.6079	0.24	L10			Since one slope is negative
-0.3158	0.24	L11			and other positive
0.0572	0.25	L12			we avoid transforming
-0.3921	0.25	L13			volatile acidity
0.9215	0.25	L14			
-1.2479	0.27	L15			
-0.8666	0.28	L16			
0.9893	0.28	L17			
-0.2479	0.3	M1	Y-mid	X-mid	
0.7605	0.3	M2	0.404622	0.395	
1.0740	0.3	M3			
-0.0785	0.33	M4		Slope-right	
-1.5362	0.36	M5		-5.04617	

ANOVA

	df	SS	MS	F	Significance F
Regression	2	59.01940565	29.50970282	33.84421468	7.86301E-10
Residual	47	40.98059435	0.871927539		
Total	49	100			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Ipper 95.0%
Intercept	-0.437230522	1.169498601	-0.373861518	0.710188619	-2.78995824	1.91549719	-2.789958239	1.915497
Volatile_Acidity	-2.051159046	0.606446626	-3.382258157	0.001456806	-3.27117229	-0.8311458	-3.271172292	-0.83115
Alcohol	0.698723841	0.103447926	6.754353306	1.92857E-08	0.490613457	0.90683423	0.490613457	0.906834

both are significant

RESIDUAL OUTPUT

Observation	Predicted Quality	Resid(Y-2var)
1	4.630844418	-0.630844418
2	4.751330895	-0.751330895
3	5.276959517	-1.276959517

9. Check correlation of Residual (y-2Var) with all other variables

	Fixed_Acidity	Citric_Acid	Residual_Sugar	Chlorides	Free_Sulphur	Total_Sulfur_Dioxide	Density	pH	Sulphates	lor_(1=whit	/olatile_Acidity
Fixed_Acidity	1										
Citric_Acid	0.608272517	1									
Resid(Y-1var)	0.076555472	0.188299815	0.022585475	-0.244874535	0.156089073	-0.058620529	-0.124919201	-0.16611	0.066768161	0.071023	-0.43496767
Quality	-0.006596111	0.1786953	-0.221608132	-0.346707472	0.07568974	-0.131581489	-0.437053972	0.047696	0.138005381	0	-0.43864505
Resid(Y-2var)	0.113998038	0.022891407	-0.112665536	0.060557564	-0.03560969	-0.274927915	-0.006652579	-0.00422	0.119914203	-0.200011	1.27343E-16

Residual (y-2Var) has highest absolute correlation with Total Sulfur Dioxide. Check if it needs transformation.

10. No transformation required for Total Sulfur dioxide (similar to step 6) since one slope positive and other negative
11. Columns readjusted for alcohol, volatile acidity and total sulfur dioxide
    Run Quality = f(Alcohol, Volatile acidity, total sulfur dioxide)

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	62.87612792	20.95870931	25.9698295	5.58234E-10
Residual	46	37.12387208	0.807040697		
Total	49	100			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Ipper 95.0%*
Intercept	0.808030423	1.261123015	0.640722922	0.52488196	-1.73047854	3.346539388	-1.730478543	3.346539
Total_Sulfur_Di(	-0.004669477	0.002136027	-2.186056917	0.033937956	-0.00896908	-0.000369877	-0.008969077	-0.00037
Volatile_Acidity	-2.655687096	0.645663751	-4.113111647	0.000159645	-3.95534082	-1.356033372	-3.95534082	-1.35603
Alcohol	0.64927147	0.102062905	6.361483323	8.29877E-08	0.443829497	0.854713444	0.443829497	0.854713

Note that all t* for each var given others are sig

Look at correl of Resid(Y-3var) vs all other var

RESIDUAL OUTPUT

Observation	*Predicted Quality*	*Resid(Y-3var)*
1	4.618400569	-0.618400569
2	4.446959084	-0.446959084
3	5.025350124	-1.025350124

We observe all three variables are significant

12. Check correlation of Residual (y-3Var) with all other variables

	Fixed_Acidity	*Citric_Acid*	*Residual_Sugar*	*Chlorides*	*Free_Sulphur*	*Density*	*pH*	*Sulphates*	*Color_(1=white)*	*l_Sulfur_Dic*	*olatile_Acidity*
Fixed_Acidity	1										
Citric_Acid	0.608272517	1									
Residual_Sugar	-0.125891886	0.20405057	1								
Chlorides	0.333858403	-0.110663321	-0.219551289	1							
Free_Sulphur	-0.455611895	0.036655181	0.444615608	-0.522443718	1						
Quality	-0.006596111	0.1786953	-0.221608132	-0.346707472	0.07568974	-0.437053972	0.047696376	0.138005	0	-0.131581	-0.43864505
Resid(Y-2var)	0.113998038	0.022891407	-0.112665536	0.060557564	-0.03560969	-0.006652579	-0.004217753	0.119914	-0.200010554	-0.274928	1.27343E-16
Resid(Y-3var)	-0.050610566	-0.047697311	-0.032502882	-0.044103427	0.169778461	-0.131524027	0.017068491	0.033811	0.004832173	1.44E-16	9.38015E-17

Residual (y-3Var) has highest absolute correlation with free sulfur. Check if it needs transformation.

13. No transformation required since one slope positive and other negative
14. Columns readjusted for alcohol, volatile acidity and total sulfur dioxide, free sulfur.
15. Run Quality = f(Alcohol, Volatile acidity, total sulfur dioxide, free sulfur)

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	65.52763879	16.3819097	21.38484021	6.16887E-10
Residual	45	34.47236121	0.766052471		
Total	49	100			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Ipper 95.0%*
Intercept	0.684187846	1.230482422	0.556032198	0.580943233	-1.79413097	3.162506661	-1.79413097	3.162507
Free_Sulphur	0.01838945	0.009884422	1.860447616	0.06936468	-0.0015188	0.038297698	-0.001518799	0.038298
Total_Sulfur_Di(	-0.008587203	0.002960619	-2.900475027	0.00574553	-0.0145502	-0.002624209	-0.014550196	-0.00262
Volatile_Acidity	-2.45543932	0.638195963	-3.847469216	0.000373626	-3.74083197	-1.170046669	-3.740831972	-1.17005
Alcohol	0.644107093	0.099476071	6.474995306	6.14172E-08	0.443752002	0.844462185	0.443752002	0.844462

t* for Free_Sulphur is no longer significant given other variables so remove it

Each of other t* are sig given all other variables

Run Y=f(Alcohol,Volatile_Acidity,Total_Sulfur_Dioxide)

RESIDUAL OUTPUT

Observation	*Predicted Quality*	*Resid(Y-4var)*
1	4.463048392	-0.463048392
2	4.905613334	-0.905613334
3	4.795888692	-0.795888692

T stat from table = 2.01

T* for Free Sulfur < T stat from table. Therefore, not significant enough, given other variables. Therefore, drop it.

Quality = f(Alcohol, Volatile acidity, total sulfur dioxide) is the final equation
Final Anova and dataset:

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.792944689							
R Square	0.628761279							
Adjusted R Squa	0.604550058							
Standard Error	0.898354439							
Observations	50							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	3	62.87612792	20.95870931	25.9698295	5.58234E-10			
Residual	46	37.12387208	0.807040697					
Total	49	100						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Ipper 95.0%*
Intercept	0.808030423	1.261123015	0.640722922	0.52488196	-1.73047854	3.346539388	-1.730478543	3.346539
Total_Sulfur_Dic	-0.004669477	0.002136027	-2.186056917	0.033937956	-0.00896908	-0.000369877	-0.008969077	-0.00037
Volatile_Acidity	-2.655687096	0.645663751	-4.113111647	0.000159645	-3.95534082	-1.356033372	-3.95534082	-1.35603
Alcohol	0.64927147	0.102062905	6.361483323	8.29877E-08	0.443829497	0.854713444	0.443829497	0.854713
			Each t* is sig with all other variables					
			Look at correl vs Resid(Y-3var2)					
RESIDUAL OUTPUT								
Observation	*Predicted Quality*	*Resid(Y-3var2)*			**Total_Sulfur _Dioxide**	**Volatile_Acidity**	**Alcohol**	**Quality**
1	4.618400569	-0.618400569			97	0.595	9	4
2	4.446959084	-0.446959084			189	0.4	8.6	4
3	5.025350124	-1.025350124			189	0.28	9	4
4	5.369907257	-1.369907257			138	0.68	10.8	4
5	5.343856391	-1.343856391			196	0.27	9.5	4

f. TEST OF ASSUMPTIONS : NORMALITY

Total_Sulfur_Dioxide (x1)	Volatile_Acidity (x2)	Alcohol (x3)	Quality (y)	number of observation	y-ŷ(e) *Residual(Y-3var2)*	rank	normality formula = area= (k-.375)/(n+.25)	expected z = point = xe	normality = √MSE(z(norm))
97	0.595	9	4		-0.6184	12	0.23	-0.73	-0.660
189	0.4	8.6	4		-0.4470	15	0.29	-0.55	-0.494
189	0.28	9	4		-1.0254	7	0.13	-1.12	-1.004

Test of Hypothesis

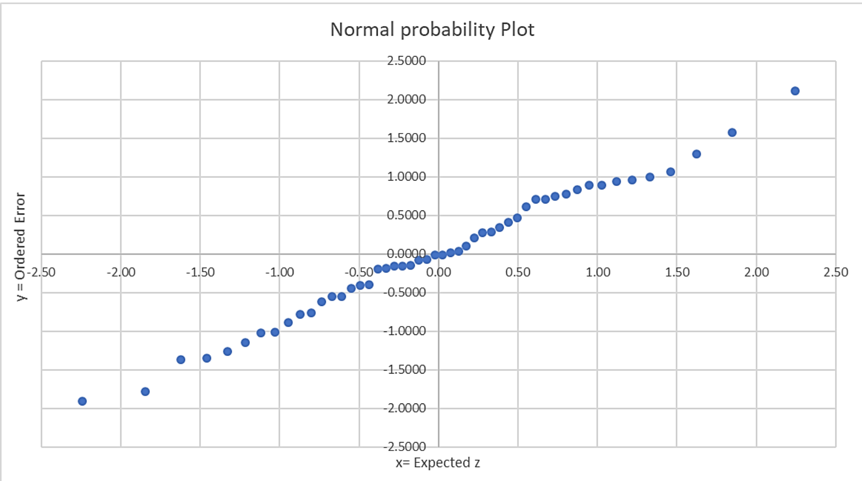
Ho = Normal; if r >= r table conclude Ho; **Ha** = Not Normal; if r < r table conclude Ha

r (correlation coefficient) = 0.995

r from r table at .05 significance for 50 observations = 0.977

Since r > r table, therefore, errors are normal

The graph of ordered error with expected z shoes following observations:



1. Symmetrical error term distribution and not skewed much
2. Heavy tails, i.e. higher probabilities in the tails than normal distribution

g. TEST OF ASSUMPTIONS : HOMOSKEDASTICITY

Calculated by ratio of standard deviation of errors for lowest half of x and standard deviation of errors for upper half of x

y-ŷ(e) Residual(Y-3var2)	ordered residual	std dev of top and bottom half of errors	
-0.6184		**0.7891**	**s1**
-0.4470			
-1.0254			
-1.3699			
-1.3439			

1.2944			
0.9561			
0.8945			
0.7758			
-0.0812		**0.9613**	**s2**
-1.0115			

Test of Hypothesis

Ho = Homoskedastic; if 0.5 < $s1/s2 < 2$ conclude Ho

Ha = Hetroskedastic; if not above conclude Ha

$s1/s2 = 0.8209$

So variability of the residuals in the regression model is constant. The error term does not vary much as the value of predictor variable changes.

h. CROSS VALIDATION

Training set = 70% = 35 observations

Test data = 30% = 15 observations

- New coefficients with training data set and residuals from ANOVA table
- Use the coefficients to predict y hat for test data set.
- Subract y- y hat to get the predicted errors.
- Compare errors with real errors obtained from original 100% observation dataset.

It is observed that the predicted errors of test data are larger than real errors of the best equation data.

189	0.28	9	4	-1.025		-1.10155	3	
110	0.63	9.4	5	0.276		0.333086	32	
170	0.17	11.8	8	0.776		0.820909	25	
28	0.56	9.4	6	0.707		0.807728	38	
87	0.49	14	8	-0.190		0.068067	49	
29	0.38	11.3	8	1.000		1.155008	50	
103	0.18	9.3	5	-0.887		-0.90893	8	
191	0.25	11	6	-0.394		-0.38583	11	**<--Training data**
16	0.4	12.5	8	0.213	7.565	0.435	48	**<--Test data**
15	0.49	9.2	5	-0.410	5.322	-0.322	35	
196	0.27	9.5	4	-1.344	5.404	-1.404	5	
189	0.4	8.6	4	-0.447	4.520	-0.520	2	
23	0.915	10.2	7	2.107	4.690	2.310	43	
29	0.61	9.1	5	0.039	4.866	0.134	31	
238.5	0.3	9.5	5	-0.066	5.149	-0.149	7	
126	0.63	9.5	5	0.285	4.663	0.337	33	
207	0.46	9.8	5	0.017	5.003	-0.003	6	
138	0.33	11.9	7	-0.014	6.914	0.086	18	
218.5	0.37	9.8	6	0.832	5.212	0.788	15	
46	0.61	9.3	4	-1.011	4.918	-0.918	27	
130	0.24	11	8	1.294	6.657	1.343	22	
24	0.21	9.2	7	0.888	6.079	0.921	44	
148	0.18	11.5	8	0.894	7.057	0.943	24	
Total_Sulfur_Dioxide	**Volatile_Acidity**	**Alcohol**	**Quality**	**Real Errors Residual (Y-3var) from best equation**	**y hat**	**Predicted Error (y-yhat)**	**random series**	

SUMMARY OUTPUT with training dataset of 35 observations

Regression Statistics	
Multiple R	0.796490736
R Square	0.634397493
Adjusted R Squ	0.599016605
Standard Error	0.894773126
Observations	35

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	43.06652692	14.35551	17.93051	6.211E-07
Residual	31	24.81918736	0.800619		
Total	34	67.88571429			

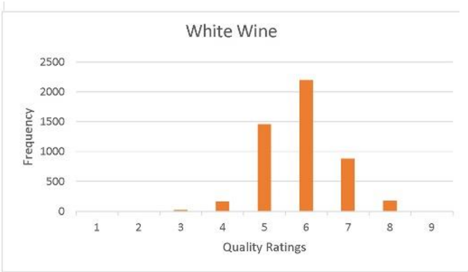
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	1.215679777	1.468999508	0.827556	0.414247	-1.780364	4.211724	-1.7803645	4.21172403
Total_Sulfur_D	-0.003989142	0.002894849	-1.37801	0.178065	-0.009893	0.0019149	-0.0098932	0.00191494
Volatile_Acidit	-2.832119501	0.725847661	-3.90181	0.00048	-4.312496	-1.3517434	-4.3124956	-1.3517434
Alcohol	0.603646241	0.117845867	5.122337	1.51E-05	0.363298	0.8439945	0.363298	0.84399447

RESIDUAL OUTPUT

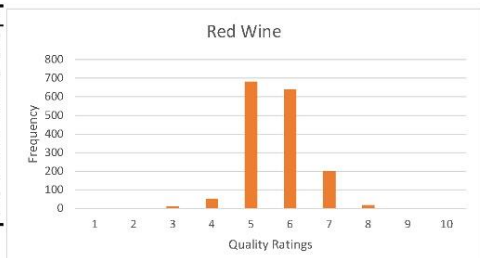
Observation	*Predicted Quality*	*Residuals*
1	4.631273721	0.368726279
2	7.103587659	-0.103587659
3	5.234915483	-0.234915483

DATASET with 50 observations

rating	Frequency
1	0
2	0
3	20
4	163
5	1457
6	2198
7	880
8	175
9	5
10	0



rating	Frequency
1	0
2	0
3	10
4	53
5	681
6	638
7	199
8	18
9	0
10	0



To limit the scope of work, combine and take a random sample of 50 from the dataset based on average quality rating (4 to 8). Therefore 6 random samples each for range 4 to 8.

fixed acidity g/l	volatile acidity g/l	citric acid g/l	residual sugar g/l	chlorides g/l	free sulfur dioxide mg/l	total sulfur dioxide mg/l	density g/cm3	pH	sulphates g/l	alcohol % Volume	quality 0-10	Wine Color (1= white, 0 = red)
6.4	0.595	0.14	5.2	0.058	15	97	0.9951	3.38	0.36	9	4	1
7.2	0.4	0.62	10.8	0.041	70	189	0.9976	3.08	0.49	8.6	4	1
6.1	0.28	0.25	12.9	0.054	34	189	0.9979	3.25	0.43	9	4	1
8.2	0.68	0.3	2.1	0.047	17	138	0.995	3.22	0.71	10.8	4	1
6.4	0.27	0.19	1.9	0.085	21	196	0.99516	3.49	0.64	9.5	4	1
6.2	0.46	0.25	4.4	0.066	62	207	0.9939	3.25	0.52	9.8	5	1
6.8	0.3	0.23	4.6	0.061	50.5	238.5	0.9958	3.32	0.6	9.5	5	1
7.4	0.18	0.3	8.8	0.064	26	103	0.9961	2.94	0.56	9.3	5	1
7.1	0.43	0.61	11.8	0.045	54	155	0.9974	3.11	0.45	8.7	5	1
7.4	0.25	0.37	13.5	0.06	52	192	0.9975	3	0.44	9.1	5	1
6.9	0.25	0.35	1.3	0.039	29	191	0.9908	3.13	0.52	11	6	1
7.9	0.21	0.4	1.2	0.039	38	107	0.992	3.21	0.54	10.8	6	1
7.3	0.41	0.24	6.8	0.057	41	163	0.9949	3.2	0.41	9.9	6	1
7.4	0.21	0.27	1.2	0.041	27	99	0.9927	3.19	0.33	9.8	6	1
6	0.37	0.32	1	0.053	31	218.5	0.9924	3.29	0.72	9.8	6	1
7.1	0.21	0.37	2.4	0.026	23	100	0.9903	3.15	0.38	11.4	7	1
6.8	0.21	0.27	2.1	0.03	26	139	0.99	3.16	0.61	12.6	7	1
7.3	0.33	0.4	6.85	0.038	32	138	0.992	3.03	0.3	11.9	7	1
6.9	0.23	0.38	8.3	0.047	47	162	0.9954	3.34	0.52	10.5	7	1
6.4	0.28	0.29	1.6	0.052	34	127	0.9929	3.48	0.56	10.5	7	1
5.2	0.44	0.04	1.4	0.036	43	119	0.9894	3.36	0.33	12.1	8	1
6.2	0.24	0.29	13.3	0.039	49	130	0.9952	3.33	0.46	11	8	1
7.3	0.25	0.36	2.1	0.034	30	177	0.99085	3.25	0.4	11.9	8	1
6.5	0.18	0.34	1.6	0.04	43	148	0.9912	3.32	0.59	11.5	8	1
5.8	0.17	0.34	1.8	0.045	96	170	0.99035	3.38	0.9	11.8	8	1
5.7	1.13	0.09	1.5	0.172	7	19	0.994	3.5	0.48	9.8	4	0
8.8	0.61	0.3	2.8	0.088	17	46	0.9976	3.26	0.51	9.3	4	0
7	0.975	0.04	2	0.087	12	67	0.99565	3.35	0.6	9.4	4	0
9.9	0.5	0.24	2.3	0.103	6	14	0.9978	3.34	0.52	10	4	0
10.1	0.935	0.22	3.4	0.105	11	86	1.001	3.43	0.64	11.3	4	0
7.8	0.61	0.29	1.6	0.114	9	29	0.9974	3.26	1.56	9.1	5	0
7.5	0.63	0.12	5.1	0.111	50	110	0.9983	3.26	0.77	9.4	5	0
6.8	0.63	0.12	3.8	0.099	16	126	0.9969	3.28	0.61	9.5	5	0
8.2	0.57	0.26	2.2	0.06	28	65	0.9959	3.3	0.43	10.1	5	0
11.7	0.49	0.49	2.2	0.083	5	15	1	3.19	0.43	9.2	5	0
7.8	0.6	0.14	2.4	0.086	3	15	0.9975	3.42	0.6	10.8	6	0
7.3	0.39	0.31	2.4	0.074	9	46	0.9962	3.41	0.54	9.4	6	0
7.8	0.56	0.12	2	0.082	7	28	0.997	3.37	0.5	9.4	6	0
10.2	0.36	0.64	2.9	0.122	10	41	0.998	3.23	0.66	12.5	6	0
8.2	0.24	0.34	5.1	0.062	8	22	0.9974	3.22	0.94	10.9	6	0
7.5	0.52	0.16	1.9	0.085	12	35	0.9968	3.38	0.62	9.5	7	0
12.8	0.3	0.74	2.6	0.095	9	28	0.9994	3.2	0.77	10.8	7	0
7.7	0.915	0.12	2.2	0.143	7	23	0.9964	3.35	0.65	10.2	7	0
15	0.21	0.44	2.2	0.075	10	24	1.00005	3.07	0.84	9.2	7	0
15.6	0.685	0.76	3.7	0.1	6	43	1.0032	2.95	0.68	11.2	7	0
9.4	0.3	0.56	2.8	0.08	6	17	0.9964	3.15	0.92	11.7	8	0
5	0.42	0.24	2	0.06	19	50	0.9917	3.72	0.74	14	8	0
9.1	0.4	0.5	1.8	0.071	7	16	0.99462	3.21	0.69	12.5	8	0
5.5	0.49	0.03	1.8	0.044	28	87	0.9908	3.5	0.82	14	8	0
7.2	0.38	0.31	2	0.056	15	29	0.99472	3.23	0.76	11.3	8	0

10. CITATIONS

"Breaking Down the Booze: Wine Alcohol Levels Explored." *Www.whichwinery.com*, https://www.whichwinery.com/ask-the-somm/breaking-down-booze-wine-alcohol-levels-explored/.

"Chloride concentration in red wines: Influence of terroir and grape type" *Coli, Marina & Rangel, Angelo & Souza, Elizangela & Oliveira, Margareth & Chiaradia, Ana.* (2015). Food Science and Technology (Campinas). 35. 95-99. 10.1590/1678-457X.6493.

"Salt in grapes and wine a common issue. " *Www.awri.com.au*, https://www.awri.com.au/wp-content/uploads/2018/08/s1530.pdf

"What Is Residual Sugar in Wine?" *Wine Folly*, 20 Sept. 2019, https://winefolly.com/review/what-is-residual-sugar-in-wine/.

"Wine-Tasting by Numbers: Using Binary Logistic Regression to Reveal the Preferences of Experts." *Minitab, Inc.*, https://www.minitab.com/en-us/Published-Articles/Wine-Tasting-by-Numbers--Using-Binary-Logistic-Regression-to-Reveal-the-Preferences-of-Experts/.