Wine Quality Exploration by Heather Rauch

This report will explore data about 11 chemical properties and the average rating for 4898 different white wines.

Univariate Plots Section

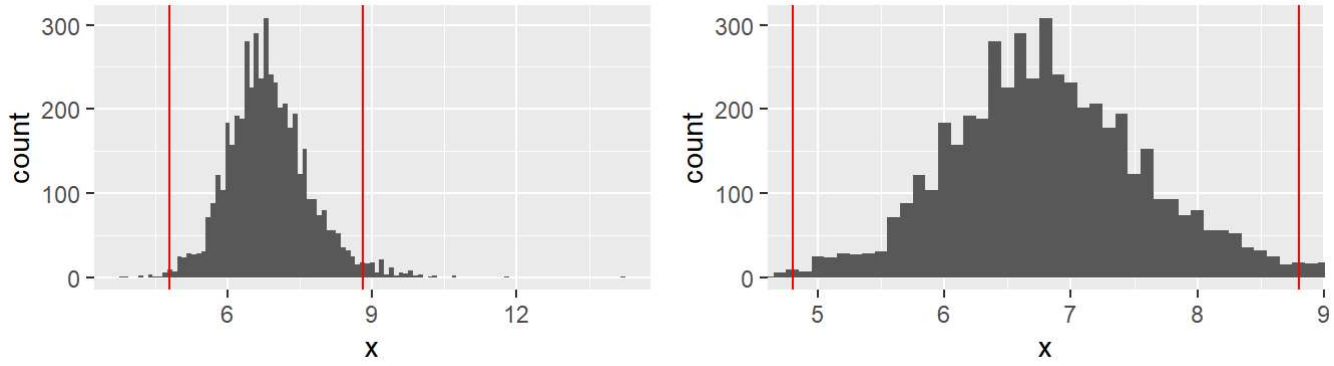
```
## [1] 4898   13
```

```
## '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
##  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 total.sulfur.dioxide
##  Min.   : 0.600   Min.   :0.00900   Min.   :  2.00      Min.   :  9.0
##  1st Qu.: 1.700   1st Qu.:0.03600   1st Qu.: 23.00      1st Qu.:108.0
##  Median : 5.200   Median :0.04300   Median : 34.00      Median :134.0
##  Mean   : 6.391   Mean   :0.04577   Mean   : 35.31      Mean   :138.4
##  3rd Qu.: 9.900   3rd Qu.:0.05000   3rd Qu.: 46.00      3rd Qu.:167.0
##  Max.   :65.800   Max.   :0.34600   Max.   :289.00      Max.   :440.0
##     density             pH          sulphates         alcohol
##  Min.   :0.9871   Min.   :2.720   Min.   :0.2200   Min.   : 8.00
##  1st Qu.:0.9917   1st Qu.:3.090   1st Qu.:0.4100   1st Qu.: 9.50
##  Median :0.9937   Median :3.180   Median :0.4700   Median :10.40
##  Mean   :0.9940   Mean   :3.188   Mean   :0.4898   Mean   :10.51
##  3rd Qu.:0.9961   3rd Qu.:3.280   3rd Qu.:0.5500   3rd Qu.:11.40
##  Max.   :1.0390   Max.   :3.820   Max.   :1.0800   Max.   :14.20
##     quality
##  Min.   :3.000
##  1st Qu.:5.000
##  Median :6.000
##  Mean   :5.878
##  3rd Qu.:6.000
##  Max.   :9.000
```

The dataset consists of 13 variables with 4898 observations.

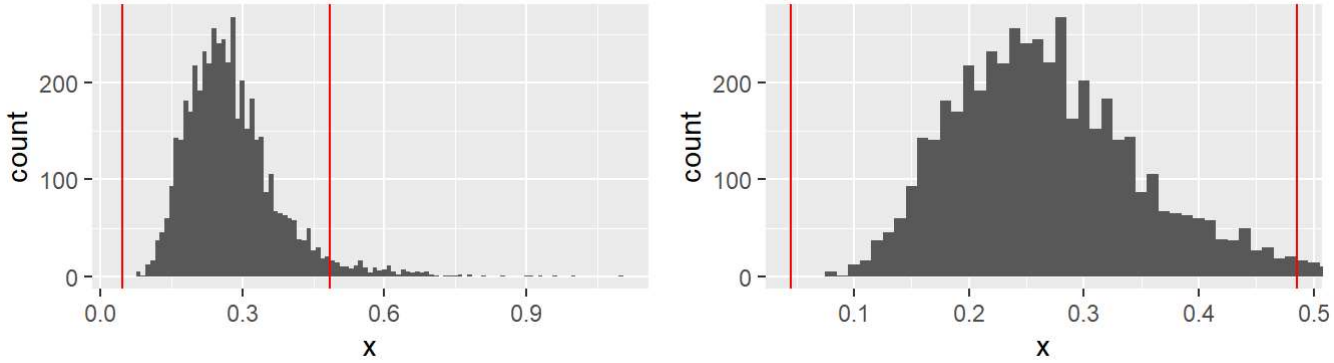
fixed.acidity




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.800   6.300   6.800   6.855   7.300  14.200
```

The fixed acidity distribution is slightly skewed in the positive direction. The red lines were added to detect statistical outliers with a lower limit of 4.8 and an upper limit of 8.8. When zoomed to exclude outliers, the distribution looks much more normal with the majority of values falling between about 6.3 and 7.3.

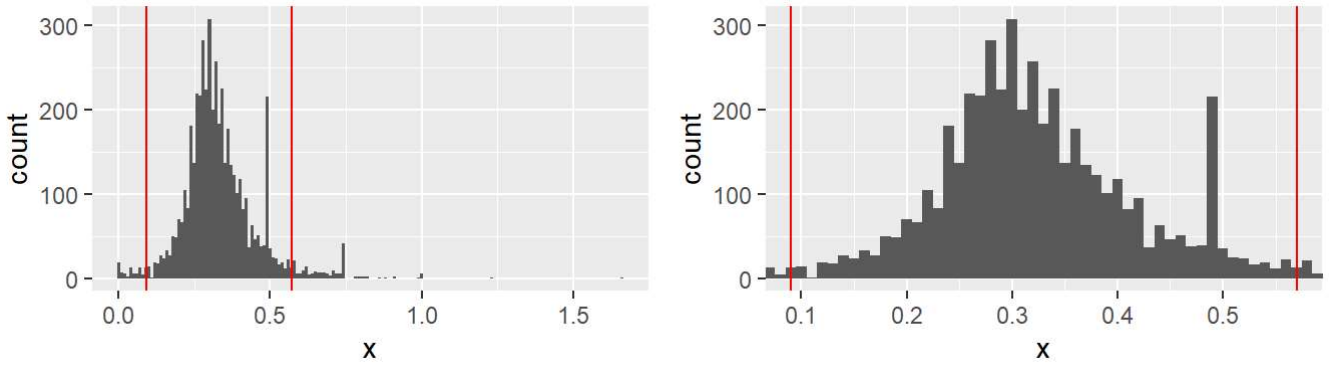
volatile.acidity




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0800  0.2100  0.2600  0.2782  0.3200  1.1000
```

The volatile.acidity distribution is far more skewed in the positive direction than the fixed.acidity distribution. The majority of values appear to fall between about 0.21 and 0.32.

citric.acid



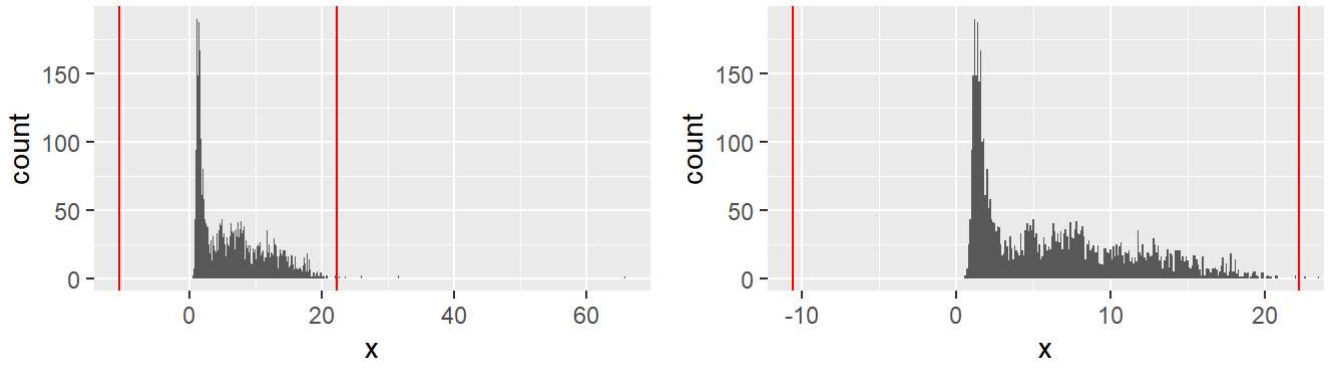

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.2700  0.3200  0.3342  0.3900  1.6600
```

```
##        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 total.sulfur.dioxide
##  Min.   : 0.600   Min.   :0.00900   Min.   :  2.00      Min.   :  9.0
##  1st Qu.: 1.700   1st Qu.:0.03600   1st Qu.: 23.00      1st Qu.:108.0
##  Median : 5.200   Median :0.04300   Median : 34.00      Median :134.0
##  Mean   : 6.391   Mean   :0.04577   Mean   : 35.31      Mean   :138.4
##  3rd Qu.: 9.900   3rd Qu.:0.05000   3rd Qu.: 46.00      3rd Qu.:167.0
##  Max.   :65.800   Max.   :0.34600   Max.   :289.00      Max.   :440.0
##     density             pH          sulphates         alcohol
##  Min.   :0.9871   Min.   :2.720   Min.   :0.2200   Min.   : 8.00
##  1st Qu.:0.9917   1st Qu.:3.090   1st Qu.:0.4100   1st Qu.: 9.50
##  Median :0.9937   Median :3.180   Median :0.4700   Median :10.40
##  Mean   :0.9940   Mean   :3.188   Mean   :0.4898   Mean   :10.51
##  3rd Qu.:0.9961   3rd Qu.:3.280   3rd Qu.:0.5500   3rd Qu.:11.40
##  Max.   :1.0390   Max.   :3.820   Max.   :1.0800   Max.   :14.20
##     quality
##  Min.   :3.000
##  1st Qu.:5.000
##  Median :6.000
##  Mean   :5.878
##  3rd Qu.:6.000
##  Max.   :9.000
```

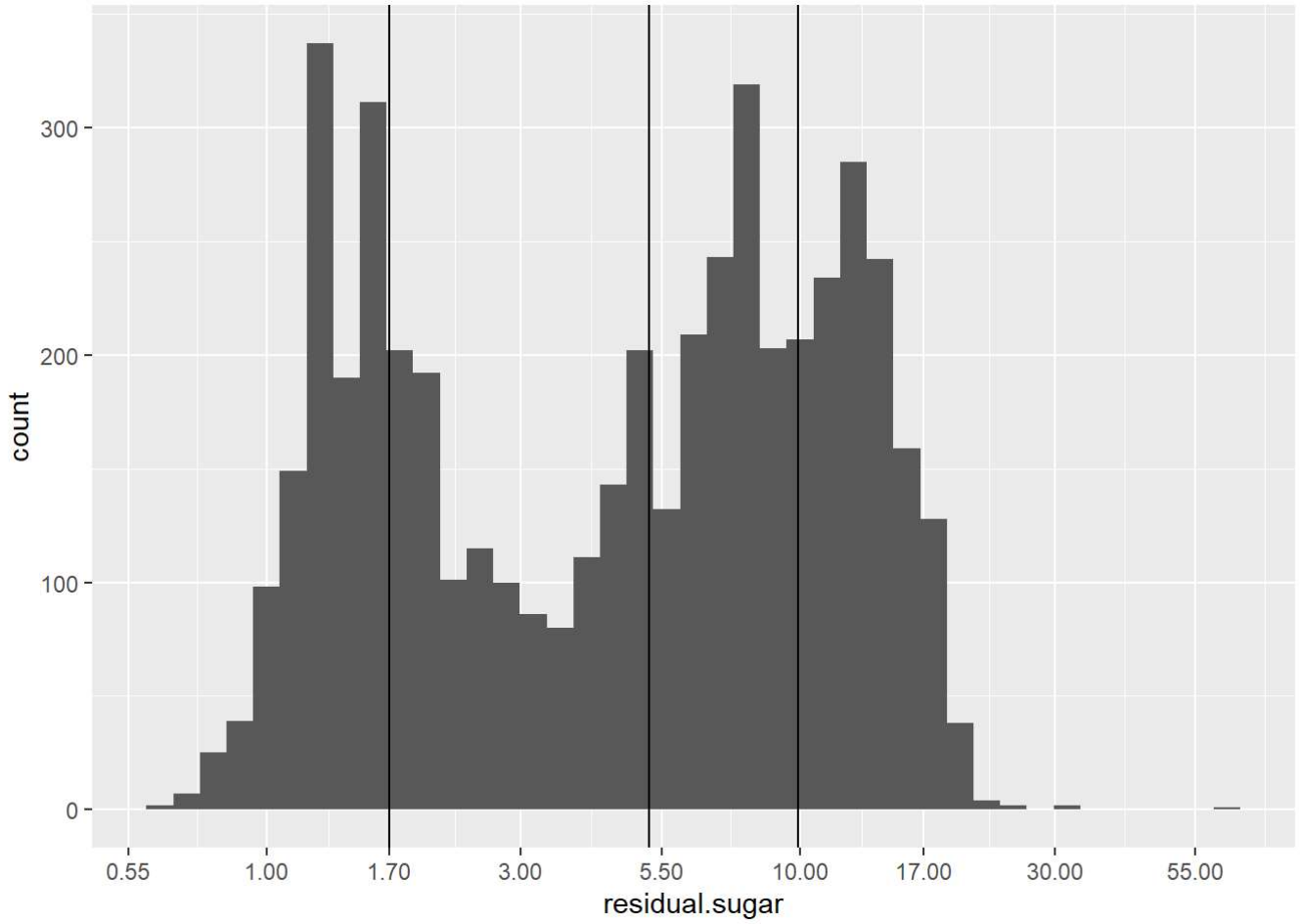
```
##        X        fixed.acidity    volatile.acidity  citric.acid
##  Min.   : 280   Min.   : 5.600   Min.   :0.0800   Min.   :0.49
##  1st Qu.:1478   1st Qu.: 6.800   1st Qu.:0.2000   1st Qu.:0.49
##  Median :1554   Median : 7.400   Median :0.2500   Median :0.49
##  Mean   :1710   Mean   : 7.489   Mean   :0.2629   Mean   :0.49
##  3rd Qu.:1626   3rd Qu.: 8.000   3rd Qu.:0.3000   3rd Qu.:0.49
##  Max.   :4679   Max.   :14.200   Max.   :0.8500   Max.   :0.49
##  residual.sugar     chlorides       free.sulfur.dioxide total.sulfur.dioxide
##  Min.   : 0.900   Min.   :0.02700   Min.   : 3.00       Min.   : 18.0
##  1st Qu.: 1.500   1st Qu.:0.03600   1st Qu.:21.50       1st Qu.:113.5
##  Median : 5.000   Median :0.04400   Median :32.00       Median :138.0
##  Mean   : 5.793   Mean   :0.04558   Mean   :33.61       Mean   :141.2
##  3rd Qu.: 8.100   3rd Qu.:0.05100   3rd Qu.:45.00       3rd Qu.:164.0
##  Max.   :23.500   Max.   :0.23900   Max.   :87.00       Max.   :247.0
##     density             pH          sulphates         alcohol
##  Min.   :0.9893   Min.   :2.850   Min.   :0.2700   Min.   : 8.50
##  1st Qu.:0.9928   1st Qu.:3.065   1st Qu.:0.3750   1st Qu.: 9.70
##  Median :0.9940   Median :3.140   Median :0.4500   Median :10.50
##  Mean   :0.9943   Mean   :3.163   Mean   :0.4623   Mean   :10.48
##  3rd Qu.:0.9956   3rd Qu.:3.240   3rd Qu.:0.5300   3rd Qu.:11.20
##  Max.   :1.0024   Max.   :3.650   Max.   :0.9800   Max.   :13.00
##     quality
##  Min.   :4.000
##  1st Qu.:5.000
##  Median :6.000
##  Mean   :5.893
##  3rd Qu.:6.000
##  Max.   :9.000
```

Like the other acidity attributes, the citric.acid distribution is skewed in the positive direction with the majority of values falling between about 0.27 and 0.39. However, there is a very noticeable spike at 0.49 with over 200 observations. However subsetting this data and comparing the summary to the original does not reveal any major differences.

residual.sugar

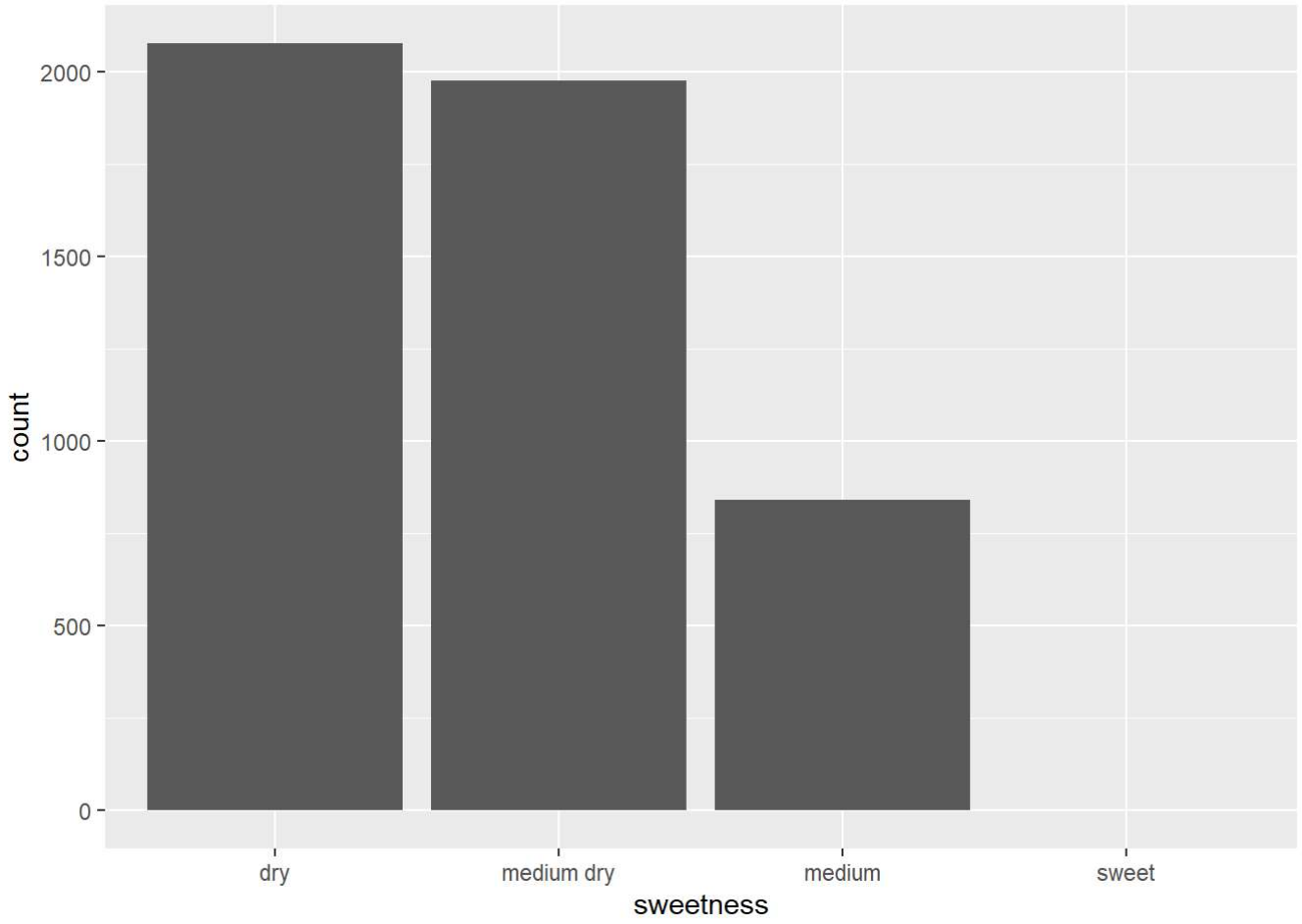


This distribution is most definitely not normal and standard outlier detection techniques did not help. It looks like a more custom approach needs to be taken.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.600   1.700   5.200   6.391   9.900  65.800
```

Viewing the same data on a log10 transformed scale allows us to get a better view of the distribution. Additionally, there are three lines visible on the plot. They are, in order from left to right, quartile 1, median, and quartile 3. The transformed distribution appears bimodal with peaks near quartiles 1 and 3. I wonder how this would look by sweetness category.

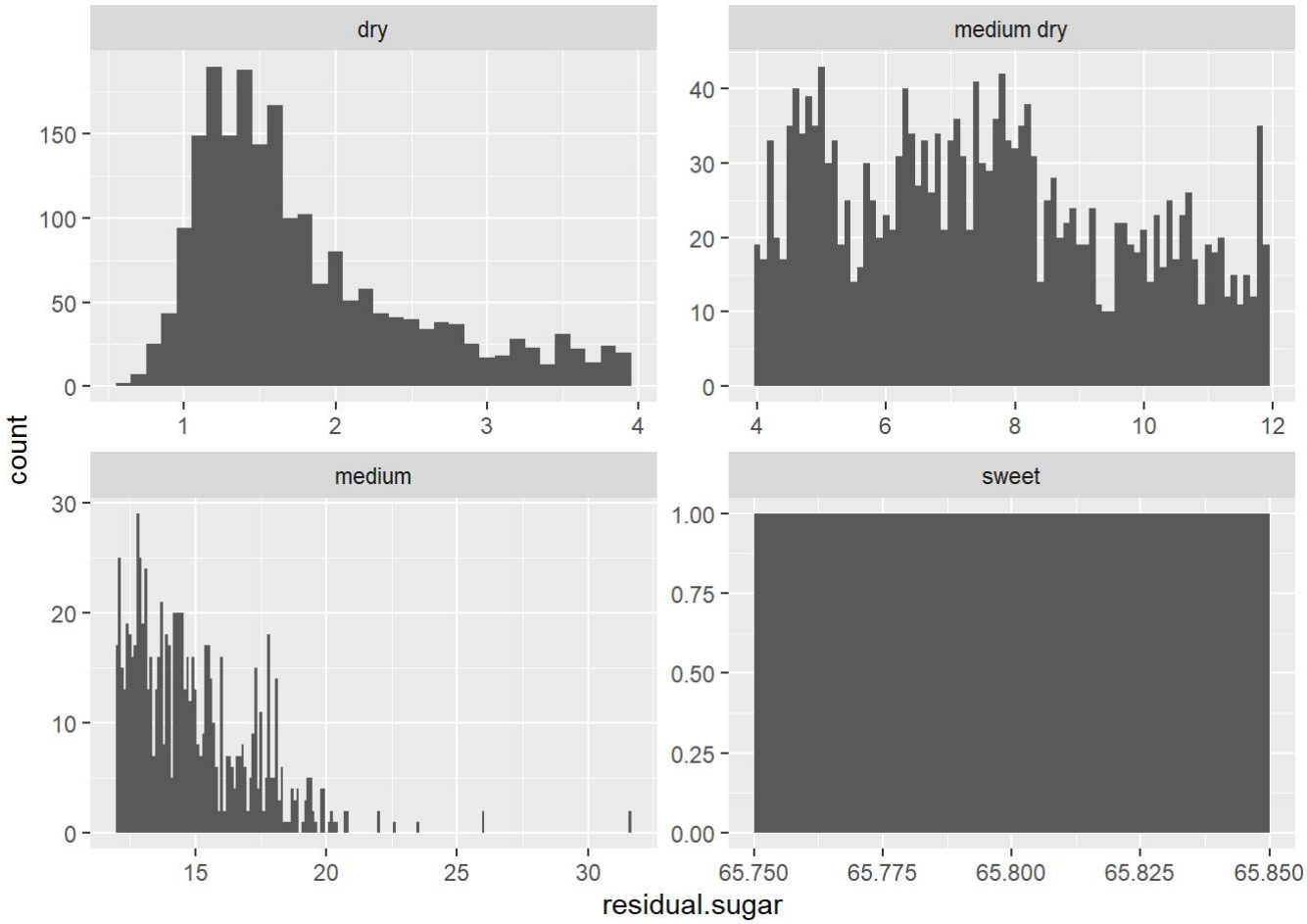


The European Union defines the sweetness terms of wine as follows:

- Dry < 4 g/L of sugar
- Medium dry 4-12 g/L of sugar
- Medium 12-45 g/L of sugar
- Sweet > 45 g/L of sugar

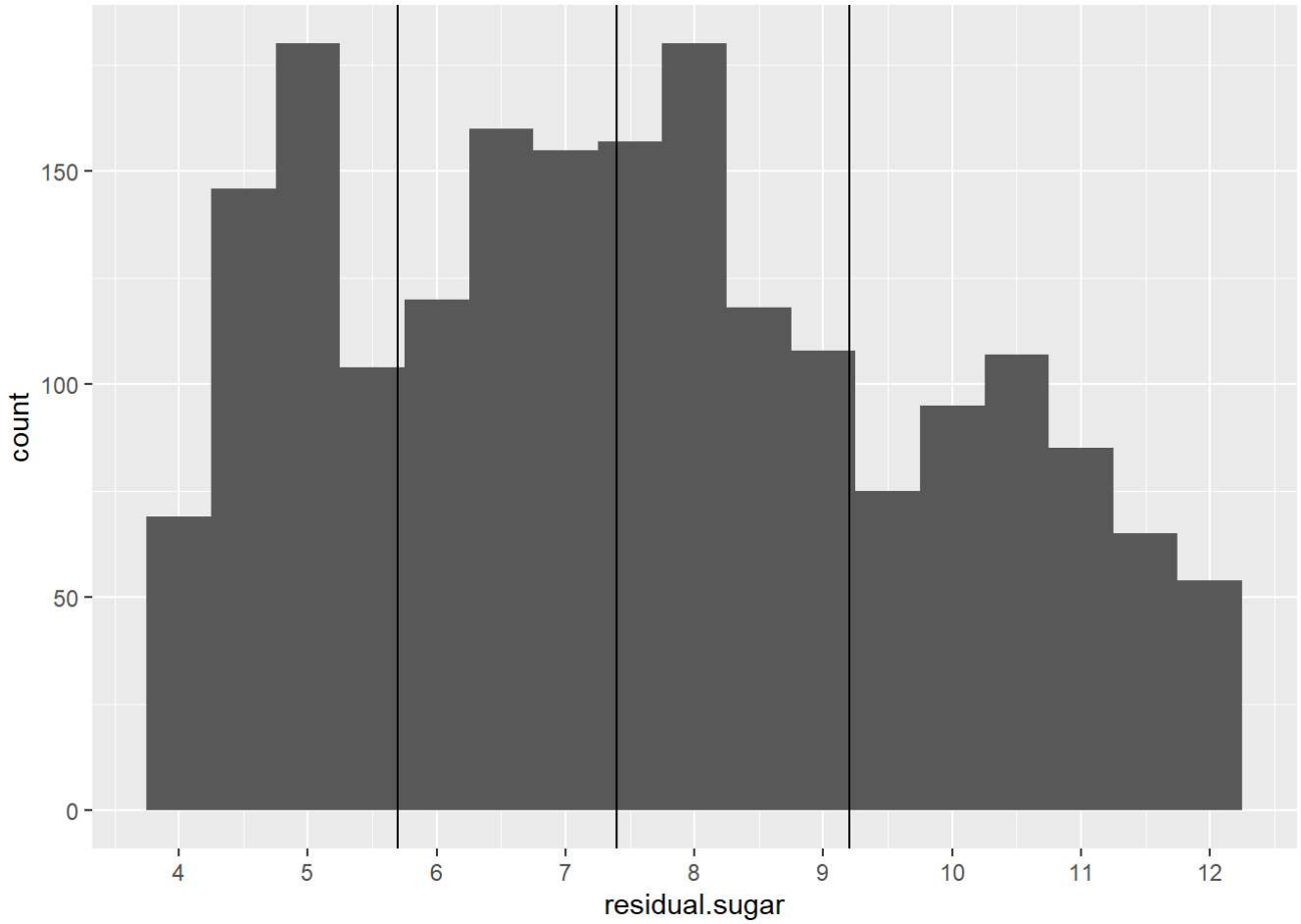
Official Journal of the European Union (https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2006:384:0038:0043:EN:PDF)

The majority of the wines in this dataset are either dry or medium dry. There is only a single observation in the sweet category.



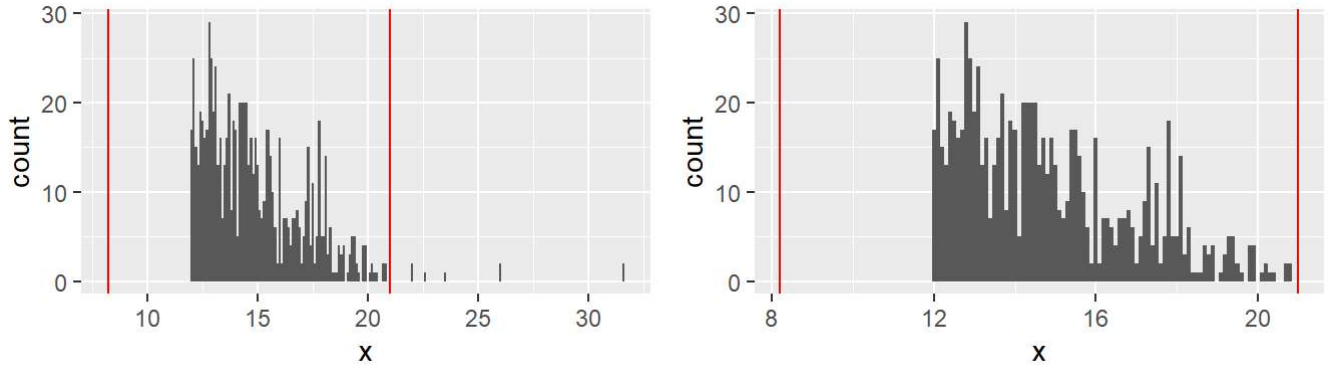

```
## wine$sweetness: dry
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.600   1.300   1.600   1.803   2.200   3.950
## ------------------------------------------------------------
## wine$sweetness: medium dry
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.00    5.70    7.40    7.57    9.20   11.95
## ------------------------------------------------------------
## wine$sweetness: medium
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.00   13.00   14.40   14.89   16.20   31.60
## ------------------------------------------------------------
## wine$sweetness: sweet
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    65.8    65.8    65.8    65.8    65.8    65.8
```

There is only one wine in the sweet grouping, but the dry, medium dry, and medium groups all have much different distributions than just residual.sugar plotted alone. The residual.sugar in dry wines appears to capture the left mode of the initial plot. This distribution is slightly skewed in the positive direction, but is mostly normal. The medium dry and medium plots appear to need a bit more work. I'll look at each individually to see if we can make better sense of it.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.00    5.70    7.40    7.57    9.20   11.95
```

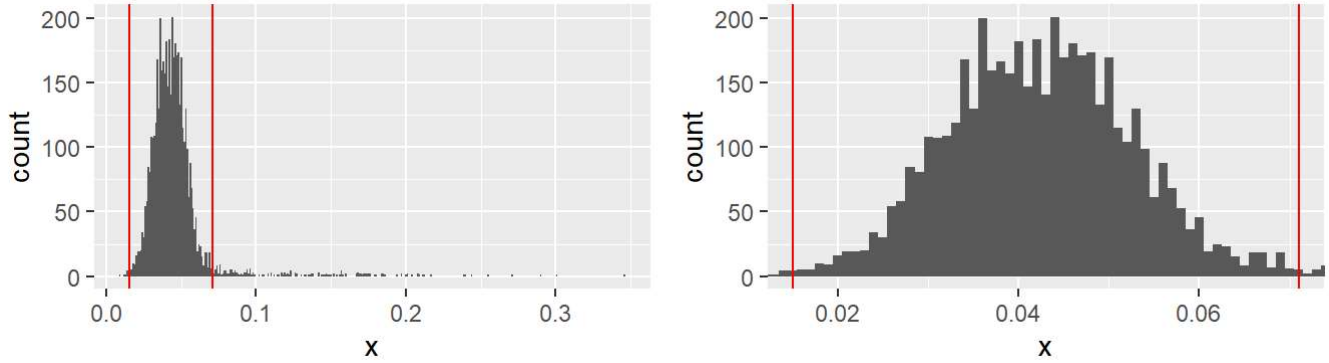
The residual.sugar distribution within the medium dry sweetness group is much more normal than the unfiltered plot, however, it is still positively skewed. The majority of medium dry wines have between 5.7 and 9.2 g/L of residual.sugar.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.00   13.00   14.40   14.89   16.20   31.60
```

The residual.sugar distribution for medium sweetness wines is far more skewed in the positive direction than medium dry wines. When zooming in to view non-outliers, the distribution appears slightly less skewed. The majority of values fall between 13 and 16.2 g/L of residual.sugar.

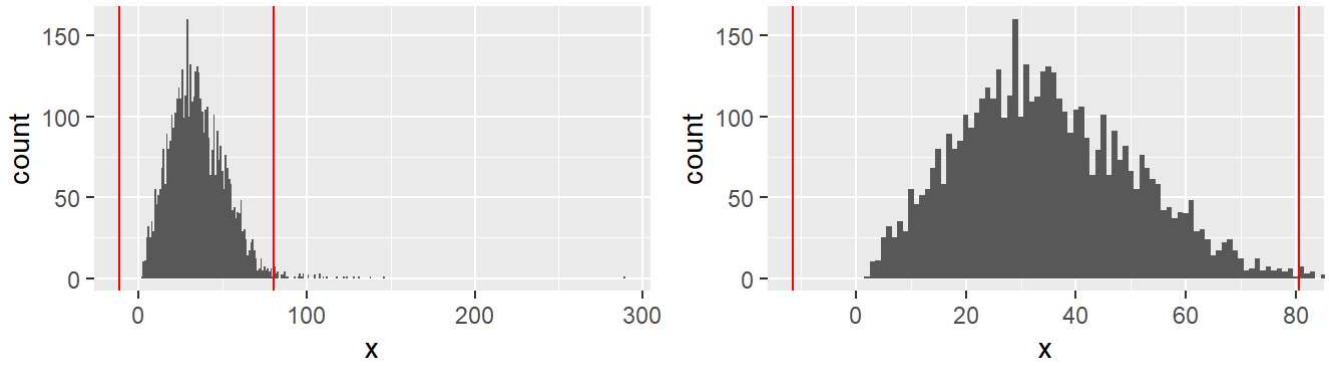
chlorides




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00900 0.03600 0.04300 0.04577 0.05000 0.34600
```

The chloride distribution is heavily skewed in the positive direction. Zooming in on the non-outliers, we see that most values fall between approximately 0.036 and 0.05.

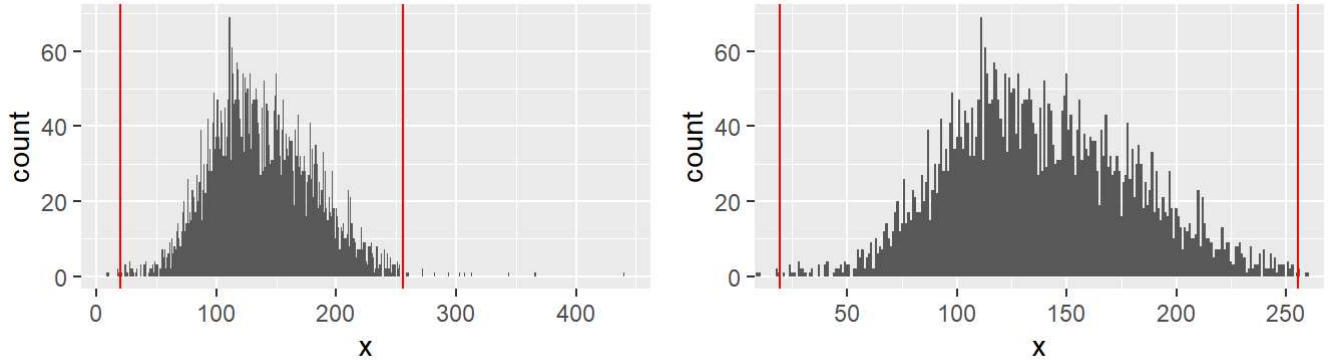
free.sulfur.dioxide




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00   23.00   34.00   35.31   46.00  289.00
```

The distribution of free.sulfur.dioxide is positively skewed by a number of outliers. Aside from the outliers, the distribution is mostly normal, with the majority of values landing between 23 and 46.

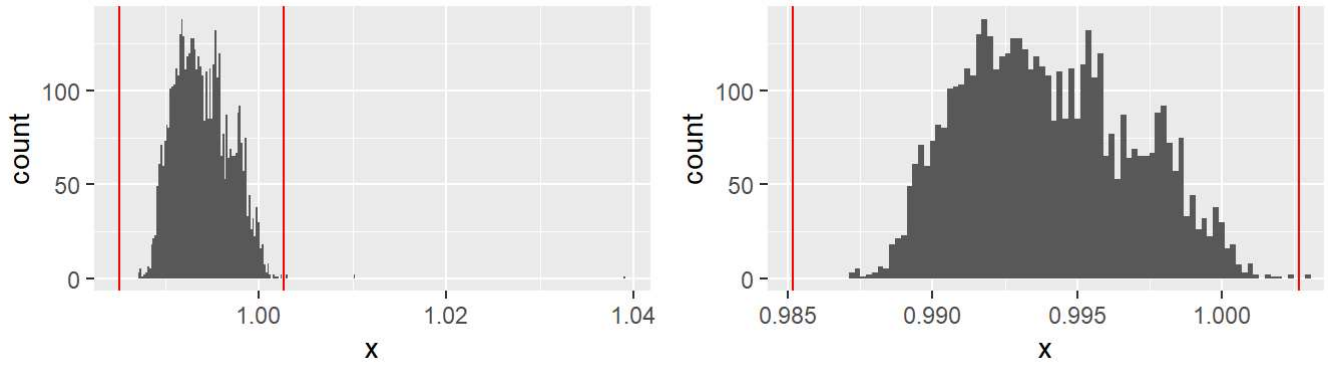
total.sulfur.dioxide




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     9.0   108.0   134.0   138.4   167.0   440.0
```

Once again, aside from the outliers pulling positively, the distribution is normal with most values falling between 108 and 168.

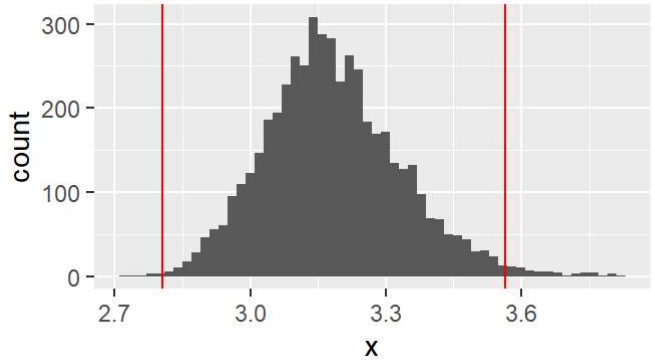
density

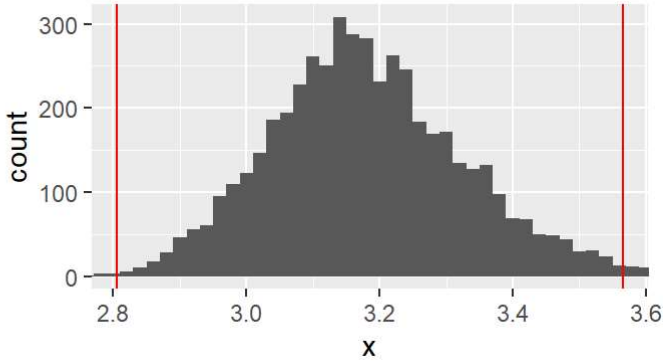



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9871  0.9917  0.9937  0.9940  0.9961  1.0390
```

The density is dependent on the alcohol and sugar content, so that will be investigated later. Other than that, the distribution is positively skewed by outliers. Viewing only the non-outliers, this distribution still isn't quite normal. The right side has the familiar stairstep pattern seen elsewhere.

pH

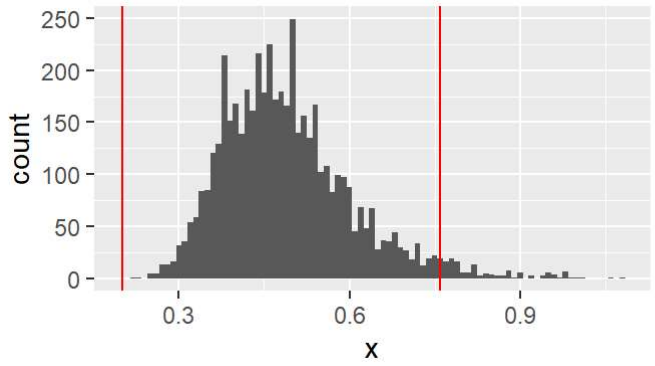


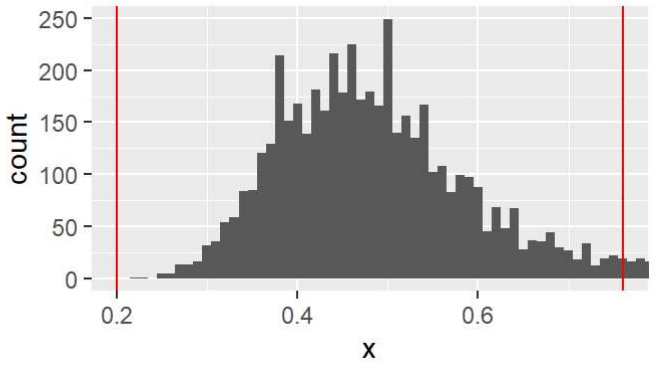



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.720   3.090   3.180   3.188   3.280   3.820
```

The right tail of the pH distribution is a bit longer than the left. Other than that, it is normally distributed with the majority of values between 3.09 and 3.28. a very faint stair-step pattern in seen in this plot on the right side. This same pattern is evident in the sugar and alcohol plots, suggesting a possible relationship.

sulphates

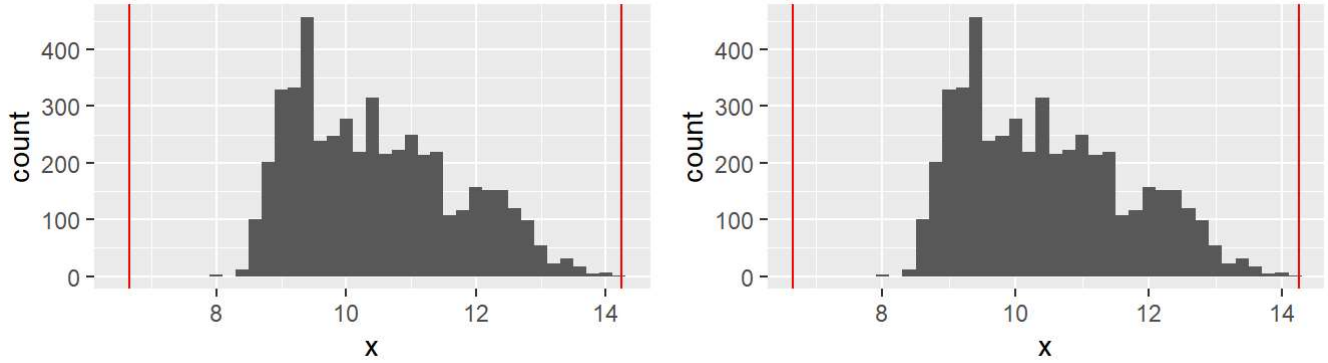





```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2200  0.4100  0.4700  0.4898  0.5500  1.0800
```

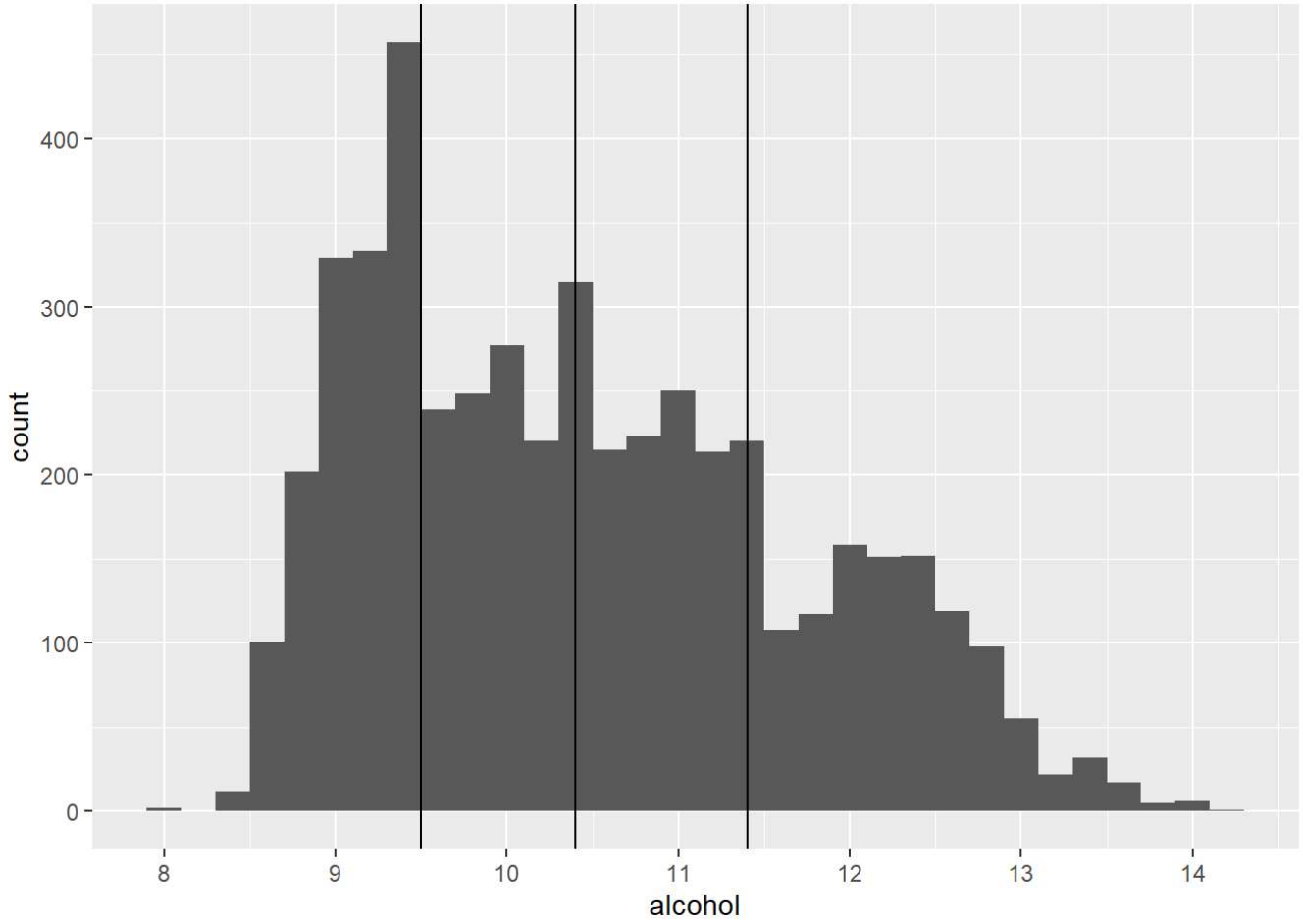
The sulphates distribution is fairly normal with a slight tail on the right consisting mostly of outliers. The majority of values fall between 0.41 and 0.55.

alcohol




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00    9.50   10.40   10.51   11.40   14.20
```

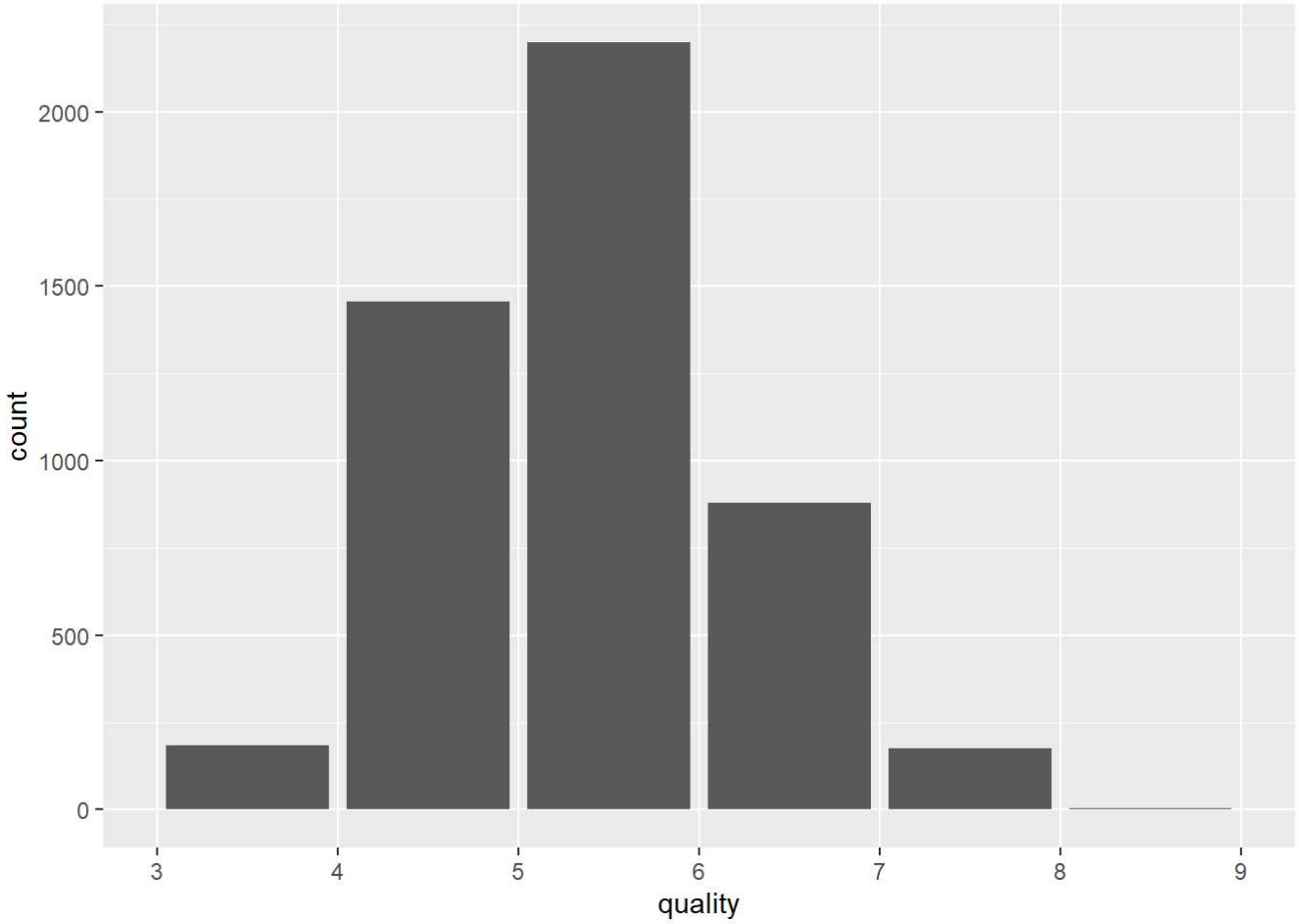
The distribution of alcohol does not contain any outliers. Because of this, the zoomed in plot is nearly identical to the initial plot. This one needs a custom plot to get a better view of the distribution.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00    9.50   10.40   10.51   11.40   14.20
```

The same techniques were used in this plot as previous custom plots, with the 1st quartile, median and 3rd quartile displayed with black lines. The distribution for alcohol has a stair-step pattern like we saw earlier in the density distribution. This makes sense since the density is dependent on alcohol. The majority of values fall between 9.5 and 11.4.

quality




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   5.000   6.000   5.878   6.000   9.000
```

All of the previous variables have been continuous and held information about properties of the wine. This variable, however, is discreet, since it represents the ten point scale on which the wines were rated. The distribution appears to be normal with most values between 5 and 6.

Univariate Analysis

What is the structure of your dataset?

This dataset contains observations about 4,898 rated white wines. 11 of the variables are attributes of the wine itself, like the acidity and sulphate content. One of the variables is the average of three ratings for each wine on a scale of one to ten.

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

The primary features of interest in this dataset are the quality rating and the wine attributes which contribute to it. This analysis will identify which variables have the most impact on the wine quality as well as whether those variables themselves are related.

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

The variables which probably have the most impact on the rating will also contribute to the taste of the wine. Some of those attributes are volatile.acidity, which can add a vinegar taste in large quantities; citric.acid, which is known to give a wine a “fresh” flavor; residual.sugar and chlorides since sugar and salt directly impact taste, total.sulfur.dioxide, which also contributes to the nose and taste; and the alcohol content, because alcohol is usually an evident taste in wine.

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

I created the ordered factor, sweetness, to align with the European Union’s definition of wine sweetness categories by residual sugar measured in grams of sugar per Liter of wine.

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?

To generally asses each wine attribute, I created a function which calculates the variable’s outlier thresholds using the following equations:

- Lower Limit = Quartile 1 - 1.5 * IQR
- Upper Limit = Quartile 3 + 1.5 * IQR

I created a second function which produces two arranged plots per variable. The first plot displays a histogram of the variable with red lines indicating where the outlier thresholds are located. The second plot displays the same histogram with the x-axis zoomed in to view only non-outliers. Using this method, I only identified one conspicuous finding. In the citric.acid histogram, there was a very noticeable spike at 0.49 with over 200 observations. This value is higher than both median and mean at 0.32 and 0.33, respectively. Upon further investigation, there did not appear to be noticeable differences between the observations with this citric.acid measurement and the rest of the dataset.

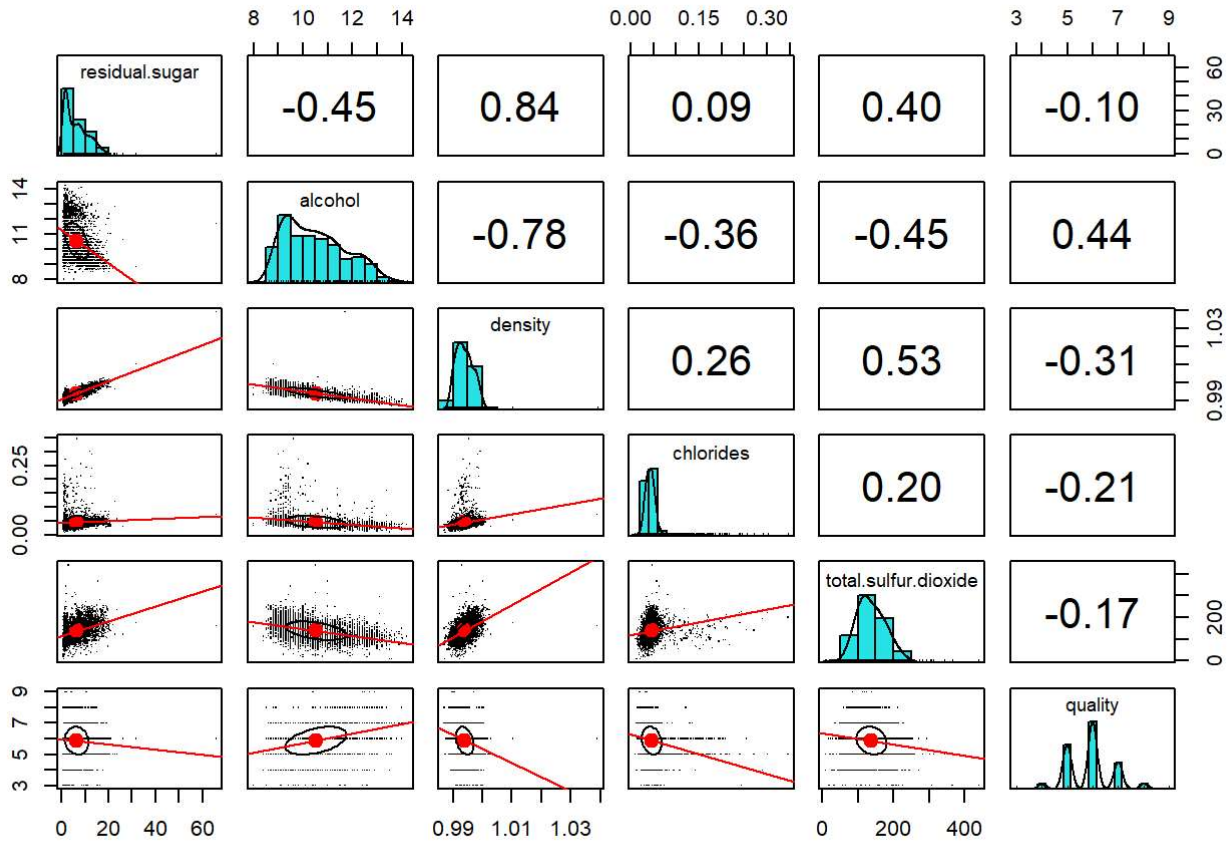
For a couple of the variables, the outlier method was not effective. In these cases, I created custom plots to better view the distribution and quartiles. This was done for residual.sugar and alcohol. With residual.sugar, I performed a log10 transformation to get a better view of the right tail. This produced a bimodal distribution with the modes located near the first and third quartiles. I then used the sweetness categories to view residual.sugar across the different groupings. This grouping will be useful to assess residual.sugar against other variables during the bivariate and multivariate analyses.

With alcohol, the distribution did not contain outliers, so the zoom did nothing. Instead, I adjusted the binwidth, x-axis breaks and added quartile lines. This brought out a stairstep-like pattern on the right side, which looked familiar. After reviewing a few of the other plots, I realized this pattern is evident in a number of the distributions.

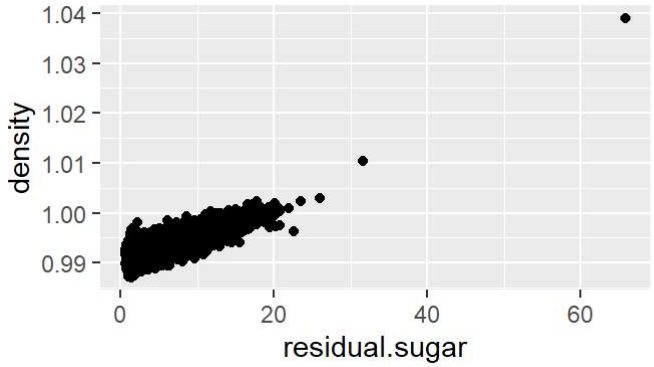
Bivariate Plots Section

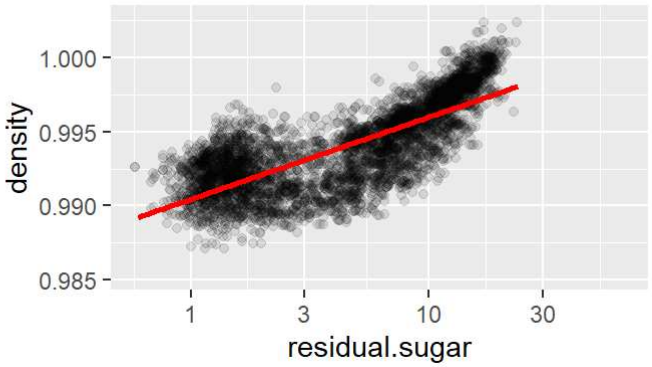
```
##              row               column        cor
## 1 residual.sugar              density  0.8389665
## 2        alcohol              density -0.7801376
## 3        density total.sulfur.dioxide  0.5298813
## 4 residual.sugar              alcohol -0.4506312
## 5        alcohol total.sulfur.dioxide -0.4488921
## 6        alcohol              quality  0.4355747
## 7 residual.sugar total.sulfur.dioxide  0.4014393
## 8        alcohol            chlorides -0.3601887
## 9        density              quality -0.3071233
```

The table above contains the rows and columns of the correlation matrix for the variables of interest in this dataset. It is sorted by the descending absolute value of the correlation for each column-row pair and filtered to display only correlations stronger than plus or minus 0.3, resulting in 9 rows.



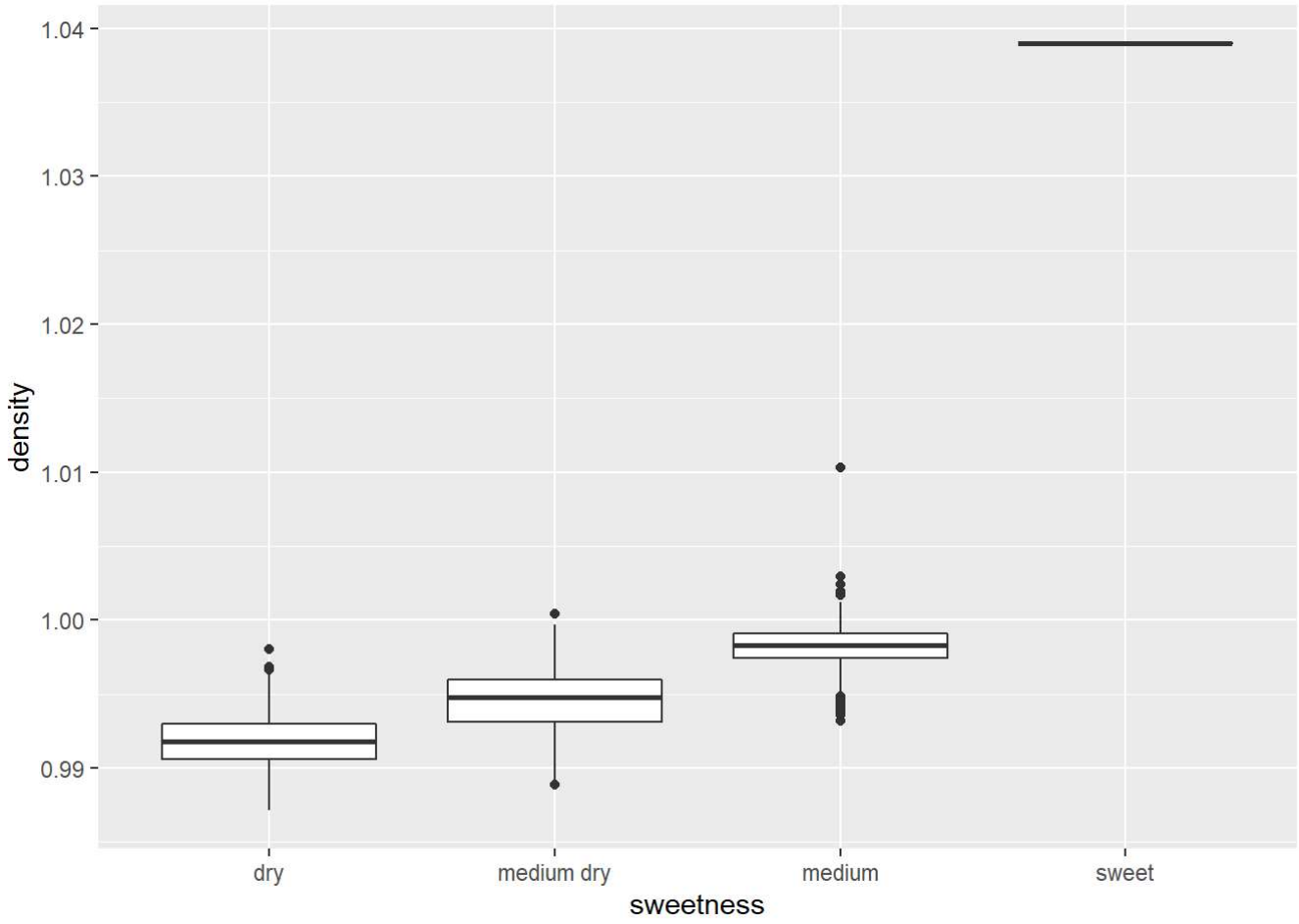
The matrix above displays the relationships between the variables of interest. I first want to look at the relationships between density, residual.sugar, and alcohol since the dataset documentation specifically mentions that density is dependent on both.



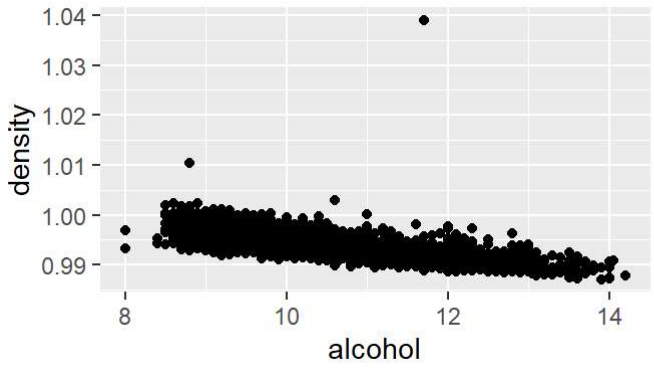


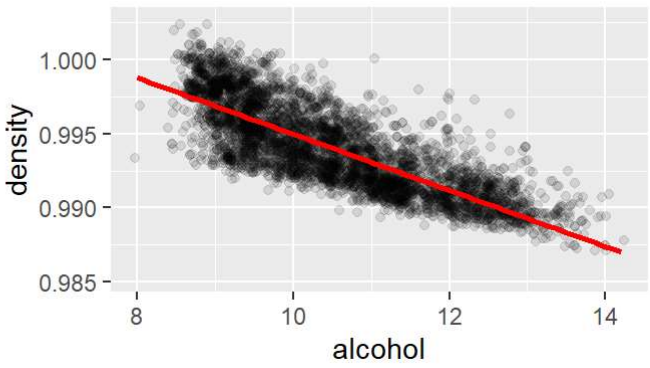

```
##
##  Pearson's product-moment correlation
##
## data:  wine.sub$residual.sugar and wine.sub$density
## t = 107.87, df = 4896, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8304732 0.8470698
## sample estimates:
##       cor
## 0.8389665
```

The initial plot does display the positive correlation, but there is a lot of overplotting and the scale could be better. In the second plot, I applied the same transformations as were applied to the univariate plots. The y-axis was zoomed in to eliminate outliers and the x-axis was log10 transformed to better view the right tail of the distribution. Here two clusters are visible which align with the bimodal shape of the transformed residual.sugar distribution. With a strong, positive correlation of 0.839, as residual.sugar increases, so does density. This should translate to a positive trend across the sweetness categories.



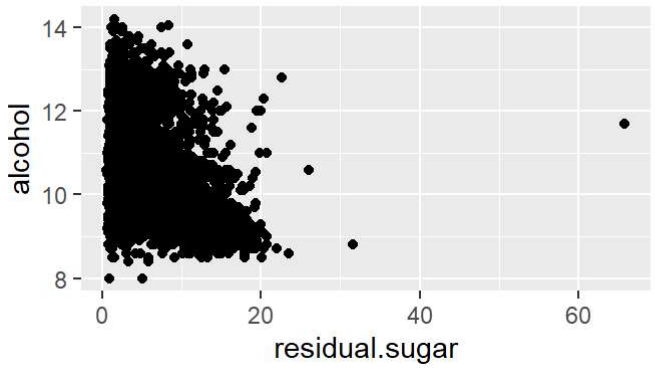
As expected from the previous plot and test, sweeter wines are more dense.

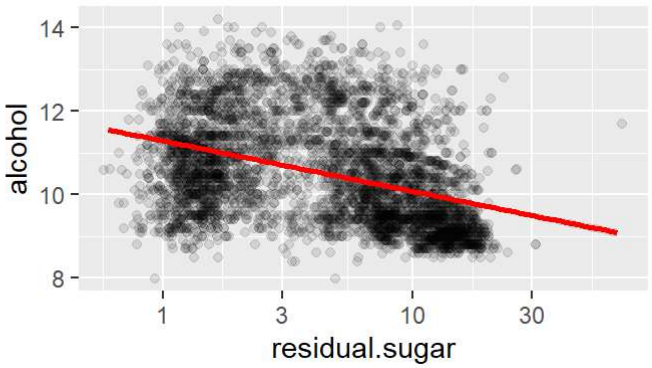





```
##
##  Pearson's product-moment correlation
##
## data:  wine.sub$alcohol and wine.sub$density
## t = -87.255, df = 4896, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.7908646 -0.7689315
## sample estimates:
##       cor
## -0.7801376
```

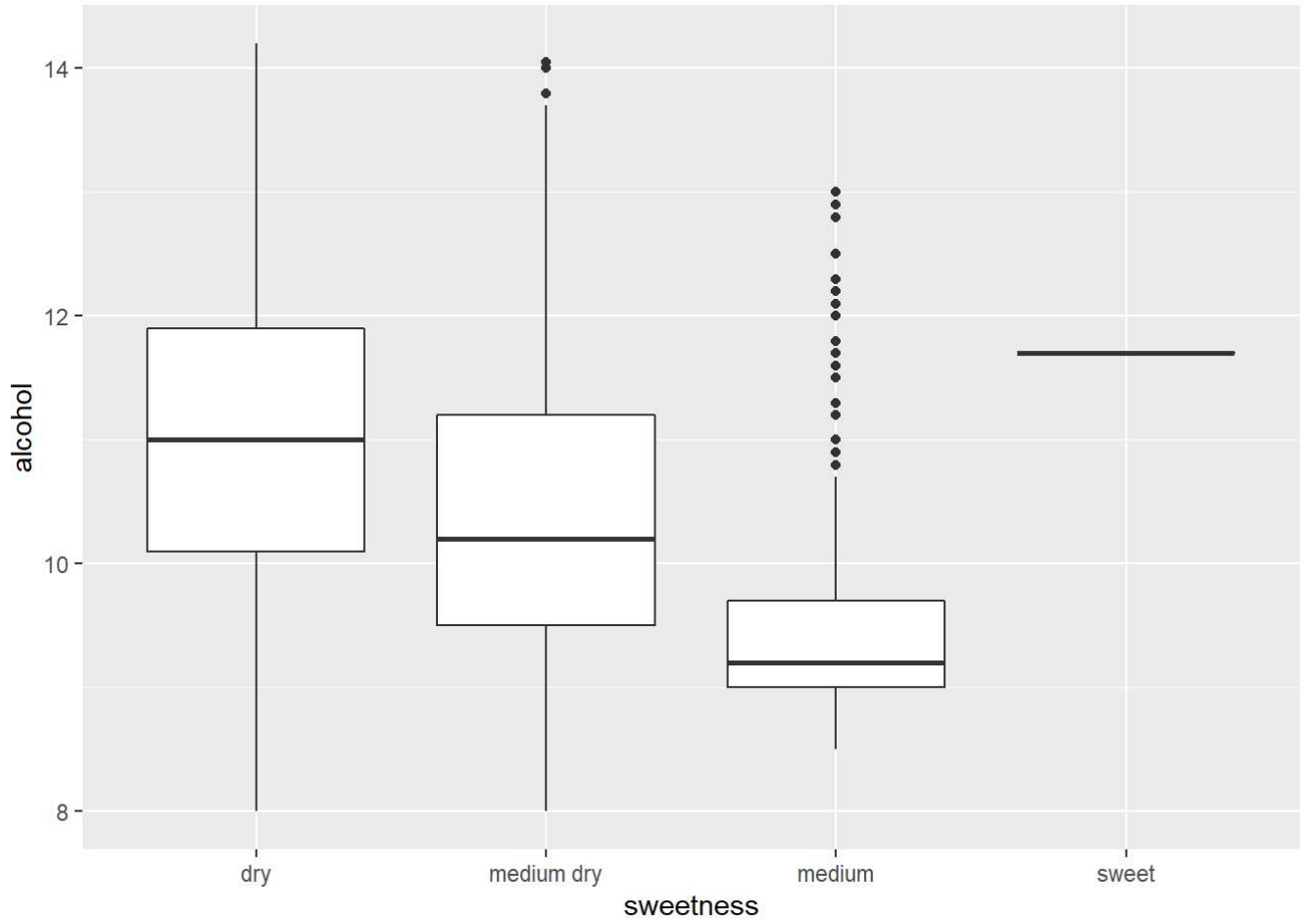
Again, the initial plot is not very clear, so I applied the outlier transformation like before. Here we can see a strong negative correlation of -0.78. As alcohol increases, density decreases. I also want to look at the relationship between alcohol and residual.sugar since they are products of the same fermentation process.



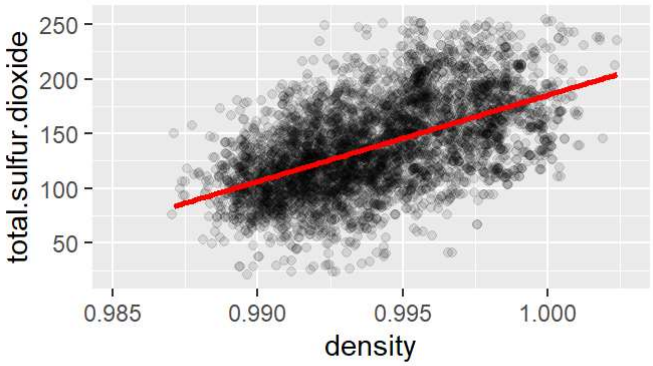


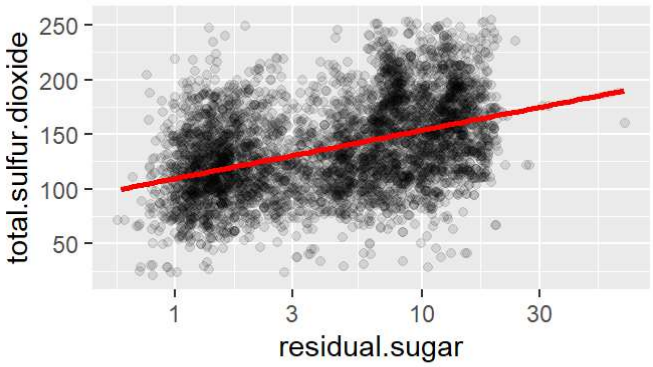

```
##
##  Pearson's product-moment correlation
##
## data:  wine.sub$residual.sugar and wine.sub$alcohol
## t = -35.321, df = 4896, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4726723 -0.4280267
## sample estimates:
##       cor
## -0.4506312
```

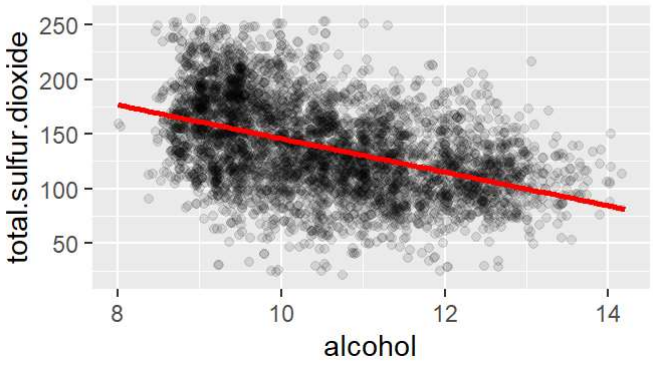
There is a moderate negative correlation of -0.451 between alcohol and residual.sugar. As residual.sugar increases, alcohol decreases. A similar pattern should be visible in a plot of alcohol and sweetness.



The negative correlation is far more apparent here. The exception is the sweet grouping, which does not contain enough data to summarize (n = 1). Next, I want to look at total.sulfur dioxide and how it relates to density, residual.sugar, and alcohol.





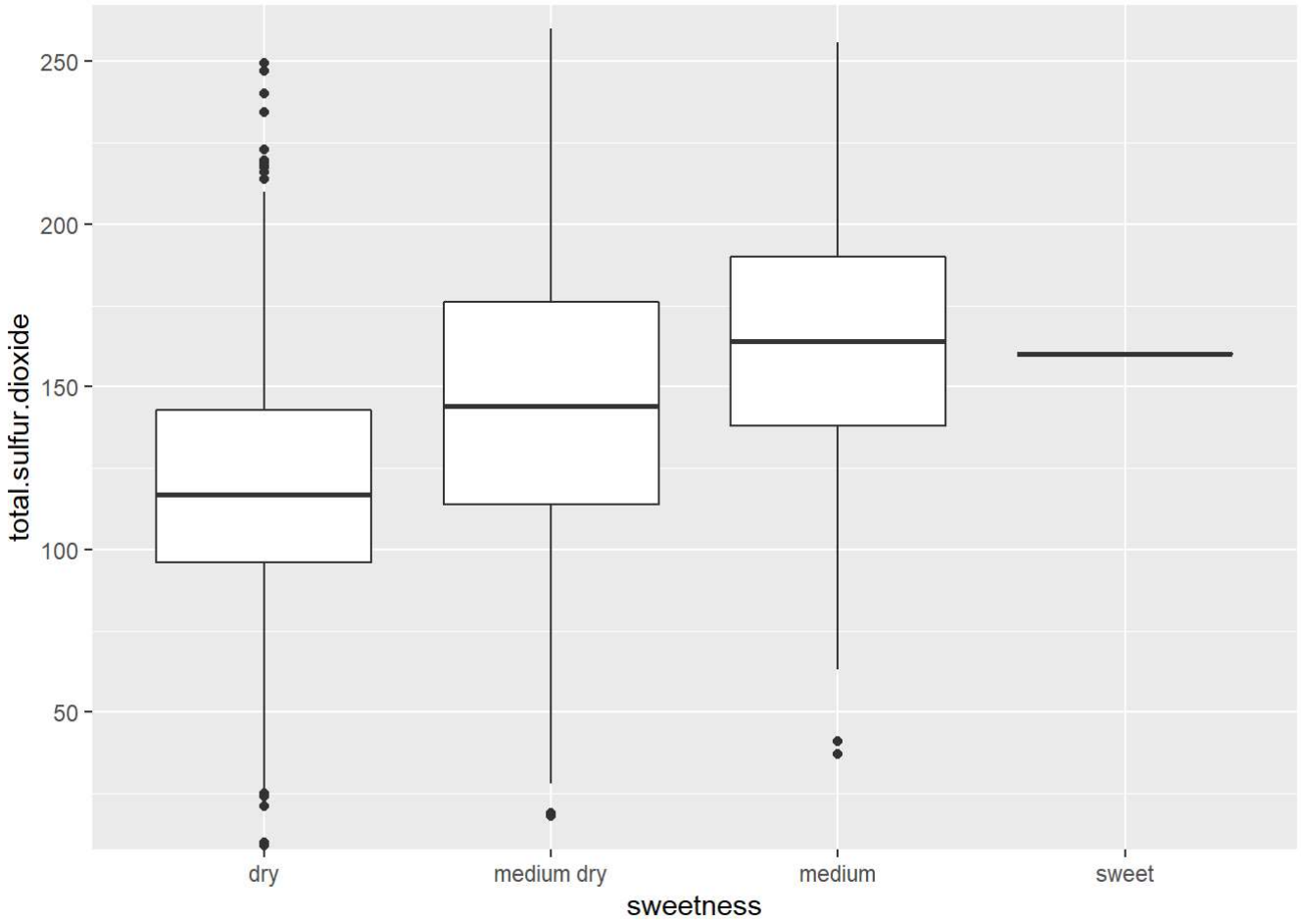



```
##
##  Pearson's product-moment correlation
##
## data:  wine.sub$density and wine.sub$total.sulfur.dioxide
## t = 43.719, df = 4896, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5094349 0.5497297
## sample estimates:
##       cor
## 0.5298813
```

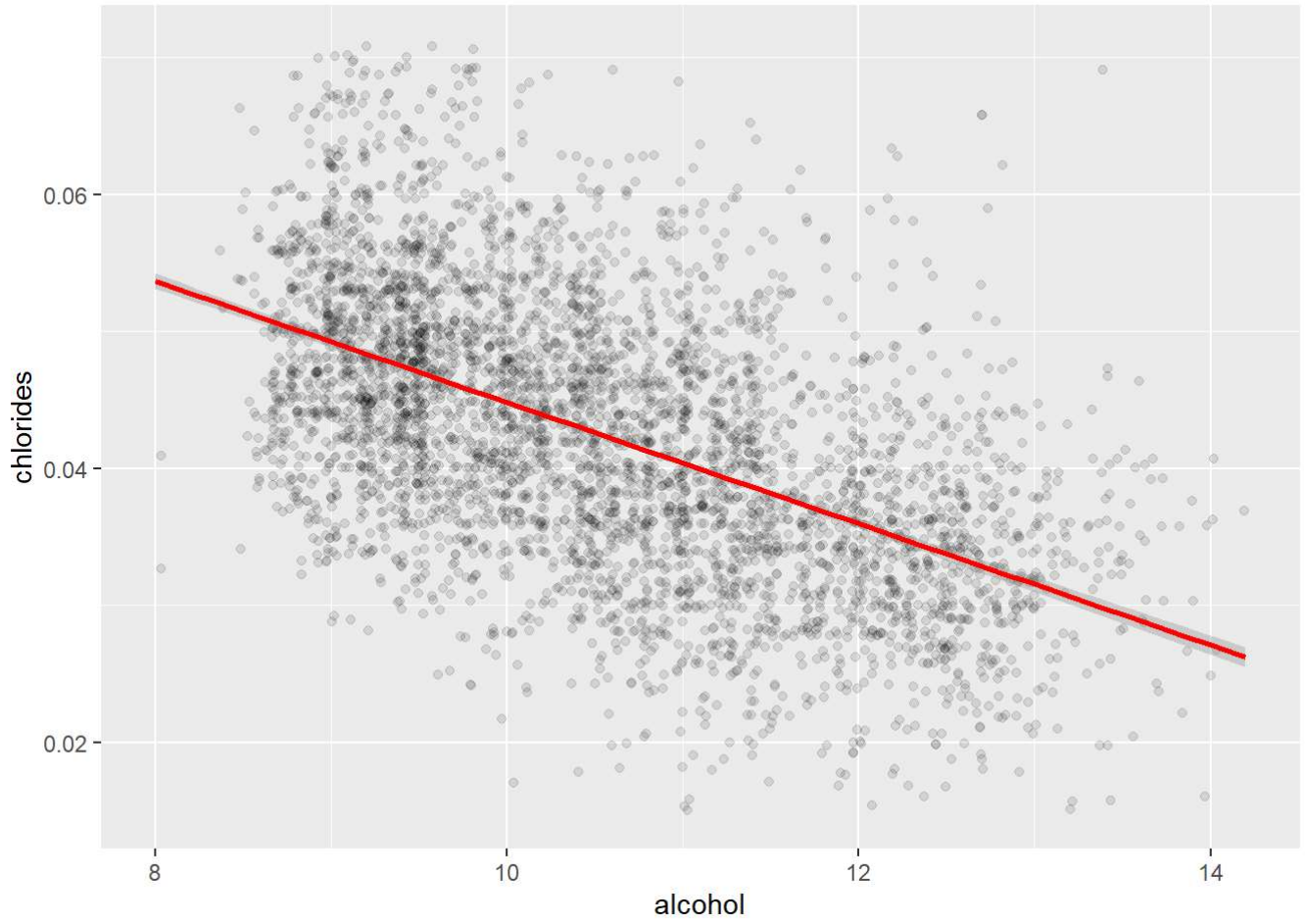
```
##
##  Pearson's product-moment correlation
##
## data:  wine.sub$residual.sugar and wine.sub$total.sulfur.dioxide
## t = 30.669, df = 4896, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3776791 0.4246712
## sample estimates:
##       cor
## 0.4014393
```

```
##
##  Pearson's product-moment correlation
##
## data:  wine.sub$alcohol and wine.sub$total.sulfur.dioxide
## t = -35.15, df = 4896, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4709775 -0.4262443
## sample estimates:
##        cor
## -0.4488921
```

density has a moderate correlation of 0.53 with total.sulfur.dioxide. It can be seen on the plot that as density increases, so does total.sulfur.dioxide. Similarly, residual sugar is positively correlated with total.sulfur.dioxide, but the strength is low. With a correlation of 0.401, as residual.sugar increases, total.sulfur.dioxide also increases. alcohol, on the other hand, has a low negative correlation with total.sulfur.dioxide of -0.449. As alcohol increases, total.sulfur.dioxide decreases. I also want to see the relationship total.sulfur dioxide has with sweetness. It should follow the same correlation pattern as residual.sugar like was seen in previous comparisons.



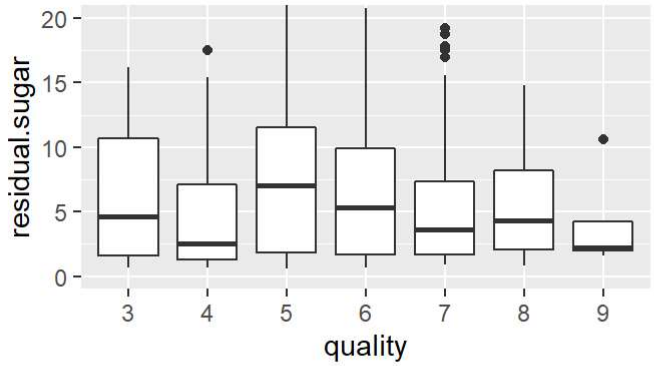
I zoomed this plot in to get a better view of the bulk of the data. Sure enough, there is a positive shift when moving from dryer to sweeter wines. Again, there is not enough information to consider results in the sweet category (n = 1). The last relationship I would like to look at before diving into the relationships with quality is alcohol and chlorides.

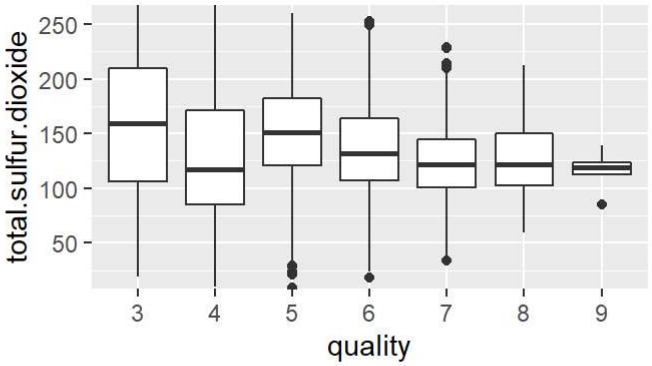


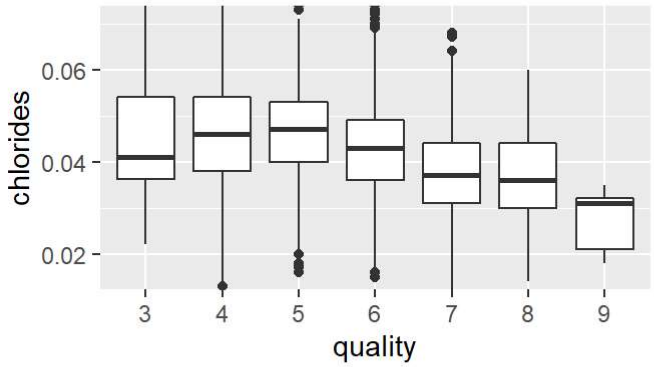

```
##
##  Pearson's product-moment correlation
##
## data:  wine.sub$alcohol and wine.sub$chlorides
## t = -27.016, df = 4896, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3843183 -0.3355673
## sample estimates:
##        cor
## -0.3601887
```

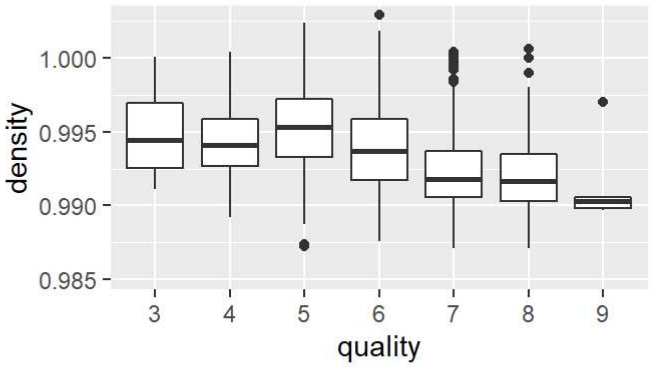
Applying the same outlier detection methods as were used in the univariate analysis, this plot is zoomed in to eliminate outliers. A low negative correlation is visible in the plot and reflected in the test with a value of -0.36. Now that the interaction between the attributes have been inspected, it is time to see how they measure up against the quality ratings.

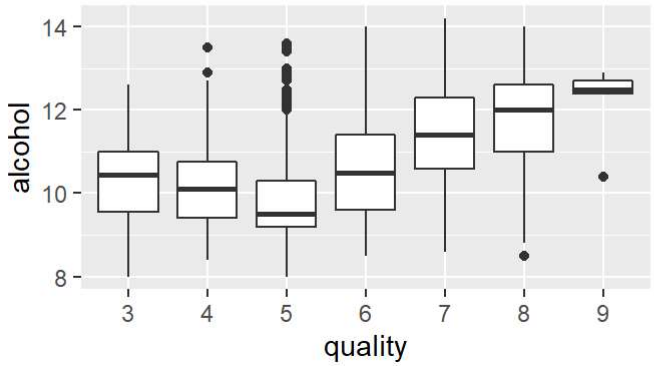
```
##                    row  column         cor
## 1       residual.sugar quality -0.09757683
## 2 total.sulfur.dioxide quality -0.17473722
## 3            chlorides quality -0.20993441
## 4              density quality -0.30712331
## 5              alcohol quality  0.43557472
```











From the correlation table, only density and alcohol are noteworthy. density is negatively correlated with quality (-0.307). This can be seen in the plot. It is weak, but there is a general downward trend. density decreases as quality increases. alcohol, however, has a positive correlation with quality (0.436). As alcohol increases, quality does too. The residual.sugar and total.sulfur.dioxide plots have no visible trend. The chloride plot appears to have a downward trend, but the correlation is very weak at -0.21.

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?

There are no strong correlations between the wine quality ratings and the other variables. However, there are a few low to moderate correlations. The attribute with the strongest relationship to quality is alcohol (0.436). As alcohol increases, quality ratings increase.

density has the next strongest relationship with quality (-0.307). As density increases, quality decreases.

chlorides also appears to have a relationship with quality when looking at the boxplots where quality decreases as chlorides increase, However, the correlation is only -0.21.

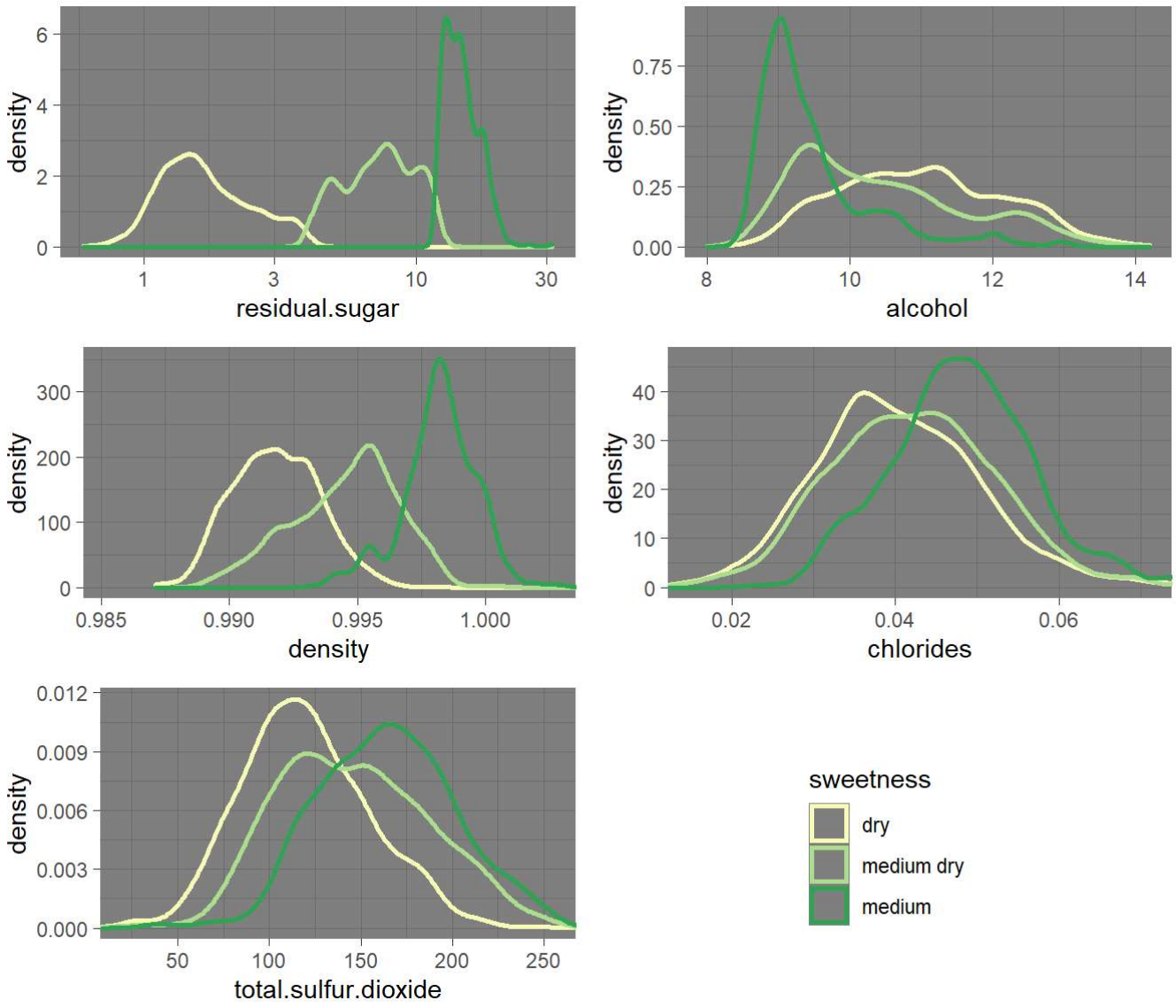
Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

The bimodal shape of the residual.sugar distribution, which was revealed in the univariate plots, is also evident in the multivariate plots. The left hand mode of the residual.sugar distribution appears to align with the dry sweetness category, with the remaining categories captured by the right hand mode. There is only one observation in the sweet sweetness category. This category will be excluded from further analyses.

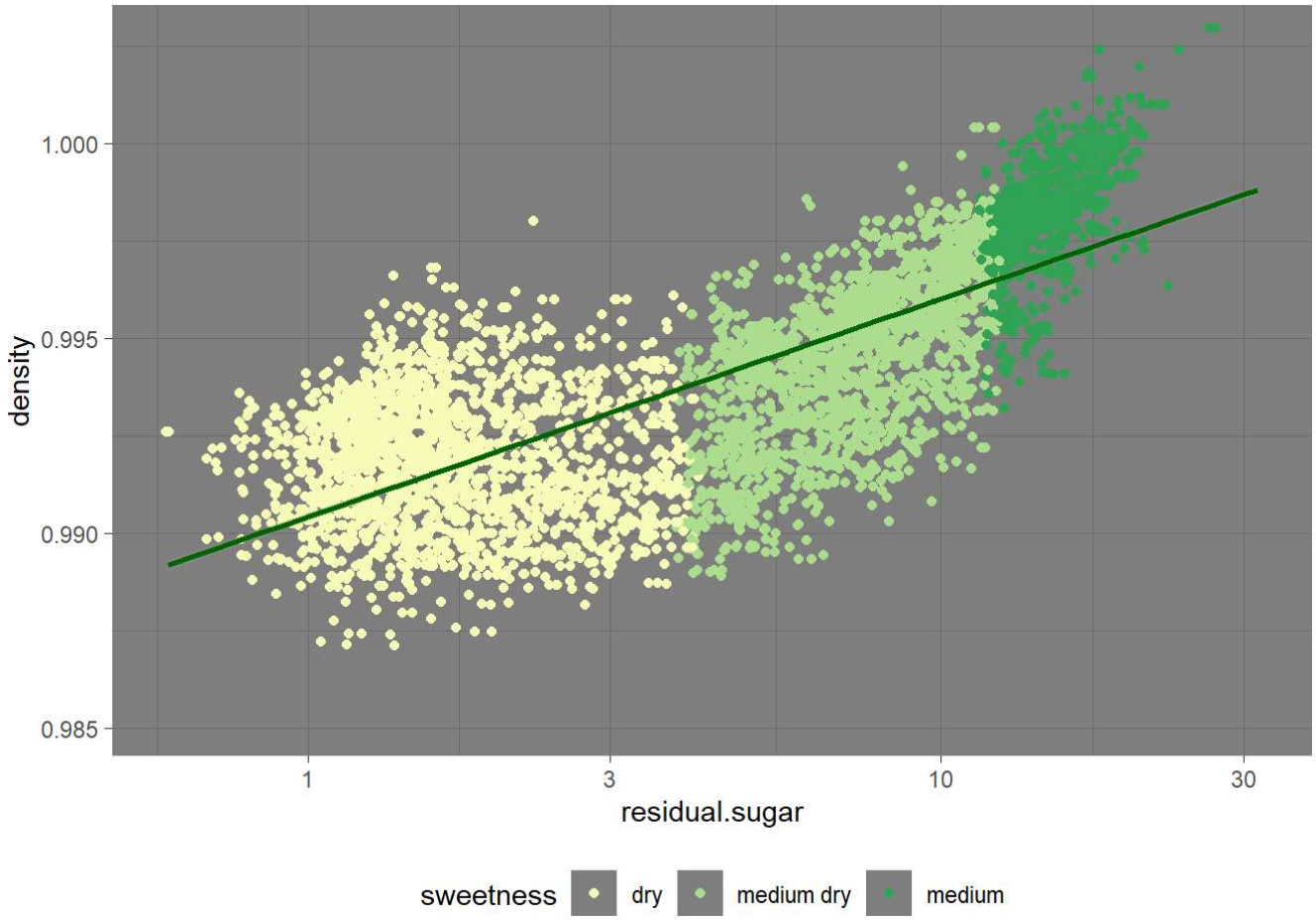
What was the strongest relationship you found?

As was noted in the dataset documentation, density is dependent on residual.sugar and alcohol, which is easily seen in the correlation matrix and tables, as well as the plots. residual.sugar and density have a strong correlation of 0.84. As residual.sugar increases, density also increases. alcohol and density, however, have a negative relationship. It is still strong at -0.78. As alcohol increases, density decreases. alcohol and residual.sugar are also correlated, but not nearly as strongly as they are with density. Their relationship is negative, at -0.45. As alcohol increases, residual.sugar decreases.

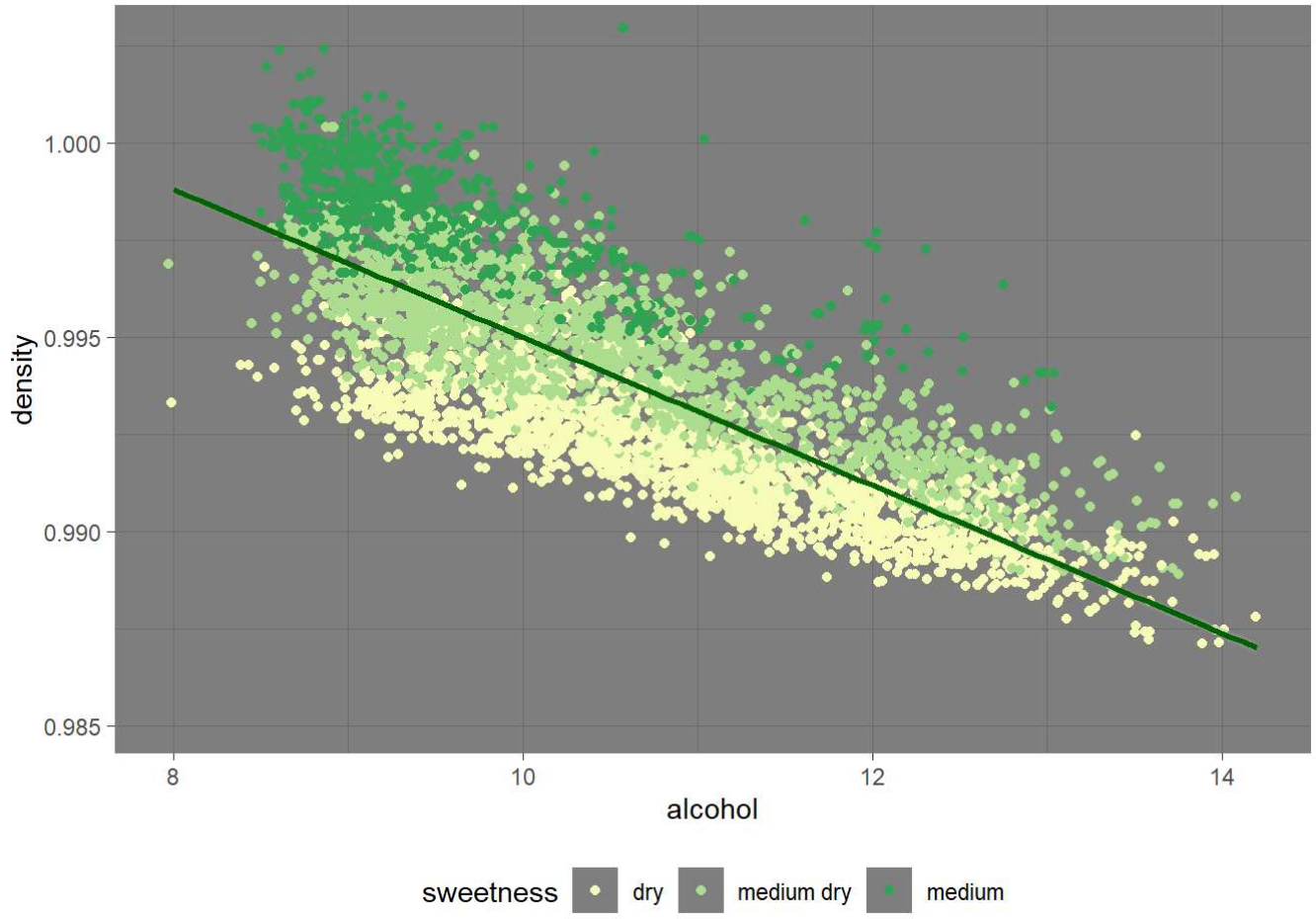
Multivariate Plots Section



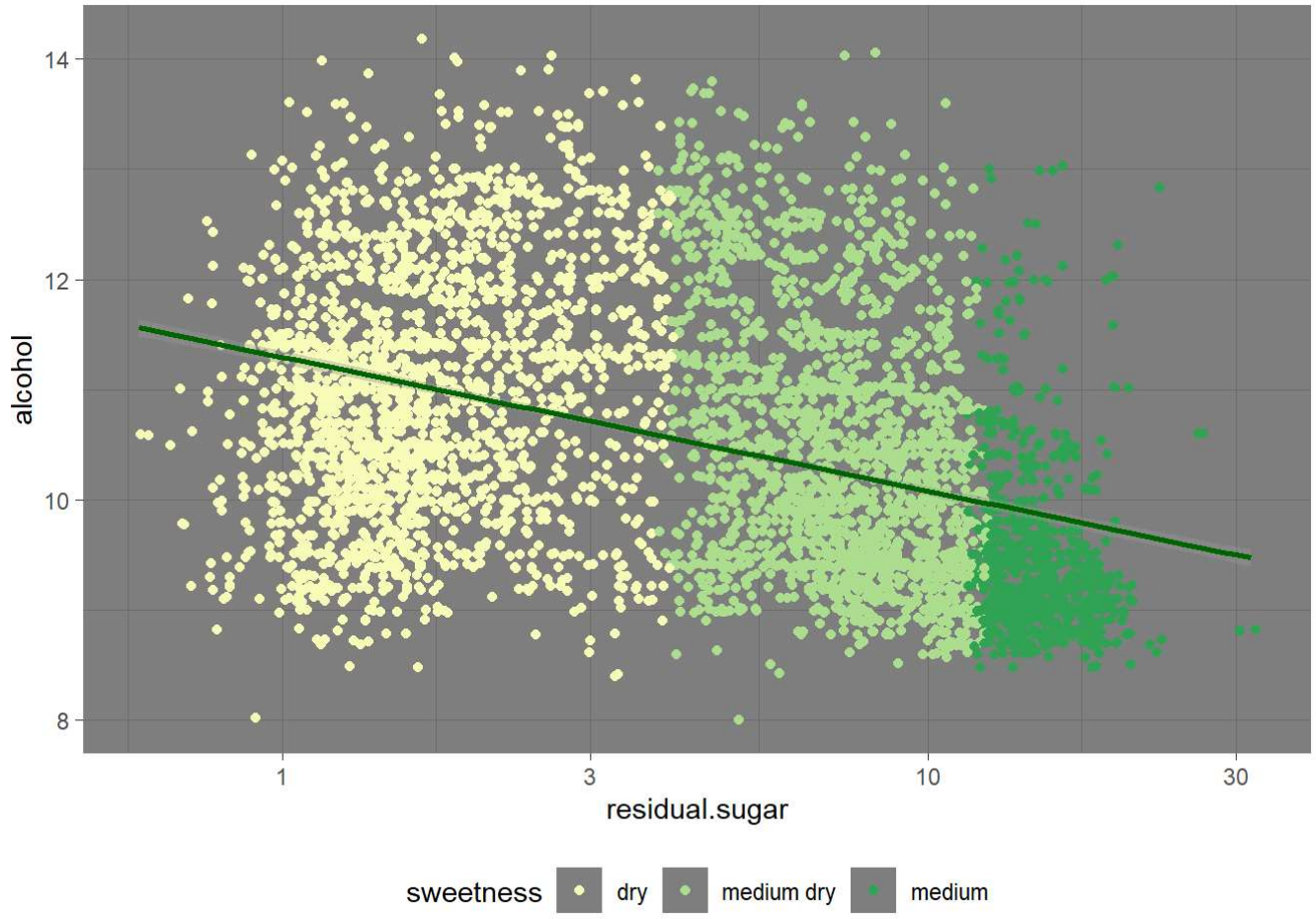
The above plots display each quantitative variable's density (not the variable, but the measurement) across the first three sweetness groupings. The sweet group has been excluded since it only contained a single observation. Transformations were performed in line with the univariate and bivariate plots. As would be expected, residual sugar has three distinct shapes that do not overlap because residual.sugar is used to determine the sweetness group. The remaining plots show that there are in fact fluctuations among the sweetness groupings for both alcohol and density. chlorides and total.sulfur.dioxide also fluctuate, but much more evenly, with medium wines occurring with higher, and dry wines with lower chlorides and total.sulfur.dioxide. Next I want to look at each of the comparisons done in the bivariate analysis with the added layer of sweetness.



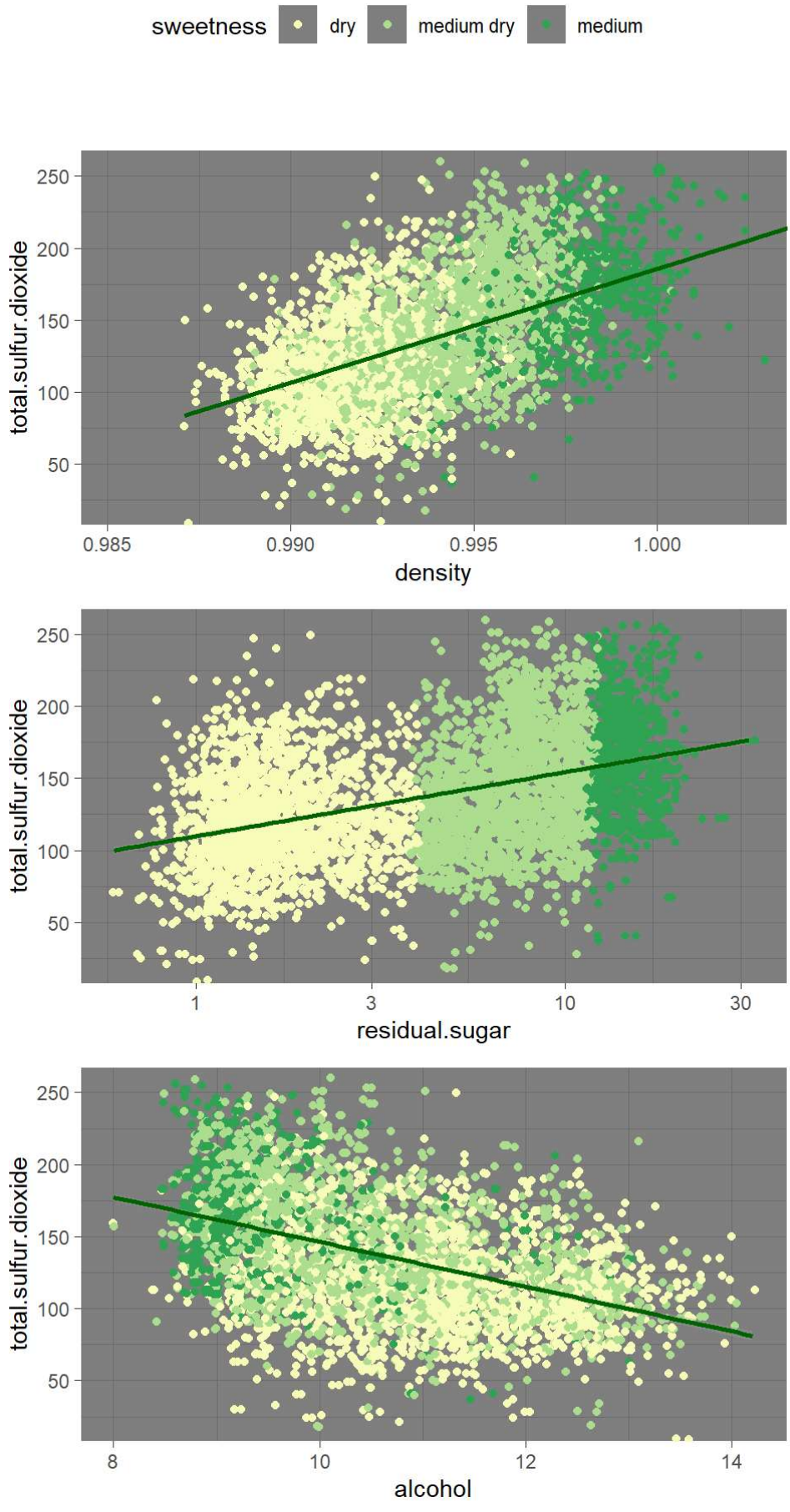
What is interesting about this plot, is the two clusters which were previously visible are now different colors. The left cluster is the dry wine grouping, the right cluster is every other grouping. This is due to the log10 scale which was applied to reduce some of the spread in the sweeter groupings. As we would expect, the groups do not overlap since residual.sugar is used to determine sweetness. As was seen before, sweeter wines are more dense.



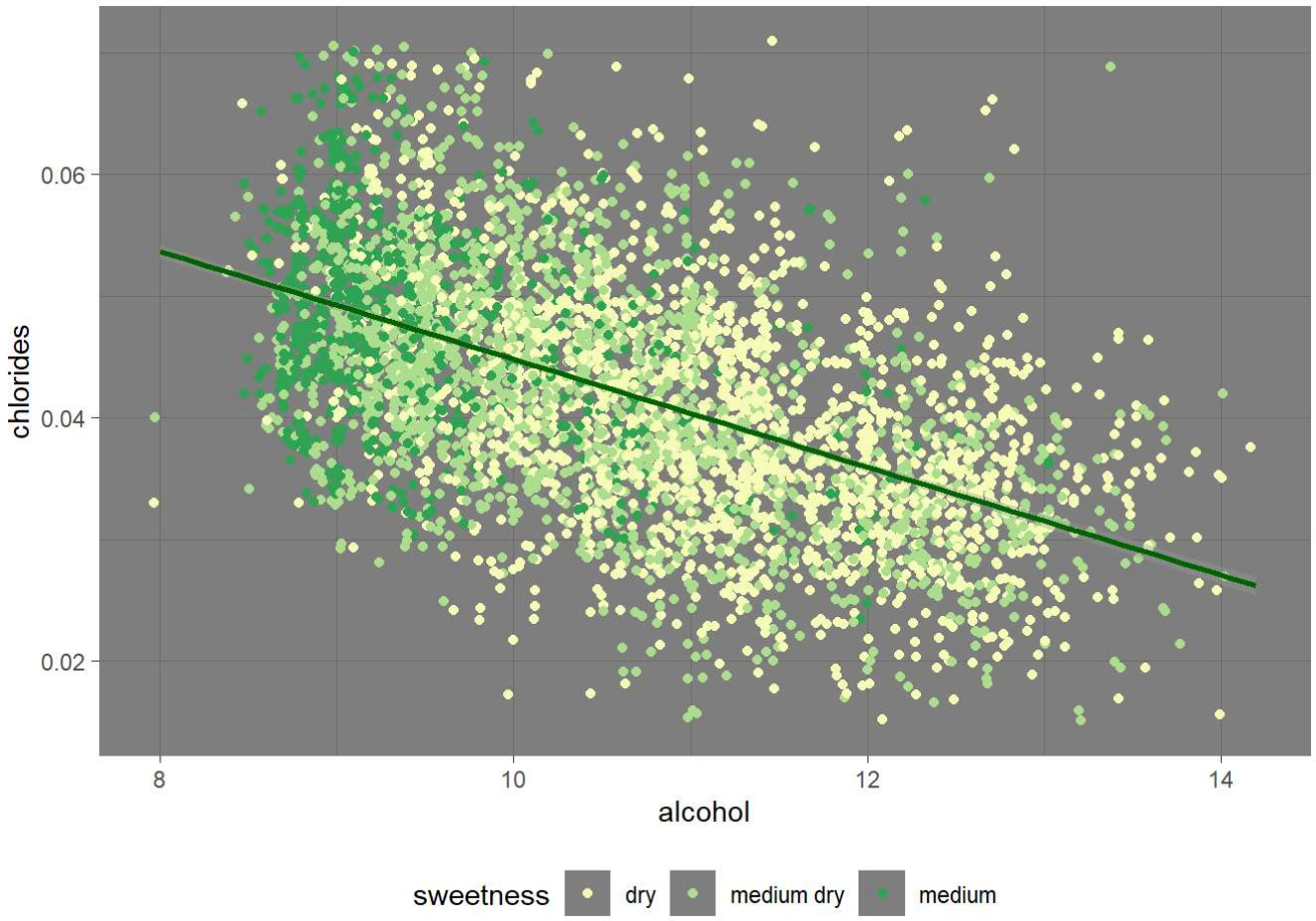
Overlaying sweetness on the density vs alcohol plot reveals what looks like three stripes of data. medium wines are in the top stripe with higher densities and dry wines are in the lower stripe with lower relative densities. The majority of the medium wines contain less alcohol than medium dry or dry wines. dry wines appear to contain more alcohol.



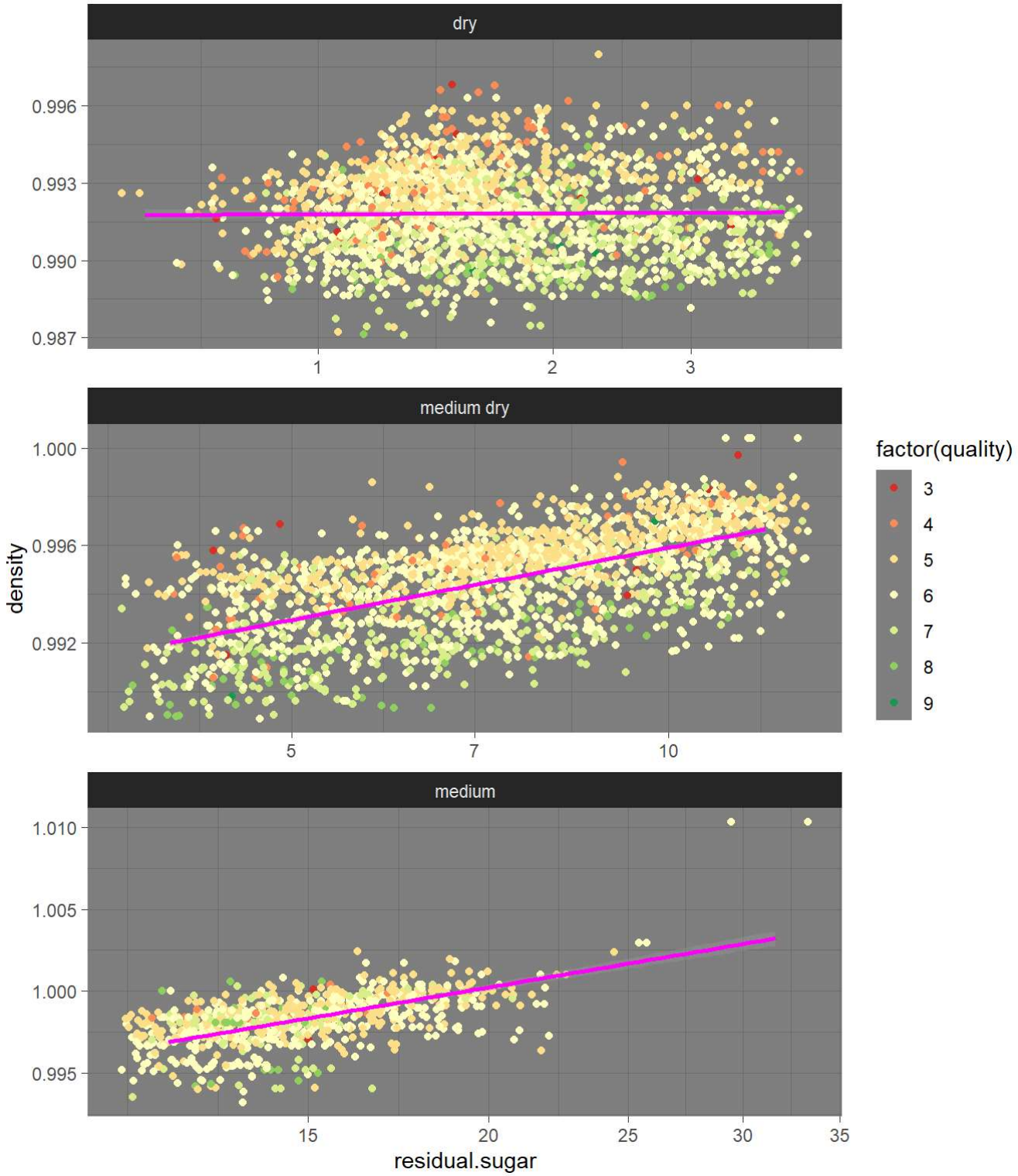
Again, the two clusters in residual.sugar are highlighted by the sweetness groupings. It is easier to see here that sweeter wines tend to have less alcohol.


sweetness
dry
medium dry
medium
total.sulfur.dioxide
250
200
150
100
50
0.985
0.990
0.995
1.000
density
total.sulfur.dioxide
250
200
150
100
50
1
3
10
30
residual.sugar
total.sulfur.dioxide
250
200
150
100
50
8
10
12
14
alcohol

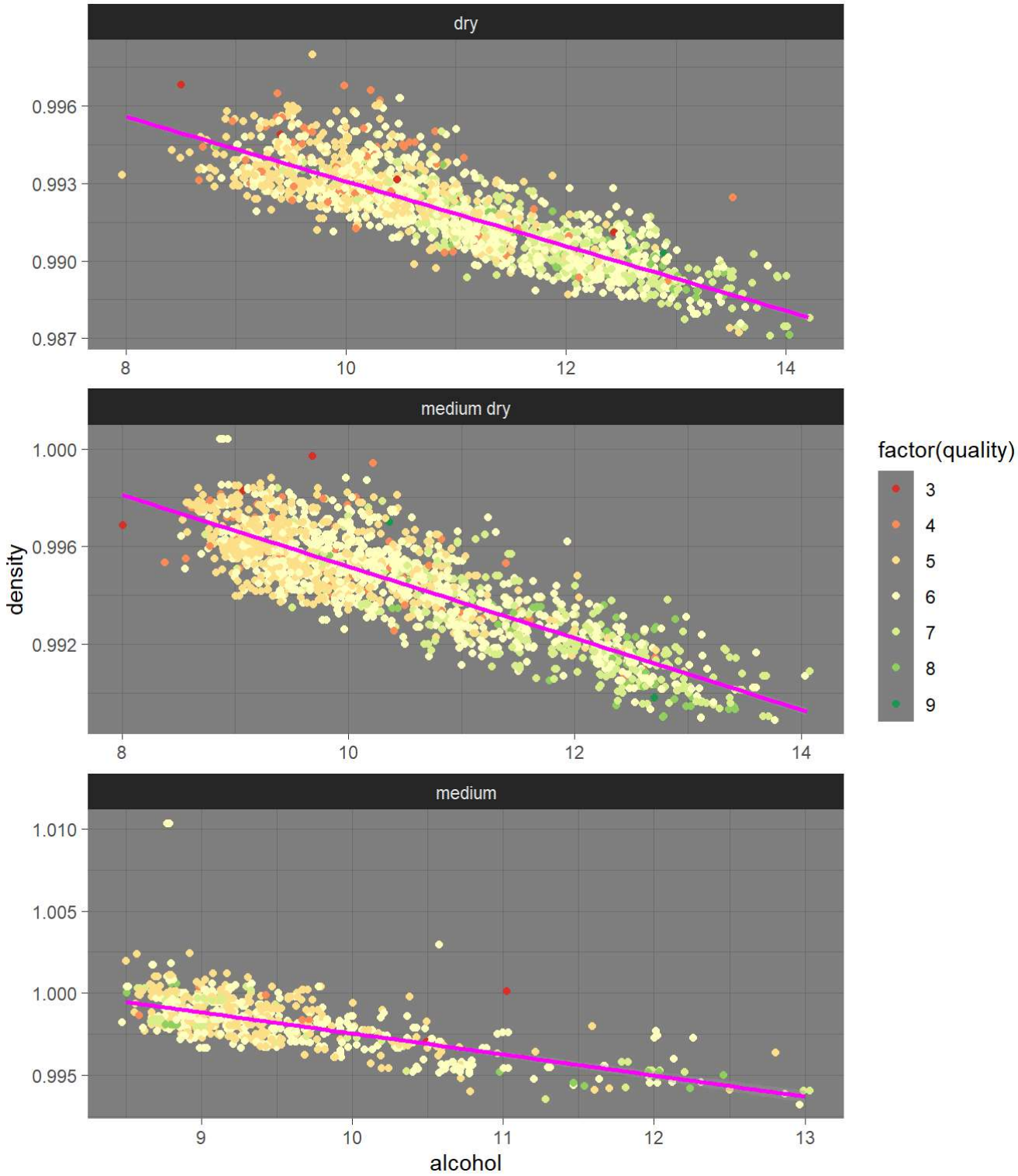
Here we can see that the sweeter wines tend to have a higher total.sulfur.dioxide regardless of whether they are compared over density, residual.sugar, or alcohol.



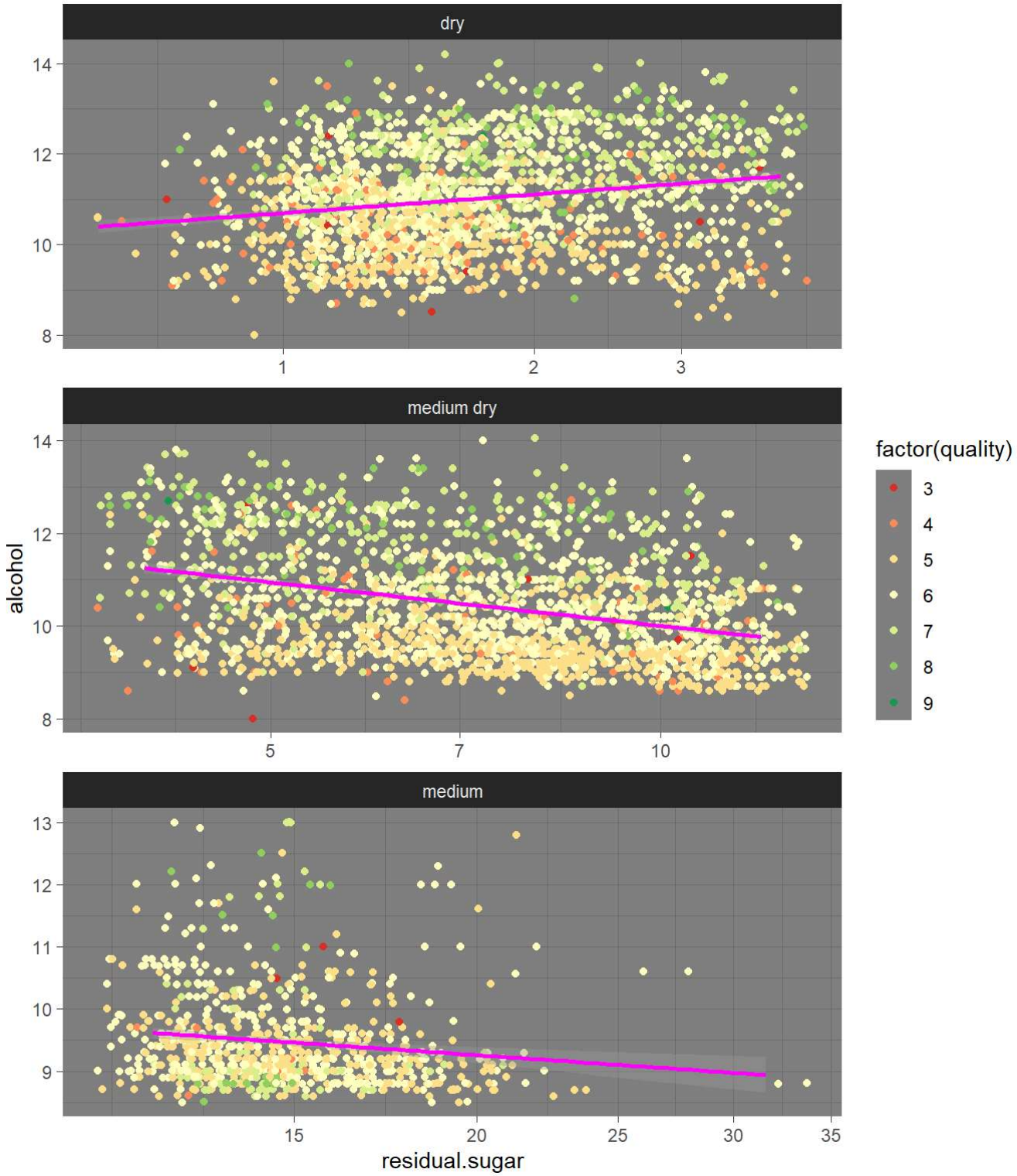
Here we can see that sweeter wines also tend to have higher chloride content. Now that we can clearly see the interplay between the variables of interest, it is time to see how they compare across the quality ratings.



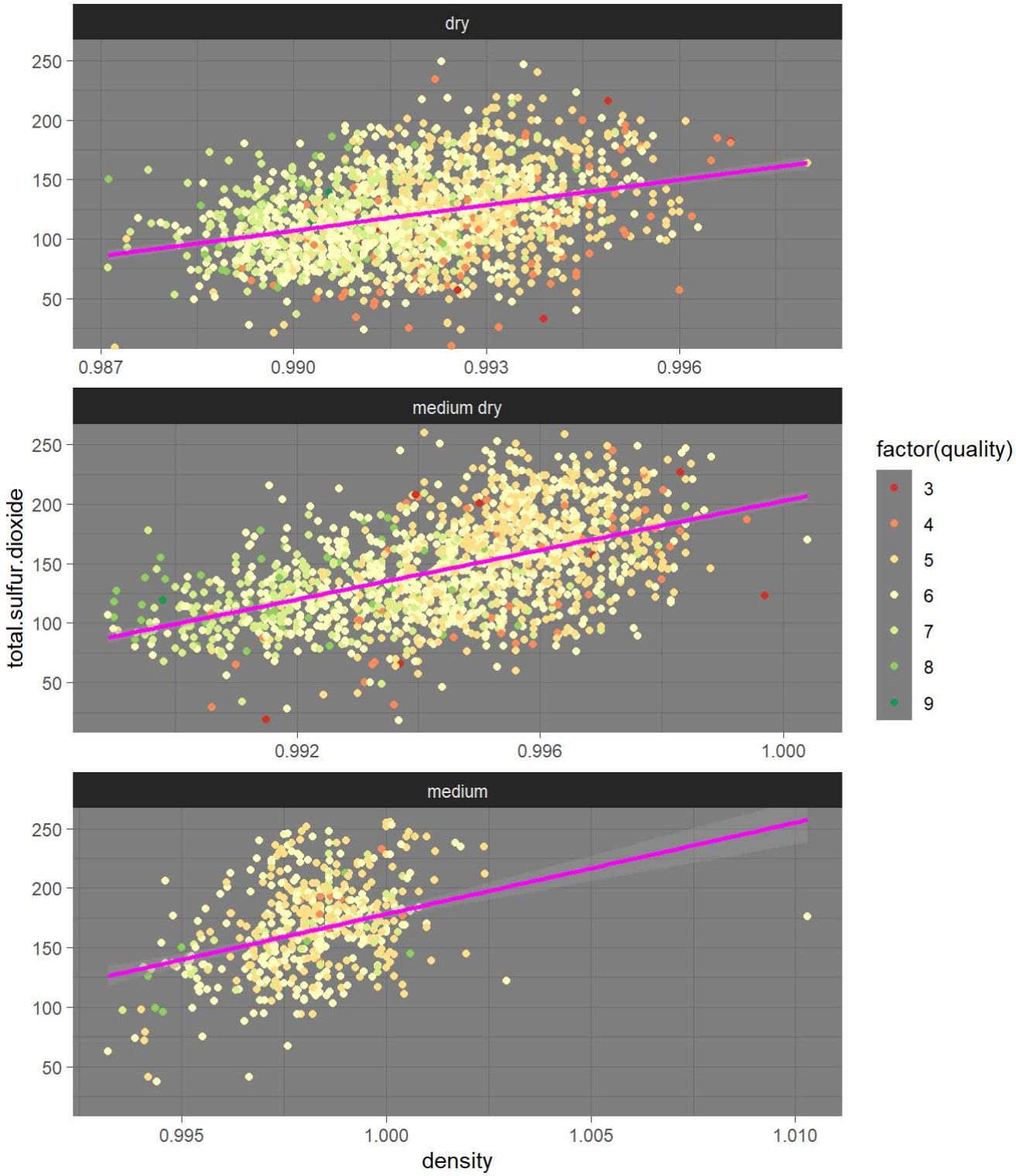
Here, the density and residual.sugar correlation is displayed. When the plots are faceted over sweetness and a diverging color scale is applied over the quality rating, we can see that the majority of higher quality ratings (7-9) fall below the mean (magenta line) and the majority of lower quality ratings (3-5) fall above the mean. This suggests that a negative relationship exists such that lower quality wines have a higher density and higher quality wines have a lower density. This is consistent with the quality boxplots from the bivariate analysis.



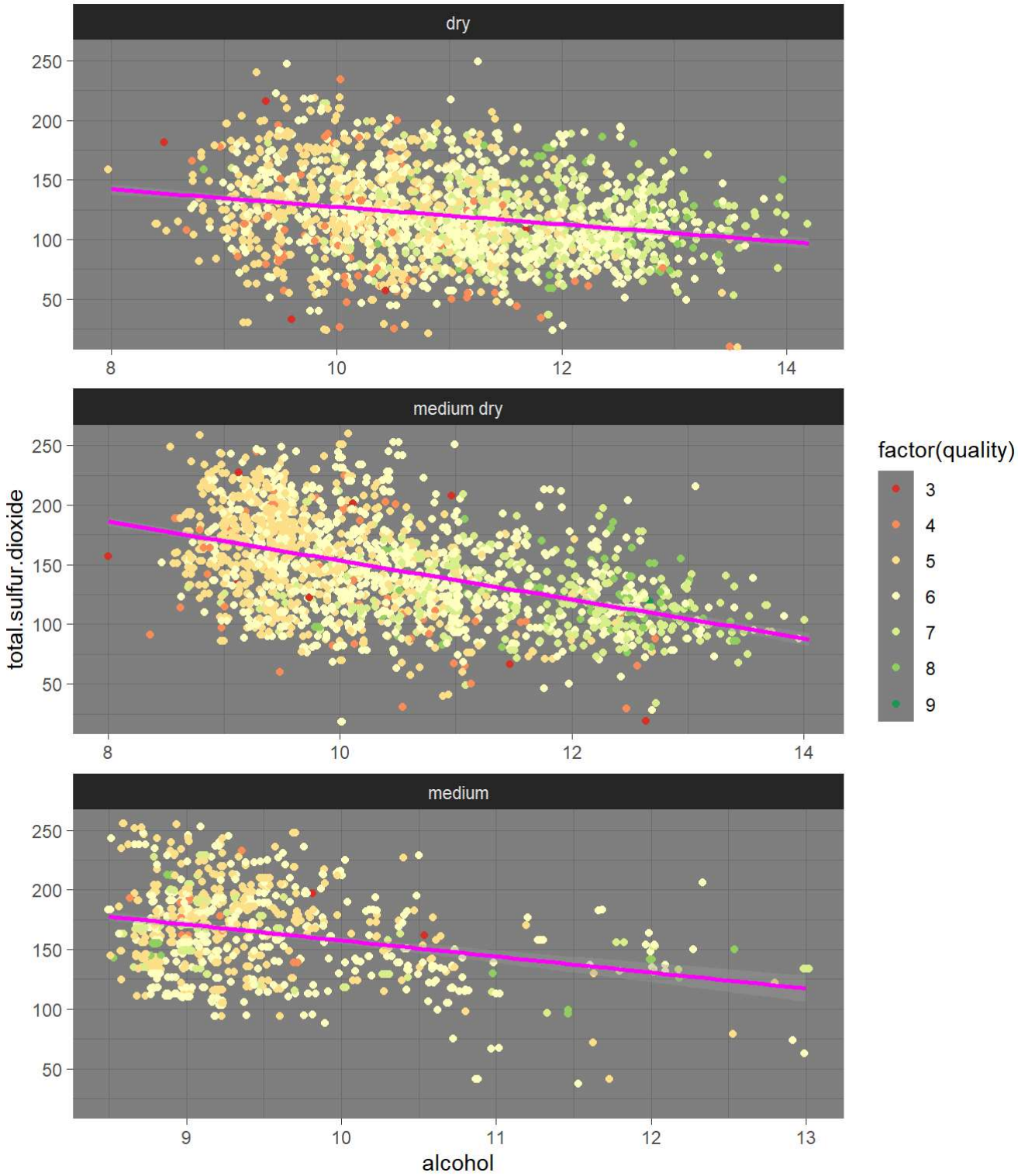
We already know to expect lower density measurements in the higher quality wines. The above plot also suggests that higher quality wines contain more alcohol. Again, this is consistent with the quality boxplots from the bivariate analysis.



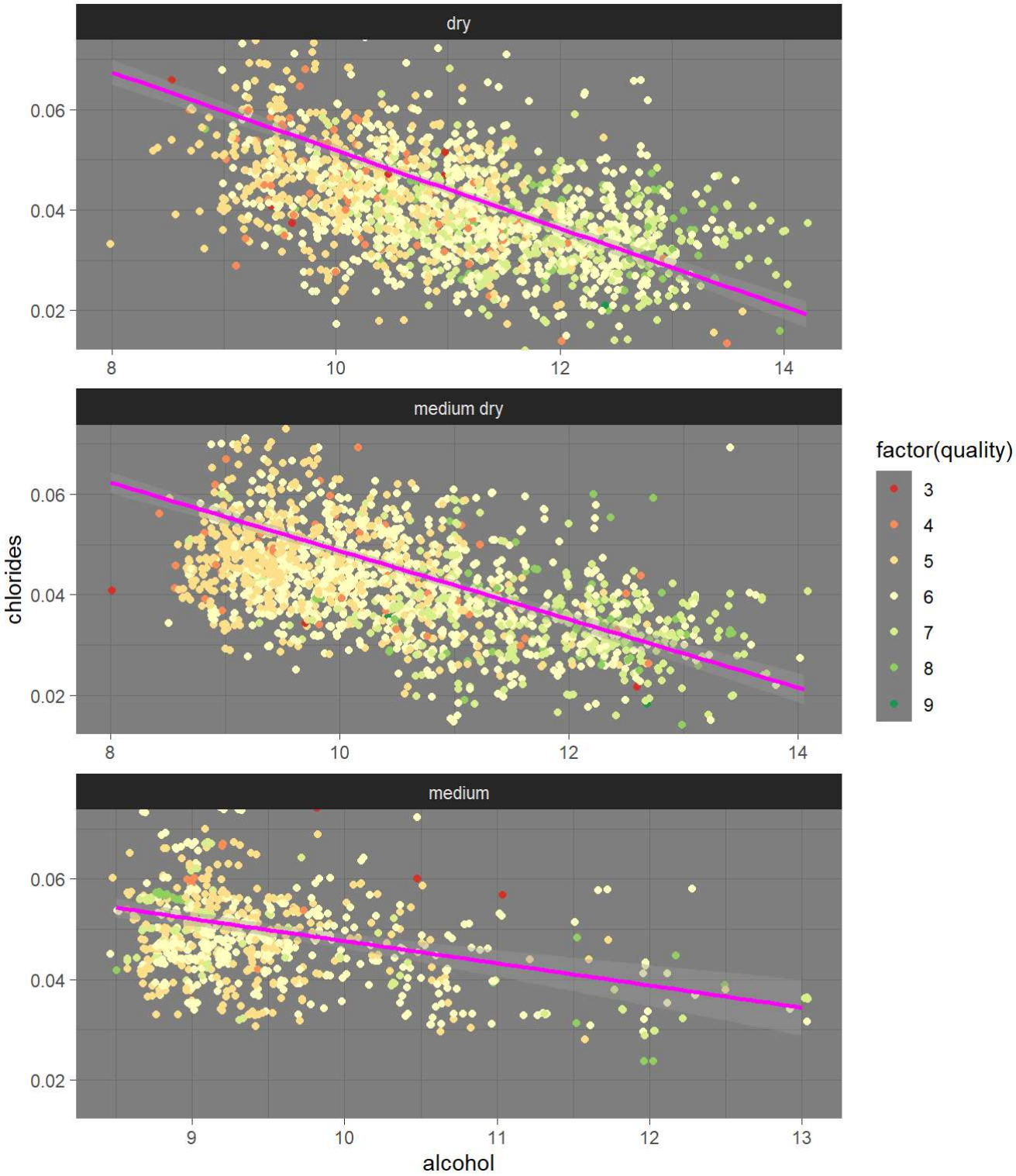
Again, as was seen in the bivariate analysis boxplots for quality, the higher quality wines have more alcohol with the exception of medium sweetness wines. In the medium group, there does not appear to be any strong patterns.



As we've seen before, the higher quality wines have a lower density and contain less total.sulfur.dioxide.



In keeping with previous observations, higher alcohol tends to align with higher quality. total.sulfur.dioxide, however, appears to have a negative relationship with quality.



The trend with alcohol continues to hold true where higher alcohol contents have higher quality ratings and lower chlorides appear to align with higher quality scores.

```
##
## Call:
## lm(formula = quality ~ residual.sugar, data = wine.model)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0054 -0.8107  0.0433  0.2624  3.1731
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     6.018058   0.020937 287.434  < 2e-16 ***
## residual.sugar -0.018034   0.002592  -6.958 3.92e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8773 on 4667 degrees of freedom
## Multiple R-squared:  0.01027,    Adjusted R-squared:  0.01006
## F-statistic: 48.42 on 1 and 4667 DF,  p-value: 3.923e-12
```

```
##
## Call:
## lm(formula = quality ~ sweetness, data = wine.model)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.97623 -0.87798  0.02377  0.21809  3.12202
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.8787063  0.0139772 420.591  < 2e-16 ***
## sweetness.L -0.1374038  0.0259649  -5.292 1.27e-07 ***
## sweetness.Q  0.0008846  0.0223161   0.040    0.968
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.879 on 4666 degrees of freedom
## Multiple R-squared:  0.006507,   Adjusted R-squared:  0.006081
## F-statistic: 15.28 on 2 and 4666 DF,  p-value: 2.43e-07
```

```
##
## Call:
## lm(formula = quality ~ total.sulfur.dioxide, data = wine.model)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2573 -0.7528  0.0373  0.3454  3.1016
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           6.3690418  0.0442608  143.90   <2e-16 ***
## total.sulfur.dioxide -0.0033859  0.0003079  -10.99   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8706 on 4667 degrees of freedom
## Multiple R-squared:  0.02525,    Adjusted R-squared:  0.02504
## F-statistic: 120.9 on 1 and 4667 DF,  p-value: < 2.2e-16
```

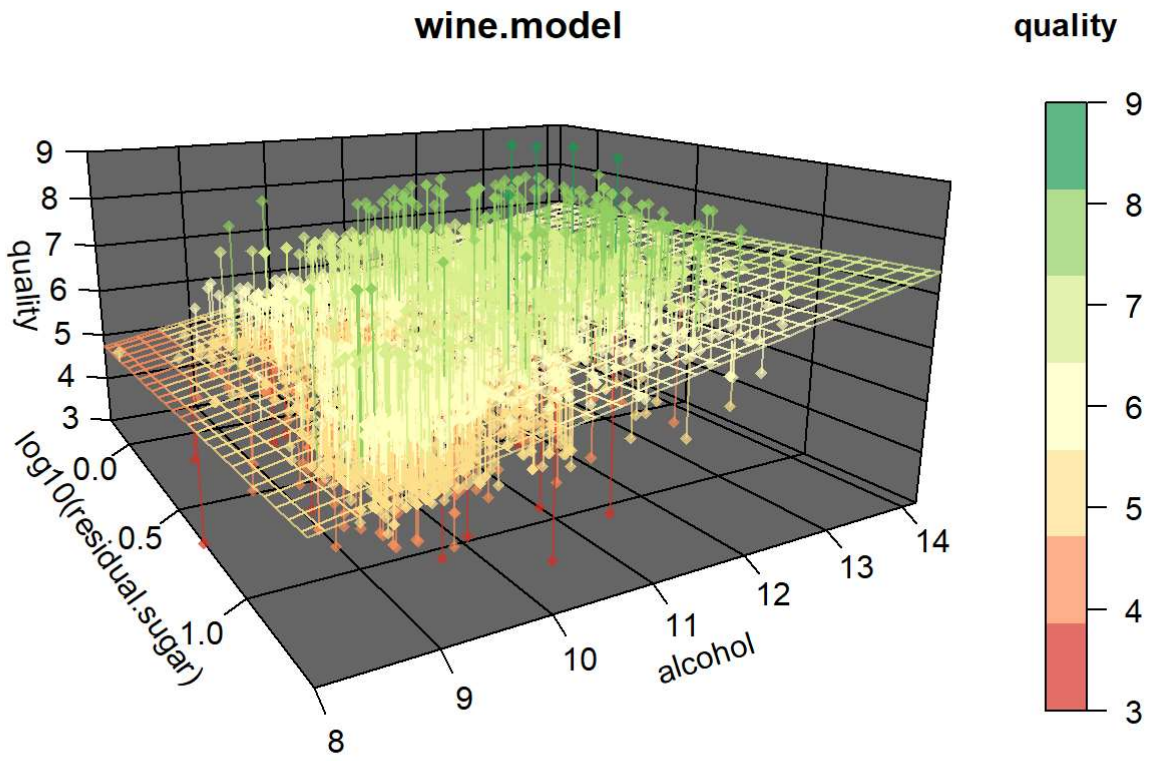
```
##
## Call:
## lm(formula = quality ~ chlorides, data = wine.model)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2382 -0.6614  0.0125  0.4640  2.9122
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.96552    0.05441  128.02   <2e-16 ***
## chlorides   -25.07854    1.25045  -20.06   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8461 on 4667 degrees of freedom
## Multiple R-squared:  0.07935,    Adjusted R-squared:  0.07915
## F-statistic: 402.2 on 1 and 4667 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = quality ~ density, data = wine.model)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1772 -0.5945 -0.0167  0.5393  3.3864
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  100.853      4.190   24.07   <2e-16 ***
## density      -95.526      4.216  -22.66   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.837 on 4667 degrees of freedom
## Multiple R-squared:  0.09911,    Adjusted R-squared:  0.09892
## F-statistic: 513.4 on 1 and 4667 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = quality ~ alcohol, data = wine.model)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4842 -0.5413 -0.0128  0.4901  3.1444
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.586974   0.100532   25.73   <2e-16 ***
## alcohol     0.314294   0.009465   33.21   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7931 on 4667 degrees of freedom
## Multiple R-squared:  0.1911, Adjusted R-squared:  0.1909
## F-statistic:  1103 on 1 and 4667 DF,  p-value: < 2.2e-16
```

It appears that alcohol has the greatest effect on quality with an r-squared value of 0.19 followed by density with an r-squared value of 0.099. By themselves, they are not that large.

```
##
## Call:
## lm(formula = quality ~ alcohol + log10(residual.sugar), data = wine.model)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3829 -0.5517  0.0052  0.4685  3.0364
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            1.97764    0.11852  16.686   <2e-16 ***
## alcohol                0.35380    0.01026  34.483   <2e-16 ***
## log10(residual.sugar)  0.29884    0.03152   9.482   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7856 on 4666 degrees of freedom
## Multiple R-squared:  0.2064, Adjusted R-squared:  0.2061
## F-statistic: 606.8 on 2 and 4666 DF,  p-value: < 2.2e-16
```



After looking at a few different formulas, `quality ~ log10(residual.sugar) + alcohol` appears to be the most representative of the data with an r-squared of 0.21. It is not a strong model,which can be seen in the plot.

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 appear to have more alcohol and lower density. When the factor, sweetness is overlayed, some clear patterns are visible. The most obvious patterns are visible when residual.sugar is included because it is used to construct the sweetness variable. However it is still helpful in identifying the source of the bimodal shape seen throughout the analyses.

When quality is overlayed the relationships with alcohol and density become even more visible with higher alcohol and lower density wines receiving higher quality ratings. While there does not appear to be any relationship between quality and sugar, the relationships between quality and alcohol appears to be enforced by faceting the plots over sweetness. This makes sense since the sweetness categories are used when rating wine quality.

Were there any interesting or surprising interactions between features?

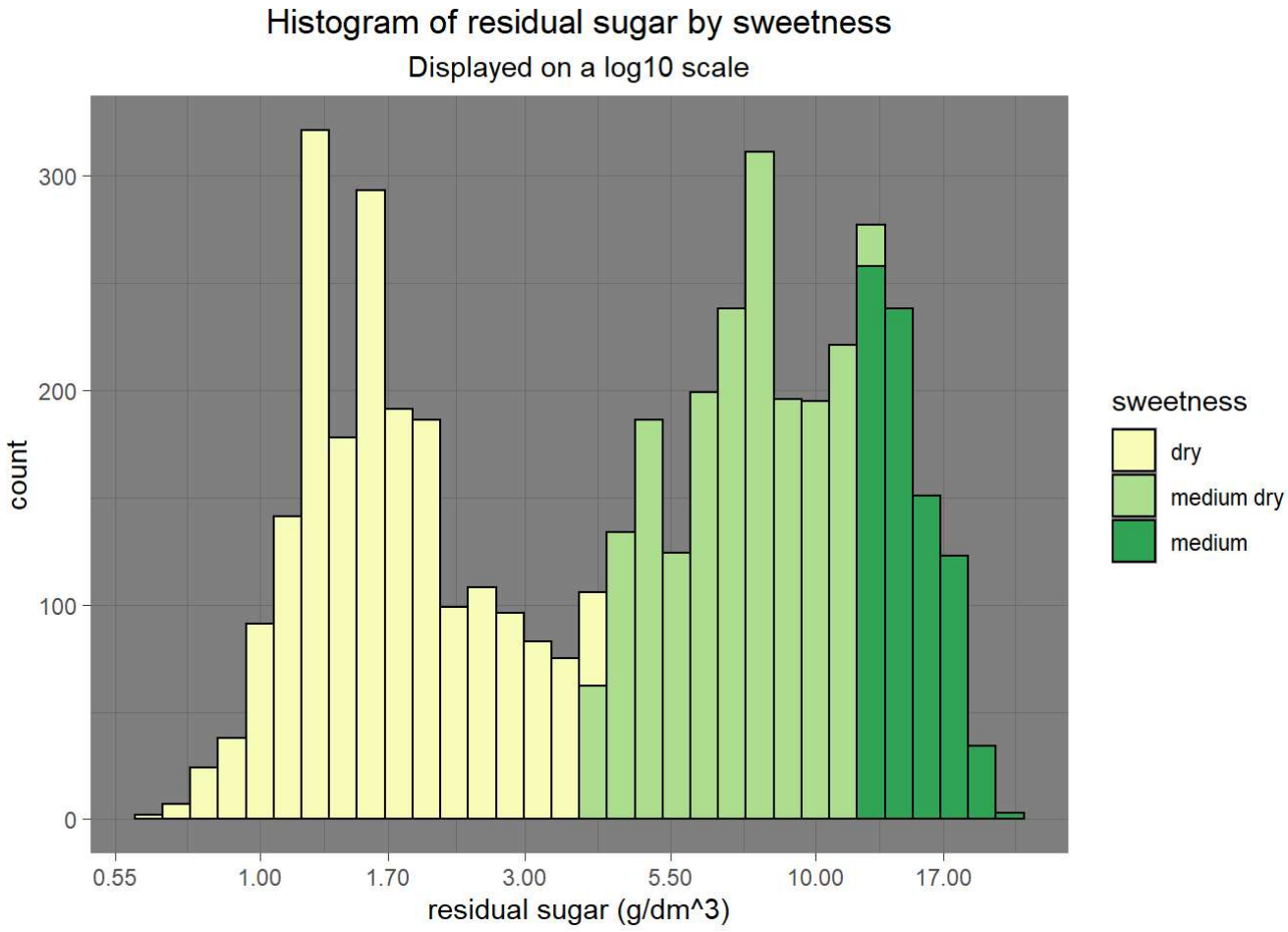
It was surprising that residual.sugar did not seem to affect quality. However, splitting the data into the sweetness groupings using residual.sugar does appear to strengthen a number of the relationships with quality.

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

I did create a linear model for quality using the log10 transformed residual.sugar and alcohol. It appears to capture some but not all of the shape of the data when looking at the associated plot, accounting for 21% of the variance in quality. The model could be improved by including additional variables and/or performing additional transformations.

Final Plots and Summary

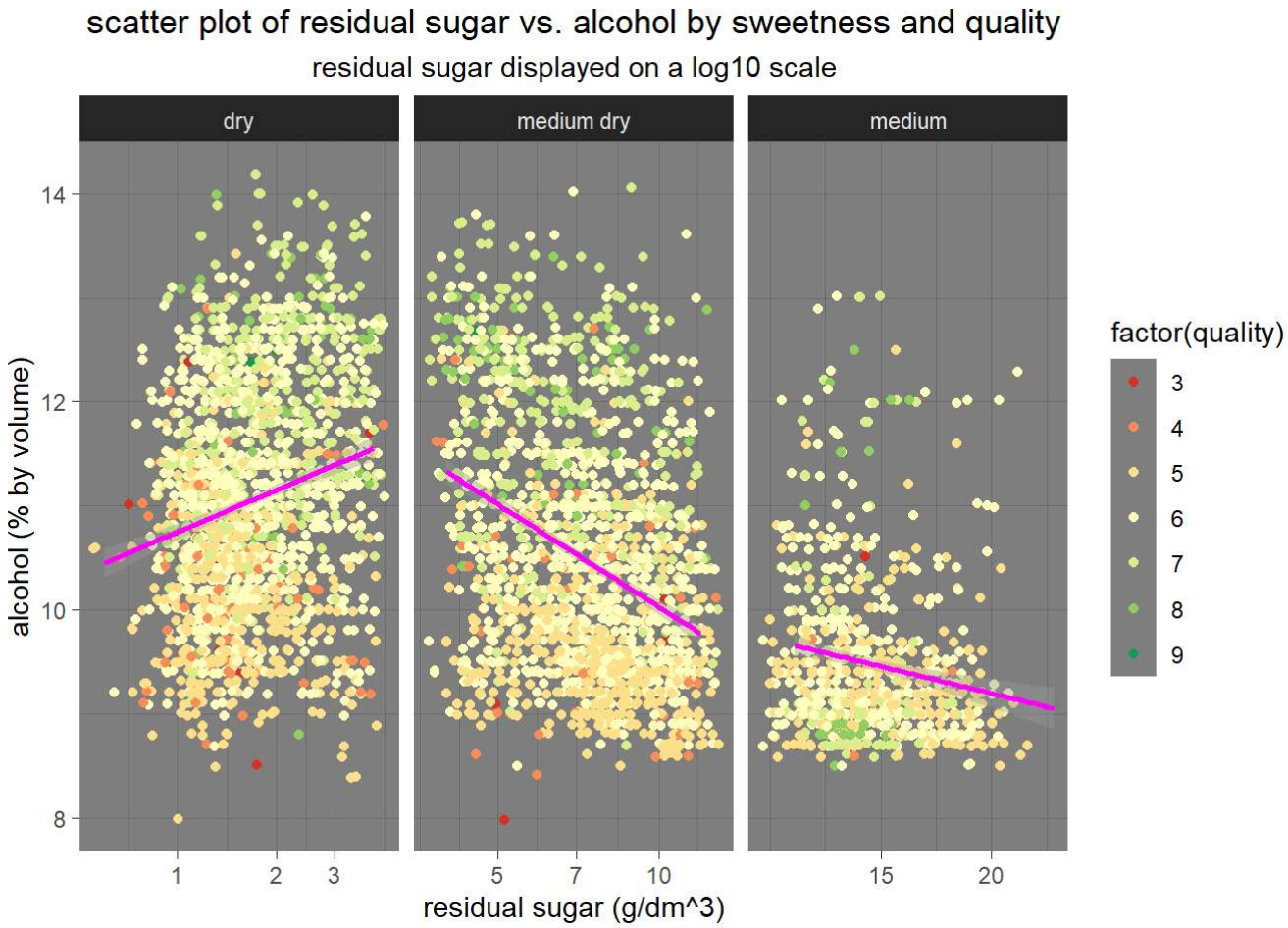
Plot One



Description One

This plot uses the filtered dataset used for modeling. The residual sugar distribution displayed on a log10 scale by sweetness category explains the bimodal distribution identified during the univariate analysis. The left mode is the dry sweetness category. the left mode actually appears to be two modes coving the medium dry and medium sweetness categories.

Plot Two




```
##
##  Pearson's product-moment correlation
##
## data:  log10(subset(wine.model, wine.model$sweetness == "dry")$residual.sugar) and subset(win
e.model, wine.model$sweetness == "dry")$alcohol
## t = 9.1355, df = 1975, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1586769 0.2432793
## sample estimates:
##       cor
## 0.2013536
```

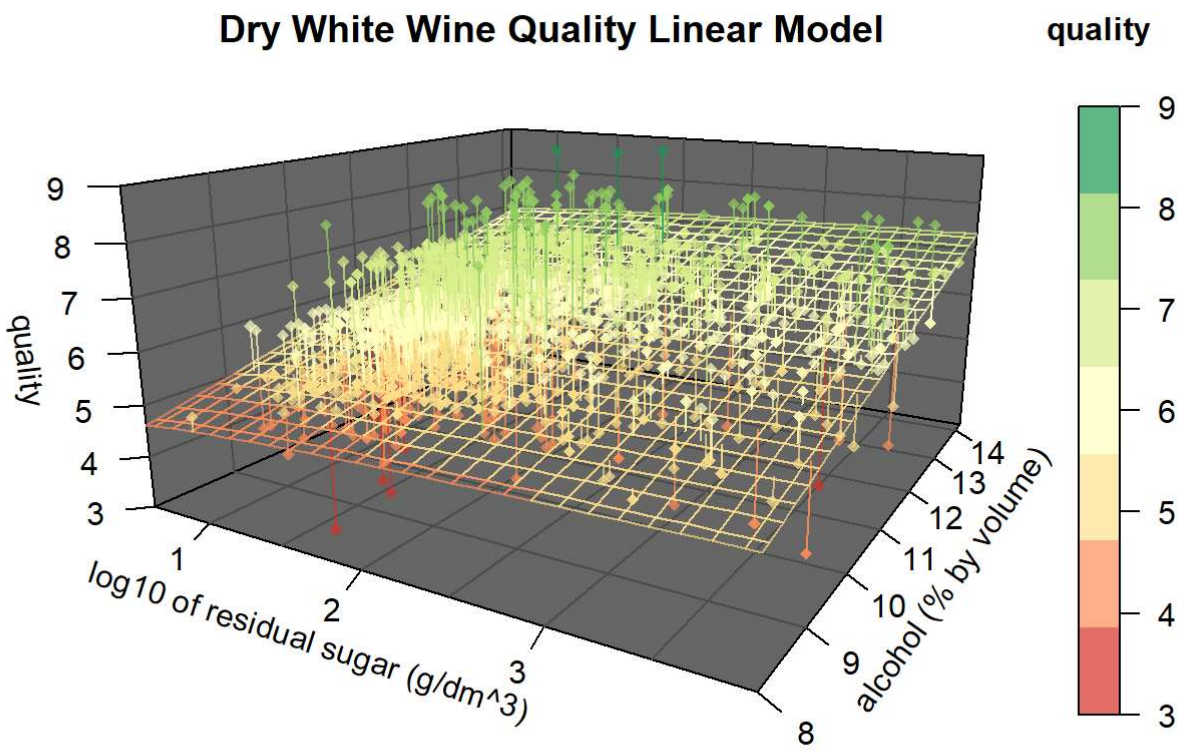
```
##
##  Pearson's product-moment correlation
##
## data:  log10(subset(wine.model, wine.model$sweetness == "medium dry")$residual.sugar) and sub
set(wine.model, wine.model$sweetness == "medium dry")$alcohol
## t = -16.465, df = 1883, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3935920 -0.3146383
## sample estimates:
##        cor
## -0.3547475
```

```
##
##  Pearson's product-moment correlation
##
## data:  log10(subset(wine.model, wine.model$sweetness == "medium")$residual.sugar) and subset
(wine.model, wine.model$sweetness == "medium")$alcohol
## t = -4.414, df = 805, p-value = 1.153e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.22039932 -0.08562043
## sample estimates:
##        cor
## -0.1537248
```

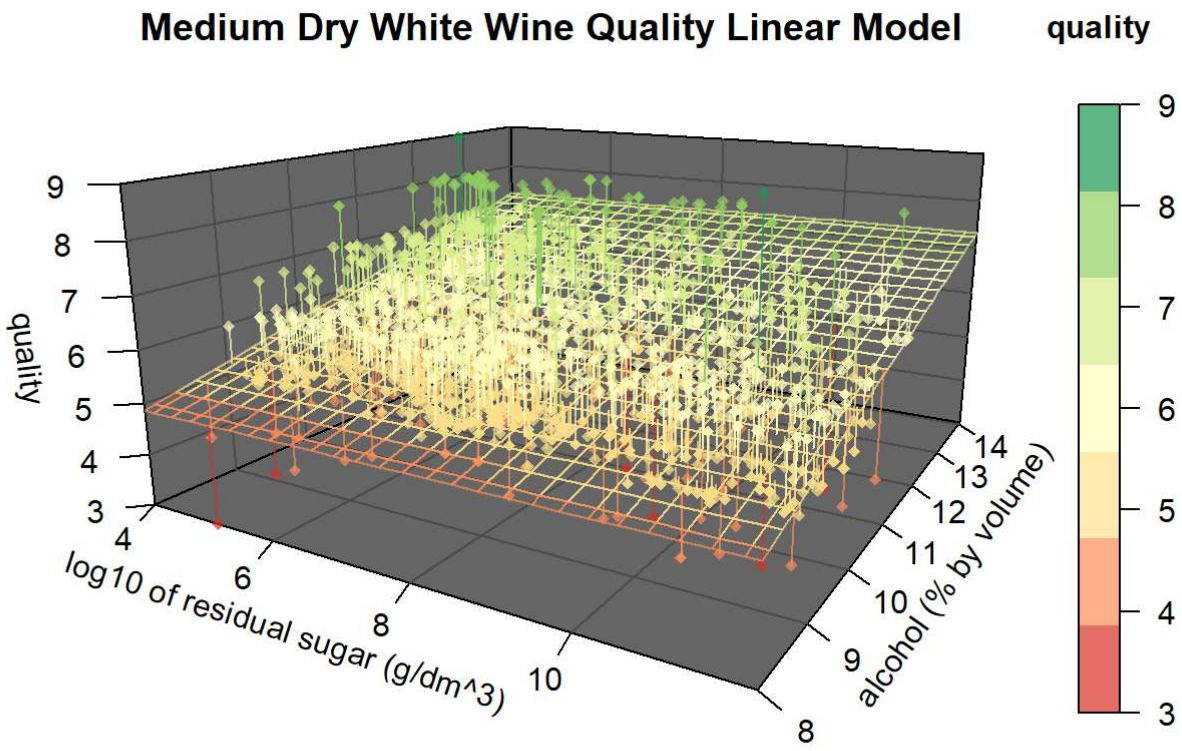
Description Two

The highest quality wines tend to have higher alcohol content regardless of residual sugar or sweetness category. The majority of the lowest quality wines (3 and 4) are in the dry and medium dry wine categories, which could be due to the the skew in the residual sugar distribution. Wines in the medium sweetness group tend to have less alcohol. It is also noticeable that the correlation patterns appear to change between sweetness categories. There is a very weak positive correlation visible in the dry category of 0.201, but medium dry and medium wines have negative correlations between alcohol and residual.sugar with strengths of -0.355 and -0.154 respectively. The strongest correlation appears in the medium sweetness group with residual sugar decreasing as alcohol increases.

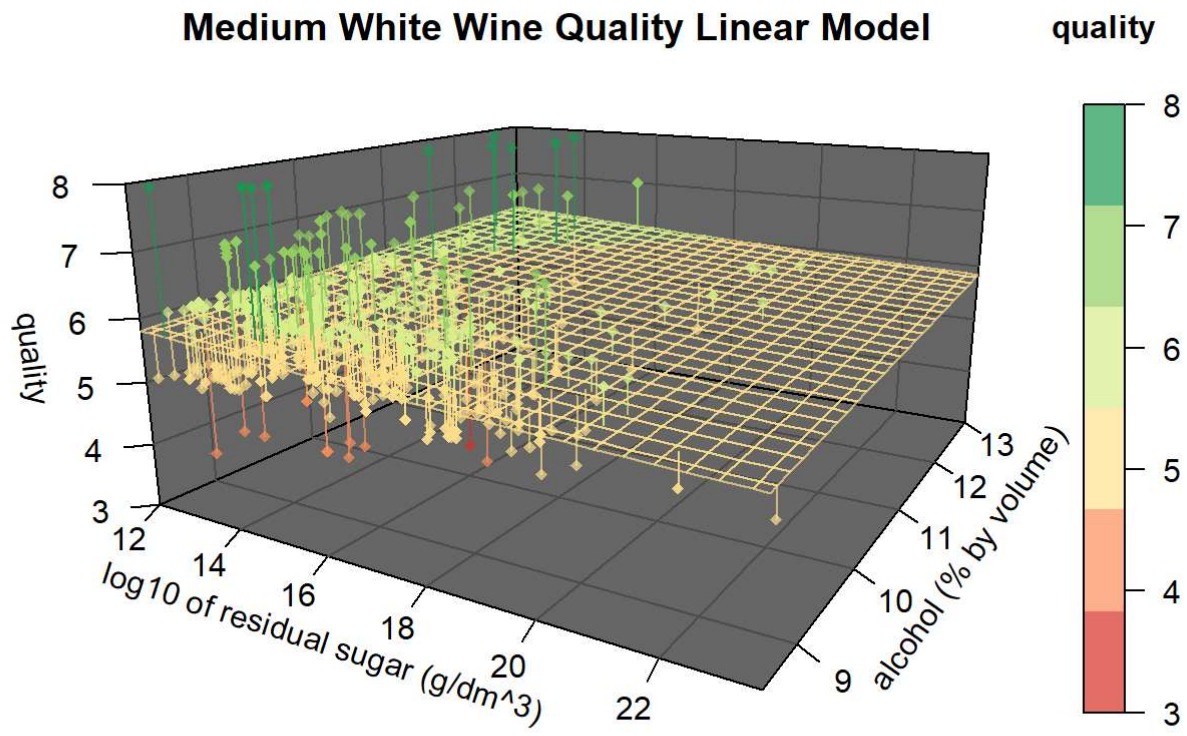
Plot Three




```
##
## Call:
## lm(formula = formula)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3933 -0.5321 -0.0114  0.5091  2.7608
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             1.8027     0.1832   9.841  < 2e-16 ***
## log10(residual.sugar)   0.5614     0.1127   4.980 6.92e-07 ***
## alcohol                 0.3663     0.0168  21.797  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8093 on 1974 degrees of freedom
## Multiple R-squared:  0.223,  Adjusted R-squared:  0.2222
## F-statistic: 283.3 on 2 and 1974 DF,  p-value: < 2.2e-16
```




```
##
## Call:
## lm(formula = formula)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3068 -0.4704  0.0015  0.4821  3.1327
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            1.59081    0.23274   6.835  1.1e-11 ***
## log10(residual.sugar)  0.13945    0.14045   0.993    0.321
## alcohol                0.39746    0.01524  26.077  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7344 on 1882 degrees of freedom
## Multiple R-squared:  0.2871, Adjusted R-squared:  0.2864
## F-statistic:   379 on 2 and 1882 DF,  p-value: < 2.2e-16
```




```
##
## Call:
## lm(formula = formula)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8624 -0.7313  0.1277  0.3252  2.2616
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            6.49579    0.69096   9.401  < 2e-16 ***
## log10(residual.sugar) -1.40615    0.47637  -2.952  0.00325 **
## alcohol                0.09757    0.03521   2.771  0.00572 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7929 on 804 degrees of freedom
## Multiple R-squared:  0.02352,    Adjusted R-squared:  0.02109
## F-statistic: 9.682 on 2 and 804 DF,  p-value: 7.002e-05
```

Description Three

This plot displays the mathematical model of the formula, `quality ~ log10(residual.sugar) + alcohol` for each sweetness grouping.The model is not that strong with an R-squared values of 0.223, 0.2871, and 0.2871 for dry, medium dry, and medium wines respectively. However, it does indicate that higher quality white wines tend to have higher alcohol content by volume. As was seen in the 2D plot, the residual sugar - alcohol correlation tends to be negative with the exception of the dry sweetness group. The variation in correlation across the groups suggests that the overall model is not linear.

Reflection

The wine quality data set contains almost 4,900 wines with 12 attributes. 11 of the attributes represent measurements of the chemical properties of each wine. The last attribute is an average of three quality ratings, judged on a 10 point scale. Throughout this analysis, my aim was to identify which of the measurements contribute the most to quality, as well as to each other. I began by plotting and refining the plots for the distribution of each quantitative variable. The bimodal shape of the log10 residual sugar measurement led me to create the sweetness category based on the EU definition. This made sense since each of the groups are judged separately. The limitation with the implementation for this analysis is that tartaric acid is typically used with residual sugar measurements to classify the sweetness. This analysis uses residual sugar alone.

The European Union defines the sweetness terms of wine as follows:

- Dry < 4 g/L of sugar
- Medium dry 4-12 g/L of sugar
- Medium 12-45 g/L of sugar
- Sweet > 45 g/L of sugar

Official Journal of the European Union (https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2006:384:0038:0043:EN:PDF)

Even without using tartaric acid to categorize sweetness, there are some clear differences between the groups, not only in the strength and direction of the correlations, but also in the amount of variance explained by the model. This leads me to believe that the model is not linear. The model for medium dry wines alone has the highest r squared value at 0.28. This still accounts for less than a third of the variance in a portion of the data.

Further analyses could strengthen the mathematical model developed here by more accurately categorizing the sweetness groups. This should be done using tartaric acid, captured by the variable, citric.acid. Other regression functions should also be investigated since this model does not appear to fit well to a straight line. Additional attributes and/or transformations should also be tested since there appear to be a complex set of relationships contributing to the quality of a wine within a specific sweetness category.