P4_final_project: EDA of the White Wine dataset

CWT

08/01/2017

Introduction

This report explores a dataset containing quality and attributes for approximately 4900 white wines. The dataset is related to white variants of the Portuguese “Vinho Verde” wine [1]. The inputs include objective tests (e.g. PH values) and the output is based on sensory data (median of at least 3 evaluations made by wine experts). Each expert graded the wine quality between 0 (very bad) and 10 (very excellent).

The Input Variables

The eleven input variables of this white wine dataset based on physicochemical tests are listed below with their respective units of measurement.

1 - Fixed Acidity (tartaric acid - g/dm^3)

2 - Volatile Acidity (acetic acid - g/dm^3)

3 - Citric Acid (g/dm^3)

4 - Residual Sugar (g/dm^3)

5 - Chlorides (sodium chloride - g/dm^3)

6 - Free Sulfur Dioxide (mg/dm^3)

7 - Total Sulfur Dioxide (mg/dm^3)

8 - Density (g/cm^3)

9 - pH (potential of hydrogen)

10 - Sulphates (potassium sulphate - g/dm^3)

11 - Alcohol (% by volume)

Note: dm^3 stands for cubic decimeter which is equivalent to 1 Liter

Output Variable

The single output variable in this dataset is quality. This attribute is subjective based on expert sensory data. Each expert graded the wine quality between 0 (very bad) and 10 (very excellent). The quality variable for each white wine in this study is based on the median of at least 3 evaluations made by wine experts. Throughout this exploratory data analysis the input variables are measured against the quality variable in an attempt to determine what chemical attributes make a high quality white wine.

More about the Analysis of What Makes a High Quality Wine

The Sugar Content

Sugar content is an essential indicator of vinification (the conversion of grape juice or other vegetable extract into wine by fermentation). Measured during maturity inspections, the sugar content allows the date of the harvest to be anticipated. It takes approximately 16.83 g/l of sugar for a yeast to produce 1% volume of alcohol. It also allows the alcoholic fermentation to be monitored (therefore the transformation of the sugar into alcohol), and it can give the quantity of residual sugars in a wine. The residual sugars contribute to the balance of a wine [2].

Residual Sugar (g/l)

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.

Alcohol (% by volume)

The Alcohol of each wine in the study is the percent alcohol content of the wine. The alcohol content is also an indicator, the gustatory approach to a wine with 11% alcohol by volume will be different to a wine with 14% volume. But note that the alcohol content alone is not an indicator of the quality of the wine. A wine with 14% volume does not necessarily translate into a higher quality than a wine with 14% volume. Everything is a question of balance with the other parameters (sugars, acidity, etc.) and material.

Total Acidity

This is the sum of the volatile acidity and the fixed acidity.

Wine contains a certain number of mineral and organic acids. Some of these acids in the wine are entirely combined with alkali - they are in salt form and therefore are not involved in the perception of the acidity of the wine. However, certain organic acids are only partially saturated by alkali or turned into salts. Some of their molecules are in salt form, others are free. The sum of the free acid functions and the partly free acids constitutes the acidity of the wine. The three main acids in wine are tartaric, malic and citric acid. Tartaric acid is the specific acid of the grape and of wine - it is found very little in natural form except in vines. Malic acid is an organic acid which is very common in the plant world, it is the main acid in many fruits. Citric acid exists in grapes of every variety and in greater quantity in most concentrated by gray fungus that is deliberately cultivated on grapes to enhance the making of certain sweet wines.

Fixed Acidity

Most acids involved with wine are fixed or nonvolatile (do not evaporate readily).

Volatile Acidity

The volatile acidity of a wine consists of the part of the fatty acids belonging to the acetic series, which is found in wine either in the free state or in salt form. Volatile acidity gives the wine its characteristic taste, but when the dose of volatile acid is too high the wine is said to be "sour", characterized by notes of glue or vinegar.

Therefore the amount of acetic acid in wine at too high of levels can lead to an unpleasant vinegar taste.

The pH

Parallel to the total acidity, the free acid functions are partially dissociated or ionized and release H+ ions into the liquid, which represent the “real” acidity, whose concentration is expressed as a pH. It depends to a large extent on tartaric acid.

pH is a scale of acidity from 0 to 14. It tells how acidic or alkaline a substance is. More acidic solutions, have lower pH. More alkaline solutions, have higher pH. Substances that aren’t acidic or alkaline (that is, neutral solutions) usually have a pH of 7.

Sulfur Dioxide (SO2)

The use of sulfur dioxide is widely accepted as a useful winemaking aide. It is used as a preservative because of its anti-oxidative and anti-microbial properties in wine, but also as a cleaning agent for barrels and winery facilities [3]. To make high quality wines, a winemaker needs to learn how to effectively manage the sulfite levels. When done correctly, maintaining the proper amount of sulfites in a wine protects it from oxidation and microbial contamination during ageing and storage. The amount of SO2 needed to protect a wine is based on the wine’s pH. The higher the pH the more SO2 will be needed, and conversely, the lower the pH the less SO2 will be needed to attain the ideal level [4]. SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident to the smell and taste of wine.

Sodium Chloride

The amount of salt in the wine. Generally moderate to large concentrations of salt give the wine a salty flavor which may turn away potential customers. Generally a wine with less than .06 g/dm^3 is generally referred to as a low salt level [5].

My Exploratory Data Analysis of the White Wine Dataset

The first order of this EDA is to look at the summary of the wine attributes. This information provides the basic statistics to be used throughout the analysis, helping with bin sizing and scaling as appropriate. Along with the mean and median, I always like to have the standard deviations of the variables as well.

As I scan through the various attributes that were given about wines I notice a categorical variable called ‘quality’. I focus on this variable because my thought is a higher quality wine would be more broadly appealing and therefore have a higher value and command a higher price.

As mentioned in the introduction, the quality parameter is between 0 and 10 however within this dataset the minimum is 3 and there are no 10’s. As the focus of this EDA is on the quality parameter and attempts to fit a multiple linear regression model to help explain what makes a high quality white wine, I partitioned the quality parameter into 3 bins; High, Med, and Low. High is defined as quality values of 7, 8, and 9, Med defined as a quality value of 6, and Low is defined as quality values of 3, 4, and 5. The focus will be on the High quality.

The Summary of the Wine Attributes

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

The Standard Deviation of Each Wine Attributes

```
##        fixed.acidity     volatile.acidity          citric.acid
##          0.843868228          0.100794548          0.121019804
##       residual.sugar            chlorides  free.sulfur.dioxide
##          5.072057784          0.021847968         17.007137325
## total.sulfur.dioxide              density                   pH
##         42.498064554          0.002990907          0.151000600
##            sulphates              alcohol              quality
##          0.114125834          1.230620568          0.885638575
```

Correlation Matrix

The following matrix provides the correlation between each pair of attributes. This is a very important first step when doing EDA because it provides a quick way to determine which pair of variables have the strongest relationships, Each cell in the matrix provides the correlation coefficient (r) which measures the strength and direction of a linear relationship between pair of variables. The value of r is always between +1 and-1. r = 0 indicate no relationship, r = +/- 1 indicate perfect positive and negative relationship respectively, below +/- .50 indicate weak relationship, between +/-.50 and +/- .70 a moderate relationship, and greater than +/- .70 a

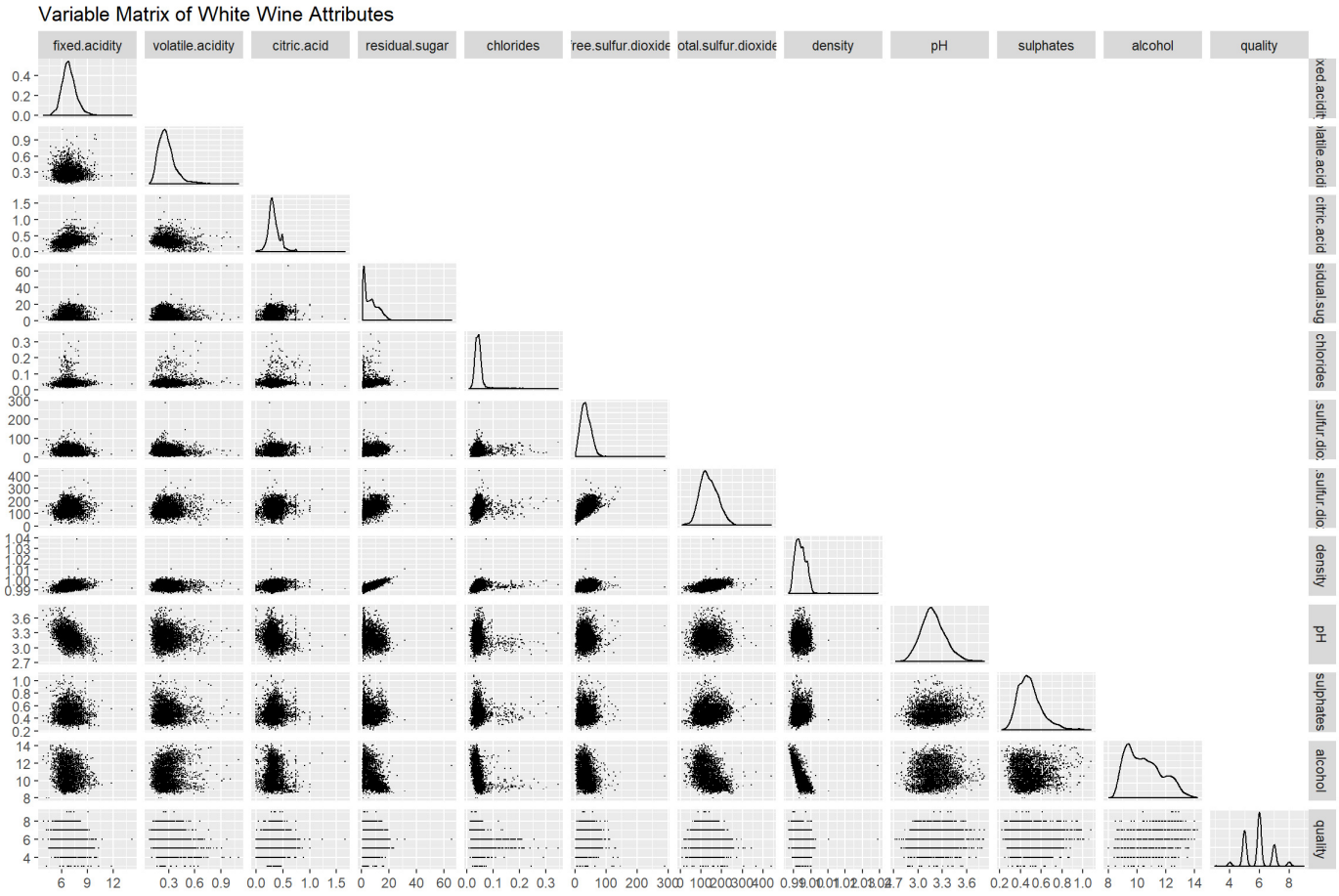
strong relationship.

Reviewing correlation between white wine attributes and specifically focusing on the quality attribute column, the strongest relationship is with alcohol although at .44 it is a weak relationship. Including all other pairs in the matrix the strong relationships are between density vs. residual sugar at .84, and density vs. alcohol with -.78.

Matrix of the Scatterplots, and Univariate Analysis

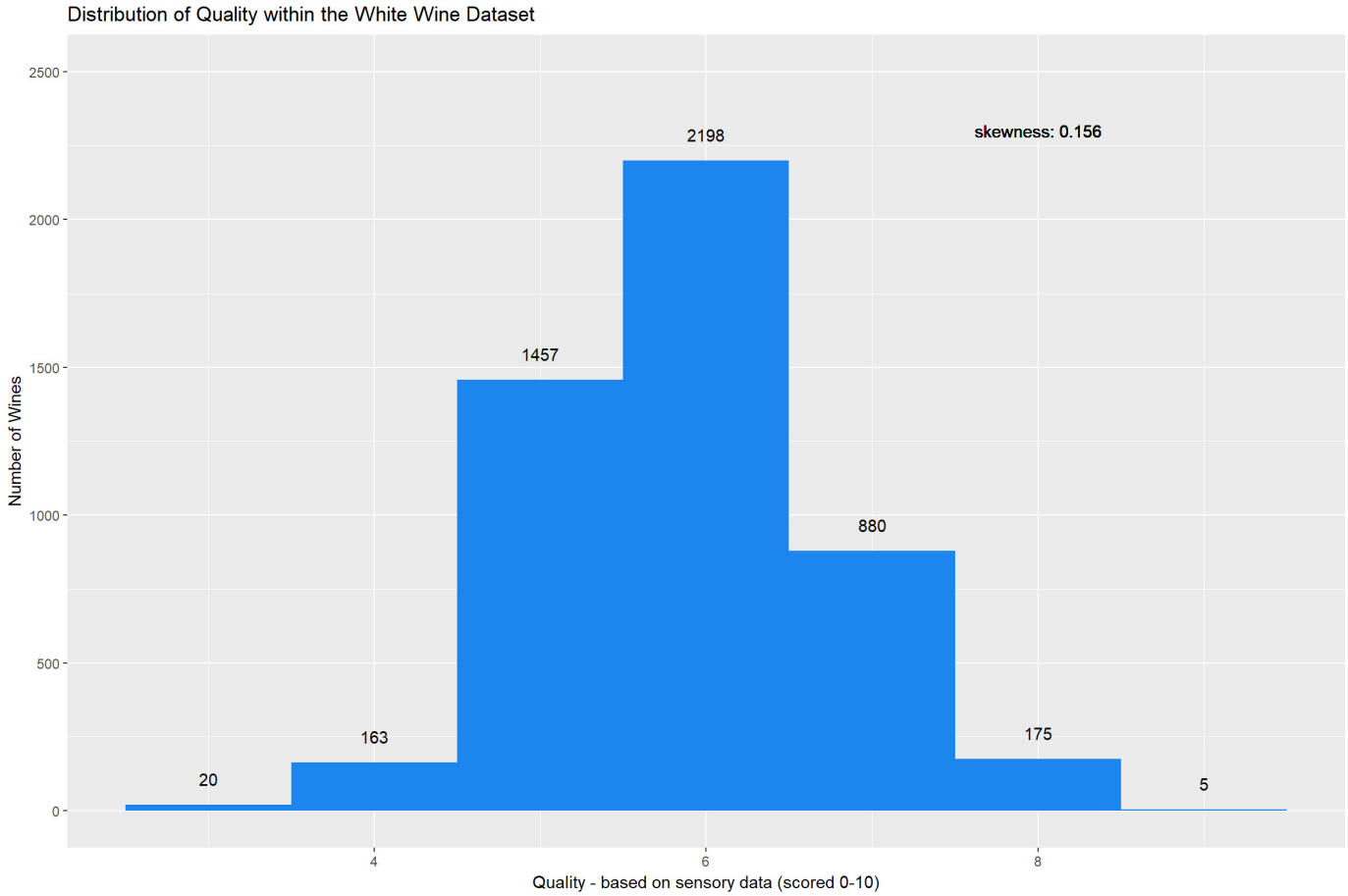
Along with the correlation matrix, I find it very helpful when in the discovery mode to look at a very cursory view of the scatterplots of all pairs with my study. The following matrix provides this matrix using the ggpairs function. Rather than just the single correlation coefficient this matrix provides a more visual look at the relationships. Also the diagonal of this matrix provide a compact univariate analysis of all the variables. It illustrates the distribution of each variable. We will go into more depth of each distribution next but I find having all variables together in one figure very helpful to gain a better understanding of the variables. I will note that with ggpairs the correlations and scatterplots can be combined into one figure however I chose to separate them.

Most of the white wine variables exhibit a normal distribution with the exception of residual sugar which has a positive skew. Alcohol is also a little bit different than a normal distribution however I would not say it is dramatically skewed.



The “Quality”" Variable Distribution

The distribution of the quality variable is given in the figure below. Recall in this dataset the quality values are discrete whole numbers from 3 to 9. We see within this histogram the quality of 6 squarely in the middle of the distribution is the dominate value at 2198. And we note there are more wines below this mean than above indicating a slight positive skew [8] - calculating the skew with the moments library results in a skew of 0.156 as depicted in the drawing which is very slight as the visual suggests.



Distribution of Variables (Univariate and Bivariate Plots)

Now let's dive into each variable with more detail. As I explore this data further I wanted to add the component of quality into each variable. So within each of the bins in the distribution I use quality levels (High, Med, and Low) to visualize how the quality levels are distributed.

Quality Levels:

High = 7, 8, and 9

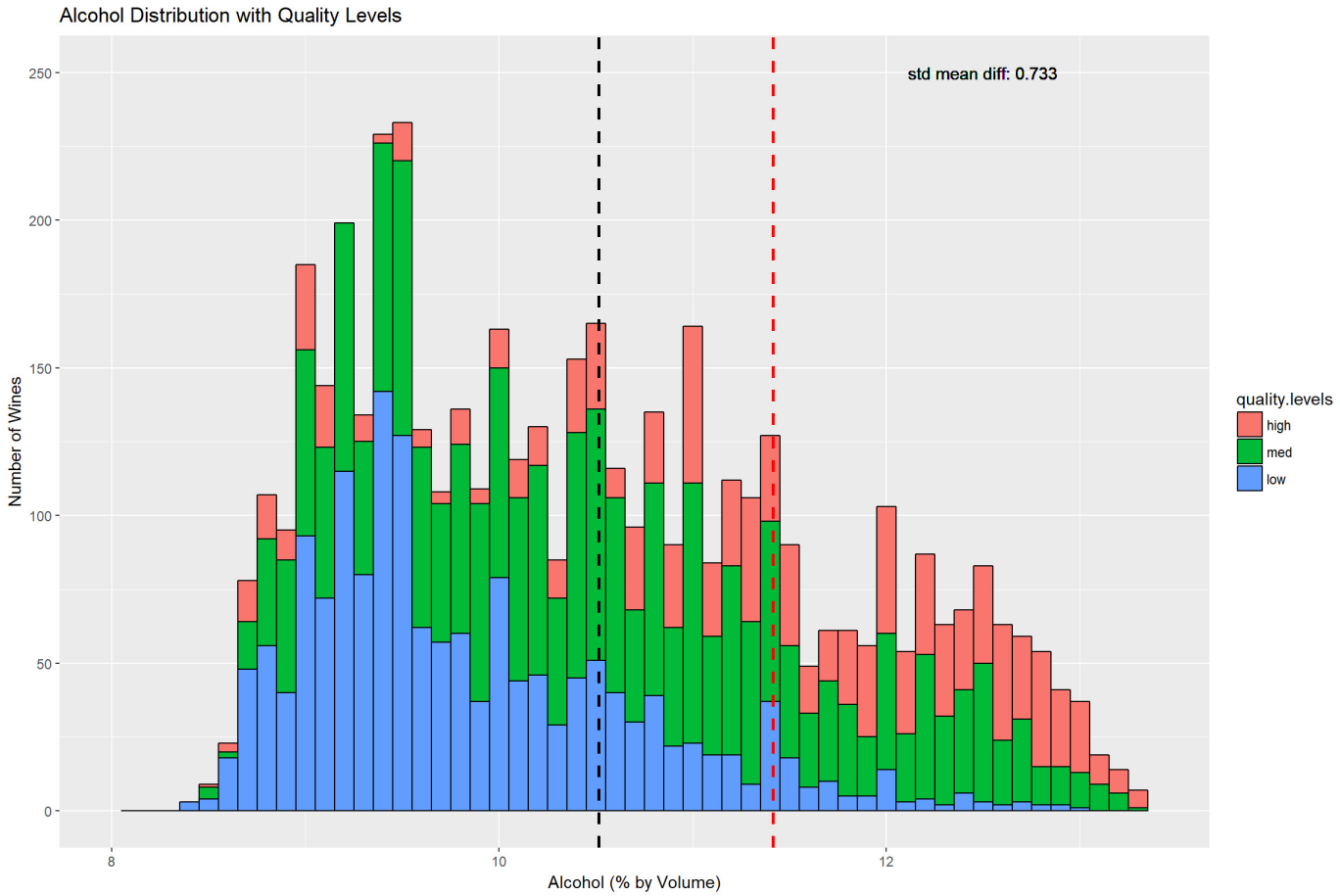
Med = 6

Low = 3, 4, and 5

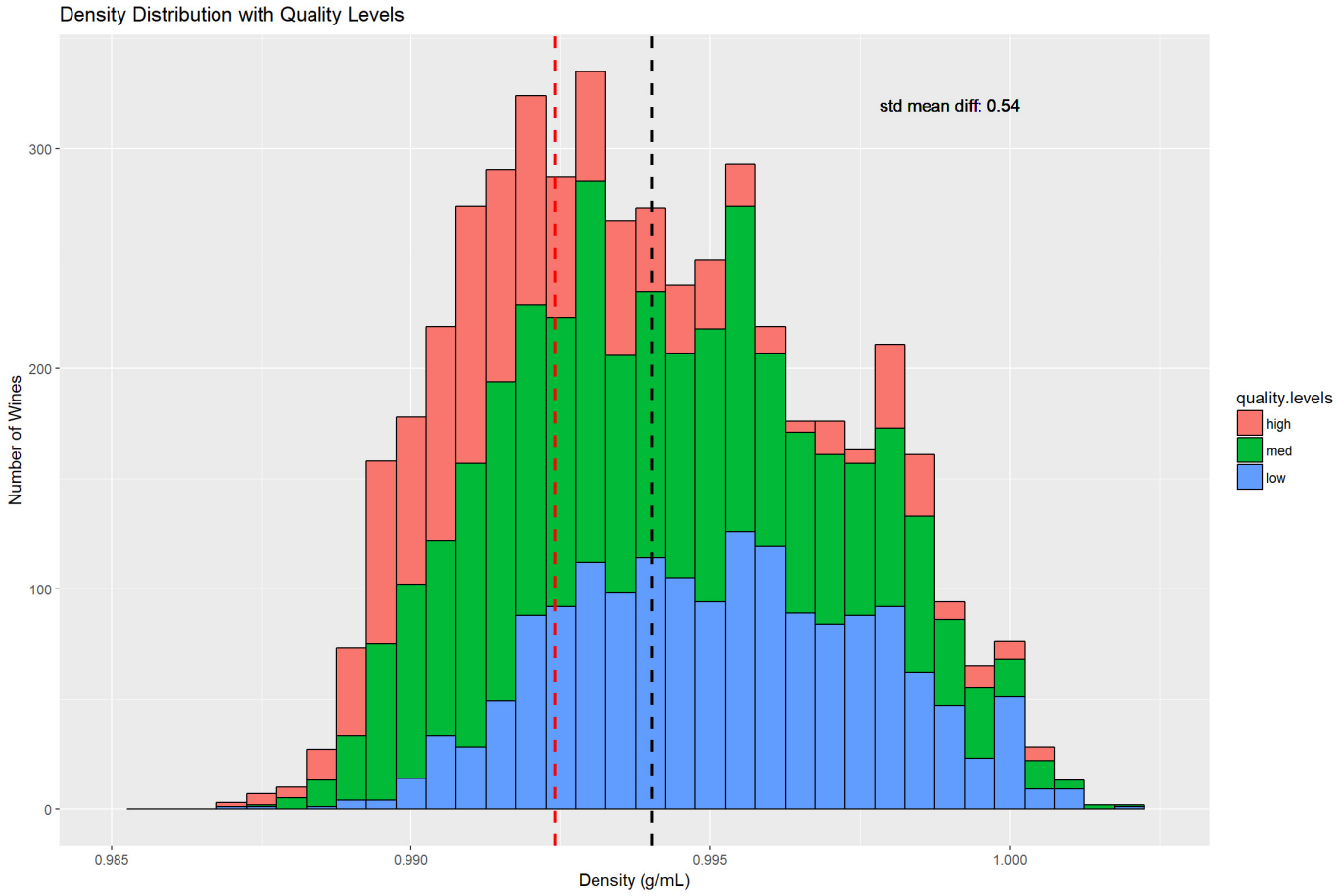
Furthermore, I calculated the mean of the distribution for all wines and mean of the distribution for the high quality wines only. For each graph I have the means depicted with a vertical dotted lines, the black line is the mean of the variable for all wines and red is the mean of the variable for taking into account only the high quality wines. With the means I wanted to see how much difference there were between the two means to determine how much of an impact a particular variable has on the high quality wines. The thought is if there is a significant difference between the two means then it would give me more information about how much of an effect a variable would have on high quality. Each graph provides a 'std mean diff' number than is the standardized difference between the two means. The standardized number is calculated by dividing the absolute value of the difference between the two means with the standard deviation associated with that variable.

Alcohol

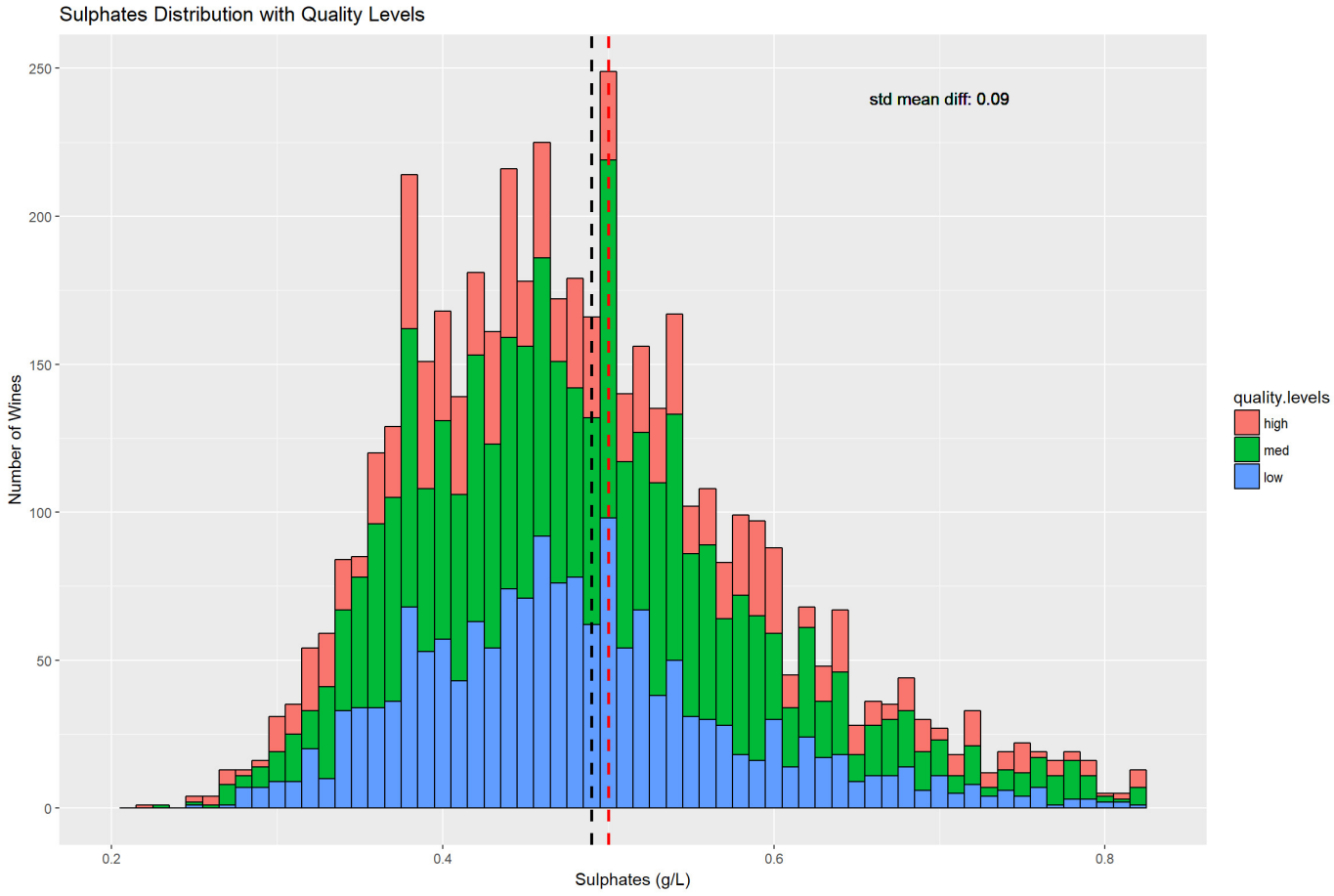
The mean of the high quality wines are greater than the overall mean and the std diff mean is .733 which means it's less than one standard deviation. This indicates that while there is some difference between these two populations they are not significant. This agrees with the correlation study where the correlation between alcohol and quality was .44, a weak relationship. I expect the rest of the variables will have an even weaker relationship but let's go through each of them to see how much influence they have on the high quality wines.



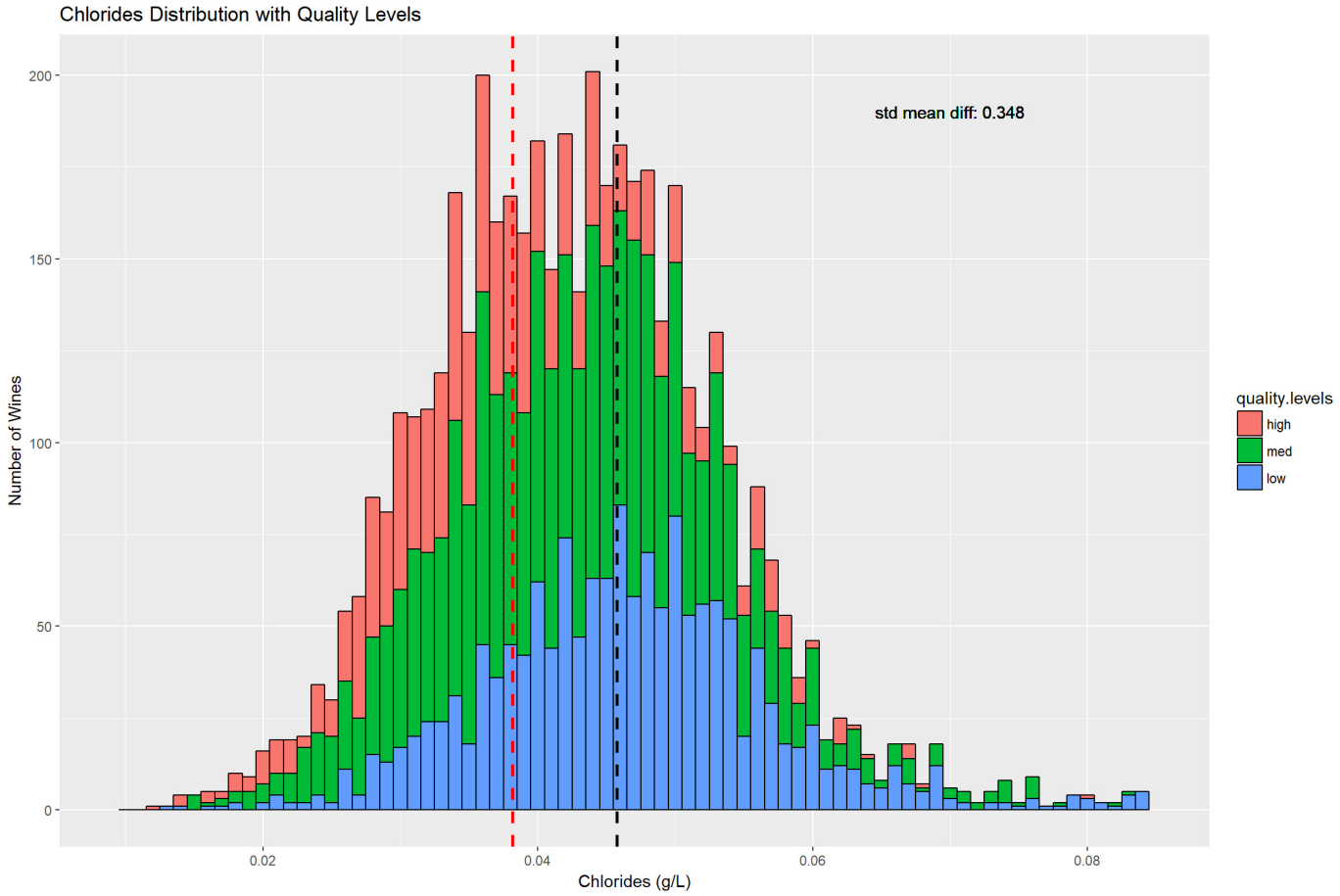
Density


Density Distribution with Quality Levels
std mean diff: 0.54
Number of Wines
300
200
100
0
0.985
0.990
0.995
1.000
Density (g/mL)
quality.levels
high
med
low

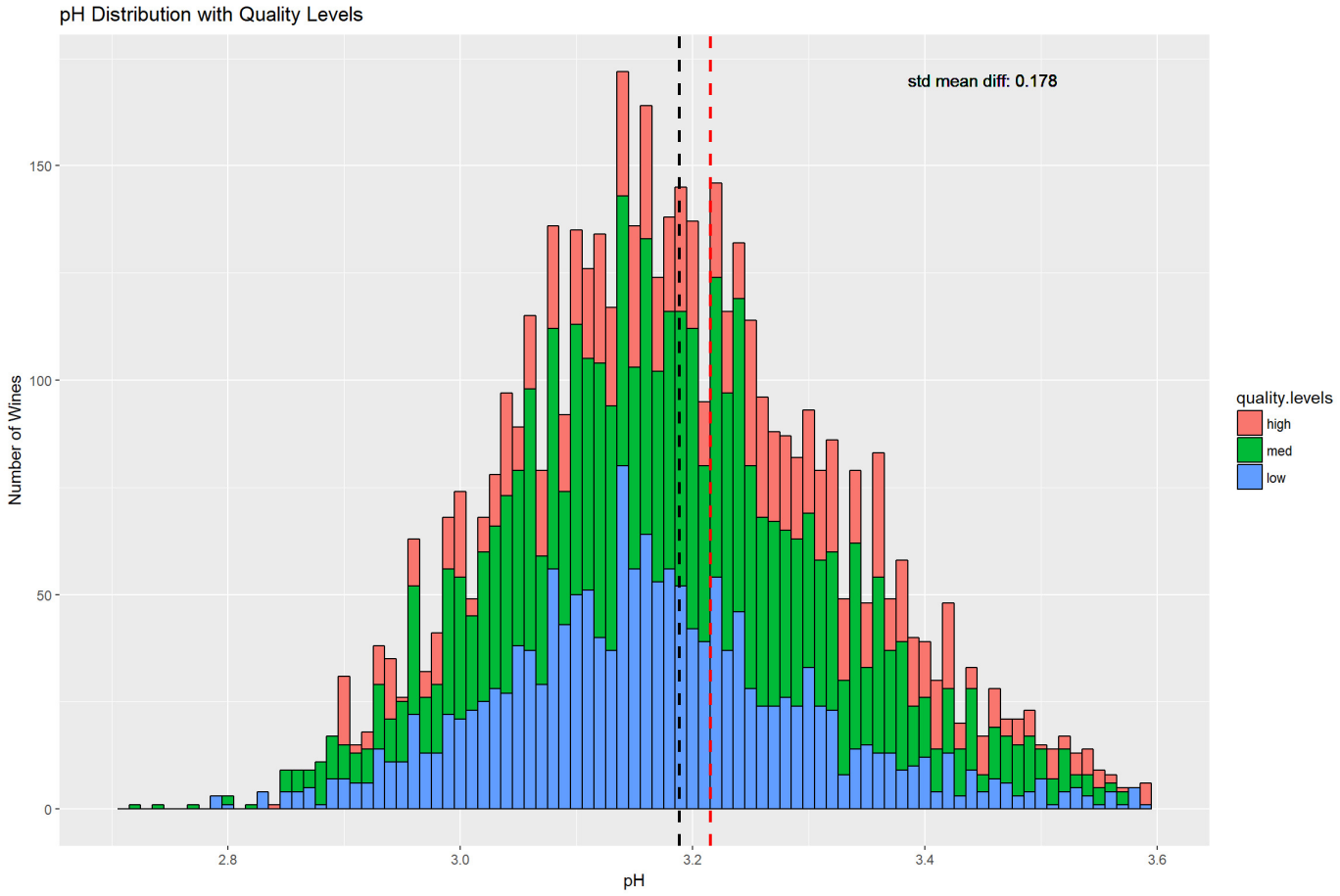
Sulphates


Sulphates Distribution with Quality Levels
std mean diff: 0.09
Number of Wines
250
200
150
100
50
0
0.2
0.4
0.6
0.8
Sulphates (g/L)
quality.levels
high
med
low

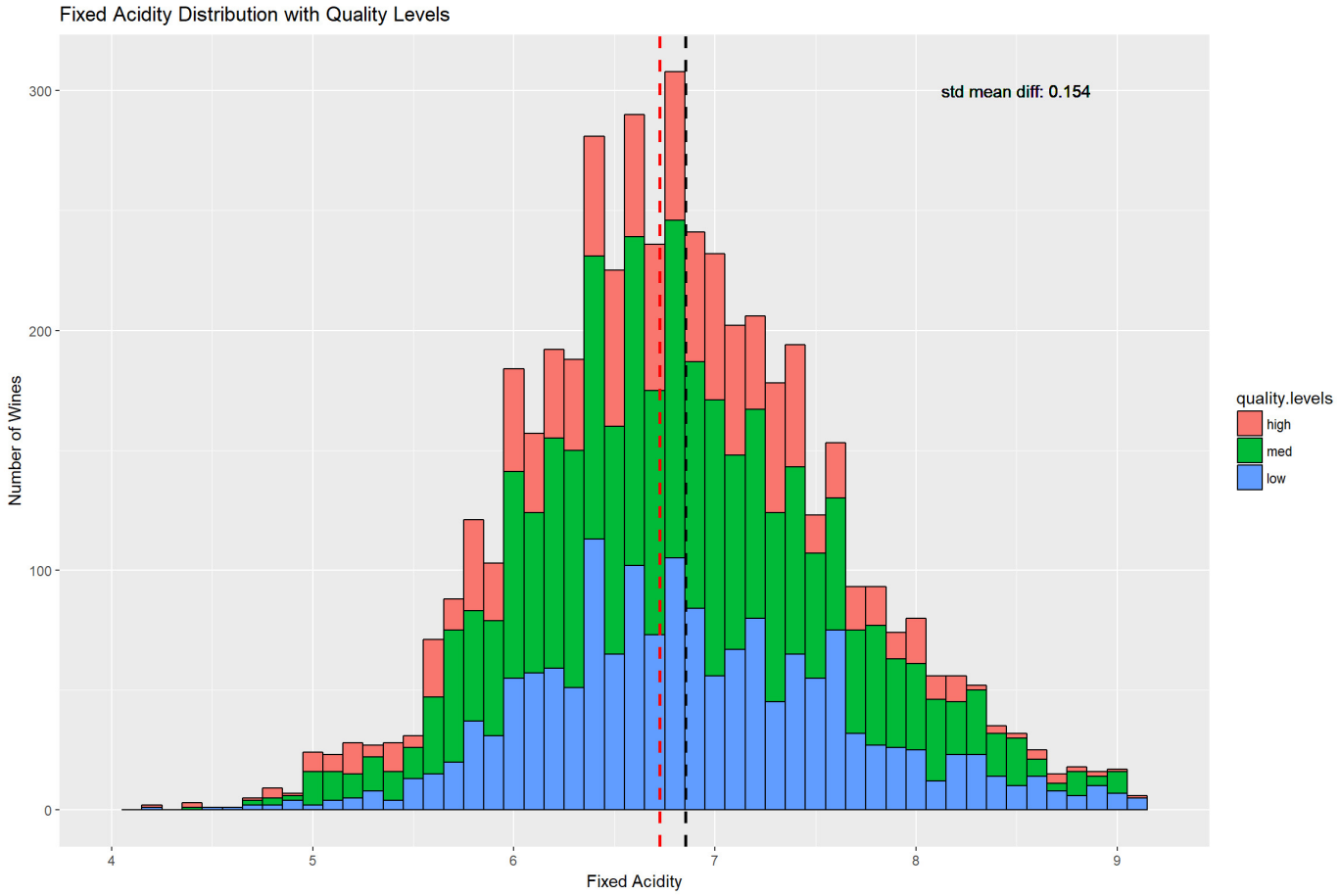
Chlorides



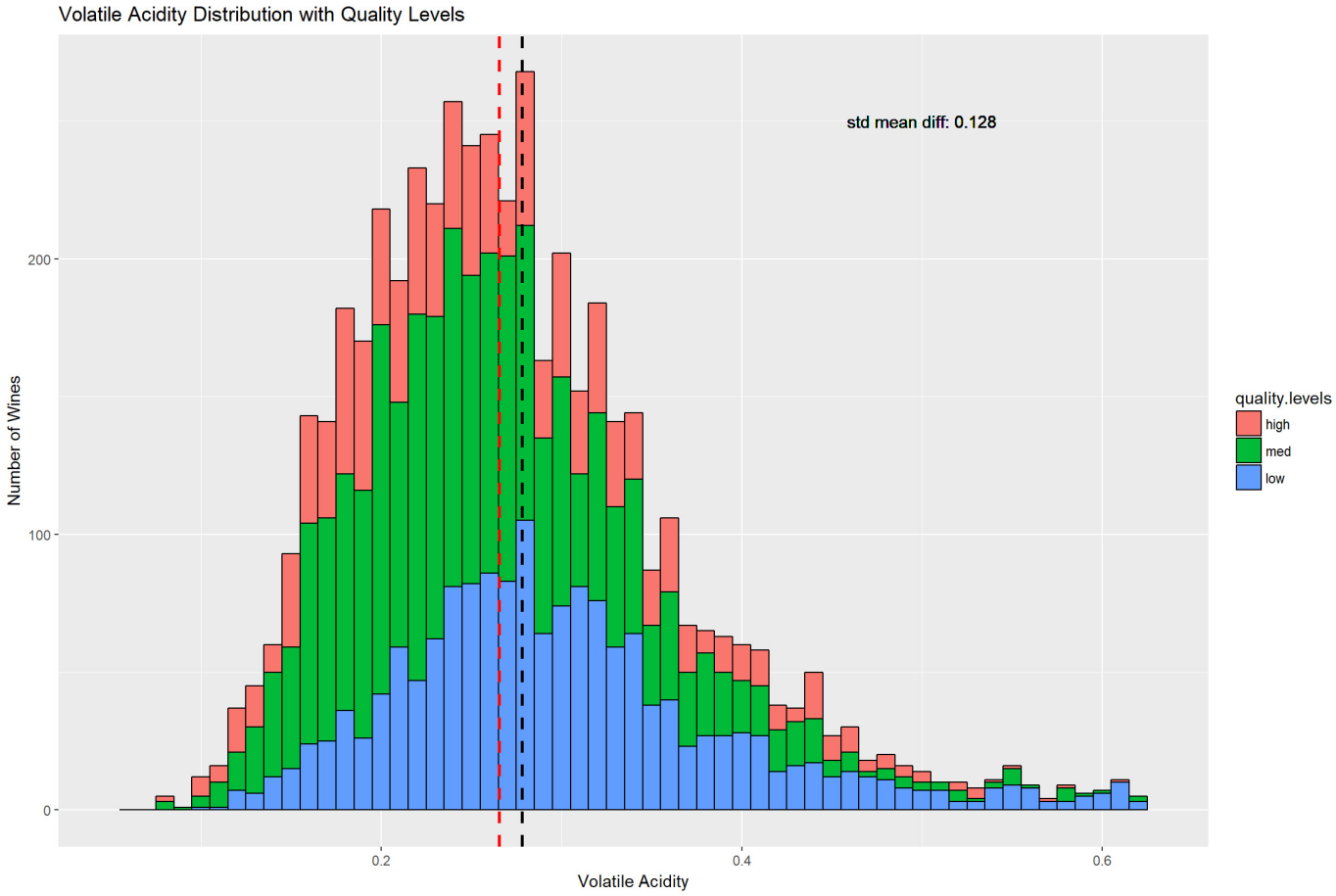
pH



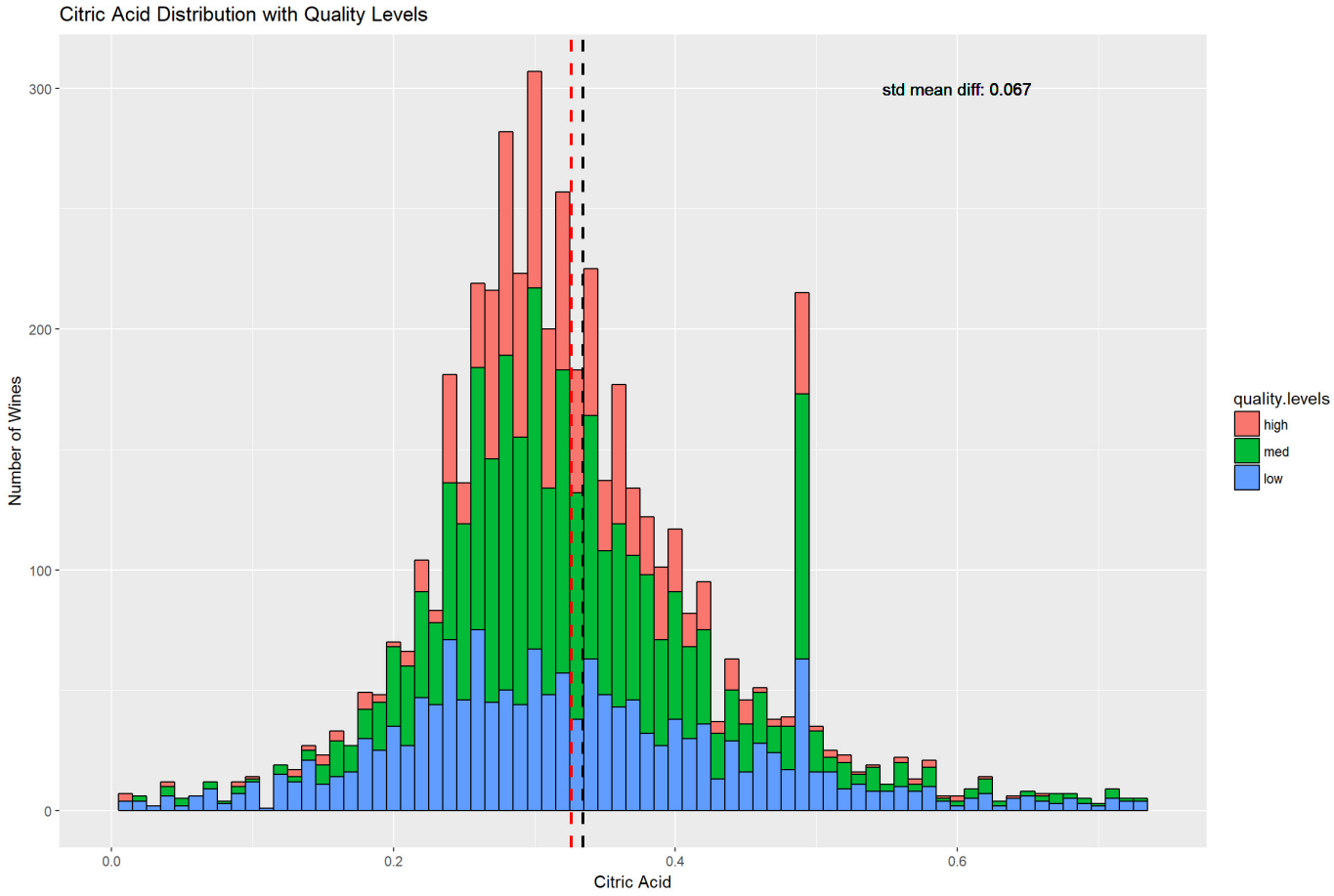
Fixed Acidity



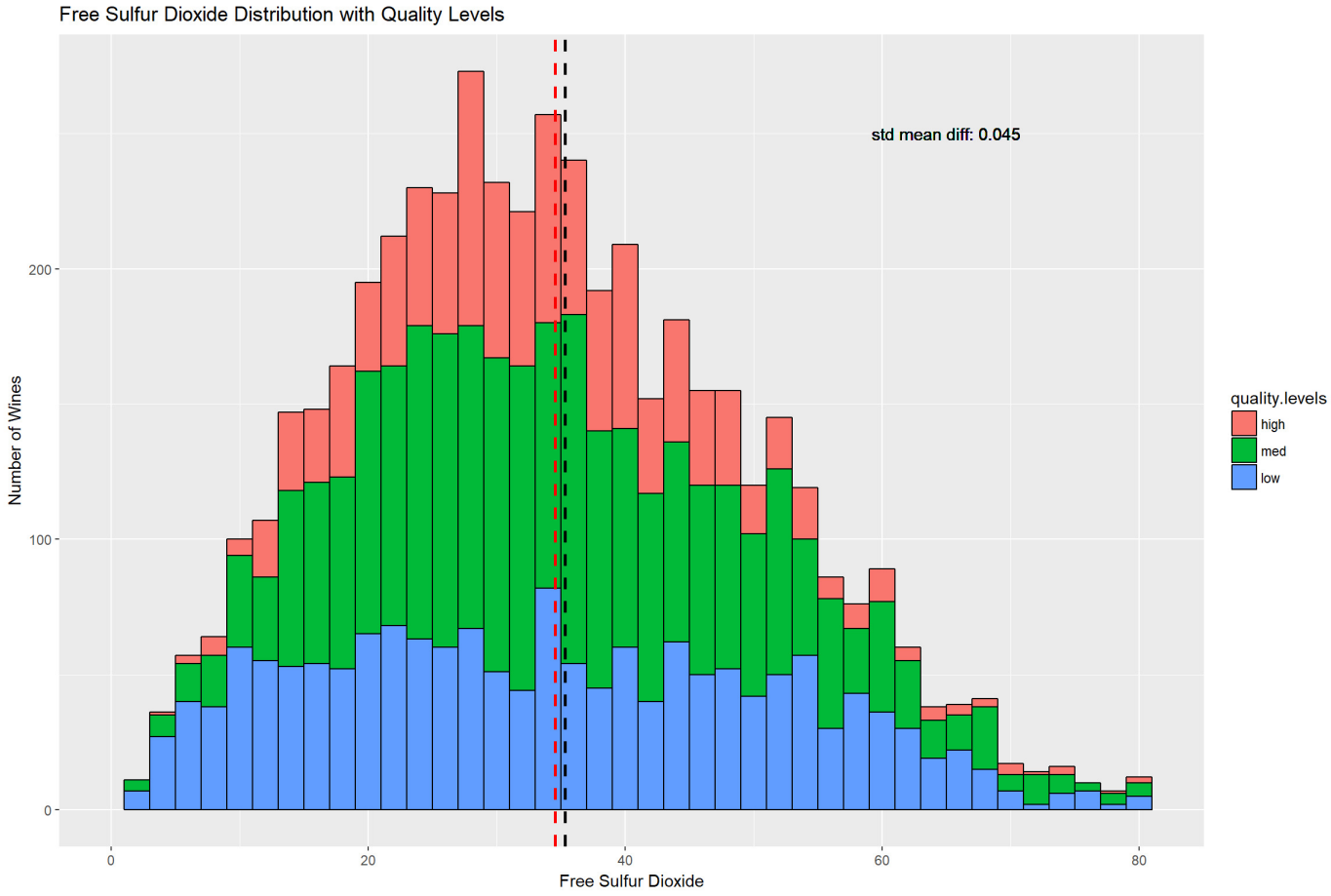
Volatile Acidity



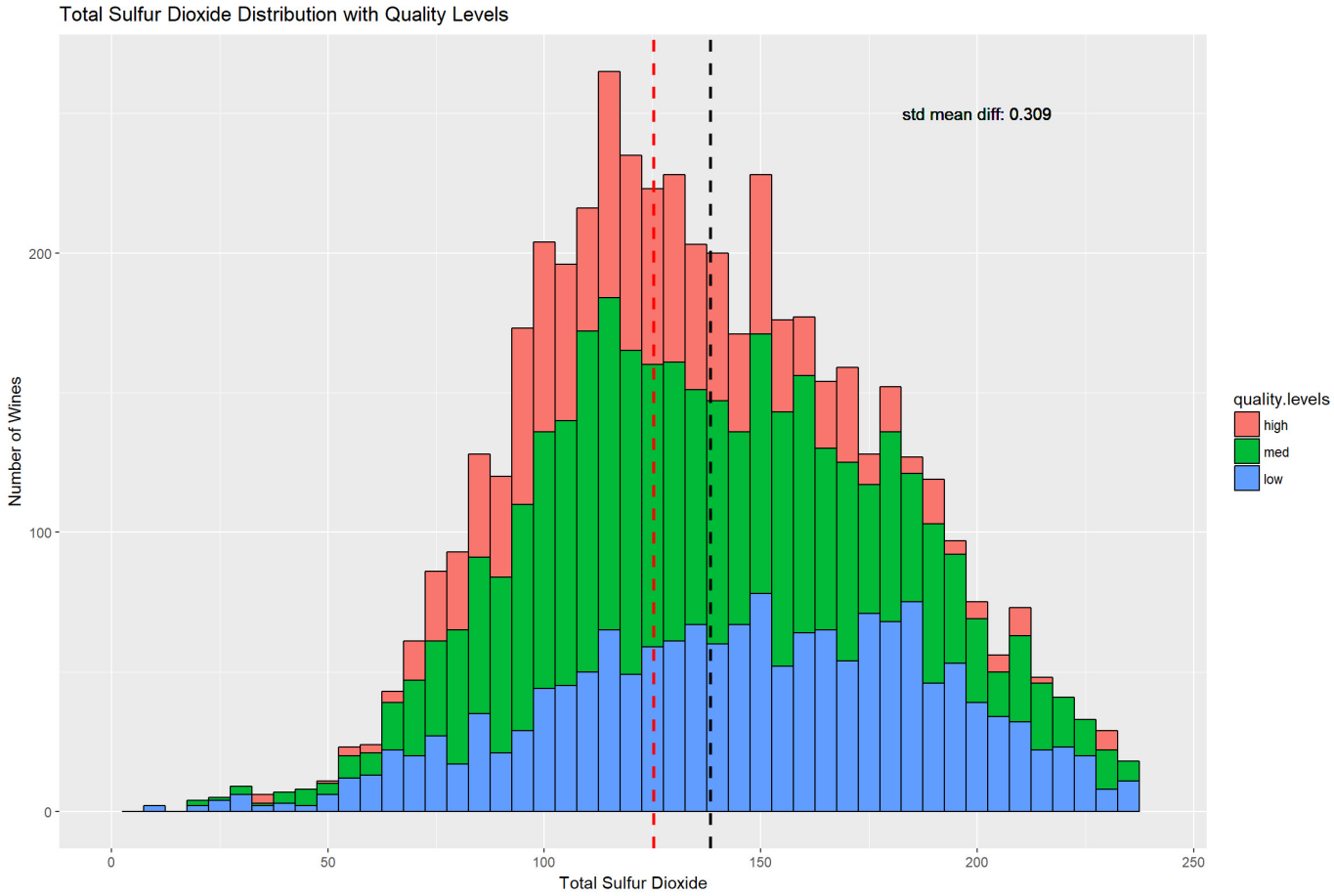
Citric Acidity



Free Sulfur Dioxide

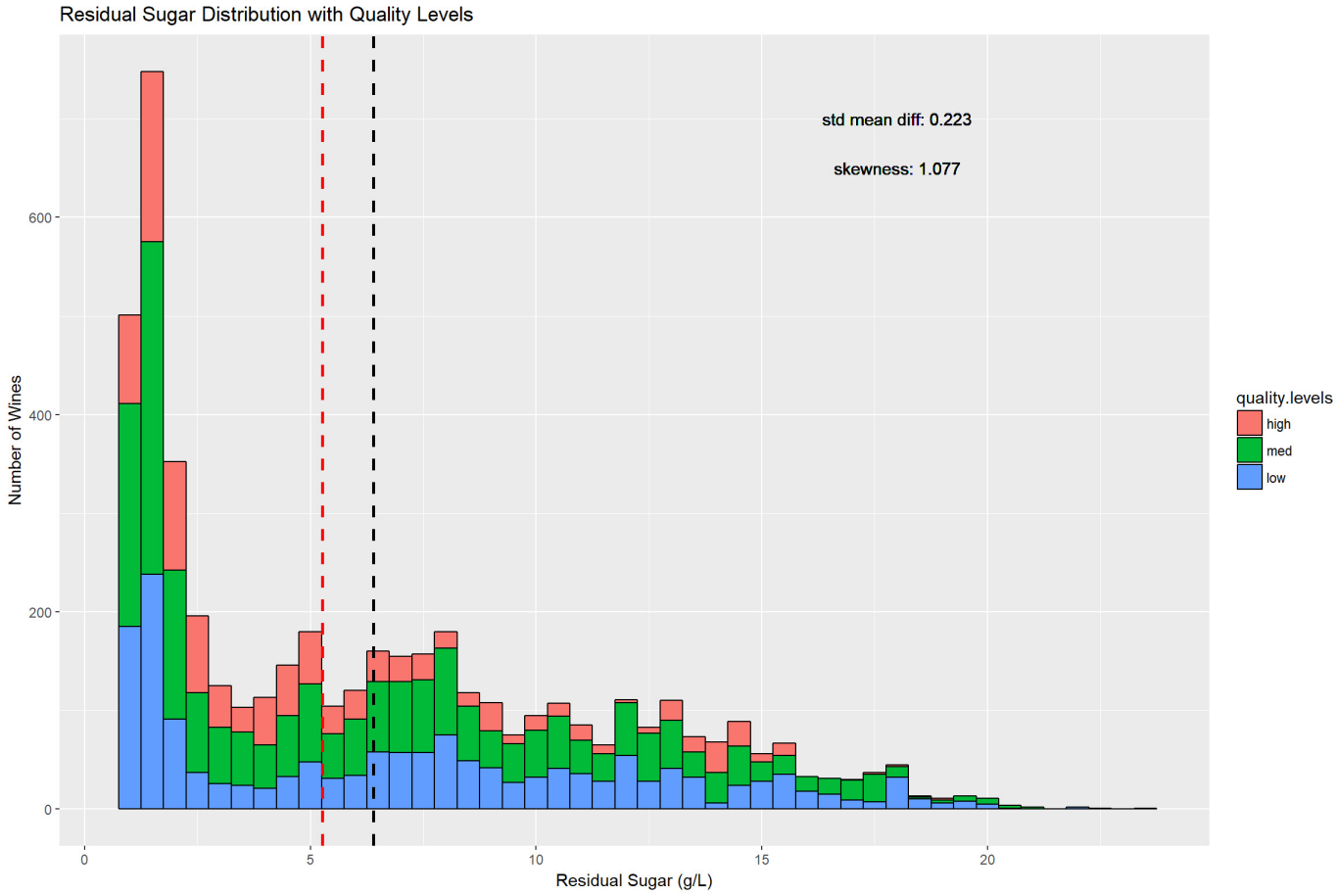


Total Sulfur Dioxide

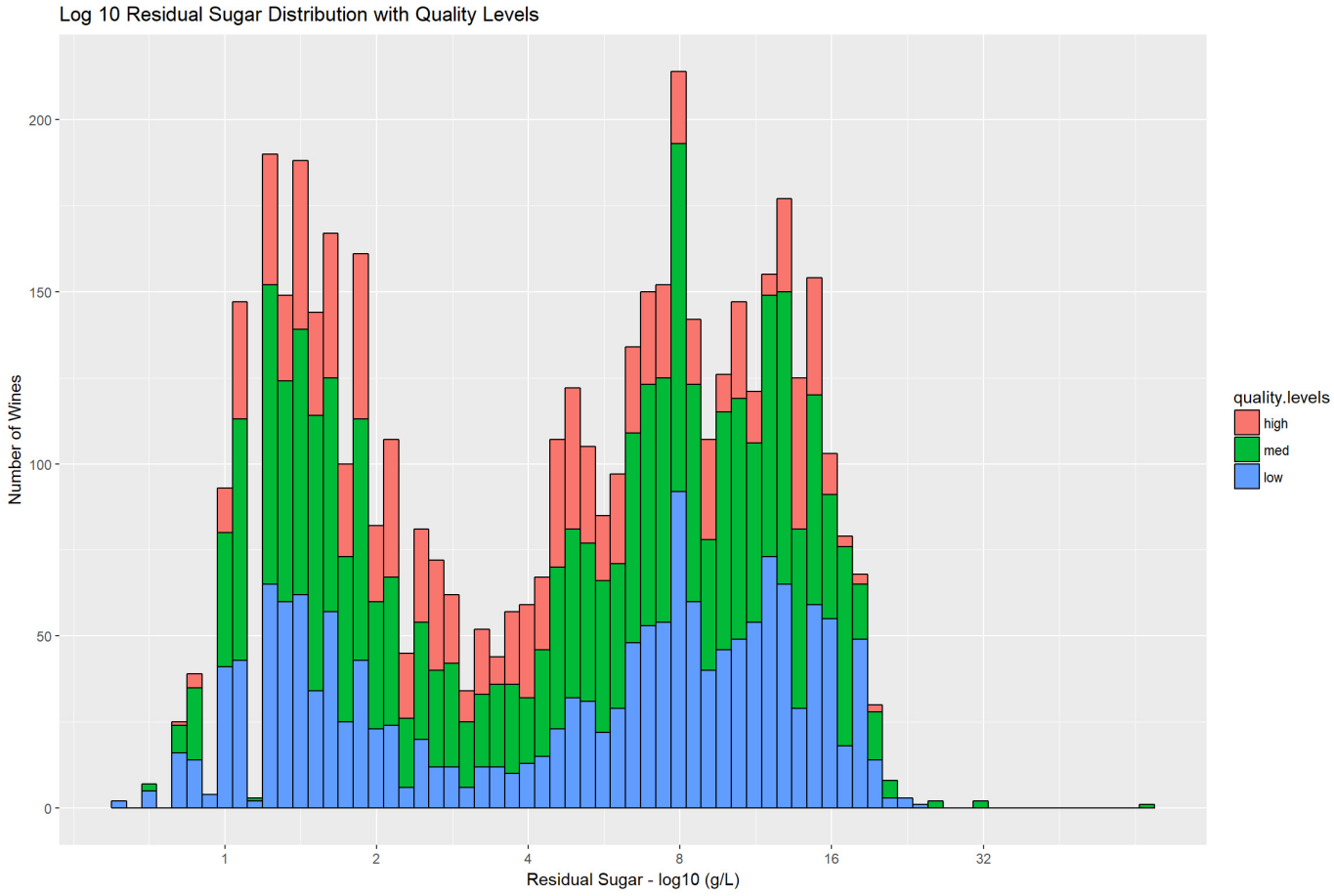


Residual Sugar

Residual Sugar is a unique attribute in this dataset because it's a dramatically skewed distribution. We deal with this kind distribution by transforming the scale to log 10. The succeeding plot is the log 10 distribution and Residual Sugar, which turns it into a bimodal distribution.



Residual Sugar (log 10)



Standardized Mean Difference (a.k.a. Cohen's d)

The following list provides the standardized mean difference for each variable. This is also referred to as Cohen's d. However a typical Cohen's d relationship is performed between means of different populations. Here I use the relationship between the mean of the entire population versus the mean of high quality population (which is a subset of the entire population). Therefore we'll have to take this into account when interpreting the effect size of the results. Even though this Cohen's d analysis does not use independent populations I feel like the results are helpful in determining the relative relationship between the white wine variable and high quality.[9]

1 - Fixed Acidity = 0.154

2 - Volatile Acidity = 0.128

3 - Citric Acid = 0.067

4 - Residual Sugar = 0.233

5 - Chlorides = 0.348

6 - Free Sulfur Dioxide = 0.045

7 - Total Sulfur Dioxide = 0.309

8 - Density = 0.540

9 - pH = 0.178

10 - Sulphates = 0.090

11 - Alcohol = 0.733

A Note about Effect Size.

Cohen originally gave guidelines as to what constitutes a small, medium, and large effect. He gave the guidelines of 0.2 as small, 0.5 as medium, and 0.8 and higher as large. However, these guidelines are only potentially useful if you are doing research in an area that has no history of effect sizes, or that you are unable to evaluate for yourself whether the obtained effect is practically meaningful.

So when interpreting effect sizes we will want to contextualize the effects in relation to similar research in other areas to give a sense of where the result stands relative to the field of study. Otherwise the reviewer will not know how to evaluate the findings. There are usually no absolutely "big" or absolutely "small" effects - there are only meaningful effects relative to the research area under investigation.[10]

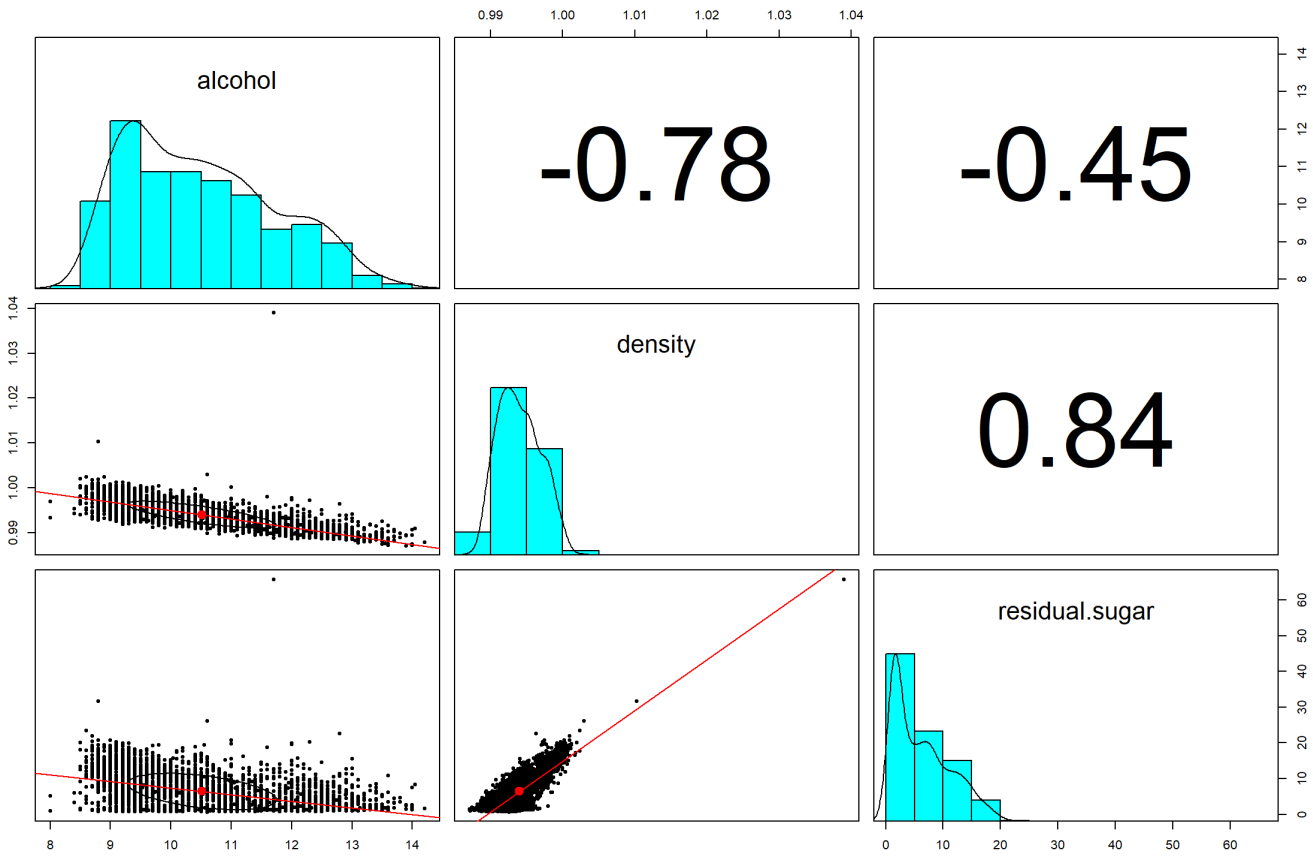
Effect Size of the White Wine Attributes

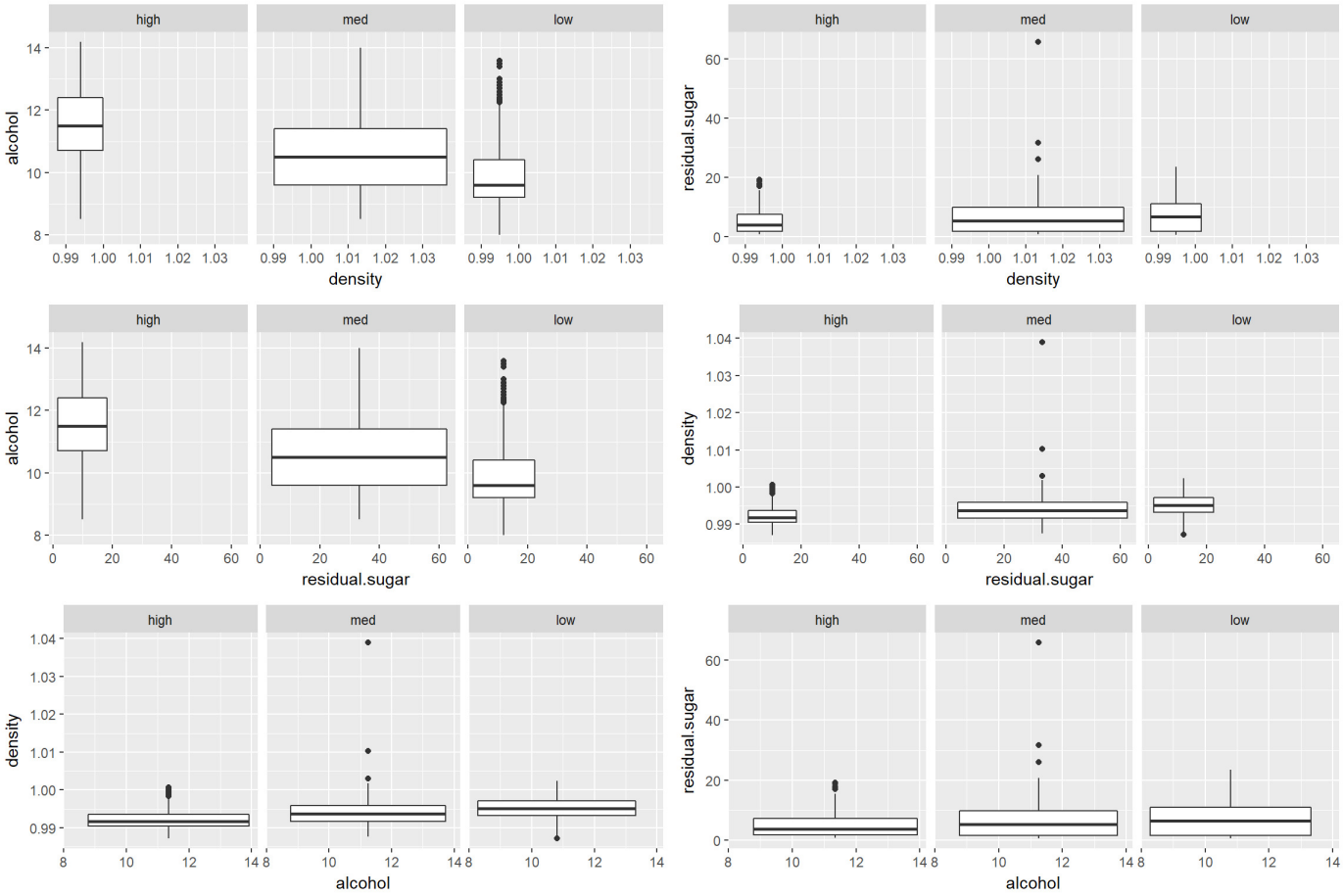
The determination of the degree of effect size in this study is to compare each white wine attribute against the others. For example alcohol and density have a highest effect size of the attributes, while citric acid and free sulfur dioxide have the smallest. Therefore we would expect alcohol and density to be greater predictors of quality level then citric acid and free sulfur dioxide.

Bivariate Plots

Now we turn our attention to the Bivariate Plots by focusing on the variables that have the highest correlation - alcohol, density, and residual.sugar. High quality wine is strongly correlated with the higher alcohol wines with low density and low sugar. Based on R-squared these three variables explain about 21% of the variance in

quality.


alcohol
-0.78
-0.45
density
0.84
residual.sugar

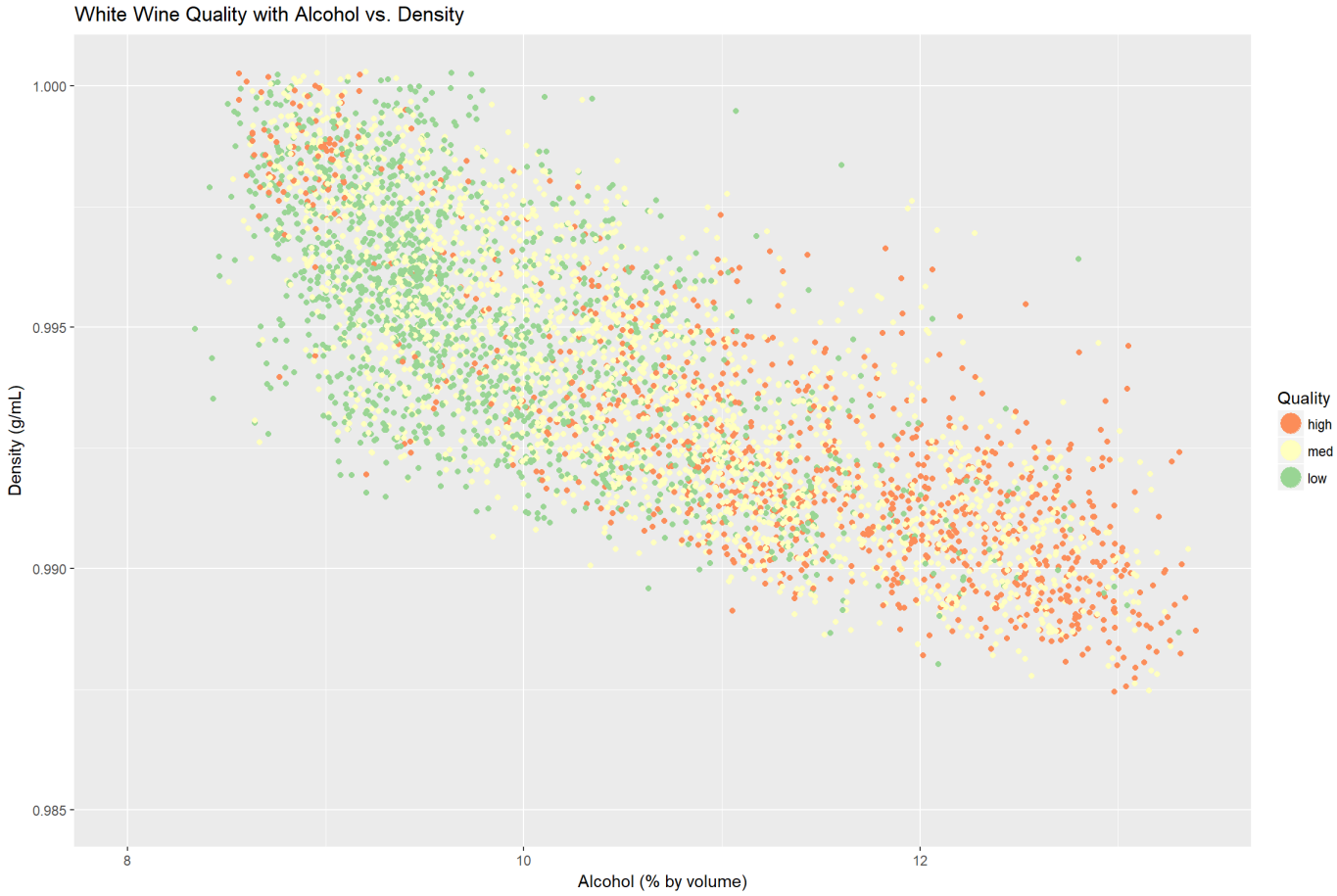

high
med
low
alcohol
density
residual.sugar

Multivariate Plots

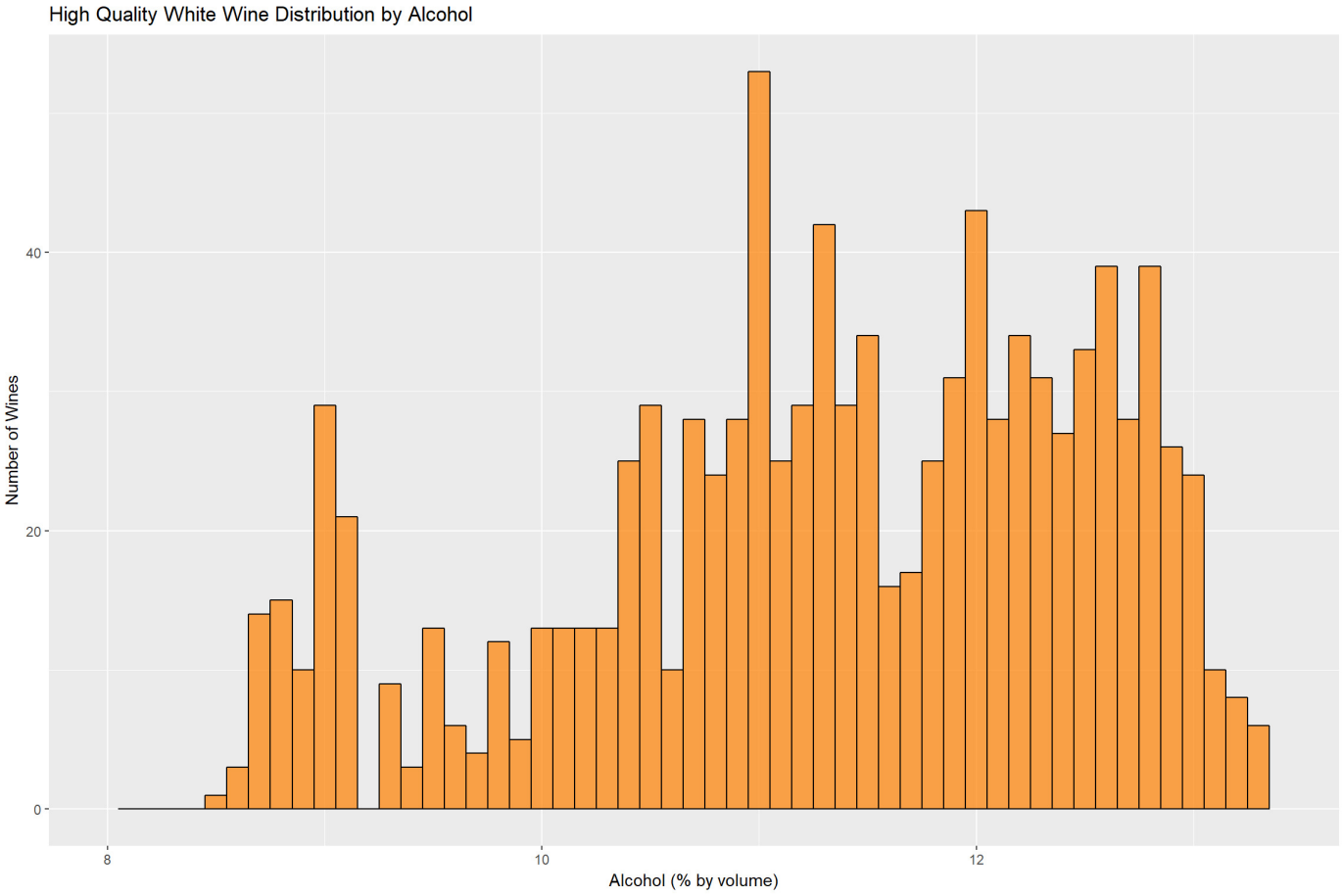
Let's go back to the correlation matrix and explore the two relationships that exhibited the highest correlation: Alcohol vs. Density (-0.78), and Residual Sugar vs. Density (0.84). But this time we'll add quality as the 3rd variable.

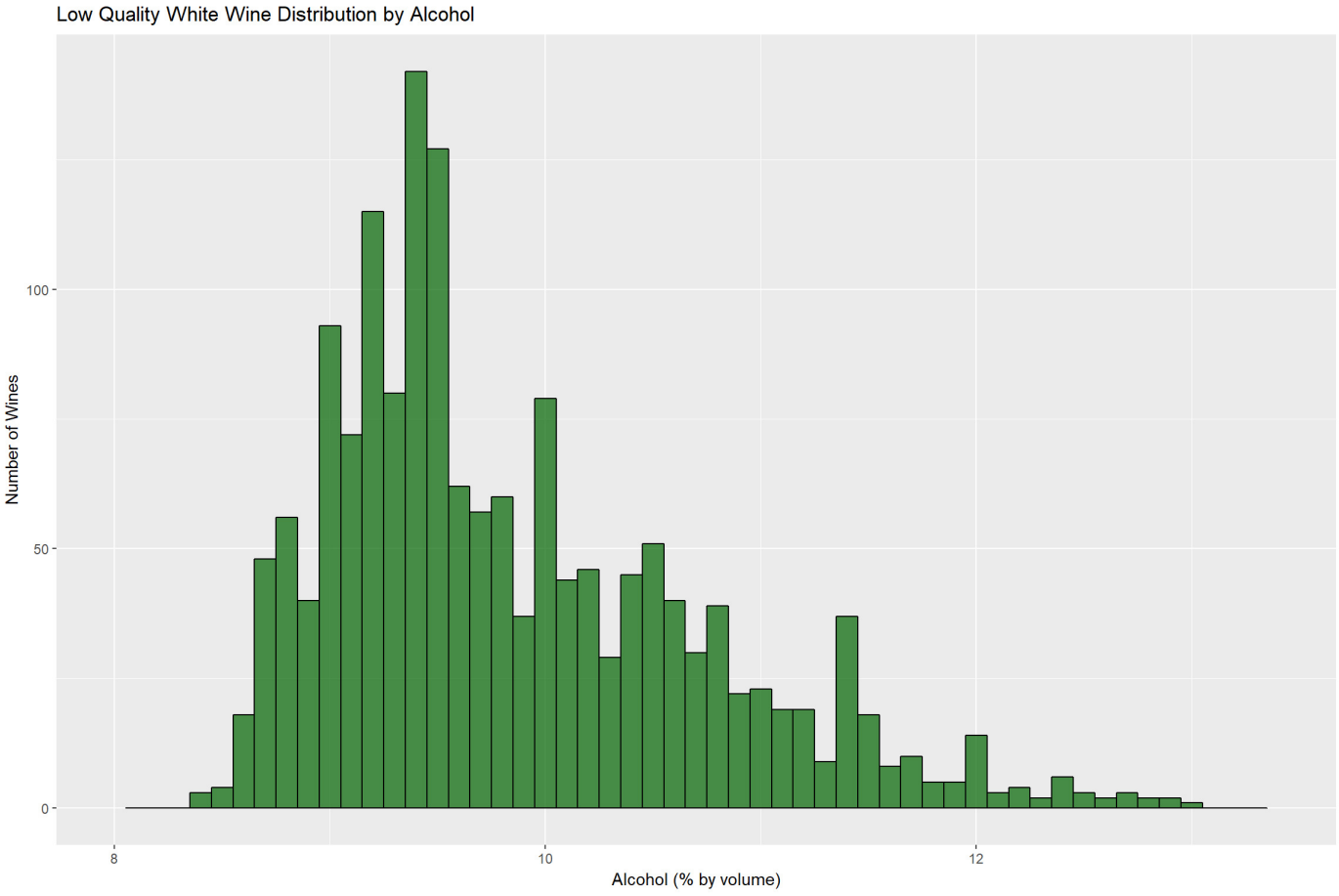
White Wine Quality with Alcohol vs. Density (-0.78)

The following graph is a scatterplot of Alcohol vs. Density. I notice a negative linear trend which is expected given the negative correlation. With respect to quality I notice a clustering of high quality wine samples toward the higher alcohol volumes. I also find it interesting that there is a smaller clustering of high quality wine samples near and at the bottom of the alcohol volume spectrum. The low quality wine samples appear to be clustered from the middle to lower alcohol volumes - not so many toward the high volume portion.



The next two figures plot the high and low quality distributions of the Alcohol variable respectfully. The high quality histogram shows a multimodal distribution which could be bimodal or even tri-modal if there is such a concept. We can see within this profile the clustering at the lower volume levels as noted before. Whereas the low quality distribution is much more normal with a small positive skew.





White Wine Quality with Residual Sugar vs. Density (0.84)

The following graph is a scatterplot of Alcohol vs. Density. I notice a positive linear trend which is expected given the positive correlation. In this plot the distribution of low and high quality across Residual Sugar is fairly even, but I notice a split of quality where generally high quality is on the upper half of the scatterplot and low quality on the lower half. I further add best fit linear model lines on the graph per quality. I notice at lower residual sugar levels the density is more of a predictor of high quality wine than it is at the higher residual sugar levels.

A Linear Model to Predict Quality

Here I will develop a multiple regression model utilizing a simple linear model.

Quality = B0 + B1(var1) + B2(var2) + … BN(varN) + k

Where quality is the dependent variable and vars are the various attributes of the white wine within this study (independent variables). B0 is the constant or intercept term. B1, B2, etc. are the coefficients or slope parameters. k is the error parameter.[6]

This exercise is an attempt to predict quality based on all of the attributes provided within the given white wine dataset.

This regression analysis will use a stepwise selection of variables by backwards elimination where all candidate variables are considered at first then are eliminated one at a time. After each iteration the model is assessed based on the properties to indicate the quality of the model. The model will try to estimate the coefficients of the variables in the linear equation such that the overall errors of the quality estimation could be minimized.

Below is the results from the first iteration of running the model with all the variables.

As we look at the summary of the linear model fit. We find a collection of estimates of the coefficients of the formula. Each of these values is estimated with some degree of confidence. We have a probability statement of how well we can trust the coefficients. The estimates with 3 stars indicate the probability of rejecting value is low, therefore we keep this value in the model. However there are other probability values that are higher (2 stars or no stars) which indicate a probability these values are wrong and therefore we will want to remove these variables from the model.

We also have an r-squared measure which tells us how well the model explains the variation in the data which is not random. So this model explains around 28% of errors or residuals from the linear model - which tells me this is a very poor model of wine quality.

```
##
## Call:
## lm(formula = quality ~ fixed.acidity + volatile.acidity + citric.acid +
##     residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + pH + sulphates + alcohol, data = white_wine)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8348 -0.4934 -0.0379  0.4637  3.1143
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.502e+02  1.880e+01   7.987 1.71e-15 ***
## fixed.acidity         6.552e-02  2.087e-02   3.139  0.00171 **
## volatile.acidity     -1.863e+00  1.138e-01 -16.373  < 2e-16 ***
## citric.acid           2.209e-02  9.577e-02   0.231  0.81759
## residual.sugar        8.148e-02  7.527e-03  10.825  < 2e-16 ***
## chlorides            -2.473e-01  5.465e-01  -0.452  0.65097
## free.sulfur.dioxide   3.733e-03  8.441e-04   4.422 9.99e-06 ***
## total.sulfur.dioxide -2.857e-04  3.781e-04  -0.756  0.44979
## density              -1.503e+02  1.907e+01  -7.879 4.04e-15 ***
## pH                    6.863e-01  1.054e-01   6.513 8.10e-11 ***
## sulphates             6.315e-01  1.004e-01   6.291 3.44e-10 ***
## alcohol               1.935e-01  2.422e-02   7.988 1.70e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7514 on 4886 degrees of freedom
## Multiple R-squared:  0.2819, Adjusted R-squared:  0.2803
## F-statistic: 174.3 on 11 and 4886 DF,  p-value: < 2.2e-16
```

So now we're going to eliminate the variable we trust least, or the variable with the highest probability that it is wrong. So we will eliminate citric acid and run the model again. And we notice there is no change to r-squared.

```
##
## Call:
## lm(formula = quality ~ fixed.acidity + volatile.acidity + residual.sugar +
##     chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + pH + sulphates + alcohol, data = white_wine)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8388 -0.4907 -0.0370  0.4665  3.1153
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.499e+02  1.876e+01   7.991 1.66e-15 ***
## fixed.acidity         6.611e-02  2.071e-02   3.192  0.00142 **
## volatile.acidity     -1.868e+00  1.121e-01 -16.668  < 2e-16 ***
## residual.sugar        8.140e-02  7.519e-03  10.827  < 2e-16 ***
## chlorides            -2.338e-01  5.434e-01  -0.430  0.66701
## free.sulfur.dioxide   3.740e-03  8.435e-04   4.434 9.46e-06 ***
## total.sulfur.dioxide -2.822e-04  3.777e-04  -0.747  0.45501
## density              -1.500e+02  1.903e+01  -7.882 3.94e-15 ***
## pH                    6.843e-01  1.050e-01   6.517 7.90e-11 ***
## sulphates             6.324e-01  1.003e-01   6.306 3.12e-10 ***
## alcohol               1.940e-01  2.411e-02   8.046 1.06e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7513 on 4887 degrees of freedom
## Multiple R-squared:  0.2819, Adjusted R-squared:  0.2804
## F-statistic: 191.8 on 10 and 4887 DF,  p-value: < 2.2e-16
```

Next we'll eliminate all variables that do not have 3 stars. Run the model again and notice still the r-squared does not improve beyond the 28%.

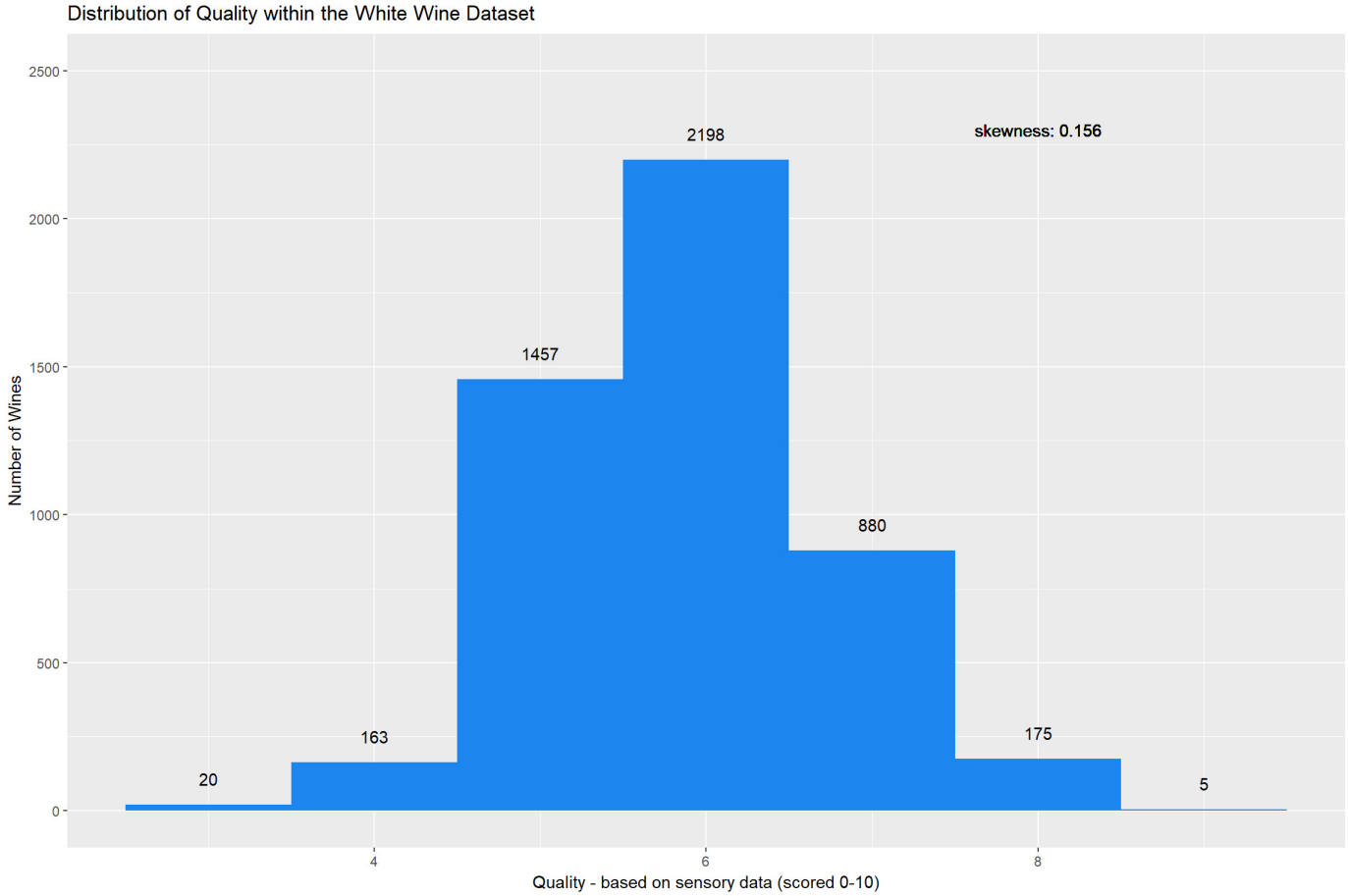
```
##
## Call:
## lm(formula = quality ~ volatile.acidity + residual.sugar + free.sulfur.dioxide +
##     density + pH + sulphates + alcohol, data = white_wine)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8107 -0.4999 -0.0375  0.4636  3.2180
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.112e+02  1.273e+01   8.734  < 2e-16 ***
## volatile.acidity    -1.940e+00  1.085e-01 -17.872  < 2e-16 ***
## residual.sugar       6.637e-02  5.358e-03  12.386  < 2e-16 ***
## free.sulfur.dioxide  3.283e-03  6.770e-04   4.849 1.28e-06 ***
## density             -1.103e+02  1.274e+01  -8.653  < 2e-16 ***
## pH                   4.619e-01  7.638e-02   6.046 1.59e-09 ***
## sulphates            5.708e-01  9.856e-02   5.791 7.42e-09 ***
## alcohol              2.438e-01  1.870e-02  13.035  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.752 on 4890 degrees of freedom
## Multiple R-squared:  0.2801, Adjusted R-squared:  0.2791
## F-statistic: 271.8 on 7 and 4890 DF,  p-value: < 2.2e-16
```

Next I elimated all but the three variables used in the Bivariate portion of the analysis - alcohol, density, and sugar. We find these three variable explain about 21% of the variance in quality. While 21% is not a high percentage it is most of the 28% including all other variables.

```
##
## Call:
## lm(formula = quality ~ residual.sugar + density + alcohol, data = white_wine)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5173 -0.5368 -0.0093  0.4739  3.1870
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     90.31292   12.37418   7.298 3.38e-13 ***
## residual.sugar   0.05332    0.00509  10.476  < 2e-16 ***
## density        -87.88589   12.31680  -7.135 1.11e-12 ***
## alcohol          0.24587    0.01825  13.474  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7873 on 4894 degrees of freedom
## Multiple R-squared:  0.2102, Adjusted R-squared:  0.2097
## F-statistic: 434.1 on 3 and 4894 DF,  p-value: < 2.2e-16
```

Final Plots and Summary

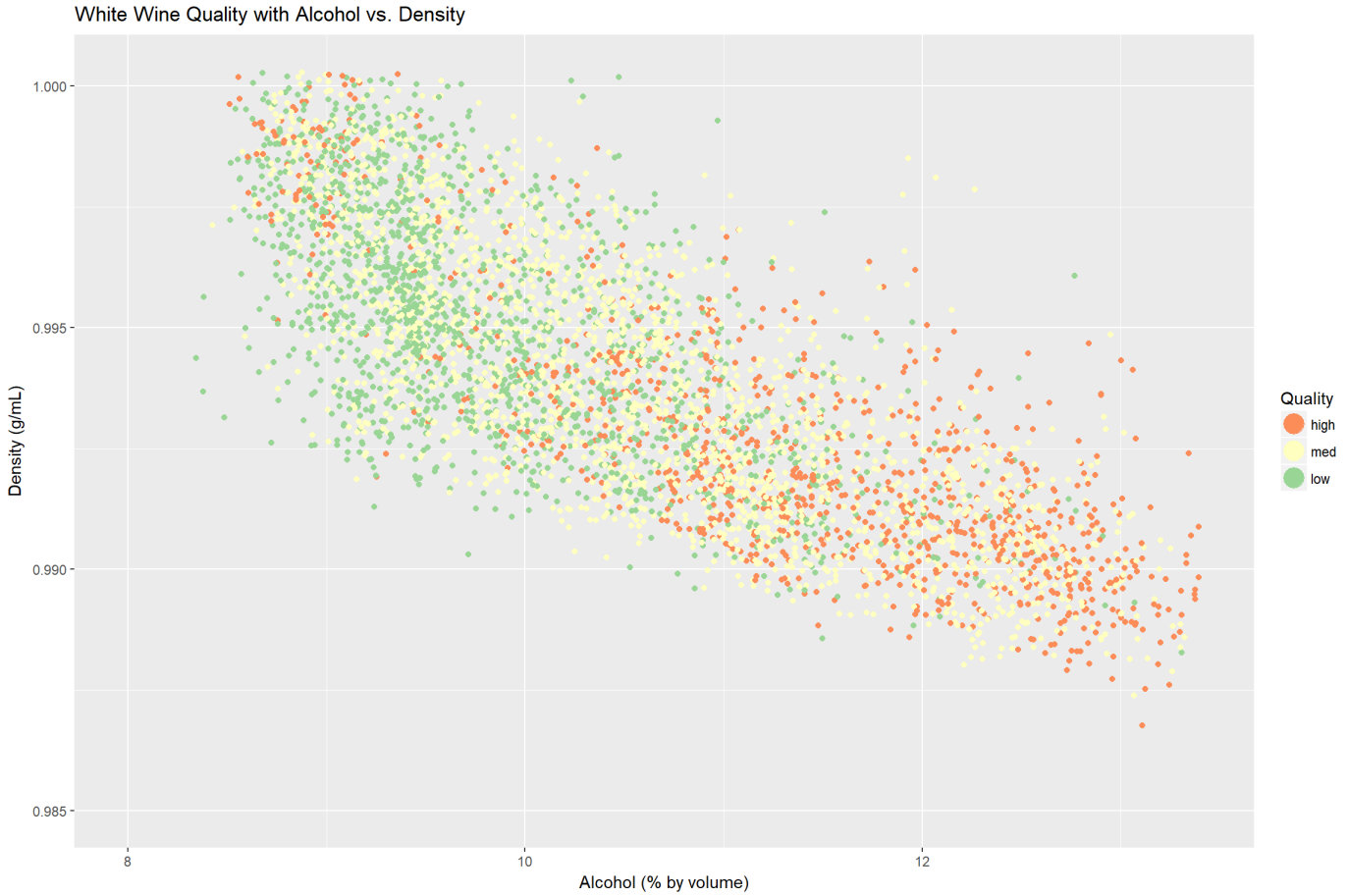
Plot One



Discription One

Quality is based on sensory data with scores between 0 and 10 is noramlly distributed with a quality score of 6 in the middle. There is a little skew toward more wines below 6 than above. This EDA analysis focuses on the attributes that explain the variability of Quality.

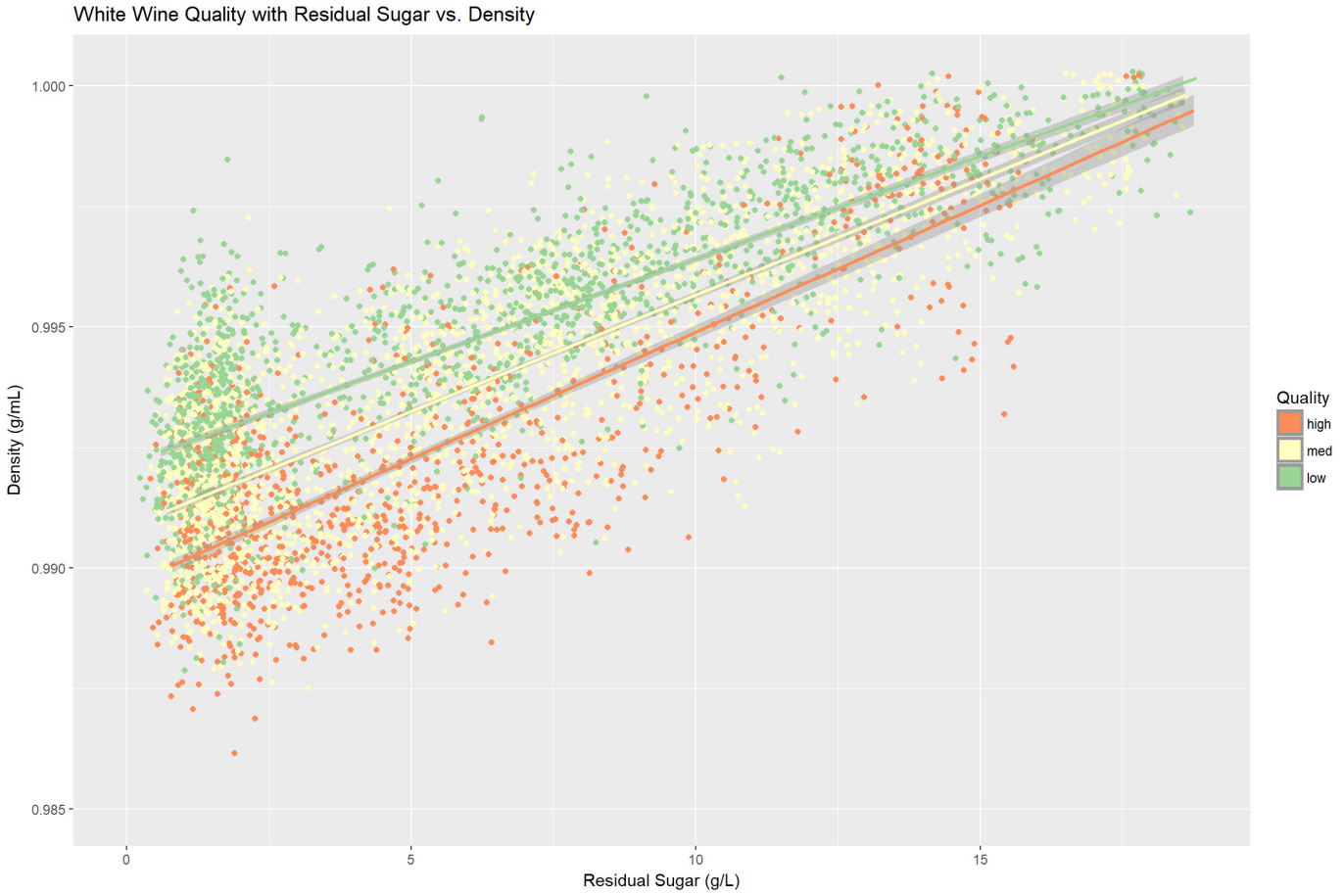
Plot Two



Discription Two

Alcohol is the most defining variable for high quality wines, as we can clustering of the high quality wines at the high end of alcohol content. I also note that there is a small cluster of high quality wines at the low end of alcohol as well.

Plot Three



Discription Three

High quality wine is also very highly correlated to density and sugar as well. Wines are considered high quality are those with the lowest density and sugar content.

Refection

The quality factor was the median of at least 3 expert evaluations. To me ‘quality’ was the key attribute in this dataset because I believe it directly relates to price or value of the wine. As I explored the data there were several times I wanted to know more about the breakdown of the expert evaluations. For example I’d like to know more about the evaluators of the high quality wines that were clustered at the low end of the Alcohol continuum - wondering if they are the same clustered group that chose high quality wine at the high end of the Residual Sugar continuum.

I do realize that I transformed the residual sugar scales to log 10 but never used that transformed scale in the linear model. With the linear model r-squared at 28% I didn’t really find the motivation to add more complexity into the model - ever though the added complexity would have been slight.

The Cohen’s d analysis I just kind of stumbled on. It made intuitive sense to me that comparing the means of the low versus high quality attributes would provide some insight. Also in order to compare the attributes together on a level basis I knew I had to normalize the differences. So I just kind of backed into the Cohen’s d analysis. Kind of interesting and I something maybe to follow up with for any additional analysis.

I was disappointed about the linear model only yielding a 28% r-squared factor, perhaps an additional set of attributes would be needed to get value that will better explain quality.

References:

[1] For more details, consult: http://www.vinhoverde.pt/en/ (http://www.vinhoverde.pt/en/) or the reference [Cortez et al., 2009]. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.).

[2] Wine Analysing by the Cleebourg Vineyards - http://www.cave-cleebourg.com/en/wine-analysing/ (http://www.cave-cleebourg.com/en/wine-analysing/)

[3] https://winobrothers.com/2011/10/11/sulfur-dioxide-so2-in-wine/ (https://winobrothers.com/2011/10/11/sulfur-dioxide-so2-in-wine/)

[4] http://morewinemaking.com/public/pdf/so2.pdf (http://morewinemaking.com/public/pdf/so2.pdf)

[5] http://www.oiv.int/public/medias/2604/oiv-ma-d1-03.pdf (http://www.oiv.int/public/medias/2604/oiv-ma-d1-03.pdf)

[6] http://statlab.stat.yale.edu/workshops/IntroRegression/StatLab-IntroRegressionFa08.pdf (http://statlab.stat.yale.edu/workshops/IntroRegression/StatLab-IntroRegressionFa08.pdf)

[7] http://www.bing.com/videos/search?q=r+multiple+regression+example (http://www.bing.com/videos/search?q=r+multiple+regression+example)&&view=detail&mid=ADC47975E0D7F6C52970ADC47975E0D7F6C52970&FORM=VRDGAR

[8] http://www.mathsisfun.com/data/skewness.html (http://www.mathsisfun.com/data/skewness.html)

[9] http://rpsychologist.com/d3/cohend/ (http://rpsychologist.com/d3/cohend/)

[10] http://www.bwgriffin.com/gsu/courses/edur9131/content/cohen_d_Denis.pdf (http://www.bwgriffin.com/gsu/courses/edur9131/content/cohen_d_Denis.pdf)