Red Wine Quality Exploration by Sherry Tan

Univariate Plots Section

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

```
##  [1] "X"                    "fixed.acidity"        "volatile.acidity"
##  [4] "citric.acid"          "residual.sugar"       "chlorides"
##  [7] "free.sulfur.dioxide"  "total.sulfur.dioxide" "density"
## [10] "pH"                   "sulphates"            "alcohol"
## [13] "quality"
```

```
## '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 0.06
5 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 3.3
5 ...
##  $ 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
```

The median quality is 6.0.

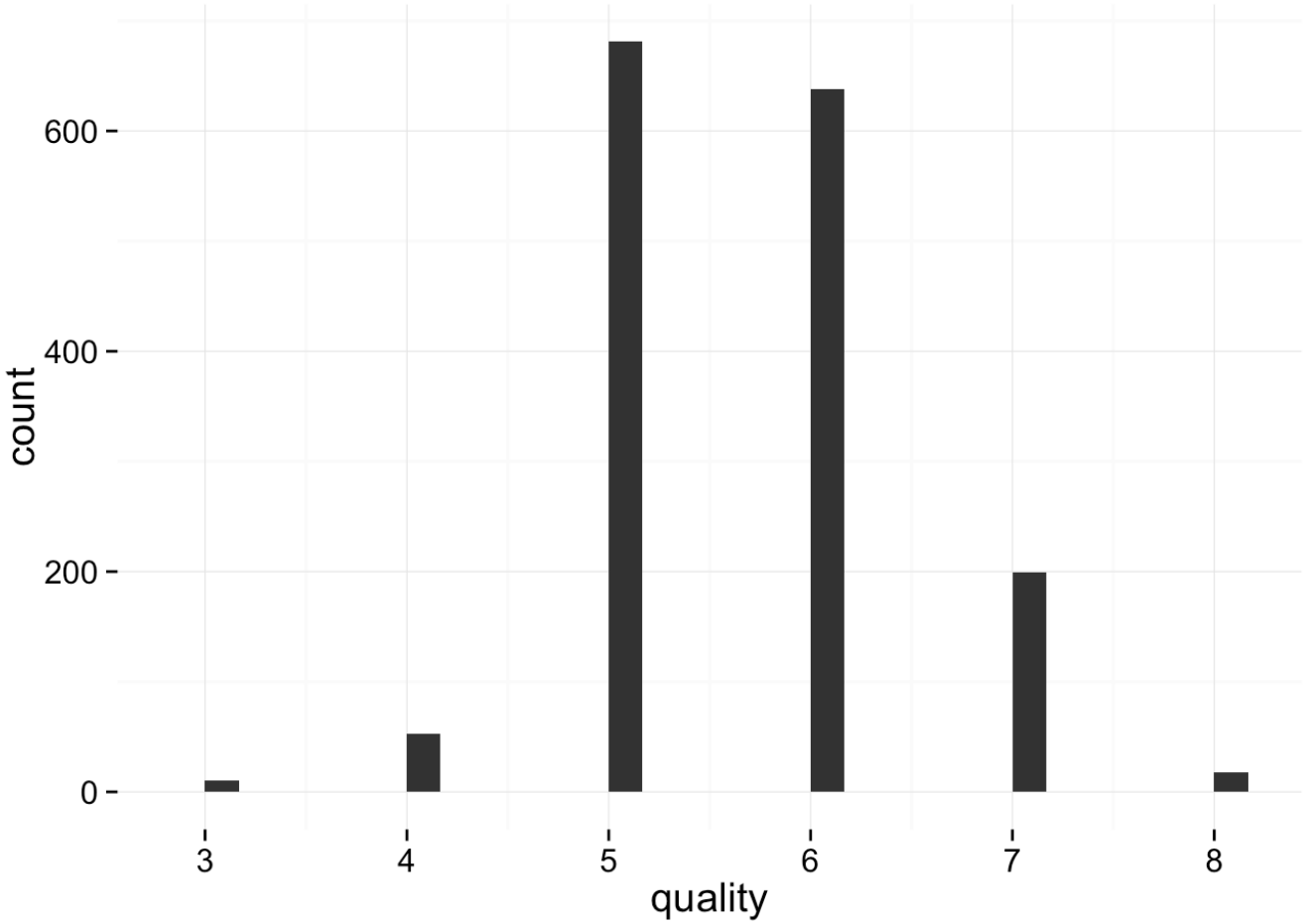
Median volatile.acidity if 0.52 with a maximum of 1.58. 75% of the wines have levels of 0.64 or lower, whereas the other 25% span a wide range from 0.64 to 1.58.

Median citric acid is 0.260 and the maximum is 1.00.

75% of wines have residual sugar of 2.6 g/l or less. Maximum residual.sugar is 15.5g/l, so all red wines in the data set are below the threshold of 45g/l for sweet wines i.e. all the wines are not sweet.

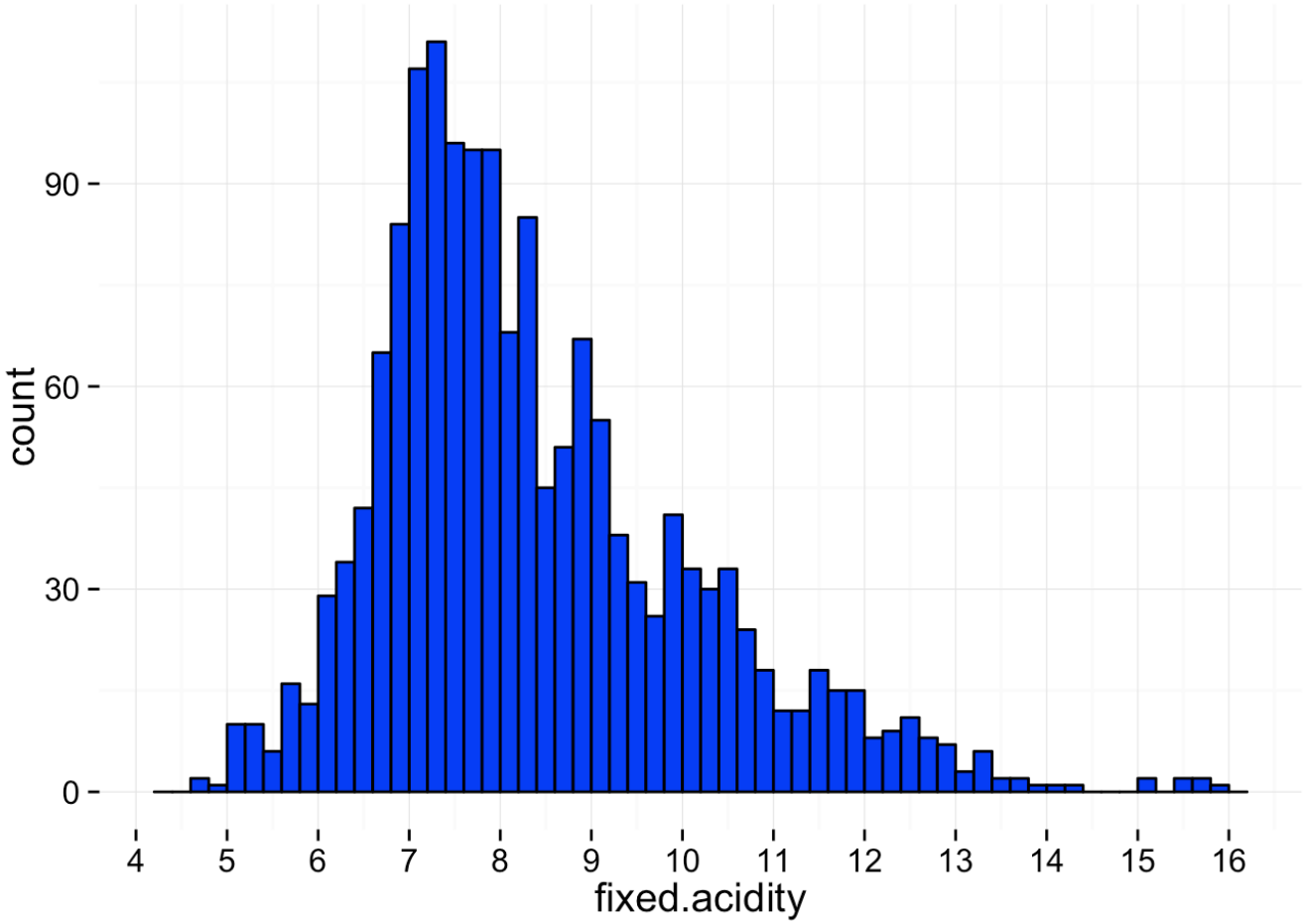
Median free sulphur dioxide is 21.00ppm and the maximum is 72.00ppm.

75% of wines have pH of 3.21 - 4.01.

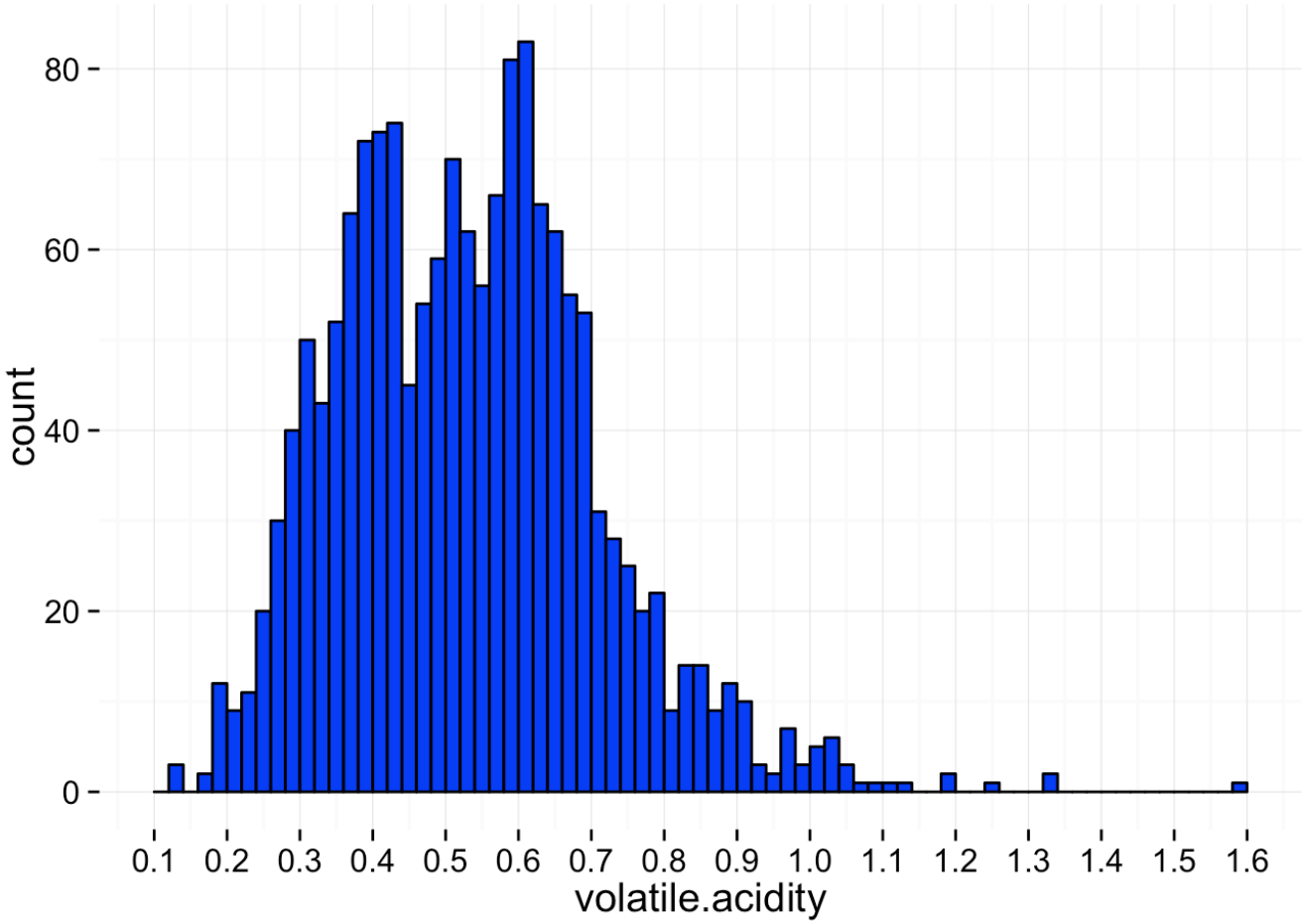



```
##
##   3   4   5   6   7   8
##  10  53 681 638 199  18
```

QUALITY:
Most wines were average and have quality of 5 and 6 with very few wines having low scores of 3 or high scores of 8. Quality appears to have a normal distribution. The quality scores are discrete. I will transform these variables later to factors so that I can do boxplots using the quality variable.

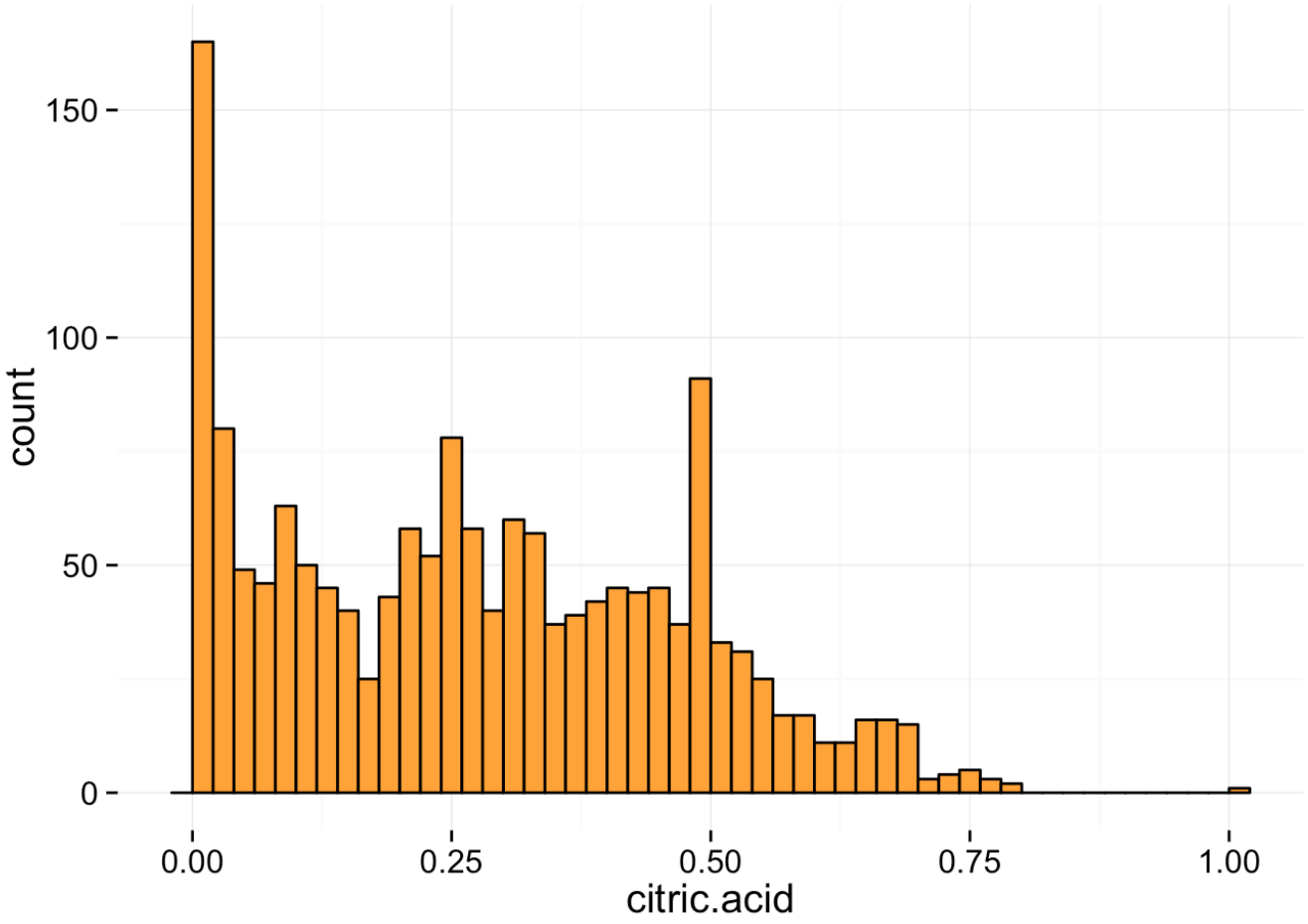


FIXED ACIDITY:
Roughly normally distributed with peak at around 7.2. There appear to be a few outliers at around 15 and 16.



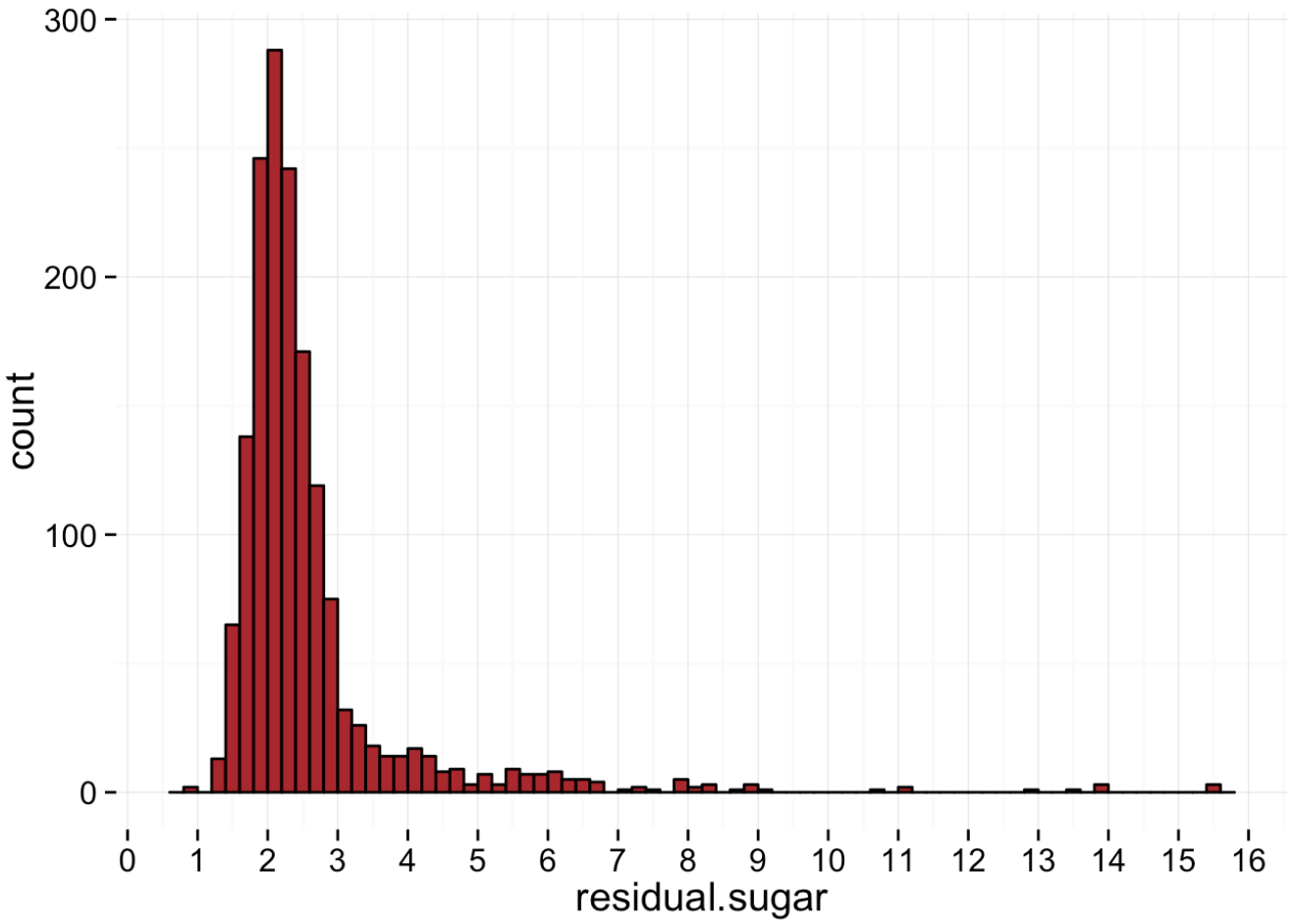
VOLATILE ACIDITY:

The distribution has 2 peaks, at 0.6 and 0.44. According to wineQualityInfo.text, high levels can lead to an unpleasant, vinegar taste. I will see later if higher volatile.acidity values correspond to lower quality.



CITRIC ACID:
There appear to be more wines with citric levels of 0, 0.25 and 0.5. Found in small quantities, citric acid adds freshness and flavor to wines. I would like to see if there is an optimal citric acid concentration beyond which there is no further improvement in taste or quality.

The differences in citric acid concentration could also be a result of the intrinsic properties of the grape varieties used.

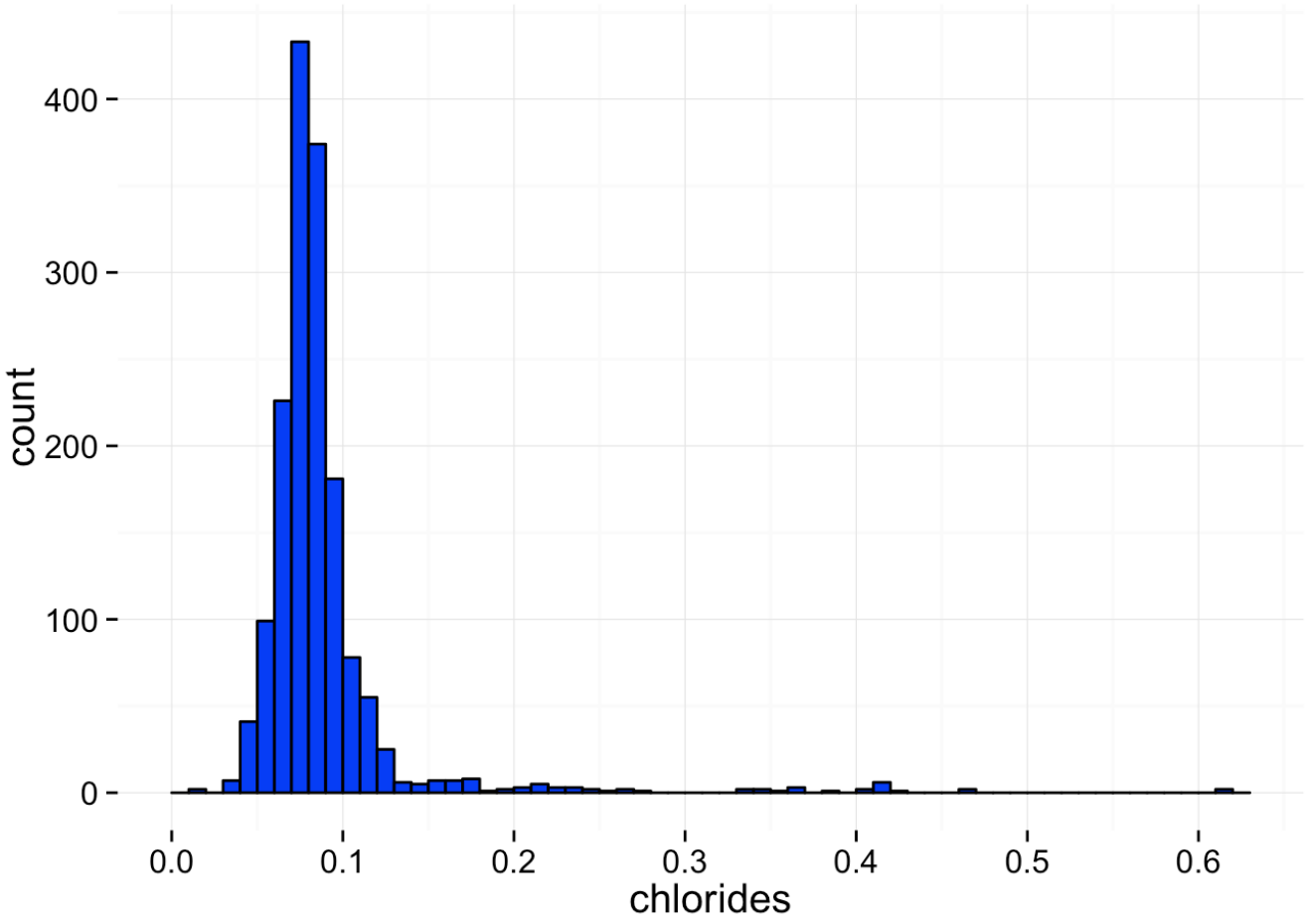



```
## [1] 0.9474672
```

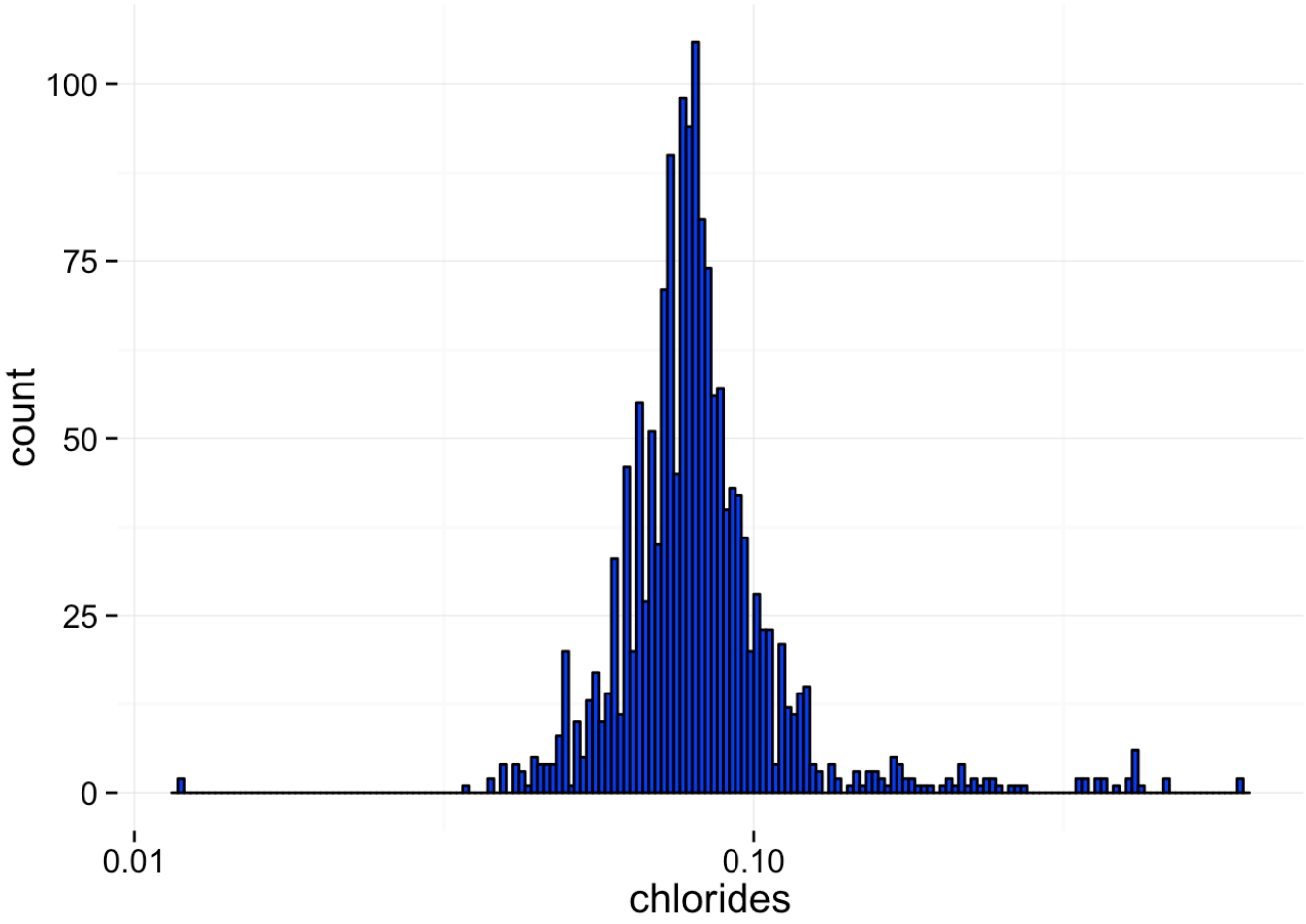
```
##
## FALSE  TRUE
##  1597     2
```

RESIDUAL SUGAR:
95% of wines have residual sugar in the range of 1g/l - 5g/l. The remaining 5% span a much wider range, from 5g/l to 15.5g/l. Could residual.sugar be a characteristic that is quality controlled in the production process, leading to the narrow spread?

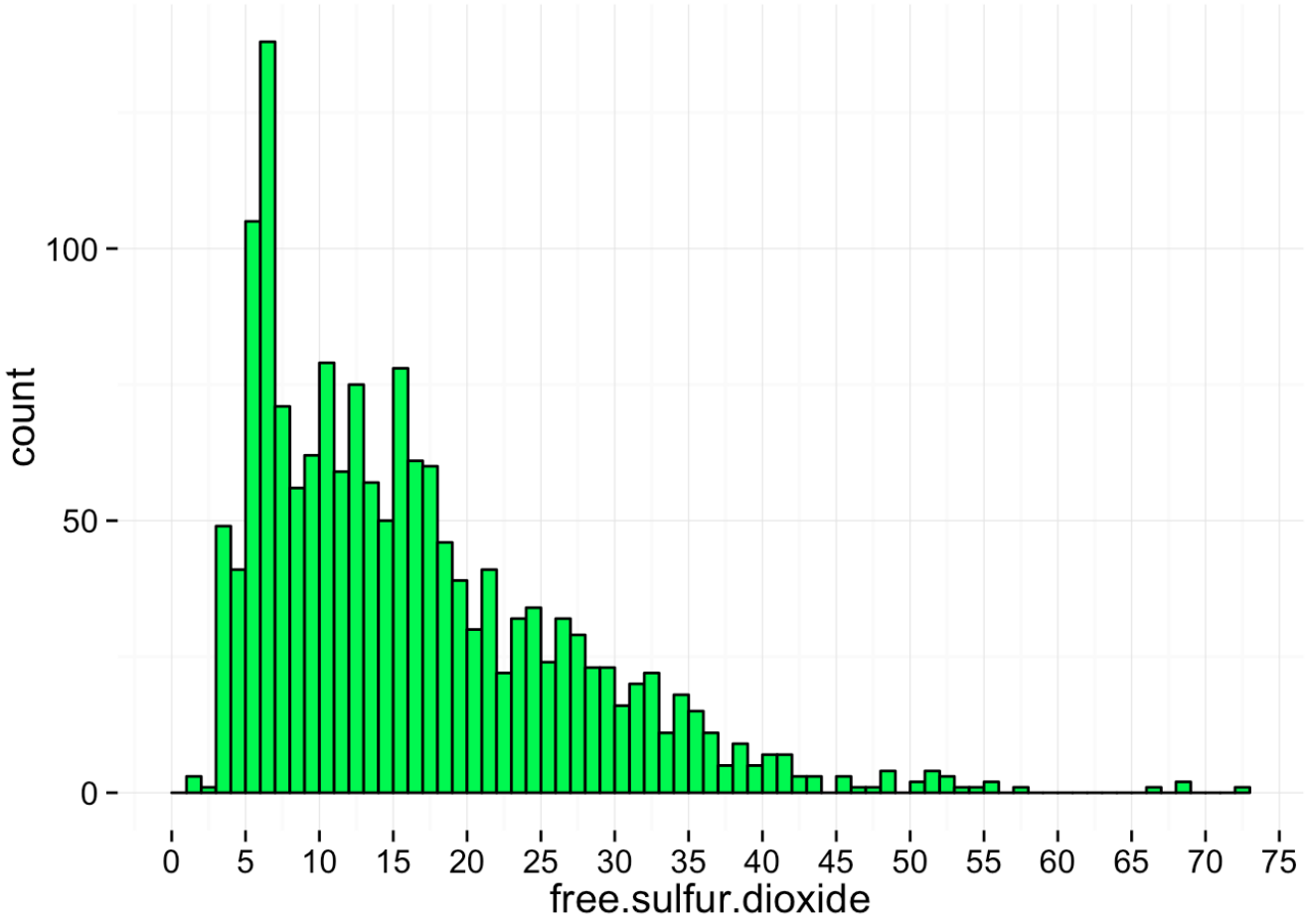
There are only 2 (out of 1599 wines) with 1g/l, which is expected has wines with such low levels are rare (according to the data set info).




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



CHLORIDES:
Most of the wines have chlorides in the range of 0.05 to 0.15. There are outliers in the data, with the most extreme value at 0.611. To better visualize the distribution, I am also plotting the x axis on a log scale. The second plot shows that the chloride has a normal distribution.




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

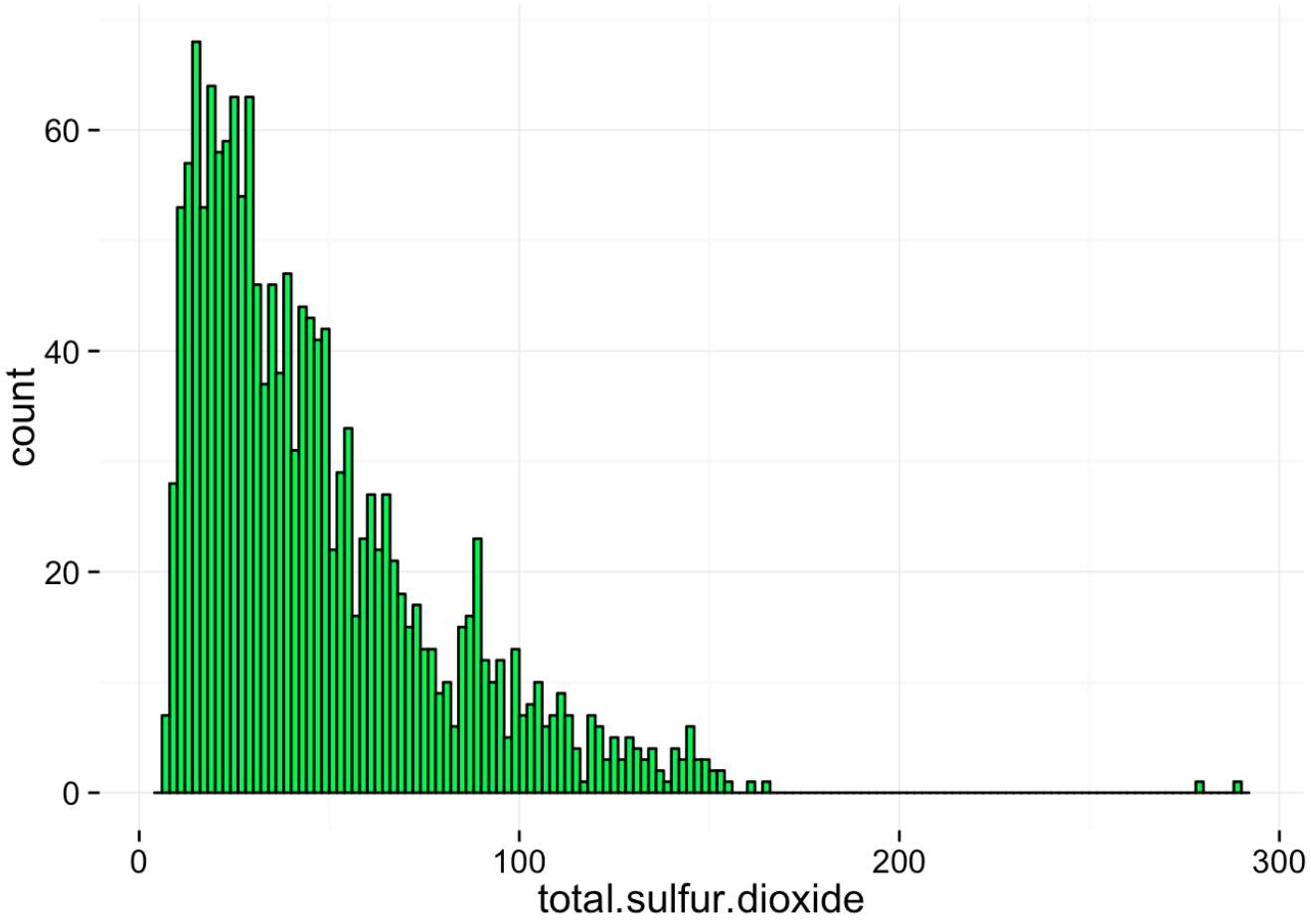
```
##
## FALSE  TRUE
##  1583    16
```

FREE SULFUR DIOXIDE:
Distribution of free sulphur dioxide was right skewed, with most of the wines having 40ppm or less free sulphur dioxide.

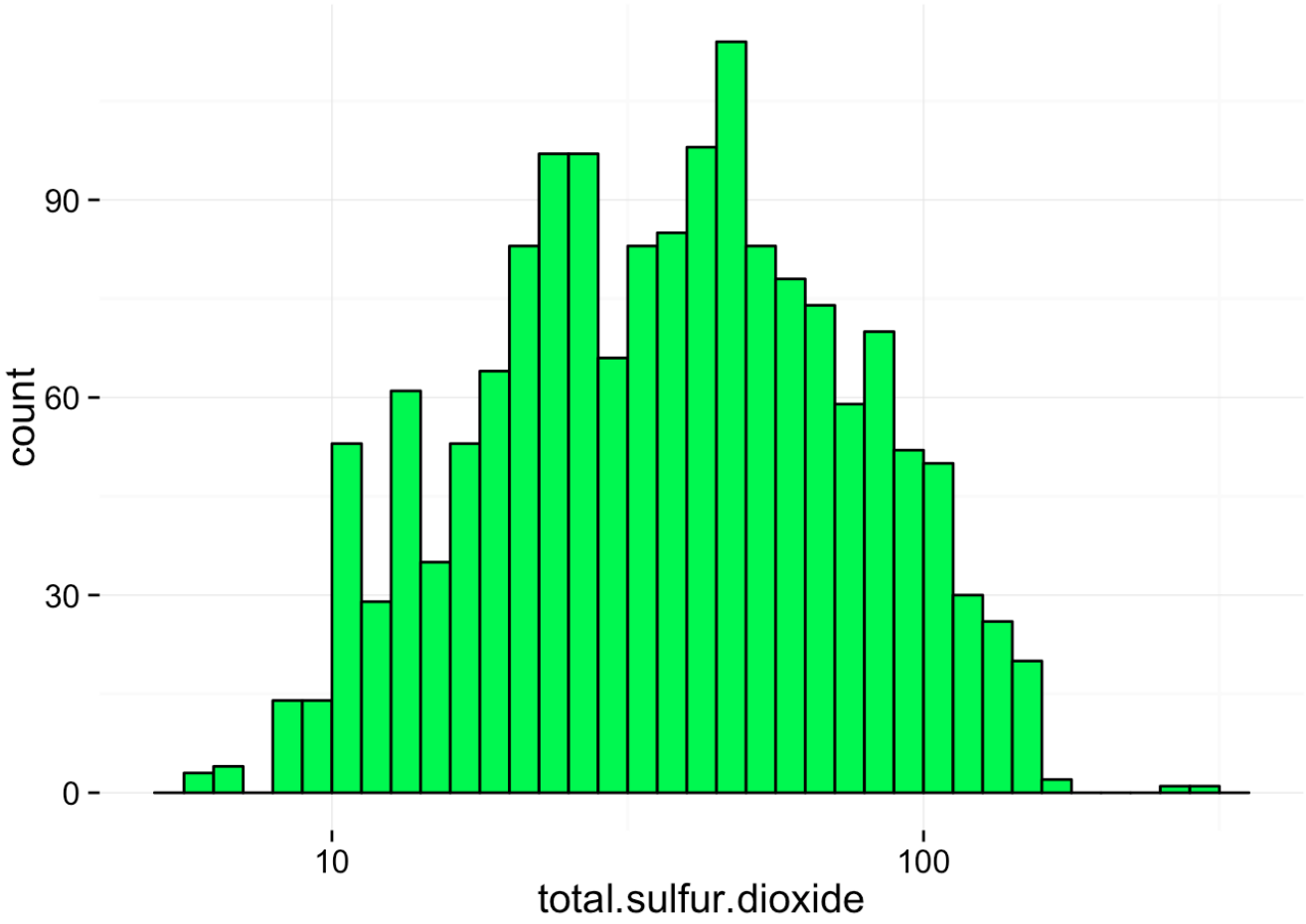
16 out of 1599 wines (1%) have free sulphur dioxide > 50ppm. According to wineQualityInfo, SO2 becomes evident in the nose and taste of wine at these levels. I expect this to lower the quality of the wines. I will look later at these wines have lower quality.

There is a peak at 7ppm. What's so special about this value? Is this a optimum level for SO2 to perform its function of preventing microbial growth and the oxidation of wine?

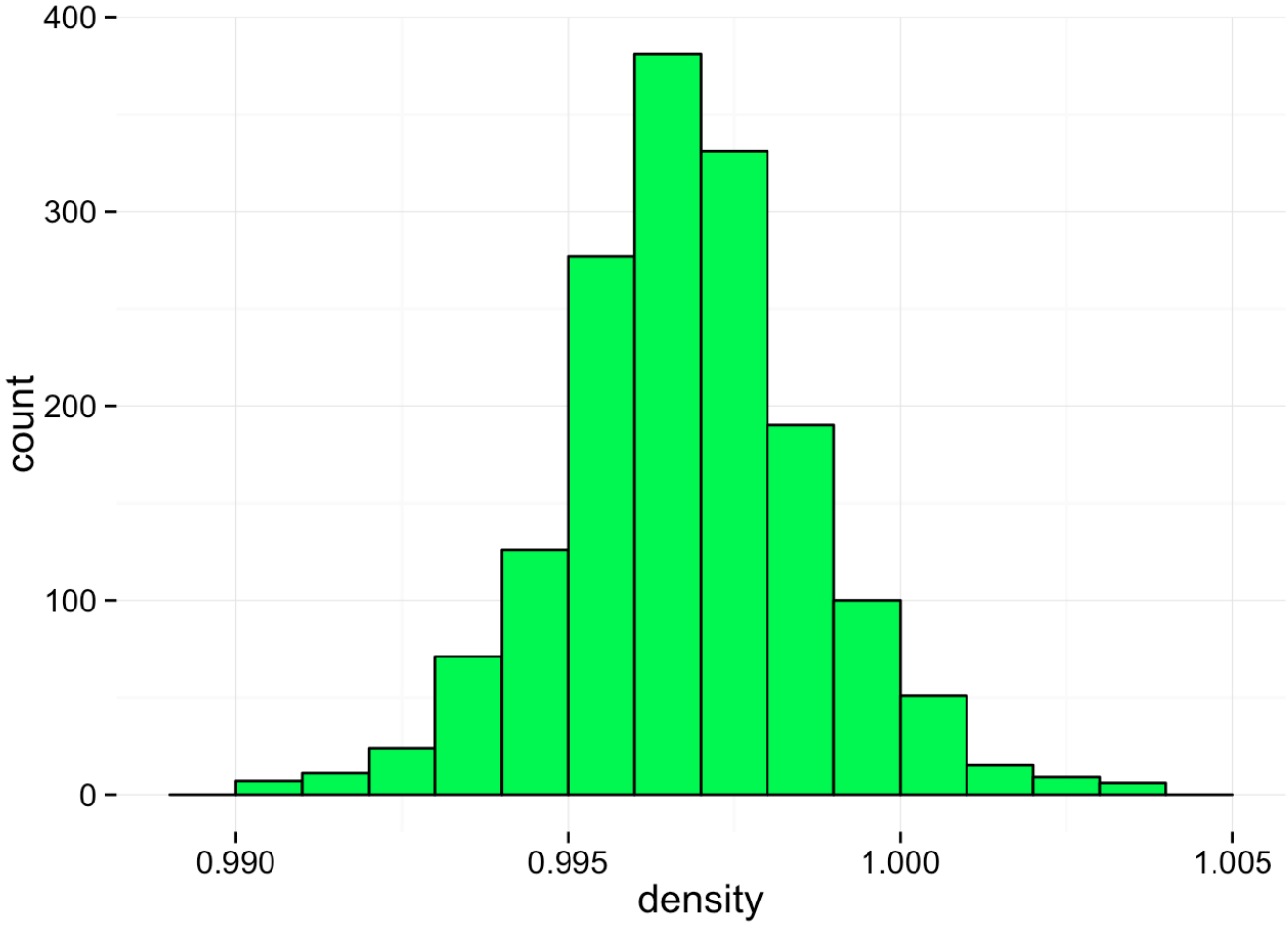
The production process involves the addition of sulphate additives, which affects SO2 levels. I would expect that this production step is tightly controlled and that the amount of sulphates added to the wines is kept constant. I will look at the relationship between sulphate additives and free SO2 later.




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

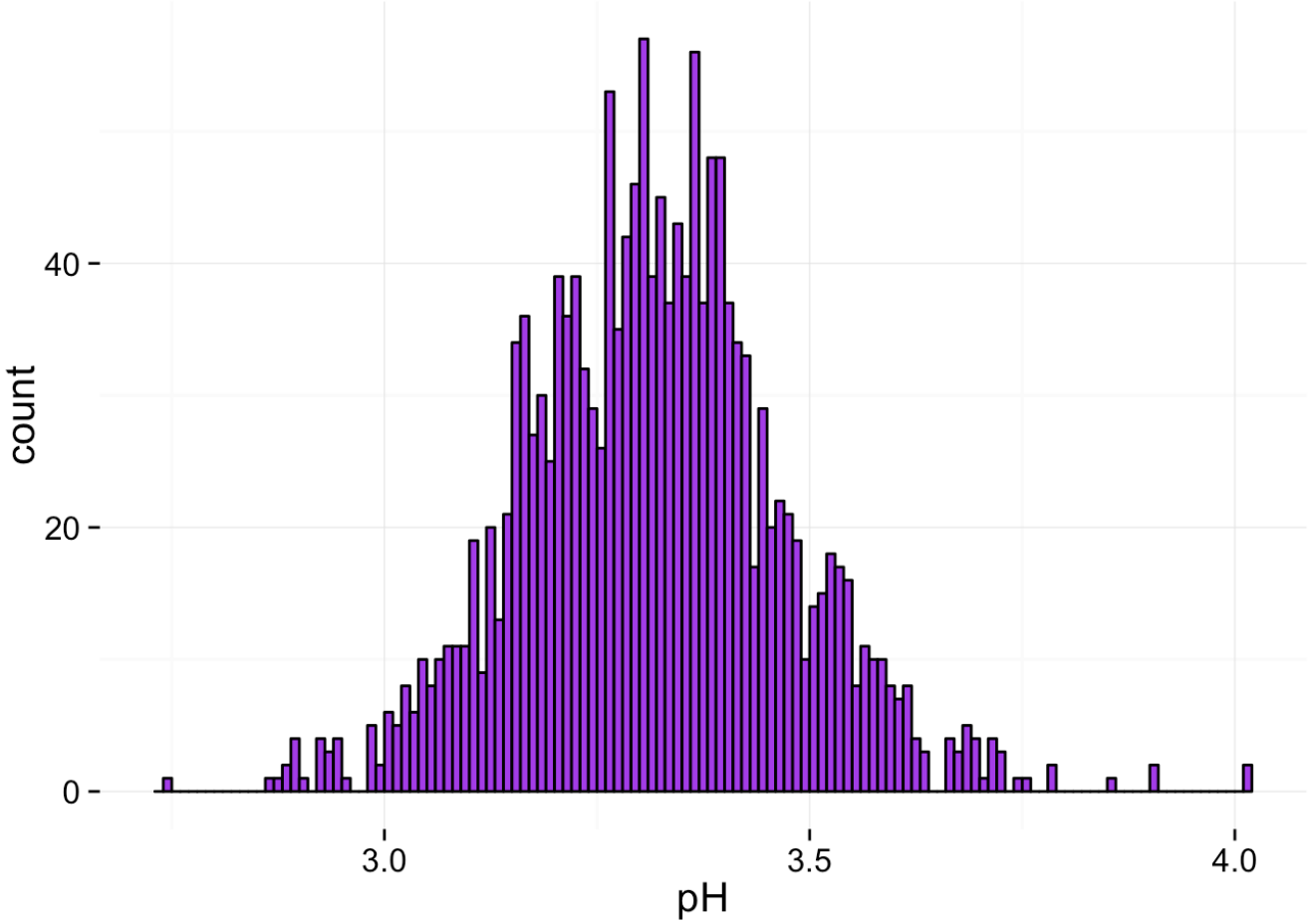


TOTAL SULFUR DIOXIDE:
The distribution is right skewed with a median of 38.00. There are a few outliers close to 300.00. To better visualize the data, I replotted using a log x scale and found that the resulting distribution was normal. I will investigate later if these extreme values coincide with high levels of free SO2.




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

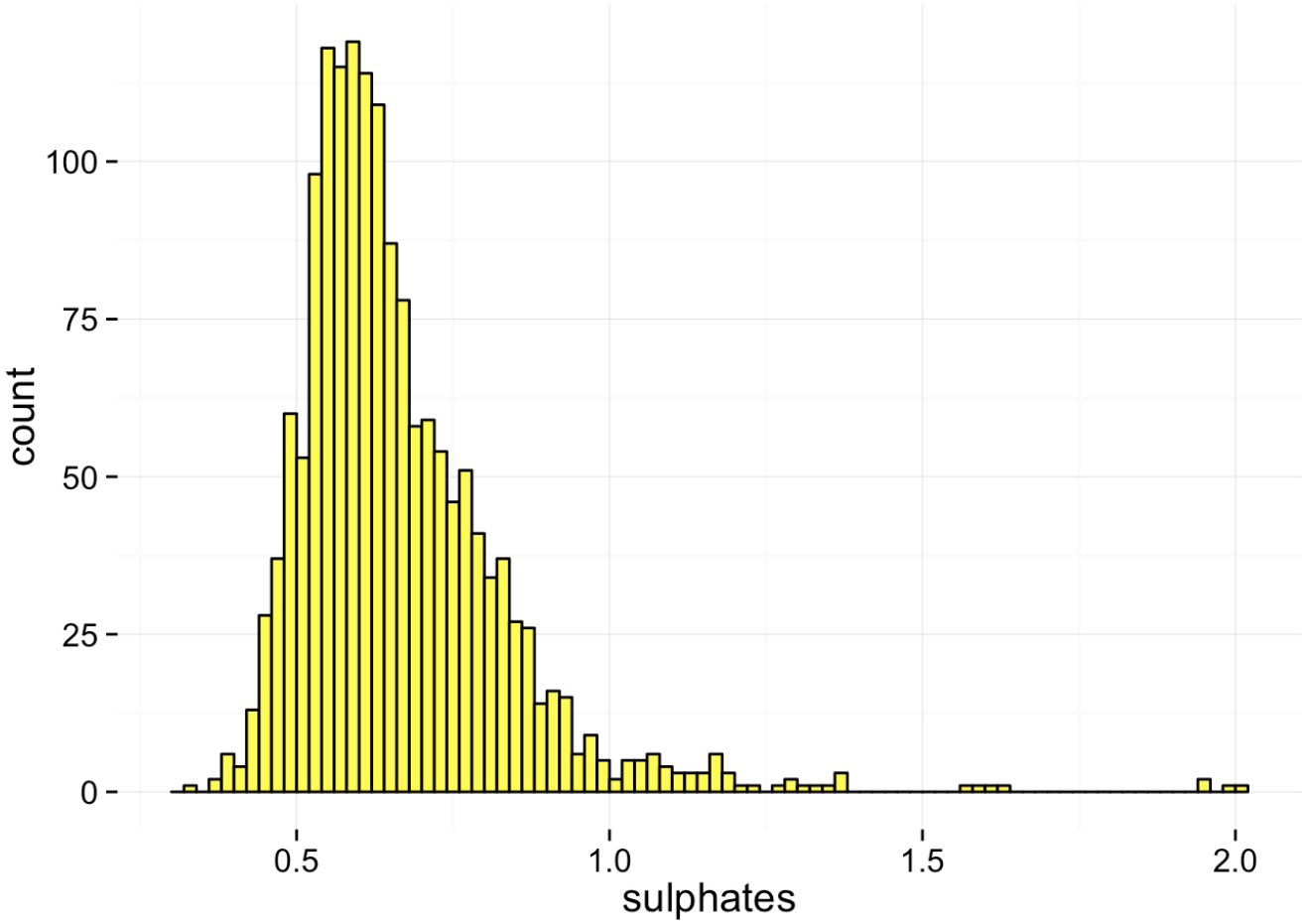
DENSITY:
Density is normally distributed with a median 0.997. Indeed, this doesn't come as a surprise. Even though the concentration of other species such as alcohol and sugar would affect density, wine is mostly water, so the impact of other species on density would not be significant. As such, I didn't expect too much deviation in density from the density of water for all wines in the data set.



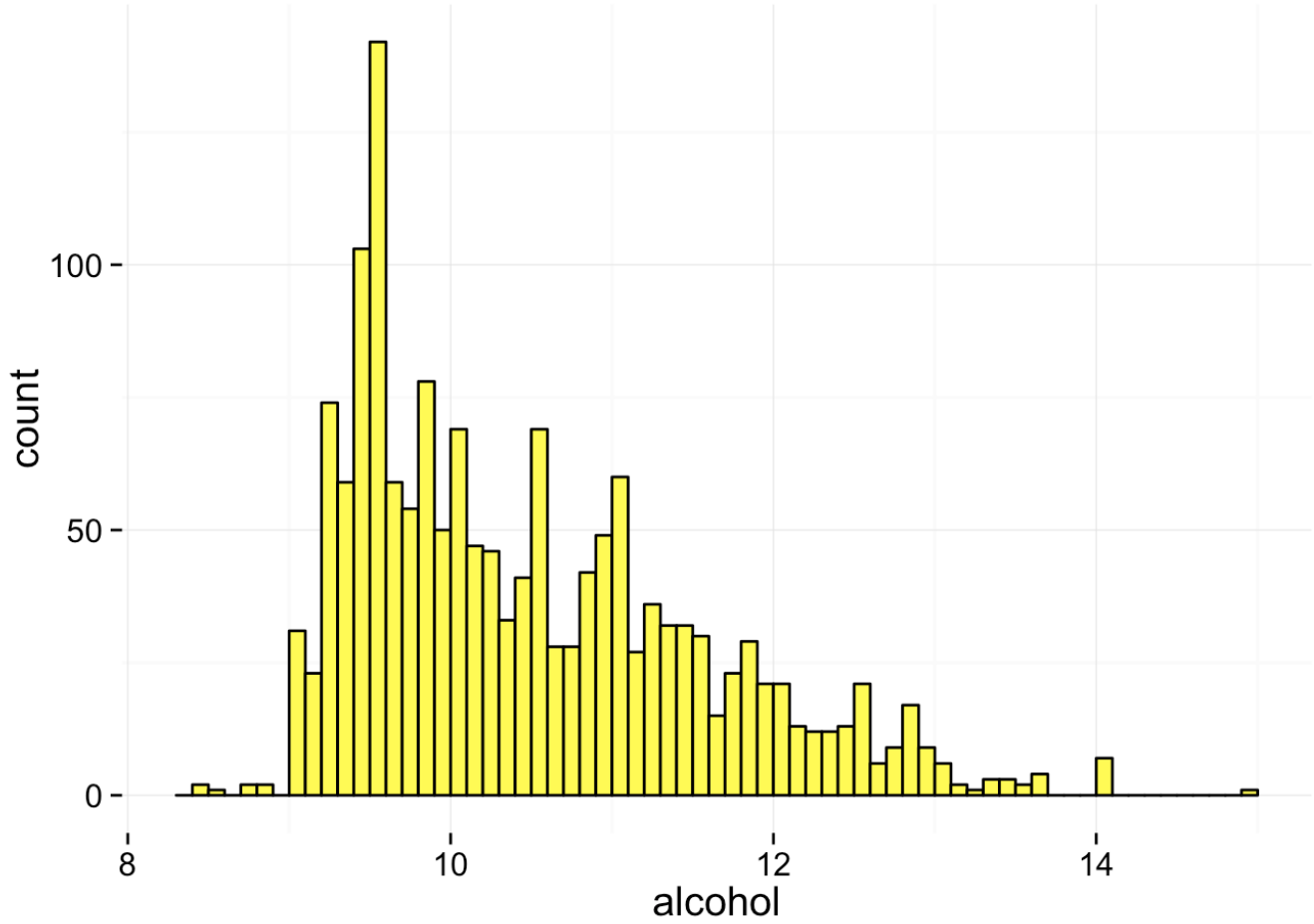

```
##
## FALSE  TRUE
##    31  1568
```

pH:
Most wines have pH between 3-3.7.

31 of 1599 wines (1.9%) had pH outside the usual range of 3-4 (ref. winequalityinfo.txt), wonder if these would correspond to the poor quality wines? I will take a closer look at the wines lying outside this normal range, and whether they correspond to lower quality.



SULPHATES:
Sulphate levels are concentrated between 0.5-1.0 with some outliers up to a value of 2. I expected the distribution to be narrower with fewer outliers since these are additives are added during production. I would have thought that the level of sulphates added would be quite tightly controlled and therefore more uniform in value.




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

ALCOHOL:
Alcohol has a right skewed distribution. There is a peak at around 9.5%. 75% of the wines have alcohol levels ranging from 8.40 - 11.1

Univariate Analysis

What is the structure of your dataset?

There are 1,599 wines in the dataset with 12 features (fixed.acidity, volatile.acidity, citric.acid, residual.sugar, chlorides, free.sulfur.dioxide, total.sulfur.dioxide, density, pH, sulphates, alcohol, quality). Apart from quality, all variables are continuous numerical variables. Quality is a discrete integer variable.

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

The main feature of interest is quality.

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

Volatile.acidity, free.sulfur.dioxide, Citric.acid, sulphates, which affect the taste of wine. I will also look at sugar content and alcohol - these may affect the quality in a more subjective sense, depending on the personal preferences of the wine taster.

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

The following variables were created to facilitate subsequent analysis:
1. qualityF - I will convert the discrete quality scores to factors later on so that I can do boxplots.
2. pH_normal- to distinguish between wines with pH 3-4 (‘1’) and otherwise (‘0’).

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?

Citric acid was unusual because of its 3 peaks at 0, 0.25 and 5 g/dm^3(perhaps a controlled characteristic during production).

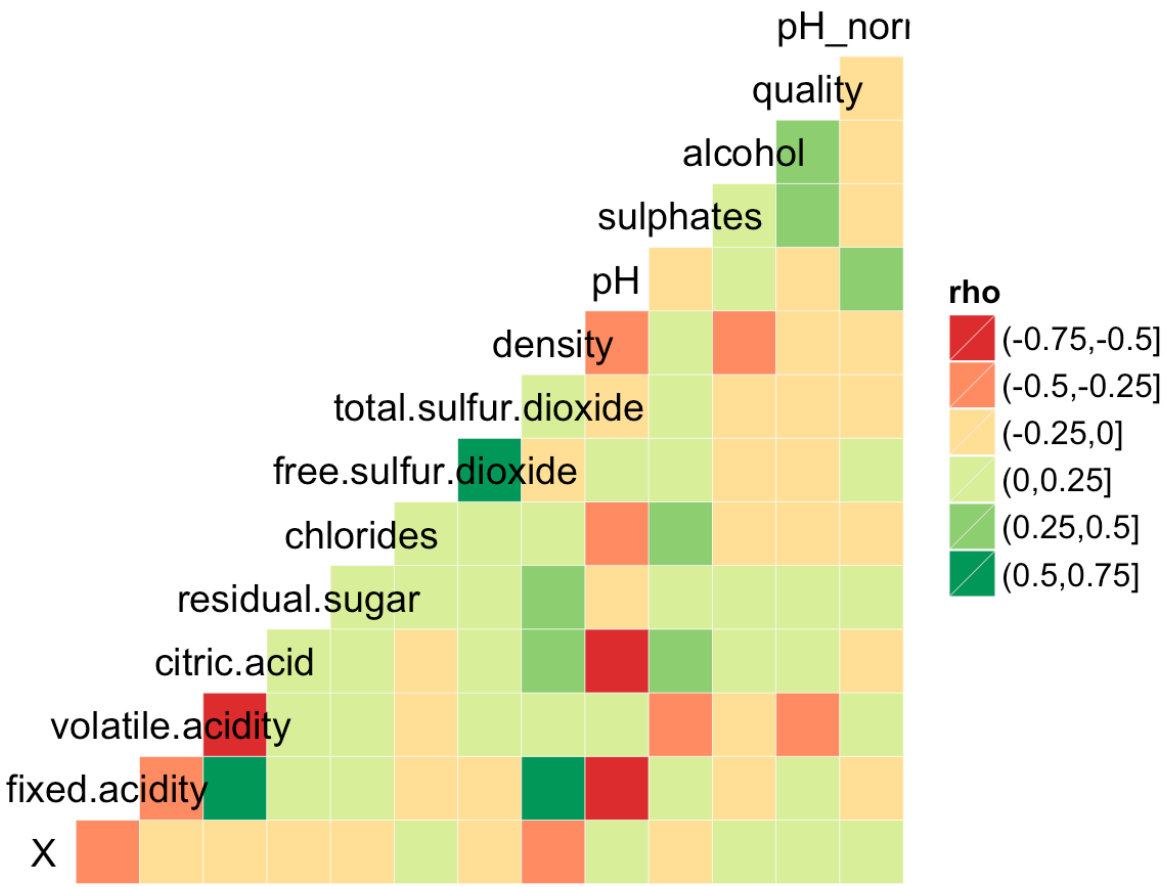
The sulphate distribution was wider than I expected, perhaps sulphate content is not affected only by the addition of additives?

Total sulfur dioxide had a right skewed distribution and was unusual because there were a few very extreme outliers - a maximum value of 289.00 mg/dm^3, more than 4 times larger than the 75th percentile value of 62.00 mg/dm^3.

Chloride showed a few extreme outliers as well - a maximum value of 0.611 g/dm^3, almost 7 times greater than the 75th percentile value of 0.09 g/dm^3.

To better visualize the data for chloride and total.sulfur.dioxide, I made plots using a log scale for the x axis. Both had normal distributions. I will investigate later whether these extreme outliers are related to low wine quality.

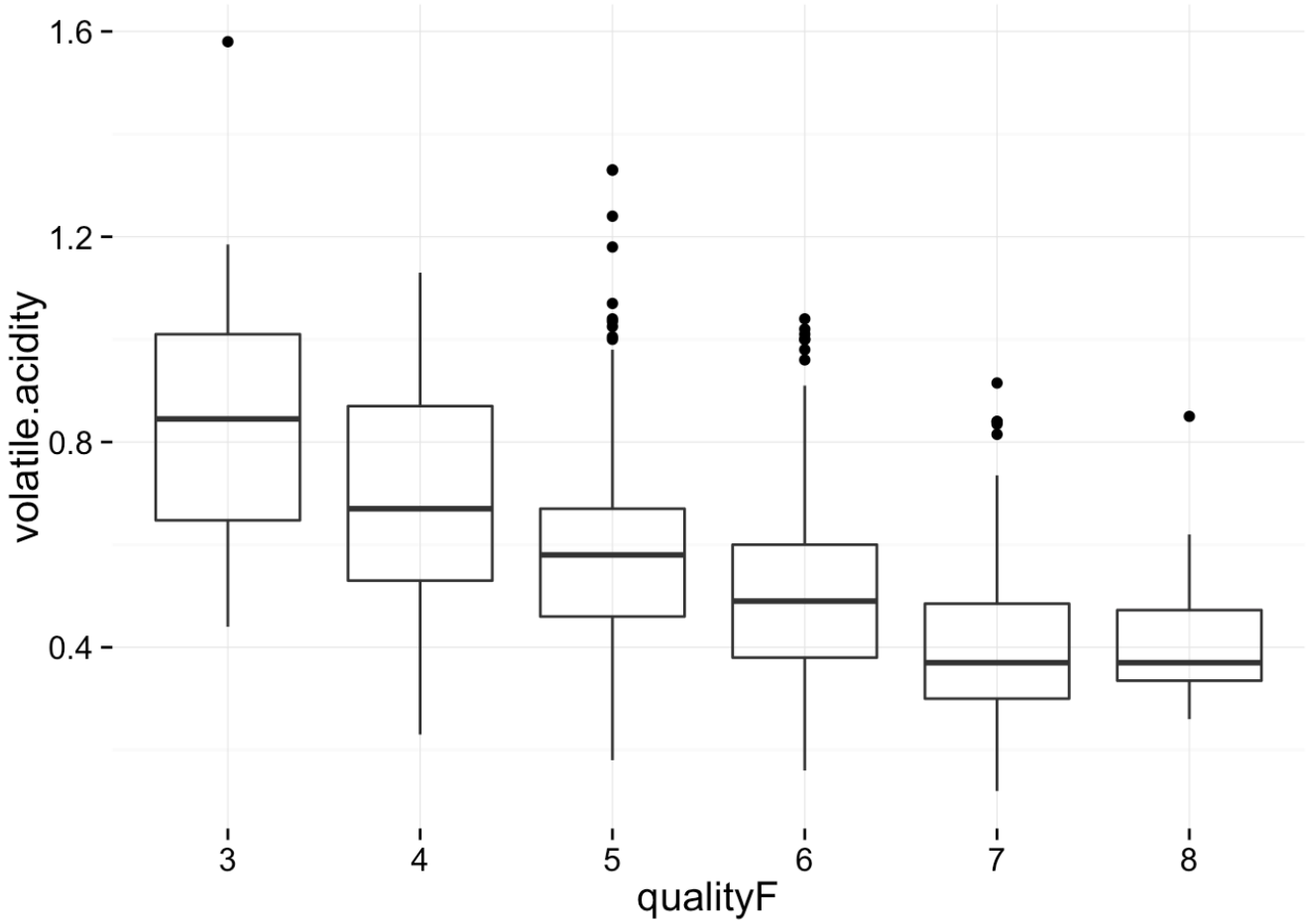
Bivariate Plots Section



This figure gives me a preliminary sense of the pairs of variables that are likely to be related. I will use this later to guide my investigation of relationships between features.

Besides investigating how quality varies with the features I identified earlier - Volatile.acidity, free.sulfur.dioxide, Citric.acid, sulphates, alcohol and residual.sugar, I will investigate whether outliers in pH, total.sulfur.dioxide and chlorides belong to lower quality wines.

Relationship between Quality and features of interest




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

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

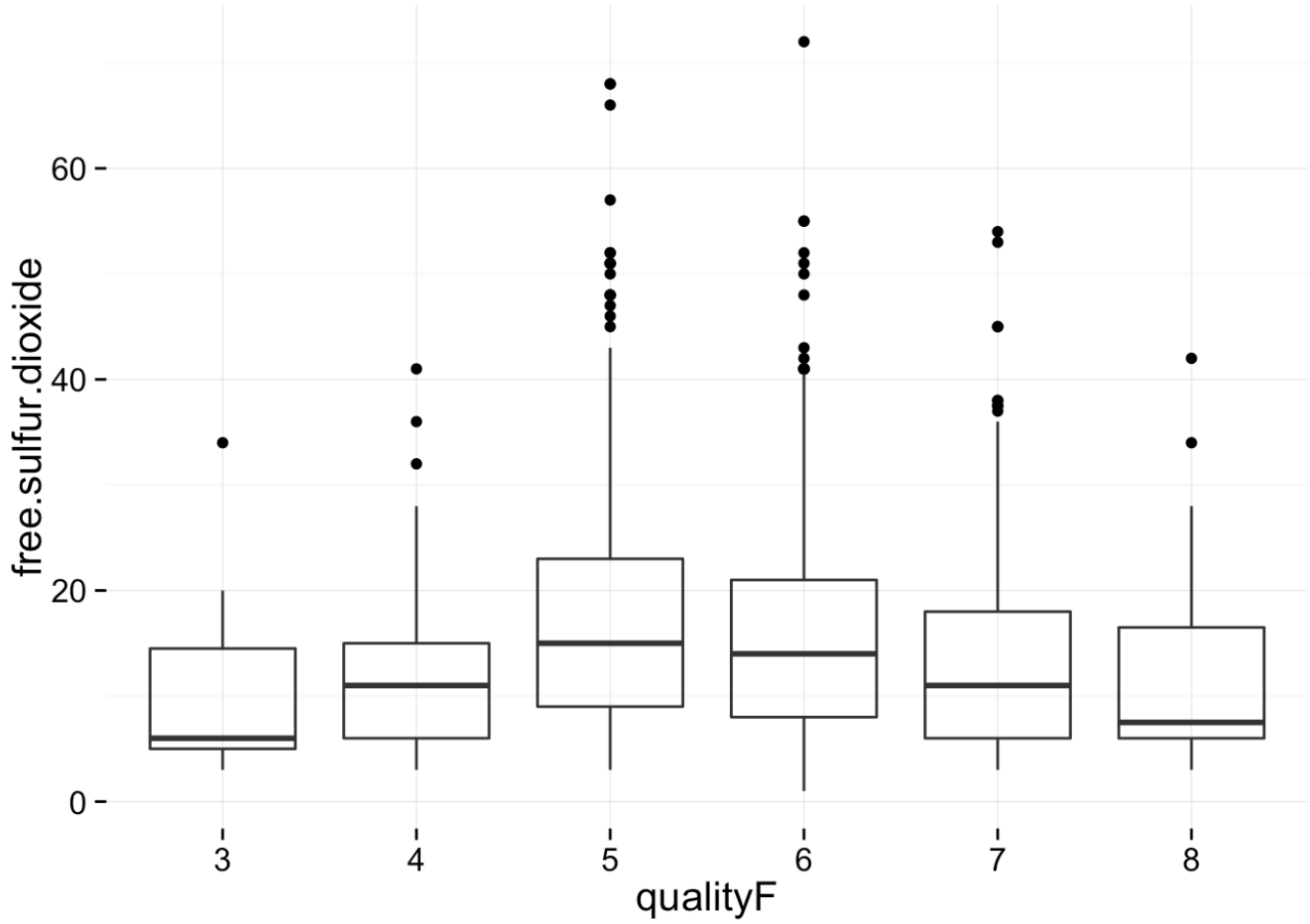
```
##
##  Pearson's product-moment correlation
##
## data:  redwine$quality and redwine$volatile.acidity
## t = -16.9542, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4313210 -0.3482032
## sample estimates:
##       cor
## -0.3905578
```

QUALITY AND VOLATILE ACIDITY:
To better visualize the spread of values, I plotted a scatter plot and superimposed a boxplot. The plots show that higher volatile.acidity is generally associated with lower quality wines. Examining the summaries for different qualities, the median is 0.845 for the lowest quality of 3. The median decreases to 0.37 as quality increases to 7.

The mean volatile.acidity for quality = 8 is 0.4233, slightly higher than that at quality = 7, and the 2 medians are the same at 0.37. It could be that there is an optimum volatile.acidity range of around 0.37 - 0.42, so there isn't much variation in higher quality wines.

In addition, this data set only has 18 wines that have quality = 8. Such a small sample may not be a good representation of the volatile.acidity levels in high quality wines.

Correlation between the volatile.acidity and quality was found to be -0.39.




```
## redwine$qualityF: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     3.0     5.0     6.0    11.0    14.5    34.0
## ------------------------------------------------------------
## redwine$qualityF: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.00    6.00   11.00   12.26   15.00   41.00
## ------------------------------------------------------------
## redwine$qualityF: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.00    9.00   15.00   16.98   23.00   68.00
## ------------------------------------------------------------
## redwine$qualityF: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    8.00   14.00   15.71   21.00   72.00
## ------------------------------------------------------------
## redwine$qualityF: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.00    6.00   11.00   14.05   18.00   54.00
## ------------------------------------------------------------
## redwine$qualityF: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.00    6.00    7.50   13.28   16.50   42.00
```

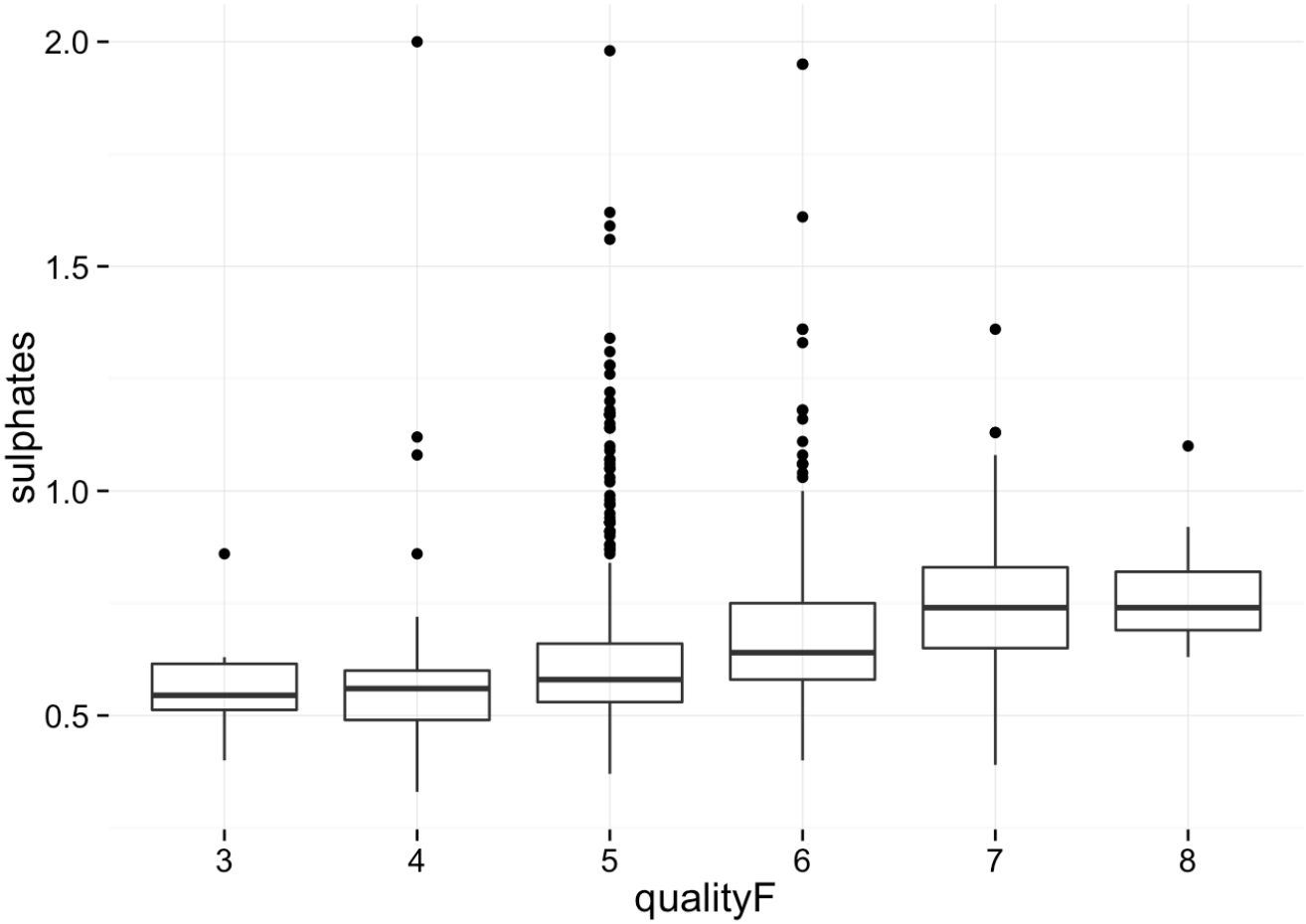
QUALITY AND FREE SULFUR DIOXIDE:

Average wines of quality 5 and 6 appear to have a wider range of free SO2 levels and higher median values of 14-15.

Wines of other qualities have free SO2 levels that are generally below 20ppm. The spread of free.SO2 values for poor quality wines (=3) and high quality wines (=8) appear to be very similar which is quite puzzling to me and all levels are below 50ppm.

Perhaps free SO2 has no impact on quality so long as it's below 50ppm? Wines rated poorly or highly may have gotten their extreme ratings as a result of other variables.

Sulphates affect the amount of SO2 so I will take a look at how quality varies with sulphate levels.




```
## continuous_scale(aesthetics = c("y", "ymin", "ymax", "yend",
##     "yintercept", "ymin_final", "ymax_final"), scale_name = "position_c",
##     palette = identity, breaks = ..2, limits = ..1, expand = expand,
##     guide = "none")
```

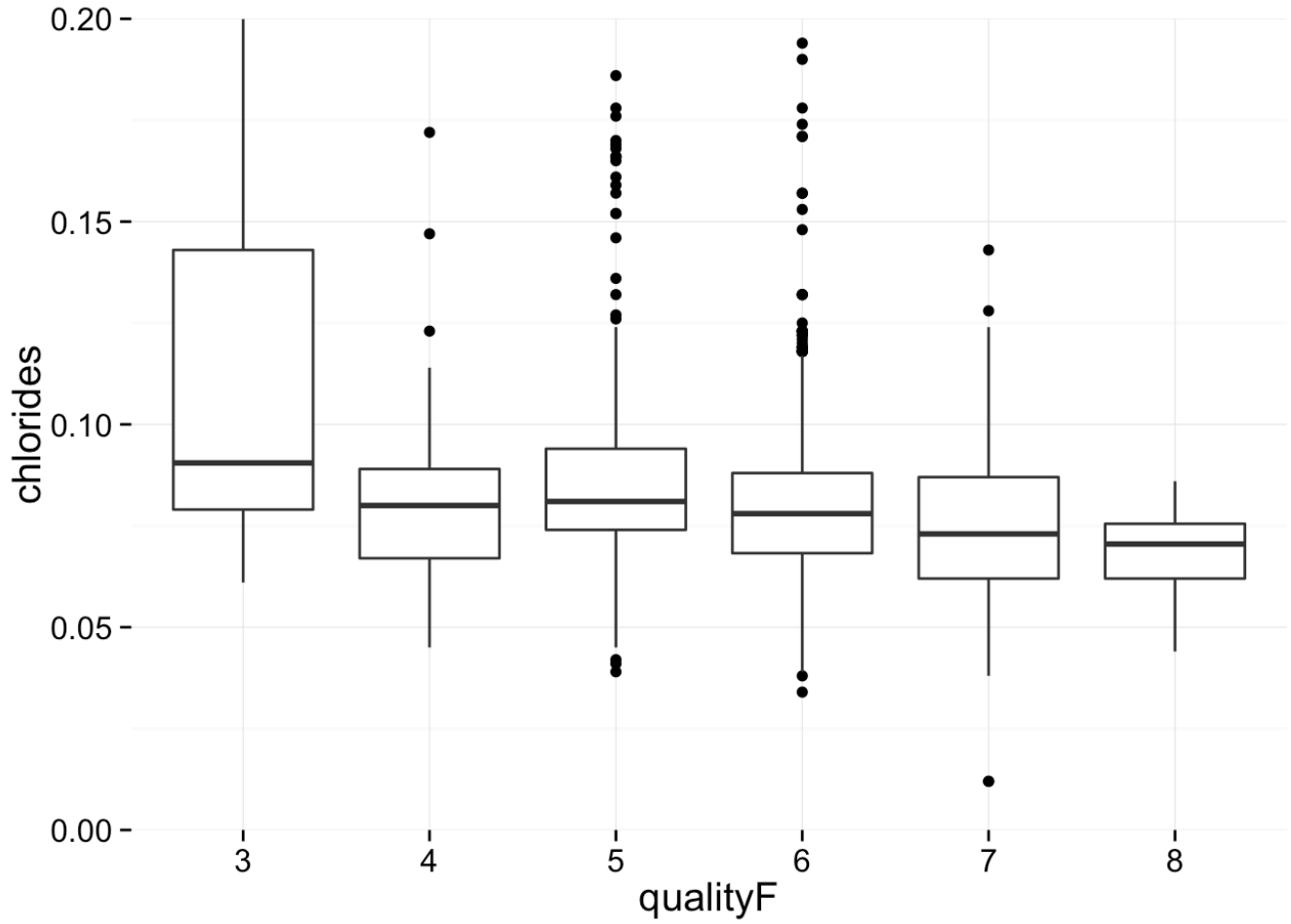
```
## redwine$qualityF: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.4000  0.5125  0.5450  0.5700  0.6150  0.8600
## ------------------------------------------------------------
## redwine$qualityF: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3300  0.4900  0.5600  0.5964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

```
##
##  Pearson's product-moment correlation
##
## data:  redwine$sulphates and redwine$quality
## t = 10.3798, 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
```

QUALITY AND SULPHATES:
Sulphate content appears to be a better predictor of the quality of wines. Generally, higher quality wines appear to have higher sulphate levels. The medians for quality = 7 and 8 are the same at 0.74, while the median for quality = 3 is 0.545.

I will use this variable in place of free SO2 for subsequent analyses to include the effect of antioxidative and microbial properties

correlation = 0.25

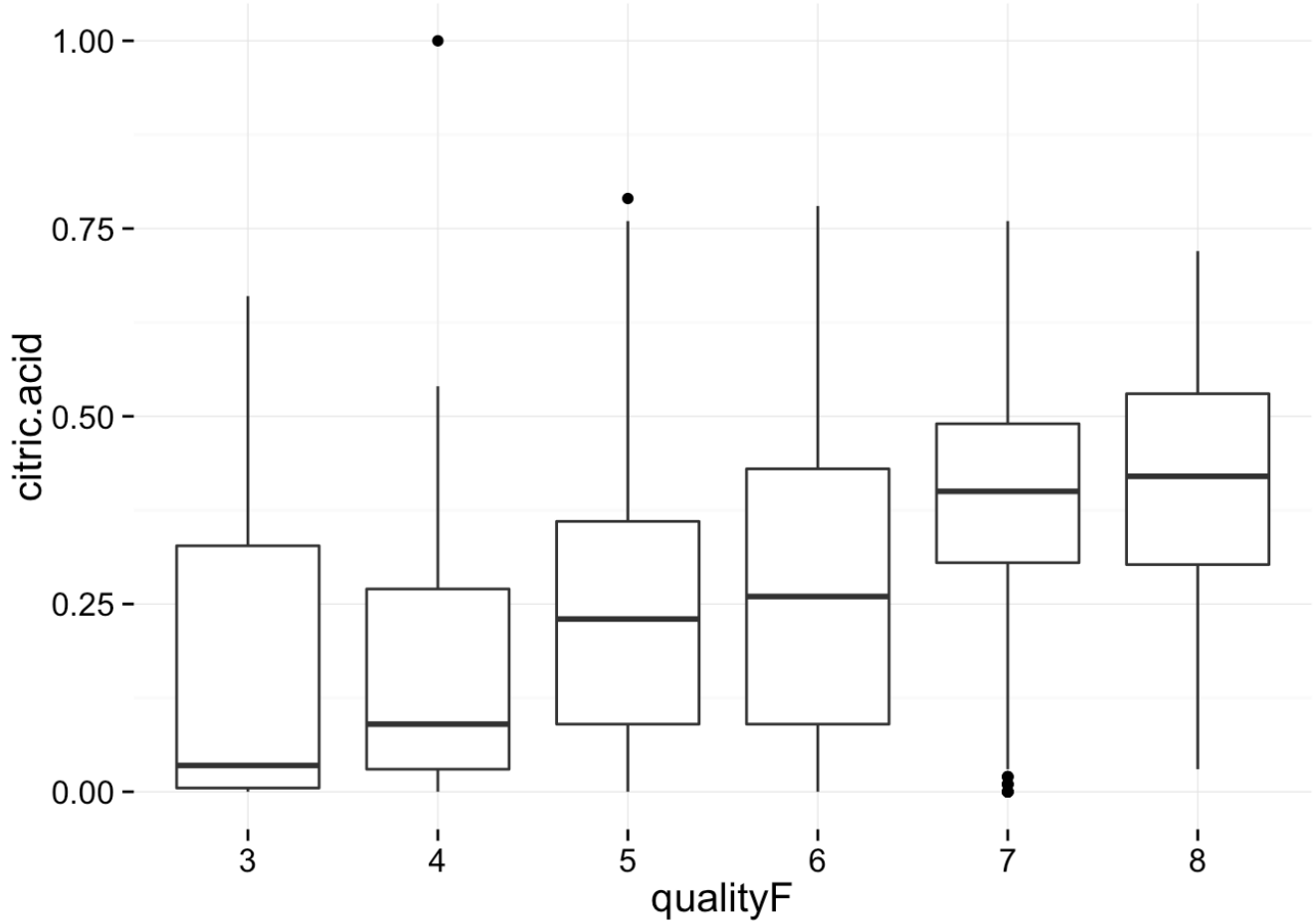



```
## redwine$qualityF: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0610  0.0790  0.0905  0.1225  0.1430  0.2670
## ------------------------------------------------------------
## redwine$qualityF: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.04500 0.06700 0.08000 0.09068 0.08900 0.61000
## ------------------------------------------------------------
## redwine$qualityF: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.03900 0.07400 0.08100 0.09274 0.09400 0.61100
## ------------------------------------------------------------
## redwine$qualityF: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.03400 0.06825 0.07800 0.08496 0.08800 0.41500
## ------------------------------------------------------------
## redwine$qualityF: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01200 0.06200 0.07300 0.07659 0.08700 0.35800
## ------------------------------------------------------------
## redwine$qualityF: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.04400 0.06200 0.07050 0.06844 0.07550 0.08600
```

QUALITY AND CHLORIDES:

The boxplot suggests that median chloride levels are relatively similar across the wines, but the spread of values is narrower for higher quality wines of 7 and 8. The extreme chloride values of around 0.6

observed in the univariate plot section are wines with quality of 4 and 5. Controlling chloride levels may be important to produce high quality wines.




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

```
##
##  Pearson's product-moment correlation
##
## data:  redwine$citric.acid and redwine$quality
## 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
```

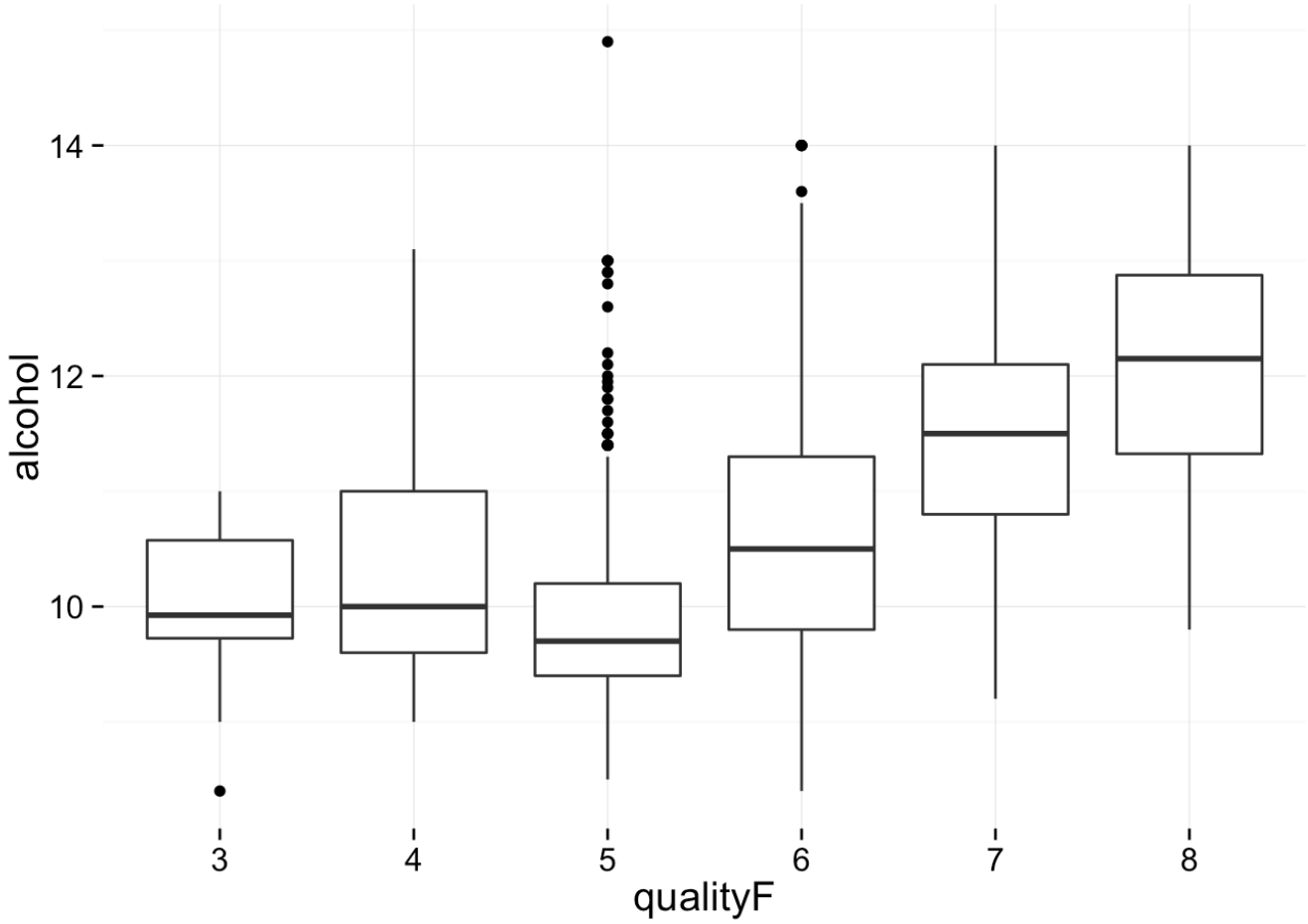
QUALITY AND CITRIC ACID:
Higher citric levels correspond to higher quality ratings. High quality wines of quality = 8 have median levels of 0.42, compared to 0.035 for low quality wines (quality = 3).

Considering high quality wines of 7/8, there isn't much difference in citric acid levels as quality improves from 7 to 8.

On the other hand, the increase in citric acid levels as quality improves from 3-6 is significant.

I suspect that a level of around 0.42 is the optimal citric acid level concentration, so we don't see appreciably higher citric levels in quality = 8 compared to 7.

correlation = 0.226

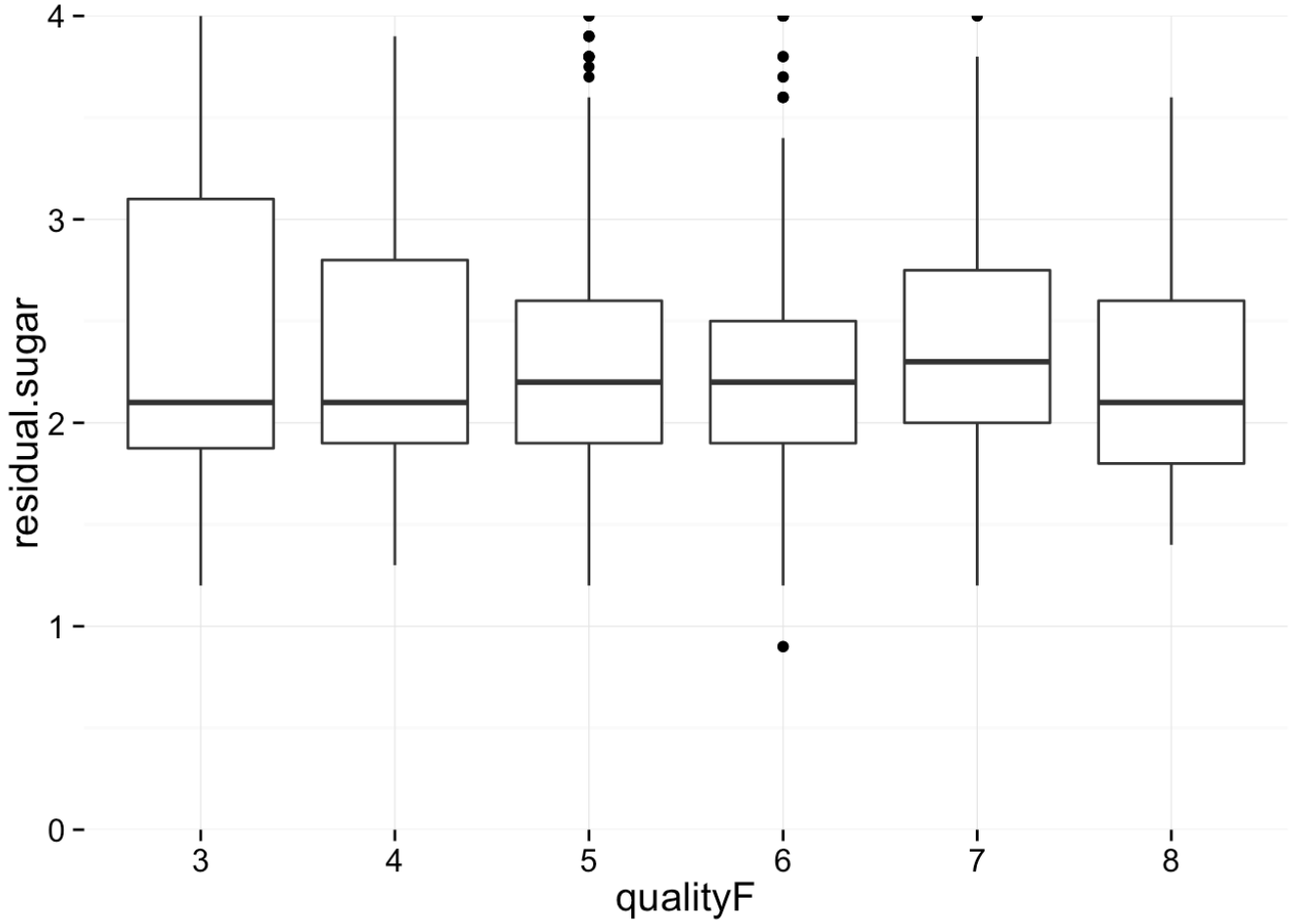


```
## redwine$qualityF: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   8.400   9.725   9.925   9.955  10.580  11.000
## ------------------------------------------------------------
## redwine$qualityF: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.00    9.60   10.00   10.27   11.00   13.10
## ------------------------------------------------------------
## redwine$qualityF: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     8.5     9.4     9.7     9.9    10.2    14.9
## ------------------------------------------------------------
## redwine$qualityF: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.80   10.50   10.63   11.30   14.00
## ------------------------------------------------------------
## redwine$qualityF: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.20   10.80   11.50   11.47   12.10   14.00
## ------------------------------------------------------------
## redwine$qualityF: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.80   11.32   12.15   12.09   12.88   14.00
```

```
##
##  Pearson's product-moment correlation
##
## data:  redwine$alcohol and redwine$quality
## t = 21.6395, 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
```

QUALITY AND ALCOHOL:
Higher alcohol content is associated with higher quality wines. Median values for quality = 6/7/8 are 10.50 - 12.15, compared to 10.00 or less for lower quality wines.

correlation = 0.476

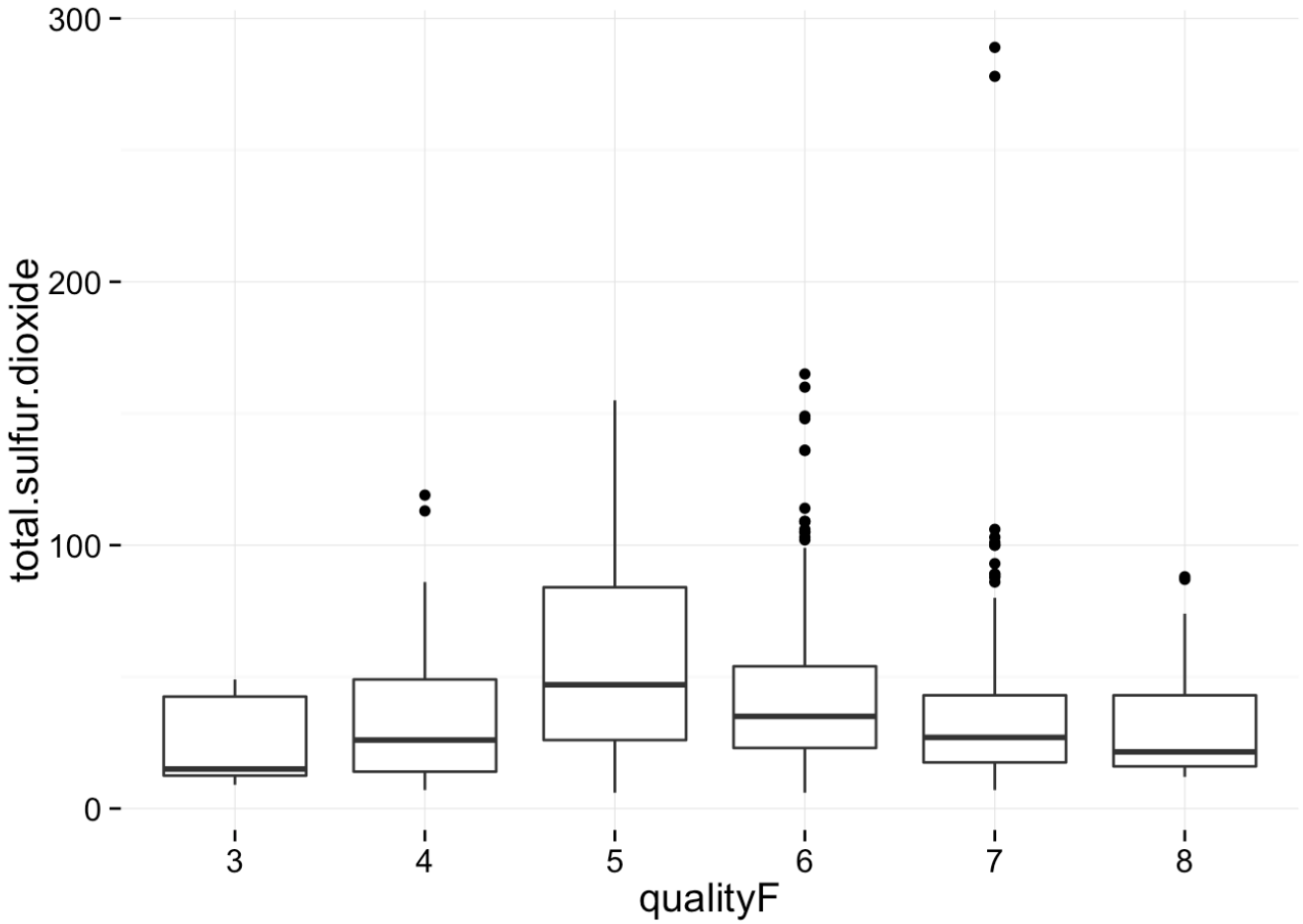



```
## redwine$qualityF: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.200   1.875   2.100   2.635   3.100   5.700
## ------------------------------------------------------------
## redwine$qualityF: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.300   1.900   2.100   2.694   2.800  12.900
## ------------------------------------------------------------
## redwine$qualityF: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.200   1.900   2.200   2.529   2.600  15.500
## ------------------------------------------------------------
## redwine$qualityF: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.900   1.900   2.200   2.477   2.500  15.400
## ------------------------------------------------------------
## redwine$qualityF: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.200   2.000   2.300   2.721   2.750   8.900
## ------------------------------------------------------------
## redwine$qualityF: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.400   1.800   2.100   2.578   2.600   6.400
```

```
##
##  Pearson's product-moment correlation
##
## data:  redwine$residual.sugar and redwine$quality
## t = 0.5488, df = 1597, p-value = 0.5832
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.03531327  0.06271056
## sample estimates:
##       cor
## 0.01373164
```

QUALITY AND RESIDUAL SUGAR:

The median values of residual sugar are similar across the different qualities. The spread of values appear to be wider for the average wines. Otherwise, residual.sugar doesn't seem to affect quality. This is supported by the low correlation value of 0.0137.

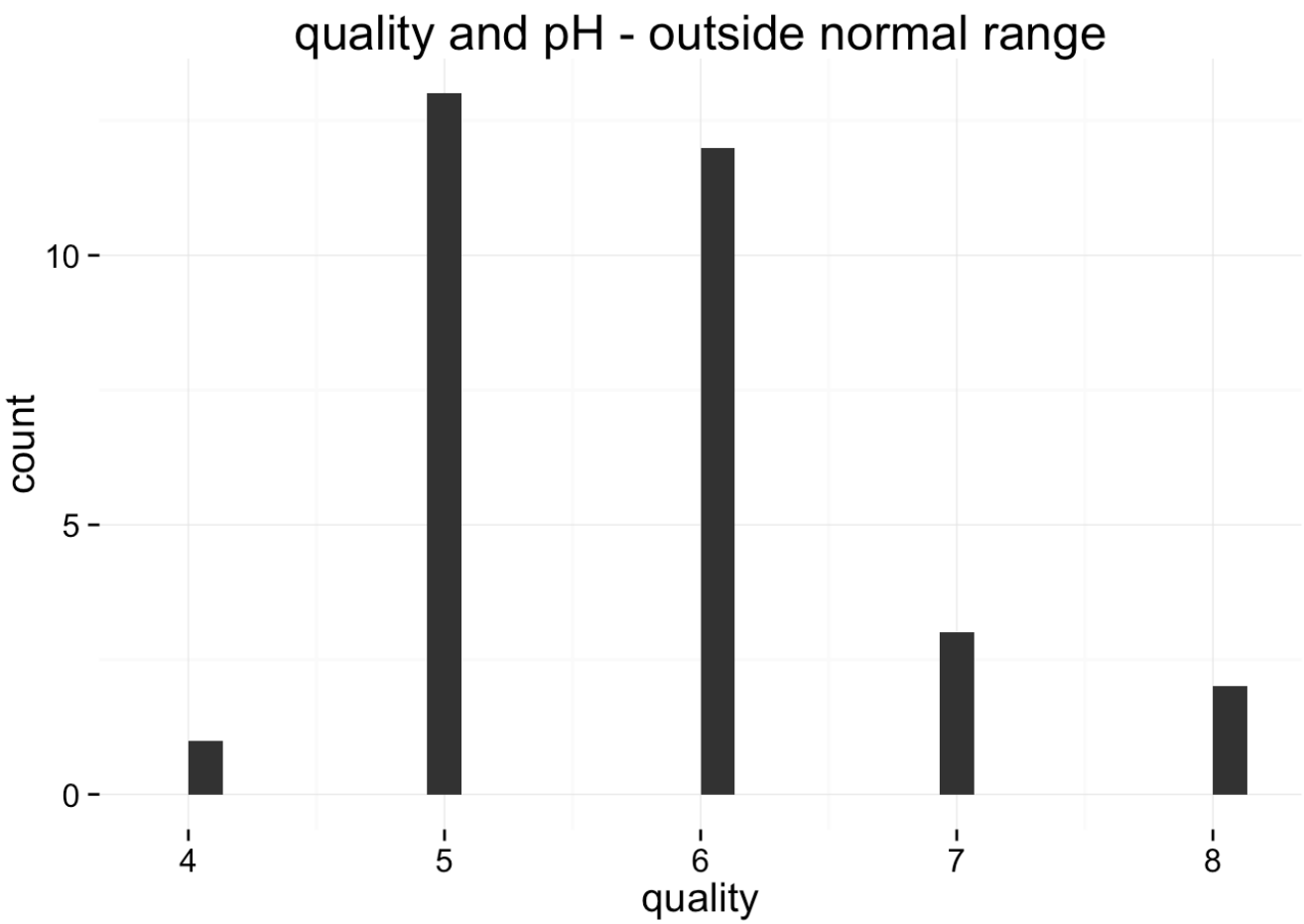


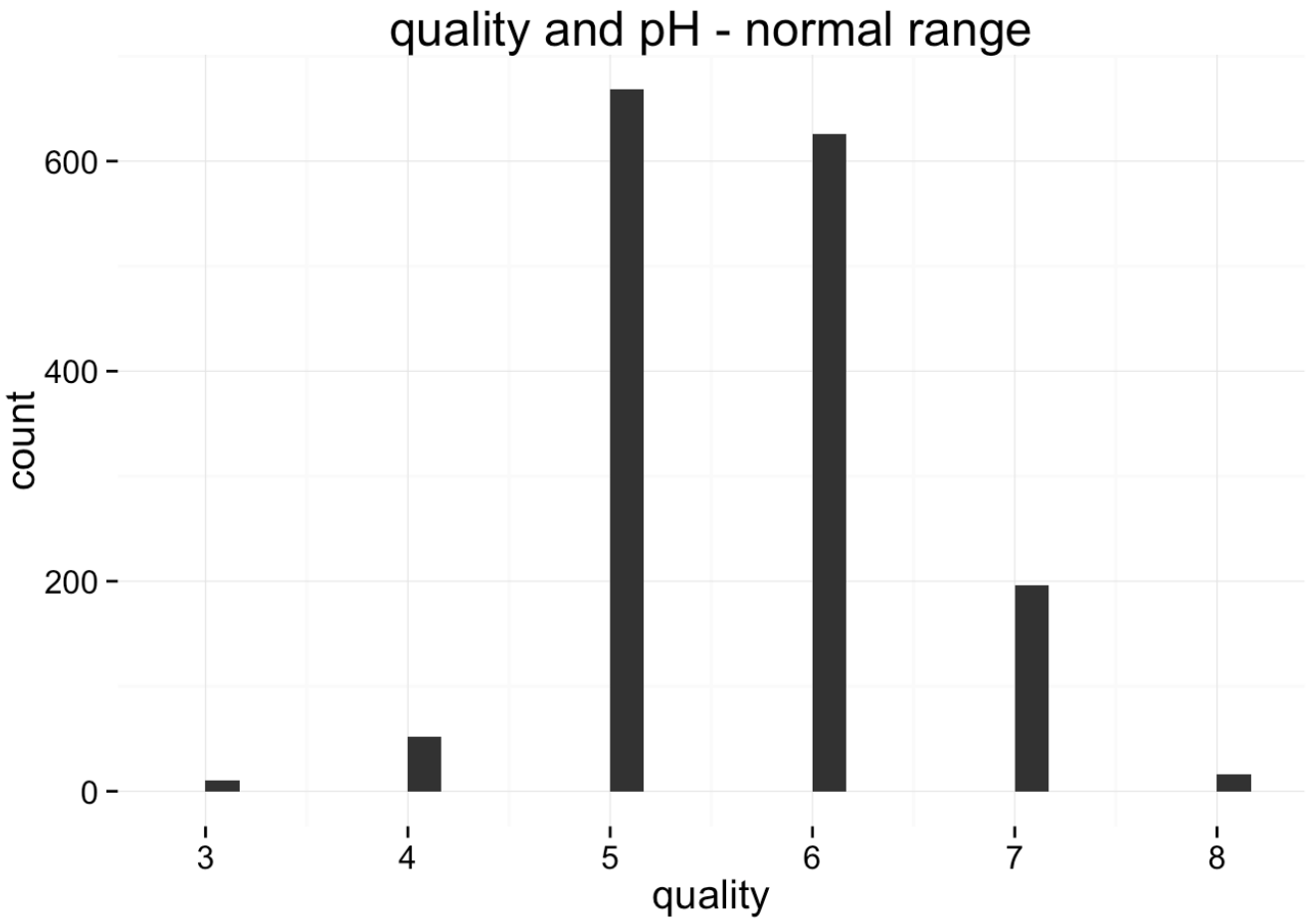
```
## redwine$qualityF: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     9.0    12.5    15.0    24.9    42.5    49.0
## ------------------------------------------------------------
## redwine$qualityF: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.00   14.00   26.00   36.25   49.00  119.00
## ------------------------------------------------------------
## redwine$qualityF: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00   26.00   47.00   56.51   84.00  155.00
## ------------------------------------------------------------
## redwine$qualityF: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00   23.00   35.00   40.87   54.00  165.00
## ------------------------------------------------------------
## redwine$qualityF: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.00   17.50   27.00   35.02   43.00  289.00
## ------------------------------------------------------------
## redwine$qualityF: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.00   16.00   21.50   33.44   43.00   88.00
```

```
##
##  Pearson's product-moment correlation
##
## data:  redwine$total.sulfur.dioxide and redwine$quality
## 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
```

QUALITY AND TOTAL SULFUR DIOXIDE:

The median total SO2 values are higher for the average wines, and poor and high quality wines have similar median values. The extreme total SO2 values observed earlier are for wines with a quality of 7. Correlation between the 2 variables is -0.185. Based on these observations, I don’t think total sulfur dioxide has an impact on quality.


quality and pH - outside normal range
count
10
5
0
4
5
6
7
8
quality

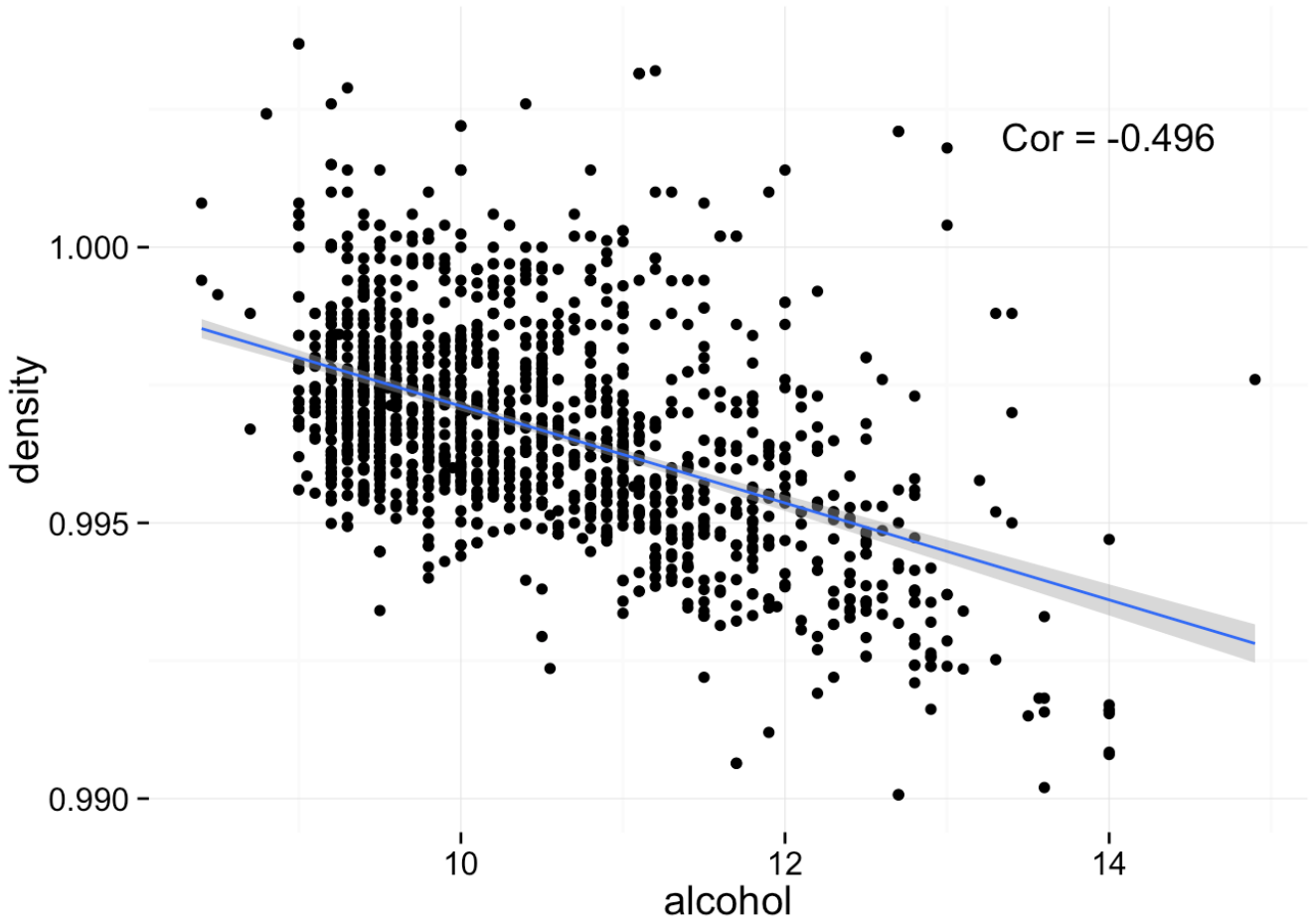

quality and pH - normal range
count
600
400
200
0
3
4
5
6
7
8
quality

QUALITY AND pH:
Wines with pH values not in the normal range of 3-4 appear to have the same quality rating distribution as wines in the normal range. Possibly small deviations from the norm do not affect quality (the min pH for this data set was around 2.7, not too far from the normal range)

Relationships between pairs of variables

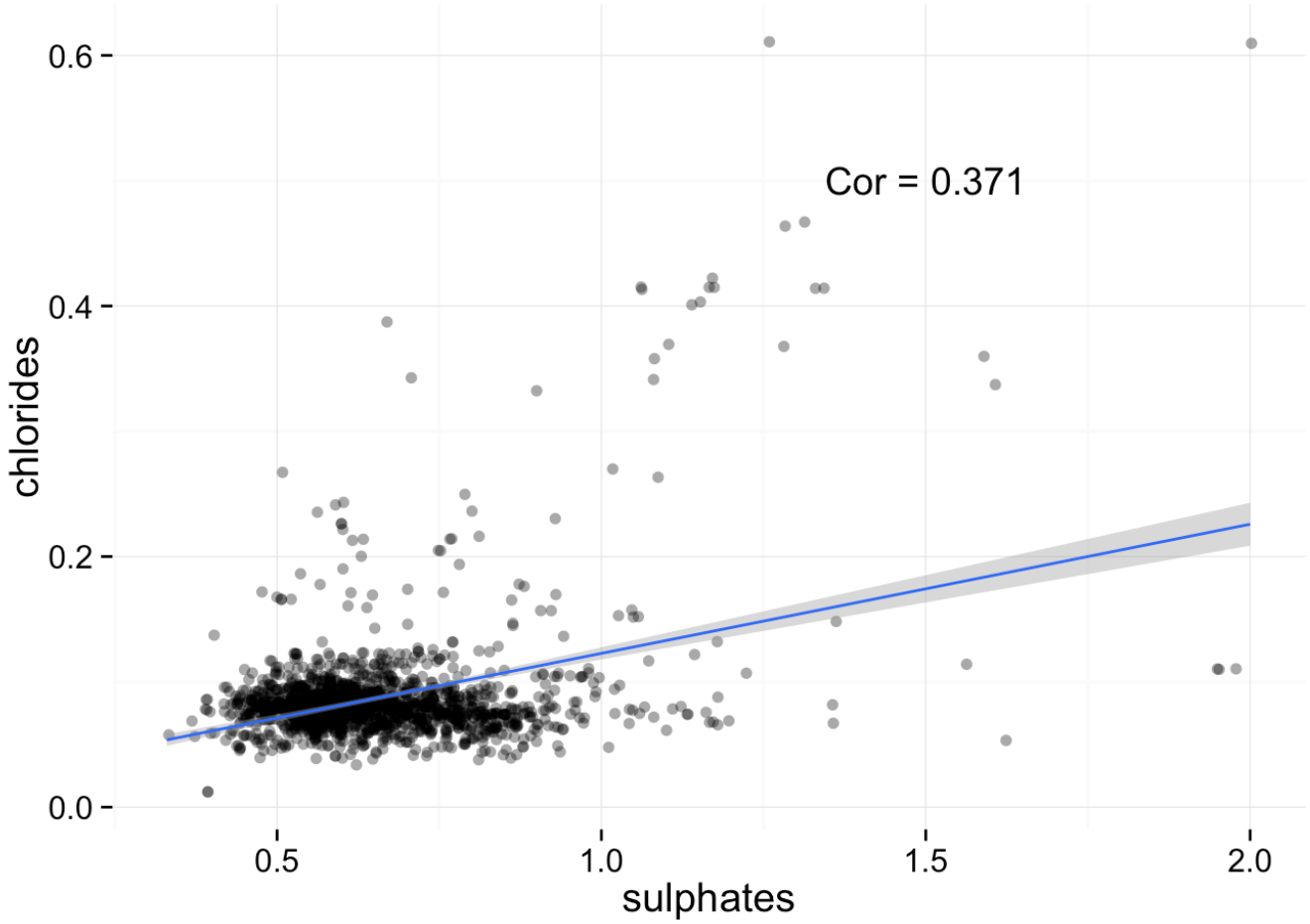
Based on the previous bivariate plots, I deduced that the following features could impact quality: volatile.acidity, sulphates, citric acid, alcohol. Using the correlation matrix, I identified pairs of variables containing these features, and investigated those with a correlation of greater than 0.25.

```
##
##  Pearson's product-moment correlation
##
## data:  redwine$alcohol and redwine$density
## t = -22.8382, 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
```

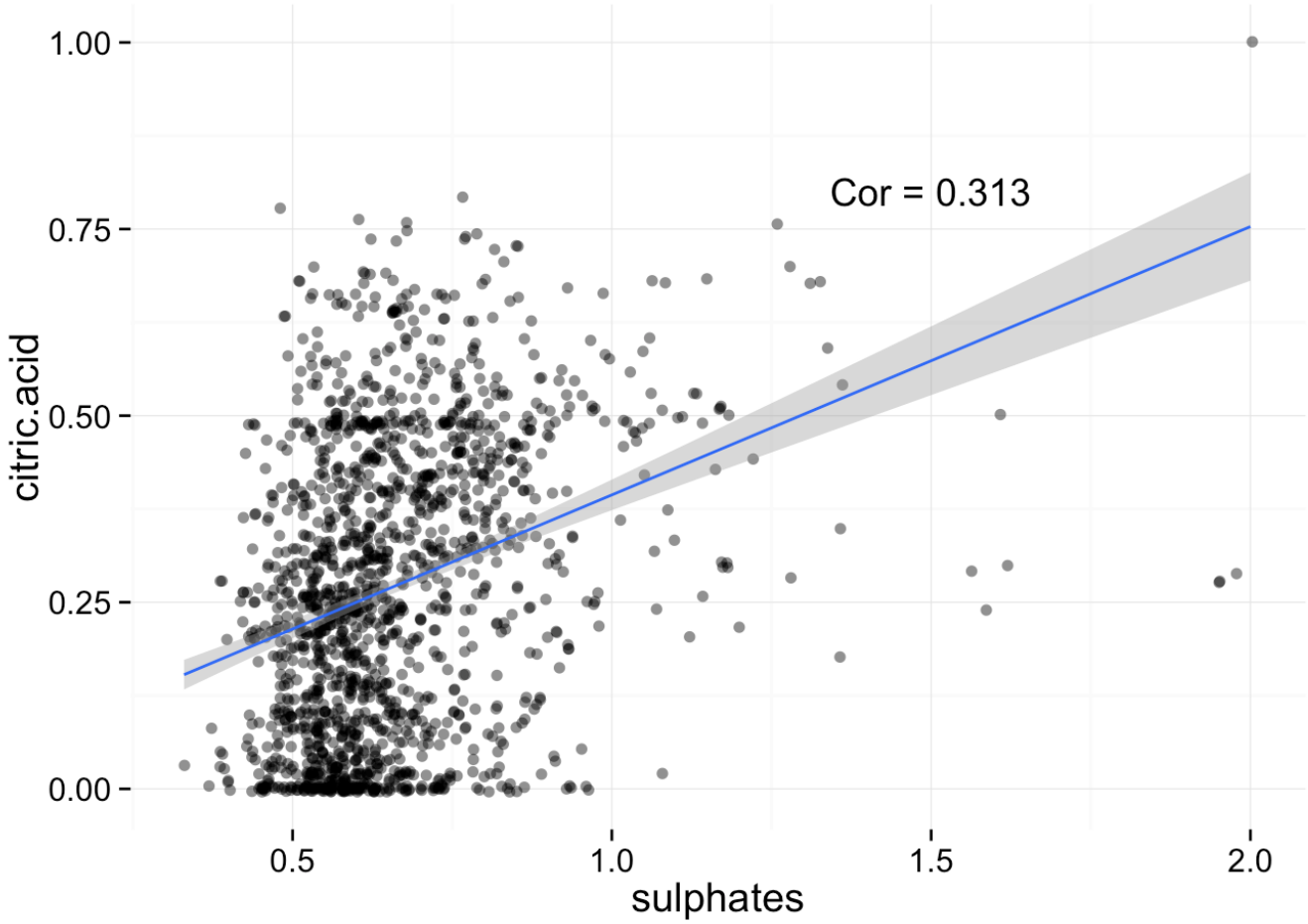


ALCOHOL AND DENSITY: Alcohol content has a negative relationship with density. No surprises here since alcohol is less dense than water.

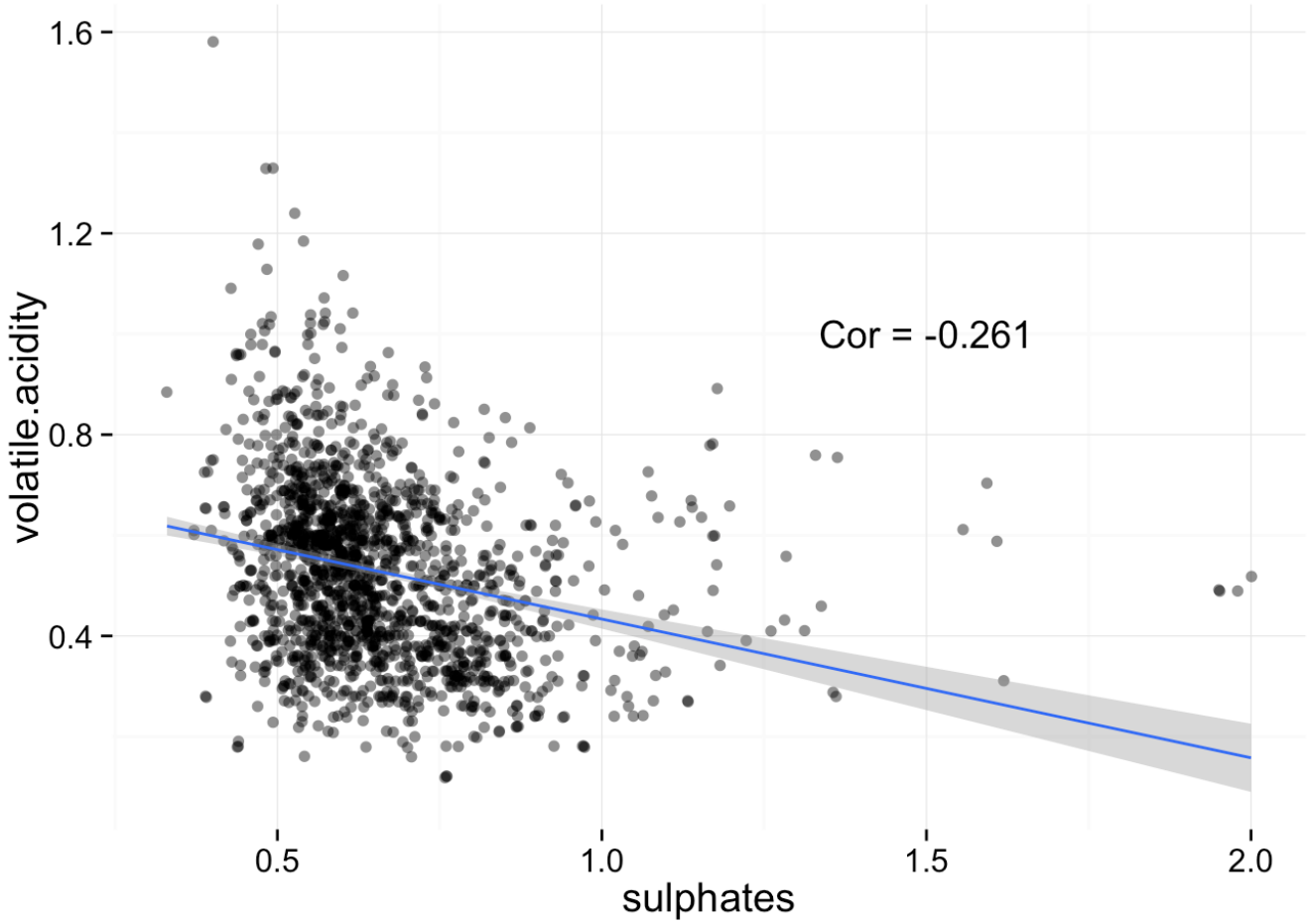
```
##
##  Pearson's product-moment correlation
##
## data:  redwine$sulphates and redwine$chlorides
## t = 15.9785, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3282127 0.4127694
## sample estimates:
##       cor
## 0.3712605
```




```
##
##  Pearson's product-moment correlation
##
## data:  redwine$sulphates and redwine$citric.acid
## t = 13.1593, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2678558 0.3563278
## sample estimates:
##     cor
## 0.31277
```

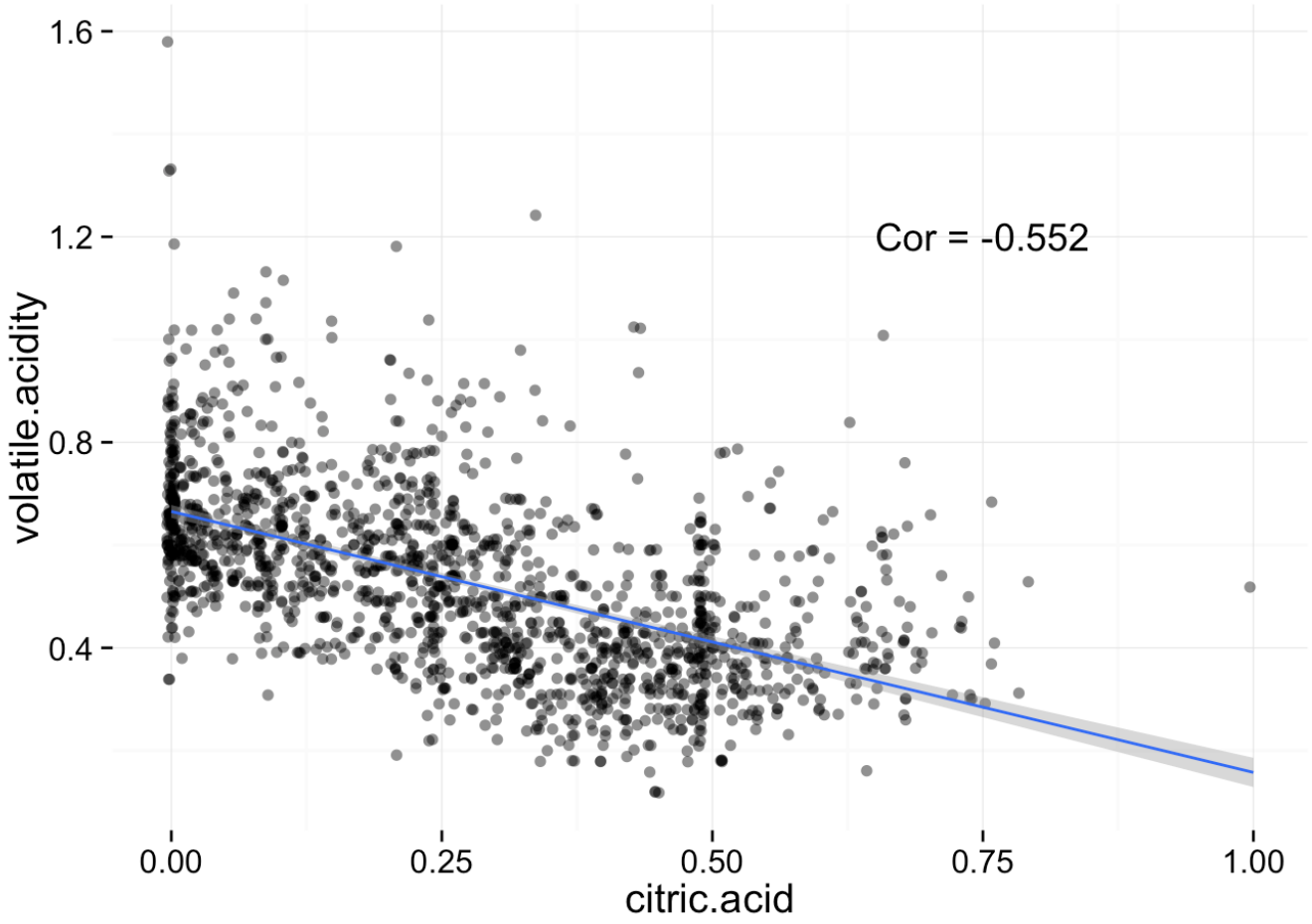



```
##
##  Pearson's product-moment correlation
##
## data:  redwine$sulphates and redwine$volatile.acidity
## t = -10.8041, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3060917 -0.2147125
## sample estimates:
##        cor
## -0.2609867
```

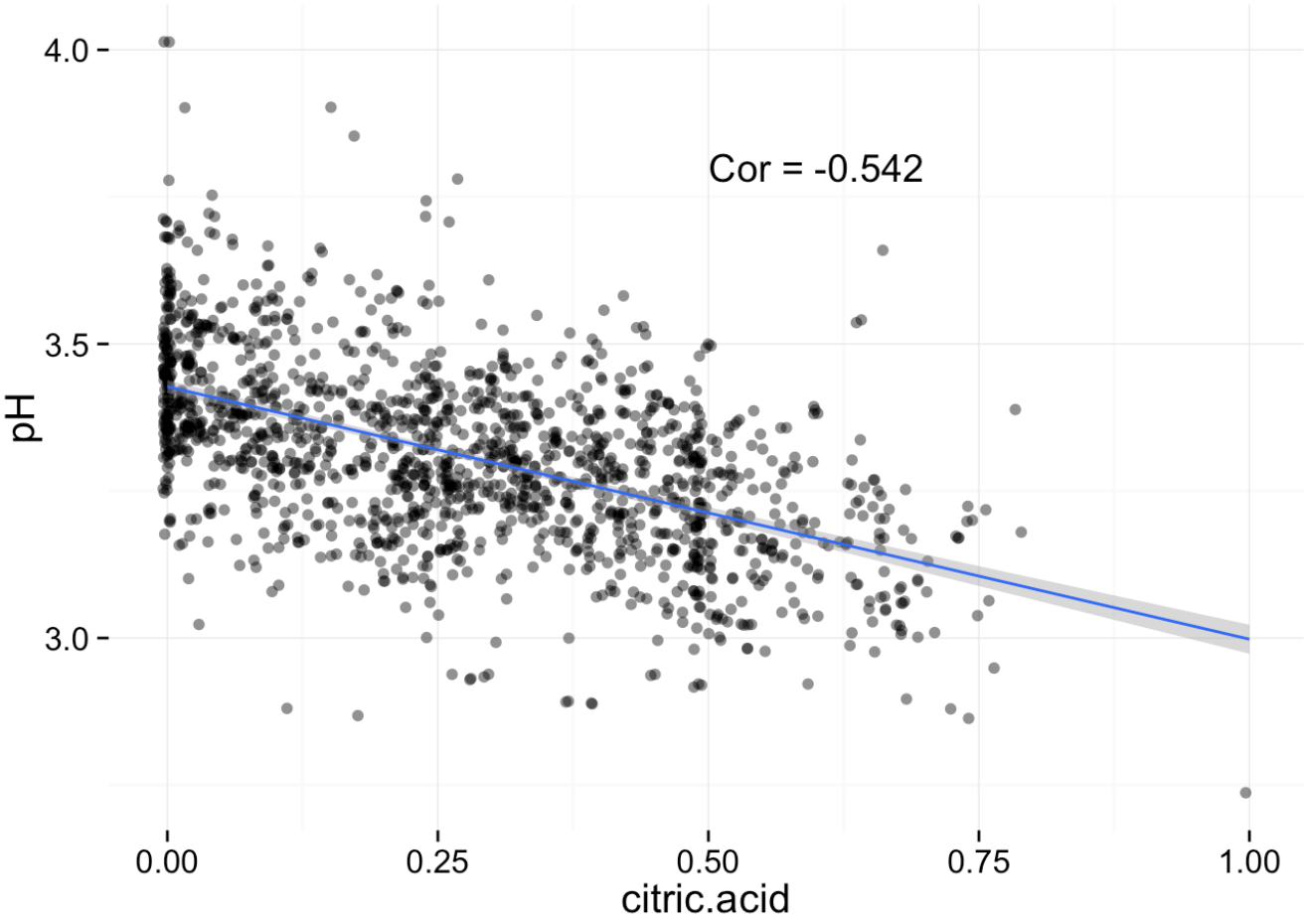


SULPHATES AND OTHER FEATURES:
Sulphates has positive relationship with chlorides and citric acid, with correlation of 0.371 and 0.313 respectively. It has an inverse relationship with volatile.acidity with correlation of -0.261. The dependencies here are probably a result of the equilibrium of H+ in wine

```
##
##  Pearson's product-moment correlation
##
## data:  redwine$volatile.acidity and redwine$citric.acid
## t = -26.4891, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5856550 -0.5174902
## sample estimates:
##        cor
## -0.5524957
```



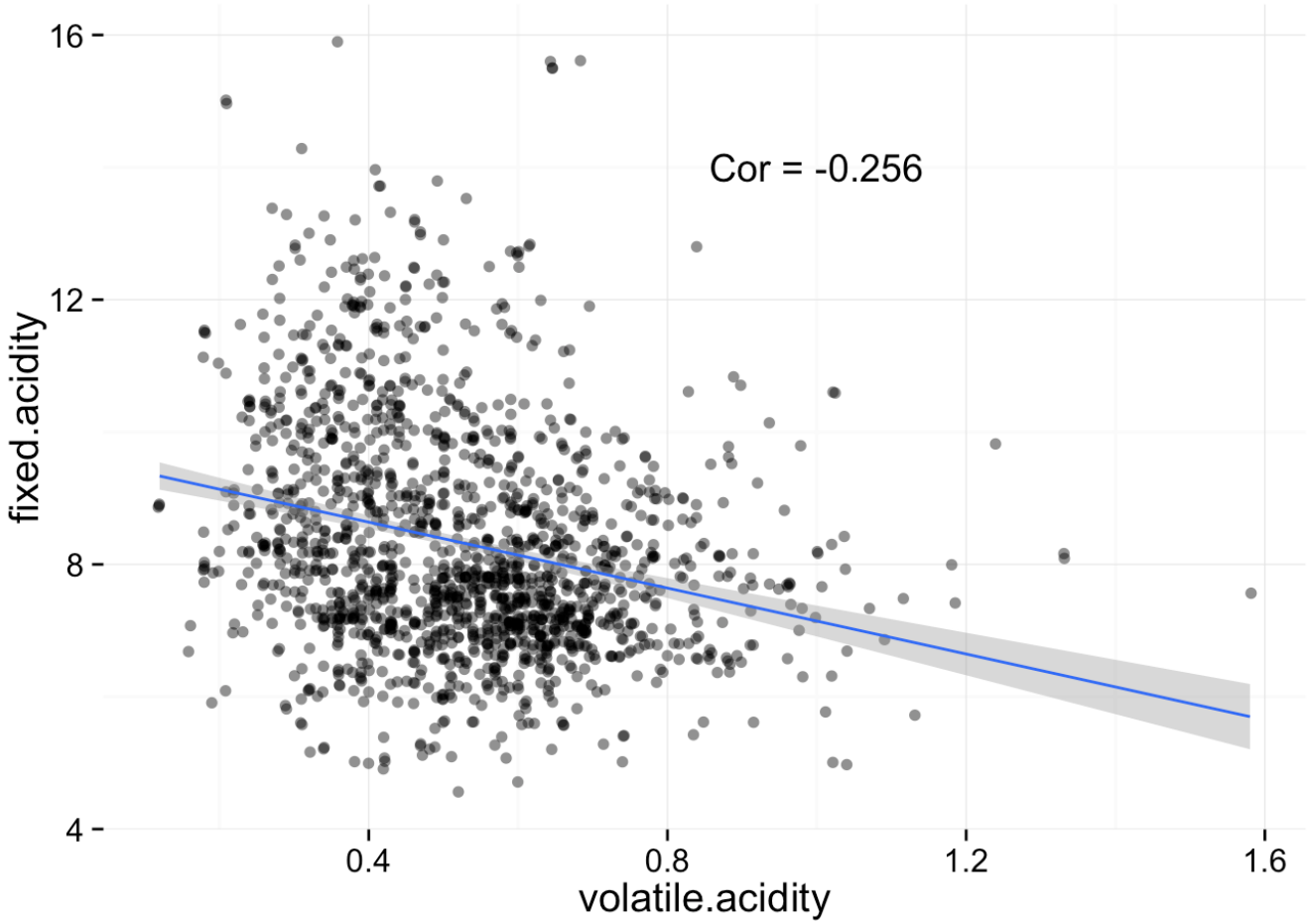

```
##
##  Pearson's product-moment correlation
##
## data:  redwine$pH and redwine$citric.acid
## t = -25.7672, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5756337 -0.5063336
## sample estimates:
##       cor
## -0.5419041
```



CITRIC ACID AND OTHER FEATURES:
Citric acid has a negative relationship with volatile acidity and pH, with correlation of -0.552 and -0.542 respectively. As citric acid concentration increases, we naturally expect that pH will decrease as wine becomes more acidic.

Volatile acidity indicates the concentration of acetic acid present. From the inverse relationship and my previous life as a chemical engineer, I understand this to be a result of chemical equilibira involving H+ ions and competing acid species. As the concentration of one acid species increases, other acid species will tend to stay in dissociated ion form, rather than exist as an acid. So increases in one acid concentration would decrease others.

```
##
##  Pearson's product-moment correlation
##
## data:  redwine$volatile.acidity and redwine$fixed.acidity
## t = -10.5888, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3013681 -0.2097433
## sample estimates:
##        cor
## -0.2561309
```



FIXED ACIDITY AND VOLATILE ACIDITY:
Volatile acidity has a negative relationship with fixed acidity, with correlation of -0.256. This can be attributed to the same reason of competing acid species.

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?

Positive correlations were found between:

- alcohol content and quality (correlation = 0.48)
- citric acid and quality (correlation = 0.23)
- sulphates and quality (correlation = 0.25)

Negative correlation was found between:

- volatile.acidity and quality (correlation = -0.39)

The correlations between alcohol/volatile.acidity and quality are more significant.

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

Citric acid has an inverse relationship with fixed and volatile acidity, while sulphates has a positive relationship with chlorides and citric acid.

Possibly, volatile acidity could be manipulated to control the citric acid concentration, to improve wine quality.

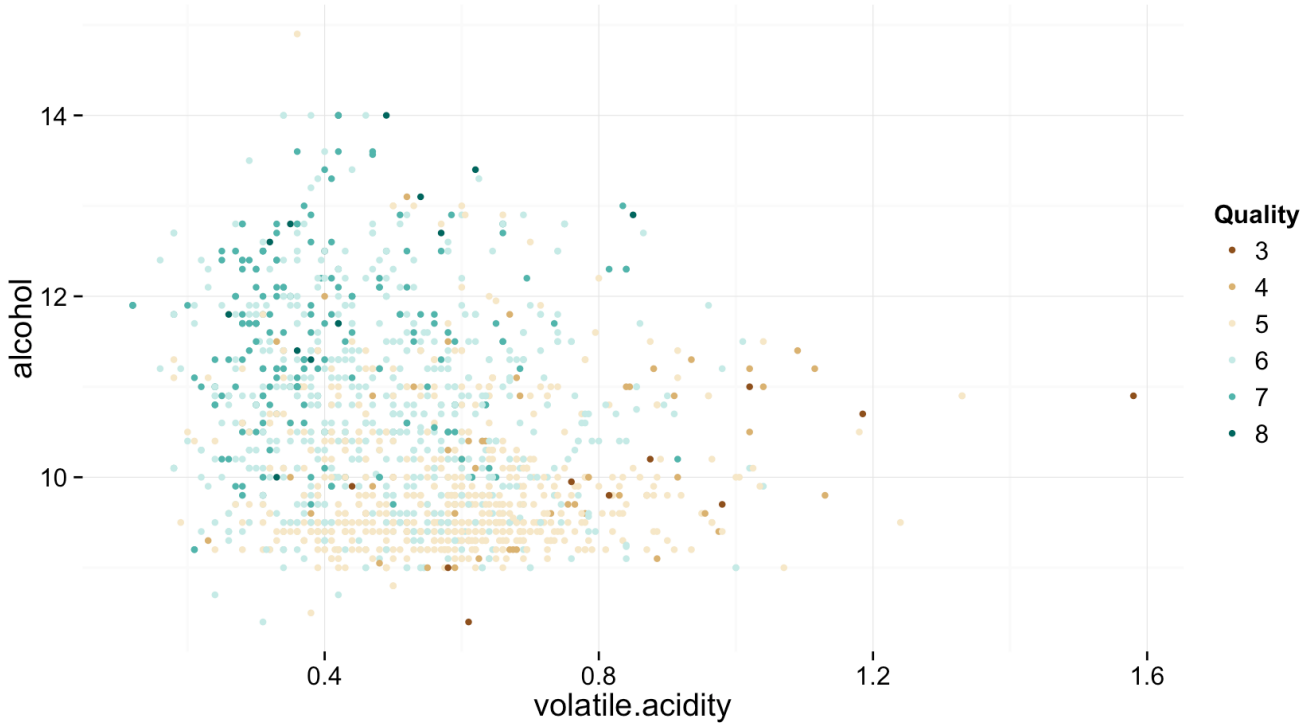
Sulphates and citric acid were found earlier to be positively related to wine quality, and are used for different purposes - antimicrobial/antioxidative functions and taste respectively. They are also positively correlated to each other so it may be interesting to see how wine quality varies as these 2 properties change.

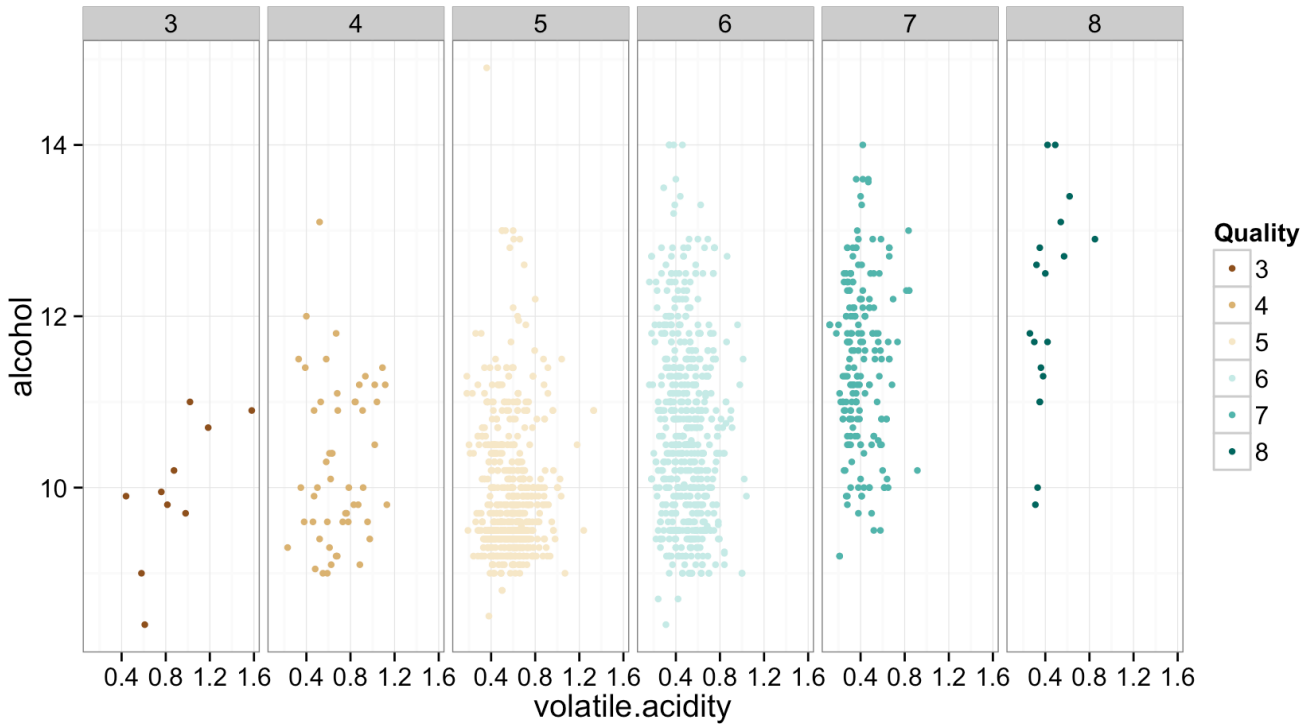
Chlorides were earlier shown to have a neutral - negative effect on wine quality. As such, even though a positive relationship exists between sulphates and chlorides, it wouldn't be prudent to increase chloride concentration for the sake of increasing sulphates as that could be detrimental for wine quality.

What was the strongest relationship you found?

Strongest relationship was between alcohol content and quality.

Multivariate Plots Section

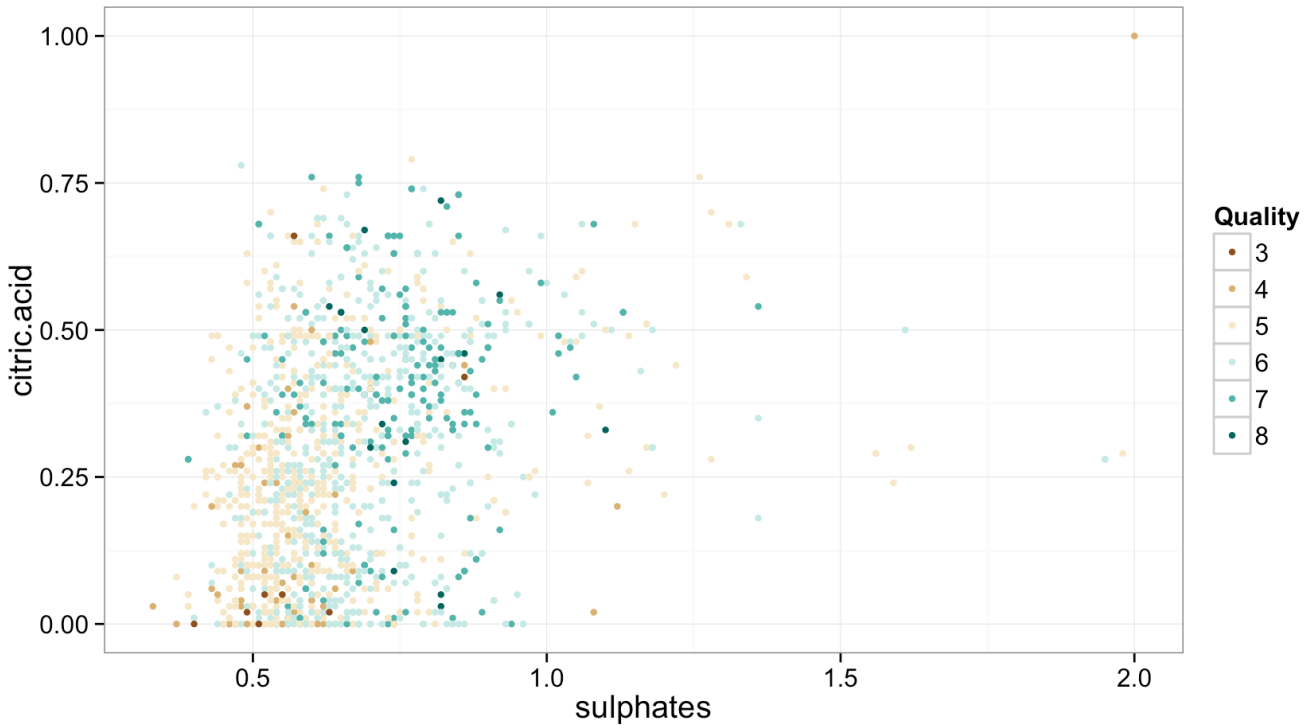


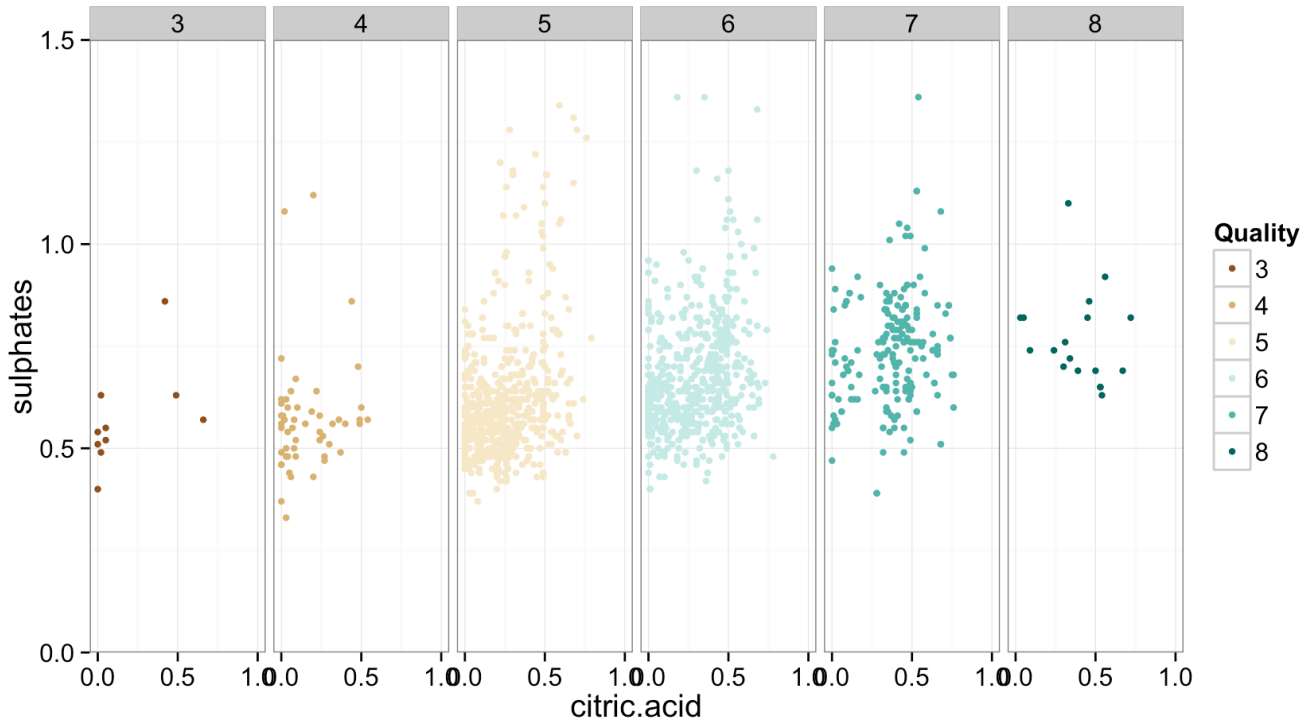


ALCOHOL, VOLATILE ACIDITY AND QUALITY:

First I plotted the 2 variables found to most strongly affect quality, alcohol and volatile acidity. The first plot shows that higher quality wines of 7/8 appear to be concentrated towards the upper left region (higher alcohol and lower volatile acidity) while lower quality wines of 3/4 seem to be concentrated in the lower right region. Average wines of 5/6 are more randomly spread on the plot, but appear to be intermediate to the 2 extreme clusters, where wines with quality 5 are clustered closer to the low quality wines and wines of quality 6 clustered closer to the high quality wines.

To better visualize the trends, I used facet_grid to create a second plot. The new plot shows that as quality increases, the data points tend to be clustered more to the upper left hand of each panel.

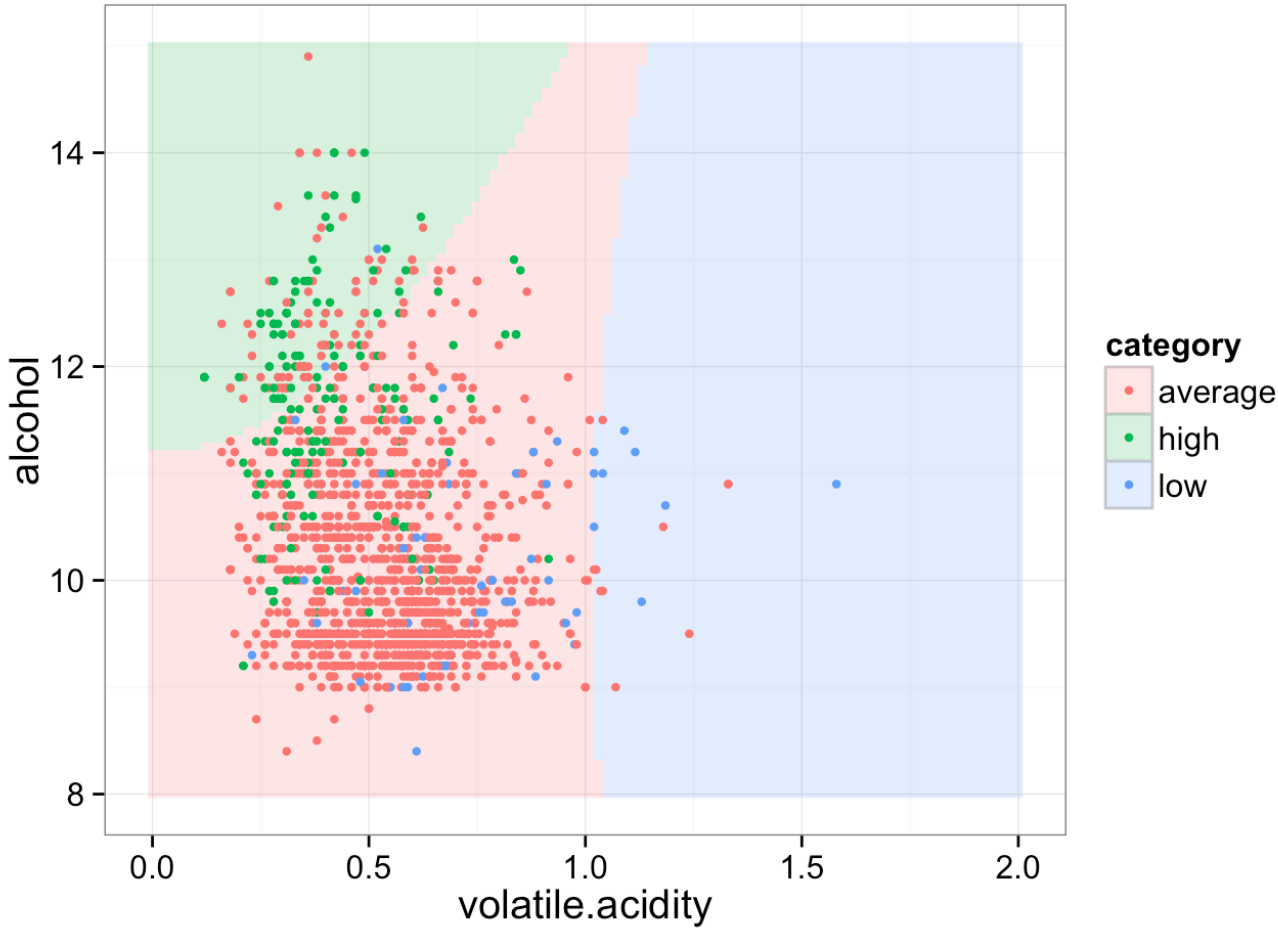




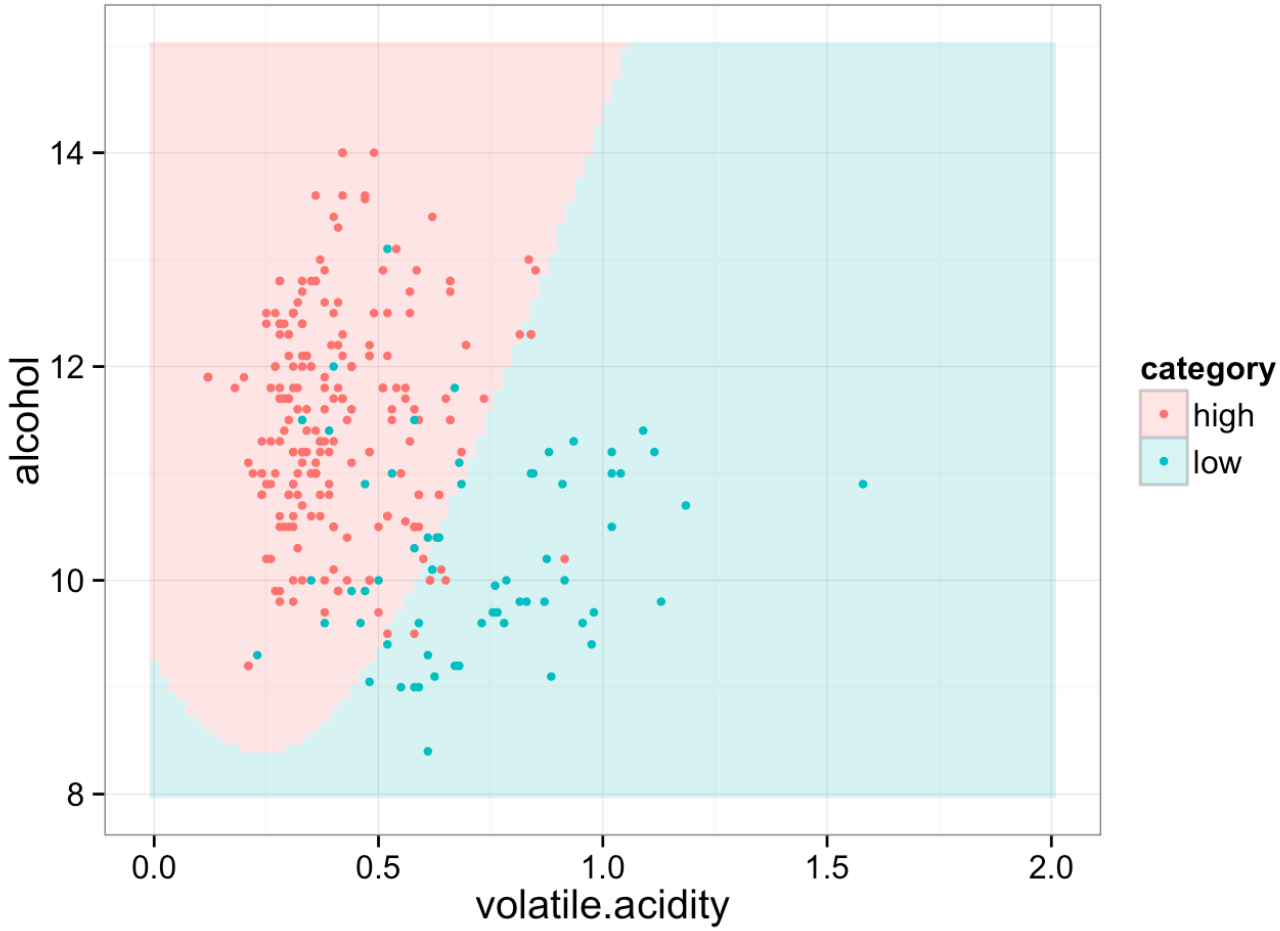
SULPHATES, CITRIC ACID AND QUALITY:

Plotting the other 2 features that affect quality, citric acid and sulphates, we see a progression towards higher citric acid and sulphate values for higher quality wines, albeit a less distinct trend compared to the alcohol vs volatile.acidity plot. I don't think this plot is as useful in charactierizing the wines as the previous plot.

```
##
## pred       average high  low
##   average     1232  141   54
##   high          77   76    2
##   low           10    0    7
```




```
##
## pred      average high low
##   average       0    0   0
##   high          0  208  18
##   low           0    9  45
```



I wanted to see if I could delineate the regions on a alcohol vs volatile acidity plot to show the range of values that wines of different qualities were likely to have.

I looked at alcohol and volatile acidity because these were the 2 variables most strongly related to quality.

To simplify analysis, I transformed the quality variable into 3 main categories, ‘low’ quality wines (quality ratings of 3/4); ‘average’ quality wines (quality ratings of 5/6), ‘high’ quality wines (quality ratings of 7/8).

Using Naive Bayes, I trained a model to predict the category of wines based on volatile.acidity and alcohol features in order to cluster regions on a plot that would likely be of a certain category. The first plot suffers from overplotting and the average wines are spread all over so it’s not that useful.

From a production angle, I assume that average wines would be acceptable. On the other hand low quality wines should be minimized and producers would strive to maximize the number of high quality wines. On this basis, I’d be more interested to distinguish high quality wines from low quality wines while I might be neutral about average wines. As such, I subsetted my data set to include only low quality and high quality wines and repeated the plot.

The resulting plot is much clearer, with the red region showing the range of alcohol and volatile.acidity that high quality wines are likely to have, and the blue region showing the same, for low quality wines.

Predictive Models

```
##
##            0   1
##   FALSE   44   8
##   TRUE    19 209
```

```
## [1] 31.7265
```

```
## [1] 0.08160089
```

```
##
## pred      average high  low
##   average    1150   78   53
##   high        133  138    2
##   low          36    1    8
```

I tried creating some models using log regression, linear regression and NaiveBayes using the variables I identified earlier as having relationships with quality. Log regression had an accuracy of 90.3% compared to Naive Bayes with 86.4%.

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?

Higher quality wines of 7/8 are related to higher alcohol and lower volatile acidity while the opposite is true of lower quality wines of 3/4. Average wines of 5/6 are more widely spread in values but generally cluster in regions intermediate to the high and low quality wines.

Were there any interesting or surprising interactions between features?

The average wines of quality 5 and 6 have wider variation across all the variables plotted and are dispersed in all of the multivariate plots. I would have expected there to be a more distinct ‘intermediate’ cluster for the average wines, transitioning from the low quality wines to the high quality wines.

I think that the differences in taste due to variation in these features could be very subtle or that they may not always be picked up - therefore resulting in many average-rated wines with a wide variation in measured properties.

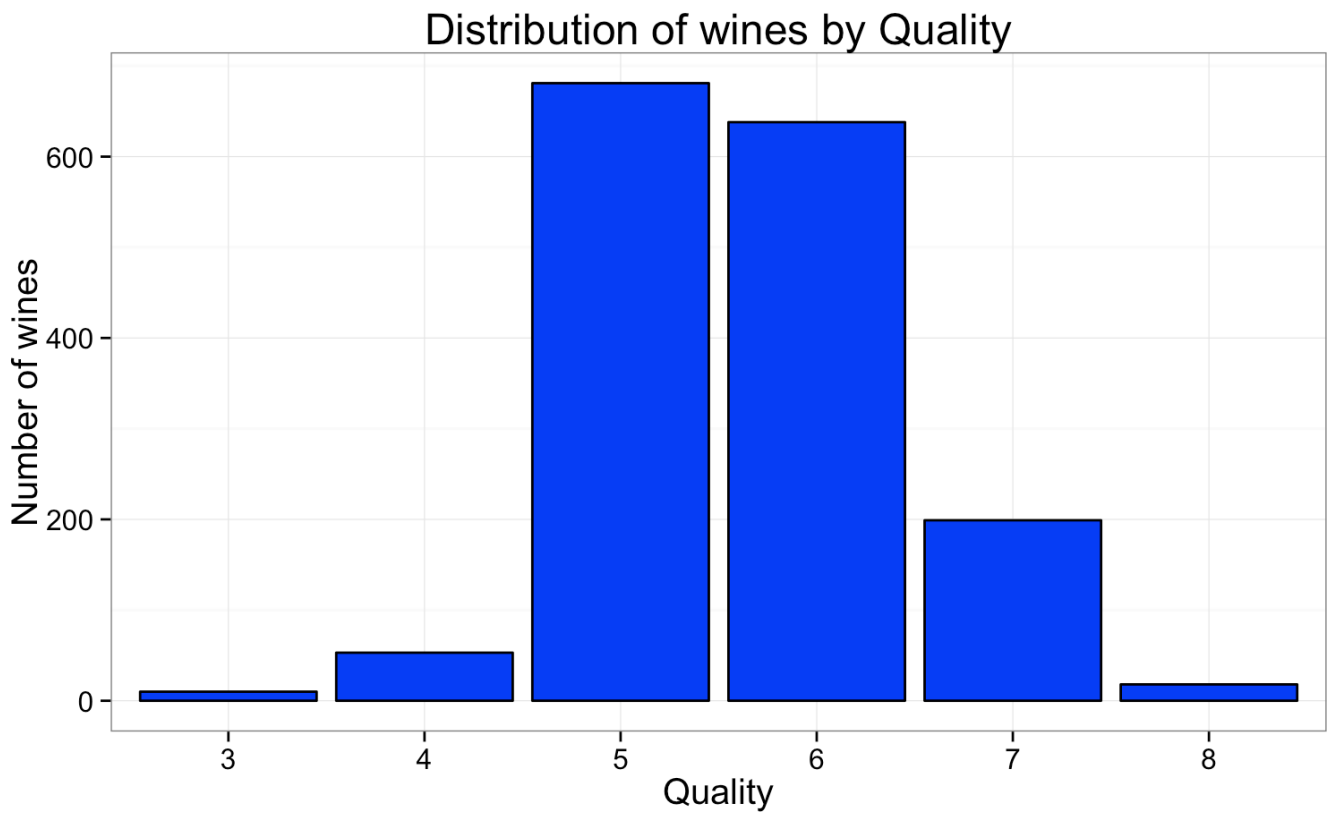
OPTIONAL: Did you create any models with your dataset? Discuss the strengths and limitations of your model.

I created 3 models using log regression, linear regression and naive Bayes.

- Log regression - It gave the highest accuracy but was only able to deal with 2 output categories, high and low. An arbitrary threshold of 0.5 was used to distinguish between high and low which may not be appropriate.
- Linear regression - It was able to model how quality ratings of 3-8 would vary with different variables. However, we know that the quality ratings were on a discrete scale, so linear regression which is used for predicting continuous variables may not be appropriate.
- Naive Bayes - It could deal with more than 2 categories for the outpput variable. However Naive Bayes assumes that the the different features are independent. With average wines spanning such a wide range of values, and there being such a large number of average wines, the model is likely to predict average wines correctly, leading to an inflated accuracy value.

Final Plots and Summary

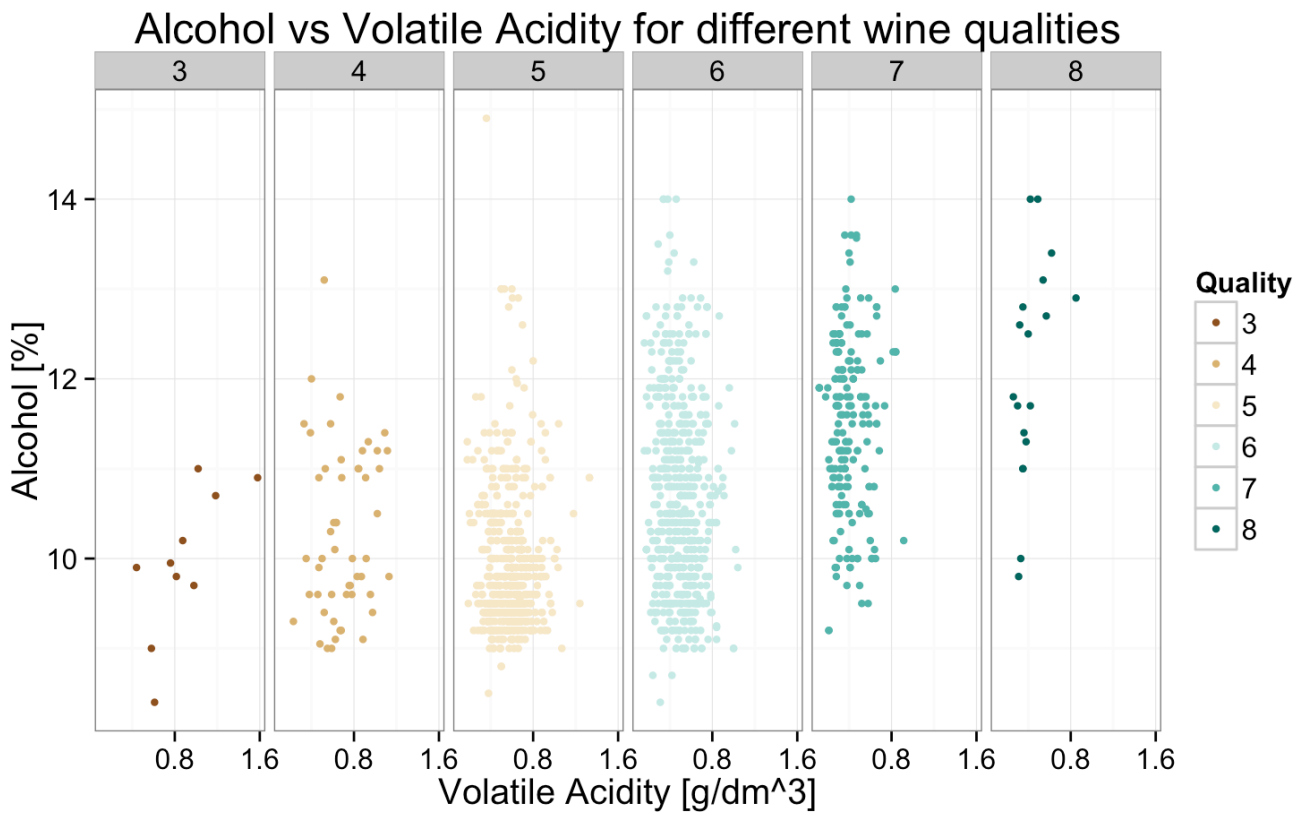
Plot One



Description One

Quality of wines is the output variable we're interested in so I chose to show its distribution. The distribution of quality in red wines is normal, with most wines being average at quality ratings of 5 or 6. There are few wines with low quality scores of 3/4 or high quality scores of 8.

Plot Two



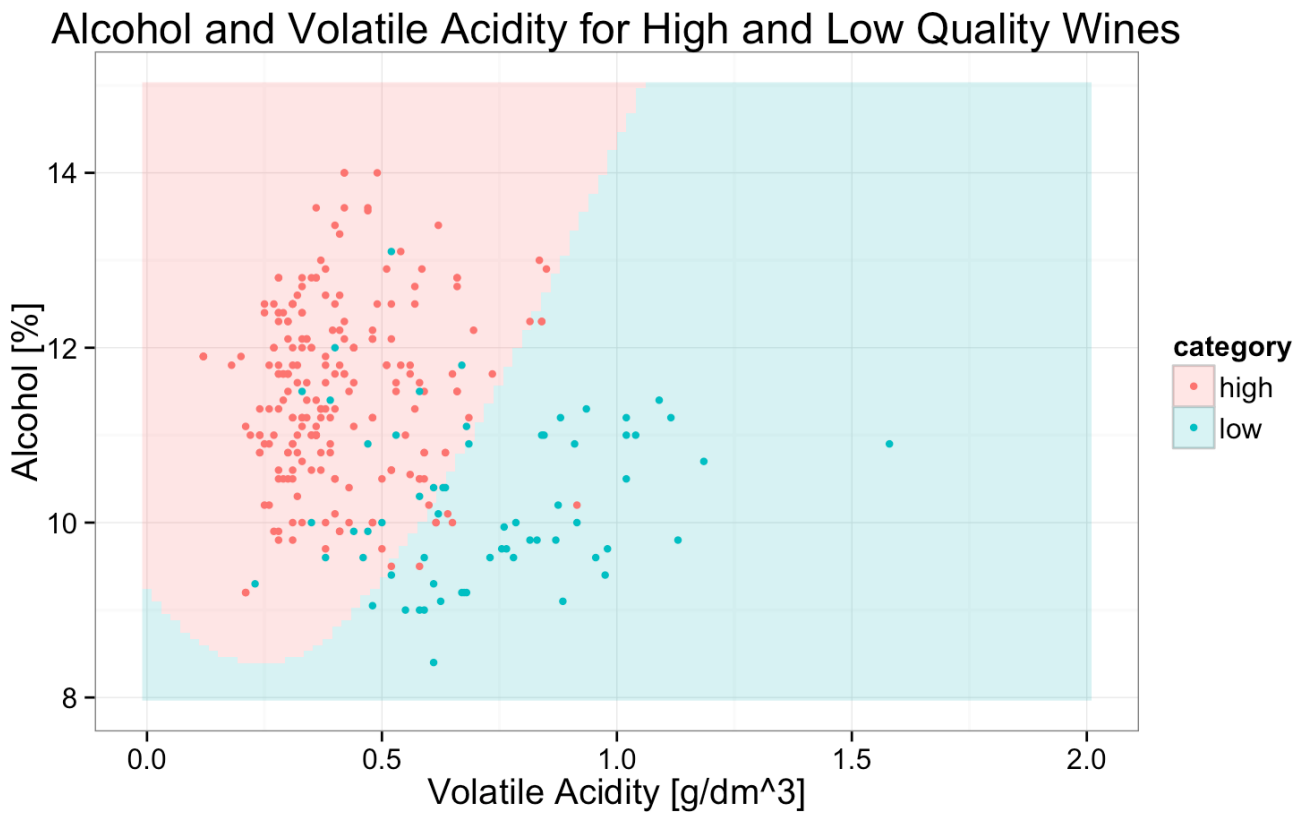
Description Two

As wine quality improves, volatile acidity tends to decrease while alcohol tends to increase. We see this in the plot, where the clusters appear to get thinner and move upwards as wine quality increases.

At lower qualities of 3/4/5, volatile acidity values are spread between 0 - 1.6g/dm^3. As quality improves to 6/7/8, these values appear to be increasingly constrained to 0.8g/dm^3 or less.

Similarly, the bulk of data points for lower quality wines are beneath the 12% alcohol mark. As quality improves to 6/7/8, the proportion of data points lying above the 12% mark increases.

Plot Three



Description Three

The plot illustrates the range of alcohol and volatile acidity levels that we expect for low quality (3/4) and high quality(7/8) wines. High quality wines tend to be concentrated toward the upper left of the plot while low quality wines tend to be concentrated towards the bottom right of the plot.

We could make a guess on whether a particular wine would be rated high or low quality if we knew its alcohol content and volatile acidity, and located that pair of values in the appropriate region.

Reflection

The wine data set contains information on 1599 wines. I started by understanding the individual variables in the data set, and then by investigating how characteristics differed for different wine qualities. I struggled with the data set because the bulk of the data points were average wines, so it was difficult elucidating trends across different qualities (and even though I did, I have to question whether the findings are representative)

Based on the final plots, I was able to show how alcohol and volatile acidity affect wine quality.

Other things that could be studied further include making more multivariate plots, for instance, looking at how other features of interest vary within each quality group, for different alcohol and volatile acidity values.

Other models could be used to cluster groups for Plot 3 to see if we can get a better separation. Different pairs of variables could also be used to see if the clusters can be better distinguished from one another, particularly if average wines are also to be included.

The main limitiation is that the output variable in this data set is a subjective rating by 3 testers Many external factors could affect these ratings, such as number of wines tasted per session, whether they ate anything etc. We don't know the vintage of the wines either, that would also have an impact on taste. It might be good to revisit this exploration with a more expansive data set that minimizes the personal preferences of the tasters (e.g. having a larger number of tasters per wine).

The wine quality depends on concentration of certain chemical species such as sulphates, acetic acid and citric acid. I would expect these concentrations to be quite sensitive to changes in the chemical equilibria in the wines. I expect the equilibrium to keep shifting as oxidation of the wine occurs with time. For the measured variables to be useful predictors of quality, would be greatly affected by the chemical equilibria in the wines, the wine tasting would have to take place shortly after the physicochemical properties are tested.