Exploring the properties of red wine

Randy Tilson

04/01/2022

Red Wine Properties Data Exploration: This dataset explores the properties of red wine as determined by the given variables displayed below

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

Total Data Entries: As shown below

```
## [1] 19188
```

Thoughts regarding variables: Given the variables, it is likely that "quality" is the best benchmark to measure against for correlations from within other variables.

Data Summary

```
##  fixed.acidity   volatile.acidity  citric.acid    residual.sugar  
##  Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900  
##  1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900  
##  Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200  
##  Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539  
##  3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600  
##  Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500  
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide    density      
##  Min.   :0.01200   Min.   : 1.00       Min.   :  6.00       Min.   :0.9901  
##  1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00       1st Qu.:0.9956  
##  Median :0.07900   Median :14.00       Median : 38.00       Median :0.9968  
##  Mean   :0.08747   Mean   :15.87       Mean   : 46.47       Mean   :0.9967  
##  3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00       3rd Qu.:0.9978  
##  Max.   :0.61100   Max.   :72.00       Max.   :289.00       Max.   :1.0037  
##        pH          sulphates         alcohol         quality     
##  Min.   :2.740   Min.   :0.3300   Min.   : 8.40   Min.   :3.000  
##  1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50   1st Qu.:5.000  
##  Median :3.310   Median :0.6200   Median :10.20   Median :6.000  
##  Mean   :3.311   Mean   :0.6581   Mean   :10.42   Mean   :5.636  
##  3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10   3rd Qu.:6.000  
##  Max.   :4.010   Max.   :2.0000   Max.   :14.90   Max.   :8.000
```

Data summarized and grouped by quality rating

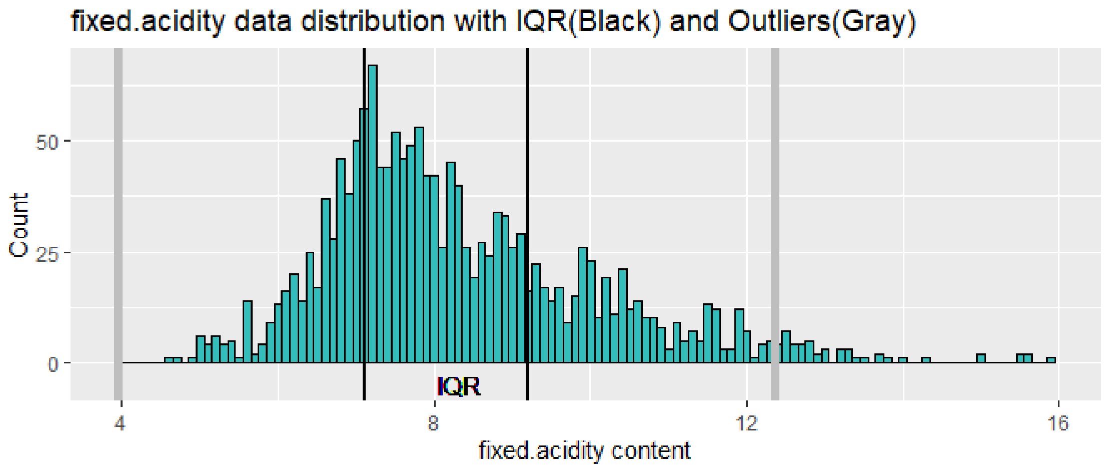
Descriptive Statistics AS GROUPED BY Quality Rating

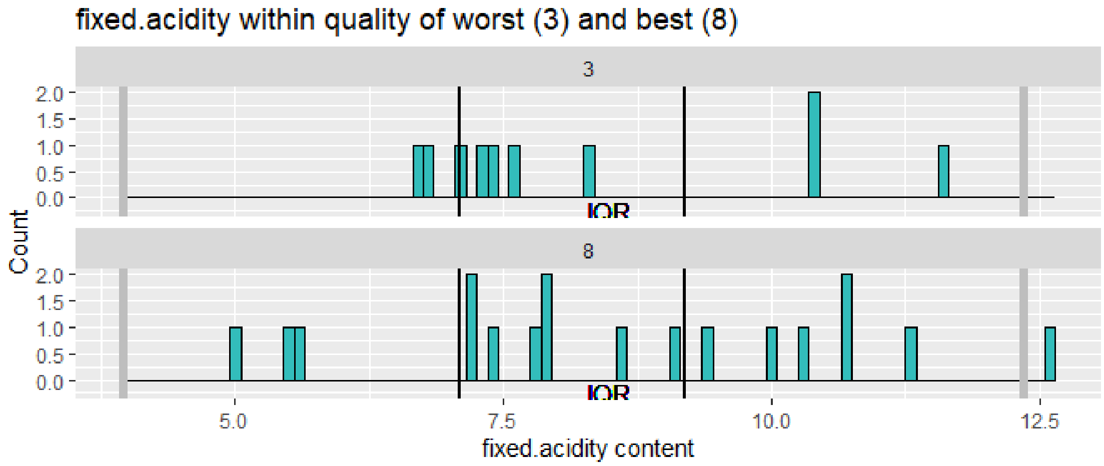
Variables As grouped by quality rating	**3**, N = 10	**4**, N = 53	**5**, N = 681	**6**, N = 638	**7**, N = 199	**8**, N = 18
fixed.acidity	7.50 (7.15, 9.88)	7.50 (6.80, 8.40)	7.80 (7.10, 8.90)	7.90 (7.00, 9.40)	8.80 (7.40, 10.10)	8.25 (7.25, 10.23)
volatile.acidity	0.84 (0.65, 1.01)	0.67 (0.53, 0.87)	0.58 (0.46, 0.67)	0.49 (0.38, 0.60)	0.37 (0.30, 0.48)	0.37 (0.33, 0.47)
citric.acid	0.04 (0.00, 0.33)	0.09 (0.03, 0.27)	0.23 (0.09, 0.36)	0.26 (0.09, 0.43)	0.40 (0.30, 0.49)	0.42 (0.30, 0.53)
residual.sugar	2.10 (1.88, 3.10)	2.10 (1.90, 2.80)	2.20 (1.90, 2.60)	2.20 (1.90, 2.50)	2.30 (2.00, 2.75)	2.10 (1.80, 2.60)
chlorides	0.090 (0.079, 0.143)	0.080 (0.067, 0.089)	0.081 (0.074, 0.094)	0.078 (0.068, 0.088)	0.073 (0.062, 0.087)	0.071 (0.062, 0.076)
free.sulfur.dioxide	6 (5, 14)	11 (6, 15)	15 (9, 23)	14 (8, 21)	11 (6, 18)	8 (6, 16)
total.sulfur.dioxide	15 (12, 42)	26 (14, 49)	47 (26, 84)	35 (23, 54)	27 (18, 43)	22 (16, 43)
density	0.9976 (0.9962, 0.9988)	0.9965 (0.9956, 0.9974)	0.9970 (0.9962, 0.9979)	0.9966 (0.9954, 0.9979)	0.9958 (0.9948, 0.9974)	0.9949 (0.9942, 0.9972)
pH	3.39 (3.31, 3.50)	3.37 (3.30, 3.50)	3.30 (3.20, 3.40)	3.32 (3.22, 3.41)	3.28 (3.20, 3.38)	3.23 (3.16, 3.35)
sulphates	0.54 (0.51, 0.62)	0.56 (0.49, 0.60)	0.58 (0.53, 0.66)	0.64 (0.58, 0.75)	0.74 (0.65, 0.83)	0.74 (0.69, 0.82)
alcohol	9.93 (9.72, 10.57)	10.00 (9.60, 11.00)	9.70 (9.40, 10.20)	10.50 (9.80, 11.30)	11.50 (10.80, 12.10)	12.15 (11.33, 12.88)

Prior to plotting any variables, the data from within this table shows a few correlations already. It appears as if the variables of alcohol, sulphates, chlorides, citric.acid, and volatile.acidity are all are somewhat correlated to the quality rating.

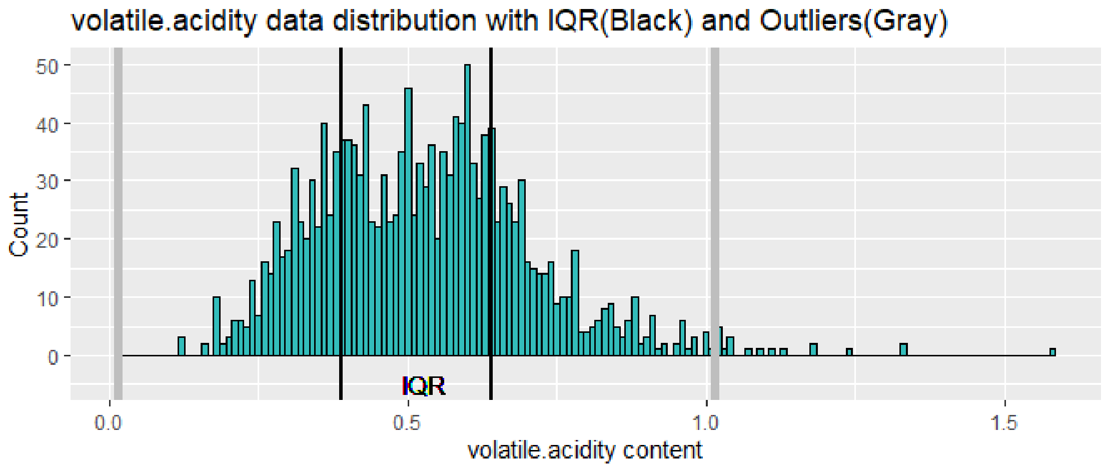
Univariate Plots Section

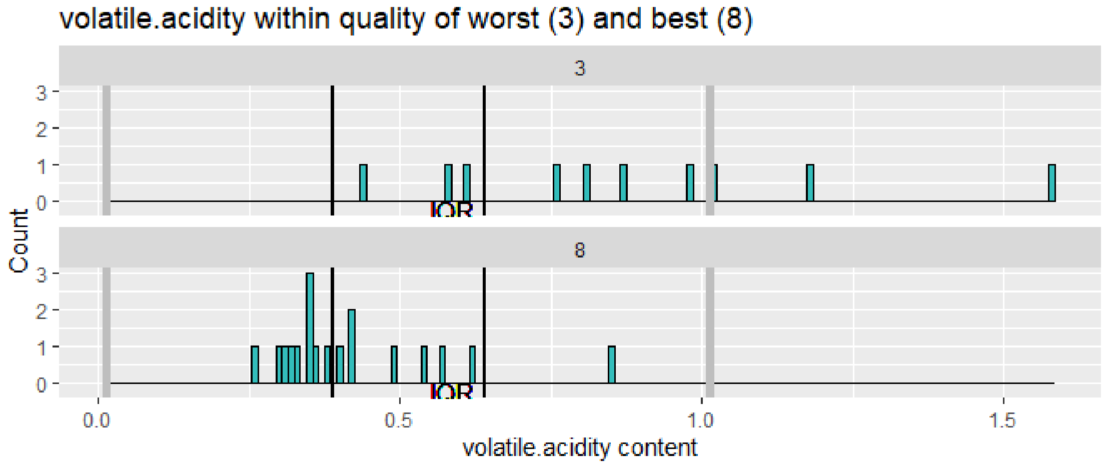
With a lack of domain knowledge and a need to better understand the distributions, all variables will be charted in combination with their IQR. Also, standardized outlier markers will be put in place for better visualization of the distribution. In addition, the quality variable will be faceted together with each individual variable.



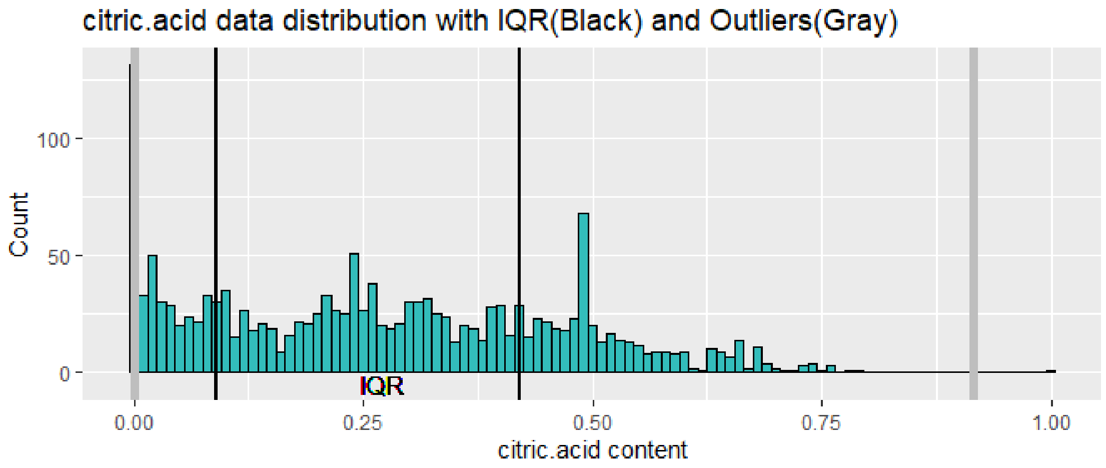


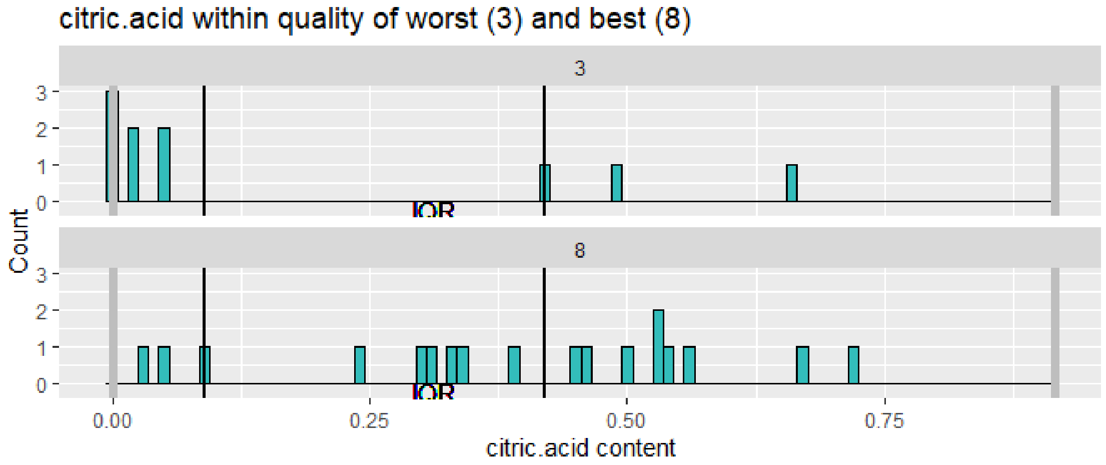
fixed.acidity: Normal / slightly positive skewed distribution with outliers within the right tail. In looking at the quality facet, I do not see that much of a correlation between this variable and quality.





volatile.acidity: Normal / slightly positive skewed distribution. Some outlier data within the right tail. It is possible that quality is affected by the content level of this variable with lower levels increasing the quality.

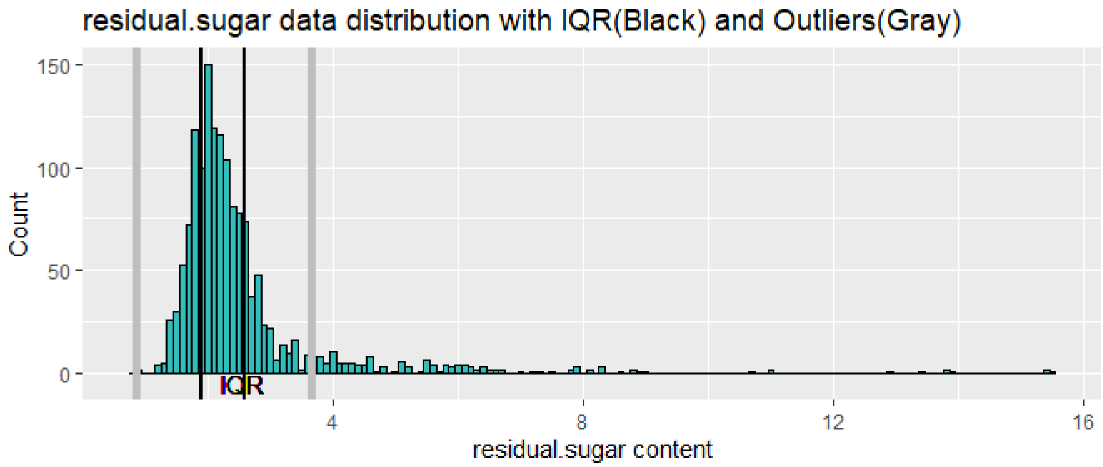


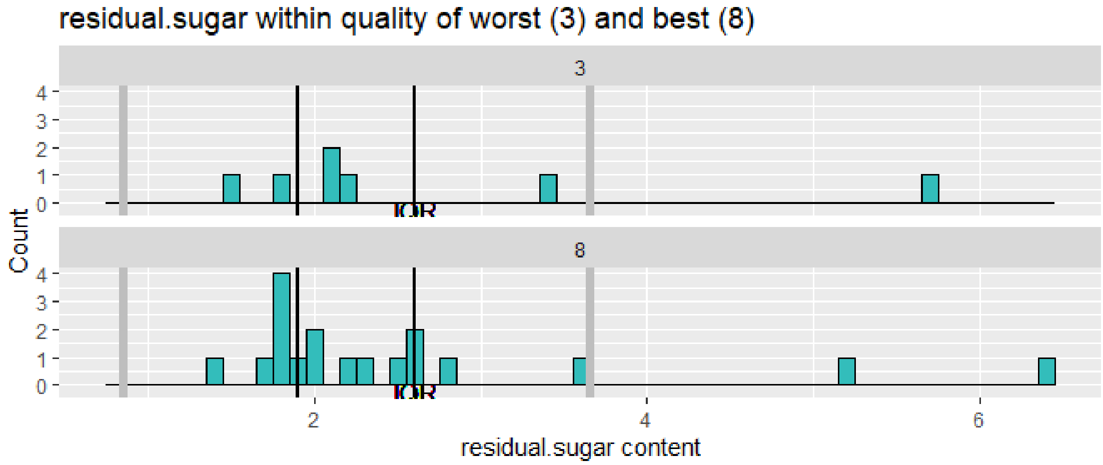


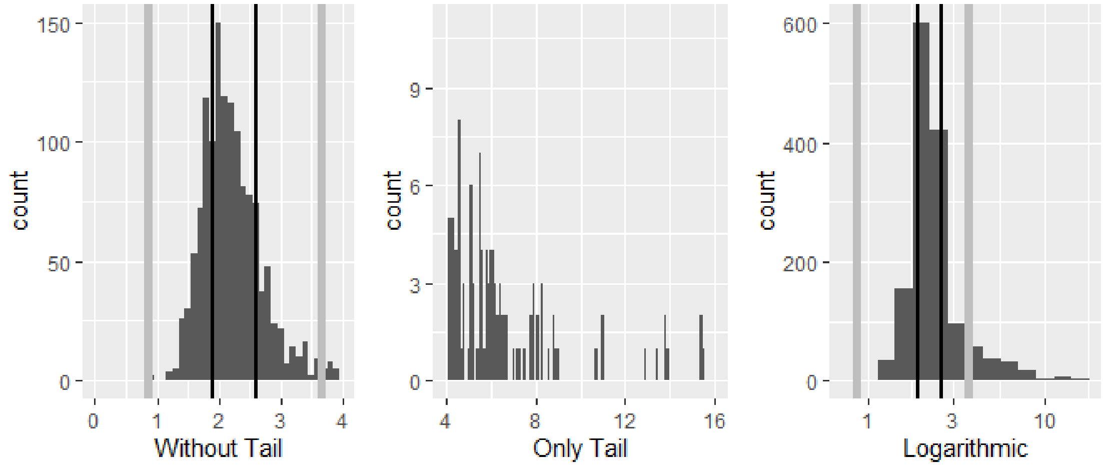
citric.acid: The first thing that I notice is the first bar in the histogram, why are there so many zero values? I am not sure if this is possible for red wine. The structure is multi-modal, which so far is a unique quality. With it being multi-modal, there is a wide range of possible values without becoming outliers. I do want to disect the rest of the data and look for other zero values.

```
##        fixed.acidity     volatile.acidity          citric.acid
##                    0                    0                    8
##       residual.sugar            chlorides  free.sulfur.dioxide
##                    0                    0                    0
## total.sulfur.dioxide              density                   pH
##                    0                    0                    0
##            sulphates              alcohol              quality
##                    0                    0                    0
```

In looking at the dataframe, 8% of all citric acid values are zero values, Whereas zero values do not exist within other variables. This does lead me to believe that this is these values should be removed given the uncertainty and uncommon nature of this.





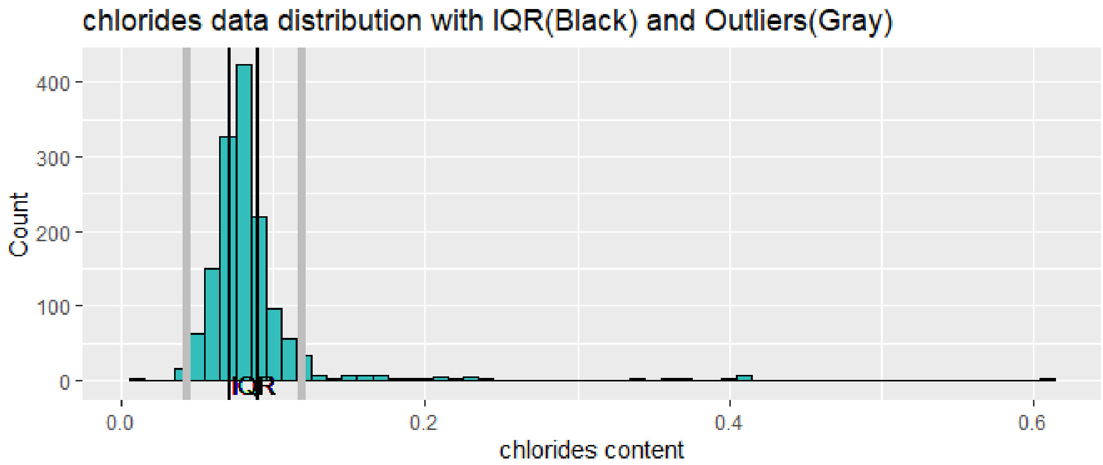


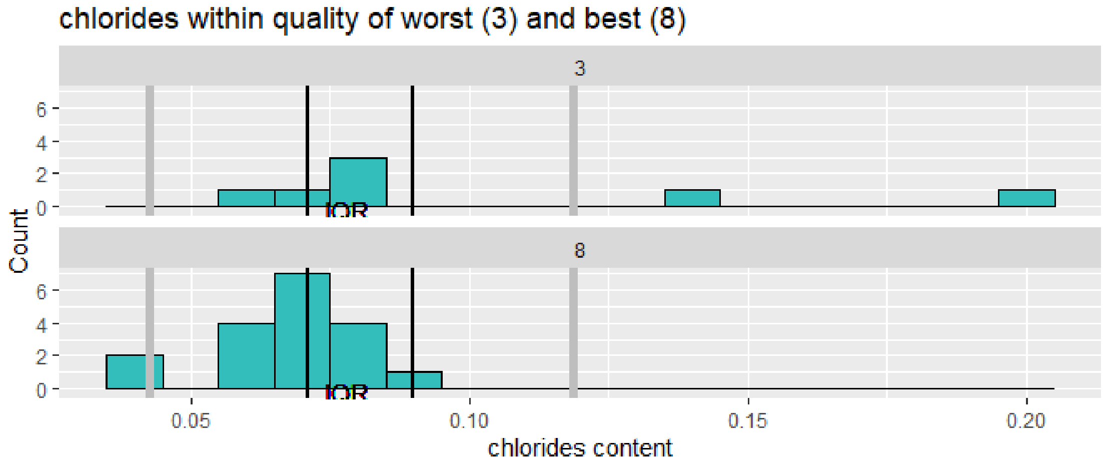
residual.sugar: Normal / slightly positive skewed distribution with one sided outliers. The amount of outliers within the tail of this variable show a possible unique characteristic that would limit their discarding. In re-scaling and looking at the tail sections, it does appear that a fair amount of data exists past the technical outlier that lies just shy of 4. The correlation with quality seems to be lacking.

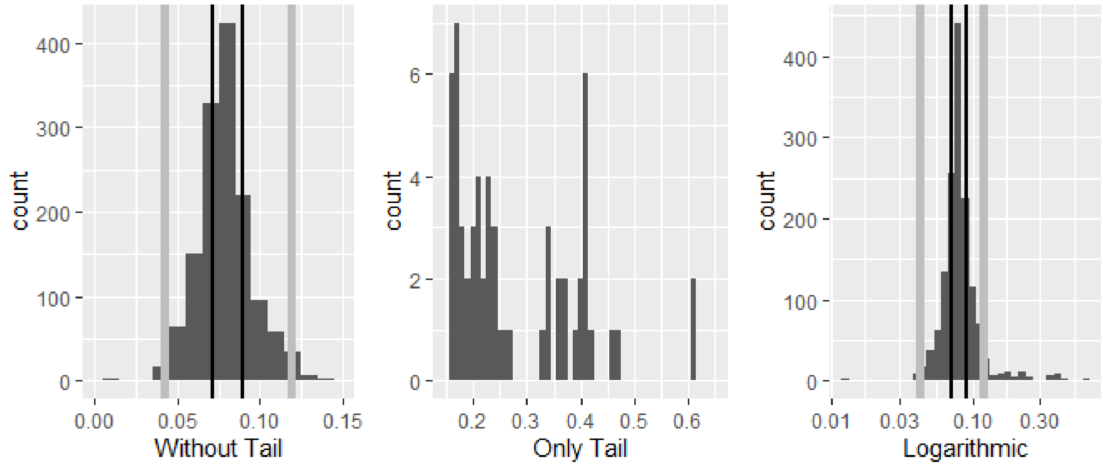
```
## Percent of original residual.sugar values that resides in the tail.
## Ordered from lowest to highest quality rating

## [1] 14.28571
## [1] 18.60465
## [1] 9.775641
## [1] 8.561644
## [1] 16.75393
## [1] 16.66667
```

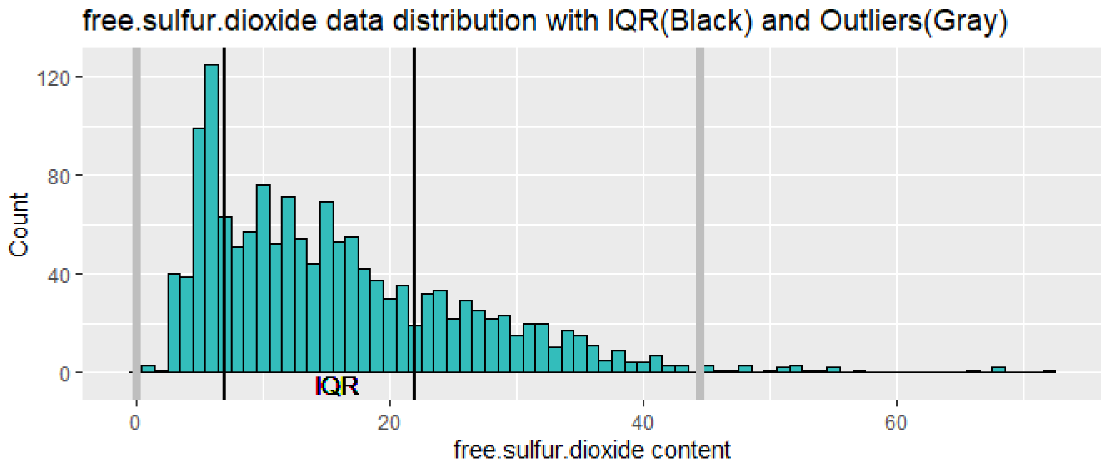
In performing a calculation on the percent of each quality that resides within the tail, it appears as if the mid grade wines have the lowest amount of data within the tail. This would infer that both the high grade and low grade wines possibly hold a higher content of this variable. This could also be a result of the law of large number as well, given that the mid grade wines have a much higher total count to where more data is available. This can be looked at in further detail within multivariate observations.

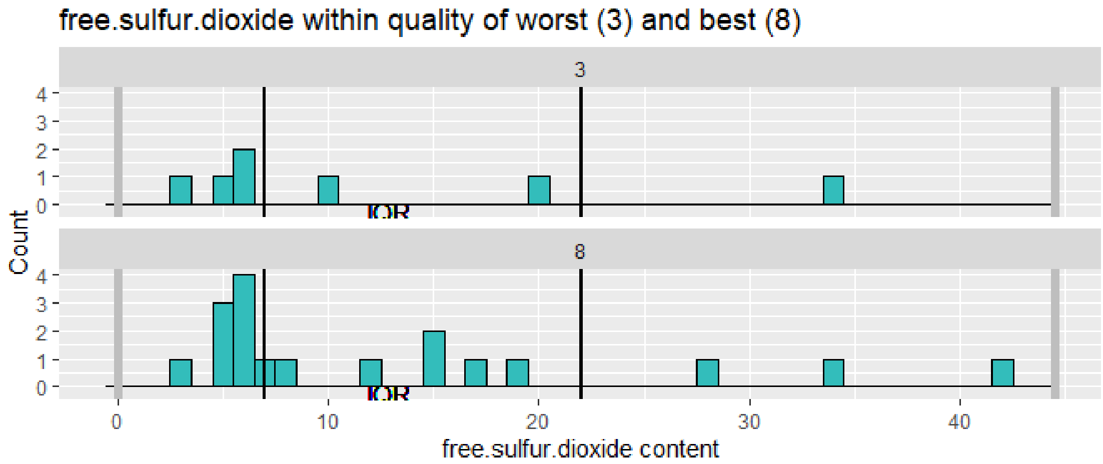


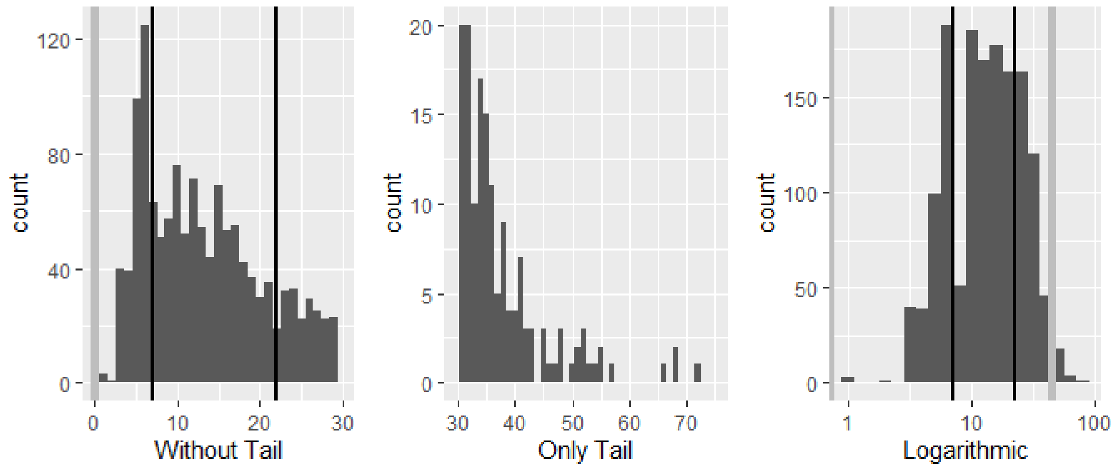




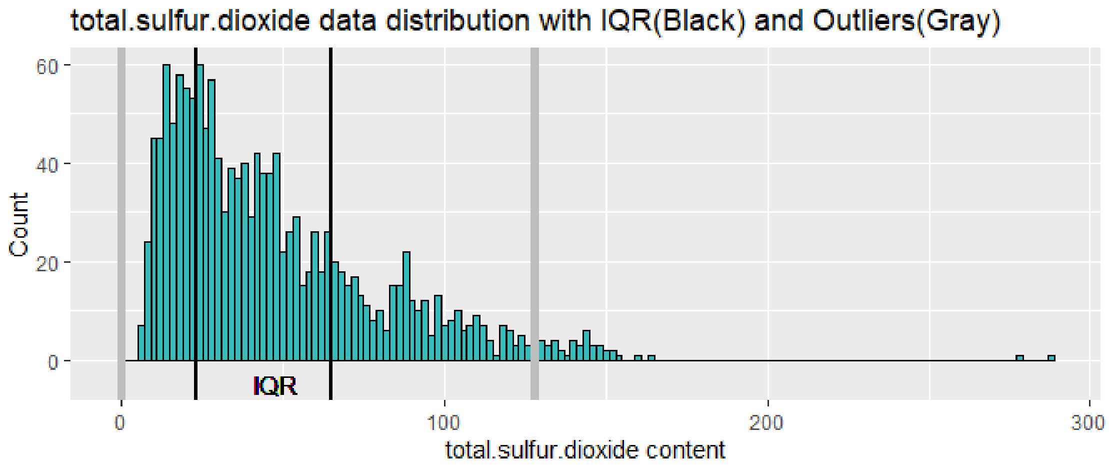
chlorides: This distribution is similiar to sugar with a very normal distribution followed by a very long tail to the right. Of difference though would be the possible correlation with quaility. Because of the tail, a logarithmic view was created, however the view seems inconclusive other than showing a small distribution past the .30 range. It would appear as if lower chlorides may lead to higher quality wines

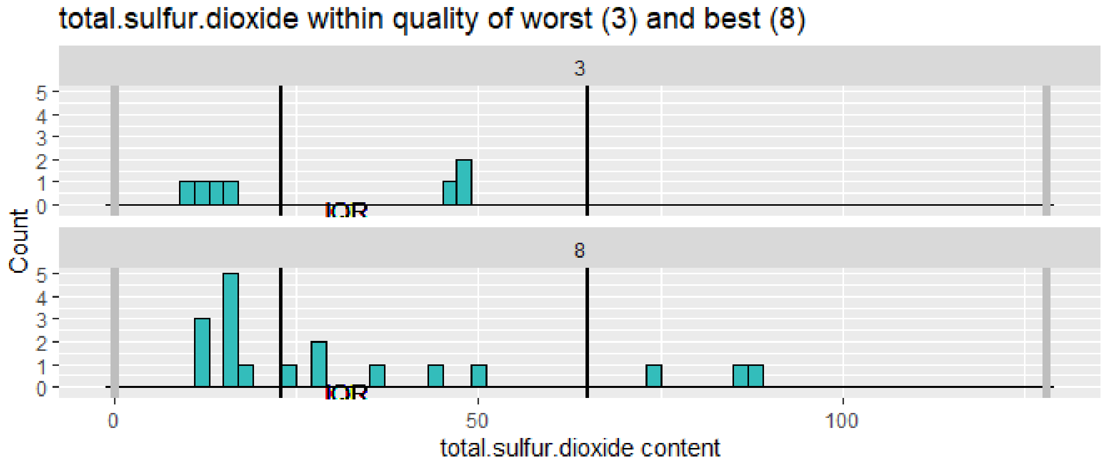


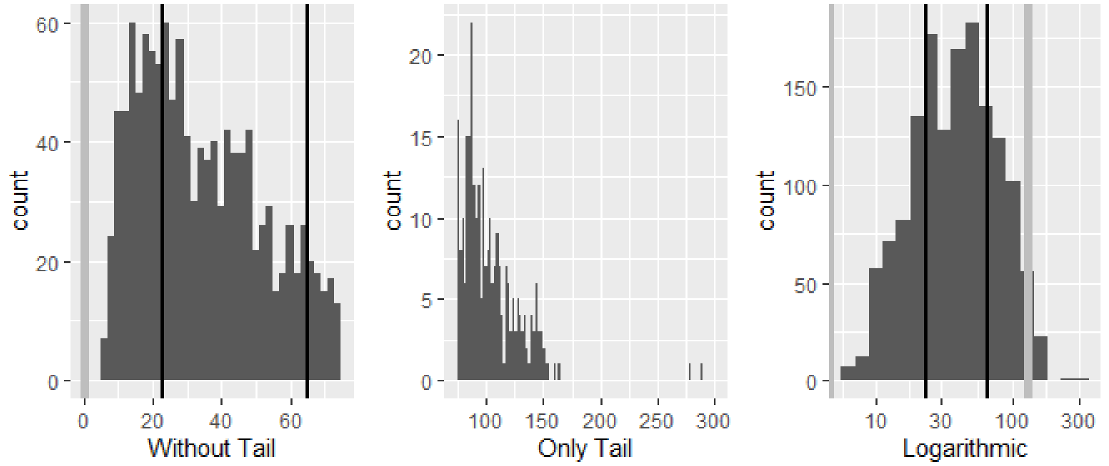




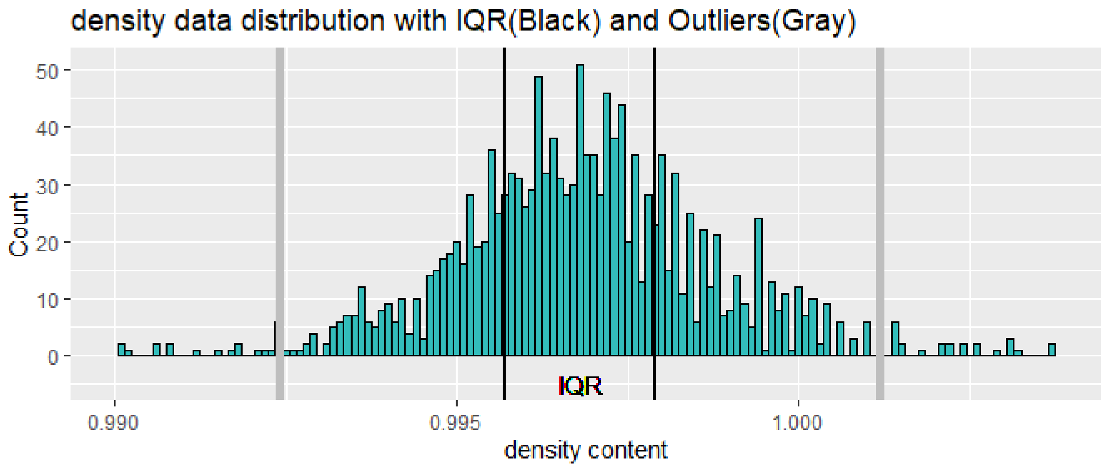
free.sulfur.dioxide: Positive skew with a large amount of data falling outside of Q1. Log scale shows a more even distribution. Also, it is a pretty even distribution amongst the low and high quality which does not lead me towards correlation with quality.

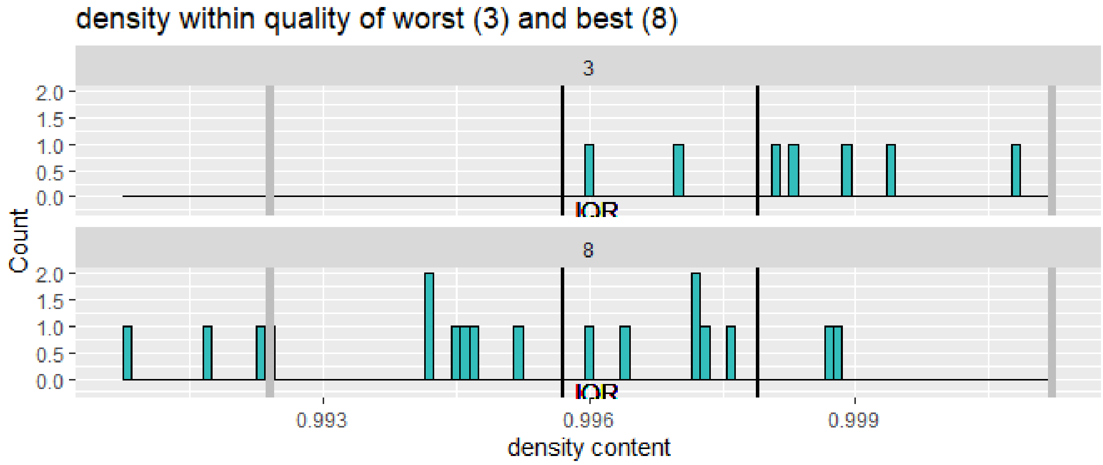




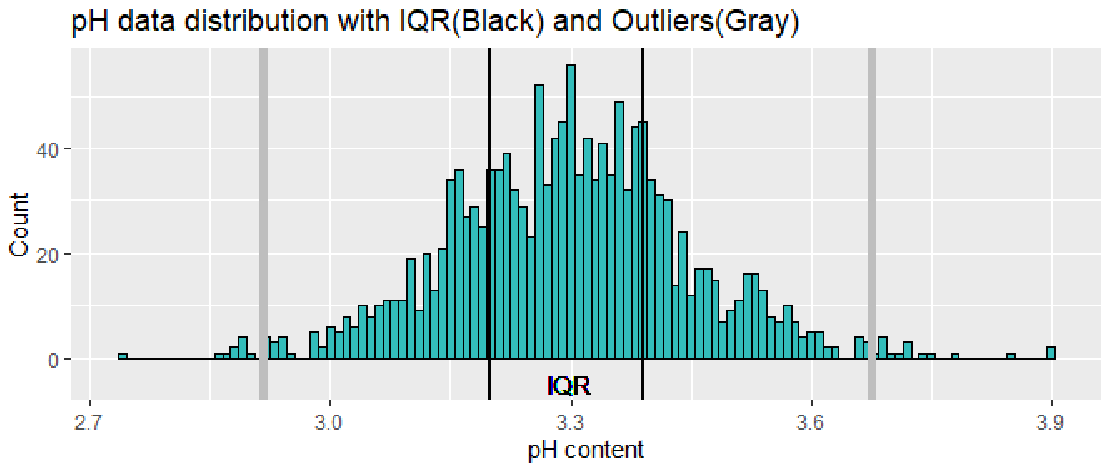


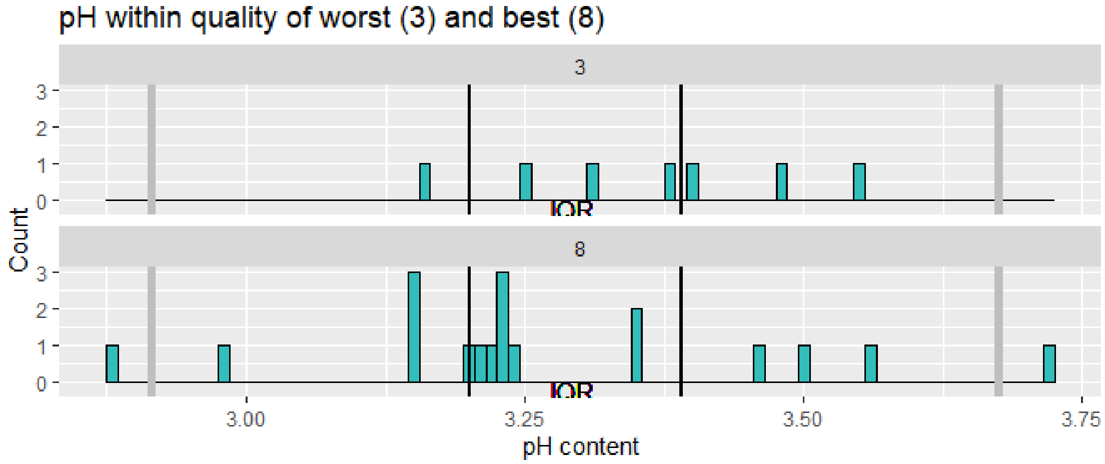
total.sulfur.dioxide: Similar distribution to that of free.sulfur.dioxide. This would make sense given that they are measurements of similar properties. Rescaling it logarithmically shows a more even distribution as well. Its possible that higher values lead to higher quality, but it is certainly not a strong correlation so far.



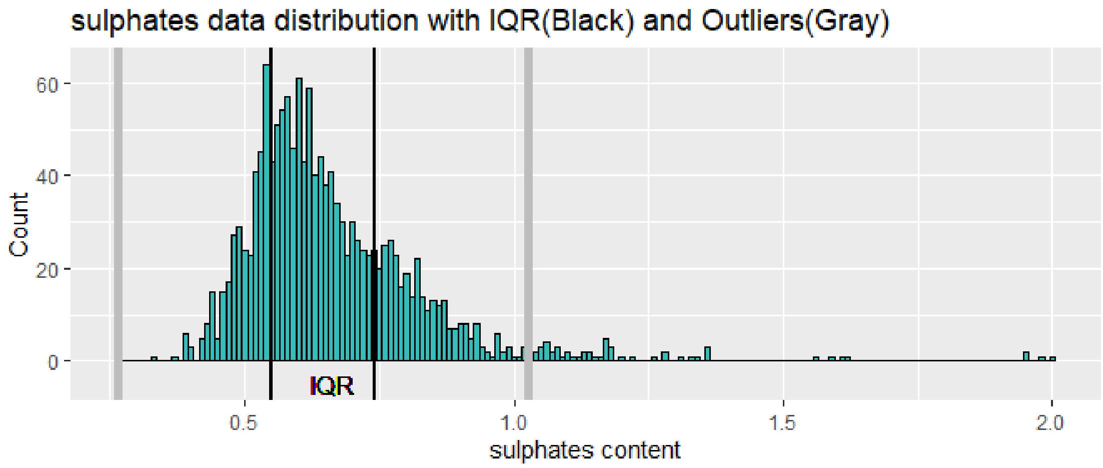


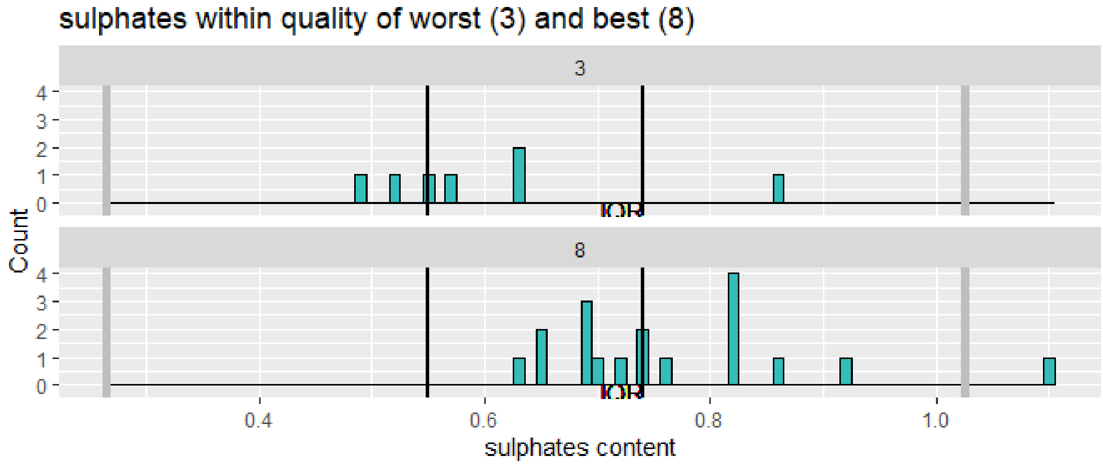
density: Very normal distribution with outliers in the left and right tail. Given the normalized distribution, I am not going to look into any tail features. It appears as if a lower density may lead to a higher quality wine.



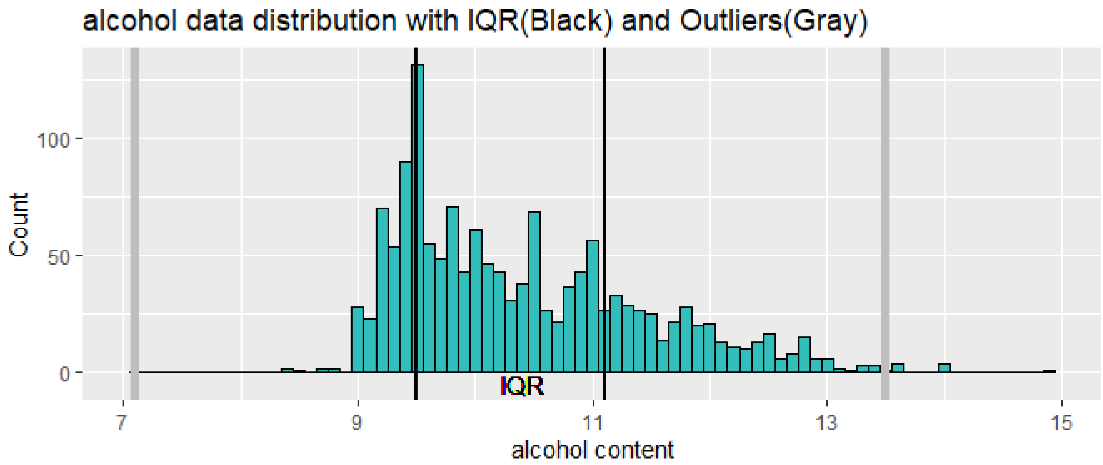


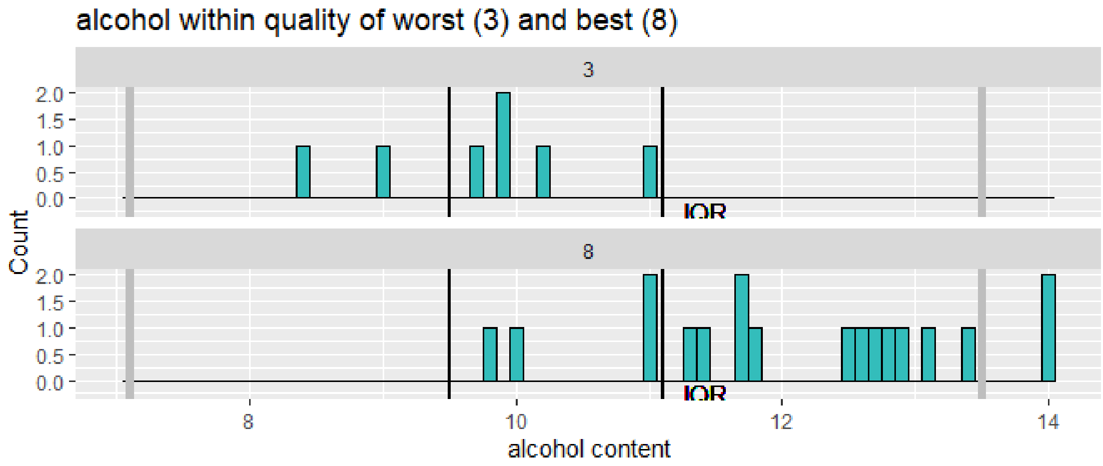
pH: Normal distribution similiar to that of density. A possible small correlation may exist though, with lower pH equaling higher quality.



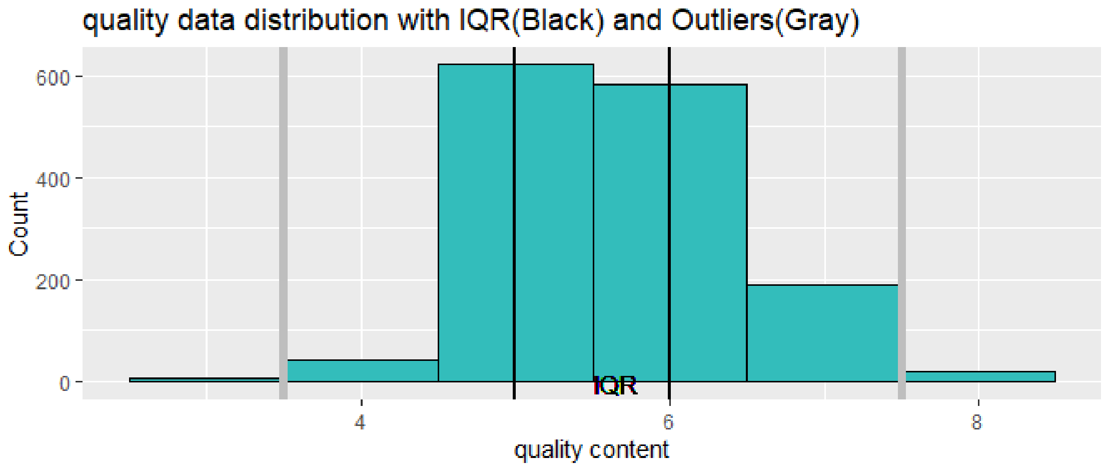


sulphates: Normal distribution with outlier data within the right tail. There is also a possible correlation with higher pH equaling higher quality.





alcohol: Long positively skewed tail. There is not much in the way of outlier data though so at this time I will not look into the tail. There does however seem to be a very large correlation between alcohol content and quality rating. Higher alcohol seems to create higher quality.



quality: Normal distribution with the majority of ratings being that of 5 and 6.

Univariate Analysis

Dataset structure:

This dataset has both normalized and positively skewed distributions. Within the skewed distributions, all outliers lie within the right tail. In some cases, the outliers appear to be more than outliers, this may point to conditions that are rare but very possible.

Main feature of interest:

The main feature of interest is the quality variable and any given correlations from within the remaining variables of the dataset.

Dataset features to help investigate the main feature of interest:

The numerous apparent correlations will be very helpful to examine, particularly within the alcohol content.

New variables:

As of now, new variables have not been created. This may change as I proceed into bivariate and multivariate, however at this time the variables seem sufficient.

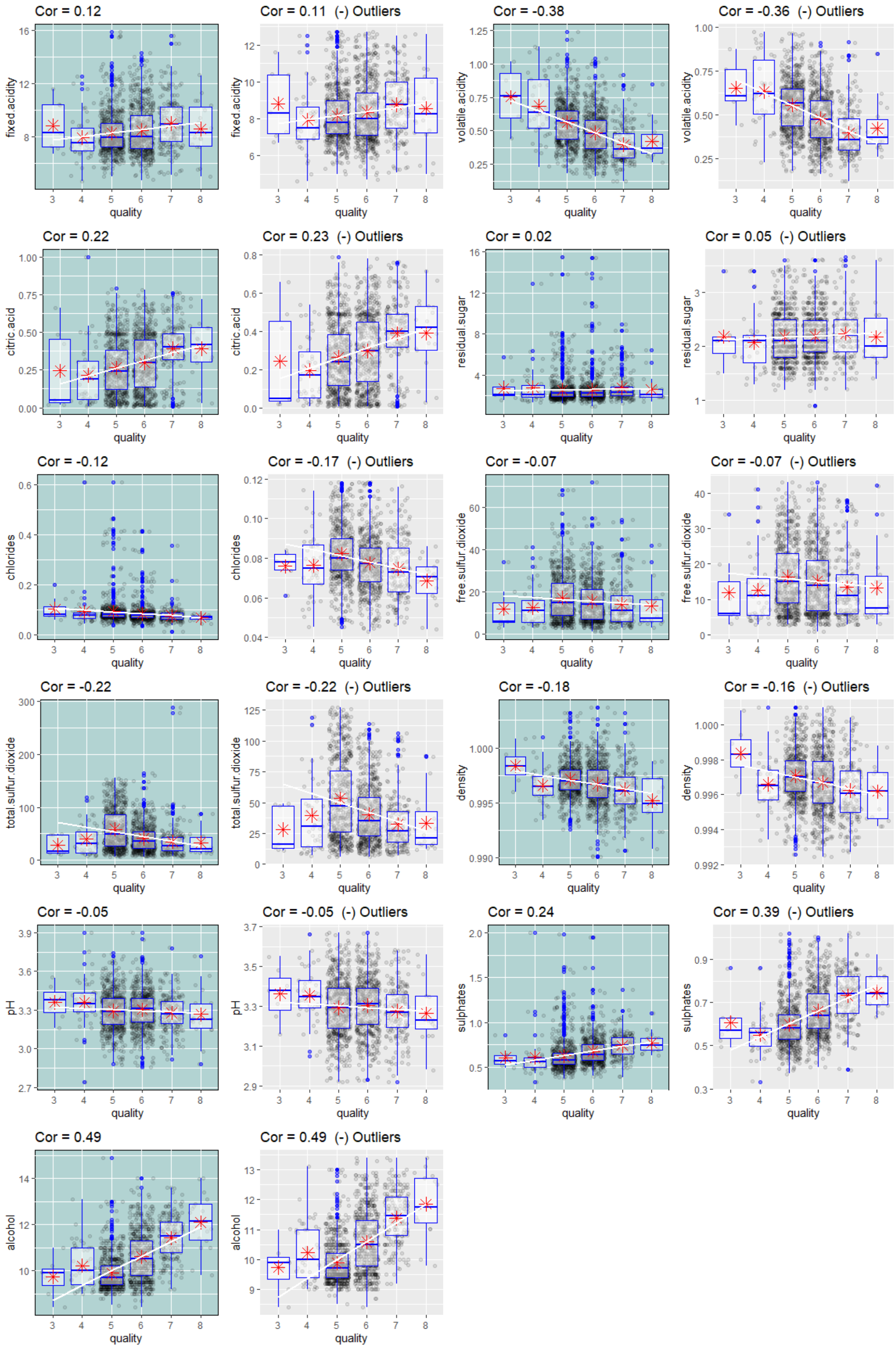
Unusual features and data cleaning:

First of all, I subsetted the dataframe to remove the x variable, which I determined not necessary for my analysis. Then The zero values within citric acid grabbed my attention. Because of this, the dataframe was examined further to ensure this anomaly was isolated within this one specific variable. Given that they were, these values were removed to ensure data completeness. Also, in separating out the numerous outlier values within residual sugar, I was surprised to see that the low and high grade wines hold a larger percent of the outlier data. As noted though, this could easily be due to a larger amount of data being available from other quality ratings, which could possibly lower their percentages to a more representative value. Lastly, it seems as if the data is either structured with a positive skew, or normalized distribution, with some of these variables looking very similar in structure.

In general, the data is and was fairly clean and ready to examine outside of the zero values from within citric.acid

Bivariate Plots Section

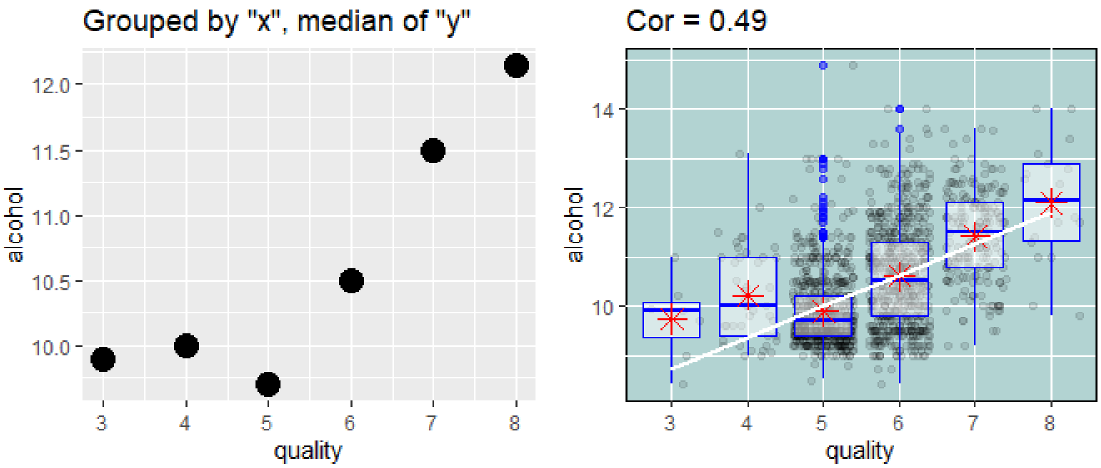
In moving forward, I would like to further examine the relationships between these variables. Based on my univariate analysis, I already have some assumptions towards possible correlations, specifically the correlation between alcohol and the quality rating. I would like to start with a full data overview of all variables as compared to quality. This may expose any other correlations within the quality rating, or within supporting variables.



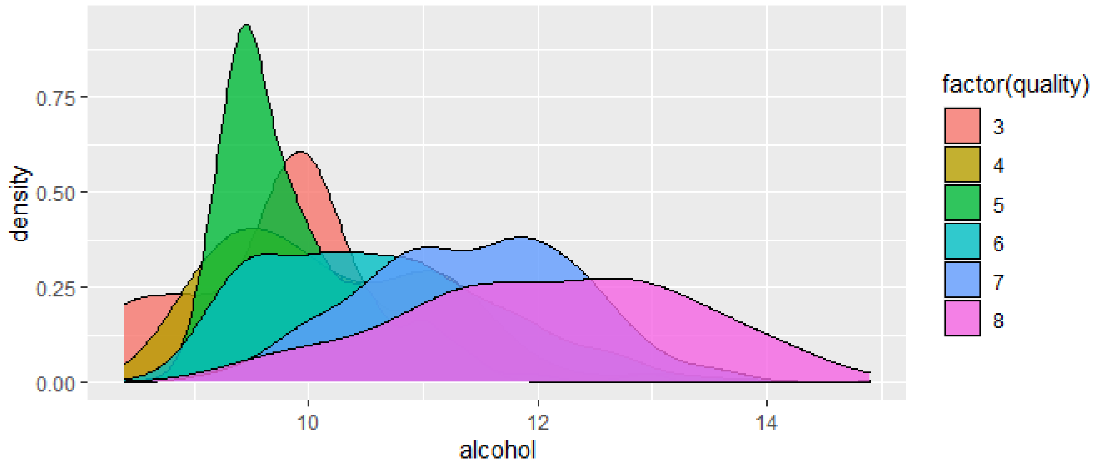
Cor = 0.12
fixed.acidity
16
12
8
quality
3
4
5
6
7
8
Cor = 0.11 (-) Outliers
fixed.acidity
12
10
8
6
quality
Cor = -0.38
volatile.acidity
1.25
1.00
0.75
0.50
0.25
quality
Cor = -0.36 (-) Outliers
volatile.acidity
1.00
0.75
0.50
0.25
quality
Cor = 0.22
citric.acid
1.00
0.75
0.50
0.25
0.00
quality
Cor = 0.23 (-) Outliers
citric.acid
0.8
0.6
0.4
0.2
0.0
quality
Cor = 0.02
residual.sugar
16
12
8
4
quality
Cor = 0.05 (-) Outliers
residual.sugar
3
2
1
quality
Cor = -0.12
chlorides
0.6
0.4
0.2
0.0
quality
Cor = -0.17 (-) Outliers
chlorides
0.12
0.10
0.08
0.06
0.04
quality
Cor = -0.07
free.sulfur.dioxide
60
40
20
0
quality
Cor = -0.07 (-) Outliers
free.sulfur.dioxide
40
30
20
10
0
quality
Cor = -0.22
total.sulfur.dioxide
300
200
100
0
quality
Cor = -0.22 (-) Outliers
total.sulfur.dioxide
125
100
75
50
25
0
quality
Cor = -0.18
density
1.000
0.995
0.990
quality
Cor = -0.16 (-) Outliers
density
1.000
0.998
0.996
0.994
0.992
quality
Cor = -0.05
pH
3.9
3.6
3.3
3.0
2.7
quality
Cor = -0.05 (-) Outliers
pH
3.7
3.5
3.3
3.1
2.9
quality
Cor = 0.24
sulphates
2.0
1.5
1.0
0.5
quality
Cor = 0.39 (-) Outliers
sulphates
0.9
0.7
0.5
0.3
quality
Cor = 0.49
alcohol
14
12
10
quality
Cor = 0.49 (-) Outliers
alcohol
13
12
11
10
9
quality

This grid plots the overall distribution of each variable as it relates to the quality rating. A linear regression line and associated correlation values are also plotted within each variable. This provides a clear view towards the positive and negative correlated values. Volatile.acidity appears to have a small negative correlation, while alcohol is closer to a moderate correlation. Also, I can see the outlier data within residual.sugar and its association between quality levels much more clear now. Most of these variables do contain a lot of outlier data.

Because of the outliers, the original distributions are compared against distributions with the outliers removed. Correleation calculations are also compared against the original distribution calculations as well. Interestingly enough, removing the outliers did not drastically change the correlations. The majority of the outliers lie within the right tail of positively skewed distributions. Given this, these outliers would equate to higher quantity levels of the given variable. Within univarariate analyzation, it was found that the tail of residual.sugar was pretty evenly dispersed among all quality levels. With this knowledge and now seeing the correlation comparisons as plotted above, it would appear as if outlier data does exist within all quality levels, thus its removal is inconsequential towards the analysis, other than to clean up the appearance of the plot. Now it would be prudent to explore further variable correlations.



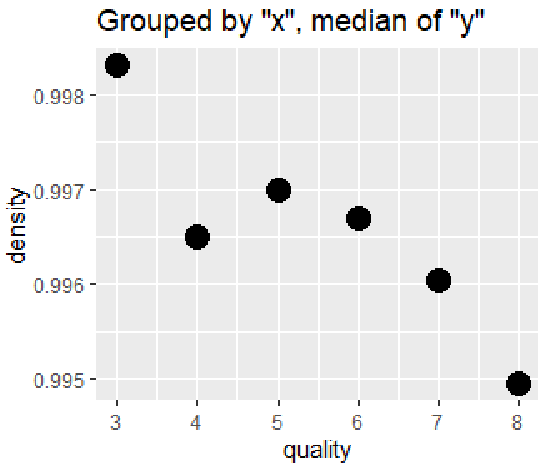
Plotting the main feature of quality rating with the variable of strongest correlation, alcohol. The chart on the left represents the median alcohol content value at each quality level. I would like to explore which supporting variables may have the highest correlation value with alcohol in an effort to isolate any dependencies.

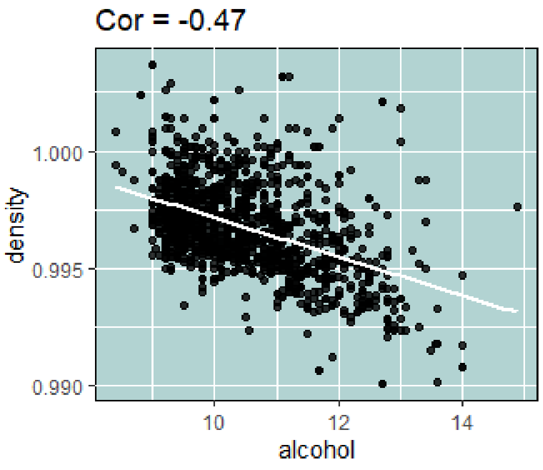


Showing the alcohol density at each quality level

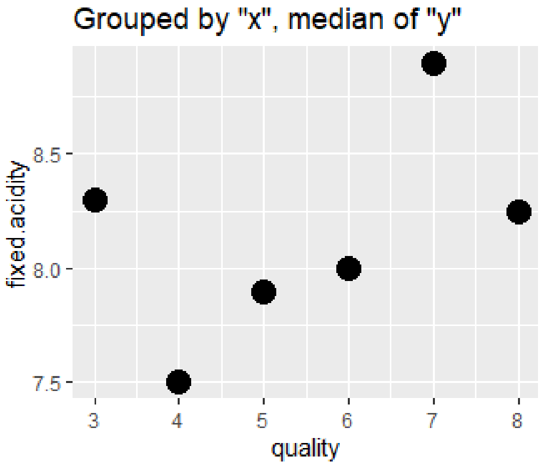
```
##                        fixed.acidity volatile.acidity citric.acid
residual.sugar
## fixed.acidity            1.00000000      -0.23591004  0.65954351
0.1018064623
## volatile.acidity        -0.23591004       1.00000000 -0.52350778
0.0150139714
## citric.acid              0.65954351      -0.52350778  1.00000000
0.1306422974
## residual.sugar           0.10180646       0.01501397  0.13064230
1.0000000000
## chlorides                0.08682118       0.05834564  0.21590904
0.0495191881
## free.sulfur.dioxide     -0.16278828       0.01838999 -0.07808537
0.1916337992
## total.sulfur.dioxide    -0.14523547       0.13801243 -0.02196773
0.1975306908
## density                  0.66455621       0.04330891  0.35988028
0.3510773624
## pH                      -0.66847091       0.18315799 -0.50499704  -
0.0686189003
## sulphates                0.17685624      -0.24317776  0.30832676
0.0003495688
## alcohol                 -0.02722621      -0.22183274  0.14293145
0.0548729369
## quality                  0.12365152      -0.37716175  0.22048370
0.0189329994
##                          chlorides free.sulfur.dioxide total.sulfur.dioxide
## fixed.acidity          0.086821179        -0.162788285          -0.14523547
## volatile.acidity       0.058345641         0.018389993           0.13801243
## citric.acid            0.215909039        -0.078085372          -0.02196773
## residual.sugar         0.049519188         0.191633799           0.19753069
## chlorides              1.000000000         0.009803183           0.04903671
## free.sulfur.dioxide    0.009803183         1.000000000           0.67243555
## total.sulfur.dioxide   0.049036711         0.672435547           1.00000000
## density                0.194901914        -0.018398693           0.06198103
## pH                    -0.272670673         0.084615857          -0.03944042
## sulphates              0.391103067         0.039847271           0.02027708
## alcohol               -0.217817835        -0.084898054          -0.22444159
## quality               -0.123382491        -0.067668038          -0.21855959
##                          density          pH     sulphates     alcohol
## fixed.acidity         0.66455621 -0.66847091  0.1768562418 -0.02722621
## volatile.acidity      0.04330891  0.18315799 -0.2431777623 -0.22183274
## citric.acid           0.35988028 -0.50499704  0.3083267628  0.14293145
## residual.sugar        0.35107736 -0.06861890  0.0003495688  0.05487294
## chlorides             0.19490191 -0.27267067  0.3911030671 -0.21781783
## free.sulfur.dioxide  -0.01839869  0.08461586  0.0398472707 -0.08489805
## total.sulfur.dioxide  0.06198103 -0.03944042  0.0202770805 -0.22444159
## density               1.00000000 -0.31741098  0.1432001261 -0.46523355
## pH                   -0.31741098  1.00000000 -0.2047777092  0.17354874
## sulphates             0.14320013 -0.20477771  1.0000000000  0.09082713
## alcohol              -0.46523355  0.17354874  0.0908271338  1.00000000
## quality              -0.17869845 -0.04984500  0.2373255704  0.48927491
##                          quality
## fixed.acidity         0.12365152
## volatile.acidity     -0.37716175
## citric.acid           0.22048370
## residual.sugar        0.01893300
## chlorides            -0.12338249
## free.sulfur.dioxide  -0.06766804
## total.sulfur.dioxide -0.21855959
## density              -0.17869845
## pH                   -0.04984500
## sulphates             0.23732557
## alcohol               0.48927491
## quality               1.00000000
```

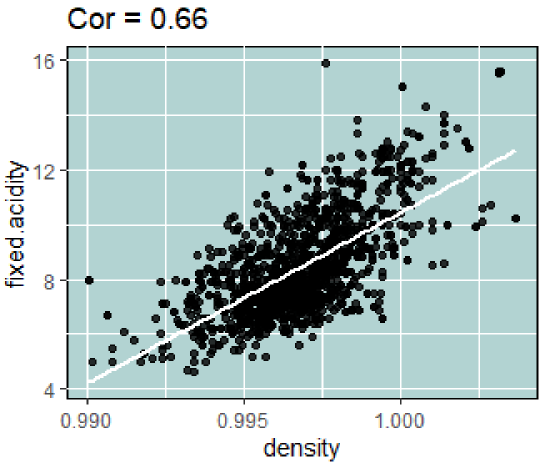
A correlation table will be useful to hone in on any other possible correlated variables. Given that alcohol has the strongetst correlation within quality, I would like to follow alcohol and identify a trail of association from within it. If possible, I would like to follow each strongest correlation to the next in an effort to unfold any possible patterns.



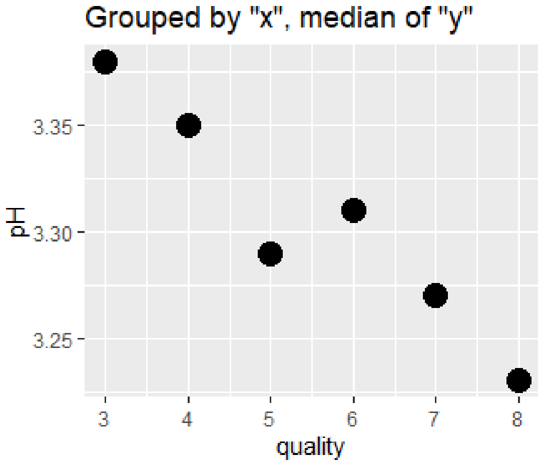


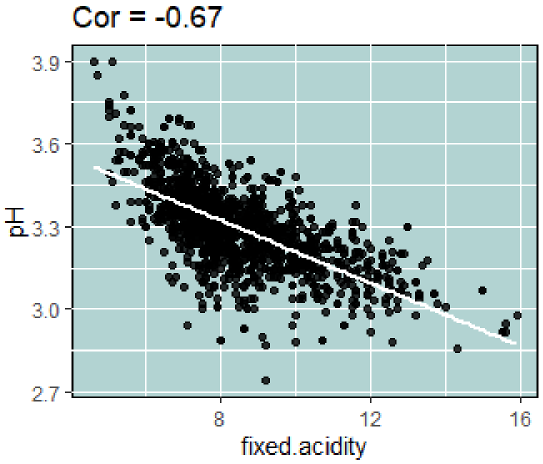
Density is the strongest correlation with alcohol, the negative association with quality can also be seen within the plot to the upper left.



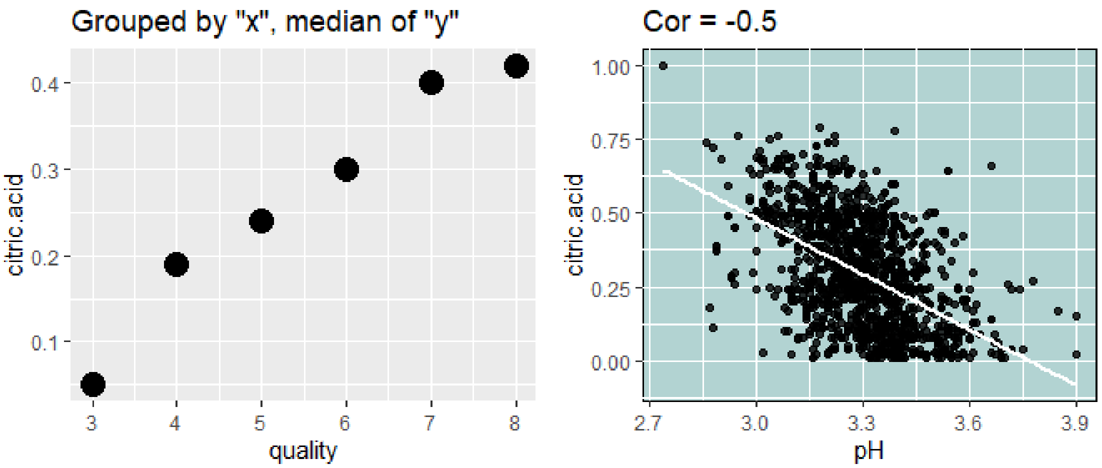


Fixed.acidity has a strong correlation with density. A slight positive assoction from within quality can also be noted within the plot to the left.

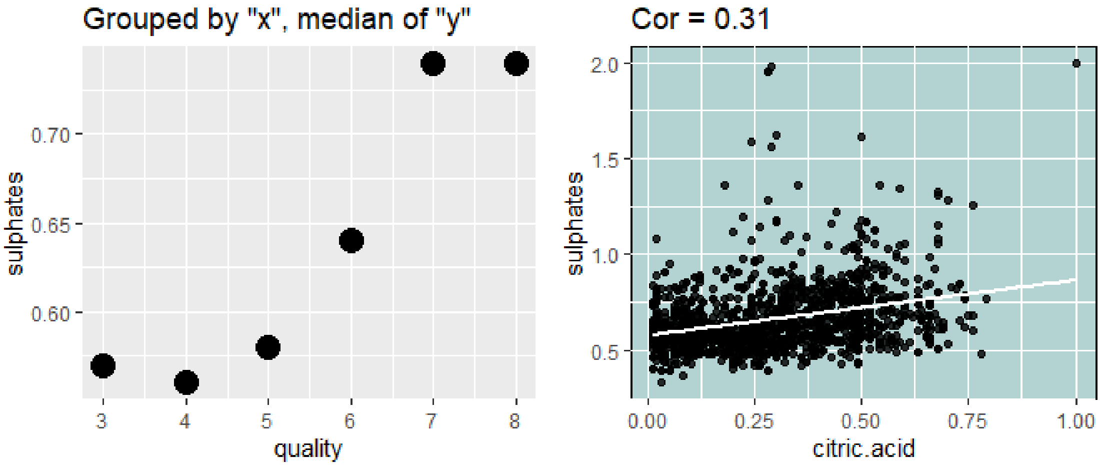




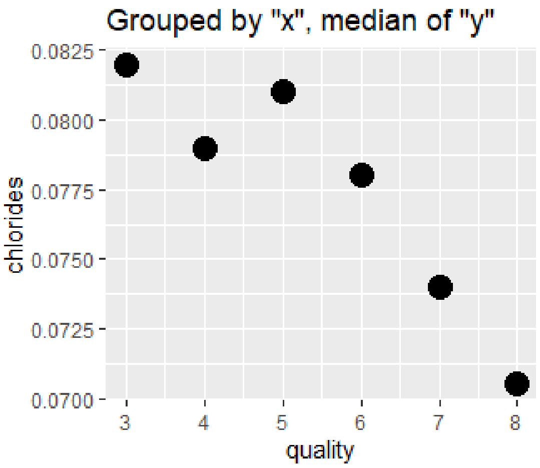
pH as shown, has a very strong relationship with fixed.acidity. Given the nature of science, this makes sense. A negative association with quality is also noted from within the leftmost plot.

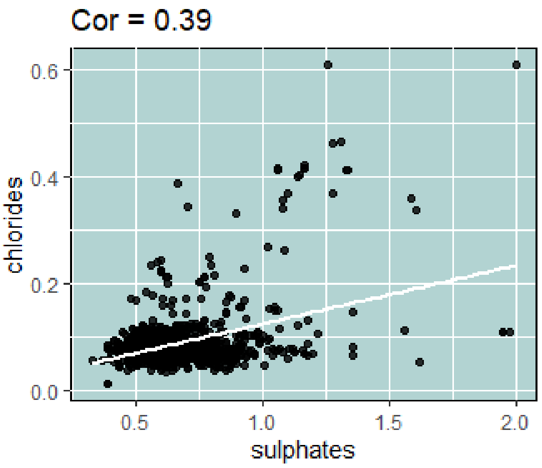


Not suprisingly, pH has a strong correlation with another acid, citric.acid. It is starting to become clear that acidity is part of a bigger picture in determining wine quality.

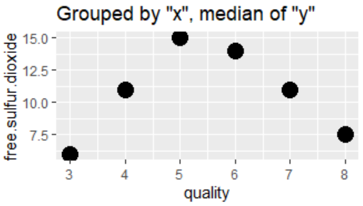


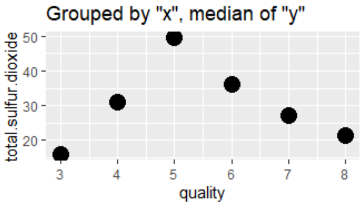
citric.acid does not contain a lot of strong associations, however I felt I should follow the path. Sulphates appear to be of importance for high quality wines.

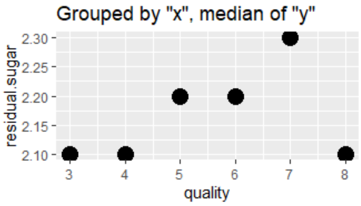




Chlorides also did not hold a strong association from within sulphates, but it seemed to be the last within this group of possible associated variables. Of note though is the negative association from within the quality rating.







Lastly, a quick look at the remaining variables of free/total.sulfur.dioxide and residual.sugar as compared to median values at each quality rating. Given the available data, an association withi quality is lacking from within these variables.

Bivariate Analysis

Bivariate explorations produced some interesting findings. Initially, the quality rating did not appear to have very many strong correlations. However, in following the variable of alcohol, qualitative associations were found from within most variables. The thought of outliers was also explored within this section. Several attempts were made to expose outliers as a potential problem within modeling the data, however it appears as if these outliers do not heavily impact the trajectory of any future regression modelling.

Relationsips among variables

Alcohol and quality do share a relationship, however it is the child relationships from within alcohol that are of the strongest from within the dataset. Not surprisingly , all of the acidic variables, and pH share a strong correlation.

Interesting relationships

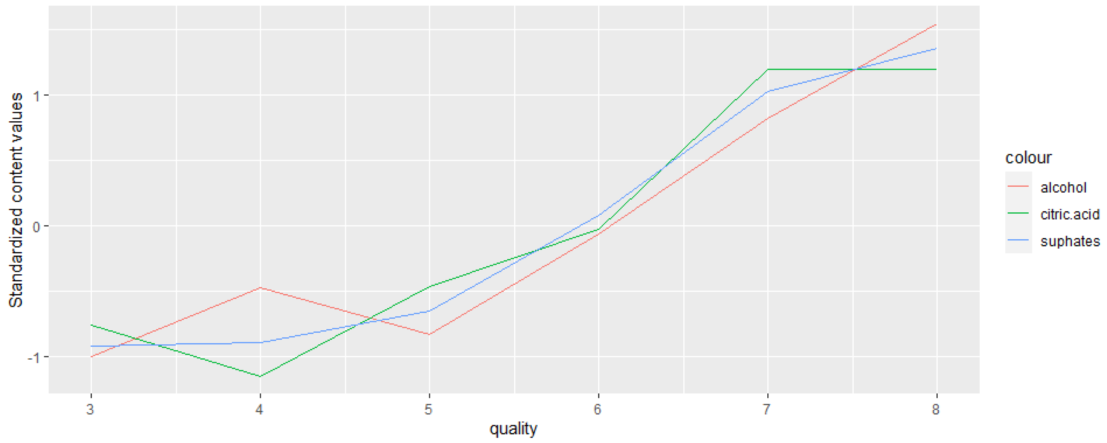
The most interesting relationship is that of density. It seems as if density falls right in the middle of alcohol and acidity levels. In essence, acting as a byproduct of the combined chemical reactions from within the other variables.

Strongest relationship

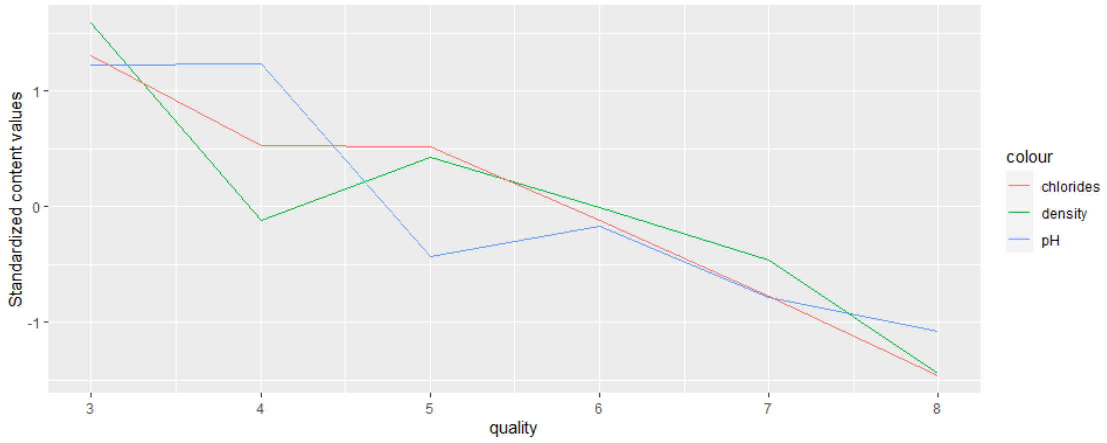
Without surprise, the strongest relationship found is that of pH and acidity levels, strongest of which would be pH/fixed.acidity.

Multivariate Plots Section

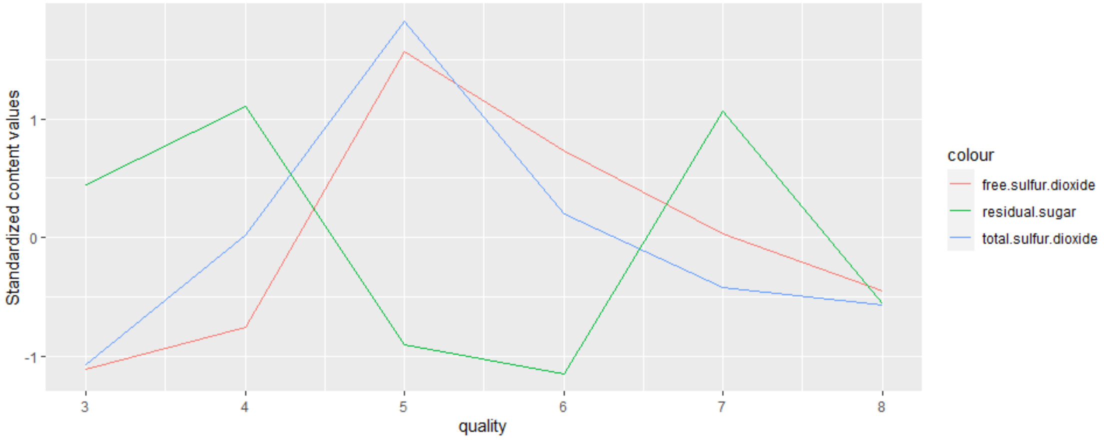
At a multivariate level, combining like variables would be helpful. The dataframe will be standardized to achieve these results.



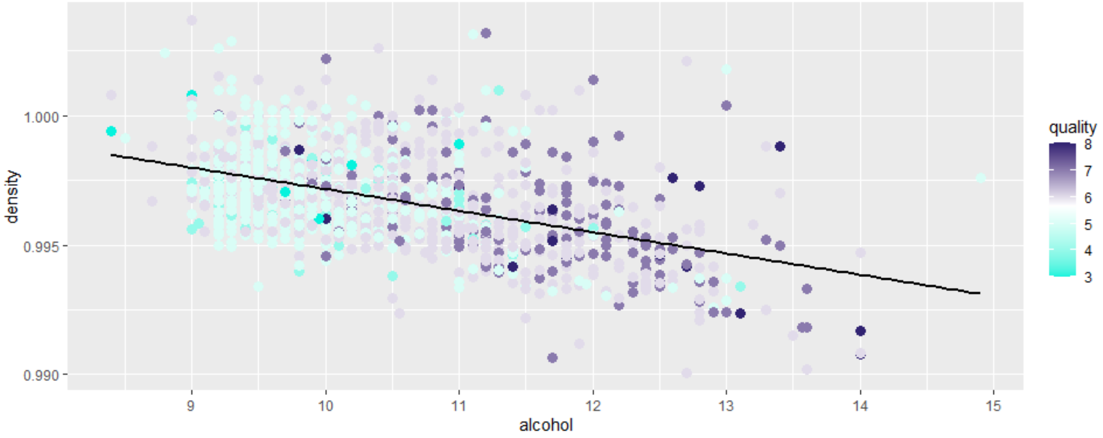
Variables with positive quality associations are grouped by their mean at each quality rating and plotted plotted within a standardized format. These variables are strongly associated with the underlying quality rating.



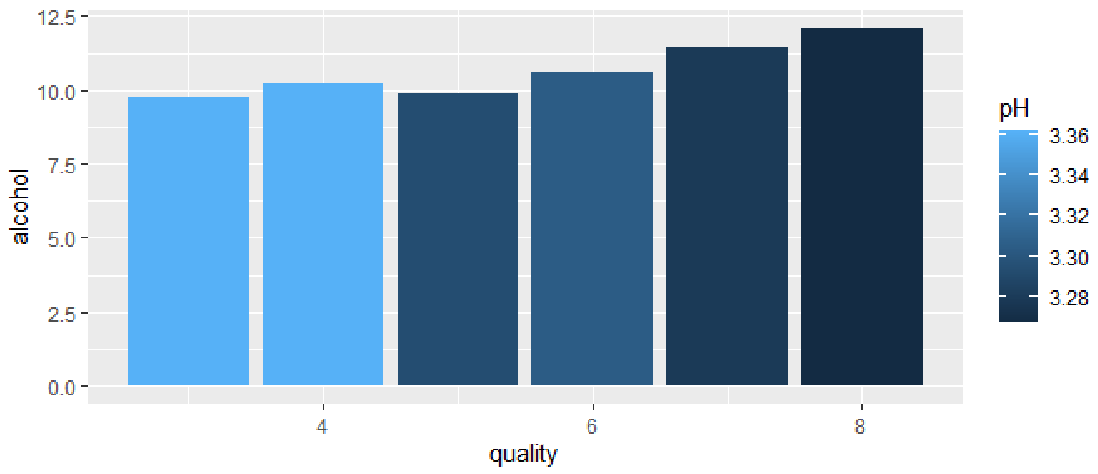
Variables with negative quality associations are grouped by their mean at each quality rating and plotted within a standardized format. These variables are also strongly associated with the underlying quality rating.



Variables that do not show an obvious association within quality are grouped by their quality rating and plotted within a standardized format. Of interest is the appearance of an inverse quality between residual.sugar and sulfur.dioxide.



Given the association between alchol and density, these variables are looked at within each quality rating. This shows the highest quality ratings are of higher alcohol and lower density, with quality being the variable of choice for color formatting.



There also seems to be a correlation between alcohol, pH, and quality. This plot is simple, but it shows mean values grouped by quality rating. Alcohol increases and pH decreases as quality ratings rise.

Multivariate Analysis

Relationships

Alcohol, density, and acidic levels all look to be decent indicators as to what quality rating a wine may be.

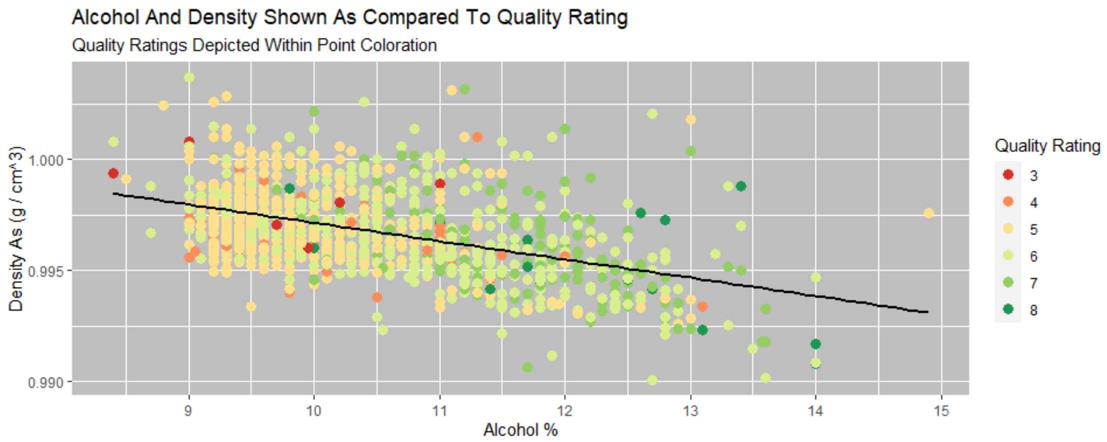
Surprising interactions

Residual.sugar and its interaction within density, and alcohol was the most surprising feature I found. Although the correlation is not strong, there does seem to be some significance to the fact that sugar levels are highest among high density values.

Final Plots and Summary

Prior observations have shown the overall unison that the majority of these variables play towards deciding the quality level of red wine. This theme will be depicted within the final plots.

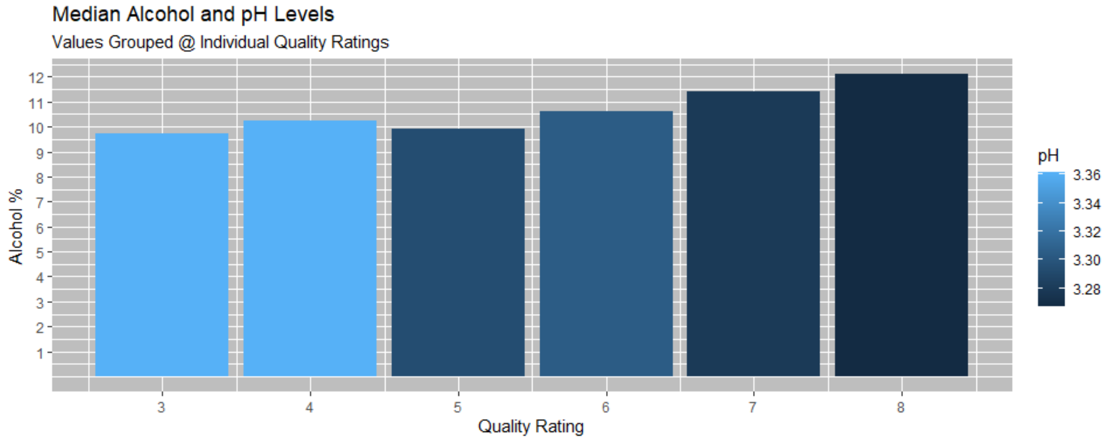
Plot One



Description One

This plot shows how alcohol percent and density work in unison to help determine the quality rating.

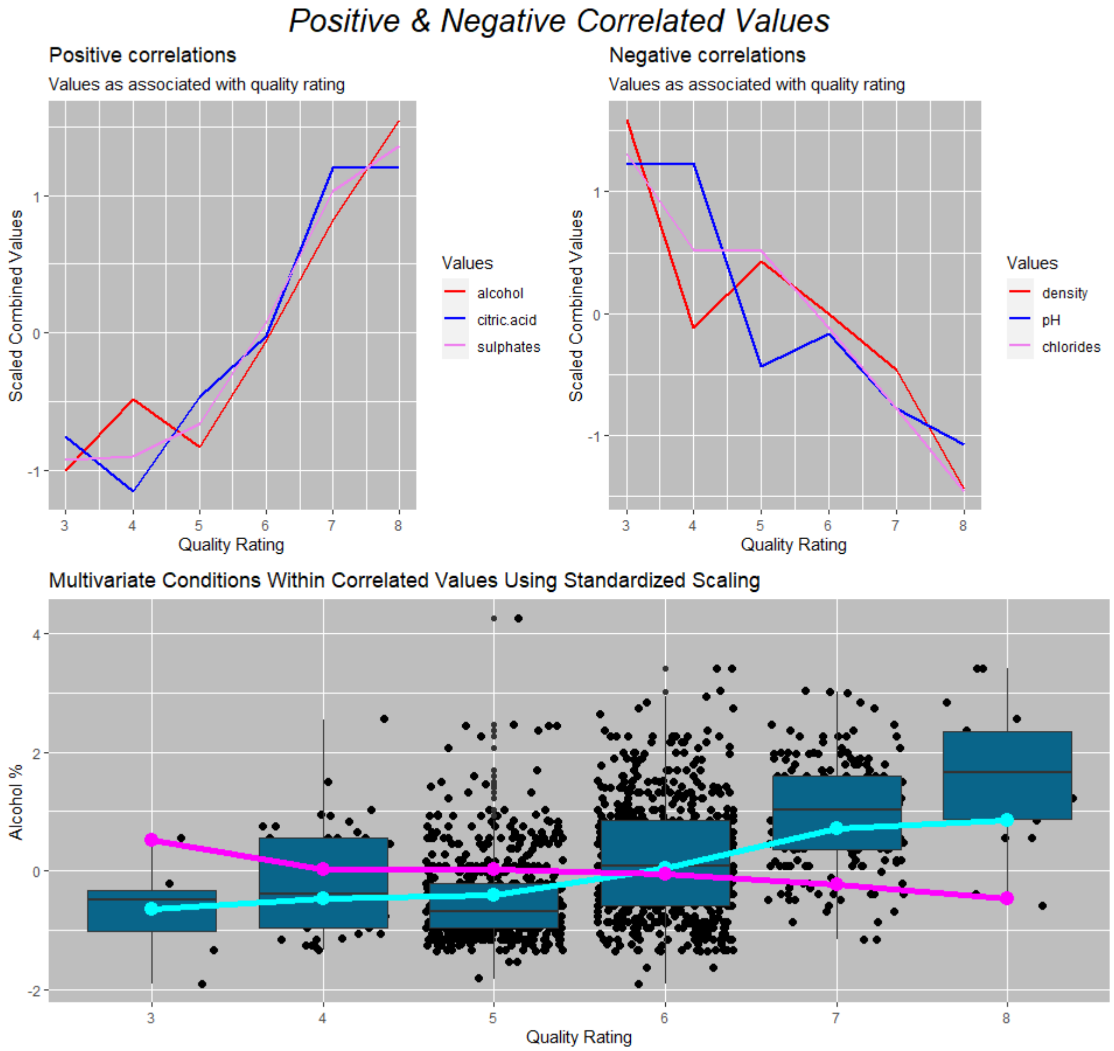
Plot Two



Description Two

This is a simple but powerful representation of how acidic levels and alcohol content influence the quality rating. Higher alcohol and lower pH equate to higher quality ratings.

Plot Three



Correlated Median Values Are Combined And Shown As:
Positive: citric.acid, sulphates, alcohol (Blue)
Negative: pH, chlorides, density (Magenta)

Description Three

This is a view an overview of the most correlated variables and how they react from within the quality rating. Values have all been standardized for ease of comparison. To begin with, conditional means of positive and negative correlations are calculated and plotted individually. Lastly, these items are combined on the lower chart into two lines, one positive, and one negative. Given that alcohol is the single strongest correlated value, it is chosen for the y axis within the boxplot. With this plot, it also becomes clear how all of these variables work in unison to produce the overall result of quality.

Reflection

The red wine dataset has 13 original variables and around 20,000 entries. Domain knowledge would be beneficial, however with exploration this dataset can provide the information needed to help predict the quality of wine. Because of my lack of domain knowledge I did need to first explore each individual variable. Once this was complete, I was able to have a better idea of how to proceed from within the bivariate and multivariate explorations.

The dominant correlation from within quality is alcohol. On its own, I do believe that alcohol can provide a fair level of predictability towards quality, however it should be noted that alcohol does not exist in unison. Without the remaining supporting variables, its predictability would undoubtedly diminish. I should also note that on a personal level, I do ponder the question as to why alcohol is such a strong predictor given the subjectivity of a human created quality rating system. Is this a matter of conditioning, where an individual is used to the taste of the higher alcohol and its given physiological effects, and thus ratings increase when this taste is discovered?   Overall, the main difficulties that I ran into were that of domain knowledge. It is more difficult to trust the results when lacking a foundation of knowledge. As an example, I was unsure of the zero values from within citric acid. I was not sure if these values were possible or not, but they did seem out of place. I tested the data with and without these entries and I did not notice a significant difference without, so I did ultimately decide upon their removal.   The primary limitations from within the dataset are from within the lack of data observations at higher and lower quality levels. It can be assumed that these ratings are unique, however a larger sample size would be nice. To achieve this, an larger overall dataset can be acquired and then randomly sampled to where all quality levels have the same number of entries.

References

Sjoberg DD, Whiting K, Curry M, Lavery JA, Larmarange J. Reproducible summary tables with the gtsummary package. The R Journal 2021;13:570–80. https://doi.org/10.32614/RJ-2021-053.   Winston Chang, Cookbook for R http://www.cookbook-r.com/