#Exploratory Data Analysis for the Red Wine Quality Data Set by Joshua Maine ##10/6/18
==========

#INTRODUCTION

This dataset interest me as its similar to some further data analysis I want to do on data on beer. My longest running hobby, for almost a decade, has been homebrewing, and there is more similarities to red wines and beer than whites.

#Univariate Explorations and Plots

To get an idea of the data and gain a better understanding, I'm going to work through a few Univariate explorations. I'll start by getting an idea of the table structure and looking at the summary.

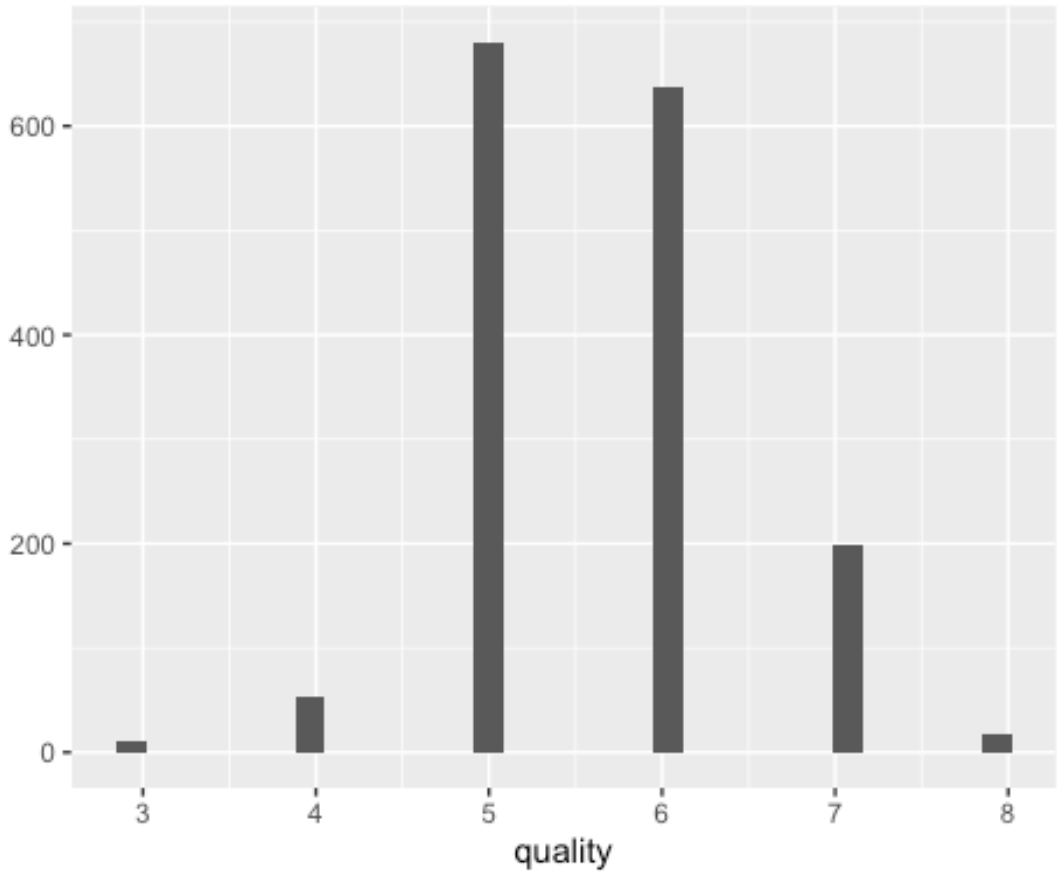
```
## 'data.frame':    1599 obs. of  13 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8
7.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65
0.58 0.5 ...
##  $ citric.acid         : num  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 data set as 1599 records with 13 variables. Taking a full look at the variables, it has:

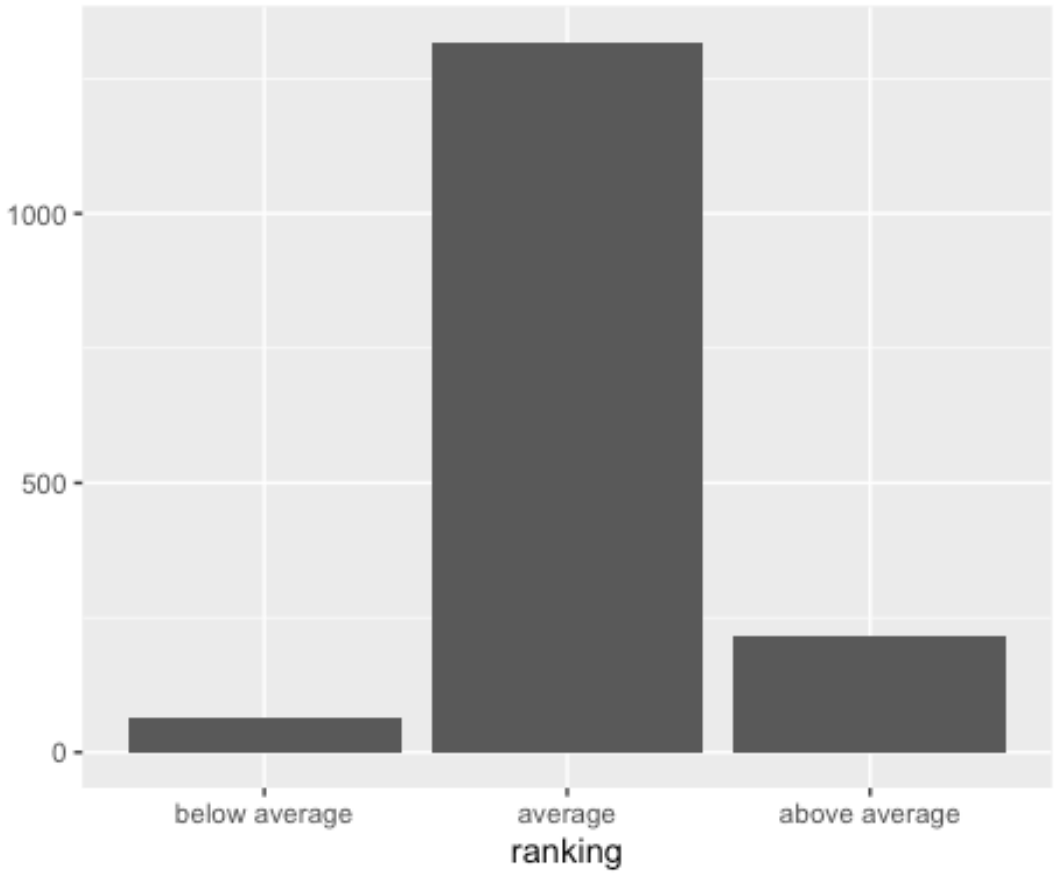
X appears to be the wine tested, the removal of name and other identifying information, shows this is likely from a judging or evaluation of the wines in the data set.

```
##        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
total.sulfur.dioxide
##  Min.   : 0.900   Min.   :0.01200   Min.   : 1.00       Min.   :
6.00
##  1st Qu.: 1.900   1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.:
22.00
##  Median : 2.200   Median :0.07900   Median :14.00       Median :
38.00
##  Mean   : 2.539   Mean   :0.08747   Mean   :15.87       Mean   :
46.47
##  3rd Qu.: 2.600   3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.:
62.00
##  Max.   :15.500   Max.   :0.61100   Max.   :72.00       Max.
:289.00
##     density             pH          sulphates         alcohol
##  Min.   :0.9901   Min.   :2.740   Min.   :0.3300   Min.   : 8.40
##  1st Qu.:0.9956   1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50
##  Median :0.9968   Median :3.310   Median :0.6200   Median :10.20
##  Mean   :0.9967   Mean   :3.311   Mean   :0.6581   Mean   :10.42
##  3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10
##  Max.   :1.0037   Max.   :4.010   Max.   :2.0000   Max.   :14.90
##     quality
##  Min.   :3.000
##  1st Qu.:5.000
##  Median :6.000
##  Mean   :5.636
##  3rd Qu.:6.000
##  Max.   :8.000
```

With all the variables, this is most certainly data used in comparing and evaluting different red wines. Looking at the summary data there are things I want to plot and get a better idea of how the data is grouped and to find any outliers for correction.

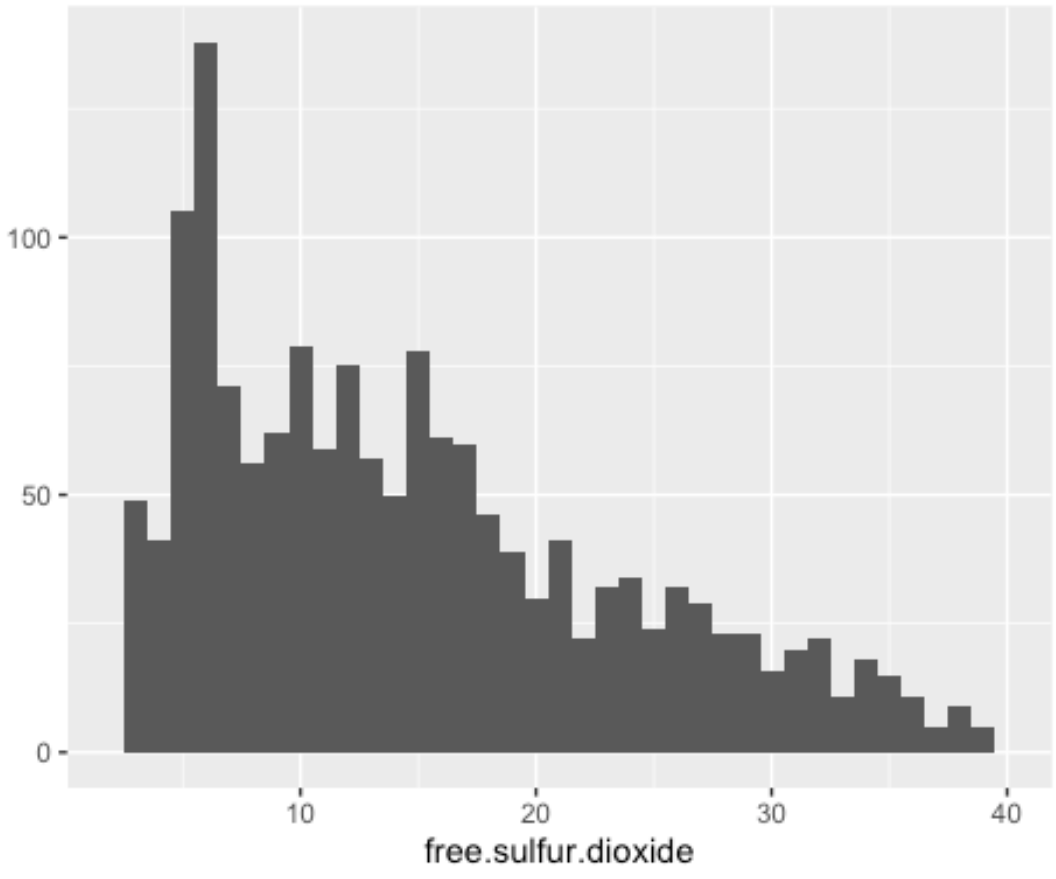


Most of the quality values exist as a 5 or a 6, so I may group these a little tigther to sort the standouts on both sides. Grouping the values by pairs to have below average, average, above average is shown below.



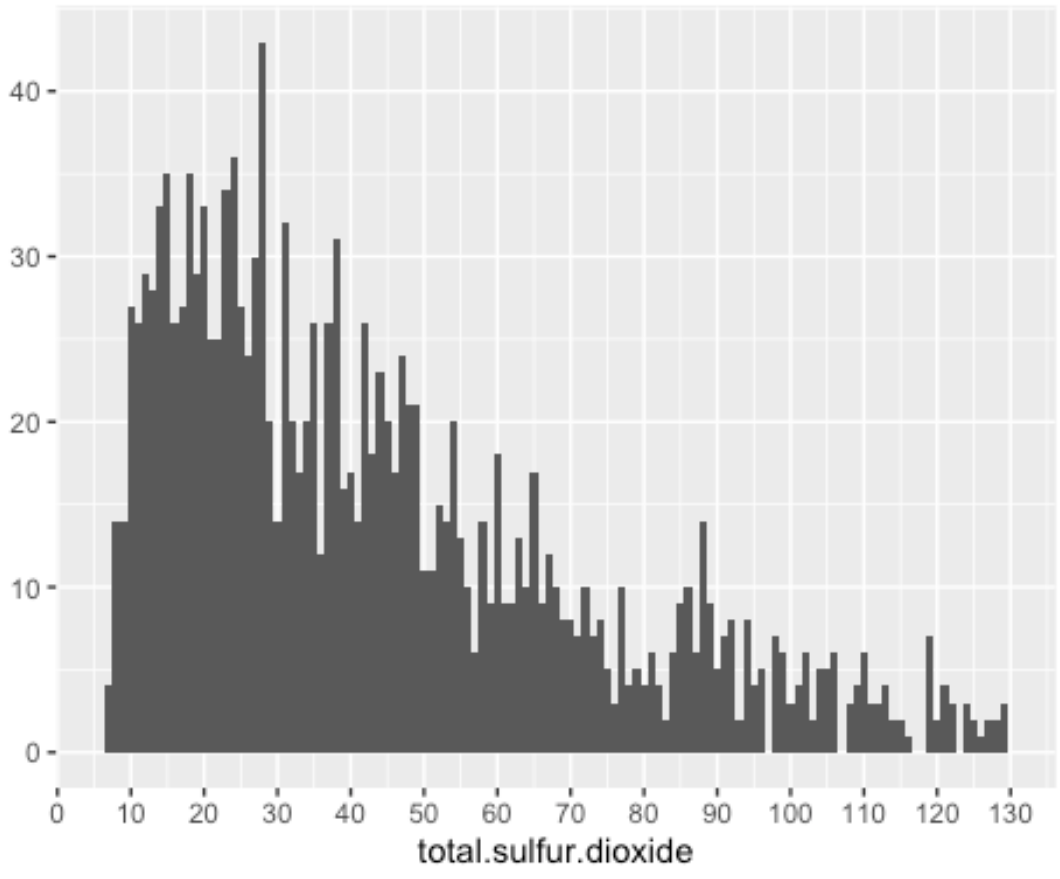

```
summary(wd$ranking)

## below average       average above average
##            63          1319           217
```

Looking at possibly related variables for trends or relations seems like a decent assumption. I'm going to look at sulfur and acidity related values first. THe relationships between acidity and pH, sugar and alchol content, and alchol and density as I suspect density in this case is referring to the gravity value of the wine.




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

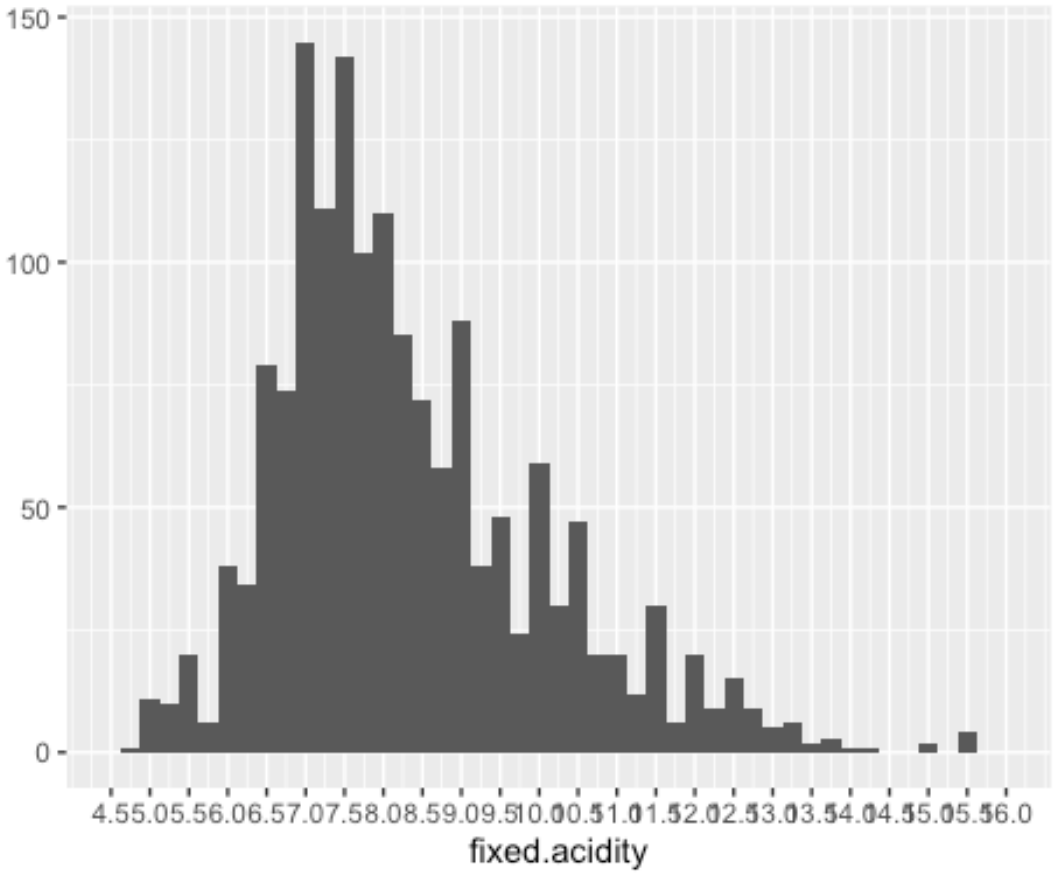
Limiting the x axis lead to removal of 40 records



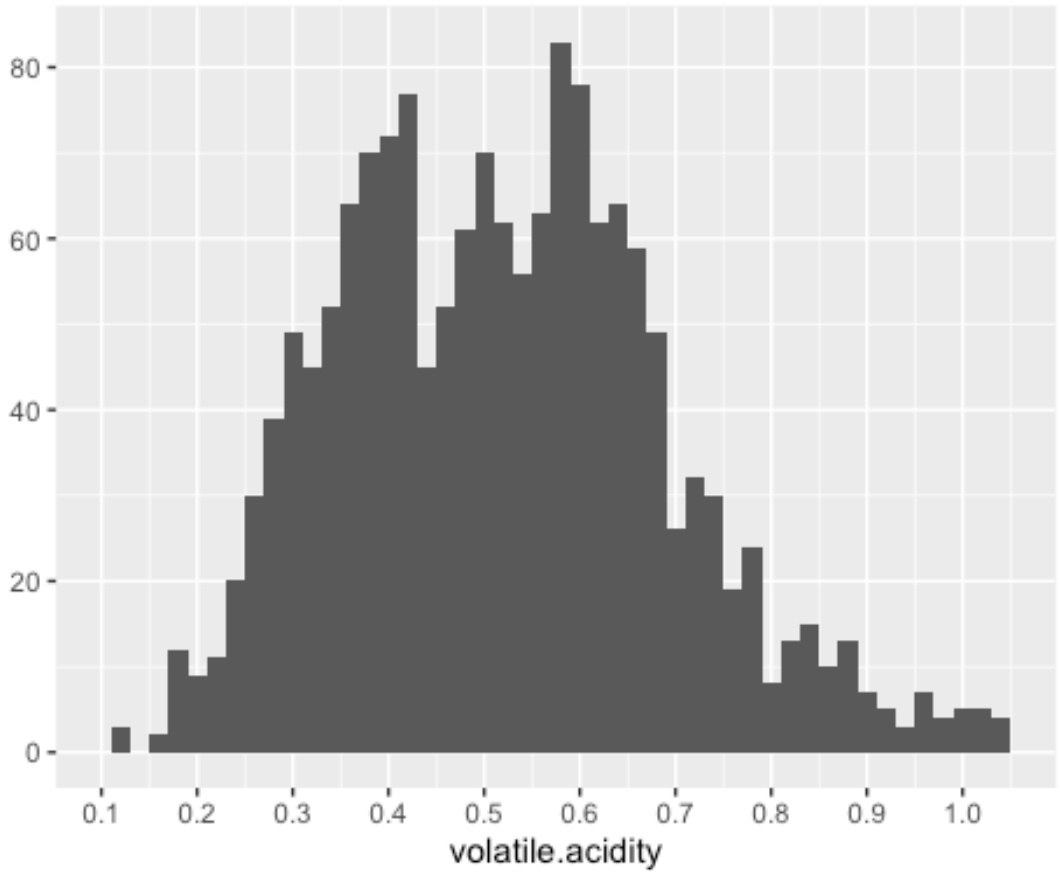

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00   22.00   38.00   46.47   62.00  289.00
```

Limiting the x axis lead to removal of 41 records

The next few charts relate to the acidity measures.

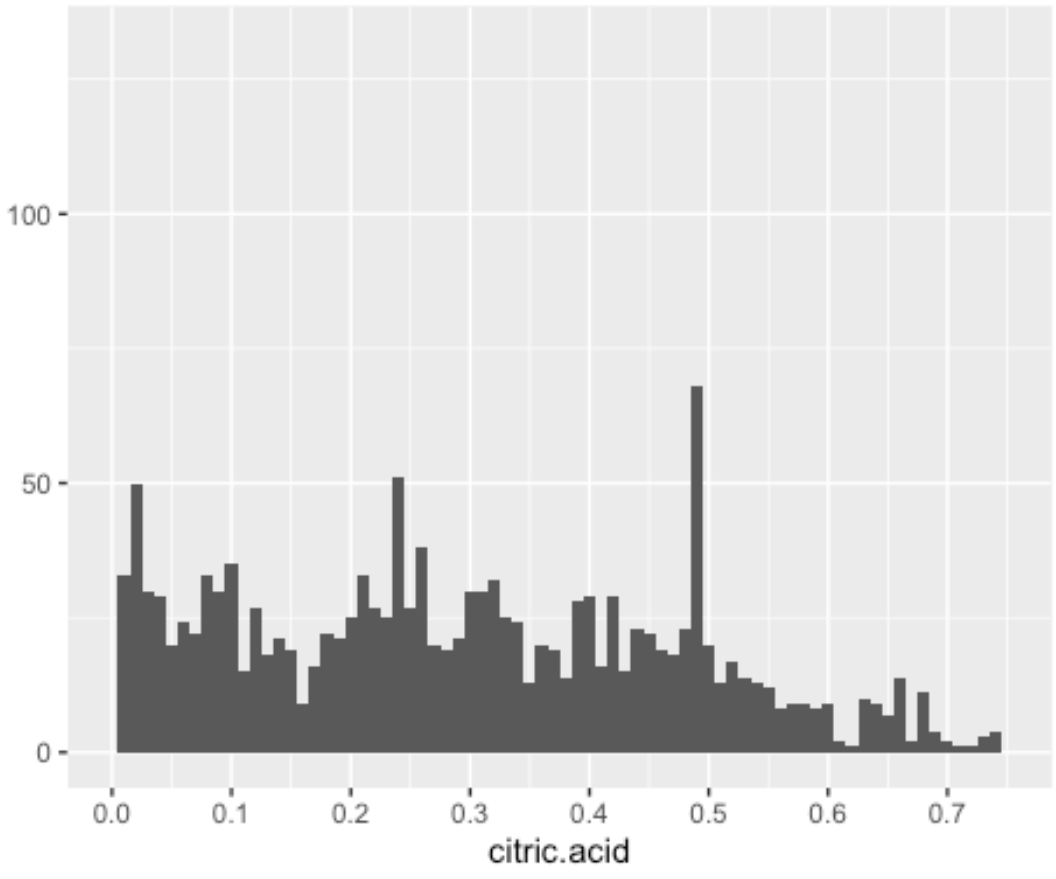



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.60    7.10    7.90    8.32    9.20   15.90
```



Limiting the x axis lead to removal of 10 records.

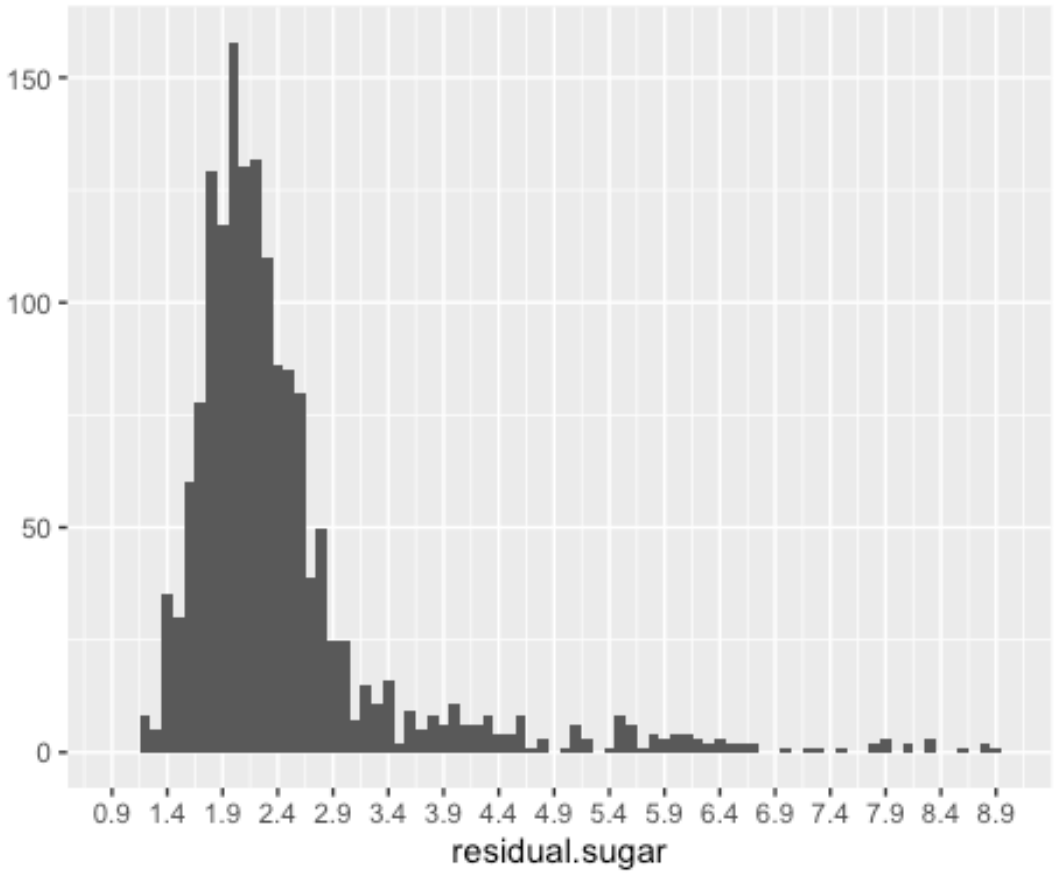
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1200  0.3900  0.5200  0.5278  0.6400  1.5800
```




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

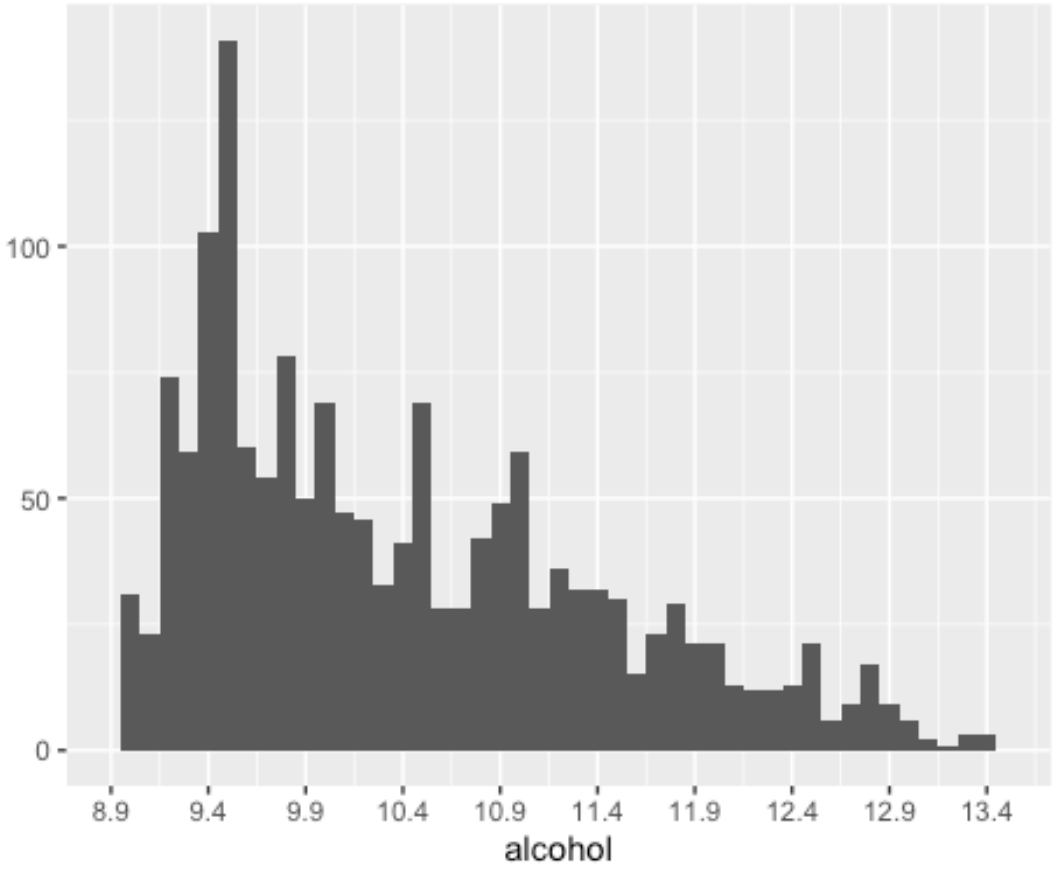
Limiting the x axis lead to removal of 6 records.

The next set of plots deal with residual sugars, alchol, and density



Limiting the x axis lead to removal of 10 records.

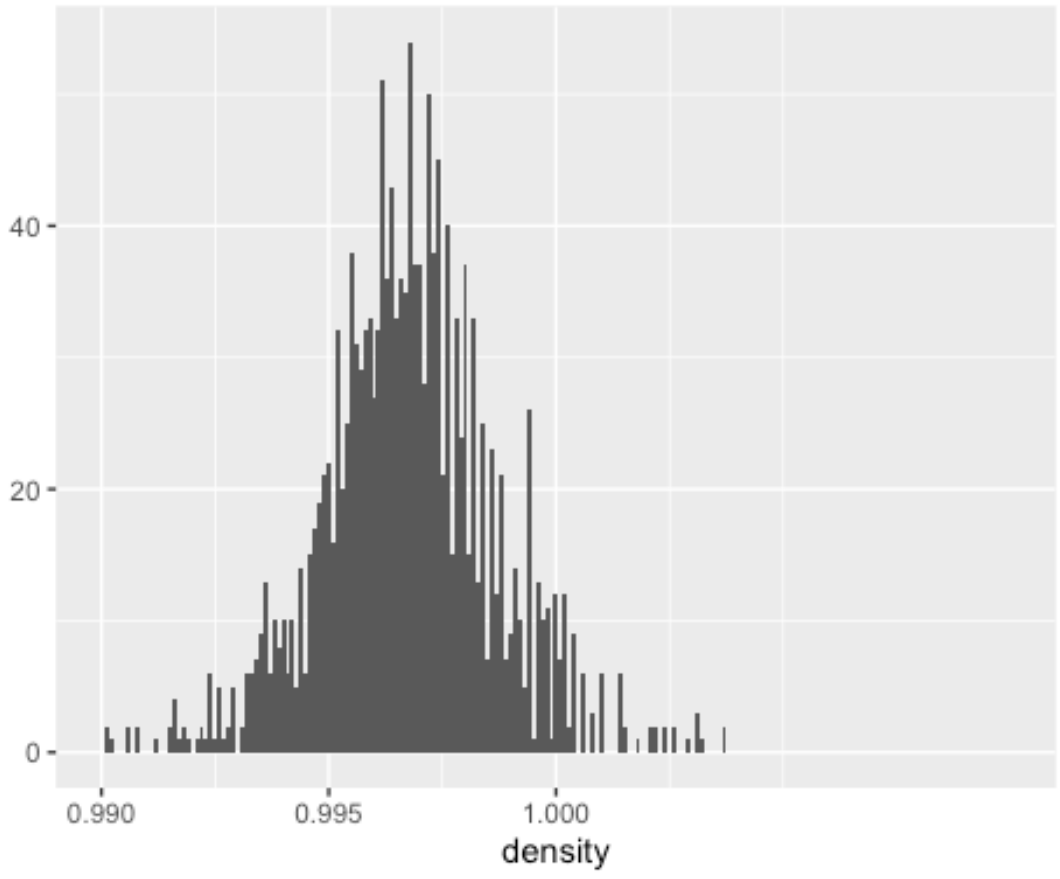
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.900   1.900   2.200   2.539   2.600  15.500
```



Limiting the x axis lead to removal of 20 records.

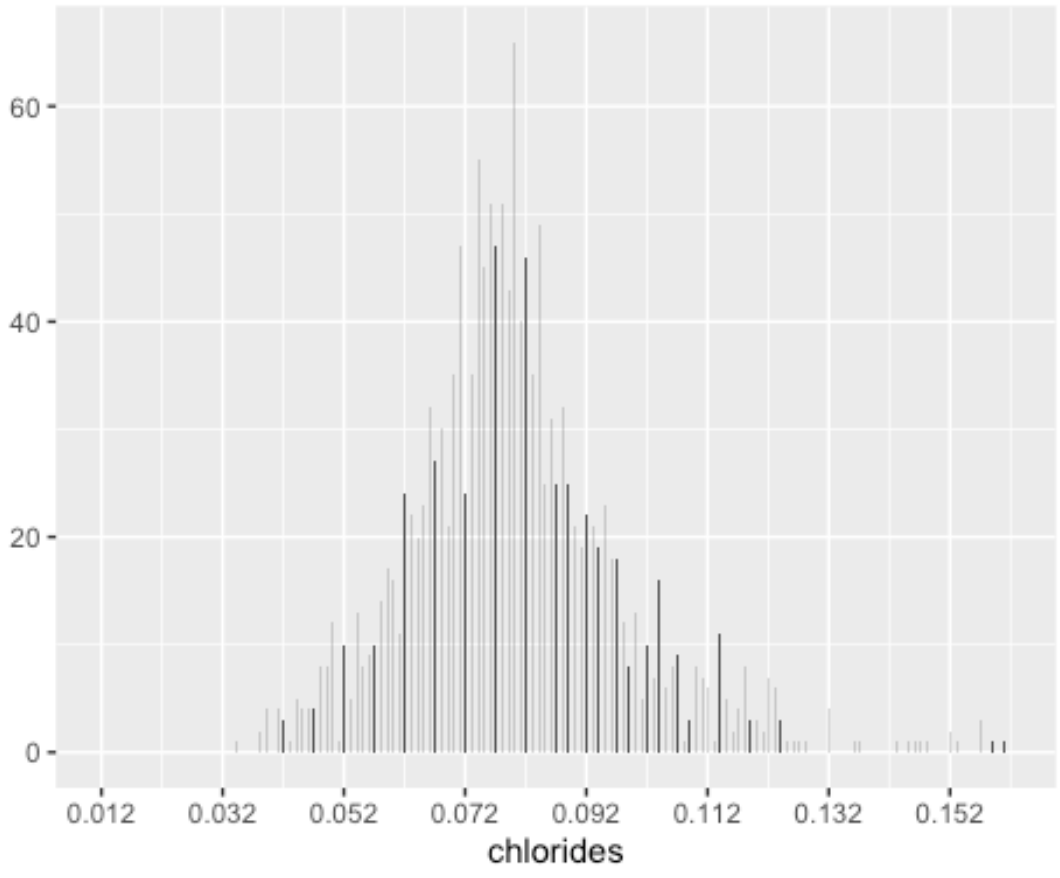
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.50   10.20   10.42   11.10   14.90
```

This is another measure related to sugar to alchol conversions in terms of fermentation potential and actual. The vinter would usually take an originial gravity reading and a final gravity, what I believe density to be, to measure fermentable effeciency.



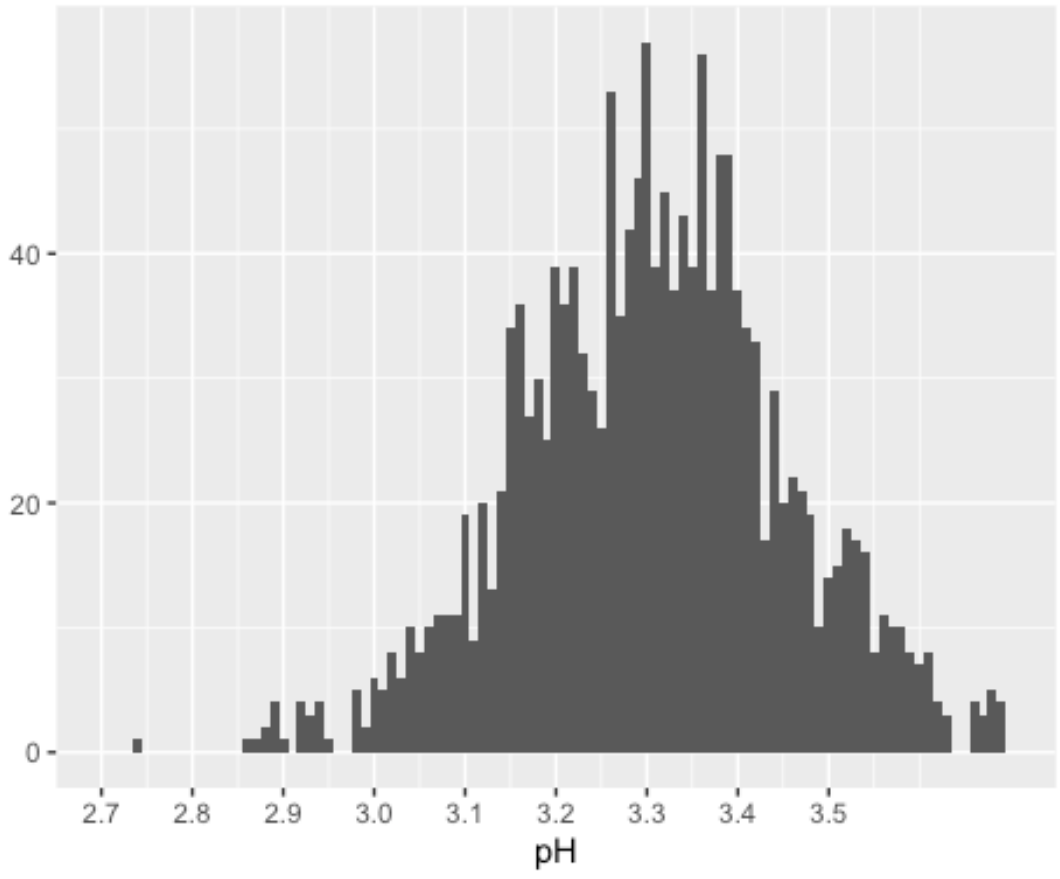

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9901  0.9956  0.9968  0.9967  0.9978  1.0037
```

The last group of variables or other measures that initially I do not know or have any expectations about their respective values.



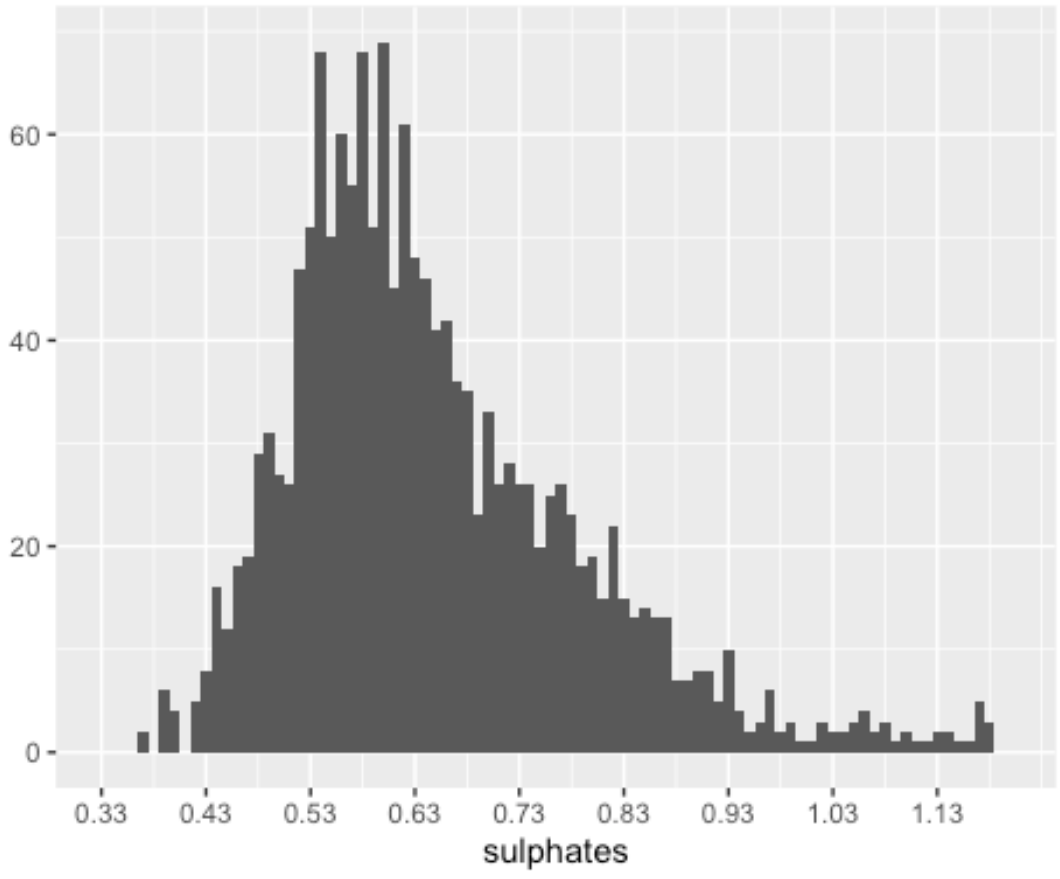
Limiting the x axis lead to removal of 59 records.

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



Limiting the x axis lead to removal of 16 records.

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



Limiting the x axis lead to removal of 18 records.

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

#Univariate Analysis

#Template Questions #————————————————————————-

##What is the structure of your dataset? The Red wine data set has 1599 items and 13 variables for each entry, one of which is a simple numerical listing of all the entries. There is one subjective value, quaility, which is how the wine was judged based on its othere values. The rest of the data set is composed of the chemical measurements of the wine. The data doesn't contain one variable that may become a lurking variable, taste notes. I

##What is/are the main feature(s) of interest in your dataset? The features that have my interest relate my an existing hobby and area of interest I already have. I'm more interested

in the fermentation related data including sugar, alchol, and density. My other area of interests is looking at the quality scores given and how it relates to the other variables.

##What other features in the dataset do you think will help support your ##investigation into your feature(s) of interest? At this point I think there will be common trends in the wines that fall into the 3 different ranking groups. Using the chemical properties, I suspect some things will be found to be indicitive of an above average wine.

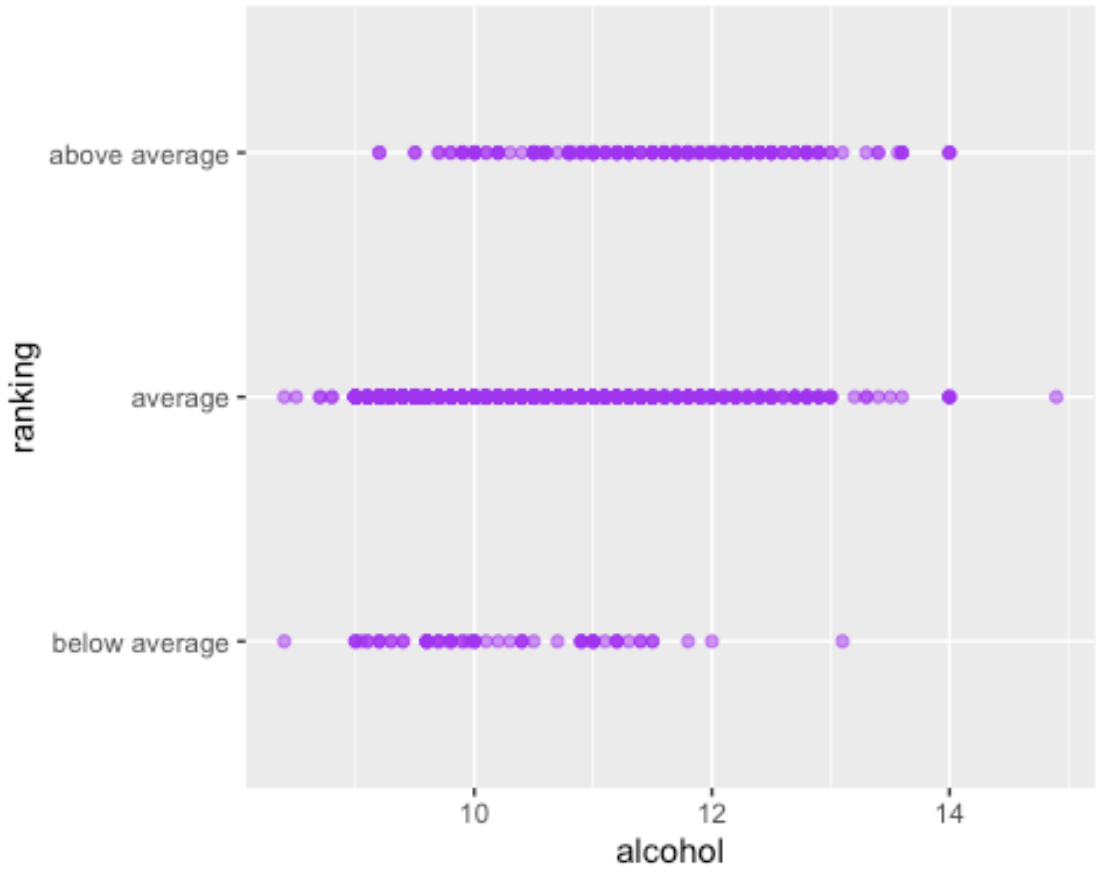
##Did you create any new variables from existing variables in the dataset? I created a ranking of the quality values into 3 categories of below average, average, and above average. I may have to create more once I get into the bivariate exploration.

##Of the features you investigated, were there any unusual distributions? Quality, density, chlorides, and volatile acidity have more uniform distributions. Both sulfur measures, sugar, alchol, fixed acidity, and sulfates have a some sort of leftward skew. pH right skewed slightly right. Surprisingly citric acid has a flatter but still slightly skewed left distribution.

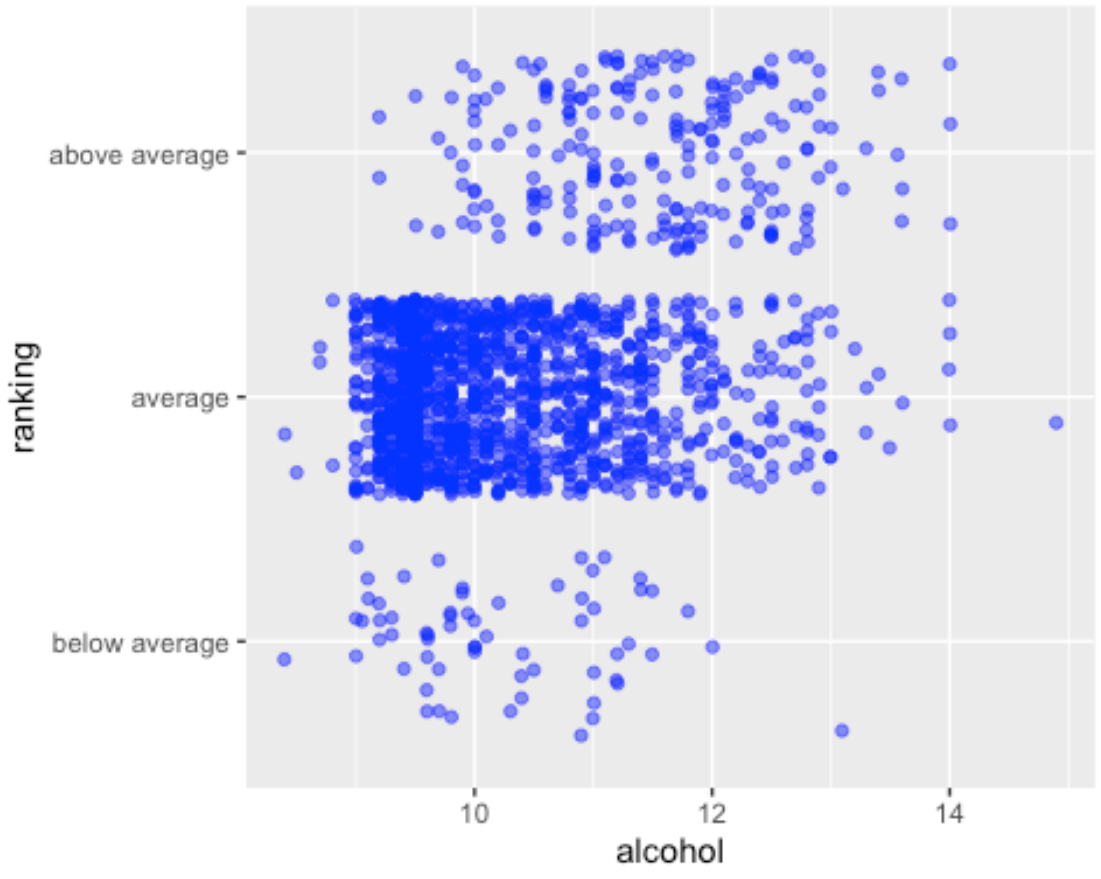
##Did you perform any operations on the data to tidy, adjust, or change the form ##of the data? If so, why did you do this? The data appears pretty tidy and easy to work with. I haven't had to change any forms yet but have to do some adjustments to compare some values more easily.

#Bivariate Plots Section

Next up, I'm going to compare different chemical properties against the subjective quality measure and against other chemical markers.



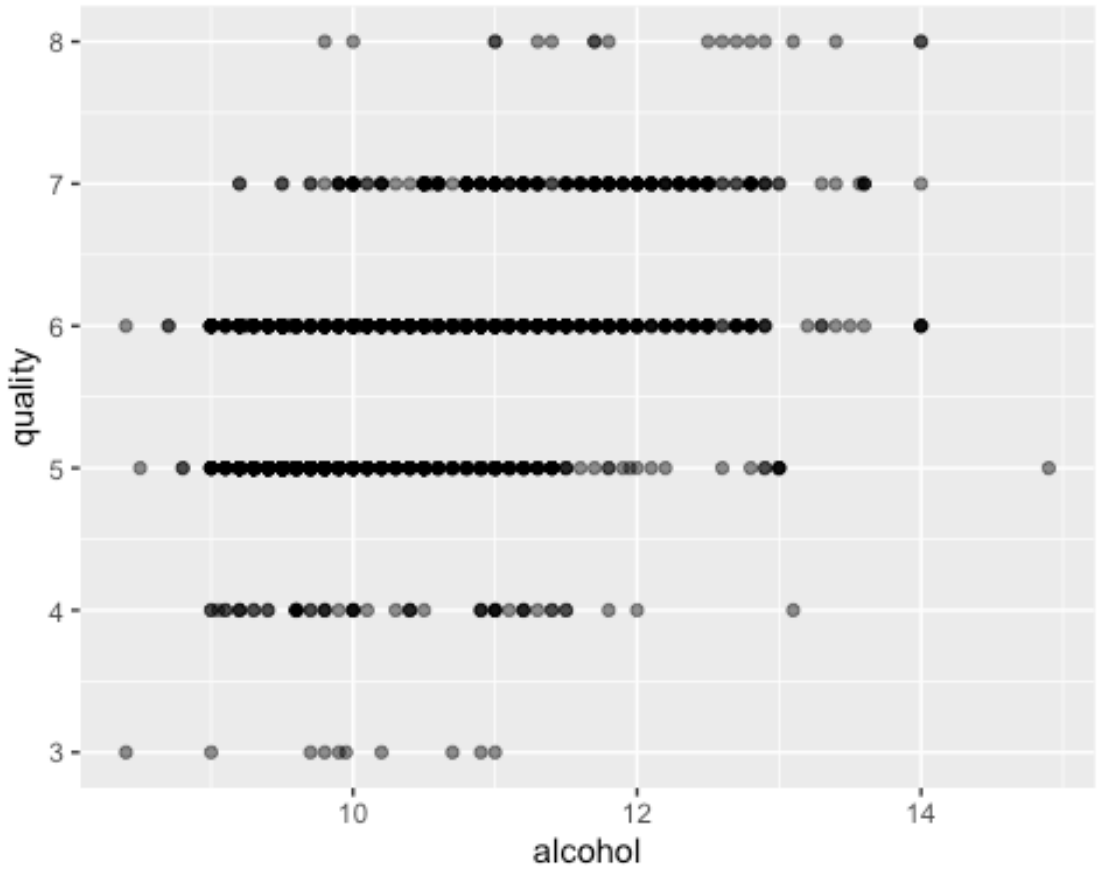
Seems a bit too tidy and ordered, going to jitter the plot to so the density of the data a bit better.

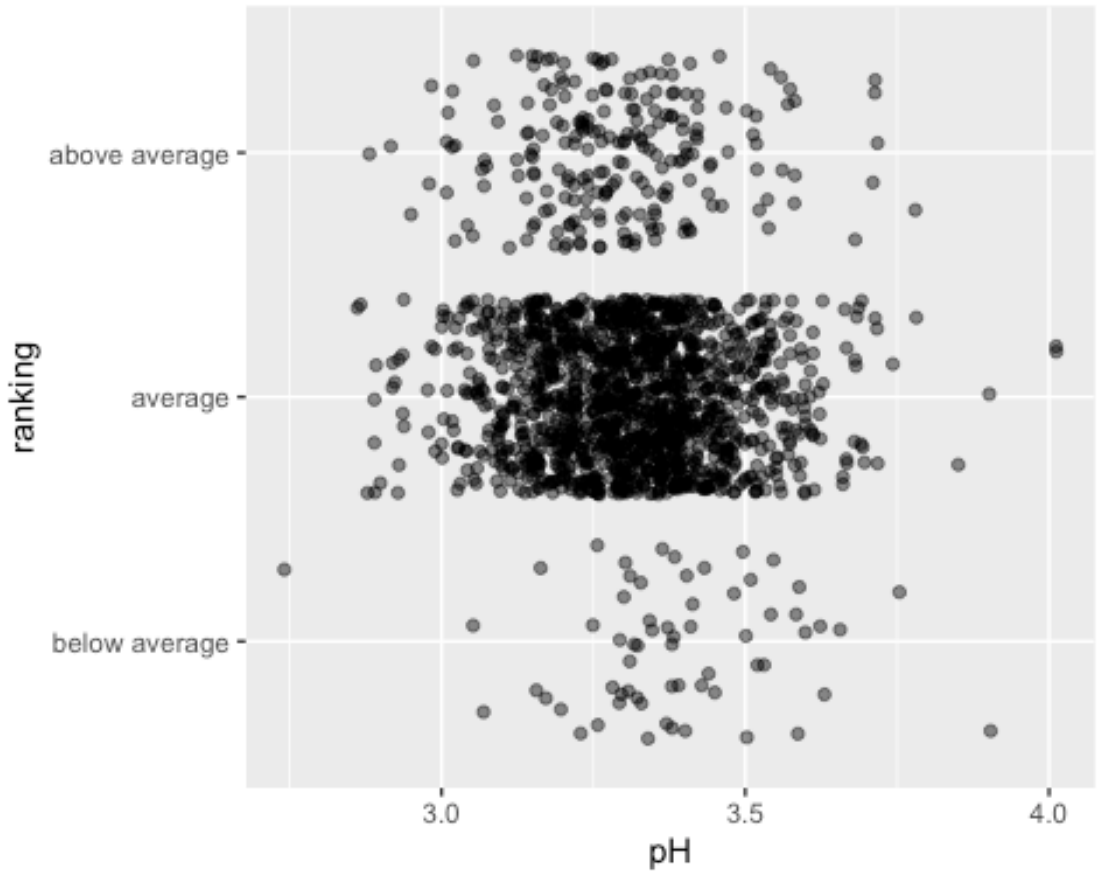


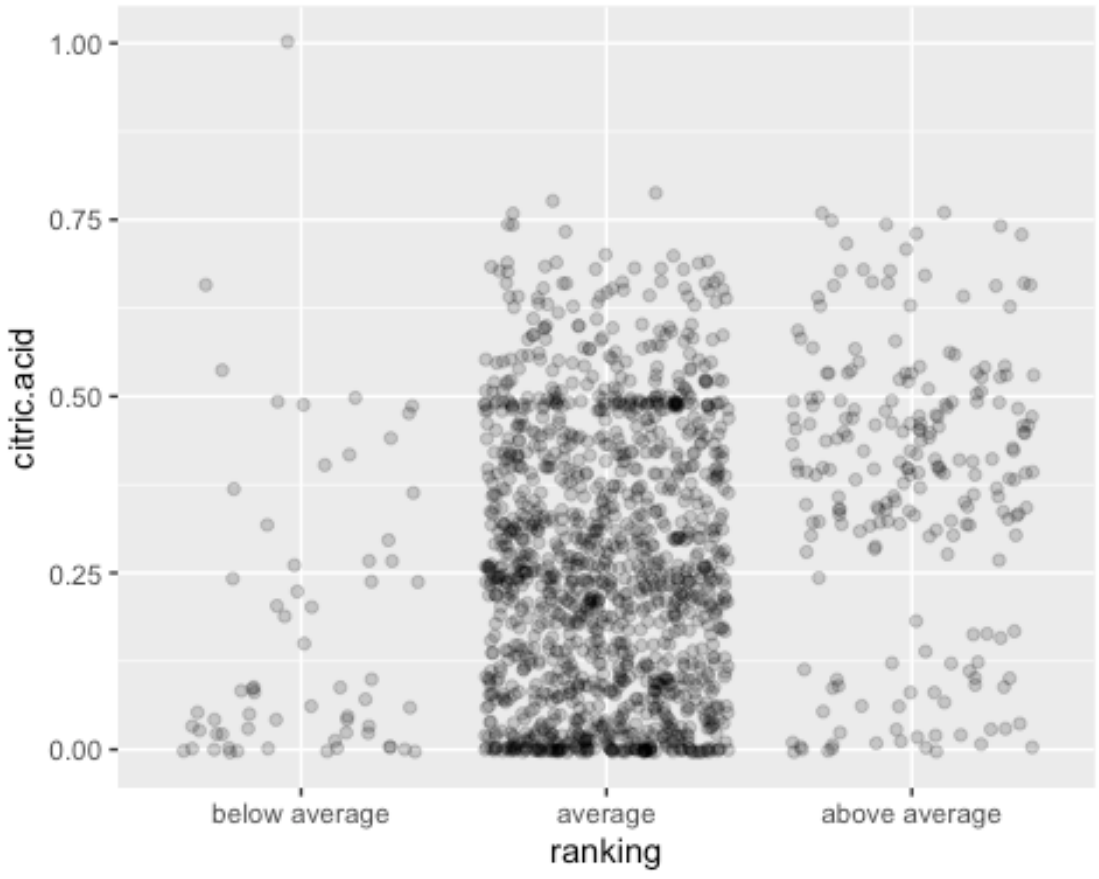
That is better in terms of relecting the data in each category.

I find it interesting that it appears the above and below average rankings had both have a more narrow band in their repective alcohol content.

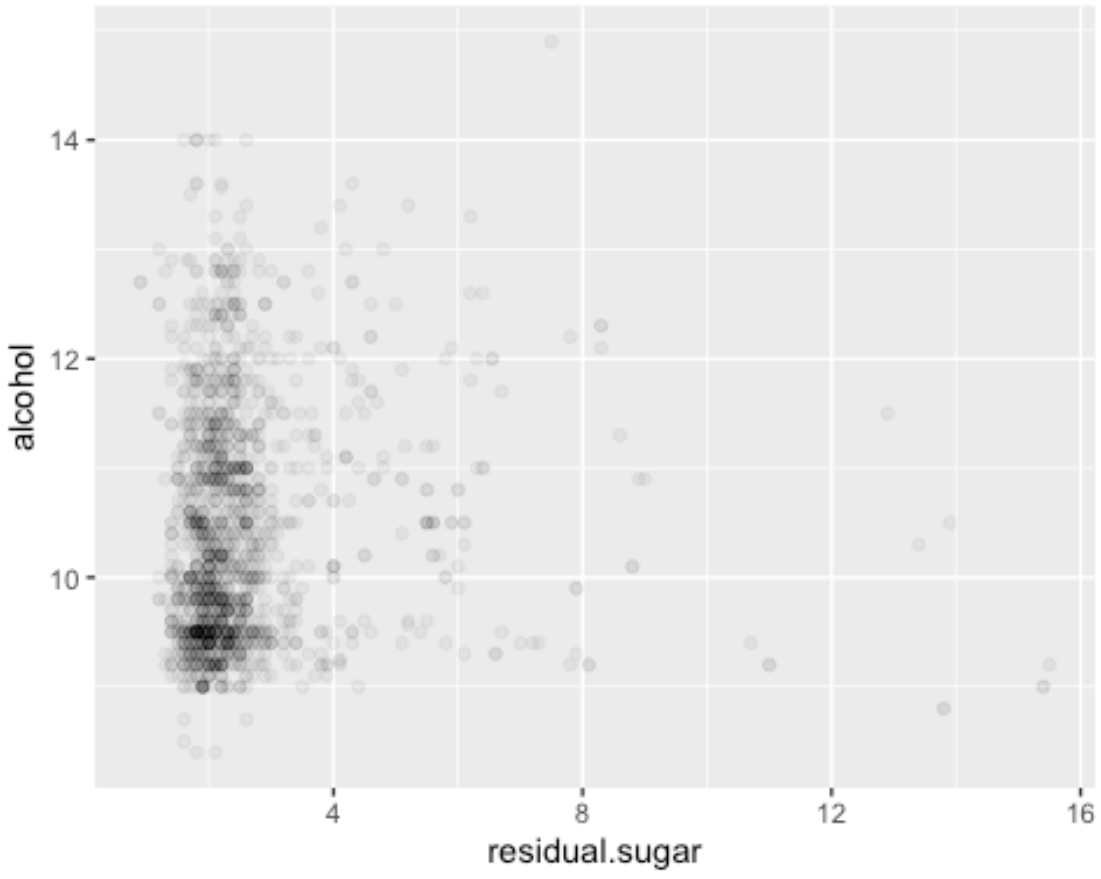
I'm going to continue plotting out the various sets I'm curious about.

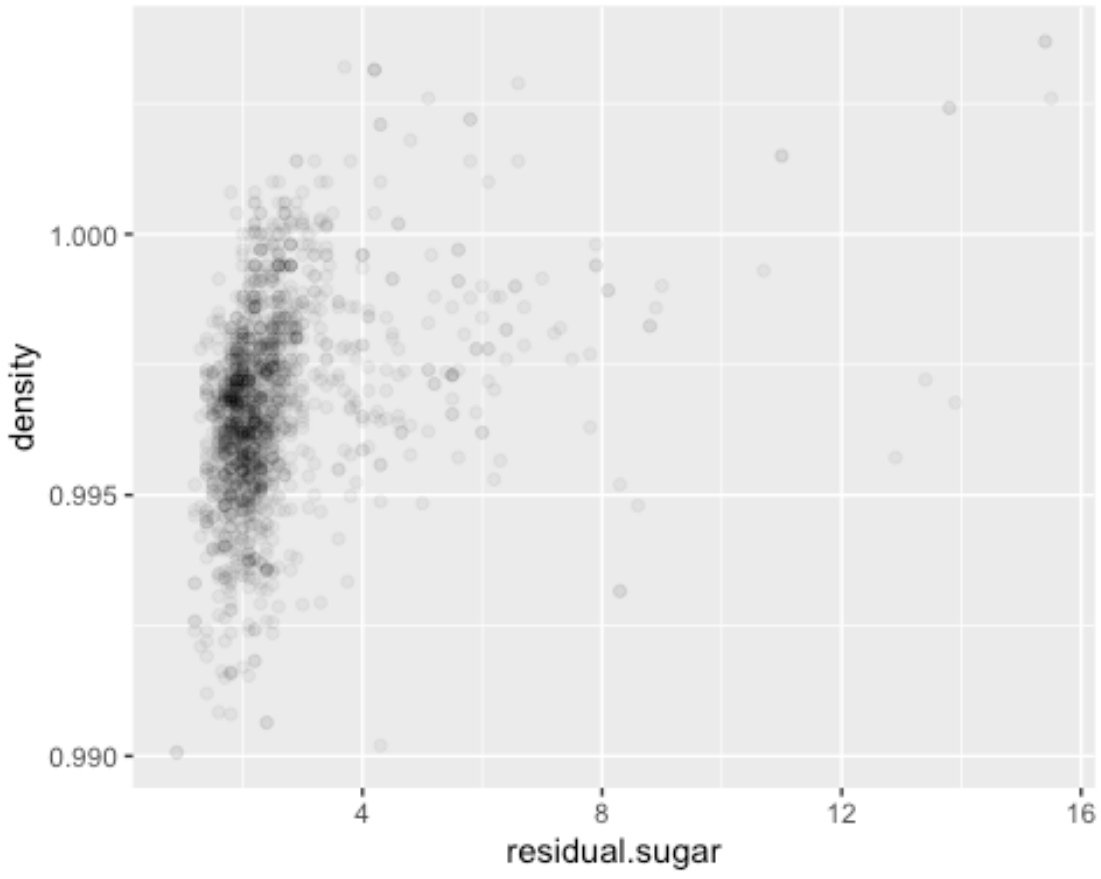

8
7
6
5
4
3
quality
10
12
14
alcohol

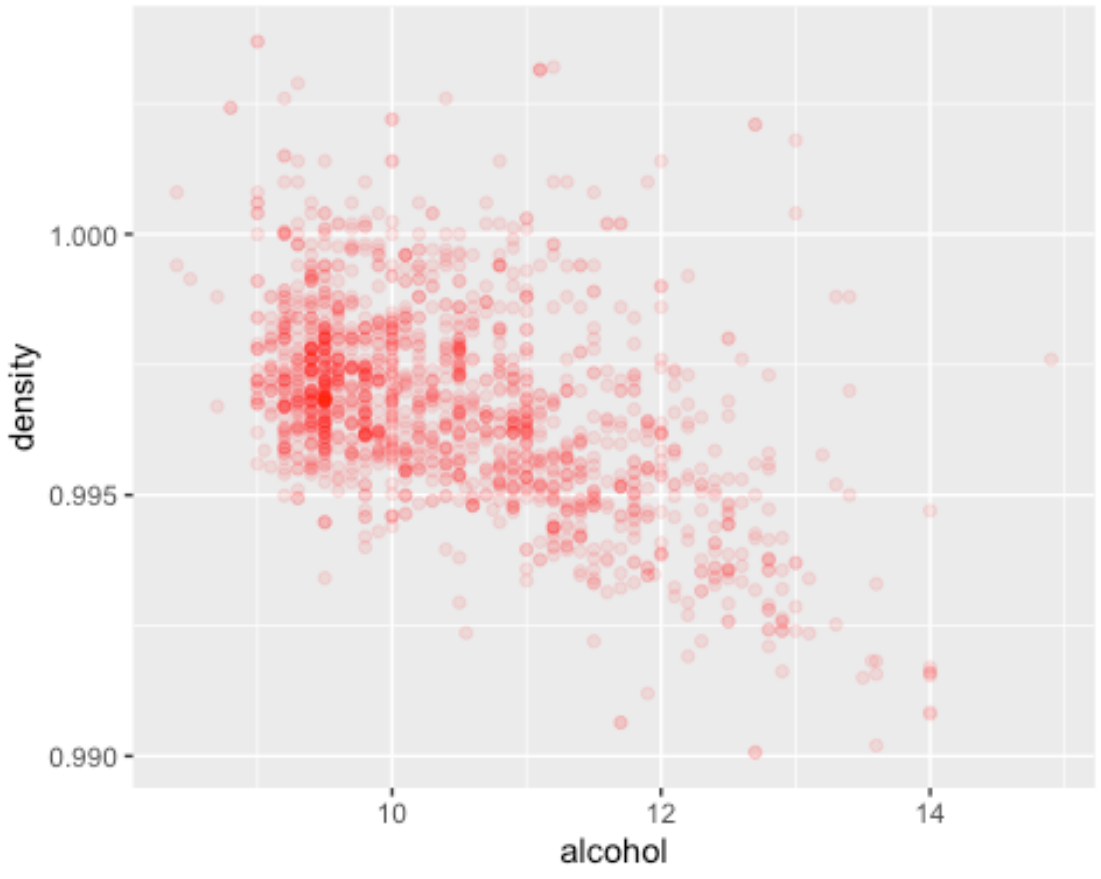

above average
average
below average
ranking
3.0
3.5
4.0
pH

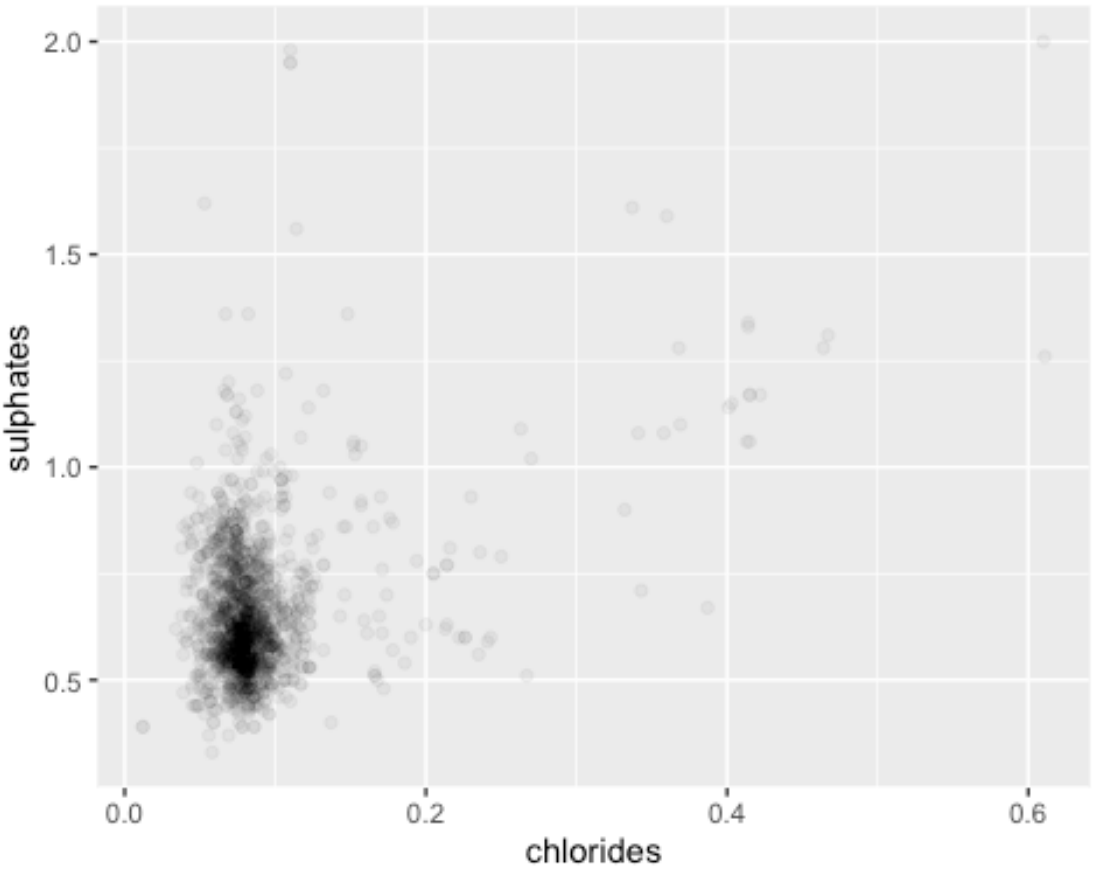


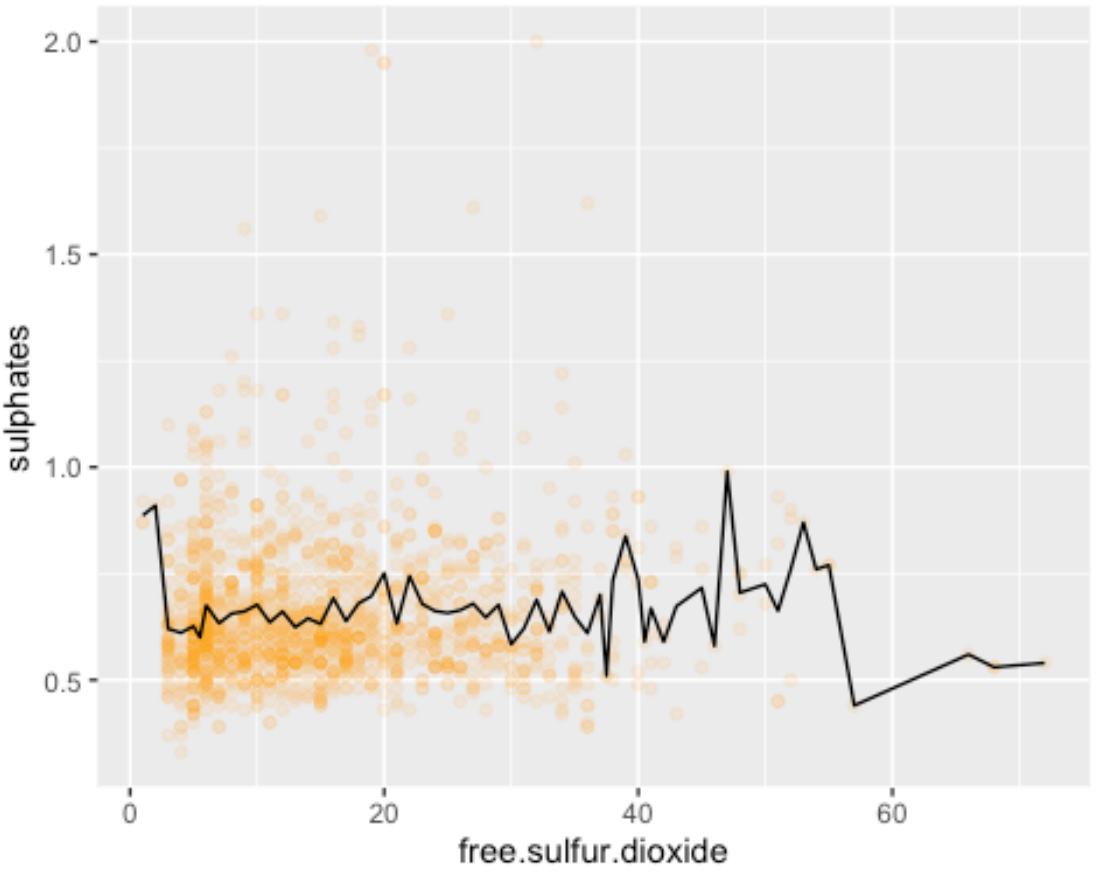
I'm going to add jitter properties to plots that appear too clustered and possibly some trend lines if they make sense for the data points.

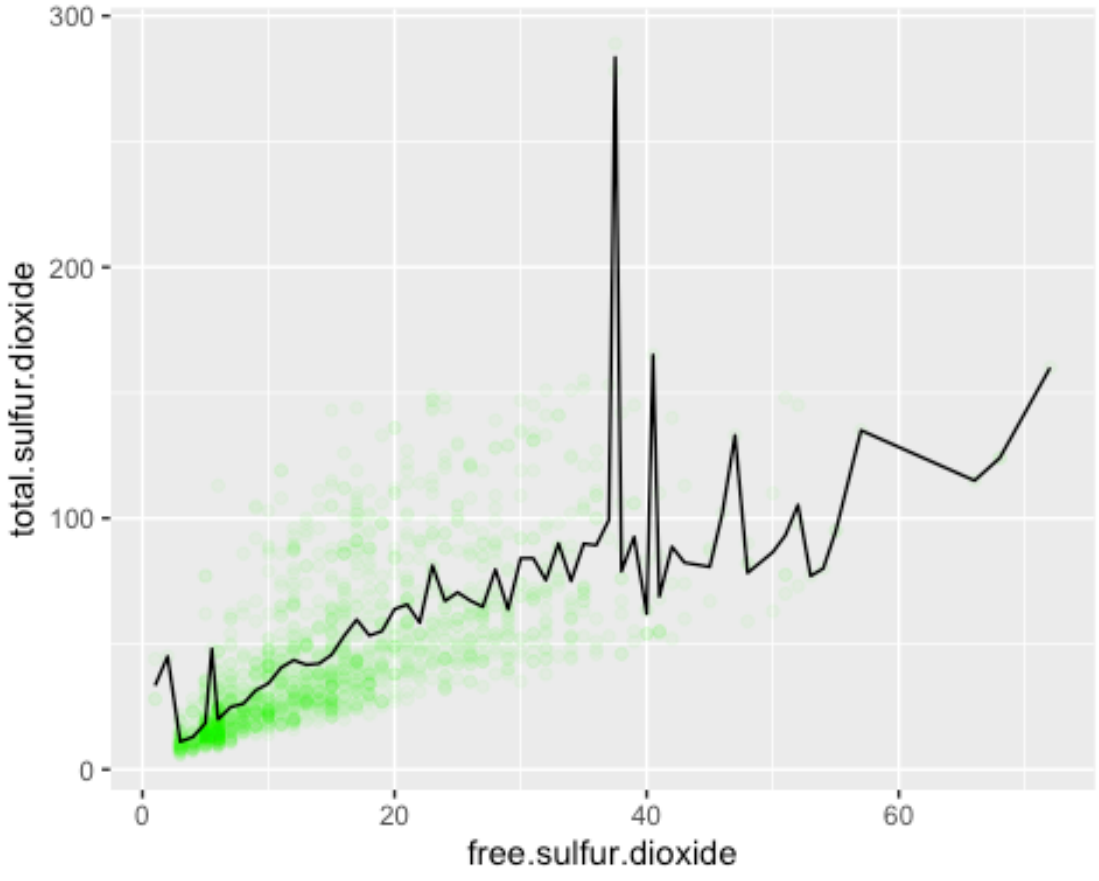

alcohol
14
12
10
4
8
12
16
residual.sugar


density
1.000
0.995
0.990
4
8
12
16
residual.sugar


density
1.000
0.995
0.990
10
12
14
alcohol


2.0
1.5
1.0
0.5
sulphates
0.0
0.2
0.4
0.6
chlorides


2.0
1.5
1.0
0.5
sulphates
0
20
40
60
free.sulfur.dioxide



#Bivariate Analysis

The first set of information that I want to check is the correlation values between the campared values. Using a pearson method caluculation as shown below I've calucated all the below values.

I'm going to have to use the raw quality scores that ranking is derived from.

```
##
##  Pearson's product-moment correlation
##
## data:  wd$alcohol and wd$quality
## t = 21.639, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4373540 0.5132081
```

```
## sample estimates:
##       cor
## 0.4761663

##
##  Pearson's product-moment correlation
##
## data:  wd$alcohol and wd$quality
## t = 21.639, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4373540 0.5132081
## sample estimates:
##       cor
## 0.4761663

##
##  Pearson's product-moment correlation
##
## data:  wd$pH and wd$quality
## t = -2.3109, df = 1597, p-value = 0.02096
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.106451268 -0.008734972
## sample estimates:
##         cor
## -0.05773139

##
##  Pearson's product-moment correlation
##
## data:  wd$citric.acid and wd$quality
## t = 9.2875, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1793415 0.2723711
## sample estimates:
##       cor
## 0.2263725

##
##  Pearson's product-moment correlation
```

```
##
## data:  wd$residual.sugar and wd$alcohol
## t = 1.6829, df = 1597, p-value = 0.09258
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.006960058  0.090909069
## sample estimates:
##       cor
## 0.04207544

##
##  Pearson's product-moment correlation
##
## data:  wd$residual.sugar and wd$density
## t = 15.189, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3116908 0.3973835
## sample estimates:
##       cor
## 0.3552834

##
##  Pearson's product-moment correlation
##
## data:  wd$alcohol and wd$density
## t = -22.838, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5322547 -0.4583061
## sample estimates:
##       cor
## -0.4961798

##
##  Pearson's product-moment correlation
##
## data:  wd$chlorides and wd$sulphates
## t = 15.978, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
```

```
##  0.3282127 0.4127694
## sample estimates:
##       cor
## 0.3712605

##
##  Pearson's product-moment correlation
##
## data:  wd$free.sulfur.dioxide and wd$sulphates
## t = 2.0671, df = 1597, p-value = 0.03888
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.002643125 0.100424406
## sample estimates:
##       cor
## 0.05165757

##
##  Pearson's product-moment correlation
##
## data:  wd$free.sulfur.dioxide and wd$total.sulfur.dioxide
## t = 35.84, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6395786 0.6939740
## sample estimates:
##       cor
## 0.6676665
```

Below is a table of the Correlation between data variables.

Variable 1 | Variable 2 | r | Strength | -/+ Alcohol | Quality | 0.48 | Moderate | + pH | Quality | -0.06 | Weak | - Citric Acid | Quality | 0.23 | Weak | + Sugar | Alcohol | 0.04 | Weak | + Sugar | Density | 0.36 | Moderate | + Alcohol | Density | -0.49 | Moderate | - Cholrides | Sulphates | 0.37 | Moderate |+ Free Sulfur | Sulphates | 0.05 | Weak | + Free Sulfur | Total Sulfur | 0.67 | Strong | +

I am surprised by a few of the correlations. The free sulfur dioxide and total sulfur dioxide have a strong positive relationship that isn't surprising as its measuring the same chemical compounds. Citric acid and pH don't have a relation to quality scores, so I can assume the judges probably had no biases to these qualities of the wine.

Sugar and density and alcohol and density both had moderate r values but the measures had opposite polarities. Sugar being the fuel for the yeast is responsible for both measures and their directions. Density is the relational measure offermentation potential and the goal range is 0 or below. Alcohol is less dense than water and sugar so a negative value denotes higher conversion. This is also shown in the moderate to strong negative correlation between alcohol and density with a r of -0.49, meaning more alcohol means a lower density but sugar leads to a higher value.

It's an interesting note in trying to objectively explain the judges quality score, that judges statistically are more inclined to rate a higher ABV with a higher score.

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?

My main point of interest are factors that may show trends in the taste judging of the samples. It seems most chemical markers have insignificant relations but there was a significant r in correlating alcohol with the scores.

Did you observe any interesting relationships between the other features

(not the main feature(s) of interest)? From my own interests, I found the mathematical relationship that shows the sugar, alcohol, and density interesting to see charted out.

What was the strongest relationship you found?

The correlation of between sulfur measurements.

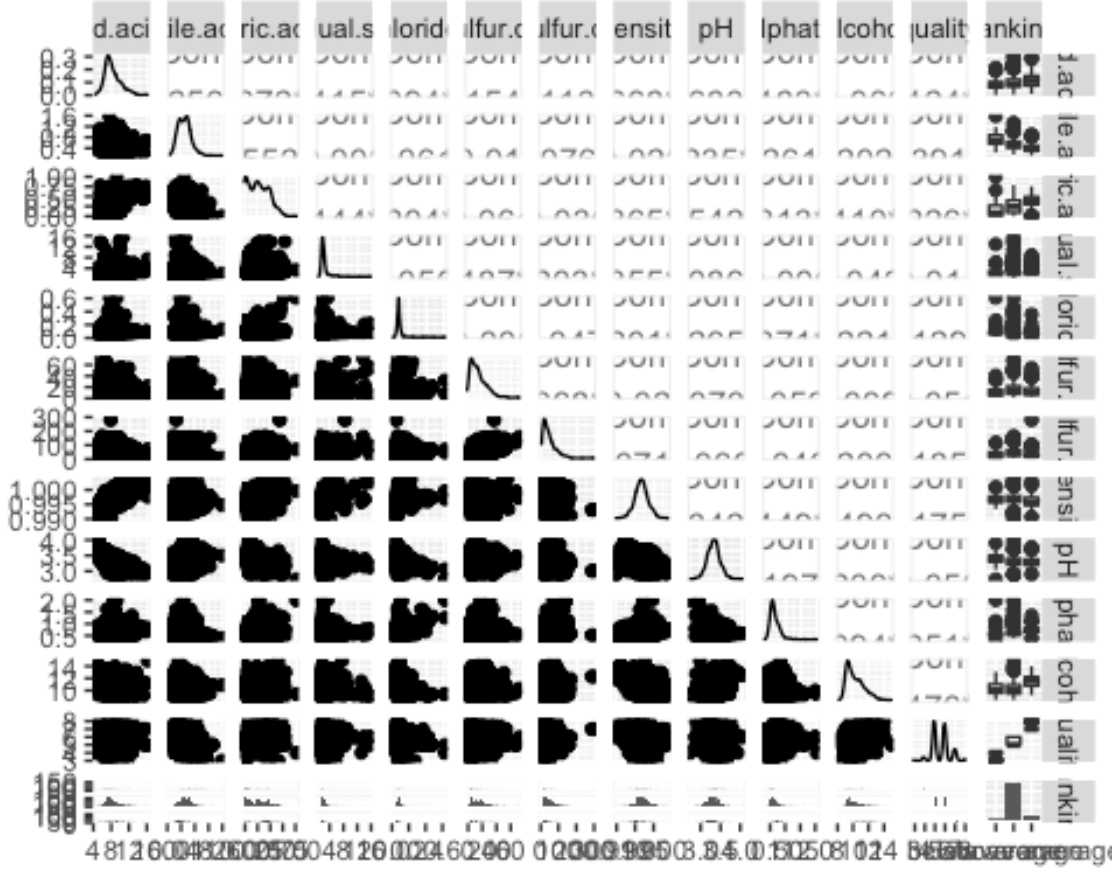
Multivariate Plots Section

To start off the multivariate investigation, it seems sound to some matrices plots and look for any additional investigations that haven't been apparent to me so far. I've created a subset to exclude the sequential sample numbers of the wines X.

```
nox <- subset(wd, select=c(
fixed.acidity, volatile.acidity, citric.acid, residual.sugar,
chlorides, free.sulfur.dioxide, total.sulfur.dioxide, density, pH,
sulphates, alcohol, quality, ranking))
```

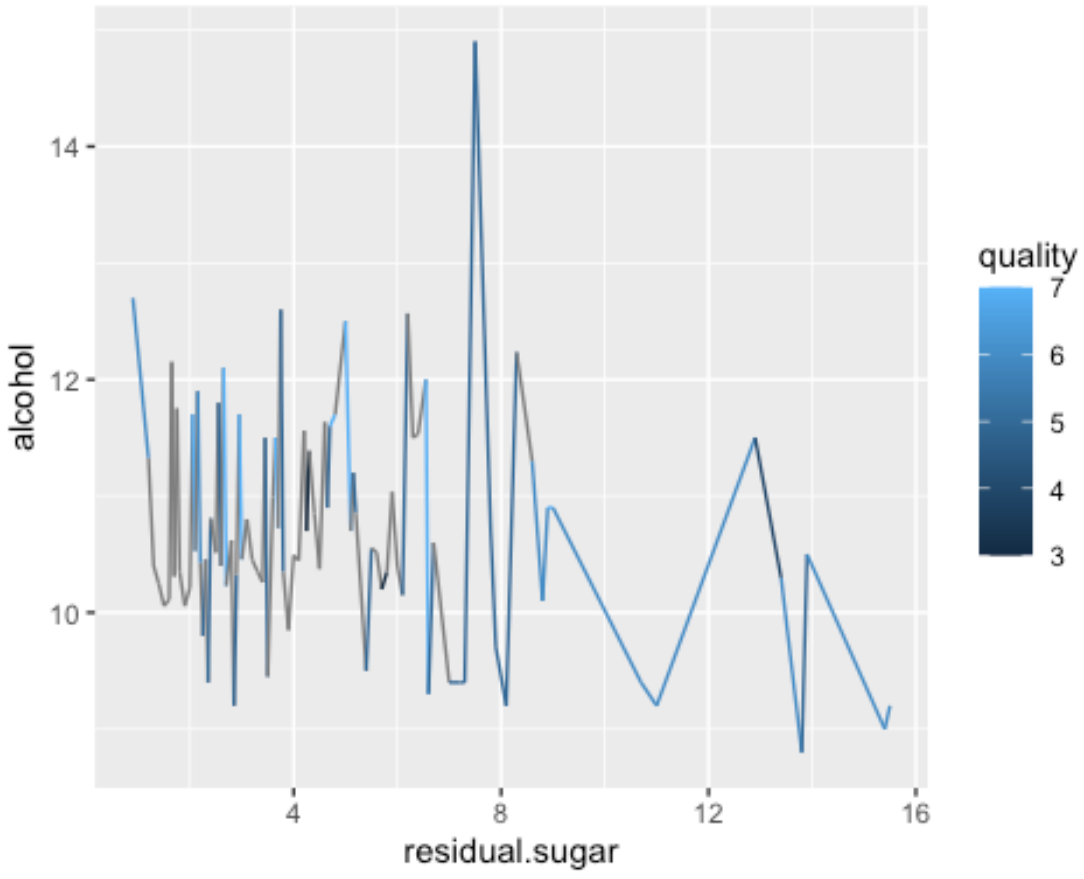
Using this subset the matrix is as follows.

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

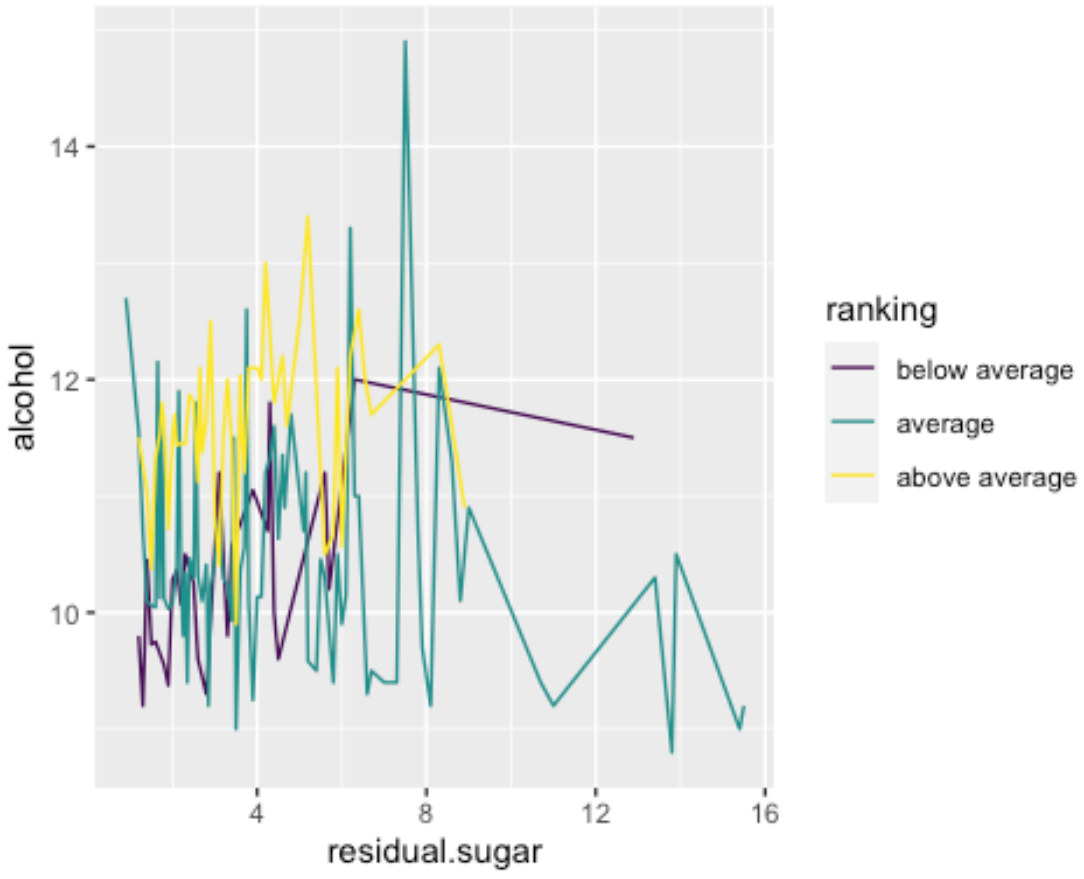


I don't see anything thats super unexpected. The residual sugar and alcohol both skew the same as expected. The citric acid to sugar distribution is a little wider than I had thought but that sharp decline on the tail is interesting, so I'm going to add that to investigative plots.

From the previous data explorations, the first relationship I want to look at is the one between residular sugar, alcohol, and quality/ranking.

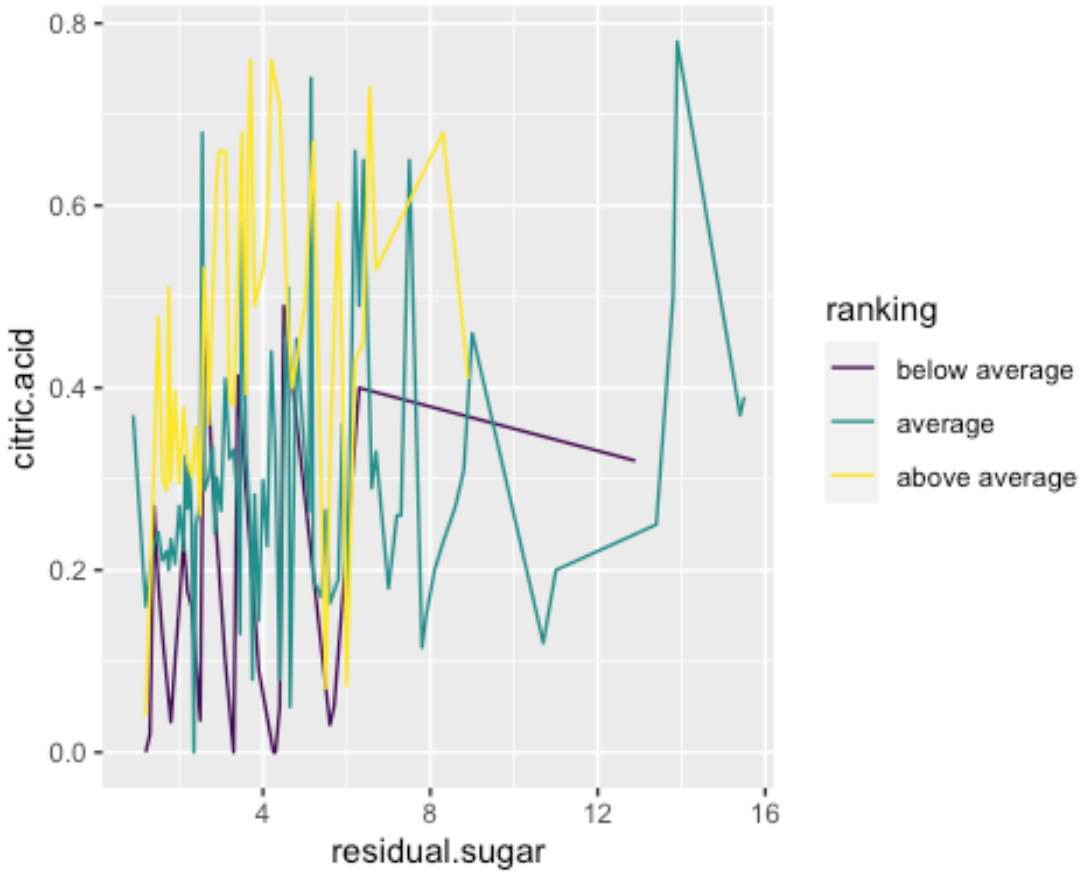


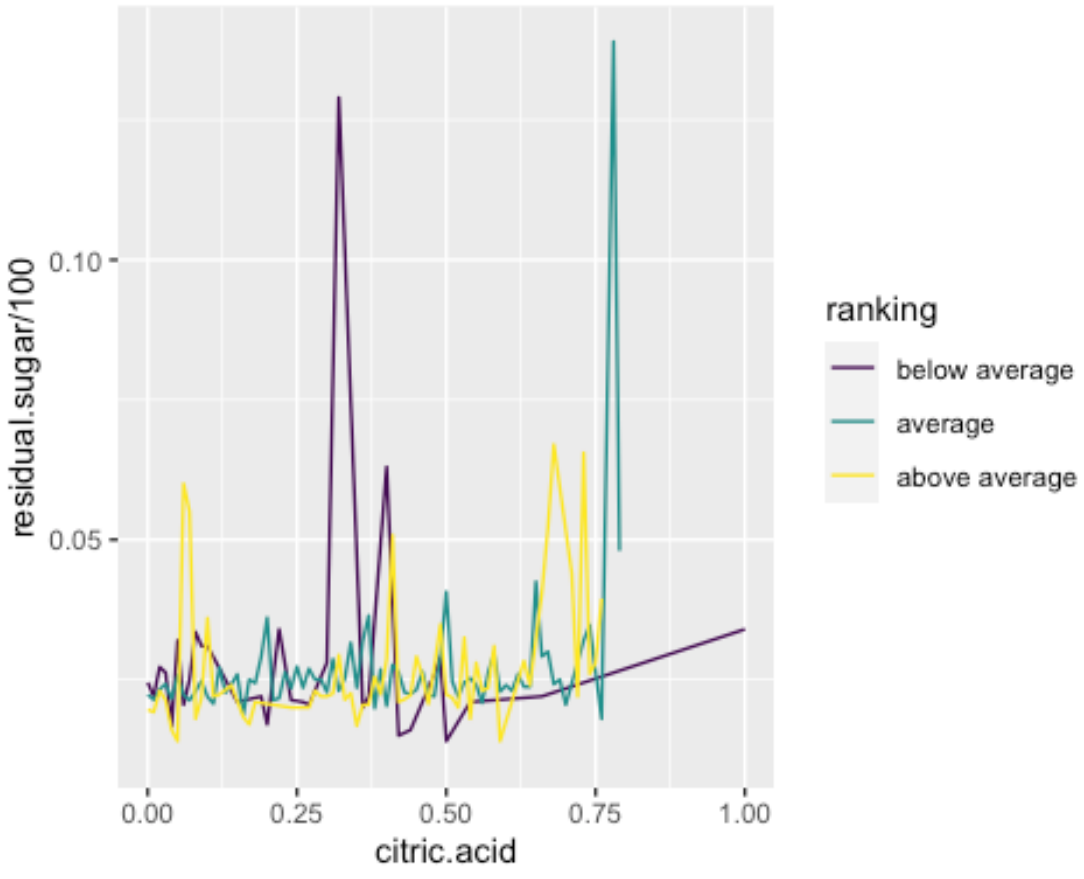
That isn't a view that is easily readable by most, so if I use the hard set ranking categories it may translate better to the veiwer.

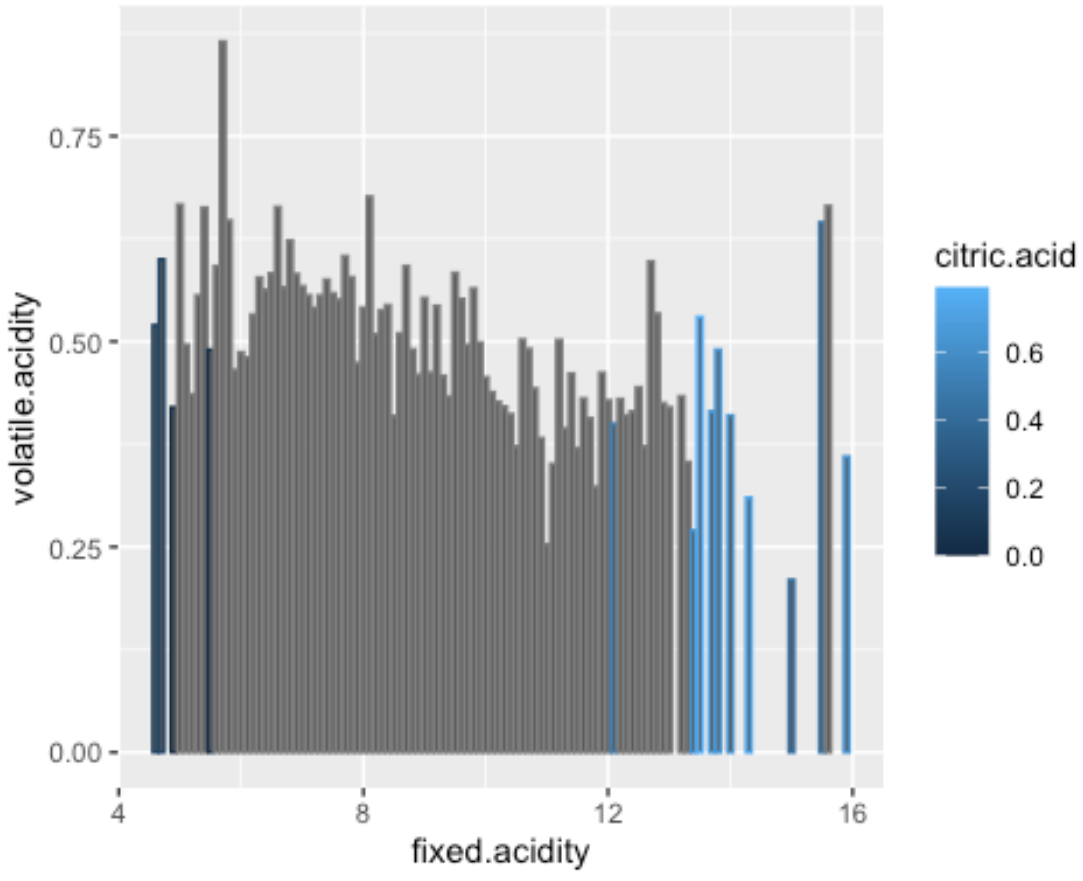


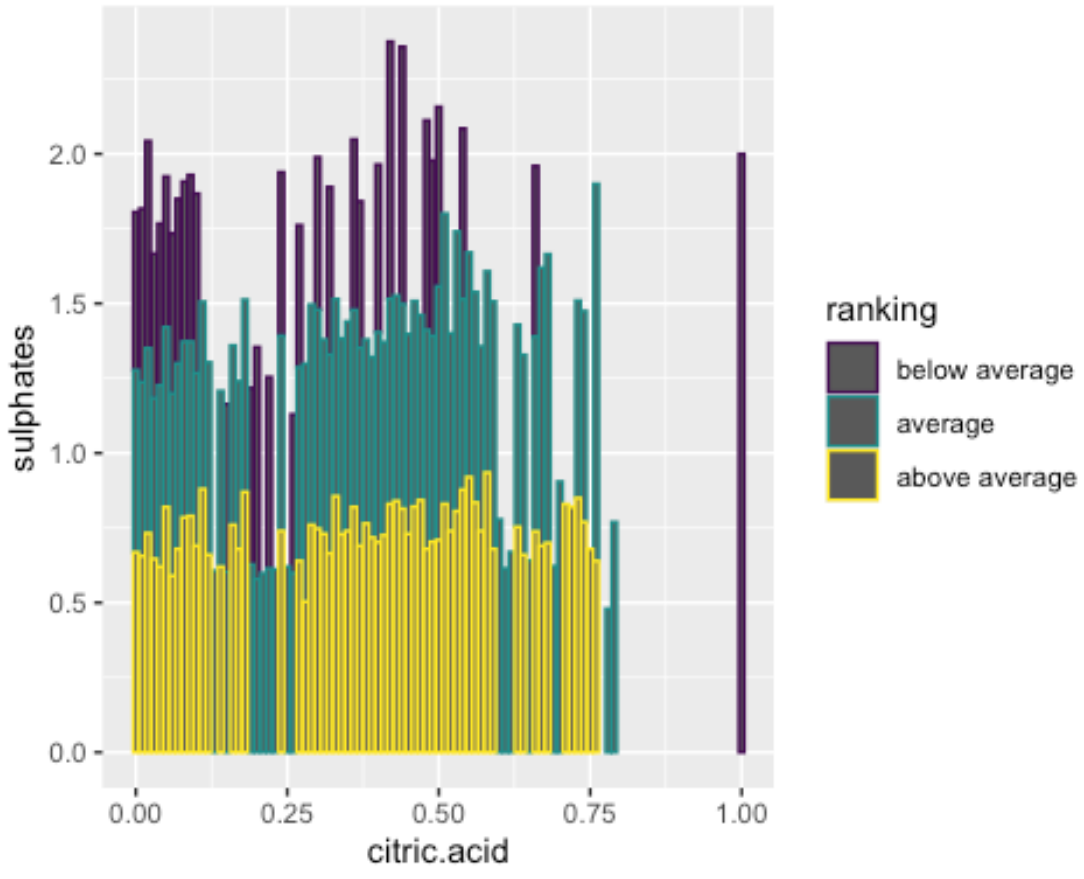
Ok that makes the rankings stand out a bit more and also shows that above the entries judged with a higher score overall plot higher in alcohol.

I want to explore some of the more posible complex chemical relations. Im just trying out a some different combinations looking for things that stand out. The citric acid and sugar relation caught my eye in the mass plotting so I want to investigate that.

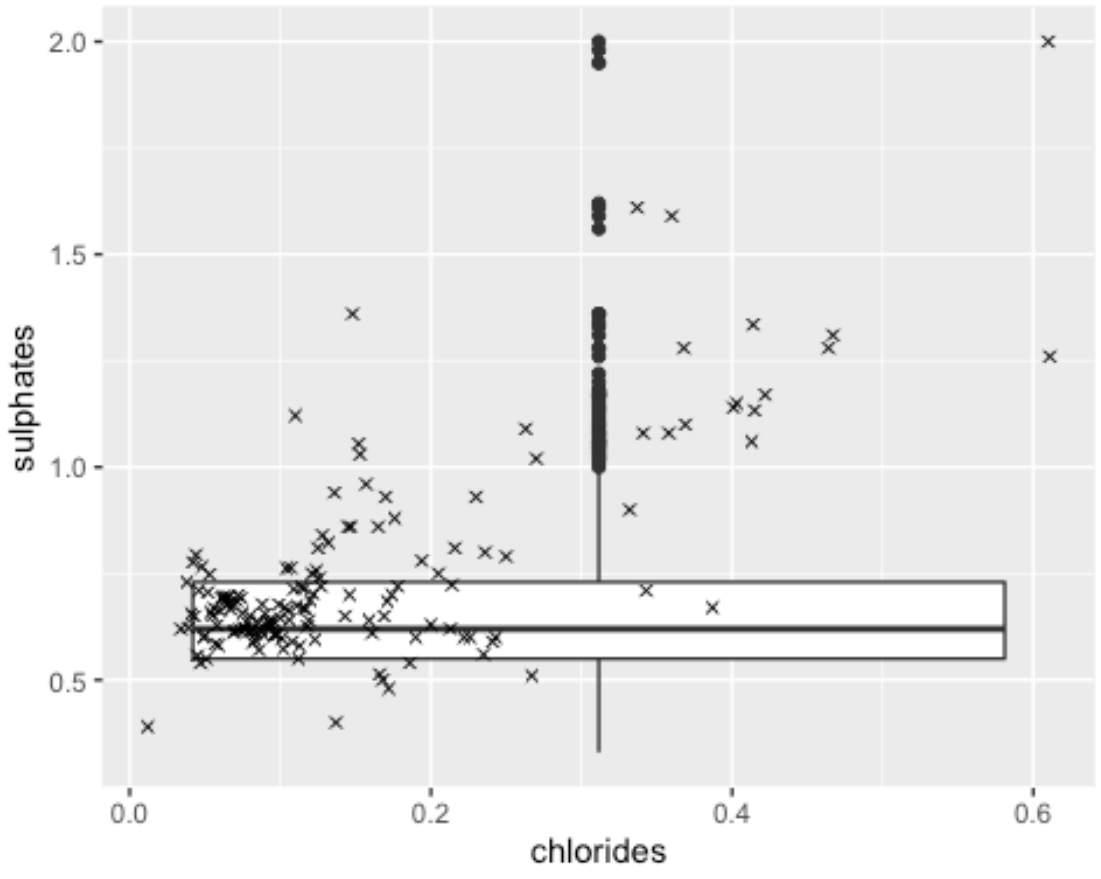

0.8
0.6
0.4
0.2
0.0
citric.acid
4
8
12
16
residual.sugar
ranking
below average
average
above average


residual.sugar/100
0.10
0.05
0.00
0.25
0.50
0.75
1.00
citric.acid
ranking
below average
average
above average

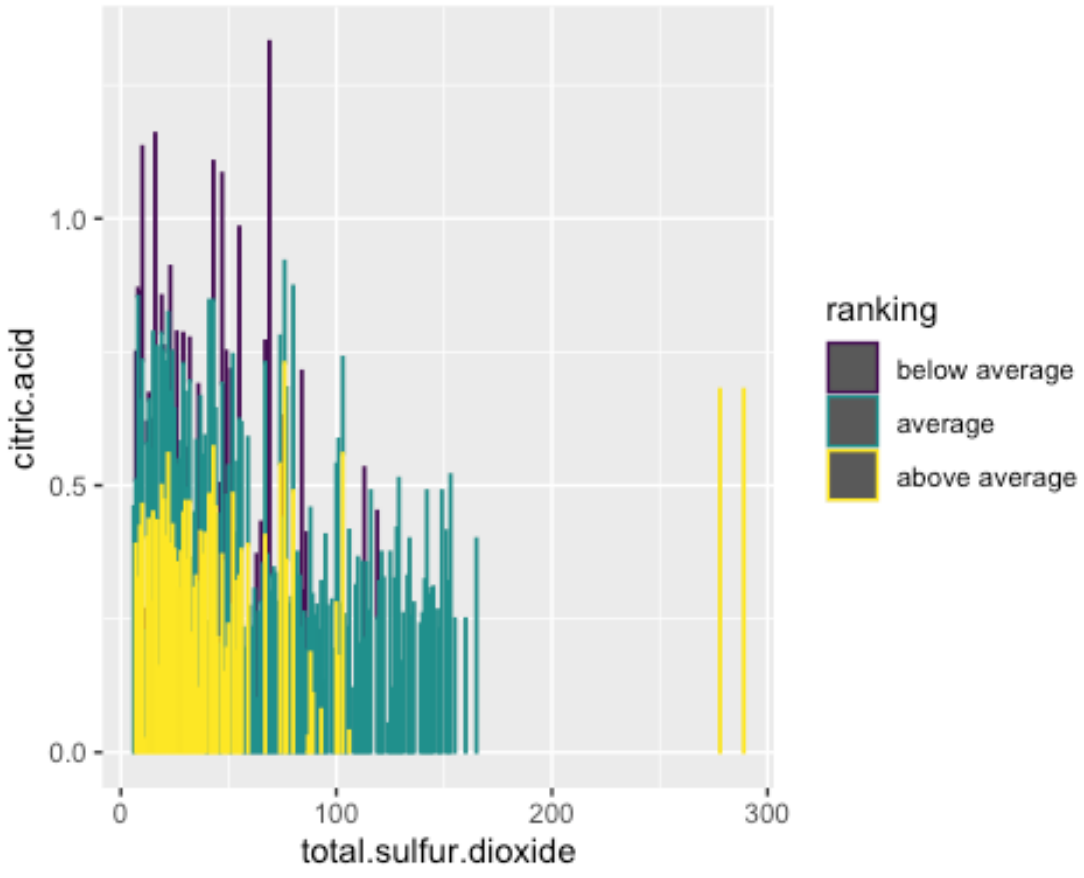

0.75
0.50
0.25
0.00
volatile.acidity
4
8
12
16
fixed.acidity
citric.acid
0.6
0.4
0.2
0.0



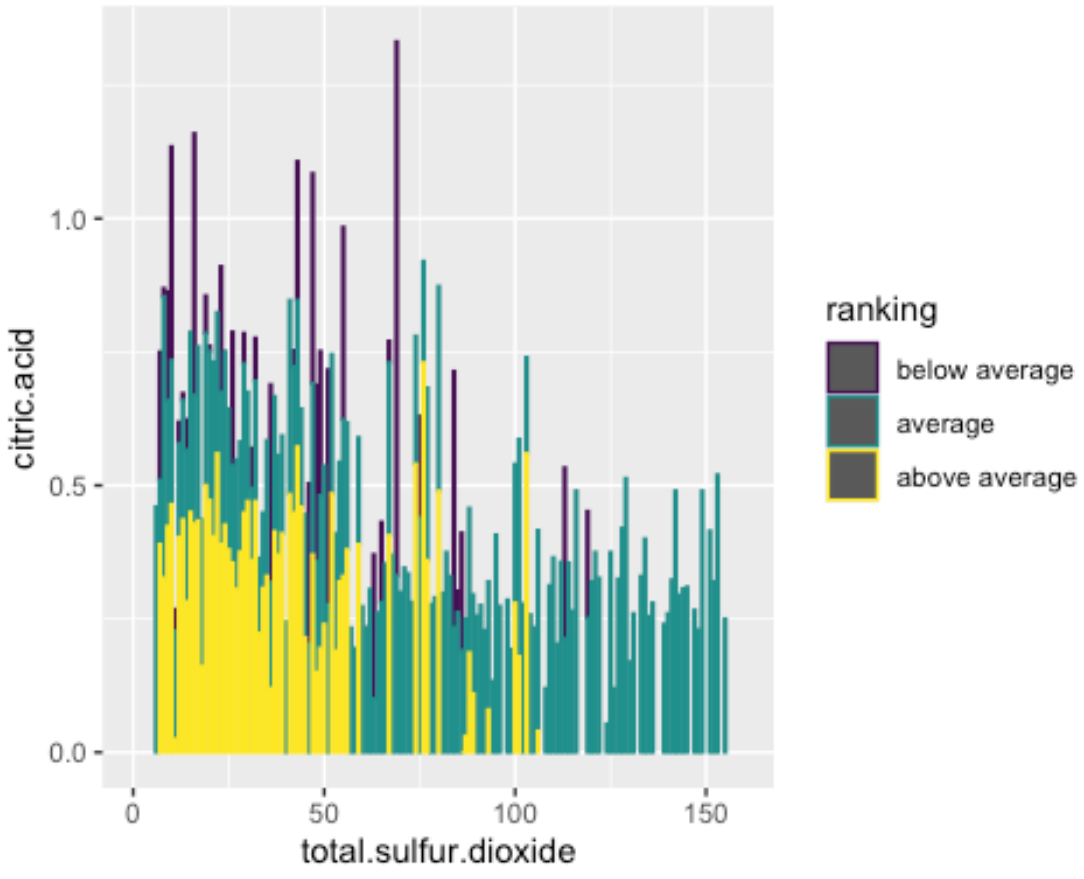
Interesting view of two no related chemical markers in respect to quality score.

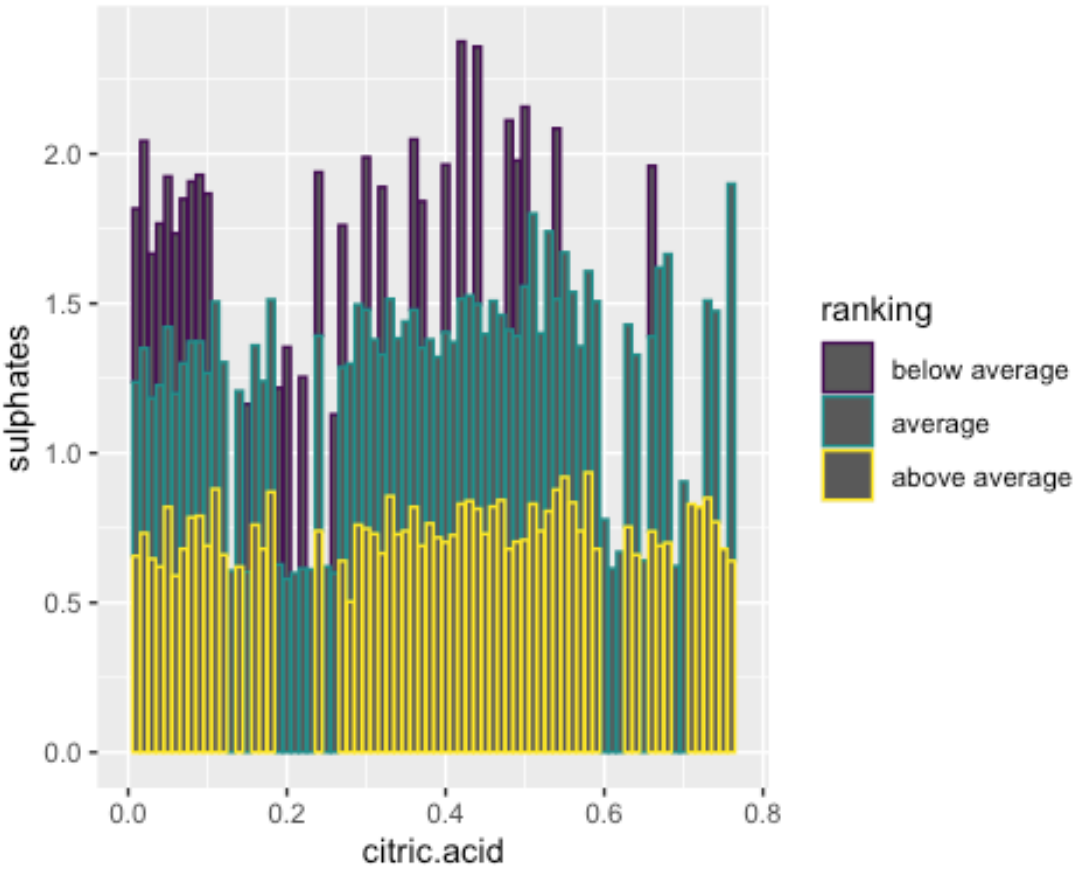


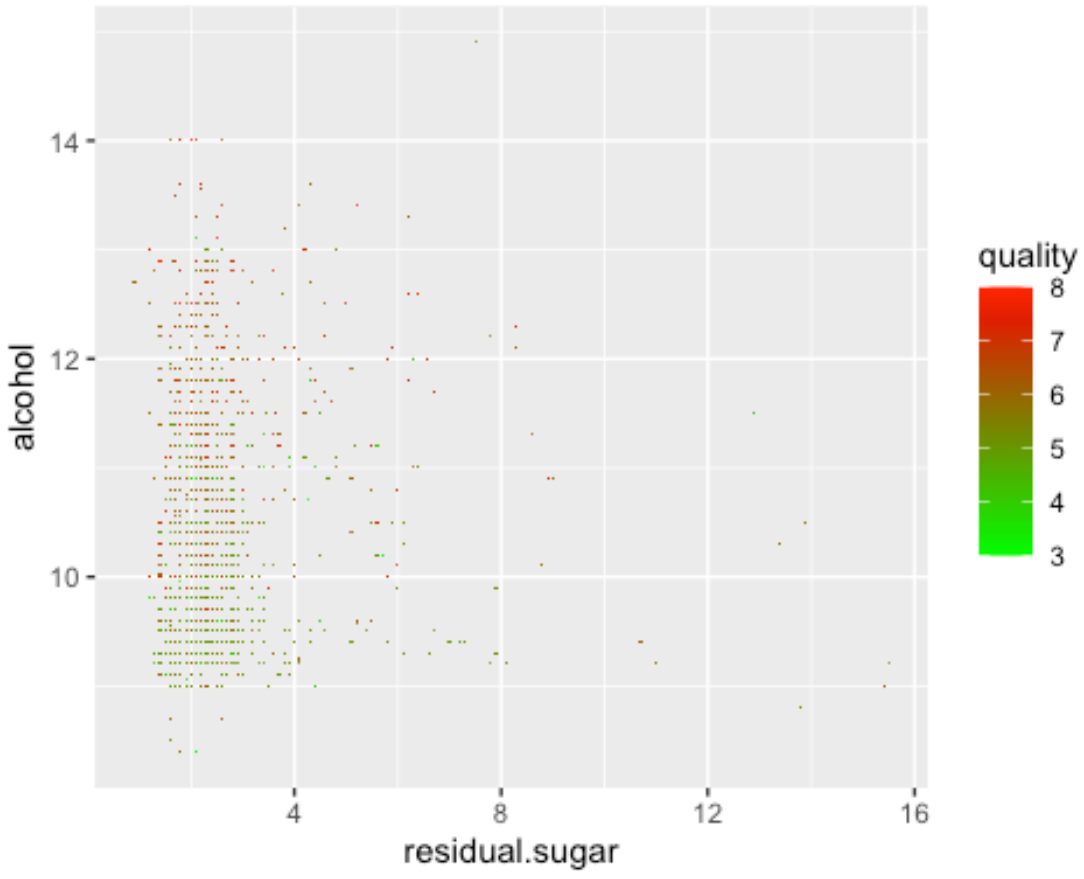
Well that doesn't really show much. I'm going to keep exploring different measures to quality.



That's more like the views Im hoping to find. Going to note that the highest sulfur samples happened to rank in the top group. I restricted the axis for better presentation so those values were lost.


citric.acid
1.0
0.5
0.0
0
50
100
150
total.sulfur.dioxide
ranking
below average
average
above average


sulphates
2.0
1.5
1.0
0.5
0.0
0.0
0.2
0.4
0.6
0.8
citric.acid
ranking
below average
average
above average



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?

The scieniftic explanation of alcohol, sugar, and density plot out nicely as expected. It appears that there are trends in higher rated wines, may not a chemical composistion but just the quanity that lead to a better taste experience. Knowing the quality ranking criteria and categorical scores but be a great data point to have in furthering this investigation.

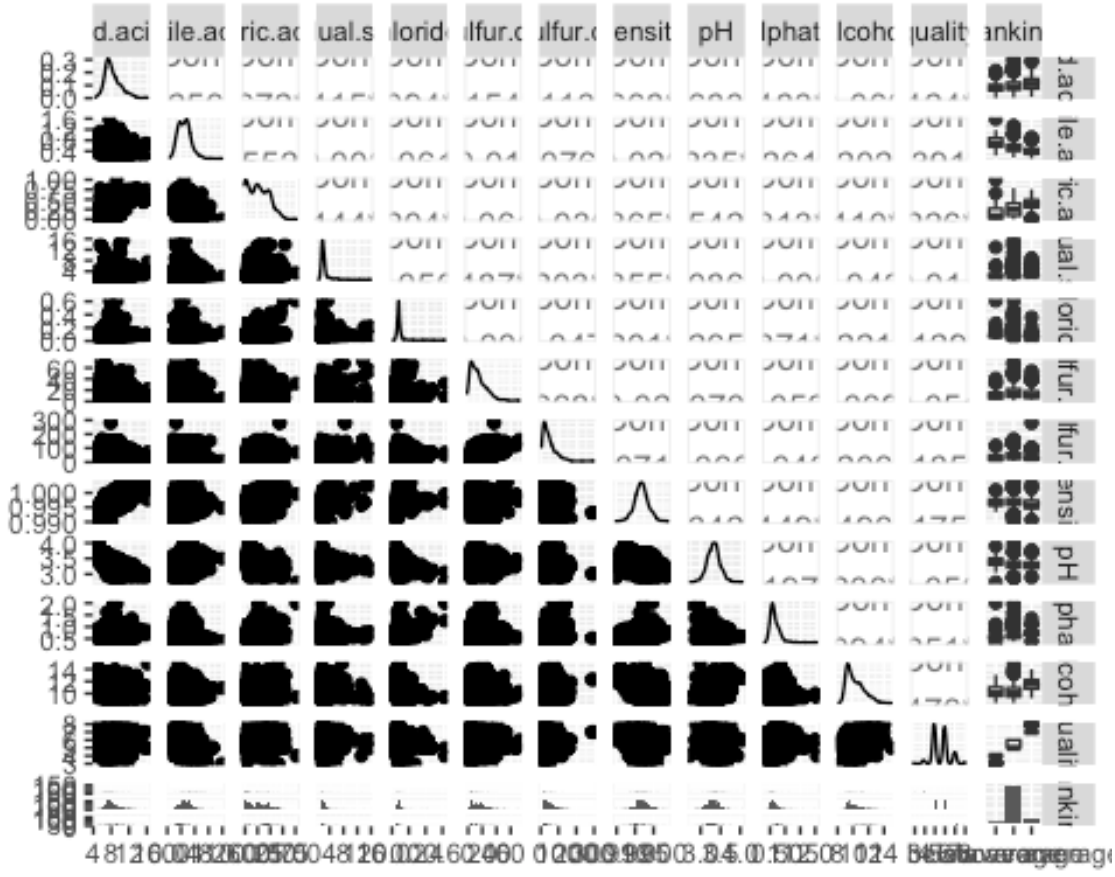
Were there any interesting or surprising interactions between features?

The overall measures of most chemical existing in higher rates in the highest rated samples has my curiousity peaked in this data. Maybe scores can't be factored out into key points, simply the higher content but still in relation to the other compounds make flavor even more complicated to account for.

Final Plots and Summary

Plot One

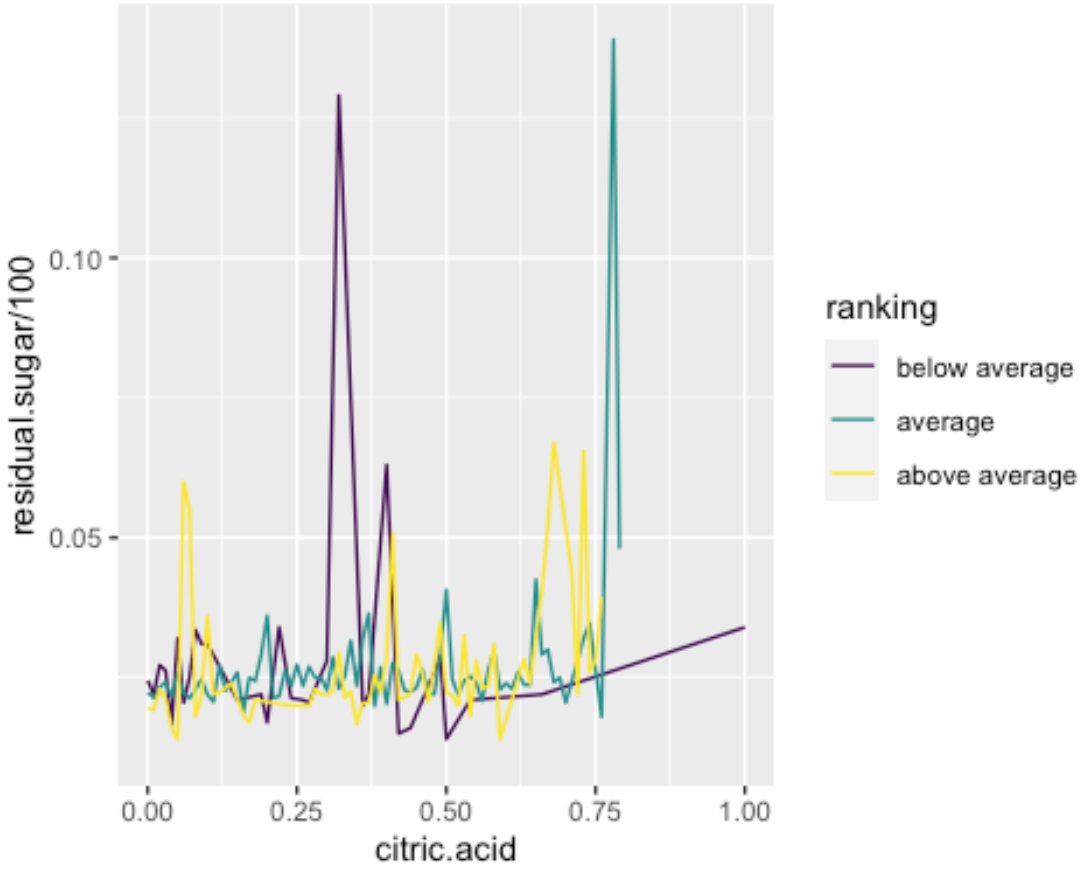
```
##  [1] "fixed.acidity"        "volatile.acidity"     "citric.acid"
##  [4] "residual.sugar"       "chlorides"
"free.sulfur.dioxide"
##  [7] "total.sulfur.dioxide" "density"              "pH"
## [10] "sulphates"            "alcohol"              "quality"
## [13] "ranking"
```



Description One

This is the overall scatterplot matrix for the data set. This can be argued that it may be very audience specific to the inclusion in presenting data with this added to it. For those that understand what is shown it adds an overall insight to the relations explored.

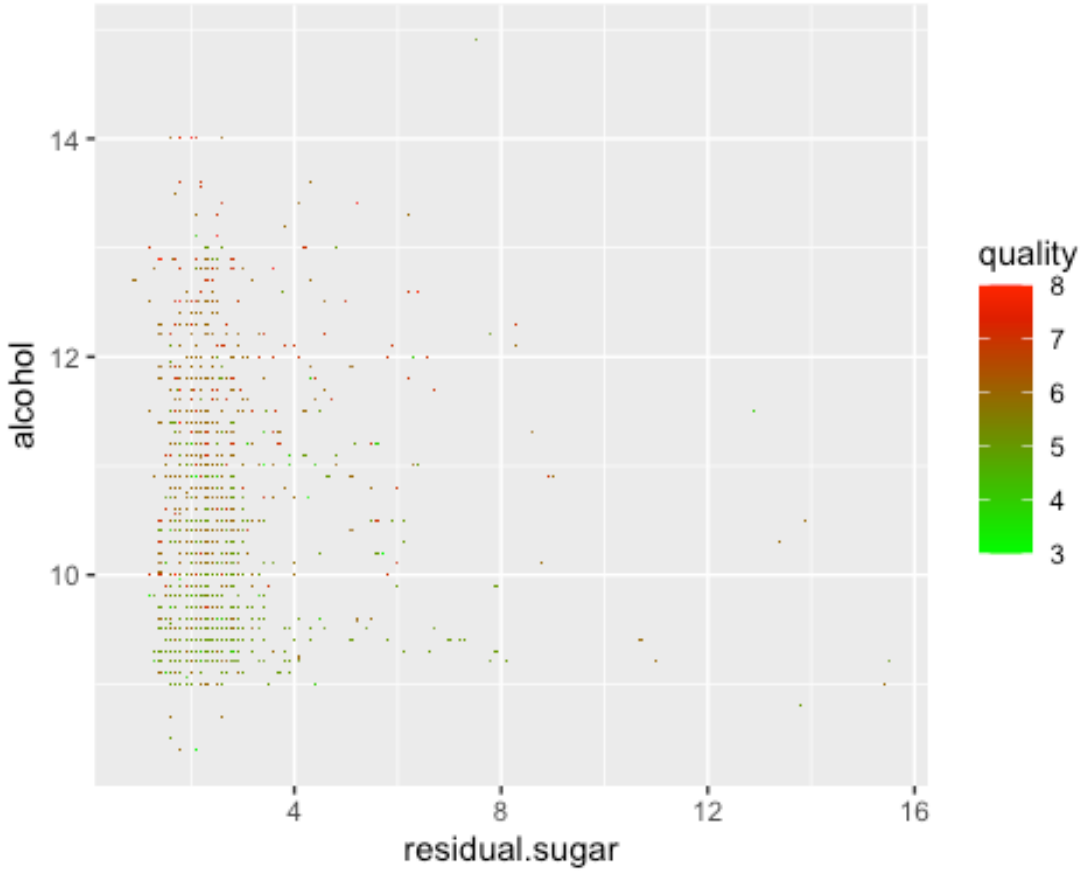
Plot Two



Description Two

This plot helps point to a recurring trend that higher quality wines average a lower resisudal sugar content. Some assumptions can be made to that line of thinking that could open a deeper analysis with a more robust data set. If I was the vinter of some juddged wines and had additional data such as age, grape variety, and other variables of the winery, this could help me improve the quaility or drive to better consistent products that are likely to be favored in judgings.

Plot Three



###

Description Three

This a related plot to some of the other discussions showing 'dryer', lower sweetness, and higher ABV measures tend to be favored more.

Reflection

I found this data set interesting but maybe a bit incomplete for the questions it aroused in myself. My largest interest is how this can be used to find the factors that impact judging of the wines. There were some trends and relations that were found that helped shed some light on the question at hand. The most common factor seem the concentration of compound including alcohol tend to score higher. This could be due to several factors that prompt further discovery. I feel that having the values that were judged to get the quality

scores, as well vinter specific inforamtion such as region, yeast strain, grape variety, and age could really round out the picture of what goes into making great wine.