White Wine Quality EDA by Adrián Vera Ros

Introduction

In this EDA we aim to analyze the white variants of the Portuguese “Vinho Verde” wine with data from the book “Modeling wine preferences by data mining from physicochemical properties.” by P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis.

Doing a preliminary research about the “Vinho Verde” denomination, we discover that it’s an Origin Denomination (OD) and not a grape variety. “Vinho verde” translates as “Young wine”, or wines that are relased 3-6 months after harvest.

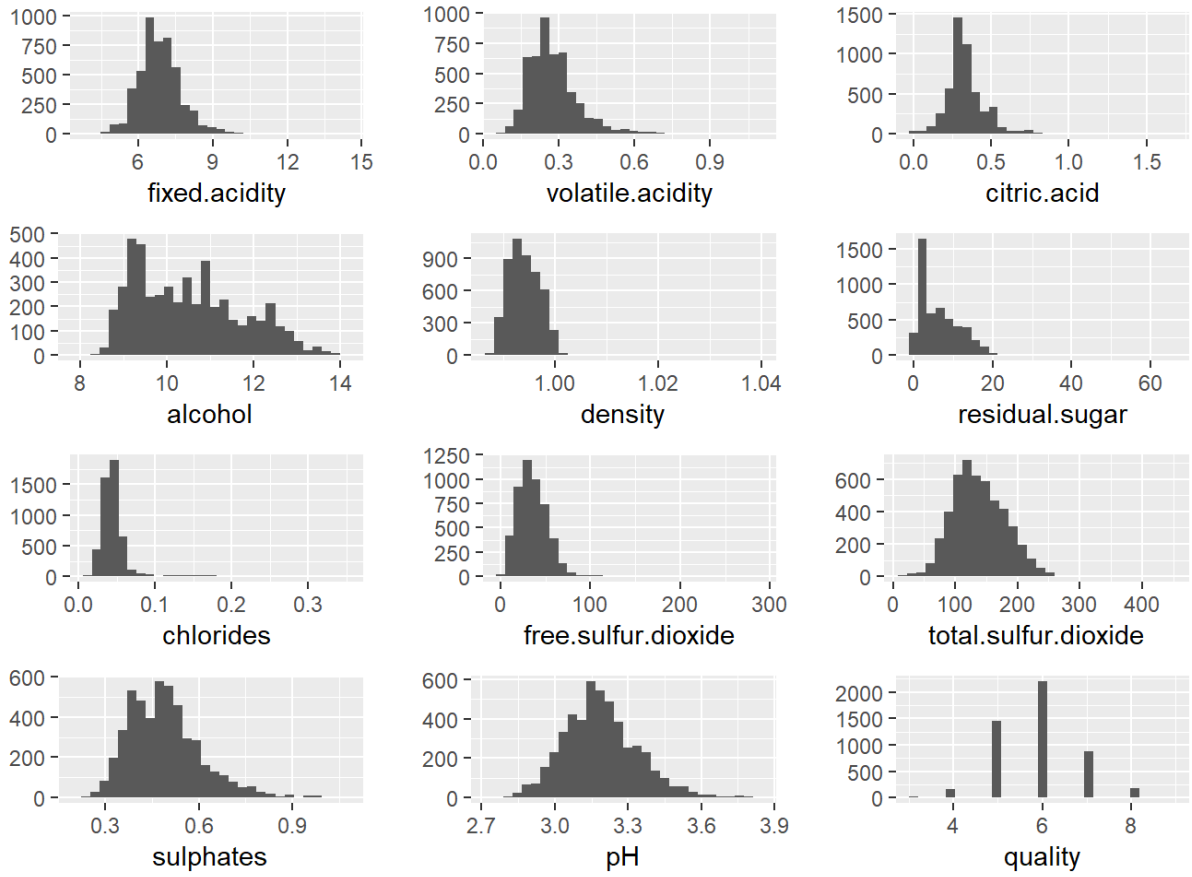
Univariate Plots Section

The dataset is comprised by 4898 observations from 13 different variables, where we have the wines ordered by a numeric key, different objective physicochemical characteristics of each wine and the median of at least 3 subjective sensory evaluations made by wine experts.

More information about the nature of the variables can be found here: https://s3.amazonaws.com/udacity-hosted-downloads/ud651/wineQualityInfo.txt (https://s3.amazonaws.com/udacity-hosted-downloads/ud651/wineQualityInfo.txt)

```
## 'data.frame':    4898 obs. of  13 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ 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 0.04
9 0.044 ...
##  $ free.sulfur.dioxide : num  45 14 30 47 47 30 30 45 14 28 ...
##  $ total.sulfur.dioxide: num  170 132 97 186 186 97 136 170 132 129 ...
##  $ density             : num  1.001 0.994 0.995 0.996 0.996 ...
##  $ pH                  : num  3 3.3 3.26 3.19 3.19 3.26 3.18 3 3.3 3.22 ...
##  $ sulphates           : num  0.45 0.49 0.44 0.4 0.4 0.44 0.47 0.45 0.49 0.45 ...
##  $ alcohol             : num  8.8 9.5 10.1 9.9 9.9 10.1 9.6 8.8 9.5 11 ...
##  $ quality             : int  6 6 6 6 6 6 6 6 6 6 ...
```

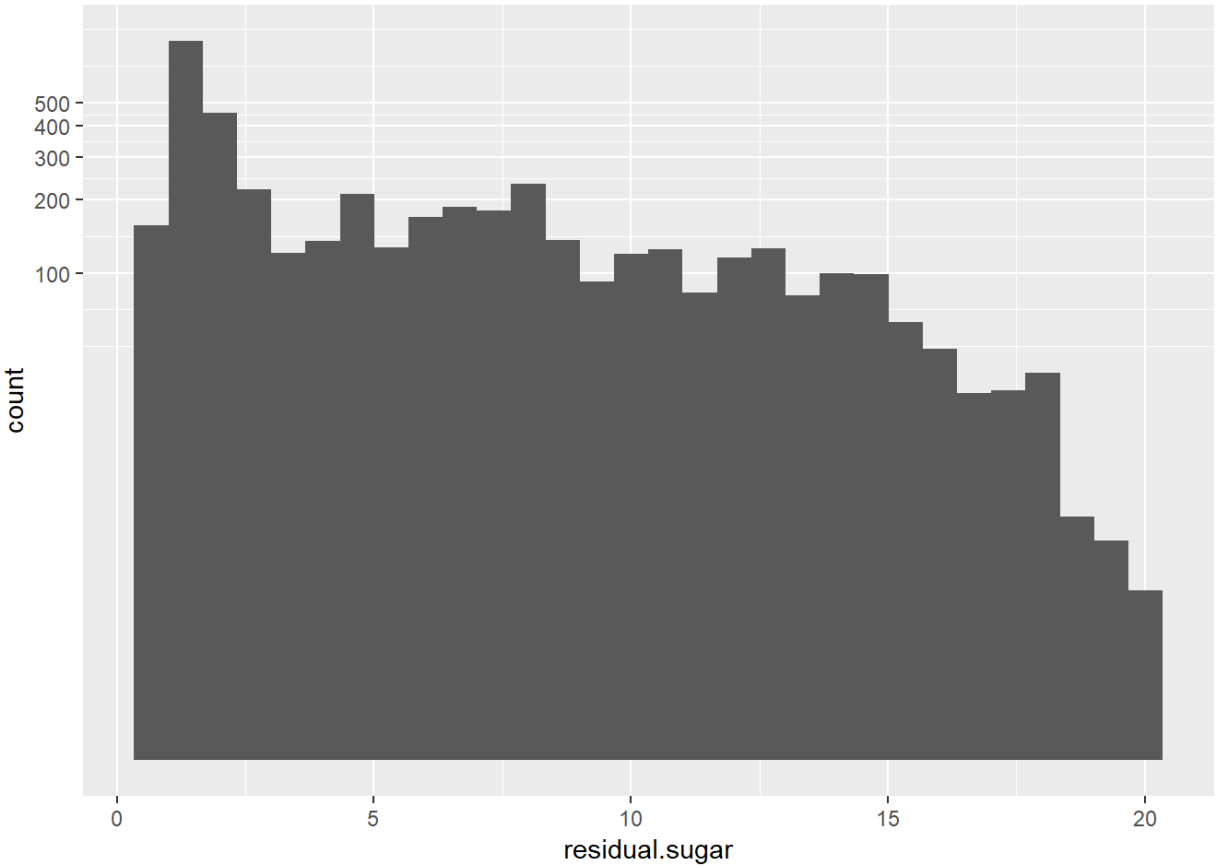
```
##        X        fixed.acidity    volatile.acidity  citric.acid
##  Min.   :   1   Min.   : 3.800   Min.   :0.0800   Min.   :0.0000
##  1st Qu.:1225   1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700
##  Median :2450   Median : 6.800   Median :0.2600   Median :0.3200
##  Mean   :2450   Mean   : 6.855   Mean   :0.2782   Mean   :0.3342
##  3rd Qu.:3674   3rd Qu.: 7.300   3rd Qu.:0.3200   3rd Qu.:0.3900
##  Max.   :4898   Max.   :14.200   Max.   :1.1000   Max.   :1.6600
##  residual.sugar     chlorides       free.sulfur.dioxide
##  Min.   : 0.600   Min.   :0.00900   Min.   :  2.00
##  1st Qu.: 1.700   1st Qu.:0.03600   1st Qu.: 23.00
##  Median : 5.200   Median :0.04300   Median : 34.00
##  Mean   : 6.391   Mean   :0.04577   Mean   : 35.31
##  3rd Qu.: 9.900   3rd Qu.:0.05000   3rd Qu.: 46.00
##  Max.   :65.800   Max.   :0.34600   Max.   :289.00
##  total.sulfur.dioxide    density             pH          sulphates
##  Min.   :  9.0        Min.   :0.9871   Min.   :2.720   Min.   :0.2200
##  1st Qu.:108.0        1st Qu.:0.9917   1st Qu.:3.090   1st Qu.:0.4100
##  Median :134.0        Median :0.9937   Median :3.180   Median :0.4700
##  Mean   :138.4        Mean   :0.9940   Mean   :3.188   Mean   :0.4898
##  3rd Qu.:167.0        3rd Qu.:0.9961   3rd Qu.:3.280   3rd Qu.:0.5500
##  Max.   :440.0        Max.   :1.0390   Max.   :3.820   Max.   :1.0800
##     alcohol         quality
##  Min.   : 8.00   Min.   :3.000
##  1st Qu.: 9.50   1st Qu.:5.000
##  Median :10.40   Median :6.000
##  Mean   :10.51   Mean   :5.878
##  3rd Qu.:11.40   3rd Qu.:6.000
##  Max.   :14.20   Max.   :9.000
```

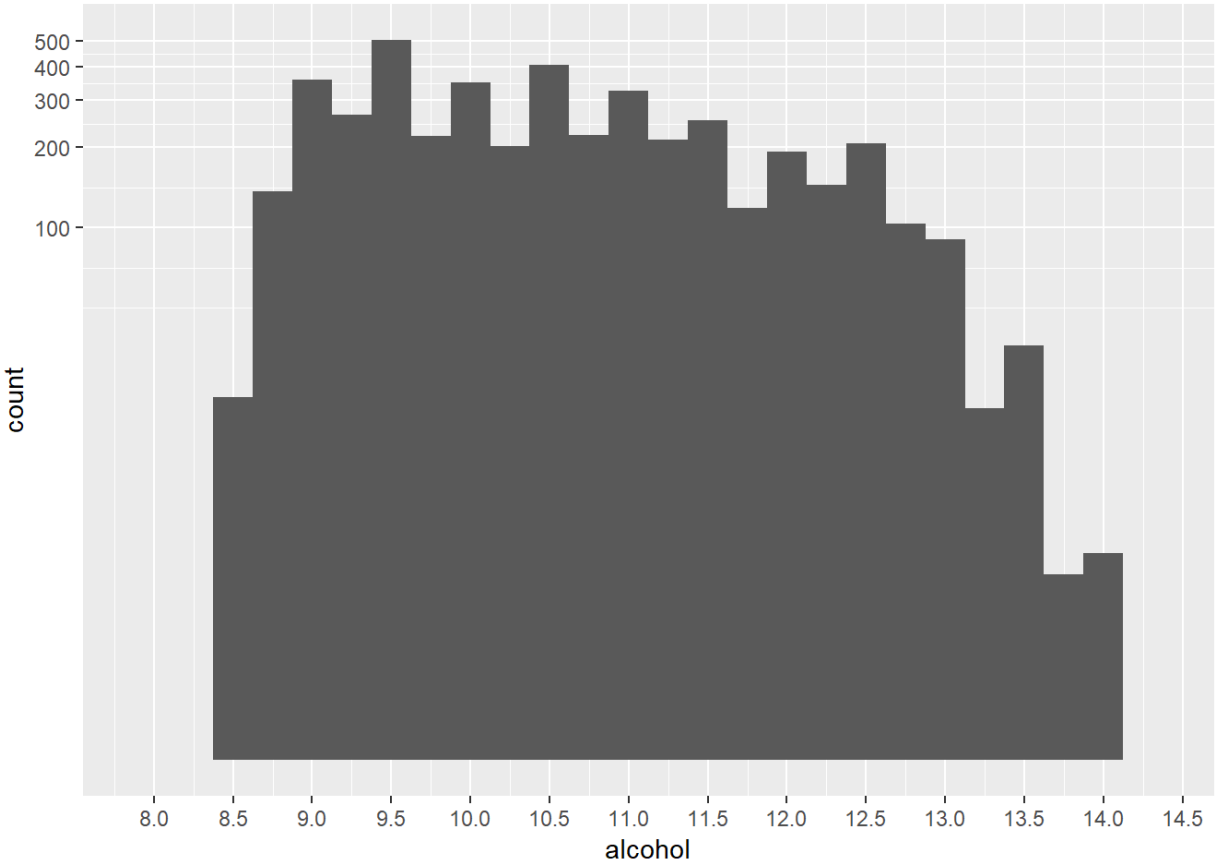


At first glance we can detect outliers in our dataset, both in the summary (with the max value far the average and the 3rd quartile) and in the graphs.

Attending to the distributions, most variables have a normal distribution not centered in the graph due to these outliers. However, in the case of residual sugar and alcohol, this situation doesn't happen and we find a left skewed distribution.

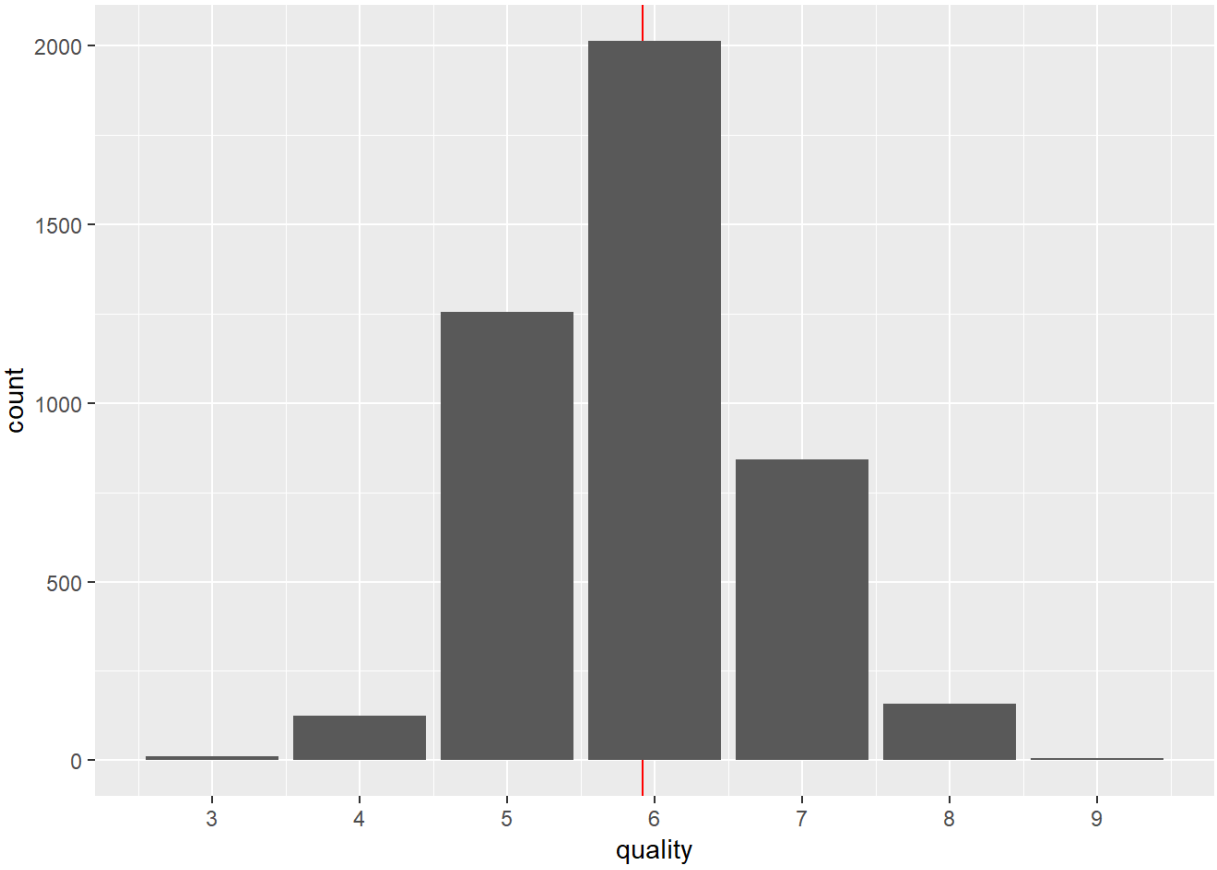
Outliers elimination and plot transformation

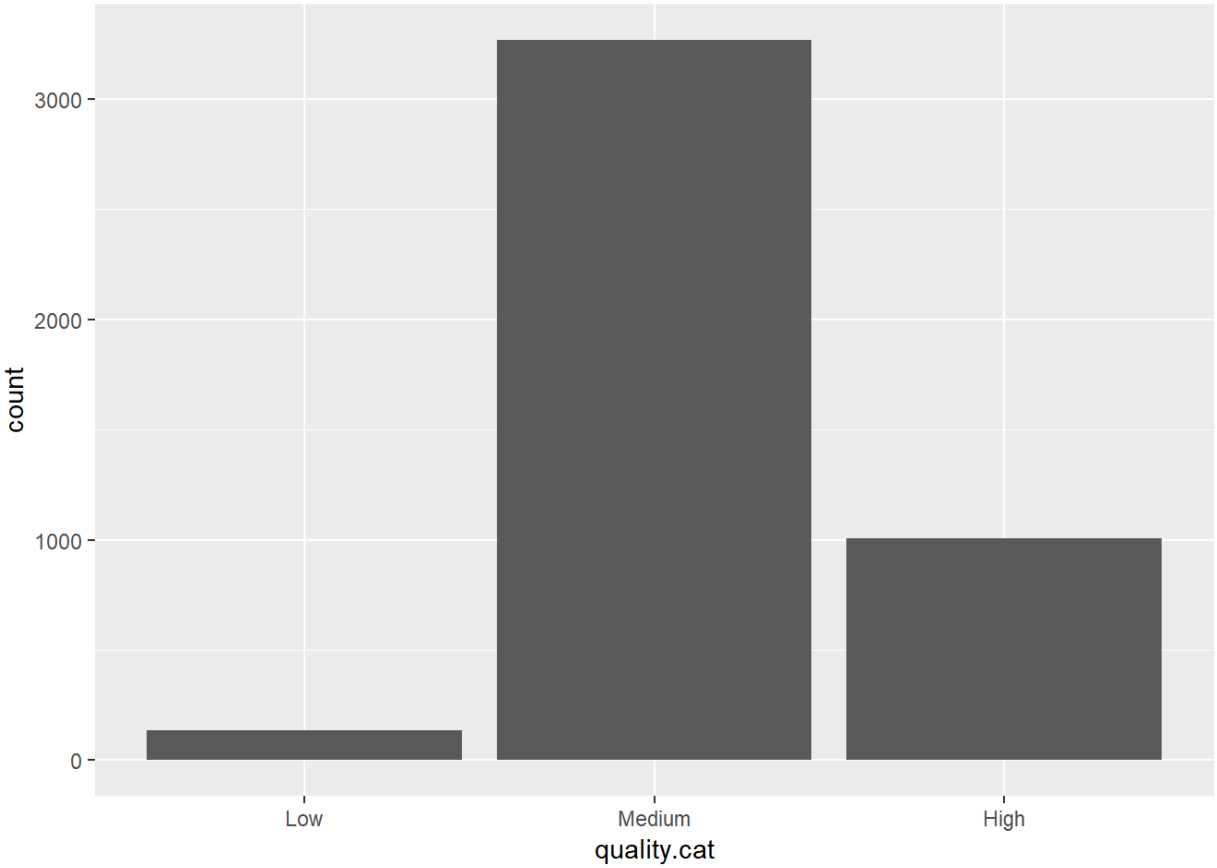




After transforming the variables we notice how sugar still mantains an skewed distribution and how alcohol has spikes in the distribution, probably due to some figures having been rounded.

Quality


2000
1500
1000
500
0
count
3
4
5
6
7
8
9
quality


3000
2000
1000
0
count
Low
Medium
High
quality.cat

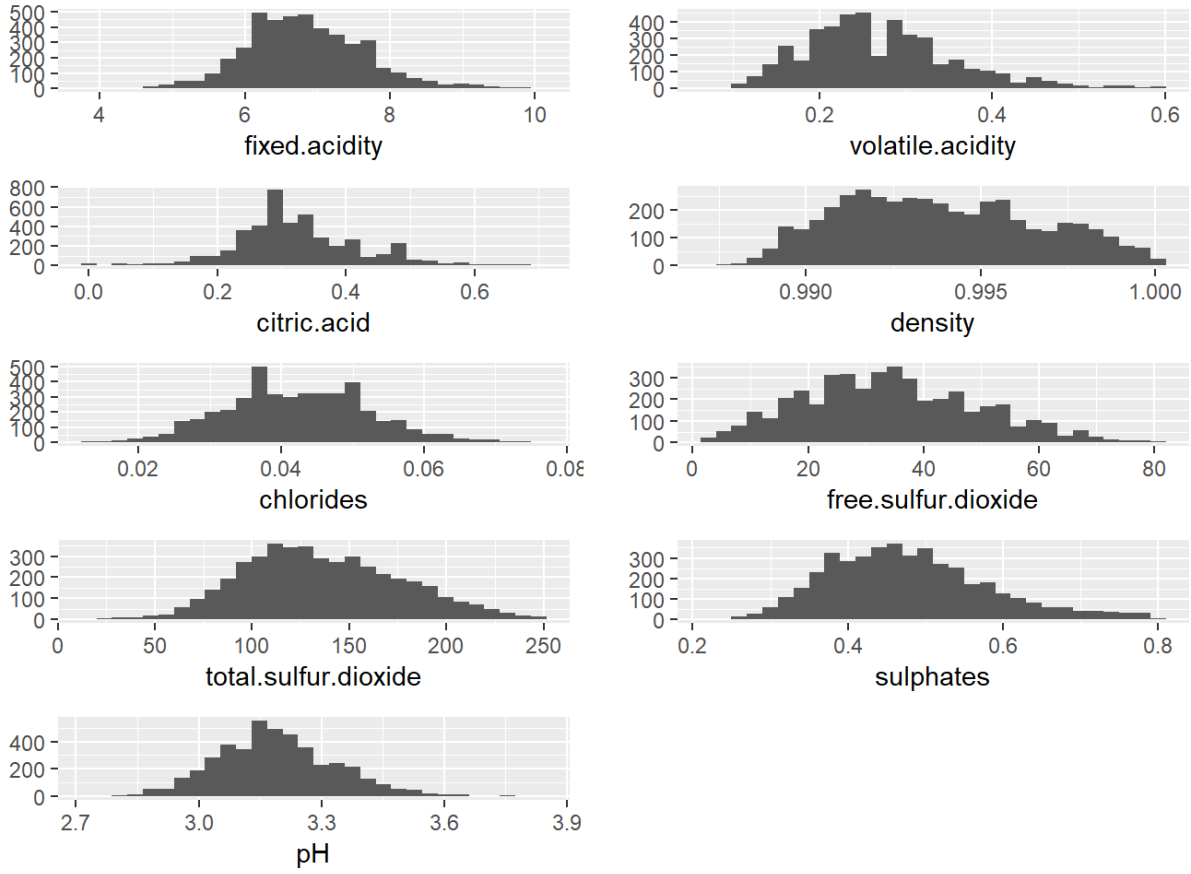
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   5.000   6.000   5.917   6.000   9.000
```

```
##
##    3    4    5    6    7    8    9
##   12  124 1254 2012  841  159    5
```

The wine score are integers, so there are no scores like 6,5. It ranges from 3 to 9, with 6 being the median with 2012 observations and a average score of 5,9. Of the 4407 observations we have after eliminating the outliers, 3.09% have a score between 3 and 4 (Low), 74.11% have a score between 5 and 6 (Medium) and 22,8% have a score of 7 or more (High). However, only 3,77% have a score of 8 or more.

```
## 'data.frame':    4407 obs. of  14 variables:
##  $ X                   : int  2 3 4 5 6 7 9 10 11 12 ...
##  $ fixed.acidity       : num  6.3 8.1 7.2 7.2 8.1 6.2 6.3 8.1 8.1 8.6 ...
##  $ volatile.acidity    : num  0.3 0.28 0.23 0.23 0.28 0.32 0.3 0.22 0.27 0.23 ...
##  $ citric.acid         : num  0.34 0.4 0.32 0.32 0.4 0.16 0.34 0.43 0.41 0.4 ...
##  $ residual.sugar      : num  1.6 6.9 8.5 8.5 6.9 7 1.6 1.5 1.45 4.2 ...
##  $ chlorides           : num  0.049 0.05 0.058 0.058 0.05 0.045 0.049 0.044 0.03
3 0.035 ...
##  $ free.sulfur.dioxide : num  14 30 47 47 30 30 14 28 11 17 ...
##  $ total.sulfur.dioxide: num  132 97 186 186 97 136 132 129 63 109 ...
##  $ density             : num  0.994 0.995 0.996 0.996 0.995 ...
##  $ pH                  : num  3.3 3.26 3.19 3.19 3.26 3.18 3.3 3.22 2.99 3.14 ...
##  $ sulphates           : num  0.49 0.44 0.4 0.4 0.44 0.47 0.49 0.45 0.56 0.53 ...
##  $ alcohol             : num  9.5 10.1 9.9 9.9 10.1 9.6 9.5 11 12 9.7 ...
##  $ quality             : int  6 6 6 6 6 6 6 6 5 5 ...
##  $ quality.cat         : Factor w/ 3 levels "Low","Medium",..: 2 2 2 2 2 2 2 2 2
2 ...
```

```
##        X        fixed.acidity    volatile.acidity  citric.acid
##  Min.   :   2   Min.   : 3.800   Min.   :0.0800   Min.   :0.0000
##  1st Qu.:1212   1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700
##  Median :2496   Median : 6.800   Median :0.2600   Median :0.3100
##  Mean   :2465   Mean   : 6.839   Mean   :0.2719   Mean   :0.3256
##  3rd Qu.:3694   3rd Qu.: 7.300   3rd Qu.:0.3200   3rd Qu.:0.3800
##  Max.   :4898   Max.   :10.000   Max.   :0.6000   Max.   :0.7000
##  residual.sugar     chlorides       free.sulfur.dioxide
##  Min.   : 0.600   Min.   :0.01200   Min.   : 2.00
##  1st Qu.: 1.700   1st Qu.:0.03500   1st Qu.:23.00
##  Median : 5.100   Median :0.04200   Median :33.00
##  Mean   : 6.207   Mean   :0.04229   Mean   :34.44
##  3rd Qu.: 9.600   3rd Qu.:0.04900   3rd Qu.:45.00
##  Max.   :19.950   Max.   :0.07500   Max.   :80.00
##  total.sulfur.dioxide    density             pH          sulphates
##  Min.   : 18.0        Min.   :0.9871   Min.   :2.720   Min.   :0.2200
##  1st Qu.:107.0        1st Qu.:0.9916   1st Qu.:3.090   1st Qu.:0.4100
##  Median :132.0        Median :0.9936   Median :3.180   Median :0.4700
##  Mean   :136.2        Mean   :0.9939   Mean   :3.192   Mean   :0.4837
##  3rd Qu.:165.0        3rd Qu.:0.9959   3rd Qu.:3.280   3rd Qu.:0.5400
##  Max.   :249.5        Max.   :1.0000   Max.   :3.820   Max.   :0.8000
##     alcohol         quality      quality.cat
##  Min.   : 8.00   Min.   :3.000   Low   : 136
##  1st Qu.: 9.50   1st Qu.:5.000   Medium:3266
##  Median :10.40   Median :6.000   High  :1005
##  Mean   :10.57   Mean   :5.917
##  3rd Qu.:11.40   3rd Qu.:6.000
##  Max.   :14.20   Max.   :9.000
```



Univariate Analysis

Most of the wines are between 9 and 11% alcohol and most of our dataset is composed by average wines while only a few score of 8 and more.

According to information coming with the dataset, most wines have a pH level between 3 and 4, which concurs with our wine's pH normal distribution.

According to EU regulation 753/2002, wine's like the ones we have in our dataset can be considered medium or medium dry depending on if the sweetness gets balanced with acidity.

Finally, the dataset states that the lower the percentage of alcohol and sugar, the closest the density of the wine is to water. During the process of fermentation, the sugar is converted to alcohol, resulting on lower levels of density. The opposite to a wine with high density (or watery) would be a wine with gravity (with higher sugar/alcohol content).

While the full alcohol range in our dataset is between 8 and 14 degrees, most of the wines fall between 9 and 11 degrees of alcohol. Since "Vinho Verde" is an OD and not only a type of grape, we are willing to accept the asumption that different grapes and fermentations processes can result in different alcohol and residual sugar contents.

Sulphates additives help preventing wine oxidation and preserving the wine's freshness through higher sulfur dioxide (SO2) gas levels. However, too much SO2 will have implications on the wine taste and potentially, on the consumer's health.

Small quantities of citric acid adds freshness to the flavor. Too much citric acid can create an unpleasant vinegar-like taste.

Salt is commonly used in the culinary world as a flavor enhacer, which might affect the wine final

flavor and rating.

All of the above variables are different measures that can affect wine flavor, so they will be the main characteristics for the analysis in the following stages.

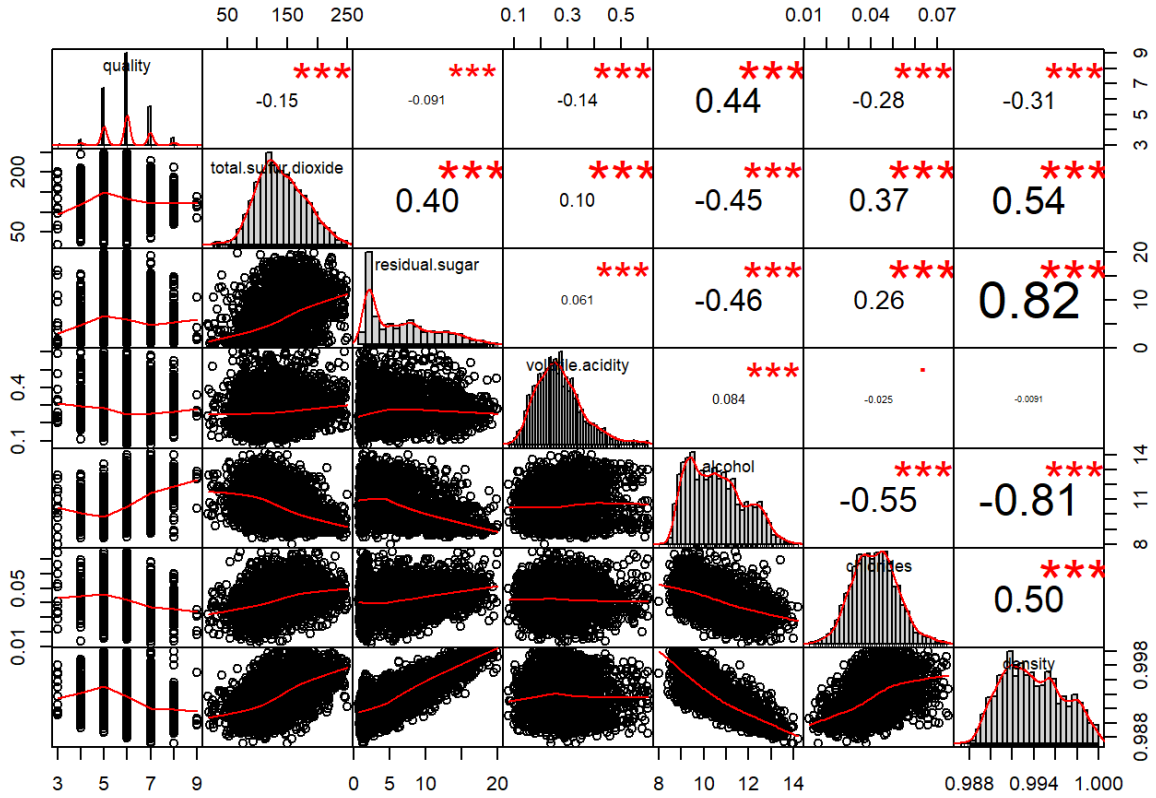
So what's next? First we are going to observe the correlation between variables to make sure that we are selecting the most relevant for the wine rating analysis.

Then, we will transform quality to a factor variable to be able to perform a better bivariate analysis and observe the relationship between characteristics and score and between the different characteristics.

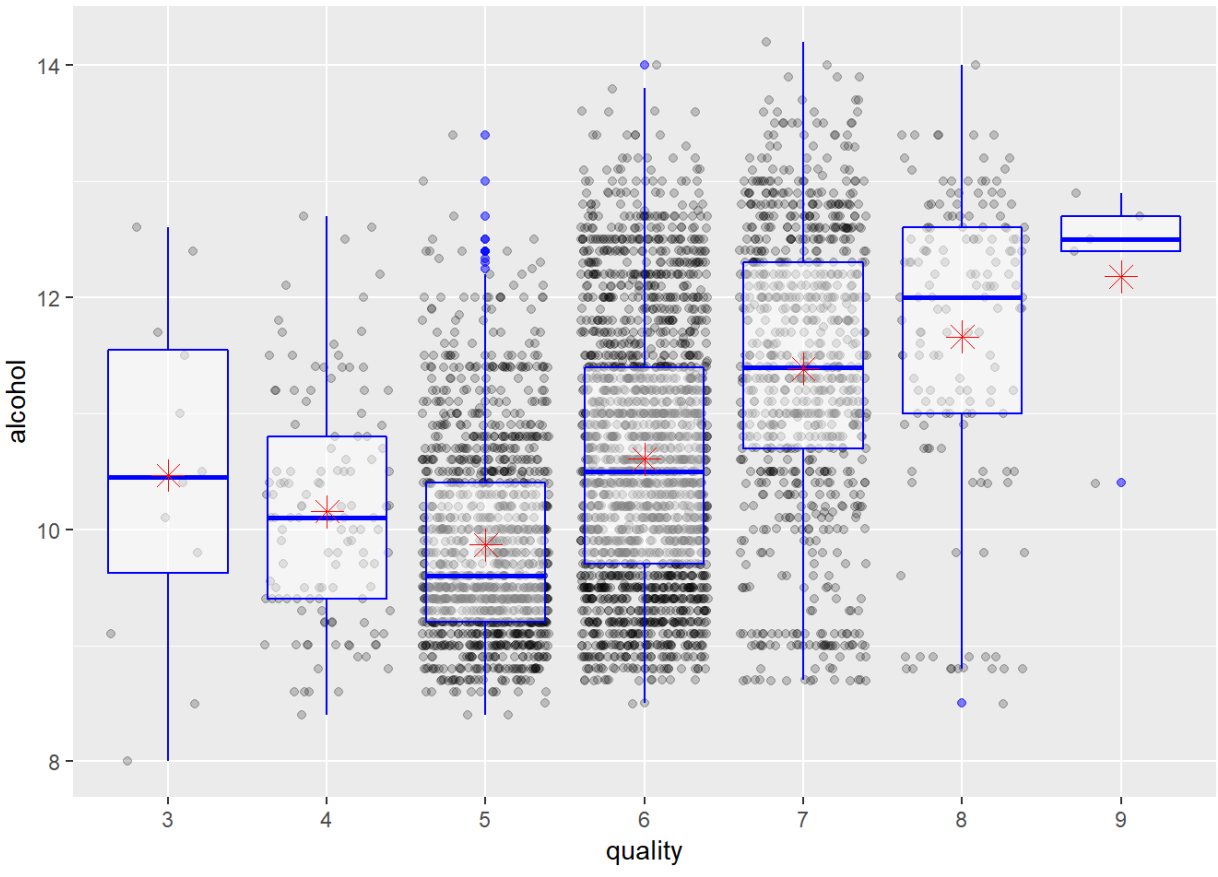
Bivariate Plots Section

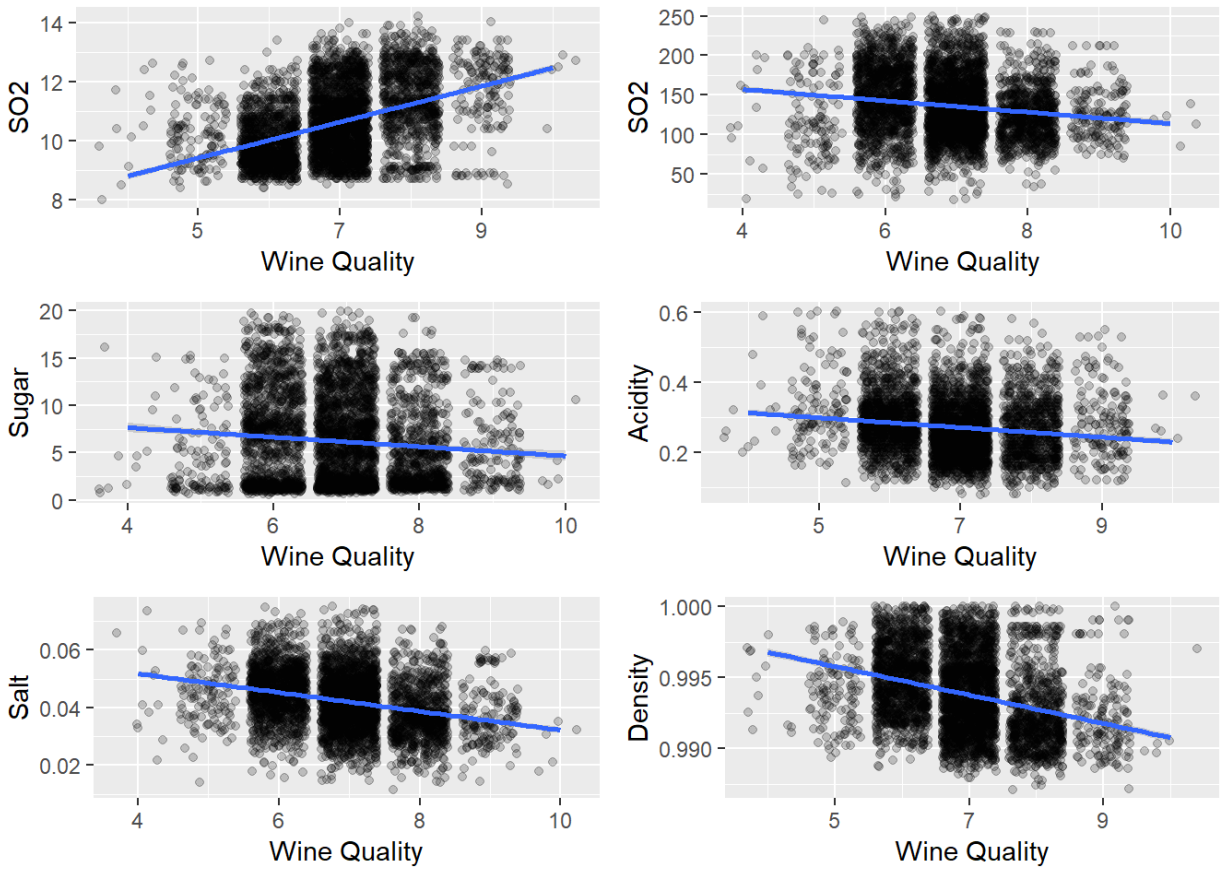
```
##                              [,1]
## fixed.acidity        -0.098914461
## volatile.acidity     -0.138726483
## citric.acid           0.001174011
## residual.sugar       -0.090958171
## chlorides            -0.281843923
## free.sulfur.dioxide   0.027770668
## total.sulfur.dioxide -0.154488846
## density              -0.311562641
## pH                    0.085547869
## sulphates             0.032478390
## alcohol               0.438875546
```

Here we can see the main correlation scores to quality. We now select the most important ones and create a correlation chart with them.



In the correlation chart we can observe the strong relationship between density and alcohol/sugar, as stated in the dataset documentation. Now is time to first analyze alcohol to wine quality score (as the main factor) and then to other relevant characteristics.





Bivariate Analysis

Surprisingly, alcohol seems to be the main driver on the quality score, with a positive correlation of 0.44. Salt, SO2 and sugar, on the other hand, have soft negative correlations with the score. At the sime time, these three variables have a moderate negative correlations with the alcohol % content and positive with density.

We can corroborate that density is opposite to alcohol and sugar in our chart, but we can also see that wines with high density (or low gravity) get worse scores.

Finally, the acidity measure seems to be mild, with no middle or strong relationship whatsoever.

In the next stage (multivariate analysis) we are going to investigate the relationship between the variables on a more complex level and try to build a predictive model to rate white wines.

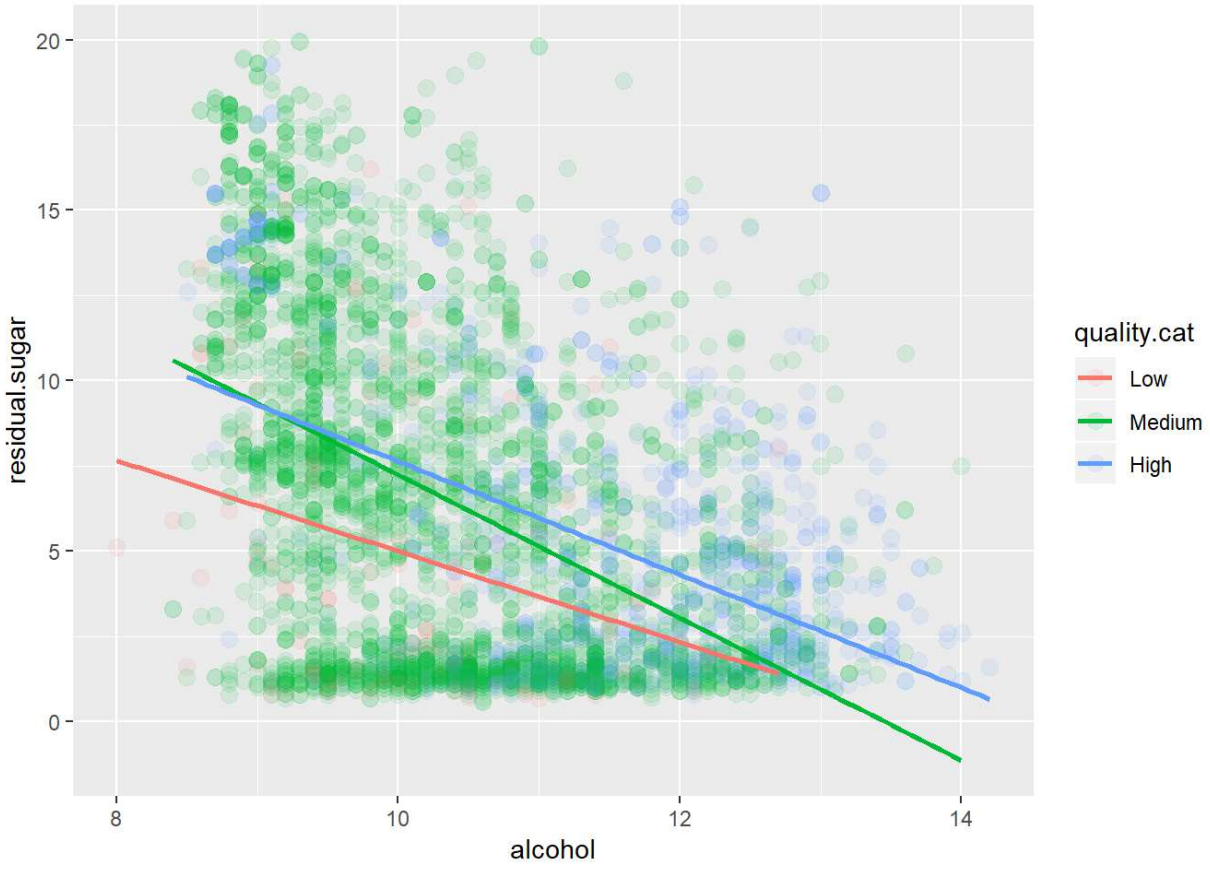
Multivariate Plots Section

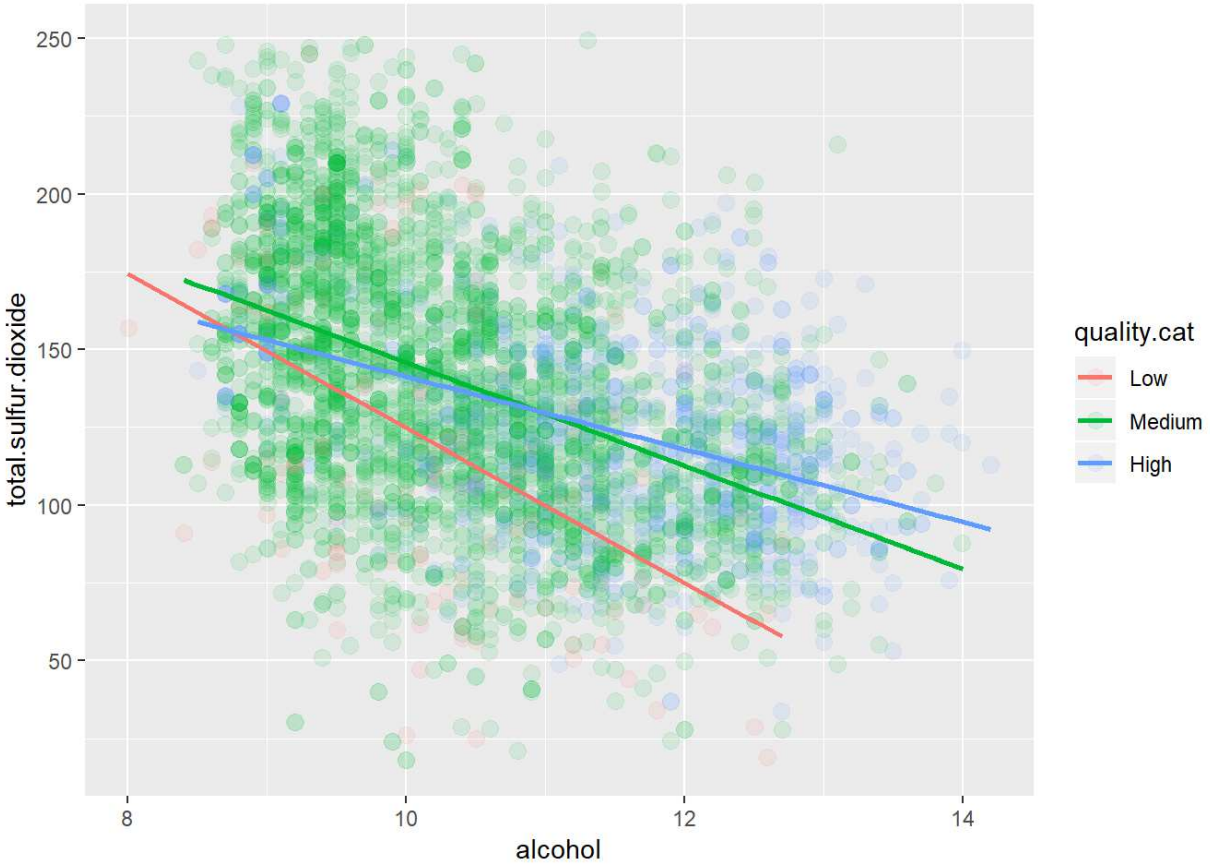
First we create a new variable, quality.cat, to classify the wines into three categories depending on the score.

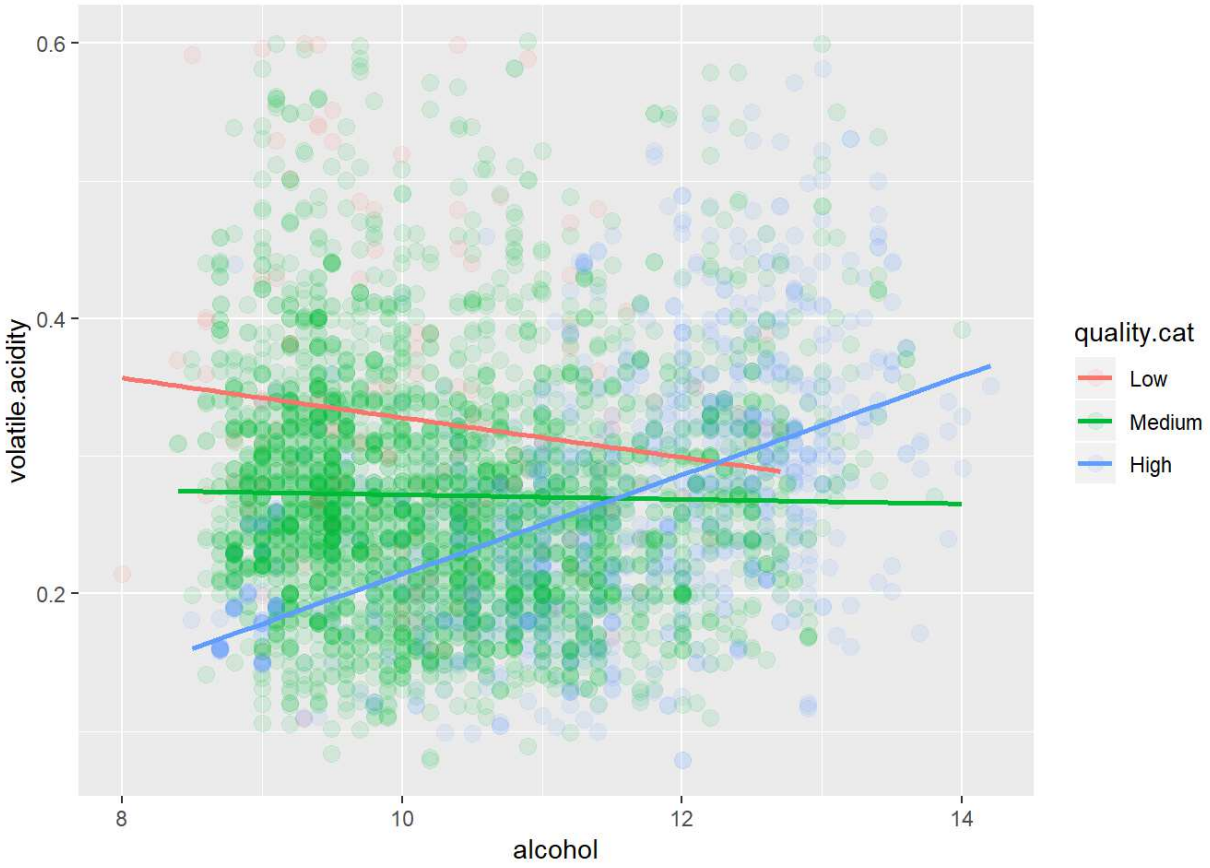
Low quality: Scores between 3 and 4 Medium quality: Scores between 5 and 6 High quality: scores between 7 and 9

```
##    Low Medium   High
##    136   3266   1005
```

Then we analyze the different relationships between quality, alcohol and other variables


residual.sugar
20
15
10
5
0
8
10
12
14
alcohol
quality.cat
Low
Medium
High


total.sulfur.dioxide
250
200
150
100
50
8
10
12
14
alcohol
quality.cat
Low
Medium
High



In the graph volatile acidity/ alcohol / quality graph above it can be seen how only for High quality wines there is a positive relationship between alcohol % content and acidity.

Now we will build our model:

```
##
## Calls:
## m1: lm(formula = as.numeric(quality) ~ alcohol, data = whitefinal)
## m2: lm(formula = as.numeric(quality) ~ alcohol + volatile.acidity,
##     data = whitefinal)
## m3: lm(formula = as.numeric(quality) ~ alcohol + volatile.acidity +
##     total.sulfur.dioxide, data = whitefinal)
## m4: lm(formula = as.numeric(quality) ~ alcohol + volatile.acidity +
##     total.sulfur.dioxide + residual.sugar, data = whitefinal)
## m5: lm(formula = as.numeric(quality) ~ alcohol + volatile.acidity +
##     total.sulfur.dioxide + residual.sugar + chlorides, data = whitefinal)
##
## ================================================================================
=============
##                               m1            m2            m3            m
4           m5
## --------------------------------------------------------------------------------
-------------
##   (Intercept)               3.582***      3.949***      3.426***      2.962**
*      3.425***
##                            (0.104)       (0.105)       (0.139)       (0.14
5)       (0.181)
##   alcohol                   0.316***      0.326***      0.355***      0.395**
*      0.372***
##                            (0.010)       (0.010)       (0.011)       (0.01
1)       (0.013)
##   volatile.acidity                       -1.768***     -1.889***     -1.993**
*     -1.992***
##                                          (0.133)       (0.134)       (0.13
3)       (0.133)
##   total.sulfur.dioxide                                  0.002***      0.001**
*      0.001***
##                                                        (0.000)       (0.00
0)       (0.000)
##   residual.sugar                                                       0.028**
*      0.028***
##                                                                      (0.00
3)       (0.003)
##   chloride
s                                                                       -5.764***
#
#
 (1.362)
## --------------------------------------------------------------------------------
-------------
##   R-squared                 0.193         0.224         0.230         0.24
7         0.250
##   adj. R-squared            0.192         0.223         0.229         0.24
6         0.249
##   sigma                     0.784         0.768         0.766         0.75
7         0.756
##   F                      1050.863       634.464       437.191       360.97
```

```
8       293.473
##   p                          0.000         0.000         0.000         0.00
0         0.000
##   Log-likelihood        -5177.505     -5091.036     -5074.419     -5023.848     -
5014.900
##   Deviance               2704.698      2600.617      2581.079      2522.517
2512.294
##   AIC                   10361.010     10190.072     10158.838     10059.695     1
0043.800
##   BIC                   10380.183     10215.636     10190.792     10098.041     1
0088.537
##   N                      4407          4407          4407          4407
4407
## ================================================================================
=============
```

Multivariate Analysis

From our plots we get that there seems to be a trade-off between alcohol and sugar, sulfates and acidity; where either the wines are high in alcohol content or in sugar/sulfates/acidity.

It can be seen as well as wines with higher quality usually have a higher amount of both parameters (alcohol + sugar, alchol + sulfates or alcohol + volatile acidity), as there seems to be a higher presence of quality wines on the upper right quadrants.

The lack of data on the right upper quadrant in the relationship between sugar and alcohol makes sense since during the fermentation, the sugar in the grapes becomes alcohol. High quality wines seem to have a sugar content higher than medium and low quality wines.

Wines with lower sulfates also seem to have lower quality scores than the average.

However, we can find that the combination of alcohol and acidity seems to hit the spot regarding the palate of our experts. According to Peynaud, a French oenologist, a dry wine can taste sweet if the alcohol level is elevated. This seems to be the make or break regarding our wines, as the high score wine show a tendence of having a balance betweeen both acidity and presence.

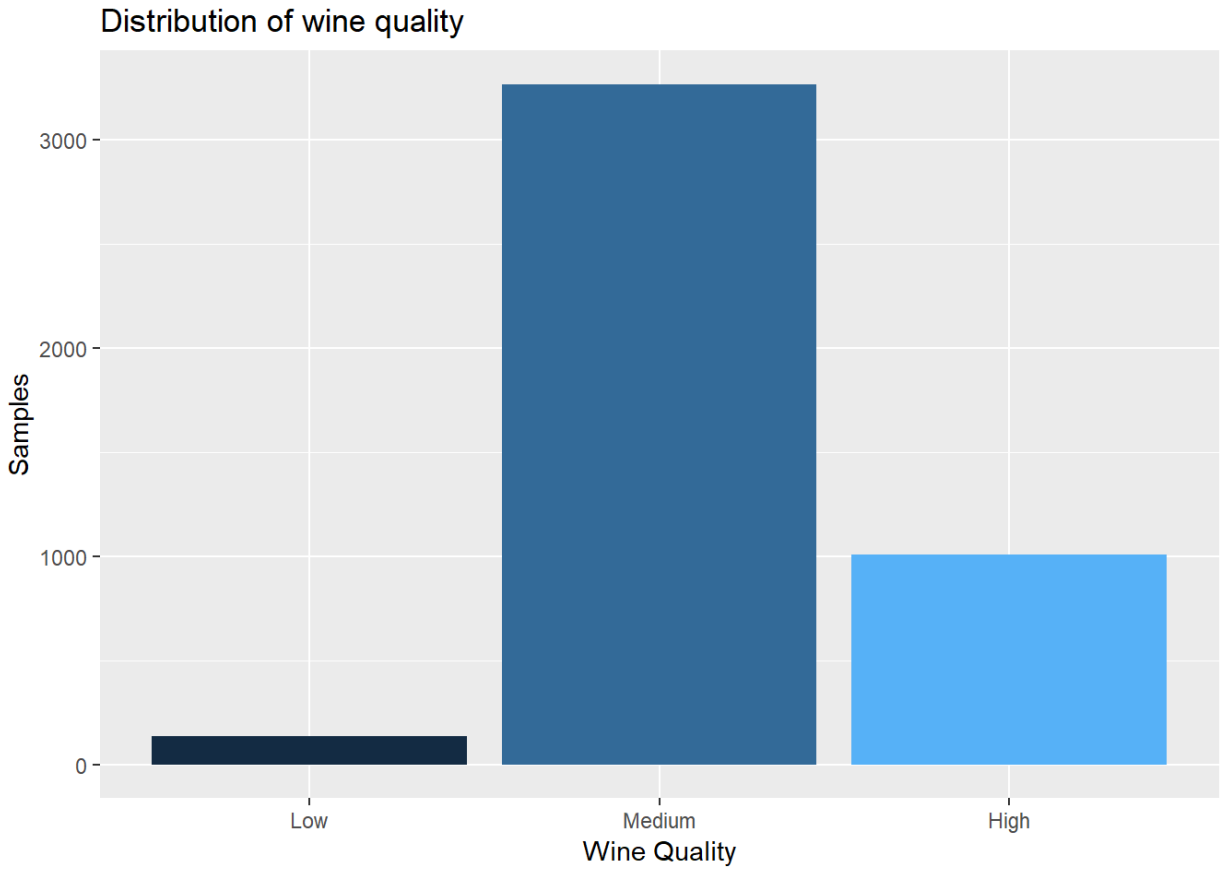
In the graph it can be seen as the upper right quadrant gets more importance on visualization; confirming that the combination of high levels of alcohol and acidity result in better wines.

Finally, after creating the linear model we can see disappointing results: the model built can only explain 25% of the score. This might be due not only to the underlying variable connections or the unbalanced dataset where most of the wines where average, but also due to the complexity of the wine rating. With no clear criteria on how to separate great wine from average or bad, it is not possible to create a predictive model.

Final Plots and Summary

We are now going to summarize the dataset exploration through three graphs and draw conclusions from our previous work

Plot One

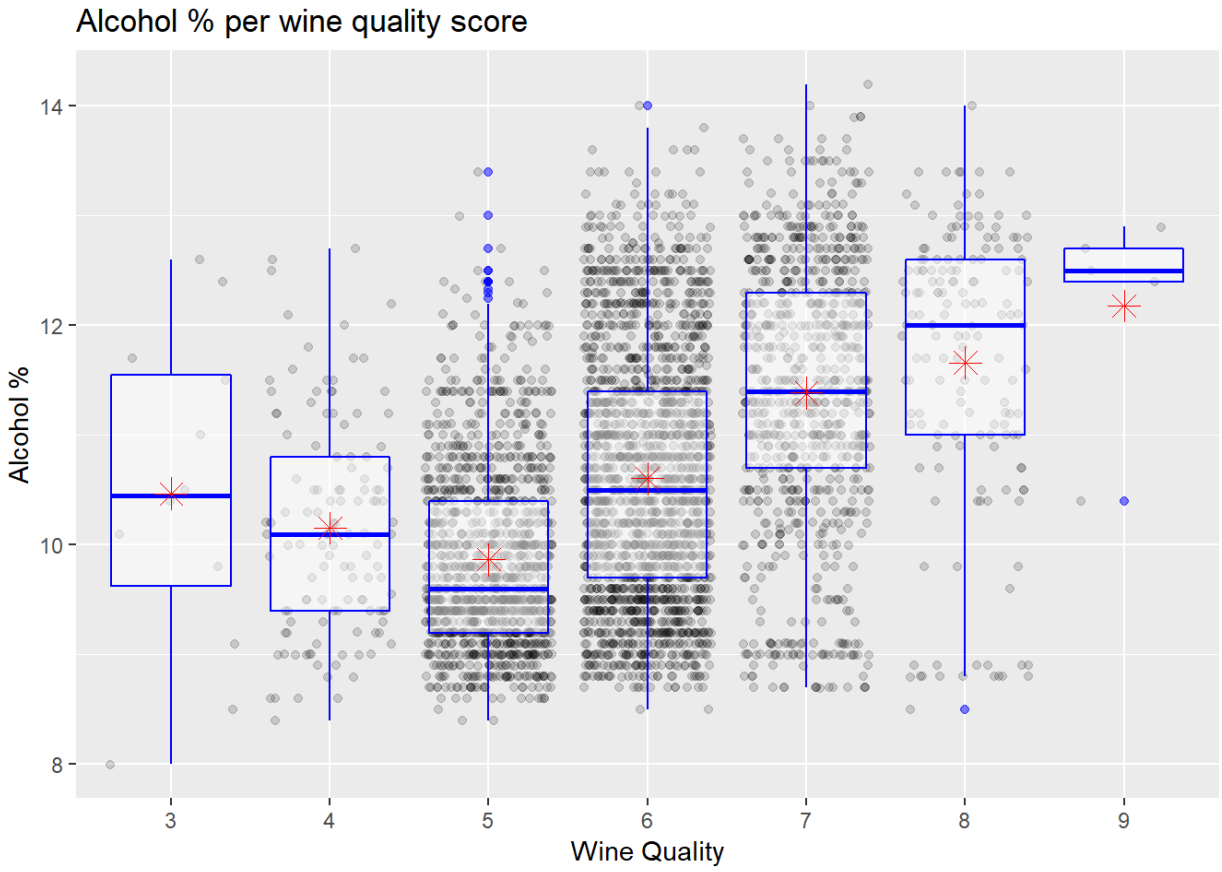


Description One

Since we aim to analyze the quality of “Vinho Verde” white wine, the first step would be to know how the quality is distributed on our dataset. Here we can see how the wine scores are distributed using our category

Of the 4407 observations we have after eliminating the outliers, 3.09% have a score of 3 to 4, 74.11% have a score between 5 and 6 and 22,8% have a score of 7 or more. However, only 3,77% have a score of 8 or more. This makes our dataset to be composed mostly by wines of medium quality, which might be a difficulty in order to find the characteristics that separates medium quality wines from premium wines.

Plot Two

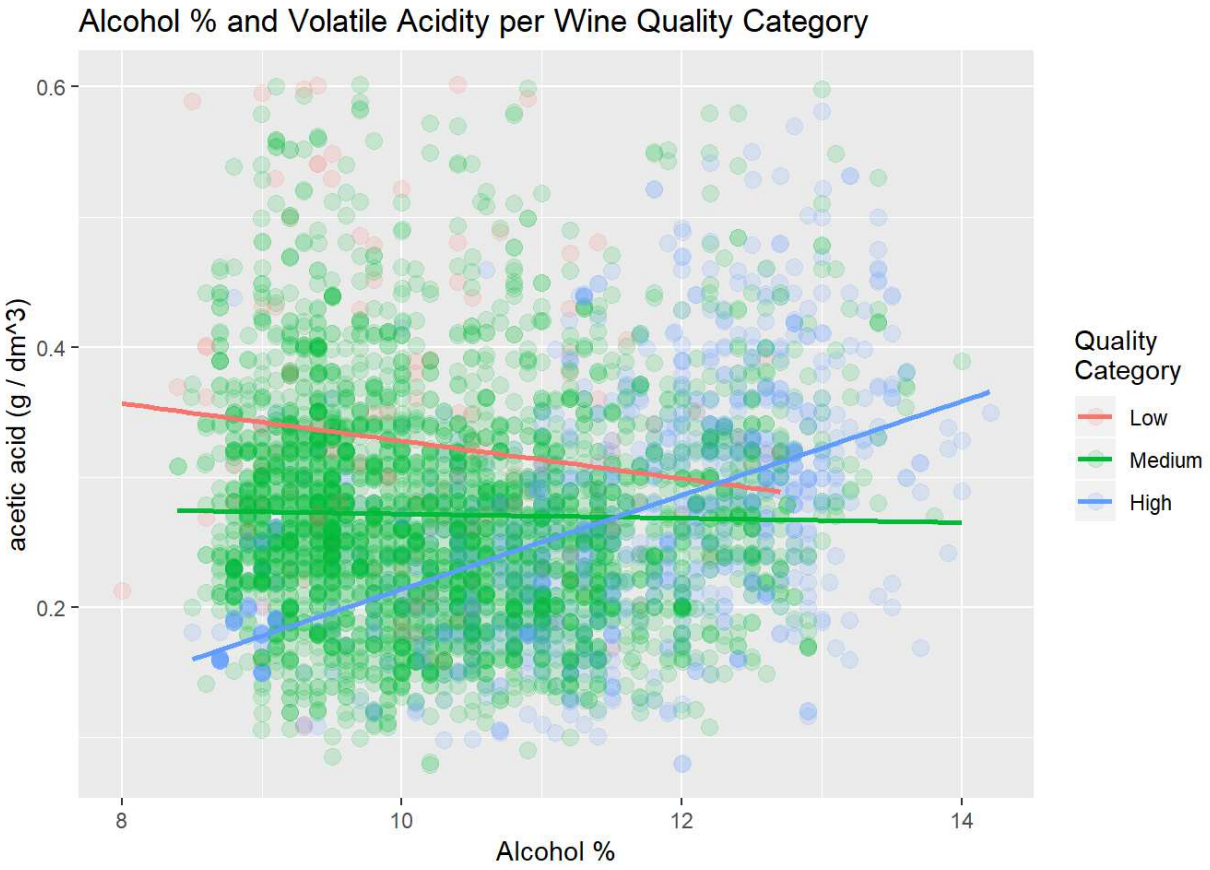


Description Two

In this second plot we can see the interaction between alcohol and wine quality, after finding that alcohol was the main driver of wine quality score, with a correlation of 0.44.

It can be seen how the alcohol content is higher in wines with a higher quality score. However, in the case of wines of high quality, we see how the average score is lower than the median, and even falls ouside of the box. This makes us think that we need a higher sample of high quality wines in order to find a more equal distribution.

Plot Three



Description Three

Finally, in this graph it can be seen the relationship between alcohol and volatile acidity, other of the variables with higher correlation with wine quality score. What makes this graph stand out is that shows more wines with high scores in the right upper quadrant than the other multivariate analysis, signaling a positive relationship between alcohol and acidity in the case of great wines.

Reflection

Regarding this analysis, the main hindrances were the faulty dataset and the complexity of how the variables are intertwined. Even analyzing the variables with higher correlation with the score we have been able to develop only a model able to explain only a 25% of the score, and it's still not clear to me how to separate good wine from bad.

Maybe a larger dataset, with more observations of extreme quality scores and and the mean score and not the median would be more useful for this kind of analysis. In that sense, I find hard to believe that the alcohol is the main driver of wine taste and I wish I knew more about wines to be able to create relationship between similar variables, like acidity, sulphates and such. Maybe that way I could do a better job determining what makes a great wine.

Overall, I enjoyed doing this project but it leaves a sour aftertaste.