The Red Wine Journey by Samuel Rodriguez

Selecting the best red wine is a tricky one. There are many variables that can make the quality of a red wine the best or the worst.

This analysis is designed to explore the factors that contribute to a bad or good red wine. The dataset includes only physicochemical variables to predict the sensory output. The sensory output is the quality of a red wine measured subjectively by at least 3 wine experts with a rating from 0 (very bad) to 10 (very excellent).

The purpose is to create a model that can differentiate a good red wine from a bad red wine using attribute information within the data provided. Let's start exploring!

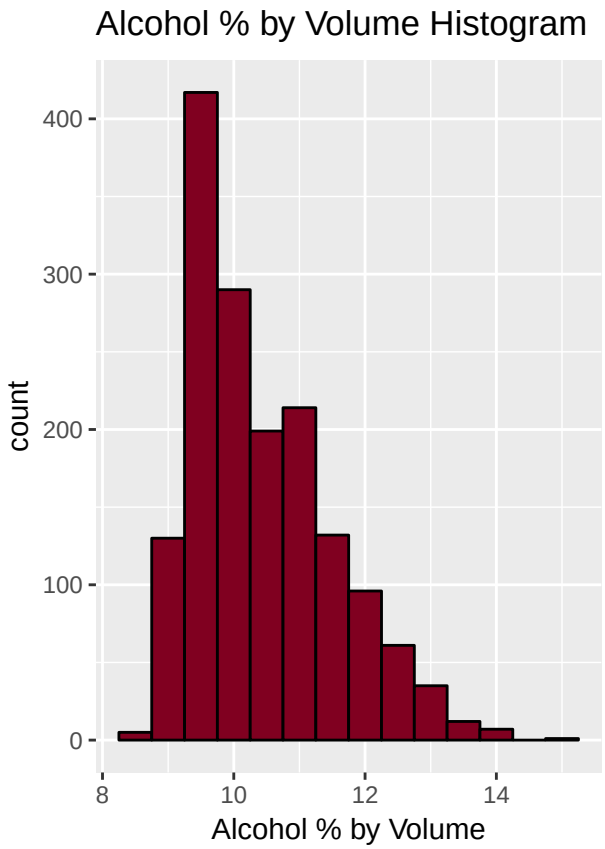
Univariate Plots Section

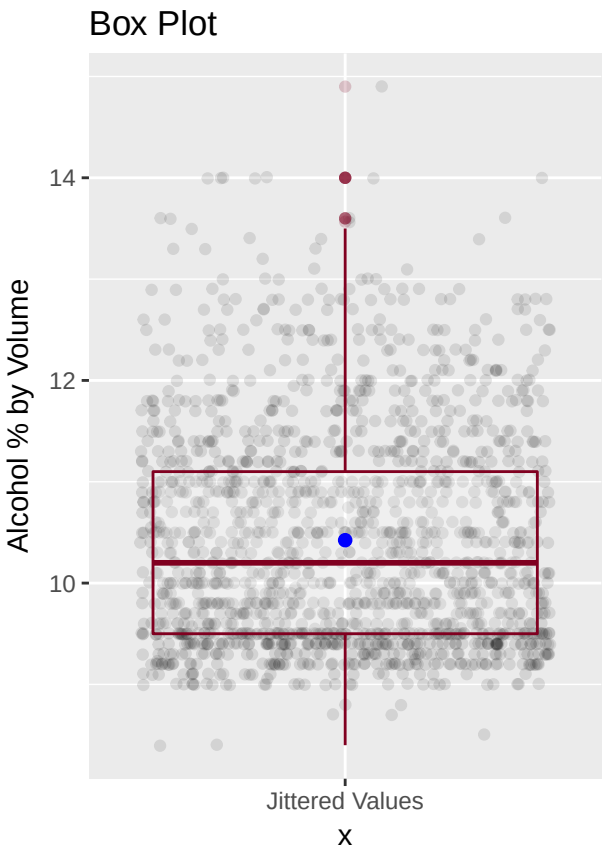
Variables Available

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

Alcohol

Histogram & Box Plot





Data Summary

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   8.400   9.500  10.200  10.423  11.100  14.900
```

```
## Values greater than or equal to 13.62 are outliers.
```

```
## There are 8 outliers in this variable.
```

```
## The statistics after outlier removal are shown below:
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   8.400   9.500  10.100  10.404  11.100  13.600
```

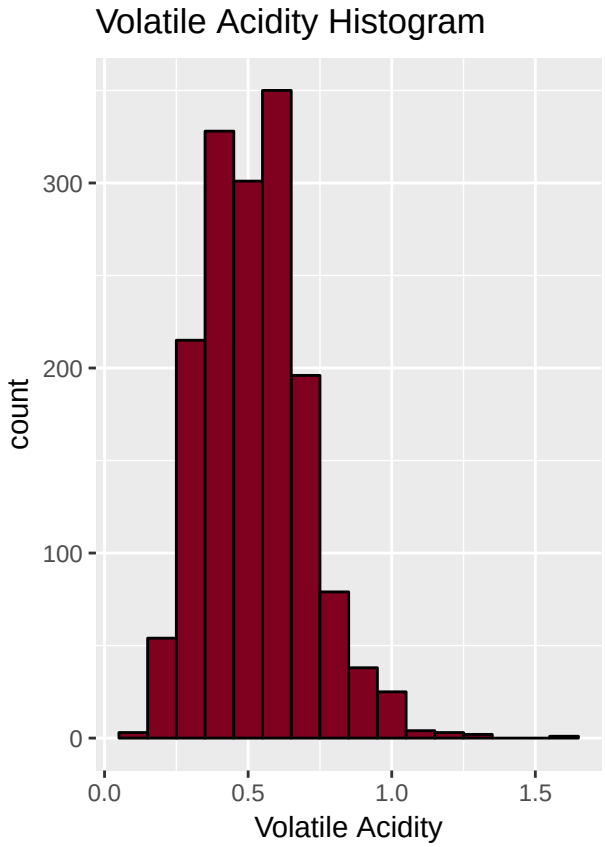
Description

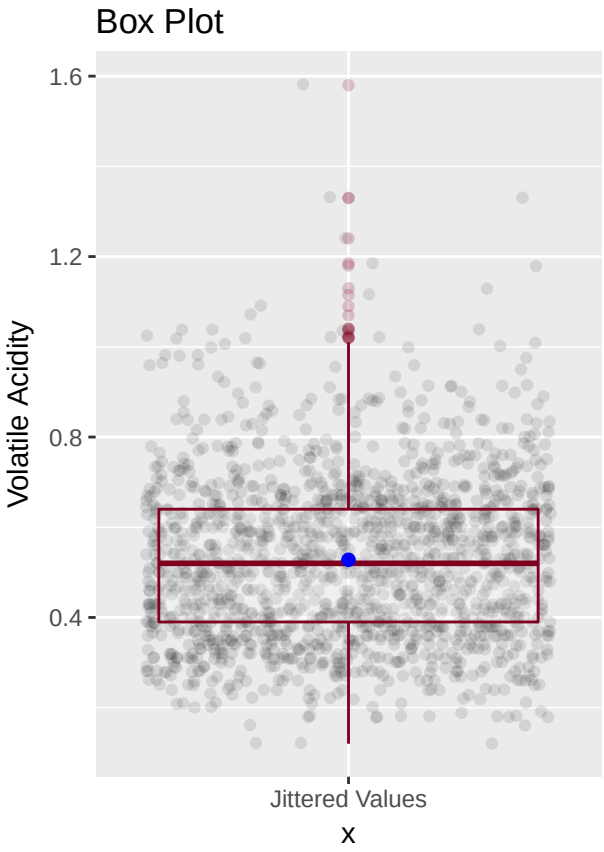
In the above grid, we see the distribution of alcohol % by volume of red wines sampled. In the box plot, we can appreciate the distribution more closely.

We can observe a more dense area around 9.5 and some outliers with higher alcohol % by volume. There is a long stretch on the top 25 percentile, which indicates that there is more alcohol % by volume variability between 11.1 and 14.9. We also noticed that there are 8 red wines with alcohol levels higher or equal to than 13.62. These values are very unlikely to happen.

Volatile Acidity

Histogram & Box Plot





Data Summary

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.12000 0.39000 0.52000 0.52782 0.64000 1.58000
```

```
## Values greater than or equal to 1.06 are outliers.
```

```
## There are 10 outliers in this variable.
```

```
## The statistics after outlier removal are shown below:
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.12000 0.39000 0.52000 0.52343 0.63500 1.04000
```

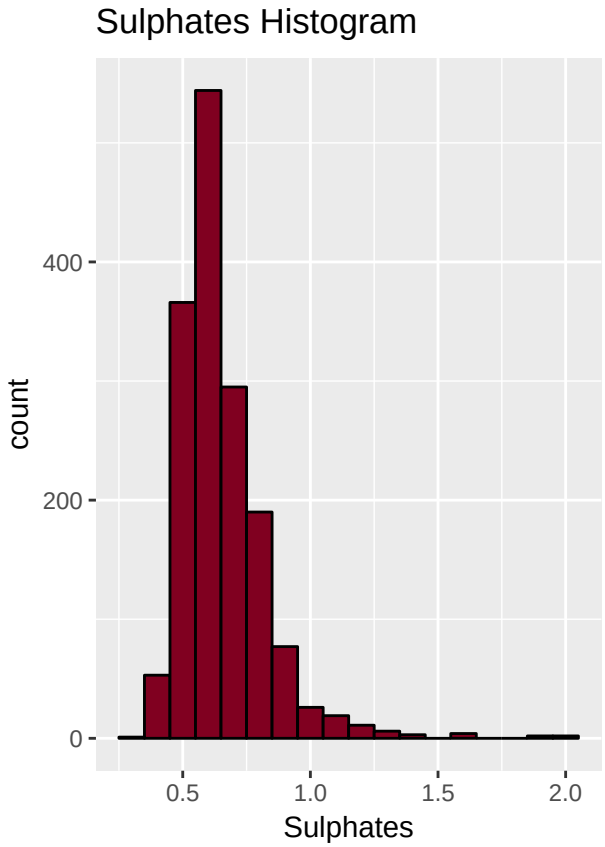
Description

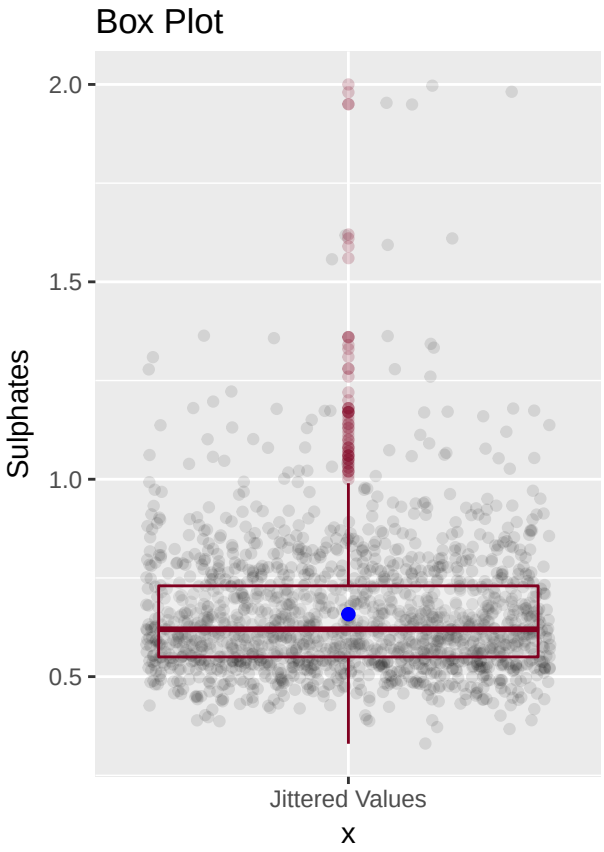
In the above grid, we see the distribution of volatile acidity of red wines sampled.

We can observe a more spread density around the inner quartile range and less of a stretch on the top 25 percentile than what we observed with alcohol % by volume. Here we also found that there are many outliers with volatile acidity greater than or equal to 1.06.

Sulphates

Histogram & Box Plot





Data Summary

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.33000 0.55000 0.62000 0.65815 0.73000 2.00000
```

```
## Values greater than or equal to 1.17 are outliers.
```

```
## There are 27 outliers in this variable.
```

```
## The statistics after outlier removal are shown below:
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.33000 0.55000 0.62000 0.64531 0.72000 1.16000
```

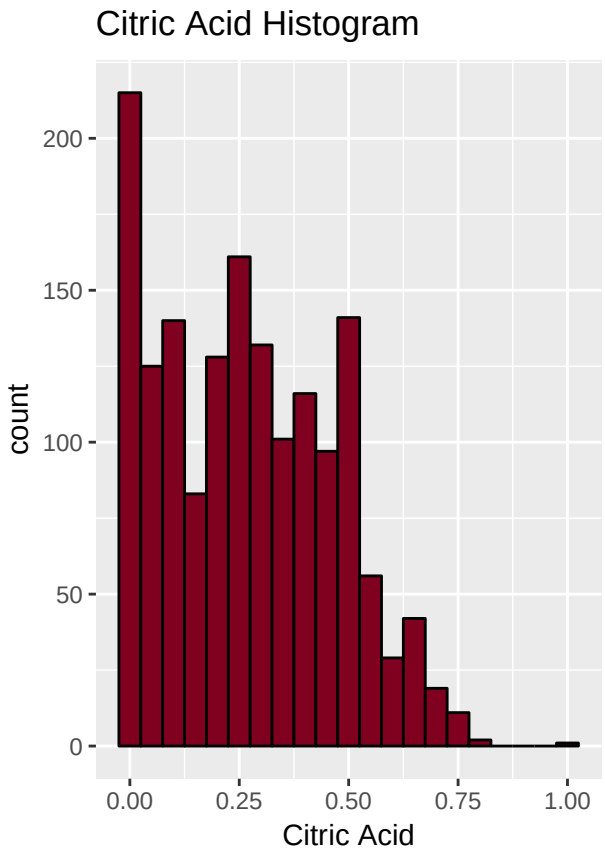
Description

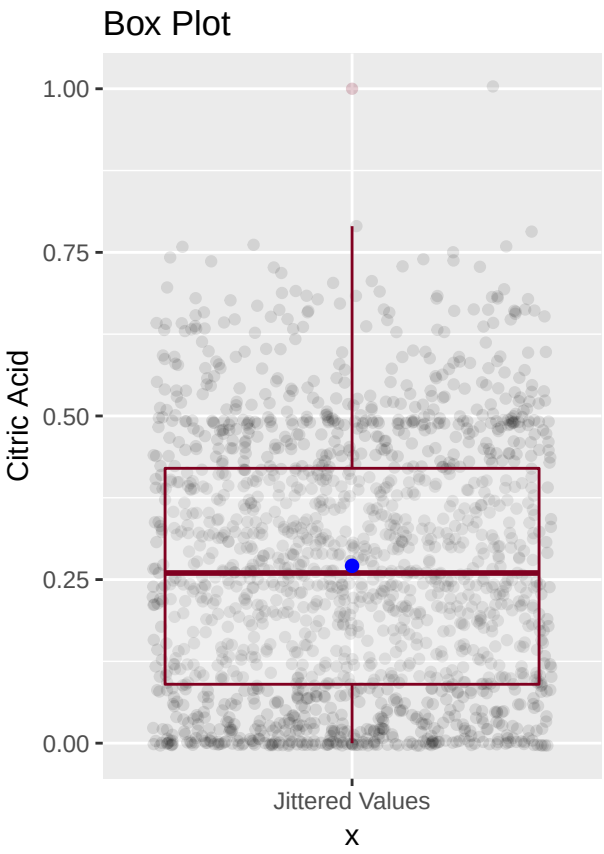
In the above grid, we see the distribution of sulphates of red wines sampled.

Contrary to volatile acidity and similar to alcohol % by volume, the sulphate content of red wines sampled have 27 outliers–19 more than alcohol % by volume. However, compared to alcohol % by volume, we can observe a more narrow distribution at lower levels of sulphates. This indicates that many red wines have low levels of sulphates while higher levels of sulphates are not very likely.

Citric Acid

Histogram & Box Plot





Data Summary

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.09000 0.26000 0.27098 0.42000 1.00000
```

```
## Values greater than or equal to 0.86 are outliers.
```

```
## There are 1 outliers in this variable.
```

```
## The statistics after outlier removal are shown below:
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.09000 0.26000 0.27052 0.42000 0.79000
```

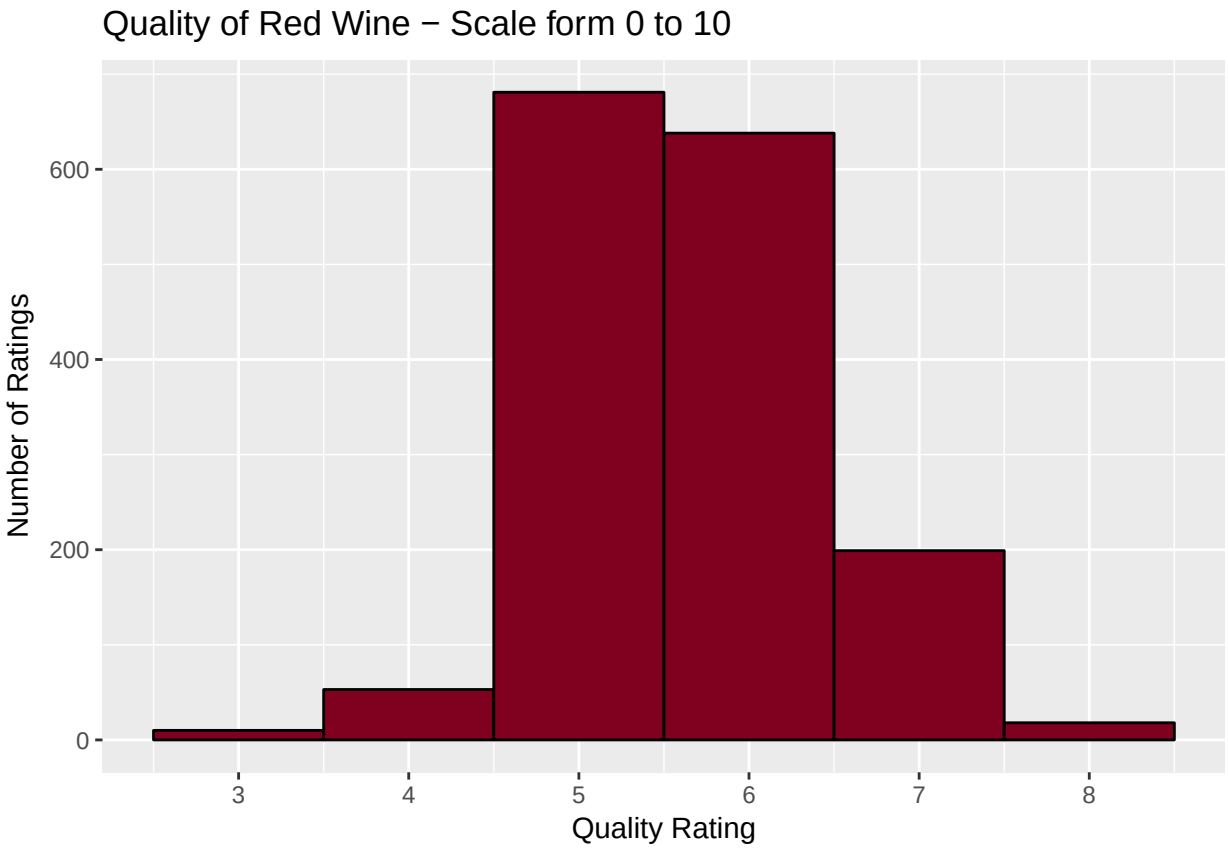
Description

In the above grid, we see the distribution of Citric Acid of red wines sampled.

This distribution seems to be scattered more normal with the exeption of high density in the lower quartile and an outlier in the top quartile.

Quality

Bar Plot



Data Summary

```
## wqr$quality
##        n  missing distinct     Info     Mean      Gmd
##     1599        0        6    0.857    5.636   0.8431
##
## Value          3     4     5     6     7     8
## Frequency     10    53   681   638   199    18
## Proportion 0.006 0.033 0.426 0.399 0.124 0.011
```

Description

In the above bar chart, we see the quantity per quality rating on a scale from 0 to 10. Missing bars indicate 0 ratings.

We can see that at least 3 wine experts mostly found within the sample that most of red wines have a quality of 5 with a proportion of 42.6%.

Univariate Analysis

Main Features Investigated:

Observation of Sampled Red Wine Quality Distribution

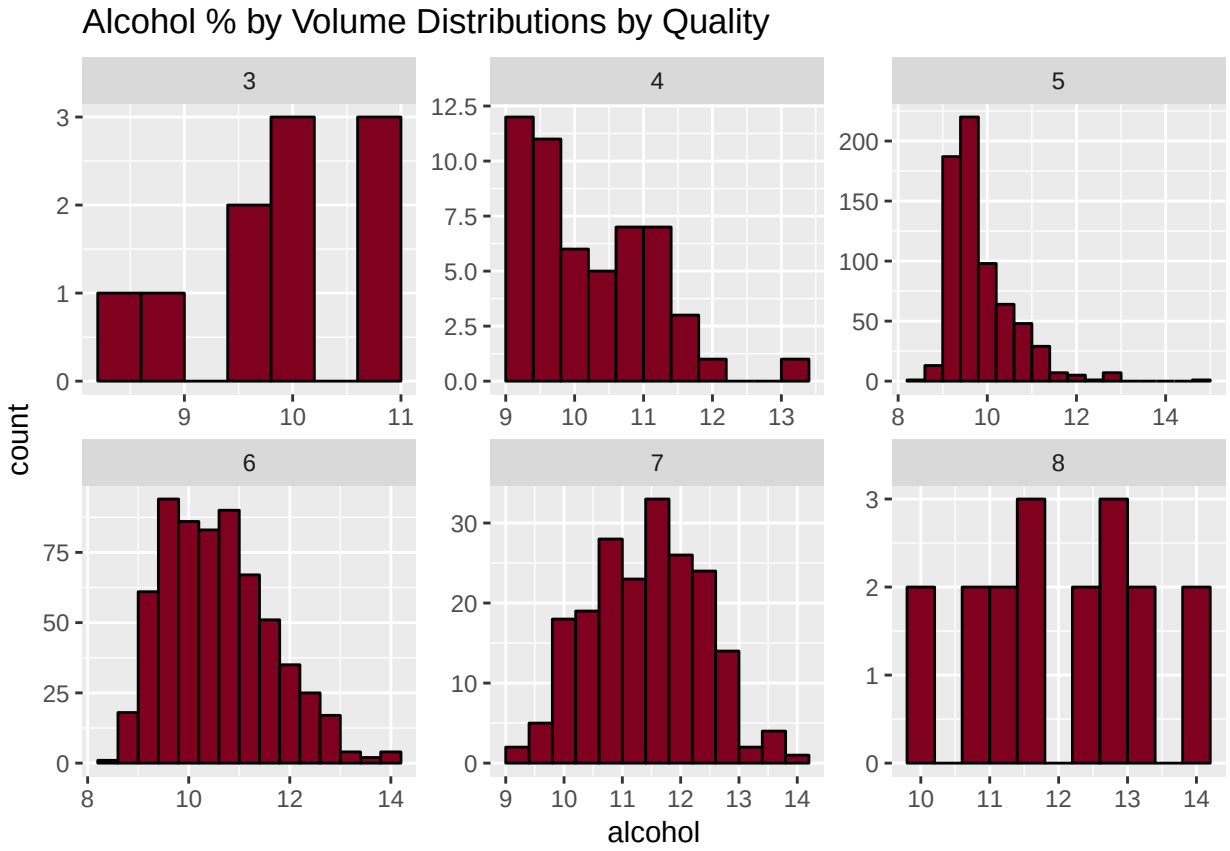
The probability distribution of the quality rating of the sampled red wines is binomial. Most wines in the distribution have a quality output of around 5 of a scale from 0 to 10. Thus, for a big selection of red wines, if we randomly chose a wine, we are most likely to get one with a quality of 5 out of 10 in the eyes of wine experts–quality will variate subjectively depending on the grader.

Observation of Continuous Distributions about Sampled Red Wine Variables

The alcohol % volume is not quite normal but robust. The average red wine has around 10.42% (10.40% after outlier removal) of alcohol by volume. So we are most likely to get a random bottle of red wine that contains around that volume. The same we can say about volatile acidity, sulphates, and citric acid; we are most likely to get a random bottle of red wine that contains around the mean values.

However, I am more interested to know what quality I am most likely to get when I choose a bottle of red wine with 10.40% of alcohol by volume and/or any other physicochemical factor. People like me do not go and randomly select a bottle of red wine–maybe I do sometimes to try something new.

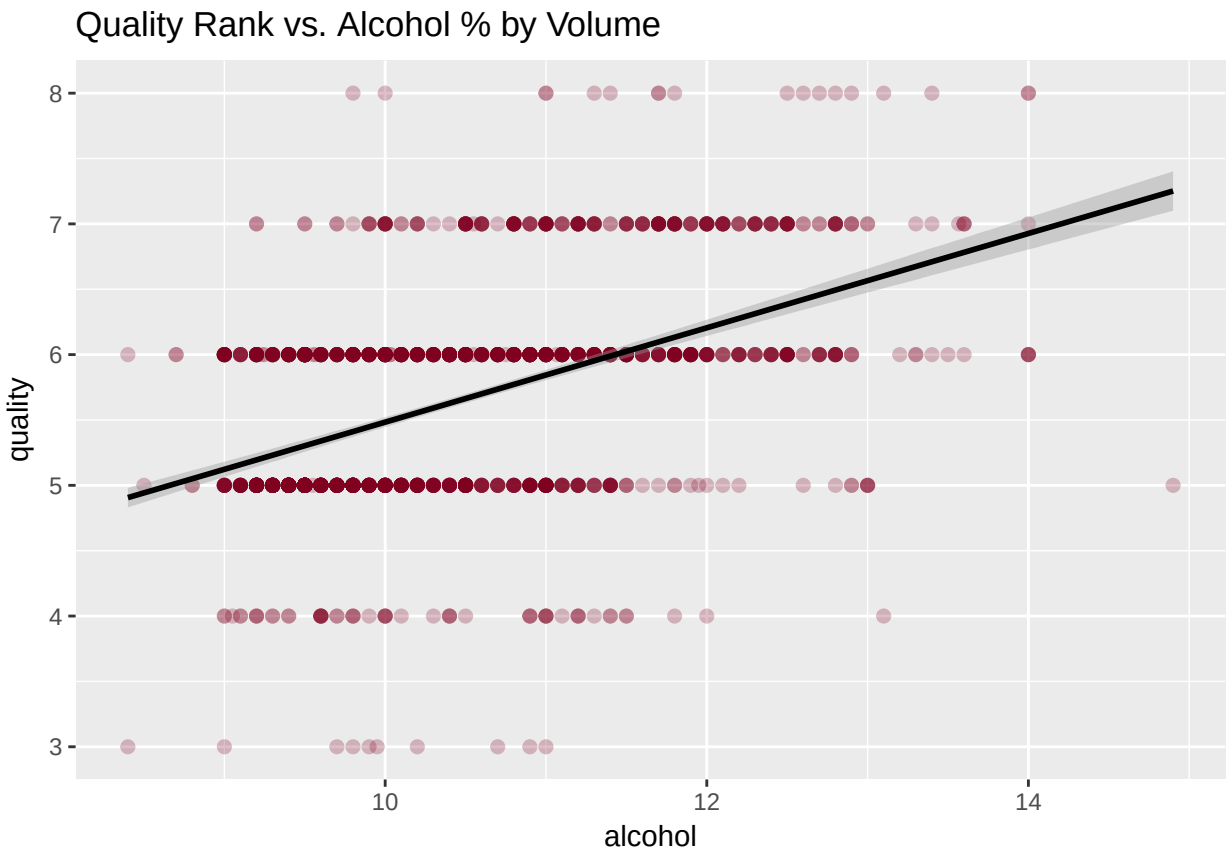
Grid of Histograms of Alcohol % by Volume



We see that there is a fairly uniform distribution for quality rank 8, which tells me that alcohol % by volume for a red wine of this rate is not too much of a driver of quality. Moreover, there is a normal distribution for quality rank 7, and a robust normal distribution for quality rank 7. Quality ranks 5 and 4 are skewed to the right. And quality rank 3 is fairly skewed to the left. The big question is "How impactful overall is the alcohol % by volume on quality?" We are continuing a bivariate analysis of these two variables next.

Bivariate Plots Section

Qualit vs. Alcohol Scatterplot Unjittered

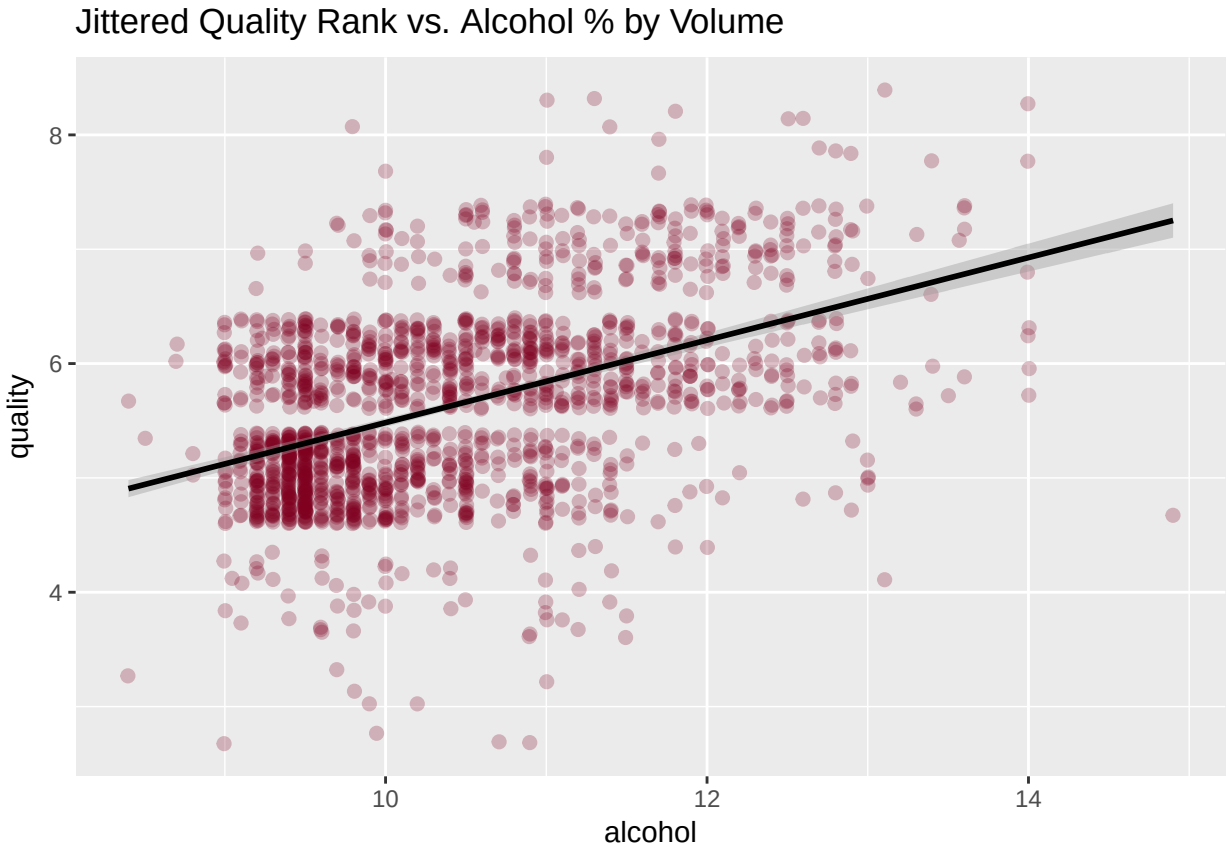


Description

The scatterplot above show us the relationship between quality and alcohol % by volume.

Since the ouput (y) is a discrete interval level of measurement, we can see overplotting in it. And as there is modedrate to wide ranges of alcohol % by volume variability per quality, we can see categorical lines on the cartesian plane. To observe the values better, we are jittering the scatterplot.

Quality vs. Alcohol Scatterplot Jittered



Description

The above scatterplot show us the same relation but jittered allowing us to appreciate the relationshipt closer. Darker colors indicate more density in the relationship.

We can see that there is more overploting in the rating 5 with about 9 to 10 alcohol % by volume. This plot includes the outliers previously mentioned in the alcohol variable. We can see the positive relationship wich seem to be moderate to me. We check to coefficient of correlation next.

Correlation

```
##
##  Pearson's product-moment correlation
##
## data:  wqr$alcohol and wqr$quality
## t = 21.6395, df = 1597, p-value < 2.22e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.43735400 0.51320805
## sample estimates:
##       cor
## 0.47616632
```

Correlation Findings

We can see that the correlation between quality and alcohol is moderate, with a 0.48 coefficient of correlation. If the wine experts were more consistent with their quality ratings and the content of alcohol, this value would be higher. By consistent, I mean that at high levels of alcohol they would grade the red wine as high and at low levels of alcohol, the quality of red wine low. However, we can tell by the scatterplot that that was not the case here. there is a wide distribution of alcohol % by volume on each quality rating, which will make the prediction of quality more complicated.

Model Statistics Description

```
##
## Call:
## lm(formula = quality ~ alcohol, data = wqr)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.84423 -0.41122 -0.16899  0.51661  2.58878
##
## Coefficients:
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 1.874975   0.174710  10.732 < 2.2e-16 ***
## alcohol     0.360842   0.016675  21.640 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.71036 on 1597 degrees of freedom
## Multiple R-squared:  0.22673,    Adjusted R-squared:  0.22625
## F-statistic: 468.27 on 1 and 1597 DF,  p-value: < 2.22e-16
```

Regression Findings

Here we can see that the model is valid since the probability that there is no relationship (p-value) is 0. However, the model does not fit perfectly as I presumed. The coefficient of determination is 0.23 when a perfect fit in the regression is 1. The other .77 needs to be determined by other factors. We check those factors next.

Bivariate Analysis

One thing to notice is that most of the relationship is around 9 to 10 % alcohol by volume as we see in the darkest circles in the jittered plot above.

According to the above scatterplots, the model does not look as strong as I assumed it would be as there is a wide distribution of alcohol % by volume on each quality rating.

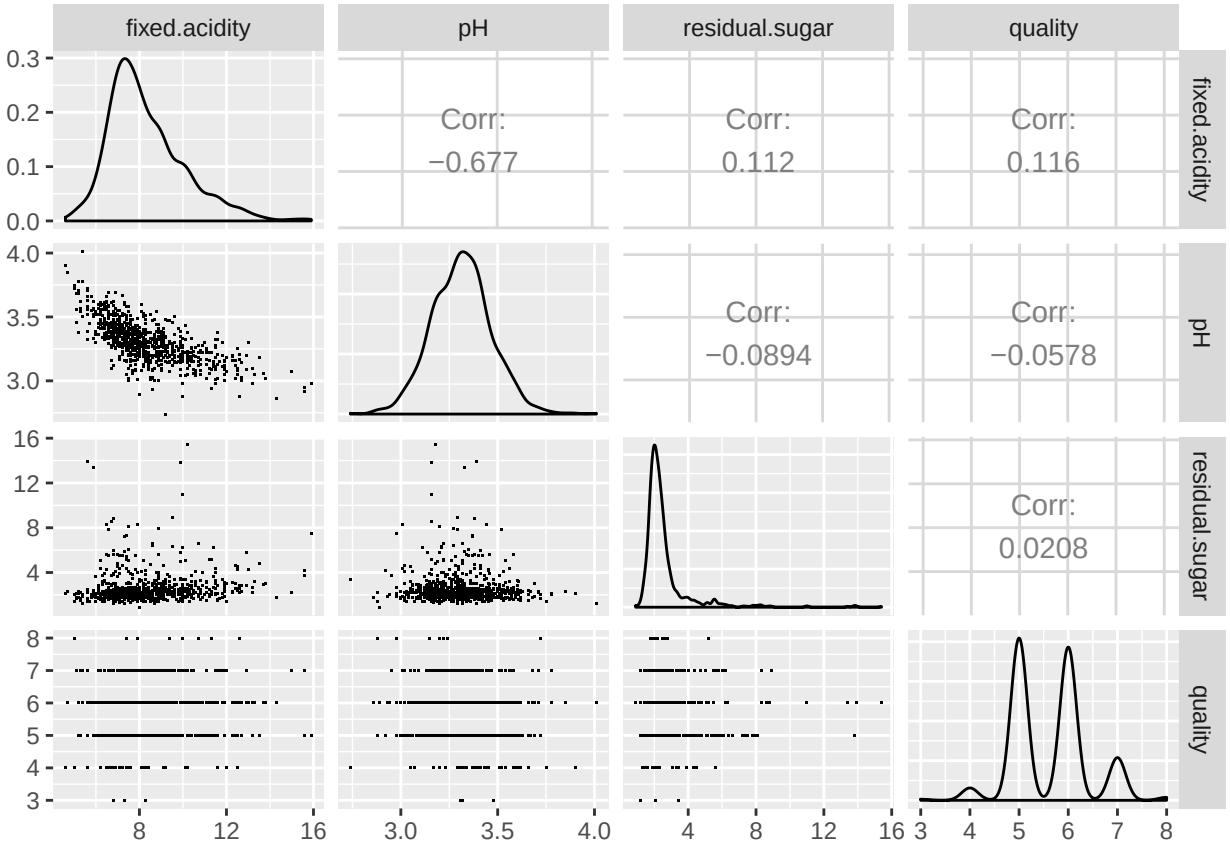
The f-value is significantly high and the p-value is 0. So we can conclude that the relationship between alcohol and quality is significant within a .01 significance level. This means that there is not enough evidence to say that the slope of the line can be zero and no relationship between the two above variables is possible. This makes the model is valid.

As we see by looking at the graph and by the correlation coefficient (0.48), we have a moderate positive relationship.

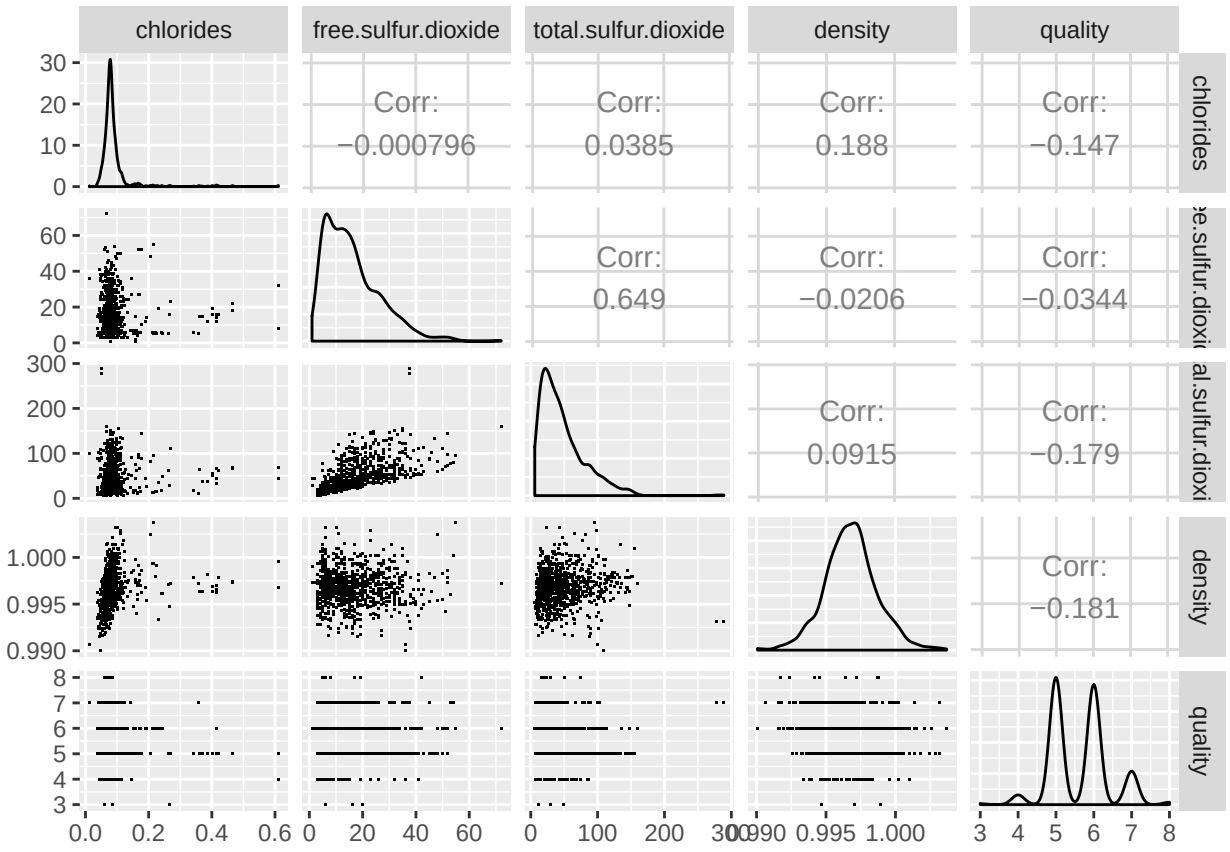
However by looking at the coefficient of determination (.227), we can say that the only 22.7% of the variability in quality can be explained by alcohol % by volume only. As we mentioned, the rest of the variability is due to other factors. On this section, we will introduce new factors into the model to attempt to increase the percentage of variability in quality that can be explained by the new multivariate model.

Identifying Other Bivariate Factors

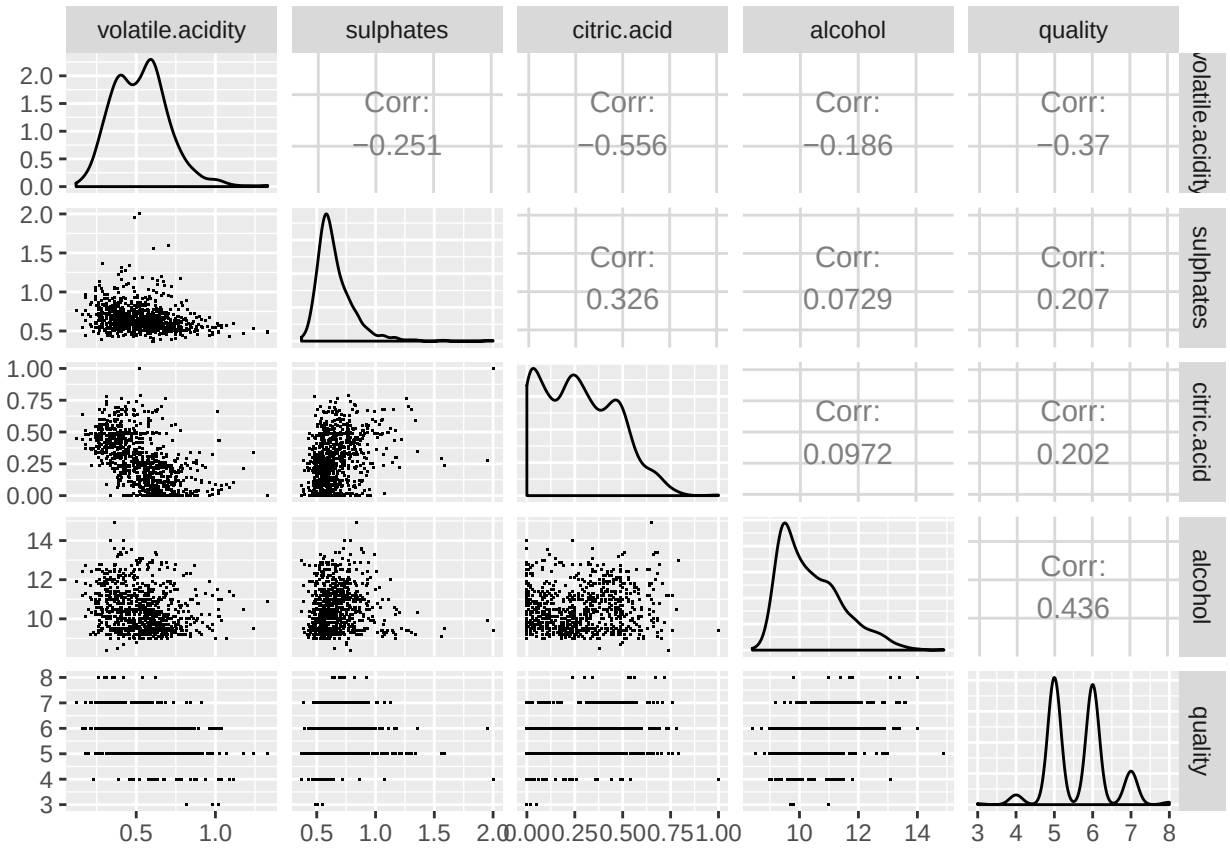
All Relationships in the Dataset with Focus on Quality



On the above grid, there is no substantial correlation with quality.



On the above grid, there is not substantial correlation with quality.



On the above grid, there is a substantial correlation with quality. We have alcohol with 0.436 and volatile acidity with -0.37. And there are two more slightly less substantial correlations with quality such as sulphates with 0.207 and citric acid with 0.202.

Findings On Additional Factors

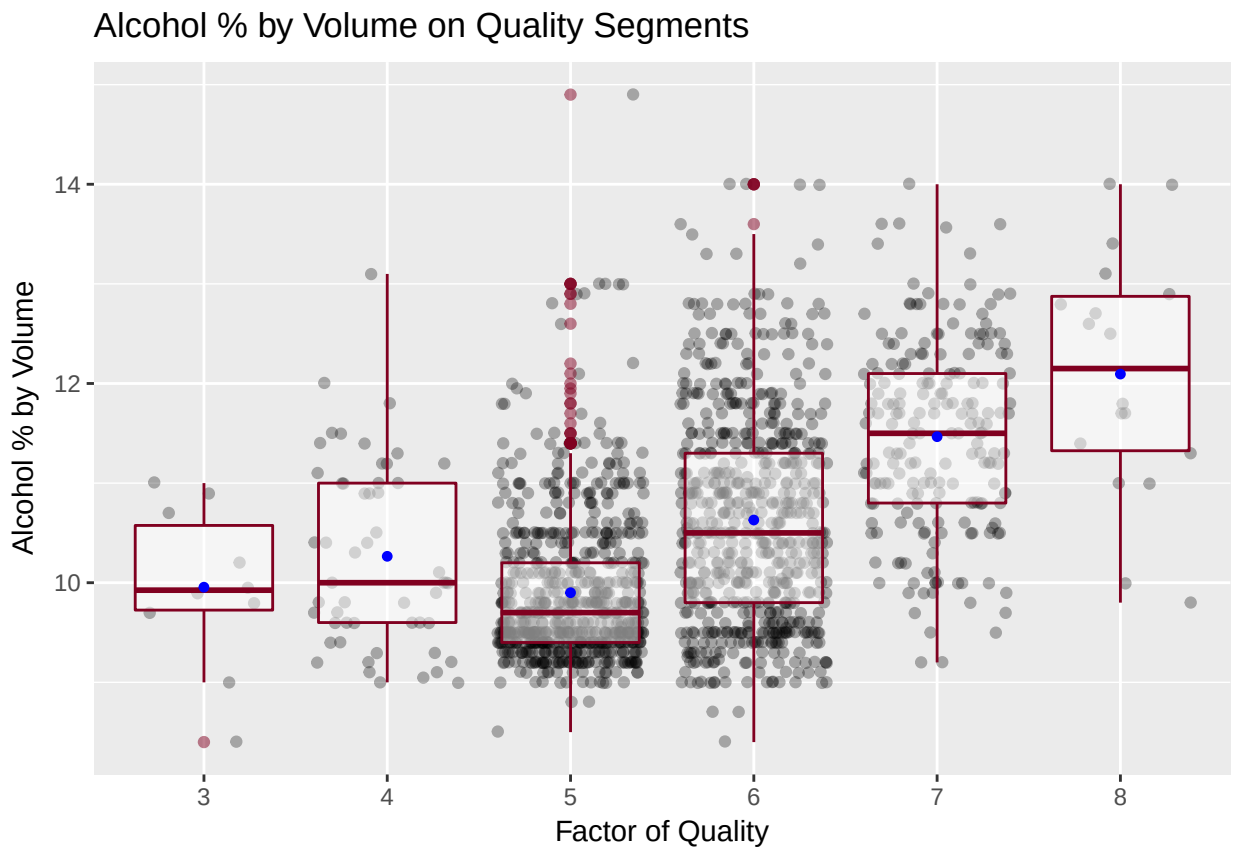
It seems there is no other significant correlation greater than alcohol and quality. The only closest significant correlation of quality is volatile acidity with a negative correlation of 0.37. We would like to add this variable to the model because the correlation with alcohol is not strong.

The model could be stronger if we add sulphates and citric acid. But the issue with using a combination of sulphates and citric acid is that the two independent variables are substantially correlated with each other, which will bring the same meaning to the model and increase the coefficient of determination.

For the model, we are using sulphates since is slightly more correlated to quality than citric acid.

Additional Bivariate Exploration on Promissing Variables

Alcohol/Quality Box Plot



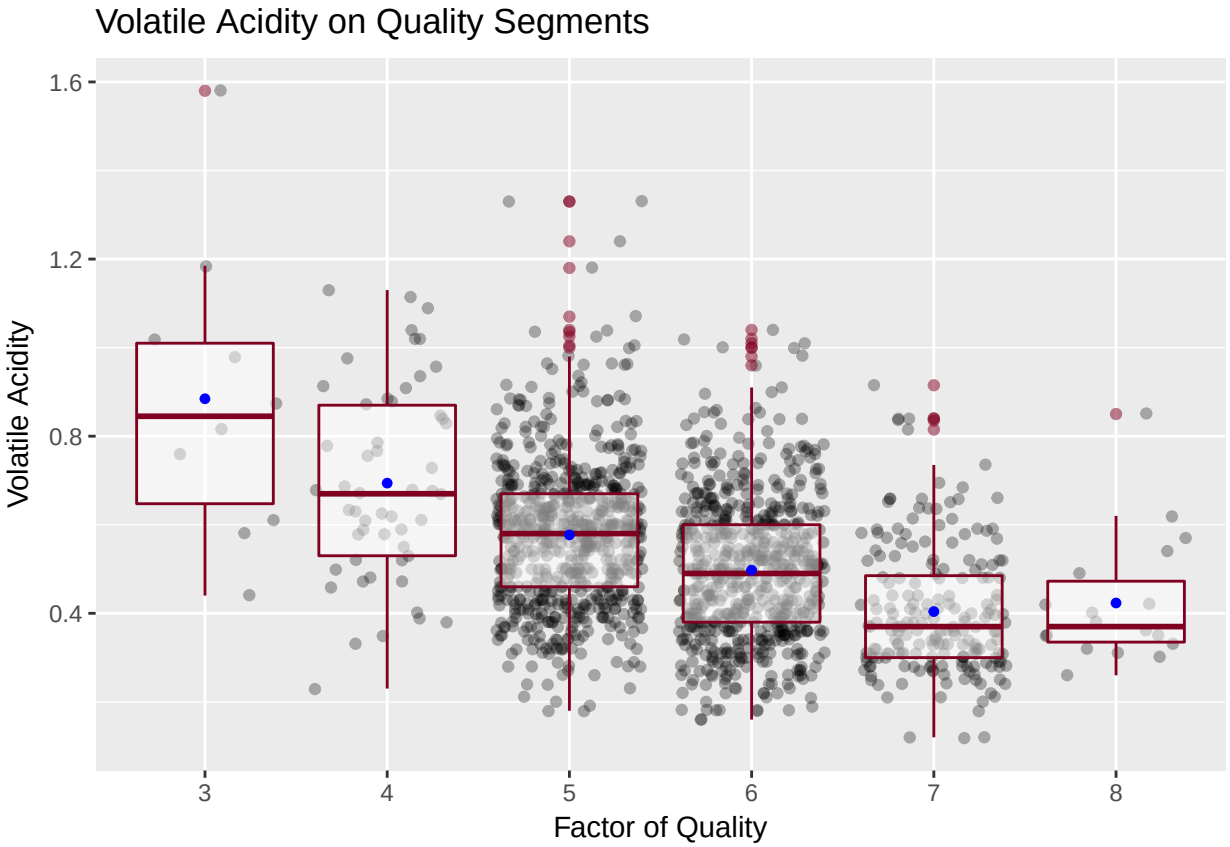
Description

In the above jittered box plot, we see the distribution of alcohol % by volume per quality rating on a scale from 0 to 10. Missing quality ratings indicate no rating.

Here we can observe the alcohol distribution per quality rating closely and separately. We noticed before that the alcohol distribution was most dense around 9.5. We can clearly see here that the density around this area is mostly for red wines with a quality rating of 5. Also, most of the red wines have qualities of 5 and 6 as the density shows and there is more alcohol content for higher qualities of red wines. We can observe outliers on ratings 5 and 6, as well.

Here we can clearly see why the correlation between alcohol and quality is not that strong as each box plot shows a fairly wide distribution with the exception of quality 3. However, this quality rating seems to have too few cases of red wine ratings.

Volatile Acidity/Quality Box Plot



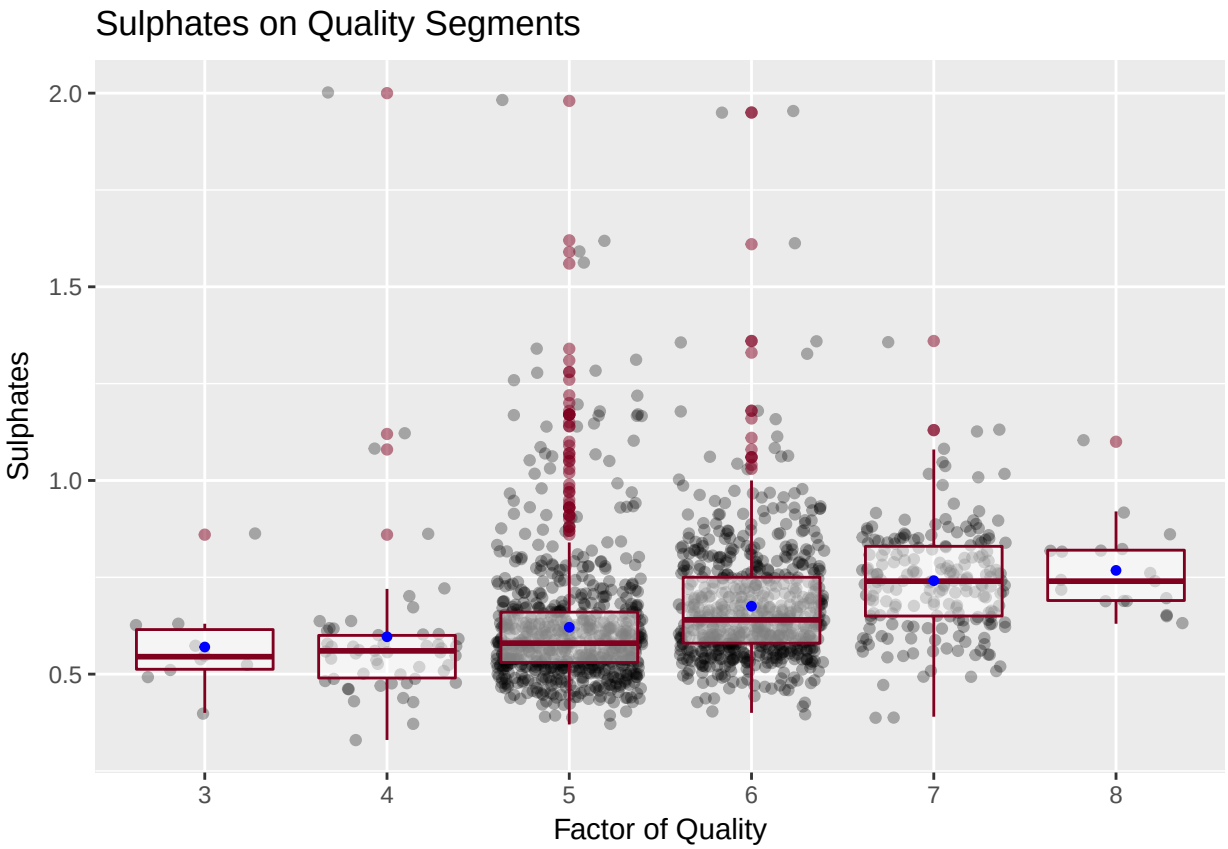
Description

In the above jittered box plot, we see the distribution of volatile acidity per quality rating on a scale from 0 to 10.

Here we can observe that volatile acidity is mostly dense on red wine qualities 5 and 6 as with alcohol % by volume; but contrary to it, there is lower volatile acidity for higher quality red wines. Also, we can observe outliers almost across all the qualities except for quality 4. These outliers should be investigated in correlation with other factors and why they are there before removal.

We can clearly see that there is also wide distributions of volatile acidity within quality ratings, and there are too few cases in quality 3 and 8.

Sulphates/Quality Box Plot



Description

In the above jittered box plot, we see the distribution of sulphates per quality rating on a scale from 0 to 10.

Here we can observe a slightly upward trend as the quality increases and also a high density in quality 5 and 6. For sulphates, there is outlier activity in every quality rating mostly in 5 and 6. They should also be investigated.

We also notice here less wide distribution within sulphates and quality ratings except for those with huge outliers, and there are also too few cases in quality rating 3 and 8.

Continuation of Bivariate Analysis

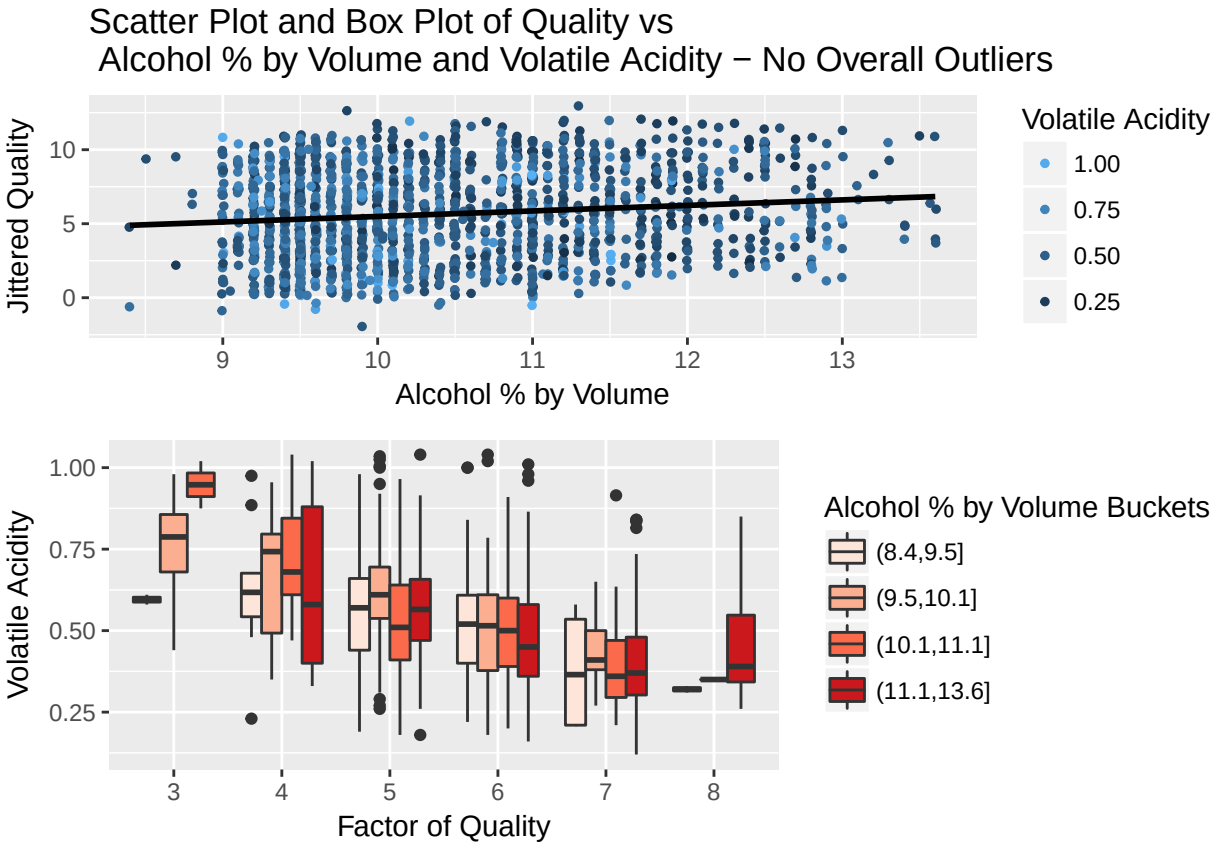
With the additional exploration, we have seen that the distributions of the other factors are not tightly fit to quality rating. This gives me to assume that the variability of the quality is mostly subjective to the wine expert reactions to these ingredients. The variables explored above where the most correlated with quality rating, so the other variables are even less of a qualifier of a good or bad red wine. Perhaps there are other unavailable variables that we are missing that can explain these wide distributions within quality ratings. We are going to explore further with more variables to determine if that is the case.

Multivariate Plots Section

Since there are overall outliers affecting our analysis, any value greater than 3 standard deviations from the respective average will be omitted in the rest of the analysis.

The prhase "overall outliers" represent the outliers of the whole distribution of a particular variable, which is not segmented by other variable.

Quality, Alcohol, and Volatile Acidity



Description

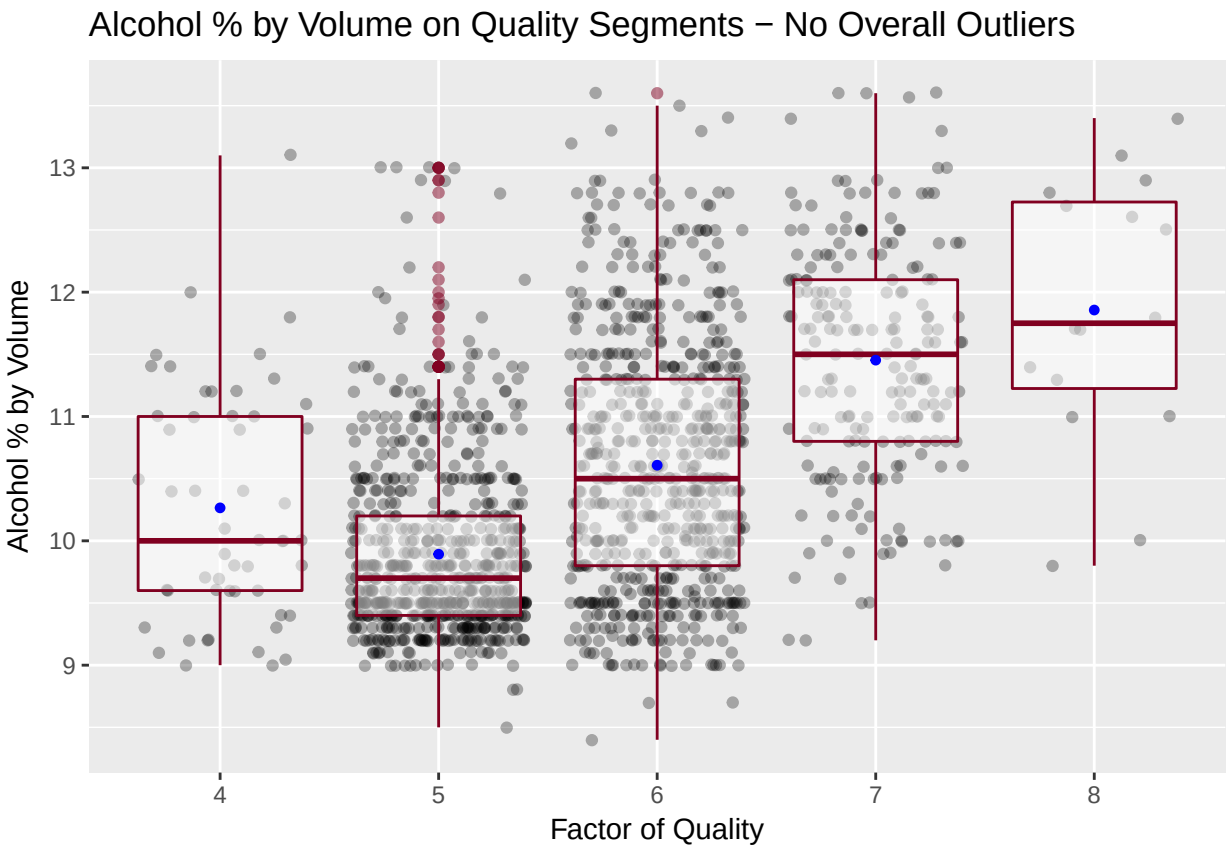
In the jittered scatterplot, we see a positive trend between quality and alcohol but we cannot observe the volatility well due to overplotting. We already know that the relationship with quality is negative as we saw before.

Looking at the box plot, we can see a downward trend between quality and volatile acidity for each quartile bucket of alcohol % by volume.

From here, I can see that high-quality red wines have low levels of volatile acidity on a median level. I can see that high alcohol levels have better quality also on a median level; we can corroborate the latter by looking at the scattered box plot below.

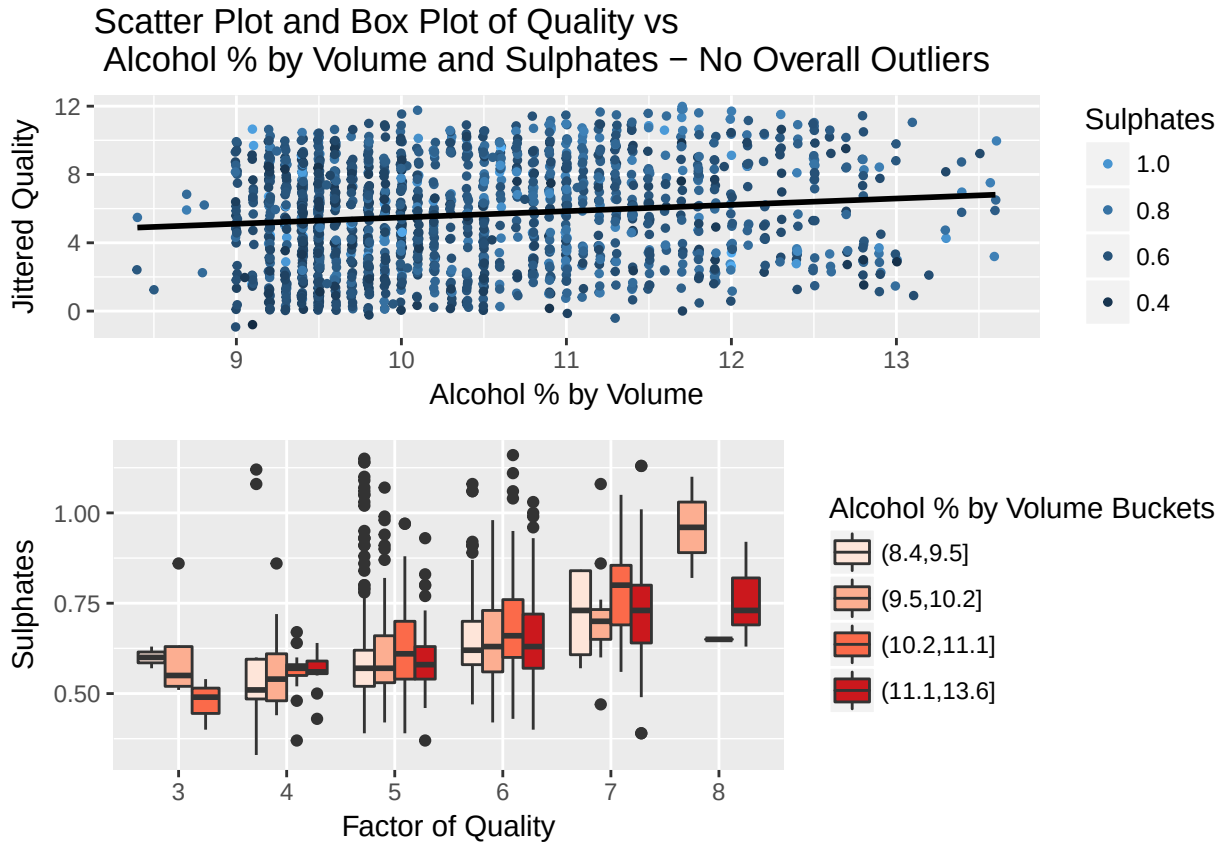
There is an issue with the quartile segmentation in the box plot above, which prevent us from comparing the distribution within the quality segments. If we see the distribution of quality 8 on the scattered box plot below, we see that there are not enough cases to provide an accurate

quantile statistics for that quality rating. One suggestion is to increase the sample size to provide us with more cases under this segment. A drawback will be that quality rating 5 and 6 will be more overplotted assuming that the rest of the population is dense under these quality ratings.



*this is the same scatterplot under the bivariate section but with no overall outliers.

Quality, Alcohol, and Sulphates



Description

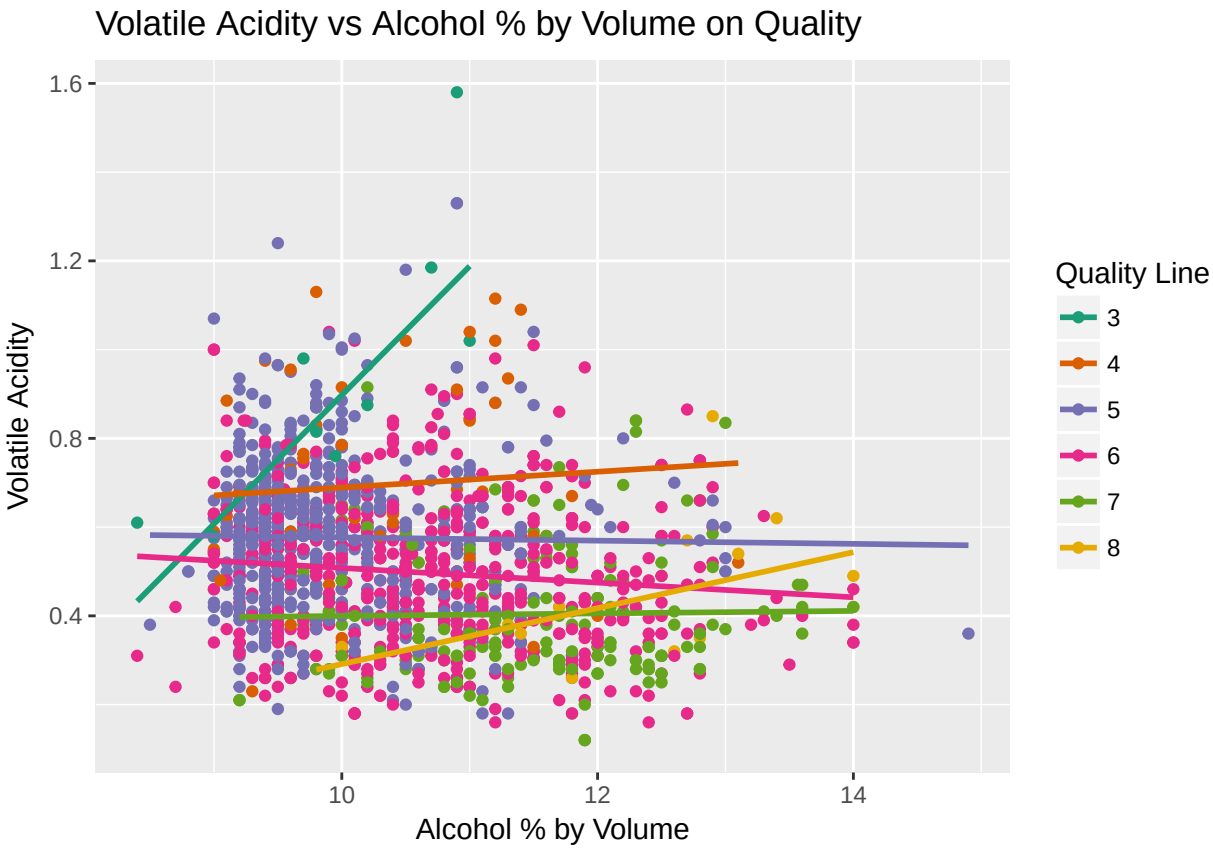
In the above scatterplot, we can see the same relationship between quality and alcohol % by volume. We have the same issue as the volatile acidity with sulphates due to overplotting. But we already know that the relationship between quality and sulphates is slightly positive.

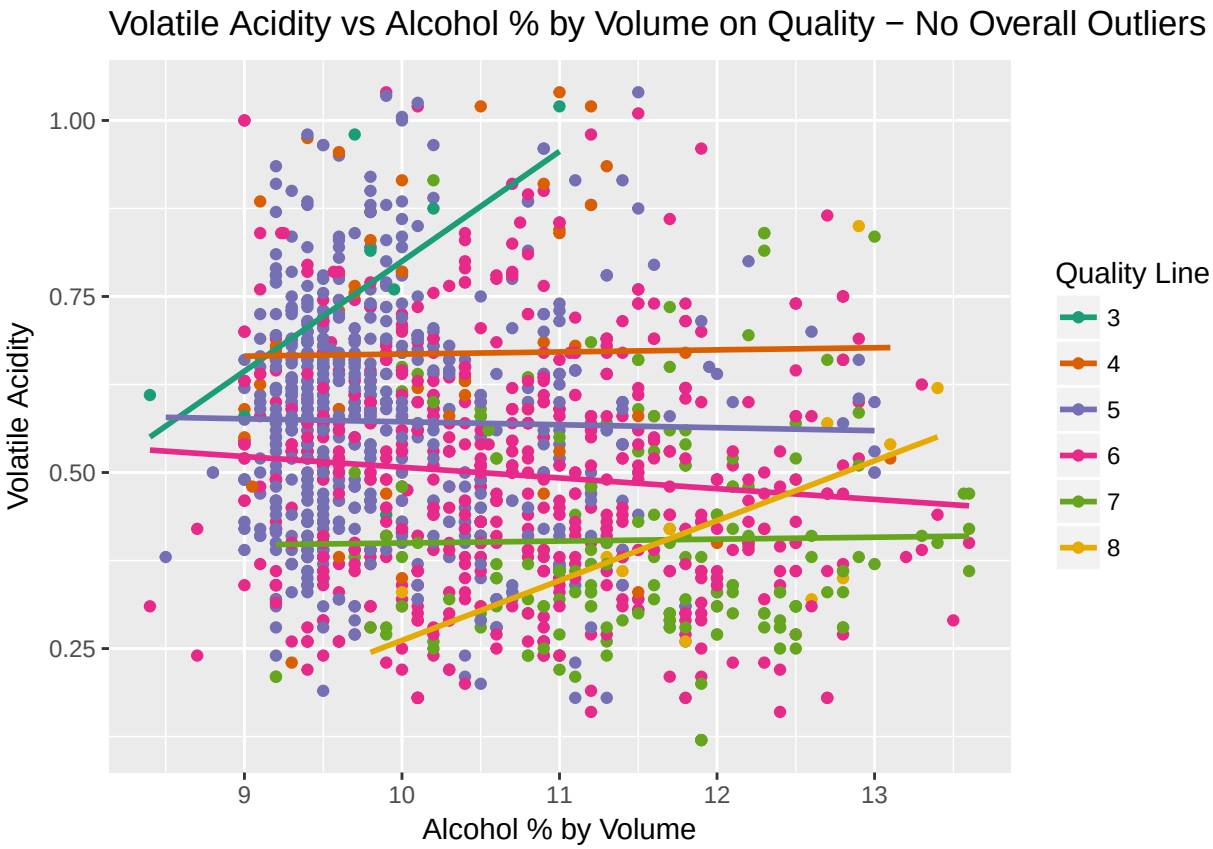
Looking at the box plot, we can tell that the relationship between quality and sulphates is positive in almost every quartile bucket of alcohol % by volume. However, we are faced with the same sample size issue in both quality rating 3 and 8; there is not enough data to compare the within quality rating distribution of sulphates.

Here there are many outliers within quality ratings and quartile alcohol % by volume buckets that are worth investigating further. I wonder how the sulphates actually affect the decision of the wine experts and if they are aware of the flavor changes specifically for red wine quality ratings of 5 and 6.

Multi-variable Scatterplot

Volatile Acidity(y) vs Alcohol(x) View


Volatile Acidity vs Alcohol % by Volume on Quality
Volatile Acidity
1.6
1.2
0.8
0.4
10
12
14
Alcohol % by Volume
Quality Line
3
4
5
6
7
8



Description

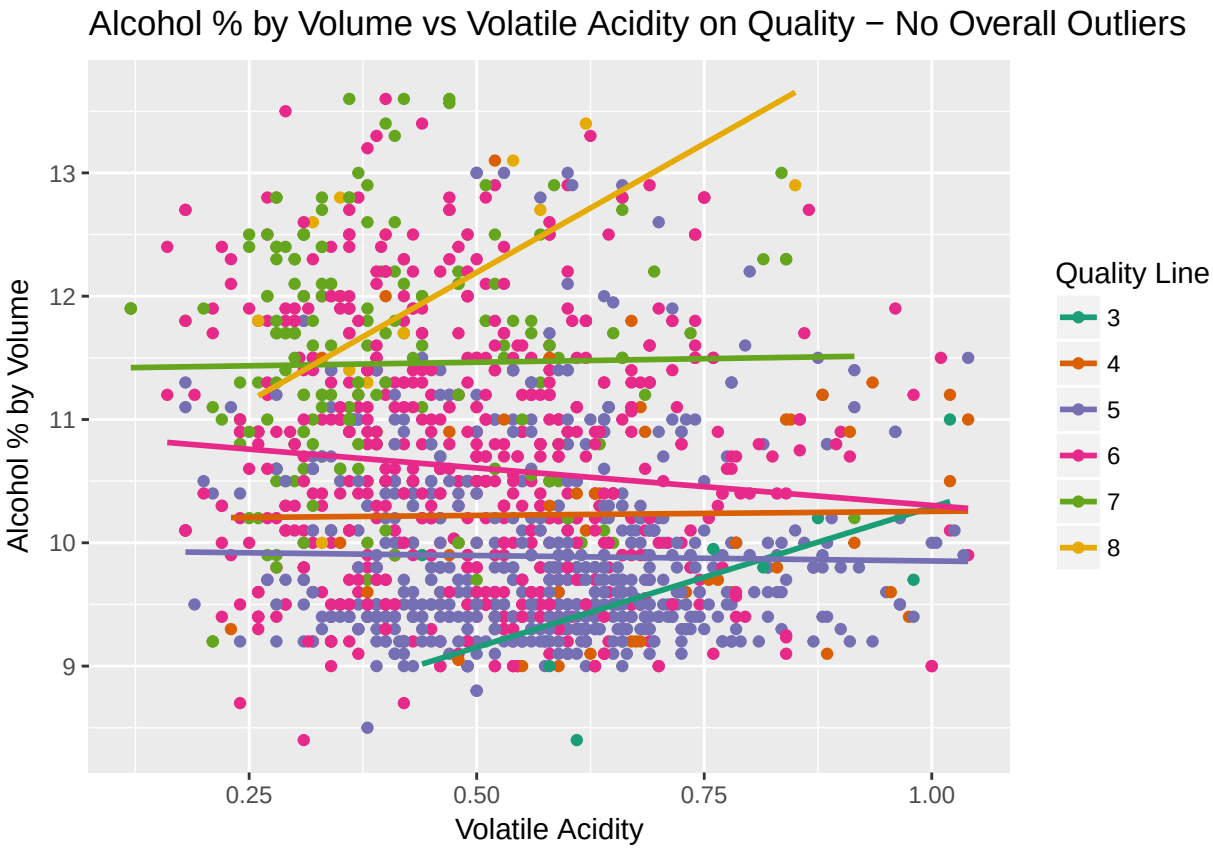
In the above multiple-variable scatterplot, we can see the relationship between volatile acidity and alcohol % by volume on a categorical line legend.

The top graph shows the relationship including the overall outliers of all quantitative variables. The bottom graph shows the same relationship but without the outliers. The purpose is to see if the outliers change the structure of the graph. Indeed it does, but only slightly. The strongest noticeable change is in the quality rating number 4 line–from a positive trend, it goes to a slightly stagnant line, showing almost no relationship.

By looking at the second graph (with no overall outliers), I can see that volatile acidity tends to not matter as much in quality ratings 3 and 8. However, this can be due to the sample issue mentioned before. Thre is just not enough cases under these quality ratings to have a more accurate representation of the trend.

In quality ratings 4 to 7, we can see that the lower the volatile acidity, the higher the rating is at any level of alcohol % by volume.

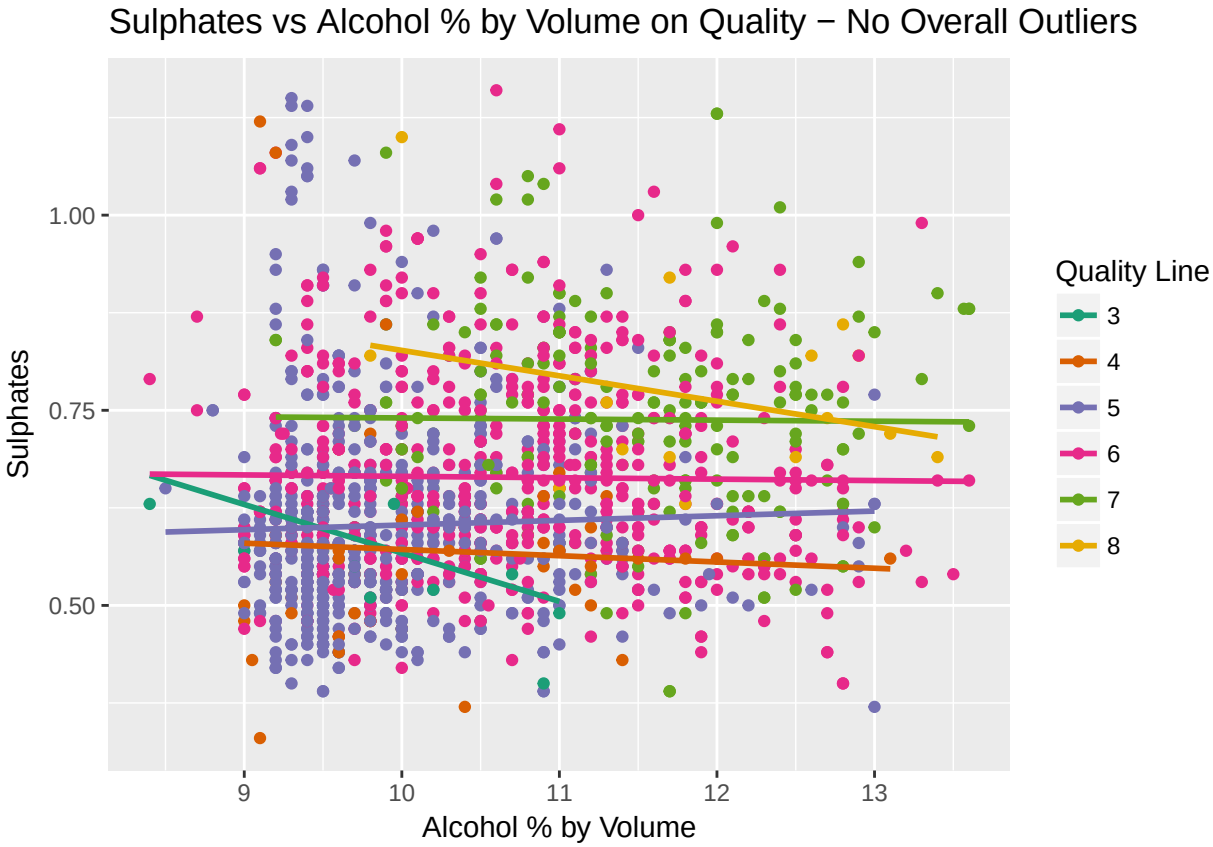
Alcohol(y) vs Volatile Acidity(x) View



Description

By reversing the previous multiple-variable scatterplot, we can appreciate the categorical relationship between quality and alcohol % by volume.

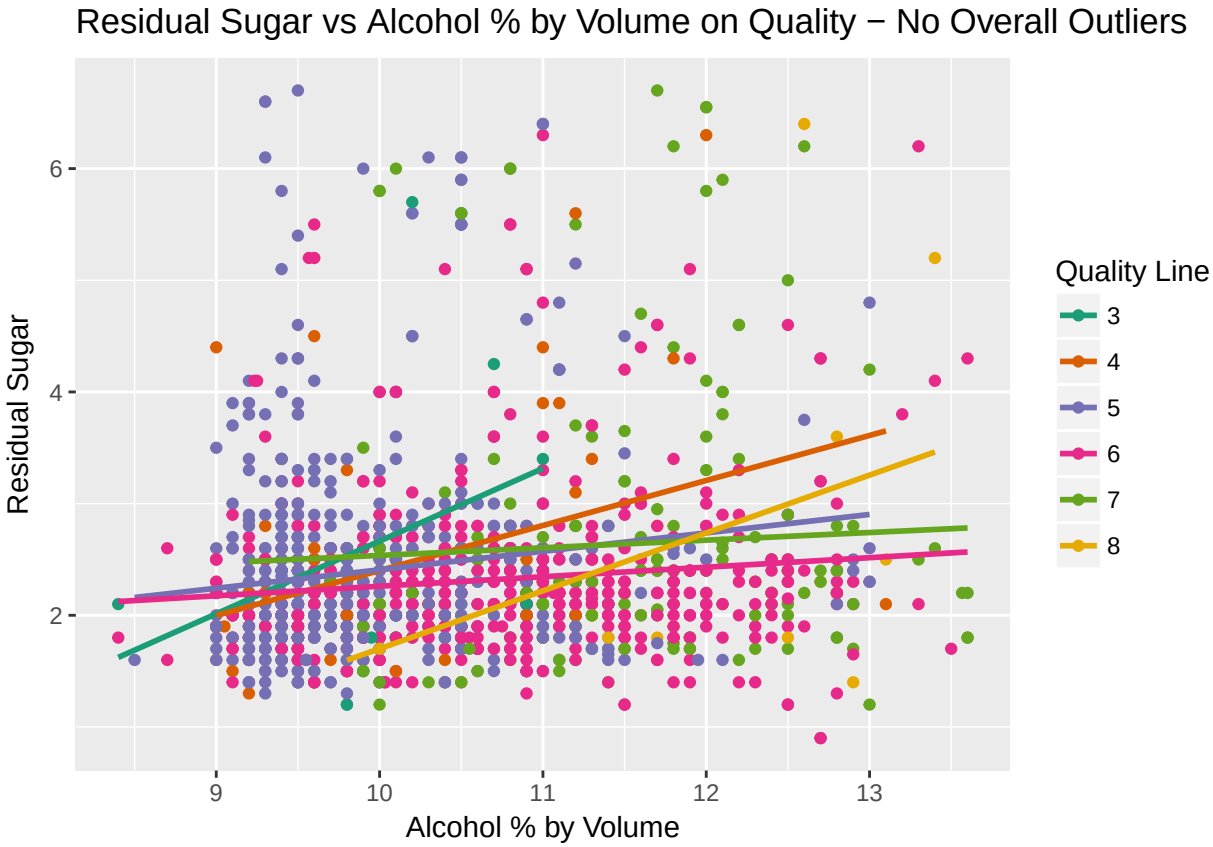
We can see the same possible issue with quality ratings 3 and 8. However, for quality ratings 4 to 7, we can tell that higher alcohol % by volume values tend to increase the quality rating of a red wine given by a wine expert. There is the exception of quality ratings 4 and 5, which are inverted. We can also see that at quality levels 4 and 6, with high levels of volatile acidity is harder to determine the quality of a red wine as it can be either or.



Description

Needless to say, we can see the horizontal pattern in the quality lines here as sulphates are correlated with quality at any level of alcohol.

To prove the concept, let us see an uncorrelated variable in the plot below. That is residual sugar.



It is impossible to determine the quality of a red wine by using residual sugar at each level of alcohol % by volume. There too many competing values and the lines are more close to each other.

Multiple Regression Model Statistics

```
##
## Calls:
## m1: lm(formula = quality ~ alcohol, data = wqr)
## m2: lm(formula = quality ~ alcohol + volatile.acidity, data = wqr)
## m3: lm(formula = quality ~ alcohol + volatile.acidity + sulphates,
##     data = wqr)
##
## ================================================================
##                           m1           m2           m3
## ----------------------------------------------------------------
##   (Intercept)           1.875***     3.095***     2.611***
##                        (0.175)      (0.184)      (0.196)
##   alcohol               0.361***     0.314***     0.309***
##                        (0.017)      (0.016)      (0.016)
##   volatile.acidity                  -1.384***    -1.221***
##                                     (0.095)      (0.097)
##   sulphates                                       0.679***
##                                                  (0.101)
```

```
## ---------------------------------------------------------------
##   R-squared             0.227         0.317         0.336
##   adj. R-squared        0.226         0.316         0.335
##   sigma                 0.710         0.668         0.659
##   F                   468.267       370.379       268.912
##   p                     0.000         0.000         0.000
##   Log-likelihood    -1721.057     -1621.814     -1599.384
##   Deviance            805.870       711.796       692.105
##   AIC                3448.114      3251.628      3208.768
##   BIC                3464.245      3273.136      3235.654
##   N                  1599          1599          1599
## ===============================================================
```

Multiple Regression Findings

Attempting to increase the variability in quality ratings that can be explained by the model, we added volatile acidity and sulphates. The model increases its coefficient of determination to 33.5% from 22.7%, and it is still a valid model. However, without a variable that is more correlated to quality in clusters, there are low chances of increasing the predictability of quality.

Multivariate Analysis

In this analysis, I corroborated that we should increase the size of the sample and get more wine expert ratings of red wines.

I also noticed that there is there are significant relationships with volatile acidity, alcohol, and sulphates and quality. We know that we can use these variables in the model to predict the quality of a wine base on wine expert grading.

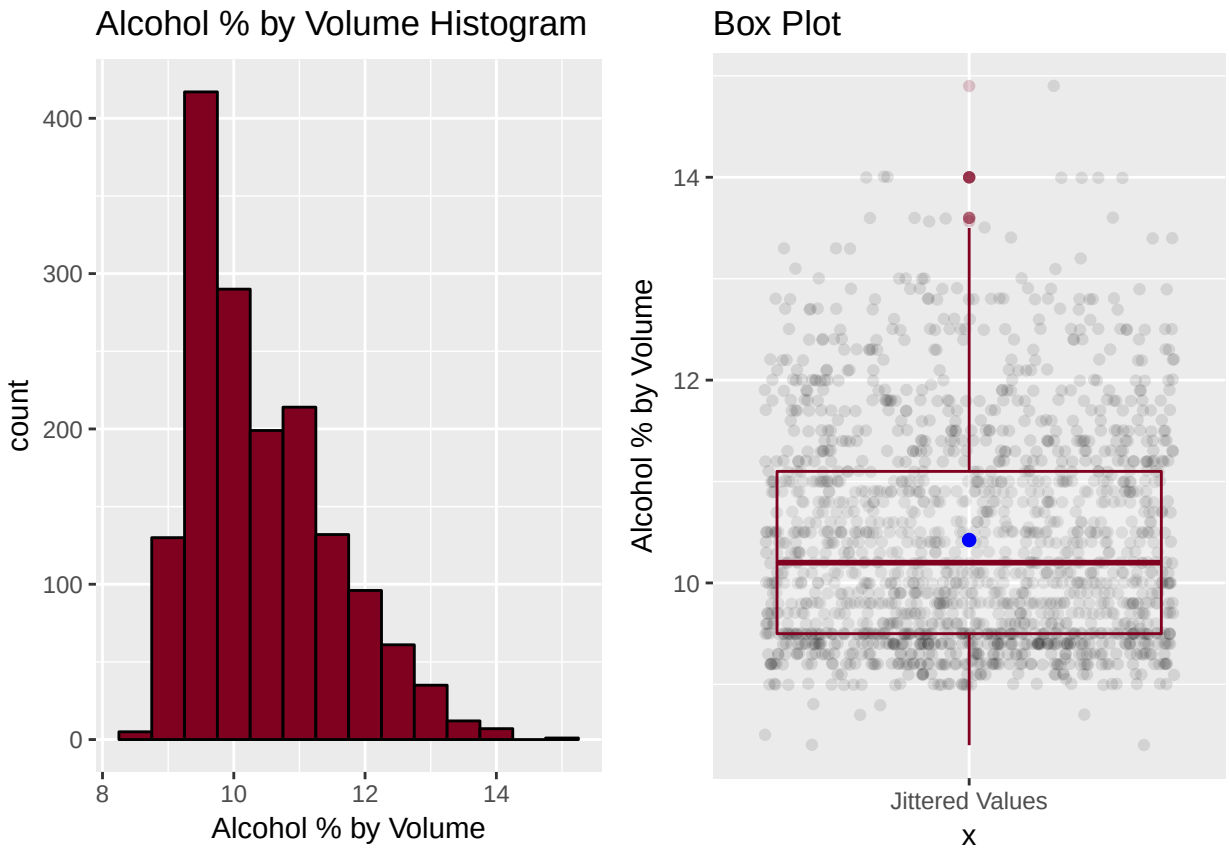
OPTIONAL:

Strengths and Limitations of Model.

The model above describes the relationship between quality and alcohol, volatile acidity and/or sulphates. The addition of sulphates to the model only adds 1% of the adjusted explanation of variability in quality using the full model with alcohol, volatile acidity, and sulphates. This model although strongly valid given by the high F-value and 0.00 p-value is very weak as a predictor of quality. I managed to fit the model a bit more from 23% to 33% (adjusted r-squared). However, it is still quite weak. Honestly, I would not use this model to predict the quality of a wine as it seems to me that the quality of a red wine is more subjective and slightly objective. We need more data, perhaps specific to preferences about expert giving the quality rating.

Final Plots and Summary

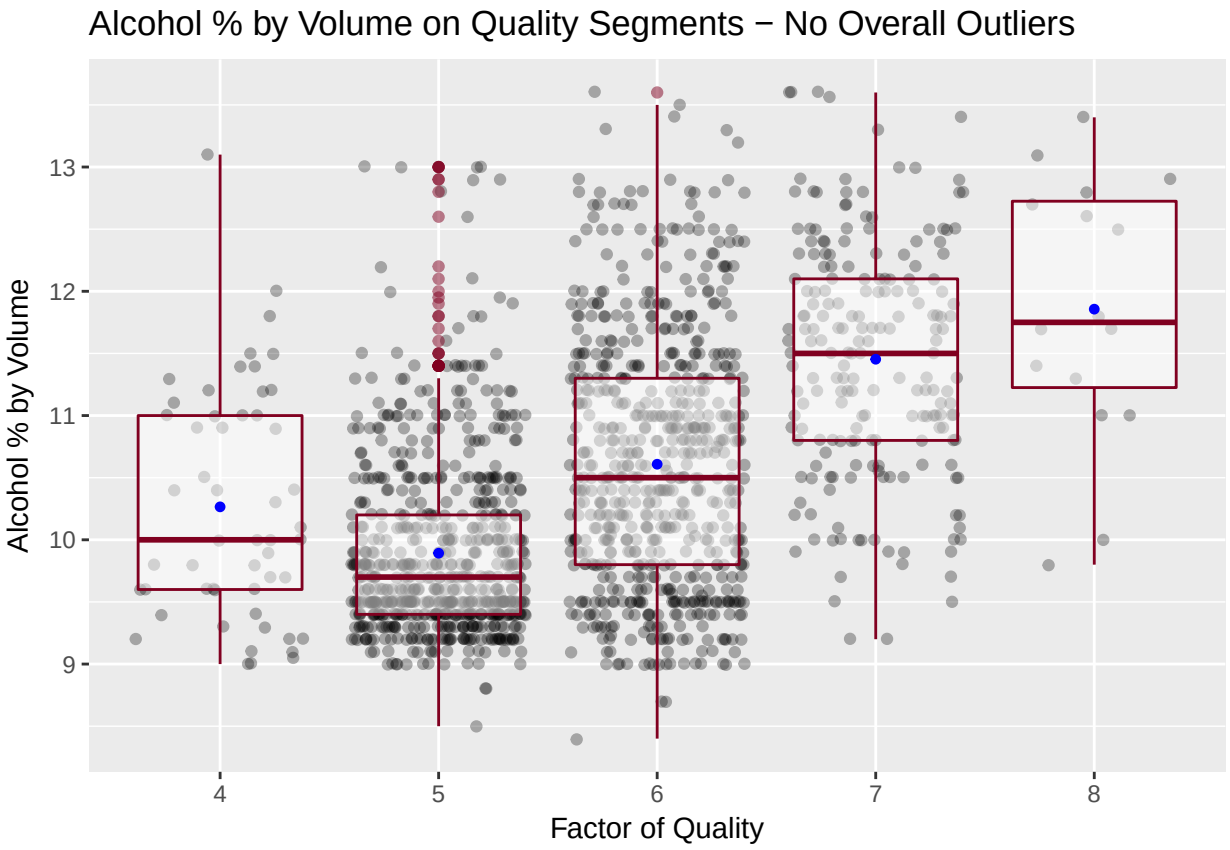
Plot One



Description One

In this grid, I determined that there are possible outliers affecting the distribution of alcohol. The same graph was done for the rest of the potential variables, which led me to conclude that an outlier removal was necessary for this analysis as they are unlikely events in which the wine expert might have been biased when giving the quality rating.

Plot Two

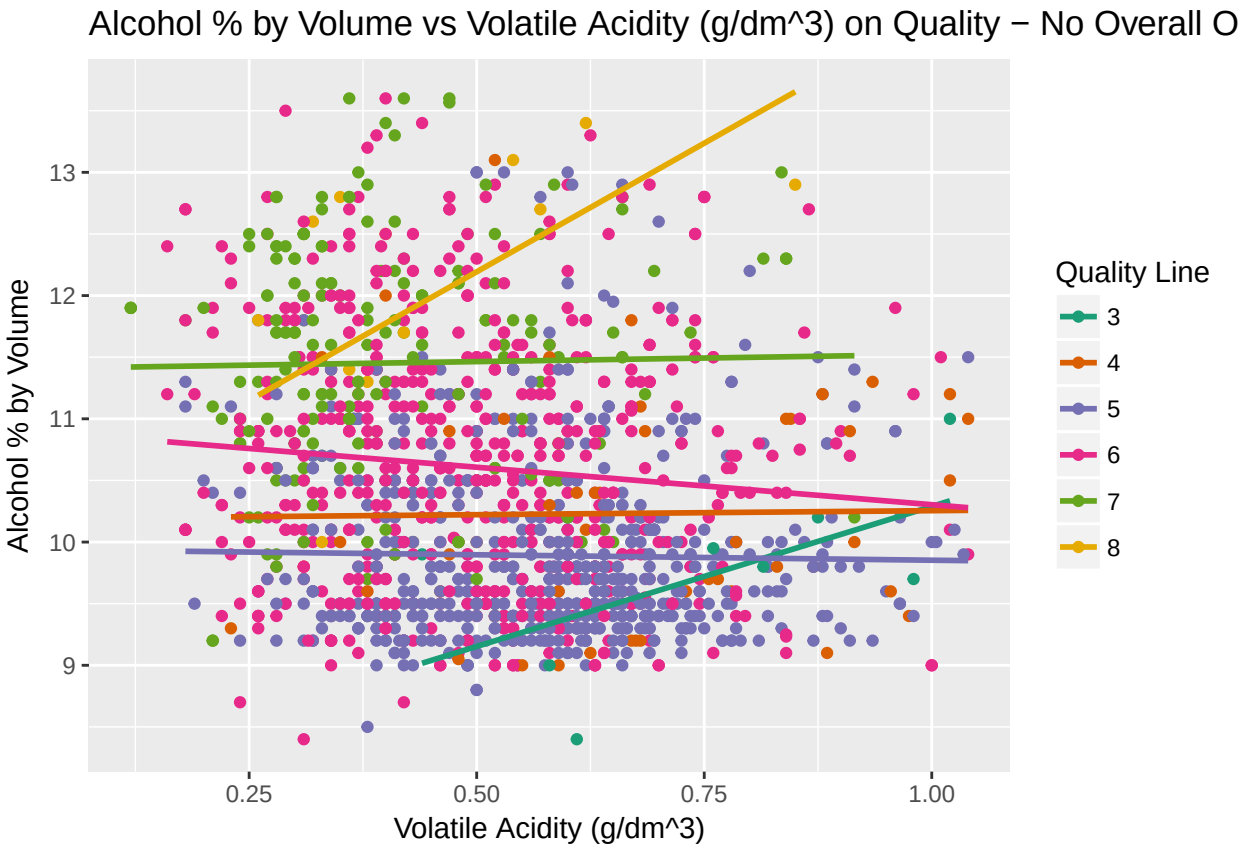


Description Two

I ended up adding volatile acidity to the multivariate analysis because it gave me a promising correlation with quality. We can see that by looking at the middle box plot. The visual relationship agree with the -0.37 coefficient of correlation value when it was previously analyzed against quality using the pearson's method.

I assumed that there was a categorical relationship between volatile acidity and quality, as well as sulphates and quality. We see that on each quality rating, there is a slight decrease in volatile acidity. Perhaps we can notice this trend better by doing a multiple-variable scatterplot and include sulphates as the gradient to see its density to attempt to identify density.

Plot Three



Description Three

Indeed, we can see the changes in alcohol % by volume on each quality rating for each continuous level of volatile acidity. And vice-versa we can see the changes in volatile acidity on each quality rating for each continuous level of alcohol % by volume. However, as I mentioned before there is an inverse relationship between quality rating 4 and 5, and there is a level of volatile acidity where the wine experts cant decide whether to choose a quality of 4 or 6. Perhaps we should investigate further with exploratory analysis to understand their choices.

Reflection

After experimenting with the red wine dataset, I gain insight about the tendency to grade a red wine 5 or 6 out of 10. I also learned that wine experts have the tendency to grade good red wines when there is low volatile acidity. The same happens with alcohol but when it has high contents of percentage by volume, with the exception of alcohol % by volume levels of around 9.5 and 10.5, which is worth investigating.

But I also learned that the limitations of the dataset due to privacy and logistics issues is detrimental to the exploratory analysis conducted here. My next step is to gather more variables such as grape types, wine brand, wine selling price, county of the wine made, year of the wine, etc. Also, we should investigate preferences of the wine experts grading the wines as well.

The reason of the density in quality 5 and alcohol % by volume of around 9 and 10 could be due to consumer preference, including many of those customers preferring higher alcohol contents on their red wine with an expert rating of 5. Perhaps these red wines are mid-priced as well. This is why I need more info about prices, etc. But to test my hypothesis of the density, I will need to do an experimental analysis on consumer preference to validate my claim. I would also do an observational study on consumer demand.

In addition, we need to increase the sample size to be able to compare values in quality ratings 3 and 8, which show a different pattern than the rest of the ratings. It seems that at these quality ratings, wine experts cannot determine if the quantity of alcohol or volatile acidity contribute to their choices for quality of a red wine. But there is not too much data under those ratings.

Another exploration is related to why these wine experts cannot decide whether a red wine with alcohol % by volume of around 10.3 and volatile acidity (g/dmˆ3) content of around 1.12 have a quality of 4 or 6. This is worth investigating further to understand the decision-making process of the wine experts and assess which other factors we should include in the analysis and future datasets.