Red Wine Analysis

Scott Jue

20 March, 2019

==========

This dataset we are about to explore is of red wine quality based on 12 different variables. There are 1599 observations in this dataset.

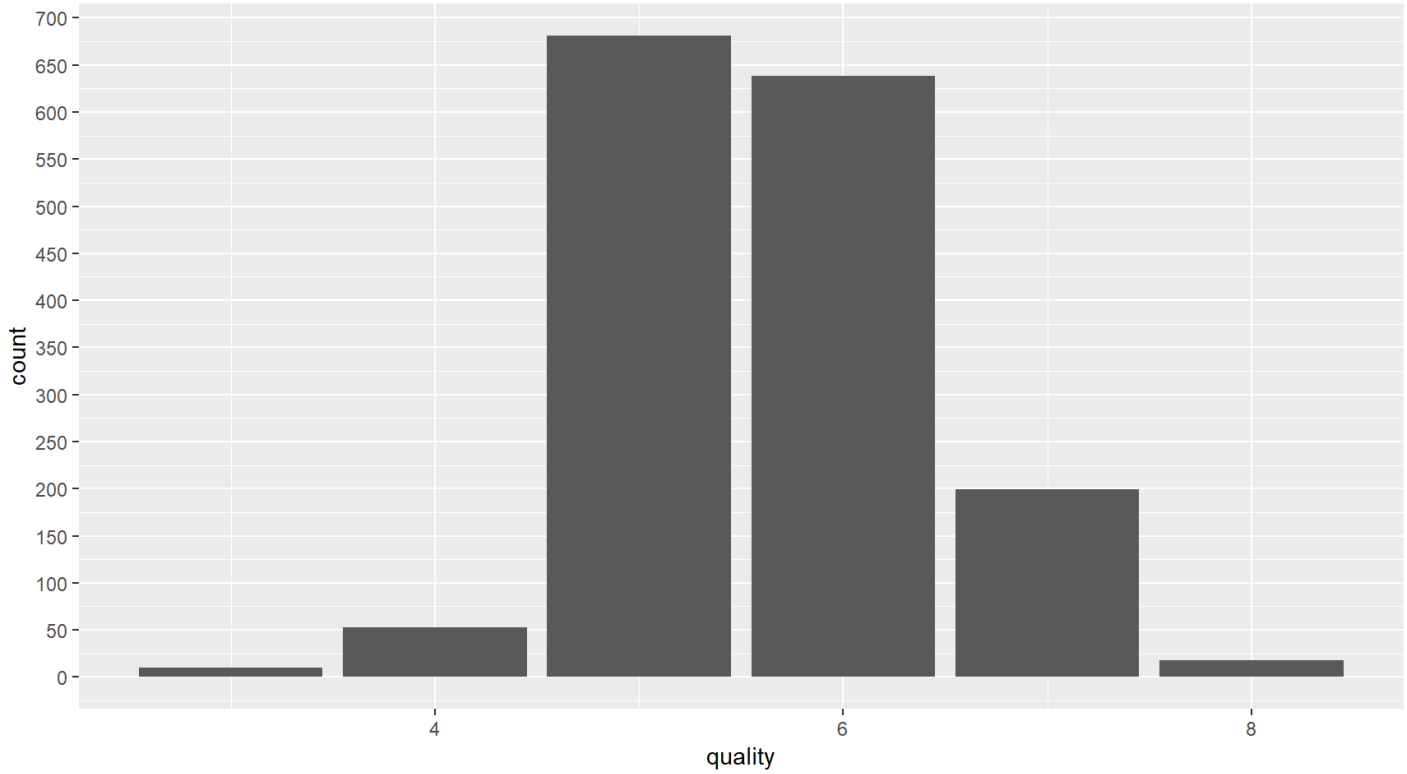
Univariate Plots Section

```
## [1] 1599   12
```

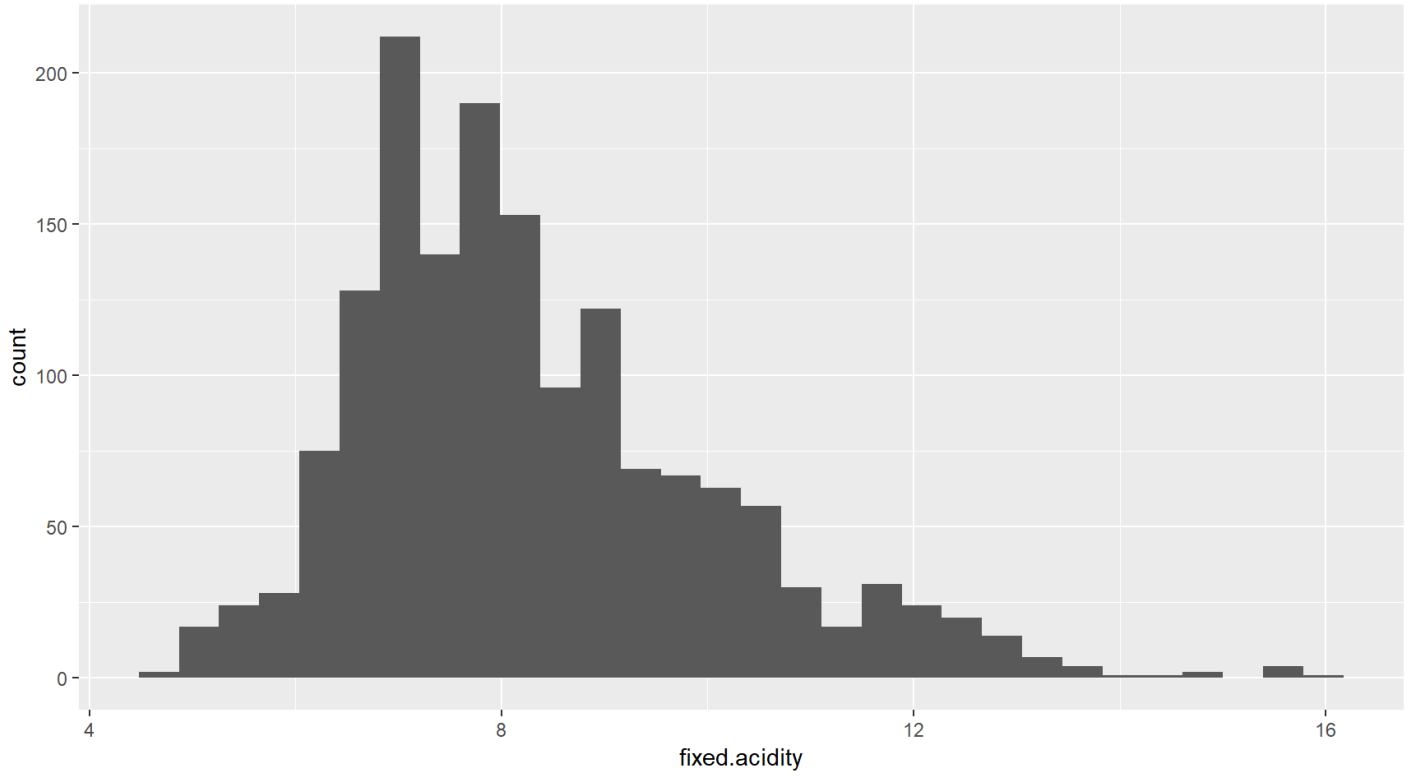
```
##  [1] "fixed.acidity"        "volatile.acidity"     "citric.acid"
##  [4] "residual.sugar"       "chlorides"            "free.sulfur.dioxide"
##  [7] "total.sulfur.dioxide" "density"              "pH"
## [10] "sulphates"            "alcohol"              "quality"
```

```
## 'data.frame':    1599 obs. of  12 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             : int  5 5 5 6 5 5 5 7 7 5 ...
```

```
##  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
##  Min.   :0.01200   Min.   : 1.00       Min.   :  6.00
##  1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00
##  Median :0.07900   Median :14.00       Median : 38.00
##  Mean   :0.08747   Mean   :15.87       Mean   : 46.47
##  3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00
##  Max.   :0.61100   Max.   :72.00       Max.   :289.00
##     density             pH          sulphates         alcohol
##  Min.   :0.9901   Min.   :2.740   Min.   :0.3300   Min.   : 8.40
##  1st Qu.:0.9956   1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50
##  Median :0.9968   Median :3.310   Median :0.6200   Median :10.20
##  Mean   :0.9967   Mean   :3.311   Mean   :0.6581   Mean   :10.42
##  3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10
##  Max.   :1.0037   Max.   :4.010   Max.   :2.0000   Max.   :14.90
##     quality
##  Min.   :3.000
##  1st Qu.:5.000
##  Median :6.000
##  Mean   :5.636
##  3rd Qu.:6.000
##  Max.   :8.000
```

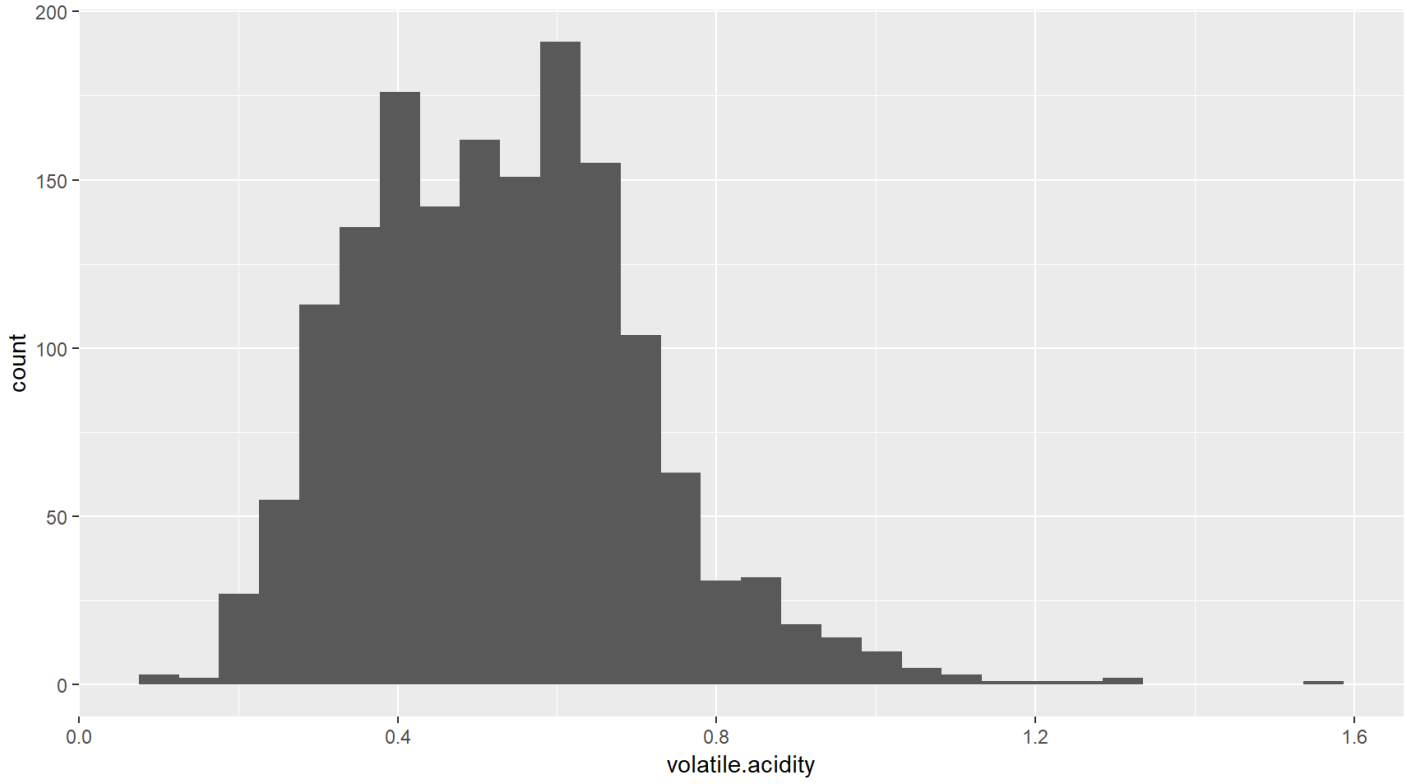


Looking at the histogram, we can notice that the distribution of quality is normal. We can also see that most of the red wine quality scored at 5 & 6 grades and no quality below 3 and there is no quality higher than 8.



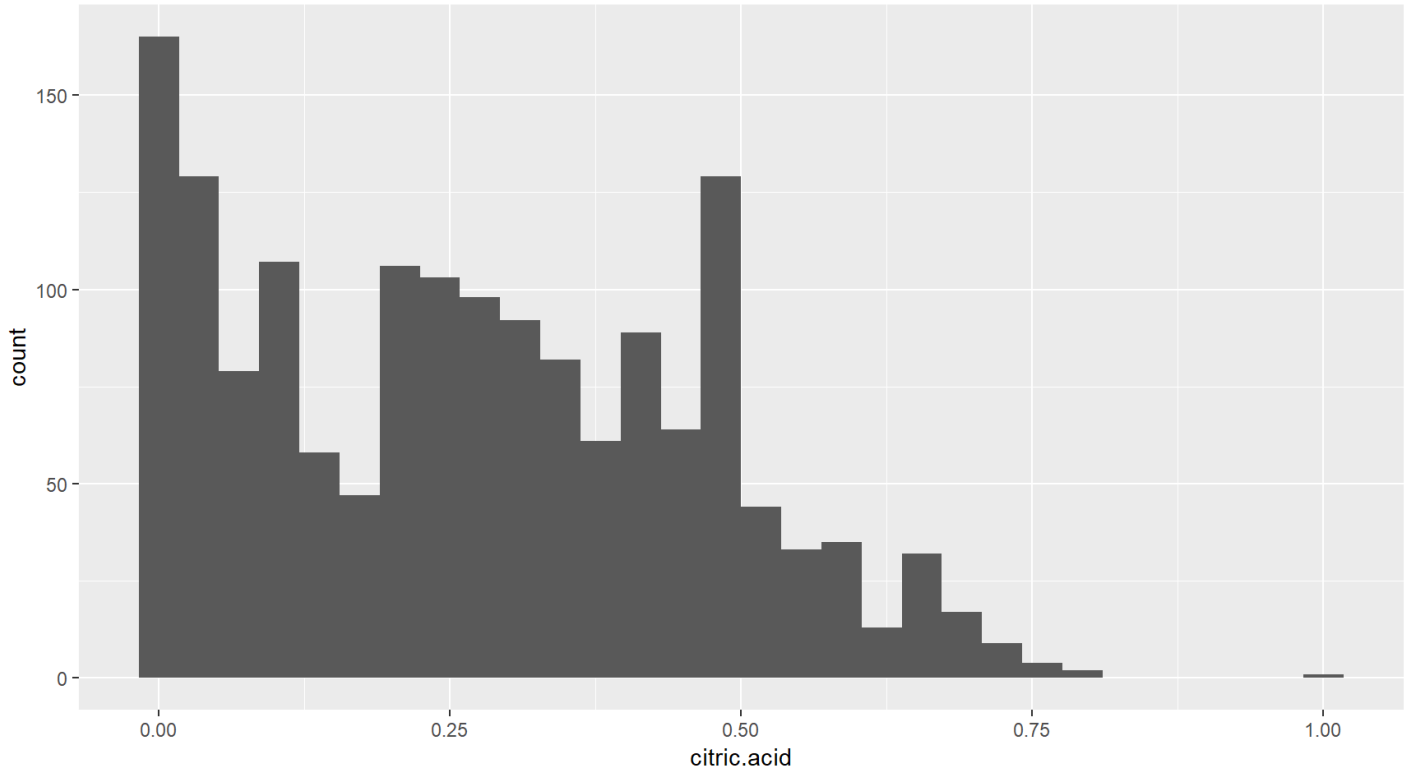

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.60    7.10    7.90    8.32    9.20   15.90
```

The distrubiton of this histogram looks to be normal with the most of the fixed acidit falling between 6 and 9. When looking at the summary we can see that there be an outlier in the data (15.9).



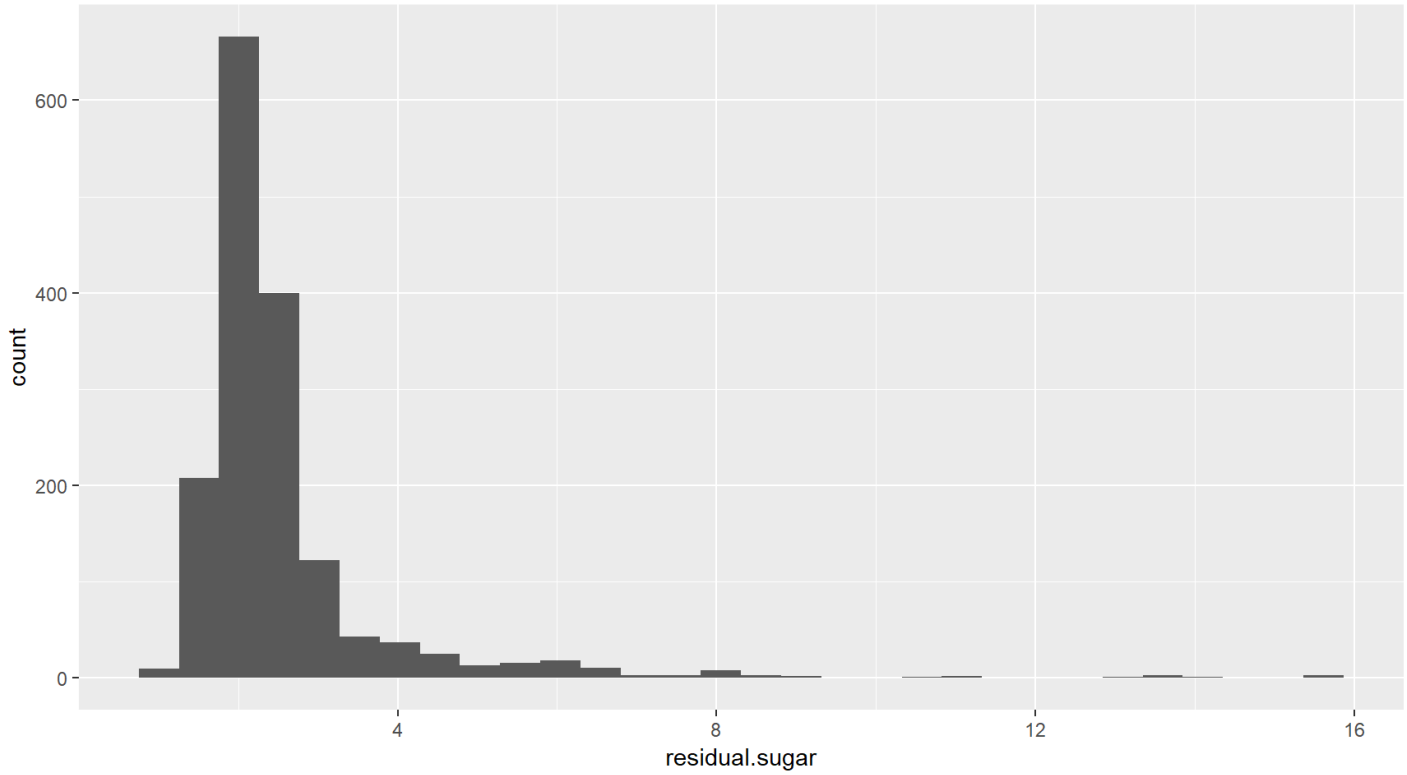
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1200  0.3900  0.5200  0.5278  0.6400  1.5800
```

The distribution of volatile acidity is normal with the mean and median values being very closed together. The majority of the volatile acidit is between 0.2 and 0.8 and some outliers are present towards the right tail.



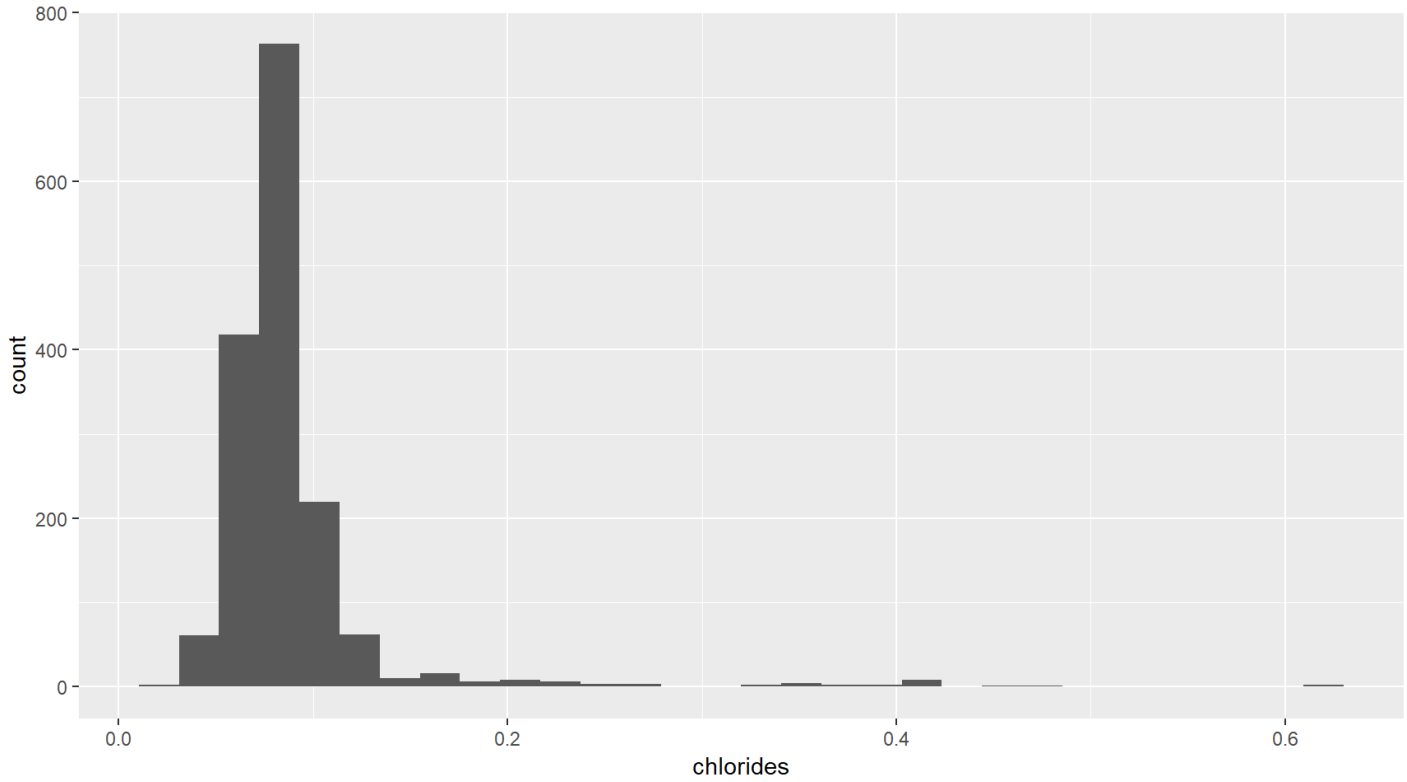

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.090   0.260   0.271   0.420   1.000
```

We can see that the mean is higher than the median which means that this distribution is positively skewed. It is also interesting to note that the mode for this distribution is 0. An outlier can also be observed with a citric acid level of 1.



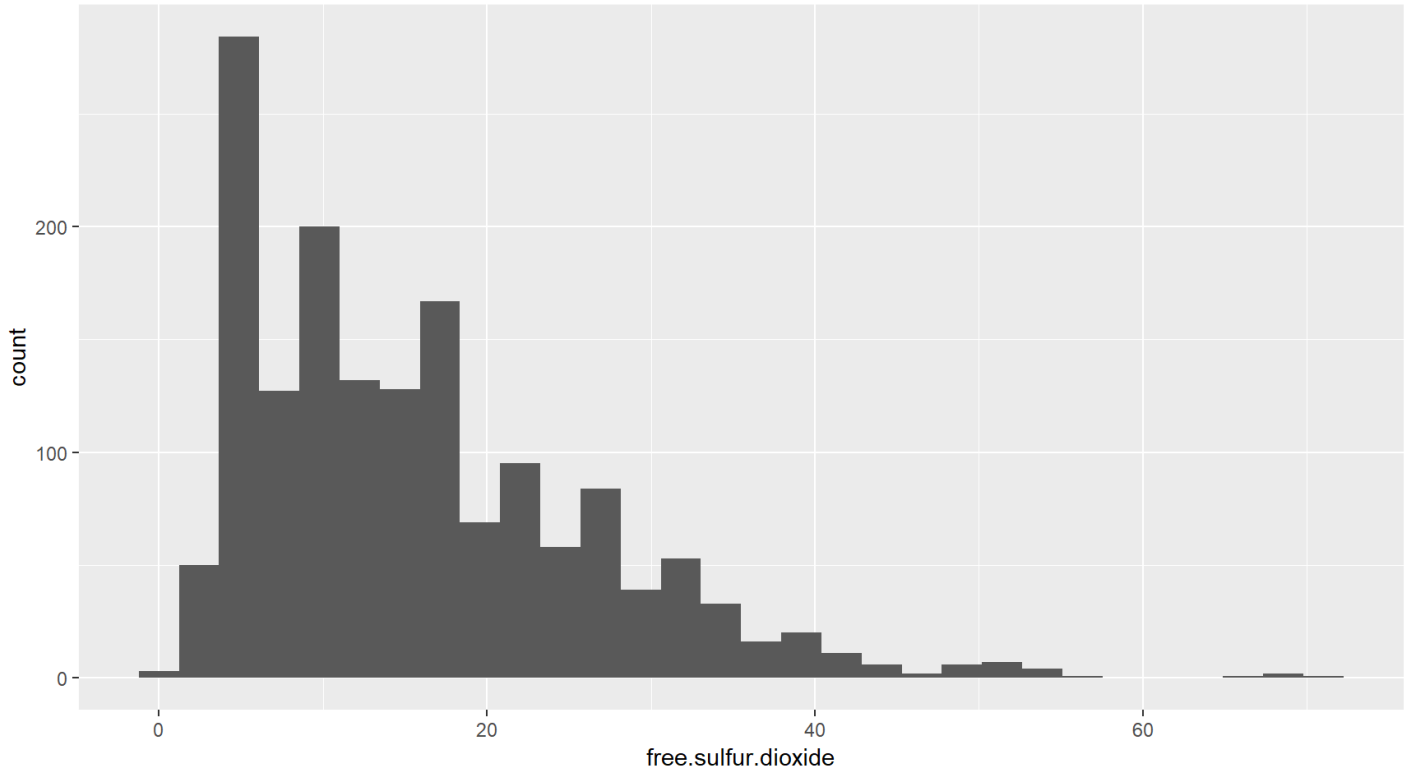

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.900   1.900   2.200   2.539   2.600  15.500
```

We can see that the mean is higher than the median which means that this distribution is positively skewed. The majority of residual sugar data falls below 2.6 and we can also see quite a few outliers with the maximum data point at 15.5.



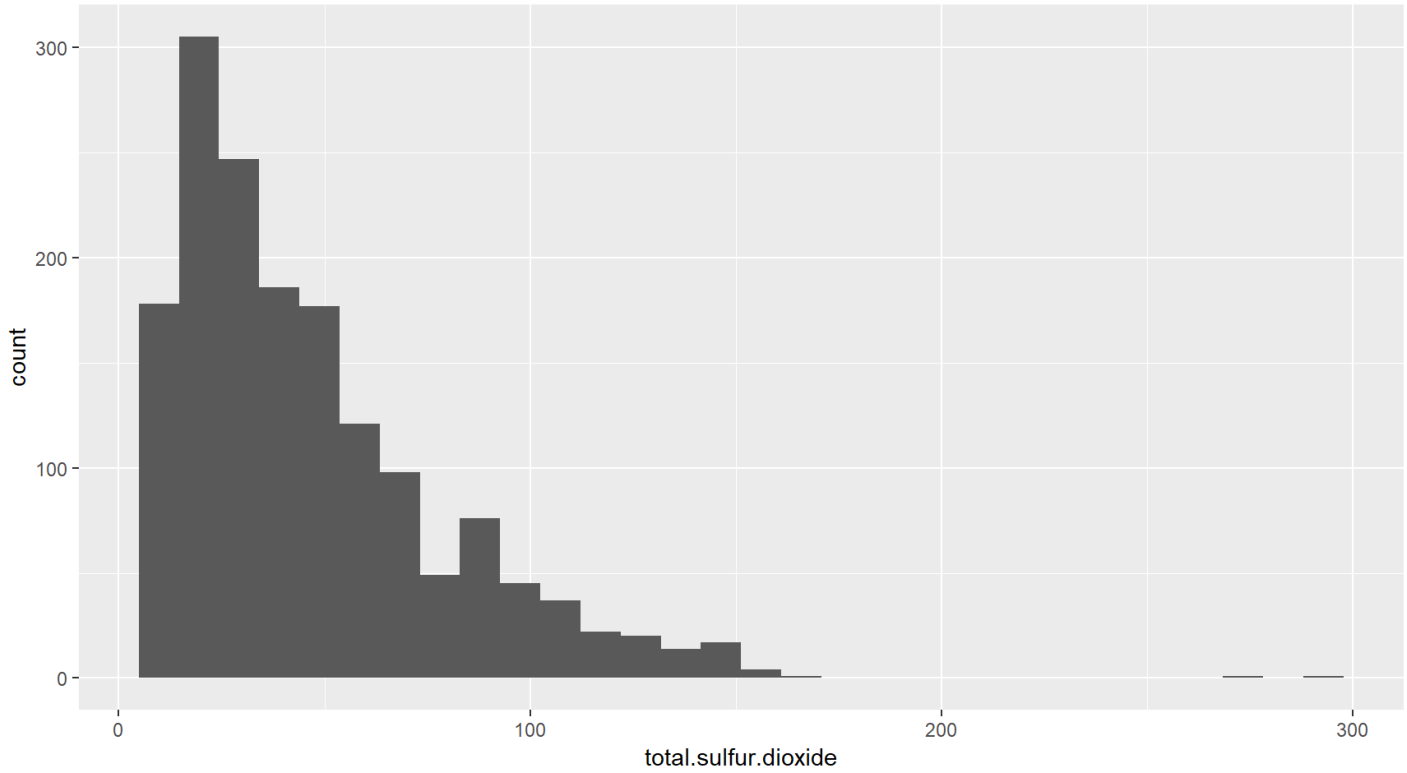
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01200 0.07000 0.07900 0.08747 0.09000 0.61100
```

The distribution of the chlorides is also positvely skewed with the majority falling around 0.1.



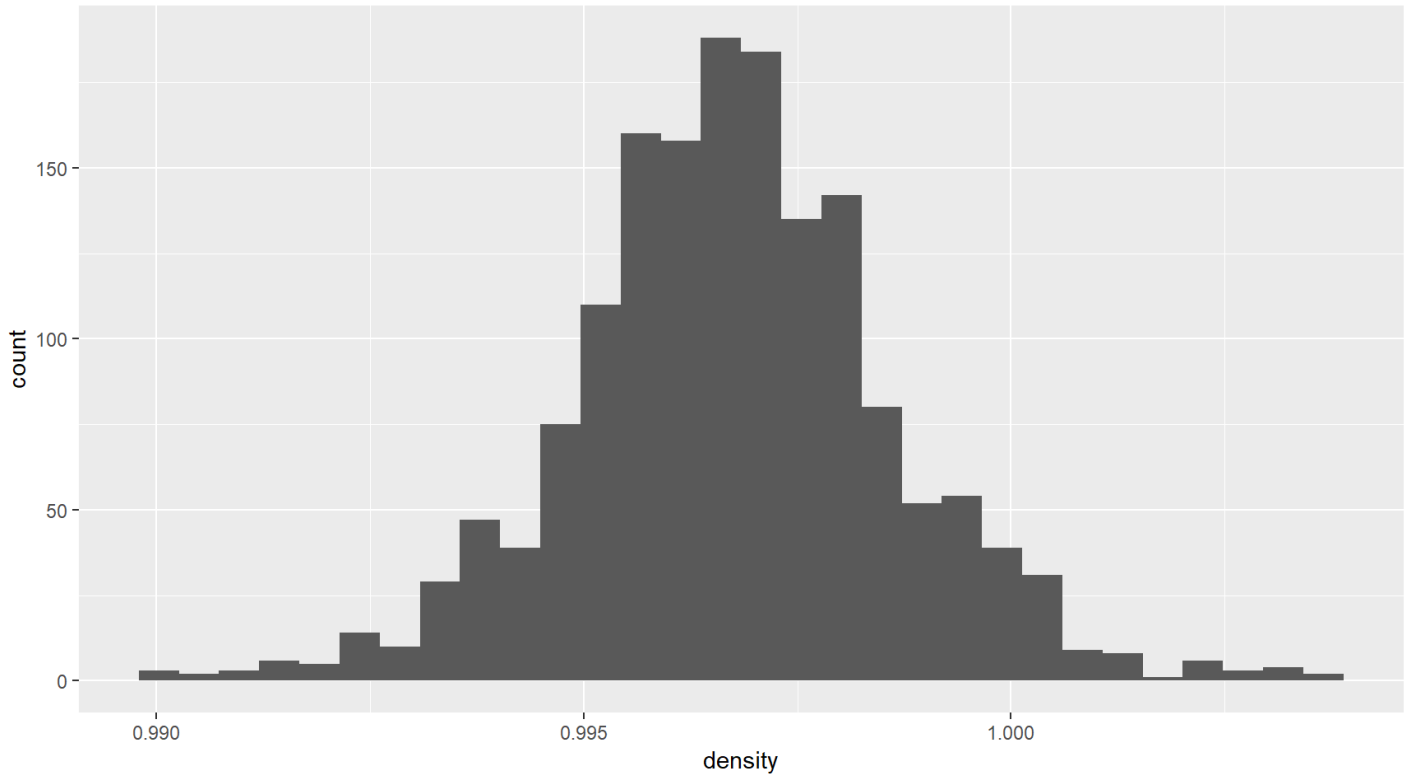

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    7.00   14.00   15.87   21.00   72.00
```

The distribution of free sulfur dioxide here is positively skewed. Most of the data falls under 21, but there can be outliers seen with a max of 72.



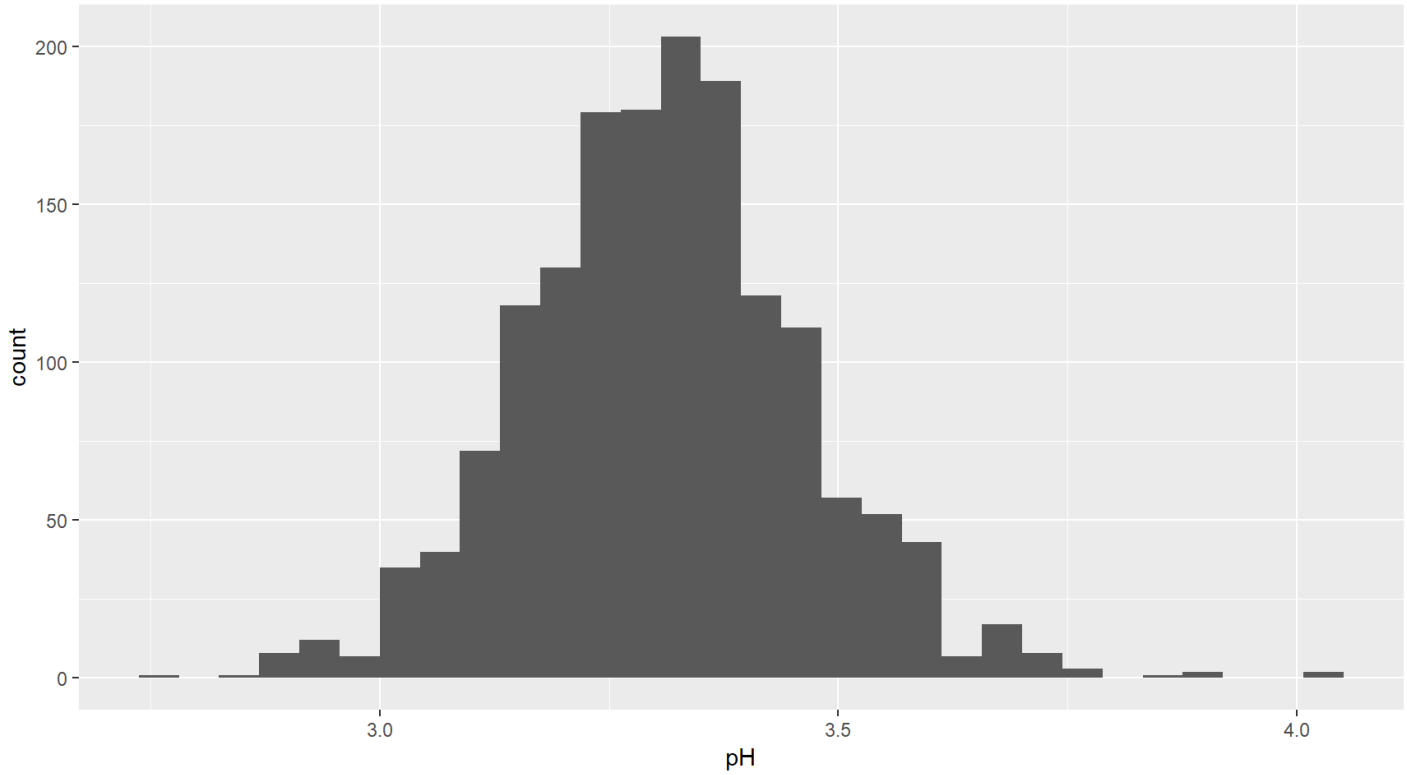

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00   22.00   38.00   46.47   62.00  289.00
```

The distribution again here is positively skewed. 3/4 of the data falls under 62, but there can be outliers observed all the way up to 289.



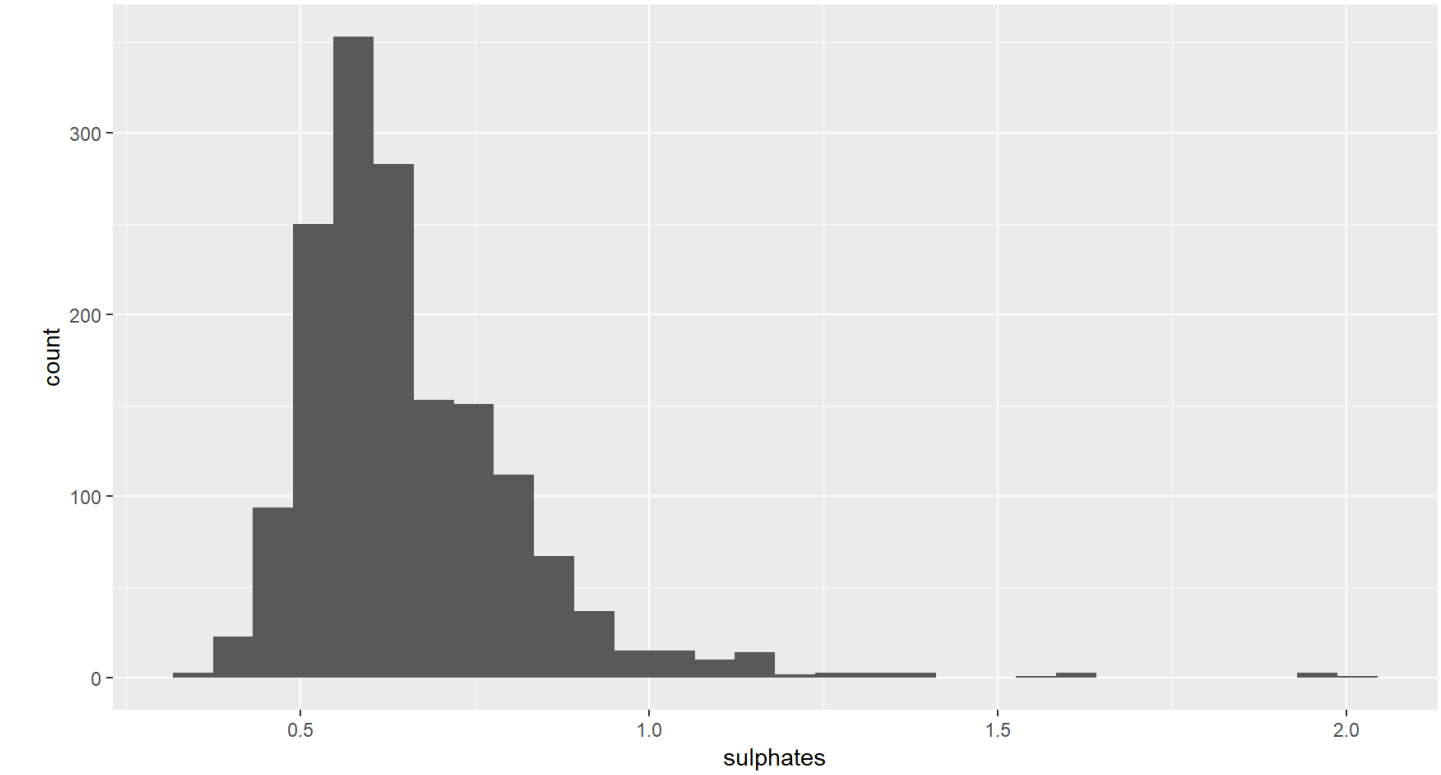
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9901  0.9956  0.9968  0.9967  0.9978  1.0037
```

The distribution is normally distributed and the mean and median are almost exactly the same.



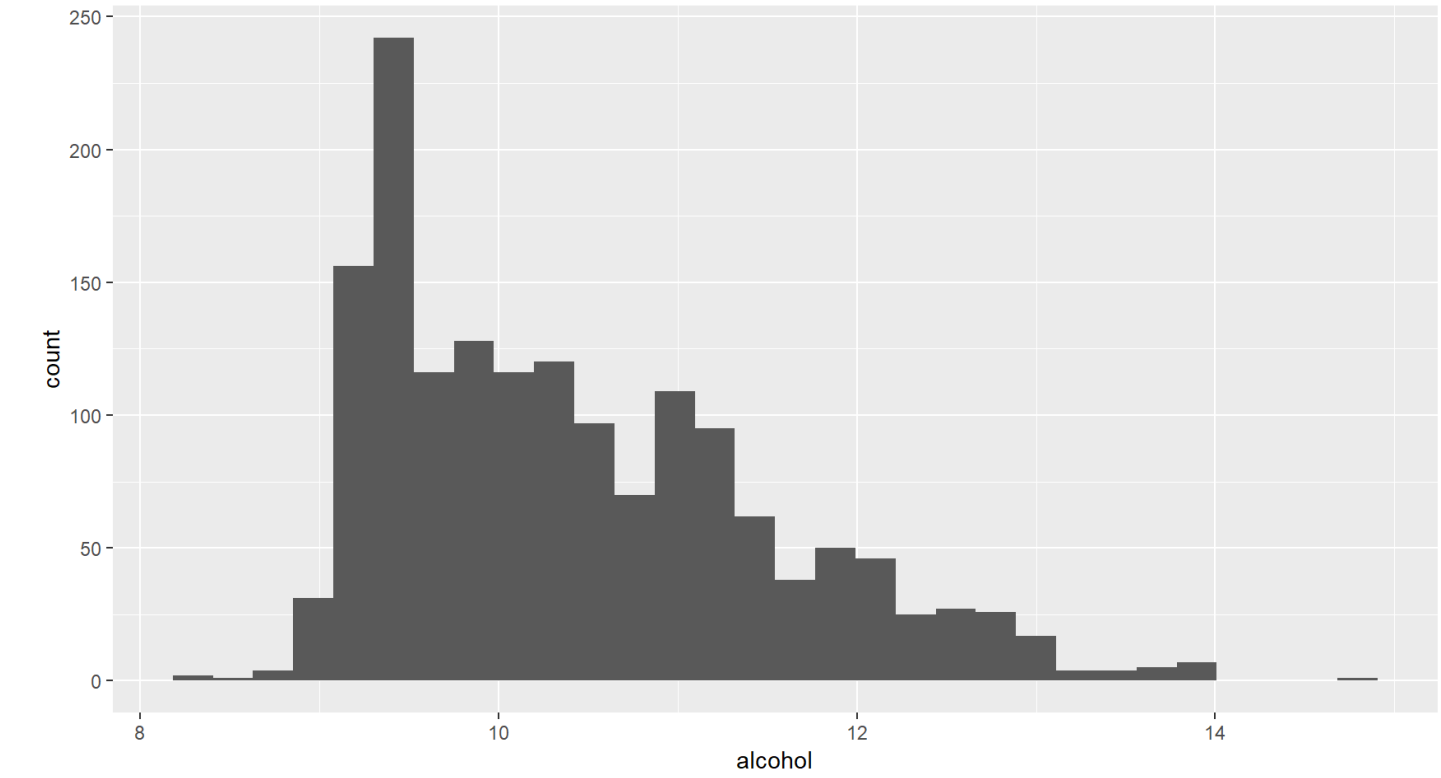

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.740   3.210   3.310   3.311   3.400   4.010
```

The distribution is normally distributed and the mean and median are almost exactly the same. Most of the data for ph is between 3.0 and 3.5.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3300  0.5500  0.6200  0.6581  0.7300  2.0000
```

This distribution is positively skewed. We can see some outliers in this sulphates dataset.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.50   10.20   10.42   11.10   14.90
```

The distibution is positively skewed. We can see an outlier value at 14.9%. The majority of wines have an alcohol percent of at least 9%.

```
##      Poor      Good Excellent
##        63      1319       217
```

In order to facilitate our analysis I have created a new variable “rating” that uses the quality data and creates a new categorical data consisting of “poor”, “good”, and “excellent”.

Univariate Analysis

What is the structure of your dataset?

Most of data are numerical. We have no string data in the original data set. There are 1599 wines with values for 12 different variables. I added a 13th variable for rating which gives categorical meaning to the quality dataset.

What is/are the main feature(s) of interest in your dataset?

The main features in the dataset are quality, however I believe the other variables in the dataset may influence the quality of red wine since these variables can affect the taste of the wine.

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

The other features of interest from this dataset that could influence the quality of wine are alcohol %, density, and acidity. I think these variables could affect the quality since they factor into how the wine tastes.

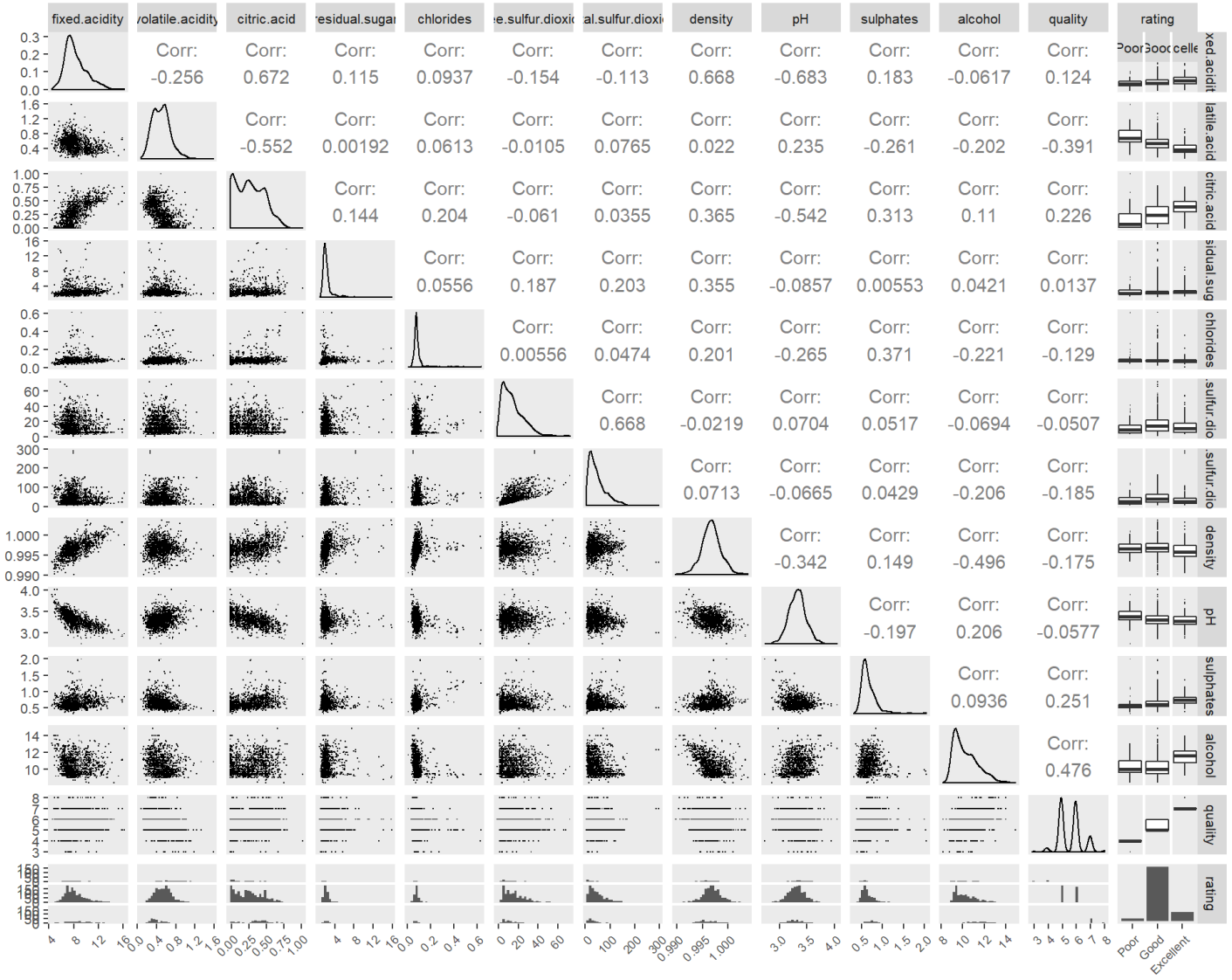
Did you create any new variables from existing variables in the dataset?

Yes, I created a new variable called “rating” which takes the quality variable breaks it up into 3 bins of Poor, Good and Excellent.

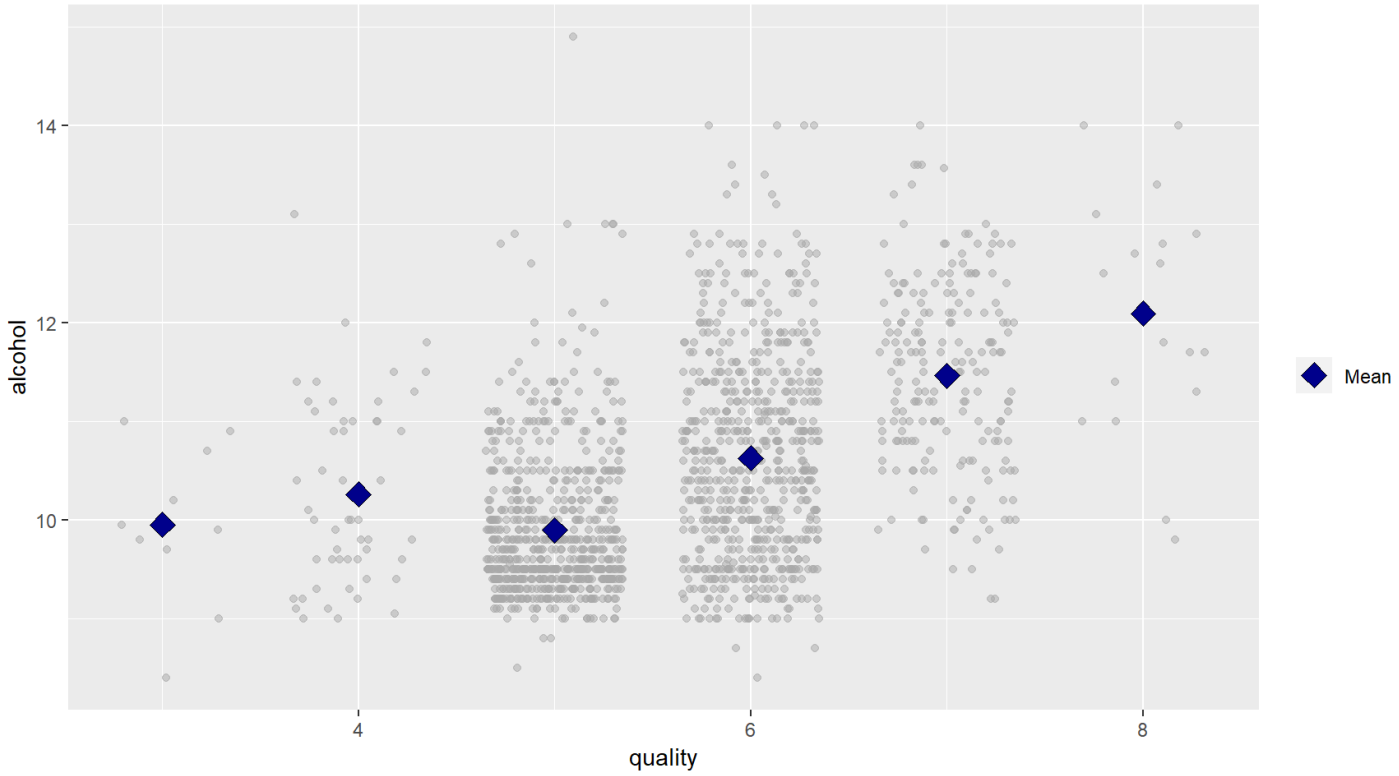
Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

Yes, I noticed positively skewed distributions in a number a different variables observed. No operations were performed on this data to tidy or adjust it because the data is already tidy.

Bivariate Plots Section

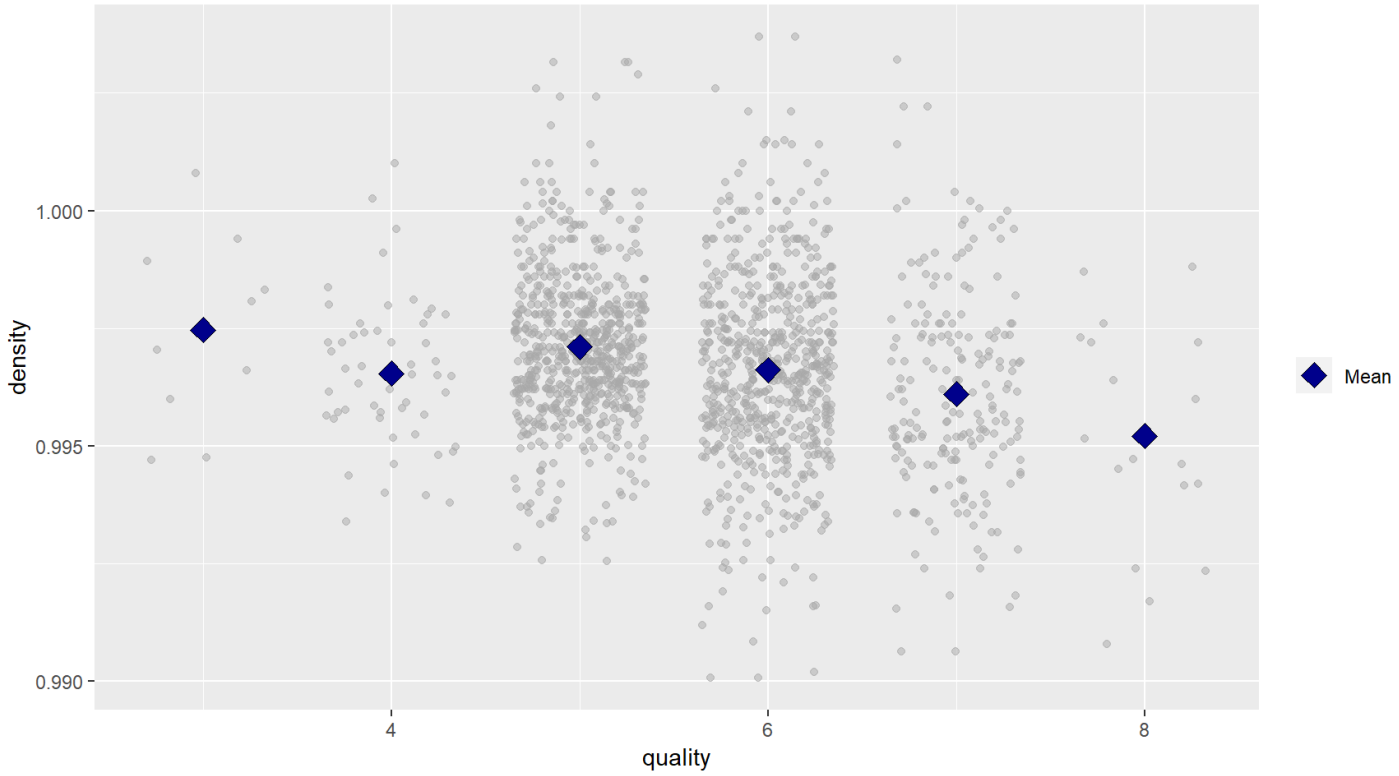


Here we can notice that both alcohol and volatile.acidity are moderately correlated with quality.



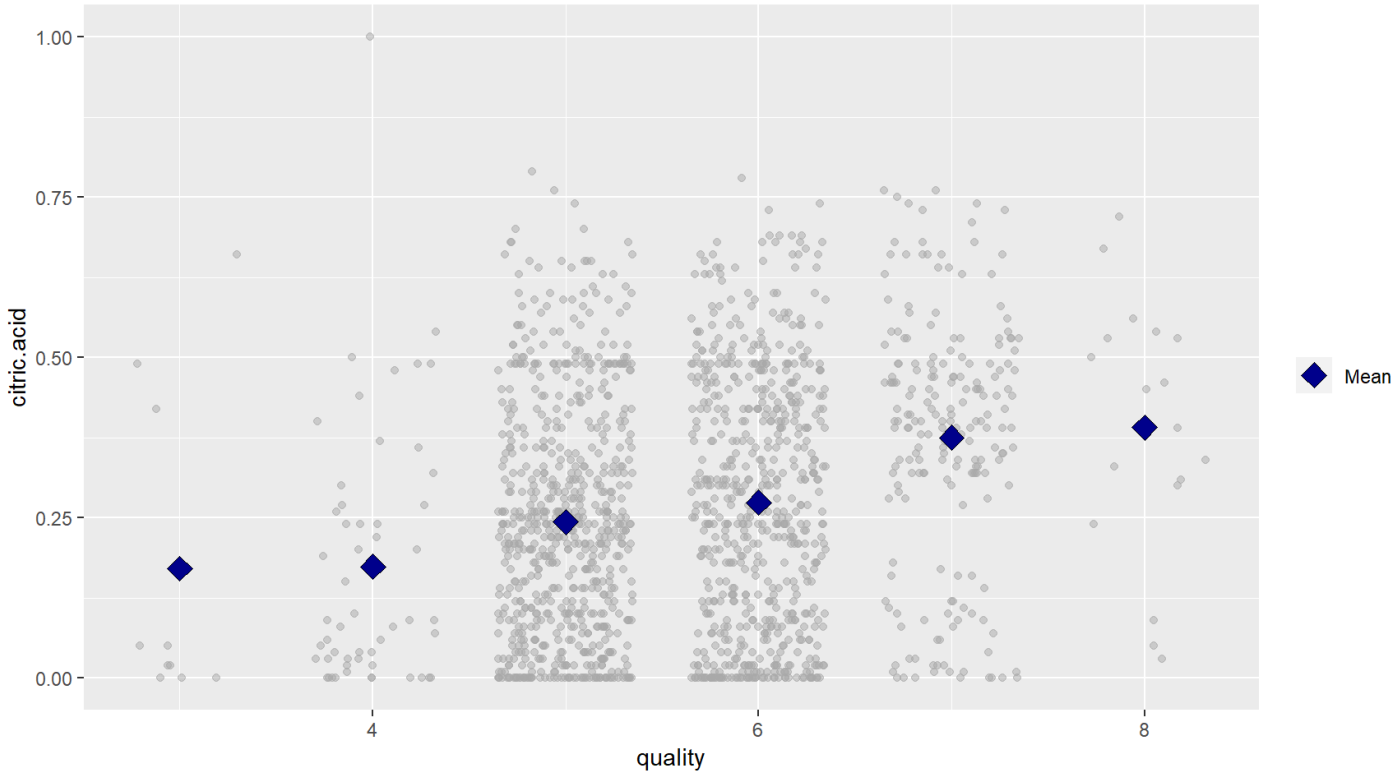

```
## wine$quality: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   8.400   9.725   9.925   9.955  10.575  11.000
## ------------------------------------------------------------
## wine$quality: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.00    9.60   10.00   10.27   11.00   13.10
## ------------------------------------------------------------
## wine$quality: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     8.5     9.4     9.7     9.9    10.2    14.9
## ------------------------------------------------------------
## wine$quality: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.80   10.50   10.63   11.30   14.00
## ------------------------------------------------------------
## wine$quality: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.20   10.80   11.50   11.47   12.10   14.00
## ------------------------------------------------------------
## wine$quality: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.80   11.32   12.15   12.09   12.88   14.00
```

Higher quality wines generally tend to have higher mean and median alcohol percent content.



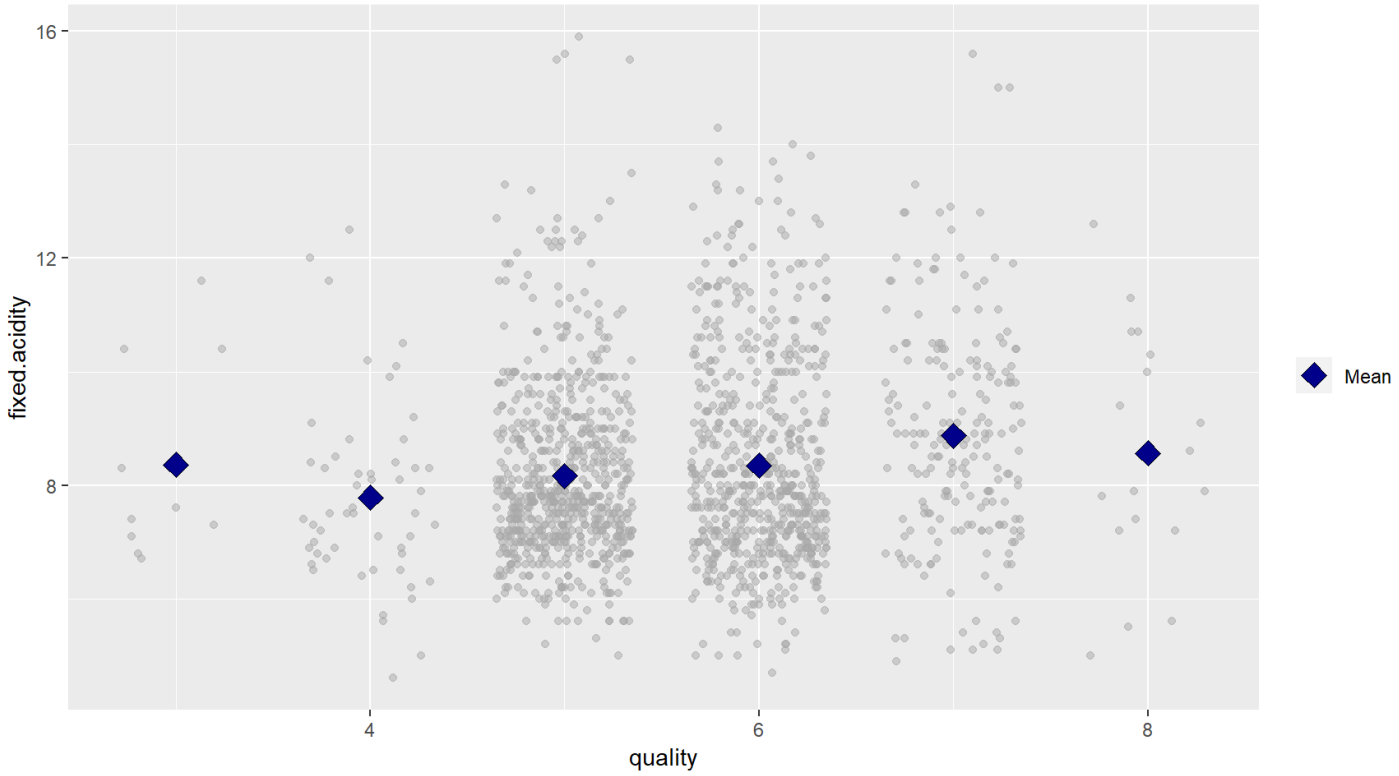

```
## wine$quality: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9947  0.9961  0.9976  0.9975  0.9988  1.0008
## ------------------------------------------------------------
## wine$quality: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9934  0.9957  0.9965  0.9965  0.9974  1.0010
## ------------------------------------------------------------
## wine$quality: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9926  0.9962  0.9970  0.9971  0.9979  1.0031
## ------------------------------------------------------------
## wine$quality: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9901  0.9954  0.9966  0.9966  0.9979  1.0037
## ------------------------------------------------------------
## wine$quality: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9906  0.9948  0.9958  0.9961  0.9974  1.0032
## ------------------------------------------------------------
## wine$quality: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9908  0.9942  0.9949  0.9952  0.9972  0.9988
```

Higher quality wines show to have a lower mean and median density level.



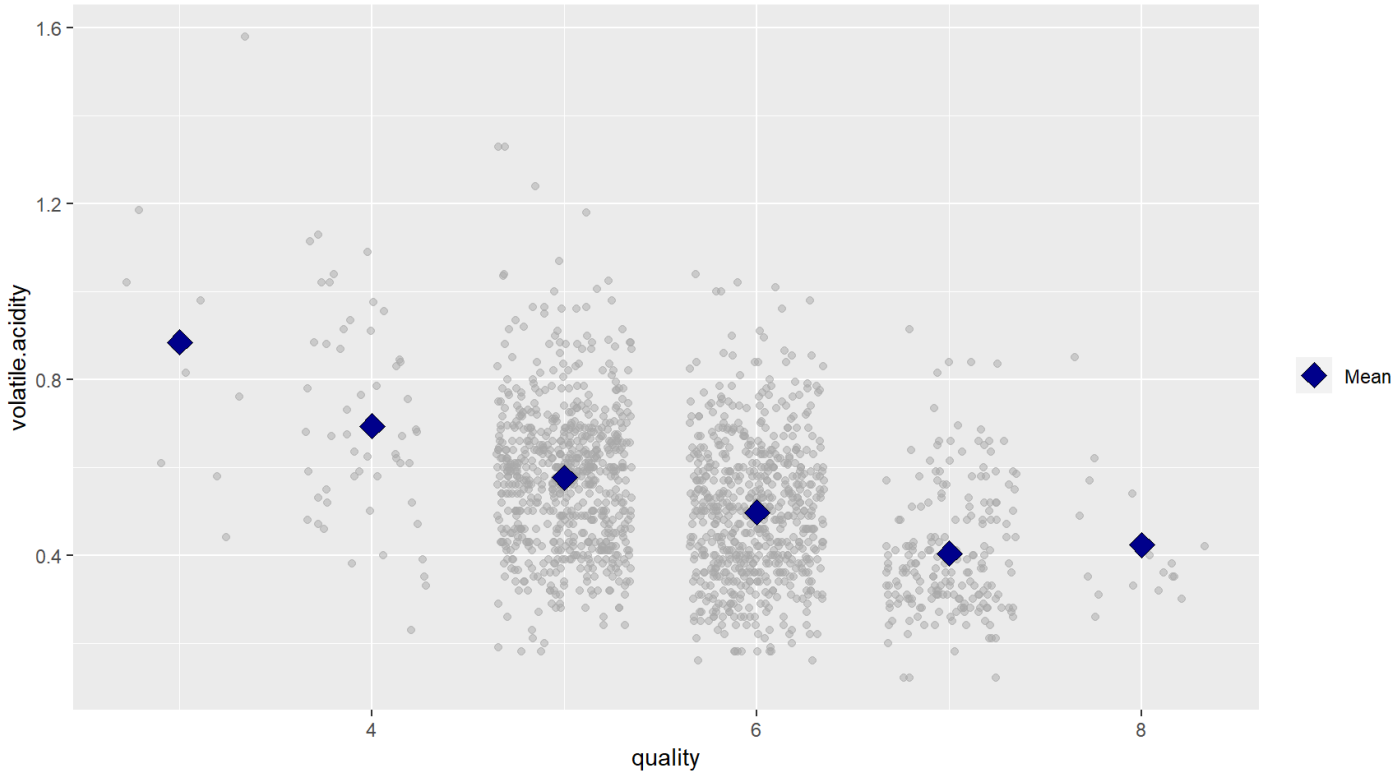

```
## wine$quality: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0050  0.0350  0.1710  0.3275  0.6600
## ------------------------------------------------------------
## wine$quality: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0300  0.0900  0.1742  0.2700  1.0000
## ------------------------------------------------------------
## wine$quality: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0900  0.2300  0.2437  0.3600  0.7900
## ------------------------------------------------------------
## wine$quality: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0900  0.2600  0.2738  0.4300  0.7800
## ------------------------------------------------------------
## wine$quality: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.3050  0.4000  0.3752  0.4900  0.7600
## ------------------------------------------------------------
## wine$quality: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0300  0.3025  0.4200  0.3911  0.5300  0.7200
```

Higher quality wines show to have a higher mean and median citric acid level.



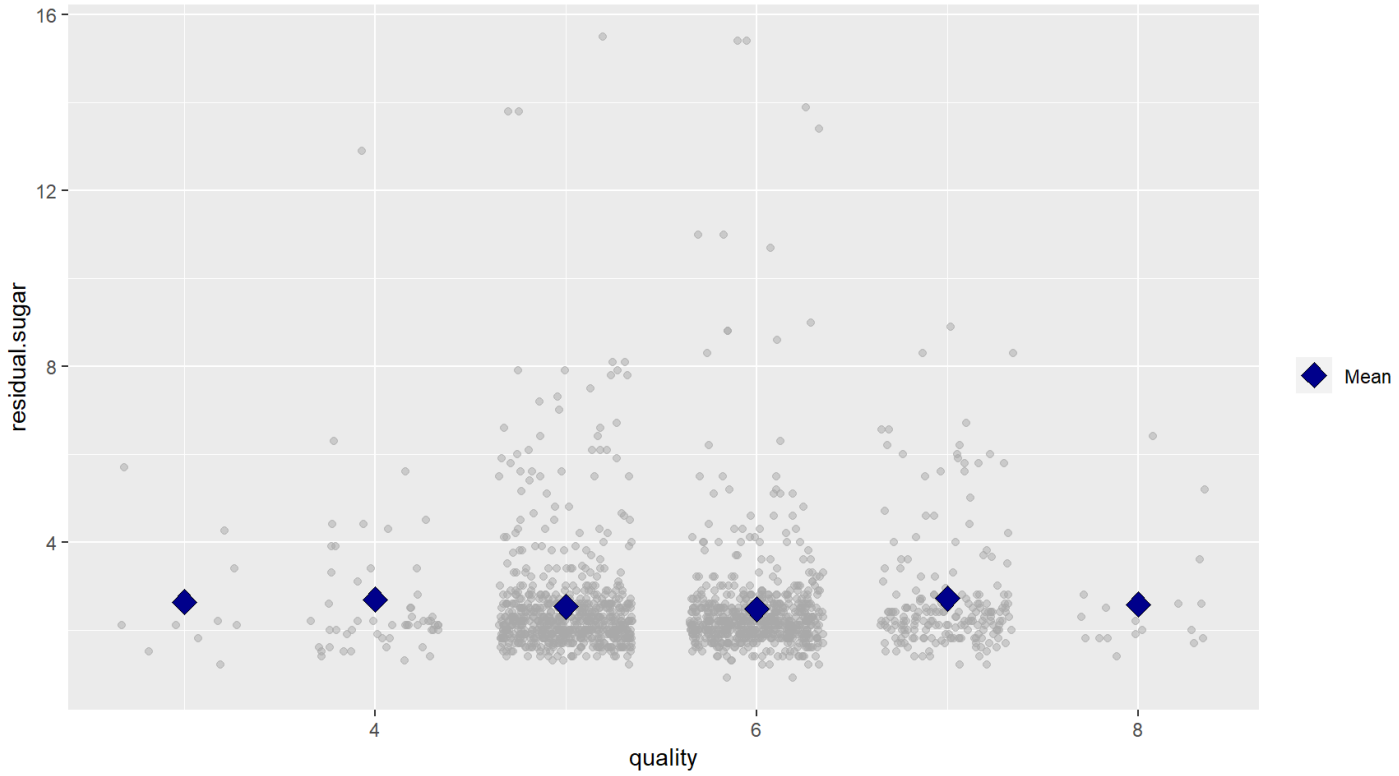

```
## wine$quality: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   6.700   7.150   7.500   8.360   9.875  11.600
## ------------------------------------------------------------
## wine$quality: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.600   6.800   7.500   7.779   8.400  12.500
## ------------------------------------------------------------
## wine$quality: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.000   7.100   7.800   8.167   8.900  15.900
## ------------------------------------------------------------
## wine$quality: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.700   7.000   7.900   8.347   9.400  14.300
## ------------------------------------------------------------
## wine$quality: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.900   7.400   8.800   8.872  10.100  15.600
## ------------------------------------------------------------
## wine$quality: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.000   7.250   8.250   8.567  10.225  12.600
```

Higher quality wines show to have a higher mean and median fixed acidity level.



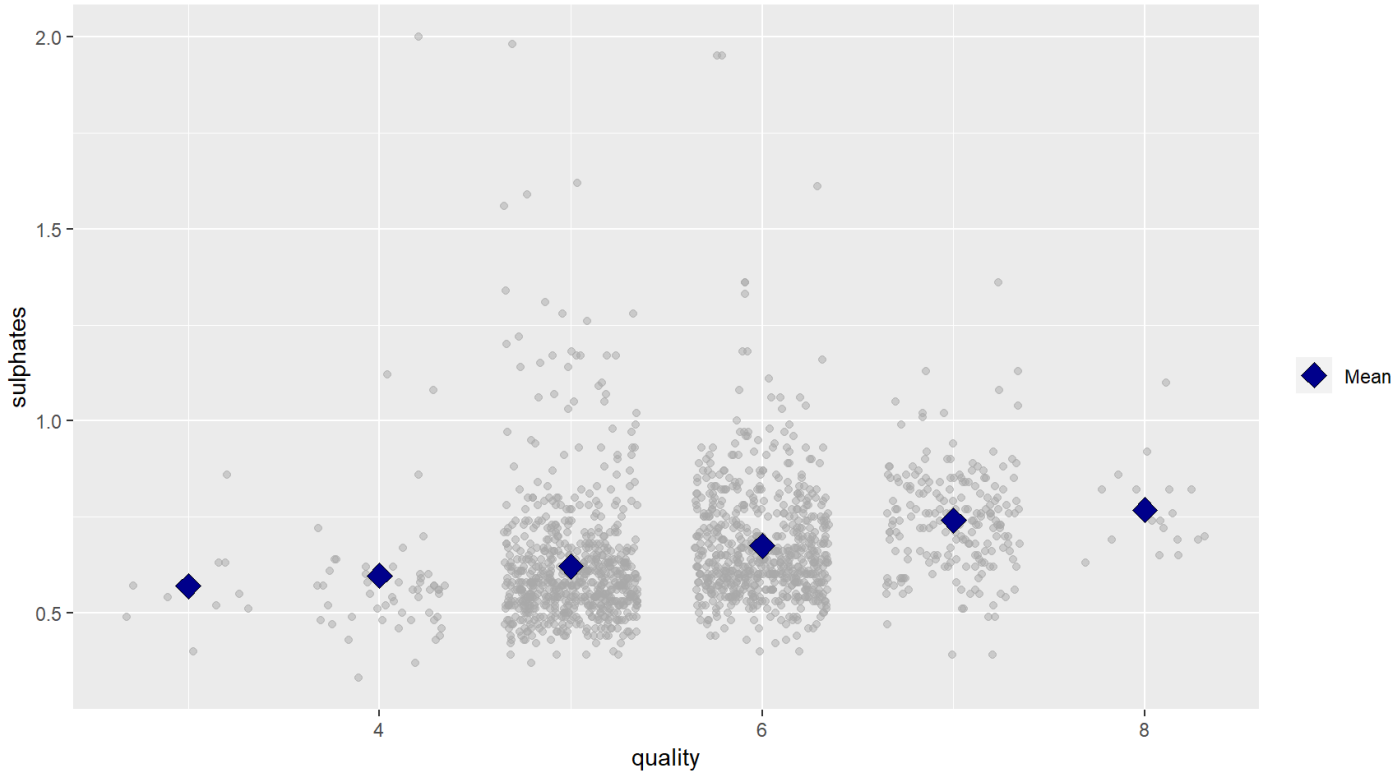

```
## wine$quality: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.4400  0.6475  0.8450  0.8845  1.0100  1.5800
## ------------------------------------------------------------
## wine$quality: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.230   0.530   0.670   0.694   0.870   1.130
## ------------------------------------------------------------
## wine$quality: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.180   0.460   0.580   0.577   0.670   1.330
## ------------------------------------------------------------
## wine$quality: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1600  0.3800  0.4900  0.4975  0.6000  1.0400
## ------------------------------------------------------------
## wine$quality: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1200  0.3000  0.3700  0.4039  0.4850  0.9150
## ------------------------------------------------------------
## wine$quality: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2600  0.3350  0.3700  0.4233  0.4725  0.8500
```

Higher quality wines show to have a lower mean and median volatile acidity level.



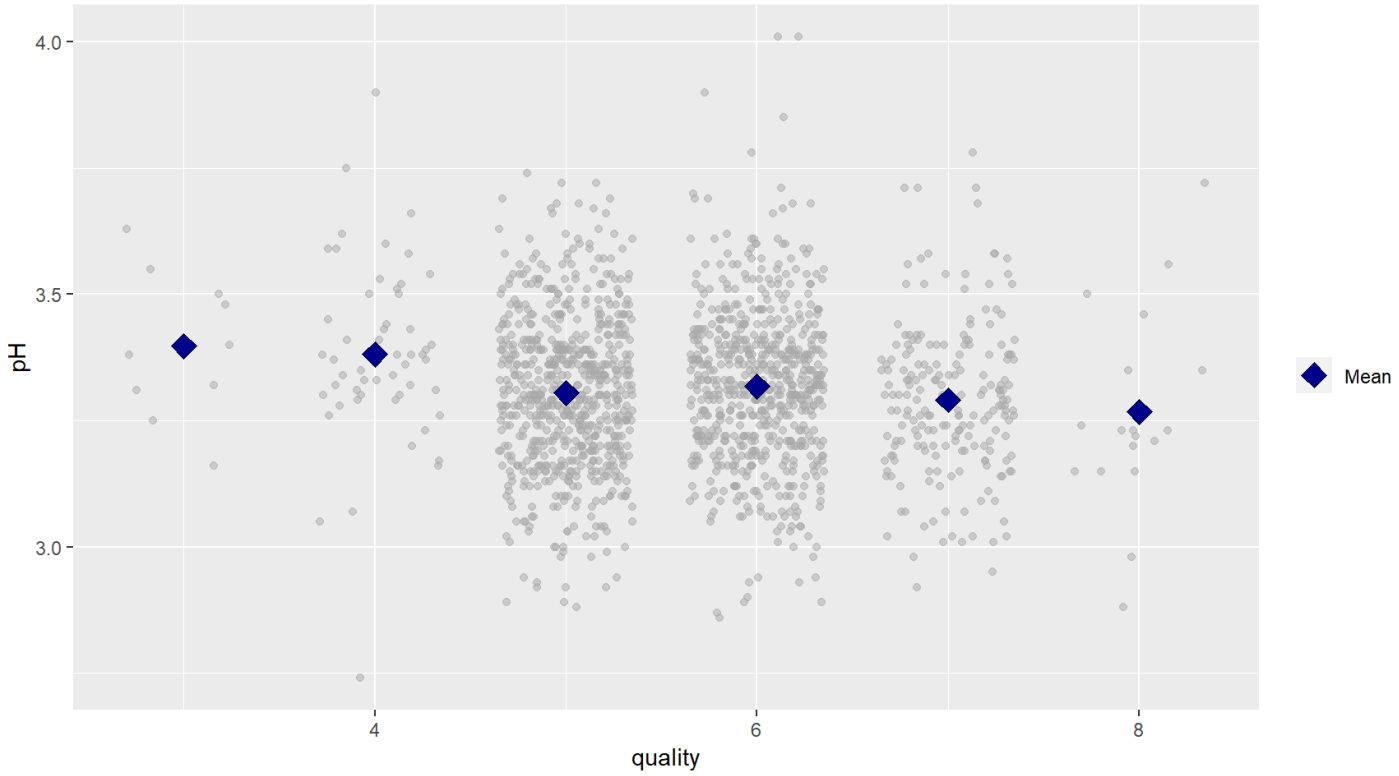

```
## wine$quality: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.200   1.875   2.100   2.635   3.100   5.700
## ------------------------------------------------------------
## wine$quality: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.300   1.900   2.100   2.694   2.800  12.900
## ------------------------------------------------------------
## wine$quality: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.200   1.900   2.200   2.529   2.600  15.500
## ------------------------------------------------------------
## wine$quality: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.900   1.900   2.200   2.477   2.500  15.400
## ------------------------------------------------------------
## wine$quality: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.200   2.000   2.300   2.721   2.750   8.900
## ------------------------------------------------------------
## wine$quality: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.400   1.800   2.100   2.578   2.600   6.400
```

This one is a bit misleading because both low quality wines and the highest quality wines appear to have low residual sugar. While the higher residual sugar wines tend to score between 5-7 on the quality scale.



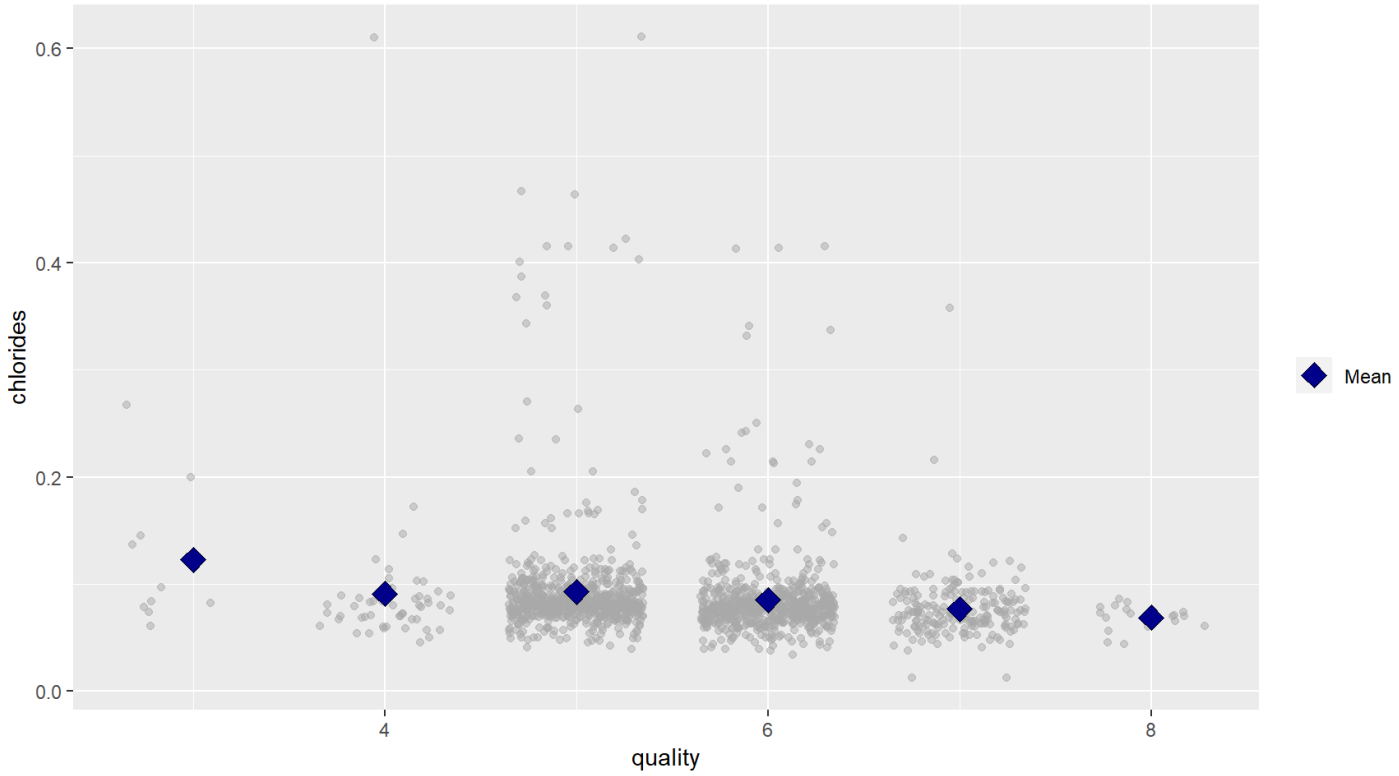

```
## wine$quality: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.4000  0.5125  0.5450  0.5700  0.6150  0.8600
## ------------------------------------------------------------
## wine$quality: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3300  0.4900  0.5600  0.5964  0.6000  2.0000
## ------------------------------------------------------------
## wine$quality: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.370   0.530   0.580   0.621   0.660   1.980
## ------------------------------------------------------------
## wine$quality: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.4000  0.5800  0.6400  0.6753  0.7500  1.9500
## ------------------------------------------------------------
## wine$quality: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3900  0.6500  0.7400  0.7413  0.8300  1.3600
## ------------------------------------------------------------
## wine$quality: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.6300  0.6900  0.7400  0.7678  0.8200  1.1000
```

Higher quality wines have higher mean and median levels of sulphates.




```
## wine$quality: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.160   3.312   3.390   3.398   3.495   3.630
## ------------------------------------------------------------
## wine$quality: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.740   3.300   3.370   3.382   3.500   3.900
## ------------------------------------------------------------
## wine$quality: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.880   3.200   3.300   3.305   3.400   3.740
## ------------------------------------------------------------
## wine$quality: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.860   3.220   3.320   3.318   3.410   4.010
## ------------------------------------------------------------
## wine$quality: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.920   3.200   3.280   3.291   3.380   3.780
## ------------------------------------------------------------
## wine$quality: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.880   3.163   3.230   3.267   3.350   3.720
```

Higher quality wines show to have lower mean and median pH level.




```
## wine$quality: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0610  0.0790  0.0905  0.1225  0.1430  0.2670
## ------------------------------------------------------------
## wine$quality: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.04500 0.06700 0.08000 0.09068 0.08900 0.61000
## ------------------------------------------------------------
## wine$quality: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.03900 0.07400 0.08100 0.09274 0.09400 0.61100
## ------------------------------------------------------------
## wine$quality: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.03400 0.06825 0.07800 0.08496 0.08800 0.41500
## ------------------------------------------------------------
## wine$quality: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01200 0.06200 0.07300 0.07659 0.08700 0.35800
## ------------------------------------------------------------
## wine$quality: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.04400 0.06200 0.07050 0.06844 0.07550 0.08600
```

Higher quality wines show to have lower mean and median chlorides level.

Bivariate Analysis

Talk about some of the relationships you observed in this part of the investigation. How did the feature(s) of interest vary with other features in the dataset?

Here are some of the relationships I observed from the above bivariate plots:

- Red wines with high quality have higher alcohol percentages.
- Red wines with high quality have lower volatile acidity.
- Red wines with high quality have higher citric acid level.
- Red wines with high quality have higher fixed acidity level.
- Red wines with high quality have lower density level.
- Red wines with high quality have higher levels of sulphates.
- Red wines with high quality have lower level of pH.
- Red wines with high quality have lower level of chlorides.

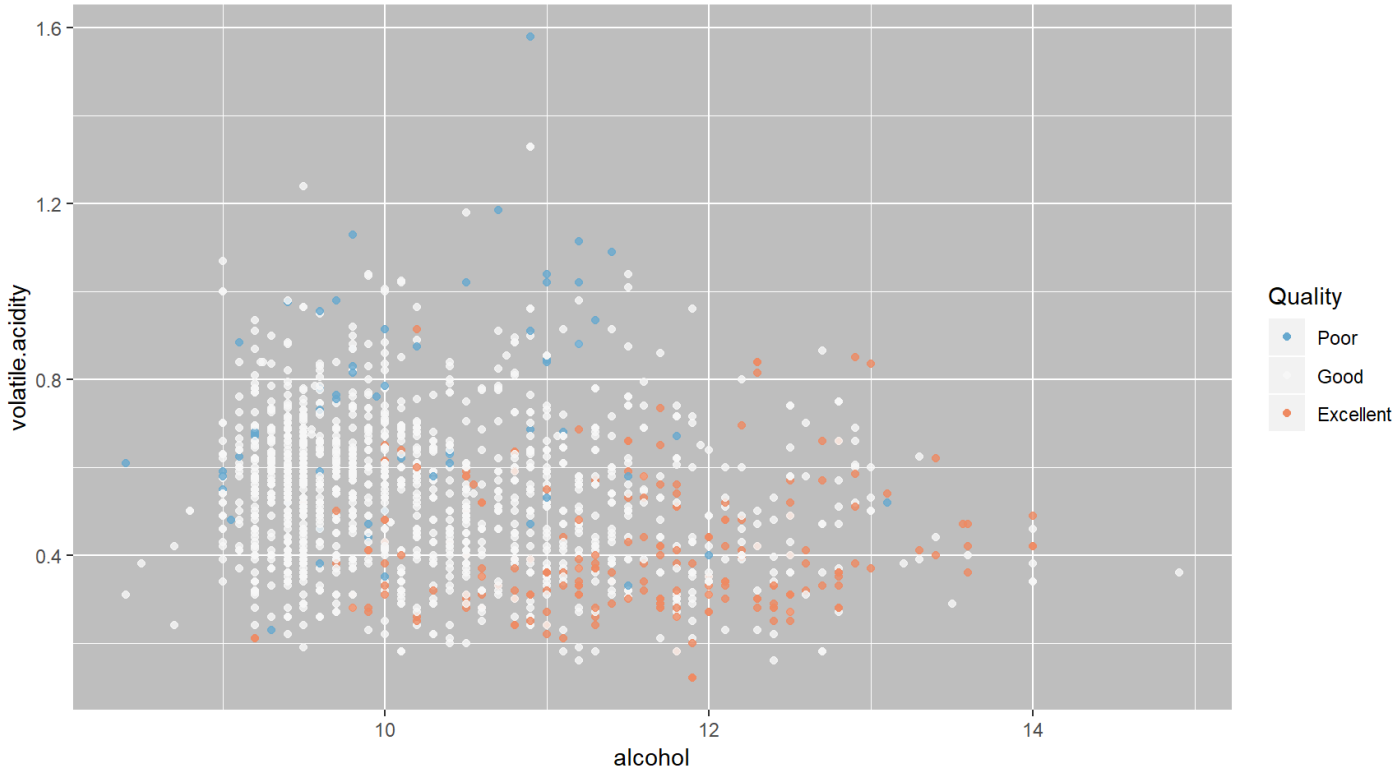
Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

Yes, in the ggpairs plot we can see a slight correlation between fixed acidity vs density, fixed acidity vs pH, and fixed acidity vs citric acid.

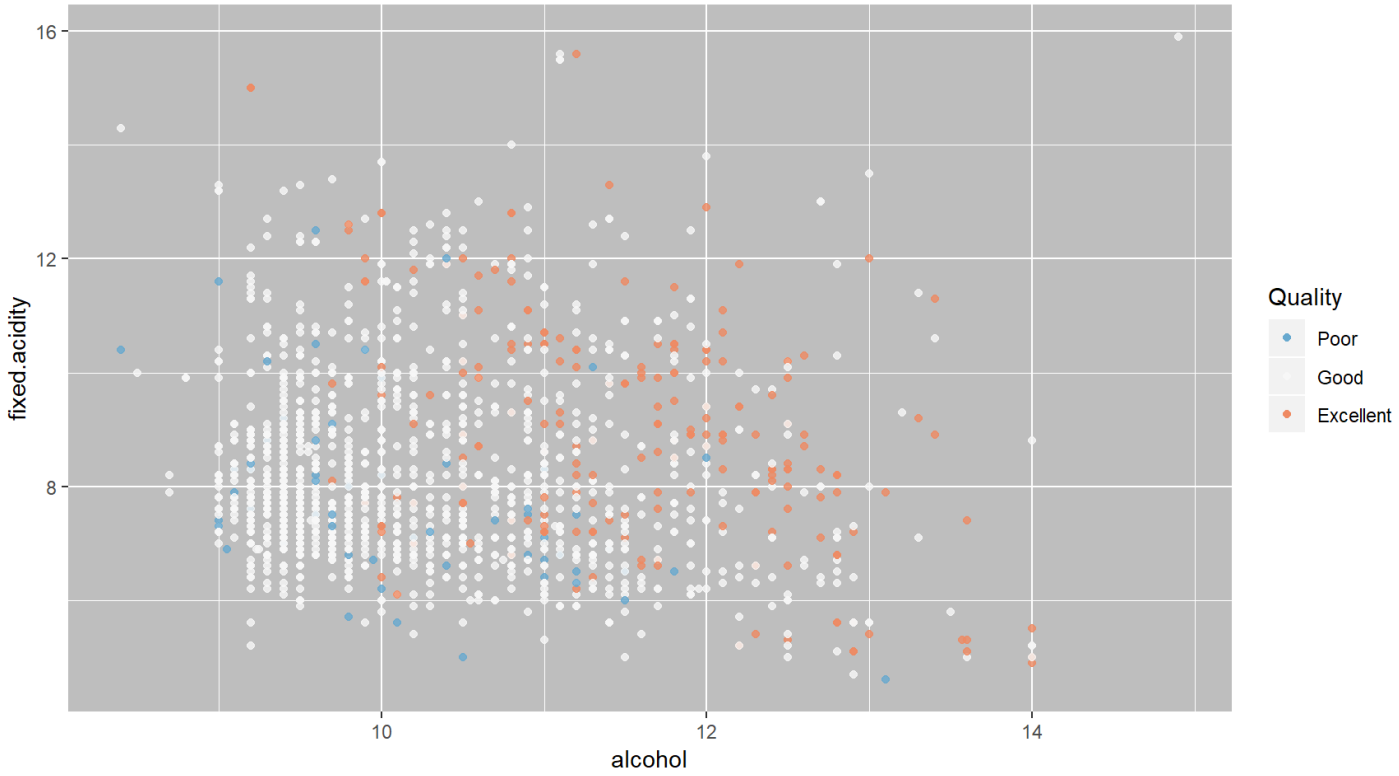
What was the strongest relationship you found?

The strongest relationship found is between fixed acidity and pH. Since I am primarily concerned with the quality feature, the strongest relationship found to quality is alcohol and volatile acidity. All the other variables had a very weak correlation coefficient when compared to quality.

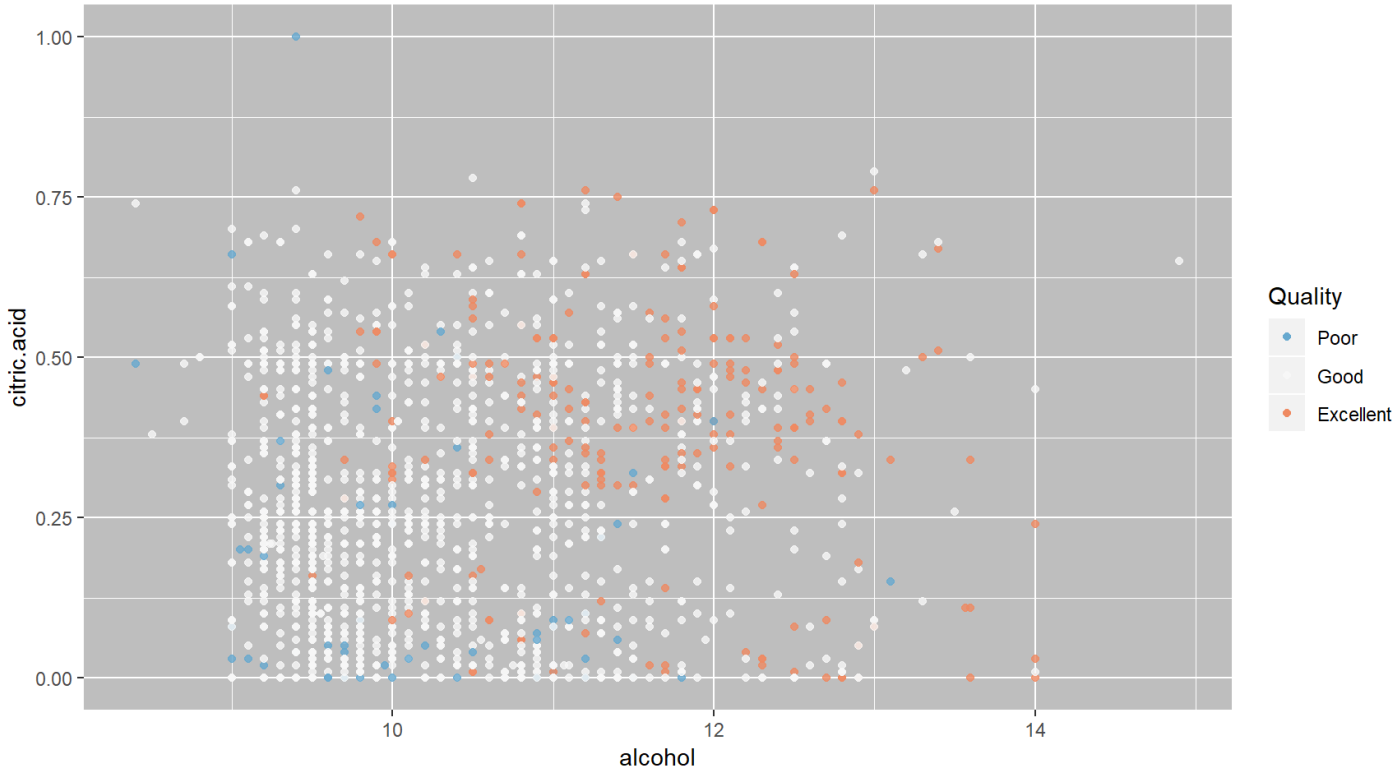
Multivariate Plots Section



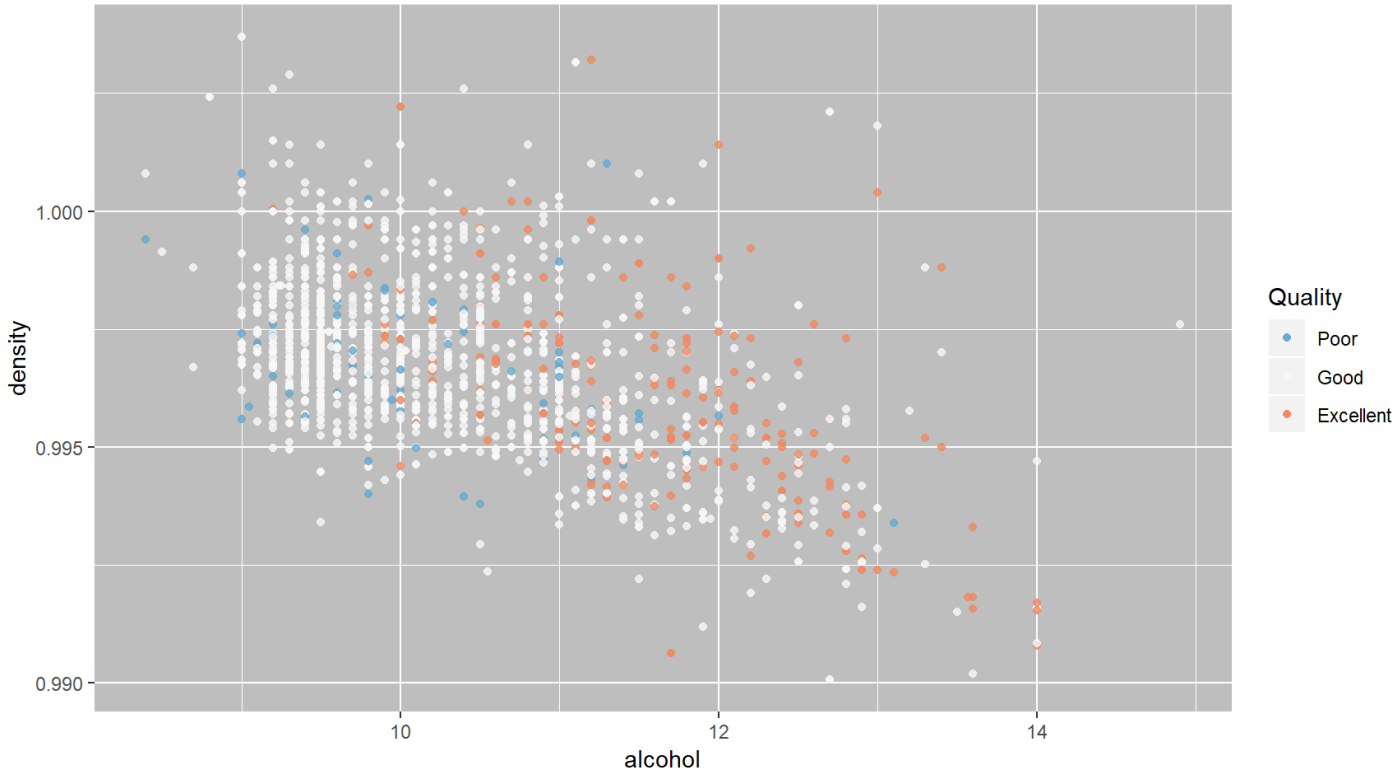
In the above scatterplot we can observe most of the “excellent” wines in the lower right portion of the plot. This further concludes our findings from the bivariate analysis that higher quality red wines (ie. “Excellent”) are associated with having higher alcohol % and lower volatile acidity.



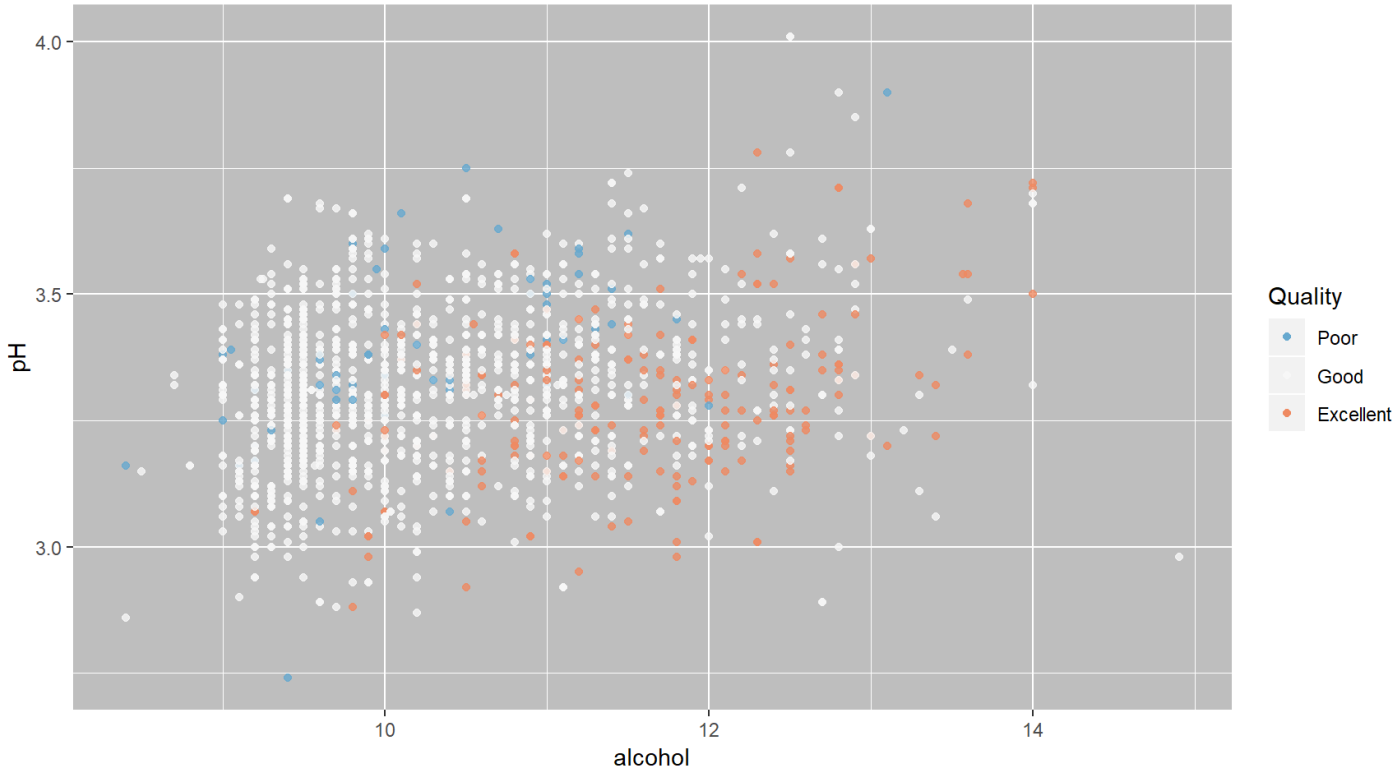
This plot doesn’t tell us much between the two variables and the rating category. We already know that higher rated wines have higher alcohol % but here we can see the “excellent” category having both low fixed acidity and high fixed acidity.



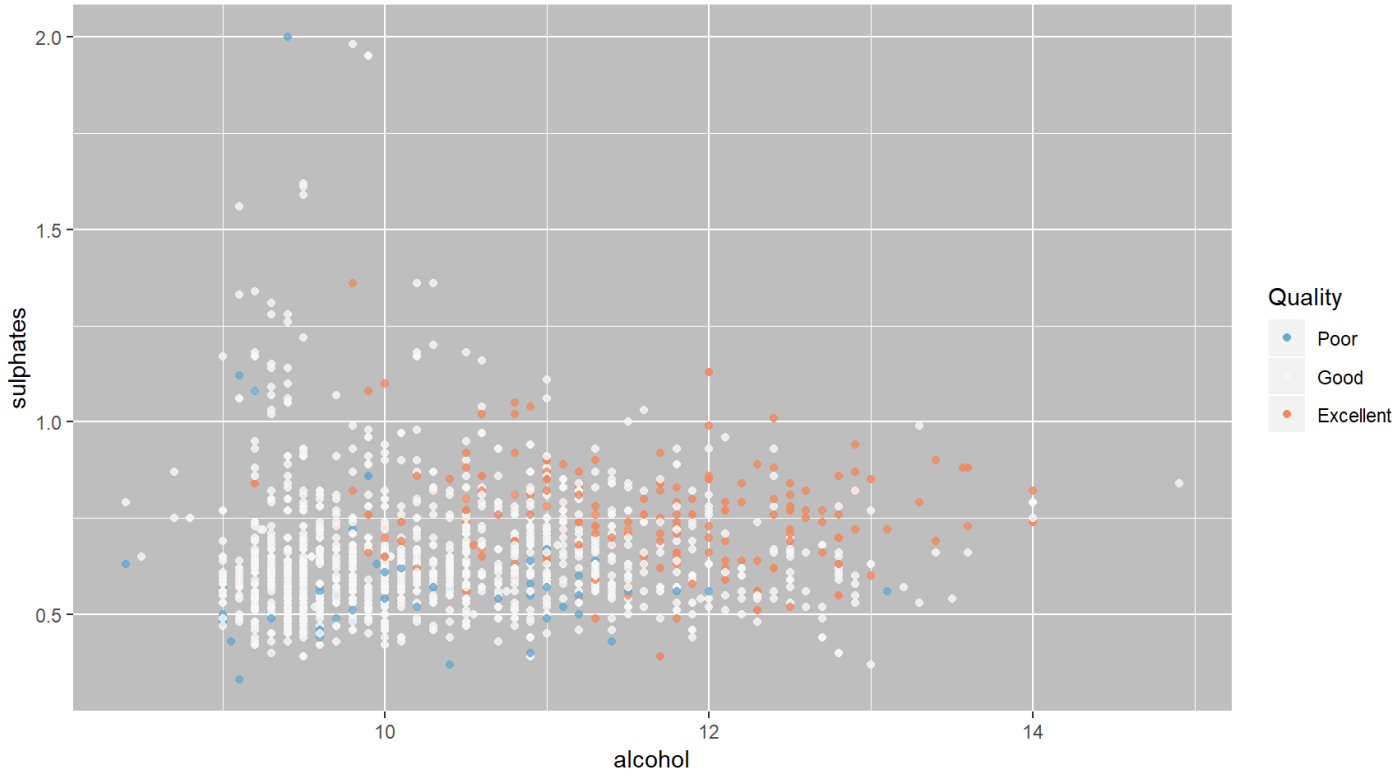
This scatter plot shows that “excellent” red wines have high alcohol % and generally higher citric acid.



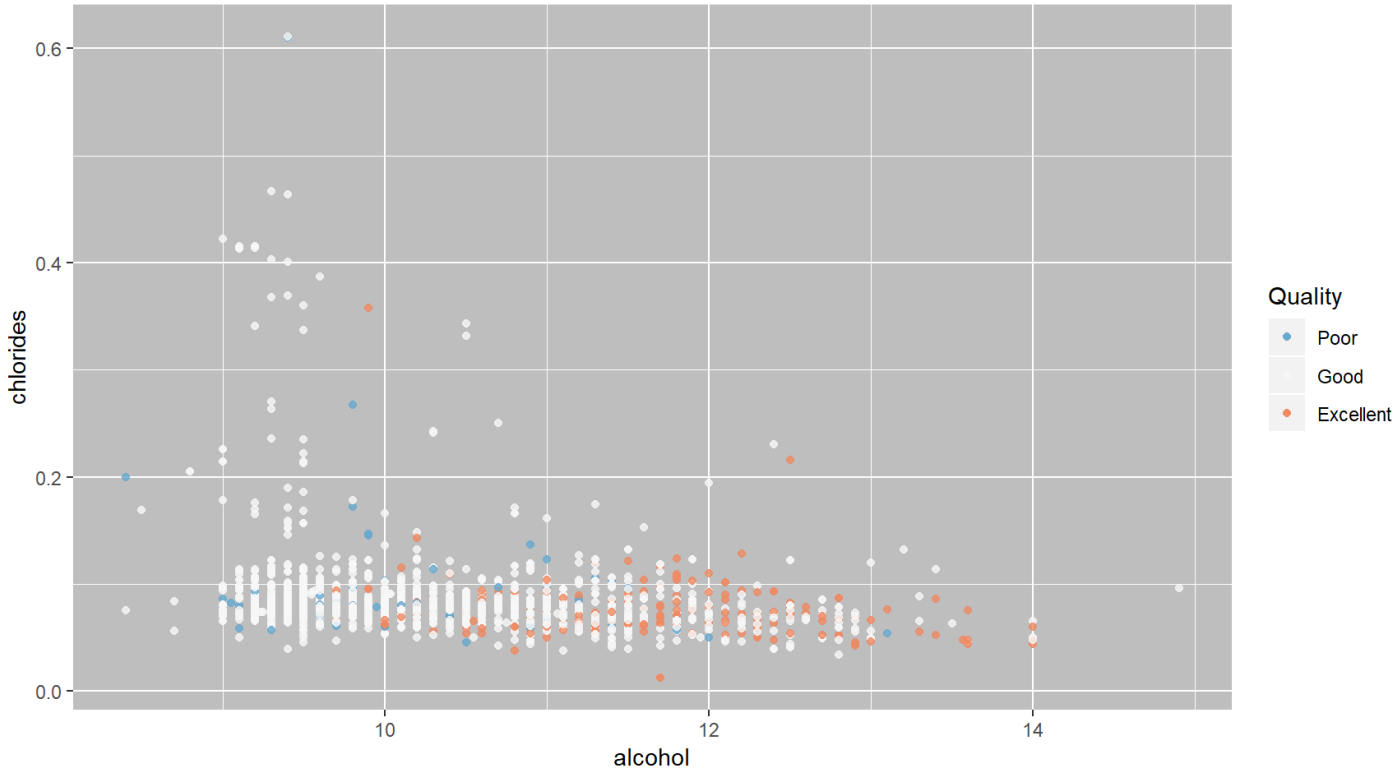
We can see most of the blue dots in the lower right section of the scatterplot which means that “excellent” wines tend to have lower density and higher alcohol %.



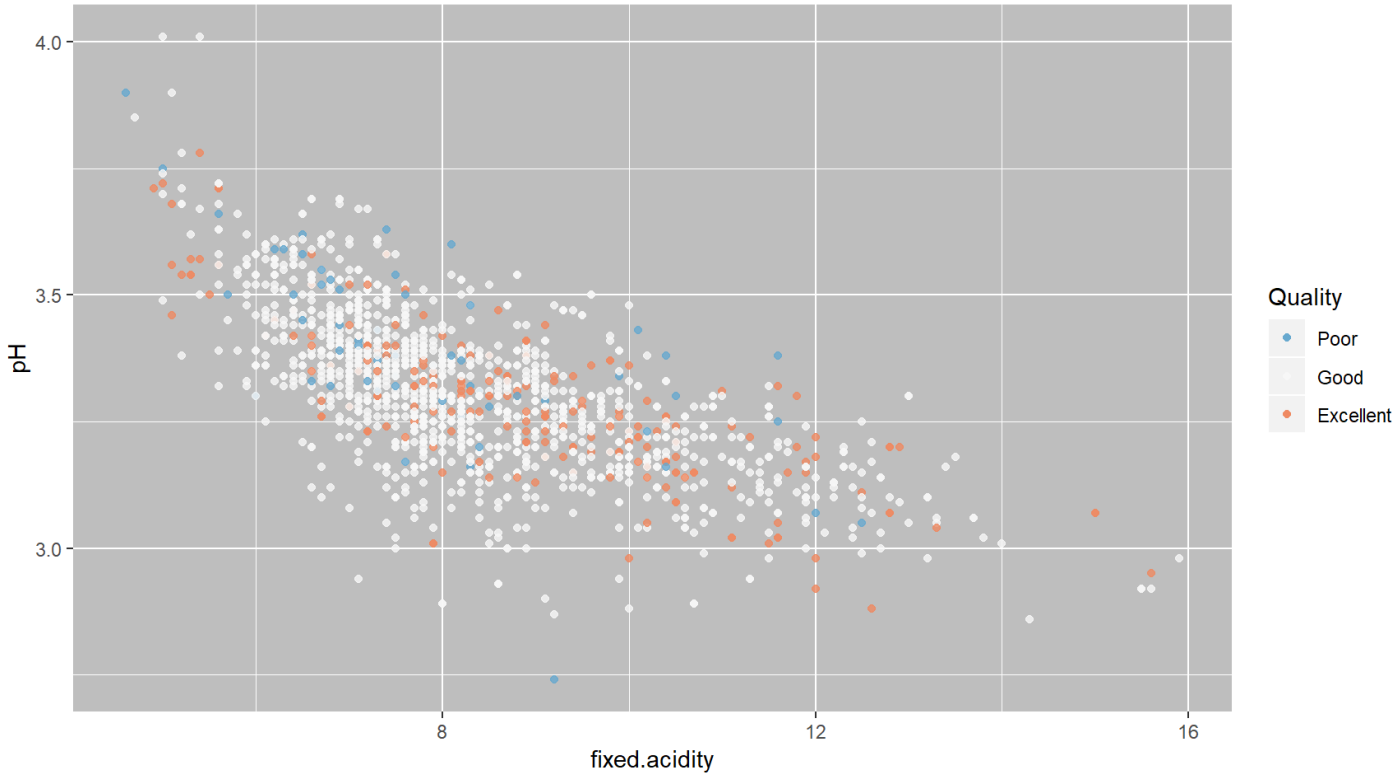
Here we can see that “excellent” wines tend to have low pH and high alcohol %.



Here we can see that “excellent” wines tend to have higher sulphates and high alcohol %.



Here we can see that “excellent” wines tend to have lower chlorides and high alcohol %.



Because we saw a strong relationship between fixed acidity and pH, I would like to explore this relationship further by adding in the rating variable. In the scatterplot above we can see the negative correlation between fixed acidity and pH that we observed earlier. However it does not seems that both variables combined together does influence the rating of the wine.

Multivariate Analysis

Talk about some of the relationships you observed in this part of the investigation. Were there features that strengthened each other in terms of looking at your feature(s) of interest?

We observed two variables that influence the quality of wine which are alcohol % and volatile acidity. The other variables did not have a strong relationship to the quality of the wine.

Were there any interesting or surprising interactions between features?

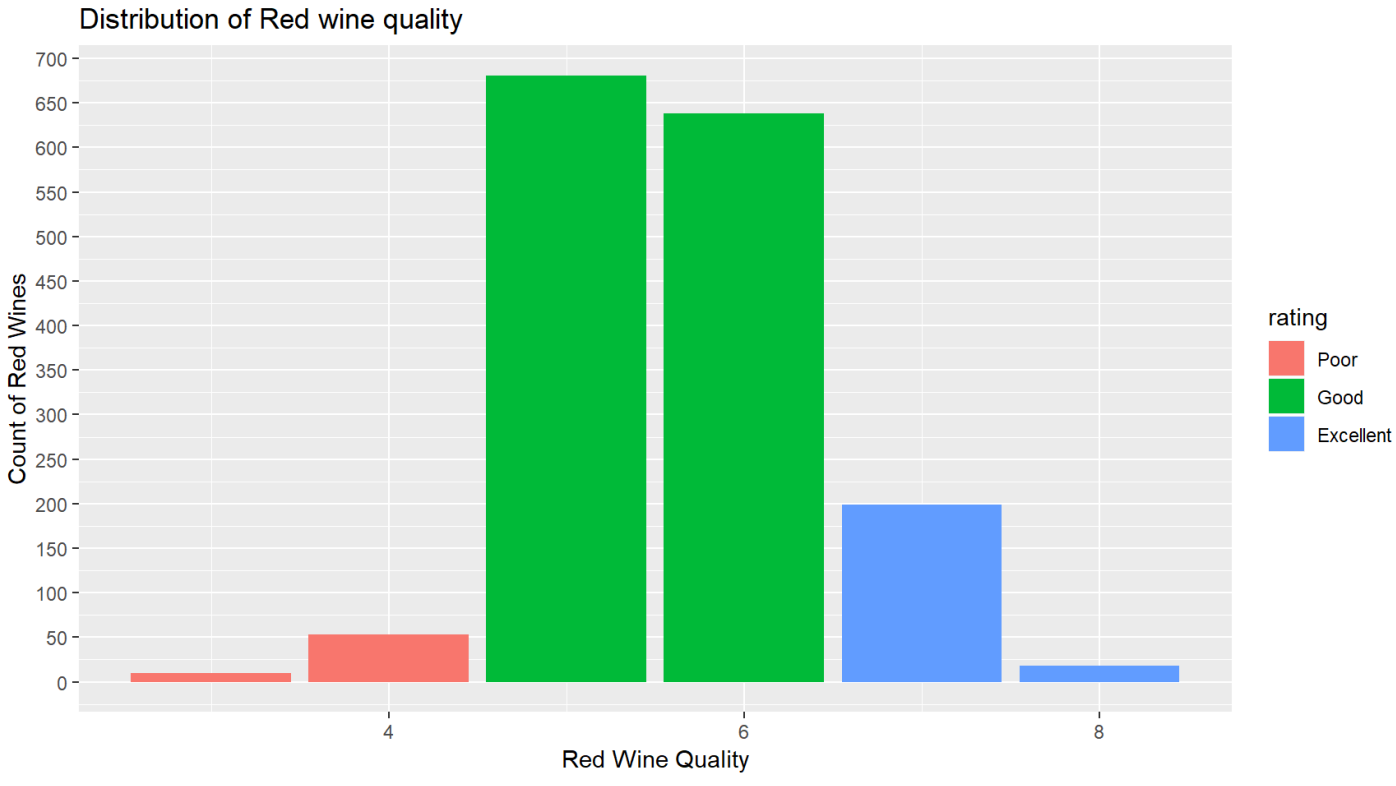
I was surprised that residual sugar did not have any meaningful relationship to the quality of wine since it influences how sweet the wine is.

OPTIONAL: Did you create any models with your dataset? Discuss the strengths and limitations of your model.

No I did not create any models with the dataset.

Final Plots and Summary

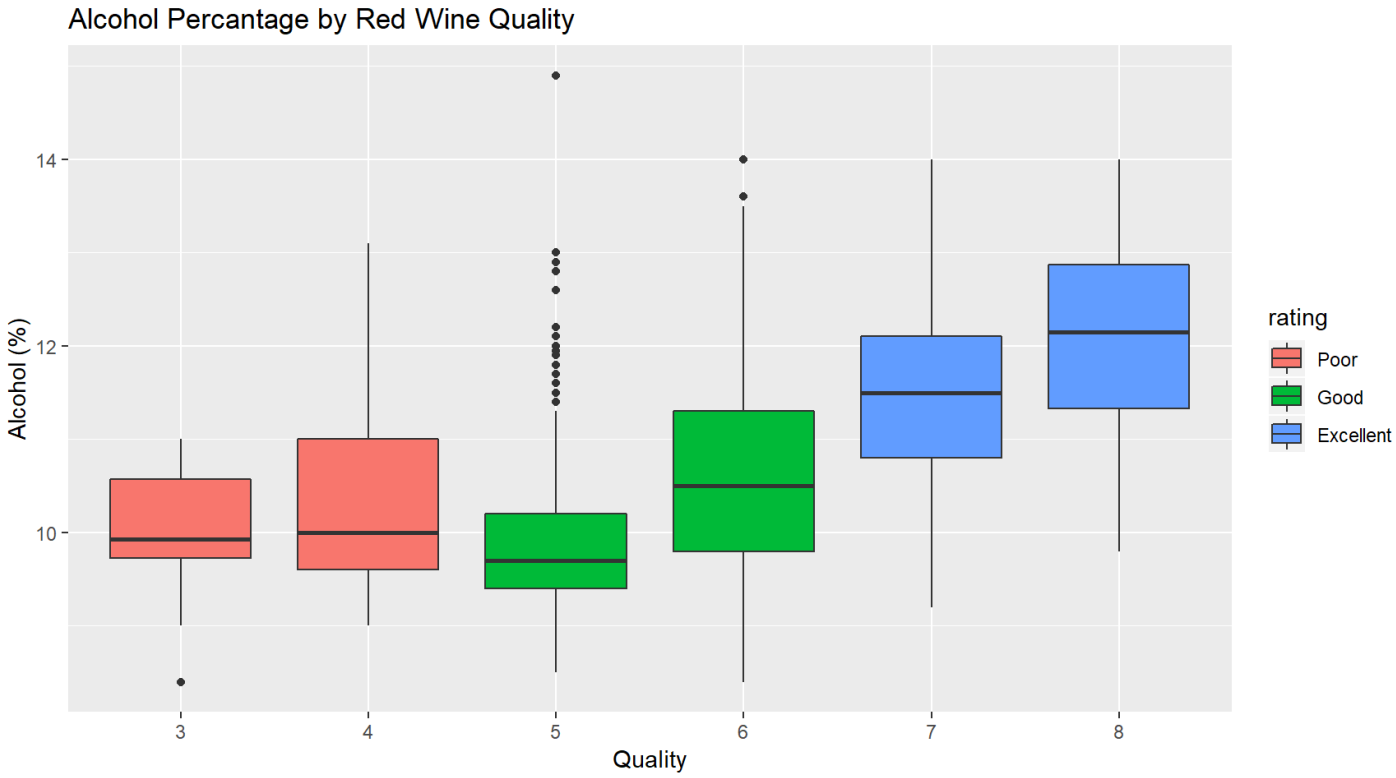
Plot One



Description One

From the histogram plot above, we can see that the majority of our data falls in the 5-6 quality scale. There are a lot more “good” wine (those with quality of 5 - 6) than the “excellent” ones (quality of 7- 8) or the “poor” ones (3 - 4). This might be because this category is the baseline for most people and that the wine has to be worse than expected to give it a low rating and better than expected for a high rating.

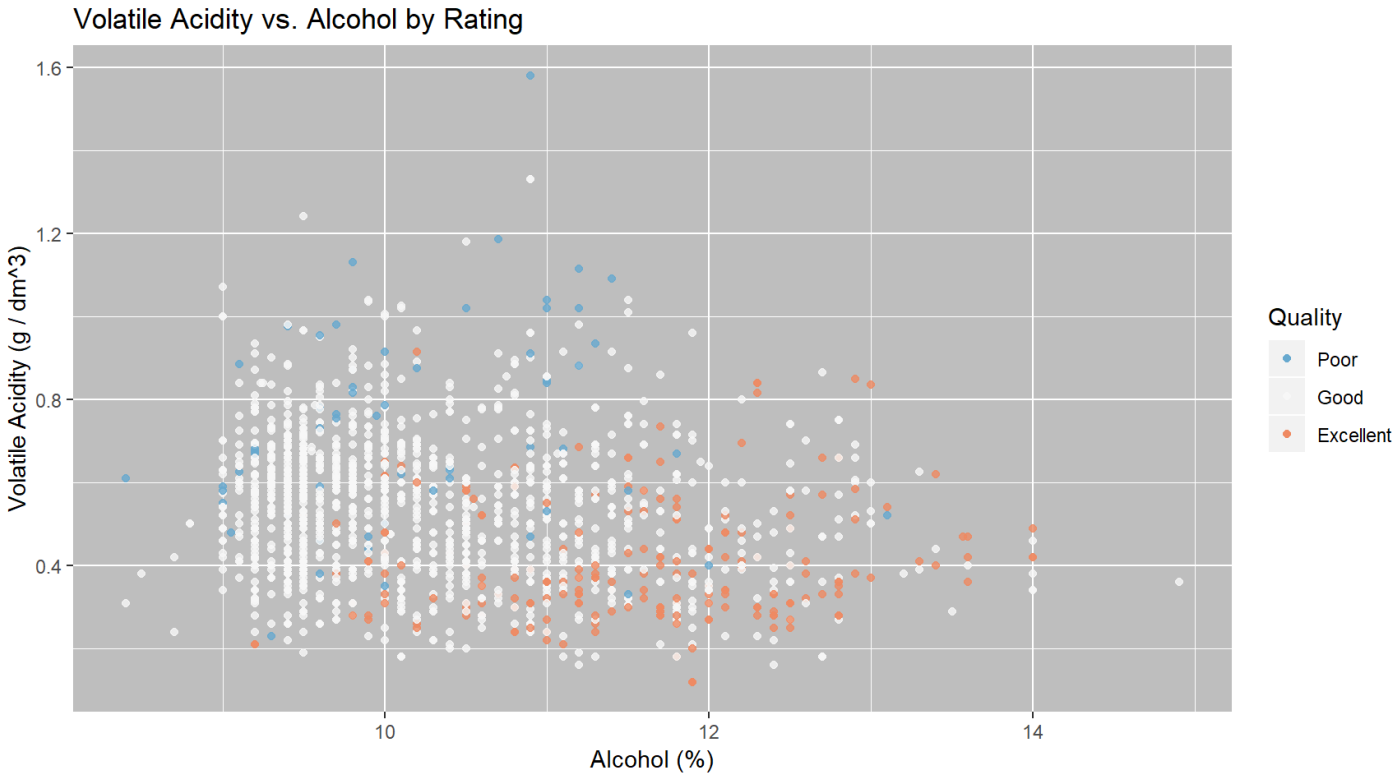
Plot Two



Description Two

The boxplots above show that higher quality red wines (“Excellent”) have largest median alcohol % compared to the other qualities and ratings. In other words as alcohol % increases the quality and rating of the wine tend to increase as well.

Plot Three



Description Three

In the above scatterplot we are looking at volatile acidity and alcohol % and how it relates to the rating. We can observe that the “Excellent” red wines do in fact tend to have higher alcohol % which confirms our finding from the boxplot in Plot #2. We can also see that the “Excellent” wines also have have lower volatile acidity. So we can infer from this scatterplot that higher alcohol % and lower volatile acidity leads to higher quality and subsequently higer rating of red wine.

Reflection

The red wines data set contains information on almost 1,600 red wines across t12 variables from 2009. I started by understanding the individual variables in the data set, and then I explored interesting questions and leads as I continued to make observations on plots. Eventually, I explored the wine quality and rating across many variables and tried to understand which variables created a higher quality/rating. The two variables I was able to see the strongest relationship to higher quality wines were alcohol % and volatile acidity.

What went well was that the data was tidy with no null values. I was surprised that residual sugar did not have a any relationship with quality. Prior to looking at the data I assumed that sweeter wines may have higher quality, but this was not the case.

Some of the struggles with the data set was that the quality was given in integers. To help analyze the data, I created another variable “rating” to transform this data into categorical data. Another struggle I had was that I am not familiar with chemistry so I am not truly able to fully understand some of the nuances in between the variables and their relationships to each other.

In the future it would be interesting to include price to this data set as I believe price may have a strong relationship to the quality of the wine.