Data Analysis of quality for 2 types of wines;

White and Red

Final Project

Course: MSBA-320 Tools for Business Analytics

School: Golden Gate University

Instructor: Dr. Siamak Zadeh

Student: Sadegh Bamohabbat Chafjiri

Introduction:

In Portuguese, “Vinho Verde” literally means “green wine” being metaphor for “young wine”(Vinho Verde, 2018). Vinho Verde is not a specific variant of grape, but it is a kind of documentation for different types of wine; red or white (Vinho Verde, 2018). According to UCI Machine Learning Repository (n.d.), the collected data is related to two sets of observations on Vinho Verde wine located in Portugal. The Datasets are belonged to two types of white and red wines. The datasets which we are working include 1599 samples for red wine and 4898 sample for white wine. Due to the market secret issues, some details about grape types, selling price are not available and datasets only includes chemical components. This project statistically investigates the impact of chemical factors on wine quality.

According to Adding Sulphur to wine. (n.d.), there is an unproved believe between the winemakers that Romans have the custom of using sulphur dioxide to preserve wines. This group of winemakers believe that it is impossible to make a good wine without sulphur dioxide. French in North Africa was one of the early pioneers in the industrial usage of Sulphur dioxide to control wine fermentation and stabilization around 100 years ago (Adding Sulphur to wine, n.d.). Based on this industrial process, majority of winemakers think sulphur dioxide maybe keep the quality of wine for long time (Adding sulphur to wine, n.d.).

However, it seems the myth of sulphur dioxide necessity for wine quality needs to be reconsidered. Interestingly, minority of winemakers specially those concern about consumers’ health have accepted risk of making wine naturally by ffermentation temperature(Adding Sulphur to wine, n.d.). This method works very well in the hot climate conditions and quickly become popular in other climates by simulating that natural temperature. However, in certain circonstances adding

Sulphur dioxide is the only option to make good flavor of wine. By the way, correcting this general view needs scientific evidence supported by statistical analysis.

The primary propose of quality assessment in this project is to distinguish the main chemical factors impacting wine quality and also to verify this believe about quality that if myth of necessity of sulphur dioxide is true or not.

In the section of descriptive statistics and exploratory data analysis, we will consider Quality as an independent (output) attribute apart from which kind of chemical factors a wine sample contains because wine quality was ranked by human tasters independently. In other words, in the first section quality is not necessarily dependent to chemical factors and Quality is only a measure of wine appreciation determined by tasters. In the section of Correlation, Analysis we will evaluate chemical factors impacting quality and will find correlation between quality and high related chemical factors.

What we are willing to add here is that in all plots which quality presented in X axis, we assume this attribute as classification measure which its scale determined by tasters independently, not as the dependent attribute resulted from chemical factors. That is also true when we present chemical factors in Y axis. Presenting chemical factors in Y axis of plots does not mean chemical factors are necessarily dependent variable of quality. It means quality was considered as an independent classification measure in X axis to categorize all chemical factors having the most impact on quality.

As a last point, a limitation of the current analysis is that the current data consists of samples collected from a specific Portugal region. A good dataset usually includes samples being taken from various regions of the world to have this potential to generalize results. However, Quality

assessment as an assurance process of production can make a competitive benefit and a good measure of product quality. We will perform multiple statistical analysis in the following steps:

- performing descriptive data analysis on all wine samples by use of histogram, bar charts, box plots.

- performing scatterplot matrix and correlation analysis between quality and chemical factors such as alcohol, density, pH, free sulfur dioxide and total sulfur dioxide to find a relationship between quality and chemical factors.

- Performing other methods of Correlation analysis on different distributions of red and white wines to clarify the strength of correlation between quality or density and other chemical factors.

- Performing Anova as a statistical technique to compare the means of more than two groups of factors and to find the statistical significance among the selected chemical factors.

- performing regression by use of backward selection method and R Squared Equation method.

Data Collection and Description :

In our data collection, number of red-wine instances is 1599 and number of white-wine instances is 4898. We merged both datasets and created a bigger dataset including one more column named “type” distinguishing red wine by “R” and white wine by “W”. Table 1 presented type, attributes and characteristic of data set. Data set includes 11 input attributes resulted by physicochemical tests and 1 output attribute of quality. For more details about the detail of datasets please consider article published by Cortez, Cerdeira, Almeida, Matos and Reis (2009).

Table 1. Data set type and characteristic (UCI Machine Learning Repository, n.d.)

Data Set Characteristics/Area :	Multivariate/Business	Number of Instances:	4898 (W) + 1599 (R)	Missing Values?	N/A
Attribute Characteristics:	Real	Number of Attributes:	12	Statistical tasks	Classification, Regression
input attributes	fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol,				
output attribute	quality (score between 0 and 10)				

This database can be considered as classification or regression task. It seems the presented class of qualities are not uniform and it seems wines are more in a mid-level rather than high or poor conditions which can be determined by outliers. As a primary task we present details of the dataset in Table 2 Based on observations. It can be seen that the quality levels mostly vary between 5 to 8:

Table 2. details of Attributes in dataset

```
## 'data.frame':    6497 obs. of  13 variables:
##  $ fixed.acidity       : num  7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...
##  $ volatile.acidity    : num  0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 ...
##  $ citric.acid         : num  0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34 0.43 ...
##  $ residual.sugar      : num  20.7 1.6 6.9 8.5 8.5 6.9 7 20.7 1.6 1.5 ...
##  $ chlorides           : num  0.045 0.049 0.05 0.058 0.058 0.05 0.045 0.045 0.049 0.044 ...
##  $ free.sulfur.dioxide : num  45 14 30 47 47 30 30 45 14 28 ...
##  $ total.sulfur.dioxide: num  170 132 97 186 186 97 136 170 132 129 ...
##  $ density             : num  1.001 0.994 0.995 0.996 0.996 ...
##  $ pH                  : num  3 3.3 3.26 3.19 3.19 3.26 3.18 3 3.3 3.22 ...
##  $ sulphates           : num  0.45 0.49 0.44 0.4 0.4 0.44 0.47 0.45 0.49 0.45 ...
##  $ alcohol             : num  8.8 9.5 10.1 9.9 9.9 10.1 9.6 8.8 9.5 11 ...
##  $ quality             : int  6 6 6 6 6 6 6 6 6 6 ...
##  $ Type                : Factor w/ 2 levels "R","W": 2 2 2 2 2 2 2 2 2 2 …
```

Before starting analysis, we should show that our data set has no missing value. Table 3 confirmed our claim that all attributes completely filled out with valid values and we can start explanatory data analysis.

Table 3. Number of missing values

```
##       fixed.acidity    volatile.acidity         citric.acid
##                   0                   0                   0
##      residual.sugar           chlorides free.sulfur.dioxide
##                   0                   0                   0
## total.sulfur.dioxide             density                  pH
##                   0                   0                   0
##           sulphates             alcohol             quality
##                   0                   0                   0
##                Type
##                   0
```

Descriptive Statistics – Exploratory Data Analysis:

Summary of data set in Table 4 shows all variables with their statistical characteristics such as minimum and maximum values, median and mean and 1st and 3rd quartile values. Based on available characteristics, total sulfur dioxide has the biggest range from 0 to 440 while density varies in the smallest interval between 0.9871 to 1.0390.

Table 4. Summary

```
##  fixed.acidity      volatile.acidity   citric.acid      residual.sugar
##  Min.   : 3.800    Min.   :0.0800    Min.   :0.0000   Min.   : 0.600
##  1st Qu.: 6.400    1st Qu.:0.2300    1st Qu.:0.2500   1st Qu.: 1.800
##  Median : 7.000    Median :0.2900    Median :0.3100   Median : 3.000
##  Mean   : 7.215    Mean   :0.3397    Mean   :0.3186   Mean   : 5.443
##  3rd Qu.: 7.700    3rd Qu.:0.4000    3rd Qu.:0.3900   3rd Qu.: 8.100
##  Max.   :15.900    Max.   :1.5800    Max.   :1.6600   Max.   :65.800
##   chlorides        free.sulfur.dioxide total.sulfur.dioxide
##  Min.   :0.00900   Min.   :  1.00     Min.   :  6.0
##  1st Qu.:0.03800   1st Qu.: 17.00     1st Qu.: 77.0
##  Median :0.04700   Median : 29.00     Median :118.0
##  Mean   :0.05603   Mean   : 30.53     Mean   :115.7
##  3rd Qu.:0.06500   3rd Qu.: 41.00     3rd Qu.:156.0
##  Max.   :0.61100   Max.   :289.00     Max.   :440.0
##   density           pH              sulphates       alcohol
##  Min.   :0.9871    Min.   :2.720     Min.   :0.2200   Min.   : 8.00
##  1st Qu.:0.9923    1st Qu.:3.110     1st Qu.:0.4300   1st Qu.: 9.50
##  Median :0.9949    Median :3.210     Median :0.5100   Median :10.30
##  Mean   :0.9947    Mean   :3.219     Mean   :0.5313   Mean   :10.49
##  3rd Qu.:0.9970    3rd Qu.:3.320     3rd Qu.:0.6000   3rd Qu.:11.30
##  Max.   :1.0390    Max.   :4.010     Max.   :2.0000   Max.   :14.90
##   quality          Type
##  Min.   :3.000     R:1599
##  1st Qu.:5.000     W:4898
##  Median :6.000
##  Mean   :5.818
##  3rd Qu.:6.000
##  Max.   :9.000
```

In addition, the alcohol range of variation is between 8 to 14.9. The quality and pH levels respectively vary from 3 to 9 and 2.72 to 4.01.

In Fig 1 we can observe quality distribution of 2 types wine separately. The graph with red color is for red wine instances and the graph with green color is for white ones. The quality range is classified in 6 levels from 3 to 9 for white wine and 3 to 8 for red wine. It can be seen that both types of wine likely have normal distributions. The peak of distribution for white wine is on 6 while red wine has a distribution peak on 5.

Moreover, it can be seen that the most of instances in both white and red wine have a mid-level quality and the number of excellent quality in white wine is significantly more than red one.

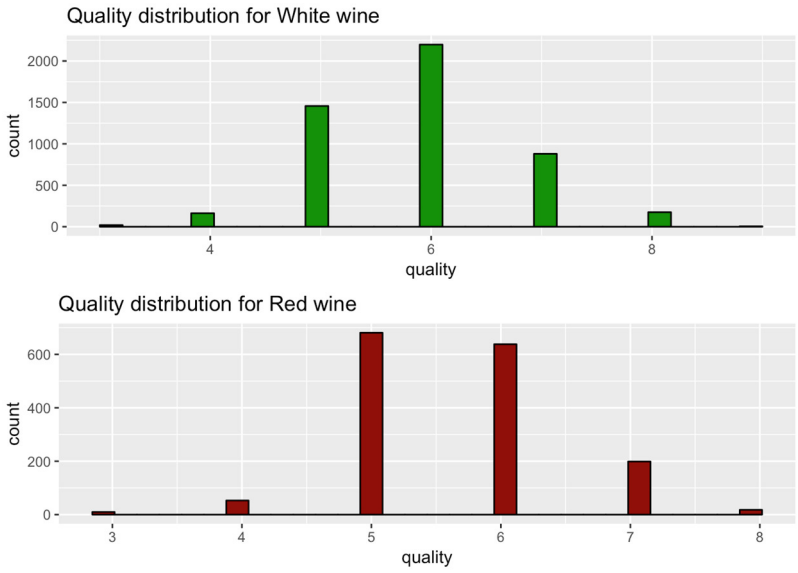


Fig 1. distribution of quality in classified levels [0 to 10]

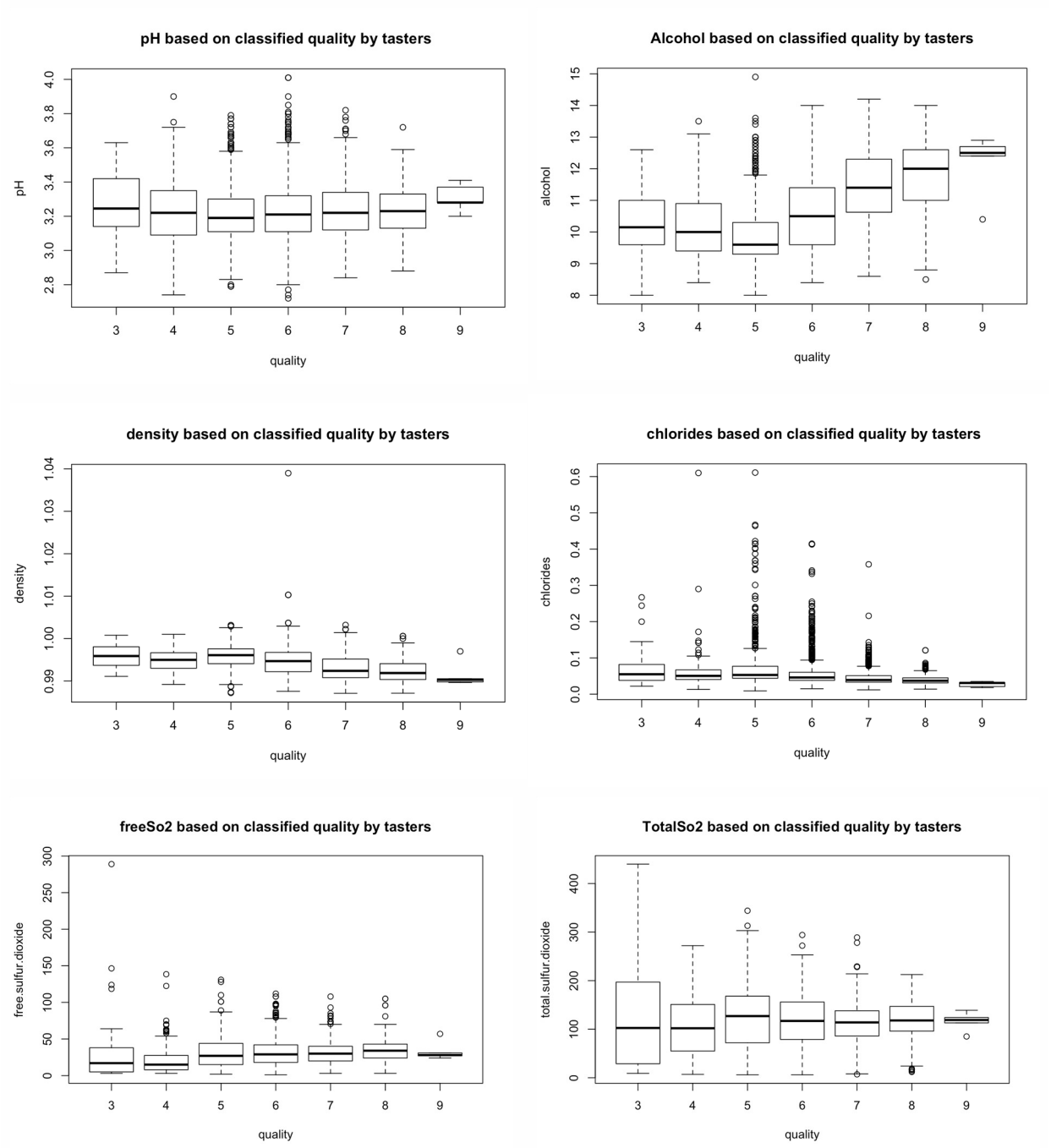


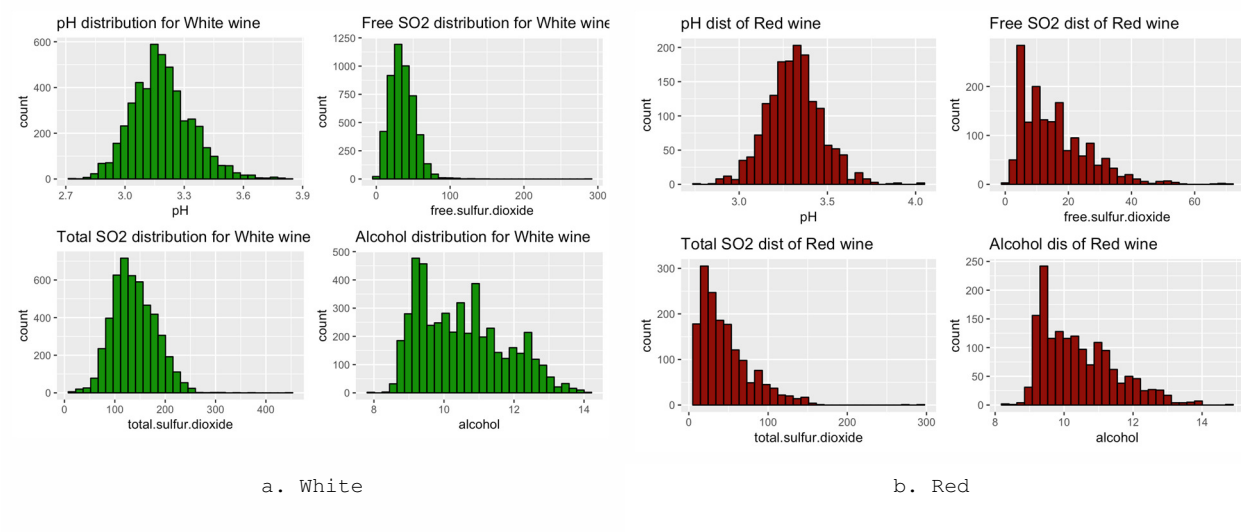
Fig 2. Box plot for different chemical factors (such as pH, alcohol, density, free sulfur dioxide and total sulfur dioxide, chlorides) based on different levels of quality

Now that a classified range of quality was presented, we would like to visualize different chemical factors such as pH, alcohol, density, free sulfur dioxide and total sulfur dioxide on this classified range by box plot. In Fig 2, Box plots depict that the mean of alcohol significantly increase by a rise in the level of quality while density and chlorides reversely experience a substantial decrease in mean. Interestingly, this phenomenon of density is supported by physics (table 5). When level of quality increases from 3 to 9 we are expecting that the percentage of alcohol will increase (to 13 for level 9 of quality) and consequently density of wine (0.99 for level 9 of quality) will become similar to density of liquors (see table 5). Last but not least, total sulfur dioxide and free sulfur dioxide and pH either with a fluctuation or without that experienced an gradual increase in mean for higher ranks of quality.

Table 5. density of some liquids less than 1

Water	1.00
Tuaca	0.98
Southern Comfort	0.97
Vodka (40%)	0.916
ethanol (pure Alcohol)	0.789

In the following, Fig 3 illustrates the distribution analysis on key factors such as pH, alcohol, density, free sulfur dioxide and total sulfur dioxide in both red and white variants of wine.



a. White b. Red

Fig 3. distribution analysis on key factors

In fig 3.a pH has a normal distribution with the most amounts of counts between 3 and 3.5 and a peak on 3.3. Free sulfur dioxide is distributed normally between 10 to 90 and this normal distribution includes a maximum amount in 50. The total sulfur dioxide has a right skewed distribution between 0 to 300 with a peak in 120. Alcohol has a two skewed pattern distributed between 8 to 14 with two peaks in 9 (bigger) and 11 (shorter).

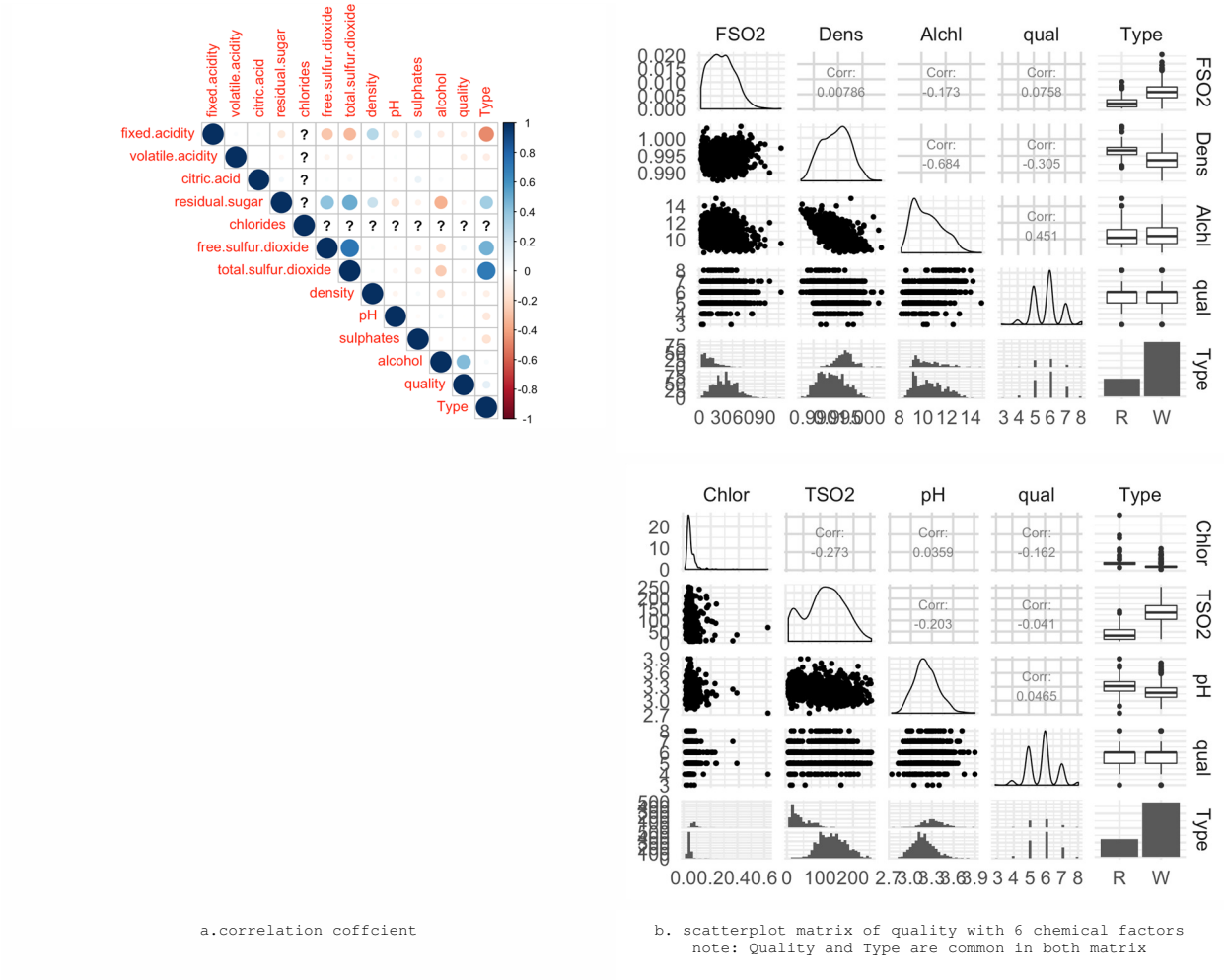
In Fig 3.b pH has a normal spread with a peak on 3.3 and higher levels of counts between 3.1 and 3.5. Free sulfur dioxide distributed in a right skewed pattern between 0 to 90 with a peak in 5 and also total sulfur dioxide has a similar distribution but it has spread between 0 to 300 with a peak in 20. Alcohol experienced a right skewed pattern between 8 to 14 with a major count in 9.

Correlation Analysis:

What I did in the previous section as a descriptive and exploratory analysis were unable to distinguish the main factors impacting quality. In this section, we will perform the correlation analysis to determine the relationship of quality ranked tasters with chemical factors analyzed by physiochemical lab. It was valuable to mention that from bar charts, we saw that alcohol and density had a significant linear change in means with a rise in quality level. In Fig 4a It seems the quality is highly correlated with alcohol and there is a positive correlation (correlation coefficient ~ 0.7) between them. Furthermore, there is a lightly reversed correlation between quality and chlorides (more than -0.1) and free sulfur dioxide (~ -0.07). Moreover, a lightly reversed linear relationship can be seen between Quality and density and total Sulphur dioxide. In other words, Quality is reversely correlated with density and total Sulphur dioxide. Moreover, there is no association between Quality and pH, citric acid and sulphates (correlation coefficient: 0).

In addition, in Fig 4b, we presented scatterplot matrix of some typical factors to evaluate their overall correlation characteristics. scatterplot matrix determined that Quality has a positive correlation with Alcohol(0.451), free sulfur dioxide (0.0758) and pH (0.0465) and reversed correlation with density (-0.305), chloride (-0.162), Total sulfur dioxide (-0.041). As a concluded result of both charts, it is clear that quality has the most correlation with alcohol, density and Chloride.

In the next step, we will investigate relationship between density and some important chemical factors in different levels of quality (3 to 9) for both red and white wines. The red color is for red wine instances and green color is for white ones.



a.correlation coffcient

b. scatterplot matrix of quality with 6 chemical factors
note: Quality and Type are common in both matrix

Fig 4. correlation analysis

Moreover, we selected density as a dependent in measure on Y axis for this 3 reasons; Firstly, it showed a semi linear behavior in mean in box plot (see Fig 2). Secondly, it had a lightly strong correlation with quality. Correlation coefficient is equal to -0.305 (see Fig 4b). Thirdly, it benefited from a short range variation between min and max (see table 4) which permit us to compare more plots to each other based on density. As you see in Fig 5, in the different levels of quality from 3 to 8 and different range of alcohol and pH, red wines have more density while by increasing the amount of sulfur dioxide (both free and total), white wine gets higher density. It means the red wine with less sulfur dioxide has higher level of density.

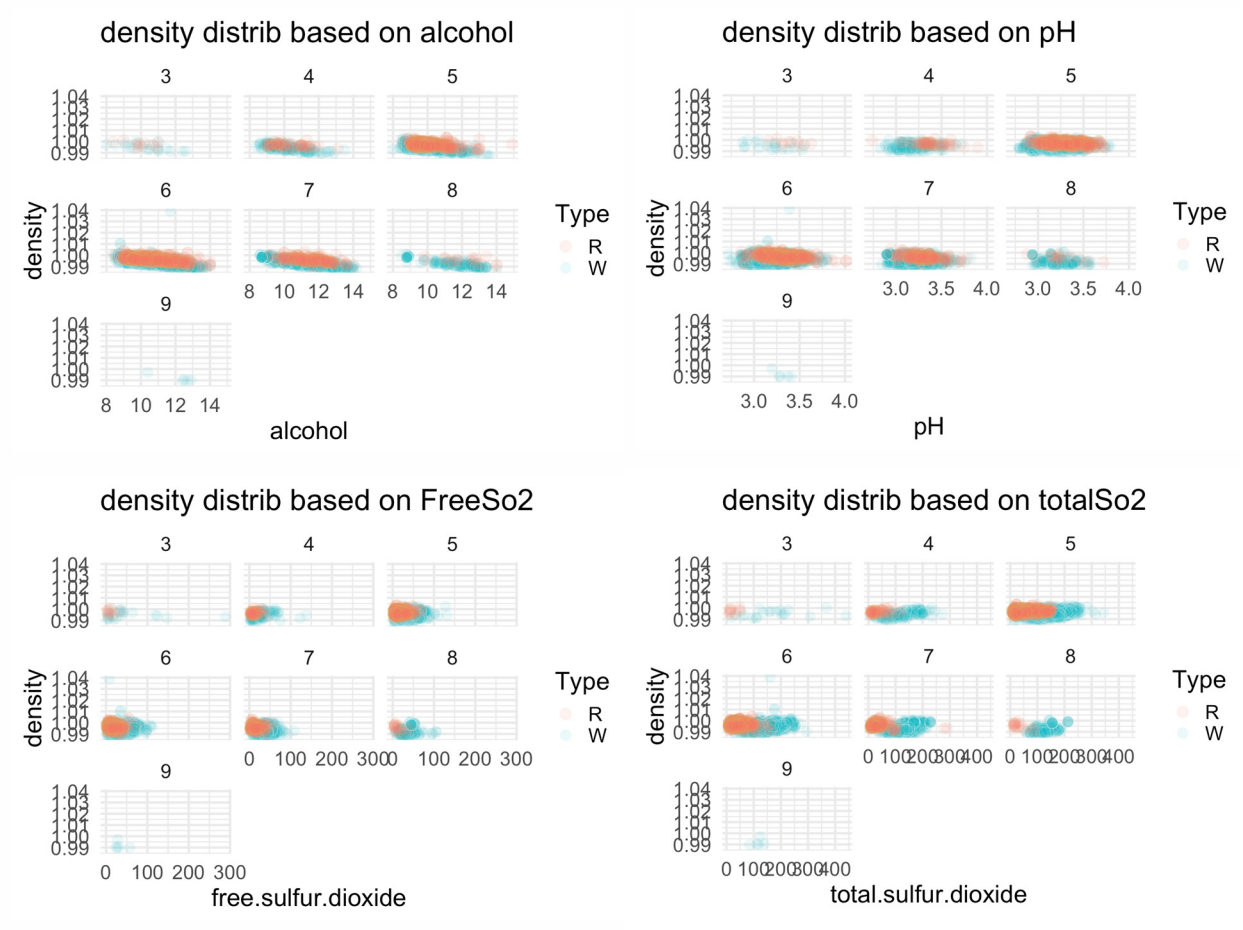


Fig 5. correlation analysis between density and other chemical factors

In the next step, we will show the quality relationship with important chemical factors for both red and white wines. The red color is for red wine instances and green color is for white ones. moreover, we will answer to the traditional question mentioned in introduction if sulfur dioxide impacting on quality or not and how much this impact is for both wine variants.

In the fig 6, we can consider density for higher qualities statistically tends to lower amounts, especially for white wine so that for a quality rank higher than 8, density converge to 0.99 asymptotically. There is a tendency to increase in alcohol to 14 asymptotically in higher ranks of quality for both red and white wines. Interestingly, pH has a likely uniform spread in highest rank

of quality (higher than 8) for both variants of wine. However, there is a common point of pH in 3.5 which both white and red wines benefited from level of 9 in quality. However, with regard to last graphs, it is clear that there is a considerable decrease in the amounts of chlorides wine, free and total sulfur dioxide in the higher levels of quality and reinforces this doubt about the positive effect which sulfur dioxide can exert on wine quality and.

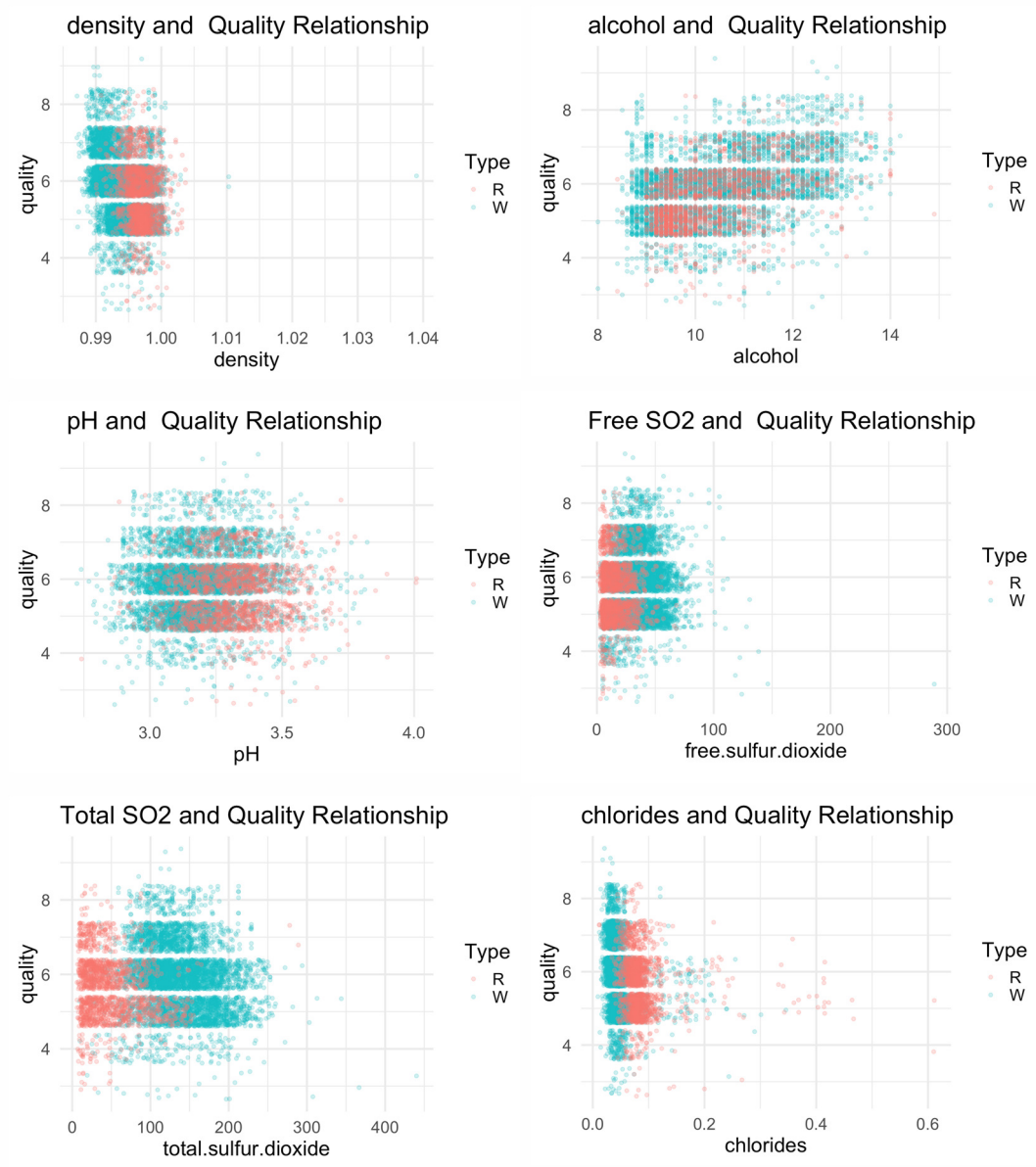


Fig 6. correlation analysis

In the plots presented in Fig 7, we will show the distribution of different chemical factors in different range of quality and density where the graph with a red color is for red wine and green color is for white wine. The last graphs show that in different levels of quality (3 to 8), free and total sulfur dioxide are significantly biased with a right skewed pattern while alcohol and pH have a spread pretty similar to normal distribution. As you can see, the peaks of correlation between density and different chemical factors for red wine are little bigger than white one in all ranges of quality vary from 3 to 8.

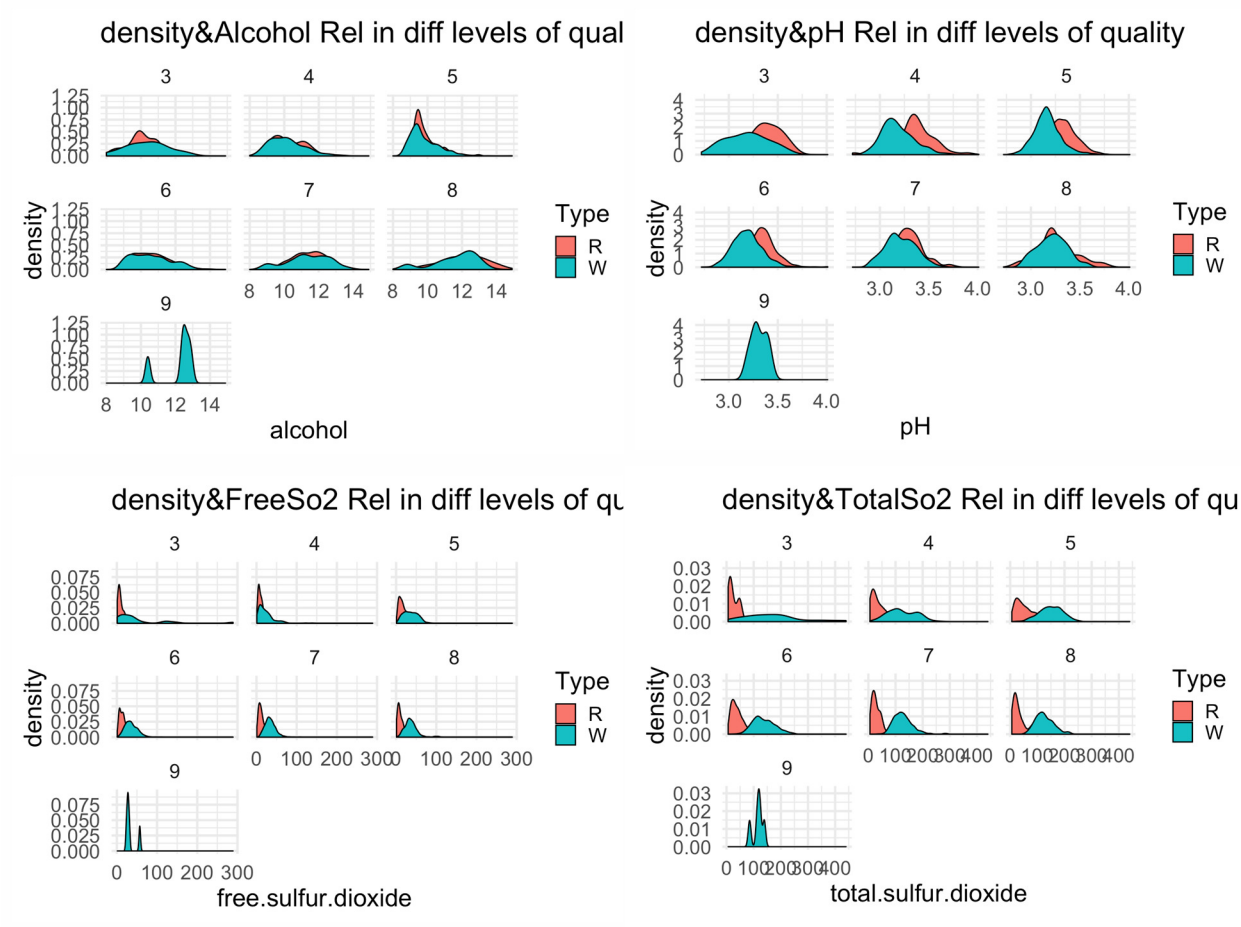


Fig 7. correlation analysis

Statistical tests:

In this section, we perform ANOVA (analysis of variance) to investigate mean difference between quality and chemical factors. We know the limitations of t-test that it can be performed on only two groups of data. Therefore, when we have more than two groups of data, it is suitable to apply one way ANOVA for comparing data groups by means and find statistical significance. In table 6, we perform ANOVA on majority of chemical factors. As you know, in data analytics 0.05 is a threshold value for statistical significance. If the p-value is less than 0.05, null hypothesis will be rejected and it means there is a statistical significance between assumed groups, otherwise, there is no statistical significance those groups.

Table 6. analysis of variance

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
alcohol	1	978	978.0	1781.020	< 2e-16	***
pH	1	6	5.9	10.814	0.00101	**
volatile.acidity	1	307	307.1	559.287	< 2e-16	***
residual.sugar	1	36	36.3	66.046	5.23e-16	***
free.sulfur.dioxide	1	3	3.2	5.822	0.01585	*
sulphates	1	60	59.7	108.638	< 2e-16	***
chlorides	1	1	0.9	1.704	0.19177	
density	1	0	0.1	0.192	0.66139	
Residuals	6488	3563	0.5			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Firstly, we will investigate statistical significance between quality and alcohol.

Null Hypothesis: There is no difference between mean of quality and alcohol

Alternative Hypothesis: There is difference between mean of quality and alcohol

After performing ANOVA we found that p-value is equal to 2e-16 which is less than 0.05. It means independent variable of alcohol is statistically significant with the quality of wine. So, the null hypothesis is rejected.

In the next step, we will investigate statistical significance between quality and pH

Null Hypothesis: There is no difference between mean of quality and pH

Alternative Hypothesis: There is difference between mean of quality and pH

After performing ANOVA It could be seen that p-value is equal to 0.00101 which is less than 0.05. It means pH as a independent variable was statistically significant with the wine quality. So, the null hypothesis is rejected.

Then, statistical significance between quality and density with a p-value equal to 0.66139 is investigated.

Null Hypothesis: There is no difference between mean of quality and density

Alternative Hypothesis: There is difference between mean of quality and density

Due to result of performing ANOVA we saw that p-value is equal to 0.66139 which is more than 0.05. It means independent variable of density is not statistically significant with the quality of wine. So, we fail to reject null hypothesis.

As a next case, statistical significance between quality and chlorides with a p-value equal to 0.19177 is investigated.

Null Hypothesis: There is no difference between mean of quality and chlorides

Alternative Hypothesis: There is difference between mean of quality and chlorides

Due to result of performing ANOVA we found that p-value is equal to 0.19177 which is more than 0.05. It means independent variable of chlorides is not statistically significant with the quality of wine. So, we fail to reject null hypothesis.

Moreover, statistical significance between quality and free sulfur dioxide with a p-value equal to 0.01585 is investigated

Null Hypothesis: There is no difference between mean of quality and free sulfur dioxide

Alternative Hypothesis: There is difference between mean of quality and free sulfur dioxide

Due to result of performing ANOVA, it is clear that for a p-value equal to 0.01585 which is less than 0.05, independent variable of free sulfur dioxide is statistically significant with the quality of wine. So, the null hypothesis is rejected

Regression Analysis:

Regression analysis is performed to evaluate the effect of one or more variables. Regression can be done with one two variables and with more than two and correlated among variables is evaluated. In our case, we had multiple continuous variables we opted for multiple regression. The nature of our dependent variable is of discrete so the only way to make the model is through making multiple regression by factorizing them and to check for the model and the accuracy.

Regression model (see table 7) is checked through distribution of the elements and variance or distribution of the error (residuals).

Table 7. first step of regression analysis

```
Call:

lm(formula = quality ~ ., data = WR)

Residuals:

    Min      1Q  Median      3Q     Max

-3.7796 -0.4671 -0.0444  0.4561  3.0211

Coefficients:

                      Estimate Std. Error t value Pr(>|t|)

(Intercept)          1.048e+02  1.414e+01   7.411 1.42e-13 ***

fixed.acidity         8.507e-02  1.576e-02   5.396 7.05e-08 ***

volatile.acidity     -1.492e+00  8.135e-02 -18.345  < 2e-16 ***

citric.acid          -6.262e-02  7.972e-02  -0.786   0.4322

residual.sugar        6.244e-02  5.934e-03  10.522  < 2e-16 ***

chlorides            -7.573e-01  3.344e-01  -2.264   0.0236 *

free.sulfur.dioxide   4.937e-03  7.662e-04   6.443 1.25e-10 ***

total.sulfur.dioxide -1.403e-03  3.237e-04  -4.333 1.49e-05 ***

density              -1.039e+02  1.434e+01  -7.248 4.71e-13 ***

pH                    4.988e-01  9.058e-02   5.506 3.81e-08 ***

sulphates             7.217e-01  7.624e-02   9.466  < 2e-16 ***

alcohol               2.227e-01  1.807e-02  12.320  < 2e-16 ***

TypeW                -3.613e-01  5.675e-02  -6.367 2.06e-10 ***

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7331 on 6484 degrees of freedom

Multiple R-squared:  0.2965,       Adjusted R-squared:  0.2952

F-statistic: 227.8 on 12 and 6484 DF,  p-value: < 2.2e-16
```

In order to perform regression, we used R Statistical language and used linear model technique. First step in performing a regression is to make factors for each set of similar values in the dependent variable. Our dependent variable is Quality which has 5 sets of values in the range 4-8. Then we have merged all other continuous variables which are continuous in nature and contain

decimal values too. To analyze the effect of all the variables taking quality as dependent, we made a regression model and checked for the significance of each variable through p value provided automatically by the result. To check for the co linearity, there are multiple methods available such as VIF, AIC etc. Our selection and criteria was to go with backward selection method. With the help of backward selection method, we kept on removing one variable on each step by checking the p value. First we removed density variable and then some other variable.

Finally, by following this process (see table 8), we were left with seven variables which were found to be significant at p value < 0.05. They are alcohol, volatile acidity, residual sugar, free sulfur dioxide, sulphates and pH variables found to be significant at this point. The intercept value for the dependent variable was also significant.

Table 8. final step of regression analysis by backward selection method

```
Call:
lm(formula = quality ~ volatile.acidity + residual.sugar + total.sulfur.dioxide +
    sulphates + pH + alcohol, data = WR)
Residuals:
    Min      1Q  Median      3Q     Max
-3.4516 -0.4656 -0.0359  0.4620  3.0359
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          1.7821910  0.2192818   8.127 5.21e-16 ***
volatile.acidity    -1.4973949  0.0637025 -23.506  < 2e-16 ***
residual.sugar       0.0247746  0.0023469  10.557  < 2e-16 ***
total.sulfur.dioxide -0.0010085  0.0002074  -4.863 1.18e-06 ***
sulphates            0.6070991  0.0654805   9.271  < 2e-16 ***
pH                   0.2070334  0.0613033   3.377 0.000737 ***
alcohol              0.3371972  0.0084686  39.817  < 2e-16 ***
```

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7403 on 6490 degrees of freedom

Multiple R-squared: 0.2821, Adjusted R-squared: 0.2814

F-statistic: 425 on 6 and 6490 DF, p-value: < 2.2e-16

R Squared Equation:

*Quality=1.78219 + 0.33719172*alcohol - 1.4973949*volatile acidity+0.0247746*residual sugar - 0.0010085*total sulfur dioxide + 0.6070991*sulphates + 0.2070334*ph*

For each unit increase in alcohol there will be 1.78219 percent increase in the total quality of the wine. For each unit increase in volatile acidity there will be 1.4973949 percent decrease in the total quality of the wine. For each unit increase in residual sugar there will be 0.0247746 percent increase in the total quality of the wine. For each unit increase in total sulfur dioxide there will be 0.0010085 percent decrease in the total quality of the wine. For each unit increase in 0.6070991*sulphates, there will be 0.6070991 percent increase in the total quality of the wine. For each unit increase in pH, there will be 0.2070334percent increase in the total quality of the wine.

Final model:

After checking and go through the final R equation the r squared found to be 0.25. In order to make increase in the R squared value the error distribution in analyzed which was found to be normally distributed and then some outliers' values were checked with cook's distance method but removal of outliers values had not improved that much so the final model had R squared value 0.2571.

*Quality=1.78219+0.33719172*alcohol-1.4973949*volatile acidity+0.0247746*residual sugar-*

*0.0010085*total sulfar dioxide+0.6070991*sulphates+0.2070334*ph.*

Conclusion:

In this project firstly we performed descriptive data analysis on wine quality and selected chemical factors such as alcohol, pH, density, sulfur dioxide (free and total), chlorides and presented sort of graphs. We showed distribution of quality, pH, alcohol, sulfur dioxide (free and total) by histogram and noticed that quality, pH and alcohol had normal distribution where sulfur dioxide (free and total) showed a right skewed characteristic and also we depicted mean behavior of alcohol, pH, density, sulfur dioxide (free and total), chlorides by box plots. We found that by increasing the level of quality, level of alcohol will increase sharply to 13 and level of density and chlorides will go down. Then, we performed scatterplot matrix and correlation analysis to find strongly correlated chemical factors. Based on correlation coefficient alcohol was highly correlated and density was moderately strongly correlated and chlorides being in the third rank was only correlated with quality. We could investigate relationship between quality or density with some selected chemical factors. In the quality case, we included top rank correlated factors and also pH and sulfur dioxide (free and total). Moreover, we Performed ANOVA as a statistical technique to compare the means of chemical factors. We found that chlorides and density had the p-value more than 0.05 and they are not statistical significance. Finally, by performing backward selection method we could do regression. First we removed density variable and then some other variable to suggest our final model for quality.

References

Vinho Verde. (2018). Retrieved from https://en.wikipedia.org/wiki/Vinho_Verde

UCI Machine Learning Repository: Wine Quality Data Set. (n.d.). Retrieved from http://archive.ics.uci.edu/ml/datasets/Wine%20Quality

Cortez, P. & Cerdeira, A. & Almeida, F. & Matos, T. & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. Retrieved from https://www.sciencedirect.com/science/article/pii/S0167923609001377

Adding sulphur to wine. (n.d.). Retrieved from http://www.morethanorganic.com/sulphur-in-wine

Alcohol density chart - the most comprehensive list available. (2013, May 08). Retrieved from https://bartenderly.com/tips-tricks/alcohol-density-chart/

Appendix: The code used for data analysis

```{r}
library(ggplot2)
```
```{r}
library (gridExtra)
```
```{r}
library(GGally)
```
```{r}
#Reading Data set
WR <- read.csv('WR.csv')
head(WR)
```
```{r}
dim(WR)
```
```{r}
sapply(WR, function(x) { sum(is.na(x))})
```
```{r}
names(WR)
```
```{r}
str(WR)
```
```{r}
summary(WR)
```
```{r}
boxplot(pH~quality,data=WR, main="pH based on classified quality by tasters ",
        xlab="quality", ylab="pH")
boxplot(free.sulfur.dioxide~quality,data=WR, main="freeSo2 based on classified quality by tasters",
        xlab="quality", ylab="free.sulfur.dioxide")
boxplot(total.sulfur.dioxide~quality,data=WR, main="TotalSo2 based on classified quality by tasters",
        xlab="quality", ylab="total.sulfur.dioxide")
boxplot(alcohol~quality,data=WR, main="Alcohol based on classified quality by tasters",
        xlab="quality", ylab="alcohol")
boxplot(density~quality,data=WR, main="density based on classified quality by tasters",
        xlab="quality", ylab="density")
boxplot(chlorides~quality,data=WR, main=" chlorides based on classified quality by tasters",
        xlab="quality", ylab="chlorides")
```
```{r}
q1<-ggplot(aes(x=pH),
       data =  subset(WR,Type %in% c("W")))+
   geom_histogram(color =I('black'),fill = I('#099009'))+
  ggtitle('pH distribution for White wine')
q2<-ggplot(aes(x=free.sulfur.dioxide),
       data =  subset(WR,Type %in% c("W")))+
   geom_histogram(color =I('black'),fill = I('#099009'))+
  ggtitle('Free SO2 distribution for White wine')
q3<-ggplot(aes(x=total.sulfur.dioxide),
       data =  subset(WR,Type %in% c("W")))+
   geom_histogram(color =I('black'),fill = I('#099009'))+
  ggtitle('Total SO2 distribution for White wine')
q4<-ggplot(aes(x=alcohol),
       data =  subset(WR,Type %in% c("W")))+
```

```
   geom_histogram(color =I('black'),fill = I('#099009'))+
  ggtitle('Alcohol distribution for White wine')

grid.arrange(q1,q2,q3,q4,ncol=2)

```
```{r}
p1<-ggplot(aes(x=pH),
       data =  subset(WR,Type %in% c("R")))+
   geom_histogram(color =I('black'),fill = I('#900909'))+
  ggtitle('pH dist of Red wine')
p2<-ggplot(aes(x=free.sulfur.dioxide),
       data =  subset(WR,Type %in% c("R")))+
   geom_histogram(color =I('black'),fill = I('#900909'))+
  ggtitle('Free SO2 dist of Red wine')
p3<-ggplot(aes(x=total.sulfur.dioxide),
       data =  subset(WR,Type %in% c("R")))+
   geom_histogram(color =I('black'),fill = I('#900909'))+
  ggtitle('Total SO2 dist of Red wine')
p4<-ggplot(aes(x=alcohol),
       data =  subset(WR,Type %in% c("R")))+
   geom_histogram(color =I('black'),fill = I('#900909'))+
  ggtitle('Alcohol dis of Red wine')

grid.arrange(p1,p2,p3,p4,ncol=2)
```


```{r}
r1<-ggplot(aes(x=quality),
       data =  subset(WR,Type %in% c("W")))+
   geom_histogram(color =I('black'),fill = I('#099009'))+
  ggtitle('Quality distribution for White wine')
r2<-ggplot(aes(x=quality),
       data =  subset(WR,Type %in% c("R")))+
   geom_histogram(color =I('black'),fill = I('#900909'))+
  ggtitle('Quality distribution for Red wine')

grid.arrange(r1,r2,ncol=1)
```
```{r}
#Running Scatterplot matrice 1
WR1 <- read.csv('WR1.csv')
head(WR1)
```
```{r}
library(GGally)
theme_set(theme_minimal(20))
set.seed(2200)
ggpairs(WR1[sample.int(nrow(WR1),1000),])
```
```{r}
#Running Scatterplot matrice 2
WR2 <- read.csv('WR2.csv')
head(WR2)
```
```{r}
library(GGally)
theme_set(theme_minimal(20))
set.seed(2200)
ggpairs(WR2[sample.int(nrow(WR2),1000),])
```

```{r}
ggplot(aes(x=alcohol),data =WR) +
 geom_density(aes(fill = Type))+
 facet_wrap(~quality)+
 ggtitle('density&Alcohol Rel in diff levels of quality')
```
```{r}
ggplot(aes(x=pH),data =WR) +
 geom_density(aes(fill = Type))+
 facet_wrap(~quality)+
 ggtitle('density&pH Rel in diff levels of quality')
```
```{r}
ggplot(aes(x=free.sulfur.dioxide),data =WR) +
 geom_density(aes(fill = Type))+
 facet_wrap(~quality)+
 ggtitle('density&FreeSo2 Rel in diff levels of quality ')
```
```{r}
ggplot(aes(x=total.sulfur.dioxide),data =WR) +
 geom_density(aes(fill = Type))+
 facet_wrap(~quality)+
 ggtitle('density&TotalSo2 Rel in diff levels of quality')
```

```{r}
ggplot(aes(x=alcohol,y=density),data = WR) +
  geom_jitter(aes(color = Type,bg = Type),alpha=1/10,,pch=21,cex=4)+
  facet_wrap(~quality)+
   scale_color_brewer(type = 'div')+
  ggtitle('density distrib based on alcohol')
```

```{r}
ggplot(aes(x=pH,y=density),data = WR) +
  geom_jitter(aes(color = Type,bg = Type),alpha=1/10,,pch=21,cex=4)+
  facet_wrap(~quality)+
   scale_color_brewer(type = 'div')+
  ggtitle('density distrib based on pH')
```

```{r}
ggplot(aes(x=free.sulfur.dioxide,y=density),data = WR) +
  geom_jitter(aes(color = Type,bg = Type),alpha=1/10,,pch=21,cex=4)+
  facet_wrap(~quality)+
   scale_color_brewer(type = 'div')+
  ggtitle('density distrib based on FreeSo2')
```
```{r}
ggplot(aes(x=total.sulfur.dioxide,y=density),data = WR) +
  geom_jitter(aes(color = Type,bg = Type),alpha=1/10,,pch=21,cex=4)+
  facet_wrap(~quality)+
   scale_color_brewer(type = 'div')+
  ggtitle('density distrib based on totalSo2')
```
```{r}
a1<-ggplot(aes(x=total.sulfur.dioxide,y=quality),
       data = WR)+
   geom_density(aes(color=Type),stat='summary',fun.y=median)
 a2<-ggplot(aes(x=free.sulfur.dioxide,y=quality),
       data = WR)+
```

```
   geom_density(aes(color=Type),stat='summary',fun.y=median)
 a3<-ggplot(aes(x=chlorides,y=quality),
      data = WR)+
   geom_density(aes(color=Type),stat='summary',fun.y=median)


grid.arrange(a1,a2,a3,ncol=2)
```
```{r}

   a4<-ggplot(aes(x=density,y=quality),
      data = WR)+
   geom_density(aes(color=Type),stat='summary',fun.y=median)
   a5<-ggplot(aes(x=alcohol,y=quality),
      data = WR)+
   geom_density(aes(color=Type),stat='summary',fun.y=median)
    a6<-ggplot(aes(x=pH,y=quality),
      data = WR)+
   geom_density(aes(color=Type),stat='summary',fun.y=median)

  grid.arrange(a4,a5,a6,ncol=2)
```

```{r}
df1<- WR
for(i in 1:ncol(df1)){
  df1[,i]<-as.integer(df1[,i])
}
library(corrplot)
corrplot(cor(as.matrix(df1)),type = "upper")
cor(df1)
```
```{r}
ggplot(aes(x =density , y =quality ), data = WR) +
  geom_point(aes(color=Type),alpha=1/4, position = 'jitter')+
  ggtitle(' density and  Quality Relationship')
```
```{r}
ggplot(aes(x =alcohol , y =quality ), data = WR) +
  geom_point(aes(color=Type),alpha=1/4, position = 'jitter')+
  ggtitle(' alcohol and  Quality Relationship')
```
```{r}
ggplot(aes(x =pH , y =quality ), data = WR) +
  geom_point(aes(color=Type),alpha=1/4, position = 'jitter')+
  ggtitle(' pH and  Quality Relationship')
```
```{r}
ggplot(aes(x =free.sulfur.dioxide , y =quality ), data = WR) +
  geom_point(aes(color=Type),alpha=1/4, position = 'jitter')+
  ggtitle(' Free SO2 and  Quality Relationship')
```
```{r}
ggplot(aes(x =total.sulfur.dioxide , y = quality), data = WR) +
  geom_point(aes(color=Type),alpha=1/4, position = 'jitter')+
  ggtitle('Total SO2 and Quality Relationship')
```

```{r}
ggplot(aes(x =chlorides , y = quality), data = WR) +
  geom_point(aes(color=Type),alpha=1/4, position = 'jitter')+
  ggtitle('chlorides and Quality Relationship')
```

```
```{r}
aov <- aov(quality ~ alcohol + pH + volatile.acidity + residual.sugar + free.sulfur.dioxide + sulphates + chlorides+ density , data=WR)
summary(aov)

```
```{r}
Q1 <- lm(quality ~ ., data=WR)
```
```{r}
summary(Q1)
```
```{r}
Q2 <- lm(quality ~ fixed.acidity + volatile.acidity + residual.sugar + free.sulfur.dioxide+ total.sulfur.dioxide +density+ sulphates+ pH +alcohol, data=WR)
```

```{r}
# Linear Regression
summary(Q2)
```
```{r}
Q3 <- lm(quality ~ volatile.acidity + residual.sugar +  total.sulfur.dioxide +sulphates+ pH +alcohol, data=WR)
```

```{r}
# Linear Regression
summary(Q3)
```
```