Red Wines Exploration by Gourav Aich

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553. ISSN: 0167-9236.

Available at: [@Elsevier] http://dx.doi.org/10.1016/j.dss.2009.05.016 (http://dx.doi.org/10.1016/j.dss.2009.05.016)
[Pre-press (pdf)] http://www3.dsi.uminho.pt/pcortez/winequality09.pdf (http://www3.dsi.uminho.pt/pcortez/winequality09.pdf)
[bib] http://www3.dsi.uminho.pt/pcortez/dss09.bib (http://www3.dsi.uminho.pt/pcortez/dss09.bib)

This report explores a dataset containing 1599 instances of red wine and its 12 attributes. Out of these 12 attributes, eleven are based on physiochemical tests and are termed as input variables. On the other hand, the twelfth attribute "quality", scored between 0(very bad) and 10(excellent) is based on sensory data and is termed as the output variable.

Univariate Plots Section

```
## [1] 1599   13
```

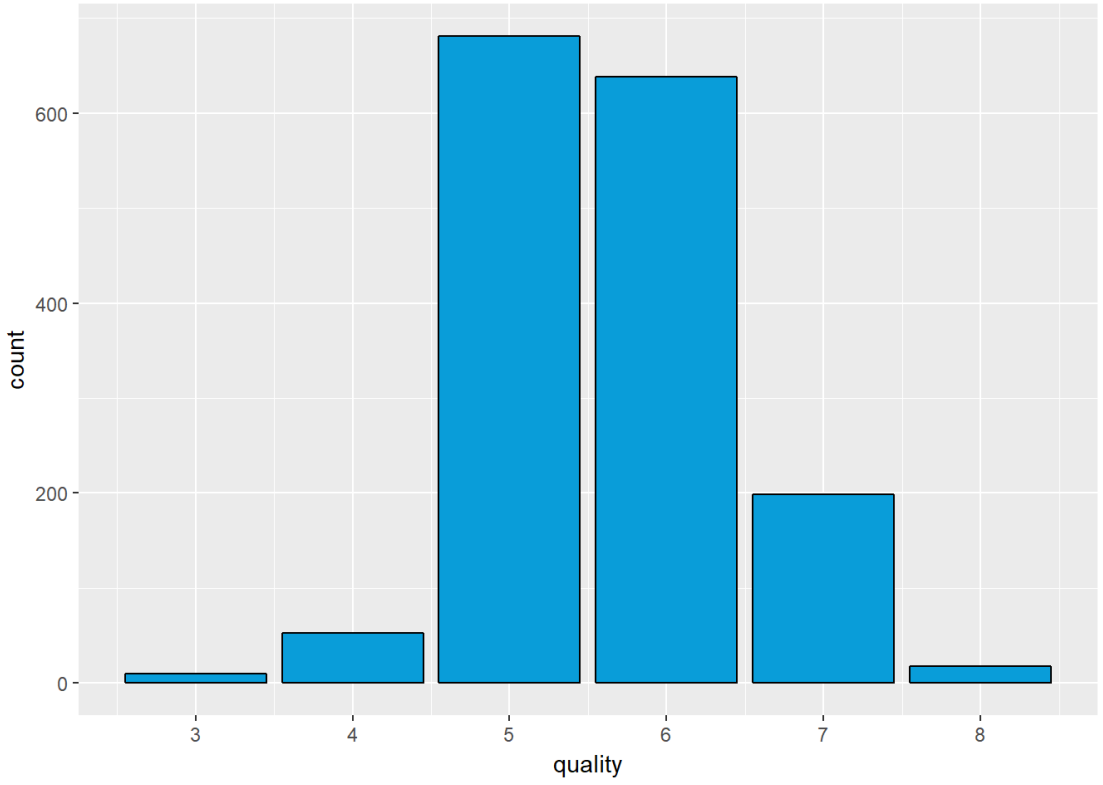
```
## 'data.frame':    1599 obs. of  13 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
##  $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
##  $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071 ...
##  $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
##  $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
##  $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
##  $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...
##  $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...
##  $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
##  $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...
```

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

Our dataset consists of thirteen variables, with 1599 observations. The first variable “X” denotes the row number, and therefore, can be ignored for further analysis. Out of the remaining 12 variables, 11 are of numeric/floating data types and are also the input variables. The twelfth variable “quality” is of integer data type and is the output variable.

Just by looking at the summary, the following variables seem to have some extreme max. values/outliers when compared to their means and 3rd Quartile values:

- residual.sugar
- chlorides
- free.sulfur.dioxide
- total.sulfur.dioxide
- sulphates

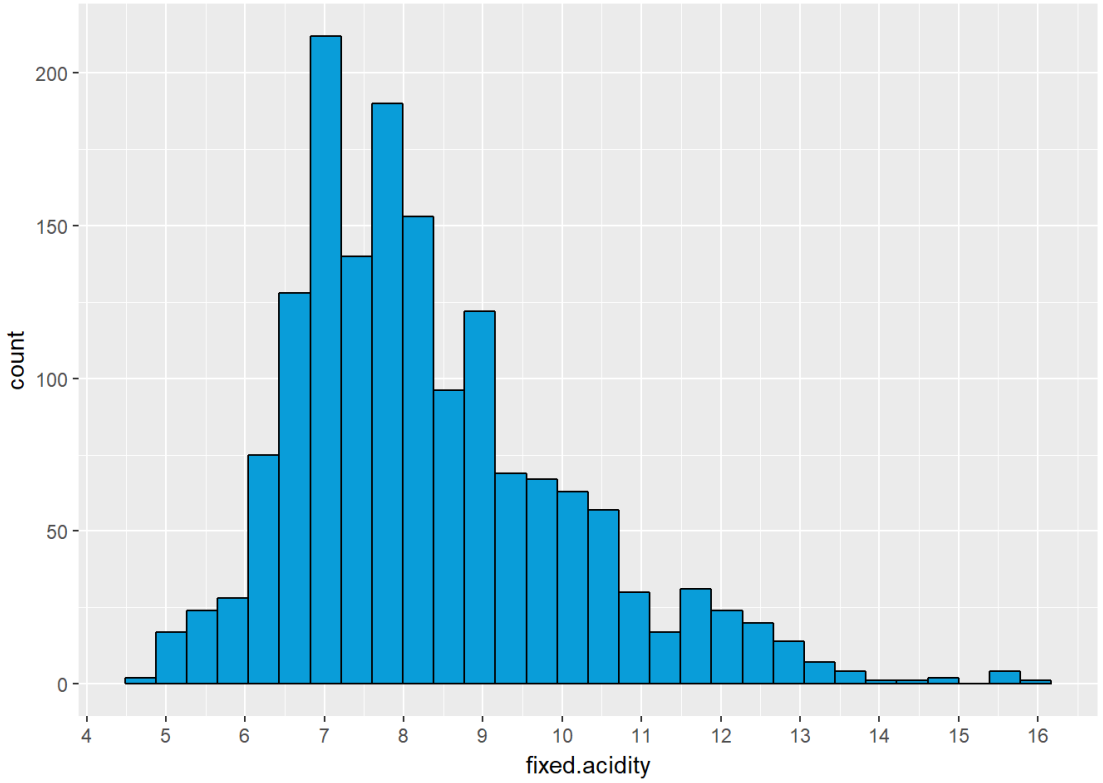



```
##
##   3   4   5   6   7   8
##  10  53 681 638 199  18
```

The **quality** histogram above has a normal distribution, peaking at quality equal to 5 and 6. It’s worth mentioning here that a lot of common statistical techniques, like linear regression (which we will use as we plan to seek correlation between quality and other variables), are based on the assumption that variables have normal distributions.

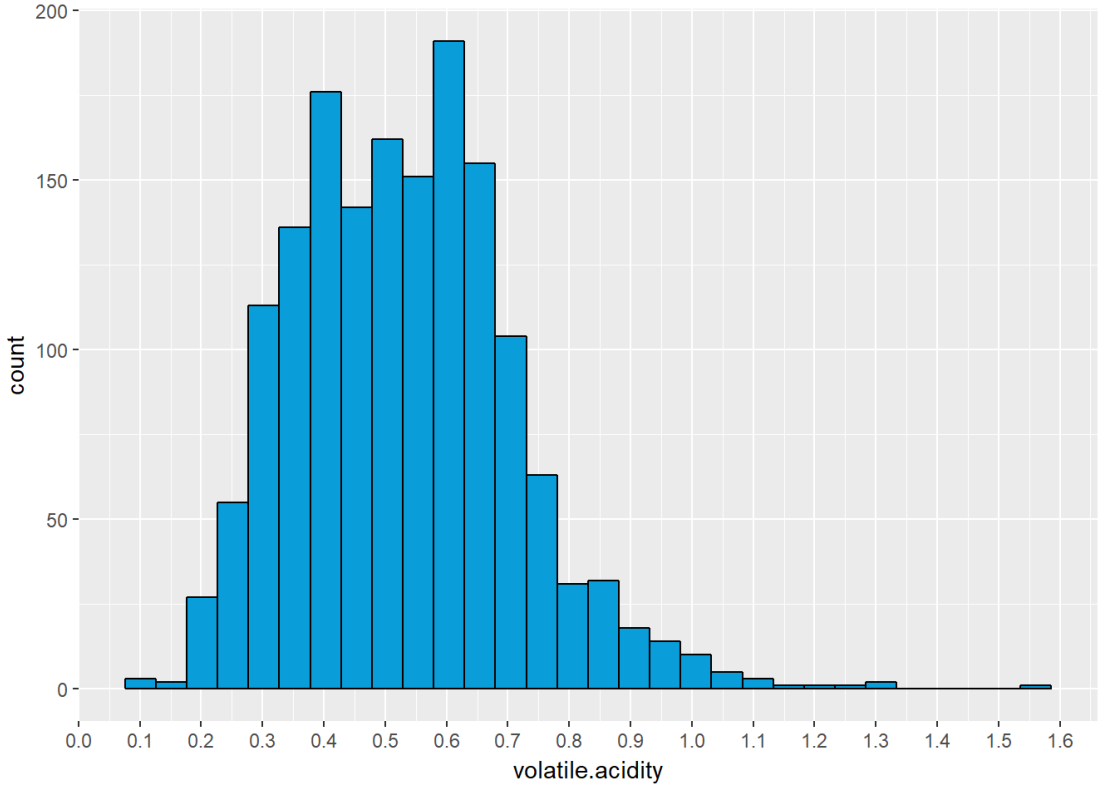
The table below the histogram shows us that more than 80% of the wine samples were graded average by experts.

Now, let’s plot histograms and boxplots for the input variables to understand their distributions and outliers respectively.



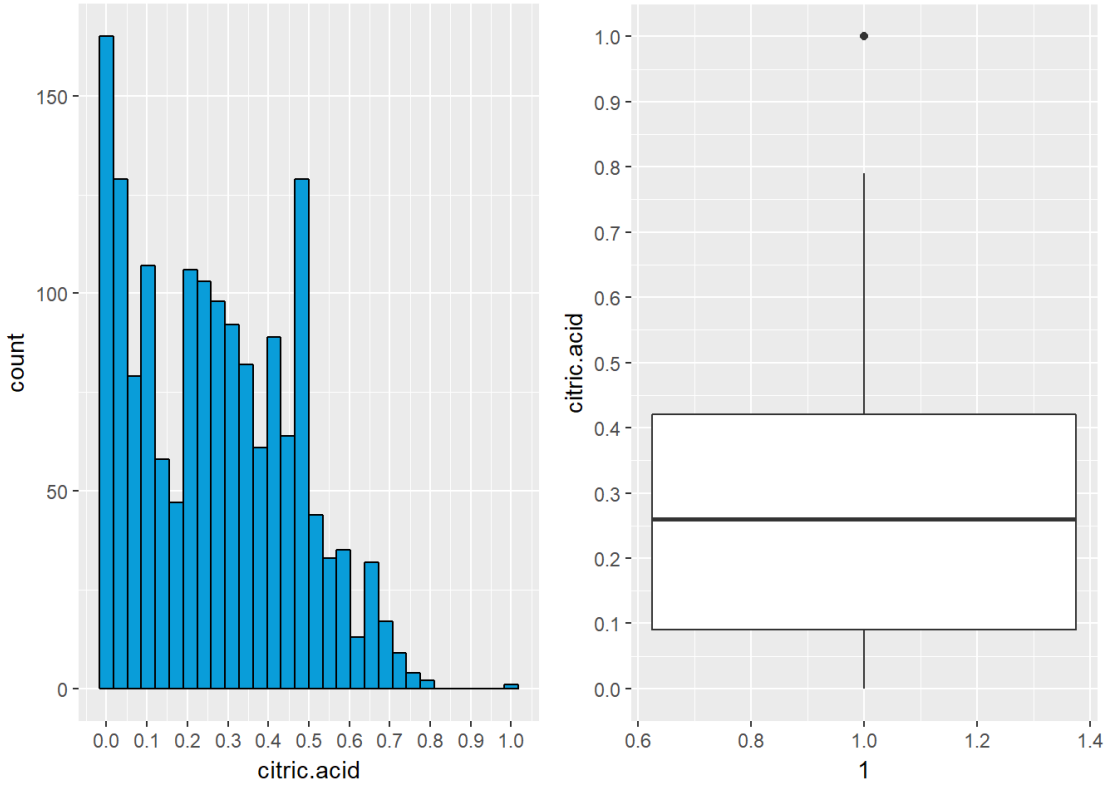

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.60    7.10    7.90    8.32    9.20   15.90
```

The distribution of **fixed.acidity** appears bimodal and peaks at 7 and 7.75 (*approx.*).




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1200  0.3900  0.5200  0.5278  0.6400  1.5800
```

Similar to fixed.acidity, **volatile.acidity** also appears bimodal and peaks at 0.4 and 0.6.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.090   0.260   0.271   0.420   1.000
```

The histogram of **citric.acid** is positively skewed. The log transformation doesn't help much either and just reverses the direction of skewness. However, one interesting fact to notice is the tall bar for citric.acid=0. Let's find out the number of wines having zero citric acid.

```
## [1] 132
```

So, 132 wines are reported having no citric acid, at all. This might be an issue of non-reporting. However, a little bit of digging on the internet does reveal that many types of red wines do not contain any citric acid, according to tests conducted by the Food Standards Australia New Zealand website [http://www.livestrong.com/article/189520-what-drinks-do-not-contain-citric-acid/ (http://www.livestrong.com/article/189520-what-drinks-do-not-contain-citric-acid/)].
So, I really can't be sure about these 132 wines. Moving on, let's take a look at the distribution of quality for wines with citric.acid=0 and otherwise.

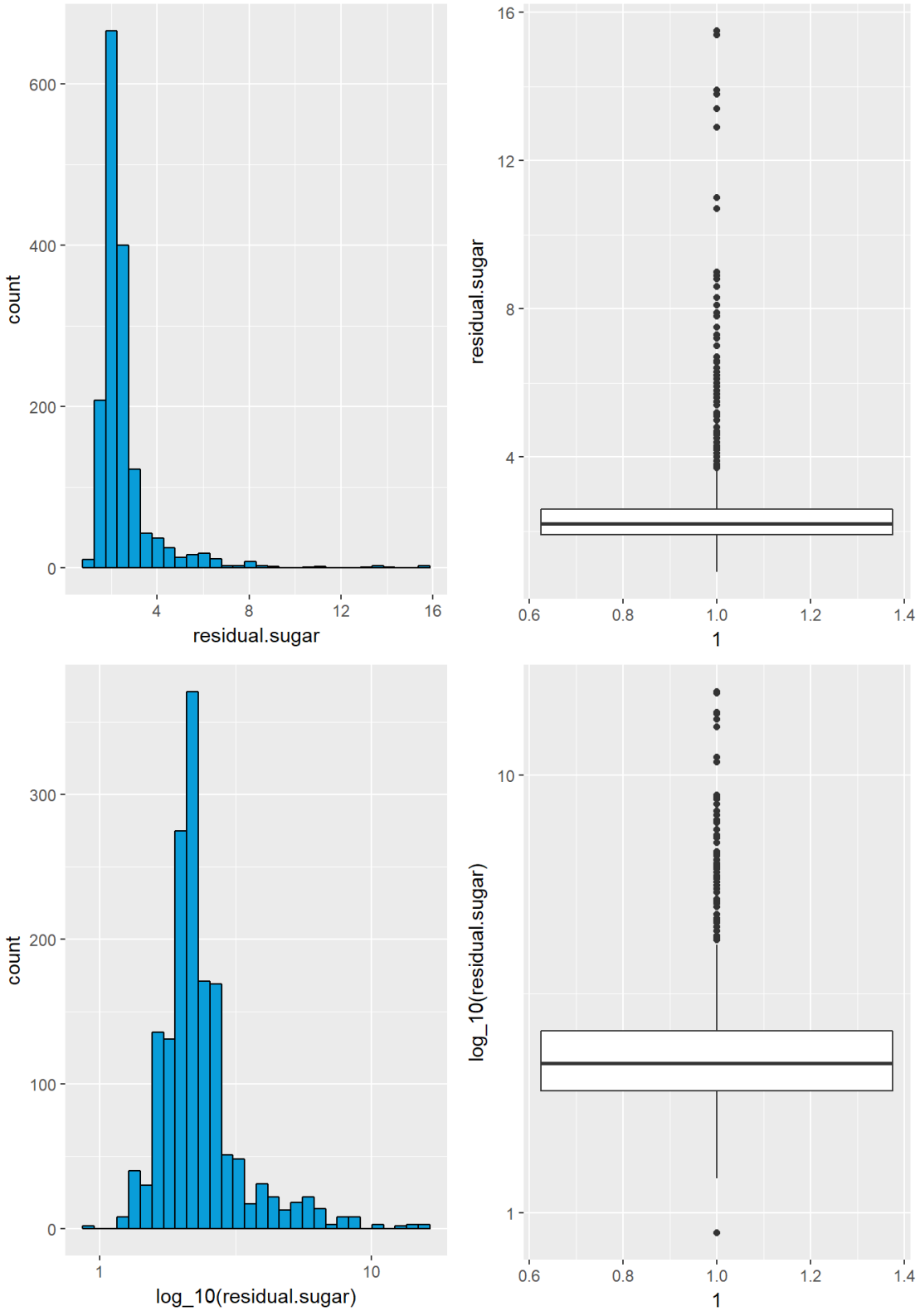
```
## redwines$citric.acid == 0: FALSE
##
##   3   4   5   6   7   8
##   7  43 624 584 191  18
## ------------------------------------------------------------
## redwines$citric.acid == 0: TRUE
##
##  3  4  5  6  7
##  3 10 57 54  8
```

The quality of red wines seems to be proportionately distributed across both the categories. However, none of the wines with citric.acid=0 are graded 8 on the quality scale.

Coming to the boxplot, it's quite interesting to see that there is probably only 1 outlier (with citric.acid=1). Let's extract the record(s).

```
##     X fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1 152           9.2             0.52           1            3.4      0.61
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  32                   69  0.9996 2.74         2     9.4
##   quality
## 1       4
```

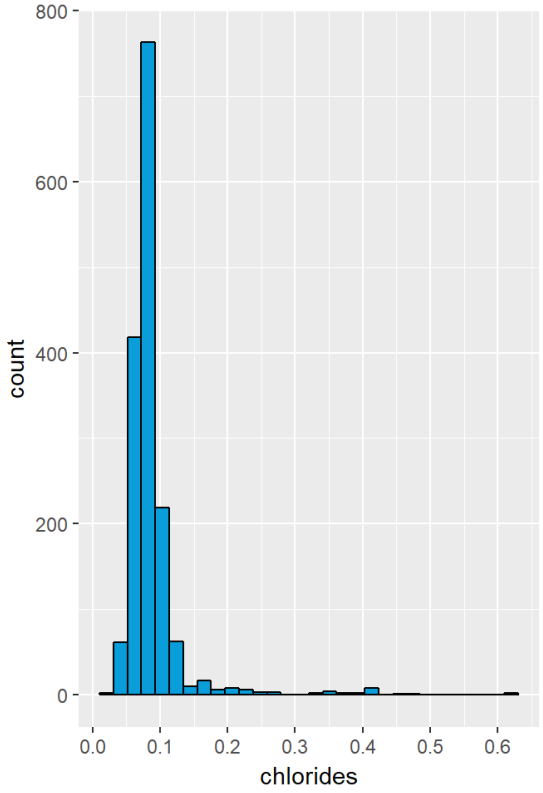
Well, as speculated, there is only 1 wine having citric.acid=1 and it's quality has been graded 4 (*that's quite poor*). Does this mean that higher citric acid reduces the quality, or, being an outlier, this is probably erroneous data? It would be interesting to report our findings when I plot quality against citric.acid during our bivariate analysis.

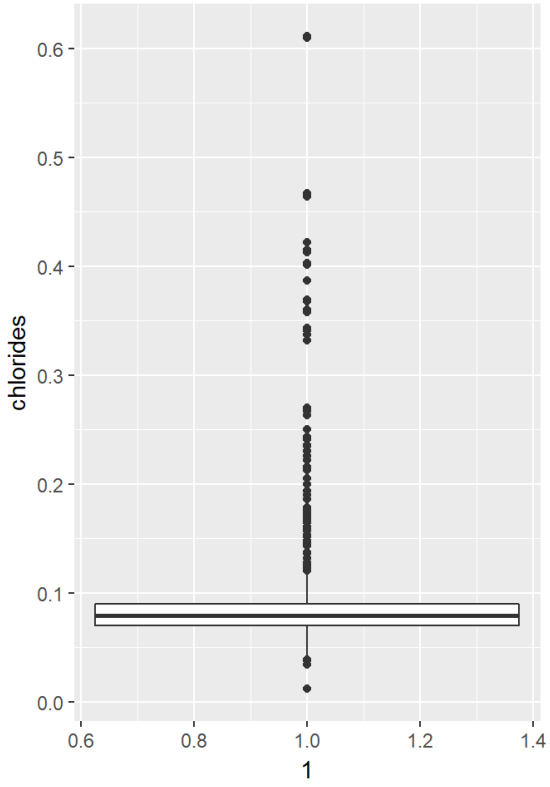



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.900   1.900   2.200   2.539   2.600  15.500
```

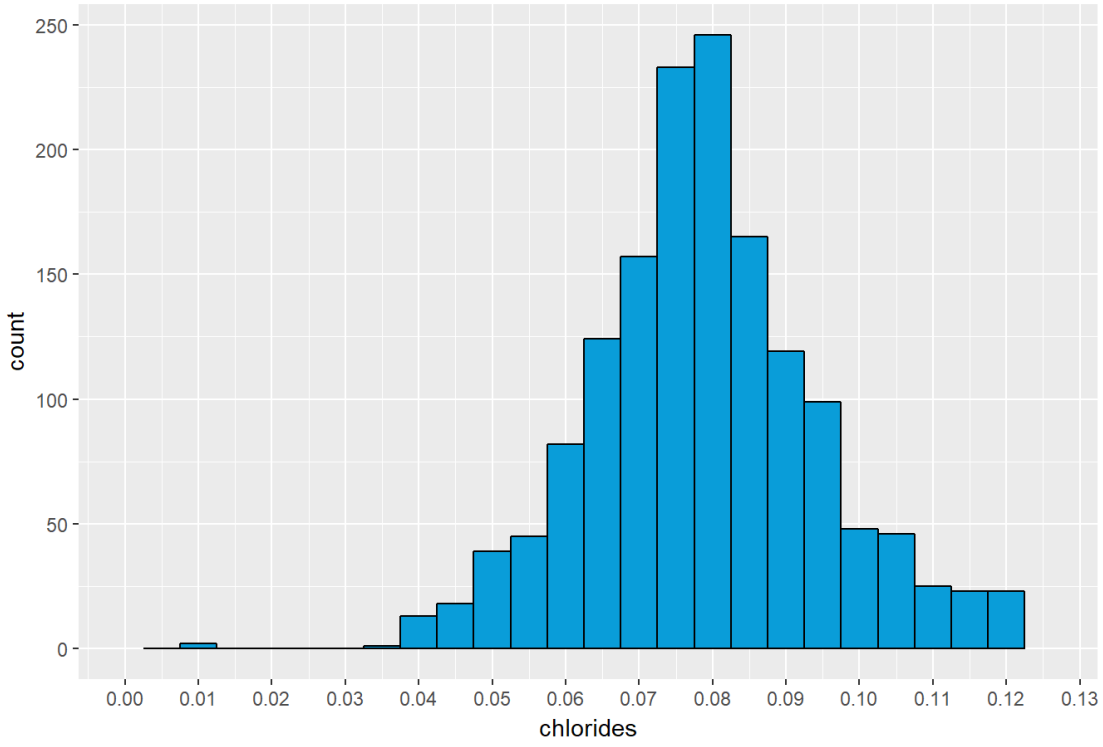
```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.04576  0.27880  0.34240  0.36930  0.41500  1.19000
```

Transformed the long-tail data to better understand the distribution of **residual.sugar**. The log-transformed distribution is relatively less skewed and more normal. As a result of which, the second data-summary (for the log-transformed data) looks much better than the first one.





95% quartile of chlorides

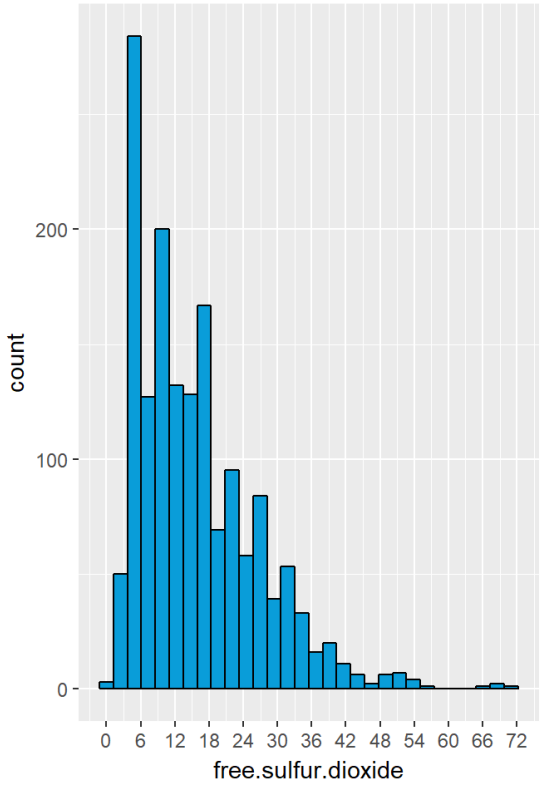


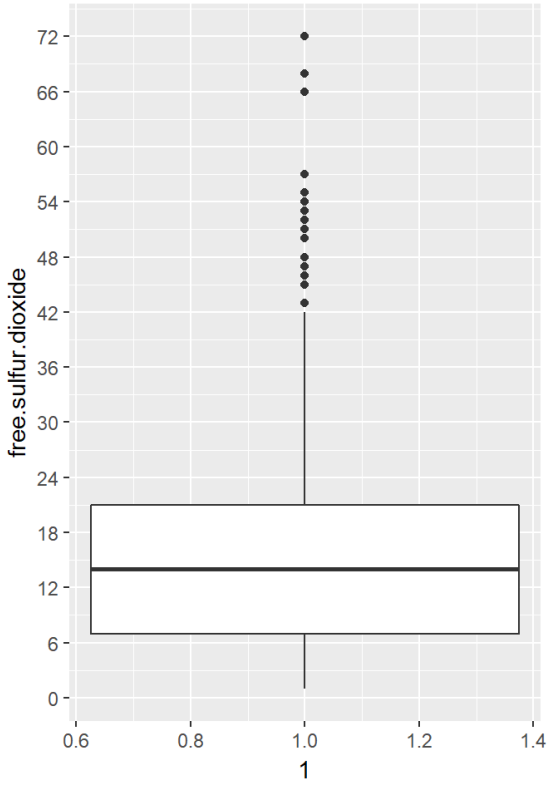

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01200 0.07000 0.07900 0.08747 0.09000 0.61100
```

The distribution of **chlorides** is extremely right skewed. The huge difference between 3rd quartile and the max. value supports this fact as does the boxplot (*which shows several outliers*). So, I omitted the top 5% of the values (95th percentile truncated) and adjusted the binwidth. Now, the distribution appears normal. However, I can see a tiny bar at 0.01. Let's find out these wines having chlorides < 0.02.

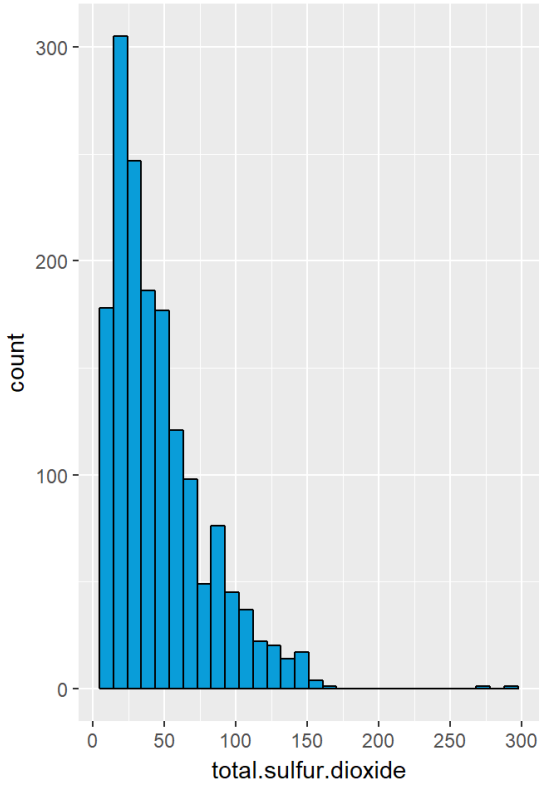
```
##     X fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1 837           6.7             0.28        0.28            2.4     0.012
## 2 838           6.7             0.28        0.28            2.4     0.012
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  36                  100 0.99064 3.26      0.39    11.7
## 2                  36                  100 0.99064 3.26      0.39    11.7
##   quality
## 1       7
## 2       7
```

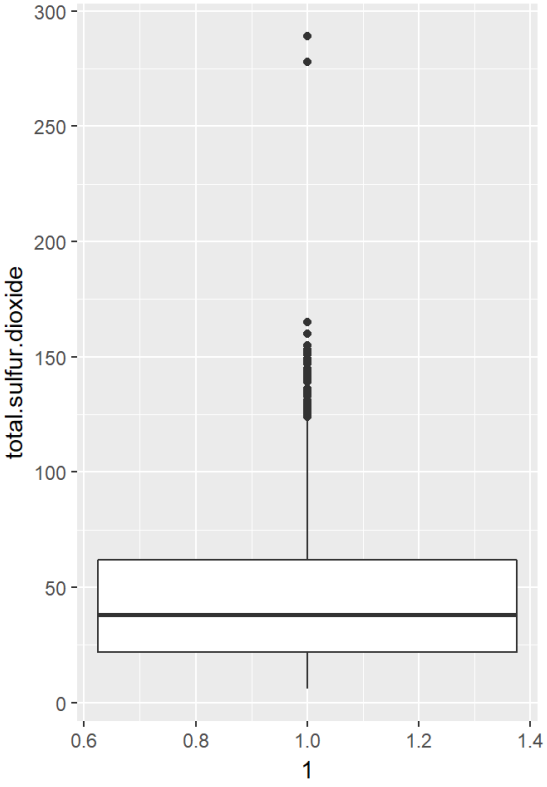
Well, I have two wine samples with identical data. Do I have more such records? As I took a closer look at the dataset, I found that there are many such groups of wine samples with exactly the same data for all the columns. This is probably due to the same wine variant being graded the same by the experts.






```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    7.00   14.00   15.87   21.00   72.00
```

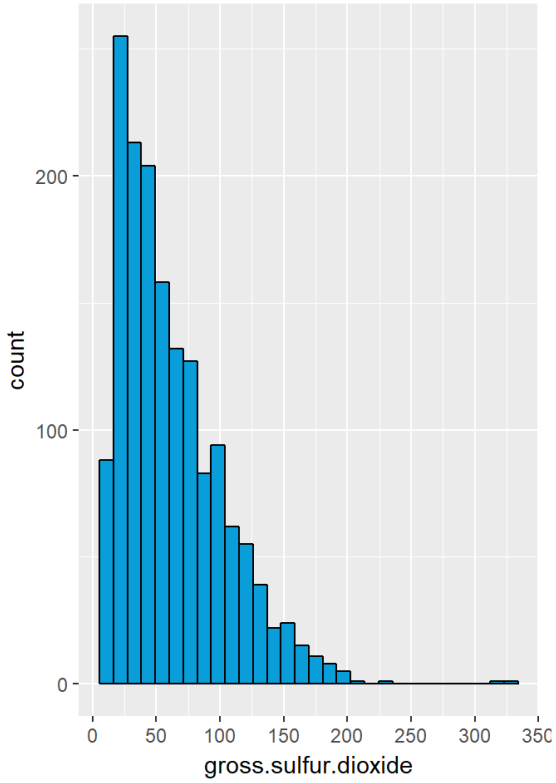


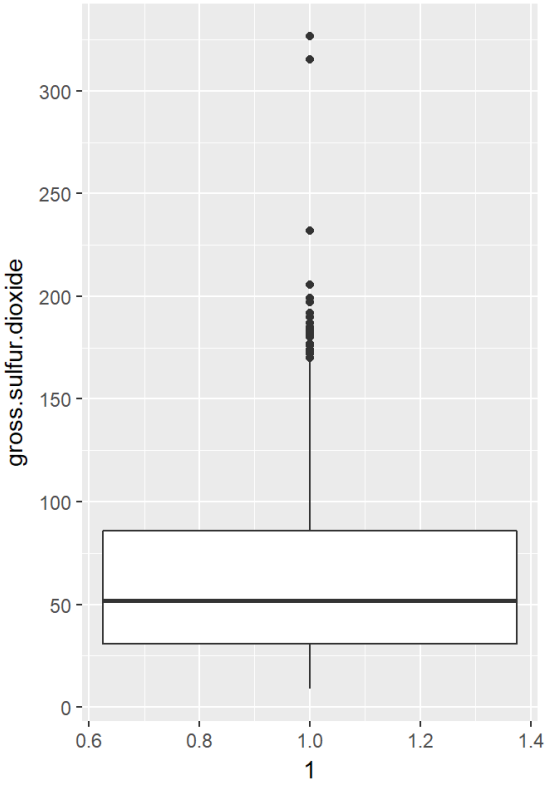


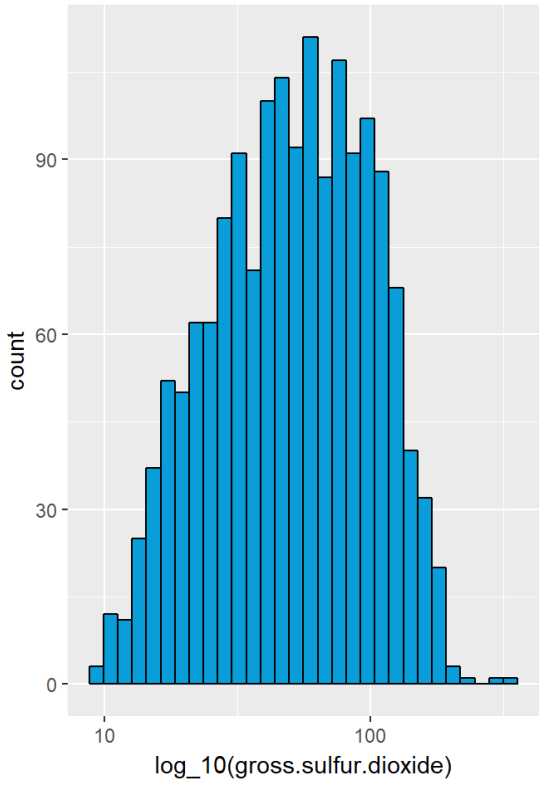

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00   22.00   38.00   46.47   62.00  289.00
```

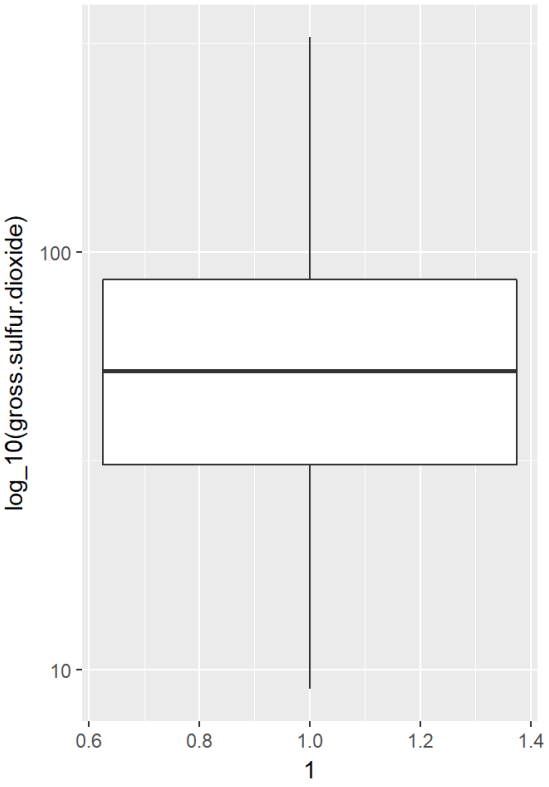
The **free.sulfur.dioxide** is again long-tailed and it's max. value is way beyond the 3rd quartile. The log transformed plot doesn't make the plot any more normal. However, it reduces the number of outliers drastically to just one.

Similar to the previous plot, **total.sulfur.dioxide** distribution is right-skewed, too. Rather than studying these 2 related sulfur.dioxide features separately, I think it would be a good idea to combine them both.





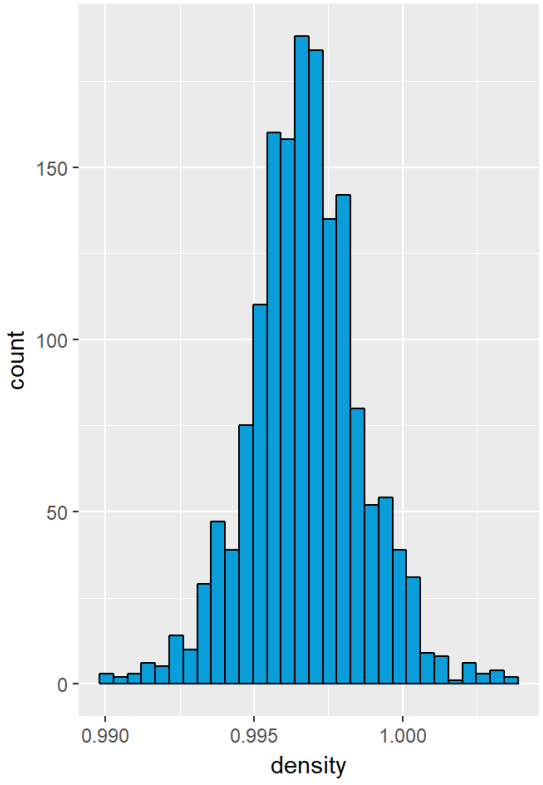


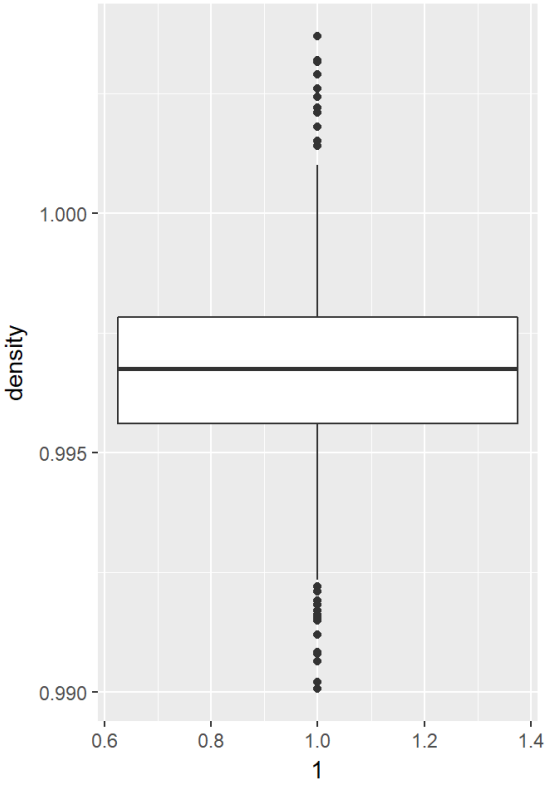



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.00   31.00   52.00   62.34   86.00  326.50
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9542  1.4910  1.7160  1.7020  1.9340  2.5140
```

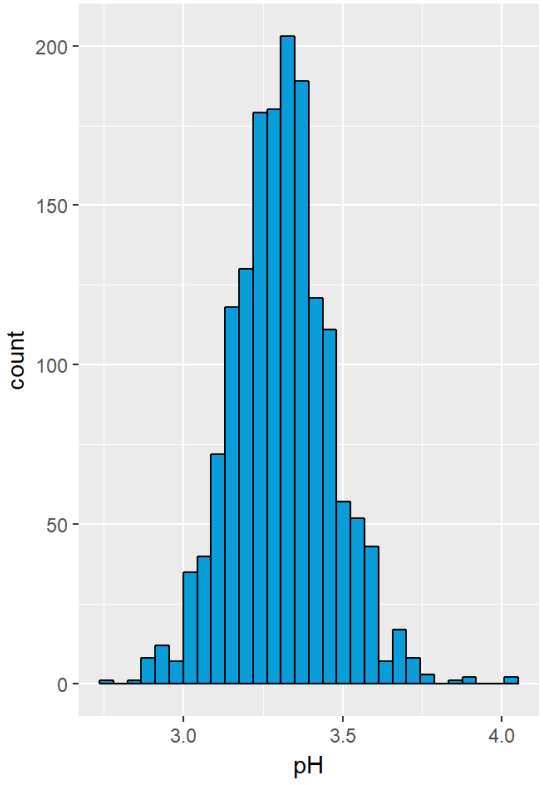
I summed up both the sulfur dioxides to create a new variable gross.sulfur.dioxide.
The right skewed **gross.sulfur.dioxide** distribution is log transformed and results in a somewhat normal distribution (check both the data summaries above, the 2nd one being the log-transformed one). One interesting observation is that the boxplot of log-transformed data shows zero outliers.

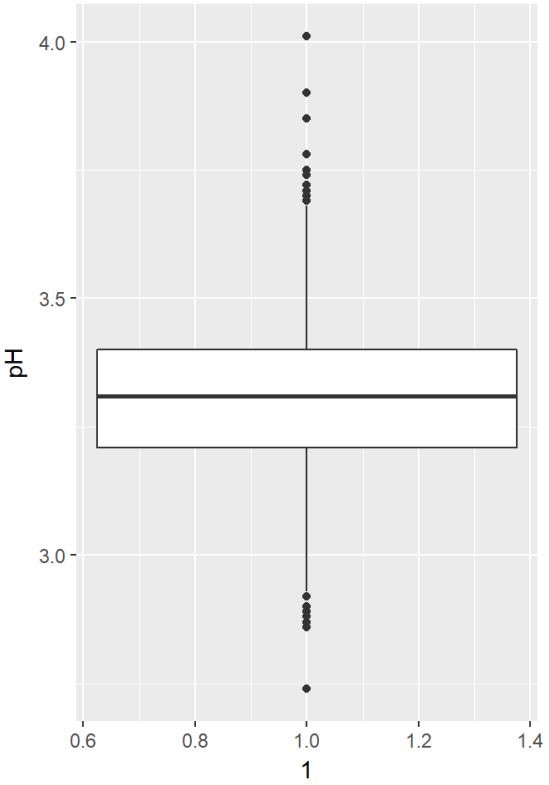





```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9901  0.9956  0.9968  0.9967  0.9978  1.0040
```

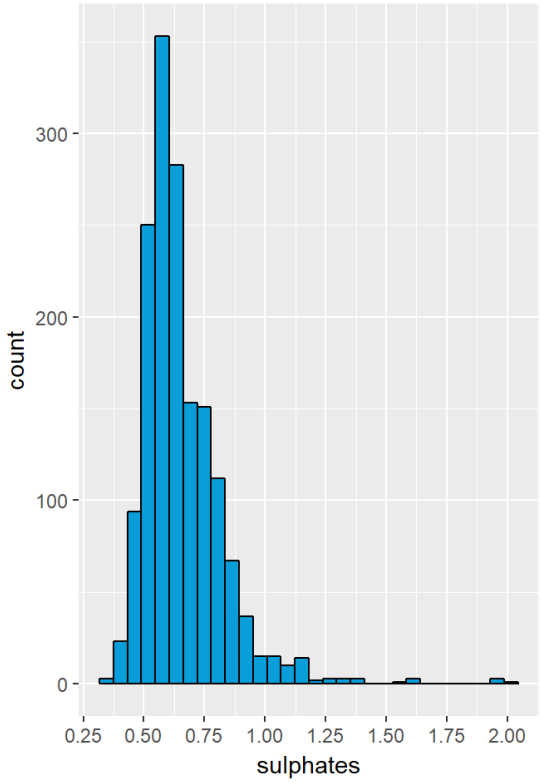
The histogram of **density** is normal with outliers appearing on both sides of the boxplot. One fact worth mentioning is that density depends on alcohol percentage and sugar content. So, it would be worth investigating their relation during further analysis.

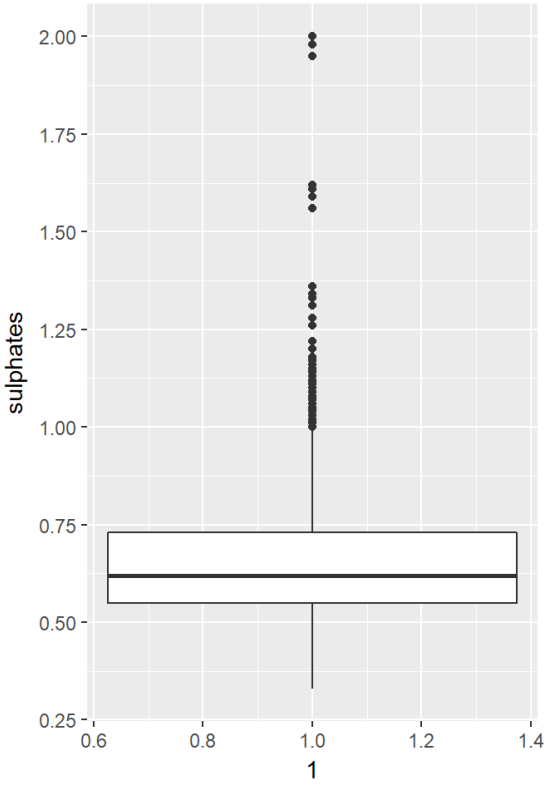


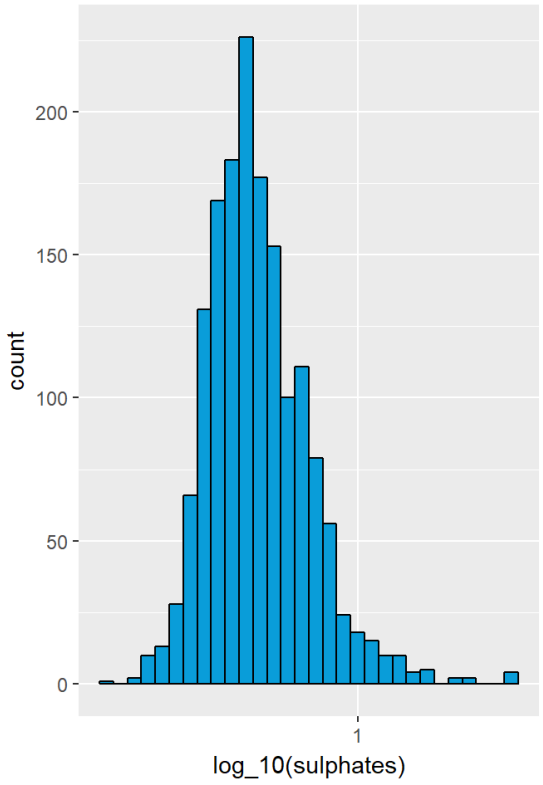


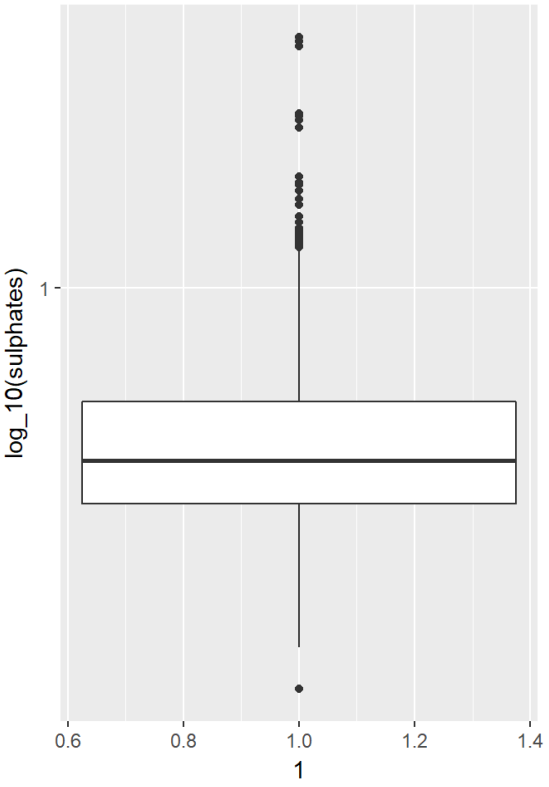

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.740   3.210   3.310   3.311   3.400   4.010
```

Similar to density, **pH**, too, has a normal distribution with outliers at both ends. pH, being a measure of acidity, might have a correlation with fixed.acidity and is definitely worth investigating.



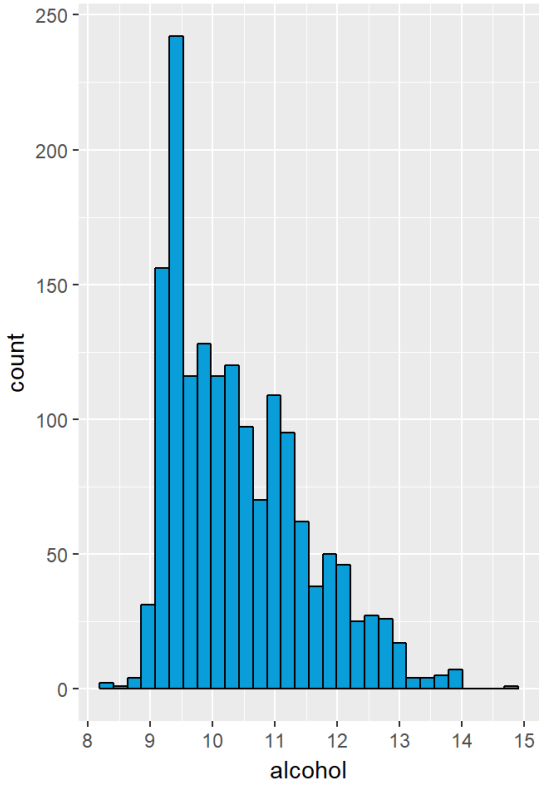


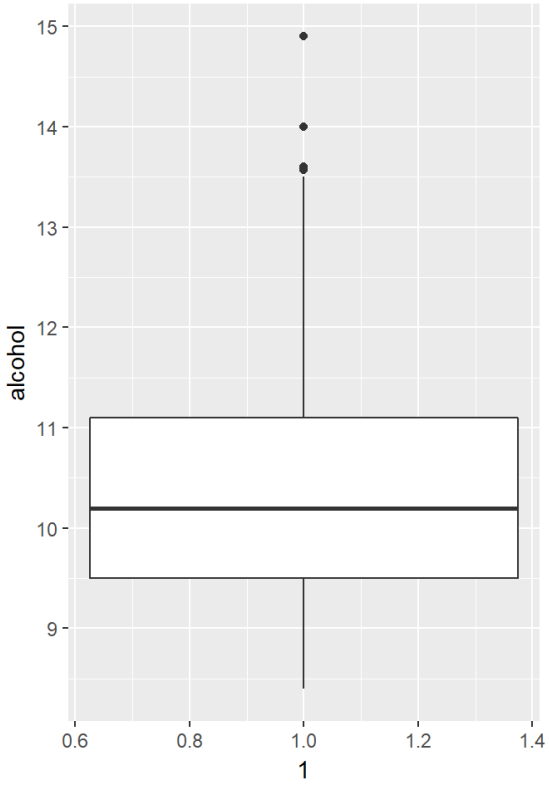


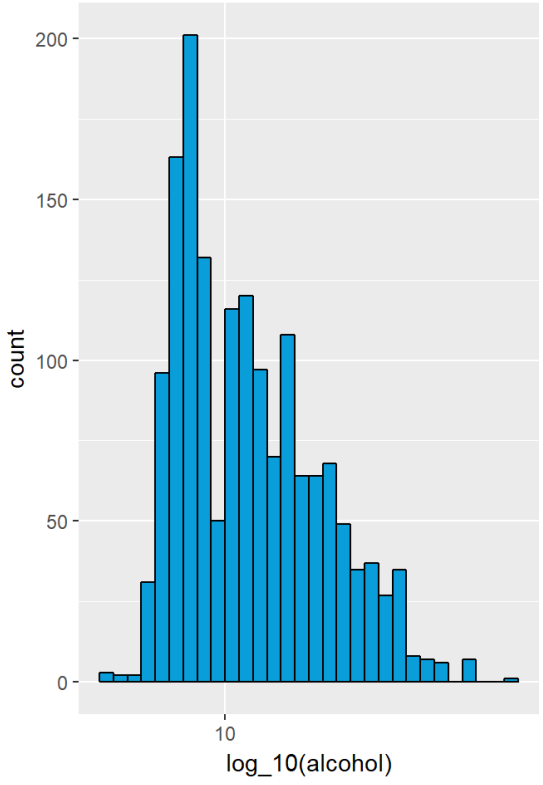


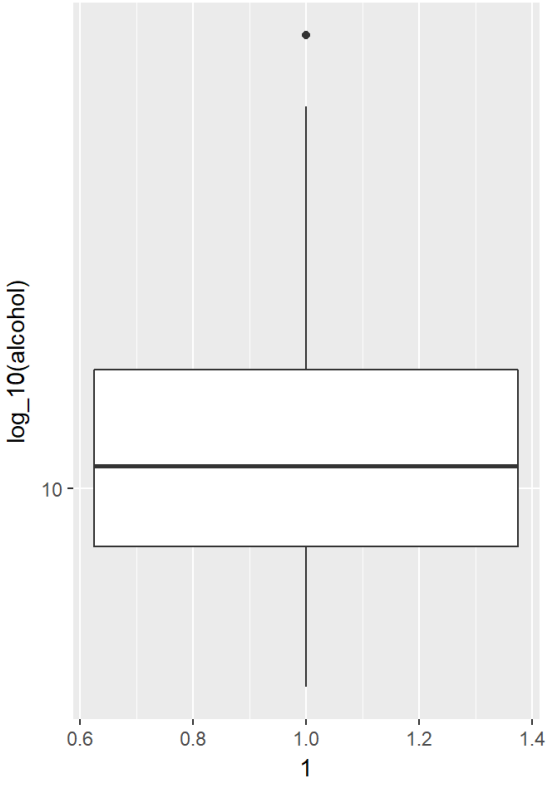

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3300  0.5500  0.6200  0.6581  0.7300  2.0000
```

The histogram of **sulphates** is definitely long-tailed, whereas, the log-transformed data gives us a relatively normal distribution. The outliers though, still exist in large numbers.










```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.50   10.20   10.42   11.10   14.90
```

```
percent of wines containing less than 12% alcohol -->  [1] 89.86867
```

Alcohol is long-tailed with a very few outliers. Almost 90% of the wines have less than 12% alcohol. The log-transformed distribution, though resembles the actual one, reduces the outliers to just one.

Univariate Analysis

What is the structure of your dataset?

This dataset contains 1599 instances of red wine and its 12 attributes. Out of these 12 attributes, 11 are based on physiochemical tests and are termed as input variables. On the other hand, the twelfth attribute “quality”, scored between 0(very bad) and 10(excellent) is based on sensory data and is termed as the output variable.

Some primary observations:

- More than 80% of the wine samples are graded average (i.e. rated 5 or 6).
- 132 wines are reported having no citric acid, at all.
- Only 1 wine is reported to have a maximum value of citric acid (i.e. 1), and graded 4 on quality (poor). This happens to be the only outlier, too.
- Many wine samples in our dataset have attribute values identical to 1 or more samples.
- No outliers are found in the log-transformed distribution of total.sulfur.dioxide.
- Almost 90% of the wines have less than 12% alcohol.

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

The main feature of interest is definitely ***quality***. I’d like to find out which attributes most likely affect the quality of red wine.

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

Based on my research, these are the features that seem to contribute to the quality of a red wine:

1. ***fixed.acidity*** - *imparts the sourness or tartness that is a fundamental feature in wine taste*
2. ***residual.sugar*** - *often indicates the level of dryness (an important feature in wine taste)*
3. ***alcohol*** - *higher alcohols, though, can have an aromatic effect, does not have many sensory effects in wines* (we’ll find out)
4. ***citric.acid*** - *can add ‘freshness’ and flavor to wines*
5. ***chlorides*** - *gives the wine its salty taste; higher concentration is undesirable, though*
6. ***gross.sulfur.dioxide*** - *(sum of free and total SO_2) at free sulfur dioxide (SO_2) concentrations over 50 ppm, SO_2 becomes evident in the nose and taste of wine*

Another relevant feature of my research has been the relationship of sweetness (*imparted due to residual.sugar*) with acidity and alcohol content. It seems sweetness and acidity have an inverse relationship. Whereas, sweetness has a direct relationship with the potential alcohol in the wine.

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

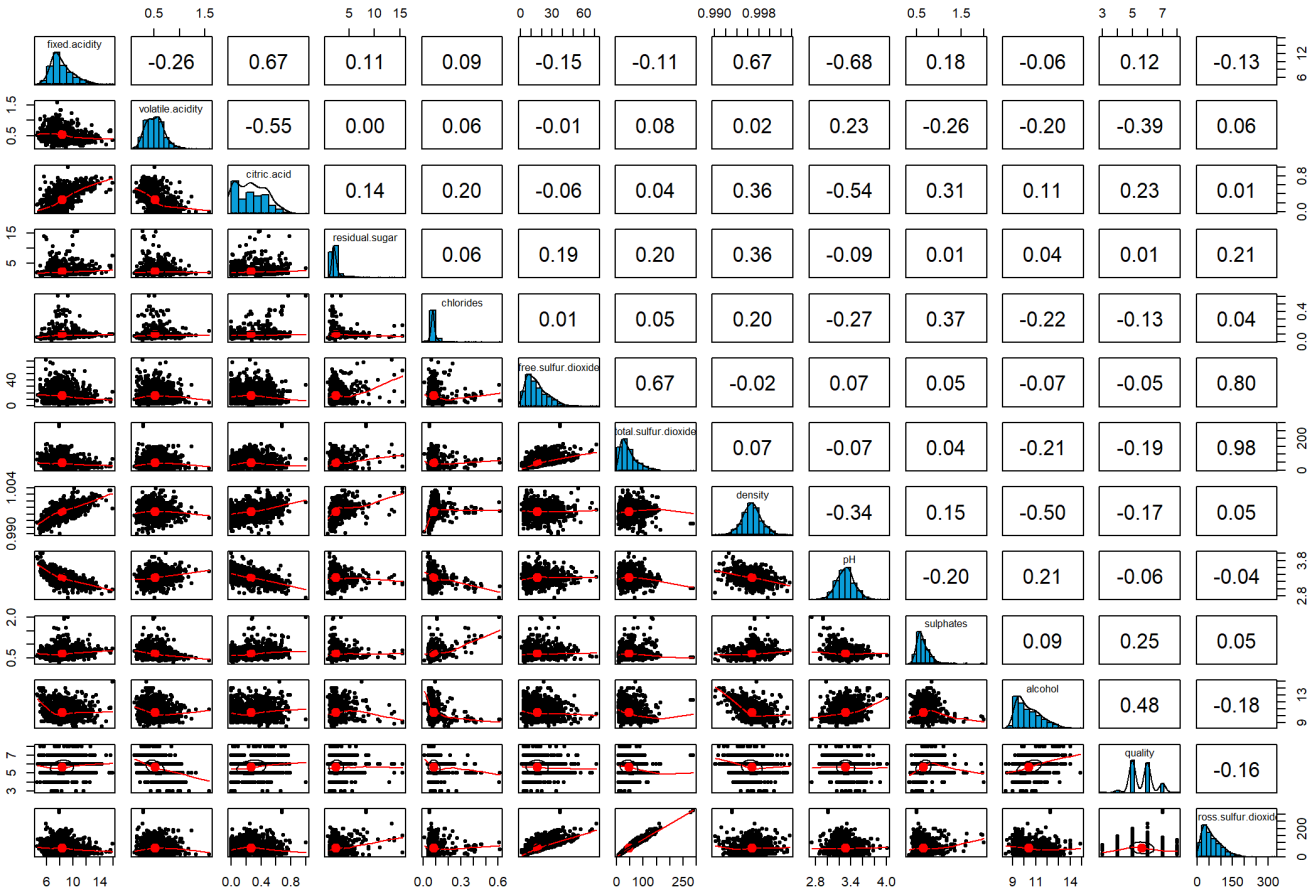
I created gross.sulfur.dioxide by summing up free.sulfur.dioxide and total.sulfur.dioxide.

Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

I log-transformed the right-skewed **residual.sugar**, **gross.sulfur.dioxide** and **sulphates** distributions. The transformed distributions are less skewed and appear somewhat normal.

The distribution of **chlorides** being extremely right-skewed and its boxplot revealing several outliers, I omitted the top 5% of its values (95th percentile truncated) and adjusted the binwidth. The transformed distribution appears normal.

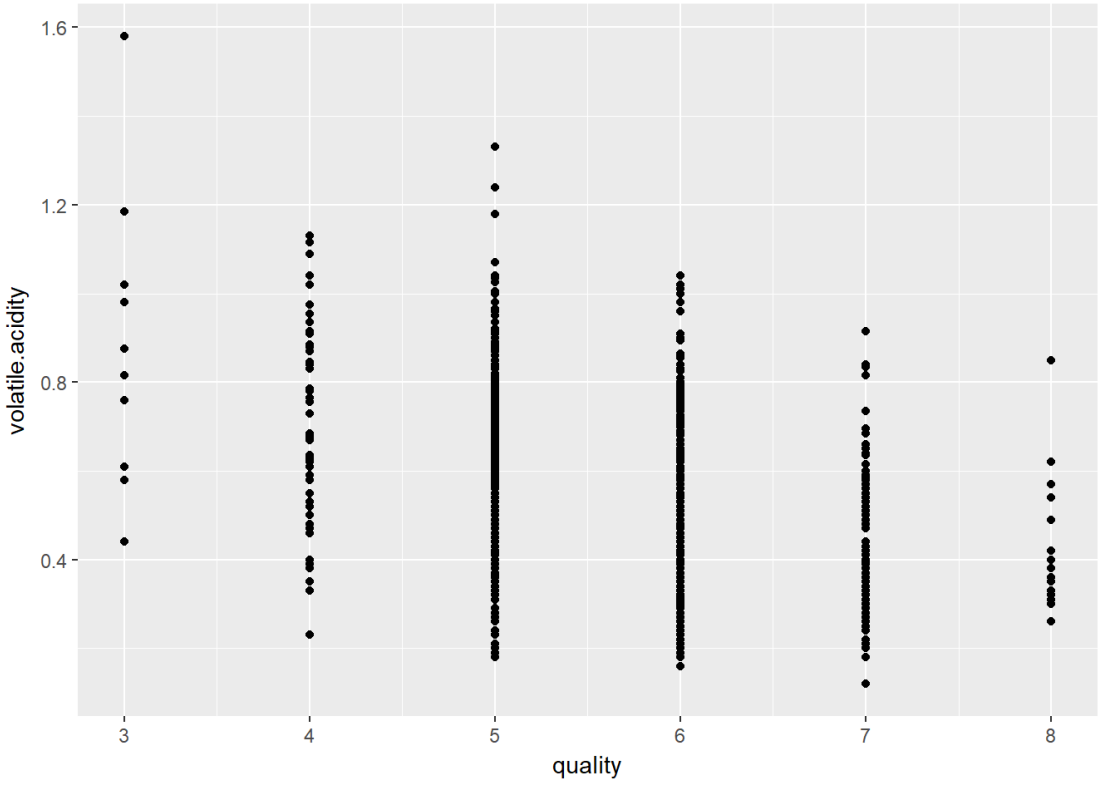
Bivariate Plots Section



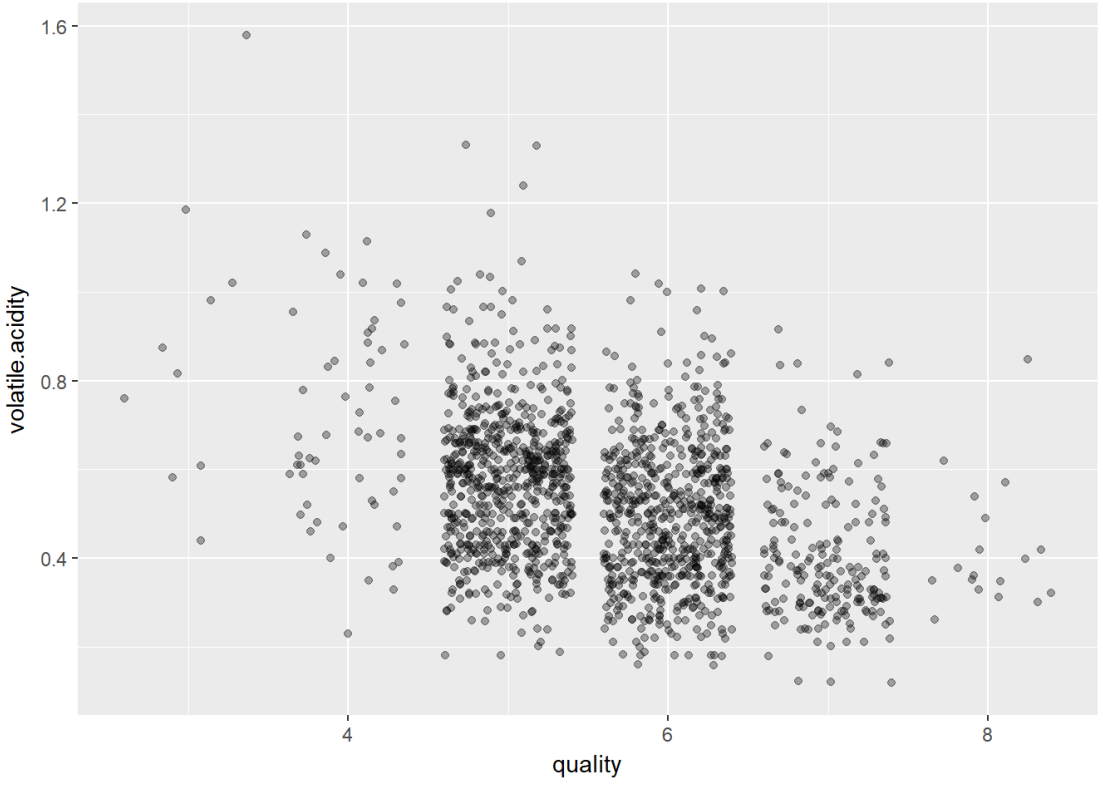
From the above scatterplot matrix, I can see that quality has a moderate positive correlation with alcohol (0.48) and a moderate negative correlation with volatile.acidity (-0.39). Surprisingly, the above matrix shows no considerable correlation between quality and fixed.acidity (0.12). A slightly higher correlation than this is reported by citric.acid (0.23) and surprisingly, sulphates (0.25). Let's not forget that I have log-transformed the sulphates distribution and that's what I need to plot against quality.

Coming to sulfur dioxides, quality has a better correlation value with total.sulfur.dioxide (-0.19) compared to my newly created variable, gross.sulfur.dioxide (-0.16). Contrary to what I have assumed, there is hardly any correlation between quality and residual.sugar (0.01). Similar is the case with chlorides (-0.13) but the -ve sign shows that it has an inverse relationship with quality. This, in a way, supports our assumption that higher concentration of chlorides (i.e. saltiness) is undesirable. As chlorides, too, is transformed (95th percentile truncated), I want to look closer at scatter plots involving quality and the transformed chlorides data.

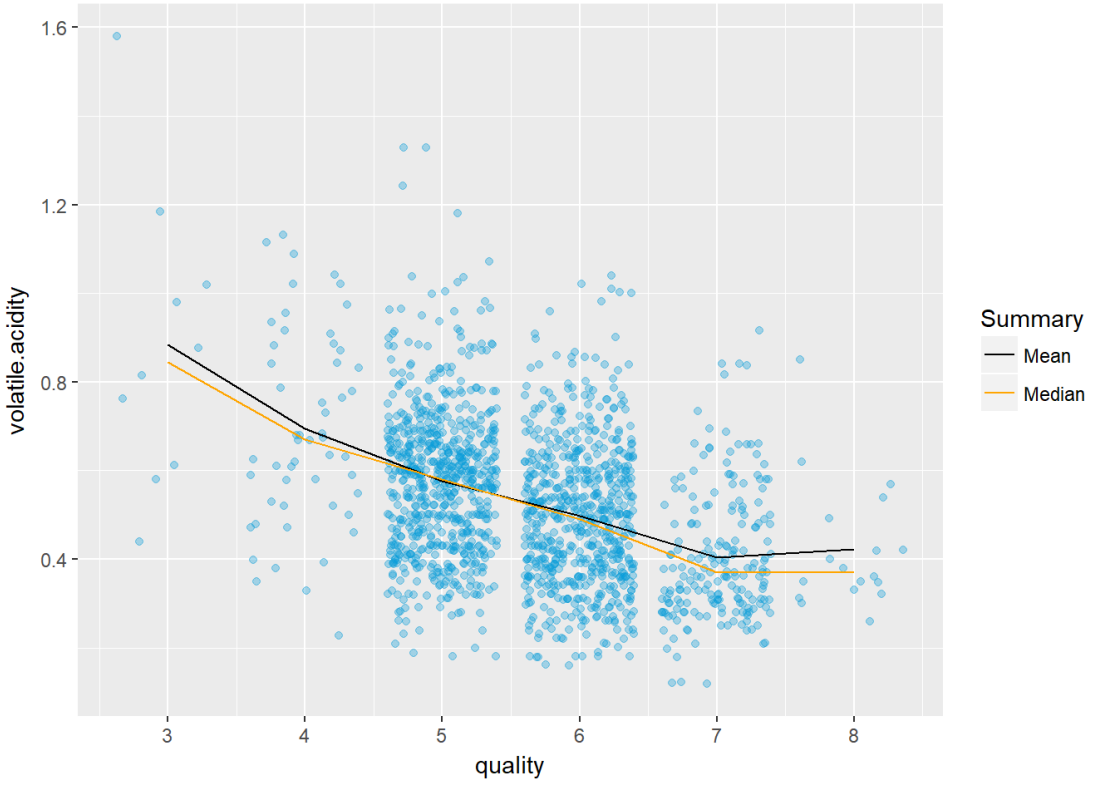
The other matters of interest are the significant correlation values of density with fixed.acidity (0.67) and alcohol (-0.50). Also, interestingly, pH is positively correlated with volatile.acidity. Whereas, it is inversely related to other types of acids, such as fixed and citric (*as it should be*).



The basic scatterplot between quality and volatile.acidity isn't as dispersed as I want it to be. Let's add some jitter and set the transparency level at alpha=1/3.



Now that I have my scatter plot, I want to overlay the summary of mean and median values of volatile.acidity.



The above scatter plot reveals the decrease in volatile.acidity for wines with higher quality and a slight increase in mean volatile.acidity right after quality=7. However, the median volatile.acidity remains constant for the highest quality wines of my dataset.

Looking at mean and median summaries, I can see that the best quality wines have a volatile.acidity around 0.4.

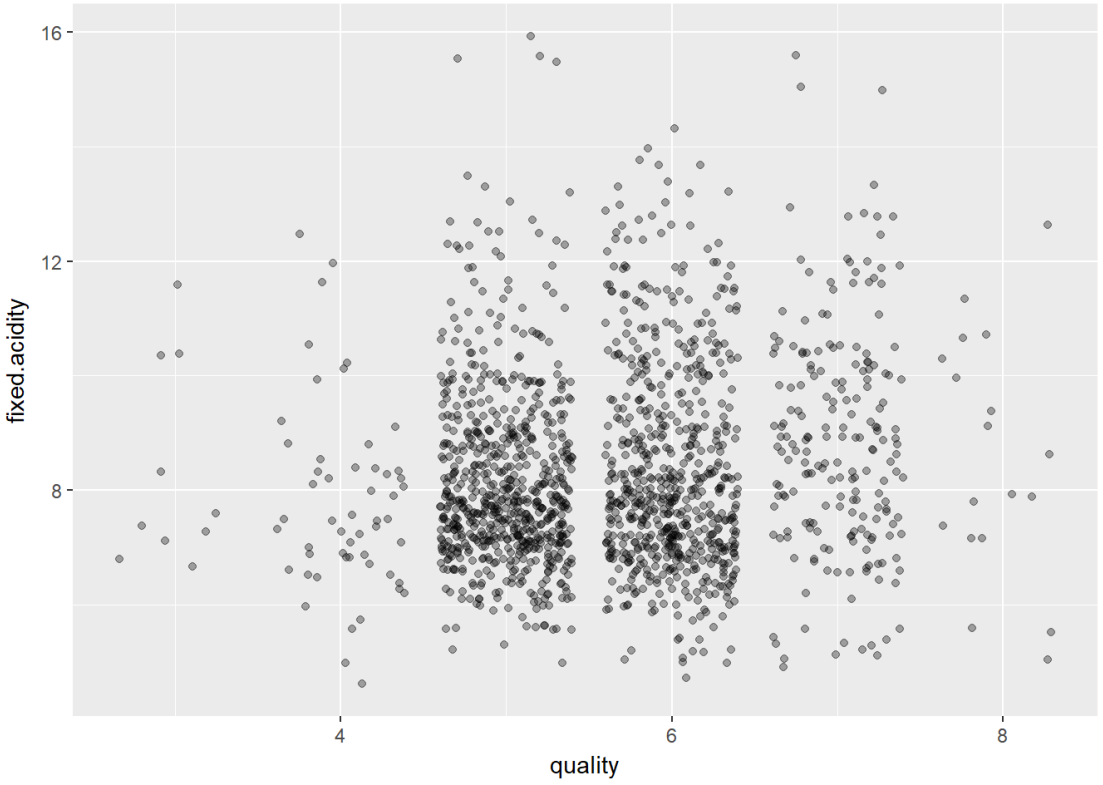
In order to take a closer look at the count of wines and how the average volatile.acidity exactly vary over quality, I have come up with the following table for each wine quality.

```
## # A tibble: 6 × 4
##   quality   va_mean va_median     n
##     <int>     <dbl>     <dbl> <int>
## 1       3 0.8845000     0.845    10
## 2       4 0.6939623     0.670    53
## 3       5 0.5770411     0.580   681
## 4       6 0.4974843     0.490   638
## 5       7 0.4039196     0.370   199
## 6       8 0.4233333     0.370    18
```

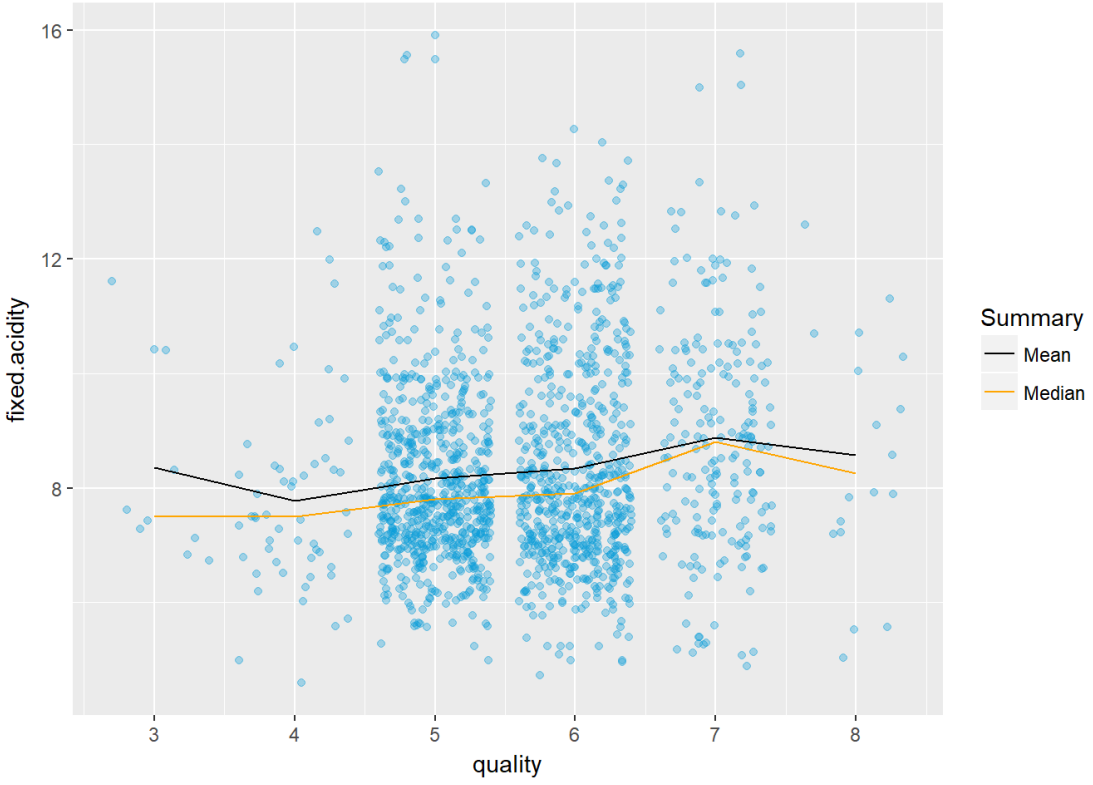
The mean volatile.acidity keeps on decreasing with increasing quality and the wines with a mean range of volatile.acidity between 0.4 - 0.42 are graded the best on quality scale. However, the best quality wines (7 and 8) have a median volatile.acidity of 0.37.

Let's now take a look at the other acidity features (fixed.acidity and citric.acid). I will start with the scatterplots between quality and these 2 features.

I will add some jitter to the fixed.acidity plot and set the transparency level at alpha=1/3.

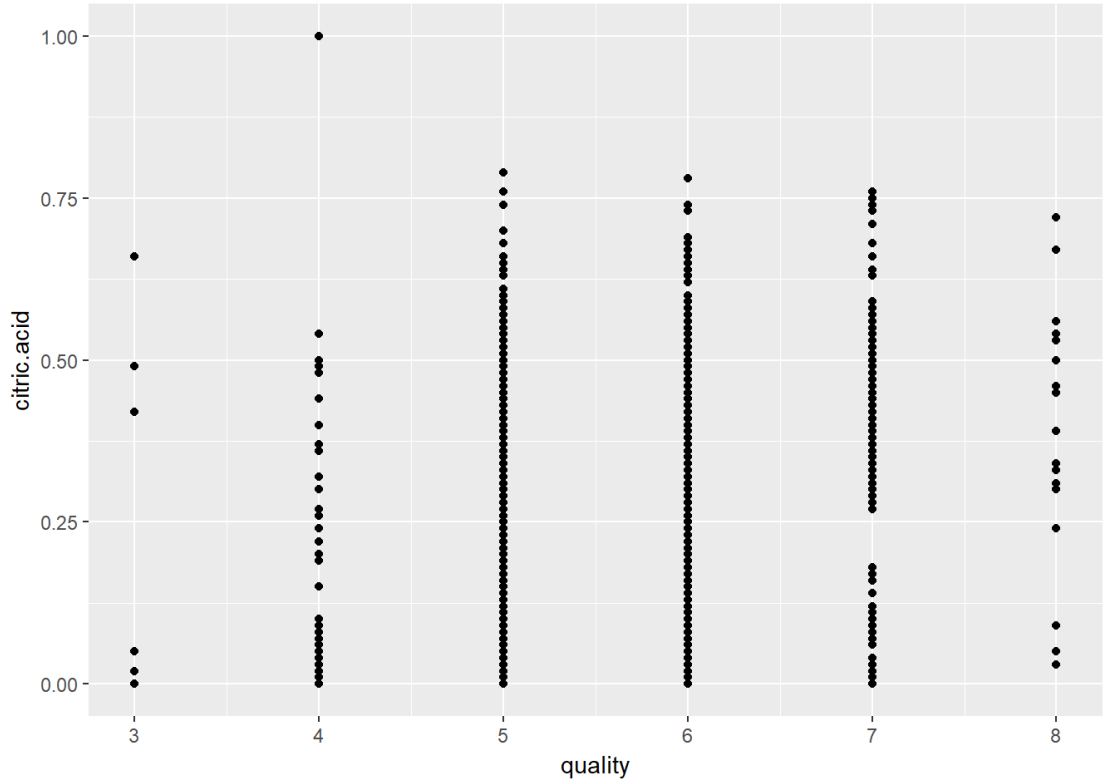


So, wines having a fixed.acidity more than 12 are mostly graded between 4 and 7. On the other hand, almost all the highest-graded wines have a fixed.acidity less than 11.

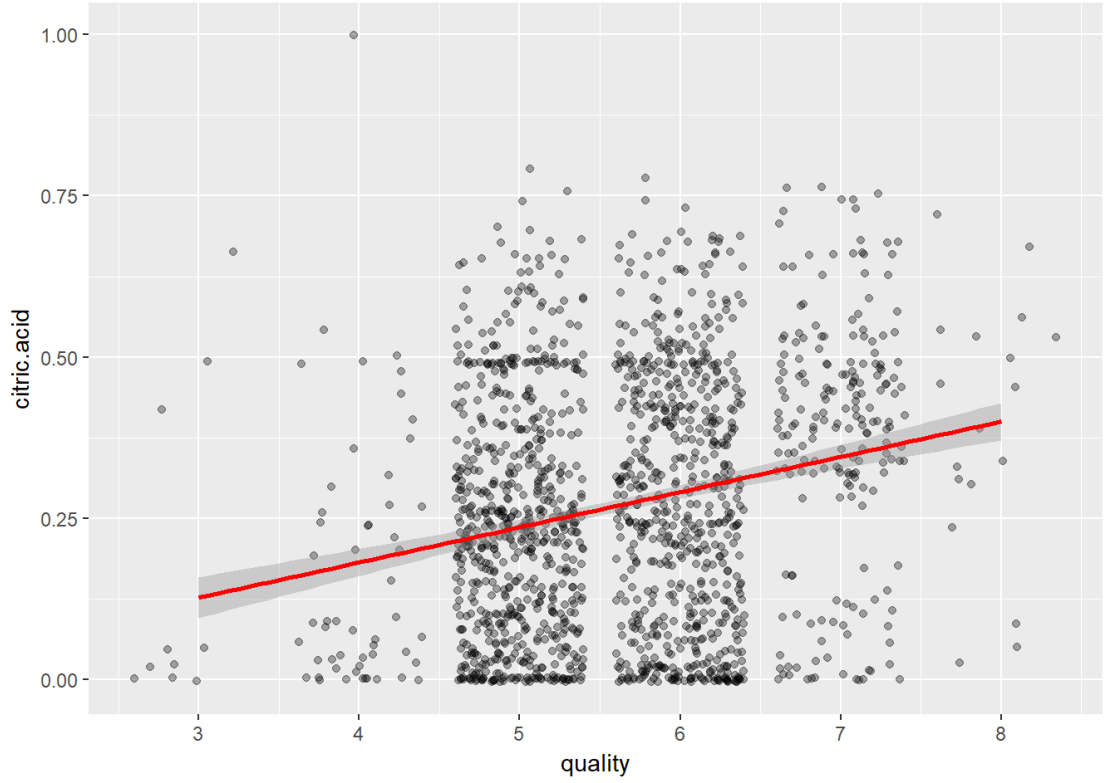



```
## # A tibble: 6 × 4
##   quality  fa_mean fa_median     n
##     <int>    <dbl>     <dbl> <int>
## 1       3 8.360000      7.50    10
## 2       4 7.779245      7.50    53
## 3       5 8.167254      7.80   681
## 4       6 8.347179      7.90   638
## 5       7 8.872362      8.80   199
## 6       8 8.566667      8.25    18
```

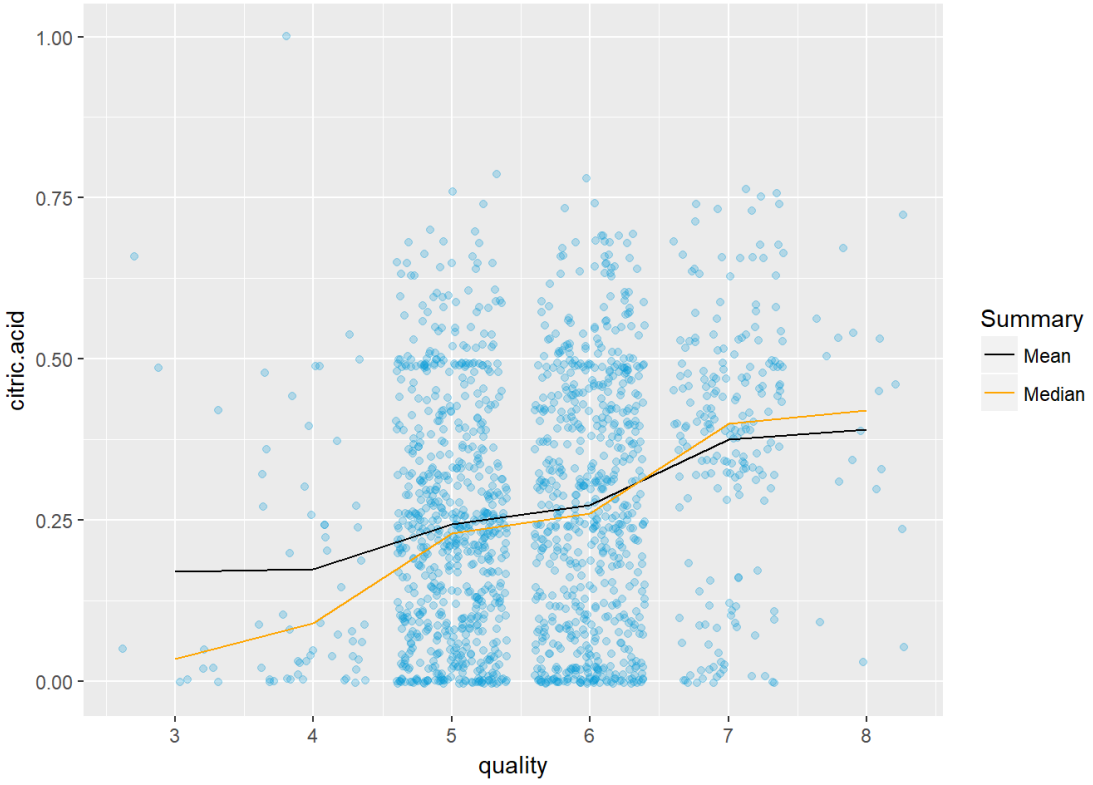
There isn't really much of a correlation between fixed.acidity and quality. However, a fixed.acidity slightly greater than 8 seems to be essential for a good quality wine.



The basic scatterplot between quality and citric.acid isn't as dispersed as we want it to be. One interesting point is that though there are many wines in our dataset that do not contain any citric.acid, none of them is graded 8 (highest) on quality.
I will now add some jitter to fix overplotting and set the transparency level at alpha=1/3. I will also fit a linear model to check the linear relationship between citric.acid and quality.



So, citric.acid and quality has a weak linear relationship where citric.acid increases with quality. I will analyse it further by overlaying the summary of mean and median values of citric.acid.

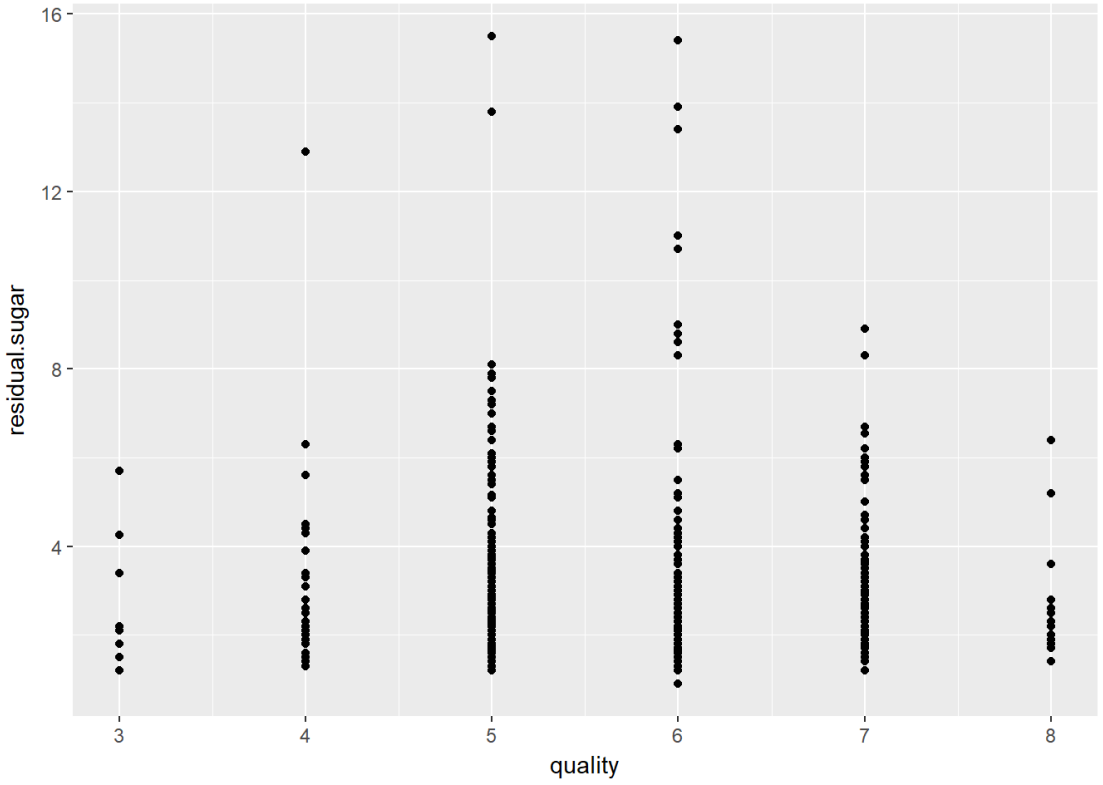



```
##
##  Pearson's product-moment correlation
##
## data:  quality and citric.acid
## t = 9.2875, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1793415 0.2723711
## sample estimates:
##       cor
## 0.2263725
```

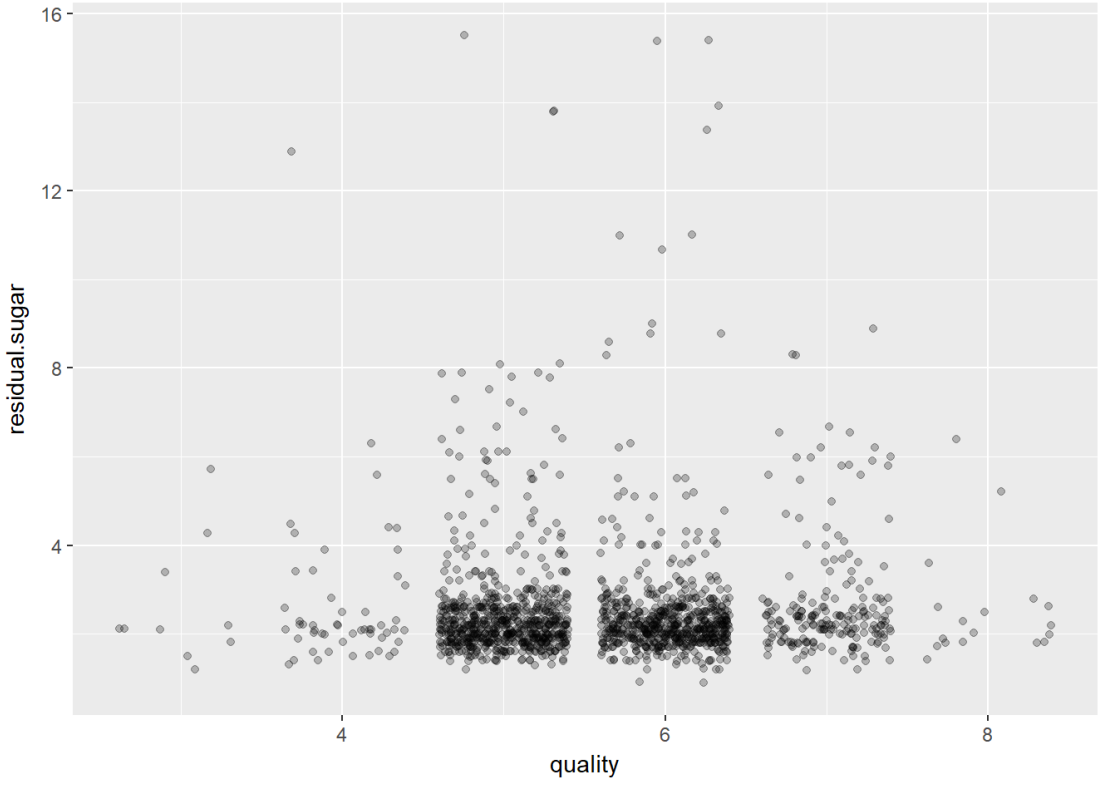
A positive correlation of 0.22 between citric.acid and quality and the mean and median summaries peaking at the highest value of quality shows that citric.acid can improve the wine's quality. This makes sense as citric.acid is known to add freshness and flavor to wines.
I would now like to take a closer look at the exact mean and median values for different quality wines.

```
## # A tibble: 6 × 4
##   quality   ca_mean ca_median     n
##     <int>     <dbl>     <dbl> <int>
## 1       3 0.1710000     0.035    10
## 2       4 0.1741509     0.090    53
## 3       5 0.2436858     0.230   681
## 4       6 0.2738245     0.260   638
## 5       7 0.3751759     0.400   199
## 6       8 0.3911111     0.420    18
```

The above table supports the positive direct relationship between citric.acid and quality. A citric.acid content of 0.4 g/dm^3 seems to be ideal for a higher quality wine.

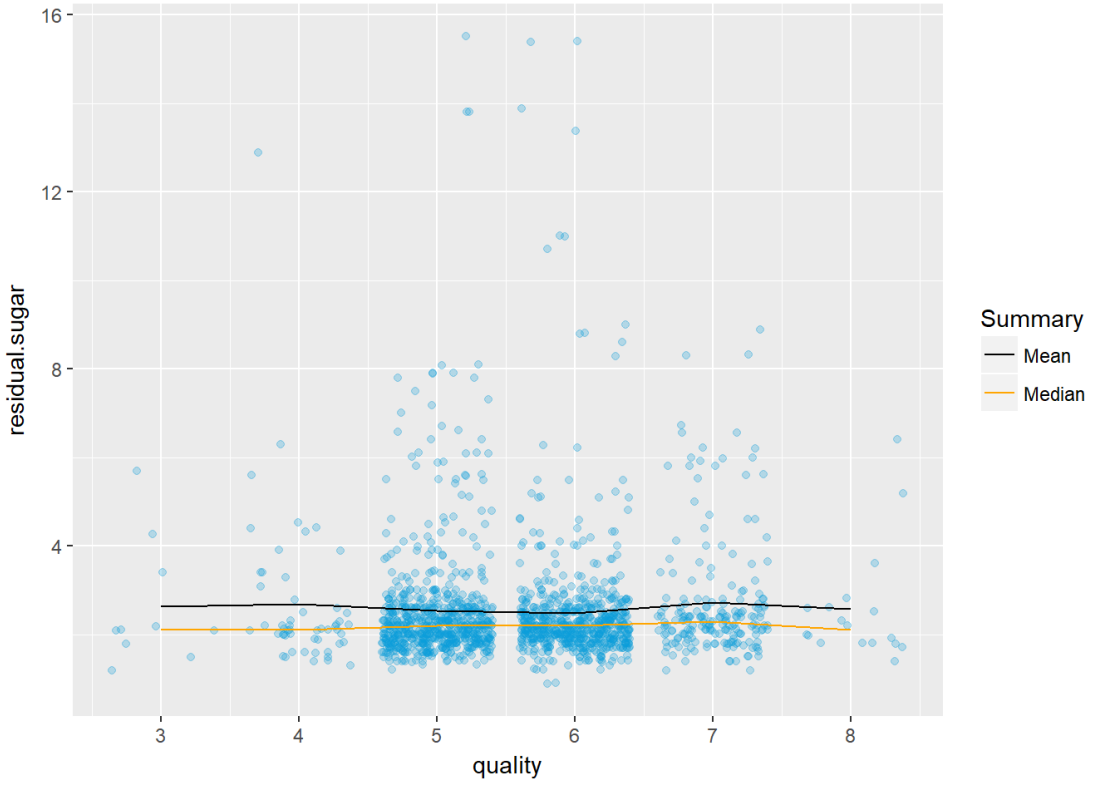


The basic scatterplot between residual.sugar and quality reveals that majority of the wines graded 6 to 8 have residual.sugar less than 7. To get a better understanding of their relationship, I will plot it again with some jitter and set the transparency level at alpha=1/4.

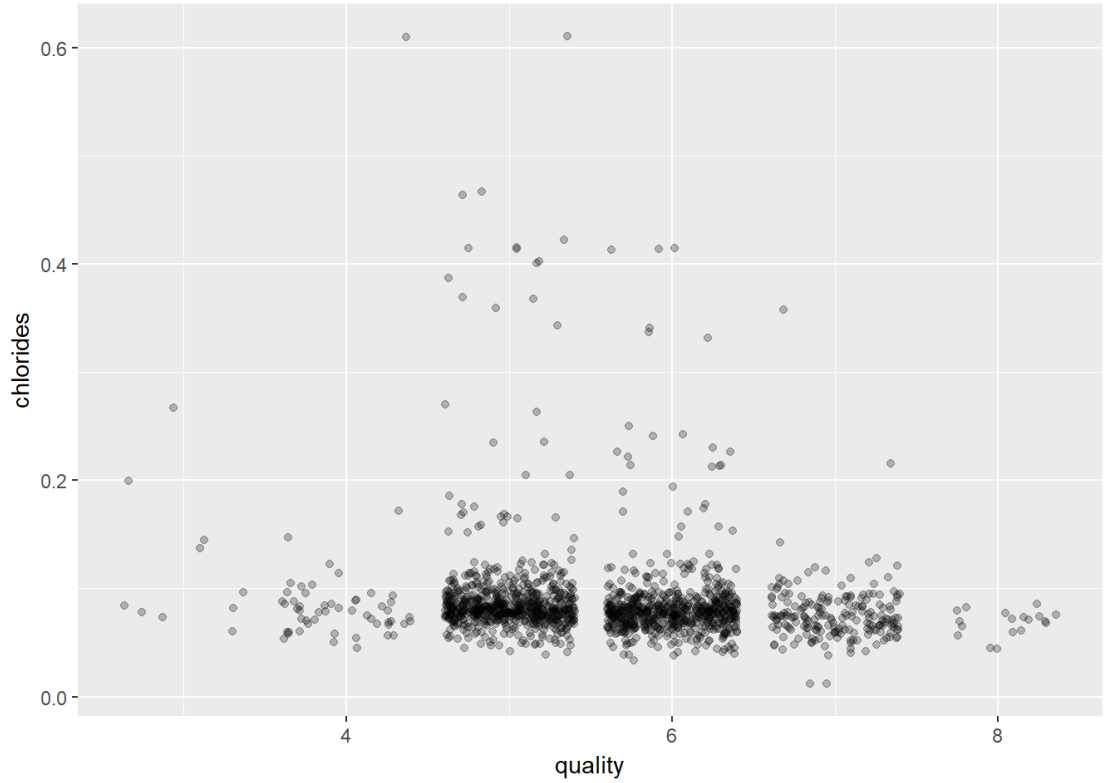


The circles are heavily concentrated towards the base of the plot and the darker ones are hovering around residual.sugar = 2. In fact, anything above the value of 8 for residual.sugar seems to be mostly outliers. There seems to be hardly any correlation between the variables.

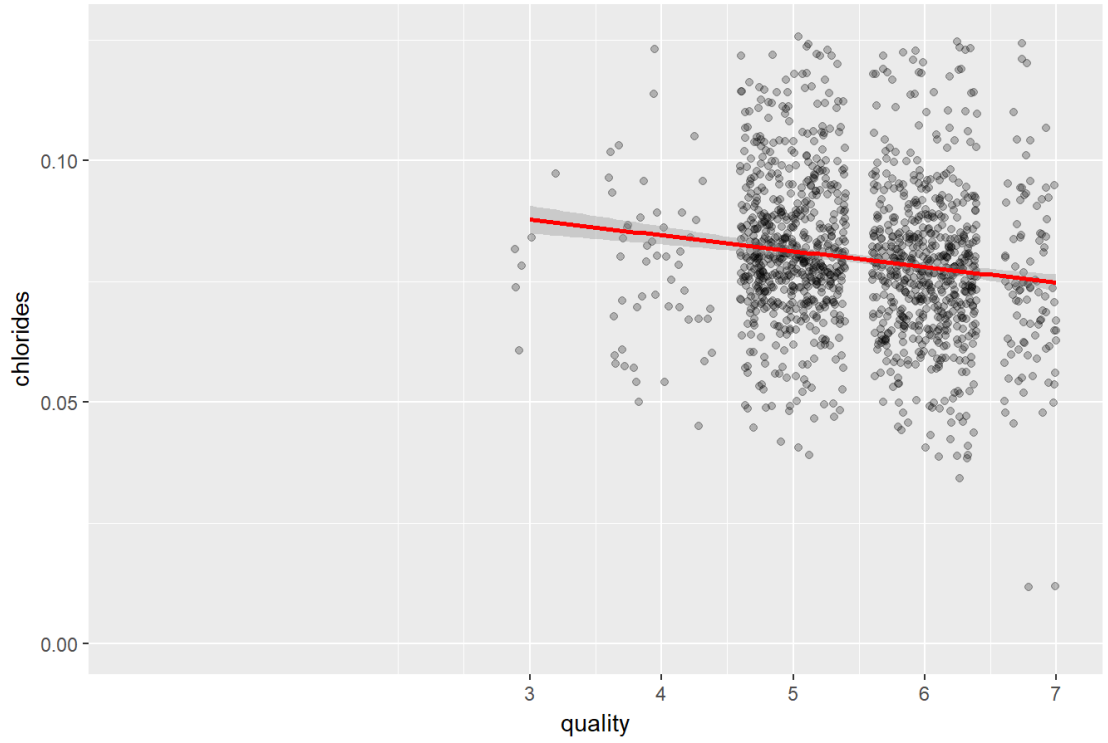
But, let's not forget that residual.sugar imparts a certain sweetness to red wine which we had earlier assumed to impact its quality.



The mean and median values of residual.sugar fluctuate and there is no way we can establish a pattern here. Based on the above plots, residual.sugar seems to have no considerable impact on quality.



Quality vs 95% quartile Chlorides



The first plot is a scatterplot between chlorides and quality where I have added some jitter and set the transparency level at alpha=1/4. Though there are some medium-quality wines (i.e. quality = 5 or 6) with higher levels of chlorides, all the higher quality wines have very low chlorides.

The second plot is created by removing the top 5% of both the variables as I have normalised chlorides data by omitting its top 5% during univariate analysis. I have smoothened the relationship and plotted a linear model, which shows an inverse relationship between these 2 variables.

```
## # A tibble: 6 × 4
##   quality    cl_mean cl_median     n
##     <int>      <dbl>     <dbl> <int>
## 1       3 0.12250000    0.0905    10
## 2       4 0.09067925    0.0800    53
## 3       5 0.09273568    0.0810   681
## 4       6 0.08495611    0.0780   638
## 5       7 0.07658794    0.0730   199
## 6       8 0.06844444    0.0705    18
```

```
##
##  Pearson's product-moment correlation
##
## data:  quality and chlorides
## t = -5.1948, df = 1597, p-value = 2.313e-07
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.17681041 -0.08039344
## sample estimates:
##        cor
## -0.1289066
```

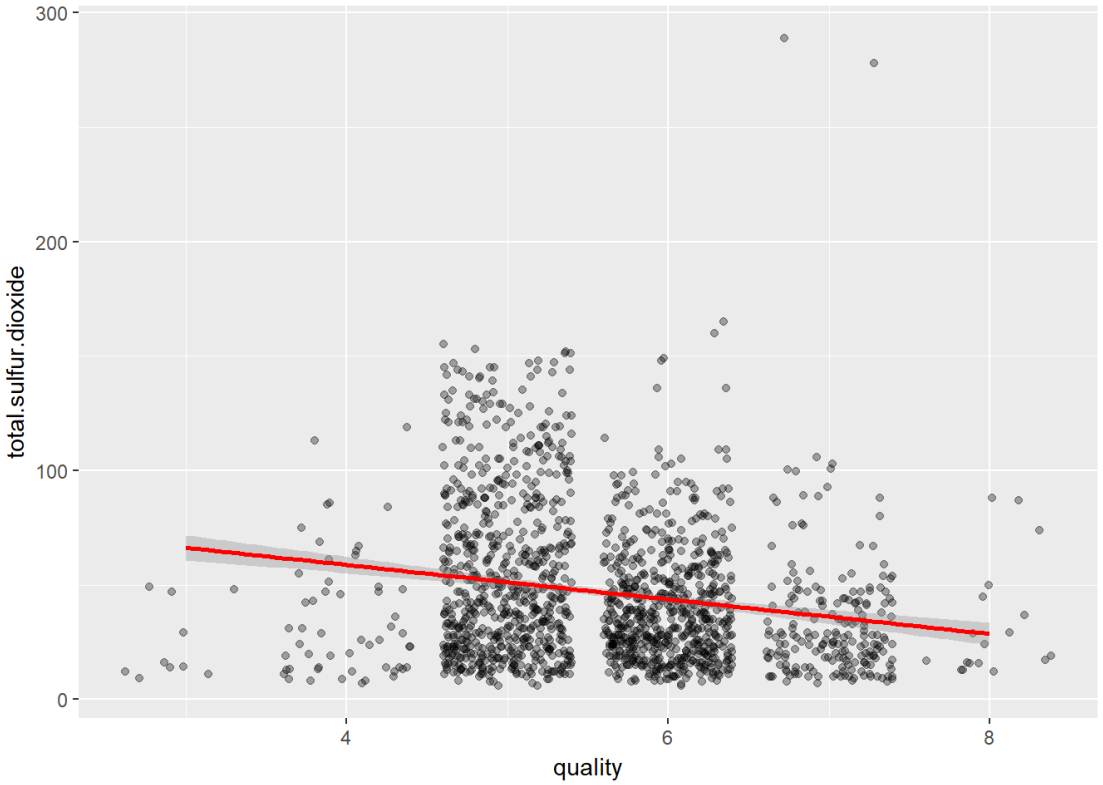
The above table summarises the mean and median values of chlorides by quality. The decreasing chloride values with increasing quality shows some negative correlation. The Pearson's correlation coefficient is -0.129 which is very nominal but nevertheless, supports the assumption that higher amounts of salt imparted by chlorides is undesirable in wine.

```
##
##  Pearson's product-moment correlation
##
## data:  quality and total.sulfur.dioxide
## t = -7.5271, df = 1597, p-value = 8.622e-14
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2320162 -0.1373252
## sample estimates:
##        cor
## -0.1851003
```

```
##
##  Pearson's product-moment correlation
##
## data:  quality and gross.sulfur.dioxide
## t = -6.5978, df = 1597, p-value = 5.659e-11
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2102371 -0.1147902
## sample estimates:
##        cor
## -0.1628947
```

```
##
##  Pearson's product-moment correlation
##
## data:  quality and log10(gross.sulfur.dioxide)
## t = -5.8727, df = 1597, p-value = 5.205e-09
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.19303982 -0.09706524
## sample estimates:
##        cor
## -0.1453946
```

The log-transformation of gross.sulfur.dioxide shows a decrease in its correlation with quality. Quality continues to have a better correlation with total.sulfur.dioxide and that's I want to explore further.



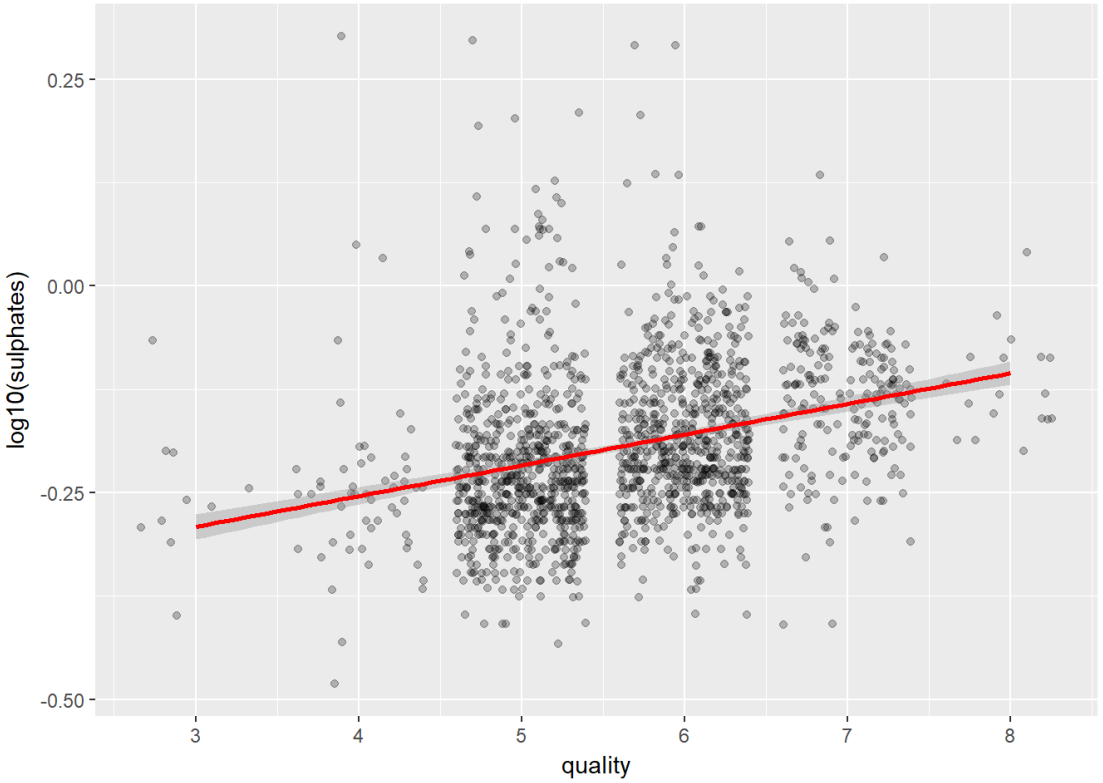

```
## # A tibble: 6 × 4
##   quality so2_mean so2_median     n
##     <int>    <dbl>      <dbl> <int>
## 1       3 24.90000       15.0    10
## 2       4 36.24528       26.0    53
## 3       5 56.51395       47.0   681
## 4       6 40.86991       35.0   638
## 5       7 35.02010       27.0   199
## 6       8 33.44444       21.5    18
```

The mean and median total.sulfur.dioxide increases till quality=5 and then shows a downward trend as quality increases. Hence, the negative correlation. The scatterplot, to a certain extent, shows that the best quality(=8) red wines have less total.sulfur.dioxide compared to the next best(=7).

```
##
##  Pearson's product-moment correlation
##
## data:  quality and sulphates
## t = 10.38, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2049011 0.2967610
## sample estimates:
##       cor
## 0.2513971
```

```
##
##  Pearson's product-moment correlation
##
## data:  quality and log10(sulphates)
## t = 12.967, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2636092 0.3523323
## sample estimates:
##       cor
## 0.3086419
```

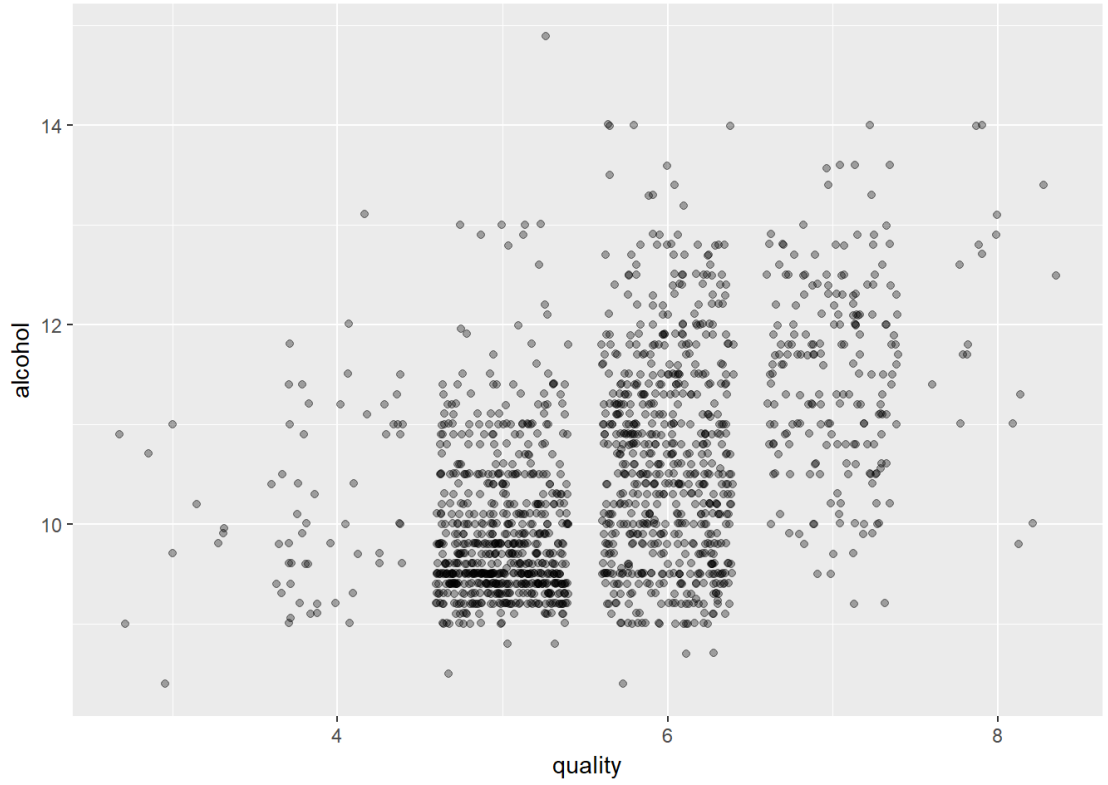
For sulphates, I had log-transformed its values to create a normalised distribution during univariate analysis. The Pearsons correlation is more between quality and log-transformed sulphates which is 0.31, as compared to 0.25 (quality and sulphates).



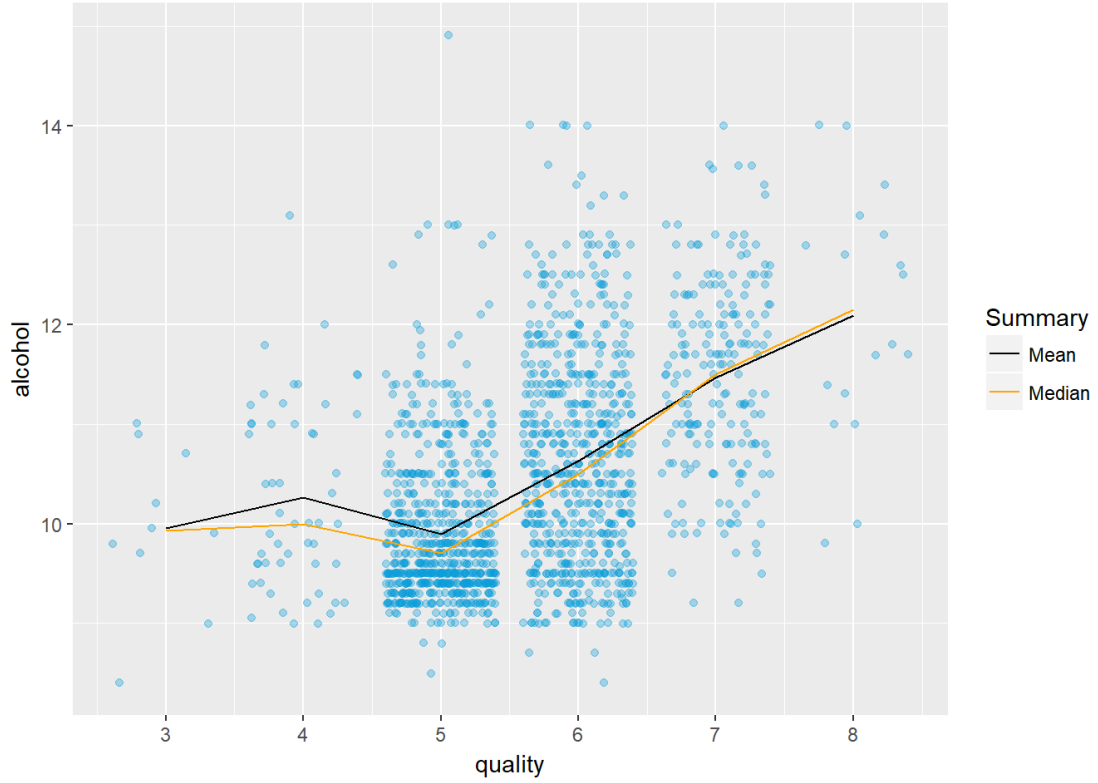
The above scatterplot between quality and log-transformed sulphates (smoothened using linear method) shows some correlation between sulphates and quality. In fact, the best red wines have higher sulphate levels compared to the mediocre and poor quality wines.

```
## # A tibble: 6 × 4
##   quality    s_mean s_median     n
##     <int>     <dbl>    <dbl> <int>
## 1       3 0.5700000    0.545    10
## 2       4 0.5964151    0.560    53
## 3       5 0.6209692    0.580   681
## 4       6 0.6753292    0.640   638
## 5       7 0.7412563    0.740   199
## 6       8 0.7677778    0.740    18
```

The above summary table (mean and median values of sulphates against quality) shows that sulphate levels increase with increasing quality and the median value of 0.74 seems to be optimum for the highest quality wines in our dataset.



Based on the above scatterplot (alpha=1/3) between quality and alcohol content, I can say that the highest quality wines of our dataset have an alcohol content of more than 10%. Wines with higher alcohol levels, like greater than 13%, are mostly rated 6 or above.

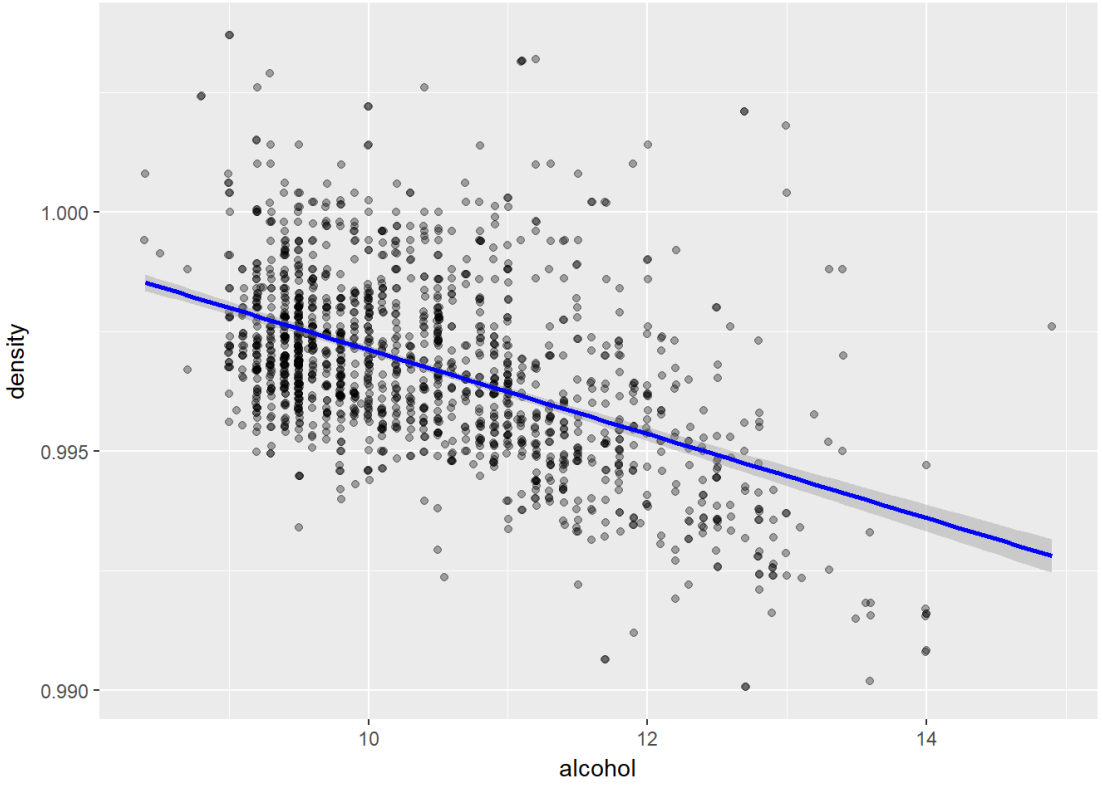


From the above plot, I can see that the mean and median values of alcohol just shoots up from levels below 10% to more than 12% as quality improves from 5 to 8. Let's find the Pearson's correlation coefficient between these two variables and take a closer look at the data on which the above plot is based.

```
##
##  Pearson's product-moment correlation
##
## data:  quality and alcohol
## t = 21.639, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4373540 0.5132081
## sample estimates:
##       cor
## 0.4761663
```

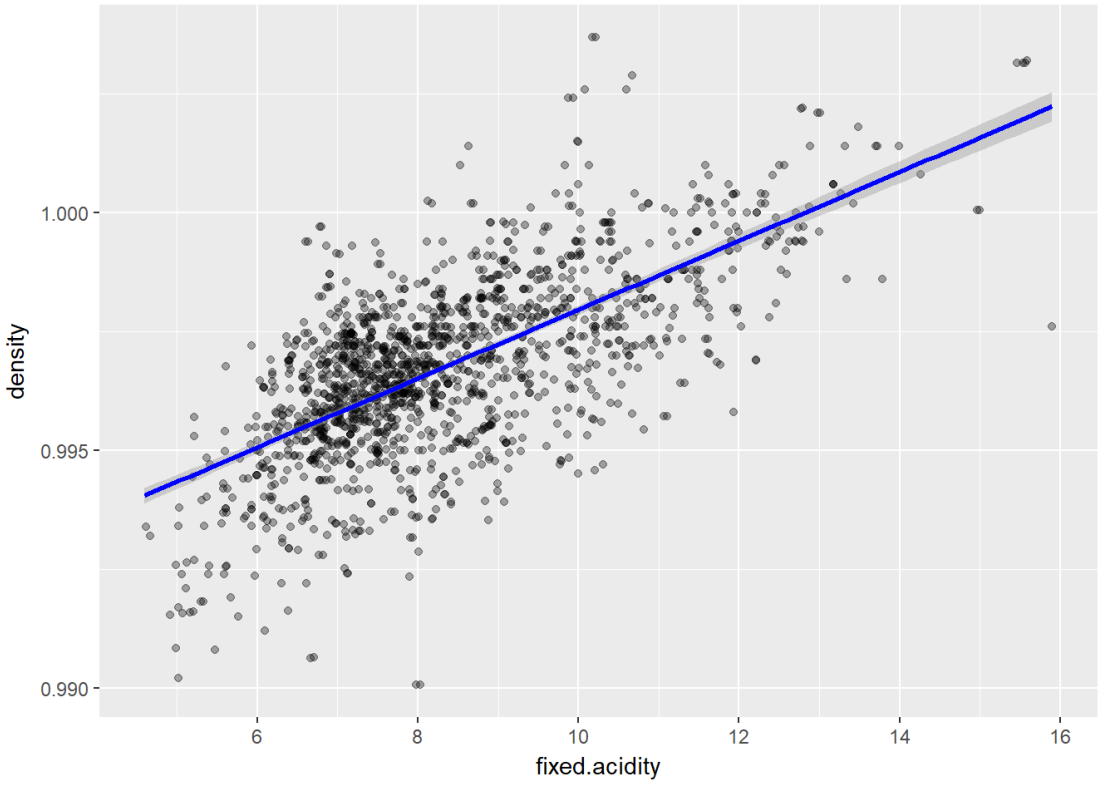
```
## # A tibble: 6 × 4
##   quality    a_mean a_median     n
##     <int>     <dbl>    <dbl> <int>
## 1       3  9.955000    9.925    10
## 2       4 10.265094   10.000    53
## 3       5  9.899706    9.700   681
## 4       6 10.629519   10.500   638
## 5       7 11.465913   11.500   199
## 6       8 12.094444   12.150    18
```

A correlation of 0.47 is moderately high and the highest among all the input variables of our dataset. In a way, alcohol has probably the highest impact on quality of red wines. Now when I look at the summary table of mean and median values of alcohol, I can see that wine experts would probably prefer red wines with alcohol content above 11%, to say the least.




```
##
##  Pearson's product-moment correlation
##
## data:  alcohol and density
## t = -22.838, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5322547 -0.4583061
## sample estimates:
##       cor
## -0.4961798
```

Alcohol content has an inverse relationship with density (correlation ~ -0.5). Some of the least dense wines (density <= 0.9925) have greater than 11% alcohol content, which is greater than both the mean and median values of alcohol. The relationship looks moderately linear.




```
##
##  Pearson's product-moment correlation
##
## data:  fixed.acidity and density
## t = 35.877, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6399847 0.6943302
## sample estimates:
##       cor
## 0.6680473
```

Summary of fixed.acidity:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.60    7.10    7.90    8.32    9.20   15.90
```

Fixed acidity has a high correlation with density (correlation ~ 0.67). In fact, for fixed.acidity between 6 and 10, the linear smooth seems to be a very good fit.

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?

Quality of red wines is moderately correlated with alcohol and volatile.acidity.

As alcohol content increases, quality of a red wine increases. Based on the scatterplot between alcohol and quality, I found that the mean and median values of alcohol just shoots up from levels below 10% to more than 12% as quality improves from 5 to 8. Based on the summary table of mean and median values of alcohol, I can see that wine experts preferred red wines with alcohol content above 11%.

On the other hand, volatile.acidity has an inverse relationship with quality. As volatile acidity decreases, quality of red wine increases. However, there is a slight increase in mean volatile.acidity right after quality=7 (8 being the highest in this dataset). However, the median volatile.acidity remains constant (0.37) for the highest quality wines.

Contrary to what I had earlier assumed, there isn't much of a correlation between fixed.acidity and quality. However, citric.acid (*known to add 'freshness' and flavor to wines*) showed some correlation with quality. In fact, its mean and median summaries peaked for the best red wines. A citric.acid content of 0.4 g/dm^3 seems to be ideal for a good red wine.

On analysing the impact of chlorides on quality, I didn't see much of a correlation, except for the fact that better quality red wines have low chloride levels. This, in a way, supports my assumption that higher amounts of salt imparted by chlorides is undesirable in wine. The relationship between total.sulfur.dioxide and quality is similar to chlorides.

Sulphates data, after being log-transformed, recorded a slight increase in its correlation with quality. Based on my analysis, sulphate levels increase with increasing quality and its median value of 0.74 seems to be optimum for the best red wines.

Very surprisingly, I did not find any impact of residual.sugar on quality. Most wines, irrespective of quality, have lower levels of residual.sugar. The higher levels, which is greater than 8 are mostly outliers.

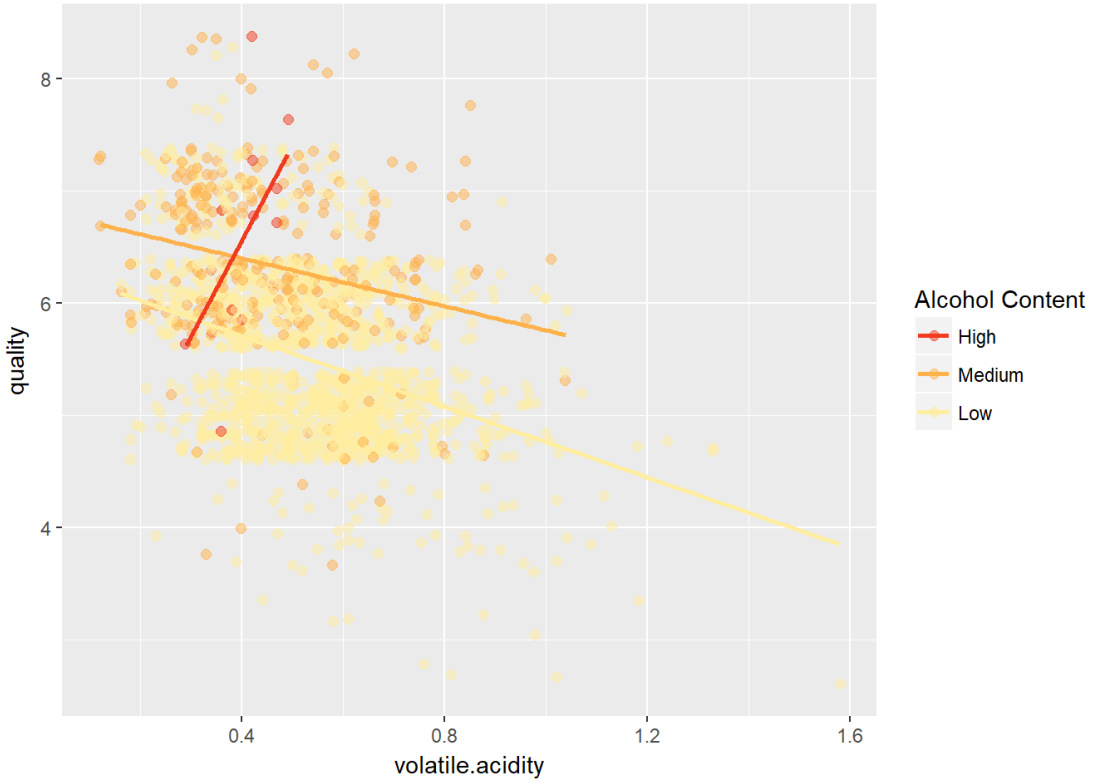
Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

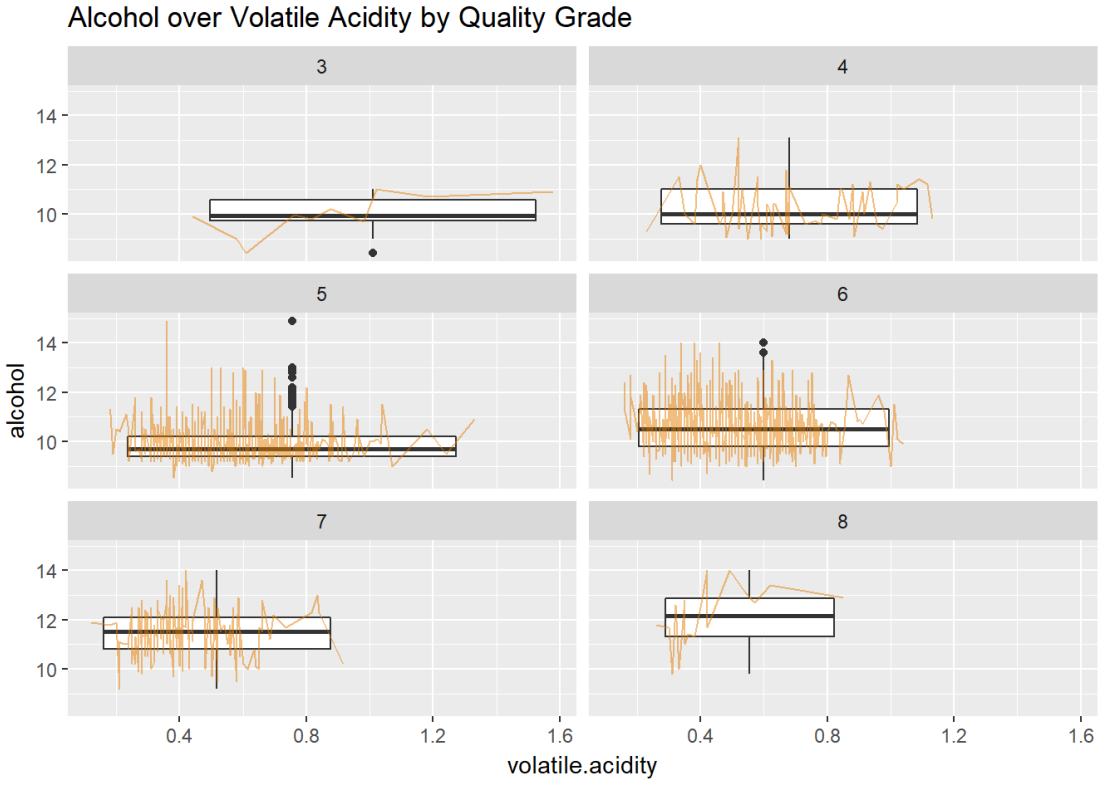
Alcohol content has an inverse relationship with density (correlation ~ -0.5). Whereas, fixed.acidity has a high correlation with density (correlation ~ 0.67).

What was the strongest relationship you found?

Among all the features, alcohol seems to have the strongest impact on the quality of a red wine. It has a positive and a moderately high correlation with quality. Sulphates and citric acid are the next in line to be positively correlated with quality, but less as compared to alcohol. On the other hand, quality is negatively and moderately correlated with volatile.acidity. Chlorides, too, has a negative, but a weak correlation with quality.

Multivariate Plots Section

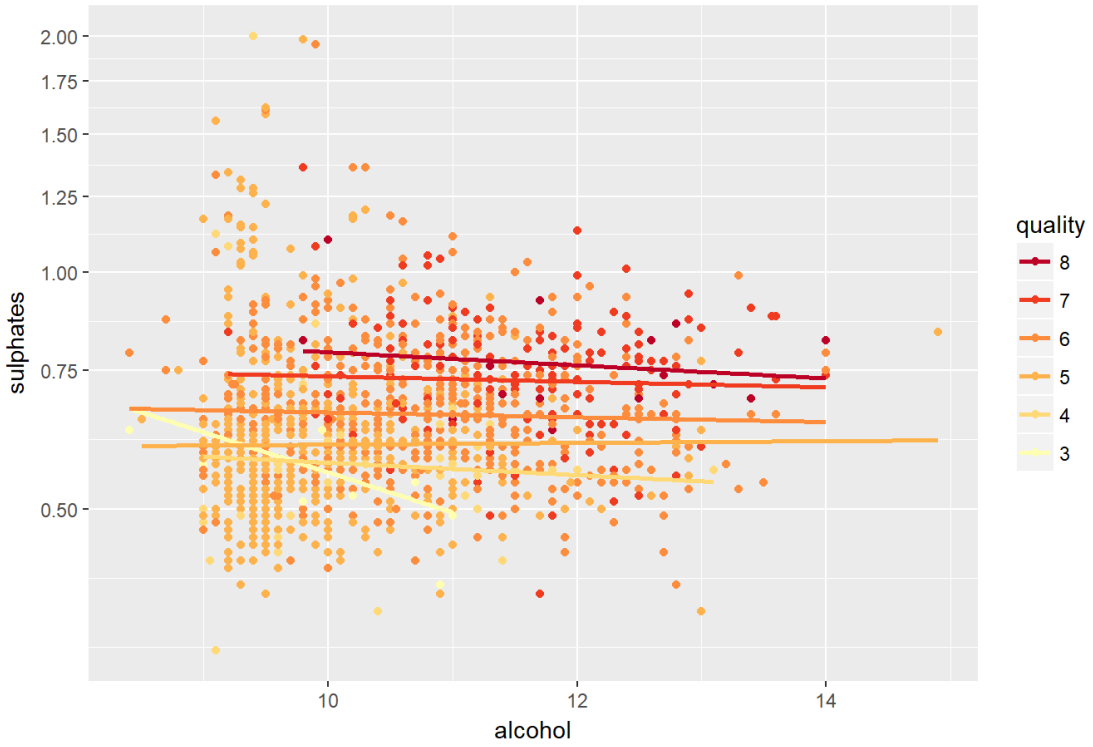




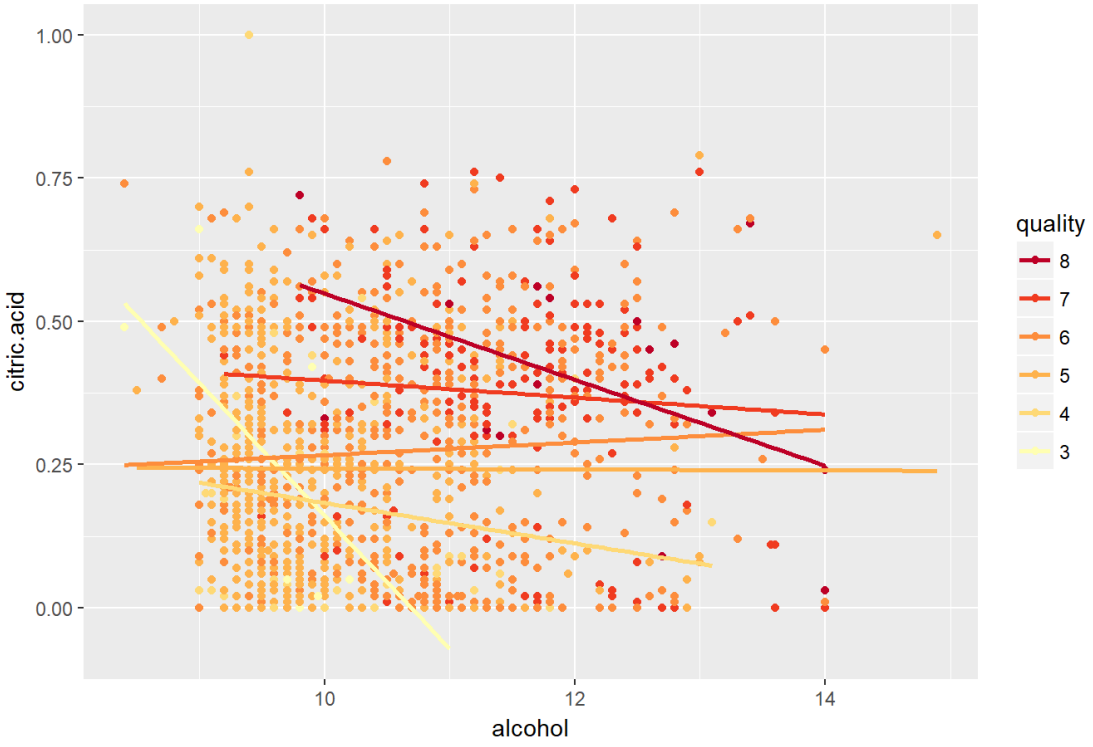
Based on my research about the categorisation of alcohol content in red wines, I divided this dataset into Low (less than 11.5%), Medium (11.5% to less than 13.5%) and High (anything equal to or more than 13.5%) alcohol-content wines. The first scatter plot has been smoothened and it shows that red wines, low on alcohol content, keeps dropping in quality with increasing volatile.acidity. Whereas, wines with high alcohol-content are mostly rated 6 or above on quality and their volatile.acidity never crosss the 0.5 mark. Wines in the medium alcohol-content category initially shows an increase in quality with increased volatile.acidity but soon shows a decline in quality which eventually improves at around volatile.acidity = 0.8. Looking at the orange dots on the scatterplot, I would say that quality of “medium” alcohol-content wines with a volatile.acidity between 0.4 and 0.6 seems to be graded extremely well.

The second plot shows boxplots and line plots of alcohol over volatile.acidity faceted over each quality grade. Some of the lowest quality(=3) wines have the lowest levels of alcohol and highest levels of volatile.acidity. Most of the red wines graded highest (7 and 8) on quality have a median alcohol of around 12% and a volatile.acidity of around 0.5. Another thing I notice for these wines is that there are no outliers and therefore their patterns and trends tend to be less affected compared to the other quality wines.

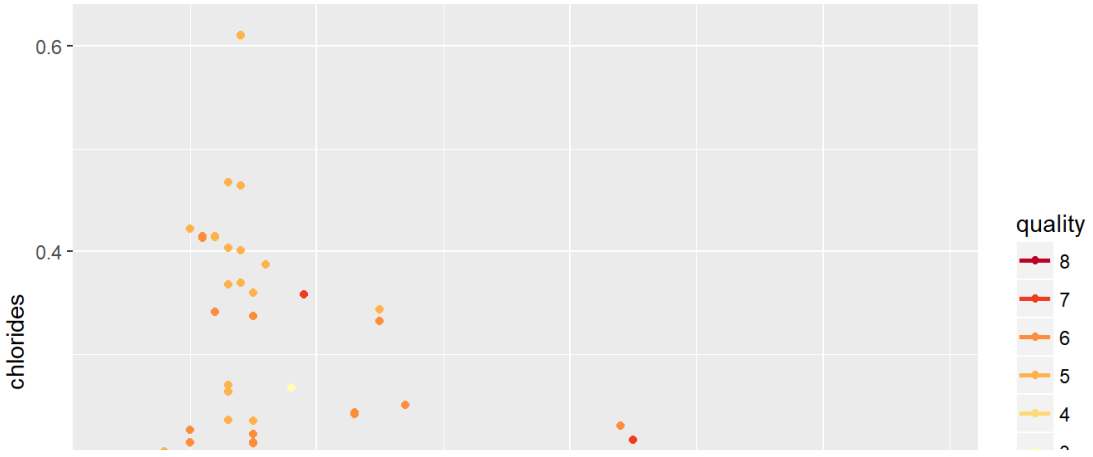
Quality by Alcohol and Sulphates (log_10)

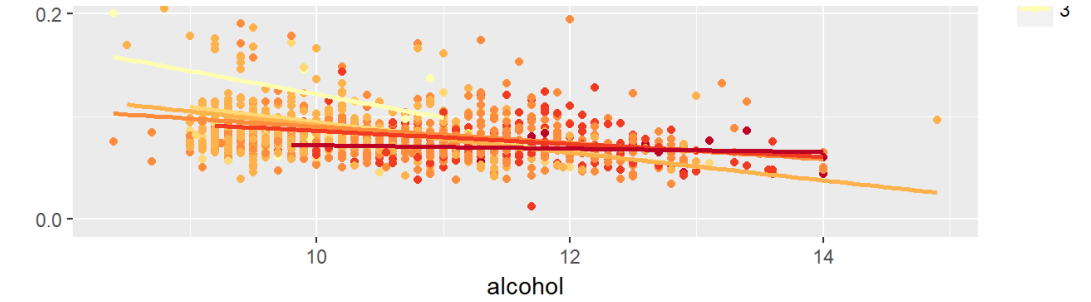


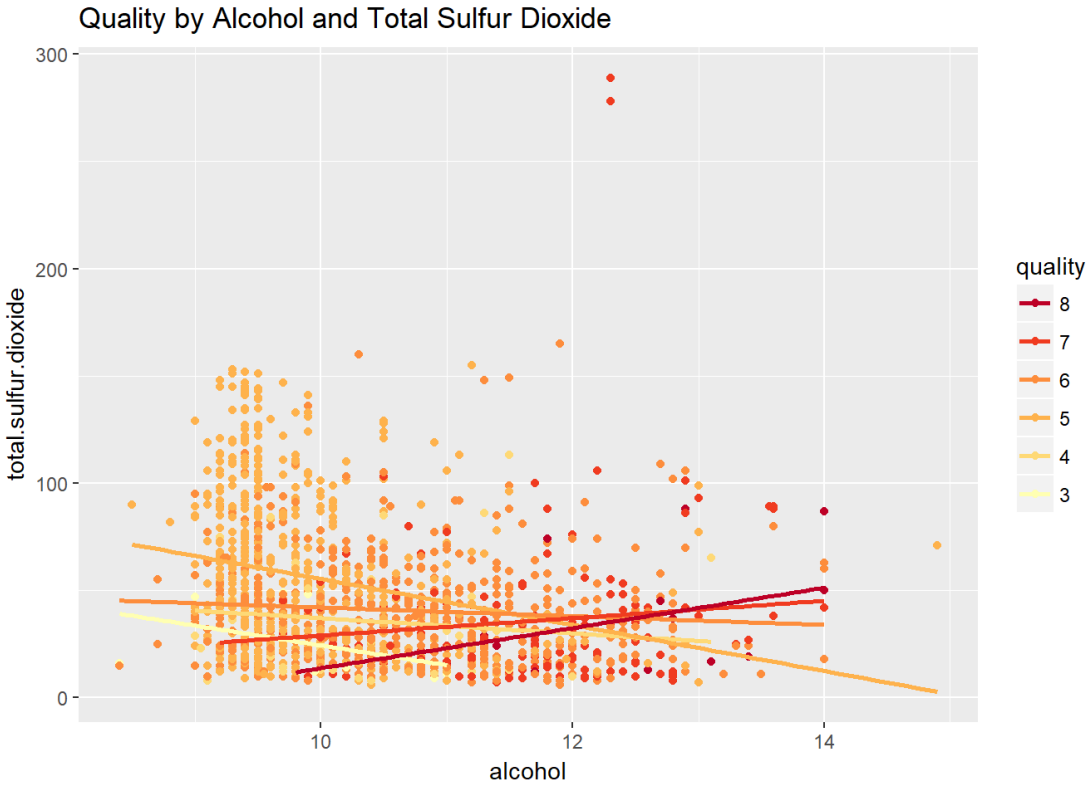
Quality by Alcohol and Citric Acid



Quality by Alcohol and Chlorides

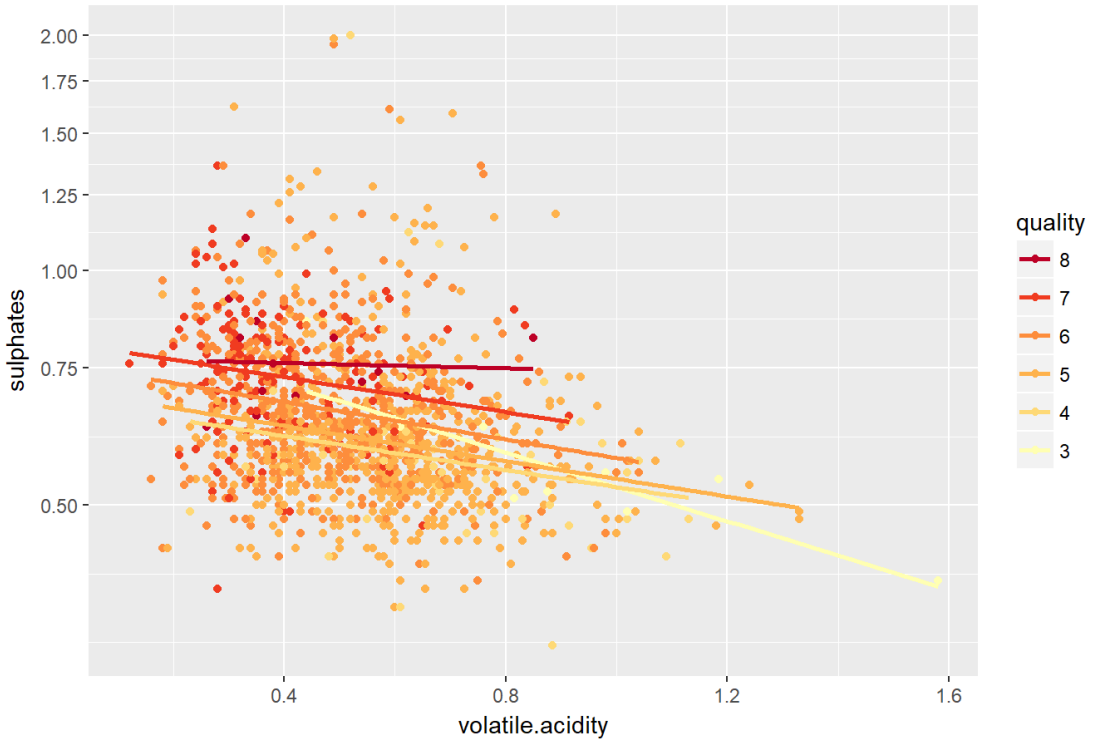




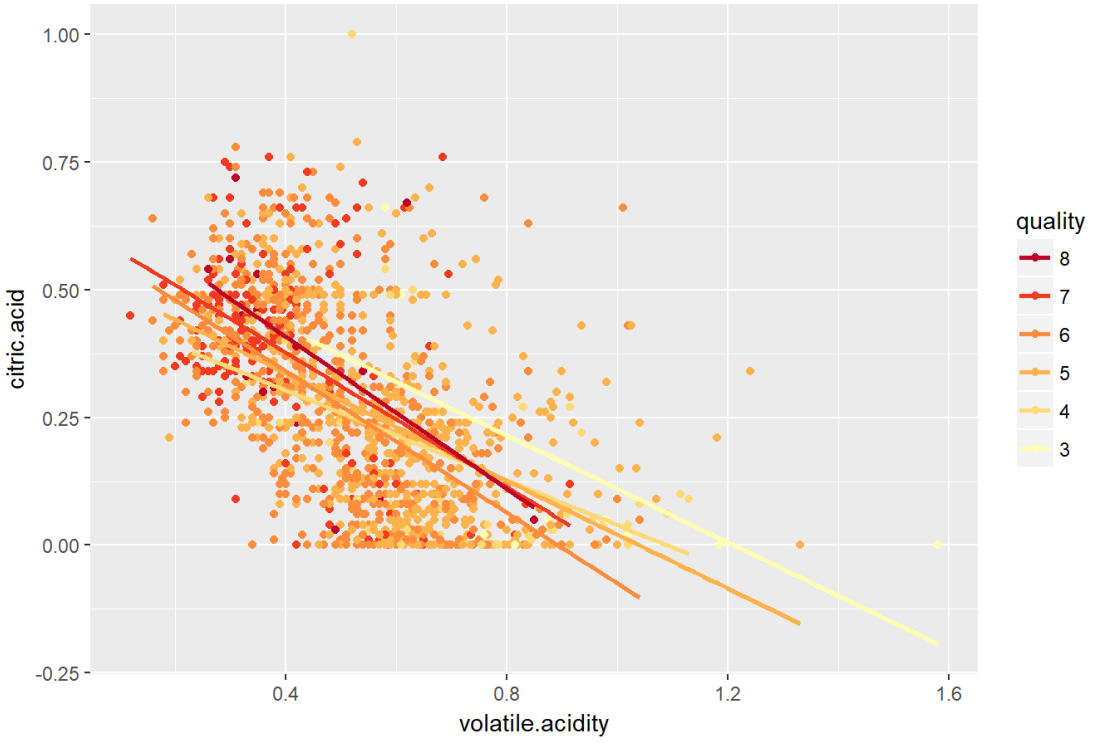


Keeping alcohol constant, I can see that higher levels of sulphates do have a positive effect on quality. Similar is the case for citric.acid. However, chlorides and total.sulfur.dioxide affects quality inversely for wines whose alcohol content is greater than 12%.

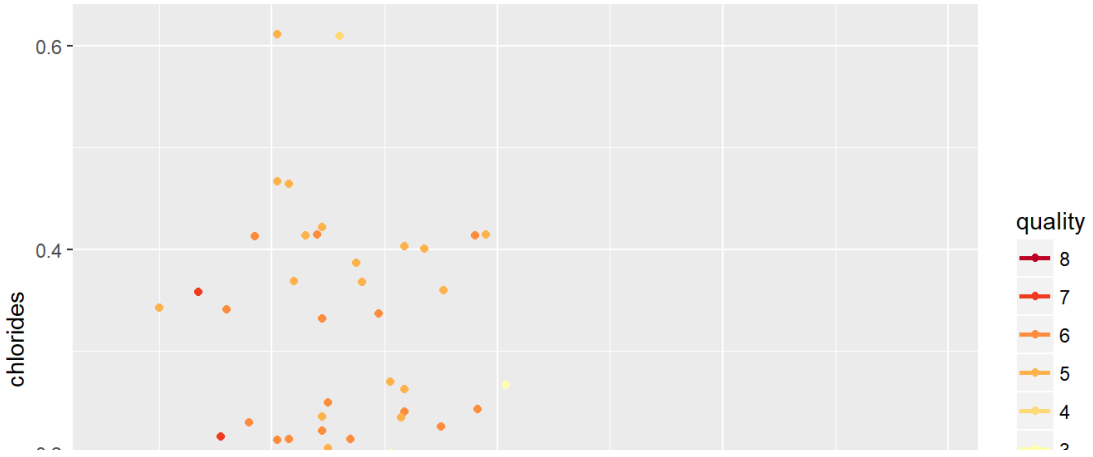
Quality by Volatile Acidity and Sulphates (log_10)

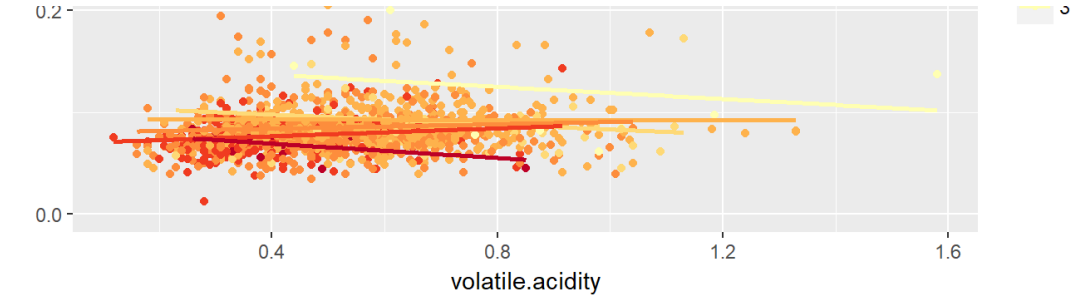


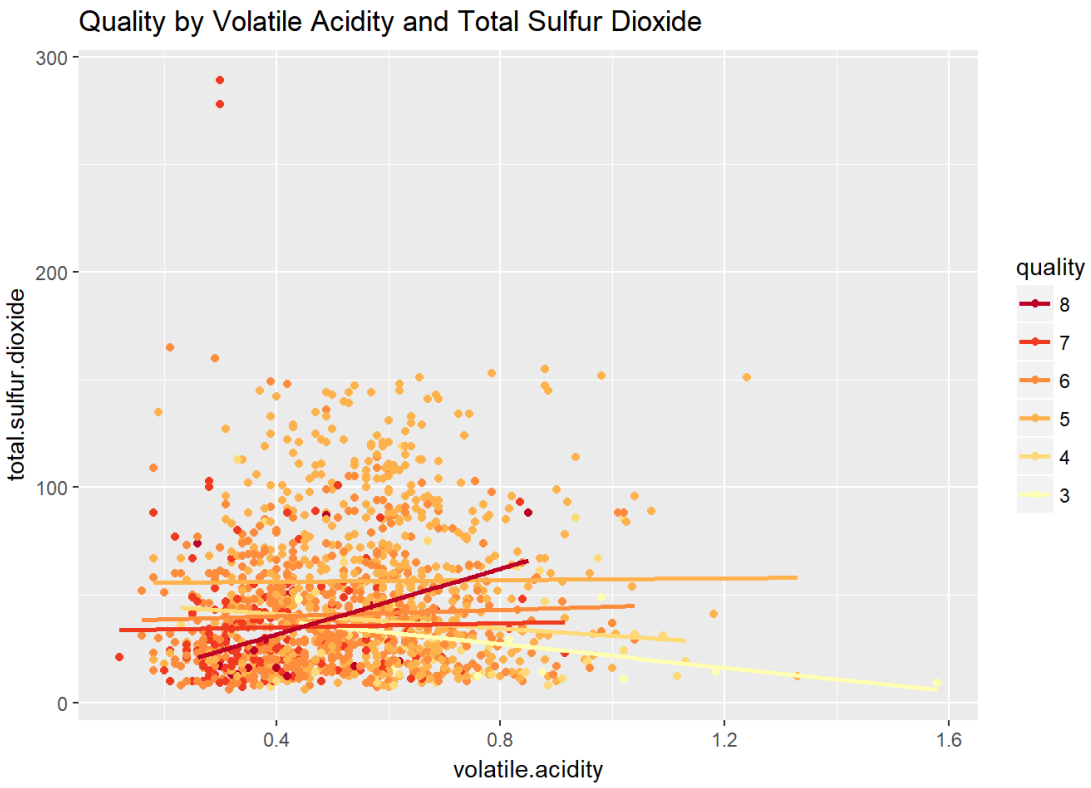
Quality by Volatile Acidity and Citric Acid



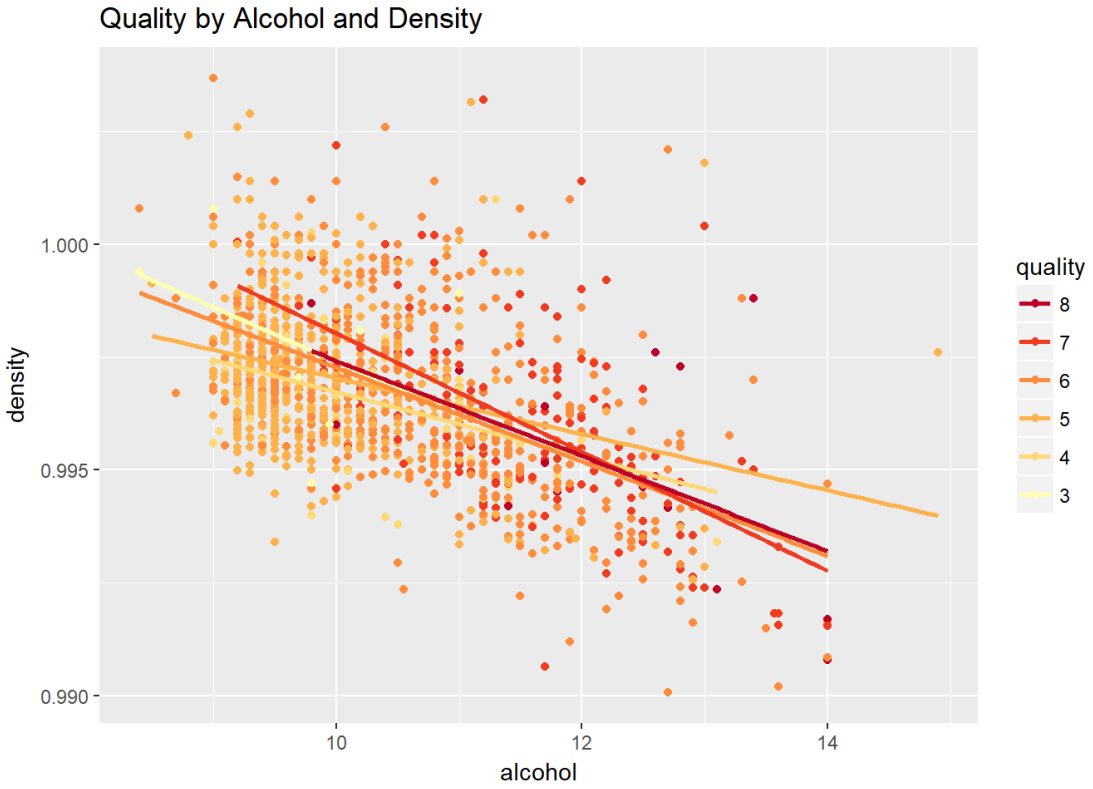
Quality by Volatile Acidity and Chlorides

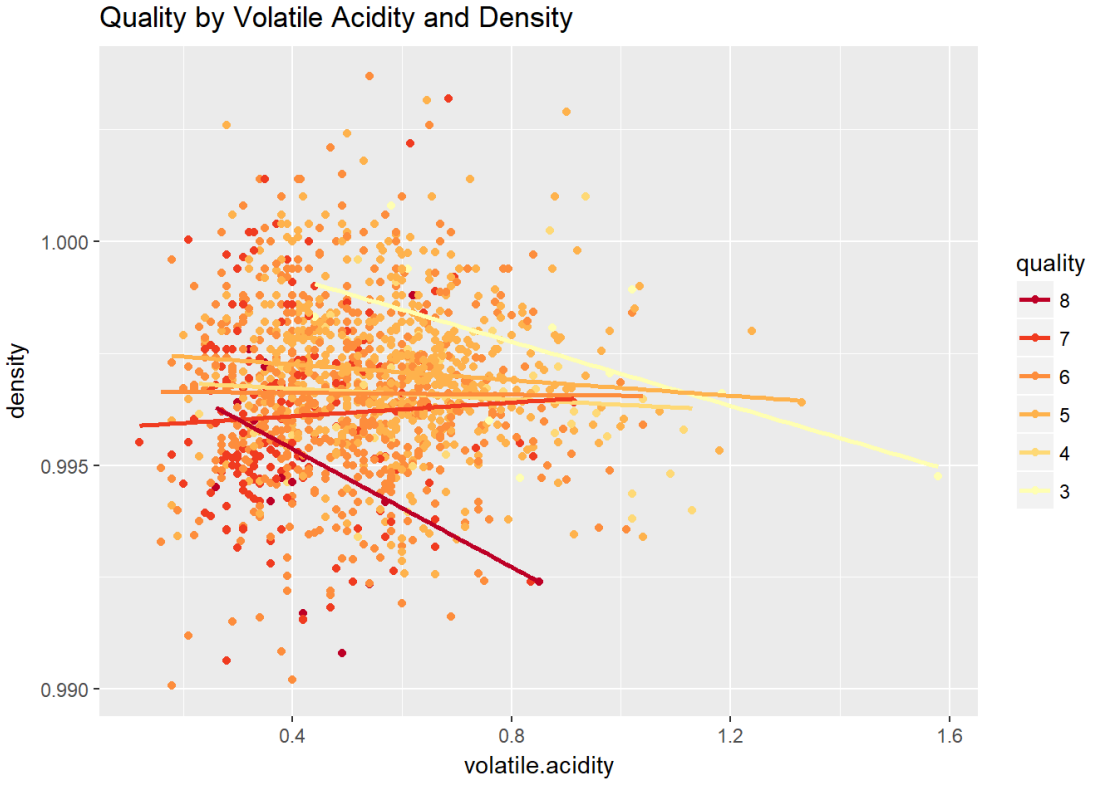




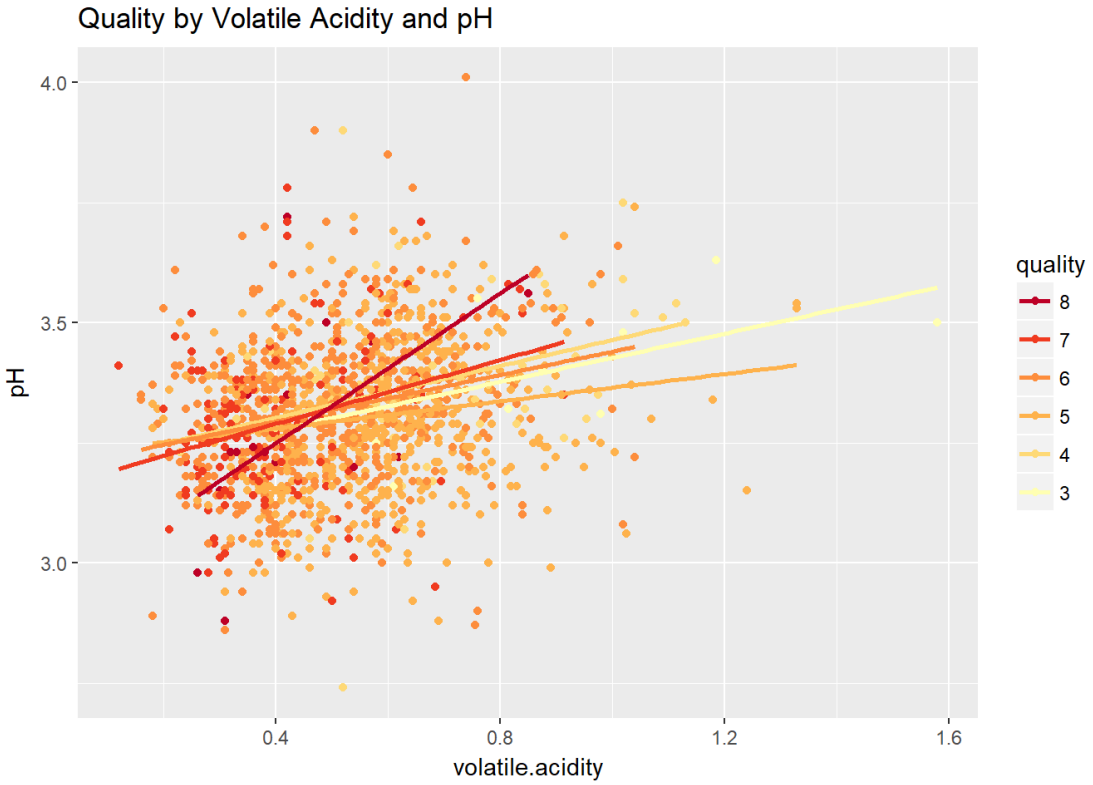


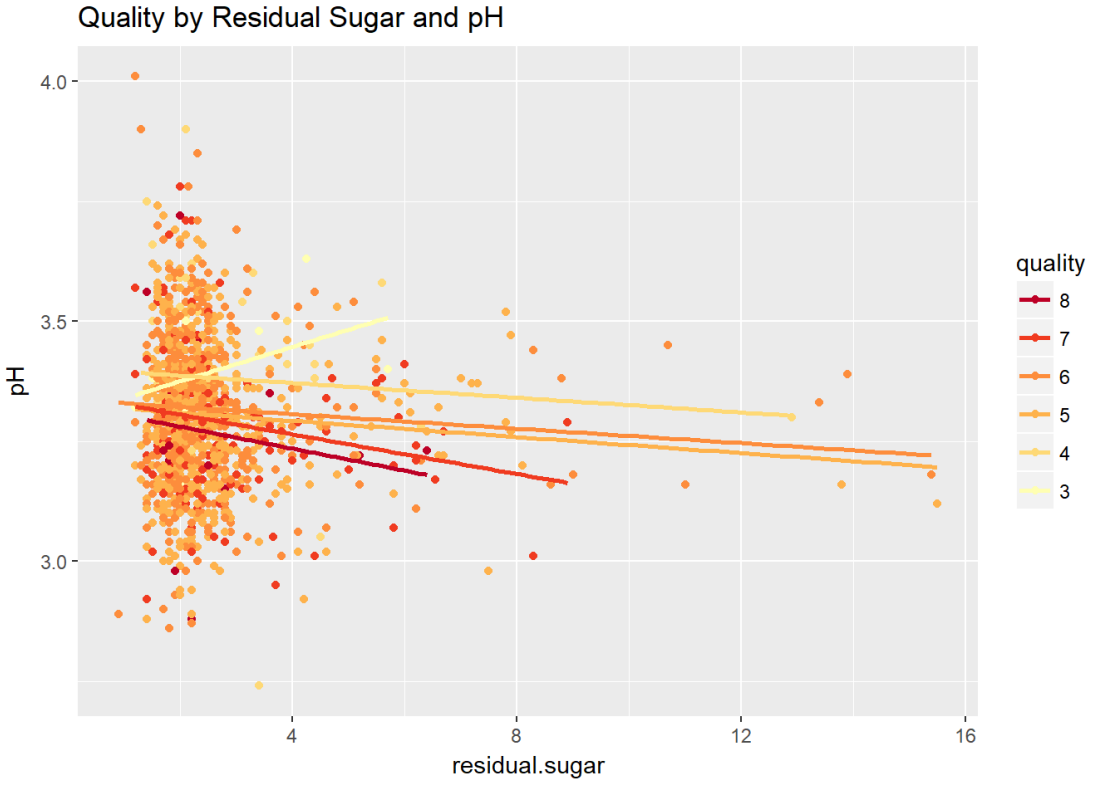
Here, too, sulphates and citric acid seem to have some positive effect on quality. This time, I see that red wines with lower level of chlorides and total.sulfur.dioxide have a better quality.





Effect of density on quality is not visible in the alcohol plot. However, when plotted with volatile.acidity, I see that a lot of better-quality wines are less dense compared to poor-quality wines.

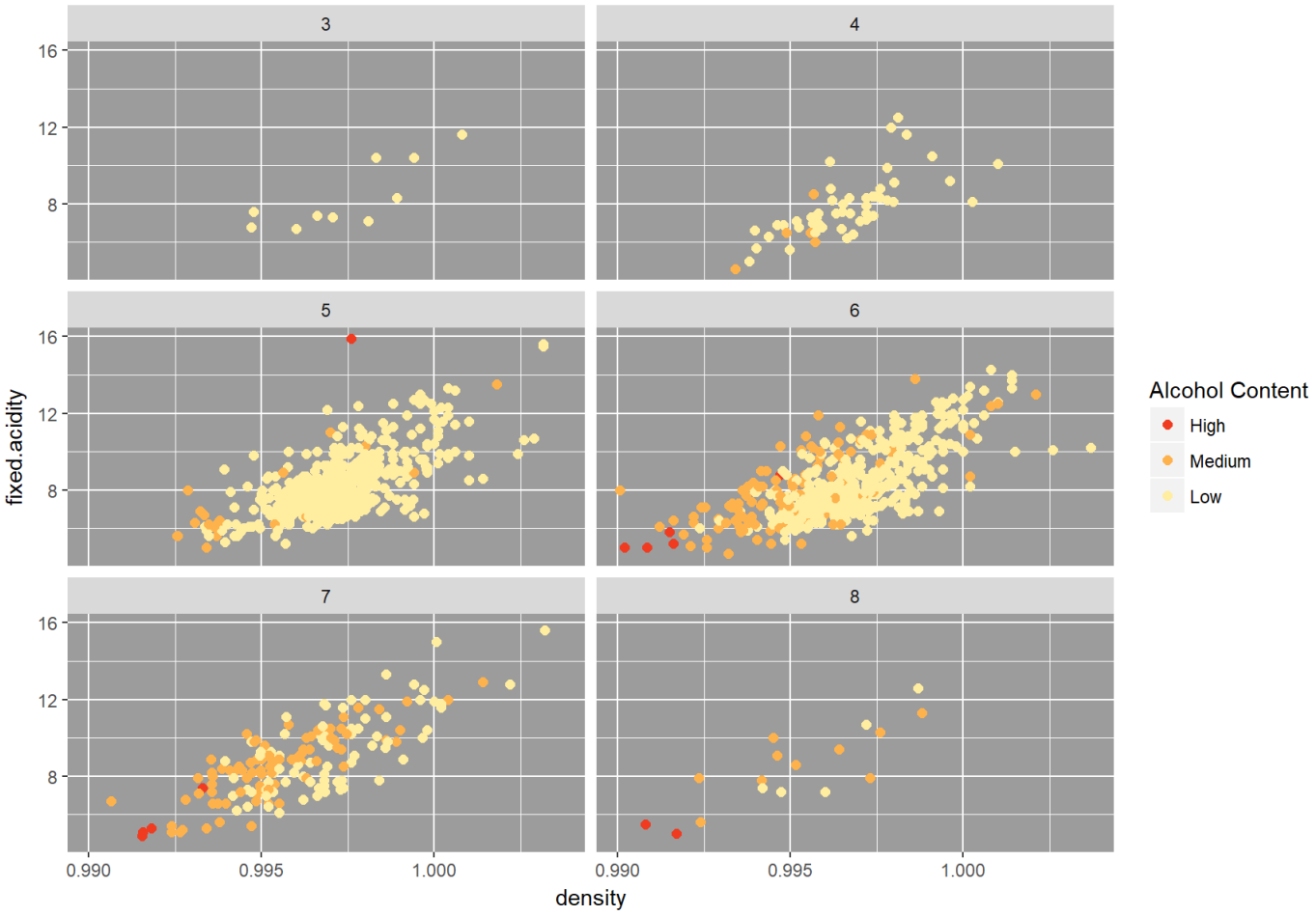




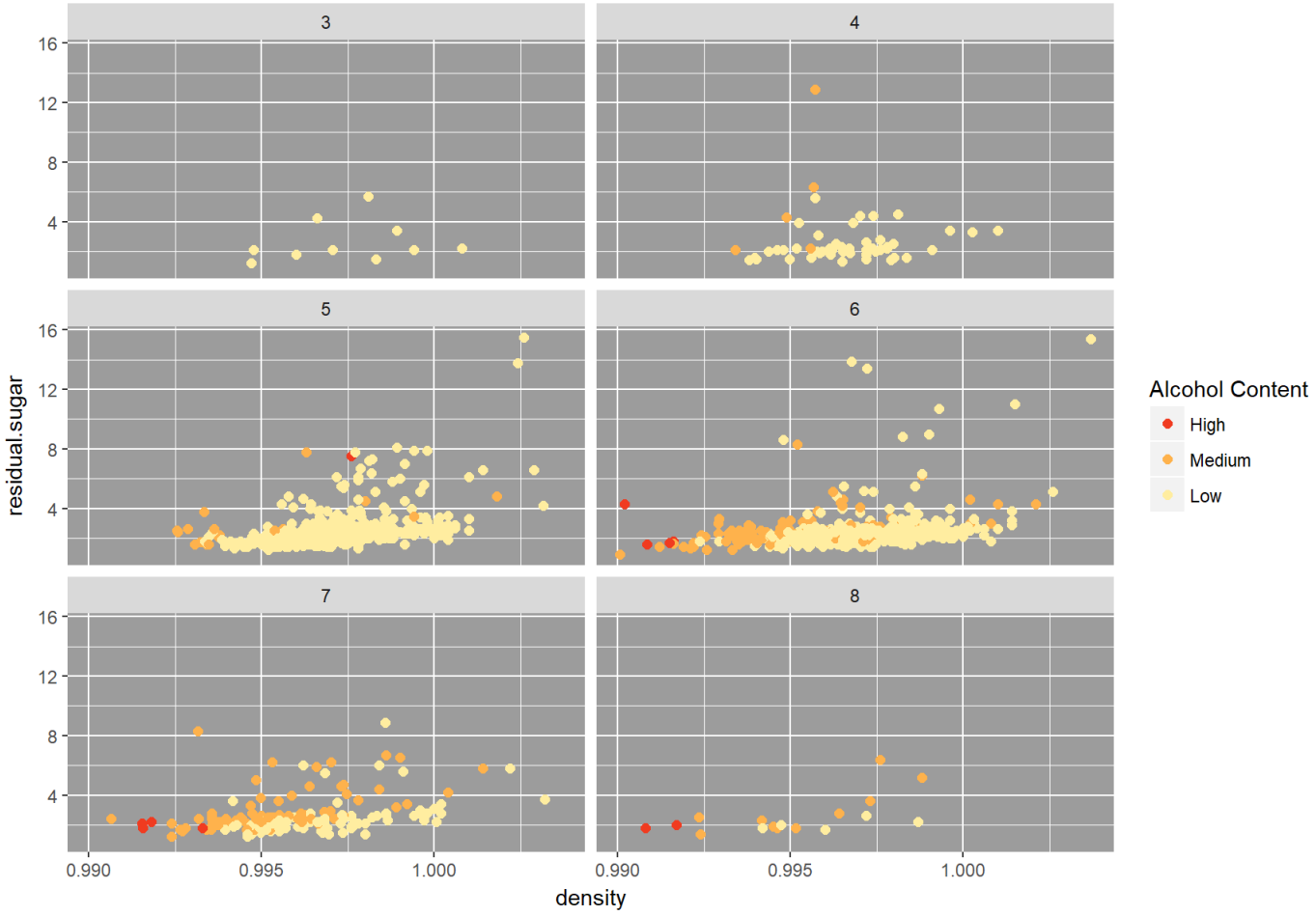
Again, volatile.acidity is showing positive correlation with pH which is quite the opposite of the expected relationship between pH and acidity. Residual sugar and pH doesn't seem to be correlated, though.

From the first plot, most of the wines with lower volatile.acidity expected to be graded good in quality, shows a decrease of quality with increase of pH. The second plot furthers this argument by showing an increase in pH with decrease in wine quality. So, I can say that the better-quality wines have an inverse relationship with pH.

Alcohol Content by Density and Fixed Acidity - Facet wrap by Quality



Alcohol Content by Density and Residual Sugar - Facet wrap by Quality



Density is highly correlated with fixed.acidity (positively) as well as alcohol (negatively). Coming to quality, fixed.acidity doesn't seem to have much of an effect. Though, high-quality wines seem to be less dense, there are lower quality wines exhibiting the same. Similarly, residual sugar, too, does not affect quality in any considerable way.

```
##
## Calls:
## m1: lm(formula = I(quality) ~ I(alcohol), data = redwines)
## m2: lm(formula = I(quality) ~ I(alcohol) + volatile.acidity, data = redwines)
## m3: lm(formula = I(quality) ~ I(alcohol) + volatile.acidity + log10(sulphates),
##     data = redwines)
## m4: lm(formula = I(quality) ~ I(alcohol) + volatile.acidity + log10(sulphates) +
##     total.sulfur.dioxide, data = redwines)
## m5: lm(formula = I(quality) ~ I(alcohol) + volatile.acidity + log10(sulphates) +
##     total.sulfur.dioxide + chlorides, data = redwines)
## m6: lm(formula = I(quality) ~ I(alcohol) + volatile.acidity + log10(sulphates) +
##     total.sulfur.dioxide + chlorides + pH, data = redwines)
## m7: lm(formula = I(quality) ~ I(alcohol) + volatile.acidity + log10(sulphates) +
##     total.sulfur.dioxide + chlorides + pH + citric.acid, data = redwines)
##
## ================================================================================================
##                           m1         m2         m3         m4         m5         m6         m7
## ------------------------------------------------------------------------------------------------
##   (Intercept)            1.875***   3.095***   3.369***   3.612***   3.998***   5.369***   5.819***
##                         (0.175)    (0.184)    (0.184)    (0.191)    (0.208)    (0.395)    (0.461)
##   I(alcohol)             0.361***   0.314***   0.303***   0.290***   0.270***   0.285***   0.290***
##                         (0.017)    (0.016)    (0.016)    (0.016)    (0.016)    (0.017)    (0.017)
##   volatile.acidity                 -1.384***  -1.156***  -1.134***  -1.076***  -0.960***  -1.061***
##                                    (0.095)    (0.097)    (0.097)    (0.097)    (0.101)    (0.114)
##   log10(sulphates)                             1.477***   1.519***   1.843***   1.825***   1.864***
##                                               (0.177)    (0.176)    (0.189)    (0.189)    (0.190)
##   total.sulfur.dioxide                                    -0.002***  -0.002***  -0.002***  -0.002***
##                                                           (0.001)    (0.001)    (0.001)    (0.001)
##   chlorides                                                          -1.729***  -2.077***  -1.968***
##                                                                      (0.380)    (0.388)    (0.392)
##   pH                                                                            -0.468***  -0.590***
##                                                                                 (0.115)    (0.132)
##   citric.acid                                                                              -0.226
##                                                                                            (0.120)
## ------------------------------------------------------------------------------------------------
##   R-squared                0.227      0.317      0.345      0.353      0.361      0.368      0.369
##   adj. R-squared           0.226      0.316      0.344      0.352      0.359      0.366      0.367
##   sigma                    0.710      0.668      0.654      0.650      0.646      0.643      0.643
##   F                      468.267    370.379    280.646    217.574    180.338    154.524    133.167
##   p                        0.000      0.000      0.000      0.000      0.000      0.000      0.000
##   Log-likelihood       -1721.057  -1621.814  -1587.752  -1578.324  -1568.023  -1559.722  -1557.943
##   Deviance               805.870    711.796    682.108    674.111    665.482    658.608    657.144
##   AIC                   3448.114   3251.628   3185.503   3168.648   3150.046   3135.443   3133.885
##   BIC                   3464.245   3273.136   3212.389   3200.911   3187.686   3178.460   3182.280
##   N                     1599       1599       1599       1599       1599       1599       1599
## ================================================================================================
```

The variables in this linear model can account for 36.9% of the variance in the quality of red wines. With only alcohol and volatile acidity, we can account for 31.7% of the variance. However, with log transformation of sulphates, we can account for 34.5% of the variance.

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?

Red wines with high levels of alcohol content (11.5 - 13.5 %) and lower levels of volatile acidity are graded the highest on the quality scale. Many such wines have a median alcohol of around 12% and a volatile.acidity of around 0.5. Another notable thing for these wines is the absence of outliers, thus eliminating the possibility of skewedness to a large extent.

Keeping alcohol constant, I can see that higher levels of sulphates do have a positive effect on quality. Similar is the case for citric.acid. However, chlorides and total.sulfur.dioxide inversely affects the quality of wines whose alcohol content is more than 12%.

When plotted with volatile.acidity and residual.sugar, pH was found to be inversely related with the quality of wines.

Were there any interesting or surprising interactions between features?

pH, which is expected to have an inverse relationship with acidic features, showed positive correlation with volatile acidity. Also, contrary to the seemingly popular belief that sugar can result in an increase in pH value, residual sugar showed absolutely no signs of correlation with pH.

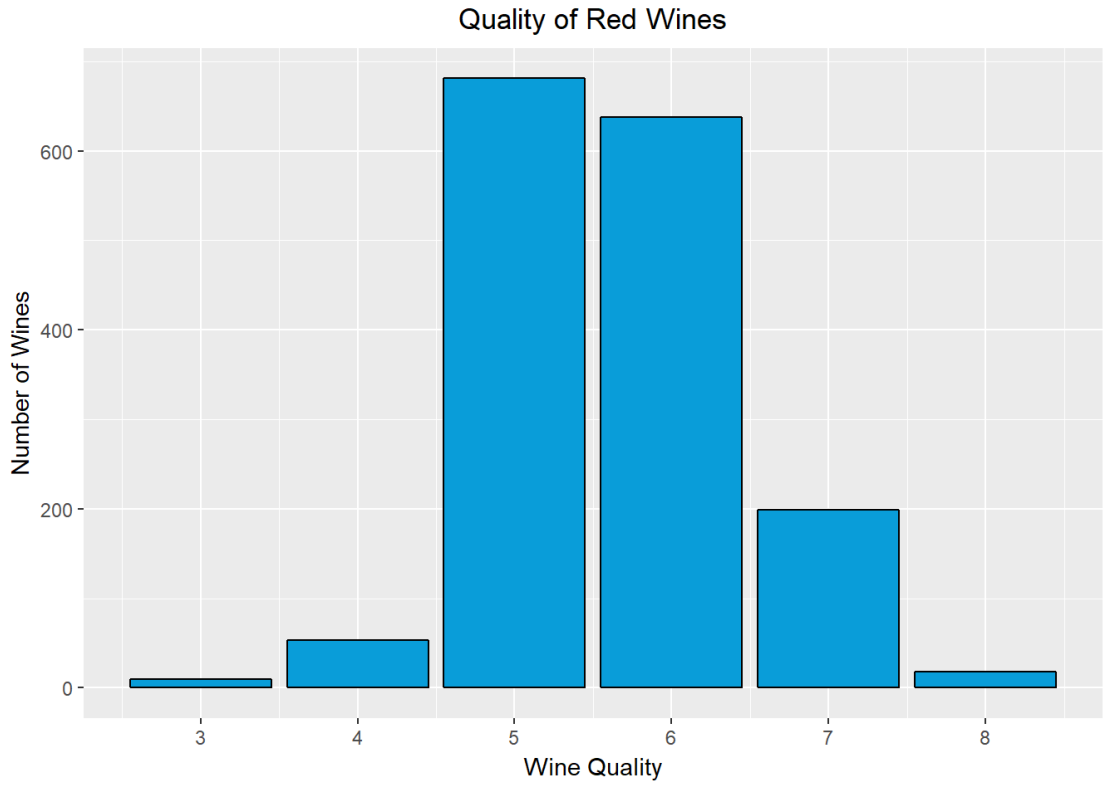
OPTIONAL: Did you create any models with your dataset? Discuss the strengths and limitations of your model.

Yes, I created a linear model starting from alcohol and volatile acidity.

These 2 variables accounted for 31.7% of the variance. Addition of log-transformed sulphates and total sulfur dioxide improved it to 34.4%. After adding the low-correlated features like chlorides, pH and citric acid, I was able to account for approximately 37% variance in the quality of red wines.

Final Plots and Summary

Plot One



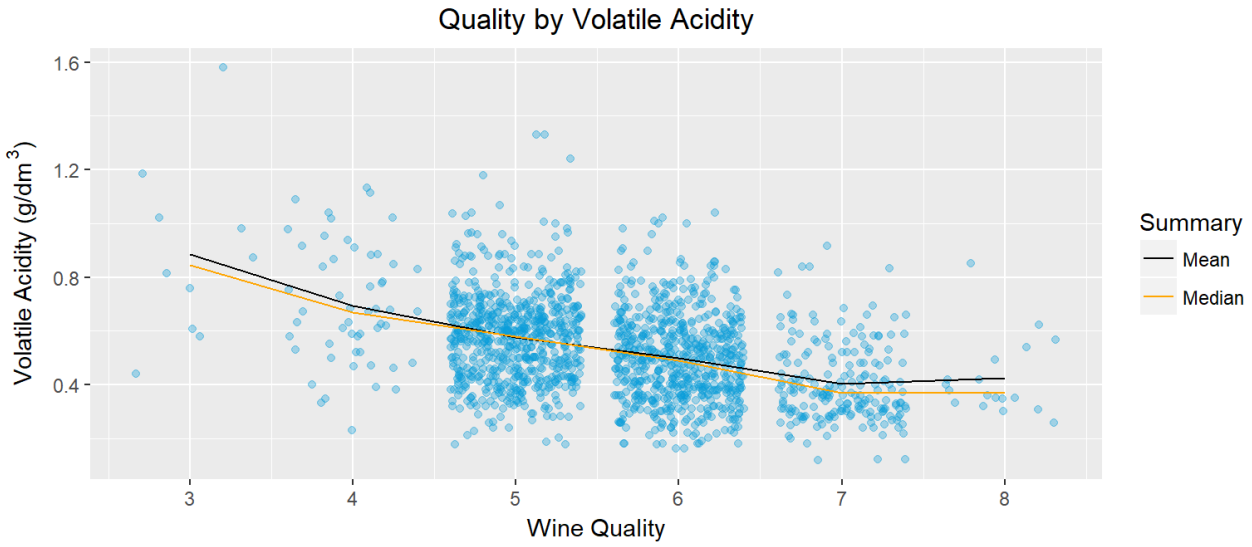
Count of wines by Quality (3 to 8):

```
##
##   3   4   5   6   7   8
##  10  53 681 638 199  18
```

Description One

The distribution of quality of red wines appears to be a normal distribution, peaking at quality equal to 5 and 6. Additionally, the table below the histogram tells us that more than 80% of the wine samples were graded average by experts.

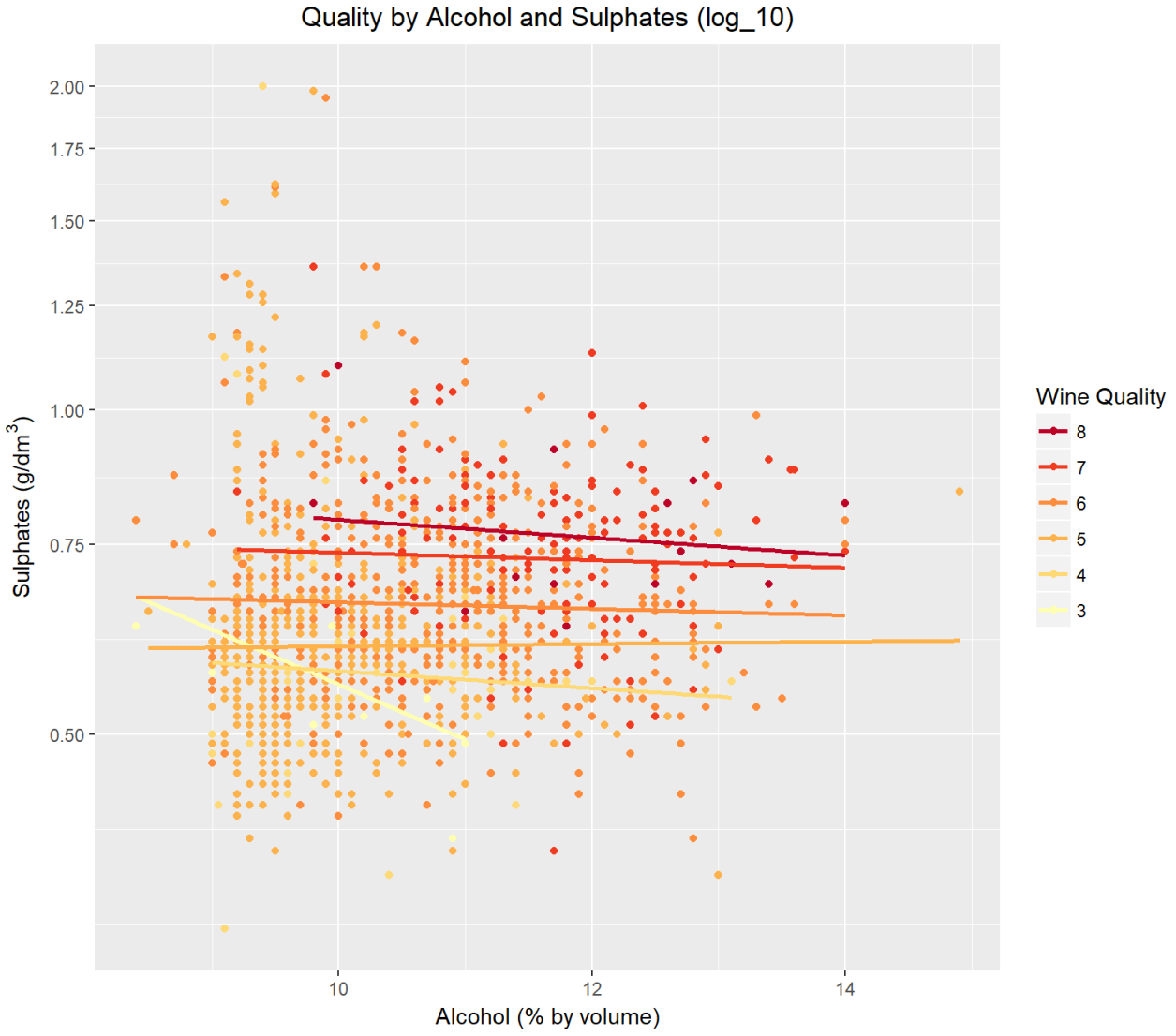
Plot Two



Description Two

The mean and median values of alcohol just shoots up from levels below 10% to more than 12% as quality improves from 5 to 8. On the other hand, volatile acidity decreases as quality of wine increases. From the mean and median summaries of volatile acidity, it looks like the best quality wines have a volatile acidity of around 0.4 g/dm^3.

Plot Three



Description Three

After having determined that alcohol content and volatile acidity have some effect on the quality of red wine, sulphates was the next feature to have maximum correlation with quality. Log-transformation of sulphates led to more correlaton and resulted in a more normalised distribution. This plot shows that keeping alcohol constant, the yellow dots slowly turn to orange and then to red with increasing value of sulphates. So, higher level of sulphates do have a positive effect on quality

---

Reflection

The red wines data set contains 1599 instances of red wine and its 12 attributes. Out of these 12 attributes, eleven are based on physiochemical tests and are termed as input variables. The twelfth attribute, “quality”, is scored between 0(very bad) and 10(excellent) based on sensory data and is termed as the output variable. I started by understanding the individual variables in the data set, and then I explored interesting questions and leads as I continued to make observations on plots. Eventually, I explored the quality of red wines across many variables and created a linear model that can account for 36.9% variance in the quality of red wines.

Out of all the input variables, alcohol and volatile acdity showed some degree of impact on the quality of a red wine. While higher alcohol content positively affected the wine quality, volatile acidity showed a negative correlation with the quality of red wines. I was surprised that fixed acidity and residual sugar (transformed or not), known to add tartness and dryness in wine taste, did not show any worthwhile correlation with quality. I categorised wines based on 3 levels of alcohol-content, which helped me further to establish these findings.

I also thought that chlorides and citric acid will show some substantial correlation with quality as their presence in wines is supposed to affect its taste considerably. Though both these features showed “some” impact, it was definitely less than what I had expected. What met my expectations though, was the fact that chlorides, known to add an undesirable salty taste to wines, was negatively correlated. Whereas, citric acid, which can add freshness and flavor to wines showed a positive correlation with wine quality, as expected.

Sulphates showed some correlation with quality which improved a bit after log-transformation. Keeping alcohol constant, I found higher levels of sulphates to have a positive effect on quality. Now, sulphates are known to add sulfur dioxide gas levels which, beyond certain concentraions, becomes evident in the nose and taste of wine. So, I naturally progressed to explore both “free” and “total” sulfur dioxide features expecting either or both of them to have a similar (i.e. positive) effect on quality. Though free sulfur dioxide has no effect, I was rather surprised to find that total sulfur dioxide has a negative correlation with quality. pH, too, showed an inverse relationship but, with better-quality wines having lower levels of volatile acidity.

At the end, I created a linear model using 7 (*i.e. alcohol, volatile.acidity, sulphates, total.sulfur.dioxide, chlorides, pH and citric acid*) of the 11 attributes that can account for 36.9% of the variance in the quality of red wines. This is pretty low as compared to what I would have preferred. However, I feel that this will improve if we have more records for lowest and highest quality wines that will evenly distribute the dataset across all the ratings. This current dataset has more than 80% of the wine samples graded as average (i.e. rated 5 or 6) whereas the remaining 20% are distributed among the other 4 ratings. Also, information regarding grape types, wine brand, wine selling price (*currently unavailable due to privacy and logistic issues*) may account for better results in predicting the quality of red wines.