# Sai Lakshman Rao Duggirala – M15848995

#### Wine Project Part A

**library**(tidyverse)

wine<-read.csv(file="D:/Applied Stat methods/winequality-red.csv", header=T, sep = ";")

# Question a

1. What is the sample size?

We can determine sample size of the data by running count and str function. Therefore, Sample size n = 1599.

count(wine)

## n

## 1 1599

str(wine)

## 'data.frame': 1599 obs. of 12 variables:

## $ fixed.acidity : num 7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...

## $ volatile.acidity : num 0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...

## $ citric.acid : num 0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...

## $ residual.sugar : num 1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...

## $ chlorides : num 0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071 ...

## $ free.sulfur.dioxide : num 11 25 15 17 11 13 15 15 9 17 ...

## $ total.sulfur.dioxide: num 34 67 54 60 34 40 59 21 18 102 ...

## $ density : num 0.998 0.997 0.997 0.998 0.998 ...

## $ pH : num 3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...

## $ sulphates : num 0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...

## $ alcohol : num 9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...

## $ quality : int 5 5 5 6 5 5 5 7 7 5 ...

# Question b

1. Any outliers? Do you have any concerns about the data quality

By looking at the summary of data, we identify the Mean and Median values of all the data fields. From this, we can identify the difference between Mean and Median values in total.sulfur.dioxide has the highest difference among other data fields. There are no NA values, no blank/empty values. We can identify few data points with the value 0, which do not affect the quality of data.

summary(wine)

## fixed.acidity volatile.acidity citric.acid residual.sugar

## Min. : 4.60 Min. :0.1200 Min. :0.000 Min. : 0.900

## 1st Qu.: 7.10 1st Qu.:0.3900 1st Qu.:0.090 1st Qu.: 1.900

## Median : 7.90 Median :0.5200 Median :0.260 Median : 2.200

## Mean : 8.32 Mean :0.5278 Mean :0.271 Mean : 2.539

## 3rd Qu.: 9.20 3rd Qu.:0.6400 3rd Qu.:0.420 3rd Qu.: 2.600

## Max. :15.90 Max. :1.5800 Max. :1.000 Max. :15.500

## chlorides free.sulfur.dioxide total.sulfur.dioxide density

## Min. :0.01200 Min. : 1.00 Min. : 6.00 Min. :0.9901

## 1st Qu.:0.07000 1st Qu.: 7.00 1st Qu.: 22.00 1st Qu.:0.9956

## Median :0.07900 Median :14.00 Median : 38.00 Median :0.9968

## Mean :0.08747 Mean :15.87 Mean : 46.47 Mean :0.9967

## 3rd Qu.:0.09000 3rd Qu.:21.00 3rd Qu.: 62.00 3rd Qu.:0.9978

## Max. :0.61100 Max. :72.00 Max. :289.00 Max. :1.0037

## pH sulphates alcohol quality

## Min. :2.740 Min. :0.3300 Min. : 8.40 Min. :3.000

## 1st Qu.:3.210 1st Qu.:0.5500 1st Qu.: 9.50 1st Qu.:5.000

## Median :3.310 Median :0.6200 Median :10.20 Median :6.000

## Mean :3.311 Mean :0.6581 Mean :10.42 Mean :5.636

## 3rd Qu.:3.400 3rd Qu.:0.7300 3rd Qu.:11.10 3rd Qu.:6.000

## Max. :4.010 Max. :2.0000 Max. :14.90 Max. :8.000

# Question c

1. How can you summarize the data of each variable in a concise way? What statistics are you going to present?

We can summarize the data by using the summary() function for each field in the data frame.

**attach**(wine)

summary(fixed.acidity)

## Min. 1st Qu. Median Mean 3rd Qu. Max.

## 4.60 7.10 7.90 8.32 9.20 15.90

summary(volatile.acidity)

## Min. 1st Qu. Median Mean 3rd Qu. Max.

## 0.1200 0.3900 0.5200 0.5278 0.6400 1.5800

summary(citric.acid)

## Min. 1st Qu. Median Mean 3rd Qu. Max.

## 0.000 0.090 0.260 0.271 0.420 1.000

summary(residual.sugar)

## Min. 1st Qu. Median Mean 3rd Qu. Max.

## 0.900 1.900 2.200 2.539 2.600 15.500

summary(chlorides)

## Min. 1st Qu. Median Mean 3rd Qu. Max.

## 0.01200 0.07000 0.07900 0.08747 0.09000 0.61100

summary(free.sulfur.dioxide)

## Min. 1st Qu. Median Mean 3rd Qu. Max.

## 1.00 7.00 14.00 15.87 21.00 72.00

summary(total.sulfur.dioxide)

## Min. 1st Qu. Median Mean 3rd Qu. Max.

## 6.00 22.00 38.00 46.47 62.00 289.00

summary(density)

## Min. 1st Qu. Median Mean 3rd Qu. Max.

## 0.9901 0.9956 0.9968 0.9967 0.9978 1.0037

summary(pH)

## Min. 1st Qu. Median Mean 3rd Qu. Max.

## 2.740 3.210 3.310 3.311 3.400 4.010

summary(sulphates)

## Min. 1st Qu. Median Mean 3rd Qu. Max.

## 0.3300 0.5500 0.6200 0.6581 0.7300 2.0000

summary(alcohol)

## Min. 1st Qu. Median Mean 3rd Qu. Max.

## 8.40 9.50 10.20 10.42 11.10 14.90

summary(quality)

## Min. 1st Qu. Median Mean 3rd Qu. Max.

## 3.000 5.000 6.000 5.636 6.000 8.000

With the summary of each data frame, we can confirm that the difference in mean and median is highest in total.sulfur.dioxide . To check for outliers more precisely, we can use histogram and boxplot data to confirm the data is symmetry or skewed.

Variance of each data field from the dataset.

var(fixed.acidity)

## [1] 3.031416

var(volatile.acidity)

## [1] 0.03206238

var(citric.acid)

## [1] 0.03794748

var(residual.sugar)

## [1] 1.987897

var(chlorides)

## [1] 0.002215143

var(free.sulfur.dioxide)

## [1] 109.4149

var(total.sulfur.dioxide)

## [1] 1082.102

var(density)

## [1] 3.562029e-06

var(pH)

## [1] 0.02383518

var(sulphates)

## [1] 0.02873262

var(alcohol)

## [1] 1.135647

var(quality)

## [1] 0.6521684

**How can we measure wine:**

Wine can be measured with the use of few factors like taste, quality, level of alcohol, use of acid levels to generate the overall essence. From the data set we have for red-wine, we can use pH to determine the level of acidity used, residual sugar can determine the taste and alcohol can be helpful to identify the percentage of alcohol used.

1. Using pH to determine acid value. Activity of a good wine should be in the range of 2 to 3.5. Hence it is of range 89%.
2. Using residual.sugar determines the sweetness value in wine. It is of 99%.
3. With the use of alcohol field from the dataset, we identify the level of alcohol used in a wine which lies between 12% to 15%. But from our dataset, only 10% of alcohol levels are used.
4. I have calculated the mean and standard deviation in quality data field to determine the percentage that lies in -1 and +1 which is 82%, -2 and +2 is 94% and -3 and +3 is 99%. With this most of the red-wines lies between 5 and 6 which are the approximate mean and median for quality data field from the given dataset.

sum((pH>=2 & pH<=3.5)==TRUE)/count(wine)

## n

## 1 0.8949343

sum(( residual.sugar<= 10)==TRUE)/count(wine)

## n

## 1 0.9931207

sum ((alcohol >= 12 & alcohol <= 15)==TRUE)/count(wine)

## n

## 1 0.1013133

lw <- mean(quality)-1\*sd(quality) *#lower end*

up <- mean(quality)+1\*sd(quality) *#upper end*

sum((quality > lw & quality<up)==TRUE)/count(wine)

## n

## 1 0.8248906

lw <- mean(quality)-2\*sd(quality) *#lower end*

up <- mean(quality)+2\*sd(quality) *#upper end*

sum((quality > lw & quality<up)==TRUE)/count(wine) *#deleting the outliers*

## n

## 1 0.9493433

lw <- mean(quality)-3\*sd(quality) *#lower end*

up <- mean(quality)+3\*sd(quality) *#upper end*

sum((quality > lw & quality<up)==TRUE)/count(wine)

## n

## 1 0.9937461

# Question d & e.

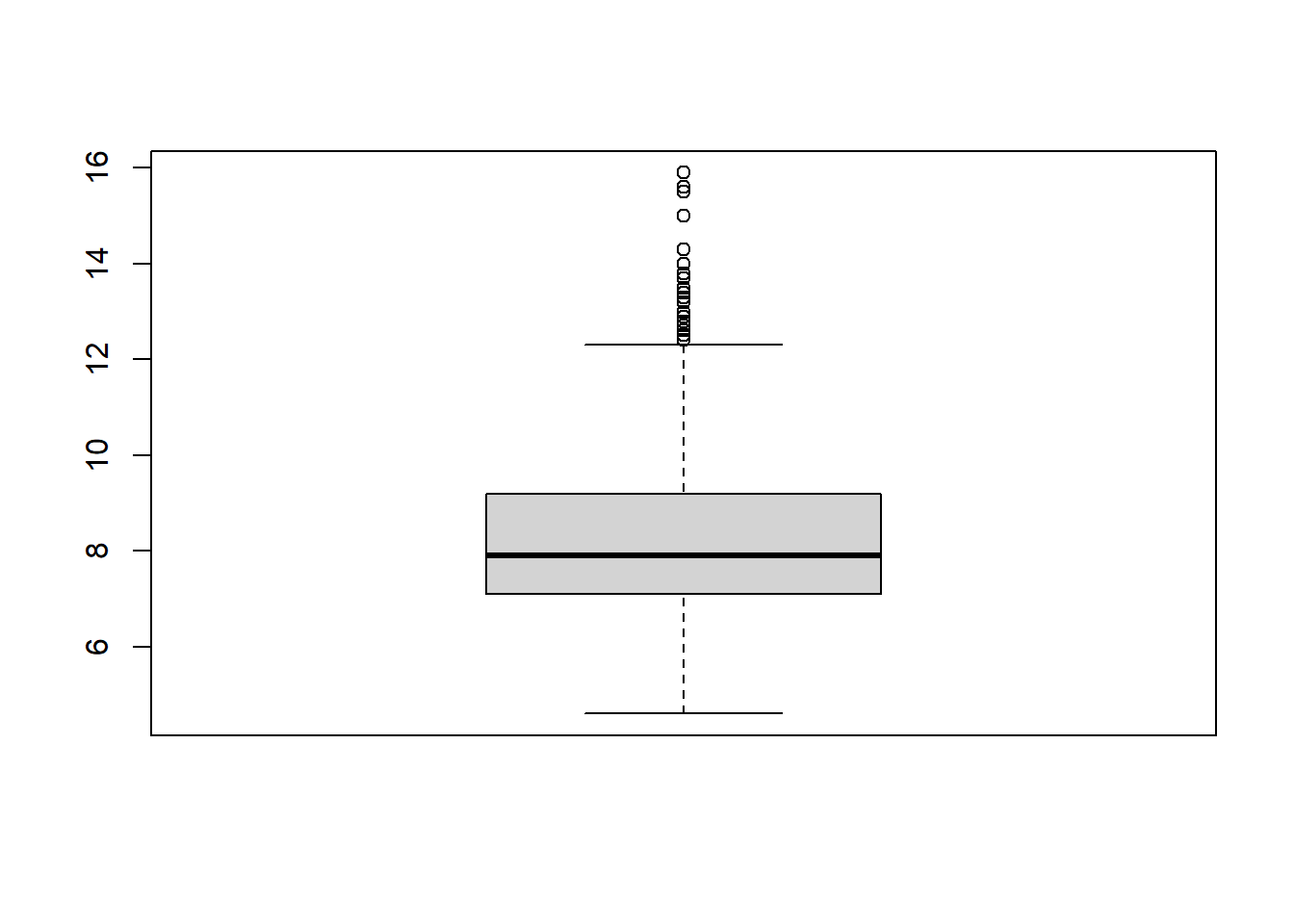
D. How can you visualize the distribution of each variable?

E. Do you see any skewed distributions?

With the use of Boxplots and Histogram plots we can identify each variable and also we can identify the overall distribution.

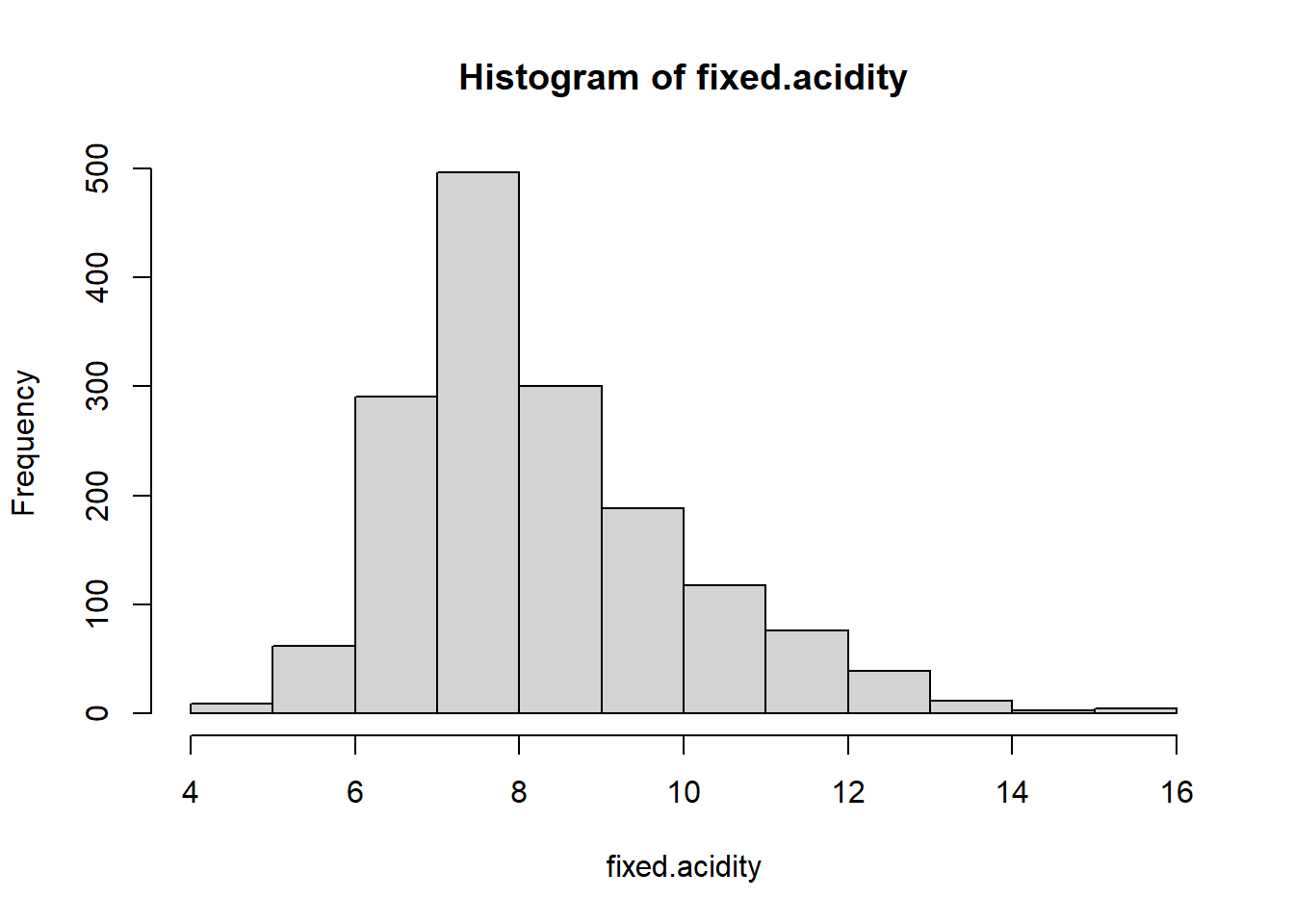
**attach**(wine)

boxplot(fixed.acidity)



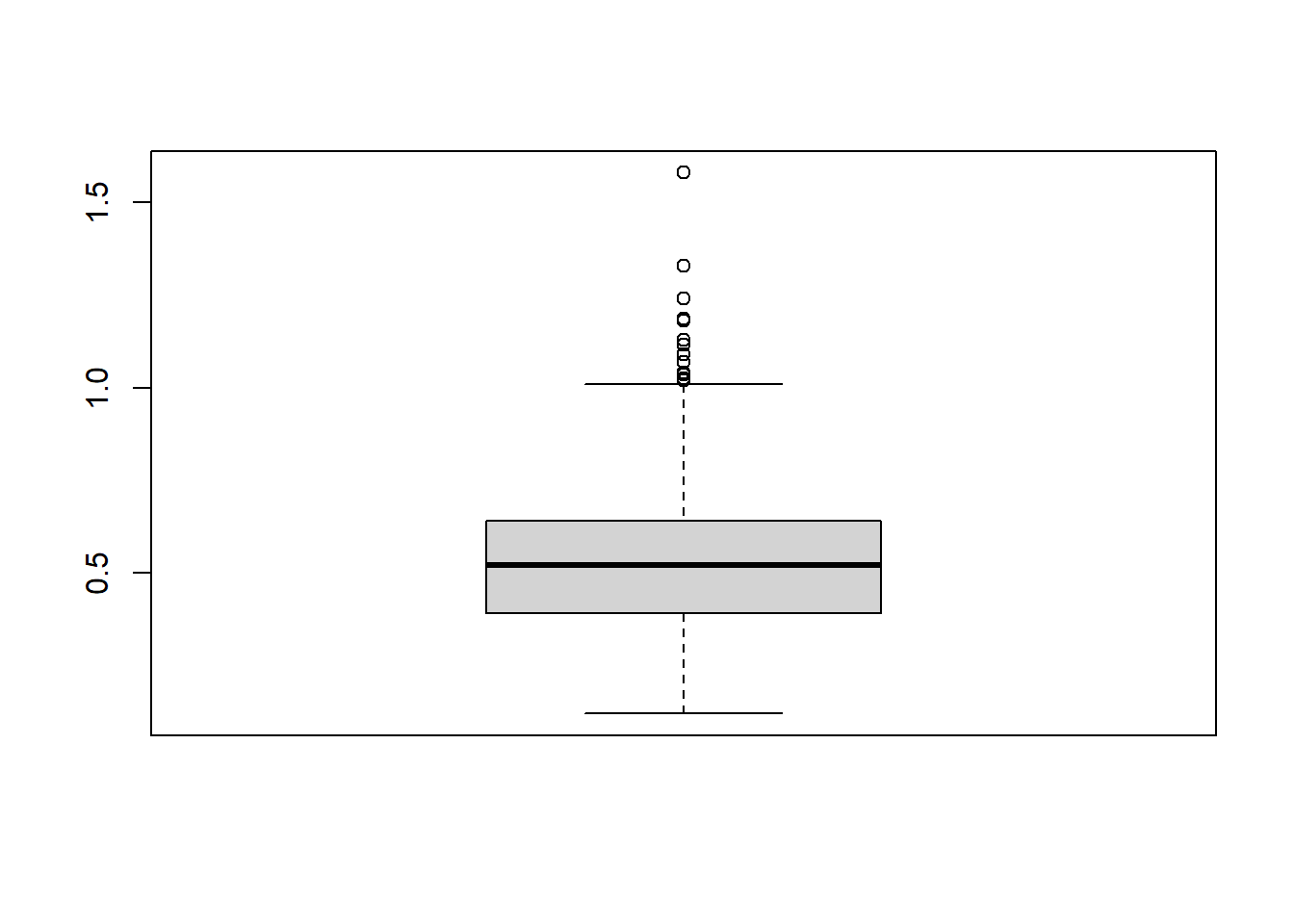
From the above boxplot we can identify Outliers.

hist(fixed.acidity)



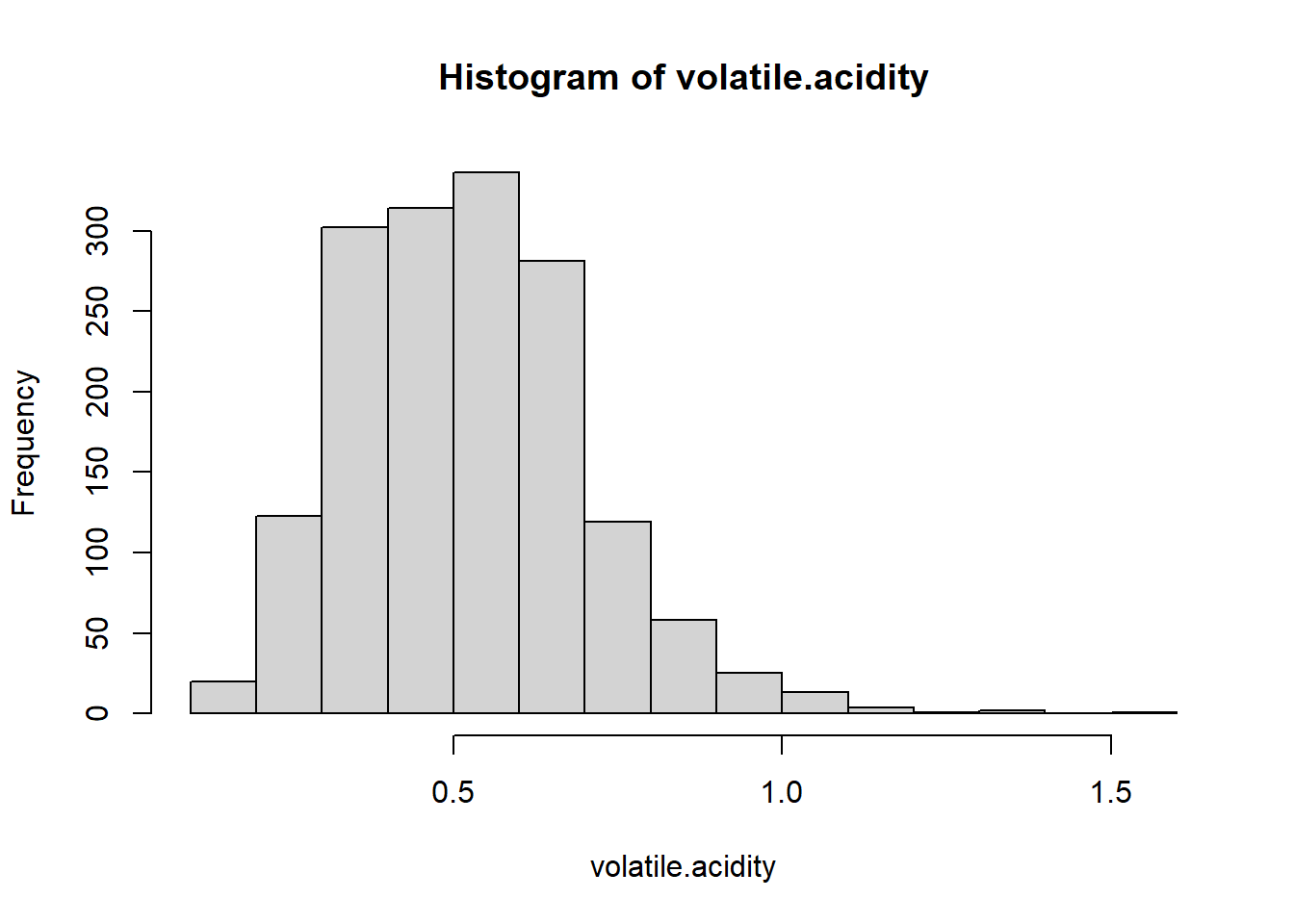
We can identify that the distribution is skewed to the right from the histogram

boxplot(volatile.acidity)



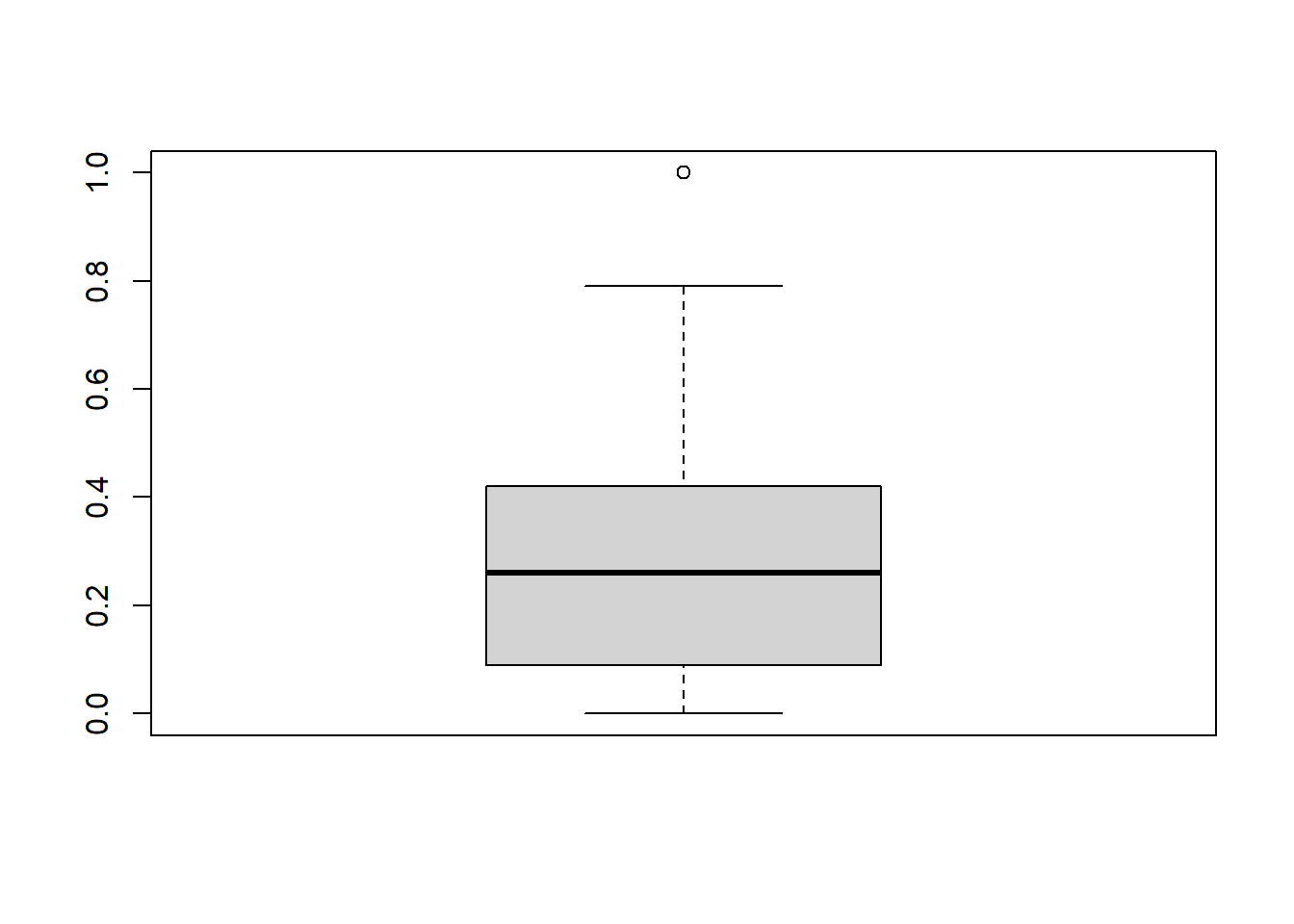
From the above boxplot we can identify Outliers.

hist(volatile.acidity)



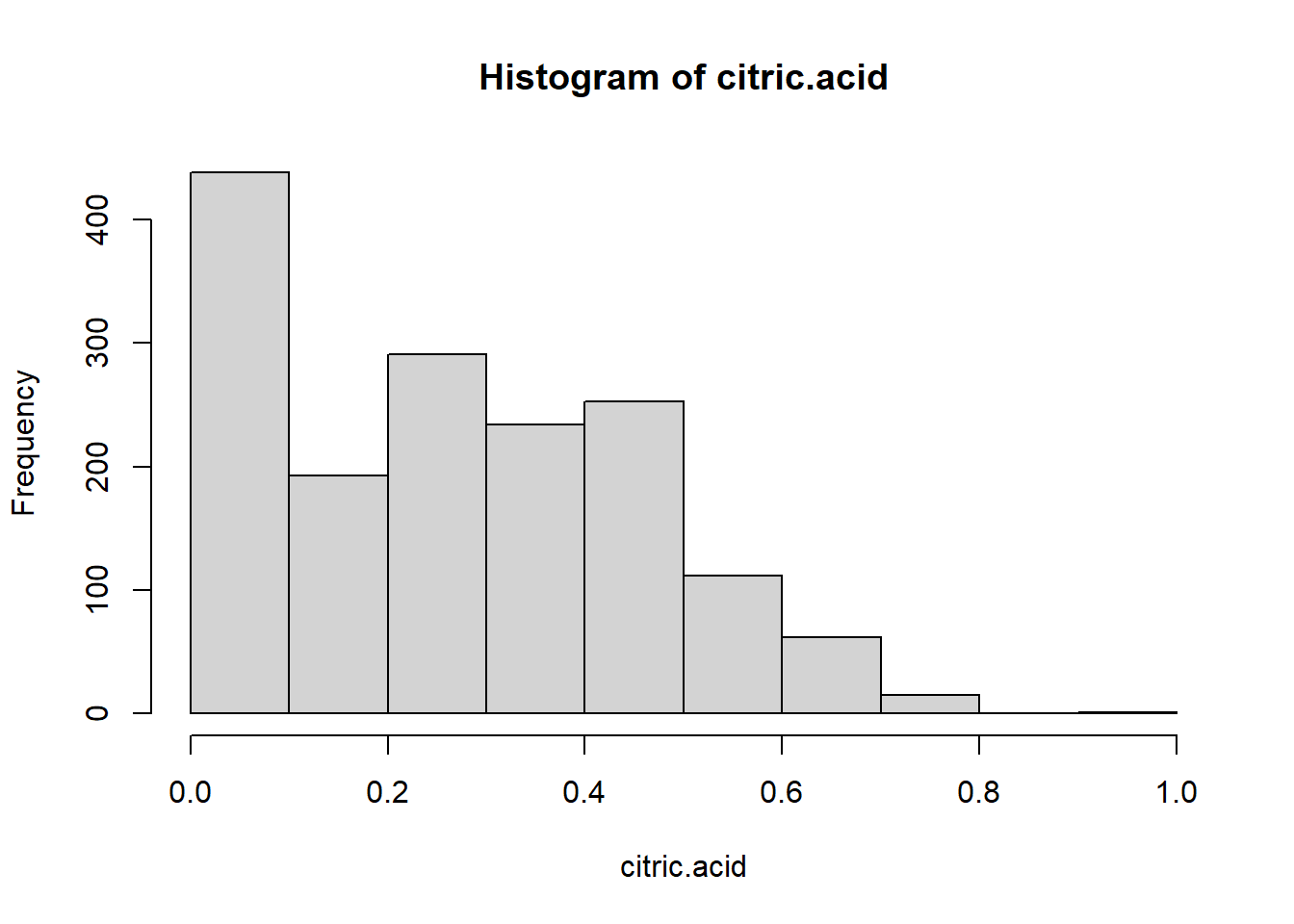
We can identify that the distribution is skewed to the right from the histogram

boxplot(citric.acid)



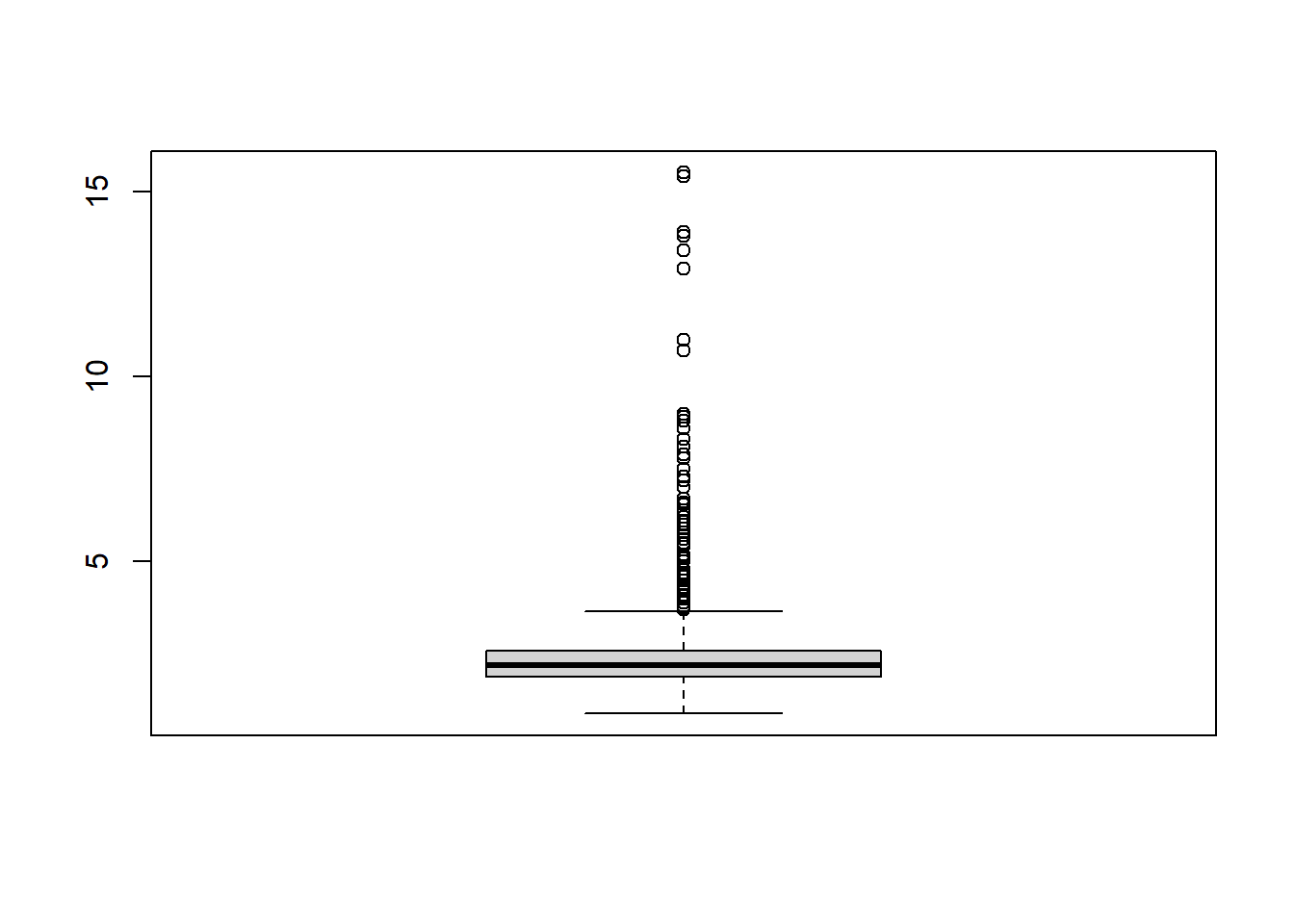
We can identify Outliers from the boxplot

hist(citric.acid)



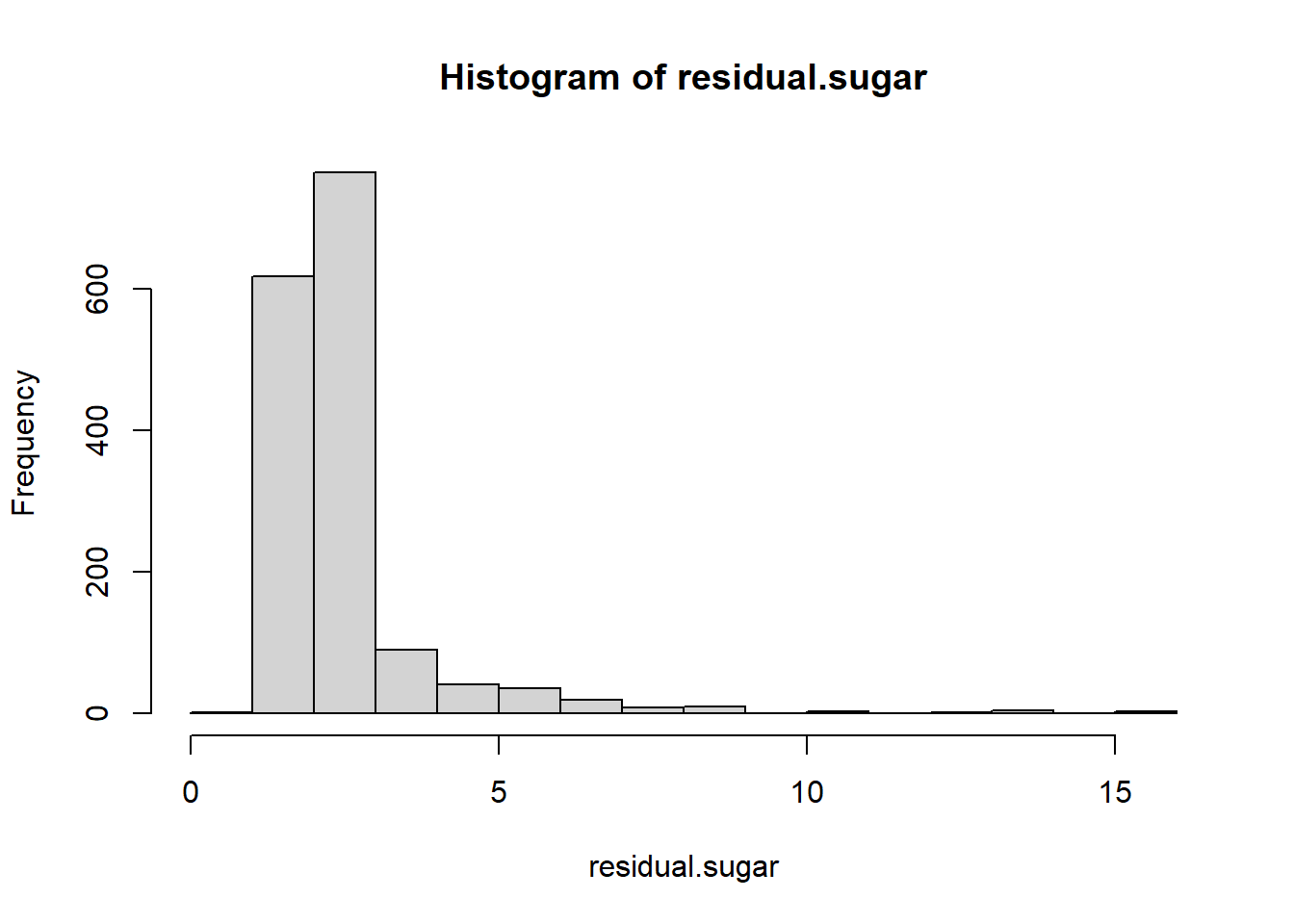
We can identify that the distribution is skewed to the right.

boxplot(residual.sugar)



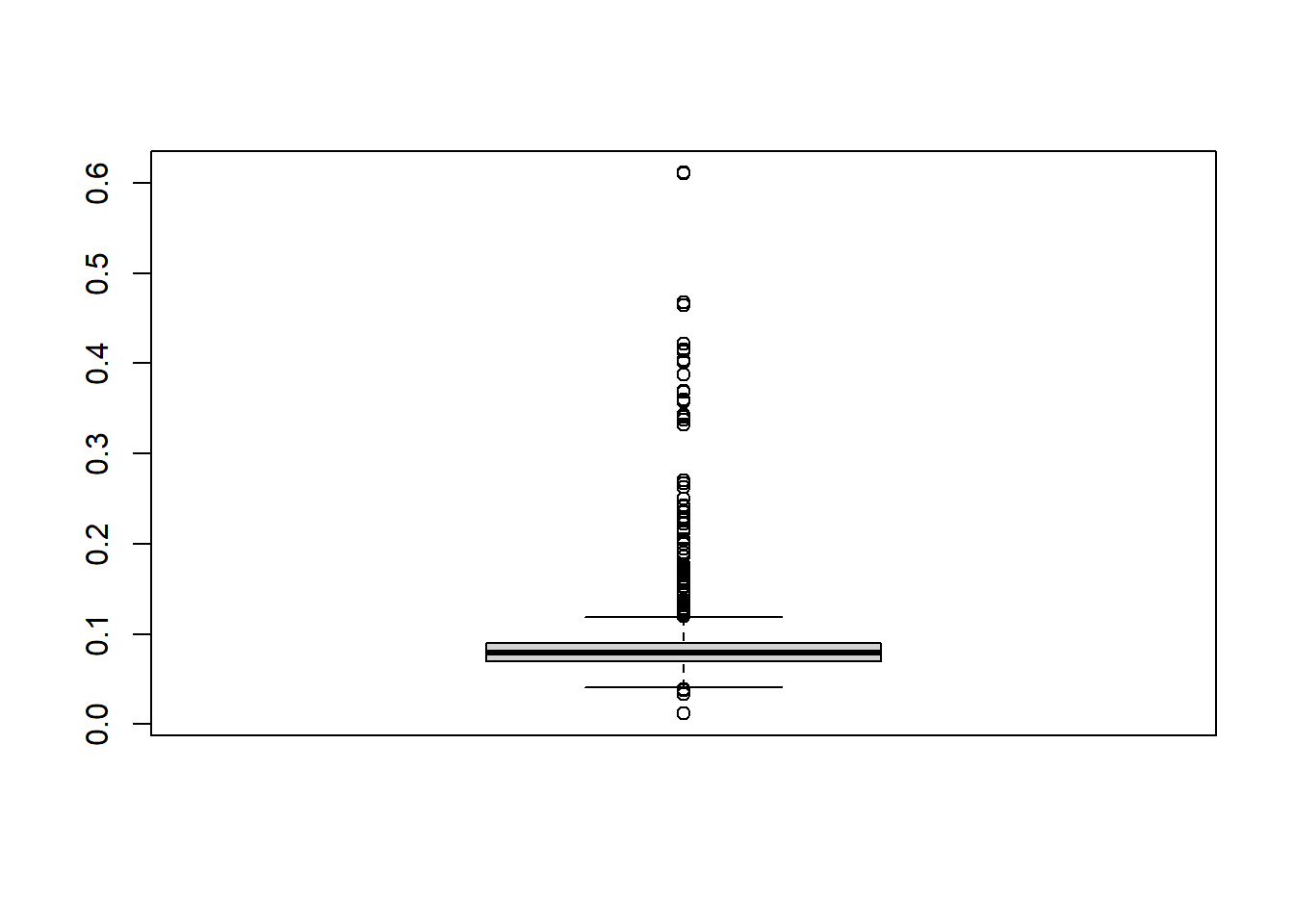
We can identify Outliers from the boxplot.

hist(residual.sugar)



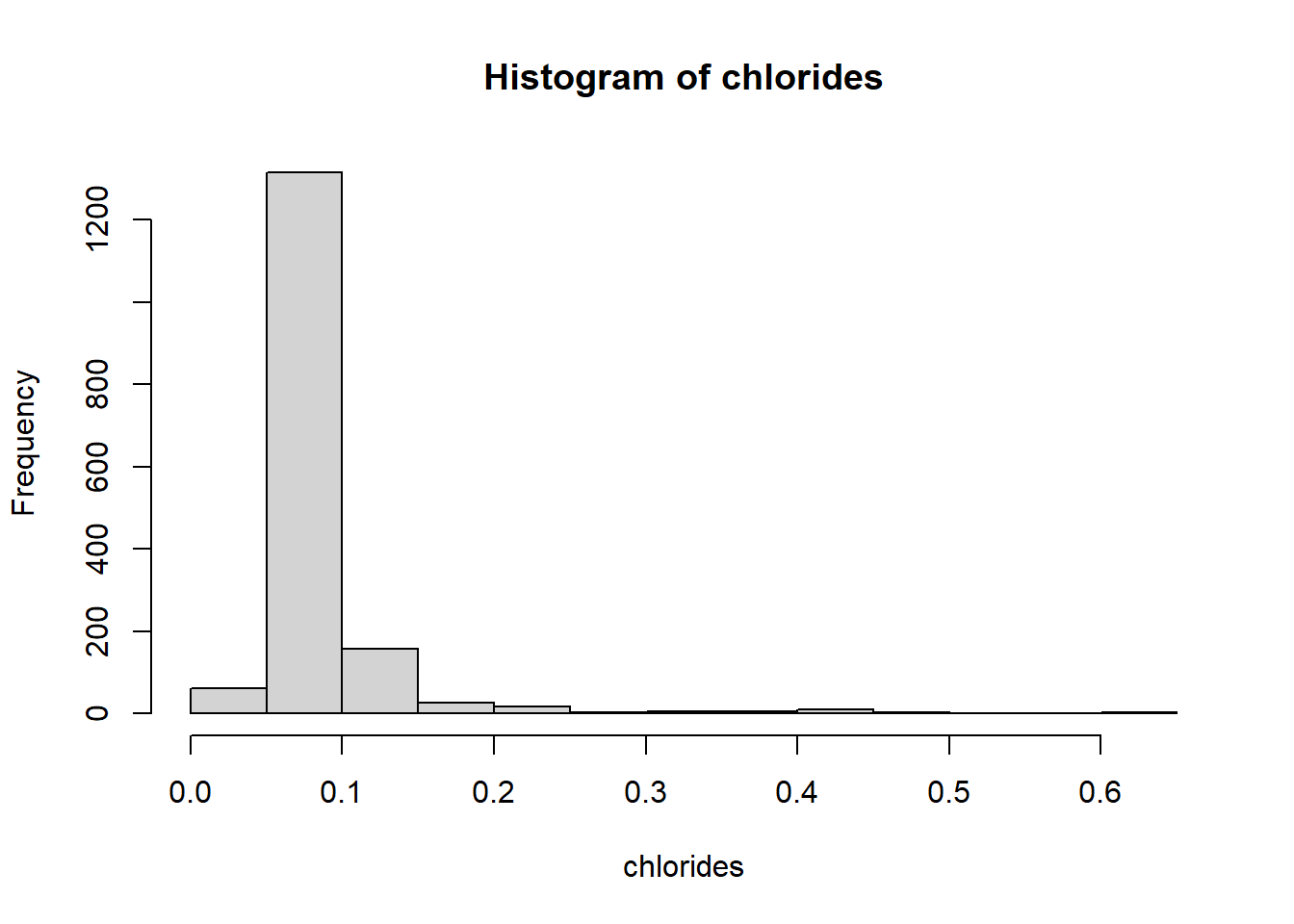
We can identify that the distribution is skewed to the right.

boxplot(chlorides)



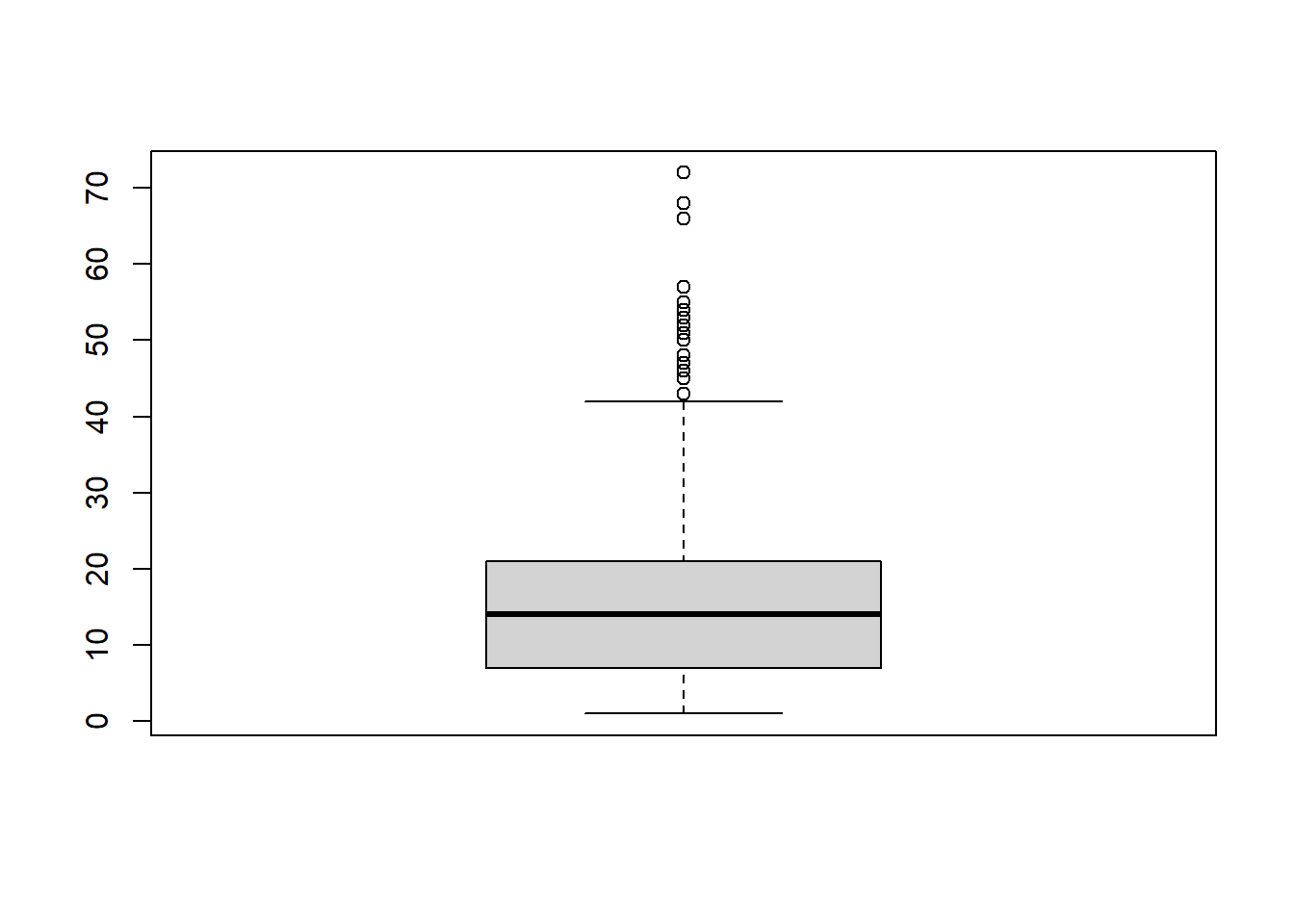
We can identify Outliers from the boxplot

hist(chlorides)

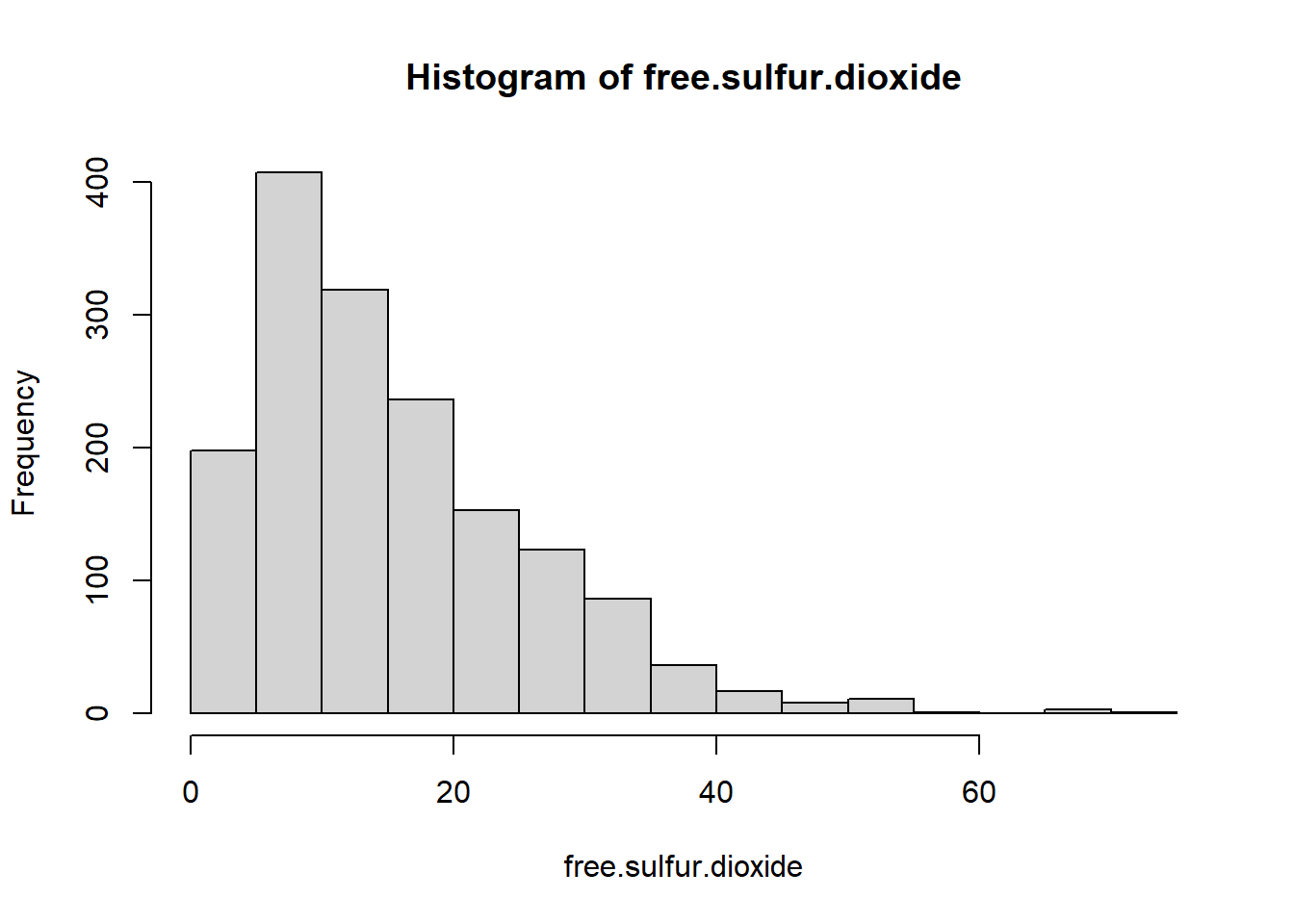


We can identify that the distribution is skewed to the right

boxplot(free.sulfur.dioxide)

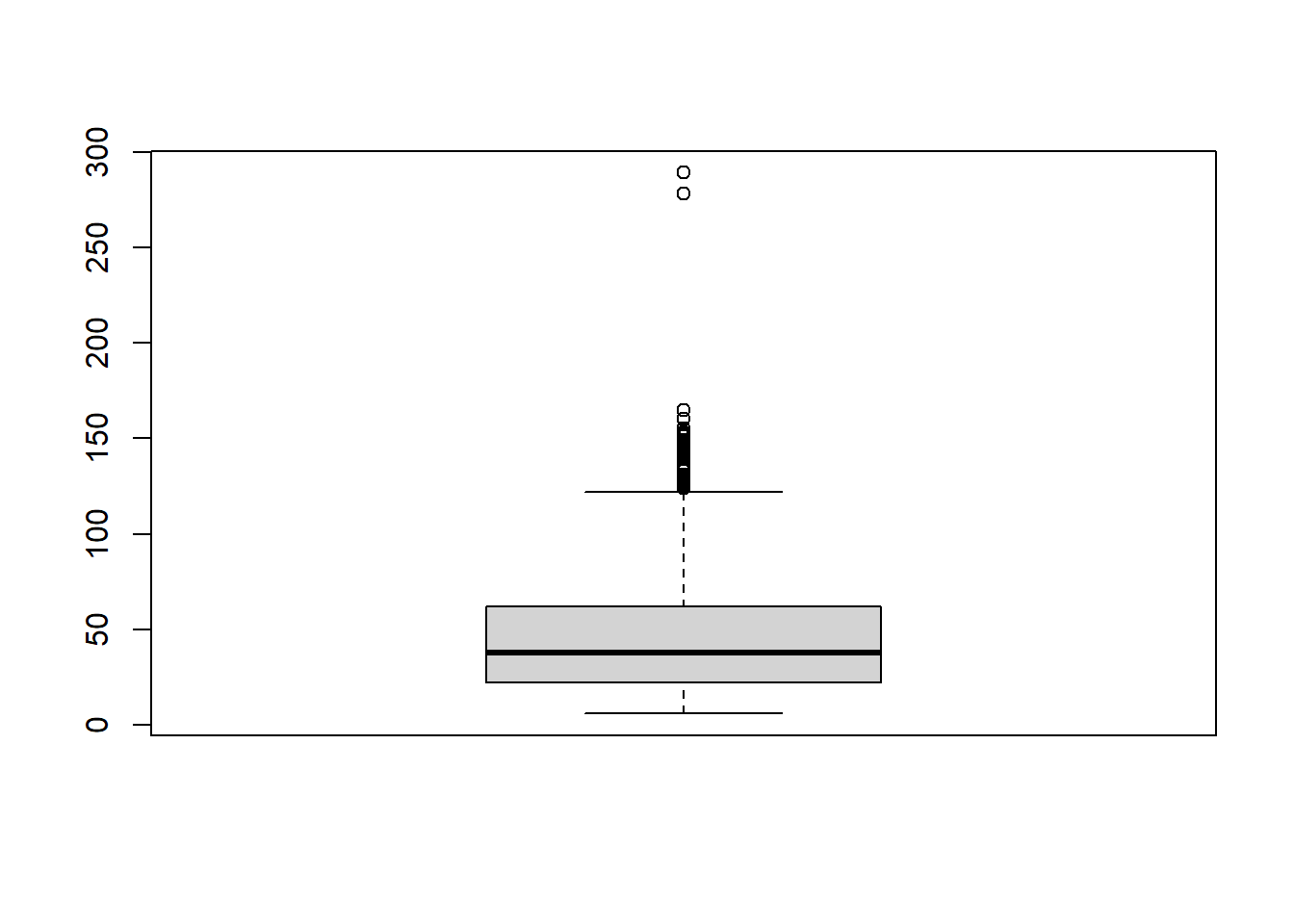
We can identify Outliers from the above box plot.

hist(free.sulfur.dioxide)



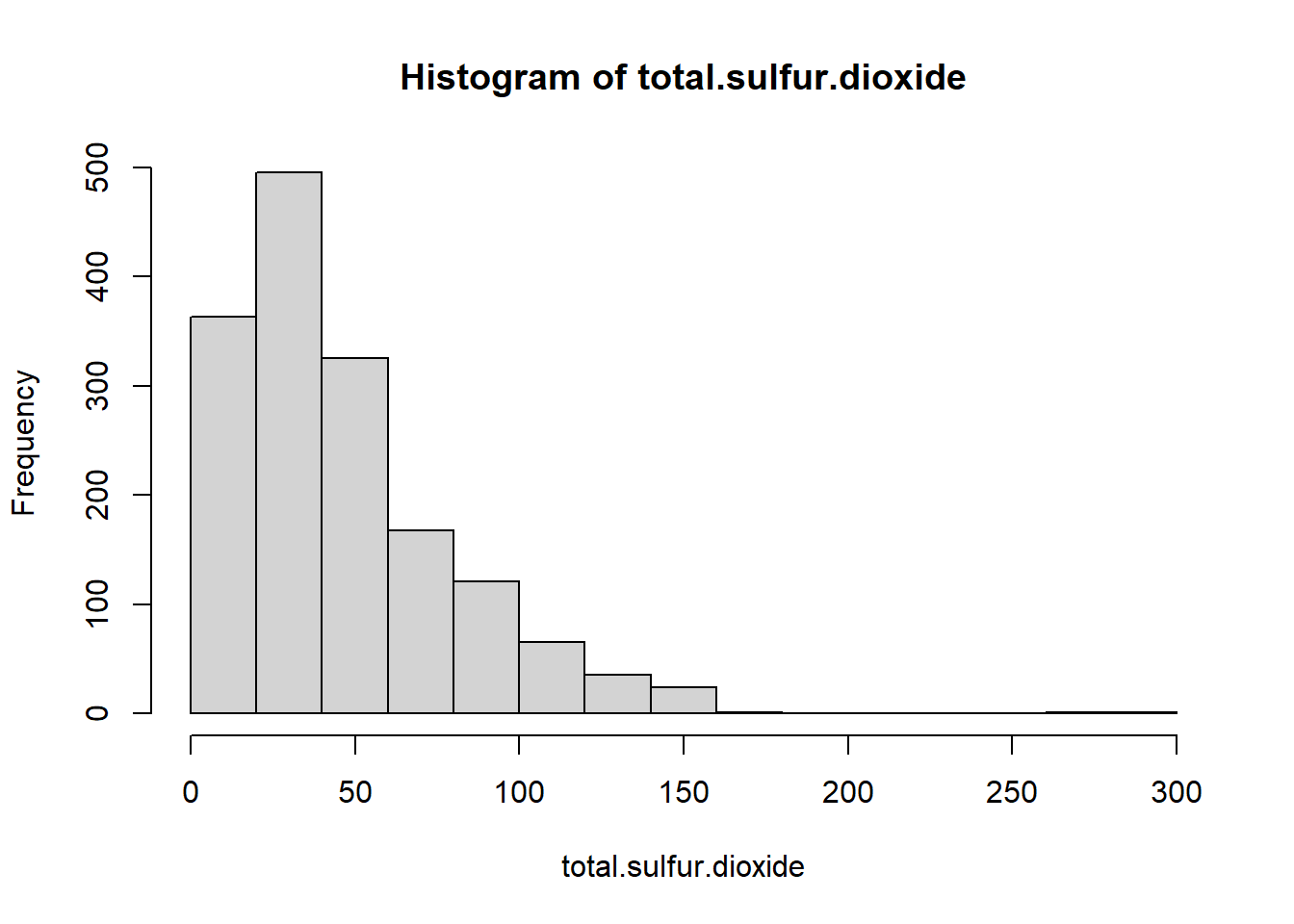
We can identify that the distribution is right skewed.

boxplot(total.sulfur.dioxide)



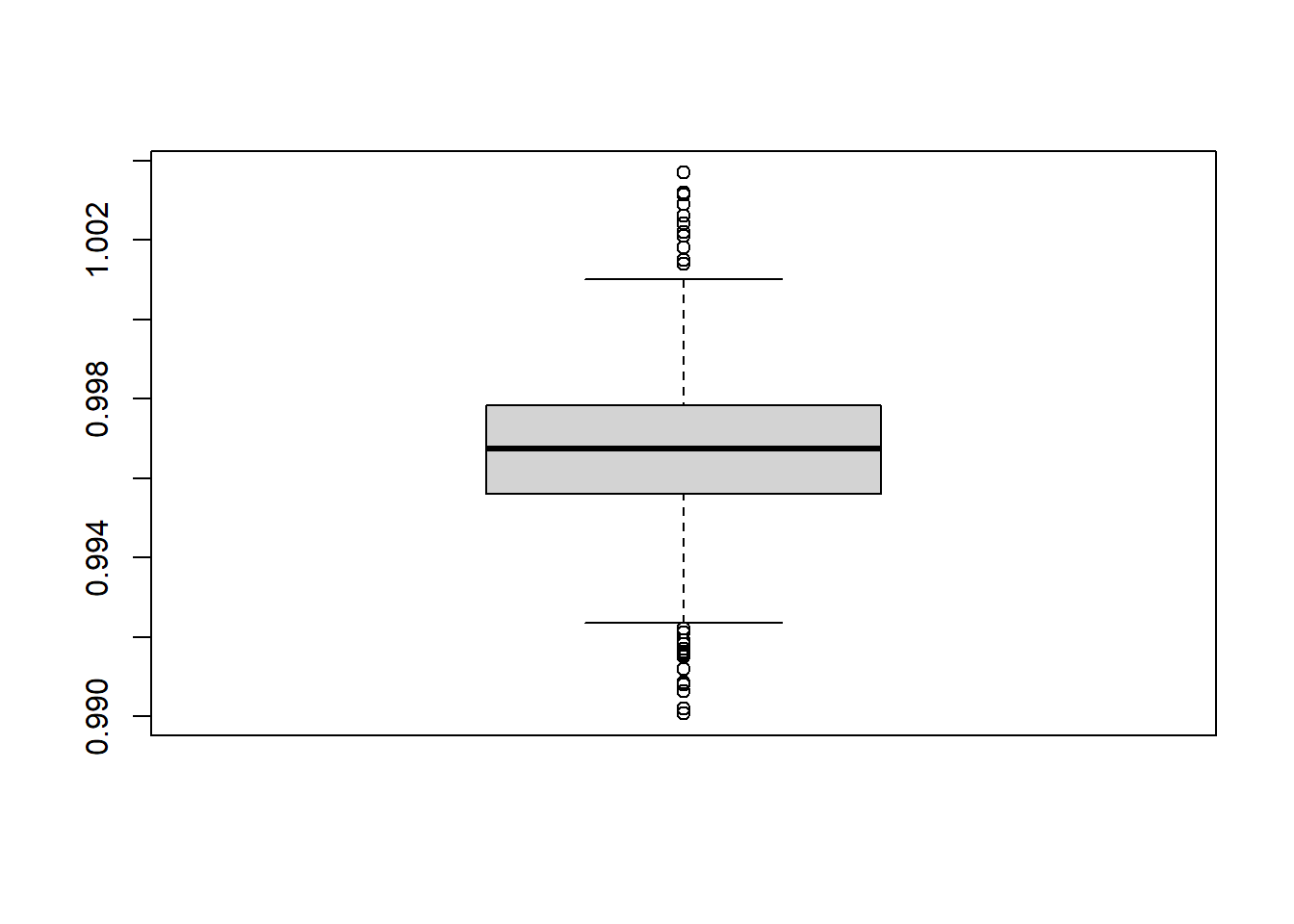
We can identify 2 Outliers from the boxplot.

hist(total.sulfur.dioxide)

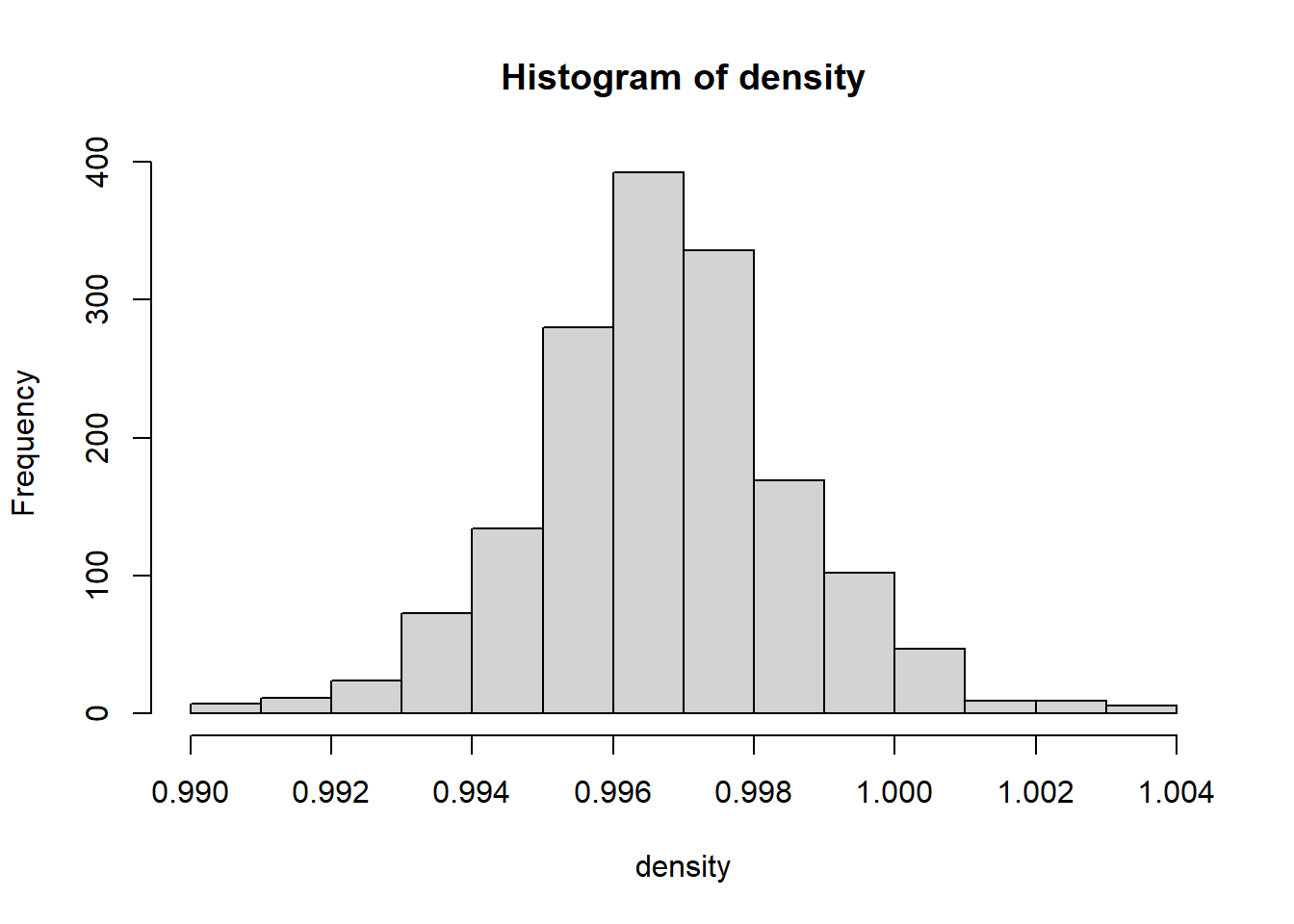


We can identify that the distribution is right skewed.

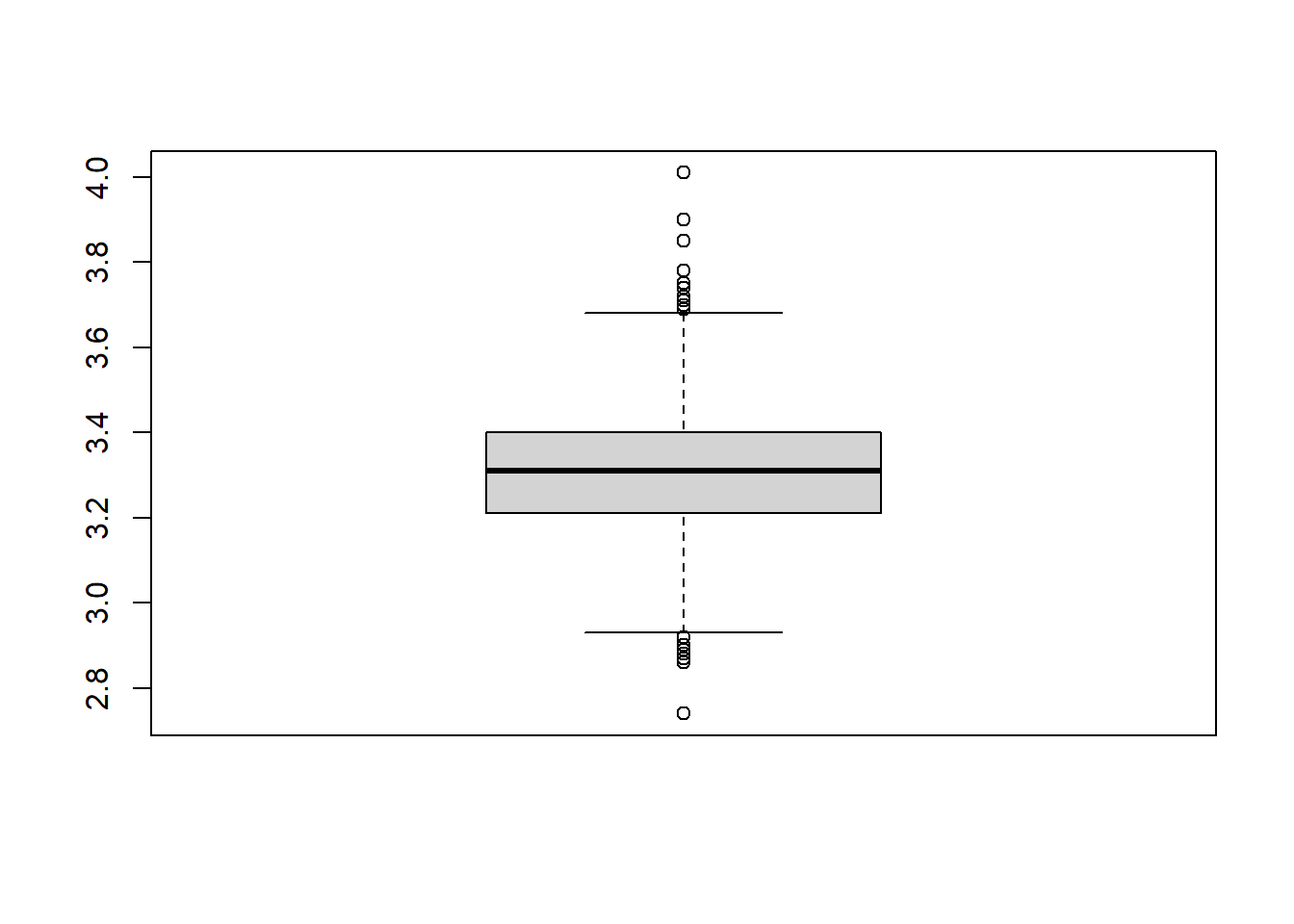
boxplot(density)

We can identify Outliers from the above box plot.

hist(density)

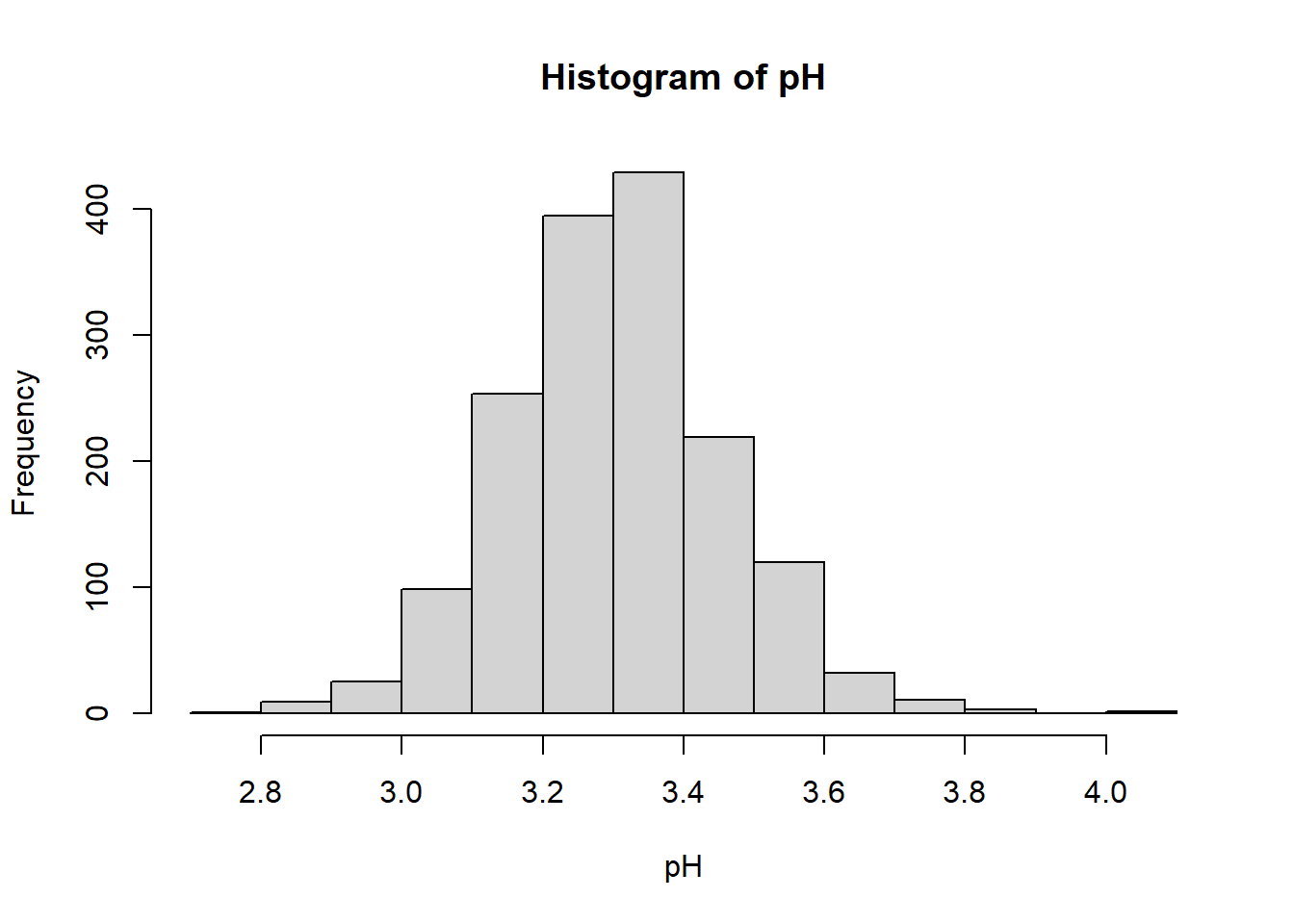


We can identify that the distribution is Normal distribution.

boxplot(pH)

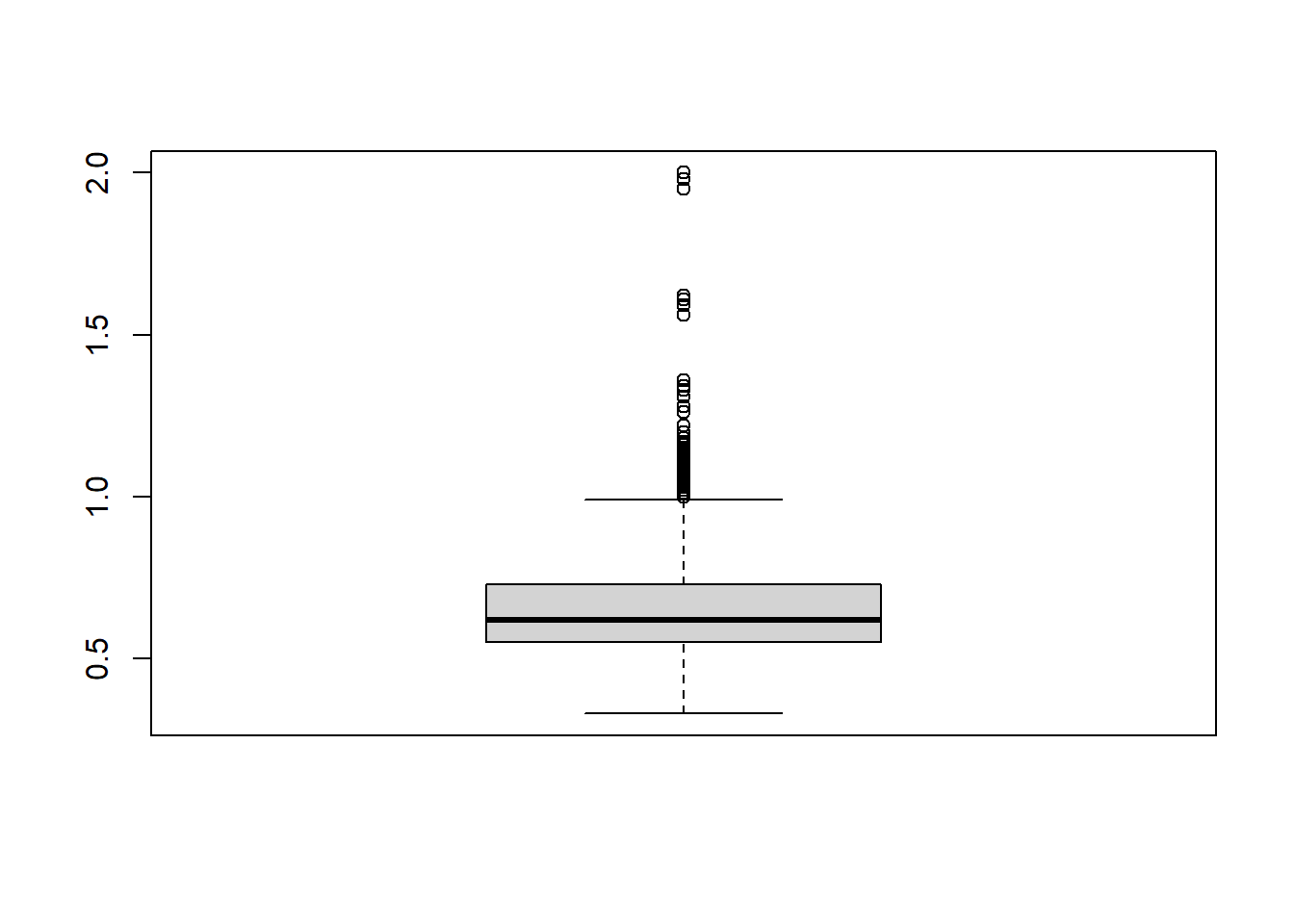
We can identify Outliers from boxplot.

hist(pH)



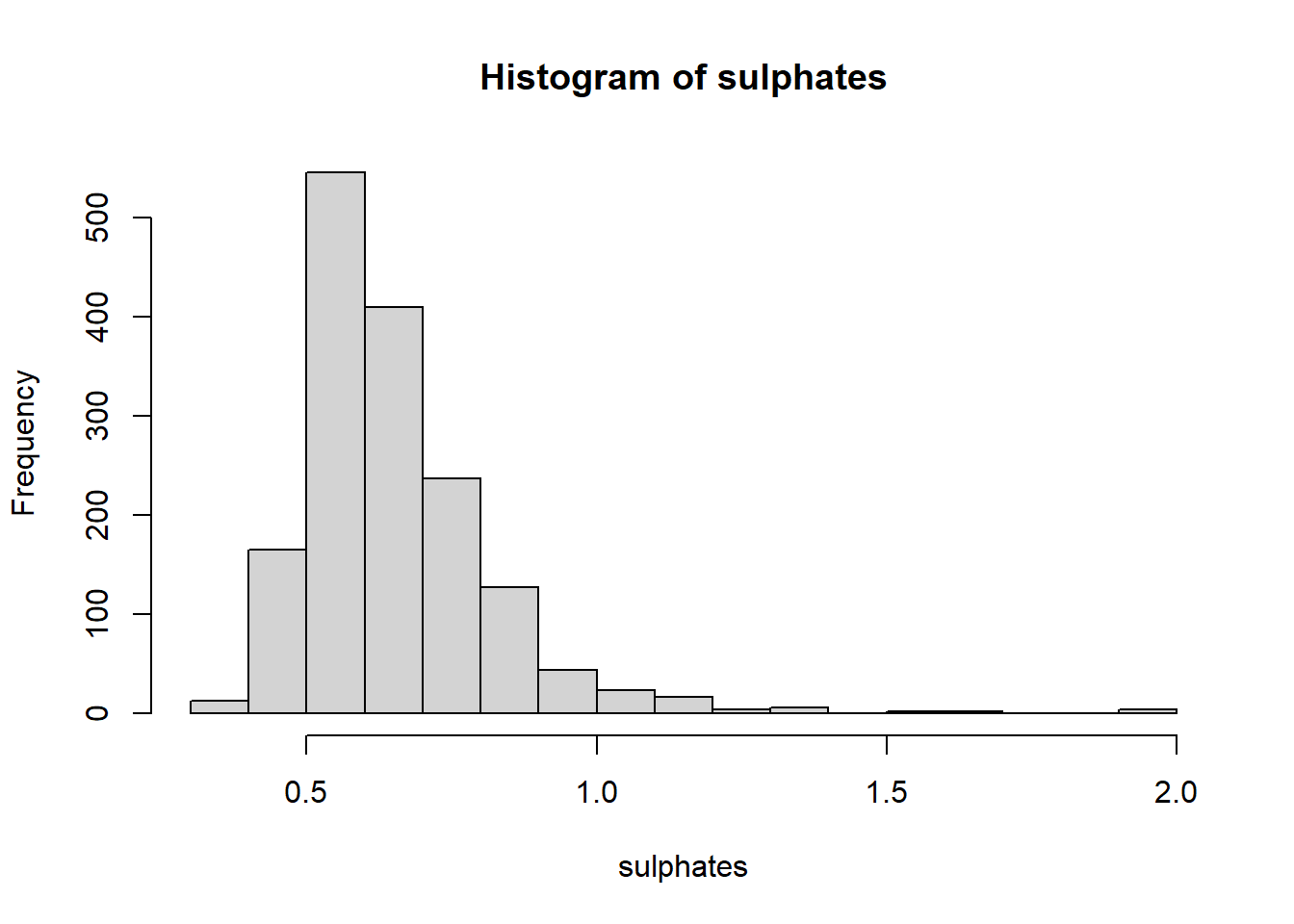
We can identify that the distribution is normal distribution.

boxplot(sulphates)



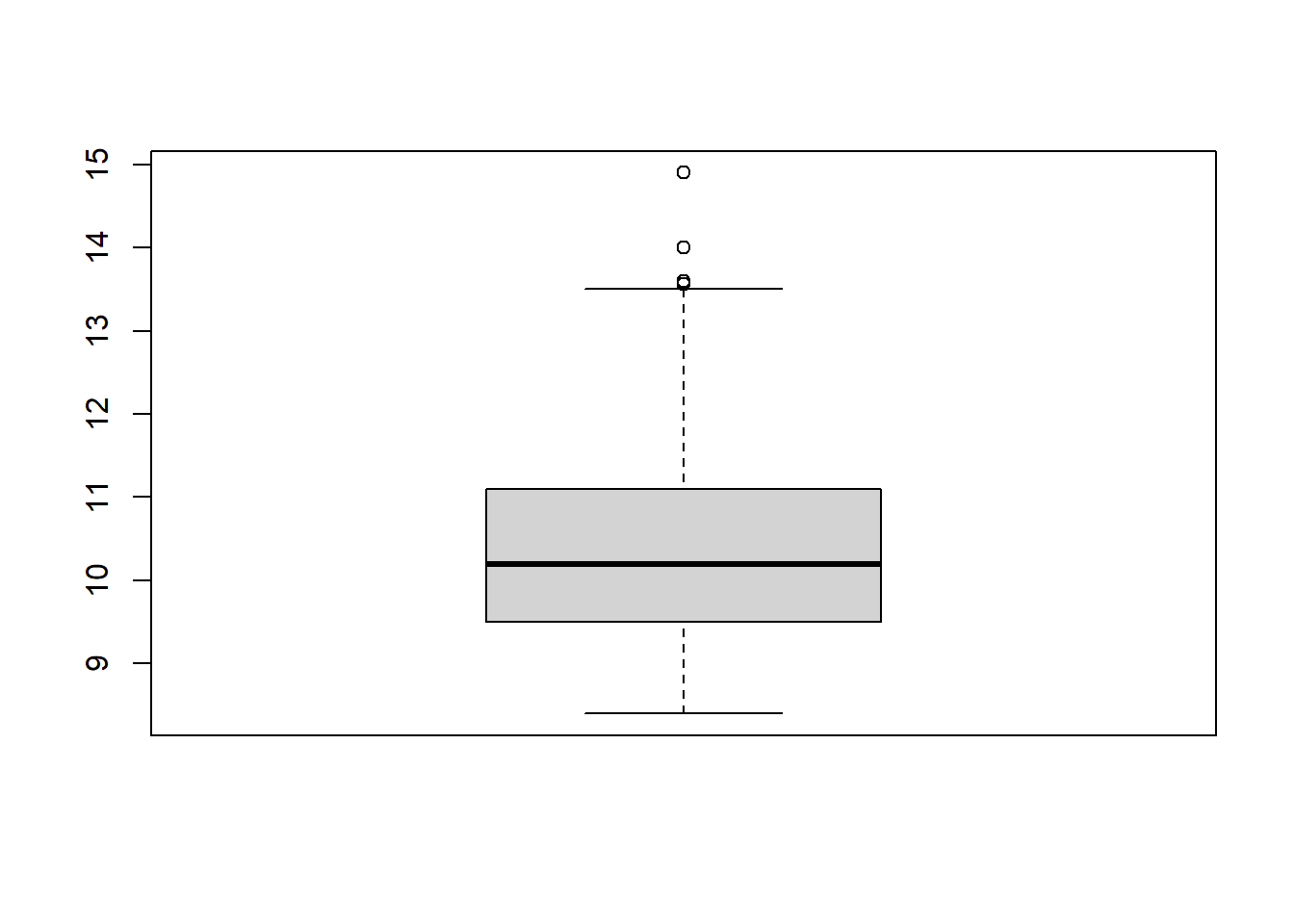
We can identify outliers from the boxplot

hist(sulphates)



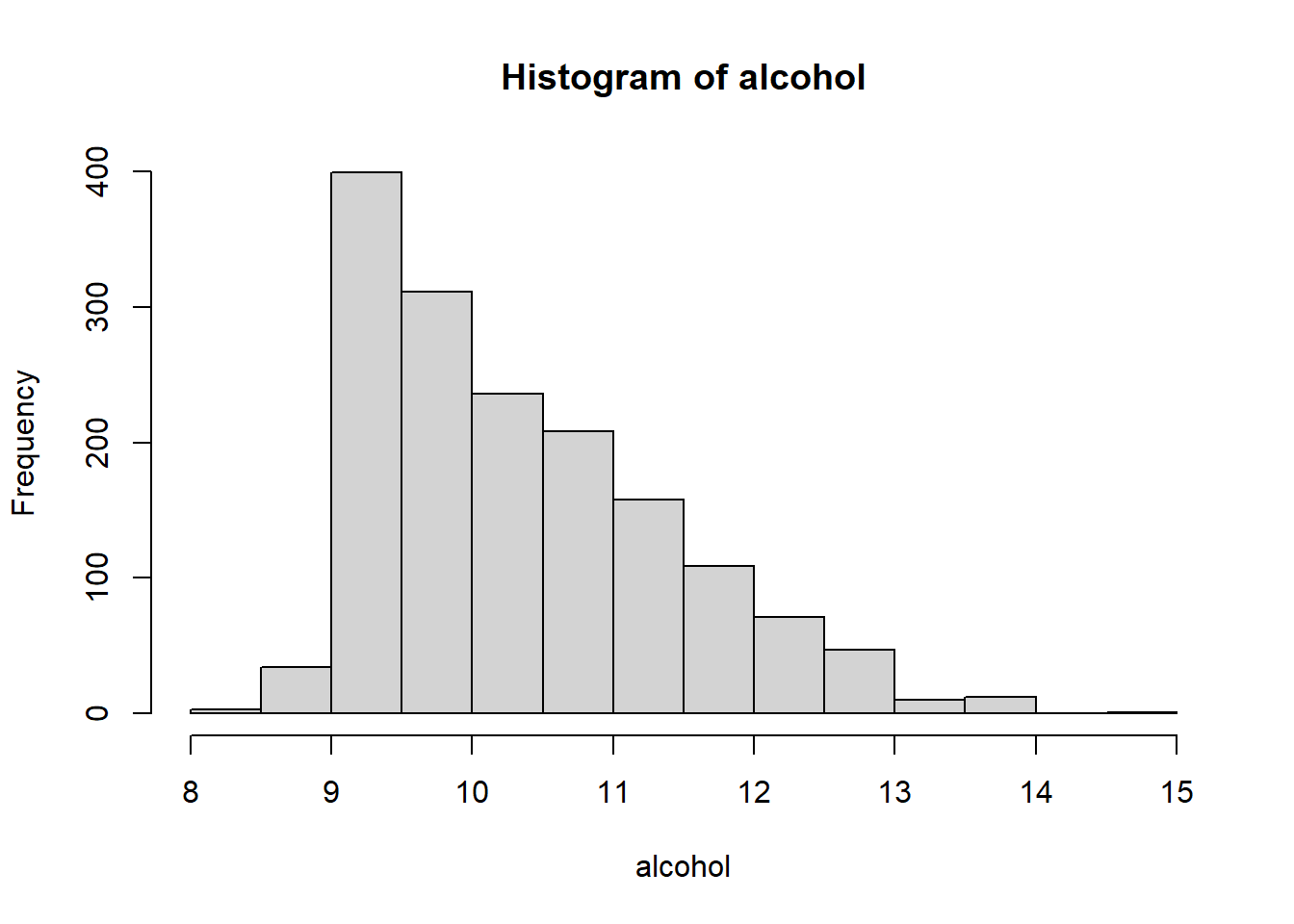
We can identify that the distribution is skewed to the right.

boxplot(alcohol)



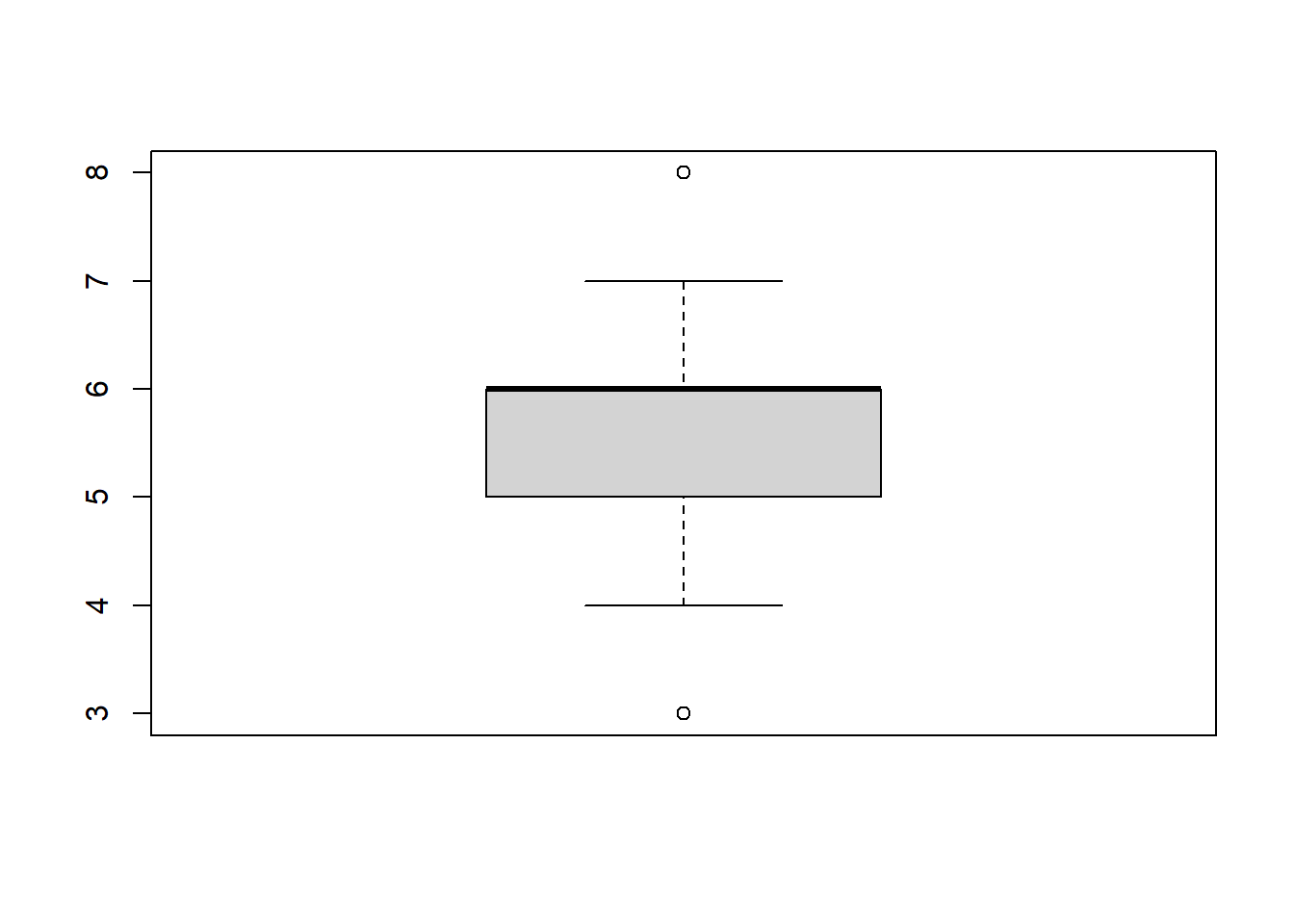
We can identify three Outliers from the boxplot

hist(alcohol)



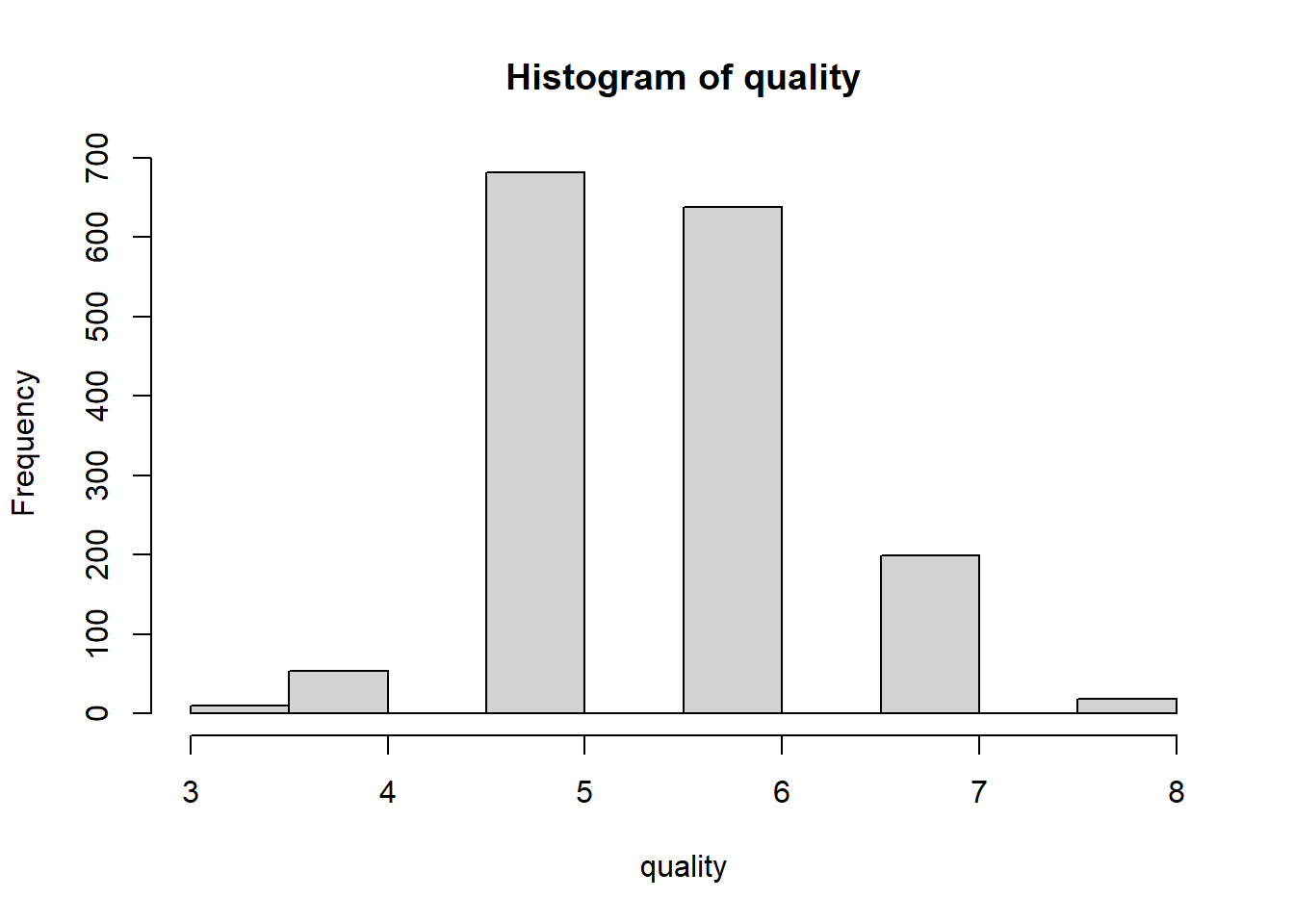
We can identify that the data is skewed to the right.

boxplot(quality)



We can identify one Outlier from the box plot.

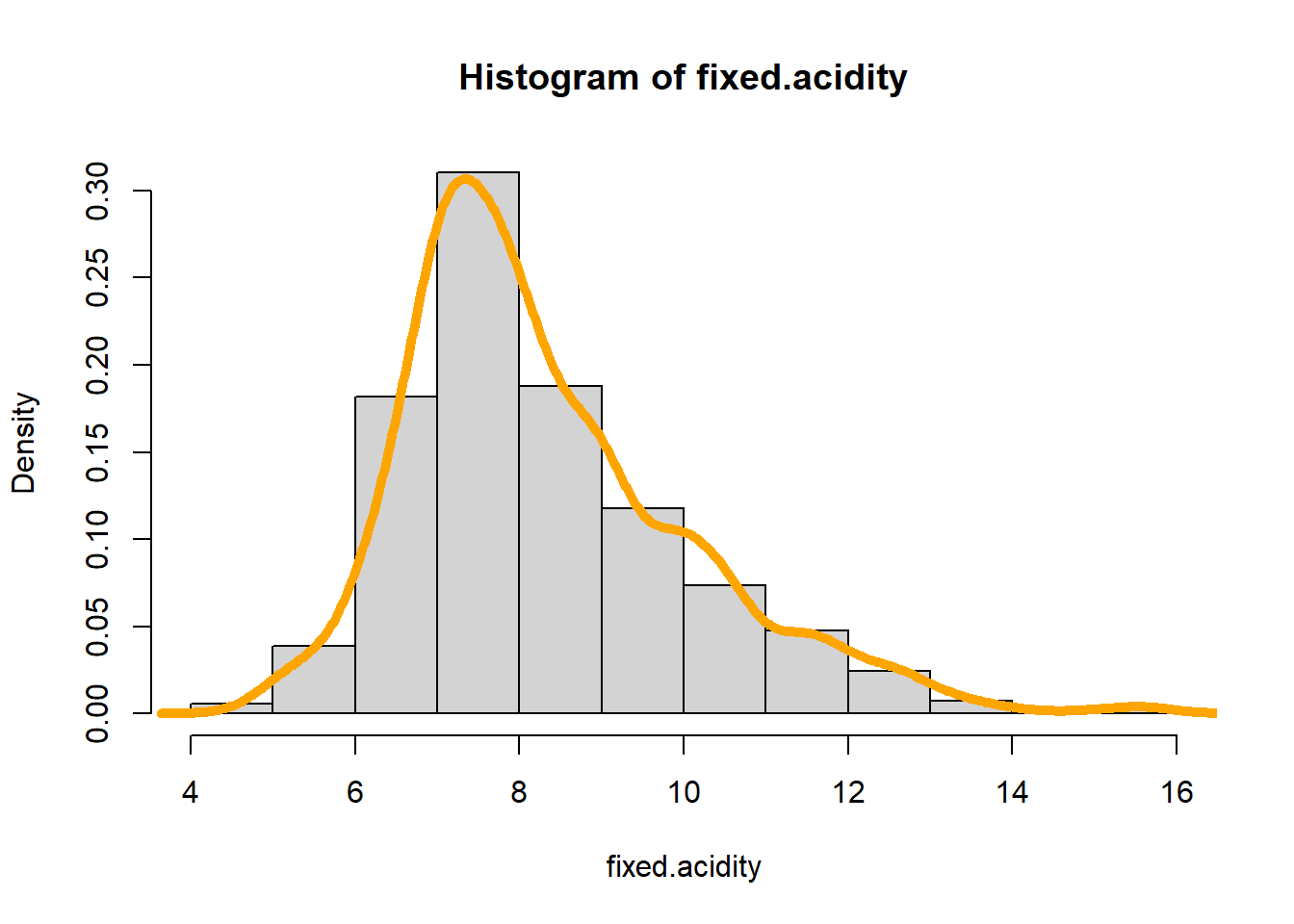
hist(quality)



We can identify that the distribution is normal

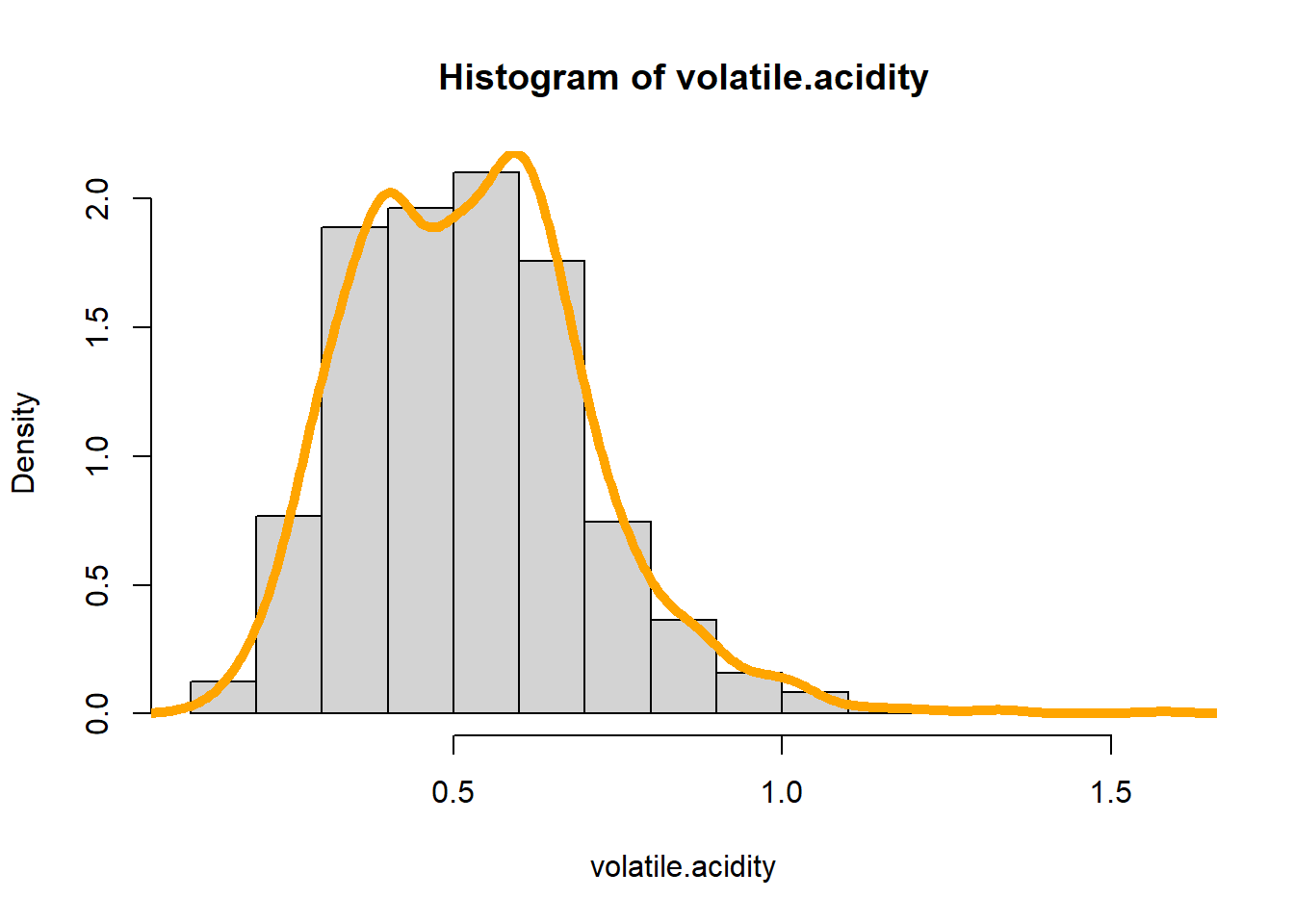
hist(fixed.acidity, freq = FALSE,ylim= c())

lines(density(fixed.acidity), lwd=5, col='orange')



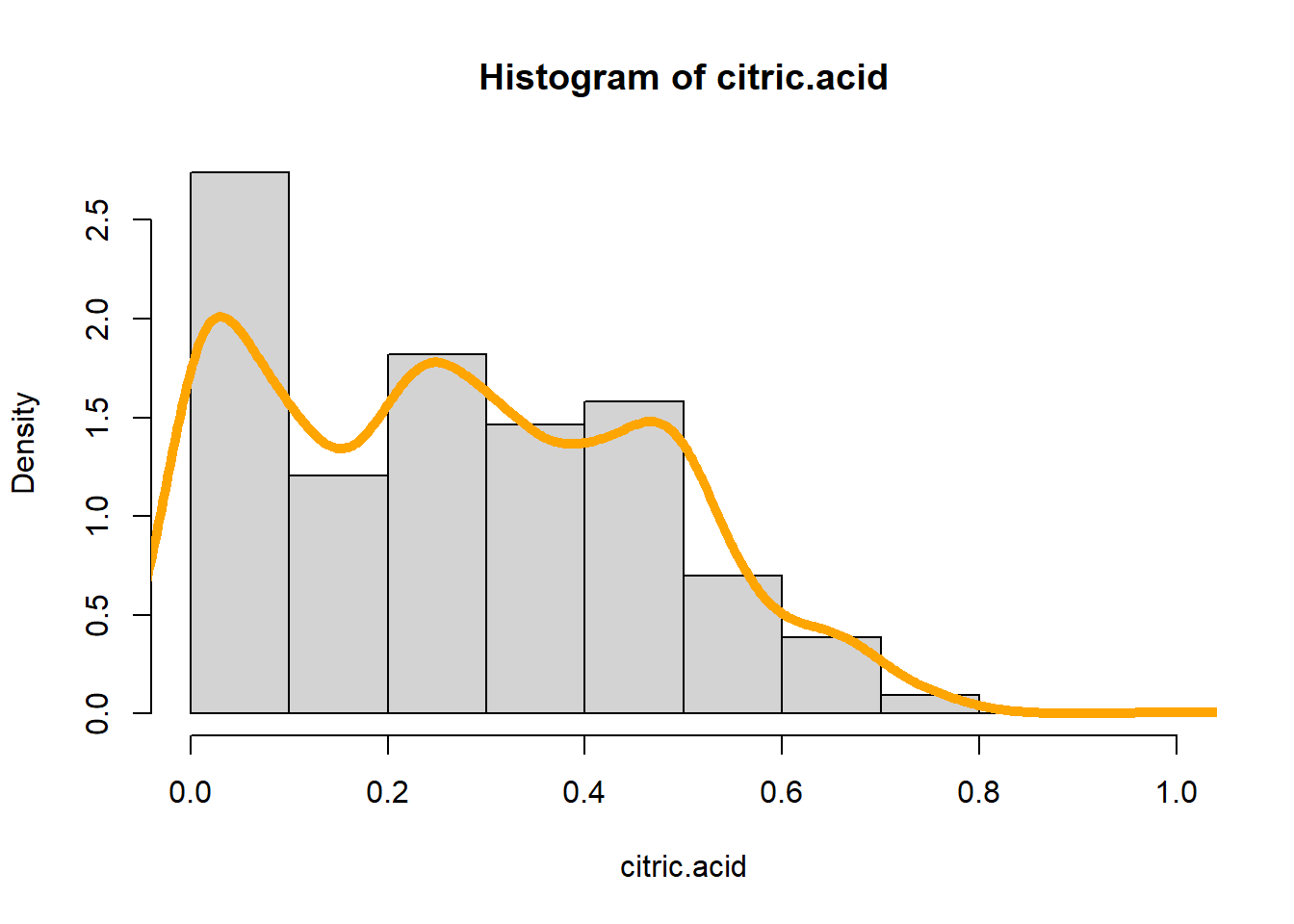
hist(volatile.acidity, freq = FALSE,ylim= c())

lines(density(volatile.acidity), lwd=5, col='orange')



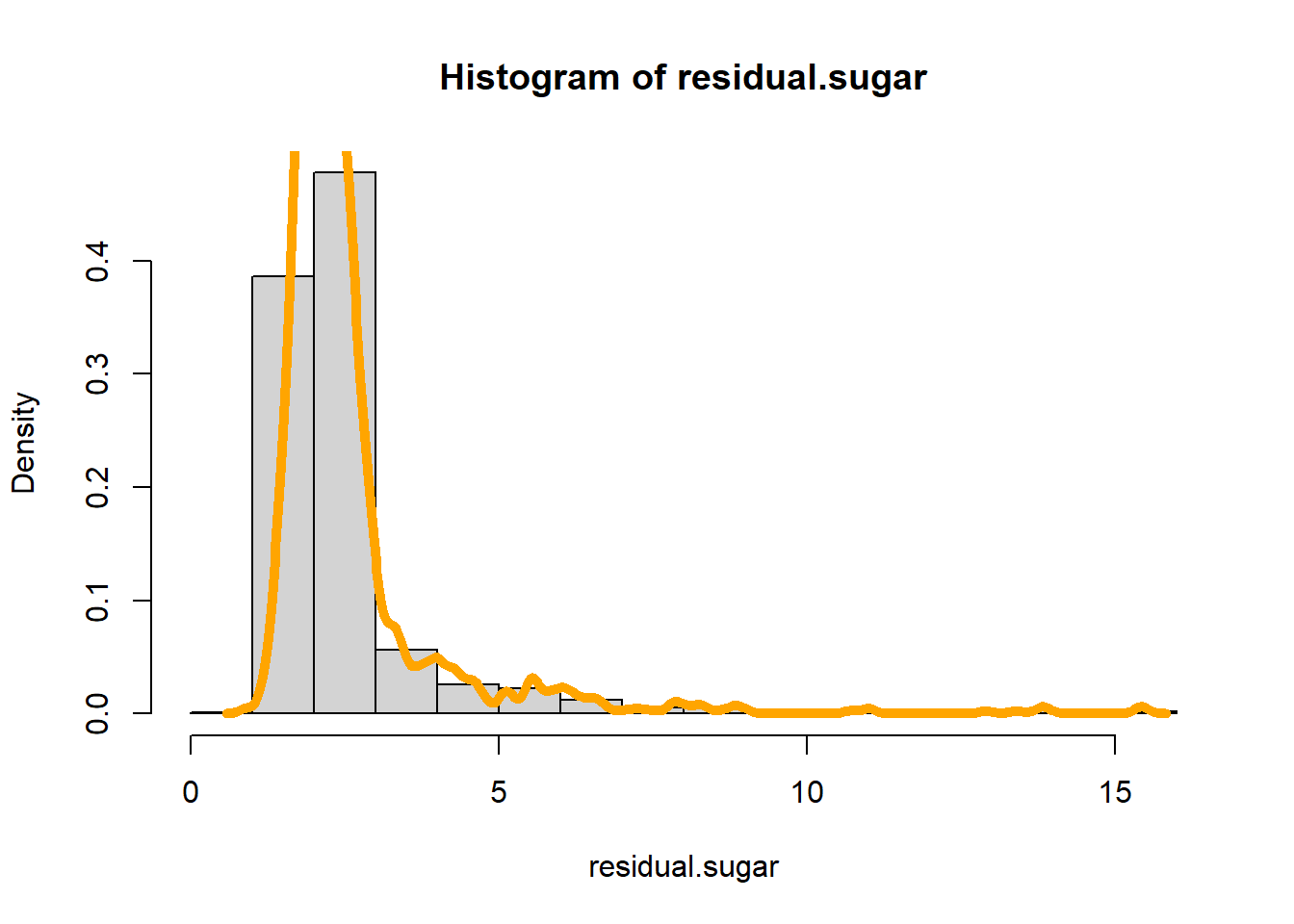
hist(citric.acid, freq = FALSE,ylim= c())

lines(density(citric.acid), lwd=5, col='orange')



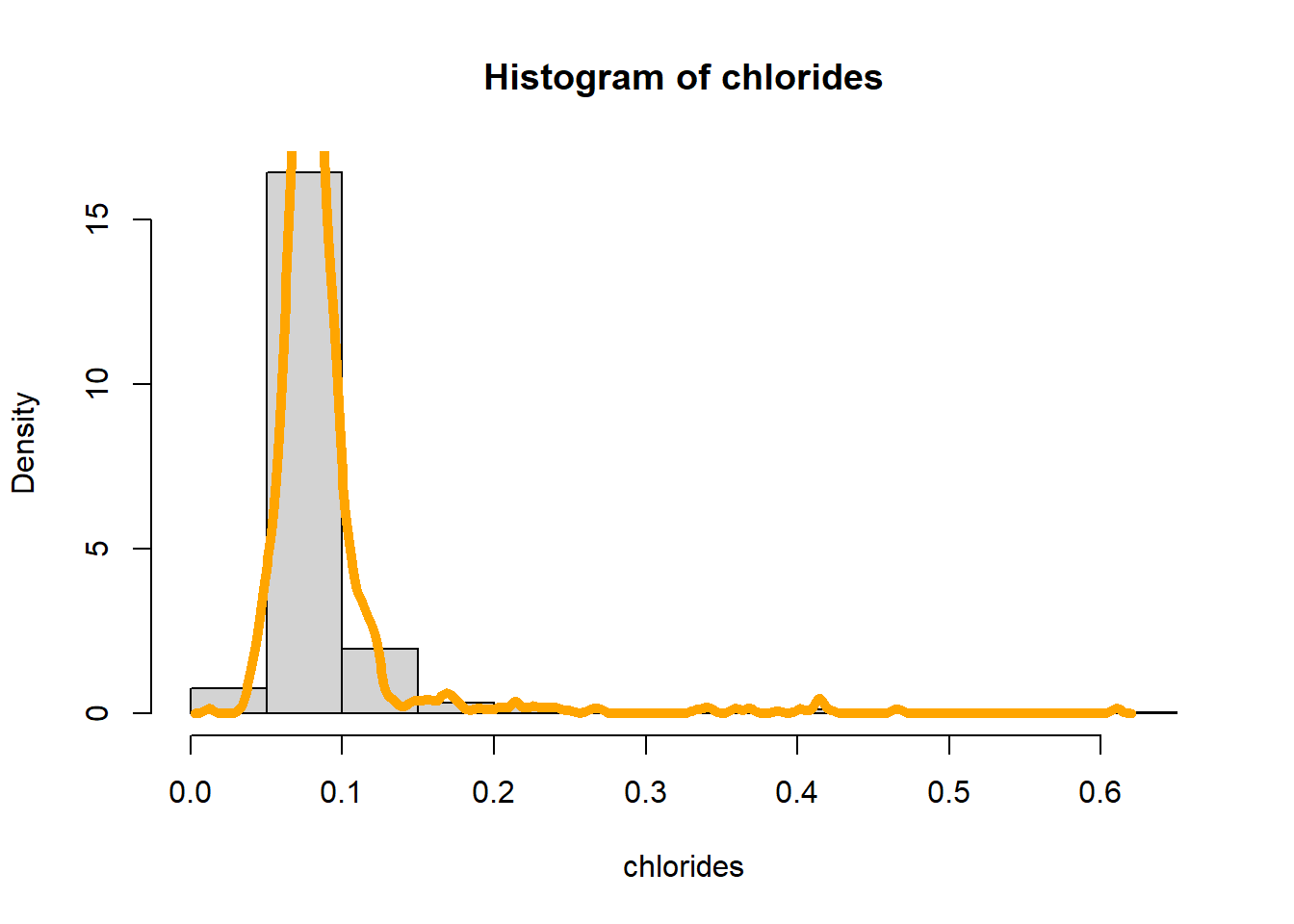
hist(residual.sugar, freq = FALSE,ylim= c())

lines(density(residual.sugar), lwd=5, col='orange')



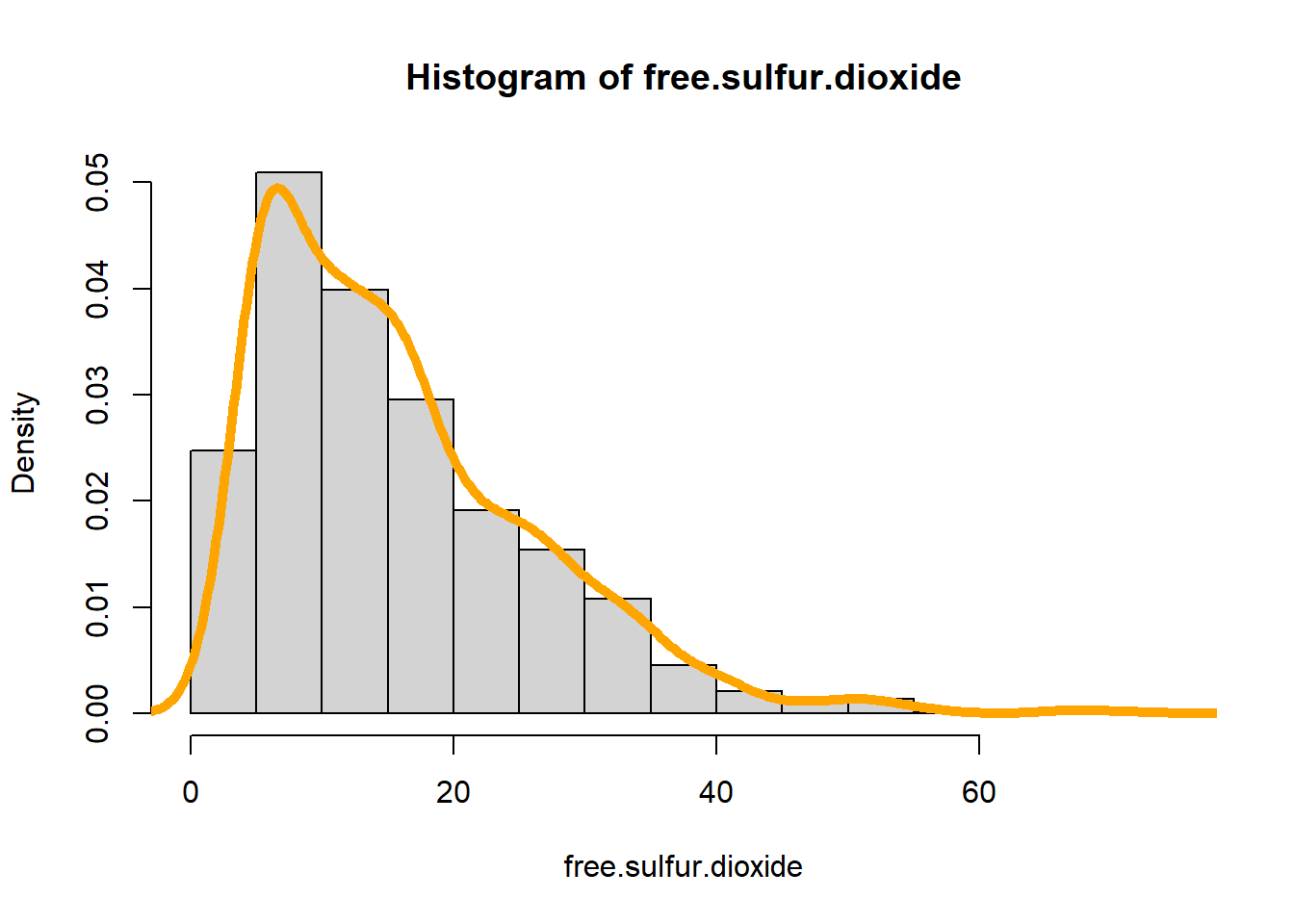
hist(chlorides, freq = FALSE,ylim= c())

lines(density(chlorides), lwd=5, col='orange')



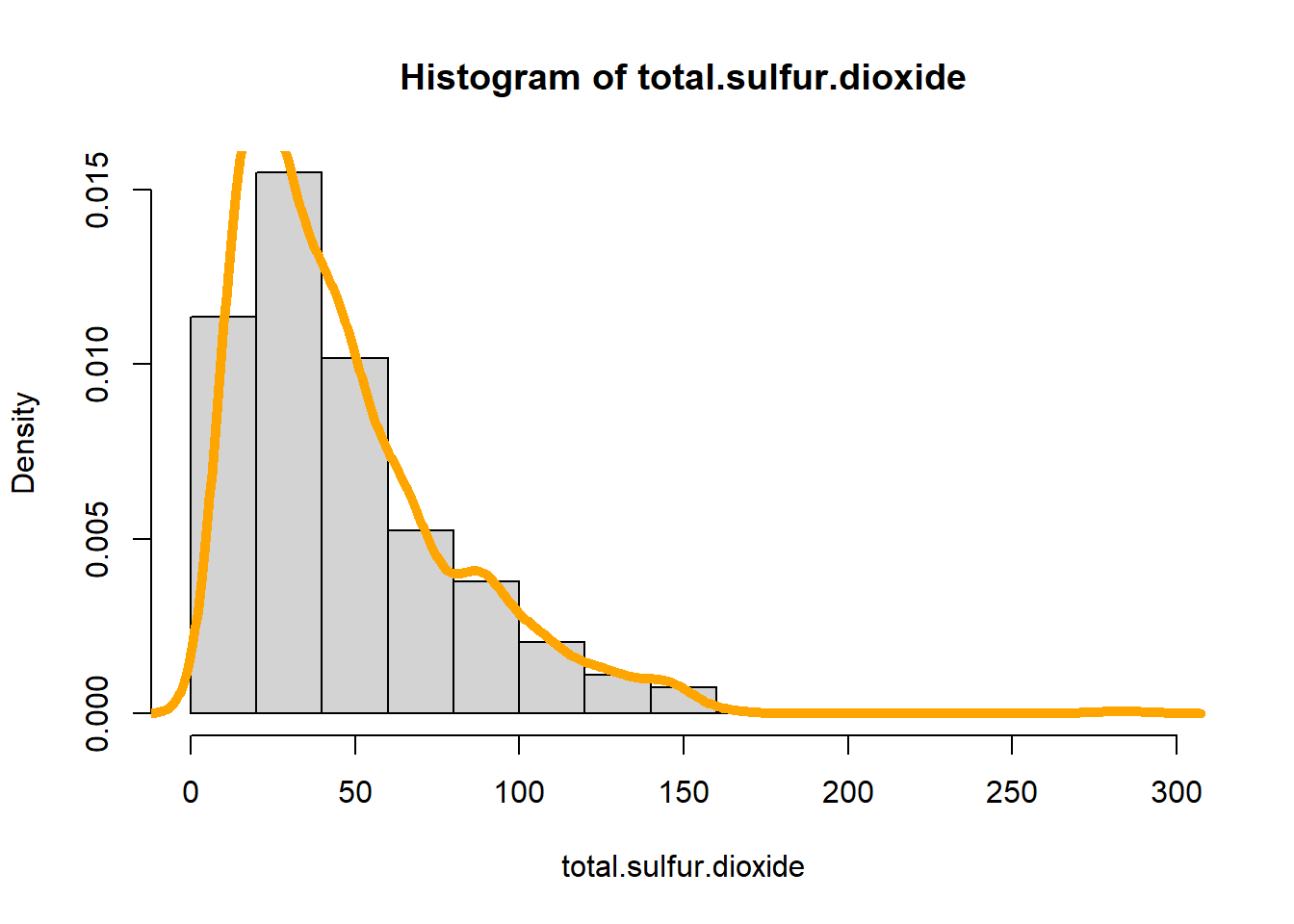
hist(free.sulfur.dioxide, freq = FALSE,ylim= c())

lines(density(free.sulfur.dioxide), lwd=5, col='orange')



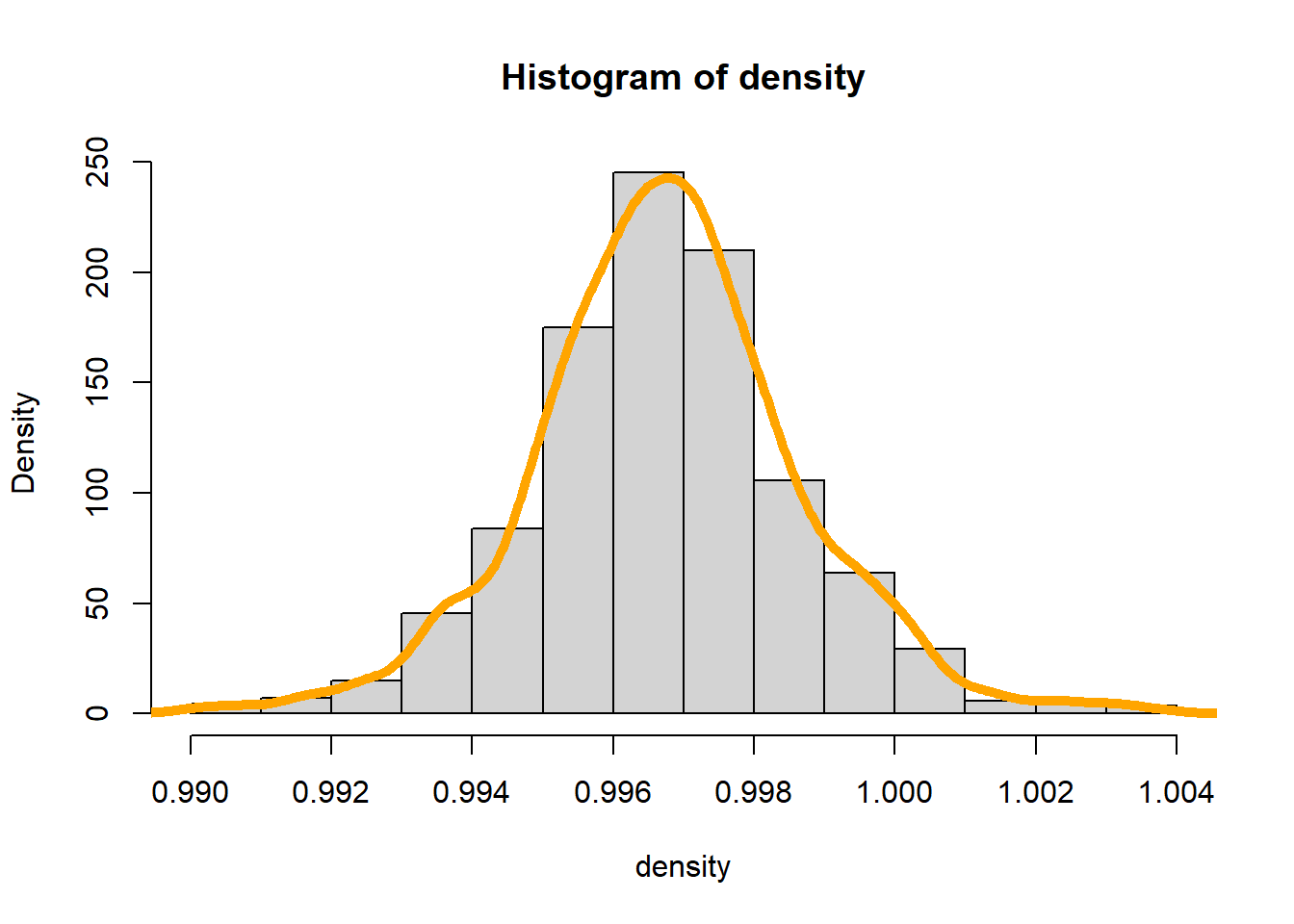
hist(total.sulfur.dioxide, freq = FALSE,ylim= c())

lines(density(total.sulfur.dioxide), lwd=5, col='orange')



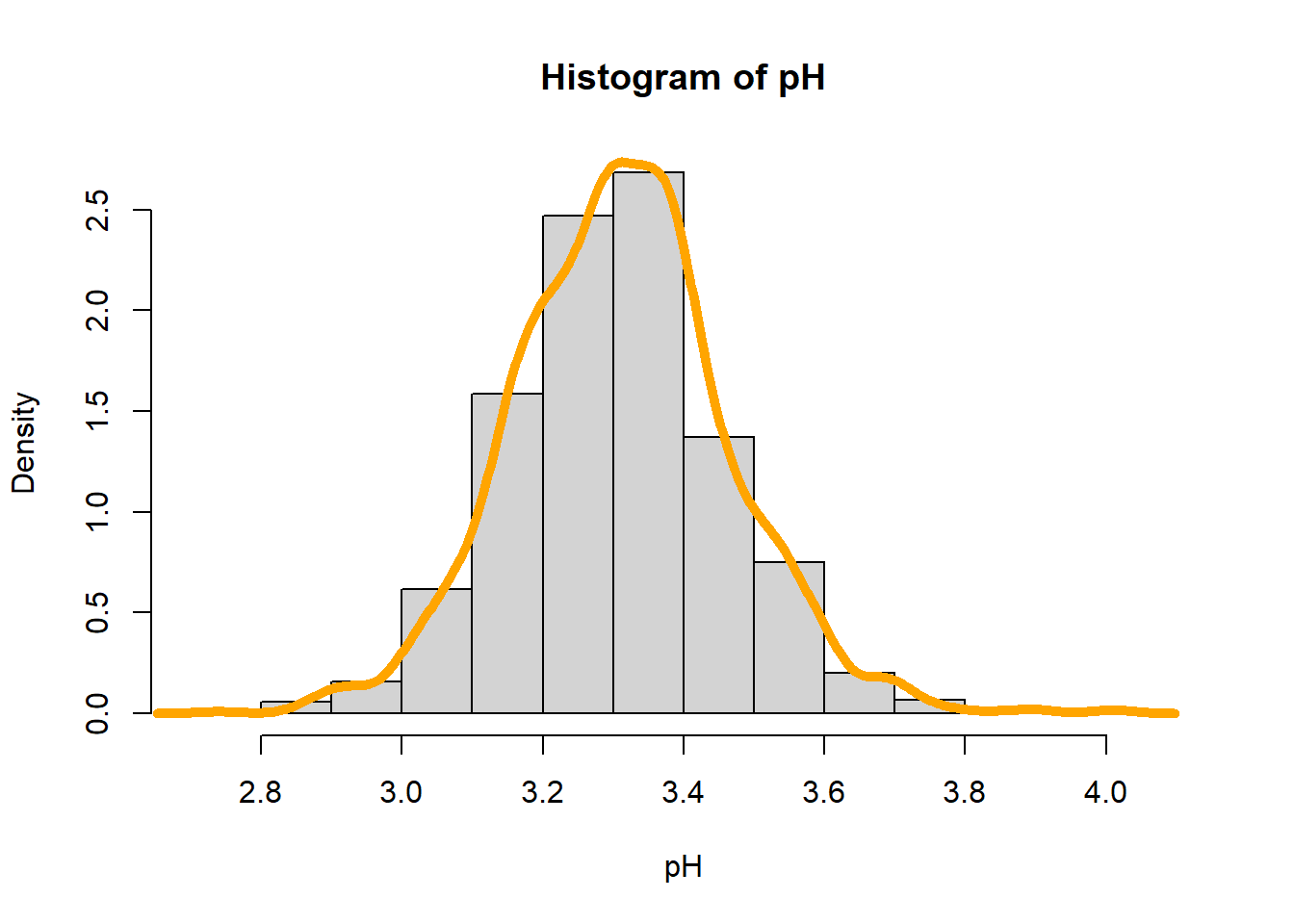
hist(density, freq = FALSE,ylim= c())

lines(density(density), lwd=5, col='orange')

****

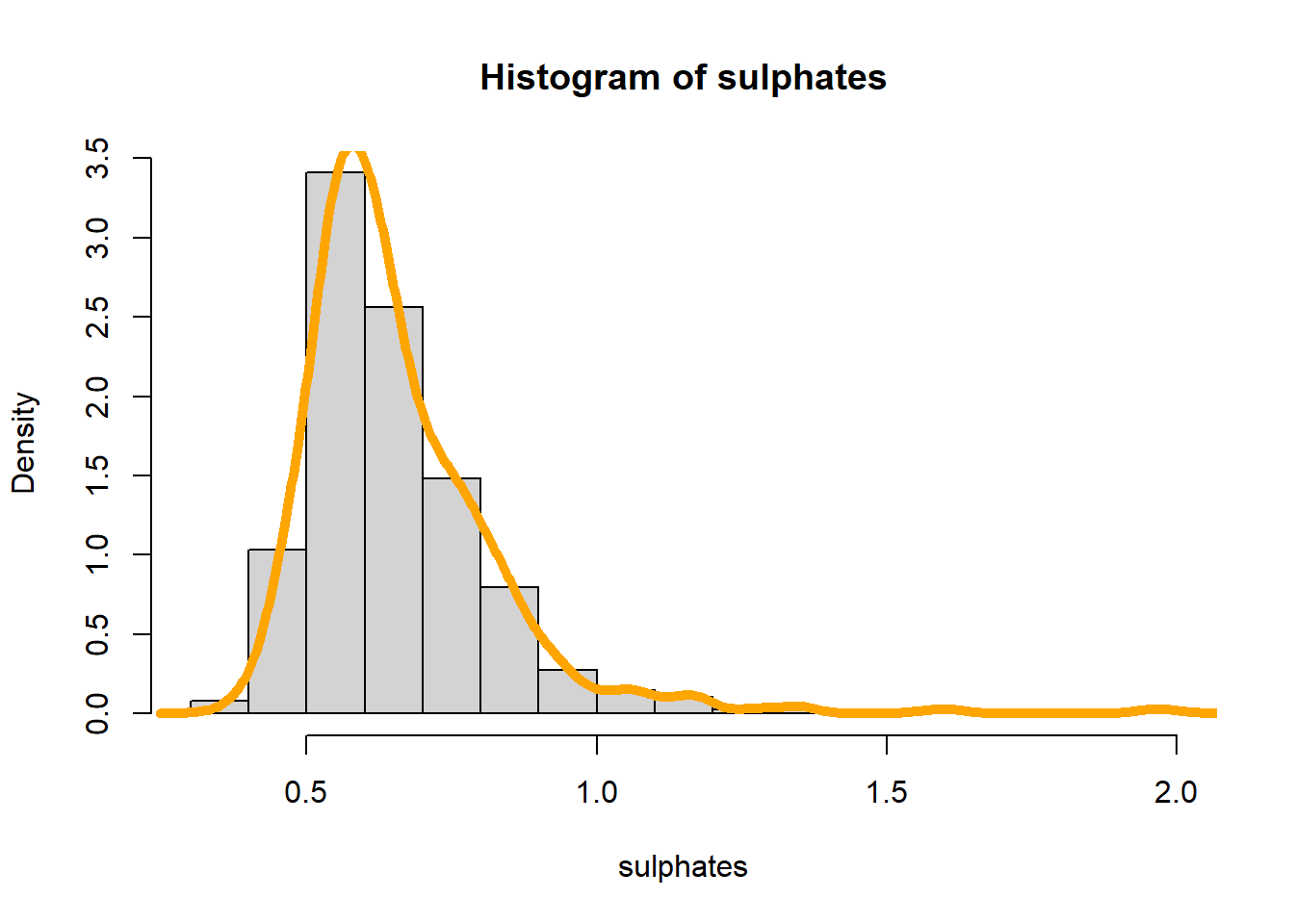
hist(pH, freq = FALSE,ylim= c())

lines(density(pH), lwd=5, col='orange')



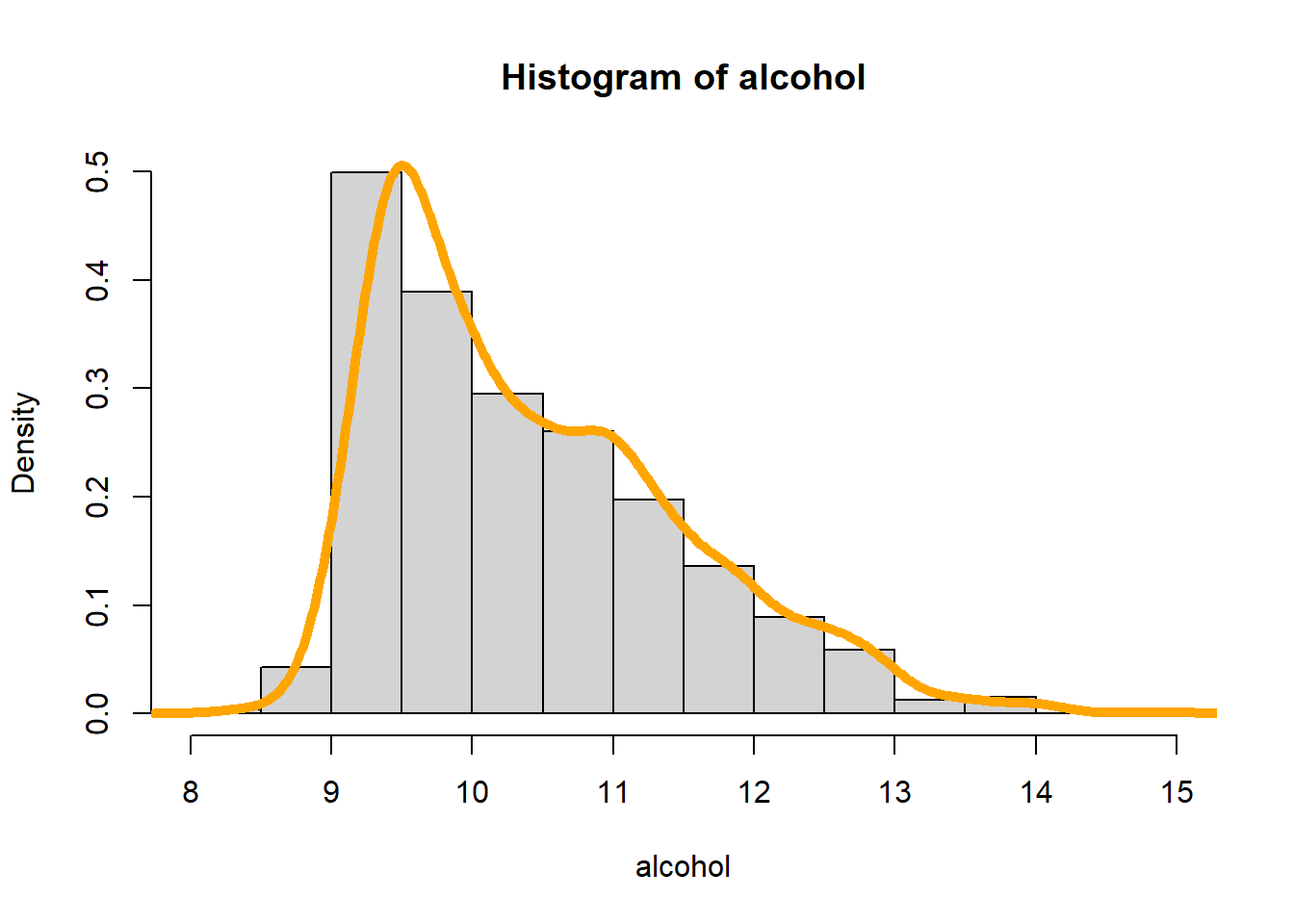
hist(sulphates, freq = FALSE,ylim= c())

lines(density(sulphates), lwd=5, col='orange')



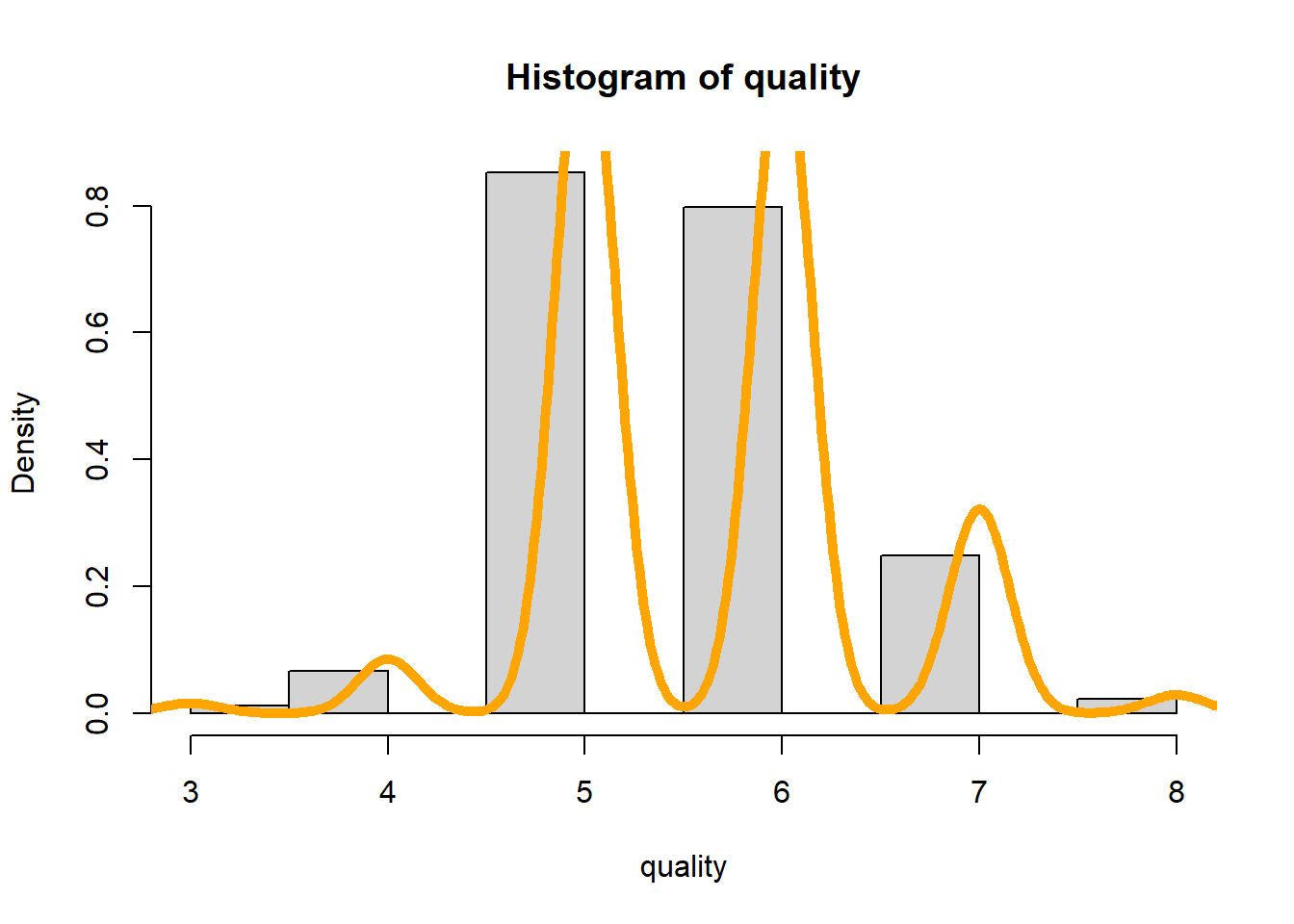
hist(alcohol, freq = FALSE,ylim= c())

lines(density(alcohol), lwd=5, col='orange')



hist(quality, freq = FALSE,ylim= c())

lines(density(quality), lwd=5, col='orange')



**Final Observations**:

1. When we initially measured the summary for dataset, we can identify the values of mean and median for each data fields.
2. With the help of visualization of boxplots and histogram plots, we were able to plot each data fields with its individual outliers and distribution and we were able to identify that the distribution of overall dataset is right skewed.
3. We can observe that the difference in mean and median is highest in total.sulfur.dioxide when compared to other fields.
4. Also calculated the variance of each data field from the dataset.
5. I have also measured the density() to draw the curve for distance of every field in dataset and it can be observed and confirmed that the distribution is right skewed.
6. With the help of boxplot and histogram plots we can identify that the Interquartile range(IOR) is not at the centre for the datasets and hence the nature of it is asymmetric.
7. With the help of measuring the factors of alcohol, we determine the different levels of acid, sugar, quality, and alcohol percentages used in red wine from our dataset.