Red Wine Exploratory Data Analysis

by Javier Soto

==========

Introduction

This analysis explores the univariate, bivariate, & multivariate relationships between the variables in the provided tidy Red Wine data set using RStudio. The complete dataset can be found here. (https://www.google.com/url?q=https://s3.amazonaws.com/udacity-hosted-downloads/ud651/wineQualityReds.csv&sa=D&ust=1539234033126000)

Red Wine Data - Load and Assessment

First step is to load the data and do some basic assessments. This includes displaying the internal structure of the initial data frame.

```
## 'data.frame':    1599 obs. of  13 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
##  $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
##  $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071
...
##  $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
##  $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
##  $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
##  $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...
##  $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...
##  $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
##  $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...
```

Red Wine Data - Transforming Quality Integer Data to a Factor

Next, programatically transformed the integer Quality data to a Factor, and output the head and tail results for review.

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

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

Red Wine Data - Create New Rating Factor

Generated a new Factored Variable named 'Rating,' and output the head and tail results for review.

```
## [1] average average average average average average
## Levels: bad < average < good
```

```
## [1] average average average average average average
## Levels: bad < average < good
```

Red Wine Data - Display New Internal Structure

Display the internal structure after creating the new Factored Variable: Rating.

```
## 'data.frame':    1599 obs. of  13 variables:
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
##  $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
##  $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071
...
##  $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
##  $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
##  $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
##  $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...
##  $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...
##  $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
##  $ quality             : Ord.factor w/ 6 levels "3"<"4"<"5"<"6"<..: 3 3 3 4 3 3 3 5 5 3
...
##  $ rating              : Ord.factor w/ 3 levels "bad"<"average"<..: 2 2 2 2 2 2 2 3 3 2
...
```

Red Wine Data - Load and Assessment

In addition, listed out the column (variable) names of interest for reference. Superfluous columns were excluded.

```
##
## * fixed.acidity
## * volatile.acidity
## * citric.acid
## * residual.sugar
## * chlorides
## * free.sulfur.dioxide
## * total.sulfur.dioxide
## * density
## * pH
## * sulphates
## * alcohol
## * quality
## * rating
##
## <!-- end of list -->
```

Red Wine Data - Display Head of Selected Columns

Output of the head for selected columns.

```
##     pH sulphates alcohol quality  rating
## 1 3.51      0.56     9.4       5 average
## 2 3.20      0.68     9.8       5 average
## 3 3.26      0.65     9.8       5 average
## 4 3.16      0.58     9.8       6 average
## 5 3.51      0.56     9.4       5 average
```

Red Wine Data - Display Tail of Selected Columns

Output of the tail for selected columns.

```
##        pH sulphates alcohol quality  rating
## 1595 3.45      0.58    10.5       5 average
## 1596 3.52      0.76    11.2       6 average
## 1597 3.42      0.75    11.0       6 average
## 1598 3.57      0.71    10.2       5 average
## 1599 3.39      0.66    11.0       6 average
```

Red Wine Data - Display Summary Statistics

Output of the Summary Statistics together with the Counts for the Factored Variables:

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

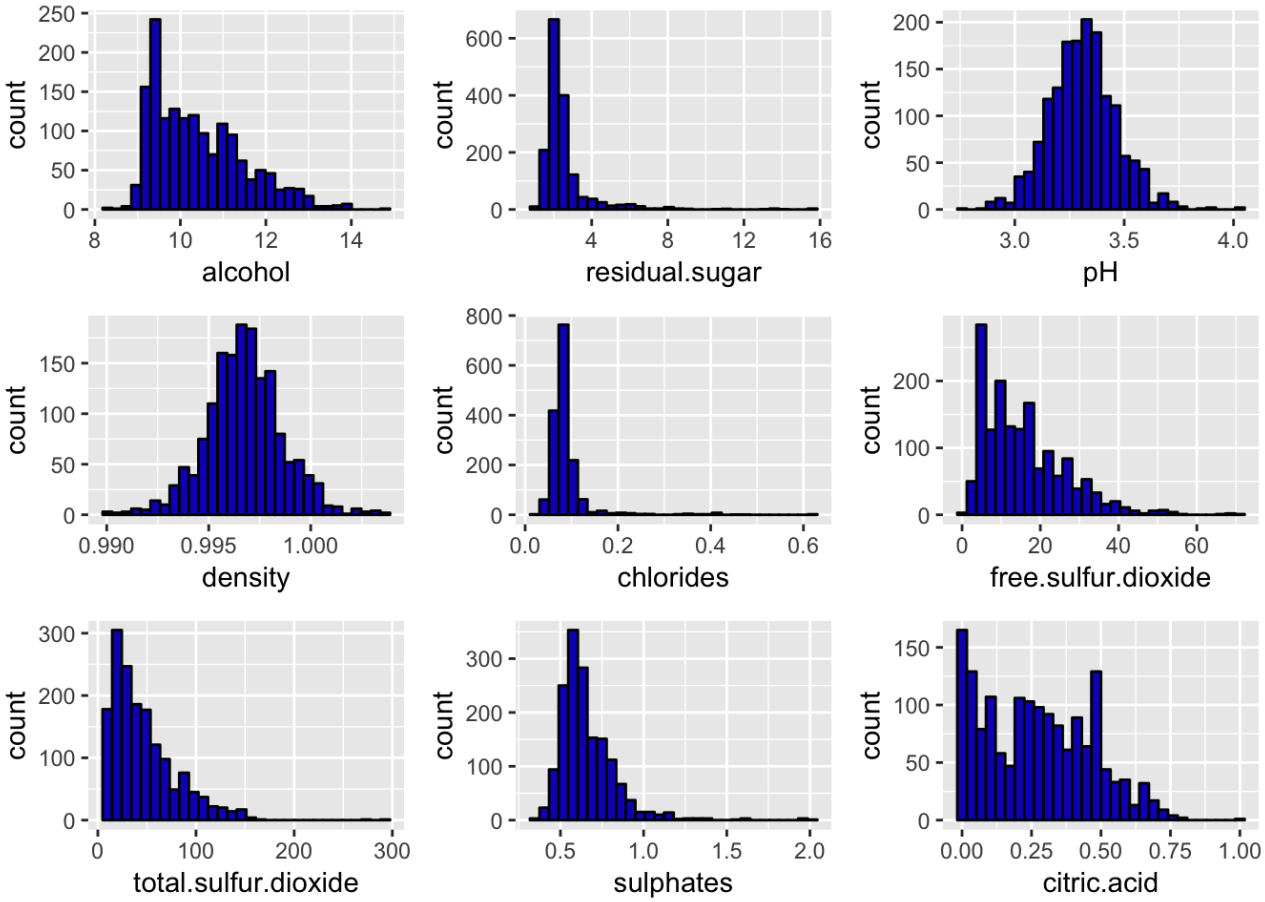
chlorides	free.sulfur.dioxide	total.sulfur.dioxide	density
Min. :0.01200	Min. : 1.00	Min. : 6.00	Min. :0.9901
1st Qu.:0.07000	1st Qu.: 7.00	1st Qu.: 22.00	1st Qu.:0.9956
Median :0.07900	Median :14.00	Median : 38.00	Median :0.9968
Mean :0.08747	Mean :15.87	Mean : 46.47	Mean :0.9967
3rd Qu.:0.09000	3rd 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
Min. :2.740	Min. :0.3300	Min. : 8.40
1st Qu.:3.210	1st Qu.:0.5500	1st Qu.: 9.50
Median :3.310	Median :0.6200	Median :10.20
Mean :3.311	Mean :0.6581	Mean :10.42
3rd Qu.:3.400	3rd Qu.:0.7300	3rd Qu.:11.10
Max. :4.010	Max. :2.0000	Max. :14.90

quality	rating
3: 10	bad : 63
4: 53	average:1319
5:681	good : 217
6:638	NA
7:199	NA
8: 18	NA

Univariate Plots Section

First lets plot the distribution of each variable to get a sense of the data. Now, lets remove some outliers and explore these individual univariate plots in detail!



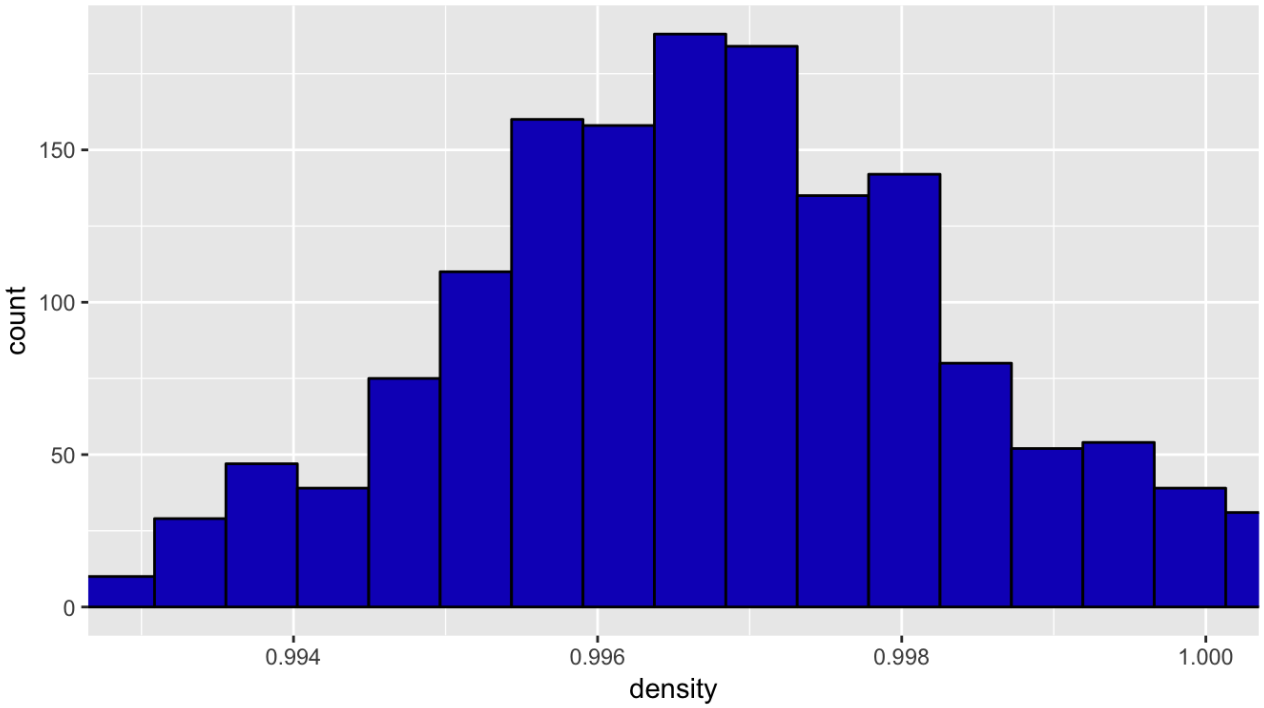
Univariate Plots - pH and Density Normally Distributed

Both distribution plots for pH and Density appear to be normally distributed.

The mean typically is equal to the median.

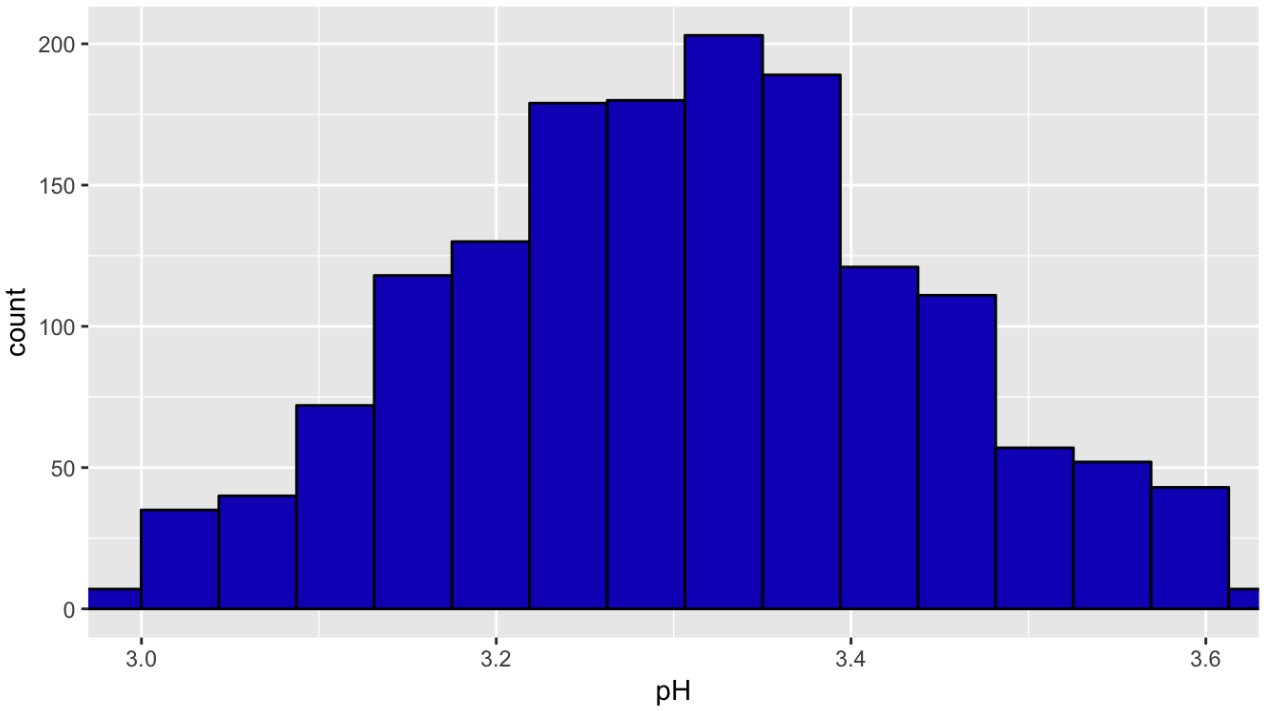
Density Histogram between 0.9930 and 1.0000

Mean and Median equal .997



pH Histogram between 3.00 and 3.60

Mean and Median equal 3.31



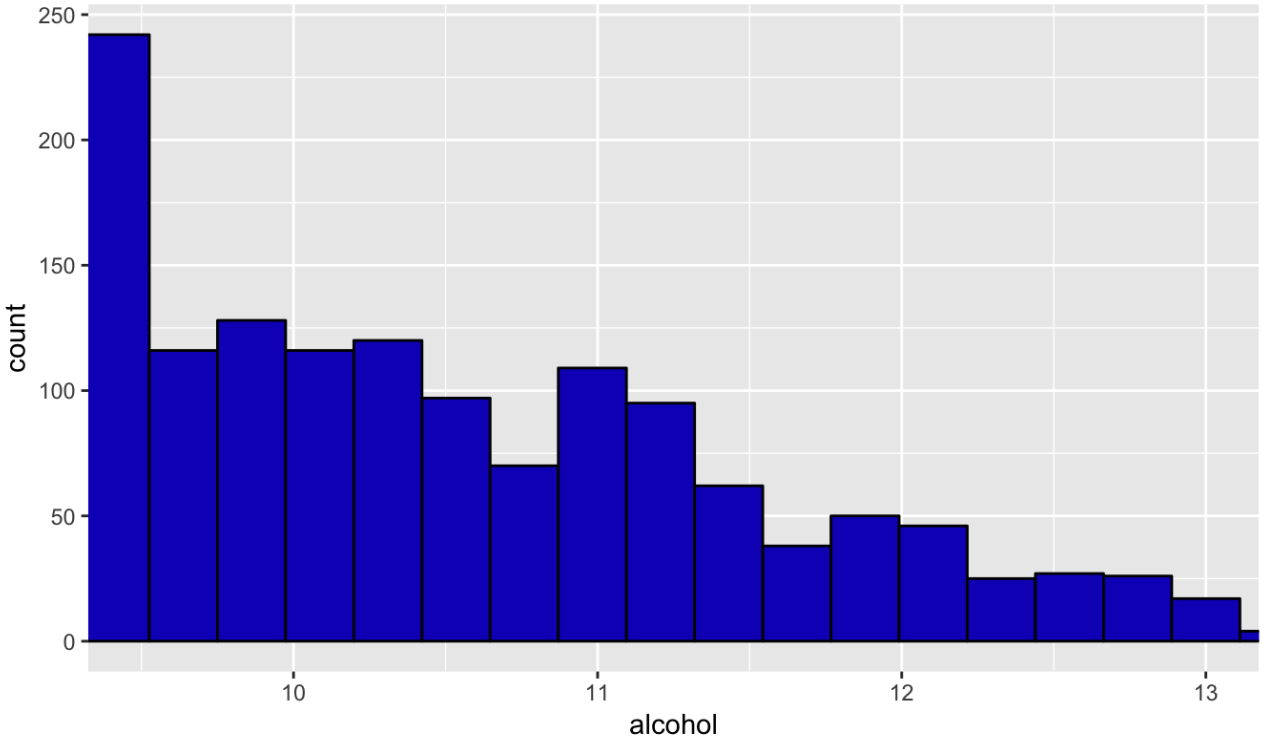
Univariate Plots - Right Skewed Distributions

The distribution plots for alcohol, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, and sulphates appear to be rightly skewed.

For a right skewed distribution, the mean is typically greater than the median. Also, the tail of the distribution on the right hand (positive) side is longer than on the left hand side.

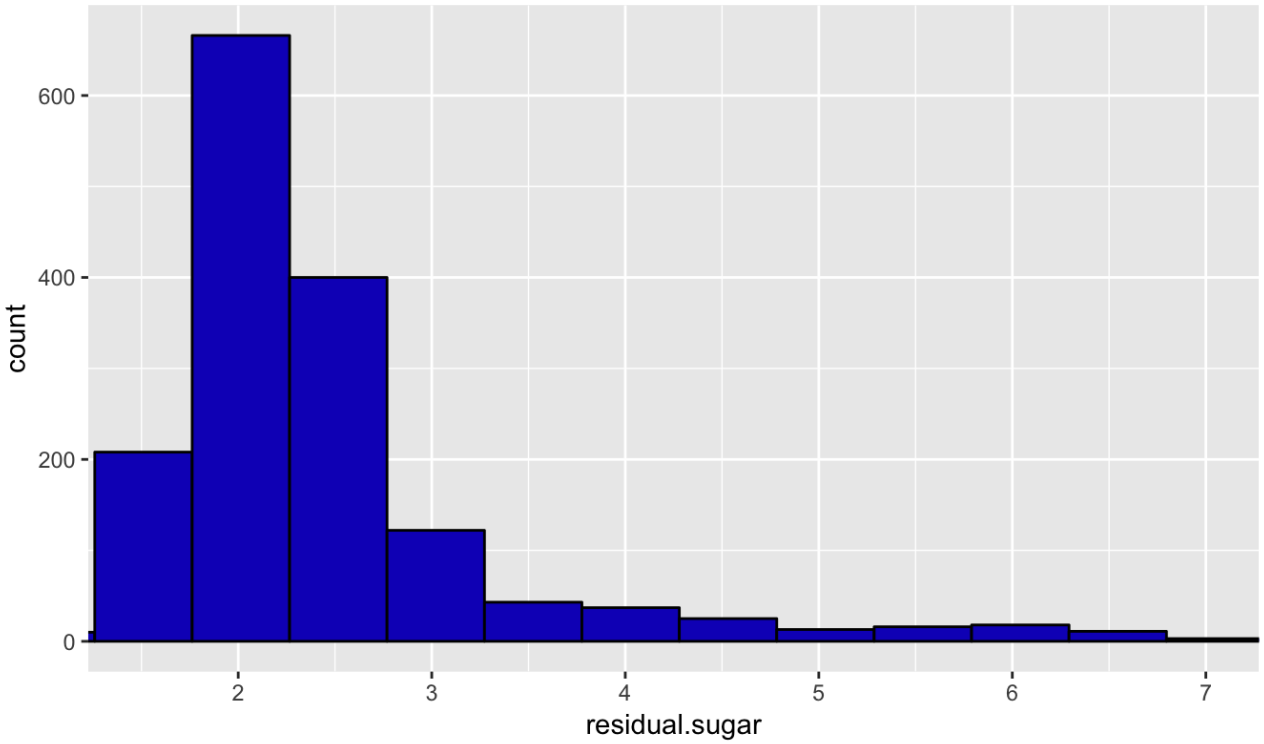
Alcohol Histogram between 9.00 and 13.00

Mean 10.42 is greater than median 10.20



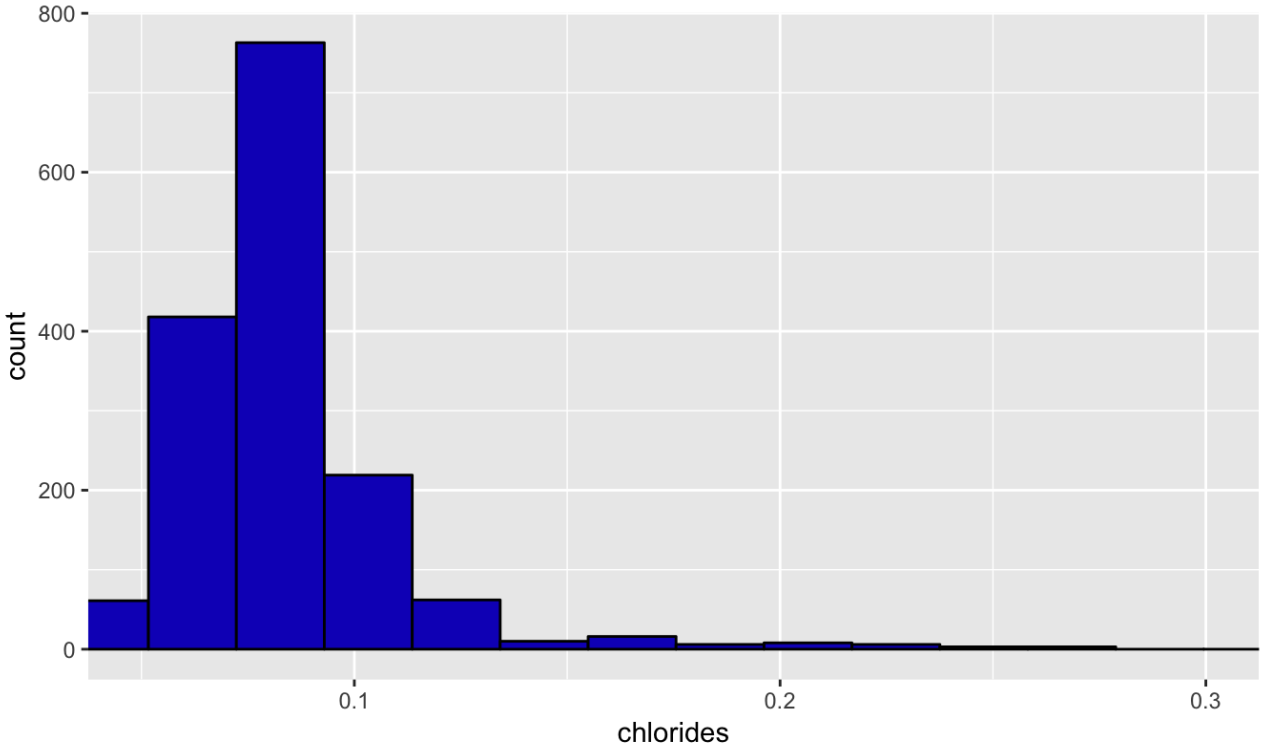
Residual Sugar Histogram between 1.50 and 7.00

Mean 2.539 is greater than median 2.20



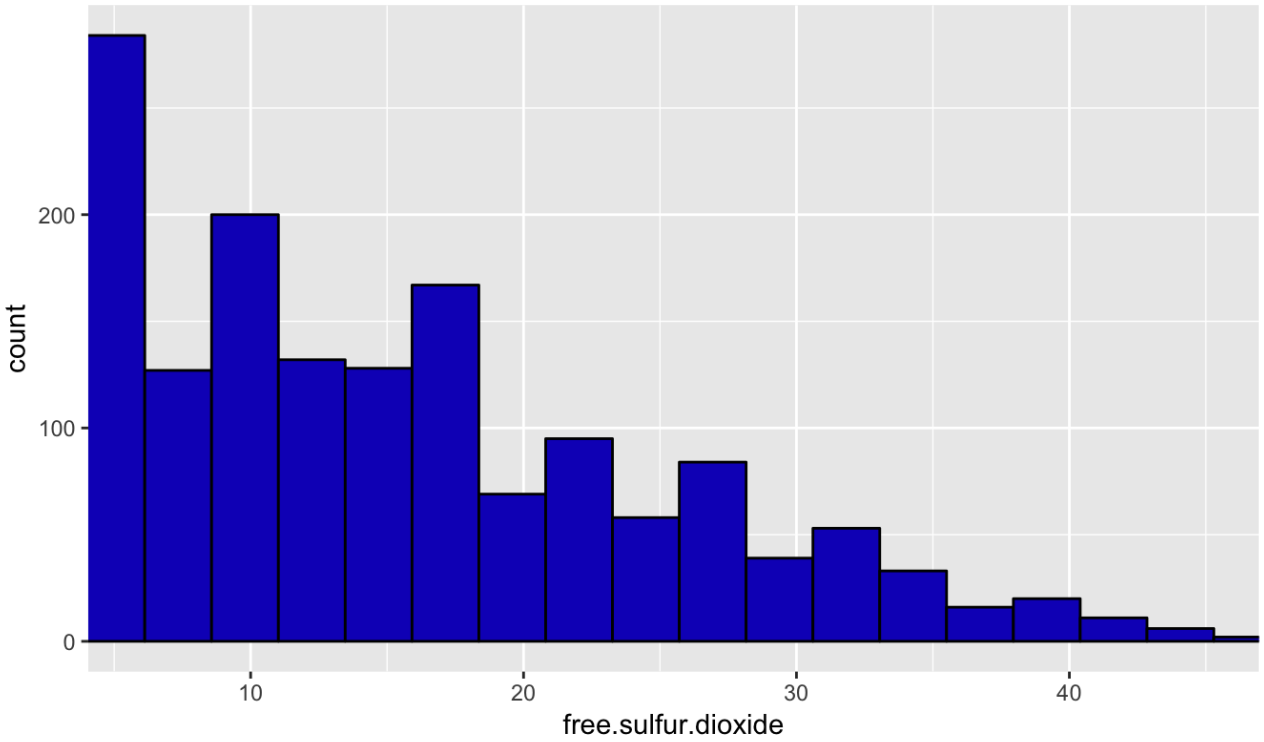
Chlorides Histogram between 0.05 and 0.30

Mean 0.087 is greater than median 0.079



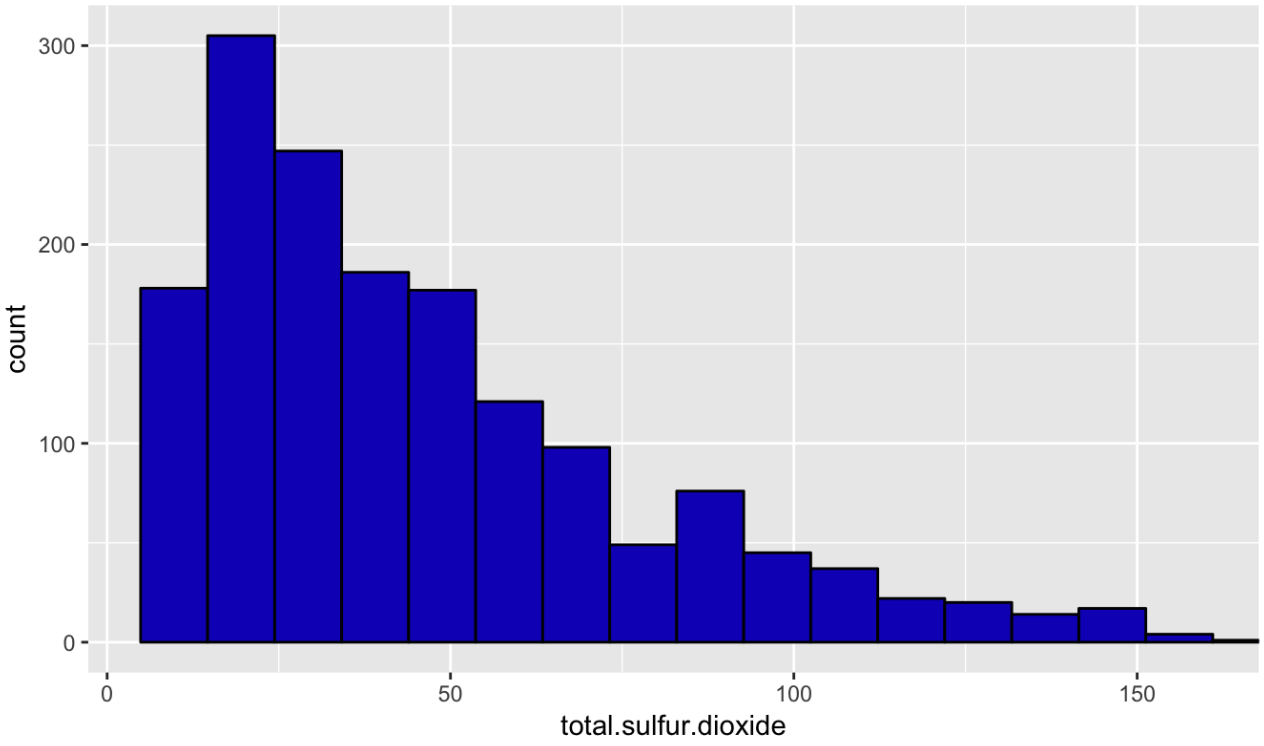
Free Sulfur Dioxide Histogram between 6.00 and 45.00

Mean 15.87 is greater than median 14.00



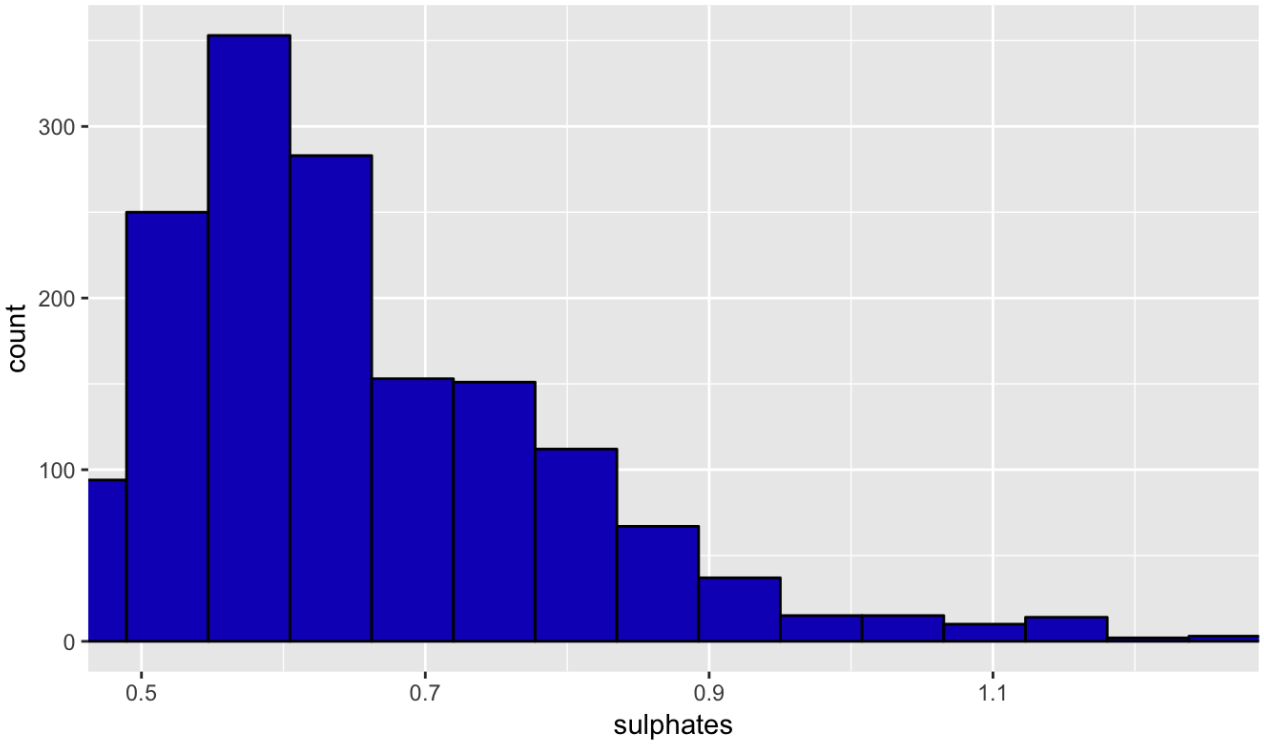
Total Sulfur Dioxide Histogram between 5.00 and 160.00

Mean 46.47 is greater than median 38.00



Sulphates Histogram between 0.50 and 1.25

Mean 0.6581 is greater than median 0.6200



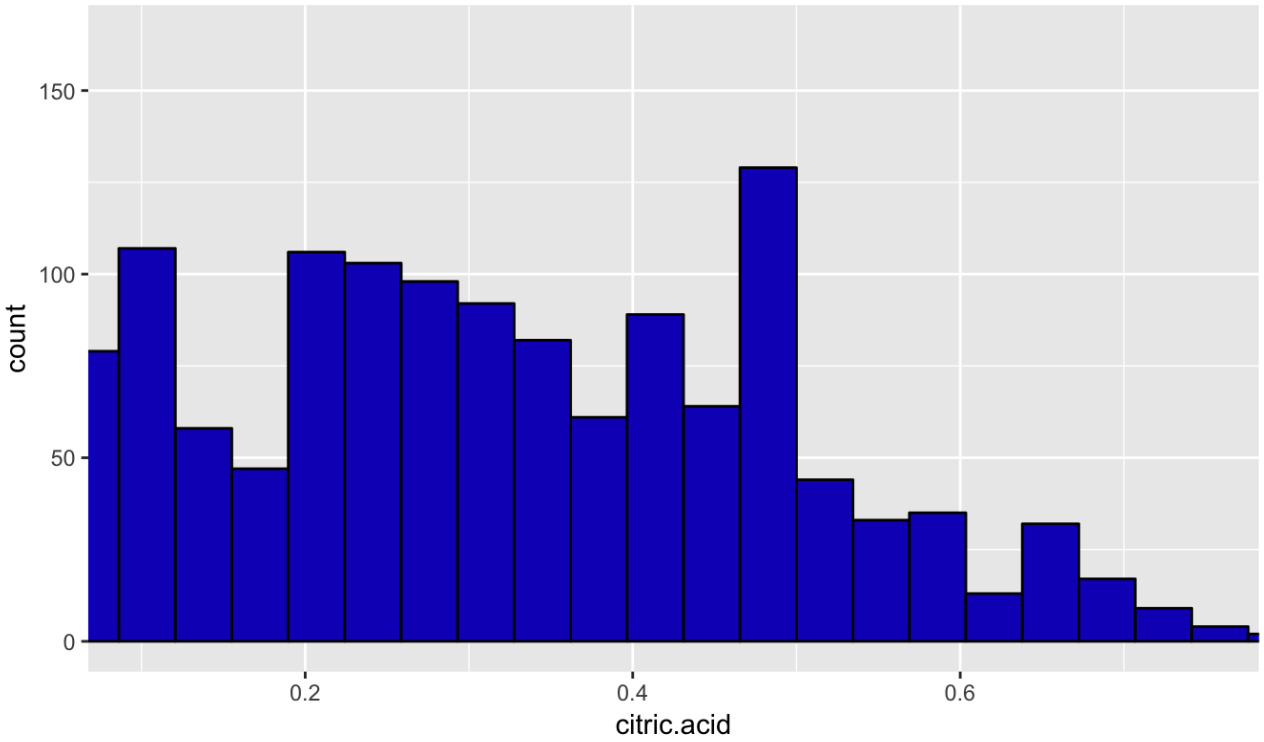
Univariate Plots - Citric Acid is Bimodally Distributed

The distribution plot for Citric Acid appears to be Bimodally Distributed.

In statistics, a bimodal distribution appears as distinct peaks.

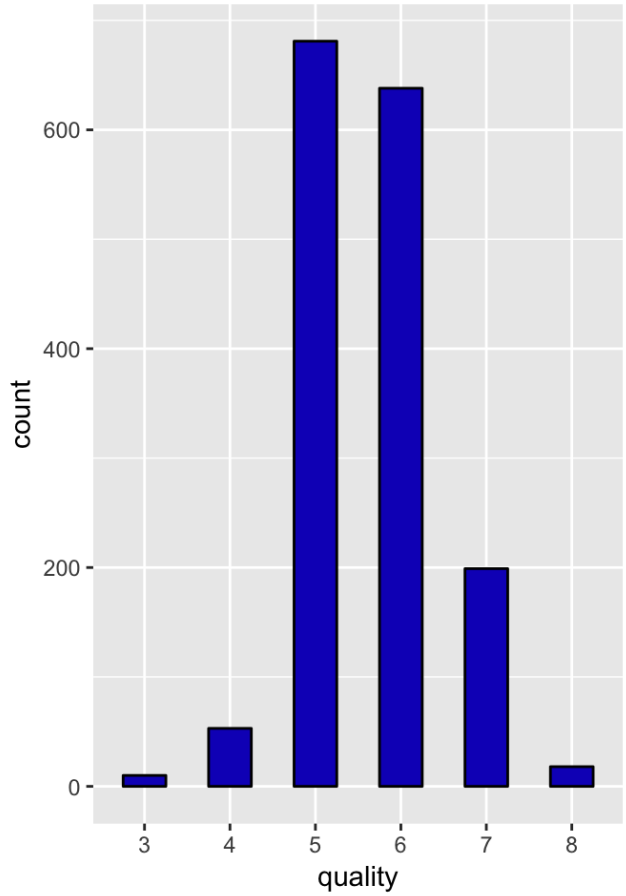
Sulphates Histogram between 0.10 and 0.75

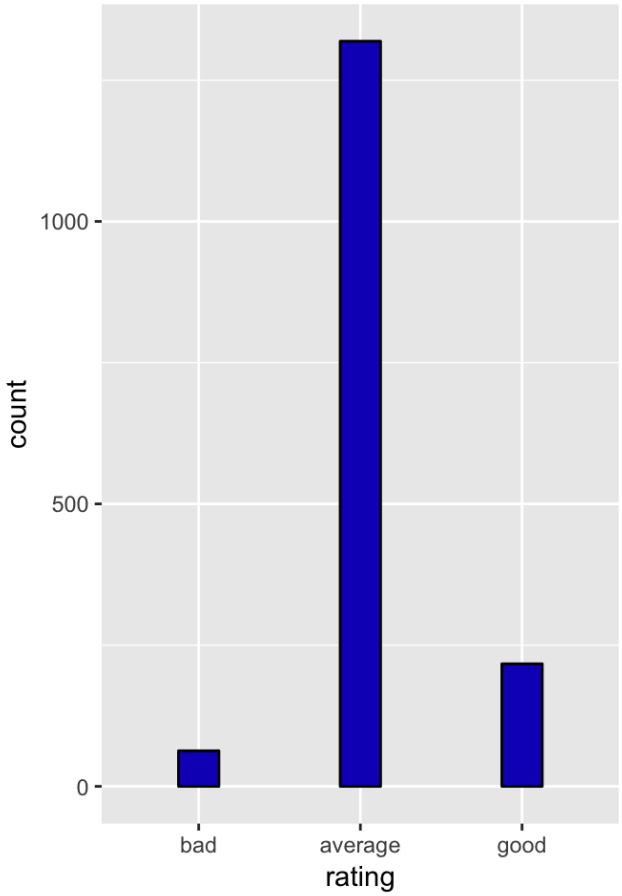
Citric Acid is Bimodally Distributed



Univariate Plots - Quality and Ratings

Most of the wines in the dataset are average quality wines. I need to determine the main variables that are driving these results.





Analysis of the Univariate Plots

Structure of the dataset?

The Red Wine Dataset has 1599 rows and 13 columns. The categorical variables are ‘quality’ and ‘rating’, and the remainder are numerical variables that indicate the physical and chemical properties of the wine.

It was observed that most of the wines are categorized as ‘average’ quality with a few ‘bad’ and ‘good’ wines, respectively, in the tals. Also, all variables showed outliers. This suggests that the data is of poor predictive value.

Main Feature

My main point of interest in this dataset is the average ‘rating’, and why so many wines fell into that bucket. I would like to determine which factors determine the quality of a wine and what factors are critical for a good wine versus a bad wine.

Other Features

The variables related to acidity (fixed, volatile, citric.acid and pH) might explain some of the variance. The different acid concentrations might affect the taste of the wine. Also,residual sugar determines how sweet a wine is, so it also may influence the taste.

New Variable Created

A rating variable was created.

Unusual Distributions

Citric.acid stood out from the other distributions. It had outliers and what appeared to be a bimodal distribution. Also, most of the variables were rightly skewed. Outliers were removed to confirm that the observed bimodal and skewed distributions were valid, and not the result of a few outliers.

Bivariate Plots Section

First step is to generate a correlation table, for all of the relevant variables in the provided red wine data set, to better understand the relationships between them.

```
##
##
## +-------------------------+---------------+------------------+-------------+
## |                    | fixed.acidity | volatile.acidity | citric.acid |
## +=========================+===============+==================+=============+
## | **fixed.acidity**       |             1 |            -0.26 |    **0.67** |
## +-------------------------+---------------+------------------+-------------+
## | **volatile.acidity**    |         -0.26 |                1 |   **-0.55** |
## +-------------------------+---------------+------------------+-------------+
## | **citric.acid**         |      **0.67** |        **-0.55** |           1 |
## +-------------------------+---------------+------------------+-------------+
## | **residual.sugar**      |          0.11 |                0 |        0.14 |
## +-------------------------+---------------+------------------+-------------+
## | **chlorides**           |          0.09 |             0.06 |         0.2 |
## +-------------------------+---------------+------------------+-------------+
## | **free.sulfur.dioxide** |         -0.15 |            -0.01 |       -0.06 |
## +-------------------------+---------------+------------------+-------------+
## | **total.sulfur.dioxide** |         -0.11 |             0.08 |        0.04 |
## +-------------------------+---------------+------------------+-------------+
## | **density**             |      **0.67** |             0.02 |    **0.36** |
## +-------------------------+---------------+------------------+-------------+
## | **pH**                  |     **-0.68** |             0.23 |   **-0.54** |
## +-------------------------+---------------+------------------+-------------+
## | **sulphates**           |          0.18 |            -0.26 |    **0.31** |
## +-------------------------+---------------+------------------+-------------+
## | **alcohol**             |         -0.06 |             -0.2 |        0.11 |
## +-------------------------+---------------+------------------+-------------+
## | **quality**             |          0.12 |        **-0.39** |        0.23 |
## +-------------------------+---------------+------------------+-------------+
##
## Table: Table continues below
##
##
##
## +-------------------------+----------------+-----------+
## |                    | residual.sugar | chlorides |
## +=========================+================+===========+
## | **fixed.acidity**       |           0.11 |      0.09 |
## +-------------------------+----------------+-----------+
## | **volatile.acidity**    |              0 |      0.06 |
## +-------------------------+----------------+-----------+
## | **citric.acid**         |           0.14 |       0.2 |
## +-------------------------+----------------+-----------+
## | **residual.sugar**      |              1 |      0.06 |
## +-------------------------+----------------+-----------+
## | **chlorides**           |           0.06 |         1 |
## +-------------------------+----------------+-----------+
## | **free.sulfur.dioxide** |           0.19 |      0.01 |
## +-------------------------+----------------+-----------+
## | **total.sulfur.dioxide** |            0.2 |      0.05 |
## +-------------------------+----------------+-----------+
## | **density**             |       **0.36** |       0.2 |
## +-------------------------+----------------+-----------+
## | **pH**                  |          -0.09 |     -0.27 |
## +-------------------------+----------------+-----------+
## | **sulphates**           |           0.01 |  **0.37** |
```

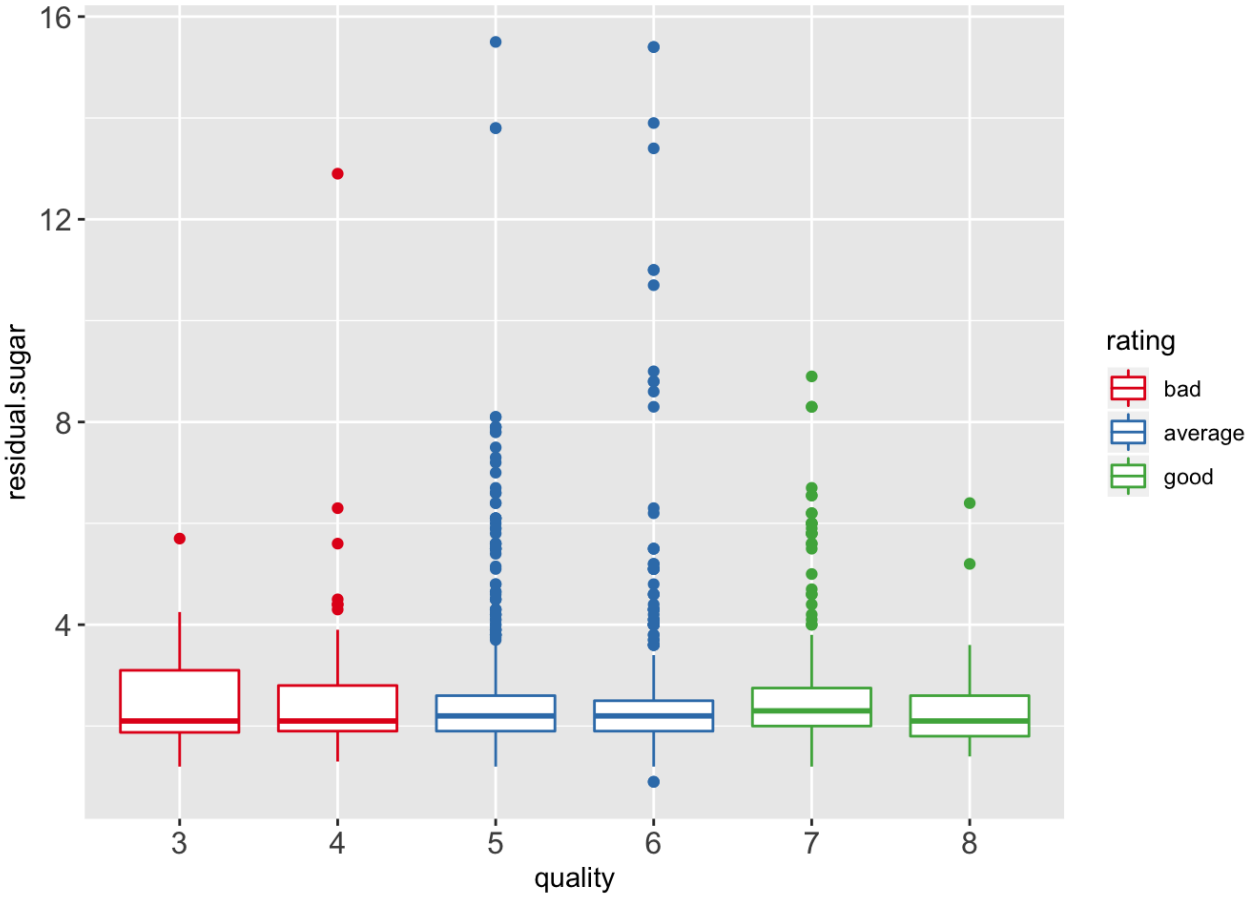
```
## +--------------------------+----------------+-----------+
## | **alcohol**              |           0.04 |     -0.22 |
## +--------------------------+----------------+-----------+
## | **quality**              |           0.01 |     -0.13 |
## +--------------------------+----------------+-----------+
##
## Table: Table continues below
##
##
##
## +--------------------------+--------------------+----------------------+
## |                     | free.sulfur.dioxide | total.sulfur.dioxide |
## +==========================+====================+======================+
## | **fixed.acidity**        |              -0.15 |                -0.11 |
## +--------------------------+--------------------+----------------------+
## | **volatile.acidity**     |              -0.01 |                 0.08 |
## +--------------------------+--------------------+----------------------+
## | **citric.acid**          |              -0.06 |                 0.04 |
## +--------------------------+--------------------+----------------------+
## | **residual.sugar**       |               0.19 |                  0.2 |
## +--------------------------+--------------------+----------------------+
## | **chlorides**            |               0.01 |                 0.05 |
## +--------------------------+--------------------+----------------------+
## | **free.sulfur.dioxide**  |                  1 |             **0.67** |
## +--------------------------+--------------------+----------------------+
## | **total.sulfur.dioxide** |            **0.67** |                    1 |
## +--------------------------+--------------------+----------------------+
## | **density**              |              -0.02 |                 0.07 |
## +--------------------------+--------------------+----------------------+
## | **pH**                   |               0.07 |                -0.07 |
## +--------------------------+--------------------+----------------------+
## | **sulphates**            |               0.05 |                 0.04 |
## +--------------------------+--------------------+----------------------+
## | **alcohol**              |              -0.07 |                -0.21 |
## +--------------------------+--------------------+----------------------+
## | **quality**              |              -0.05 |                -0.19 |
## +--------------------------+--------------------+----------------------+
##
## Table: Table continues below
##
##
##
## +--------------------------+-----------+-----------+-----------+----------+
## |                     |   density |        pH | sulphates |  alcohol |
## +==========================+===========+===========+===========+==========+
## | **fixed.acidity**        |  **0.67** | **-0.68** |      0.18 |    -0.06 |
## +--------------------------+-----------+-----------+-----------+----------+
## | **volatile.acidity**     |      0.02 |      0.23 |     -0.26 |     -0.2 |
## +--------------------------+-----------+-----------+-----------+----------+
## | **citric.acid**          |  **0.36** | **-0.54** |  **0.31** |     0.11 |
## +--------------------------+-----------+-----------+-----------+----------+
## | **residual.sugar**       |  **0.36** |     -0.09 |      0.01 |     0.04 |
## +--------------------------+-----------+-----------+-----------+----------+
## | **chlorides**            |       0.2 |     -0.27 |  **0.37** |    -0.22 |
## +--------------------------+-----------+-----------+-----------+----------+
## | **free.sulfur.dioxide**  |     -0.02 |      0.07 |      0.05 |    -0.07 |
## +--------------------------+-----------+-----------+-----------+----------+
```

```
## | **total.sulfur.dioxide** |      0.07 |     -0.07 |      0.04 |    -0.21 |
## +--------------------------+-----------+-----------+-----------+----------+
## | **density**              |         1 | **-0.34** |      0.15 | **-0.5** |
## +--------------------------+-----------+-----------+-----------+----------+
## | **pH**                   | **-0.34** |         1 |      -0.2 |     0.21 |
## +--------------------------+-----------+-----------+-----------+----------+
## | **sulphates**            |      0.15 |      -0.2 |         1 |     0.09 |
## +--------------------------+-----------+-----------+-----------+----------+
## | **alcohol**              |  **-0.5** |      0.21 |      0.09 |        1 |
## +--------------------------+-----------+-----------+-----------+----------+
## | **quality**              |     -0.17 |     -0.06 |      0.25 | **0.48** |
## +--------------------------+-----------+-----------+-----------+----------+
##
## Table: Table continues below
##
##
##
## +--------------------------+-----------+
## |                     |   quality |
## +==========================+===========+
## | **fixed.acidity**        |      0.12 |
## +--------------------------+-----------+
## | **volatile.acidity**     | **-0.39** |
## +--------------------------+-----------+
## | **citric.acid**          |      0.23 |
## +--------------------------+-----------+
## | **residual.sugar**       |      0.01 |
## +--------------------------+-----------+
## | **chlorides**            |     -0.13 |
## +--------------------------+-----------+
## | **free.sulfur.dioxide**  |     -0.05 |
## +--------------------------+-----------+
## | **total.sulfur.dioxide** |     -0.19 |
## +--------------------------+-----------+
## | **density**              |     -0.17 |
## +--------------------------+-----------+
## | **pH**                   |     -0.06 |
## +--------------------------+-----------+
## | **sulphates**            |      0.25 |
## +--------------------------+-----------+
## | **alcohol**              |  **0.48** |
## +--------------------------+-----------+
## | **quality**              |         1 |
## +--------------------------+-----------+
```

Second step is to create and leverage a function to programmatically generate summary tables alongside selected bivariate plots between these variables for analysis of the selected correlations.

Residual Sugar and Quality

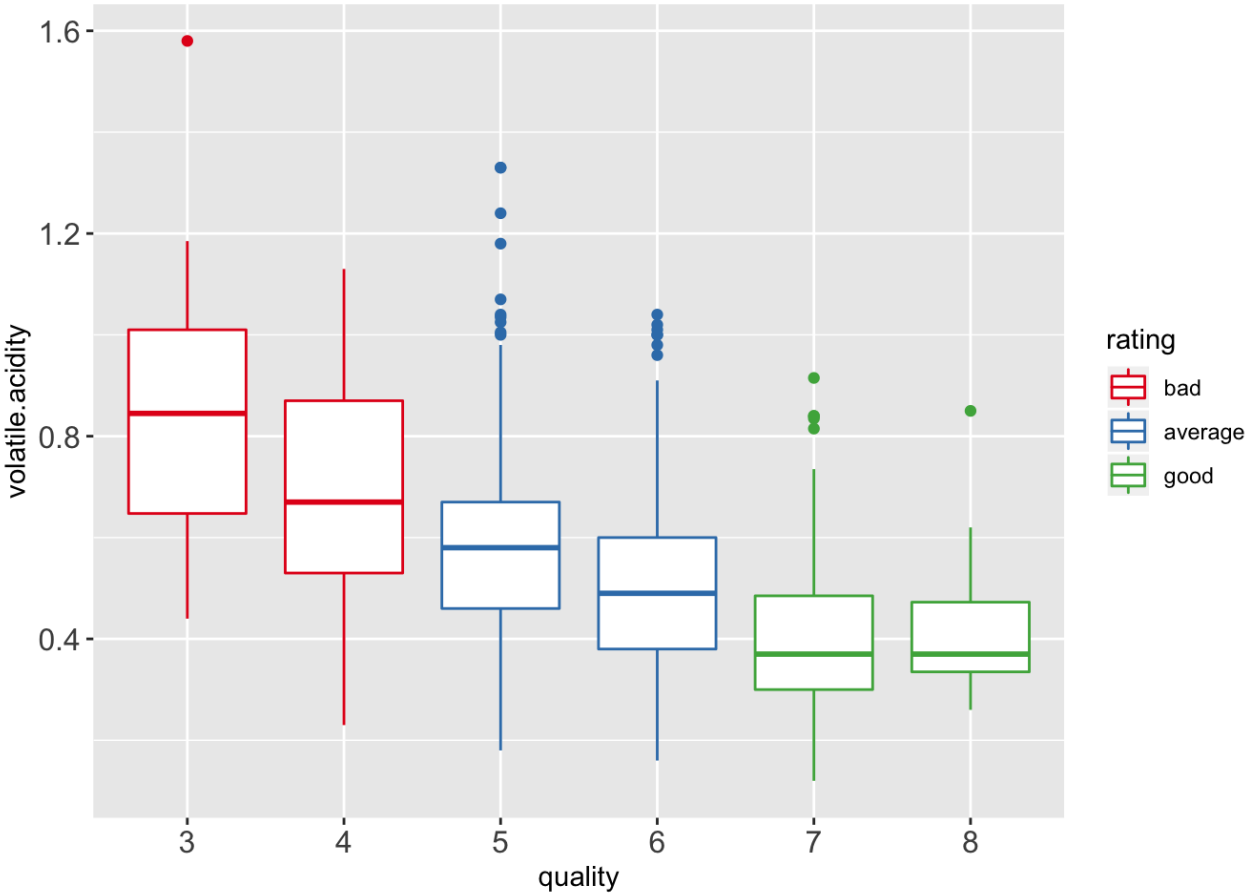
Surprisingly residual sugar and quality had a very weak positive quality correlation of only 0.01. This suggests residual sugar impact on quality is negligible.




```
##
## --------------------------
##  rating    mean    median
## --------- ------- --------
##    bad     2.685    2.1
##
##  average   2.504    2.2
##
##   good     2.709    2.3
## --------------------------
##
## Table: Summaries for residual.sugar grouped by rating
```

Volatile Acidity and Quality

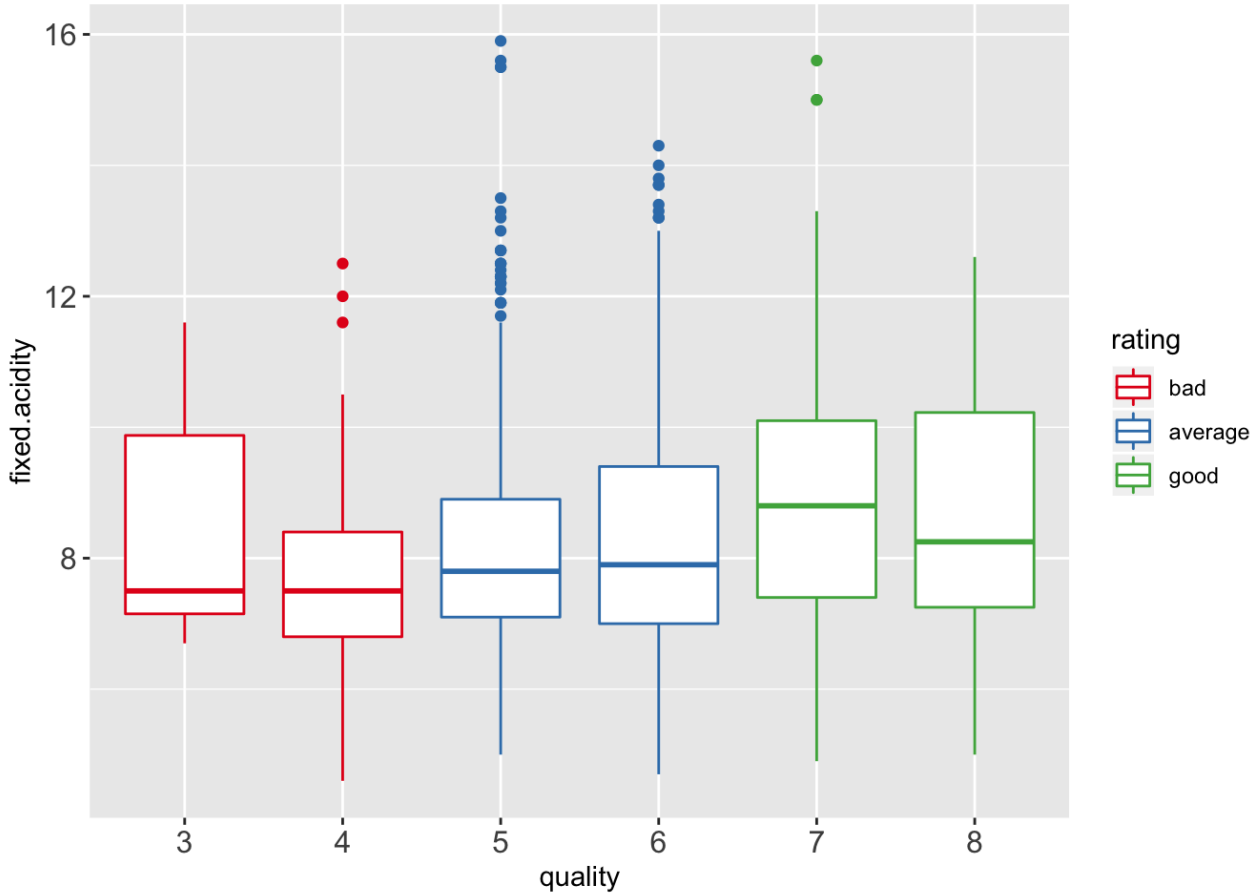
Volatile acidity and quality have a -0.39 moderate negative correlation. This suggests red wine quality decreases as volatile acidity increases.




```
##
## ---------------------------
##  rating     mean    median
## --------- -------- --------
##    bad     0.7242    0.68
##
##  average   0.5386    0.54
##
##   good     0.4055    0.37
## ---------------------------
##
## Table: Summaries for volatile.acidity grouped by rating
```

Fixed Acidity and Quality

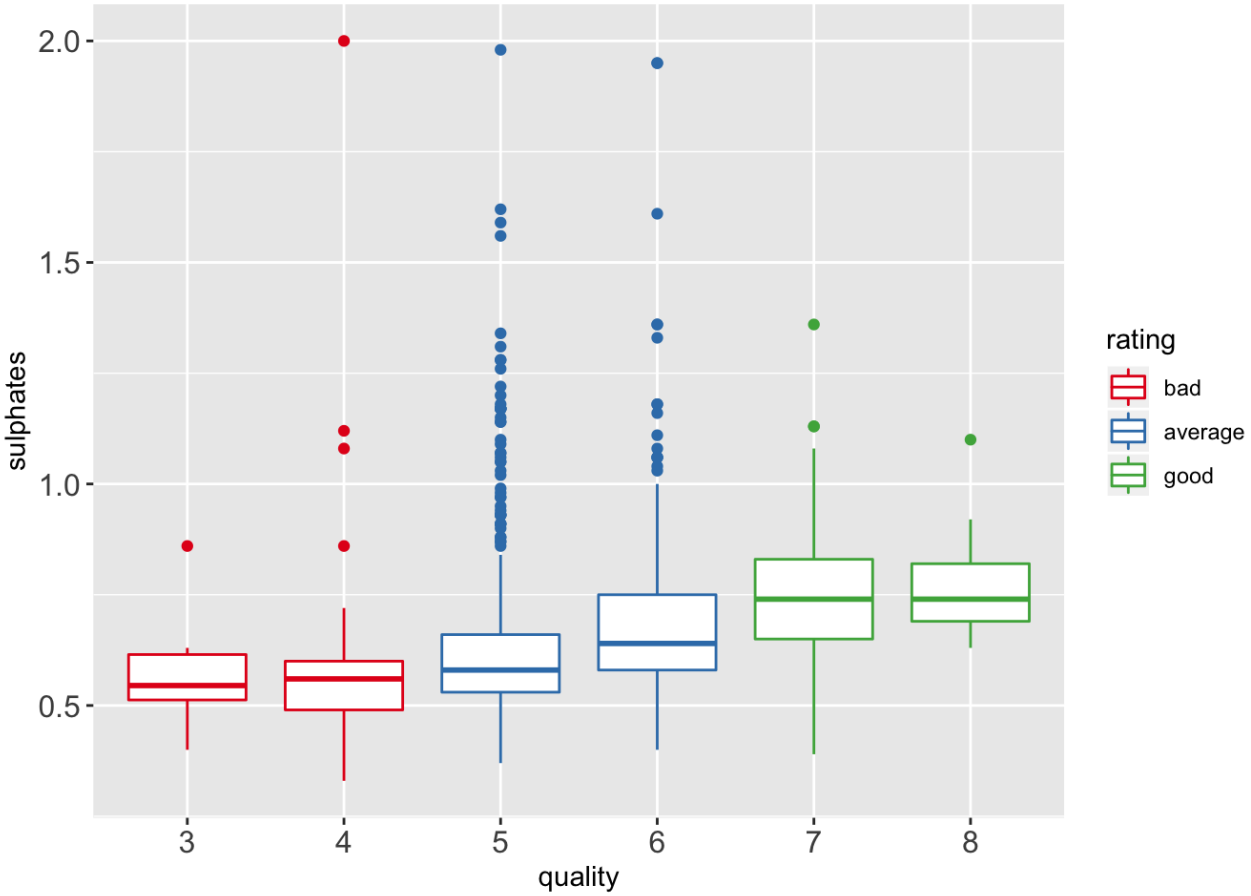
However, fixed acidity and quality has a weak positive correlation of 0.11. Fixed Acidity may have some minimal impact on wine quality.



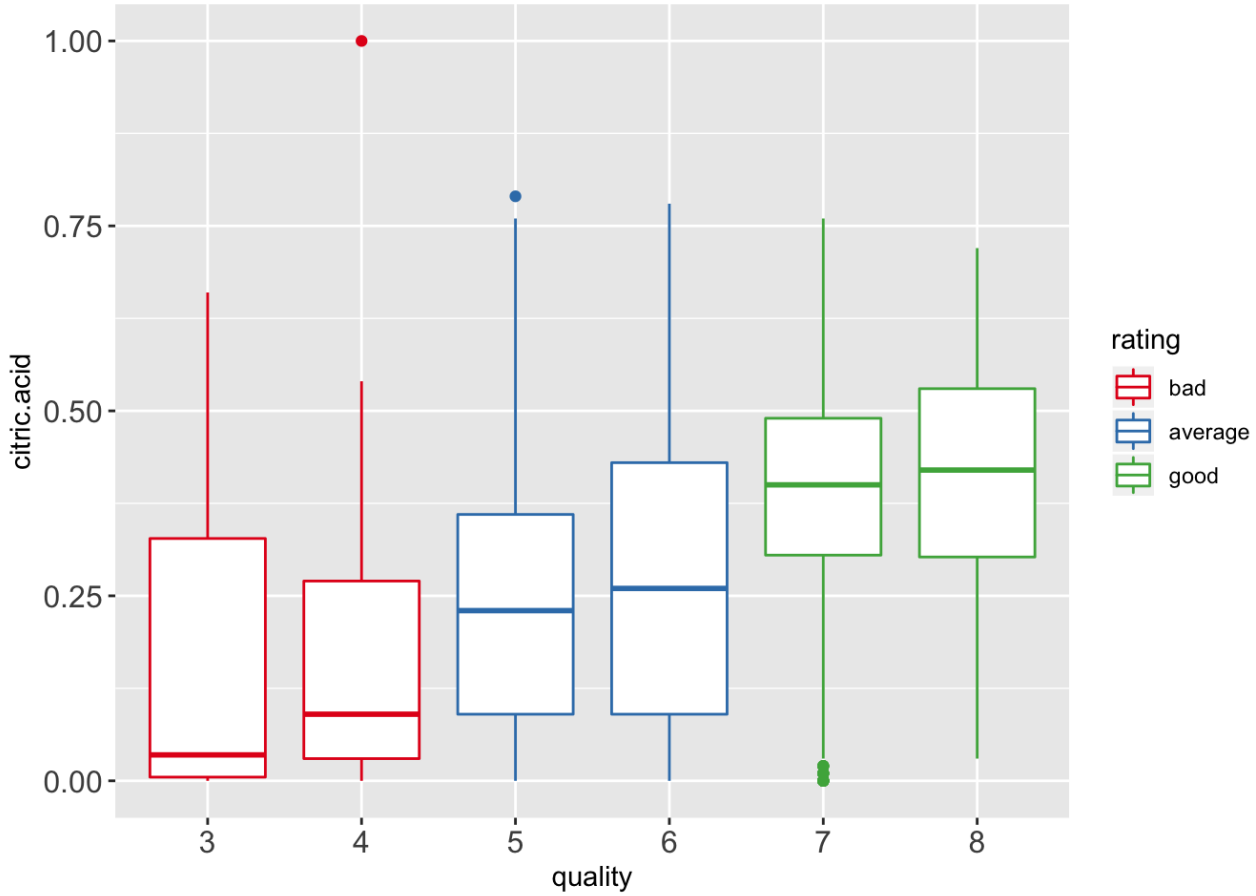

```
##
## --------------------------
##  rating    mean    median
## --------- ------- --------
##    bad     7.871    7.5
##
##  average   8.254    7.8
##
##   good     8.847    8.7
## --------------------------
##
## Table: Summaries for fixed.acidity grouped by rating
```

Quality and Sulphates & Quality and Citric acid.

Furthermore, there are weak positive correlations for both (1) quality and sulphates at .25 and (2) quality and citric acid at .23. Also, ratings trends in the same direction for both. This suggests that better wines may have a stronger concentration of sulphates; and also, better wines may have higher citric acid.



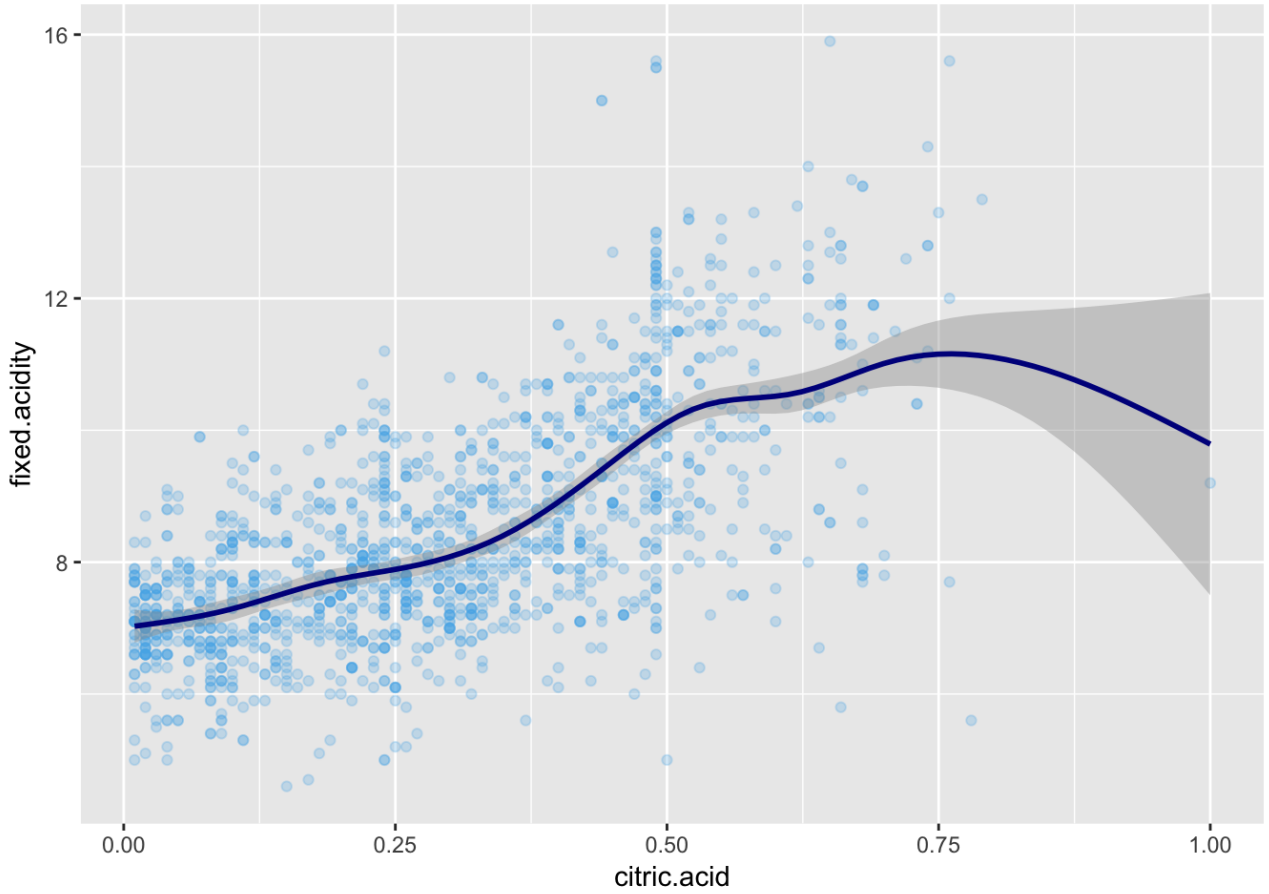

```
##
## ----------------------------
##  rating      mean     median
## --------- -------- --------
##    bad      0.5922     0.56
##
##  average    0.6473     0.61
##
##   good      0.7435     0.74
## ----------------------------
##
## Table: Summaries for sulphates grouped by rating
```




```
##
## ----------------------------
##  rating      mean     median
## --------- -------- --------
##    bad      0.1737     0.08
##
##  average    0.2583     0.24
##
##   good      0.3765     0.4
## ----------------------------
##
## Table: Summaries for citric.acid grouped by rating
```

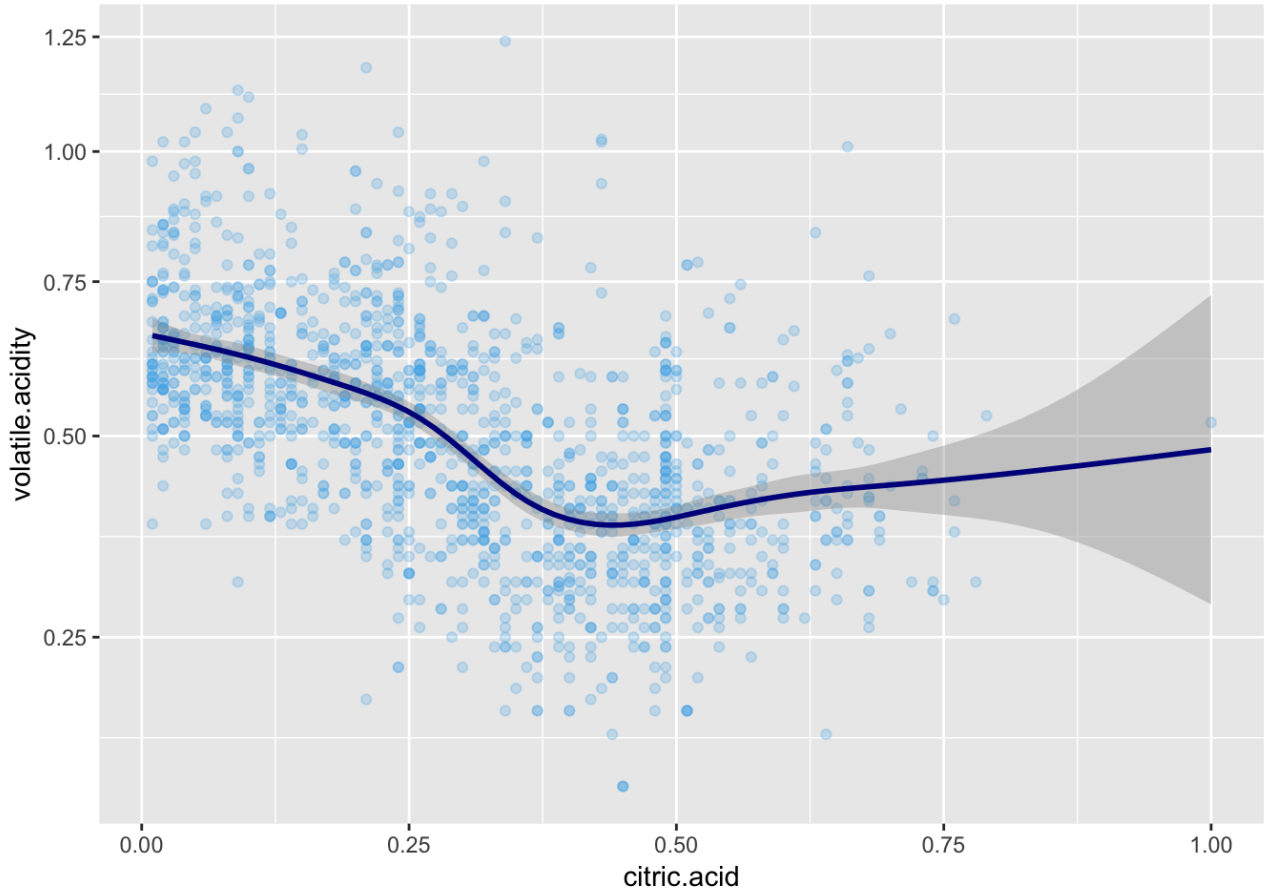
Citric Acid and Fixed Acidity

As expected, citric acid and fixed acidity have a strong positive correlation of 0.67



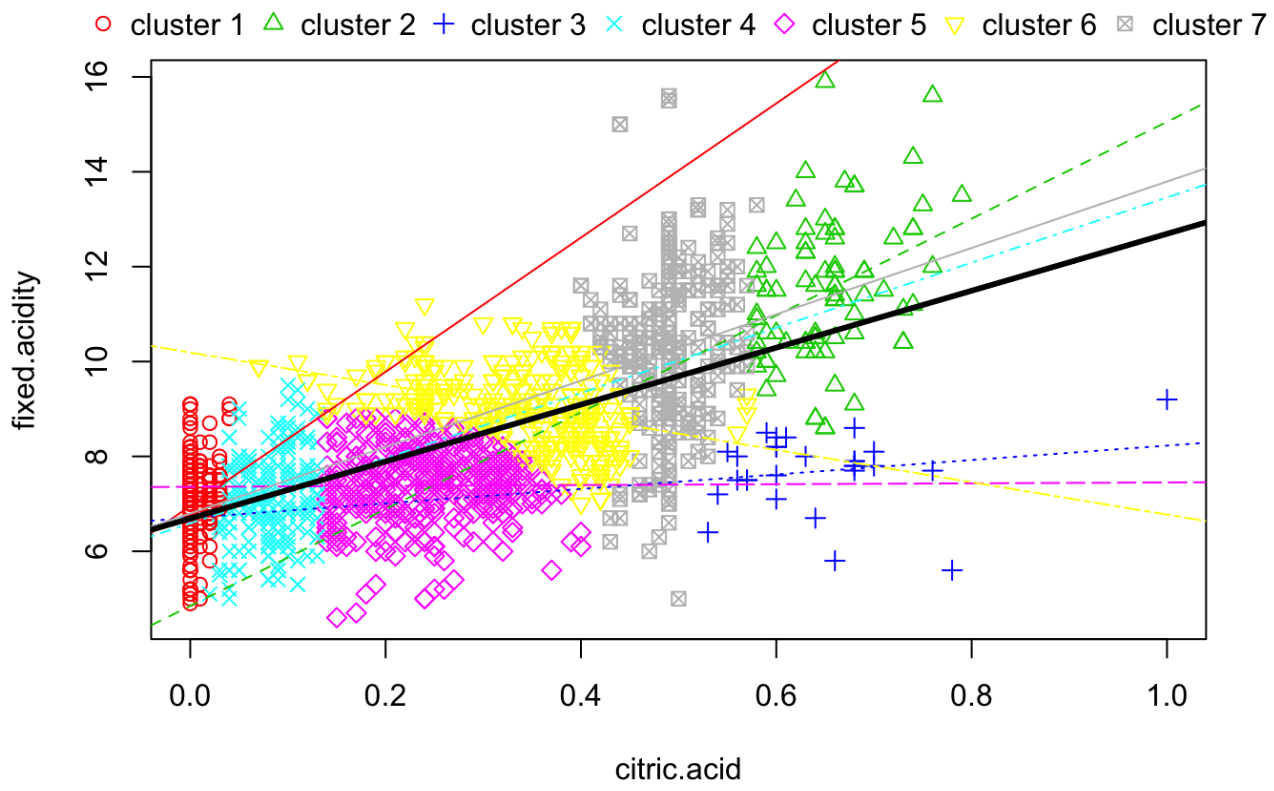
Moderate Negative Correlation between Volatile Acidity and Citric Acid

Also, you can see a moderate negative correlation between volatile acidity and citric acid where the volatile acidity y values scale with the square root function. This is not so surprising given that we observed that red wine quality decreases as volatile acidity increases, and better wines have higher citric acid.



The Simpson function against citric acid and fixed acidity detects 7 clusters.

Running the Simpson function against citric acid and fixed acidity detected 7 clusters. Only two clusters correlated in the same direction as the group. The overall trend for the subgroups reversed or disappeared when the subgroups were combined.



Bivariate Analysis

Some of the relationships observed

Volatile acidity and quality have a -0.39 moderate negative correlation. This suggests red wine quality decreases as volatile acidity increases. However, fixed acidity and quality have a weak positive correlation. Fixed Acidity has almost no impact on wine quality. Furthermore, there are weak positive correlations for both (1) quality and sulphates and (2) quality and citric acid. Also, ratings trends in the same direction for both. This suggests that better wines have a stronger concentration of sulphates. Also, better wines have higher citric acid. As expected, citric acid and fixed acidity have a strong positive correlation of 0.67. Also, you can see a moderate negative correlation between volatile acidity and citric acid where the volatile acidity y values scale with the square root function. This is not so surprising given that we observed that red wine quality decreases as volatile acidity increases, and better wines have higher citric acid.

Interesting Relationships

Simpson's Paradox (https://en.wikipedia.org/wiki/Simpson%27s_paradox) is a "phenomenon in probability and statistics, in which a trend appears in several different groups of data but disappears or reverses when these groups are combined." Running the Simpson function against citric acid and fixed acidity detected 7 clusters. Only two clusters correlated in the same direction as the group. The overall trend for the subgroups reversed or disappeared when the subgroups were combined.

Strongest Relationship

The strongest relationship this analysis focused on was that between citric acid and fixed volatility.

Multivariate Plots Section

Lets use multivariate plots to answer some questions that came to light from the above bivariate plot analysis and to look for other relationships in the data.

---

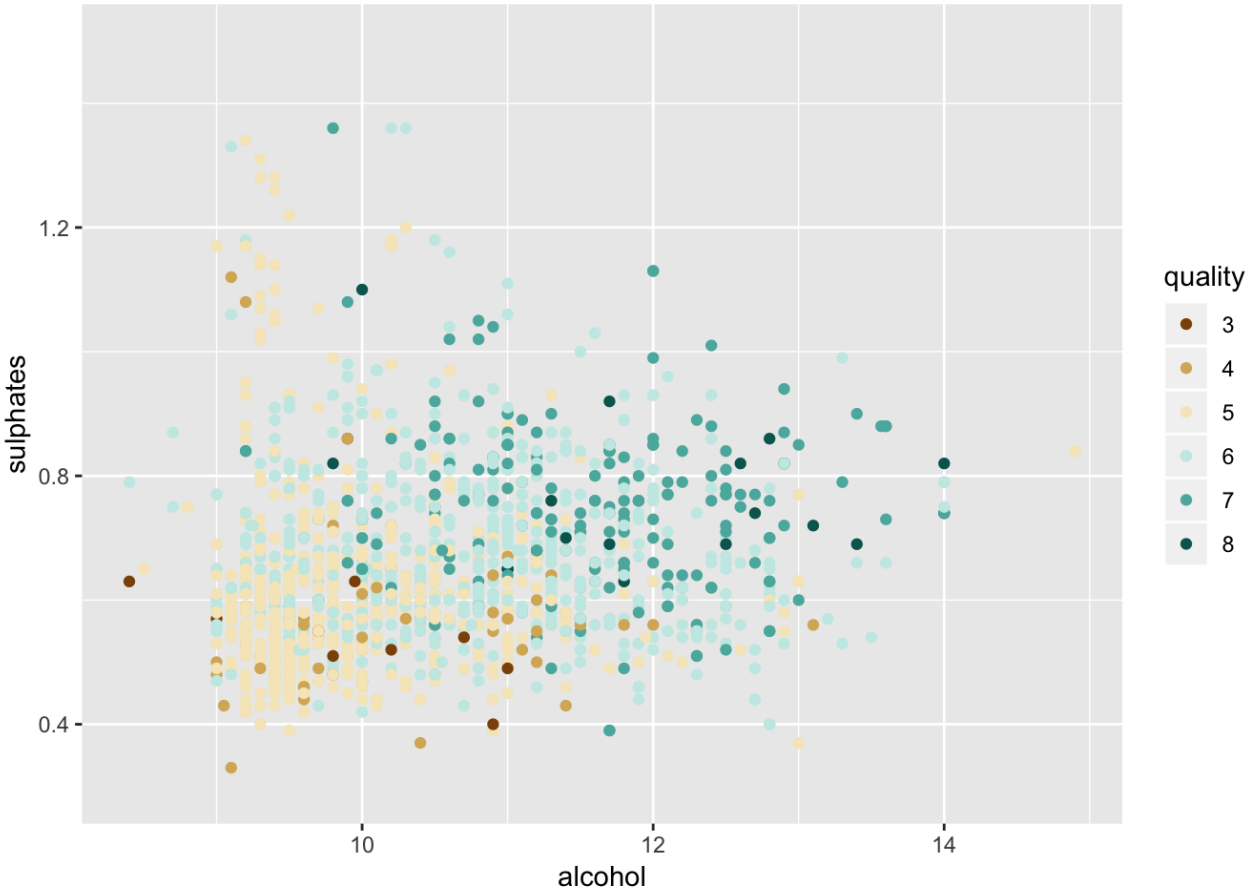
pH and Acid Concentration

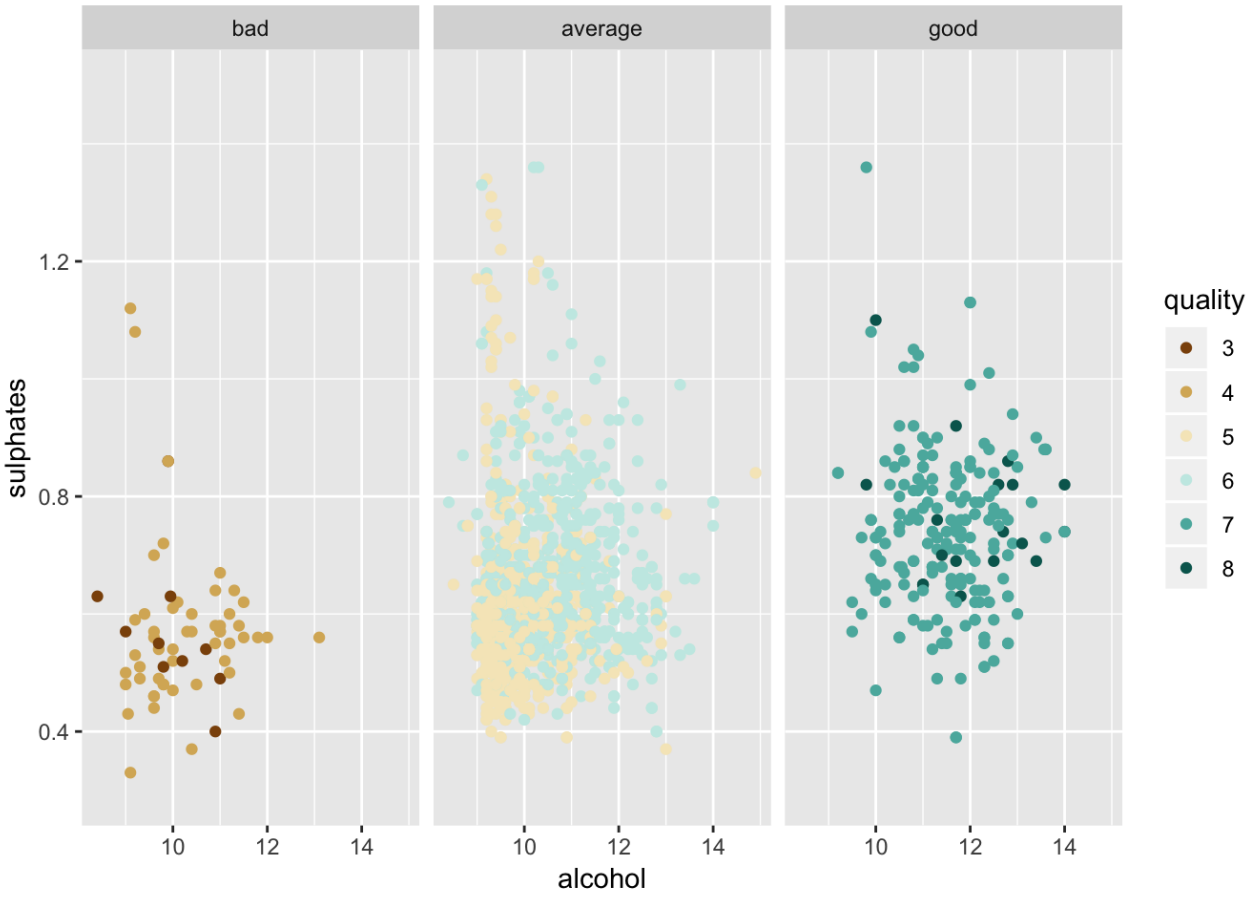
pH measures acid concentration using a log scale. Therefore, there are stronger correlations between pH and the log of the acid concentrations. We use a linear model to investigate how much of the variance in pH is explained by citric acid, fixed acidity, and volatile acidity. With R-squared equal to 0.4876, it seems that the three acidity variables can only explain about half the variance in pH. This suggests there are other more relevant variables that affect acidity.

```
##
## Call:
## lm(formula = pH ~ I(log10(citric.acid)) + I(log10(volatile.acidity)) +
##     I(log10(fixed.acidity)), data = subset(rw, citric.acid >
##     0))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.47184 -0.06318 -0.00003  0.06447  0.32265
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 4.230862   0.040578 104.266  < 2e-16 ***
## I(log10(citric.acid))      -0.052187   0.008797  -5.933 3.72e-09 ***
## I(log10(volatile.acidity)) -0.049788   0.021248  -2.343   0.0193 *
## I(log10(fixed.acidity))    -1.071983   0.038987 -27.496  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1068 on 1463 degrees of freedom
## Multiple R-squared:  0.4876, Adjusted R-squared:  0.4866
## F-statistic: 464.1 on 3 and 1463 DF,  p-value: < 2.2e-16
```

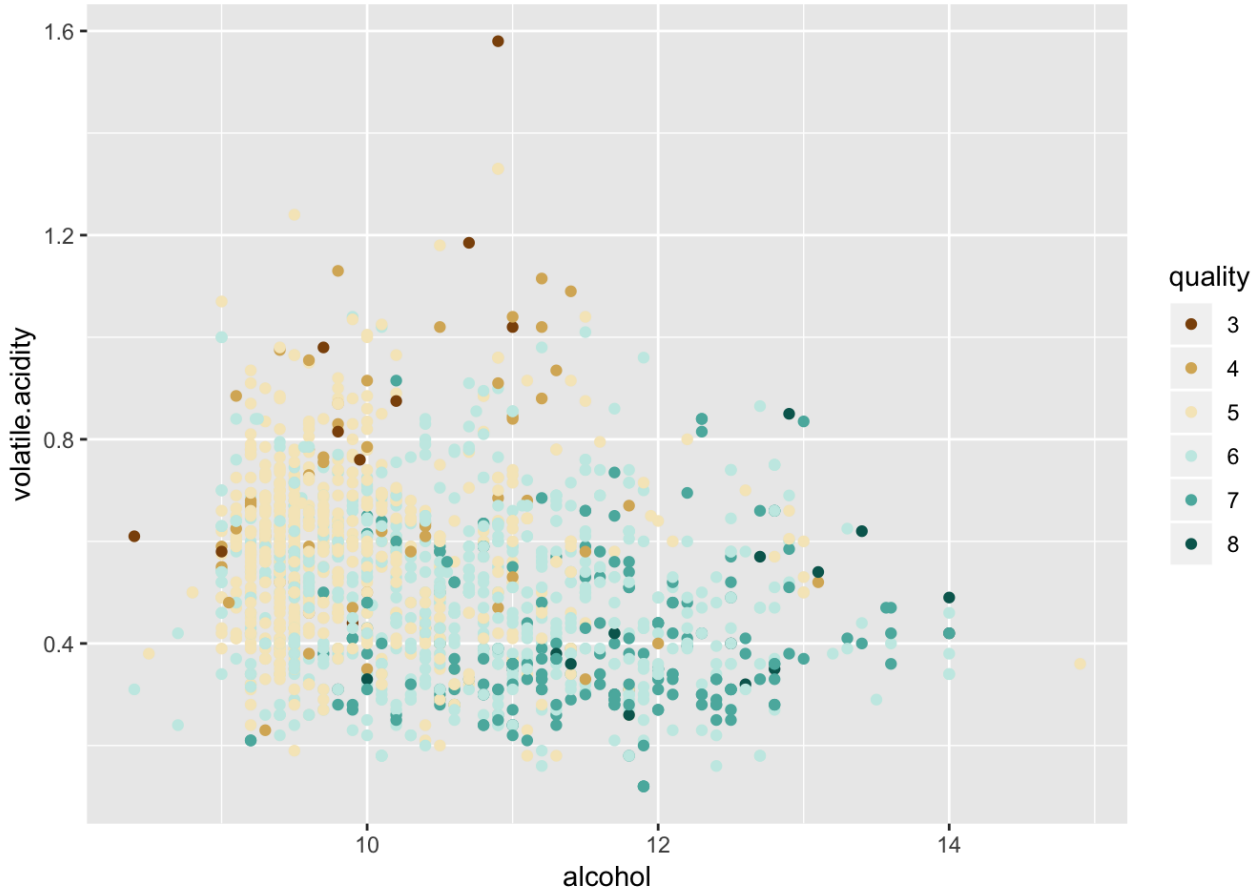
Alcohol Content and Sulphates

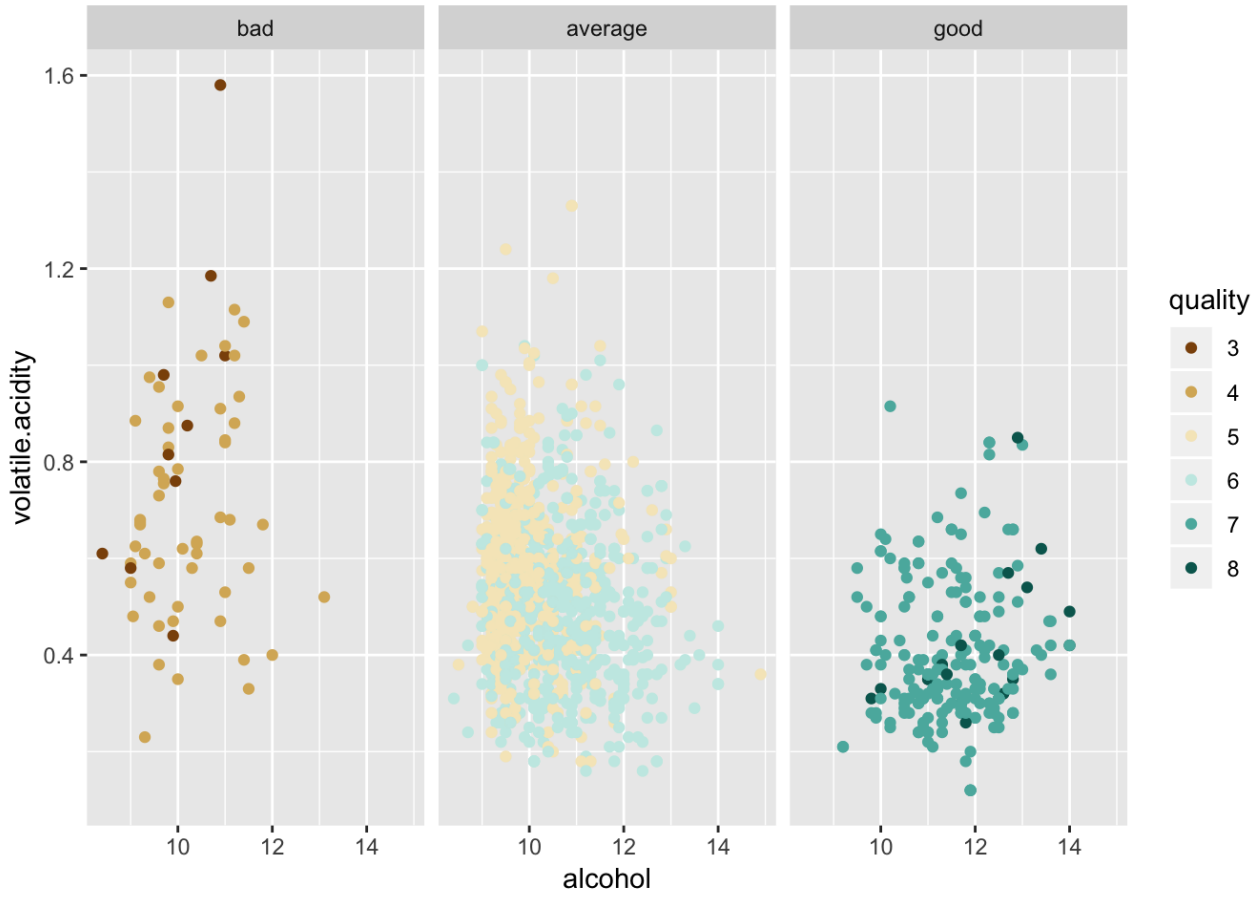
The plots below indicate that for wines with high alcohol content, having a higher concentration of sulphates produces better wines.





Also, the inverse seems to hold for acidity. For example, having less volatile acidity on higher concentrations of alcohol seems to produce better wines.



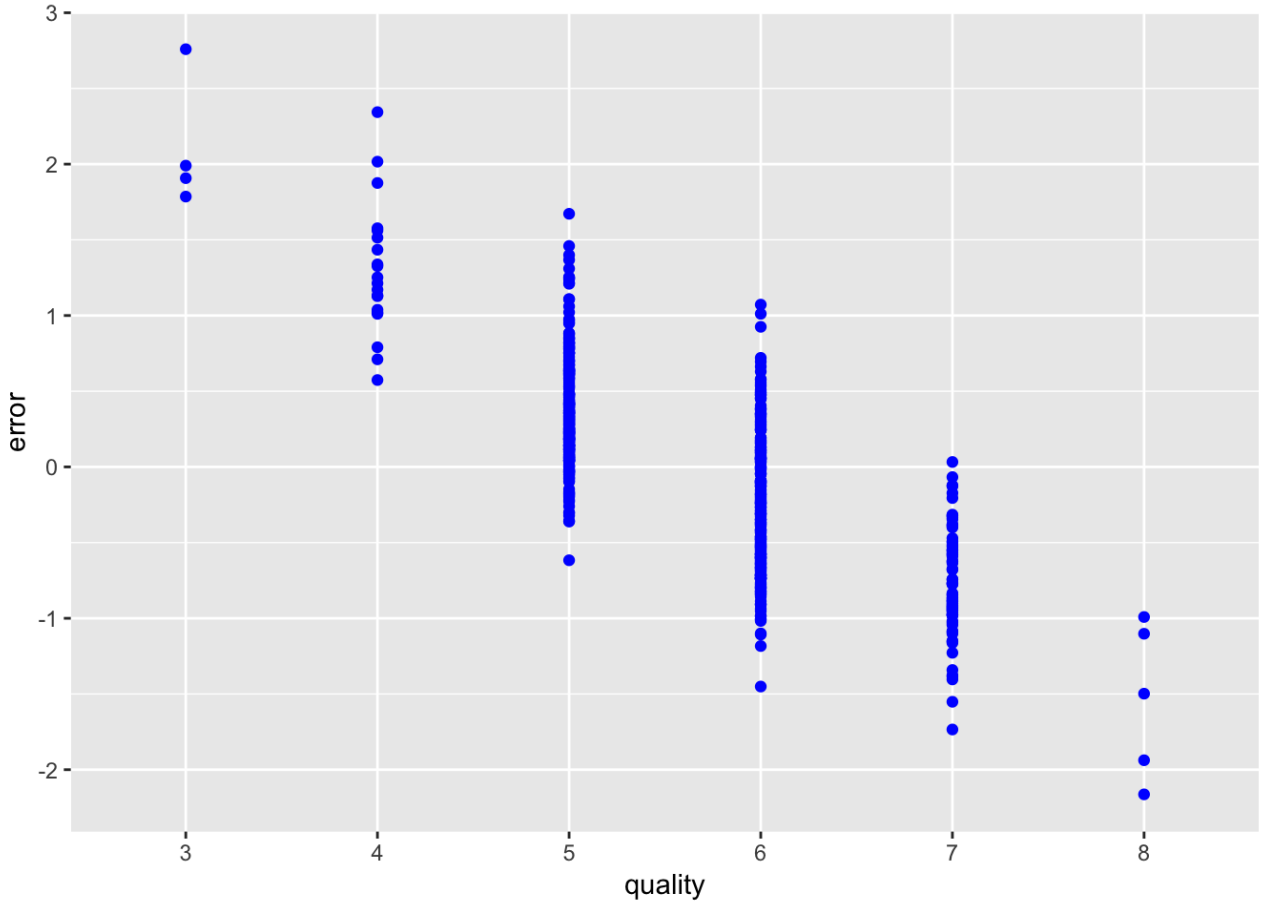


Key Variables Linear Models

A few selected key variables (alcohol, sulphates, and acidity) were used to generate some linear models for comparison. The pH variable was excluded to avoid issues with perfect multicollinearity (also collinearity). Multicollinearity is a “phenomenon in which one predictor variable in a multiple regression model can be linearly predicted from the others with a substantial degree of accuracy.” The results were disappointing, with R-squared (https://en.wikipedia.org/wiki/Coefficient_of_determination), the coefficient of determination statistic used to measure the proportion of the variance in the dependent variable that is explained by the independent variable(s), coming in low (highest reported R-square figure was .348)

```
##
## Calls:
## m1: lm(formula = as.numeric(quality) ~ alcohol, data = training_data)
## m2: lm(formula = as.numeric(quality) ~ alcohol + sulphates, data = training_data)
## m3: lm(formula = as.numeric(quality) ~ alcohol + sulphates + volatile.acidity,
##     data = training_data)
## m4: lm(formula = as.numeric(quality) ~ alcohol + sulphates + volatile.acidity +
##     citric.acid, data = training_data)
## m5: lm(formula = as.numeric(quality) ~ alcohol + sulphates + volatile.acidity +
##     citric.acid + fixed.acidity, data = training_data)
## m6: lm(formula = as.numeric(quality) ~ alcohol + sulphates + pH,
##     data = training_data)
##
## ============================================================================================
==========
##                          m1            m2           m3           m4           m5
m6
## --------------------------------------------------------------------------------------------
----------
##   (Intercept)          -0.369        -0.877***     0.197        0.227       -0.246
1.270*
##                       (0.228)       (0.233)      (0.255)      (0.261)      (0.292)
(0.508)
##   alcohol              0.384***      0.370***     0.341***     0.341***     0.352***
0.396***
##                       (0.022)       (0.021)      (0.021)      (0.021)      (0.021)
(0.022)
##   sulphates                          0.990***     0.716***     0.733***     0.751***
0.839***
##                                     (0.135)      (0.134)      (0.138)      (0.137)
(0.138)
##   volatile.acidity                                -1.115***    -1.152***    -1.239***

##                                                   (0.125)      (0.145)      (0.147)

##   citric.acid                                                   -0.069       -0.509**

##                                                                 (0.135)      (0.183)

##   fixed.acidity                                                               0.061***

##                                                                             (0.017)

##   pH                                                                                    -
0.699***
##
(0.148)
## --------------------------------------------------------------------------------------------
----------
##   R-squared            0.245         0.285        0.339        0.339        0.348
0.301
##   adj. R-squared       0.244         0.283        0.337        0.337        0.345
0.299
##   sigma                0.712         0.694        0.667        0.667        0.663
0.686
##   F                  309.758       190.078      163.441      122.550      101.742      13
```

```
7.052
##   p                    0.000         0.000         0.000         0.000         0.000
0.000
##   Log-likelihood    -1034.673     -1008.598     -970.437      -970.308      -964.026      -99
7.448
##   Deviance            485.816       460.103       424.905       424.790       419.262       44
9.528
##   AIC                2075.346      2025.195      1950.874      1952.615      1942.053      200
4.896
##   BIC                2089.944      2044.659      1975.203      1981.810      1976.114      202
9.226
##   N                   959           959           959           959           959           95
9
## ============================================================================================
==========
```



Multivariate Analysis

Observed Relationships

We used multivariate plots to answer some questions that came to light from the earlier bivariate plot analysis and to look for other relationships in the data. We know that pH measures acid concentration using a log scale; therefore, there are stronger correlations between pH and the log of the acid concentrations. We used a linear model to investigate how much of the variance in pH is explained by citric acid, fixed acidity, and volatile acidity. With R-squared equal to 0.4876, it seems that the three acidity variables can only explain about half the variance in pH. This suggests there are other more relevant variables that affect acidity.

Interesting or Surprising Interactions between Features

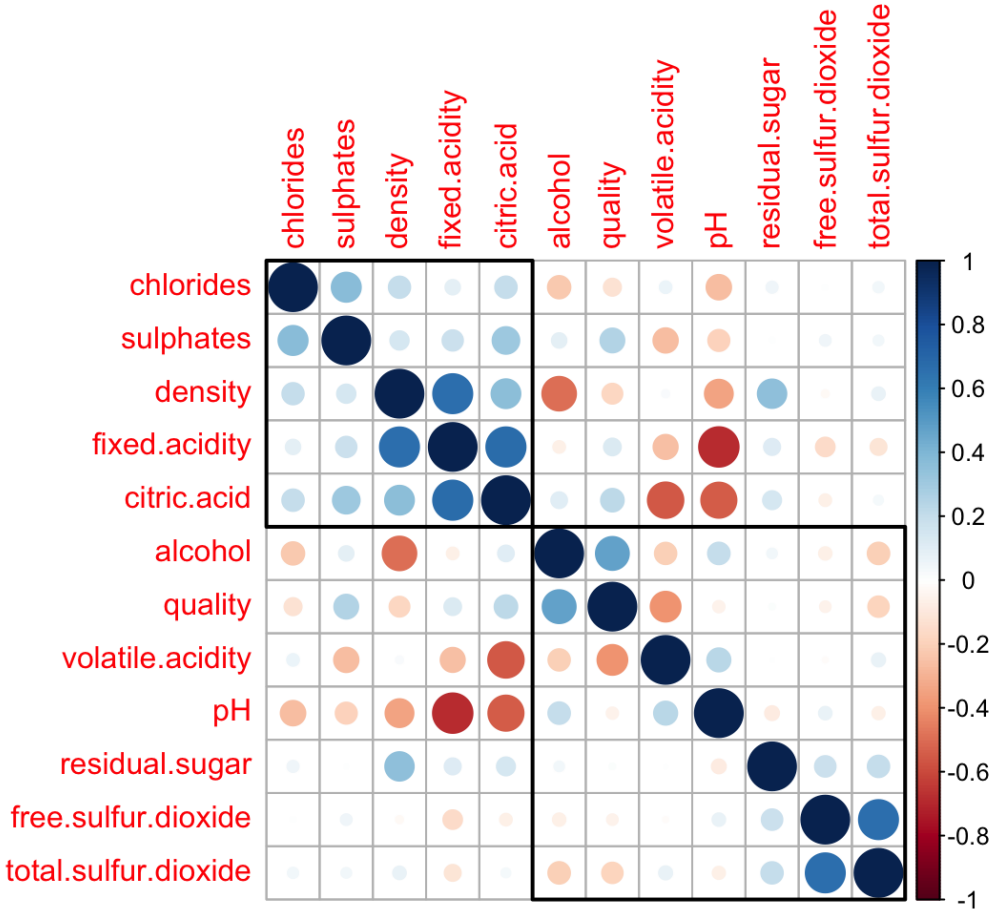
It was shown that wines with high alcohol content, having a higher concentration of sulphates produces better wines. Also, the inverse seems to hold for acidity. For example, having less volatile acidity on higher concentrations of alcohol seems to produce better wines.

Linear Models

We also used some key variables (alcohol, sulphates, and acidity) to generate a few linear models for comparison. The pH variable was excluded to avoid issues with perfect multicollinearity (also collinearity). Multicollinearity (https://en.wikipedia.org/wiki/Multicollinearity) is a “phenomenon in which one predictor variable in a multiple regression model can be linearly predicted from the others with a substantial degree of accuracy.” Also the low R-squared scores suggest that there are missing variables that can be better used to predict quality.

Final Plots and Summary

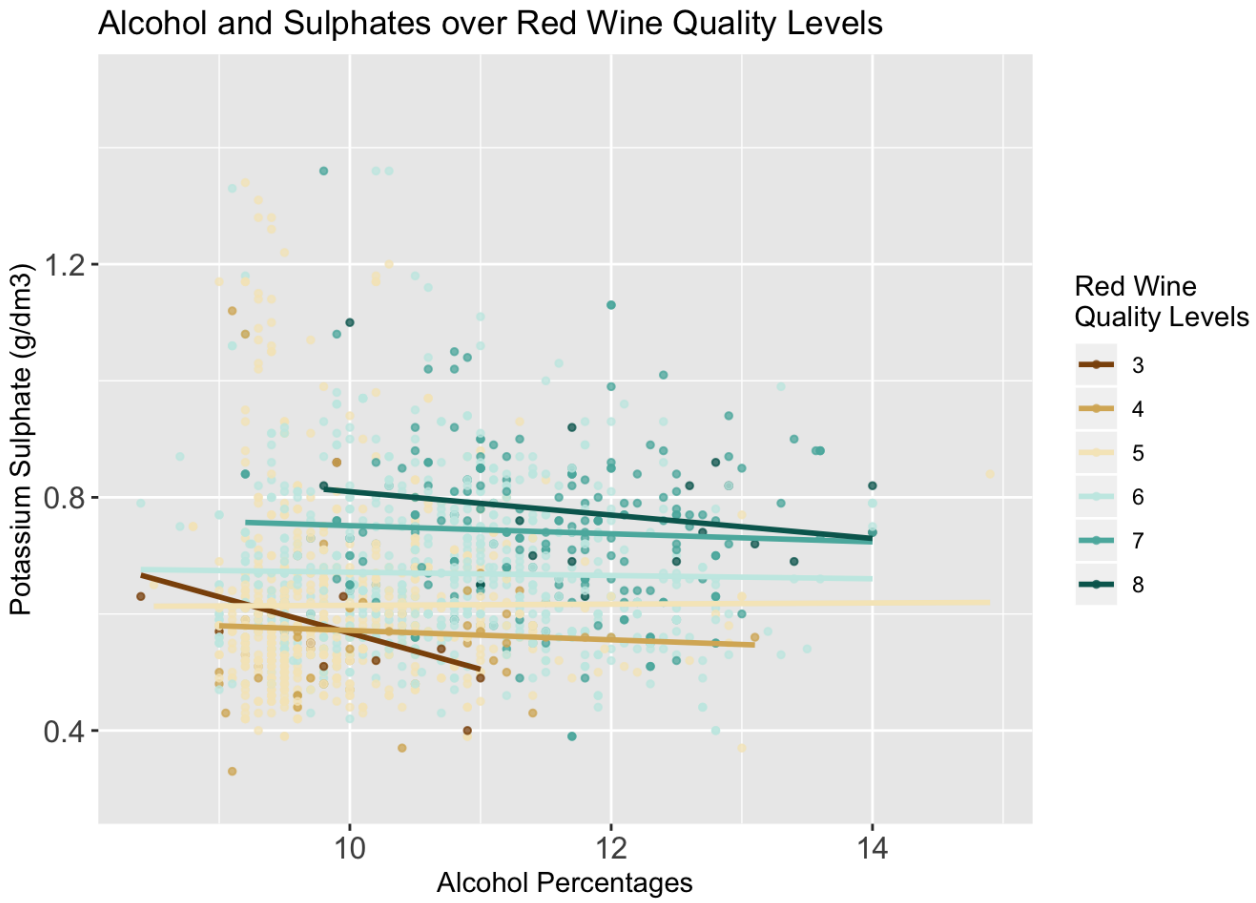
Plot One



Description One

For my initial analysis, originally a table was used to display all the correlations for the red wine data set; but plotting the correlation matrix makes it even easier to identify both positive and negative correlations greater than an absolute value of 0.2.

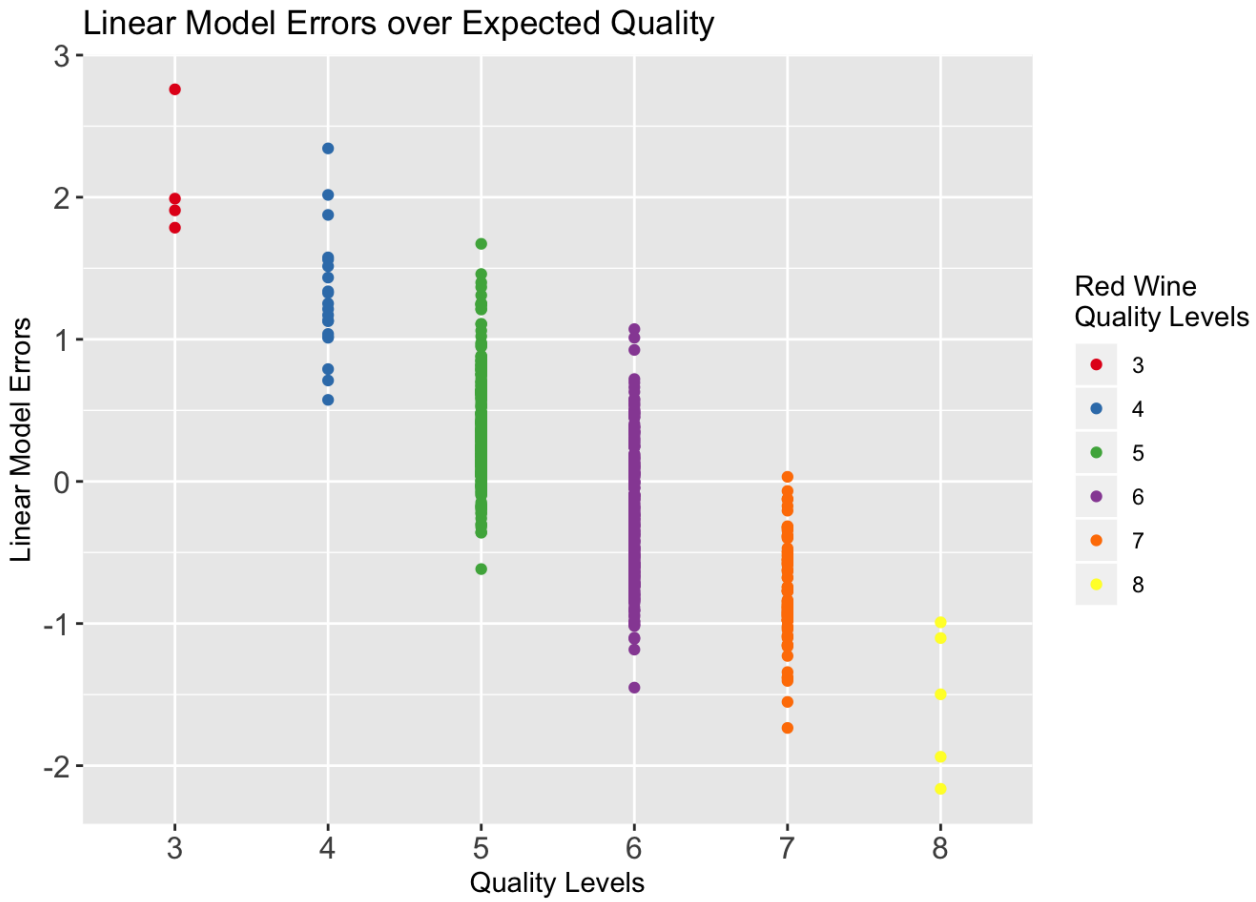
Plot Two



Description Two

This plot shows that the best quality wines have high concentrations fof both alcohol and sulphates. This implies that high alcohol contents and high sulphate concentrations together seem to produce better wines. This is something to definitely look for when shopping for red wines.

Plot Three



Description Three

The linear model with the highest R-squared value could only explain approximately 35% of the variance in quality. This plot suggests that there are missing variables needed to better predict quality wines.

Reflection

This analysis explored the univariate, bivariate, & multivariate relationships between the variables in the provided tidy Red Wine data set The data set contained information on the chemical properties of a selection of red wines. Also, a new Factored Variable named ‘Rating’ was added.

The first step was to do a univariate analysis of all the variables. A series of plots showed which variables were normally distributed and skewed. Focusing on the plots for quality and ratings showed that most wines in the dataset were of average quality. Why?

Continuing with the bivariate analysis it was shown that better wines have a stronger concentration of sulphates, together with higher counts of citric acid. Then following through with a multivariate analysis it was shown that wines with high alcohol content and having a higher concentration of sulphates produced better wines. Also, the inverse seemed to hold for acidity. For example, having less volatile acidity on higher concentrations of alcohol seemed to produce better wines. However, the generated liniear models using the alcohol, sulphates, and acidity variables did not well explain the variance in quality.

Future studies, should include larger data sets and more variables. For example, a winefolly.com (https://winefolly.com/review/understanding-acidity-in-wine/) article speaks to the importance of differences in tastes between unoaked versus oaked wines and the role the aging process plays in a wine’s malic acid conversion to lactic acid. Also, the data set can include categorized wine critic reviews and ratings, to determine if relying on such critical ratings (https://www.winespectator.com/wineratings) can truly lead to selecting better tasting wines.

Sources of Inspiration

https://stackoverflow.com/questions/7458796/how-to-suppress-qplots-binwidth-warning-inside-a-function (https://stackoverflow.com/questions/7458796/how-to-suppress-qplots-binwidth-warning-inside-a-function**)

https://cran.r-project.org/web/packages/gridExtra/vignettes/tableGrob.html (https://cran.r-project.org/web/packages/gridExtra/vignettes/tableGrob.html**)

https://rapporter.github.io/pander/pandoc_table.html (https://rapporter.github.io/pander/pandoc_table.html**)

https://github.com/pcasaretto/udacity-eda-project/blob/master/wine.Rmd (https://github.com/pcasaretto/udacity-eda-project/blob/master/wine.Rmd**)

http://www.sthda.com/english/wiki/ggplot2-colors-how-to-change-colors-automatically-and-manually (http://www.sthda.com/english/wiki/ggplot2-colors-how-to-change-colors-automatically-and-manually**)

https://rpubs.com/jasonmedina/219996 (https://rpubs.com/jasonmedina/219996**)

https://github.com/pcasaretto/udacity-eda-project/blob/master/wine.Rmd (https://github.com/pcasaretto/udacity-eda-project/blob/master/wine.Rmd**)

https://ggplot2.tidyverse.org (https://ggplot2.tidyverse.org**)