White Wine Quality Exploration by Swain Tseng

The report explore a dataset containing quality and attributes for 4898 white wines. This dataset has 13 variables which contains 9 variables of ingredient, 2 variables of physical properity, 1 variable of quantity and 1 of quality. The reason why I pick this dataset is to analysize what ingredients would influence quality of white wine and find some relationship between every variables and if I could do that maybe one day we can use computer to measure quality in the future.

Univariate Plots Section

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

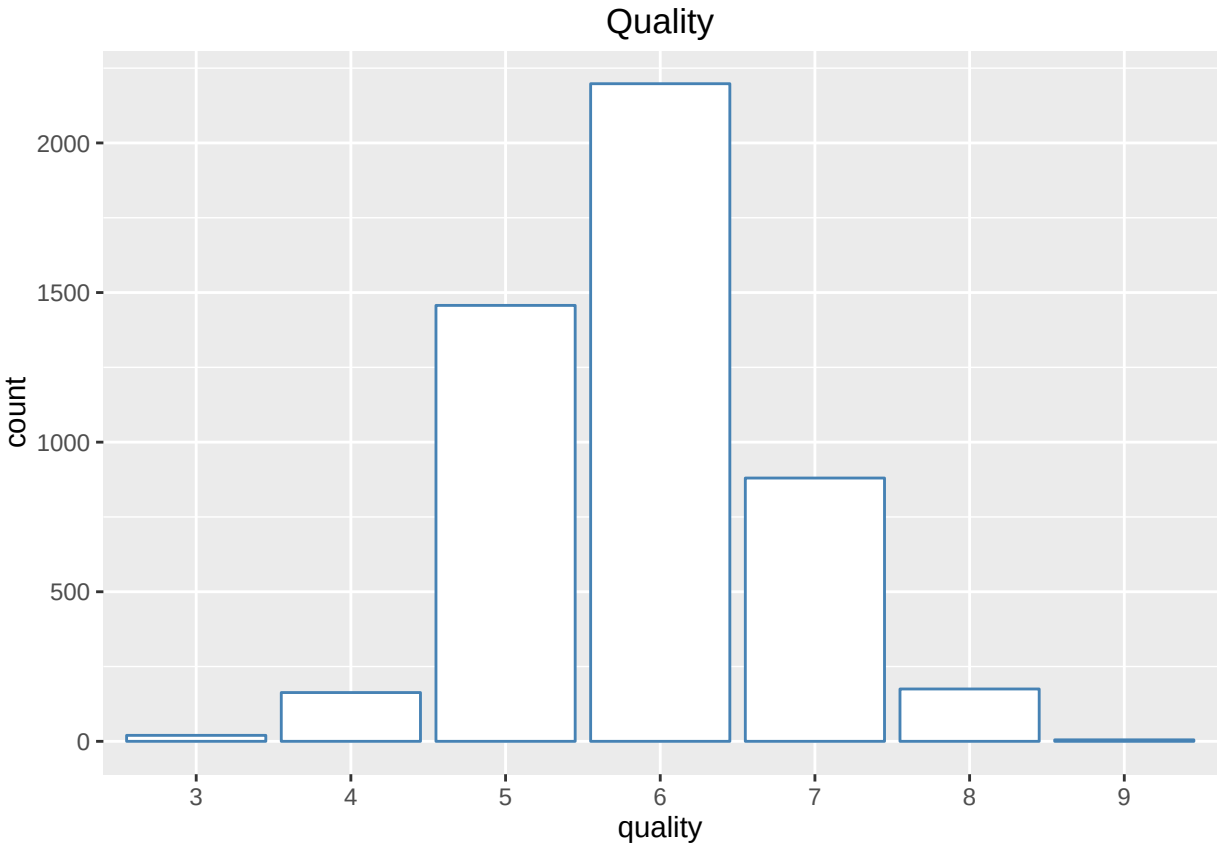
```
##  [1] "X"                    "fixed.acidity"        "volatile.acidity"
##  [4] "citric.acid"          "residual.sugar"       "chlorides"
##  [7] "free.sulfur.dioxide"  "total.sulfur.dioxide" "density"
## [10] "pH"                   "sulphates"            "alcohol"
## [13] "quality"
```

```
## 'data.frame':    4898 obs. of  13 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ 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 ...
```

```
##        X        fixed.acidity    volatile.acidity  citric.acid
##  Min.   :   1   Min.   : 3.800   Min.   :0.0800   Min.   :0.0000
##  1st Qu.:1225   1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700
##  Median :2450   Median : 6.800   Median :0.2600   Median :0.3200
##  Mean   :2450   Mean   : 6.855   Mean   :0.2782   Mean   :0.3342
##  3rd Qu.:3674   3rd Qu.: 7.300   3rd Qu.:0.3200   3rd Qu.:0.3900
##  Max.   :4898   Max.   :14.200   Max.   :1.1000   Max.   :1.6600
##  residual.sugar     chlorides       free.sulfur.dioxide
##  Min.   : 0.600   Min.   :0.00900   Min.   :  2.00
##  1st Qu.: 1.700   1st Qu.:0.03600   1st Qu.: 23.00
##  Median : 5.200   Median :0.04300   Median : 34.00
##  Mean   : 6.391   Mean   :0.04577   Mean   : 35.31
##  3rd Qu.: 9.900   3rd Qu.:0.05000   3rd Qu.: 46.00
##  Max.   :65.800   Max.   :0.34600   Max.   :289.00
##  total.sulfur.dioxide    density             pH          sulphates
##  Min.   :  9.0        Min.   :0.9871   Min.   :2.720   Min.   :0.2200
##  1st Qu.:108.0        1st Qu.:0.9917   1st Qu.:3.090   1st Qu.:0.4100
```

```
##  Median :134.0        Median :0.9937   Median :3.180   Median :0.4700
##  Mean   :138.4        Mean   :0.9940   Mean   :3.188   Mean   :0.4898
##  3rd Qu.:167.0        3rd Qu.:0.9961   3rd Qu.:3.280   3rd Qu.:0.5500
##  Max.   :440.0        Max.   :1.0390   Max.   :3.820   Max.   :1.0800
##     alcohol         quality
##  Min.   : 8.00   Min.   :3.000
##  1st Qu.: 9.50   1st Qu.:5.000
##  Median :10.40   Median :6.000
##  Mean   :10.51   Mean   :5.878
##  3rd Qu.:11.40   3rd Qu.:6.000
##  Max.   :14.20   Max.   :9.000
```

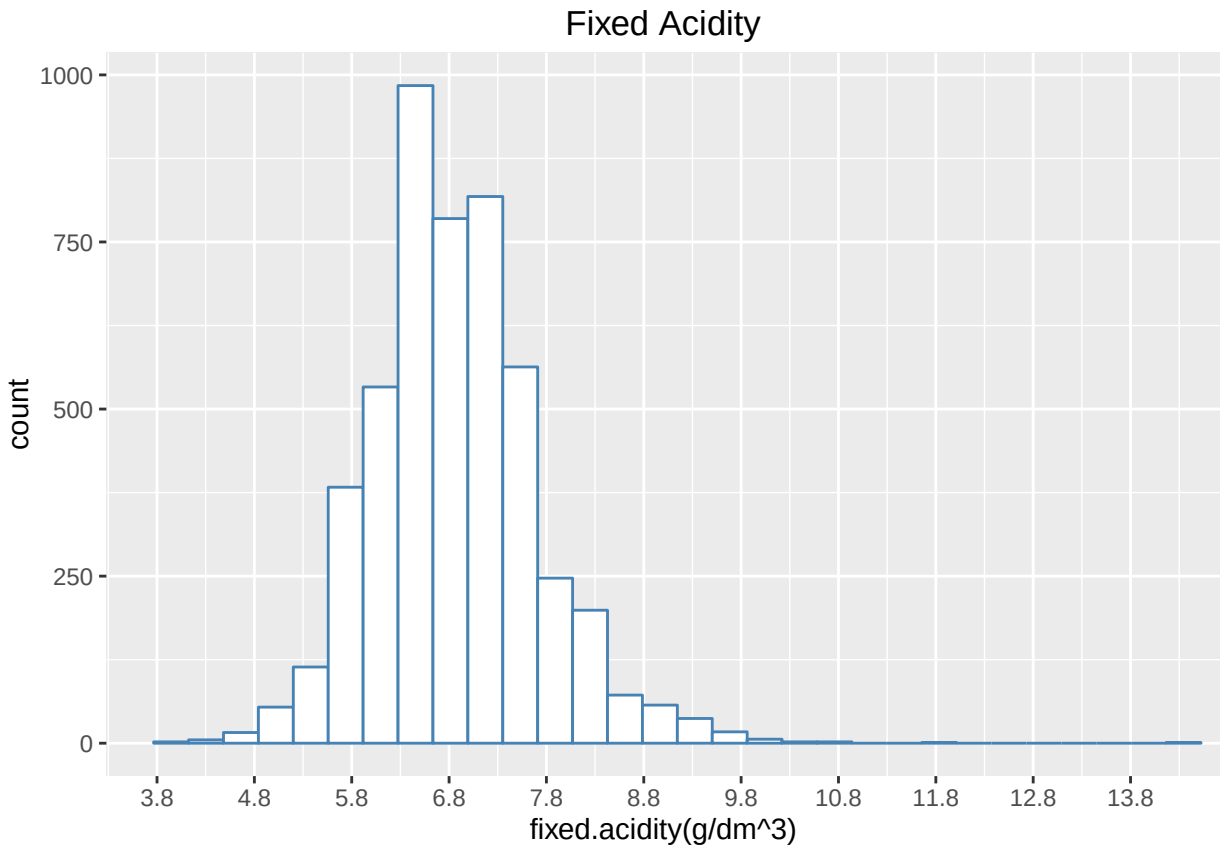
This dataset contain 13 variables and 4898 objects.




```
##    3    4    5    6    7    8    9
##   20  163 1457 2198  880  175    5
```

The reason I created this plot is to know the distribution of white wine in quality.

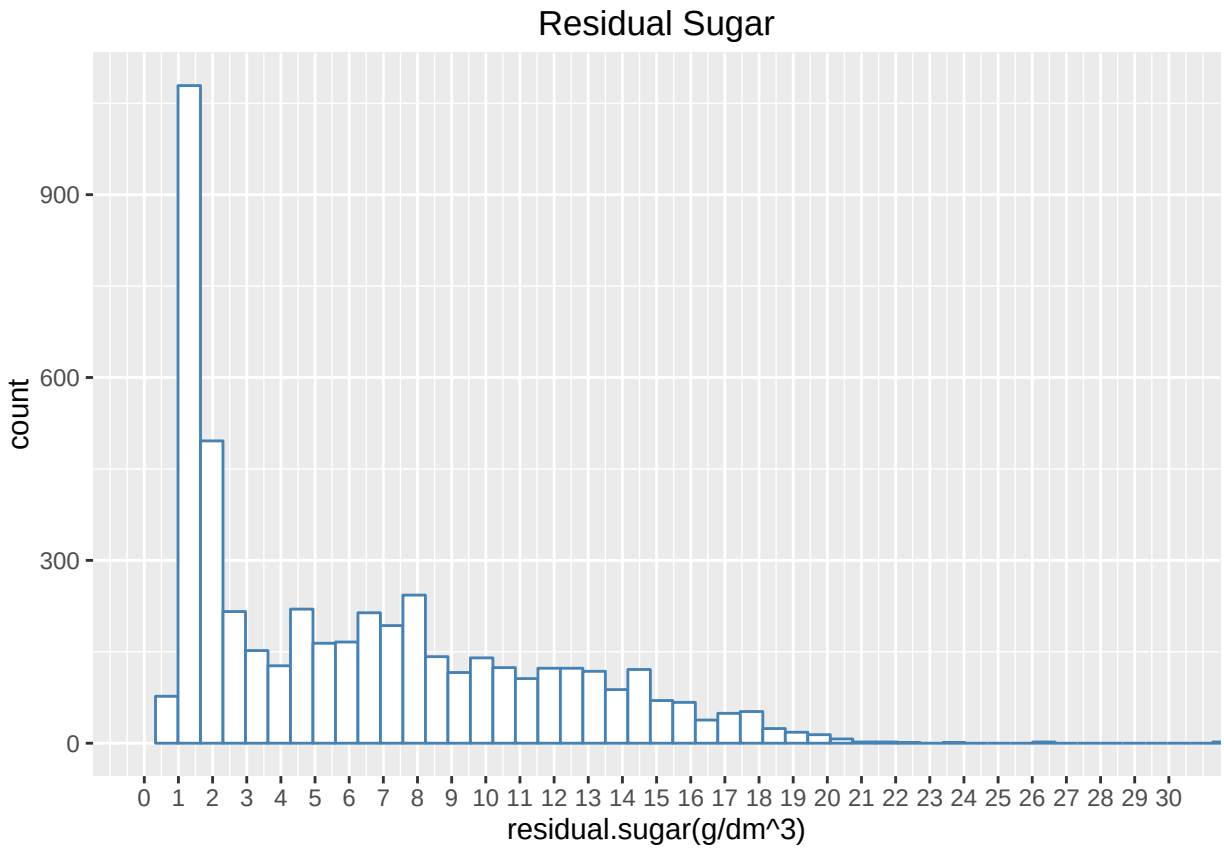
This bar chart displays that most of white wines are quality 6 then 5,7 and we want to know what ingredients will influence the quality.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.800   6.300   6.800   6.855   7.300  14.200
```

The reason I created this plot is to know the distribution of white wine in fixed acidity.

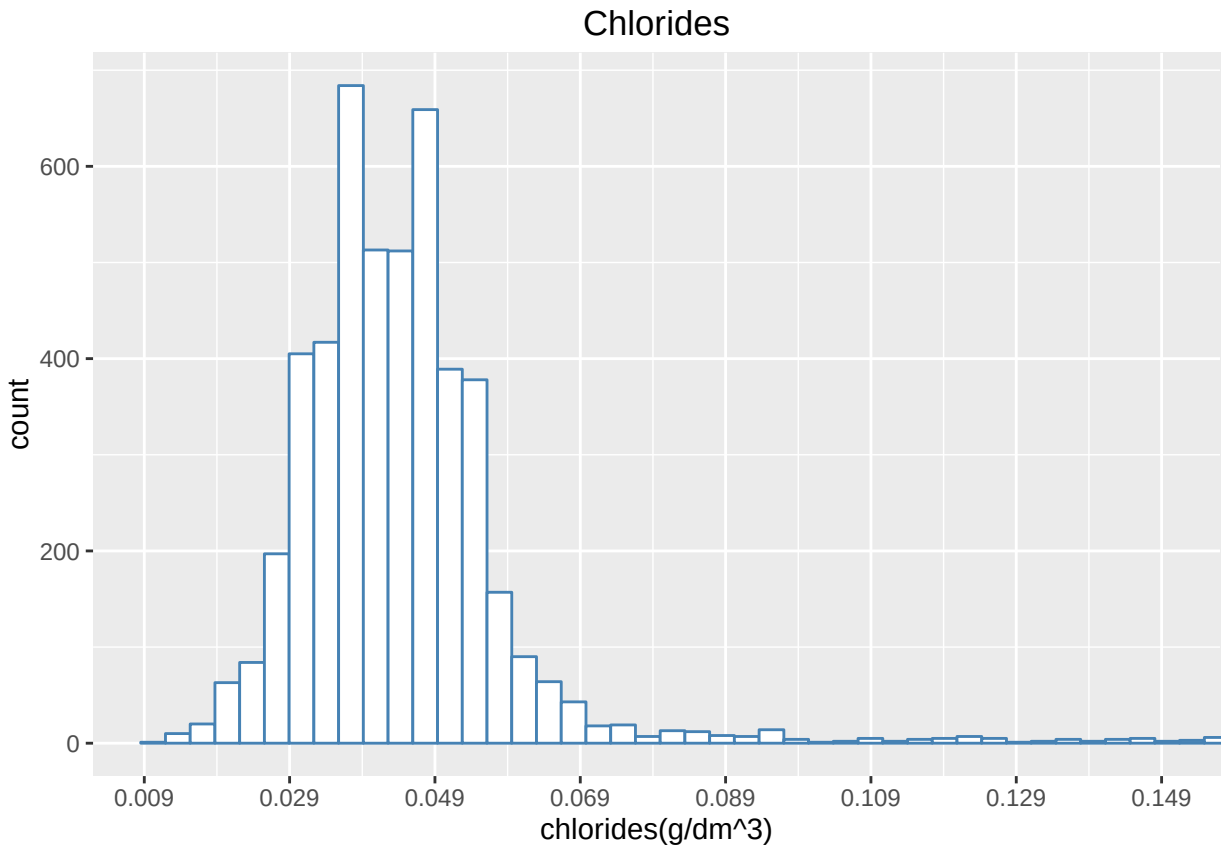
Most wines have fixed acidity between 5.8(g/dm^3) - 7.8(g/dm^3)




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.600   1.700   5.200   6.391   9.900  65.800
```

The reason I created this plot is to know the distribution of white wine in residual sugar.

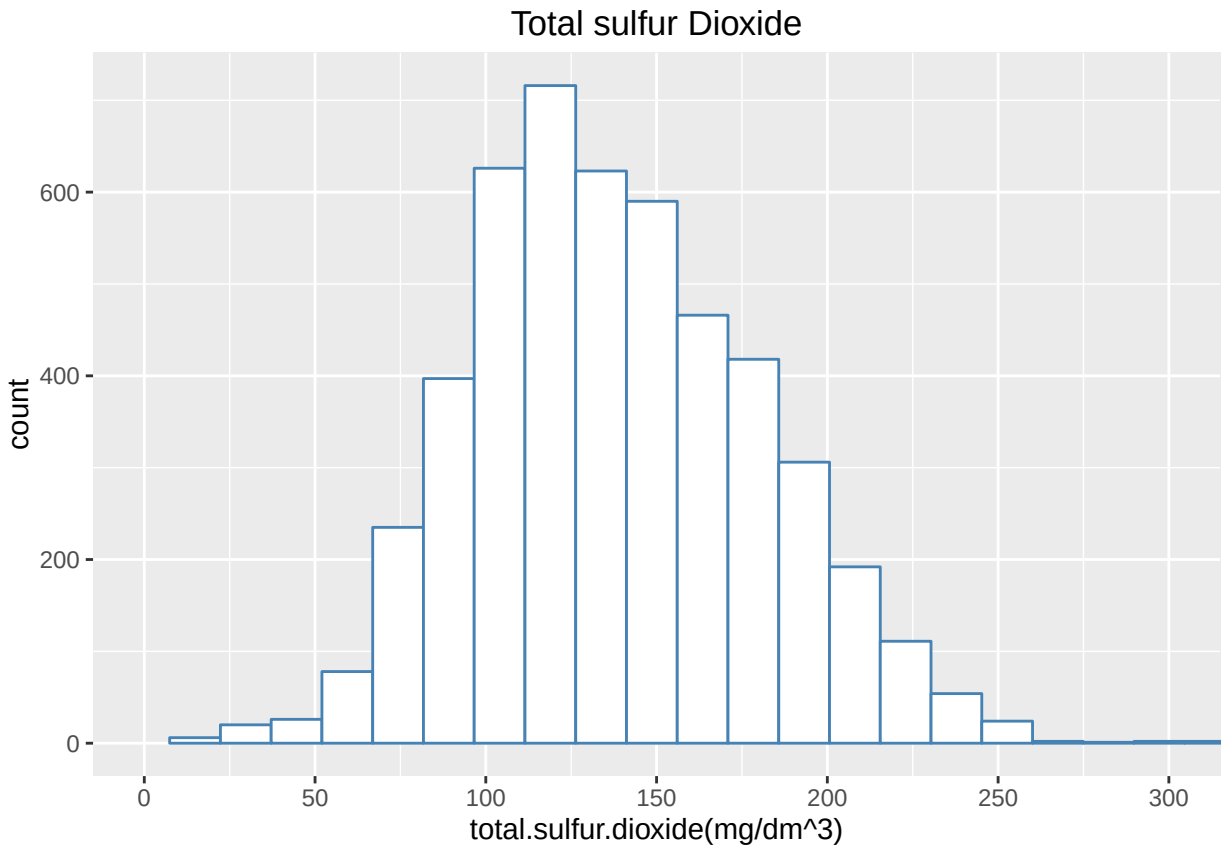
Most wines have residual sugar at 2(g/dmˆ3) and according to the summary result, it shows that Max:65.8(g/dmˆ3) is a outlier.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00900 0.03600 0.04300 0.04577 0.05000 0.34600
```

The reason I created this plot is to know the distribution of white wine in chlorides.

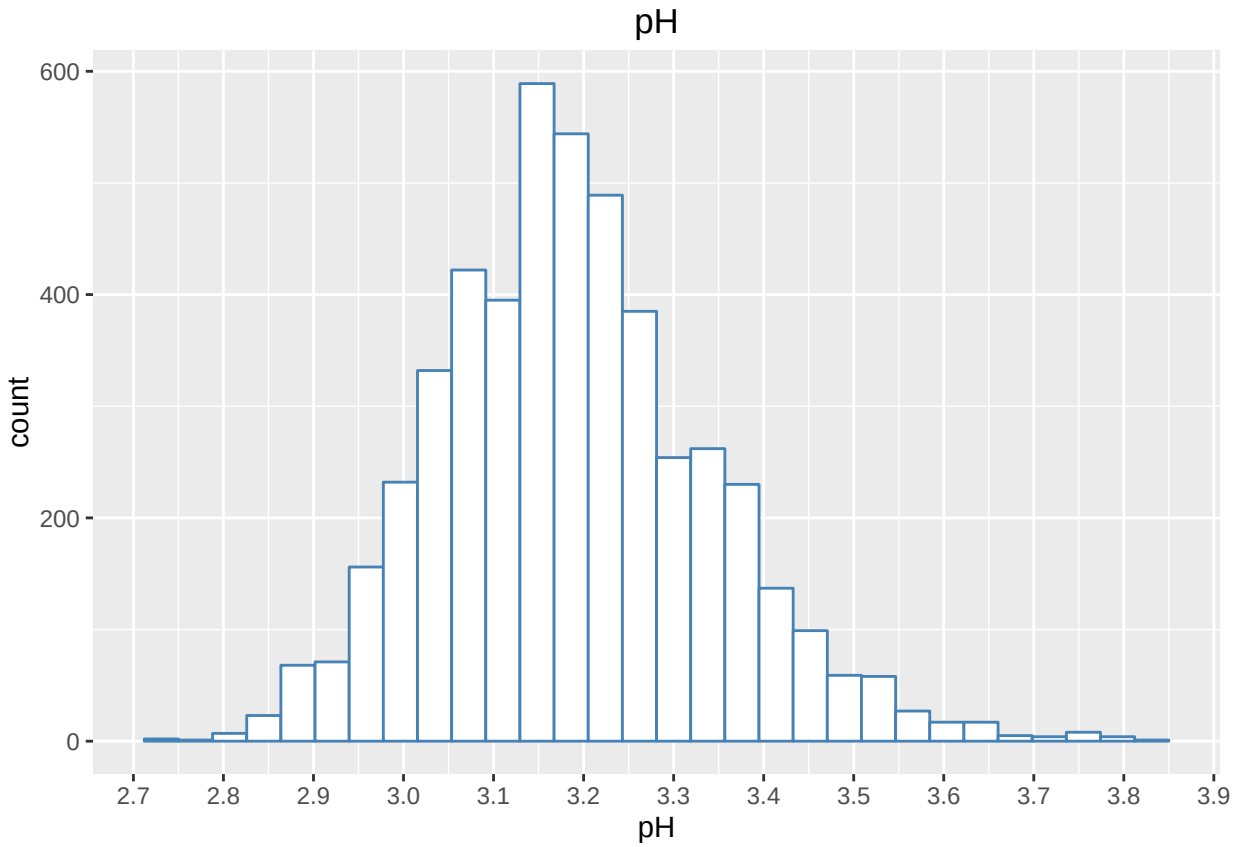
Most wines have chlorides at 0.029(g/dm^3) - 0.059(g/dm^3) and there are some outliers.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     9.0   108.0   134.0   138.4   167.0   440.0
```

The reason I created this plot is to know the distribution of white wine in total sulfur dioxide.

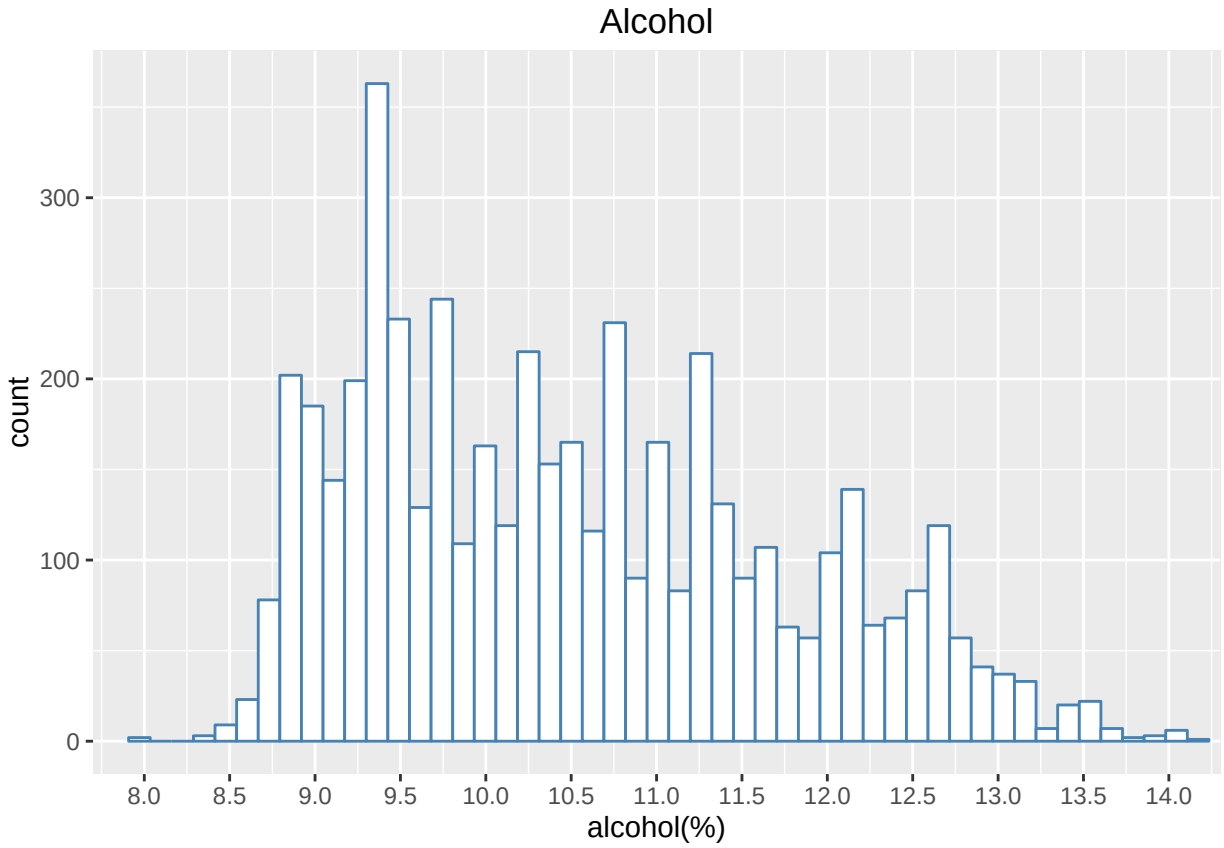
The mean is 138.4 and the median is 134 so we can know total sulfur dioxide is near a normal distribution.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.720   3.090   3.180   3.188   3.280   3.820
```

The reason I created this plot is to know the distribution of white wine in pH.

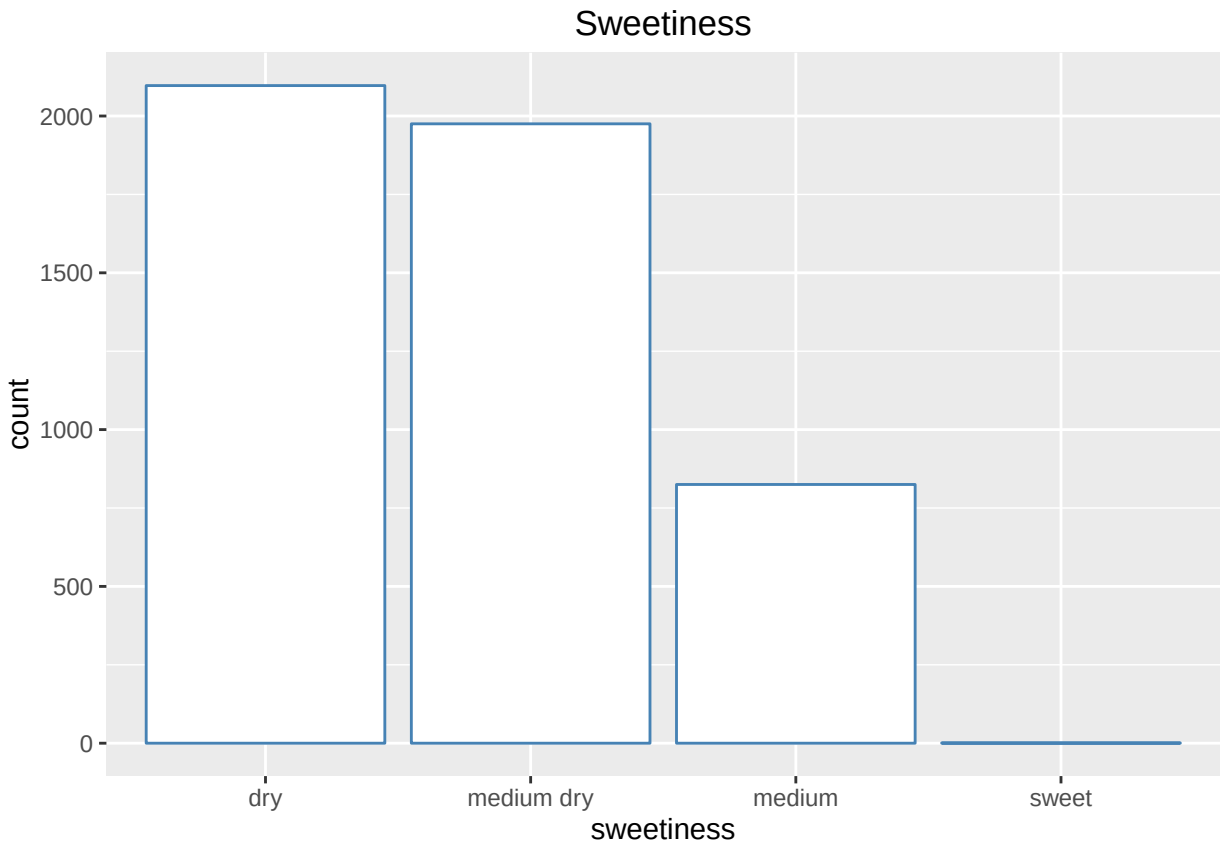
Because The mean is 3.188 and the median is 3.18, we can infer that pH is closer to normal distribution than total sulfur dioxide.




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

The reason I created this plot is to know the distribution of white wine in alcohol.

The plot is an interesting one that is different from others because it doesn't look like a normal distribution. This diagram shows that alcohol distributes more averagely than other variables.




```
##        dry medium dry     medium      sweet
##       2097       1975        825          1
```

The reason I created this plot is to know the distribution of white wine in sweetiness.

Most white wines are dry and medium dry. Dry and medium dry take up 83% of all white wines.

Univariate Analysis

What is the structure of your dataset?

There are 4849 white wines in the dataset with 13 variables. (fixed acidity, volatile acidity, titric acid, residual sugar, chlorides, free sulfur dixoxide, total sulfur dioxide, density, pH, sulphates, alcohol, quality) There are all numeric variables and the most important variable quality is also so I transfer it into factor.

worst ———-> best

quality 0,1,2,3,4,5,6,7,8,9,10

Other observations:

· **Most white wines are at quality 6**

· **Lots of white wines have near 1.5(g/dmˆ3) residual sugar**

· **Alcohol is not a normal distribution but its' mean and median are almost the same**

· **Dry and medium dry take up 83% of all white wines**

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

The main feature of interest in this dataset is quality because I want to find what ingredients(fixed acidity, residual sugar, chlorides, total sulfur dioxide, alcohol) would exactually affect the quality of white wines.

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

The reason why I choose the ingredients is that ingredients will influence the taste of wine and the quality variable is graded by expert.

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

I create a sweetiness variable to separate white wines by sweetiness because it will be more clear to know the sweetiness of white wines. There are four classes of sweetiness(dry, medium dry, medium, sweet).

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?

The quality variable displays that a normal distribution. There are almost 5000 white wines in the dataset and there is no 0,1,2,10. As the result, I think that it might be a bias because the expert who graded white wines doesn't want to rank too high or low.

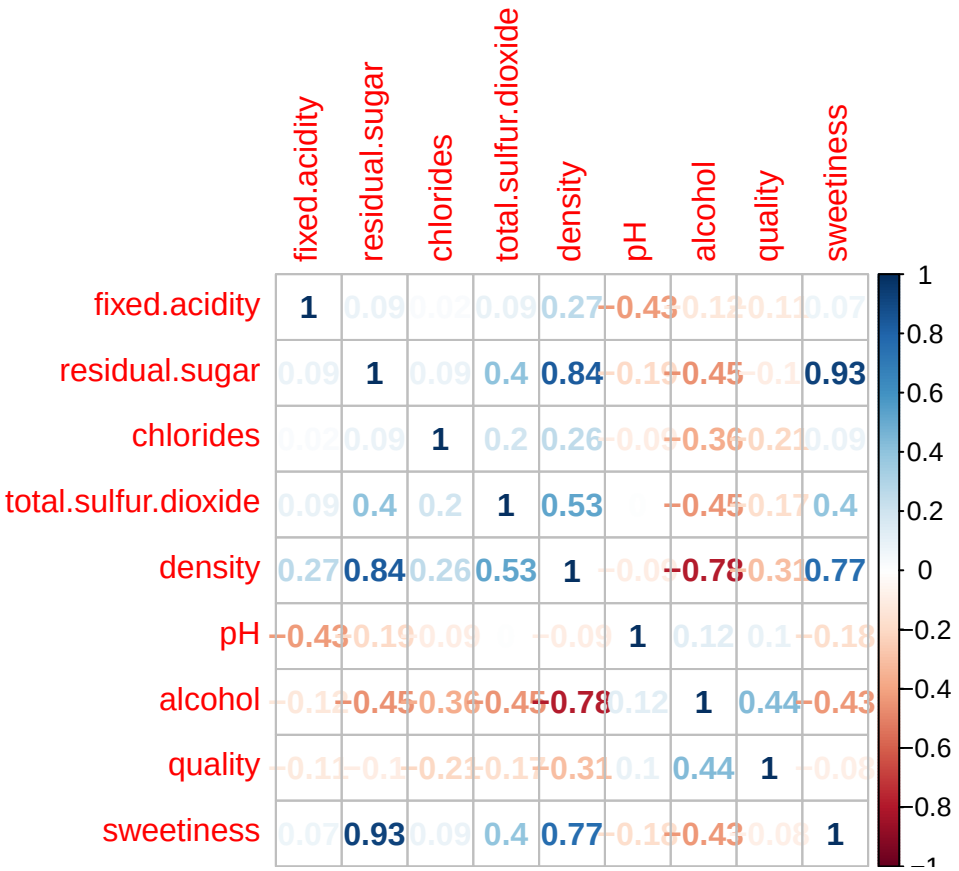
The residual sugar variable shows that most wines is near 1.5(g/dmˆ3). That is the point we can discuss and make some questions on it.

The alcohol variable isn't a normal distribution but its' mean and median are so close because alcohol at 9.5% are much more than other percentage.

Bivariate Plots Section

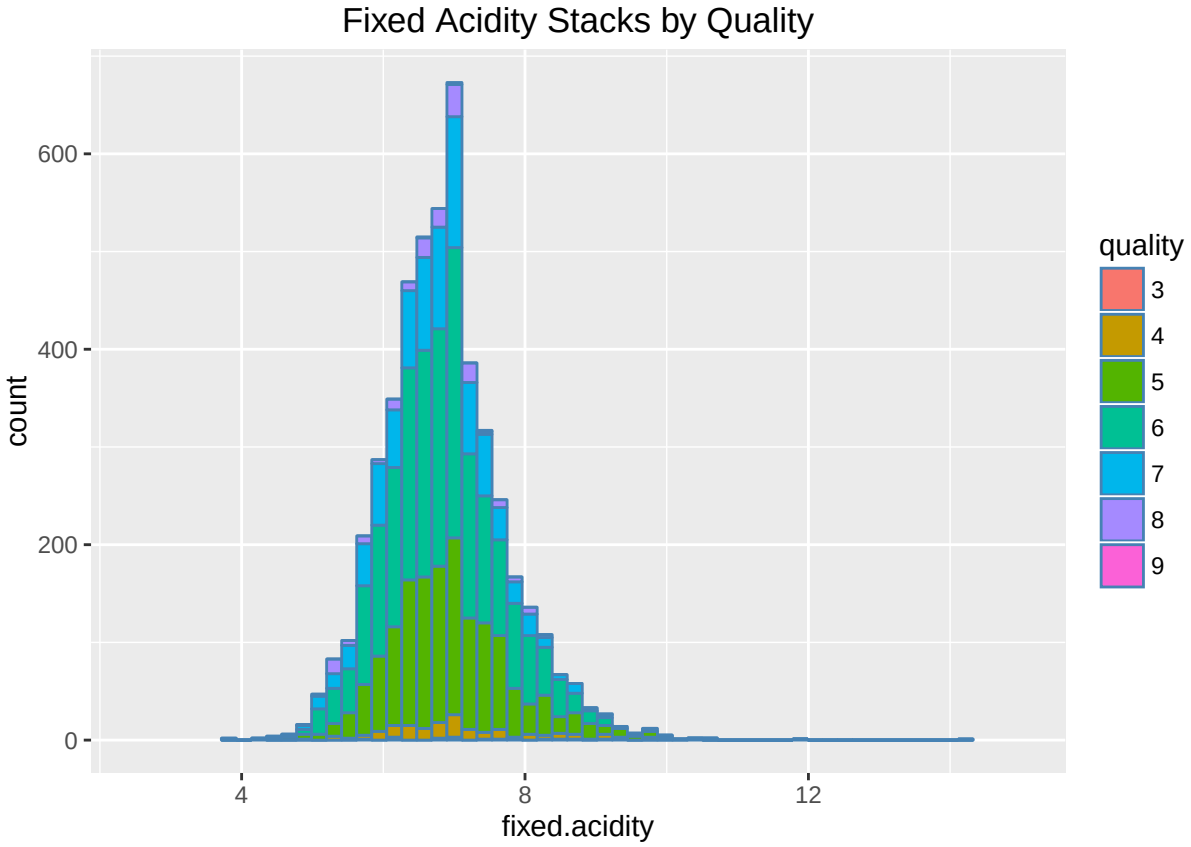
```
##                      fixed.acidity residual.sugar chlorides
## fixed.acidity                1.000          0.089     0.023
## residual.sugar               0.089          1.000     0.089
## chlorides                    0.023          0.089     1.000
```

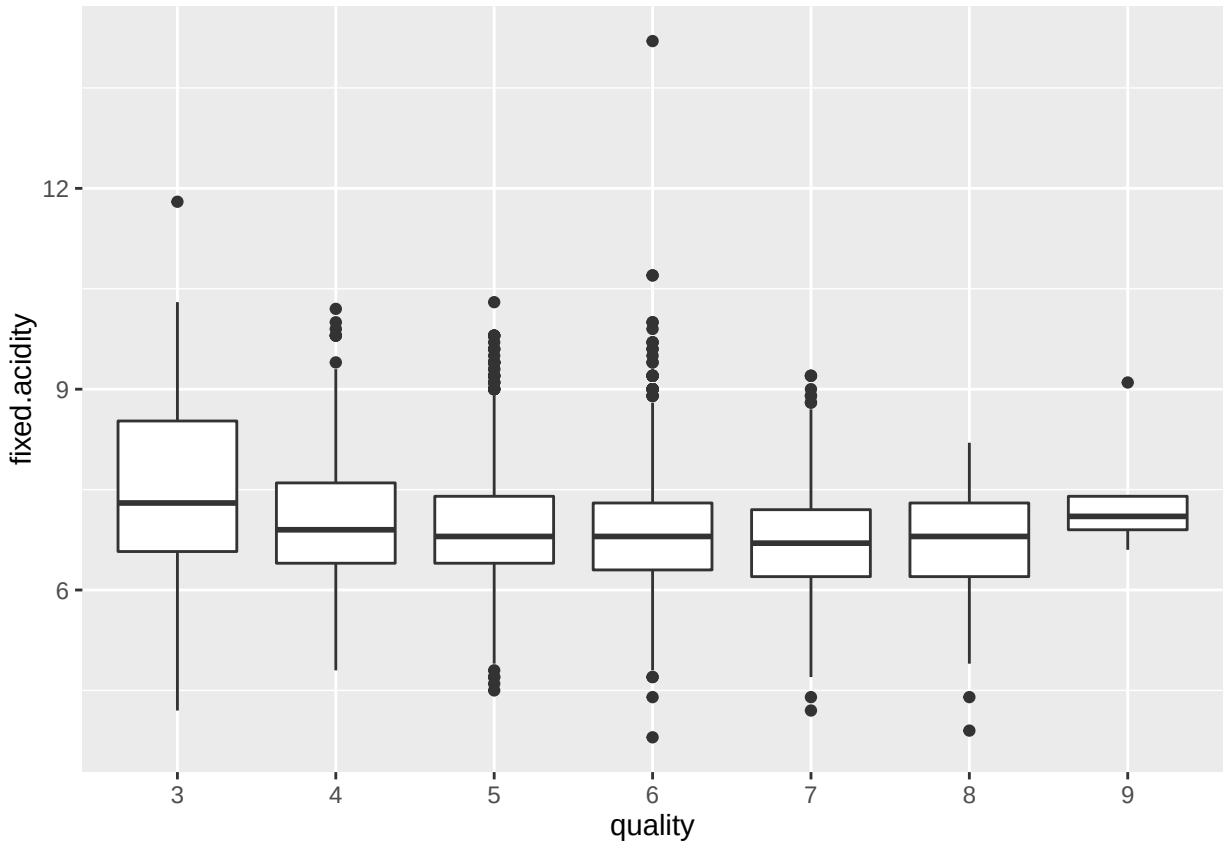
```
## total.sulfur.dioxide         0.091          0.401     0.199
## density                      0.265          0.839     0.257
## pH                          -0.426         -0.194    -0.090
## alcohol                     -0.121         -0.451    -0.360
## quality                     -0.114         -0.098    -0.210
## sweetiness                   0.071          0.926     0.090
##                      total.sulfur.dioxide density     pH alcohol quality
## fixed.acidity                       0.091   0.265 -0.426  -0.121  -0.114
## residual.sugar                      0.401   0.839 -0.194  -0.451  -0.098
## chlorides                           0.199   0.257 -0.090  -0.360  -0.210
## total.sulfur.dioxide                1.000   0.530  0.002  -0.449  -0.175
## density                             0.530   1.000 -0.094  -0.780  -0.307
## pH                                  0.002  -0.094  1.000   0.121   0.099
## alcohol                            -0.449  -0.780  0.121   1.000   0.436
## quality                            -0.175  -0.307  0.099   0.436   1.000
## sweetiness                          0.395   0.768 -0.177  -0.431  -0.080
##                      sweetiness
## fixed.acidity             0.071
## residual.sugar            0.926
## chlorides                 0.090
## total.sulfur.dioxide      0.395
## density                   0.768
## pH                       -0.177
## alcohol                  -0.431
## quality                  -0.080
## sweetiness                1.000
```



The reason why I create this plot is to know the correlation of coefficient between every variables.

I create a new dataFrame named wq_cor and drew this plot which tells us every correlation coefficient between variables.

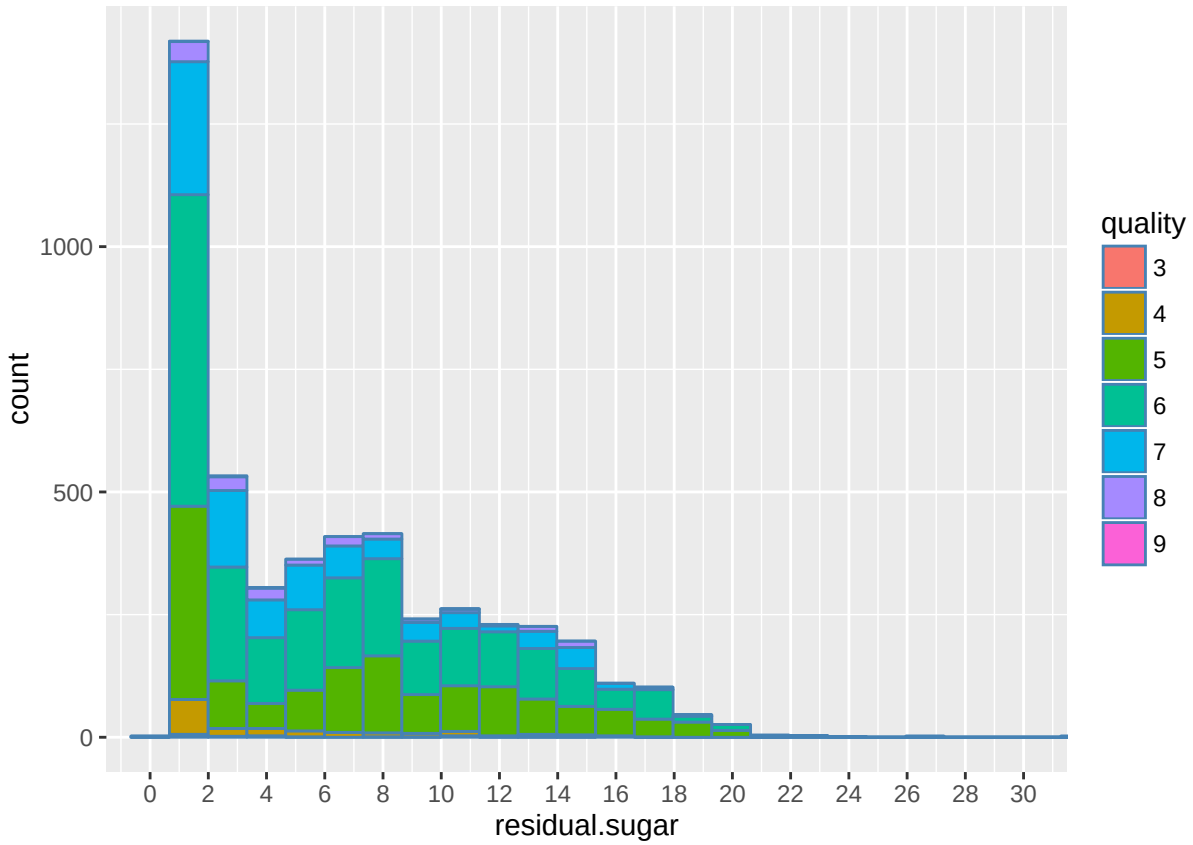





```
##    3    4    5    6    7    8    9
##   20  163 1457 2198  880  175    5
```

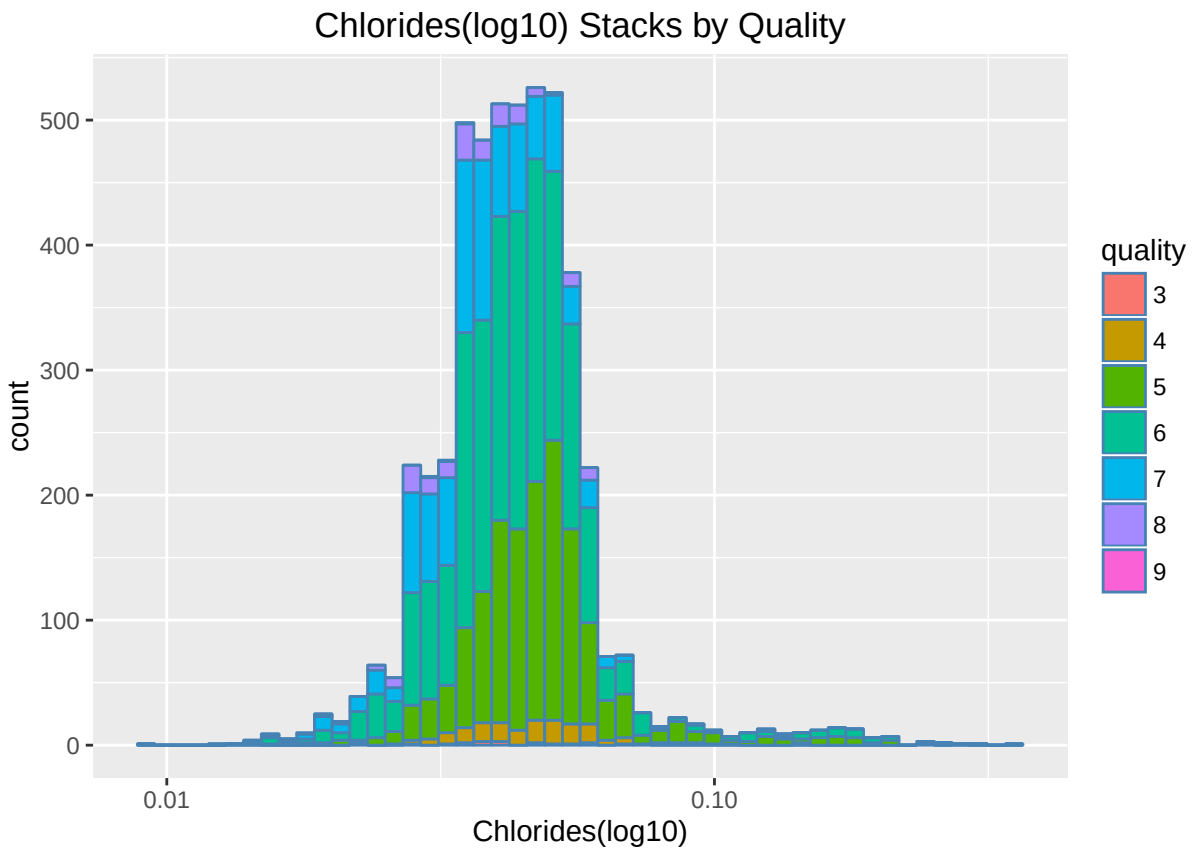
The reason why I create these plot are to know how quality distribute in fixed acidity and every qualitys' fixed acidity status.

These two plots tell that most quality are distributed averagely, but quality 3 has two times range than others at fixed acidity and quality 9 has a half range than others.



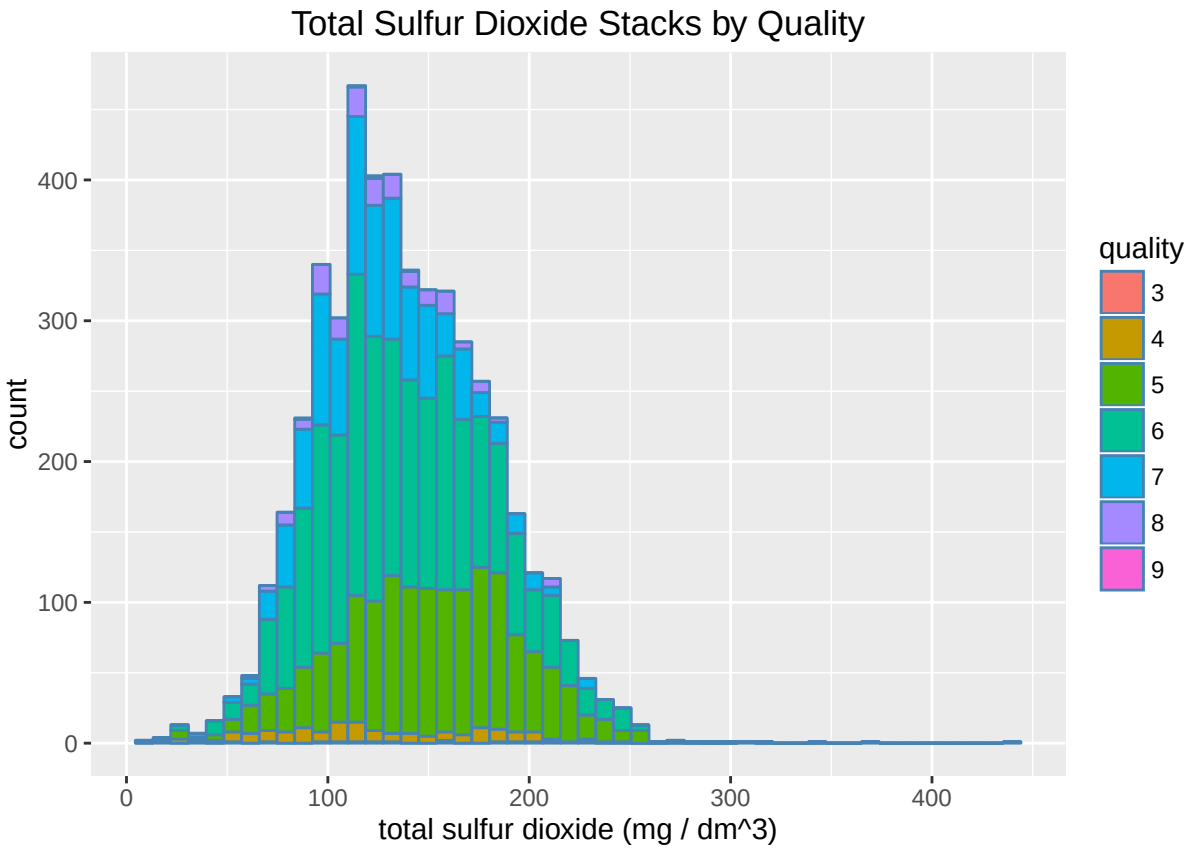
The reason why I create this plot is to know how quality distribute in residual sugar.

The plot shows that every quality distribute normally in residual sugar.



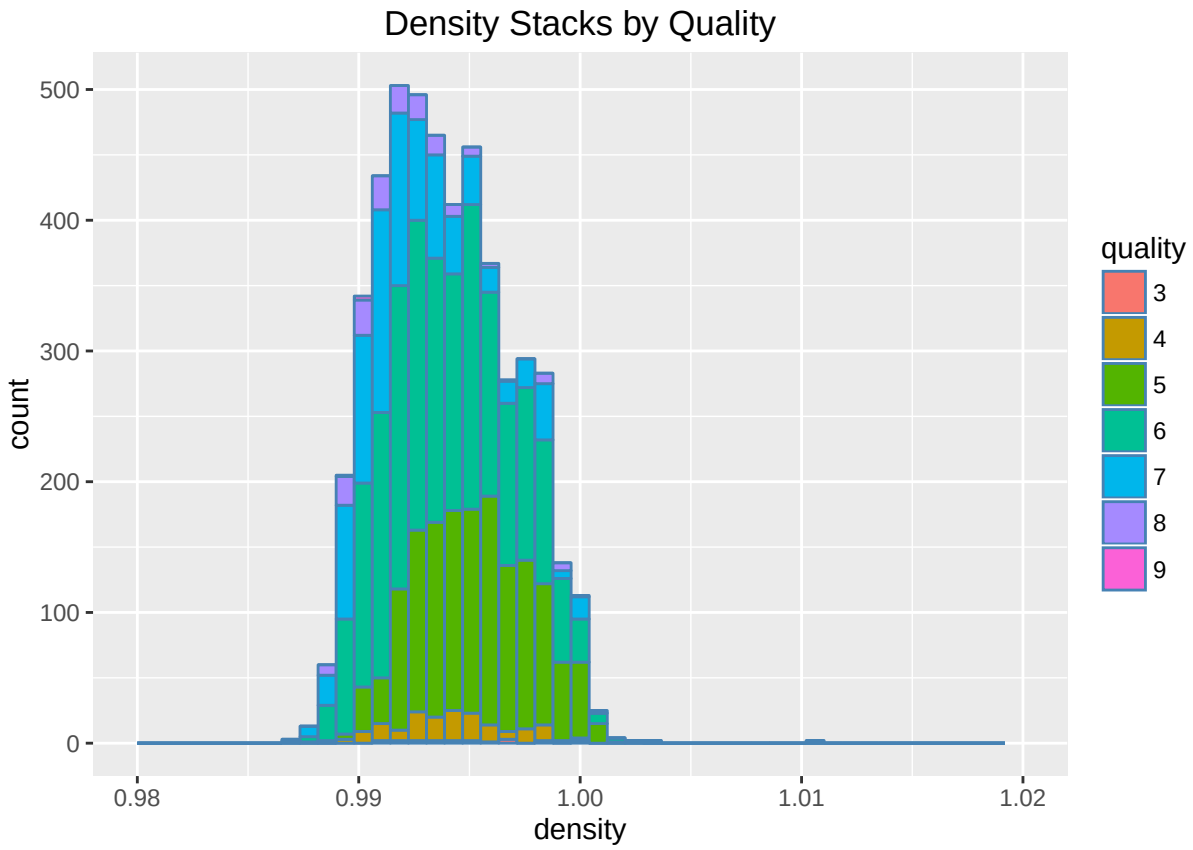
The reason why I create this plot is to know how quality distribute in chlorides.

The plot shows that every quality distribute normally in chlorides(log10)



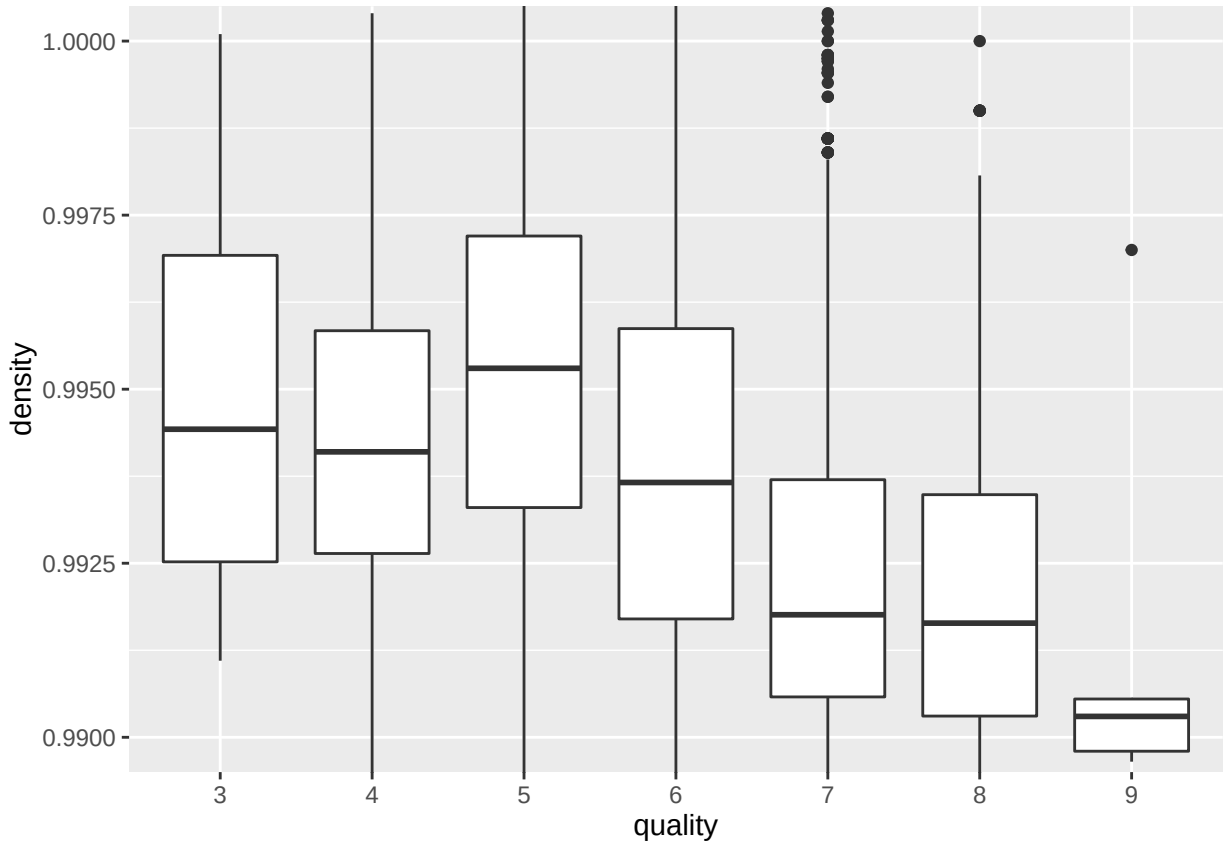
The reason why I create this plot is to know how quality distribute in total sulfur dioxide.

The plot also tells that every quality distribute normally in total sulfur dioxide.



The reason why I create this plot is to know how quality distribute in density.

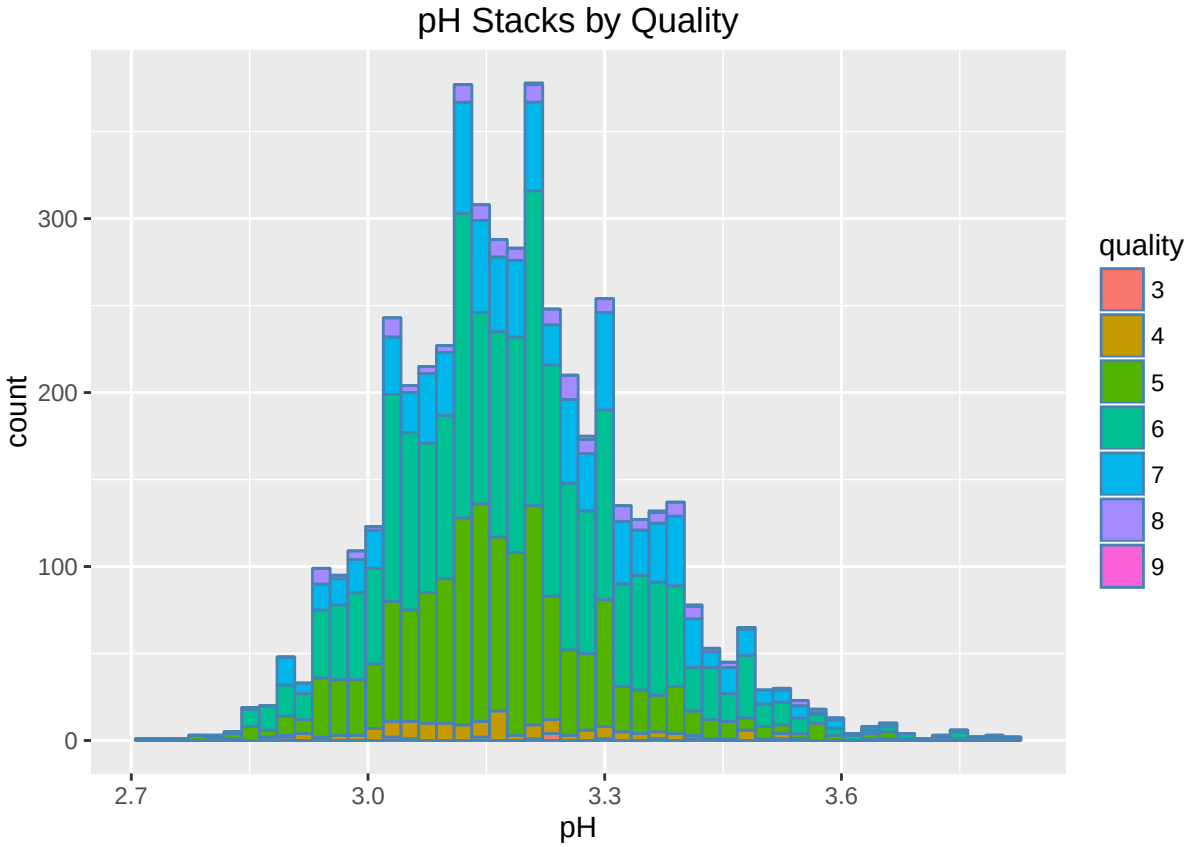
The plot display that every quality distribute averagely in density




```
## wq$quality: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9911  0.9925  0.9944  0.9949  0.9969  1.0000
## ------------------------------------------------------------
## wq$quality: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9892  0.9926  0.9941  0.9943  0.9958  1.0000
## ------------------------------------------------------------
## wq$quality: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9872  0.9933  0.9953  0.9953  0.9972  1.0020
## ------------------------------------------------------------
## wq$quality: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9876  0.9917  0.9937  0.9940  0.9959  1.0390
## ------------------------------------------------------------
## wq$quality: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9871  0.9906  0.9918  0.9925  0.9937  1.0000
## ------------------------------------------------------------
## wq$quality: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9871  0.9903  0.9916  0.9922  0.9935  1.0010
## ------------------------------------------------------------
## wq$quality: 9
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9896  0.9898  0.9903  0.9915  0.9906  0.9970
```

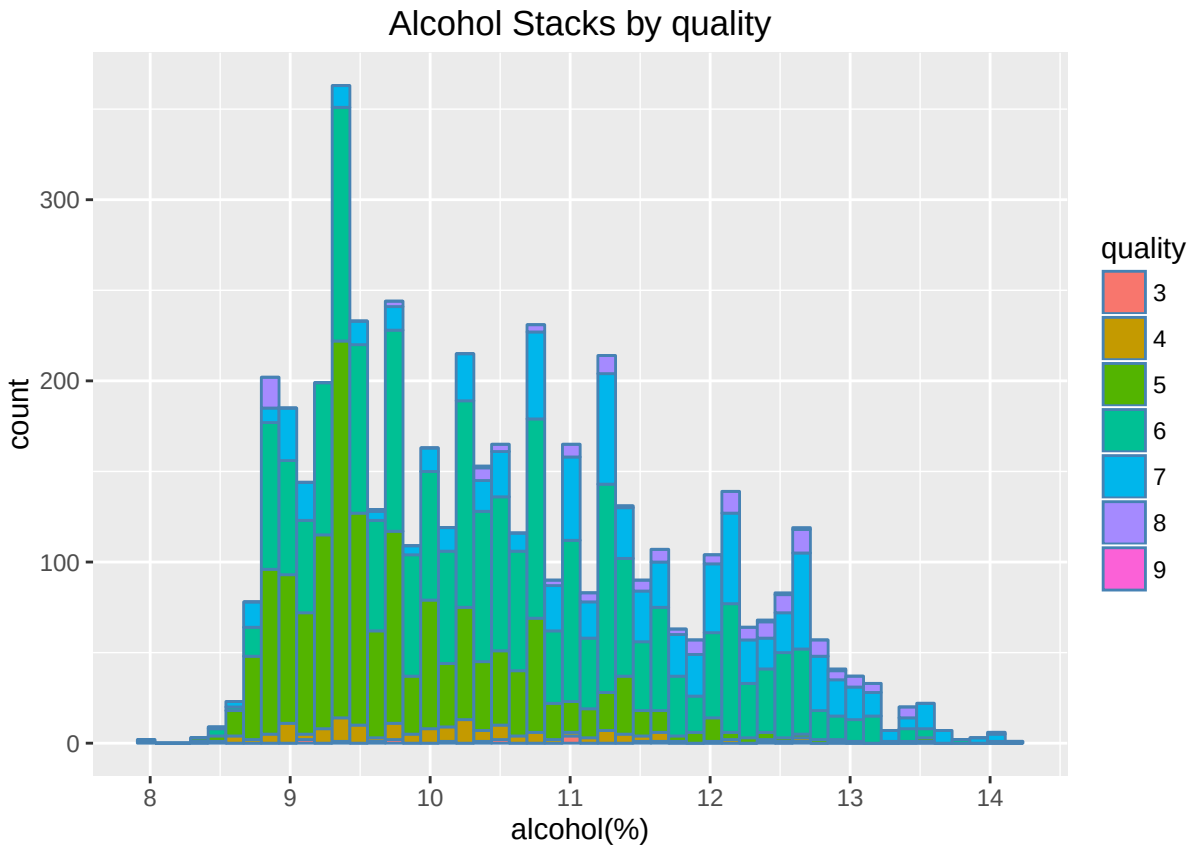
The reason why I create this plot is to know what composition of quality in density.

It is interesting that 9 is the highest quality and it's density is the lowest.



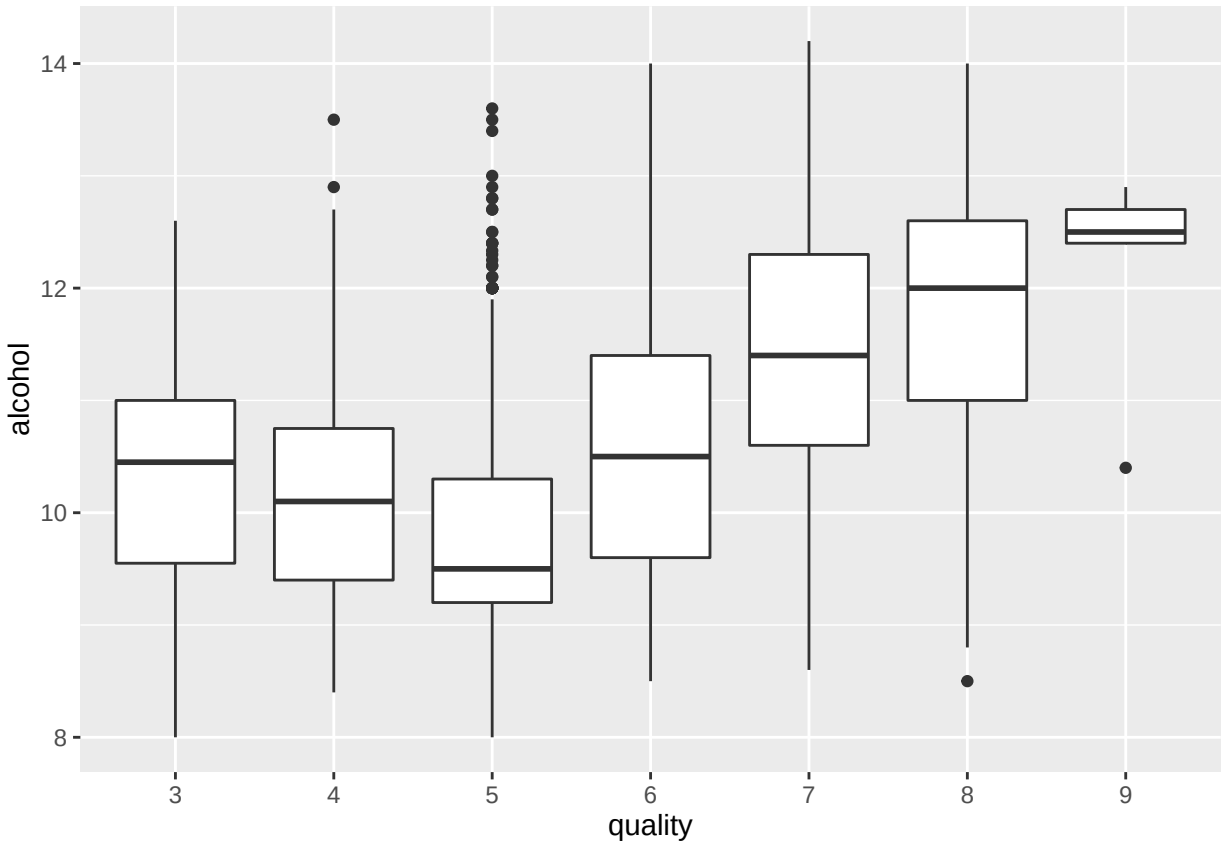
The reason why I create this plot is to know how quality distribute in pH.

It shows almost like normal distribution.



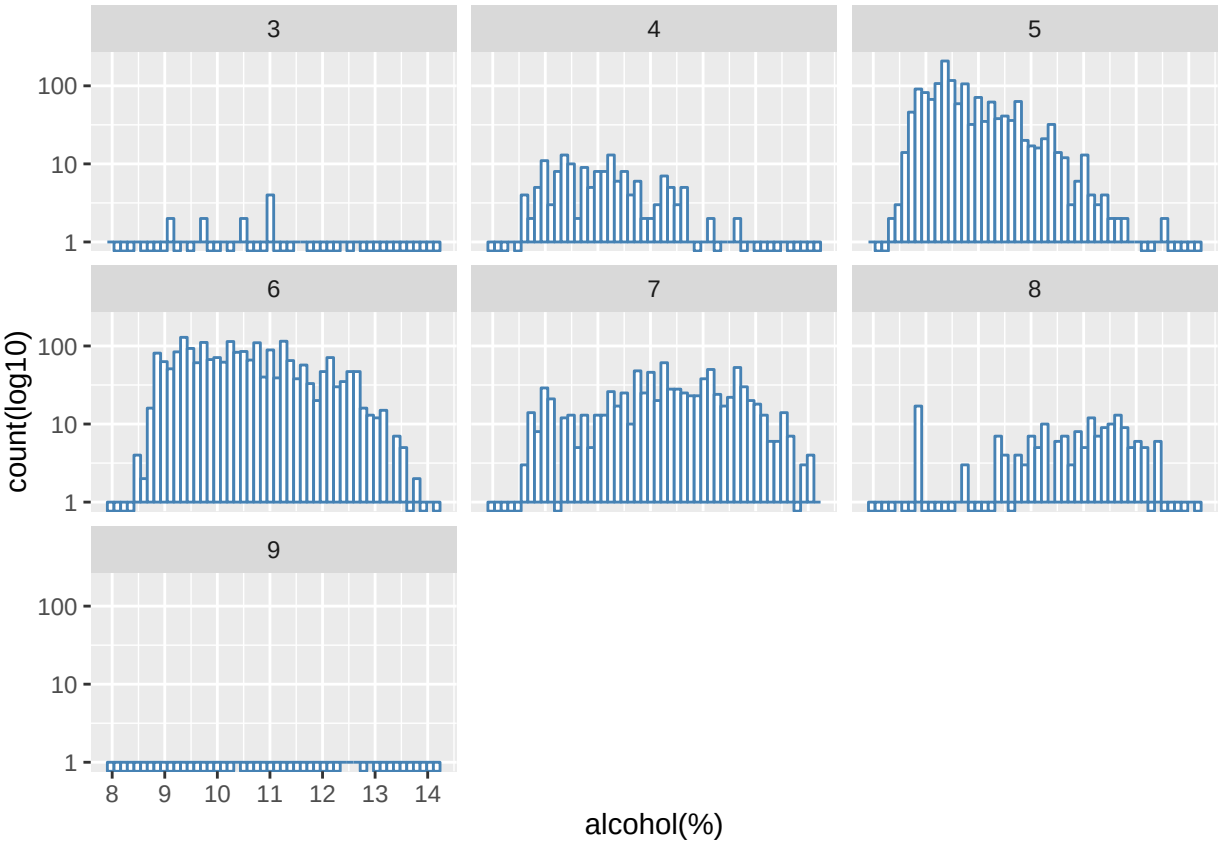
The reason why I create this plot is to know how quality distribute in alcohol.

The plot is the most different that quality quality distribute normally in the range 8-12 but white wines in 12-14 are above quality 5.

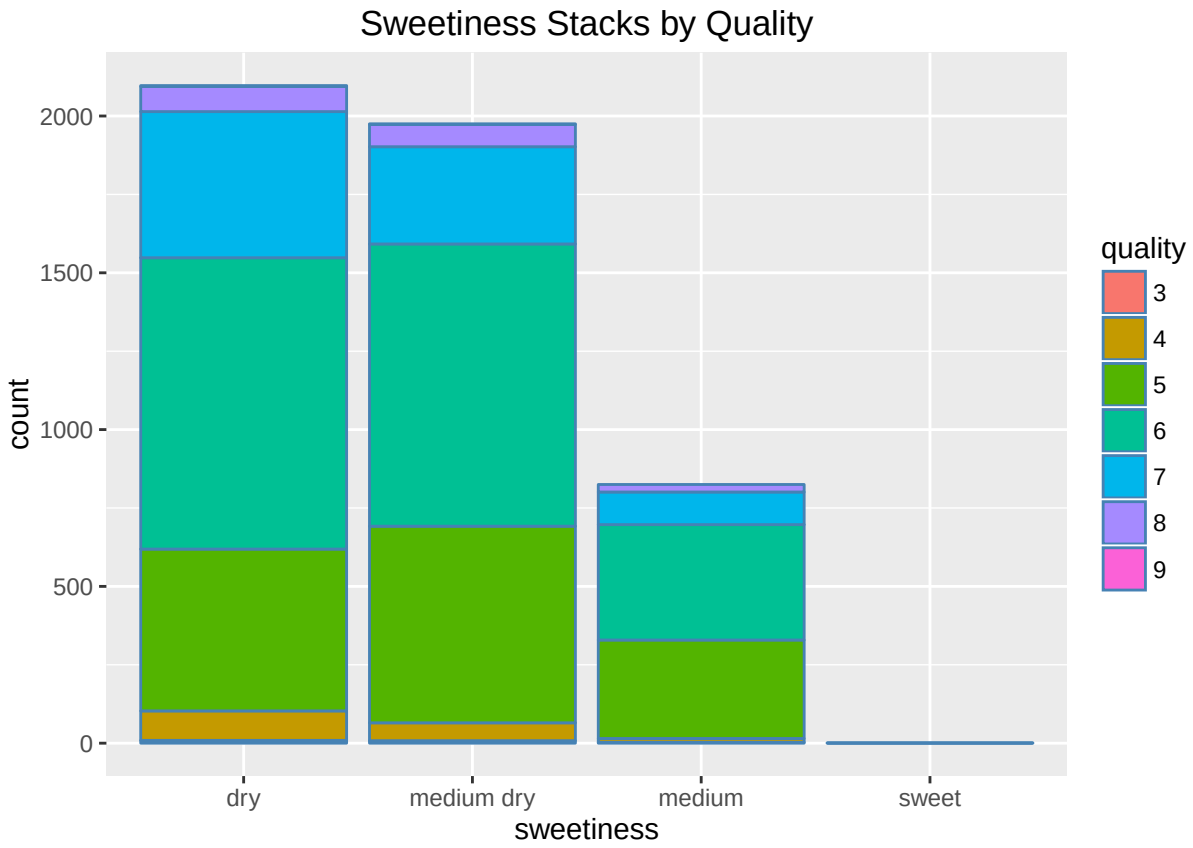



```
## wq$quality: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00    9.55   10.45   10.34   11.00   12.60
## ------------------------------------------------------------
## wq$quality: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.40   10.10   10.15   10.75   13.50
## ------------------------------------------------------------
## wq$quality: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   8.000   9.200   9.500   9.809  10.300  13.600
## ------------------------------------------------------------
## wq$quality: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.50    9.60   10.50   10.58   11.40   14.00
## ------------------------------------------------------------
## wq$quality: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.60   10.60   11.40   11.37   12.30   14.20
## ------------------------------------------------------------
## wq$quality: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.50   11.00   12.00   11.64   12.60   14.00
## ------------------------------------------------------------
## wq$quality: 9
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.40   12.40   12.50   12.18   12.70   12.90
```

With this plot and statistic, quality 9 is different from others. We might speculate that high percentage of alcohol could be high quality.



The histogram of alcohol and quality shows that most white wines' quality separate averagely below 12%, but quality 5-9 also show at 12%-14%. It tells that high alcohol percentage might be easier to be high quality.




```
## wq$quality: 3
##        dry medium dry     medium      sweet
##          9          8          3          0
## ------------------------------------------------------------
## wq$quality: 4
##        dry medium dry     medium      sweet
##         94         57         12          0
## ------------------------------------------------------------
## wq$quality: 5
##        dry medium dry     medium      sweet
##        516        627        314          0
## ------------------------------------------------------------
## wq$quality: 6
##        dry medium dry     medium      sweet
##        929        900        368          1
## ------------------------------------------------------------
## wq$quality: 7
##        dry medium dry     medium      sweet
##        466        310        104          0
## ------------------------------------------------------------
## wq$quality: 8
##        dry medium dry     medium      sweet
##         80         71         24          0
## ------------------------------------------------------------
## wq$quality: 9
##        dry medium dry     medium      sweet
##          3          2          0          0
```

This diagram tells that every quality of sweetiness takes up almost the same percentage but the quality 7 is a little different. The quality 7 takes up more percentage at dry above other sweetiness.

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?

1. Density has strongly positive relationship with residual sugar and sweetiness but it has strongly negative relationship with alcohol.

2. The reason why residual sugar correlates strongly with sweetiness because it was used to create sweetiness.

3. Quality is no obvious relationship with other variables. Alcohol is the one which has the highest correlation of coefficient with quality and it's 0.44

4. There doesn't have a specific variable that really affect quality because quality distributes normally at every variables.

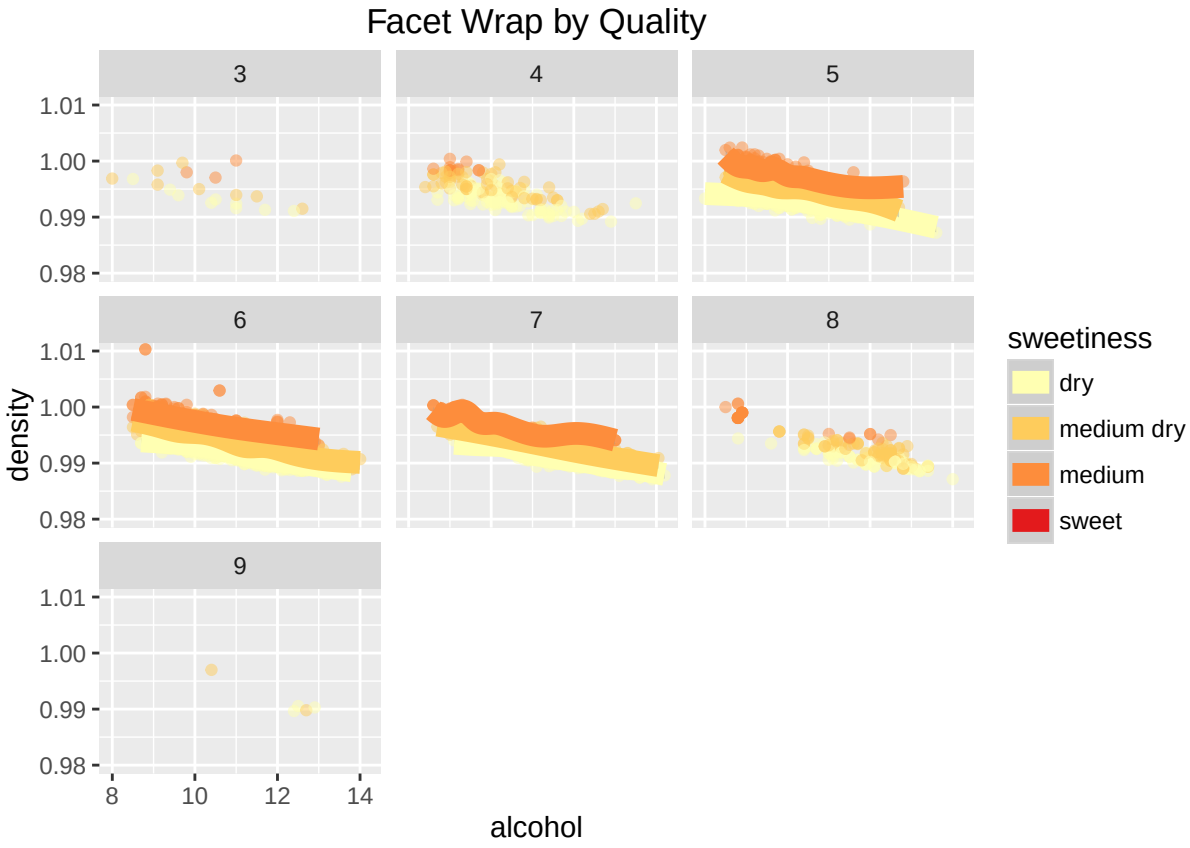
Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

I didn't find any interesting relationships. The relationships between every variables are normal. Density has high relationships with residual sugar and total sulfur dioxide and low relationship with alcohol.

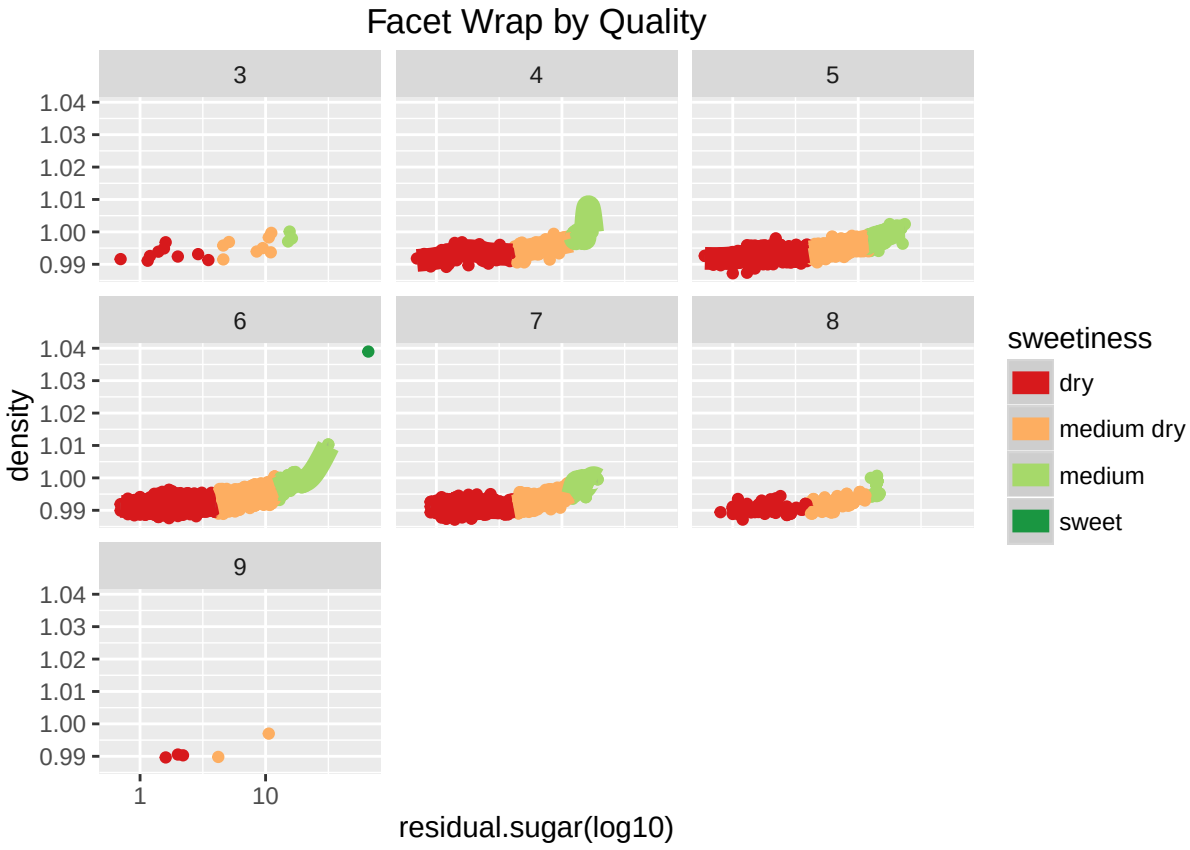
What was the strongest relationship you found?

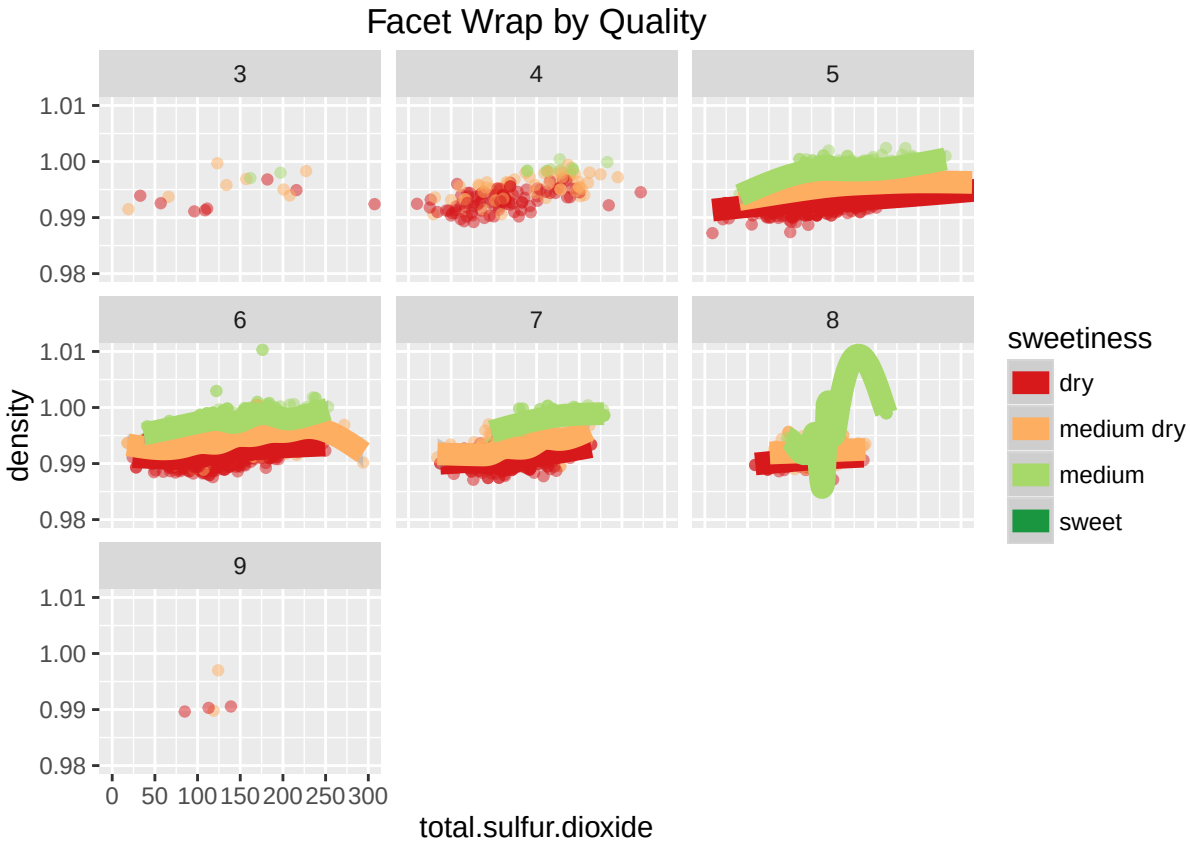
The strongest relationship I found is between residual sugar and sweetiness cause I used it to create sweetiness. The second strong relationship is density and residual sugar. That makes sense because residual sugar correlates strongly with density and so does sweetiness.

Multivariate Plots Section



This picture shows that the density is lower and alcohol is higher. The wine is sweeter and the density will be higher but there is no relationship with quality.


Facet Wrap by Quality
3
4
5
6
7
8
9
density
1.04
1.03
1.02
1.01
1.00
0.99
1
10
residual.sugar(log10)
sweetiness
dry
medium dry
medium
sweet


Facet Wrap by Quality
3
4
5
6
7
8
9
density
1.01
1.00
0.99
0.98
0
50
100
150
200
250
300
total.sulfur.dioxide
sweetiness
dry
medium dry
medium
sweet

Sugar and sulfur dioxide does not affect the quality of white wine but they have positive relationship with density

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?

There are not strong relationship with these variables.

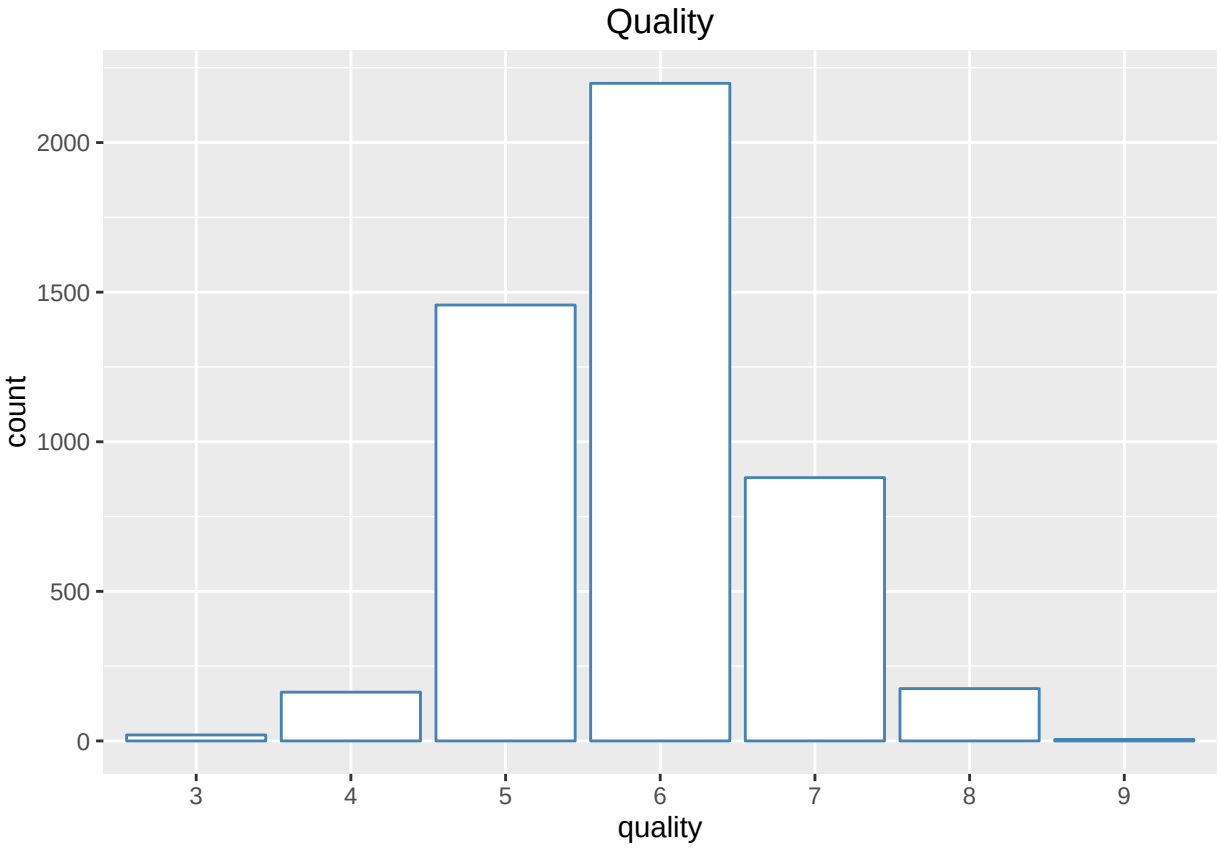
Were there any interesting or surprising interactions between features?

That are not any interesting or surprising interactions between feafures.

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

Final Plots and Summary

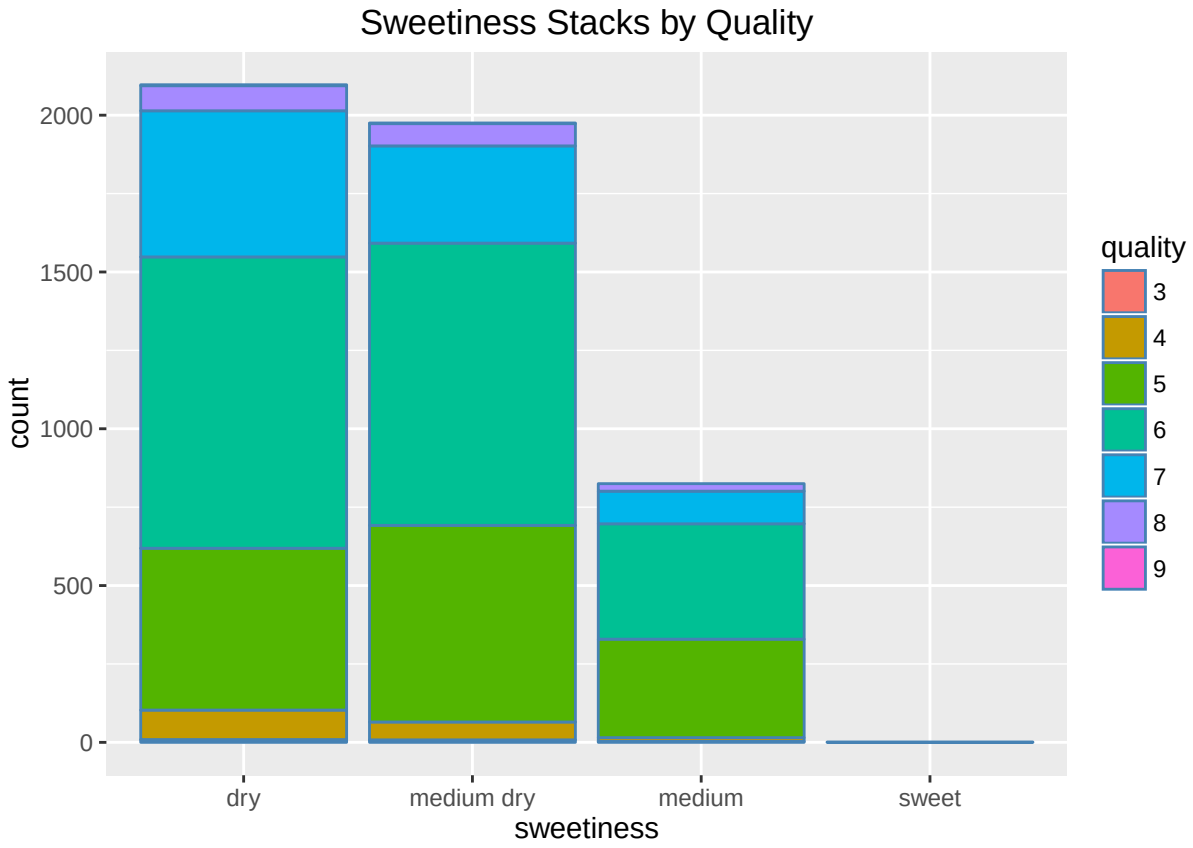
Plot One



Description One

This plot is interesting because quality 5,6,7 take up 92.5% of this sample and there does not have quality 1,2 and 10.

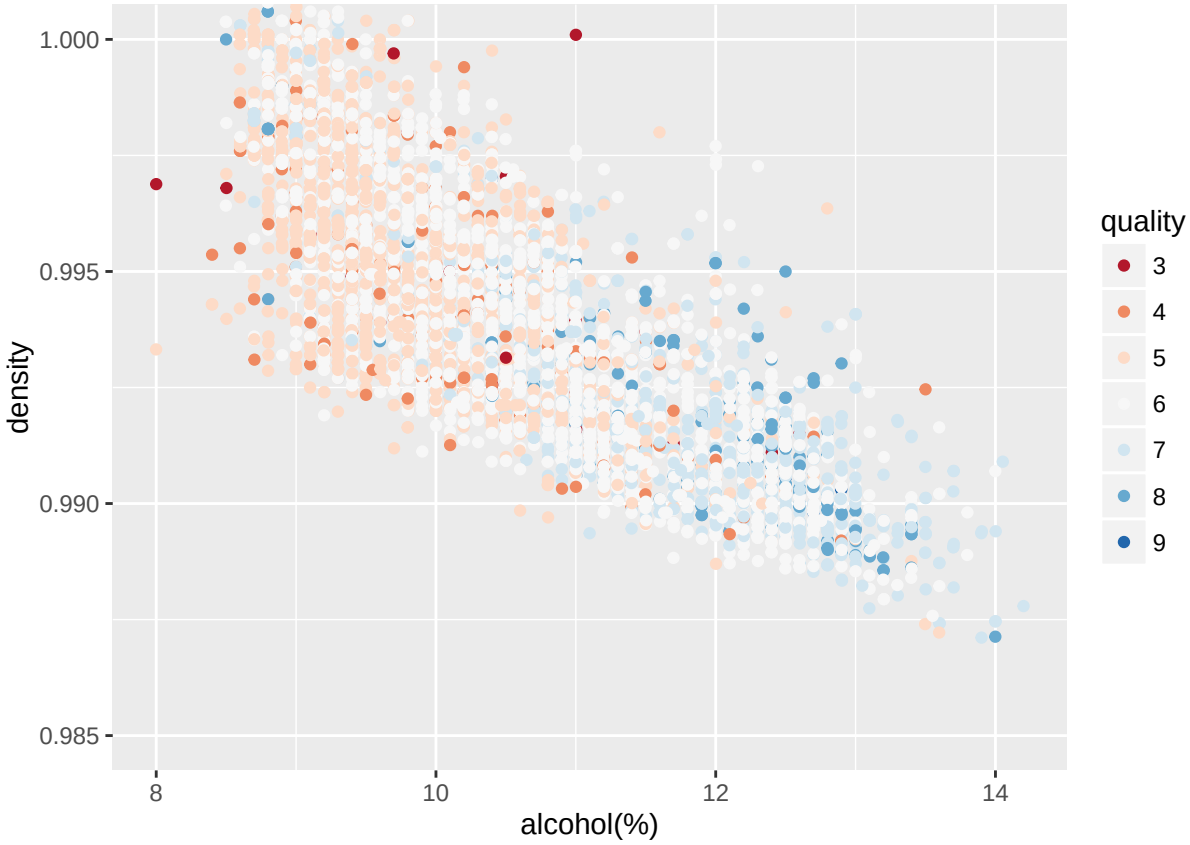
Plot Two



Description Two

This plot shows that dry and medium dry take up 83% of all white wines and quality is alloted averagely at dry, medium dry and medium.

Plot Three



Description Three

The picture shows that alcohol and density have strongly negitive correlation but having nothing to do with quality. It indicate that we can't use alcohol and density to predict quality but we could do a linear regression model with density and alcohol. It can help us to put one of them into the model and predict another one.

Reflection

The project help me to review the skill in lesson 4 and I create lots of plots, histogram, barplot, boxplot, scatter plot and so on. It also let me know how to deal with a dataset and find something interesting from it.

The primary goal of my research is to find which variable would affect the quality of white wine but I think that it is very difficult to do because the quality is defined by people. It is tough to understand how people graded white wines. It can not only use ingredients to predict quality cause people would grade by it's smell, taste, appearence and human's favor although the people who graded are experts.

But, we also can find some relationships between all ingredients such that what ingredients would affect another one. Creating linear model is also an important way we can use to. For example, we could use density to predict what percentage of alcohol is.