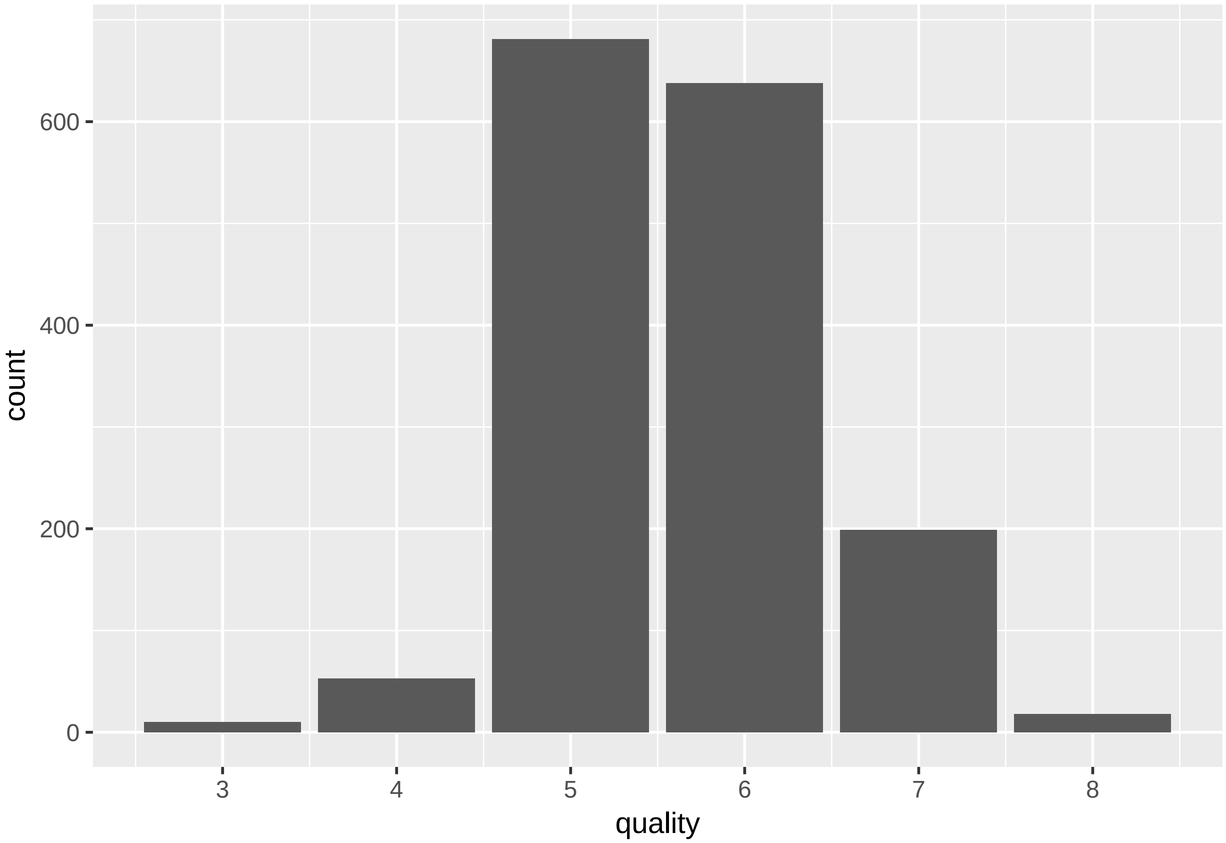
Exploratory data analysis on the red wines dataset by Badrinath Thirumalchari

Lets take a look the summary of the entire dataset.

```
##        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
```

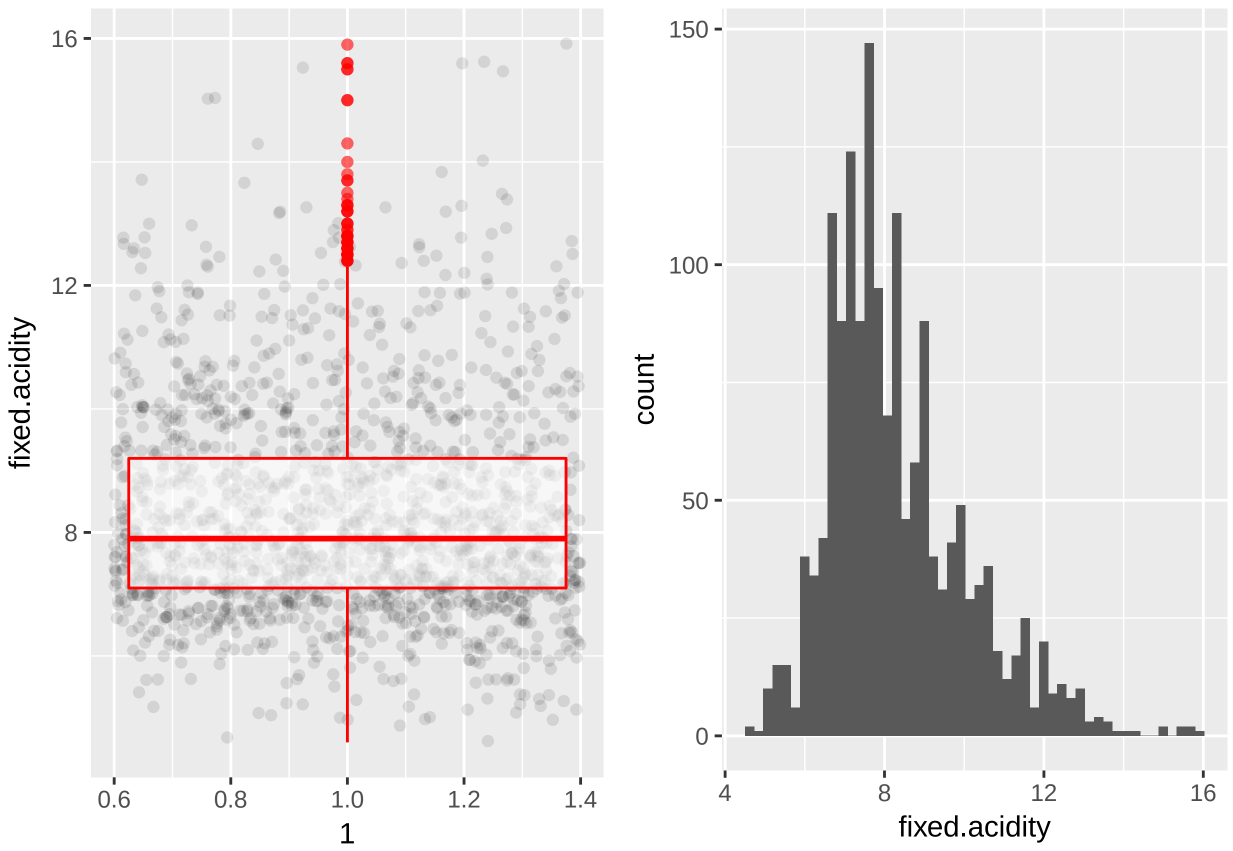
Univariate Plots Section



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   5.000   6.000   5.636   6.000   8.000
```

I first wanted to look at the distribution of quality metric, because eventually we will be finding what are all the factors that responsible for determining the quality of red wine. We can see that over 1200 wines have a rating of 5 or 6 in the dataset ie.. over 75% of the wines in the dataset are either rated 5 or 6. It is also useful to note that none of the wines have a rating of less then 3 or greater then 8.

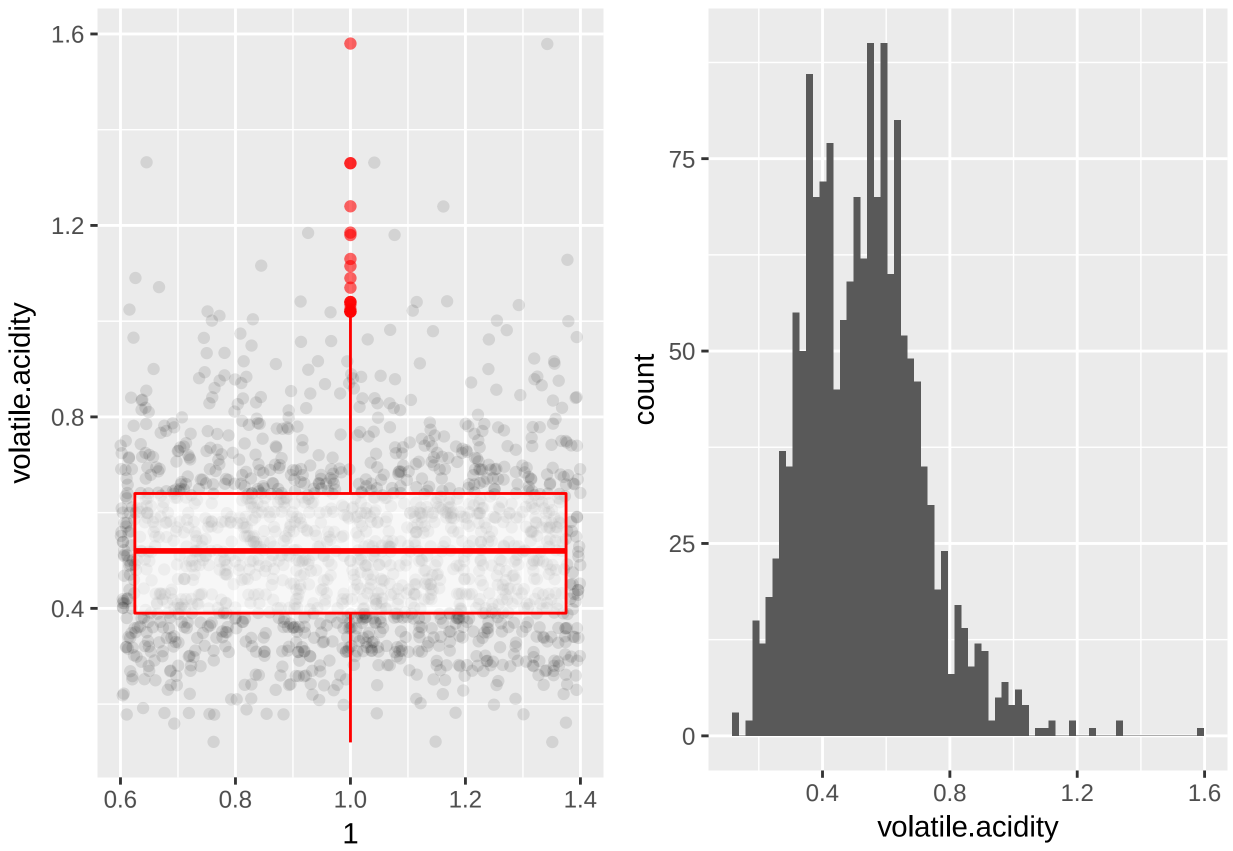
Lets start investigating the features starting with fixed acidity.



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

THe distribution approximately looks normal with a few outliers. The mean of the distribution is 8.32 g/dm^3 and median is 7.90 g/dm^3 with a maximum value of 15.90 g/dm^3.

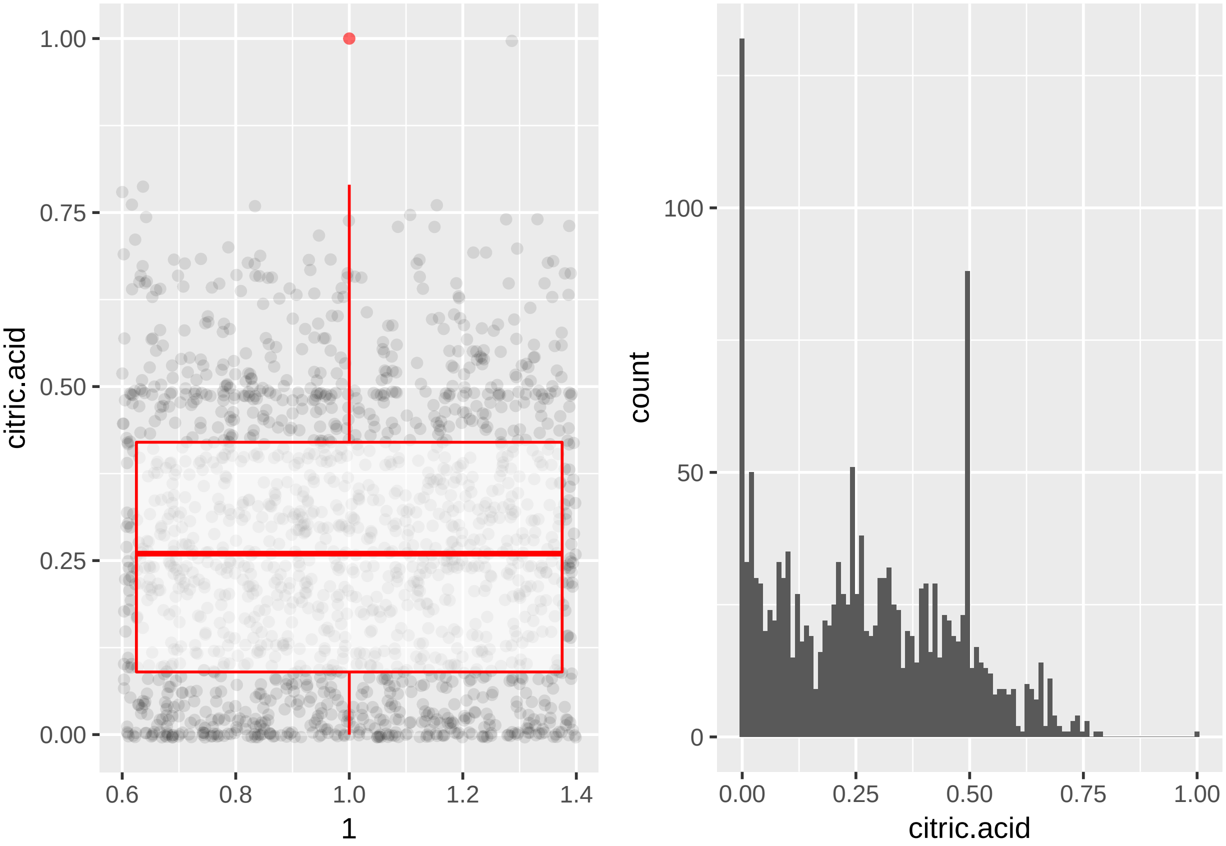
Lets investigate volatile acidity.



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

This distribution looks like it has two peaks and large valued outliers. Most of the distribution is between .2 g/dmˆ3 and 1 g/dmˆ3 but we see some large valued outliers. The maximum value of the outlier is 1.58 g/dmˆ3.

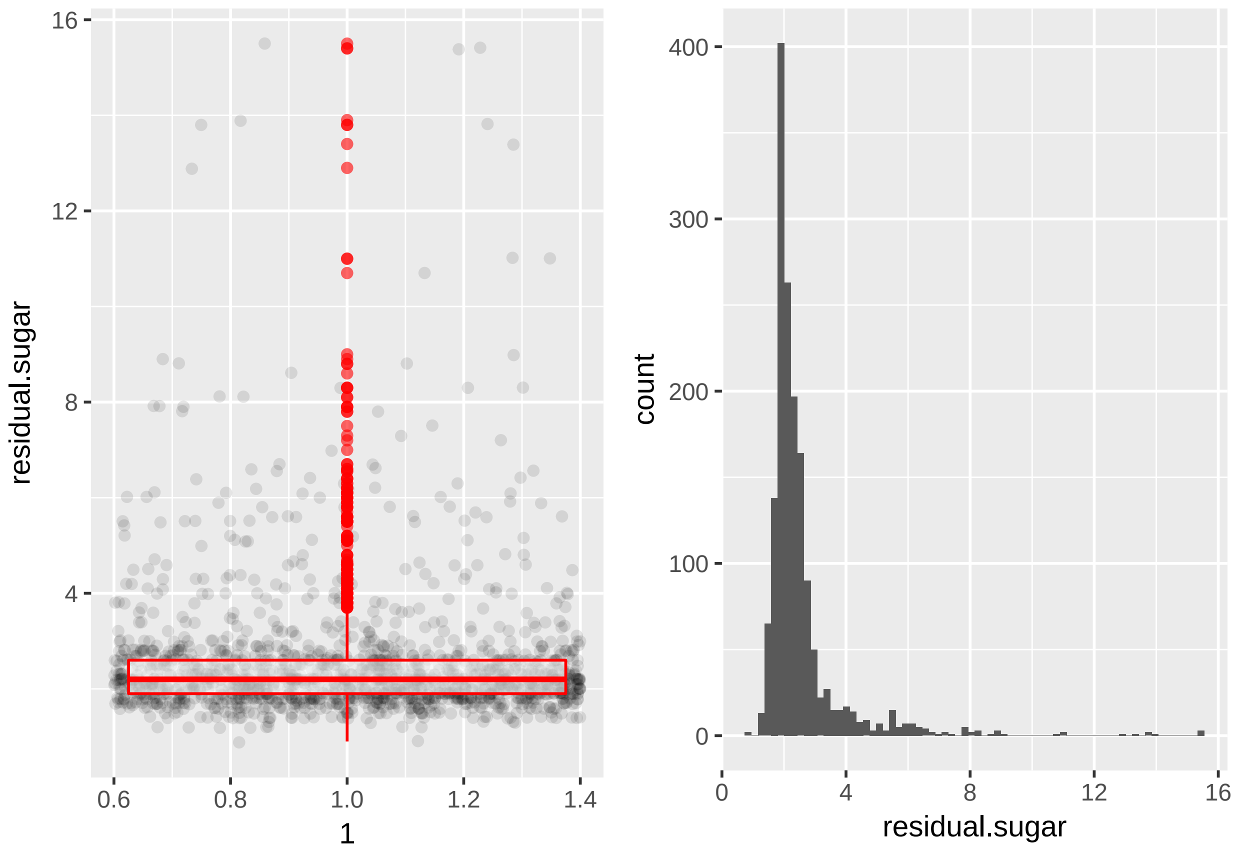
Lets investigate citric acid.



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

This distribution is a bit different because it has two significant peaks at 0 g/dmˆ3 and .49 g/dmˆ3. Most of the wines have a value of 0 g/dmˆ3 or .49 g/dmˆ3. The mean value of the feature is .271 g/dmˆ3 with a maximum value of 1 g/dmˆ3.

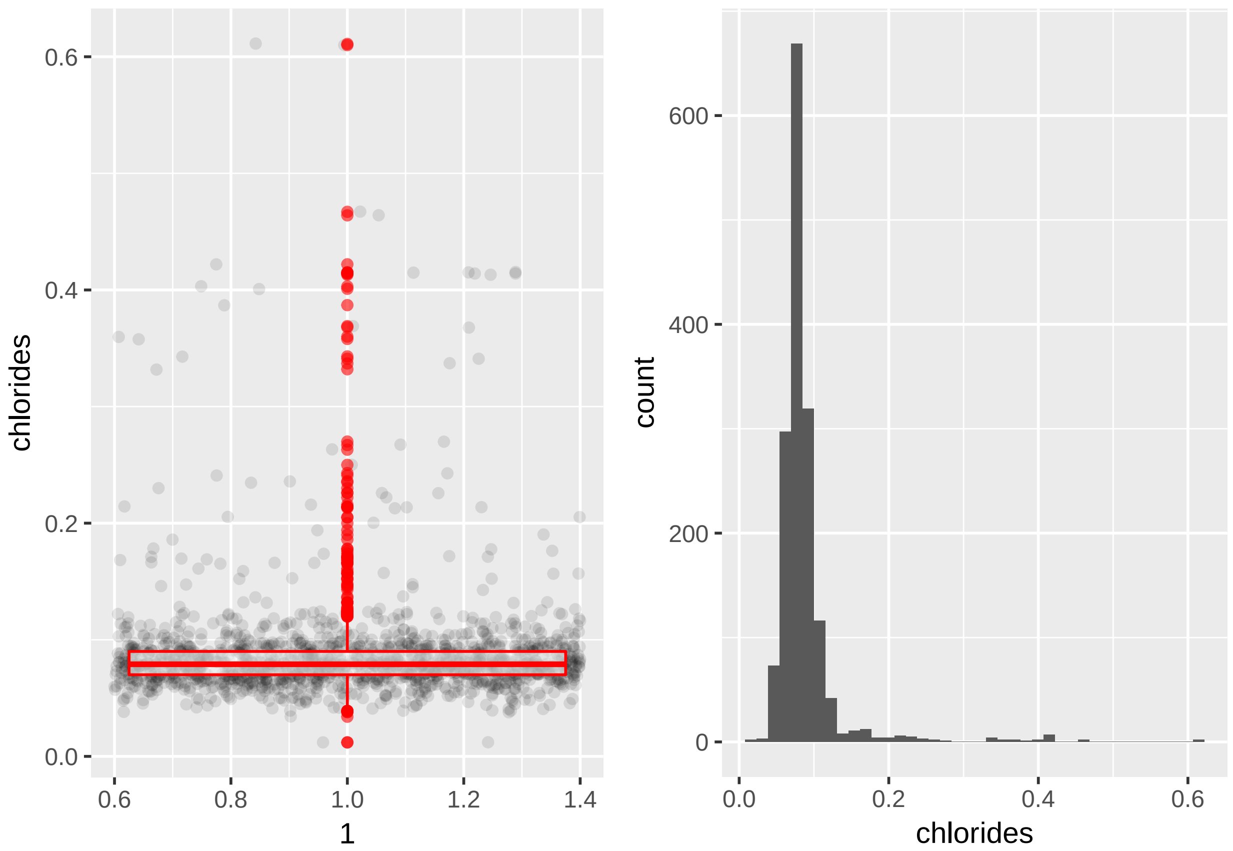
Lets investigate residual sugar in the wines.



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

The distribution has a very long tail so I made another plot showing only 95% of the data. From the plot we see that there is a sharp peak at 2 g/dm^3 and it looks like a normal distribution with a mean of 2.539 g/dm^3 and maximum value of 15.5 g/dm^3.

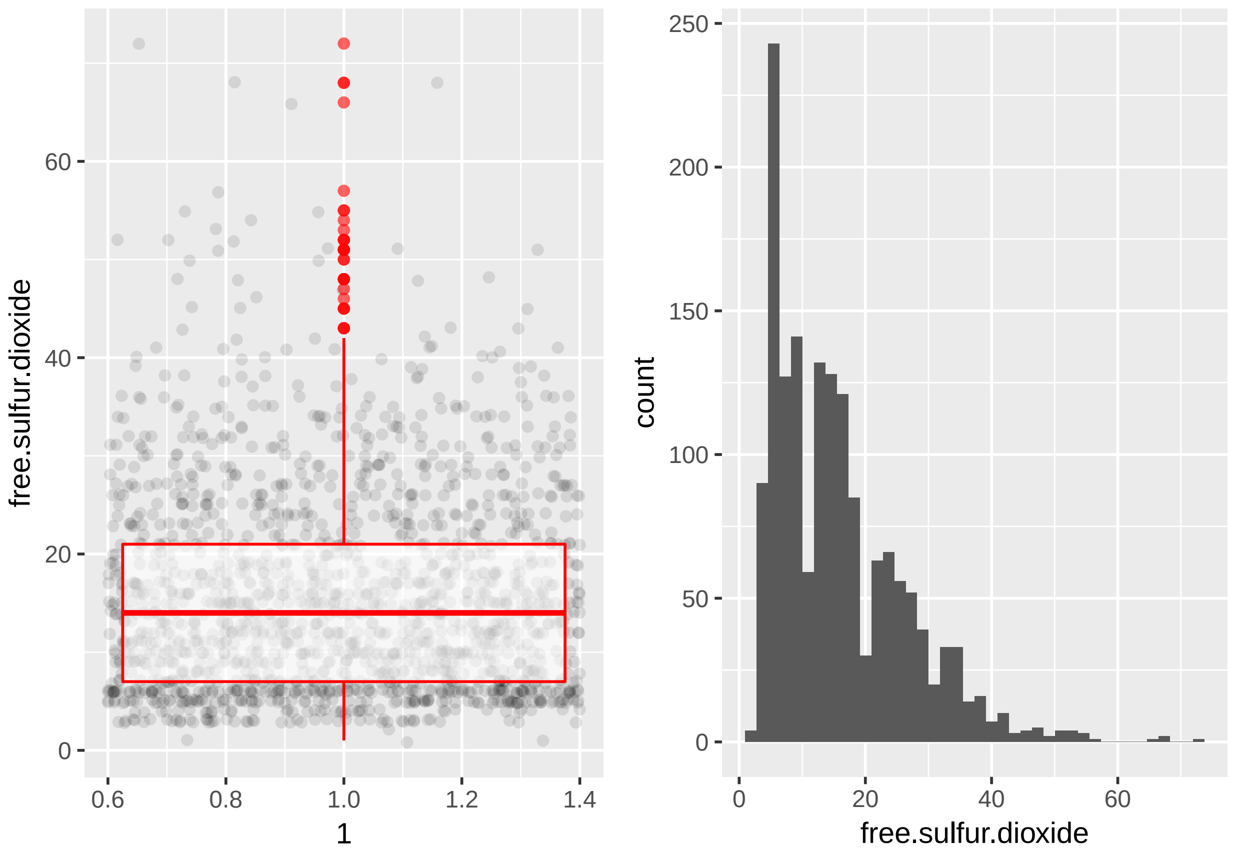
Lets investigate chlorides in the wine.



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

The distribution above looks like it has a very long tail with outliers, so I decided to make a plot with 96% of the data. The distribution looks normal with a mean of 0.087 g/dmˆ3 and maximum value 0.611 g/dmˆ3.

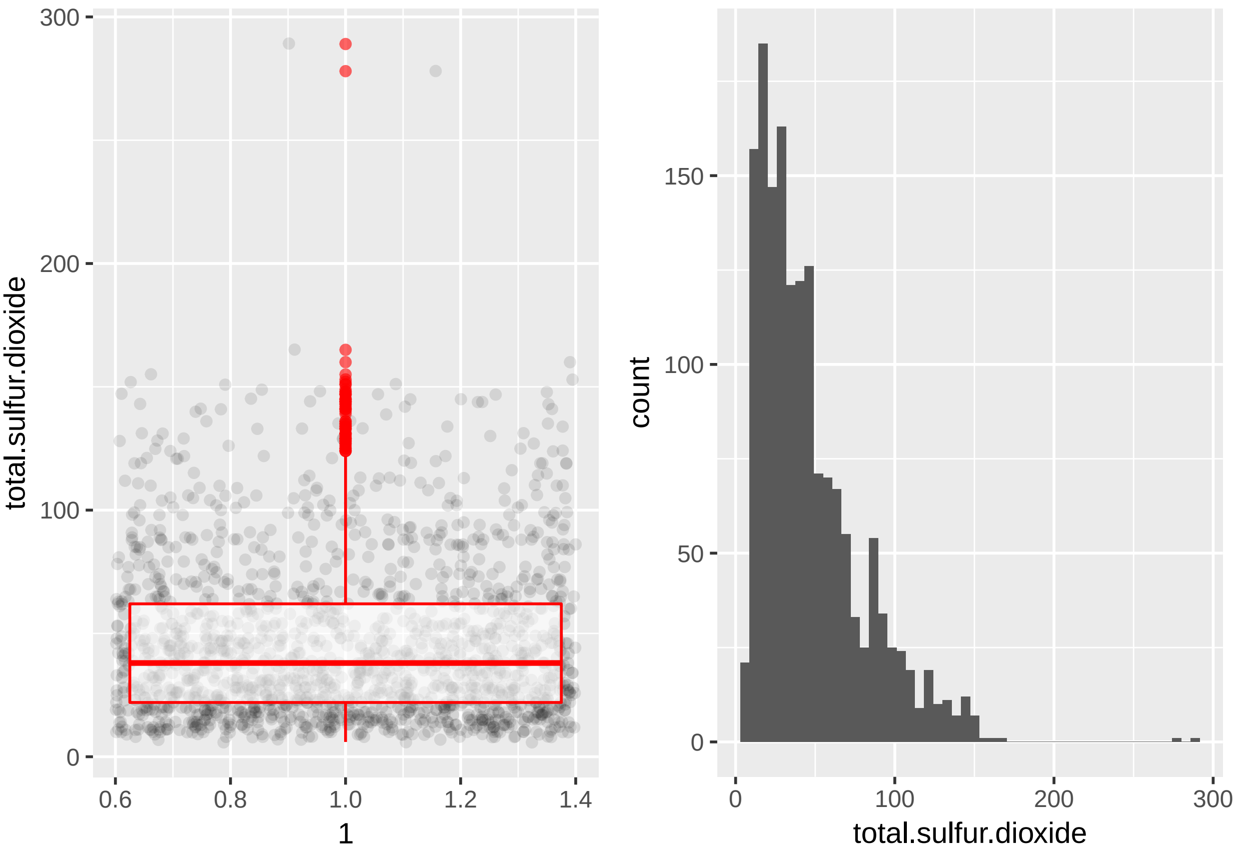
Lets investigate free sulfur dioxide content in the red wine dataset.



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

The distribution above looks skewed with a mean of 15.87 mg/dmˆ3 and a maximum value of 72 mg/dmˆ3.

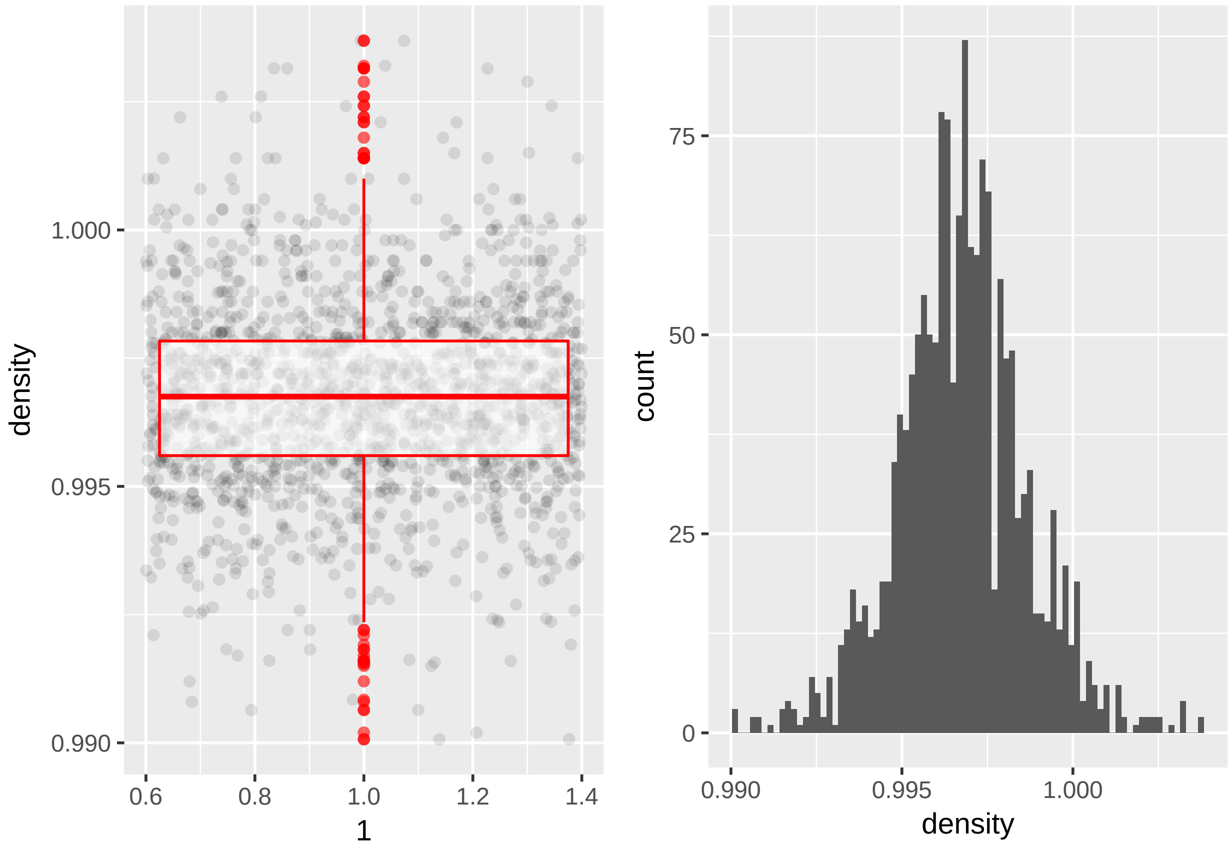
Lets investigate total sulfur dioxide feature in the dataset.



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

The distribution is a heavy skew towards lower values and also consists of a few large outliers. The following plot consisting of 99% of the data gets rid of the outliers and we can see the distribution clearly, It sill has a heavy skew. The mean of the distribution is 46.67 mg/dm^3 and the median is 38 mg/dm^3, which clearly shows that the distribution is skewed and has outliers.

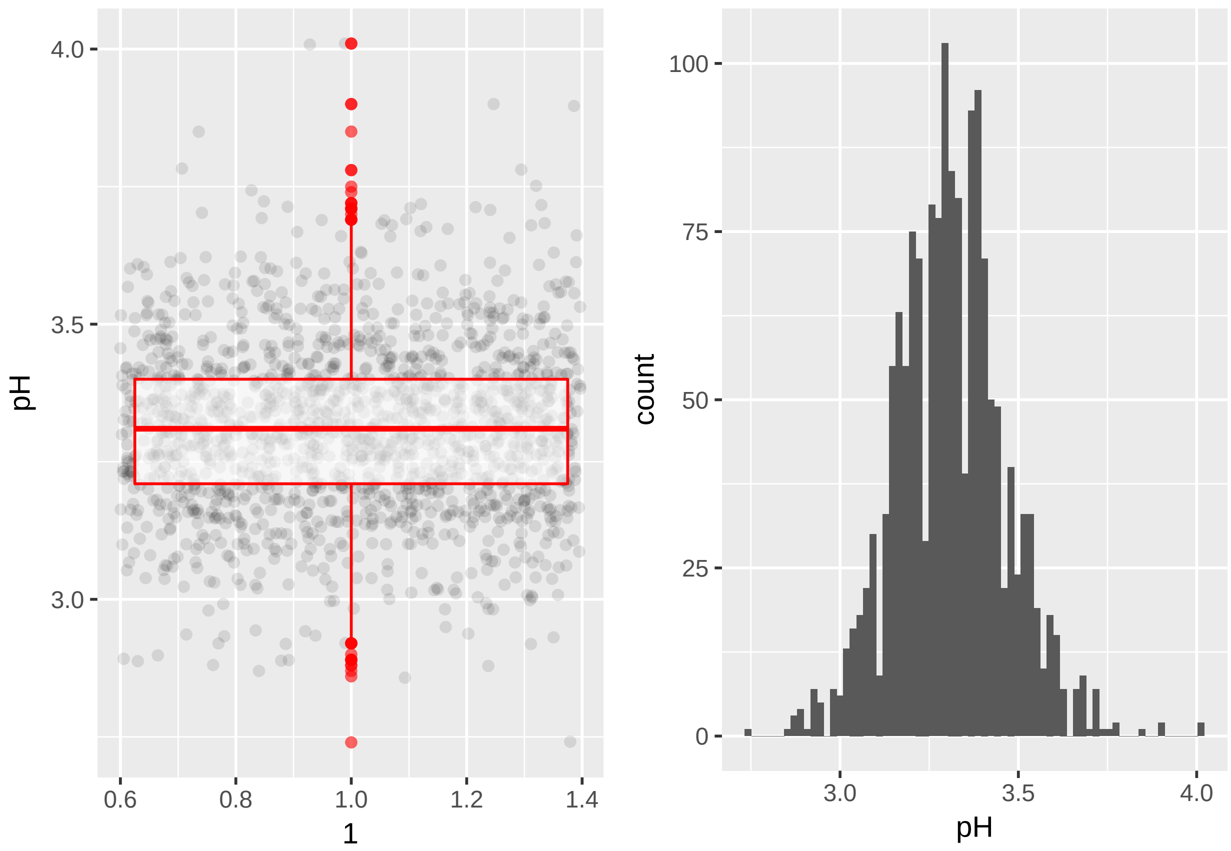
Lets investigate the density parameter in red wine dataset.



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9901  0.9956  0.9968  0.9967  0.9978  1.0037
```

The distribution looks like a normal distribution with peaks at certain values. The range of density values are not large, it varies from 0.9901 g/dmˆ3 to 1.0037 g/dmˆ3. The mean of the distribution is 0.9967 g/dmˆ3 with a median value of 0.9968 g/dmˆ3.

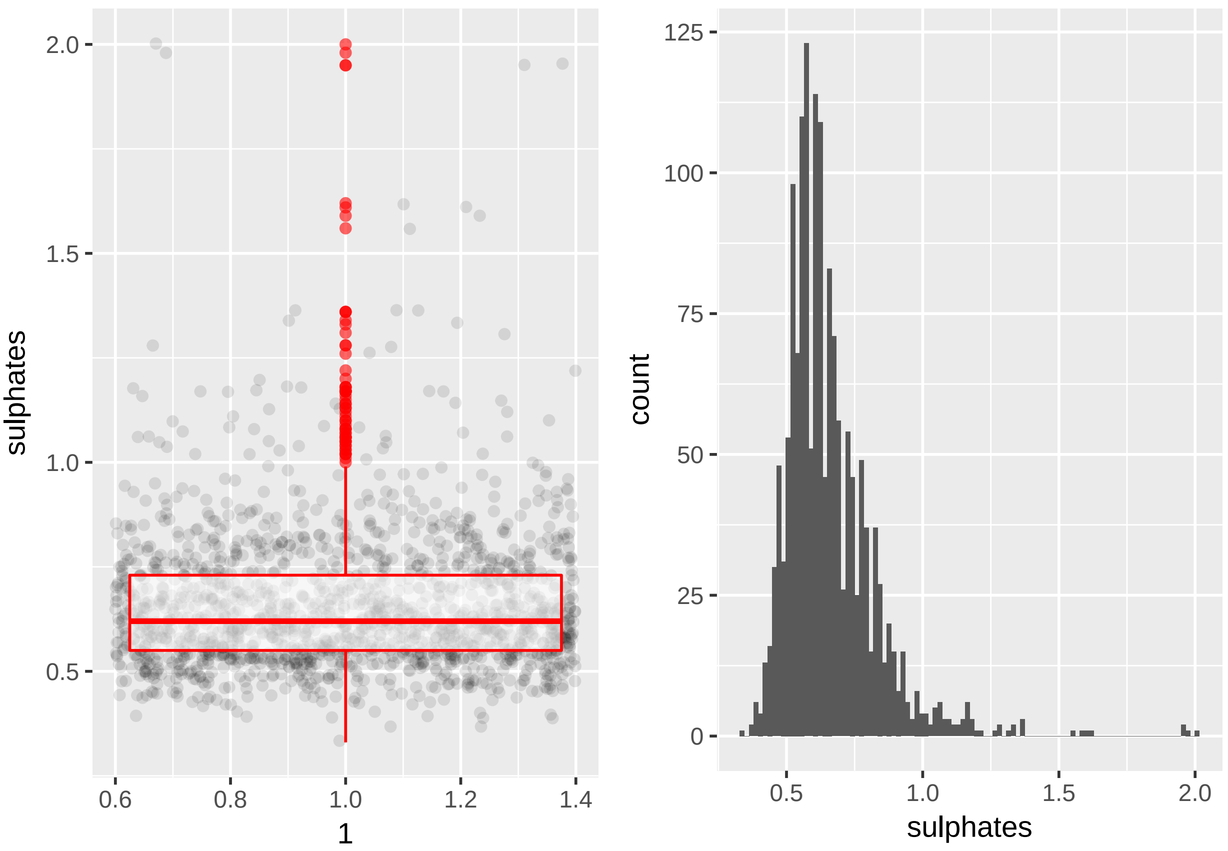
Lets investigate the pH of red wines in the dataset.



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

The distribution looks normally distributed with all the wines in the dataset begin acidic. We see some peaks between 3 and 3.5 where most of the wine data is found. The distribution has a mean of 3.311, a median of 3.310 with a maximum value of 4.010 and a minimum value of 2.74.

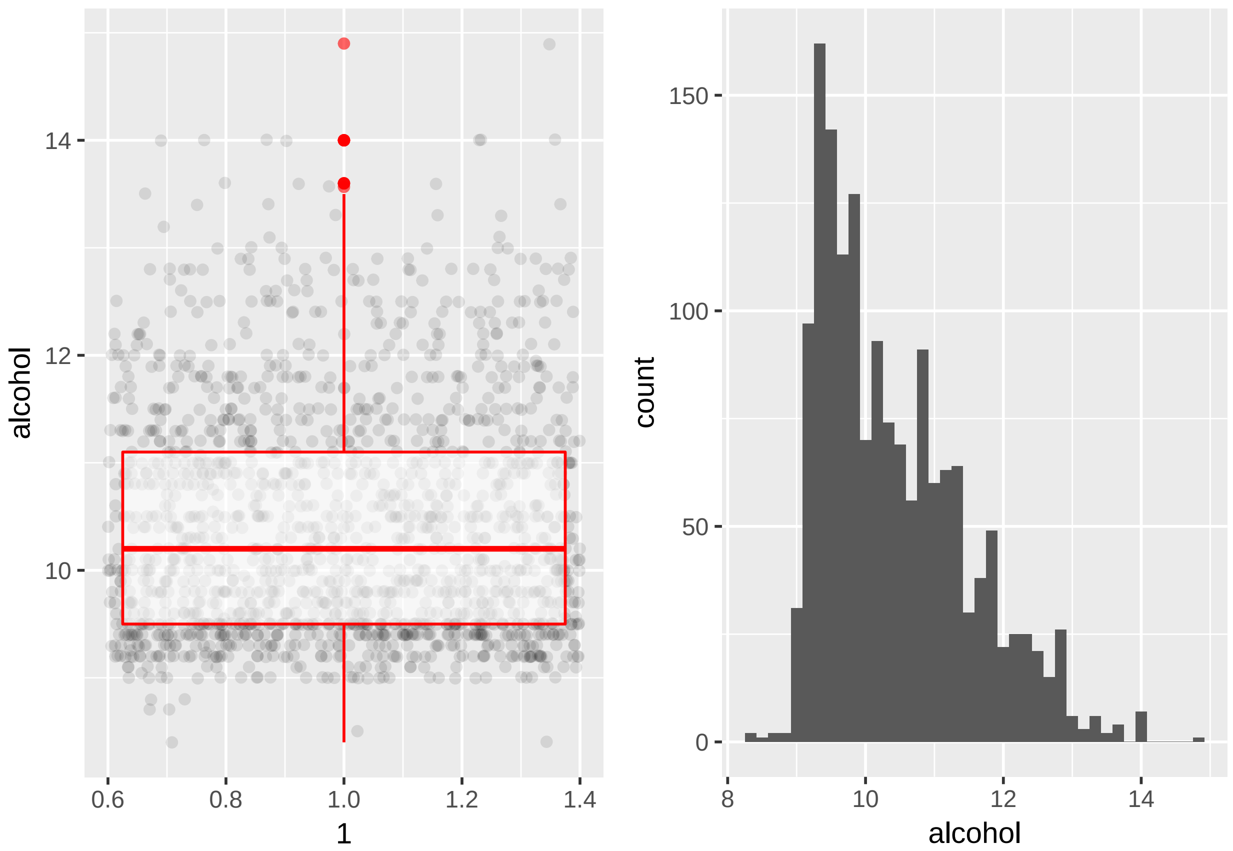
Lets investigate sulphates in the wine data.



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

The distribution has a few large outliers, to look at the distribution I will only plot 99% of the data and we see that the distribution is fairly normal. The mean of the distribution is 0.6581 g/dmˆ3 with a maximum value of 2 g/dmˆ3.

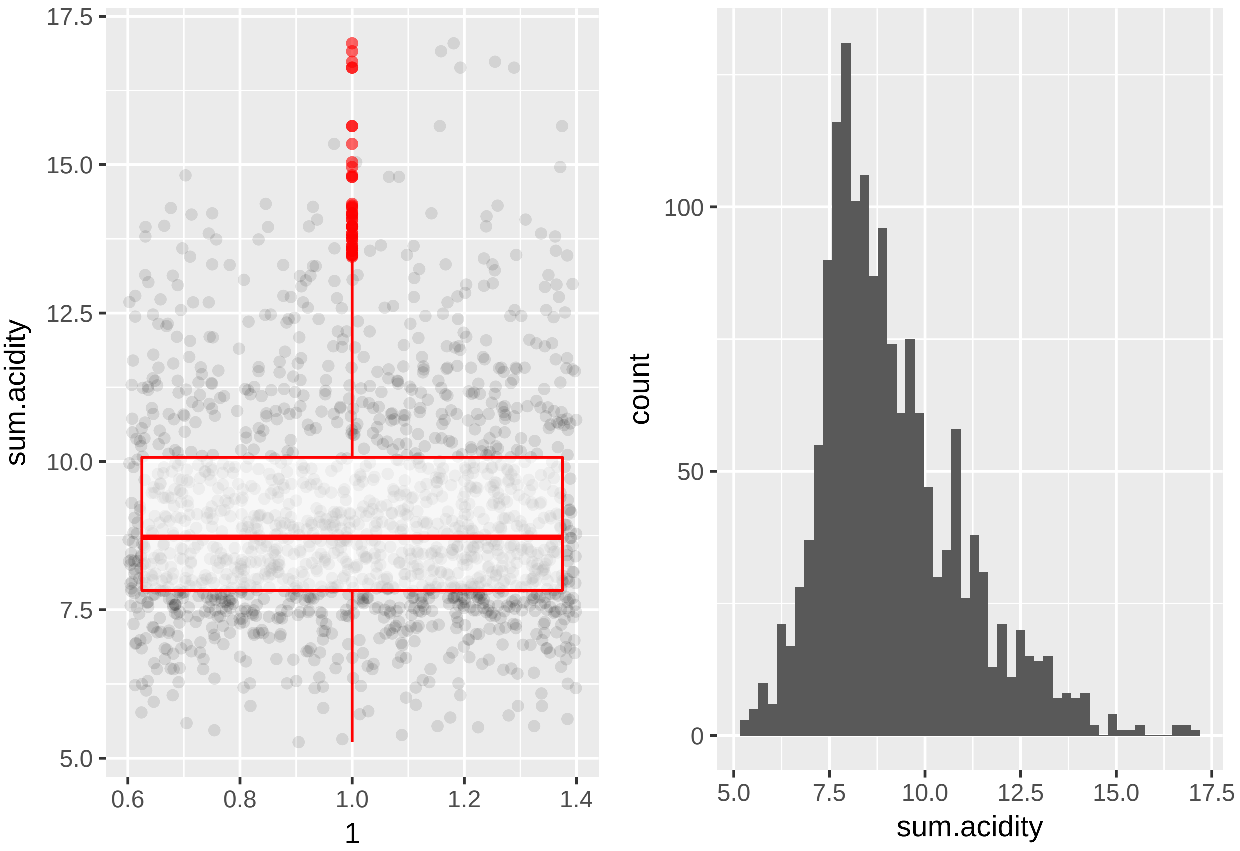
Lets investigate the alcohol percentage in red wines.



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

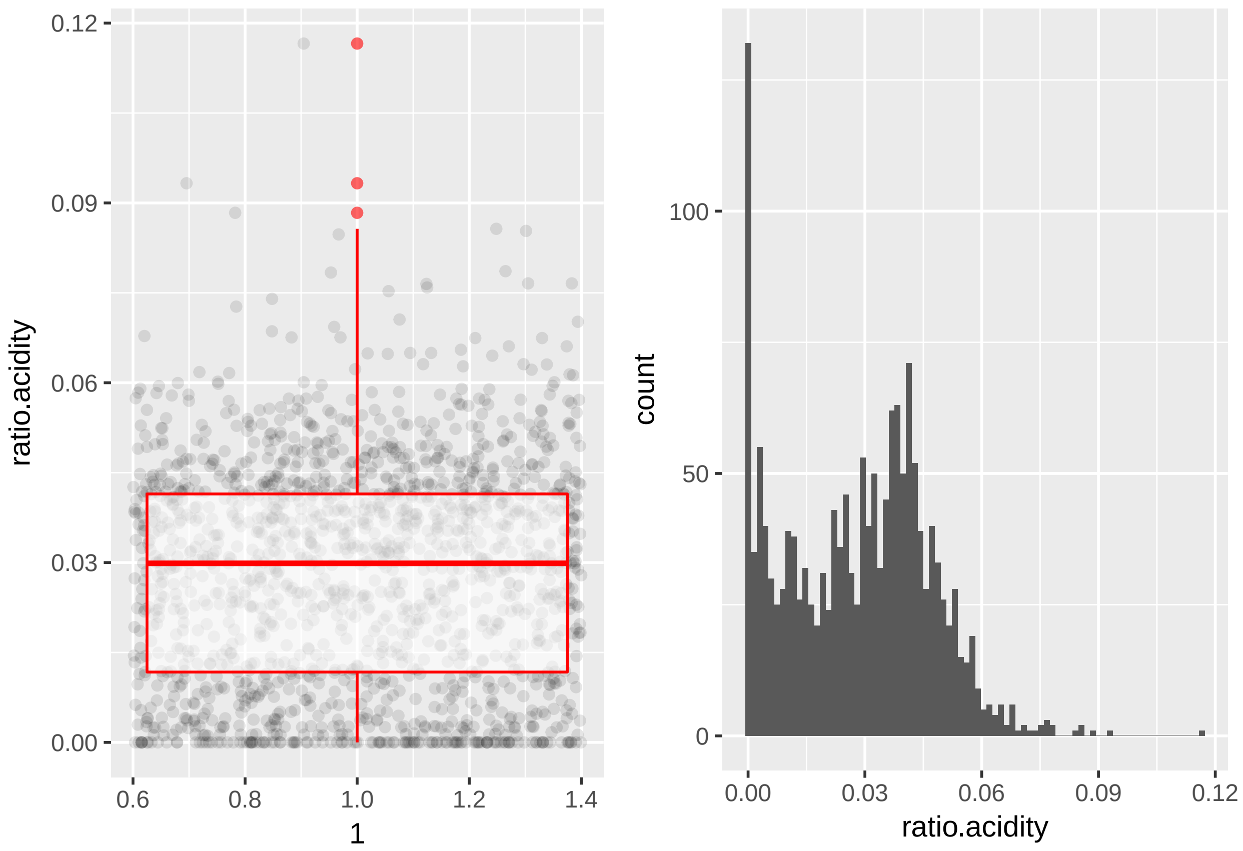
The alcohol percentage distribution looks skewed with over 75% of the wines having alcohol percentage less then 11.1%. The mean of the distribution is 10.42% with a maximum alcohol % at 14.90%.

I am creating a new feature called sum.acidity, which is the sum of volatile acidity to the fixed acidity. I wanted to see how this acidity would impact the quality of the wine.



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.270   7.827   8.720   9.118  10.070  17.045
```

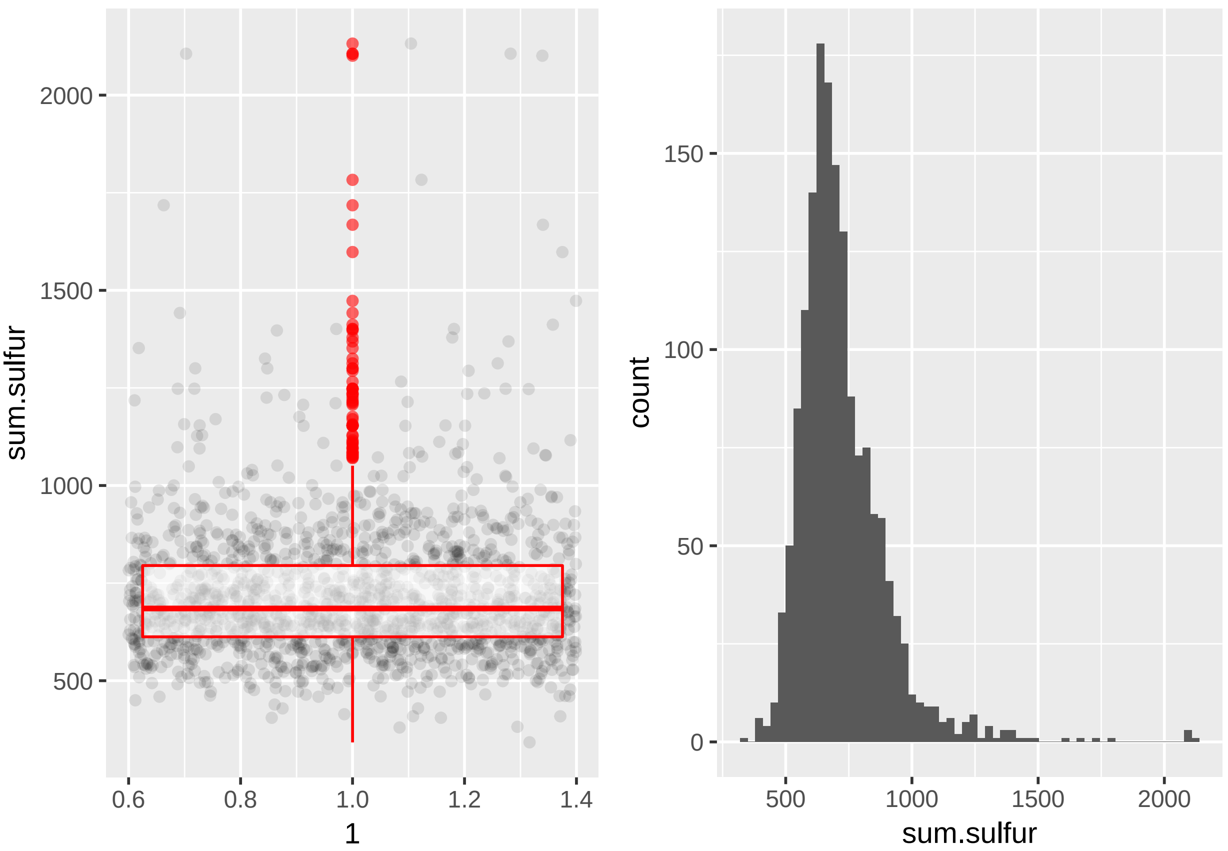
The distribution above behaves like the fixed acidity distribution.



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.01173 0.02988 0.02807 0.04145 0.11659
```

Ratio acidity distribution behaves like the citric acid distribution.

I am creating a new feature called sum.sulfur, which is the sum of all sulfur dioxide content in the wine as it can impact the taste of the wine and the quality rating. The unit of sulphates is different and must be converted before adding them up.



```
## Length  Class   Mode
##      0   NULL   NULL
```

The distribution above behaves like the sulphates feature in the wine dataset.

Univariate Analysis

What is the structure of your dataset?

The dataset consists of 1599 observations with 13 columns, of which 12 features were related to wine characteristics.

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

The most important feature of interest in the dataset is quality. We are trying to find what features affect the quality parameter in the dataset.

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

Looking at the description of each feature, I can see that volatile acidity might play an important role because higher levels can lead to an unpleasant taste. I do consume a lot of wine and from what I have experienced higher concentrations of sulphates might also cause a pungent smell or unpleasant taste.

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

I created three new variables. I wanted to look at the concentration of suphates as a whole so I added all the types of sulphates in the wine and created a new feature called sum.sulfur. I added the volatile acidity, fixed acidity and citric.acid to created a new feature called sum.acidity because these acids together contribute to acidity/pH value in the wine. I also created a third variable called ratio.acidity, the features is the ration of citric acid levels to the sum of volatile acidity and fixed acidity. I created this variable because citric acid levels determine the fresh flavor in the wine, I wanted to see the ration of freshness to acidity in the wine. It might help us better investigate the quality rating.

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?

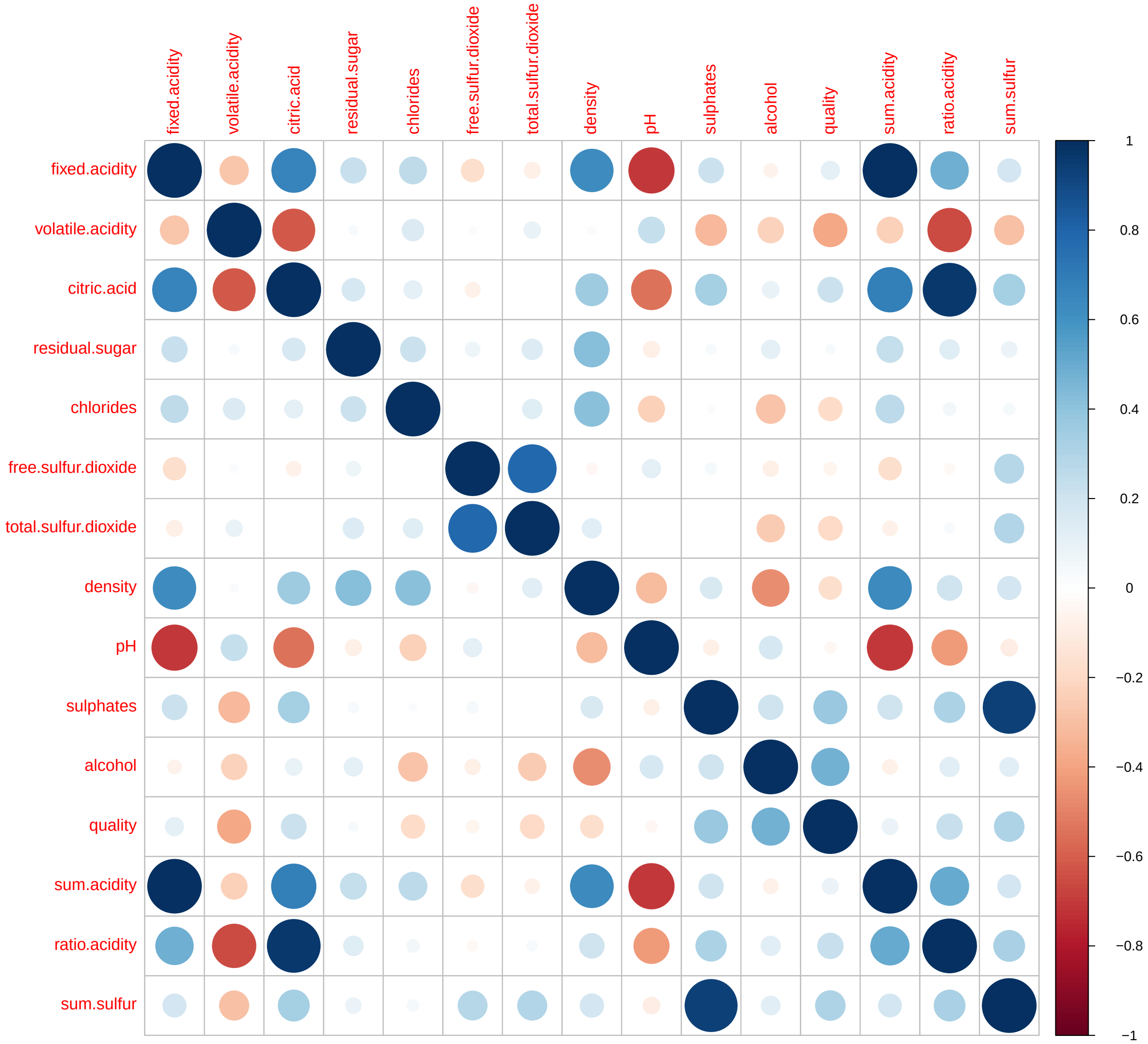
The data provided to us was already well formatted and cleaned so I did not do any further changes, the only thing I did is remove the X variable as its not important to our analysis.

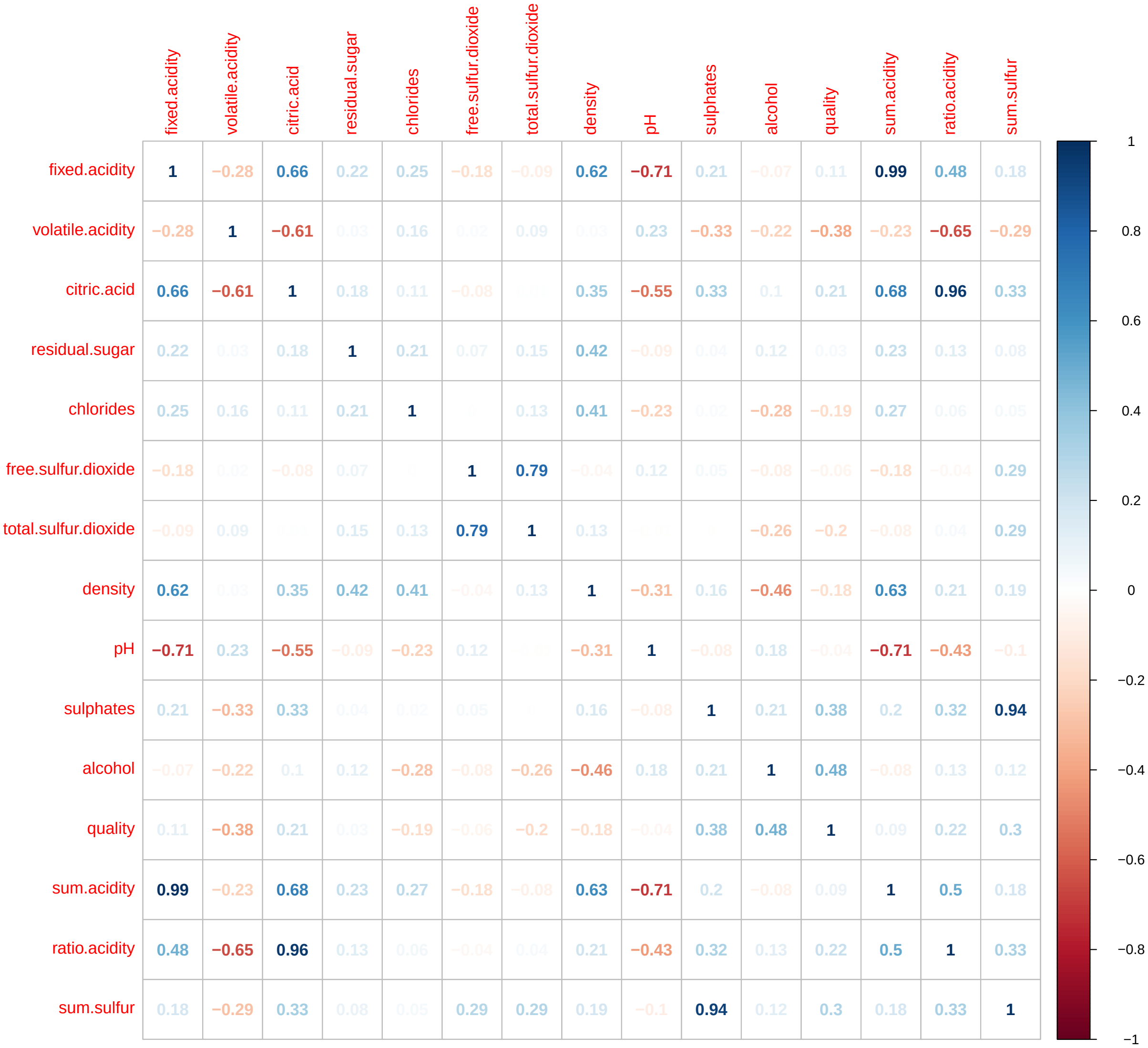
The distribution that caught my eye is the citric acid levels in the wine. The distribution has two large peaks at 0 and 0.49.

Bivariate Plots Section

The best way to look at relationships between two variables in by plotting a correlation matrix plot with all the features. We are interested in the quality rating feature to by looking at the correlation plot we can see how each feature correlates with the quality rating. It might be a simple way of looking at the relationship but its effective for Bivariate analysis.

```r
library(corrplot)
wine_df <- subset(df, select = - X)
corr_mat=cor(wine_df,method="s")
corrplot(corr_mat)
```

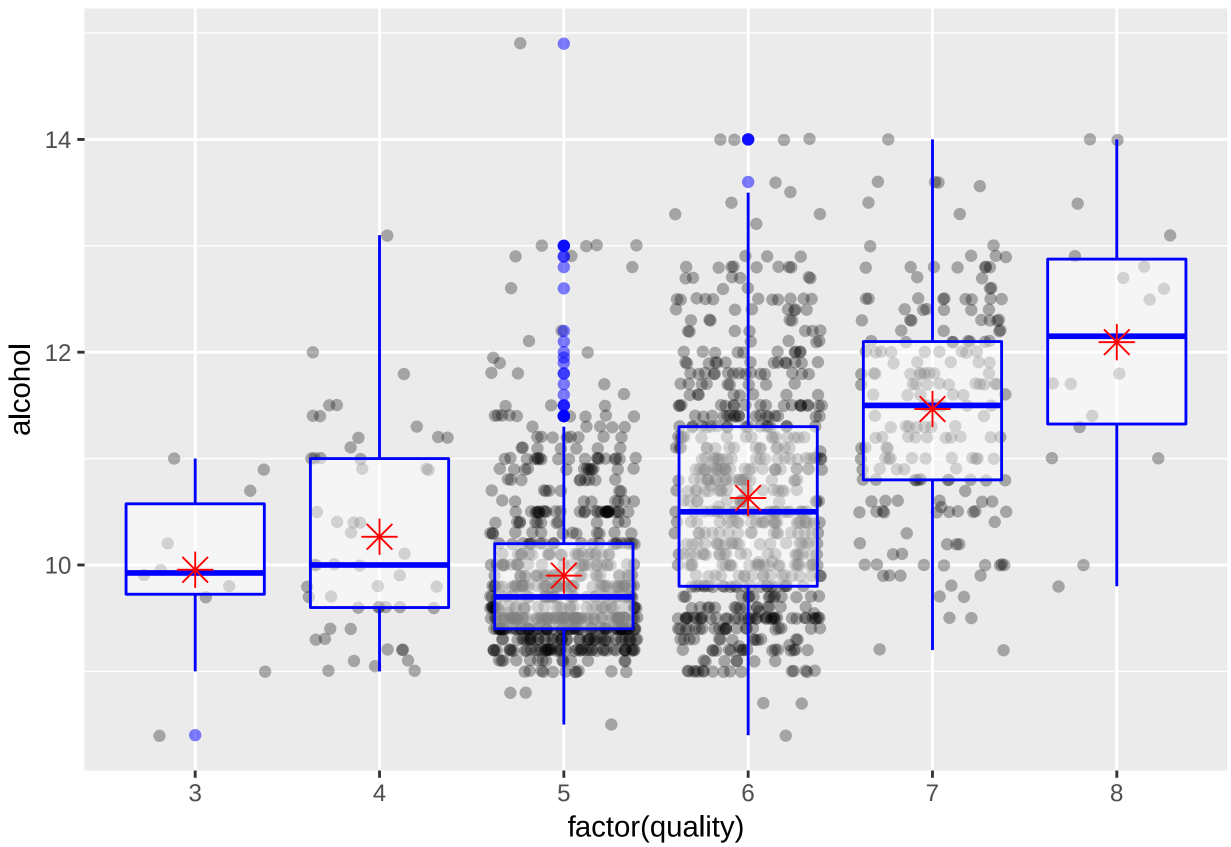

fixed.acidity
volatile.acidity
citric.acid
residual.sugar
chlorides
free.sulfur.dioxide
total.sulfur.dioxide
density
pH
sulphates
alcohol
quality
sum.acidity
ratio.acidity
sum.sulfur
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1



	fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides	free.sulfur.dioxide	total.sulfur.dioxide	density	pH	sulphates	alcohol	quality	sum.acidity	ratio.acidity	sum.sulfur
fixed.acidity	1	−0.28	0.66	0.22	0.25	−0.18	−0.09	0.62	−0.71	0.21	−0.07	0.11	0.99	0.48	0.18
volatile.acidity	−0.28	1	−0.61	0.03	0.16	0.02	0.09	0.03	0.23	−0.33	−0.22	−0.38	−0.23	−0.65	−0.29
citric.acid	0.66	−0.61	1	0.18	0.11	−0.08	0.01	0.35	−0.55	0.33	0.1	0.21	0.68	0.96	0.33
residual.sugar	0.22	0.03	0.18	1	0.21	0.07	0.15	0.42	−0.09	0.04	0.12	0.03	0.23	0.13	0.08
chlorides	0.25	0.16	0.11	0.21	1	0	0.13	0.41	−0.23	0.02	−0.28	−0.19	0.27	0.06	0.05
free.sulfur.dioxide	−0.18	0.02	−0.08	0.07	0	1	0.79	−0.04	0.12	0.05	−0.08	−0.06	−0.18	−0.04	0.29
total.sulfur.dioxide	−0.09	0.09	0.01	0.15	0.13	0.79	1	0.13	−0.01	0	−0.26	−0.2	−0.08	0.04	0.29
density	0.62	0.03	0.35	0.42	0.41	−0.04	0.13	1	−0.31	0.16	−0.46	−0.18	0.63	0.21	0.19
pH	−0.71	0.23	−0.55	−0.09	−0.23	0.12	−0.01	−0.31	1	−0.08	0.18	−0.04	−0.71	−0.43	−0.1
sulphates	0.21	−0.33	0.33	0.04	0.02	0.05	0	0.16	−0.08	1	0.21	0.38	0.2	0.32	0.94
alcohol	−0.07	−0.22	0.1	0.12	−0.28	−0.08	−0.26	−0.46	0.18	0.21	1	0.48	−0.08	0.13	0.12
quality	0.11	−0.38	0.21	0.03	−0.19	−0.06	−0.2	−0.18	−0.04	0.38	0.48	1	0.09	0.22	0.3
sum.acidity	0.99	−0.23	0.68	0.23	0.27	−0.18	−0.08	0.63	−0.71	0.2	−0.08	0.09	1	0.5	0.18
ratio.acidity	0.48	−0.65	0.96	0.13	0.06	−0.04	0.04	0.21	−0.43	0.32	0.13	0.22	0.5	1	0.33
sum.sulfur	0.18	−0.29	0.33	0.08	0.05	0.29	0.29	0.19	−0.1	0.94	0.12	0.3	0.18	0.33	1

Looking at the correlation plots and focusing on the quality rating we can see that the features don't really have a very high correlation. Alcohol content has a correlation of 0.48 followed by suphates and volatile acidity with a correlation of 0.38 and -0.38 respectively. Citric acid level has a correlation of 0.21 and the feature I created ratio.acidity as correlation of 0.22 with the quality rating. Sum.sulfur has a correlation of 0.3, which is a feature I created.

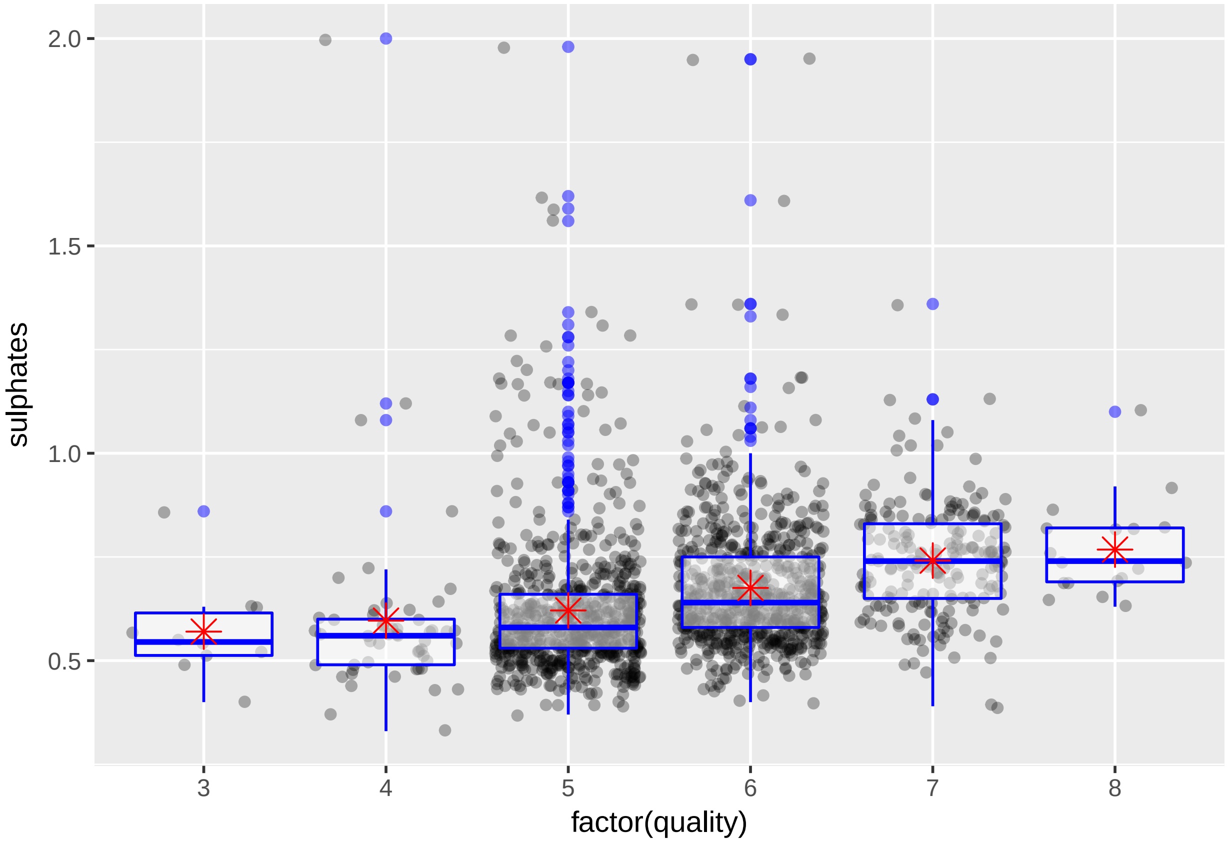
Lets first investigate the feature alcohol, because it has the highest correlation with quality and see how it looks like and also print out the correlation. Box plot might give us a better perspective of the correlation between the two parameters, by treating the quality feature as a factor.



```
## factor(quality): 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   8.400   9.725   9.925   9.955  10.575  11.000
## --------------------------------------------------------
## factor(quality): 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.00    9.60   10.00   10.27   11.00   13.10
## --------------------------------------------------------
## factor(quality): 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     8.5     9.4     9.7     9.9    10.2    14.9
## --------------------------------------------------------
## factor(quality): 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.80   10.50   10.63   11.30   14.00
## --------------------------------------------------------
## factor(quality): 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.20   10.80   11.50   11.47   12.10   14.00
## --------------------------------------------------------
## factor(quality): 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.80   11.32   12.15   12.09   12.88   14.00
```

Looking at the box plots we can see that the highest quality wine have a high average level of alcohol content, which can also be inferred from the summary. The median percentage if alcohol in wines with a rating of 8 is 12.15%, where as the median percentage of alcohol in wines with a rating of 3 is 9.925%. A few ouliers are present in wines with quality rating of 5 and 6, that is even though the alcohol content was higher the rating was mediocre. Hence we cannot just decide that alcohol is responsible for the quality rating. Other factors might be responsible for it.

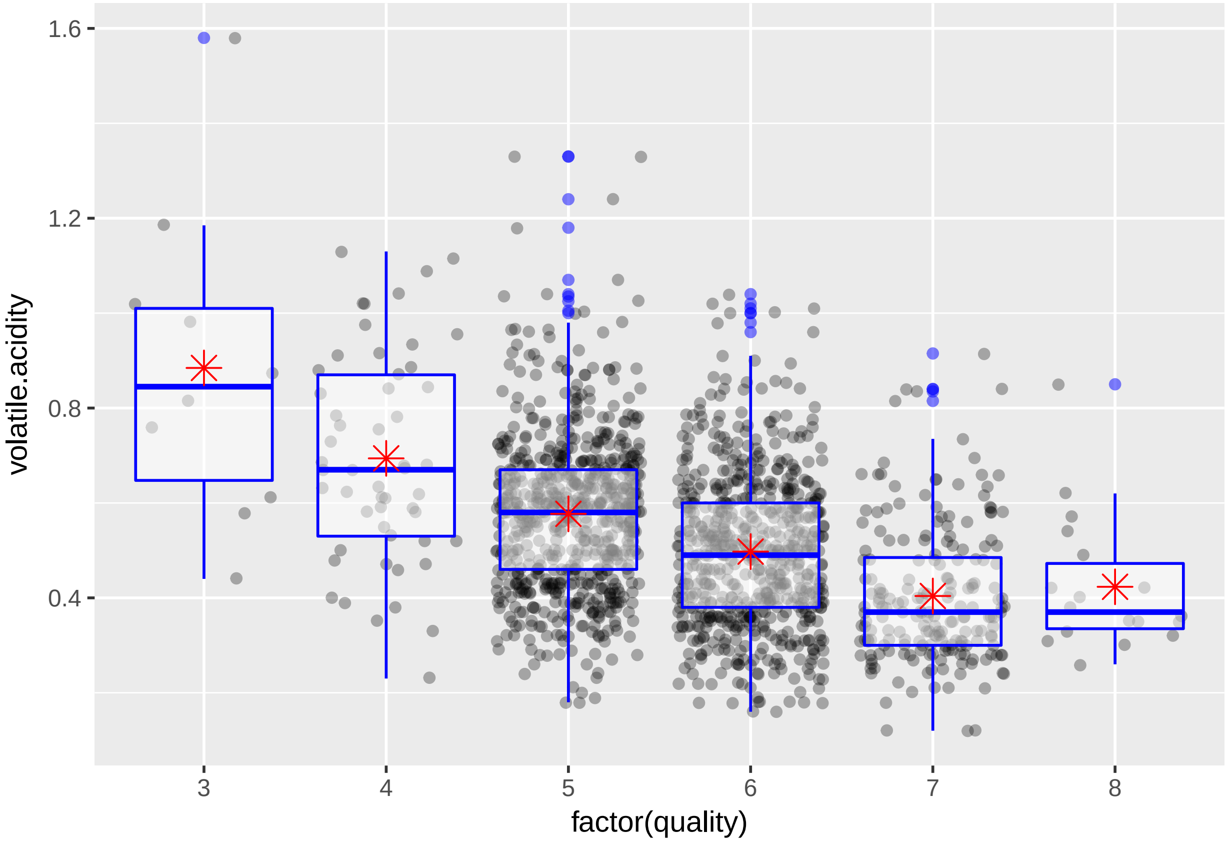
Now lets investigate the sulphate concentration in the wine. Box plot might give us a better perspective of the correlation between the two parameters, by treating the quality feature as a factor.



```
## factor(quality): 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.4000  0.5125  0.5450  0.5700  0.6150  0.8600
## --------------------------------------------------------
## factor(quality): 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3300  0.4900  0.5600  0.5964  0.6000  2.0000
## --------------------------------------------------------
## factor(quality): 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.370   0.530   0.580   0.621   0.660   1.980
## --------------------------------------------------------
## factor(quality): 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.4000  0.5800  0.6400  0.6753  0.7500  1.9500
## --------------------------------------------------------
## factor(quality): 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3900  0.6500  0.7400  0.7413  0.8300  1.3600
## --------------------------------------------------------
## factor(quality): 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.6300  0.6900  0.7400  0.7678  0.8200  1.1000
```

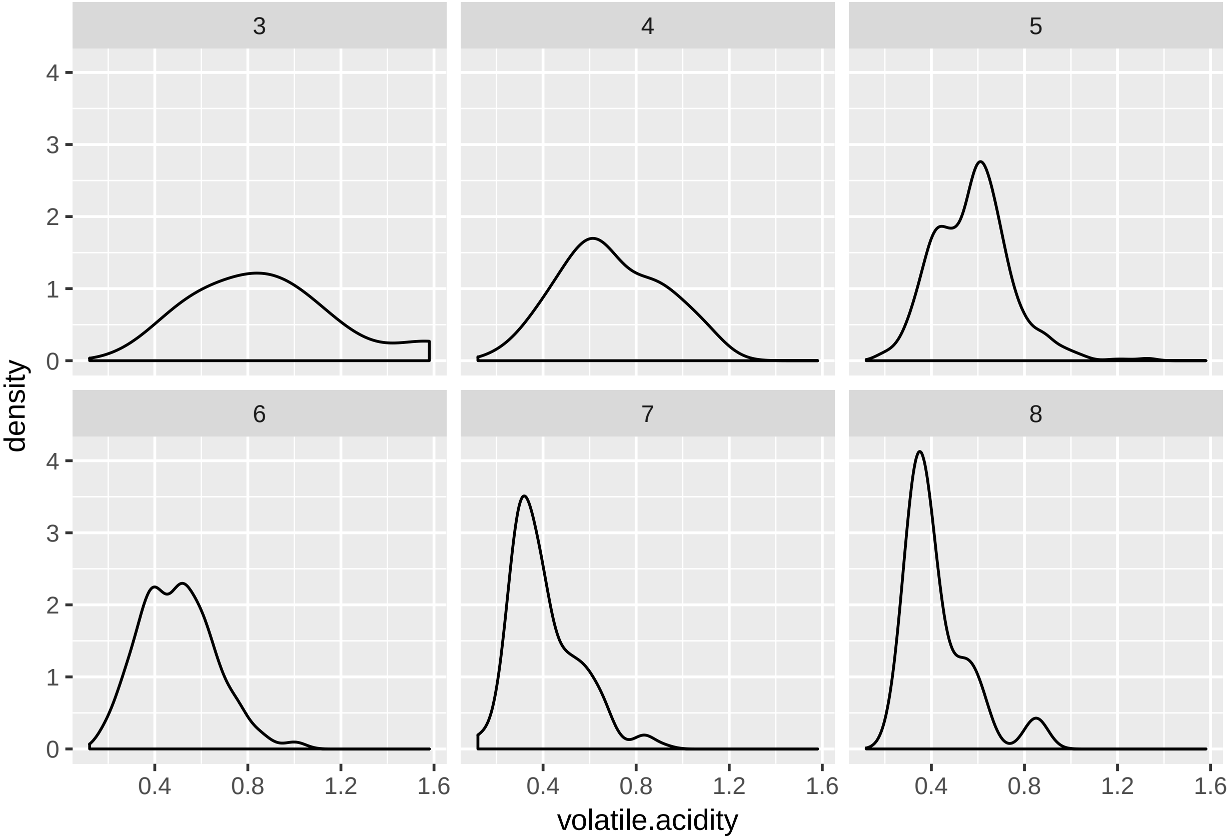
Well looks like higher rated wines had more median sulphates amounts then the other wines, which was against my intuition. We can see a lot of outliers in the data. This might be the reason we see some correlation in the first place. If we take a look at the summery of the data, we can see that the wine with quality rating of 4, 5 and 6 have a maximum value of 2 g / dm3, but the higher rated wines had a lower maximum sulphate content in that category.

Now lets investigate the volatile acidity content in the wine. Box plot might give us a better perspective of the correlation between the two parameters, by treating the quality feature as a factor.



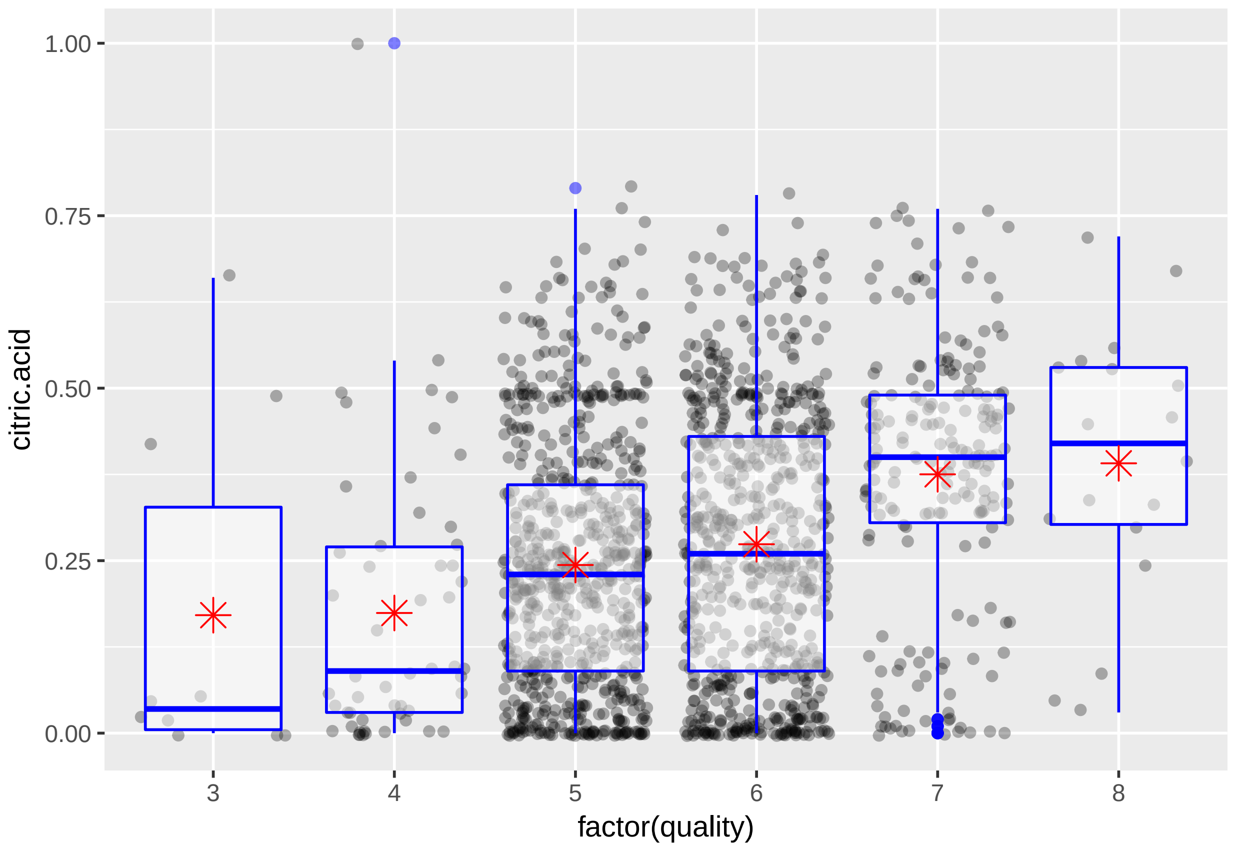
```
## factor(quality): 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.4400  0.6475  0.8450  0.8845  1.0100  1.5800 
## -------------------------------------------------------- 
## factor(quality): 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.230   0.530   0.670   0.694   0.870   1.130 
## -------------------------------------------------------- 
## factor(quality): 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.180   0.460   0.580   0.577   0.670   1.330 
## -------------------------------------------------------- 
## factor(quality): 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.1600  0.3800  0.4900  0.4975  0.6000  1.0400
## --------------------------------------------------------
## factor(quality): 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1200  0.3000  0.3700  0.4039  0.4850  0.9150
## --------------------------------------------------------
## factor(quality): 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2600  0.3350  0.3700  0.4233  0.4725  0.8500
```

This plot makes it easier to infer the correlation. from the description of volatile acidity more concentration lends to unpleasant taste, we can clearly see that here. The wines with ratings 7 and 8 have significantly less volatile acids then the other wines. If we look at the median values of the volatile acid content in each quality category we can see a drop as the quality increases. This can also be seen clearly in a density plot, lets plot a density plot.



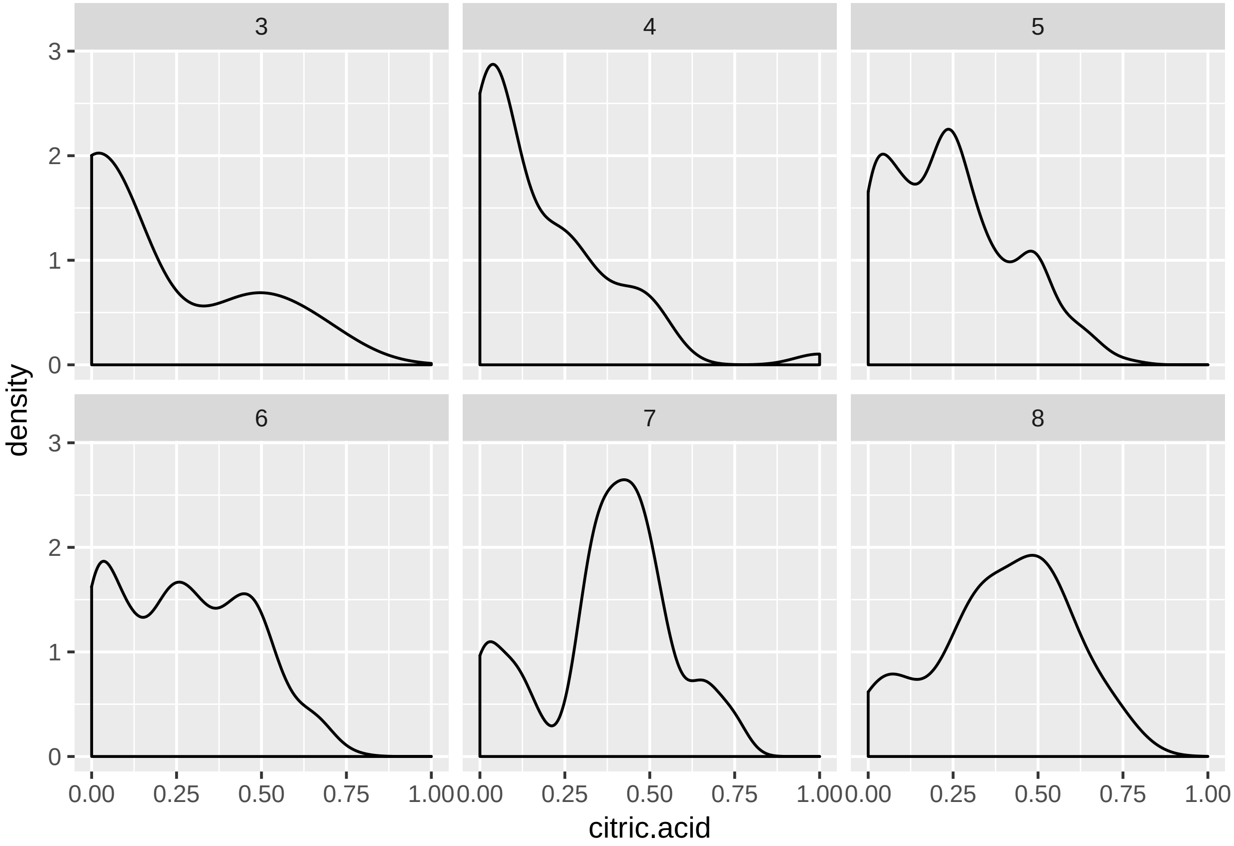
We can clearly see that the peaks of the distribution of volatile acidity for various wines of different quality and the peaks of the wines with higher qualities are more to the lower concentration levels.

Now lets investigate the citric acid content in the wine. Box plot might give us a better perspective of the correlation between the two parameters, by treating the quality feature as a factor.

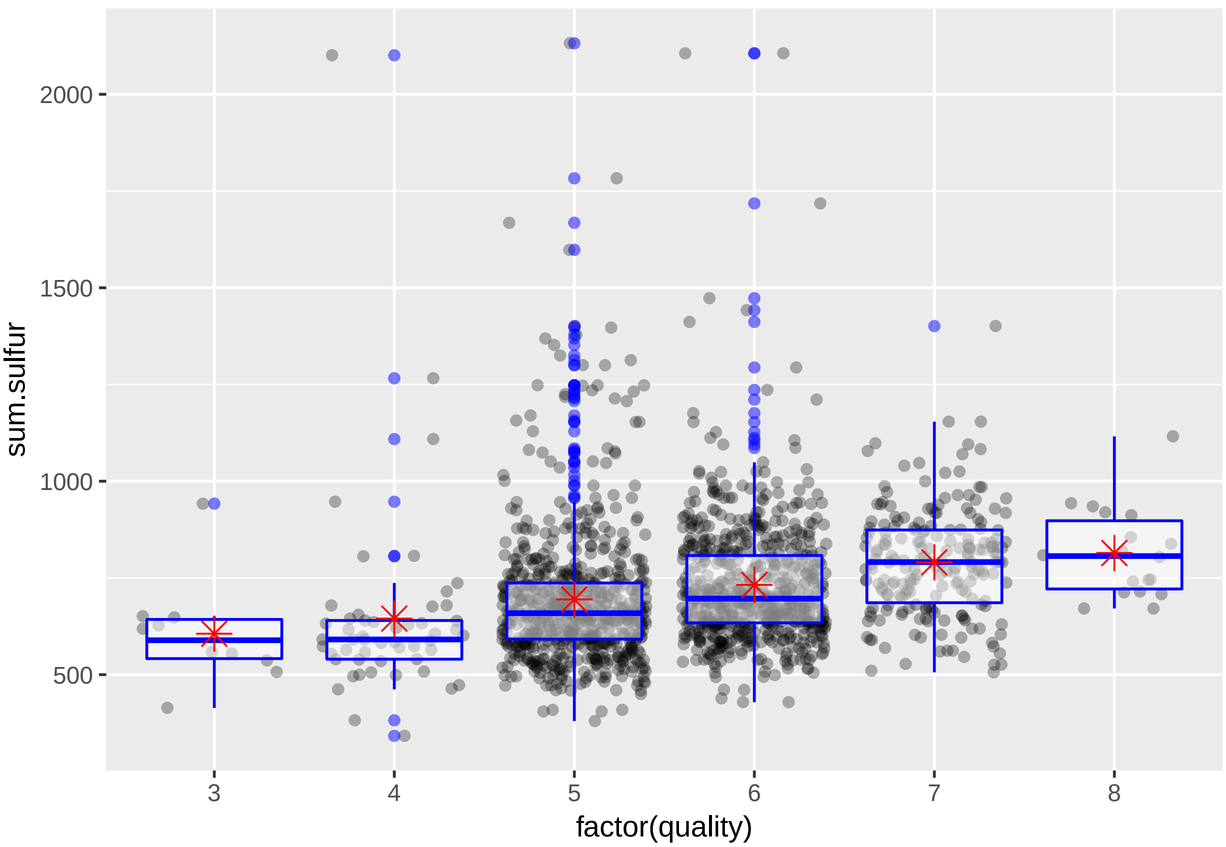


```
## factor(quality): 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0050  0.0350  0.1710  0.3275  0.6600
## --------------------------------------------------------
## factor(quality): 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0300  0.0900  0.1742  0.2700  1.0000
## --------------------------------------------------------
## factor(quality): 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0900  0.2300  0.2437  0.3600  0.7900
## --------------------------------------------------------
## factor(quality): 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0900  0.2600  0.2738  0.4300  0.7800
## --------------------------------------------------------
## factor(quality): 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.3050  0.4000  0.3752  0.4900  0.7600
## --------------------------------------------------------
## factor(quality): 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0300  0.3025  0.4200  0.3911  0.5300  0.7200
```

This plot makes it easier to infer the correlation. from the description of citric acid levels more concentration lends to freshness in taste, we can clearly see that here. The wines with ratings 7 and 8 have higher citric acids then the other wines. If we look at the median values of the citric acid content in each quality category we can see a rise as the quality increases. This can also be seen clearly in a density plot, lets plot a density plot.



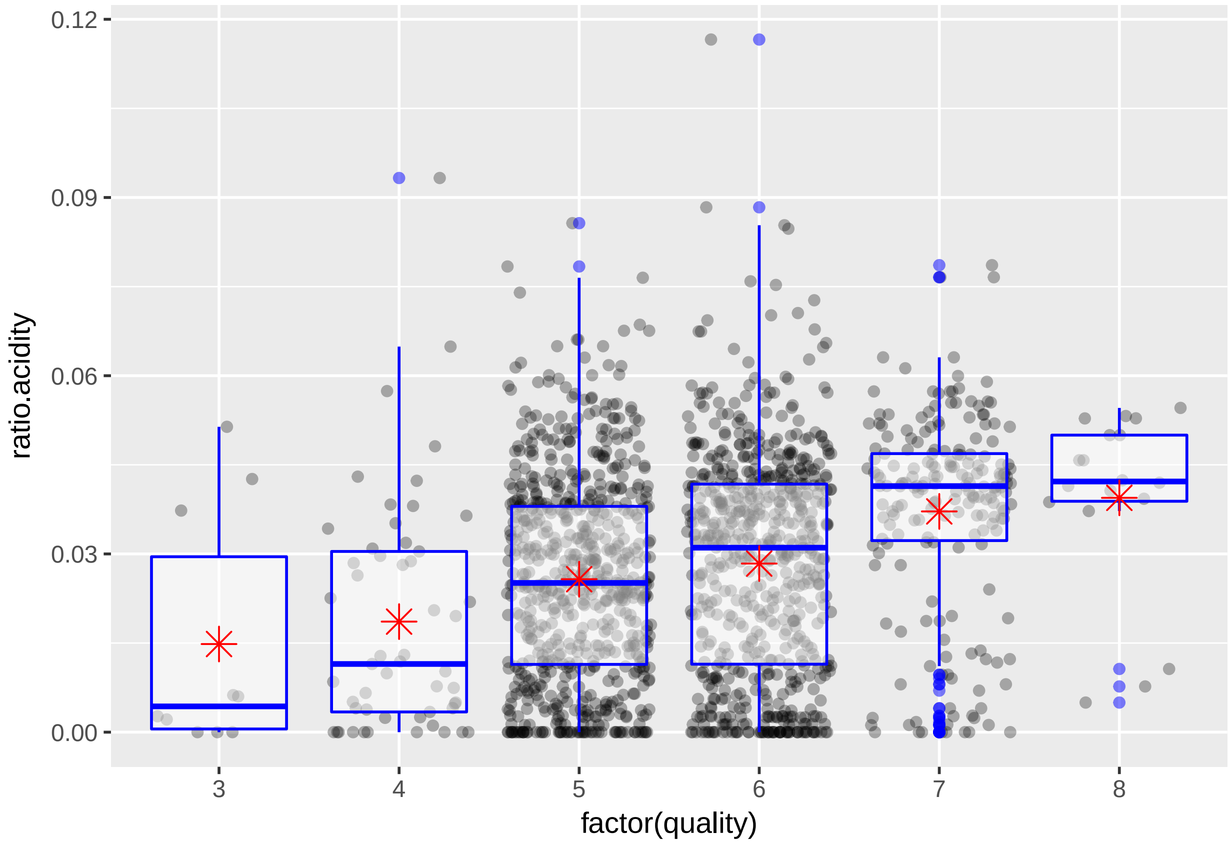
Lets investigate a feature I created sum.sulfur.



```
## factor(quality): 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   414.0   541.5   589.0   605.9   642.8   942.0
## --------------------------------------------------------
## factor(quality): 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   342.0   540.0   591.0   644.9   640.0  2101.0
## --------------------------------------------------------
## factor(quality): 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   380.0   592.0   659.0   694.5   737.0  2132.0
## --------------------------------------------------------
## factor(quality): 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   429.0   634.0   696.5   731.9   808.0  2106.0
## --------------------------------------------------------
## factor(quality): 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   506.0   686.0   791.0   790.3   874.0  1401.0
## --------------------------------------------------------
## factor(quality): 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   671.0   721.5   806.5   814.5   897.8  1116.0
```

The behavior of this feature is similar to the suphates feature in the wine, no surprises here.

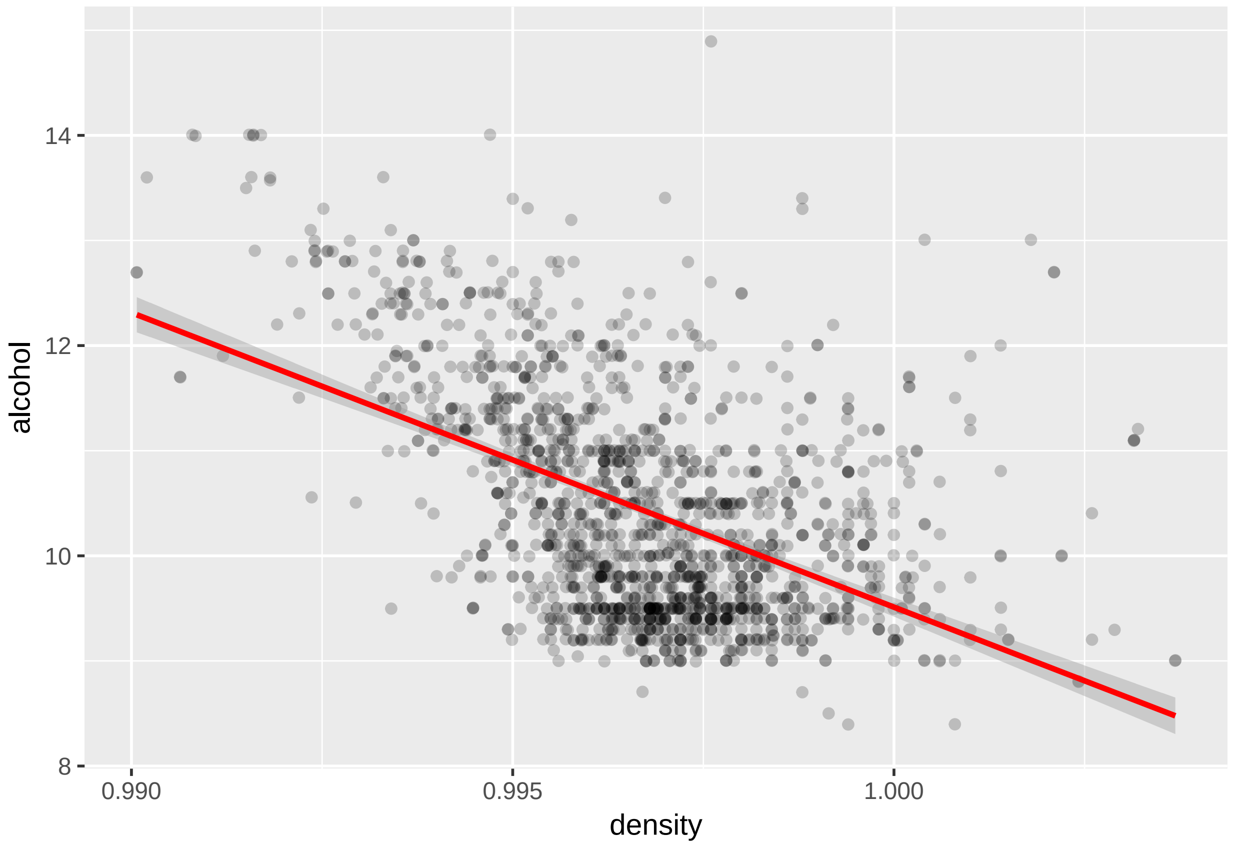
Lets investigate the ratio acidity feature I created to capture the features that cause acidity (sum.acidity) and the feature that gives freshness to the wine (citric.acid).

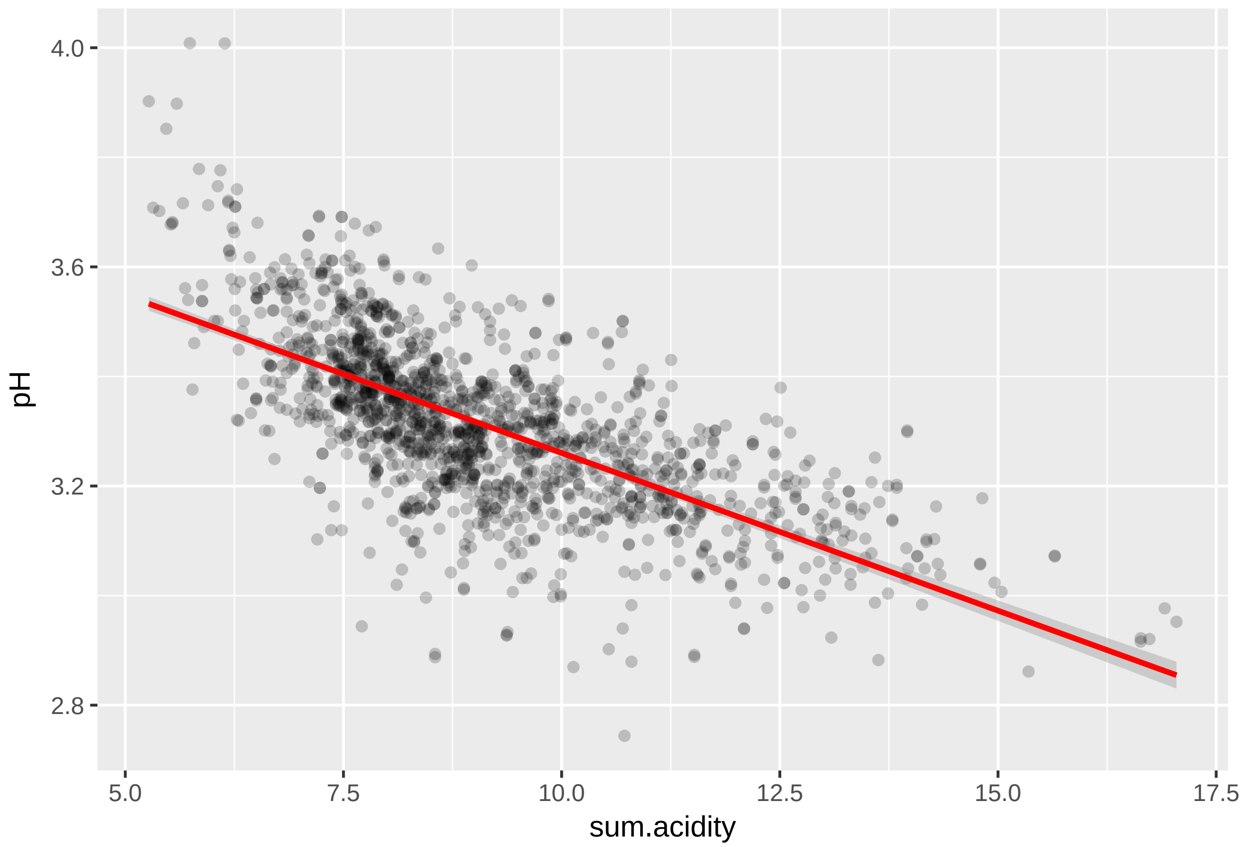


```
## factor(quality): 3
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.0000000 0.0005353 0.0043381 0.0148359 0.0295328 0.0514019
## --------------------------------------------------------
## factor(quality): 4
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.000000 0.003403 0.011474 0.018609 0.030418 0.093284
## --------------------------------------------------------
## factor(quality): 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.01141 0.02512 0.02575 0.03800 0.08568
## --------------------------------------------------------
## factor(quality): 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.01144 0.03105 0.02838 0.04175 0.11659
## --------------------------------------------------------
## factor(quality): 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.03225 0.04142 0.03716 0.04689 0.07861
## --------------------------------------------------------
## factor(quality): 8
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.004983 0.038866 0.042194 0.039445 0.050000 0.054581
```

We can see that higher ratios of citric acid concentration to the entire acid levels in the wine get better quality ratings. Wines with a quality of 8 have the highest ration of citric acid to the tonal acid content in the wine at about .0422.

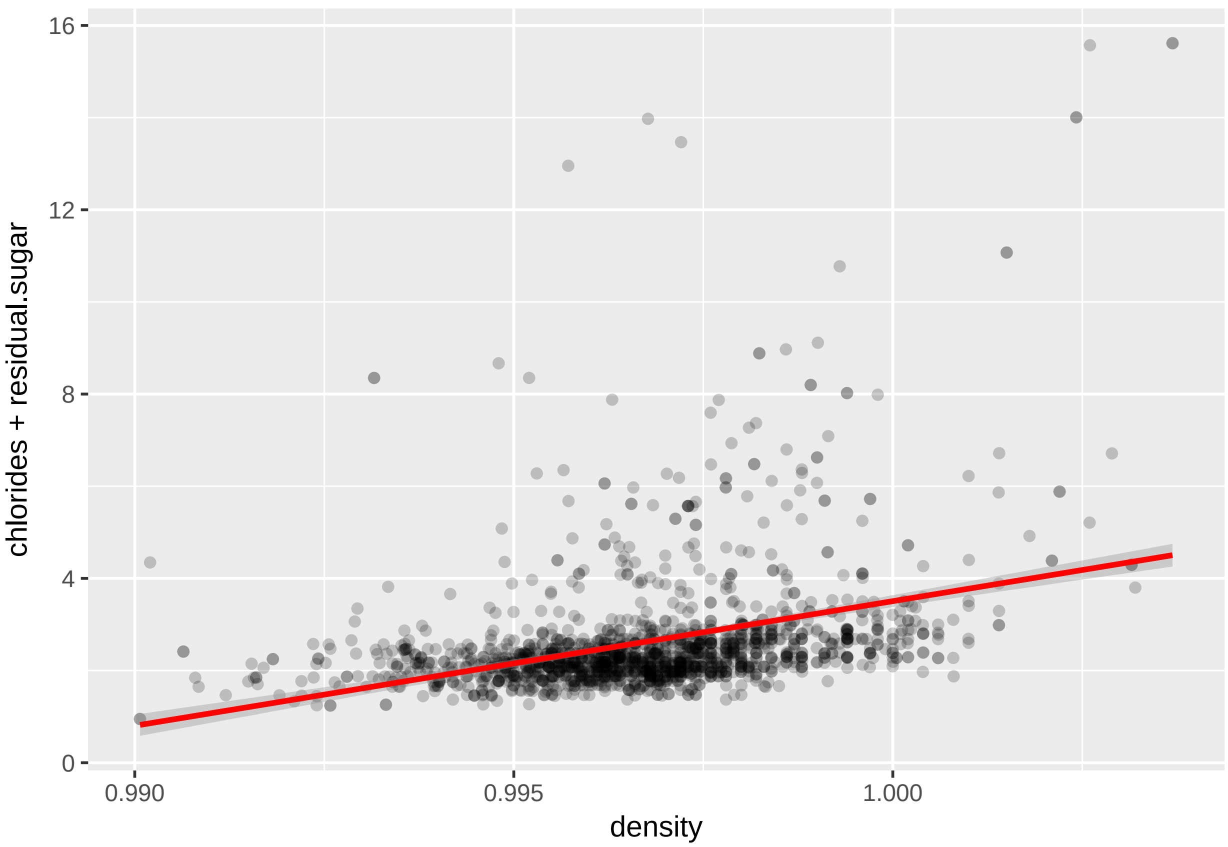
These factors had a correlation values of greater then .2 with quality. I also want to investigate other variables correlating very strongly with each other. I will investigate a few, there are a few obvious ones which we easily observe and expect, like negative correlation in density vs alcohol and negative correlation in sum.acidity vs pH.

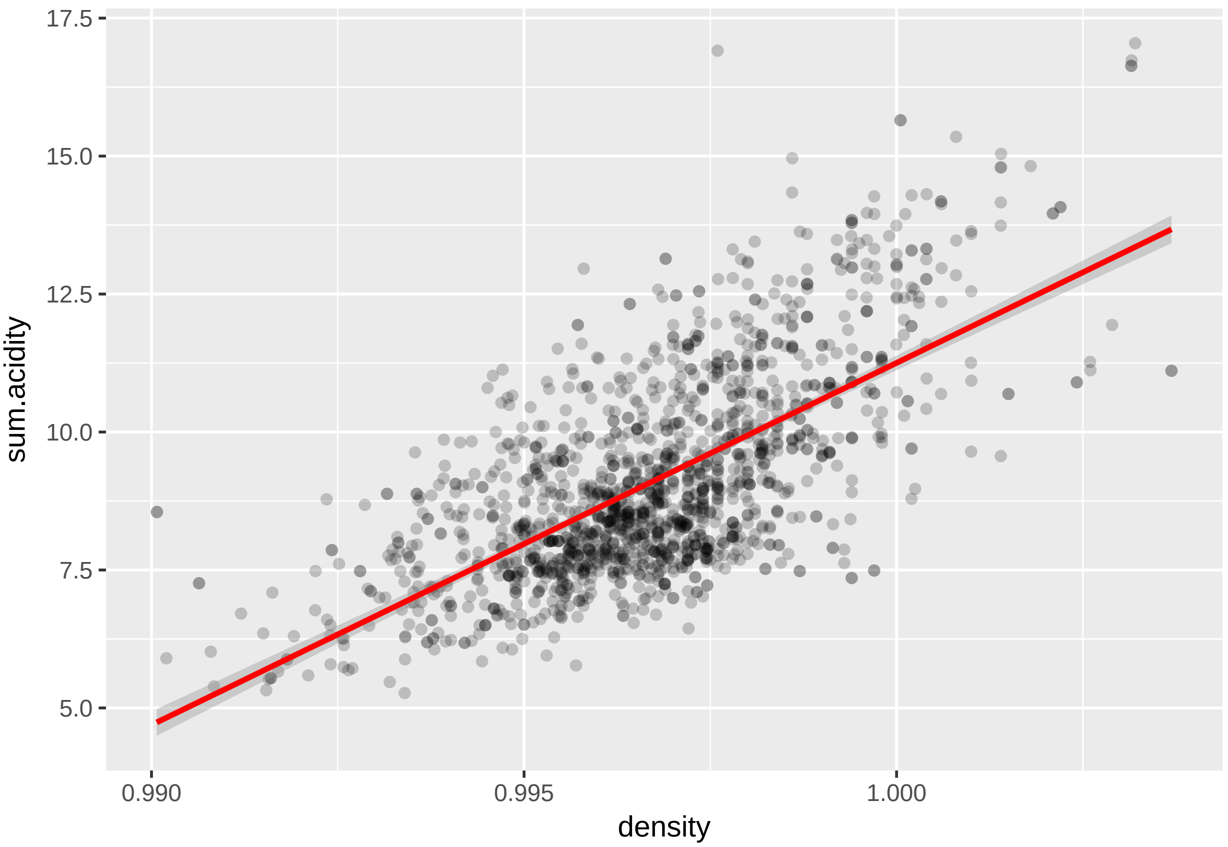


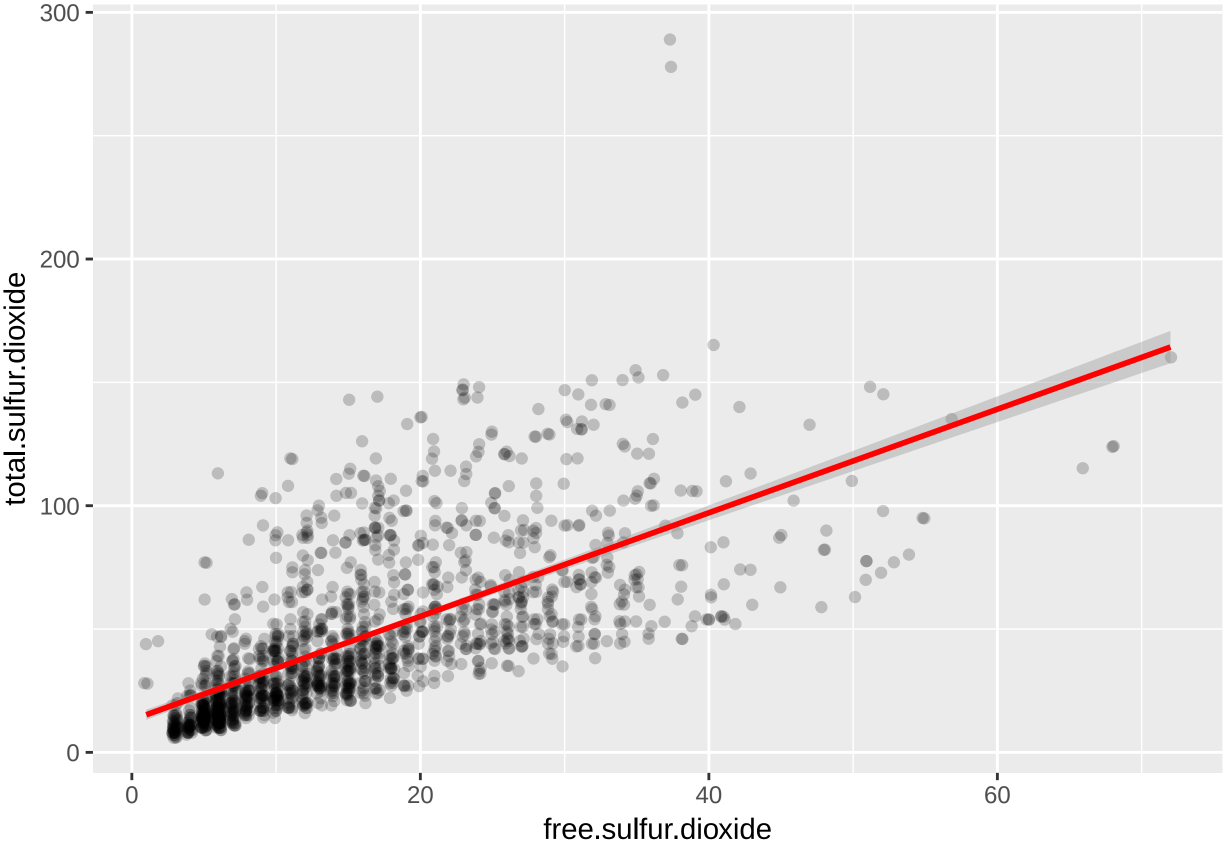


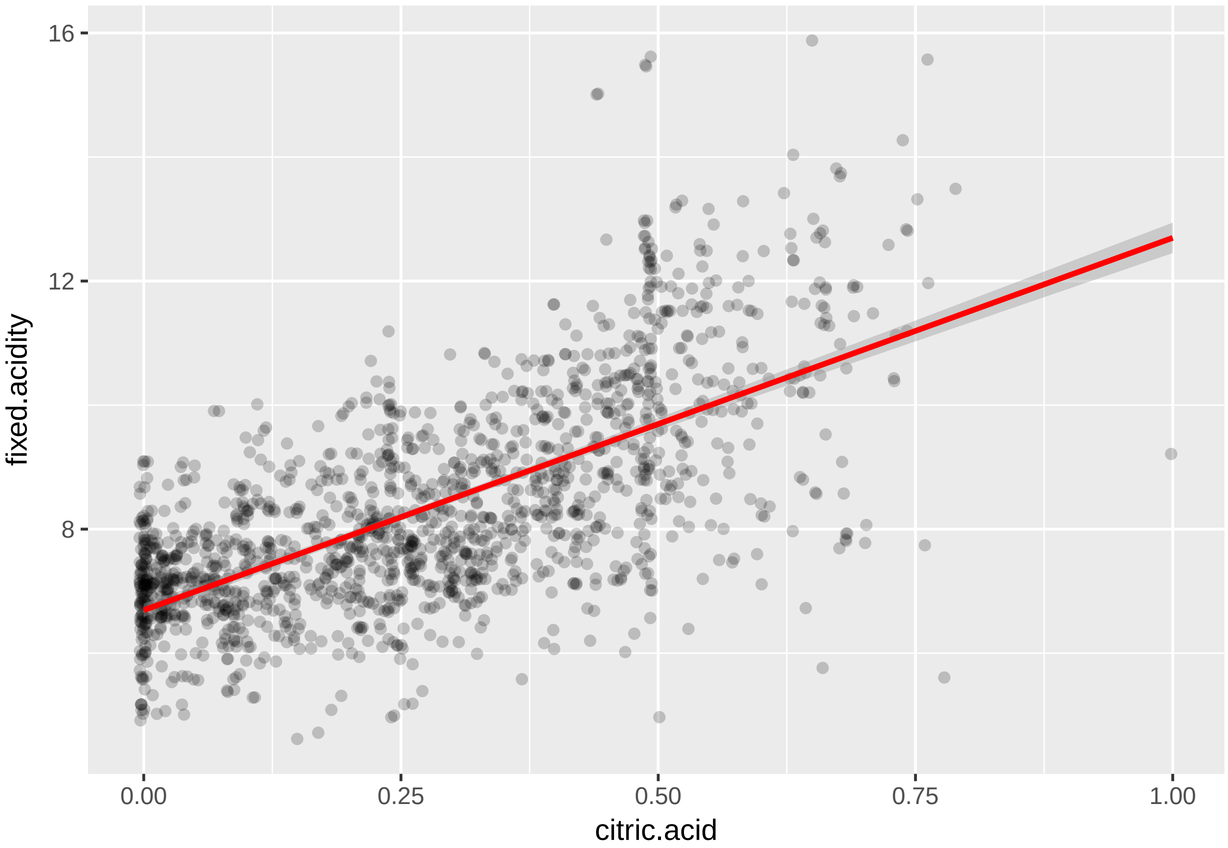
We expect negative correlations in the above features because alcohol is less denser then water and more alcohol would mean less density and for sum acidity feature more acids in the wine means lower value of pH.

The plots below show some positive correlations I found in the features. There was a very strong positive correlation between free sulfur dioxide and total sulfur dioxide.


16
12
8
4
0
chlorides + residual.sugar
0.990
0.995
1.000
density


17.5
15.0
12.5
10.0
7.5
5.0
sum.acidity
0.990
0.995
1.000
density


300
200
100
0
0
20
40
60
total.sulfur.dioxide
free.sulfur.dioxide



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?

The feature I focused on is quality and how other features are responsible for determining the quality of wine. Of all the features some features correlated with quality. Quality had the highest correlation with alcohol at about 0.48 followed by sulphates with a correlation of 0.38, volatile acidity with a correlation of -0.38, sum sulfur a feature I created has a correlation of 0.3, citric acid has a correlation of 0.2 and ratio acidity has a correlation of 0.22.

I have clearly mentioned how the feature of interest varies with the other features under each plot.

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

I found strong correlations between free sulfur dioxide and total sulfur doxide, density and alcohol, sum acidity and pH, density and sum acidity. Finally the features I created had very strong correlations with the features used to create them.

What was the strongest relationship you found?

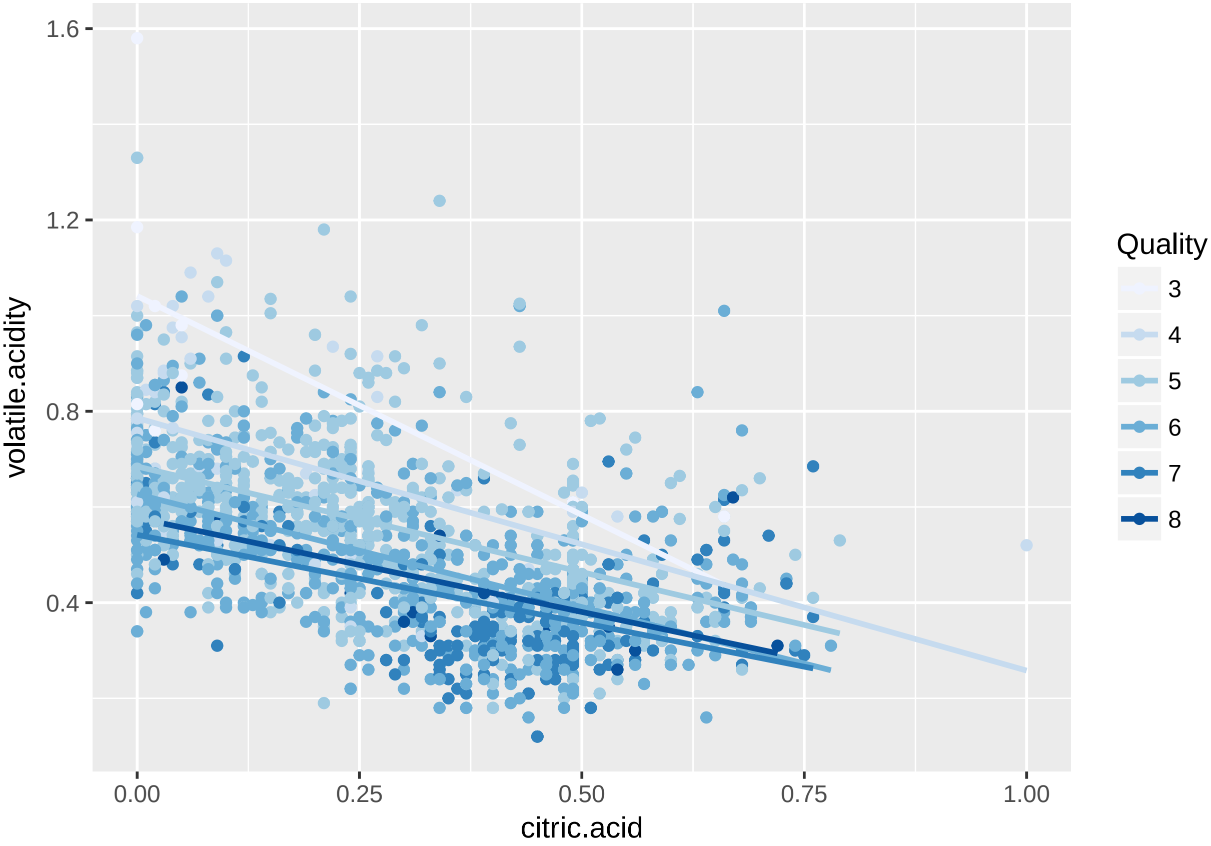
The strongest correlations I found were between the features I created and the features used to create them. This makes sense because the features I created are some mathematical function of the features used to create them and end up correlating with the variables used to create them in the first place.

Otherwise the strongest correlations I found in the data set are between free sulfur dioxide and total sulfur dioxide (0.79), pH and fixed acidity (-0.71).

Multivariate Plots Section

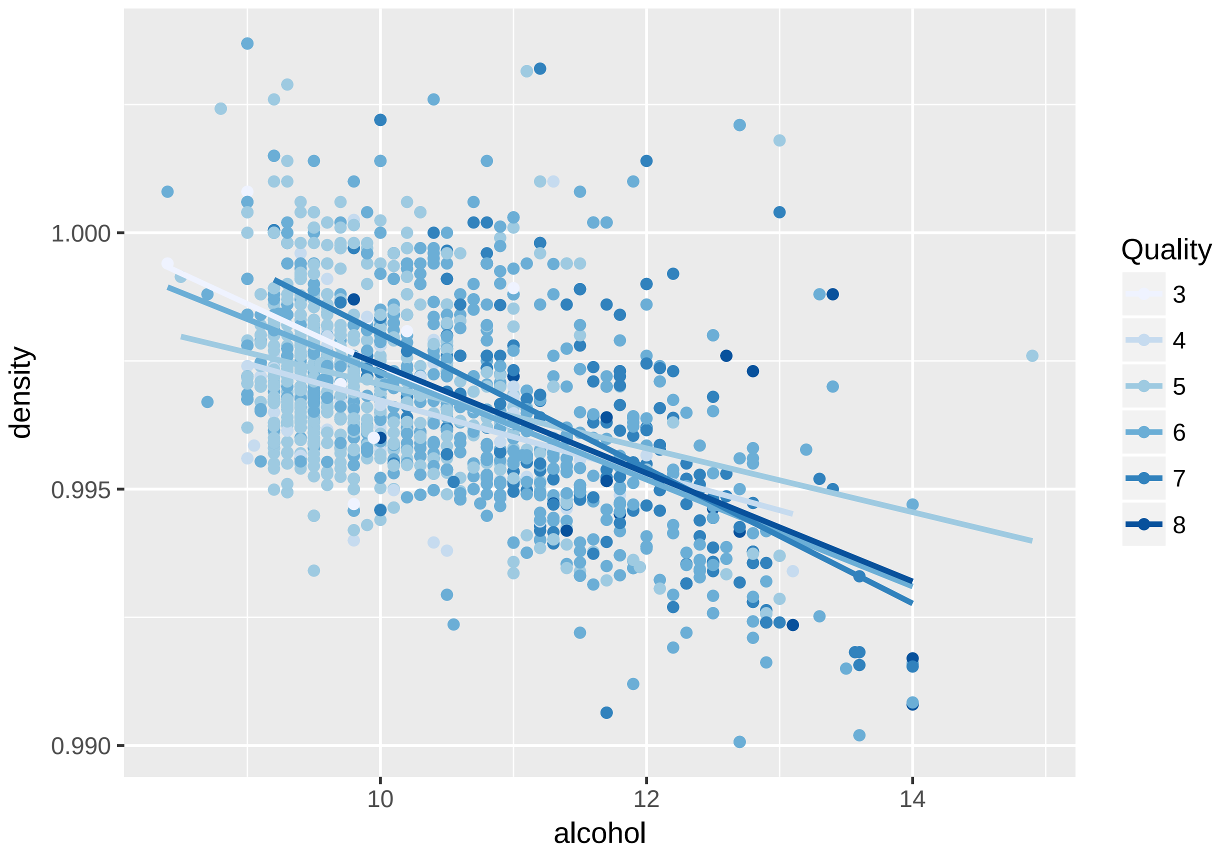
I will focus on features that have high correlations amongst each other and with quality so we can see how two features can impact the quality rating on a plot.

Lets look at how the quality of wine is impacted by citric acid and volatile acidity, because I wanted to see how freshness and unpleasant taste fare into the quality rating. Coordinate transformation might help to look at the data more clearly.



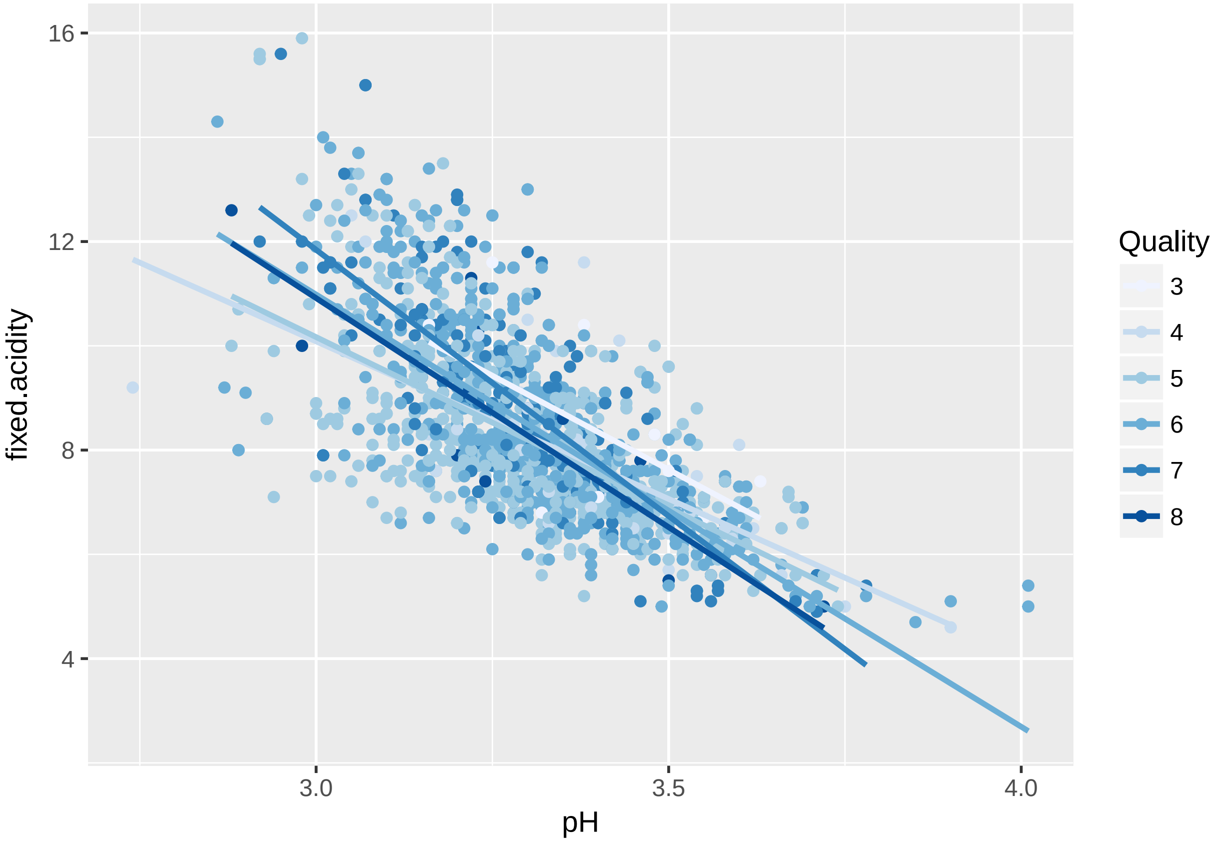
Analyzing the plot above we can see that wines with higher citric acid content and lower volatile acidity have higher ratings, this makes sense.

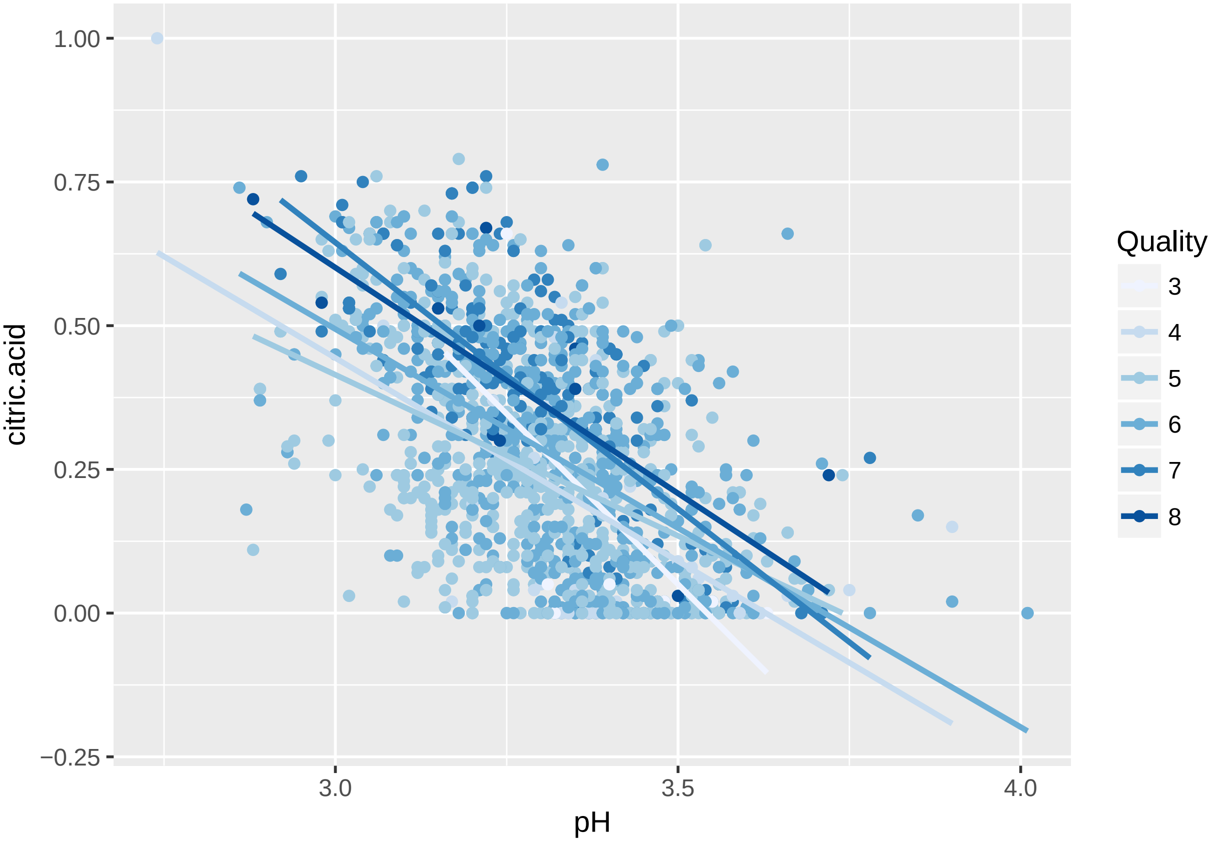
Alcohol has a pretty strong correlation with density, lets look at how quality fare in that plot.



From the plot above we can see that wines with higher quality ratings have a higher alcohol content and lower density.

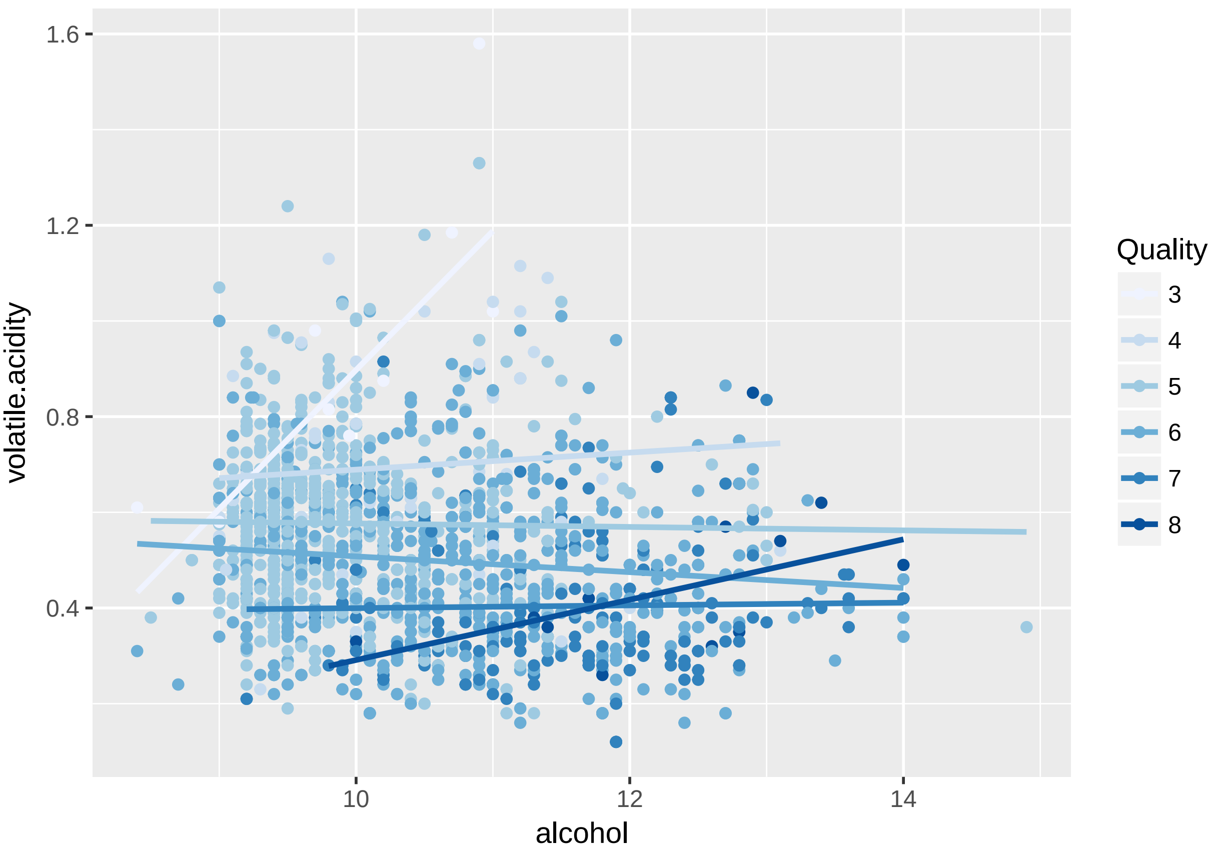
Lets see how the different acids in the wine correlate with pH, then color the plot based on quality rating.

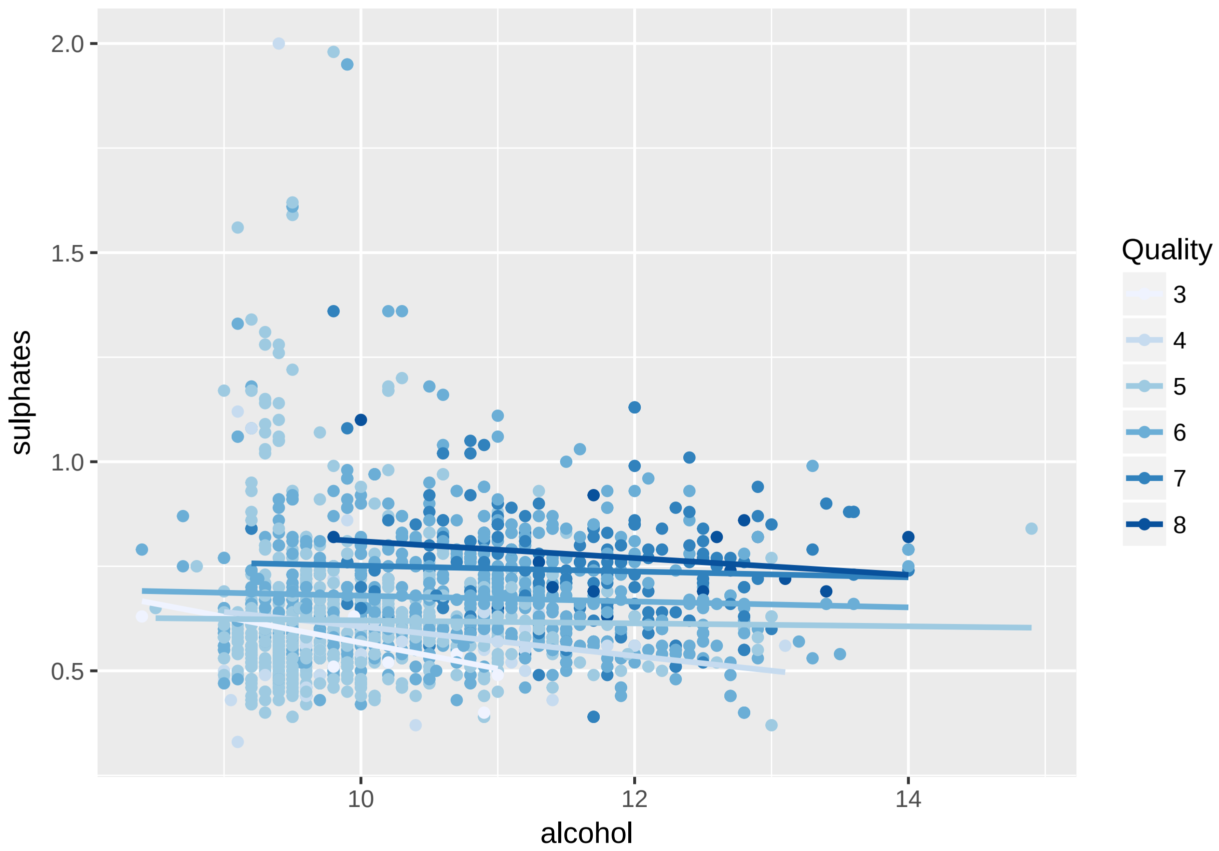

16
12
8
4
fixed.acidity
3.0
3.5
4.0
pH
Quality
3
4
5
6
7
8

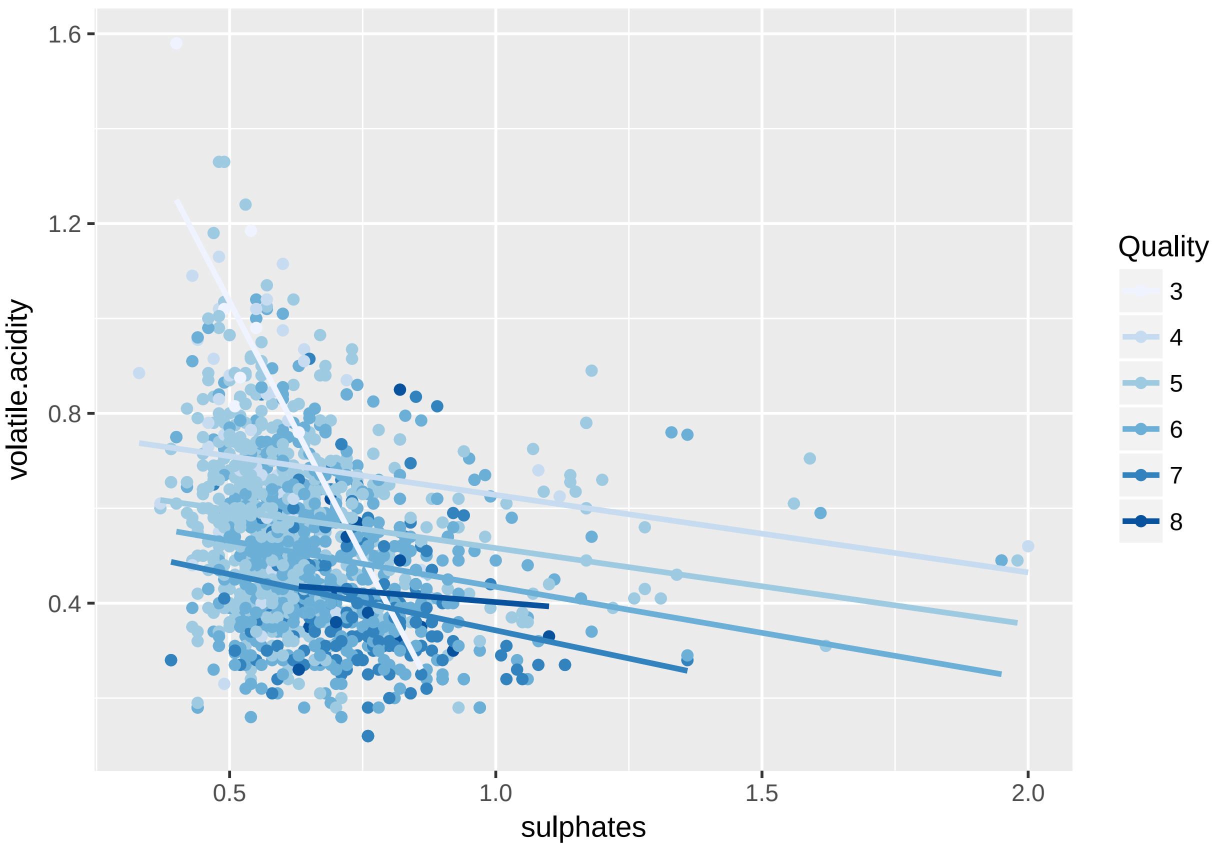


I don't see a clear pattern emerge in pH vs fixed acidity but the plot if pH vs citric acid is interesting. Generally speaking wines that are rated higher have higher citric acid content and lower pH.

Lets pick the two features that correlate the strongest with quality and then do a multivariate analysis. The features that correlated with quality are alcohol, volatile acidity and sulphates.


1.6
1.2
0.8
0.4
volatile.acidity
10
12
14
alcohol
Quality
3
4
5
6
7
8


sulphates
2.0
1.5
1.0
0.5
10
12
14
alcohol
Quality
3
4
5
6
7
8



From the above plots we can see some patterns emerge. The features used to create these plots have some correlation with the quality factor we are interested in. From the alcohol vs volatile acidity we can see that wines with higher alcohol and lower volatile acidity are rated higher, alcohol vs sulphates plot we can see that wines with higher alcohol and generally higher sulphate content is rated higher, sulphates vs volatile acidity plot we can see that wines with lower volatile acidity and generally higher sulphate content between a range has higher rating.

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?

I have in detailed explained the relationship I observed below each plot. Performing a multivariate analysis gave me a good perspective of how quality rating is affected by both the features of interest. Most of the features behaved as expected, the features that strengthened each other are effect of alcohol percentage, volatile acidity and citric acid levels in the wine. They seem to be most prominent in determining the quality rating.

Were there any interesting or surprising interactions between features?

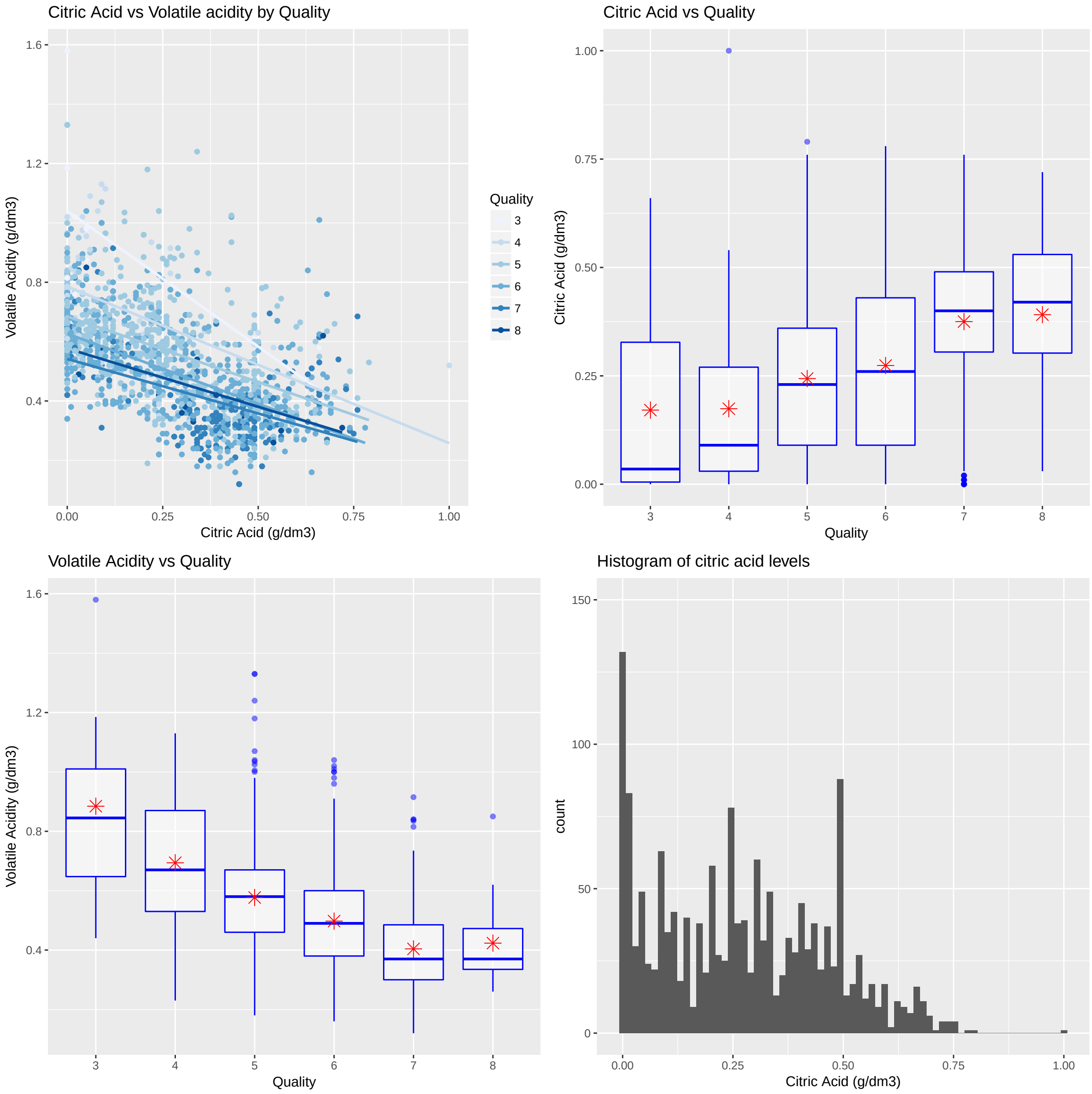
I found the volatile acidity vs citric acid plot very interesting as it clearly shows the importance of these features in deciding the quality rating of the wine.

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

No I did not create any models.

Final Plots and Summary

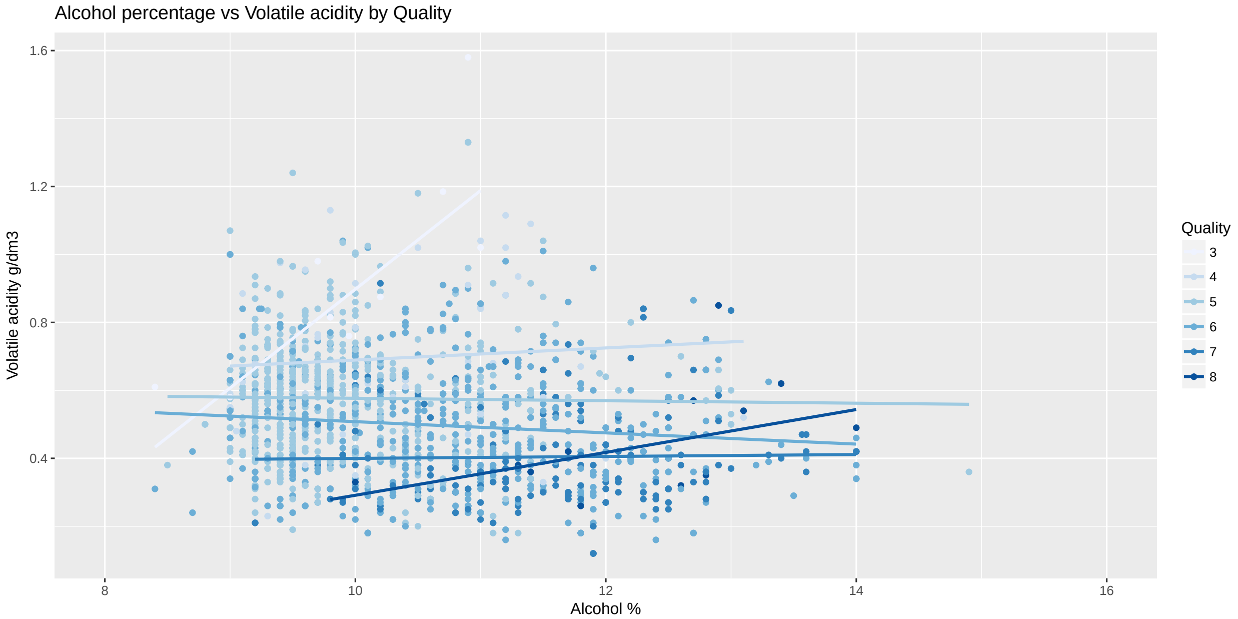
Plot One

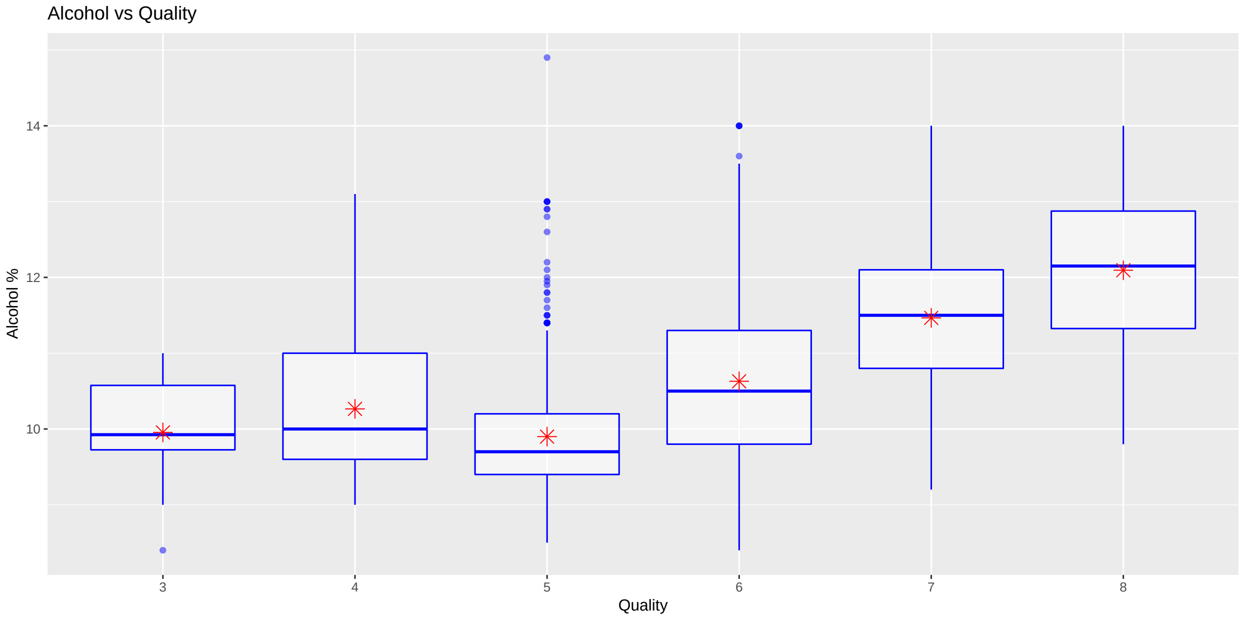


Description One

The plot shown above is the relationship between citric acid vs Volatile acidity by quality, I have added the other plots of the features to justify my reasoning. If we look at the histogram of citric acid levels we see that most of the wines have a citric acid level of 0 and 0.5 g/dm3 it is a bit odd that many wines have no citric acid levels at all, increase in the levels of citric acid lends to a freshness component in the taste of the wine and can led to a higher rating. From the box plot of quality vs citric acid, most of the wines that have a rating of 6 and higher have a mean citric acid level higher then 0.25 g/dm3, there are a few outliers though. The box plot of volatile acidity vs quality behaves differently, because volatile acids cause a unpleasant taste we can clearly observe that wines with higher ratings have lower levels of volatile acidity. Wines with ratings 7 and 8 have a mean volatile acidity of less then 0.4 g/dm3 with exceptions of a few outliers. The first subplot basically takes all these variations into account and we can clearly see a negative correlation between citric acid and volatile acidity, I have colored the points with the quality so we can see that most of the higher rated wines fall in the area of the plot where volatile acidity is less and citric acid levels are higher. These two factors play an important role in determining the quality of the red wine, but does not give us the entire story.

Plot Two

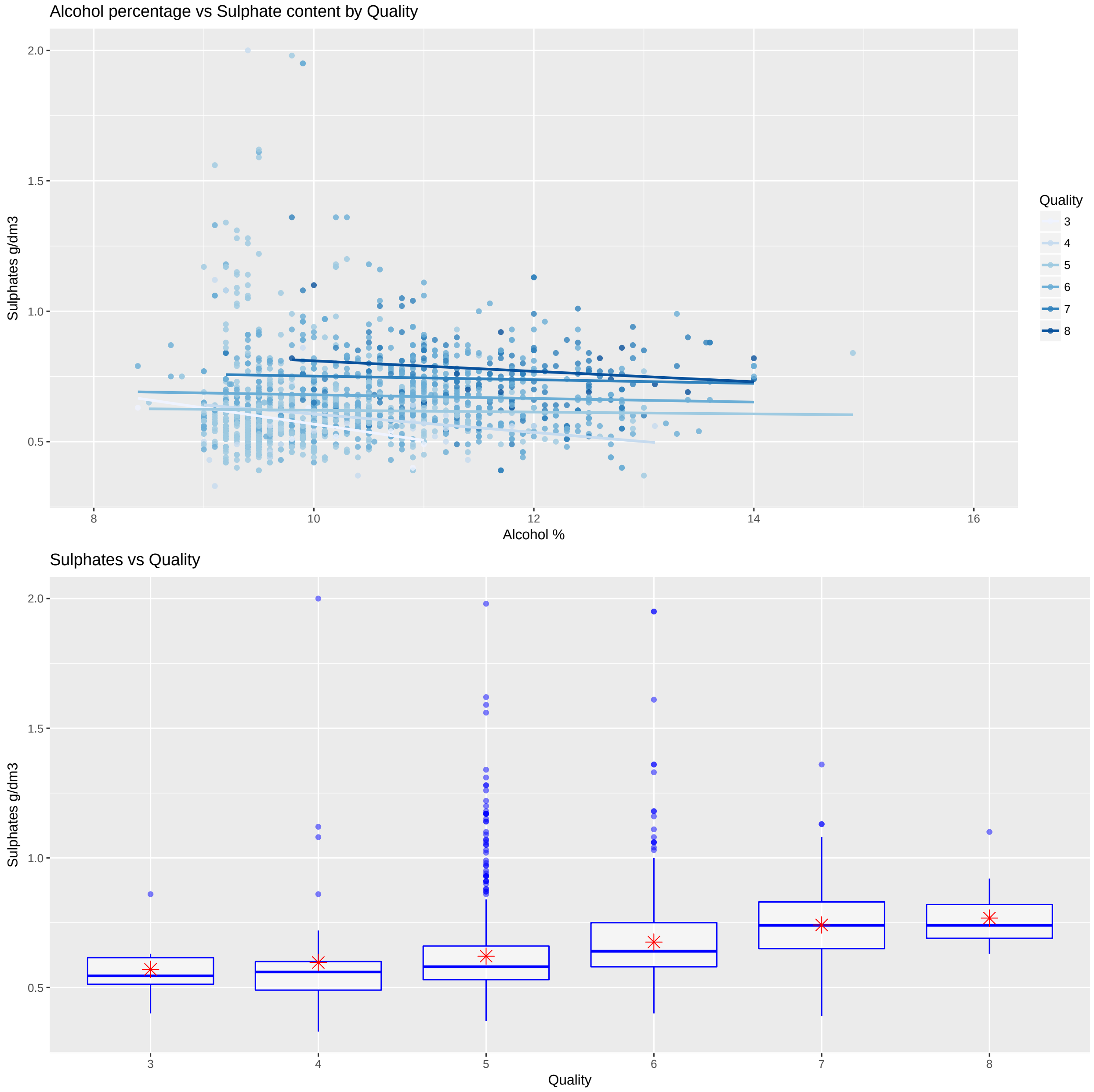




Description Two

The plot shown above is the relationship between Alcohol vs volatile acidity by quality and alcohol had the highest correlation with quality, I have added the other plot of the features to justify my reasoning. We can see that wines with higher alcohol percentage and lower volatile acidity have higher ratings. The trend we can observe in the box plot is that wines with more alcohol percentage got a better rating and the wines with the highest ratings have the highest mean alcohol percentage and from the lot we can say that wines with a mean alcohol percentage of 11 or more have a rating of 7 and higher. From the first subplot its clear that wines with greater alcohol content and lower volatile acidity get a better rating.

Plot Three



Description Three

The feature sulphates had a correlation of 0.38 with quality. The first subplot shows a plot of alcohol content vs sulphates by quality. We can clearly see that wines with higher alcohol content and higher sulphate content get a better quality rating. Performing some research online I found that sulphates behaves as a preservative, antioxidant and antibacterial in the wine thus preserving the freshness of the wine. This makes more sense if we look at the box plot of wines, quality rating of 7 and 8 have a median sulphate content of 0.74 g/dm3 greater then the other quality wines.

Reflection

The data I used to perform EDA consisted of 12 important features of red wines. The entire dataset consisted of 1599 red wines and ratings given to them from a sommelier. My task was to determine what factors are responsible for these quality ratings. I performed extensive EDA on the dataset and found that Alcohol percentage, Volatile acidity content, Sulphates and citric acid had some correlations with the quality rating that the wine received.

The final plots I selected to explain the behavior of the entire dataset were reasonable. We can clearly see how the factors that had some correlation with quality are responsible for the quality rating that the wine was given, neglecting outliers. To summarize this EDA to a reasonable bound, wines with higher alcohol percentage, higher citric acid content, higher sulphate content and lower volatile acidity received a good rating.

The problems I faced while analyzing the dataset is that there are too many variables that correlate with each other and might have played a role in contributing to the quality factor of the wine but I had to decide and investigate more on the features that correlated with the target feature under consideration. After the EDA on the dataset I now to some extent know what factors are responsible for wine quality, I was surprised about the importance of sulphates in red wines to determine the quality.

Future work on the dataset can be to build a machine learning classifier to predict the rating of the wine depending on the features we have discussed in the EDA, this might be useful in investigating the importance of the features themselves.

References

1. http://www.thekitchn.com/the-truth-about-sulfites-in-wine-myths-of-red-wine-headaches-100878
2. https://s3.amazonaws.com/udacity-hosted-downloads/ud651/wineQualityInfo.txt
3. https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html
4. https://www.nceas.ucsb.edu/~frazier/RSpatialGuides/colorPaletteCheatsheet.pdf