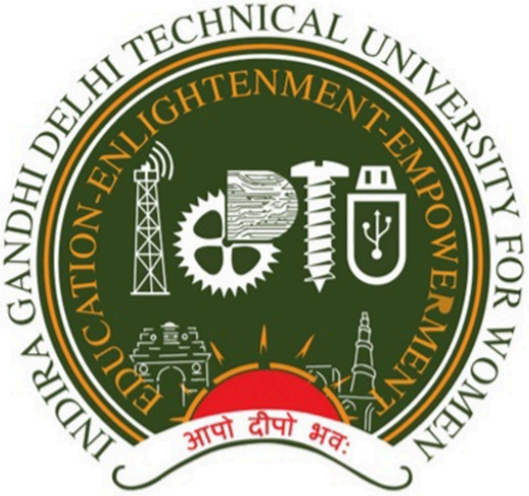
INDIRA GANDHI DELHI TECHNICAL UNIVERSITY FOR WOMEN



IT WORKSHOP PROJECT

RED WINE QUALITY
Exploratory Data Analysis

Submitted To:
Mr. Santanoo Patnaik

Submitted by:
Chhavi Verma - 019
Devanshi- 020

IT WORKSHOP

Red Wine Quality

Exploratory Data Analysis

Introduction

In this project, we will analyse the Red Wine Data and try to understand which variables are responsible for the quality of the wine.

Once viewed as a luxury good, nowadays wine is increasingly enjoyed by a wider range of consumers. Portugal is a top ten wine exporting country, with 3.17% of the market share in 2005. Exports of its vinho verde wine (from the northwest region) have increased by 36% from 1997 to 2007. To support its growth, the wine industry is investing in new technologies for both wine making and selling processes. Wine certification and quality assessment are key elements within this context. Certification prevents the illegal adulteration of wines (to safeguard human health) and assures quality for the wine market. Quality evaluation is often part of the certification process and can be used to improve wine making (by identifying the most influential factors) and to stratify wines such as premium brands (useful for setting prices).

Wine certification is generally assessed by physicochemical and sensory tests. Physicochemical laboratory tests routinely used to characterise wine include determination of density, alcohol or pH values, while sensory tests rely mainly on human experts. It should be stressed that taste is the least understood of the human senses thus wine classification is a difficult task. Moreover, the relationships between the physicochemical and sensory analysis are complex and still not fully understood.

Advances in information technologies have made it possible to collect, store and process massive, often highly complex datasets. All this data holds valuable information such as trends and patterns, which can be used to improve decision making and optimise chances of success. Data mining (DM) techniques aim at extracting high-level knowledge from raw data. There are several DM algorithms, each one with its own advantages. When modelling continuous data, the linear/multiple regression (MR) is the classic approach.

Aim of the Project

Before attempting to determine the association between the variables and the wine quality with other factors included, we will first attempt to gain a sense of the variables on their own. Finally, a linear model will be developed to forecast the results of test set data.

Code and its Explanation

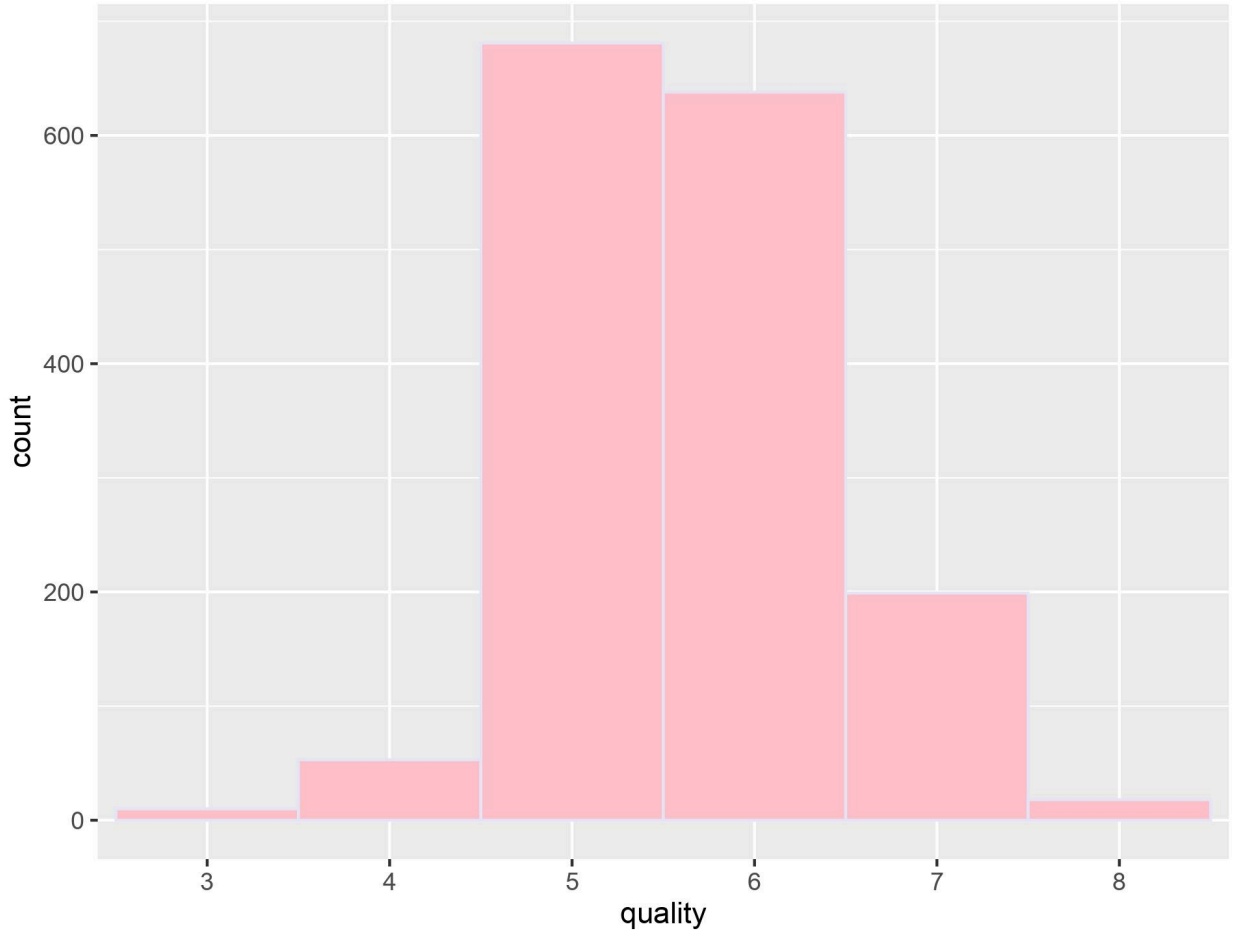
Structure and summary of the Dataframe

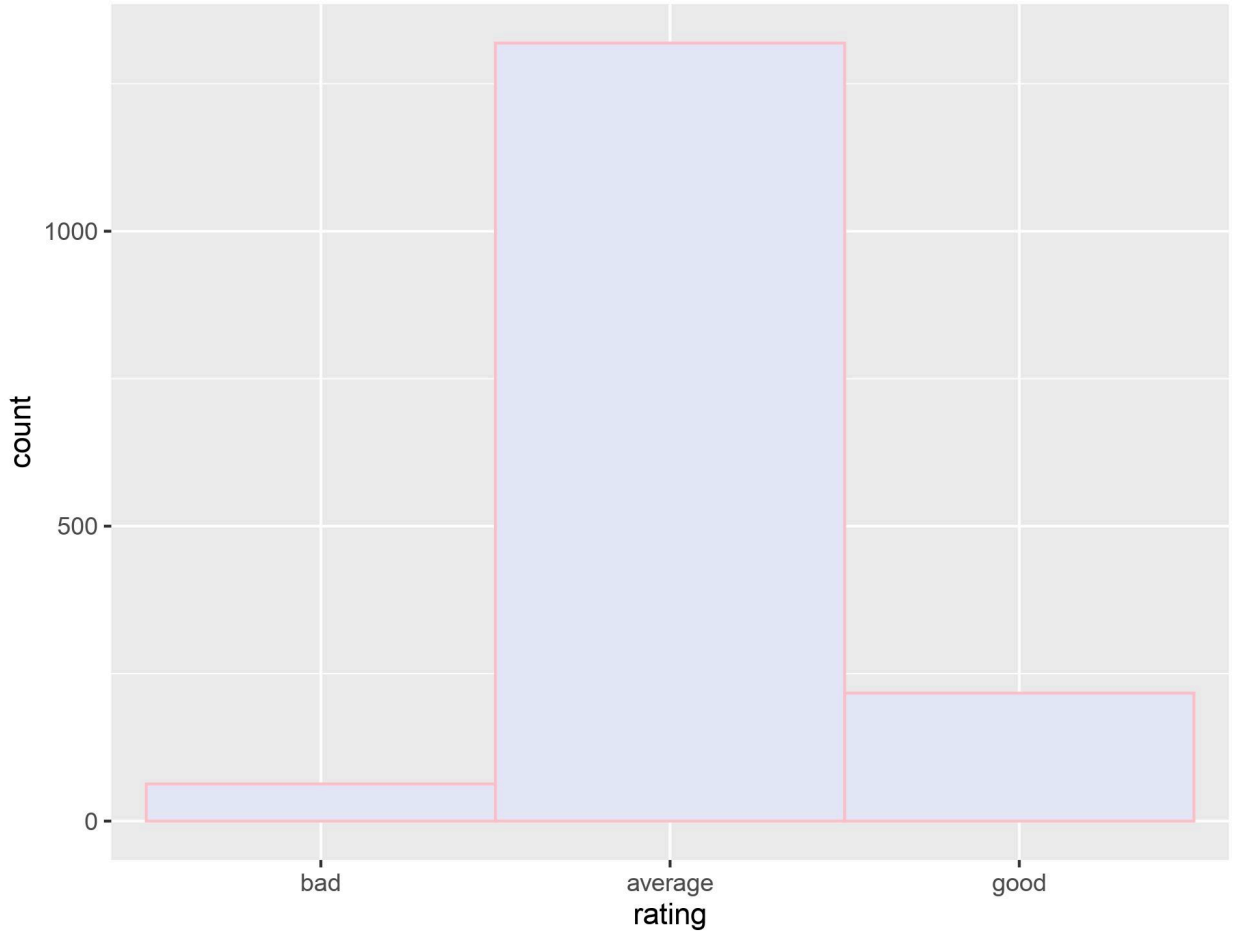
```
> summary(wine)
       X          fixed.acidity   volatile.acidity  citric.acid    residual.sugar
 Min.   :   1.0   Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900
 1st Qu.: 400.5   1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900
 Median : 800.0   Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200
 Mean   : 800.0   Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539
 3rd Qu.:1199.5   3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600
 Max.   :1599.0   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     rating
 Min.   :2.740   Min.   :0.3300   Min.   : 8.40   3: 10   bad    :  63
 1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50   4: 53   average:1319
 Median :3.310   Median :0.6200   Median :10.20   5:681   good   : 217
 Mean   :3.311   Mean   :0.6581   Mean   :10.42   6:638
 3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10   7:199
 Max.   :4.010   Max.   :2.0000   Max.   :14.90   8: 18
```

```
> str(wine)
'data.frame':   1599 obs. of  14 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             : 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 ...
```

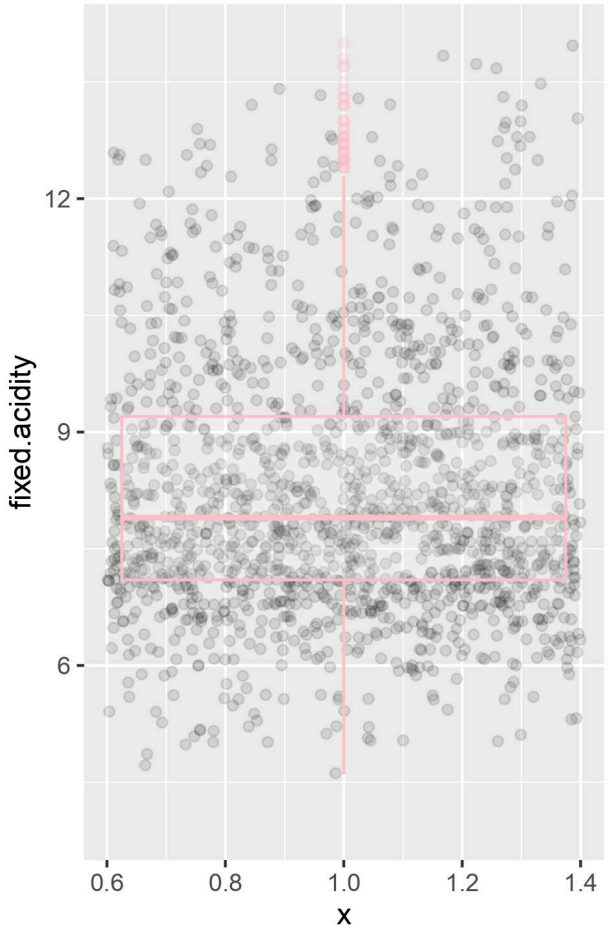
Univariate Plots

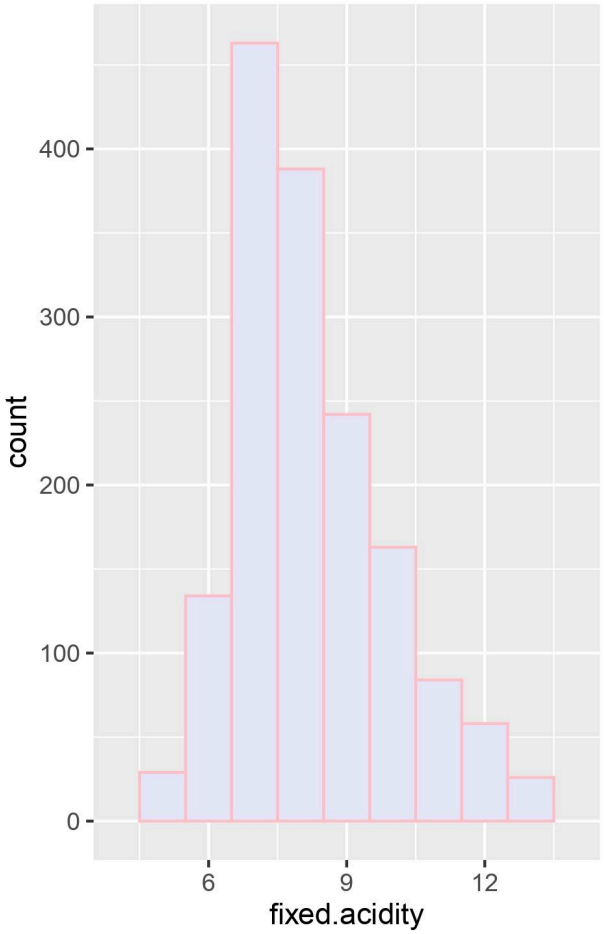
Prior to performing any analysis between the variables, we will first plot the distribution of each variable since we want to first acquire a sense of them. This will also assist us in gaining some insight into what to anticipate when we plot various variables against one another based on the distribution shape, such as Normal, Positive Skew, or Negative Skew. This dataset has severe outliers for numerous variables as well. For a more reliable analysis, we shall exclude the extreme outliers from those.



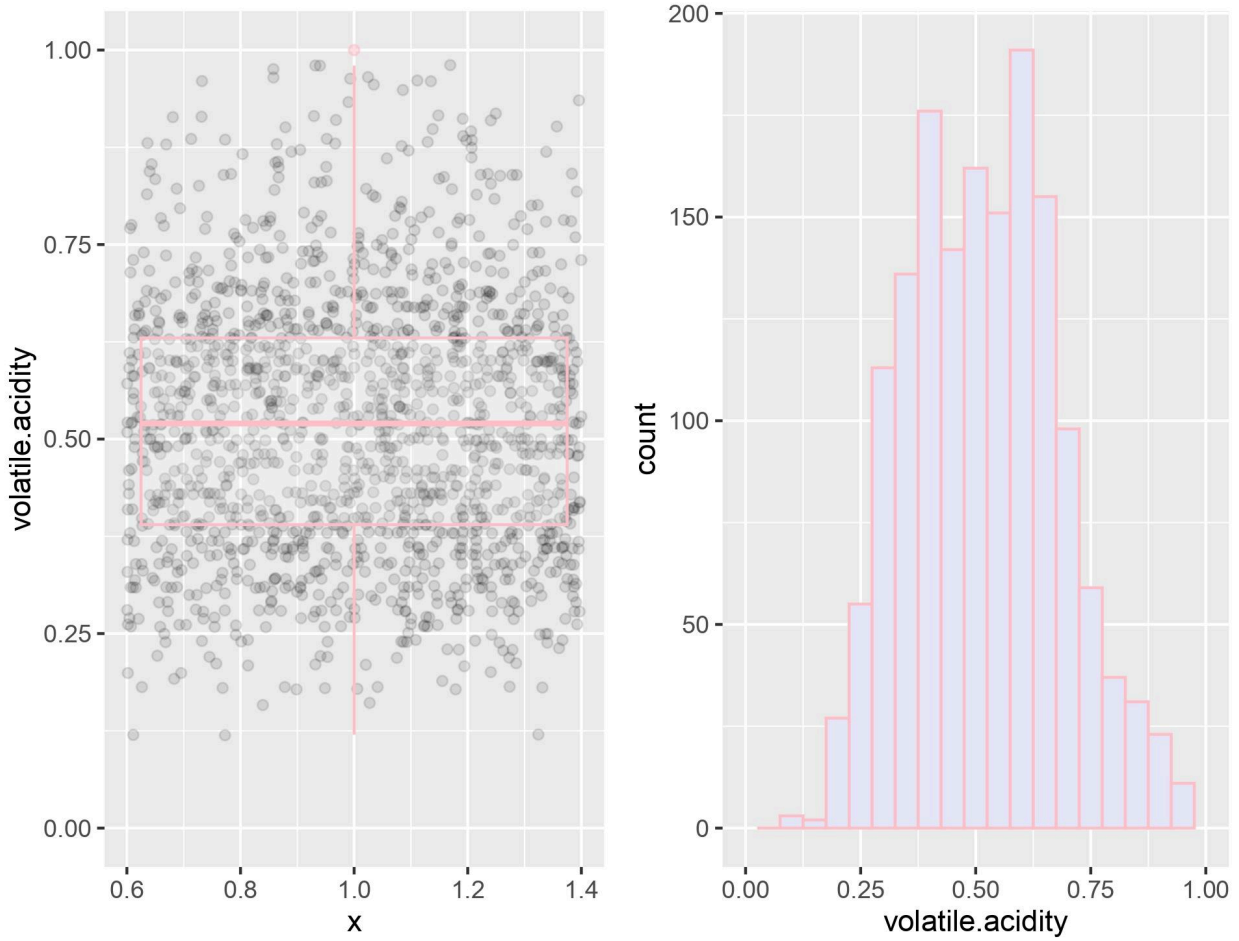


The majority of the wines in the sample are of average quality, as can be seen from the two plots above. We thus question whether the information gathered is accurate and thorough or not. Was this information gathered from a particular geographic area? Or did it cover a significant area? It could be challenging to obtain a precise model of wine quality since the high- and low-quality wines are virtually like outliers in this situation. Let's examine the further graph plots.

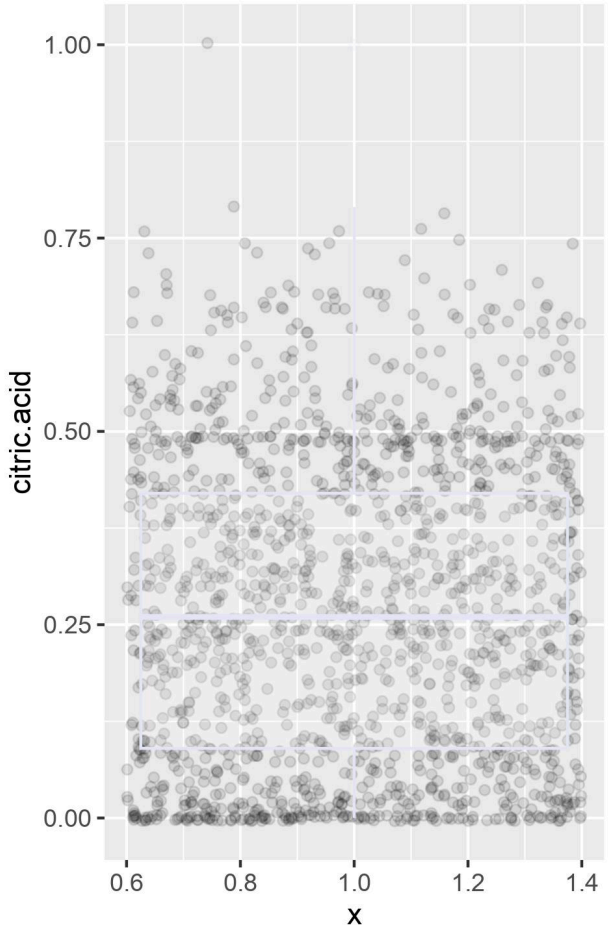




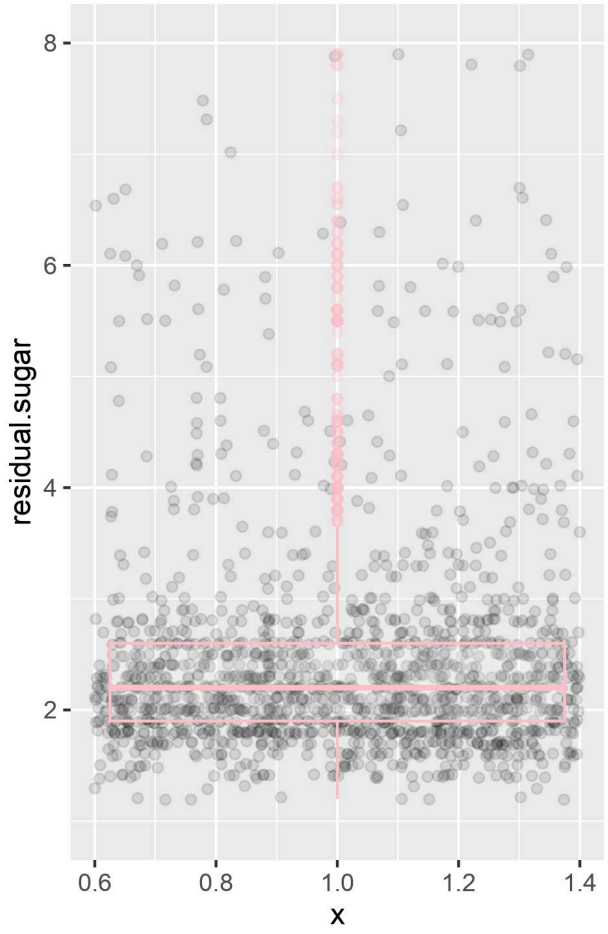
The distribution of Fixed Acidity is positively skewed.With a high concentration of wines with fixed acidity, the median is approximately 8, but the mean has been dragged down to approximately 9.4 as a result of some outliers. The high outliers have been removed by downscaling the image.

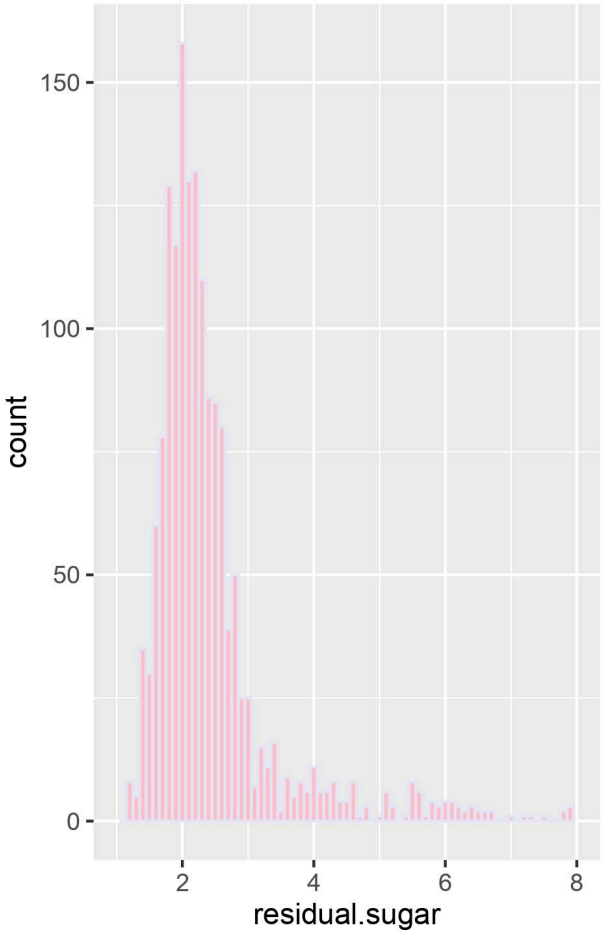


The distribution of Volatile acidity looks like Bimodal with two peaks around 0.4 and 0.6.

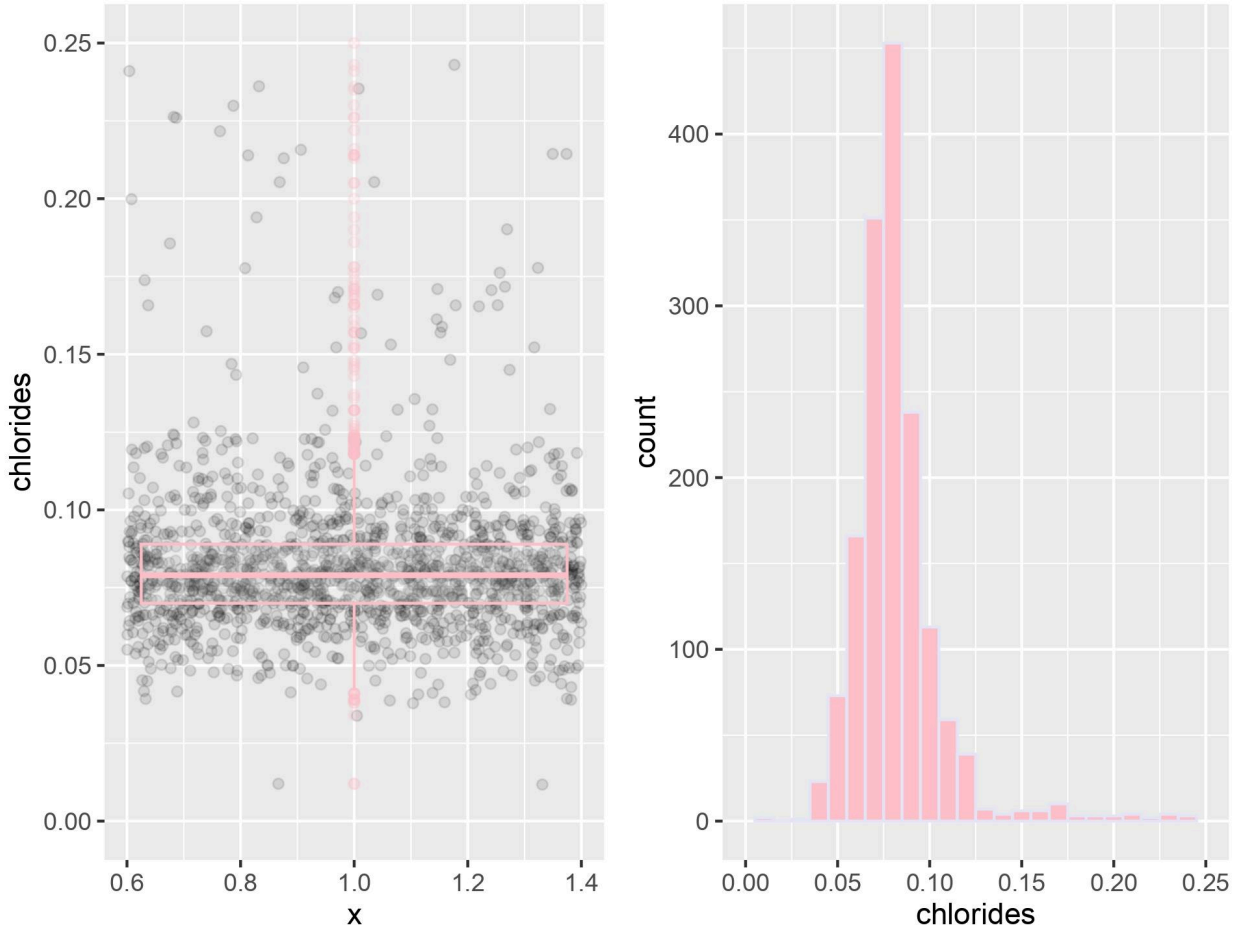


The distribution of citric acid appears odd, with a few outliers. Aside from some higher values for which there is no data at all, the distribution appears to be nearly rectangular. Maybe the data was inaccurate, or maybe the information was not all there?

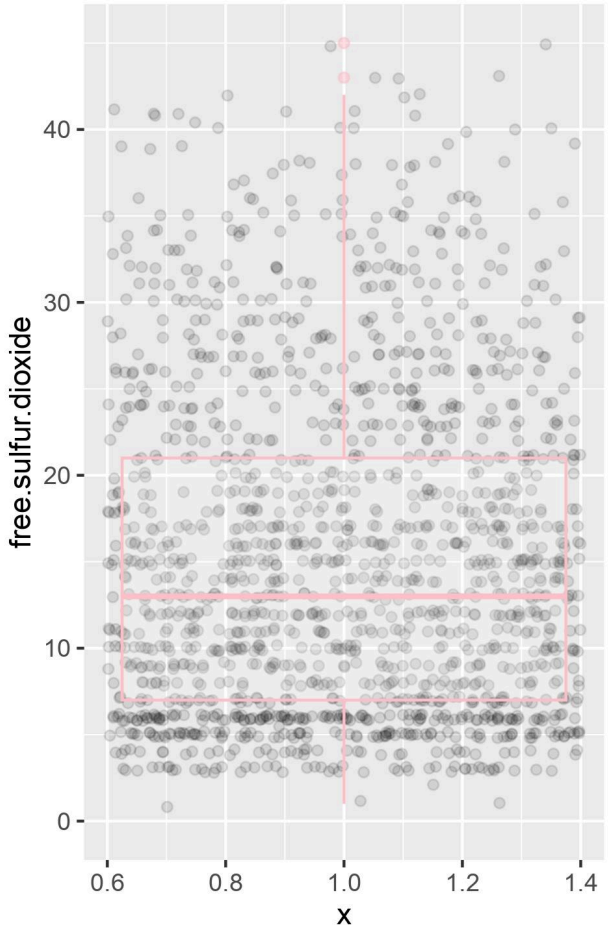


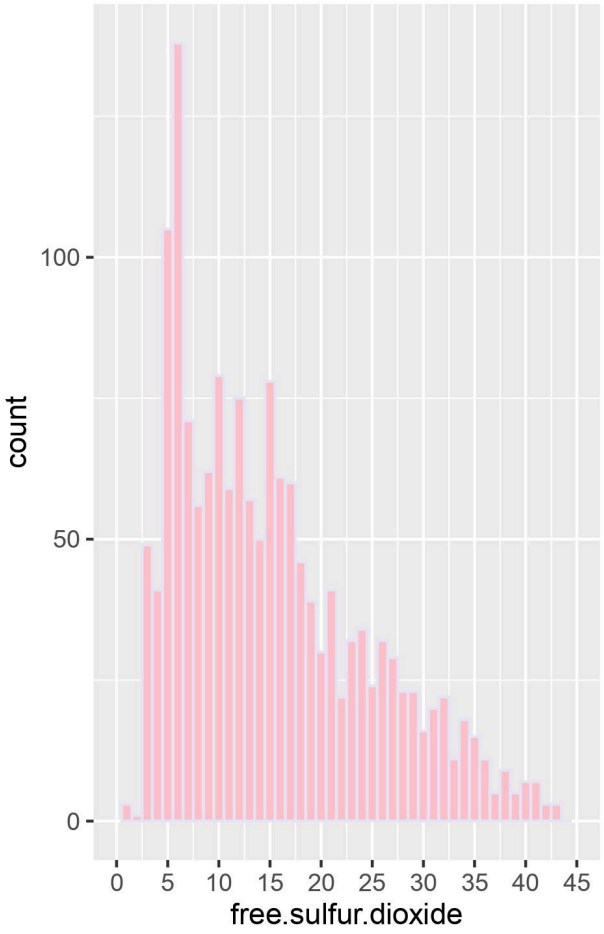


The Residual Sugar distribution is once more positively skewed, with high peaks at about 2.3 and a lot of outliers at the higher ranges.

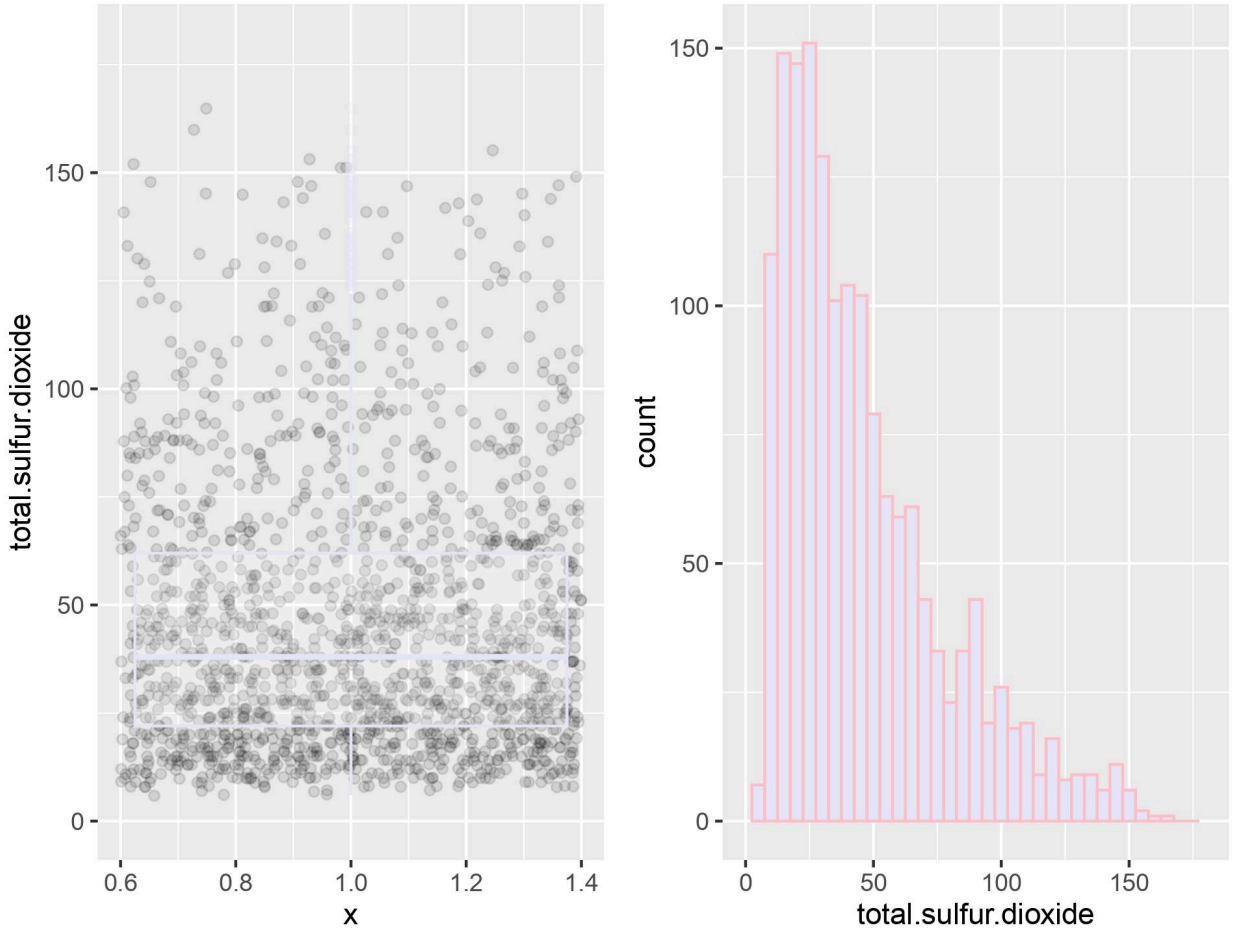


We observe a similar distribution for chlorides to residual sugar. Extreme outliers in this image have been removed.

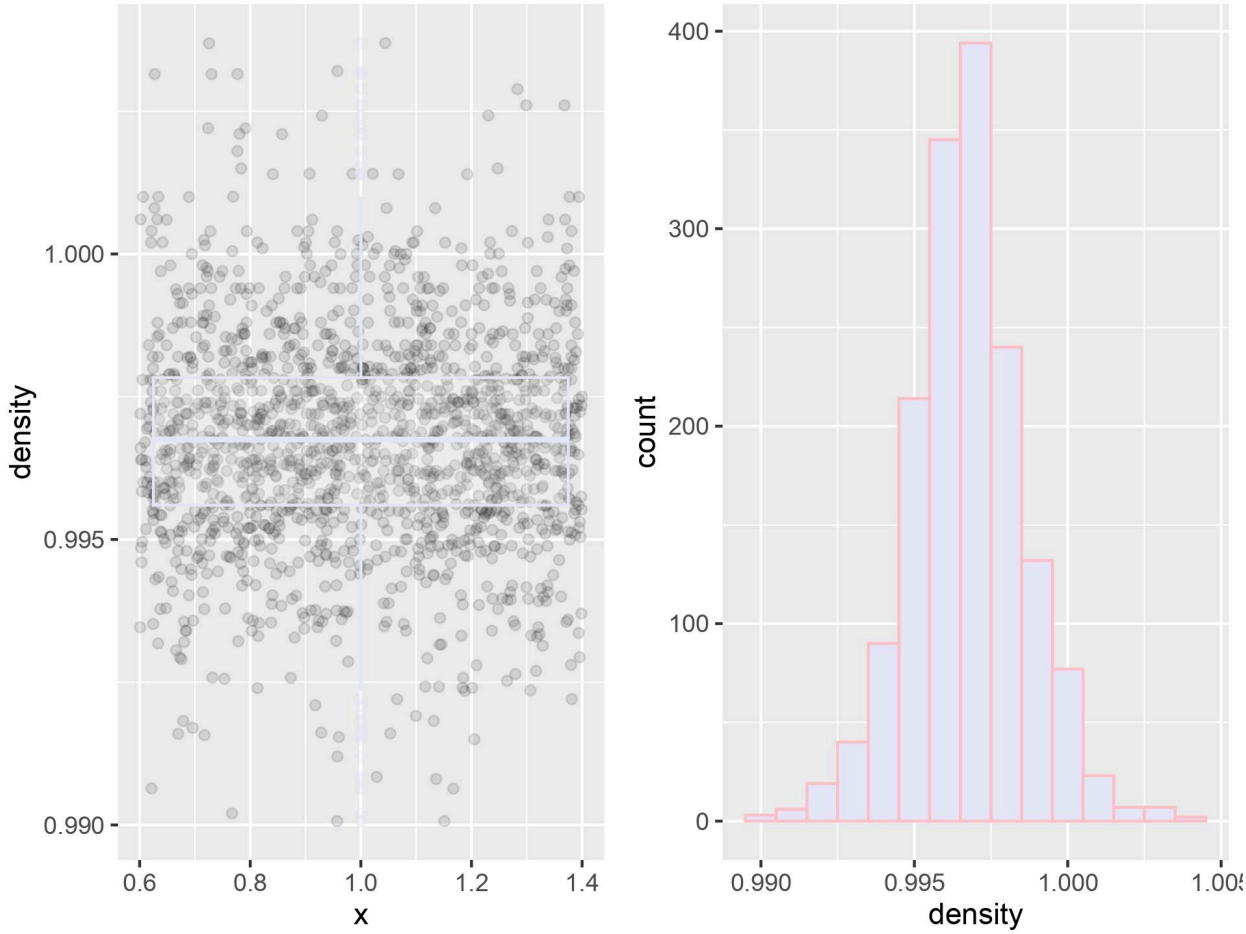




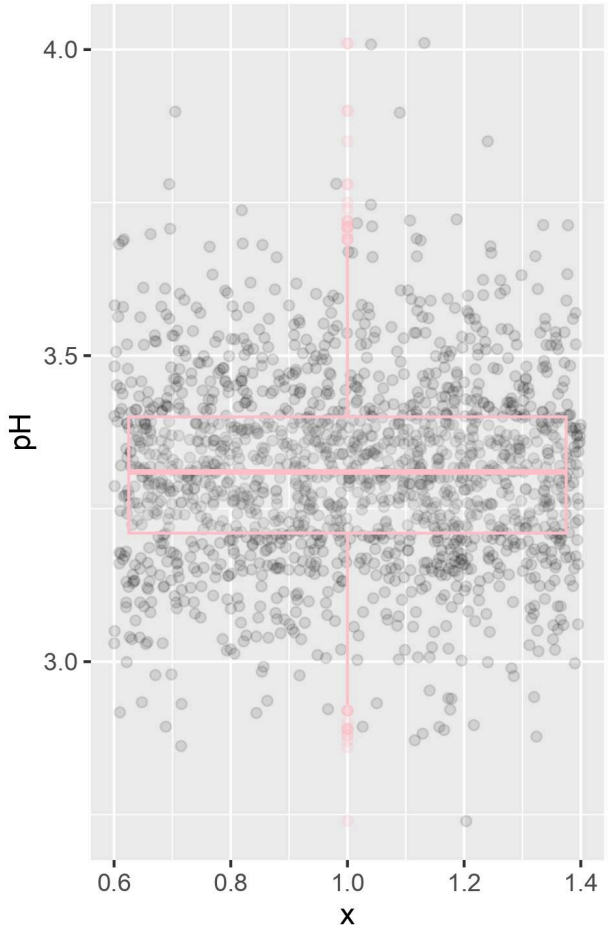
Free sulphur dioxide has a high peak at 7, but it then returns to its typical positively skewed long tailed patterns with a few outliers in the high range.

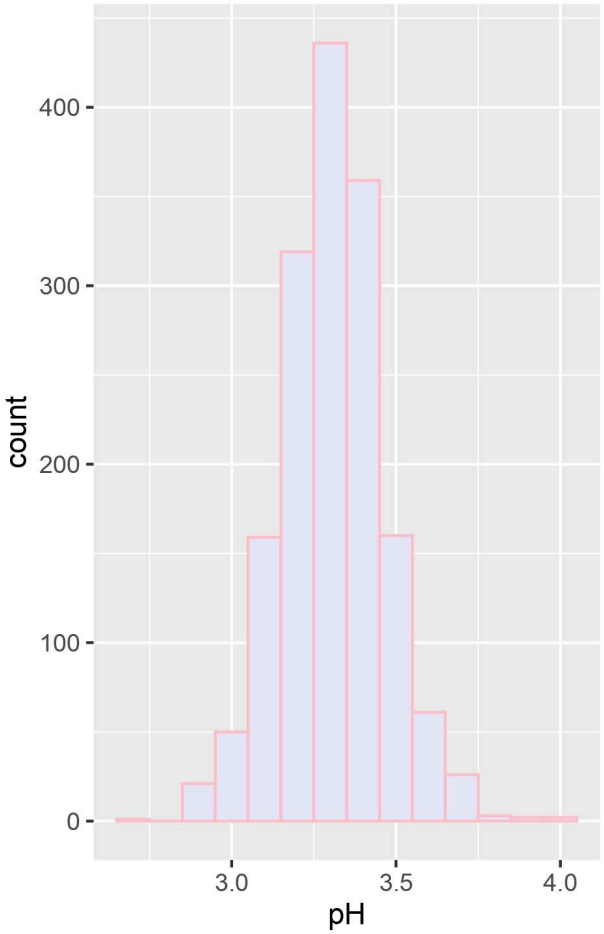


Total Sulphur Dioxide, which is a superset of the previous variable, also exhibits the same pattern.

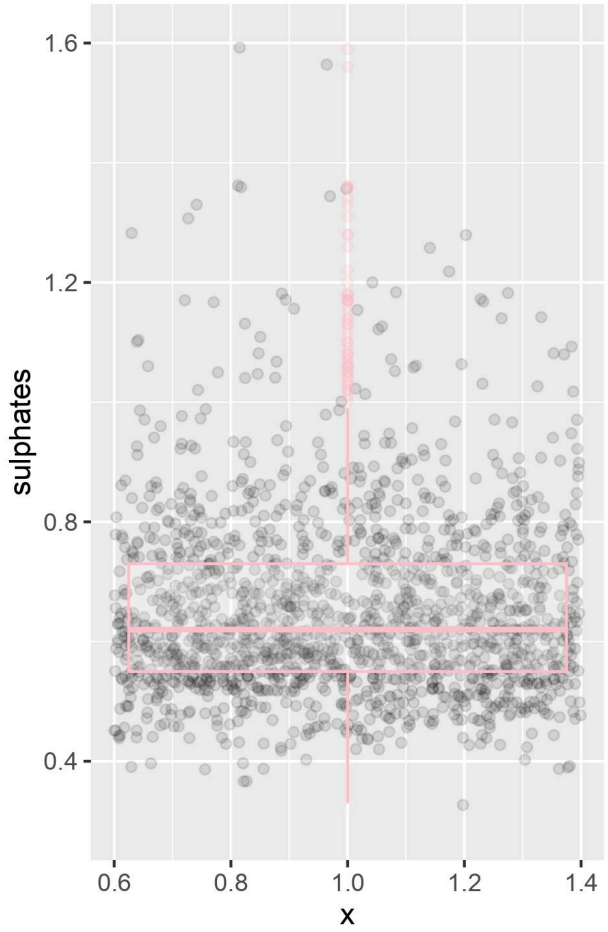


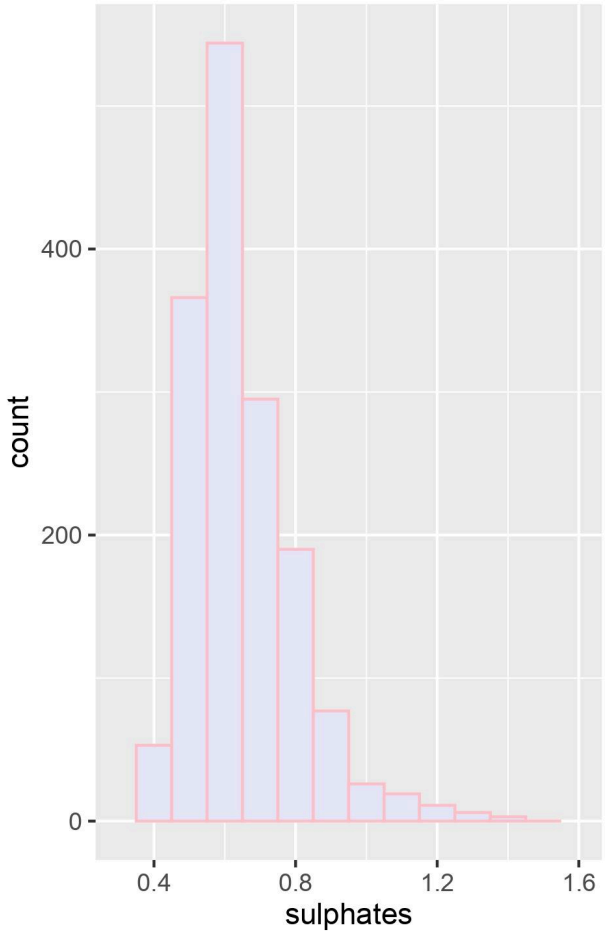
We observe a brand-new phenomenon for the density variable. A nearly perfect Normal Distribution can be found for this variable.



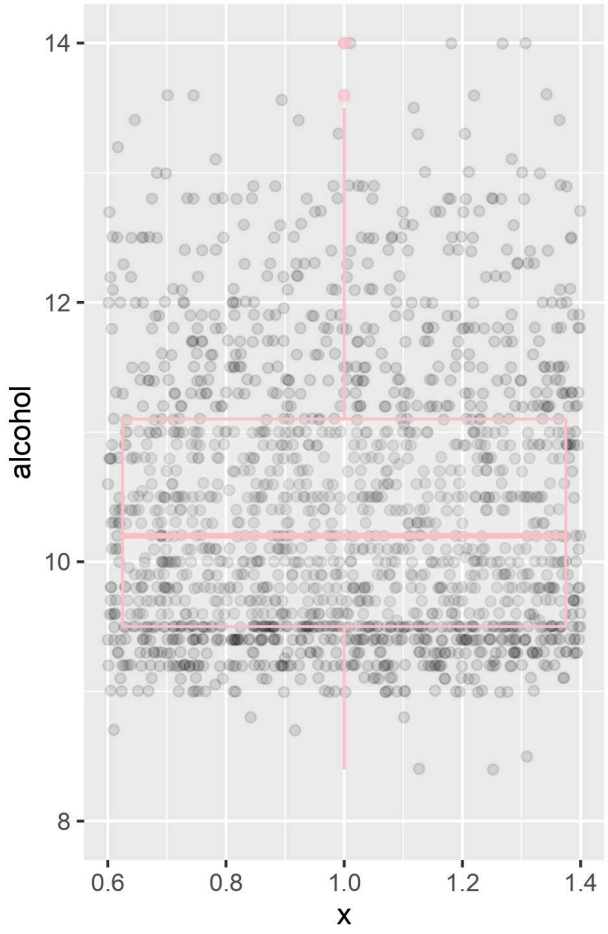


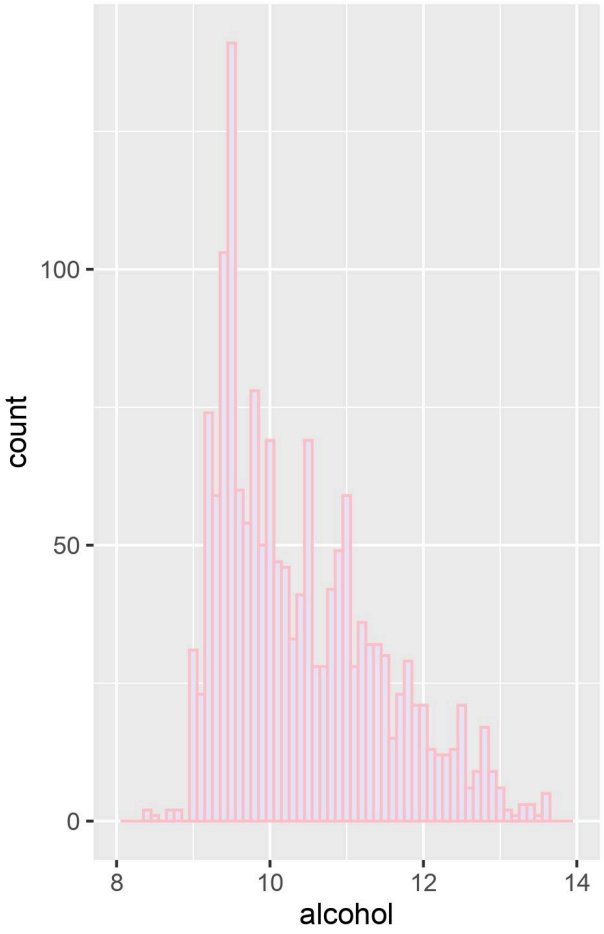
pH also has a very Normally distributed shape.





The distribution of sulphates is also long-tailed, just like that of chlorides or free/total sulphur dioxide. There are comparatively fewer outliers.





Alcohol has a skewed distribution as well, but it is less skewed than residual sugars or chlorides.

Analysis of Univariate Plots

Dataset Structure

Originally, the Red Wine Dataset had 1599 rows and 12 columns. The number of columns increased to 13 after we added a new column called "rating." Here, "quality" serves as our categorical variable, and the remaining variables are numerical variables that reflect the wine's physical and chemical characteristics.

We also observe that there are relatively few "bad" and "good" wines in this dataset, with the majority of the wines being of "average" quality. This makes us question whether or not the dataset is complete once more. Building a predictive model may be difficult due to the lack of this data because we don't have enough information about both good and bad quality wines.

Point of Interest

The 'quality' of this dataset is what interests us the most. We are trying to figure out what factors affect a wine's quality.

Initial Hypothesis

Without examining the data, we believe that the acidity of the wine (whether it be fixed, volatile, or citric) may alter the wine's quality depending on their values. Additionally, the quality may be impacted by pH and acidity. It would also be intriguing to observe how the various acids in the wine affect the pH and whether the overall pH has an impact on the quality of the wine. As sugar determines how sweet the wine will be and may negatively impact the taste of the wine, we also believe that residual sugar will have an impact on the wine's quality.

Unique Features of the Dataset

Compared to the other numeric variables, the distribution of citric acid is distinct. With the exception of a few outliers, it almost has a rectangular shape. If we compare this distribution of citric acid to the distribution of wine quality, it is incredibly unexpected and may even be the result of insufficient data collection.

Distribution and Outliers

1. With few outliers, density and pH appear to be normally distributed.
2. The extreme outliers in residual sugar and chloride appear to be present.
3. For the outliers present, fixed and volatile acidity, total and free sulphur dioxides, alcohol, and sulphates appear to have long tails.
4. There are many zero values in citric acid. we wonder if incorrect data entry is to blame.

Bivariate Plots

In order to gain a general idea of which variables might be associated with one another, we will first generate a correlation table between the variables included in this dataset.

	fixed.acidity	volatile.acidity	citric.acid
fixed.acidity	1	-0.2561	**0.6717**
volatile.acidity	-0.2561	1	**-0.5525**
citric.acid	**0.6717**	**-0.5525**	1
residual.sugar	0.1148	0.001918	0.1436
chlorides	0.09371	0.0613	0.2038
free.sulfur.dioxide	-0.1538	-0.0105	-0.06098
total.sulfur.dioxide	-0.1132	0.07647	0.03553
density	**0.668**	0.02203	**0.3649**
pH	**-0.683**	0.2349	**-0.5419**
sulphates	0.183	-0.261	**0.3128**
alcohol	-0.06167	-0.2023	0.1099
quality	0.1241	**-0.3906**	0.2264

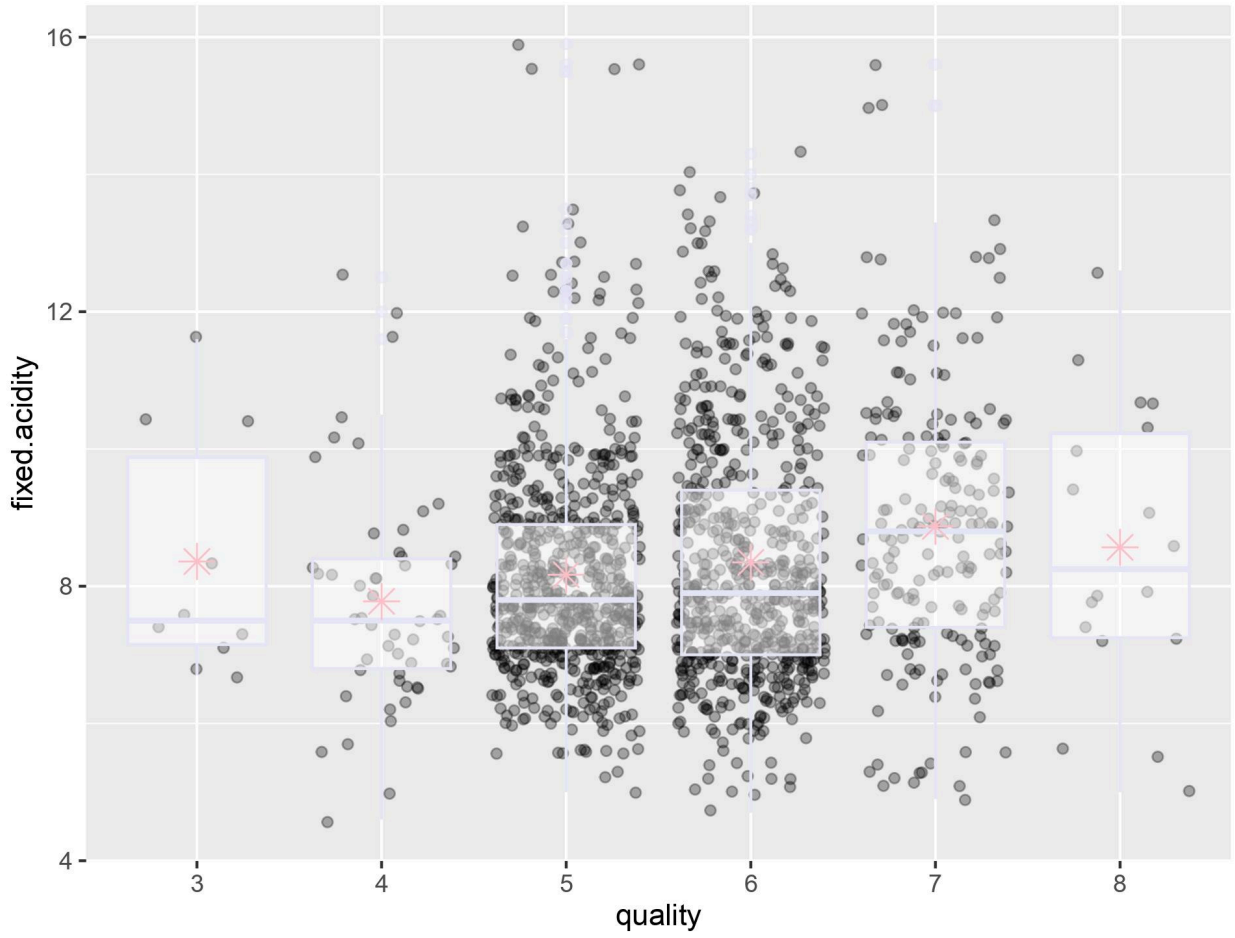
	sulphates	alcohol	quality
fixed.acidity	0.183	-0.06167	0.1241
volatile.acidity	-0.261	-0.2023	**-0.3906**
citric.acid	**0.3128**	0.1099	0.2264
residual.sugar	0.005527	0.04208	0.01373
chlorides	**0.3713**	-0.2211	-0.1289
free.sulfur.dioxide	0.05166	-0.06941	-0.05066
total.sulfur.dioxide	0.04295	-0.2057	-0.1851
density	0.1485	**-0.4962**	-0.1749
pH	-0.1966	0.2056	-0.05773
sulphates	1	0.09359	0.2514
alcohol	0.09359	1	**0.4762**
quality	0.2514	**0.4762**	1

	residual.sugar	chlorides	free.sulfur.dioxide
fixed.acidity	0.1148	0.09371	-0.1538
volatile.acidity	0.001918	0.0613	-0.0105
citric.acid	0.1436	0.2038	-0.06098
residual.sugar	1	0.05561	0.187
chlorides	0.05561	1	0.005562
free.sulfur.dioxide	0.187	0.005562	1
total.sulfur.dioxide	0.203	0.0474	**0.6677**
density	**0.3553**	0.2006	-0.02195
pH	-0.08565	-0.265	0.07038
sulphates	0.005527	**0.3713**	0.05166
alcohol	0.04208	-0.2211	-0.06941
quality	0.01373	-0.1289	-0.05066

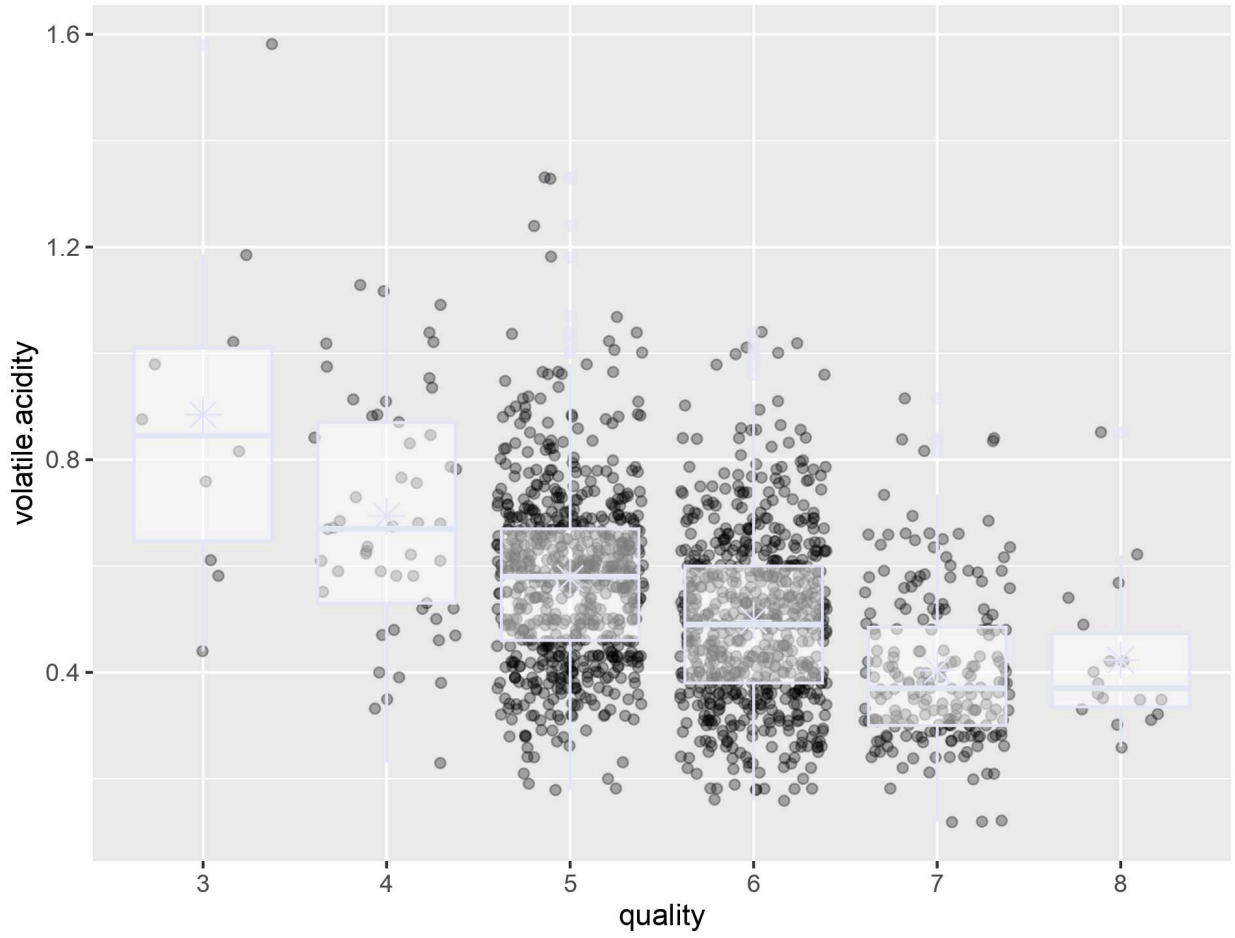
	total.sulfur.dioxide	density	pH
fixed.acidity	-0.1132	**0.668**	**-0.683**
volatile.acidity	0.07647	0.02203	0.2349
citric.acid	0.03553	**0.3649**	**-0.5419**
residual.sugar	0.203	**0.3553**	-0.08565
chlorides	0.0474	0.2006	-0.265
free.sulfur.dioxide	**0.6677**	-0.02195	0.07038
total.sulfur.dioxide	1	0.07127	-0.06649
density	0.07127	1	**-0.3417**
pH	-0.06649	**-0.3417**	1
sulphates	0.04295	0.1485	-0.1966
alcohol	-0.2057	**-0.4962**	0.2056
quality	-0.1851	-0.1749	-0.05773

1. The first item in this table that attracted our attention is the positive link between volatile acidity and pH. But how is that even conceivable? We are aware that acidity rises when pH decreases. Does this mean that a Simpson's Paradox could be at work here? I'll look into this strange quality in more detail.
2. Fixed Acidity and Density have a very strong correlation.
3. Volatile Acidity and Alcohol have a strong correlation with quality.
4. Alcohol and density are not correlated. The fact that water has a higher density than alcohol serves as a clear indication of this.

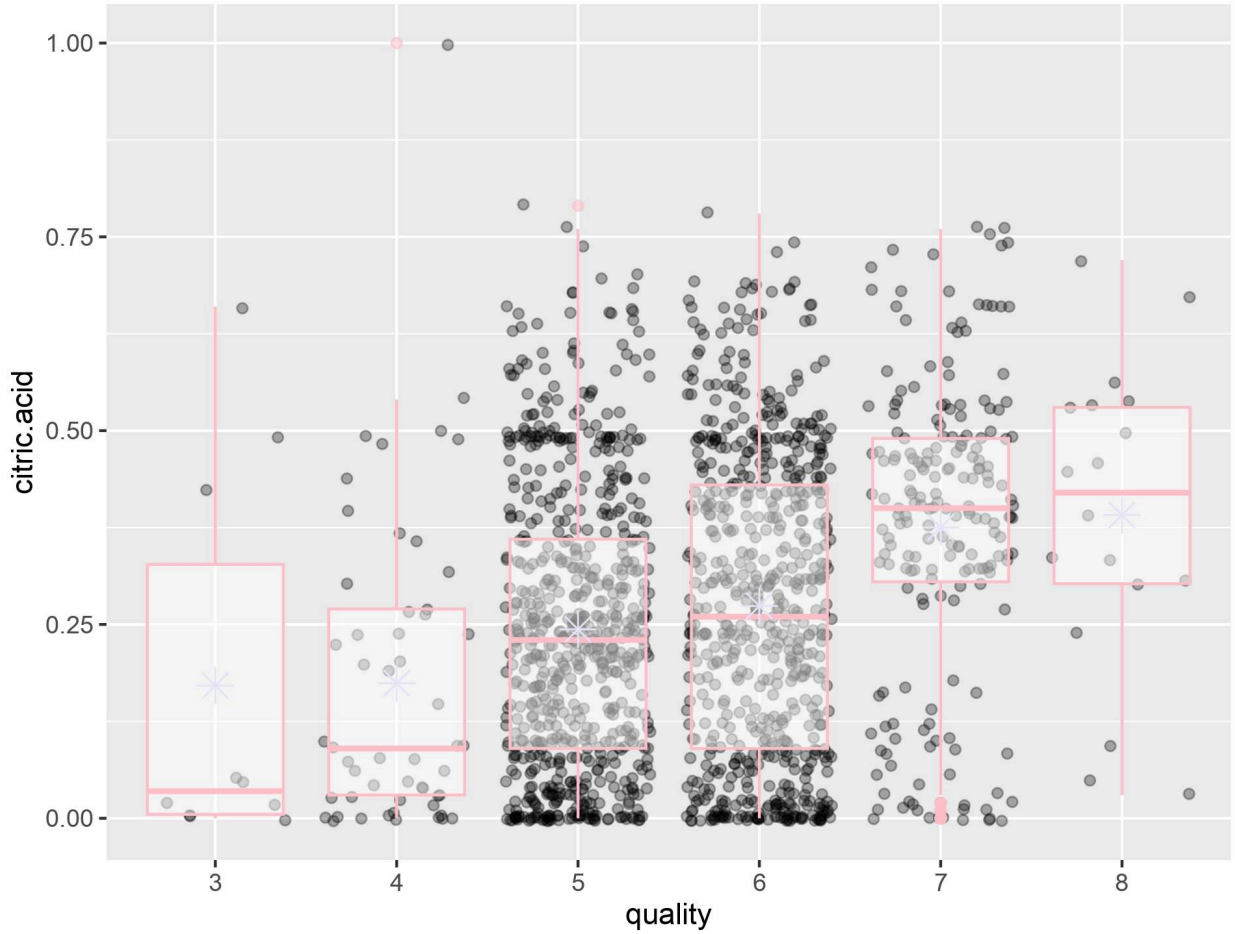
To determine whether we missed anything from the correlation table, let's now plot some Box plots between these variables.



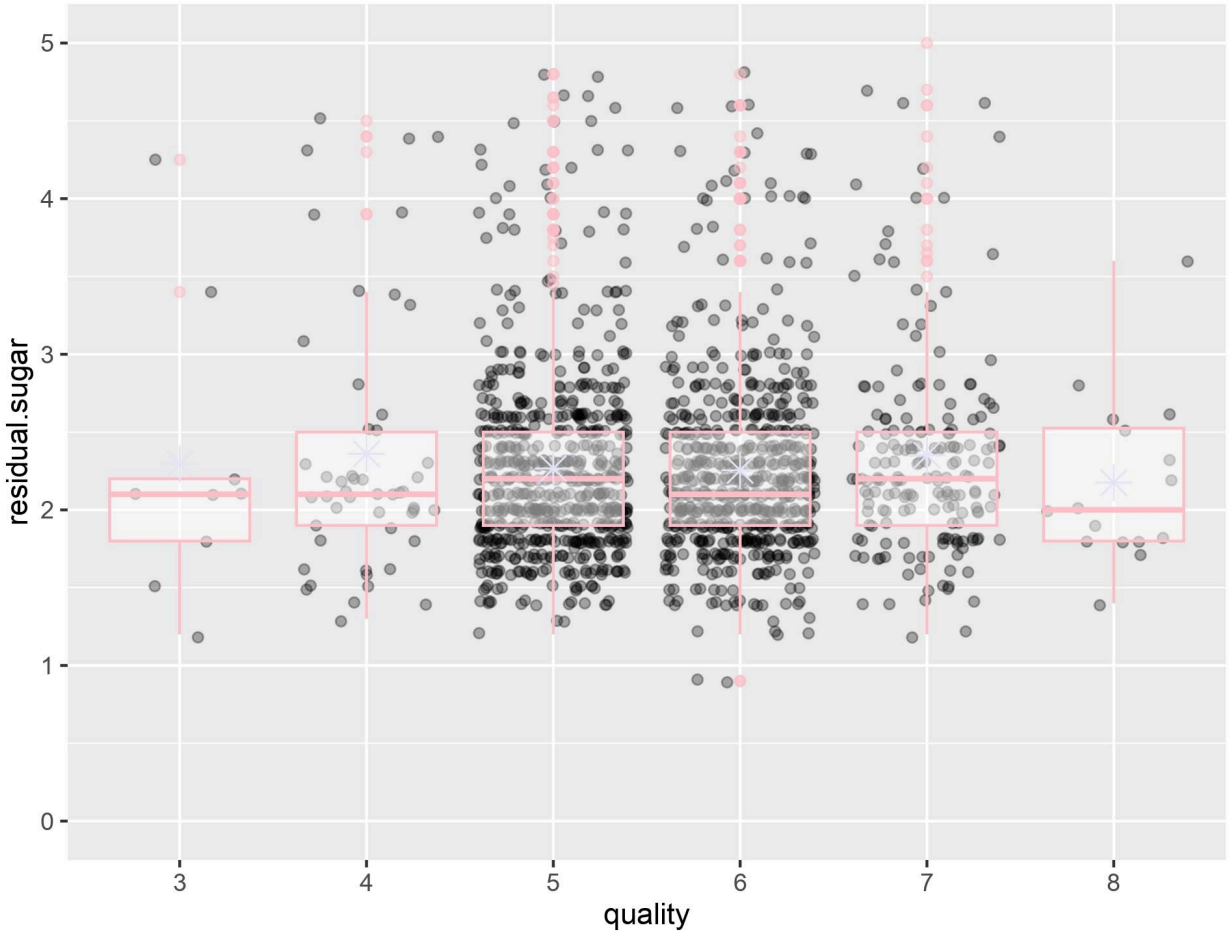
As can be seen, Fixed Acidity hardly has any impact on Quality. With an improvement in quality, the mean and median values of fixed acidity essentially remain unchanged.



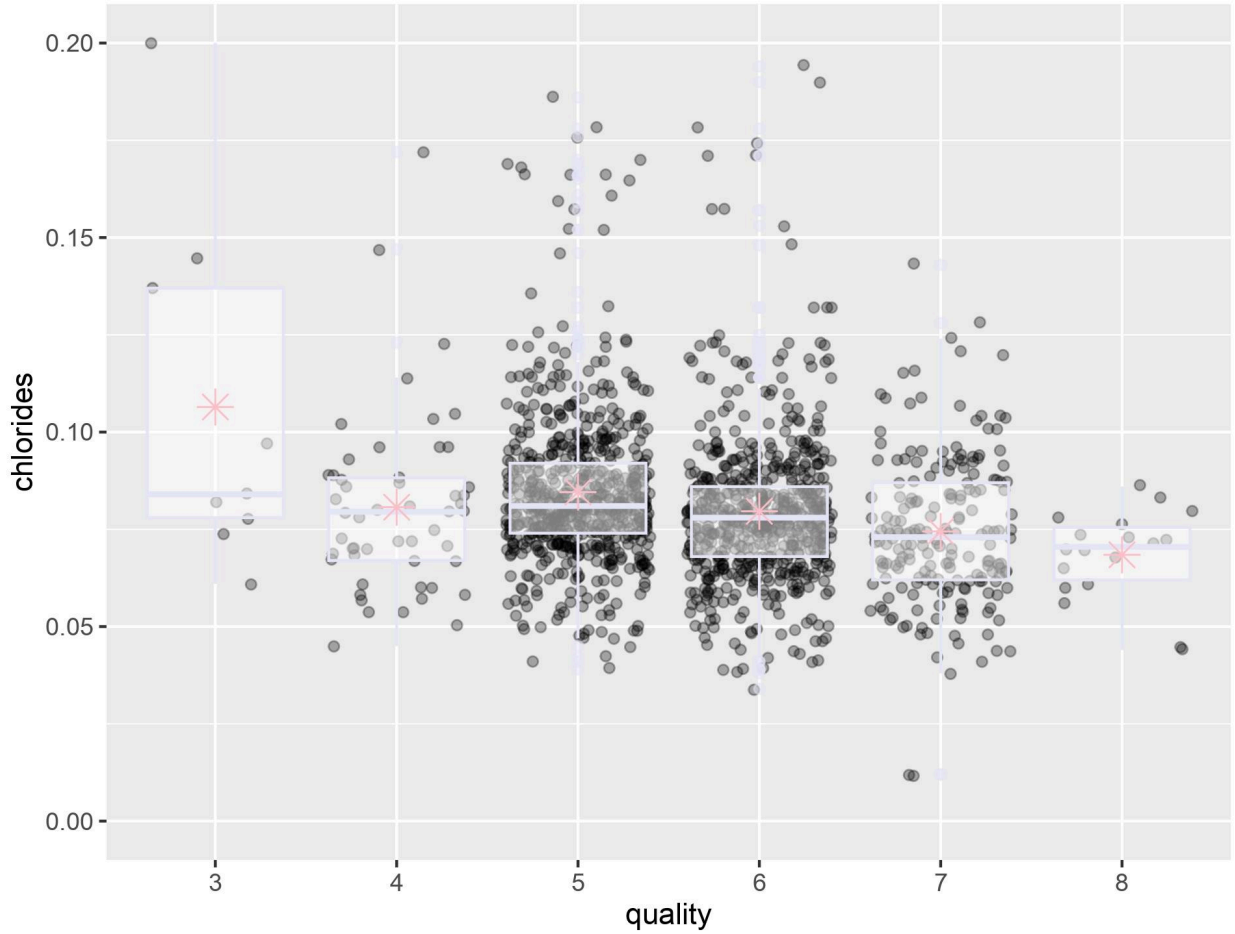
Volatile acid appears to have a detrimental effect on the wine's quality. The quality of the wine declines as the volatile acid level rises.



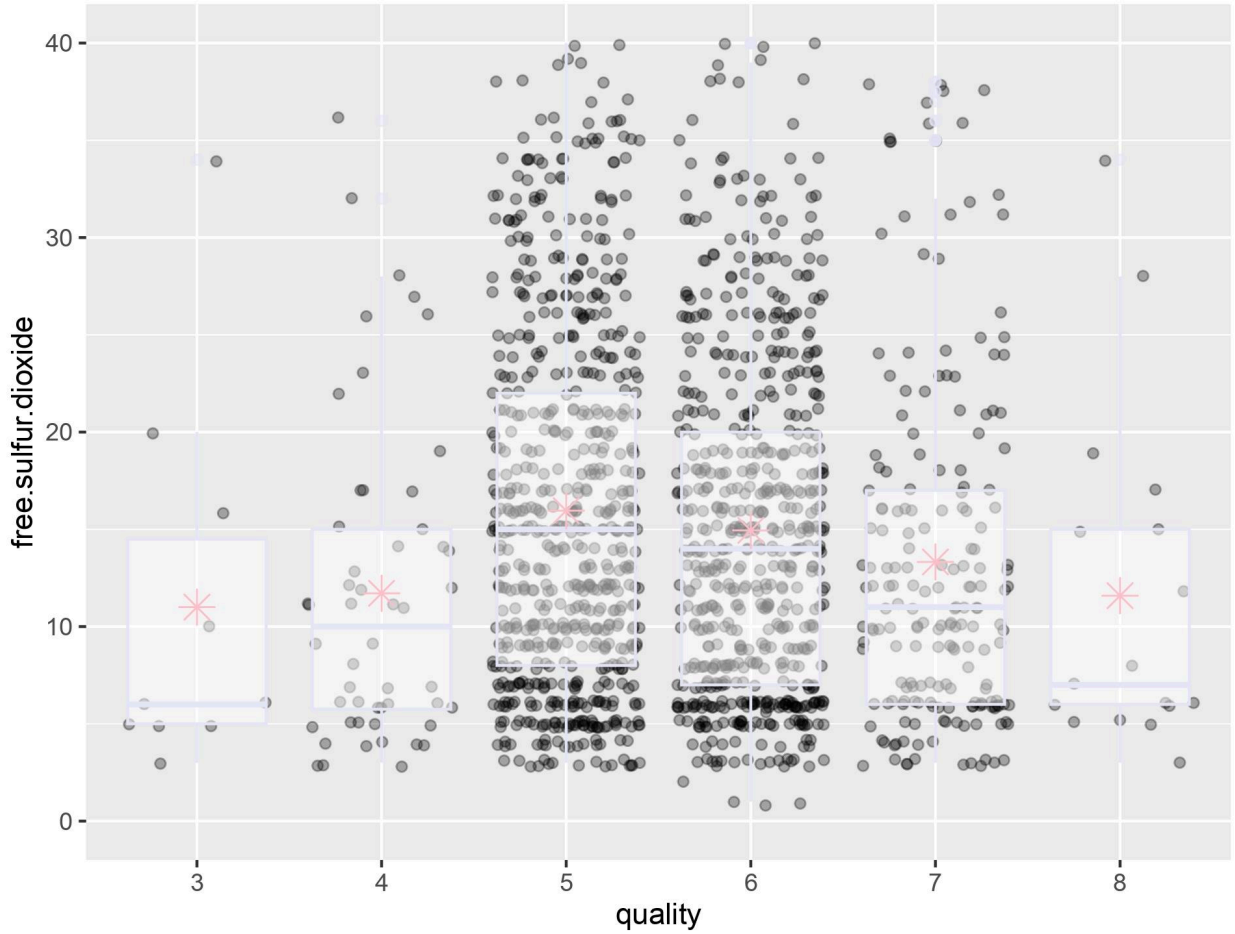
Wine quality and citric acid appear to be positively correlated. Higher Citric Acid is a sign of better wine.



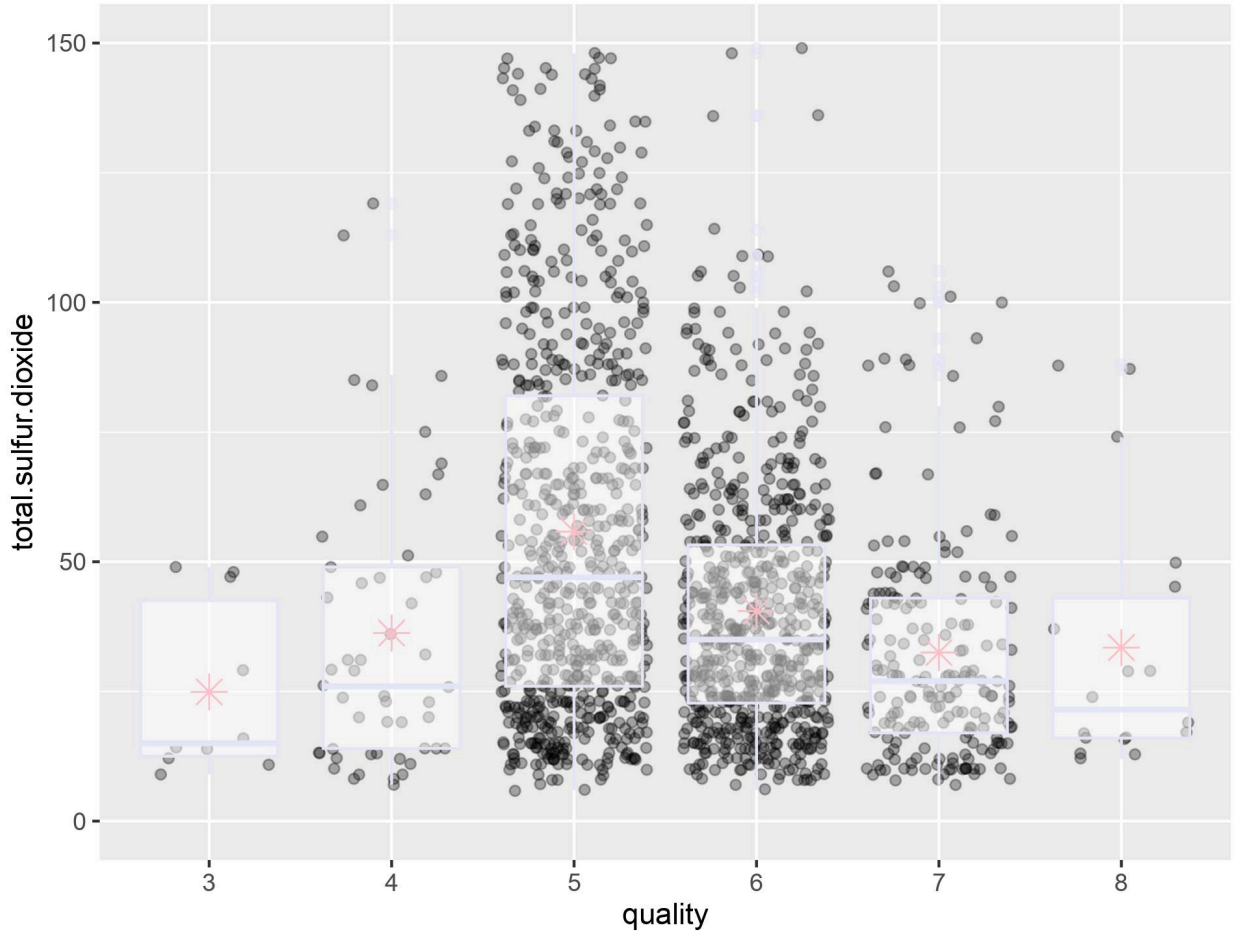
At one time, we believed that residual sugar might have an impact on the taste of the wine. This plot, however, defies that notion and demonstrates that residual sugar virtually has no impact on the quality of the wine. For every wine quality, the mean residual sugar values are essentially the same.



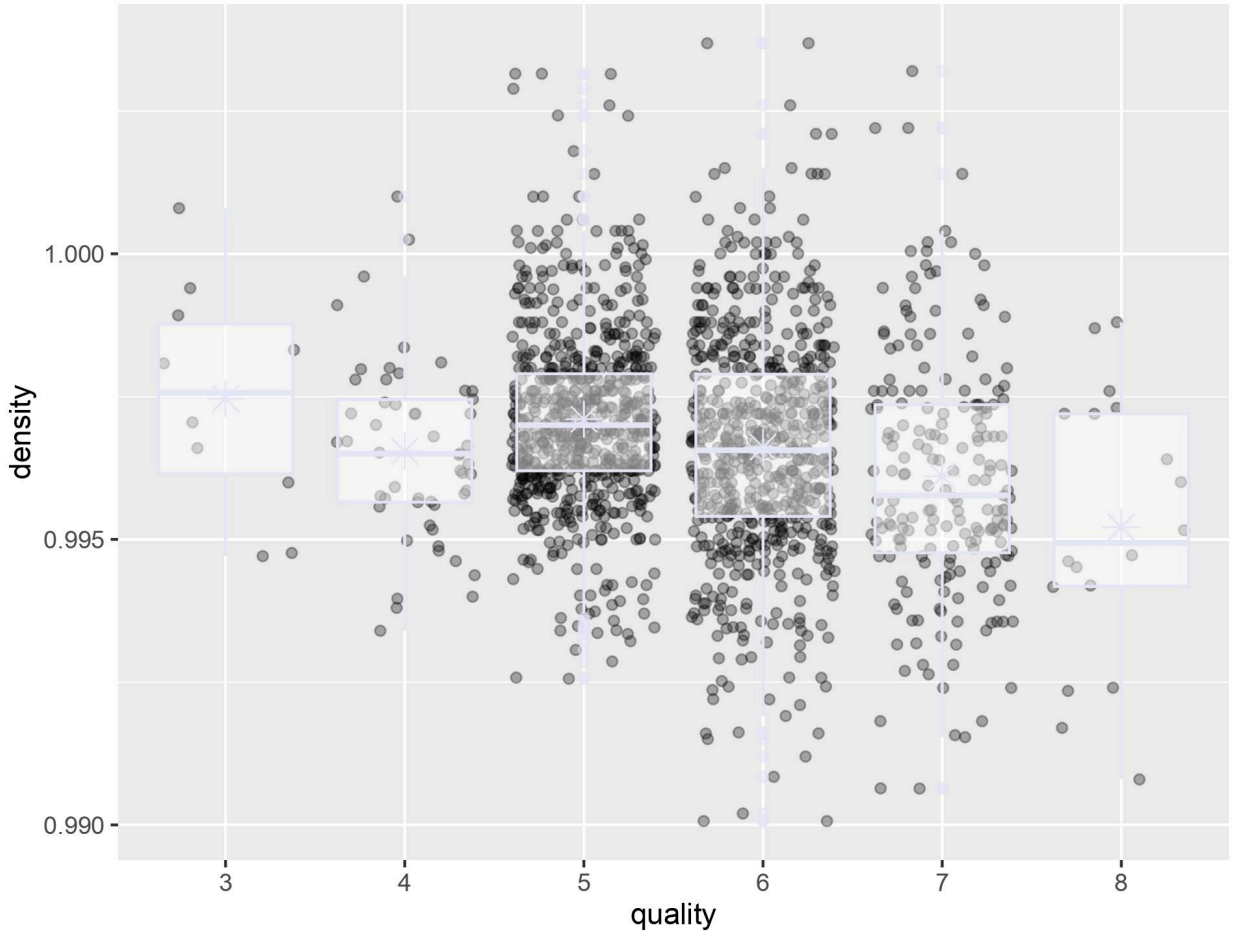
It appears that lower percentages of chloride seem to produce better wines, even though there is only a weak correlation between the decline in median values of the chlorides and increase in quality.



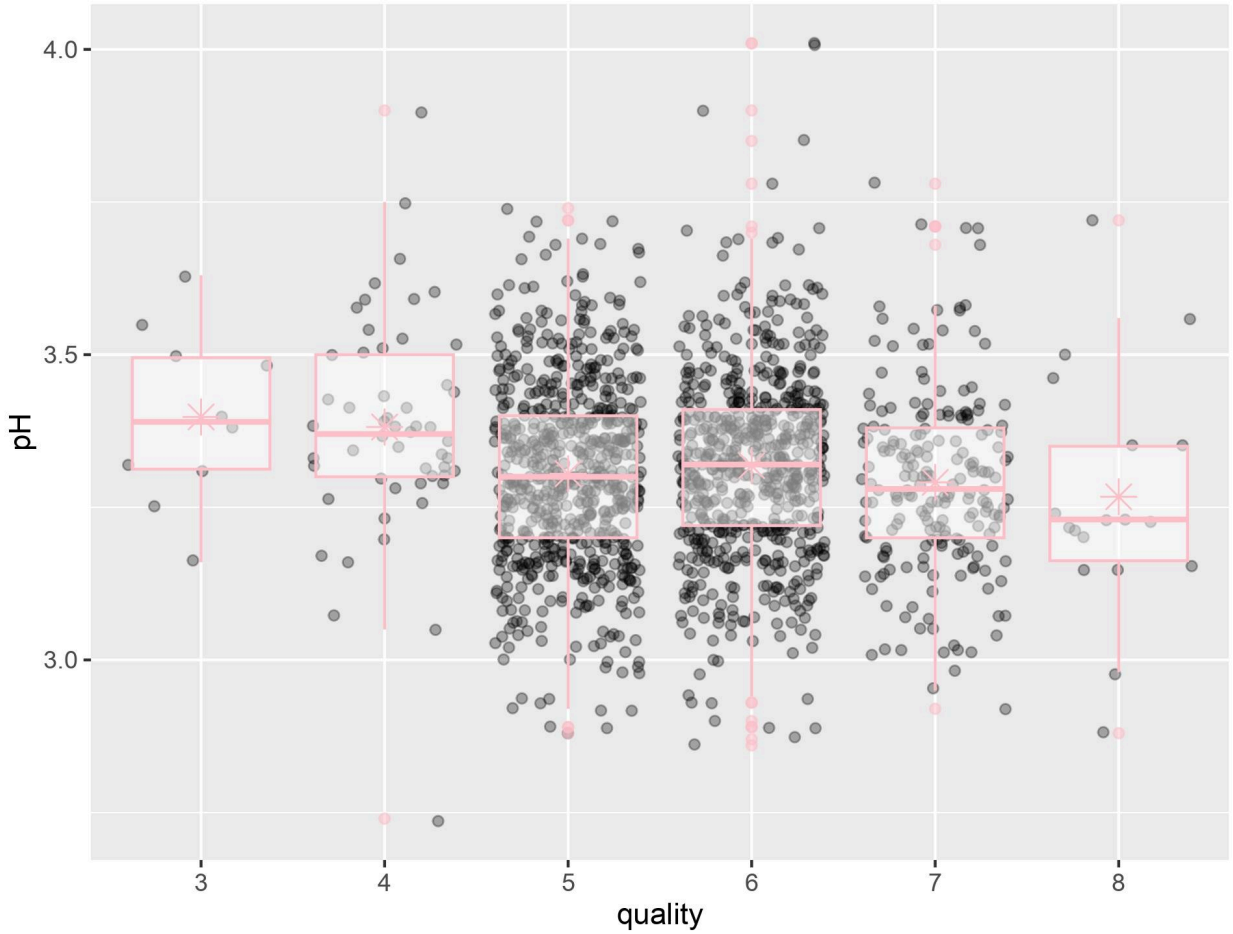
This is a fascinating observation, to be sure. We can see that too much free sulphur dioxide results in average wine while too little sulphur dioxide results in poor wine.



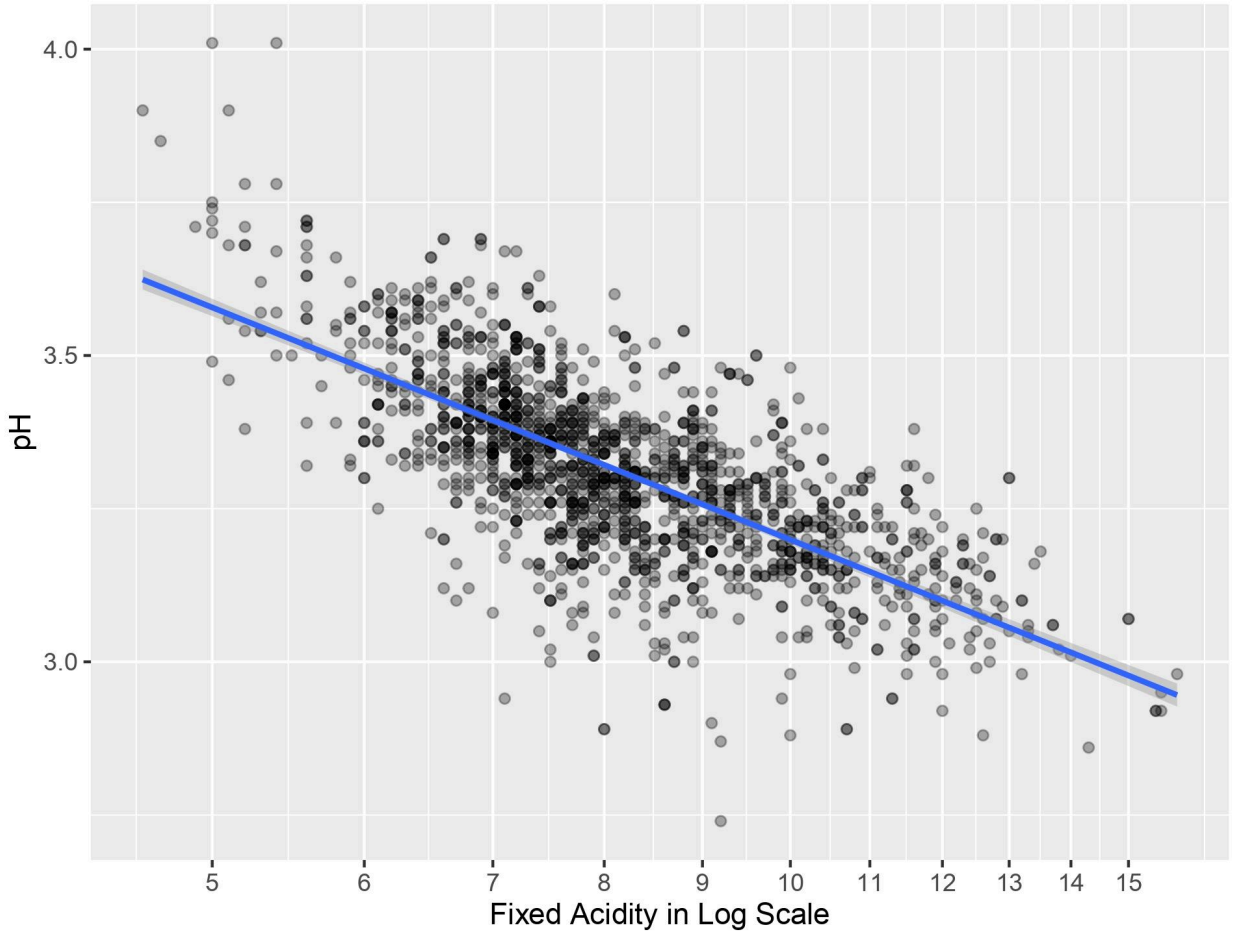
As this is a Subset of Free Sulphur Dioxide, we see a similar pattern here.

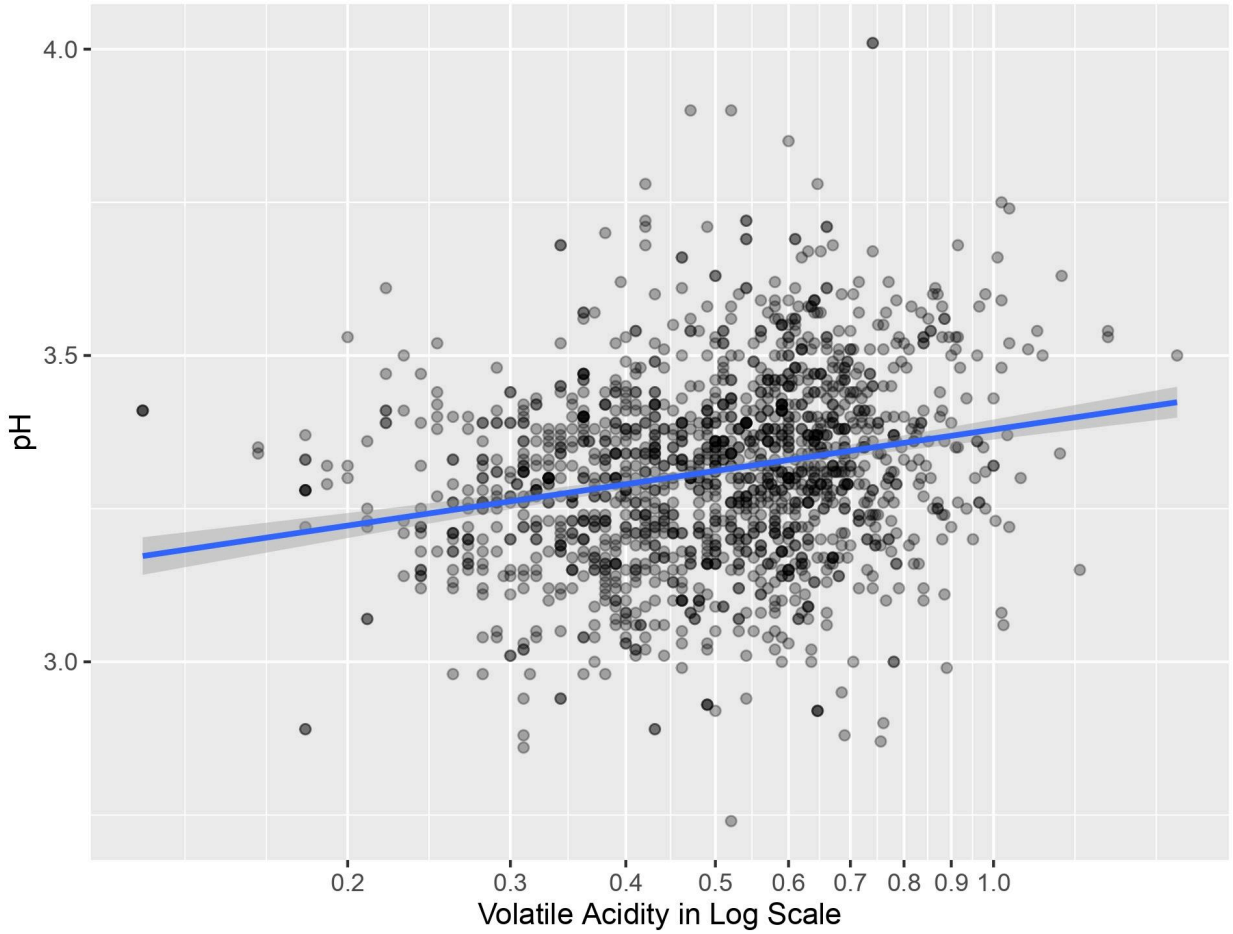


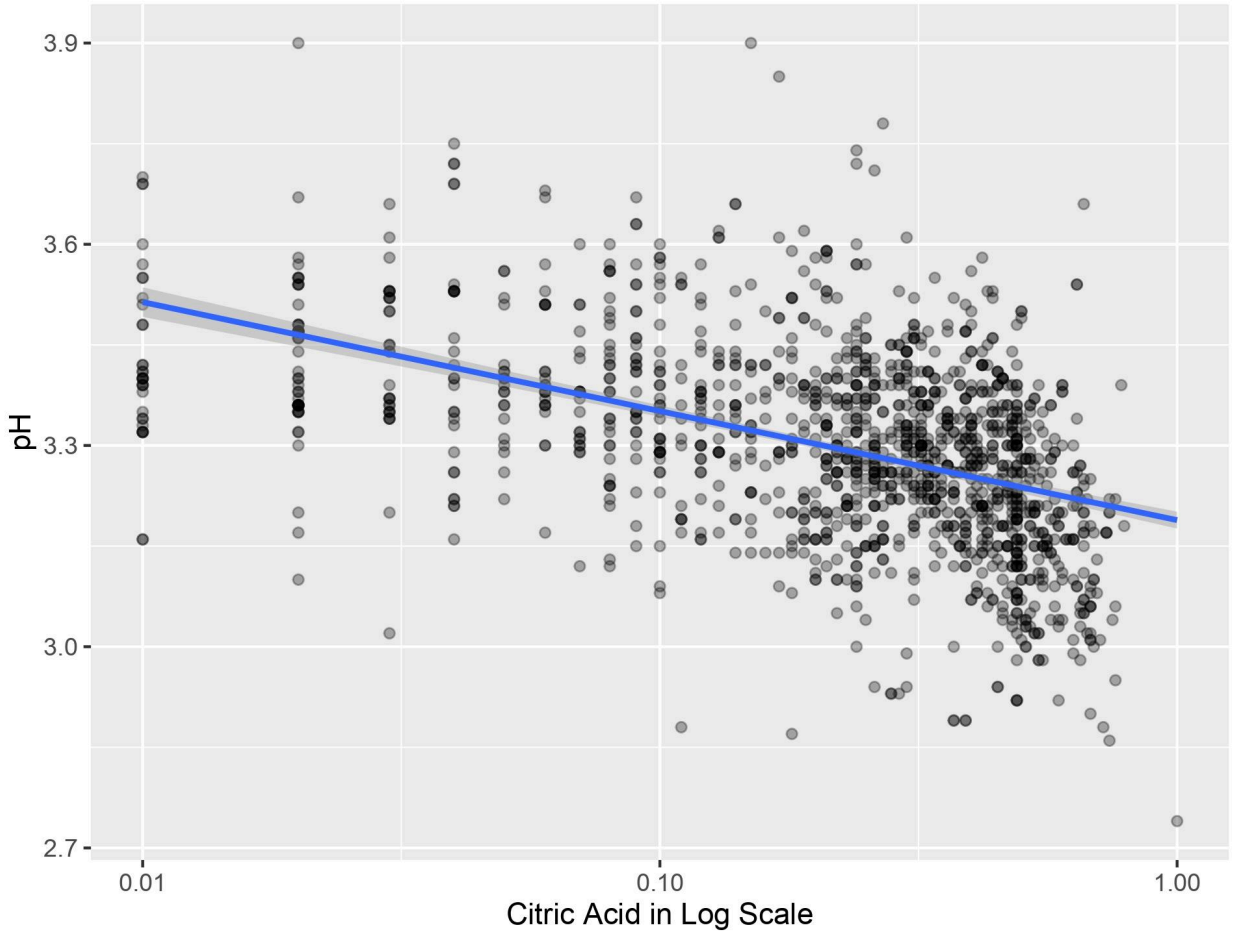
The densities of better wines appear to be lower. However, it might be best not to make any assumptions at this point. Because it's possible that the low density is caused by a higher alcohol content, which is what makes better wines in the first place.



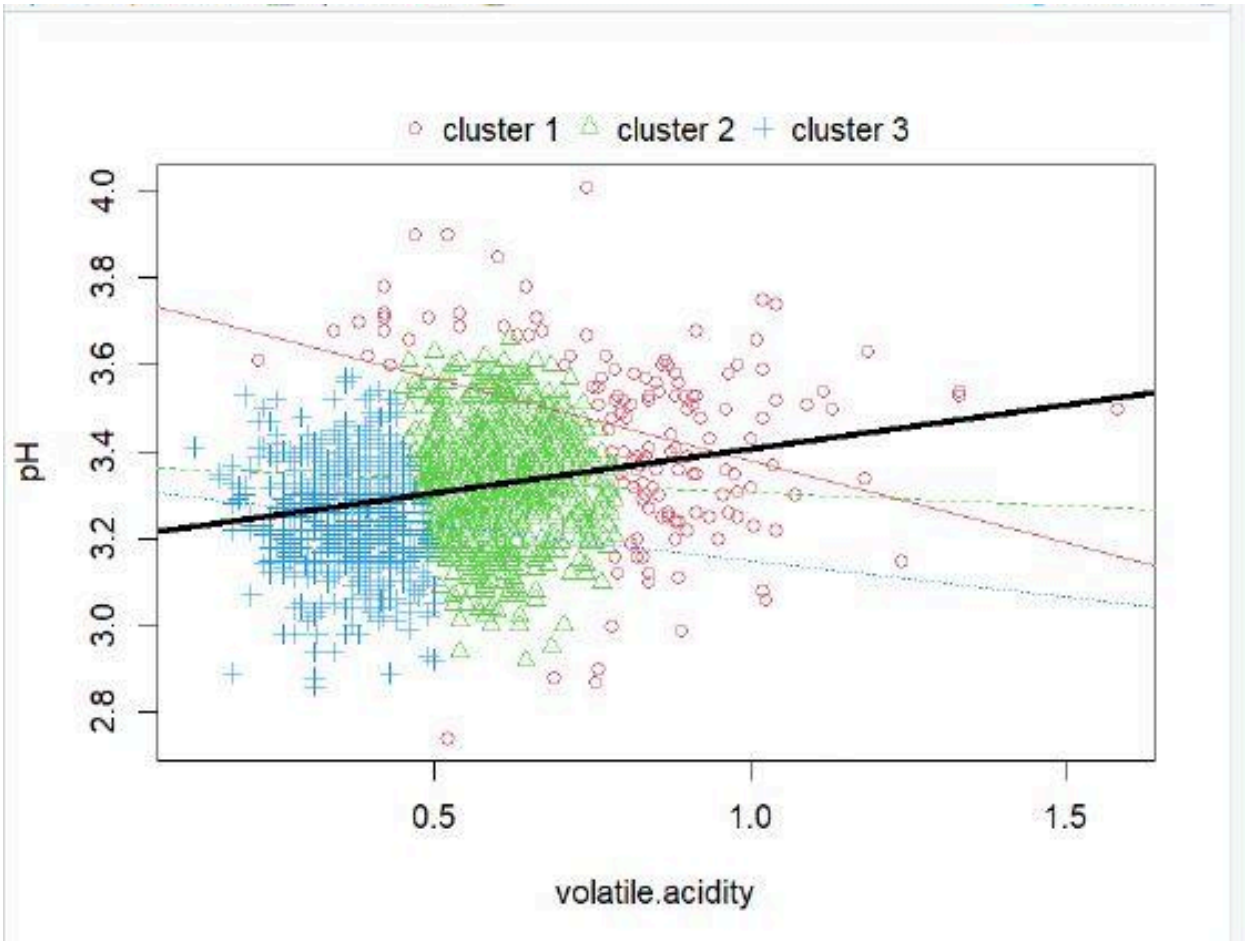
Better wines appear to be more acidic and to have a lower pH. But there are a lot of outliers in this situation. The logical step after that might be to investigate how each acid affects pH.


4.0
3.5
3.0
pH
5
6
7
8
9
10
11
12
13
14
15
Fixed Acidity in Log Scale

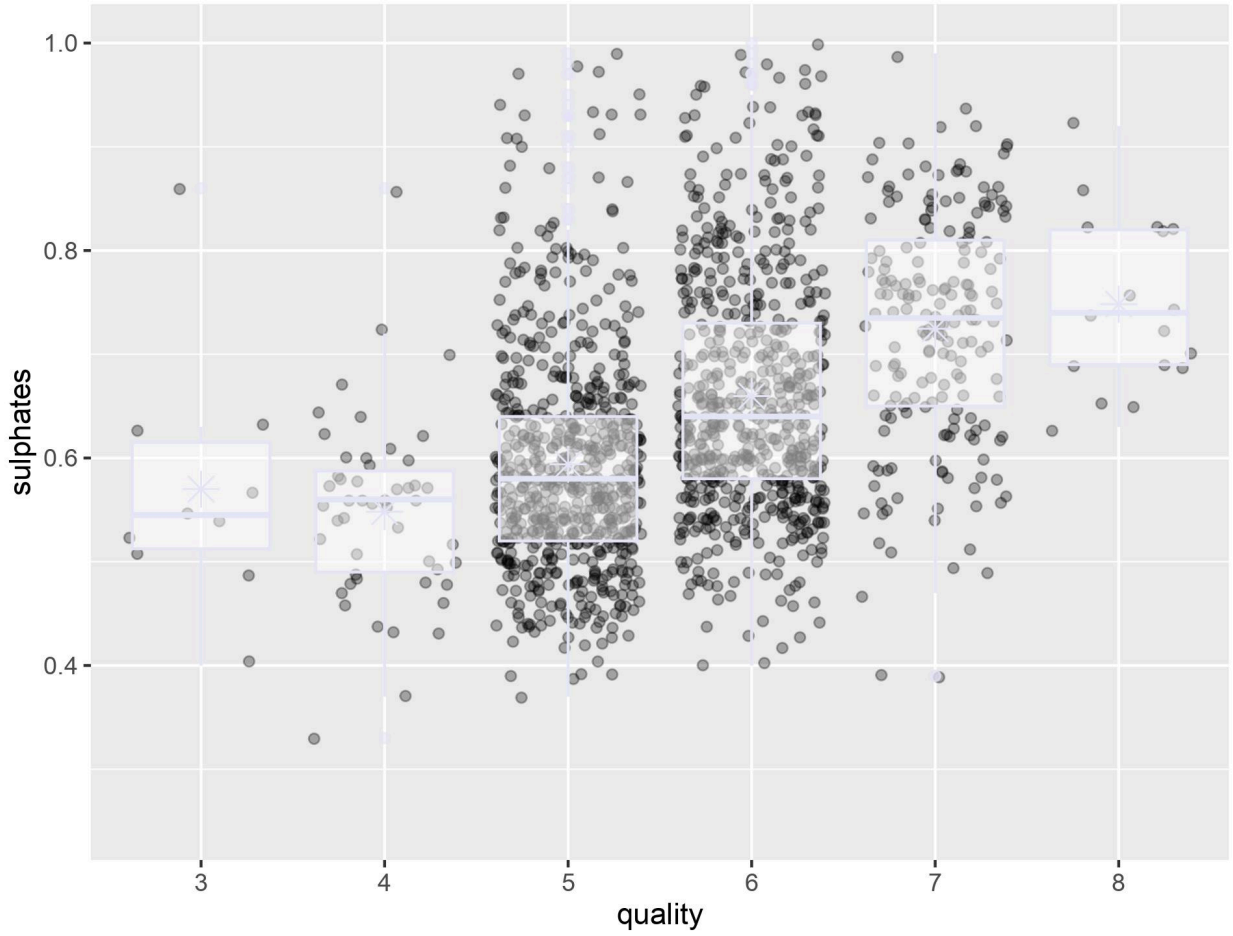

4.0
3.5
3.0
pH
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Volatile Acidity in Log Scale



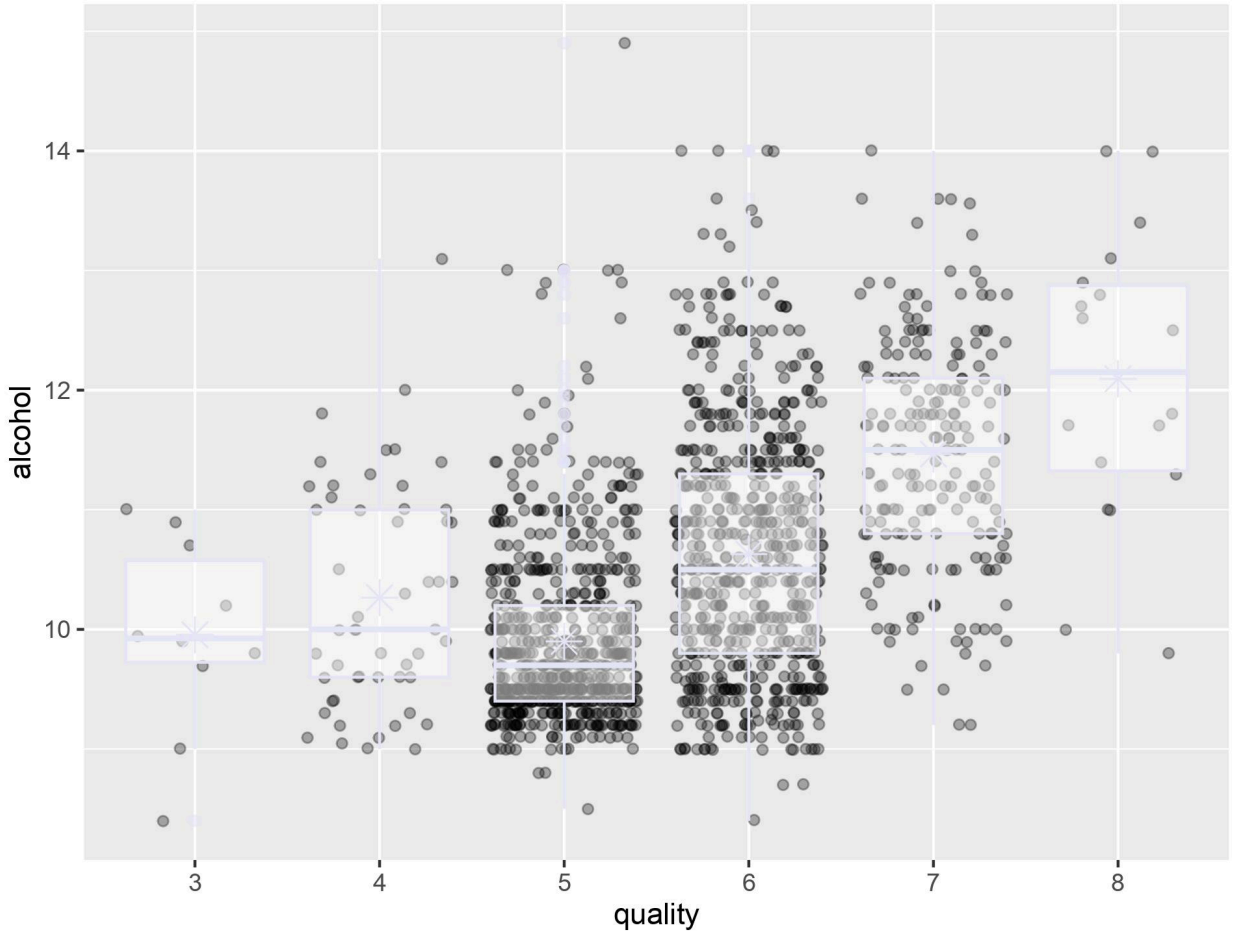
These three stories bring us full circle to the original query. Remember how pH and volatile acid have a positive correlation? However, we are aware that pH is negatively correlated with acidity. So, is it possible that a Simpson's Paradox is at work in this situation? Let's look into it.



Wow! Therefore, Simpson's paradox was the reason for the trend reversal in Volatile Acid vs. pH. After segmenting the data into three parts, we calculated the regression coefficient. The signals have actually altered, as is evident. This is due to a hidden variable that alters the total coefficient.



Although there are several outliers in the "Average" quality wine, it seems that better wines have a higher sulphate concentration.



Here, the connection is quite clear. It should be obvious that better wines have more alcohol in them. But this data has a lot of anomalies. It is therefore likely that alcohol does not determine whether a wine is of excellent quality on its own. Let's attempt to create a simple linear model to collect the statistics.

```
Call:
lm(formula = as.numeric(quality) ~ alcohol, data = wine)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8442 -0.4112 -0.1690  0.5166  2.5888

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.12503    0.17471  -0.716    0.474
alcohol      0.36084    0.01668  21.639   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7104 on 1597 degrees of freedom
Multiple R-squared:  0.2267,    Adjusted R-squared:  0.2263
F-statistic: 468.3 on 1 and 1597 DF,  p-value: < 2.2e-16
```

According to the R squared value, only roughly 22% of the wine's quality may be attributed to alcohol. Therefore, there must be other factors at work here. To create a more accurate regression model, we must ascertain them.

We will now conduct a correlation test between each variable and the wine's quality.

```
correlations
     fixed.acidity     volatile.acidity          citric.acid log10.residual.sugar
        0.12405165          -0.39055778           0.22637251           0.02353331
   log10.chlordies  free.sulfur.dioxide total.sulfur.dioxide              density
        -0.17613996          -0.05065606          -0.18510029          -0.17491923
                pH      log10.sulphates              alcohol
        -0.05773139           0.30864193           0.47616632
```

The following variables appear to have a higher association to wine quality based on the correlation test results.

1. Alcohol
2. Sulphates(log10)
3. Volatile Acidity

4. Citric Acid

Analysis of Bivariate Plots

Observations

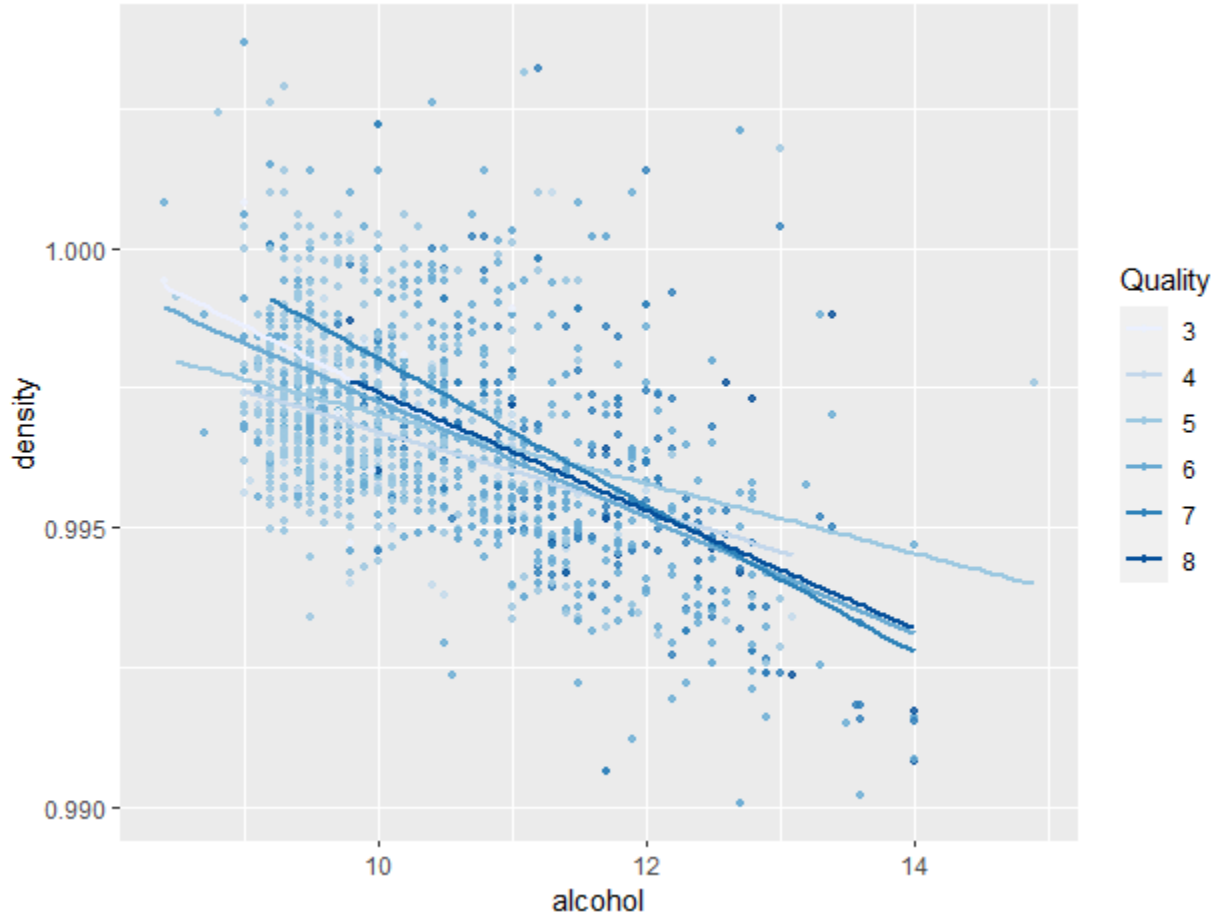
1. Almost no correlation exists between fixed acidity and quality.
2. Volatile Acidity and quality appear to be negatively correlated.
3. Better wines appear to have more citric acid in them.
4. Higher alcohol content seems to be a hallmark of better wines. However, after building a linear model around it, I discovered that alcohol only accounts for about 20% of the variance in quality. Therefore, there could be other things at work here.
5. Even though there is little evidence to support it, wines with less chloride appear to be of higher quality.
6. Less dense wines appear to be better wines. The higher alcohol level in them, however, might also be to blame for this.
7. The acidity in better wines seems to be higher.
8. The wine's quality is essentially unaffected by residual sugar.

Special feature

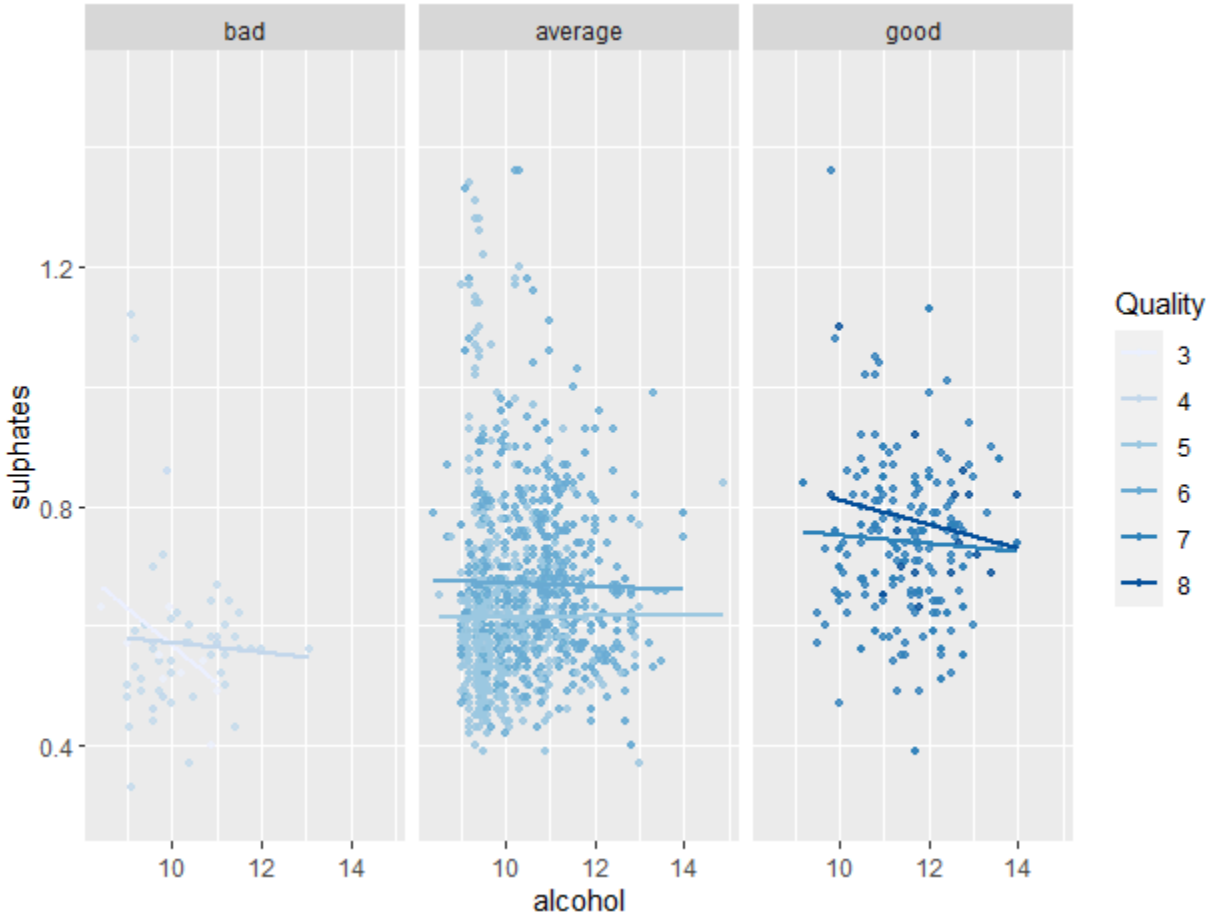
The Simpson's Paradox led to a positive connection between volatile acidity and pH.

Multivariate Plots

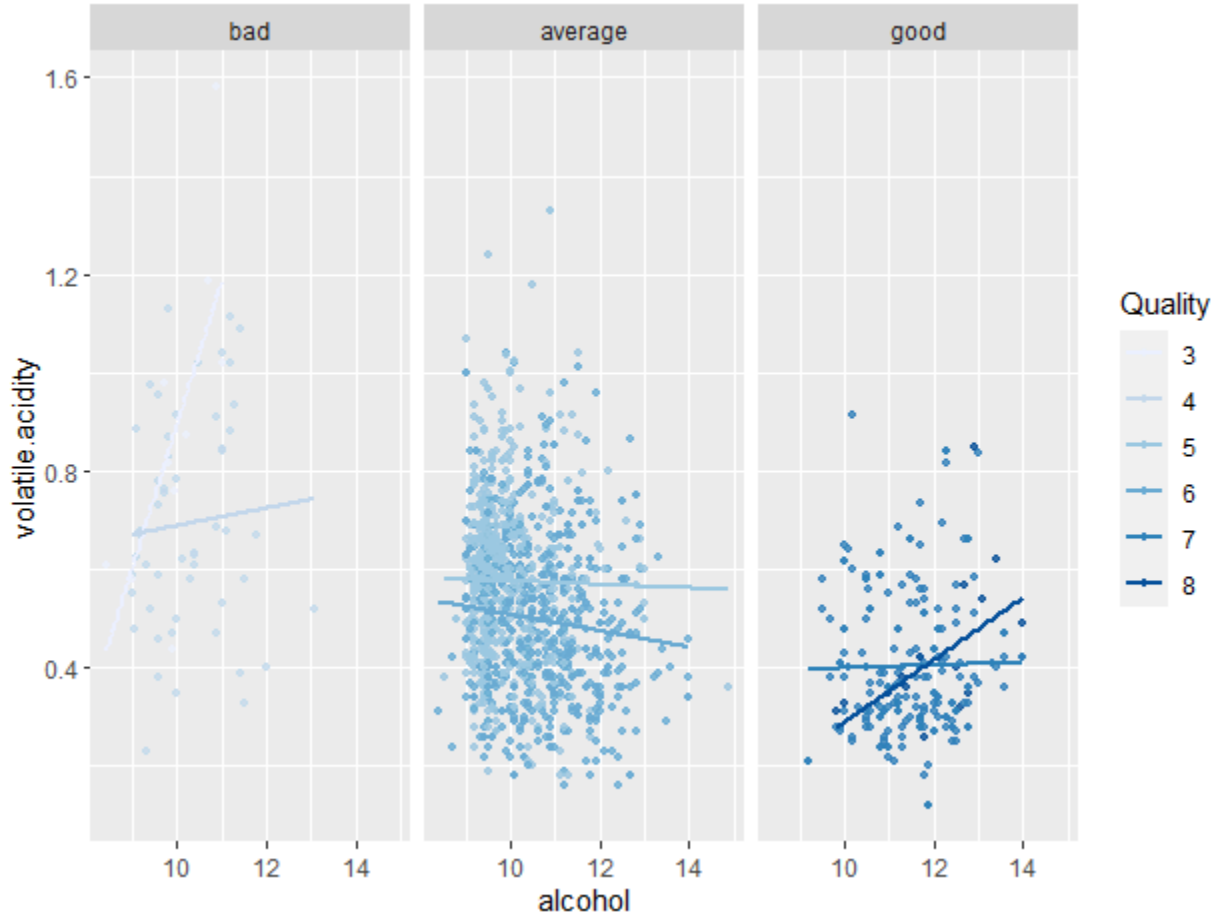
As we saw, alcohol, which was thought to have a significant role in the quality of the wine, actually contributes only 22% of the entire quality. Therefore, we will first hold alcohol constant and then try to introduce a few additional variables to see whether they have any other effects on the quality as a whole.



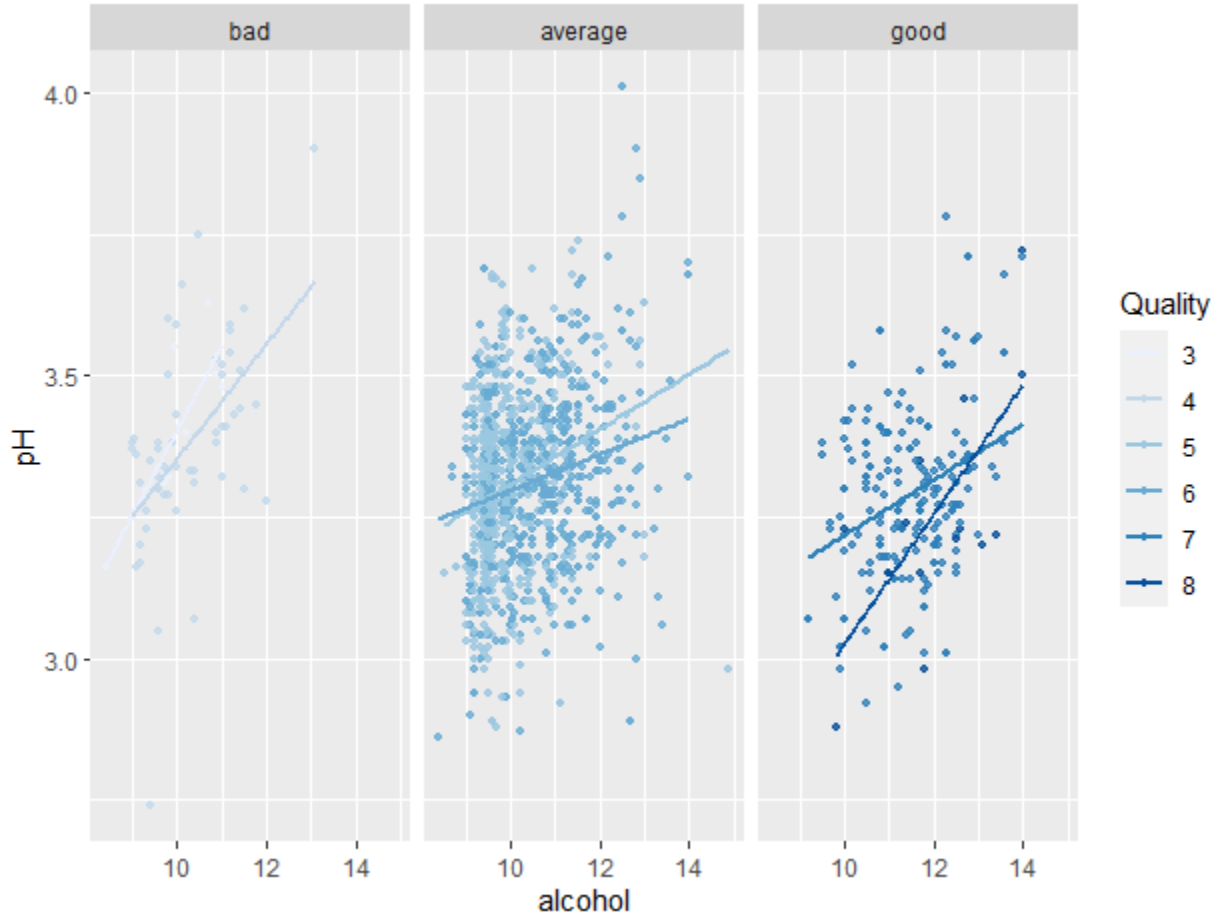
The correlation between density and quality that we had previously seen must therefore be valid, and the cause of this correlation was probably the alcohol percentage.



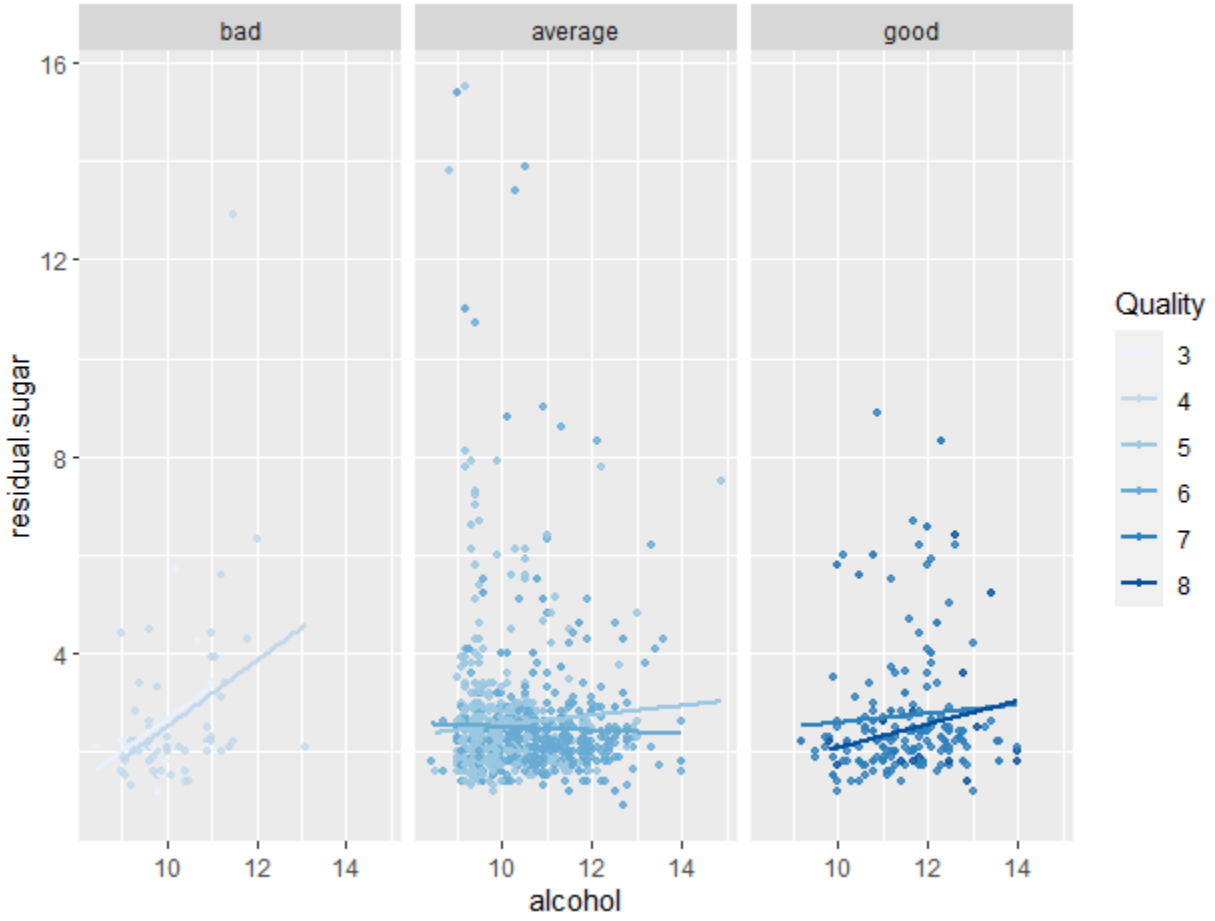
It appears that wines with more alcohol make superior wine if they include more sulphates.



Wines appear to be produced better when volatile acid levels are lower and alcohol levels are higher.



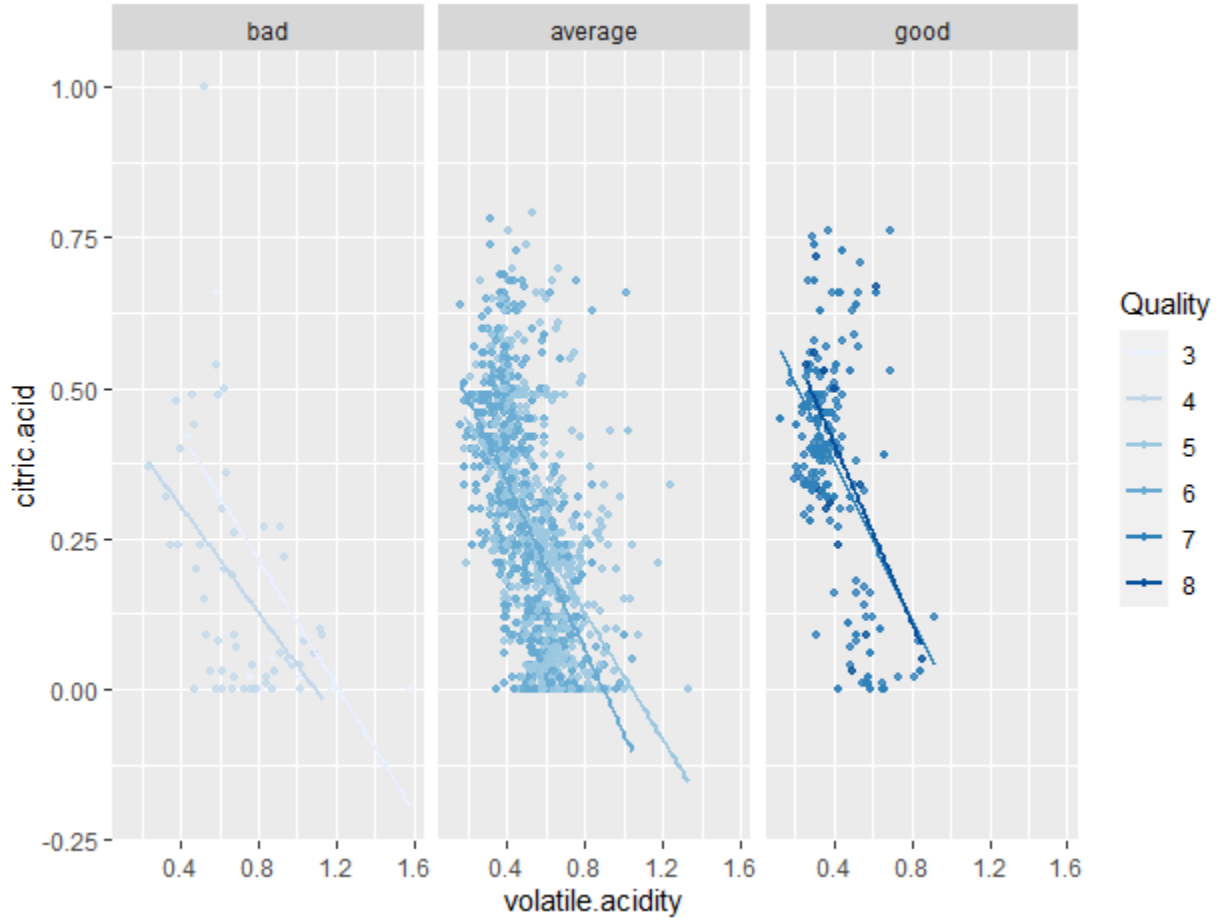
Low pH and a high alcohol content tend to result in better quality wines in this case as well.



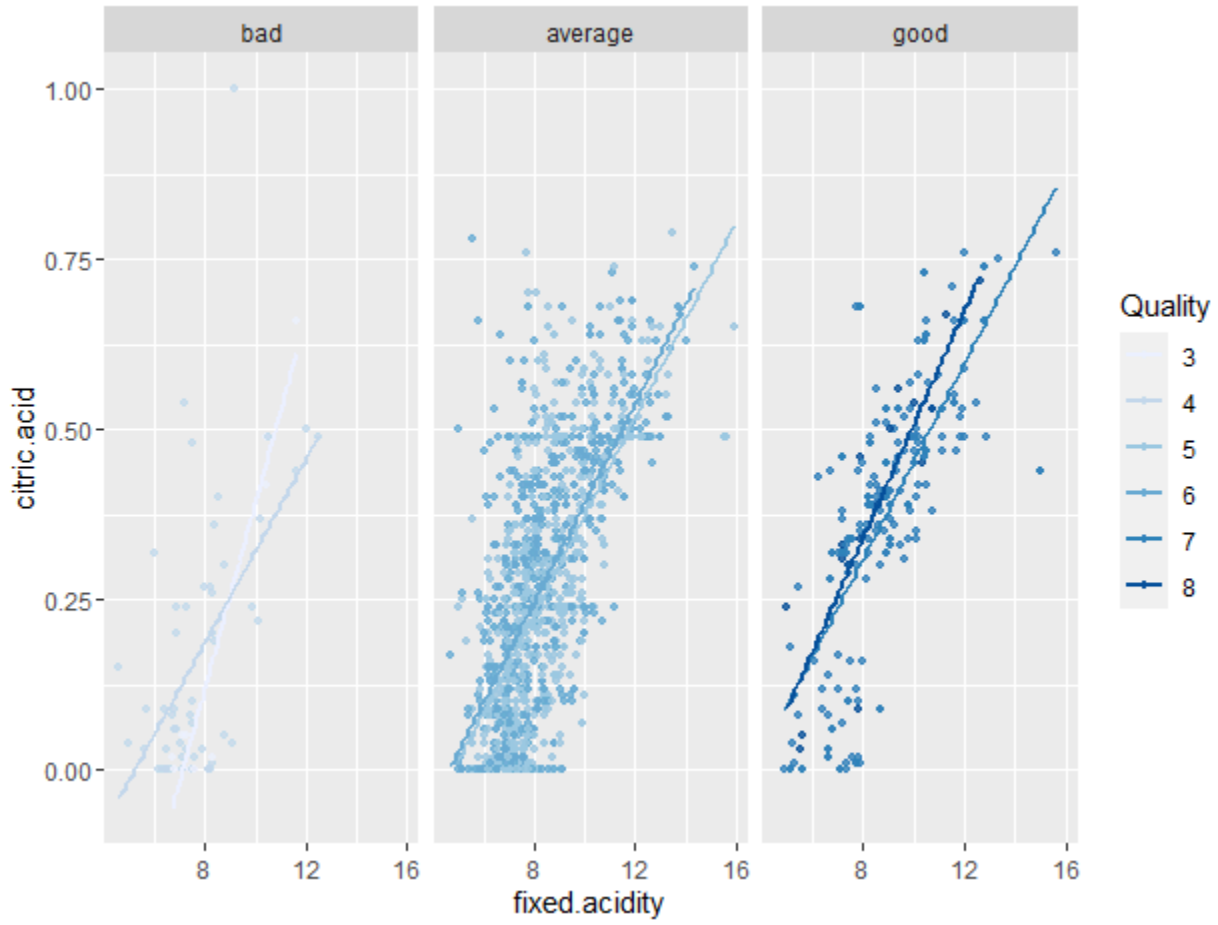
There is no connection between quality and residual sugar.

Even if there are some notable outliers for better wine with high Sulphur Dioxide, lower Sulphur Dioxide generally seems to make better wine.

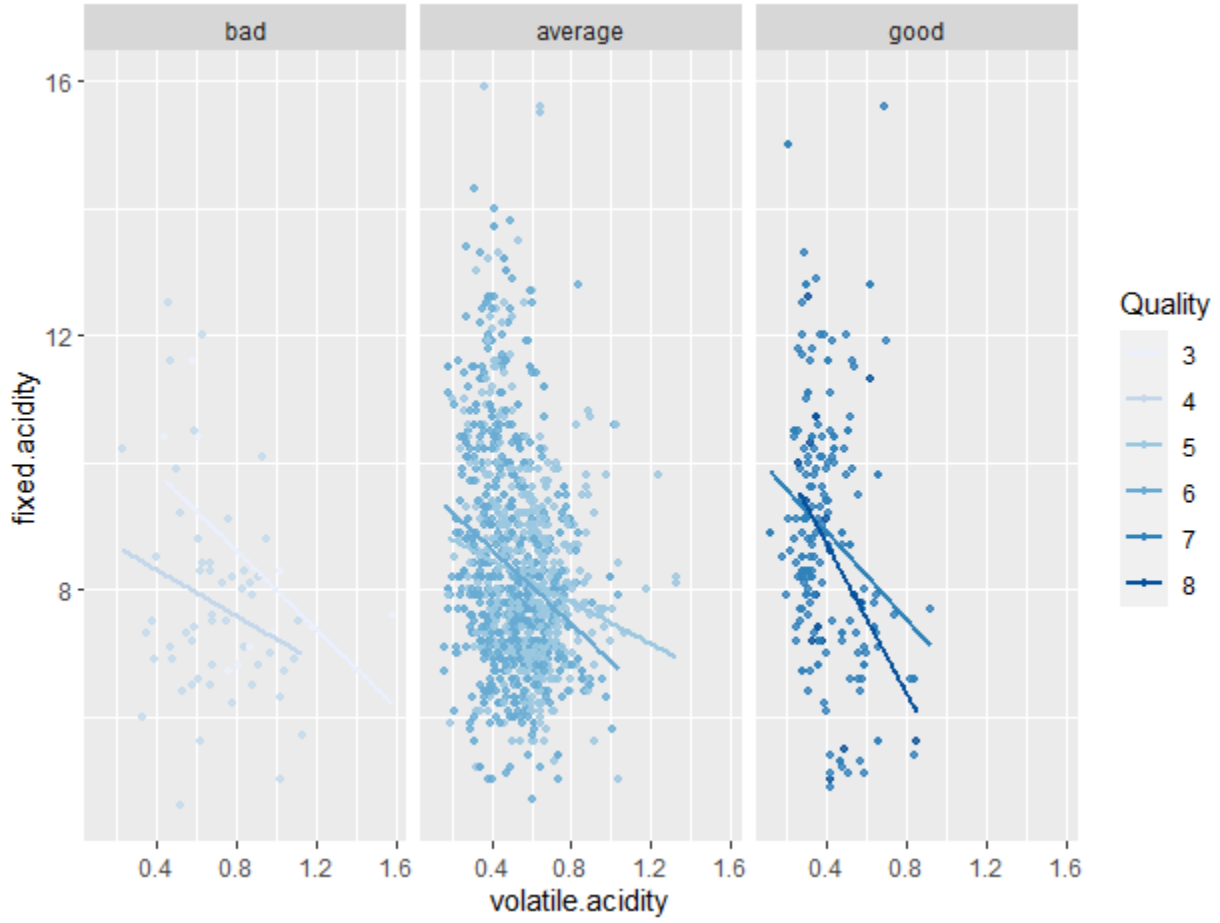
Let's now attempt to study how acids affect the quality of wines.



Better wines appear to be produced when the acidity is higher and the acidity is lower.



Not much correlations can be seen here.



Once more, there is little correlation between this and quality.

Linear modelling

Following all of the analysis, we will create a linear model using the factors that have the strongest correlations with wine quality.

Summaries of Linear Models:

```
> summary(m1)

Call:
lm(formula = as.numeric(quality) ~ alcohol, data = training_data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8163 -0.3974 -0.1581  0.5255  2.5255

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.05638    0.22593    0.25    0.803
alcohol      0.34181    0.02159   15.83   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7149 on 957 degrees of freedom
Multiple R-squared:  0.2075,    Adjusted R-squared:  0.2067
F-statistic: 250.6 on 1 and 957 DF,  p-value: < 2.2e-16
```

```
> summary(m2)

Call:
lm(formula = as.numeric(quality) ~ alcohol + sulphates, data = train

Residuals:
    Min      1Q  Median      3Q     Max
-2.6489 -0.3840 -0.1042  0.4965  2.2036

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.39107    0.22907  -1.707   0.0881 .
alcohol      0.32620    0.02117  15.411  < 2e-16 ***
sulphates    0.92188    0.12946   7.121  2.1e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6971 on 956 degrees of freedom
Multiple R-squared:  0.2474,    Adjusted R-squared:  0.2459
F-statistic: 157.2 on 2 and 956 DF,  p-value: < 2.2e-16
```

```
> summary(m3)

Call:
lm(formula = as.numeric(quality) ~ alcohol + sulphates + volatile.acidity,
    data = training_data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.70122 -0.36852 -0.06423  0.46883  2.00844

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.79493    0.25391   3.131   0.0018 **
alcohol            0.29030    0.02065  14.061   <2e-16 ***
sulphates          0.65682    0.12728   5.160    3e-07 ***
volatile.acidity -1.21044    0.13029  -9.290   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6679 on 955 degrees of freedom
Multiple R-squared:  0.3098,    Adjusted R-squared:  0.3076
F-statistic: 142.9 on 3 and 955 DF,  p-value: < 2.2e-16
```

```
> summary(m4)

Call:
lm(formula = as.numeric(quality) ~ alcohol + sulphates + volatile.acidity +
    citric.acid, data = training_data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.70230 -0.36661 -0.06223  0.46692  2.00525

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.78522    0.26197   2.997  0.00279 **
alcohol            0.29036    0.02066  14.055  < 2e-16 ***
sulphates          0.65281    0.13005   5.020 6.17e-07 ***
volatile.acidity -1.19883    0.15113  -7.932 6.00e-15 ***
citric.acid        0.02047    0.13484   0.152  0.87937
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6682 on 954 degrees of freedom
Multiple R-squared:  0.3098,    Adjusted R-squared:  0.3069
F-statistic: 107.1 on 4 and 954 DF,  p-value: < 2.2e-16


> summary(m5)

Call:
lm(formula = as.numeric(quality) ~ alcohol + sulphates + volatile.acidity +
    citric.acid + fixed.acidity, data = training_data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.76007 -0.37620 -0.05664  0.46783  2.05235

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.36939    0.28934   1.277 0.202035
alcohol            0.29990    0.02075  14.450  < 2e-16 ***
sulphates          0.66421    0.12942   5.132 3.47e-07 ***
volatile.acidity -1.28763    0.15273  -8.431  < 2e-16 ***
citric.acid      -0.38618    0.18193  -2.123 0.034043 *
fixed.acidity      0.05569    0.01683   3.309 0.000972 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6648 on 953 degrees of freedom
Multiple R-squared:  0.3177,    Adjusted R-squared:  0.3141
F-statistic: 88.73 on 5 and 953 DF,  p-value: < 2.2e-16
```

```
> summary(m6)

Call:
lm(formula = as.numeric(quality) ~ alcohol + sulphates + pH,
    data = training_data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.68737 -0.38339 -0.08869  0.49670  2.13450

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.6541     0.5242   3.155  0.00165 **
alcohol       0.3466     0.0215  16.124  < 2e-16 ***
sulphates     0.7869     0.1320   5.961 3.53e-09 ***
pH           -0.6564     0.1517  -4.328 1.66e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6907 on 955 degrees of freedom
Multiple R-squared:  0.2619,    Adjusted R-squared:  0.2596
F-statistic:   113 on 3 and 955 DF,  p-value: < 2.2e-16
```

Analysis of the Multivariate Plots

Observations

1. Wines with higher sulphate and alcohol levels seem to age better.
2. Despite their slight correlation, citric acid and wine quality are associated.

Linear Models Created

A few linear models have been built by us. However, there were insufficient statistics to provide the equations that were generated, a meaningful level of confidence. Due to the low R squared value, we were able to determine that alcohol accounts for just 22% of the wine quality and that the majority of the variables converged on wines of Average quality.This may be because the

majority of the wines in our dataset were of 'Average' quality, and the training dataset contained very little information about the wines of 'Good' and 'Bad' quality, making it challenging to forecast statistics for the edge cases. Perhaps a larger dataset would have made it easier to forecast the higher range numbers.

Final Plots and Summary

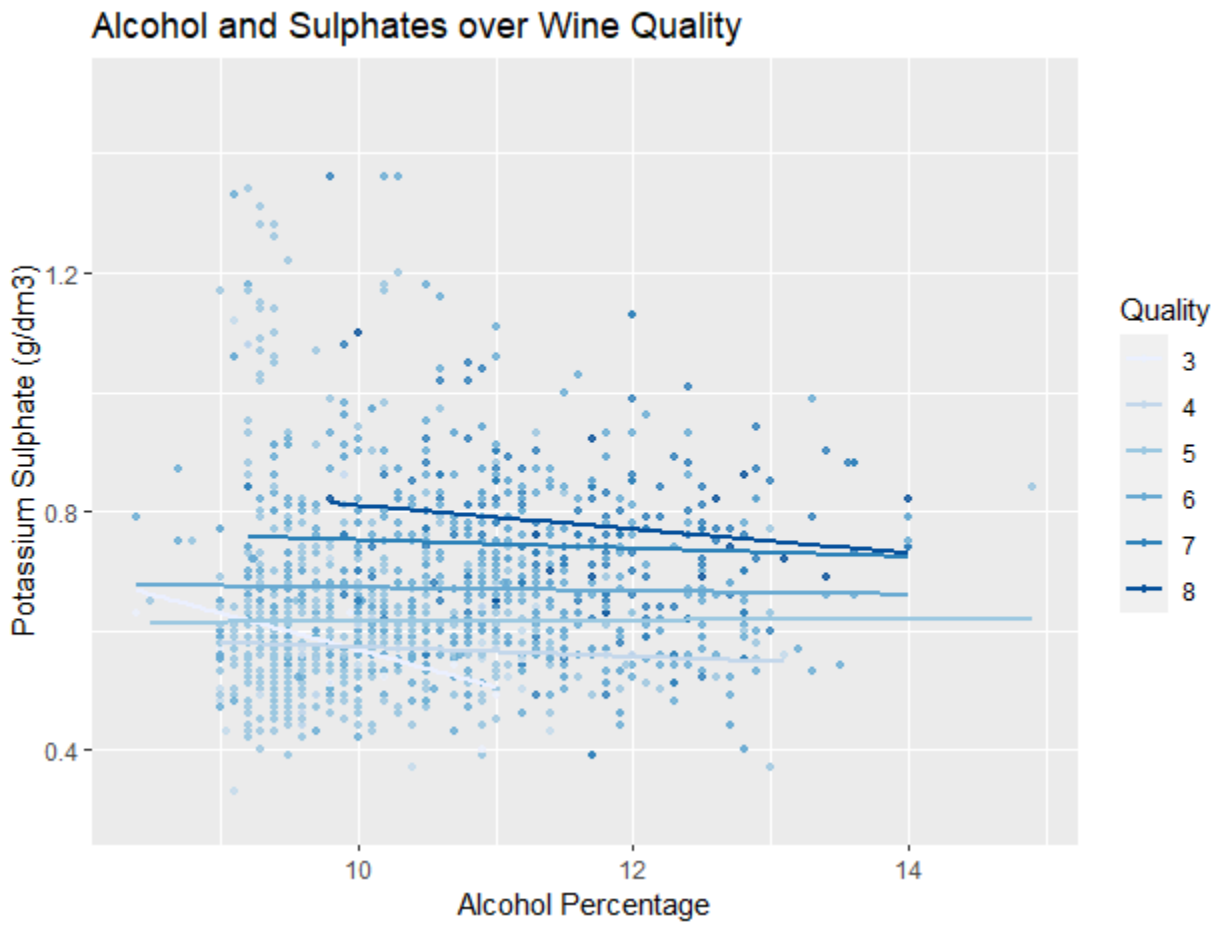
We saw that a key factor in assessing the quality of alcohol was the presence of sulphates. These three plots are absolutely essential to this project. Therefore, we have chosen to include these three plots in the section on final plots and summary.

Plot 1

This plot tells us that Alcohol percentage has played a big role in determining the quality of Wines. The higher the alcohol percentage, the better the wine quality. In this dataset, even though most of the data pertains to average quality wine, we can see from the above plot that

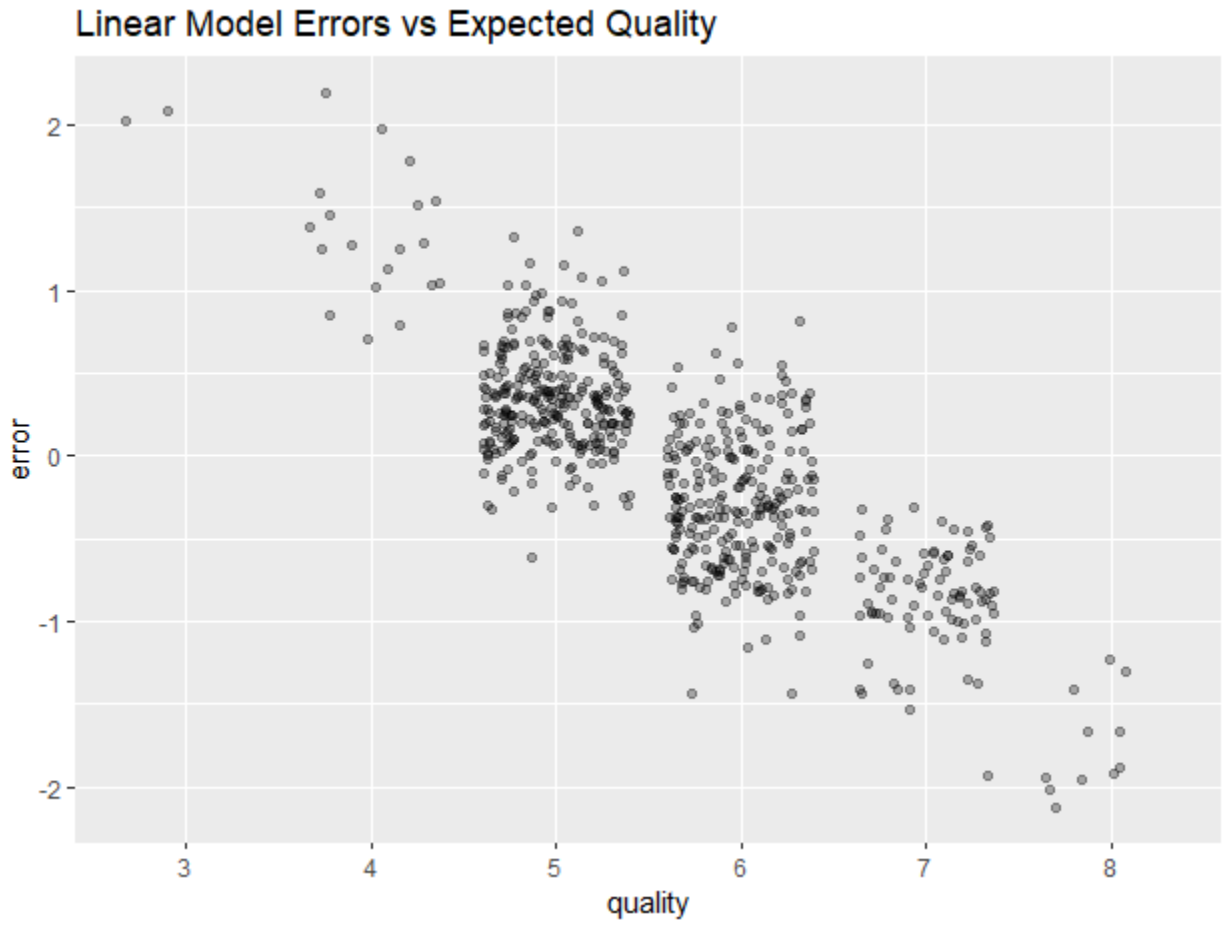
the mean and median coincides for all the boxes implying that for a particular Quality it is very normally distributed. So a very high value of the median in the best quality wines imply that almost all points have a high percentage of alcohol. But previously from our linear model test, we saw from the R Squared value that alcohol alone contributes to about 22% in the variance of the wine quality.So alcohol is not the only factor which is responsible for the improvement in Wine Quality.

Plot 2



The best wines, according to this plot, have high values for both the alcohol percentage and the sulphate concentration, suggesting that higher alcohol contents and higher sulphate concentrations work together to generate superior wines. Although there is a very little downward slope, this may be because the alcohol percentage in the finest wines is somewhat higher than the sulphate content.

Plot 3



We can see that compared to the 'Good' and 'Bad' grade wines, the error is substantially more prevalent in the 'Average' quality segment. This is clear from the fact that the majority of the wines in our dataset are of 'Average' quality, and there isn't a lot of information in the extreme ranges. Only around 33% of the quality variation could be accounted for by the linear model with the R squared value for m5.

Conclusion

The conclusion and inferences that can be drawn from this project are as follows:

Wine Quality Analysis: Through exploratory data analysis, we examined the various attributes of red wine and their distribution. We observed that the quality of wine is subjective and can be influenced by multiple factors

Correlation Analysis: We conducted correlation analysis to understand the relationships between different variables and wine quality. We identified several variables, such as alcohol, volatile acidity, sulphates, and citric acid, that showed significant correlations with wine quality.

Linear Model: We built a linear regression model to predict wine quality based on the selected variables. The model demonstrated reasonable predictive power, as indicated by the coefficients and statistical significance of the predictors. However, it is important to note that the model's performance can be further improved by exploring alternative algorithms and feature engineering techniques.

Predictive Analysis: Using the trained model, we made predictions on a test dataset and evaluated the performance of the model by comparing the predicted wine quality with the actual quality. The scatter plot of the model errors showed a relatively random distribution around zero, indicating that the model was making reasonably accurate predictions.

Future Scope: The project has several future directions, such as exploring alternative modelling techniques, incorporating external data sources, conducting more in-depth feature selection, and applying ensemble methods to improve the predictive accuracy. Additionally, the model can be deployed in real-world scenarios, and further domain-specific analysis can be conducted to gain deeper insights into the factors affecting wine quality.

Scope of improvement

Model Improvement: The current project used a linear regression model to predict wine quality. Further exploration can involve trying different machine learning algorithms, such as decision trees, random forests, or support vector machines, to see if they yield better predictive performance. Additionally, feature engineering techniques can be applied to create new variables or combinations of variables that might improve the model's accuracy.

Feature Selection: Conducting a more comprehensive feature selection process can help identify the most important variables that contribute to wine quality prediction. This can involve techniques such as forward/backward stepwise selection, LASSO regularisation, or recursive feature elimination.

Cross-Validation and Model Evaluation: Implement cross-validation techniques, such as k-fold cross-validation, to assess the stability and generalizability of the model. Calculate additional evaluation metrics, such as mean squared error (MSE), root mean squared error (RMSE), or R-squared, to gain a more comprehensive understanding of the model's performance.

Ensemble Methods: Explore the use of ensemble methods, such as bagging or boosting, to combine multiple models and improve prediction accuracy. This can involve training multiple models on different subsets of the data or applying weighted voting schemes to make predictions.

Domain-Specific Analysis: Conduct further analysis to understand the relationships between wine quality and specific variables in more depth. For example, analyse the effect of specific combinations of variables or interactions between variables on wine quality. This can provide insights into the underlying factors that contribute to wine quality and help in refining the predictive models.

External Data Sources: Consider incorporating external data sources, such as climate data, soil characteristics, or grape variety information, to enhance the predictive power of the model. This additional information can provide valuable insights into how environmental factors impact wine quality.

Deployment and Application: Once the model is refined and validated, consider deploying it in a real-world setting. This can involve creating a user-friendly interface or integrating it into a larger wine quality assessment system. Monitoring the performance and feedback of the deployed model can provide insights for further improvements.

These are just a few potential avenues for future development in the Red Wine Data project. The scope can be further expanded based on specific research objectives, available resources, and domain-specific considerations.

Reference Request

The public can use this dataset for study. [Cortez et al., 2009] describes the specifics.If you want to use this database, kindly cite it as follows:

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

Available at: [@Elsevier] http://dx.doi.org/10.1016/j.dss.2009.05.016

[Pre-press (pdf)] http://www3.dsi.uminho.pt/pcortez/winequality09.pdf

[bib] http://www3.dsi.uminho.pt/pcortez/dss09.bib

Title: Wine Quality

Sources

Created by: Paulo Cortez (Univ. Minho), Antonio Cerdeira, Fernando Almeida, Telmo Matos and Jose Reis (CVRVV) @ 2009