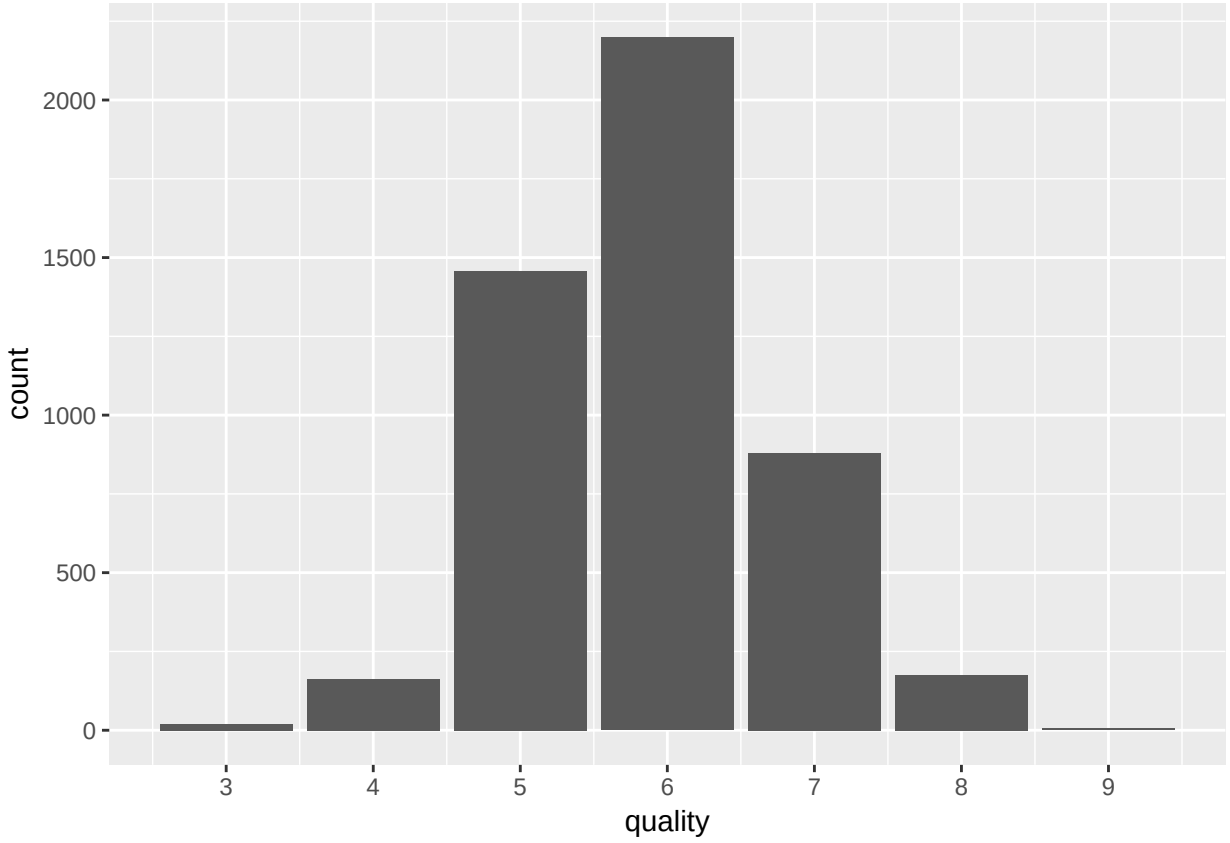
Wine Quality

Onesmas Ngugi

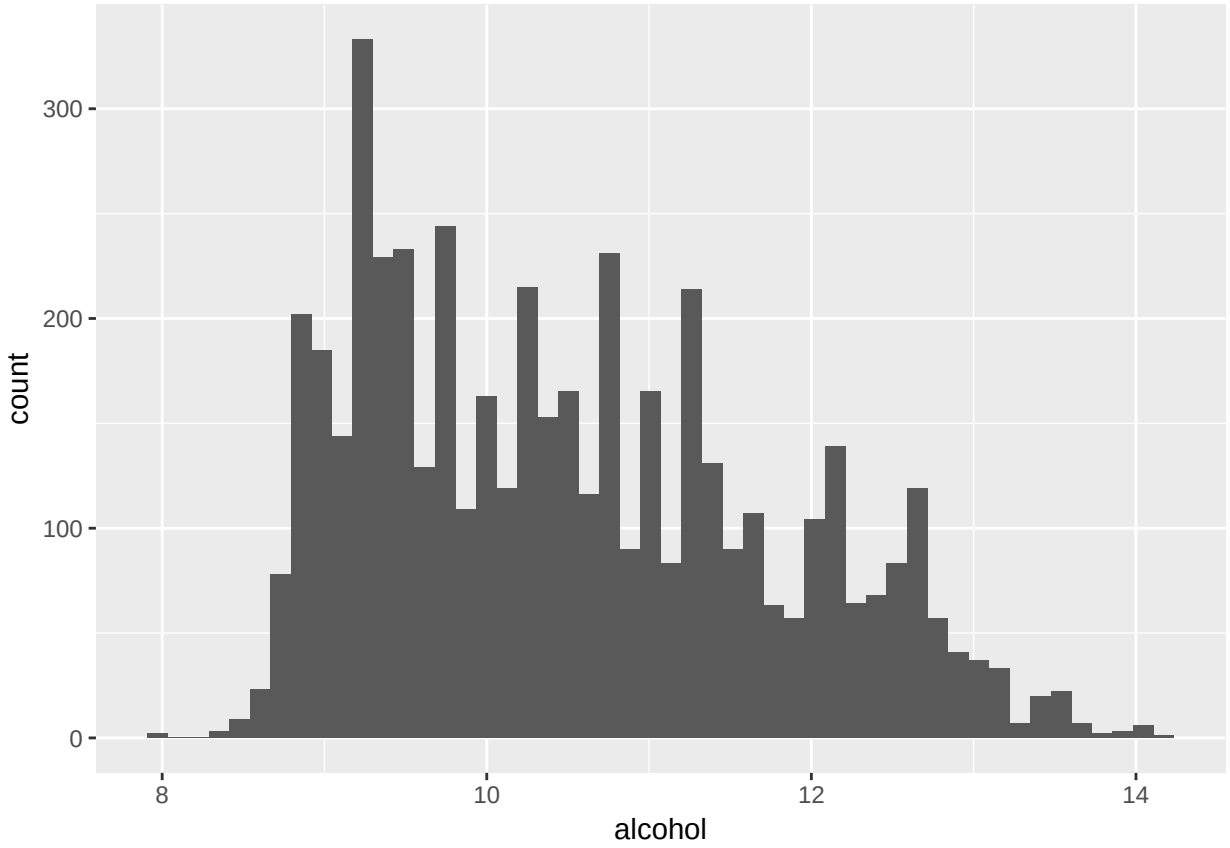
Abstract

Wine quality is usually very subjective and can vary greatly among drinkers. However, some wines can have high average quality ratings from multiple different drinkers and wine experts. This would seem to suggest that wine quality, rated by wine experts in this dataset, can be partially explained by different physical attributes, such as alcohol content, acidity, e.t.c. of the wine. I wanted to explore which attributes affect wine quality the most. Regression analysis can be used to estimate relationships between wine quality and different attributes of the wine.

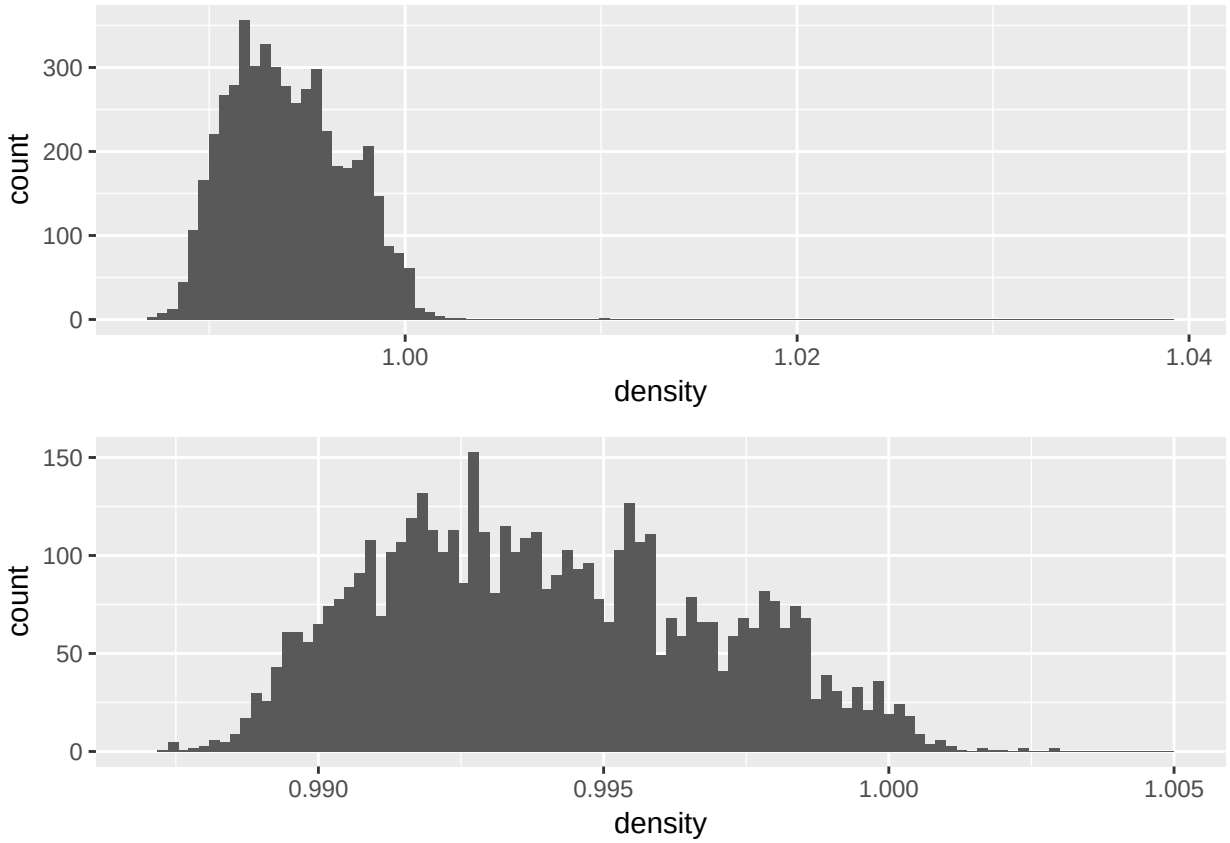
Univariate Plots Section



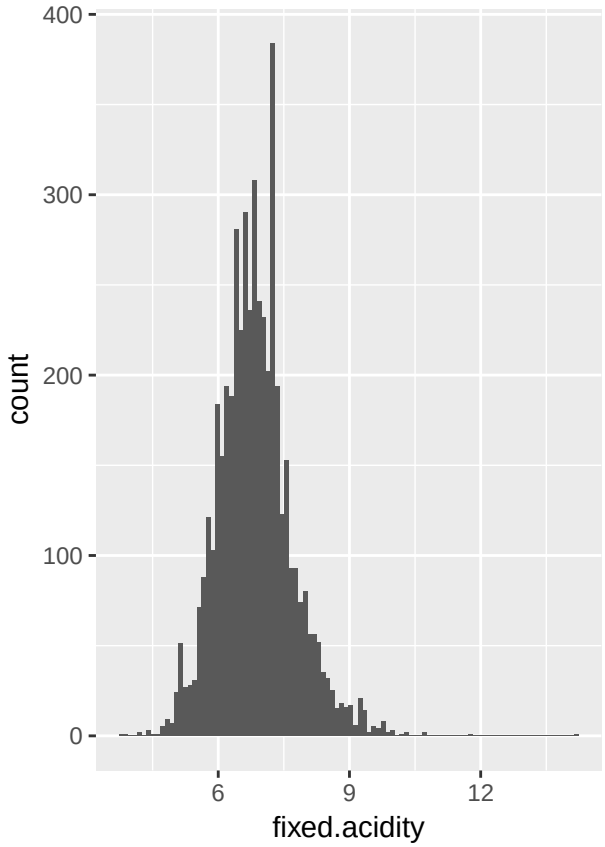
Distribution of wine quality ratings. Majority if wines rated between 5 and 8.

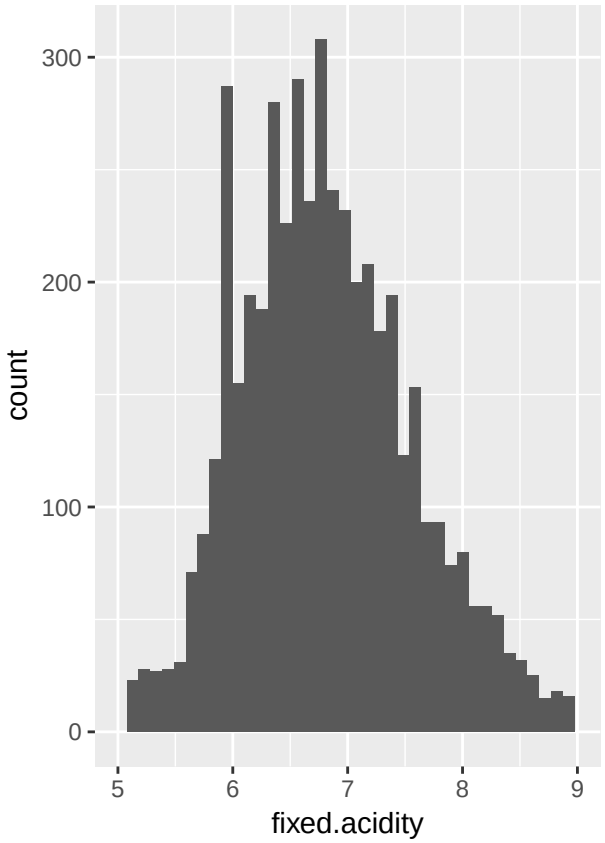


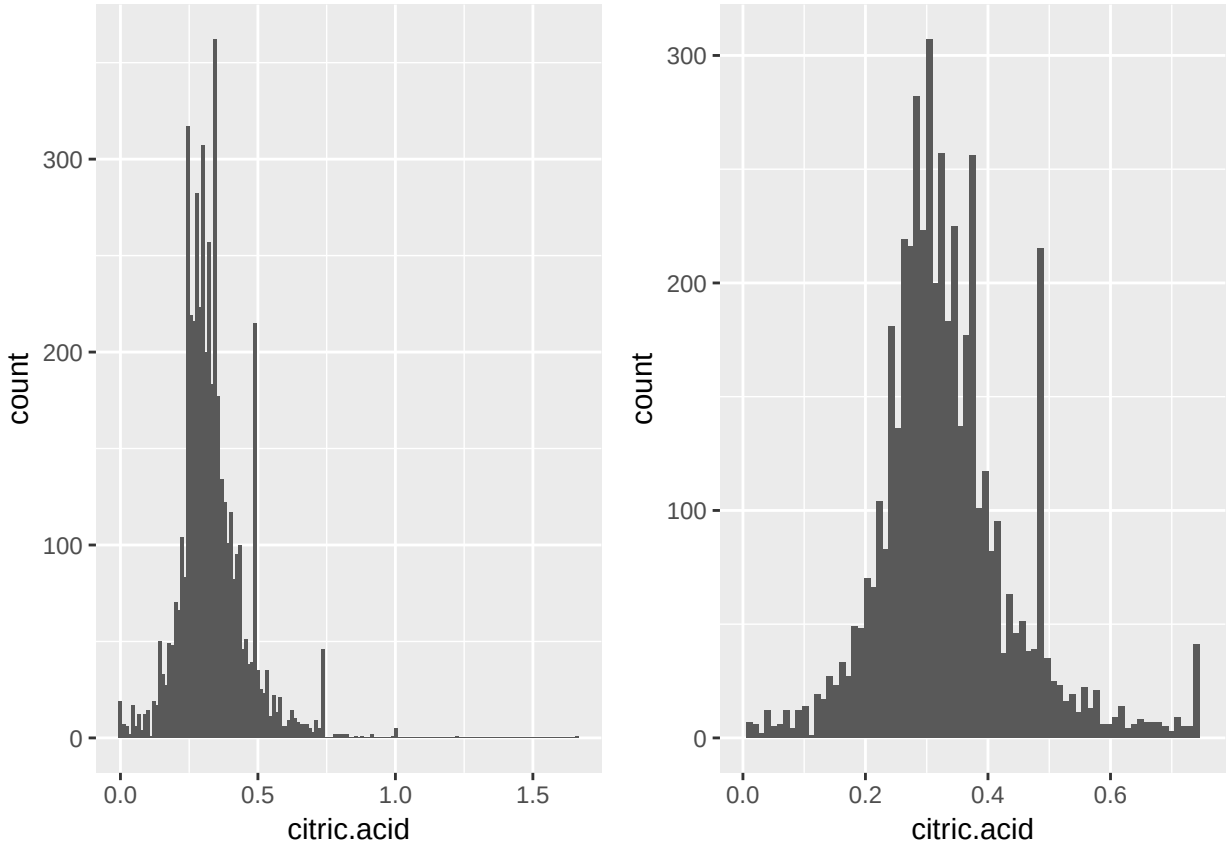
Alcohol content for most wines is between 8 and 10 % by volume.



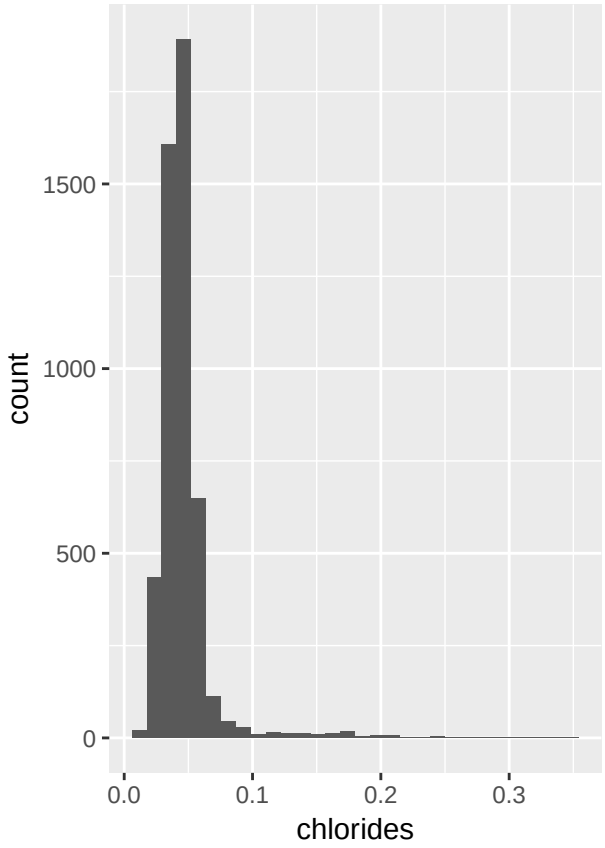
Most wines have densities between 0.985 and 1.005 g/cm^3. The first plot however, shows that the are some outliers in the data, while the second plot limits the scale to exclude the outliers.

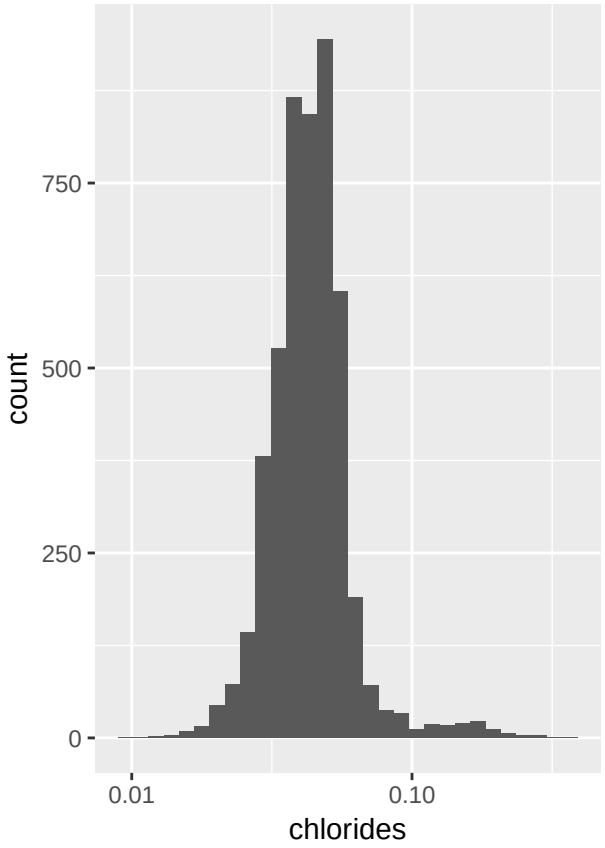

400
300
200
100
0
count
6
9
12
fixed.acidity

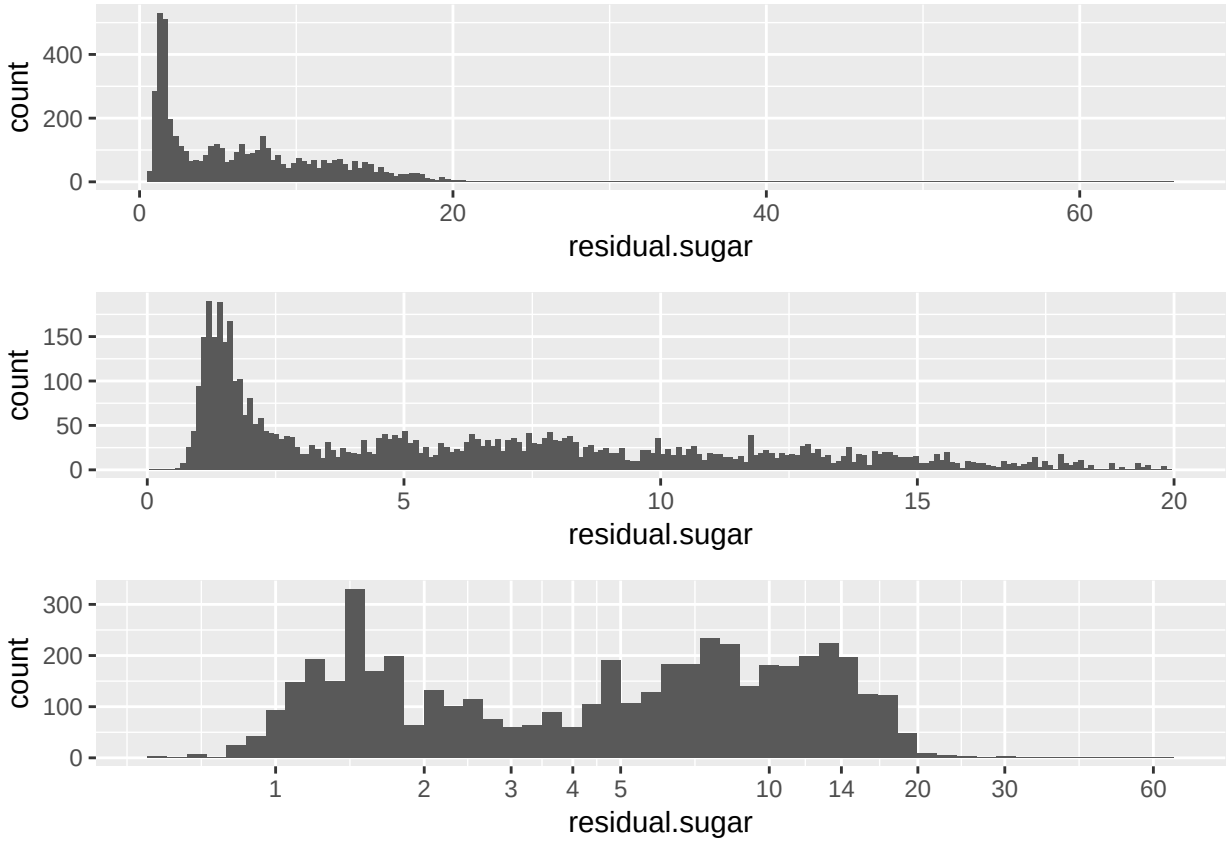

300
200
100
0
count
5
6
7
8
9
fixed.acidity



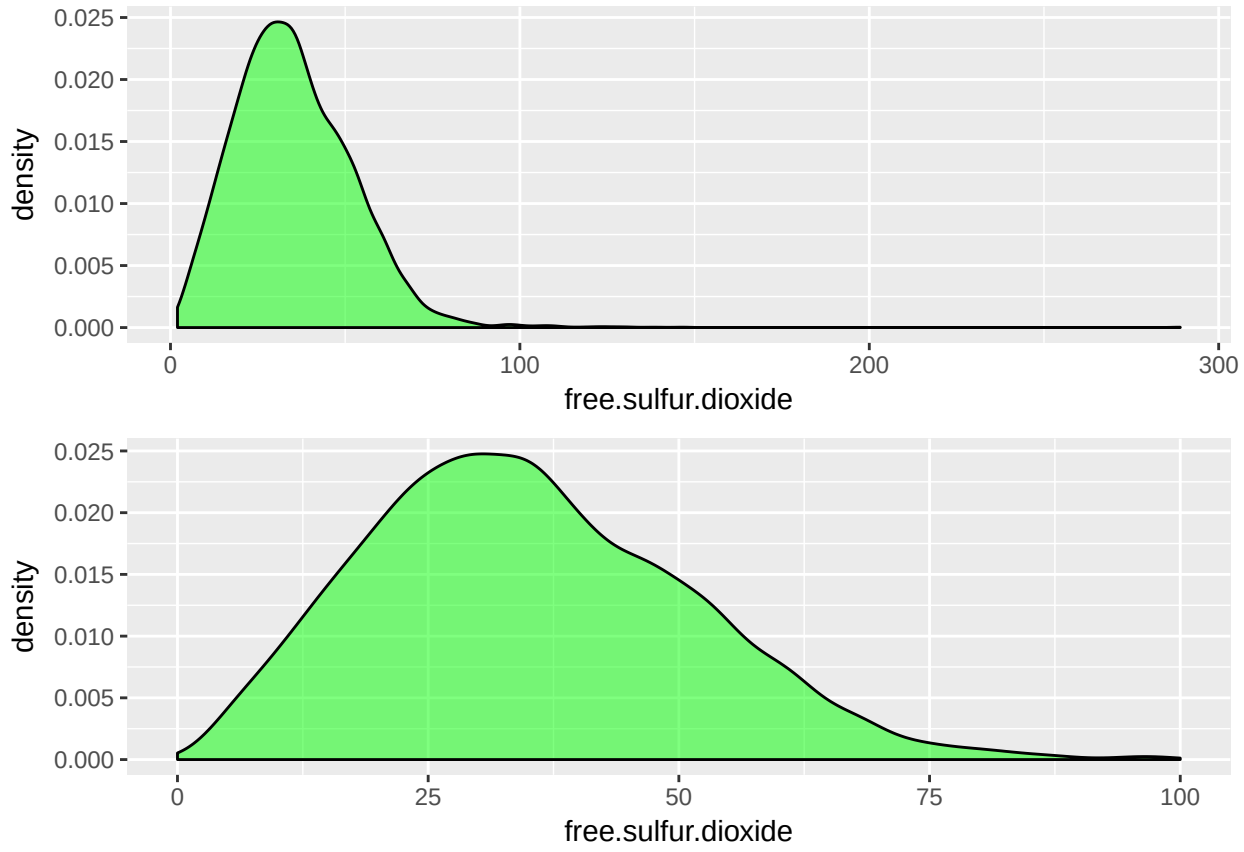
The two plots above show distributions for fixed acidity and citric acid with and without ouliers. No transformations besides limiting the scale were done on the plots.


count
1500
1000
500
0
0.0
0.1
0.2
0.3
chlorides


count
750
500
250
0
0.01
0.10
chlorides



For residual sugar and chlorides, the distribution of the data is long-tailed meaning there are a lot of values on the extreme right. This is different from the other plots that had a few outliers that could be dealt with by simply limiting the scale. In long-tail distribution, scale transformatios are necessary. I used log10 transformation of the scales as showen in the the plots above. For residual sugar, I included an original plot, a plot with scale limit and a third plot with scale transformation. The rransformation shows bimodal distribution with peaks at approximately 1.5 and 14 (g/dm^3).



Different kind of plot showing distribution of free sulfur dioxide.

```
## 'data.frame':    4898 obs. of  13 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ fixed.acidity       : num  7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...
##  $ volatile.acidity    : num  0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 ...
##  $ citric.acid         : num  0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34 0.43 ...
##  $ residual.sugar      : num  20.7 1.6 6.9 8.5 8.5 6.9 7 20.7 1.6 1.5 ...
##  $ chlorides           : num  0.045 0.049 0.05 0.058 0.058 0.05 0.045 0.045 0.049 0.044 ...
##  $ free.sulfur.dioxide : num  45 14 30 47 47 30 30 45 14 28 ...
##  $ total.sulfur.dioxide: num  170 132 97 186 186 97 136 170 132 129 ...
##  $ density             : num  1.001 0.994 0.995 0.996 0.996 ...
##  $ pH                  : num  3 3.3 3.26 3.19 3.19 3.26 3.18 3 3.3 3.22 ...
##  $ sulphates           : num  0.45 0.49 0.44 0.4 0.4 0.44 0.47 0.45 0.49 0.45 ...
##  $ alcohol             : num  8.8 9.5 10.1 9.9 9.9 10.1 9.6 8.8 9.5 11 ...
##  $ quality             : int  6 6 6 6 6 6 6 6 6 6 ...
```

```
##   X fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1 1           7.0             0.27        0.36           20.7     0.045
## 2 2           6.3             0.30        0.34            1.6     0.049
## 3 3           8.1             0.28        0.40            6.9     0.050
## 4 4           7.2             0.23        0.32            8.5     0.058
## 5 5           7.2             0.23        0.32            8.5     0.058
## 6 6           8.1             0.28        0.40            6.9     0.050
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  45                  170  1.0010 3.00      0.45     8.8
## 2                  14                  132  0.9940 3.30      0.49     9.5
```

```
## 3                 30                  97  0.9951 3.26      0.44    10.1
## 4                 47                 186  0.9956 3.19      0.40     9.9
## 5                 47                 186  0.9956 3.19      0.40     9.9
## 6                 30                  97  0.9951 3.26      0.44    10.1
##   quality
## 1       6
## 2       6
## 3       6
## 4       6
## 5       6
## 6       6
```

```
## [1] FALSE
```

```
##        X        fixed.acidity    volatile.acidity  citric.acid
##  Min.   :   1   Min.   : 3.800   Min.   :0.0800   Min.   :0.0000
##  1st Qu.:1225   1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700
##  Median :2450   Median : 6.800   Median :0.2600   Median :0.3200
##  Mean   :2450   Mean   : 6.855   Mean   :0.2782   Mean   :0.3342
##  3rd Qu.:3674   3rd Qu.: 7.300   3rd Qu.:0.3200   3rd Qu.:0.3900
##  Max.   :4898   Max.   :14.200   Max.   :1.1000   Max.   :1.6600
##  residual.sugar     chlorides       free.sulfur.dioxide
##  Min.   : 0.600   Min.   :0.00900   Min.   :  2.00
##  1st Qu.: 1.700   1st Qu.:0.03600   1st Qu.: 23.00
##  Median : 5.200   Median :0.04300   Median : 34.00
##  Mean   : 6.391   Mean   :0.04577   Mean   : 35.31
##  3rd Qu.: 9.900   3rd Qu.:0.05000   3rd Qu.: 46.00
##  Max.   :65.800   Max.   :0.34600   Max.   :289.00
##  total.sulfur.dioxide    density             pH          sulphates
##  Min.   :  9.0        Min.   :0.9871   Min.   :2.720   Min.   :0.2200
##  1st Qu.:108.0        1st Qu.:0.9917   1st Qu.:3.090   1st Qu.:0.4100
##  Median :134.0        Median :0.9937   Median :3.180   Median :0.4700
##  Mean   :138.4        Mean   :0.9940   Mean   :3.188   Mean   :0.4898
##  3rd Qu.:167.0        3rd Qu.:0.9961   3rd Qu.:3.280   3rd Qu.:0.5500
##  Max.   :440.0        Max.   :1.0390   Max.   :3.820   Max.   :1.0800
##     alcohol         quality
##  Min.   : 8.00   Min.   :3.000
##  1st Qu.: 9.50   1st Qu.:5.000
##  Median :10.40   Median :6.000
##  Mean   :10.51   Mean   :5.878
##  3rd Qu.:11.40   3rd Qu.:6.000
##  Max.   :14.20   Max.   :9.000
```

```
## # A tibble: 7 x 2
##   quality     n
##     <int> <int>
## 1       3    20
## 2       4   163
## 3       5  1457
## 4       6  2198
## 5       7   880
## 6       8   175
## 7       9     5
```

Univariate Analysis

Data structure

The dataset was created by Paulo Cortez (Univ. Minho), Antonio Cerdeira, Fernando Almeida, Telmo Matos and Jose Reis (CVRVV) @ 2009 for their "Modeling wine preferences by data mining from physicochemical properties" paper.

From the data structure above, we can see that the data set has 13 variables including quality and 4898 observations and that the are no null values in the data set. The quality variable was based on sensory data (median of at least 3 evaluations made by wine experts) while all other input variables were objective tests.

The main object of interest is quality. I will explore how the other objective attributes affect quality.

Observations

The following observations are from the summaries above:

- Some of the variable seem to have outliers as evidenced by the much higher max value compared to corresponding mean and median values. For example, free.sulfur.dioxide max value is 289.00 while the mean and median are 35.31 and 34.00 repectively.
- Most wines had quality ratings between 4 and 8. Only twenty wines were rated 3 and only five wines were rated 9. There were no wines rated 1, 2 or 10.

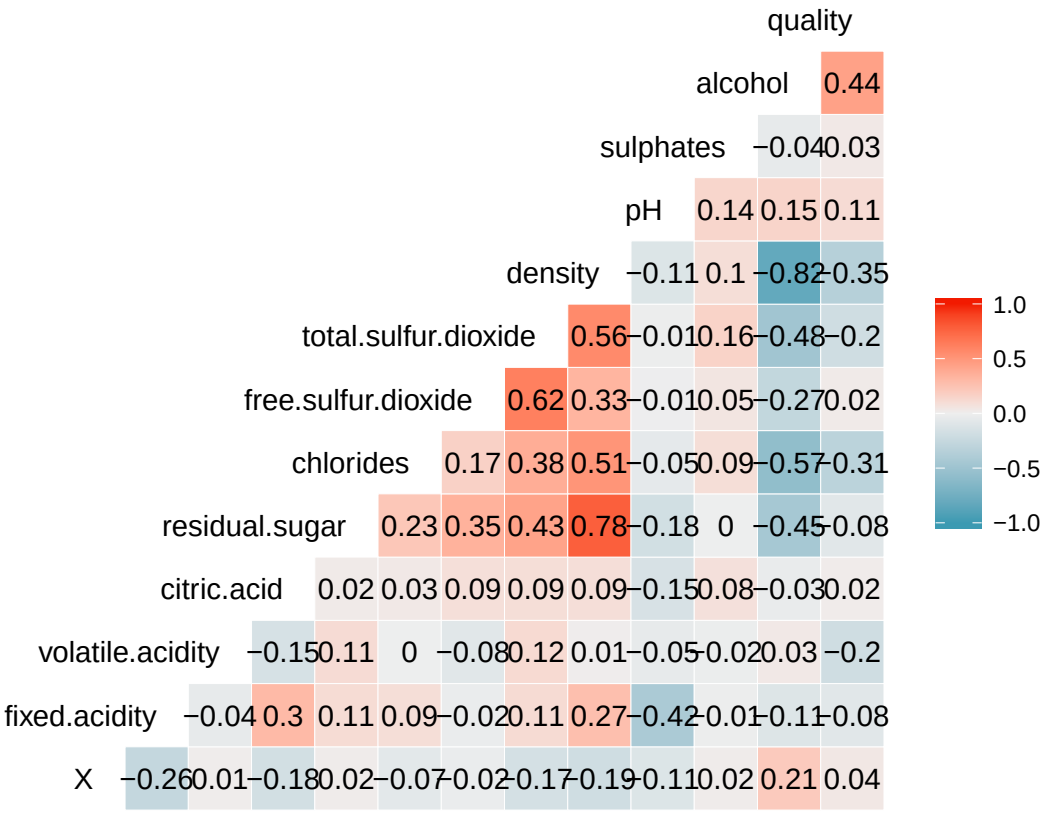
Transformations

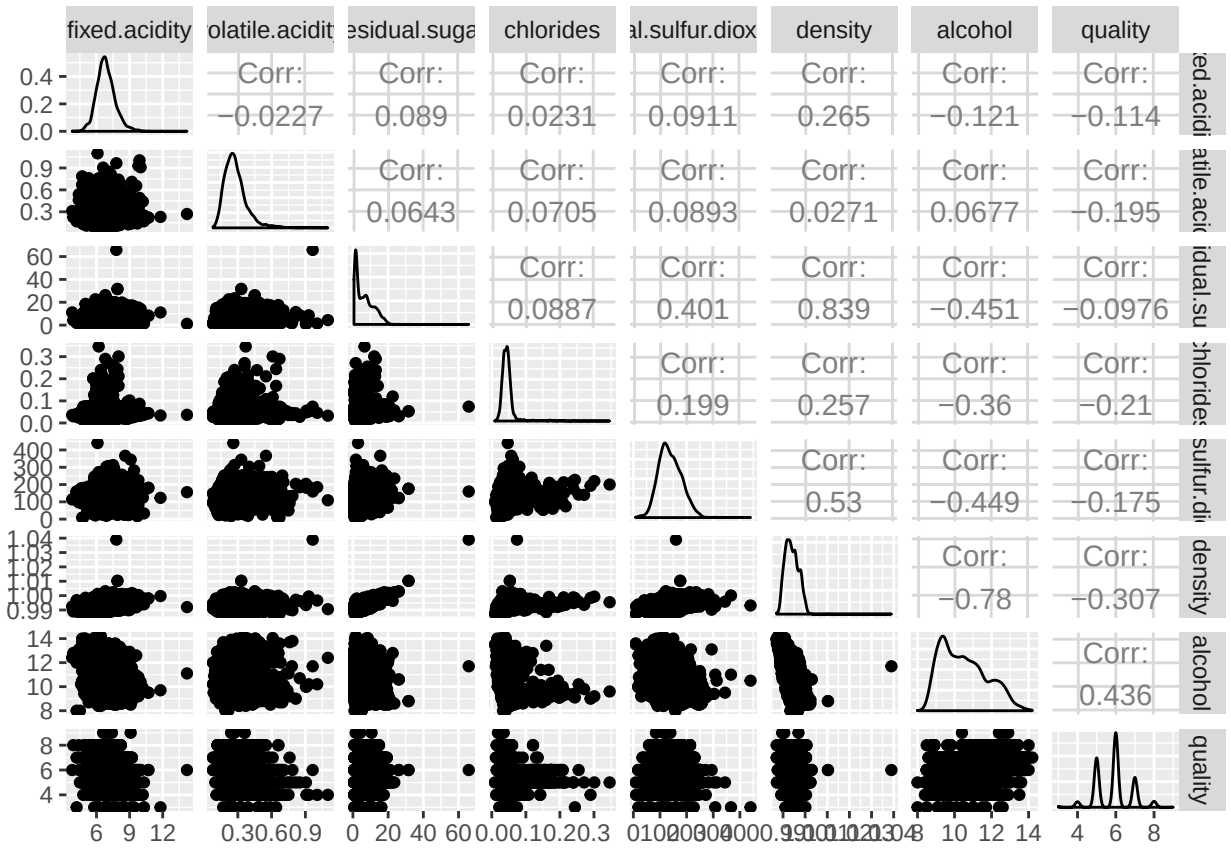
Some of the variables had outliers that I transformed using scale limits since there were few oulier values. However, for chlorides and residual sugar, I used log10 to transform the data as distribution was heavily right skewed. The transformed residual sugar has bimodal distribution.

Bivariate Plots Section

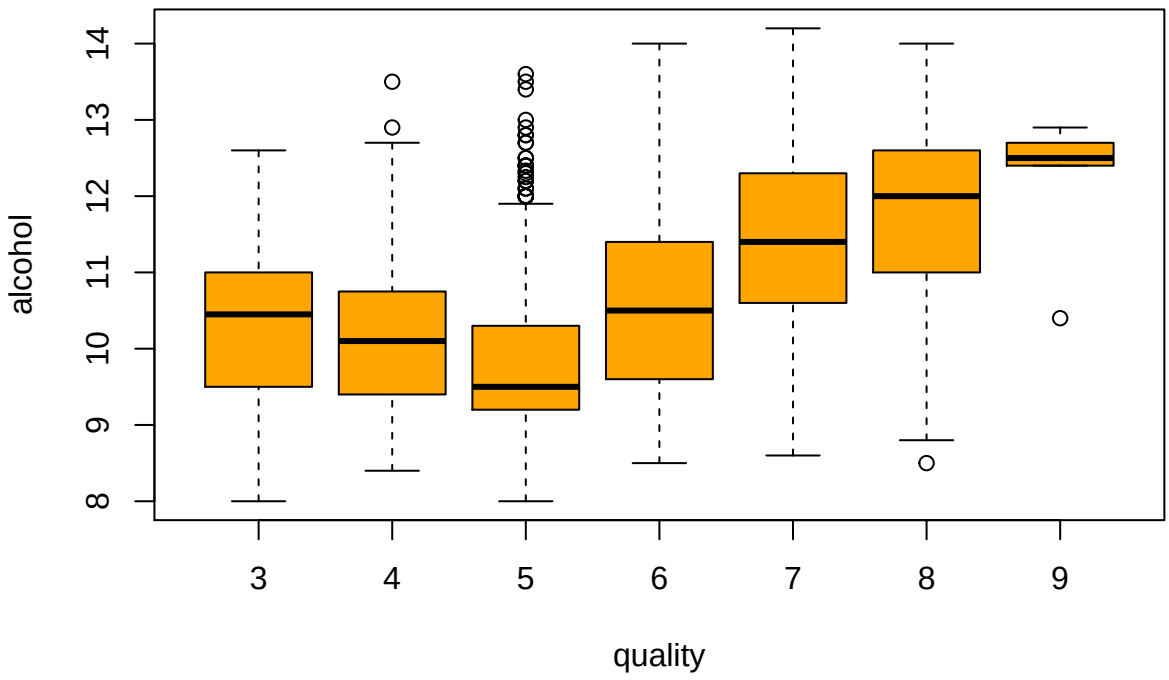
Bivariate analysis allows us to see relationships between two variables which we can use to see which attributes affect wine quality and the extent of these effects.

Using ggpairs, we can get a snapshot of relationships between variables in our data set. However, since the data set has 13 variables, ggpairs plot might be too congested and not clear enough to infer any valuable information from, thus I picked a few variables that were the most correlated with quality.


quality
alcohol 0.44
sulphates −0.04 0.03
pH 0.14 0.15 0.11
density −0.11 0.1 −0.82 −0.35
total.sulfur.dioxide 0.56 −0.01 0.16 −0.48 −0.2
free.sulfur.dioxide 0.62 0.33 −0.01 0.05 −0.27 0.02
chlorides 0.17 0.38 0.51 −0.05 0.09 −0.57 −0.31
residual.sugar 0.23 0.35 0.43 0.78 −0.18 0 −0.45 −0.08
citric.acid 0.02 0.03 0.09 0.09 0.09 −0.15 0.08 −0.03 0.02
volatile.acidity −0.15 0.11 0 −0.08 0.12 0.01 −0.05 −0.02 0.03 −0.2
fixed.acidity −0.04 0.3 0.11 0.09 −0.02 0.11 0.27 −0.42 −0.01 −0.11 −0.08
X −0.26 0.01 −0.18 0.02 −0.07 −0.02 −0.17 −0.19 −0.11 0.02 0.21 0.04
1.0
0.5
0.0
−0.5
−1.0



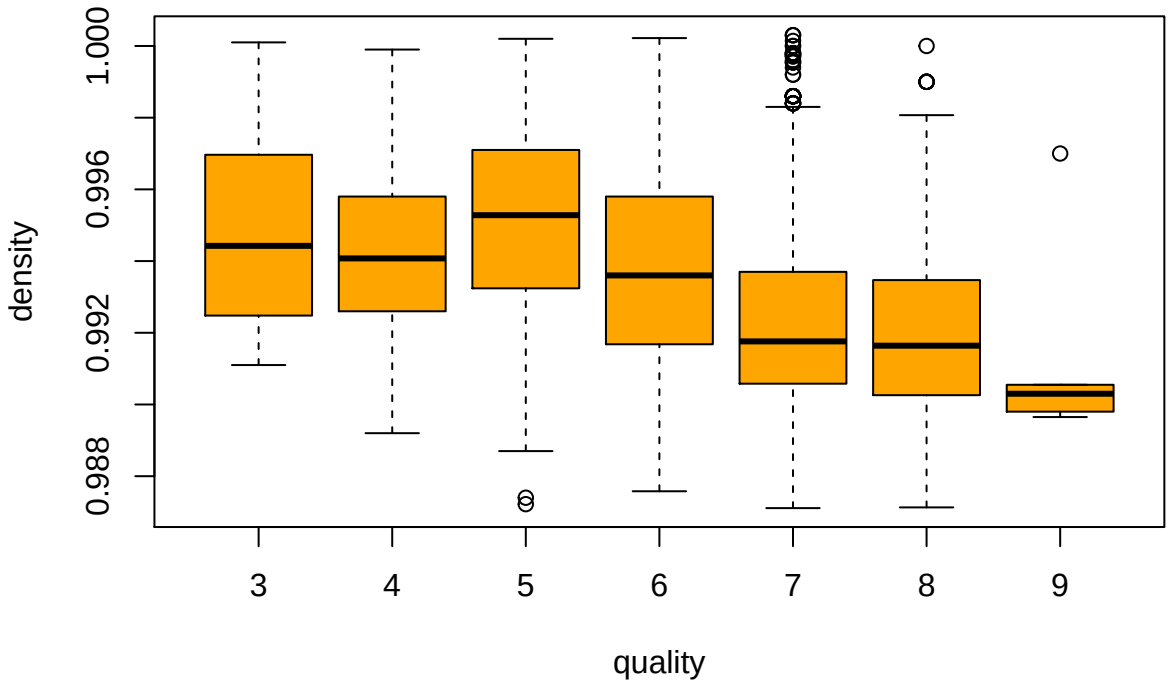
None of the variables are too highly correlated with quality with the highest correlation being alcohol correlation coefficient of 0.440. However, there are still ouliers in the data which might be affecting correlation.



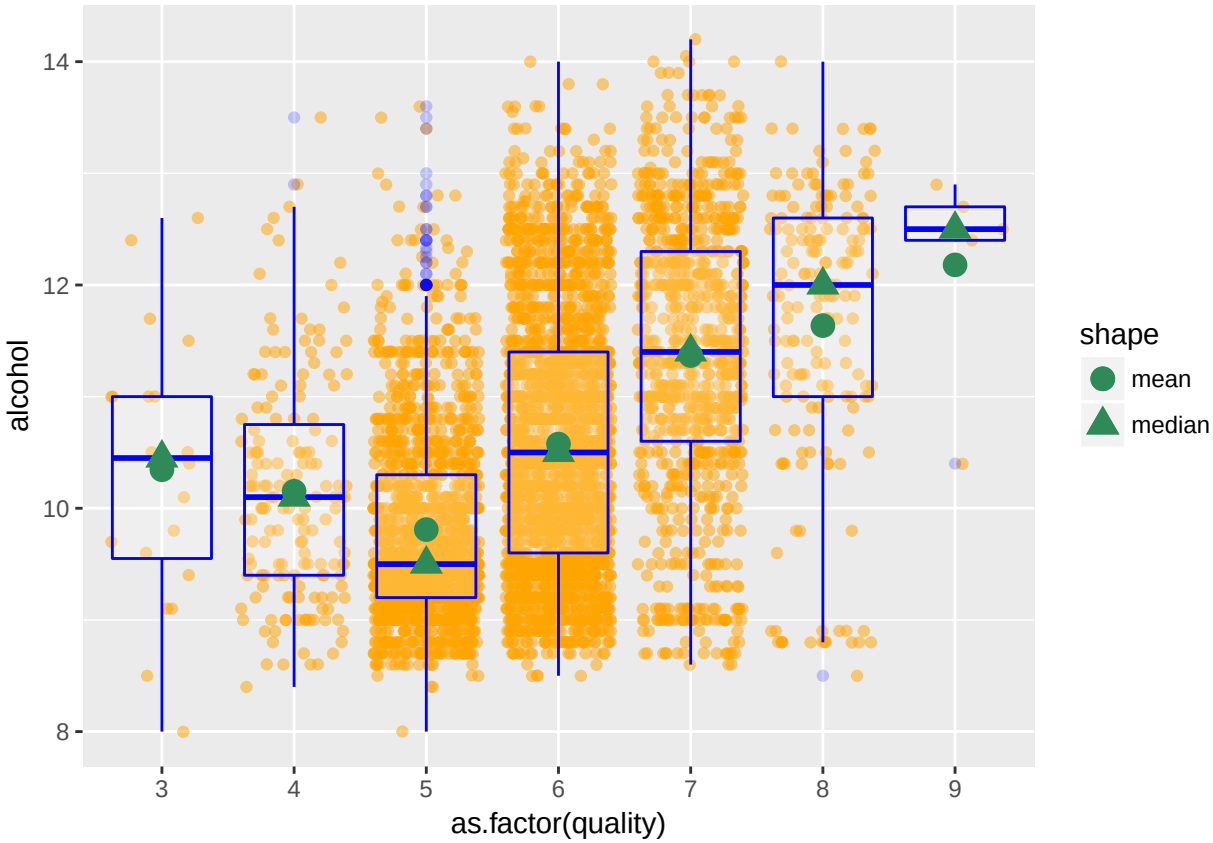

```
## wines$quality: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00    9.55   10.45   10.35   11.00   12.60
## ------------------------------------------------------------
## wines$quality: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.40   10.10   10.15   10.75   13.50
## ------------------------------------------------------------
## wines$quality: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   8.000   9.200   9.500   9.809  10.300  13.600
## ------------------------------------------------------------
## wines$quality: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.50    9.60   10.50   10.58   11.40   14.00
## ------------------------------------------------------------
## wines$quality: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.60   10.60   11.40   11.37   12.30   14.20
## ------------------------------------------------------------
## wines$quality: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.50   11.00   12.00   11.64   12.60   14.00
## ------------------------------------------------------------
## wines$quality: 9
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

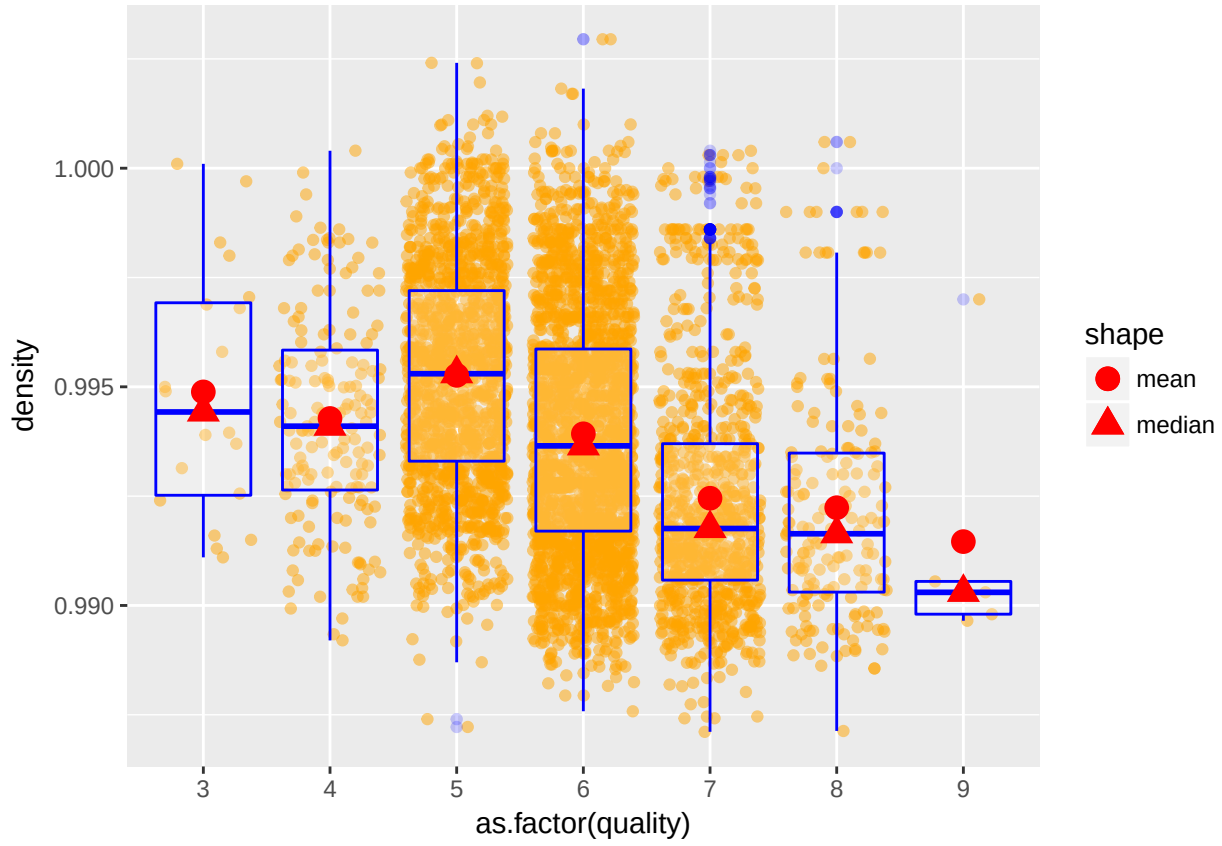
```
##   10.40   12.40   12.50   12.18   12.70   12.90
```

The boxplots above show alcohol content distribution for wines grouped by quality and the data is summarized above.



In the above boxplots, I removed extreme outliers either by subsetting the data or by limiting the scale. Subsetting the data changes the shape of the boxplots as calcuations exclude outliers, while limiting the scale only affects the plot but not the underlying calculations.


14
12
10
8
alcohol
3
4
5
6
7
8
9
as.factor(quality)
shape
mean
median



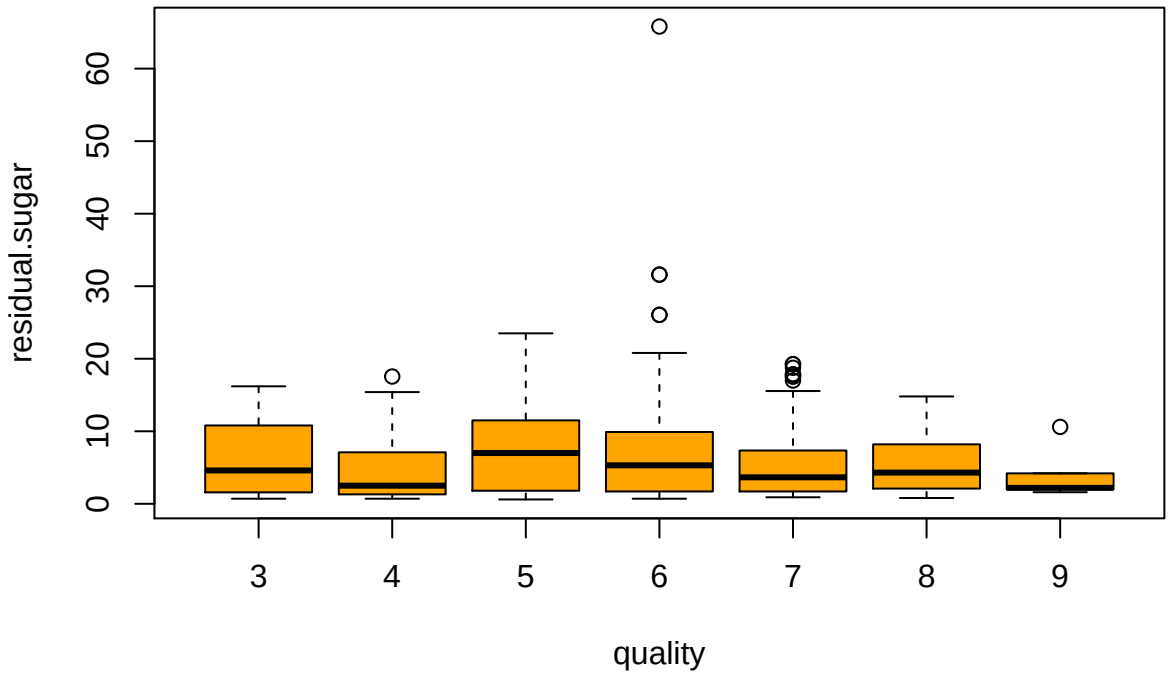
Above are scatter plots of alcohol vs quality and density vs quality with boxplots overlaid to show variance within each quality group. Using the scatter plots alone, it is hard to see the relationship between the variables, but the boxplots with the mean and median for each quality group, help illustrate these relationships

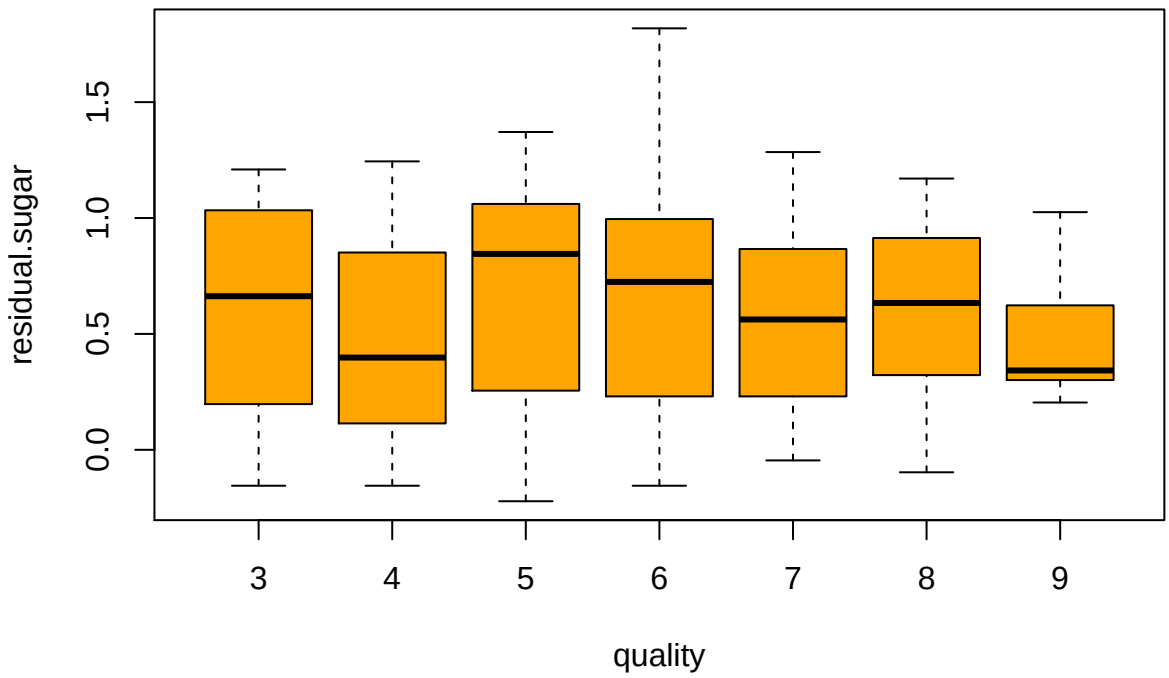
Mean and median alcohol increase across quality while mean and median density decrease with higher quality groups. However, wines rated 5 seem to break these trend for both alcohol and density.

```
##
## Call:
## lm(formula = quality ~ alcohol + density, data = wines)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5670 -0.5242 -0.0003  0.4881  3.0898
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -22.49170    6.16503  -3.648 0.000267 ***
## alcohol       0.36036    0.01478  24.389  < 2e-16 ***
## density      24.72842    6.07937   4.068 4.82e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.796 on 4895 degrees of freedom
## Multiple R-squared:  0.1925, Adjusted R-squared:  0.1921
## F-statistic: 583.3 on 2 and 4895 DF,  p-value: < 2.2e-16
```

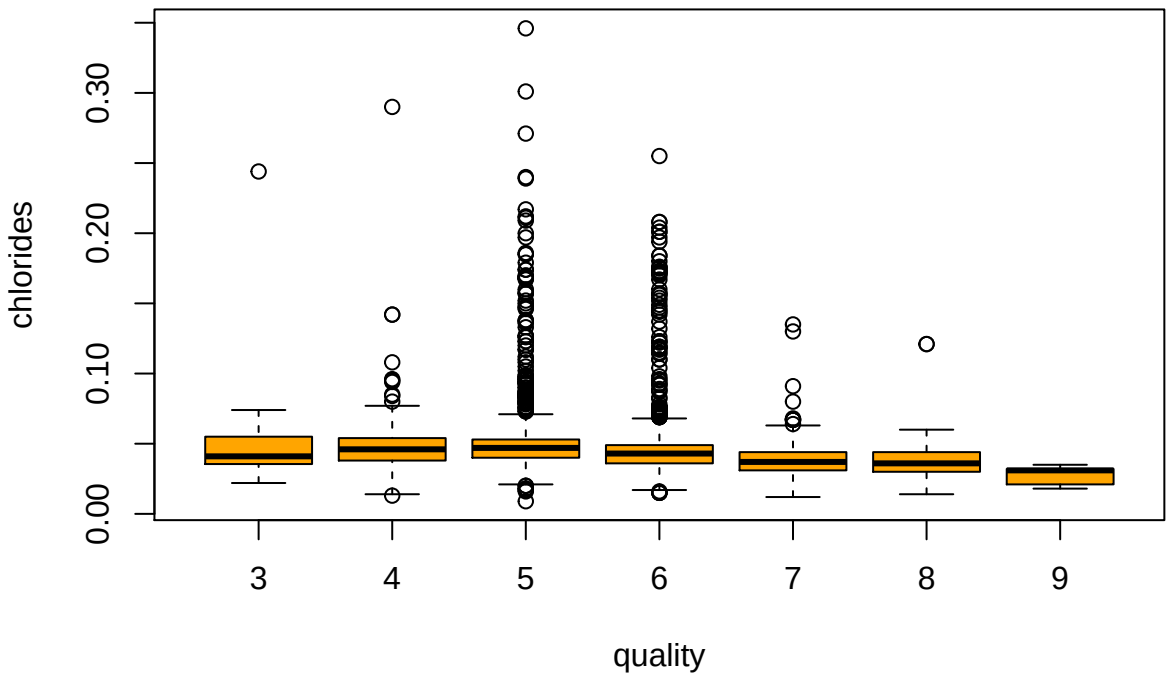
A simple linear model between wine quality, alcohol and density, with adjusted R^2 of 0.1921.

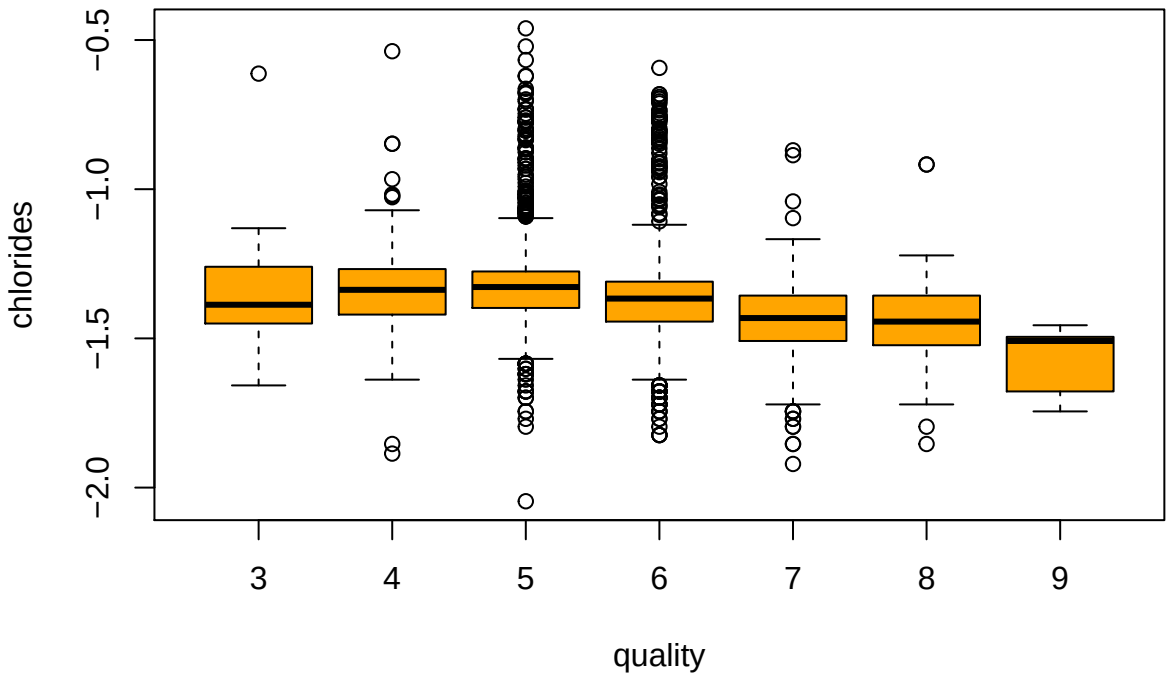
Below are a few other visuliazations I found interesting. I explored relationships between other variables that might not affect quality directly. For example, those with high correlations to alcohol and density.



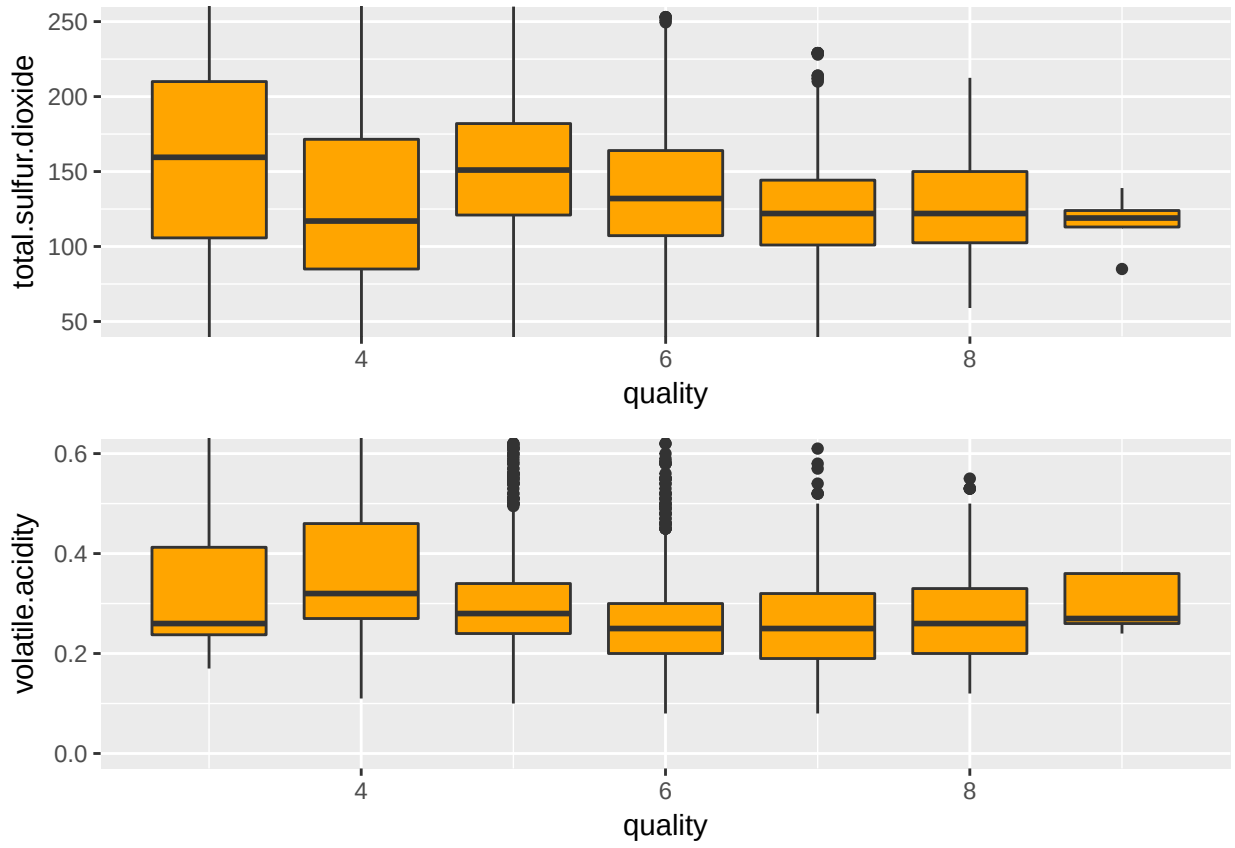


Residual sugar box plots with and without scale transformation.


chlorides
0.00
0.10
0.20
0.30
3
4
5
6
7
8
9
quality



The boxplots of chlorides vs quality show a lot of values outside the quartiles. This still true even after transforming the scale using log10.



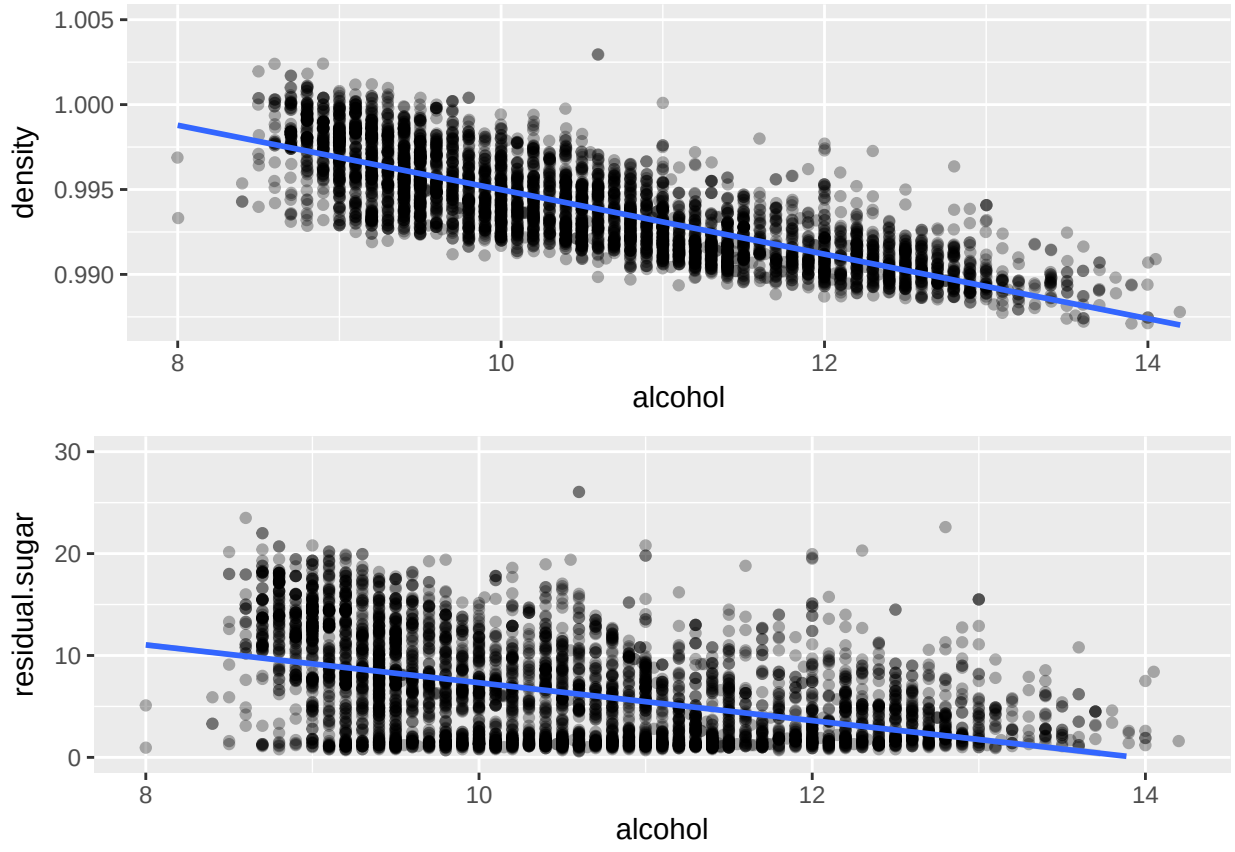

```
##
## Calls:
## m1: lm(formula = quality ~ alcohol, data = wines)
## m2: lm(formula = quality ~ alcohol + volatile.acidity, data = wines)
## m3: lm(formula = quality ~ alcohol + volatile.acidity + chlorides,
##     data = wines)
## m4: lm(formula = quality ~ alcohol + volatile.acidity + chlorides +
##     total.sulfur.dioxide, data = wines)
## m5: lm(formula = quality ~ alcohol + volatile.acidity + chlorides +
##     total.sulfur.dioxide + residual.sugar, data = wines)
##
## ================================================================================================
##                              m1           m2           m3           m4           m5
## ------------------------------------------------------------------------------------------------
##   (Intercept)              2.582***     3.017***     3.179***     2.811***     2.334***
##                           (0.098)      (0.098)      (0.114)      (0.141)      (0.148)
##   alcohol                  0.313***     0.324***     0.315***     0.335***     0.374***
##                           (0.009)      (0.009)      (0.010)      (0.011)      (0.011)
##   volatile.acidity                      -1.979***    -1.948***    -2.012***    -2.108***
##                                         (0.110)      (0.110)      (0.111)      (0.110)
##   chlorides                                          -1.491**     -1.566**     -0.966
##                                                      (0.544)      (0.543)      (0.541)
##   total.sulfur.dioxide                                             0.001***     0.001
##                                                                   (0.000)      (0.000)
##   residual.sugar                                                                0.025***
##                                                                                (0.003)
```

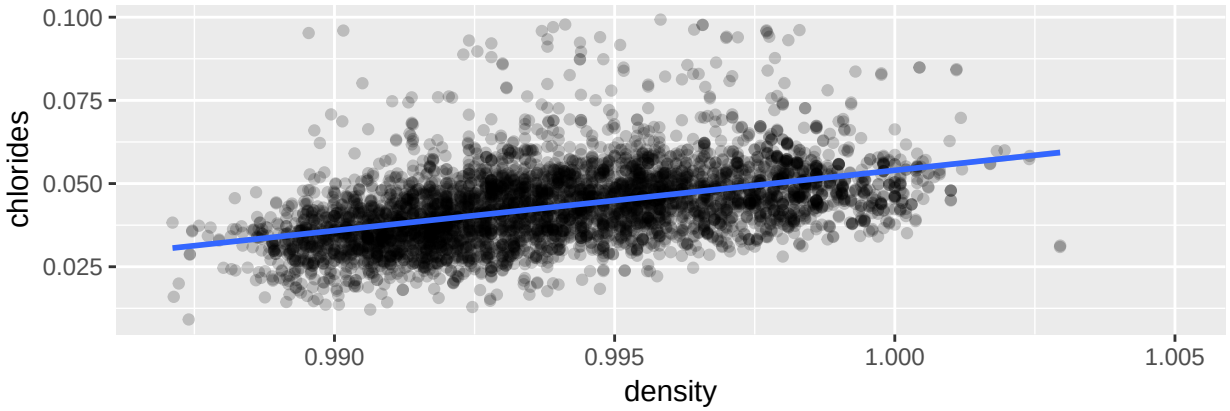
```
## ---------------------------------------------------------------------------------------------
##   R-squared                0.190         0.240         0.241         0.244         0.260
##   adj. R-squared           0.190         0.240         0.241         0.244         0.259
##   sigma                    0.797         0.772         0.772         0.770         0.762
##   F                     1146.395       773.875       519.107       395.710       342.876
##   p                        0.000         0.000         0.000         0.000         0.000
##   Log-likelihood       -5839.391     -5681.776     -5678.019     -5668.229     -5618.852
##   Deviance              3112.257      2918.264      2913.791      2902.166      2844.239
##   AIC                  11684.782     11371.552     11366.039     11348.458     11251.704
##   BIC                  11704.272     11397.538     11398.522     11387.438     11297.180
##   N                     4898          4898          4898          4898          4898
## =============================================================================================
```

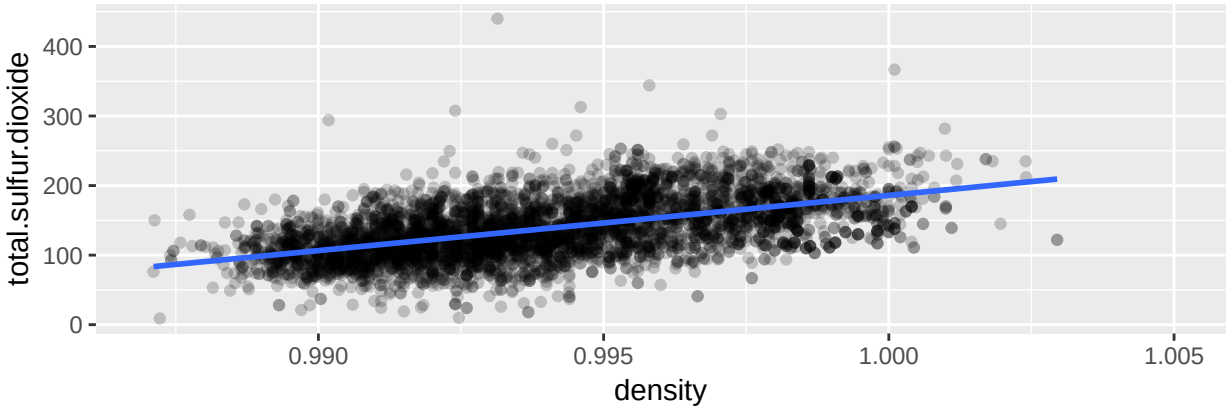
Linear model with more attributes increases adjusted R^2 only marginally from 0.190 to 0.260.

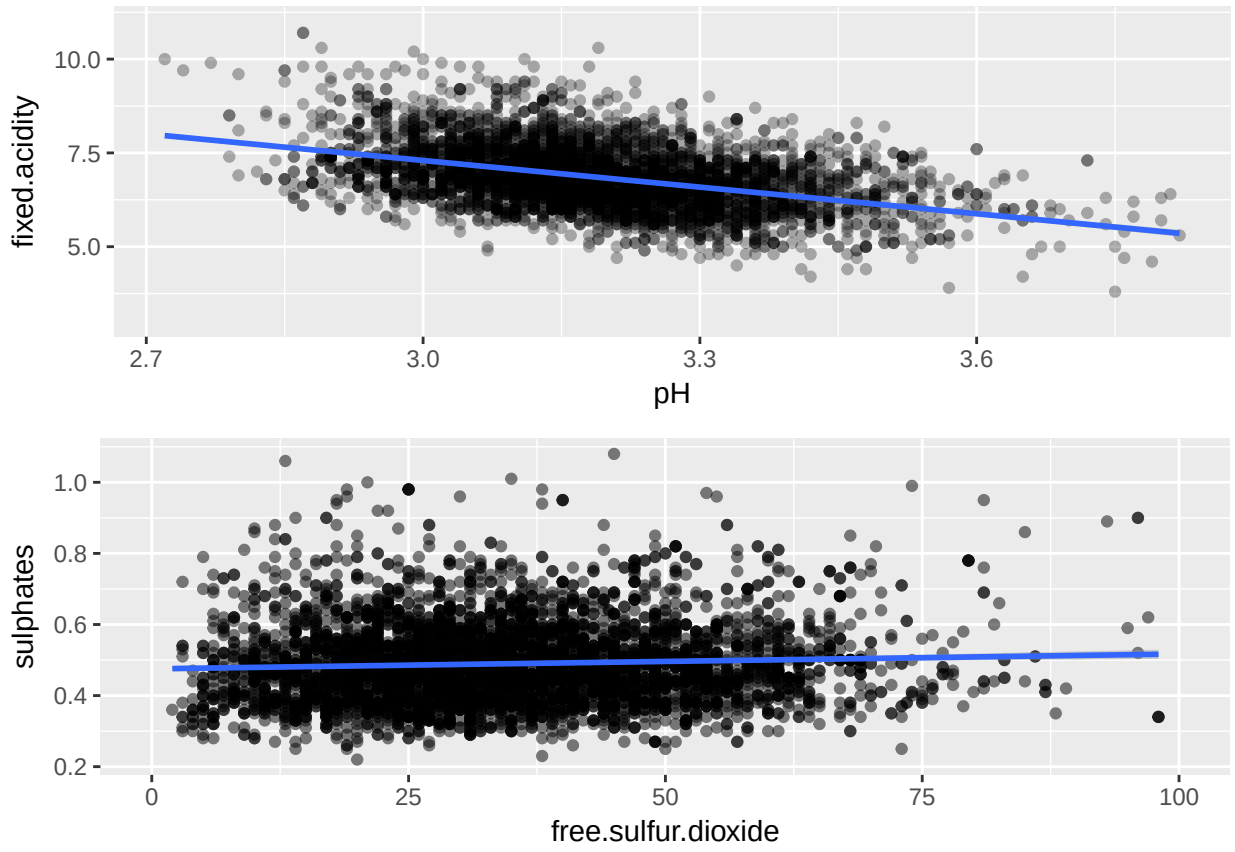
```
##              X fixed.acidity volatile.acidity citric.acid residual.sugar
## [1,] 0.2136562    -0.1208811       0.06771794 -0.07572873     -0.4506312
##       chlorides free.sulfur.dioxide total.sulfur.dioxide    density
## [1,] -0.3601887          -0.2501039           -0.4488921 -0.7801376
##             pH   sulphates alcohol   quality
## [1,] 0.1214321 -0.01743277       1 0.4355747
```

```
##               X fixed.acidity volatile.acidity citric.acid residual.sugar
## [1,] -0.1859761      0.265331       0.02711385   0.1495026      0.8389665
##      chlorides free.sulfur.dioxide total.sulfur.dioxide density
## [1,] 0.2572113           0.2942104            0.5298813       1
##                pH  sulphates    alcohol    quality
## [1,] -0.09359149 0.07449315 -0.7801376 -0.3071233
```




0.100
0.075
0.050
0.025
chlorides
0.990
0.995
1.000
1.005
density


400
300
200
100
0
total.sulfur.dioxide
0.990
0.995
1.000
1.005
density



The correlation statistics and the plots show that both density and residual sugar have high negative correlations with alcohol while chlorides and total sulfur dioxide are positively correlated with density.

Last plot showed a weak correlation between fixed acidity and pH, and between sulphates and free sulphur dioxide.

Bivariate Analysis

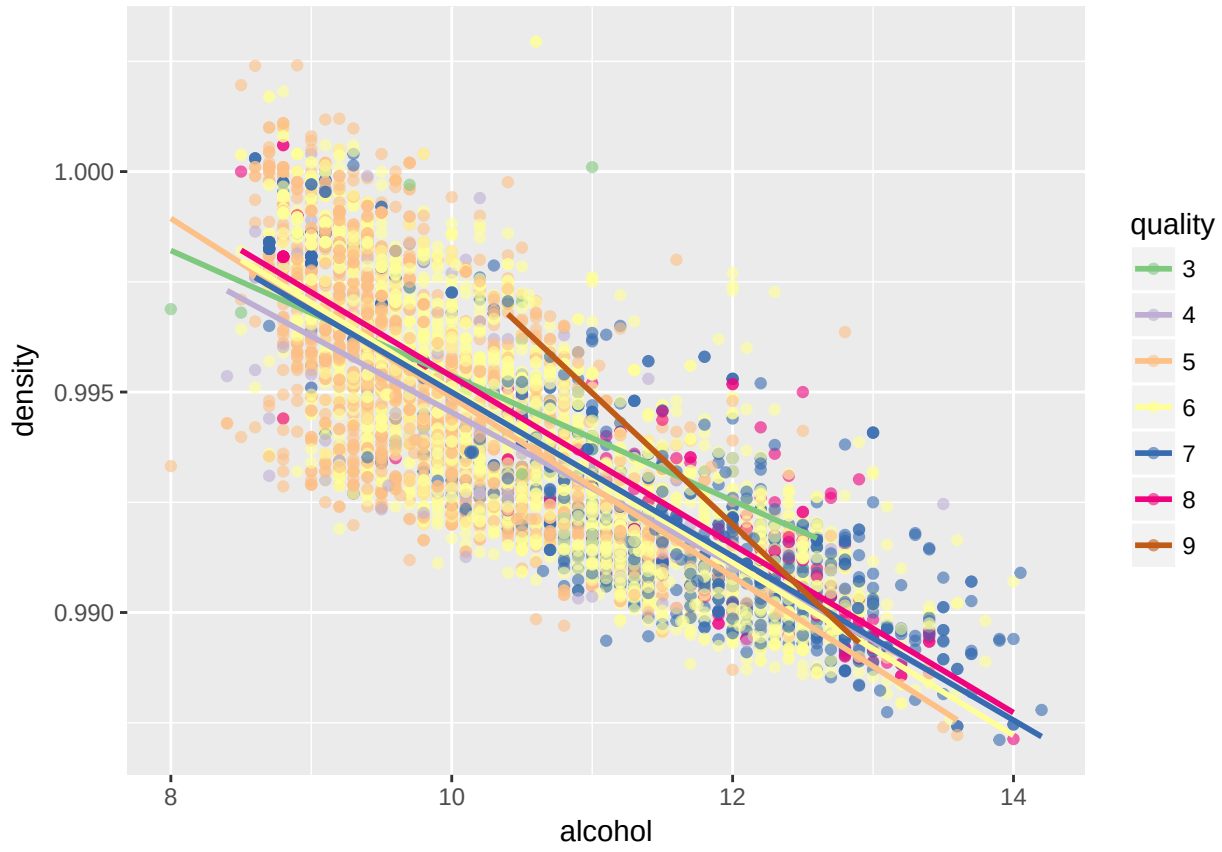
No single attribute correlated strongly with wine quality. However, the plots and statistics show a weak positive relationship between alcohol and quality and weak negative relationsip between density and alcohol with correlation coeffecients of 0.440 and -0.348 respectively.

The linear model shows alcohol and density explain a little of the variation in quality, though not a lot as adjusted R^2 is only 0.192.This means that other attributes besides alcohol and density affect wine quality. Linear model increases adjusted R^2 only marginally from 0.192 to 0.260 by including other attributs such as volatile acidity, chlorides, residual sugar and total sulfur dioxide.

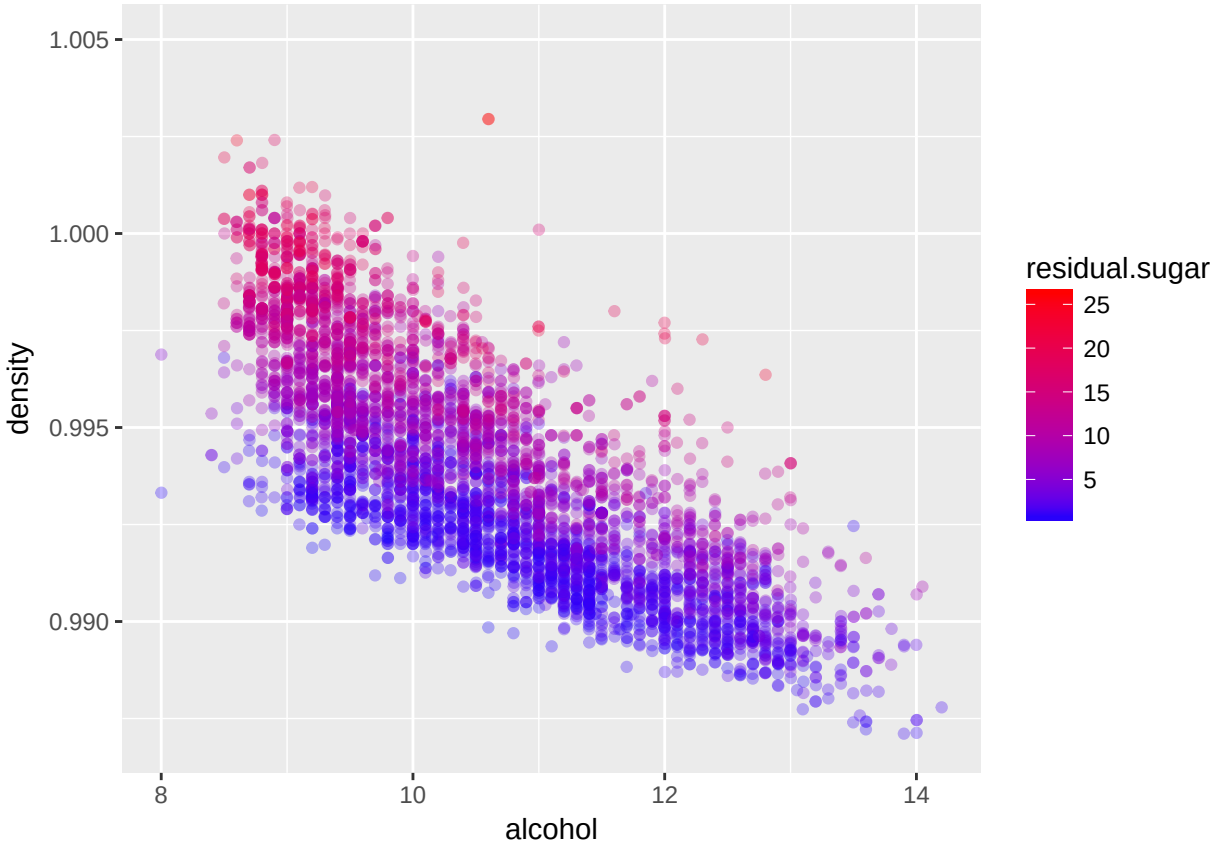
Other strong relationships were between alcohol and density with r = -0.780, meaning the higher the alcohol content the lower the density of the wine. Density was also highly correlated with residual sugar with r = 0.839.

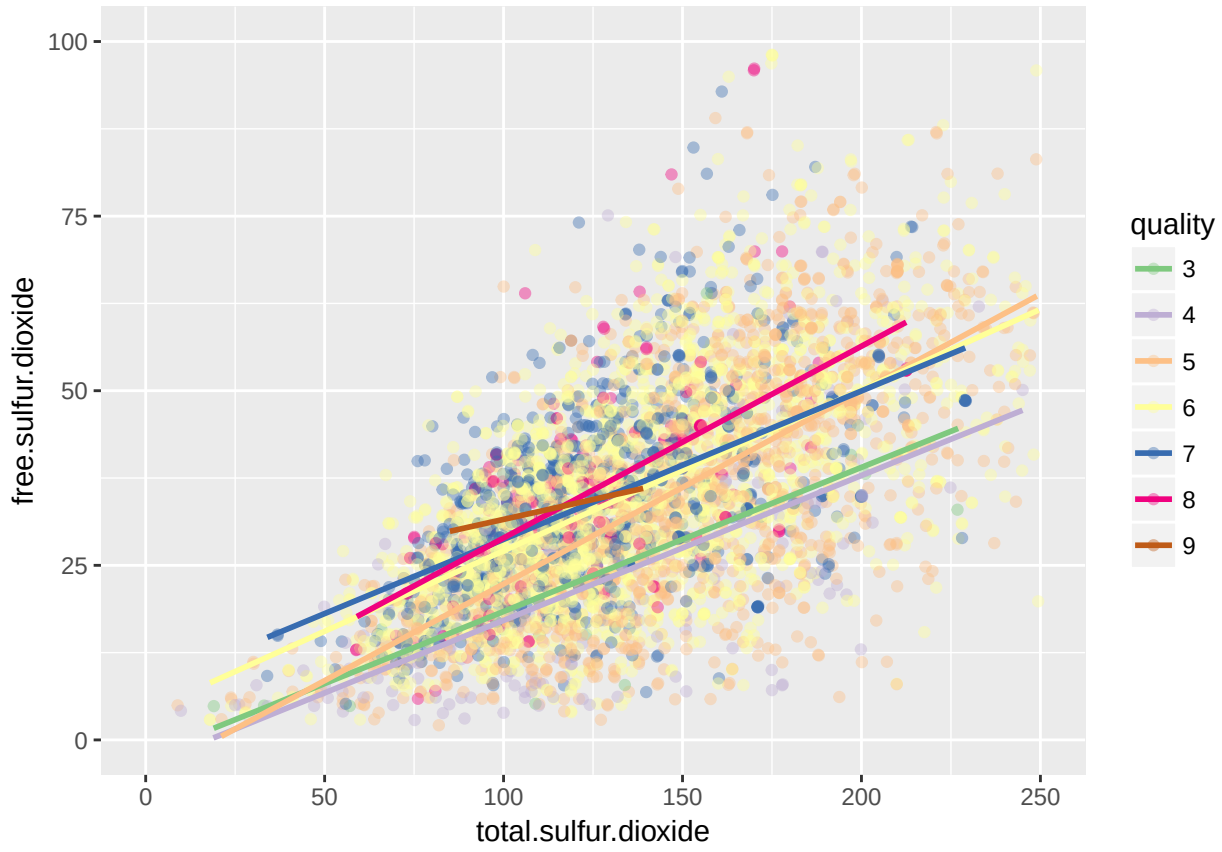
Multivariate Analysis and Plots

In this section I analysed relationships among at least 3 variables for each plot.



The scatter plot shows the very high correlation between density and alcohol and is colored using quality as the third variable.

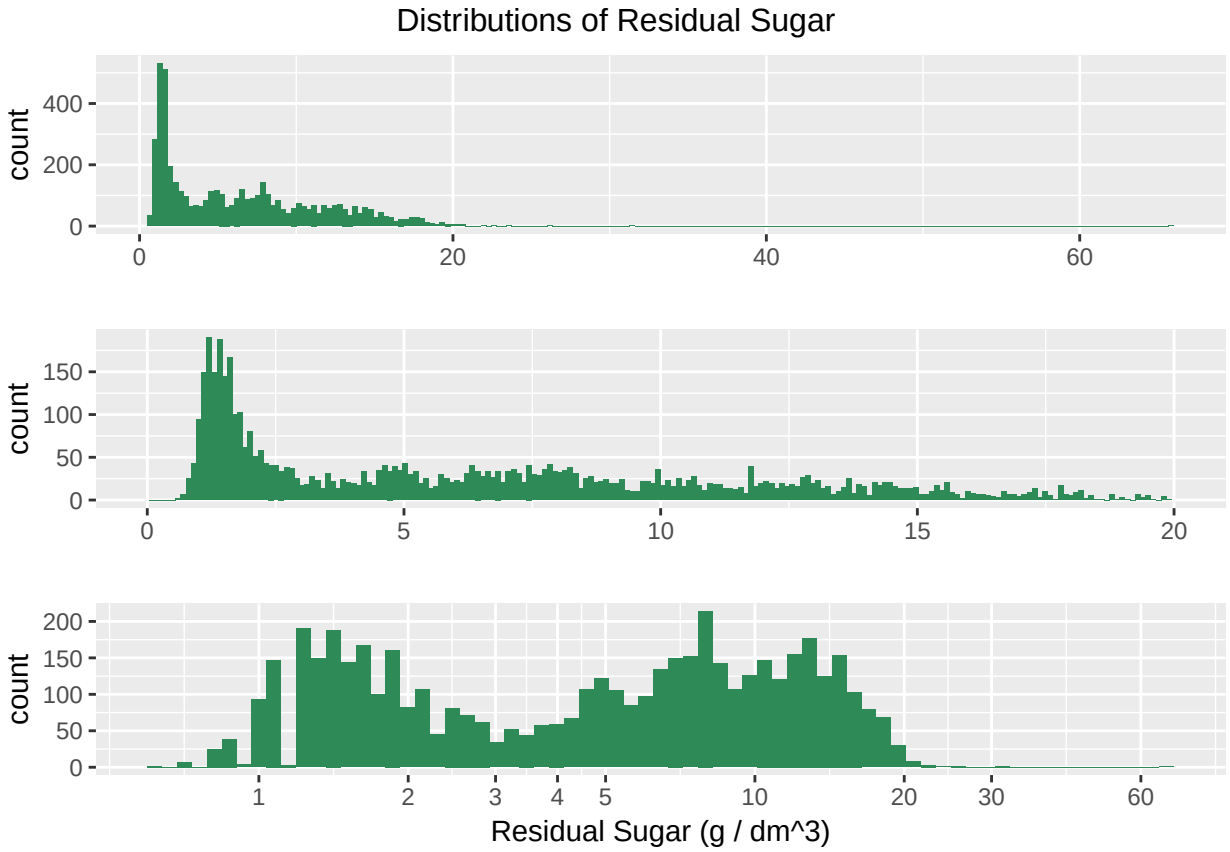

1.005
1.000
0.995
0.990
density
8
10
12
14
alcohol
residual.sugar
25
20
15
10
5



Altough not very clear, quality seems to be weakly correlated with free sulfur dioxide as seen by higher quality colors being higer in the y scale.

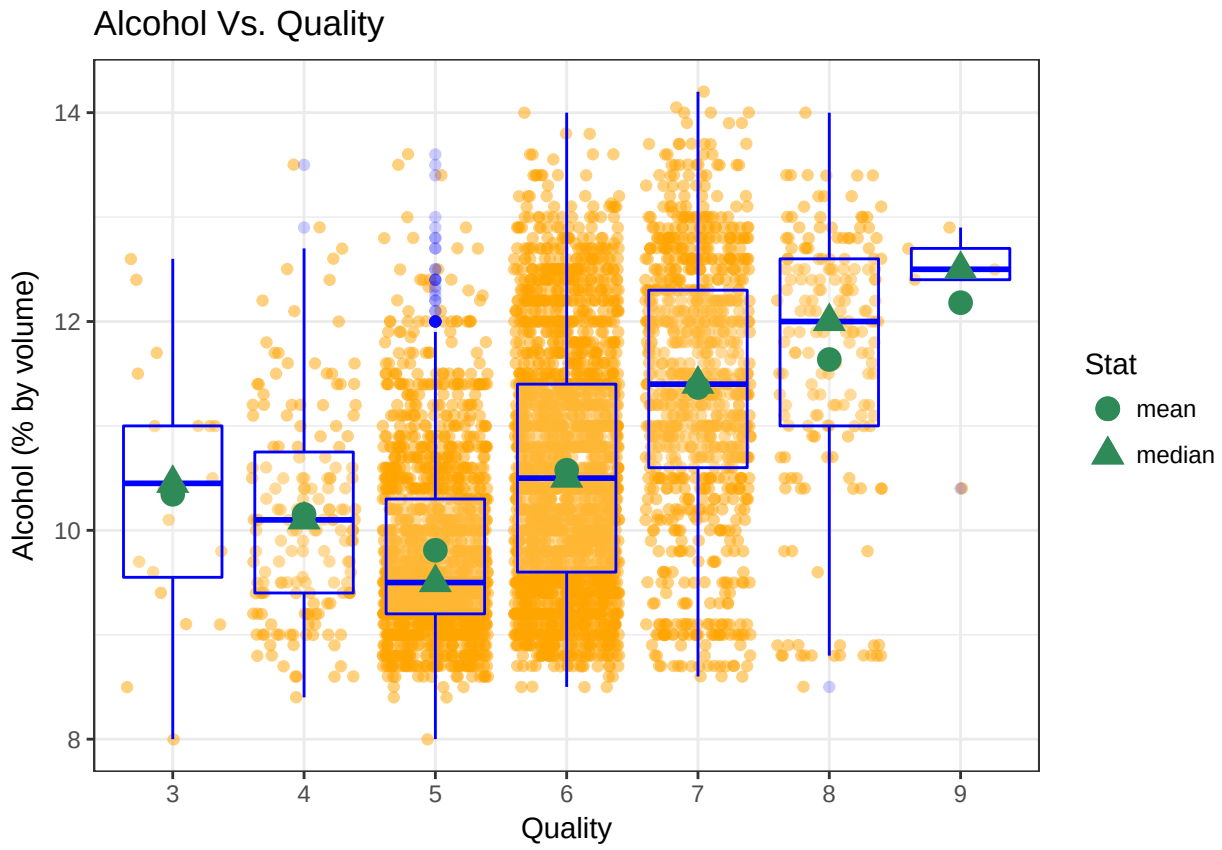
Final Plots and Summary

Plot 1



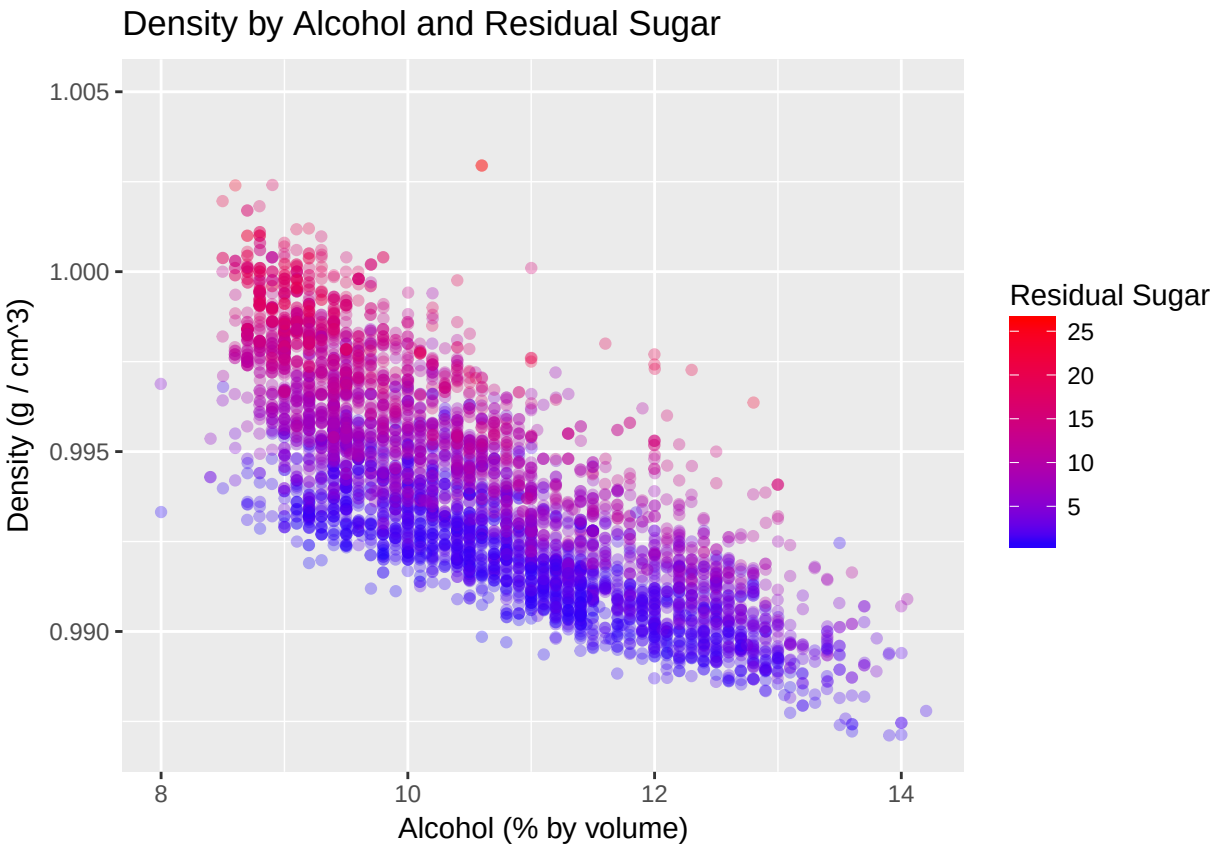
I chose this plot to help illustrare the distribution of residual sugar values. The top plot shows all the data including exterme outliers while the middle plot limits the scale to eliminate the outliers. Even with ouliers removed, the data distribution was still skewed, meaning there is a large number of values on the far right of the center, thus the need for scale transformation. The bottom plot shows log transformed bimodal distribution of residual sugars. The log transform also explains the x scale in the plot which looks uneven.

Plot 2



The above is a scatter plot of alcohol vs quality overlaid with boxplots for each quality rating. I aslo included the mean and median alcohol for each group shown by the marks in the boxplots. Mean and median alcohol increase across quality, although there is a break in this trend for quality rating of 5. Another observation is that the mean and median are close and within the interquartile range for all quality ratings except for quality rating 9. This might be due to very few data points for this group as only 5 wines have a quality rating of 9.

Plot 3



This plot illustrates the strong correlation between density and alcohol, and density and residual sugar as shown by the very well defined colors as density increases. The general trend of the scatter plot indicates the negative relationship between alcohol and density and the color scheme shows relationship between density and residual sugars. According to Khan Academy, fermentation is a process in which sugars are turned to alcohol by yeast. Thus the negative relationship between residual sugars and alcohol for the wines in this dataset can be partlially explained by fermentation. Higher alcohol levels mean more sugars were turned into alcohol, thus less residual sugars. Ths could also mean that the alcohol in the wines is less dense than the sugars and hence more alcohol and less sugars means lower density.

Reflections

The data set is composed of 12 objective variables and one output variable, quality based on sensory data (median of at least 3 evaluations made by wine experts) on a scale 0f 0 to 10. I wanted to explored the data set and determine which of the attributes contribute most to quality raitings and by how much. I found out that alcohol (% by volume) is the best predictor of quality but not a good predictor as the linear model showed that alcohol explained about 19% of the variation in quality. Other attributes contributed marginally to the output variable. This could have been because the output variable is ordinal and a linear model is not a good fit for ordinal data. There are definitely other variables that affect quality that may not be in the data such as age of wine and price. Also, ratings by humans maybe considered subjective as people have different tastes.

One issue with the data is that the majority of wines (92.6%) have quality ratings between 5 and 7. Only five wines with a quality rating of 9 and 0 wines with a quality rating of 0, 1, 2 and 10.