# Wine Quality Analysis

2023-02-16

## Contents


**Chapter 1: Introduction** **2**

**Chapter 2: Description of the data** **4**

Source Data . . . . . 4

Data Descriptive Summary . . . . . 4

Data Description . . . . . 5

Prior research and background on domain . . . . . 5

**Chapter 3: Smart Questions** **6**

**Chapter 4: Analyses** **6**

Wine Quality distribution . . . . . 6

Alcohol distribution . . . . . 7

Answering SMART Questions . . . . . 11

Does sugar level affect the alcohol level and how this relationship affects the quality? . . . . . 11

Are white wines more acidic compared to Red wines? And how does this affect the quality? . 17

Are there different acidity levels that are heavily correlated and how do they affect quality? . 19

Are there any independent variables that are highly correlated with each other ? . . . . . 25

Chloride . . . . . 26

Density . . . . . 27

Sulphates . . . . . 29

Conclusion . . . . . 30

**Modeling** **31**

Correlation test between Total sulfur dioxide and Free sulfur dioxide . . . . . 31

Correlation between Residual Sugar and Density . . . . . 32

Is alcohol level the same for red and white wine . . . . . 33

Volatile Acidity . . . . . 34

Point Biserial Correlation . . . . . 34

Correlation between independent variables and dependent variables . . . . . 34

Correlation between wine.type and other variables excluding quality . . . . . 35
Is density statistically significant between two wine types . . . . . 36

**Future Work** **42**

Connection between residual sugar, alcohol and quality . . . . . 42
Clustering . . . . . 44
Multicollinearity . . . . . 44

**References** **44**

```
library(ggplot2)
library(ggExtra)
library(gridExtra)
library(ezids)
library(lattice)
library(corrplot)
library(tidyverse)
library(latex2exp)
```

# Chapter 1: Introduction

Wine is an alcoholic beverage that has been produced for over thousands of years and was originally founded in Georgia from 6000BC. People consume wine for its wonderful sweet taste and nutritious properties that are beneficial to health. While there are many types of wines, two of the most famous variants would be the Red and White versions produced mostly in the Europe. Wines that were once expensive to buy, are increasingly enjoyed by consumers around the world and the market is expected to grow annually by 7.22% (CAGR 2022-2025)Statista. For this analysis, we are dealing with wine made in Vinho Verde, a region in Portuguese and is one of the top ten wine exporting countries with approximately 3% of the market share in 2005 and whose export has seen substantial growth between 1997 and 2007 Cortez et. al. The sales of this wine has reached a whopping €64 million just outside the Portugal market and one of the biggest importers of this wine would be The U.S The Portugal News

Wine industries are researching different methods involved in the process of creating and selling wines at the same time retaining its quality to support its growth. Wines are sold at different price ranges depending on its quality, while the least expensive could be as low as $4, the pinnacle of wines could cost beyond $200 per bottle winefolly. Given a bottle of wine could be sold for over $200, we decided to conduct analyses to better understand what gives rise to the quality by assessing publicly available dataset downloaded from UCI Machine Learning Repository. Perhaps this work will lay the foundation for future research analysis in understanding what defines a better quality **White** wine. This analysis is restricted to **White** wine although I do make comparisons between White and Red wines to have an overall better perception.

**Summary of the dataset after cleaning**

```
summary(wine.orig)
```

```
fixed.acidity volatile.acidity citric.acid residual.sugar
Min. :4.800 Min. :0.0800 Min. :0.100 Min. : 0.600
1st Qu.:6.300 1st Qu.:0.2100 1st Qu.:0.270 1st Qu.: 1.700
Median :6.800 Median :0.2600 Median :0.310 Median : 5.200
Mean :6.806 Mean :0.2658 Mean :0.325 Mean : 6.354
```

3rd Qu.:7.300 3rd Qu.:0.3200 3rd Qu.:0.380 3rd Qu.: 9.850
Max. :8.800 Max. :0.4850 Max. :0.570 Max. :22.000
NA's :119 NA's :186 NA's :270 NA's :7
chlorides free.sulfur.dioxide total.sulfur.dioxide density
Min. :0.01500 Min. : 2.00 Min. : 21 Min. :0.9871
1st Qu.:0.03500 1st Qu.:23.00 1st Qu.:108 1st Qu.:0.9917
Median :0.04200 Median :33.00 Median :134 Median :0.9937
Mean :0.04237 Mean :34.63 Mean :138 Mean :0.9940
3rd Qu.:0.04900 3rd Qu.:45.00 3rd Qu.:167 3rd Qu.:0.9961
Max. :0.07100 Max. :80.00 Max. :255 Max. :1.0024
NA's :208 NA's :50 NA's :19 NA's :5
pH sulphates alcohol quality
Min. :2.820 Min. :0.2200 Min. : 8.00 Min. :3.000
1st Qu.:3.080 1st Qu.:0.4100 1st Qu.: 9.50 1st Qu.:5.000
Median :3.170 Median :0.4700 Median :10.40 Median :6.000
Mean :3.183 Mean :0.4807 Mean :10.51 Mean :5.878
3rd Qu.:3.270 3rd Qu.:0.5400 3rd Qu.:11.40 3rd Qu.:6.000
Max. :3.560 Max. :0.7600 Max. :14.20 Max. :9.000
NA's :75 NA's :124

As we can now see, there are many NA's after removing the outliers. From what I have understood, I think it's best to replace NA's with 2Q (median) thereby not affecting our statistics.

```
tmp.cnames <- cnames
# Let's get rid of quality for now as it would not make sense to replace them
tmp.cnames <- tmp.cnames[-length(tmp.cnames)]
for(cname in tmp.cnames)
{
  tmp_median <- median(wine.orig[, cname], na.rm=T)
  wine.orig[, cname] <- wine.orig[, cname] %>% replace_na(tmp_median)
}

for(cname in tmp.cnames)
{
  tmp_median <- median(redwine.orig[, cname], na.rm=T)
  redwine.orig[, cname] <- redwine.orig[, cname] %>% replace_na(tmp_median)
}
```

**Summary of the dataset after replacing NA's by median**

```
summary(wine.orig)
```

fixed.acidity volatile.acidity citric.acid residual.sugar
Min. :4.800 Min. :0.0800 Min. :0.1000 Min. : 0.600
1st Qu.:6.300 1st Qu.:0.2100 1st Qu.:0.2700 1st Qu.: 1.700
Median :6.800 Median :0.2600 Median :0.3100 Median : 5.200
Mean :6.806 Mean :0.2656 Mean :0.3242 Mean : 6.352
3rd Qu.:7.300 3rd Qu.:0.3100 3rd Qu.:0.3700 3rd Qu.: 9.838
Max. :8.800 Max. :0.4850 Max. :0.5700 Max. :22.000
chlorides free.sulfur.dioxide total.sulfur.dioxide density
Min. :0.01500 Min. : 2.00 Min. : 21 Min. :0.9871
1st Qu.:0.03600 1st Qu.:23.00 1st Qu.:108 1st Qu.:0.9917
Median :0.04200 Median :33.00 Median :134 Median :0.9937
Mean :0.04236 Mean :34.62 Mean :138 Mean :0.9940

3rd Qu.:0.04900 3rd Qu.:45.00 3rd Qu.:167 3rd Qu.:0.9961
Max. :0.07100 Max. :80.00 Max. :255 Max. :1.0024
pH sulphates alcohol quality
Min. :2.820 Min. :0.2200 Min. : 8.00 Min. :3.000
1st Qu.:3.090 1st Qu.:0.4100 1st Qu.: 9.50 1st Qu.:5.000
Median :3.170 Median :0.4700 Median :10.40 Median :6.000
Mean :3.182 Mean :0.4804 Mean :10.51 Mean :5.878
3rd Qu.:3.270 3rd Qu.:0.5400 3rd Qu.:11.40 3rd Qu.:6.000
Max. :3.560 Max. :0.7600 Max. :14.20 Max. :9.000

```
wine <- wine.orig
redwine <- redwine.orig

wine$quality <- as.factor(wine$quality)
redwine$quality <- as.factor(redwine$quality)
```

# Chapter 2: Description of the data

## Source Data

The dataset contains 4898 instances with 11 explanatory and 1 dependent variable - *quality* all of which are numerical. Also, the dataset does not appear to contain any missing field which appeared to be a good fit for the analyses. The features in the dataset represent various chemical levels associated with different quality level. A glimpse of the dataset is as follows:

```
show(str(wine))
```

```
## 'data.frame':    4898 obs. of  12 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             : Factor w/ 7 levels "3","4","5","6",..: 4 4 4 4 4 4 4 4 4 4 ...
## NULL
```

## Data Descriptive Summary

```
library(psych)
describe(wine)
```

```
##                       vars    n   mean    sd median trimmed   mad   min    max
## fixed.acidity            1 4898   6.81  0.74   6.80    6.79  0.74  4.80   8.80
```

```
## volatile.acidity        2 4898   0.27  0.08   0.26    0.26  0.07  0.08   0.48
## citric.acid             3 4898   0.32  0.09   0.31    0.32  0.07  0.10   0.57
## residual.sugar          4 4898   6.35  4.95   5.20    5.79  5.34  0.60  22.00
## chlorides               5 4898   0.04  0.01   0.04    0.04  0.01  0.01   0.07
## free.sulfur.dioxide     6 4898  34.62 15.33  33.00   34.06 16.31  2.00  80.00
## total.sulfur.dioxide    7 4898 138.00 41.31 134.00  136.80 43.00 21.00 255.00
## density                 8 4898   0.99  0.00   0.99    0.99  0.00  0.99   1.00
## pH                      9 4898   3.18  0.14   3.17    3.18  0.13  2.82   3.56
## sulphates              10 4898   0.48  0.10   0.47    0.47  0.10  0.22   0.76
## alcohol                11 4898  10.51  1.23  10.40   10.43  1.48  8.00  14.20
## quality*               12 4898   3.88  0.89   4.00    3.85  1.48  1.00   7.00
##                       range skew kurtosis   se
## fixed.acidity          4.00 0.15    -0.07 0.01
## volatile.acidity       0.40 0.45    -0.10 0.00
## citric.acid            0.47 0.43     0.13 0.00
## residual.sugar        21.40 0.73    -0.51 0.07
## chlorides              0.06 0.12    -0.10 0.00
## free.sulfur.dioxide   78.00 0.33    -0.40 0.22
## total.sulfur.dioxide 234.00 0.24    -0.32 0.59
## density                0.02 0.25    -0.76 0.00
## pH                     0.74 0.21    -0.22 0.00
## sulphates              0.54 0.48    -0.03 0.00
## alcohol                6.20 0.49    -0.70 0.02
## quality*               6.00 0.16     0.21 0.01
```

## Data Description

| Variable | Description |
|---|---|
| fixed.acidity | nonvolatile acidity represents the acetic acid in wine |
| volatile.acidity | represents the amount of aromatic flavour |
| citric.acid | level of fresh flavor added to a finished wine |
| residual.sugar | amount of sugar left unfermented in a finished wine. This affects the sweetness of the wine |
| chlorides | Saltiness in wine |
| free.sulfur.dioxide | portion of sulfur dioxide that is free in wine |
| total.sulfur.dioxide | the portion of sulfur dioxide that is free plus the portion that is bound to other chemicals |
| density | level of consistency |
| pH | represents how acidic or basic the water is |
| sulphates | Amount of antimicrobials in wine to prevent oxidation and spoilage |
| alcohol | Alcohol level in wine |
| quality | represents rating for the quality of the wine |

## Prior research and background on domain

After thorough research, information collected from various sources has been put together in a table.

| Alcohol level | Typical ABV | Made in, for instance | Avg climate | Impact in harvesting | Acidity level | Sweetness |
|---|---|---|---|---|---|---|
| Low | 10-11.5% | Germany and Italy | Usually colder | Under ripeness | High | Semi-Sweet |
| Medium | 11.5-13.5% | Italy and CA | Warm | Under ripeness | High | Semi-Sweet |
| High | 20% | Portugal | Hot | Ripe | Low | Sweet |

As it can be inferred from the above table, grapes that are harvested before the ripe stage typically contain high acidic level that is reflected on the fresh flavor found in white wines. These wines are also low in alcohol level. The alcohol present in these wines are typically produced by the fermentation process during which the yeast added to the grape juice (white wine) converts the sugar in the grape into ethanol and carbon dioxide. So higher the level of sweetness in the wine, the higher the level of alcohol is. Now, it is not surprising why high alcohol wine, also known as hot wine, is considered better in quality and more expensive as it is high in sweetness and takes longer time for the grapes to ripe than those that are made with under-ripe grapes.

"Despite concerns over the pace of change, many independent producers are experimenting more now and concentrating predominantly on the quality of the wine rather than on quantity. This is leading to drier, richer and less spritzy styles with greater ripeness of fruit evident on the palate and improved export sales, particular to the US, with a growing base of new consumers. Whether or not we want to admit it, statistically, America does have a sweet tooth" vinepair

The association of alcohol level with wine quality is more profound with red wine. The yeast is not applied only to the grape juice instead to the entire grapes including skin, which gives red wine the color, aroma, and the sweet flavor it has. Acids are crucial in boosting the effects of sulfur dioxide $SO_2$, which ensures the wine has longer shelf life protecting it from becoming rotten. A good acidity level also fends off most unwanted bacteria, as these compounds are unable to survive in low pH solutions. vivino

It is also believed, without drifting far from the scope of this analysis, acidity and Ph levels have correlation between them.

# Chapter 3: Smart Questions

1. Does sugar level affect the alcohol level and how this relationship affects the quality?

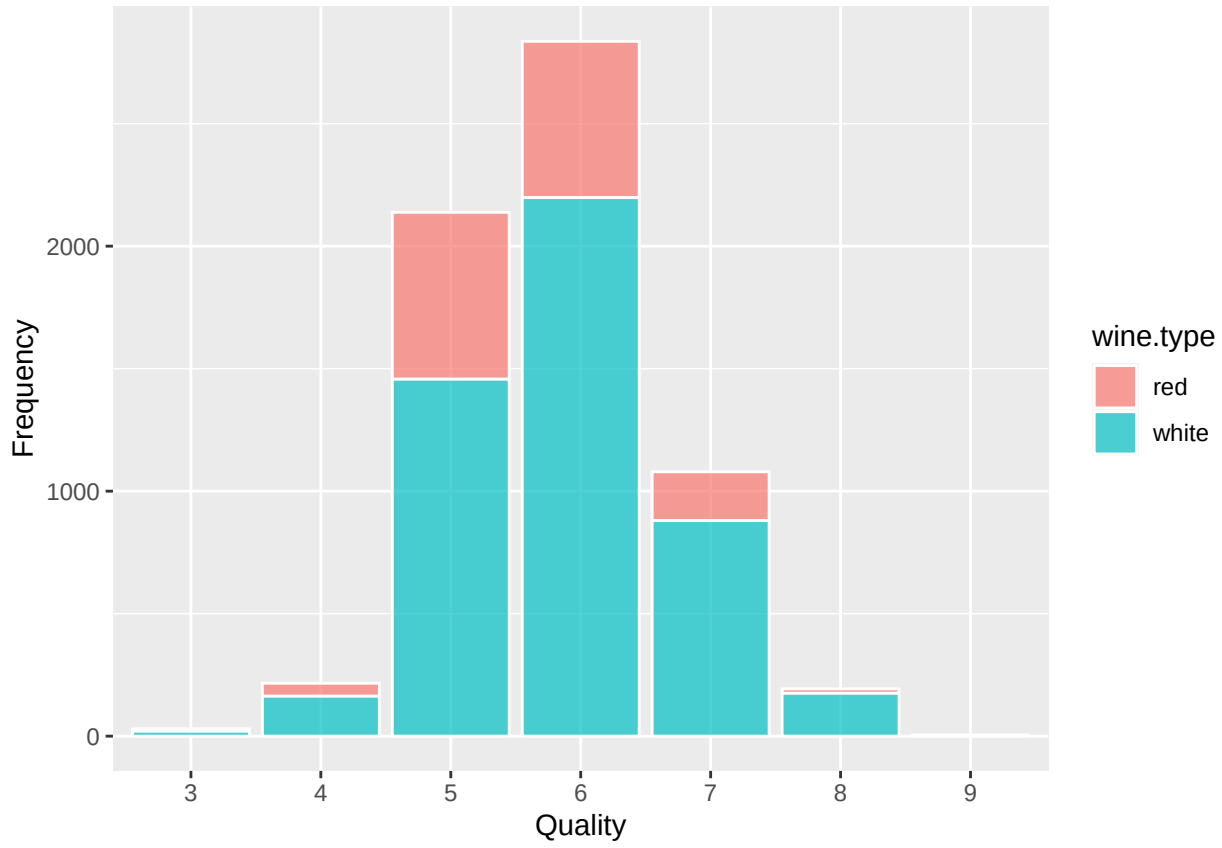
2. Are white wines more acidic compared to Red wines? And how does this affect the quality?

3. Are there different acidity levels that are heavily correlated and how do they affect quality?
4. Are there any independent variables that are highly correlated with each other ?

# Chapter 4: Analyses

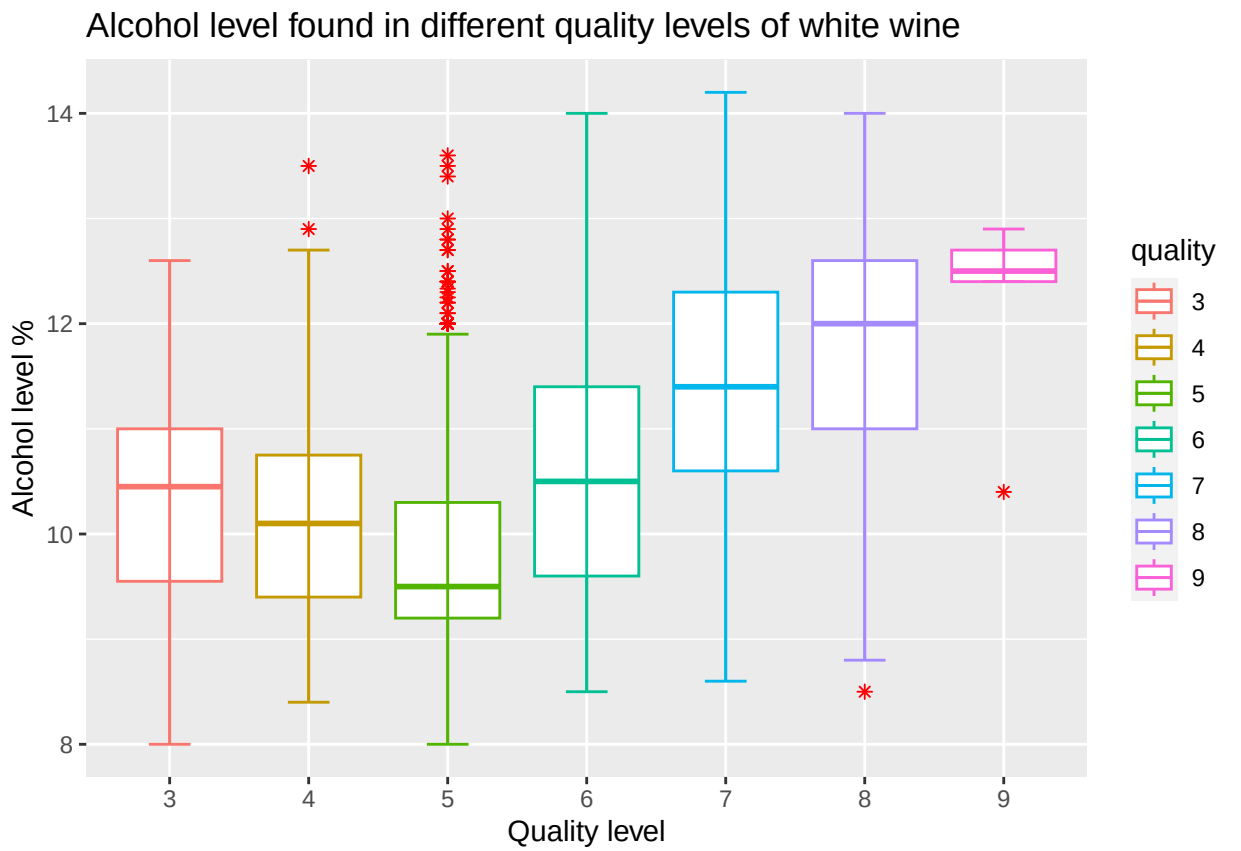
## Wine Quality distribution

```
# May be we should start using bar instead of histogram as this looks so ugly
ggplot(wine.collated, aes(quality, fill=wine.type)) +
  geom_bar(color='white', alpha=0.7) +
  xlab('Quality') +
  ylab('Frequency') +
  labs('Quality distribution')
```




### Alcohol distribution

```
# , fig.width=11, fig.height=4
bp.al <- ggplot(wine, aes(x=quality, y=alcohol, color=quality, group=quality)) +
         geom_boxplot(outlier.shape=8, outlier.color='red') +
         stat_boxplot(geom='errorbar', width=0.3) +
         xlab('Quality level') +
         ylab("Alcohol level %") +
         ggtitle('Alcohol level found in different quality levels of white wine')

bp.al
```

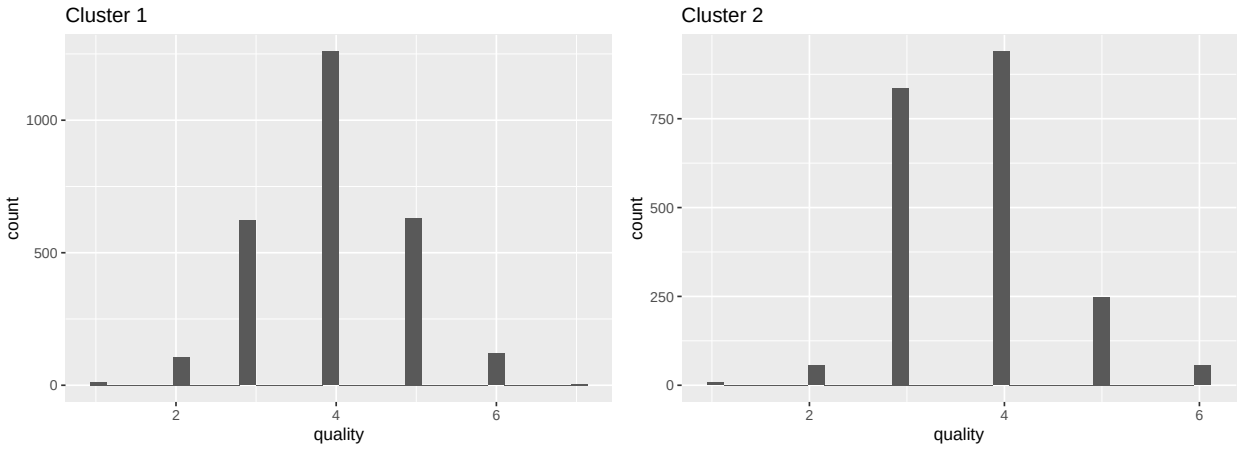



It is somewhat muddy to draw conclusion from the above plot as the first two quality levels are negatively correlated with alcohol, while the levels from 6 are positive correlated. We believe it would be best to group the quality levels so that we can compare the difference between the second quartile of low and high quality wines. Before proceeding to grouping, it is essential to first understand which levels to combine and the best solution we can think of now is to use a clustering model that will cluster data points that are closer in terms of Euclidean distance. Then any quality levels that are highly clustered can throw light onto finding a threshold to combine different group together.

```
library(class)

# Let's remove quality column so that kmeans can handle continuous data
tmp_cnames <- colnames(wine)
tmp_cnames <- tmp_cnames[-length(tmp_cnames)]

# I don't think 1000 iterations would be required, by the way, I wanted
# a more robust clustering. Hence, I chose a really high number
# For even better clustering, we should consider using more mature
# algorithms such as GMM that clusters datapoints of different statistical
# parameter such as not just mean, but also variance.
# For now, I think this is fine.
tmp_model <- kmeans(wine[, tmp_cnames], centers=2, iter.max = 1000)

ncluster1 <- sum(tmp_model$cluster == 1)
ncluster2 <- sum(tmp_model$cluster == 2)

cluster1 <- data.frame(matrix(nrow=ncluster1, ncol=length(colnames(wine))))
```

```
cluster2 <- data.frame(matrix(nrow=ncluster2, ncol=length(colnames(wine))))

colnames(cluster1) <- cnames
colnames(cluster2) <- cnames

c1_idx <- 1
c2_idx <- 1
idx <- 1
for(x in tmp_model$cluster)
{
  if(x == 1){
    cluster1[c1_idx,] <- wine[idx,]
    c1_idx <- c1_idx + 1
  }
  else{
    cluster2[c2_idx,] <- wine[idx,]
    c2_idx <- c2_idx + 1
  }
  idx <- idx + 1
}
```

```
cluster1 <- ggplot(cluster1, aes(x=quality, fill=quality)) +
            geom_histogram() +
            ggtitle('Cluster 1')

cluster2 <- ggplot(cluster2, aes(x=quality, fill=quality)) +
            geom_histogram() +
            ggtitle('Cluster 2')

grid.arrange(cluster1, cluster2, nrow=1)
```

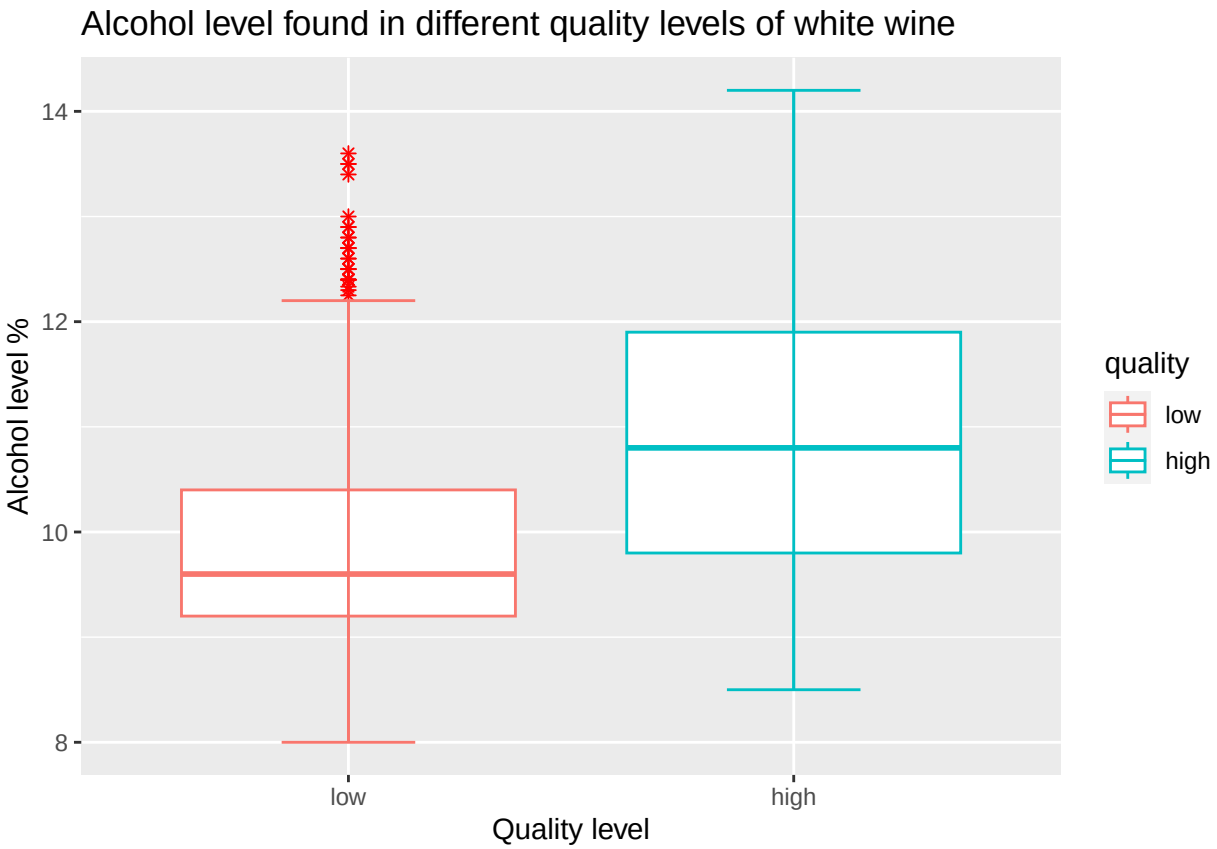



We did attempted KMeans, but the results were unsatisfactory so instead of grouping by KMeans clustering, we had decided to group levels by random binning technique. For any data points with quality level $<= 5$ will be categorized as low quality and for those that have quality $> 5$ will be grouped as high quality wines.

```
# Variable name wine refers to white wine, while redwine explicitly means red wine
# wine.collated means it has both white and red wine in it
# grouped term refers to quality 3-6 combined as low and 7-9 combined as high
```

```
# We use wine.orig to capture since quality in wine.orig is numeric instead of factors
bad <- wine[wine.orig[,'quality'] <= 5,]
good <- wine[wine.orig[,'quality'] > 5,]

bad$quality <- 'low'
good$quality <- 'high'
bad$quality <- as.factor(bad$quality)
good$quality <- as.factor(good$quality)

wine.grouped <- rbind(bad, good)
wine.grouped <- cbind(wine.grouped, rep('white', nrow(wine.grouped)))
cnames_wtype <- c(cnames, 'wine.type')
colnames(wine.grouped) <- cnames_wtype
# Since doing this at a later stage, I wish all analyses will reflect on this
# grouped data from this point onward.
wine <- wine.grouped

bad.redwine <- redwine[redwine.orig[,'quality'] <= 6,]
good.redwine <- redwine[redwine.orig[,'quality'] > 6,]

bad.redwine$quality <- 'low'
good.redwine$quality <- 'high'
bad.redwine$quality <- as.factor(bad.redwine$quality)
good.redwine$quality <- as.factor(good.redwine$quality)

redwine.grouped <- rbind(bad.redwine, good.redwine)
redwine.grouped <- cbind(redwine.grouped, rep('red', nrow(redwine.grouped)))
colnames(redwine.grouped) <- cnames_wtype
redwine <- redwine.grouped

wine.collated.grouped <- rbind(wine.grouped, redwine.grouped)
wine.collated.grouped$wine.type <- as.factor(wine.collated.grouped$wine.type)
```

```
# , fig.width=11, fig.height=4
bp.al <- ggplot(wine, aes(x=quality, y=alcohol, color=quality, group=quality)) +
         geom_boxplot(outlier.shape=8, outlier.color='red') +
         stat_boxplot(geom='errorbar', width=0.3) +
         xlab('Quality level') +
         ylab("Alcohol level %") +
         ggtitle('Alcohol level found in different quality levels of white wine')

bp.al
```

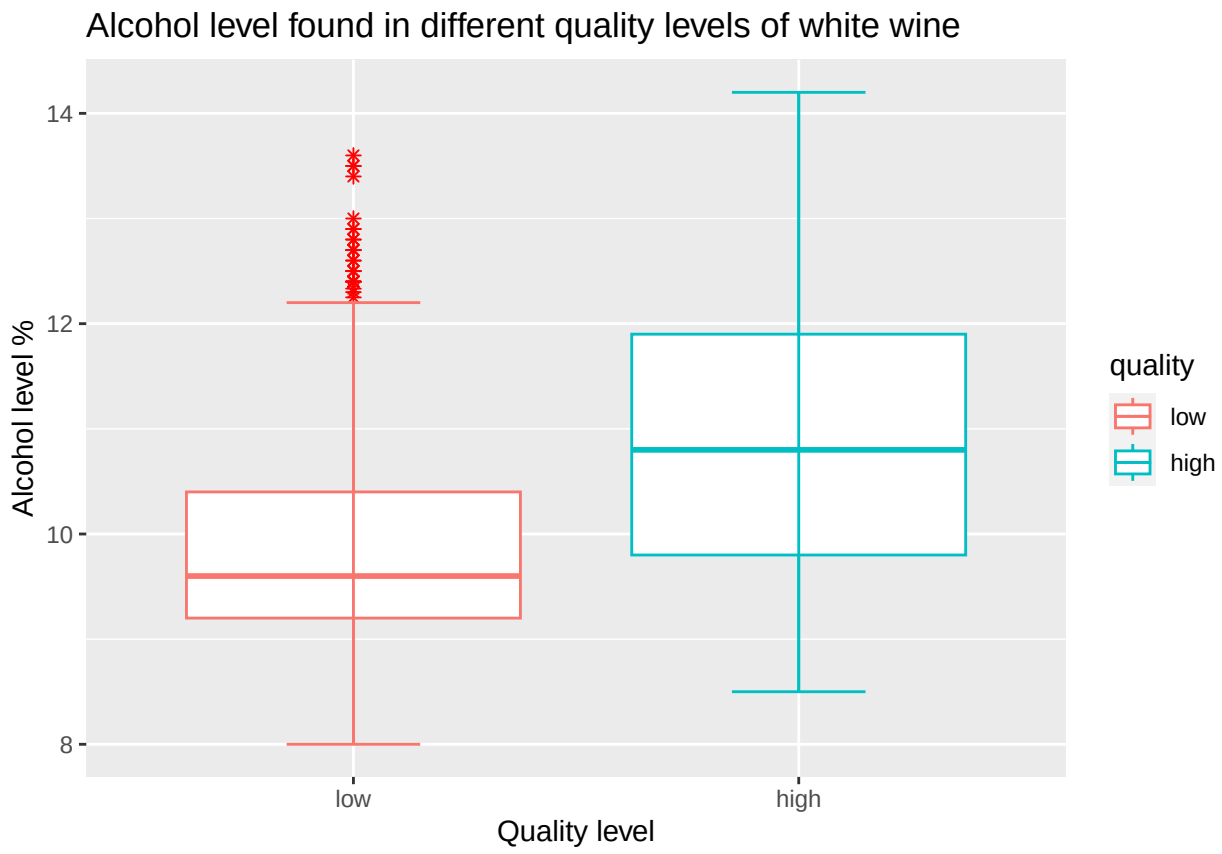



It is believed this plot makes it easier to draw conclusions on what level of ABV defines a better quality wine.

## Answering SMART Questions

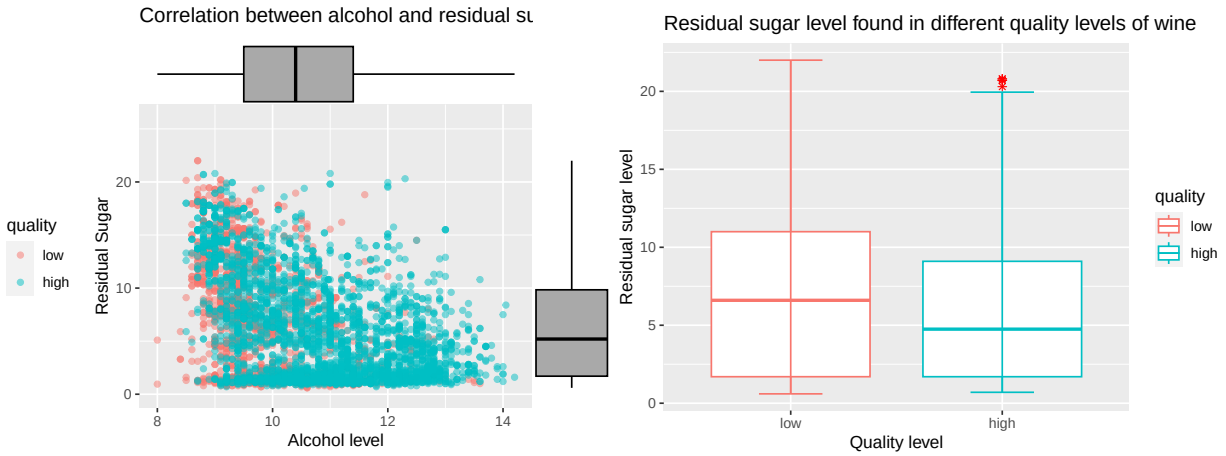
**Does sugar level affect the alcohol level and how this relationship affects the quality?**

```
bp.al
```




The above box and whisker plot confirms that better quality is indeed associated with increased level of alcohol level as was also our finding from the prior research. Now, it is time to understand the connection between alcohol level and residual sugar as it was understood from the prior research that alcohol is produced by the fermentation process by converting grape sugar into ethanol.

```
sp.al_rs <- ggplot(wine.grouped, aes(x=alcohol, y=residual.sugar, color=quality)) +
            geom_point(alpha=0.5) +
            xlab('Alcohol level') +
            ylab('Residual Sugar') +
            ggtitle('Correlation between alcohol and residual sugar levels') +
            ylim(0, 26) +
            theme(legend.position='left')
sp.al_rs <- ggMarginal(sp.al_rs, type='boxplot', fill='darkgray')

bp.rs <- ggplot(wine.grouped, aes(x=quality, y=residual.sugar, color=quality, group=quality)) +
         geom_boxplot(outlier.shape=8, outlier.color='red') +
         stat_boxplot(geom='errorbar', width=0.3) +
         xlab('Quality level') +
         ylab('Residual sugar level') +
         ggtitle('Residual sugar level found in different quality levels of wine')

grid.arrange(sp.al_rs, bp.rs, nrow=1)
```




From the macroscopic level, the correlation may not be very visible. Perhaps a correlation test can provide strong confidence on the correlation number as was taught in the lectures that lower correlation coefficient does not necessarily mean weak correlation. Only correlation test results can be reliable.

**Correlation between alcohol and residual sugar in high quality wines**

```
cor.test(good$alcohol, good$residual.sugar)
```

```
##
##  Pearson's product-moment correlation
##
## data:  good$alcohol and good$residual.sugar
## t = -31.523, df = 3256, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5094333 -0.4567975
## sample estimates:
##       cor
## -0.4835524
```

**Correlation between alcohol and residual sugar in low quality wines**

```
cor.test(bad$alcohol, bad$residual.sugar)
```

```
##
##  Pearson's product-moment correlation
##
## data:  bad$alcohol and bad$residual.sugar
## t = -19.815, df = 1638, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4779607 -0.3998358
## sample estimates:
##       cor
## -0.4397297
```

It is surprising that our expectation was not backed by the results conveyed by the plots above, but we think there are two things that must be held accountable for this finding:

1. Residual sugar is not the same as sugar level in the grapes.

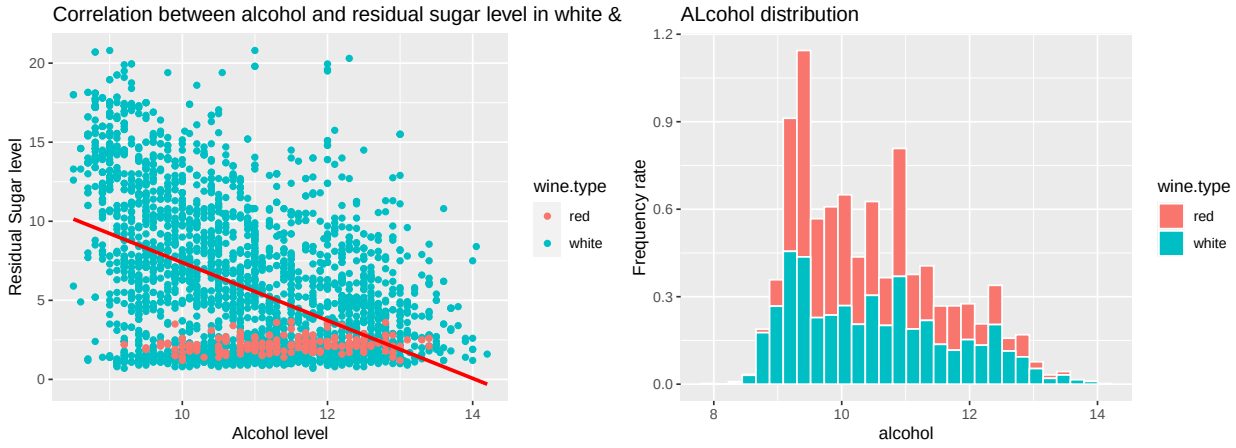
**Terminology Alert**
Residual sugar represents the amount of sugar that is left after the fermentation process.

We are not sure at this point whether this can directly represent the level of sugar in the wine and strongly believe this would be inversely proportional. What we mean by that is the residual sugar level does not necessary allow us to infer what the original sugar level was in the grape juice prior to the fermentation process. Hence, we believe it would be safe to presume that the lesser the residual sugar level is the higher the sugar got decomposed into alcohol and this is evident from the above plot. Also the correlation between alcohol level and the residual sugar is -0.4616576

2. Analyzing the how alcohol and residual sugar are correlated in red wine can provide more information.

```
good.collated.grouped <- wine.collated.grouped[wine.collated.grouped[, 'quality']=='high',]
bad.collated.grouped <- wine.collated.grouped[wine.collated.grouped[, 'quality']=='low',]
```

```
al_rs.model <- lm(residual.sugar~alcohol, data=good.collated.grouped)
alcohol_range <- seq(min(good.collated.grouped$alcohol), max(good.collated.grouped$alcohol), length.out=
tmp.df <- data.frame(alcohol=alcohol_range)
pred_rs <- predict(al_rs.model, tmp.df)
tmp.df <- data.frame(alcohol=alcohol_range, residual.sugar=pred_rs)

sp.al_rs <- ggplot(good.collated.grouped, aes(x=alcohol, y=residual.sugar, color=wine.type)) +
            geom_point() +
            geom_line(data=tmp.df, color='red', lwd=1.1) +
            xlab('Alcohol level') +
            ylab('Residual Sugar level') +
            ggtitle('Correlation between alcohol and residual sugar level in white & red wines')

alc.hist <- ggplot(wine.collated, aes(x=alcohol, y=..density.., fill=wine.type, group=wine.type)) +
            geom_histogram(color='white') +
            ylab('Frequency rate') +
            ggtitle('ALcohol distribution')

grid.arrange(sp.al_rs, alc.hist, nrow=1)
```



```
tmp_wine <- NULL # I don't want this to accidentally affect other results
```

**P.S: The red line in the scatter represents the fit for white wine alone.**

The scatter plot confirms the hypothesis we made in the last statement. As mentioned earlier, we should note that lighter white wines i.e., those with ABV <= 11.5% tend to have more residual sugar and exhibit fresher notes with high level of acidity. This is the case in particular with white wine and is evident from the scatter plot above that red wines contain less residual sugar compared to that found in white wine.

```
bp.rs <- ggplot(wine.grouped, aes(x=quality, y=residual.sugar, color=quality, group=quality)) +
         geom_boxplot(outlier.shape=8, outlier.color='red') +
         stat_boxplot(geom='errorbar', width=0.3) +
         xlab('Quality level') +
         ylab('Residual sugar level') +
         ggtitle('Residual sugar level in White Wine')

bp.rs.red <- ggplot(redwine.grouped, aes(x=quality, y=residual.sugar, color=quality, group=quality)) +
           geom_boxplot(outlier.shape=8, outlier.color='red') +
           stat_boxplot(geom='errorbar', width=0.3) +
           xlab('Quality level') +
           ylab('Residual sugar level') +
           ggtitle('Residual sugar level in Red Wine')

bp.rs.r.w <- ggplot(good.collated.grouped, aes(x=wine.type, y=residual.sugar, color=wine.type)) +
             geom_boxplot(outlier.shape=8, outlier.color='red') +
             stat_boxplot(geom='errorbar', width=0.3) +
             xlab('Quality level') +
             ylab('Residual sugar level') +
             ggtitle('Residual sugar level in Red and White')

grid.arrange(bp.rs, bp.rs.red, bp.rs.r.w, nrow=1)
```

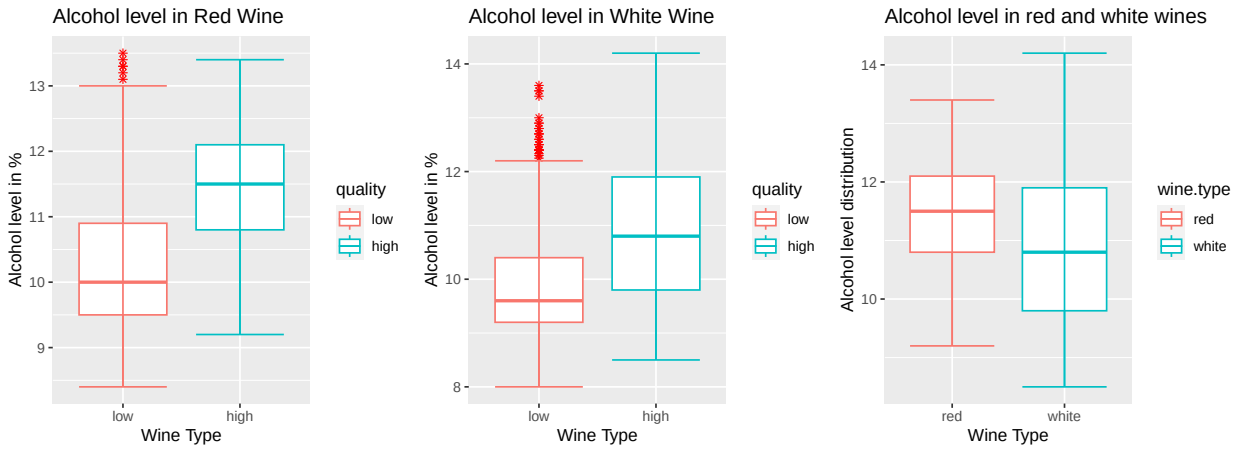

It is our hypothesis that high quality wines are always associated to wines with lesser residual sugar. I believe a hypothesis test `t.test` would shed more light onto this. Given a low quality wine has a mean residual sugar of 7.0326829 we would like to assess whether high quality wine's residual sugar level on average is lesser than that of low quality ones (white wine).

```
ttest.res.good.bad <- t.test(good$residual.sugar, mu=mean(bad$residual.sugar), alternative='greater', co
ttest.res.good.bad
```

```
##
##  One Sample t-test
##
## data:  good$residual.sugar
## t = -12.29, df = 3257, p-value = 1
## alternative hypothesis: true mean is greater than 7.032683
## 95 percent confidence interval:
##  5.873153      Inf
## sample estimates:
## mean of x
##  6.010052
```

And since the pvalue for `t.test` for residual sugar between two subgroups low and high quality is 1 I can conclude from the basis of hypothesis test results that good quality wines tend to have residual sugar less than 7.0326829.

```
bp.alc.red1 <- ggplot(redwine.grouped, aes(x=quality, y=alcohol, color=quality)) +
                geom_boxplot(outlier.shape=8, outlier.color='red') +
                stat_boxplot(geom='errorbar') +
                xlab('Wine Type') +
                ylab('Alcohol level in %') +
                ggtitle('Alcohol level in Red Wine')

bp.alc.white1 <- ggplot(wine.grouped, aes(x=quality, y=alcohol, color=quality)) +
                   geom_boxplot(outlier.shape=8, outlier.color='red') +
                   stat_boxplot(geom='errorbar') +
                   xlab('Wine Type') +
                   ylab('Alcohol level in %') +
                   ggtitle('Alcohol level in White Wine')

bp.alc.red.white1 <- ggplot(good.collated.grouped, aes(x=wine.type, y=alcohol, color=wine.type)) +
                   geom_boxplot(outlier.shape=8, outlier.color='red') +
                   stat_boxplot(geom='errorbar') +
                   xlab('Wine Type') +
                   ylab('Alcohol level distribution') +
                   ggtitle('Alcohol level in red and white wines')

grid.arrange(bp.alc.red1, bp.alc.white1, bp.alc.red.white1, nrow=1)
```




It is understood that red wine typically has more alcohol than the white wine Masterclass and the above plot confirms our hypothesis once again that red wines tend to have higher ABV compared to white wines. In terms of quality, ABV higher than 10.5% along with other properties exhibit better quality in wine.

Given a low quality wine has a mean ABV of 9.8495305% we would like to assess whether high quality wine's alcohol level on average is higher than that of low quality ones (white wine).

```
ttest.alc.good.bad <- t.test(good$alcohol, mu=mean(bad$alcohol), alternative='less', conf.level=0.95)
ttest.alc.good.bad
```

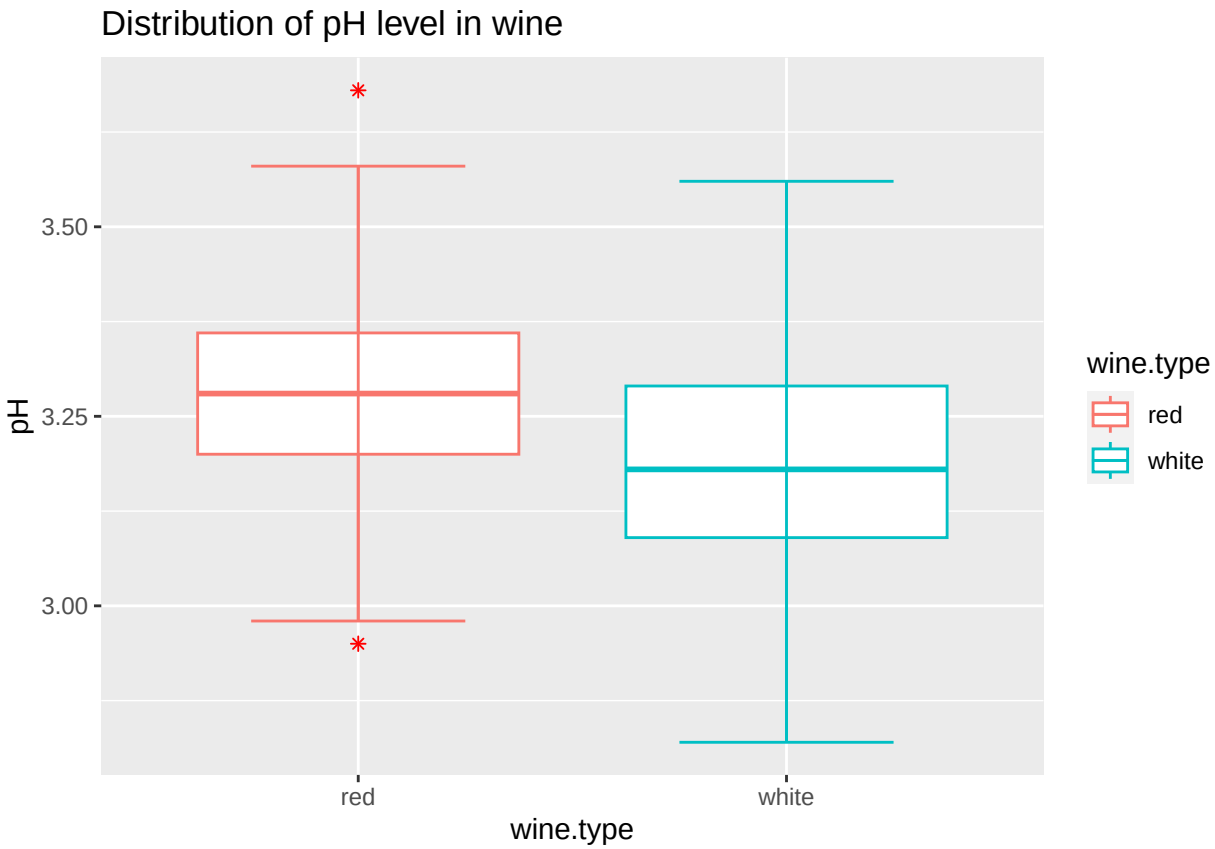
```
##
##  One Sample t-test
##
## data:  good$alcohol
## t = 45.727, df = 3257, p-value = 1
## alternative hypothesis: true mean is less than 9.84953
## 95 percent confidence interval:
##      -Inf 10.88484
## sample estimates:
## mean of x
##  10.84888
```

Statistical results once again confirm what we can from the boxplots that high quality wines on average tend to have ABV higher than 9.8495305% as the p value of the `t.test` was 1

**Are white wines more acidic compared to Red wines? And how does this affect the quality?**

**Terminology Alert**
pH level refers to how basic or acidic the liquid is. While a lower pH level means the liquid is more acidic, higher level represents how basic or alkaline the water is.

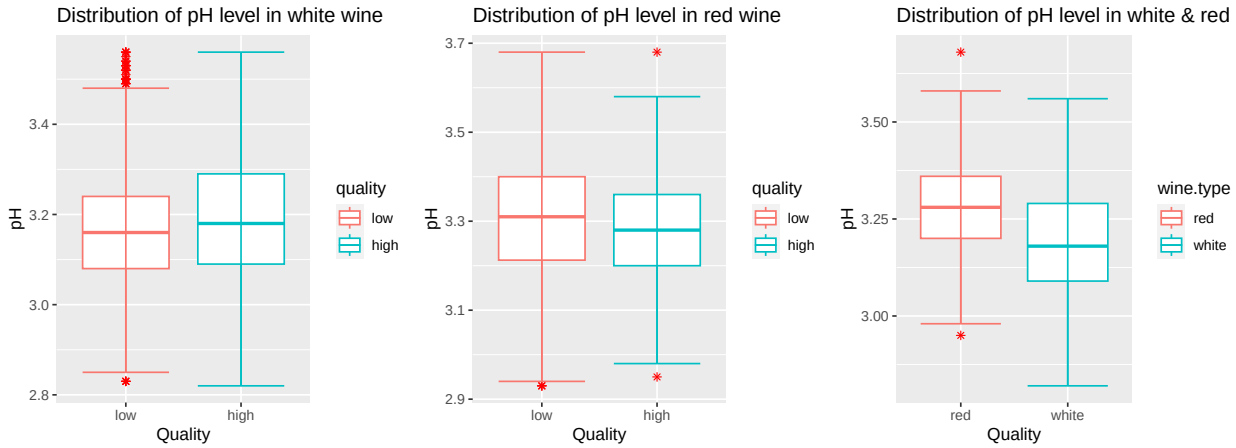
```
ggplot(good.collated.grouped, aes(x=wine.type, y=pH, color=wine.type)) +
geom_boxplot(outlier.shape=8, outlier.color='red') +
stat_boxplot(geom='errorbar', width=0.5) +
ggtitle('Distribution of pH level in wine')
```




The above boxplot between white and red wine confirms our hypothesis that white wine is more acidic compared to that of red wine, but what constitutes to the quality is the question. In order to answer, it would be first relevant to see how the pH level varies across different quality levels.

```
bp.ph.white <- ggplot(wine.grouped, aes(x=quality, y=pH, color=quality)) +
               geom_boxplot(outlier.shape=8, outlier.color='red') +
               stat_boxplot(geom='errorbar') +
               xlab('Quality') +
               ylab('pH') +
               ggtitle('Distribution of pH level in white wine')

bp.ph.red <- ggplot(redwine.grouped, aes(x=quality, y=pH, color=quality)) +
               geom_boxplot(outlier.shape=8, outlier.color='red') +
               stat_boxplot(geom='errorbar') +
               xlab('Quality') +
               ylab('pH') +
               ggtitle('Distribution of pH level in red wine')

tmp.wine.collated.grouped.good <- wine.collated.grouped[wine.collated.grouped[, 'quality']=='high',]

bp.ph.red.white <- ggplot(tmp.wine.collated.grouped.good, aes(x=wine.type, y=pH, color=wine.type)) +
                   geom_boxplot(outlier.shape=8, outlier.color='red') +
                   stat_boxplot(geom='errorbar') +
                   xlab('Quality') +
                   ylab('pH') +
                   ggtitle('Distribution of pH level in white & red')
```

```
grid.arrange(bp.ph.white, bp.ph.red, bp.ph.red.white, nrow=1)
```




It is commonly believed that many of the Americans prefer sweeter wines (vinepair)[https://vinepair.com/articles/sweet-wine-dry-culture/] and it makes sense why higher pH level on average is linked to high quality while low average pH is linked to low quality wines. We believe the process of winemaking has reflected on consumer preferences.

Given a low quality wine has a mean pH level of 3.164378 we would like to assess whether high quality wine's pH level on average is higher than that of low quality ones (white wine).

```
ttest.pH.good.bad <- t.test(good$pH, mu=mean(bad$pH), alternative='less', conf.level=0.95)
ttest.pH.good.bad
```

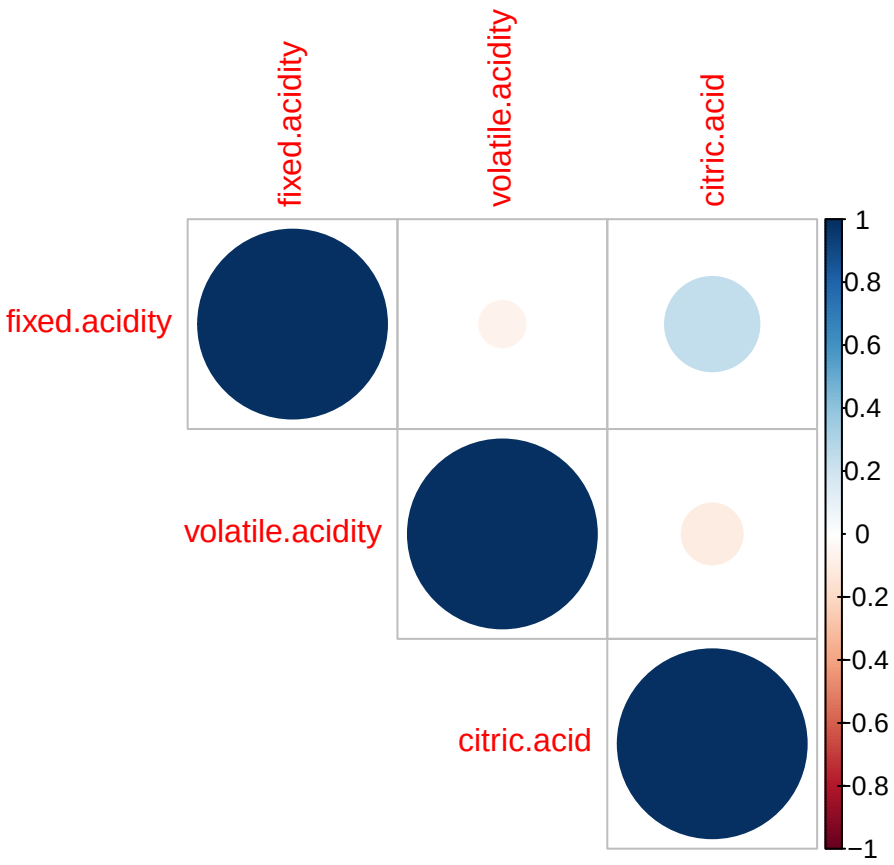
```
##
##  One Sample t-test
##
## data:  good$pH
## t = 10.848, df = 3257, p-value = 1
## alternative hypothesis: true mean is less than 3.164378
## 95 percent confidence interval:
##      -Inf 3.195671
## sample estimates:
## mean of x
##   3.19155
```

Statistical results once again confirm what we can from the boxplots that high quality wines on average tend to have pH level higher than 3.164378 as the p value of the `t.test` was 1

### Are there different acidity levels that are heavily correlated and how do they affect quality?

For the sake of correctness in extracting the correlation between variables, I have ensured the correlation plot below refers to the subgroup good quality wines alone.

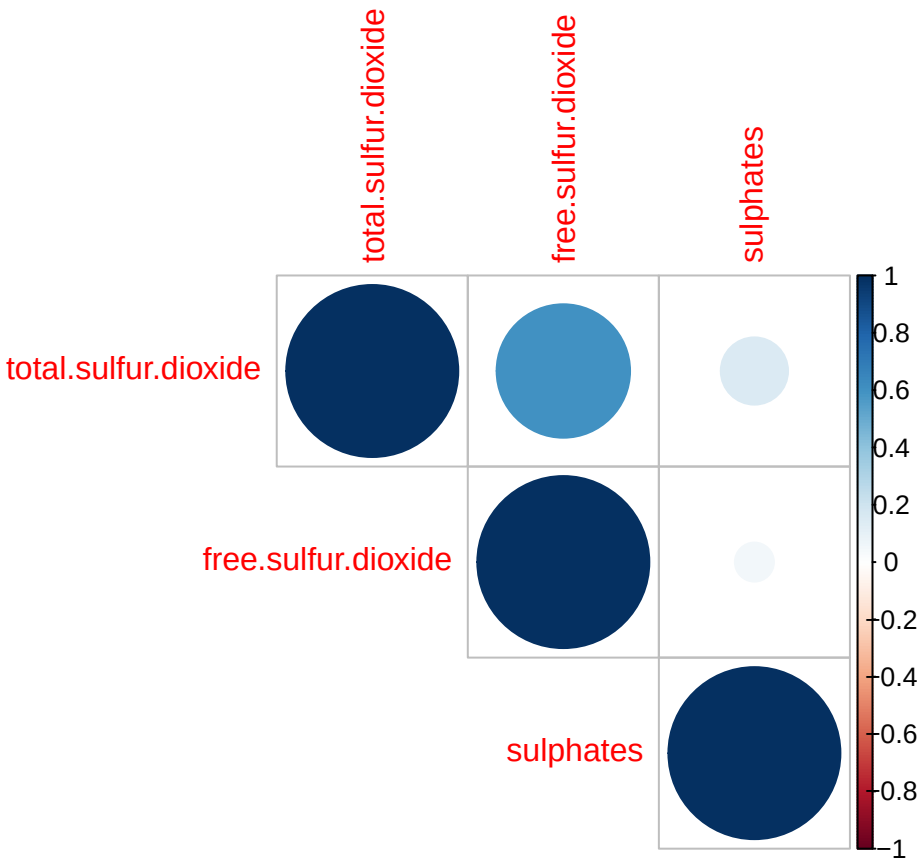
```
acidity_cnames <- c('fixed.acidity', 'volatile.acidity', 'citric.acid')
cplot <- corrplot(cor(good[, acidity_cnames]),
                  type='upper',
                  method='circle')
```




As mentioned earlier in the Data description section, fixed acidity refers to the level of vinegar like taste in the wine and it should be intuitive to think about fixed acidity being correlated with citric acidity as they both refer to bitter taste found in the wine.

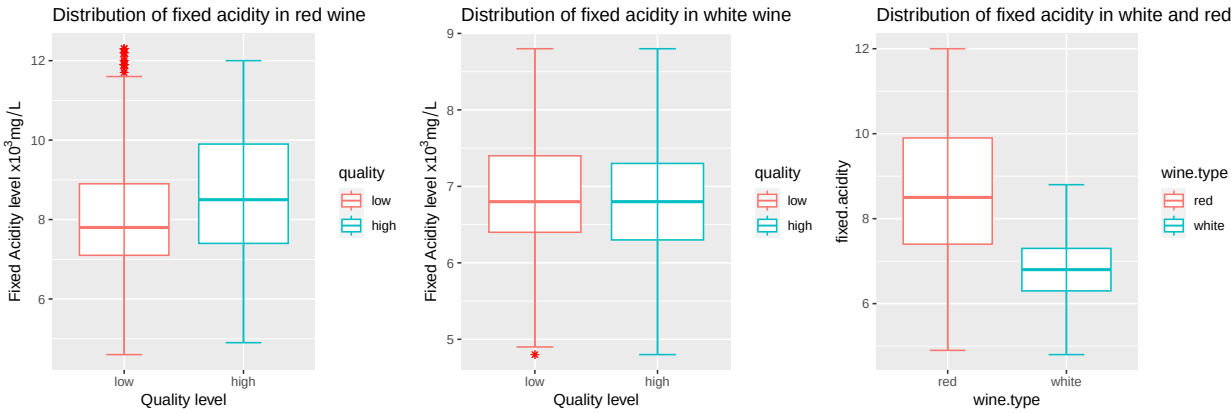
When it comes to total sulfur dioxide, free sulfur dioxide and sulphates, I personally found it a little daunting to understand these variables especially the distinction between them as most of the resources I referred to were either pointing to the same definition or chemistry-related explanation. However, in simple terms, Sulphates refer to the level of antimicrobial elements in the wine that prevents oxidation and spoilage. Oxidation is a process of altering a chemical's properties when exposed to oxygen $O_2$ over period of time. This may involve losing the wine's color, aromatic falvor, etc., Although it may make sense to perceive Sulphate as being correlated with the overall quality, it may not make much sense to conclude as far as the quality of taste is concerned given we consume the wine at a bar, where it is expected to finish on the same day it is opened.

```
sulfur_cnames <- c('total.sulfur.dioxide', 'free.sulfur.dioxide', 'sulphates')
s.cplot <- corrplot(cor(wine[, sulfur_cnames]),
                    type='upper',
                    method='circle')
```




Not surprised by the results, they do have a lot in common.

```
bp.fa <- ggplot(wine, aes(x=quality, y=fixed.acidity, color=quality, group=quality)) +
         geom_boxplot(outlier.shape=8, outlier.color='red') +
         stat_boxplot(geom='errorbar', width=0.3) +
         xlab('Quality level') +
         ylab(TeX('Fixed Acidity level $x 10^{3} mg/L $')) +
         ggtitle("Distribution of fixed acidity in white wine")


bp.fa.red <- ggplot(redwine, aes(x=quality, y=fixed.acidity, color=quality, group=quality)) +
             geom_boxplot(outlier.shape=8, outlier.color='red') +
             stat_boxplot(geom='errorbar', width=0.3) +
             xlab('Quality level') +
             ylab(TeX('Fixed Acidity level $x 10^{3} mg/L $')) +
             ggtitle("Distribution of fixed acidity in red wine")

bp.fa_wr <- ggplot(wine.collated.grouped[wine.collated.grouped[, 'quality']=='high',], aes(x=wine.type,
            geom_boxplot(outlier.shape=8, outlier.color='red') +
            stat_boxplot(geom='errorbar', width=0.3) +
            ggtitle('Distribution of fixed acidity in white and red')

grid.arrange(bp.fa.red, bp.fa, bp.fa_wr, nrow=1)
```




### Fixed Acidity

```
# bp.fa_wr
```

It is an interesting finding that high level of fixed acidity in red wine is high quality, while in case of white wine the 2Q levels are somewhat similar. And the last plot answers the question "whether red wine is more vinegar like in taste along with the sweet?" and as it appears red wines with more vinegar like taste and white wines with less vinegar like taste is considered better in quality.

```
ttest.fa.good.bad <- t.test(good$fixed.acidity, bad$fixed.acidity, var.equal=F)
ttest.fa.good.bad
```

```
##
##  Welch Two Sample t-test
##
## data:  good$fixed.acidity and bad$fixed.acidity
## t = -4.3116, df = 3362.8, p-value = 1.667e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1386984 -0.0519859
## sample estimates:
## mean of x mean of y
##  6.773987  6.869329
```

```
anova.fa.good.bad <- aov(fixed.acidity~quality, data=wine.grouped)
anova.fa.good.bad
```

```
## Call:
##    aov(formula = fixed.acidity ~ quality, data = wine.grouped)
##
## Terms:
##                   quality Residuals
## Sum of Squares     9.9162 2657.1952
## Deg. of Freedom         1      4896
##
## Residual standard error: 0.7367006
## Estimated effects may be unbalanced
```

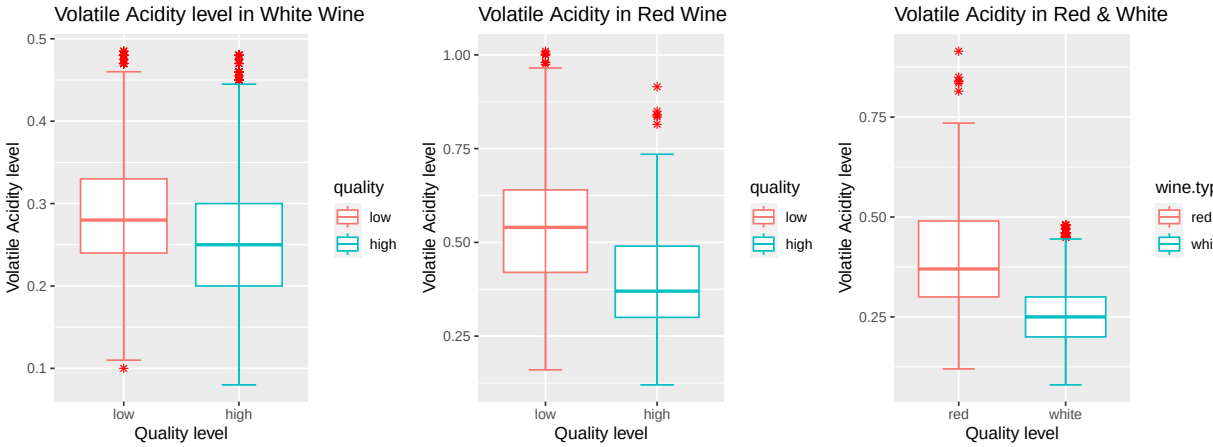
```
summary(anova.fa.good.bad)
```

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## quality        1    9.9   9.916   18.27 1.95e-05 ***
## Residuals   4896 2657.2   0.543
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

I wish not to go too much into explaining why I sometimes prefer running an `ANOVA` instead of a `t.test`, but superficially, `ANOVA` is more robust in terms of the assumptions generally made by the hypothesis testing algorithms. I do understand `t.test` is generally preferred over `ANOVA` when two groups are concerned and the flexibility in defining the alternative hypothesis.

Talking about the results, both `ANOVA` and `t.test` do point that quality does have an effect on the fixed acidity levels between the two subgroups low and high quality in **White** wines. For this particular test, I did not define an alternative hypothesis since I can see from the boxplots above for White Wine that both low and high quality wines tend to have almost similar levels of fixed acidity and I believe it would be appropriate to define the alternate hypothesis $H_1 :=$ Difference in mean between the two subgroups is not equal to 0.

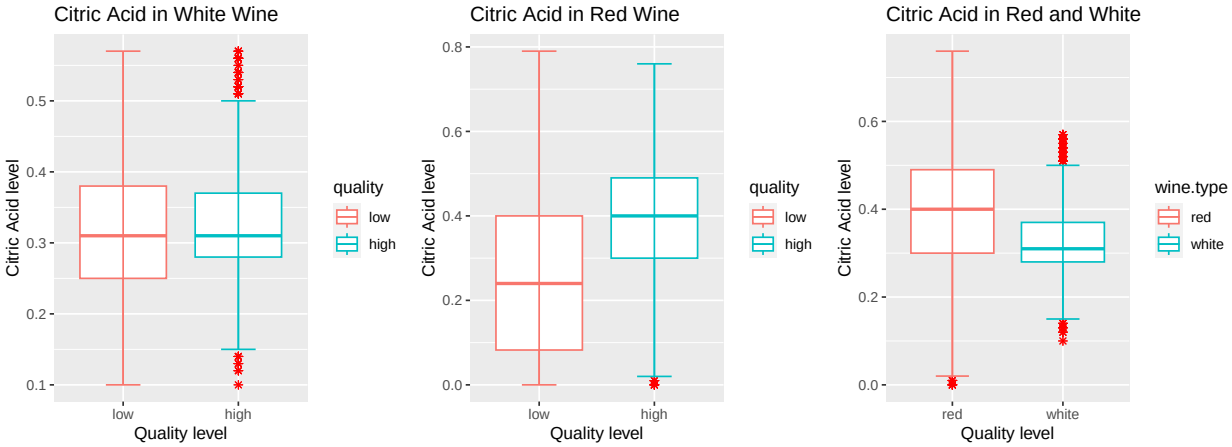
```
bp.va <- ggplot(wine, aes(x=quality, y=volatile.acidity, color=quality)) +
          geom_boxplot(outlier.shape=8, outlier.color='red') +
          stat_boxplot(geom='errorbar', width=0.3) +
           xlab('Quality level') +
           ylab('Volatile Acidity level') +
           #ylim(0.05, 1) +
           ggtitle('Volatile Acidity level in White Wine')

bp.va.red <- ggplot(redwine, aes(x=quality, y=volatile.acidity, color=quality)) +
              geom_boxplot(outlier.shape=8, outlier.color='red') +
              stat_boxplot(geom='errorbar', width=0.3) +
               xlab('Quality level') +
               ylab('Volatile Acidity level') +
               # ylim(0.05, 1) +
               ggtitle('Volatile Acidity in Red Wine')

bp.va.red.white <- ggplot(wine.collated.grouped[wine.collated.grouped[, 'quality']=='high',], aes(x=wine
                   geom_boxplot(outlier.shape=8, outlier.color='red') +
                   stat_boxplot(geom='errorbar', width=0.3) +
                    xlab('Quality level') +
                    ylab('Volatile Acidity level') +
                    # ylim(0.05, 1) +
                    ggtitle('Volatile Acidity in Red & White')

grid.arrange(bp.va, bp.va.red, bp.va.red.white, nrow=1)
```




### Volatile acidity

As mentioned earlier, too high or too low of volatile acidity can give bad smell and leave the consumer with an unpleasant experience. Perhaps this confirms our initial idea. I can see for both red and white wine that better quality wine always has relatively lesser volatile acidity level. In White Wine, the high quality wine is associated with volatile acidity less than 0.28. I wish not to run a hypothesis test as the difference is obvious.

```
bp.ca <- ggplot(wine.grouped, aes(x=quality, y=citric.acid, color=quality)) +
          geom_boxplot(outlier.shape=8, outlier.color='red') +
          stat_boxplot(geom='errorbar', width=0.3) +
         xlab('Quality level') +
         ylab('Citric Acid level') +
         ggtitle('Citric Acid in White Wine')

bp.ca.red <- ggplot(redwine.grouped, aes(x=quality, y=citric.acid, color=quality)) +
             geom_boxplot(outlier.shape=8, outlier.color='red') +
             stat_boxplot(geom='errorbar', width=0.3) +
             xlab('Quality level') +
             ylab('Citric Acid level') +
             ggtitle('Citric Acid in Red Wine')

bp.ca.red.white <- ggplot(wine.collated.grouped[wine.collated.grouped[, 'quality']=='high',], aes(x=wine
                  geom_boxplot(outlier.shape=8, outlier.color='red') +
                  stat_boxplot(geom='errorbar', width=0.3) +
                  xlab('Quality level') +
                  ylab('Citric Acid level') +
                  ggtitle('Citric Acid in Red and White')

grid.arrange(bp.ca, bp.ca.red, bp.ca.red.white, nrow=1)
```




### Citric acid

Once again I cannot see much difference in the 2nd quartile between the two subgroups (in White wine).

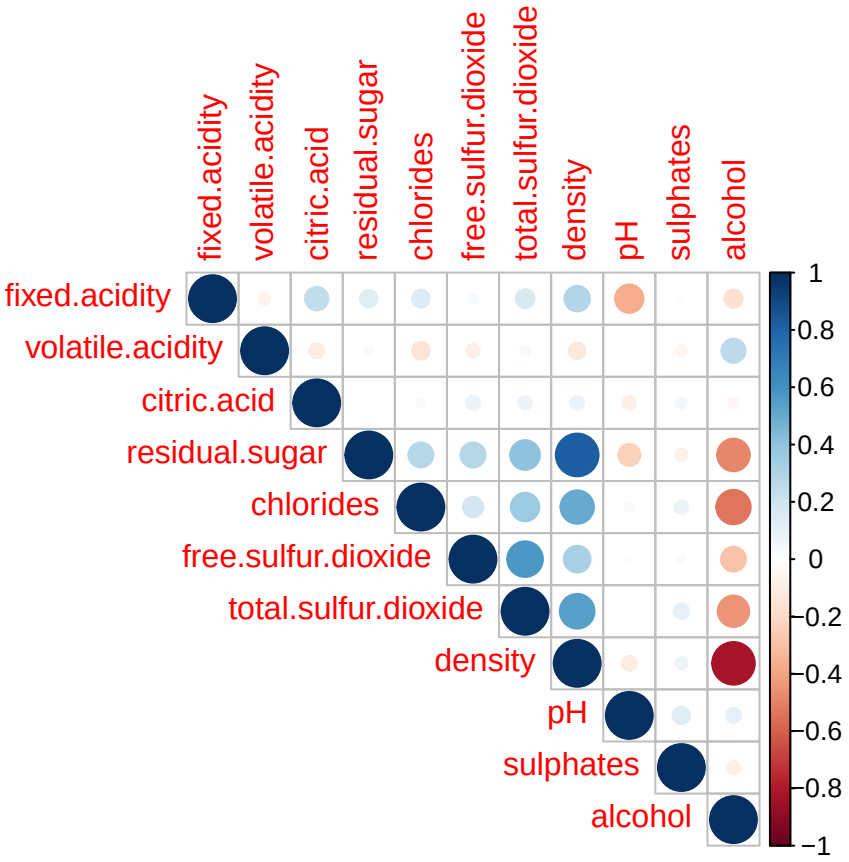
```
ttest.citric.good.bad <- t.test(good$citric.acid, bad$citric.acid, var.equal=F)
ttest.citric.good.bad
```

```
##
##  Welch Two Sample t-test
##
## data:  good$citric.acid and bad$citric.acid
## t = 1.4976, df = 2675, p-value = 0.1344
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.001317812  0.009838325
## sample estimates:
## mean of x mean of y
## 0.3256200 0.3213598
```

Since there was a confusion in understanding whether citric acid is similar to both subgroups low and high quality in White wine, I performed the hypothesis test that produced a pvalue of 0.1343546. Since the pvalue is higher than the significance threshold 0.05, I fail to reject the null hypothesis and conclude that there is no statistical significance between them. In other words, quality has no effect on the citric acid level in low and high quality **White** wines. In terms of the comparison between red and the white wine's citric acid levels, the above boxplot confirms our preconceived opinion that white wines tend to contain more citric acid in them as the grapes are harvested before they are fully ripe.

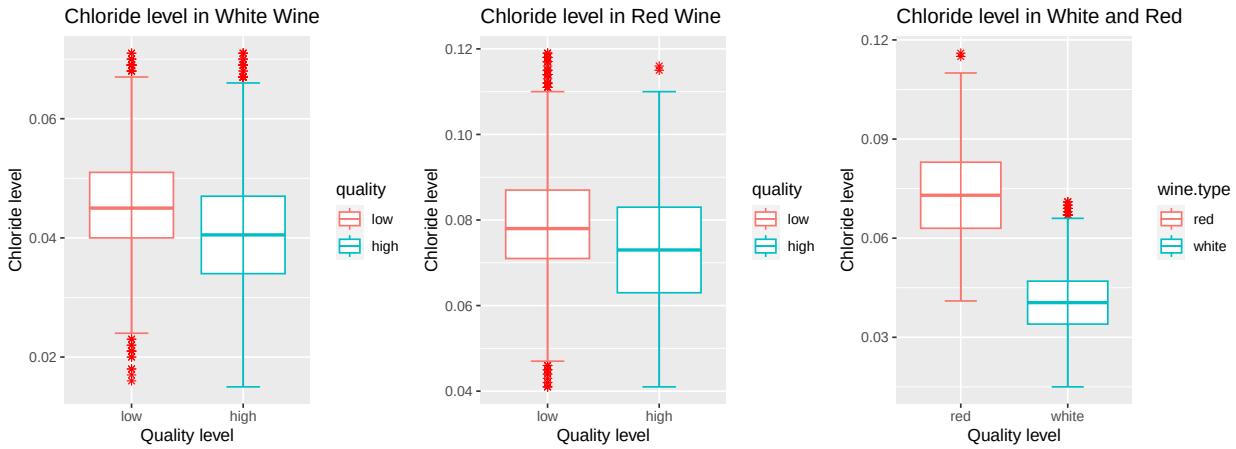
### Are there any independent variables that are highly correlated with each other ?

```
ind_cnames <- colnames(good)
ind_cnames <- ind_cnames[1:(length(ind_cnames)-1)]
corrplot(cor(good[, ind_cnames]),
         type='upper',
         method='circle')
```




### Chloride

```
bp.ch <- ggplot(wine.grouped, aes(x=quality, y=chlorides, color=quality)) +
          geom_boxplot(outlier.shape=8, outlier.color='red') +
          stat_boxplot(geom='errorbar', width=0.3) +
          xlab('Quality level') +
          ylab('Chloride level') +
          ggtitle('Chloride level in White Wine')

bp.ch.red <- ggplot(redwine.grouped, aes(x=quality, y=chlorides, color=quality)) +
              geom_boxplot(outlier.shape=8, outlier.color='red') +
              stat_boxplot(geom='errorbar', width=0.3) +
              xlab('Quality level') +
              ylab('Chloride level') +
              ggtitle('Chloride level in Red Wine')

bp.ch.red.white <- ggplot(wine.collated.grouped[wine.collated.grouped[, 'quality']=='high',], aes(x=wine
                   geom_boxplot(outlier.shape=8, outlier.color='red') +
                   stat_boxplot(geom='errorbar', width=0.3) +
                   xlab('Quality level') +
                   ylab('Chloride level') +
                   ggtitle('Chloride level in White and Red')

grid.arrange(bp.ch, bp.ch.red, bp.ch.red.white, nrow=1)
```




The above boxplot suggest that high quality wines tend to have less salty taste.

```
ttest.chloride.good.bad <- t.test(good$chlorides, mu=mean(bad$chlorides), alternative='greater')
ttest.chloride.good.bad
```

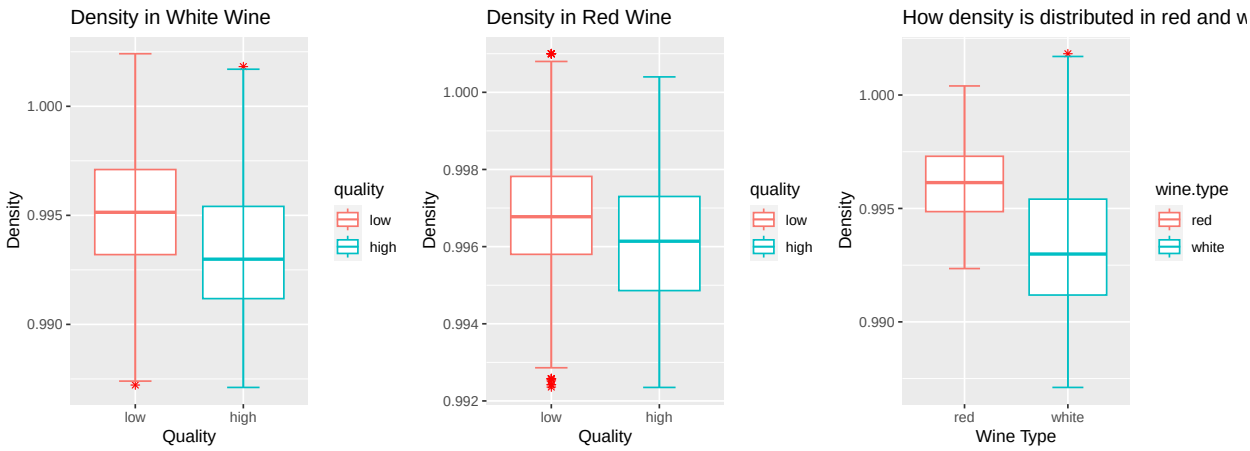
```
##
##  One Sample t-test
##
## data:  good$chlorides
## t = -28.582, df = 3257, p-value = 1
## alternative hypothesis: true mean is greater than 0.04556463
## 95 percent confidence interval:
##  0.04046716        Inf
## sample estimates:
##  mean of x
## 0.04074463
```

As the pvalue for the above `t.test` is 1 that is greater than significance threshold 0.05, I conclude that on average chlorides level in better quality White wine must be $< 0.0455646$.

## Density

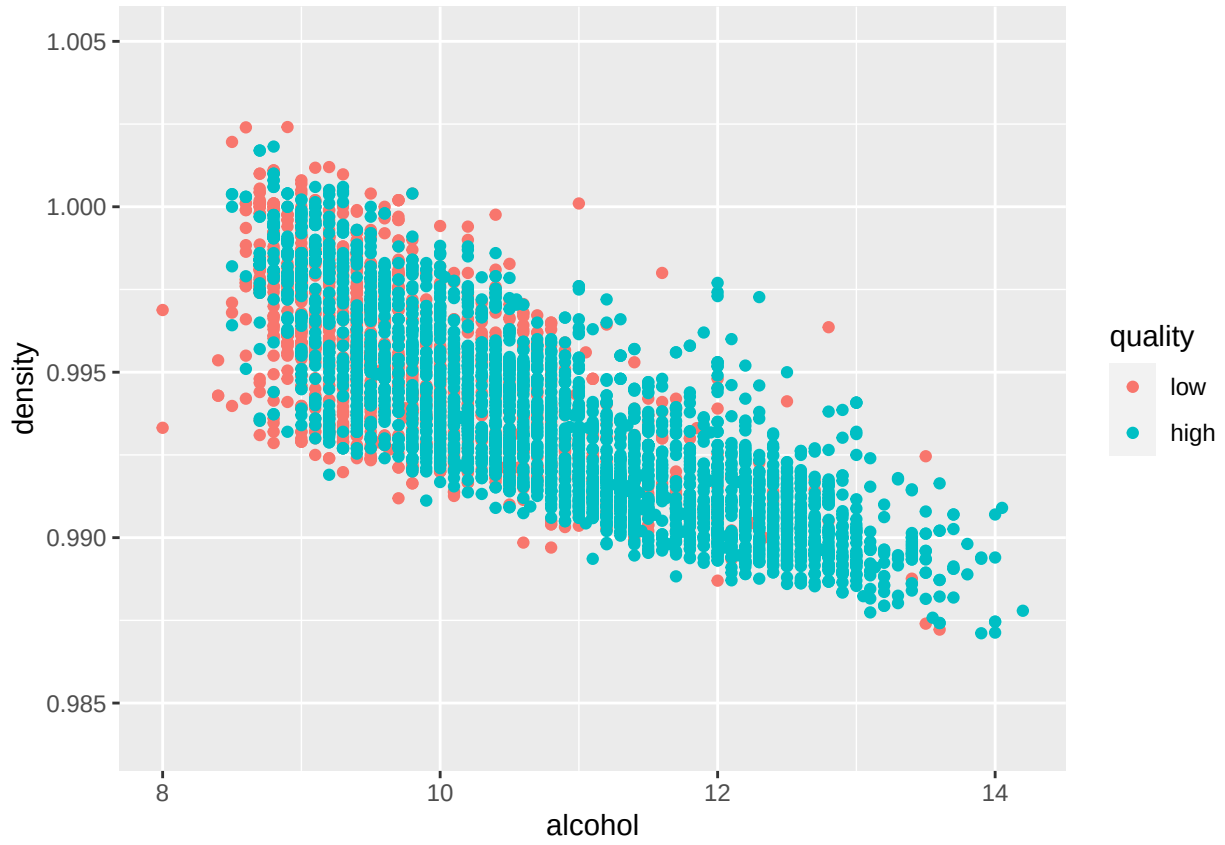
```
bp.density.w.quality <- ggplot(wine, aes(x=quality, y=density, color=quality)) +
                        geom_boxplot(outlier.shape=8, outlier.color='red') +
                        stat_boxplot(geom='errorbar', width=0.3) +
                        xlab('Quality') +
                        ylab('Density') +
                        # ylim(0.985, 1.005) +
                        ggtitle('Density in White Wine')

bp.density.r.quality <- ggplot(redwine, aes(x=quality, y=density, color=quality)) +
                          geom_boxplot(outlier.shape=8, outlier.color='red') +
                          stat_boxplot(geom='errorbar', width=0.3) +
                          xlab('Quality') +
                          ylab('Density') +
                          # ylim(0.985, 1.005) +
```

```
                        ggtitle('Density in Red Wine')

bp.density <- ggplot(wine.collated.grouped[wine.collated.grouped[, 'quality']=='high',], aes(x=wine.type
              geom_boxplot(outlier.shape=8, outlier.color='red') +
              stat_boxplot(geom='errorbar', width=0.3) +
              xlab('Wine Type') +
              ylab('Density') +
              # ylim(0.985, 1.005) +
              ggtitle('How density is distributed in red and white wine')
# bp.density
grid.arrange(bp.density.w.quality, bp.density.r.quality, bp.density, nrow=1)
```




Overall high quality is associated with less density. Remember (more sugar) lesser the residual sugar, lesser the density. Then lesser the density, higher the alcohol level is.

```
ggplot(wine, aes(x=alcohol, y=density, color=quality)) +
geom_point() +
ylim(0.984, 1.005)
```




```
ggtitle('How density and alcohol are correlated')
```

### Sulphates

```
bp.sul.w <- ggplot(wine, aes(x=quality, y=sulphates, color=quality)) +
                      geom_boxplot(outlier.shape=8, outlier.color='red') +
                      stat_boxplot(geom='errorbar', width=0.3) +
                      xlab('Quality') +
                      ylab('Sulphates') +
                      # ylim(0.985, 1.005) +
                      ggtitle('Sulphates in White Wine')

bp.sul.r <- ggplot(redwine, aes(x=quality, y=sulphates, color=quality)) +
                        geom_boxplot(outlier.shape=8, outlier.color='red') +
                        stat_boxplot(geom='errorbar', width=0.3) +
                        xlab('Quality') +
                        ylab('Sulphates') +
                        # ylim(0.985, 1.005) +
                        ggtitle('Sulphates in Red Wine')

bp.sul <- ggplot(wine.collated.grouped[wine.collated.grouped[, 'quality']=='high',], aes(x=wine.type, y=
              geom_boxplot(outlier.shape=8, outlier.color='red') +
              stat_boxplot(geom='errorbar', width=0.3) +
```

```
              xlab('Wine Type') +
              ylab('Sulphates') +
              # ylim(0.985, 1.005) +
              ggtitle('Sulphates in red and white wine')

grid.arrange(bp.sul.w, bp.sul.r, bp.sul, nrow=1)
```


The sulphate is a preservative and red wine typically possess a lot of aroma and fixed acidity than white wine and in order to preserve the chemical properties, I think it makes sense to see the association between high quality wine and high level of Sulphate being more profound in red wine than in the white wine. We will check if there is statistical difference between two subgroups of sulphate in low and high quality level.

```
ttest.sulphates.good.bad <- t.test(good$sulphates, bad$sulphates, var.equal=F)
ttest.sulphates.good.bad
```

```
##
##  Welch Two Sample t-test
##
## data:  good$sulphates and bad$sulphates
## t = 1.5506, df = 3507.8, p-value = 0.1211
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.001191081  0.010198769
## sample estimates:
## mean of x mean of y
## 0.4819368 0.4774329
```

As the pvalue for the above `t.test` is 0.1210933 that is greater than the significance threshold 0.05 I conclude there is not statistical significance in sulphates level between the two subgroups low and high quality in **White** wines.

## Conclusion

From the analyses conducted above, we can conclude a good quality Vinho Verde White wine should possess the following properties:

1. Residual sugar below 7.0326829.

2. ABV of higher than 9.8%
3. pH level of higher than 3.164
4. Chlorides less than 0.04
5. Volatile acidity less than 0.28
6. Density less than 0.995

I don't think it would be a good idea to list all exhaustive criteria as we will quickly run into something called overfitting. I believe the above stated constraints act as a general criteria to illustrate a good quality Vinho Verde **White** wine.

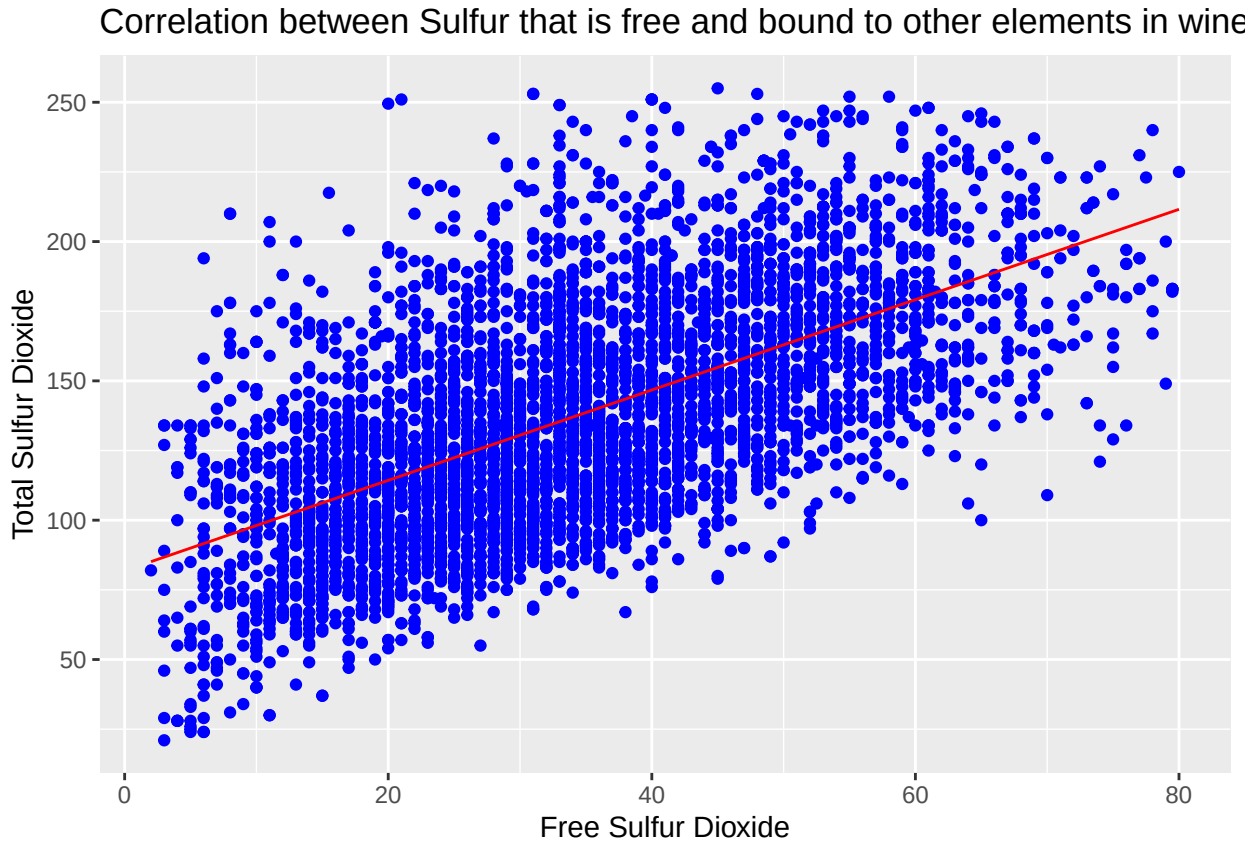
# Modeling

## Correlation test between Total sulfur dioxide and Free sulfur dioxide

```
tmp_model <- lm(total.sulfur.dioxide ~ free.sulfur.dioxide, data=wine)
tmp_x <- seq(min(wine$free.sulfur.dioxide), max(wine$free.sulfur.dioxide), length.out=50)
tmp_y <- tmp_x * tmp_model$coefficients[['free.sulfur.dioxide']] + tmp_model$coefficients[['(Intercept)
tmp_d <- cbind(tmp_x, tmp_y)
colnames(tmp_d) <- c('free.sulfur.dioxide', 'total.sulfur.dioxide')
tmp_d <- data.frame(tmp_d)

cor.test(wine$free.sulfur.dioxide, wine$total.sulfur.dioxide)
```

```
##
##  Pearson's product-moment correlation
##
## data:  wine$free.sulfur.dioxide and wine$total.sulfur.dioxide
## t = 52.693, df = 4896, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5833884 0.6191414
## sample estimates:
##       cor
## 0.6015661
```

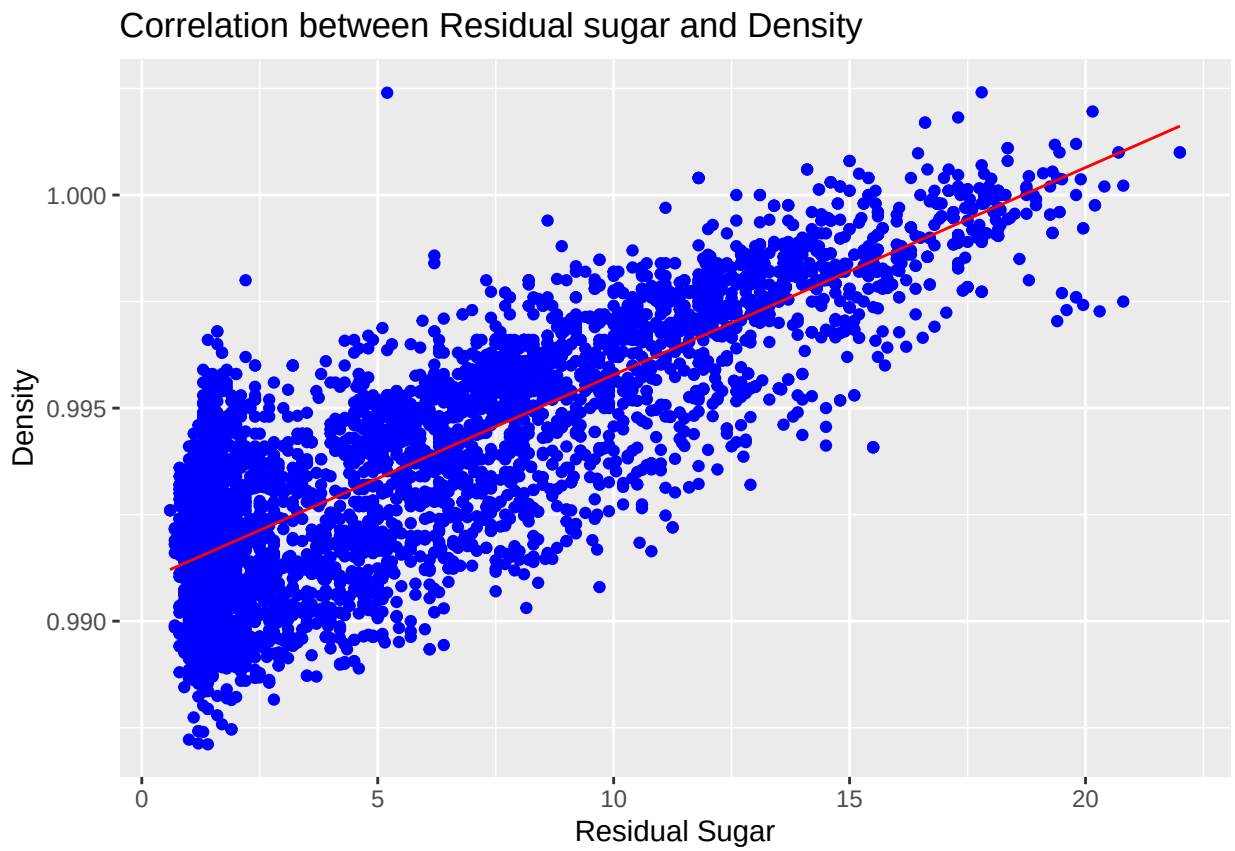
```
ggplot(wine, aes(x=free.sulfur.dioxide, y=total.sulfur.dioxide)) +
geom_point(color='blue') +
geom_line(data=tmp_d, color='red') +
xlab('Free Sulfur Dioxide') +
ylab('Total Sulfur Dioxide') +
ggtitle('Correlation between Sulfur that is free and bound to other elements in wine')
```




As I showed earlier in the correlation plot, cor.test does show that there is strong correlation between Free Sulfur Dioxide and Total Sulfur Dioxide with 95 confidence level.

### Correlation between Residual Sugar and Density

```
tmp_model <- lm(density ~ residual.sugar, data=wine)
tmp_x <- seq(min(wine$residual.sugar), max(wine$residual.sugar), length.out=50)
tmp_y <- tmp_x * tmp_model$coefficients[['residual.sugar']] + tmp_model$coefficients[['(Intercept)']]
tmp_d <- cbind(tmp_x, tmp_y)
colnames(tmp_d) <- c('residual.sugar', 'density')
tmp_d <- data.frame(tmp_d)

cor.test(wine$residual.sugar, wine$density)

ggplot(wine, aes(x=residual.sugar, y=density)) +
geom_point(color='blue') +
geom_line(data=tmp_d, color='red') +
xlab('Residual Sugar') +
ylab('Density') +
ggtitle('Correlation between Residual sugar and Density')
```




### Is alcohol level the same for red and white wine

It is one of our hypothesis that a good quality wine would reflect the true nature of the type of wine. For instance, white wines are generally sparkling, spritzy wines with more fresh notes compared to red wine that is more sweet with high ABV in percentage.

```
aov.alcohol_winetype <- aov(alcohol~wine.type, data=wine.collated)
summary(aov.alcohol_winetype)
```

```
##               Df Sum Sq Mean Sq F value  Pr(>F)
## wine.type      1     18  18.025   12.91 0.00033 ***
## Residuals   6495   9071   1.397
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
TukeyHSD(aov.alcohol_winetype)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = alcohol ~ wine.type, data = wine.collated)
##
## $wine.type
##                diff        lwr       upr     p adj
## white-red 0.1222825 0.05555777 0.1890072 0.0003298
```

From the results above, I can conclude that the alcohol level between the two subgroups red and white wine are statistically significant. In other words they do have differences between them that is due to more than just a chance.

## Volatile Acidity

As we may have seen in the boxplot from the EDA part, the difference in volatile acidity between two subgroups high and low quality wines were not visible from the macroscopic level. Hence I would like to run an ANOVA.

```
aov.volatile.acidity_winetype <- aov(volatile.acidity~wine.type, data=wine.collated.grouped)
summary(aov.volatile.acidity_winetype)
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## wine.type       1  78.36   78.36    7024 <2e-16 ***
## Residuals    6495  72.46    0.01
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Volatile acidity refers to the aromatic flavor. I can see that the two subgroups red and white wines are statistically significant.

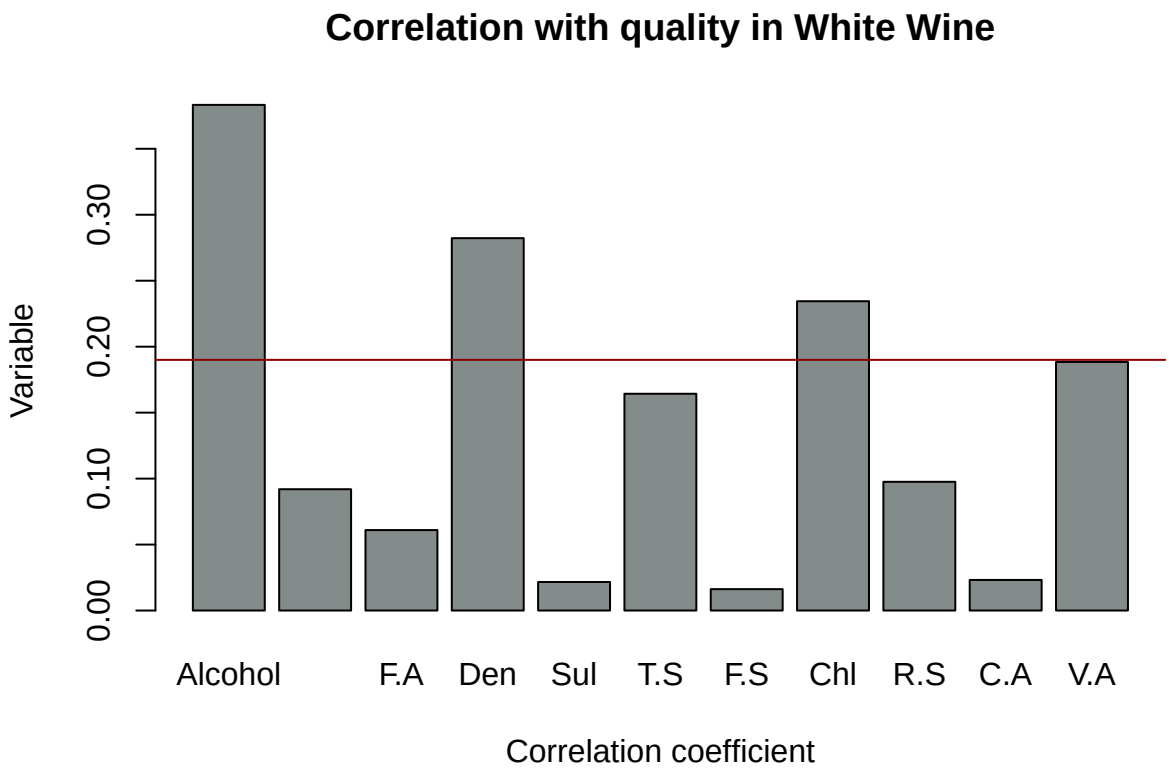
## Point Biserial Correlation

### Correlation between independent variables and dependent variables

As you know, we are dealing with a response variable that is dichotomous and a set of independent variables that are continuous in nature. In order to compute the correlation and to understand the variables that most affect the quality I have computed this.

```
library(ltm)
alcohol_corr <- biserial.cor(wine.grouped$alcohol, wine.grouped$quality, use = c("all.obs"), level=2)
ph_corr <- biserial.cor(wine.grouped$pH, wine.grouped$quality, use=c("all.obs"), level=2)
fixed.acidity_corr <- biserial.cor(wine.grouped$fixed.acidity, wine.grouped$quality, use=c("all.obs"),
density_corr <- biserial.cor(wine.grouped$density, wine.grouped$quality, use=c("all.obs"), level=2)
sulphates_corr <- biserial.cor(wine.grouped$sulphates, wine.grouped$quality, use=c("all.obs"), level=2)
total.sulfur.dioxide_corr <- biserial.cor(wine.grouped$total.sulfur.dioxide, wine.grouped$quality, use=
free.sulfur.dioxide_corr <- biserial.cor(wine.grouped$free.sulfur.dioxide, wine.grouped$quality, use=c(
chlorides_corr <- biserial.cor(wine.grouped$chlorides, wine.grouped$quality, use=c("all.obs"), level=2)
residual.sugar_corr <- biserial.cor(wine.grouped$residual.sugar, wine.grouped$quality, use=c("all.obs")
citric.acid_corr <- biserial.cor(wine.grouped$citric.acid, wine.grouped$quality, use=c("all.obs"), level
volatile.acidity_corr <- biserial.cor(wine.grouped$volatile.acidity, wine.grouped$quality, use=c("all.ob


vars.affect.quality <- c(abs(alcohol_corr), abs(ph_corr), abs(fixed.acidity_corr), abs(density_corr), ab
barplot(vars.affect.quality,
main = "Correlation with quality in White Wine",
xlab = "Correlation coefficient",
ylab = "Variable",
names.arg = c("Alcohol", "pH", "F.A", "Den", "Sul", "T.S", "F.S", "Chl", "R.S", "C.A", "V.A"),
col = "azure4",
```

```
horiz = F)
abline(h=0.19, col=c('darkred'))
```

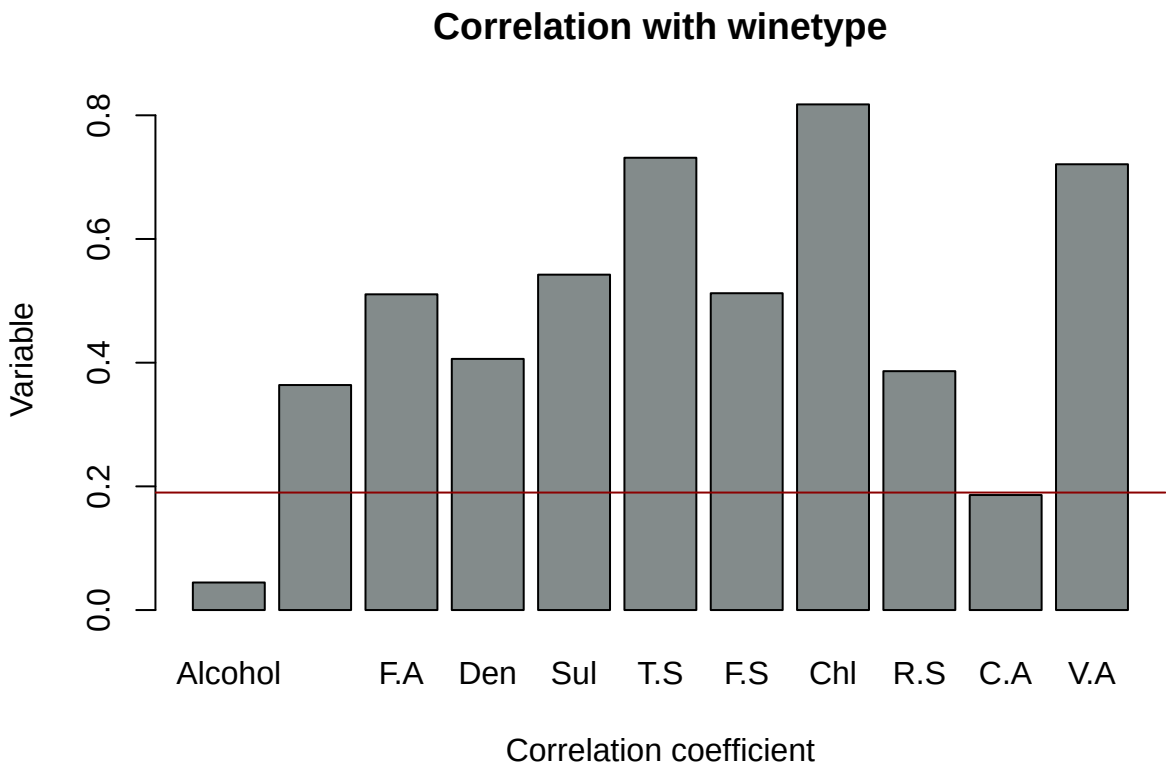



**Correlation between wine.type and other variables excluding quality**

I thought it would also be interesting to show what variables play a critical role in deciding the type of wine.

```
alcohol_corr <- biserial.cor(wine.collated.grouped$alcohol, wine.collated.grouped$wine.type, use = c("al
ph_corr <- biserial.cor(wine.collated.grouped$pH, wine.collated.grouped$wine.type, use=c("all.obs"), lev
fixed.acidity_corr <- biserial.cor(wine.collated.grouped$fixed.acidity, wine.collated.grouped$wine.type
density_corr <- biserial.cor(wine.collated.grouped$density, wine.collated.grouped$wine.type, use=c("all
sulphates_corr <- biserial.cor(wine.collated.grouped$sulphates, wine.collated.grouped$wine.type, use=c(
total.sulfur.dioxide_corr <- biserial.cor(wine.collated.grouped$total.sulfur.dioxide, wine.collated.grou
free.sulfur.dioxide_corr <- biserial.cor(wine.collated.grouped$free.sulfur.dioxide, wine.collated.groupe
chlorides_corr <- biserial.cor(wine.collated.grouped$chlorides, wine.collated.grouped$wine.type, use=c(
residual.sugar_corr <- biserial.cor(wine.collated.grouped$residual.sugar, wine.collated.grouped$wine.typ
citric.acid_corr <- biserial.cor(wine.collated.grouped$citric.acid, wine.collated.grouped$wine.type, use
volatile.acidity_corr <- biserial.cor(wine.collated.grouped$volatile.acidity, wine.collated.grouped$wine

vars.affect.winetype <- c(abs(alcohol_corr), abs(ph_corr), abs(fixed.acidity_corr), abs(density_corr), a
barplot(vars.affect.winetype,
main = "Correlation with winetype",
xlab = "Correlation coefficient",
ylab = "Variable",
```

```
names.arg = c("Alcohol", "pH", "F.A", "Den", "Sul", "T.S", "F.S", "Chl", "R.S", "C.A", "V.A"),
col = "azure4",
horiz = F)
abline(h=0.19, col=c('darkred'))
```




I believe the abline may not be entirely correct here, but I will learn in the coming classes on how to use it more appropriately.

### Is density statistically significant between two wine types

```
ttest.density.red.white <- t.test(good$density, mu=mean(good.redwine$density), alternative='greater')
ttest.density.red.white
```

```
##
##  One Sample t-test
##
## data:  good$density
## t = -53.441, df = 3257, p-value = 1
## alternative hypothesis: true mean is greater than 0.9961289
## 95 percent confidence interval:
##  0.9933444       Inf
## sample estimates:
## mean of x
## 0.9934275
```

It would be my hypothesis based on the research that White wines must be higher in density compared to Red wines as Red wines are the ones with more ABV and both `t.test` and boxplot defied this. I believe this is due to the level of other chemicals present in the composition.

```
test.model <- glm(quality ~ ., data=wine.grouped[-length(colnames(wine.grouped))], family='binomial') #
summary(test.model)
```

```
##
## Call:
## glm(formula = quality ~ ., family = "binomial", data = wine.grouped[-length(colnames(wine.grouped))]
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9477  -0.9025   0.4360   0.8014   2.3746
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)           4.191e+02  5.813e+01   7.210 5.60e-13 ***
## fixed.acidity         1.938e-01  6.488e-02   2.987  0.00282 **
## volatile.acidity     -7.338e+00  4.979e-01 -14.738  < 2e-16 ***
## citric.acid           4.929e-01  4.038e-01   1.221  0.22224
## residual.sugar        2.145e-01  2.185e-02   9.818  < 2e-16 ***
## chlorides            -5.065e+00  4.334e+00  -1.169  0.24259
## free.sulfur.dioxide   1.685e-02  2.991e-03   5.635 1.75e-08 ***
## total.sulfur.dioxide -9.331e-04  1.220e-03  -0.765  0.44447
## density              -4.338e+02  5.879e+01  -7.379 1.59e-13 ***
## pH                    1.696e+00  3.171e-01   5.349 8.86e-08 ***
## sulphates             2.023e+00  3.977e-01   5.088 3.62e-07 ***
## alcohol               5.239e-01  7.924e-02   6.612 3.79e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6245.4  on 4897  degrees of freedom
## Residual deviance: 4953.0  on 4886  degrees of freedom
## AIC: 4977
##
## Number of Fisher Scoring iterations: 5
```

We can clearly the highly correlated variables shows no statistical significance.

Starting off with highly correlated variable

```
model1 <- glm(quality ~ alcohol, data=wine.grouped, family='binomial')
summary(model1)
```

```
##
## Call:
## glm(formula = quality ~ alcohol, family = "binomial", data = wine.grouped)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -2.6245  -1.0843   0.5021   0.8730   1.5310
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.82259    0.33627  -23.26   <2e-16 ***
## alcohol      0.82603    0.03312   24.94   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6245.4  on 4897  degrees of freedom
## Residual deviance: 5434.2  on 4896  degrees of freedom
## AIC: 5438.2
##
## Number of Fisher Scoring iterations: 4
```

I do not want to use density as it is highly correlated with alcohol. The next highly correlated predictor variable with response variable is chlorides.

```
model2 <- glm(quality ~ alcohol+chlorides, data=wine.grouped, family='binomial')
summary(model2)
```

```
##
## Call:
## glm(formula = quality ~ alcohol + chlorides, family = "binomial",
##     data = wine.grouped)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6580  -1.0964   0.5052   0.8878   1.5442
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.97575    0.48135  -14.49   <2e-16 ***
## alcohol      0.78480    0.03702   21.20   <2e-16 ***
## chlorides   -9.73366    4.00614   -2.43   0.0151 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6245.4  on 4897  degrees of freedom
## Residual deviance: 5428.3  on 4895  degrees of freedom
## AIC: 5434.3
##
## Number of Fisher Scoring iterations: 4
```

```
xkablevif(model2)
```

Table 3: VIFs of Model: quality  alcohol + chlorides

| alcohol | chlorides |
|---|---|
| 10.161 | 7.4 |

Table 4: VIFs of Model: quality  alcohol + chlorides + free.sulfur.dioxide

| alcohol | chlorides | free.sulfur.dioxide |
|---|---|---|
| 11.448 | 7.478 | 5.797 |

```
model3 <- glm(quality ~ alcohol+chlorides+free.sulfur.dioxide, data=wine.grouped, family='binomial')
summary(model3)
```

```
##
## Call:
## glm(formula = quality ~ alcohol + chlorides + free.sulfur.dioxide,
##     family = "binomial", data = wine.grouped)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8926  -1.0559   0.4818   0.8850   1.7095
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -8.661442   0.521974 -16.594  < 2e-16 ***
## alcohol               0.879609   0.039289  22.388  < 2e-16 ***
## chlorides           -10.416564   4.027021  -2.587  0.00969 **
## free.sulfur.dioxide   0.021421   0.002244   9.544  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6245.4  on 4897  degrees of freedom
## Residual deviance: 5334.0  on 4894  degrees of freedom
## AIC: 5342
##
## Number of Fisher Scoring iterations: 4
```

I did not include Total Sulfur Dioxide as it did not have statistical significance.

```
xkablevif(model3)
```

Not going to account for free.sulphur.dioxide as it has higher VIF score.

```
model4 <- glm(quality ~ alcohol+chlorides+residual.sugar, data=wine.grouped, family='binomial')
summary(model4)
```

```
##
## Call:
```

Table 5: VIFs of Model: quality  alcohol + chlorides + residual.sugar

| alcohol | chlorides | residual.sugar |
|---|---|---|
| 13.458 | 7.506 | 6.64 |

```
## glm(formula = quality ~ alcohol + chlorides + residual.sugar,
##     family = "binomial", data = wine.grouped)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7245  -1.0759   0.4873   0.9134   1.7529
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -8.860279   0.545459 -16.244  < 2e-16 ***
## alcohol          0.933700   0.042600  21.918  < 2e-16 ***
## chlorides      -10.593138   4.034568  -2.626  0.00865 **
## residual.sugar   0.060559   0.007443   8.136 4.07e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6245.4  on 4897  degrees of freedom
## Residual deviance: 5360.3  on 4894  degrees of freedom
## AIC: 5368.3
##
## Number of Fisher Scoring iterations: 4
```

```
xkablevif(model4)
```

You can clearly see residual sugar is over the limit. This is because, residual sugar is correlated with alcohol. Remember the results I conveyed in the EDA part ?

```
model5 <- glm(quality ~ alcohol+chlorides+sulphates, data=wine.grouped, family='binomial')
summary(model5)
```

```
##
## Call:
## glm(formula = quality ~ alcohol + chlorides + sulphates, family = "binomial",
##     data = wine.grouped)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6071  -1.0849   0.5148   0.8779   1.5168
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -7.47644    0.50801 -14.717  < 2e-16 ***
## alcohol       0.78487    0.03697  21.228  < 2e-16 ***
## chlorides   -10.54500    4.01576  -2.626  0.00864 **
## sulphates     1.11084    0.34938   3.179  0.00148 **
```

Table 6: VIFs of Model: quality  alcohol + chlorides + sulphates

| alcohol | chlorides | sulphates |
|---|---|---|
| 10.138 | 7.436 | 5.777 |

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6245.4  on 4897  degrees of freedom
## Residual deviance: 5418.1  on 4894  degrees of freedom
## AIC: 5426.1
##
## Number of Fisher Scoring iterations: 4
```

```
xkablevif(model5)
```

I will ignore the slight increase in VIF score.

```
model6 <- glm(quality ~ alcohol+chlorides+sulphates+pH, data=wine.grouped, family='binomial')
summary(model6)
```

```
##
## Call:
## glm(formula = quality ~ alcohol + chlorides + sulphates + pH,
##     family = "binomial", data = wine.grouped)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6260  -1.0836   0.5128   0.8723   1.5731
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -9.10490    0.85749 -10.618  < 2e-16 ***
## alcohol       0.77185    0.03726  20.718  < 2e-16 ***
## chlorides   -10.99154    4.02552  -2.730  0.00632 **
## sulphates     1.00447    0.35165   2.856  0.00428 **
## pH            0.57627    0.24330   2.369  0.01786 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6245.4  on 4897  degrees of freedom
## Residual deviance: 5412.5  on 4893  degrees of freedom
## AIC: 5422.5
##
## Number of Fisher Scoring iterations: 4
```

Table 7: VIFs of Model: quality  alcohol + chlorides + sulphates + pH

| alcohol | chlorides | pH | sulphates |
|---|---|---|---|
| 10.294 | 7.472 | 5.64 | 5.852 |

```
xkablevif(model6)
```

This appears like a good model eventually.

```
an.md <- anova(model1, model2, model5, model6)
an.md
```

```
## Analysis of Deviance Table
##
## Model 1: quality ~ alcohol
## Model 2: quality ~ alcohol + chlorides
## Model 3: quality ~ alcohol + chlorides + sulphates
## Model 4: quality ~ alcohol + chlorides + sulphates + pH
##   Resid. Df Resid. Dev Df Deviance
## 1      4896     5434.2
## 2      4895     5428.3  1   5.9073
## 3      4894     5418.1  1  10.1932
## 4      4893     5412.5  1   5.6301
```

```
summary(an.md)
```

```
##    Resid. Df      Resid. Dev         Df       Deviance
##  Min.   :4893   Min.   :5412   Min.   :1   Min.   : 5.630
##  1st Qu.:4894   1st Qu.:5417   1st Qu.:1   1st Qu.: 5.769
##  Median :4894   Median :5423   Median :1   Median : 5.907
##  Mean   :4894   Mean   :5423   Mean   :1   Mean   : 7.244
##  3rd Qu.:4895   3rd Qu.:5430   3rd Qu.:1   3rd Qu.: 8.050
##  Max.   :4896   Max.   :5434   Max.   :1   Max.   :10.193
##                                NA's   :1   NA's   :1
```

Now I am somewhat surprised that ANOVA summary for the model does not show statistical significance level or p adjusted values. Perhaps this is due to the use of GLM instead of LM. Due to the limited scope of this project, we will end the analysis at this point and will look forward learning more on interpreting ANOVA results with GLM models.

# Future Work

## Connection between residual sugar, alcohol and quality

```
library(pracma)
library(plotly)
```

```r
get_prob <- function(logit){
  odds <- exp(logit)
  return(odds/(1+odds))
}

tmp.model2 <- glm(quality~residual.sugar+alcohol, data=wine.grouped, family='binomial')
al_range <- seq(min(good$alcohol), max(good$alcohol), length.out=50)
rs_range <- seq(min(good$residual.sugar), max(good$residual.sugar), length.out=50)
M <- expand.grid(al_range, rs_range)
al <- M$Var1
rs <- M$Var2
c <- tmp.model2$coefficients[['(Intercept)']]
b1 <- tmp.model2$coefficients[['residual.sugar']]
b2 <- tmp.model2$coefficients[['alcohol']]
logits <- c + b1*rs + b2*al
q.probs <- get_prob(logits)
q.probs <- as.matrix(q.probs) %>% Reshape(50, 50)

# tmp.z <- c + b1*wine.grouped$residual.sugar + b2*wine.grouped$alcohol
# tmp.z.probs <- get_prob(tmp.z)

fig <- plot_ly(x=rs_range, y=al_range, z=q.probs) %>% layout(scene=list()) %>%
       # add_trace(data=wine.grouped, x=~residual.sugar, y=~alcohol, z=tmp.z.probs, mode = "markers", t
       #      marker = list(size = 5, colors=c("blue", "red"), symbol = 104)) %>%
       add_surface()
fig
```


The curve represents something called `sigmoid` a so called an activation function that maps logits that are unbounded $[-\infty, +\infty]$ to $[0, 1]$ that is often interpreted as probability. Labels on axis did not work for some reason so **x->residual sugar, y->alcohol and z->prob of being high quality**

### Clustering

In the due course, I wish to further understand how clustering such as Gaussian Mixture Model can be used to categorize different quality levels. Perhaps it would help have an even better understanding.

### Multicollinearity

We have seen in the Modeling section that there were several variables that were highly correlated with each other and did not allow us to include them in the model. In the further course, I wish to use Principal Component Analysis with sphering to analyze and observe how multicollinearity can be avoided and at the same improve model performance with less dimensions included.

## References

1. Wine - worldwide: Statista market forecast. Statista. (n.d.). Retrieved April 3, 2022, from https://www.statista.com/outlook/cmo/alcoholic-drinks/wine/worldwide
2. Hagan , M. K. (2021, April 20). The powers of wine fermentation: What it does besides create alcohol. A large glass of real wine, by the bottle. Retrieved April 3, 2022, from https://usualwines.com/blogs/knowledge-base/wine-fermentation

3. Goldsmith, R. (2014, April 23). Vinho Verde – a splash of summer vinous joy! Vinho Verde – a Splash of Summer Vinous Joy! Retrieved April 3, 2022, from https://www.alcoholprofessor.com/blog-posts/blog/2014/04/23/vinho-verde-a-splash-of-summer-vinous-joy

4. Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. Decision Support Systems, 47(4), 547–553.

5. Exports of portuguese 'Vinho verde' surpass 50% of total sales. The Portugal News. (2019, March 14). Retrieved April 3, 2022, from https://www.theportugalnews.com/news/exports-of-portuguese-vinho-verde-surpass-50-of-total-sales/48712?fbclid=IwAR2Q1Uk_EzV0HKMv8RMscMEPC6m20PosvzppC7qIfv8gzBeVDlE

6. Puckette, M. (2021, January 25). Reality of wine prices (what you get for what you spend). Wine Folly. Retrieved April 3, 2022, from https://winefolly.com/lifestyle/reality-of-wine-prices-what-you-get-for-what-you-spend/

7. Suckling, J. (2021, July 29). Learn about alcohol content in wine: Highest to lowest ABV wines - 2022. MasterClass. Retrieved April 3, 2022, from https://www.masterclass.com/articles/learn-about-alcohol-content-in-wine-highest-to-lowest-abv-wines#what-does-alcohol-by-volume-abv-mean

8. Saladino, E., & Grinberg, D. (2019, April 22). Let's all say it just once: There's nothing wrong with liking sweet wine. VinePair. Retrieved April 3, 2022, from https://vinepair.com/articles/sweet-wine-dry-culture/

9. Deviations, O. T. S. (2018, September 14). An overview of correlation measures between categorical and continuous variables. Medium. Retrieved April 3, 2022, from https://medium.com/@outside2SDs/an-overview-of-correlation-measures-between-categorical-and-continuous-variables-4c7f85610365

10. Kumar, A. (2022, January 23). Python - replace missing values with mean, median & mode. Data Analytics. Retrieved April 3, 2022, from https://vitalflux.com/pandas-impute-missing-values-mean-median-mode/