Red Wine Quality by Trevor Cook

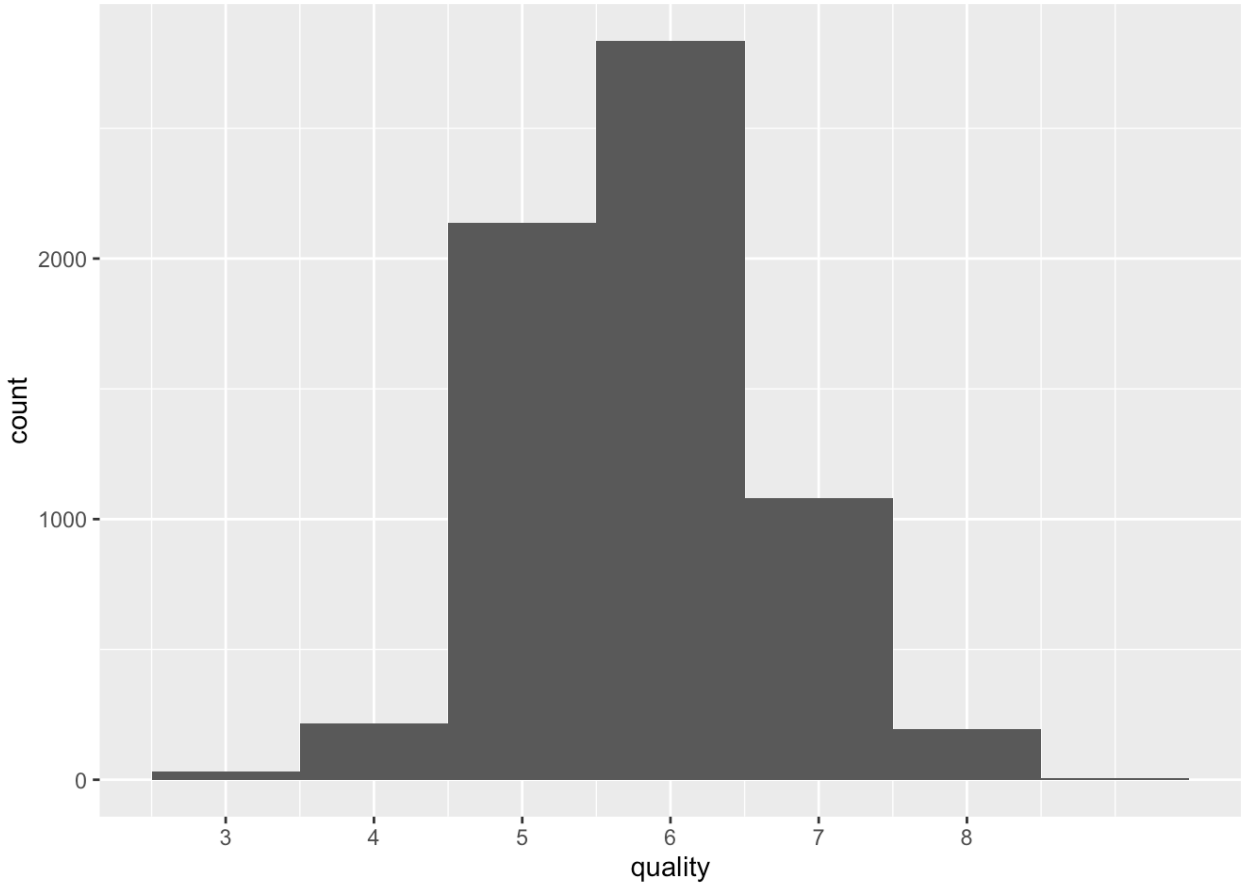
This report focuses on analyzing a dataset of the characteristics and ratings of approxiamtely 6500 red and white wines.

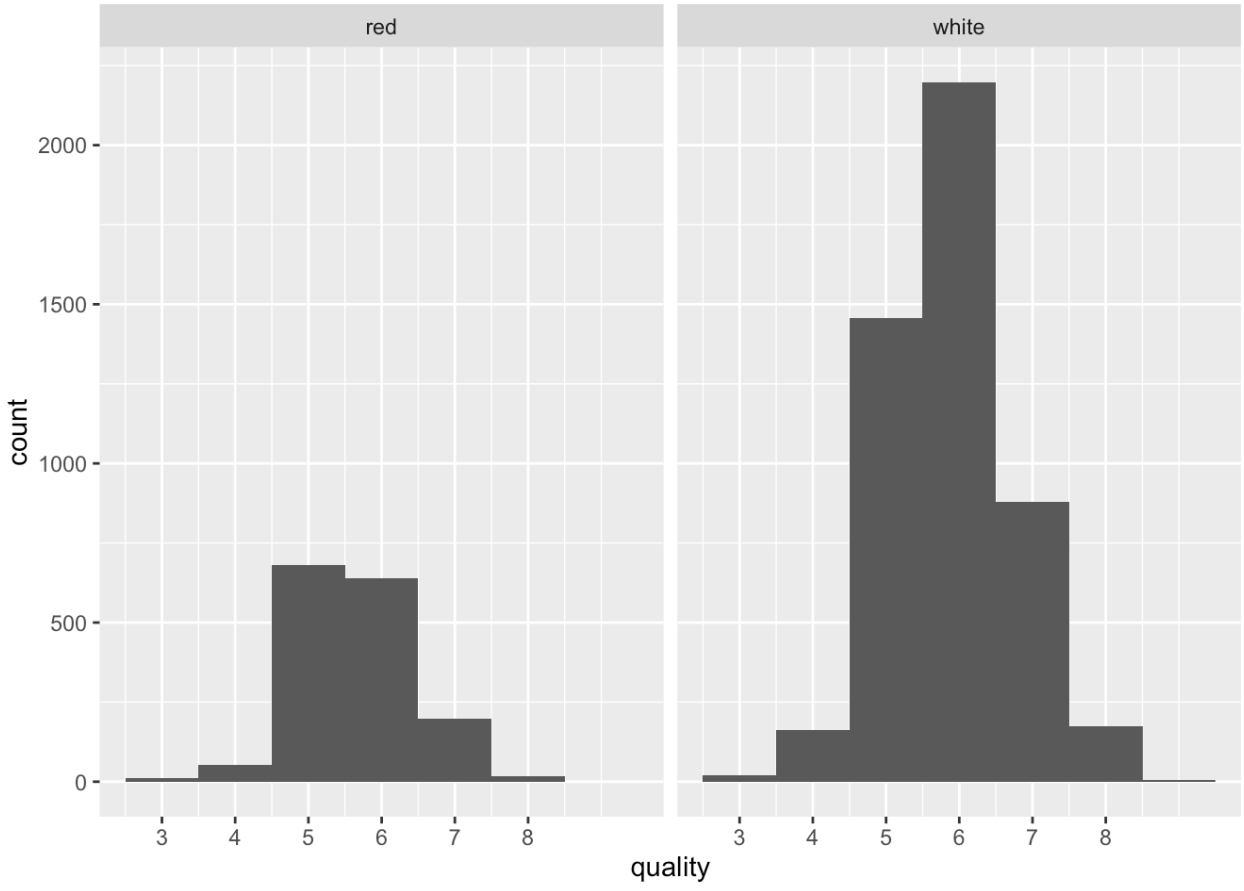
```
## 'data.frame':    6497 obs. of  14 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
##  $ citric.acid         : num  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 0.07
3 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 ...
##  $ color               : Factor w/ 2 levels "red","white": 1 1 1 1 1 1 1 1 1 1 ...
```

```
##        X        fixed.acidity    volatile.acidity  citric.acid
##  Min.   :   1   Min.   : 3.800   Min.   :0.0800   Min.   :0.0000
##  1st Qu.: 813   1st Qu.: 6.400   1st Qu.:0.2300   1st Qu.:0.2500
##  Median :1650   Median : 7.000   Median :0.2900   Median :0.3100
##  Mean   :2044   Mean   : 7.215   Mean   :0.3397   Mean   :0.3186
##  3rd Qu.:3274   3rd Qu.: 7.700   3rd Qu.:0.4000   3rd Qu.:0.3900
##  Max.   :4898   Max.   :15.900   Max.   :1.5800   Max.   :1.6600
##  residual.sugar     chlorides       free.sulfur.dioxide
##  Min.   : 0.600   Min.   :0.00900   Min.   :  1.00
##  1st Qu.: 1.800   1st Qu.:0.03800   1st Qu.: 17.00
##  Median : 3.000   Median :0.04700   Median : 29.00
##  Mean   : 5.443   Mean   :0.05603   Mean   : 30.53
##  3rd Qu.: 8.100   3rd Qu.:0.06500   3rd Qu.: 41.00
##  Max.   :65.800   Max.   :0.61100   Max.   :289.00
##  total.sulfur.dioxide    density             pH          sulphates
##  Min.   :  6.0        Min.   :0.9871   Min.   :2.720   Min.   :0.2200
##  1st Qu.: 77.0        1st Qu.:0.9923   1st Qu.:3.110   1st Qu.:0.4300
##  Median :118.0        Median :0.9949   Median :3.210   Median :0.5100
##  Mean   :115.7        Mean   :0.9947   Mean   :3.219   Mean   :0.5313
##  3rd Qu.:156.0        3rd Qu.:0.9970   3rd Qu.:3.320   3rd Qu.:0.6000
##  Max.   :440.0        Max.   :1.0390   Max.   :4.010   Max.   :2.0000
##     alcohol         quality        color
##  Min.   : 8.00   Min.   :3.000   red  :1599
##  1st Qu.: 9.50   1st Qu.:5.000   white:4898
##  Median :10.30   Median :6.000
##  Mean   :10.49   Mean   :5.818
##  3rd Qu.:11.30   3rd Qu.:6.000
##  Max.   :14.90   Max.   :9.000
```

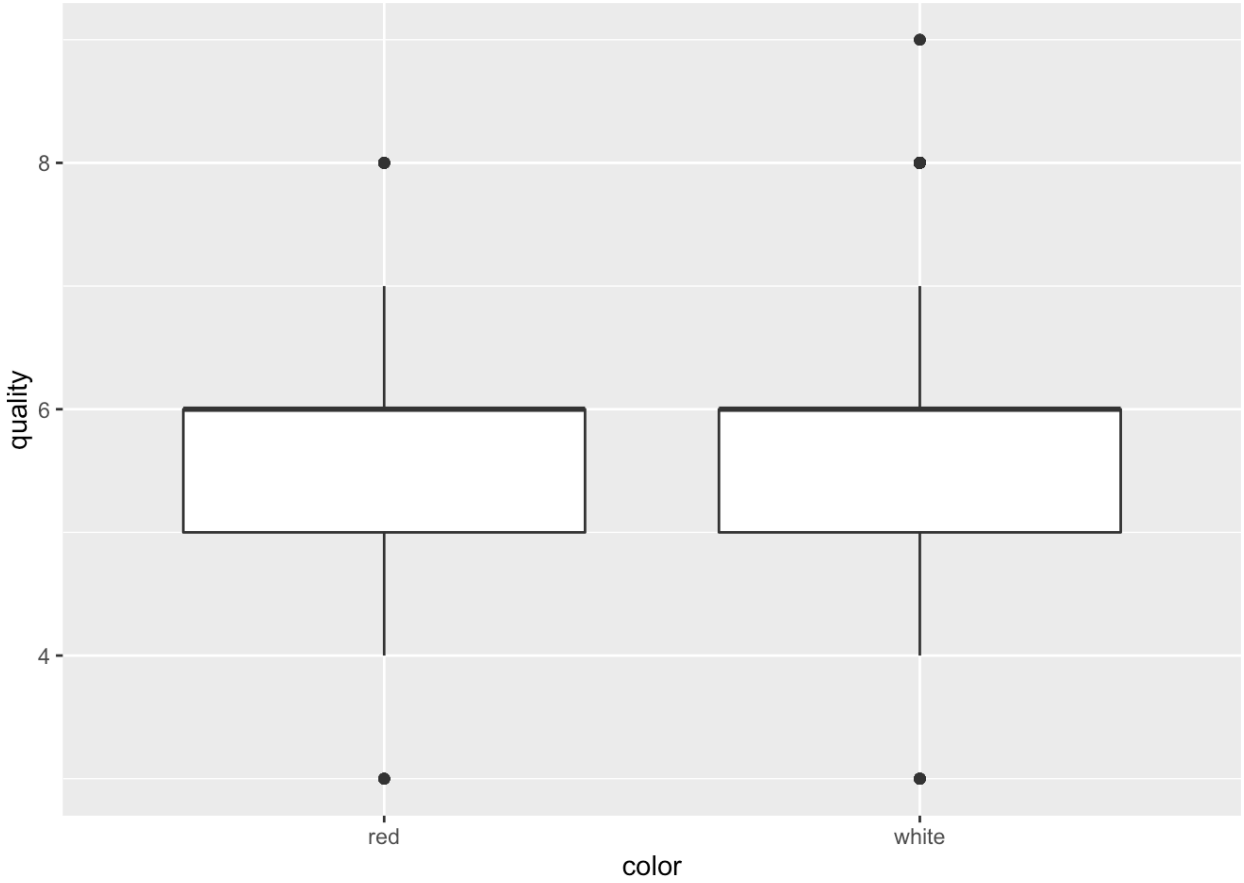
The combined dataset of red and white wine data contains 14 variables and 6497 observations.

Univariate Plots Section

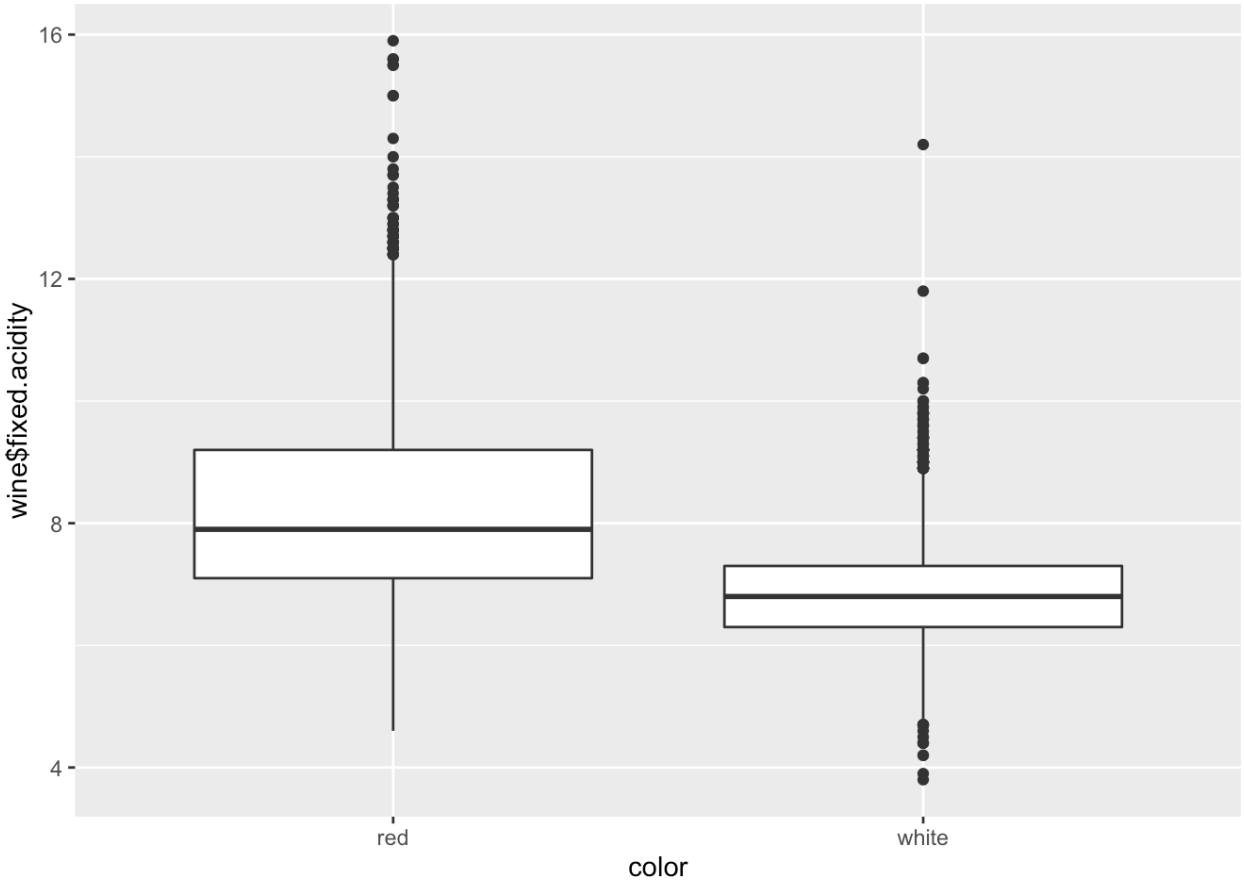


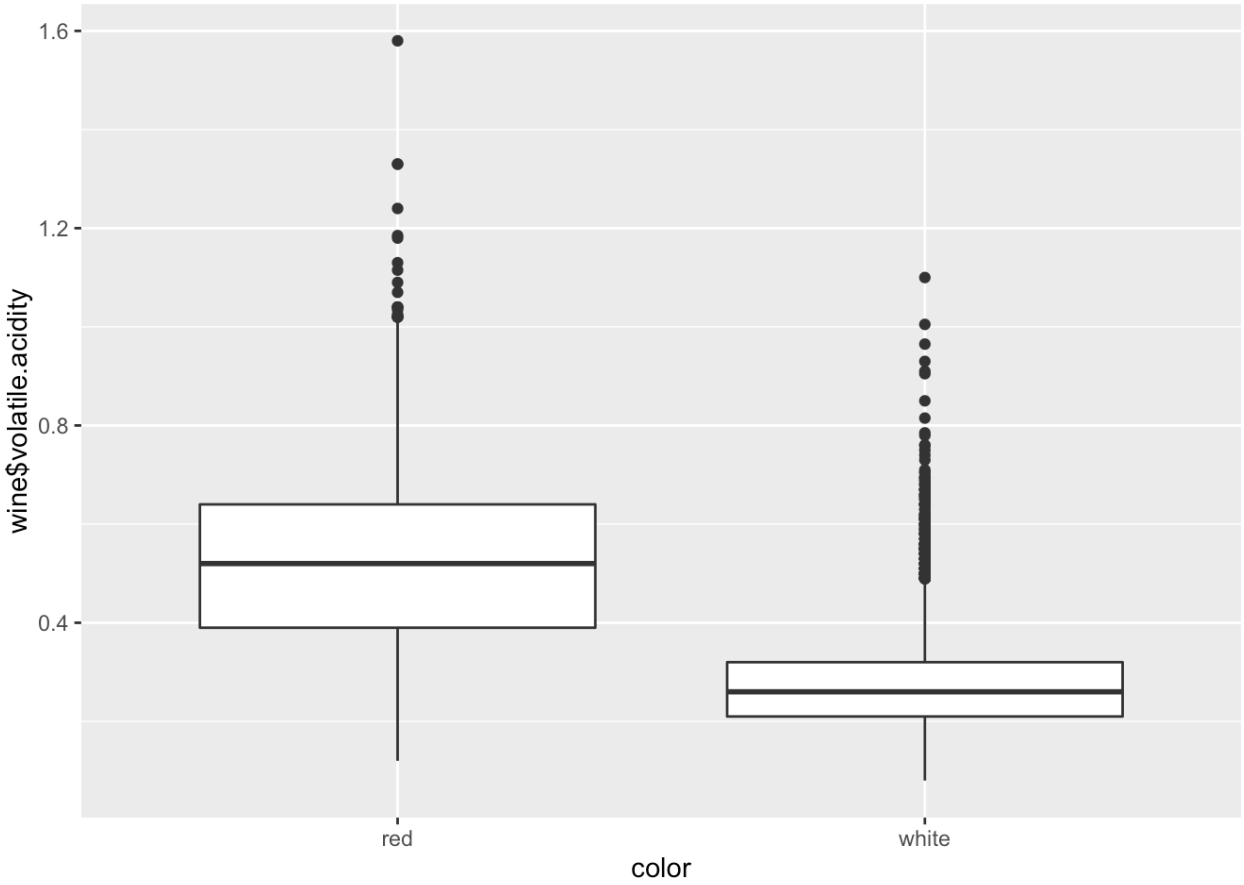


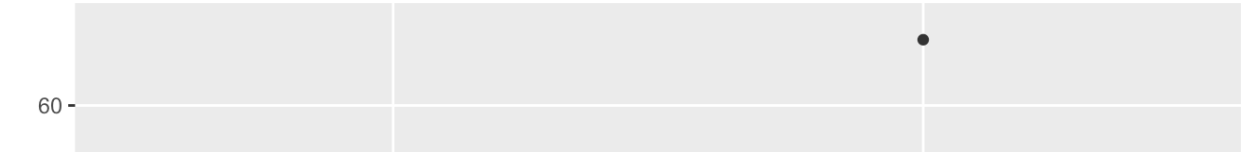
The distribution of wine ratings for both red and white wine appears to follow a normal distribution. The majority of wines receive a quality score of around 5 or 6. We can see from the above histogram that the amount of white wine data is more than double the amount of red.

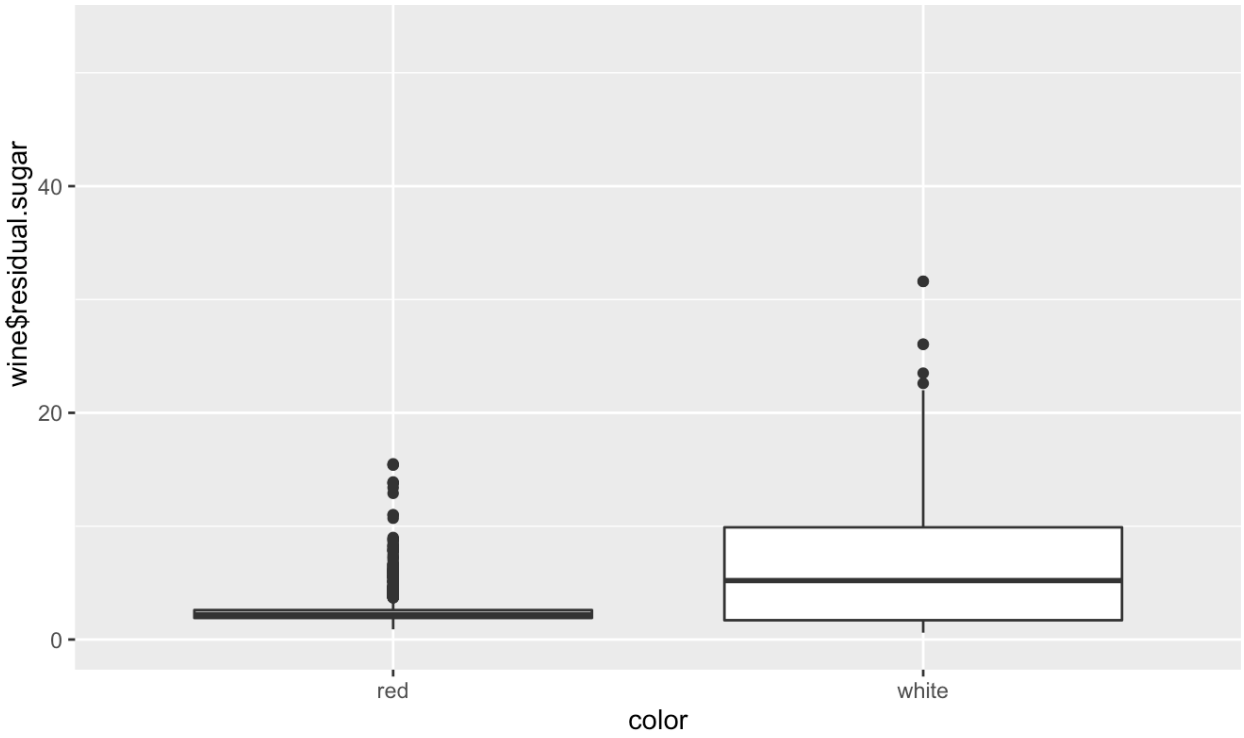


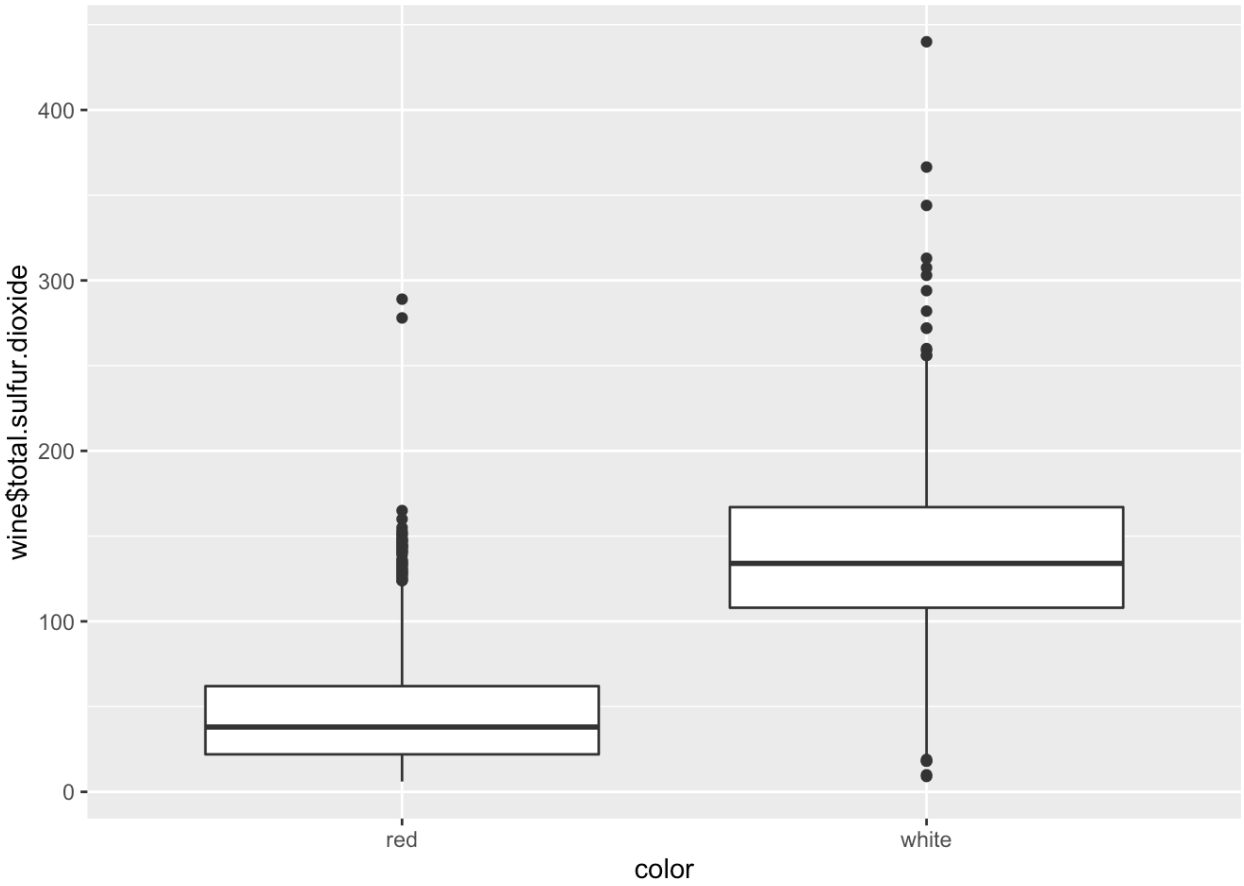
Displaying the quality of wines as a boxplot shows that both wines have a median rating of 6, and most wines are rated as either 5 or 6.

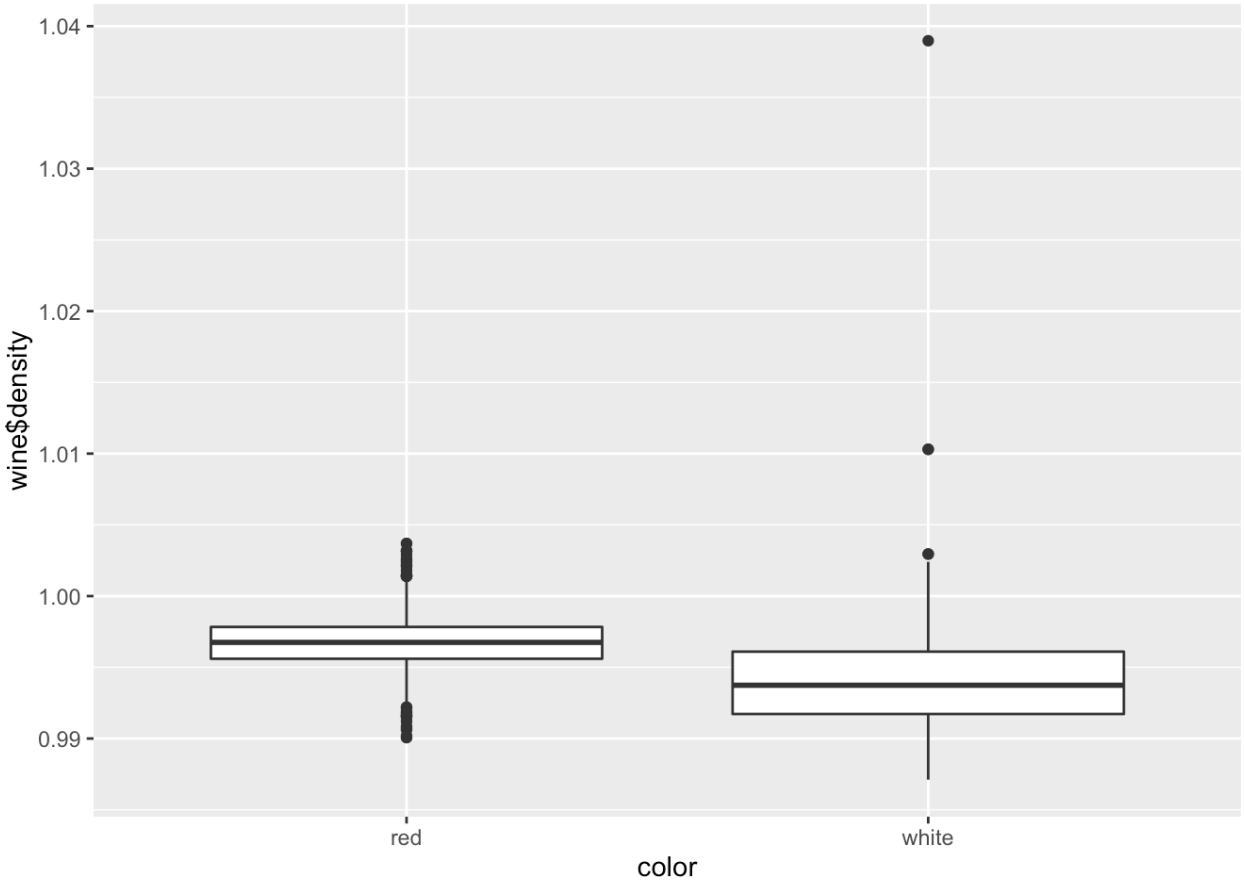

16
12
8
4
wine$fixed.acidity
red
white
color

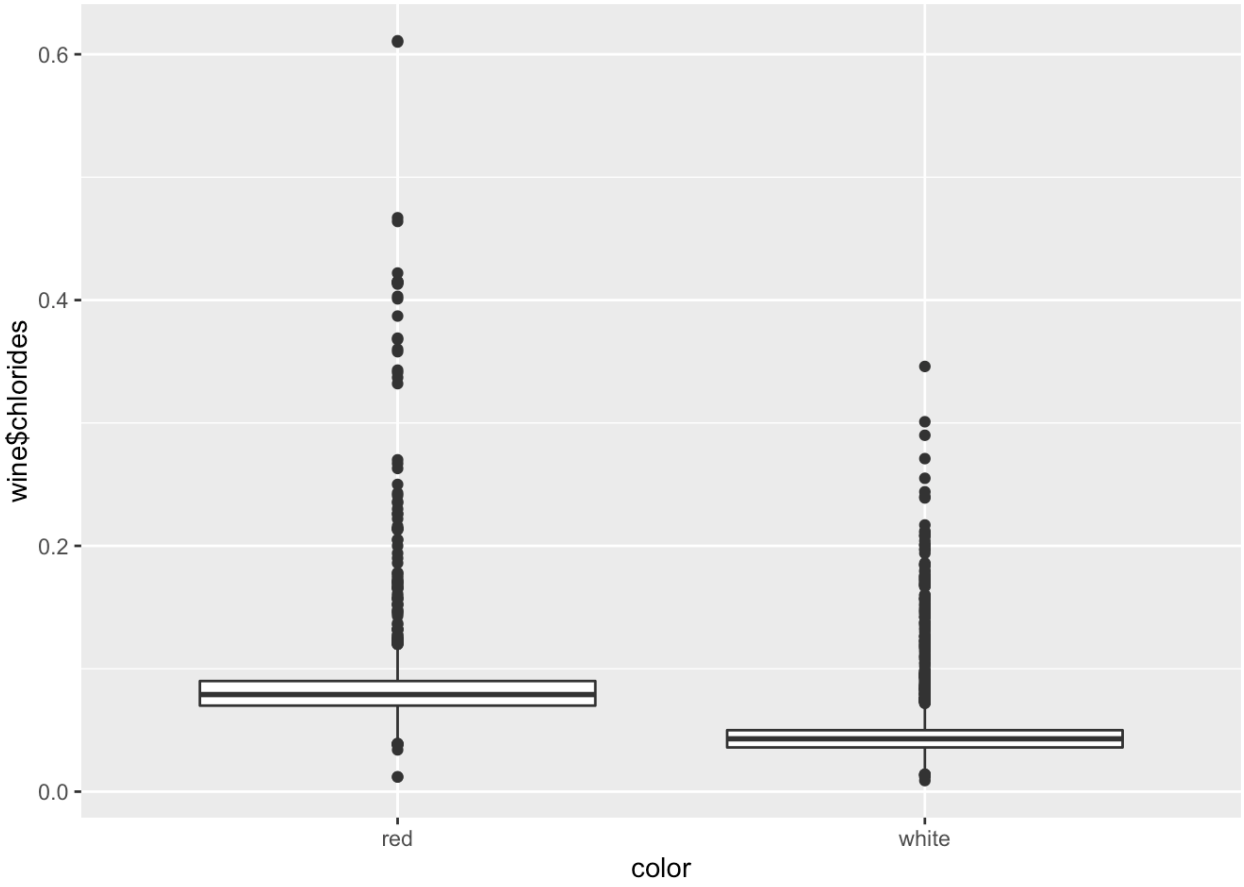

1.6
1.2
0.8
0.4
wine$volatile.acidity
red
white
color

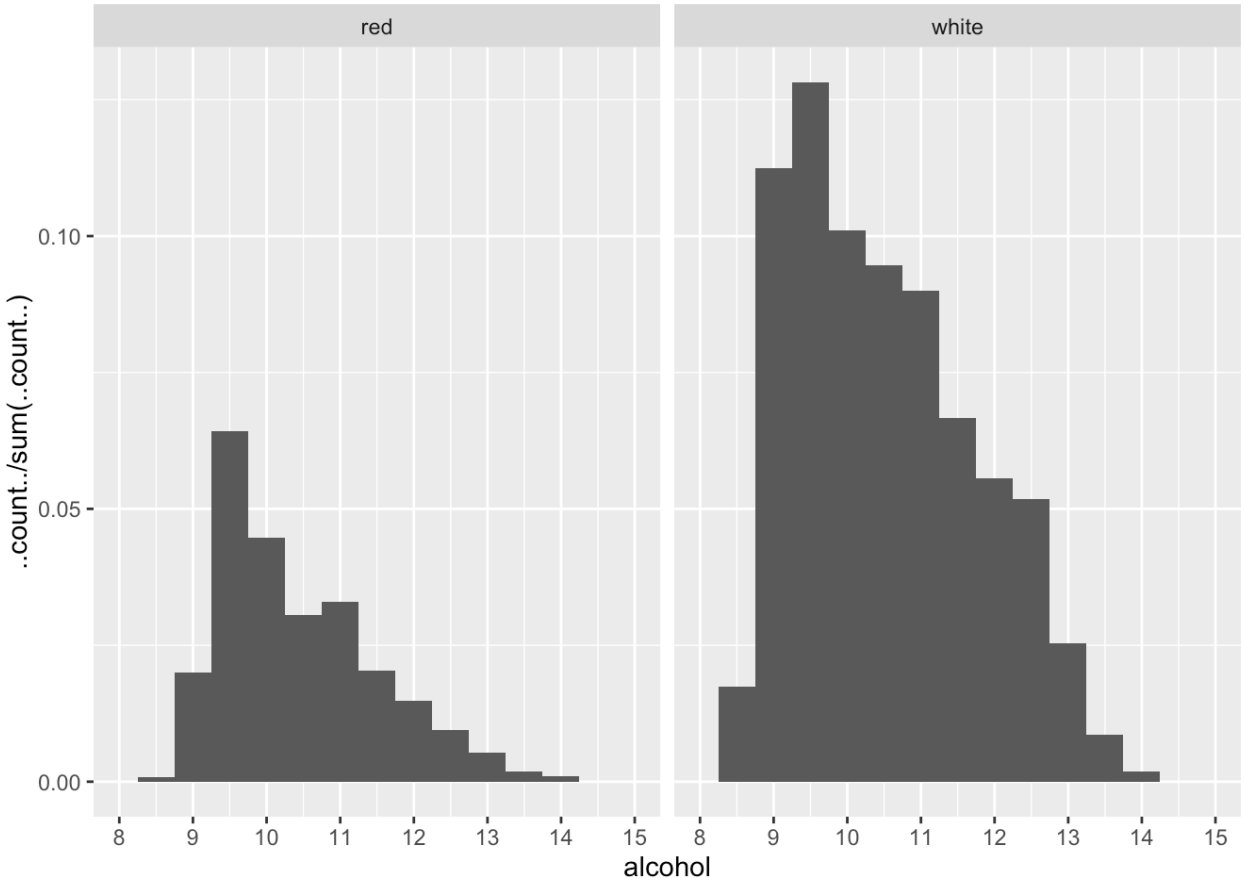

60

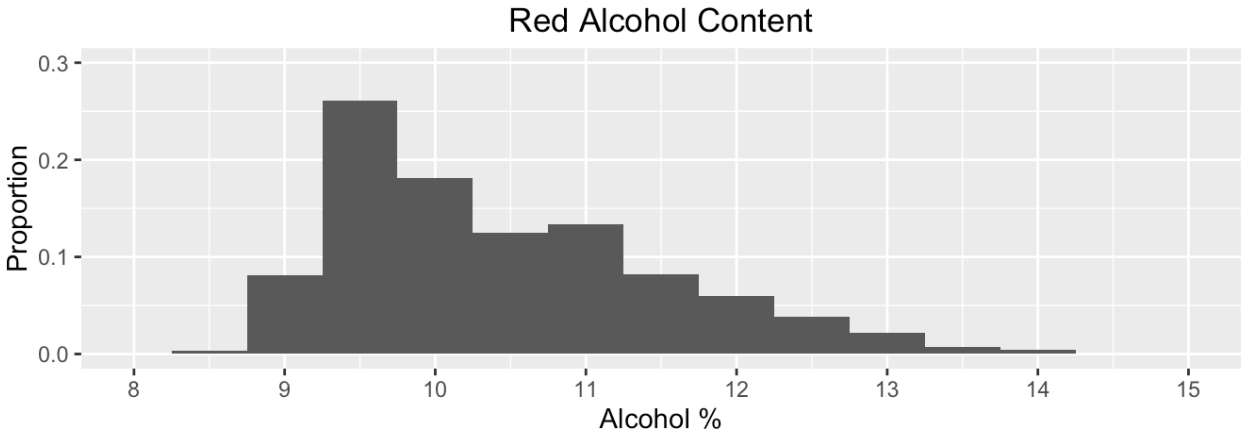

wine$residual.sugar
40
20
0
red
white
color

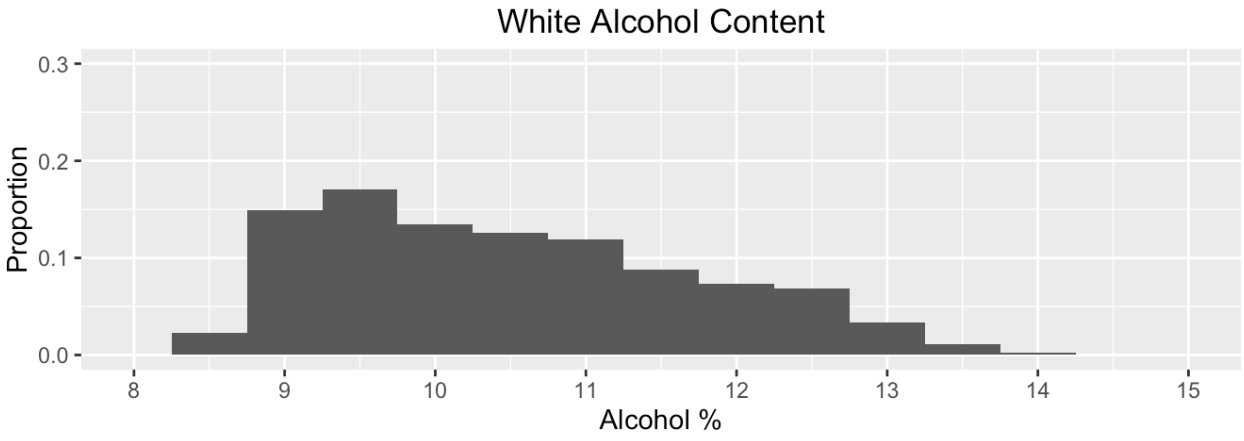

wine$total.sulfur.dioxide
400
300
200
100
0
red
white
color


1.04
1.03
1.02
1.01
1.00
0.99
wine$density
red
white
color


0.6
0.4
0.2
0.0
wine$chlorides
red
white
color







By comparing the distribution of red and white wines across several variables, we can see that there are many instances where the characteristics of these wines differ. For example, on average, red wine is more acidic and contains less sugar than white wine. For all wines included in this dataset, the majority have an alcohol

level of 9.5%.

Univariate Analysis

What is the structure of your dataset?

There are two datasets that are being analyzed for this project. One corresponds to the data for white wines, the other for red wines. There are 12 variables that describe these wines; fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, sulfur dioxide, density, PH, sulphates, alcohol, and quality. Since these variables are identical across both dataset for red and white wines, I was able to combine them together and add an additional variable to differentiate between the type of wine. The red wine dataset contained 1599 observations and the white wine contained 4898, making a combined dataset of 6497 wines.

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

The main feature of interest for the dataset is how the quality of the wine, based on a score between 0 - 10, is affected by the other variables. Additionally, I would like to determine what characteristics differentiate a red wine from a white wine.

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

I think that all the variables (fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, sulfur dioxide, density, PH, and sulphates) will have an effect on how a wine tastes. However, the ideal level of these variables for red and white wines are likely to be different when it comes to rating wine.

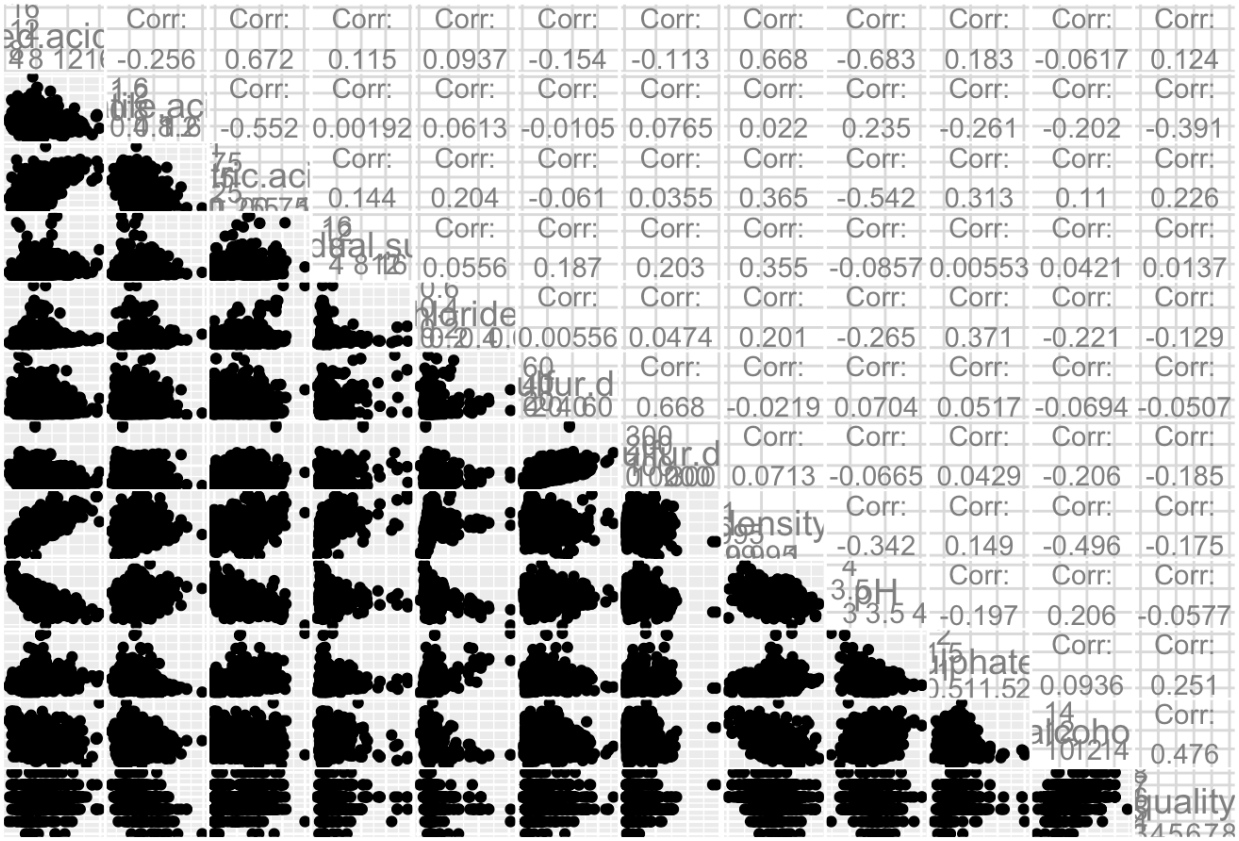
Did you create any new variables from existing variables in the dataset?

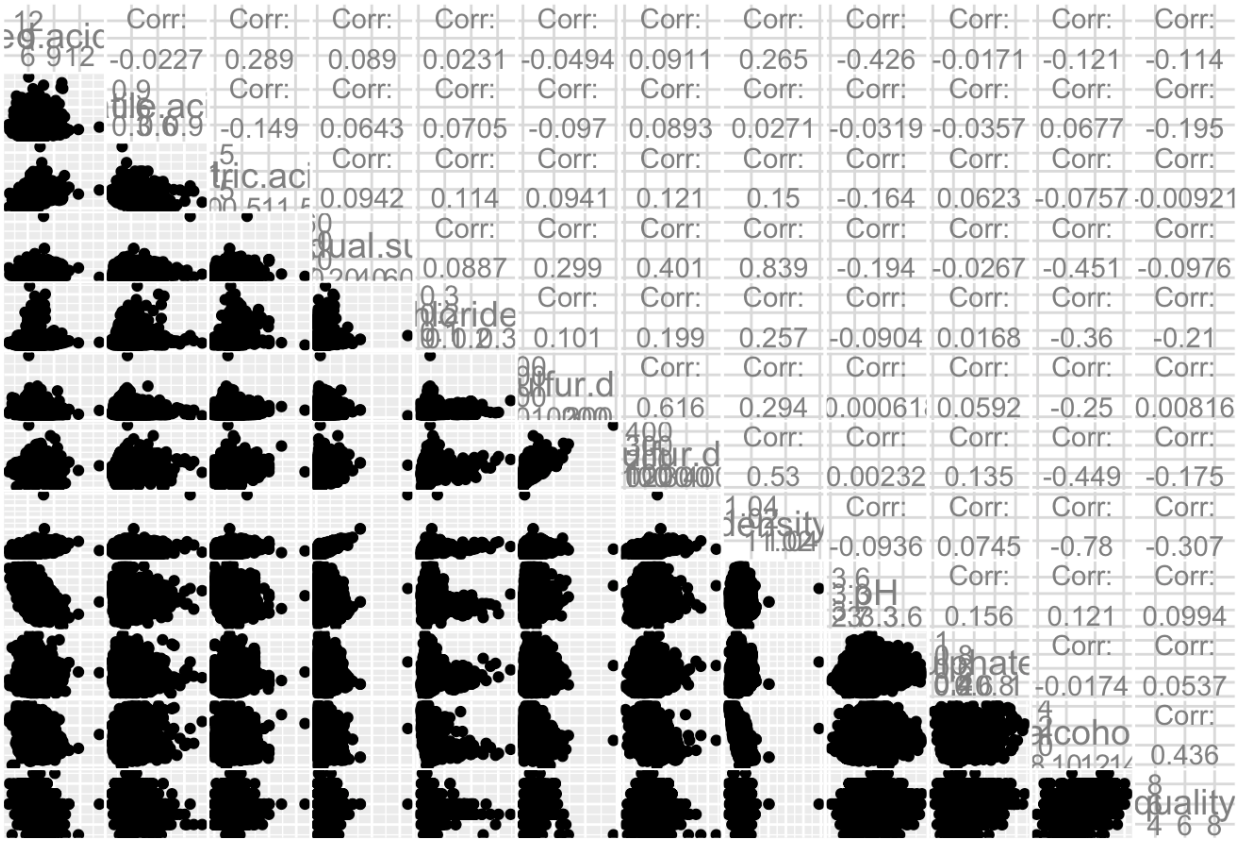
Yes, by joining the two datasets on red and white wines together, I created a new variable called “color”. This indicated whether the wine in the new dataset is a red or white wine.

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?

Since there are an unequal amount of data when comparing the red and white datasets, I decided to display most of my visualizations in this section as boxplots. I think that this method allows us to better alalyze the data as it displays the data in relative instead of absolute values. Using histograms of overall count are more difficult to visually compare since the white wine dataset is more than double the amount of red wine dataset.

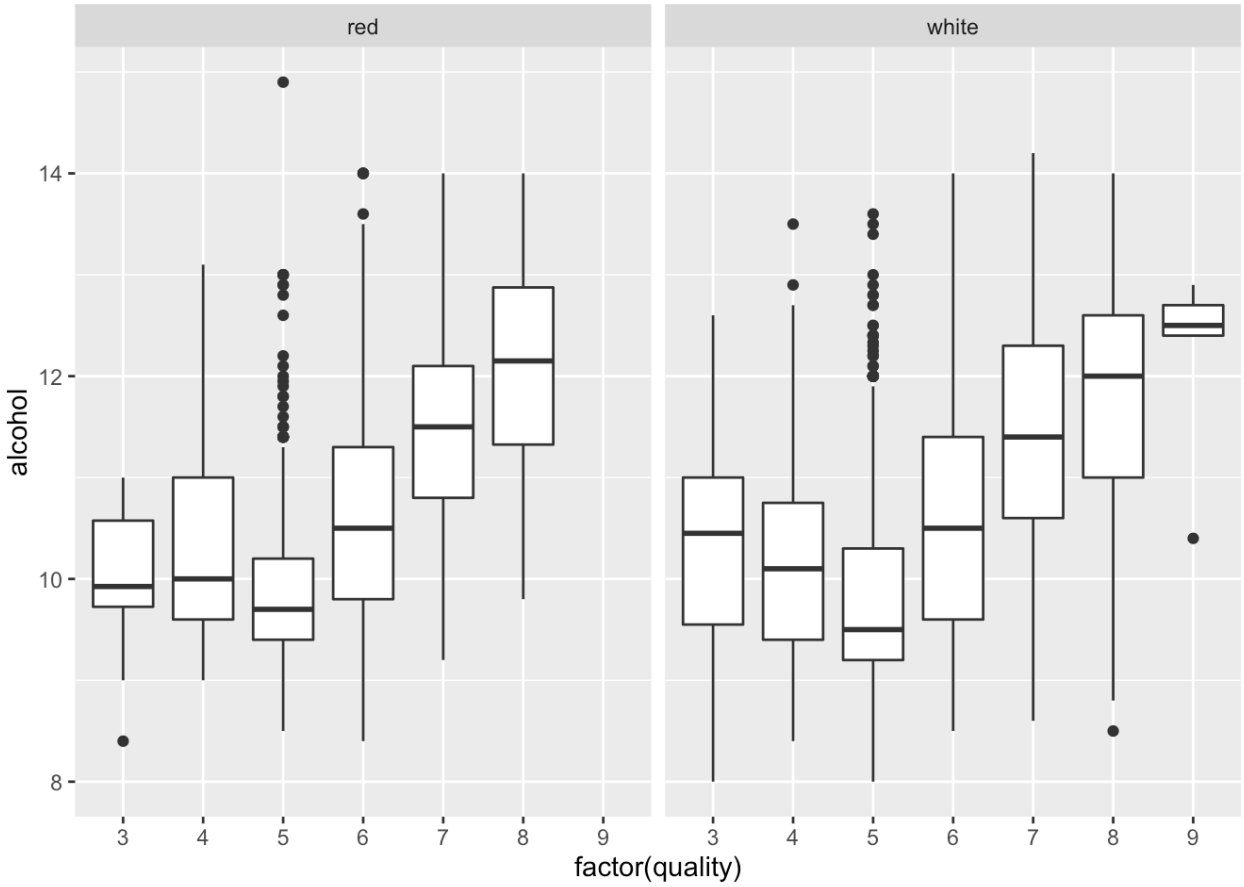
Bivariate Plots Section

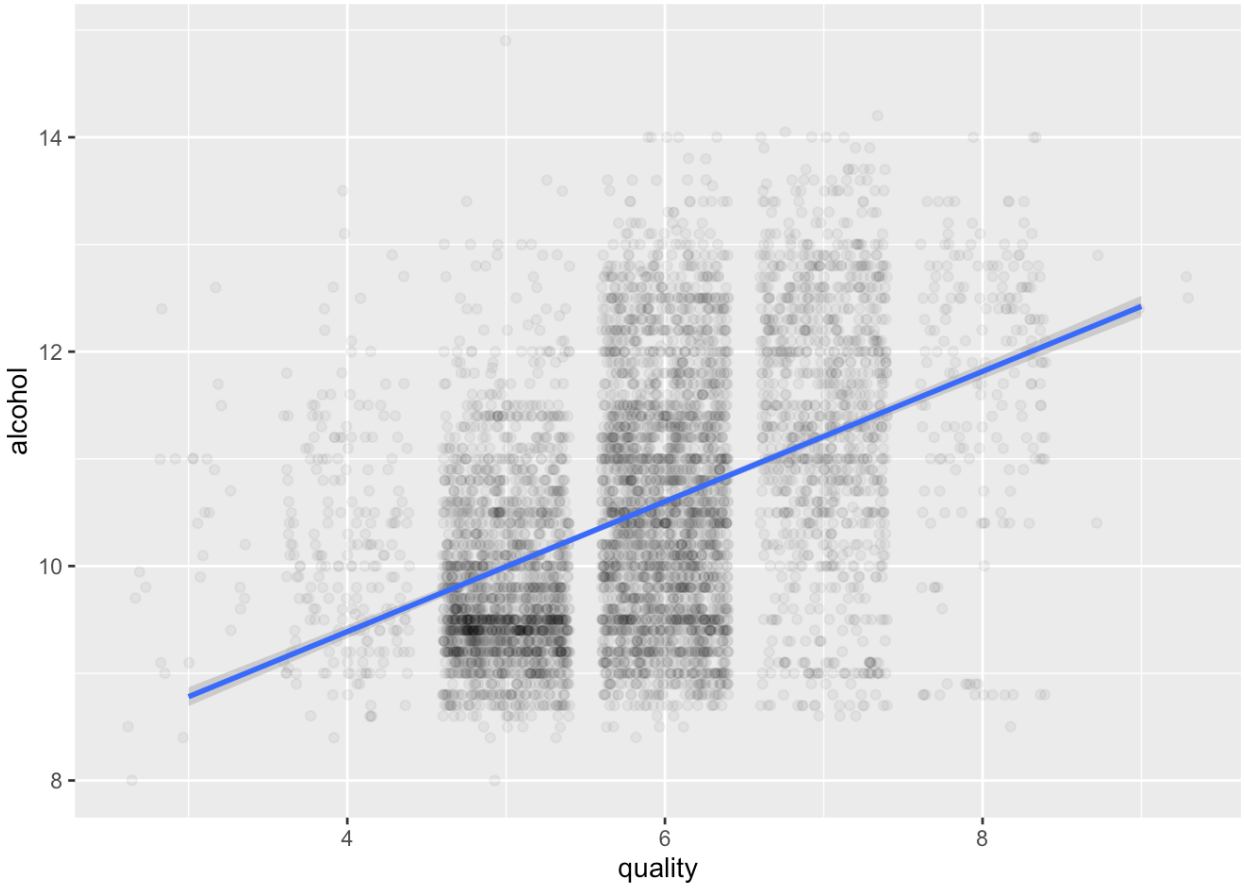


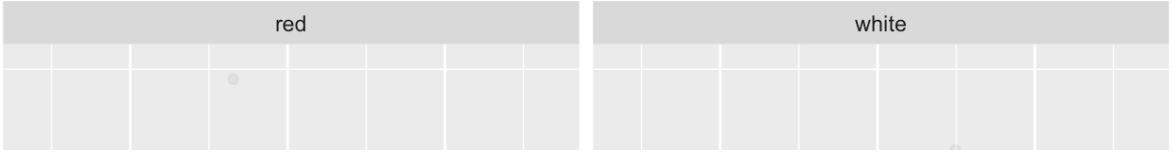


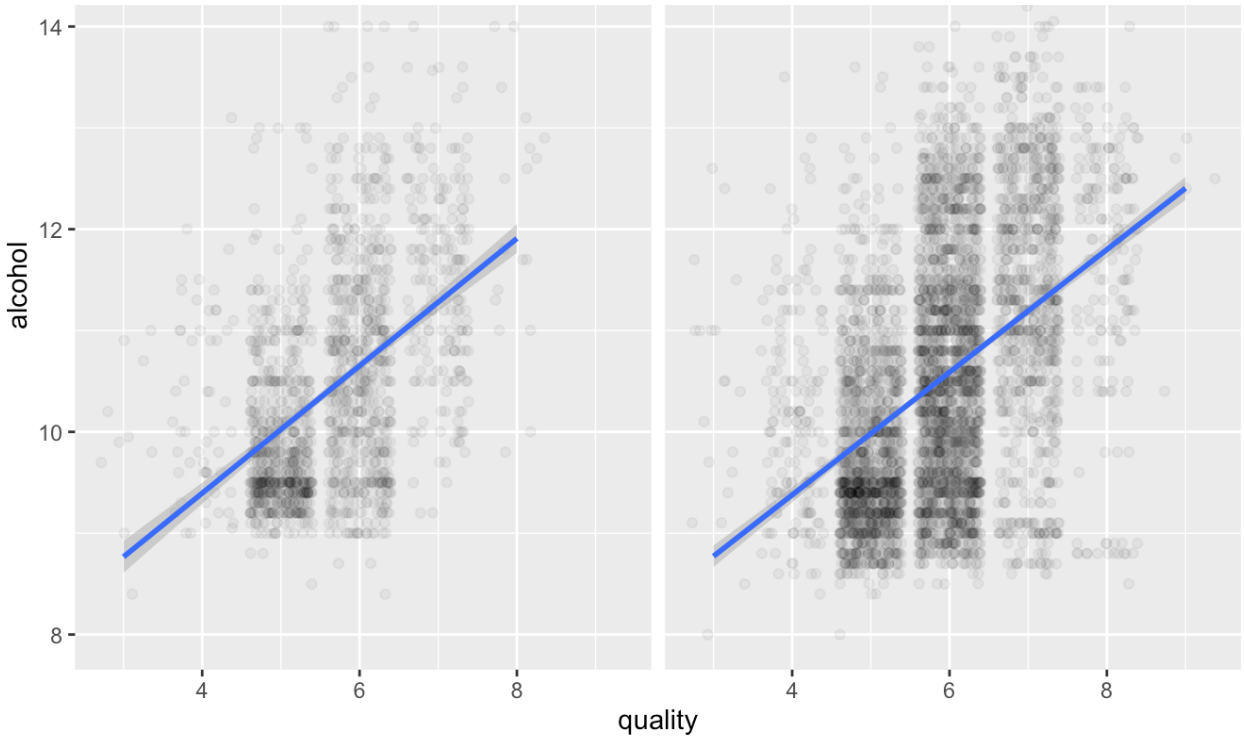
For both red and white wines, the variables that most strongly correlates with quality is alcohol. Density also seems to have a negative correlation with the quality of wines.

Additionally, volatile acidity seems to correlate negatively with red wine quality, but the same is not true for white wines.

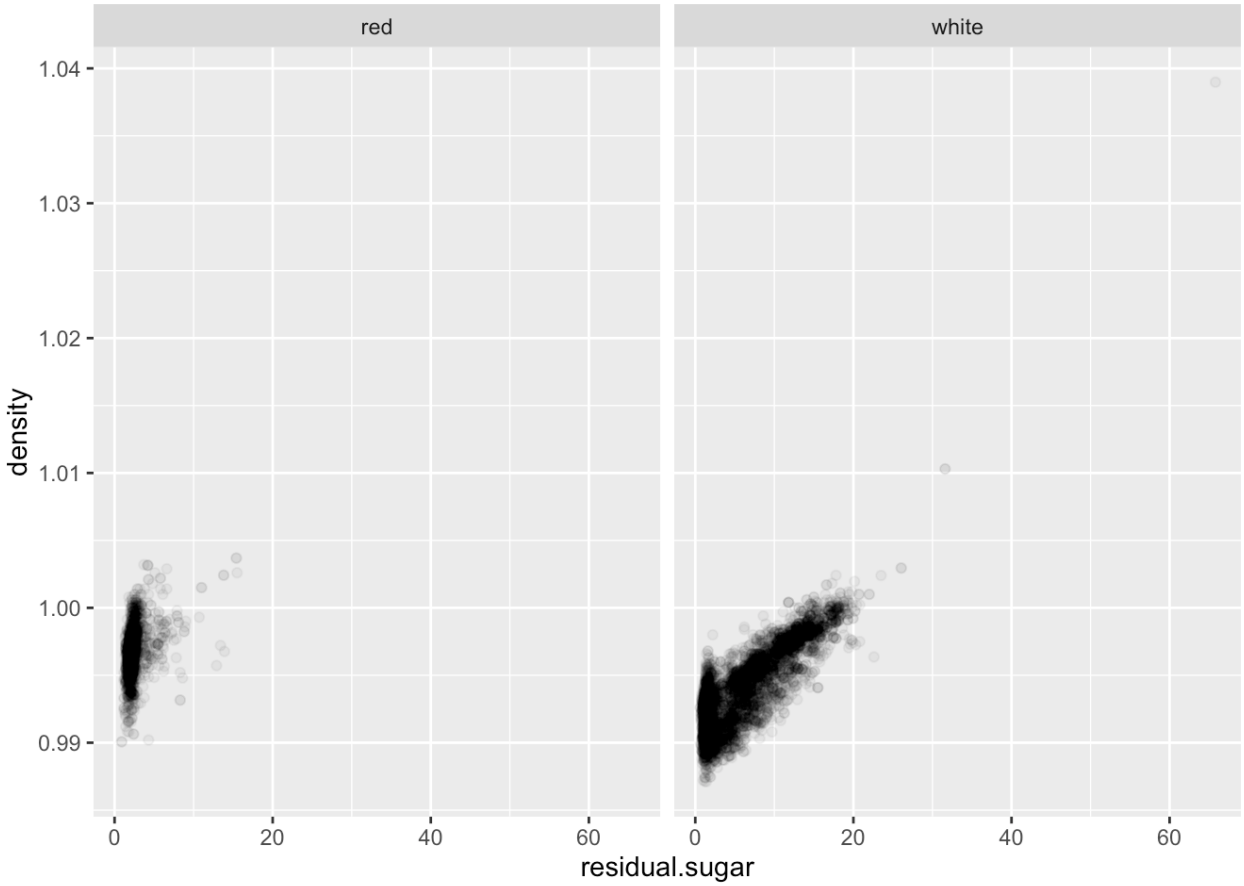

red
white
alcohol
14
12
10
8
3
4
5
6
7
8
9
factor(quality)

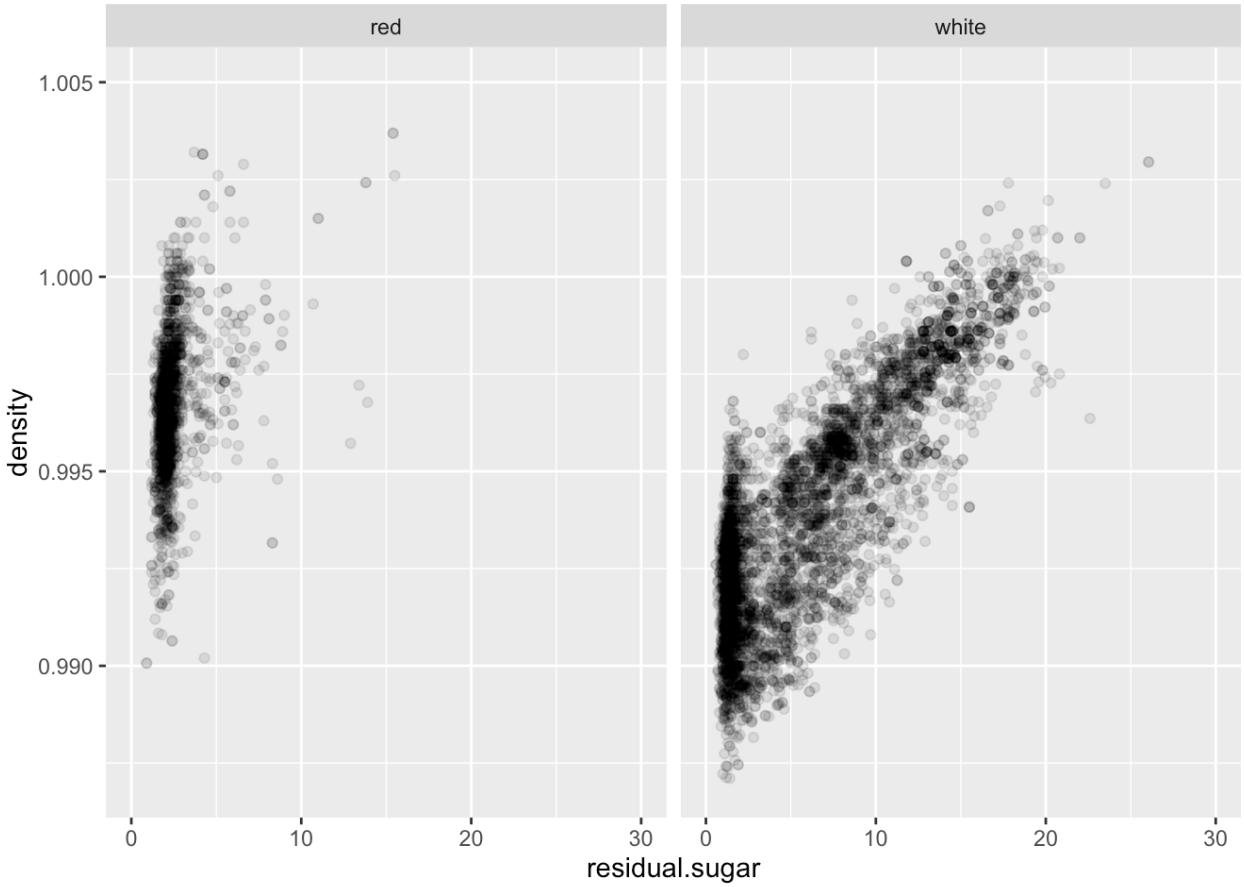

alcohol
14
12
10
8
4
6
8
quality

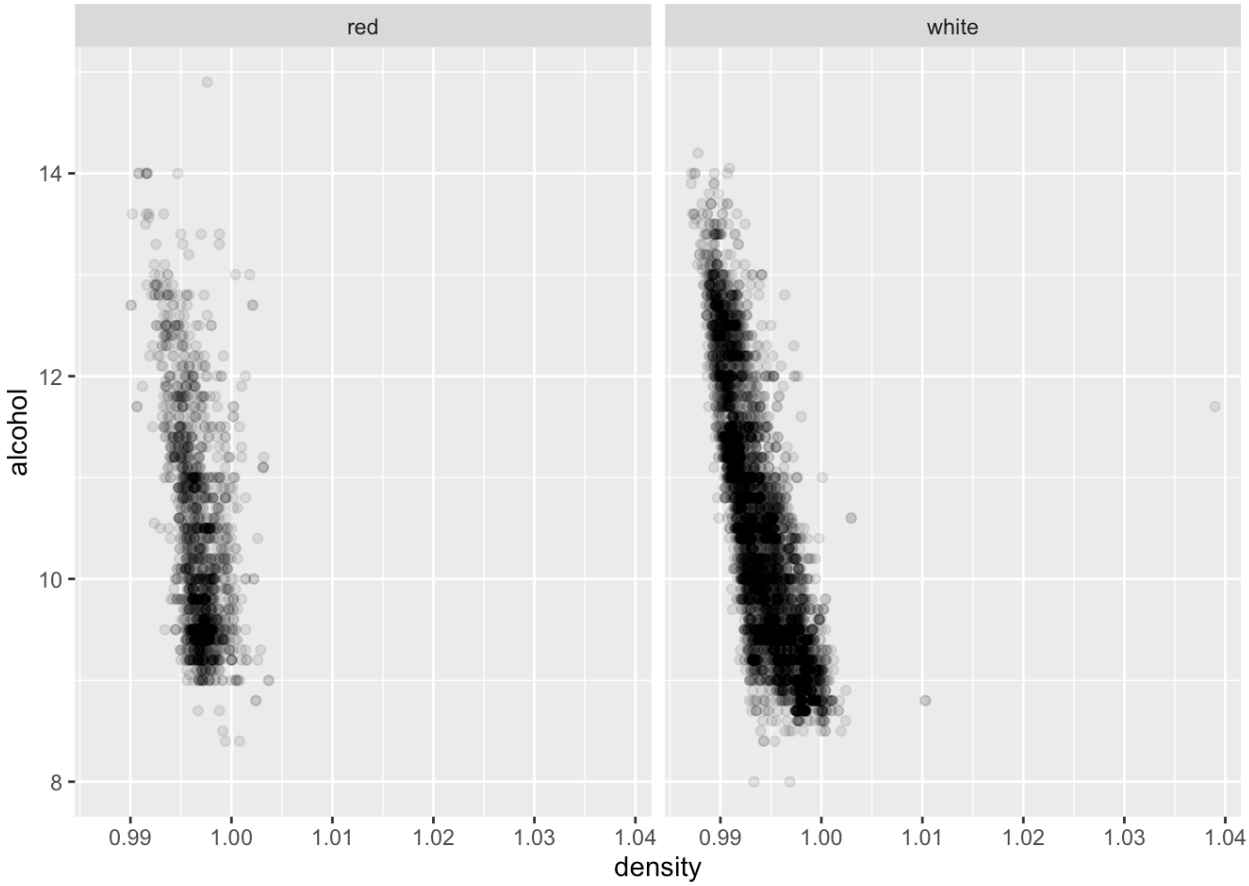

red
white

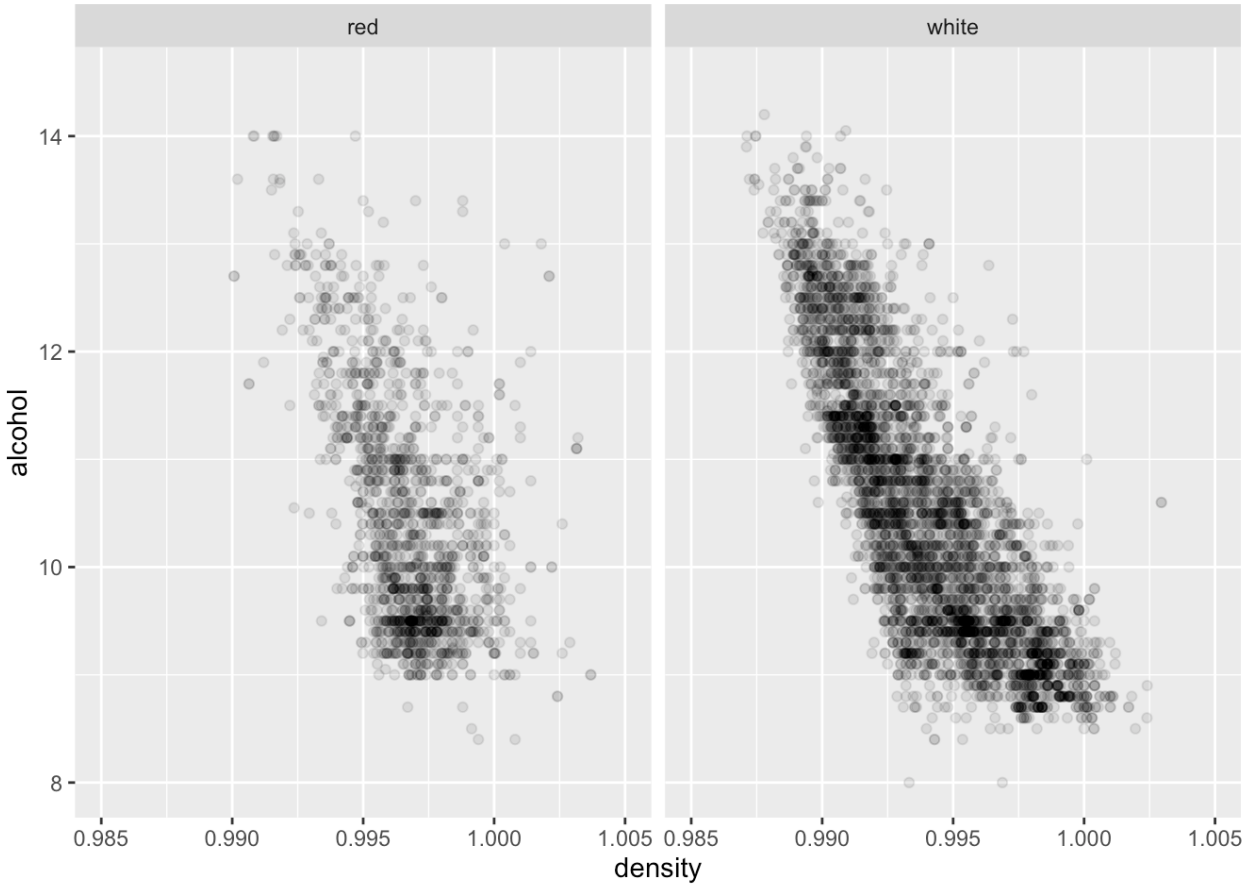


There seems to be a positive relationship between alcohol content and quality of wine. However, this relationship is only apparent for wines that are rated above a level of 5.

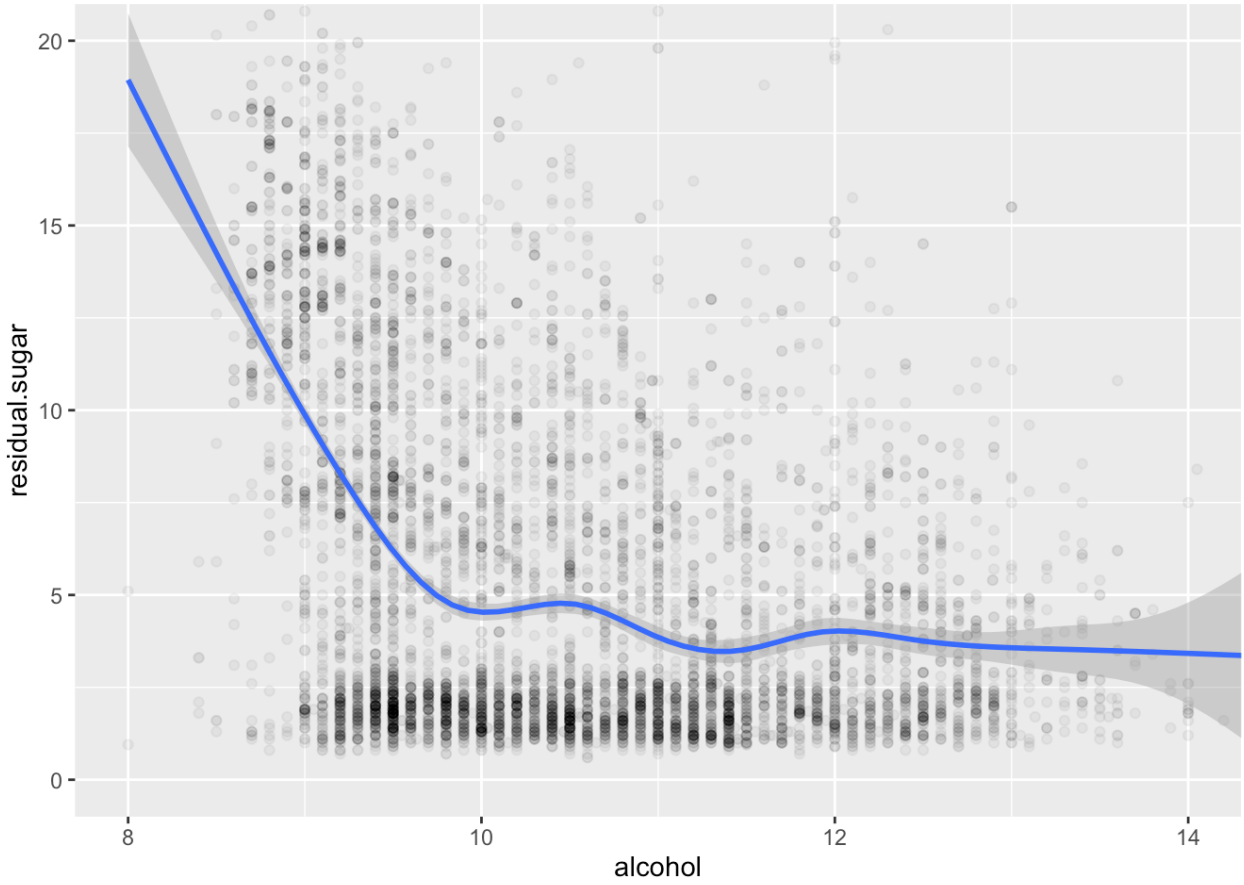

red
white
density
1.04
1.03
1.02
1.01
1.00
0.99
0
20
40
60
residual.sugar

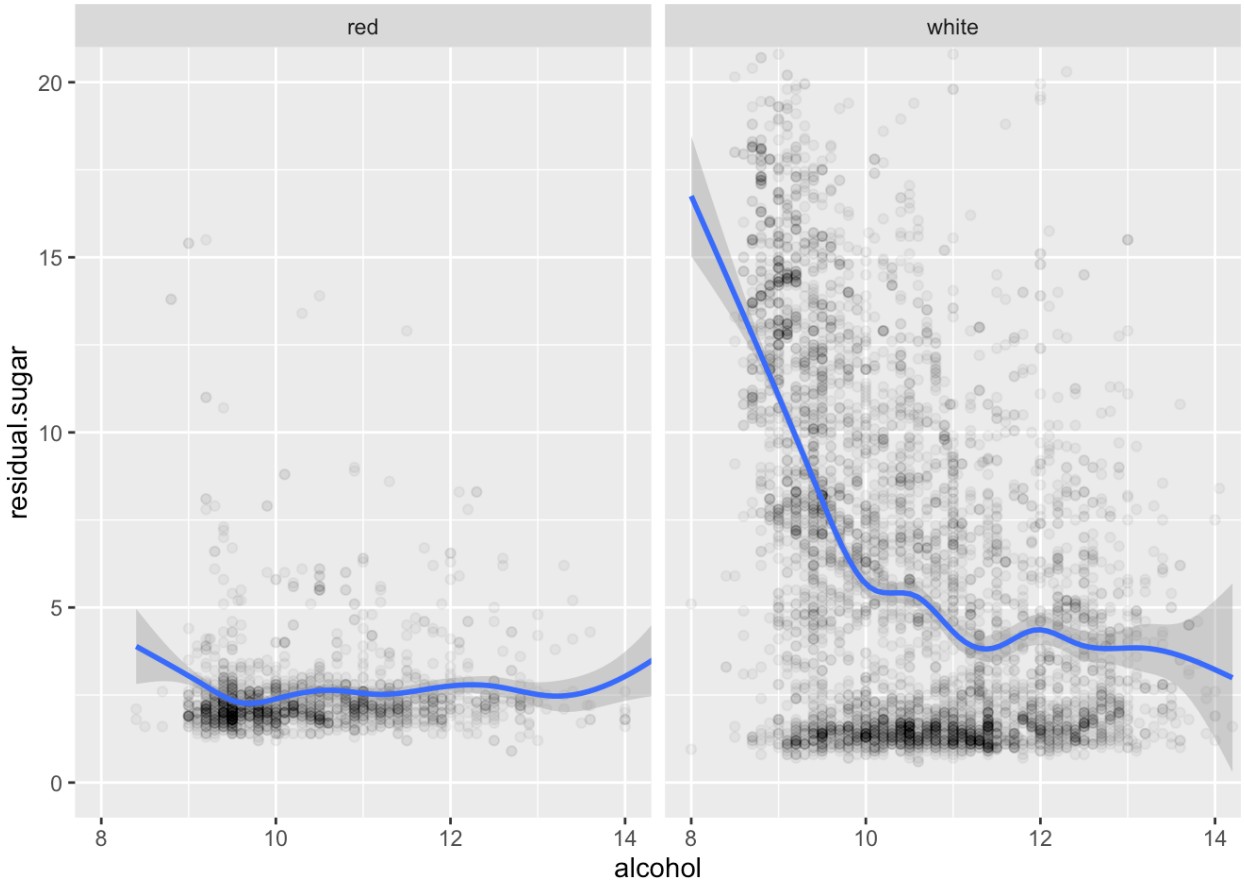

red
white
density
1.005
1.000
0.995
0.990
0
10
20
30
residual.sugar





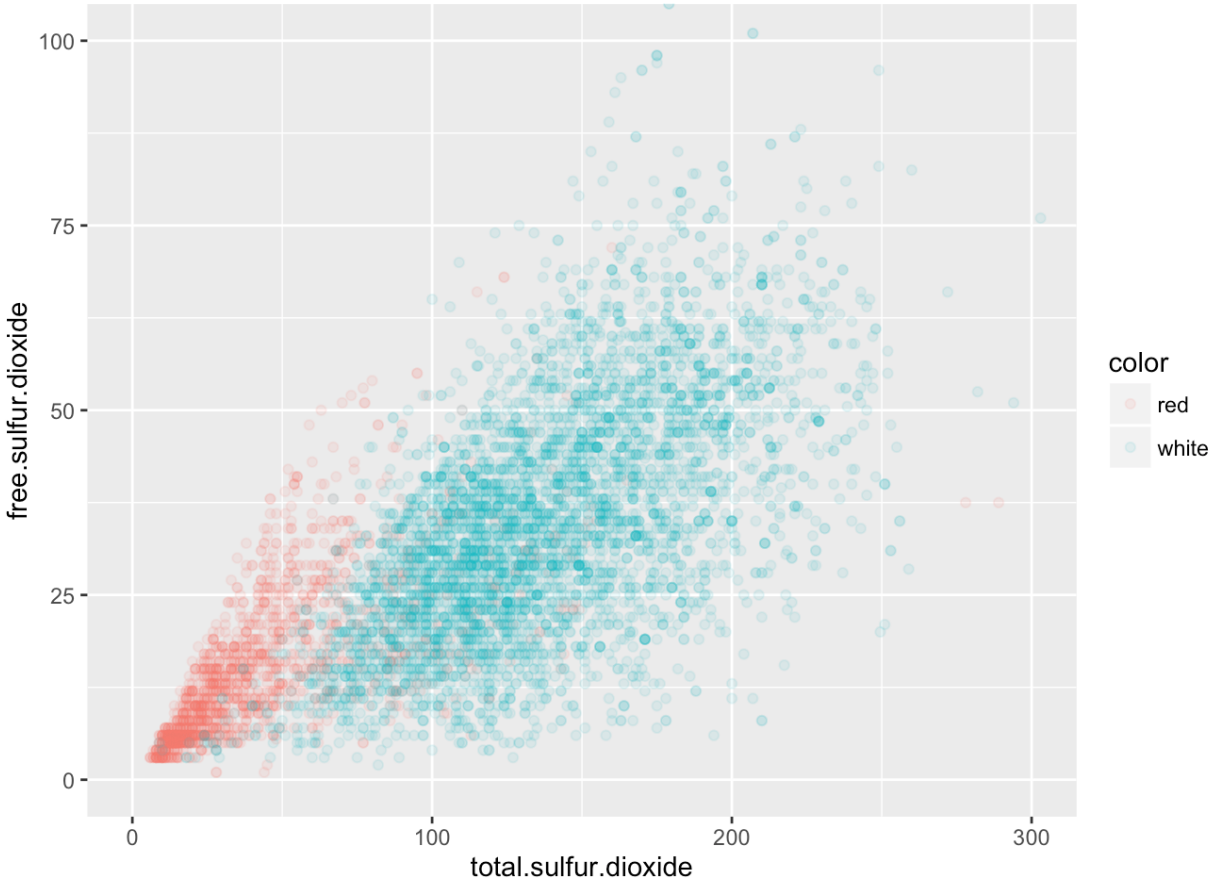
The density of wine is affected by the alcohol level and sugar content. Higher sugar levels will increase the density, while alcohol decreases density.

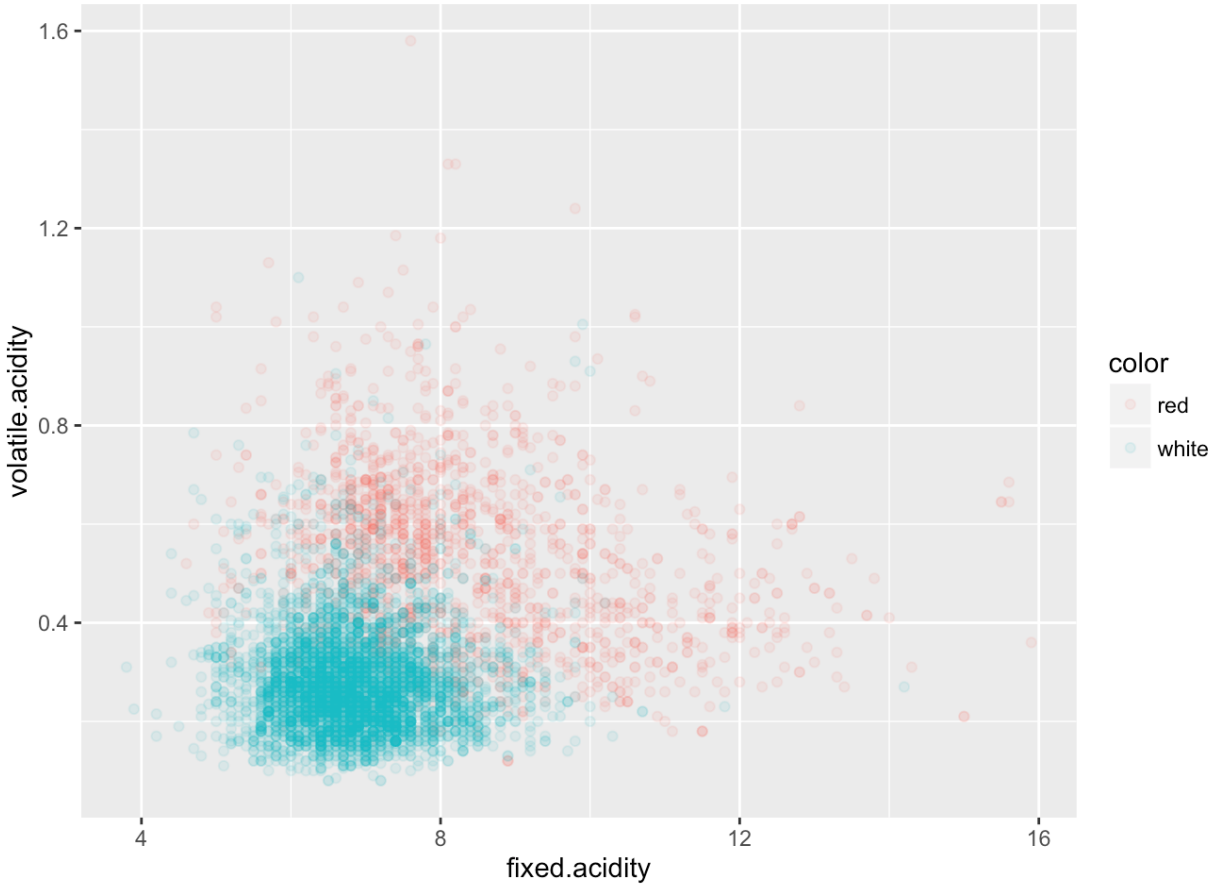




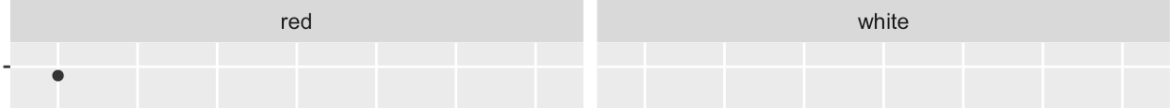
During wine fermentation, yeast turns the sugar from grape juice into alcohol. I expected there to be a negative relationship between residual sugar and alcohol because residual sugar usually refers to the sugar remaining after fermentation stops. It is interesting to note that this negative relationship is present for the

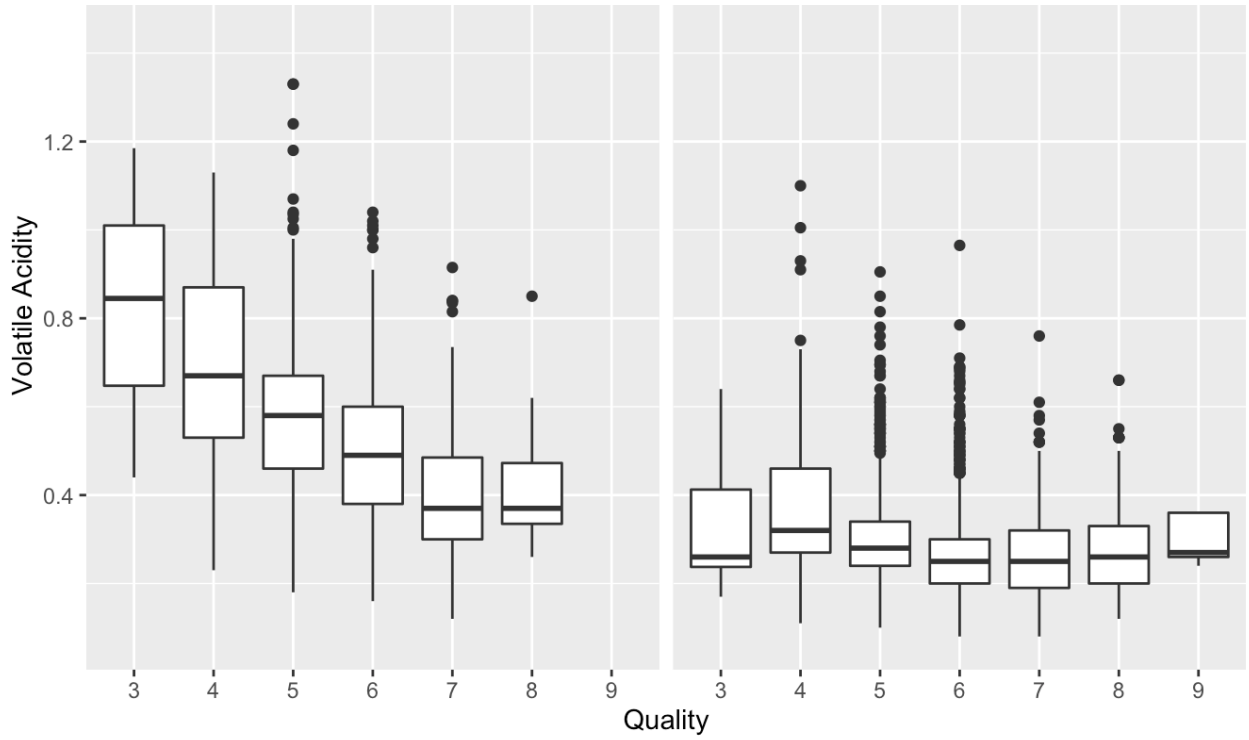
white wine, but not for the red.





Volatile Acidity and Wine Quality





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?

The quality of wines most strongly correlates with higher levels of alcohol. Density also has a negative correlation with wine quality, however; we saw during our analysis of alcohol and density that there is also a negative relationship between these two variables. Since a lower density is the result of higher alcohol levels, and higher alcohol contents are positively correlated with wine quality, it would be interesting to investigate how this correlation is shared between these two variables.

Residual sugar also has an effect on the alcohol content of wine, as well as the density of the wine. This is due to the fermentation process that turns sugars into alcohol. After investigating the dataset, I noticed that as residual sugar decreases in white wines, alcohol content increases. The same relationship is not true regarding red wine.

Other relationships that I looked at include total sulfur dioxide vs free sulfur dioxide, and fixed acidity vs volatile acidity. It makes sense that there is a linear relationship between free sulfur dioxide and total sulfur dioxide, as the total includes the free sulfur dioxide amounts. The visualizations show that red wines are more acidic and contain less sulfur dioxide than white wines. There is also a negative relationship between volatile acidity and the quality of red wines.

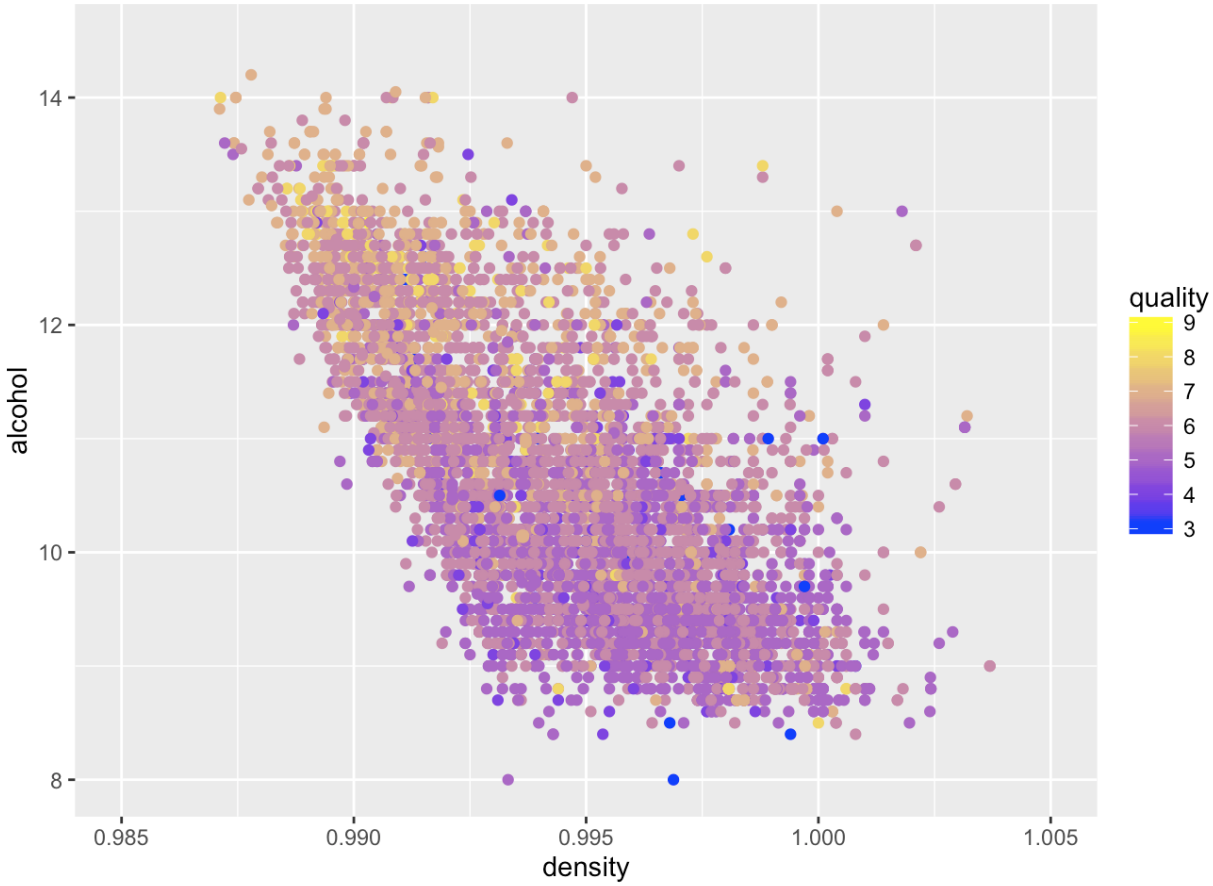
Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

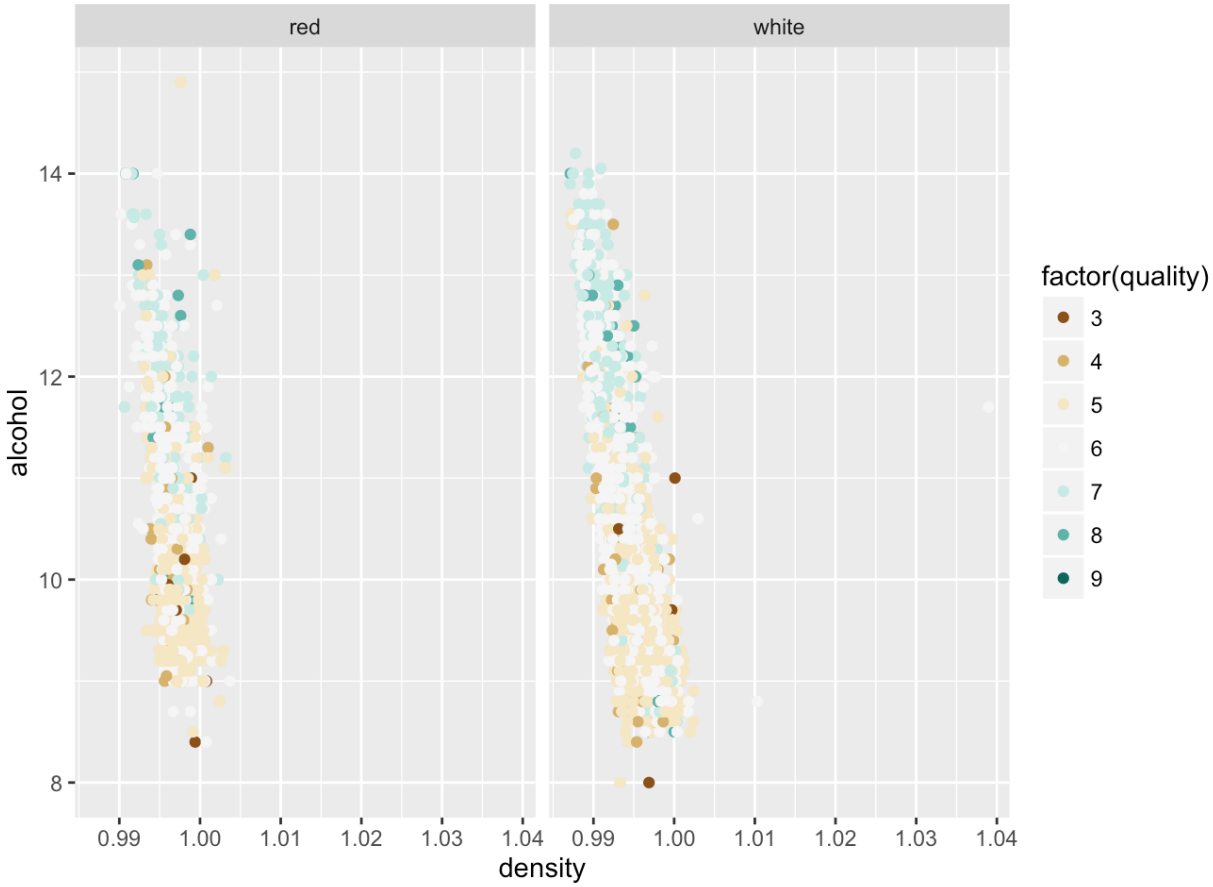
I noticed that density, residual sugar, and alcohol all relate to one another. After looking into this relationship between these three variables, I noticed that density depends on the level of sugar and alcohol contents of the wine.

What was the strongest relationship you found?

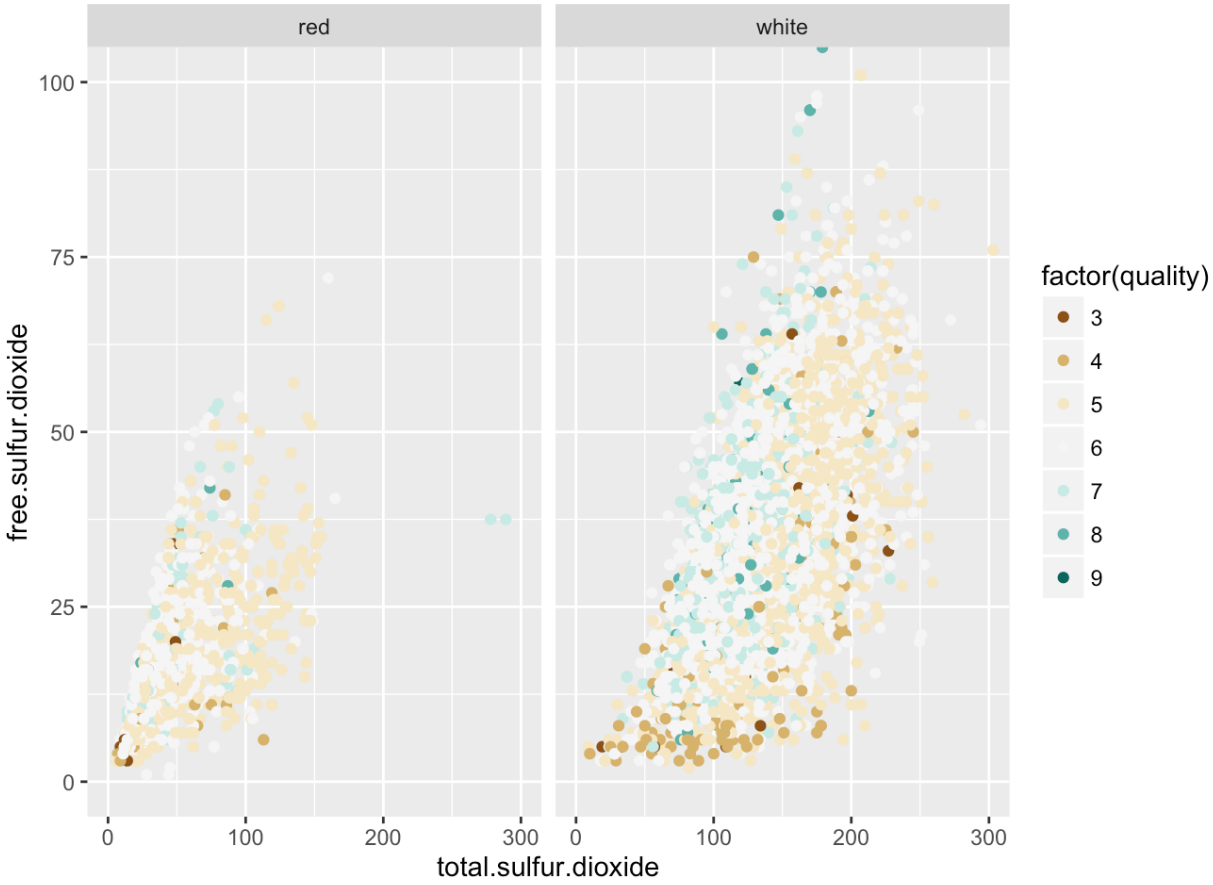
The strongest relationship I found with respect to the quality of both red and white wines was the alcohol content. Alcohol and quality has a 0.476 correlation among red wine and 0.436 for white wine.

Multivariate Plots Section

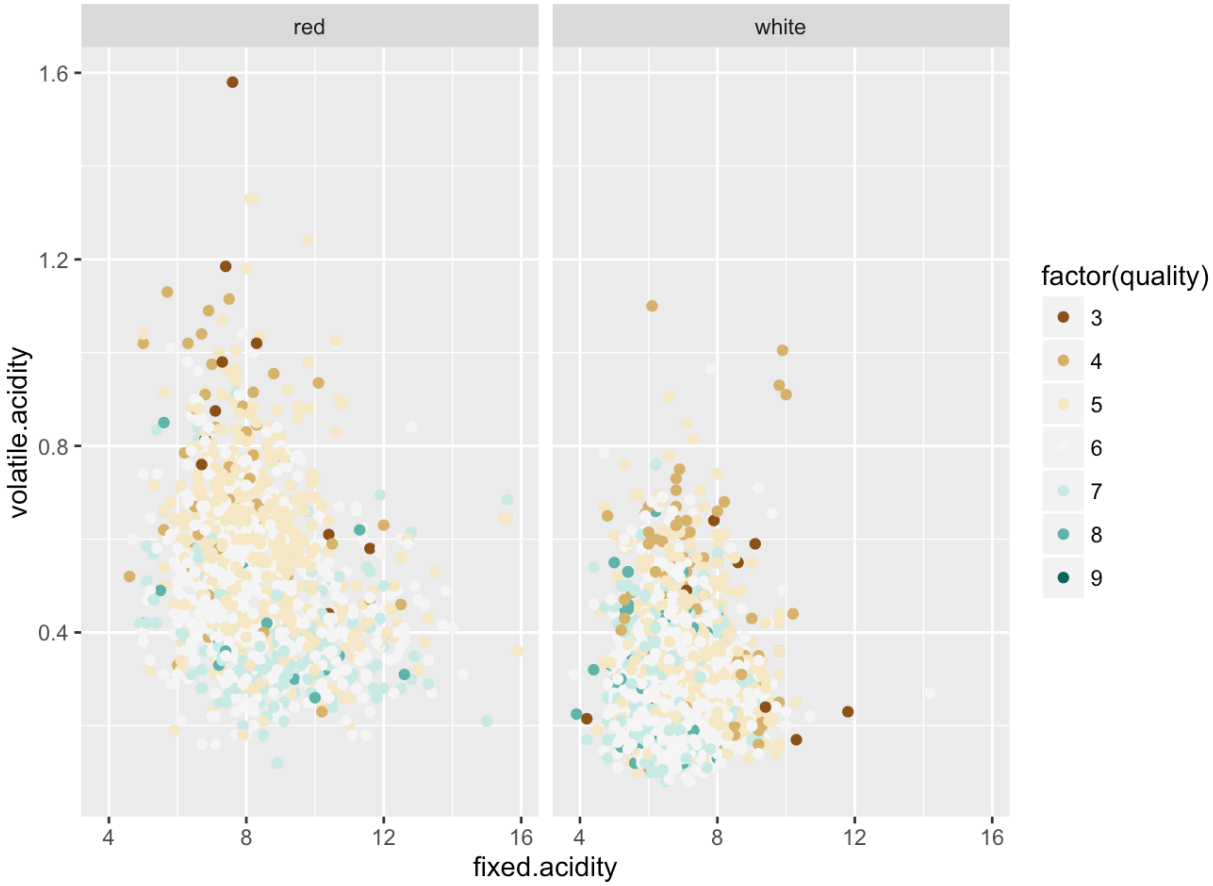




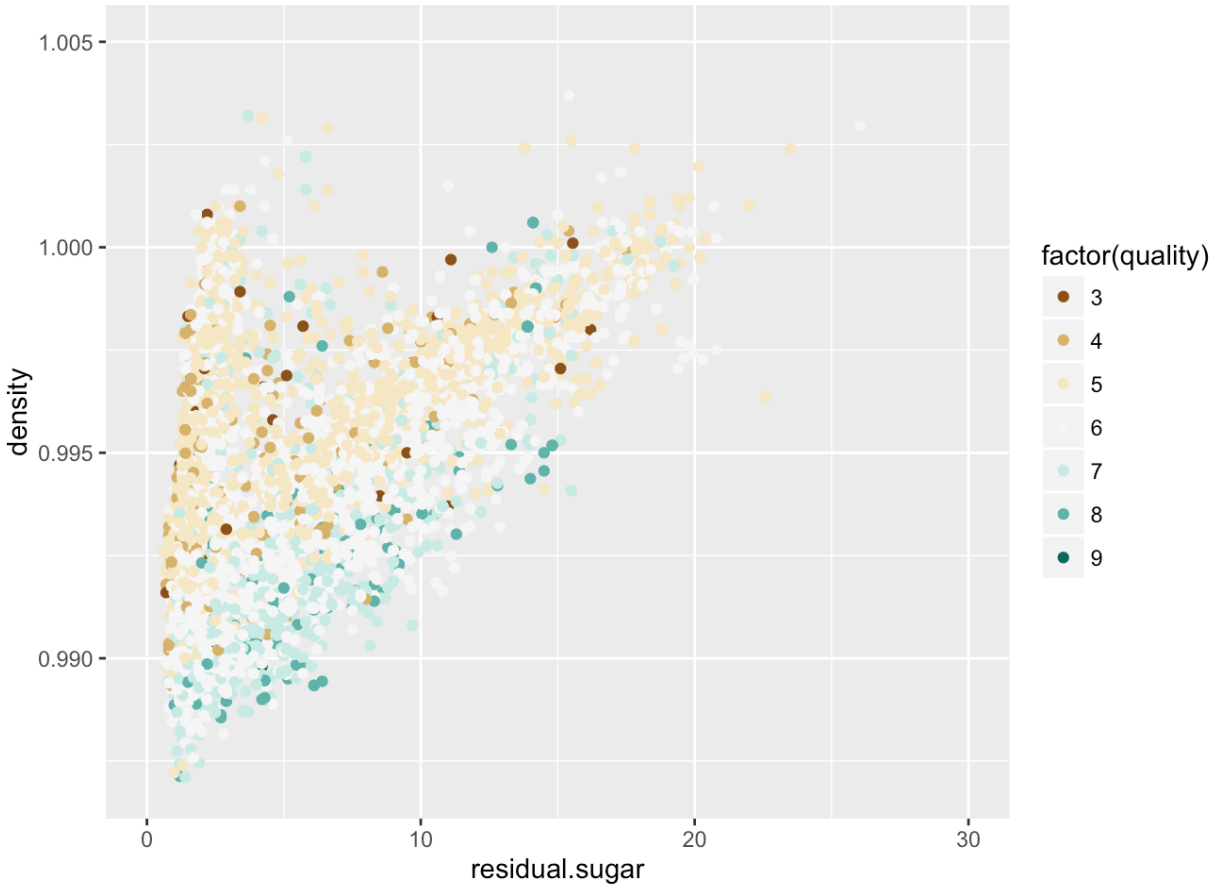
Higher rated wines tend to have higher alcohol contents and lower densities.

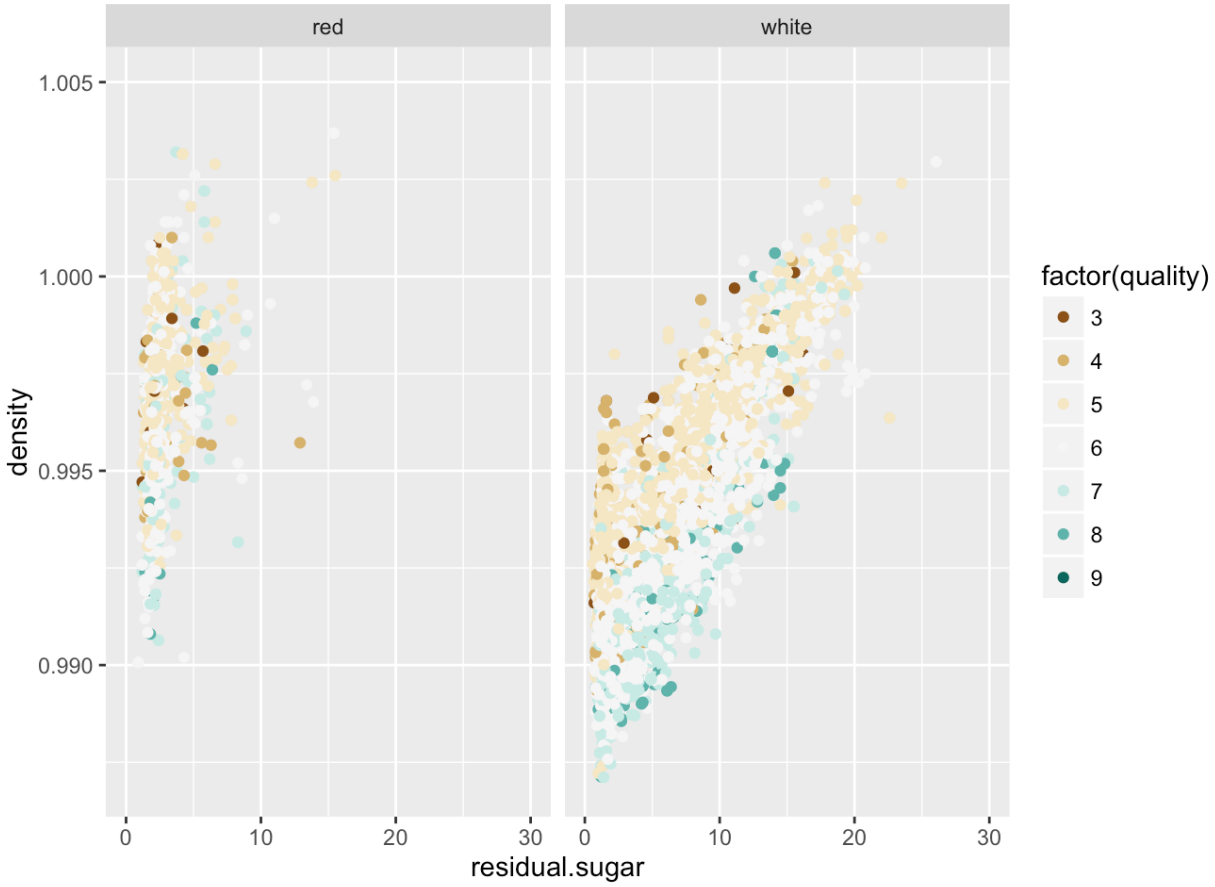


The amount of sulfur dioxide in white wines varies much more than red wines.

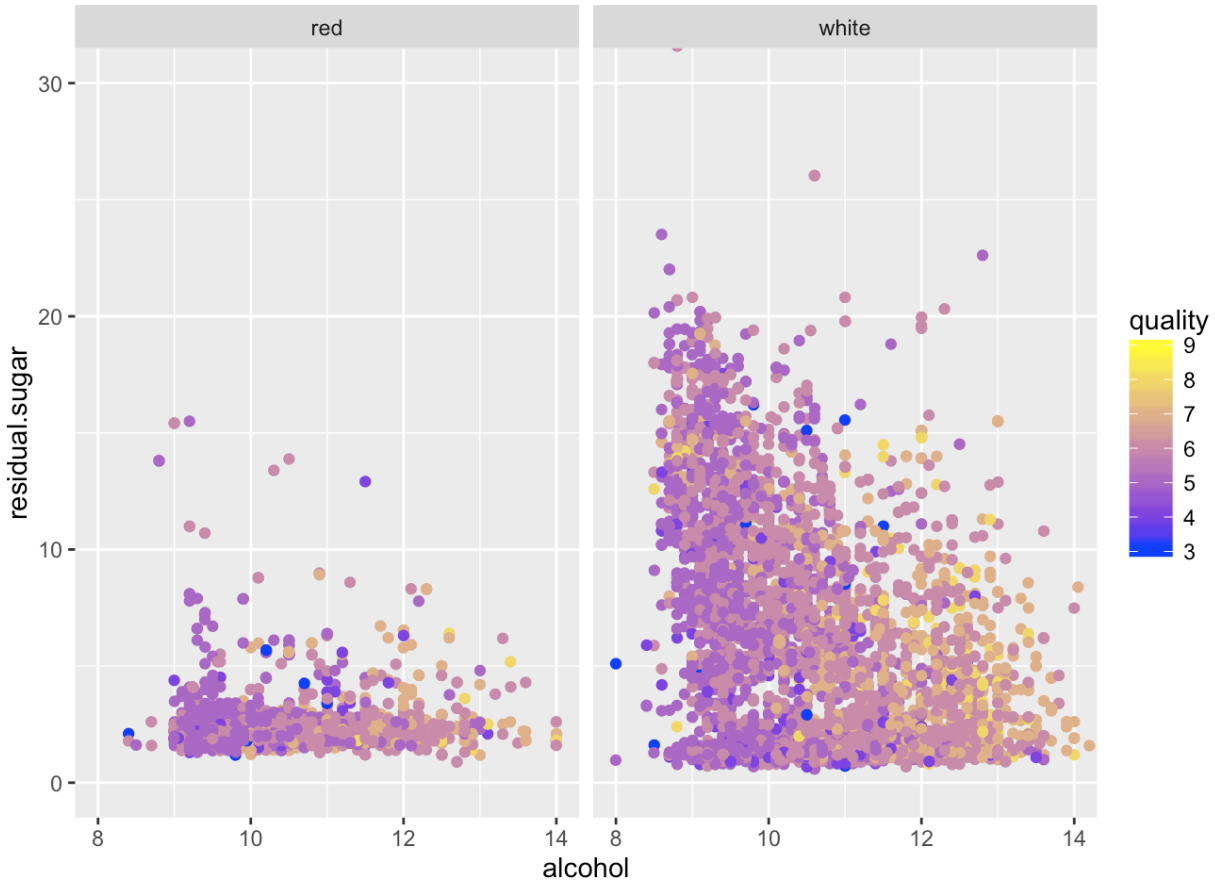


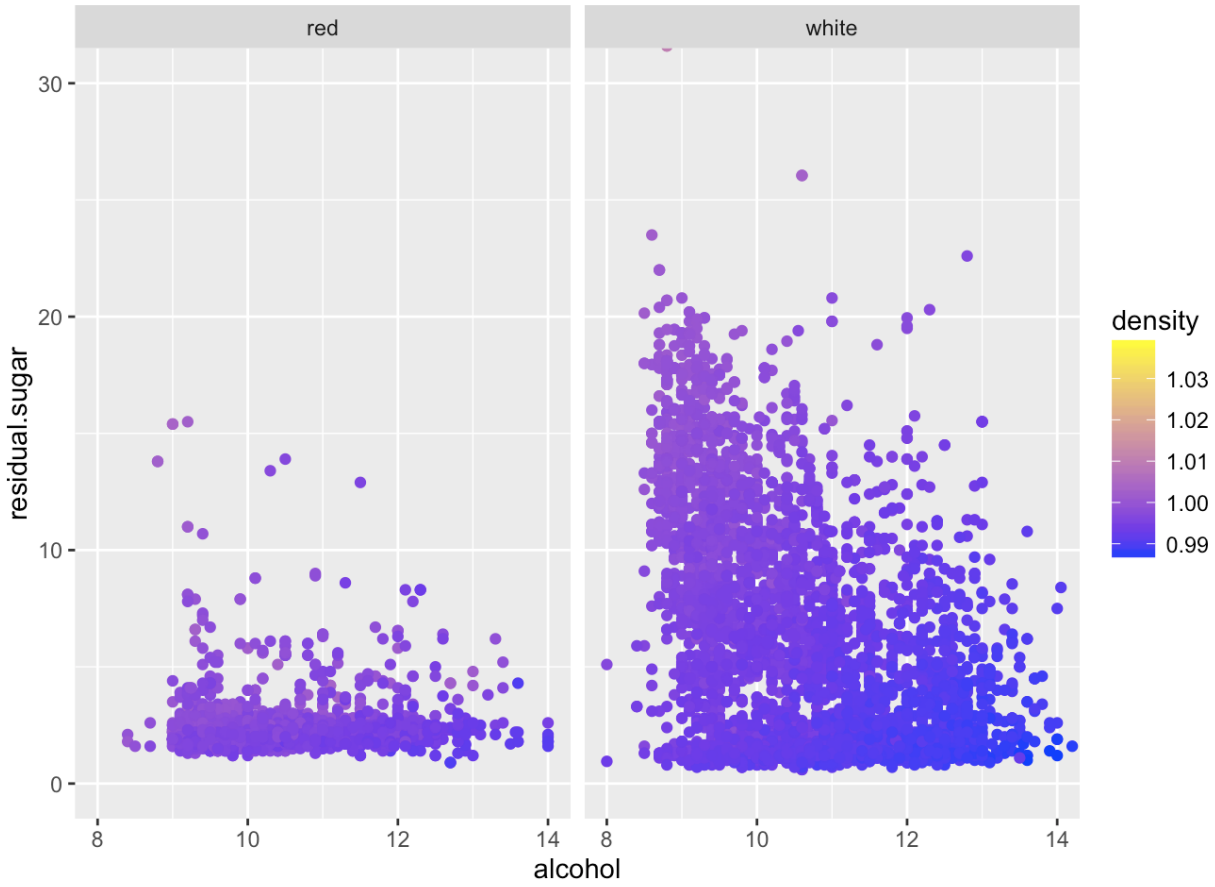
White wines are much less acidic than red wines. The higher rated red wines seem to have lower levels of acidity.





Density seems to have a larger affect on the quality of white wines compared to red. The majority of white wines also have lower densities than red wines, and their residual sugar is dispersed over a much larger range.





```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9901  0.9956  0.9968  0.9967  0.9978  1.0040
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9871  0.9917  0.9937  0.9940  0.9961  1.0390
```

Most of the wines in this dataset have a density that falls in the range of 0.99 to to 1.04. The first table above represents the summary of densities for red wines, while the second corresponds to whites.

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?

The main variables that affect wine quality were alcohol, residual sugar, and density. The density of wine decreases as alcohol rises, but increases as the level of residual sugar goes up. Most of the wines in this dataset have densities that fall between the range of 0.99 to 1.01

Total and free sulfur dioxide vary on a much broader scale for white wines compared to red wines. There is not a clear relationship between the amount of sulfur dioxide and quality of wine.

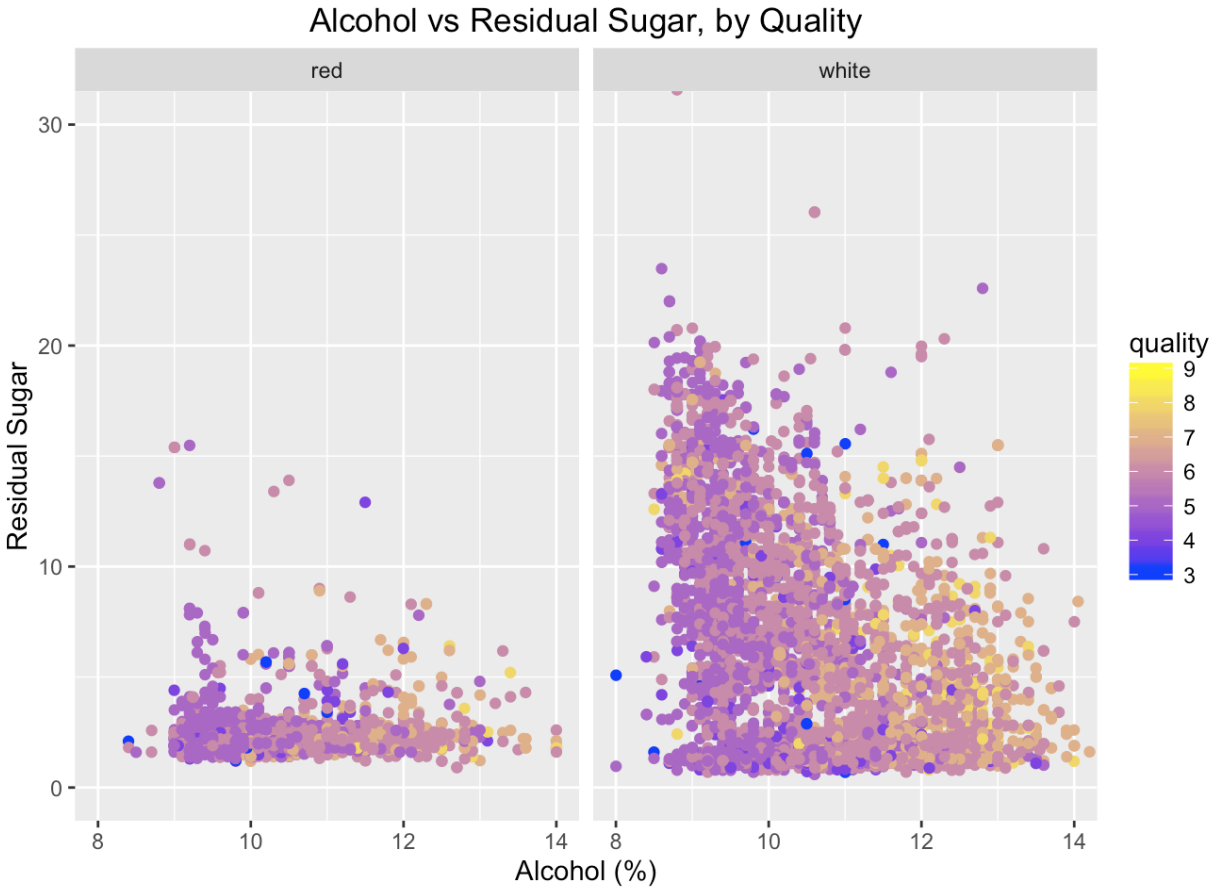
There is not much of a relationship between fixed acidity and volatile acidity. However, the less acidic red wines seem to score a higher quality rating.

Were there any interesting or surprising interactions between features?

I found it interesting that most of the red wines have residual sugar levels of around 2.2 to 2.5, regardless of the level of alcohol for that wine. Wine fermentation converts sugar into alcohol, so I would expect a negative relationship between these two variables, as shown for the white wines.

Final Plots and Summary

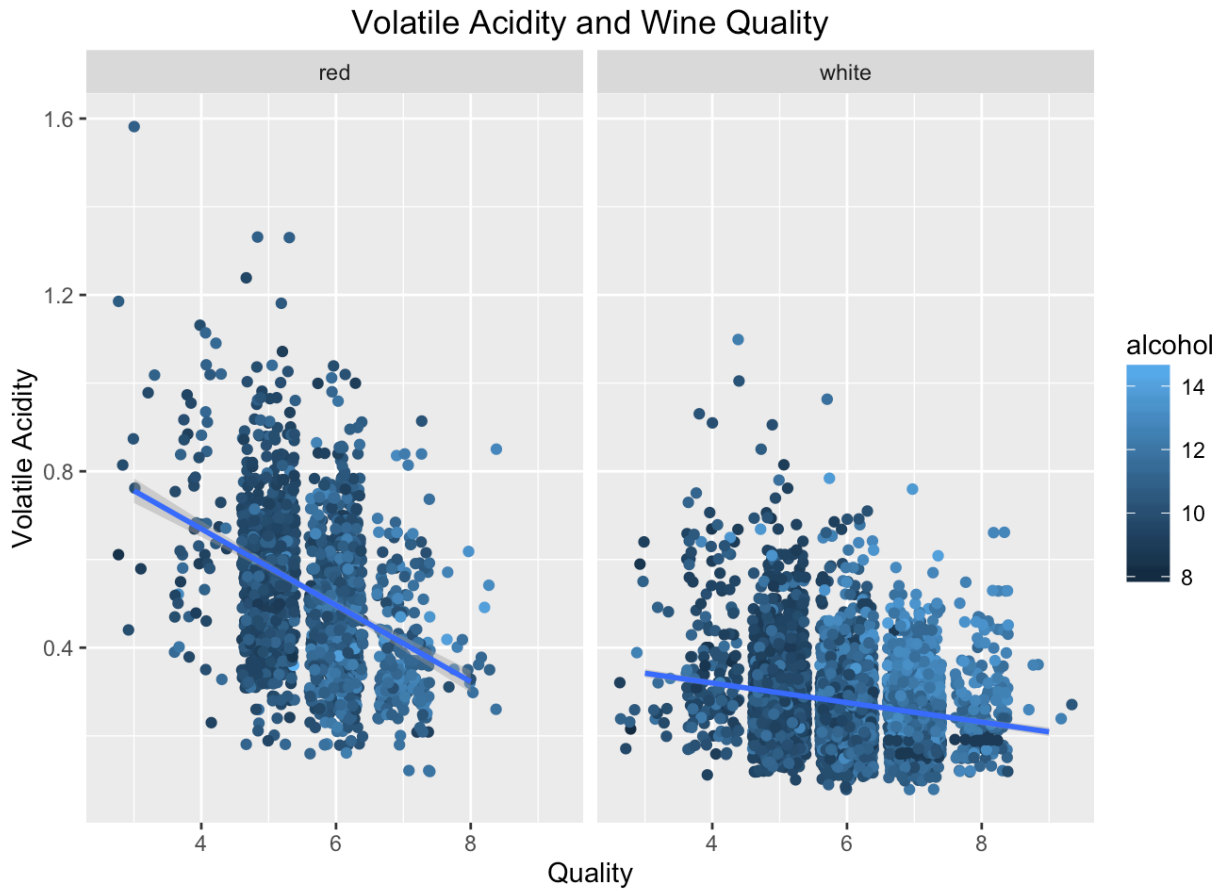
Plot One



Description One

The level of alcohol increases among white wines as residual sugar decreases. The red wine residual sugars vary much less than white wines, even as alcohol content increases. The highest rated wines are more concentrated at the right of these two plots, as the alcohol content increases.

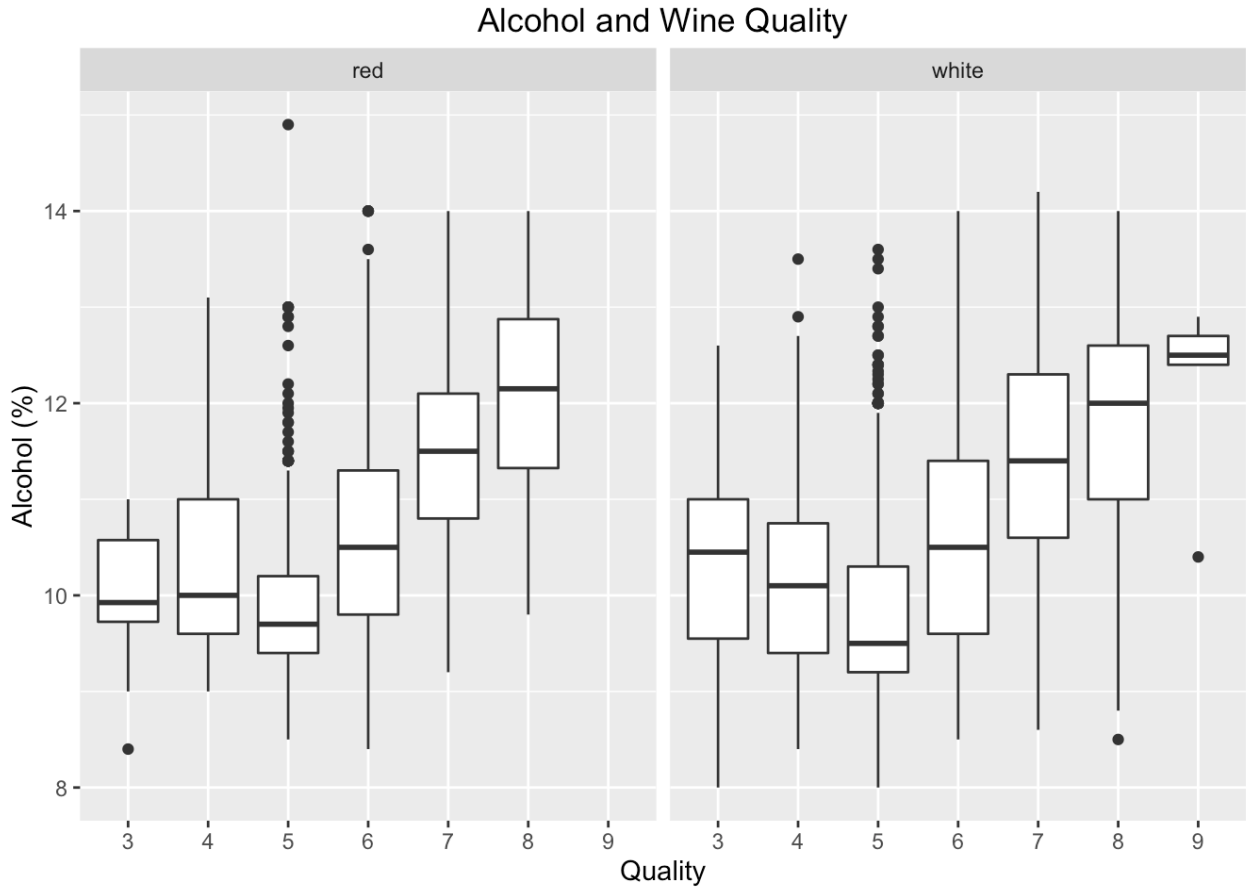
Plot Two



Description Two

This plot shows how there is a negative relationship between volatile acidity and quality for red wines. As volatile acidity decreases, most of the red wines are given a higher rating. This relationship is not as evident when it comes to rating white wines.

Plot Three



Description Three

There appears to be a positive relationship between alcohol content and the quality of wine for both red and white. However, this relationship only becomes apparent for wines with a quality of 6 and above. These higher rated wines have a median alcohol content of at least 10.5%

Reflection

The wine dataset used for this project contains close to 6500 wines with 14 different variables relating to each of them. It was created by combining two similar datasets of red and white wines, and adding an additional variable to distinguish between whether the wine was red or white. The variables relating to the wines included certain characteristcs, as well as an overall rating of the quality of the wine. By exploring the data, I set out to find relationships between the variables, uncover some of the differences between red and white wines, and find out why certain wines are rated higher than others.

The first step was to analyze single variables within this dataset. This was useful for finding similarities and differences between the variables of red and white wines. For example, red wine is usually more acidic and contains less sugar than white wine. The ratings of wines also follow a normal distribution.

After comparing relationships between multiple variables, I discovered several interesting patterns. I noticed that density is a function of the wine's alcohol level and residual sugar content. I also noticed that red wines are more acidic than white wines, and that a lower volatile acidity is related to higher quality ratings in red wine.

After doing some research on the wine making process, I found out that residual sugar is the amount of sugar left over after wine fermentation turns sugar from grapes into alcohol. This led me to believe that there should be a negative relationship between residual sugar and alcohol. What I found interesting was that this negative relationship was true for white wines; however, for red wines, as alcohol increased, residual sugar remained unchanged. Out of all the variables in this dataset, the one with the strongest correlation to alcohol quality for both red and white wines was alcohol content.

One of the limitations to this data is that rating a particular wine based on how it tastes is a matter of preference. Although these ratings are performed by experts, not everyone will have the same preferences on how wine tastes. To further investigate this dataset, I would be interested in building a linear model that would be able to predict wine quality based on the value of its other variables. It would also be interesting to look at the prices of these wines to determine whether it is worth spending money on wine, and what would be the optimal price to pay for a bottle.