INST627

Data Analytics for Information Professionals

White Paper:

Analysis of Red Wines: Data Analytics Project

Submitted by:
Prajakta Arolker, parolker@umd.edu
Osheen Shrivastav, oshrivas@umd.edu
Anuj Shah, ashah126@umd.edu
Anuj Sharma, asharm24@umd.edu

CONTENTS

List of Figures

List of Tables

Table of Contents

PROBLEM DOMAIN

To analyze the components of red wines with respect to its quality in order to enable the industry to produce better red wines.

RESEARCH QUESTION

Which chemical components influence the quality of red wines?

MOTIVATION

According to International Organization of Vine and Wine, US has overtaken France as the world's biggest market for wine. The purpose of this research is to study the effect of components of red wines on quality and provide recommendations to improve the quality with respect to its aroma, taste and flavor based on the results.

The chemical components chosen to perform statistical analysis on (not in any chronological order of importance) are:

1 - pH value
2 – Alcohol

QUESTIONS ATTEMPTED TO ANSWER BY THIS RESEARCH PROJECT

1 - Why pH and Alcohol are chosen for analysis to determine the effects of chemical components on red wines?
2 - Descriptives on the chemical components chosen.
3 - What are the statistical tests performed?
4 - Why are the chosen tests performed?
5 - What are our observations and inferences from the same?

1 - Why pH and Alcohol are chosen for analysis to determine the effects of chemical components on red wines?

1. pH content

- The pH value affects flavor, aroma, color of the wine
- It also affects the chemical reactions that take place in a wine during and after fermentation
- When the acid levels are too low wine will lack body, the mouthfeel will be off, and it will taste weak
- The pH equally affects the stability of red wine. Most types of bacteria and a few types of fungi do not survive at pH levels of 3.0 to 3.75 which may act as a natural protection against spoilage microorganisms

2. Alcohol:

- Alcohol plays a very important role in the structure of the and mouthfeel quality of wine
- It acts as a preservative too and contributes to the flavor of wine
- Alcohol is the main carrier of aroma and bouquet and hence flavors of wine
- Alcohol provides balance to a wine

DATASET:

This dataset is public available for research. The details are described in [Cortez et al., 2009].

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553. ISSN: 0167-9236

Number of Instances: red wine - 1599

Number of Attributes: 11 + output attribute.

The inputs include objective tests (e.g. PH values) and the output is based on sensory data. Each expert graded the wine quality between 0 (very bad) and 10 (very excellent).

Input variables:

1 - fixed acidity (tartaric acid - g / dm^3)
2 - volatile acidity (acetic acid - g / dm^3)
3 - citric acid (g / dm^3)
4 - residual sugar (g / dm^3)
5 - chlorides (sodium chloride - g / dm^3
6 - free sulfur dioxide (mg / dm^3)
7 - total sulfur dioxide (mg / dm^3)
8 - density (g / cm^3)
9 - pH
10 - sulphates (potassium sulphate - g / dm3)
11 - alcohol (% by volume)
Output variable (based on sensory data):
12 - quality (score between 0 and 10)

Missing Attribute Values: None

Statistics

		Sr. No.	fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides	free.sulfur.dioxide	total.sulfur.dioxide	density	pH Value	sulphates	Alcohol Content	Quality of the wine	pH Value (Binned)	pH Value (Binned)	Alcohol Content
N	Valid	1599	1599	1599	1599	1599	1599	1599	1599	1599	1599	1599	1599	1599	1599	1599	1599
	Missing	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

DESCRIPTIVE ANALYSIS:

Quality, pH and Alcohol

Quality

Quality (score between 0 and 10)

Statistics

Quality

N	Valid	1599
	Missing	0
Mean		5.64
Median		6.00
Mode		5
Std. Deviation		.808
Minimum		3
Maximum		8

Table 1: Descriptive for Quality

Quality

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	3	10	.6	.6	.6
	4	53	3.3	3.3	3.9
	5	681	42.6	42.6	46.5
	6	638	39.9	39.9	86.4
	7	199	12.4	12.4	98.9
	8	18	1.1	1.1	100.0
	Total	1599	100.0	100.0	

Table 2: Descriptive with respect to range of Quality

Observations and Inferences:

Quality rating ranges from 3 to 8

Mode = 5

Maximum number of wines have quality rating of 5 with the average quality rating of the wines being 5.64. Only 1.1% of the wines have the highest rating of 8.

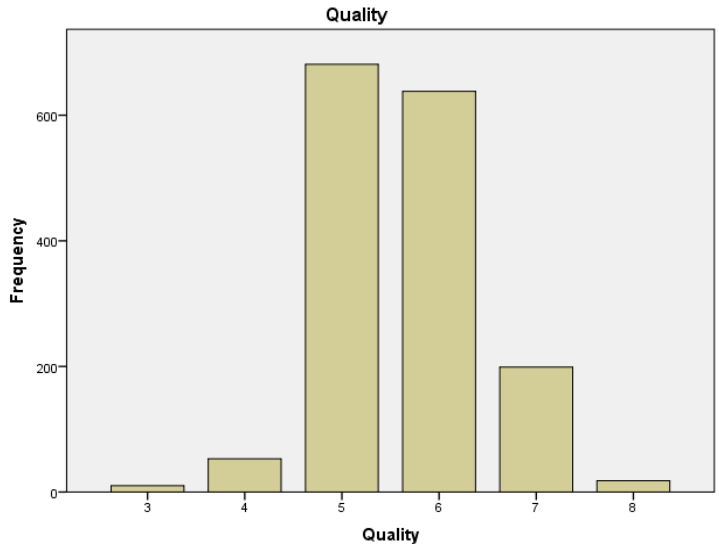


Figure 1: Bar Graph for Quality variable

pH Content

pH: describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale

Statistics

pH

N	Valid	1599
	Missing	0
Mean		3.3111
Std. Error of Mean		.00386
Median		3.3100
Mode		3.30
Std. Deviation		.15439
Variance		.024
Skewness		.194
Std. Error of Skewness		.061
Kurtosis		.807
Std. Error of Kurtosis		.122
Range		1.27
Minimum		2.74
Maximum		4.01

Table 3: Descriptive for pH variable

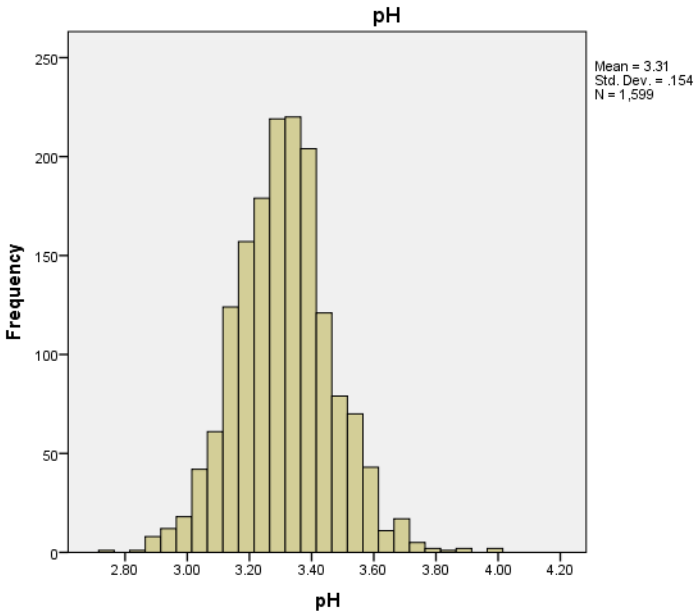


Figure 2: Histogram for pH variable

Observations and Inferences:

pH ranges from 2.74 to 4.01.

The mean and mode are approximately equal ay 3.30.

Maximum number of red wines in the data set have pH 3.3

Alcohol Content

Alcohol: the percent alcohol content of the wine

Statistics

alcohol

N	Valid	1599
	Missing	0
Mean		10.4230
Std. Error of Mean		.02665
Median		10.2000
Mode		9.50
Std. Deviation		1.06567
Variance		1.136
Skewness		.861
Std. Error of Skewness		.061
Kurtosis		.200
Std. Error of Kurtosis		.122
Range		6.50
Minimum		8.40
Maximum		14.90

Table 4: Descriptive for Alcohol variable

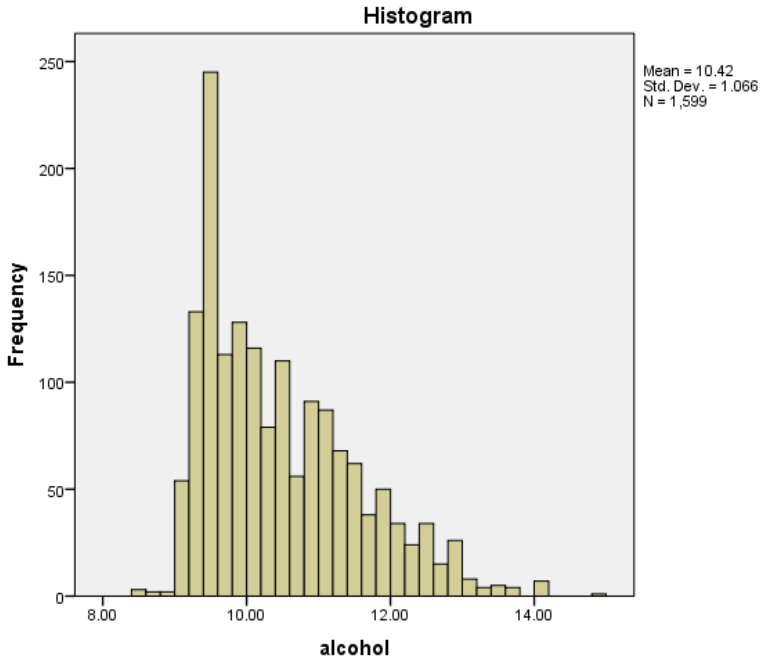


Figure 3: Histogram for Alcohol variable

Observations and Inferences:

Alcohol content in the wines ranges from 8.40 -14.90 with average being 10.4

Maximum number of red wines in the data set have alcohol 9.5

TEST FOR NORMALITY OF DATA:

Normality of Data has been tested in two ways:

1. Skewness and Kurtosis: It should be near to 0, less than 1.
2. Histograms and Normal Q-Q Plots: They should visually indicate that the data is approximately normally distributed

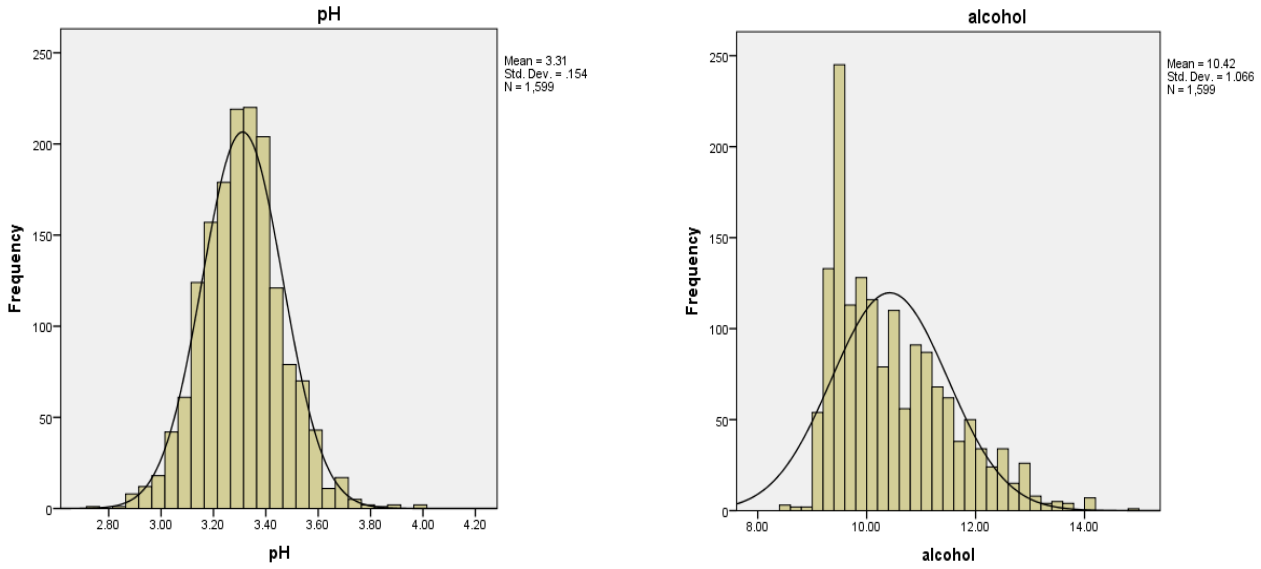


Figure 4: Normal curve over histogram for pH and Alcohol variable

Statistics

		pH	alcohol
N	Valid	1599	1599
	Missing	0	0
Mean		3.3111	10.4230
Median		3.3100	10.2000
Mode		3.30	9.50
Skewness		.194	.861
Std. Error of Skewness		.061	.061
Kurtosis		.807	.200
Std. Error of Kurtosis		.122	.122
Range		1.27	6.50

Table 5: Descriptives with Skewness and Kurtosis for pH and Alcohol

Observations and Inferences:

Skewness for pH is 0.194 and alcohol is 0.861. This indicates a weak positive skewness. Since for both, pH and alcohol content the values are less than 1 and closer to 0. The data is close to normal distribution.
There is not much departure from normality.

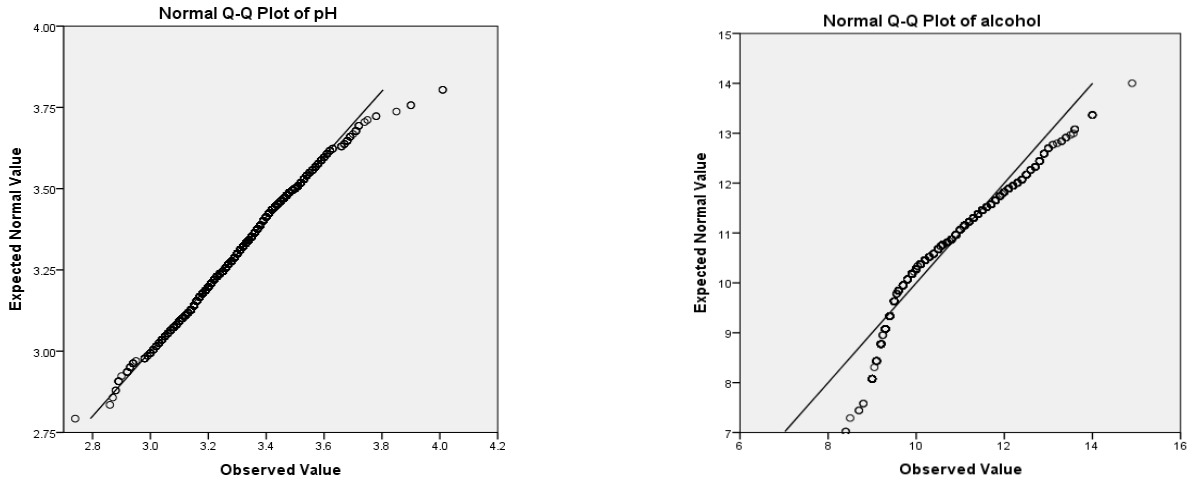


Figure 4: Normal Q-Q Plots for pH and Alcohol variables

The Q-Q Plots also show that the plot is approximately around the normal distribution line, with a few outliers in alcohol.

CORRELATION AND REASON FOR USING CORRELATION:

Correlation is a statistical measure that indicates the extent to which two or more variables fluctuate together.

Here, we want to find a correlation between pH (predictor) and Quality (dependent variable), and Alcohol (predictor) and Quality (dependent variable), and find out how these variables fluctuate together.

Pearson's correlation measures the degree between the two variables. By linear relationship we mean that the relationship can be well-characterized by a straight line.

Correlation between pH content and Quality

Correlation with extreme values (outliers):
Correlation is denoted by *r* and ranges from -1.0 to 1.0.
Level of significance chosen: $\alpha = 0.05$

Correlations

		Quality	pH content
Quality	Pearson Correlation	1	-.058
	Sig. (2-tailed)		.021
	N	1599	1599
pH content	Pearson Correlation	-.058	1
	Sig. (2-tailed)	.021	
	N	1599	1599

Table 6: Correlation between pH and Quality

Here, r = - 0.058 which indicates that observed correlation is negative. There is weak negative correlation between the two variables. This means as the pH content decreases, Quality increases.
The Significance value, i.e. p value = 0.021. Since p < 0.05 , above relationship between pH content and Quality is statistically valid and significant.

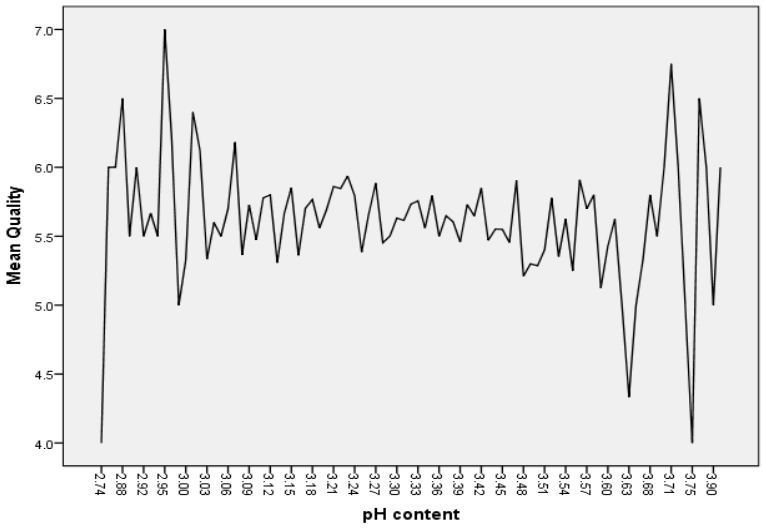


Figure 5: Line graph of pH versus Mean quality

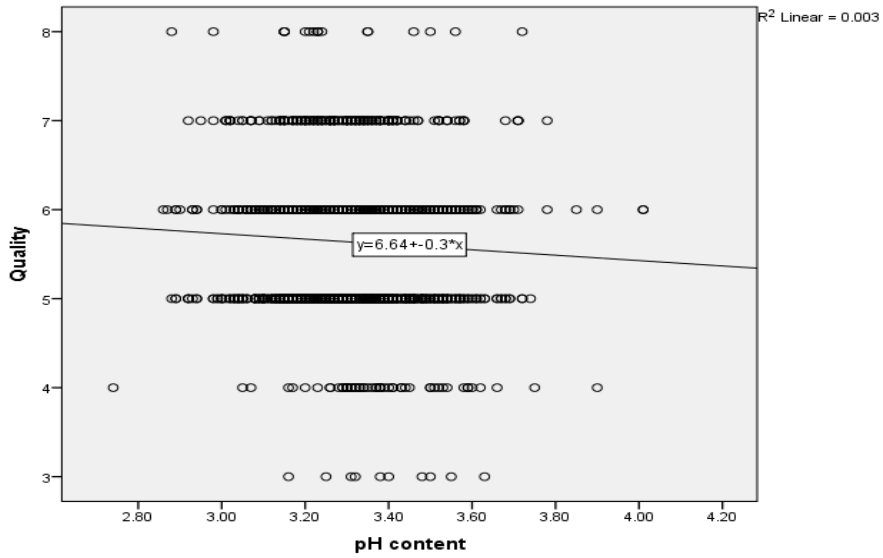


Figure 6: Scatter plot of pH versus Quality with line of best fit

The line graph and the scatter plot drawn above clearly represent a negative correlation between pH and Quality of the wine. It is observed from the descriptive and the above correlation analysis that majority of the red wines are rated as 5 and 6 in terms of their quality and these wines are neither too acidic nor too basic i.e. they neither have a very high pH value nor a very low pH value. It is also observed that wines with quality ratings 7 and 8, again have pH moderate pH values.

A wine can neither be too acidic nor too basic for it to have a good quality i.e. good aroma, flavor and content.

Outliers:

An outlier is an observation whose dependent-variable value is unusual or extreme given its values on the predictor variables. An outlier may indicate a sample peculiarity or may indicate a data entry error or other problem.
There are mild outliers and extreme outliers.

Examining extreme outliers for pH:

Extreme outliers are any data values which lie more than 3.0 times the interquartile range below the first quartile Q1 or above the third quartile Q3.

Using John Tukey's method of leveraging the Interquartile Range,

an observation *x* is an extreme outlier if:
x < Q1 - 3 * IQR

or

x > Q3 + 3 * IQR

Percentiles

		Percentiles						
		5	10	25	50	75	90	95
Weighted Average(Definition 1)	pH	3.0600	3.1200	3.2100	3.3100	3.4000	3.5100	3.5700
Tukey's Hinges	pH			3.2100	3.3100	3.4000		

For x > Q3 + 3 * IQR
3.4 + (3 * (4.3 - 3.21)
= 3.97

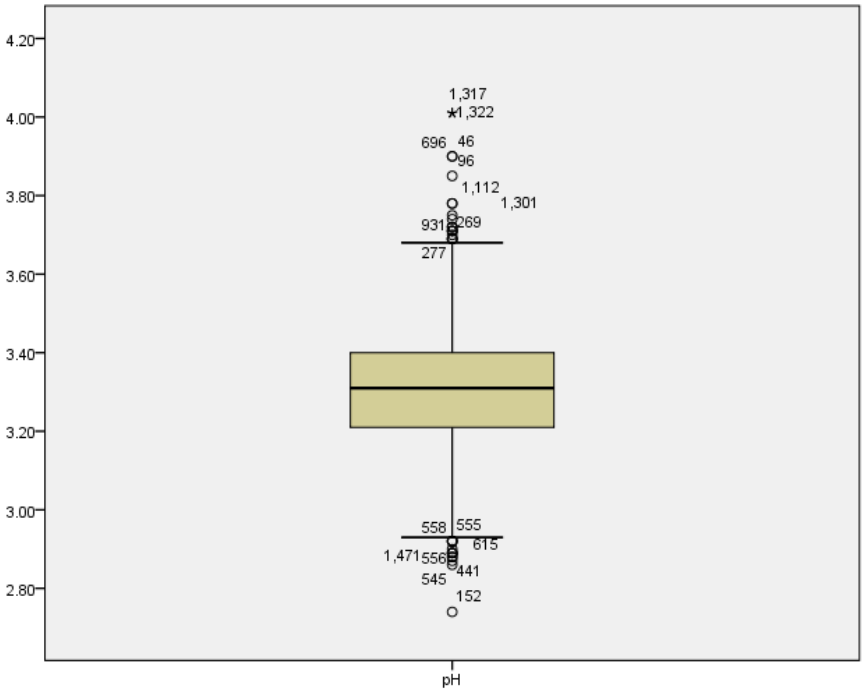
Any value that is greater than 3.97 can be considered as an extreme outlier.

For x < Q1 - 3 * IQR
3.21 – (3 * (4.3 - 3.21)
= 2.64

Any value less than 2.64 can be considered as an extreme outlier. Since minimum value of pH in the dataset is 2.74 there is no extreme outlier below the first Quartile (Q1)

Extreme Values

			Case Number	Value
pH	Highest	1	1317	4.01
		2	1322	4.01
		3	46	3.90
		4	696	3.90
		5	96	3.85
	Lowest	1	152	2.74
		2	545	2.86
		3	615	2.87
		4	1471	2.88
		5	441	2.88



Removing value 4.01 i.e. case numbers 1317 and 1322 from the data set since it is greater than 3.97 as calculated above. Extreme outlier is denoted by asterisk (*) in SPSS which can be seen above in the box-plot.

Correlation (without extreme outliers) :

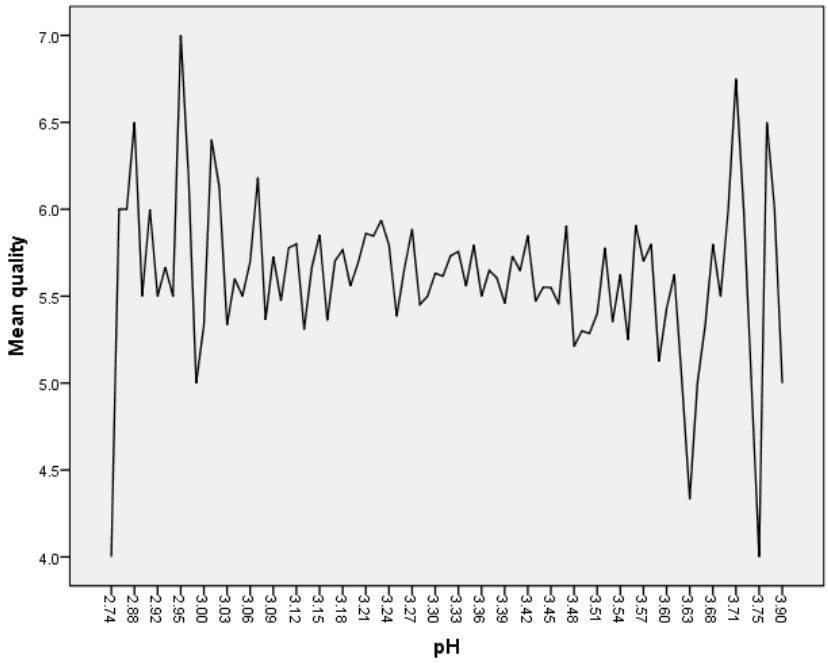
Correlations

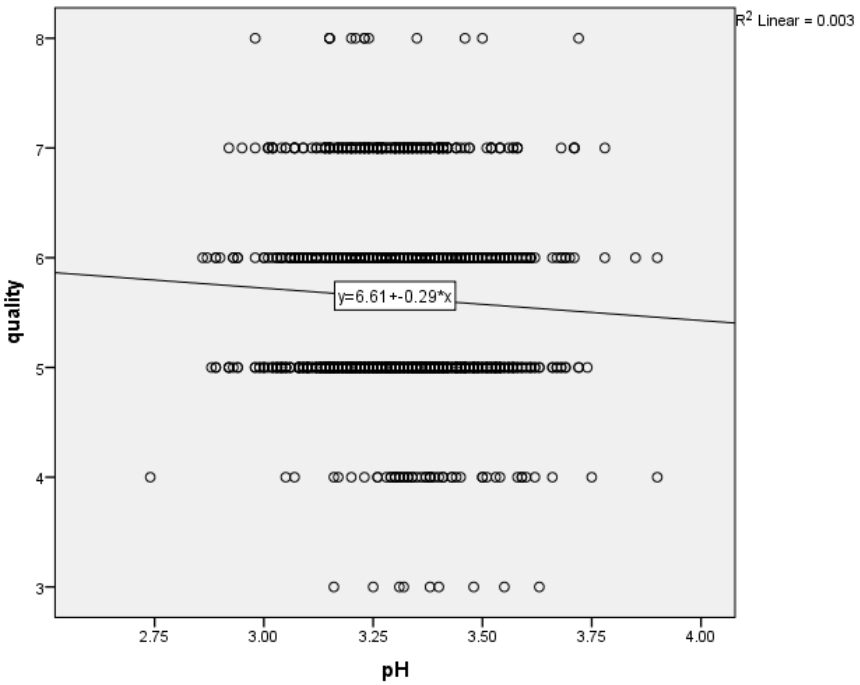
		pH	quality
pH	Pearson Correlation	1	-.061
	Sig. (2-tailed)		.015
	N	1597	1597
quality	Pearson Correlation	-.061	1
	Sig. (2-tailed)	.015	
	N	1597	1597

Two extreme cases removed. New N = 1597

Here r = - 0.061, the observed correlation is negative. There is weak negative correlation between the two variables. This indicates that's as the pH content decreases, Quality increases.

The Significance value, p = 0.015. Since p < 0.05, above relationship between pH content and Quality is statistically valid and significant.





The line graph and the scatter plot drawn above clearly represent a negative correlation between pH and Quality of the wine. It is observed from the descriptive and the above correlation analysis that majority of the red wines are rated as 5 and 6 in terms of their quality and these wines are neither too acidic nor too basic i.e. they neither have a very high pH value nor a very low pH value. It is also observed that wines with quality ratings 7 and 8, again have pH moderate pH values.

A wine can neither be too acidic nor too basic for it to have a good quality i.e. good aroma, flavor and content.

Inferences:

1. It can be observed that before and after the removal of extreme outliers
    a. The trend remains the same i.e. a weak negative correlation
    b. The quality of wine decreases weakly (since weak correlation) with an increase in pH value
2. Since it is not a strong relation, there is not enough evidence to infer that the quality of wine strongly depends on pH
3. There are other variables on which the quality rating depends
4. Similar analysis is to be performed on other variables

Correlation between Alcohol content and Quality

Correlation with extreme values (outliers)

Correlations

		Quality	Alcohol content
Quality	Pearson Correlation	1	.476
	Sig. (2-tailed)		.000
	N	1599	1599
Alcohol content	Pearson Correlation	.476	1
	Sig. (2-tailed)	.000	
	N	1599	1599

Table 7: Correlation between Alcohol and Quality

Here r = 0.476, observed correlation is positively. There is moderate positive correlation between the two variables. This indicates as the Alcohol content increases, Quality also increases.

The Significance value, p = 0.000. Since p < 0.05, the relationship between Alcohol content and Quality is statistically significant.

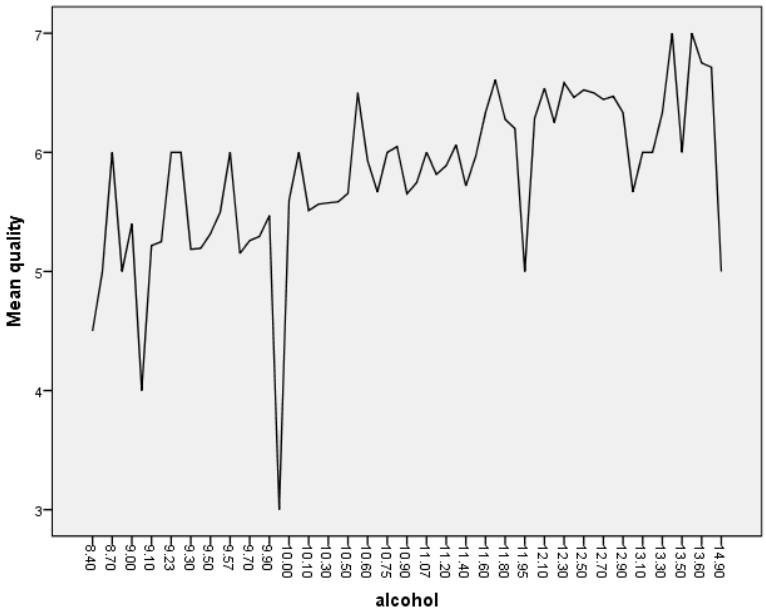


Figure 7: Line graph of Alcohol versus Mean Quality

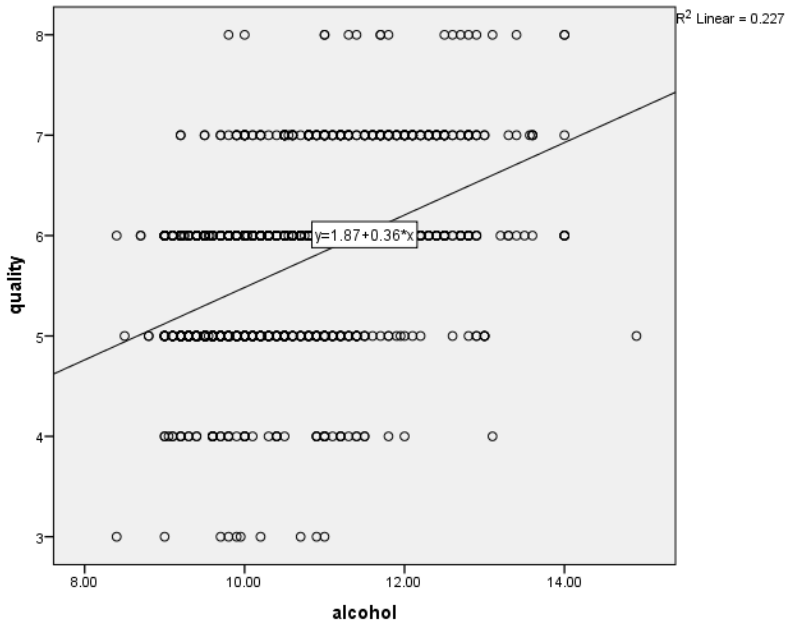


Figure 8: Scatter plot of Alcohol versus Quality with line of best fit

It is observed that wines with quality ratings 7 and 8 have high alcohol contents. The line graph and the scatter plot drawn above clearly represents a strong positive correlation between alcohol and Quality of the wine.

A wine with higher levels of alcohol content is rated as a good quality wine i.e. with good aroma, flavor and content.

Examining extreme outliers for Alcohol:

Extreme outliers are any data values which lie more than 3.0 times the interquartile range below the first quartile or above the third quartile.

For extreme outliers for alcohol:

Percentiles

		Percentiles						
		5	10	25	50	75	90	95
Weighted Average(Definition 1)	alcohol	9.200	9.300	9.500	10.200	11.100	12.000	12.500
Tukey's Hinges	alcohol			9.500	10.200	11.100		

For x > Q3 + 3 * IQR
11.1 + (3 * (11.1 – 9.5))
= 15.9

Any value that is greater than 3.97 can be considered as an extreme outlier.

For x < Q1 - 3 * IQR
9.5 – (3 * (11.1-9.5))
= 4.7

Descriptives

			Statistic	Std. Error
alcohol	Mean		10.423	.0267
	95% Confidence Interval for Mean	Lower Bound	10.371	
		Upper Bound	10.475	
	5% Trimmed Mean		10.356	
	Median		10.200	
	Variance		1.136	
	Std. Deviation		1.0657	
	Minimum		8.4	
	Maximum		14.9	
	Range		6.5	
	Interquartile Range		1.6	
	Skewness		.861	.061
	Kurtosis		.200	.122

For Alcohol, minimum value is 8.4 and maximum value is 14.9. Hence, the extreme outliers 4.7 and 15.9 do not exist in this data set for alcohol.

REGRESSION AND REASON FOR USING REGRESSION:

Regression analysis is another statistical process for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, with the focus on the relationship between a dependent variable and one or more independent variables.

Here, the focus is pH (predictor) and Quality (dependent variable), and Alcohol (predictor) and Quality (dependent variable).

Multiple Regression on predictors' pH and Alcohol to find the effect of each on Quality

Case I: With extreme outliers included:
Total data N = 1599

a: Order in which variables added: pH, Alcohol

Descriptive Statistics

	Mean	Std. Deviation	N
quality	5.64	.808	1599
alcohol	10.423	1.0657	1599
pH	3.3111	.15439	1599

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	pH content[b]	.	Enter
2	Alcohol content[b]	.	Enter

a. Dependent Variable: Quality

b. All requested variables entered.

Table 8: Table of variables entered hierarchically

The two variables, pH and Alcohol were added hierarchically to perform multiple regression.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.058[a]	.003	.003	.806
2	.502[b]	.252	.251	.699

a. Predictors: (Constant), pH content

b. Predictors: (Constant), pH content, Alcohol content

Table 9: Model Summary

This table shows the percent of variability in the dependent variable that can be accounted for by the two predictors together i.e. R^2. The weighted change in is a way to evaluate how much predictive power was added to the model by the addition of another variable in each step or block. Here, we can see that after adding Alcohol to the predictors the percent variability increases. In this case, the % of variability accounted for went up from **0.3 to 25.2%.**

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3.473	1	3.473	5.340	.021[b]
	Residual	1038.692	1597	.650		
	Total	1042.165	1598			
2	Regression	262.657	2	131.328	268.888	.000[c]
	Residual	779.508	1596	.488		
	Total	1042.165	1598			

a. Dependent Variable: Quality

b. Predictors: (Constant), pH content

c. Predictors: (Constant), pH content, Alcohol content

Table 10: ANOVA values on pH and Alcohol

Since Sig. values p, 0.21 and 0.00 are less than α (0.05), **there is enough evidence to show that the tests are statistically significant.**

b: Order in which variables added: Alcohol, pH

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	alcohol[b]	.	Enter
2	pH[b]	.	Enter

a. Dependent Variable: quality

b. All requested variables entered.

The two variables, Alcohol and pH were added hierarchically to perform multiple regression.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.476[a]	.227	.226	.710
2	.502[b]	.252	.251	.699

a. Predictors: (Constant), alcohol

b. Predictors: (Constant), alcohol, pH

This table shows you the percent of variability in the dependent variable that can be accounted for by the two predictors together i.e. R^2. The weighted change in is a way to evaluate how much predictive power was added to the model by the addition of another variable in each step or block. Here, we can see that after adding pH to the predictors the percent variability increases, but as significantly as when pH was added first and then Alcohol in 'Case I a'. In this case, the % of variability accounted for went up from **22.7 to 25.2%.**

Hence, alcohol has more significant effect on quality of red wines.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	236.295	1	236.295	468.267	.000[b]
	Residual	805.870	1597	.505		
	Total	1042.165	1598			
2	Regression	262.657	2	131.328	268.888	.000[c]
	Residual	779.508	1596	.488		
	Total	1042.165	1598			

a. Dependent Variable: quality

b. Predictors: (Constant), alcohol

c. Predictors: (Constant), alcohol, pH

Since Sig. values p for both is 0.00 which is less than α (0.05), **there is enough evidence to show that the tests are statistically significant.**

Multiple Regression on predictors pH and Alcohol to find the effect of each on Quality

Case II: Without extreme outliers included:
Total data N = 1597

a: Order in which variables added: pH, Alcohol

Descriptive Statistics

	Mean	Std. Deviation	N
quality	5.64	.808	1597
alcohol	10.420	1.0638	1597
pH	3.3102	.15249	1597

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	pH[b]	.	Enter
2	alcohol[b]	.	Enter

a. Dependent Variable: quality

b. All requested variables entered.

The two variables, pH and Alcohol were added hierarchically to perform multiple regression.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.061[a]	.004	.003	.807
2	.502[b]	.252	.251	.699

a. Predictors: (Constant), pH

b. Predictors: (Constant), pH, alcohol

This table shows you the percent of variability in the dependent variable that can be accounted for by the two predictors together i.e. R^2. The weighted change in is a way to evaluate how much predictive power was added to the model by the addition of another variable in each step or block. Here, we can see that after adding Alcohol to the predictors the percent variability increases. In this case, the % of variability accounted for went up from **0.4 to 25.2%.**

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3.888	1	3.888	5.974	.015[b]
	Residual	1038.012	1595	.651		
	Total	1041.900	1596			
2	Regression	262.442	2	131.221	268.348	.000[c]
	Residual	779.458	1594	.489		
	Total	1041.900	1596			

a. Dependent Variable: quality

b. Predictors: (Constant), pH

c. Predictors: (Constant), pH, alcohol

Since Sig. values p in the table are 0.015 and 0.00 which is less than α (0.05), **there is enough evidence to show that the tests are statistically significant.**

a: Order in which variables added: Alcohol, pH

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	alcohol[b]	.	Enter
2	pH[b]	.	Enter

a. Dependent Variable: quality

b. All requested variables entered.

The two variables, Alcohol and pH were added hierarchically to perform multiple regression.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.476[a]	.227	.226	.711
2	.502[b]	.252	.251	.699

a. Predictors: (Constant), alcohol

b. Predictors: (Constant), alcohol, pH

This table shows you the percent of variability in the dependent variable that can be accounted for by the two predictors together i.e. R^2. The weighted change in is a way to evaluate how much predictive power was added to the model by the addition of another variable in each step or block. Here, we can see that after adding pH to the predictors the percent variability increases, but as significantly as when pH was added first and then Alcohol in 'Case I a'. In this case, the % of variability accounted for went up from **22.7 to 25.2%.**

Hence, alcohol has more significant effect on quality of red wines.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	236.328	1	236.328	467.921	.000[b]
	Residual	805.571	1595	.505		
	Total	1041.900	1596			
2	Regression	262.442	2	131.221	268.348	.000[c]
	Residual	779.458	1594	.489		
	Total	1041.900	1596			

a. Dependent Variable: quality

b. Predictors: (Constant), alcohol

c. Predictors: (Constant), alcohol, pH

Since Sig. values p for both is 0.00 which is less than α (0.05), **there is enough evidence to show that the tests are statistically significant.**

SPLITTING THE DATA SETS FOR TESTS

Splitting pH and Alcohol into two sets with respect to Quality

Wines having a pH value of 3.25- are grouped as 1 while Wines having a pH value of 3.25+ are grouped as 2, with 1 being the poor (low) quality wines and 2 being the rich (high) quality wines.

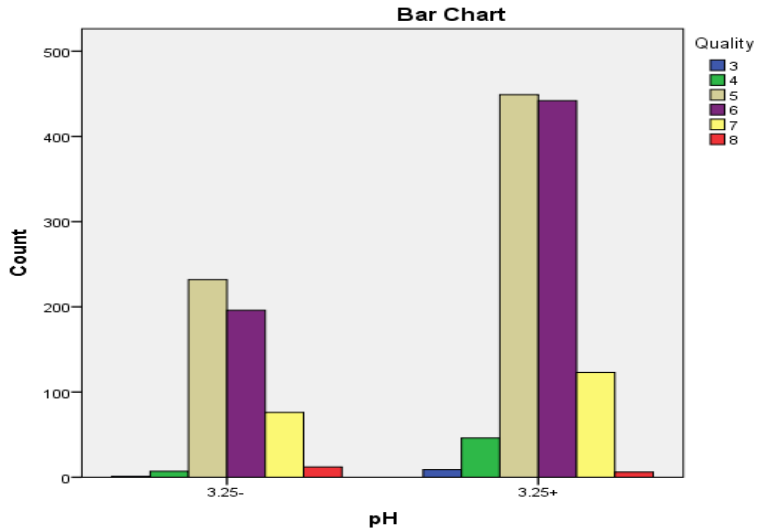


Figure 9: pH variable split

Wines having mean alcohol content less than 10 are grouped as 1 and with mean alcohol content as more than 10 value of 3.25+ are grouped as 2, with 1 being the poor (low) quality wines and 2 being the rich (high) quality wines.

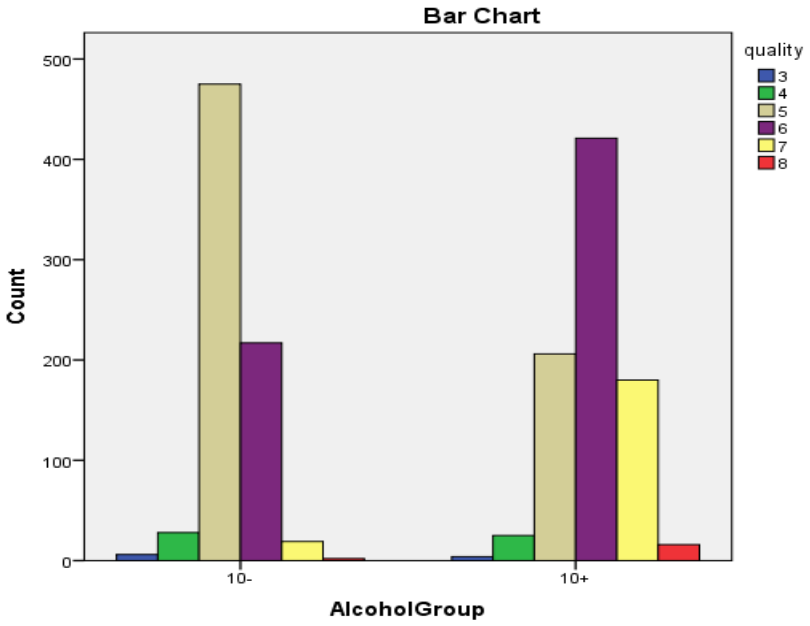


Figure 10: Alcohol variable split

T-TEST AND REASON FOR USING T-TEST:

T-test is used to test whether the means of two samples are significantly different from each other. T test is performed when Independent variables are categorical and dependent variables are continuous. Since, our independent variables in the dataset are continuous, we had to convert them to categories to perform this test. By performing this test, we will be going to prove if there is any significant statistical difference between the mean of pH values less than 3.25 and pH values more than 3.25. And, if there is any significant statistical difference between the mean of alcohol value less than 10 and alcohol value more than 10.

T-Test for pH

Analysis will be conducted using two variables: pH and quality of the dataset of the project to test their means.

Wines having a pH value of 3.25- are grouped as 1 while Wines having a pH value of 3.25+ are grouped as 2.

Null Hypothesis: H_0: There is no significant statistical difference between the mean of the two groups
Alternative Hypothesis: Ha: There is a significant statistical difference between the mean of the two groups
Assuming level of significance to be 95%:
α=0.05

Group Statistics

	pH	N	Mean	Std. Deviation	Std. Error Mean
Quality	pH3.25-	524	5.72	.819	.036
	pH3.25+	1075	5.60	.800	.024

Table 11: Group statistics for pH

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Quality	Equal variances assumed	.227	.634	2.758	1597	.006	.1184	.0429	.0342	.2027
	Equal variances not assumed			2.736	1015.352	.006	.1184	.0433	.0335	.2034

Table 12: T-Test on pH variable

The resulting p-value of Levene's test is more than α=0.05
Significance value:
p-value = 0.634

Since p value 0.634 > 0.05, hence we fail to reject our null hypothesis and there is no significant difference between the mean of our groups.

Result: An independent T-test was used to perform the effectiveness of pH content on Quality of wine, t = 2.758, p = 0.634, but significant difference was found (pH with poor (bad quality) rated as 1 having pH value 3.25- and alcohol with rich (high quality) rated as 2 having pH value 3.25+).

One Way ANOVA and reason for using one way ANOVA:

One Way analysis of variance (also known as **one-way ANOVA**) is a technique used to compare means of three or more samples (using the F distribution). One way ANOVA test is performed when Independent variables are categorical and dependent variables are continuous. Since, our independent variables in the dataset are continuous, we had to convert them to categories to perform this test. By performing this test, we will be going to prove if there is any significant statistical interaction between the pH value and quality.

Null Hypothesis: H_0: There is no significant interaction between the two groups
Alternative Hypothesis: Ha: There is a significant interaction between the two groups
Assuming level of significance to be 95%:
α=0.05

Descriptives

Quality

		N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum	Between-Component Variance
						Lower Bound	Upper Bound			
pH=3.25-		524	5.716	.8188	.0358	5.645	5.786	3.0	8.0	
pH=3.25+		1075	5.597	.7996	.0244	5.549	5.645	3.0	8.0	
Total		1599	5.636	.8076	.0202	5.596	5.676	3.0	8.0	
Model	Fixed Effects			.8059	.0202	5.596	5.676			
	Random Effects				.0618	4.851	6.421			.0061

Table 13: Descriptives for pH group with Quality

Test of Homogeneity of Variances

Quality

Levene Statistic	df1	df2	Sig.
.227	1	1597	.634

Table 14: Levene's statistics for Quality

This test for homogeneity of variance which provides an F statistic and a significance value (p-value). Here, p=value is greater than 0.05, so our group variances can be treated as equal and we have not violated the assumption of homogeneity of variances and the tests hold statistical significance.

ANOVA

Quality

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	4.942	1	4.942	7.609	.006
Within Groups	1037.223	1597	.649		
Total	1042.165	1598			

Table 15: ANOVA for Quality with respect to pH

Significance value, p-value = 0.006

Result: Since p value 0.006 < 0.05, we reject the null hypothesis i.e. there is no significant interaction between the two groups.

T-Test for alcohol

Here, alcohol quality = 1 represents the alcohol values less than 10 and alcohol quality = 2 represents the alcohol values greater than 10.

Null Hypothesis: H_0: There is no significant statistical difference between the mean of the two groups
Alternative Hypothesis: Ha: There is a significant statistical difference between the mean of the two groups
Assuming level of significance to be 95%:
α=0.05

Group Statistics

	quality	N	Mean	Std. Deviation	Std. Error Mean
alcohol	1	744	9.9265	.75801	.02779
	2	855	10.8550	1.10611	.03783

Table 16: Group statistics on Alcohol variable

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
alcohol	Equal variances assumed	136.894	.000	-19.292	1597	.000	-.92855	.04813	-1.02296	-.83415
	Equal variances not assumed			-19.782	1516.753	.000	-.92855	.04694	-1.02062	-.83648

Table 17: T-Test on Alcohol variable

The resulting p-value of Levene's test is less than α=0.05

Significance value:
p-value = 0.00

Since p value 0.00 < 0.05, our group variances can be treated as equal.

Result: An independent T-test was used to perform the effectiveness of alcohol content on Quality of wine, t = -19.292, p = 0.00, but significant difference was found (wines with poor (bad quality) rated as 1 having pH value 10- and wines with rich (high quality) rated as 2 having pH value 10+). So alcohol does have a prominent effect on the quality of red wines

LIMITATIONS

1. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.).
2. Conditions of testing are unknown: Conditions like temperature, origin of the wine etc. are not included in the dataset which could affect the quality of the wine
3. Time period through which the tests were conducted are unknown: Whether the wines were tested during a small time period or during an elongated time period is not specified

FUTURE WORK AND RECOMMENDATIONS

1. Analysis can be conducted on the remaining chemicals in order to understand their effects on the quality of the wine
2. Analysis of the wines can be done in various environments to see if the change in chemical components would change the quality of the wines