Red Wine Quality

Made by majafoi

UPDATED: 1/8/2021

MADE IN R

Let's start with the codes – loading the csv file into the R environment/project, checking missing values and attaching the dataset.

```
library (readr)
RWQ<-read.csv ("winequality-red.csv")

## Parsed with column specification:
## cols(
##   `fixed acidity` = col_double(),
##   `volatile acidity` = col_double(),
##   `citric acid` = col_double(),
##   `residual sugar` = col_double(),
##   chlorides = col_double(),
##   `free sulfur dioxide` = col_double(),
##   `total sulfur dioxide` = col_double(),
##   density = col_double(),
##   pH = col_double(),
##   sulphates = col_double(),
##   alcohol = col_double(),
##   quality = col_double()
## )

cs = complete.cases (RWQ) # since there is a possibility that this original dataset has some missing data or NAs (not availa
ble data), I will use the function complete.cases in order to remove that part of the dataset.
RWQ = RWQ [cs,]
attach (RWQ) # I will only work with a complete dataset, or dataset without missing data and NAs.

Dfrm <- as.data.frame (RWQ)

colnames (dfrm) <- c ("fixed.acidity", "volatile.acidity", "citric.acid", "residual.sugar", "chlorides", "free.sulfur.dioxide", "t
otal.sulfur.dioxide", "density", "pH", "sulphates", "alcohol", "quality")
# it is easier to work with variables without separation between the names
attach (dfrm) # this function attaches the new data table with all changes made above.
```

Hello, all!

In this R notebook, I will show you all the details about statistical analysis I performed based on a dataset called Red Wine Quality, and the dataset is found on the following link: https://archive.ics.uci.edu/ml/datasets/wine+quality (it is also on Kaggle website).

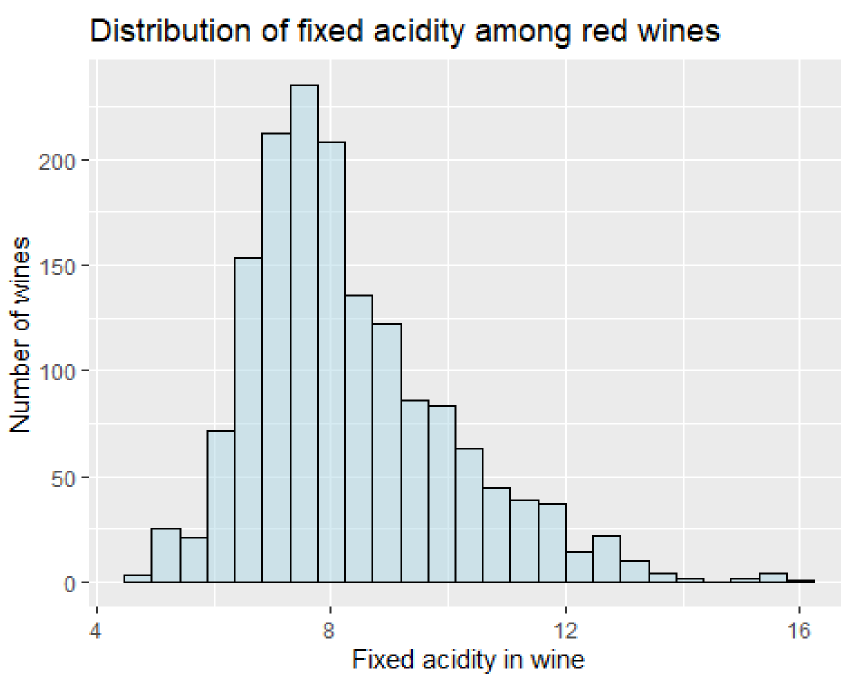
Citation info: P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

The main goal of this analysis is to perform an EDA (Exploratory Data Analysis), and to try to create a model in order to predict the Quality variable of the red wines in the sample.

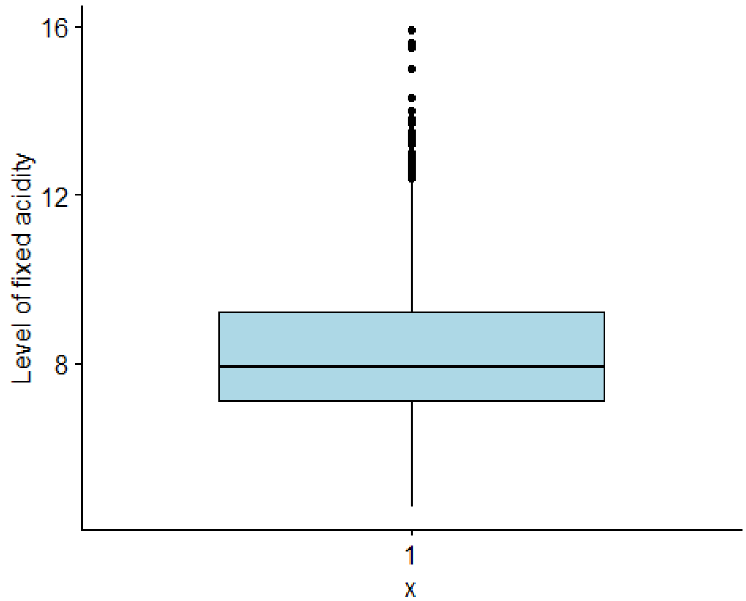
Summaries of the variables

In this chapter, I am going to create summaries of the variables in the sample. The sample itself has 1599 observations or wines, and 12 variables which have been presented and explained above.

```
summary (fixed.acidity)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.60    7.10    7.90    8.32    9.20   15.90

gf_histogram (~fixed.acidity, fill="lightblue", color="black", xlab="Fixed acidity in wine", ylab="Number of wines", title=
"Distribution of fixed acidity among red wines")
```



```r
ggboxplot (fixed.acidity, fill="lightblue", color="black", ylab="Level of fixed acidity")
```




```r
shapiro.test (fixed.acidity)
```

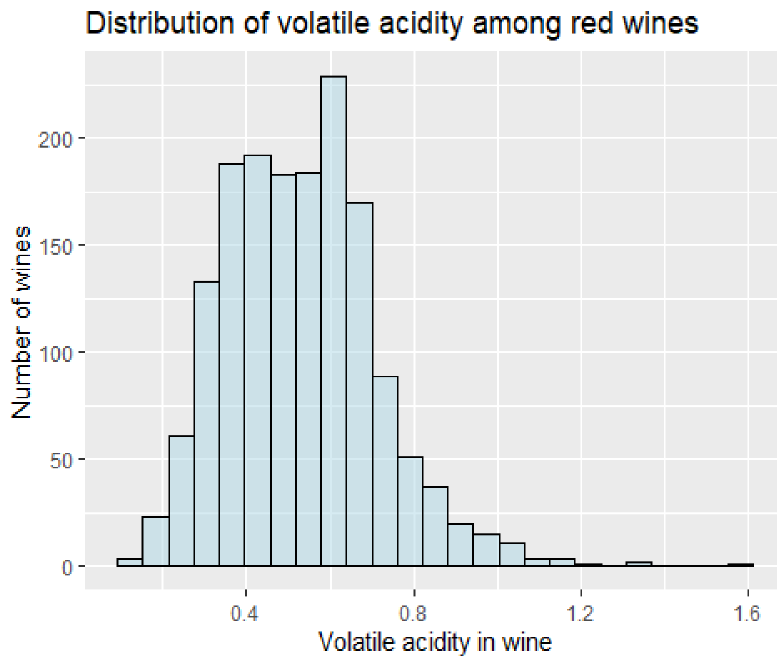
```
##
##  Shapiro-Wilk normality test
##
## data:  fixed.acidity
## W = 0.94203, p-value < 2.2e-16
```

The average fixed acidity level in the sample's red wines is 8.32, but the distribution has positive skewness, as it is last a longer tail on its right side (which means that there are wines in the sample with fixed acidity level greater than 12). Those are being considered as outliers (different than others), which are also shown as black dots on boxplot. Right away, I performed Shapiro test which shows that this variable distribution hasn't come from a normal population distribution, or that the population itself has outliers too. Shapiro test is a statistical test which is used to determine whether a data sample is from a normally distributed population, as that is a prerequisite for many statistical tests. If a p value is lower than 0.05 that indicates that the population is not normally distributed, whereas a p value greater than 0.05 provides no such evidence.

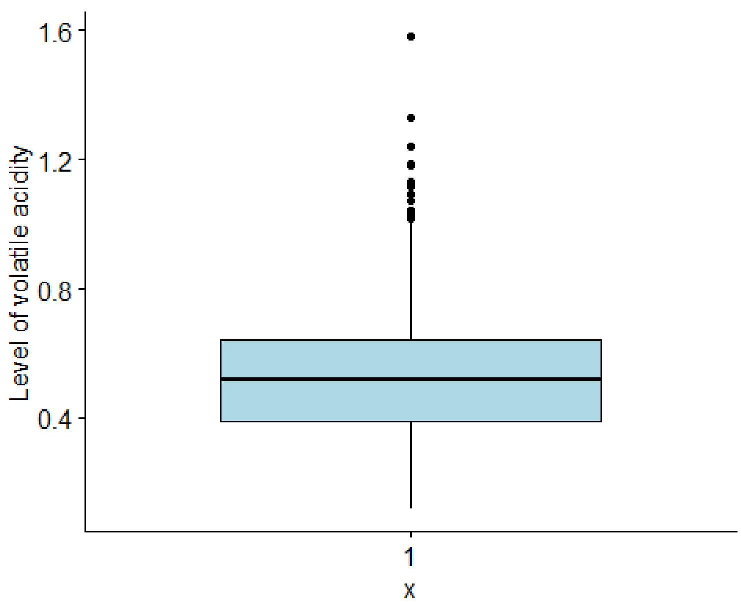
```r
summary (volatile.acidity)
```

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

```r
gf_histogram (~volatile.acidity, fill="lightblue", color="black", xlab="Volatile acidity in wine", ylab="Number of wines", t
itle="Distribution of volatile acidity among red wines")
```




```
ggboxplot (volatile.acidity, fill="lightblue", color="black", ylab="Level of volatile acidity")
```

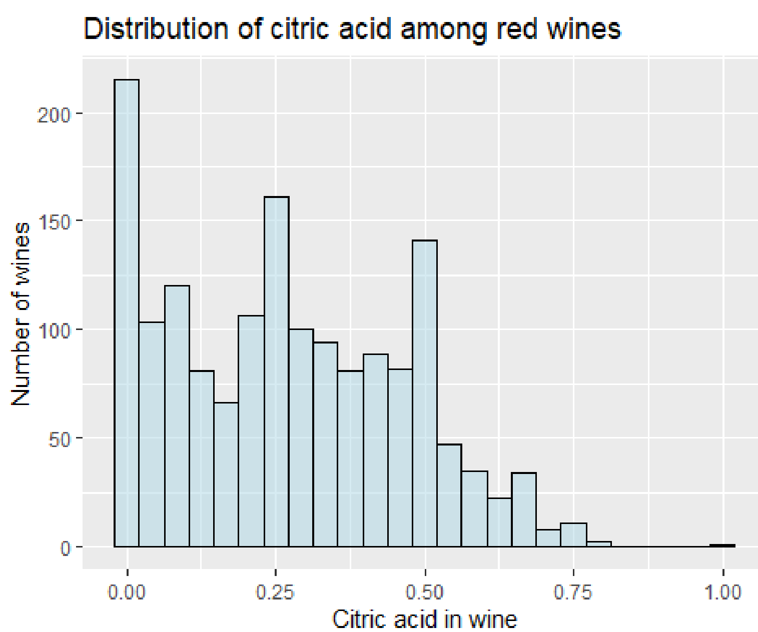


```
shapiro.test (volatile.acidity)

##
##  Shapiro-Wilk normality test
##
## data:  volatile.acidity
## W = 0.97434, p-value = 2.693e-16
```

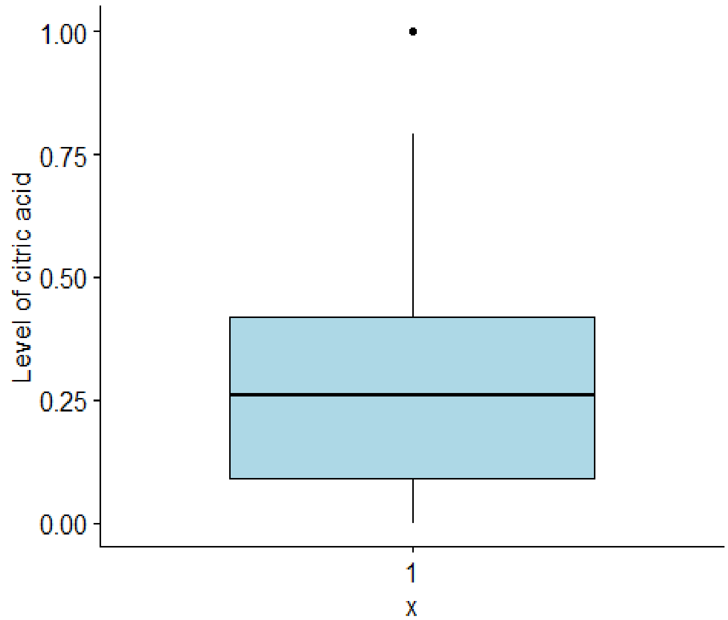
Volatile acidity is the amount of acetic acid in wine which, at too high of levels, can lead to an unpleasant and vinegar taste. The average volatile acidity of red wines in the sample is 0.52, but there is a big difference between minimum and maximum value (points that some wines have too high level of acetic acid in it). At histogram, we can see that the distribution has positive skewness and outliers on boxplot (points on the same things and conclusion as for the first variable). Also, one outlier on the boxplot is away from other outliers (point at 1.6). Again, Shapiro test shows that the distribution of this variable has come from a population which also doesn't have a normal distribution.

```
summary (citric.acid)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.090   0.260   0.271   0.420   1.000
```

```
gf_histogram (~citric.acid, fill ="lightblue", color ="black", xlab ="Citric acid in wine", ylab ="Number of wines", title ="D
istribution of citric acid among red wines")
```

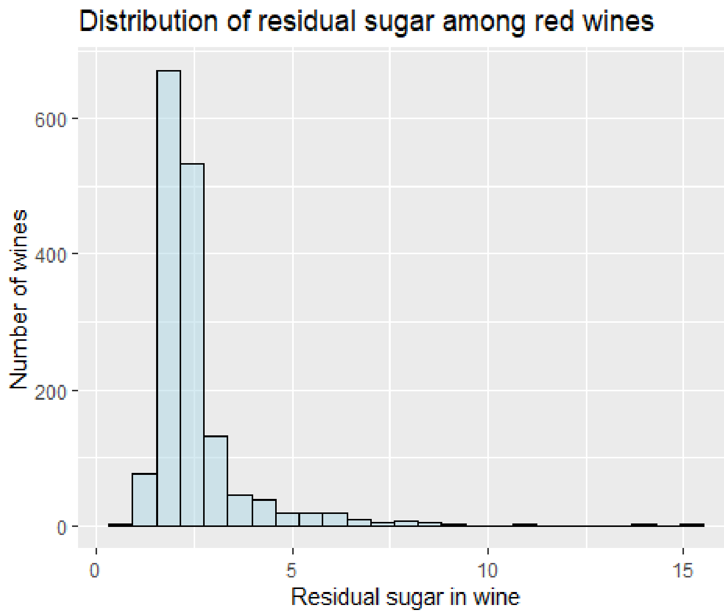


```
ggboxplot (citric.acid, fill="lightblue", color="black", ylab="Level of citric acid")
```

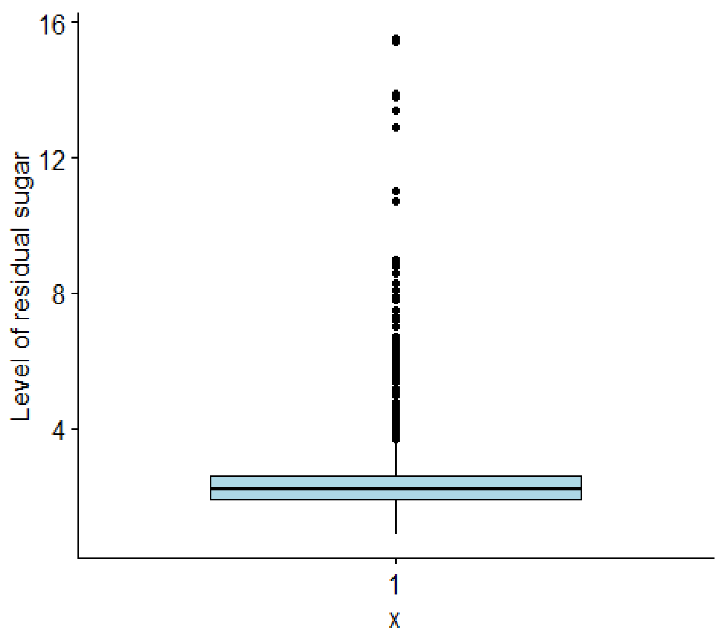


Citric acid is found in small quantities in wines, adding “freshness” and flavor to wines. So, we can obviously see that the levels of this acid are much smaller than the amounts of acetic acid. Most red wines in the sample have a level of citric acid between 0 and 0.5, with the average value of 0.27. On the boxplot it is visible that only one red wine is considered as outlier, with the level of citric acid around 1. Again, Shapiro test shows that the distribution of this variable has come from a population which also doesn’t have a normal distribution.

```
summary (residual.sugar)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.900   1.900   2.200   2.539   2.600  15.500

sd (residual.sugar)

## [1] 1.409928

gf_histogram (~residual.sugar, fill ="lightblue", color ="black", xlab ="Residual sugar in wine", ylab ="Number of wines",
title="Distribution of residual sugar among red wines")
```




```
ggboxplot (residual.sugar, fill="lightblue", color="black", ylab="Level of residual sugar")
```



```
shapiro.test (residual.sugar)
```

```
## 
## Shapiro-Wilk normality test
## 
## data:  residual.sugar
## W = 0.56608, p-value < 2.2e-16
```

Residual sugar is the amount of sugar remaining after fermentation stops, it is very rare to find wines with less than 1g/L and wines with greater than 45g/L (those are considered sweet). In this sample's distribution, we can see that the minimum value is 0.9, which makes that wine – rare, when speaking of residual sugar, but the maximum value is below 45g/L - only 15.5. The average amount of residual sugar in these red wines is 2.53, with standard deviation of 1.4, which means that outliers haven't had much influence on the distribution. Looking at the histogram distribution, that is not expected because we can see a huge positive skewness. But, as we can see on boxplot, there are a lot of red wine's residual sugar which are outliers. Again, Shapiro test shows that the distribution of this variable has come from a population which also doesn't have a normal distribution.

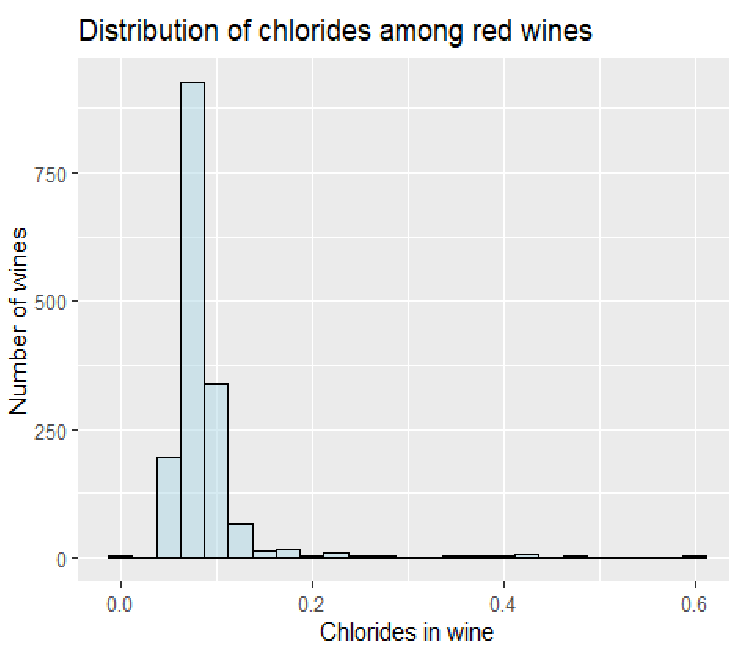
```
summary (chlorides)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
## 0.01200 0.07000 0.07900 0.08747 0.09000 0.61100
```

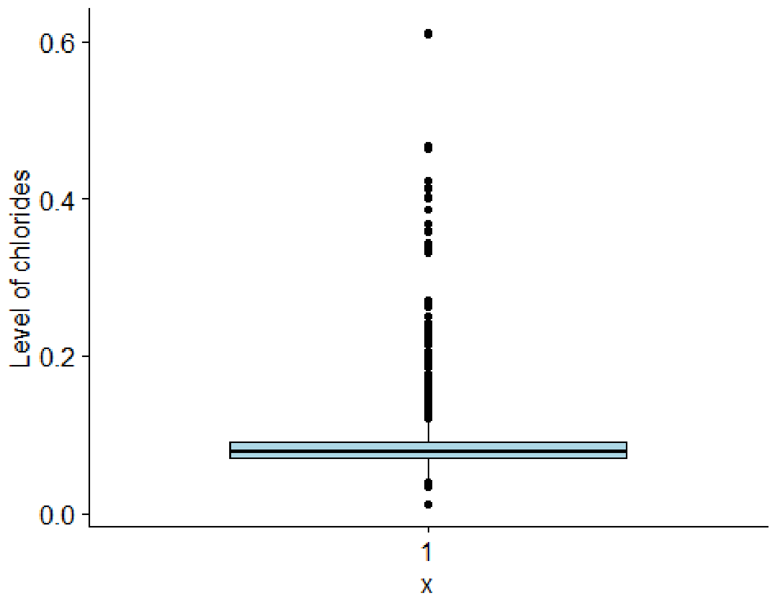
```
sd (chlorides)
```

```
## [1] 0.0470653
```

```
gf_histogram (~chlorides, fill="lightblue", color="black", xlab="Chlorides in wine", ylab="Number of wines", title="Distribution of chlorides among red wines")
```



```
ggboxplot (chlorides, fill="lightblue", color="black", ylab="Level of chlorides")
```

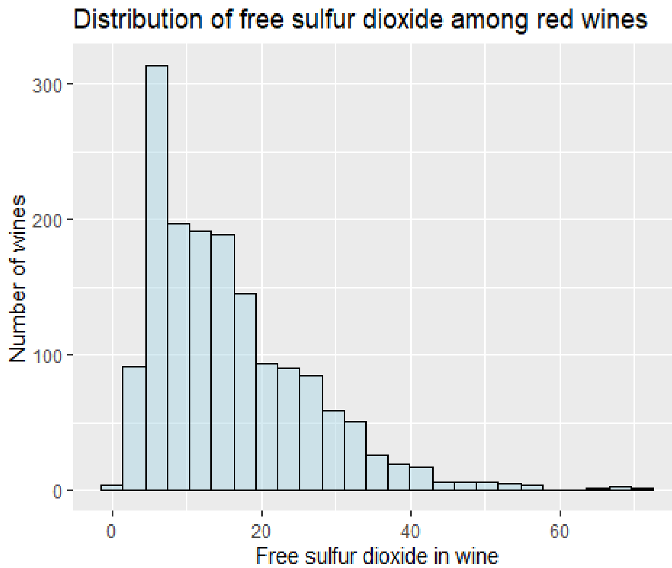



```
shapiro.test(chlorides)

##
## 	Shapiro-Wilk normality test
##
## data:  chlorides
## W = 0.48425, p-value < 2.2e-16
```

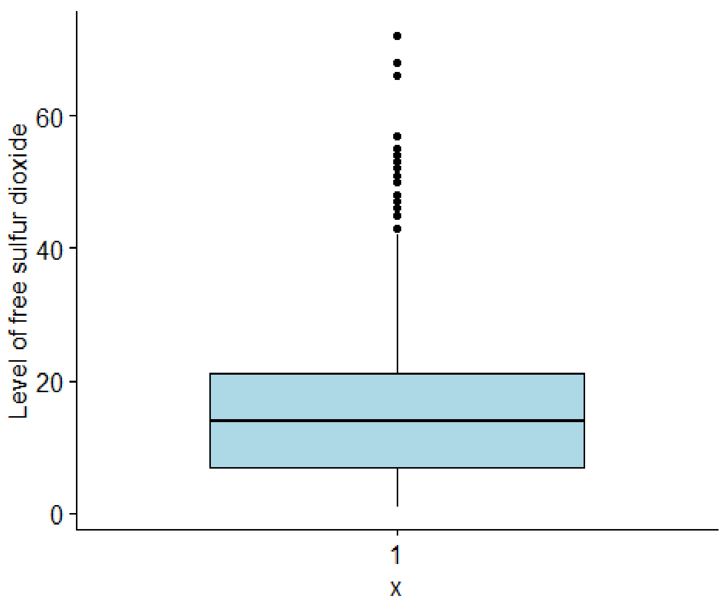
Chlorides represent the amount of salt in the wine. The average amount of chlorides in these red wines is 0.08, but again - the minimum and maximum value differentiate largely. This time, we also have a positive skewness of the distribution, but also outliers on both sides of the medium and average value. Again, Shapiro test shows that the distribution of this variable has come from a population which also doesn't have a normal distribution.

```
summary (free.sulfur.dioxide)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    7.00   14.00   15.87   21.00   72.00
```

```
gf_histogram (~free.sulfur.dioxide, fill="lightblue", color="black", xlab="Free sulfur dioxide in wine", ylab="Number of wines", title="Distribution of free sulfur dioxide among red wines")
```



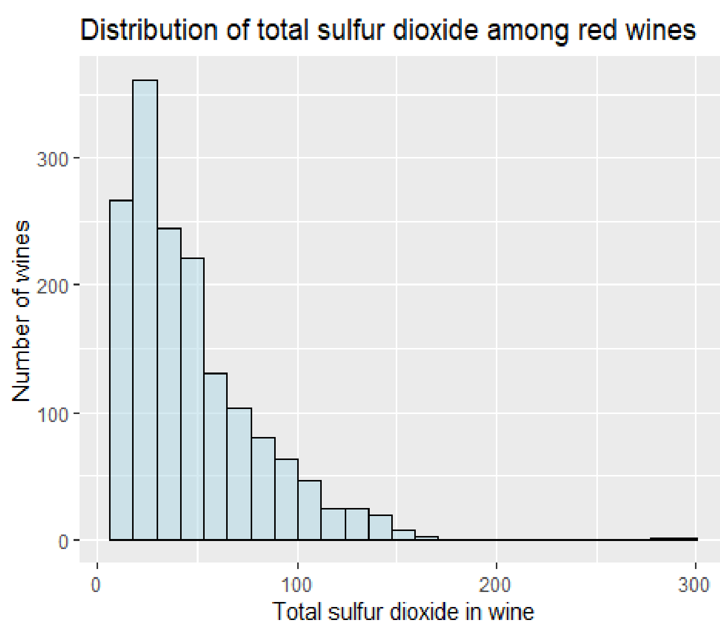

```
ggboxplot (free.sulfur.dioxide, fill="lightblue", color="black", ylab="Level of free sulfur dioxide")
```



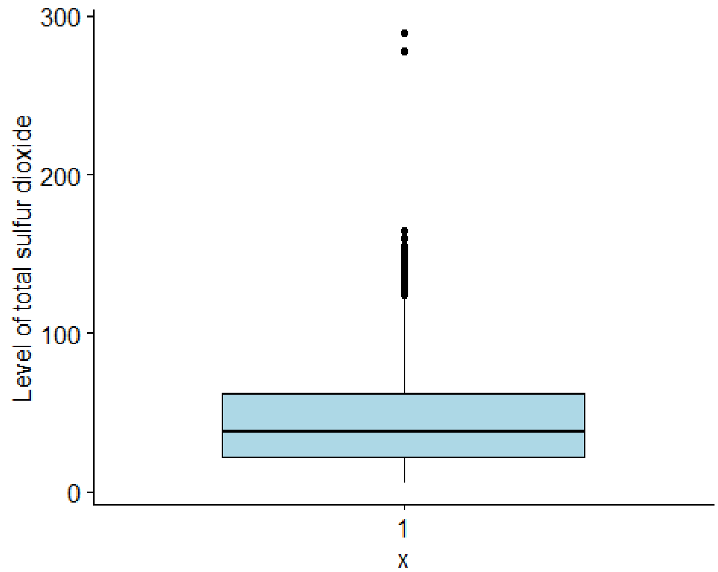
```
shapiro.test (free.sulfur.dioxide)

##
##  Shapiro-Wilk normality test
##
## data:  free.sulfur.dioxide
## W = 0.90184, p-value < 2.2e-16
```

Free sulfur dioxide is the free form of SO2 which exists in the equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion, it also prevents microbial grown and the oxidation of the wine. The minimum and maximum value differentiate highly here too, whereas the average value of free sulfur dioxide is 15.87. This distribution has a hard positive skewness, with outliers of some wines above 40 amount of free sulfur dioxide. Again, Shapiro test shows that the distribution of this variable has come from a population which also doesn't have a normal distribution.

```
summary (total.sulfur.dioxide)

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

sd (total.sulfur.dioxide)

## [1] 32.89532

gf_histogram (~total.sulfur.dioxide, fill="lightblue", color="black", xlab="Total sulfur dioxide in wine", ylab="Number of wines", title="Distribution of total sulfur dioxide among red wines")
```



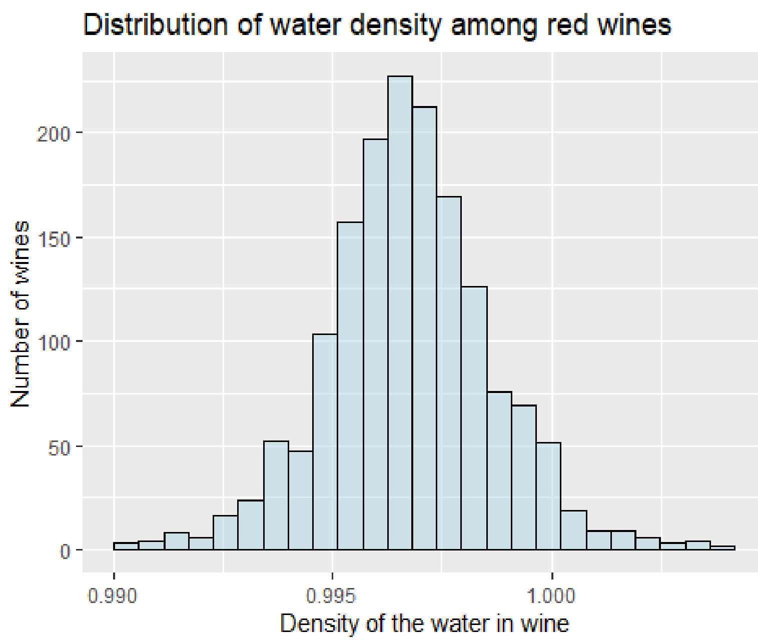
```
ggboxplot (total.sulfur.dioxide, fill="lightblue", color="black", ylab="Level of total sulfur dioxide")
```



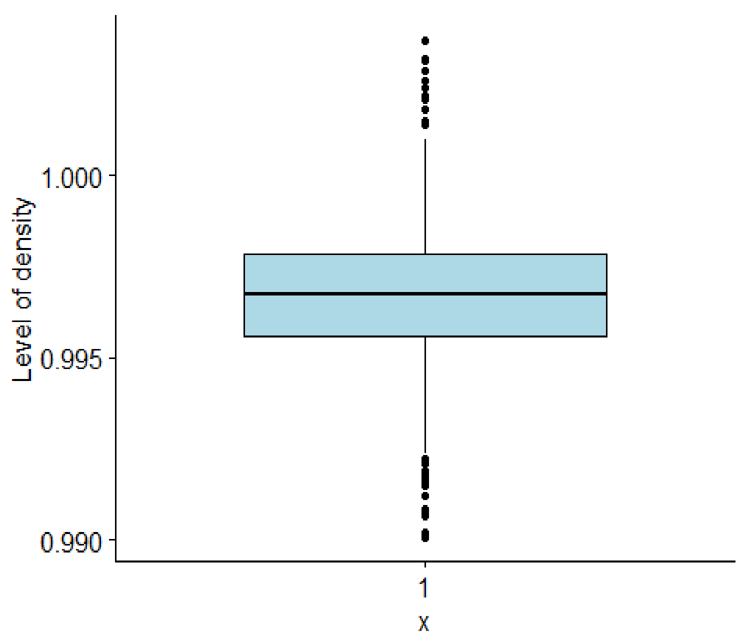

```
shapiro.test (total.sulfur.dioxide)

##
##  Shapiro-Wilk normality test
##
## data:  total.sulfur.dioxide
## W = 0.87322, p-value < 2.2e-16
```

Total sulfur dioxide is the amount of free and bound forms of SO_2. In low concentrations, SO_2 is mostly undetectable in wine, but at free SO_2 concentrations over 50 ppm, SO_2 becomes evident in the nose of people and taste of wine. Here, some wines have that concentration above 50 ppm, even the average value is pretty close. Maximum value is astonishing 289 ppm! This distribution has a big positive skewness, with outliers with values around 120 ppm, but there are some extreme outliers around 300 ppm. We should take care of this variable's distribution, as it has the biggest outliers so far, and that can distort future statistical tests or make those tests insignificant. Again, Shapiro test shows that the distribution of this variable has come from a population which also doesn't have a normal distribution.

```
summary(density)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.9901  0.9956  0.9968  0.9967  0.9978  1.0037

gf_histogram(~density, fill="lightblue", color="black", xlab="Density of the water in wine", ylab="Number of wines", titl
e="Distribution of water density among red wines")
```



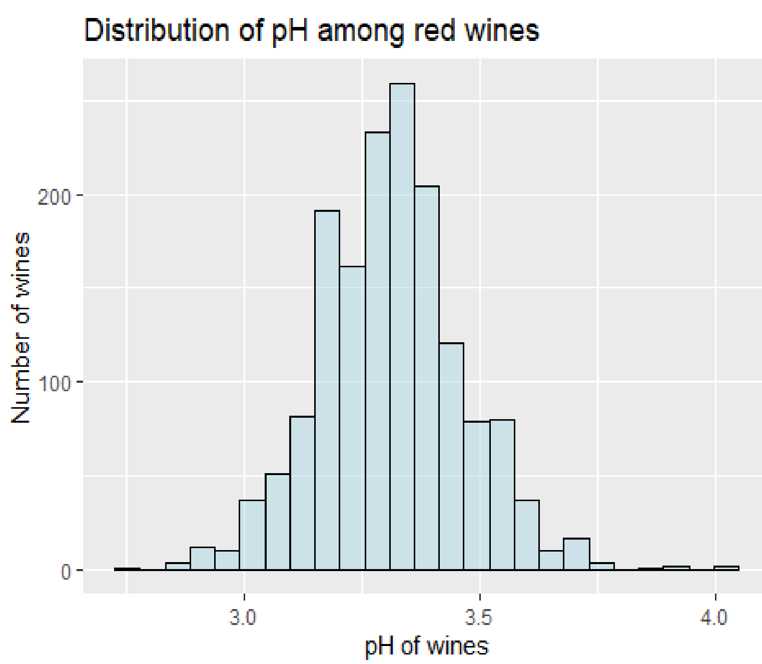

```
ggboxplot(density, fill="lightblue", color="black", ylab="Level of density")
```



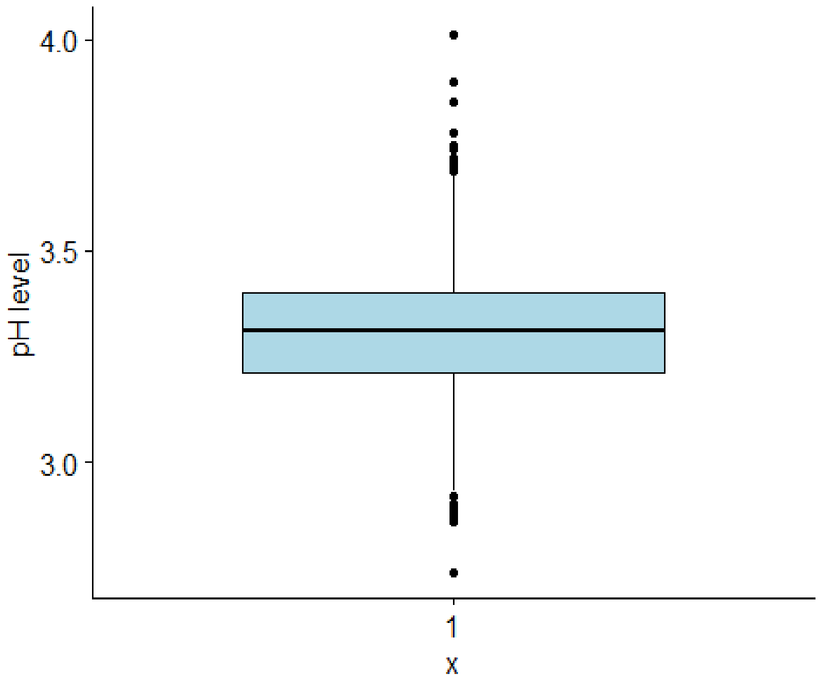
```
shapiro.test (density)

##
##  Shapiro-Wilk normality test
##
## data:  density
## W = 0.99087, p-value = 1.936e-08
```

Density represents density of water, which is close to that of the regular water itself, depending on the percent alcohol and sugar content. The mean and the median almost are the same, and the histogram is showing normal distribution. The boxplot shows that there are some outliers on both sides, but those aren't outliers visible on the histogram, and the summary proves it. Again, Shapiro test shows that the distribution of this variable has come from a population which also doesn't have a normal distribution, whereas we can say that this variable' distribution in the sample is normal. But this difference between distributions in the sample and population can happen when you are extracting numerous samples from one big population. Some samples tend to have many outliers, whereas some samples are normally distributed.

```
summary (pH)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.740   3.210   3.310   3.311   3.400   4.010

gf_histogram (~pH, fill="lightblue", color="black", xlab="pH of wines", ylab="Number of wines", title="Distribution of p
H among red wines", xlim=c(40,95), ylim=c(0,50))
```



```r
ggboxplot(pH, fill="lightblue", color="black", ylab="pH level")
```



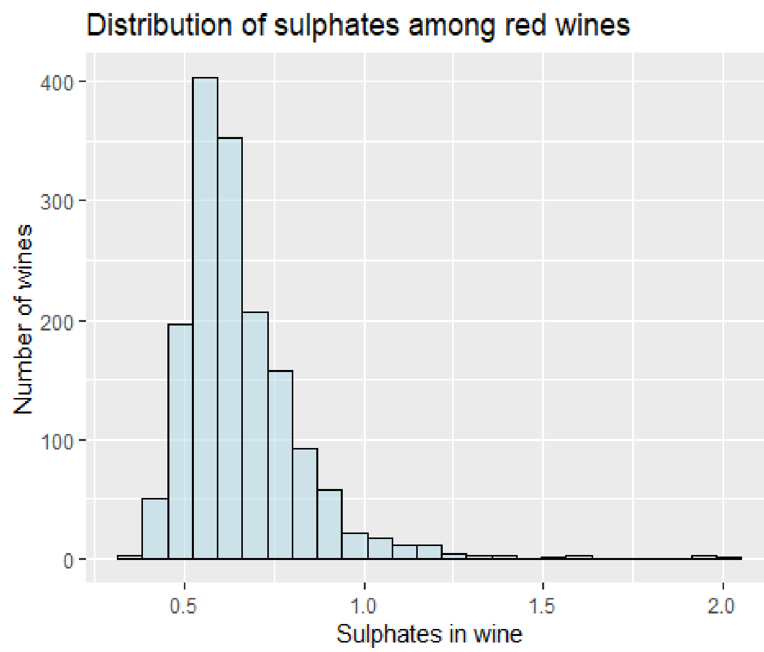

```r
shapiro.test(pH)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  pH
## W = 0.99349, p-value = 1.712e-06
```

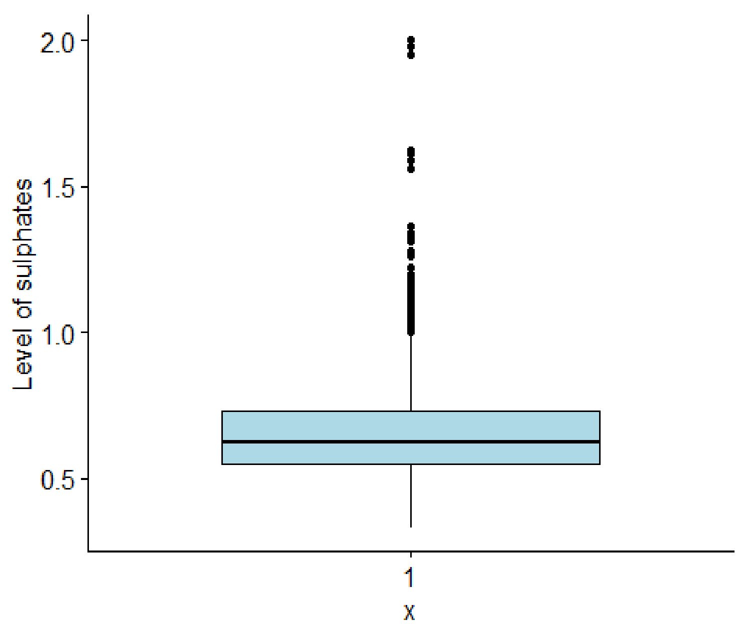
pH describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic), and most wines are between 3-4 on the pH scale. That is also visible here on the histogram, which almost has a normal looking distribution too, as a variable density before. The mean and median do differentiate, but just slightly. Boxplot does show that there are outliers on both sides of the distribution, but the histogram and the summary shows us that those aren't real outliers. Again, Shapiro test shows that the distribution of this variable has come from a population which also doesn't have a normal distribution.

```
summary (sulphates)

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

```
gf_histogram (~sulphates, fill="lightblue", color="black", xlab="Sulphates in wine", ylab="Number of wines", title="Distribution of sulphates among red wines", xlim=c(40,95), ylim=c(0,50))
```




```
ggboxplot (sulphates, fill="lightblue", color="black", ylab="Level of sulphates")
```

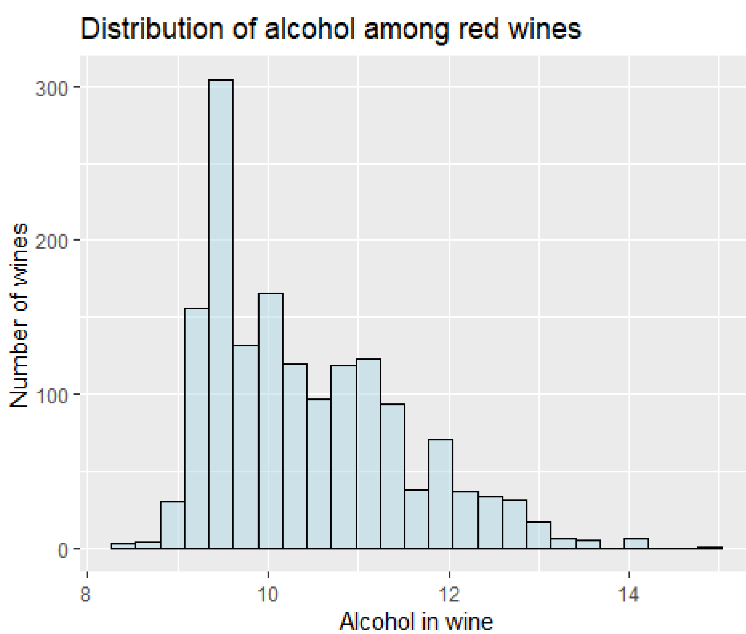


```
shapiro.test (sulphates)

##
##  Shapiro-Wilk normality test
##
## data:  sulphates
## W = 0.83304, p-value < 2.2e-16
```

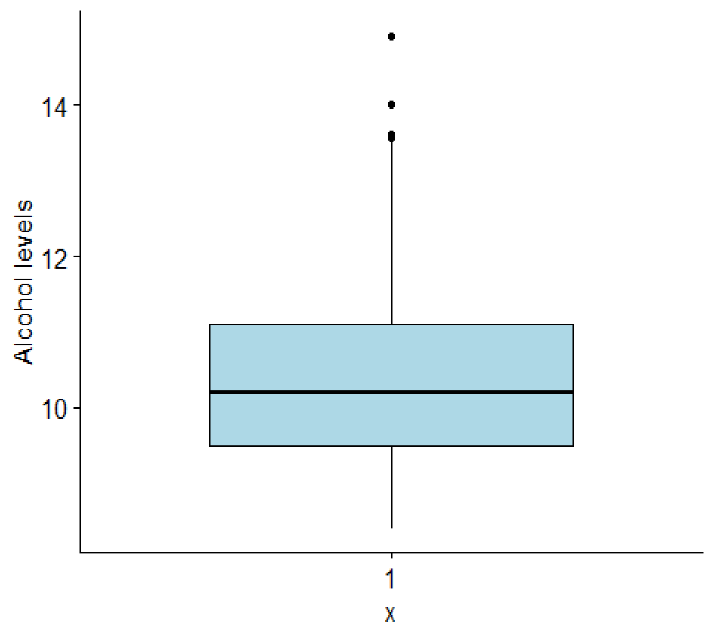
Sulphates are a wine additive which can contribute to sulfur dioxide gas (SO_2) levels, which acts as an antimicrobial and antioxidant. Mean and the median are close together, but not enough to say it is a normal distribution of the variable like for the last two variables. Also, there is a big difference between minimum and maximum value and if we look at the histogram, we can see that the distribution has a negative skewness. On boxplot, we can also see that there are outliers from level 1 and beyond. Again, Shapiro test shows that the distribution of this variable has come from a population which also doesn't have a normal distribution.

```
summary (alcohol)

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

```
gf_histogram (~alcohol, fill="lightblue", color="black", xlab="Alcohol in wine", ylab="Number of wines", title="Distribution of alcohol among red wines", xlim=c(40,95), ylim=c(0,50))
```



```
ggboxplot(alcohol,fill="lightblue",color="black",ylab="Alcohol levels")
```

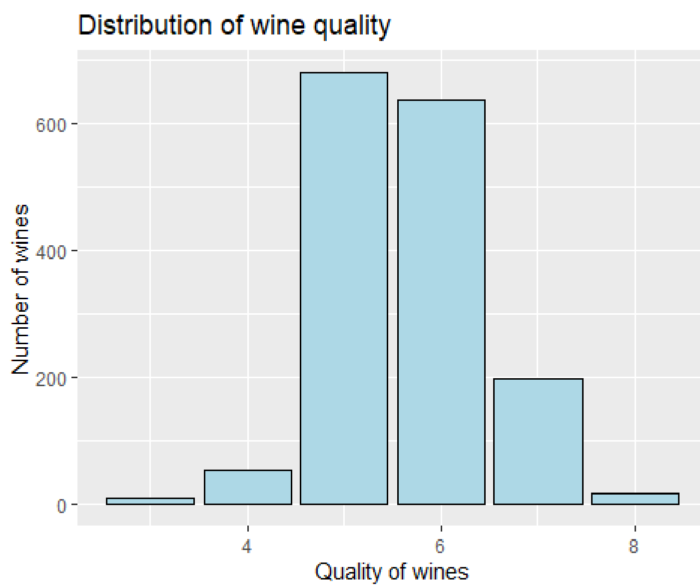



```
shapiro.test (alcohol)

##
## Shapiro-Wilk normality test
##
## data: alcohol
## W = 0.92884, p-value < 2.2e-16
```

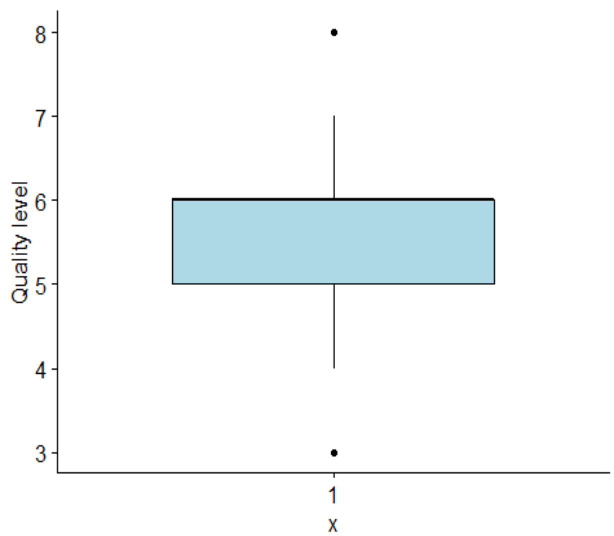
Alcohol variable represents the percent alcohol of the wine. The average percent of alcohol in the wines in the sample is 10.42, but the median isn't far from the average value/mean. But, the difference between minimum and maximum value is high, which means that there is some (positive) skewness visible on histogram. The boxplot shows that there are some outliers after a value of approximate 13.5%. Again, Shapiro test shows that the distribution of this variable has come from a population which also doesn't have a normal distribution.

```
summary (quality)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   5.000   6.000   5.636   6.000   8.000
```

```
gf_bar (~quality, fill="lightblue", color="black", xlab="Quality of wines", ylab="Number of wines", title="Distribution of wine quality", xlim=c(40,95), ylim=c(0,50))
```




```
ggboxplot (quality, fill="lightblue", color="black", ylab="Quality level")
```



The quality variable is an output variable, based on sensory data, and it scores between 0 and 10. Unfortunately, the average quality grade is 5.63. The difference is existent between minimum and maximum value, but the histogram seems to show pretty normal distribution. Boxplot shows outliers on both sides, which, again, shouldn't be real outliers.

With this, I have finished summaries of all variables in the sample. Since we want to predict the quality variable, this is going to be our y-variable or dependent variable, whereas all other variables will be independent variables (x-variables). More info on that in the next chapter.