What Makes an Excellent Vinho Verde Red Wine?

By: Leslie Culliton

1. Introduction

The following is an analysis of data from a paper published by Cortez, et.al. in 2009[1]. The data consists of 11 attributes of Portugese Vinho Verde Red Wine with a quality score for each wine. A list of the attributes and their effects on a wine can be found in the notes at the end of the analysis.

The analysis explores how the attributes relate to quality score.

2. Summary

Below is a summary of the data.

```
##        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
```

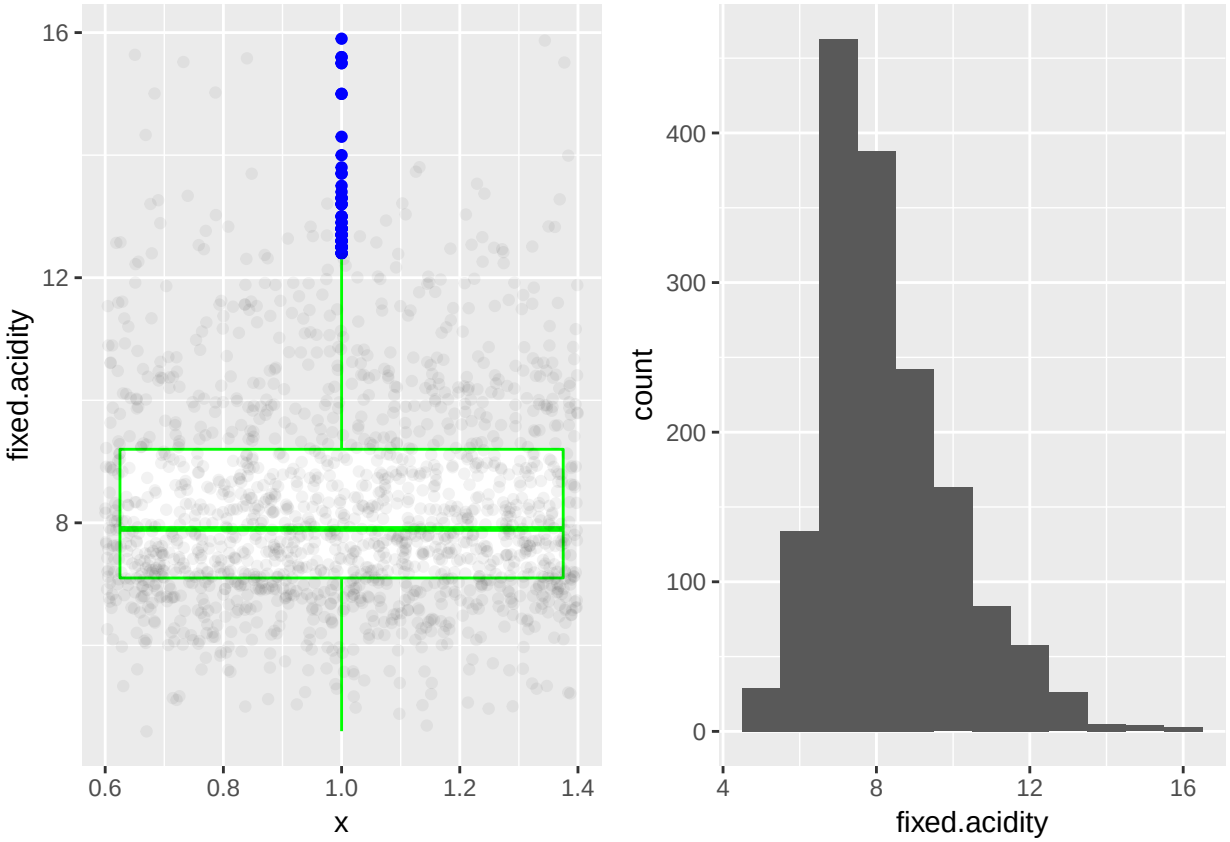
```
## 'data.frame':    1599 obs. of  13 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
##  $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
##  $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071 ...
```

```
##  $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
##  $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
##  $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
##  $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...
##  $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...
##  $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
##  $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...
```

The dataframe consists of 1599 rows and 13 columns. The original dataframe did not have a column name for the wine ID #. It is now added. See below.

```
##   ID fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1  1           7.4             0.70        0.00            1.9     0.076
## 2  2           7.8             0.88        0.00            2.6     0.098
## 3  3           7.8             0.76        0.04            2.3     0.092
## 4  4          11.2             0.28        0.56            1.9     0.075
## 5  5           7.4             0.70        0.00            1.9     0.076
## 6  6           7.4             0.66        0.00            1.8     0.075
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  11                  34  0.9978 3.51      0.56     9.4
## 2                  25                  67  0.9968 3.20      0.68     9.8
## 3                  15                  54  0.9970 3.26      0.65     9.8
## 4                  17                  60  0.9980 3.16      0.58     9.8
## 5                  11                  34  0.9978 3.51      0.56     9.4
## 6                  13                  40  0.9978 3.51      0.56     9.4
##   quality
## 1       5
## 2       5
## 3       5
## 4       6
## 5       5
## 6       5
```

3. Univariate Plots Section



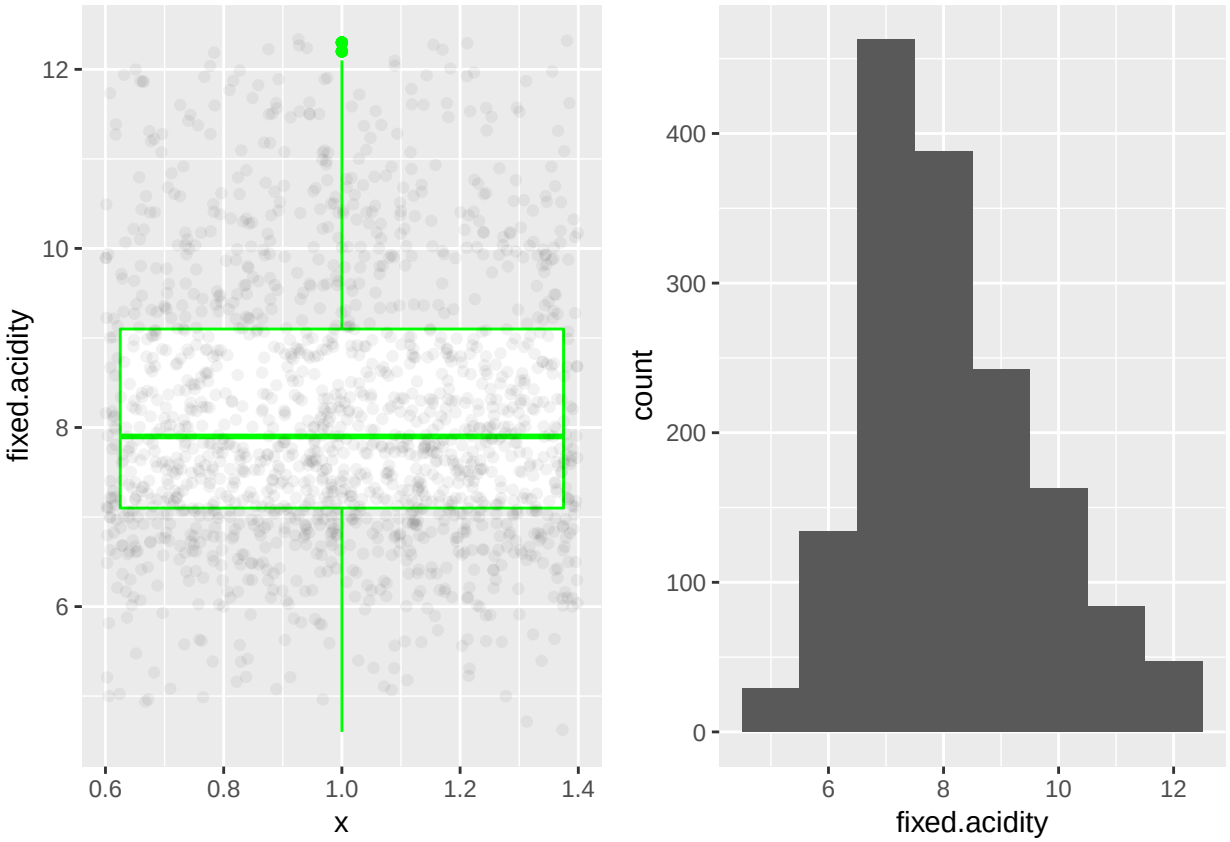
Plot 3.1

On the right is a histogram of fixed.acidity. This plot shows the non-volatile acidity of the wines. The pH scale is 1 - 14 with pH 7 being neutral, acids below 7, and bases above 7.

On the left is a boxplot of the fixed acidity. Points outside of 1.5 * the Interquartile Range (IQR) are considered outliers and are shown on the boxplot in blue. The IQR is the difference in the 3rd and 1st quantiles, which, for fixed acidity, are 9.2 and 7.1, respectively. These values can be found in the above Summary section. The IQR = 2.1. Multiplying IQR by 1.5 gives 3.15. Outliers are values that lie below 7.1 - 3.15 = 3.95 and above 9.2 + 3.15 = 12.35. These values will be removed.

Statistics for fixed acidity:

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

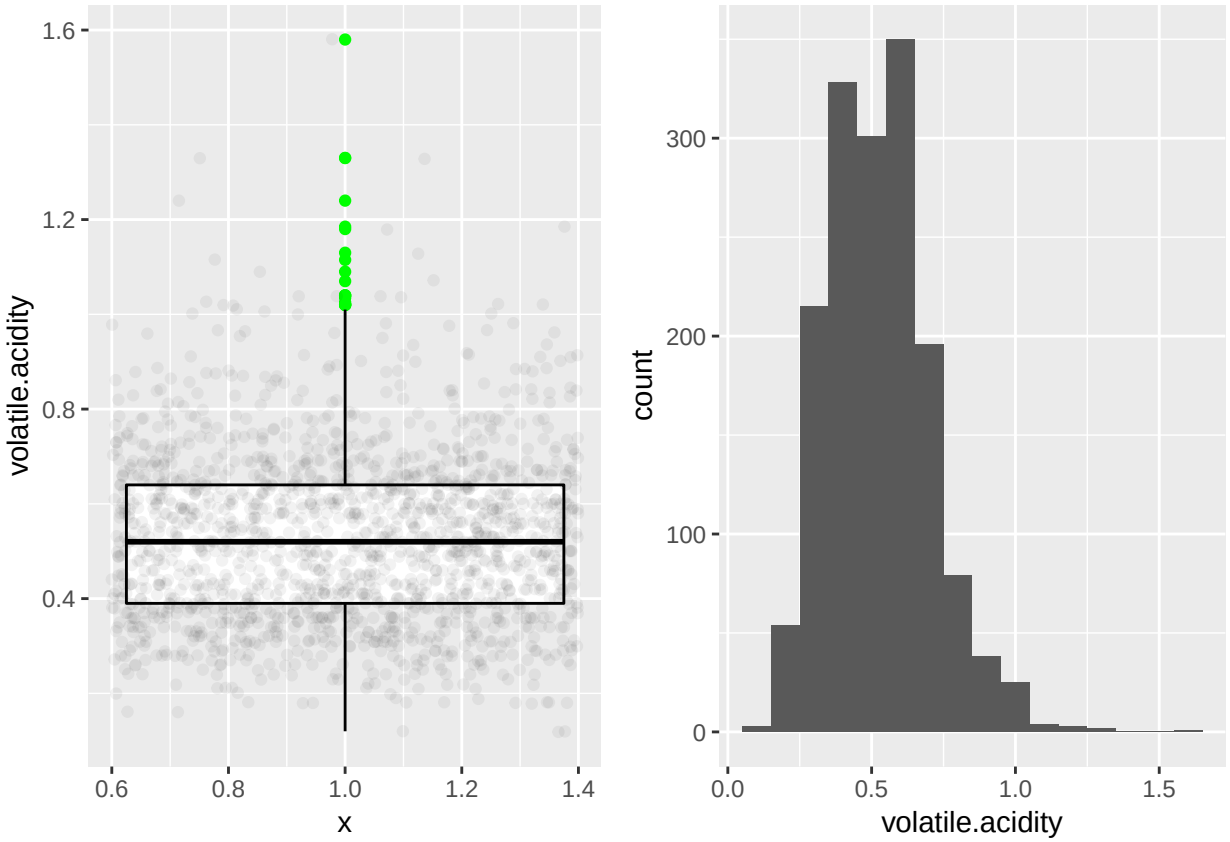


Plot 3.2

The plot on the left shows fixed acidity with outliers removed.

Statistics for fixed acidity after outlier removal:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   4.600   7.100   7.900   8.163   9.100  12.300      49
```



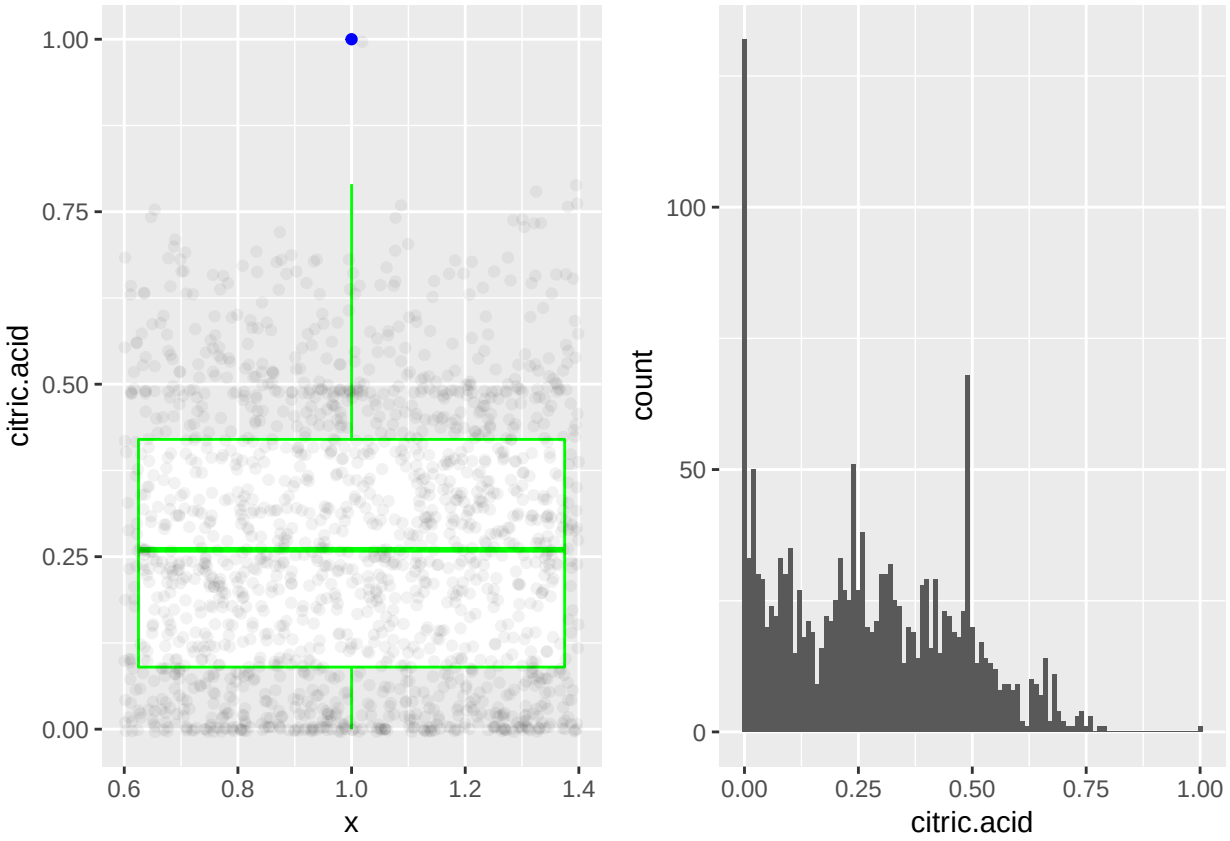
Plot 3.3

The above plot shows the volatile acidity of the wines, or the amount of acetic acid in the wine. If the amount is too high, the wine will have a vinegar taste.[1]

Outliers are shown in green. However, because higher levels of volatile acids are associated with a negative consumer experience, the outliers will not be removed as they affect the quality rating. Also, the range of volatile acidity concentration is very small, so the data points may not actually be outliers.

Summary of volatile acidity data:

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

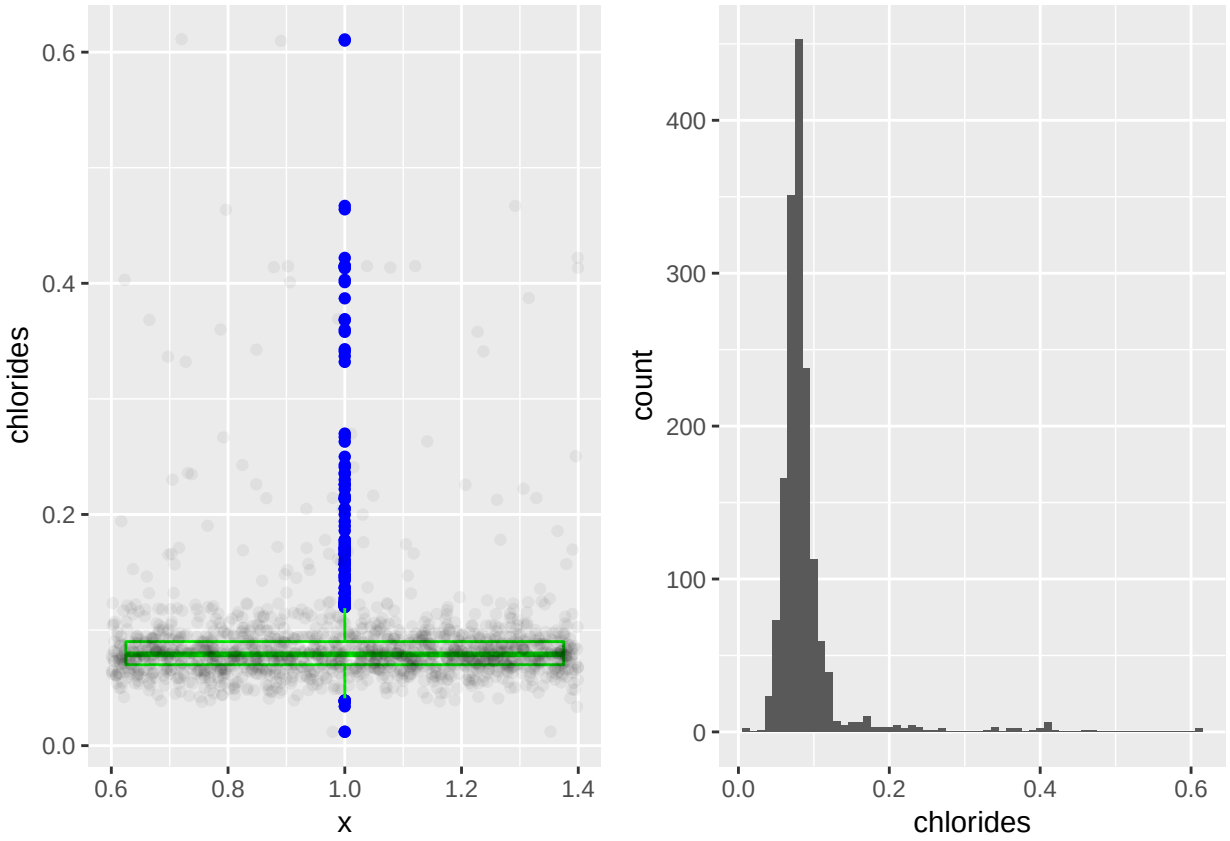


Plot 3.4

The amount of citric acid in the wines. Adds a freshness to the flavor.[1] Because the range of concentrations of citric acid are so small, the outliers will not be removed.

Summary of citric acid data:

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

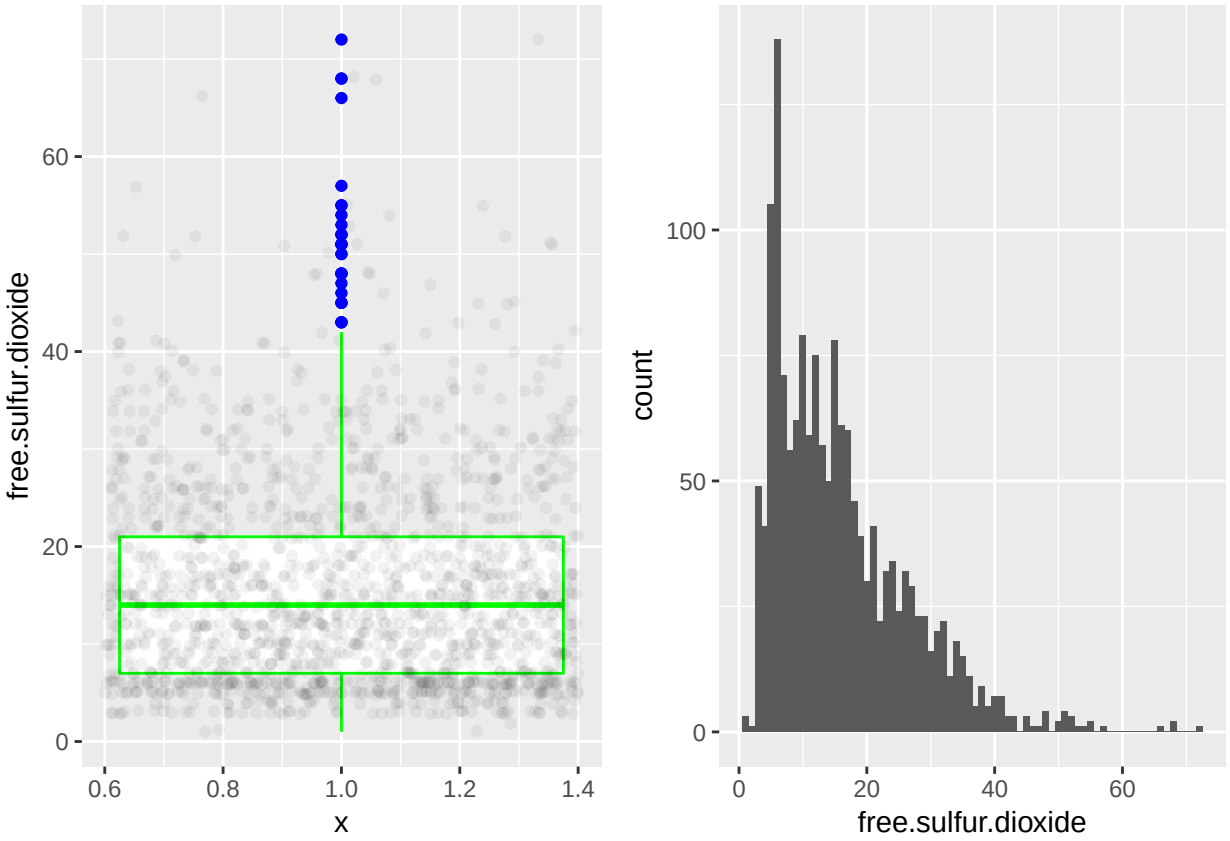


Plot 3.5

The amount of chlorides, or salt, in the wines. Because the concentration of chlorides is so small, the outliers will not be removed.

Summary choloride data:

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

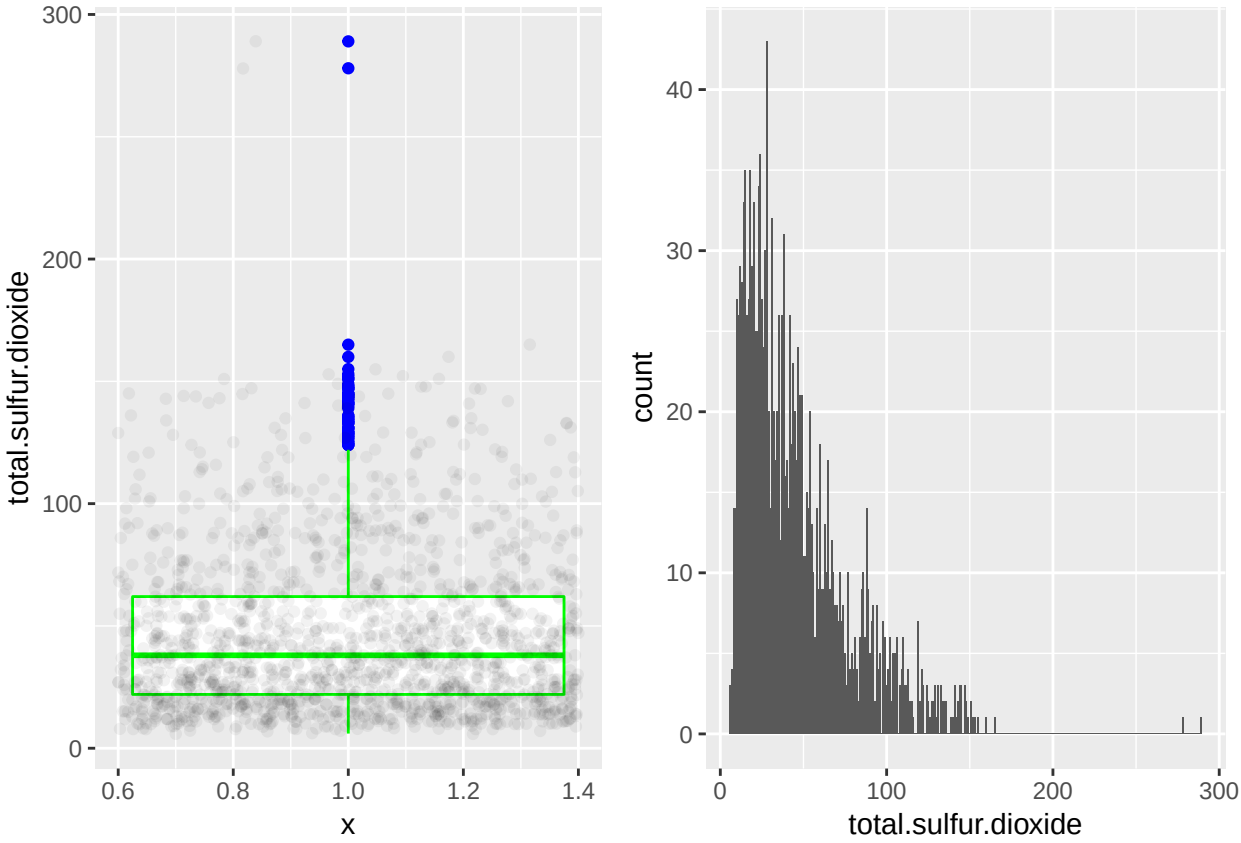


Plot 3.6

The amount of free sulfur dioxide in the wines. This prevents microbial growth and oxidation.[1] The plot on the left shows that there are outliers. However, because there is also data for total sulfur dioxide concentration, the outliers for free sulfur dioxide will not be removed.

Summary free sulfur dioxide:

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

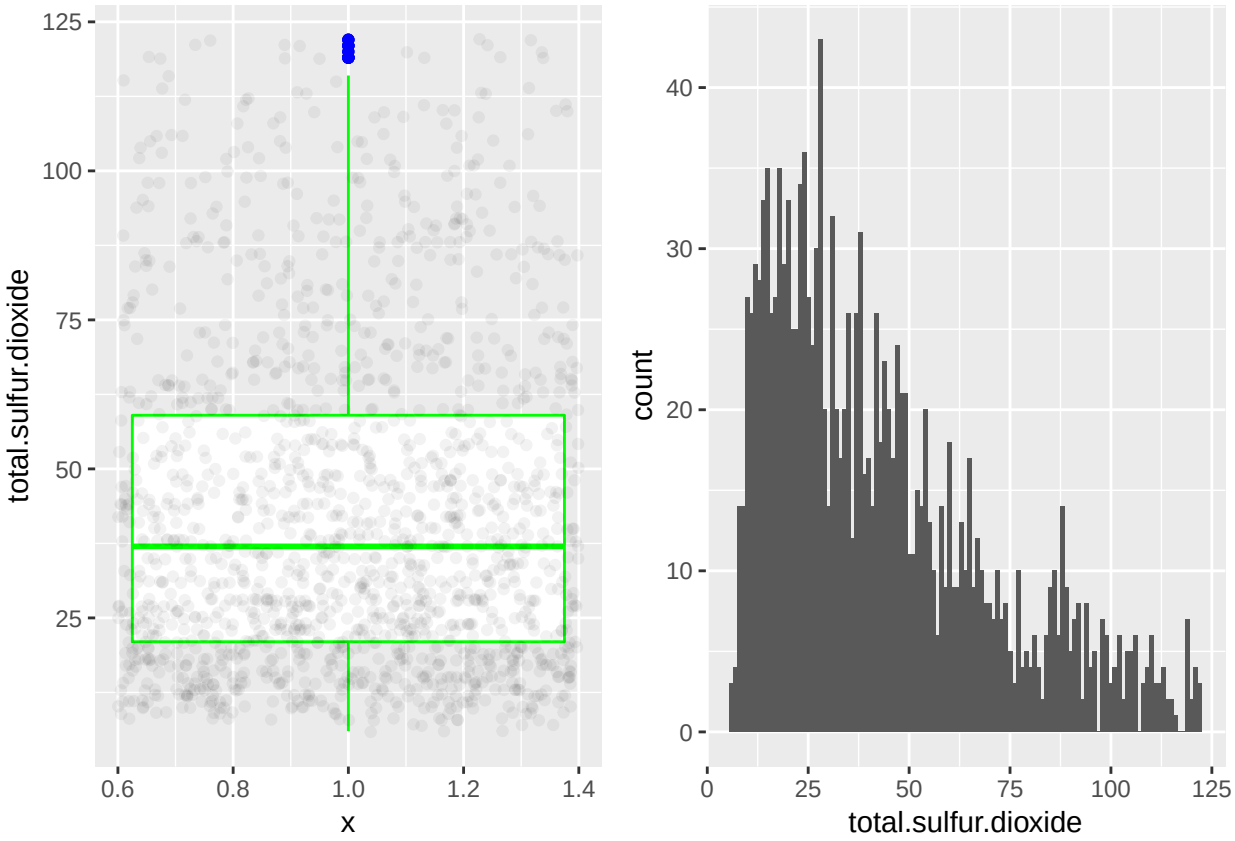


Plot 3.7

Total amount of sulfur dioxide in the wines. Outliers are blue.

Statistics for sulfur dioxde:

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

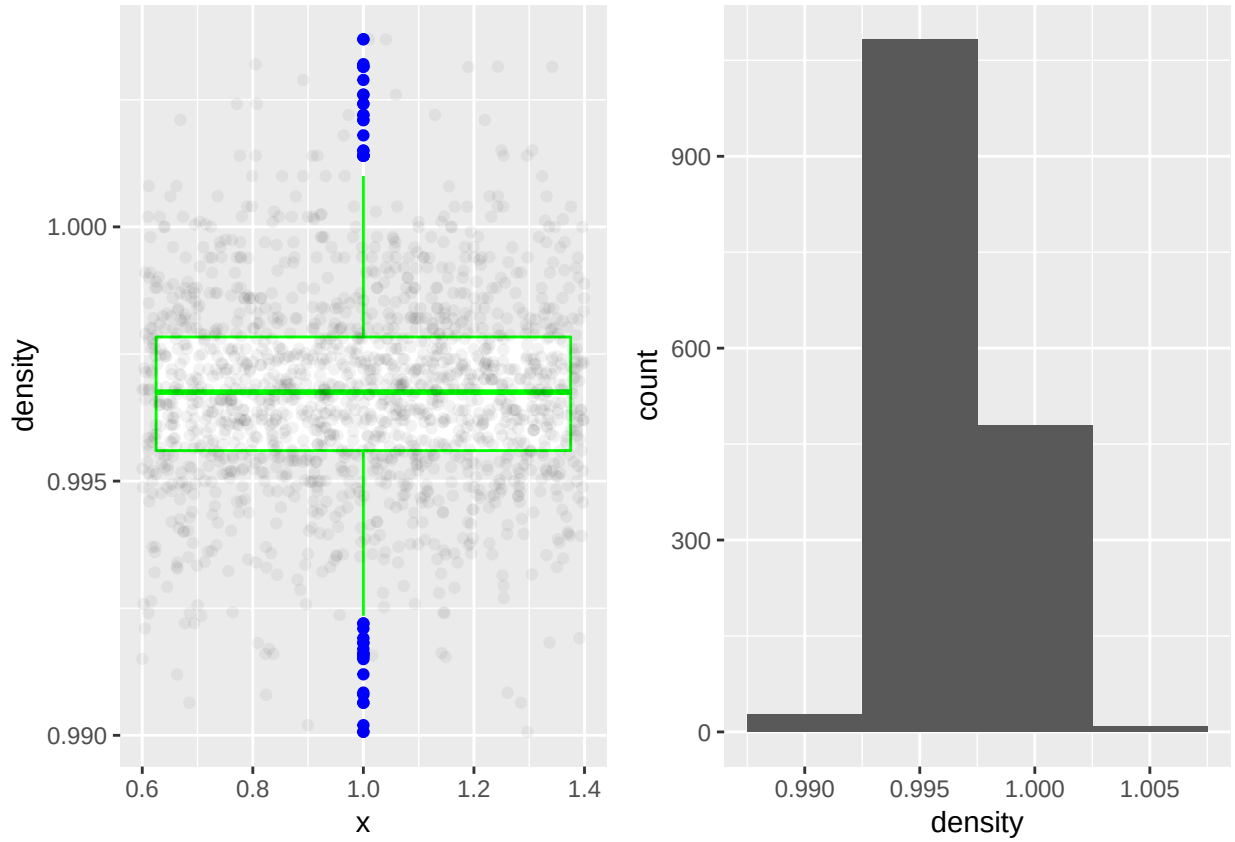


Plot 3.8

Total sulfur dioxide concentrations with outliers removed.

Summary of total sulfur dioxide data after removal of outliers:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       6      21      37      43      59     122      55
```

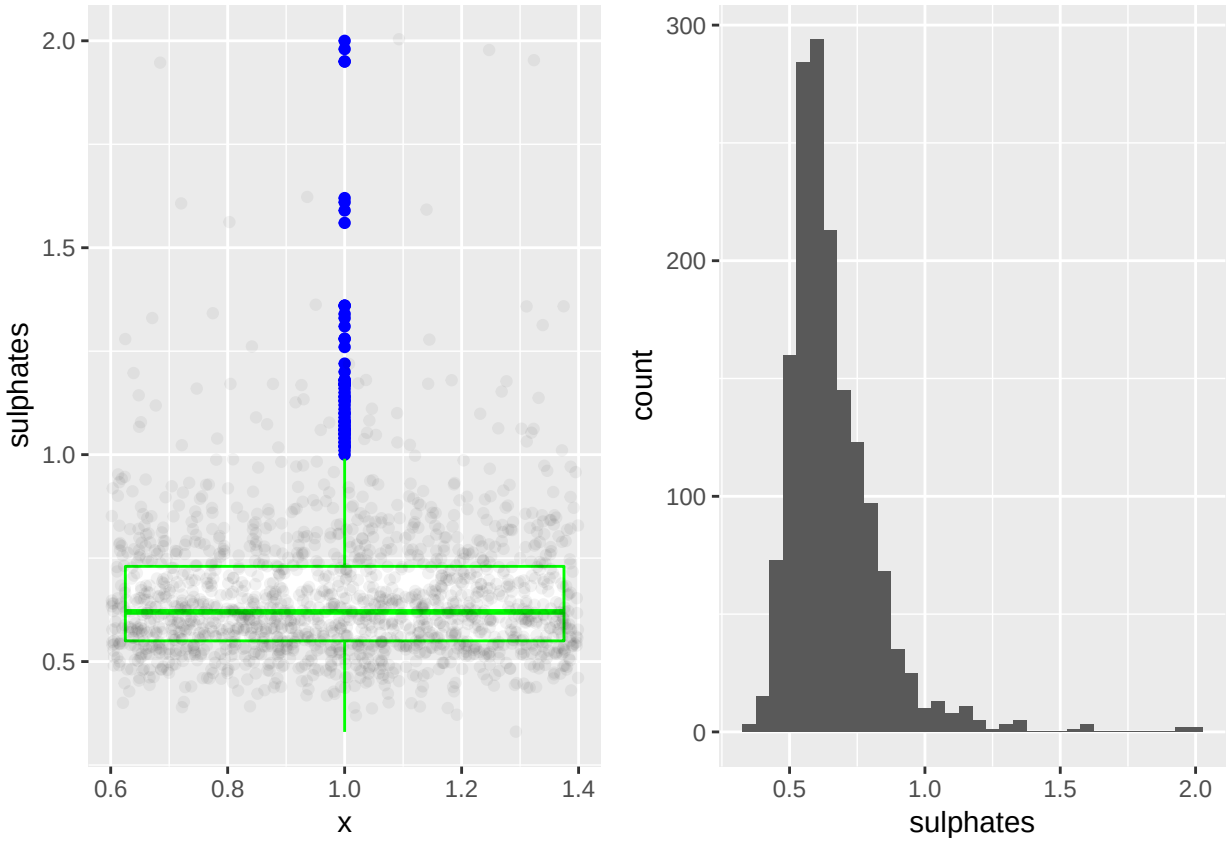


Plot 3.9

Density across all of the wines. Because the range of density values is so small, the outliers will not be removed.

Summary of density data:

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

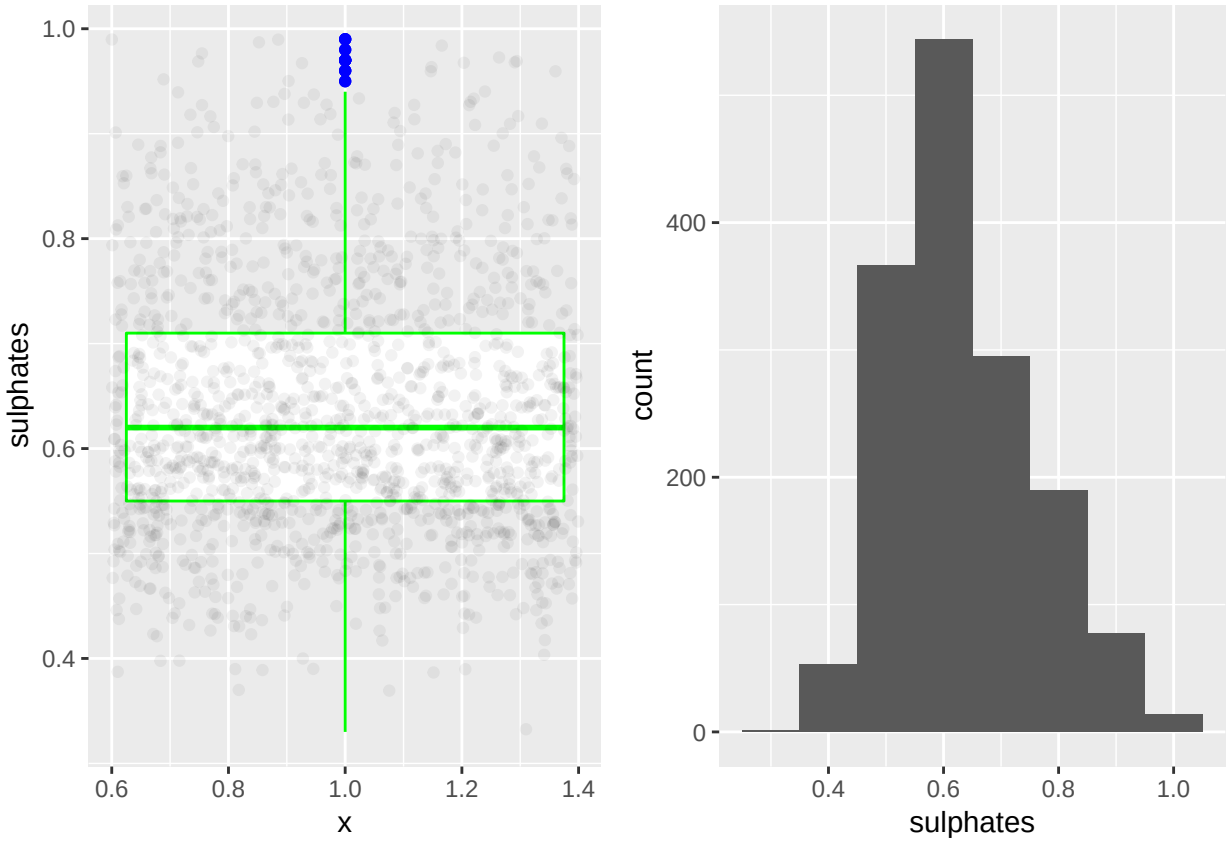


Plot 3.10

Amount of sulphates in the wines, which are antimicrobials and antioxidants.

Statistics for the sulphate data:

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

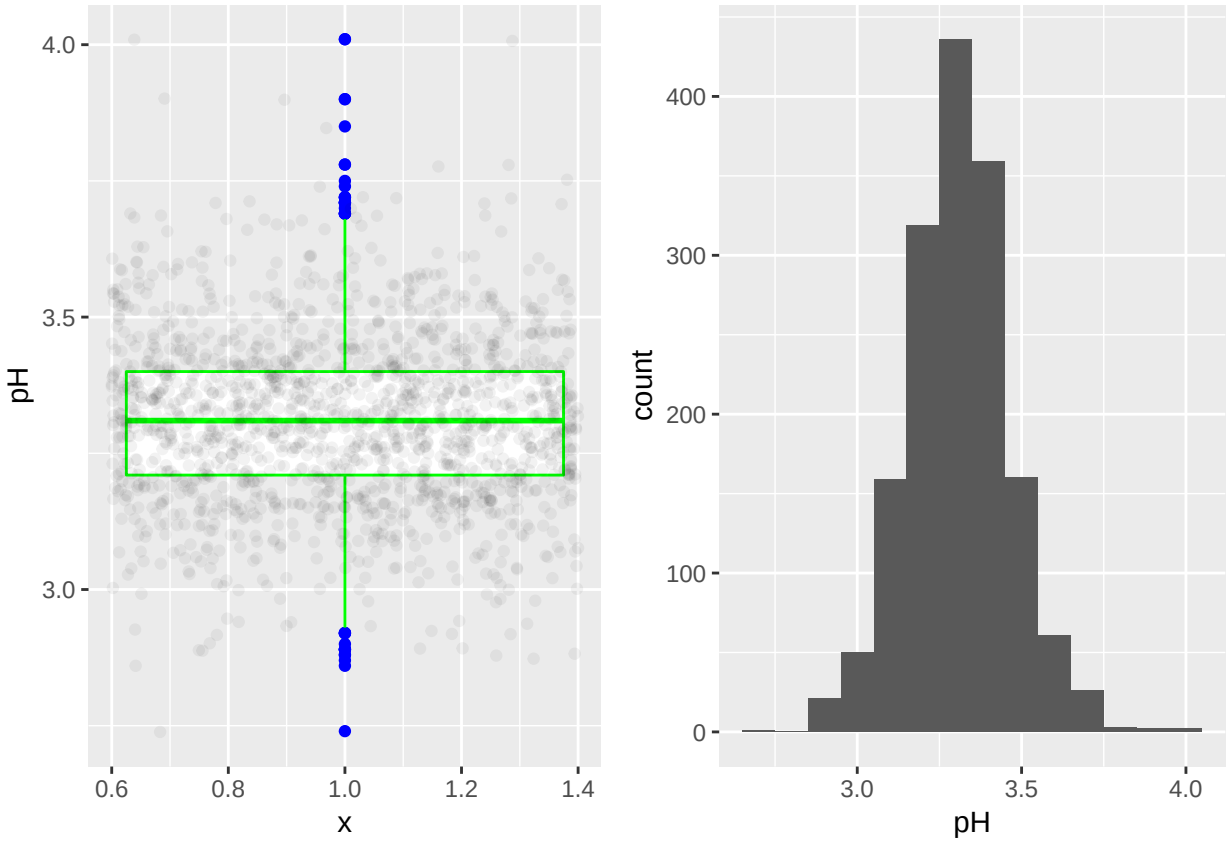


Plot 3.11

Sulphates concentration with outliers removed.

Summary of sulphates data after removal of outliers:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.3300  0.5500  0.6200  0.6364  0.7100  0.9900      59
```

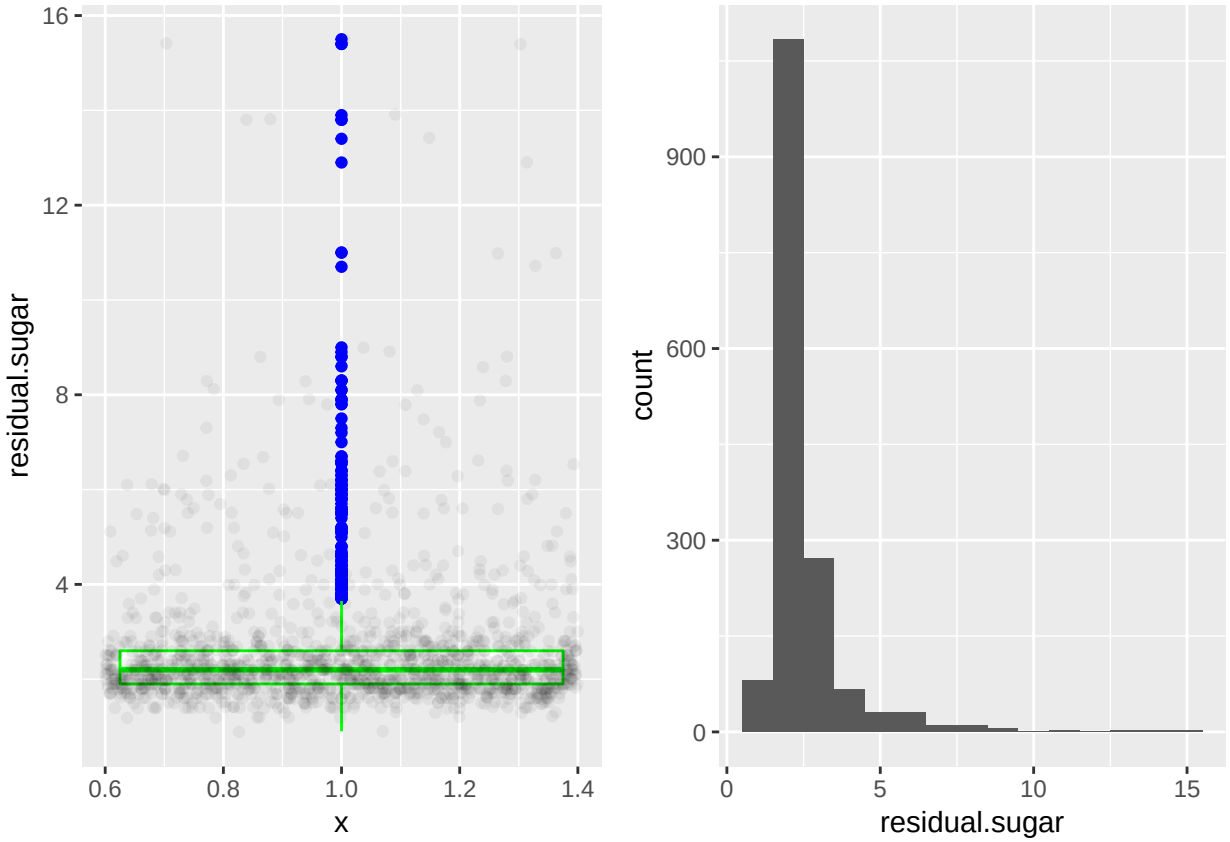


Plot 3.12

The pH of the wines. The outliers will not be removed because the range of values is small and the values in blue may not actually be outliers.

Summary of pH data:

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

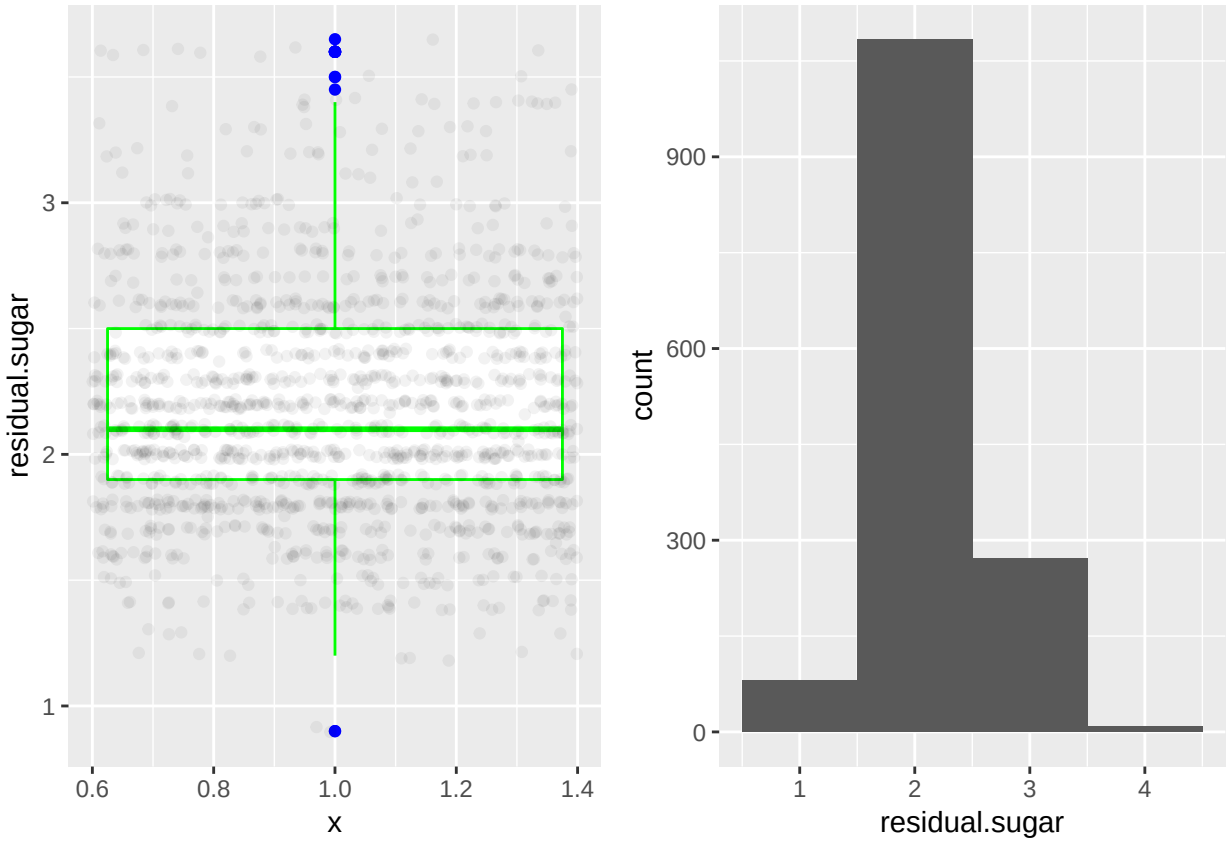


Plot 3.13

Amount of residual sugar in the wines. The outliers are shown in blue on the left.

Statistics for the residual sugar data:

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

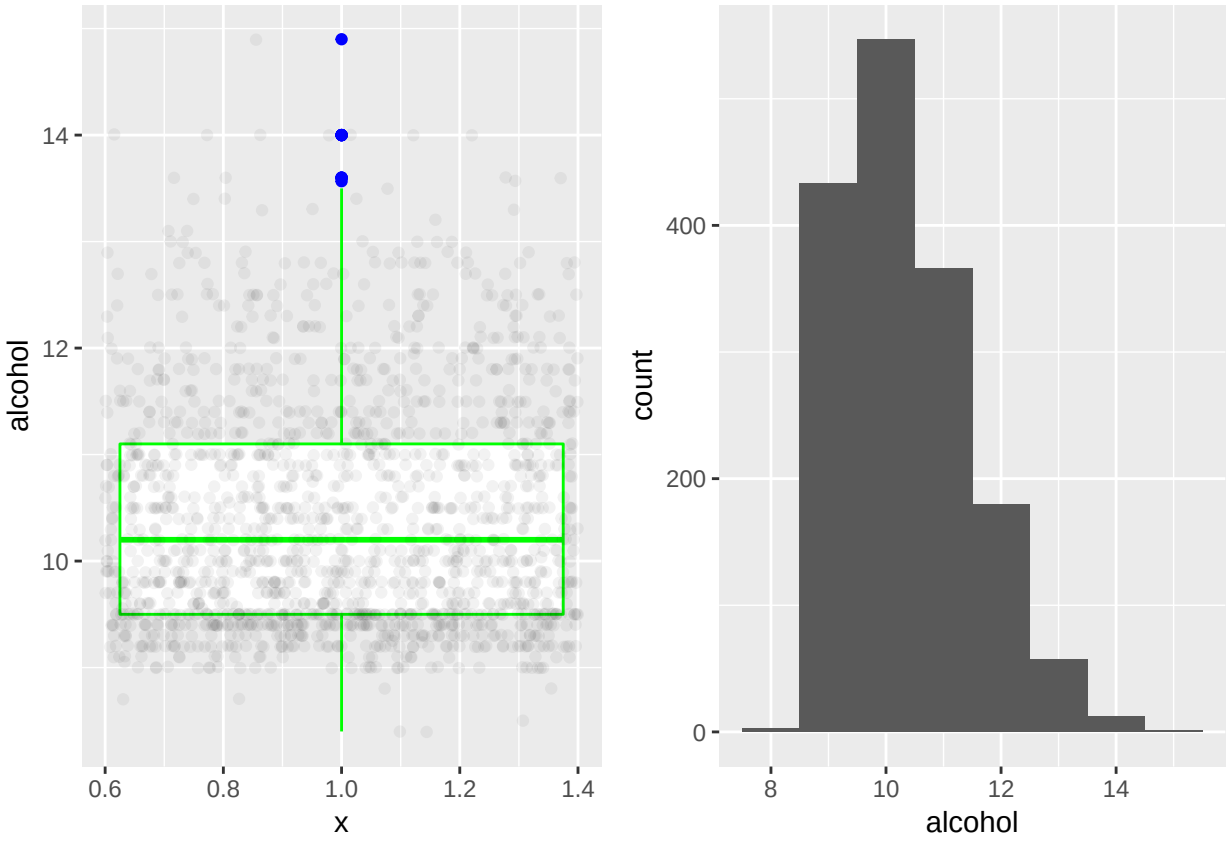


Plot 3.14

Residual sugar concentration with outliers removed.

Summary of residual sugar data after the removal of outliers:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.90    1.90    2.10    2.18    2.50    3.65     155
```

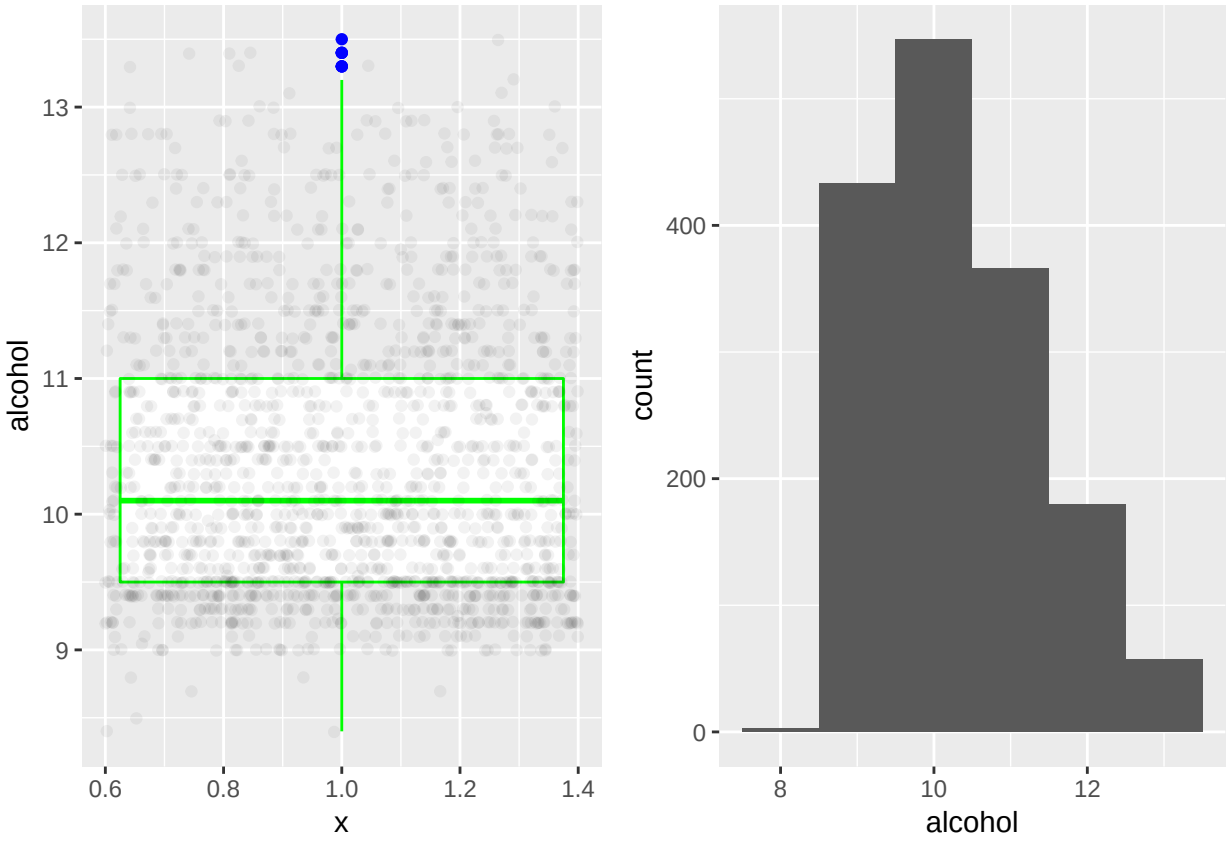


Plot 3.15

Amount of alcohol in the wines.

Statistics for the alcohol data:

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

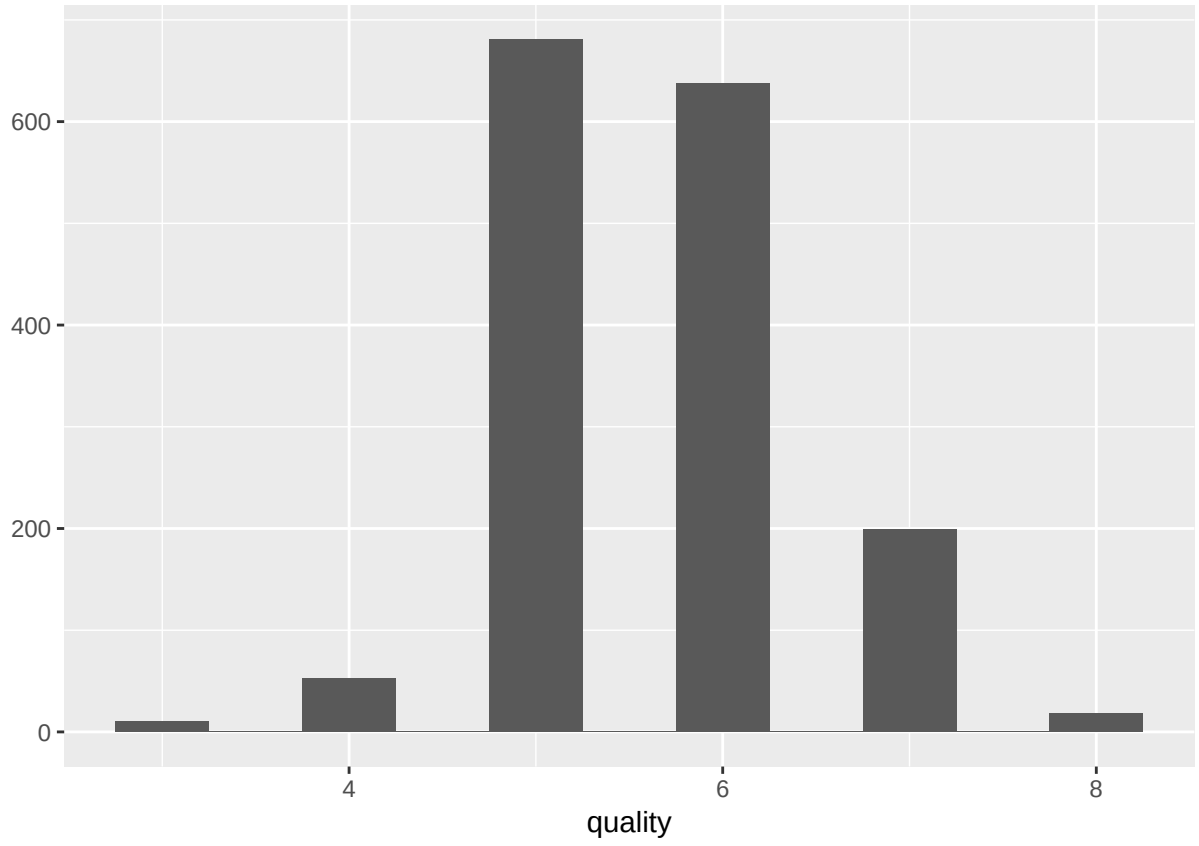


Plot 3.16

Plot of alcohol concentrations with outliers removed.

Summary alcohol data after outlier removal:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    8.40    9.50   10.10   10.39   11.00   13.50      13
```



Plot 3.17

Wine quality, on a scale of 0 - 10.

Summary of quality data:

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

4. Univariate Analysis

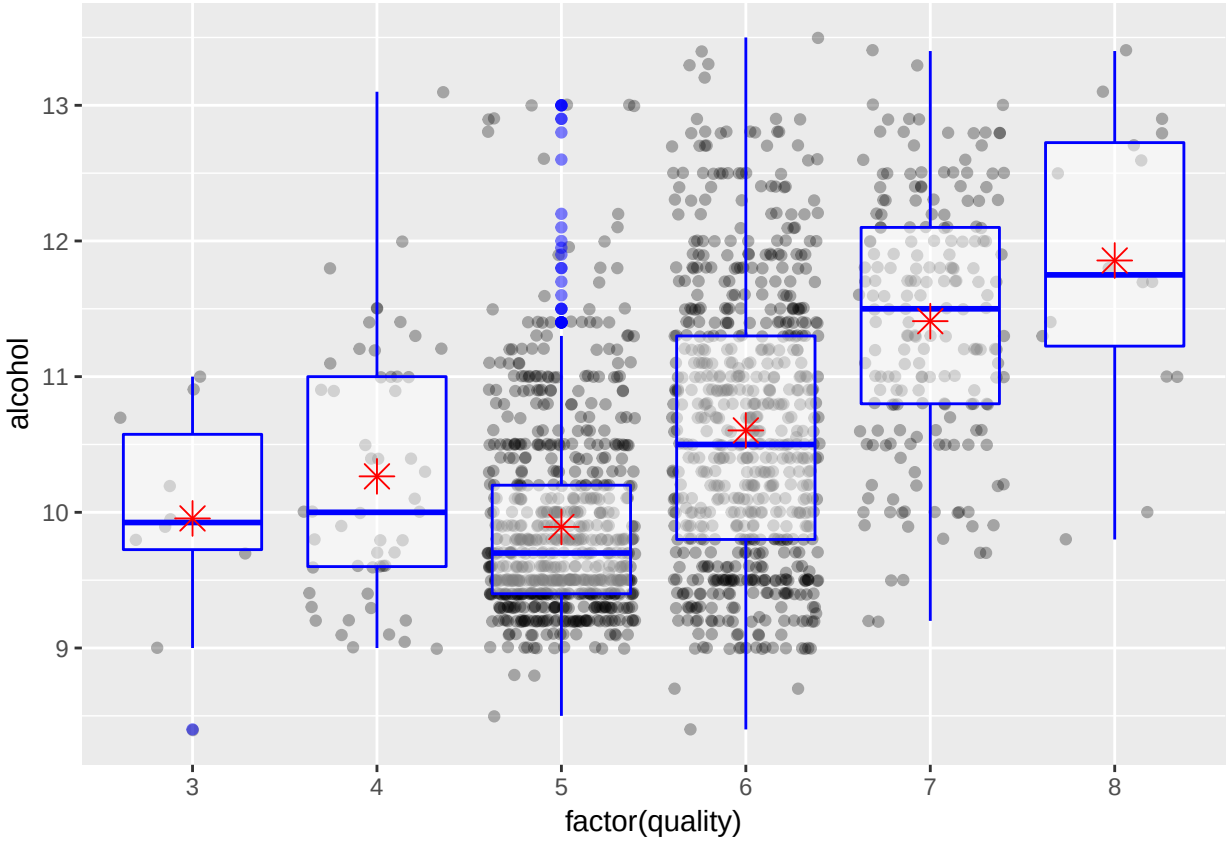
A quick look at the data shows that the dataset is 1599 rows with 13 columns. One column is the wine ID number and one column is the wine quality. The other 11 columns are attributes that contribute to the quality of wine.

The plots show the number of wines that contain an amount of each compound, except for the last plot, which shows the number of wines with each quality score. Most of the wines in the dataset have scores of 5, 6, or 7. The following analysis will explore what attributes contribute most to the quality of the wines.

Of interest are the variables fixed.acidity, residual.sugar, total.sulfur.dioxide, sulphates, and alcohol. The outliers for these attributes have been removed.

5. Bivariate Plots Section

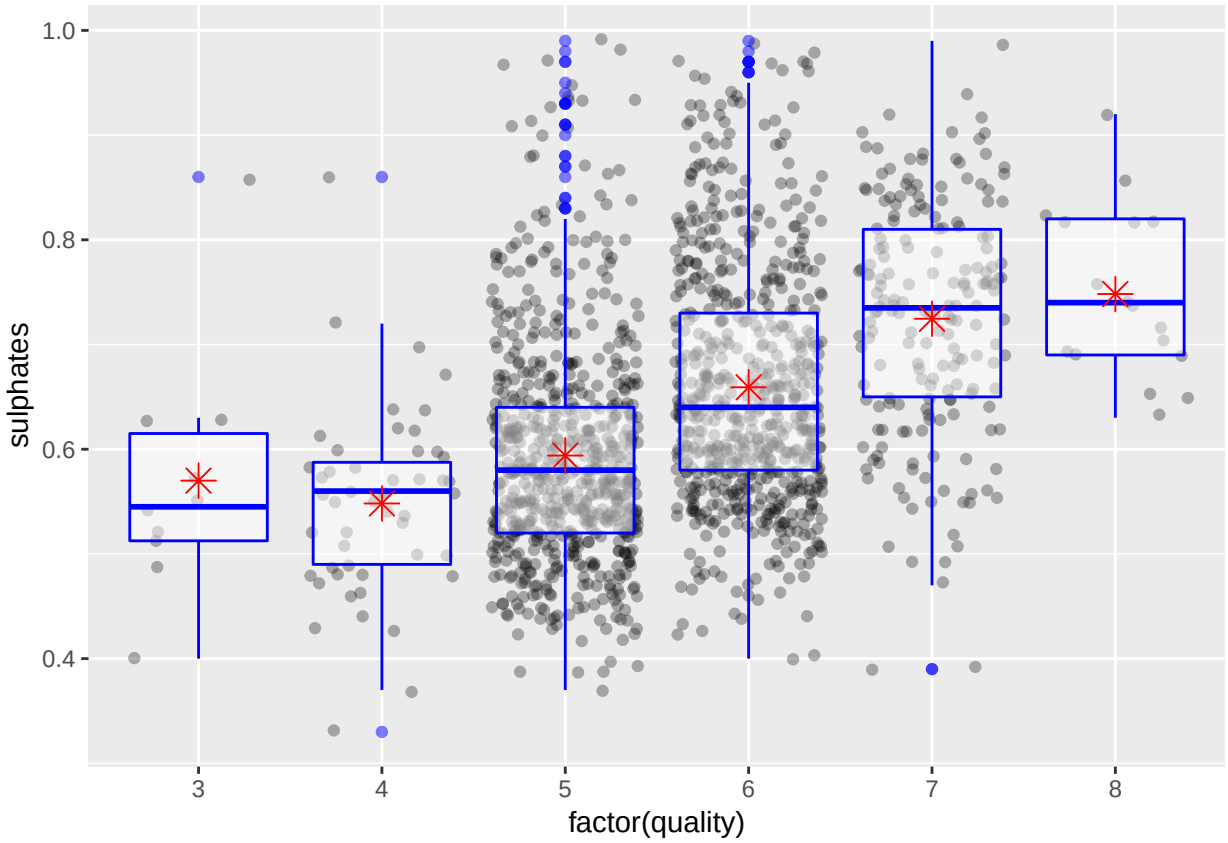
Plots in this section explore the relationships between two of the wine attributes listed in the paper by P. Cortez, et. al.[1]. The purpose was to gather enough information about the attributes to be able to begin to draw some conclusions about the role of each in wine preference.



Plot 5.1

This plot shows the relationshop between alcohol and quality. Generally, the quality increases as the concentration of alchol increases. There is a medium positive correlation between the two.

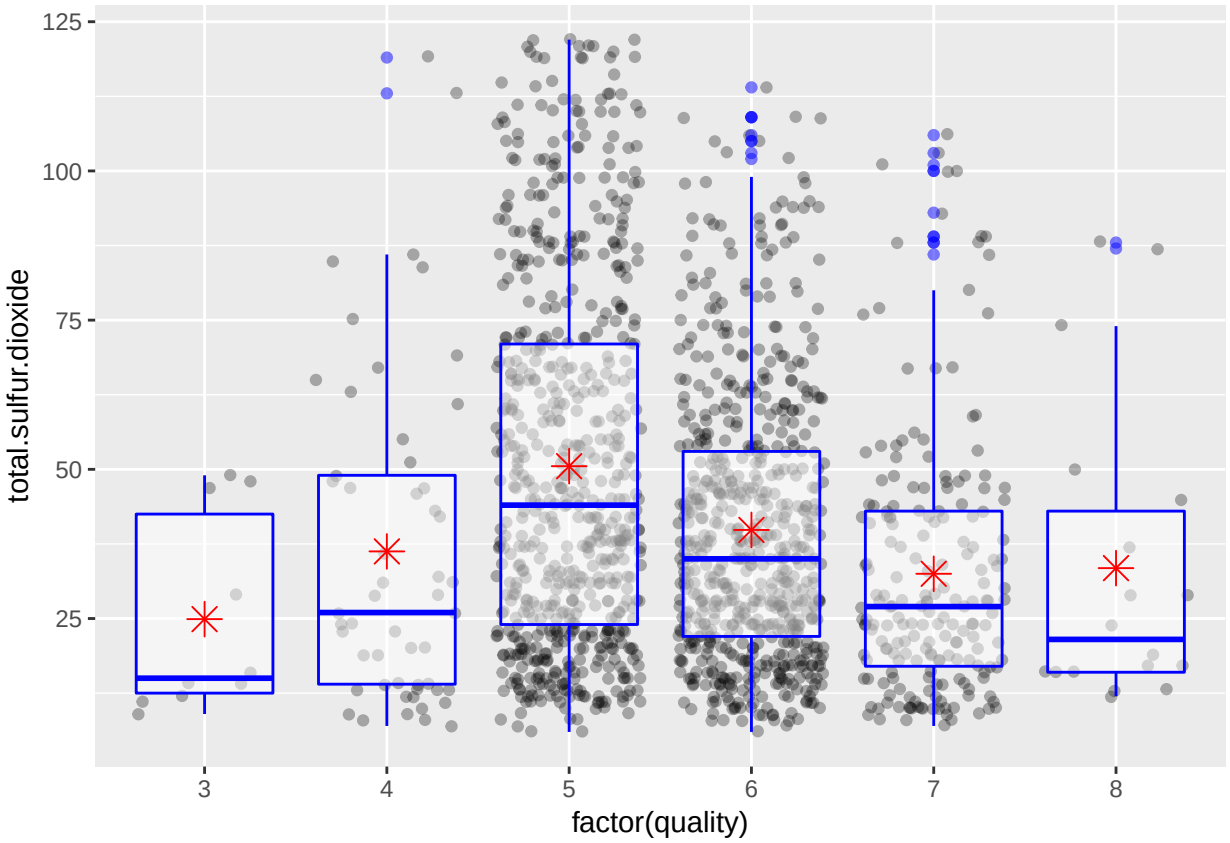
```
##
##  Pearson's product-moment correlation
##
## data:  df$quality and df$alcohol
## t = 21.252, df = 1584, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4318135 0.5084571
## sample estimates:
##       cor
## 0.4710238
```



Plot 5.2

The plot shows the relationship between qUality and sulphates. There is a medium positive correlation between the two.

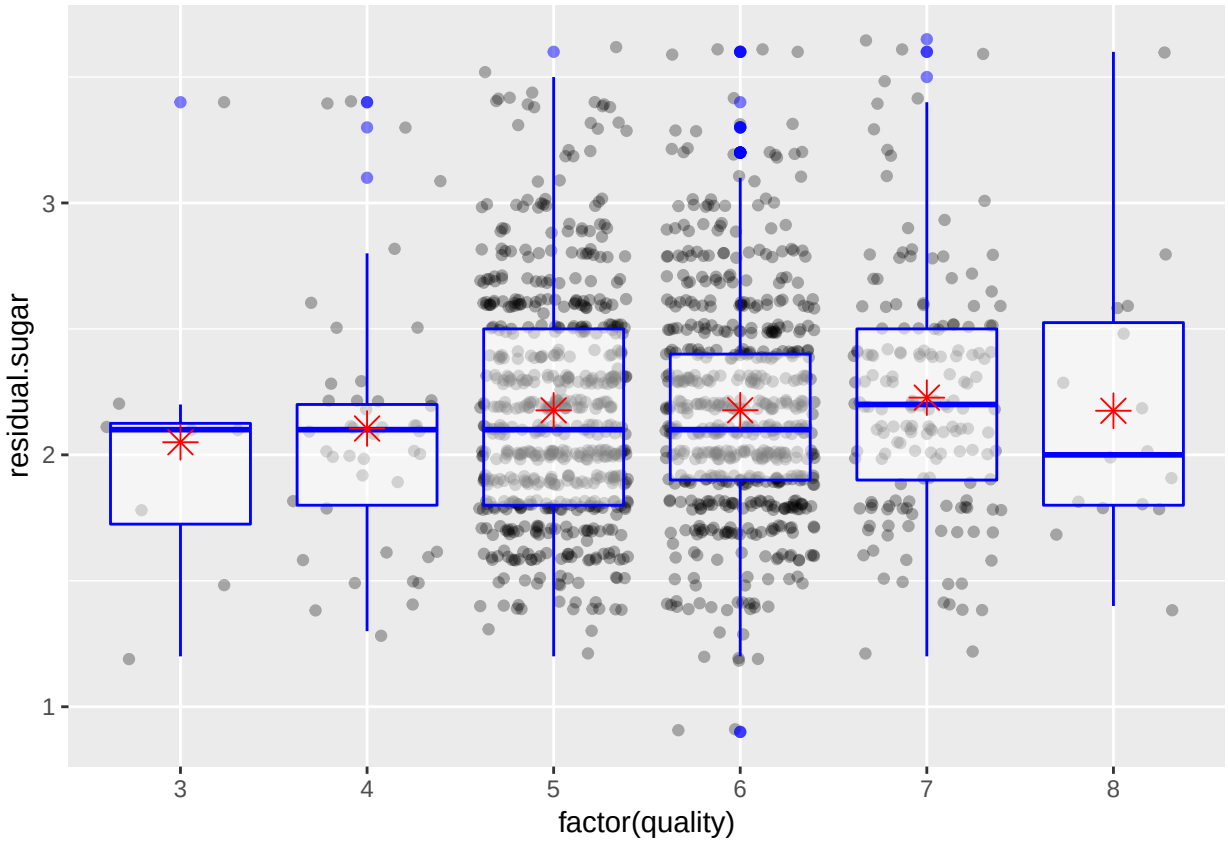
```
##
##  Pearson's product-moment correlation
##
## data:  df$quality and df$sulphates
## t = 16.815, df = 1538, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3510234 0.4354456
## sample estimates:
##       cor
## 0.3940654
```



Plot 5.3

This plot shows the relationship of total.sulfur.dioxide to quality. Total sulfur dioxide increases until quality score of 5 and then decreases. There is a weak negative correlation between total sulfur dioxide and quality.

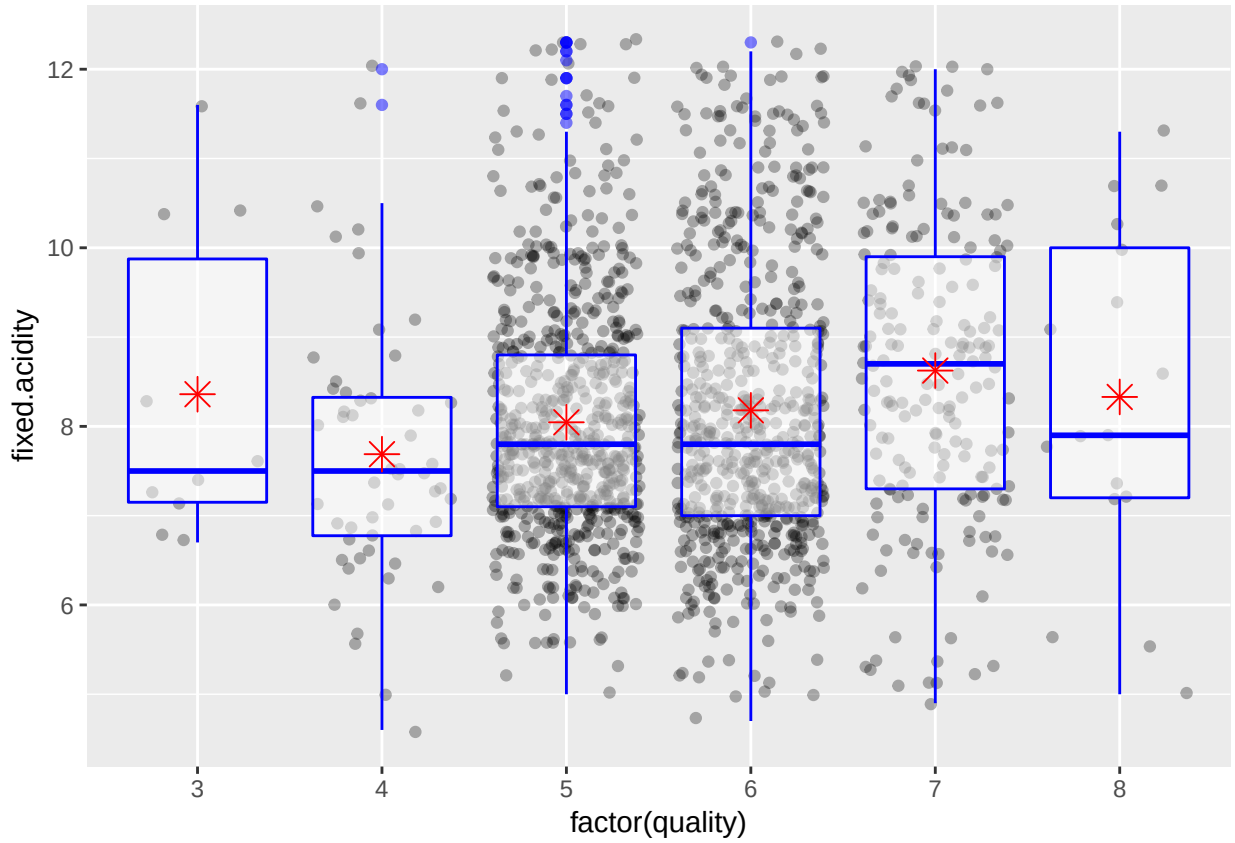
```
##
##  Pearson's product-moment correlation
##
## data:  df$quality and df$total.sulfur.dioxide
## t = -6.9275, df = 1542, p-value = 6.27e-12
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2216970 -0.1249275
## sample estimates:
##        cor
## -0.1737316
```



Plot 5.4

This plot shows the relationship between quality and residual sugar. The mean of residual sugar is between 2 and ~ 2.25 across all quality scores. There is a very weak positive correlation between residual sugar and qaulity.

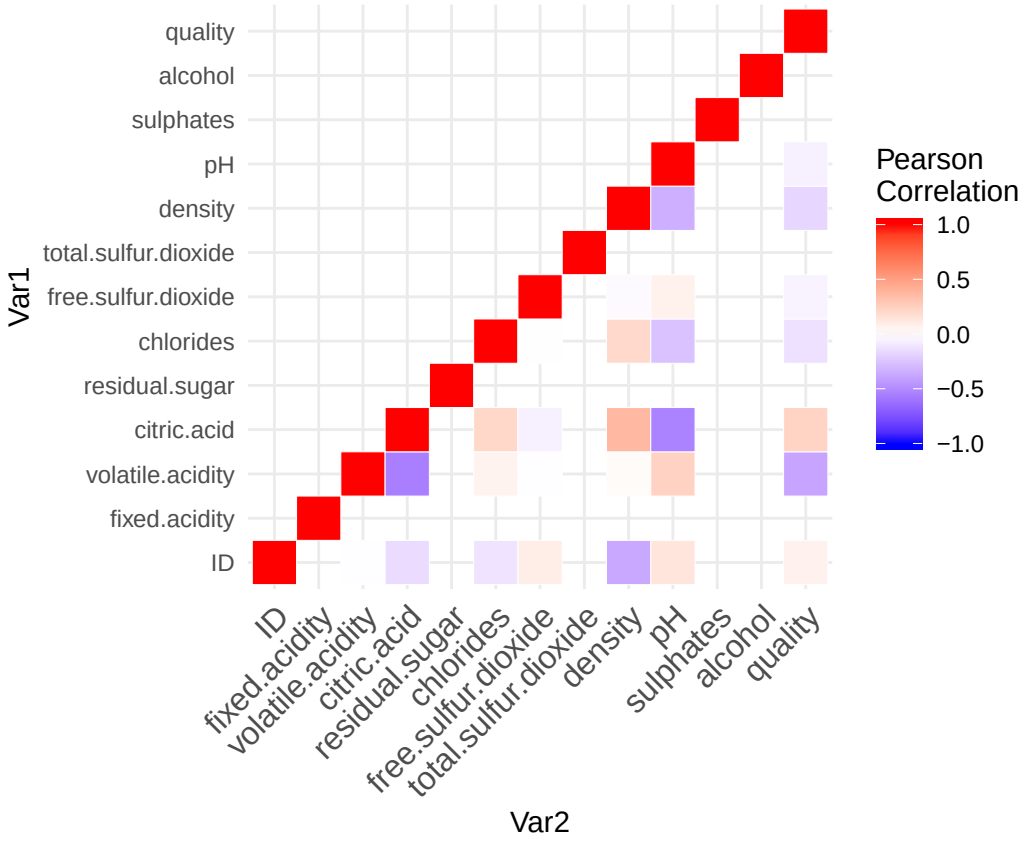
```
##
##  Pearson's product-moment correlation
##
## data:  df$quality and df$residual.sugar
## t = 1.448, df = 1442, p-value = 0.1478
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.01350746  0.08951479
## sample estimates:
##        cor
## 0.03810492
```



Plot 5.5

This plot shows the relationship between fixed.acidity and quality. There is a weak positive correlation between the two.

```
##
##  Pearson's product-moment correlation
##
## data:  df$quality and df$fixed.acidity
## t = 4.4912, df = 1548, p-value = 7.607e-06
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.06398476 0.16228738
## sample estimates:
##       cor
## 0.1134136
```



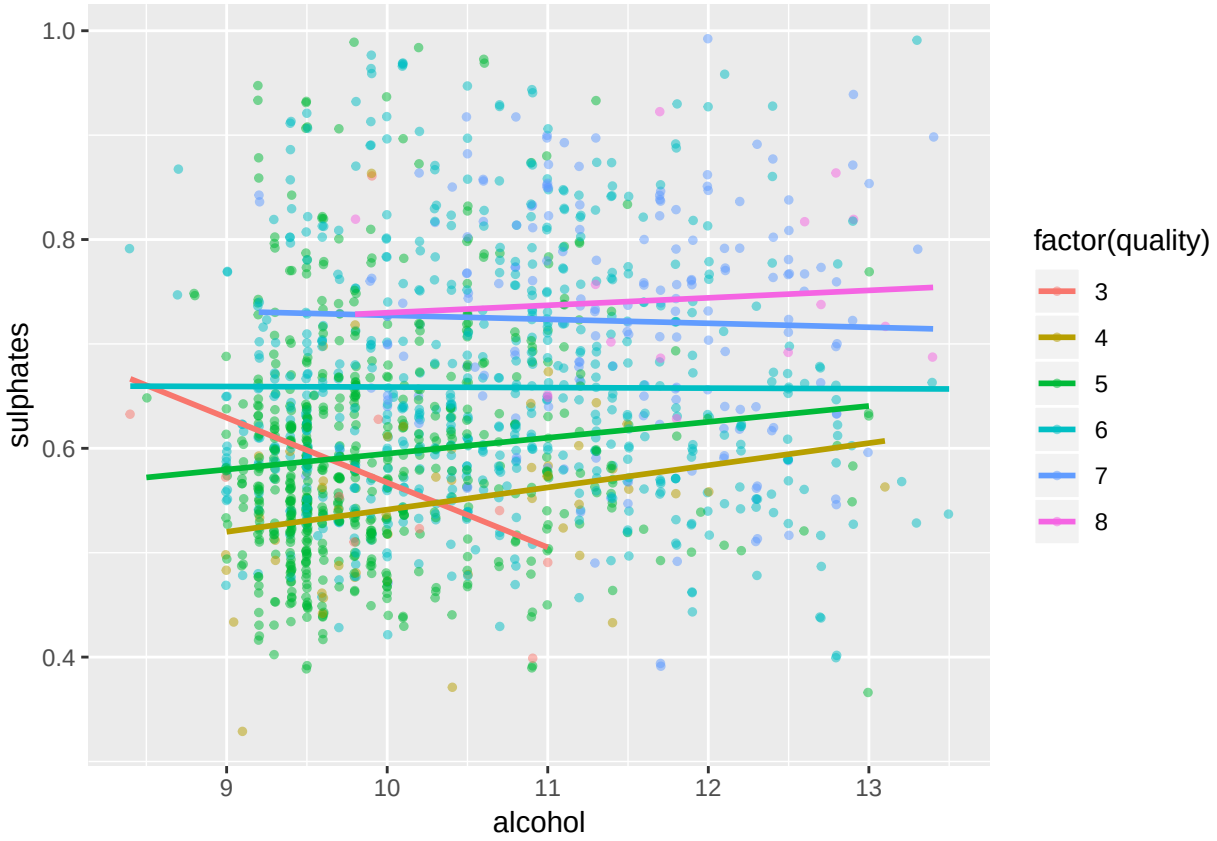
Plot 5.6

This plot shows the correlations between all of the attributes in the data. The red correlations are positive and the blue ones are negative.

6. Bivariate Analysis

In order to answer the question of what makes a quality vinho verde red wine, fixed.acidity, residual.sugar, total.sulfur.dioxide, sulphates, and alcohol were each plotted against quality. The plots were all the same style in order to more easily compare the relationship of each attribute with quality to each other. Lastly, a plot showing the correlations of all of the attributes to each other shows that there are no strong correlations across the attributes (and quality) in this data set.

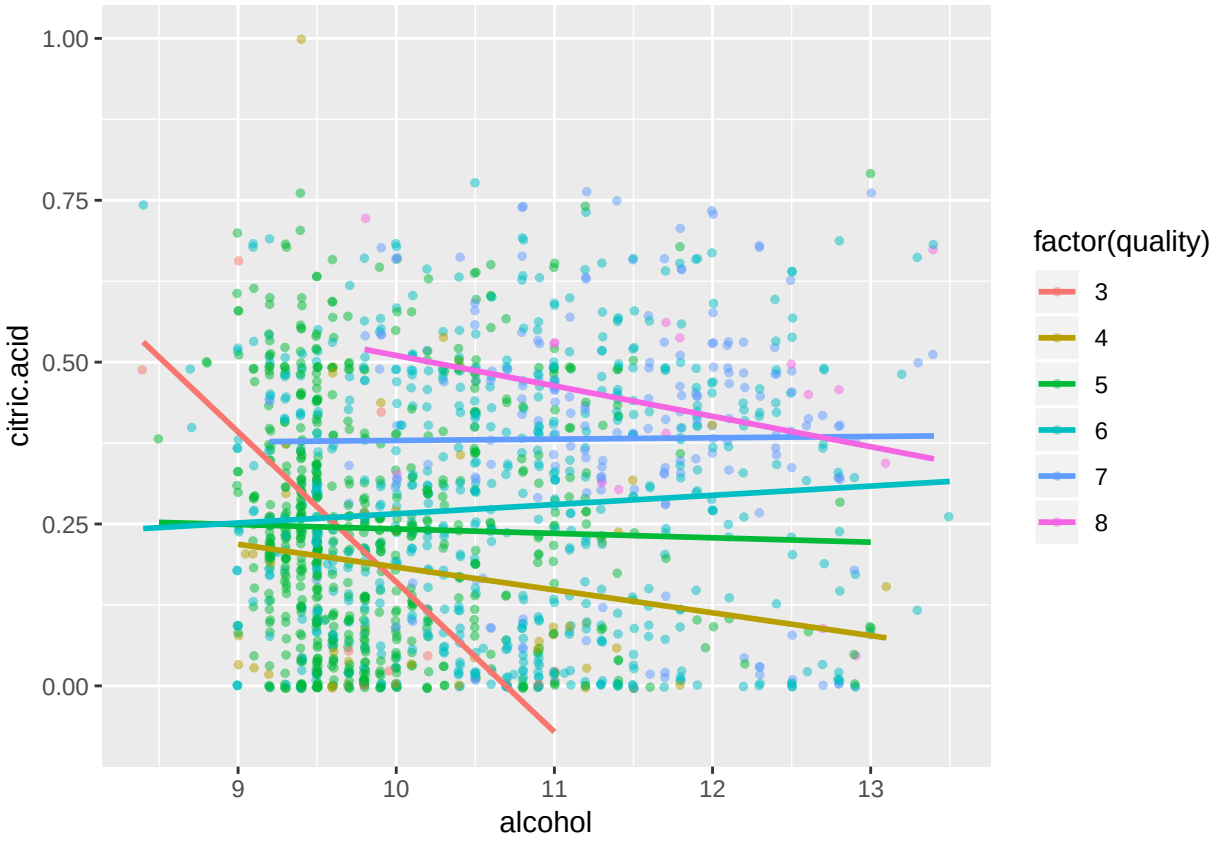
7. Multivariate Plots Section




```
##
##  Pearson's product-moment correlation
##
## data:  df$alcohol and df$sulphates
## t = 8.6861, df = 1525, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1687974 0.2644067
## sample estimates:
##       cor
## 0.2171227
```

Plot 7.1

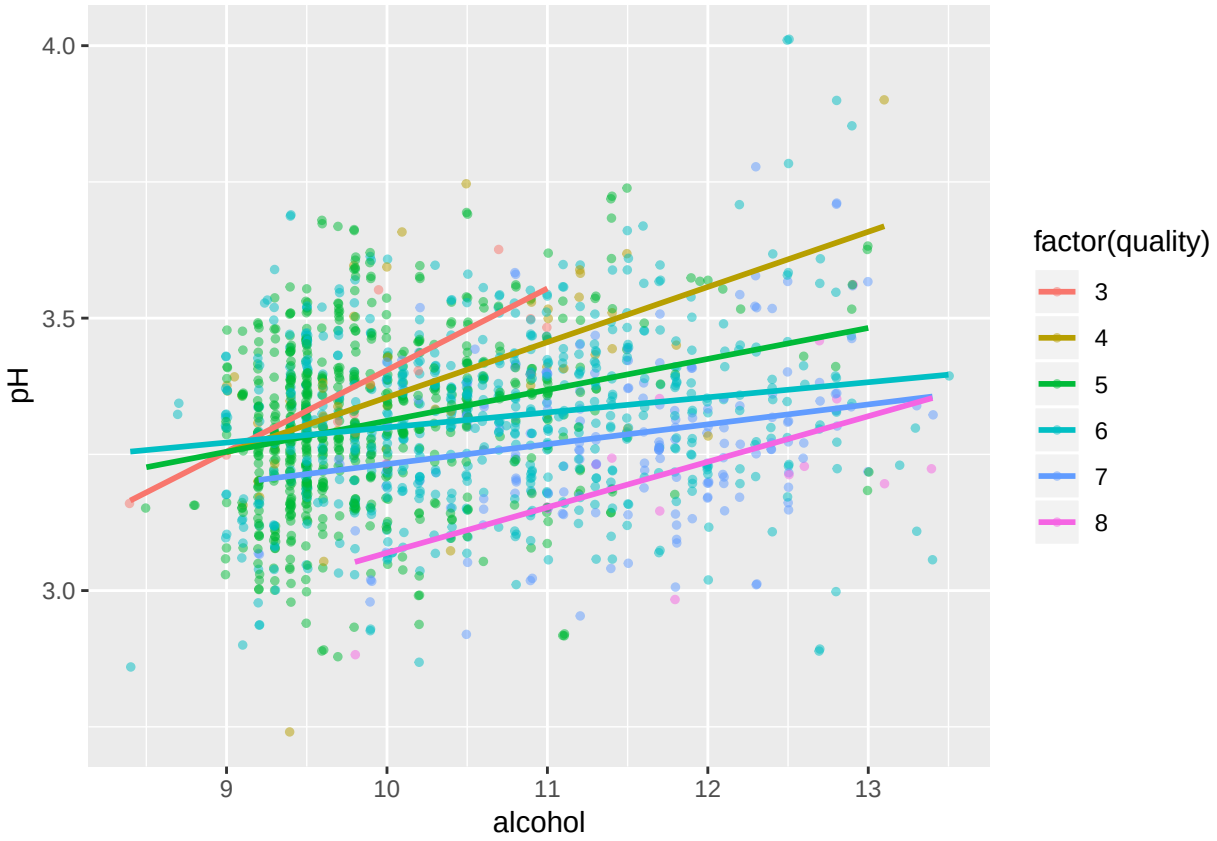
This plot shows the relationship of alcohol to sulphates. The trend lines show the quality scores across this data. Although alcohol and sulphates both have a medium correlation with quality, there is only a weak positive correlation between the two. This plot also shows that wines with quality scores of 4 and 5 have a positive trend in both alcohol and sulphates. The other scores are fairly constant across the wines or negative with increasing quality.




```
##
##  Pearson's product-moment correlation
##
## data:  df$alcohol and df$citric.acid
## t = 5.085, df = 1584, p-value = 4.113e-07
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.07800075 0.17486671
## sample estimates:
##       cor
## 0.1267359
```

Plot 7.2

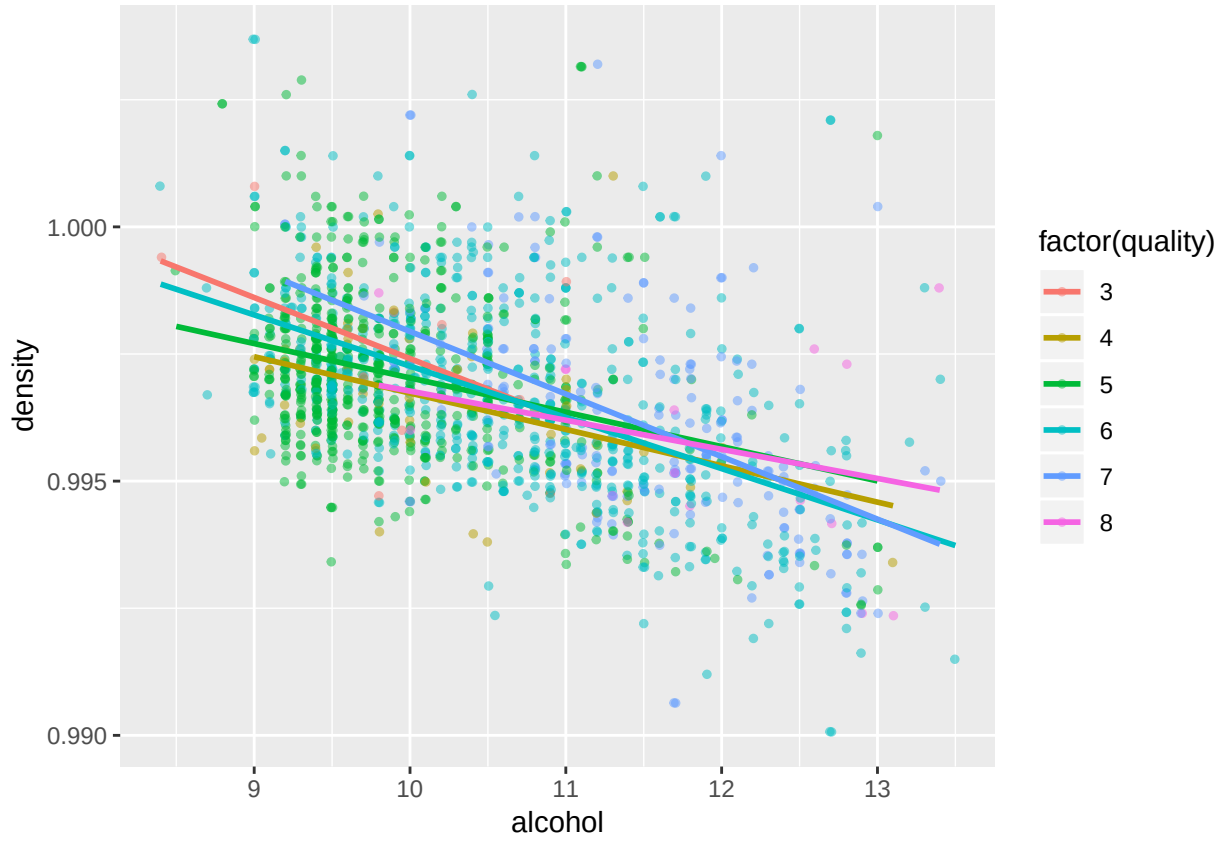
This plot shows the relationship between alcohol and citric acid, overlayed by quality across all wines. The plot shows that this relationship has a decreased affect on quality in the lower and upper quality wines, but has very little effect on quality in the mid-range of the scores. Because the pH of ethanol is slightly basic, it is logical that acidity decreases and concentration of alcohol increases. There is a weak positive correlation between the two.




```
##
##  Pearson's product-moment correlation
##
## data:  df$alcohol and df$pH
## t = 7.2762, df = 1584, p-value = 5.38e-13
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1317852 0.2270522
## sample estimates:
##       cor
## 0.1798404
```

Plot 7.3

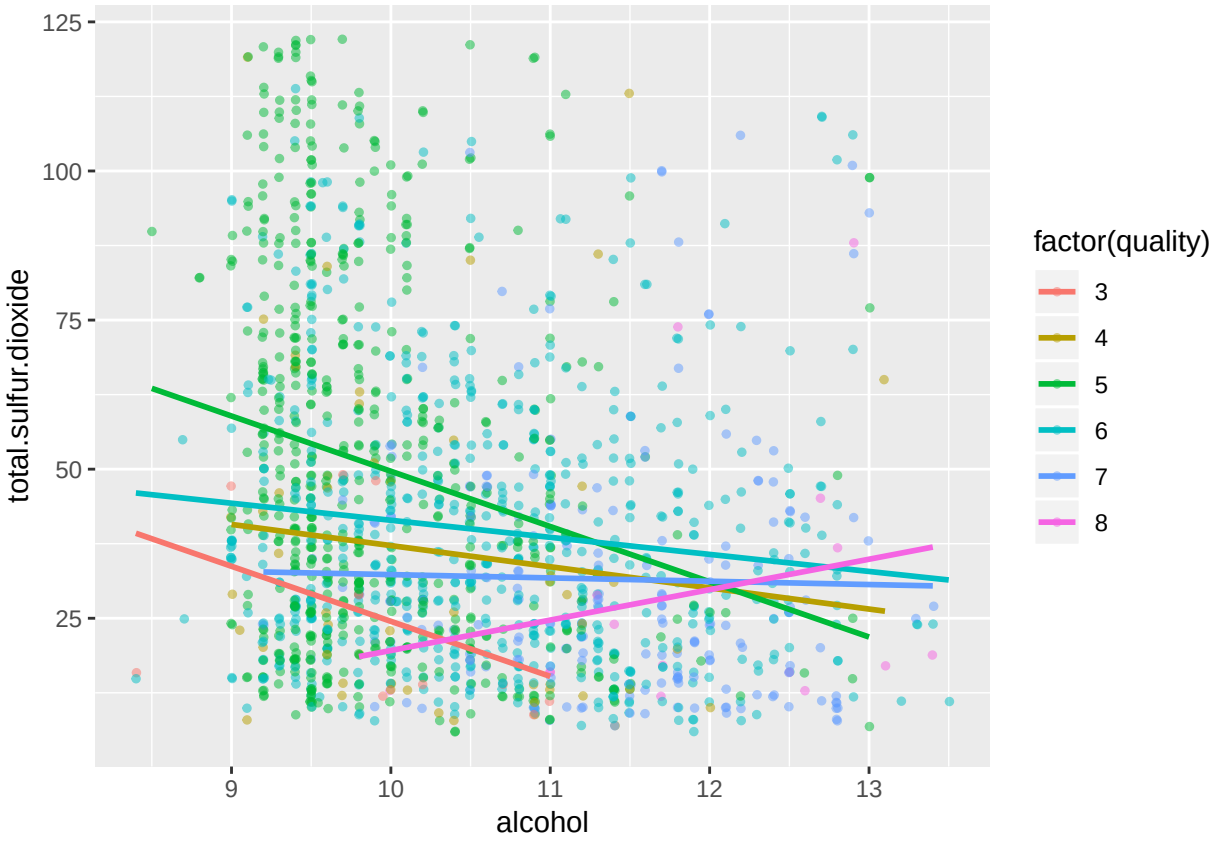
This plot shows the relationshop between alcohol and pH and is overlayed by quality. Here, quality increases with the increase in alcohol and the increase in pH. Because the pH of ethanol is slightly basic, it is logical that pH increases as the percentage of alcohol increases and that the quality increases with an increase in pH, as alcohol has a positive correlation with quality. There is a weak positive correlation between alcohol and pH.




```
##
##  Pearson's product-moment correlation
##
## data:  df$alcohol and df$density
## t = -21.048, df = 1584, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5051070 -0.4281391
## sample estimates:
##        cor
## -0.4675087
```

Plot 7.4

This plot shows the relationship between alcohol and density overlayed with quality. Here, there is a negative correlation between alcohol and density. Also, quality decreases over these wines. The density of wine is near 1, which is the density of water. The density of ethanol is 0.79. As the amount of alcohol increases, the density decreases, as does the quality with denisity. There is a medium negative correlation between alcohol concentraion and density.




```
##
##  Pearson's product-moment correlation
##
## data:  df$alcohol and df$total.sulfur.dioxide
## t = -9.7087, df = 1529, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2875984 -0.1932063
## sample estimates:
##        cor
## -0.2409721
```

Plot 7.5

This plot shows the relationship of alcohol to total sulfur dioxide with an overlay of quality. Generally, there is a decrease in sulfur dioxide and quality as alcohol increases. Because sulfur dioxide gas becomes sulfuric acid in water, it is logical that the amount of total sulfur dioxide would decrease as the amount of alcohol increases, making the wine more basic. There is a weak negative correlation between the two.

Plot 7.6

Below is the linear model for the comparison of all of the attributes and the effects they have on quality. It is not a traditonal plot, but it provides information to support the previously plotted data.

```
##
```

```
## Call:
## lm(formula = quality ~ alcohol + fixed.acidity + volatile.acidity +
##     citric.acid + residual.sugar + chlorides + free.sulfur.dioxide +
##     total.sulfur.dioxide + density + pH + sulphates, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.74325 -0.37598 -0.04084  0.42221  1.87447
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           32.765901  26.062605   1.257 0.208907
## alcohol                0.270697   0.032889   8.231 4.48e-16 ***
## fixed.acidity          0.010873   0.029781   0.365 0.715093
## volatile.acidity      -0.937247   0.131471  -7.129 1.67e-12 ***
## citric.acid           -0.312811   0.161576  -1.936 0.053085 .
## residual.sugar         0.065325   0.049872   1.310 0.190477
## chlorides             -1.713722   0.658840  -2.601 0.009397 **
## free.sulfur.dioxide    0.002525   0.002475   1.020 0.307759
## total.sulfur.dioxide  -0.002497   0.000927  -2.694 0.007153 **
## density              -28.268802  26.541850  -1.065 0.287044
## pH                    -0.703615   0.207512  -3.391 0.000718 ***
## sulphates              1.758337   0.166996  10.529  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6185 on 1300 degrees of freedom
##   (287 observations deleted due to missingness)
## Multiple R-squared:  0.3848, Adjusted R-squared:  0.3796
## F-statistic: 73.91 on 11 and 1300 DF,  p-value: < 2.2e-16
```

Plot 7.7

Model coefficients.

```
##                          Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept)           32.765901470 26.062605050  1.2571998 2.089071e-01
## alcohol                0.270697346  0.032889032  8.2306267 4.482106e-16
## fixed.acidity          0.010873131  0.029780845  0.3651048 7.150925e-01
## volatile.acidity      -0.937247001  0.131471151 -7.1289176 1.673092e-12
## citric.acid           -0.312811177  0.161576220 -1.9359976 5.308456e-02
## residual.sugar         0.065324985  0.049871956  1.3098541 1.904766e-01
## chlorides             -1.713722311  0.658840249 -2.6011196 9.397474e-03
## free.sulfur.dioxide    0.002525334  0.002475004  1.0203351 3.077594e-01
## total.sulfur.dioxide  -0.002497162  0.000926964 -2.6939151 7.152705e-03
## density              -28.268801587 26.541850367 -1.0650652 2.870442e-01
## pH                    -0.703614992  0.207511938 -3.3907206 7.180067e-04
## sulphates              1.758336694  0.166996011 10.5292138 6.182528e-25
```

Plot 7.8

Confidence intervals.

```
##                              2.5 %        97.5 %
```

```
## (Intercept)          -18.363468997 83.8952719367
## alcohol                0.206175956  0.3352187367
## fixed.acidity         -0.047550648  0.0692969097
## volatile.acidity      -1.195165854 -0.6793281487
## citric.acid           -0.629789868  0.0041675138
## residual.sugar        -0.032513344  0.1631633138
## chlorides             -3.006228839 -0.4212157830
## free.sulfur.dioxide   -0.002330106  0.0073807730
## total.sulfur.dioxide  -0.004315672 -0.0006786532
## density              -80.338350953 23.8007477800
## pH                    -1.110709935 -0.2965200479
## sulphates              1.430725510  2.0859478775
```

8. Multivariate Analysis

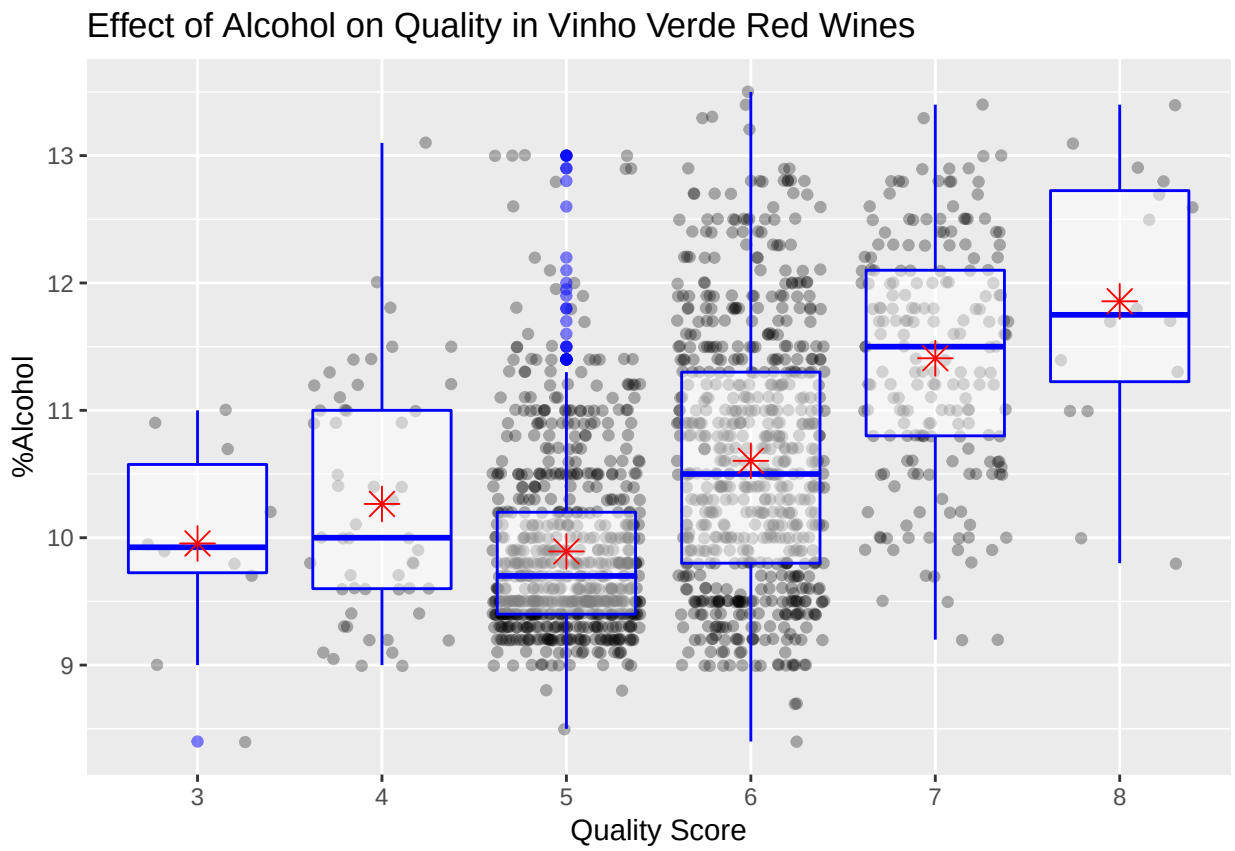
The multivariate plots show the relationships of the attributes to quality. They also show how each attribute behaves across the spectrum of taste preference. Quality is not correlated to any one attribute strongly. Neither are any attributes strongly correlated with each other. There are some medium and weak correlations.

Of interest is how the attributes behave in the presence of alcohol. Because the density of alcohol is less than the density of water and the pH of the alcohol is greater than the pH of water, acidity decreases and pH increases as the concentration of alcohol increases.

A regression analysis indicates that the attribute with the lowest impact on the quality of wines is volatile acidity. All of the other attributes are more significant. However, it is important to note that pH and density will change as the concentration of alcohol changes.

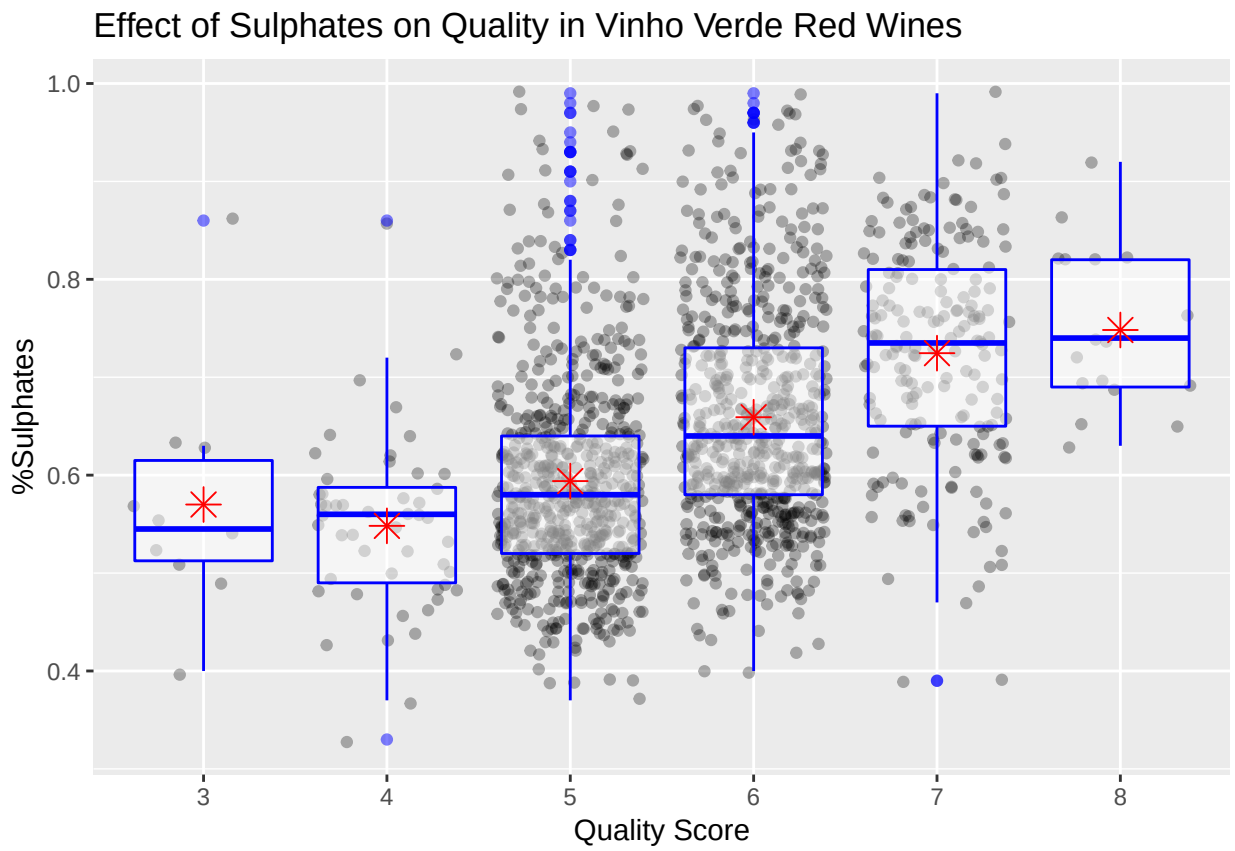
9. Final Plots and Summary

The preceding analysis has provided some insight into consumer preference regarding Vinho Verde Red Wines, the most remarkable being that no one attribute of the wines has a strong effect on perceived quality. Below are some findings of interest.



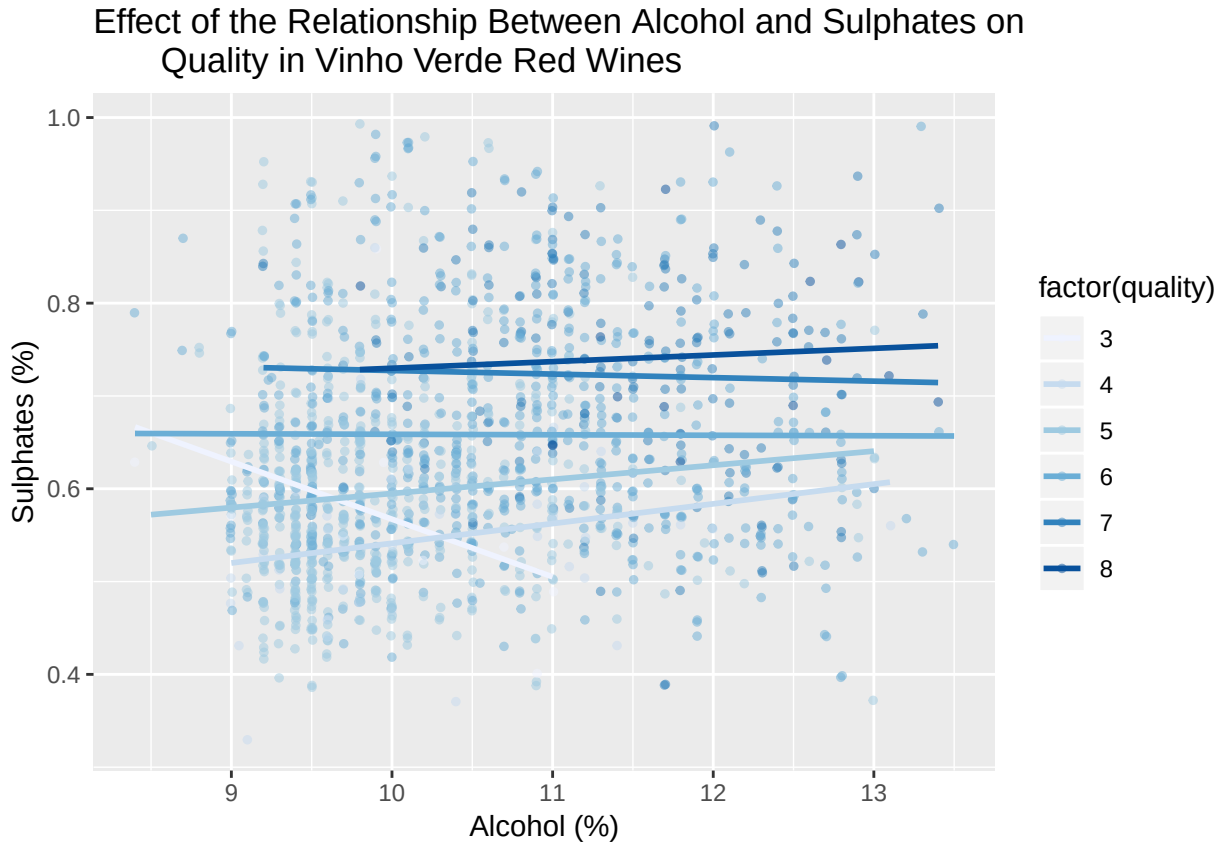
Plot 9.1

Alcohol has the strongest correlation with quality of any of the attributes. The above plot shows the relationship between alcohol concentration and quality score. Most of the wines fall into the 5 and 6 quality category. Generally, quality score increases with alcohol concentration increase.



Plot 9.2

Sulphate concentration has one of the strongest correlations with quality of any of the attributes. The above plot shows the relationship between sulphate concentration and quality score. Most of the wines fall into the 5 and 6 quality category. Generally, quality score increases with sulphate concentration increase.



Plot 9.3

The above plot shows how alcohol and sulphates combine to affect quality. The highest quality of the wines contains the highest concentrations of both alcohol and sulphates.

Summary

The analysis of how wine attributes relate to consumer preference, or quality, shows that none of the attributes in the data set are strongly correlatetd with quality. The linear model constructed in 7.5 confirms this. All of the attributes are significant in their relationships to quality.

While some of the attributes decrease with increase in quality score and some increase with increase in quality score, they all contribute to the taste of the wine. Therefore, the conclusion is the it is the mixture of all of the attributes that affect consumer preferences for wine.

11. Reflection

This analysis shows how certain attributes of vinho verde red wines affect consumer preference, which is summarized in a quality score. The analysis was deceptively difficult. None of the attributes had a strong correlation with quality score or with the other attributes. And, some of the attributes are influenced by other attributes; i.e. pH and density, which are affected by alcohol concentration.

Because quality is a discrete data attribute, it was difficult to compare quality to the other attributes, as they are continuous and the range of values of the attributes do not fit cleanly within the quality score ranges.

The linear model confirms this by showing that all of the attributes are significant when determining wine quality. Additionally, quality is subjective and dependent on individual persons.

Future analysis might include the consumer preferences of the wines over time. How do the attributes of the wine change with age in the barrel and after purchase? Might there be other attributes that add to quality of the wine? What is the optimal age for these wines? When does the taste begin to decline? It may also be valuable to compare red wine quality to white wine quality.

12. Notes

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

13. References used

jdatalab.com

dummies.com for correlation coefficient

stackoverflow and RPubs for plotting 2 lines

Google search for normalizing data

ggplot2 Elegant Graphics for Data Analysis by Hadley Wickham

Google search for densities of water and ethanol

todo.science.blogspot.com

stackoverflow for printing test with variables

http://www.sthda.com/ for correlation matrix heatmap and multivariate linear models

stackoverflow for font size

stackoverflow for suppressing warnings and output