P4: Explore and Summarize Data by Julia Kudinovich

Intro to the dataset

This report explores a tidy dataset containing 4,898 white wines with 11 variables on quantifying the chemical properties of each wine. At least 3 wine experts rated the quality of each wine, providing a rating between 0 (very bad) and 10 (very excellent).

In the report I will be investigating which chemical properties influence the quality of white wines.

Univariate Analysis

Statistics on the dataset

First, let's run some statistics on the dataset.

```
## '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 ...
```

What is the structure of your dataset?

The dataset contains 4898 observation of 13 variables: 12 features (fixed.acidity, volatile.acidity, citric.acid, residual.sugar, chlorides, free.sulfur.dioxide, total.sulfur.dioxide, density, pH, sulphates, alcohol and quality) and 1 id column

All variables are numeric. Quality is categorical discrete variable (rating form 0 to 10), all the other variables are continuous.

Below is the summary for each variable in the dataset:

```
##        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
```

75% of wines have rating 6 or below with 3 being the lowest rating and 9 - highest.

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

Quality is an integer field. It can only have values between 0 and 10. For better representation I would like to create factor variable out of quality.

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

Futhermore, I want to group quality factor and create 3 rating groups out of it. Wines with rating 3-5 will have 'low' factor, 6 'average' and 7-9 will be 'high' qualiuty factor.

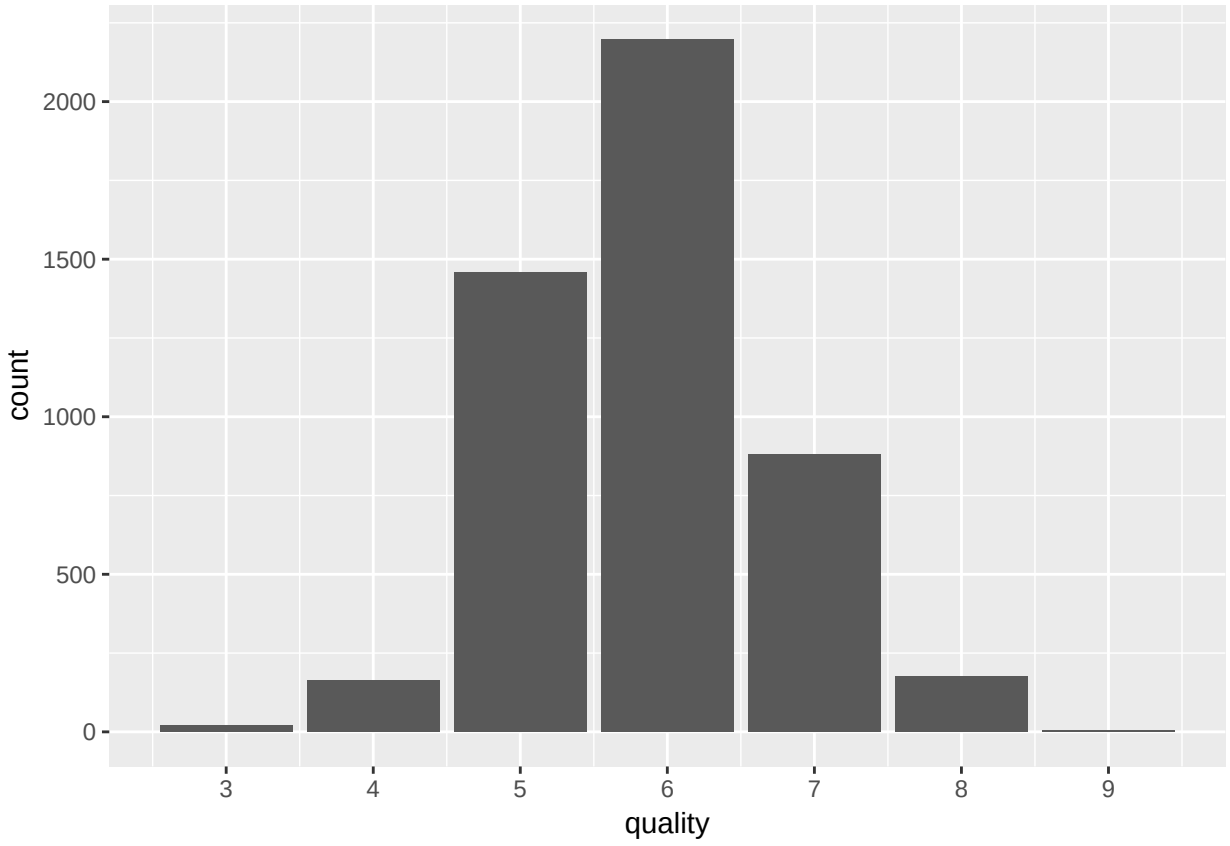
```
##     low average    high
##    1640    2198    1060
```

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

The main feature in the dataset is `quality` which determines how good or bad the wine is. I would like to investigate which chemical features influence the main feature

Quality

Let's graph quality:



From the graph above we can see that the quality of wines has a bell shaped normal distribution with most of the wines having 5 or 6 rating, which is consistent with 5.878 mean for wine quality we got from the summary. Very low number of wines have rating of 9 and none have very excellent (10) rating. Similarly, the lowest rating is 3 with very little number of wines having this rating.

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

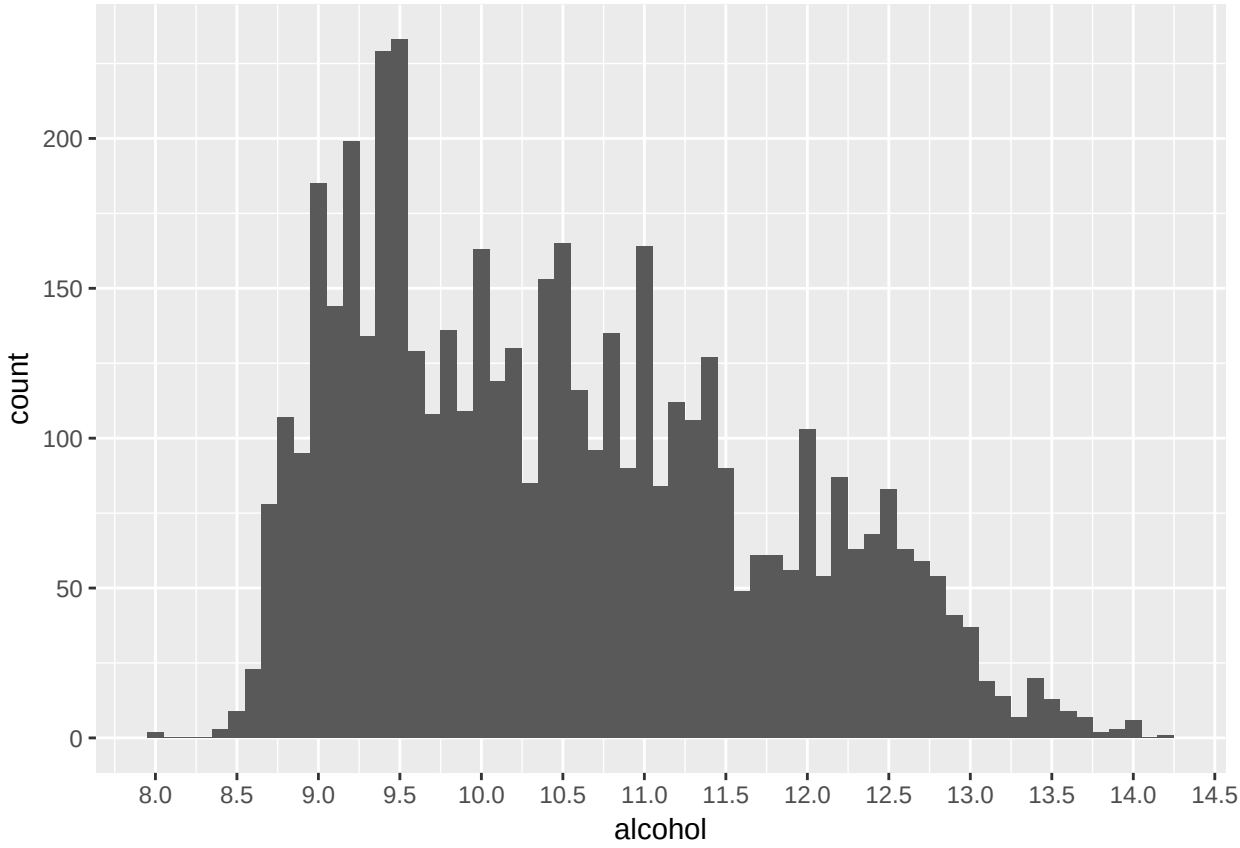
I will investigate the chemical properties that affect the taste or smell since the wine rating was given based on sensory data. Namely, features I am interested are: `citric acid` (adds freshness and flavor to wines), `residual sugar` (determines how sweet wine is), `alcohol` (can alter the taste), `pH` (describes how acidic wine is).

Alcohol

Summary:

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

Let's plot alcohol content:



From the above graph we can see that alcohol content distribution is positively skewed. More wines have lower alcohol.

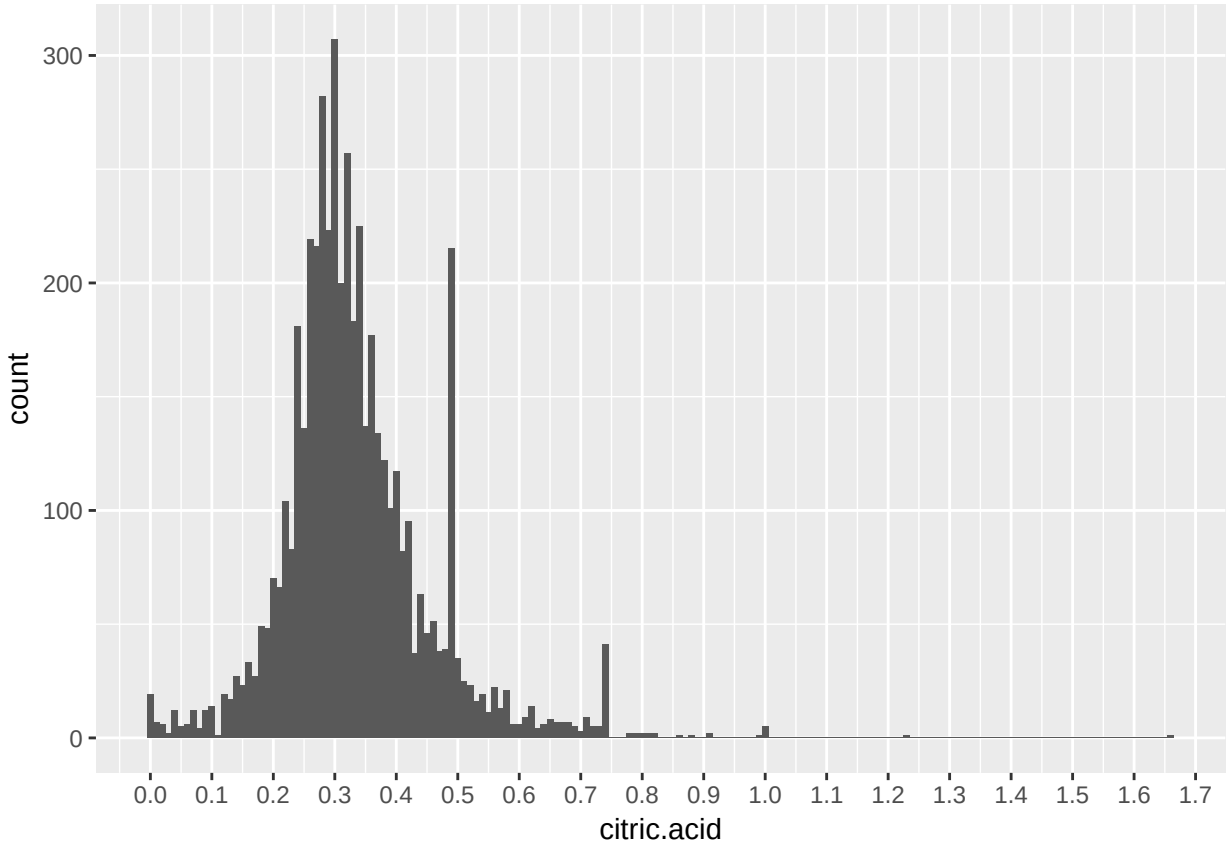
Median alcohol content is 10.4 with most of the wines having alcohol content between 8 and 11.40

Citric Acid

Summary:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.2700  0.3200  0.3342  0.3900  1.6600
```

Plot:



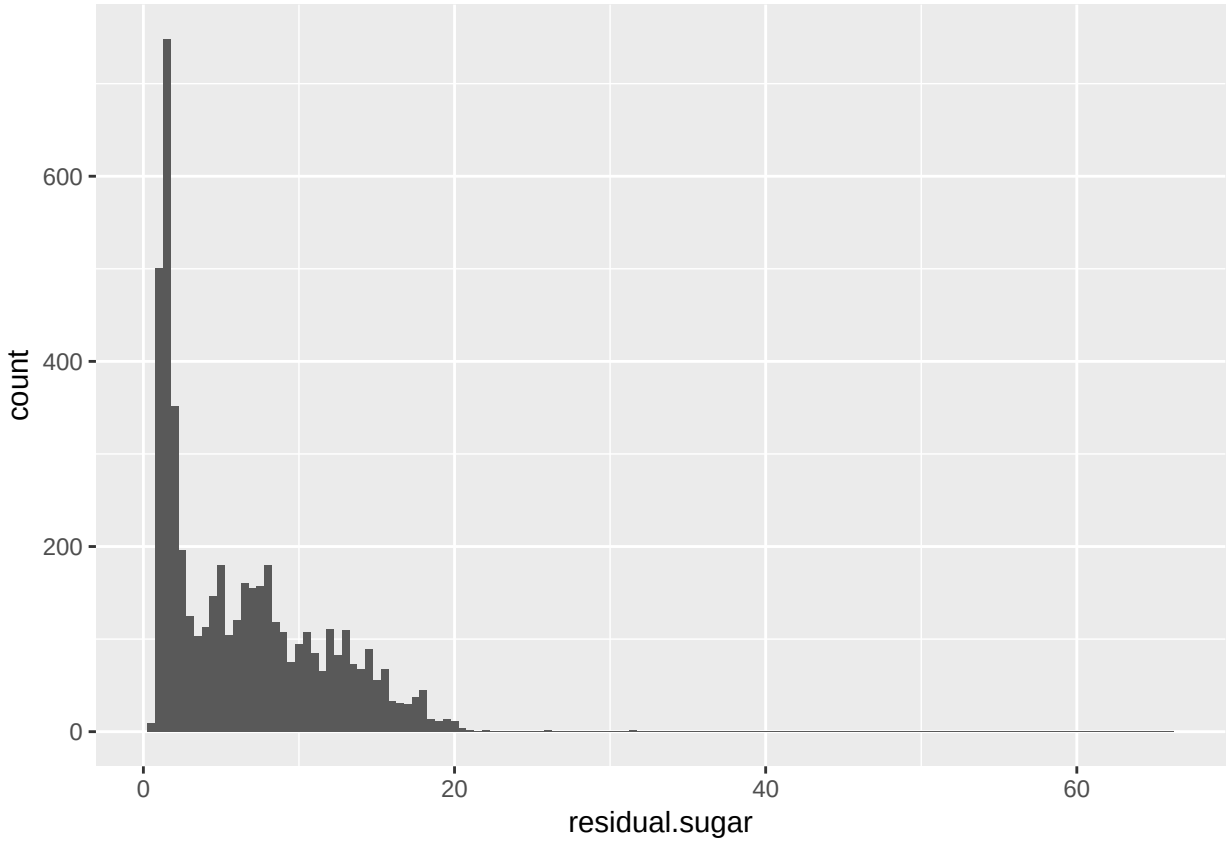
Most wines have citric acid between 0 and 0.39. 3rd quantile of 0.39 differs a lot from the maximum value of 1.66. We can see that there are outliers with values 0.7 and above. Also, there are a peaks around 0.5 and 0.75 values.

Residual sugar

Summary:

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

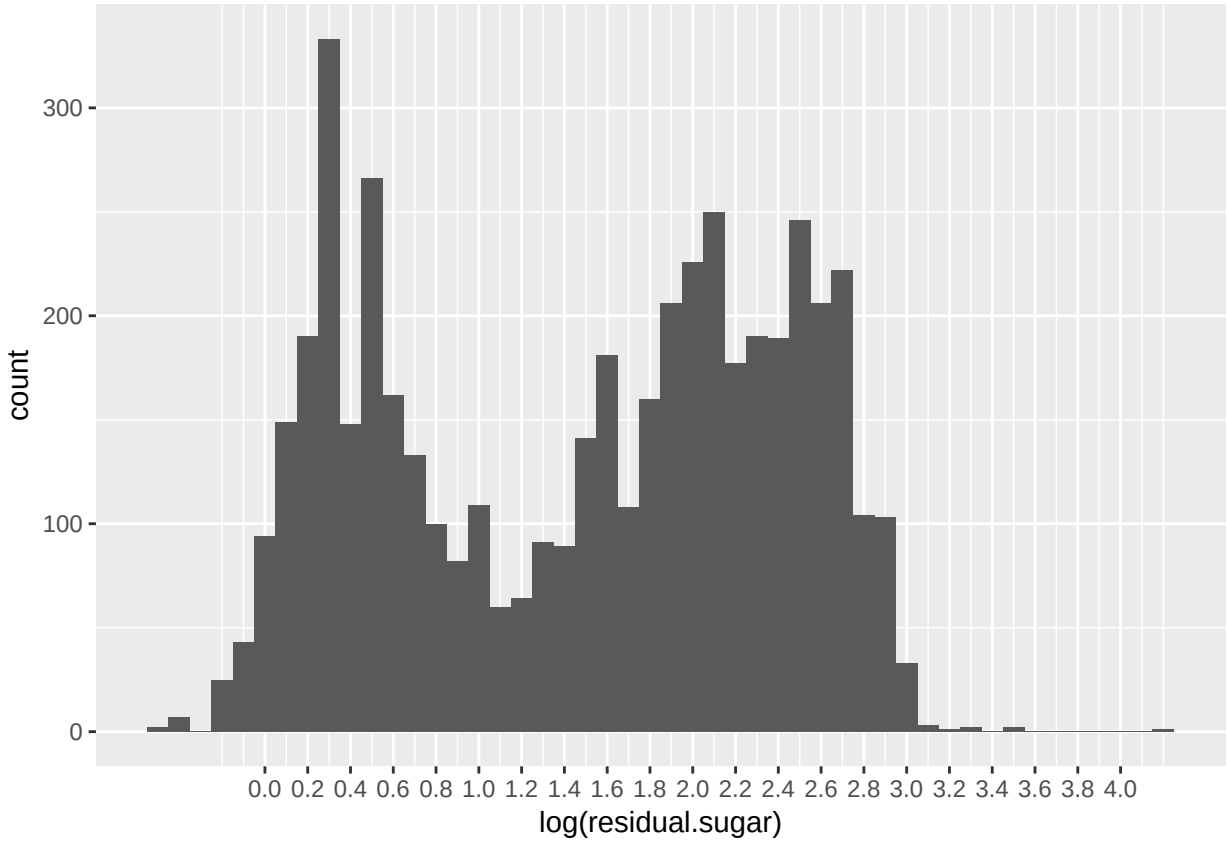
Plot:



Residual sugar is positively skewed and has a long right tail and several vary large valued outliers. 3rd quantile ha the value of 9.9 while the maximum value for the variable is 65.8

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 found that residual.sugar distribution is positively skewed and has a long right tail. I am going to transform long tailed distribution using log scale



After transformation we can see that distribution is bimodal with peaks around 0.2 and 2.

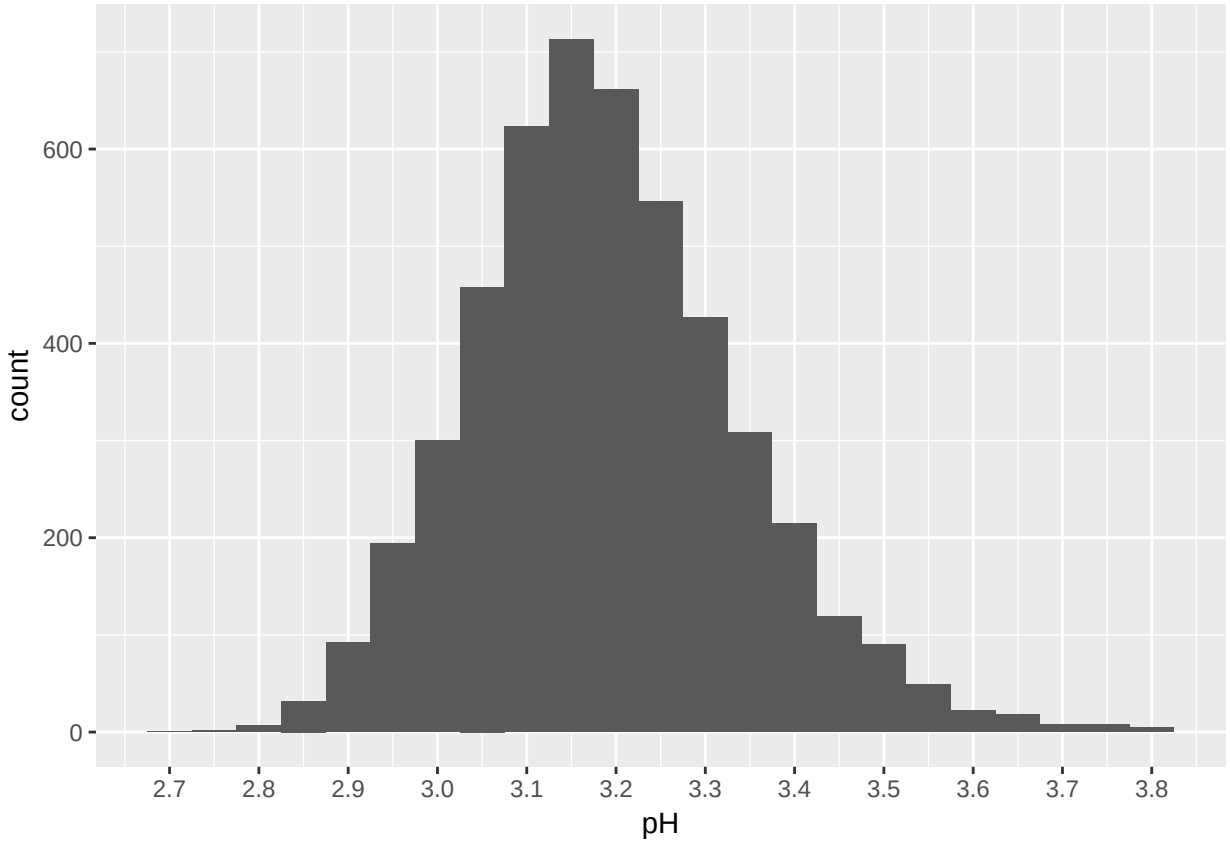
pH

Summary:

```
summary(wine$pH)
```

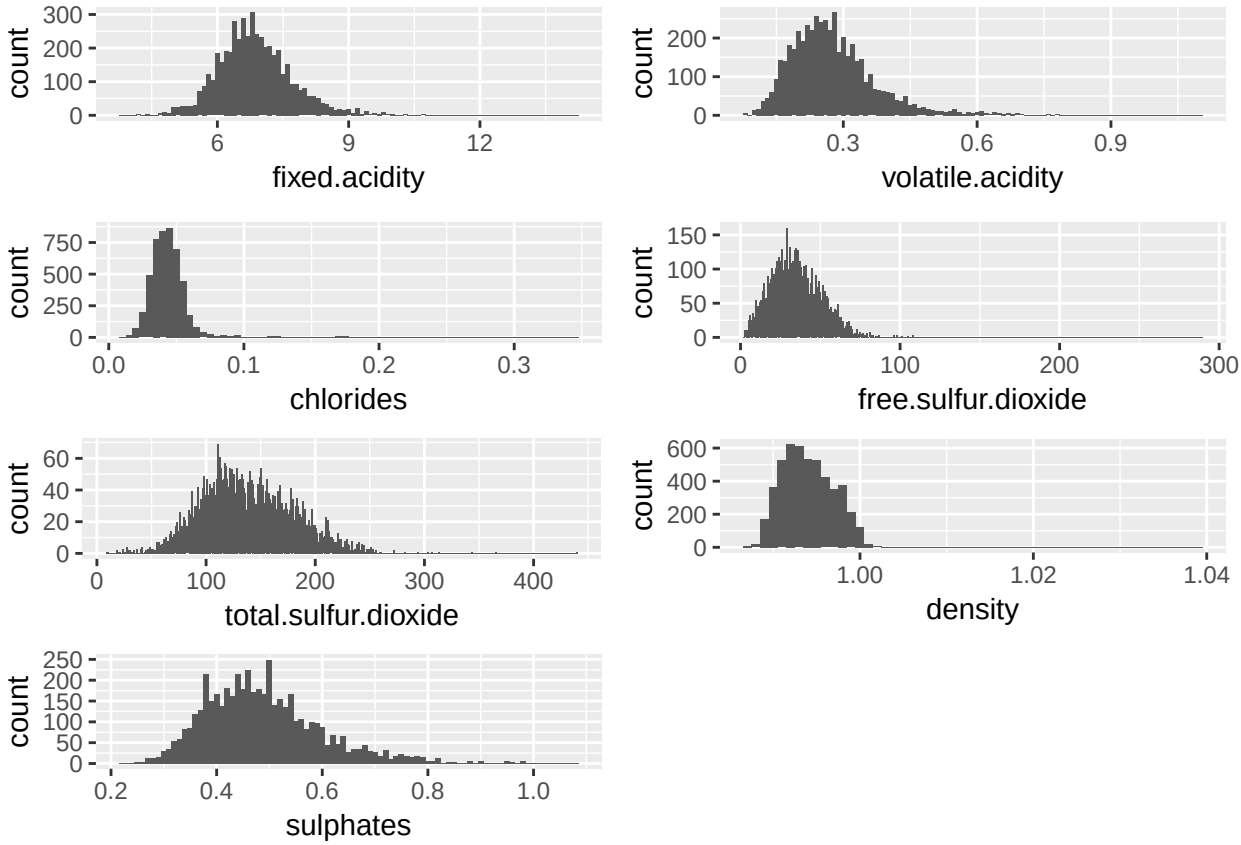
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.720   3.090   3.180   3.188   3.280   3.820
```

Plot:



pH has a bell shaped normal distribution with a peak around 3.1. Since pH is a logarithmic scale the wine with maximum value of 3.82 pH is ten times more acidic than wine with minimum value of 2.72.

Below are several more graphs exploring remaining variables in the dataset.

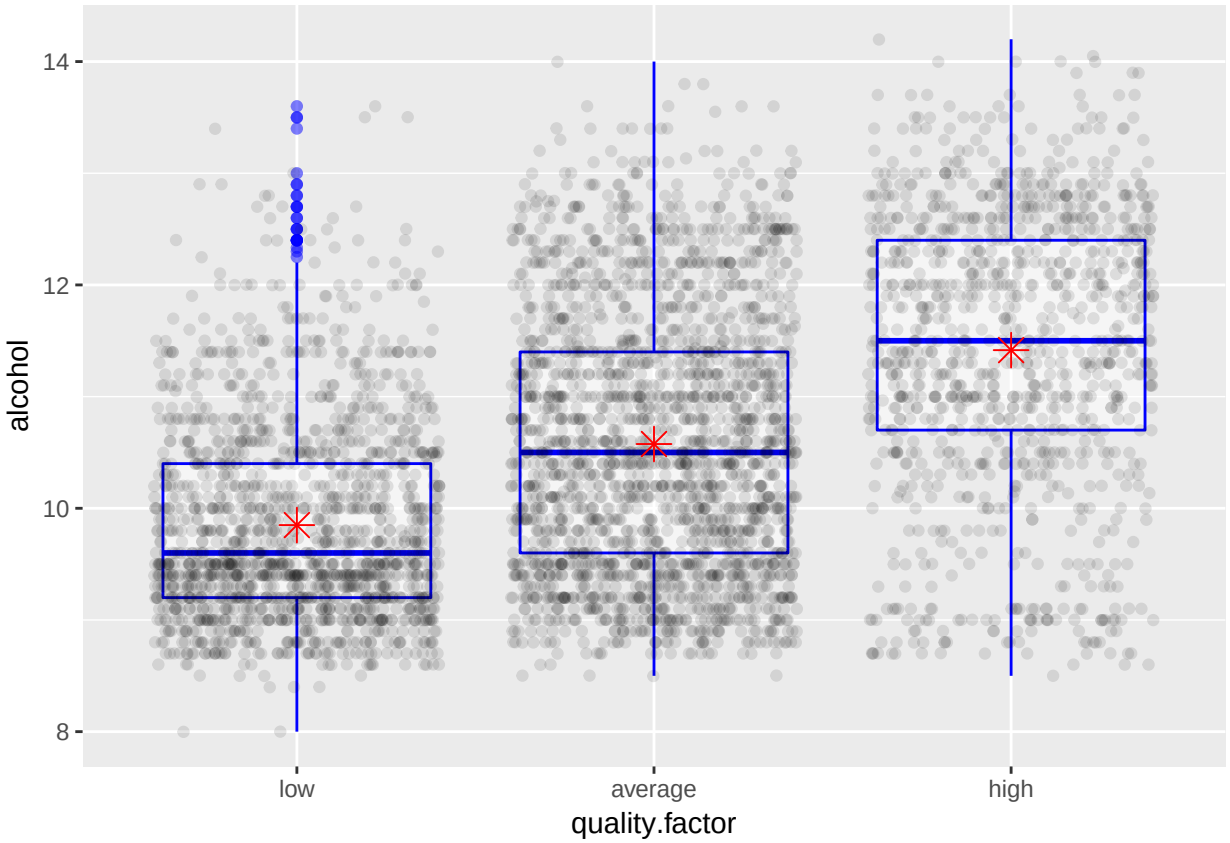


At the first glance, most of the above variables have a bell shaped distribution with some having right tails.

Bivariate Plots Section

Quality vs alcohol

First, I am interested in relationship between quality and alcohol. Let's plot them below:



From the plot we can see apparent correlation between alcohol and quality. Better wines have higher alcohol.

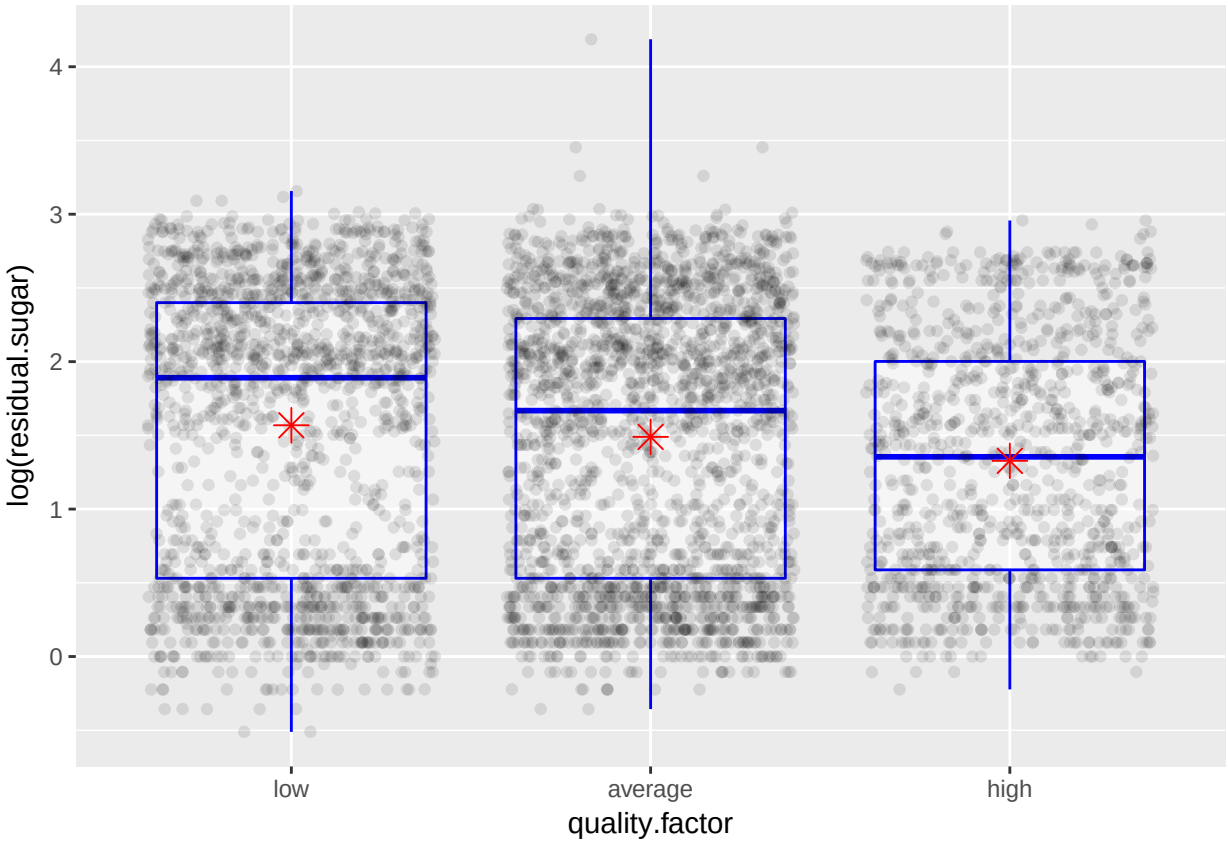
Let's check correlation coefficient between alcohol and quality:

```
##
##  Pearson's product-moment correlation
##
## data:  wine$alcohol and wine$quality
## t = 33.858, df = 4896, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4126015 0.4579941
## sample estimates:
##       cor
## 0.4355747
```

Correlation coefficient of 0.4355747 means that there is a positive meaningful correlation between alcohol content and quality.

Quality vs residual sugar

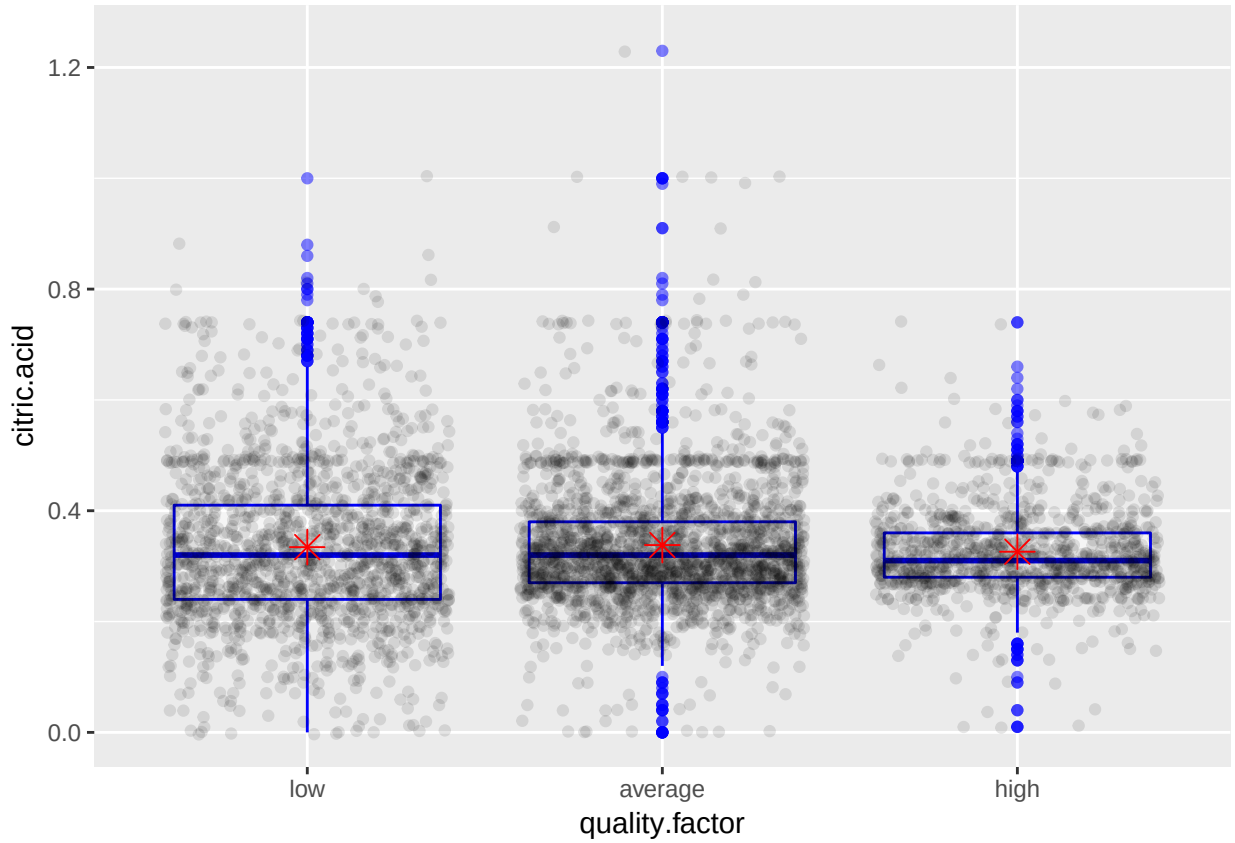
Next, onto relationship between residual sugar and quality:




```
##
##  Pearson's product-moment correlation
##
## data:  wine$residual.sugar and wine$quality
## t = -6.8603, df = 4896, p-value = 7.724e-12
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.12524103 -0.06976101
## sample estimates:
##         cor
## -0.09757683
```

Even though from the graph we may think there is a negative relationship between residual sugar and quality, the correlations coefficient is very low which tells me that these 2 variables little if any correlation.

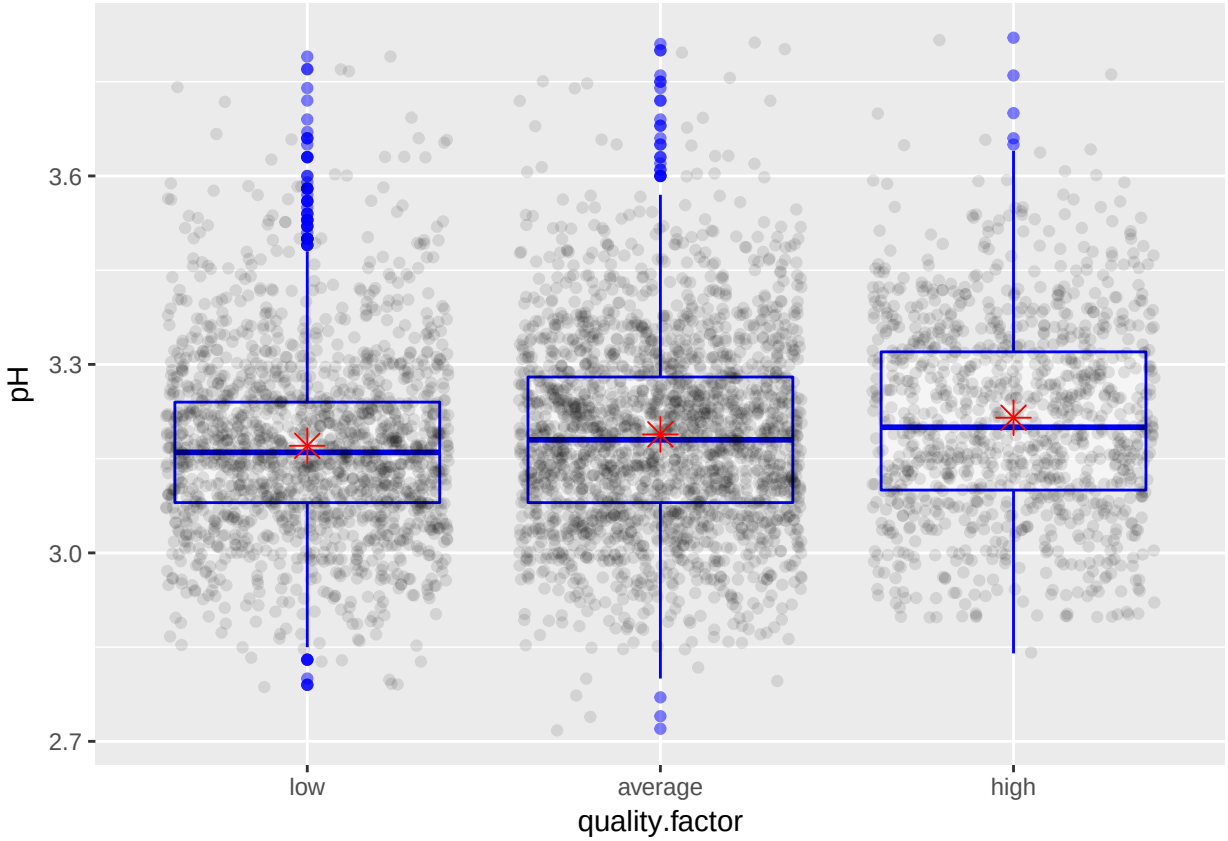
Quality vs Citric acid




```
##
##  Pearson's product-moment correlation
##
## data:  wine$citric.acid and wine$quality
## t = -0.6444, df = 4896, p-value = 0.5193
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.03720595  0.01880221
## sample estimates:
##          cor
## -0.009209091
```

Form the graph and correlation coefficient calculation we can conclude that there is little if any correlation between quality and citric acid.

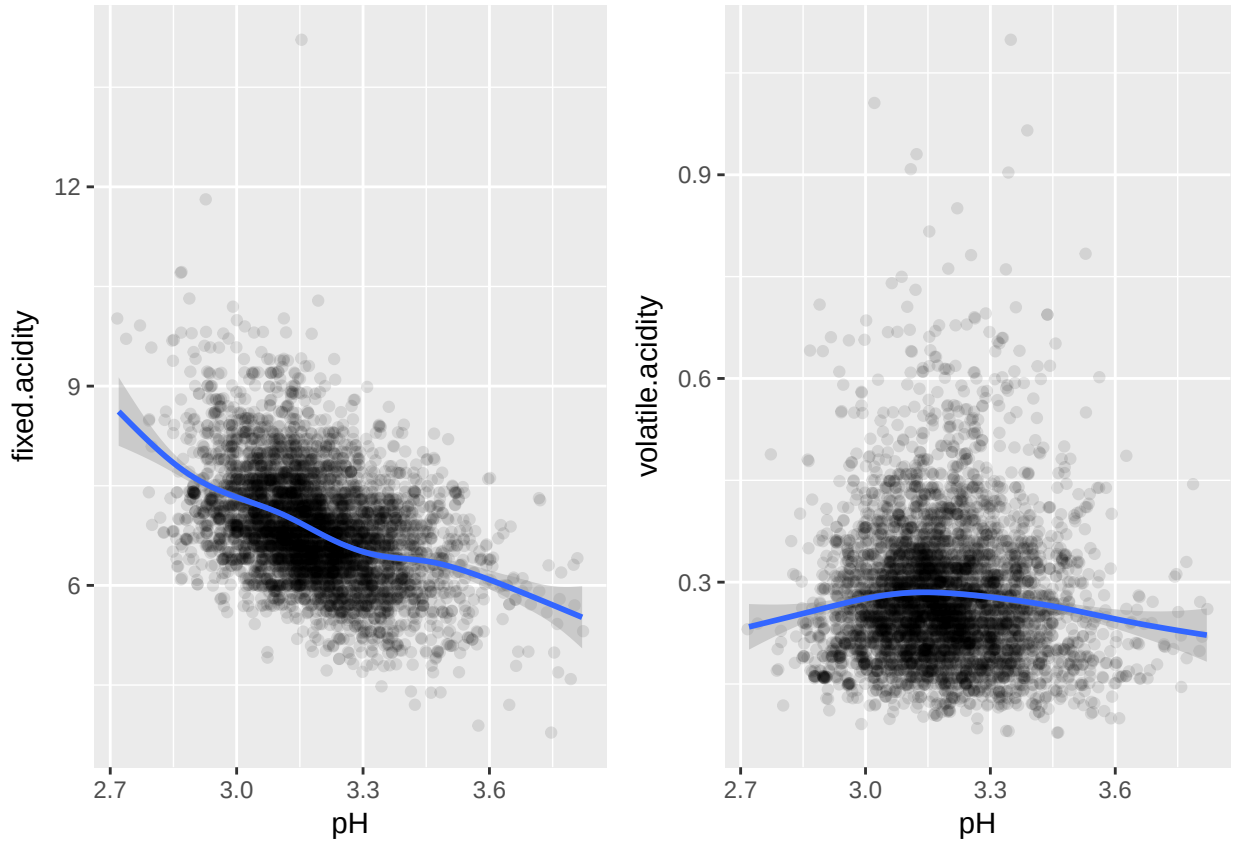
Quality vs pH




```
##
##  Pearson's product-moment correlation
##
## data:  wine$pH and wine$quality
## t = 6.9917, df = 4896, p-value = 3.081e-12
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.07162022 0.12707983
## sample estimates:
##       cor
## 0.09942725
```

Form the graph and correlation coefficient calculation we can conclude that there is little if any correlation between quality and pH.

pH vs acidity




```
cor.test(wine$pH, wine$fixed.acidity)
```

```
##
##  Pearson's product-moment correlation
##
## data:  wine$pH and wine$fixed.acidity
## t = -32.934, df = 4896, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4485154 -0.4026542
## sample estimates:
##        cor
## -0.4258583
```

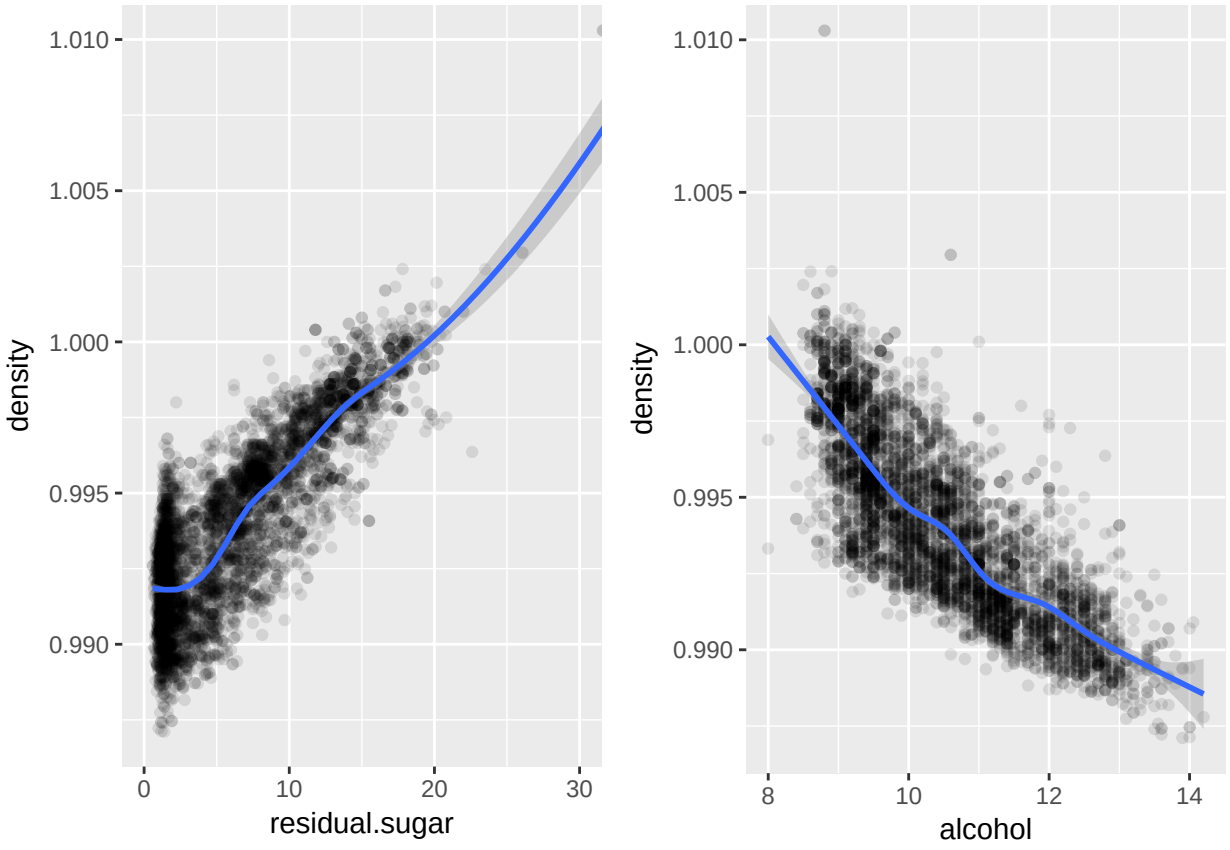
```
cor.test(wine$pH, wine$volatile.acidity)
```

```
##
##  Pearson's product-moment correlation
##
## data:  wine$pH and wine$volatile.acidity
## t = -2.2343, df = 4896, p-value = 0.02551
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.059868312 -0.003912409
## sample estimates:
```

```
##         cor
## -0.03191537
```

pH and fixed acidity have moderate correlation. The lower the fixed acidity the higher is pH. Which is not surprising since pH describes how acidic the wine is on a scale from 0 (very acidic) to 14 (very basic). On the other hand volatile acidity and pH have little to no correlation.

Dencity vs Alcohol and Residual sugar




```
##
##  Pearson's product-moment correlation
##
## data:  wine$density and wine$residual.sugar
## t = 107.87, df = 4896, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8304732 0.8470698
## sample estimates:
##       cor
## 0.8389665
```

```
##
##  Pearson's product-moment correlation
##
## data:  wine$density and wine$alcohol
## t = -87.255, df = 4896, p-value < 2.2e-16
```

```
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.7908646 -0.7689315
## sample estimates:
##       cor
## -0.7801376
```

There is a strong positive almost linear relationship between residual sugar and density and negative relationship between alcohol and density with a couple of outliers. Correlation coefficients in both cases are quite large.

Remaining variables

Next, I want to calculate correlation coefficients between quality and remaining variables to see the possibility of relationships I could have missed.

```
##
##  Pearson's product-moment correlation
##
## data:  wine$quality and wine$fixed.acidity
## t = -8.005, df = 4896, p-value = 1.48e-15
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.14121974 -0.08592991
## sample estimates:
##       cor
## -0.1136628
```

```
##
##  Pearson's product-moment correlation
##
## data:  wine$quality and wine$chlorides
## t = -15.024, df = 4896, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2365501 -0.1830039
## sample estimates:
##       cor
## -0.2099344
```

```
##
##  Pearson's product-moment correlation
##
## data:  wine$quality and wine$free.sulfur.dioxide
## t = 0.57085, df = 4896, p-value = 0.5681
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.01985292  0.03615626
## sample estimates:
##        cor
## 0.008158067
```

```
##
##  Pearson's product-moment correlation
##
## data:  wine$quality and wine$total.sulfur.dioxide
```

```
## t = -12.418, df = 4896, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2017563 -0.1474524
## sample estimates:
##        cor
## -0.1747372
```

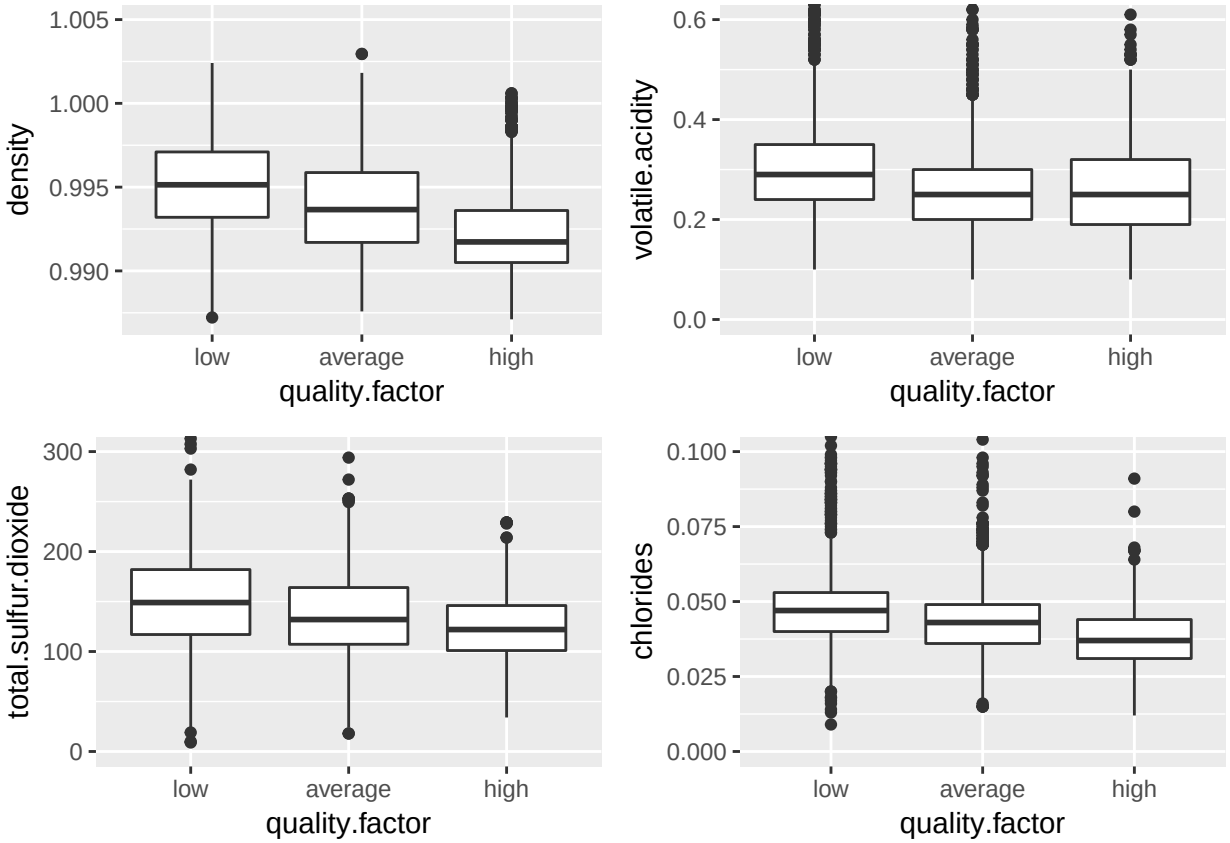
```
##
##  Pearson's product-moment correlation
##
## data:  wine$quality and wine$density
## t = -22.581, df = 4896, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3322718 -0.2815385
## sample estimates:
##        cor
## -0.3071233
```

```
##
##  Pearson's product-moment correlation
##
## data:  wine$quality and wine$volatile.acidity
## t = -13.891, df = 4896, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2215214 -0.1676307
## sample estimates:
##       cor
## -0.194723
```

```
##
##  Pearson's product-moment correlation
##
## data:  wine$quality and wine$sulphates
## t = 3.7613, df = 4896, p-value = 0.000171
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.02571007 0.08156172
## sample estimates:
##        cor
## 0.05367788
```

Out of remaining variables volatile acidity, density, total sulfur dioxide and chlorides have small negative correlation with quality. Density has the most significant coefficient among these variables.

I will graph variables mentioned above:



From above we see that all the variables have negative correlation with quality. The lower the variable value the higher the quality.

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 discovered some relationships between quality and several chemical features.

For example, there is a correlation between alcohol and quality. Better wines have higher alcohol content.

Density and quality have some correlation: the lower the density higher the quality.

Also, quality have negative relationships with volatile acidity, total sulfur dioxide and chlorides.

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

I discovered that residual sugar and alcohol have strong relationships with density.

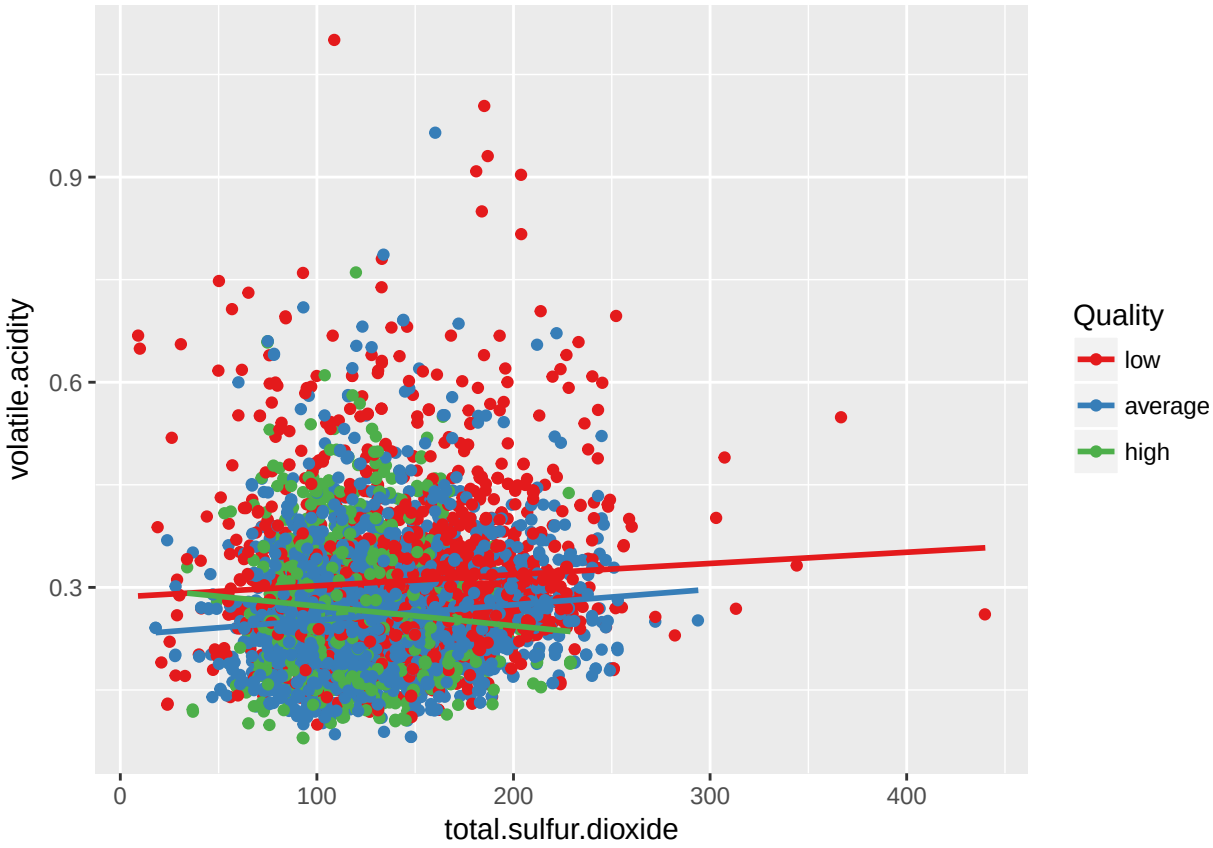
What was the strongest relationship you found?

The strongest relationship was density with residual sugar. Correlation coefficient for the relationship is 0.8389665 which means high positive correlation: the higher residual sugar is the higher density is.

Multivariate Plots Section

Volatile acidity vs total sulfur dioxide by quality

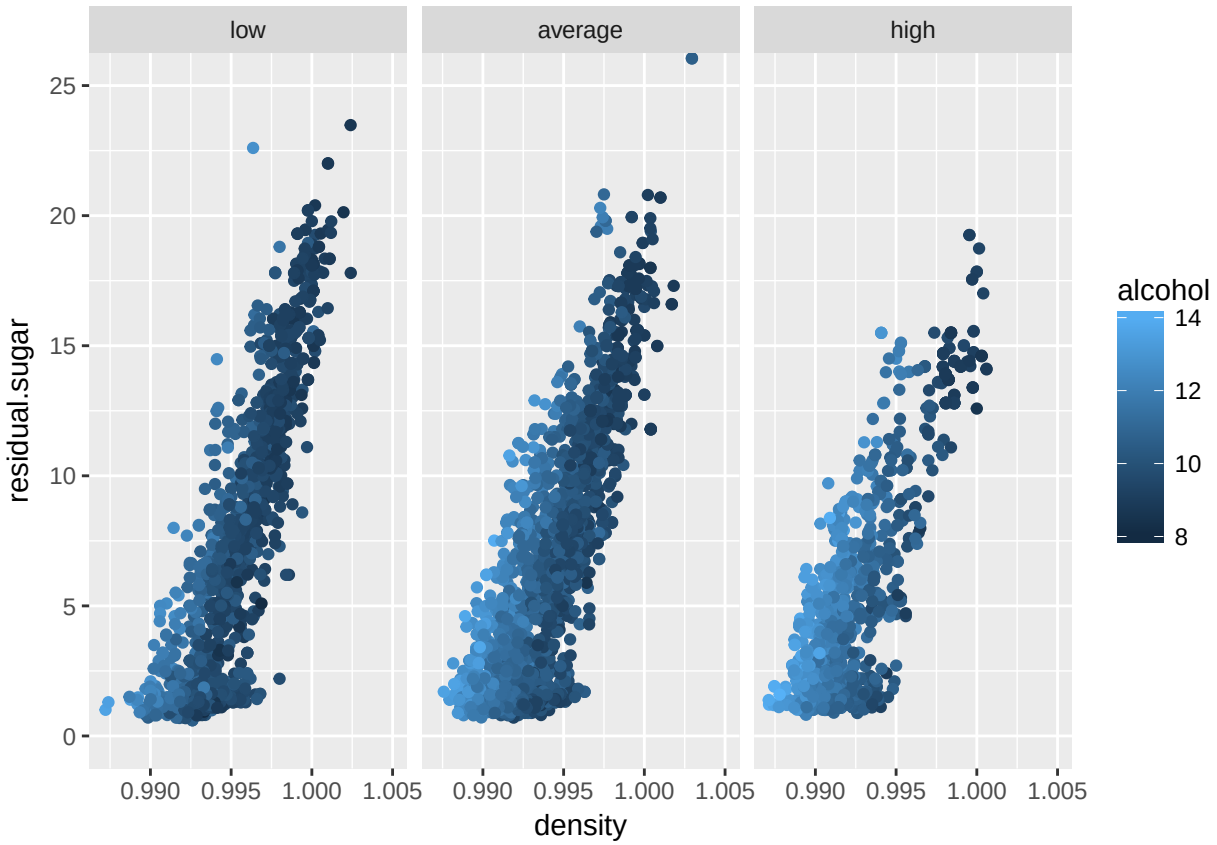
First I want to examine relationship between volatile acidity and total sulfur dioxide with quality.



It looks like better quality wines tend to concentrate closer the lower left corner than lower quality wines This means that better wines have lower volatile acidity and total sulfur dioxide.

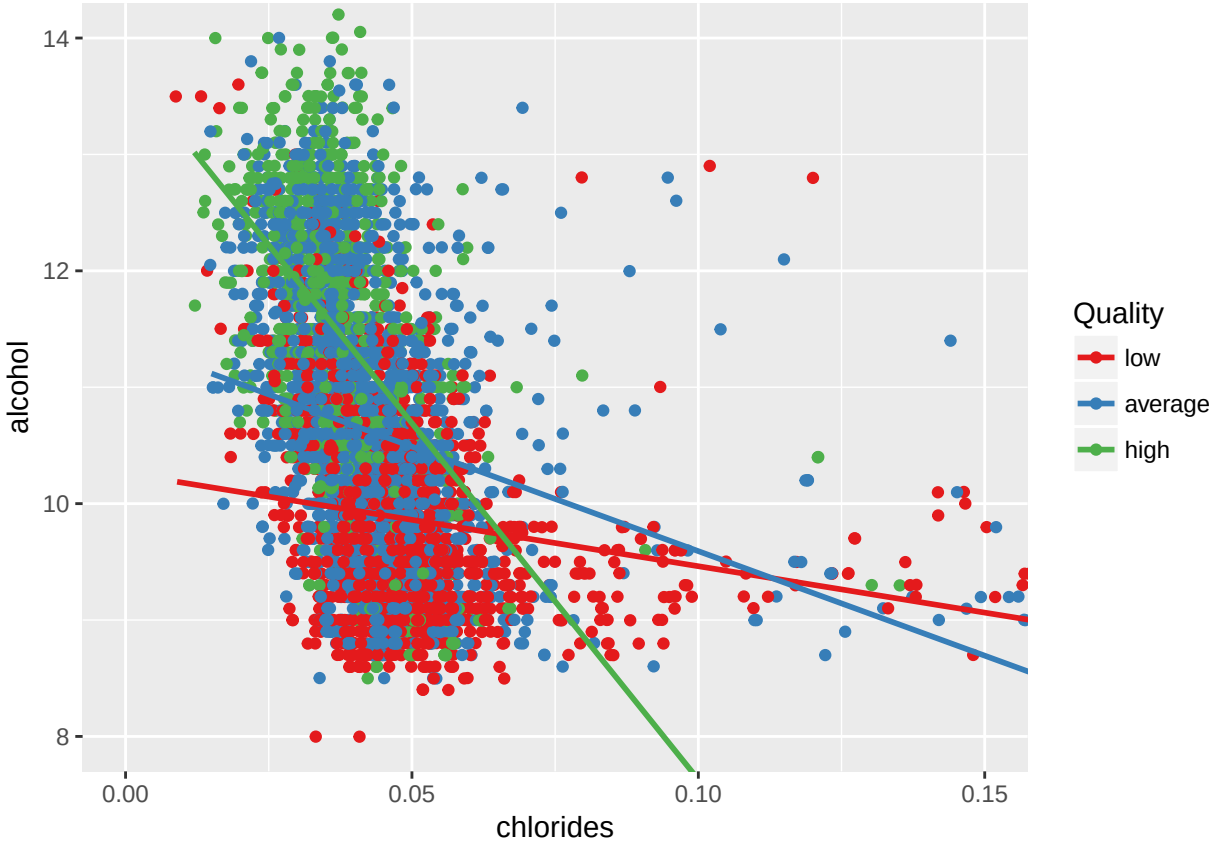
Density vs alcohol vs residual sugar

In bivariate section I discovered that both alcohol and residual sugar have strong relationships with density. Now I want to see how these 3 variables interact together.



From the graph we see that wines with lower residual sugar and higher alcohol are less dense than sweeter wines with less alcohol. What is interesting is that relationship between these 3 variables do not depend on quality of wines. For different levels of quality, we see almost identical plots.

Quality vs chlorides vs alcohol



Quality seem to increase with increasing alcohol and descreasing chlorides

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?

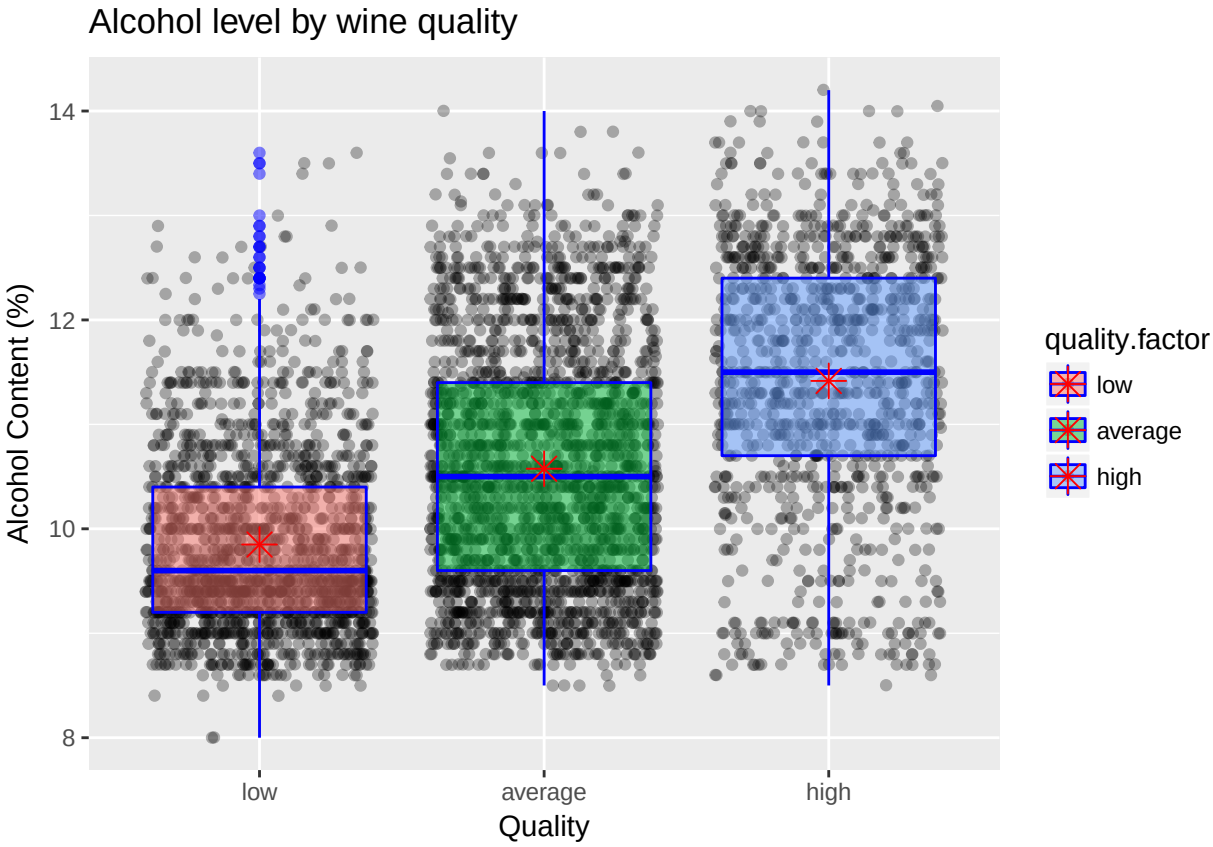
It became clear that low volatile acidity, low total sulfur dioxide, high alcohol and low chlorides contribute to the better wine quality

Were there any interesting or surprising interactions between features?

What I found interesting is that relationship between density, alcohol and residual sugar is independent of wine quality.

Final Plots and Summary

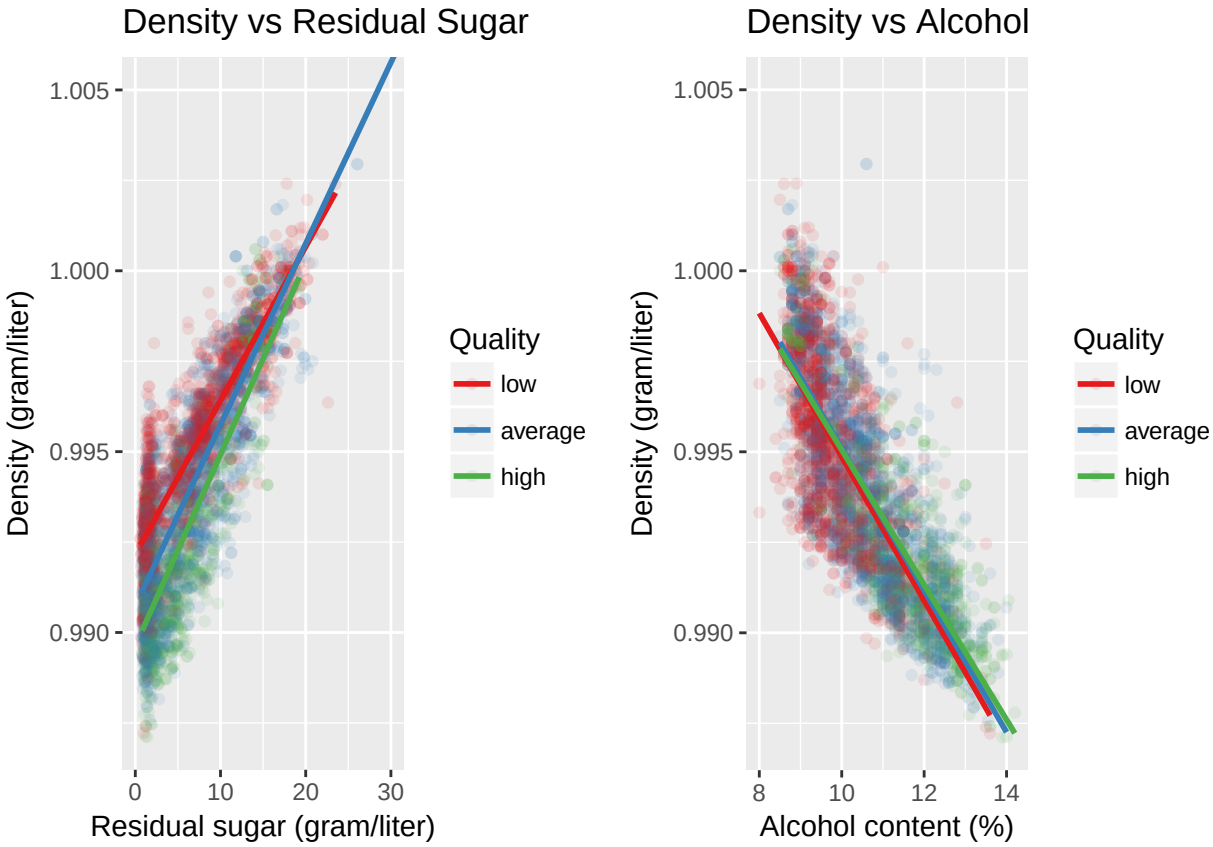
Plot One: Alcohol vs Quality Plot

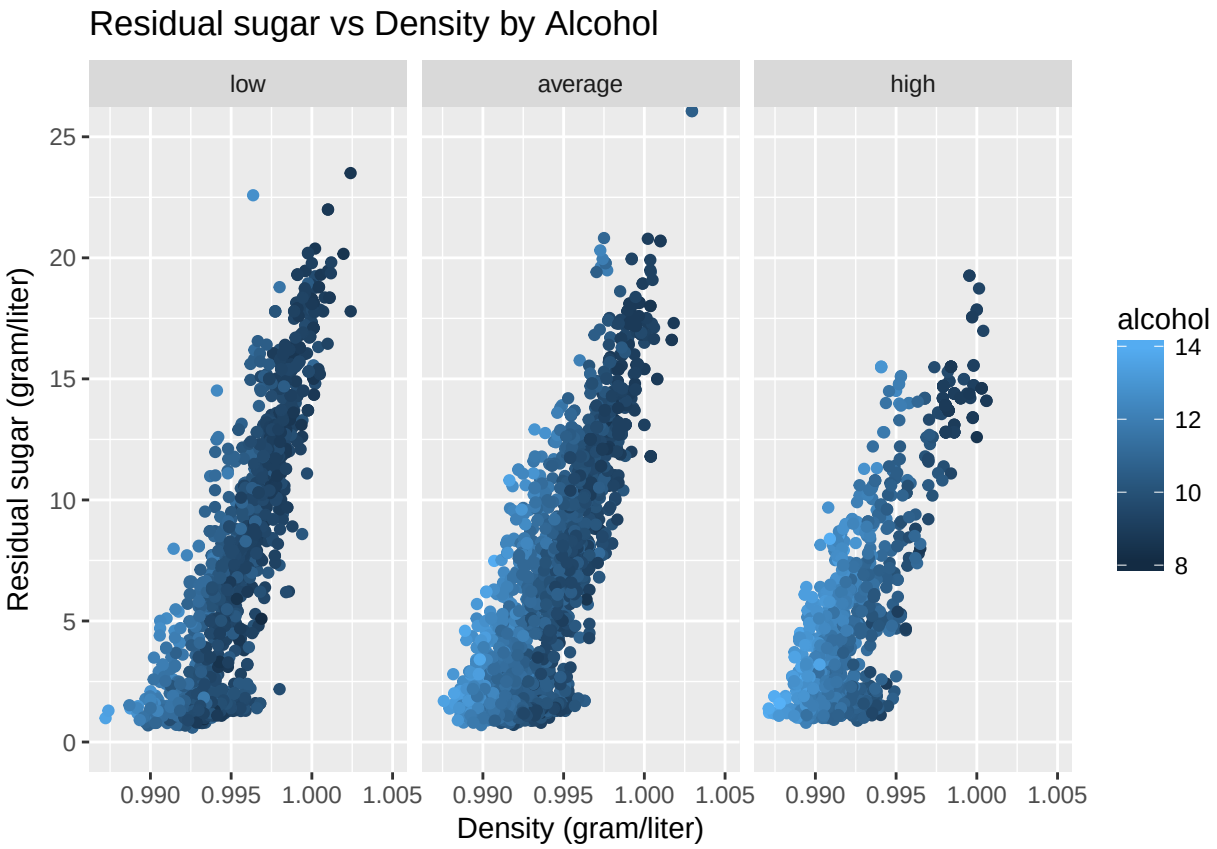


Description

I chose this plot because it represents the strongest relationship out of all I have found. It shows that with increasing alcohol content quality of wine goes up

Plot Two: Density vs alcohol and residual sugar



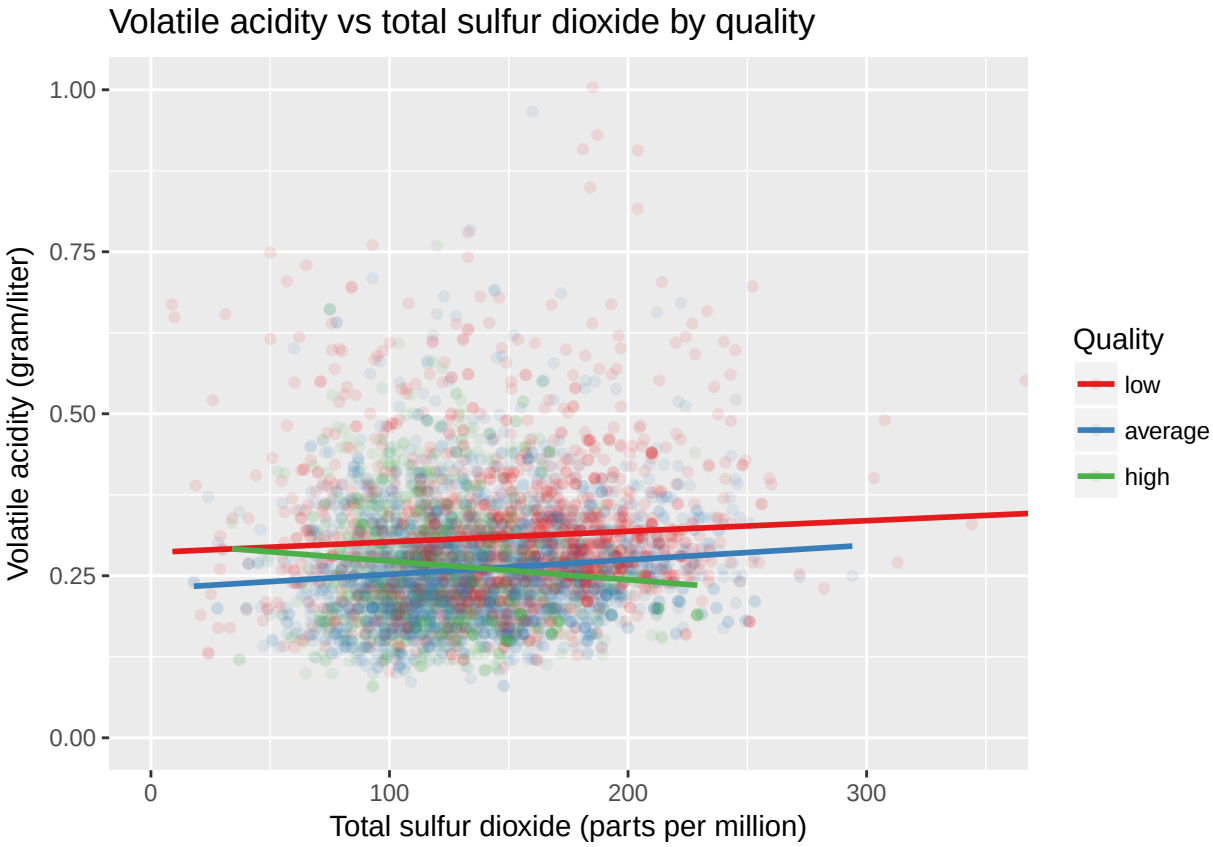


Description Two

I found 2 very strong relationships between density and alcohol (negative) and between density and residual sugar (positive). Both these relationships have high close to linear correlation.

What is more interesting, when put all these 3 variables against each other we see that relationship between these 3 variables do not depend on quality of wines. For different levels of quality, we see almost identical plots.

Plot Three: Volatile acidity vs total sulfur dioxide by quality



Description Three

Plot shows relationship between volatile acidity and total sulfur dioxide for different wine quality. Better wines have lower volatile acidity and total sulfur dioxide.

Reflection

It is common to think that wine tastings are subjective to the individual. However, in the analysis I had success in finding relationships between wine quality and several chemical factors.

The main factors that influence wine quality are alcohol (strong, positive relationship), density. volatile acidity, total sulfur dioxide and chlorides (all having negative relationship).

What surprised me is that I have not found a relationship between quality and citric acid which is supposed to add freshness and taste to wines. I believe further analysis should be done on citric acid to possibly uncover the relationship.

Also, I did not find a relationship between pH level and quality. That particular variable was among the ones I thought would have a great impact on quality.

One of the struggles I came across was the choice of variables that might influence wine quality. Without having a chemistry and/or wine background knowledge I found it hard to pick variables. From the description

of chemical factors and common sense I have selected several that I thought might influence quality, However, only after I explored each variable one at a time I found out which of those variables actually had relationships.

Also, I need to mention that correlation does not mean causation. All of the findings suggest that there might be a relationship however we cannot say that particular chemical factor caused better wine quality. For the possible future analyses it would be beneficial to carry a controlled expertise to identify if there is a causation.