Project - Exploratory Data Analysis with R

Annabel Lippold

22 October 2017

Qualityanalysis of White Wine by Annabel Lippold

Load wineQualityWhits.csv Data into R

```
##   X fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1 1           7.0             0.27        0.36           20.7     0.045
## 2 2           6.3             0.30        0.34            1.6     0.049
## 3 3           8.1             0.28        0.40            6.9     0.050
## 4 4           7.2             0.23        0.32            8.5     0.058
## 5 5           7.2             0.23        0.32            8.5     0.058
## 6 6           8.1             0.28        0.40            6.9     0.050
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  45                  170  1.0010 3.00      0.45     8.8
## 2                  14                  132  0.9940 3.30      0.49     9.5
## 3                  30                   97  0.9951 3.26      0.44    10.1
## 4                  47                  186  0.9956 3.19      0.40     9.9
## 5                  47                  186  0.9956 3.19      0.40     9.9
## 6                  30                   97  0.9951 3.26      0.44    10.1
##   quality
## 1       6
## 2       6
## 3       6
## 4       6
## 5       6
## 6       6
```

Analyzing structure of the data set

```
str(whiteWine)
```

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

The dataset of white Wines has 12 different variables and 4898 observations.
To analyze the quality of white Wine the variable Quality should splitted in different levels (bad, average, good).

Create a new variable for Quality

```
subdivided.quality = cut(whiteWine$quality, 3, labels = c('Bad','Average','Good'))
head(subdivided.quality)
```

```
## [1] Average Average Average Average Average Average
## Levels: Bad Average Good
```

I create a new Variable (subdivided.quality), to splite the factor variable quality into three categories (bad, average, good).

Add new Variable to the dataset of White Wine

```
quality_whiteWine <- whiteWine %>%
  mutate(subdivided.quality)
head(quality_whiteWine)
```

```
##   X fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1 1           7.0             0.27        0.36           20.7     0.045
## 2 2           6.3             0.30        0.34            1.6     0.049
## 3 3           8.1             0.28        0.40            6.9     0.050
## 4 4           7.2             0.23        0.32            8.5     0.058
## 5 5           7.2             0.23        0.32            8.5     0.058
## 6 6           8.1             0.28        0.40            6.9     0.050
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  45                  170  1.0010 3.00      0.45     8.8
## 2                  14                  132  0.9940 3.30      0.49     9.5
## 3                  30                   97  0.9951 3.26      0.44    10.1
## 4                  47                  186  0.9956 3.19      0.40     9.9
## 5                  47                  186  0.9956 3.19      0.40     9.9
## 6                  30                   97  0.9951 3.26      0.44    10.1
##   quality subdivided.quality
## 1       6            Average
## 2       6            Average
## 3       6            Average
## 4       6            Average
## 5       6            Average
## 6       6            Average
```

To use the new created variable, they have to be added to the original data set.
Now the data set has 13 variables.
In some cases it is better to create a new dataset, in case some Analyzes are not need the new variable. The new dataset with the created Variable subdivided.quality has the name quality_whitWine.

```
summary(whiteWine)
```

```
##        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
```

Fixed acidity has a wide range vom 3.8 to 14.2 compared to volatile acidity and citric acid they have a smaller range.
Also free and total sulfur dioxide has a wide range.
The alcohol of the different tested wines are between 8% and 14.20%.

```
summary(quality_whiteWine)
```

```
##        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      subdivided.quality
```

```
##  Min.   : 8.00   Min.   :3.000   Bad    :1640
##  1st Qu.: 9.50   1st Qu.:5.000   Average:3078
##  Median :10.40   Median :6.000   Good   : 180
##  Mean   :10.51   Mean   :5.878
##  3rd Qu.:11.40   3rd Qu.:6.000
##  Max.   :14.20   Max.   :9.000
```

I created the summary again with the new generated dataset, to look at the new variable. The summary of the variable "subdivided.quality" shows me that exist 1640 wines with a bad quality, 3078 with a average quality and only 180 Wines with a very good quality. This means the most Wines has a average resulat but 33% of the all tested Wines are bad.
When I looked deeper at the new variable, I realized that R divided the quality in the following step 0-5 are bad, 6-7 are average and 8-10 are good Wines.

Description of the different variables

Here follows a short description of the variables. To make a little bit more clear what the different chemical components of Wine are mean:

- fixed.acidity = Fixed acid or nonvolatile acid is an acid produced in the body from sources other than carbon dioxide, and is not excreted by the lungs. They are produced from e.g. an incomplete metabolism of carbohydrates, fats, and proteins.
- volatile.acidity = Volatile acid or Carbonic acid is a chemical compound with the chemical formula H2CO3 (equivalently OC(OH)2). It is also a name sometimes given to solutions of carbon dioxide in water (carbonated water), because such solutions contain small amounts of H2CO3.
- citric.acid = Citric acid is a weak organic tricarboxylic acid having the chemical formula C6H8O7. It occurs naturally in citrus fruits.
- residual.sugar = Among the components influencing how sweet a wine will taste is residual sugar. It is usually measured in grams of sugar per litre of wine, often abbreviated to g/l or g/L. Residual sugar typically refers to the sugar remaining after fermentation stops, or is stopped, but it can also result from the addition of unfermented must (a technique practiced in Germany and known as Süssreserve) or ordinary table sugar.
- chlorides = The chloride is the anion (negatively charged ion) Cl.
  It is formed when the element chlorine (a halogen) gains an electron or when a compound such as hydrogen chloride is dissolved in water or other polar solvents.
- free.sulfur.dioxide = Sulfur dioxide (also sulphur dioxide) is the chemical compound with the formula SO^2. At standard atmosphere, it is a toxic gas with a pungent, irritating smell. The triple point is 197.69 K and 1.67 kPa. It is released naturally by volcanic activity.
- total.sulfur.dioxide = Sulfur dioxide exists in wine in free and bound forms, and the combinations are referred to as total SO2. Binding, for instance to the carbonyl group of acetaldehyde, varies with the wine in question. The free form exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion, which is in turn in equilibrium with sulfite ion.
- density = The density, of a substance is its mass per unit volume.

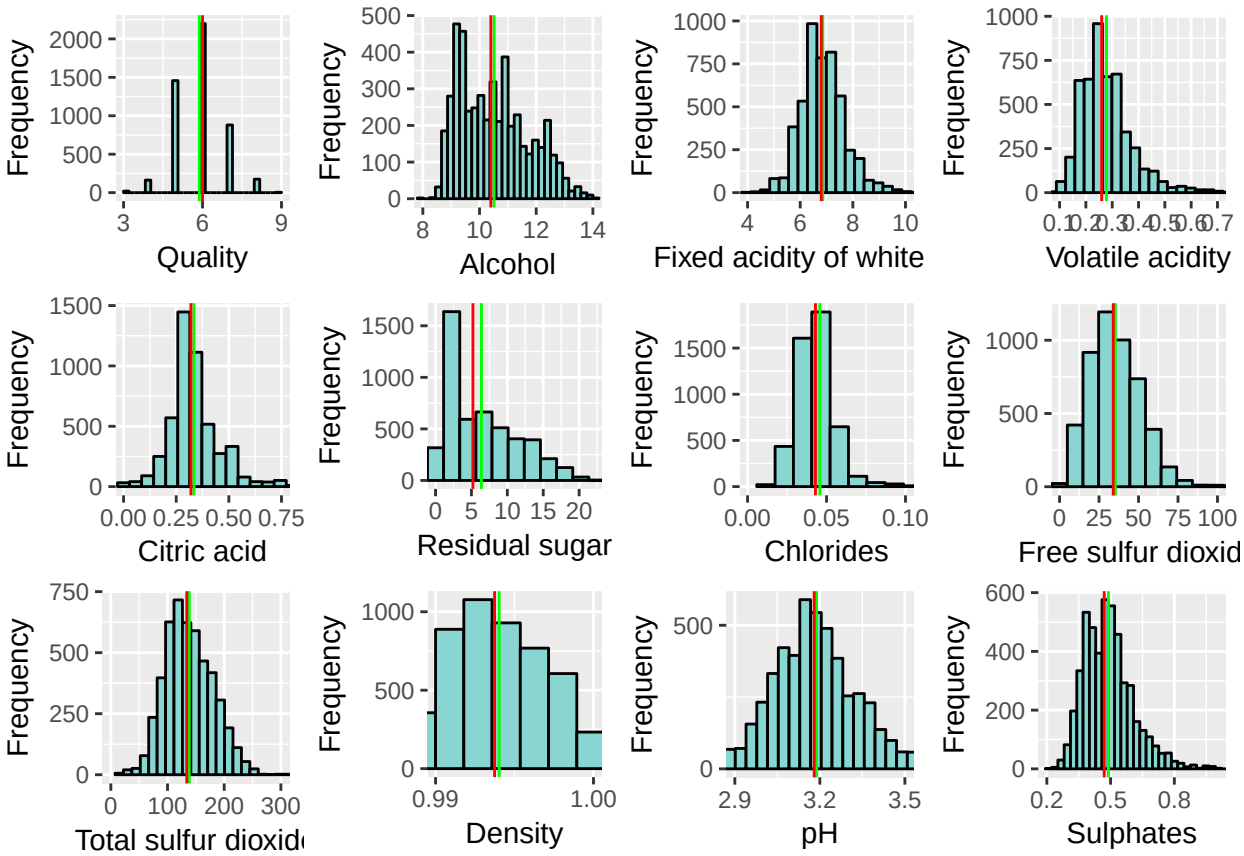
- pH = pH is a numeric scale used to specify the acidity or basicity.
- sulphates = A sulphate is a salt of sulphuric acid.
- alcohol = Alcohol by volume is a standard measure of how much alcohol. Ethanol is contained in a given volume of an alcoholic beverage. Wine has in the most cases between 9% and 16% of alcohol.
- quality = Defines the Quality of a wine. 0 means the Quality of a wine is very bad; 10 the wine has the best quality.

Introduction - what will be analyzed

- White Wine has a lot of different chemical components. In the following analysis I will show how, these variables / component influence each other and try to analyze the following questions by exploring the data set of White Wine:
  - 1. How are the frequencys of Distribution for the variable quality and alcohol?
  - 2. Which variables influence the vol. of alcohol?
  - 3. How many white Wines get a bad, average or a good quality?
  - 4. Which chemical properties influence the quality of white wines?

Univariate Plots Section

1. Showing distribution of the different variables for White Wine.



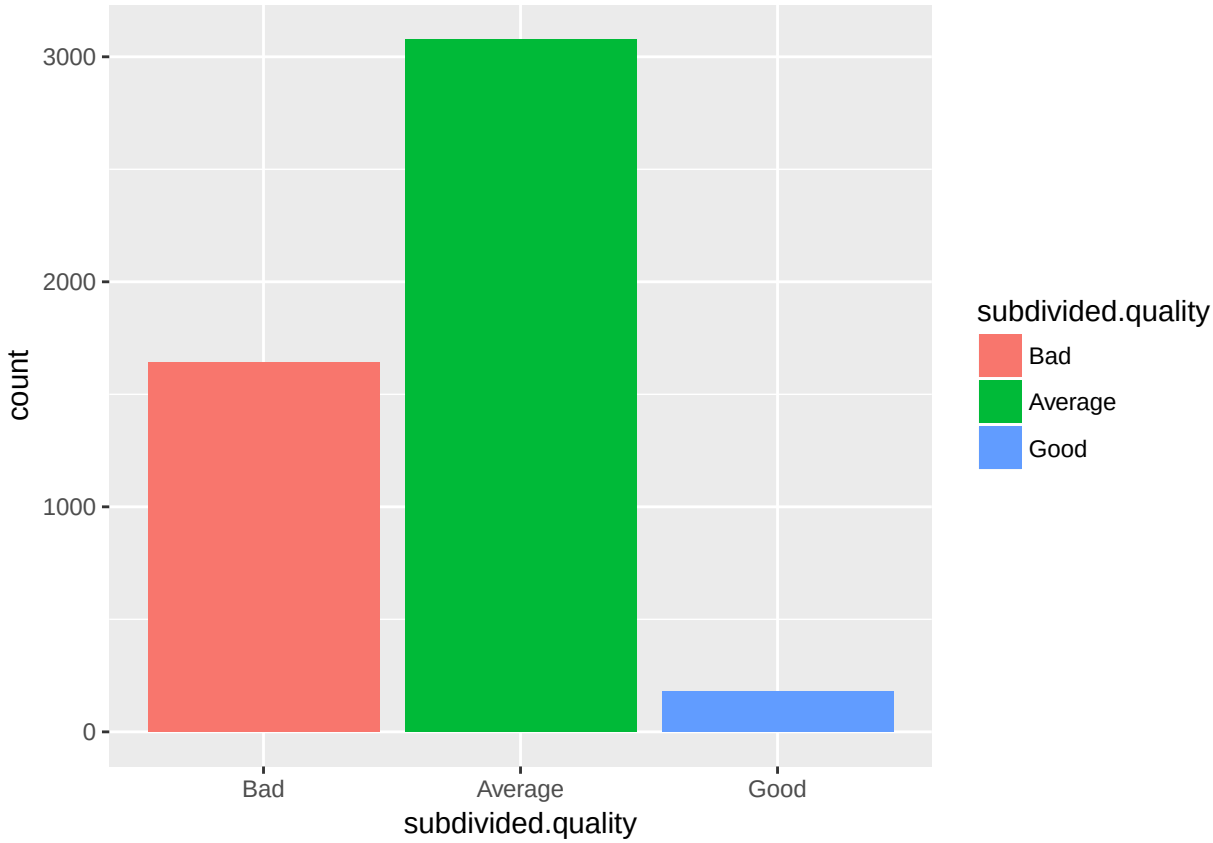
Alcohol and Residual Sugar have a bimodal distribution, all other variables shows a normal distribution.
The red line in the histograms are the median. The green are the mean.
Mean and Median are in the most cases pretty close, only in residual sugar they are not so close, this could be the reason of the bimodal distribution.

Distribution of the new created variable "Subdivided Quality"

```
count_sub_quality <- count(quality_whiteWine, subdivided.quality)
print(count_sub_quality)
```

```
## # A tibble: 3 x 2
##   subdivided.quality     n
##                <fctr> <int>
## 1                Bad  1640
## 2            Average  3078
## 3               Good   180
```

```
ggplot(data = quality_whiteWine, aes(x=subdivided.quality)) +
  geom_bar(aes(fill = subdivided.quality))
```



The Distribution of the variable "subdivided quality" shows that the most wines has a average quality (2/3). Only a few wines has a good quality. A bad quality has 1/3 of the tested wines.

Univariate Analysis

What is the structure of your dataset?

The dataset has 11 variables, there are 10 numerical and ond 1 integer variable.
The dataset of White Wine contains 4898 observations.
The most Wines have a average quality (mean = 5.878, median = 6) and a volume of alcohol around 10%.
Interessting is that sugar has a mean of 6.39 and a meidan of 5.2 but a max of 65.8. This means one ore more wines has 10 times more sugar than the average white Wine.

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

The interessting variables of the dataset are quality, alcohol, density and sugar. With Correlation test I will look at the other components, if they influence these influence the three interessting variables.

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

Acid and sulfur dioxide.

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

Yes, i created the variable subdivided.quality, which gives the quality three levels.
1. Level is "bad" for Wines with a Quality between 0 and 5.
2. Level is "average" with a Quality between 6 and 7.
3. Level is "good" with a Quality of 8 or better. Max Quality that a Wine can earn is 10.

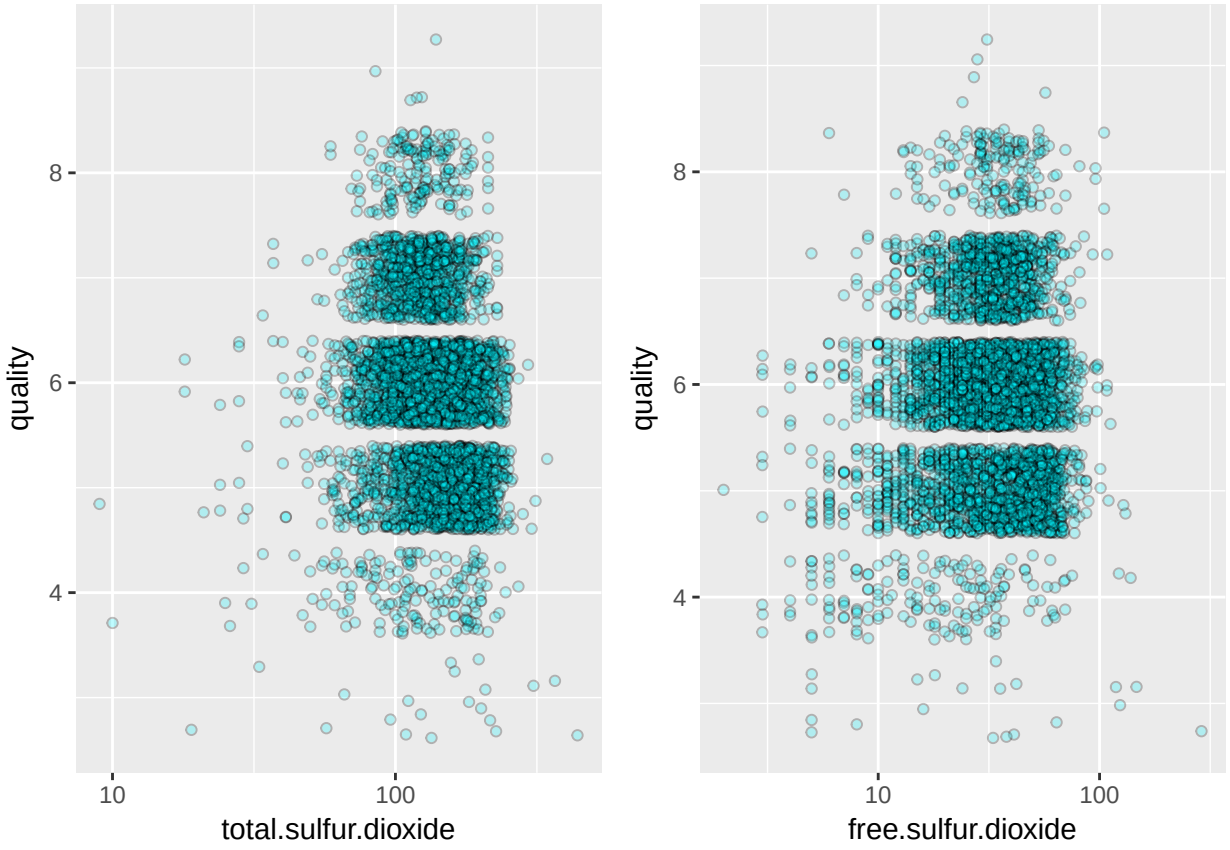
Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

No I didn't found an unusual distribution.

Bivariate Plots Section

Compare total and free sulfor dioxid

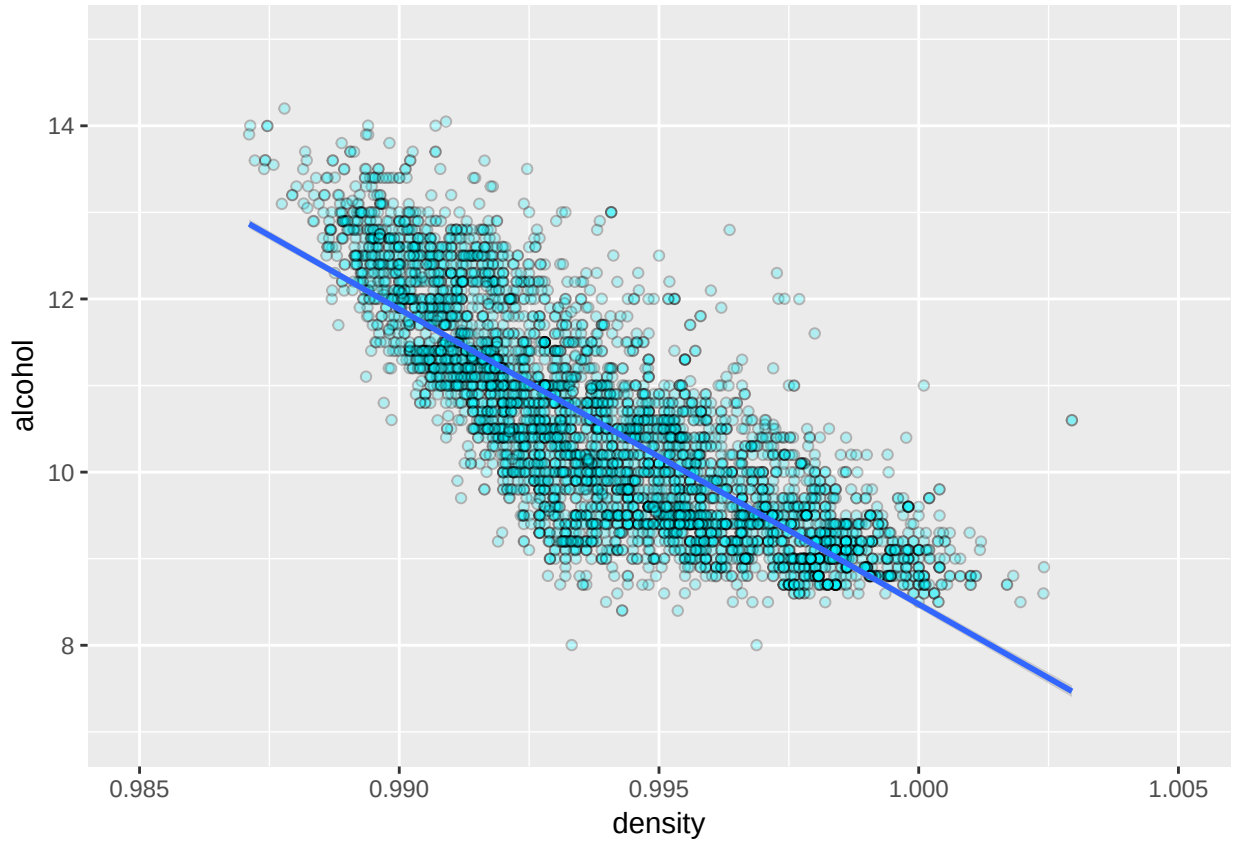
```
sd1 <- ggplot(data = whiteWine, aes(x = total.sulfur.dioxide, y = quality))+
  geom_jitter(alpha = 1/4, shape = 21, fill = '#00f5ff')+
  scale_x_log10()


sd2 <- ggplot(data = whiteWine, aes(x = free.sulfur.dioxide, y = quality))+
  geom_jitter(alpha = 1/4, shape = 21, fill = '#00f5ff')+
  scale_x_log10()


grid.arrange(sd1, sd2, ncol = 2)
```



Total sulfur dioxide is near double of free sulfur dioxide
because free is part of total, but what is the other KPI in total?

Density and alcohol

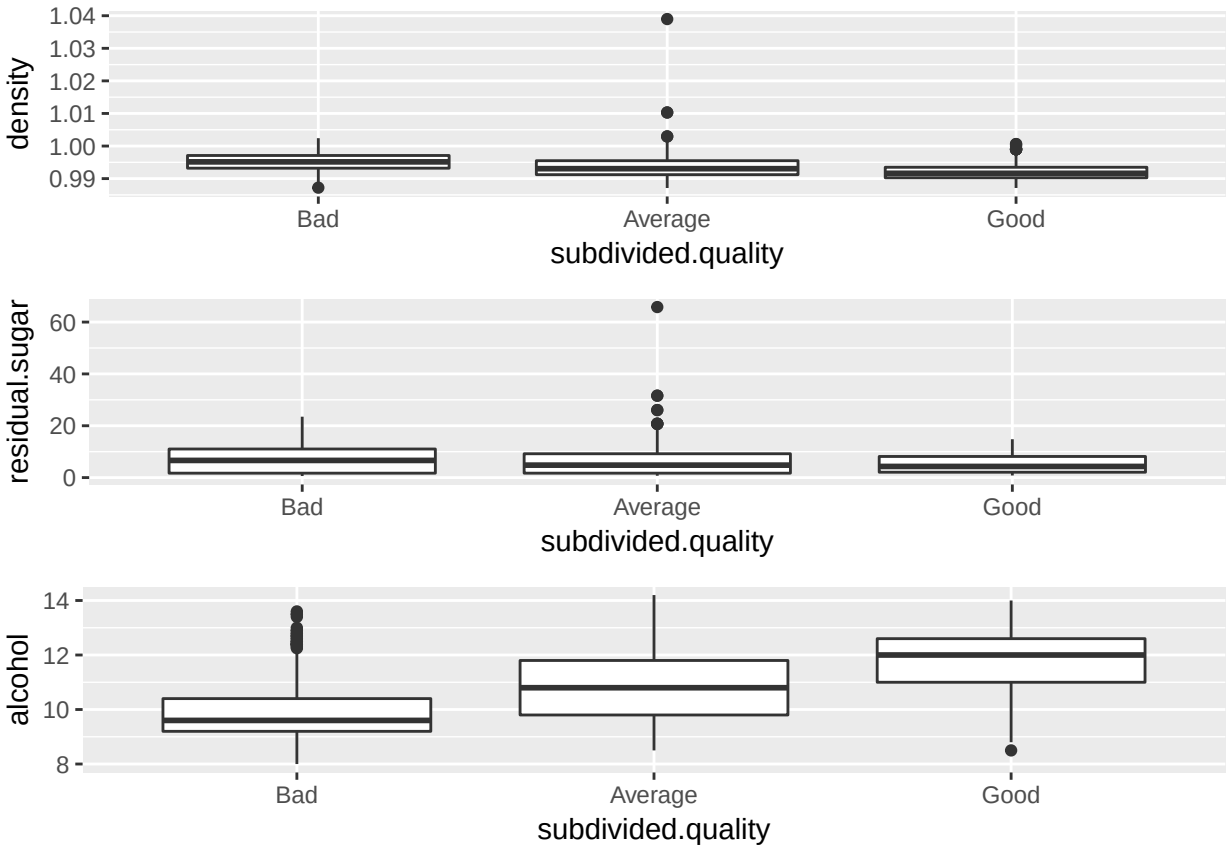
```
ggplot(whiteWine, aes(x = density, y = alcohol))+
  geom_jitter(alpha = 1/4, shape = 21, fill = '#00f5ff') +
  geom_smooth(method = "lm") +
  xlim(c(0.985, 1.005))+
  ylim(c(7, 15))
```



Density and alcohol shows a trend for a lower density a wine has a higher volume of alcohol and for a higher density a lower volume of alcohol.
It looks like that alcohol and densita are influent variables. This should be tested with a correlation test in the next steps.

Outlier for Density, Sugar and Alcohol

```
b1 <- ggplot(data = quality_whiteWine, aes(x = subdivided.quality, y = density))+
  geom_boxplot()+
  guides(fill=FALSE)

b2 <- ggplot(data = quality_whiteWine, aes(x = subdivided.quality, y = residual.sugar))+
  geom_boxplot()+
  guides(fill=FALSE)

b3 <- ggplot(data = quality_whiteWine, aes(x = subdivided.quality, y = alcohol))+
  geom_boxplot()+
  guides(fill=FALSE)

grid.arrange(b1, b2, b3)
```



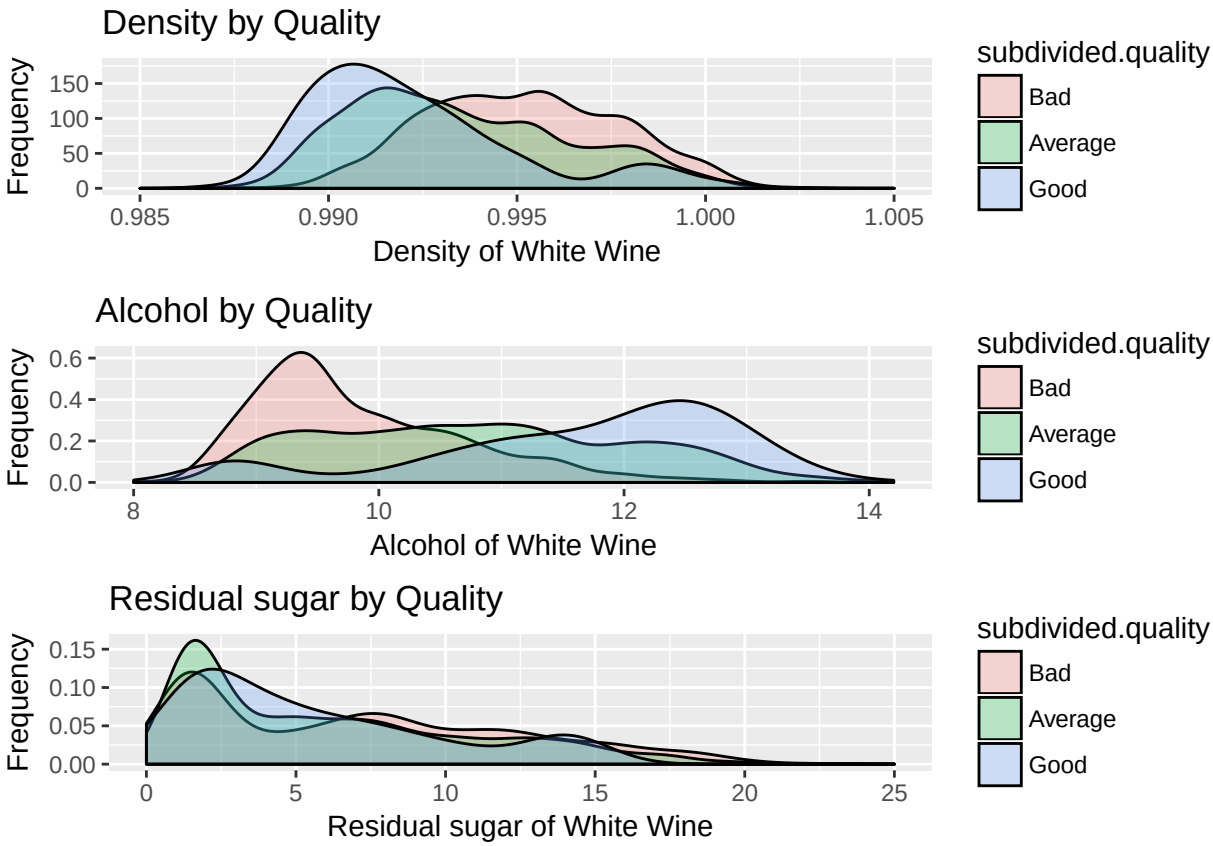
Outlier of variables are easy to visible with boxplots.
All three variables (Density, Sugar and Alcohol) have outliers.
For Sugar and density, wine with an average quality has the most and extremst outliers, alcohol has the most outliers in the bad charge of wines.
For alcohol is also very good visible that bad wines has less alcohol compared to average wines and average wines has less alcohol compared to good wines. This could mean if a wines has more alcohol, it has a better quality.

Quality Distribution for importent Wine Variables

In the Analysis of white Wine, we can see that the variables, Density Alcohol and Sugar are interessting variables.
In the follwing step I look at these three variables and try to conpare how they are distributed divided in the three defined quality ranges ("subdivided.quality").

```
dist1 <- ggplot(quality_whiteWine, aes(density, fill = subdivided.quality)) +
  geom_density(alpha = 1/4)+
  xlim(c(0.985, 1.005))+
  ggtitle("Density by Quality")+
  xlab("Density of White Wine")+
  ylab("Frequency")

dist2 <- ggplot(quality_whiteWine, aes(alcohol, fill = subdivided.quality)) +
  geom_density(alpha = 1/4)+
```

```
  ggtitle("Alcohol by Quality")+
  xlab("Alcohol of White Wine")+
  ylab("Frequency")

dist3 <- ggplot(quality_whiteWine, aes(residual.sugar, fill = subdivided.quality)) +
  geom_density(alpha = 1/4)+
  xlim(c(0, 25))+
  ggtitle("Residual sugar by Quality")+
  xlab("Residual sugar of White Wine")+
  ylab("Frequency")

grid.arrange(dist1, dist2, dist3)
```



Density:
For the KPI Density, white Wines with a good Quality has a smaller range than wines with a bad or average quality. It is also visible that a good wine has in average a smaller density than a bad wine.
It looks like that ***density influence the quality*** of white Wine!
This means as smaller Density of a white Wine is as better is the quality.
Happens this in general?
Another point that you have to keep in mind is, if a wine has a density below 0.987 we can't say a wine will be better than the tested Wines. Is a density below 0.985 possible?

Alcohol:
The distirbutions for alcohol shows that wine with a small volume of alcohol

has a bad quality and a wine with a good quality has more volume. Wine with a average quality has a uniform distribution this means that for every % of alcohol the count of Wines are pretty close.
Like for Density it looks like that ***alcohol is also a variable that has an influent factor on quality.***

Sugar:
Sugar compared with different qualities shows, that they have similar distributions. The most wines (for every quality type) has a high frequency for sugar between 0 and 5 g. As more sugar increase the Frequency decrease.
This means that ***sugar hasn't a high influence on 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?

To analyze relationships between the variables I will use correlation tests. Creating Cor-tests for all variables takes a lot of time and it give not a nice overview, so I decided to create a Corplot which shows all correlations between each variable, this makes it a lot of easier to analyze which variable influence another one.

```
whiteWine$X <- NULL
W <- cor(x = whiteWine)
head(round(W,2))
```

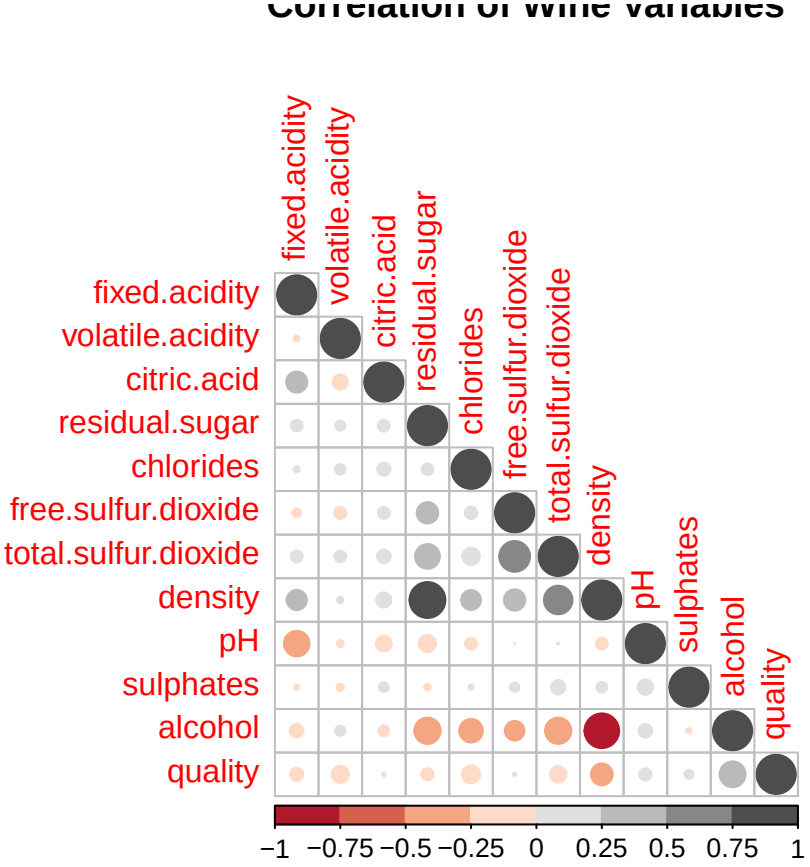
```
##                      fixed.acidity volatile.acidity citric.acid
## fixed.acidity                 1.00            -0.02        0.29
## volatile.acidity             -0.02             1.00       -0.15
## citric.acid                   0.29            -0.15        1.00
## residual.sugar                0.09             0.06        0.09
## chlorides                     0.02             0.07        0.11
## free.sulfur.dioxide          -0.05            -0.10        0.09
##                      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 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
##                      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
```

```
library(RColorBrewer)
```

```
## Warning: package 'RColorBrewer' was built under R version 3.4.1
```

```
corrplot(W, method = "circle", type = "lower",
         col=brewer.pal(n = 8, name = "RdGy"), title = "Correlation of Wine Variables")
```



Were there any interesting or surprising interactions between features?

This Plot shows the correlation of all variables in the original data set and how they influence each other.
There is visible that the variable quality has no big correlation to other variables. The highes coorelation would found for alcohol and density.

More interessting is that density has high correlations to two other variables.
+ 1. Density and alcohol has a high correlation close to -1.
+ 2. Density and residual sugar has a high positice correlation to 1.

Alcohol is the variable that influenced the most other variables (like to sugar, chlorides, free- and total sulfur dioxide), but not as much as density.

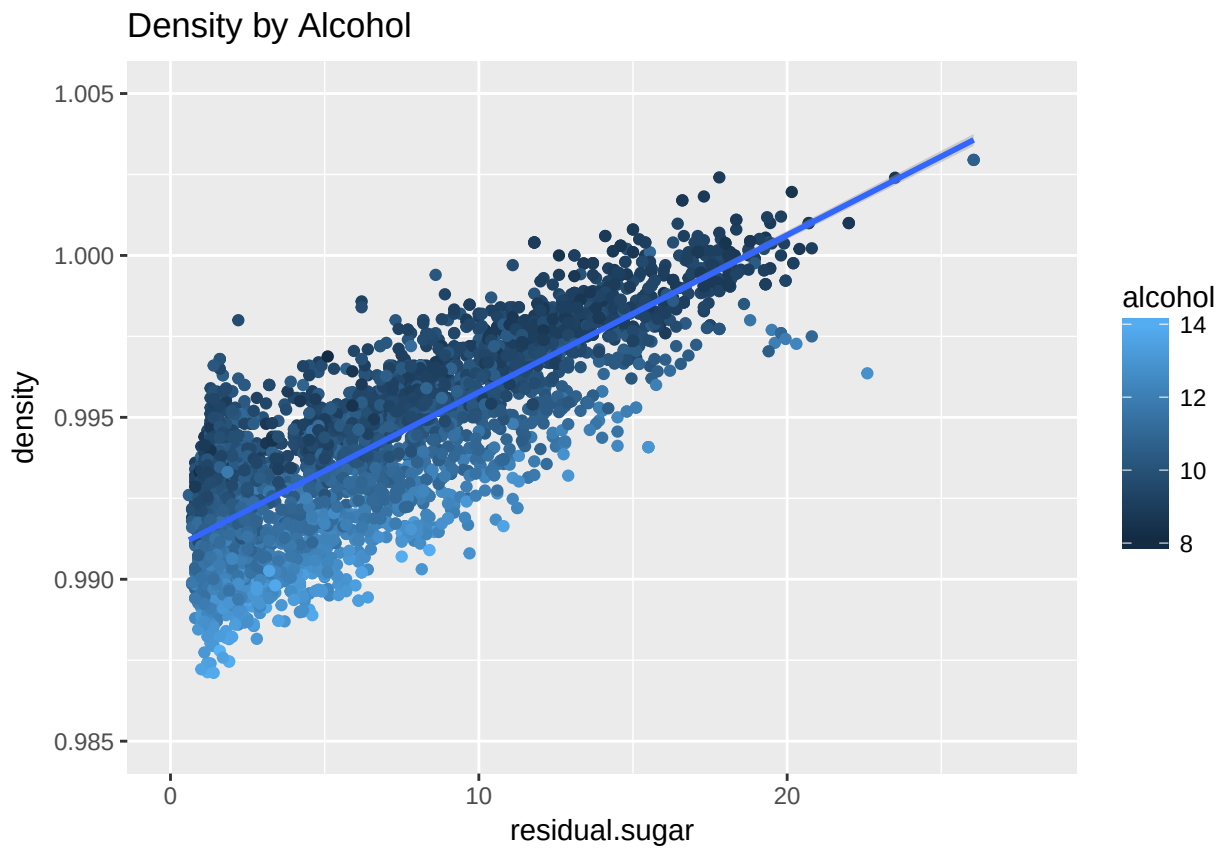
Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

Interessting can be also the taste of a wine, or the price. For price it can be interessting to see if a good wine can also be cheap or they are only expensive ones?. Same for bad wines. Or what is the average price of a Wine with a average quality?
Country of Production of wines and grape variety can also be interessting factor which influence the quality of wine.

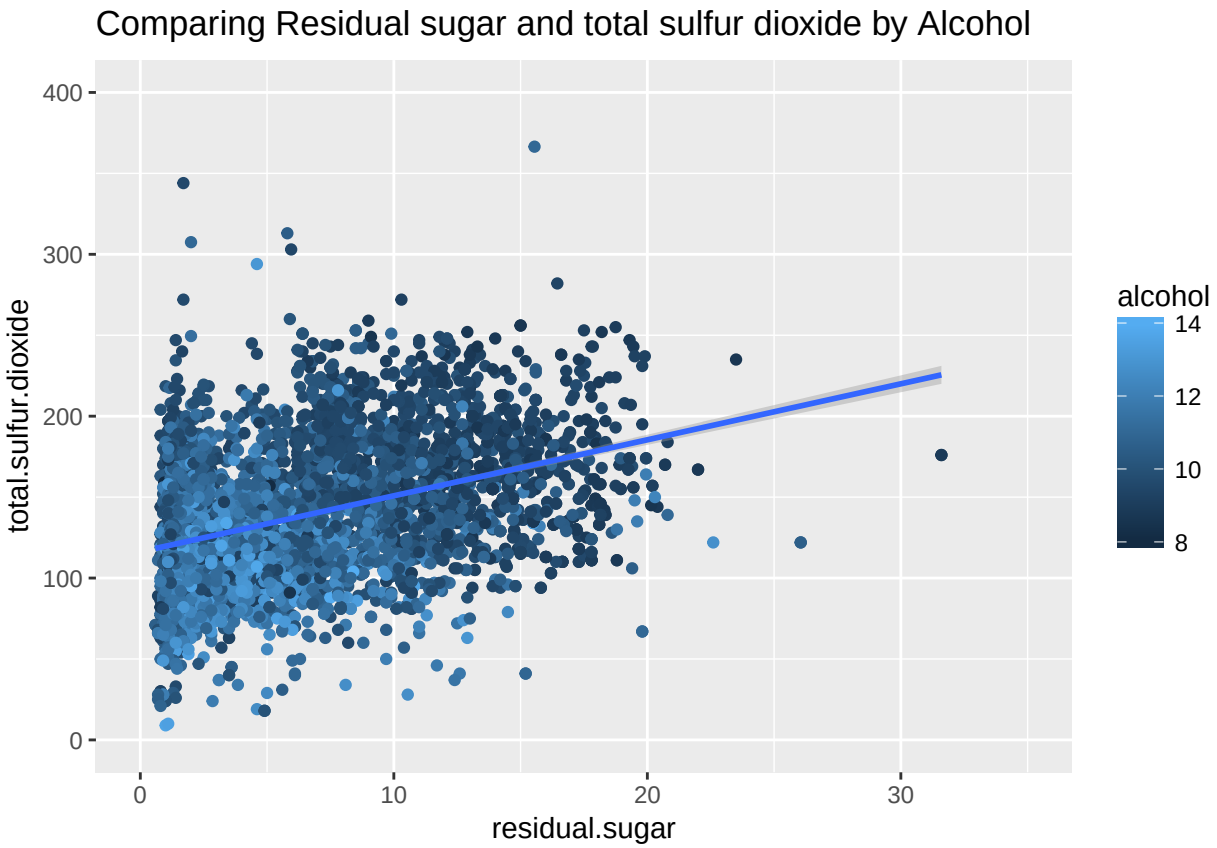
What was the strongest relationship you found?

The strongest relationship between two variables I found was residual.sugar and density.
Alcohol and density have also a strong relationship but not as strong as density and sugar.

Multivariate Plots Section



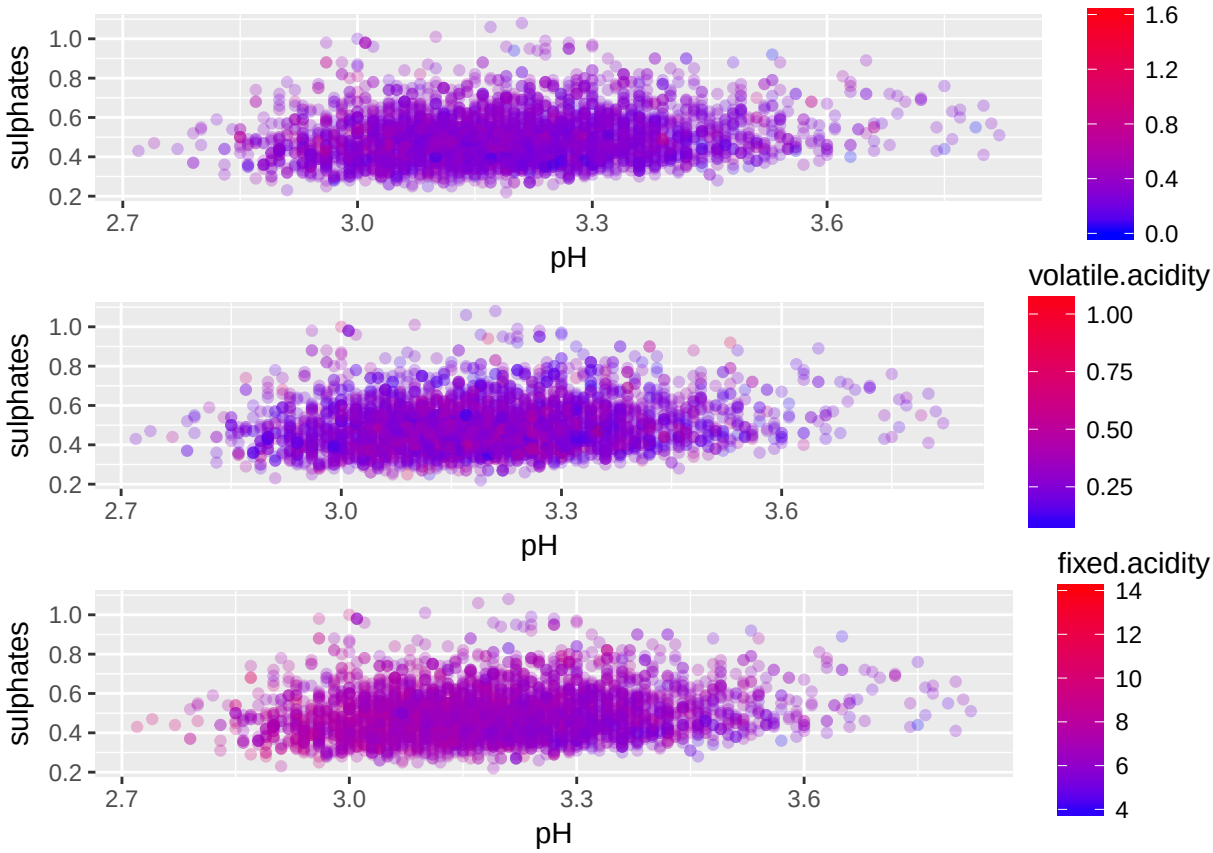
The Plot Density by Alcohol shows that a wine with high alcohol has less sugar and a lower density, a wine with a low volume of alcohol has a higher density and in many cases more gram of sugar.



White wines with a high volume of alcohol has less sugar and a smaller total sulfur dioxide range than white wines with a lower volume of alcohol.

Comparing different Acids in Wine

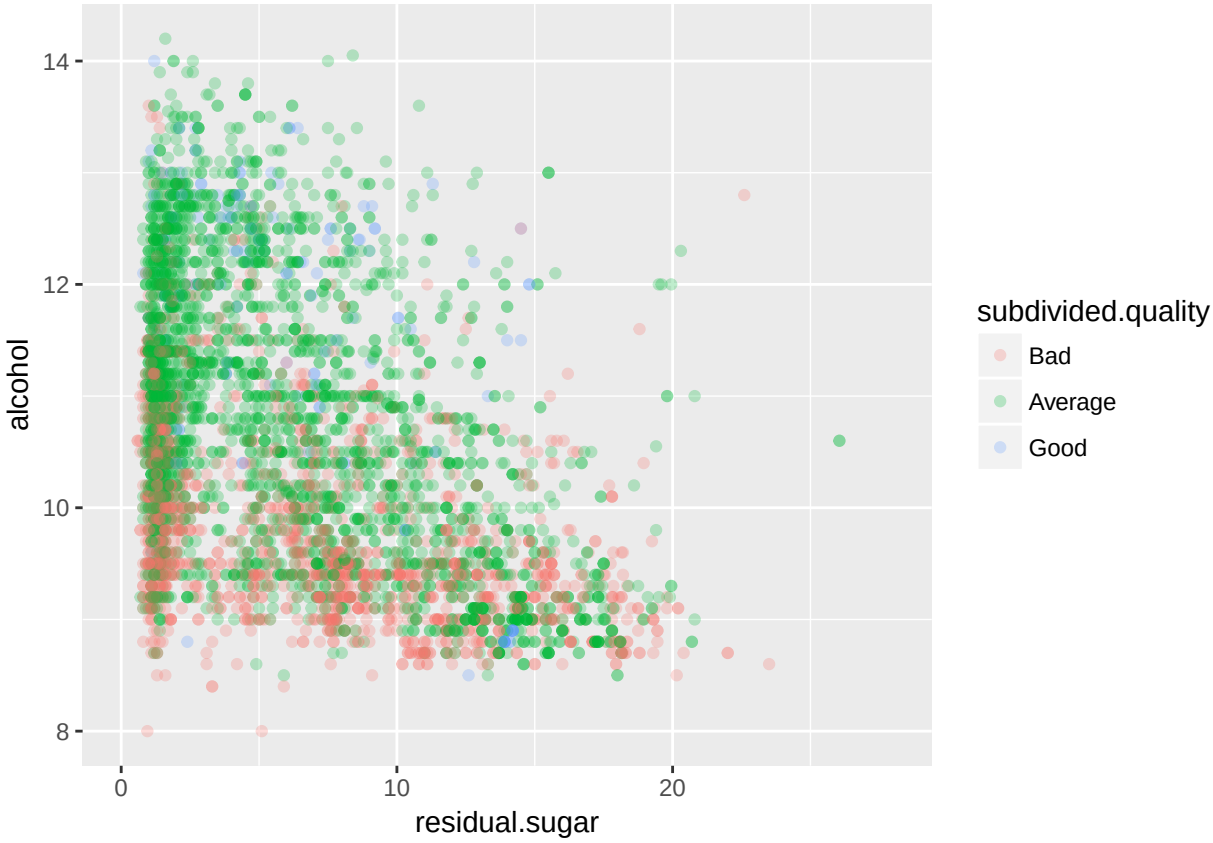
```
a1 <- ggplot(data = whiteWine, aes(x = pH, y = sulphates, color = citric.acid))+
  geom_point(alpha = 1/4)+
  scale_color_gradient(low = "blue", high = "red")

a2 <- ggplot(data = whiteWine, aes(x = pH, y = sulphates, color = volatile.acidity))+
  geom_point(alpha = 1/4)+
  scale_color_gradient(low = "blue", high = "red")

a3 <- ggplot(data = whiteWine, aes(x = pH, y = sulphates, color = fixed.acidity))+
  geom_point(alpha = 1/4)+
  scale_color_gradient(low = "blue", high = "red")


grid.arrange(a1, a2, a3, ncol = 1)
```



When comparing the different acids with variables which describes by acid (ph and suphates), there we can see they have the same distribution but for different acid ranges (citric: 0 - 1.6; volatile: 0.25 - 1.1; fixed: 4 - 14).

Which variables influence the vol. of alcohol?

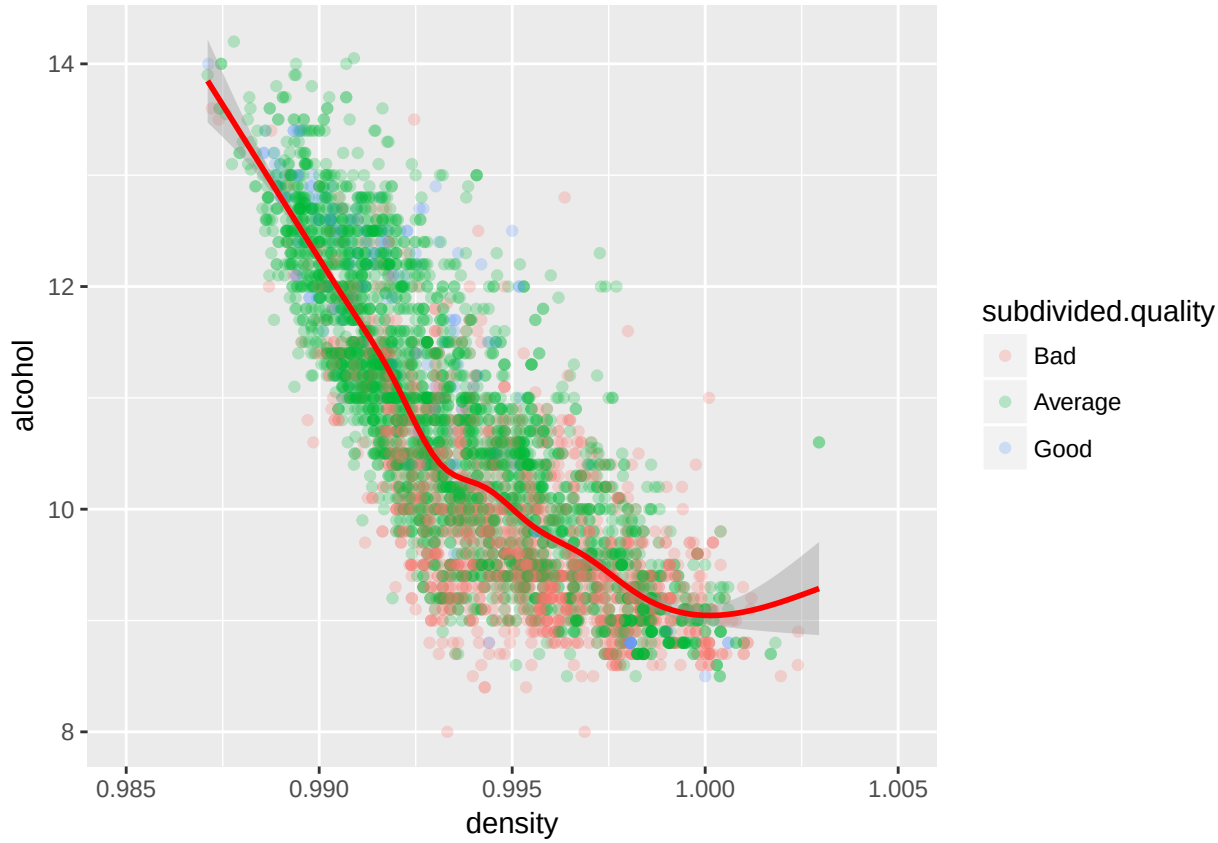
Sugar



The comparison of sugar and alcohol shows that Wine with less or average quality have a higher proportion of Sugar but a smaller volume of alcohol compared to Wines with a higher quality, they have a less proportion of sugar but more voulume of alcohol.

Density

```
ggplot(data = quality_whiteWine, aes(x = density, y = alcohol, color = subdivided.quality))+
  xlim(c(0.985, 1.005))+
  geom_point(alpha = 1/4)+
  geom_smooth(color = "red")
```

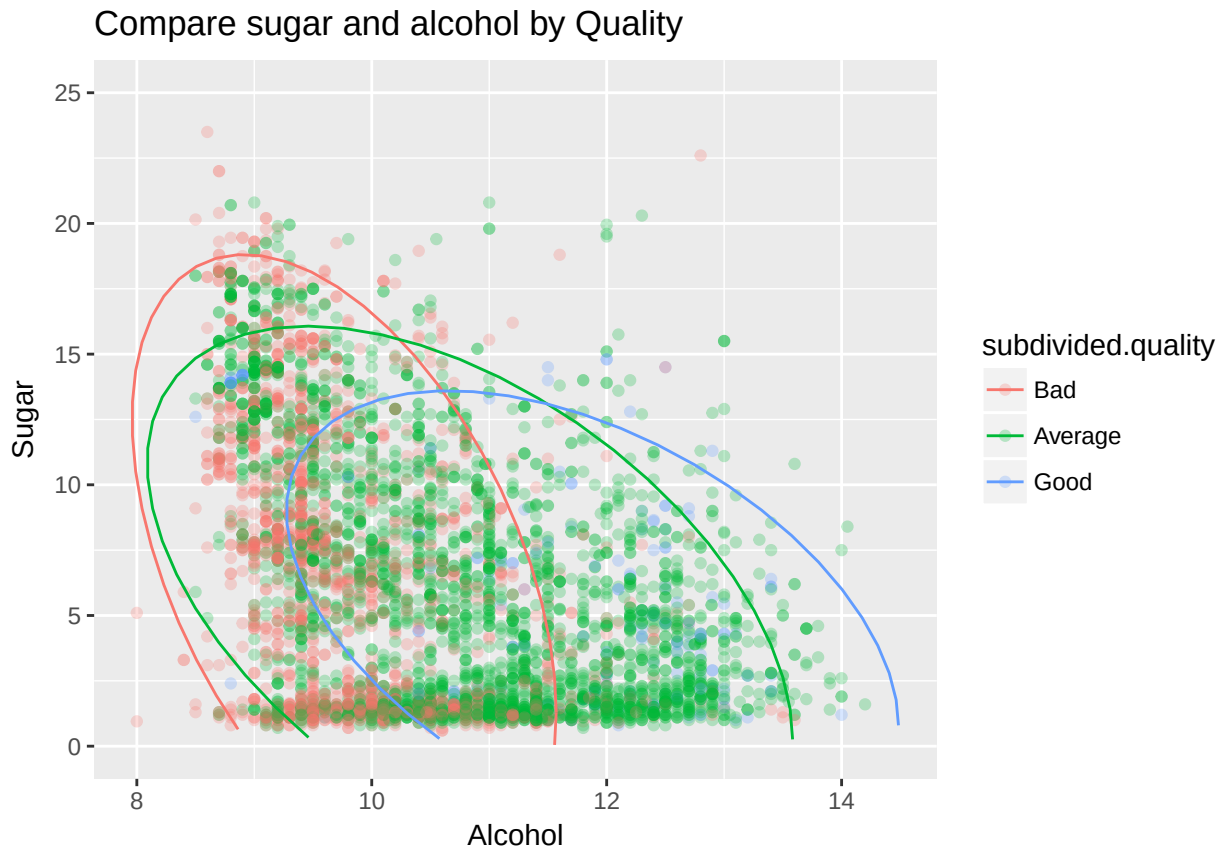


The comparison of density and alcohol shows that Wine with less or average quality have a higher density but a smaller volume of alcohol compared to Wines with a higher quality.

Comparison sugar and alcohol

The analysis of sugar by quality shows that, sugar hasn't a high influence on quality but sugar is an important variable for wine. In this part the comparison of sugar and alcohol will be analyzed.

```
ggplot(quality_whiteWine, aes(y = residual.sugar, x = alcohol, color = subdivided.quality)) +
  geom_point(alpha = 1/4)+
  stat_ellipse()+
  ylim(c(0,25))+
  ggtitle("Compare sugar and alcohol by Quality")+
  xlab("Alcohol")+
  ylab("Sugar")
```



I created a scatter plot with the variables sugar and alcohol. They are clustered in the three quality ranges. Bad wine is red, average wine is green and good wine is blue.
The scatter plot as his own makes it difficult to analyze how the both variables influence each other by quality, because there are not exists so much good wines as bad or average wines.
To get a better overview I addes ellipses to the plot. They makes it easier to analyse it. It is visible that bad alcohol has more sugar and less alcohol compared to average or good wine. Good wines hasn't a lot of sugar but a higher volume of alcohol.
The ellipses shows also how wide a range is for sugar / alcohol per quality. Bad wines has a smaller range of alcohol, average wines has the biggest range.

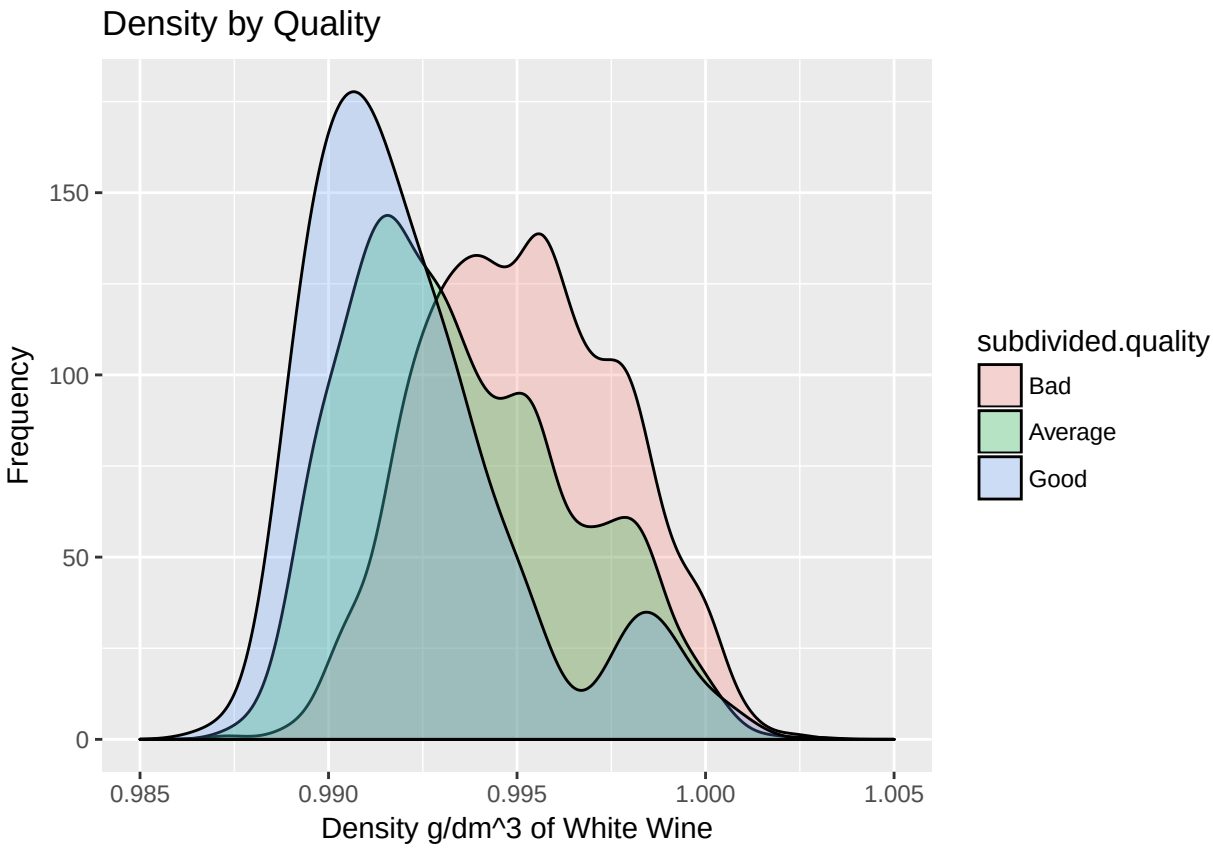
Multivariate Analysis

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 Plot Density Alcohol shows that sugar and density influence the volume of Alcohol within wines. These are variables that influence the quality of a wine.

Final Plots and Summary

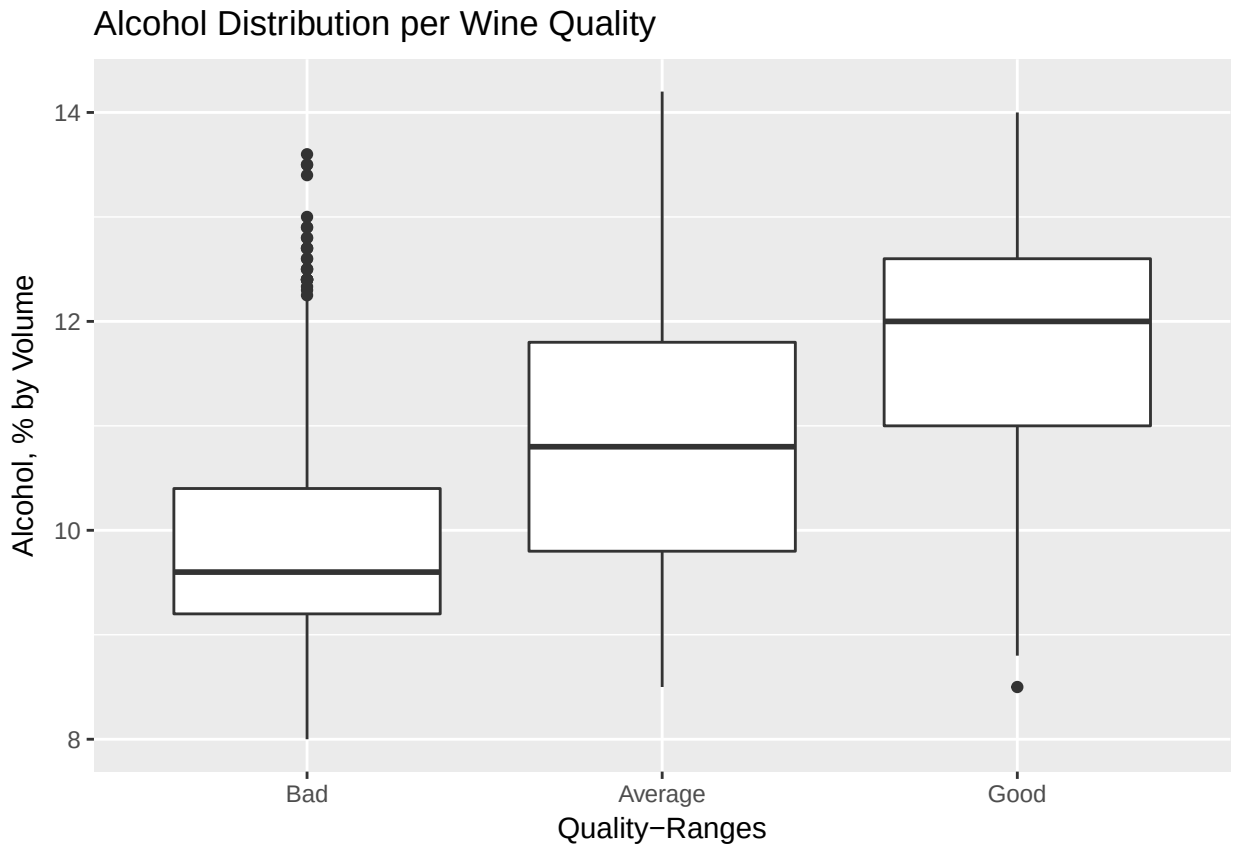
Plot One



Description One

The Density by Quality Plot shows in a very nice way how the quality is influened by the variable density. If the density is low than the wine has a good quality, if the density is higher the quality of a wine gets worst.

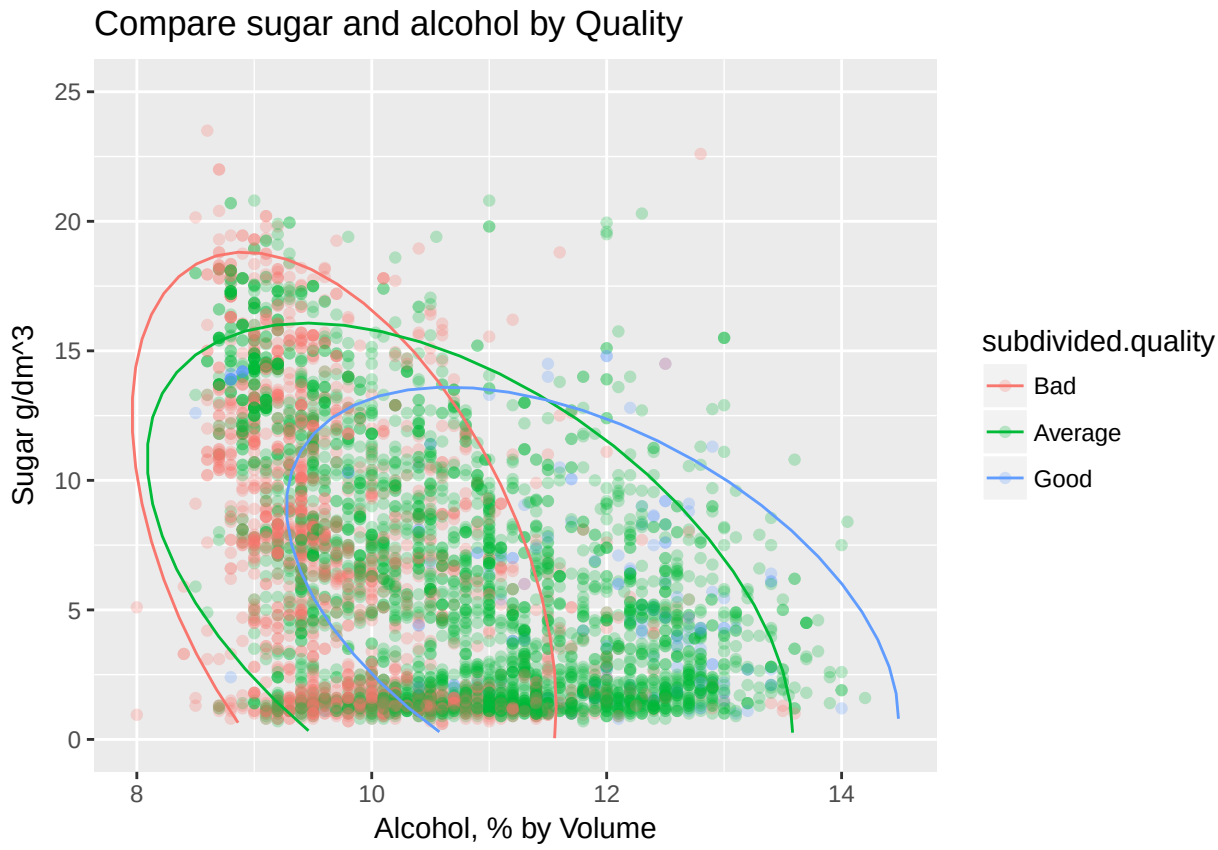
Plot Two



Description Two

This boxplot shows how the variable alcohol influence the quality. Wines with a worst quality have less alcohol than wines with a average or good quality.
The best wines has the most volume of alcohol.

Plot Three



Description Three

I choose this Plot because it shows the opposite of my assumption. Before analysing this data set I thought that a lot of sugar generate a higher volume of alcohol, but in case of this analized data set it shows the opposite. Wines with a lot of sugar has a lower volume of alcohol than wines with less sugar. Also is here very nice visible that high sugar is not an important variable of good wine.

Reflection

The Analyze shows that white wine with *higher alcohol has a better quality.* Also surprising was that sugar influence the volume in the oppoiste way I thought. *high sugar can be found in wines with a low volume of alcohol.* These to results bring me to the end result that a white wine with low alcohol and high sugar has a bad quality and a white wine with a high volume of alcohol and less sugar are a better wine.

Limitations of this dataset are that it compares only around 5000 wines and not of all regions of the world. I think that region can be also an important factor

for quality because wheater can have a high influence of the growing process of grapes. Also interessting can be the grape variety or the age of an grape tree.

For future analysis it will be interessting to predict a quality of wine, by region or grape variety.