EDA of Red Whine Data Set by Padmaraj

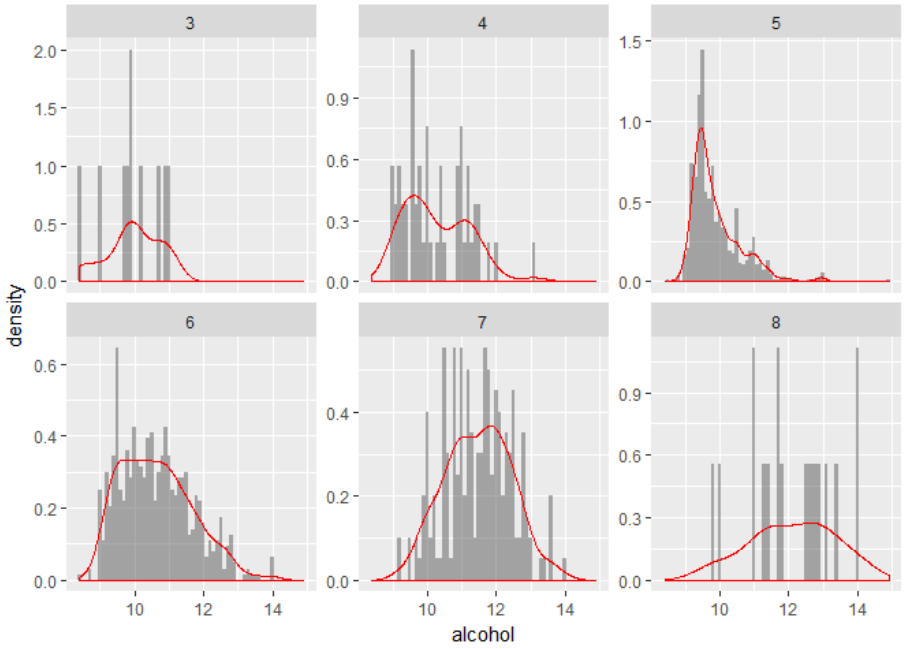
Abstract:

This tidy data set contains 1,599 red wines with 11 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).

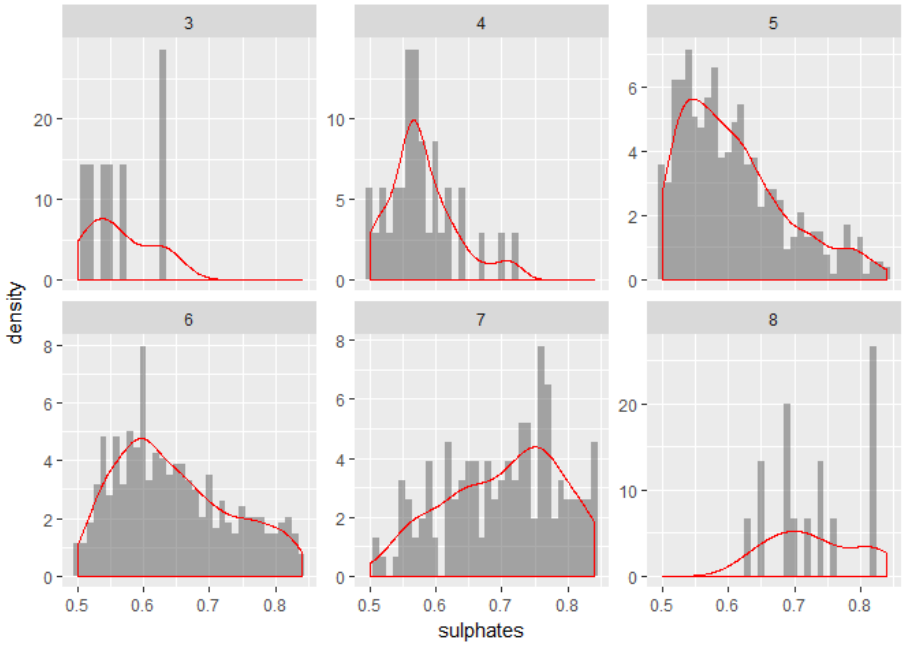
Univariate Plots Section

In the below Density plots, variables indicates positive or negative influence of on the 'quality' variables. It is evident from the below univariate plots that Sulphates, alcohol & fixed.acidity variables have positive influence i.e. increase in these variable value increases the value for the quality variable. We can see the hump like structure marked in red) moving to the right side of the below plots.

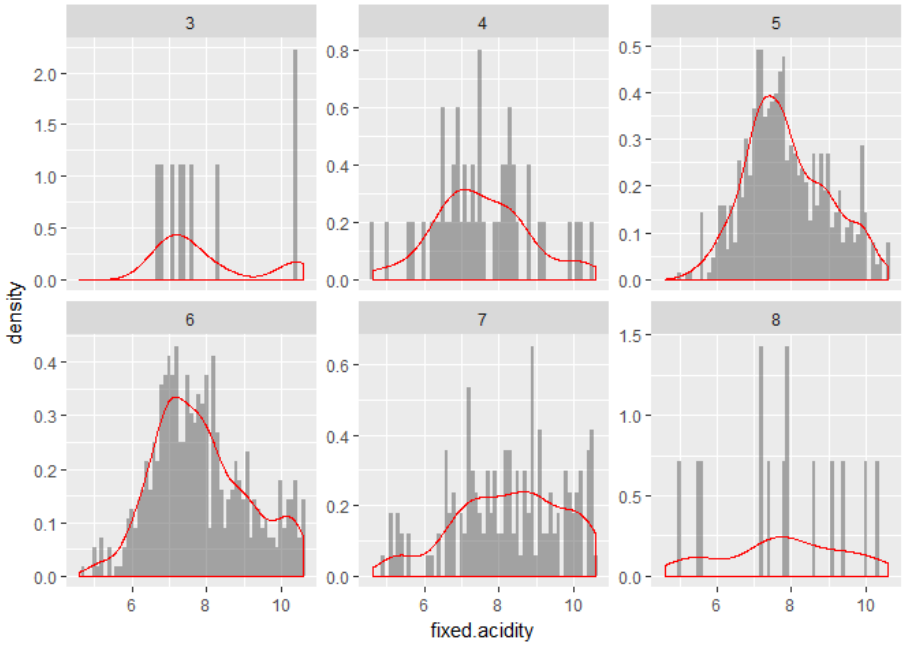
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.50   10.20   10.42   11.10   14.90
```



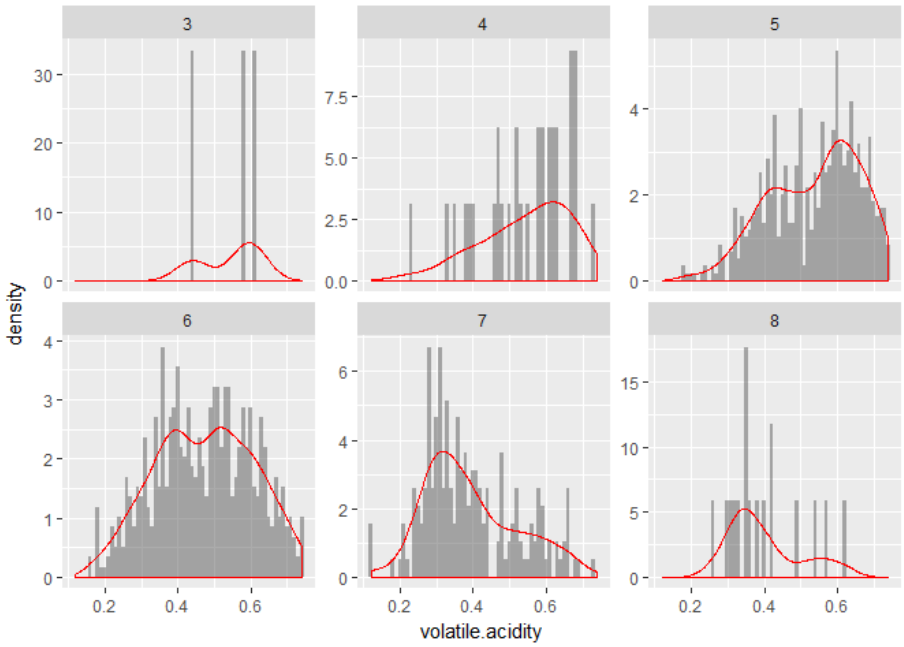
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3300  0.5500  0.6200  0.6581  0.7300  2.0000
```




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

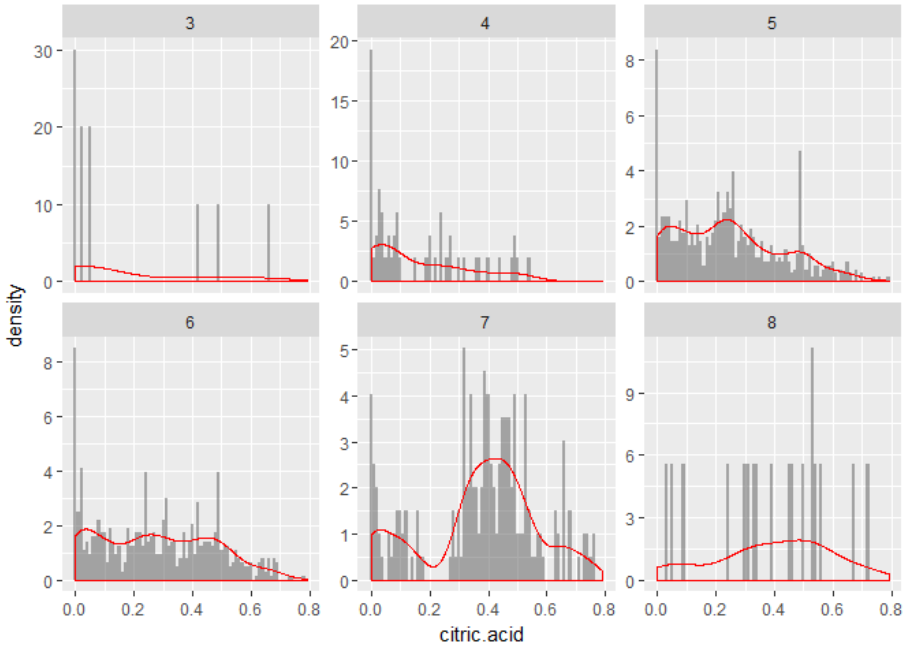



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1200  0.3900  0.5200  0.5278  0.6400  1.5800
```

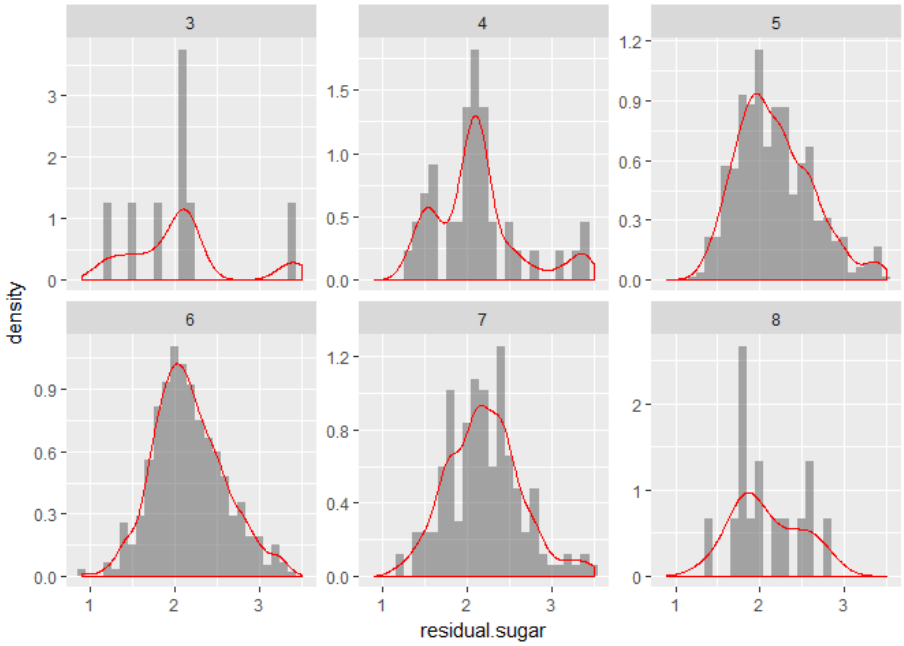


```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.090   0.260   0.271   0.420   1.000
```

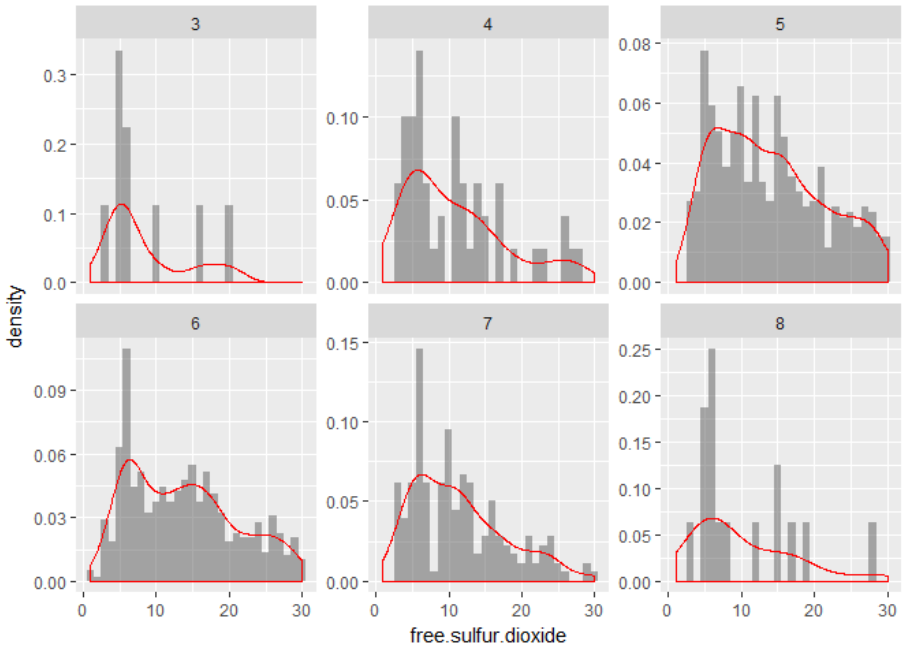
```
## Warning in citric.acid < quantile(0.9): longer object length is not a
## multiple of shorter object length
```



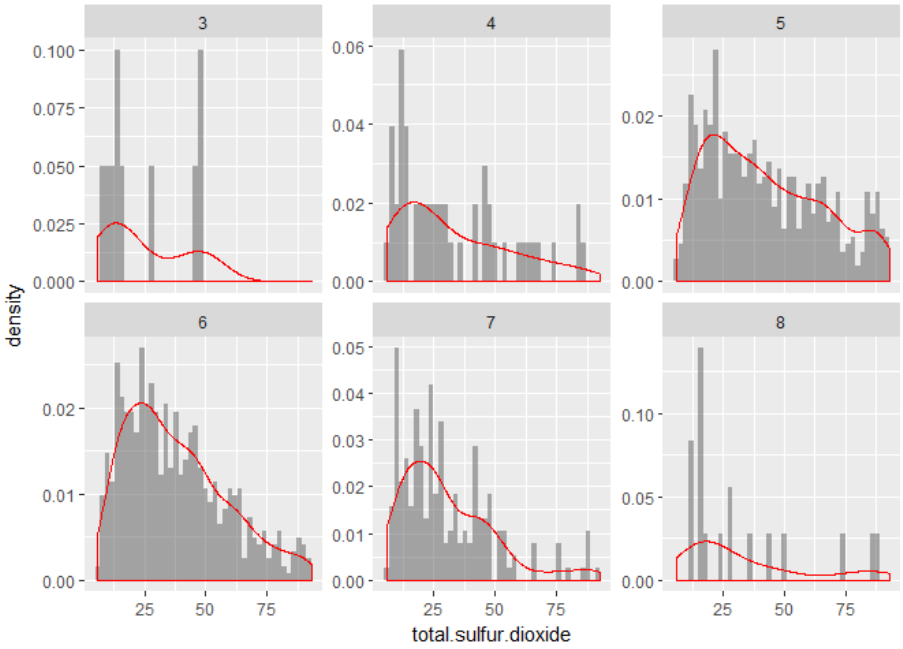

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.900   1.900   2.200   2.539   2.600  15.500
```



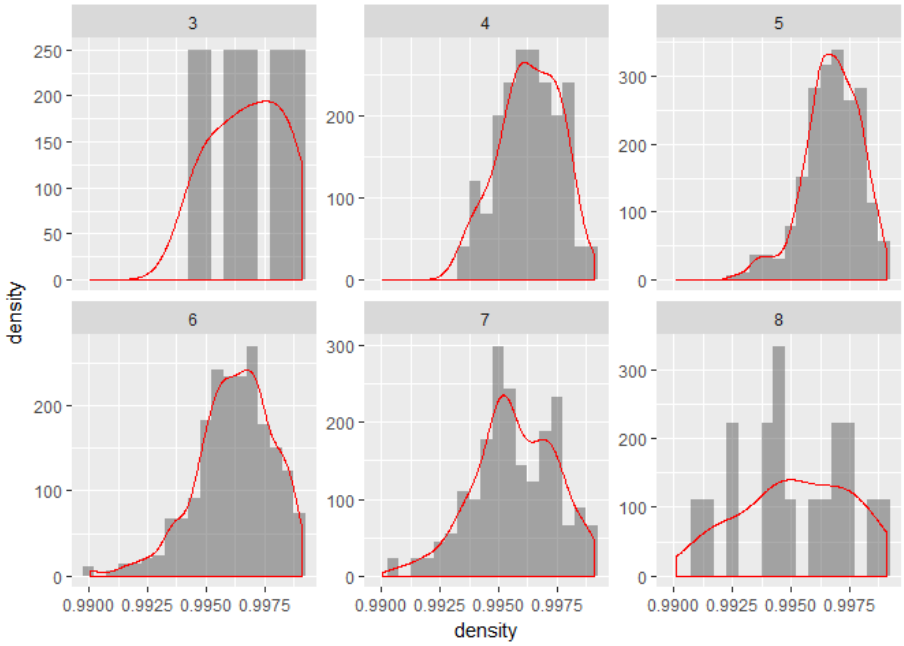

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    7.00   14.00   15.87   21.00   72.00
```



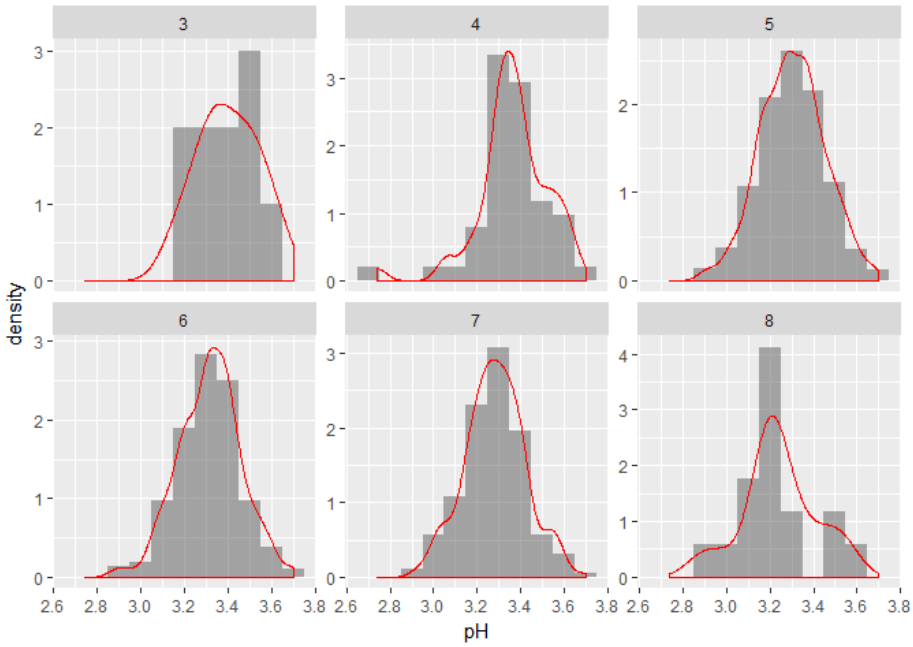
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00   22.00   38.00   46.47   62.00  289.00
```




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




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



Univariate Analysis

What is the structure of your dataset?

The Dataset has 13 variables; out of which first one is just the serial number and can be omitted. There are no categorical variables. However, quality can be viewed as both categorical or continuous numerical values.

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

Main feature of interest is 'quality'. The task is to identify the right combination available variables to come up with the formula to have the higher quality certificate from the expert's review.

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

Density histogram indicates the relation between the quality and variable. Couple of variable show clear sign of impact to the quality; while other show equal impact on the different quality category.

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

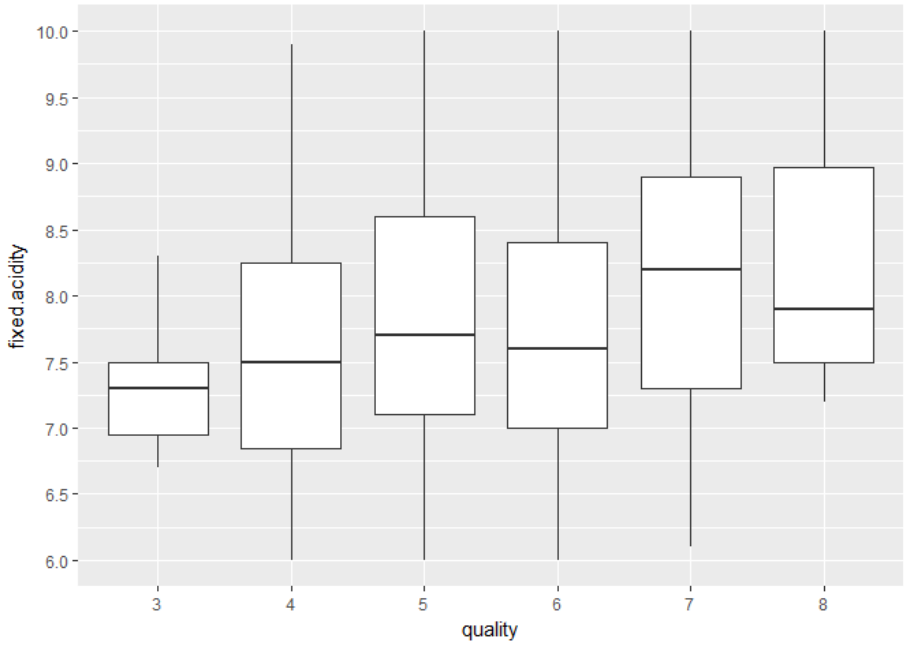
No. There were no new variables created.

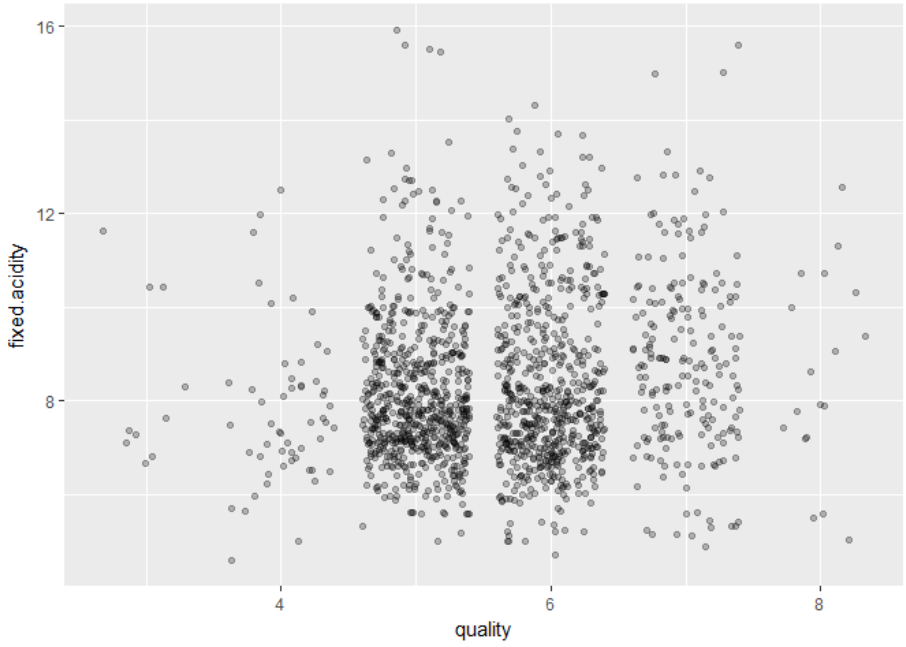
**Of the features you investigated, were there any unusual distributions?
Did you perform any operations on the data to tidy, adjust, or change the form
of the data? If so, why did you do this?**

I have changed the type of 'quality' variable to factor. There are 2 reasons for the same. One is to be able to use as categorial variable to identify the impact of other variables on it through the use of facet_wrap function.

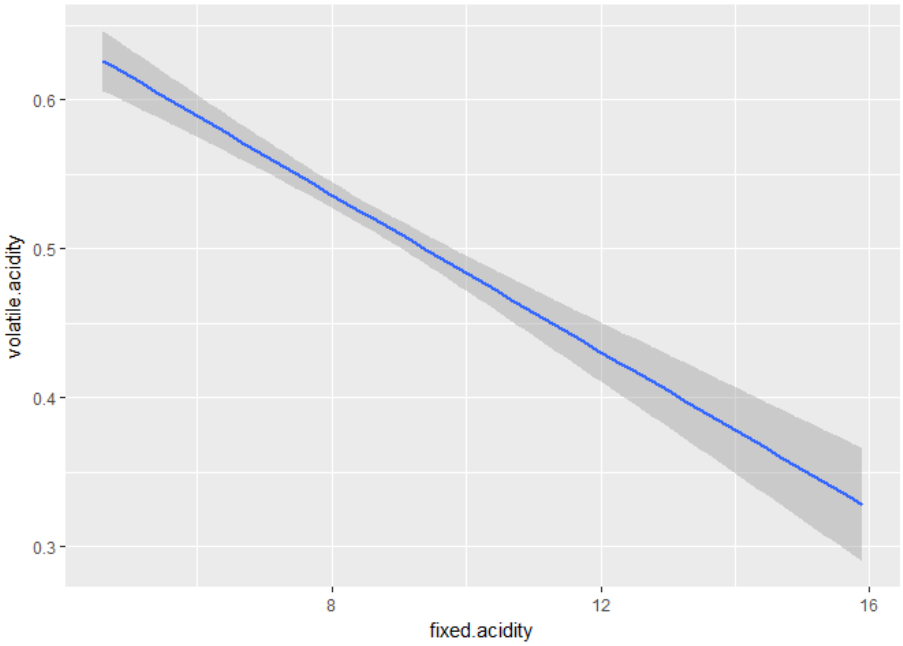
The other reason is to clearly be able to draw boxplot. Boxplot allows not only to scope of the variable values but also the marks the mean of variable values which help in comparing with other categorical values.

Bivariate Plots Section

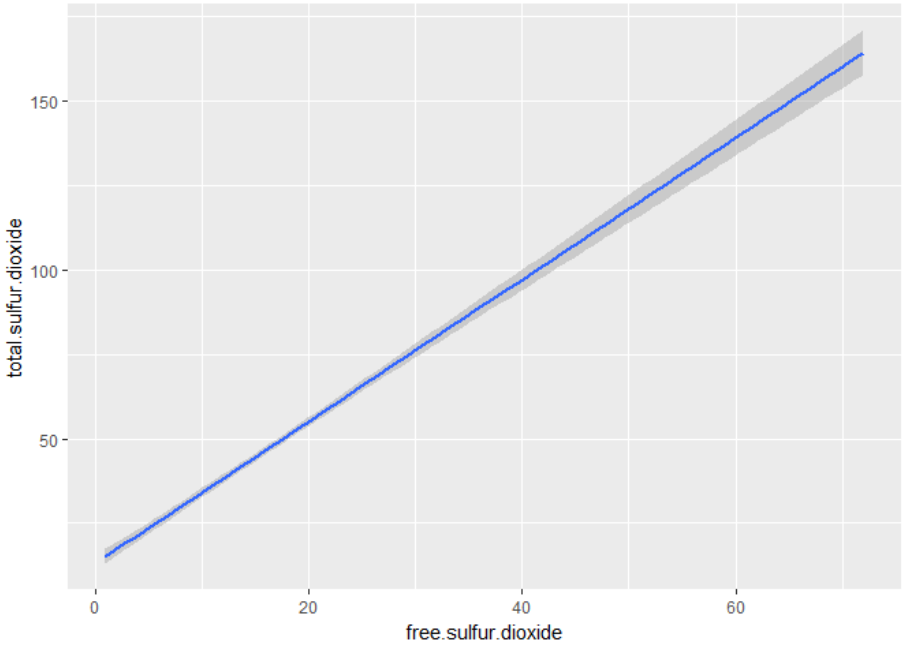


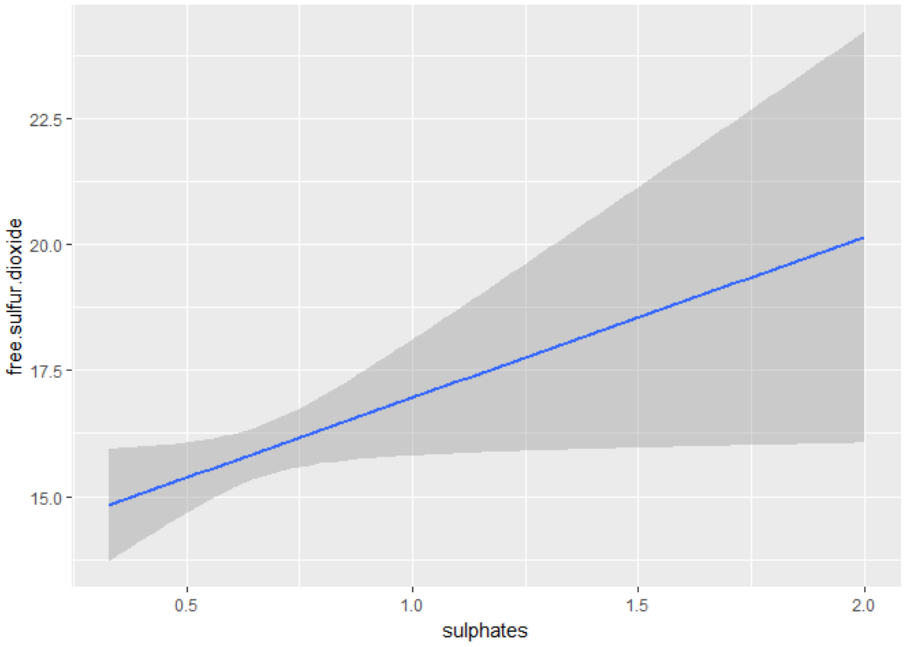


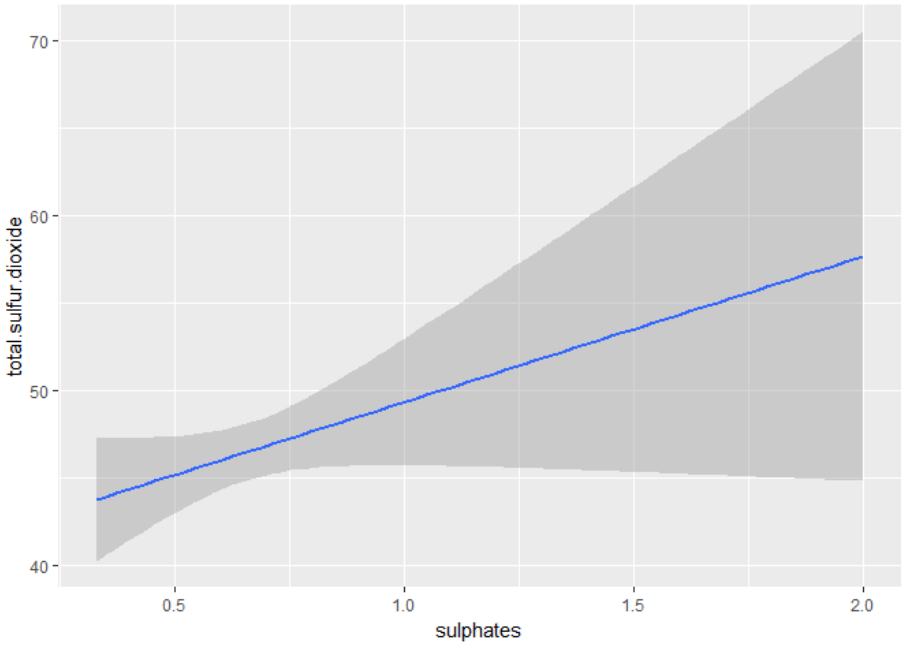
Factoring of quality helps us to understand more precisely impact of acidity on qulaity. In the boxplot above, we can see the different median values more clearly than the jitter plot indicating the density.



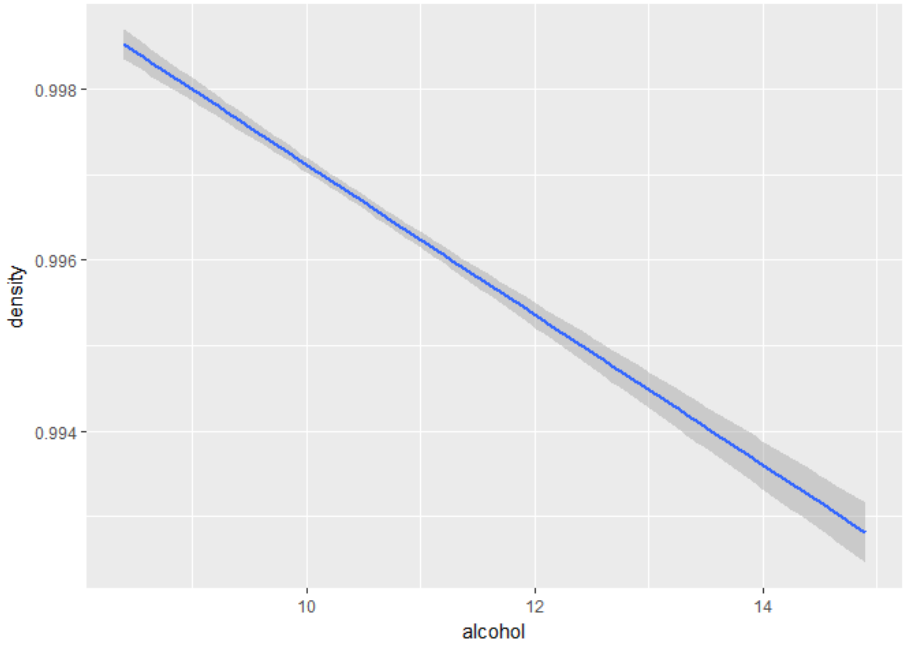
Clearly the above plot indicates that 2 acidities are mutually exclusive.





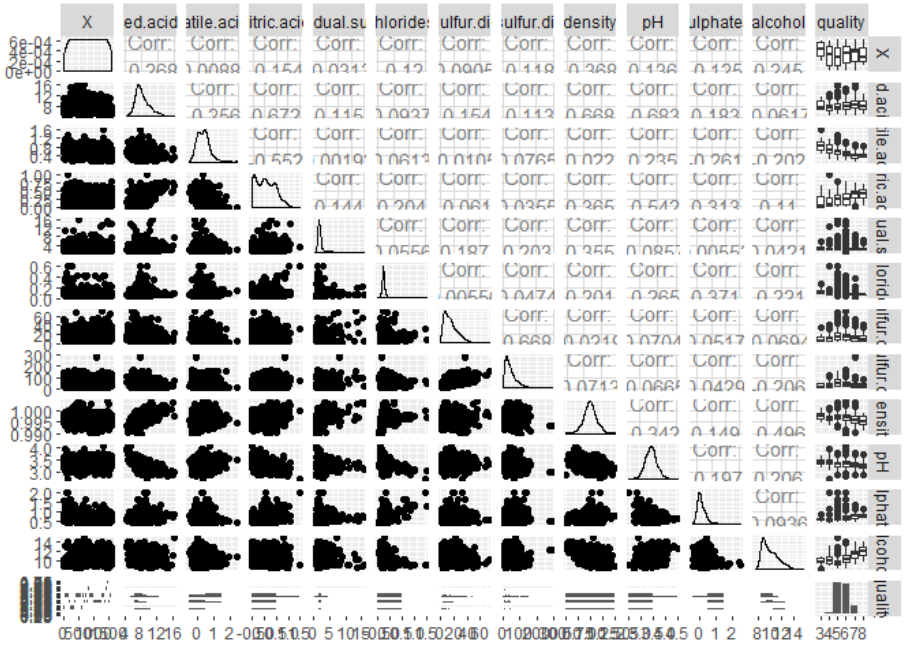


Free sulfur and total sulfur complement each other. Increase in one variable is directly proportional to other. Intutively, increase in sulphates



increases dioxides.

The above plot indicates that density and alcohol are inversly proportional.



Let us observe couple of visible relation between all the variable in the above plot. Couple of relation will be discussed in the below section.

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 analysis of couple of features indicates that histogram correctly identfies the relation between the quality and the other variables values.For instance, Alcohol and density relation clearly complements the observation we had on the respective Univariable analysis.

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

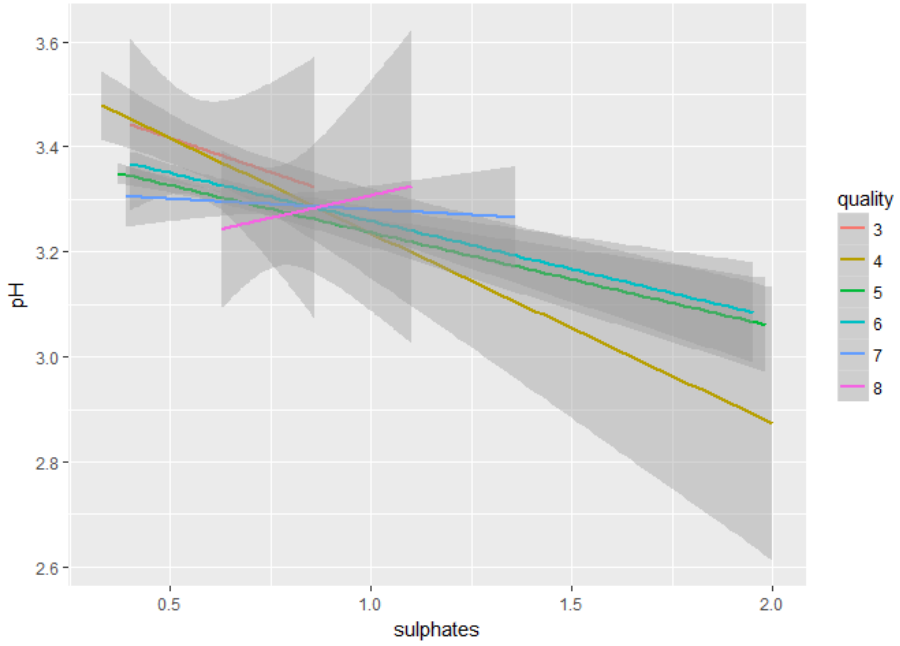
There are couple interesting relationship which complements the co-relation value between the two can also be easily seen from the ggpair plot.

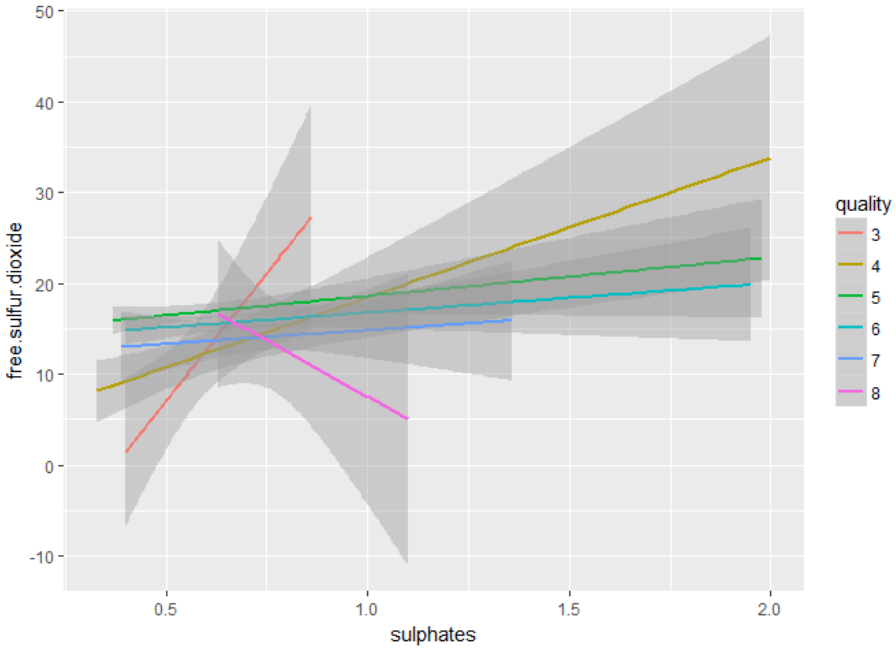
Fixed acidity vs citric.acid , fixed acidity vs density, fixed acidity vs ph are few examples which shows their relationship easily in the plot. They almost fall on a linear curve. This in simple would mean that we have redundant feature when it comes to modelling.

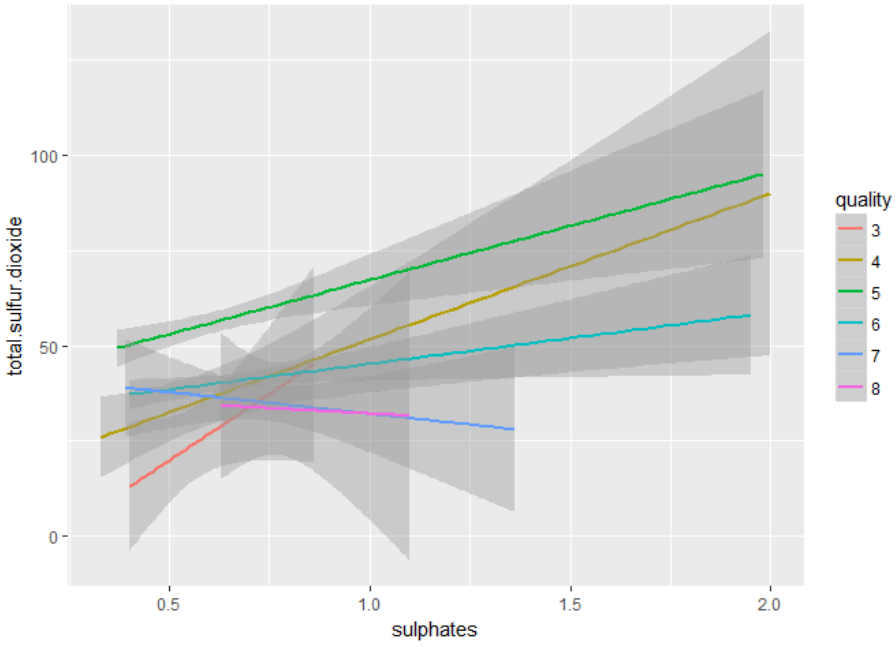
What was the strongest relationship you found?

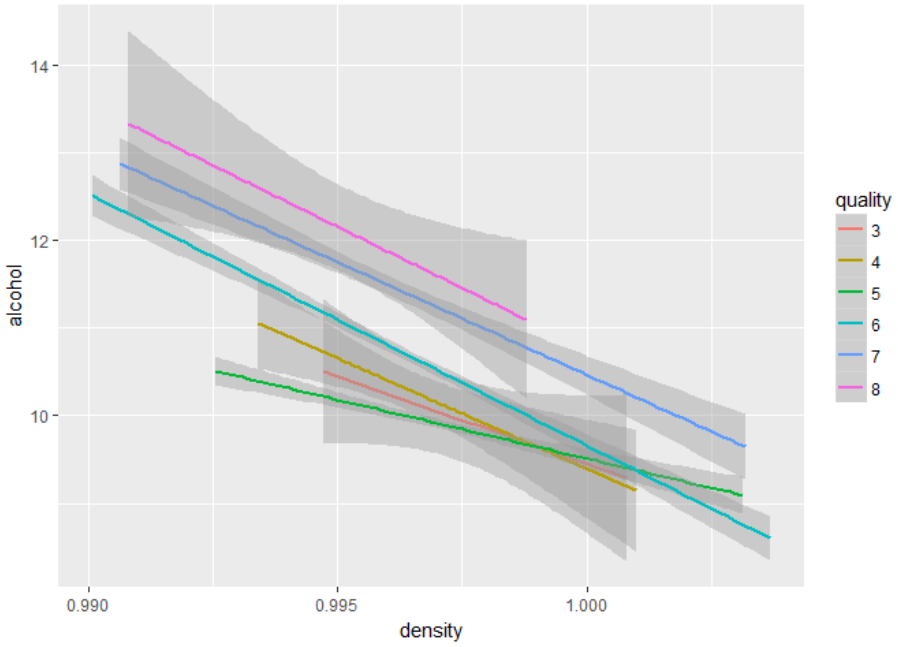
Fixed Acidity vs Citric Acid has the stongest co relation value.

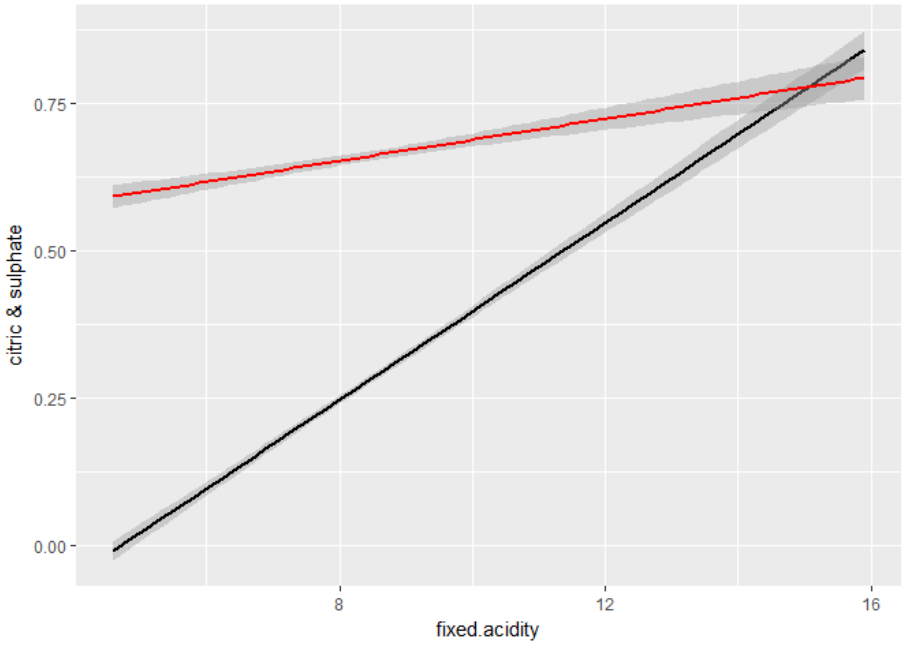
Multivariate Plots Section




50
40
30
20
10
0
-10
free.sulfur.dioxide
0.5
1.0
1.5
2.0
sulphates
quality
3
4
5
6
7
8


100
50
0
total.sulfur.dioxide
0.5
1.0
1.5
2.0
sulphates
quality
3
4
5
6
7
8


14
12
10
alcohol
0.990
0.995
1.000
density
quality
3
4
5
6
7
8


0.75
0.50
0.25
0.00
citric & sulphate
8
12
16
fixed.acidity

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?**

Plotted relationship between sulphates vs (dioxides, ph and acidities). It was based intution that suphate might impact to pH through formation of sulphuric acid and hence also on di oxides. We see that increase in the sulphate actually makes it acidic in most of the cases. But higher quality whines actually maintains acidity.

Alcohol vs density show inverse relationship irrespective of quality.

Contributor to Fixed acidity can also be easily seen in the 5th plot.Citric acid indicated through black line and Sulphates in red line. Both indicate that increase in values increases fixed acidity.

Were there any interesting or surprising interactions between features?

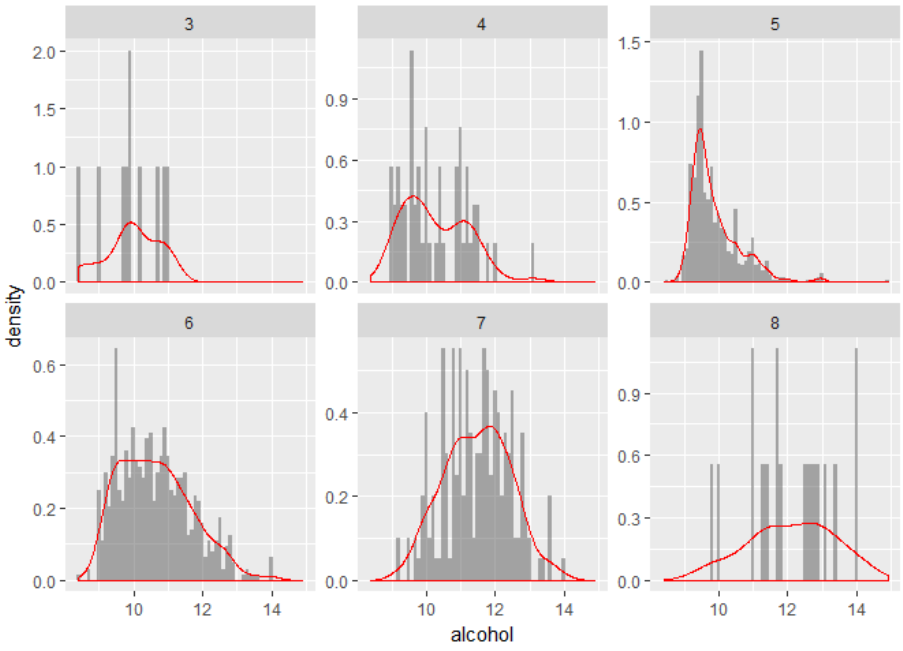
Sulphates vs free.sulfur.dioxide shows almost orthogonal variation between quality equals to 3 and quality equals to 8.

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

I wrote a python program to build a decision tree regression model. The model has very low accuray(28-30%). The model feature_importances_ also gave similar conclusion as that of above in regards to the variables impacting the quality.

Final Plots and Summary

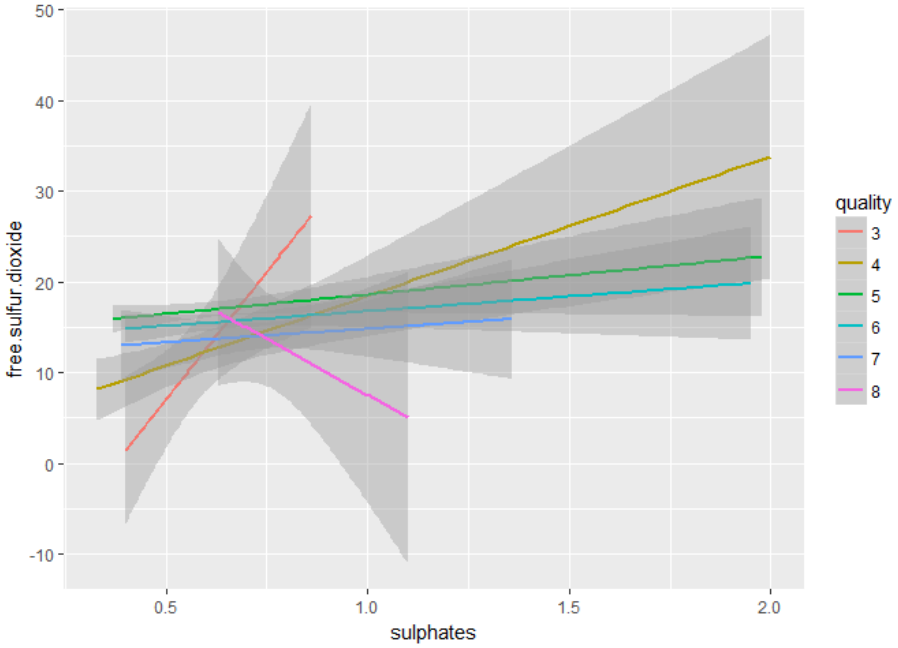
Plot One



Description One

The above model clearly shows the variation of the alcohol variable on the quality. I choose to be the most important because the variable shows clear variation across quality.

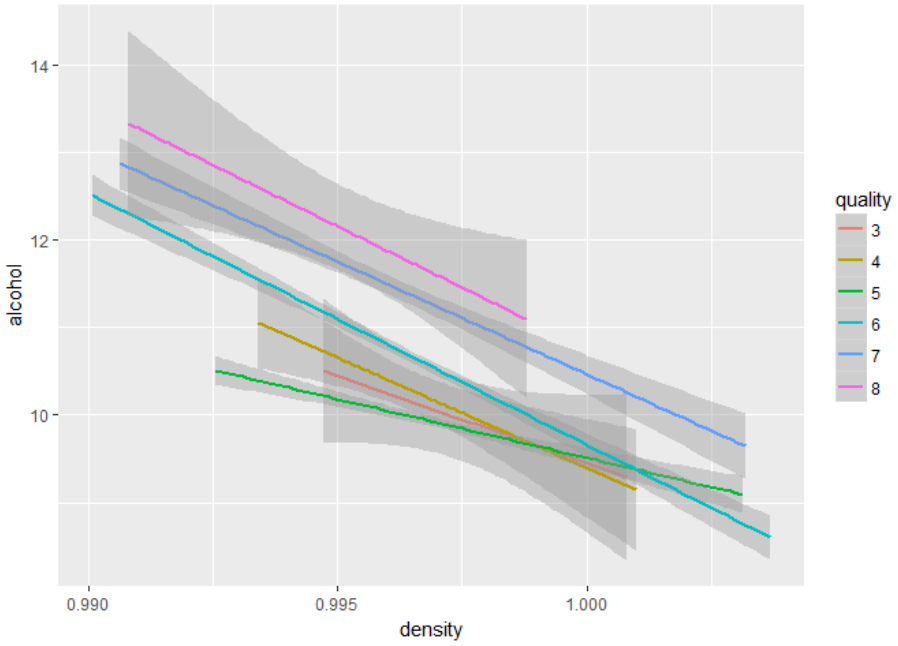
Plot Two



Description Two

Though geom_smooth in the bi variable plot indicates that free sulfur dioxide to be directly proportional to sulphates. Multivariate plot indicates that when devided into quality the influence of sulphates on quality as the quality increases. We almost see a clockwise rotation; i.e. for quality = 3 we see that the line can be seen as pointing to 2'o clock and when quality = 8 then we may see the line plot as indicating 4'o clock.

Plot Three



Description Three

The only 2 variables which show consistent variation over different "quality"

Reflection

The intension of the EDA on red whine data set was to check if it can be used to model for a quality prediction.

There are 11 variables but still as most of them are co related with each other,the only few are useful for modelling.

Even if we have 1599 observations, the observations are not uniformly distributed with respect to quality. As we have very less observation relating high quality red whines very fine quality whines might be classified as outlier in our model.

Need more quality (different variables) and quantity (homogeneous mixture of "quality" variables.) to build a model to predict quality of the red whine.