White Wine Exploration by Bo Rong

Oct. 20th 2017

This report explores a dataset containing quality and attributes for 4898 white wine. My main task is that analyze the white wine data, and try to find which variables are responsible for the quality.Then create a linear model for wine quality prediction.

Univariate Plots Section

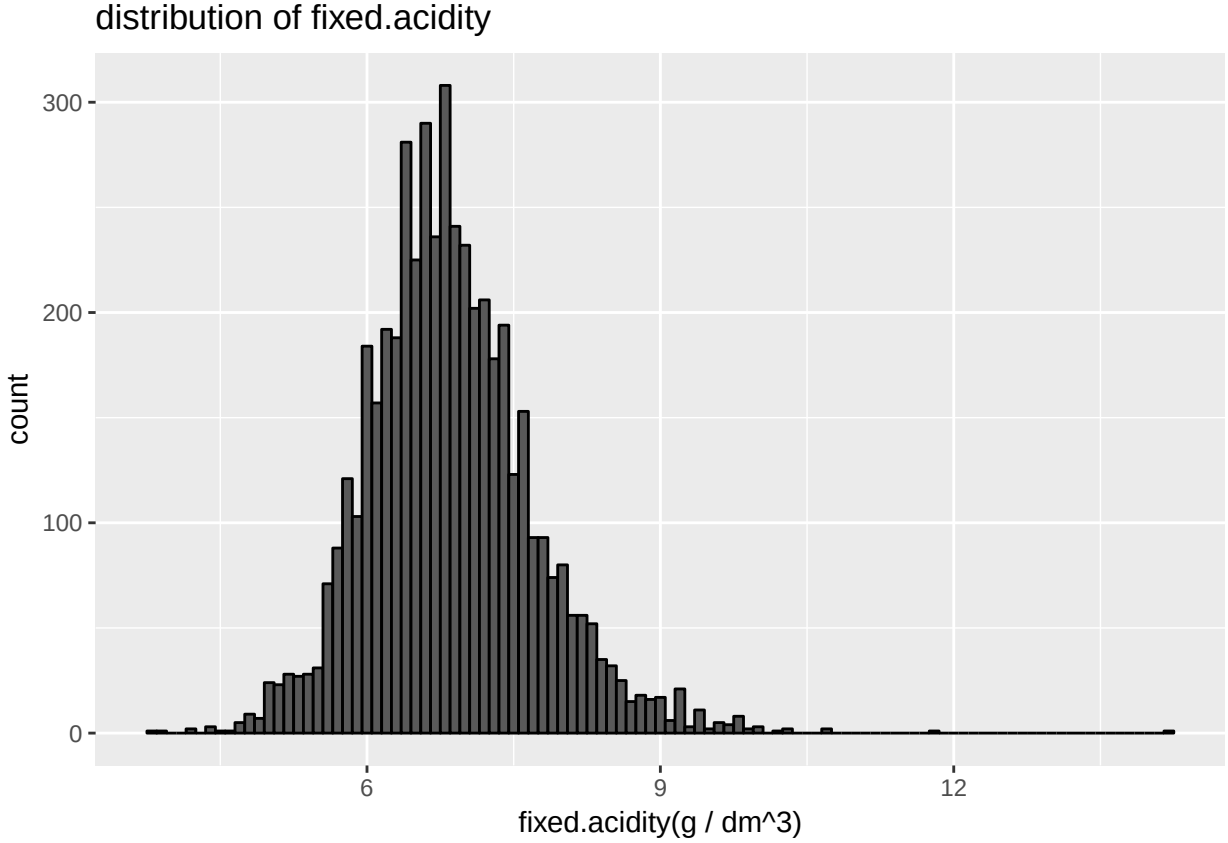
```
## [1] 4898   13
```

```
## 'data.frame':    4898 obs. of  13 variables:
##  $ fixed.acidity       : num  7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...
##  $ volatile.acidity    : num  0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 ...
##  $ citric.acid         : num  0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34 0.43 ...
##  $ residual.sugar      : num  20.7 1.6 6.9 8.5 8.5 6.9 7 20.7 1.6 1.5 ...
##  $ chlorides           : num  0.045 0.049 0.05 0.058 0.058 0.05 0.045 0.045 0.049 0.044 ...
##  $ free.sulfur.dioxide : num  45 14 30 47 47 30 30 45 14 28 ...
##  $ total.sulfur.dioxide: num  170 132 97 186 186 97 136 170 132 129 ...
##  $ density             : num  1.001 0.994 0.995 0.996 0.996 ...
##  $ pH                  : num  3 3.3 3.26 3.19 3.19 3.26 3.18 3 3.3 3.22 ...
##  $ sulphates           : num  0.45 0.49 0.44 0.4 0.4 0.44 0.47 0.45 0.49 0.45 ...
##  $ alcohol             : num  8.8 9.5 10.1 9.9 9.9 10.1 9.6 8.8 9.5 11 ...
##  $ quality             : int  6 6 6 6 6 6 6 6 6 6 ...
##  $ rate                : Factor w/ 3 levels "average","bad",..: 1 1 1 1 1 1 1 1 1 1 ...
```

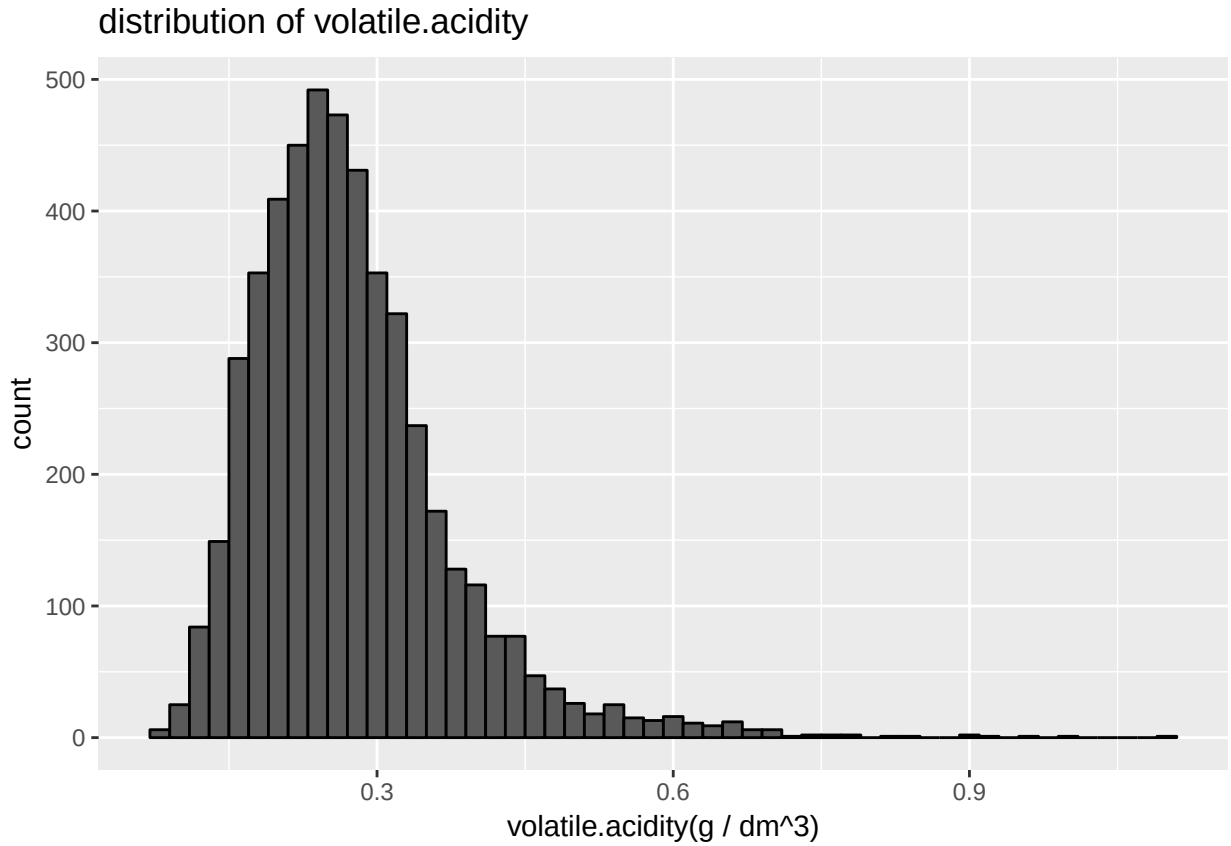
```
##  fixed.acidity    volatile.acidity  citric.acid     residual.sugar
##  Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600
##  1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700   1st Qu.: 1.700
##  Median : 6.800   Median :0.2600   Median :0.3200   Median : 5.200
##  Mean   : 6.855   Mean   :0.2782   Mean   :0.3342   Mean   : 6.391
##  3rd Qu.: 7.300   3rd Qu.:0.3200   3rd Qu.:0.3900   3rd Qu.: 9.900
##  Max.   :14.200   Max.   :1.1000   Max.   :1.6600   Max.   :65.800
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide
##  Min.   :0.00900   Min.   :  2.00      Min.   :  9.0
##  1st Qu.:0.03600   1st Qu.: 23.00      1st Qu.:108.0
##  Median :0.04300   Median : 34.00      Median :134.0
##  Mean   :0.04577   Mean   : 35.31      Mean   :138.4
##  3rd Qu.:0.05000   3rd Qu.: 46.00      3rd Qu.:167.0
##  Max.   :0.34600   Max.   :289.00      Max.   :440.0
##     density             pH          sulphates         alcohol
##  Min.   :0.9871   Min.   :2.720   Min.   :0.2200   Min.   : 8.00
##  1st Qu.:0.9917   1st Qu.:3.090   1st Qu.:0.4100   1st Qu.: 9.50
##  Median :0.9937   Median :3.180   Median :0.4700   Median :10.40
##  Mean   :0.9940   Mean   :3.188   Mean   :0.4898   Mean   :10.51
##  3rd Qu.:0.9961   3rd Qu.:3.280   3rd Qu.:0.5500   3rd Qu.:11.40
##  Max.   :1.0390   Max.   :3.820   Max.   :1.0800   Max.   :14.20
##     quality          rate
##  Min.   :3.000   average:3818
##  1st Qu.:5.000   bad    :  20
##  Median :6.000   good   :1060
```

```
##  Mean   :5.878
##  3rd Qu.:6.000
##  Max.   :9.000
```

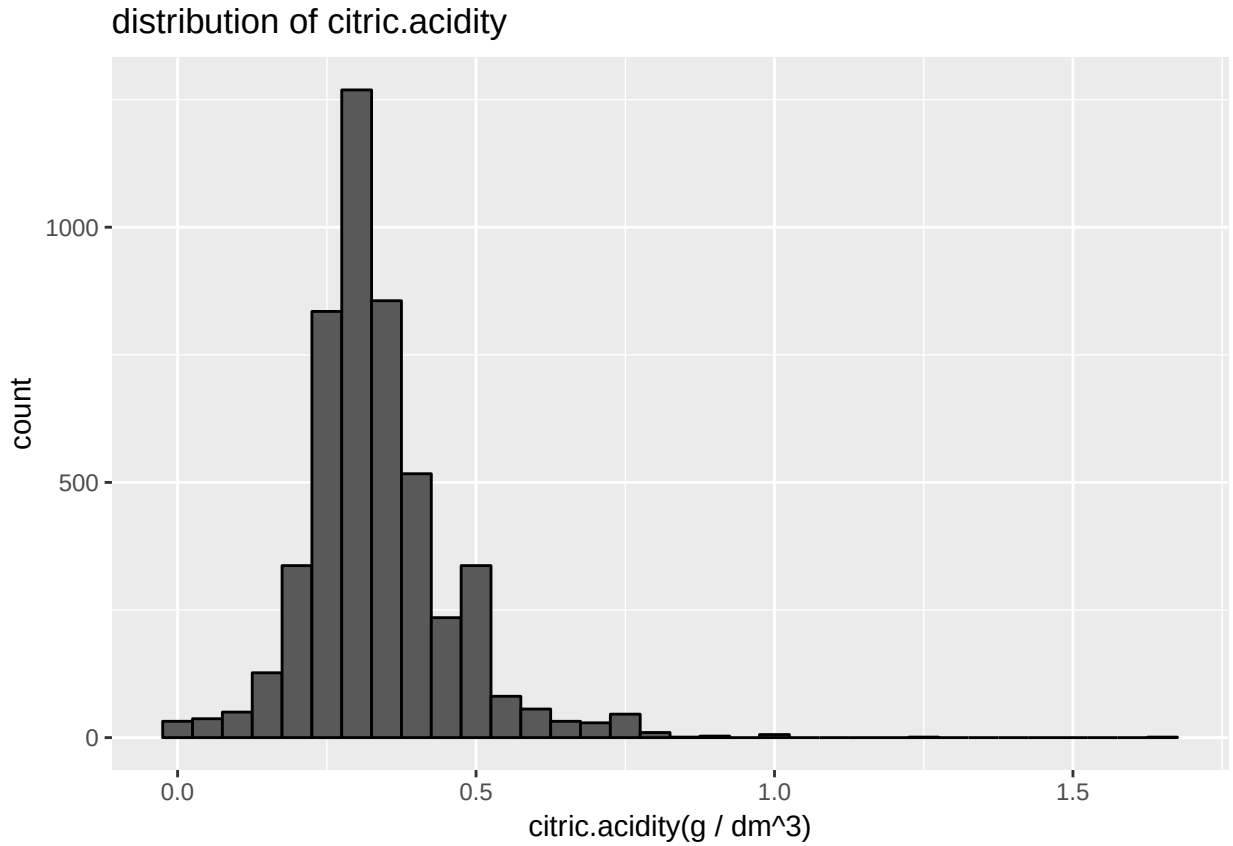
Our dataset consists of 13 variables, with 4898 observations.



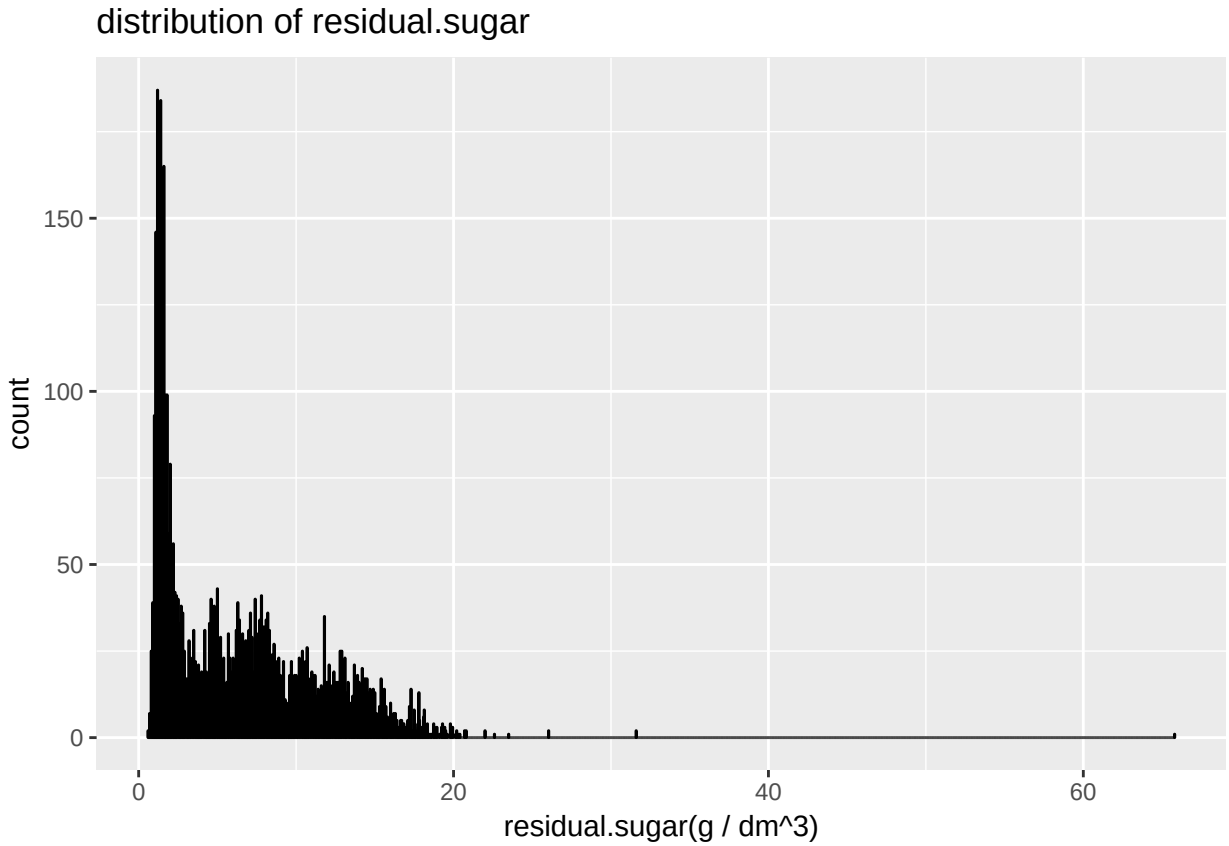
The distribution of fixed.acidity seems normally.



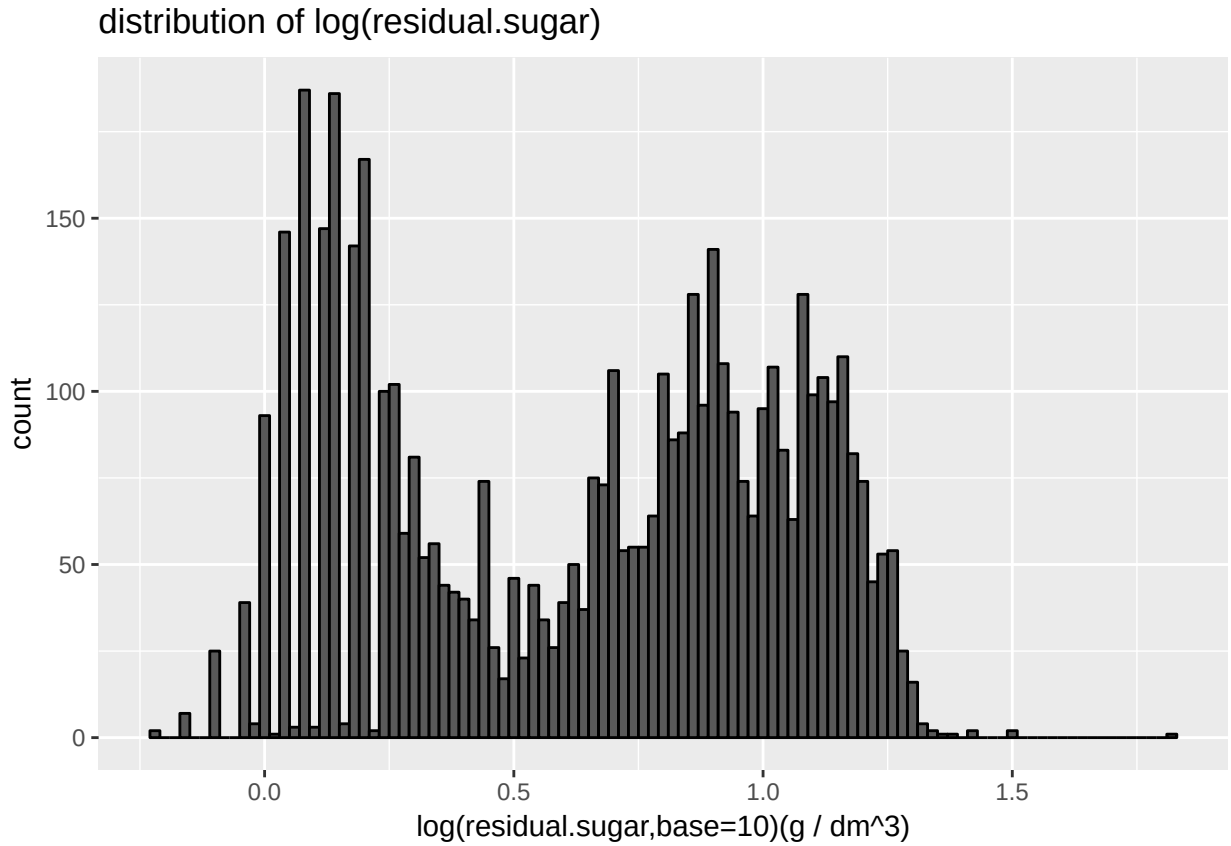
The distribution of volatile.acidity is a little bit right skewed. A few outliers beyond 0.7.



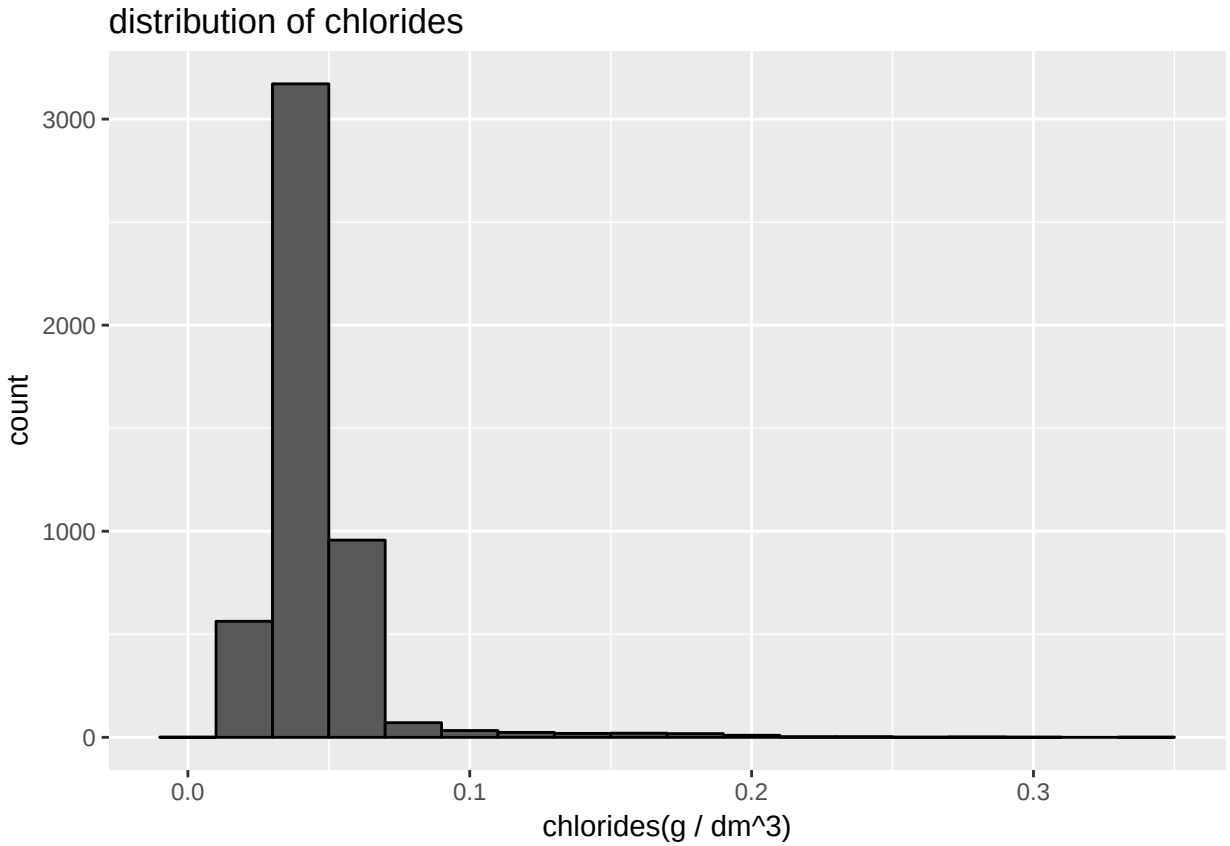
The distribution of citric.acid is normal with high peaks at around 0.35.



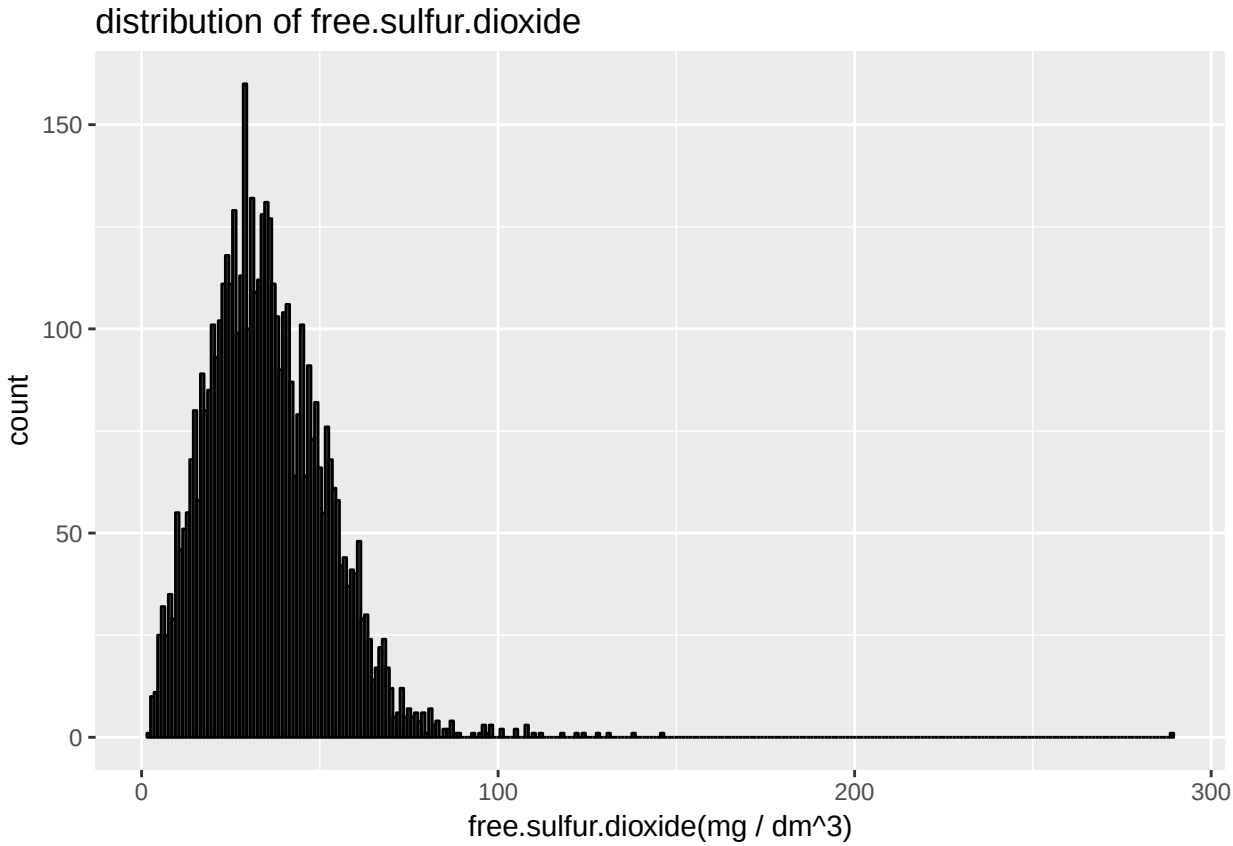
The distribution of residual.sugar is positively skewed with high peaks at around 2 with many outliers present at the higher ranges.



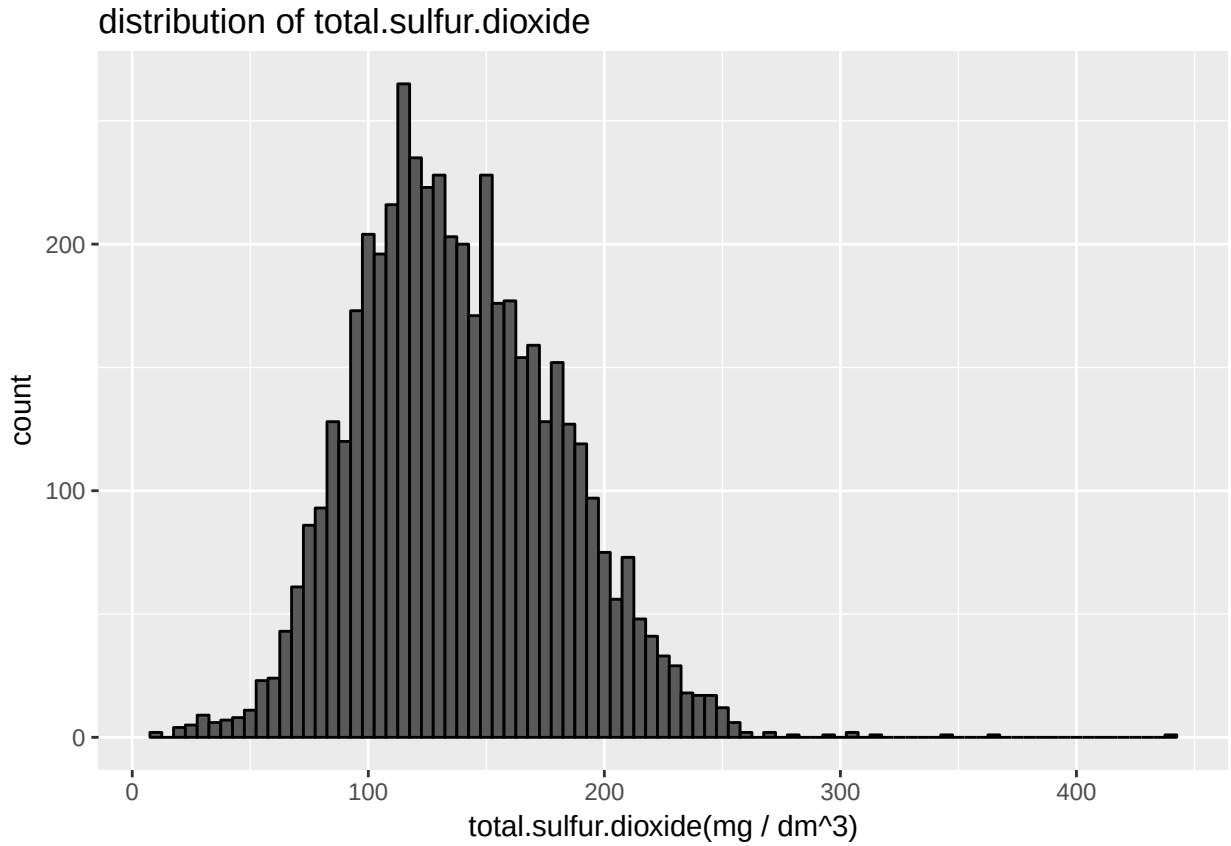
If we use log10 transformation, the result looks like a binormal distribution with peak at about 0.1 and 0.8.



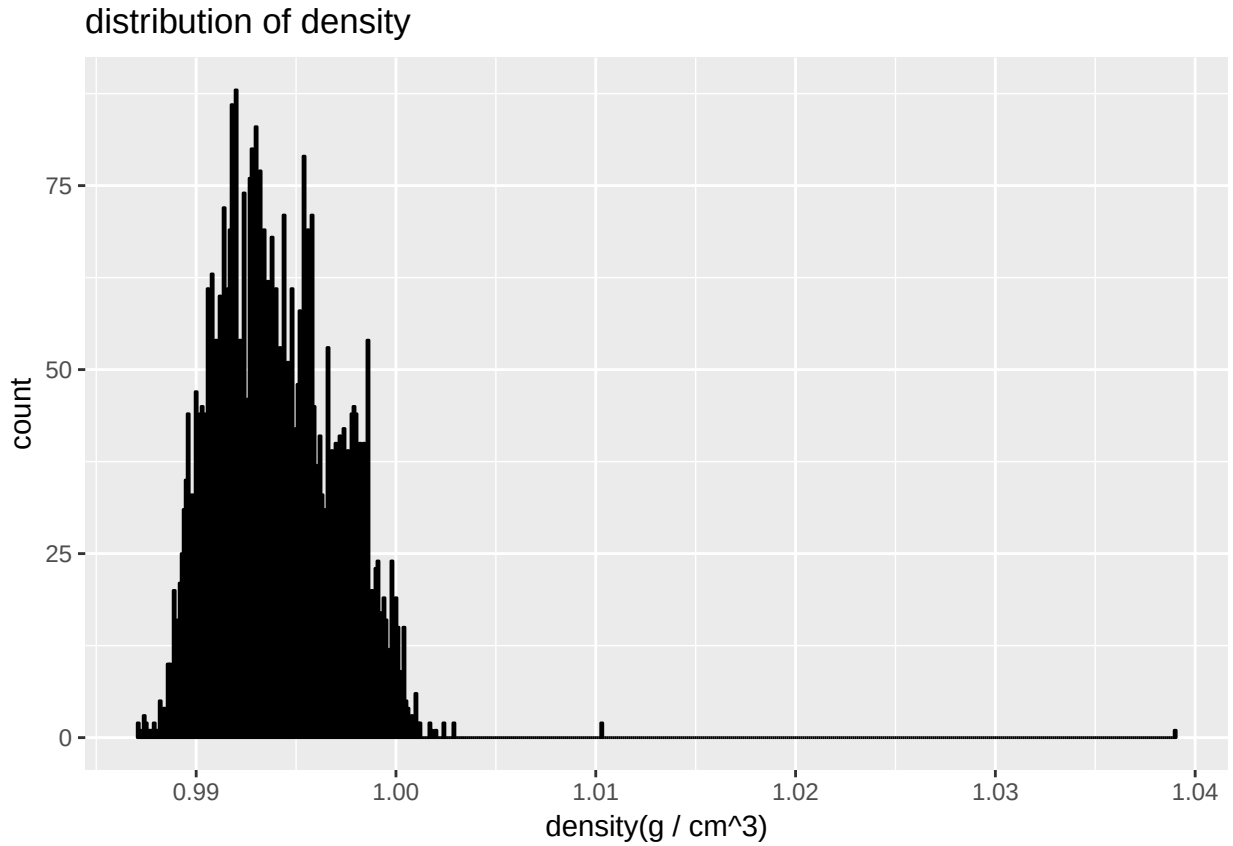
A lot of outliers present betweent 0.15 to 0.35.



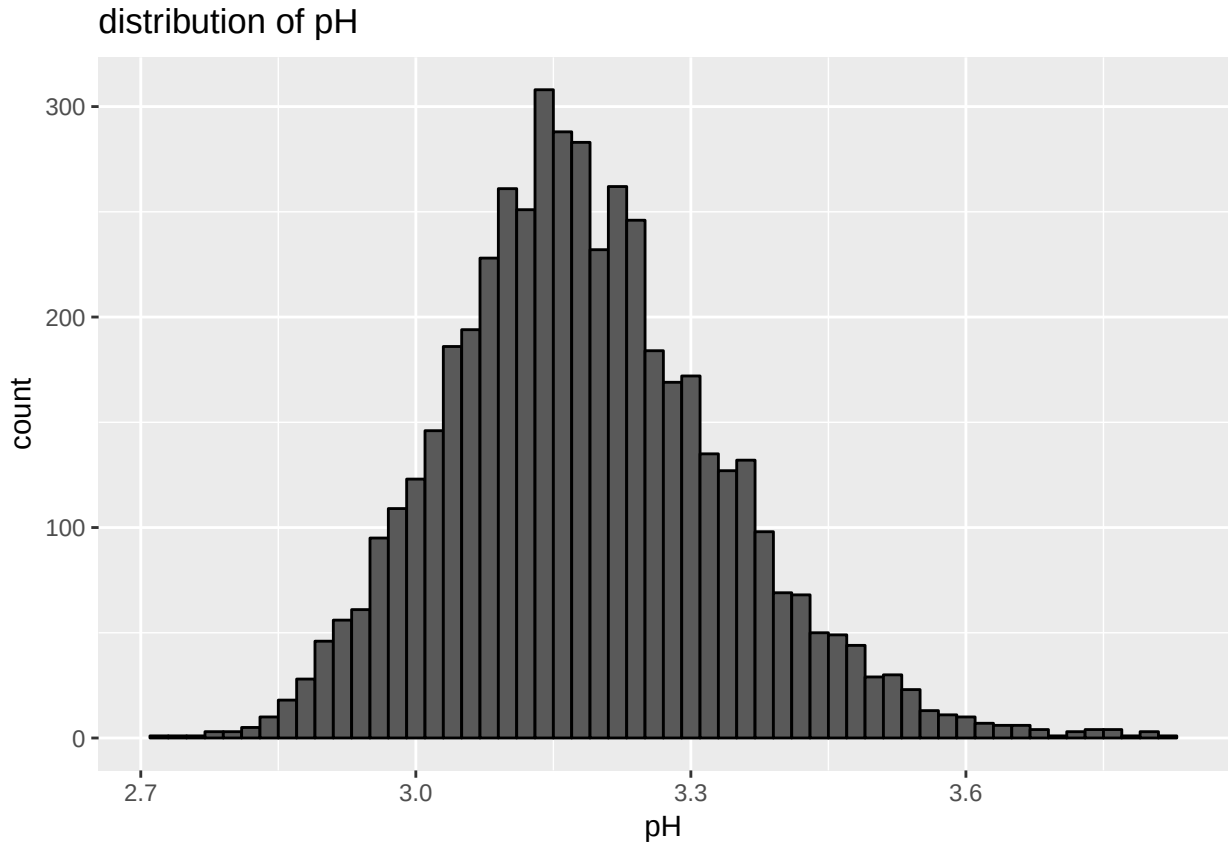
The distribution of free.sulfur.dioxide is normally, again a lot of outliers.



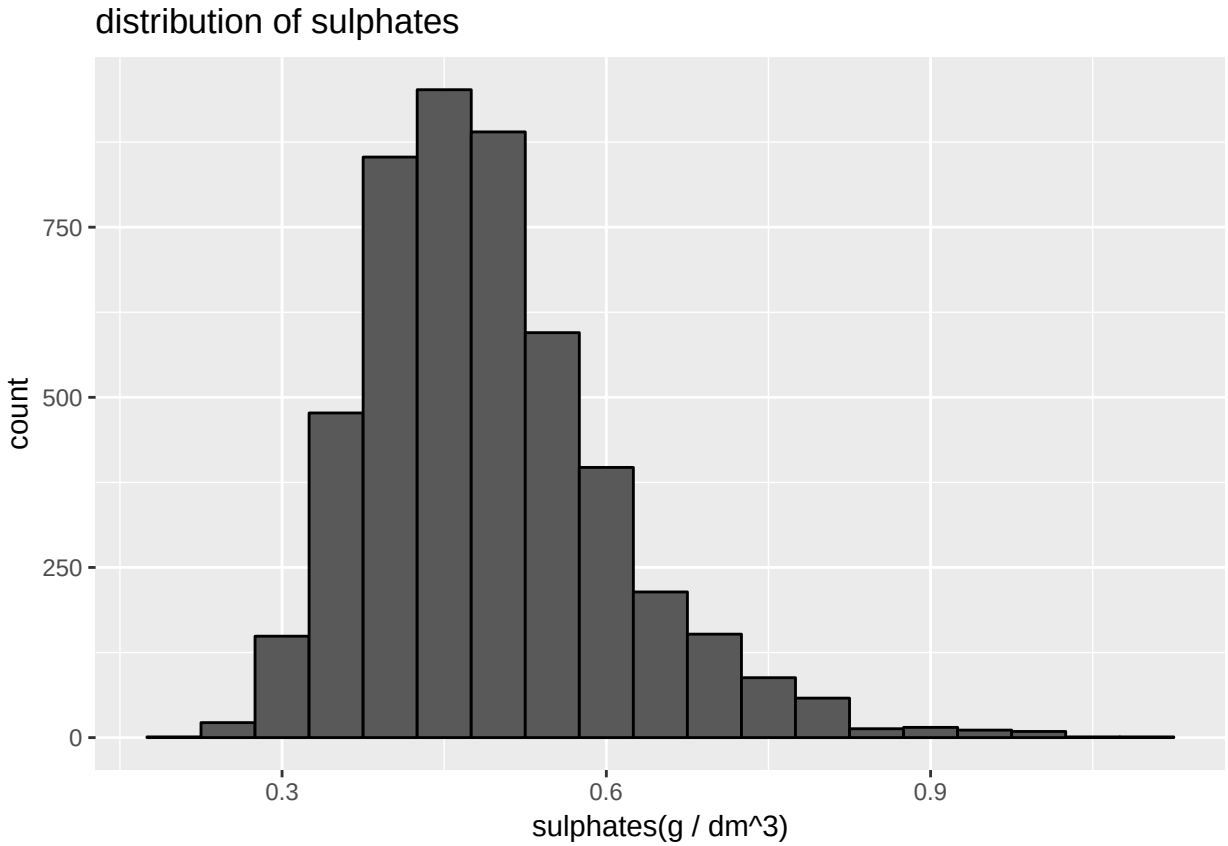
The distribution of total.sulfur.dioxide looks normally.



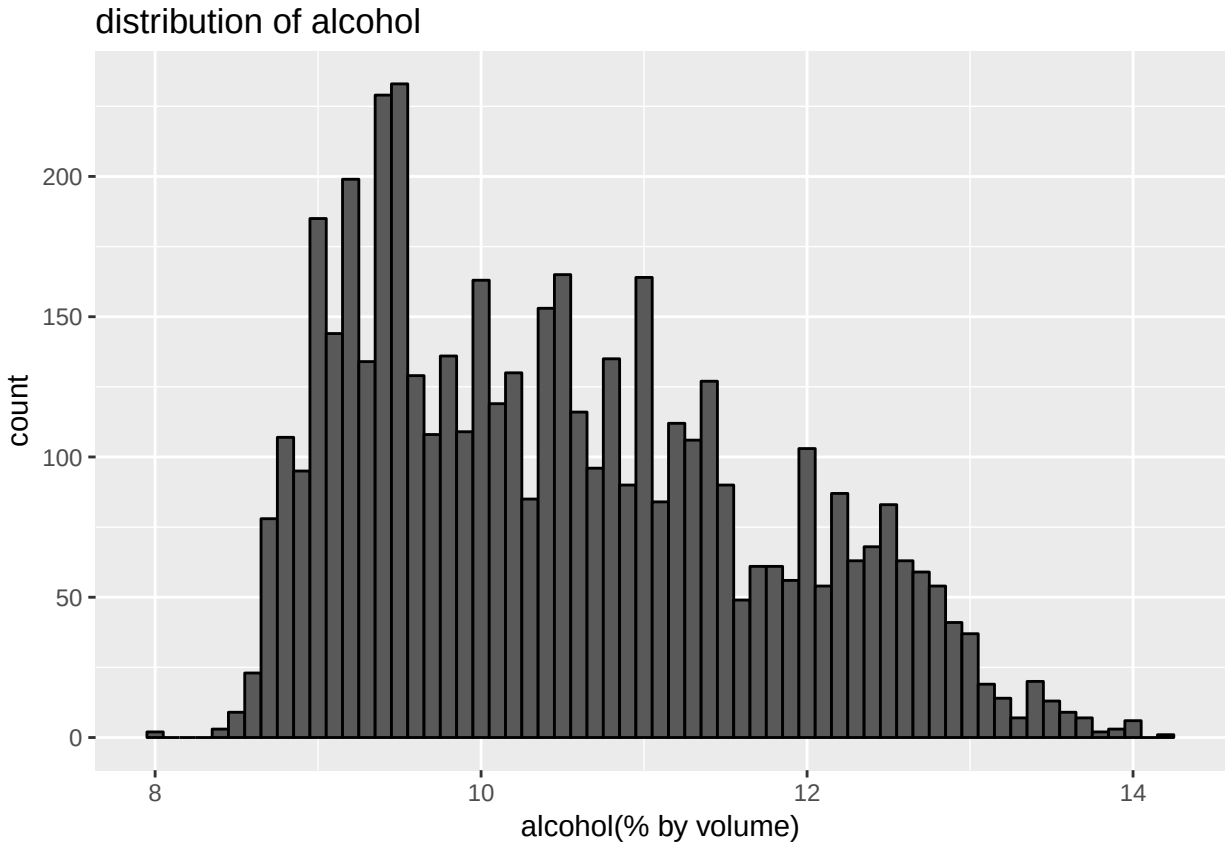
The distribution of density seems normal with peak around 0.992, and most of data are between 0.99 and 1.00



pH has a very Normally distributed shape.

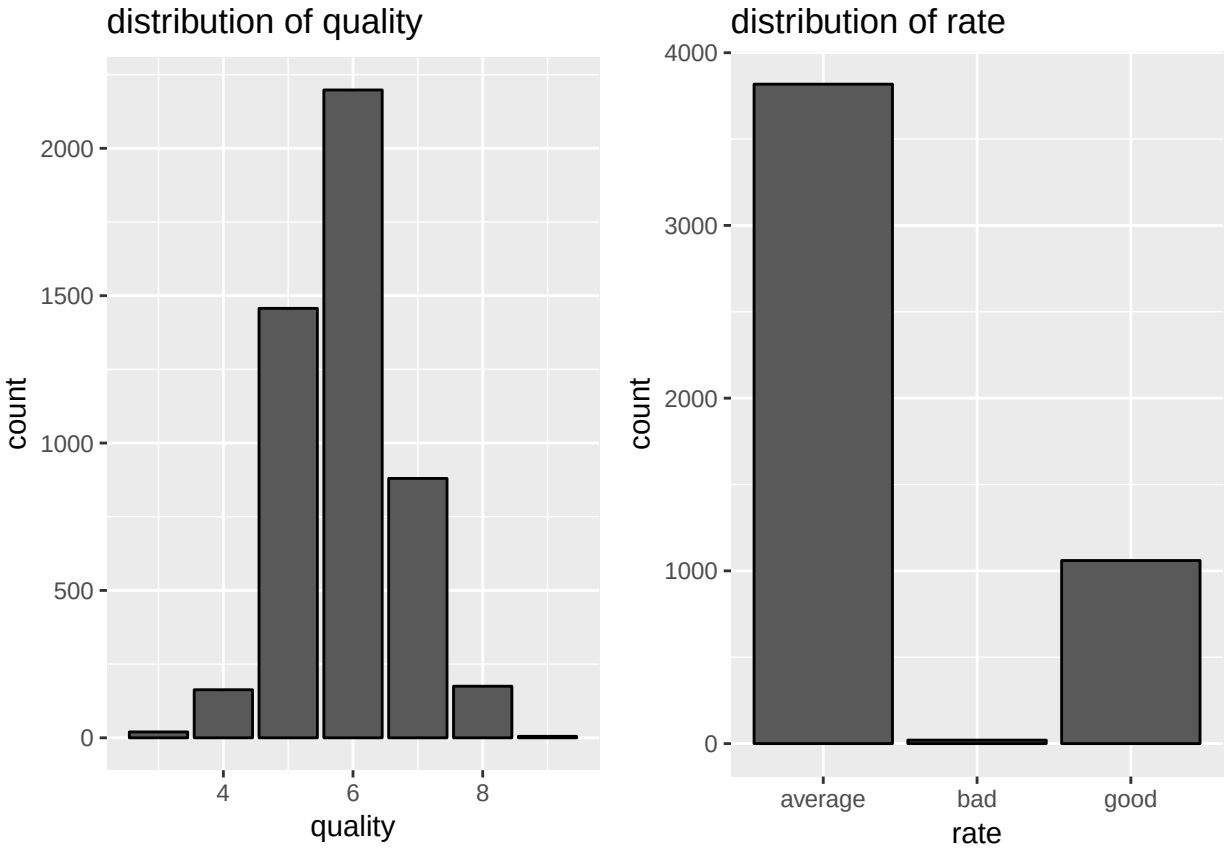


Sulphates has a long tailed distribution. It has a few outliers.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00    9.50   10.40   10.51   11.40   14.20
```

Alcohol follows a right skewed distribution.



Based on the result, we can see that most of the wine are averagr quality, just a few have bad quality.

Univariate Analysis

What is the structure of your dataset?

There are 4898 white wine in the dataset with 13 features (fixed.acidity, volatile.acidity, citric.acid, residual.sugar, chlorides, free.sulfur.dioxide, total.sulfur.dioxide, density, pH,sulphates, alcohol, quality and rate). The variable quality and rate are factor variables with the following levels, and the rest of the variables are numerical variables. (worst) -> (best) quality: 3, 4, 5, 6, 7, 8, 9 rate: bad, average, good

Other observations:

Most white wine are of average rate. The median quality is 6. All PH value of wine are less than 4 and greater than 2.7. The density are around 0.99.

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

The main features in the data set are quality and rate, I would like to determine which features are best for predicting the quality of white wine.

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

fixed.acidity, volatile.acidity, free.sulfur.dioxide, total.sulfur.dioxide and density are likely to contribute to quality of wine.

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

Yes, I create a new variable rate to the dataset which distributes the sample into 3 quality bins (0,3], [4,6] and [7,9].

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 residual.sugar distributions. The tranformed distribution is a binormal.

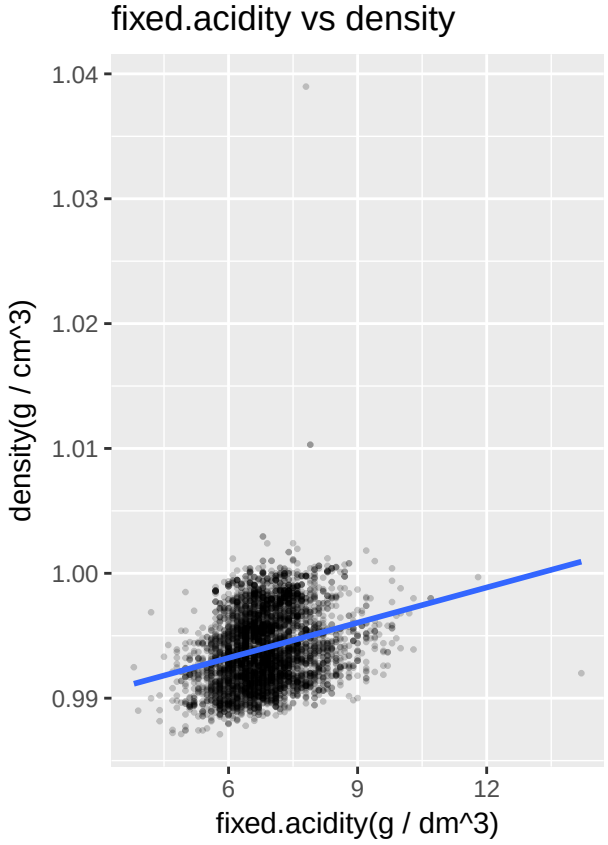
Bivariate Plots Section

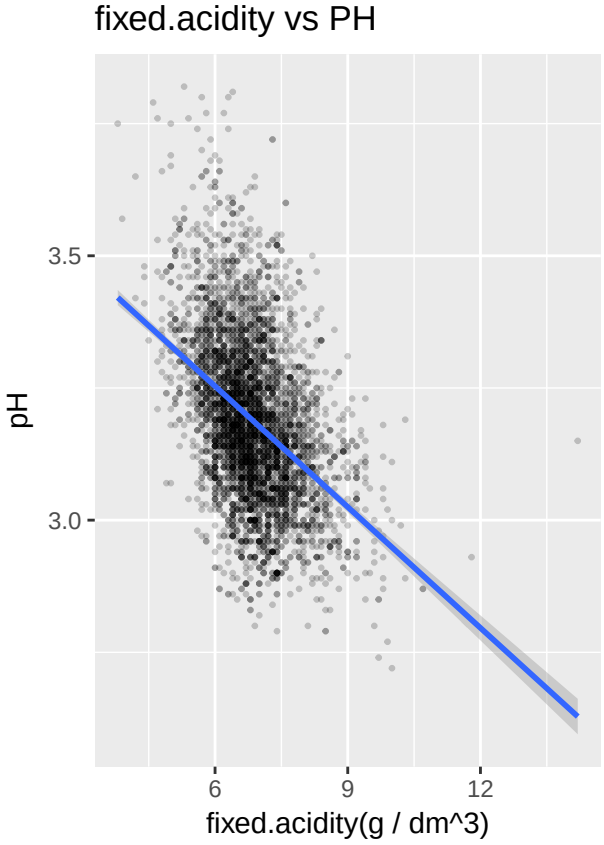
Correlation coefficient table

```
##                       fixed.acidity volatile.acidity citric.acid
## fixed.acidity                  1.00            -0.02        0.29
## volatile.acidity              -0.02             1.00       -0.15
## citric.acid                    0.29            -0.15        1.00
## residual.sugar                 0.09             0.06        0.09
## chlorides                      0.02             0.07        0.11
## free.sulfur.dioxide           -0.05            -0.10        0.09
## total.sulfur.dioxide           0.09             0.09        0.12
## density                        0.27             0.03        0.15
## pH                            -0.43            -0.03       -0.16
## sulphates                     -0.02            -0.04        0.06
## alcohol                       -0.12             0.07       -0.08
## quality                       -0.11            -0.19       -0.01
##                       residual.sugar chlorides free.sulfur.dioxide
## fixed.acidity                   0.09      0.02                -0.05
## volatile.acidity                0.06      0.07                -0.10
## citric.acid                     0.09      0.11                 0.09
## residual.sugar                  1.00      0.09                 0.30
## chlorides                       0.09      1.00                 0.10
## free.sulfur.dioxide             0.30      0.10                 1.00
## total.sulfur.dioxide            0.40      0.20                 0.62
## density                         0.84      0.26                 0.29
## pH                             -0.19     -0.09                 0.00
## sulphates                      -0.03      0.02                 0.06
## alcohol                        -0.45     -0.36                -0.25
## quality                        -0.10     -0.21                 0.01
##                       total.sulfur.dioxide density    pH sulphates alcohol
## fixed.acidity                         0.09    0.27 -0.43     -0.02   -0.12
## volatile.acidity                      0.09    0.03 -0.03     -0.04    0.07
## citric.acid                           0.12    0.15 -0.16      0.06   -0.08
## residual.sugar                        0.40    0.84 -0.19     -0.03   -0.45
## chlorides                             0.20    0.26 -0.09      0.02   -0.36
## free.sulfur.dioxide                   0.62    0.29  0.00      0.06   -0.25
## total.sulfur.dioxide                  1.00    0.53  0.00      0.13   -0.45
## density                               0.53    1.00 -0.09      0.07   -0.78
## pH                                    0.00   -0.09  1.00      0.16    0.12
## sulphates                             0.13    0.07  0.16      1.00   -0.02
```

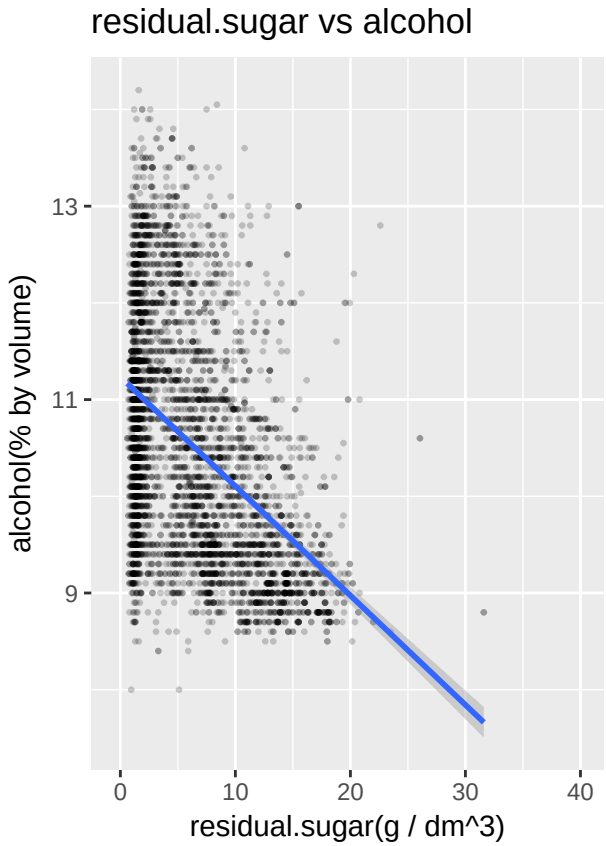
```
## alcohol                                -0.45   -0.78  0.12     -0.02    1.00
## quality                                -0.17   -0.31  0.10      0.05    0.44
##                       quality
## fixed.acidity           -0.11
## volatile.acidity        -0.19
## citric.acid             -0.01
## residual.sugar          -0.10
## chlorides               -0.21
## free.sulfur.dioxide      0.01
## total.sulfur.dioxide    -0.17
## density                 -0.31
## pH                       0.10
## sulphates                0.05
## alcohol                  0.44
## quality                  1.00
```

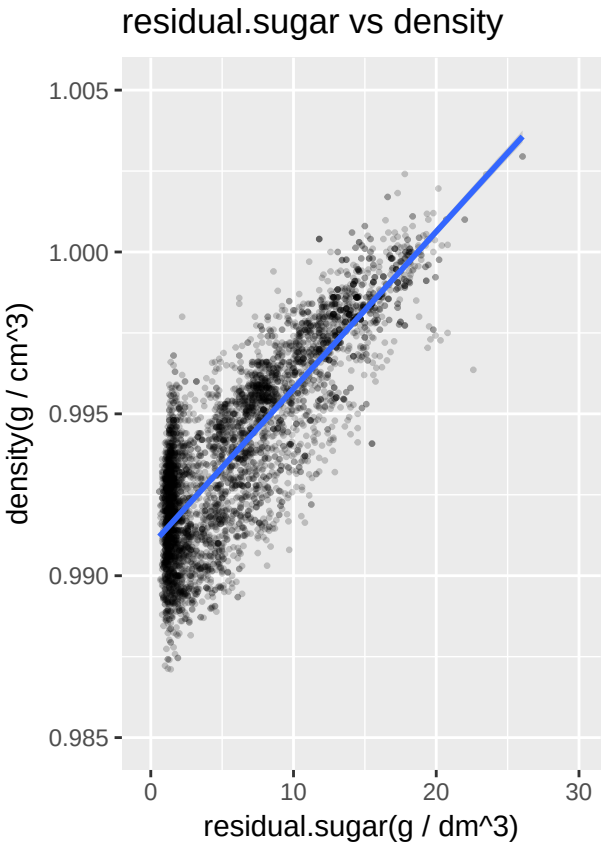
The result shows that: 1.fixed.acidity has positive relationship with density and citric.acid, while it has negative relationship with PH. 2.residual.sugar has strong relationship with density, and negative relationship with alcohol. 3.density has negative relationship with alcohol and quality. 4.It looks like that density and alcohol are important to determine the quality.



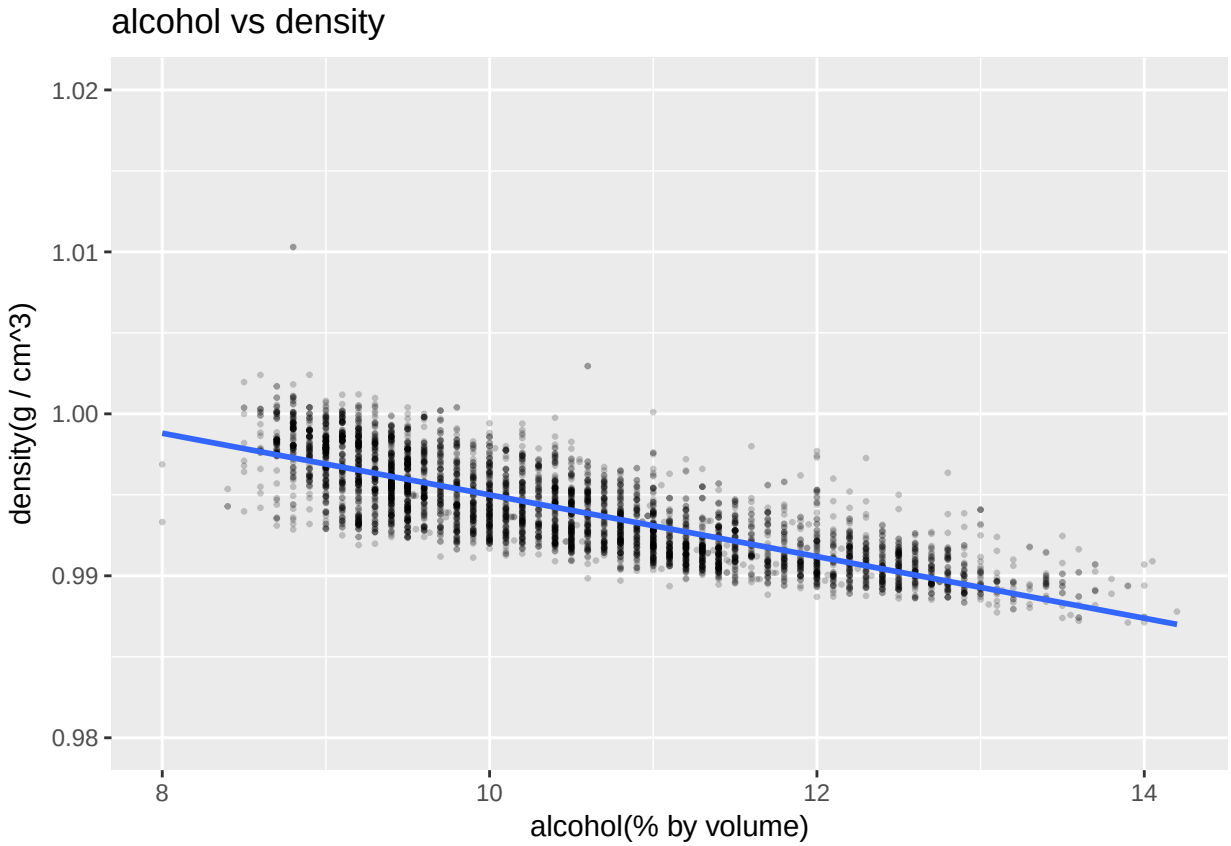


fixed.acidity and density has a positive correlation, and there are some outliers. And fixed.acidity has negative relationship with PH.

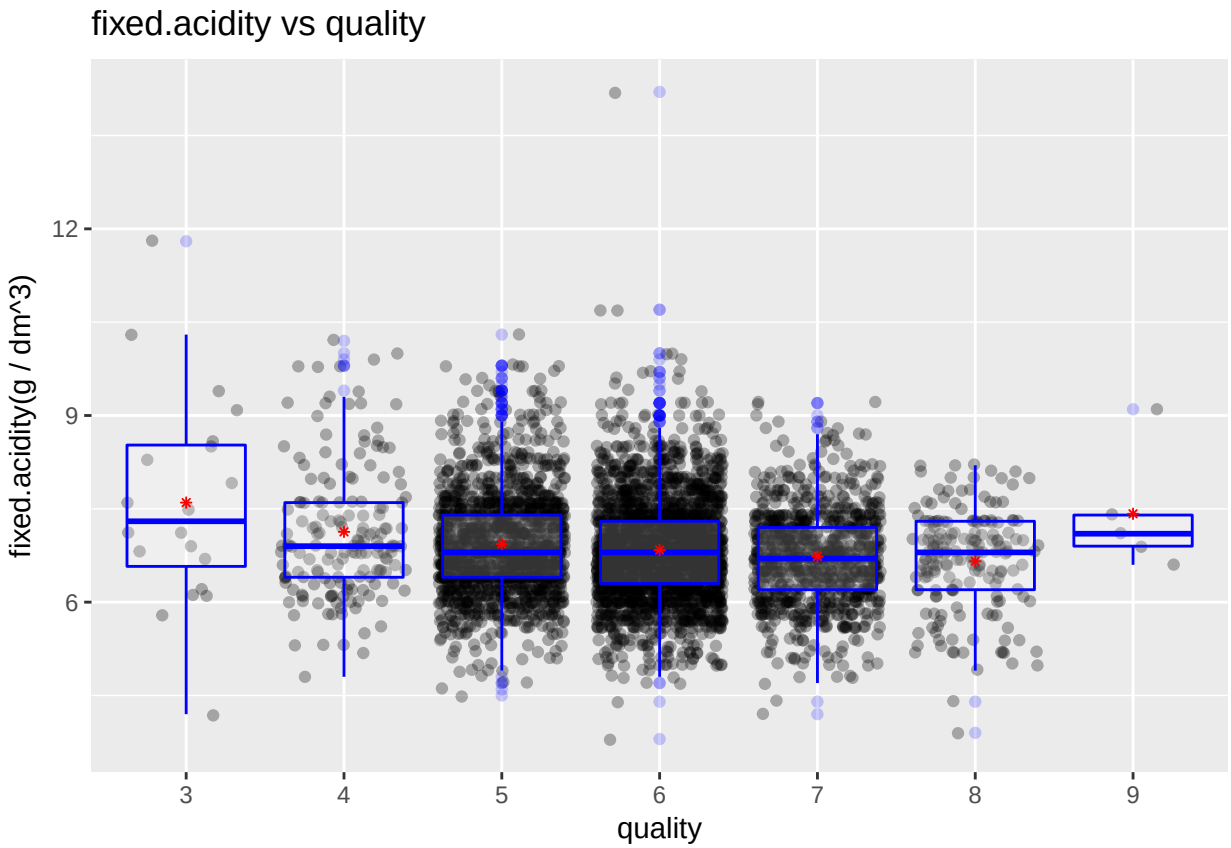




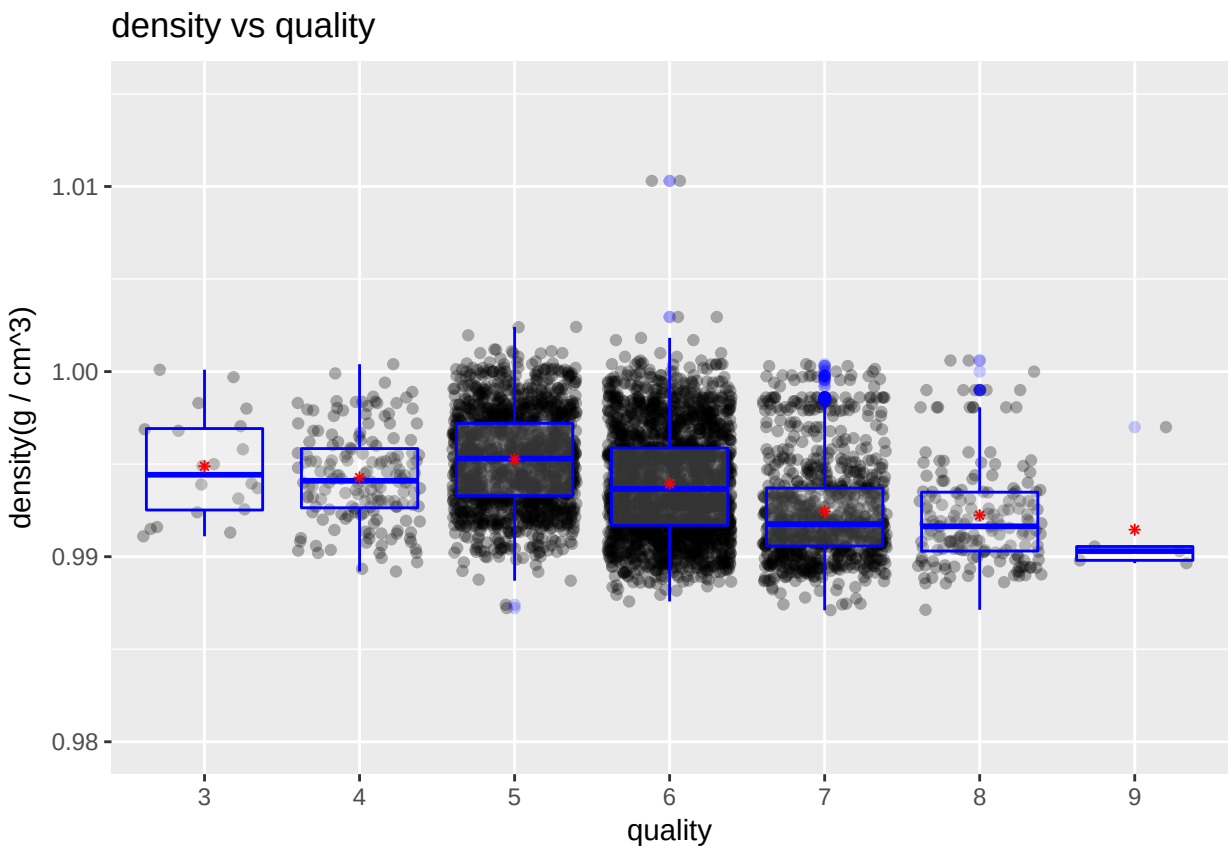
Based on the results, residual.sugar has strong relationship with density, and negative relationship with alcohol.



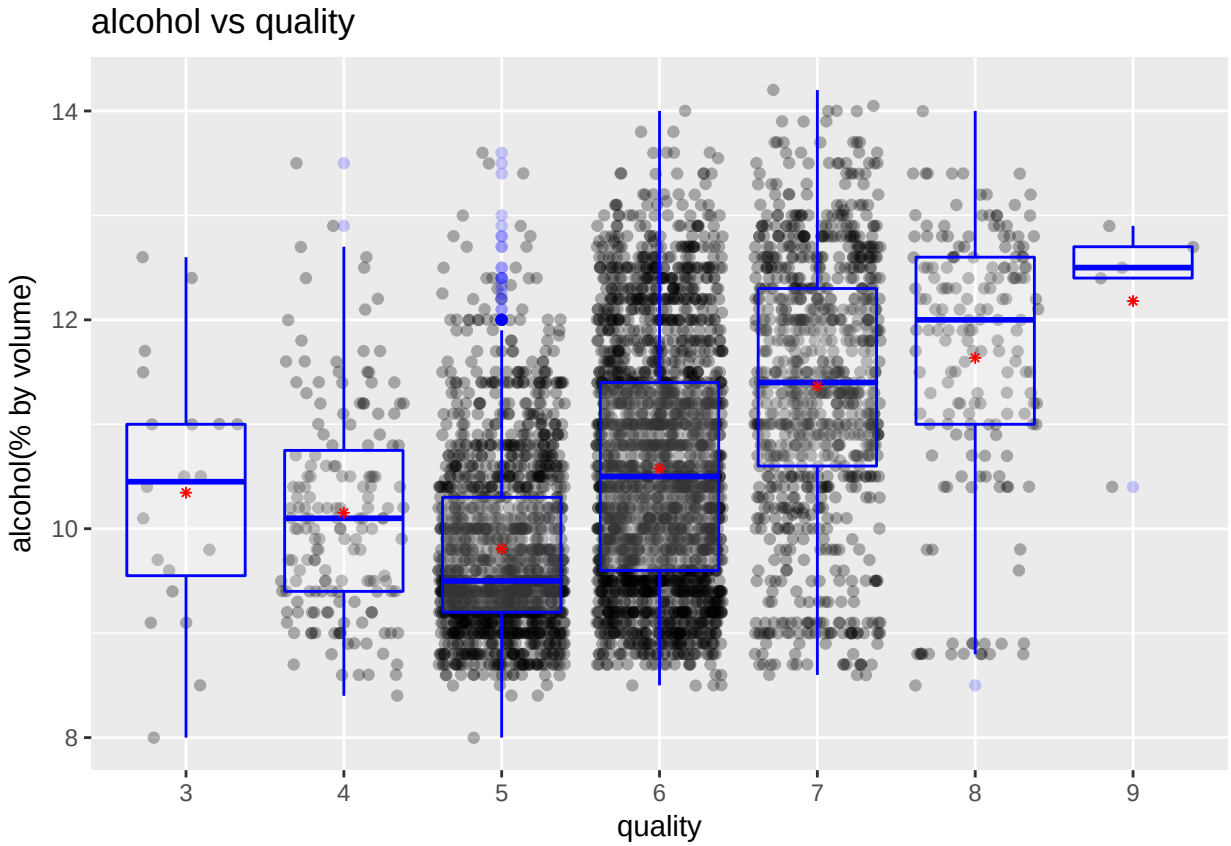
Alcohol and density have negative relationship.



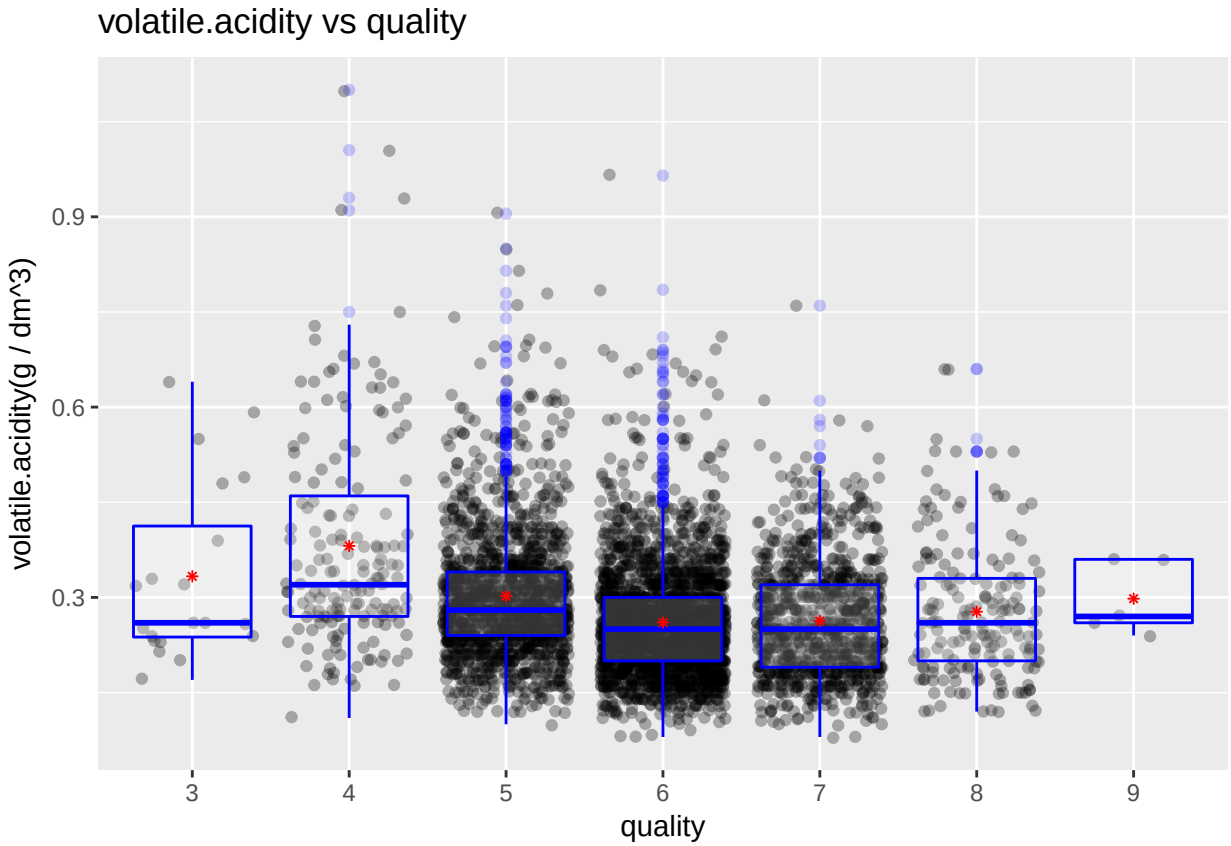
There is no significant change in the mean value for each quality.



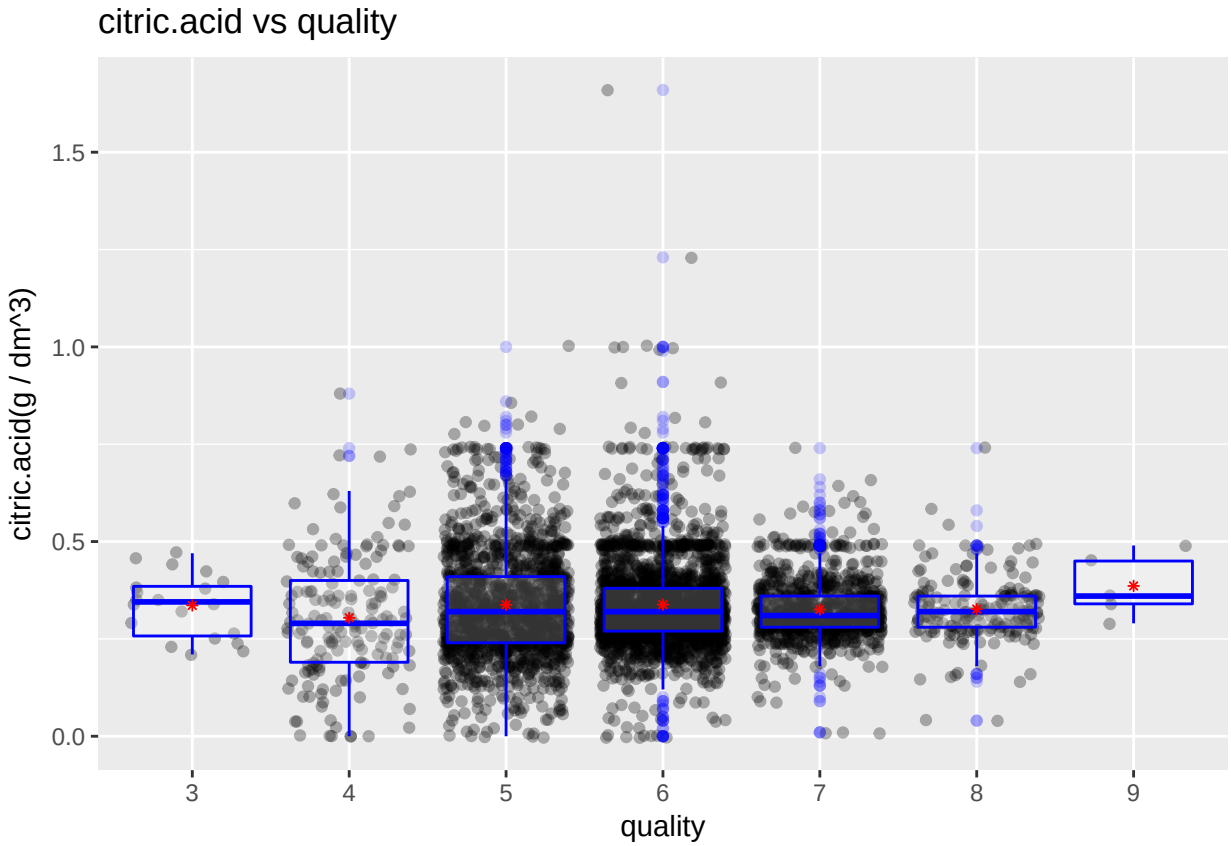
Density and quality have weak negative correlation. The higher density, the better quality.



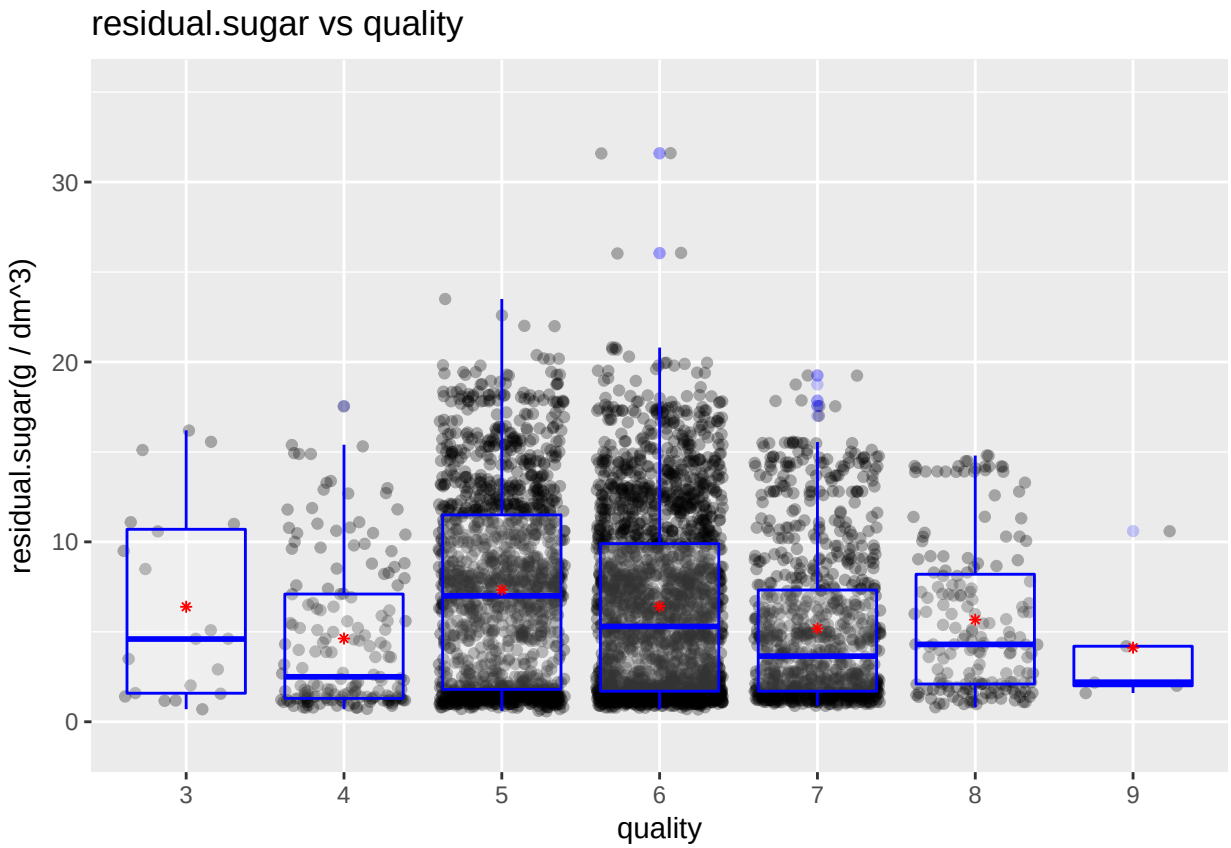
Alcohol and quality have positive correlation, the better quality, the higher alcohol value.



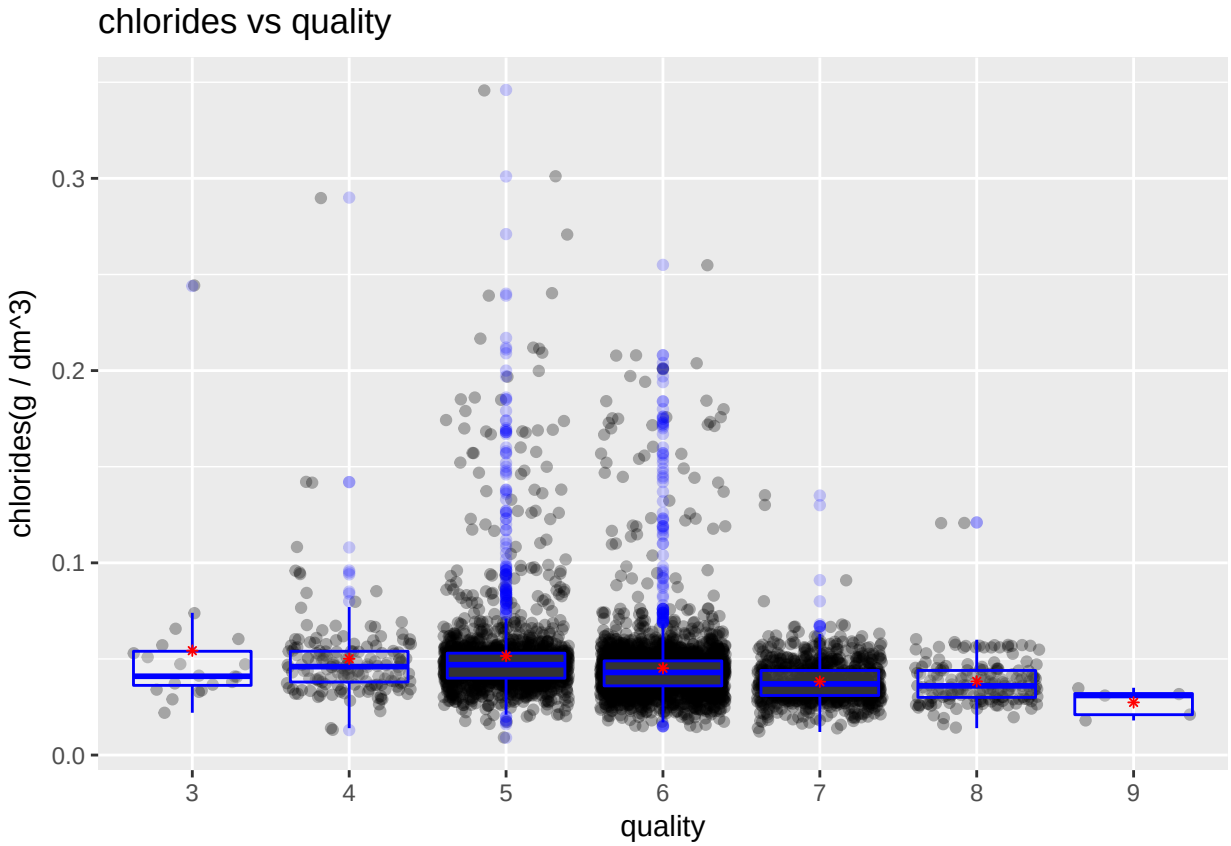
Most of the outliers are in the average quality range.There is no significant change in the mean value for each quality.



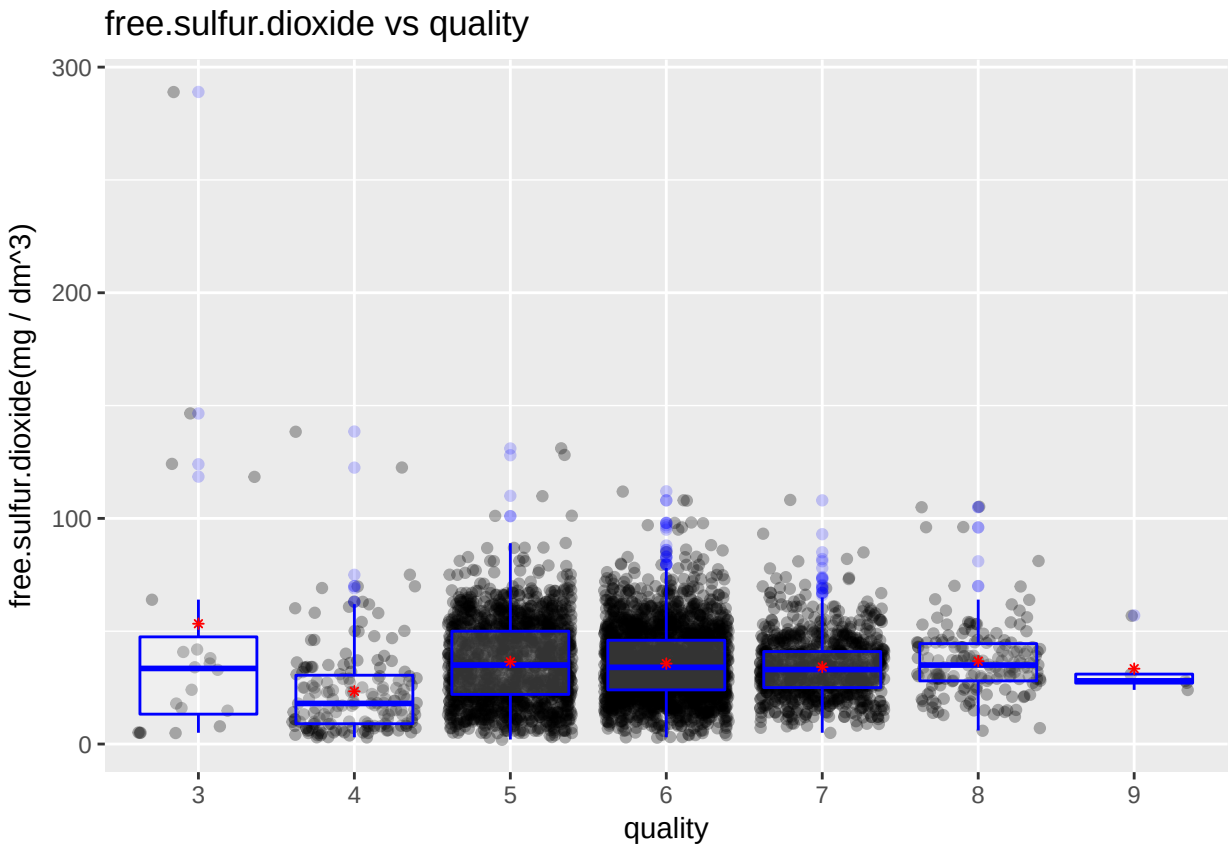
Most of the outliers are in the average quality range.There is no significant change in the mean value for each quality.



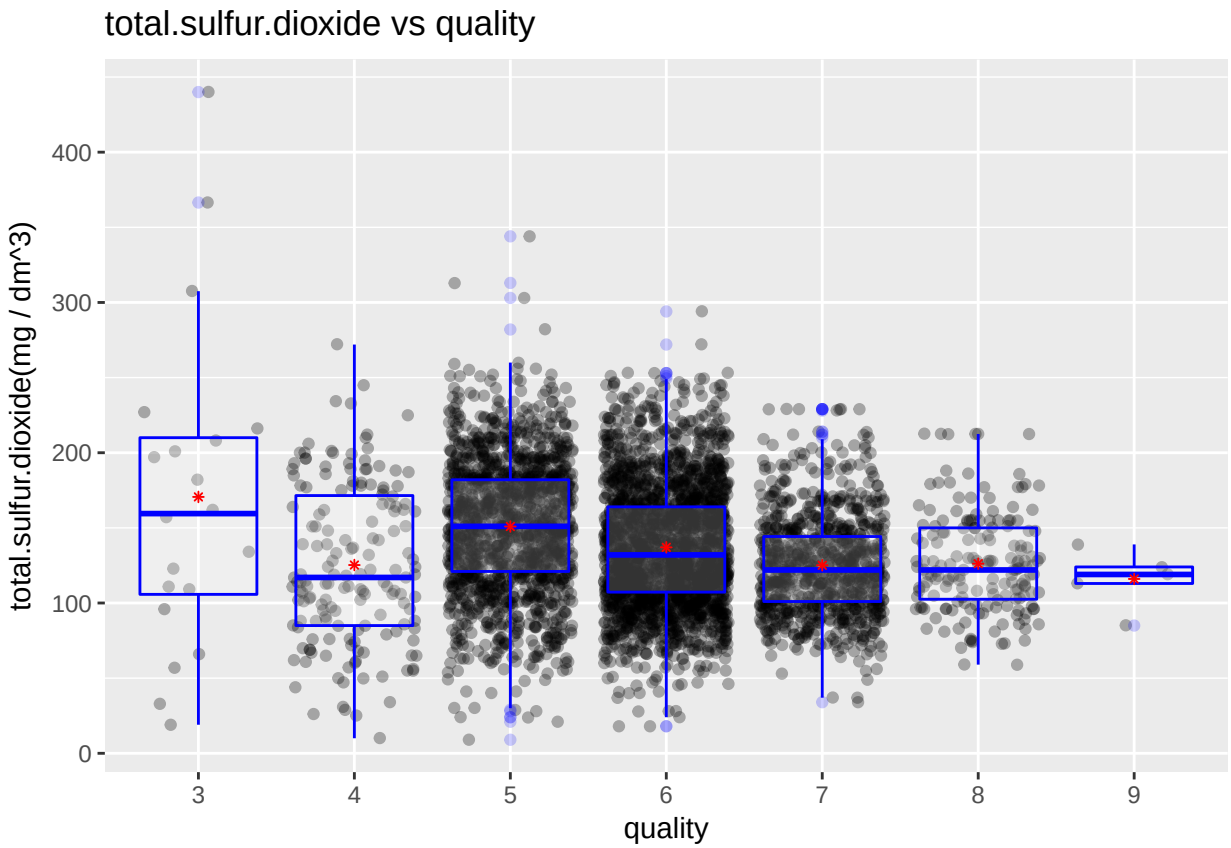
The Residual Sugar almost has no effect on the quality of the white Wine.



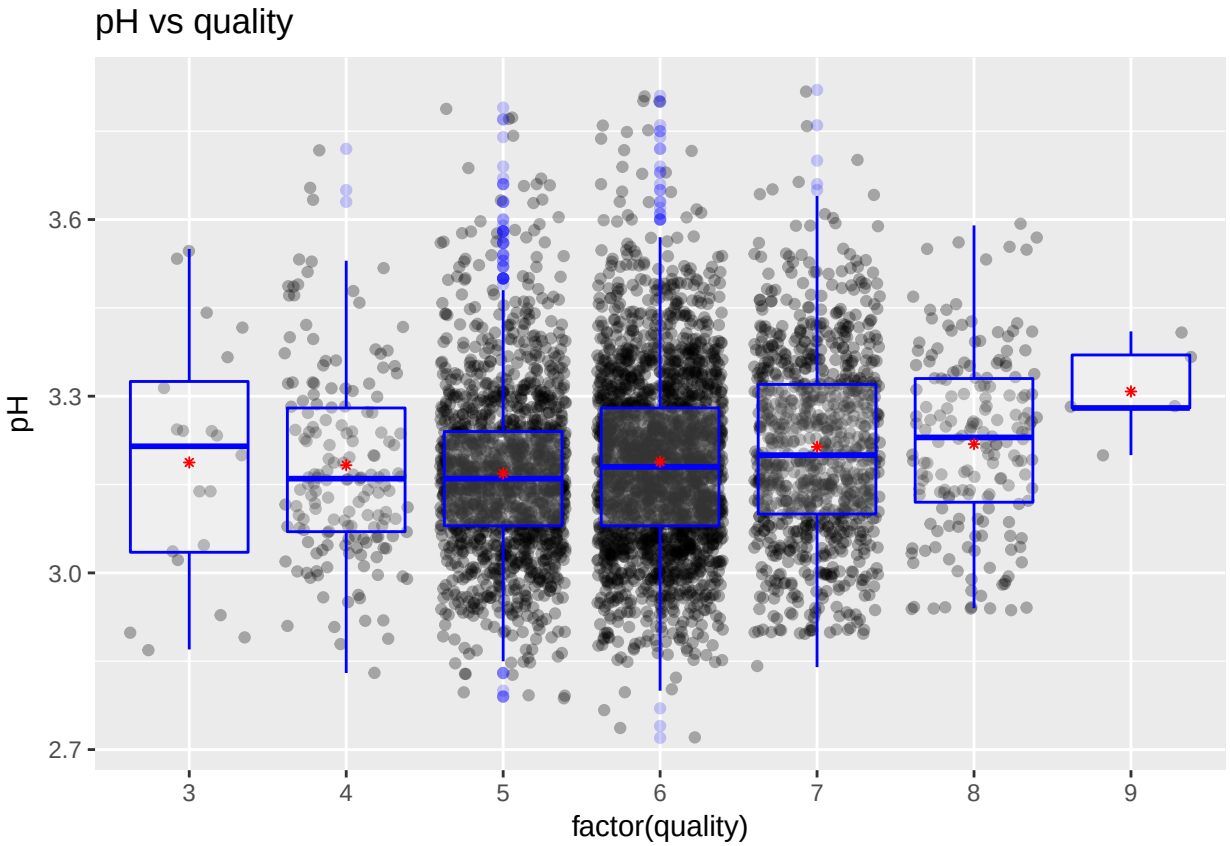
It looks like that lower chlorides create better quality wine.



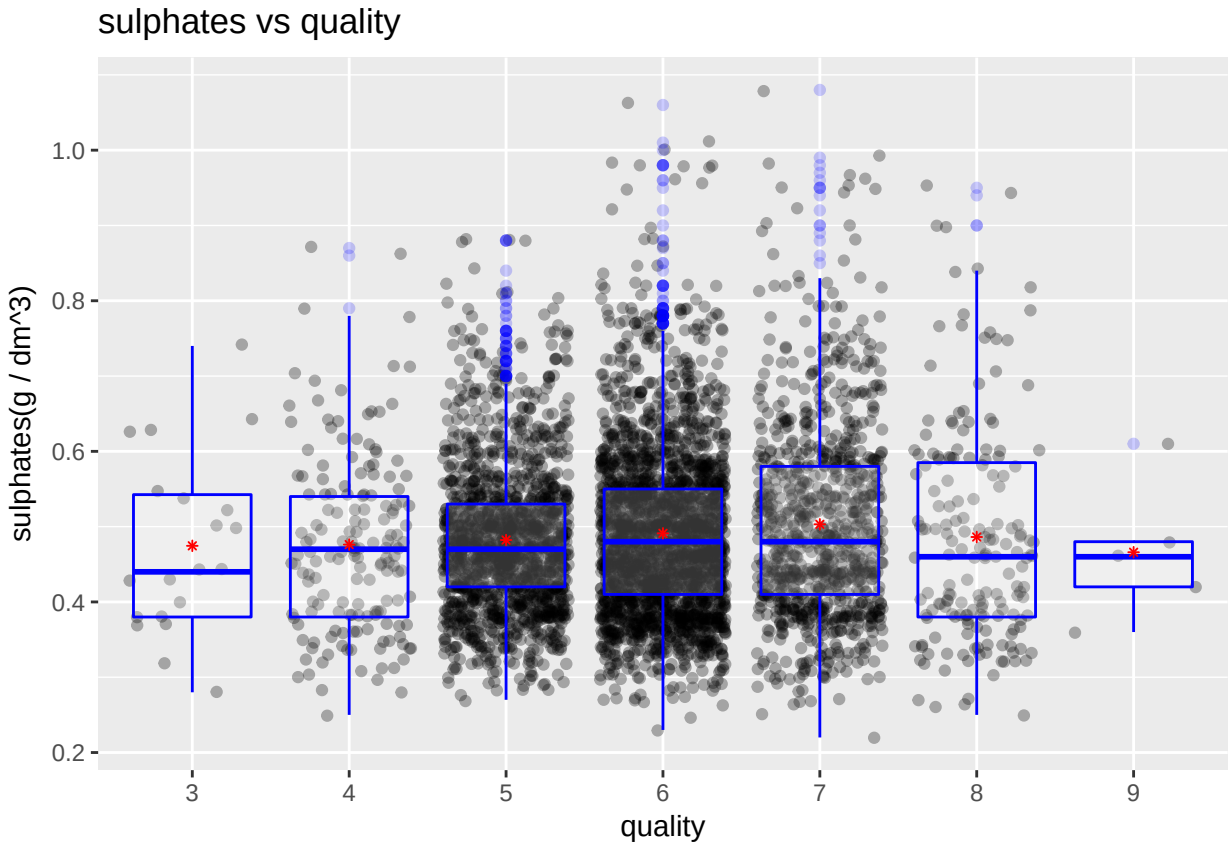
There is no significant change in the mean value for each quality.



The poorest quality has the highest total.sulfur.dioxide.



Most of the outliers are in the average quality range. pH almost has no effect on the quality of the Wine.



Again most of the outliers are in the average quality range. sulphates almost has no effect on the quality of the 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?

fixed.acidity and density(citric.acid and pH) are significant correlated. residual.sugar has strong relationship with density, and negative relationship with alcohol. density has negative relationship with alcohol and quality.

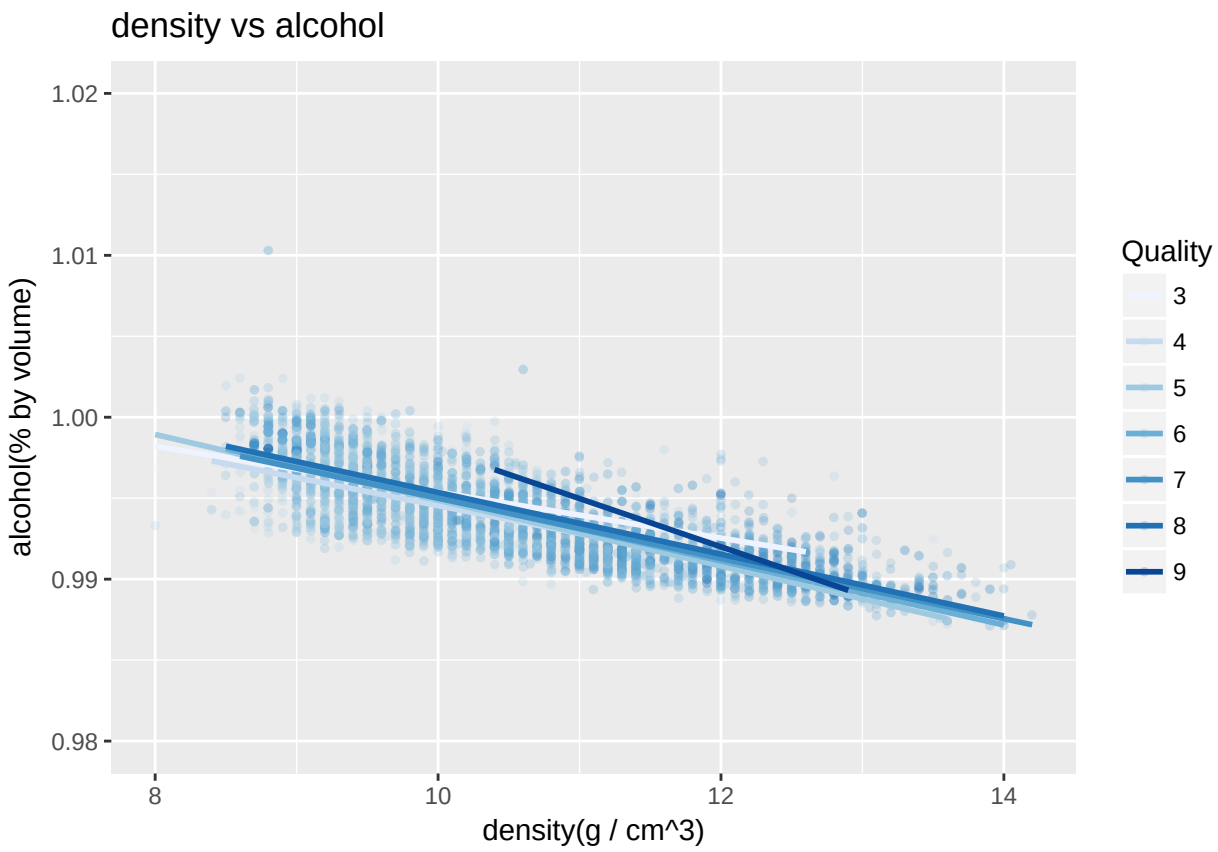
Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

fixed.acidity has positive relationship with density and citric.acid, while it has negative relationship with PH.residual.sugar has strong relationship with density, and negative relationship with alcohol.

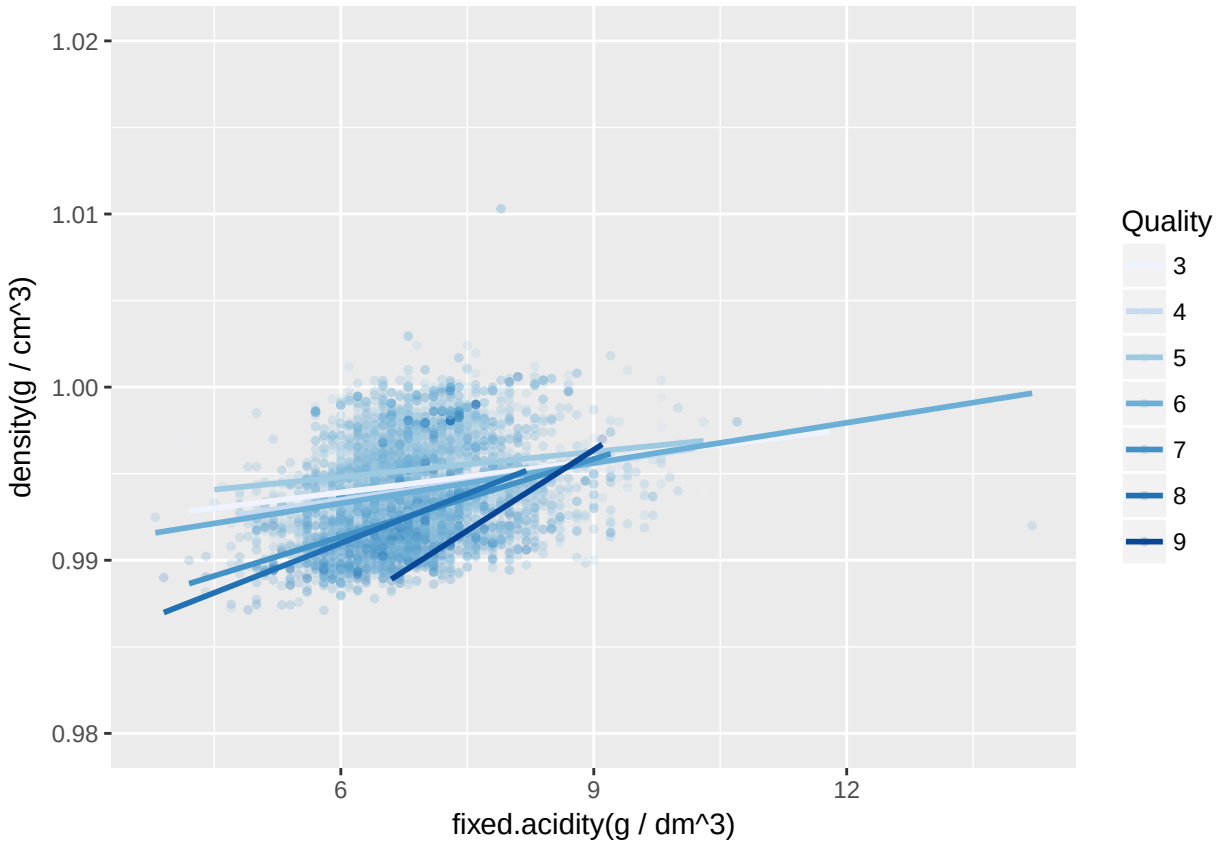
What was the strongest relationship you found?

residual.sugar has strong relationship with density.

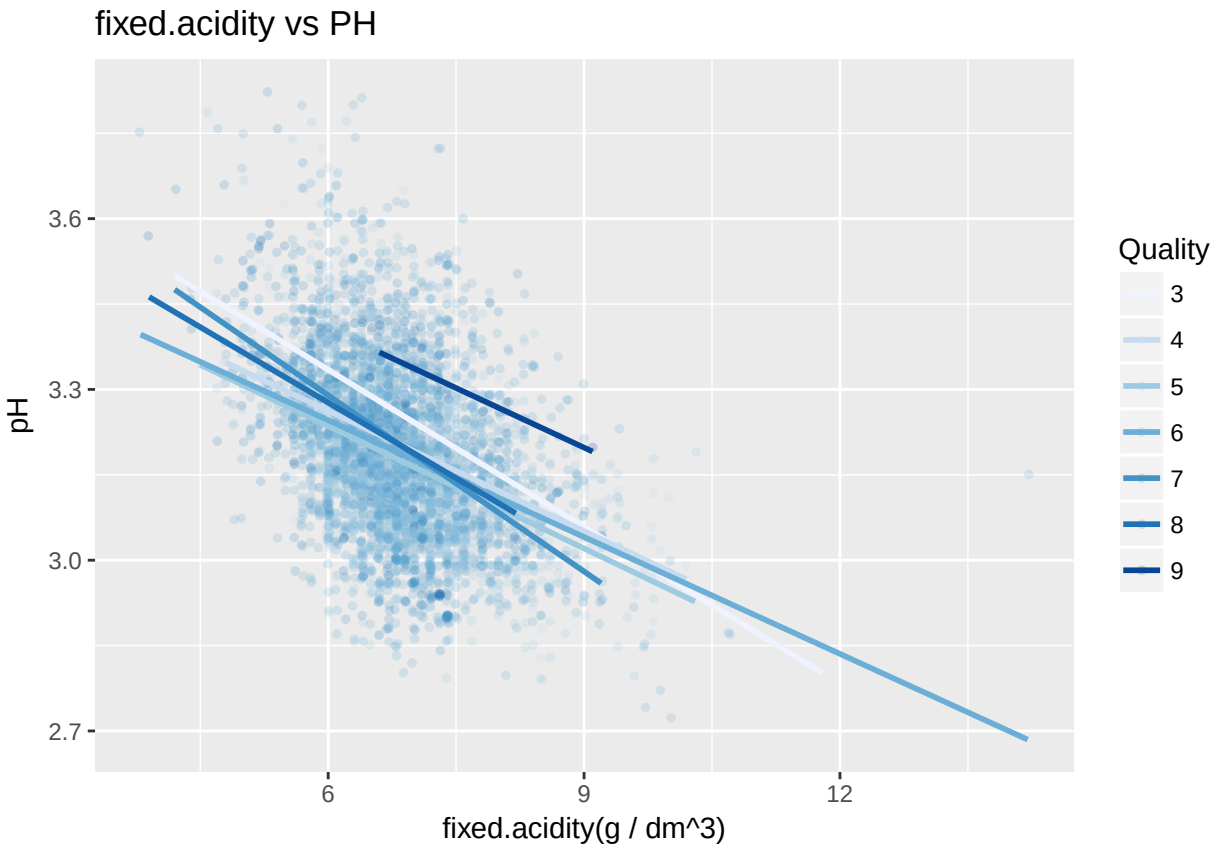
Multivariate Plots Section



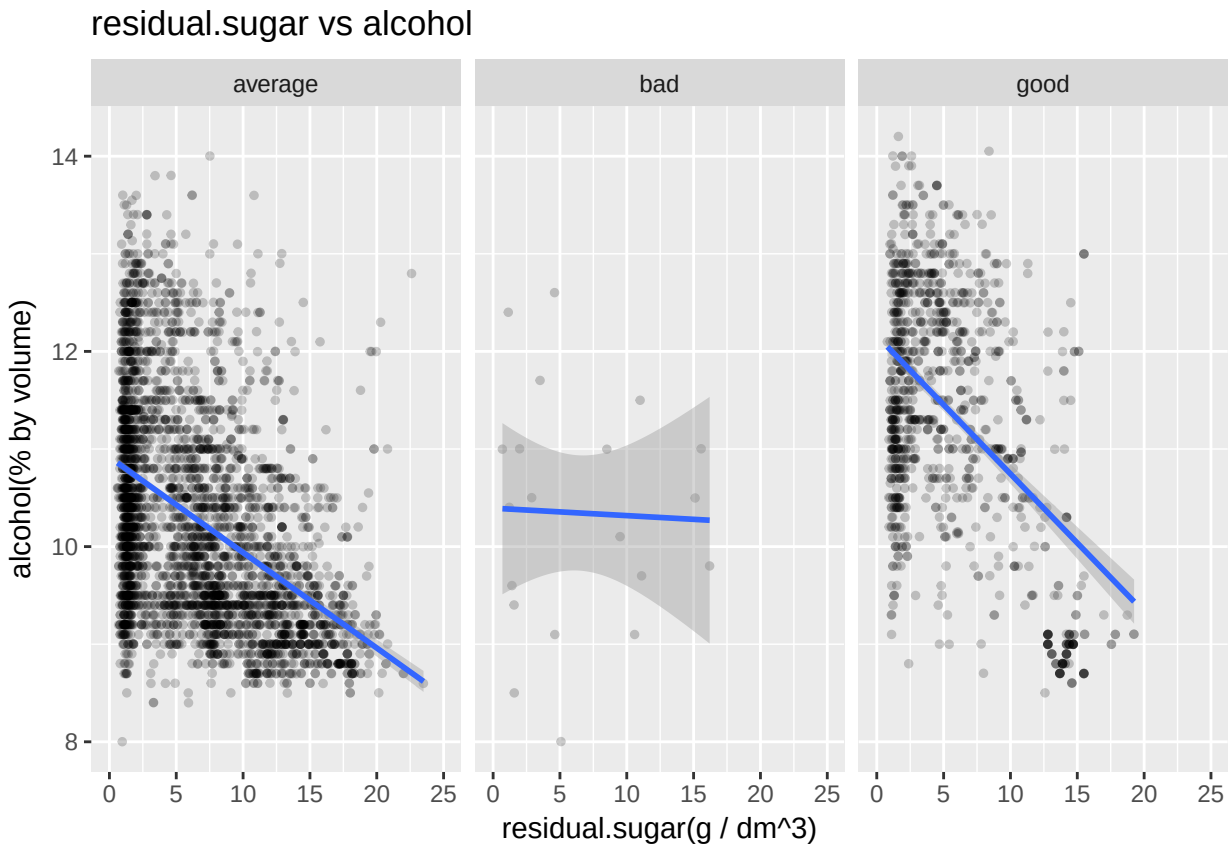
Negative correlation is observed. We don't see significant change will we apply quality g roup.



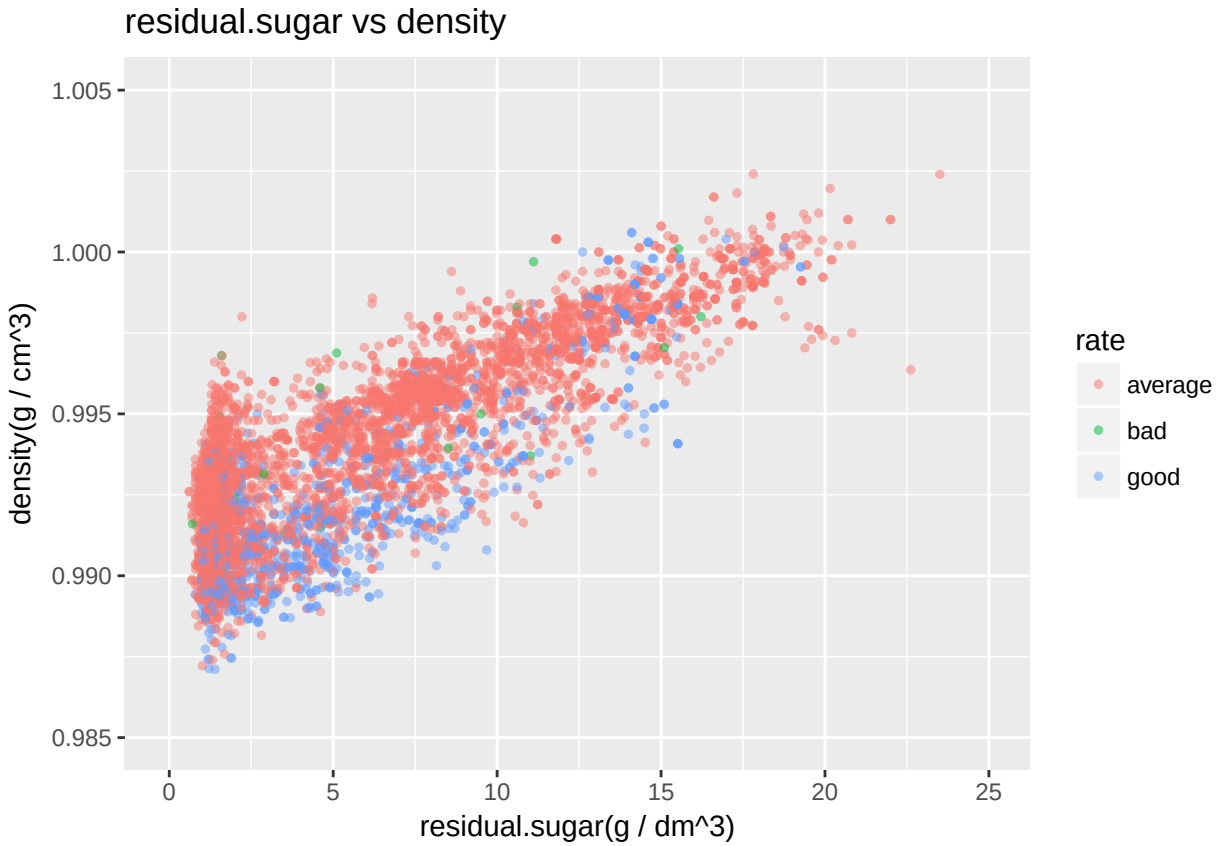
Good quality seems change the most on the density,fixed.acidity change rate.



Quality doesn't have effect on fixed.acidity and pH change rate, all quality groups show almost the same strong negative relationship between pH and fixed.acidity.



The better quality, the strongest negative correlation between alcohol and residual.sugar.



Quality seems don't effect on correlation of density and residual.sugar.

Linear regression model

```
## [[1]]
## [1] 3673   13
##
## [[2]]
## [1] 1225   13
```

Set Seed so that same sample can be reproduced in future also, now Selecting 75% of data as sample from total 'n' rows of the data

```
## Start:  AIC=-2117.07
## quality ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
##     chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + pH + sulphates + alcohol
##
##                        Df Sum of Sq    RSS     AIC
## - citric.acid           1     0.036 2050.6 -2119.0
## - total.sulfur.dioxide  1     0.077 2050.6 -2118.9
## - chlorides             1     0.534 2051.0 -2118.1
## <none>                              2050.5 -2117.1
## - fixed.acidity         1     2.164 2052.7 -2115.2
## - free.sulfur.dioxide   1    12.183 2062.7 -2097.3
```

```
## - sulphates             1    13.943 2064.4 -2094.2
## - pH                    1    15.649 2066.2 -2091.1
## - density               1    19.093 2069.6 -2085.0
## - residual.sugar        1    40.019 2090.5 -2048.1
## - alcohol               1    40.553 2091.1 -2047.1
## - volatile.acidity      1   107.991 2158.5 -1930.5
##
## Step:  AIC=-2119
## quality ~ fixed.acidity + volatile.acidity + residual.sugar +
##     chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + pH + sulphates + alcohol
##
##                         Df Sum of Sq    RSS     AIC
## - total.sulfur.dioxide  1     0.074 2050.6 -2120.9
## - chlorides             1     0.510 2051.1 -2120.1
## <none>                              2050.6 -2119.0
## - fixed.acidity         1     2.271 2052.8 -2116.9
## - free.sulfur.dioxide   1    12.263 2062.8 -2099.1
## - sulphates             1    14.032 2064.6 -2095.9
## - pH                    1    15.648 2066.2 -2093.1
## - density               1    19.068 2069.6 -2087.0
## - residual.sugar        1    39.988 2090.5 -2050.1
## - alcohol               1    41.140 2091.7 -2048.0
## - volatile.acidity      1   111.724 2162.3 -1926.1
##
## Step:  AIC=-2120.87
## quality ~ fixed.acidity + volatile.acidity + residual.sugar +
##     chlorides + free.sulfur.dioxide + density + pH + sulphates +
##     alcohol
##
##                        Df Sum of Sq    RSS     AIC
## - chlorides             1     0.513 2051.1 -2121.9
## <none>                              2050.6 -2120.9
## - fixed.acidity         1     2.287 2052.9 -2118.8
## - sulphates             1    13.958 2064.6 -2097.9
## - pH                    1    15.674 2066.3 -2094.9
## - free.sulfur.dioxide   1    17.161 2067.8 -2092.3
## - density               1    20.170 2070.8 -2086.9
## - residual.sugar        1    40.921 2091.5 -2050.3
## - alcohol               1    41.075 2091.7 -2050.0
## - volatile.acidity      1   117.726 2168.3 -1917.8
##
## Step:  AIC=-2121.95
## quality ~ fixed.acidity + volatile.acidity + residual.sugar +
##     free.sulfur.dioxide + density + pH + sulphates + alcohol
##
##                        Df Sum of Sq    RSS     AIC
## <none>                              2051.1 -2121.9
## - fixed.acidity         1     2.731 2053.9 -2119.1
## - sulphates             1    14.192 2065.3 -2098.6
## - free.sulfur.dioxide   1    16.856 2068.0 -2093.9
## - pH                    1    17.279 2068.4 -2093.1
## - density               1    22.081 2073.2 -2084.6
## - alcohol               1    41.149 2092.3 -2051.0
```

```
## - residual.sugar       1    44.996 2096.1 -2044.2
## - volatile.acidity     1   119.915 2171.1 -1915.3
```

```
##
## Call:
## lm(formula = quality ~ fixed.acidity + volatile.acidity + residual.sugar +
##     free.sulfur.dioxide + density + pH + sulphates + alcohol,
##     data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3692 -0.5005 -0.0340  0.4613  3.1294
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.295e+02  2.036e+01   6.357 2.30e-10 ***
## fixed.acidity        5.122e-02  2.319e-02   2.209   0.0273 *
## volatile.acidity    -1.867e+00  1.276e-01 -14.636  < 2e-16 ***
## residual.sugar       7.423e-02  8.280e-03   8.965  < 2e-16 ***
## free.sulfur.dioxide  4.418e-03  8.051e-04   5.487 4.36e-08 ***
## density             -1.296e+02  2.064e+01  -6.280 3.77e-10 ***
## pH                   6.570e-01  1.183e-01   5.556 2.96e-08 ***
## sulphates            5.730e-01  1.138e-01   5.035 5.01e-07 ***
## alcohol              2.321e-01  2.707e-02   8.574  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7482 on 3664 degrees of freedom
## Multiple R-squared:  0.2855, Adjusted R-squared:  0.284
## F-statistic:   183 on 8 and 3664 DF,  p-value: < 2.2e-16
```

So our model is quality = 129.5+0.05122*fixed.acidity-1.867*volatile.acidity+0.07423*residual.sugar+0.004418*free.sulfur.dioxide-129.6*density+0.657*pH+0.573*sulphates+0.2321*alcohol

```
## [1] 0.7902041
```

Use this model for test data set, find the fitted quality and rate value, then compare with the real value, compute the accuracy is 79%.

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 first three plots from the Multivariate section suggest that I can build a linear model and use those variables in the model to predict the quality of white wine.

Were there any interesting or surprising interactions between features?

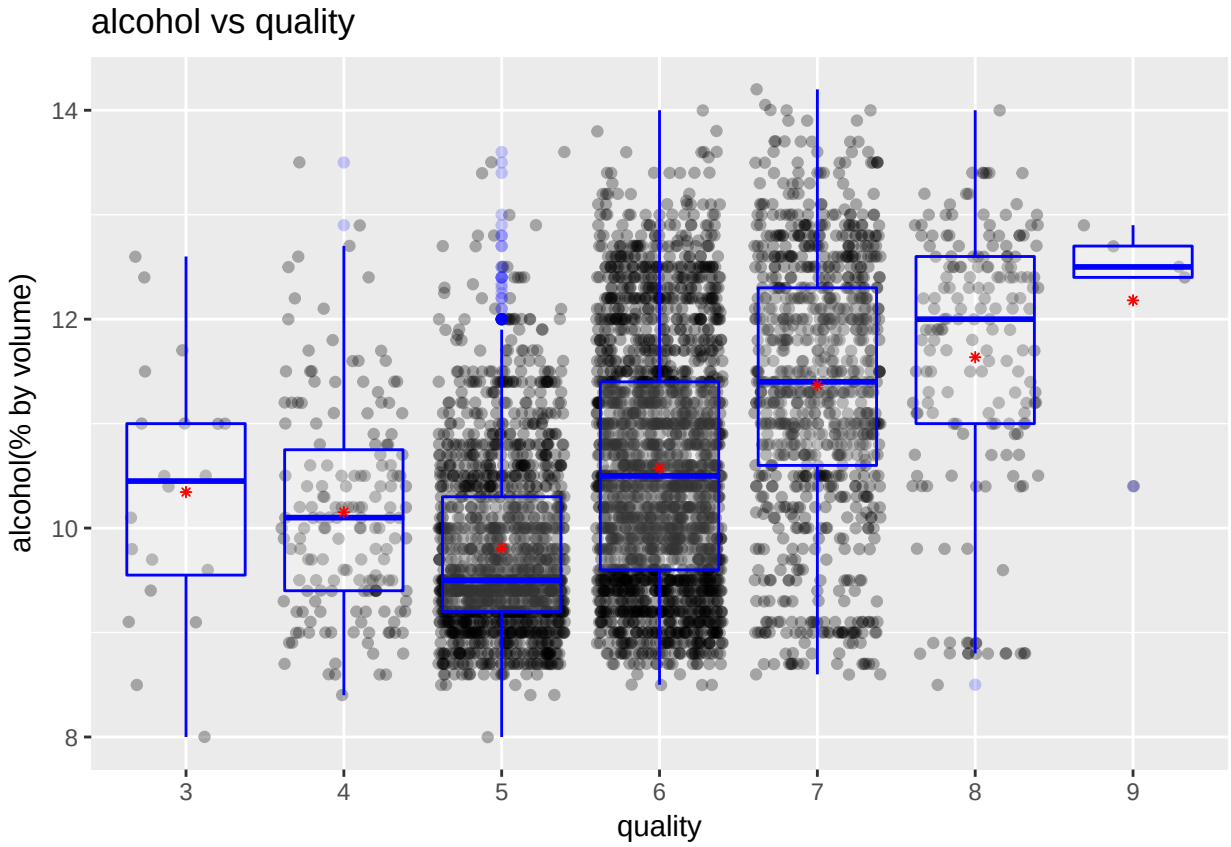
Good quality seems change the most on the density,fixed.acidity change rate.

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

I create a linear model to predict the quality, which contains fixed.acidity, volatile.acidity, residual.sugar, free.sulfur.dioxide, density, pH, sulphates and alcohol, based on the fitted result, my model has an accuracy of 79%, which is not bad. The limitation is that I don't consider the interaction terms between variables, and also the non-linear model may work well than mine.

Final Plots and Summary

Plot One



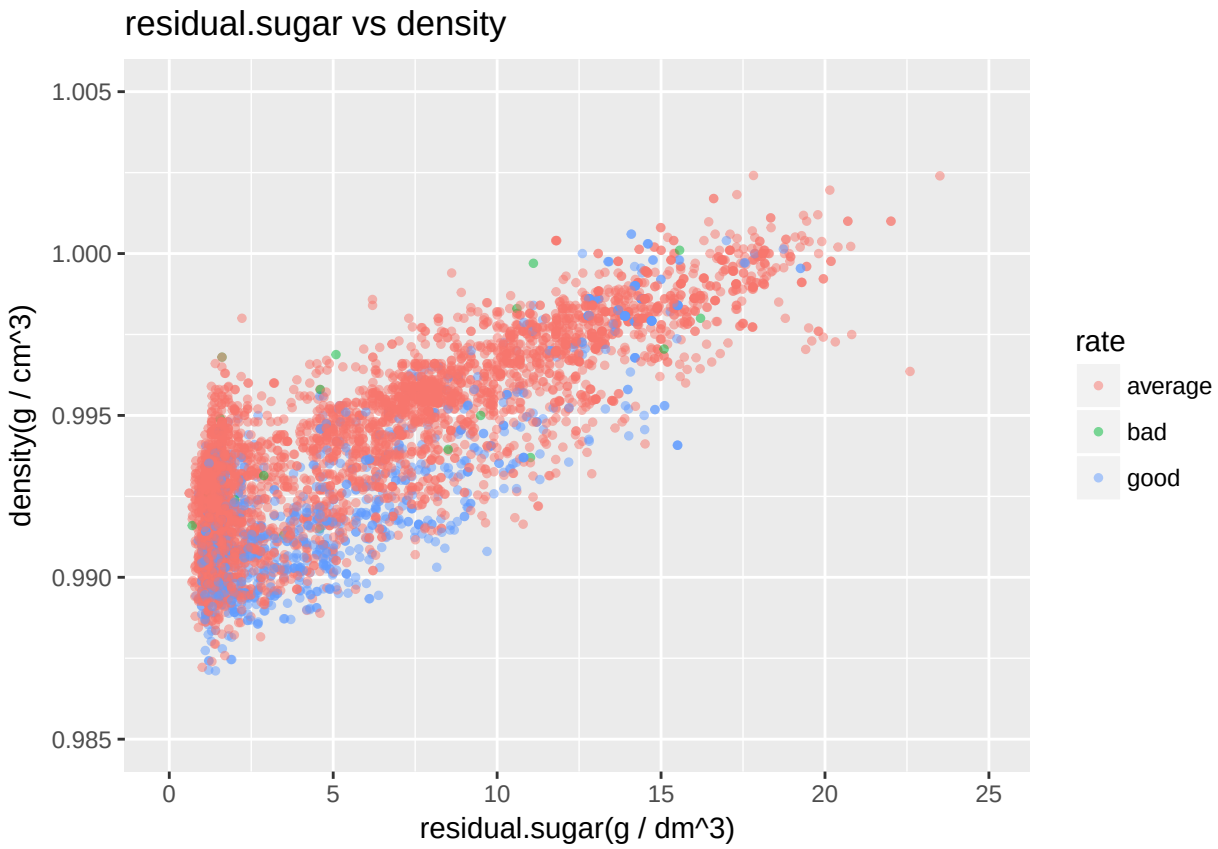
Description One

When we mention wine, the first word come into my brain will be alcohol. This plot shows for each quality score, what's summarize of the corresponding alcohol. In each step we can see the positive influence of alcohol

in a wine's quality score.

Plot Two

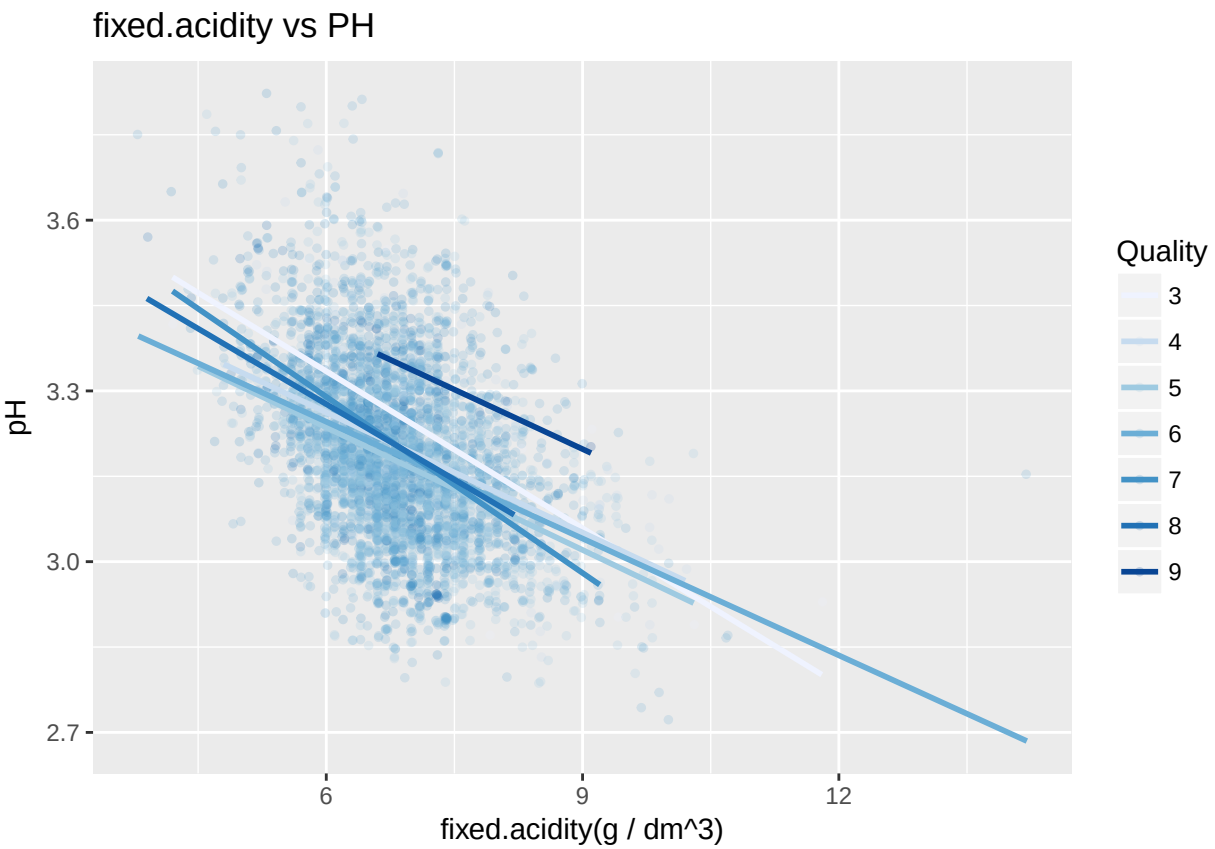
```
## Warning: Removed 5 rows containing missing values (geom_point).
```



Description Two

This plot shows the relationships between the other features(not the main feature). residual.sugar and density show the strongest correlation among all wine parameters. No matter how the quality of wine change, the strong correlation stay the same.

Plot Three



Description Three

pH and fixed.acidity have negative correlation, most of the pH scores are between 2.75 and 3.75, and fixed.acidity values are between 6 and 9 g/dmˆ3. For each quality score, we don't see significant change of correlation between pH and fixed.acidity.

Reflection

The data set contains information on 4898 white wine sampels with 12 variables, I started with the univariate analysis to know the structure of the data set, and features of each variables. Then explore the correlations between each variables to find their relationships. Finally find the variables which make effects on quality score, and create a linear model to predict the quality score.

Alcohol and density have clear correlation with quality, the better quality, the higher alcohol value. I was surprised that pH did not have a strong positive correlation with quality. In the good rate range, it loos like the mean pH value increase with quality score increase. But it doesn't show the same trend in bad and average group. For the linear model part, the accuracy is 79%, I think it is not bad.

A limitation of the current analysis is that the current data consists of samples collected from 2009, which is too old. And also the number of data is a little bit small. In future, I would like to improve linear models for

prediction of wine quality and let the prediction results more accurate. I May try other models like logistic regression and non-linear regression.

Reference

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553. ISSN: 0167-9236.