White Wine Quality EDA by Prithwish Ganguly

This report explores a dataset containing quality and attributes for 4898 white wines. This dataset is a variant of the Portuguese “Vinho Verde” wine.

Univariate Plots Section

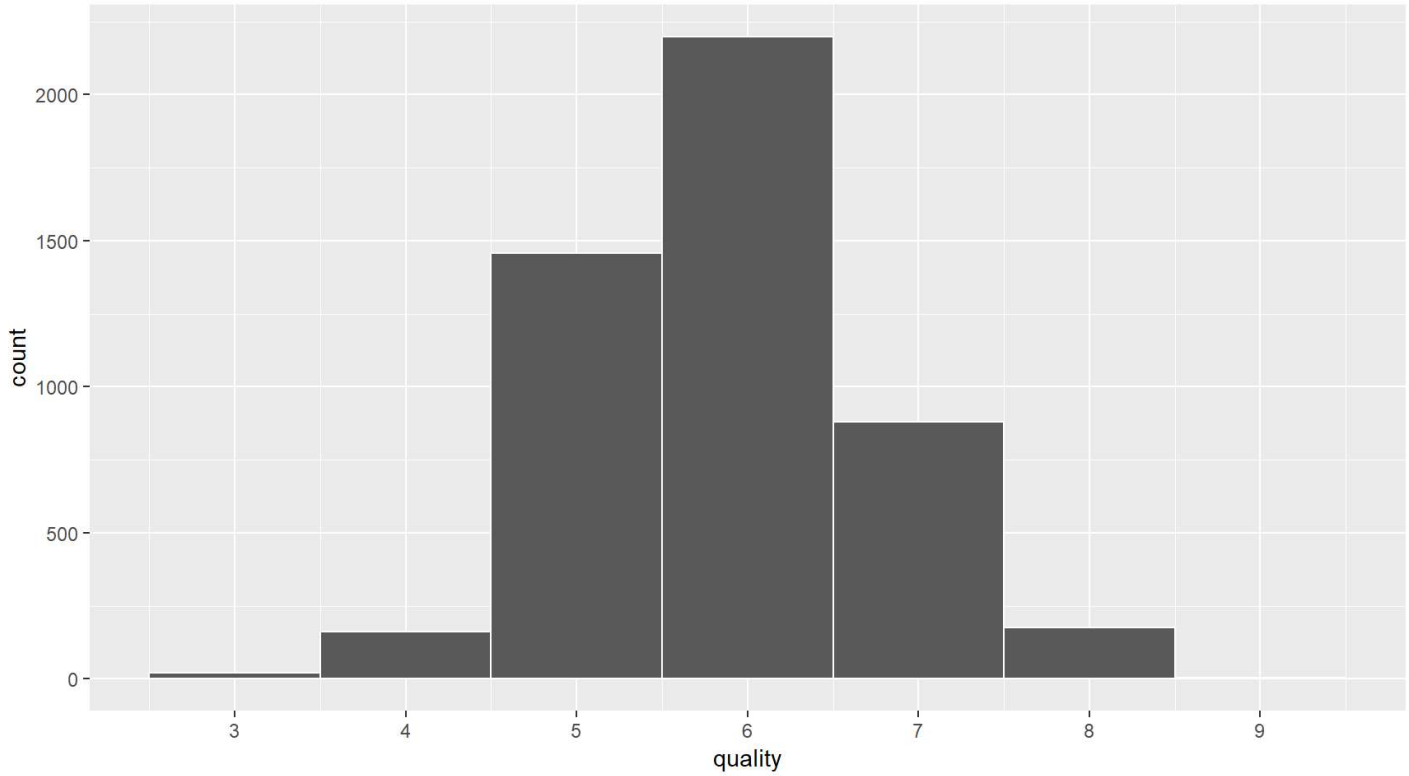
```
## [1] 4898   12
```

Our dataset has 4898 observations and 13 variables namely:

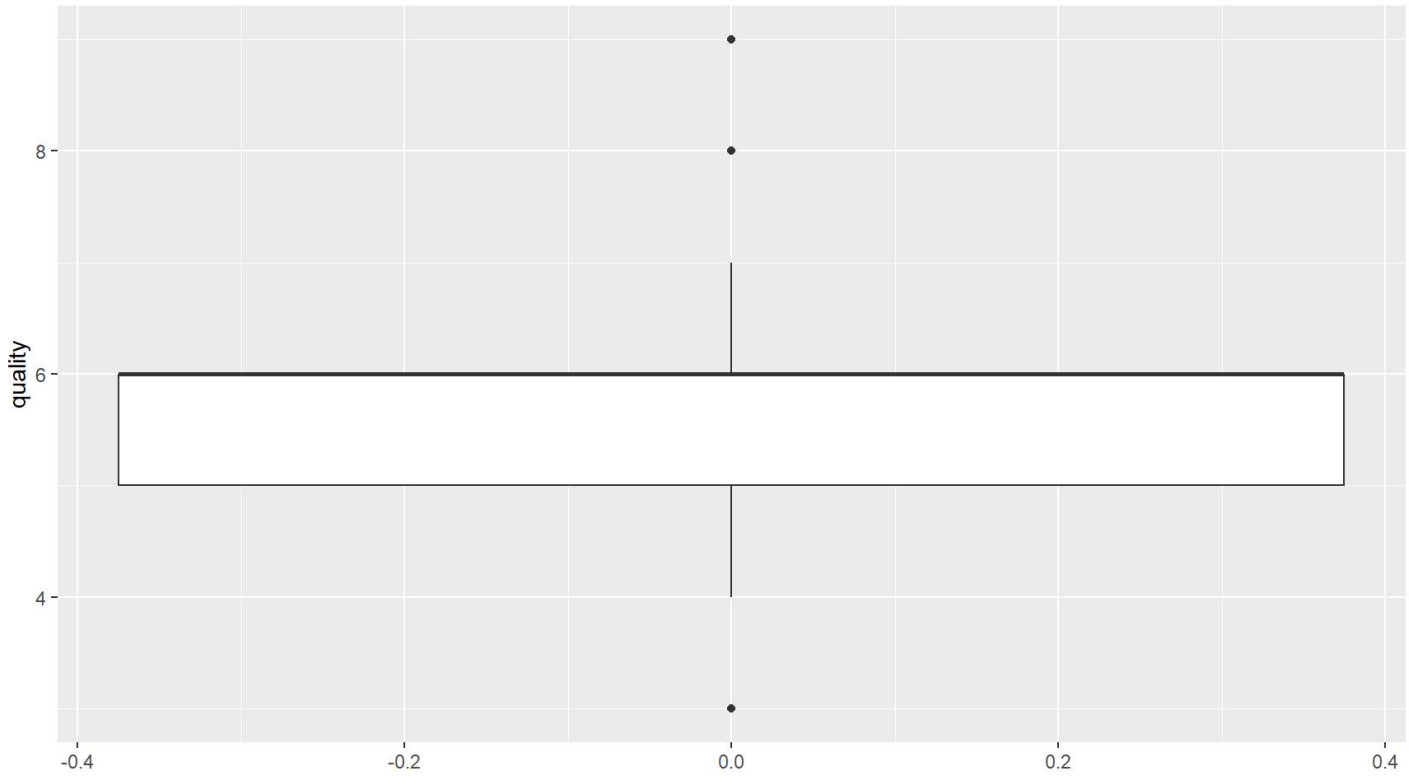
```
## 'data.frame':    4898 obs. of  12 variables:
##  $ fixed.acidity       : num  7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...
##  $ volatile.acidity    : num  0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 ...
##  $ citric.acid         : num  0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34 0.43 ...
##  $ 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             : int  6 6 6 6 6 6 6 6 6 6 ...
```

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

'x' is just the rowname so its summary is irrelevant. One of the things that caught my eye at first was the 5point summary of $quality.



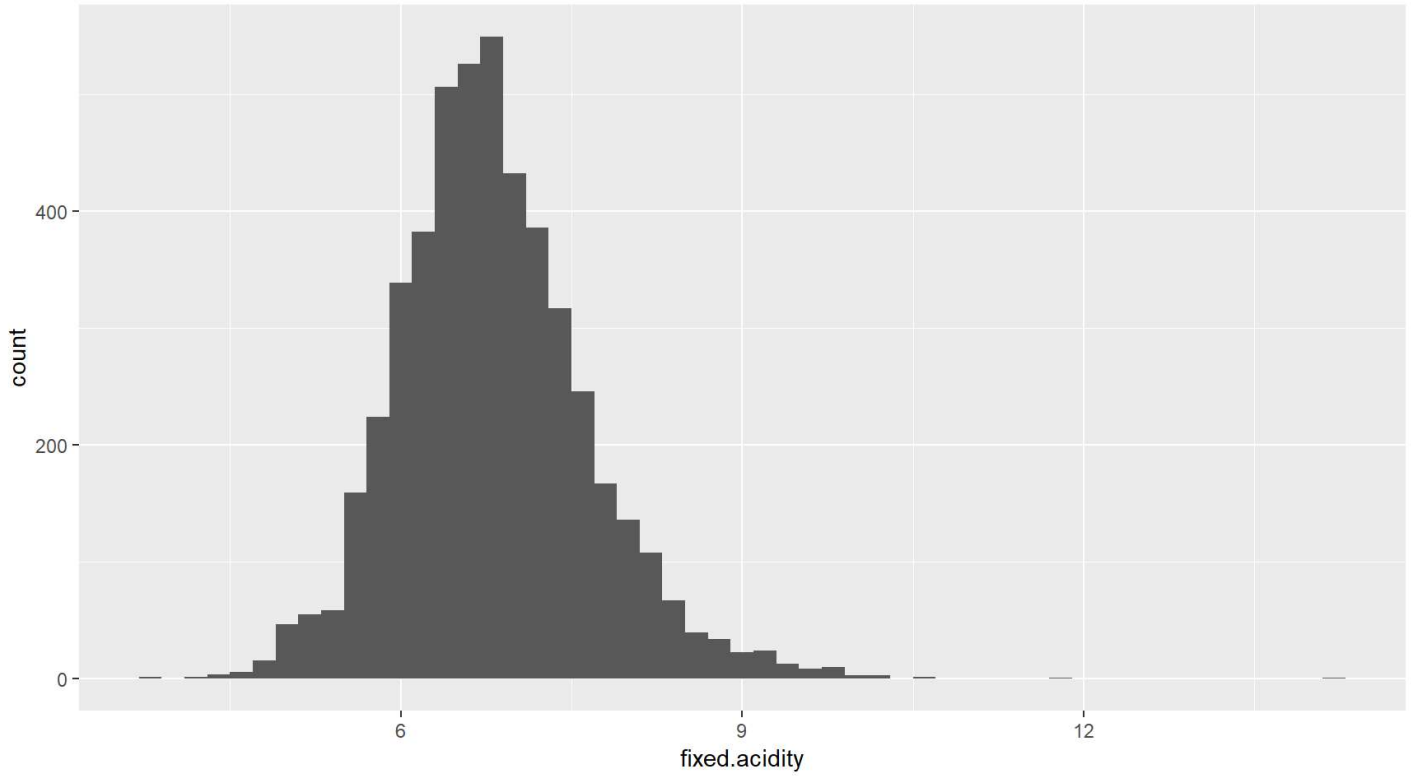
So, quality is measured in discrete values and has a near perfect normal distribution. The quality of wines in this dataset ranging between 3 and 9 and most wines have a rating of 6.



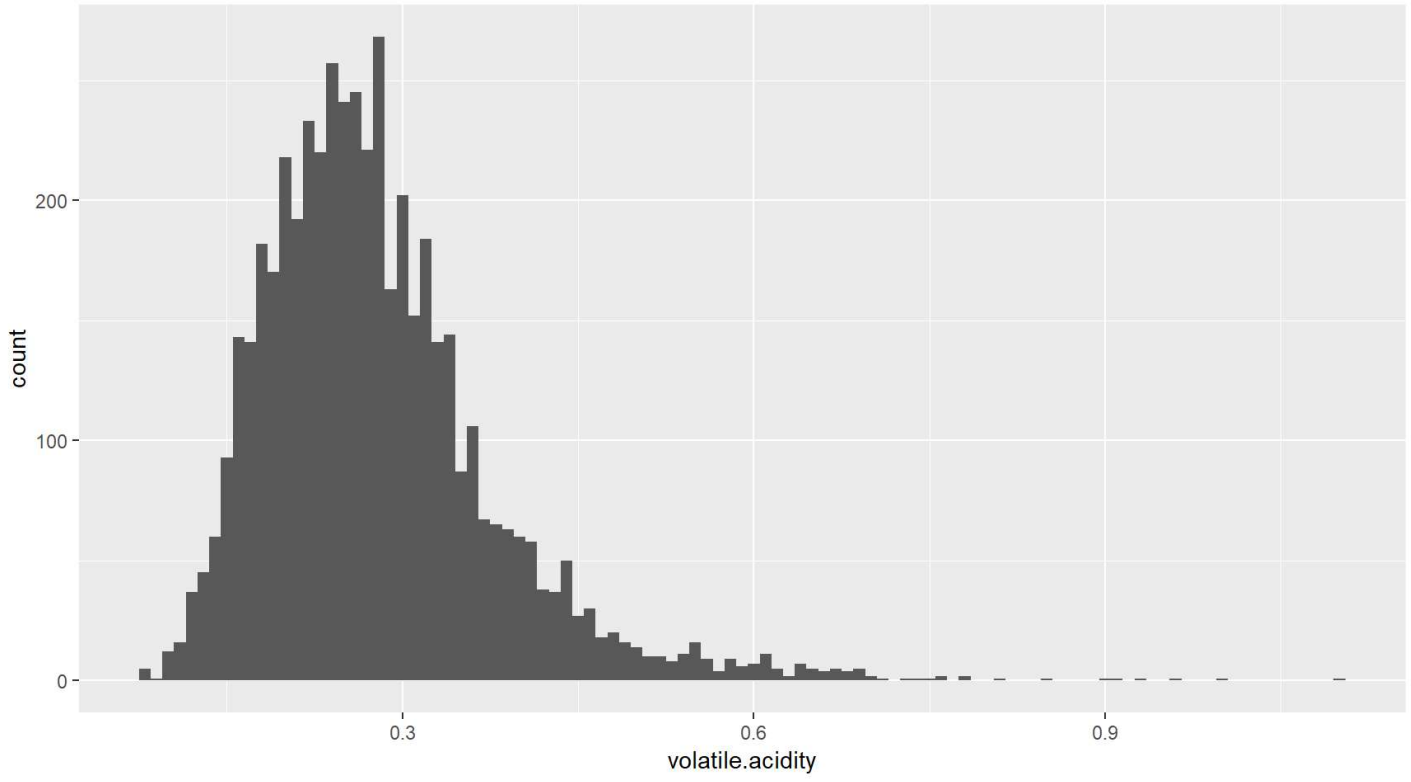
We can observe here that, there are easily more 'normal quality' wines that 'poor quality' or 'excellent quality'. The outliers at the lower end being the 'poor quality' wines and upper end being the 'excellent quality'. It would be interresting to group the wines based on these 3 categories rather than their discrete values.

```
##
##    3    4    5    6    7    8    9
##   20  163 1457 2198  880  175    5
```

Considering any rating below 4 being a poor wine and any quality 8 and above being excellent wine. We have 20 poor wines(1%) and 180 excellent wines(9.5%) against 1698 normal wines(89.5%)

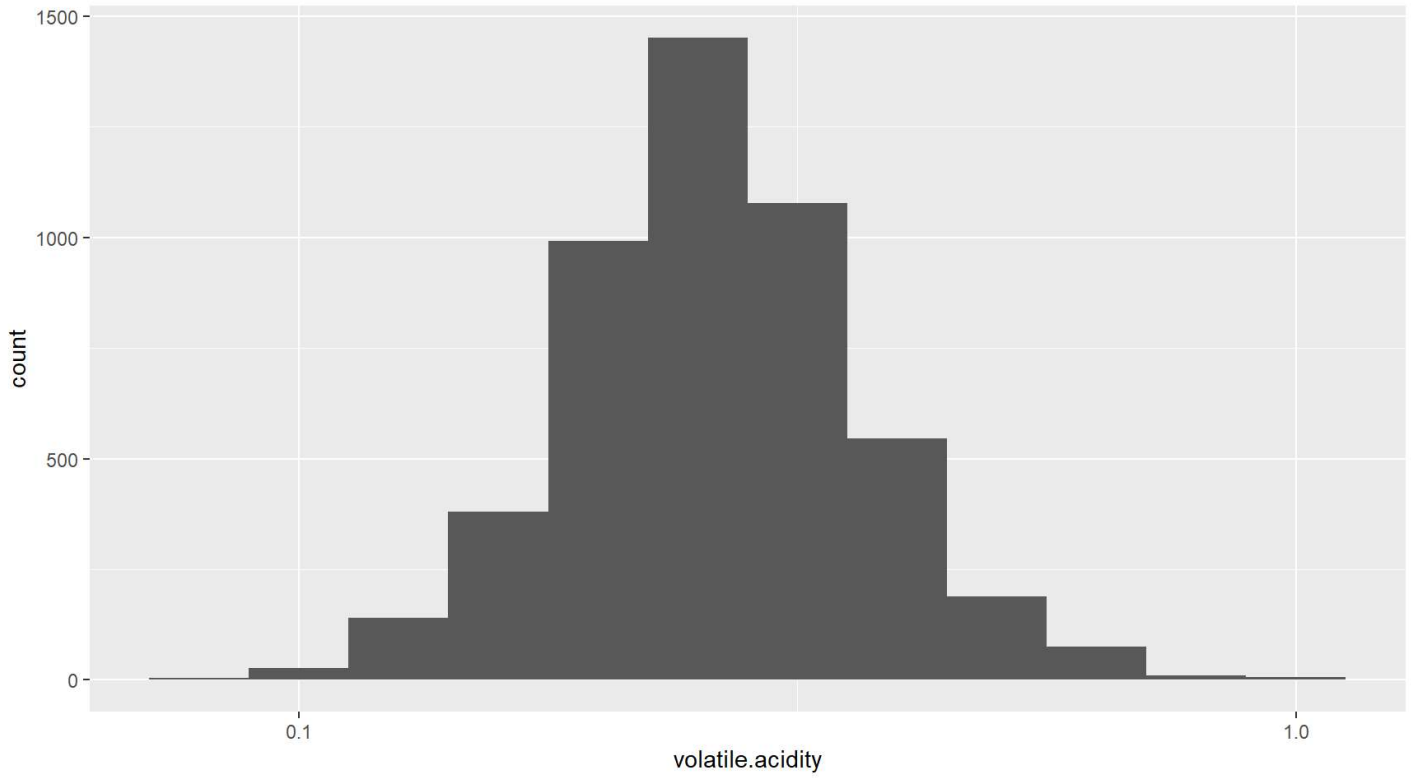


The fixed.acidity distribution is normally distributed about a mean of 6.855.

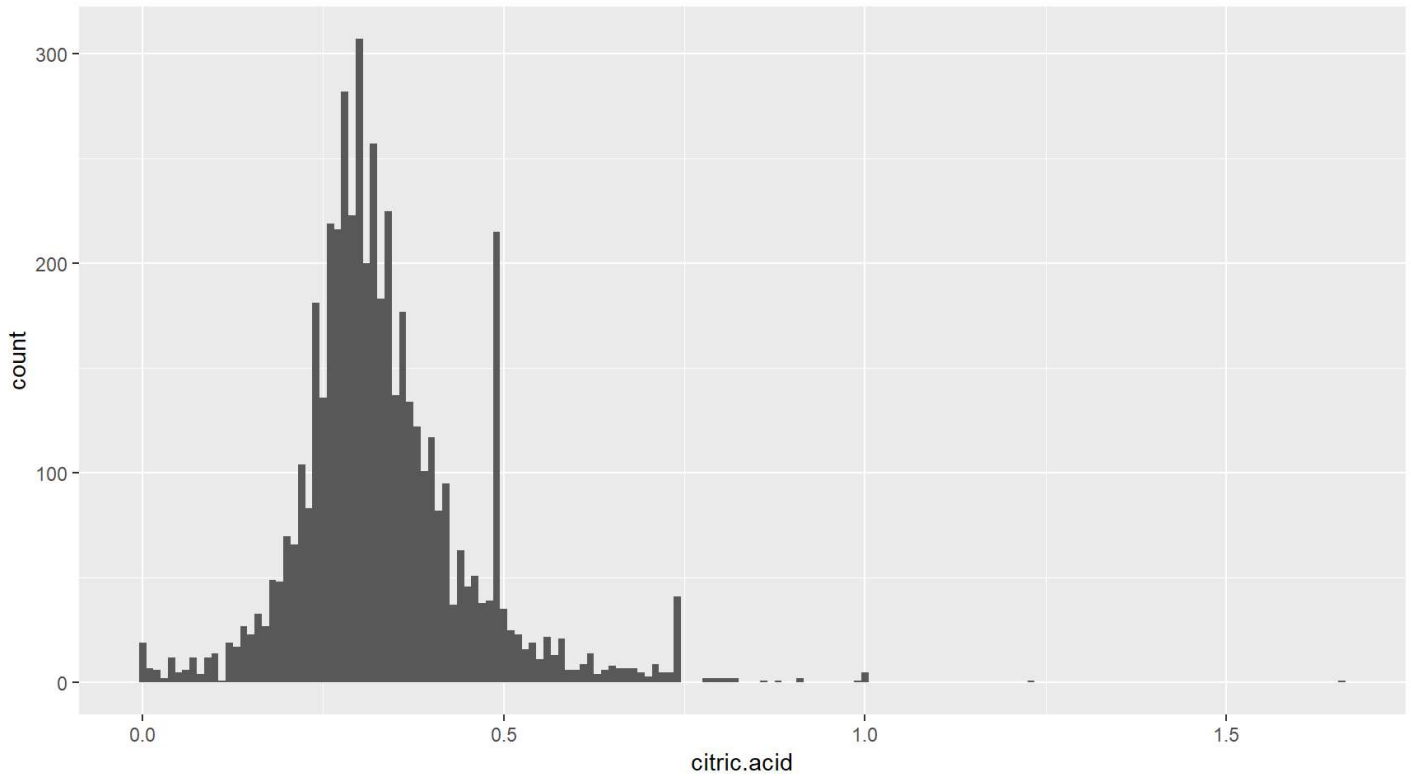


```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0800  0.2100  0.2600  0.2782  0.3200  1.1000
```

The volatile.acidity distribution looks about normal about the mean 0.2782.

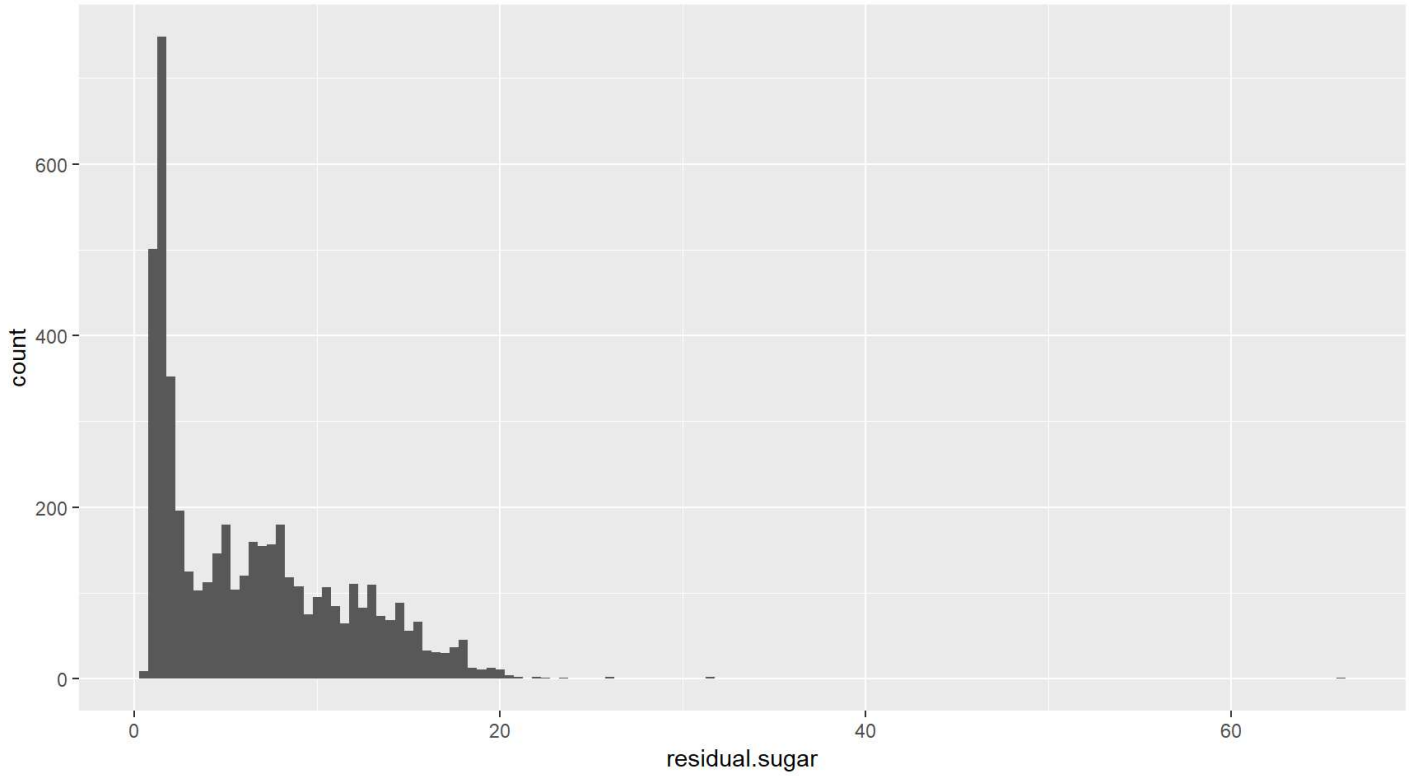


I transformed the volatile.acidity data to look at its distribution and its transformation perfectly normally distributed.

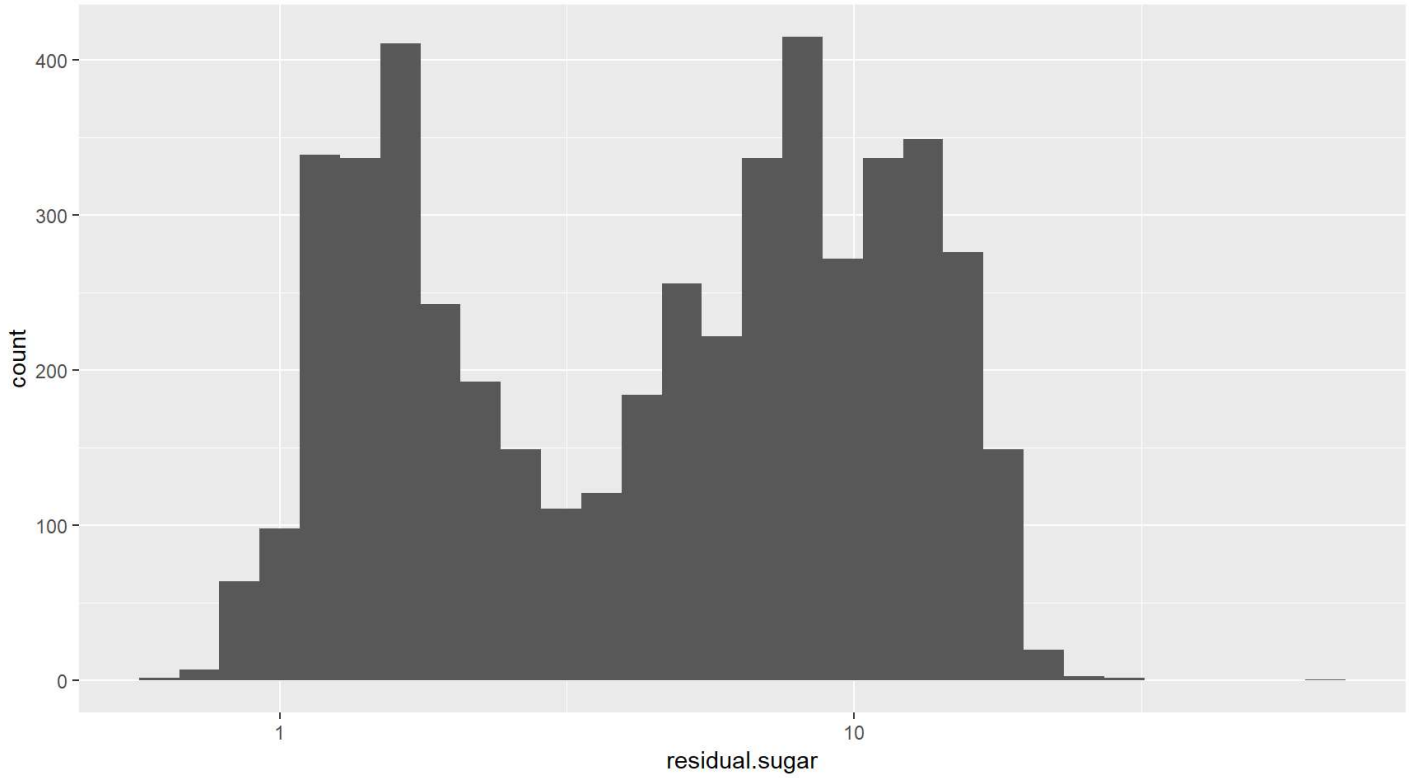


```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.2700  0.3200  0.3342  0.3900  1.6600
```

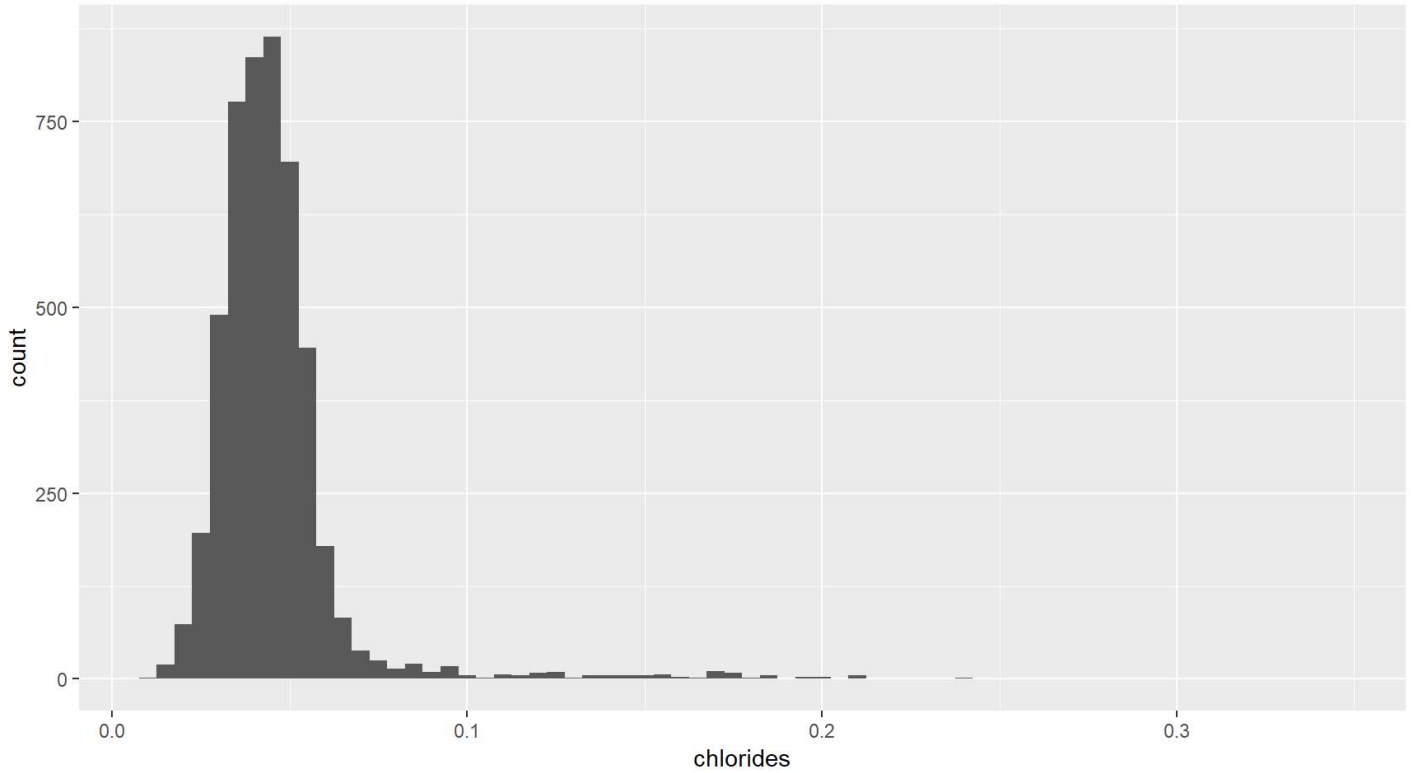
The distribution looks normal about a mean of 0.3342 but there is a wierd peak at 0.5.



The residual sugar has a long right tail, it would be interesting to transform it to a log scale

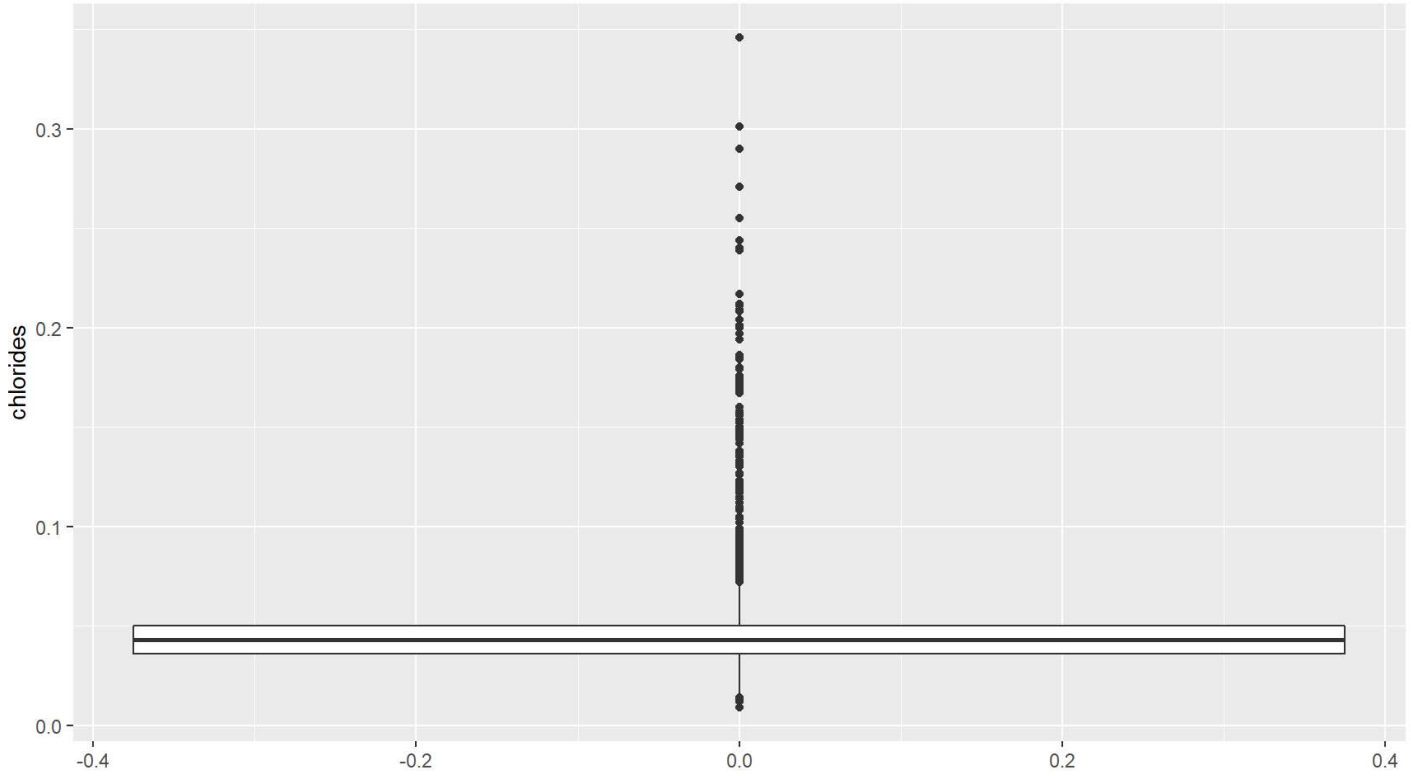


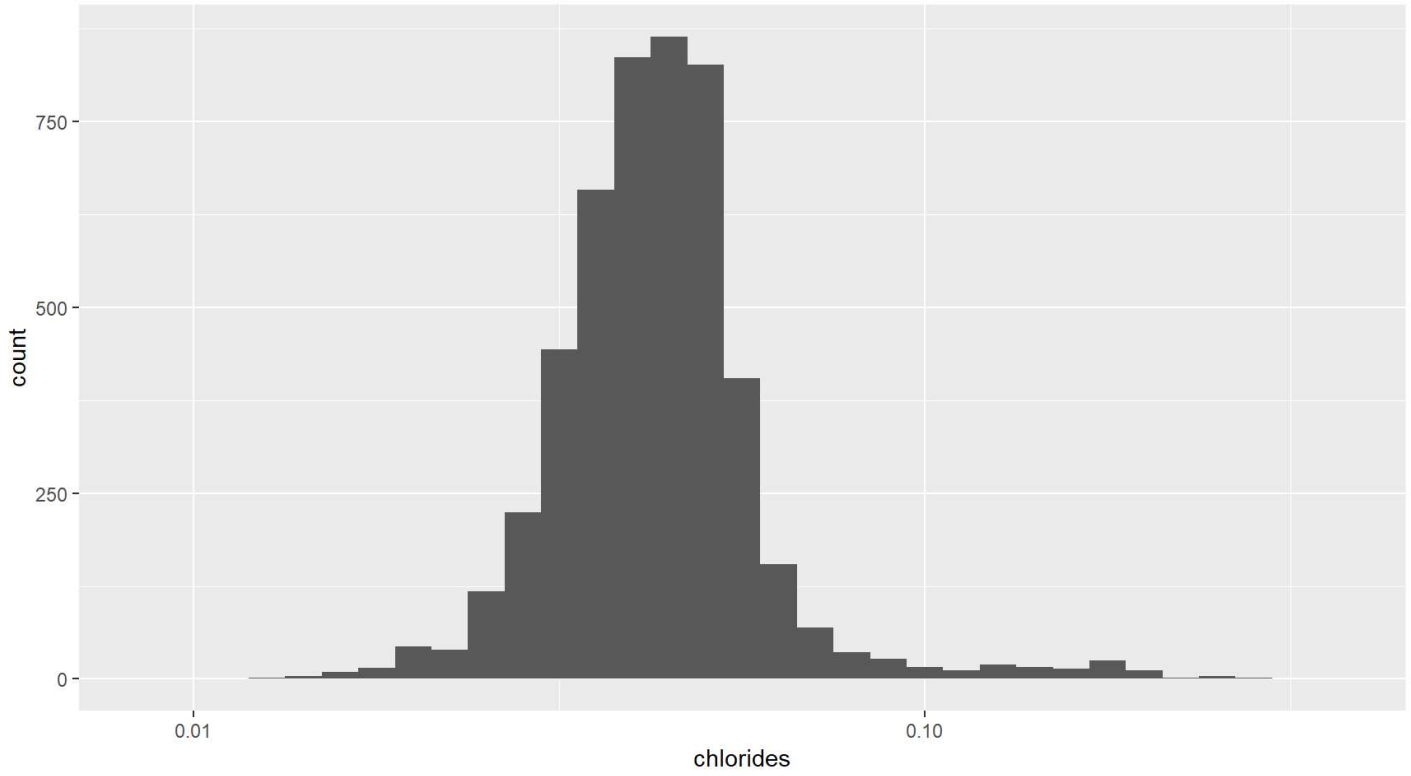
This transformation of 'residual.sugar' appears to be bimodal in nature, peaking first at around 1.75 then again at around 9.



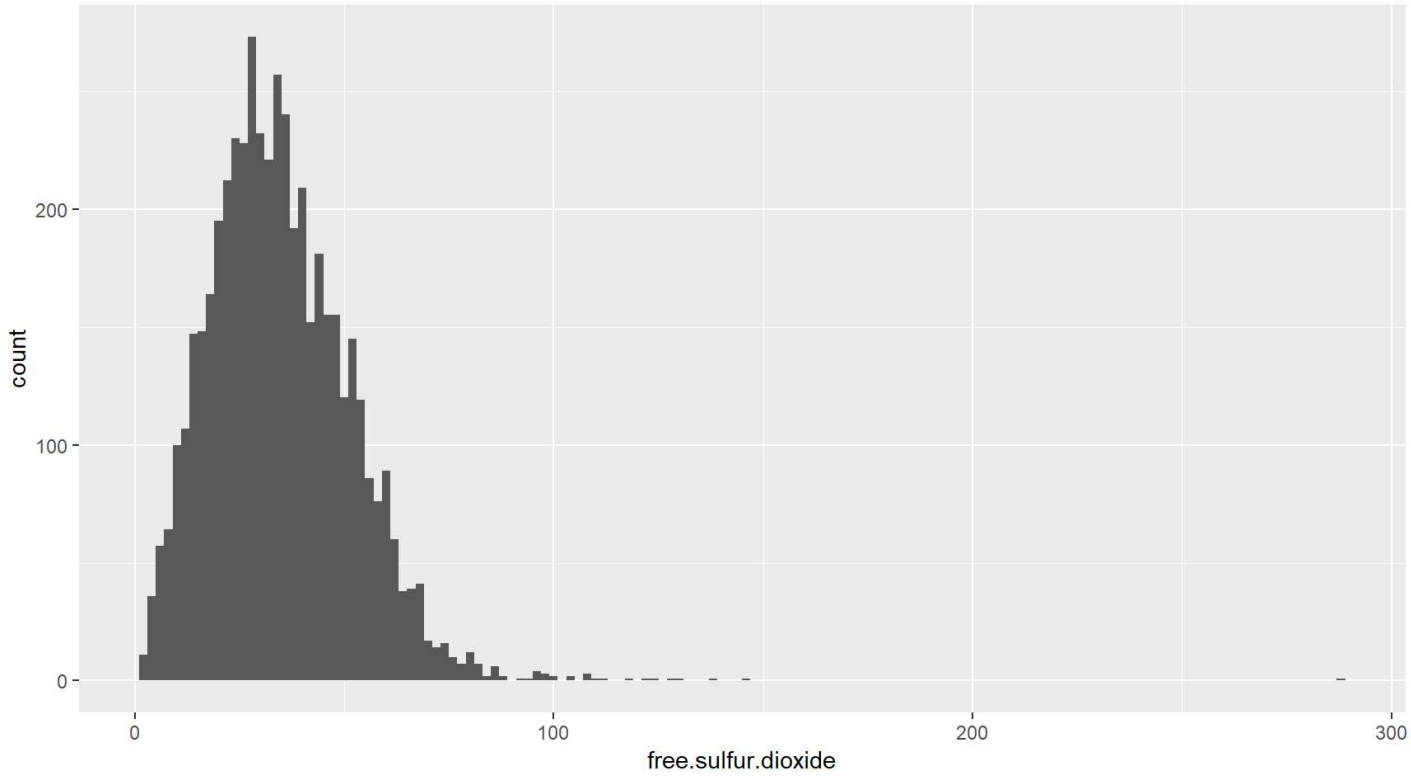

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00900 0.03600 0.04300 0.04577 0.05000 0.34600
```

The chlorides distribution is also normal but has a lot of outliers as show below.





The log transformation of chlorides is also normally distributed.

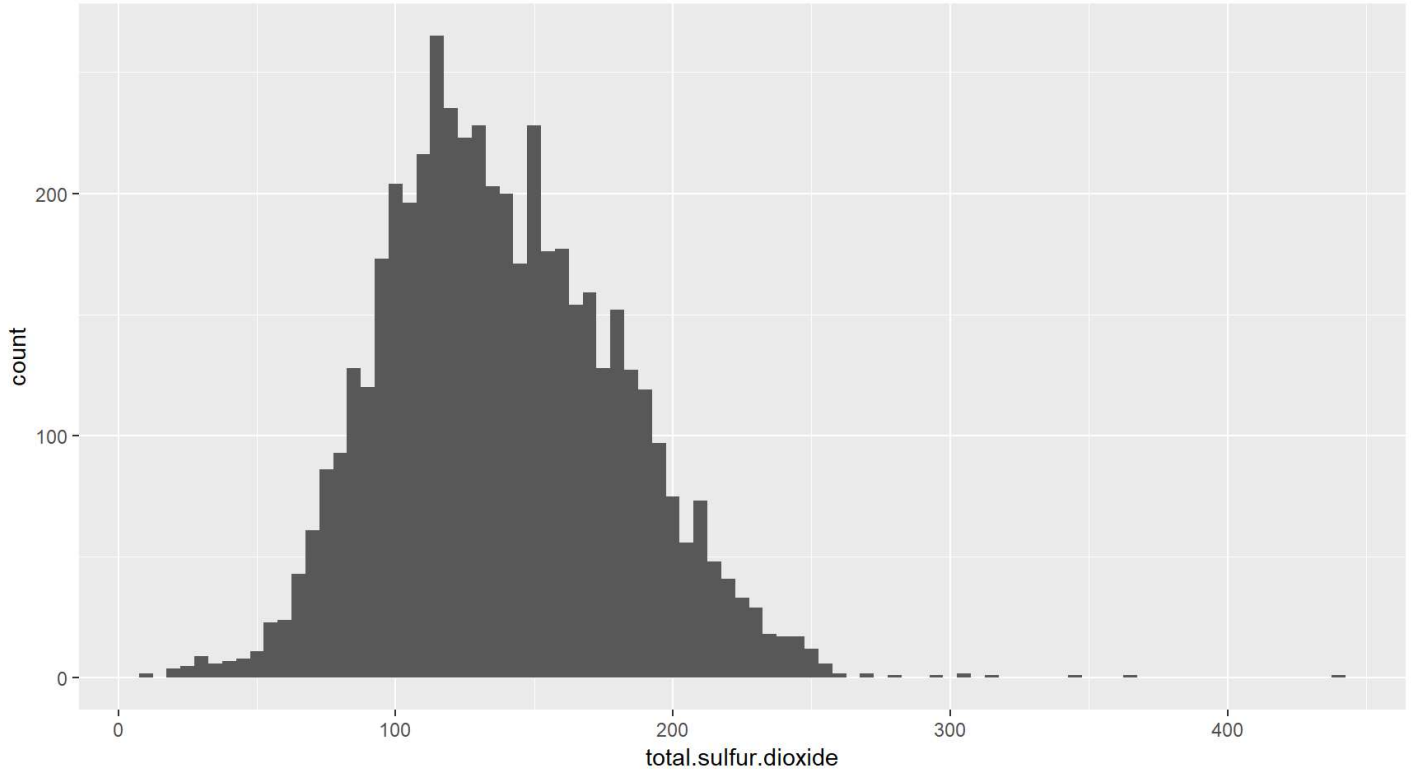



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00   23.00   34.00   35.31   46.00  289.00
```

The free.sulfur.dioxide distribution is also normal about a mean of 35.31. There are quite a few outliers but there is one particularly high one at 289. I am really interested by the one wine that had 289 free.sulfur.dioxide.

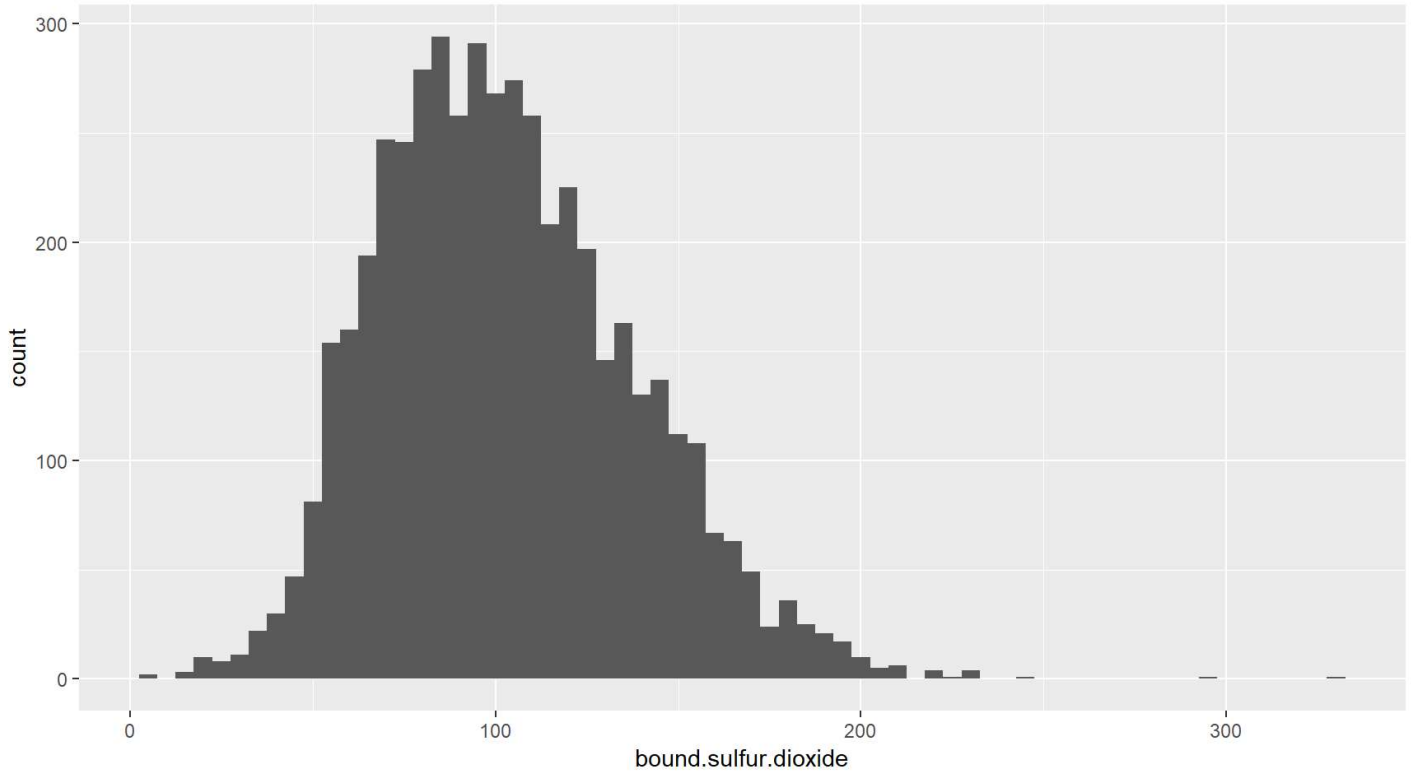
```
##      fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 4746           6.1             0.26        0.25            2.9     0.047
##      free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates
## 4746                 289                  440 0.99314 3.44      0.64
##      alcohol quality
## 4746    10.5       3
```

Oh, this belonged to one of the low quality wines! This wine also has the highest total.sulfur.dioxide in our dataset which in itself is another huge outlier. Maybe presence of high quantities of sulfur dioxide reduces the quality of wine? Will have to look into this later!



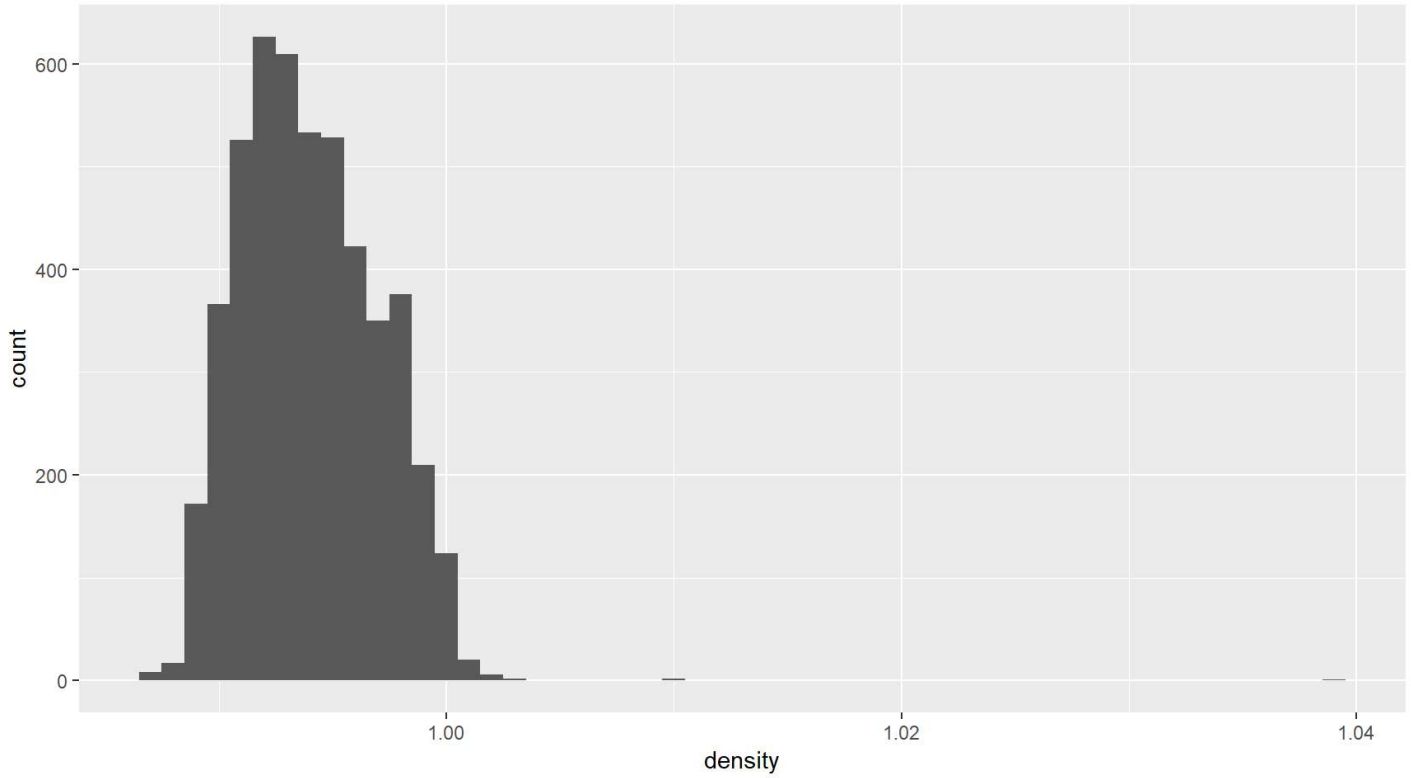

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     9.0   108.0   134.0   138.4   167.0   440.0
```

Again total.sulfur.dioxide also seems to be normally distributed with about a mean of 138.4. Now by definition, total.sulfur.dioxide = free.sulfur.dioxide + bound.sulfur.dioxide. I will now make a variable called bound.sulfur.dioxide = total.sulfur.dioxide - free.sulfur.dioxide

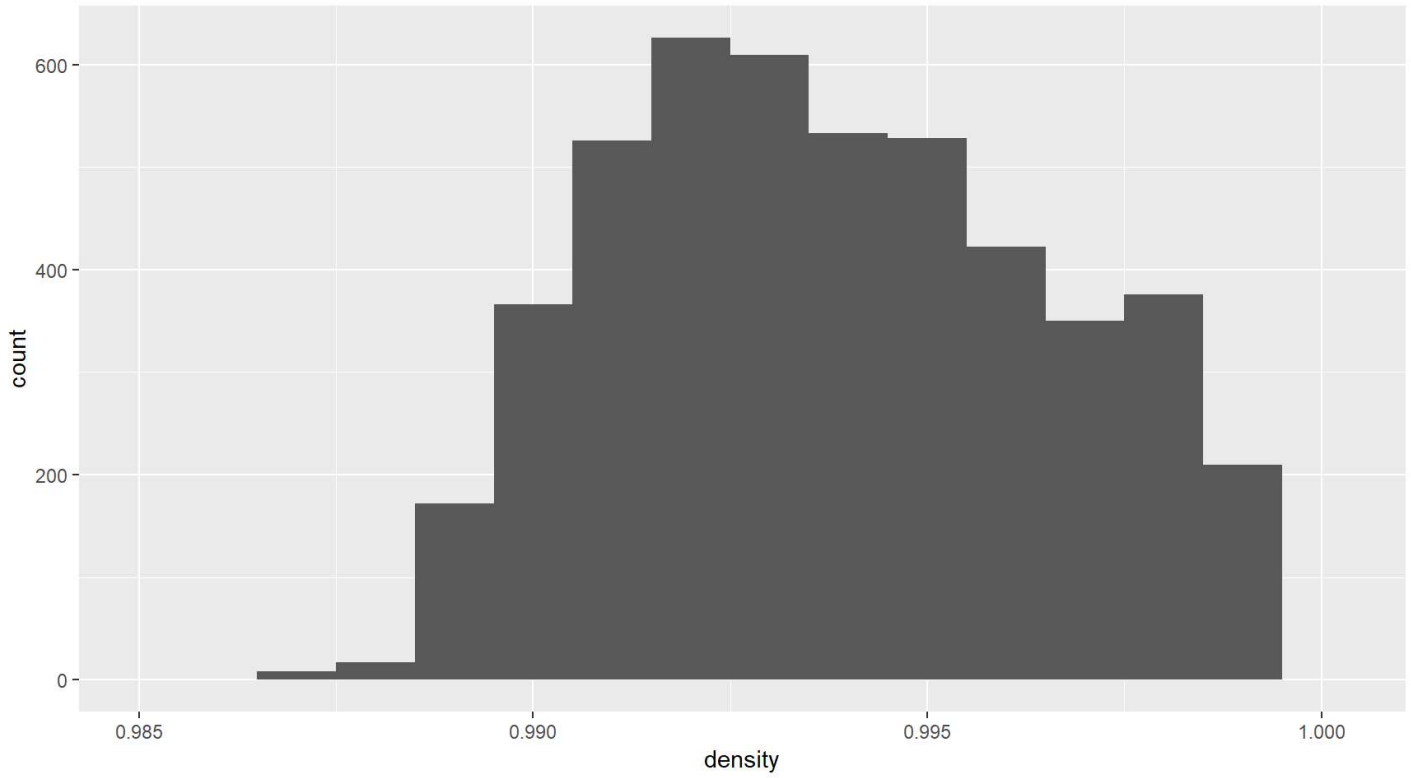



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     4.0    78.0   100.0   103.1   125.0   331.0
```

Bound.sulfur.dioxide is also normally distributed and looks similar to the other sulfur dioxide distributions. The mean of this distribution is 103.1.



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9871  0.9917  0.9937  0.9940  0.9961  1.0390
```

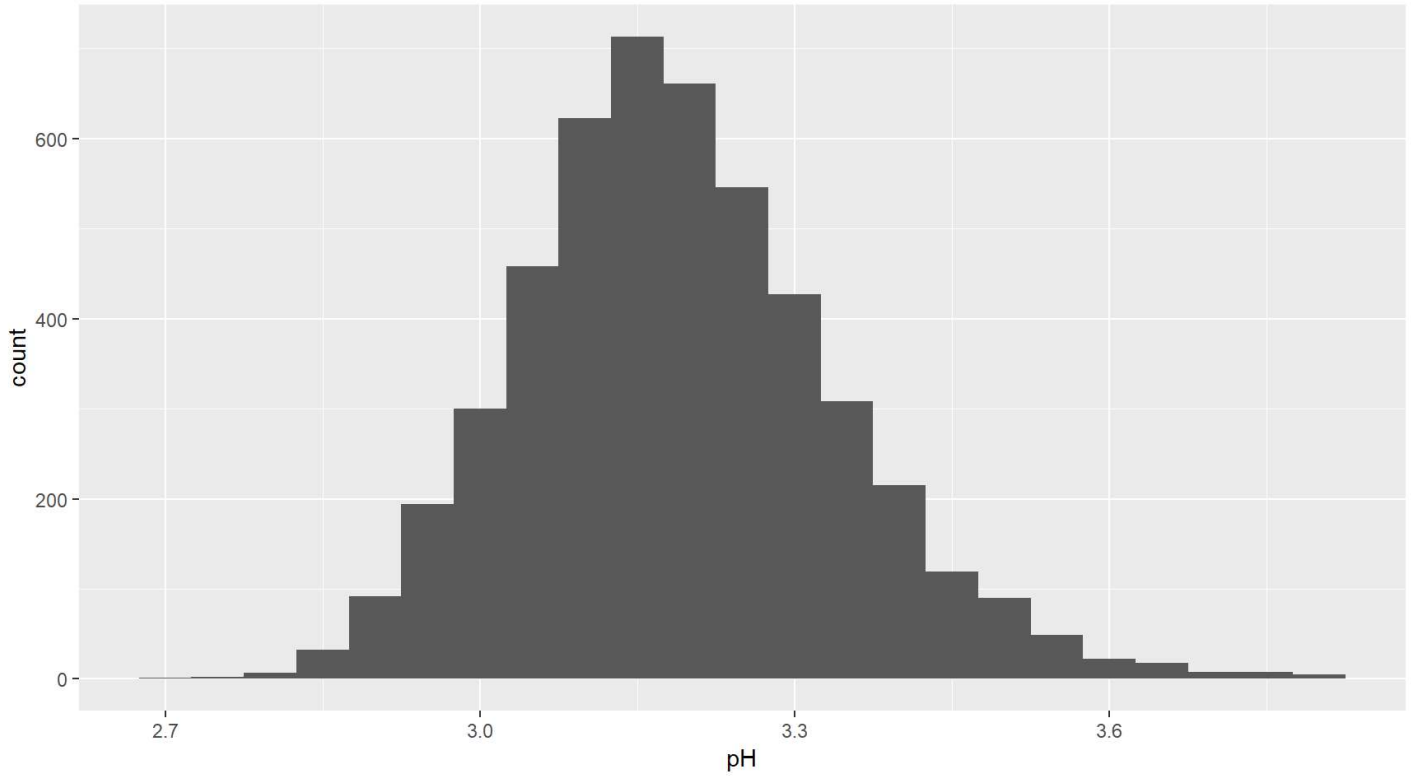


After zooming in we observe that the distribution is approx. normal, despite having uneven tail lengths.

Now, about the outliers of density.

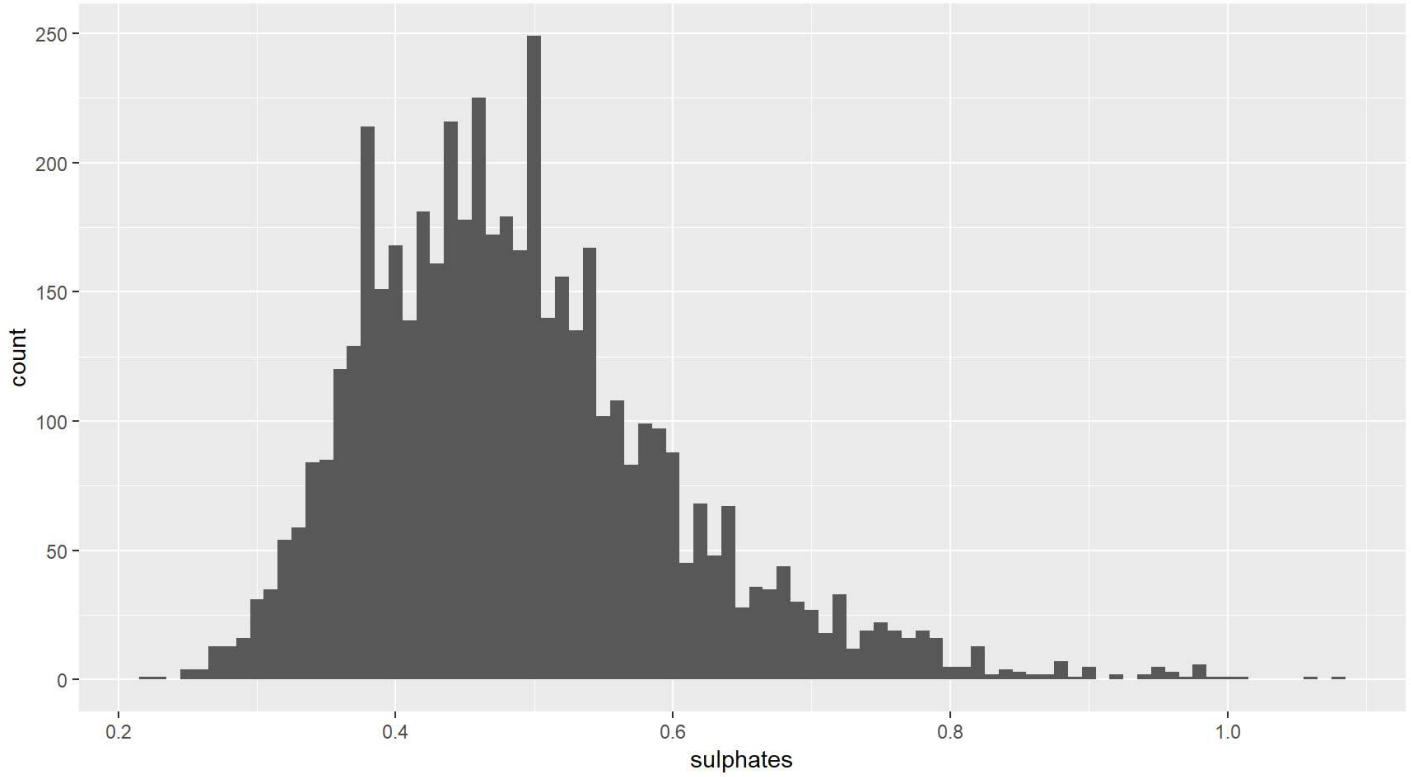
```
##      fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1654           7.9            0.330        0.28           31.6     0.053
## 1664           7.9            0.330        0.28           31.6     0.053
## 2782           7.8            0.965        0.60           65.8     0.074
##      free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates
## 1654                  35                  176 1.01030 3.15      0.38
## 1664                  35                  176 1.01030 3.15      0.38
## 2782                   8                  160 1.03898 3.39      0.69
##      alcohol quality bound.sulfur.dioxide
## 1654     8.8       6                  141
## 1664     8.8       6                  141
## 2782    11.7       6                  152
```

All 3 of them are of normal quality. I guess having higher densities than usual alone didn't affect negatively to their quality too much.

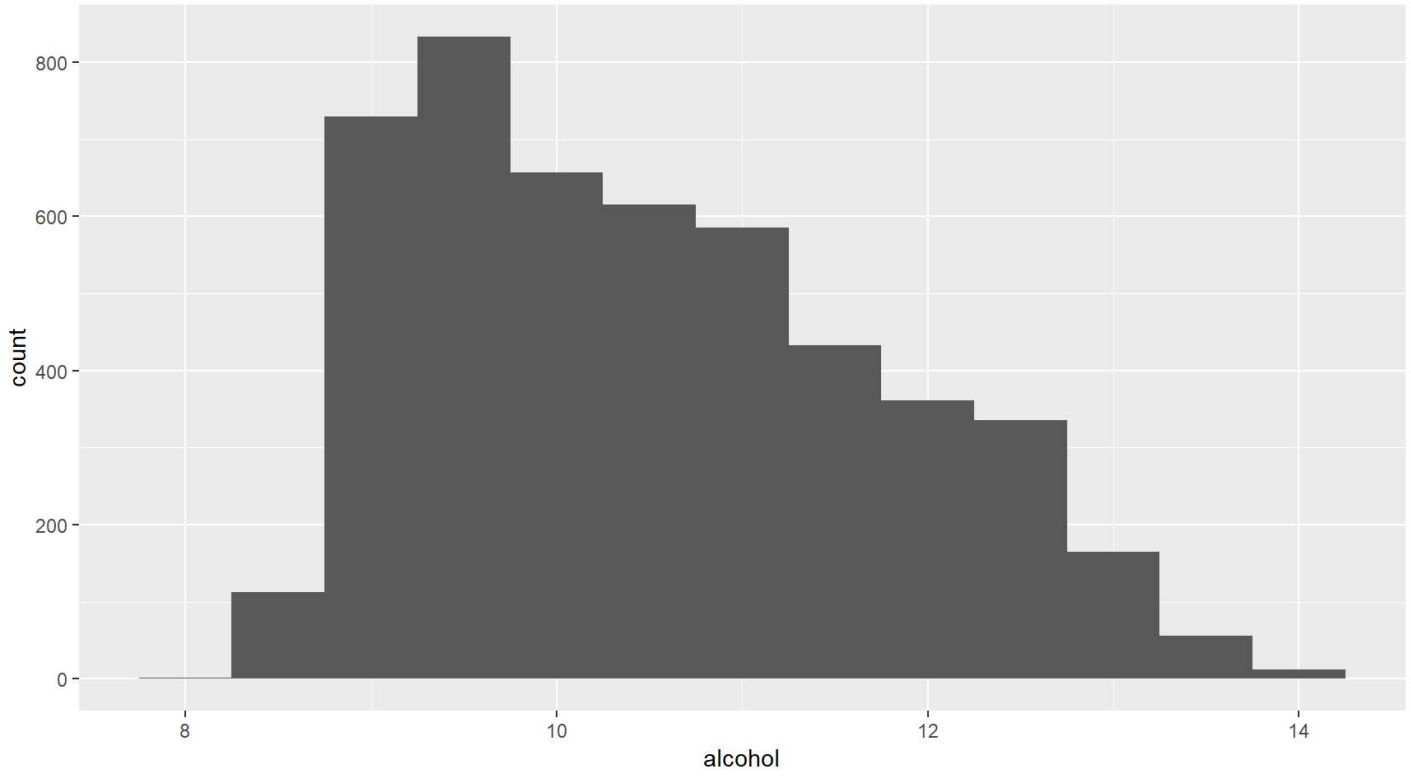



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.720   3.090   3.180   3.188   3.280   3.820
```

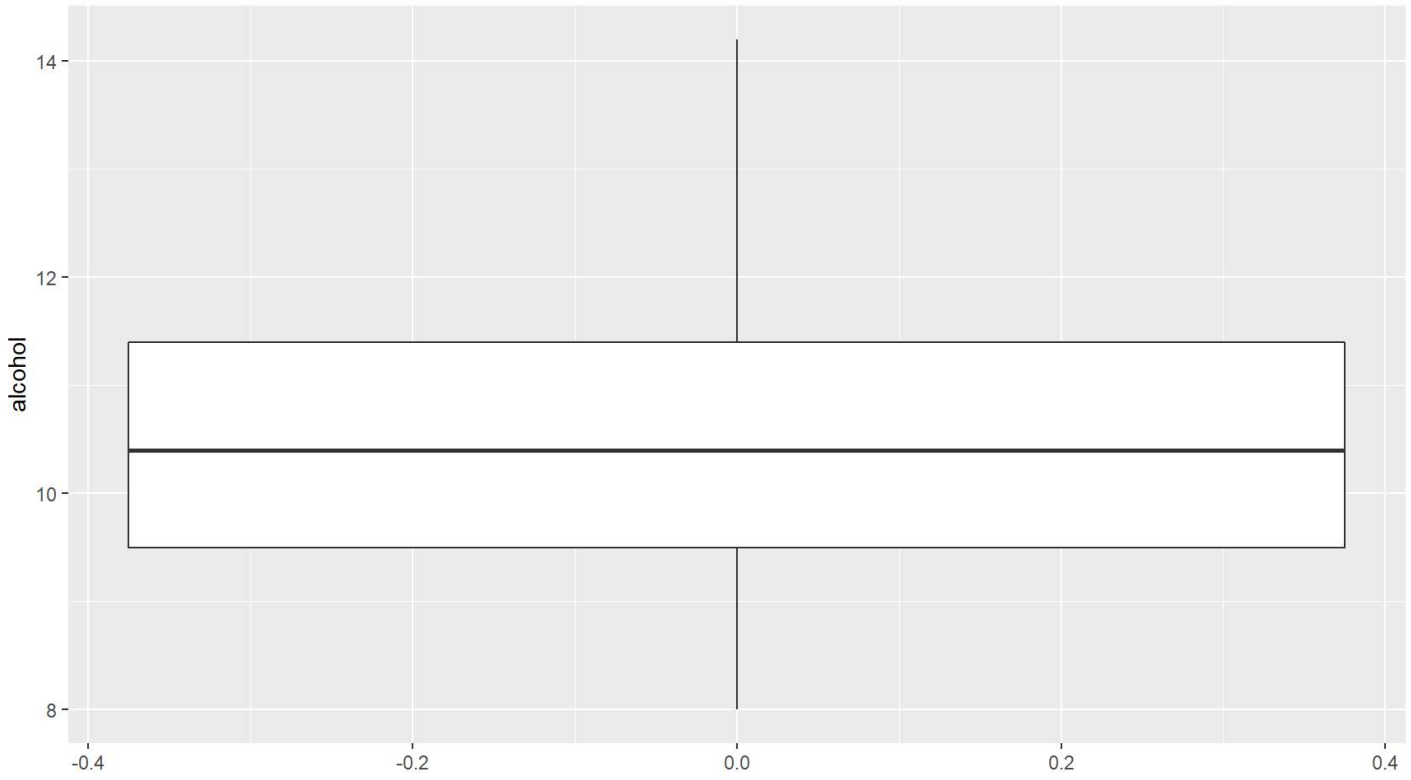
The pH looks normally distributed about the mean of 3.188.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2200  0.4100  0.4700  0.4898  0.5500  1.0800
```




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00    9.50   10.40   10.51   11.40   14.20
```



The alcohol distribution looks approximately normal with a slightly long right tail. It has a mean of 10.51.

Univariate Analysis

What is the structure of your dataset?

There are 4898 white wines in the dataset with 15 features(x, fixed.acidity, volatile.acidity, citric.acid, residual.sugar, chlorides, free.sulfur.dioxide, total.sulfur.dioxide, density, pH, sulphates, alcohol, quality). x is just the row number, and all the other variables are quantitative variables. We must also remember that the lower the pH the more acidic.

Other observations:

- We have 20 poor wines(1%) and 180 excellent wines(9.5%) against 1698 normal wines(89.5%). Quality in this dataset ranges from 3 to 9.
- After applying a log transformation to residual.sugar, its distribution appears bimodal. This probably means there are 2 groups of wine in our data set. One of the groups has more sugar and one of the groups has low sugar.
- The pH values is distributed in a small range between 2.7 to 3.8 but the majority of the values lie in an even smaller range of about 3.0 - 3.3.
- Based on the guide lines, only one white wine in this data set is sweet (>45 g/liter)

What is/are the main feature(s) of interest in your dataset?

For now, the main features will have to be quality and alcohol. I have a feeling alcohol is proportional to quality. I would like to figure out the features that are required to predict the quality of the white wine. residual.sugar also seems very interesting because of its bimodal logarithmic nature.

What other features in the dataset do you think will help support your investigation into your feature(s) of interest?

For now I can’t rule out any of the features that might help me predict quality of the wine, but I would like to pay extra attention to sulfur dioxide contents, pH, chlorides and residual sugar.

Did you create any new variables from existing variables in the dataset?

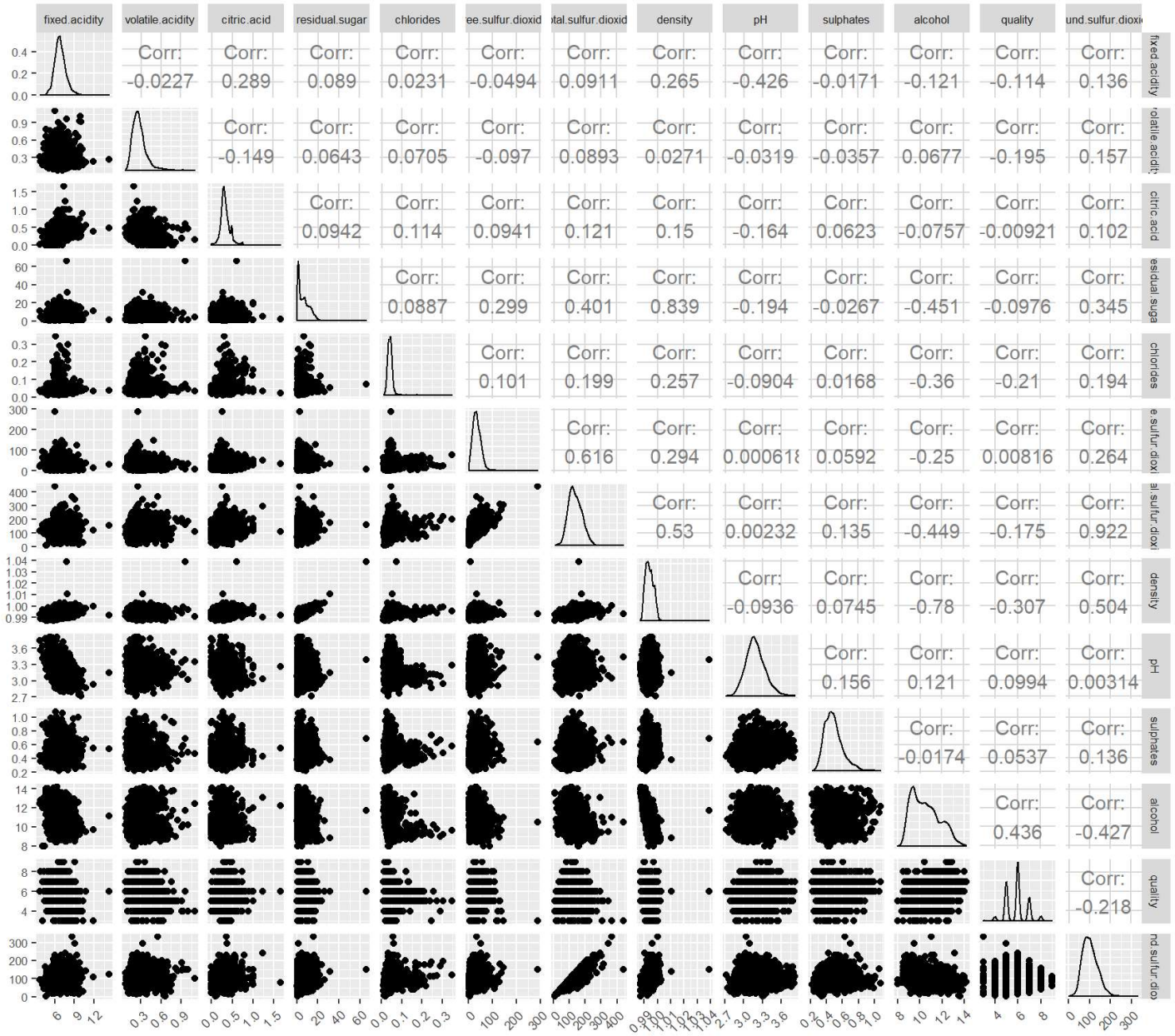
I created a numerical variable, bound.sulfur.dioxide. This will help me track the remaining sulfur dioxide in the wine besides free.sulfur.dioxide.

I also plan to create a new variable quality.ord, which would be quality but as an ordinal variable, ranking 9 as the highest quality down to 3 as the lowest quality. (since that is the range of quality in our dataset)

Of the features you investigated, were there any unusual distributions?

residual.sugar had a very long right-tail so I did a log10 transformation on it. This resulted in a bimodal graph which could signify two clear classes of wine subtypes in the dataset based on sugar content.

Bivariate Plots Section



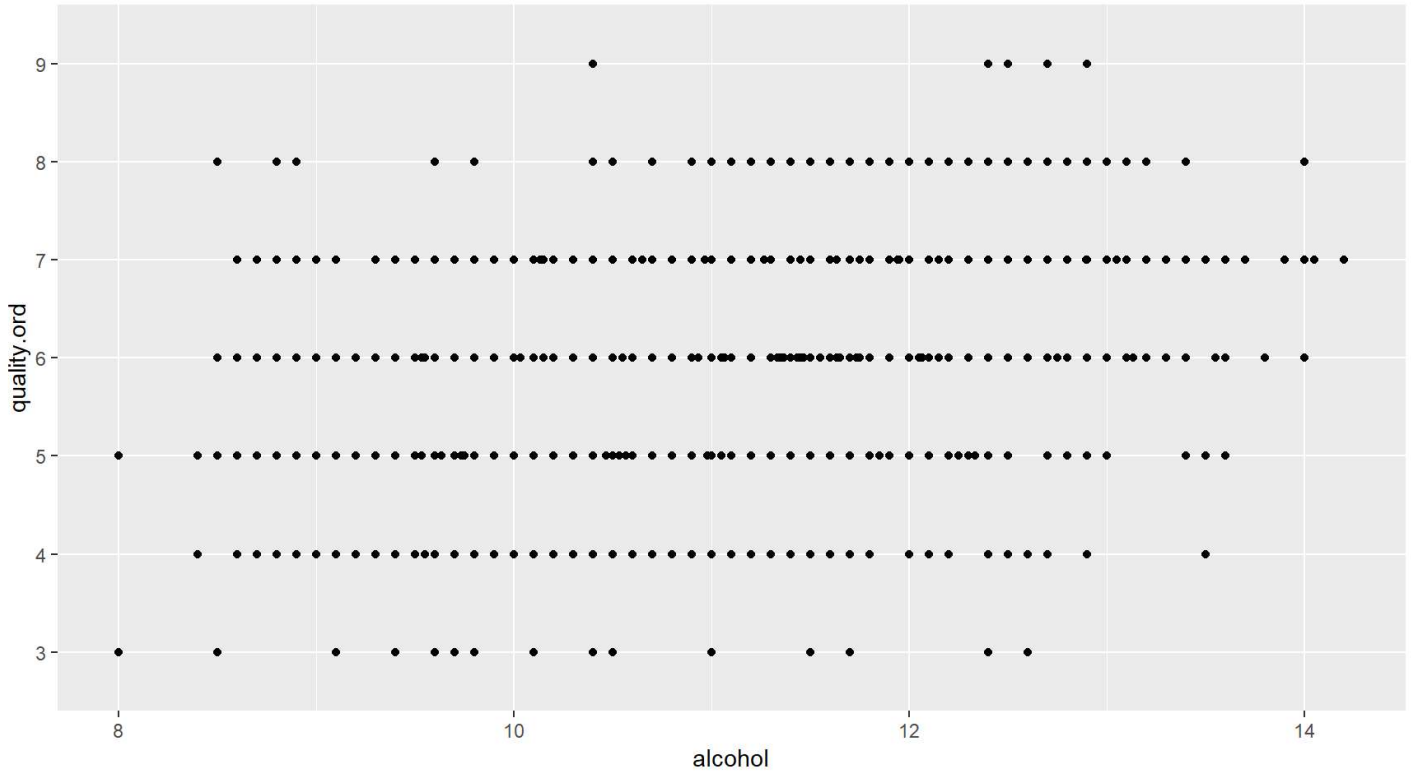
A few things to notice from here are the correlation coefficients. They give us a good indication of which bivariate relationships to focus on. Something interesting to notice is that there are only 2 meaningful correlations with “quality”:

- Quality & Alcohol : 0.436
- Quality & Density : -0.307

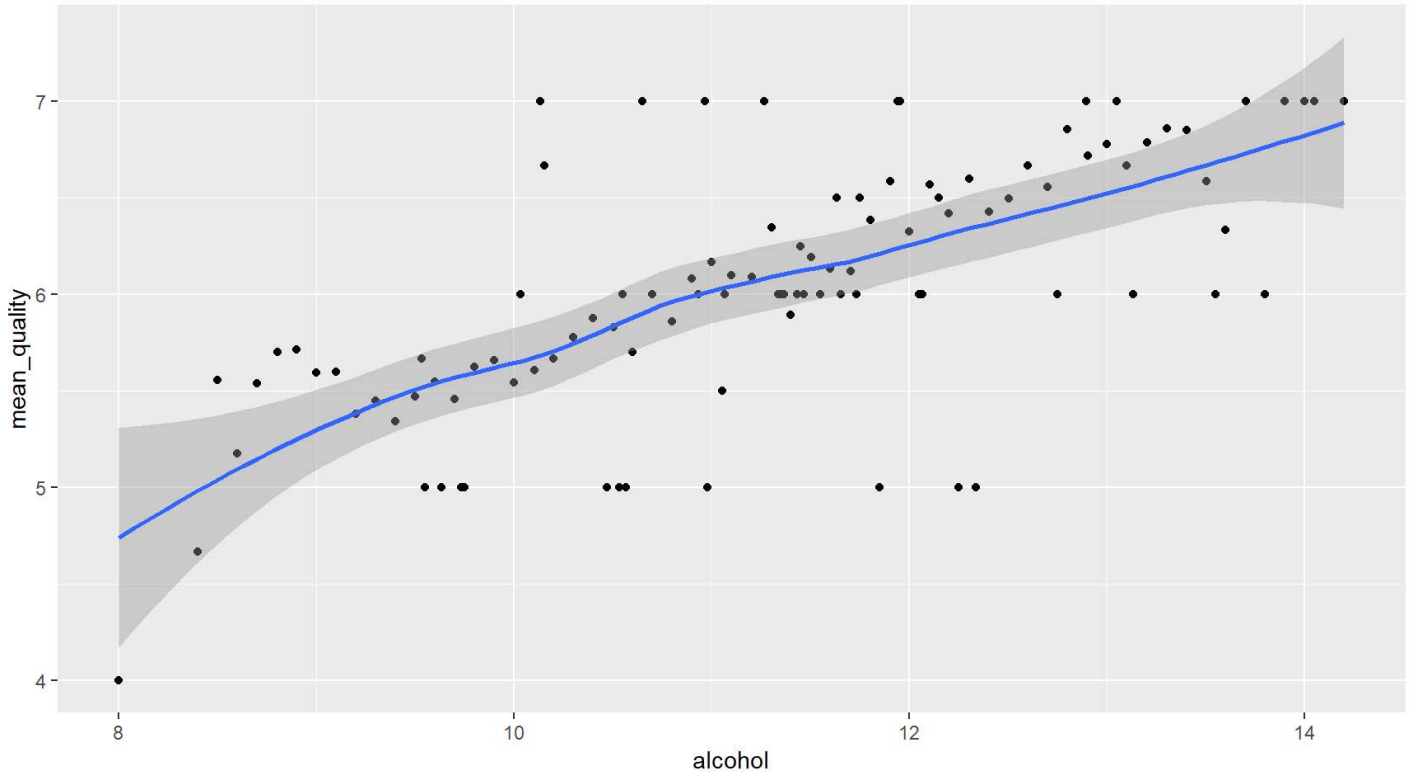
Neither of these can be considered as strong correlations, the one with Alcohol can be considered moderate as most.

Main relations

‘quality.ord’ contains quality as an ordinal variable.

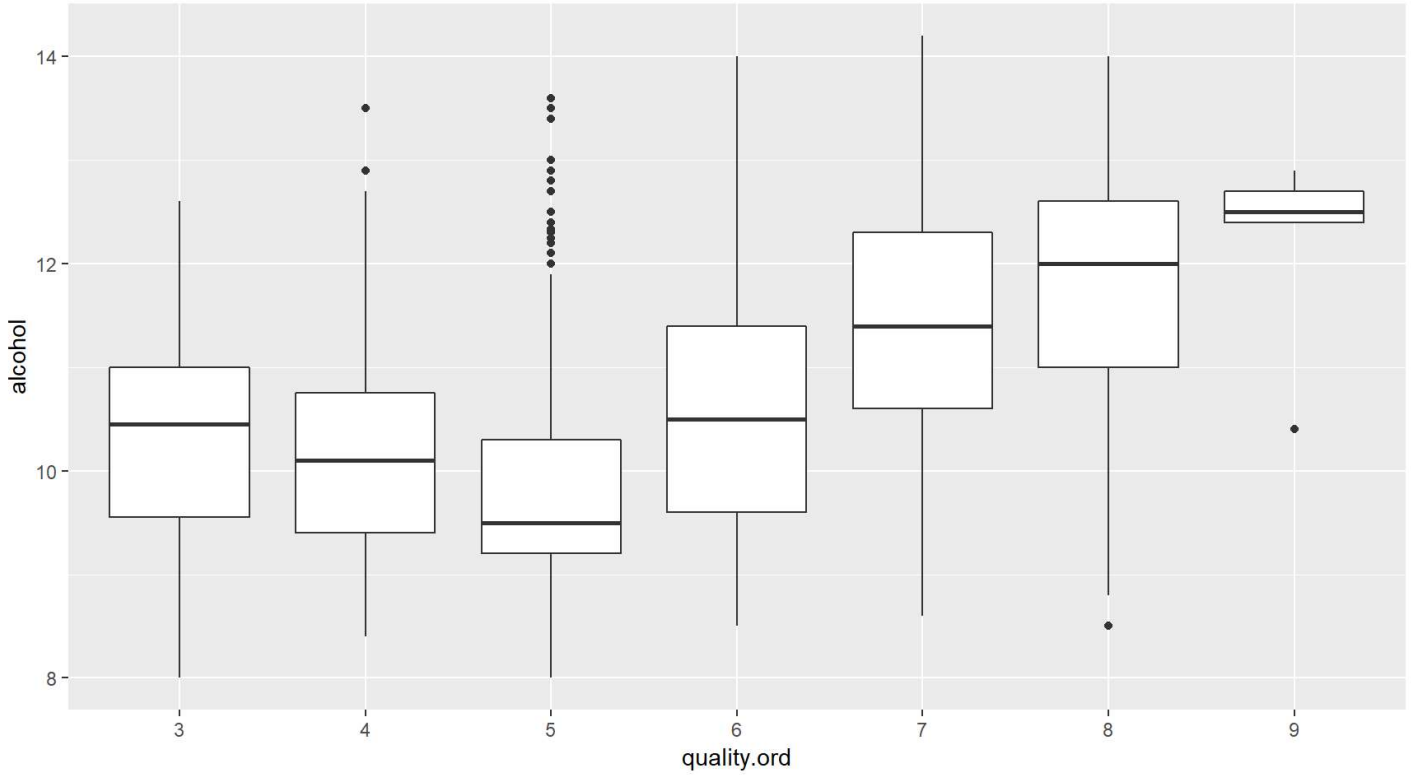


Since, quantity is a discrete value its hard for me to see a proper trend . There does seem to be a lot more points of alcohol around 11.5 the closer we get to a 6 quality but that isn't substantial information



```
##
##  Pearson's product-moment correlation
##
## data:  df$alcohol and df$quality
## t = 33.858, df = 4896, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4126015 0.4579941
## sample estimates:
##       cor
## 0.4355747
```

This seems like a better view of the data. Here we observe a clear positive linear trend between alcohol and mean_quality with a few outliers.

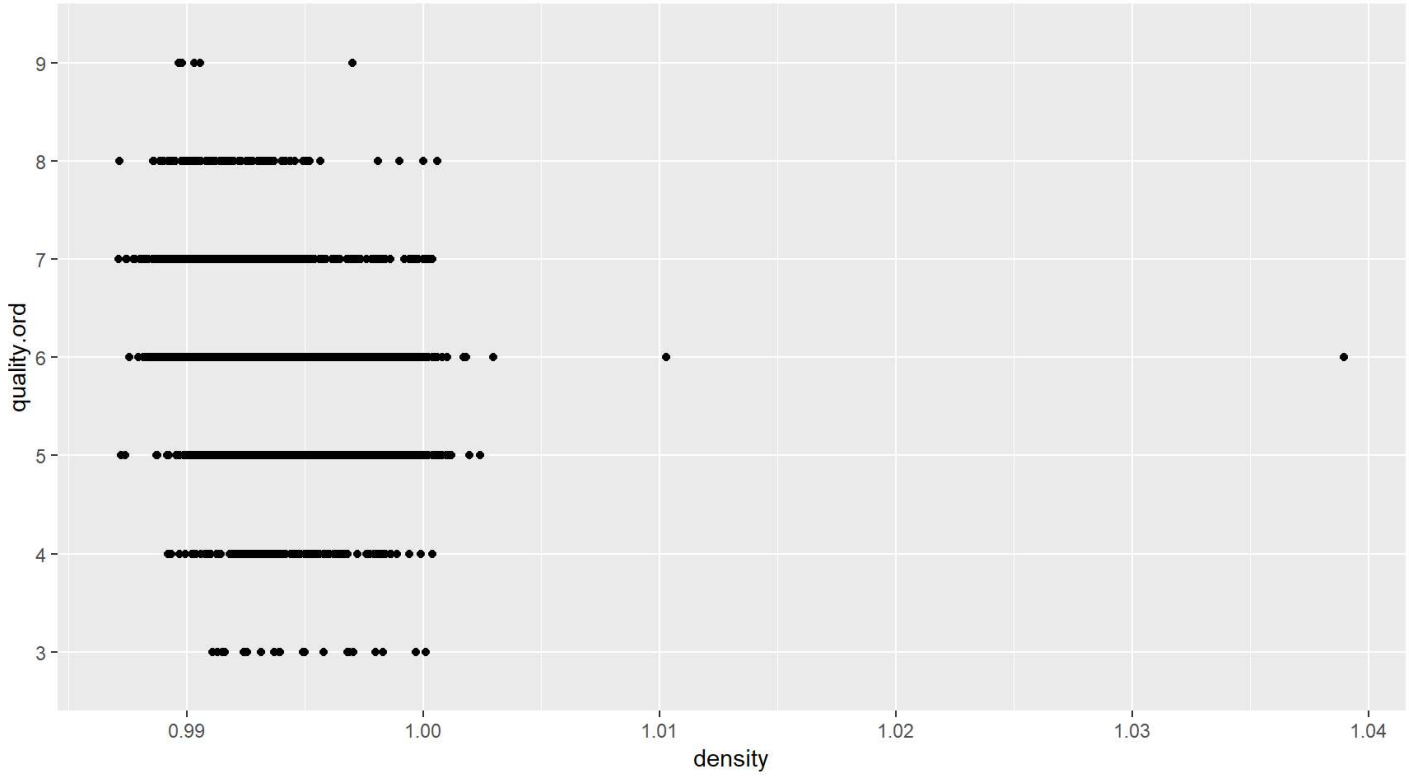


We can also see from the boxplot that the median alcohol keeps increasing as the quality increases from 5 to 9. This alongside the correlation connection between alcohol and quality.

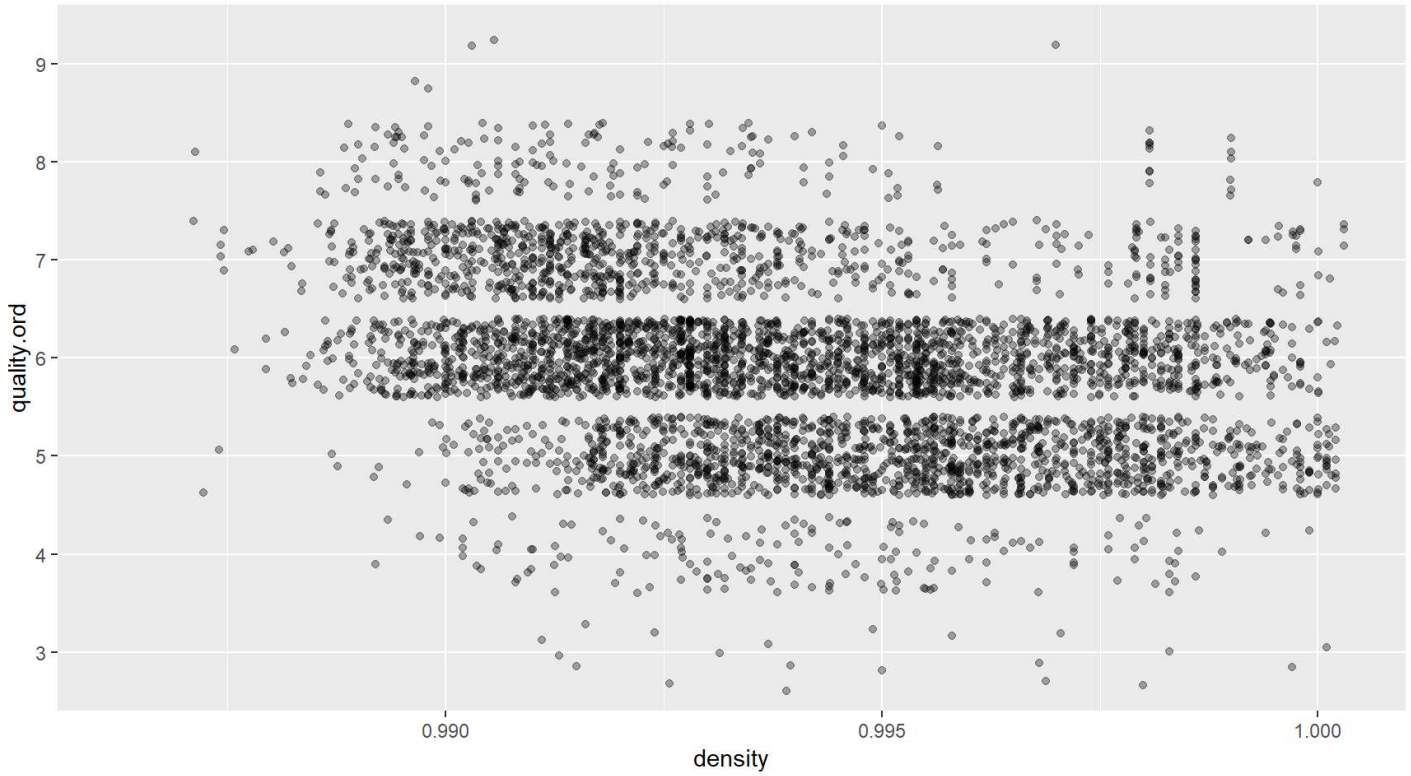
```
##
## Call:
## lm(formula = quality ~ alcohol, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5317 -0.5286  0.0012  0.4996  3.1579
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.582009   0.098008   26.34   <2e-16 ***
## alcohol     0.313469   0.009258   33.86   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7973 on 4896 degrees of freedom
## Multiple R-squared:  0.1897, Adjusted R-squared:  0.1896
## F-statistic:  1146 on 1 and 4896 DF,  p-value: < 2.2e-16
```

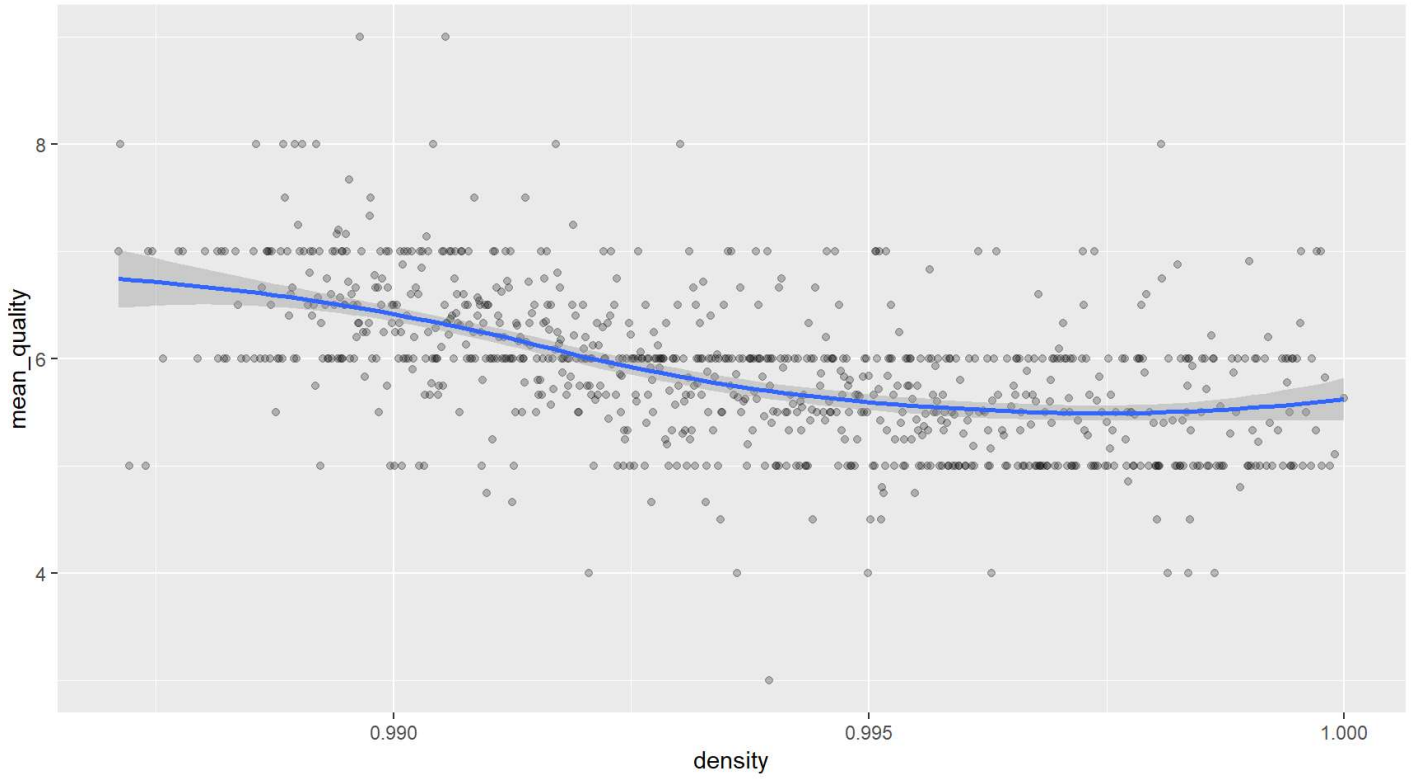
Goign by the R-squared value, alcohol only explains 18.97% of the change in quality. This clearly shows that despite there being a trend between them it isn't a very strong one.

Now let's look at the connection between quality and density:



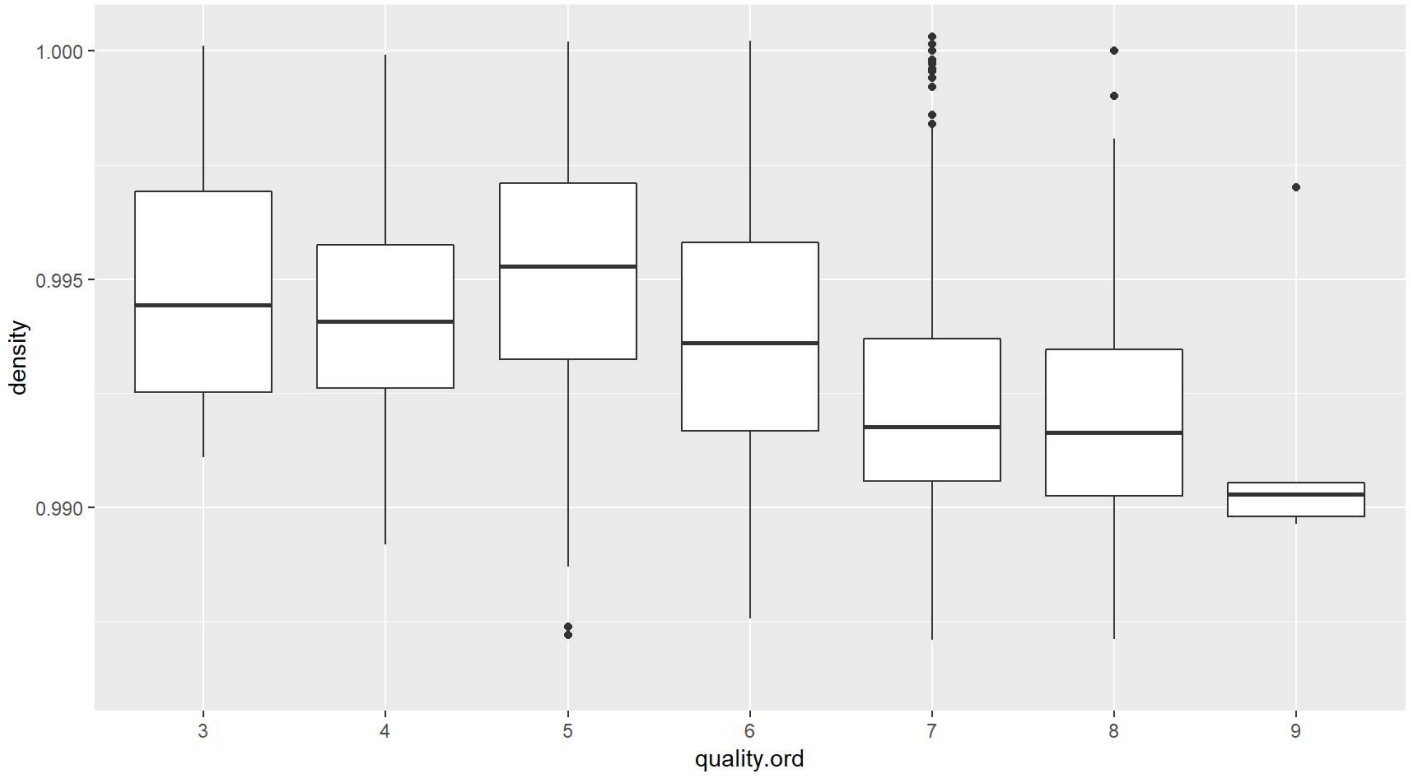
Let's remove the outliers by limiting the x-axis to remove the top 1% of data and add some jitter.


quality.ord
9
8
7
6
5
4
3
0.990
0.995
1.000
density


mean_quality
8
6
4
0.990
0.995
1.000
density

```
##
##  Pearson's product-moment correlation
##
## data:  df$density and df$quality
## t = -22.581, df = 4896, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3322718 -0.2815385
## sample estimates:
##       cor
## -0.3071233
```

We can clearly see in the above plots that their is a negative trend between density and quality.



Here we observe clearly that the median density reduces as the quality increases from 5 to 9.

```
##
## Call:
## lm(formula = quality ~ density, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1441 -0.6258  0.0005  0.5162  4.2102
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   96.277      4.003   24.05   <2e-16 ***
## density      -90.942      4.027  -22.58   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8429 on 4896 degrees of freedom
## Multiple R-squared:  0.09432,    Adjusted R-squared:  0.09414
## F-statistic: 509.9 on 1 and 4896 DF,  p-value: < 2.2e-16
```

Despite showing presence of a clear trend density only accounts for 9.432% of the quality which is weaker than the value we got for alcohol.

Considering these were our most correlated variables with quality, it might be difficult to make a simple linear model with the variables in this dataset.

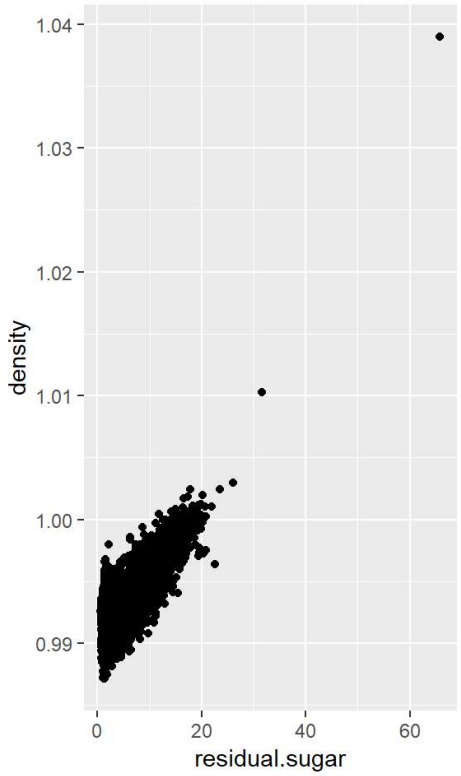
Other relations

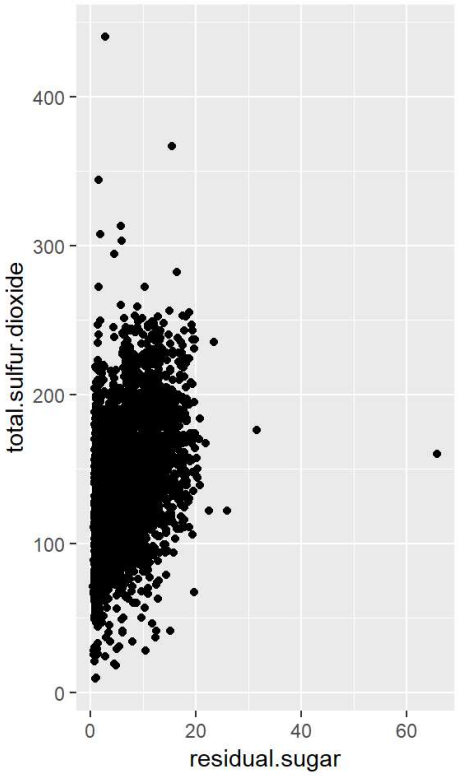
There are also some other correlations interrelated between our features that DO NOTE include our main feature of interest(quality). These are:

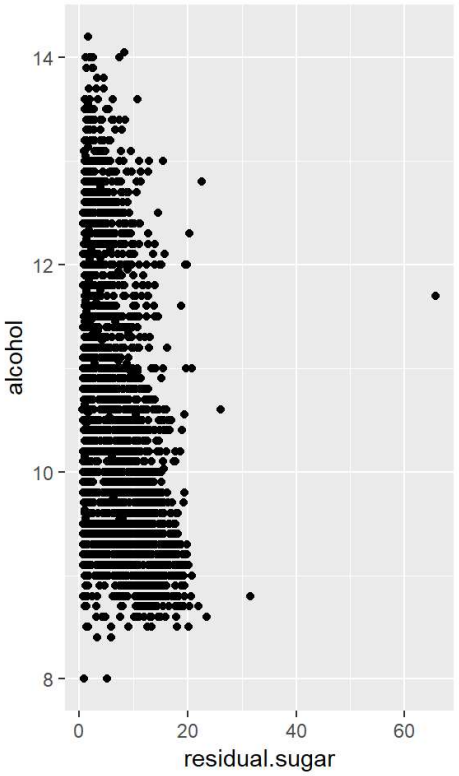
- residual.sugar & density: 0.839
- residual.sugar & total.sulfur.dioxide. : 0.401
- residual.sugar & alcohol : -0.451

- total.sulfur.dioxide & density : 0.53
- total.sulfur.dioxide & alcohol : -0.449
- total.sulfur.dioxide & bound.sulfur.dioxide : 0.922
- total.sulfur.dioxide & free.sulfur.dioxide : 0.616

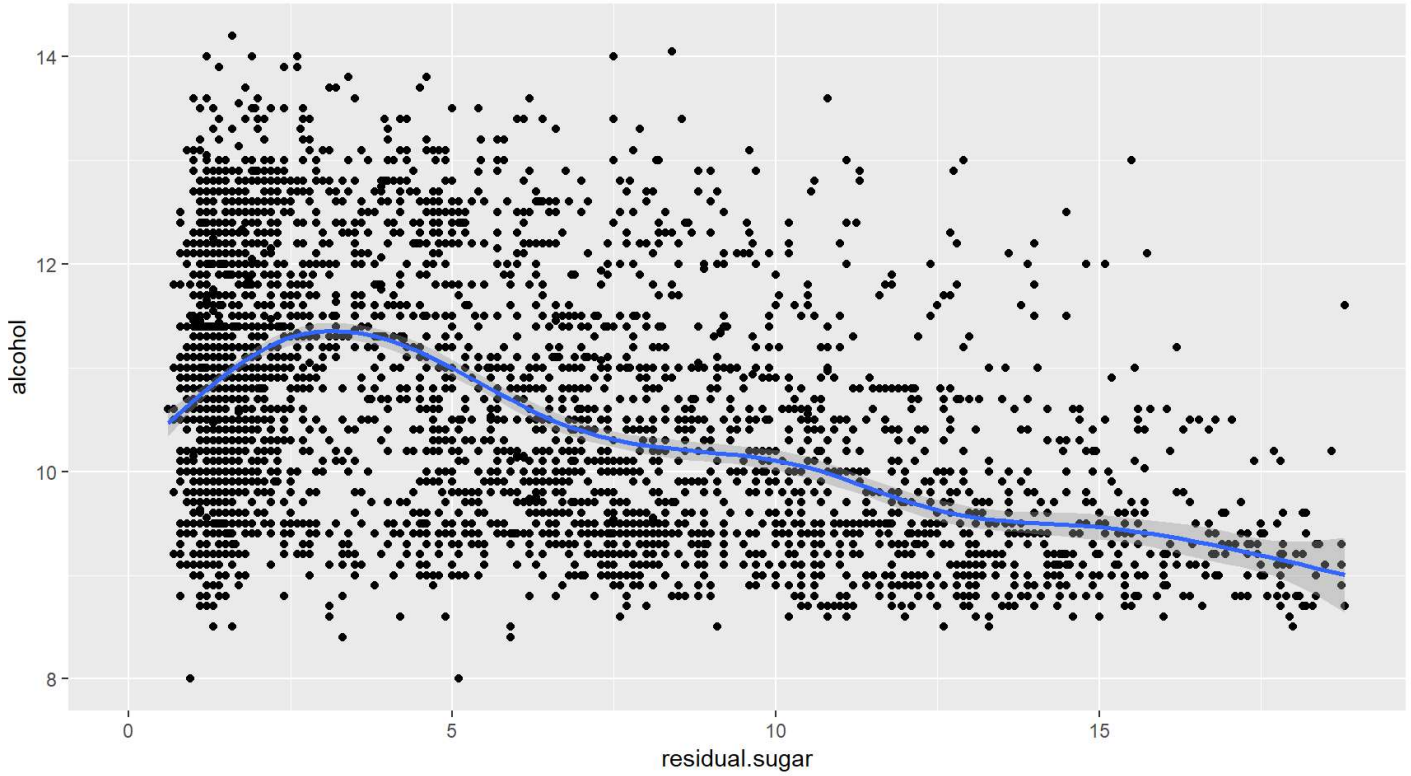
- density & alcohol : -0.307



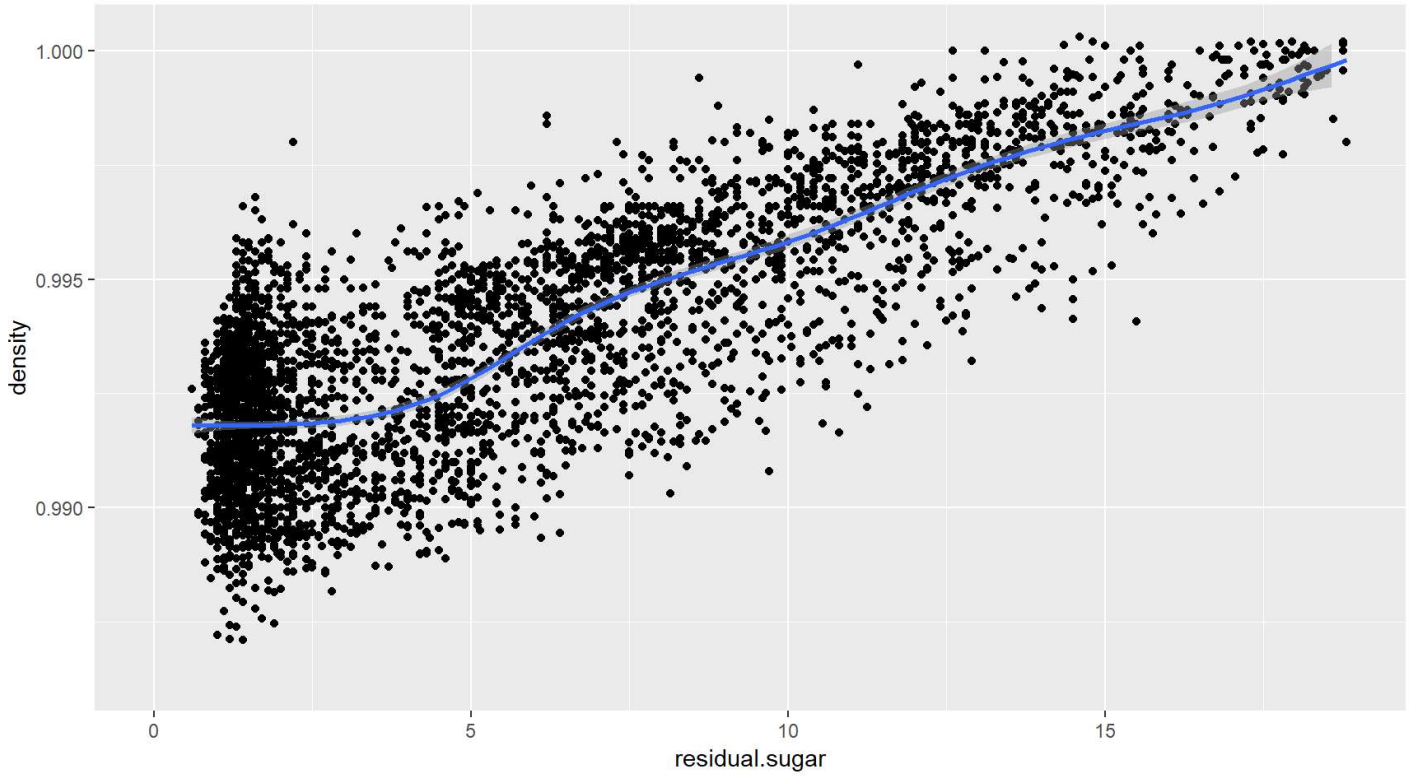




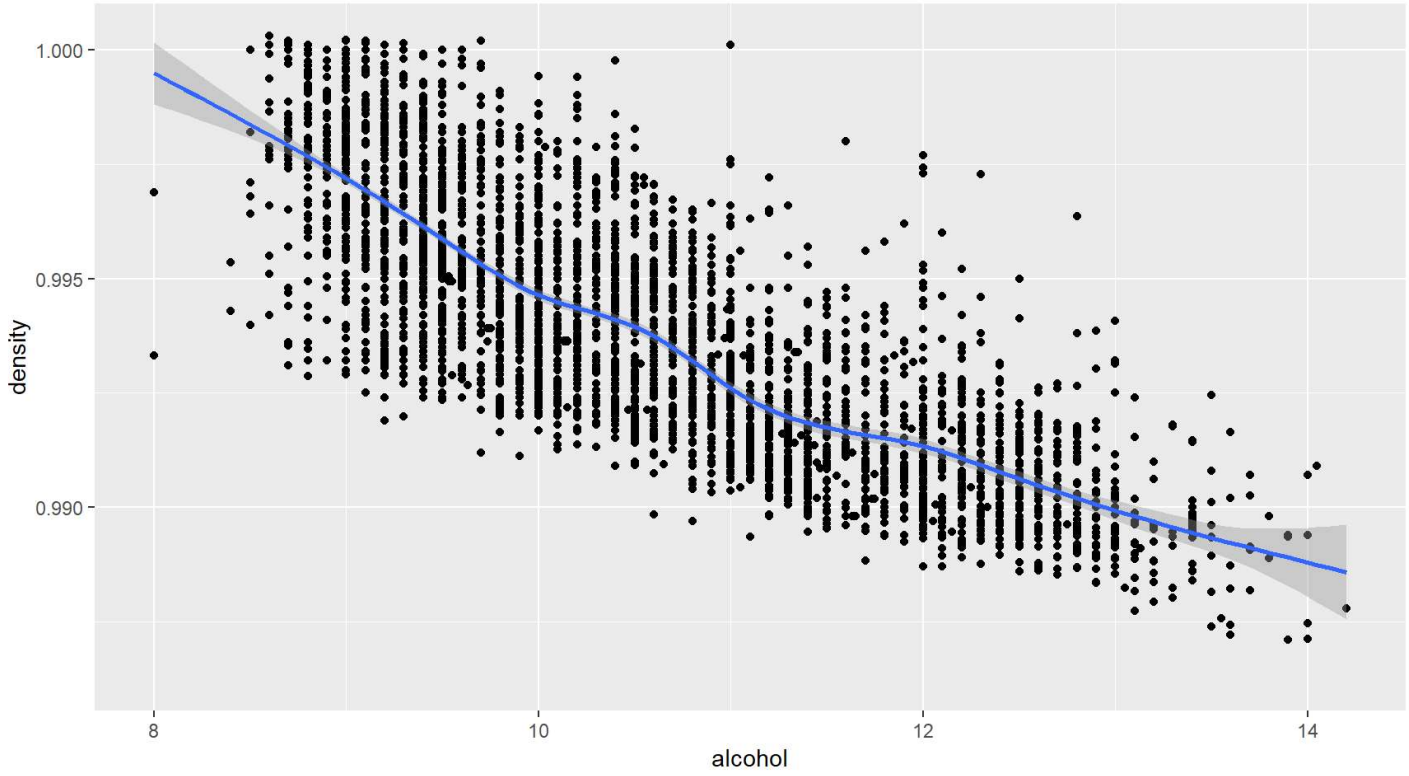
Let's remove the outliers.



The connection between residual.sugar and alcohol doesn't feel logical directly, maybe there is a hidden variable? Maybe density!



This one inherently makes sense, the more sugar in the wine, the more dense it will become!

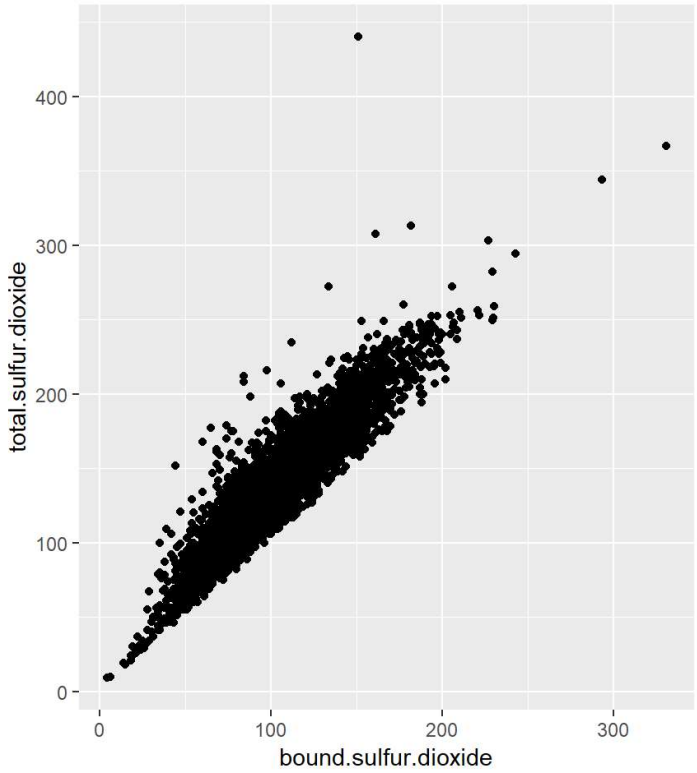


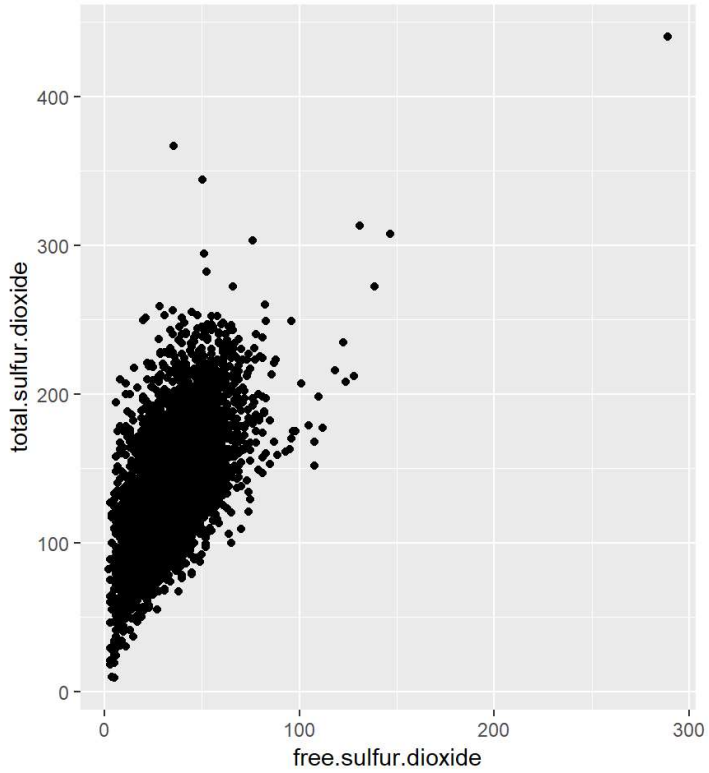
A clear negative linear trend between alcohol and density shows when alcohol increases density of the wine decreases. Maybe density is the hidden variable!

alcohol and residual.sugar have to be kept at a delicate balance to maintina density of the wine.

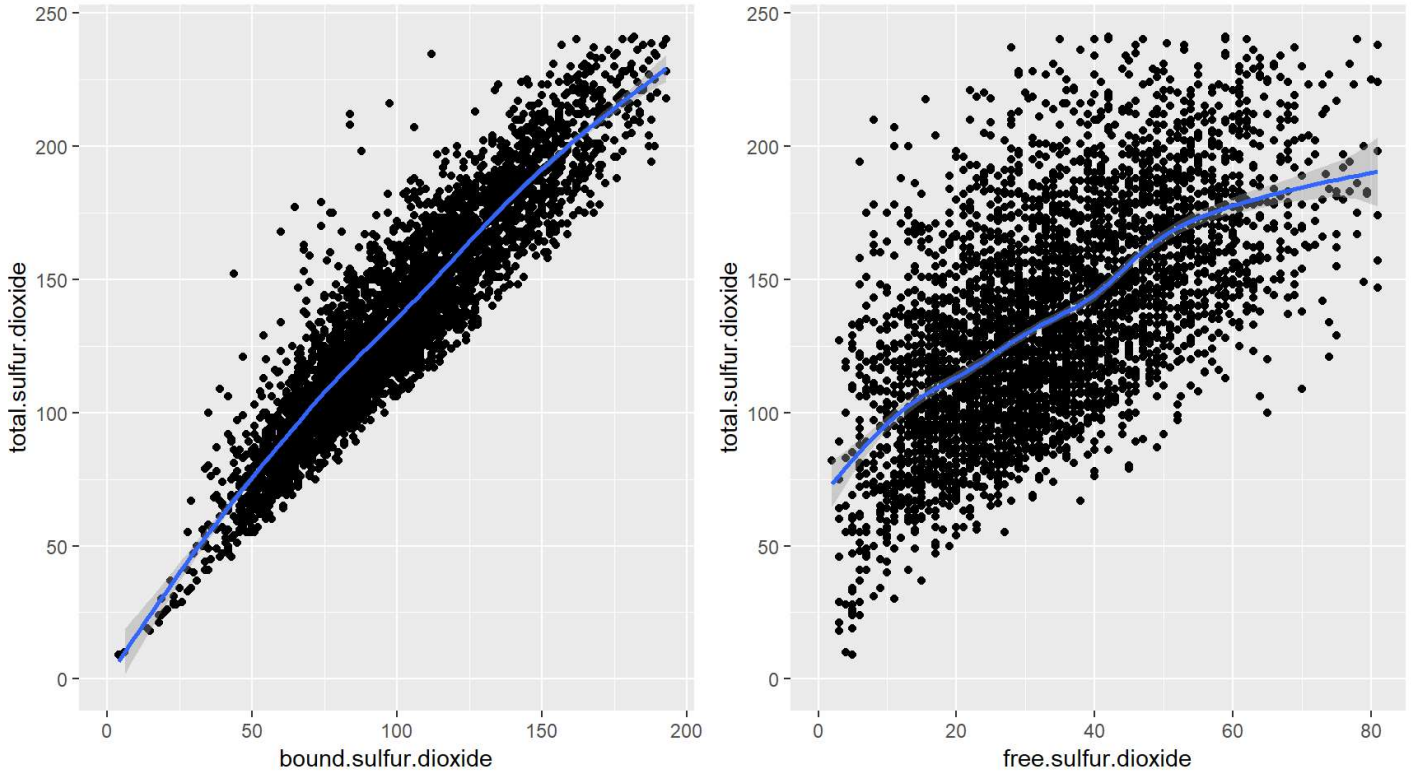
```
##
## Call:
## lm(formula = density ~ alcohol + residual.sugar, data = df)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0020196 -0.0005852 -0.0001403  0.0004674  0.0249805
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.005e+00  1.349e-04  7446.7   <2e-16 ***
## alcohol        -1.226e-03  1.189e-05  -103.2   <2e-16 ***
## residual.sugar  3.607e-04  2.884e-06   125.1   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0009137 on 4895 degrees of freedom
## Multiple R-squared:  0.9067, Adjusted R-squared:  0.9067
## F-statistic: 2.379e+04 on 2 and 4895 DF,  p-value: < 2.2e-16
```

Going by the R-squared value, alcohol and residual.sugar together account for 90.67% of the density value which works alongside our predictions.





Let us remove the outliers

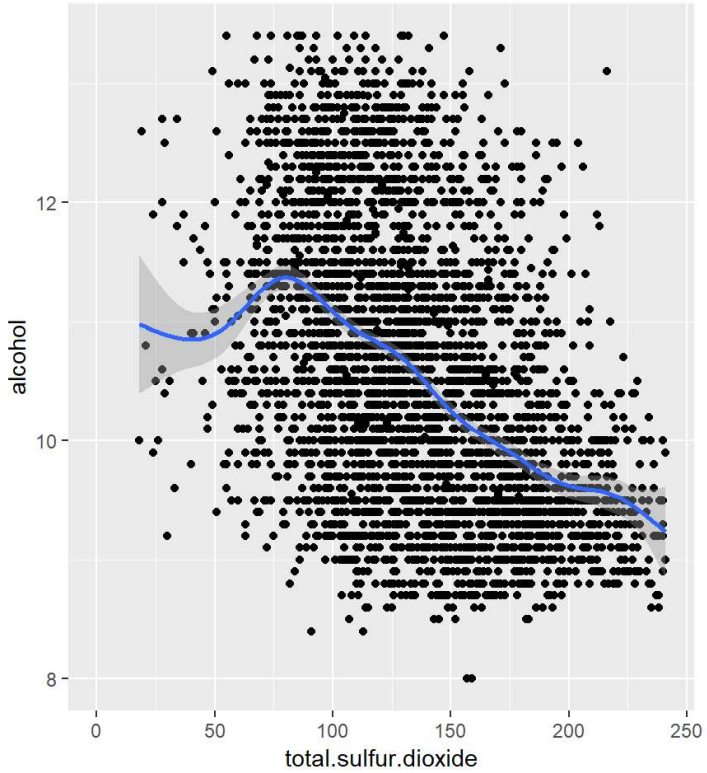


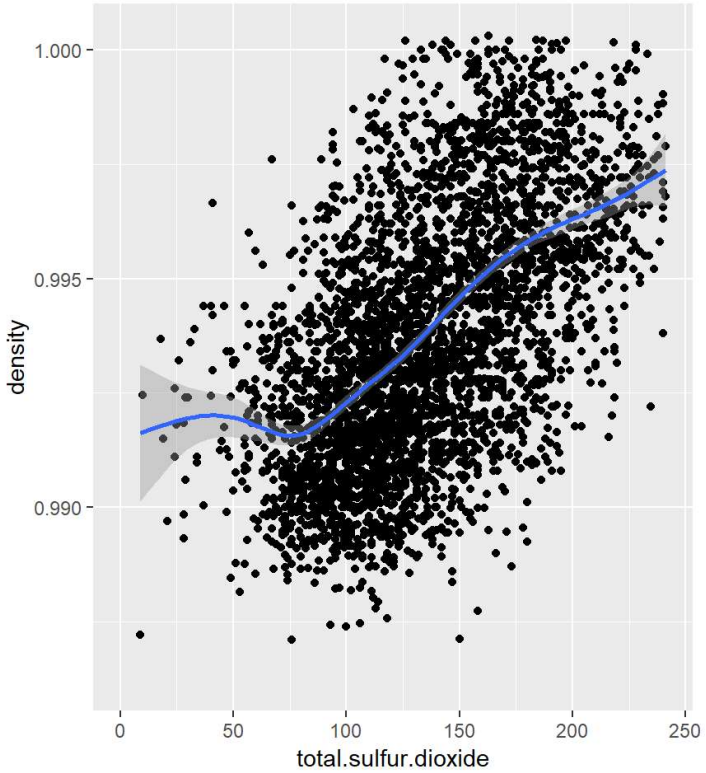
Linear trends of total.sulfur.dioxide is inherent since total.sulfur.dioxide is the sum of bound.sulfur.dioxide and free.sulfur.dioxide

```
##
## Call:
## lm(formula = total.sulfur.dioxide ~ free.sulfur.dioxide, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -98.163 -24.891  -3.273  21.407 227.844
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         84.05553    1.10304   76.20   <2e-16 ***
## free.sulfur.dioxide  1.53804    0.02815   54.65   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.5 on 4896 degrees of freedom
## Multiple R-squared:  0.3788, Adjusted R-squared:  0.3787
## F-statistic:  2986 on 1 and 4896 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = total.sulfur.dioxide ~ bound.sulfur.dioxide, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.273 -11.269  -1.528   9.700 247.503
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          22.005650   0.734316   29.97   <2e-16 ***
## bound.sulfur.dioxide  1.129084   0.006753  167.20   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.41 on 4896 degrees of freedom
## Multiple R-squared:  0.851,  Adjusted R-squared:  0.8509
## F-statistic: 2.796e+04 on 1 and 4896 DF,  p-value: < 2.2e-16
```

So going by the R-squared values, bound.sulfur.dioxide is the higher in proportion generally in any wine.





Maybe total.sulfur.dioxide and alcohol have an intertwinned relationship with density too?

```
##
## Call:
## lm(formula = density ~ alcohol + residual.sugar + free.sulfur.dioxide +
##     bound.sulfur.dioxide, data = df)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0020775 -0.0005564 -0.0001124  0.0004080  0.0247499
##
## Coefficients:
##                         Estimate Std. Error  t value Pr(>|t|)
## (Intercept)           1.003e+00  1.536e-04 6530.419  < 2e-16 ***
## alcohol              -1.151e-03  1.202e-05  -95.798  < 2e-16 ***
## residual.sugar        3.532e-04  2.850e-06  123.937  < 2e-16 ***
## free.sulfur.dioxide  -5.406e-06  7.821e-07   -6.912 5.39e-12 ***
## bound.sulfur.dioxide  8.942e-06  4.079e-07   21.925  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0008708 on 4893 degrees of freedom
## Multiple R-squared:  0.9153, Adjusted R-squared:  0.9152
## F-statistic: 1.322e+04 on 4 and 4893 DF,  p-value: < 2.2e-16
```

Well that proves it! Though a low increase in R-squared free.sulfur.dioxide and bound.sulfur.dioxide do contribute to density.

Bivariate Analysis

Talk about some of the relationships you observed in this part of the investigation. How did the feature(s) of interest vary with other features in the dataset?

I focussed on the variables with meaningful correlations. quality seems to be loosely related with alcohol and density.

For quality from 3 to 5, as alcohol reduces quality increases. But, for quality from 6-9, as alcohol increases quality increases.

For quality from 3 to 5, as density increased quality increases. But, for quality from 6-9, as density decreases quality increases.

The linear model for these 3 variables was not very relevant with its low R-squared value.

Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

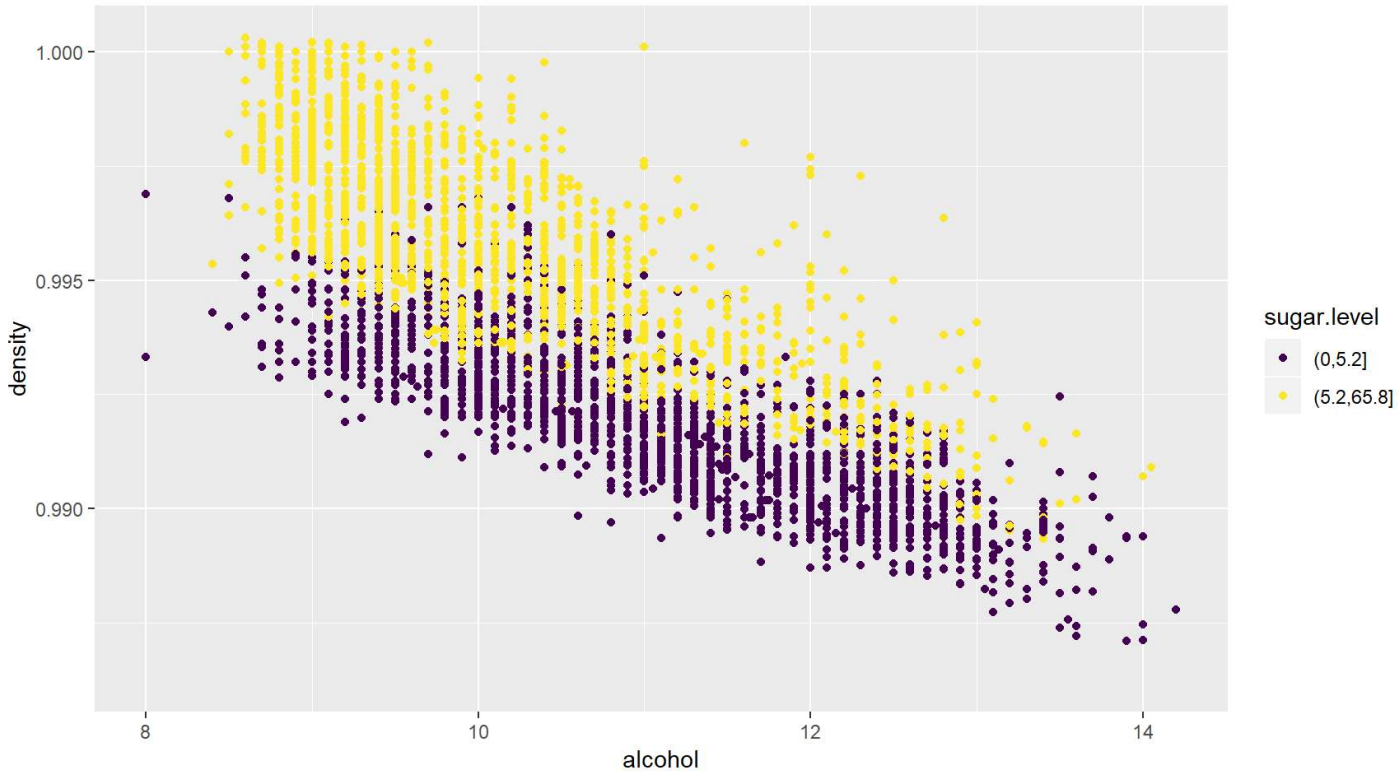
There is a relationship between alcohol, residual.sugar, density. As alcohol increases, density decreases. On the other hand as residual.sugar increases, density increases. I think a certain relationship between alcohol and residual.sugar is required to maintain a certain level of density. free.sulfur.dioixide and total.sulfur.dioxide are also related to related to density but at a much lower extent.

What was the strongest relationship you found?

The strongest relationship I found was the relationship between density & residual.sugar(correlation = 0.839) and density & alcohol(correlation = -0.78).

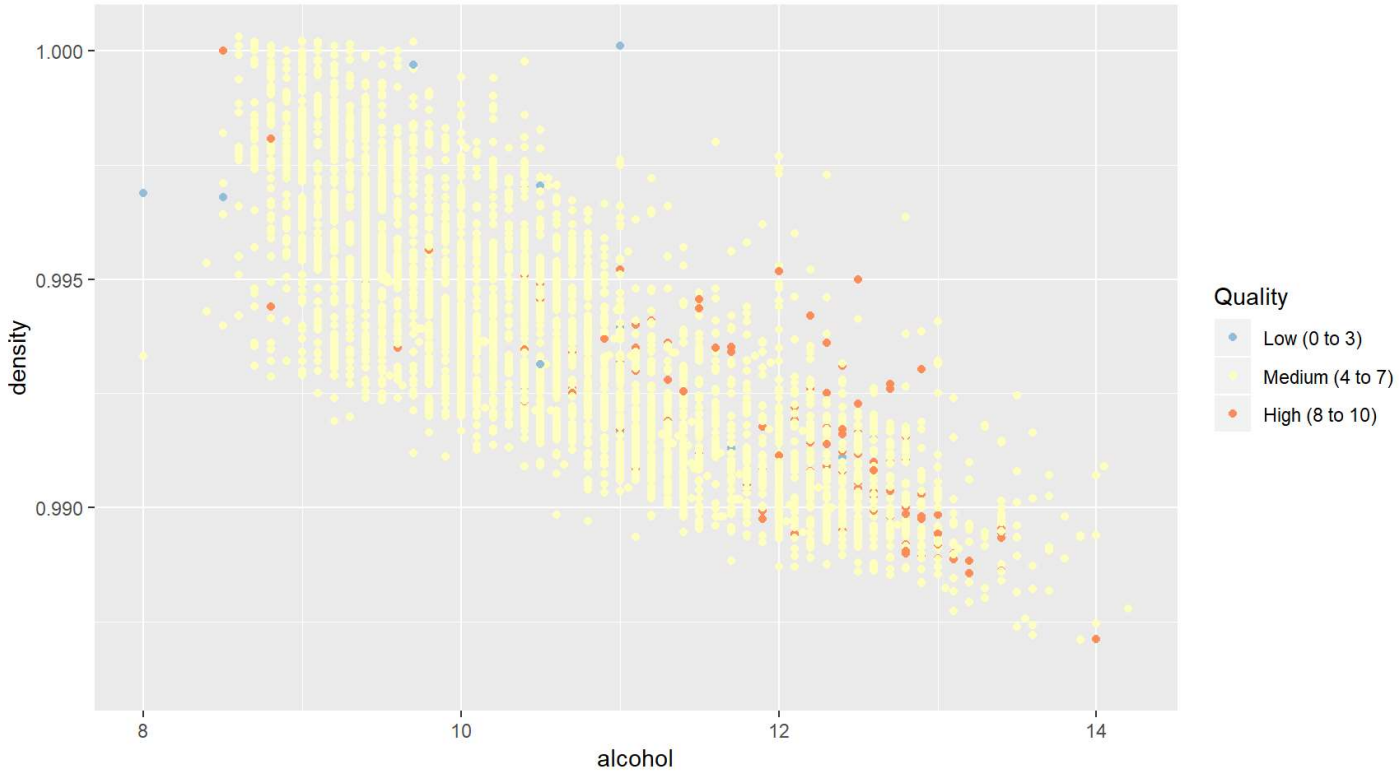
Multivariate Plots Section

I want to split residual.sugar into 2 groups due to its bimodal nature. I will split it by it's median



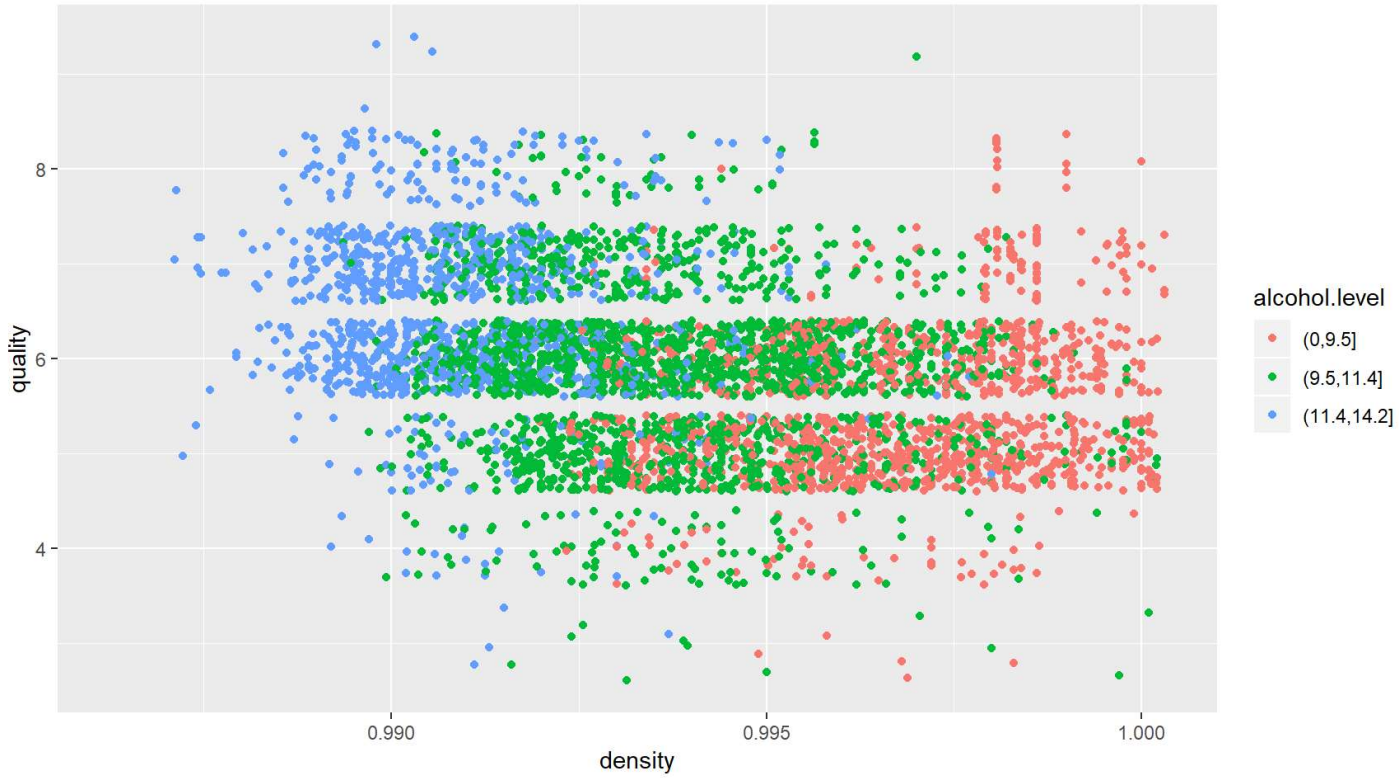
Here we observe a noticeable interpretation of our Bivariate analysis. As alcohol increases, density and sugar level decreases. As alcohol decreases, density and sugar level increases.

But, does this have an effect on quality?

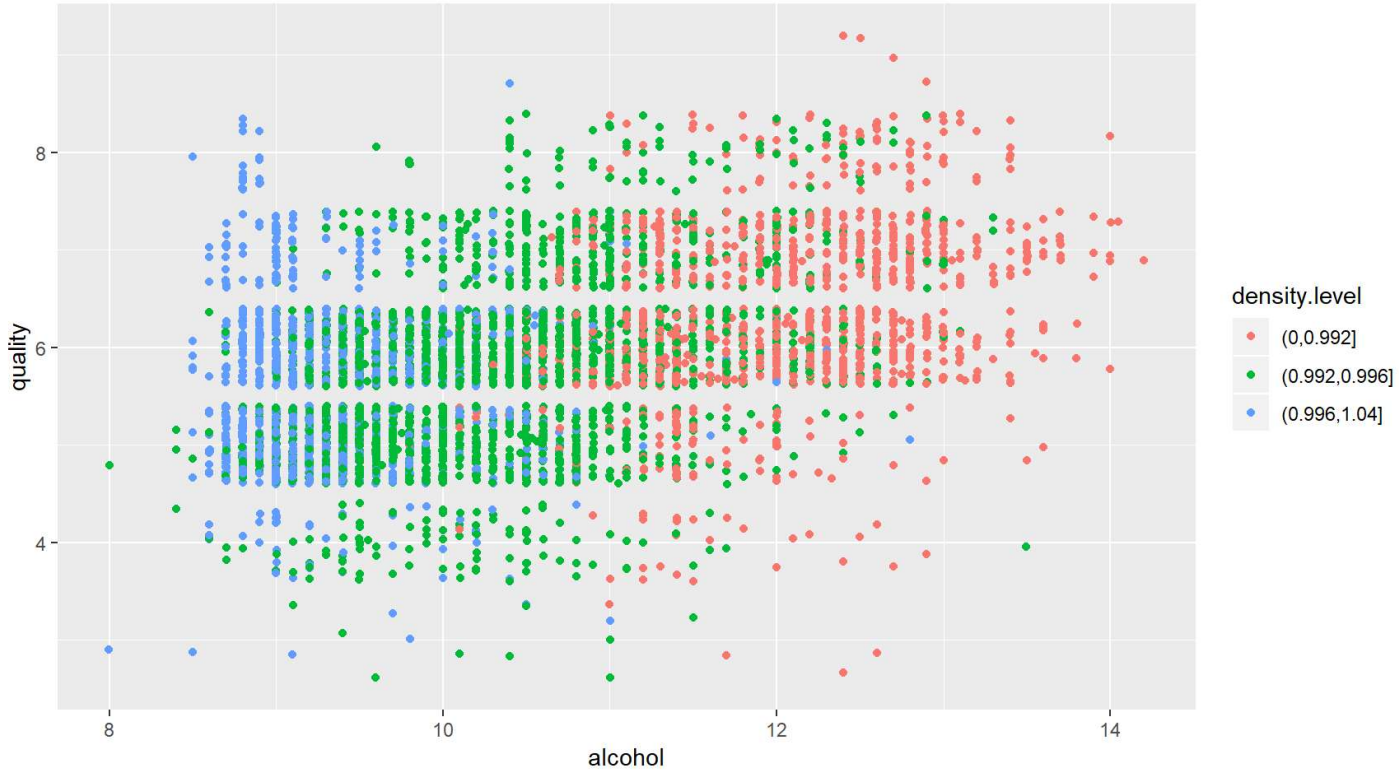


Since majority of our dataset is normal wines, this plot doesn't give us much information. Maybe I should divide my alcohol variable into groups?

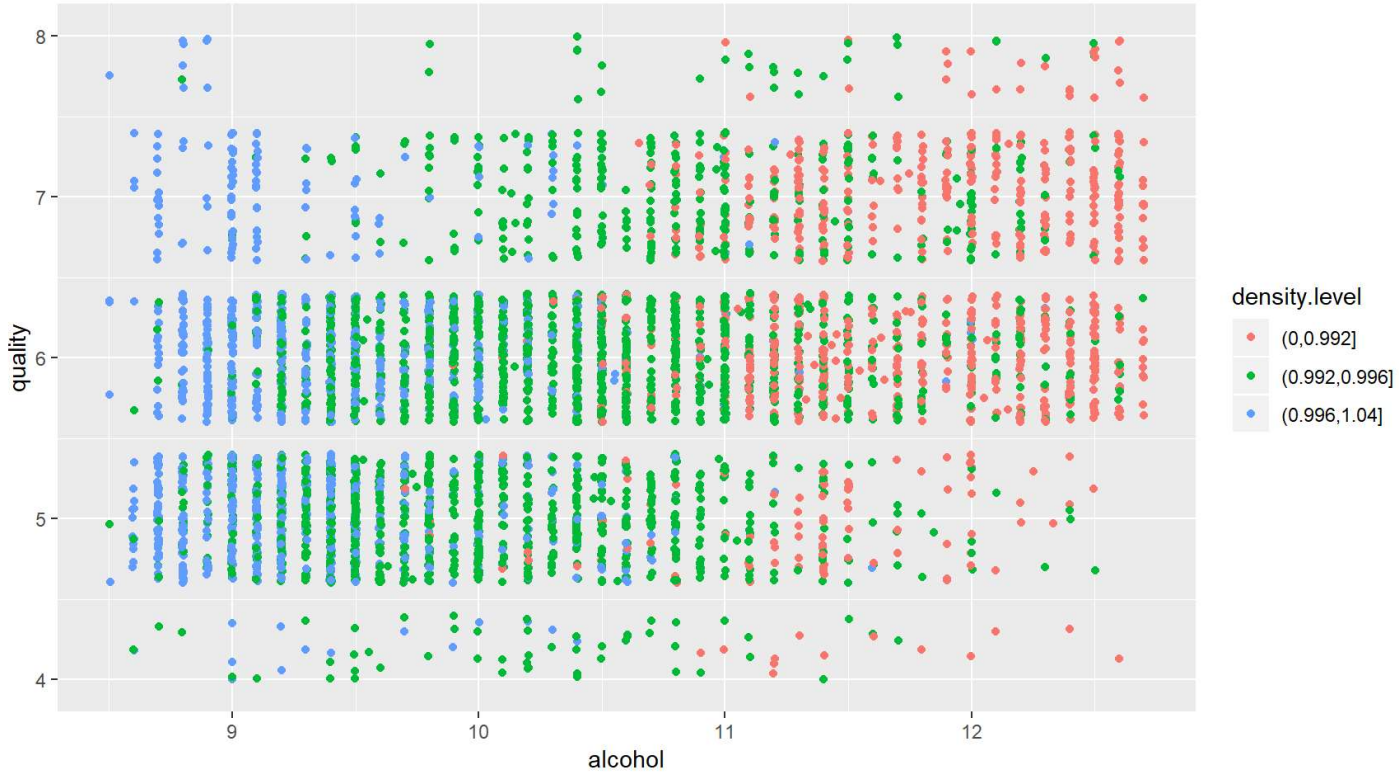
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00    9.50   10.40   10.51   11.40   14.20
```



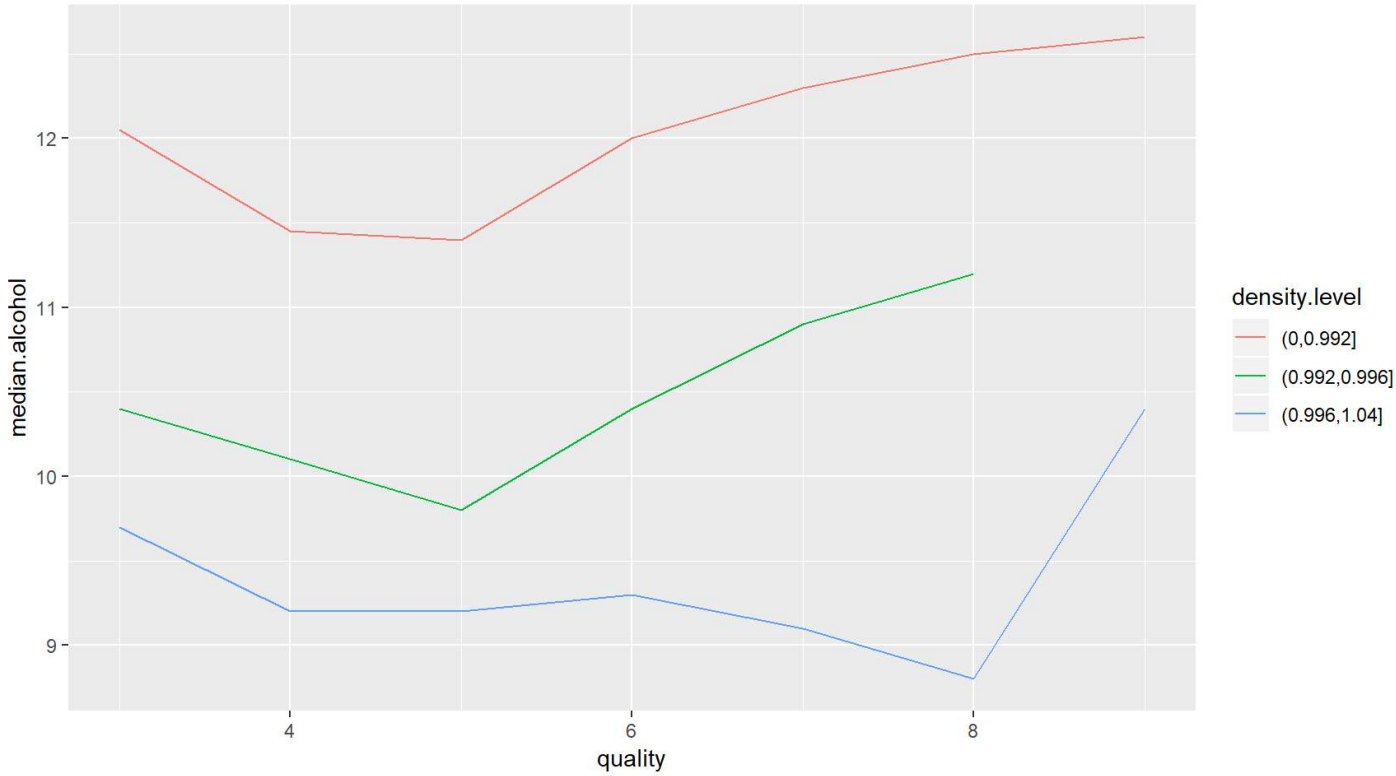
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9871  0.9917  0.9937  0.9940  0.9961  1.0390
```



Hmm, so we obviously see the relationship between alcohol and density but visually the relationship with quality doesn't seem obvious.



By zooming in we do notice that as the quality increases the number of low density high alcohol wines increase. This could be something.



Here we clearly see for the same quality as density increases, alcohol decreases!

Let us try to create a linear model including all the values:

```
##
## Calls:
## m1: lm(formula = quality ~ alcohol, data = df)
## m2: lm(formula = quality ~ alcohol + density, data = df)

## m3: lm(formula = quality ~ alcohol + density + fixed.acidity, data = df)
## m4: lm(formula = quality ~ alcohol + density + fixed.acidity + volatile.acidity,
##     data = df)
## m5: lm(formula = quality ~ alcohol + density + fixed.acidity + volatile.acidity +
##     citric.acid, data = df)
## m6: lm(formula = quality ~ alcohol + density + fixed.acidity + volatile.acidity +
##     citric.acid + I(log10(residual.sugar)), data = df)
## m7: lm(formula = quality ~ alcohol + density + fixed.acidity + volatile.acidity +
##     citric.acid + I(log10(residual.sugar)) + chlorides, data = df)
## m8: lm(formula = quality ~ alcohol + density + fixed.acidity + volatile.acidity +
##     citric.acid + I(log10(residual.sugar)) + chlorides + free.sulfur.dioxide +
##     bound.sulfur.dioxide, data = df)
## m9: lm(formula = quality ~ alcohol + density + fixed.acidity + volatile.acidity +
##     citric.acid + I(log10(residual.sugar)) + chlorides + free.sulfur.dioxide +
##     bound.sulfur.dioxide + pH, data = df)
## m10: lm(formula = quality ~ alcohol + density + fixed.acidity + volatile.acidity +
##     citric.acid + I(log10(residual.sugar)) + chlorides + free.sulfur.dioxide +
##     bound.sulfur.dioxide + pH + sulphates, data = df)
##
## ============================================================================================
=============================================================================
##                                  m1            m2            m3            m4            m5
m6            m7            m8            m9           m10
## --------------------------------------------------------------------------------------------
-----------------------------------------------------------------------------
##   (Intercept)                  2.582***    -22.492***    -32.669***    -48.571***    -48.568
***     24.044*       23.650*       20.522*       37.700***     47.757***
##                                (0.098)       (6.165)       (6.360)       (6.187)       (6.21
5)      (10.254)      (10.255)      (10.372)      (11.294)      (11.437)
##   alcohol                      0.313***      0.360***      0.373***      0.415***      0.415
***      0.337***      0.331***      0.336***      0.310***      0.296***
##                                (0.009)       (0.015)       (0.015)       (0.014)       (0.01
5)       (0.017)       (0.017)       (0.017)       (0.019)       (0.019)
##   density                                    24.728***     35.427***     51.673***     51.670
***    -21.112*      -20.603*      -17.623       -36.049**     -46.226***
##                                             (6.079)       (6.300)       (6.133)       (6.16
1)      (10.238)      (10.240)      (10.365)      (11.422)      (11.567)
##   fixed.acidity                                            -0.087***     -0.100***     -0.101
***     -0.062***     -0.064***     -0.055***     -0.020        -0.014
##                                                           (0.014)       (0.014)       (0.01
4)       (0.015)       (0.015)       (0.015)       (0.017)       (0.017)
##   volatile.acidity                                                      -2.115***     -2.114
***     -2.139***     -2.113***     -2.015***     -1.981***     -1.953***
##                                                                       (0.110)       (0.11
1)       (0.110)       (0.111)       (0.114)       (0.114)       (0.114)
##   citric.acid                                                                          0.001
0.035         0.056         0.024         0.057         0.037
##                                                                                       (0.09
6)       (0.095)       (0.096)       (0.096)       (0.096)       (0.096)
```

```
##   I(log10(residual.sugar))
0.442***      0.434***      0.391***      0.484***      0.535***
##
(0.050)       (0.050)       (0.051)       (0.056)       (0.057)
##   chlorides
-0.960        -1.029        -0.831        -0.818
##
(0.542)       (0.541)       (0.542)       (0.541)
##   free.sulfur.dioxide
0.004***      0.004***      0.003***
##
(0.001)       (0.001)       (0.001)
##   bound.sulfur.dioxide
-0.001        -0.001        -0.001*
##
(0.000)       (0.000)       (0.000)
##   pH
0.343***      0.317***
##
(0.090)       (0.090)
##   sulphates
0.502***
##
(0.099)
## ------------------------------------------------------------------------------------------
-------------------------------------------------------------------------
##   R-squared                     0.190         0.192         0.199         0.255         0.255
0.267         0.267         0.272         0.274         0.278
##   adj. R-squared                0.190         0.192         0.198         0.255         0.254
0.266         0.266         0.270         0.272         0.276
##   sigma                         0.797         0.796         0.793         0.765         0.765
0.759         0.759         0.756         0.755         0.754
##   F                          1146.395       583.290       404.474       419.207       335.297
296.933       255.074       202.639       184.336       170.797
##   p                             0.000         0.000         0.000         0.000         0.000
0.000         0.000         0.000         0.000         0.000
##   Log-likelihood            -5839.391     -5831.127     -5812.175     -5632.940     -5632.940
-5593.927     -5592.356     -5578.095     -5570.814     -5557.825
##   Deviance                   3112.257      3101.773      3077.862      2860.647      2860.647
2815.438      2813.632      2797.295      2788.992      2774.238
##   AIC                       11684.782     11670.255     11634.351     11277.879     11279.879
11203.854     11202.711     11178.189     11165.629     11141.649
##   BIC                       11704.272     11696.241     11666.834     11316.859     11325.355
11255.827     11261.181     11249.651     11243.588     11226.105
##   N                          4898          4898          4898          4898          4898
4898          4898          4898          4898          4898
## ==========================================================================================
=========================================================================
```

```
##
##    3    4    5    6    7    8    9
##   20  163 1457 2198  880  175    5
```

So the highest R-squared value we get is 0.278, so with this linear model we cannot obtain a very accurate value for quality from the given features.

Multivariate Analysis

Talk about some of the relationships you observed in this part of the investigation. Were there features that strengthened each other in terms of looking at your feature(s) of interest?

Here I added a third variable as color and was able to obtain a better idea about the relationship between:

- residual.sugar, alcohol and density
- quality, density and alcohol.

Were there any interesting or surprising interactions between features?

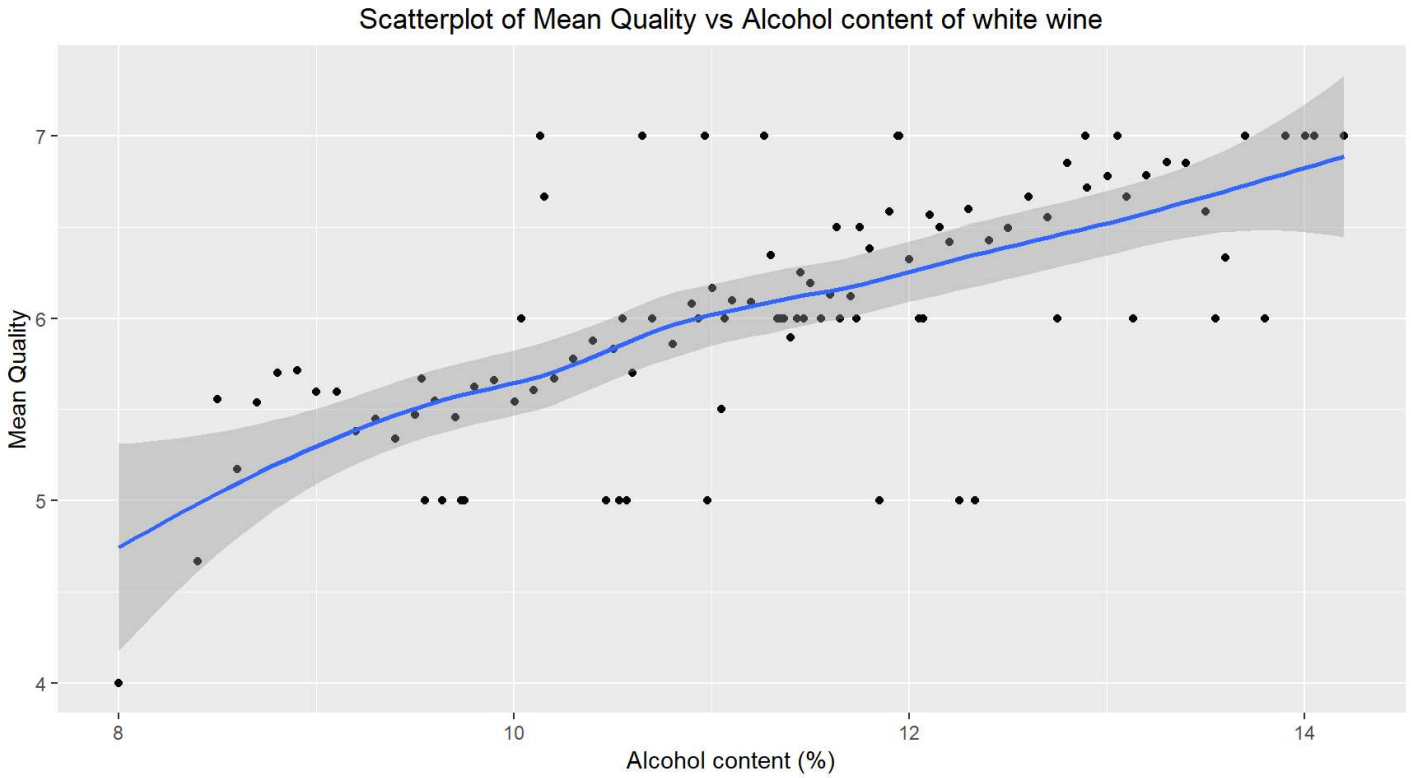
I found the interaction between quality, density and alcohol to be very interesting. Particularly the balance they need to maintain in the white wine.

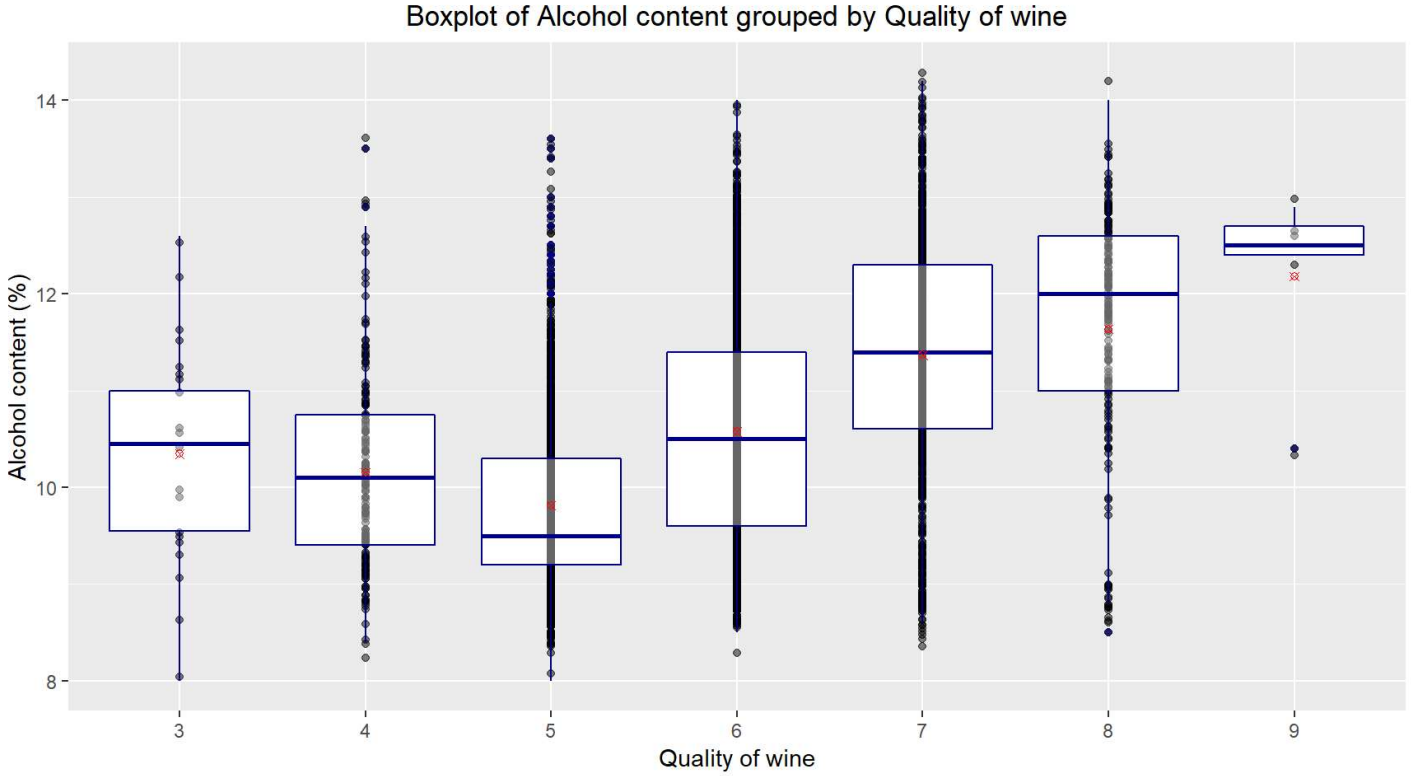
OPTIONAL: Did you create any models with your dataset? Discuss the strengths and limitations of your model.

I tried creating a linear model to predit the quality but the results I obtained show that a linear model is probably not the best way to go to fit this model, or we are missing vital features from our dataset.

Final Plots and Summary

Plot One

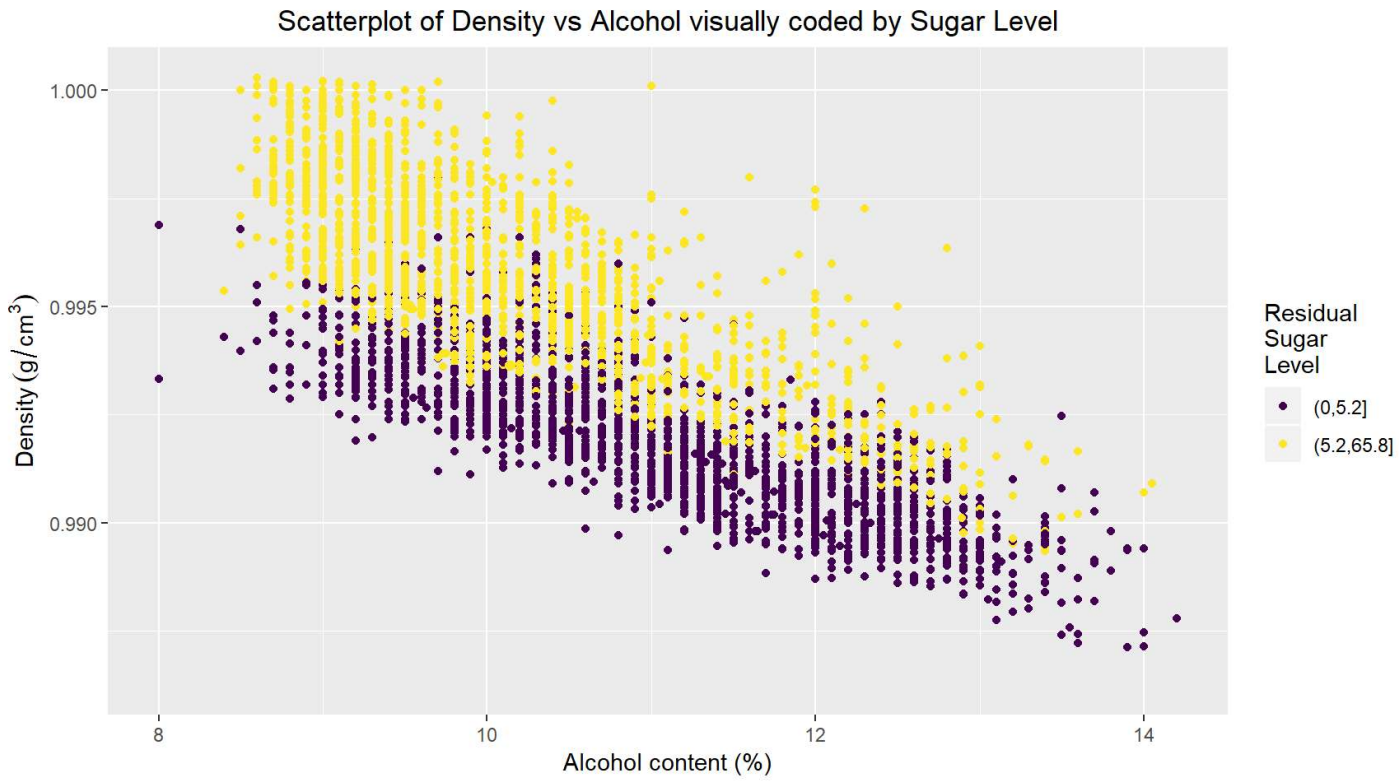




Description One

This is one of the main features I explored in relation with quality as an outcome. Alcohol has the highest correlation with quality and its relationship can easily be viewed here by the clear linear relationship between mean quality and alcohol content that can be observed in the scatterplot. There is also a change in trends between alcohol and quality as observed by the boxplot. There seems to be a threshold of alcohol(as noticed by the dip in 5 quality) after which it needs to increases in relation with other ingredients to get a better quality of wine.

Plot Two

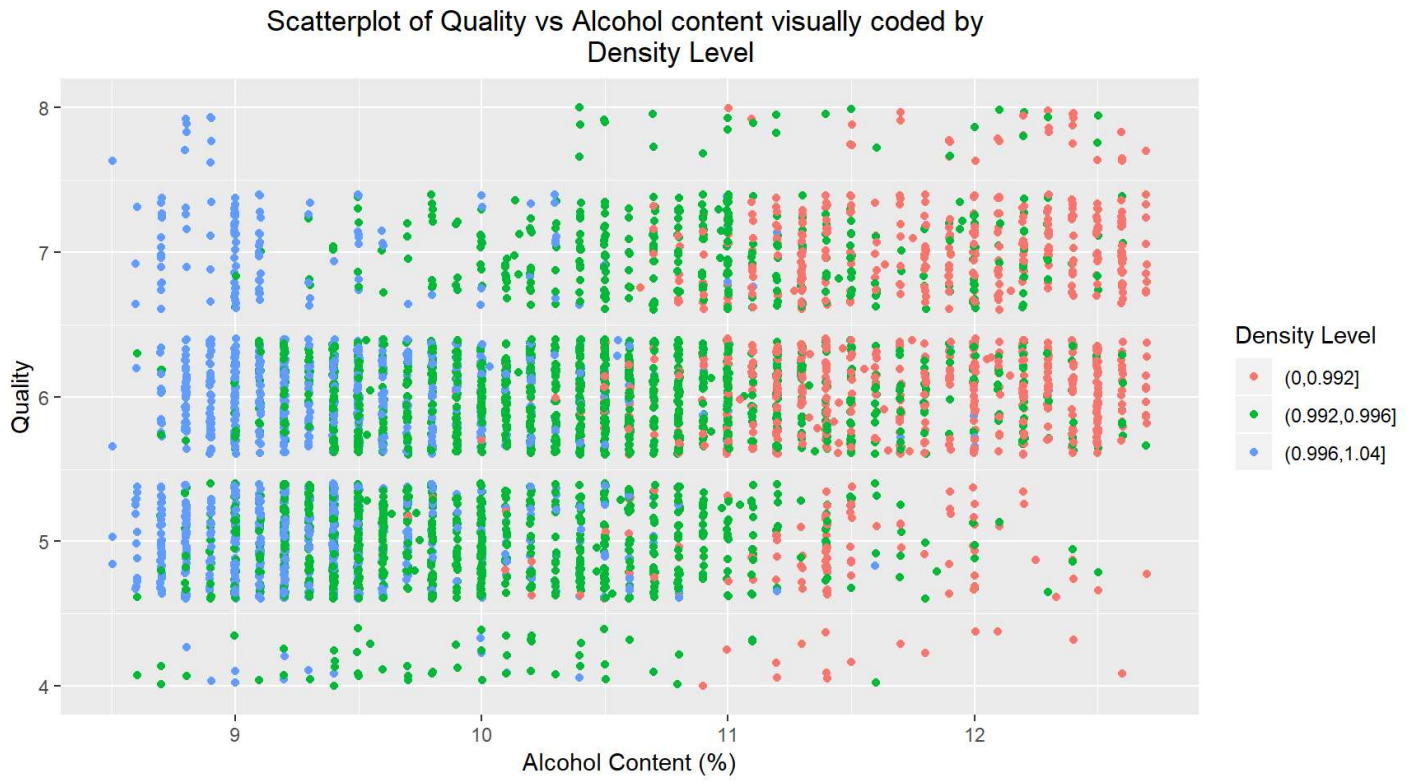


Description Two

This plot perfectly shows the relationship between residual.sugar, alcohol and density. I has split the residual.sugar into two buckets because of it's bimodal nature. I split the buckets by the median of residual.sugar.

As alcohol increases, density and sugar level decreases. As alcohol decreases, density and sugar level increases. This could be explained by the fact that, during the manufacturing of wine, fermentation converts the sugars to alcohol. This explains why better quality wine has had better fermentation and has allowed more of the sugars to convert to more alcohol.

Plot Three



Description Three

This plot is heavily connects to the previous plots. We can see by the number of the points on each level of quality that as quality increases, the number of low density high alcohol content points increase.

We can see a clear evidance of this as we move up from rating 5 with more blue dots and less red dots up to 7 with much more red dots and less blue. Note that blue dots always relate with low alcohol and high density wines while red dots relate with high alcohol and low density wines.

We can also notice this theory applying to rating 3 with more blue and green dots and very few red dots and rating 8 with more red dots and very few blue and green dots. If we had a more samples for poor and excellent wines we would have noticed this trend more vividly.

Reflection

One of my biggest struggles was settling on which features to focus on. I think even with the report I created, I still missed out on a lot of possible avenues that I could have explored. Given more time I would have liked to delve deeper into creating a better model for predicting quality of white wine but I feel I would need more domain knowledge of the making of white wines than I currently have.

Nonetheless, I had a lot of success in using different aspects of R and its vast EDA capabilities especially in it's visualizations. I found considerable success in showing connections between density, alcohol content and residual.sugars.

In the future, I would probably like to spend more time even exploring deadends and gathering side data as well, like cost of the wines and their age.