Wine Quality Exploration by Neeharika Gottipati

```
## [1] 6497   13
```

This data set contains 6497 wines with 13 variables on the chemical properties of the wine.

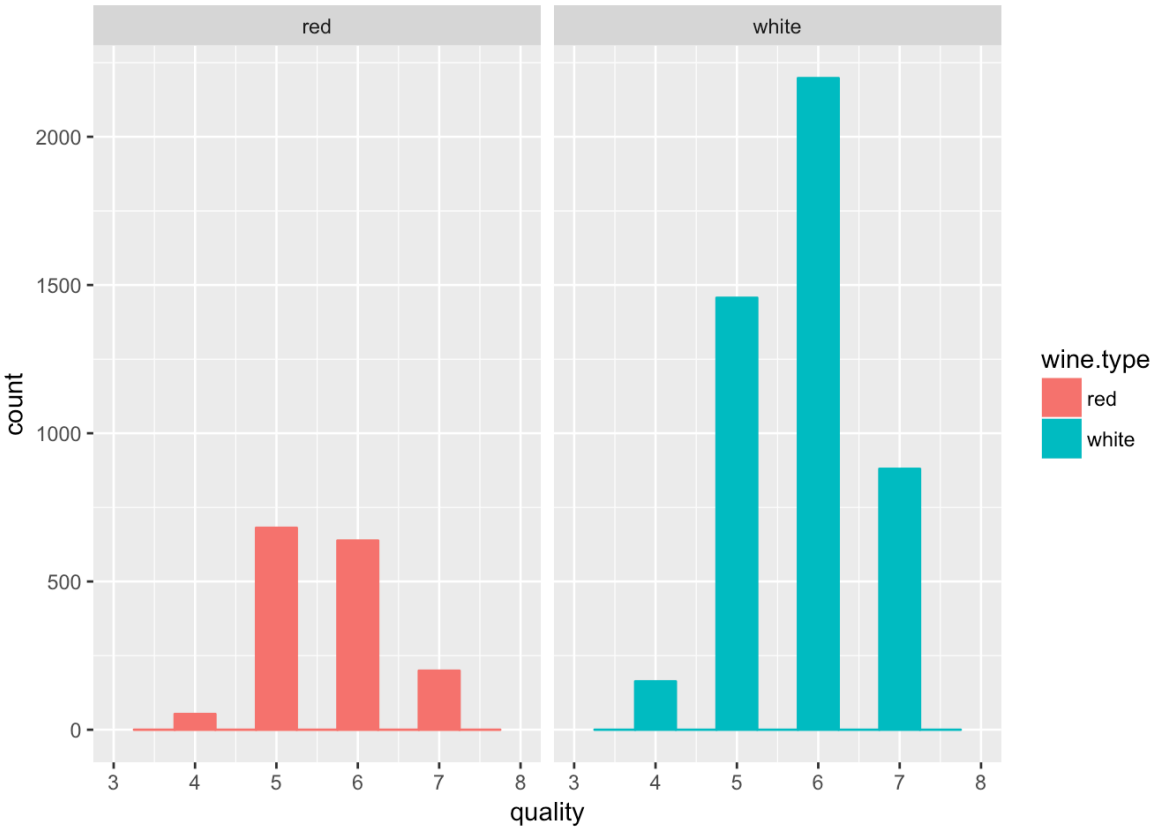
Univariate Plots

```
##
##   red white
##  1599  4898
```

This data set contains 1599 red wine observations and 4898 white wine observations.

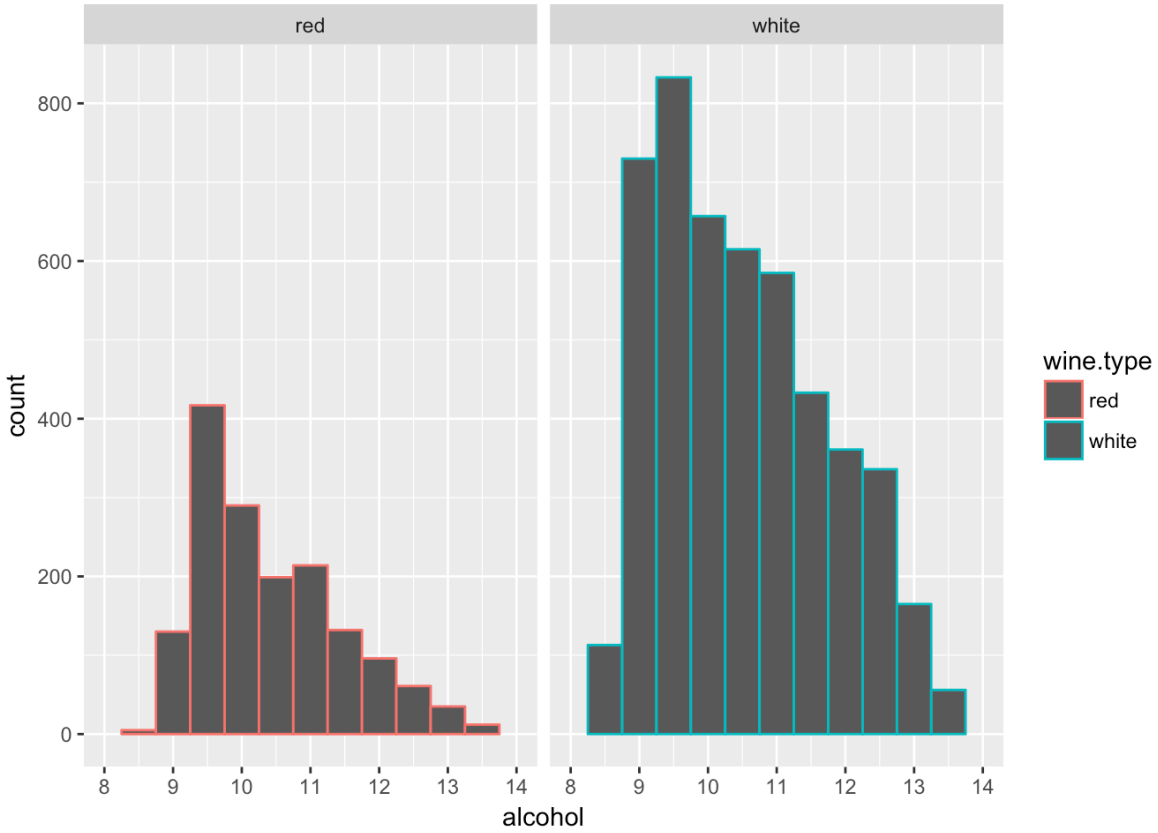
```
## 'data.frame':    6497 obs. of  13 variables:
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
##  $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
##  $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071 ...
##  $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
##  $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
##  $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
##  $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...
##  $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...
##  $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
##  $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...
##  $ wine.type           : Factor w/ 2 levels "red","white": 1 1 1 1 1 1 1 1 1 1 ...
```

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



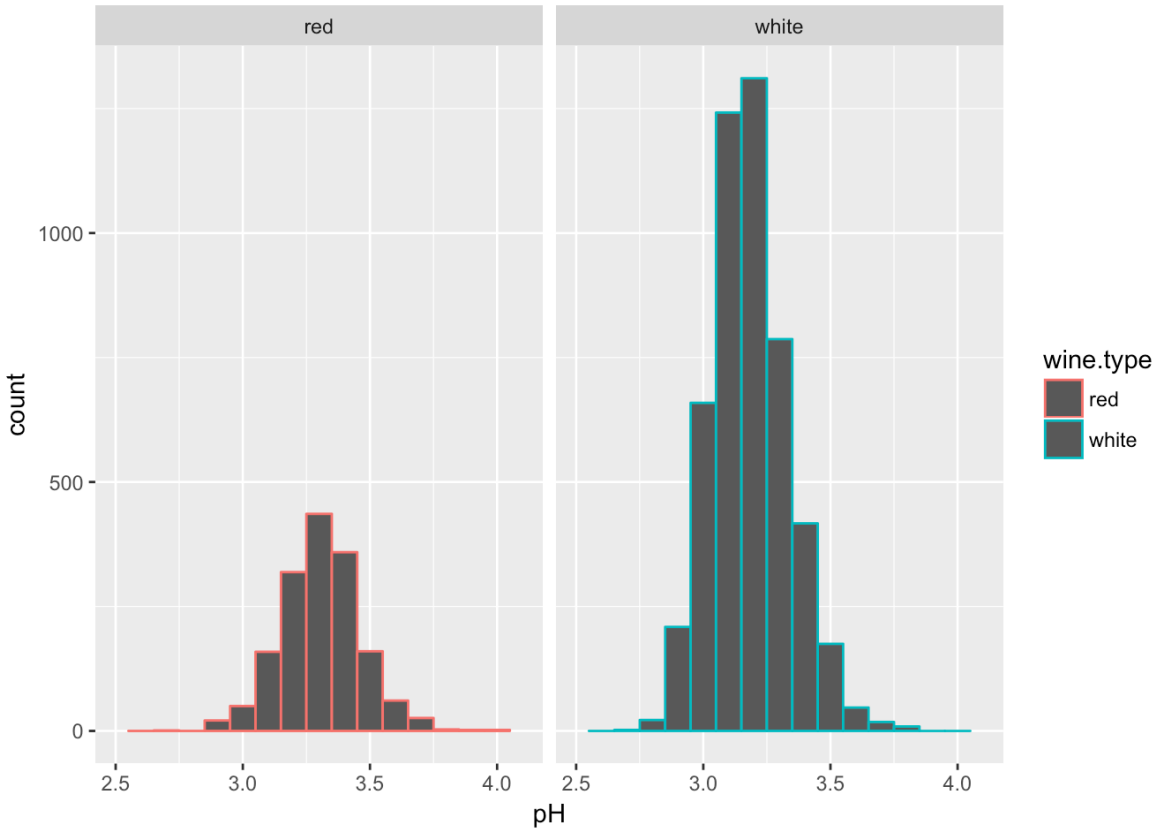
```
## wine$wine.type: red
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   5.000   6.000   5.636   6.000   8.000
## ------------------------------------------------------------
## wine$wine.type: white
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   5.000   6.000   5.878   6.000   9.000
```

- From the graph above the quality for most of the red wines is 5 and 6. Where as the quality for white wines is peaked at 6.
- From the summary above the median for both the wine types is 6.



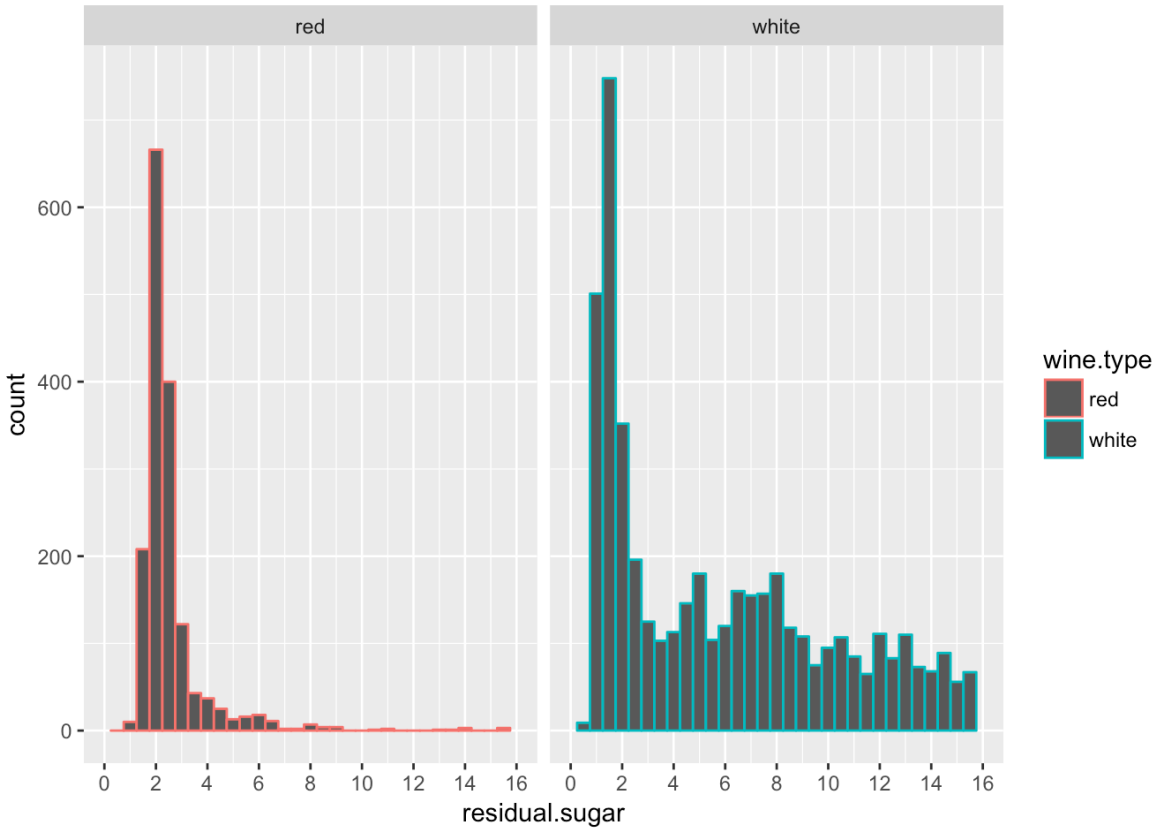

```
## wine$wine.type: red
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.50   10.20   10.42   11.10   14.90
## ------------------------------------------------------------
## wine$wine.type: white
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00    9.50   10.40   10.51   11.40   14.20
```

Most of the wines have an alcohol level of 10. The number of wines(red and white) decreases after the alcohol level 10.



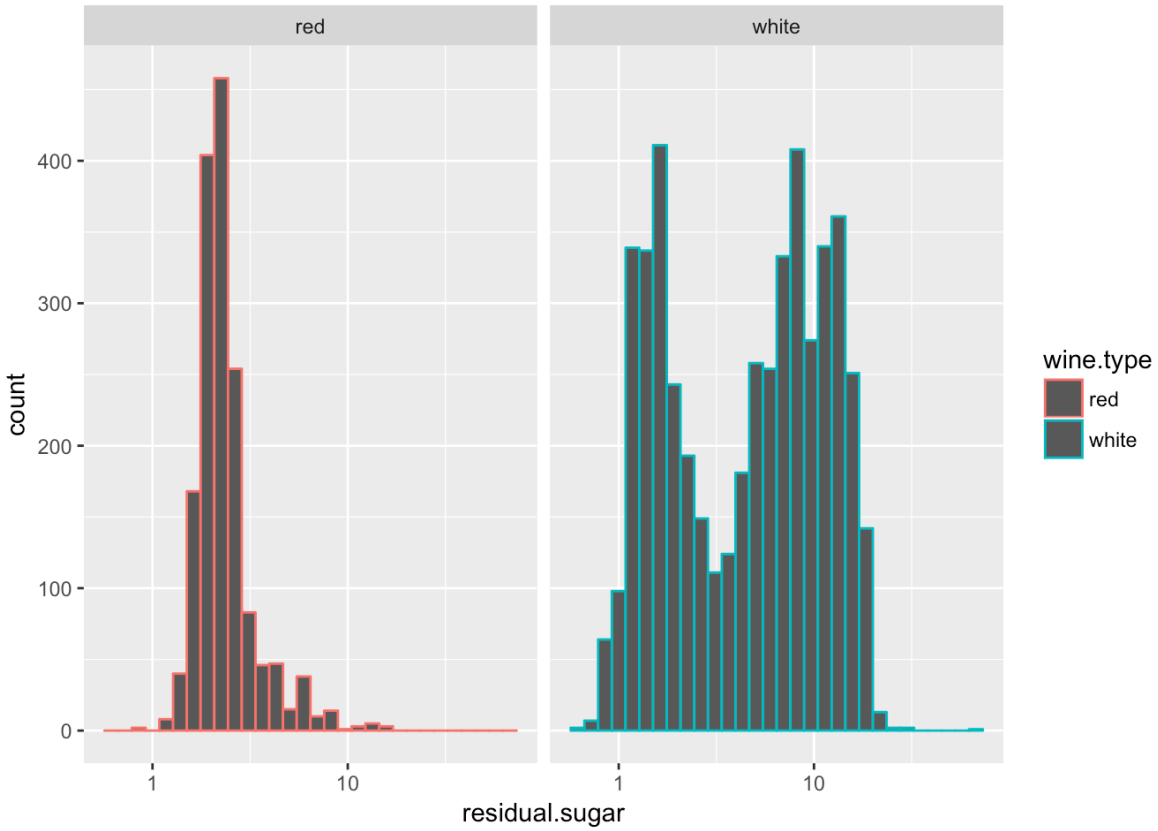

```
## wine$wine.type: red
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.740   3.210   3.310   3.311   3.400   4.010
## ------------------------------------------------------------
## wine$wine.type: white
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.720   3.090   3.180   3.188   3.280   3.820
```

The pH is normally distributed wiht a peak at 3.3.

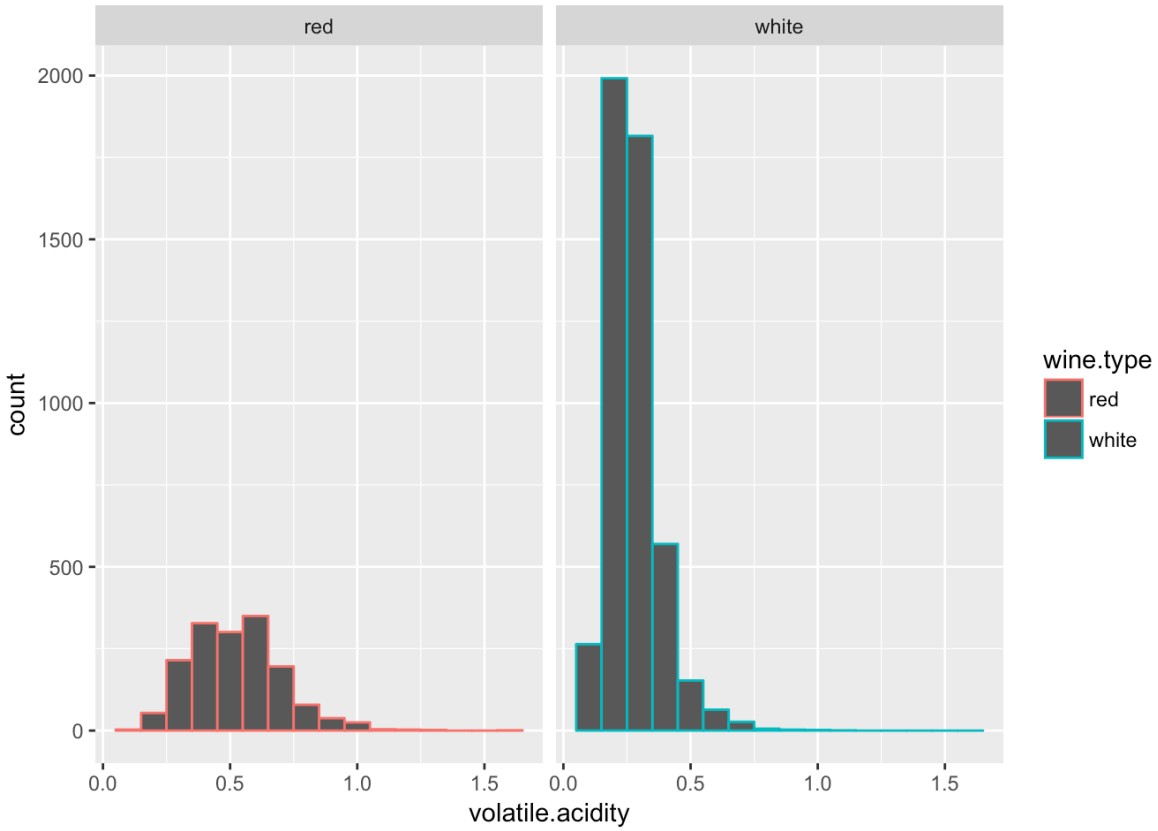



```
## wine$wine.type: red
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.900   1.900   2.200   2.539   2.600  15.500
## ------------------------------------------------------------
## wine$wine.type: white
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.600   1.700   5.200   6.391   9.900  65.800
```

The above graph is skewed to the right, with a peak at 2.

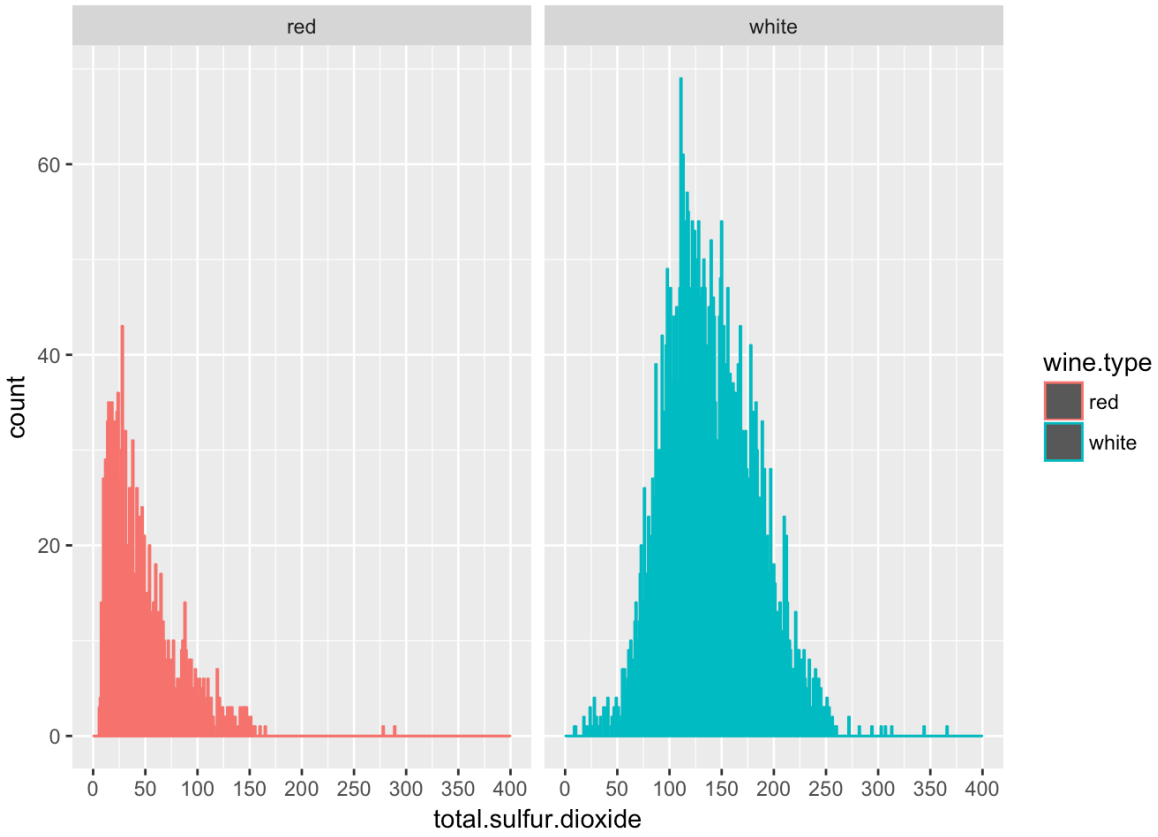


I applied log transformation to right skewed distributions and we can see two peaks for white wine and a single peak for red wine.



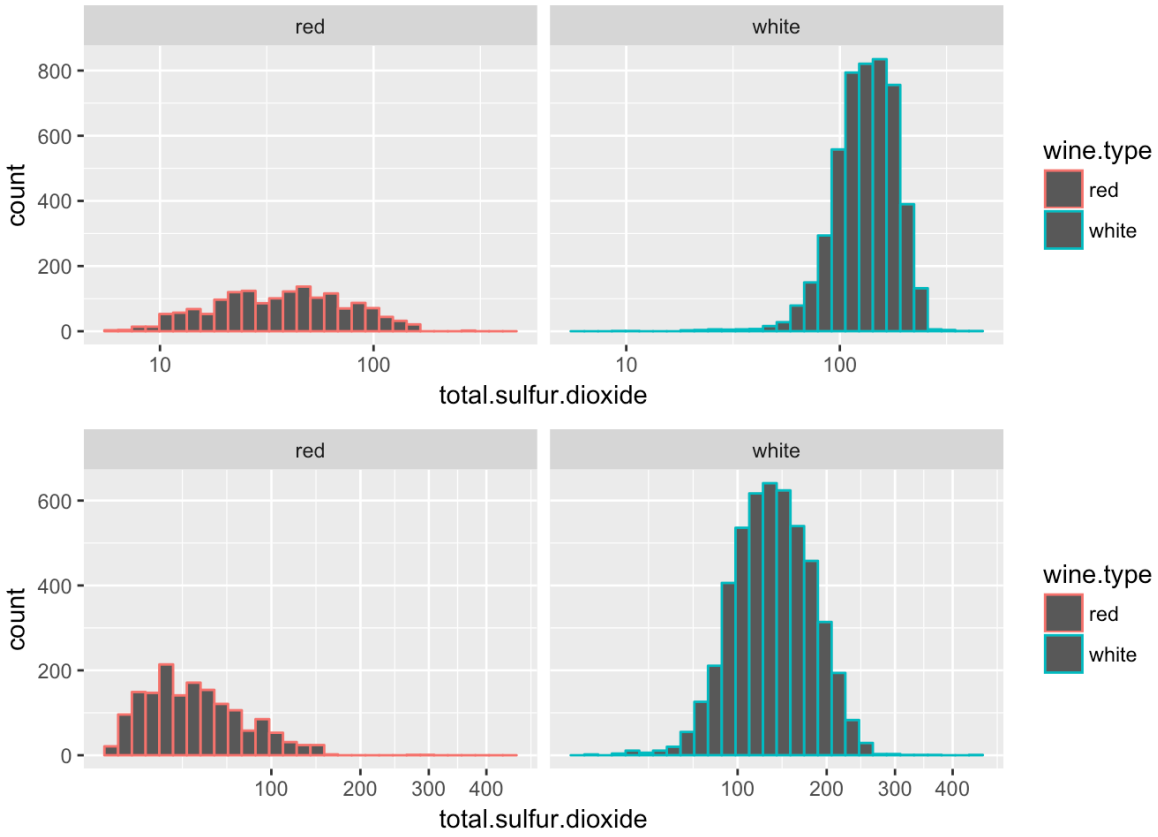
```
## wine$wine.type: red
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1200  0.3900  0.5200  0.5278  0.6400  1.5800
## ------------------------------------------------------------
## wine$wine.type: white
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0800  0.2100  0.2600  0.2782  0.3200  1.1000
```

The volatile acidity distributions are skewed to the right with a peak around 0.5 for red wine and 0.3 for white wine. The median for red wines is 0.52 and for white wines is 0.26.

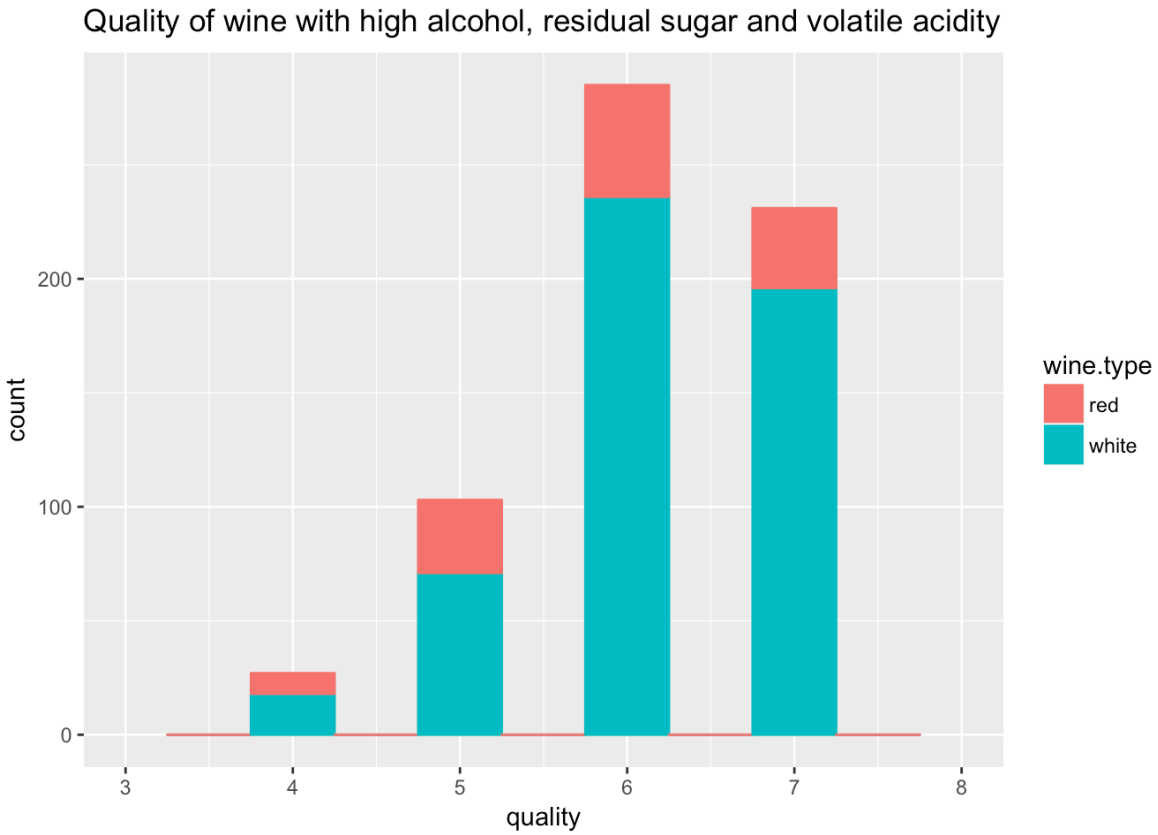



```
## wine$wine.type: red
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00   22.00   38.00   46.47   62.00  289.00
## ------------------------------------------------------------
## wine$wine.type: white
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     9.0   108.0   134.0   138.4   167.0   440.0
```

The red wine is skewed to the right with a long tail and with a peak around 50. The white wine data is right skewed with a normal distribution with a peak around 134.

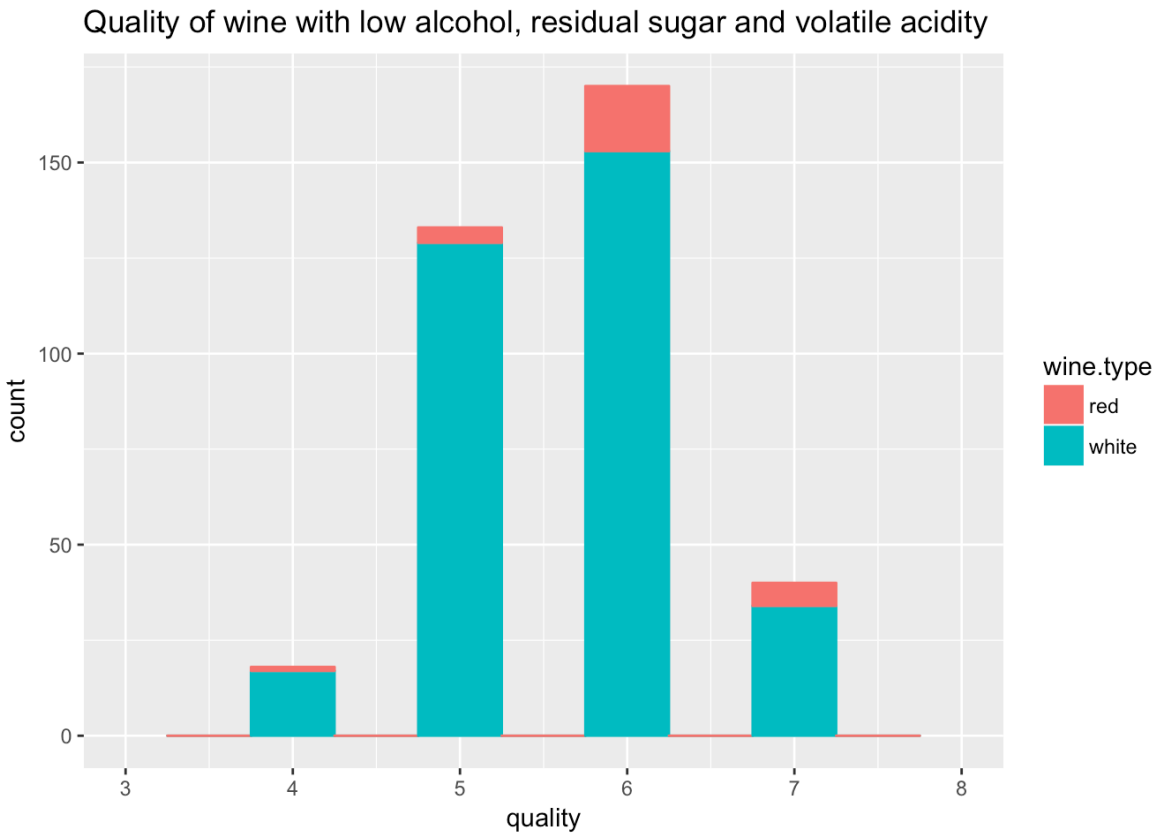


I applied log and sqrt transformation to the right skewed and distributed data. Transformed white wine data is more close to normal distribution with a peak around 130 and most of red wine data is below 100.



```
## wine1$wine.type: red
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   5.000   6.000   5.885   7.000   8.000
## ------------------------------------------------------------
## wine1$wine.type: white
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   6.000   6.000   6.301   7.000   9.000
```

From the above plot with high alcohol, residual sugar and volatile acidity, we can see that most of the wines have a quality range of 6 and a median of 6. The count of red wine is less because the volatile acidity median for red wine is 0.52, which is higher than the median 0.29 for whole wine data set




```
## wine1$wine.type: red
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.000   6.000   6.000   5.929   6.000   7.000
## ------------------------------------------------------------
## wine1$wine.type: white
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   5.000   6.000   5.542   6.000   8.000
```

In the above plot wiht low alcohol, residual sugar and volatile acidity, we can see that the count of the wines is higher for quality level 5 and both the wines have a median of 5

Univariate Analysis

What is the structure of your dataset?

This data set contains 6497 wines with 13 variables on the chemical properties of the wine. At least 3 wine experts rated the quality of each wine, providing a rating between 0 (very bad) and 10 (very excellent).

Other observations: * White wine has 4898 observations and for red wine there are only 1599 observations * pH values are normally distributed with a median of 3.2 * Most wines have a quality of 5 or 6 * The mean and median values for alcohol is around 10.2 * Most of the wines have a residual sugar of 3

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

The main features in this data set are quality and alcohol. This data set contains different chemical values to determine the wine quality. I think there is a strong relationship between alcohol and quality.

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

The other main features that determine quality of wine are residual sugar, volatile acidity, total sulfur dioxide, chlorides, sulphates and pH. I think alocohol and total sulfur dioxide probably contribute to the quality of the wine. But there are several other features that affect the wine quality are not given in this data set.

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

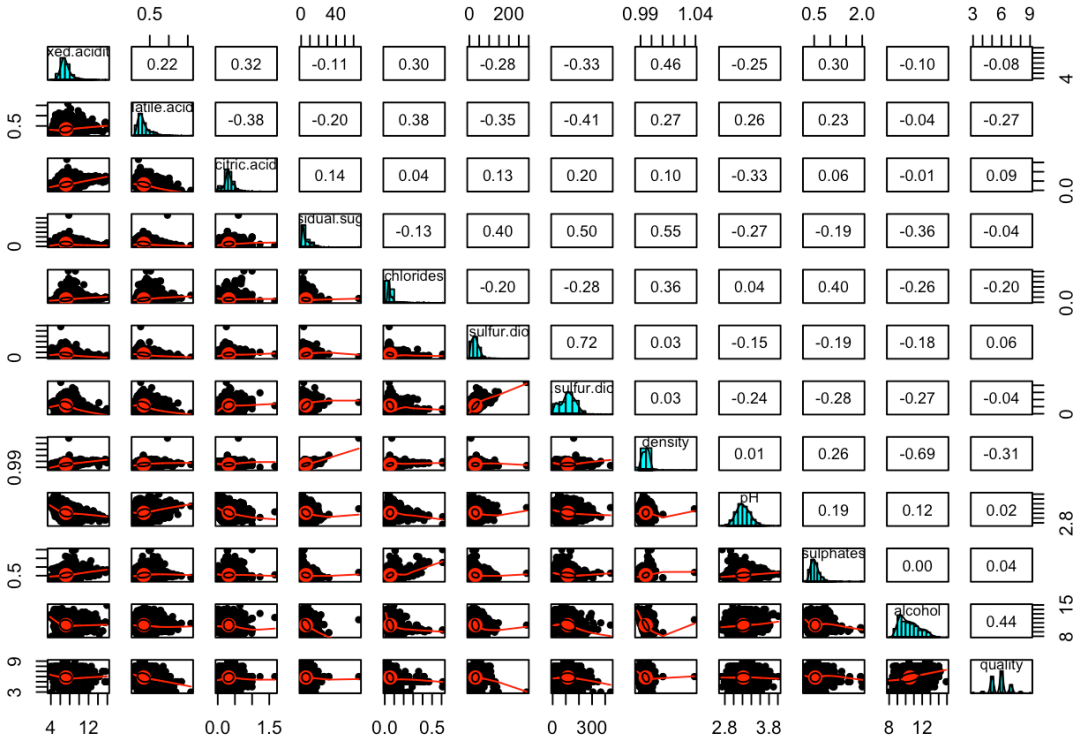
I created a categorical variable of wine type and combined both red and white wine data sets, which helped me to analyse the difference between the both wines.

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?

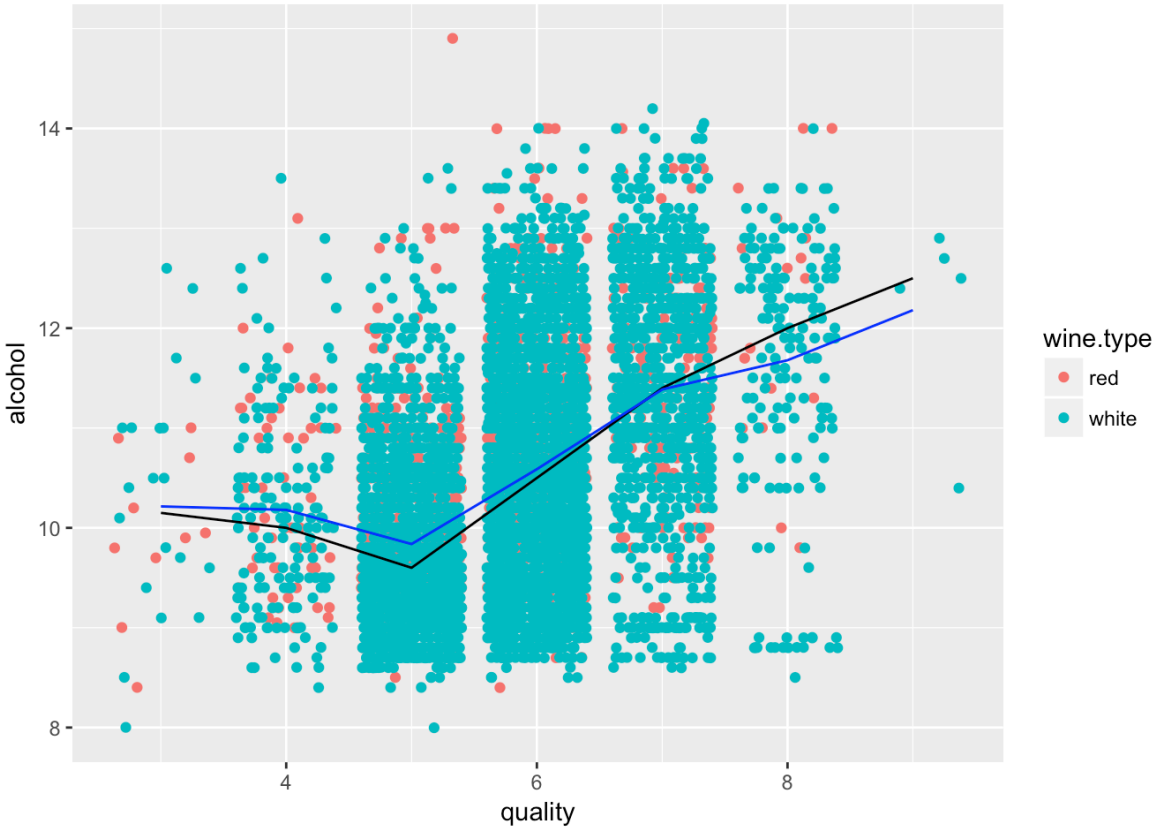
I log-transformed the right skewed residual sugar and total sulfur dioxide distributions. Transformed residual sugars for red wine distribution shows peak at 3 where as for white wine it shows peak at 3 and 10.

To examine the quality levels with high alcohol, residual sugar and fixed acidity, I subset the data with higher that median values for alcohol, residual sugar and volatile acidity. By plotting this subset data, I observed that the count of wines is higher with quality level of 6 and with low median levels the count of data is higher with quality of 5. By this there are chances that these properties define a certain amount of quality level.

Bivariate Plots Section

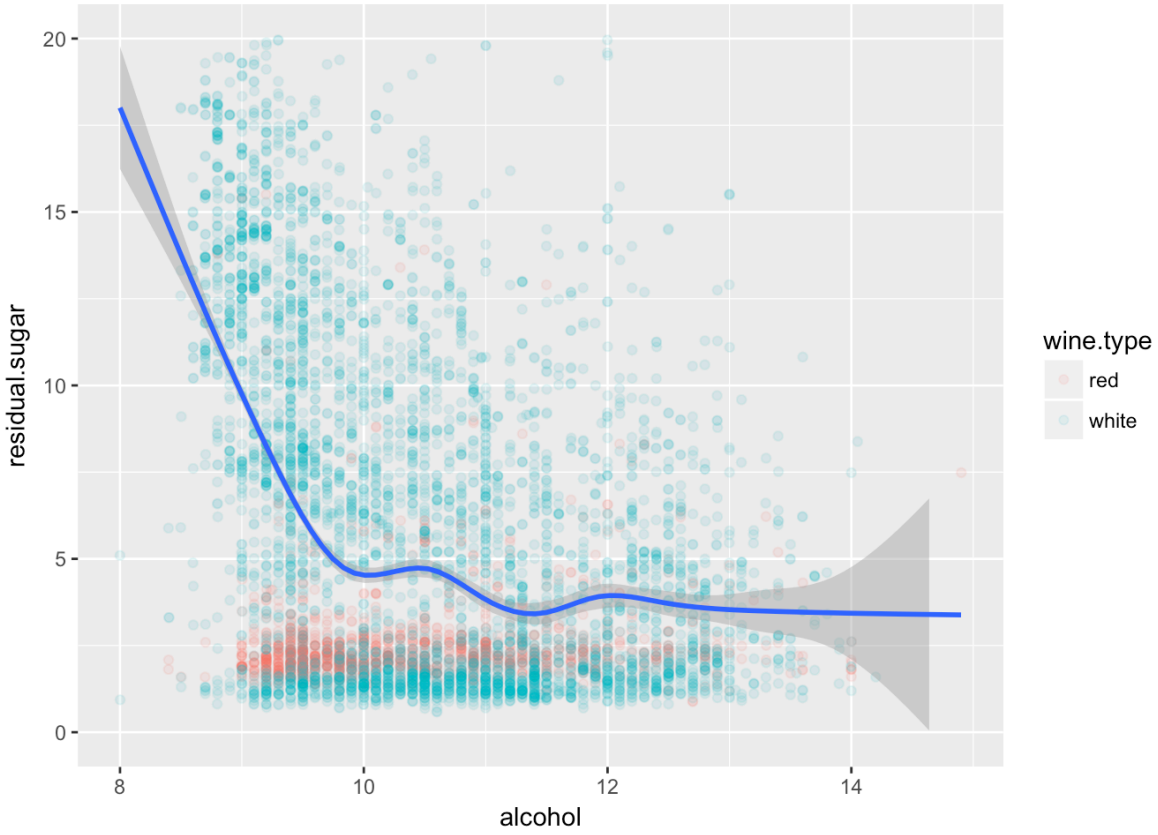


From the graph above, I see there is a correlation between alcohol and quality. I want to further examine the relationship between sulfur dioxide, residual sugar, and alcohol.

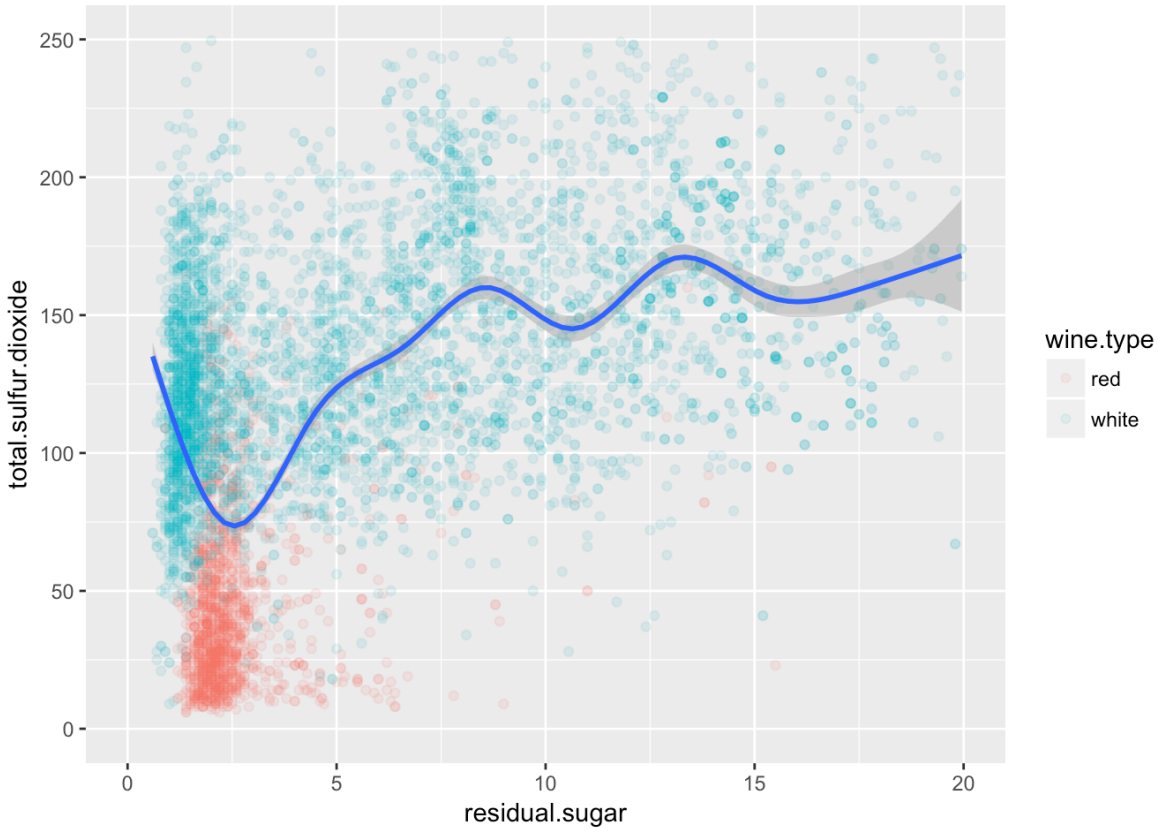


```
##
##  Pearson's product-moment correlation
##
## data:  quality and alcohol
## t = 43.091, df = 6249, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4592722 0.4975013
## sample estimates:
##       cor
## 0.4786136
```

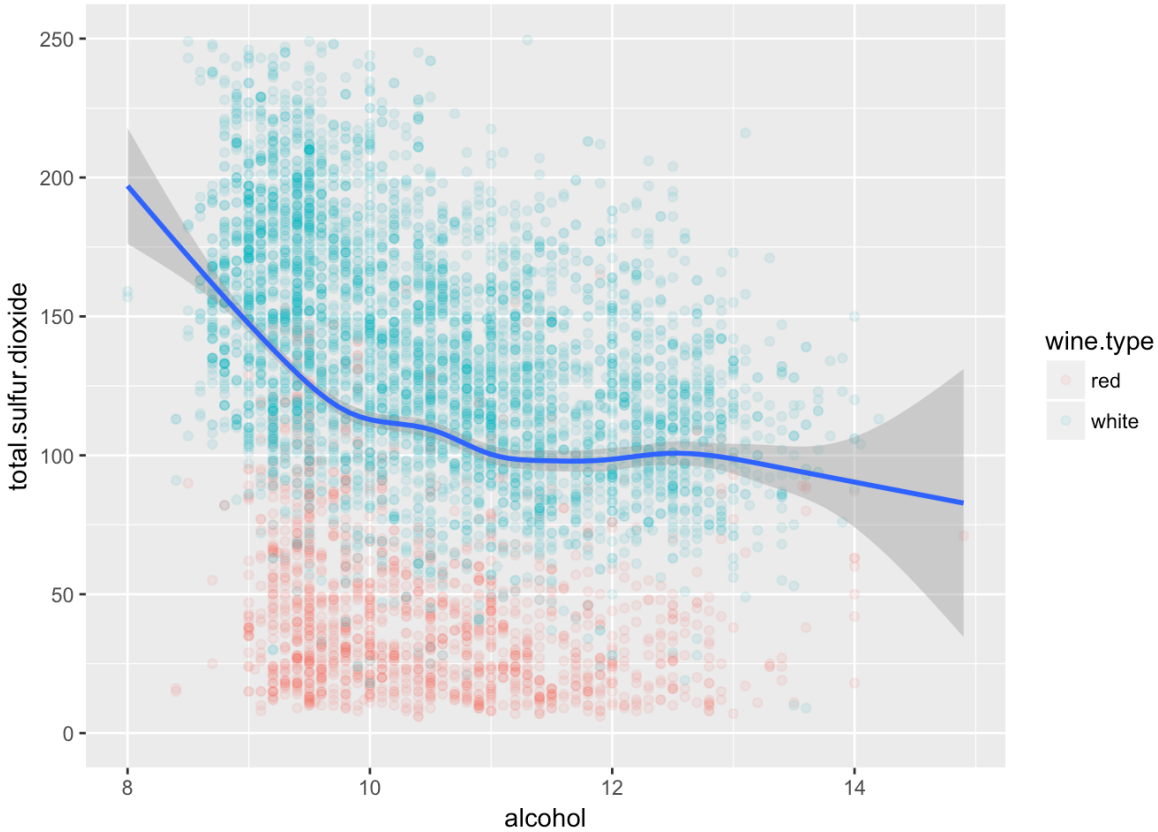
- There is no significant difference between the red wine and white wine for alcohol vs quality level.
- The correlation coefficient shows that there is some level of relationship between alcohol and quality score.



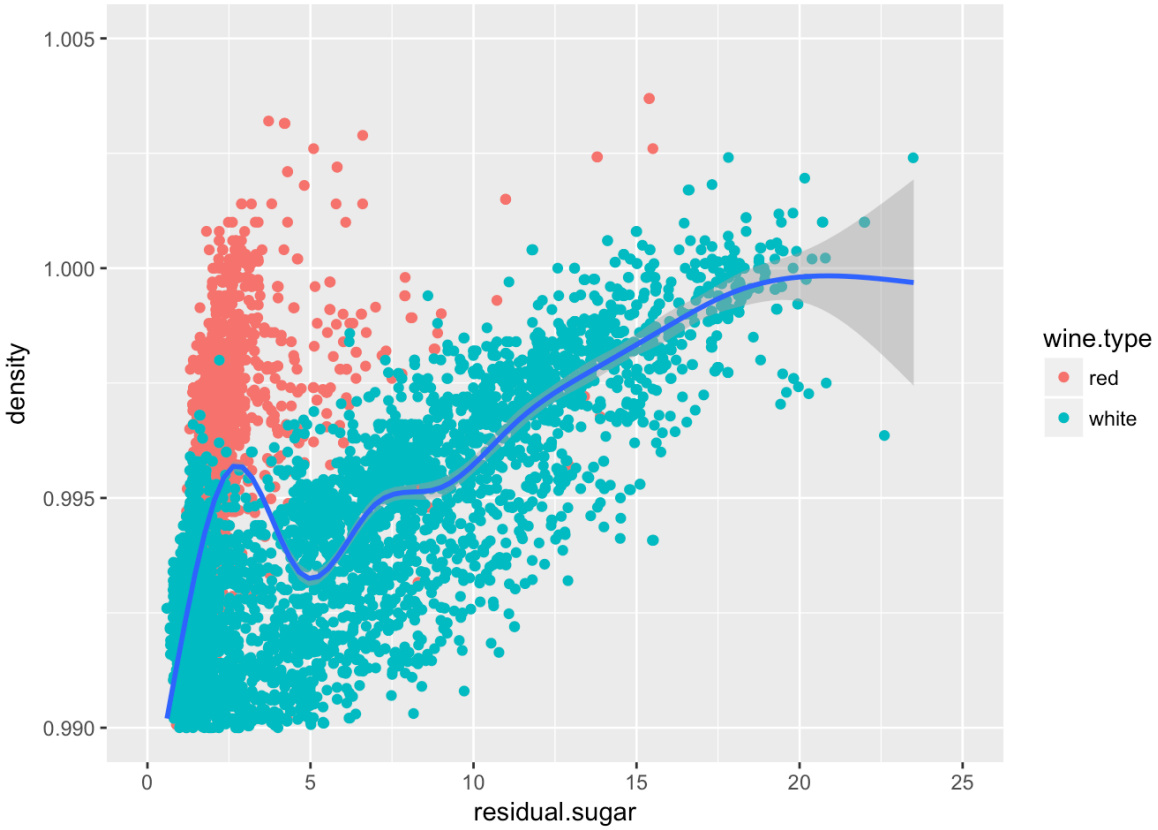
Residual sugar and alocohol has negative correlation. Residual sugar drops from ~18 to ~5 with increase in alochol level from 8 to 10 and residual sugar is at ~5 for wines that have alcohol levels above 10



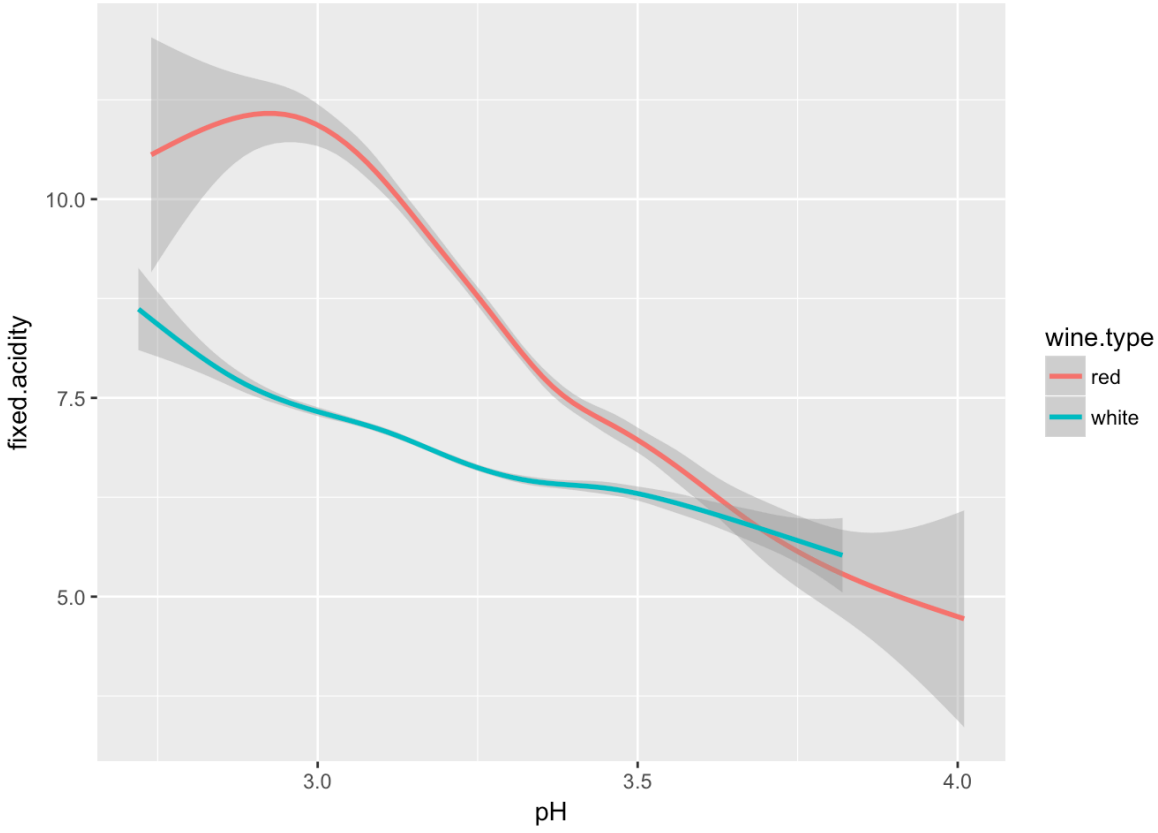
The total sulfur dioxide and residual sugar have positive correlation except for wines that have residual sugar below ~3. Most of the red wines have less residual sugar and total sulfur dioxide compared to white wines.



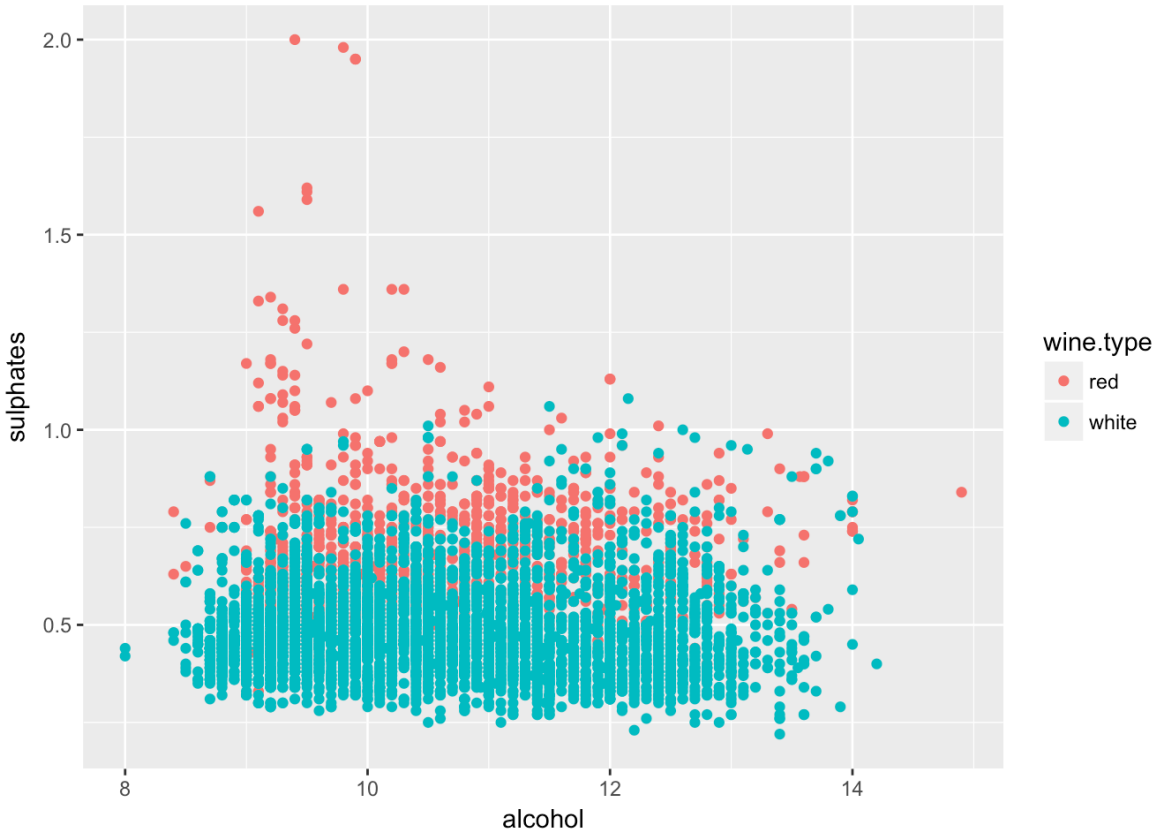
Total sulfur dioxide and alcohol have negative correlation, there is reduction of total sulfur dioxide levels with incresing alcohol. Total sulfur dioxide seems to be less for red wine compared to the white wine.



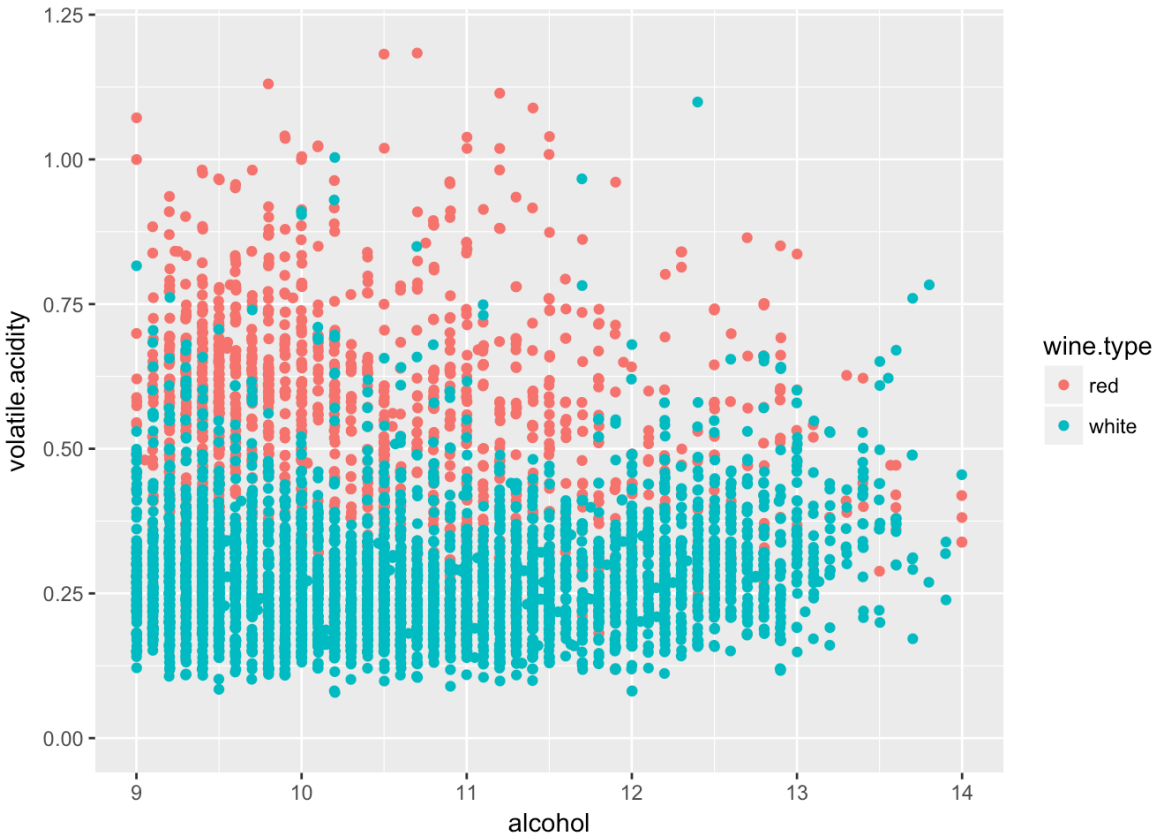
From above plot we can see that the density of wine is incresed with residual sugar. Even though the residual sugar drastically increases from 5 to 20, the density falls between 0.990 and 1.005.



In the above two plots, the pH and fixed acidity are negatively correlated, with increase in pH the fixed.acidity is decreasing. From the above observation we can see that red wines have high fixed acidity than white wines.



The amount of sulphates seems to be higher for red wines than white wines. Sulphates level are 0 to ~1 for most of the wines with alcohol levels ranging from 8 to 14.



The volatile acidity in red wine seems to be higher compared to white wine.

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?

- I see a positive correlation between the alcohol and quality levels, but not very strong correlation.
- The residual sugar level decreases with increase in alcohol level until 10. After alcohol level 10 there seems to be a slight reduction until the alcohol level 14.
- Total sulfur dioxide and residual sugar have a positive correlation.
- There is negative correlation between total sulfur dioxide and alcohol. Total sulfur dioxide is less for red wine compared to white wine.
- Density and residual sugar have a strong positive correlation. But level of density increased is very less compared to residual sugar.
- pH and fixed acidity are negatively correlated.

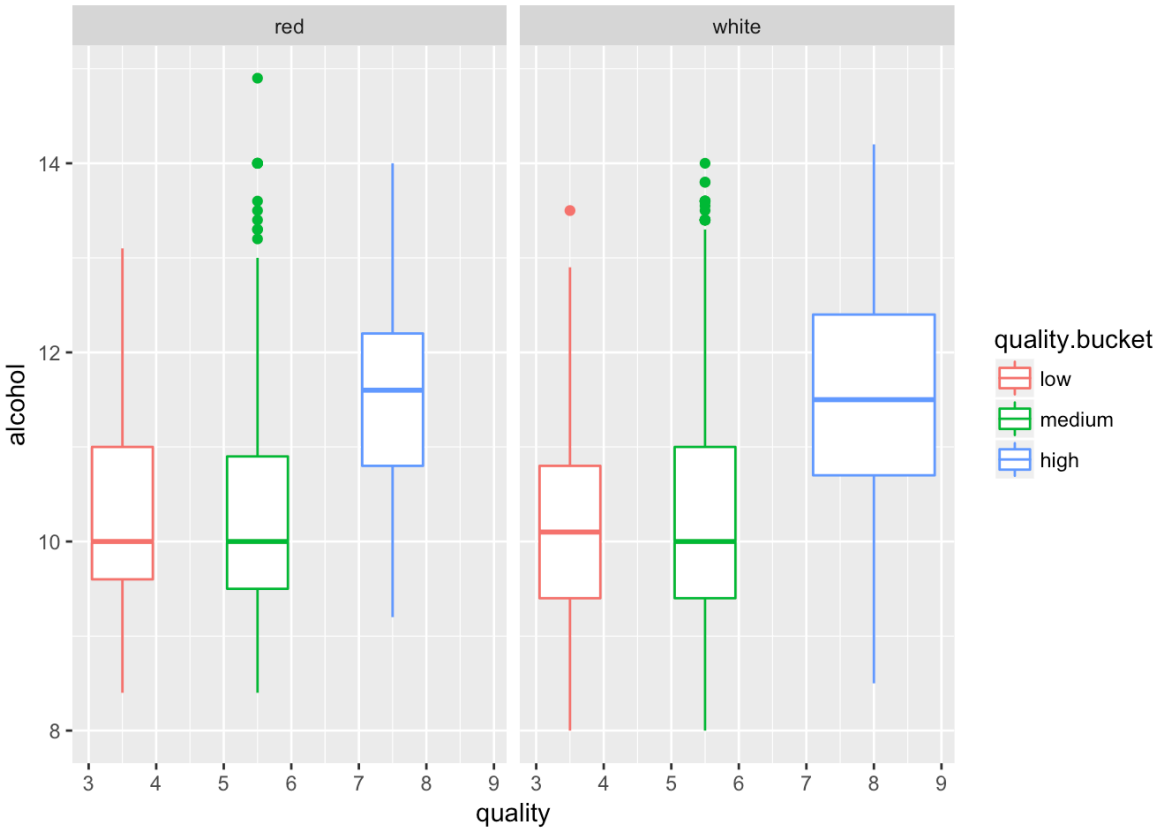
Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

Alcohol and residual sugar have negative correlation. Alcohol is produced by converting the sugars using the yeast, which explains the relationship.

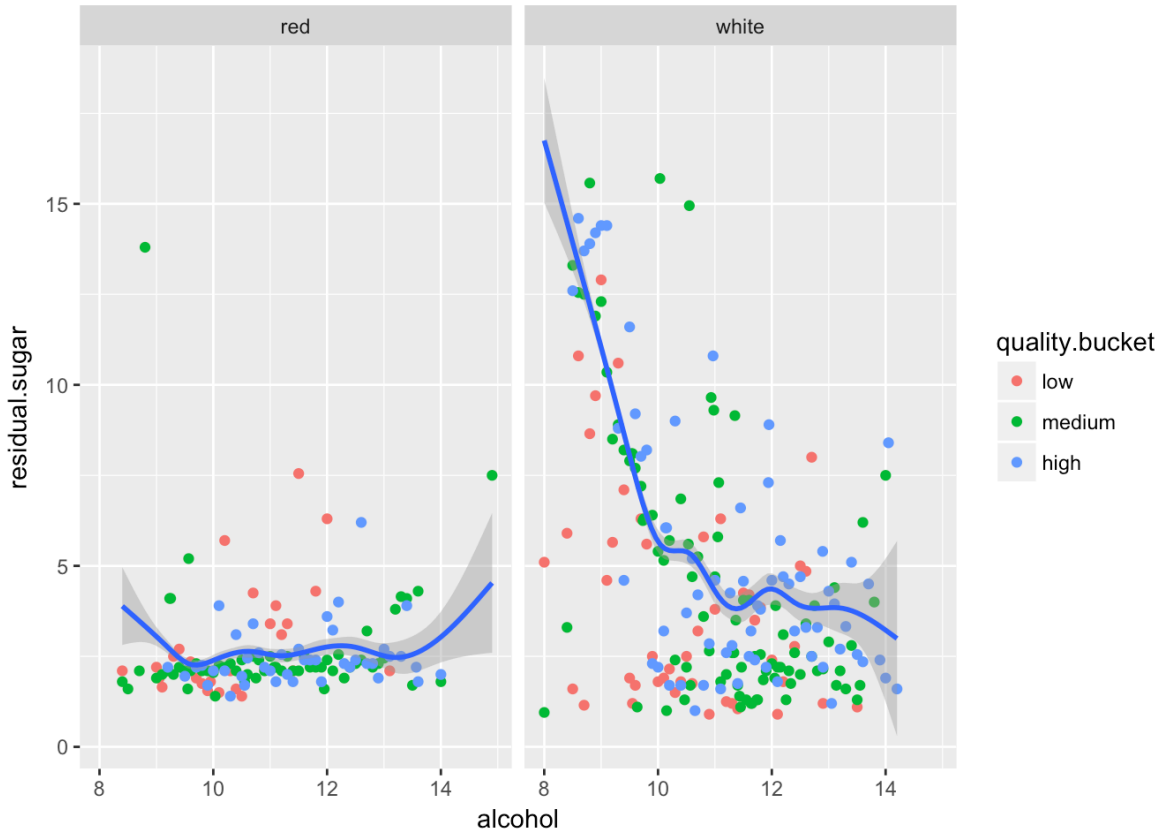
What was the strongest relationship you found?

The strongest relationship I found is between density and residual sugar. It is natural because the addition of sugar increases the liquid density. Another strong relationship I found is between pH and fixed acidity, with increase in ph the fixed acidity levels decrease.

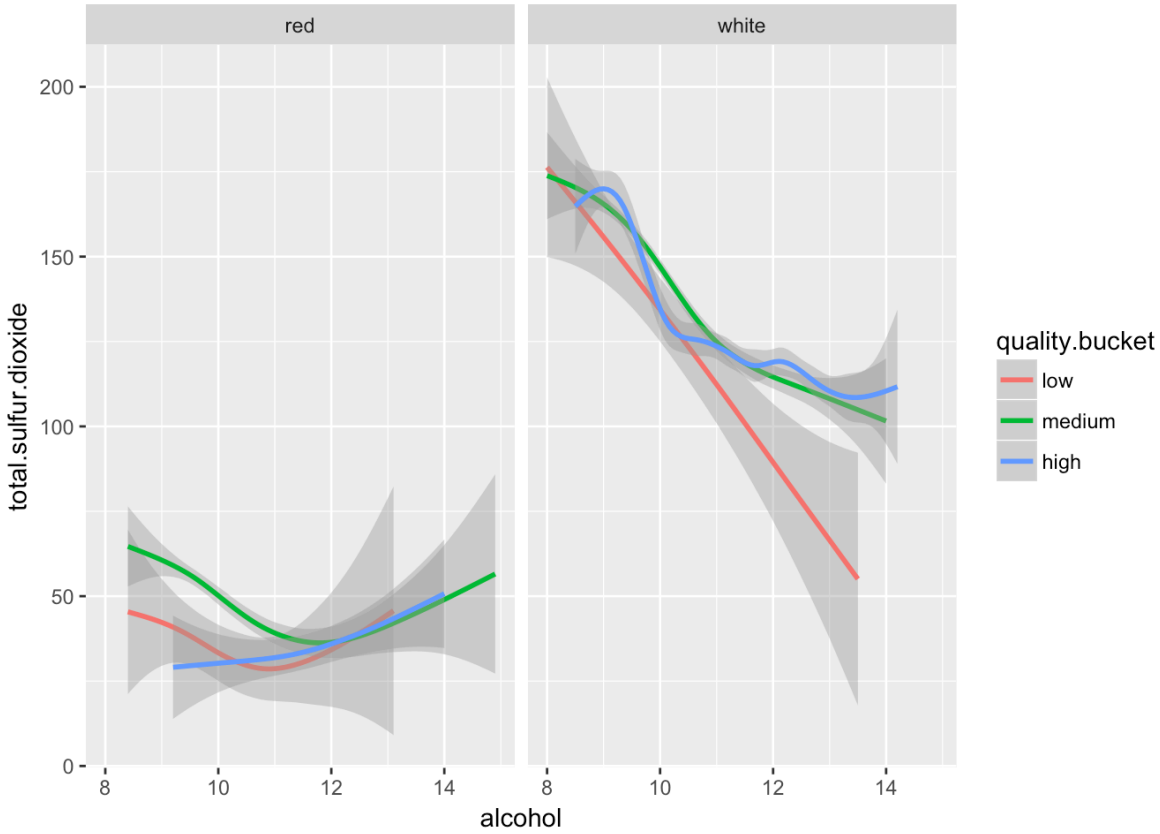
Multivariate Plots Section



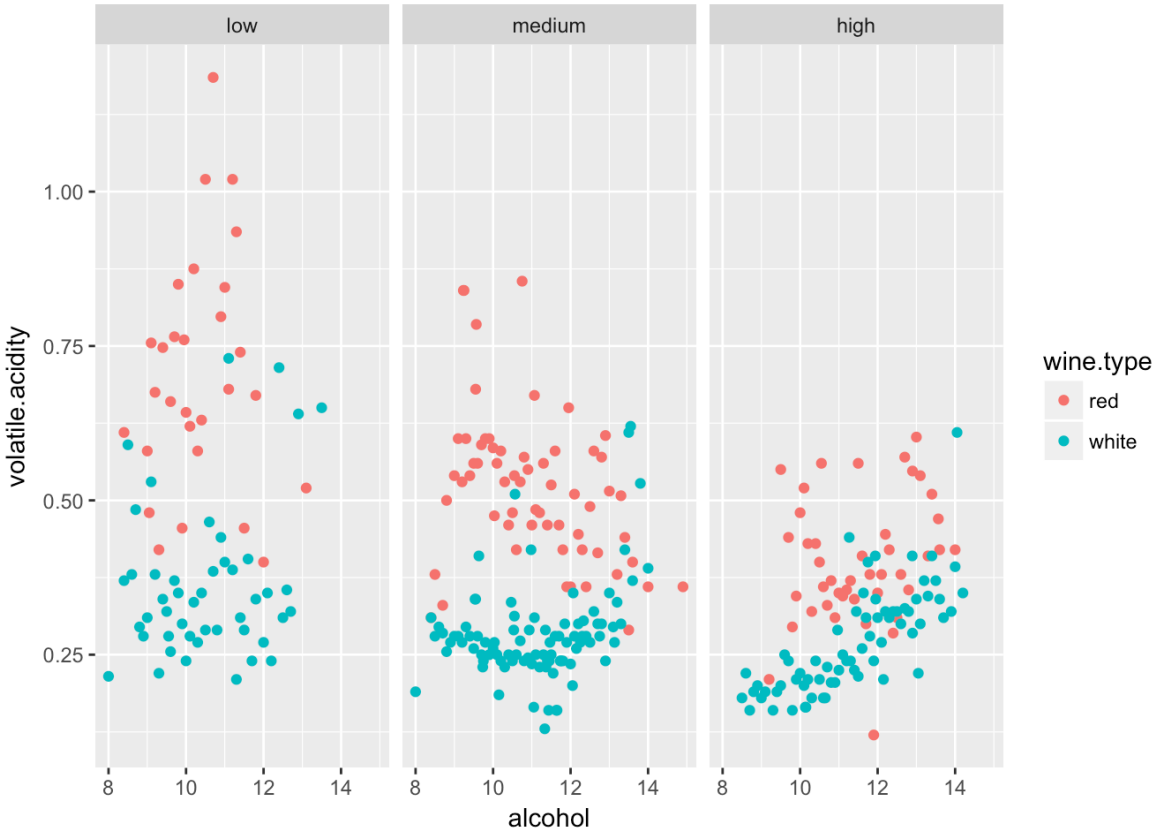
- The data has been divided to three quality buckets. Low quality bucket contains quality ratings of 3 to 4, medium contains quality ratings of 5 to 6, high bucket contains quality rantings from 7 to 9.
- The above plot shows the relation between the quality and alcohol. The high quality bucket has higher alcohol level compared to low and medium quality buckets.



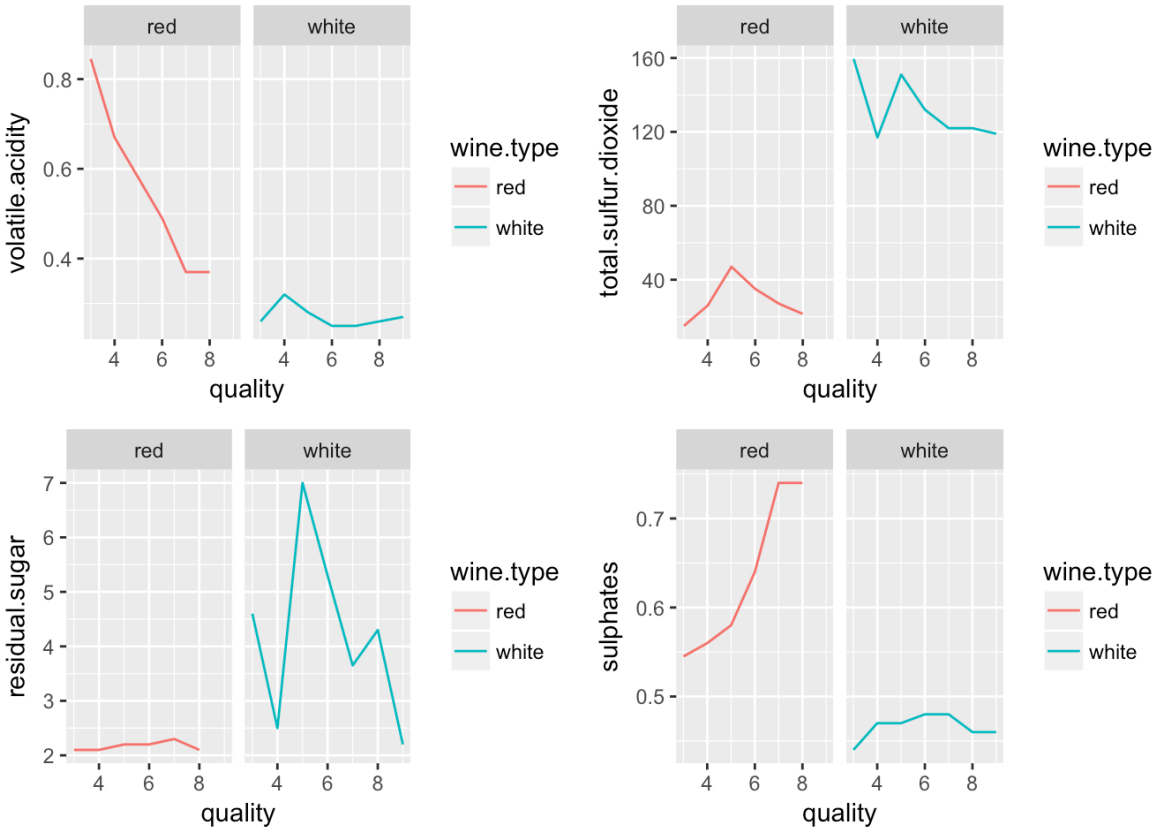
The red wine residual sugar levels doesn't seem to be varying with increase in alcohol levels, where as with white wine the residual sugar levels decreased with increase in alcohol levels.



Total sulfur dioxide levels in red wine are very low compared to white wine. Red wine total sulfur dioxide levels doesnt seem to be increasing or decreasing with alcohol levels. White wine total sulfur dioxide levels are decreasing with increase in alcohol levels.



Volatile acidity is decreasing with increase in alcohol levels. Which makes sense because volatile acidity causes the vinegar taste in wine, so the deccrease in volatile acidity range improves the taste/quality of wine.



From the above plot, it shows that the chemical properties in red wine is different compared to white wine.

- Volatile acidity levels are decreasing with increase in quality for red wine. White wine volatile acidity levels are very low compared to red wine
- Total sulfur dioxide is low in red wine and high in white wine.
- Residual sugar levels are very low and didnt change with increase in quality, where as in white wine there is an increase of sugars at quality level 5 and gradual decrease with increase in quality. Sulphate levels are high in red wine compared to white wine.

```
##
## Call:
## lm(formula = quality ~ alcohol + volatile.acidity + total.sulfur.dioxide +
##     residual.sugar + sulphates, data = wine)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4367 -0.4742 -0.0353  0.4618  3.0432
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            2.4133217  0.1148108  21.020  < 2e-16 ***
## alcohol                0.3388492  0.0084612  40.047  < 2e-16 ***
## volatile.acidity      -1.4581224  0.0626823 -23.262  < 2e-16 ***
## total.sulfur.dioxide -0.0010235  0.0002075  -4.933 8.31e-07 ***
## residual.sugar         0.0235491  0.0023205  10.148  < 2e-16 ***
## sulphates              0.6314479  0.0651345   9.695  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7408 on 6491 degrees of freedom
## Multiple R-squared:  0.2808, Adjusted R-squared:  0.2803
## F-statistic: 506.9 on 5 and 6491 DF,  p-value: < 2.2e-16
```

The variables in this linear model can account for 28% of variance in quality of wines.

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?

Alcohol and qulaity levels are higher for high qualtiy wines compared to low and medium quality wines, which shows that the alcohol is one of the factor for quality of wine. But the quality of the wine cannot be shown by just alcohol level. Residual sugar and total sulfur dioxide decreased with increase in alcohol levels, which affects the quality indirectly. Residual sugar, volatile acidity, sulphates and total sulfur dioxide works in different way for red wine compared to white wine, which is very interesting for further analysis.

Were there any interesting or surprising interactions between features?

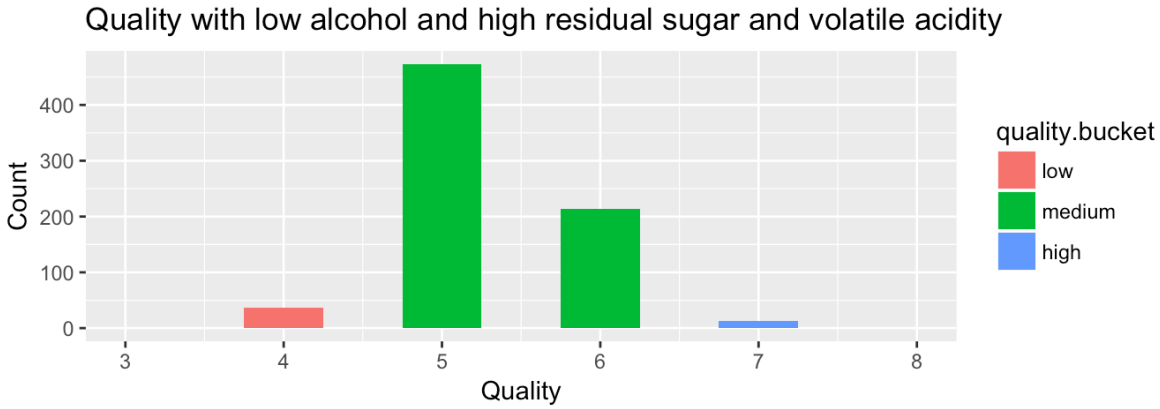
When volatile acidity is plotted against alcohol in bivariate analysis, there was no pattern. But when plotted with facet wrap by quality bucket, we can see a clear reduction of volatile acidity for high quality wines. Another interesting interaction is that red wine and white wine shows different curves for several wine chemical properties which was not I expected.

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

I created a linear model for quality and alcohol and other variables. The variables in the linear model account for 28% of the variance in the quality of wines. The addition of other varialbles increased the variance by 10%. I expected the variance to be high but after reading about the wine in Google, I understood that there are several other factors that effects wine such as the area and the temp of the grapes grown. This data set doesn't contain any information about the grapes or country the wine is from, which will help to understand about the wine quality alot better.

Final Plots and Summary

Plot One



Description One

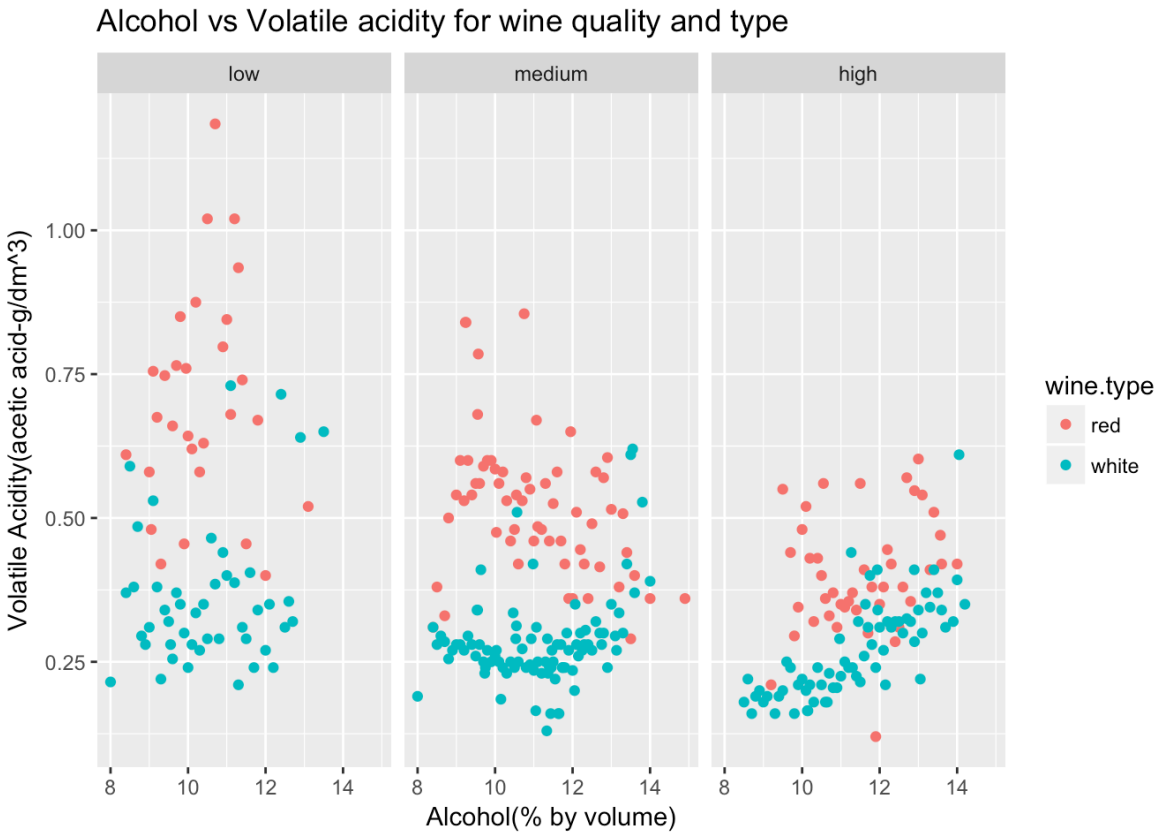
- In the first plot, most of the observations fall between quality levels of 6 and 7. Based on the multivariate plot analysis I chose alcohol levels above the median, residual sugar and volatile acidity values that are below the median as the data for the first plot.
- For second plot most of the observations fall between quality levels of 4 and 5. Based on the multivariate plot analysis I chose alcohol levels below the median, residual sugar and volatile acidity values that are above the median as the data for the second plot.
- Dividing the data based on the above and below the median values shows us that the distribution of quality levels are changing with alcohol, residual sugar and volatile acidity.

Plot Two

Description Two

With just scatter plot between alcohol and quality we dont see any relationship. But by adding median and mean lines, it shows that alcohol level is increasing from quality level 5.

Plot Three



Description Three

- The amount of volatile acidity is low for high quality wine compared to low and medium quality wine.
- Red wine has high amount of volatile acidity compared to white wine.
- There doesn't seem to be any relationship between alcohol and volatile acidity.

Reflection

The data set is very tidy and has several chemical properties about the wine. The main feature I was very interseted in this dataset is quality even though all the other variables have certain affect on wine. And the only variable which has high correlation coefficient on quality is alcohol, which is the second most important variable and has relationship with several other variables, which was shown in bivariate analysis. In multivariate analysis I created quality bucket and used it to facet wrap in alcohol and other variable analysis. In linear model the variables can account for 28% variance of quality in wine. Some limitations in this data set are that we dont have any information about the grapes that are used for making the wine and the location.

References:

1. Example project: https://s3.amazonaws.com/content.udacity-data.com/courses/ud651/diamondsExample_2016-05.html (https://s3.amazonaws.com/content.udacity-data.com/courses/ud651/diamondsExample_2016-05.html)
2. https://rstudio-pubs-static.s3.amazonaws.com/145665_96190c09f2404950ab377c937c120010.html (https://rstudio-pubs-static.s3.amazonaws.com/145665_96190c09f2404950ab377c937c120010.html)
3. https://onlinecourses.science.psu.edu/stat857/node/223 (https://onlinecourses.science.psu.edu/stat857/node/223)
4. http://www.wineskills.co.uk/winemaking/winemaking-knowledge-base/chemical-composition (http://www.wineskills.co.uk/winemaking/winemaking-knowledge-base/chemical-composition)

5. https://www.researchgate.net/publication/276444447_Chloride_concentration_in_red_wines_Influence_of_terroir_and_grape_type (https://www.researchgate.net/publication/276444447_Chloride_concentration_in_red_wines_Influence_of_terroir_and_grape_type)
6. http://data.princeton.edu/R/linearModels.html (http://data.princeton.edu/R/linearModels.html)