White Wine Quality Analysis by Jiho Tahk

Introduction

The following exploratory data analysis will explore 4,898 white wines of the Portuguese "Vinho Verde" wine variety. There are 11 attributes in the data set, and a quality score that grades the quality of the wine on a scale of 0 to 10. I will explore the data, trying to determine what attributes contribute or detract from obtaining a high quality score.

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

Available at: [@Elsevier] http://dx.doi.org/10.1016/j.dss.2009.05.016 [Pre-press (pdf)]
http://www3.dsi.uminho.pt/pcortez/winequality09.pdf [bib]
http://www3.dsi.uminho.pt/pcortez/dss09.bib

Univariate Plots Section

```
## [1] 4898   13

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

## 'data.frame':    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 ...
```

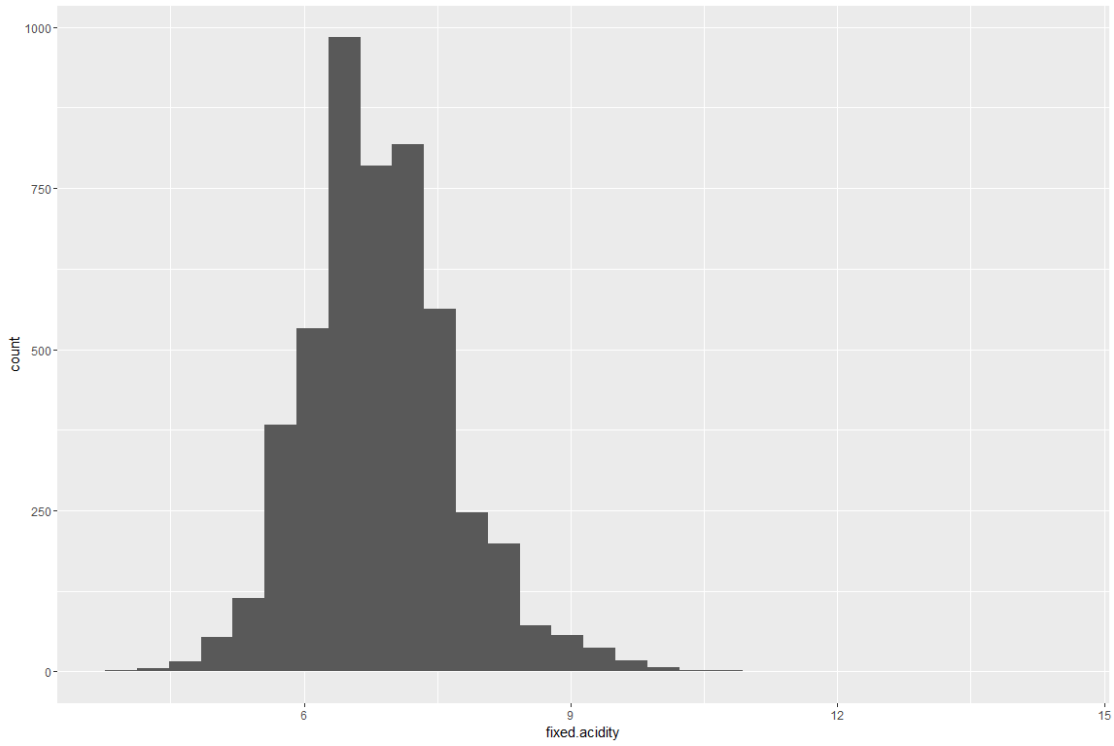
```
##  $ quality             : Ord.factor w/ 7 levels "3"<"4"<"5"<"6"<..: 4 4 4
4 4 4 4 4 4 4 ...

##        X        fixed.acidity    volatile.acidity  citric.acid
##  Min.   :   1   Min.   : 3.800   Min.   :0.0800   Min.   :0.0000
##  1st Qu.:1225   1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700
##  Median :2450   Median : 6.800   Median :0.2600   Median :0.3200
##  Mean   :2450   Mean   : 6.855   Mean   :0.2782   Mean   :0.3342
##  3rd Qu.:3674   3rd Qu.: 7.300   3rd Qu.:0.3200   3rd Qu.:0.3900
##  Max.   :4898   Max.   :14.200   Max.   :1.1000   Max.   :1.6600
##
##  residual.sugar     chlorides       free.sulfur.dioxide
##  Min.   : 0.600   Min.   :0.00900   Min.   :  2.00
##  1st Qu.: 1.700   1st Qu.:0.03600   1st Qu.: 23.00
##  Median : 5.200   Median :0.04300   Median : 34.00
##  Mean   : 6.391   Mean   :0.04577   Mean   : 35.31
##  3rd Qu.: 9.900   3rd Qu.:0.05000   3rd Qu.: 46.00
##  Max.   :65.800   Max.   :0.34600   Max.   :289.00
##
##  total.sulfur.dioxide    density             pH          sulphates
##  Min.   :  9.0        Min.   :0.9871   Min.   :2.720   Min.   :0.2200
##  1st Qu.:108.0        1st Qu.:0.9917   1st Qu.:3.090   1st Qu.:0.4100
##  Median :134.0        Median :0.9937   Median :3.180   Median :0.4700
##  Mean   :138.4        Mean   :0.9940   Mean   :3.188   Mean   :0.4898
##  3rd Qu.:167.0        3rd Qu.:0.9961   3rd Qu.:3.280   3rd Qu.:0.5500
##  Max.   :440.0        Max.   :1.0390   Max.   :3.820   Max.   :1.0800
##
##     alcohol      quality
##  Min.   : 8.00   3:  20
##  1st Qu.: 9.50   4: 163
##  Median :10.40   5:1457
##  Mean   :10.51   6:2198
##  3rd Qu.:11.40   7: 880
##  Max.   :14.20   8: 175
##                  9:   5
```

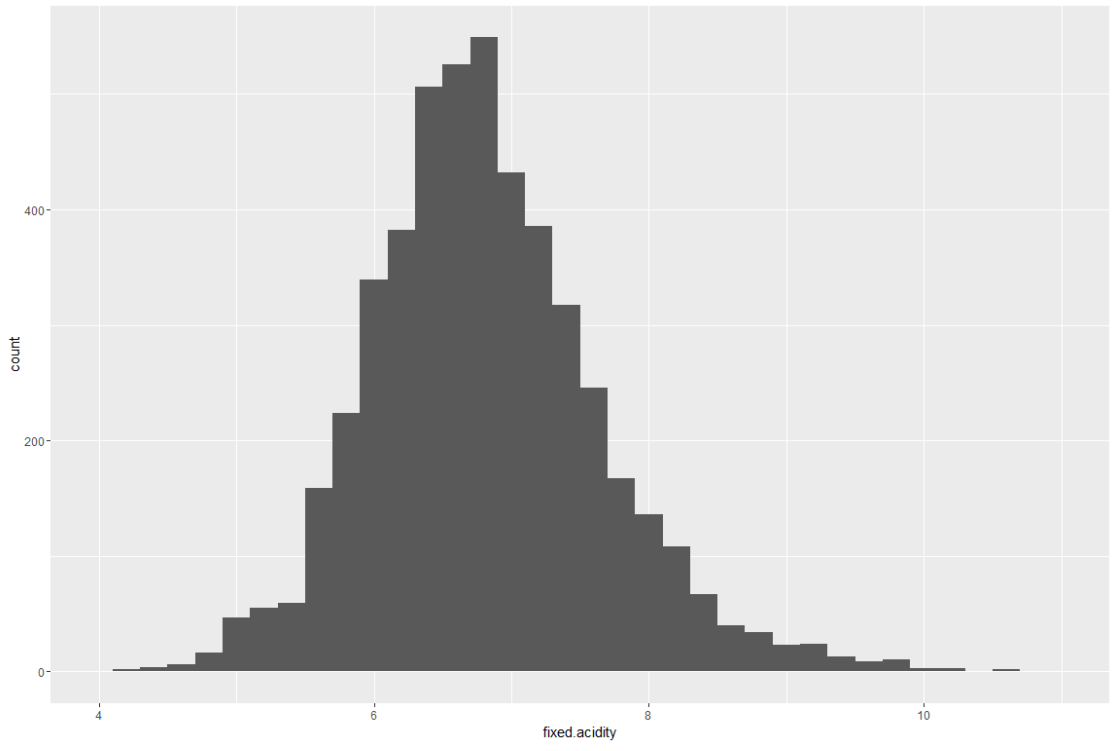
Initial thoughts on data

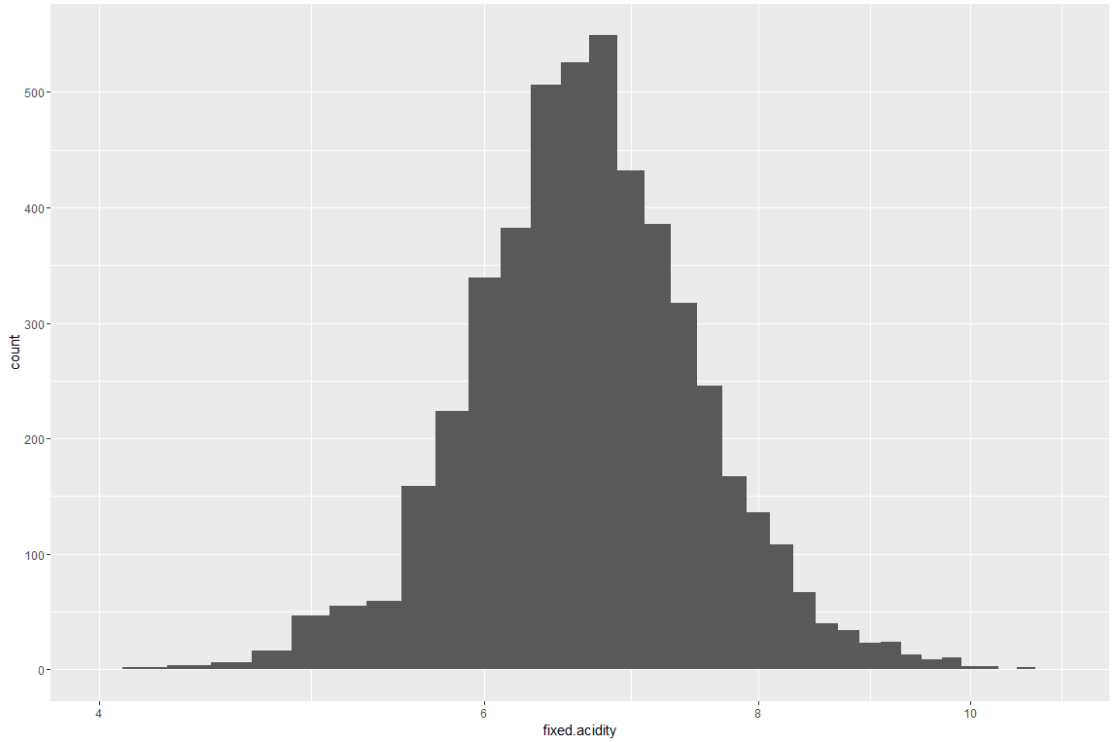
There are 12 variables of interest. Main variable of interest is 'quality,' which is an integer. I notice a lack of categorical variables, although quality could be treated as one. Noticing some outliers in some variables: max residual sugar, max chlorides, max free.sulfur.dioxide, and min & max of total sulfur dioxide. The mean/median quality score is a 6, with a mix of 3 and max of 9.

Acidity

First, I want to look at acidity. There are a few variables related to acidity: fixed acidity, volatile acidity, citric acid, and pH. Acidity has an effect on taste, so it may be correlated to the score. Let's see how the different types of acidity are distributed amongst white wines.

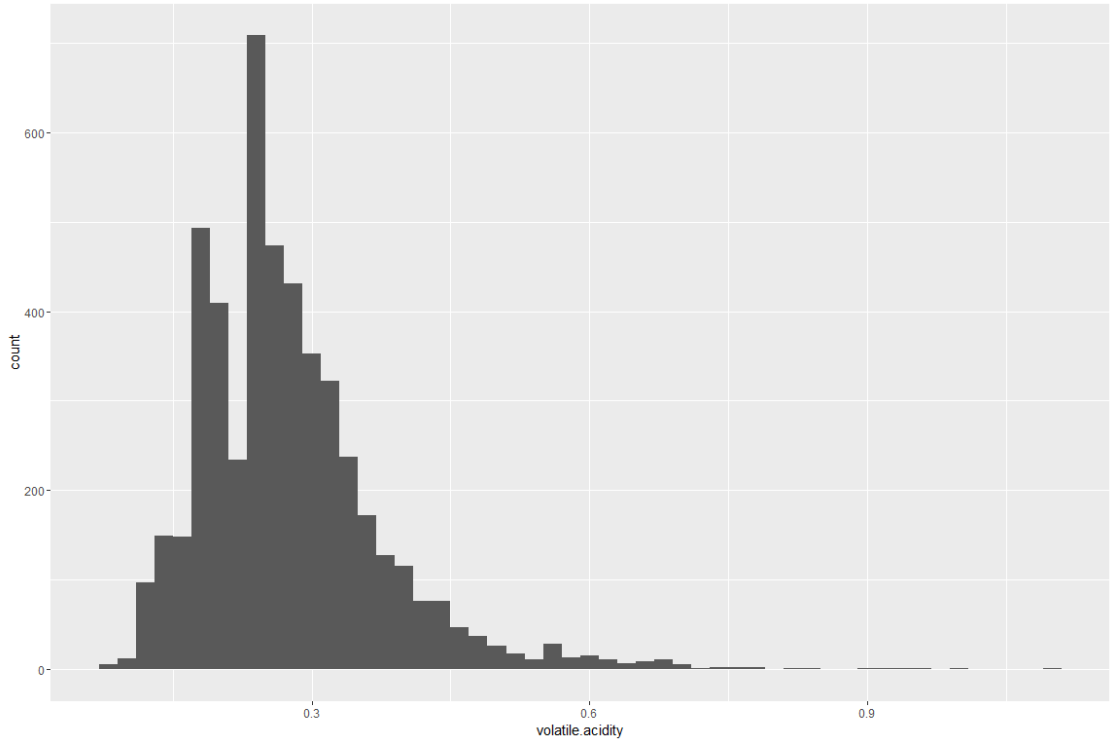

1000
750
500
250
0
count
6
9
12
15
fixed.acidity

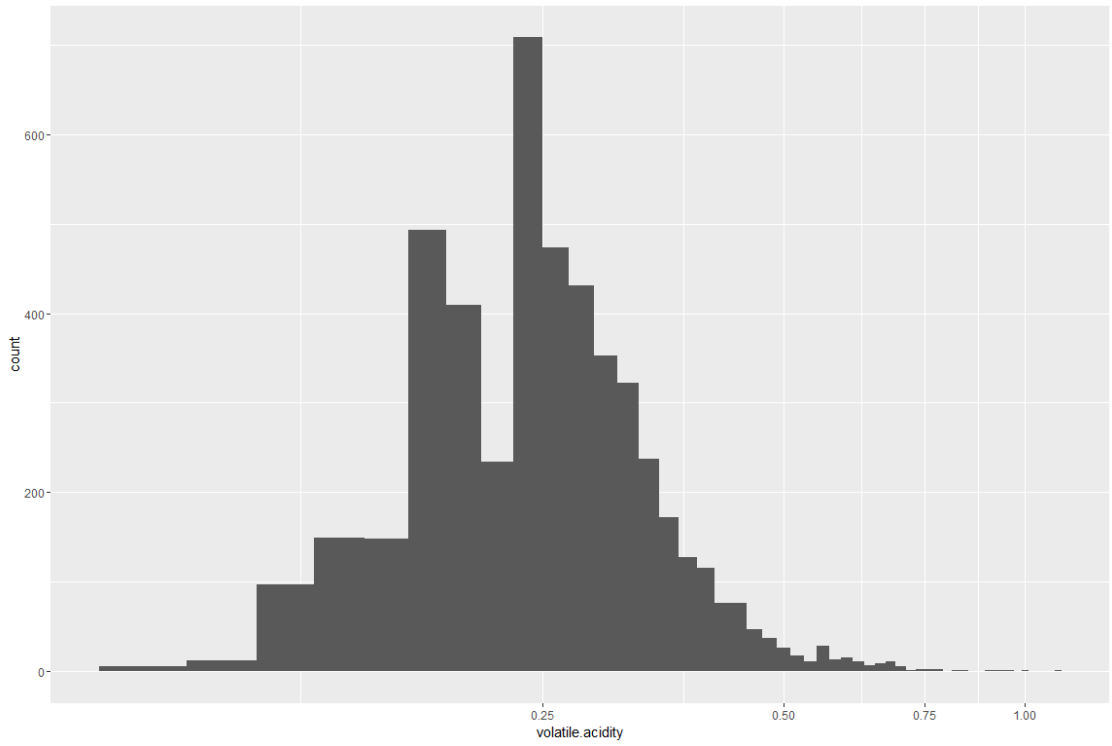

400
200
0
count
4
6
8
10
fixed.acidity




```
## [1] 4 3 5 6 6 6 3
## Levels: 3 < 4 < 5 < 6 < 7 < 8 < 9

##  [1] 6 5 5 5 7 8 3 7 5 7 7 4 7 5 5 6 6 6 7 5 5 5 6 8 8 6 6 7 6 6
## Levels: 3 < 4 < 5 < 6 < 7 < 8 < 9
```

Right skewed distribution, with mode around 7. Nearly normal with x scale transformed log10. A few outliers with acidity > 10 - only 7 data points, and none of them scored particularly well. In fact, 2 of them got the lowest quality score of 3. I am not entirely sure if this variable has an effect on taste or flavor.

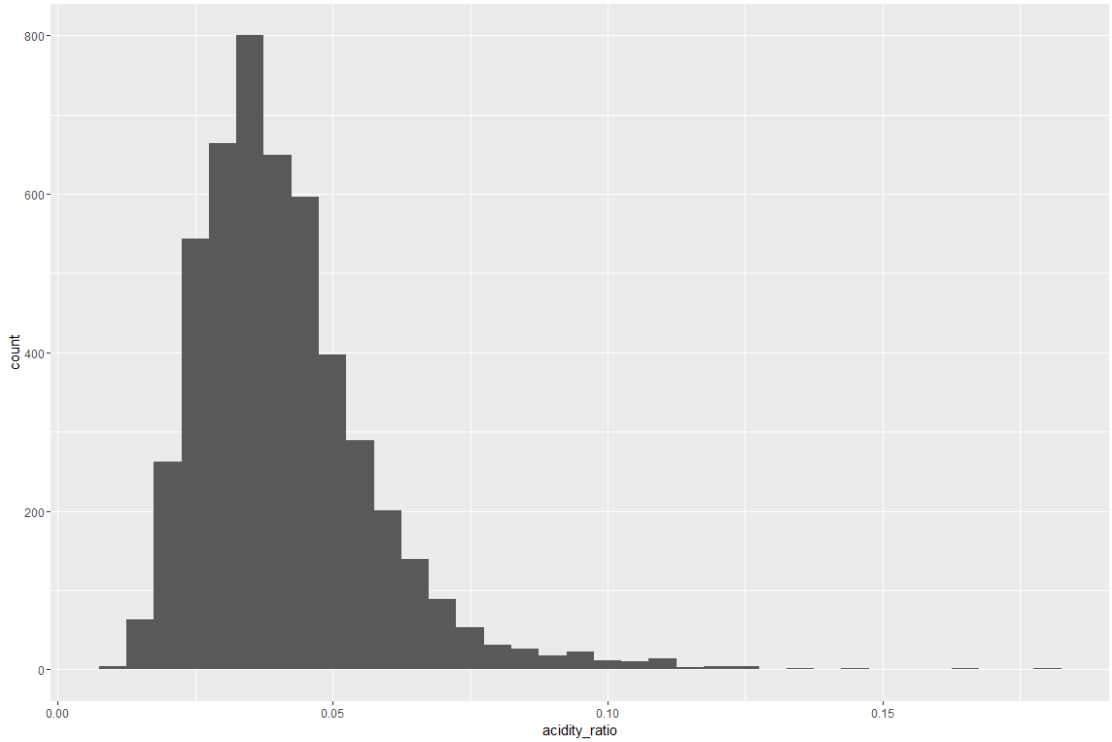





```
##  3  4  5  6  7  8  9
##  1 21 24 16  2  2  0
```

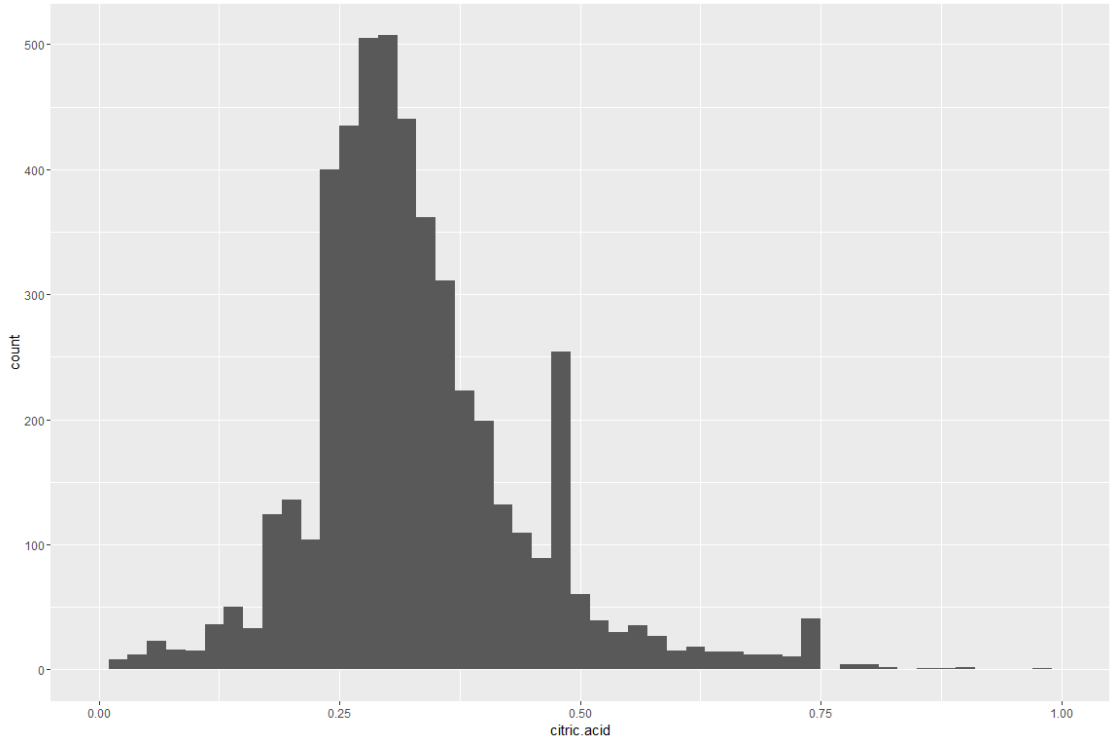
Very long right tail on volatile acidity, with mode around 0.25. Looks normal with log10 x axis transformation. Median of quality score for volatile.acidity>.6 is 5, which is less than the average for the entire dataset.

I want to see if there is any merit to comparing the ratio of volatile to fixed acidity.



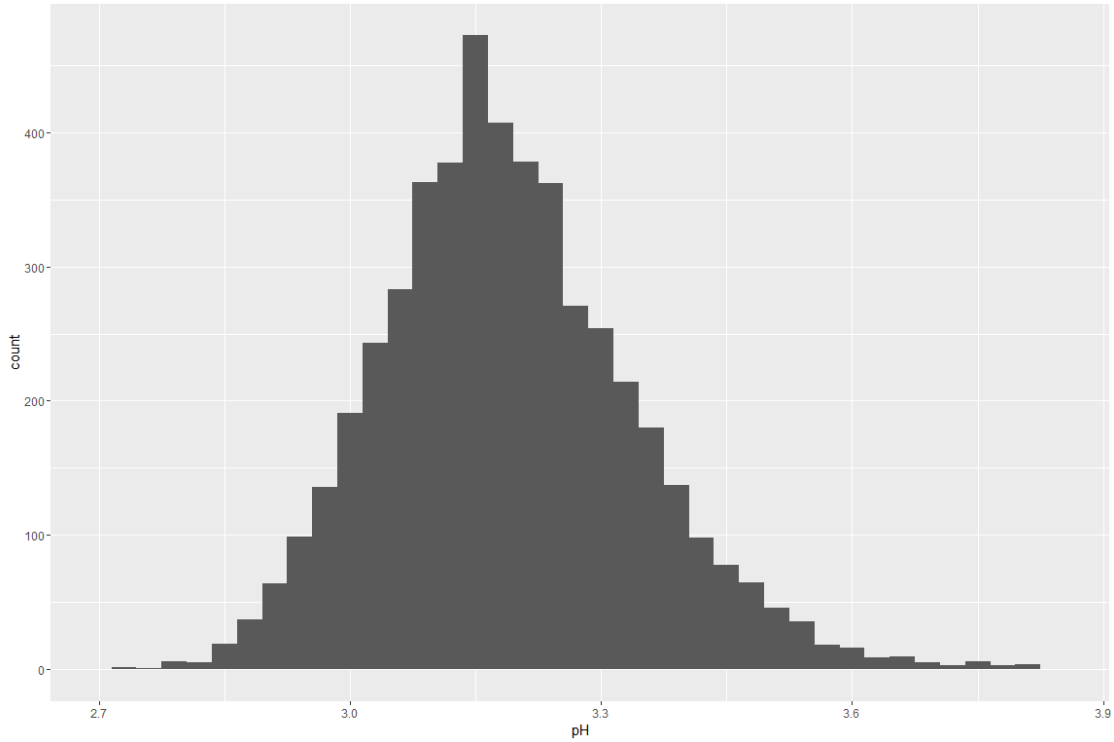

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01111 0.03030 0.03836 0.04126 0.04848 0.18030

##   3   4   5   6   7   8   9
##   7  73 376 371 208  51   2
```

Created a new ratio variable of volatile to fixed acidity. Distribution looks like the volatile acidity distribution, with a long right tail. Looking at the end of the tail quality score, it looks just like the quality score distribution of the original dataset. So nothing to note here.




```
## 3 4 5 6 7 8 9
## 0 6 5 8 0 0 0

##  3  4  5  6  7  8  9
##  0  0 16 17  2  0  0

##   3   4   5   6   7   8   9
##   0  15 151 125  20   3   0
```

Citric acid is an additive that can add freshness and flavor. I see an unusual bump for 0, .5, and .75. If citric acid is an additive, perhaps they add it in such increments. No indication on whether outliers affect quality.



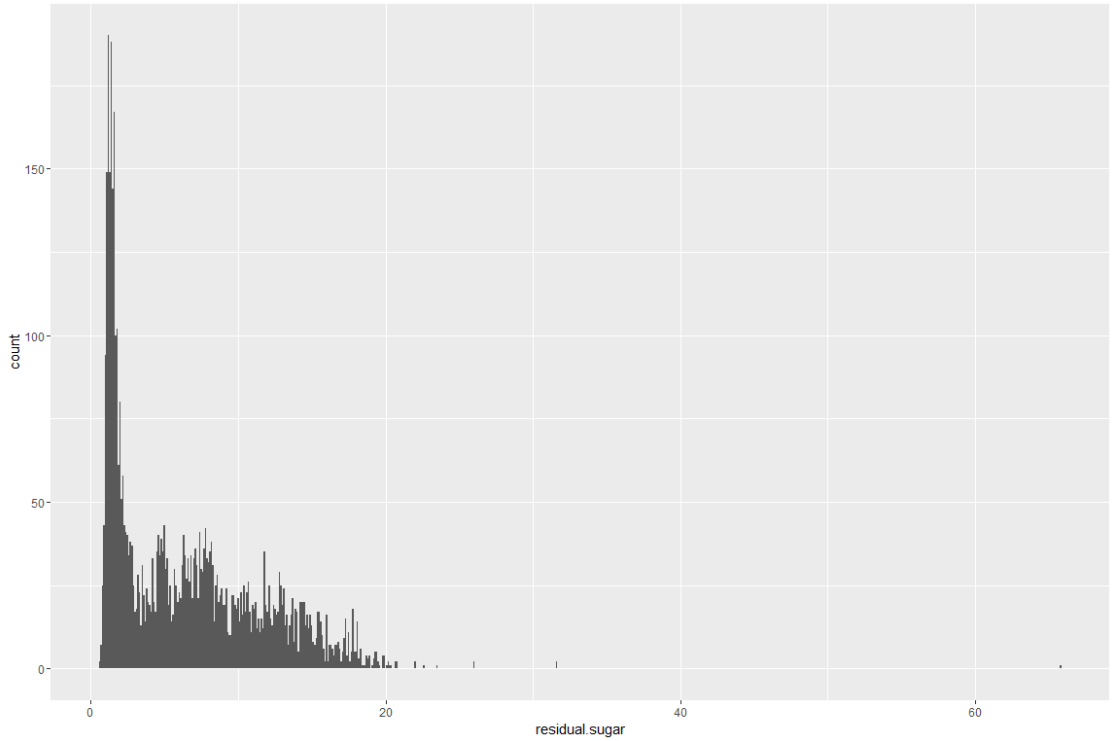

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.720   3.090   3.180   3.188   3.280   3.820
```

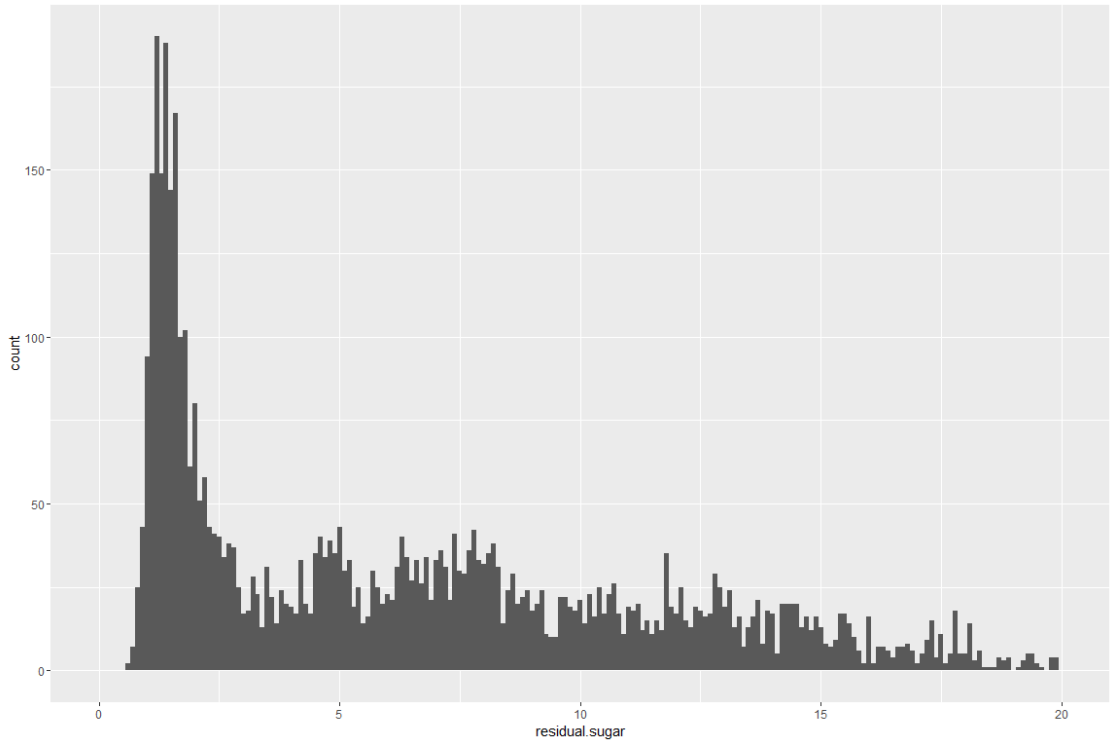
pH is between 2.72 and 3.82, with a mean and median around 3.18, and a fairly normal distribution.

I have now looked at all the acid variables - only the volatile acidity had a pronounced long tail. None of the extreme outliers seem significant for now, in terms of its effect on quality score, although fixed acidity may be worth looking into later.

Sweet & Salty

Now lets look into sweetness and saltiness.

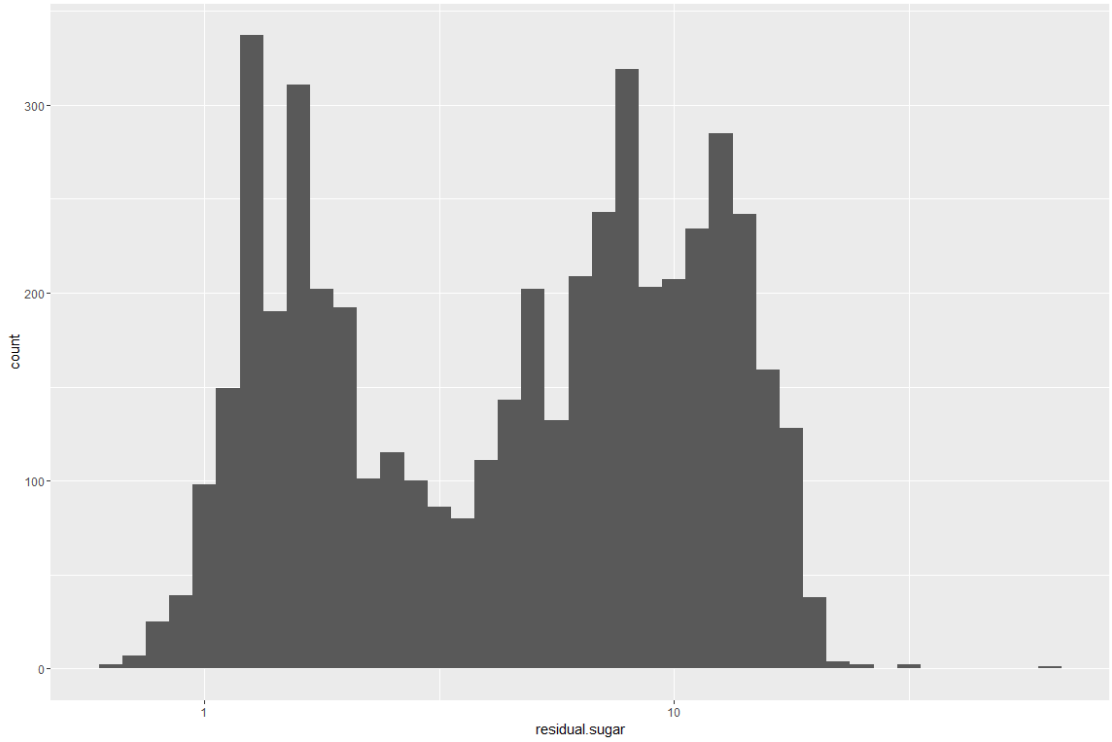





```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.600   1.700   5.200   6.391   9.900  65.800
```

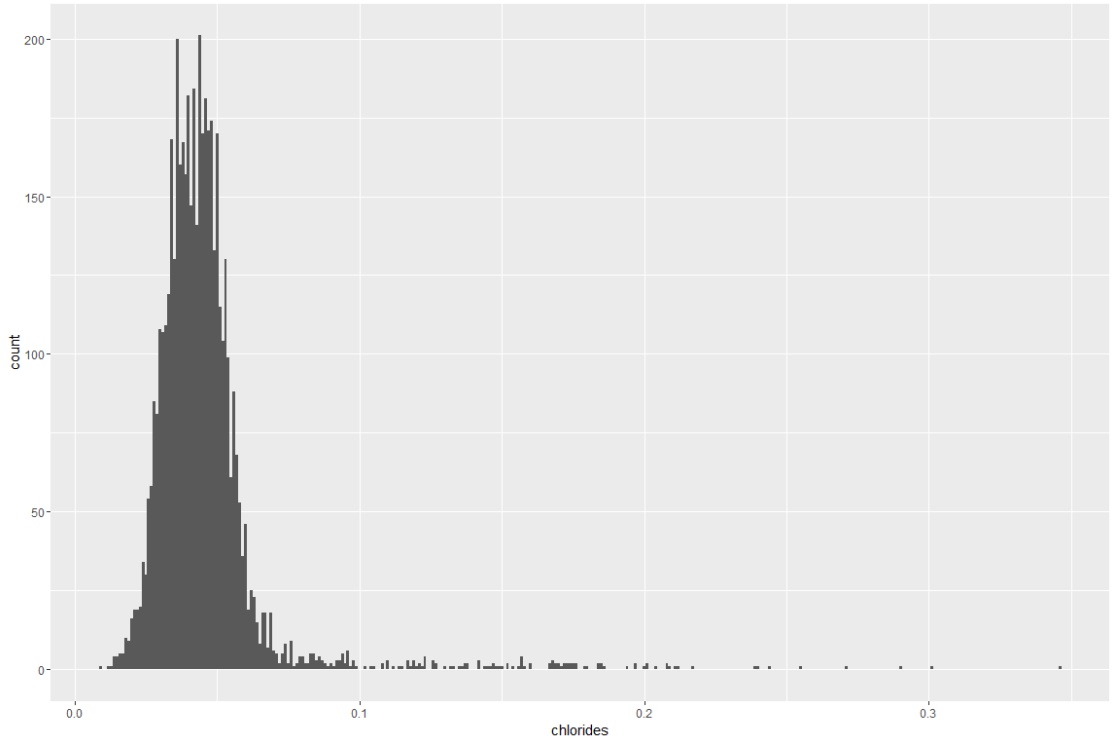
```
##         X fixed.acidity volatile.acidity citric.acid residual.sugar
## 1       1           7.0            0.270        0.36          20.70
## 8       8           7.0            0.270        0.36          20.70
## 183   183           6.8            0.280        0.40          22.00
## 192   192           6.8            0.280        0.40          22.00
## 445   445           6.9            0.240        0.36          20.80
## 1609 1609           6.9            0.270        0.49          23.50
## 1654 1654           7.9            0.330        0.28          31.60
## 1664 1664           7.9            0.330        0.28          31.60
## 2621 2621           6.5            0.280        0.28          20.40
## 2782 2782           7.8            0.965        0.60          65.80
## 2786 2786           6.4            0.240        0.25          20.20
## 2788 2788           6.4            0.240        0.25          20.20
## 3421 3421           7.6            0.280        0.49          20.15
## 3620 3620           6.8            0.450        0.28          26.05
## 3624 3624           6.8            0.450        0.28          26.05
## 3731 3731           6.2            0.220        0.20          20.80
## 4108 4108           6.8            0.300        0.26          20.30
## 4481 4481           5.9            0.220        0.45          22.60
##      chlorides free.sulfur.dioxide total.sulfur.dioxide density   pH
## 1        0.045                 45                  170 1.00100 3.00
## 8        0.045                 45                  170 1.00100 3.00
## 183      0.048                 48                  167 1.00100 2.93
## 192      0.048                 48                  167 1.00100 2.93
## 445      0.031                 40                  139 0.99750 3.20
## 1609     0.057                 59                  235 1.00240 2.98
## 1654     0.053                 35                  176 1.01030 3.15
## 1664     0.053                 35                  176 1.01030 3.15
## 2621     0.041                 40                  144 1.00020 3.14
## 2782     0.074                  8                  160 1.03898 3.39
## 2786     0.083                 35                  157 0.99976 3.17
## 2788     0.083                 35                  157 0.99976 3.17
## 3421     0.060                 30                  145 1.00196 3.01
## 3620     0.031                 27                  122 1.00295 3.06
## 3624     0.031                 27                  122 1.00295 3.06
## 3731     0.035                 58                  184 1.00022 3.11
## 4108     0.037                 45                  150 0.99727 3.04
## 4481     0.120                 55                  122 0.99636 3.10
##      sulphates alcohol quality acidity_ratio
## 1         0.45     8.8       6    0.03857143
## 8         0.45     8.8       6    0.03857143
## 183       0.50     8.7       5    0.04117647
## 192       0.50     8.7       5    0.04117647
## 445       0.33    11.0       6    0.03478261
## 1609      0.47     8.6       5    0.03913043
## 1654      0.38     8.8       6    0.04177215
## 1664      0.38     8.8       6    0.04177215
## 2621      0.38     8.7       5    0.04307692
## 2782      0.69    11.7       6    0.12371795
## 2786      0.50     9.1       5    0.03750000
```

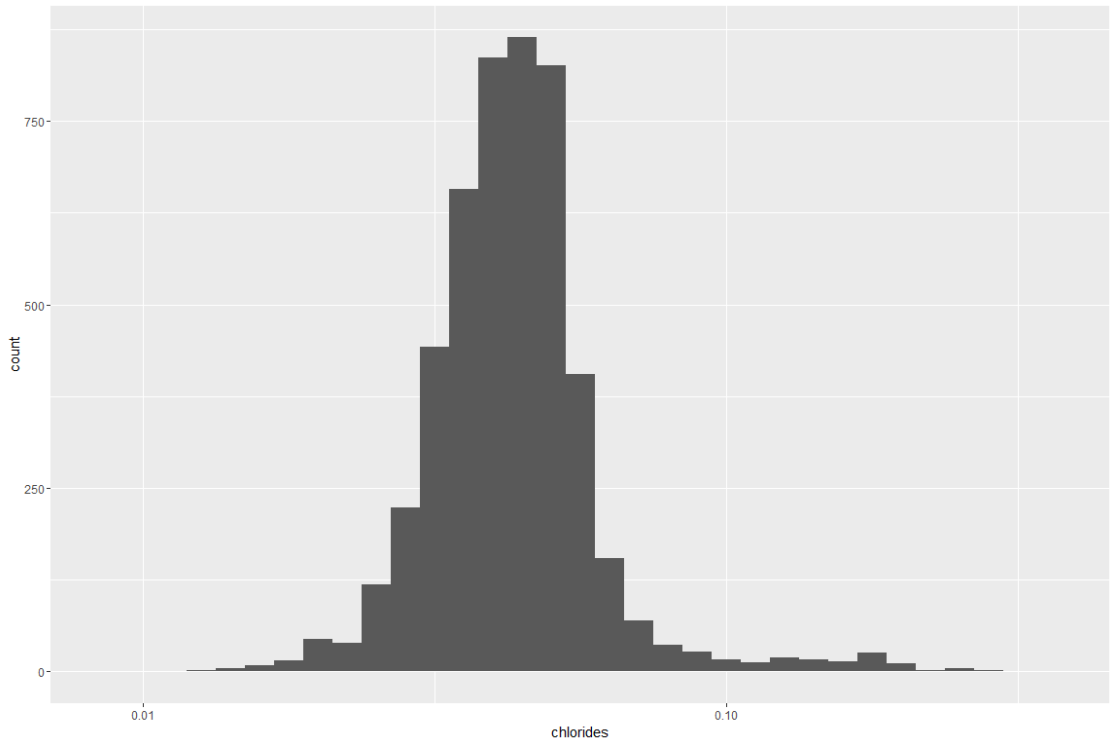
```
## 2788      0.50     9.1        5    0.03750000
## 3421      0.44     8.5        5    0.03684211
## 3620      0.42    10.6        6    0.06617647
## 3624      0.42    10.6        6    0.06617647
## 3731      0.53     9.0        6    0.03548387
## 4108      0.38    12.3        6    0.04411765
## 4481      0.35    12.8        5    0.03728814
```



Very weird distribution of residual sugar. Could be bimodal, when looking at log10 scale of the data. Majority of white wines are grouped around 1, and then a long tail with perhaps another peak around 10. This could be due to the styles of wine being produced, a sweeter variety vs a dryer one.

There is one extreme outlier, with residual sugar off the charts at 65.80 - it is also an outlier for other variables as well. I think this can be ignored because it only ranked 6 for quality, even when it was significantly different than most wines.



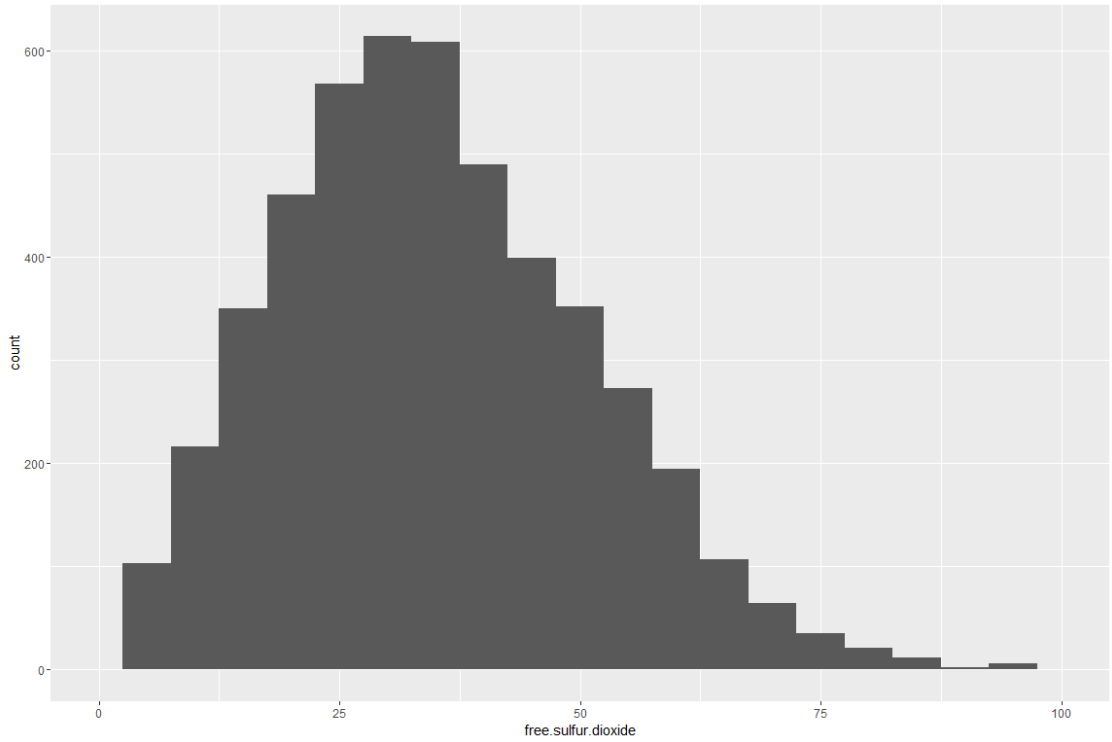



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00900 0.03600 0.04300 0.04577 0.05000 0.34600

##   3   4   5   6   7   8   9
##   7  57 509 470  97  23   0
```

Chlorides, which measures the saltiness of the wine, is centered around .043, with a very long right skewed tail. The tail quality score is representative of the entire data set, so not sure if high chloride content has any bearing on quality.

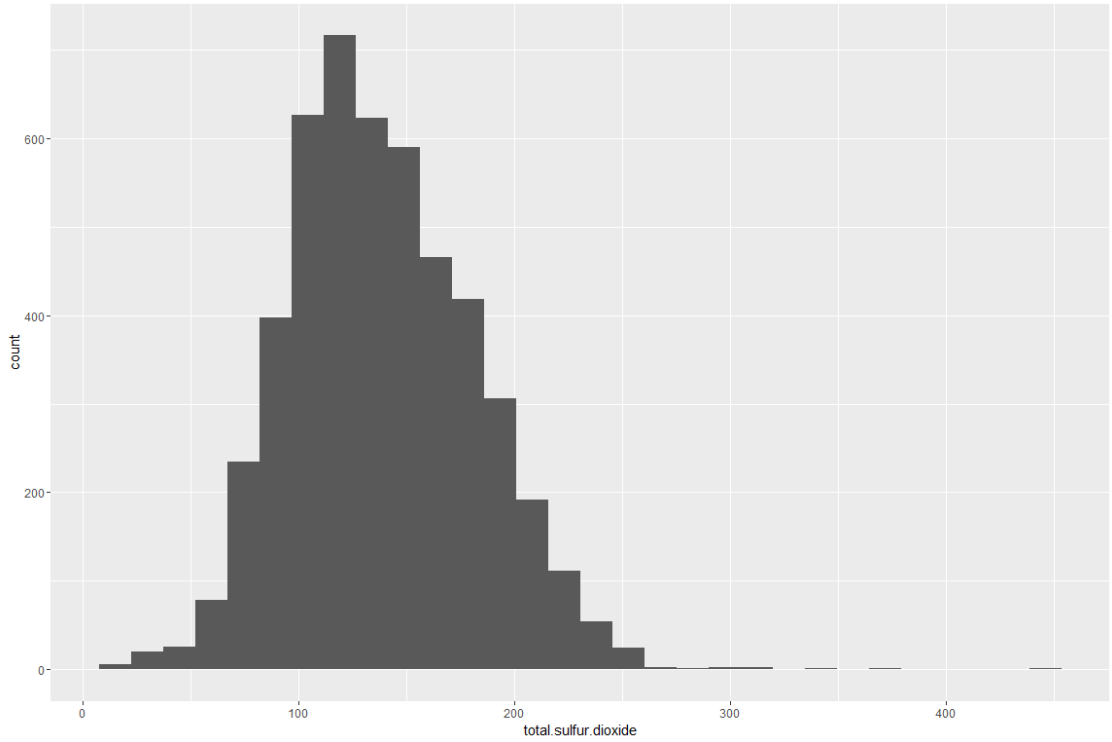
Sulfur

Let's look at free sulfur, total sulfur, and sulphates.

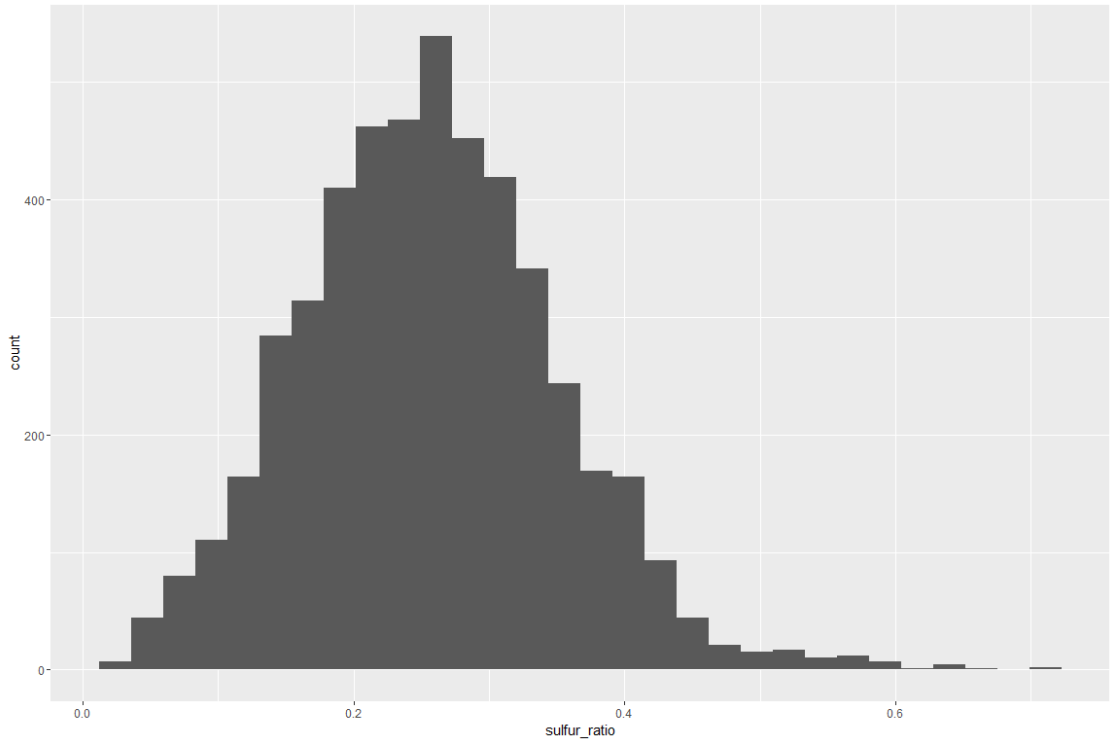



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00   23.00   34.00   35.31   46.00  289.00

## 3 4 5 6 7 8 9
## 4 2 5 3 1 2 0
```

Free sulfur dioxide is a bit right skewed with a long tail. Looking at the extreme outlier of 289, this wine got a 3 for its quality score. This could be a factor for its low score, since it may be over the limit of being detectable to taste. Something to keep an eye on.

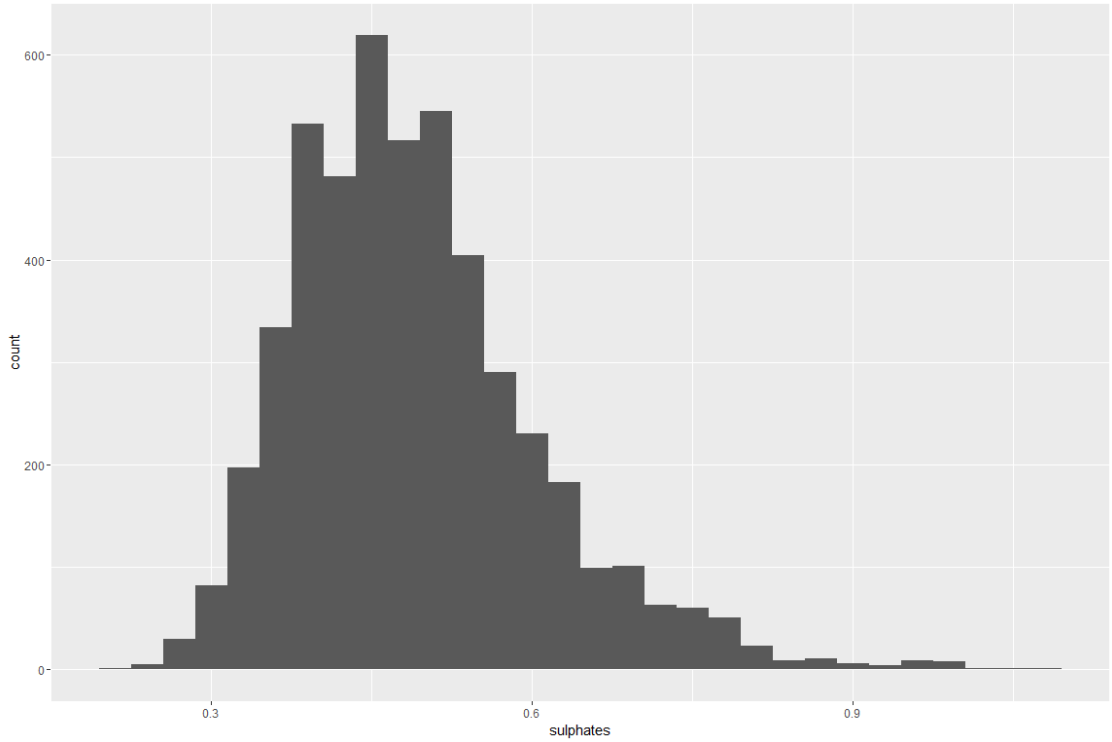



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.02362 0.19090 0.25370 0.25560 0.31580 0.71050
```




```
##   3   4   5   6   7   8   9
##   5   6  57 153  72  16   1
```

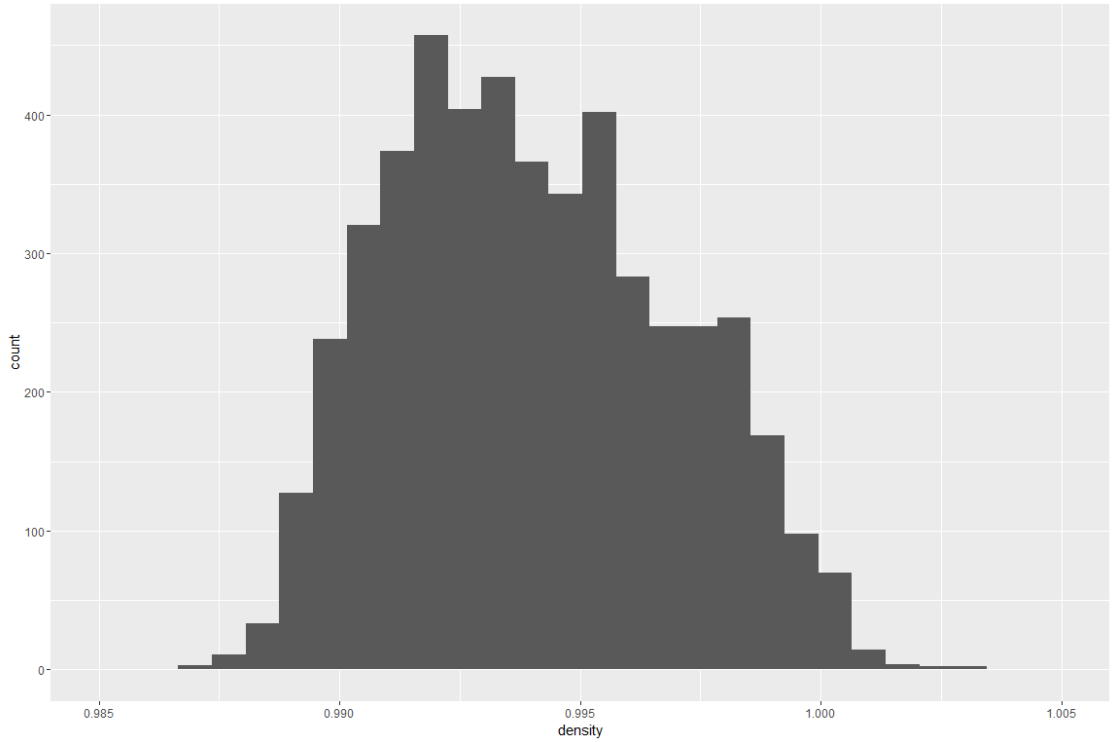
Total sulfur had a similar distribution to free sulfur, so I wanted to create a new variable of the ratio of free to total SO2. Didn't see anything unusual about the ratio - it also had the same shape distribution.




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

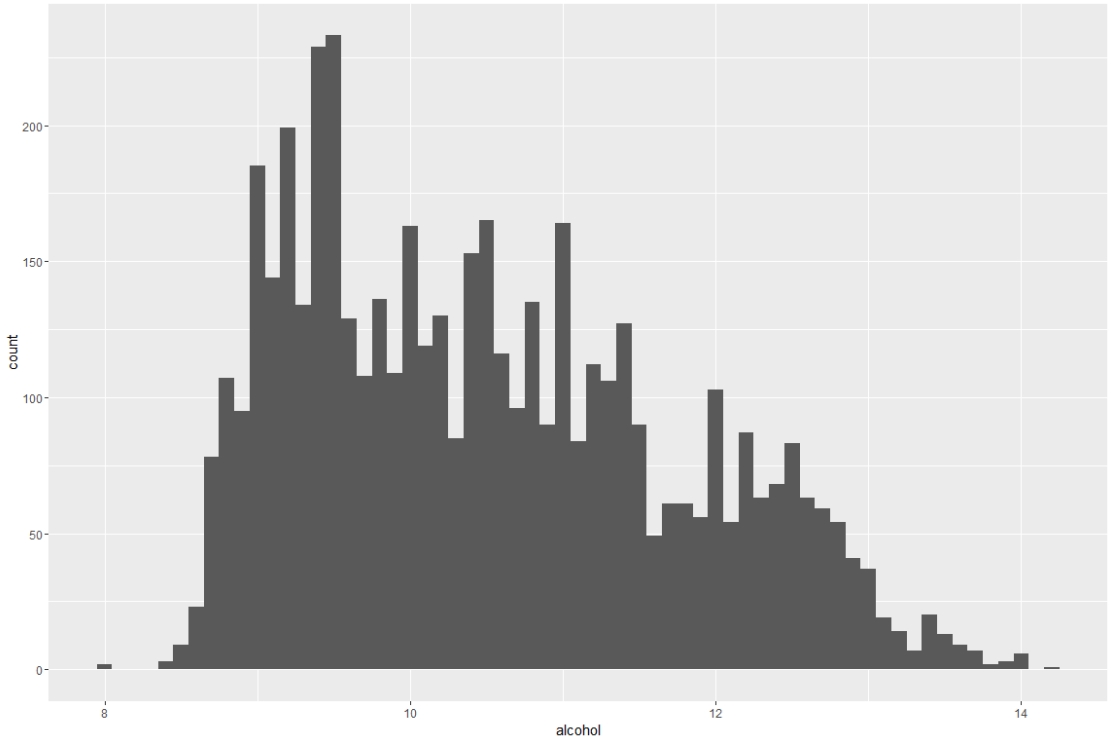
Sulphates, which is an additive, has a similar distribution of the previous sulfur variables I looked at. Most are around 0.48, with a long tail. Outliers do not seem to have an effect on quality.

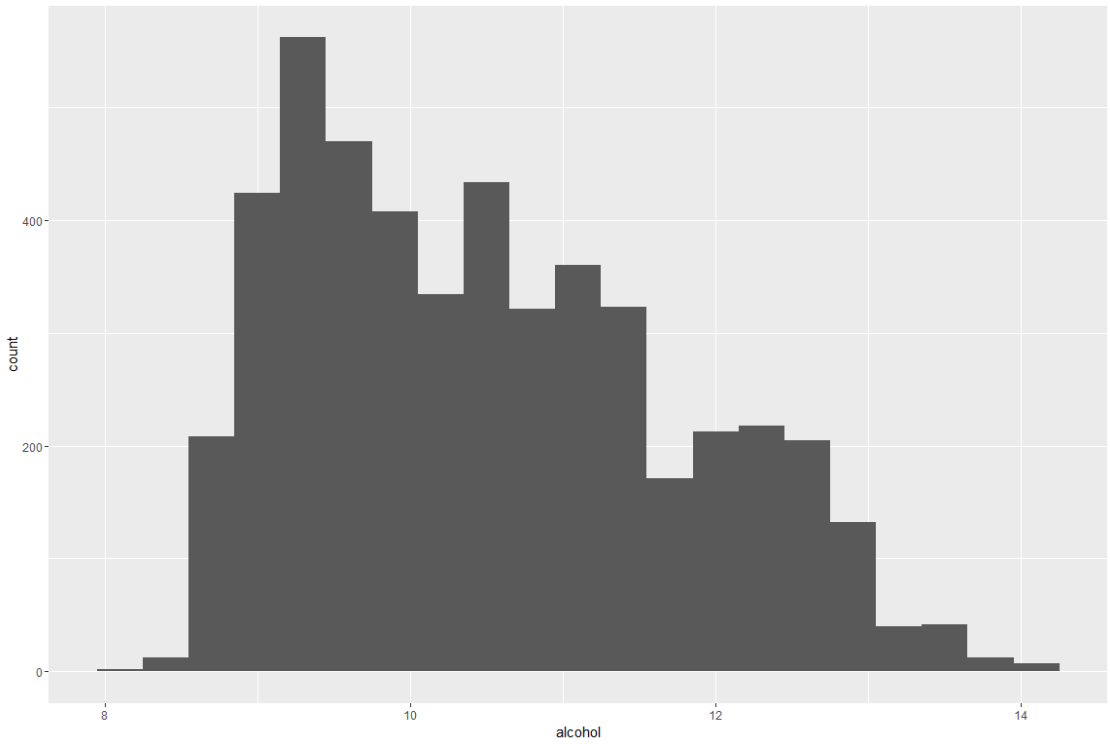
Density and Alcohol %




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

Density distribution doesn't really look normal - there seems to be a couple peaks. One outlier, but the quality score was only a 6 for that.

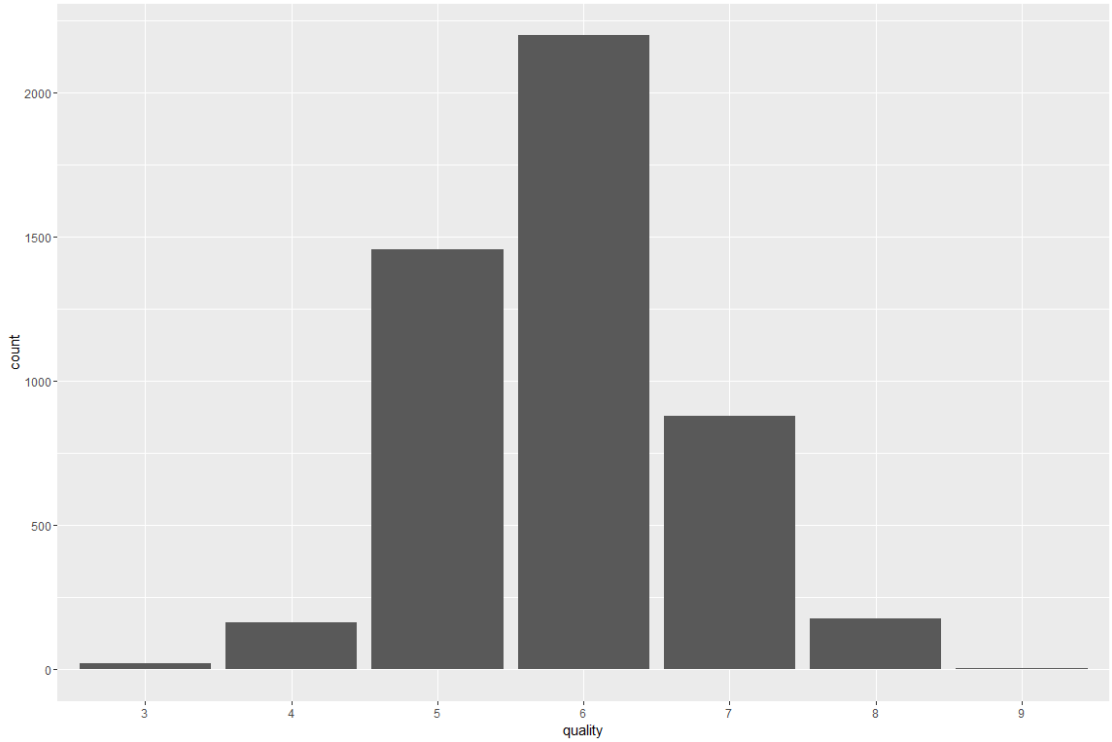





```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00    9.50   10.40   10.51   11.40   14.20
```

Alcohol % is right skewed, but the tail isn't that long. There is a mode at around 9.5, and the median and mean are around 10.5. I would be interested to see what alcohol level has to do with quality score.

Quality Score




```
##
##    3    4    5    6    7    8    9
##   20  163 1457 2198  880  175    5

##    3    4    5    6    7    8    9
##   20  163 1457 2198  880  175    5
```

50% of the data got a 5 or 6 quality score. There were 5 wines that got a high of 9, and 20 wines with the lowest score of 3. In this analysis, it is important to look at the outliers, since they may hold the key to identifying the best and worst traits of wine. I would consider 3-4 as bad, and 8-9 as good. 5-7 would be average.

Univariate Analysis

What is the structure of your dataset?

```
## 'data.frame':    4898 obs. of  15 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             : Ord.factor w/ 7 levels "3"<"4"<"5"<"6"<..: 4 4 4
4 4 4 4 4 4 4 ...
##  $ acidity_ratio       : num  0.0386 0.0476 0.0346 0.0319 0.0319 ...
##  $ sulfur_ratio        : num  0.265 0.106 0.309 0.253 0.253 ...

## [1] 0.9258881
```

There were initially 12 variables (not counting X), mostly of numerical type. I changed quality to an ordered factor, to signify that quality should be treated as an ordered categorical variable rather than just an integer value. I added a couple variables, acidity ratio and sulfur ratio, to see if there may be a relationship of the percentage of acidity or sulfur that may trigger a positive or negative taste.

The variables measure acidity, sugar levels, saltiness, sulfur levels, density, pH, and alcohol level. The quality variable is ranked on a scale of 0-10, 0 being the worst and 10 being the best. In this data set, the lowest score given was a 3, and the highest grade was a 9. 92.6% of the wines received a quality score of 5,6, or 7.

There are some extreme outliers for some of these variables. One such outlier was a wine with a residual sugar level of 65.80. It did not seem as if these extreme outliers single-handedly contributed to the quality score, positively or negatively.

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

The main feature of interest that I want to explore is the effect of various variables on the quality score of the wine. The quality of the wine is graded on a 1 to 10 scale, 1 being the worst and 10 being the best, by at least three experts and taking the median of their grade. It would be interesting to identify what aspect of wine makes it rank better.

In this dataset, I am most interested in the quality score outliers - what makes the worst wines (those that rank 3 or 4), and what makes the best wines (8 or 9).

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

A lot of these features have an aspect of taste associated with them. Residual sugar in moderation may be appealing. Too much volatile acidity or total sulfur dioxide may be repulsive. Chloride appeals to another part of the human tastes, saltiness. I don't think any one variable will determine the quality score of the wine, but perhaps a combination of

these taste profiles, each tuned to a particular part of the spectrum, may result in a wine hitting the proverbial "sweet spot" of wine.

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

I created a couple ratio variables - the acidity ratio and the sulfur ratio. The acidity ratio is the ratio of fixed acidity to volatile acidity. The sulfur ratio is the free sulfur dioxide divided by the total sulfur dioxide.

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?

Residual sugar had an unusual distribution. It had a really high peak around 1, and then a very unusual tail with the possibility of other smaller peaks. I transformed the x axis to a log10 to get a better idea of the behavior of this mysterious tail. The transformation revealed a bi-modal distribution, with a second peak around 10. I surmised that these other nodes may be a result of different wine types that may be on separate spectrums of sweetness. Each wine type may be on its own distribution of sweetness.

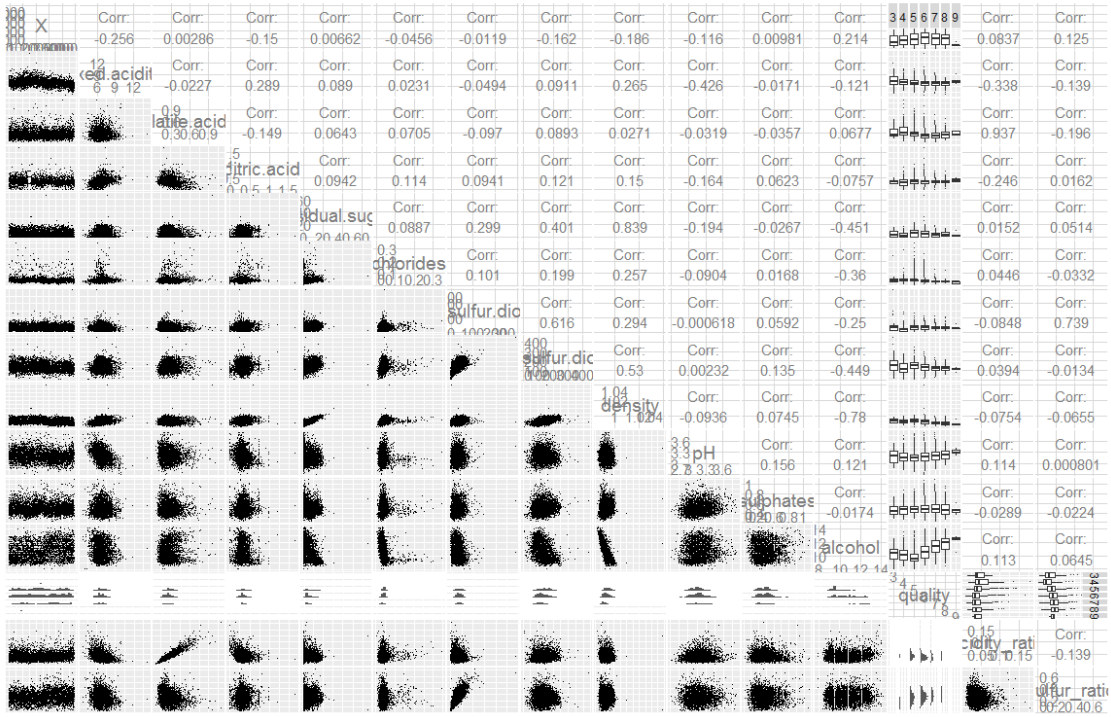
I also changed quality score to an ordered factor, so that it would be treated as a categorical variable instead of just integers.

Bivariate Plots Section

Initial Assessment

```
##                       fixed.acidity volatile.acidity citric.acid
## fixed.acidity            1.00000000      -0.02269729  0.28918070
## volatile.acidity        -0.02269729       1.00000000 -0.14947181
## citric.acid              0.28918070      -0.14947181  1.00000000
## residual.sugar           0.08902070       0.06428606  0.09421162
## chlorides                0.02308564       0.07051157  0.11436445
## free.sulfur.dioxide     -0.04939586      -0.09701194  0.09407722
## total.sulfur.dioxide     0.09106976       0.08926050  0.12113080
## density                  0.26533101       0.02711385  0.14950257
## pH                      -0.42585829      -0.03191537 -0.16374821
## sulphates               -0.01714299      -0.03572815  0.06233094
## alcohol                 -0.12088112       0.06771794 -0.07572873
##                       residual.sugar   chlorides free.sulfur.dioxide
## fixed.acidity             0.08902070  0.02308564       -0.0493958591
## volatile.acidity          0.06428606  0.07051157       -0.0970119393
## citric.acid               0.09421162  0.11436445        0.0940772210
## residual.sugar            1.00000000  0.08868454        0.2990983537
## chlorides                 0.08868454  1.00000000        0.1013923521
## free.sulfur.dioxide       0.29909835  0.10139235        1.0000000000
## total.sulfur.dioxide      0.40143931  0.19891030        0.6155009650
## density                   0.83896645  0.25721132        0.2942104109
```

```
## pH                       -0.19413345 -0.09043946       -0.0006177961
## sulphates                -0.02666437  0.01676288        0.0592172458
## alcohol                  -0.45063122 -0.36018871       -0.2501039415
##                       total.sulfur.dioxide     density            pH
## fixed.acidity                 0.091069756  0.26533101 -0.4258582910
## volatile.acidity              0.089260504  0.02711385 -0.0319153683
## citric.acid                   0.121130798  0.14950257 -0.1637482114
## residual.sugar                0.401439311  0.83896645 -0.1941334540
## chlorides                     0.198910300  0.25721132 -0.0904394560
## free.sulfur.dioxide           0.615500965  0.29421041 -0.0006177961
## total.sulfur.dioxide          1.000000000  0.52988132  0.0023209718
## density                       0.529881324  1.00000000 -0.0935914935
## pH                            0.002320972 -0.09359149  1.0000000000
## sulphates                     0.134562367  0.07449315  0.1559514973
## alcohol                      -0.448892102 -0.78013762  0.1214320987
##                         sulphates     alcohol
## fixed.acidity         -0.01714299 -0.12088112
## volatile.acidity      -0.03572815  0.06771794
## citric.acid            0.06233094 -0.07572873
## residual.sugar        -0.02666437 -0.45063122
## chlorides              0.01676288 -0.36018871
## free.sulfur.dioxide    0.05921725 -0.25010394
## total.sulfur.dioxide   0.13456237 -0.44889210
## density                0.07449315 -0.78013762
## pH                     0.15595150  0.12143210
## sulphates              1.00000000 -0.01743277
## alcohol               -0.01743277  1.00000000
```

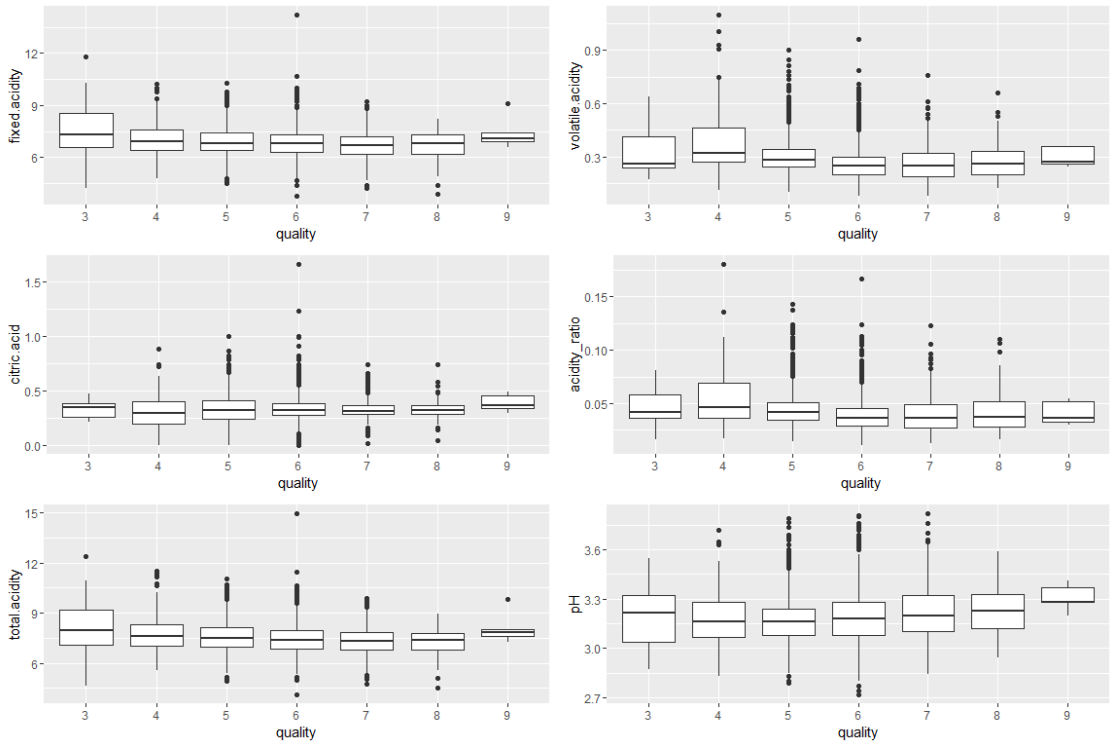


I want to take a closer look at the acidity variables, residual sugar, density, and alcohol.

Boxplots for quality scores: I see a decreasing quality score with total sulfur dioxide, and density. Also see pH and alcohol medians increase as quality score increases. Clearest indicator of quality differences is from alcohol content.

Correlation values: There is a strong correlation between residual sugar and density - seems to be a linear relationship. Alcohol seems to have a lot of strong predictor variables. Since the boxplot showed high alcohol content to be related to higher quality scores, these variables will be important to explore.

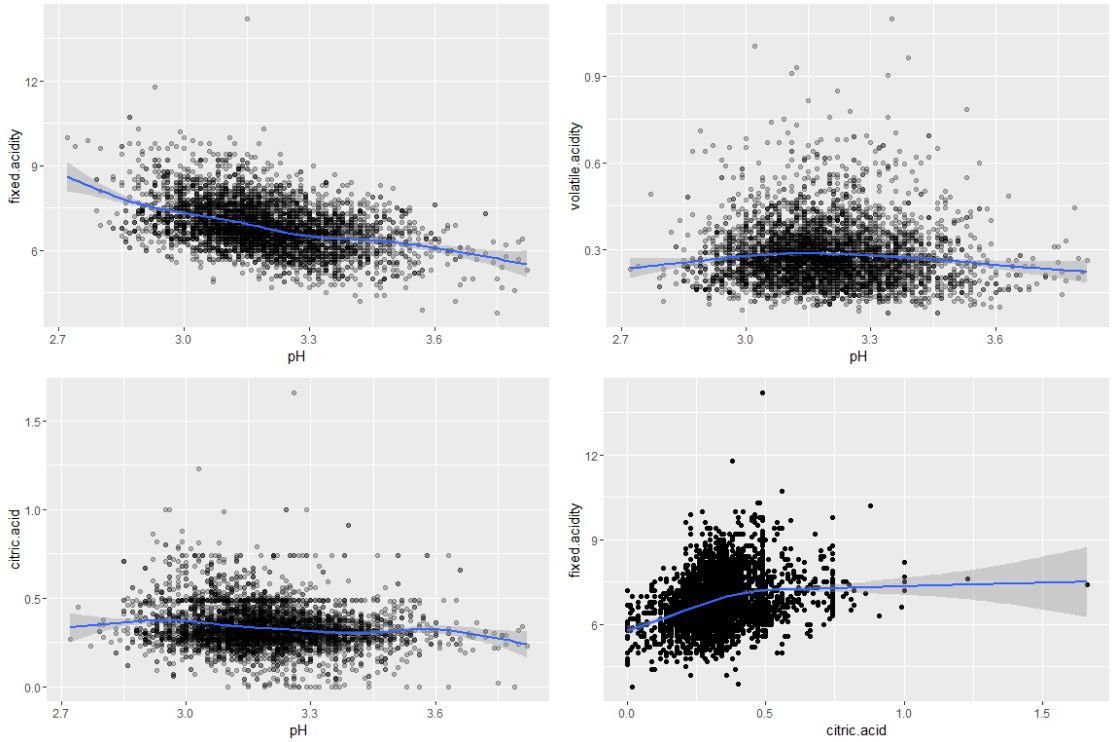
Acidity to Quality



Looking at these 6 panels related to acidity - The only thing that stands out is the pH. Increasing quality scores is related to increasing pH. I don't think the other acidity variables are related to wine scores.

I tried creating another variable - total acidity. This added up all the variables related to the acidity, and then compared it to the pH. This new variable just looks like a copy of the fixed acidity variable - without a weight to assign to each acidic variable, the fixed acidity overwhelms the new variable and renders it useless.

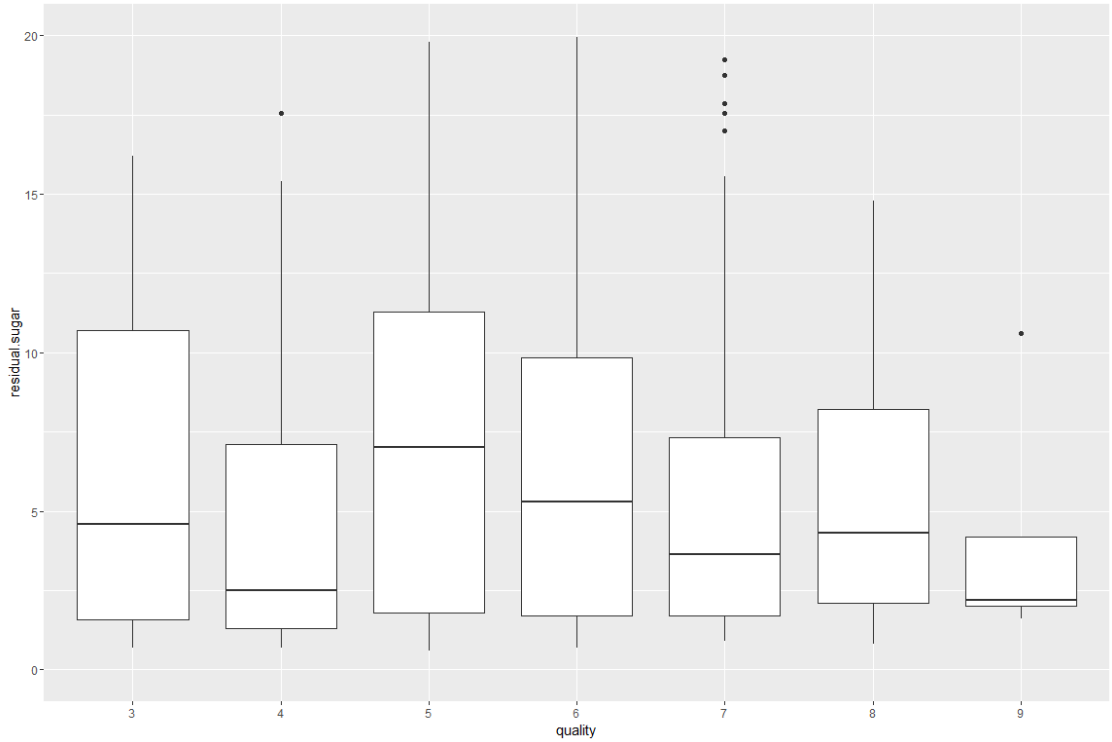
Acidity and pH



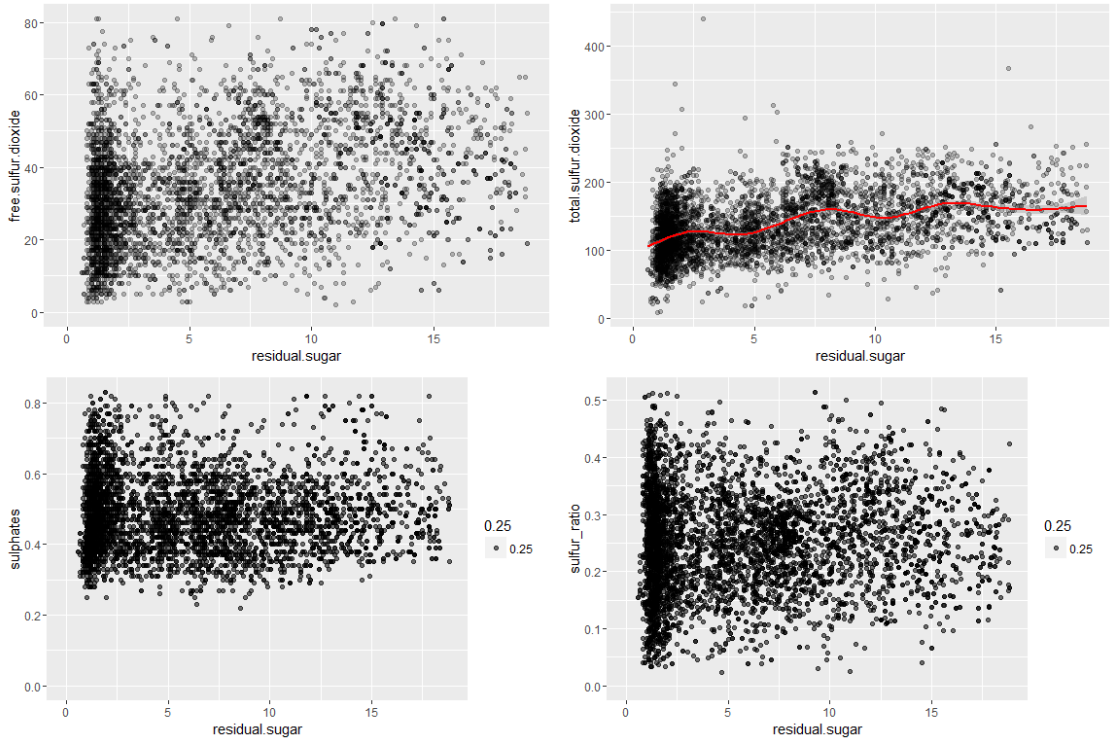
I thought there may be more of a relationship between pH and the acidic variables, but there isn't a clear relationship here. Fixed acidity has a slight relationship with pH: as fixed acidity decreases, pH gets more basic. This makes sense, but doesn't really reveal anything profound.

In the 4th panel, I put citric acid with fixed acidity. I think the positive relationship I see here is because adding citric acid will increase the fixed acidity levels, since citric acid is acidic. Nothing interesting here.

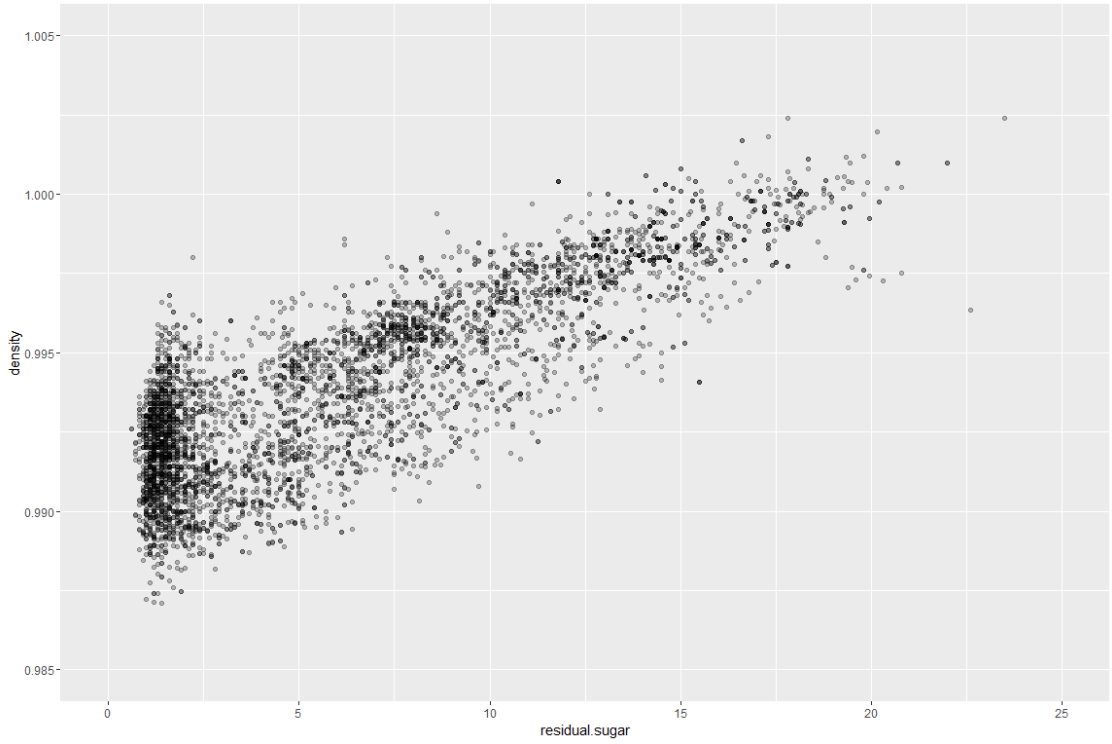
Residual Sugar Analysis

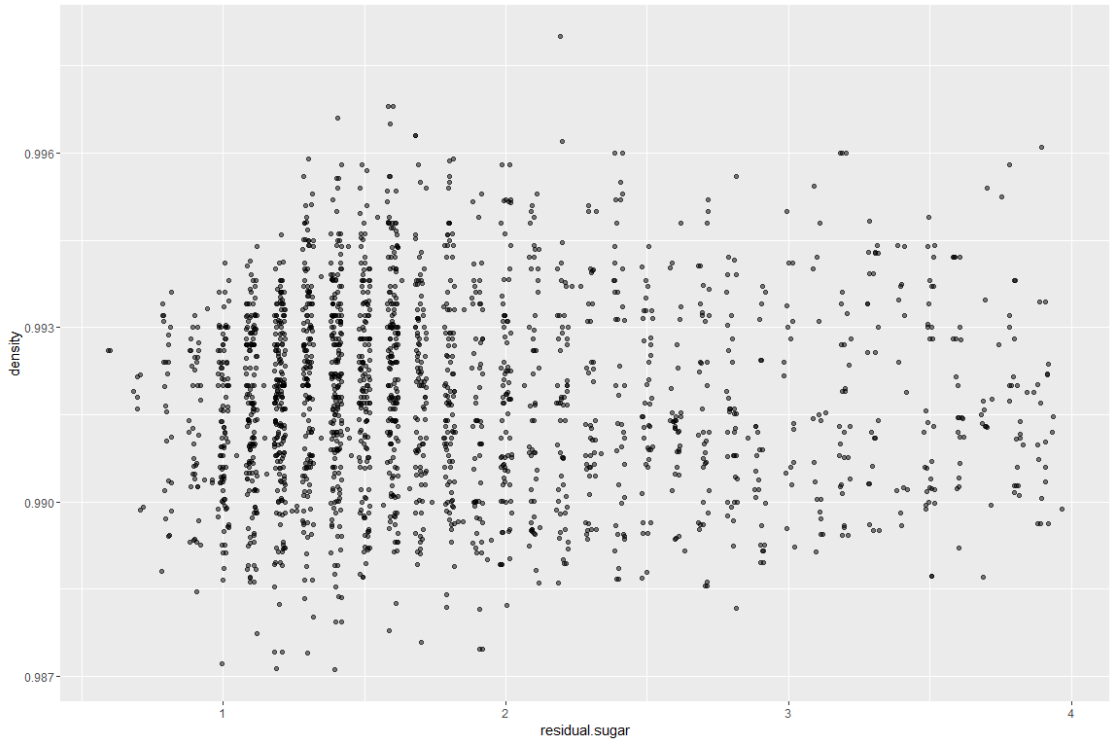


Median residual sugar levels vary with quality in no distinguishable pattern. High quality wines have lower median sugar levels, but so does lower quality wines.



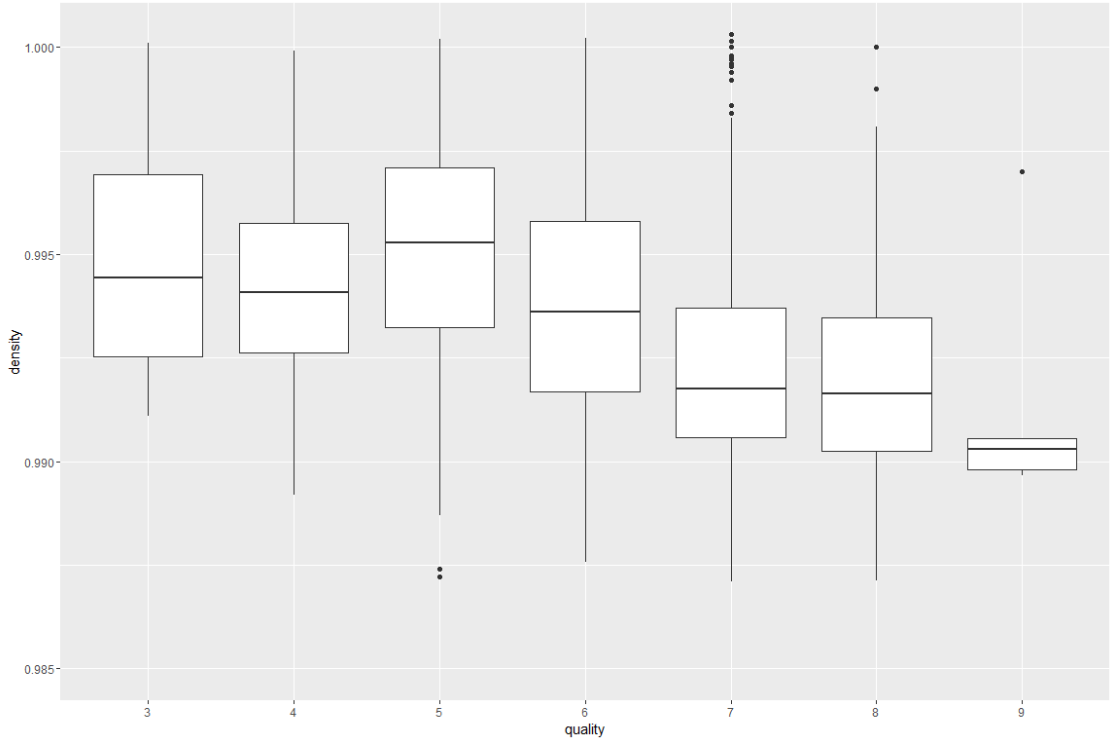
Sugar and sulfur variables seem like a dead end. I thought residual sugar and the sulfur variables would have more clear relationship. Total sulfur dioxide vs residual sugar was the best relationship between these four comparisons.





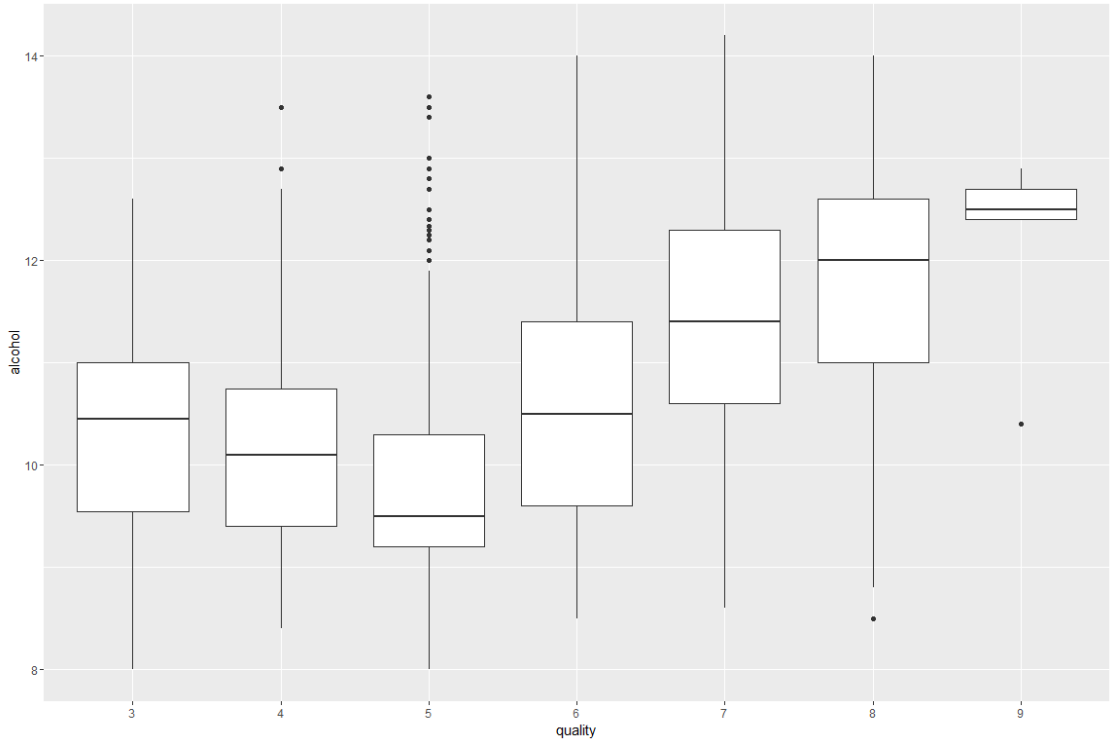
Residual sugar and density had the highest correlation value of any variable pair. There are 2 clusters of data points, one around the residual sugar value of 1, and then a long tail >4. There may be a linear relationship for the tail - as residual sugar increases, density rises. A closeup of the cluster around 1 reveals there's not much of a pattern in the set of residual sugar under 4.

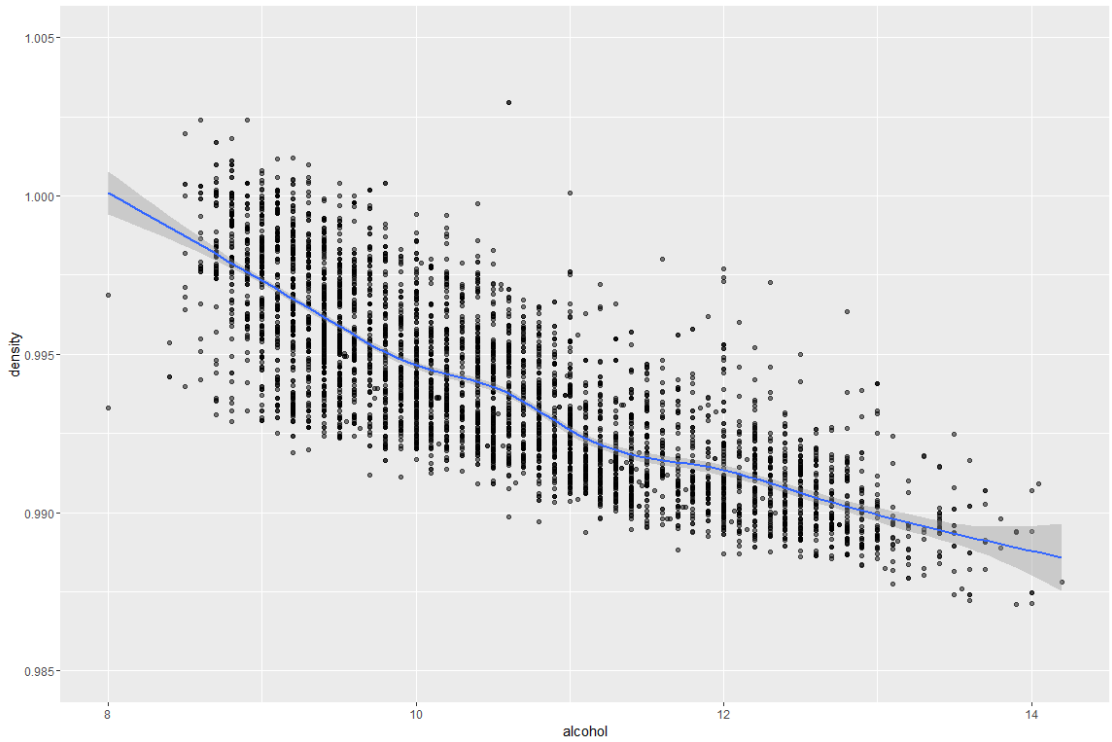
Density Bivariate Analysis



Higher quality wines are less dense. I think there might be more insight if I delve deeper into density. Density and alcohol had a high correlation.

Density and Alcohol






```
##
##  Pearson's product-moment correlation
##
## data:  wine$density and wine$alcohol
```

```
## t = -87.2549, 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
```

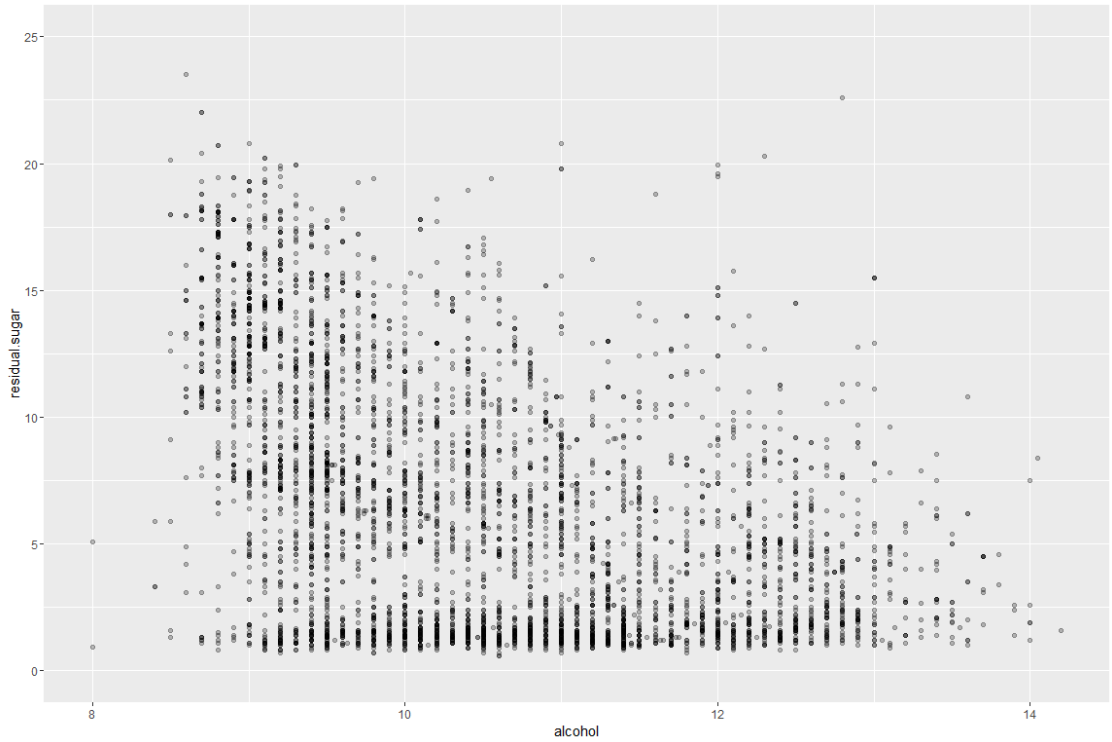
There is a clear linear relationship between density and alcohol. As alcohol content increases, density also decreases. I learned previously from the boxplot of density that high quality wines tends to have lower density. These findings are consistent. Looking at alcohol content boxplots by quality, we see that there is a very evident distinction of alcohol content in higher quality wines. I think there may be relationship between sugar, density, and alcohol content that explains the higher quality wines.

The density of sugar may explain some of these relationships as well. If there is more sugar in wine, the density will be higher. I know that higher quality wines are less dense. I saw previously from the tail end of the scatterplot that density increases as sugar increases.

Maybe the quality has something to do with the ratio of sugar to alcohol.

Alcohol Bivariate Analysis

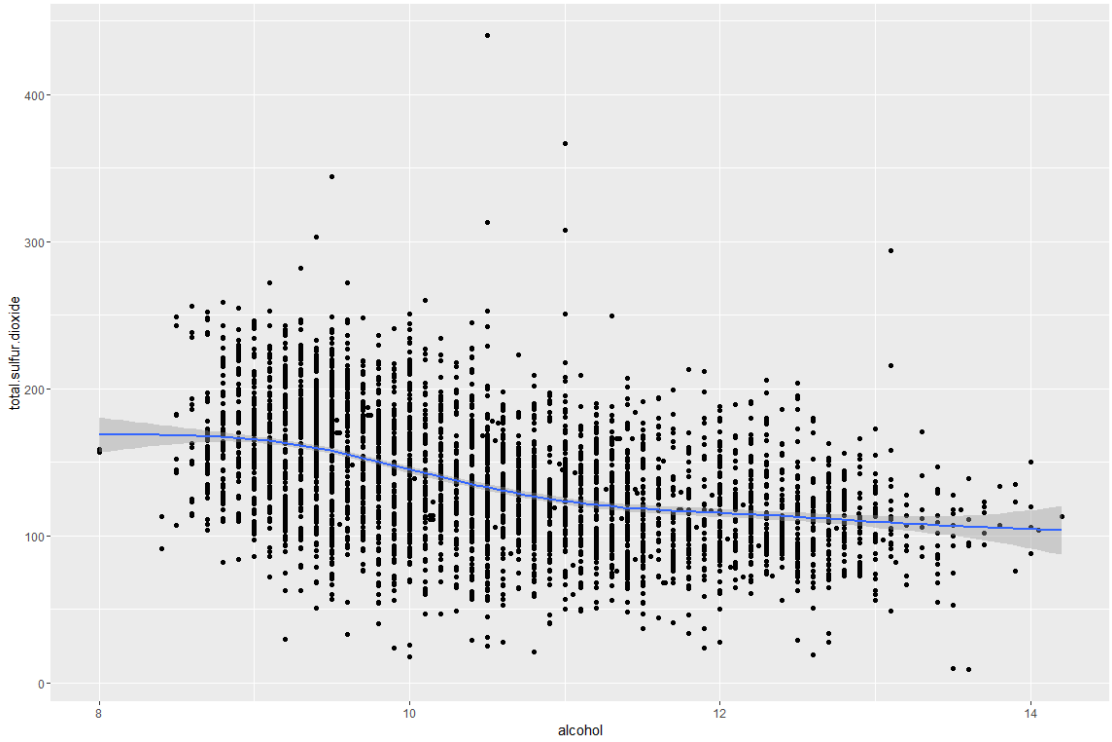
I just explored alcohol and density




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

```
## data:  wine$residual.sugar and wine$alcohol
## t = -35.3209, 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
```

Not a great relationship between alcohol and residual sugar. Cor test= -.451




```
##
##  Pearson's product-moment correlation
##
## data:  wine$total.sulfur.dioxide and wine$alcohol
## t = -35.1501, 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
```

There is a bit of a negative relationship. We know higher quality wines are located in the higher alcohol range - it makes sense that higher alcohol levels is related to lower S02 since a smaller dose of SO2 is undetectable and does not get in the way of taste. Still, not a great relationship.

I think I have identified the few variables I want to focus on in the final part of my analysis.

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?

I used boxplots to see how quality scores varied with other variables in the data set. When looking at the acidic and sulfur variables, there were not distinguishable differences between the quality. The most significant variable was alcohol - this revealed that higher quality wines had a high alcohol content. I wanted to know why.

There were some highly correlated variables when paired with alcohol, namely residual sugar, density, and total sulfur dioxide. Unfortunately, only density seemed to have a clear relationship with alcohol. Now I needed to see what related to density. Digging deeper, residual sugar is closely related to density.

I looked at most of the variables related to acidity and sulfur, but they revealed very little insight into any clear relationship with each other, or other variables.

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

After identifying alcohol as a good predictor of good wine, I needed to explore why this was so. I found interesting relationships between alcohol and density. As alcohol increased, density decreased. The cor test value was -0.78.

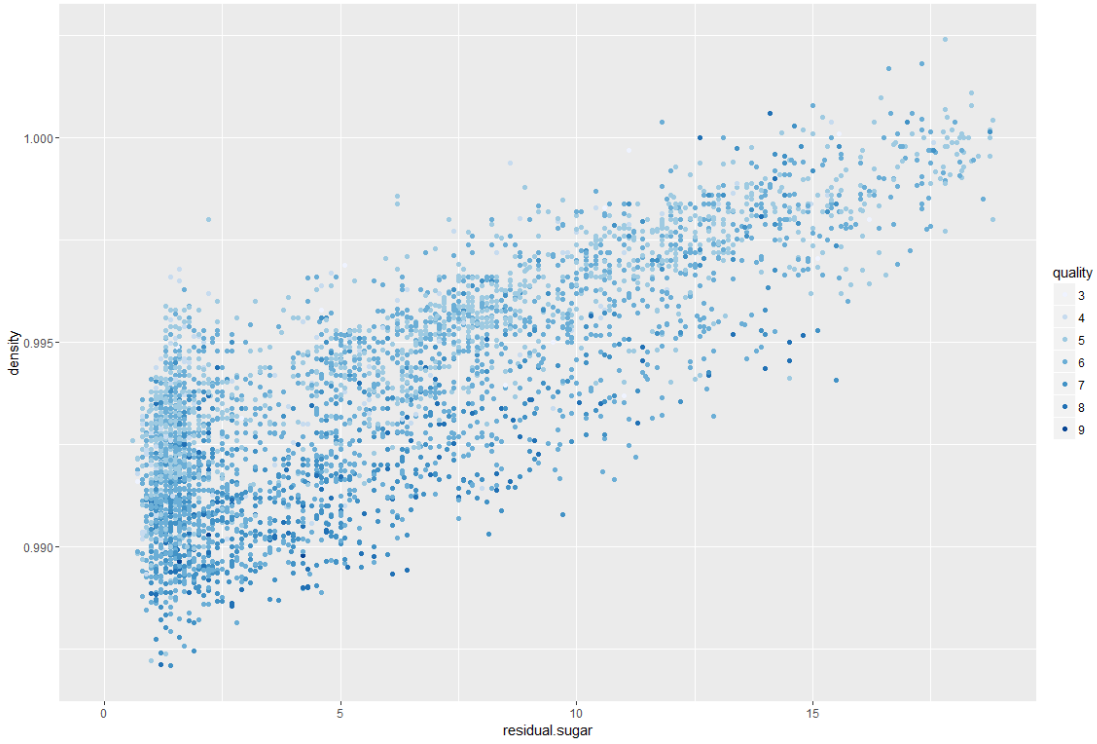
There was another interesting relationship: density and residual sugar. This one was interesting because there seemed to be two sets of data clusters. When I isolated the data cluster to focus on one part, there seemed to be a linear relationship between sugar content and density. This makes sense since more sugar would increase density.

What was the strongest relationship you found?

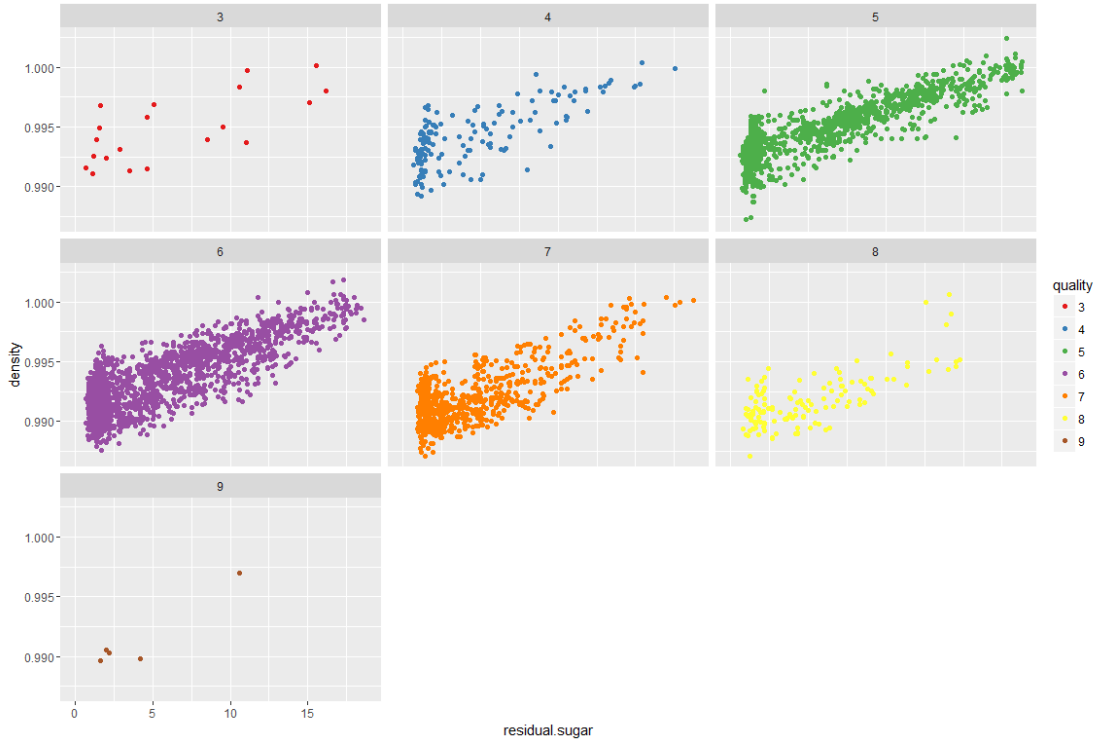
Using a correlation matrix, the highest correlation is between density and residual sugar, with a value of 0.839.

Multivariate Plots Section

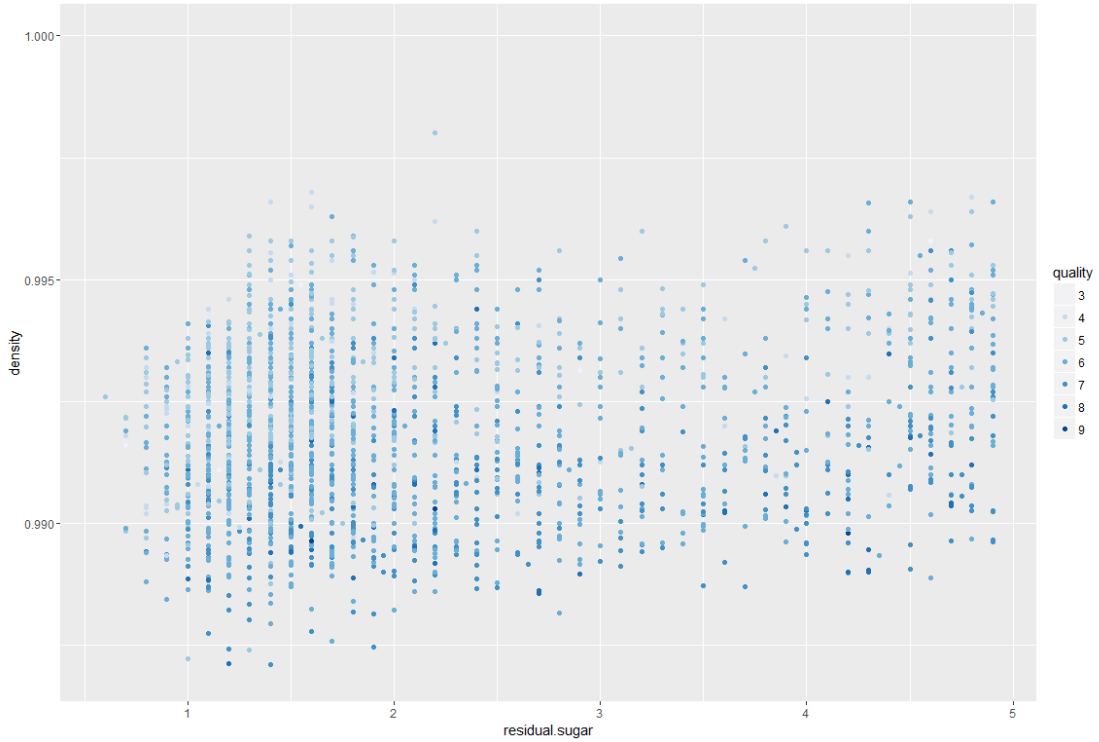
I want to look at the relationship between residual sugar and density, with respect to



quality

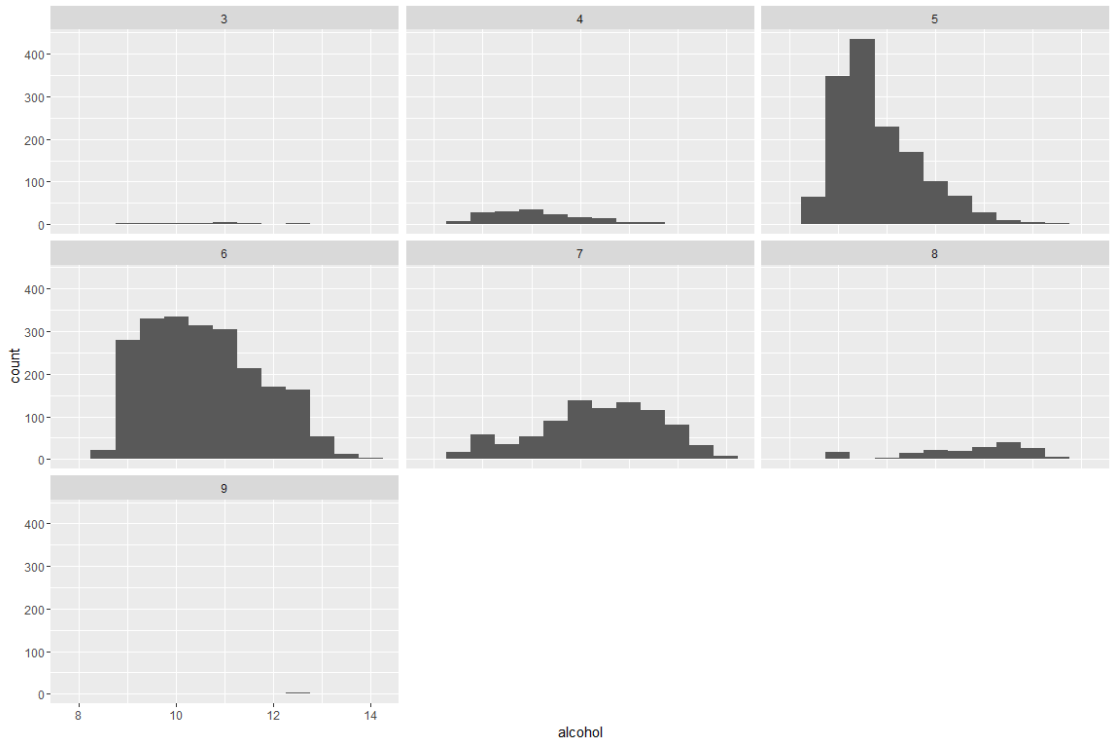


I can sort of see a dividing line between the good and bad quality wines. The bad wines are on the top half, and the better wines are on the bottom half.

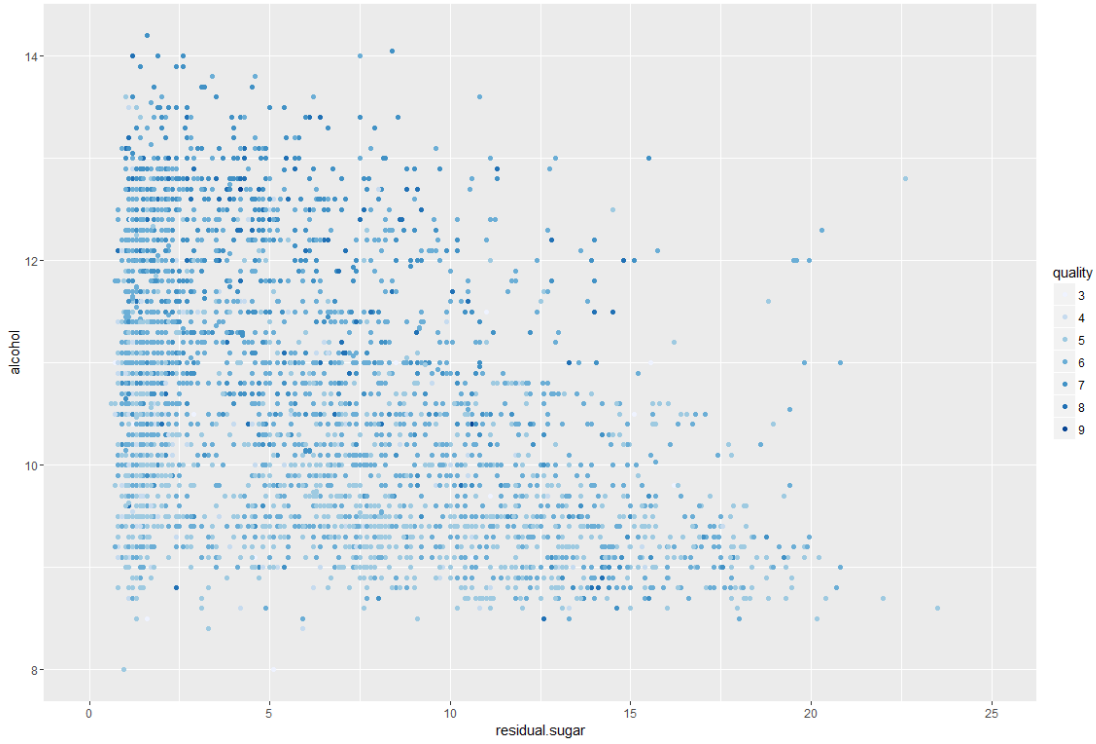


Can't figure out a quality pattern for the big mass of data points of residual sugar <5.

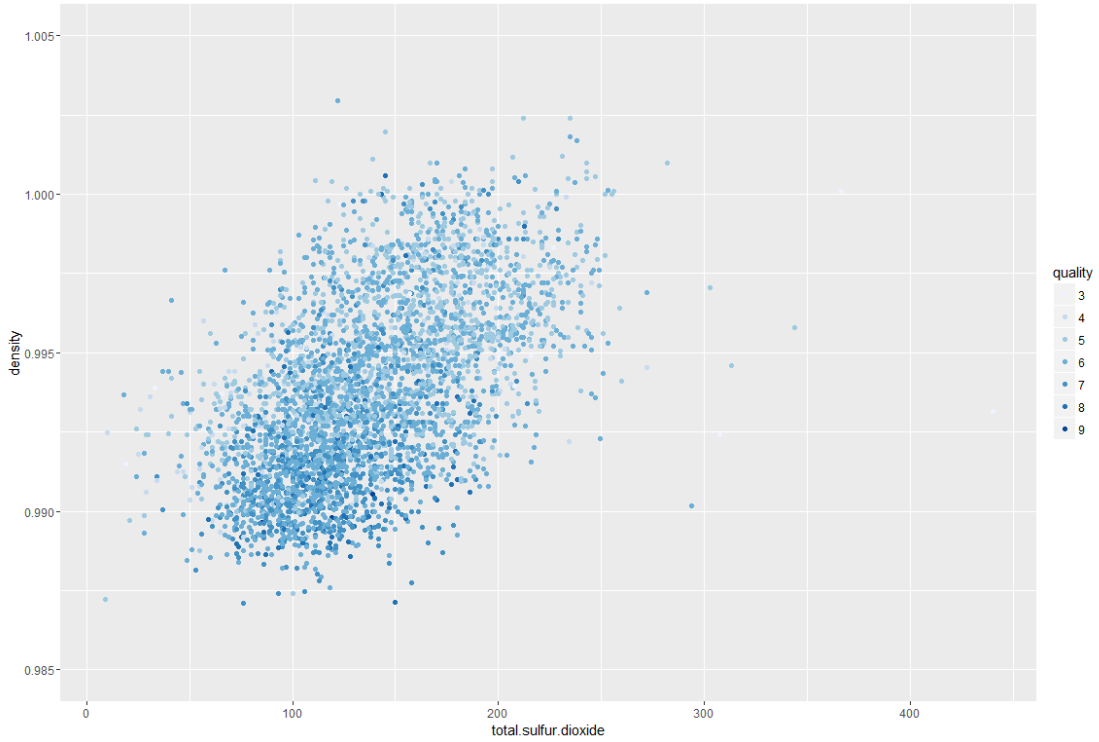
Here, we see that the higher quality wines are concentrated on the lower right tip, which corresponds to a higher alcohol content and lower density.



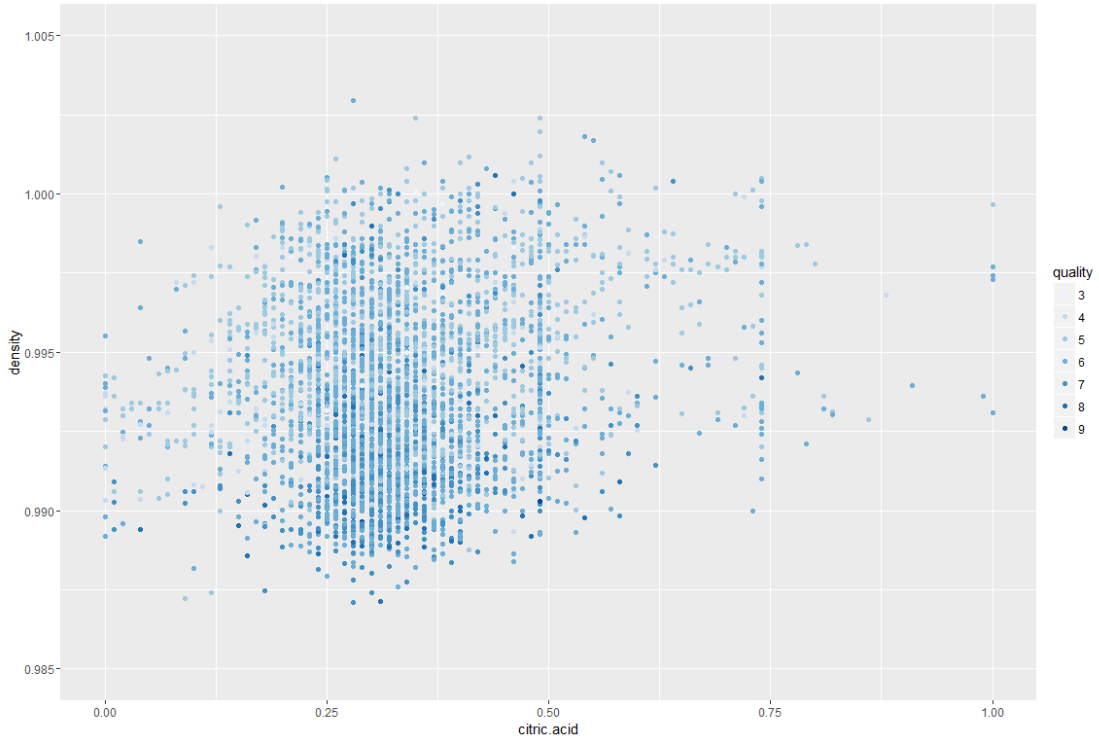
We can tell from these histograms that higher quality wines are concentrated at the higher alcohol levels, whereas the lower quality ones peak at the lower levels.



I tried to make the connection from residual sugar to alcohol - but there isn't a real pattern here.



There's not a good pattern that explains density in relation to total sulfur dioxide, even though the correlation was 0.53. I was hoping that there was a ceiling where it may have interfered with the taste, but I can't identify a max sulfur level.



After looking at some of these other variables, it is clear that the two that I concentrated on are the best examples of linear relationships in the data set.

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?

It was great to find out that the quality scores showed some segmentation in the plot between residual sugar vs density, and again in density vs alcohol. These were the two relationships that I identified in the bivariate analysis, because these showed the most linear relationship and had the highest correlation values.

In the residual sugar vs density plot, there was a dividing line that seemed to separate the good wines from the bad wines. This led me to believe that the key to a good wine was finding the correct ratio. Whether the sugar value was high or low, it had to be below the threshold.

For the density vs alcohol plot, the good wine and bad wines were at polar opposite ends of the graph. High alcohol content with low density scored well, and the opposite scored poorly. These clear distinctions really solidified the strength of these relationships.

Were there any interesting or surprising interactions between features?

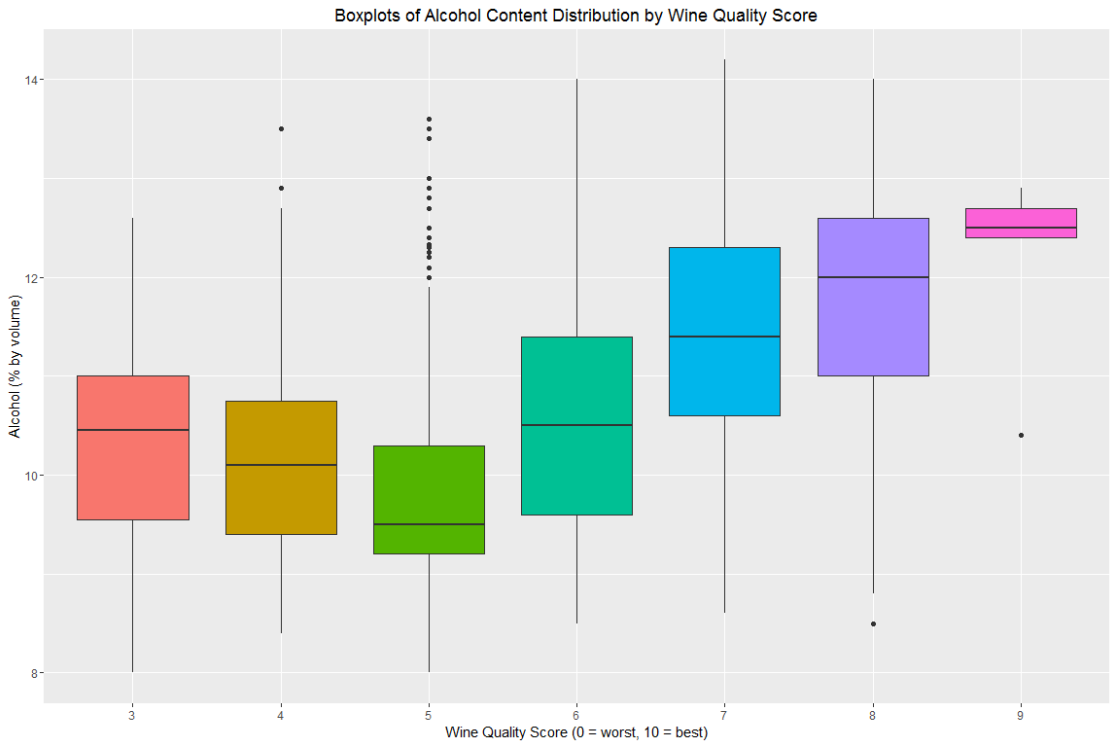
I was surprised that the acidity and sulfur variables played so little in determining the quality score. The plots showed quality scores all over the spectrum for acidic variables and sulfur variables alike.

Density being so important was surprising to me, as it seemed most irrelevant to taste. I was puzzled that density was so closely related to alcohol, and that these two variables were the clear distinguishing features for quality.

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

Final Plots and Summary

Plot One



Description One

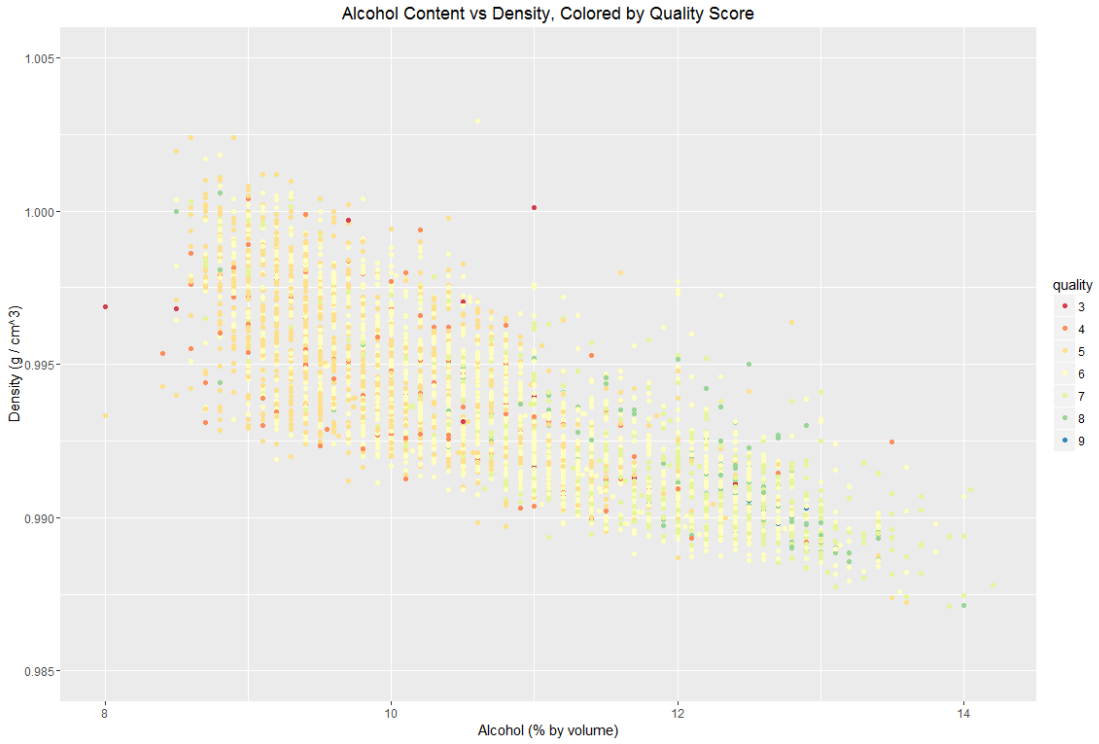
This boxplot of 'Alcohol Content Distribution by Wine Quality Score' identified alcohol % as a distinguishing factor of white wine quality. I chose this graph because it was the only bivariate boxplot that clearly distinguished the best wines from the worst wines. The mean and median of the alcohol level increases significantly as quality score increases:

```
## Source: local data frame [7 x 4]
##
##   quality alcohol_mean alcohol_median     n
##    (fctr)        (dbl)          (dbl) (int)
## 1       3     10.34500          10.45    20
## 2       4     10.15245          10.10   163
## 3       5      9.80884           9.50  1457
## 4       6     10.57537          10.50  2198
## 5       7     11.36794          11.40   880
## 6       8     11.63600          12.00   175
## 7       9     12.18000          12.50     5
```

This revelation was an important point in the investigation, because it was the key to exploring the features contributing to a good wine quality score. A higher alcohol level by itself would not give wines a good quality score, otherwise wine manufactuers would just get alcohol levels as high as possible. I surmised that a high alcohol level is a consequence

of a closely related feature that leads to a higher wine quality score, in this case, the density.

Plot Two



Description Two

The next important plot in this analysis delved into the relationship between high alcohol levels and good wine quality scores. What are the factors that are closely related to alcohol content, that may possibly explain the good wine quality scores? I identified density as a closely related feature, due to the high correlation value between the two variables.

```
##
##  Pearson's product-moment correlation
##
## data:  wine$density and wine$alcohol
## t = -87.2549, 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
```

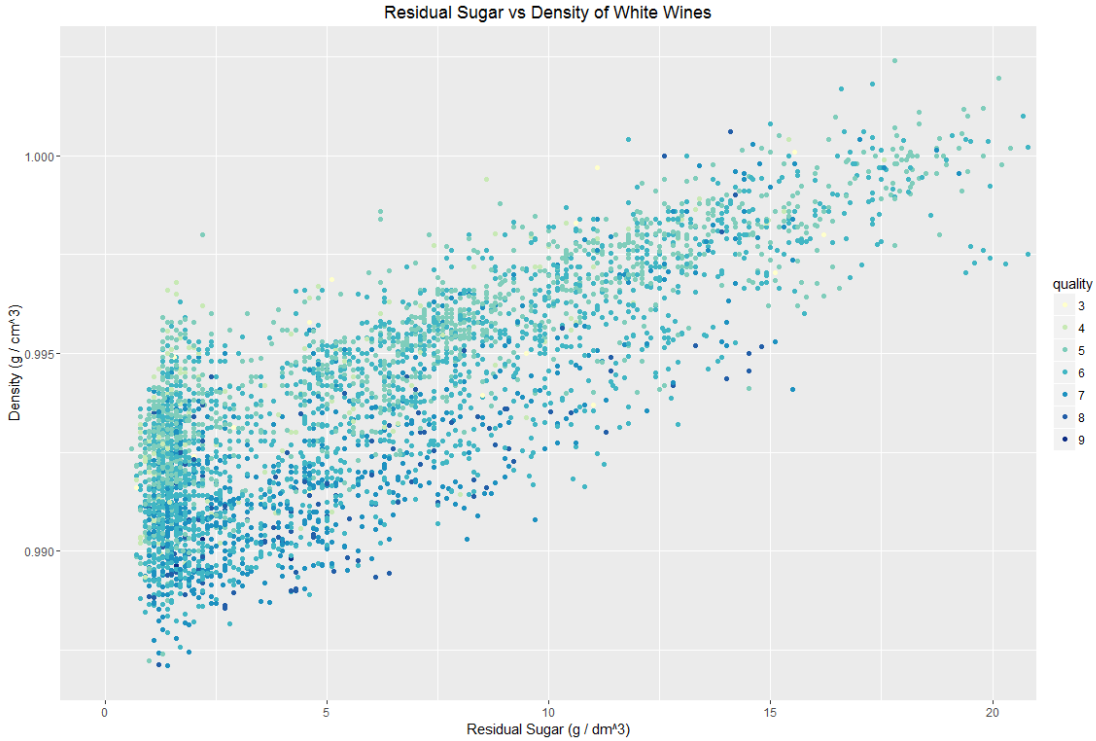
The strong negative relationship of -0.78 suggests that as alcohol increases, density decreases. This scatterplot uses red vs green to quickly illustrate that good wines and bad wines are at the opposite ends of this relationship. Good wines are at the lower right

corner, where the wine has a higher alcohol % and is less dense. Bad wines are at the top left corner, with high density and low alcohol %.

```
## Source: local data frame [7 x 4]
##
##   quality density_mean alcohol_mean     n
##    (fctr)        (dbl)        (dbl) (int)
## 1       3    0.9948840     10.34500    20
## 2       4    0.9942767     10.15245   163
## 3       5    0.9952626      9.80884  1457
## 4       6    0.9939613     10.57537  2198
## 5       7    0.9924524     11.36794   880
## 6       8    0.9922359     11.63600   175
## 7       9    0.9914600     12.18000     5
```

This table shows that as the quality increases, the average density decreases, and alcohol mean increases (in general). This was an important graph to confirm that alcohol content is the feature that distinguishes good wines from bad wines, and continues that investigation into exploring density as an explanation to why alcohol content may explain the quality score.

Plot Three



Description Three

This final plot is the culmination of my theory that the residual sugar content is the explanatory variable that contributes most towards a wine scoring high or low. My theory

is that residual sugar affects density due to the weight of the compounds, density affects alcohol content due to the fermentation of sugars that turn into alcohol, and thus this balance of alcohol to sugars results in the determining factor of whether a wine is good or bad.

In this plot, I try to highlight the density and sugar ratio "sweetspot" that separates the good wines from the bad wines. I use color to distinguish the good wines from the bad wines, because the good wine lays below the line, and bad wines is above the line.

```
## Source: local data frame [7 x 5]
##
##   quality sugar_mean   density    ratio     n
##    (fctr)      (dbl)     (dbl)    (dbl) (int)
## 1       3   6.392500 0.9948840 6.425372    20
## 2       4   4.628221 0.9942767 4.654862   163
## 3       5   7.334969 0.9952626 7.369883  1457
## 4       6   6.441606 0.9939613 6.480741  2198
## 5       7   5.186477 0.9924524 5.225920   880
## 6       8   5.671429 0.9922359 5.715806   175
## 7       9   4.120000 0.9914600 4.155488     5
```

As you can see from the table above, the ratio of mean residual sugar divided by the density, the lower quality wines have a higher ratio of density to sugars than a good scoring wine.

Reflection

Before embarking on EDA of white wines, I spent a long time trying to find my own data set to analyze. I picked a subject that I was interested in (English Premier League soccer data by season). That was the biggest struggle, and one of the reasons why I decided to abandon it in favor of the tidy white wine dataset. The main difficulties were not in the exploratory data analysis, but rather getting the data ready, and choosing how to display that data. One of the main issues I came across in that particular data set was how to deal with paired data - I had a set of data for the home team, and a matching data set for the away team. Any variable that compared the two teams resulted in a mirror image plot. In addition, I became discouraged when there were no interesting relationships between the variables.

I noticed that trying to find free data was hard to come by. In soccer, there is a lot of money in data collection. Everything is quantified and recorded, yet it costs money to have access to that data. One such company is OPTA - they have extensive in-game data that would be very fun to analyze, but the cost is very high.

In regards to EDA of this data set, I had some difficulties because I tried too hard to fit the data to my initial guesses of the features that I thought would be factors to the quality scores. I guessed that features that were relevant to tastes- like saltiness and tanginess - would be more significant to the overall quality score. Sometimes, EDA could be mundane,

because I am simply plugging different variables into the same types of graphs to see which ones stand out. It was also difficult to segregate my analysis into univariate, bivarate, and multi-variable analysis - I just wanted to go straight into multi-variable analysis.

I found success in using the ggpairs function to quickly identify pairs of data that may be of interest. Seeing all the possible combinations in one plot really helped zero-in on the features that were relevant. I may have even neglected to plot density to alcohol had I not seen the correlation numbers of that relationship.

I think this analysis would have been more interesting if the feature of interest was a continous quantitative variable rather than categorical. The diamonds example was interesting because the main feature of interest was price. Wine price would have been an interesting feature had it been available. In addition, aggregating the red and white wines would have added another interesting layer into this investigation. Does quality scores differ based on different parameters? Does the acidity and sulfur variables come into play when comparing these two different types of wine? What are some of the main distinctions between the two? I think this would have been a more interesting question than determining what influences the quality of white wines.