Wine Quality Exploration by Shanshan Wang

> **Introduction**: 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

Check the basic information of the dataset

Dimension

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

Structure

```
## 'data.frame':    1599 obs. of  12 variables:
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
##  $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
##  $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071 ...
##  $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
##  $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
##  $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
##  $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...
##  $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...
##  $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
##  $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...
```

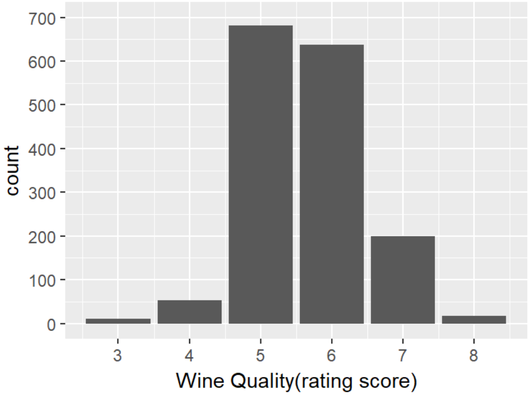
Summary of each variable

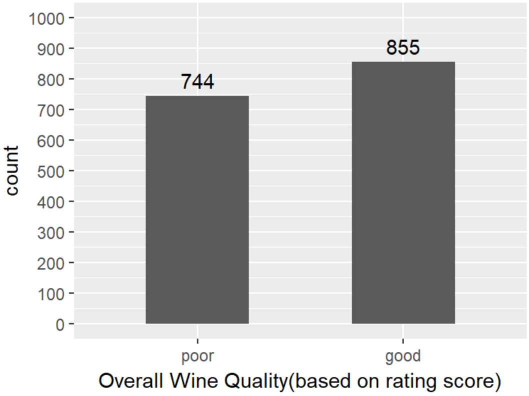
```
##  fixed.acidity   volatile.acidity  citric.acid    residual.sugar  
##  Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900  
##  1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900  
##  Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200  
##  Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539  
##  3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600  
##  Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500  
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide
##  Min.   :0.01200   Min.   : 1.00       Min.   :  6.00      
##  1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00      
##  Median :0.07900   Median :14.00       Median : 38.00      
##  Mean   :0.08747   Mean   :15.87       Mean   : 46.47      
##  3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00      
##  Max.   :0.61100   Max.   :72.00       Max.   :289.00      
##     density             pH          sulphates         alcohol     
##  Min.   :0.9901   Min.   :2.740   Min.   :0.3300   Min.   : 8.40  
##  1st Qu.:0.9956   1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50  
##  Median :0.9968   Median :3.310   Median :0.6200   Median :10.20  
##  Mean   :0.9967   Mean   :3.311   Mean   :0.6581   Mean   :10.42  
##  3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10  
##  Max.   :1.0037   Max.   :4.010   Max.   :2.0000   Max.   :14.90  
##     quality     
##  Min.   :3.000  
##  1st Qu.:5.000  
##  Median :6.000  
##  Mean   :5.636  
##  3rd Qu.:6.000  
##  Max.   :8.000  
```

Variable names

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

Quality

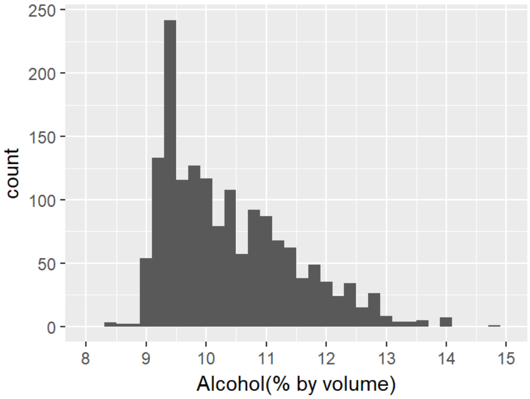


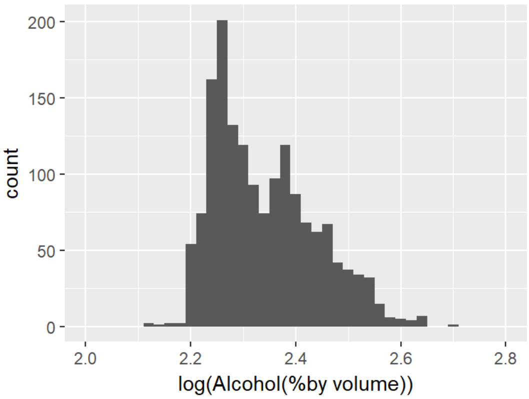


This dataset has recoreded rating scores of 1599 red wines, most of the score in the range of 5~7. Among these 1599 wines, 744 of them with a rating score equal or less than “5”, they are denoted as “poor” quality wines. Others with a rating score higher than “5”, are denoted as “good” quality wines.

Check other variables

Alcohol



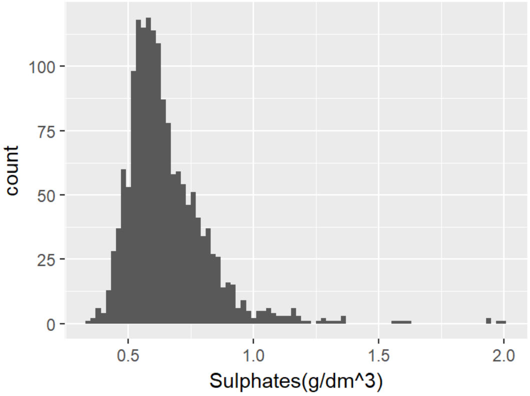


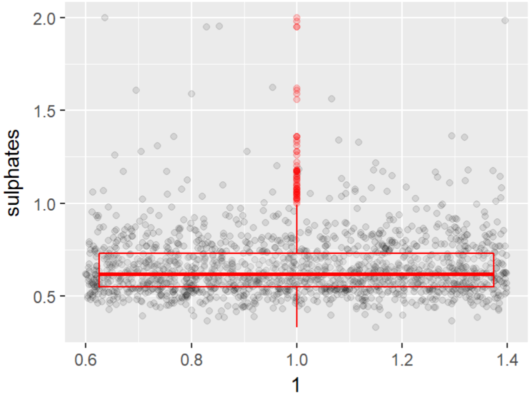
Histograph shows the distribution of Alcohol is highly right skewed, so I did a log tranform on alcohol data, but this only improved the alcohol distribution a little.

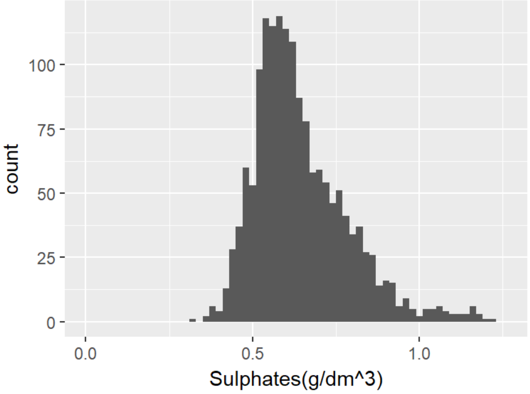
Summary of Alcohol

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

Sulphates





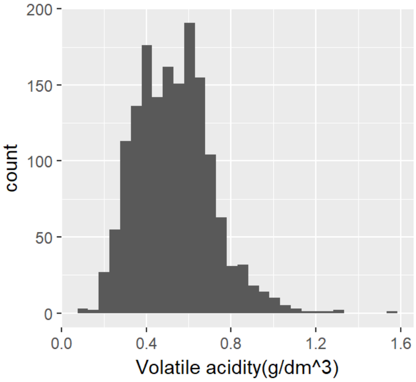


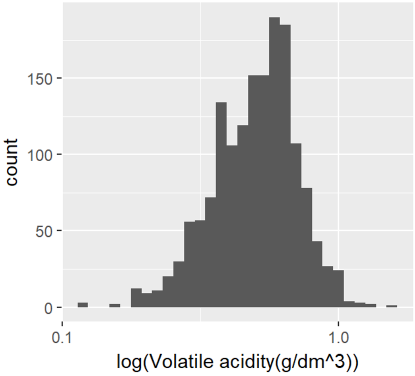
From the plot, I found there were some outliers in sulphates data. So I excluded these outliers and made the histograph again which shown a normal distibution of sulphates.

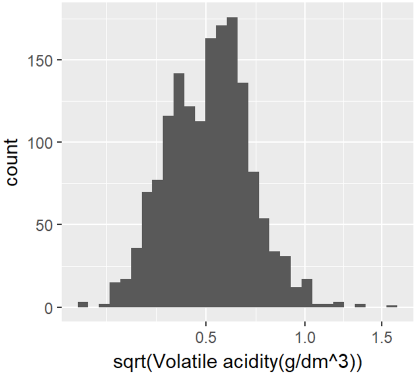
Summary of Sulphates

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

Volatile.acidity





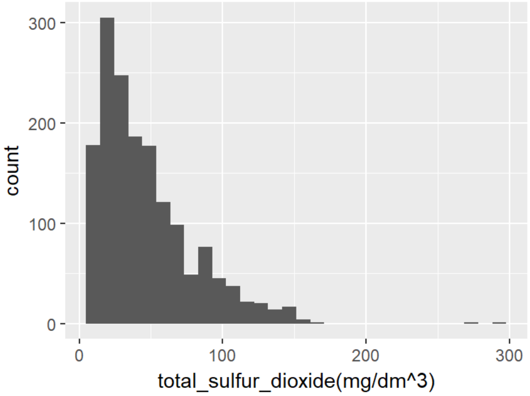


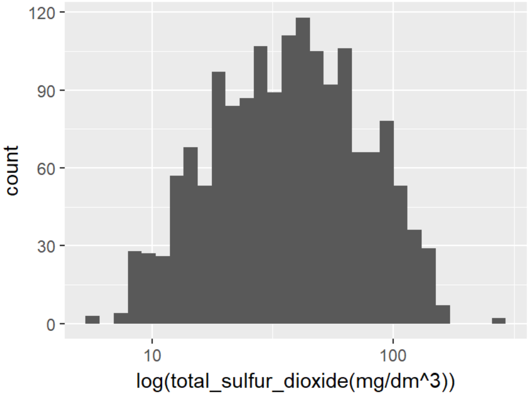
The distibution of volatile.acidity data is also right skewed. So I tried two data transformation ways such as log and sqrt, the histographs indicated sqrt transformation is better.

Summary of Volatile.acidity

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

Total.sulfur.dioxide



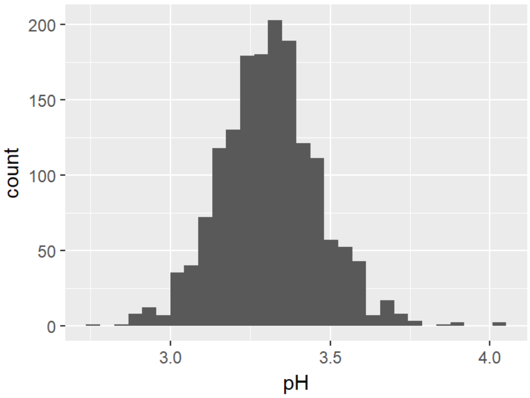


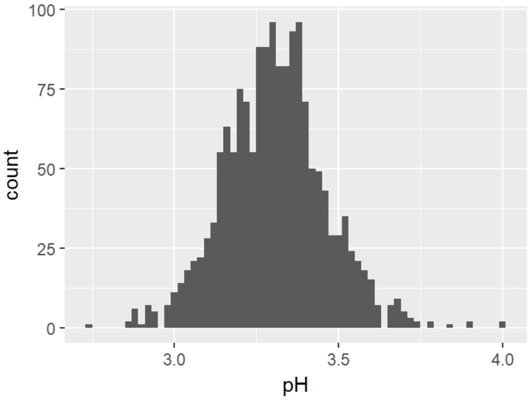
The distibution of total.sulfur.dioxide data is again right skewed. After a log transformation, it shown a normal distribution.

Summary of Total.sulfur.dioxide

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

pH value





pH data is in the shape of normal distribution. I tried different bins.

Summary of pH

```
##    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?

There are 1599 wines in this dataset with 11 features(fixed.acidity,volatile.acidity, citric.acid,residual.sugar,chlorides,free.sulfur.dioxide,total.sulfur.dioxide,density,pH,sulphates, alcohol,quality). All variables are numeric ones.

Among them:
1. Quality is in range 3-8 (rating score according to the background of dataset), the number of wines with a rating score equal or lower than 5 is almost the same as the number of wines with a rating score higher than 5.
2. Alcohol is in range 8.4-14.9(% by volume).
3. pH value is in range 2.74-4, which means these wines are very acidic.

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

The main feature of interest in this datase is quality. I'd like to determine which features are best to predict the quality of a wine, I suspect alcohol, sulphates and other variables combination can be used to build a predictive model according to the correlation plot.

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

Alcohol, sulphates, volatile.acidity are vely likely to contribute to the quality of a wine, and also pH value, residual.sugar even density may affect the wine quality.

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

I created a new variable called overall, which is the overall quality of wines. This value has two values, “poor” coorespones to wines rating lower than or equal to 5, and “good” are these with rating higher than 5.

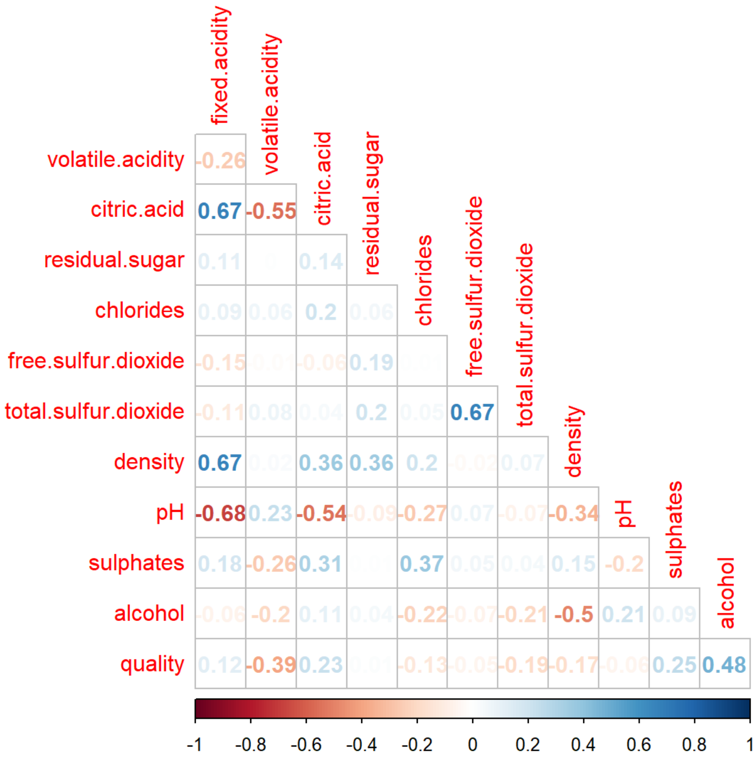
Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

I did a log-transformation to alcohol and total.sulfur.dioxide data, since they are highly right skewed.Tranformed data showed a better likely normal distibution. I did a sqrt-transformation to volatile.acidity.

Also when I was plotting sulphates data, I excluded the top1% data, the rest of the sulphates data has a nice normal distribution.

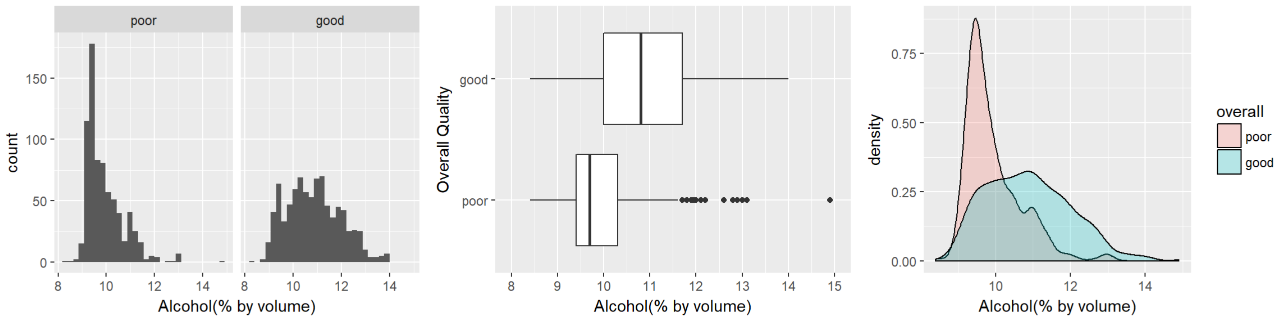
Bivariate Plots Section

Check the correlation values between each couple of numerical features.



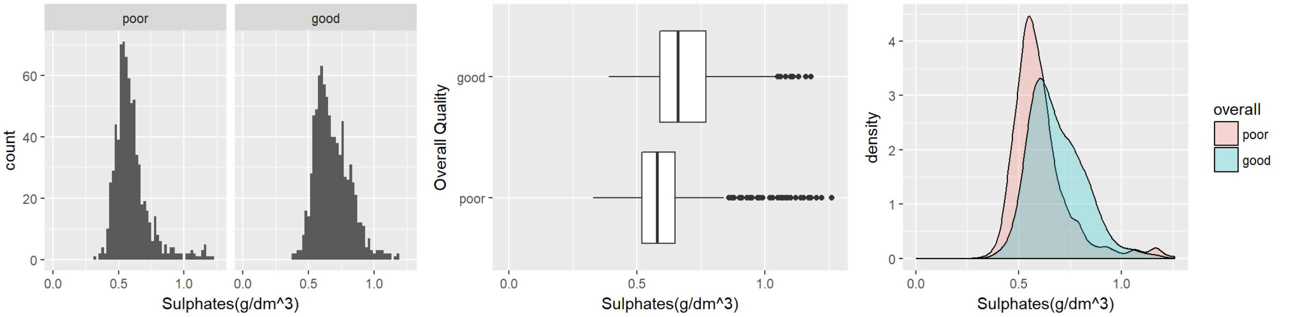
I applied a correlation plot for these data to check the correlation between each pair of numeric data. The plot indicates alcohol data has the highest correlation to quality.

Alcohol of wines in different qualities



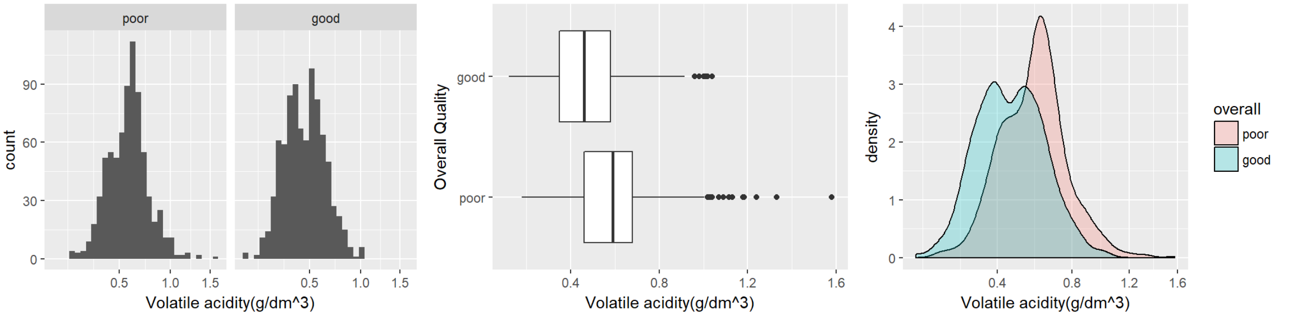
I checked alcohol data in wine groups with differnt qualities(good and poor) and found the good quality group has the higher median of alcohol. The distribution of alcohol is highly right skewed in poor quality wines while it is normal distributed in good quality wines.

Sulphates of wines in different qualities



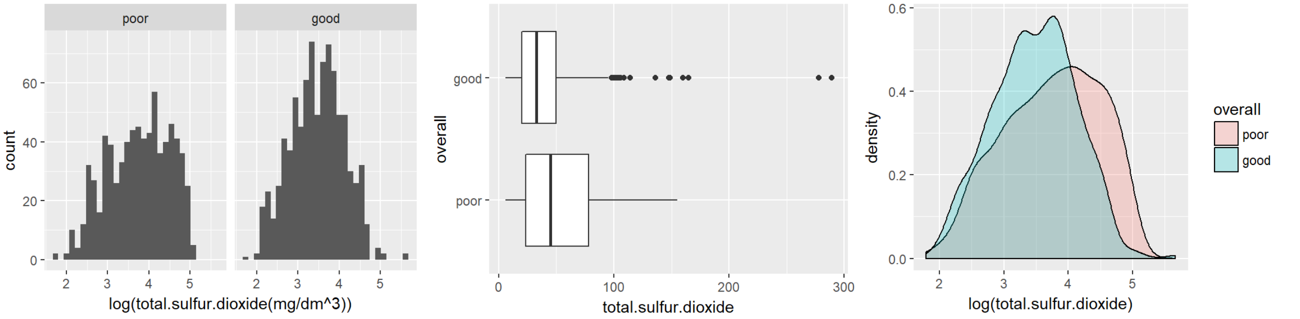
I checked sulphates data in wine groups with differnt qualities(good and poor) and found the good quality group has the higher median of sulphates.

Volatile acidity of wines in different qualities



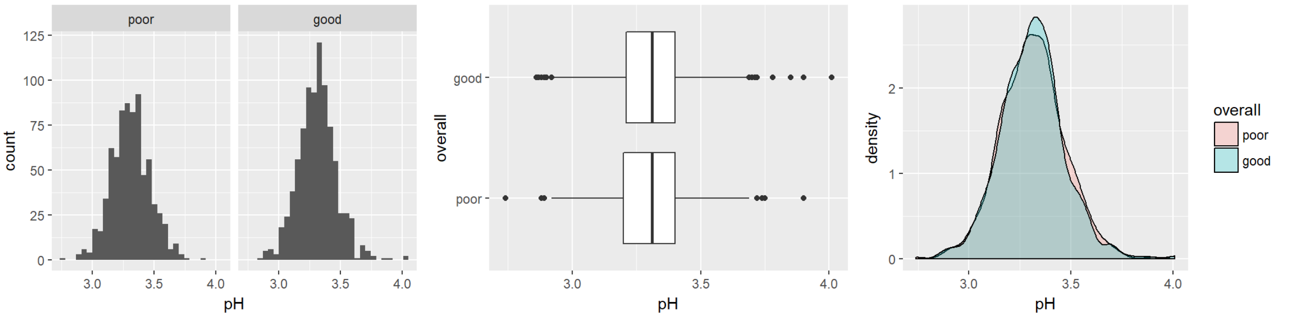
I checked volatile.acidity data in wine groups with differnt qualities(good and poor) and found the good quality group has lower median of volatile.acidity.

Total.sulfur.dioxide of wine in different qualities



I checked total.sulfer.dioxide data in wine groups with differnt qualities(good and poor) and found the good quality group has a slightly lower median of total.sulfer.dioxide.The density plot shown the peak of total.sulfer.dioxide is a litte bit larger in poor quality wines than it is in good quality wines.

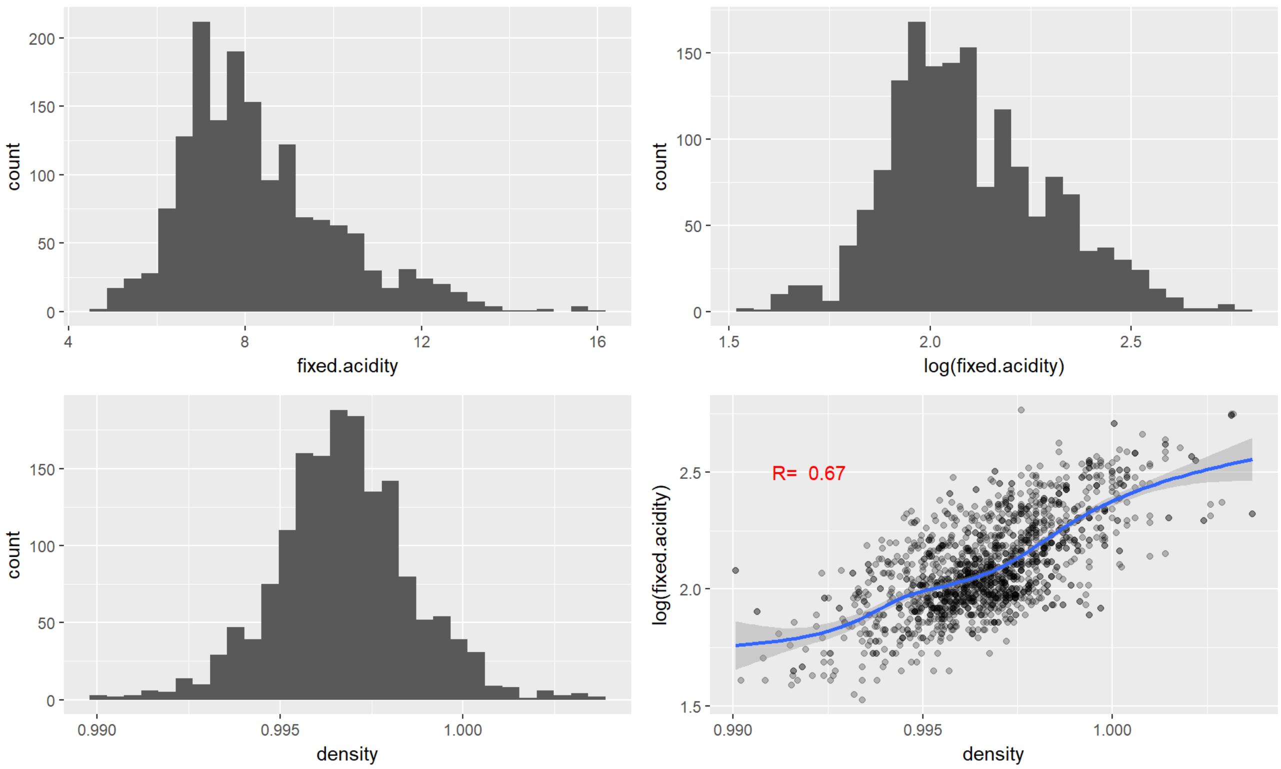
pH of wines in different qualities



I checked pH data in wine groups with differnt qualities(good and poor) and found it remained almost same in two groups.

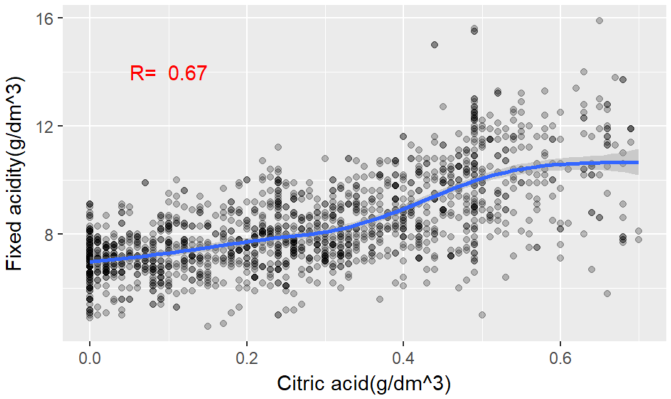
Exploring other variables

Density and fixed.acidity



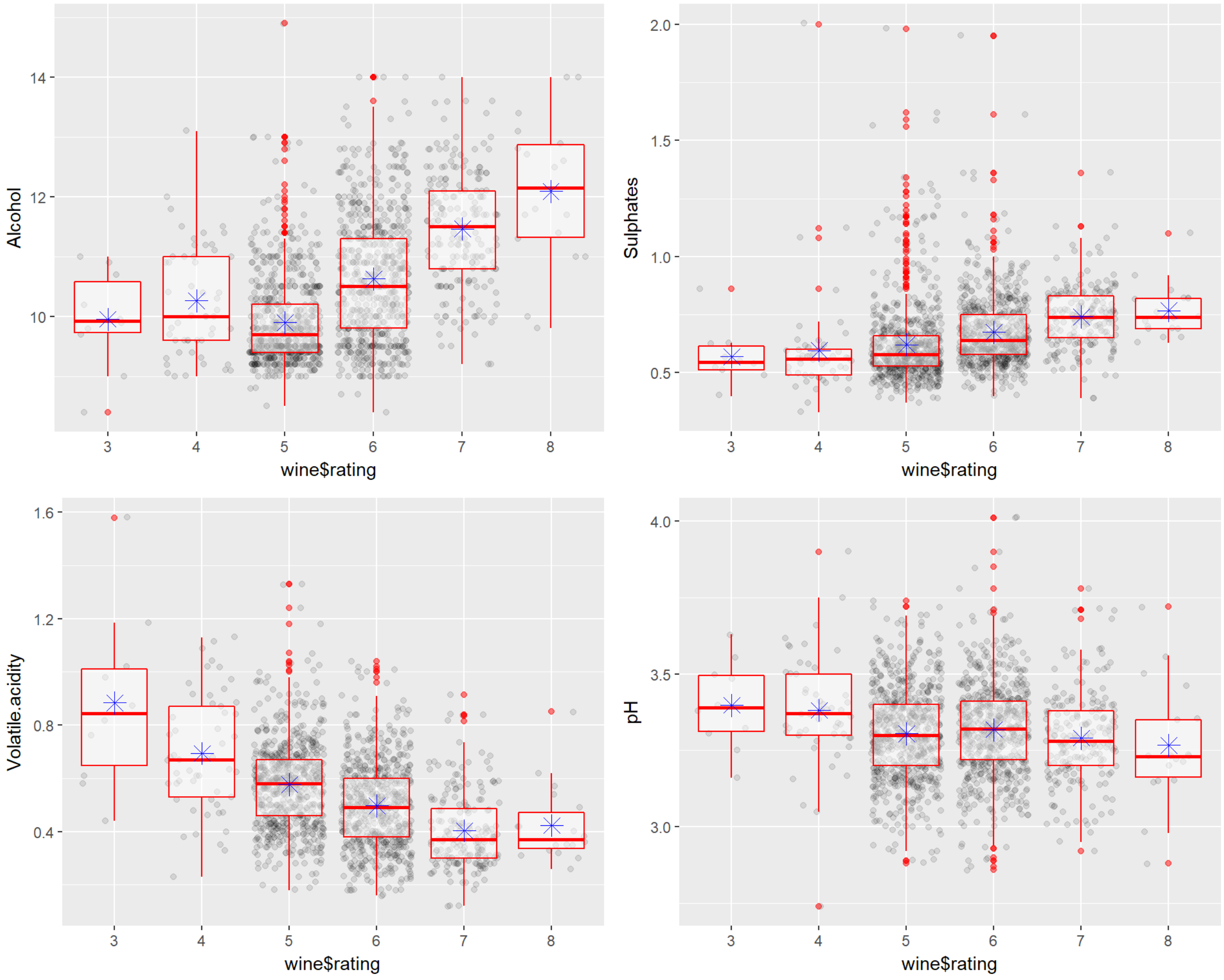
According to correlation plot ,I compared fixed.acidity and density. These two variables are correlated.

Citric.acid and fixed.acidity



Citric.acid and Fixed.acidity have a good correlation.

Exploring other variables and quality



Checked the changes of different variables in wines with differnt qualities in detail. The trend is as the same as above. Better wines have higher alcohol and sulphates amount,but lower volatile.acidity amount. pH value is similar between wines with differnt qualities.

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 plotted alcohol, sulphates, volatile.acidity, total.sulfur.dioxide and pH of wines in different quality group. These plots shows, 1.Mean of alcohol amount of good wines is higher than of poor wines(10.86 vs 9.92).We can also see from the density plot, there is a peak of alcohol amount of poor wine at around 9, while the distribution of alcohol amount of good wines has a small peak around 11.
2.Mean od sulphates in good wines is also slightly higher than in poor wines(0.69 vs.0.62).The density plot also shows some of the good wines have higher sulphates.
3.Volatile acidity has inverse changing trend to the quality. Mean of volatile in good wines is lower than in poor wines(0.47 vs. 0.59).
4.Mean of total.sulfur.dioxide in good wines is slightly lower than in poor wines(39.35 vs. 54.64).
5.While pH value is almost the same in good and poor wines.

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

I also explored other variables including density and fixed.acidity, citric.acid and fixed.acidity. I found density and fixed.acidity have a good correlation(with a R-value 0.675), looks like wines with higher fixed acidity have higher density.

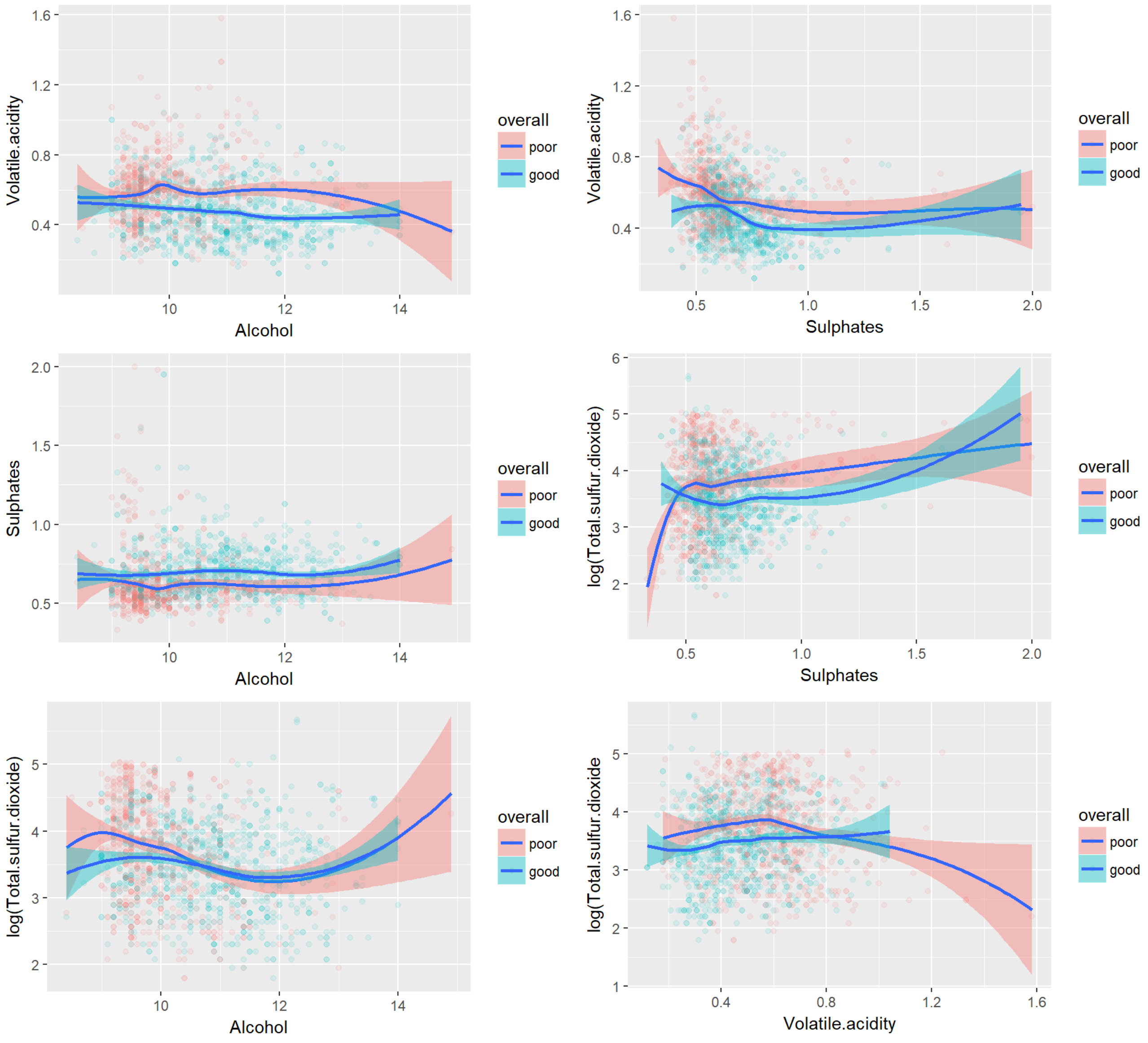
Citric.acid and fixed.acidity also have good correlation(with a R-value of 0.672). According to the dataset background, tartaric acid contributes to fixed acidity,since the data shown wines have higher fixed.acidity more likely to have higher citric.acid, which are from better quality of grapes.

What was the strongest relationship you found?

The strongest relationship I found is quality and alcohol, density and fixed.acidity, citric.acid and fixed.acidity.

Multivariate Plots Section

Highly correlated variables



Checked the changing trend between highly correlated variables in wines of differnt quality groups.

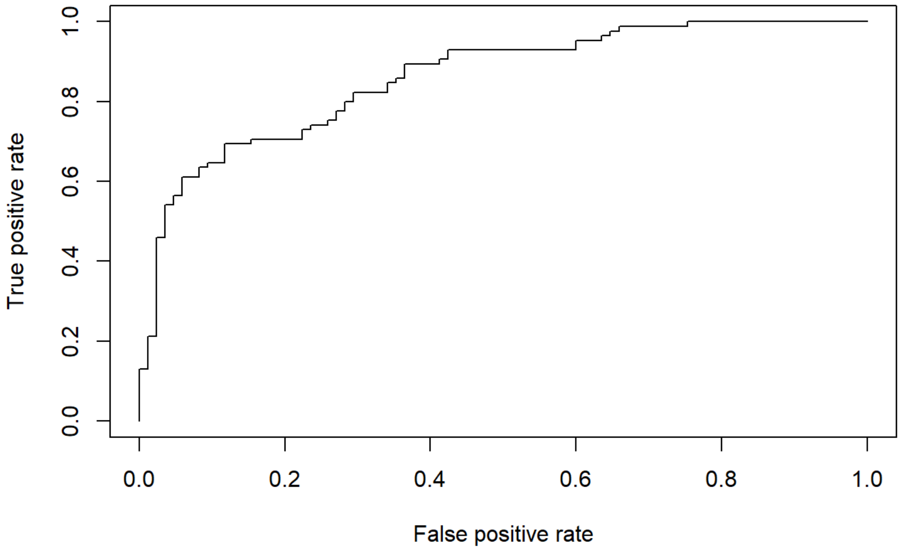
Logistic regression model

```
## 
## Call:
## glm(formula = overall ~ ., family = binomial(link = "logit"), 
##     data = train[, c(1:11, 14)])
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -3.2717  -0.8432   0.3276   0.8388   2.3217  
## 
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)    
## (Intercept)          -13.965046  84.729578  -0.165   0.8691    
## fixed.acidity          0.047367   0.105422   0.449   0.6532    
## volatile.acidity      -3.244424   0.521101  -6.226 4.78e-10 ***
## citric.acid           -1.182972   0.601604  -1.966   0.0493 *  
## residual.sugar         0.021268   0.057657   0.369   0.7122    
## chlorides             -4.426433   1.679606  -2.635   0.0084 ** 
## free.sulfur.dioxide    0.021042   0.008711   2.415   0.0157 *  
## total.sulfur.dioxide  -0.015121   0.003036  -4.981 6.34e-07 ***
## density                8.710090  86.586566   0.101   0.9199    
## pH                    -0.997040   0.782198  -1.275   0.2024    
## sulphates              2.942890   0.501821   5.864 4.51e-09 ***
## alcohol                0.884705   0.110786   7.986 1.40e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 1878.4  on 1358  degrees of freedom
## Residual deviance: 1422.3  on 1347  degrees of freedom
## AIC: 1446.3
## 
## Number of Fisher Scoring iterations: 4
```

Calculate the accruacy

```
## [1] "Accuracy 0.752941176470588"
```

ROC curve and AUC



```
## [1] "auc is 0.859377162629758"
```

ROC curve of this regression model.

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?

I plotted 6 variable-pairs colorred different wine quality, the scatter plots shown groups corresponding to different wine qualities, based on this observcation, I tried modeling these data.

Were there any interesting or surprising interactions between features?

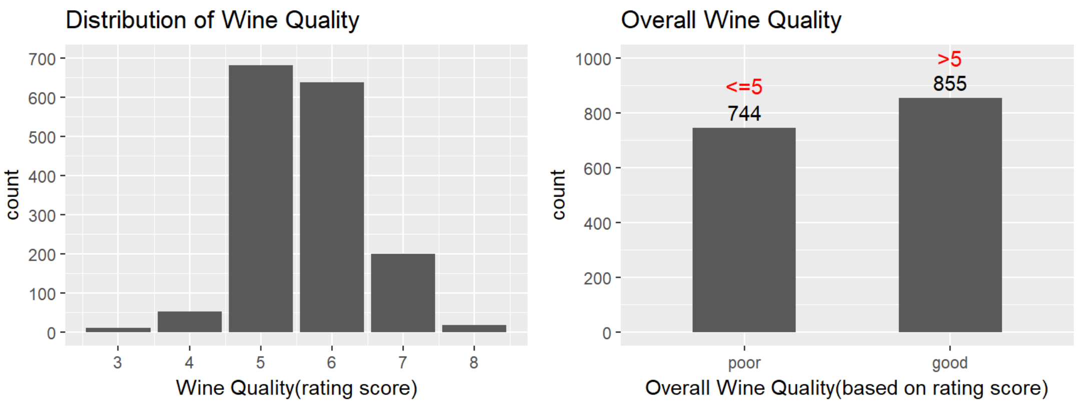
I found several acidity variables are related to wine quality, but pH value seems have nothing to do with quality. It's a little bit strange since pH value is highly related to acidity condition.

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

I tranformed the quality feature to a binary variable("<=5"-"poor"-0, while ">5"-"good"-1),then tried logistic regression model on this dataset, got a model has an accuracy of 0.75. With a dataset of only 1599 observations, I think the accuracy is acceptable.

Final Plots and Summary

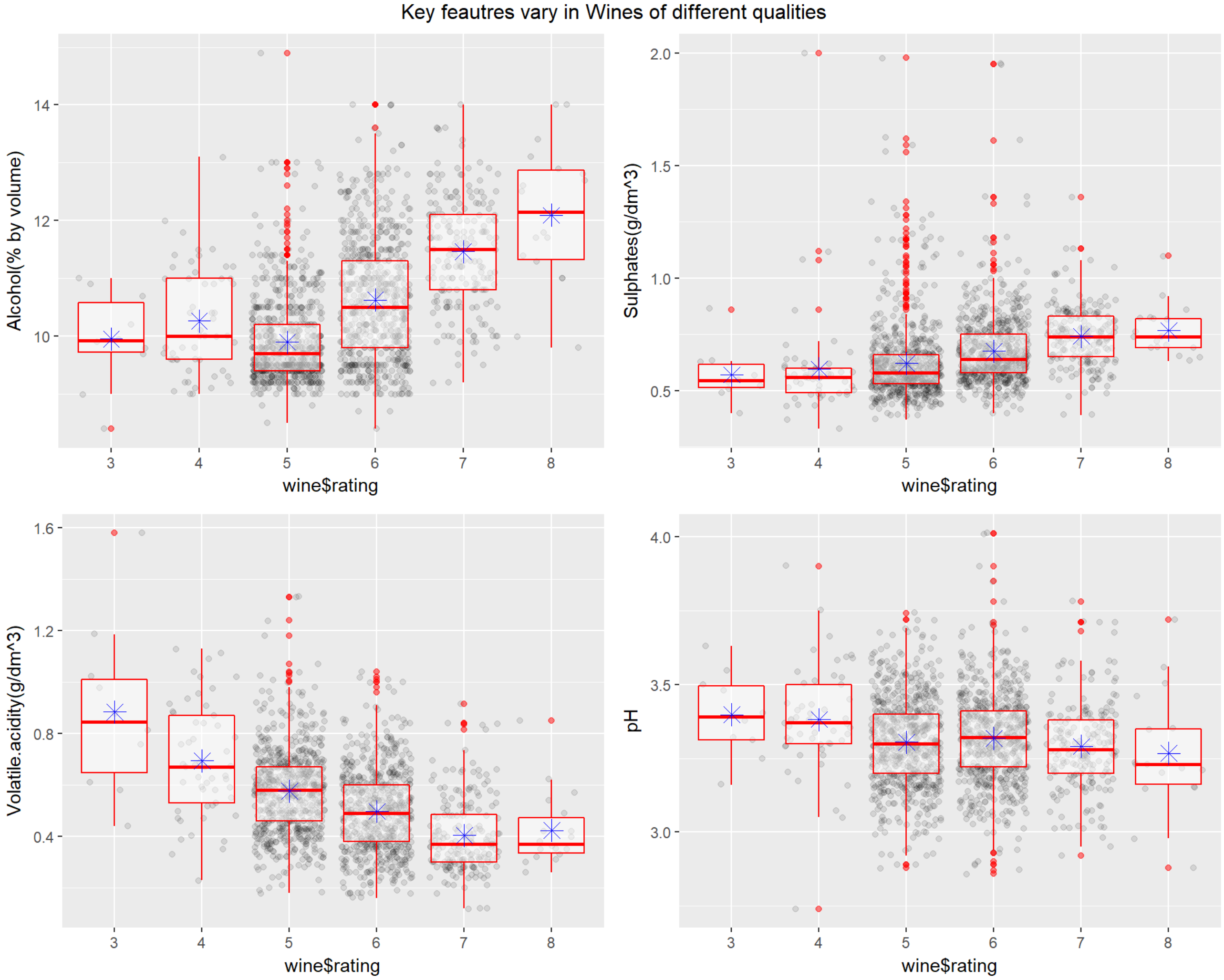
Plot One



Description One

Firstly I checked the distribution of the output variable "Quality" and found most of the wines have a rating score in the range of 5-7. For the ease of following work, I made a new variable based on quality, which is the overall feature. Wines with a rating score less or equal to "5" are labeled as "poor" and wines with a rating score higher than 5 are labeled as "good".

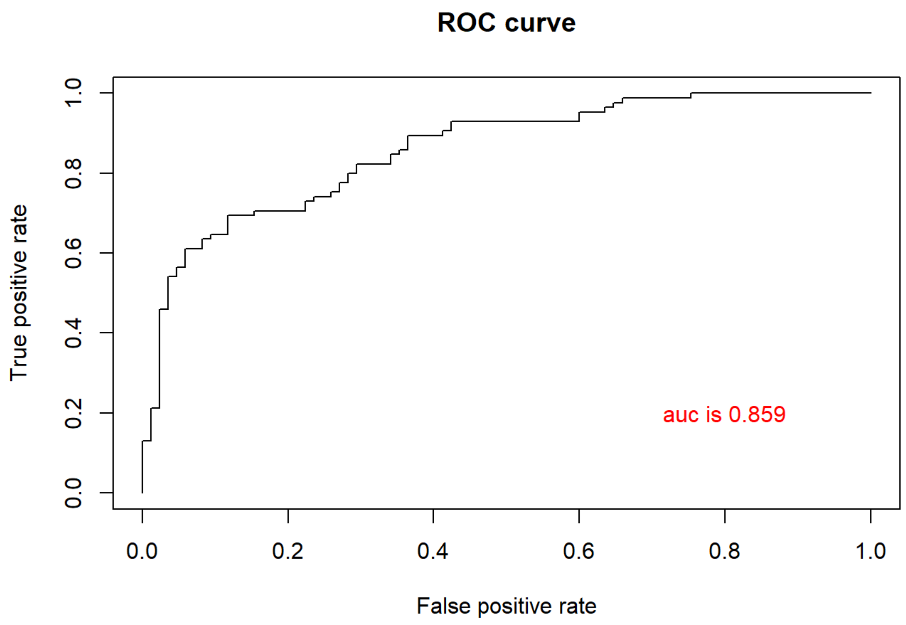
Plot Two



Description Two

In order to predict the quality of wines, I chosen some features that correlated to quality based on the correlation plot. Apparently alcohol has the strongest relationship with quality, as well as other variables such as "Volatile,.acidity","sulphates" and "total.sulfure.dioxide". I explored different vairable pairs and found based on these variables,some data points were forming into two centered groups, while some data points had overlaps.

Plot Three



Description Three

Since I made the binary variable “overall”, I tried the logistic regression model on this dataset, with 85% of the data as training set and 15% of the data as testing set.I got a predict accuracy of 0.75. And the summary of model also shown among all these variables, “Alcohol”,“sulphates”,“total.sulfur.dioxide”, and “volatile.acidity” are the most important variables to wine quality.

Reflection

This is my very first try of Explore Data Analysis, both the dataset and process are very intersting. Many features can affect the output variable “wine quality”.

I was a little bit confused about the feature type and corresponding plot type, until the first submission review came back, I revised my project according to the response, now I think I made all these plots in the right way.

I was thinking about a linear regression model at first, so I tried to explore the relationship of different variables to the main vairable of interest. But actually not a single or even some combination of variables have a very strong relationship to “quality”. So later I tranformed the output variable to a binary one, wines with a rating score lower or equal to “5” are labeled as “poor”, which are denoted to “0”. And wines with a rating score higher than “5” are labeld as “good”, which are denoted to “1” later.

Logistic regression model worked for this dataset. The very preliminary model has a predict accuracy of 0.75. With cross-validation and also expanding the size of dataet, I should be able to get a better model.

This dataset still needs more work, for example, we can try feature selection to pick up variables that contribute more to wine quality. Then we may find more interesting things.