HarvardX-Capstone Project: Red Wine Quality

Final

Jocelyn Carmen Dumlao

13/05/2021

Introduction

This report is part of the Data Science capstone project of the edx course HarvardX. The goal is to demonstrate that the student acquired skills with the R programming language in the field of Data Science to actually solve real-world problems. The aim of this project is to applying Machine Learning techniques that go beyond standard linear regression, to use exploratory data analysis (EDA) techniques to explore relationships in one variable to multiple variables and to explore selected dataset for visualization, distributions, outliers and anomalies.

I explore some new data in the UCI Machine Learning Repository and Kaggle.com. Through this website, I must select and choose my preferred project topics for the final submissions. In the end, I chose a **Red Wine Quality dataset** as my final capstone project.

Wine is an alcoholic drink typically made from fermented grape juice. Yeast consumes the sugar in the grape and converts it to ethanol, carbon dioxide, and heat. Different varieties of grapes and strains of yeasts produce different style of wine.

Executive Summary

The main aims of this project on Red Wine Quality Dataset is to predict which of the physicochemical properties of substance in making a good win, namely: (1) Fixed Acidity, (2) Volatile Acidity, (3) Citric Acid, (4) Residual Sugar, (5) Chlorides, (6) Free Sulfur Dioxide, (7) Total Sulfur Dioxide, (8) Density, (9) pH, (10) Sulphates, (11) Alcohol, with 11 variables and 1 output variable given. The conclusion is also presented.

Methodology

At first, I searched a kaggle.com website as required. Then I selected and chose my own dataset, which is **"Red Wine Quality Dataset"** for the final project - "HarvardX-capstone project." There some steps were performed to complete the project: (1) To downloaded a Red Wine Quality Dataset, (2) Exploring by investigating Red Wine Quality Analysis, (3) To determine the association between quality and chemical properties such as: (a)Fixed Acidity, (b) Volatile Acidity, (c) Citric Acid, (d) Residual Sugar, (e) Chlorides, (f) Free Sulfur Dioxide, (g) Total Sulfur, (h) Density, (i) pH, (j) Sulphates, (k) Alcohol, (4) Adding label as alcohol percentage, (5) Adding rating as Red Wine Quality, (6) Regression Analysis & Correlation Coefficient plot, (7) Presentation of different parameters, (8) By using Univariate Linear Regression -examining the effects of a singular variable on a set of data, (9) By using Bivariate Linear Regression - examining the effects of two variables on a set of data, (10) By using Multivariate Linear Regression - examining the effects of more than two variables, (12) to show the results and conclusion.

References:

https://www.kaggle.com
https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009
https://archive.ics.uci.edu/ml/dataset/wine+quality

A. Dataset

A.1 Download Dataset

```
# Note: this process could take a couple of minutes
if(!require (tidyverse)) install.packages("tidyverse", repos = "http://cran.us.r-project.org")
if(!require(caret)) install.packages("caret", repos = "http://cran.us.r-project.org")
if(!require(data.table)) install.packages("data.table", repos = "http://cran.us.r-project")
```

```
library(readr)
wineQualityReds <- read_csv("archive/wineQualityReds.csv")

library(tidyverse)
library(caret)
library(data.table)
library(corrplot)
library(kableExtra)
library(lubridate)
library(RColorBrewer)
library(ggthemes)
library(readr)
library(readxl)
library(purrr)
library(tibble)
library(ggplot2)
library(psych)
library(car)
library(memisc)
library(pseudo)
library(KMsurv)
library(geepack)
```

B. Exploratory Data Analysis

To get familiar with the Red Wine Quality dataset, with each physicochemical properties namely: “Fixed Acidity”, “Volatile Acidity”, “Citric Acid”, “Residual Sugar”, “Chlorides”, “Free Sulfur Dioxide”, “Total Sulfur Dioxide”, “Density”, “pH”, “Sulphate”, “Alcohol” as shown below.

Physicochemical Properties

ISO4_1	Fixed Acidity: most acids involved with wine or fixed or nonvolatile(do not evaporate readily)(tartaric acid - g/dm^3)
ISO4_2	Volatile Acidity : the amount of acetic acid in wine,which at too high of levels can lead to an unpleasant, vinegar taste(acetic acid - g/dm^3)
ISO4_3	Citric Acid:found in small quantities,citric acid can add 'freshness' and flavor to wines(g/dm^3)
ISO4_4	Residual Sugar: the amount of sugar remaining after fermentation stops,It's rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet.(g/dm^3)
ISO4_5	Chlorides: the amount of salt in the wine(sodium chloride - g/dm^3)
ISO4_6	Free Sulfur Dioxide:the free form of SO2 exists in equilibrium between molecular SO2(as a dissolved gas)and bisulfite ion, It prevents microbial growth and the oxidation of wine(mg/dm^3)
ISO4_7	Total Sulfur Dioxide: amount of free and bound forms of SO2:In low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm,SO2 becomes evident in the nose and taste of wine.(mg/dm^3)
ISO4_8	Density: the density of water is close to that of water depending on the percent alcohol and sugar content(g/cm^3)
ISO4_9	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
ISO4_10	Sulphates: a wine additive which can contribute to sulfur dioxide gas(SO2) levels which acts as an antimicrobial and antioxidant(potassium sulphate -g/dm3)
ISO4_11	Alcohol: the percent alcohol content of the wine(% by volume)

```
dim(wineQualityReds)
#[1] 1599  16
```

Physicochemical Class

	x
X	integer
fixed.acidity	numeric
volatile.acidity	numeric
citric.acid	numeric
residual.sugar	numeric
chlorides	numeric
free.sulfur.dioxide	numeric
total.sulfur.dioxide	numeric
density	numeric
pH	numeric
sulphates	numeric
alcohol	numeric
quality	integer
label	factor
rating	factor
quality.rank	factor

A.1 Univariate Linear Regression

Univariate Linear Regression: focuses on determining relationship between on independent (explanatory variable) variable and one dependent variable. Regression comes handy mainly in situation where the relationship between two features is not obvious to the naked eye.

Univariate Descriptive Statistics

Some ways you can describe patterns found in univariate data include central tendency (mean, mods and median) and dispersion: range, variance, maximum, minimum, quartiles (including the interquartile range), and standard deviation.

Summary Descriptive Statistics Tibble

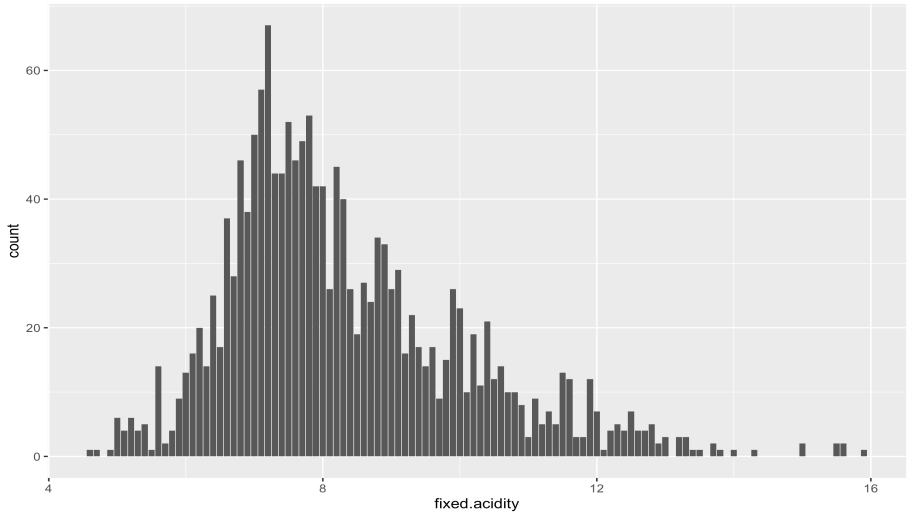
X	fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides	free.sulfur.dioxide	total.sulfur.dioxide	density	pH	sulphates	alcohol	quality	label
Min. : 1.0	Min. : 4.60	Min. :0.1200	Min. :0.000	Min. : 0.900	Min. :0.01200	Min. : 1.00	Min. : 6.00	Min. :0.9901	Min. :2.740	Min. :0.3300	Min. : 8.40	Min. :3.000	Light : 37
1st Qu.: 400.5	1st Qu.: 7.10	1st Qu.:0.3900	1st Qu.:0.090	1st Qu.: 1.900	1st Qu.:0.07000	1st Qu.: 7.00	1st Qu.: 22.00	1st Qu.:0.9956	1st Qu.:3.210	1st Qu.:0.5500	1st Qu.: 9.50	1st Qu.:5.000	Medium:1421
Median : 800.0	Median : 7.90	Median :0.5200	Median :0.260	Median : 2.200	Median :0.07900	Median :14.00	Median : 38.00	Median :0.9968	Median :3.310	Median :0.6200	Median :10.20	Median :6.000	Strong: 141
Mean : 800.0	Mean : 8.32	Mean :0.5278	Mean :0.271	Mean : 2.539	Mean :0.08747	Mean :15.87	Mean : 46.47	Mean :0.9967	Mean :3.311	Mean :0.6581	Mean :10.42	Mean :5.636	NA
3rd Qu.:1199.5	3rd Qu.: 9.20	3rd Qu.:0.6400	3rd Qu.:0.420	3rd Qu.: 2.600	3rd Qu.:0.09000	3rd Qu.:21.00	3rd Qu.: 62.00	3rd Qu.:0.9978	3rd Qu.:3.400	3rd Qu.:0.7300	3rd Qu.:11.10	3rd Qu.:6.000	NA
Max. :1599.0	Max. :15.90	Max. :1.5800	Max. :1.000	Max. :15.500	Max. :0.61100	Max. :72.00	Max. :289.00	Max. :1.0037	Max. :4.010	Max. :2.0000	Max. :14.90	Max. :8.000	NA

alcohol	quality	label	rating	quality.rank
Min. : 8.40	Min. :3.000	Light : 37	Poor:744	Low : 63
1st Qu.: 9.50	1st Qu.:5.000	Medium:1421	Good:837	Middle:1319
Median :10.20	Median :6.000	Strong: 141	NA's: 18	High : 217
Mean :10.42	Mean :5.636	NA	NA	NA
3rd Qu.:11.10	3rd Qu.:6.000	NA	NA	NA
Max. :14.90	Max. :8.000	NA	NA	NA

Descriptive Statistics Tibble

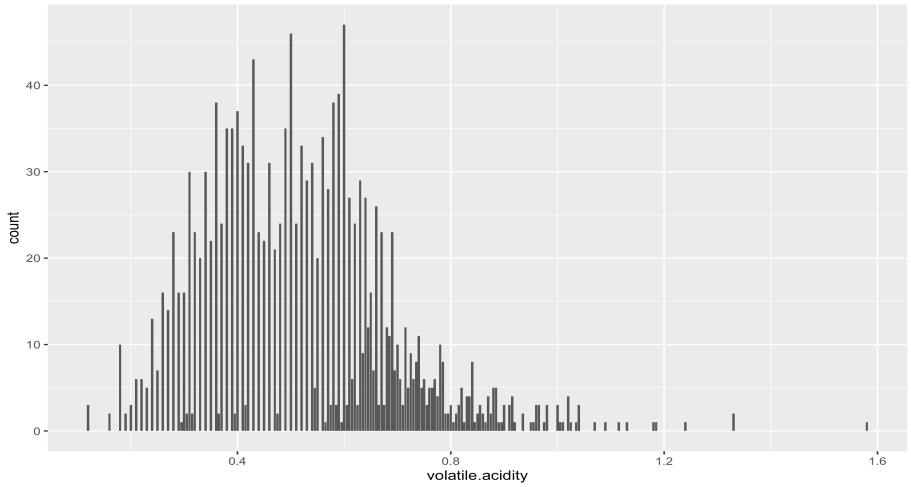
	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
X	1	1599	800.0000000	461.7358552	800.00000	800.0000000	593.0400000	1.00000	1599.00000	1.598e+03	0.0000000	-1.2022516	11.5470054
fixed.acidity	2	1599	8.3196373	1.7410963	7.90000	8.1525371	1.4826000	4.60000	15.90000	1.130e+01	0.9809084	1.1196987	0.0435410
volatile.acidity	3	1599	0.5278205	0.1790597	0.52000	0.5180679	0.1779120	0.12000	1.58000	1.460e+00	0.6703331	1.2126893	0.0044779
citric.acid	4	1599	0.2709756	0.1948011	0.26000	0.2612881	0.2520420	0.00000	1.00000	1.000e+00	0.3177403	-0.7930455	0.0048716
residual.sugar	5	1599	2.5388055	1.4099281	2.20000	2.2583528	0.4447800	0.90000	15.50000	1.460e+01	4.5321399	28.4850200	0.0352592
chlorides	6	1599	0.0874665	0.0470653	0.07900	0.0802350	0.0148260	0.01200	0.61100	5.990e-01	5.6696937	41.5259635	0.0011770
free.sulfur.dioxide	7	1599	15.8749218	10.4601570	14.00000	14.5772834	10.3782000	1.00000	72.00000	7.100e+01	1.2482220	2.0072212	0.2615857
total.sulfur.dioxide	8	1599	46.4677924	32.8953245	38.00000	41.8430913	26.6868000	6.00000	289.00000	2.830e+02	1.5126890	3.7856764	0.8226402
density	9	1599	0.9967467	0.0018873	0.99675	0.9967362	0.0016753	0.99007	1.00369	1.362e-02	0.0711540	0.9225000	0.0000472
pH	10	1599	3.3111132	0.1543865	3.31000	3.3090945	0.1482600	2.74000	4.01000	1.270e+00	0.1933203	0.7959191	0.0038609
sulphates	11	1599	0.6581488	0.1695070	0.62000	0.6374473	0.1186080	0.33000	2.00000	1.670e+00	2.4241176	11.6615285	0.0042390
alcohol	12	1599	10.4229831	1.0656676	10.20000	10.3100312	1.0378200	8.40000	14.90000	6.500e+00	0.8592144	0.1916586	0.0266500
quality	13	1599	5.6360225	0.8075694	6.00000	5.5886027	1.4826000	3.00000	8.00000	5.000e+00	0.2173931	0.2879148	0.0201955
label*	14	1599	2.0650407	0.3273474	2.00000	2.0000000	0.0000000	1.00000	3.00000	2.000e+00	1.2506663	5.4624822	0.0081862
rating*	15	1581	1.5294118	0.4992921	2.00000	1.5367589	0.0000000	1.00000	2.00000	1.000e+00	-0.1177393	-1.9873933	0.0125571
quality.rank*	16	1599	2.0963102	0.4073543	2.00000	2.0452771	0.0000000	1.00000	3.00000	2.000e+00	0.7027462	2.3565559	0.0101870

Fixed Acidity Rate



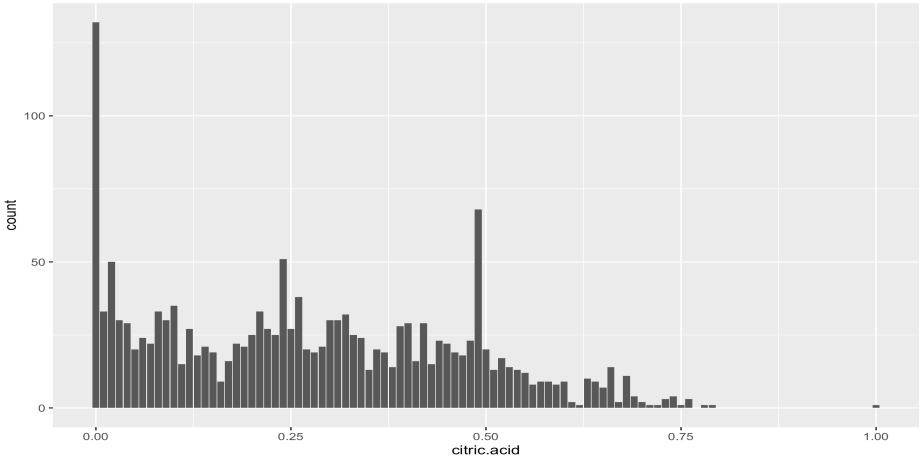
Fixed Acidity Rate: most acids involved with wine or fixed or nonvolatile (do not evaporate readily) (tartaric acid – g/dm^3) As we can see the plot there is one main peak in the plot. Fixed acidity having a range of [4.6,15.9], Median Score 50% of 7.9 and Mean Value of 8.3.

Volatile Acidity Rate



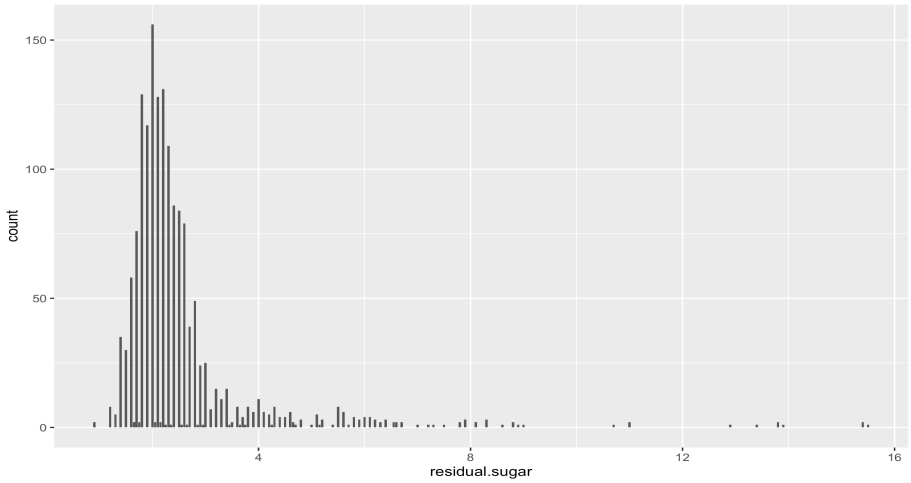
Volatile Acidity: the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste (acetic acid - g/dm^3. This attribute has Mean Value of 0.5, Median Score 50% of 0.52 and having range of [0.12,1.58].

Citric Acid Rate



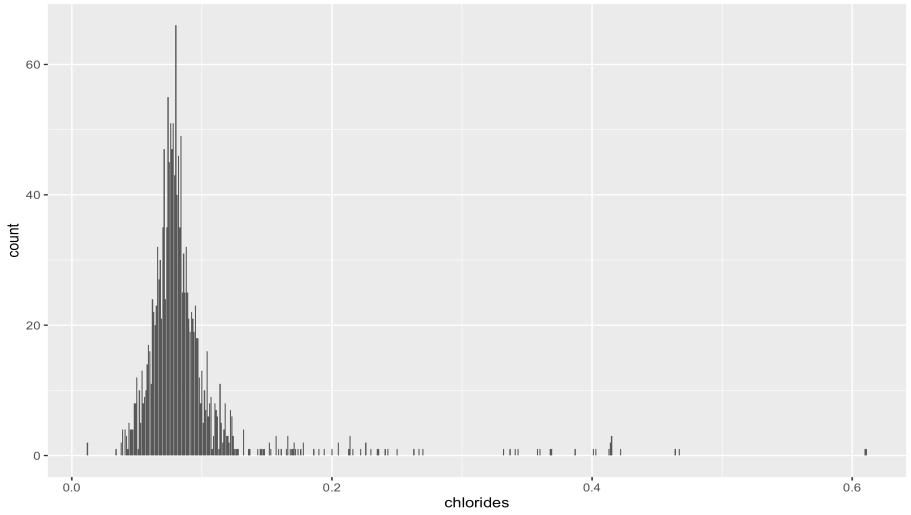
Citric Acid: found in small quantities, citric acid can add 'freshness' and flavor to wines(g/dm^3). There are two main peaks in the plot. First is between 0 and second one is 0.5 in range of [0,0.5], Mean Value of 0.2 and Main Score 50% of 0.2.

Residual Sugar Rate



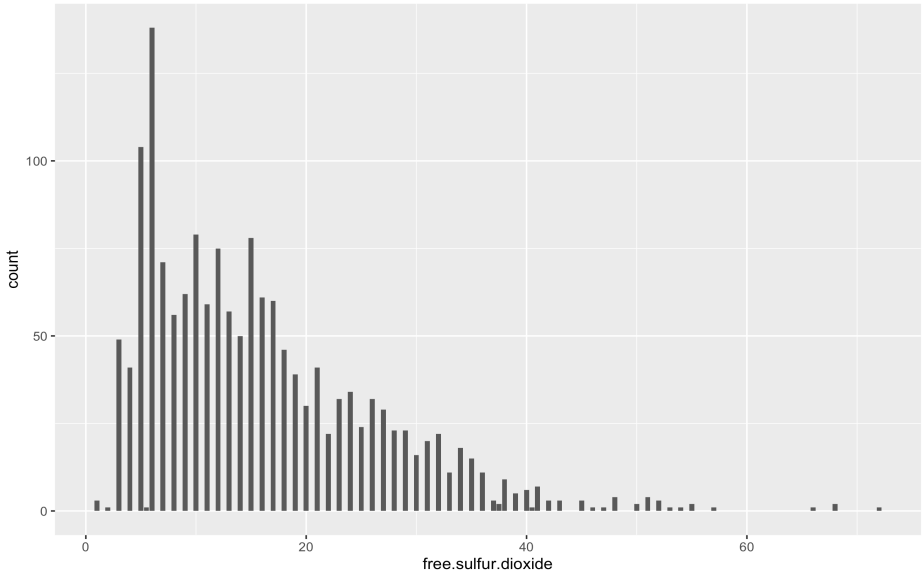
Residual Sugar: the amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet. (g/dm^3). The plot has a very long tail the third quartile of 2.6, showing that 75% of wine have a residual sugar value below 2.6 g/dm3. The Mean Value of 2.5 and Medium Score 50% of 2.2.

Chlorides Rate



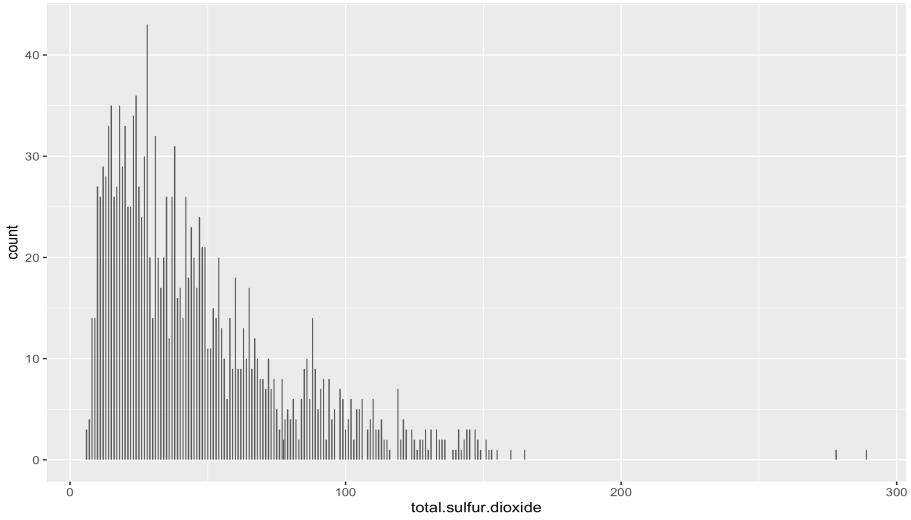
Chlorides - the amount of salt in the wine (sodium chloride - g/dm^3). It has a Mean Value of 0.08, Median Score 50% of 0.07 and range to [0.0,0.6]. This plot has a bell shape.

Free Sulfur Dioxide Rate



Free Sulfur Dioxide: the free form of SO2 exists in equilibrium between molecular SO2(as a dissolved gas) and bisulfite ion. It prevents microbial growth and the oxidation of wine(mg/dm^3). It has range of [1.0,72], Mean Value of 15.8, Median Score 50% of 14 and third quantile has 21 as shown in graph.

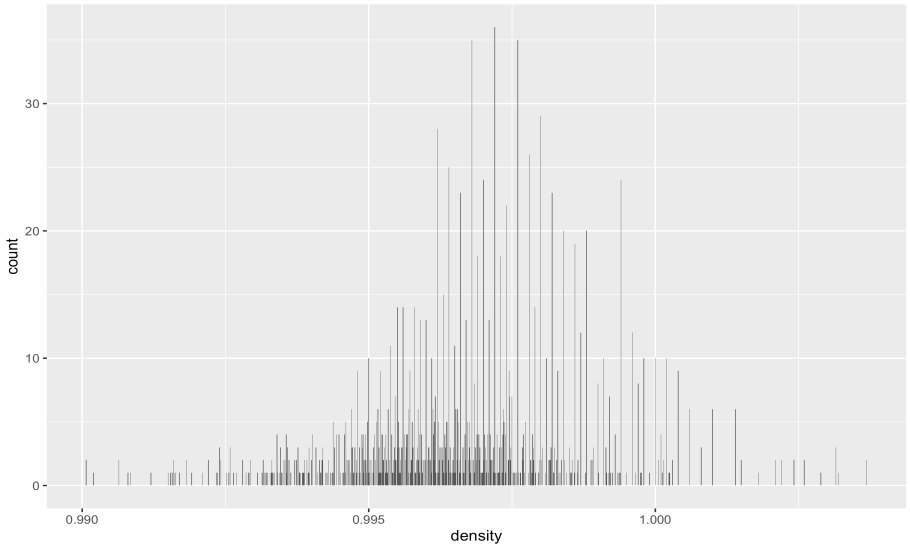
Total Sulfur Dioxide Rate



Total Sulfur Dioxide: amount of free and bound forms of SO2: In low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes

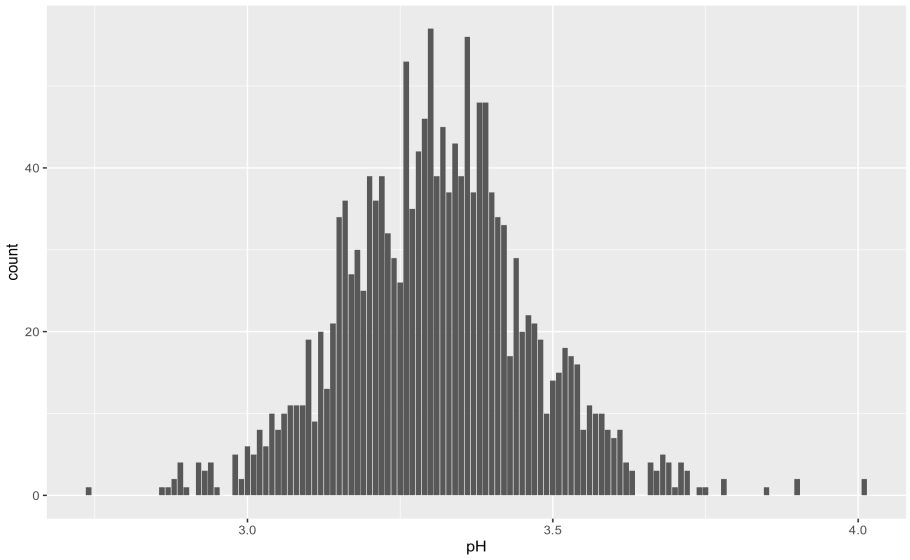
evident in the nose and taste of wine. (mg/dm^3). It has 75% of the in this dataset has a sulfur dioxide value of below 62 mg/dm^3, Mean Value of 46, Median Score 50% of 38, range of [6,289]. There is one main peak in the plot.

Density Rate



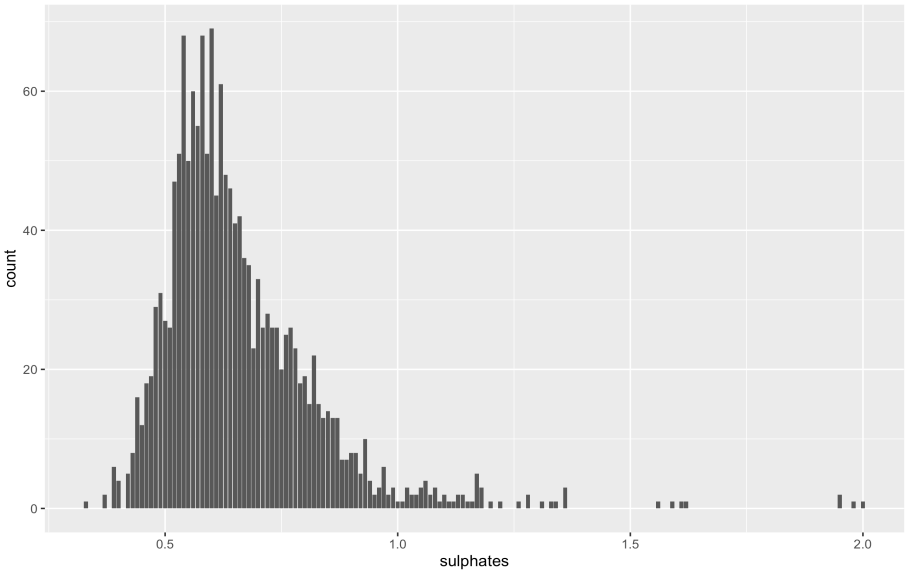
Density: the density of water is close to that of water depending on the percent alcohol and sugar content(g/cm^3). It has a sugar content of 0.9 g/cm^3, range of [0.9,1.0], Mean Value of 0.9 and Median Score 50% of 0.9.

pH Rate



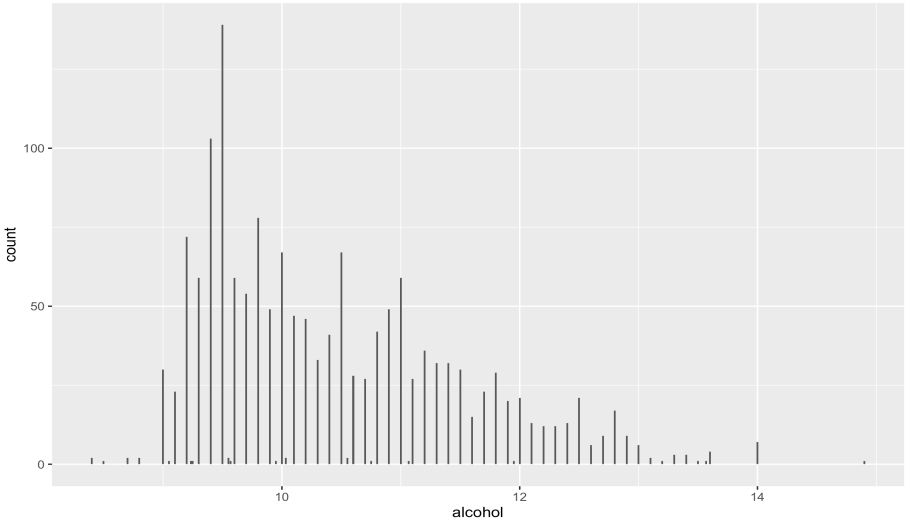
pH: describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14(very basic); most wines are between 3-4 on the pH scale. As we can see the plot it form like a bell shape the Mean Value and Median Score nearly the same 50% of 3.31. The pH level of 3.31 at most wine has. It has a value of 2.7 and range of [2.7,4] and third quantile of 3.4.

Sulphates Rate



Sulphates: a wine additive which can contribute to sulfur dioxide gas (SO2) levels which acts as an antimicrobial and antioxidant (potassium sulphate -g/dm3). It has range of [0.3,2.0], it gives a Mean Value of 0.6 and Median Score 50% of 0.6, the third quantile of 0.7.

Alcohol Rate

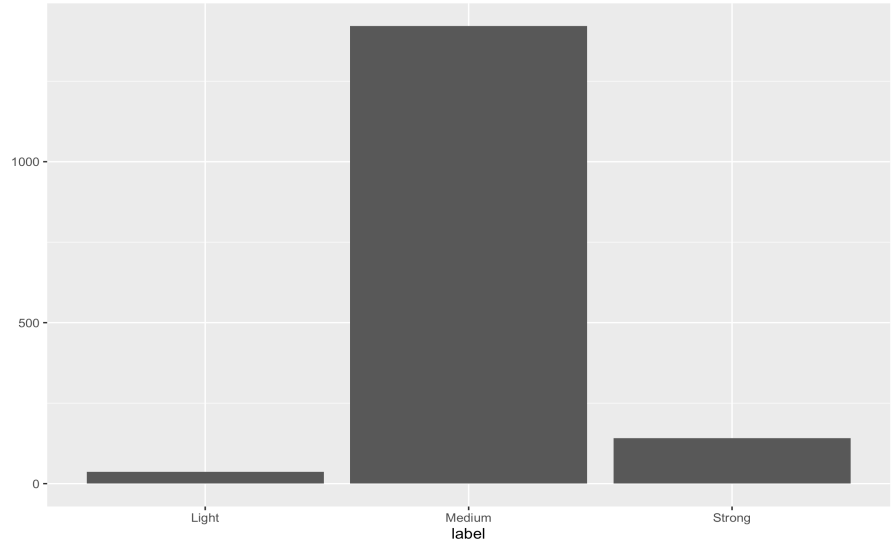


Alcohol: the percent alcohol content of the wine (% by volume). The Median Score 50% of alcohol is 10.2, the Mean Value of 10.42 and the third quantile is 11.1 as shown in the bar graph. Alcohol rate is left skewed. In these given datasets having an alcohol rate of below 11.1%.

Adding label of Alcohol percentage concentration: Light = <= 9 Medium = between 9 and 12 strong = >12

Alcohol Frequency Distribution Tibble

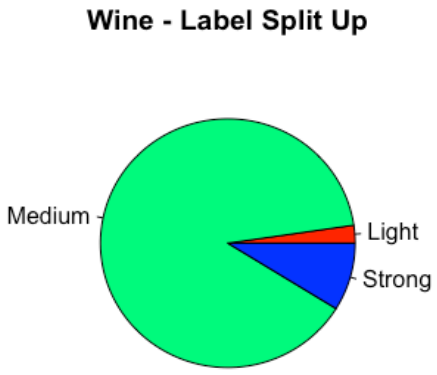
Var1	Freq
Light	37
Medium	1421
Strong	141



As shown above the histogram the highest peak represents by **"Medium label"** had 1,421% of alcohol percentage concentration, second represents by **"Strong label"** had 141% of alcohol percentage concentration and the last has a smallest "tail" represents the **"Light label"** had 37% of alcohol percentage concentration.

Pie Chart show how a total amount is divided between levels of a categorical variable as circle divided into radial slices. Each categorical value corresponds with a single slice of

the circle, and the size of each slice (both in area and arc length) indicates what proportion of the whole each category level takes.

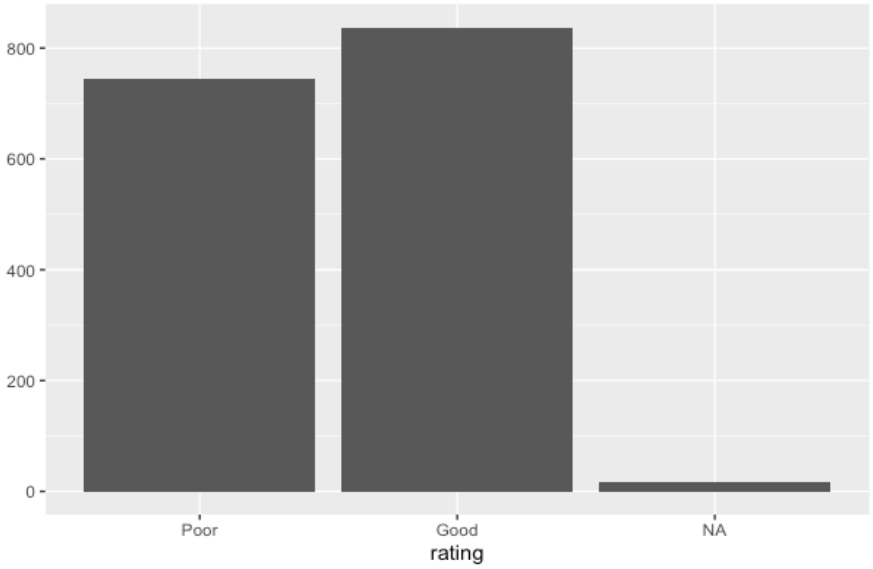


The pie chart above depicts the distribution of wine, we can see that wine label split up, represent by **"Medium label"** the first green slice, has just more than a half of the pie circle, Blue is in second represent **"Strong label"** and red is the last slice which represent the **"Light label"** smallest slice label.

Adding Rating as Red Wine Quality The Red Wine Quality Rating: Poor = < 5, Good = between 5 & 8, Excellent = 8<

Frequency Destribution of Rating as Red Wine Quality

Var1	Freq
Poor	744
Good	837



As shown in the above histogram Rating. The highest peak represents by **"Good"** had a rating of 837%, second represents by **"Poor"** had a rating of 744% and the smallest rating represents of excellent, there's no excellent rating as per given dataset.

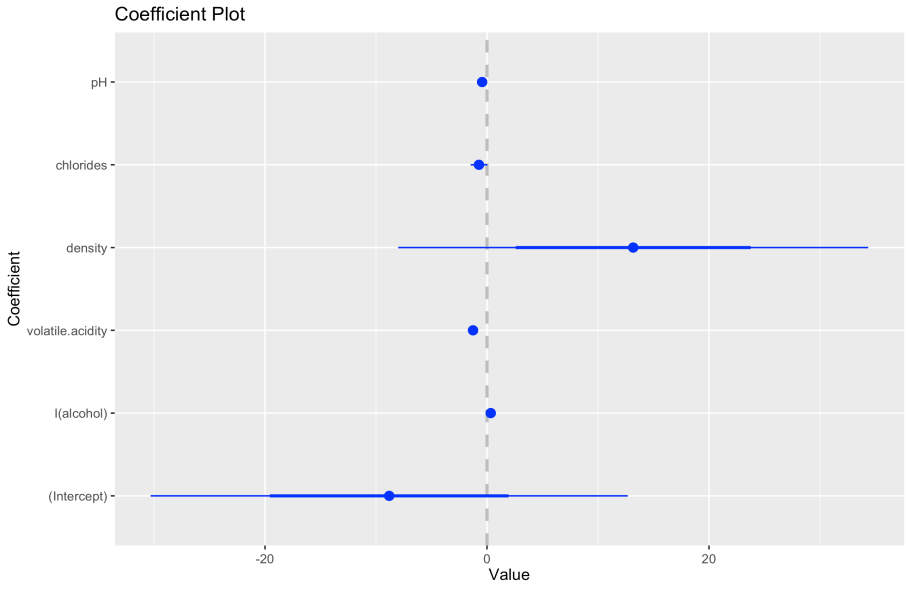
The pie chart above depicts the distribution of wine Rating. We can saw that wine-rating split up, represents by "Poor" rating the first green slice has just a half of the pie circle as well

as the red slice that represents by the “Good” rating, the Excellent rating there’s no rating as per given dataset.

Physicochemical Descriptive Statistics Tibble

X	fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides	free.sulfur.dioxide	total.sulfur.dioxide	density	pH	sulphates	alcohol	quality	label	rating	quality.rank
1	7.4	0.70	0.00	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5	Medium	Poor	Middle
2	7.8	0.88	0.00	2.6	0.098	25	67	0.9968	3.20	0.68	9.8	5	Medium	Poor	Middle
3	7.8	0.76	0.04	2.3	0.092	15	54	0.9970	3.26	0.65	9.8	5	Medium	Poor	Middle
4	11.2	0.28	0.56	1.9	0.075	17	60	0.9980	3.16	0.58	9.8	6	Medium	Good	Middle
5	7.4	0.70	0.00	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5	Medium	Poor	Middle
6	7.4	0.66	0.00	1.8	0.075	13	40	0.9978	3.51	0.56	9.4	5	Medium	Poor	Middle

Coefficient Pr (>| t|) Individual p value for each parameter to accept or reject null hypothesis, this is statistical estimate of x and y lower the p value allow us to reject null hypothesis, all type of error (true positive/negative, false positive /negative) are come to picture if we wrongly analysis p value. Asterisks mark aside p value define significance of value, lower the value have high significance.



Regression results of the following physicochemical, (1) pH having a coefficient of **-0.439**, (2) Chlorides having a coefficient of **-1.7635**, (3) Density having a coefficient of

13.173, (4) Volatile Acidity having a coefficient of **-1.260**, (5) Alcohol having a coefficient of **0.336**, Intercept had **-8.811**.

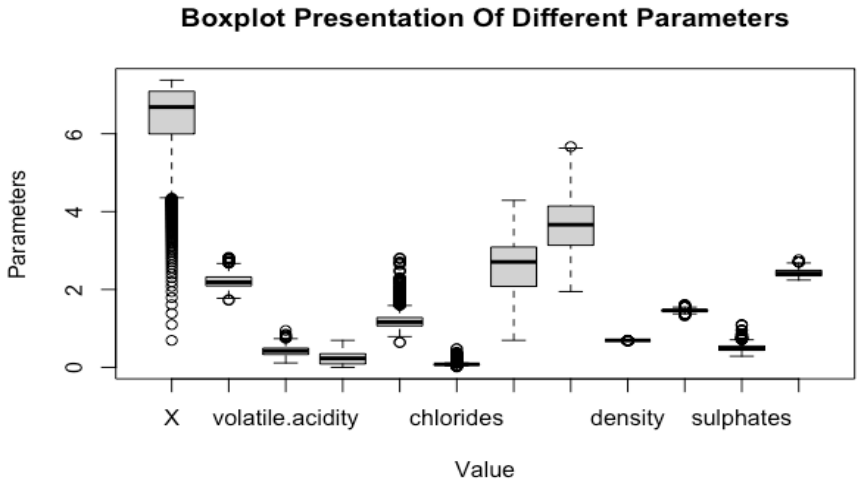
```
m1 <- lm(I(quality)~I(alcohol),data = wineQualityReds)
m2 <- update(m1, ~ . + volatile.acidity)
m3 <- update(m2, ~ . + density)
m4 <- update(m3, ~ . + chlorides)
m5 <- update(m4, ~ . + pH)

mtable(m1,m2,m3,m4,m5)
```

```
Calls:
m1: lm(formula = I(quality) ~ I(alcohol), data = wineQualityReds)
m2: lm(formula = I(quality) ~ I(alcohol) + volatile.acidity, data = wineQualityReds)
m3: lm(formula = I(quality) ~ I(alcohol) + volatile.acidity + density,
   data = wineQualityReds)
m4: lm(formula = I(quality) ~ I(alcohol) + volatile.acidity + density +
   chlorides, data = wineQualityReds)
m5: lm(formula = I(quality) ~ I(alcohol) + volatile.acidity + density +
   chlorides + pH, data = wineQualityReds)

============================================================================
                m1         m2         m3         m4         m5
----------------------------------------------------------------------------
 (Intercept)        1.875***   3.095***  -18.407    -19.637     -8.811
               (0.175)    (0.184)    (10.298)   (10.352)   (10.747)
 I(alcohol)         0.361***   0.314***   0.333***   0.330***   0.336***
               (0.017)    (0.016)    (0.018)    (0.019)    (0.019)
 volatile.acidity             -1.384***  -1.365***  -1.362***  -1.260***
                         (0.095)    (0.096)    (0.096)    (0.099)
 density                             21.360*    22.660*    13.173
                                 (10.228)   (10.289)   (10.588)
 chlorides                                      -0.422     -0.725
                                            (0.366)    (0.374)
 pH                                                        -0.439***
                                                       (0.122)
----------------------------------------------------------------------------
 R-squared          0.227      0.317      0.319      0.319      0.325
 N             1599       1599       1599       1599       1599
============================================================================
 Significance: *** = p < 0.001; ** = p < 0.01; * = p < 0.05
```

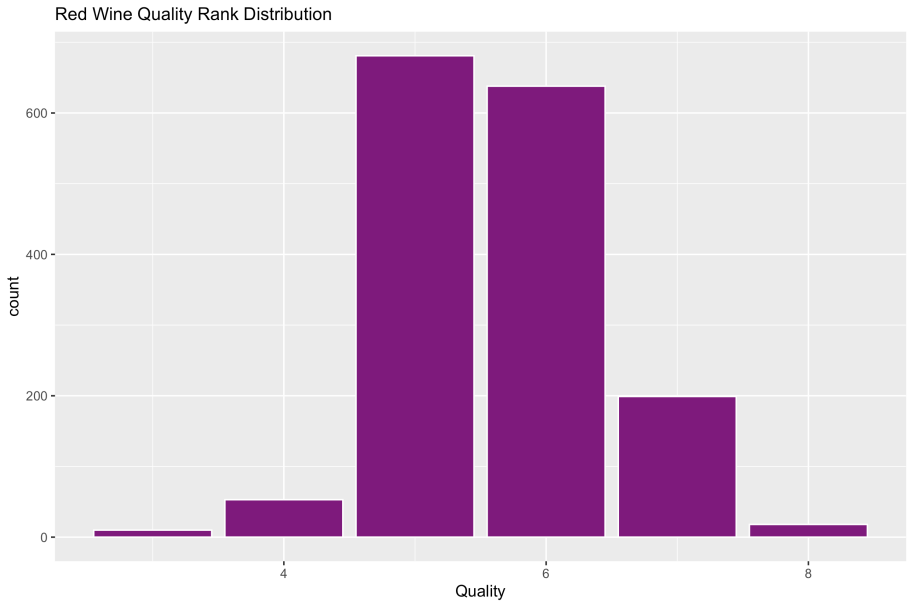
Boxplot for overall Dataset



As shown above the boxplot density stands out as having the fewest chemical substance while sulphates have the most chemical substance, the lower quartile for density is higher than the upper quartile for sulphates.

Red Wine Quality Rank

	x
Low	63
Middle	1319
High	217



As we show above the histogram of the Red Wine Quality Rank Distribution, the **"Low Rank"** had 63%, **"Middle Rank"** had 1,319% and **"High Rank"** had 217%.

Correlation Coefficient

The correlation coefficient denoted by r, tells us how closely data in a scatter plot fall along straight line. The closer that absolute value of r is to one. The better that the data are described by a linear equation. If r = 1 or r = -1 then the dataset is perfectly aligned. Datasets with values of r close to zero show little to no straight-line relationship.

Correlation coefficient formulas are used to find how strong a relationship is between data the formulas return a value between -1 and 1, where:

- 1 indicates a strong positive relationship.
- -1 indicates a strong negative relationship.
- A result of zero indicates no relationship at all

A correlation coefficient of one (1) means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other.

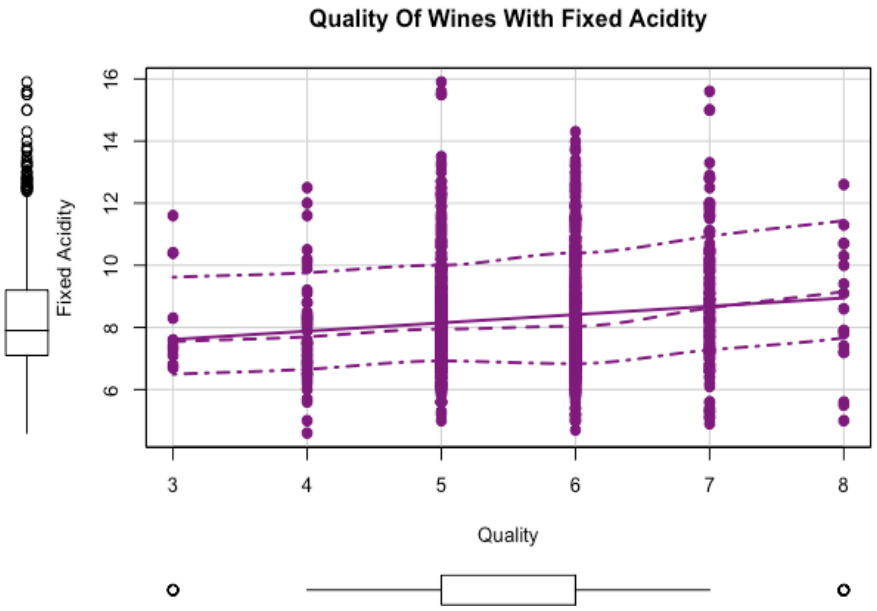
A correlation coefficient of negative one (-1) means that for every positive increase in one variable, there is a negative decrease of fixed proportion in the other.

Zero means that for every increase, there isn't a positive or negative increase. The two just aren't related.

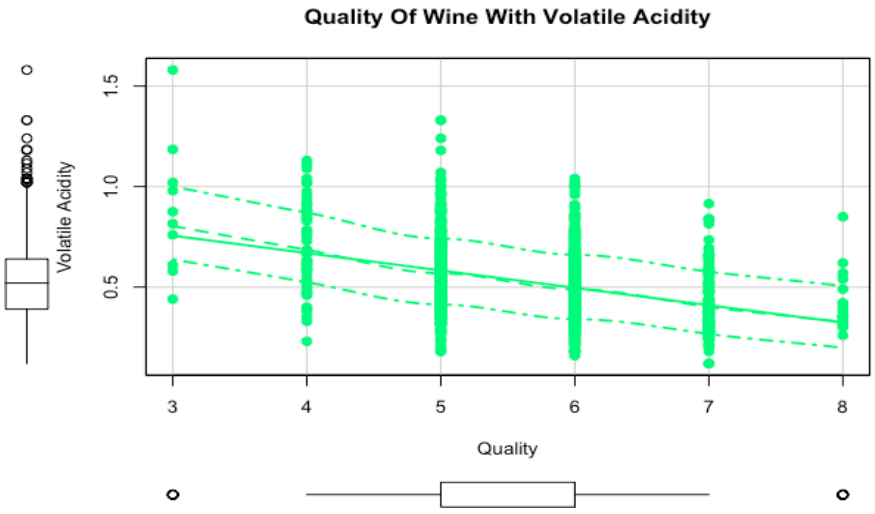
Equation:

$$r = \frac{n\,(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\,n\,\Sigma x^2 - (\Sigma x)^2\,][n\,\Sigma y^2 - (\Sigma y)^2]}}$$

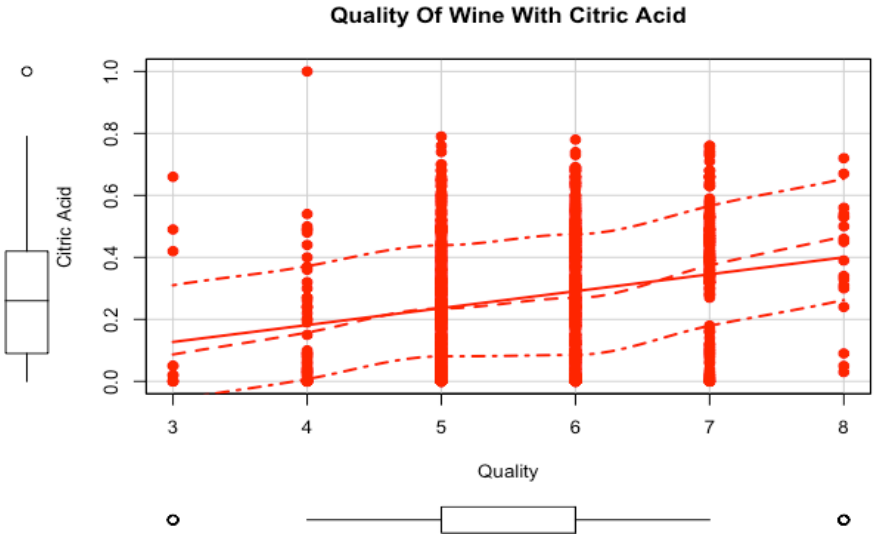
All scatter plots to determine the association between quality and other physicochemical properties:



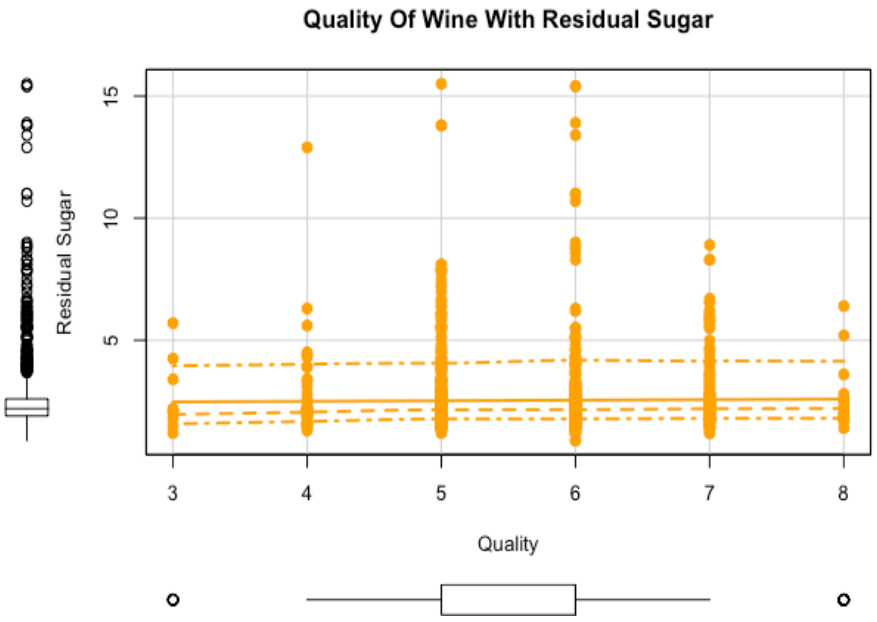
Fixed Acidity: A few things we see in this scatter plot are all about the quality & fixed acidity the regression line shows a strong positive relationship. There is a positive increase of a fixed proportion in the other. The standard deviation of 1.74 and had range of 1.130e + 01, there are few outliers on the high side.



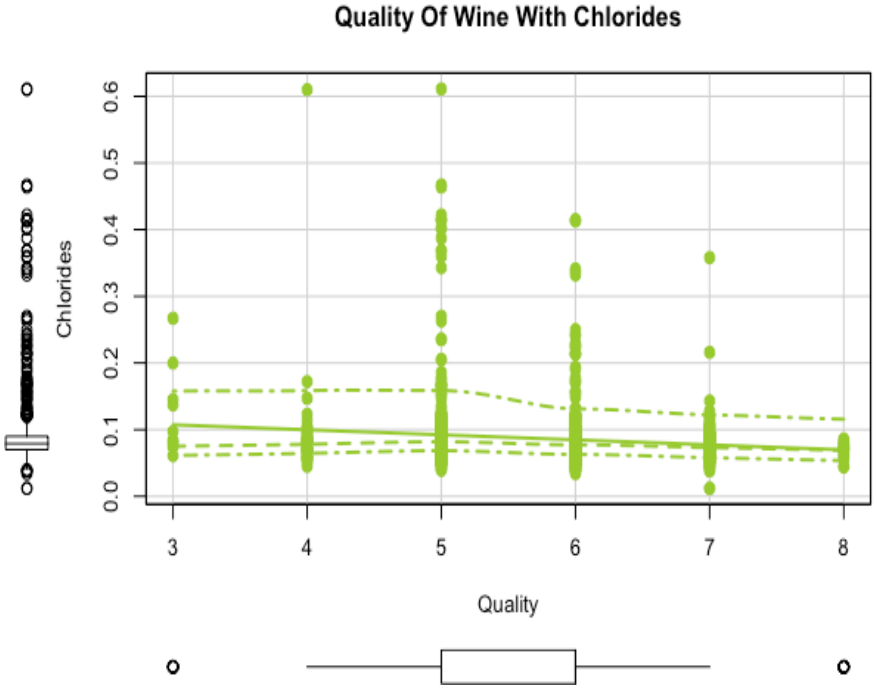
Volatile Acidity: The Quality & Volatile Acidity the regression line shows a negative correlation to wine quality. volatile acidity can contribute to acidity tastes which is often considered a wine fault. The standard deviation of 0.17 and had range of 1.460e + 00, there are few outliers on the high side.



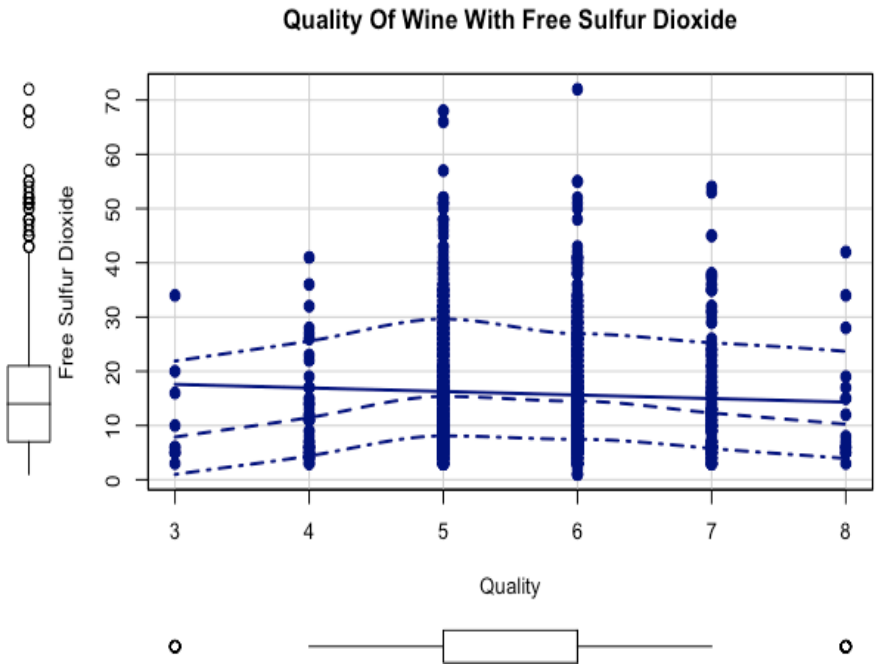
Citric Acid: The Quality & Citric Acid the regression line shows a result has positive correlation with wine quality wine maker often add citric to give a “freshness” test. However it can also bring unwanted effects through bacteria metabolism. The standard deviation of 0.19 and had range of 1.000e + 00, there are few outliers on high side.



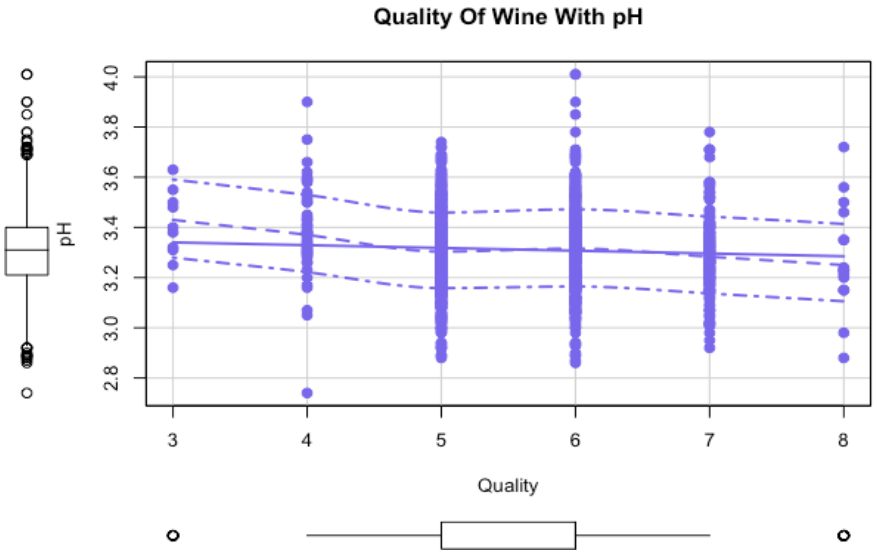
Residual Sugar: The Quality & Residual Sugar the regression line shows a result of zero indicates no relationship at all. The standard deviation of 1.40 and had range of 1.460e + 01, there are few outliers on high side.



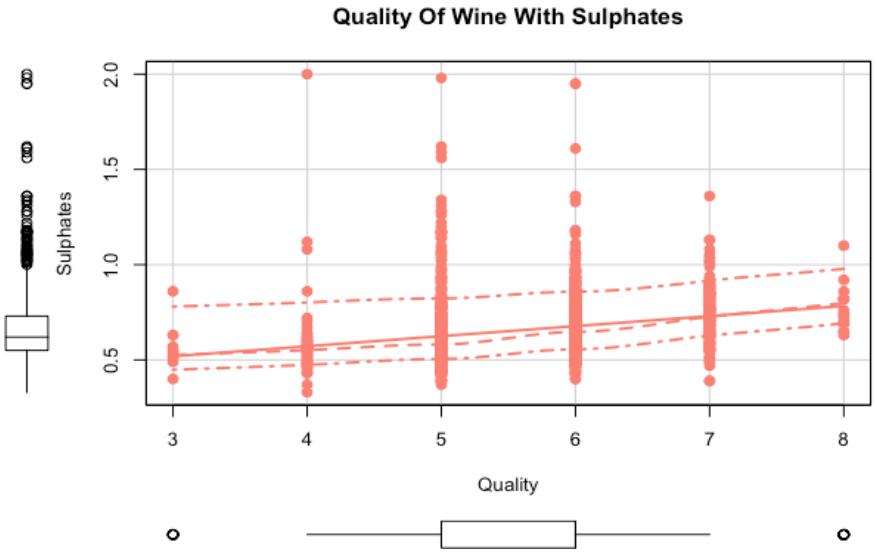
Chlorides: The Quality & Chlorides the regression line shows a result of zero indicates no relationship at all. The standard deviation of 0.04 and had a range of 5.990e -01, there are few outliers on high side.



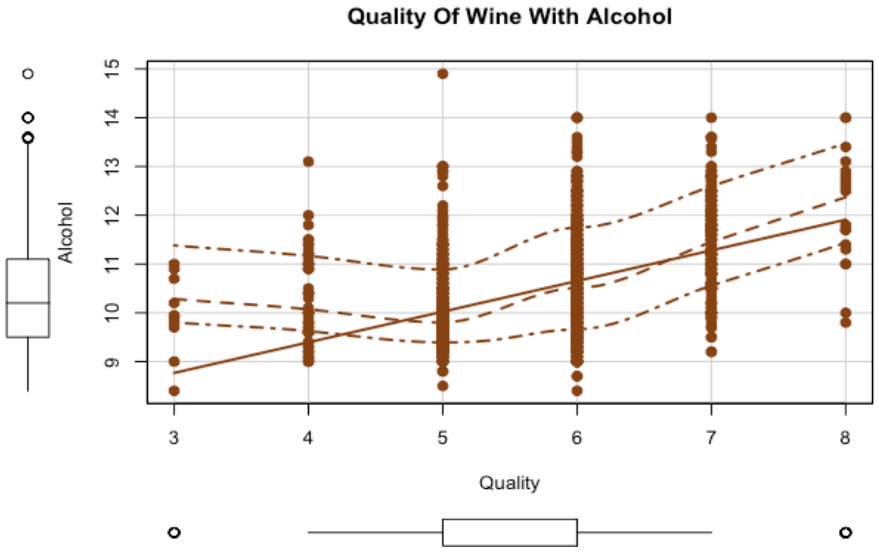
Free Sulfur Dioxide: The Quality & Free Sulfur Dioxide the regression line shows a result of a strong positive relationship. The standard deviation of 10.46 and had range of 7.100e + 01, there are few outliers on high side.



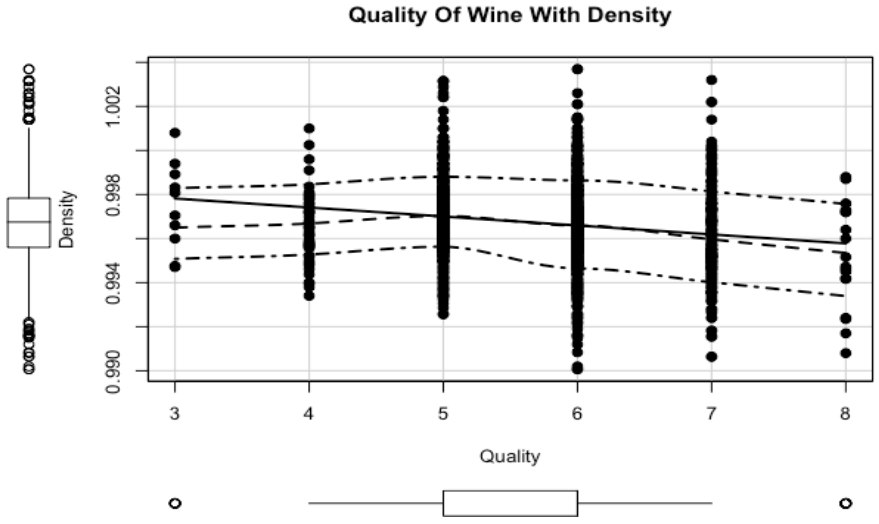
pH: The Quality & pH the regression line shows a result of strong positive relationship, standard deviation of 0.15 and had a range of 1.279e +00, there are few outliers on high side.



Sulphates: The Quality & Sulphates the regression line shows a result of strong positive relationship with wine quality, the standard deviation of 0.16 and had range of 1.670e + 00, there are few outliers on high side.



Alcohol: The Quality & Alcohol the regression line shows a one (1) indicates a strong positive relationship, the standard deviation of 1.065 and had a range of 6.500 + 00, there are few outliers on high side.



Density: The Quality & Density the regression line shows a negative one (-1) indicates a strong negative relationship, the standard deviation of 0.00 and had a range of 1.362e - 02, there are few outliers on high side.

Linear Model Assumptions

Linear regression analysis is based on six(6) fundamental assumptions:

1. The dependent and independent variables show a linear relationship between the slope and the intercept.
2. The independent variable is not random.
3. The value of the residual (error) is zero.
4. The value of the residual(error) is constant across all observations.
5. The value of the residual (error) is not correlated across all observations.
6. The residual (error) values follow the normal distribution.

Univariate Linear Regression Analysis

The tidy dataset contains 1,599 red wine observations and total of twelve (12) attributes in the dataset, eleven (11) of the attributes are numeric physicochemical test results of wine and one (1) attribute(quality) consists of sensory data ranging from 0 to 10. Some histogram plots in Rating are positively & negatively skewed and might be log-normally distributed and in the scatter plots of a quality of physicochemical some regression lines are in strong positive and strong negative relationship and some regression line at a zero result has no relationship at all.

B.2 Bivariate Linear Regression

Bivariate Linear Regression: is a linear equation describing the relationship between an explanatory variable and an outcome variable, specifically with the assumption that the explanatory variable influences the outcome variable, and not vice versa.
Simple Linear Regression - is a model that assesses the relationship between a dependent variable and an independent variable. The simple linear model is expressed using the following equation:

$$Y = a + bx + \epsilon$$

Where:

Y = Dependent variable

x = independent (explanatory) variable

a = intercept

b = slope

ϵ = Residual (error)

```
mytable <- xtabs(~quality + volatile.acidity, data= wineQualityReds)
addmargins(mytable)
```

```
volatile.acidity
quality 0.12 0.16 0.18 0.19  0.2 0.21 0.22 0.23 0.24 0.25 0.26 0.27 0.28 0.29 0.295
  3    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
  4    0    0    0    0    0    0    0    1    0    0    0    0    0    0    0
  5    0    0    2    1    1    1    0    1    2    0    2    1    5    3    0
  6    0    2    7    1    1    2    5    3    6    3   10    8    5    8    1
  7    3    0    1    0    1    3    1    0    5    4    3    5   13    5    0
  8    0    0    0    0    0    0    0    0    0    0    1    0    0    0    0
    volatile.acidity
quality  0.3 0.305 0.31 0.315 0.32 0.33 0.34 0.35 0.36 0.365 0.37 0.38 0.39 0.395  0.4
  3    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
  4    0    0    0    0    0    1    0    1    0    0    0    1    1    0    1
  5    0    0    4    0   10    3    9    6    5    2    9   10   15    0    8
  6    6    2   12    2    7    5   16    9   23    0    9   16   16    1   21
  7    9    0   13    0    5   10    5    3    9    0    6    7    3    1    6
  8    1    0    1    0    1    1    0    3    1    0    0    1    0    0    1
    volatile.acidity
quality 0.41 0.415 0.42 0.43 0.44 0.45 0.46 0.47 0.475 0.48 0.49  0.5 0.51 0.52 0.53
  3    0    0    0    0    1    0    0    0    0    0    0    0    0    0    0
  4    0    0    0    0    0    0    1    2    0    1    0    1    0    2    1
  5   16    1   12   23    6   12   16    8    0    8   16   24    2   13    7
  6   11    2   12   17   11   10   14    9    2    8   17   19   19   13   19
  7    6    0    5    3    5    0    0    2    0    7    1    2    3    5    2
  8    0    0    2    0    0    0    0    0    0    0    1    0    0    0    0
    volatile.acidity
quality 0.54 0.545 0.55 0.56 0.565 0.57 0.575 0.58 0.585 0.59 0.595  0.6 0.605 0.61
  3    0    0    0    0    0    0    0    1    0    0    0    0    0    1
  4    0    0    1    0    0    0    0    2    0    2    0    0    0    2
  5   15    0   10   21    1   12    3   15    1   20    1   31    1   15
  6   14    5    8   10    0   13    0   16    1   13    0   15    2    9
  7    1    0    1    3    0    2    0    4    1    4    0    1    0    0
  8    1    0    0    0    0    1    0    0    0    0    0    0    0    0
    volatile.acidity
quality 0.615 0.62 0.625 0.63 0.635 0.64 0.645 0.65 0.655 0.66 0.665 0.67 0.675 0.68
  3    0    0    0    0    0    0    0    0    0    0    0    0    0    0
  4    0    1    1    1    1    0    0    0    0    0    0    2    1    2
  5    4   15    1   15    3   19    6    8    7   16    3   11    2    7
  6    0    7    1   13    3    7    6    6    0    5    0   10    0    3
  7    2    0    0    0    2    1    0    2    0    5    0    0    0    0
  8    0    1    0    0    0    0    0    0    0    0    0    0    0    0
  volatile.acidity
quality 0.685 0.69 0.695  0.7 0.705 0.71 0.715 0.72 0.725 0.73 0.735 0.74 0.745 0.75
  3    0    0    0    0    0    0    0    0    0    0    0    0    0    0
  4    1    0    0    0    0    0    0    0    0    1    0    0    0    0
  5    6   14    6    6    5    2    7    4    6    5    5    5    3    3
  6    3    9    0    4    1    1    5    1    3    0    2    6    2    3
  7    1    0    1    0    0    0    0    0    0    0    1    0    0    0
  8    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    volatile.acidity
```

quality 0.755 0.76 0.765 0.77 0.775 0.78 0.785 0.79 0.795 0.8 0.805 0.81 0.815 0.82

3	0	1	0	0	0	0	0	0	0	0	0	0	1	0
4	1	0	1	0	0	1	1	0	0	0	0	0	0	0
5	1	1	2	3	3	7	3	1	1	2	1	1	1	5
6	1	3	2	3	1	2	4	1	1	1	0	1	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0	1	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0

volatile.acidity

quality 0.825 0.83 0.835 0.84 0.845 0.85 0.855 0.86 0.865 0.87 0.875 0.88 0.885 0.89

3	0	0	0	0	0	0	0	0	0	0	1	0	0	0
4	0	1	0	1	1	0	0	0	0	1	0	1	1	0
5	0	2	3	1	0	1	0	1	0	3	1	4	4	1
6	1	1	0	4	0	0	3	1	1	0	0	0	0	0
7	0	0	1	2	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	1	0	0	0	0	0	0	0	0

volatile.acidity

quality 0.895 0.9 0.91 0.915 0.92 0.935 0.95 0.955 0.96 0.965 0.975 0.98 1 1.005

3	0	0	0	0	0	0	0	0	0	0	0	1	0	0
4	0	0	1	1	0	1	0	1	0	0	1	0	0	0
5	0	2	1	2	1	1	1	0	2	3	0	1	1	1
6	1	1	1	0	0	0	0	0	1	0	0	1	2	0
7	0	0	0	1	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0

volatile.acidity

quality 1.01 1.02 1.025 1.035 1.04 1.07 1.09 1.115 1.13 1.18 1.185 1.24 1.33 1.58 Sum

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

[reached getOption("max.print") -- omitted 1 row]

For a **chi-squared test**, a p-value that is less than or equal to your significance level indicates there is sufficient evidence to conclude that the observed distribution is not the same as the expected distribution. You can conclude that a relationship exists between the categorical variables.

$$X^2 = \Sigma \frac{(O_i - E_i)^2}{E_i}$$

Where:

X^2 = Chi Squared

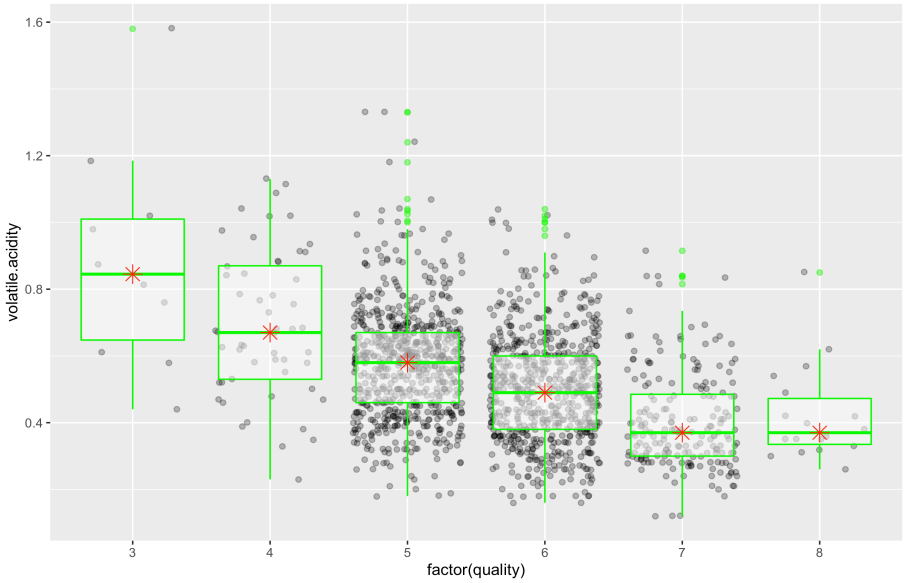
O_i = Observed Value

= Expected Value

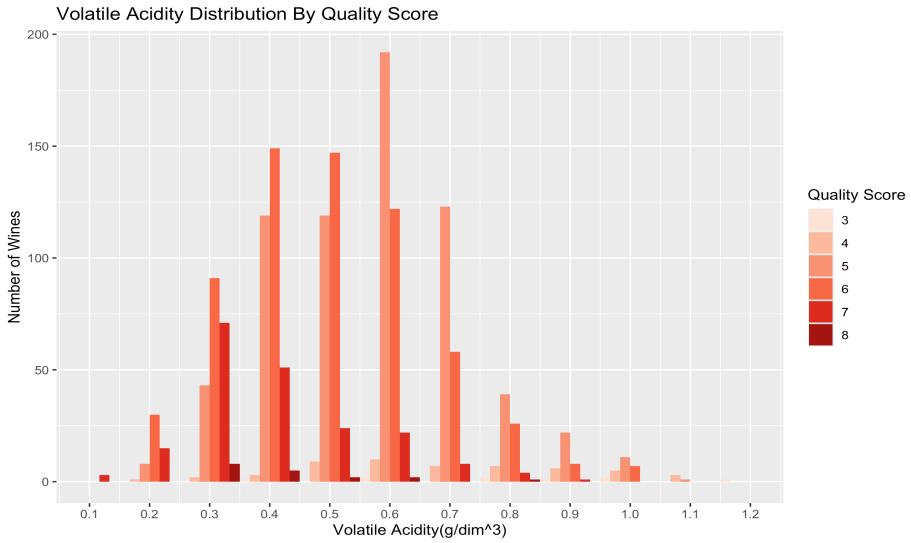
chisq.test(mytable)

Pearson's Chi-squared test

data: mytable
X-squared = 1589.2, df = 710, p-value < 2.2e-16



Factor(quality) Vs Volatile Acidity: Volatile Acidity having an outlier of 1.58 in a green dot as shown at the plot, maximum observation below upper fence had 1.58%, 75 percentiles of 0.64(upper quartile), mean value of 0.52, Median 50 percentiles of 0.52 in red asterisks, 25 percentiles of 0.3(lower quartile), and having a minimum observation of 0.12 below (lower quartile).



As shown above histogram plot, we can call this a multimodal distribution- where there are more than two peaks and very rare. it is a probability distribution with more than one "peak" or "mode". The volatile Acidity quality score light color it's represents the lowest score of quality to the dark color which represents a high score, it's gradually lowest to highest score.

```
mytable<- xtabs(~pH + alcohol,data= wineQualityReds)
addmargins(mytable)
```

```
alcohol
pH    8.4 8.5 8.7 8.8   9 9.05 9.1 9.2 9.23333333333333 9.25 9.3 9.4 9.5
 2.74   0   0   0   0   0    0   0   0                0    0   0   1   0
 2.86   1   0   0   0   0    0   0   0                0    0   0   0   0
 2.87   0   0   0   0   0    0   0   0                0    0   0   0   0
 2.88   0   0   0   0   0    0   0   0                0    0   0   0   0
 2.89   0   0   0   0   0    0   0   0                0    0   0   0   0
 2.9    0   0   0   0   0    0   1   0                0    0   0   0   0
 2.92   0   0   0   0   0    0   0   0                0    0   0   0   0
 2.93   0   0   0   0   0    0   0   0                0    0   0   0   0
 2.94   0   0   0   0   0    0   0   2                0    0   0   0   1
 2.95   0   0   0   0   0    0   0   0                0    0   0   0   0
 2.98   0   0   0   0   0    0   0   1                0    0   0   1   0
 2.99   0   0   0   0   0    0   0   0                0    0   0   0   0
 3      0   0   0   0   0    0   0   2                0    0   2   0   1
 3.01   0   0   0   0   0    0   0   0                0    0   0   1   0
 3.02   0   0   0   0   0    0   0   1                0    0   1   0   1
    alcohol
pH    9.55 9.56666666666667  9.6  9.7  9.8  9.9 9.95   10 10.0333333333333 10.1 10.2
 2.74    0                 0    0    0    0    0    0    0                 0    0    0
 2.86    0                 0    0    0    0    0    0    0                 0    0    0
 2.87    0                 0    0    0    0    0    0    0                 0    0    1
```

2.88	0	0	0	1	1	0	0	0	0	0	0	0	
2.89	0	0	2	0	0	0	0	0	0	0	0	0	
2.9	0	0	0	0	0	0	0	0	0	0	0	0	
2.92	0	0	0	0	0	0	0	0	0	0	0	0	
2.93	0	0	0	0	1	2	0	0	0	0	0	0	
2.94	0	0	0	0	0	0	0	0	0	0	0	1	
2.95	0	0	0	0	0	0	0	0	0	0	0	0	
2.98	0	0	0	0	0	1	0	0	0	0	0	0	
2.99	0	0	0	0	0	0	0	0	0	0	0	2	
3	0	0	0	0	0	0	0	0	0	0	0	0	
3.01	0	0	0	0	0	0	0	0	0	0	0	0	
3.02	0	0	0	0	0	2	0	0	0	0	0	0	

alcohol

pH	10.3	10.4	10.5	10.55	10.6	10.7	10.75	10.8	10.9	11	11.0666666666667	11.1	11.2
2.74	0	0	0	0	0	0	0	0	0	0	0	0	0
2.86	0	0	0	0	0	0	0	0	0	0	0	0	0
2.87	0	0	0	0	0	0	0	0	0	0	0	0	0
2.88	0	0	0	0	0	0	0	0	0	0	0	0	0
2.89	0	0	0	0	0	0	0	0	0	0	0	0	0
2.9	0	0	0	0	0	0	0	0	0	0	0	0	0
2.92	0	0	1	0	0	0	0	0	0	0	0	3	0
2.93	0	0	0	0	0	0	0	0	0	0	0	0	0
2.94	0	0	0	0	0	0	0	0	0	0	0	0	0
2.95	0	0	0	0	0	0	0	0	0	0	0	0	1
2.98	0	0	0	0	0	0	0	0	0	0	0	0	0
2.99	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0
3.01	0	0	0	0	0	0	0	1	0	0	0	0	0
3.02	0	0	0	0	0	0	0	0	2	0	0	0	0

alcohol

pH	11.3	11.4	11.5	11.6	11.7	11.8	11.9	11.95	12	12.1	12.2	12.3	12.4	12.5	12.6
2.74	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2.86	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2.87	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2.88	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2.89	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2.9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2.92	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2.93	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2.94	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2.95	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2.98	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
2.99	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3.01	0	0	0	0	0	1	0	0	0	0	0	2	0	0	0
3.02	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0

alcohol

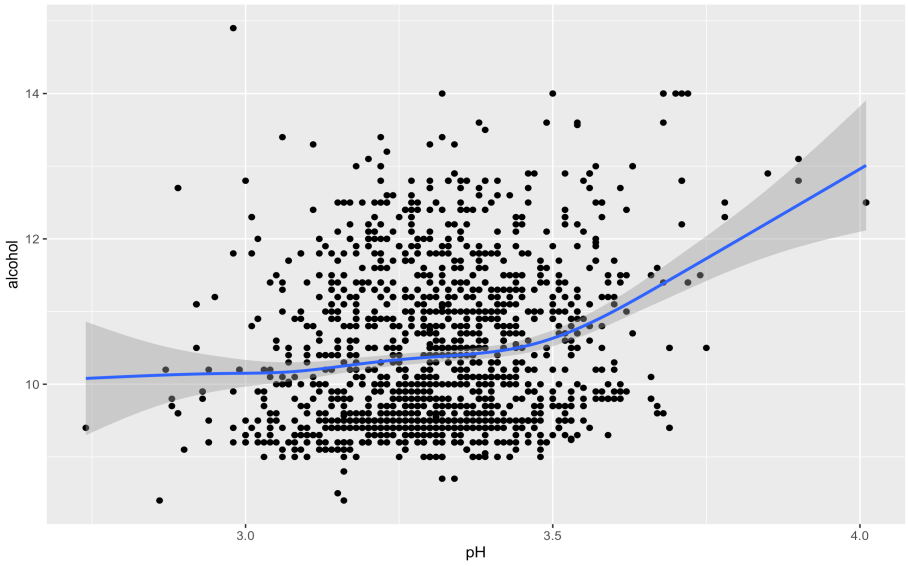
pH	12.7	12.8	12.9	13	13.1	13.2	13.3	13.4	13.5	13.5666666666667	13.6	14	14.9
2.74	0	0	0	0	0	0	0	0	0	0	0	0	0
2.86	0	0	0	0	0	0	0	0	0	0	0	0	0
2.87	0	0	0	0	0	0	0	0	0	0	0	0	0
2.88	0	0	0	0	0	0	0	0	0	0	0	0	0

```
 2.89   2   0   0   0   0   0   0   0   0          0   0   0   0
 2.9    0   0   0   0   0   0   0   0   0          0   0   0   0
 2.92   0   0   0   0   0   0   0   0   0          0   0   0   0
 2.93   0   0   0   0   0   0   0   0   0          0   0   0   0
 2.94   0   0   0   0   0   0   0   0   0          0   0   0   0
 2.95   0   0   0   0   0   0   0   0   0          0   0   0   0
 2.98   0   0   0   0   0   0   0   0   0          0   0   0   1
 2.99   0   0   0   0   0   0   0   0   0          0   0   0   0
 3      0   1   0   0   0   0   0   0   0          0   0   0   0
 3.01   0   0   0   0   0   0   0   0   0          0   0   0   0
 3.02   0   0   0   0   0   0   0   0   0          0   0   0   0
    alcohol
pH     Sum
 2.74   1
 2.86   1
 2.87   1
 2.88   2
 2.89   4
 2.9    1
 2.92   4
 2.93   3
 2.94   4
 2.95   1
 2.98   5
 2.99   2
 3      6
 3.01   5
 3.02   8
 [ reached getOption("max.print") -- omitted 75 rows ]
```

```
chisq.test(mytable)
```

Pearson's Chi-squared test

data: mytableX-squared = 9140.4, df = 5632, p-value < 2.2e-16



pH and Alcohol: the result of given above in a geom smooth plot and a line of best fit, the geom smooth plot is made up of the black points where each point represents the results of physicochemical namely pH and alcohol, so, for each point the x coordinate represent the pH had 3.4 and each point the y coordinate represent the alcohol had 13.2, and then performed linear regression the resulting linear equation would produce the blue line in a linear positive, which we call the line fo best fit, so the line of best fit in the line that best represent by given data for us to predict the quality of wine. That a prediction will always have margin of error.

```
mytable<- xtabs(~density + alcohol, data = wineQualityReds)
head(mytable)
```

```
     alcohol
density   8.4 8.5 8.7 8.8 9 9.05 9.1 9.2 9.23333333333333 9.25 9.3 9.4 9.5 9.55
  0.99007  0  0  0  0 0  0  0  0          0  0  0  0  0  0
  0.9902   0  0  0  0 0  0  0  0          0  0  0  0  0  0
  0.99064  0  0  0  0 0  0  0  0          0  0  0  0  0  0
  0.9908   0  0  0  0 0  0  0  0          0  0  0  0  0  0
  0.99084  0  0  0  0 0  0  0  0          0  0  0  0  0  0
  0.9912   0  0  0  0 0  0  0  0          0  0  0  0  0  0
     alcohol
density   9.56666666666667 9.6 9.7 9.8 9.9 9.95 10 10.0333333333333 10.1 10.2 10.3
  0.99007           0  0  0  0  0  0 0          0  0  0  0
  0.9902            0  0  0  0  0  0 0          0  0  0  0
  0.99064           0  0  0  0  0  0 0          0  0  0  0
  0.9908            0  0  0  0  0  0 0          0  0  0  0
  0.99084           0  0  0  0  0  0 0          0  0  0  0
  0.9912            0  0  0  0  0  0 0          0  0  0  0
     alcohol
density   10.4 10.5 10.55 10.6 10.7 10.75 10.8 10.9 11 11.0666666666667 11.1 11.2 11.3
```

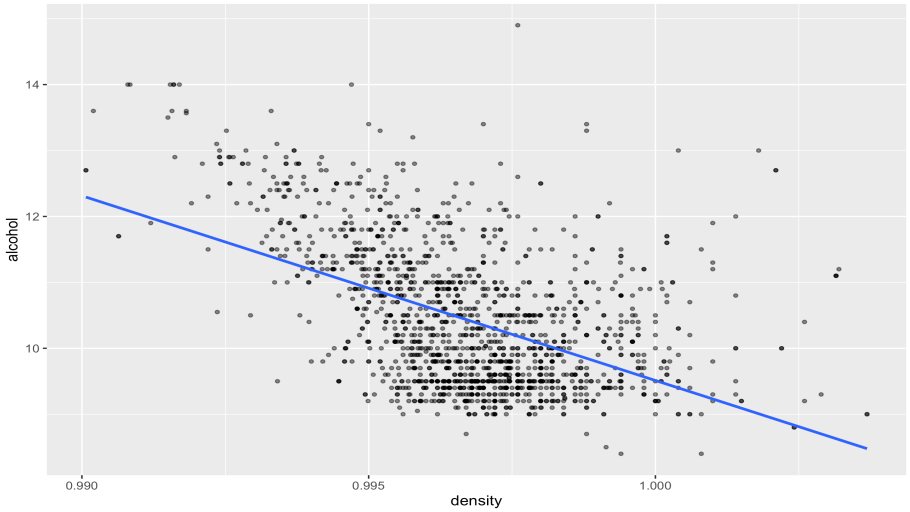
```
 0.99007   0   0   0   0   0   0   0   0 0          0   0   0   0
 0.9902    0   0   0   0   0   0   0   0 0          0   0   0   0
 0.99064   0   0   0   0   0   0   0   0 0          0   0   0   0
 0.9908    0   0   0   0   0   0   0   0 0          0   0   0   0
 0.99084   0   0   0   0   0   0   0   0 0          0   0   0   0
 0.9912    0   0   0   0   0   0   0   0 0          0   0   0   0
      alcohol
density   11.4 11.5 11.6 11.7 11.8 11.9 11.95 12 12.1 12.2 12.3 12.4 12.5 12.6 12.7
 0.99007   0   0   0   0   0   0   0 0   0   0   0   0   0   0   0   2
 0.9902    0   0   0   0   0   0   0 0   0   0   0   0   0   0   0   0
 0.99064   0   0   0   2   0   0   0 0   0   0   0   0   0   0   0   0
 0.9908    0   0   0   0   0   0   0 0   0   0   0   0   0   0   0   0
 0.99084   0   0   0   0   0   0   0 0   0   0   0   0   0   0   0   0
 0.9912    0   0   0   0   0   1   0 0   0   0   0   0   0   0   0   0
      alcohol
density   12.8 12.9 13 13.1 13.2 13.3 13.4 13.5 13.5666666666667 13.6 14 14.9
 0.99007   0   0 0   0   0   0   0   0          0   0 0   0
 0.9902    0   0 0   0   0   0   0   0          0   1 0   0
 0.99064   0   0 0   0   0   0   0   0          0   0 0   0
 0.9908    0   0 0   0   0   0   0   0          0   0 1   0
 0.99084   0   0 0   0   0   0   0   0          0   0 1   0
 0.9912    0   0 0   0   0   0   0   0          0   0 0   0
```

```
chisq.test(mytable)
```

```
    Pearson's Chi-squared test

data:  mytable

X-squared = 44200, df = 27840, p-value < 2.2e-16
```



Alcohol ~ Density: y ~ x: the result of given above is a scatter plot and a line of best fit, the scatter plot is made up of the black points where each point represents the results of physicochemical namely density and alcohol. So, for each point the x coordinate represent the density had 0.99 and each y coordinate represent the alcohol had 12.8, and then performed linear regression the resulting linear equation would produce the blue line in a linear negative, which we call the line of best fit, so the line of best fit in the line that best represent by given data for us to predict the quality of wine. That a prediction will always have margin of error.

```
mytable <- xtabs(~chlorides + alcohol, data = wineQualityReds)
addmargins(mytable)
```

```
alcohol
chlorides  8.4  8.5  8.7  8.8    9 9.05  9.1  9.2 9.23333333333333 9.25  9.3  9.4  9.5
    0.012    0    0    0    0    0    0    0    0                 0    0    0    0    0
    0.034    0    0    0    0    0    0    0    0                 0    0    0    0    0
    0.038    0    0    0    0    0    0    0    0                 0    0    0    0    0
    0.039    0    0    0    0    0    0    0    0                 0    0    0    1    0
    0.041    0    0    0    0    0    0    0    0                 0    0    0    0    0
    0.042    0    0    0    0    0    0    0    0                 0    0    0    0    0
    0.043    0    0    0    0    0    0    0    0                 0    0    0    0    0
    0.044    0    0    0    0    0    0    0    0                 0    0    0    0    0
    0.045    0    0    0    0    0    0    0    0                 0    0    0    0    1
    0.046    0    0    0    0    0    0    0    0                 0    0    0    0    0
    0.047    0    0    0    0    0    0    0    0                 0    0    0    0    0
    0.048    0    0    0    0    0    0    0    0                 0    0    0    0    0
    0.049    0    0    0    0    0    0    0    0                 0    0    0    0    0
    0.05     0    0    0    0    0    0    1    0                 0    0    0    0    1
    0.051    0    0    0    0    0    0    0    0                 0    0    0    0    0
       alcohol
chlorides 9.55 9.56666666666667  9.6  9.7  9.8  9.9 9.95   10 10.0333333333333 10.1
    0.012    0                 0    0    0    0    0    0                   0    0
    0.034    0                 0    0    0    0    0    0                   0    0
    0.038    0                 0    0    0    0    0    0                   0    0
    0.039    0                 0    0    0    0    0    0                   0    0
    0.041    0                 0    0    0    0    0    0                   0    0
    0.042    0                 0    0    0    0    0    0                   0    0
    0.043    0                 0    0    0    0    0    0                   0    0
    0.044    0                 0    0    0    0    0    0                   0    0
    0.045    0                 0    0    0    0    0    0                   0    0
    0.046    0                 0    0    0    0    0    0                   0    0
    0.047    0                 0    0    1    0    0    0                   0    0
    0.048    0                 0    0    0    0    0    0                   0    0
    0.049    0                 0    0    0    0    0    0                   0    0
    0.05     0                 0    0    0    0    0    0                   0    0
    0.051    0                 0    0    0    0    0    0                   0    0
       alcohol
chlorides 10.2 10.3 10.4 10.5 10.55 10.6 10.7 10.75 10.8 10.9   11 11.0666666666667
    0.012    0    0    0    0     0    0    0     0    0    0    0                 0
    0.034    0    0    0    0     0    0    0     0    0    0    0                 0
```

0.038	0	0	0	0	0	0	0	0	1	0	0	0
0.039	0	0	0	0	0	0	0	0	0	0	0	0
0.041	0	0	0	0	0	0	0	0	0	0	0	0
0.042	0	0	0	0	0	0	1	0	0	0	0	0
0.043	0	0	0	0	0	0	0	0	0	0	0	0
0.044	0	0	0	0	0	0	0	0	0	1	0	0
0.045	0	0	0	1	0	0	0	0	0	0	0	0
0.046	0	0	0	0	0	0	0	0	0	0	0	0
0.047	0	0	0	0	0	0	0	0	0	2	0	0
0.048	0	0	0	0	0	0	0	0	1	0	0	0
0.049	0	0	0	0	0	0	0	0	0	0	0	0
0.05	1	0	0	0	1	0	0	0	0	0	1	0
0.051	0	0	0	0	0	0	0	0	0	0	0	0

alcohol

chlorides	11.1	11.2	11.3	11.4	11.5	11.6	11.7	11.8	11.9	11.95	12	12.1	12.2	12.3	12.4
0.012	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0
0.034	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.038	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.039	0	0	0	0	1	0	0	0	0	0	0	0	0	0	2
0.041	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
0.042	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
0.043	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
0.044	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.045	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
0.046	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0
0.047	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
0.048	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
0.049	0	0	1	2	0	0	1	0	0	0	0	0	0	0	0
0.05	0	0	0	0	1	0	0	0	0	1	1	0	0	2	0
0.051	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0

alcohol

chlorides	12.5	12.6	12.7	12.8	12.9	13	13.1	13.2	13.3	13.4	13.5	13.5666666666667	13.6
0.012	0	0	0	0	0	0	0	0	0	0	0	0	0
0.034	0	0	0	1	0	0	0	0	0	0	0	0	0
0.038	0	0	0	0	0	0	0	0	0	0	0	0	0
0.039	0	0	0	0	0	0	0	0	0	0	0	0	0
0.041	3	0	0	0	0	0	0	0	0	0	0	0	0
0.042	0	0	0	0	1	0	0	0	0	0	0	0	0
0.043	0	0	0	0	0	0	0	0	0	0	0	0	0
0.044	1	0	0	0	1	0	0	0	0	0	0	0	1
0.045	0	0	0	0	1	0	0	0	0	0	0	0	0
0.046	0	0	0	0	0	1	0	0	0	0	0	0	1
0.047	0	0	0	0	0	0	0	0	0	0	0	0	0
0.048	0	0	0	1	0	0	0	0	0	0	0	1	1
0.049	0	0	2	0	0	2	0	0	0	0	0	0	0
0.05	0	0	0	0	0	0	0	0	0	0	0	0	0
0.051	0	0	0	0	0	0	0	0	0	0	0	0	0

alcohol

chlorides	14	14.9	Sum
0.012	0	0	2
0.034	0	0	1
0.038	0	0	2

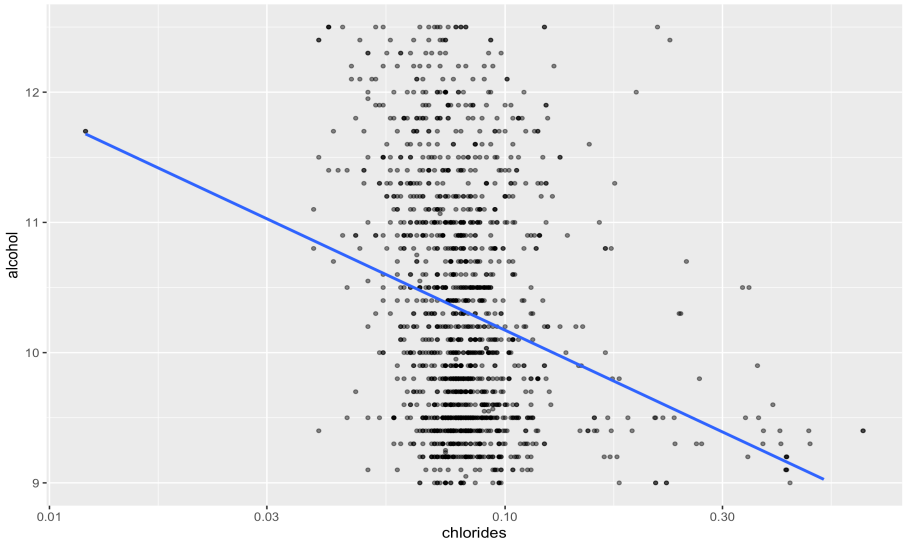
```
  0.039   0   0   4
  0.041   0   0   4
  0.042   0   0   3
  0.043   0   0   1
  0.044   1   0   5
  0.045   0   0   4
  0.046   0   0   4
  0.047   0   0   4
  0.048   2   0   8
  0.049   0   0   8
  0.05    2   0  12
  0.051   0   0   1
[ reached getOption("max.print") -- omitted 139 rows ]
```

```
chisq.test(mytable)
```

Pearson's Chi-squared test

data: mytable

X-squared = 14083, df = 9728, p-value < 2.2e-16



Alcohol ~ chlorides: y ~ x : the result of given above is a scatter plot and a line of best fit, the scatter plot is made up of the black points where each point represents the results of physicochemical namely chloride and alcohol, so, for each point the x coordinate represent the chlorides had 0.012 and each y coordinate represent the alcohol had 11.8, and then performed linear regression the resulting linear equation would produce the blue line in a linear negative, which we call the line of best fit, so the line of best fit in the line that best

represent by given data for us to predict the quality of wine. That a prediction will always have margin of error.

```
mytable<- xtabs (~residual.sugar + alcohol, data = wineQualityReds)
addmargins(mytable)
```

```
alcohol
residual.sugar  8.4  8.5  8.7  8.8    9 9.05  9.1  9.2 9.23333333333333 9.25  9.3  9.4
      0.9    0    0    0    0    0    0    0    0                   0    0    0    0
      1.2    0    0    0    0    0    0    0    0                   0    0    0    0
      1.3    0    0    0    0    0    0    0    1                   0    0    1    0
      1.4    0    0    0    0    0    0    1    3                   0    0    1    0
      1.5    0    0    0    0    0    0    1    0                   0    0    2    2
      1.6    0    1    1    0    2    0    1    5                   0    0    3    6
      1.65   0    0    0    0    0    0    0    0                   0    0    0    0
      1.7    0    0    0    0    1    0    2    3                   0    0    6    5
      1.75   0    0    0    0    0    0    0    0                   0    0    0    0
      1.8    1    0    0    0    1    0    3    6                   0    0    6    7
      1.9    0    0    0    0   11    1    0    4                   0    0    5    7
      2      0    0    0    0    1    0    5    8                   0    0    5   16
      2.05   0    0    0    0    0    0    0    0                   0    0    0    0
      2.1    1    0    0    0    0    0    0    9                   0    0    2    4
      2.15   0    0    0    0    0    0    0    0                   0    0    0    0
        alcohol
residual.sugar  9.5 9.55 9.56666666666667  9.6  9.7  9.8  9.9 9.95   10
      0.9    0    0                   0    0    0    0    0    0    0
      1.2    0    0                   0    0    0    2    0    0    1
      1.3    0    0                   0    0    0    1    0    0    0
      1.4    4    0                   0    4    2    0    0    0    3
      1.5    3    0                   0    2    0    5    3    0    0
      1.6    6    2                   0    1    1    3    3    0    3
      1.65   0    0                   0    0    0    0    0    0    0
      1.7    8    0                   0    0    3    3    0    0    9
      1.75   0    0                   0    0    0    0    0    0    0
      1.8   20    0                   0    5    0    8    7    1    3
      1.9   13    0                   0    6    8    6    4    0    5
      2     13    0                   0    7    3   10    9    0    7
      2.05   0    0                   0    0    0    0    0    0    0
      2.1   14    0                   0    4    5    6    6    0    5
      2.15   0    0                   0    0    0    0    0    0    0
        alcohol
residual.sugar 10.0333333333333 10.1 10.2 10.3 10.4 10.5 10.55 10.6 10.7 10.75 10.8
      0.9                    0    0    0    0    0    0     0    0    0     0    0
      1.2                    0    0    0    0    0    0     0    0    0     0    0
      1.3                    0    0    0    0    0    0     0    0    0     0    0
      1.4                    2    1    1    1    3    2     0    0    0     0    0
      1.5                    0    2    0    0    0    0     0    2    1     0    0
      1.6                    0    0    0    1    1    1     0    0    1     0    2
      1.65                   0    0    0    0    0    0     0    0    0     0    0
      1.7                    0    0    0    1    1    5     1    6    2     0    2
```

1.75	0	0	0	0	0	0	0	0	0	0	0
1.8	0	6	2	1	4	8	1	1	0	0	3
1.9	0	0	1	4	6	8	0	4	1	1	0
2	0	9	9	3	5	2	0	0	1	0	1
2.05	0	0	0	0	0	0	0	0	0	0	0
2.1	0	1	4	5	3	2	0	2	3	0	3
2.15	0	0	0	0	0	0	0	0	0	0	0

alcohol

residual.sugar	10.9	11	11.0666666666667	11.1	11.2	11.3	11.4	11.5	11.6	11.7	11.8	11.9
0.9	0	0	0	0	0	0	0	0	0	0	0	0
1.2	0	0	0	0	0	0	2	0	0	0	0	0
1.3	1	0	0	0	0	0	0	0	0	0	0	0
1.4	0	0	0	0	0	2	0	0	0	1	1	
1.5	4	2	0	1	0	0	0	0	0	0	0	0
1.6	1	0	0	2	2	0	1	1	0	2	0	1
1.65	0	0	0	0	0	0	1	0	0	0	0	0
1.7	2	0	0	0	1	3	2	2	1	0	1	2
1.75	0	0	0	0	0	0	0	0	0	1	1	0
1.8	2	4	0	1	3	1	2	2	1	1	3	4
1.9	5	3	0	2	3	0	1	2	0	0	2	0
2	6	1	0	2	6	5	4	3	2	3	1	3
2.05	0	0	0	0	0	0	0	0	0	2	0	0
2.1	6	7	0	3	4	2	7	1	0	0	3	0
2.15	0	0	0	0	0	1	0	0	0	0	0	0

alcohol

residual.sugar	11.95	12	12.1	12.2	12.3	12.4	12.5	12.6	12.7	12.8	12.9	13	13.1	13.2
0.9	0	0	0	0	0	0	0	0	2	0	0	0	0	0
1.2	0	0	0	0	0	0	2	0	0	0	0	1	0	0
1.3	0	0	0	0	0	0	0	0	0	1	0	0	0	0
1.4	0	0	0	1	1	0	0	0	0	0	1	0	0	0
1.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.6	1	0	1	1	0	0	0	0	0	0	0	0	0	0
1.65	0	0	0	0	0	0	0	0	0	0	1	0	0	0
1.7	0	0	0	0	1	0	1	0	0	0	1	0	0	0
1.75	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.8	0	0	1	0	1	1	1	0	0	2	0	0	0	0
1.9	0	1	0	0	1	0	1	1	0	0	0	0	0	0
2	0	1	0	1	1	1	1	0	0	0	0	0	0	0
2.05	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2.1	0	2	2	0	1	3	1	0	0	2	2	0	1	0
2.15	0	0	0	0	0	0	1	0	0	0	0	0	0	0

alcohol

residual.sugar	13.3	13.4	13.5	13.5666666666667	13.6	14	14.9	Sum
0.9	0	0	0	0	0	0	0	2
1.2	0	0	0	0	0	0	0	8
1.3	0	0	0	0	0	0	0	5
1.4	0	0	0	0	0	0	0	35
1.5	0	0	0	0	0	0	0	30
1.6	0	0	0	0	0	1	0	58
1.65	0	0	0	0	0	0	0	2
1.7	0	0	1	0	0	0	0	76
1.75	0	0	0	0	0	0	0	2

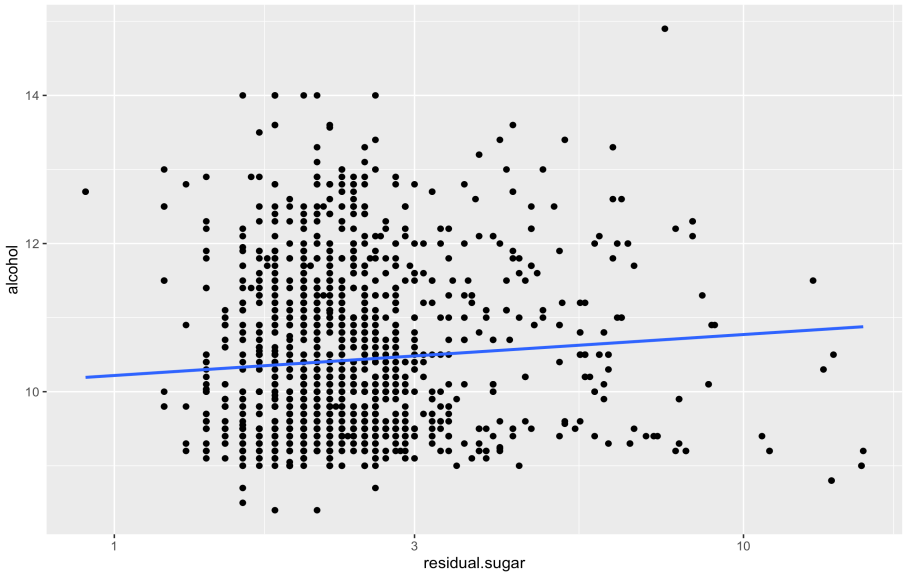
```
1.8    0   0   0          0   2   3   0 129
1.9    0   0   0          0   0   0   0 117
2      0   0   0          0   0   1   0 156
2.05   0   0   0          0   0   0   0   2
2.1    1   0   0          0   0   1   0 128
2.15   0   0   0          0   0   0   0   2
[ reached getOption("max.print") -- omitted 77 rows ]
```

```
chisq.test(mytable)
```

Pearson's Chi-squared test

data: mytable

X-squared = 11143, df = 5760, p-value < 2.2e-16



Alcohol ~ Residual Sugar: y ~ x: the result of given above is a scatter plot and a line of best fit, the scatter plot is made up of the black points where each point represents the results of physicochemical namely residual sugar and alcohol. So, for each point the x coordinate represent the residual sugar had 2.05 and each y coordinate represent the alcohol had 10.9, and then performed linear regression the resulting linear equation would produce the blue line in a linear positive, which we call the line of best fit, so the line of best fit in the line that best represent by given data for us to predict the quality of wine. That a prediction will always have margin of error.

```
mytable<- xtabs(~quality - alcohol, data = wineQualityReds)
head(table)
addmargins(mytable)
```

```
1 function (..., exclude = if (useNA == "no") c(NA, NaN), useNA = c("no",
2    "ifany", "always"), dnn = list.names(...), deparse.level = 1)
3 {
4    list.names <- function(...) {
5       l <- as.list(substitute(list(...)))[-1L]
6       nm <- names(l)
> addmargins(mytable)
     alcohol
quality  8.4  8.5  8.7  8.8    9 9.05  9.1  9.2 9.23333333333333 9.25  9.3  9.4  9.5
  3    1   0   0   0   1   0   0   0          0   0   0   0   0
  4    0   0   0   0   2   1   2   3          0   0   2   2   0
  5    0   1   0   2  11   0  14  50            0   0  44  79  97
  6    1   0   2   0  16   0   7  17            1   1  13  22  40
  7    0   0   0   0   0   0   0   2          0   0   0   0   2
  8    0   0   0   0   0   0   0   0          0   0   0   0   0
  Sum   2   1   2   2  30   1  23  72            1   1  59 103 139
     alcohol
quality 9.55 9.56666666666667  9.6  9.7  9.8  9.9 9.95   10 10.0333333333333 10.1 10.2
  3    0          0   0   1   1   1   1   0          0   0   1
  4    0          0   6   2   3   1   0   4          0   1   0
  5    1          0  38  35  49  25   0  29            0  23  21
  6    1          1  15  14  23  18   0  25            2  21  20
  7    0          0   0   2   1   4   0   8          0   2   4
  8    0          0   0   0   1   0   0   1          0   0   0
  Sum   2            1  59  54  78  49   1  67           2  47  46
     alcohol
quality 10.3 10.4 10.5 10.55 10.6 10.7 10.75 10.8 10.9   11 11.0666666666667 11.1 11.2
  3    0   0   0    0   0   1    0   0   1   1          0   0   0
  4    1   3   1    0   0   0    0   0   3   4          0   1   3
  5   13  12  31    0   8   7    0   9  13  19            0   7   8
  6   18  25  25    1  14  18    1  22  27  22            1  15  15
  7    1   1  10    1   6   1    0  11   5  11            0   4  10
  8    0   0   0    0   0   0    0   0   0   2          0   0   0
  Sum  33  41  67    2  28  27    1  42  49  59            1  27  36
     alcohol
quality 11.3 11.4 11.5 11.6 11.7 11.8 11.9 11.95   12 12.1 12.2 12.3 12.4 12.5 12.6
  3    0   0   0   0   0   0   0    0   0   0   0   0   0   0   0
  4    1   2   2   0   0   1   0    0   1   0   0   0   0   0   0
  5    5   9   3   1   1   2   1    1   1   1   1   0   0   0   1
  6   18  18  19   8   9  15  14    0  10   4   7   5   7  11   2
  7    7   2   6   6  11  10   5    0   9   8   4   7   6   9   2
  8    1   1   0   0   2   1   0    0   0   0   0   0   0   1   1
  Sum  32  32  30  15  23  29  20    1  21  13  12  12  13  21   6
     alcohol
quality 12.7 12.8 12.9   13 13.1 13.2 13.3 13.4 13.5 13.5666666666667 13.6   14 14.9
  3    0   0   0   0   0   0   0   0   0          0   0   0   0
  4    0   0   0   0   1   0   0   0   0          0   0   0   0
  5    0   1   2   4   0   0   0   0   0          0   0   0   1
```

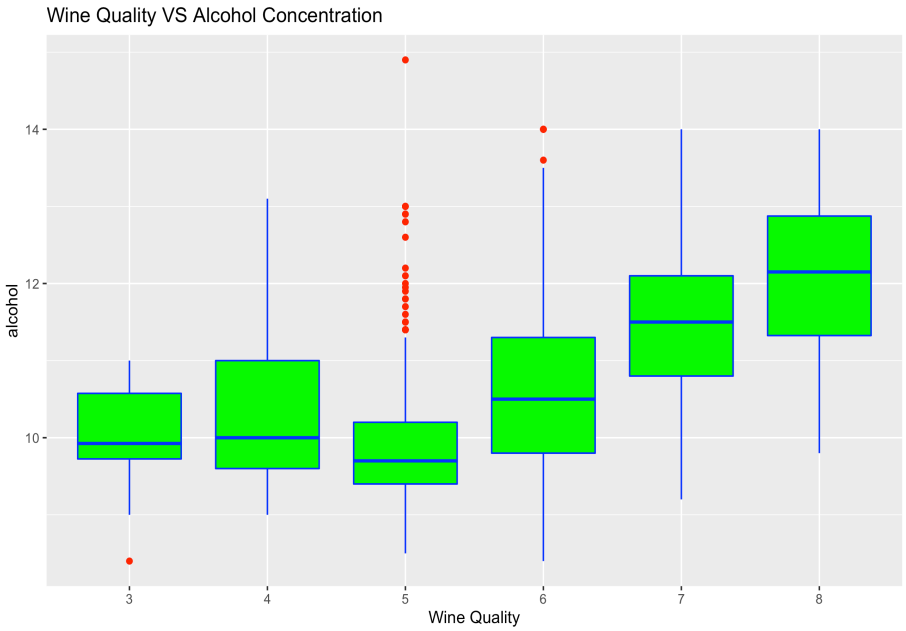
```
  6   6  8  3  0  0  1  2  1  1        0  1  4  0
  7   2  7  3  2  0  0  1  1  0        1  3  1  0
  8   1  1  1  0  1  0  0  1  0        0  0  2  0
  Sum9  17  9  6  2  1  3  3  1        1  4  7  1
    alcohol
quality  Sum
  3   10
  4   53
  5  681
  6  638
  7  199
  8   18
  Sum 1599
```

```
chisq.test(mytable)
```

Pearson's Chi-squared test

data: mytable

X-squared = 1124.5, df = 320, p-value < 2.2e-16



Wine Quality VS Alcohol concentration: As the above boxpot, the alcohol outlier more than 13.1%, maximum observation below upper fence had 14.9% in red dot as shown above, 75 percentiles of 14.9(upper quartile), Mean value of 0.52, Median had 50 percentiles of 10.20, 25 percentiles of 9.5 (lower quartile), and minimum observation of 8.40 below lower quartile.

Bivariate Linear Regression Analysis

The alcohol concentration is the most important factor to deciding the quality of red wine would be density the lower the density, the higher the alcohol concentration, and the higher the alcohol concentration the better the quality of wine.

The chi-squared test that all p-value < 2.2e-16 ($p<0.05$) we could reject the null hypothesis the variable pairs (quality & volatile acidity%), (pH & alcohol %),(chlorides content & alcohol %),(residual sugar & alcohol %), (wine quality & alcohol %) are not independent.

A.3. Multivariate Linear Regression

Multivariate Linear Regression: analysis is essentially similar to the simple linear model, with the exception that multiple independent variables are used in the model. The mathematical representation of multiple linear regression is:

$$Y = a + bX_1 + X_2 + X_3 + \epsilon$$

Where:

Y = Dependent Variable

X_1, X_2, X_3 = Independent (explanatory)Variables

a = Intercept

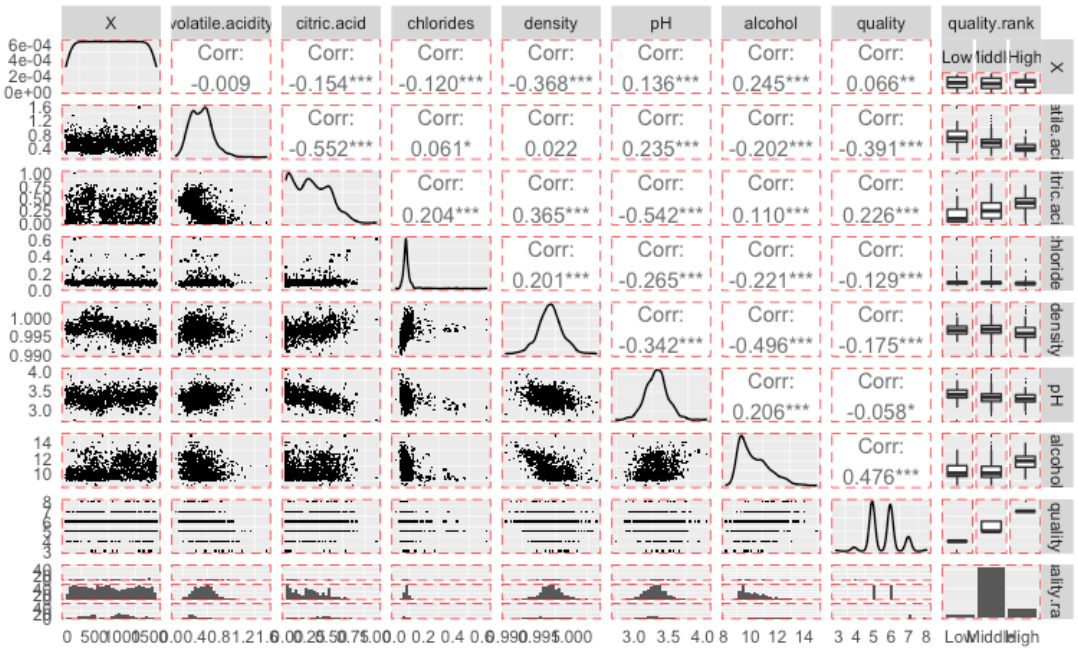
b, c, d = Slopes

ϵ = Residual (error)

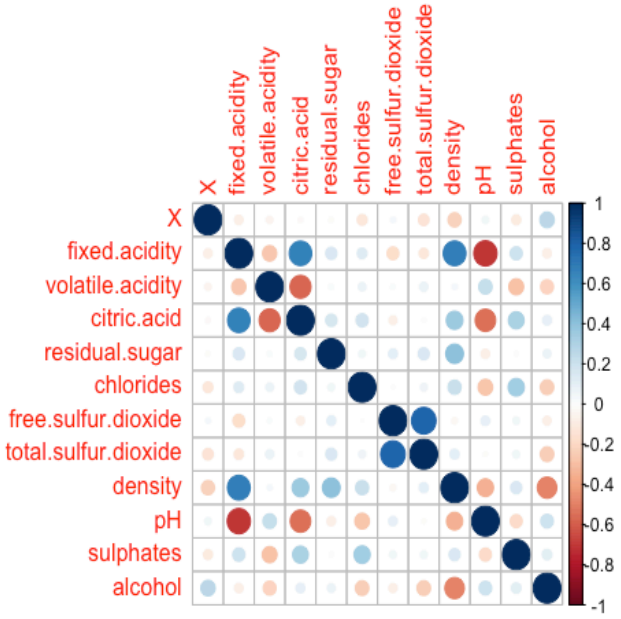
The multiple regression model is based on the following assumptions:

1. There is a linear relationship between the dependent variables and the independent variables.
2. The independent variable are not two highly correlated with each other.
3. Y_i observation is selected independently and randomly form the population.
4. Residual should be normally distributed with mean of 0 and variance σ.

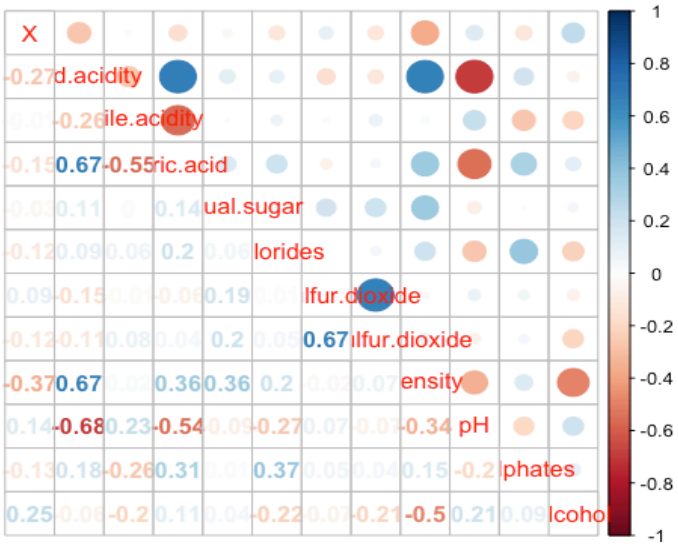
Pair wise plot This is where the pair wise plot come into the picture. As shown in the figure below, it allows the analysts to view all combinations of the variables, each in a two-dimensional plot. In this way, they can visualize all the relations and interactions among the variables on one single screen.



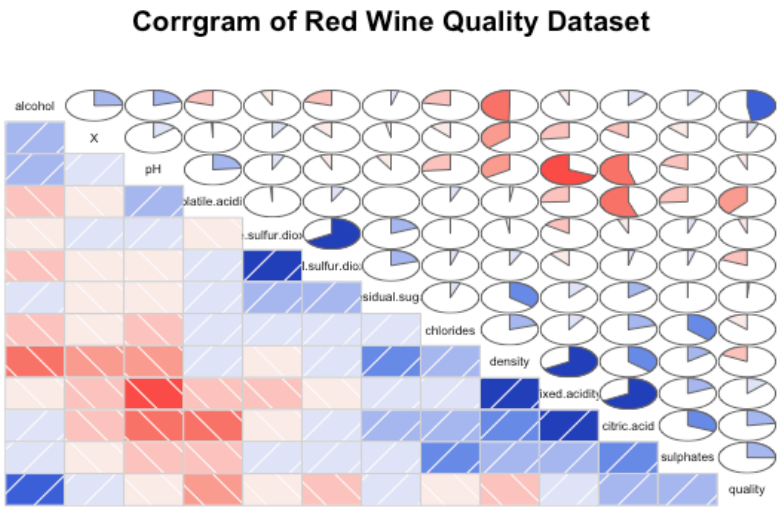
corrplot



Corrplot.mixed

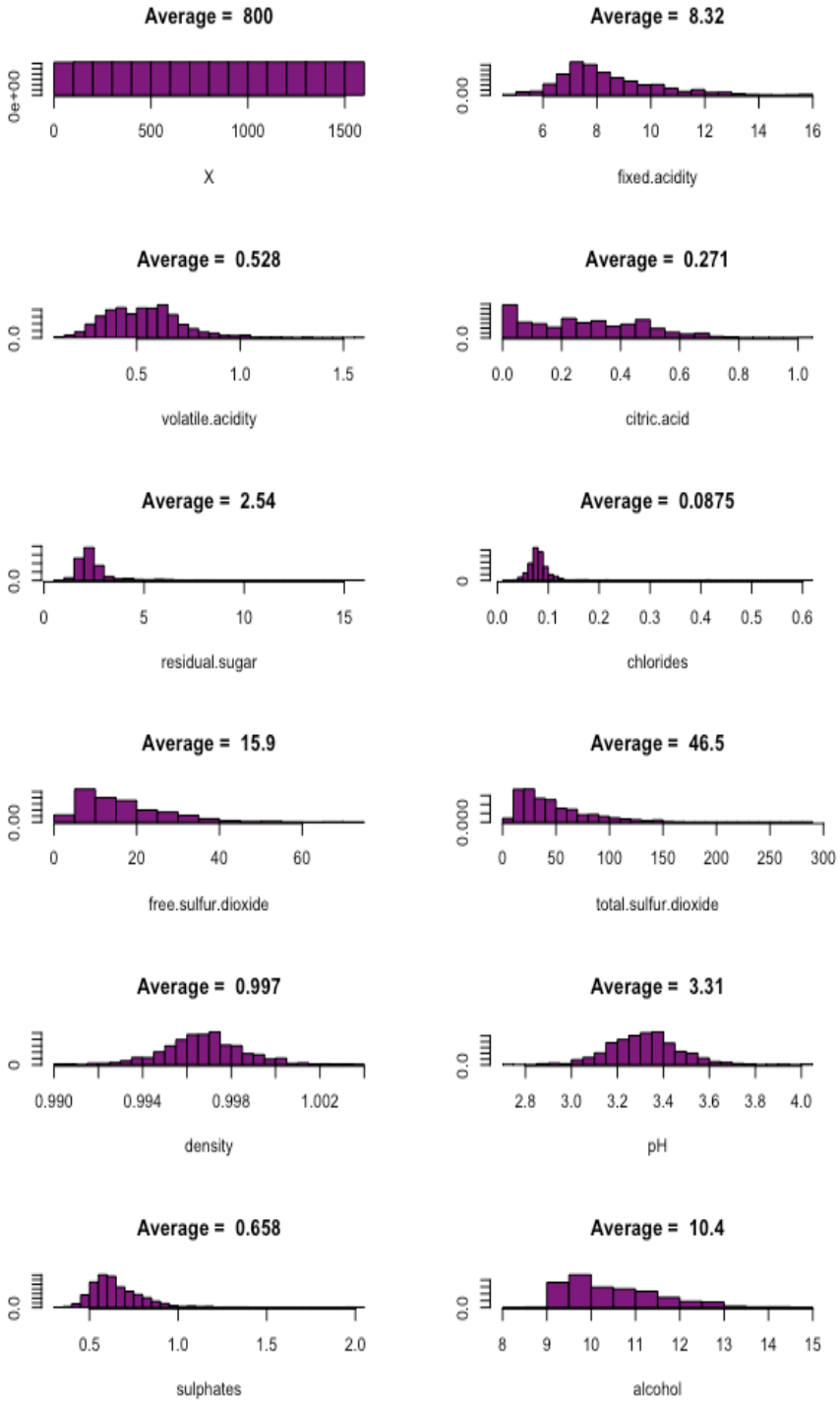


Corrgram: A corrgram, sometimes mistakenly referred to as correlogram, is just a visual display technique that helps us to represent the pattern of relations among a set of variables in terms of their correlations. Basically, a corrgram is a graphical representation of the cells of a matrix of correlations. The idea is to display the pattern of correlations in terms of their signs and magnitudes using visual thinning and correlation-based variable ordering. Moreover, the cells of the matrix can be shaded or colored to show the correlation value.



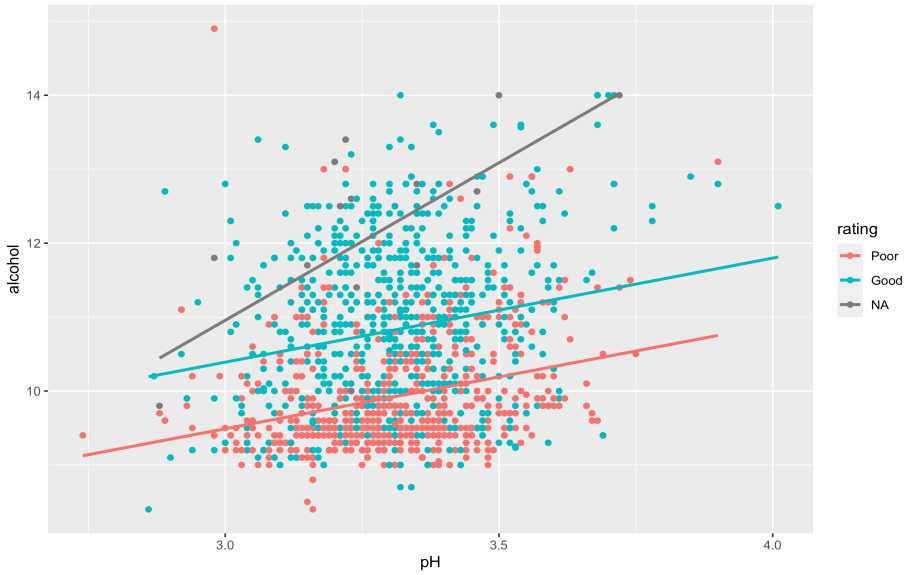
Corrgram provides another way to display the data, using both ellipses and loess lines. long, narrow ellipses represent high correlation while circular ellipses represent low correlation.

Average of Physicochemical properties

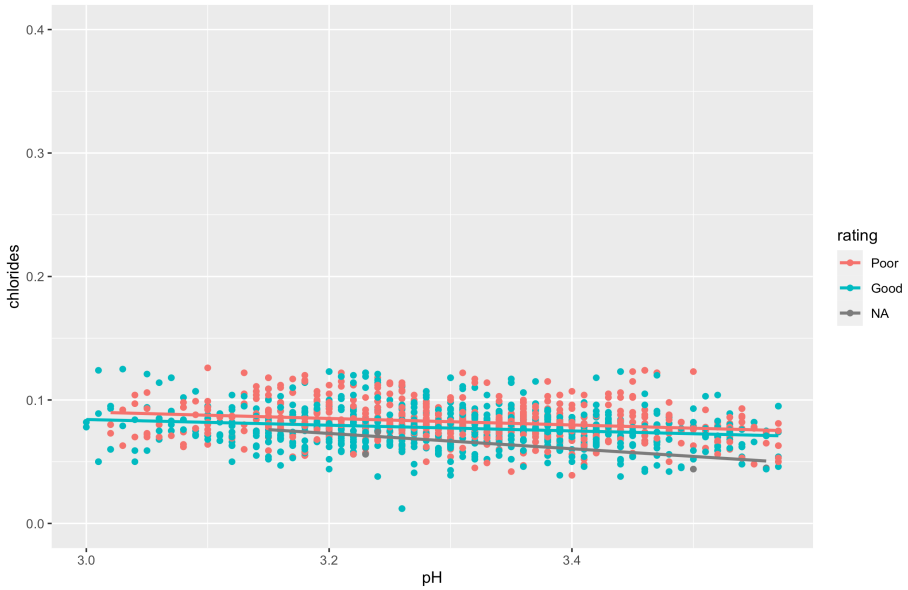


As histogram shown above the average of each physicochemical properties namely: (1) Fixed Acidity average of **8.32**, (2) Volatile Acidity average of **0.528**, (3) Citric Acid average of **0.271**, (4) Residual Sugar average of **2.54**, (5) Chlorides average of **0.087**,(6) Free Sulfur Dioxide average of **15.9**,(7) Total sulfur Dioxide **46.5**,(8) Density average of **0.997**,(9) pH average of **3.31**,(10) Sulphates average of **0.658**, and last (11)Alcohol average of **10.4**.

As shown a given scatter plot and Trend lines above. **The density & alcohol** the given results that a scatter plot is made up of the light blue and orange points where each point represents the results of physicochemical which are the density and alcohol rating. so, for each point the x - coordinate represent the density and y-coordinate represent the alcohol and then performed multiple linear regression resulting linear equation would produce trend lines, light blue represent “Good Rating”, orange represent “Poor Rating” and gray represent as excellent, but there’s no excellent rating as per given dataset, so the trend lines that best represent by given data for us to predict the ratings. Trend lines are in linear negative.



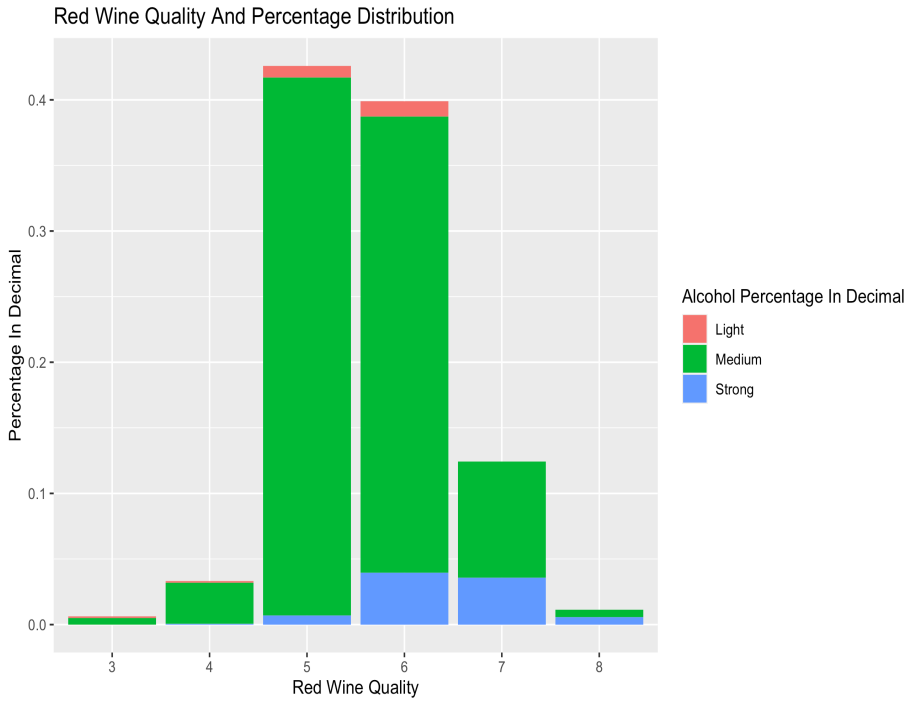
As shown a given scatter plot and Trend lines above. The **pH & alcohol** the given results that a scatter plot is made up of the light blue and orange points where each point represents the results of physicochemical which are the pH and alcohol rating. so, for each point the x - coordinate represent the pH and y-coordinate represent the alcohol and then performed multiple linear regression resulting linear equation would produce trend lines, light blue represent "Good Rating", orange represent "Poor Rating" and gray represent as excellent, but there's no excellent rating as per given dataset, so the trend lines that best represent by given data for us to predict the ratings. Trend lines are in linear negative.



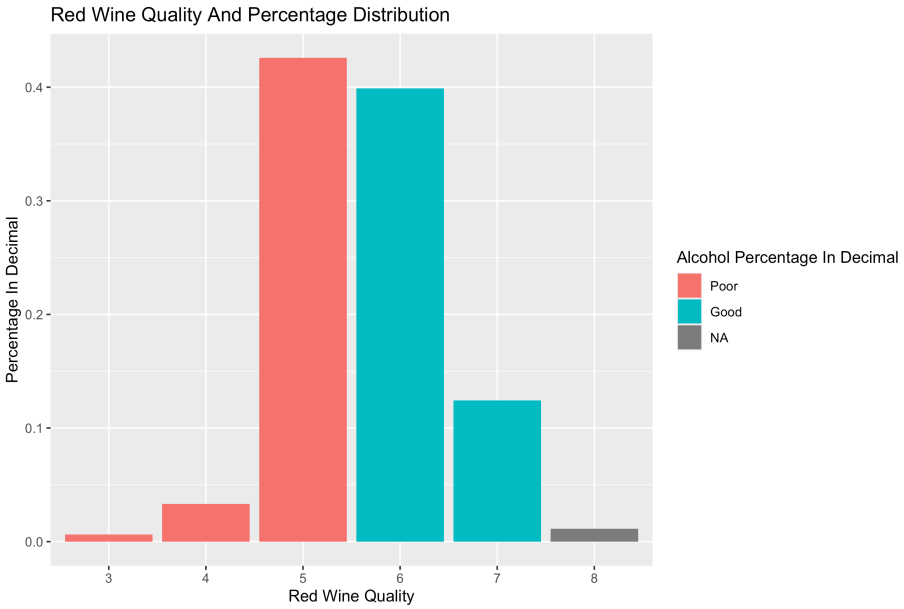
As shown a given scatter plot and Trend lines above. The **pH & chlorides** the given results that a scatter plot is made up of the light blue and orange points where each point represents the results of physicochemical which are the pH and chlorides rating. so, for each point the x - coordinate represent the pH and y-coordinate represent the chlorides and then performed multiple linear regression resulting linear equation would produce trend lines, light blue represent "Good Rating", orange represent "Poor Rating" and gray represent as excellent, but there's no excellent rating as per given dataset, so the trend lines that best represent by given data for us to predict the ratings. Trend lines are in No relationship between the pH and Chlorides.

Multivariate Linear Regression Analysis

Reject the null hypothesis because of its p-values, not all the columns are in significant variables. Multiple linear regression (MLR) model prediction of individual observation some regression trend lines are in negative and positive linear regression and no relationship at all.



As shown above the histogram of Red wine Quality in percentage in decimal distribution: the orange represents **"light"** had 0.37, the green represents **"medium"** had 14.21 and the blue represent a **"strong"** had 1.41 of alcohol label.



As shown above the histogram of Red Wine Quality in Percentage In Decimal Distribution: The orange represent of **"Poor"** had 7.44, the light blue represent **"Good"** had 8.37 and gray represent the "excellent" but there's no excellent rating in the given dataset.

Conclusion

The Red wine quality dataset contains 1,599 observation and total of twelve (12) attributes in the dataset, eleven (11) of the attributes are numeric physicochemical properties. The best result in regression model is quality:

Quality = -8.81 + 0.33(alcohol) -1.26(volatile acidity) + 12.17(density) - 0.72(chlorides) - 0.43(pH)

Therefore, the physicochemical properties are the best chemical substance in making a good wine. Future improvement of these dataset can be made a data collection on high quality and low quality wine if there is a significant correlation of chemical component in wine quality.