Exploration of White Wine Quality by Ady Oren

This project explores which chemical properties influence the quality of white wines using approximately 4900 data points. This data set is related to white variants of the Portuguese “Vinho Verde” wine. For more details, consult: http://www.vinhoverde.pt/en/ (http://www.vinhoverde.pt/en/) or the reference [Cortez et al., 2009].

Throughout this project, we will be making comparisons between the various collected variables and between quality and those variables, in an effort to identify which components have a strong relationship with the quality rating assigned.

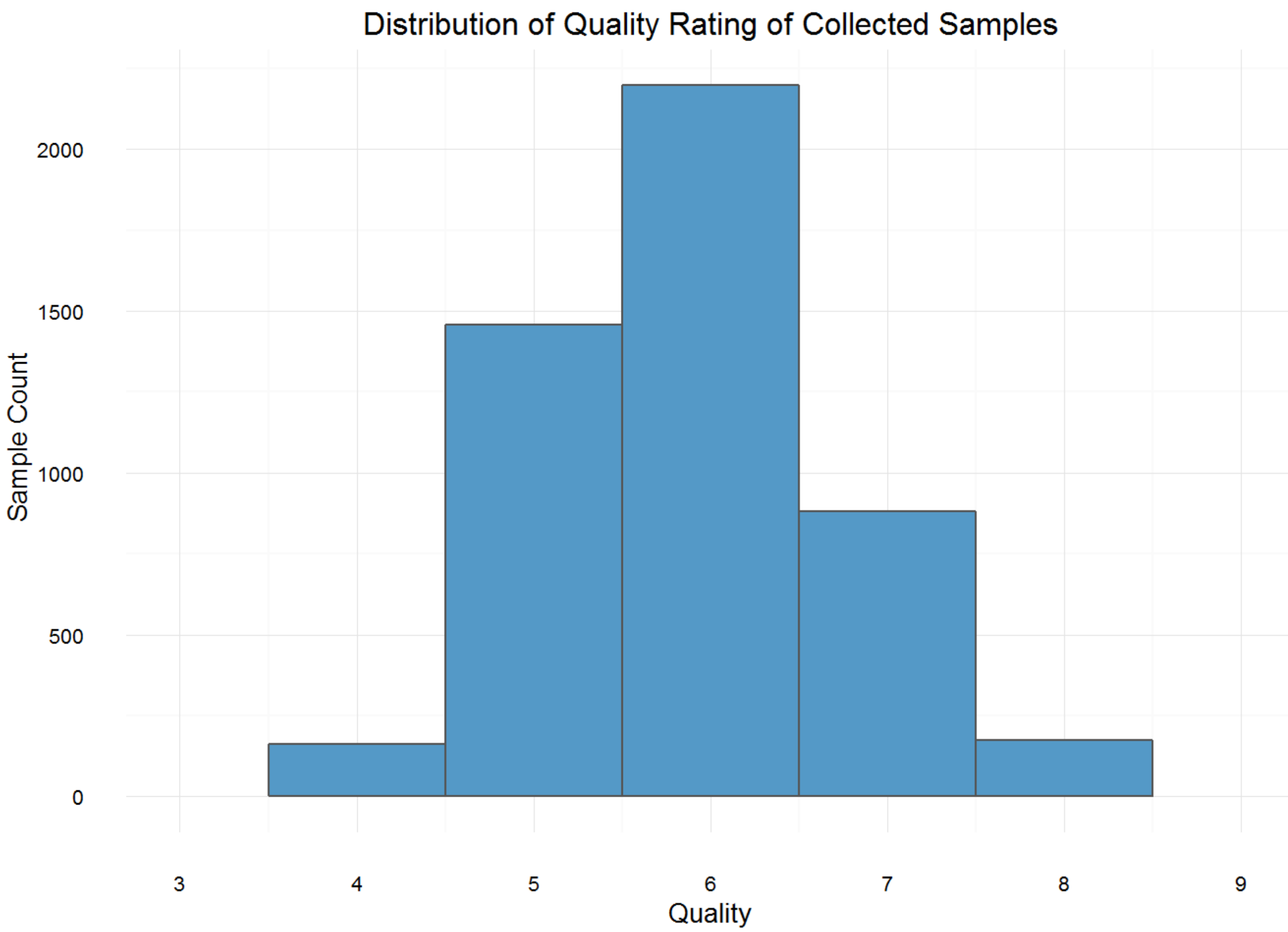
Univariate Plots and Analysis

This data set contains 4,898 white wines with 11 variables on quantifying the chemical properties of each wine. At least 3 wine experts rated the quality of each wine, providing a rating between 0 (very bad) and 10 (very excellent).

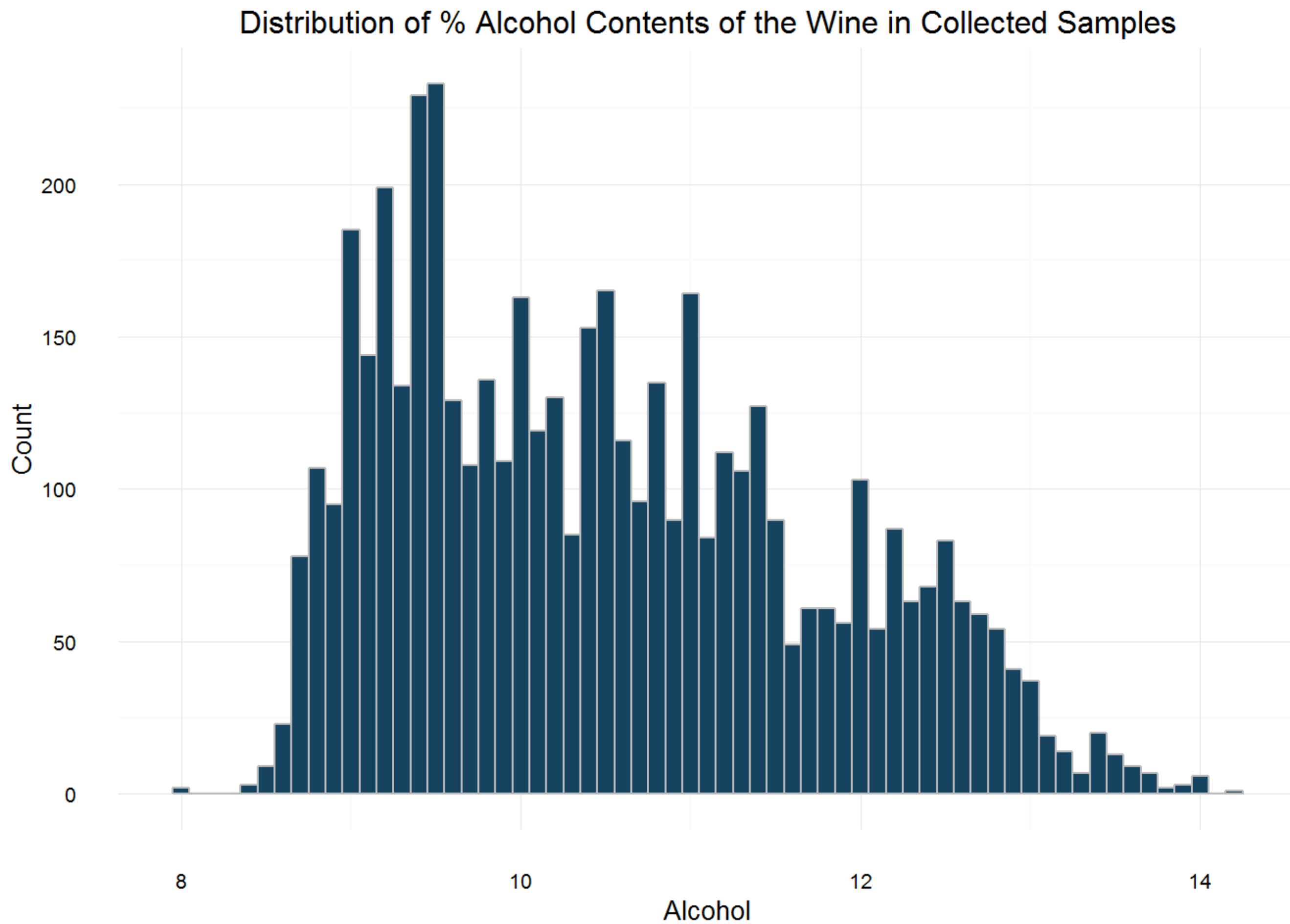
```
## [1] 4898   13
```

```
## '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 ...
```

Data visualization of counts of the various variables.




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

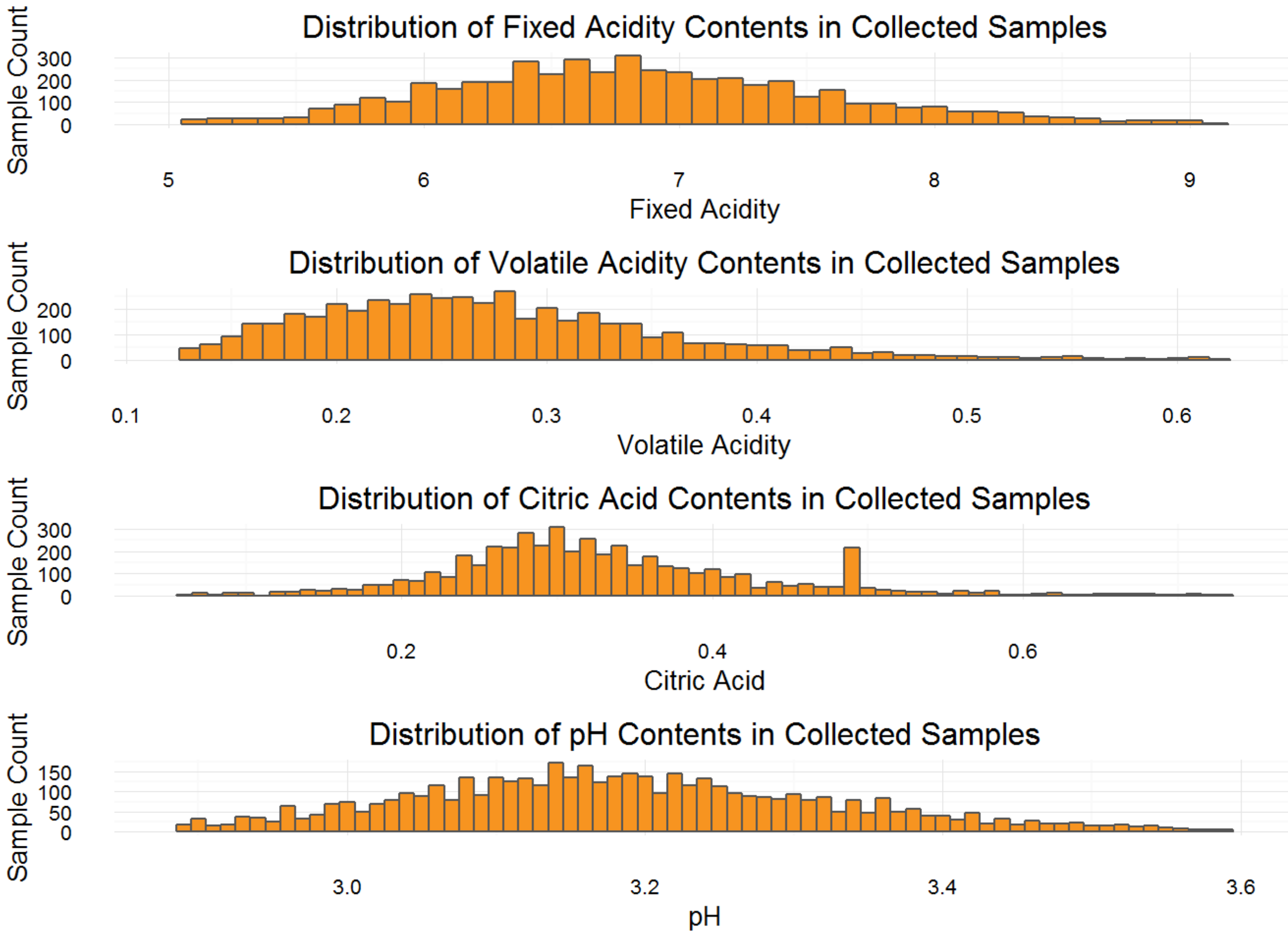



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00    9.50   10.40   10.51   11.40   14.20
```

Lets investigate acidity contents:

Please note that the bottom and top 1% data points for each of the plots below have been removed to get a more normalized distribution and exclude outliers.

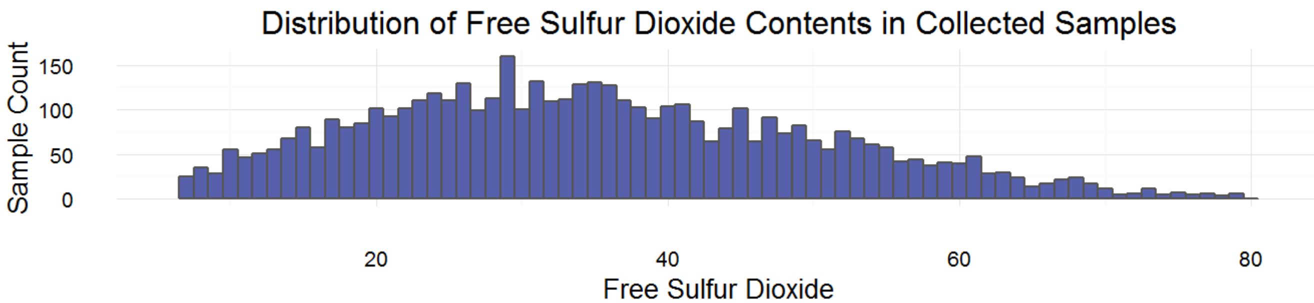
- **Fixed Acidity:** most acids involved with wine or fixed or nonvolatile (do not evaporate readily)
- **Volatile Acidity:** the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste
- **Citric Acid:** found in small quantities, citric acid can add ‘freshness’ and flavor to wines
- **pH:** describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale

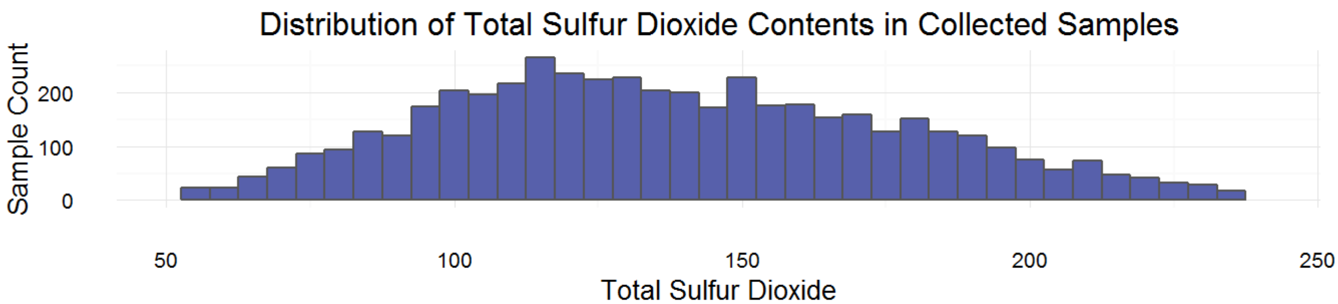


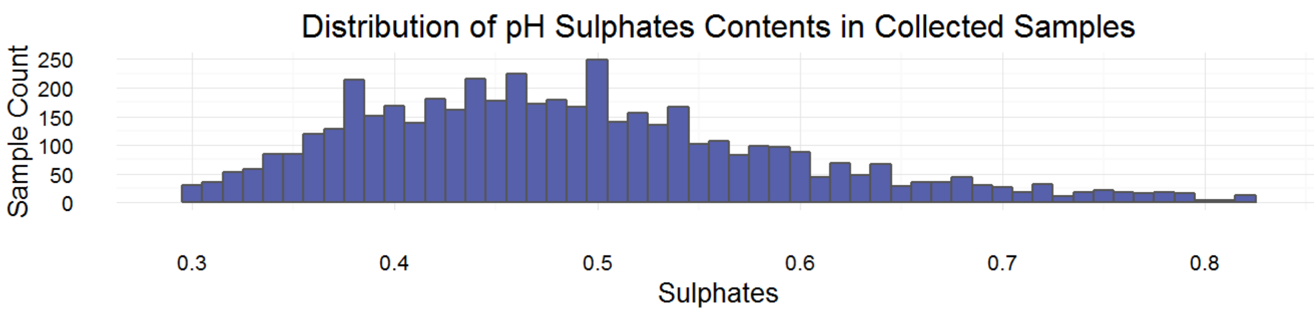

```
##  fixed.acidity    volatile.acidity  citric.acid          pH
##  Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   :2.720
##  1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700   1st Qu.:3.090
##  Median : 6.800   Median :0.2600   Median :0.3200   Median :3.180
##  Mean   : 6.855   Mean   :0.2782   Mean   :0.3342   Mean   :3.188
##  3rd Qu.: 7.300   3rd Qu.:0.3200   3rd Qu.:0.3900   3rd Qu.:3.280
##  Max.   :14.200   Max.   :1.1000   Max.   :1.6600   Max.   :3.820
```

What about sulfur contents?

- **Free Sulfur Dioxide:** the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine
- **Total Sulfur Dioxide:** amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine
- **Sulphates:** a wine additive which can contribute to sulfur dioxide gas (S02) levels, which acts as an antimicrobial and antioxidant



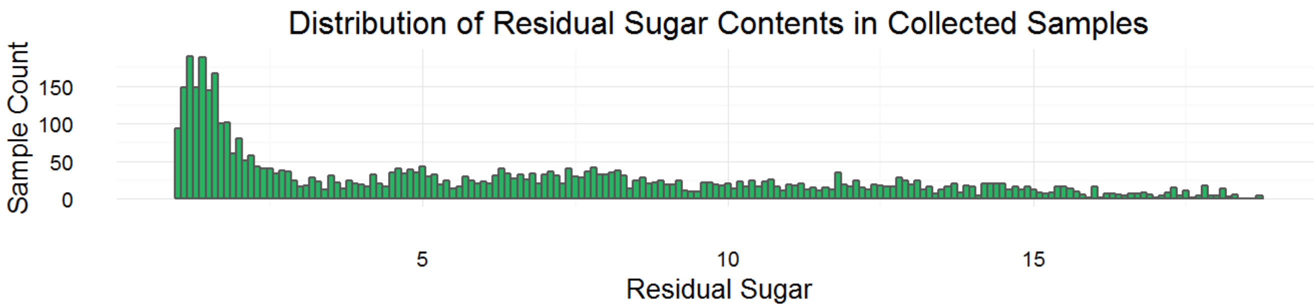


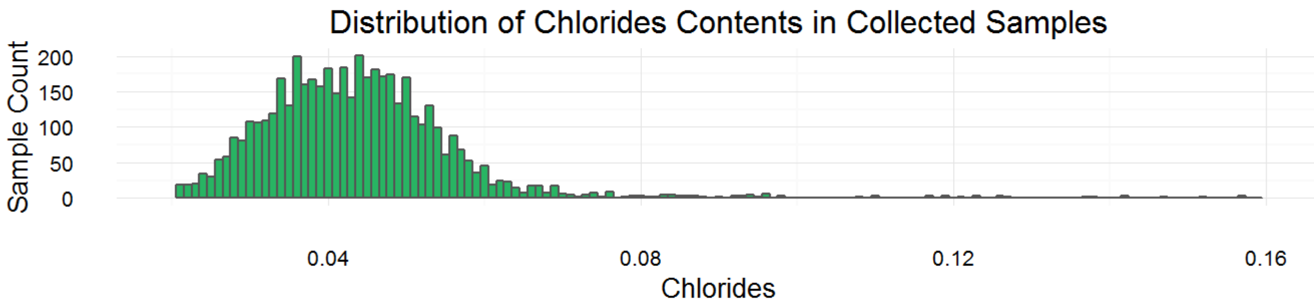


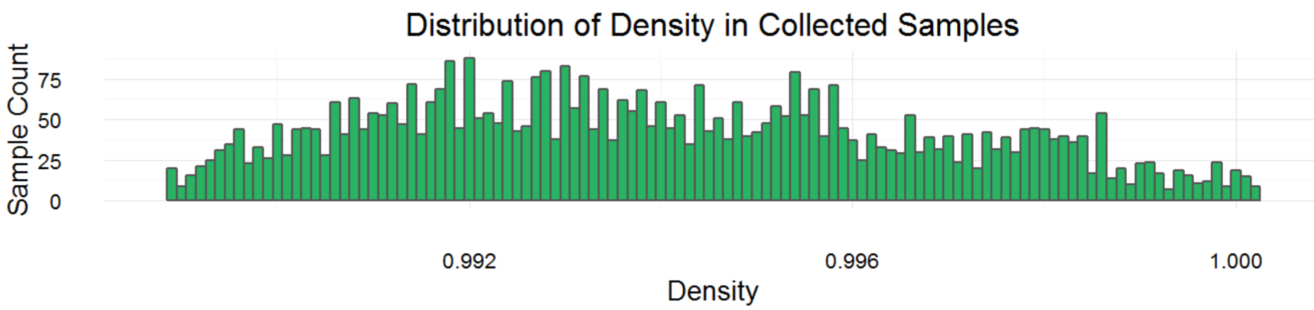

```
##  free.sulfur.dioxide total.sulfur.dioxide   sulphates
##  Min.   :  2.00      Min.   :  9.0        Min.   :0.2200
##  1st Qu.: 23.00      1st Qu.:108.0        1st Qu.:0.4100
##  Median : 34.00      Median :134.0        Median :0.4700
##  Mean   : 35.31      Mean   :138.4        Mean   :0.4898
##  3rd Qu.: 46.00      3rd Qu.:167.0        3rd Qu.:0.5500
##  Max.   :289.00      Max.   :440.0        Max.   :1.0800
```

Additional Variables:

- **Residual Sugar:** the amount of sugar remaining after fermentation stops, it’s rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet
- **Chlorides:** the amount of salt in the wine
- **Density:** the density of water is close to that of water depending on the percent alcohol and sugar content

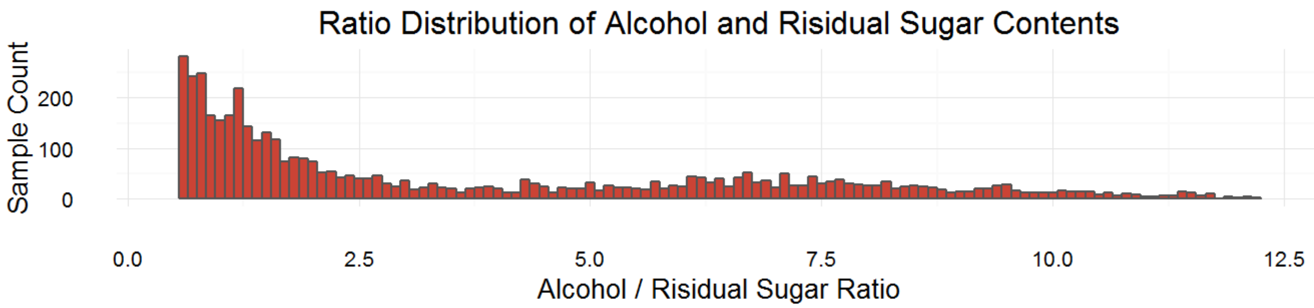


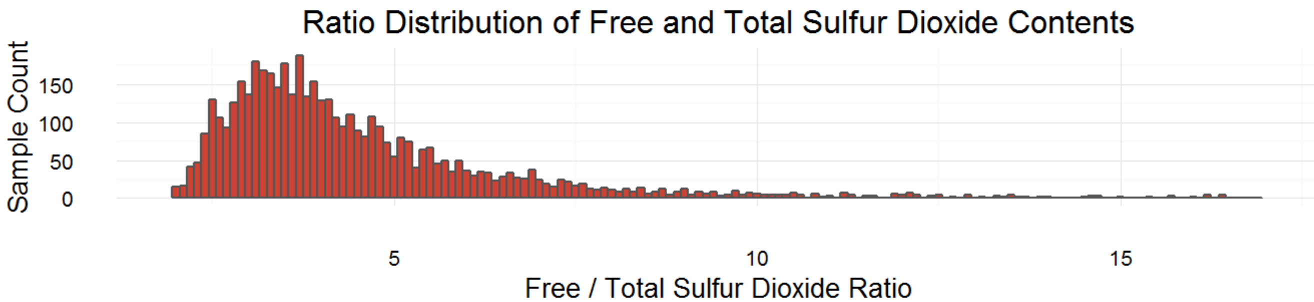


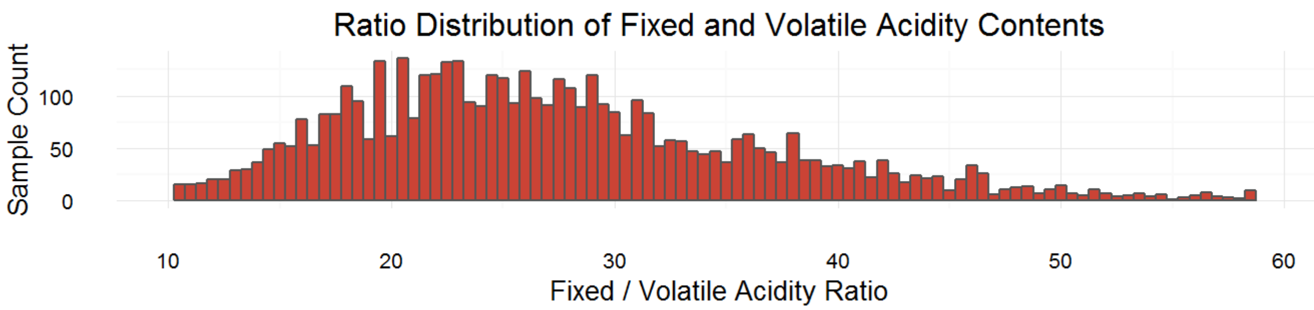



```
##  residual.sugar     chlorides           density
##  Min.   : 0.600   Min.   :0.00900   Min.   :0.9871
##  1st Qu.: 1.700   1st Qu.:0.03600   1st Qu.:0.9917
##  Median : 5.200   Median :0.04300   Median :0.9937
##  Mean   : 6.391   Mean   :0.04577   Mean   :0.9940
##  3rd Qu.: 9.900   3rd Qu.:0.05000   3rd Qu.:0.9961
##  Max.   :65.800   Max.   :0.34600   Max.   :1.0390
```

Next, lets look at the ratios between some of the variables








```
##  sugar_alcohol.ratio free_total_sulfor.ratio fixed_volatile_acidity.ratio
##  Min.   : 0.1778     Min.   : 1.407          Min.   : 5.545
##  1st Qu.: 1.0233     1st Qu.: 3.167          1st Qu.:20.627
##  Median : 2.0385     Median : 3.942          Median :26.071
##  Mean   : 3.6913     Mean   : 4.720          Mean   :27.657
##  3rd Qu.: 6.3502     3rd Qu.: 5.237          3rd Qu.:33.000
##  Max.   :17.6667     Max.   :42.333          Max.   :90.000
```

What is the structure of your dataset?

This is a tidy data set that follows a wide format. Each sample is represented by a single line that contains 11 variables containing values of fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, and quality.

What are the main features of interest in your dataset?

Per the data set instructions, I am attempting to identify which chemical properties influence the quality of white wines.

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

It would have been interesting to see the actual grading of each individual sample by each taster instead of an average quality score. Visibility to that data may have provided some insight into individual preferences among the tasters. Also, per the information provided with this data set, there is no data about grape types, wine brand or wine selling price due to privacy and logistical issues.

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

I was interested to see the distribution of the ratio of residual sugar to alcohol contents and created a new variable for that value. I also created new variables for the rations between Free and Total Sulfur Dioxide, and Fixed Acidity and Volatile Acidity.

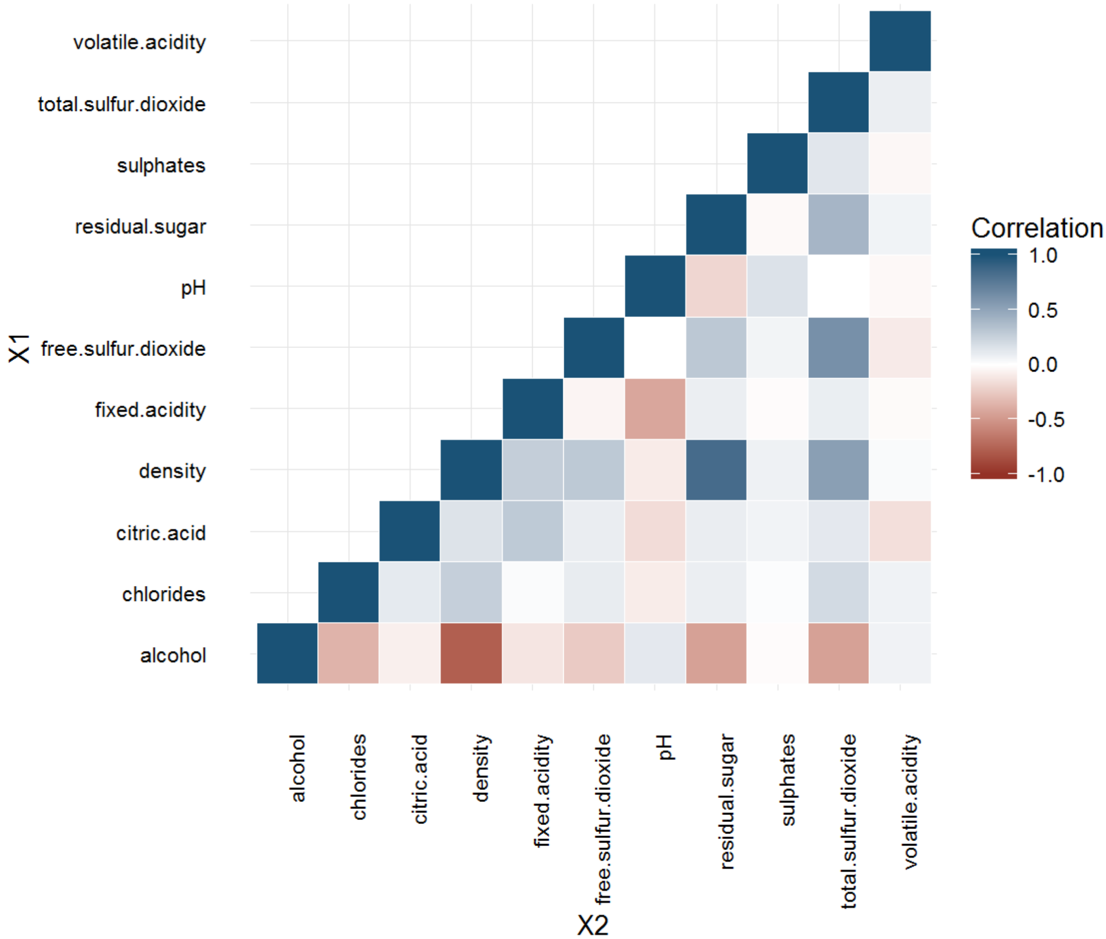
Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

I'm not a chemist, or a wine drinker for that matter, so I didn't have any expectations as to what the distributions would look like prior to creating the plots. That being said, I wasn't surprised to find that the majority of the values created a fairly consistent bell curve once the outliers were removed.

Bivariate Plots Section and Analysis

First, lets take a look at the corrolation between the variables.

```
##             X1          X2  value
## 1      alcohol     alcohol  1.000
## 12     alcohol   chlorides -0.360
## 13   chlorides   chlorides  1.000
## 23     alcohol citric.acid -0.076
## 24   chlorides citric.acid  0.114
## 25 citric.acid citric.acid  1.000
```



Three variables that stand out is the positive correlation between density and suger (r = 0.839), and the negative correlation between alcohol and density (r = -0.780). As a reminder, these three variables are described as:

- **Alcohol:** The percent alcohol content of the wine
- **Residual Sugar:** The amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet
- **Density:** The density of water is close to that of water depending on the percent alcohol and sugar content

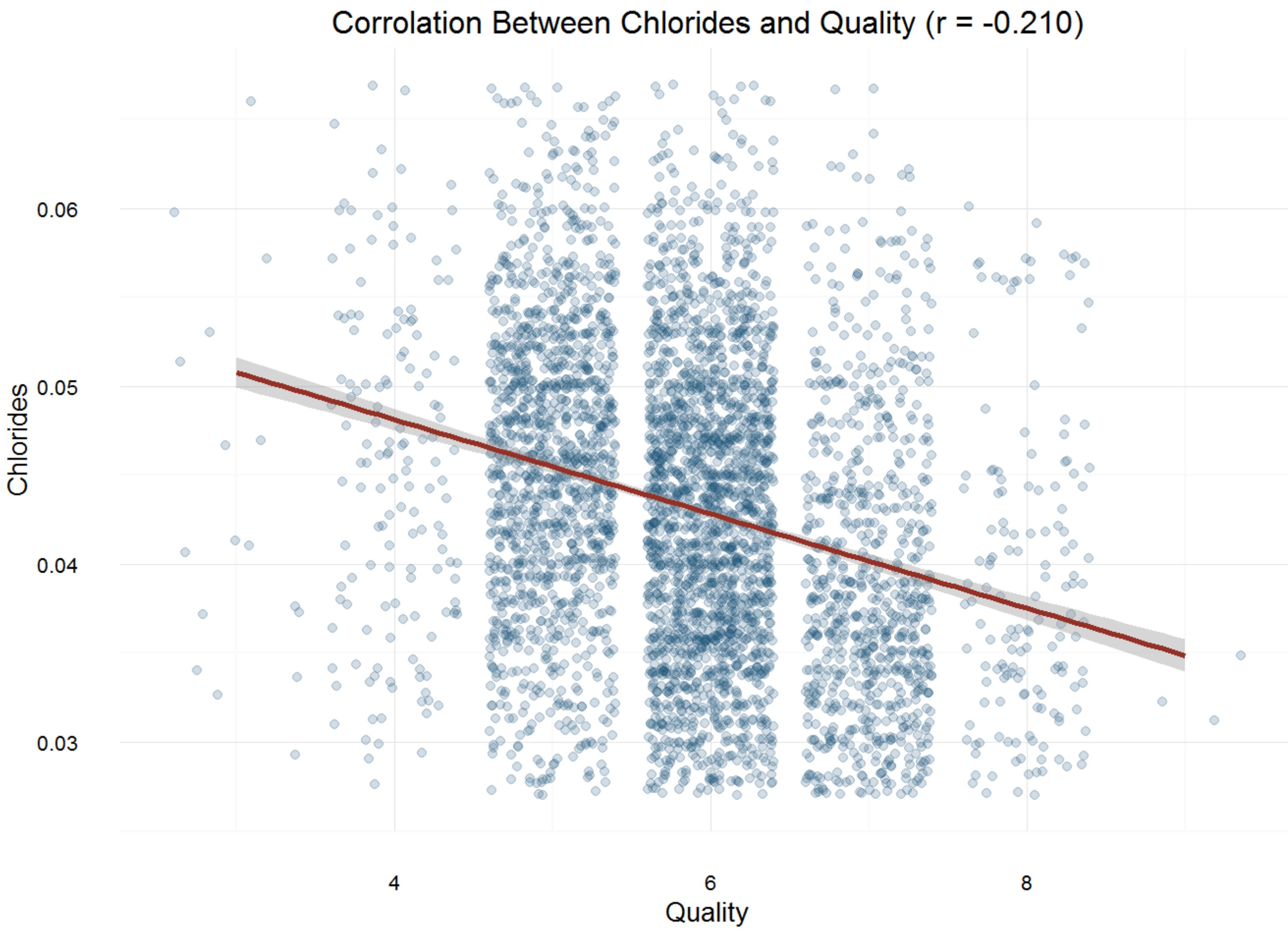
Since these three variables related to each other as part of the wine-making process, it makes sense that we have high correlations between them.

Investigating correlation between variables and the rated quality

I ran a correlation test between each one of the variables and found that the Alcohol, Chlorides and Density variables had the strongest correlations, either positive or negative, to the quality rating the sample received from the tasters. Having said that, the correlation of all three variables to the quality rating is weak to moderate at best. Nevertheless, we will concentrate on these two variables since the other variable had very weak correlation coefficients.

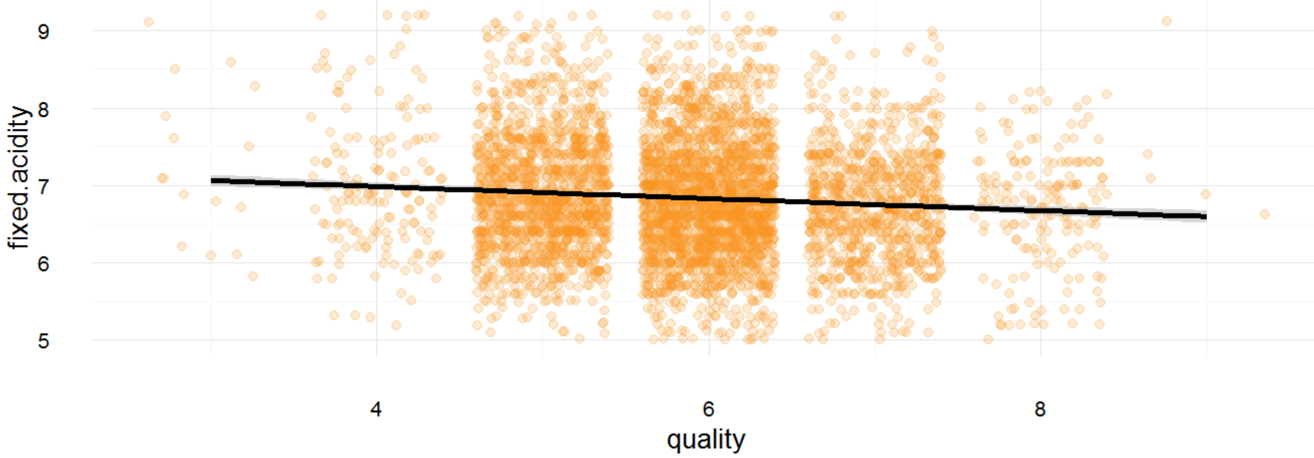

```
##
##  Pearson's product-moment correlation
##
## data:  wqw$quality and wqw$alcohol
## t = 33.858, df = 4896, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4126015 0.4579941
## sample estimates:
##       cor
## 0.4355747
```

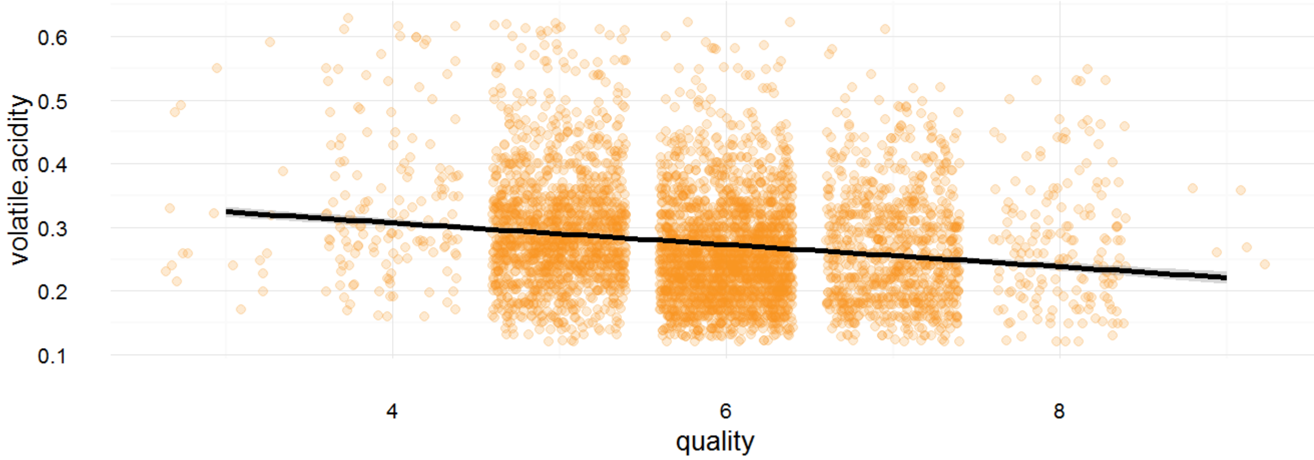

```
##
##  Pearson's product-moment correlation
##
## data:  wqw$quality and wqw$density
## t = -22.581, df = 4896, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3322718 -0.2815385
## sample estimates:
##        cor
## -0.3071233
```

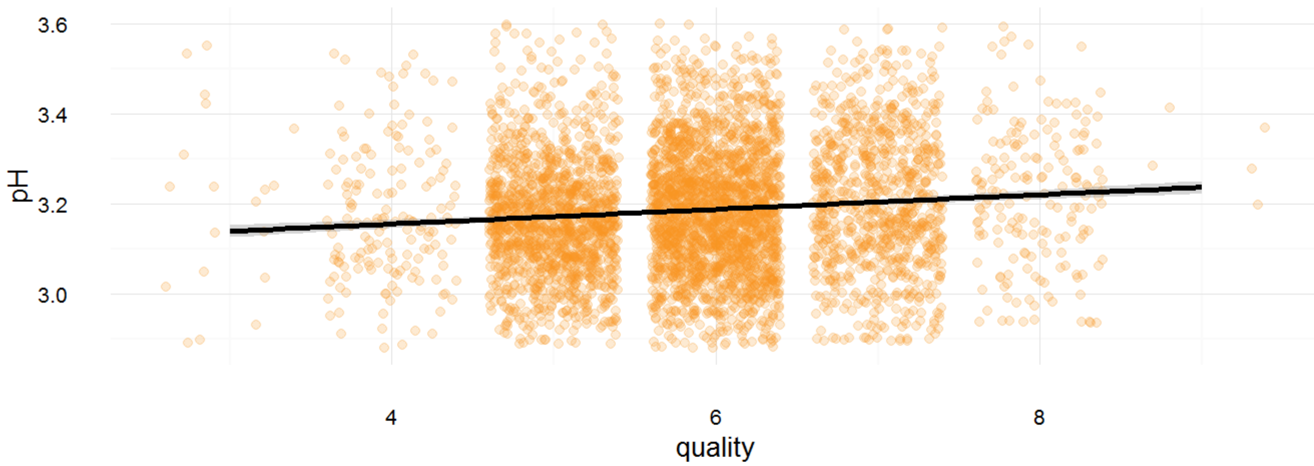


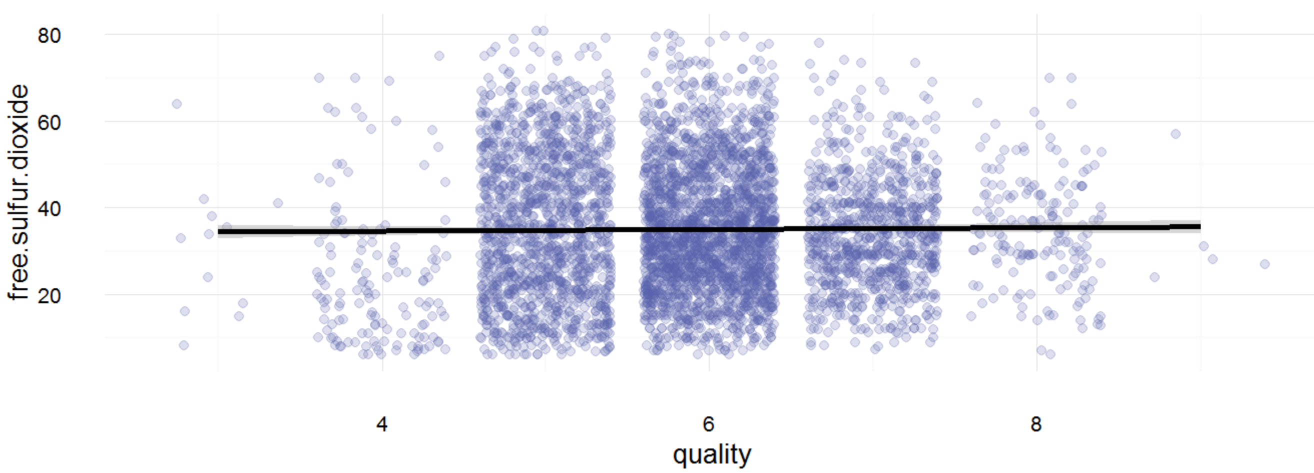

```
##
##  Pearson's product-moment correlation
##
## data:  wqw$quality and wqw$chlorides
## t = -15.024, df = 4896, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2365501 -0.1830039
## sample estimates:
##        cor
## -0.2099344
```

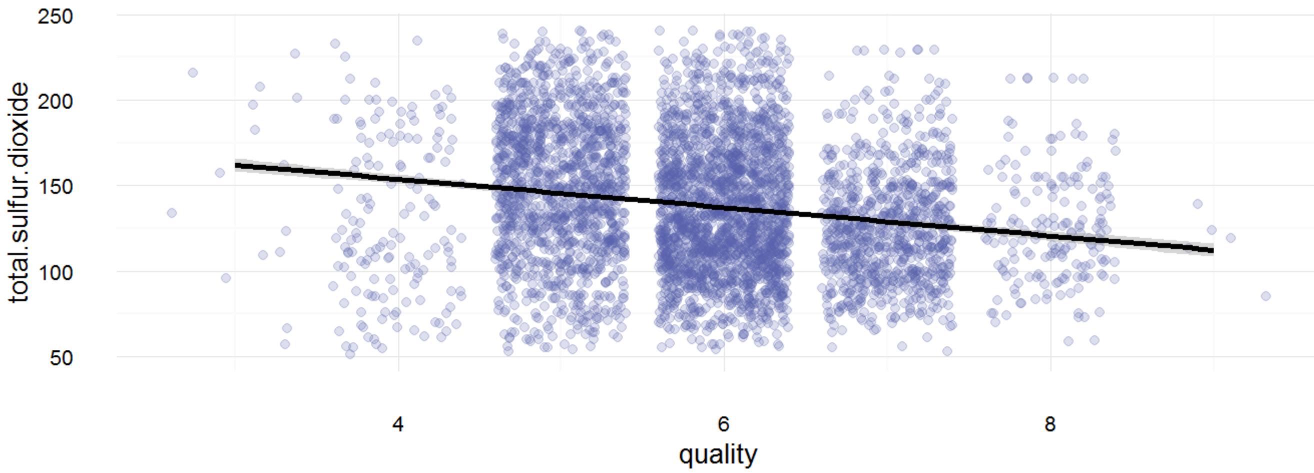
In contrast, the remaining variables showed very week to non-existent correlation coefficients.


fixed.acidity
9
8
7
6
5
4
6
8
quality

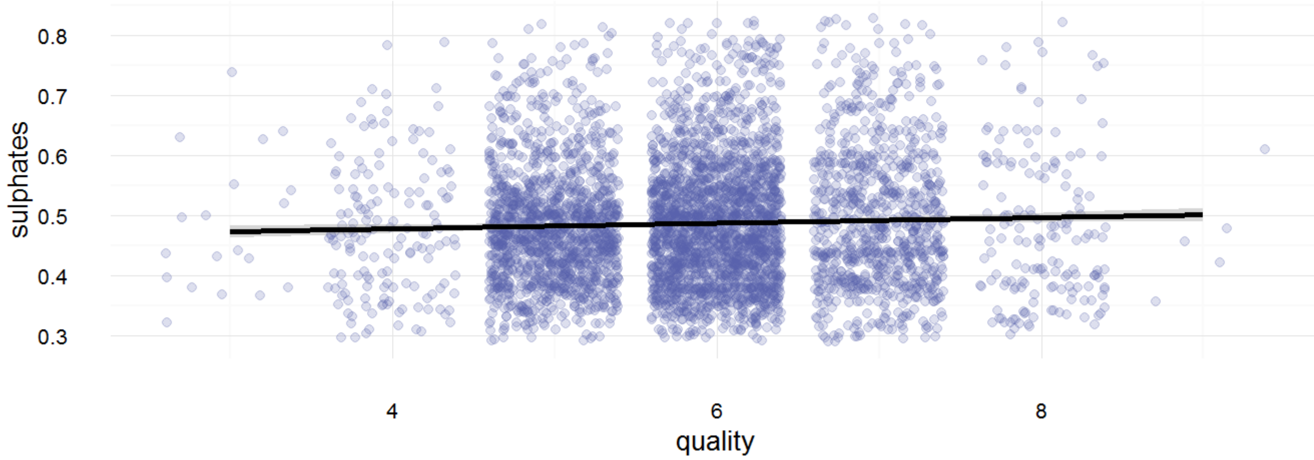

volatile.acidity
0.6
0.5
0.4
0.3
0.2
0.1
4
6
8
quality


pH
3.6
3.4
3.2
3.0
4
6
8
quality

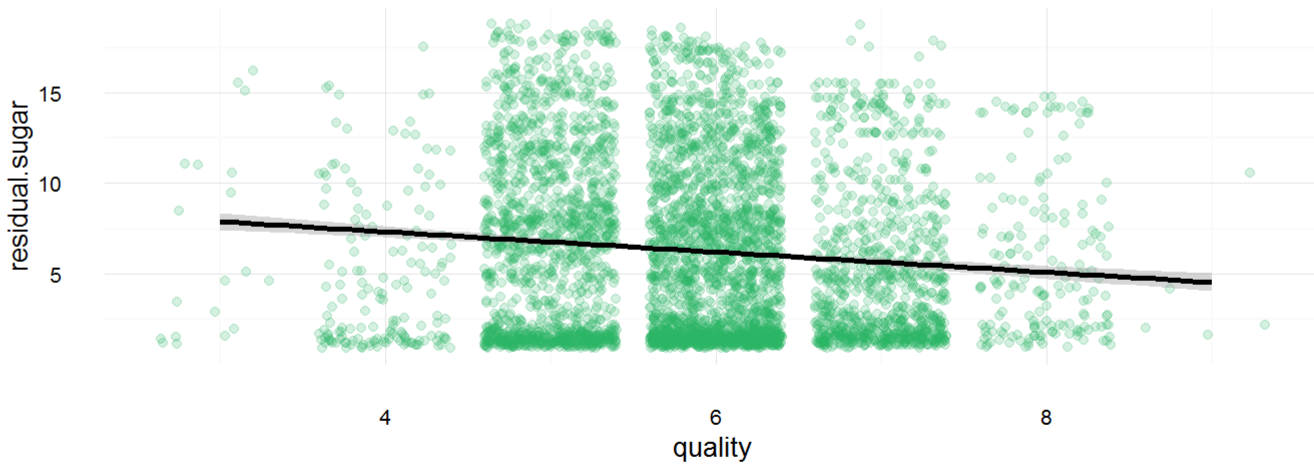

free.sulfur.dioxide
80
60
40
20
4
6
8
quality



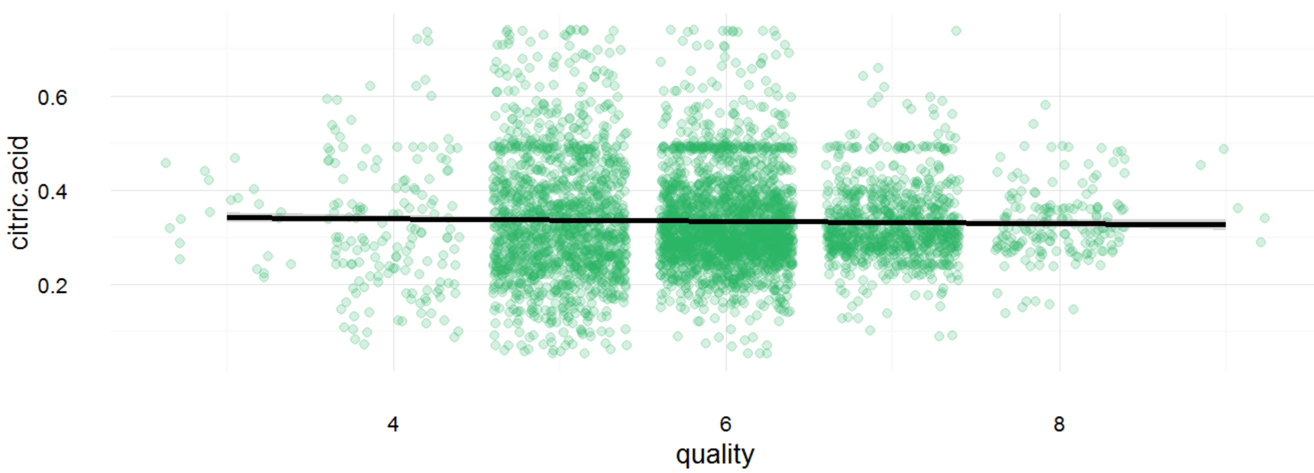
250
200
150
100
50
total.sulfur.dioxide
4
6
8
quality



0.8
0.7
0.6
0.5
0.4
0.3
sulphates
4
6
8
quality



15
10
5
residual.sugar
4
6
8
quality



0.6
0.4
0.2
citric.acid
4
6
8
quality

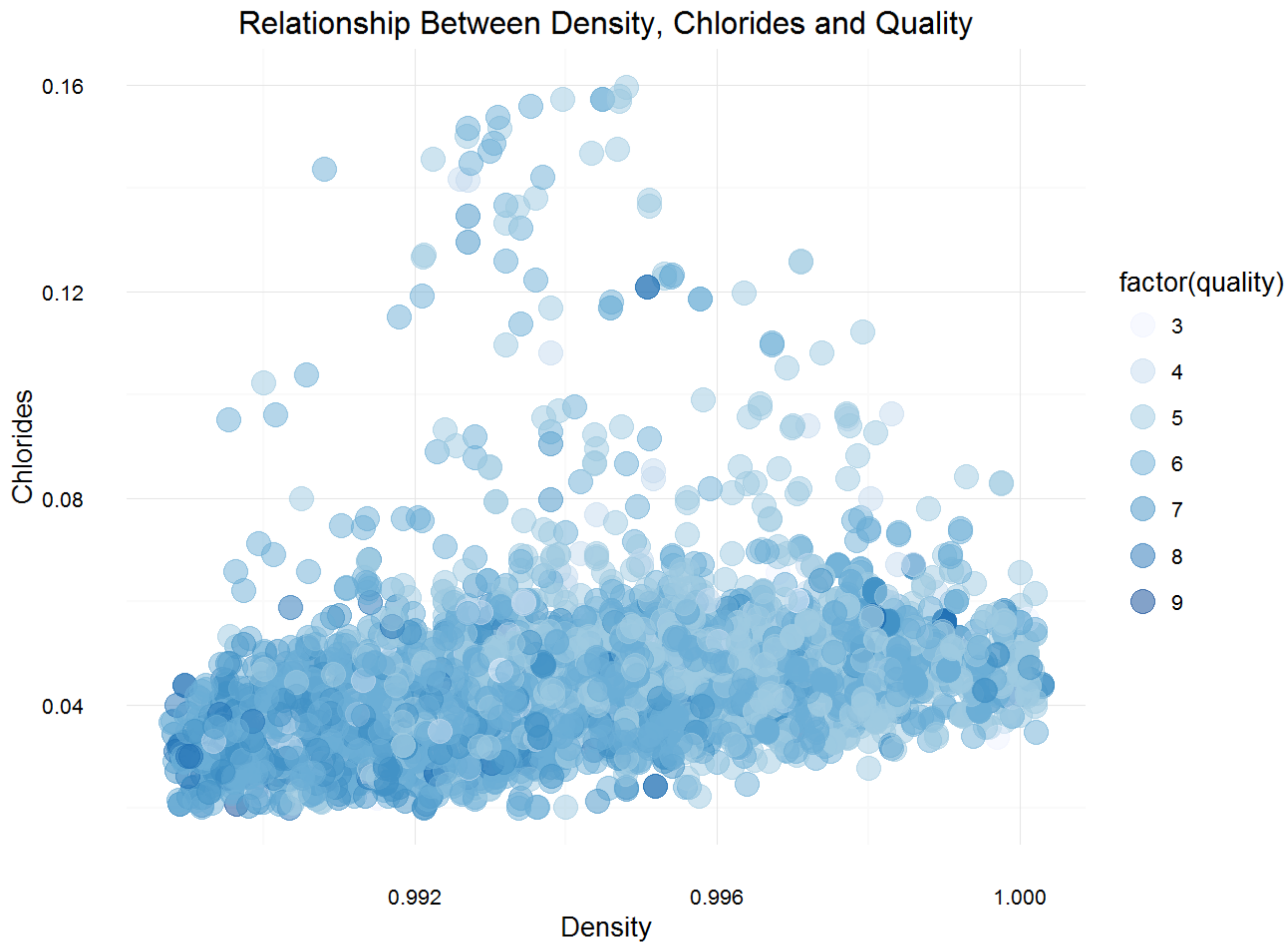
Multivariate Plots and Analysis

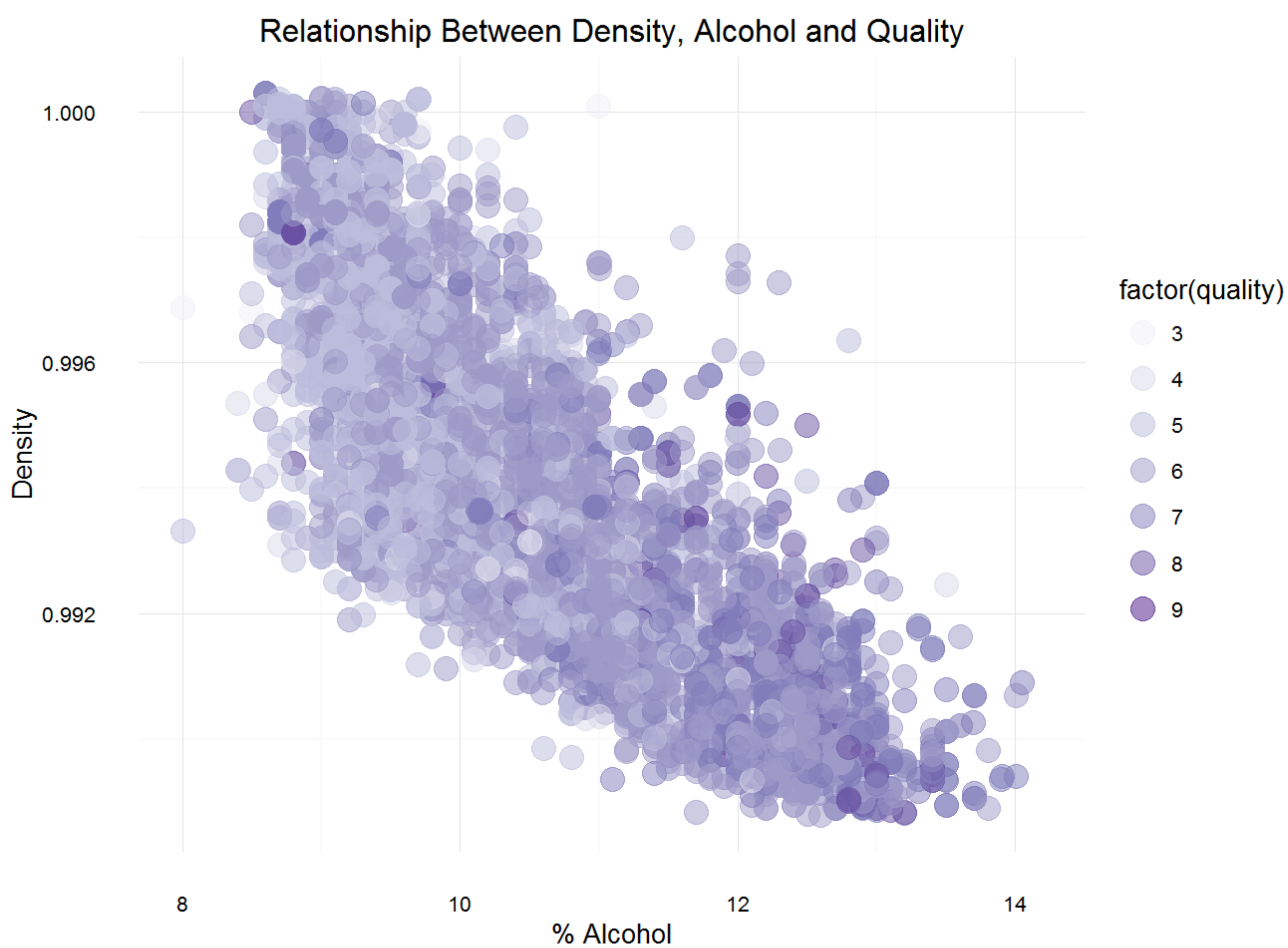
Lets start by looking at the summary data

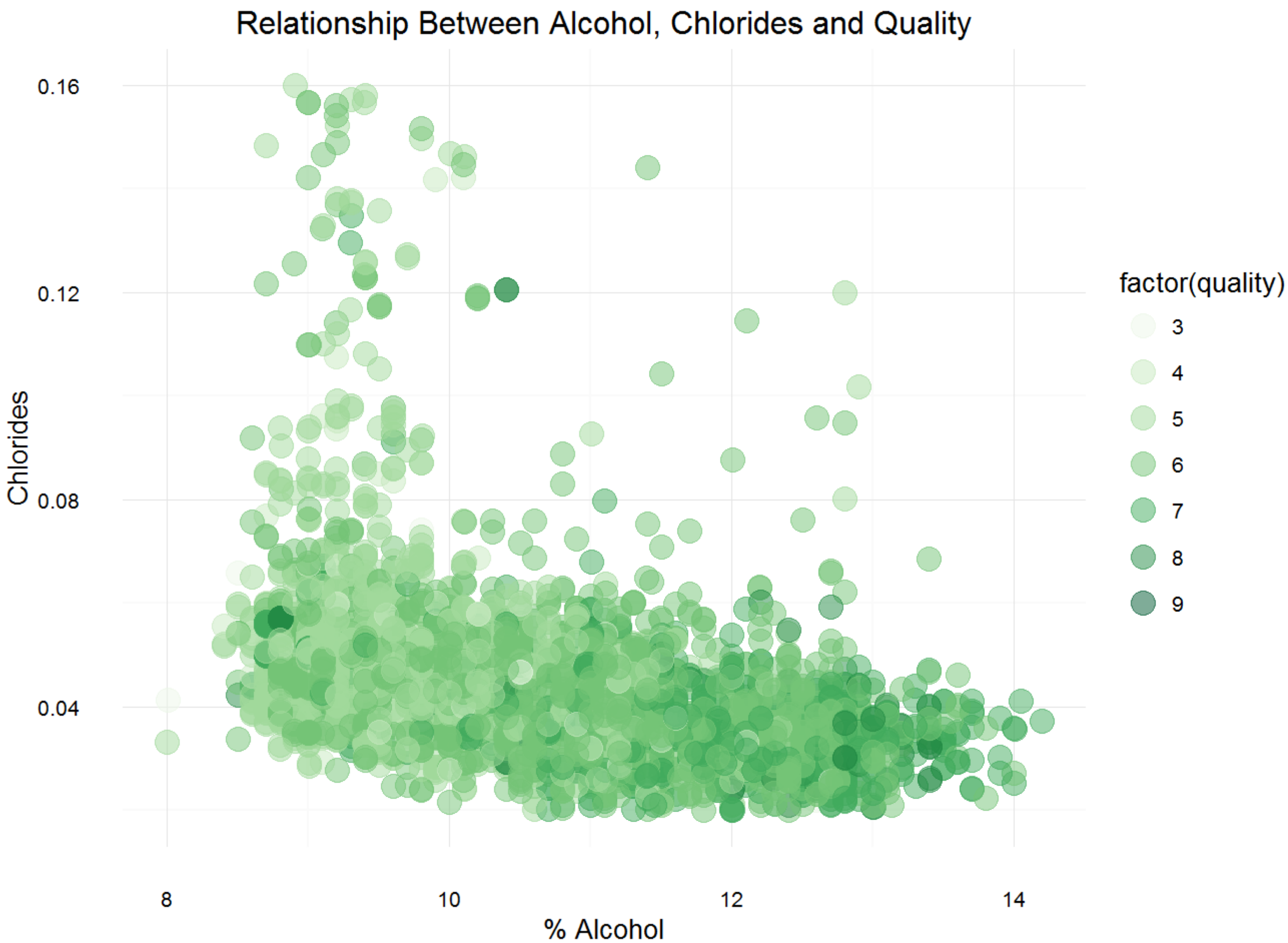
The table below shows the mean of a three main variables grouped by the quality rating (Chlorides, Alcohol and Density)

```
## # A tibble: 7 × 5
##   quality mean_chlorides mean_density mean_alcohol     n
##     <int>          <dbl>        <dbl>        <dbl> <int>
## 1       3     0.05430000    0.9948840     10.34500    20
## 2       4     0.05009816    0.9942767     10.15245   163
## 3       5     0.05154633    0.9952626      9.80884  1457
## 4       6     0.04521747    0.9939613     10.57537  2198
## 5       7     0.03819091    0.9924524     11.36794   880
## 6       8     0.03831429    0.9922359     11.63600   175
## 7       9     0.02740000    0.9914600     12.18000     5
```

Next, lets look at the relationship between the three variables with the strongest relationship (Alcohol, Chlorides and Density) and quality


Relationship Between Density, Chlorides and Quality
0.16
0.12
0.08
0.04
Chlorides
0.992
0.996
1.000
Density
factor(quality)
3
4
5
6
7
8
9


Relationship Between Density, Alcohol and Quality
1.000
0.996
0.992
Density
8
10
12
14
% Alcohol
factor(quality)
3
4
5
6
7
8
9



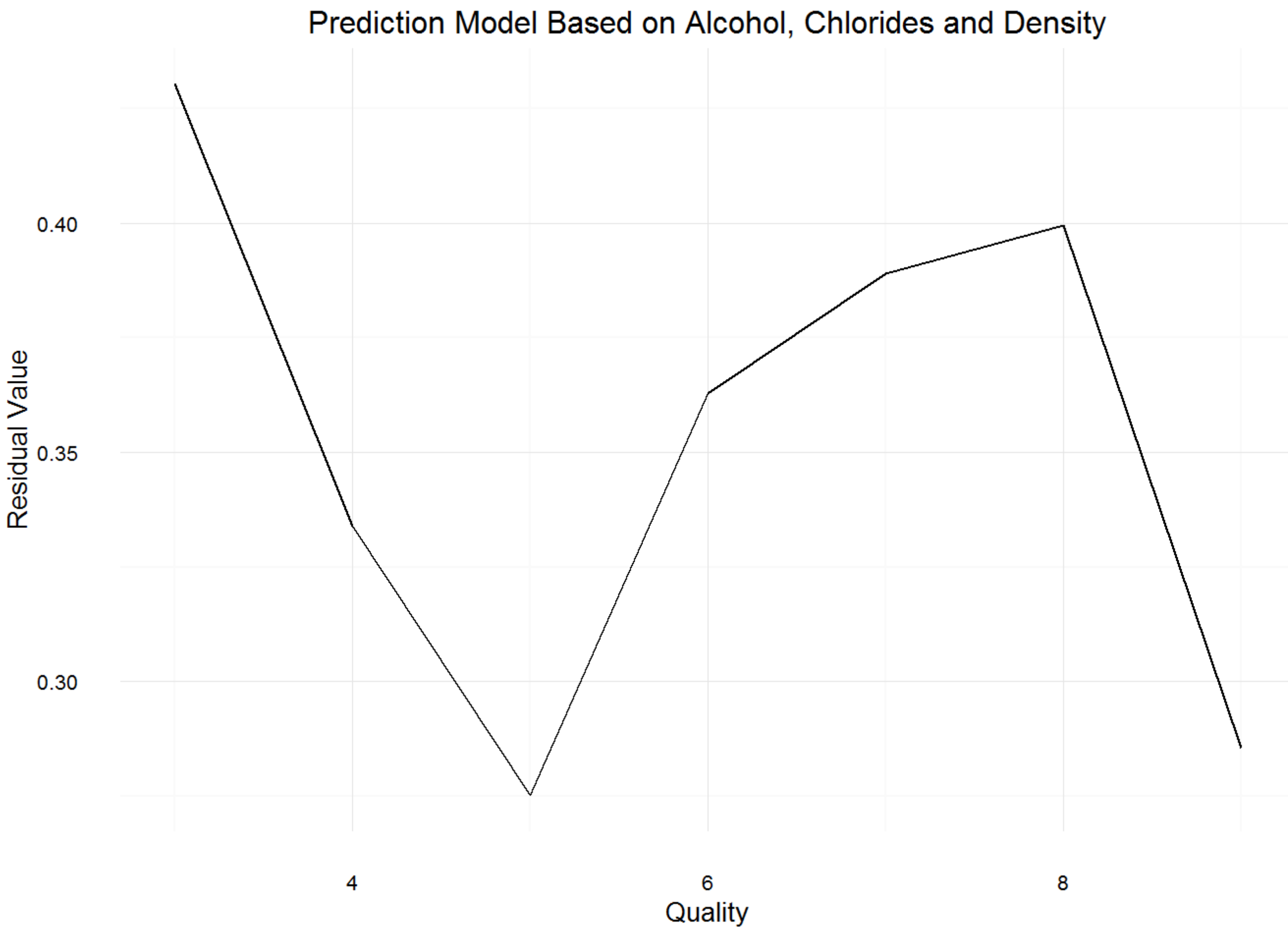
Prediction modeling of Quality based on Chlorides, Alcohol and Density

We will start by using the three main variables we identified as the one's with the most significant correlation. The purpose of this model is to decide if we can determine what quality rating a particular sample is likely to get based on these three variables (Chlorides, Density and Alcohol).

```
##
## Calls:
## model1: lm(formula = quality ~ chlorides, data = wqw)
## model2: lm(formula = quality ~ chlorides + alcohol, data = wqw)
## model3: lm(formula = quality ~ chlorides + alcohol + density, data = wqw)
##
## ====================================================
##                    model1     model2      model3
## ----------------------------------------------------
##   (Intercept)      6.267***   2.861***  -21.150***
##                   (0.029)    (0.116)     (6.162)
##   chlorides       -8.510***  -2.471***   -2.382***
##                   (0.566)    (0.558)     (0.558)
##   alcohol                     0.298***    0.343***
##                              (0.010)     (0.015)
##   density                                23.671***
##                                          (6.074)
## ----------------------------------------------------
##   R-squared           0.044      0.193      0.195
##   adj. R-squared      0.044      0.193      0.195
##   sigma               0.866      0.796      0.795
##   F                 225.727    585.182    396.315
##   p                   0.000      0.000      0.000
##   Log-likelihood  -6244.233  -5829.599  -5822.011
##   Deviance         3671.708   3099.838   3090.247
##   AIC             12494.465  11667.199  11654.021
##   BIC             12513.955  11693.185  11686.504
##   N                4898       4898       4898
## ====================================================
```

```
##        fit      lwr     upr
## 1 5.456604 3.897638 7.01557
```

```
## [1] 0.7943859
```



Interestingly enough, based on the sample data we have and the limited amount of variables, it seems that you might be able to “predict” how a taster might rate a particular sample. Since we are only using 3 out of 11 variables, this might be risky.

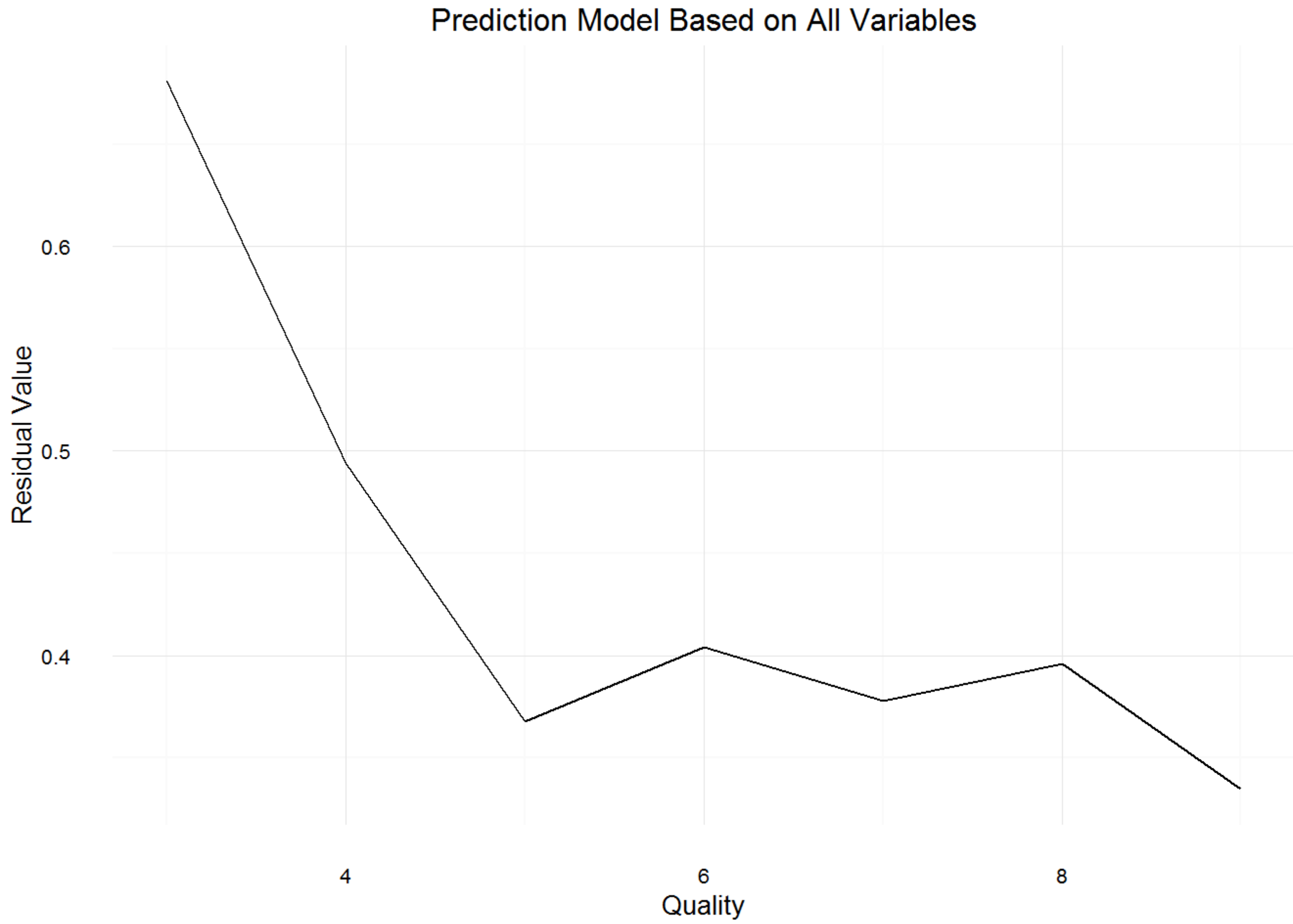
How will our model change if we add all the other variables as well?

```
##
## Calls:
## model1: lm(formula = quality ~ chlorides, data = wqw)
## model2: lm(formula = quality ~ chlorides + alcohol, data = wqw)
## model3: lm(formula = quality ~ chlorides + alcohol + density, data = wqw)
## model4: lm(formula = quality ~ chlorides + alcohol + density + fixed.acidity,
##     data = wqw)
## model5: lm(formula = quality ~ chlorides + alcohol + density + fixed.acidity +
##     volatile.acidity, data = wqw)
## model6: lm(formula = quality ~ chlorides + alcohol + density + fixed.acidity +
##     volatile.acidity + citric.acid, data = wqw)
## model7: lm(formula = quality ~ chlorides + alcohol + density + fixed.acidity +
##     volatile.acidity + citric.acid + residual.sugar, data = wqw)
## model8: lm(formula = quality ~ chlorides + alcohol + density + fixed.acidity +
##     volatile.acidity + citric.acid + residual.sugar + free.sulfur.dioxide,
##     data = wqw)
## model9: lm(formula = quality ~ chlorides + alcohol + density + fixed.acidity +
##     volatile.acidity + citric.acid + residual.sugar + free.sulfur.dioxide +
##     total.sulfur.dioxide, data = wqw)
## model10: lm(formula = quality ~ chlorides + alcohol + density + fixed.acidity +
##     volatile.acidity + citric.acid + residual.sugar + free.sulfur.dioxide +
##     total.sulfur.dioxide + pH, data = wqw)
## model11: lm(formula = quality ~ chlorides + alcohol + density + fixed.acidity +
##     volatile.acidity + citric.acid + residual.sugar + free.sulfur.dioxide +
##     total.sulfur.dioxide + pH + sulphates, data = wqw)
##
## ============================================================================================
============================================================
##                            model1     model2      model3      model4      model5      model6
   model7      model8      model9      model10      model11
## --------------------------------------------------------------------------------------------
------------------------------------------------------------
##   (Intercept)             6.267***   2.861***  -21.150***  -31.387***  -47.652***  -47.452***
  51.392***   50.901***   49.144***   118.471***   150.193***
##                          (0.029)    (0.116)     (6.162)     (6.355)     (6.195)     (6.228)
(13.802)    (13.760)    (14.279)     (18.187)     (18.804)
##   chlorides              -8.510***  -2.471***   -2.382***   -2.421***   -1.323*     -1.344*
  -0.819      -0.923      -0.920       -0.376       -0.247
##                          (0.566)    (0.558)     (0.558)     (0.555)     (0.539)     (0.544)
  (0.544)     (0.543)     (0.543)      (0.548)      (0.547)
##   alcohol                           0.298***    0.343***    0.356***    0.405***    0.404***
   0.306***    0.313***    0.313***     0.231***     0.193***
##                                    (0.010)     (0.015)     (0.015)     (0.015)     (0.015)
  (0.019)     (0.019)     (0.019)      (0.024)      (0.024)
##   density                                       23.671***   34.437***   50.909***   50.709***
-48.356***  -48.108***  -46.328**   -118.102***  -150.284***
##                                                (6.074)     (6.293)     (6.137)     (6.170)
(13.801)    (13.759)    (14.291)     (18.448)     (19.075)
##   fixed.acidity                                            -0.087***   -0.101***   -0.102***
  -0.050**    -0.042**    -0.042**      0.042*       0.066**
##                                                            (0.014)     (0.014)     (0.014)
  (0.015)     (0.015)     (0.015)      (0.021)      (0.021)
##   volatile.acidity                                                     -2.085***   -2.079***
```

```
  -2.057***   -1.994***   -1.984***    -1.910***    -1.863***
##                                                                               (0.110)     (0.112)
  (0.111)     (0.112)     (0.114)      (0.114)      (0.114)
##   citric.acid                                                                              0.031
   0.035      -0.005      -0.003        0.047        0.022
##                                                                                           (0.097)
  (0.096)     (0.096)     (0.096)      (0.096)      (0.096)
##   residual.sugar
   0.044***    0.041***    0.040***     0.069***     0.081***
##
  (0.005)     (0.005)     (0.006)      (0.007)      (0.008)
##   free.sulfur.dioxide
               0.004***    0.004***     0.004***     0.004***
##
              (0.001)     (0.001)      (0.001)      (0.001)
##   total.sulfur.dioxide
                          -0.000       -0.000       -0.000
##
                          (0.000)      (0.000)      (0.000)
##   pH
                                        0.646***     0.686***
##
                                       (0.106)      (0.105)
##   sulphates
                                                     0.631***
##
                                                    (0.100)
## -----------------------------------------------------------------------------------------------
----------------------------------------------------------------
##   R-squared                0.044      0.193      0.195       0.202       0.256       0.256
     0.266       0.270       0.271        0.276        0.282
##   adj. R-squared           0.044      0.193      0.195       0.201       0.255       0.255
     0.265       0.269       0.269        0.275        0.280
##   sigma                    0.866      0.796      0.795       0.792       0.764       0.764
     0.759       0.757       0.757        0.754        0.751
##   F                      225.727    585.182    396.315     309.222     336.912     280.725
   252.899     226.580     201.396      186.351      174.344
##   p                        0.000      0.000      0.000       0.000       0.000       0.000
     0.000       0.000       0.000        0.000        0.000
##   Log-likelihood       -6244.233  -5829.599  -5822.011   -5802.684   -5629.932   -5629.882
-5597.945   -5582.289   -5582.183    -5563.494    -5543.740
##   Deviance              3671.708   3099.838   3090.247    3065.956    2857.136    2857.077
  2820.061    2802.090    2801.968     2780.668     2758.329
##   AIC                  12494.465  11667.199  11654.021   11617.368   11273.865   11275.763
11213.891   11184.579   11186.366    11150.988    11113.480
##   BIC                  12513.955  11693.185  11686.504   11656.348   11319.341   11327.736
11272.360   11249.545   11257.828    11228.947    11197.936
##   N                     4898       4898       4898        4898        4898        4898
  4898        4898        4898         4898         4898
## ===============================================================================================
================================================================
```

```
##        fit      lwr      upr
## 1 5.562658 4.088118 7.037198
```

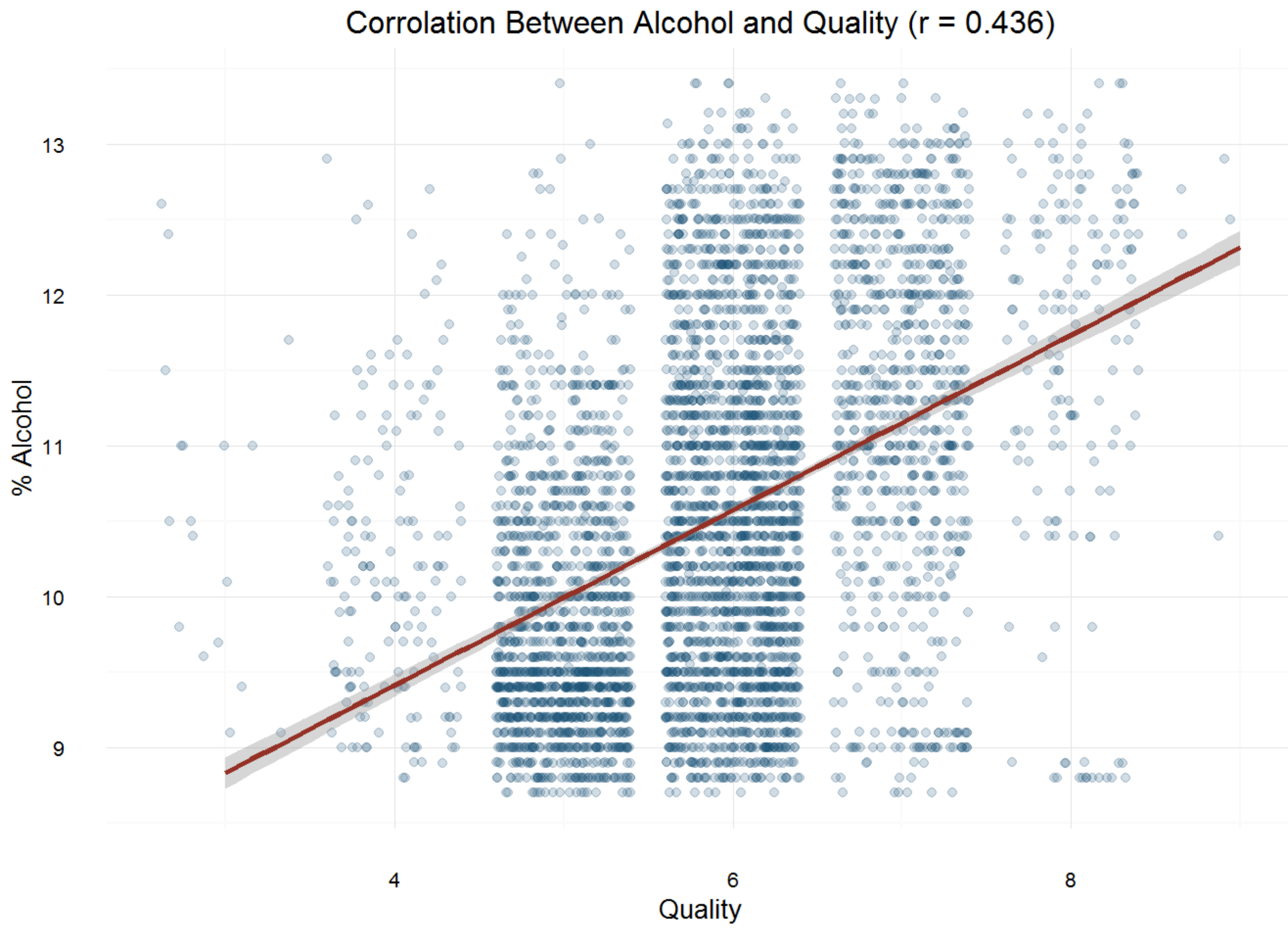
```
## [1] 0.7505125
```



Since the remaining 8 variables had almost no coloration to the rating, our best fit value changed very little. If anything, looking at the plot above, it seems that including the other variables introduced some outlier data points that are skewing our numbers.

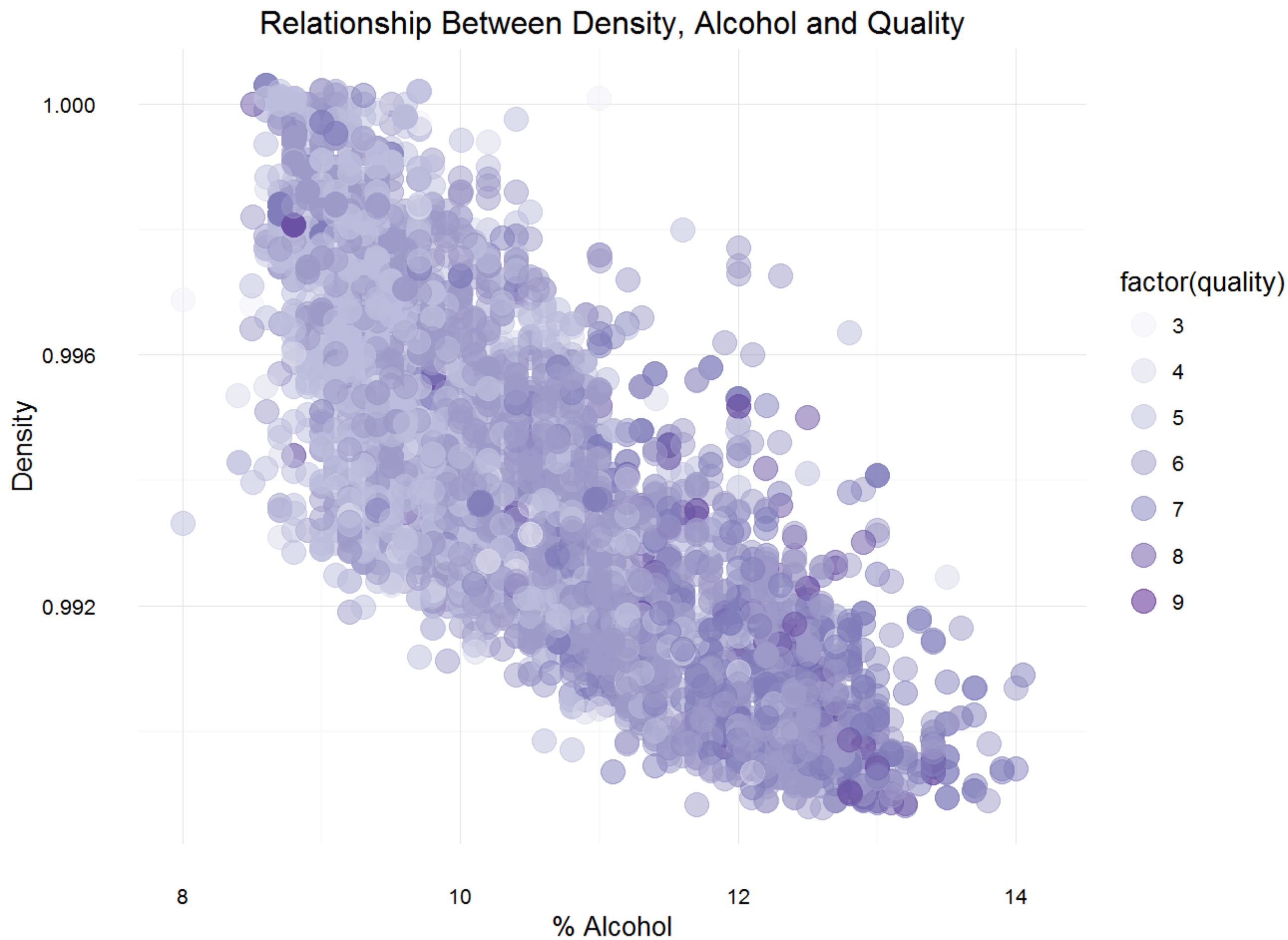
Final Plots and Summary

Plot One



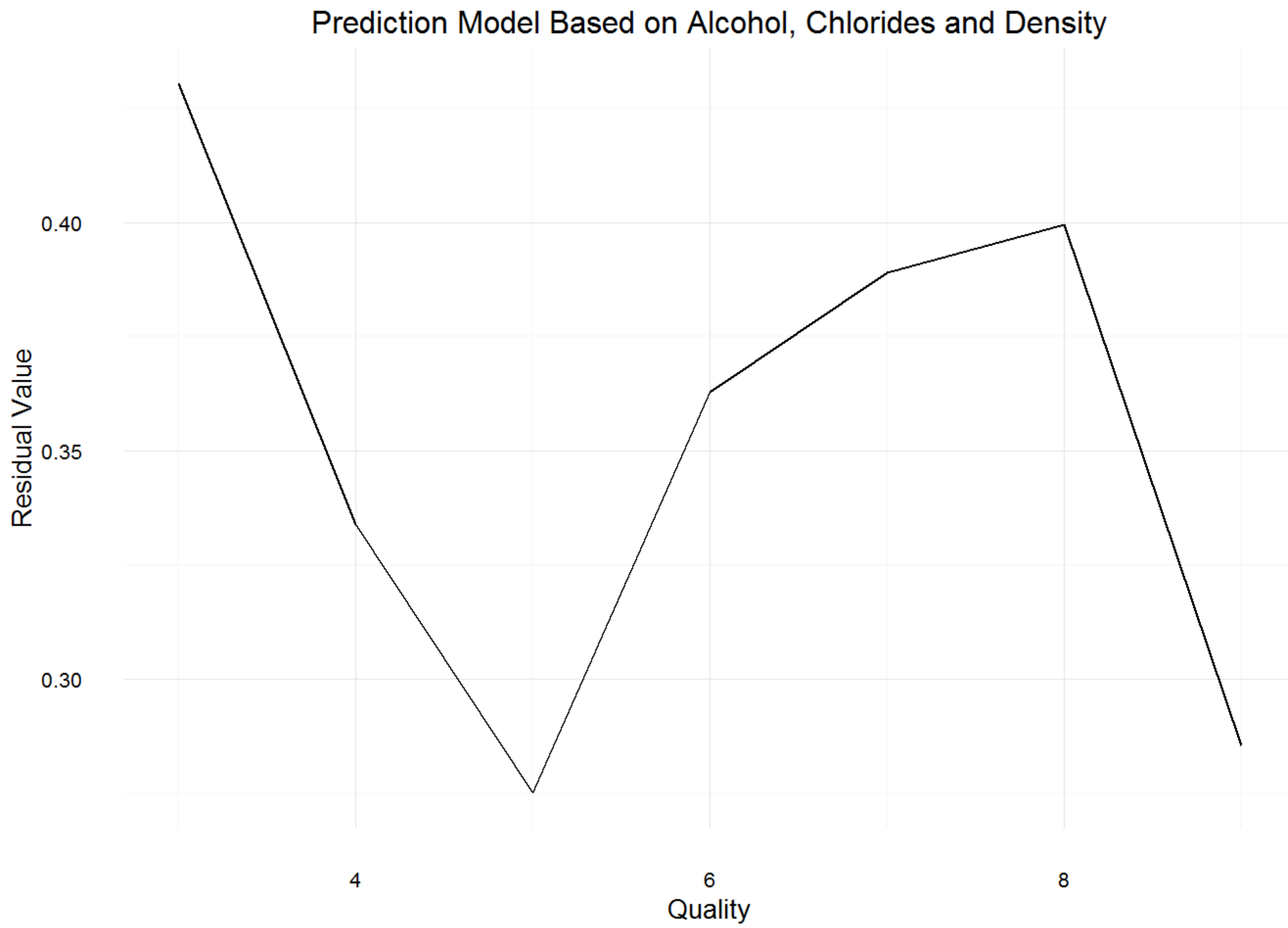
The first plot above shows to overall correlation between alcohol contents and quality ratings. I chose to focus on the alcohol contents because I was very surprised to see this come up. I'm not familiar with the ins and outs of wine tasting, but I never thought the alcohol contents will make a difference on the tasting. Or maybe this is just a bias we have in our data set?

Plot Two



The second plot Shows the multi-variant relationship between Alcohol, Density and Quality. The colors of the plot seem to darken as the density decreases and alcohol content increases, which indicates an increase in the quality rating of these samples. This supports our findings that Alcohol has a positive relationship with the quality rating and that Density has a negative relationship with Quality. It also supports the strong negative relationship we found between Alcohol contents and Density (r = -0.780).

Plot Three



The third plot displays the residual value of each data point in our data set compared to our model. Using Alcohol, Density and Chloride variables it looks like we came up with a solid model that we may be able to use to predict the rating a particular sample might get from a taster. This makes sense since we developed the model against the exact same data set we used for testing, but unfortunately I have no way to reproduce the testing with a minimum of 3 reviewers while capturing the variables of the samples they are rating.

Reflection

The main thing I took away from this project is that even though we have a working model we have no way to truly verify if it is correct because of lack of additional data not included in the data set we used for our prediction model. It would be interesting to try and take a subsection of the same data set of 4000 data points and then run evaluate the correctness of the remaining 898 samples not included in the creating of the model.

In addition, I'm not familiar with the process of making wine, but I would think that the variables captured are the result of the process used to make the wine and not the other way around. It might not be possible to accurately control these variables and "optimize" the process to make a better tasting wine.

Lastly, I would definitely **not** drink other liquids expecting a pleasant experience just because they happen to have optimal Alcohol, Residual Sugar and Density values. There are probably quite a lot of liquids out there that would not only taste horrible but are probably bad for your health as well that would score quite high based on our prediction model.

Citation

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