Wine Quality Analysis

Gunjan Shah

Analysis of Wine datasets to figure out the features that affects the wine quality, color, etc.

find the dataset here - https://archive.ics.uci.edu/ml/machine-learning-databa wine-quality/

Step 1:Initial Analysis - Let's see what we are dealing with!

We have red and white wine Datasets and we will merge it together

```
red <- read.csv("winequality-red.csv",sep = ";",header = TRUE)
dim(red)
```

```
## [1] 1599   12
```

```
str(red)
```

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

```
red$color <- 0
#Assigning 0 to red

white <- read.csv("winequality-white.csv",sep = ";",header = TRUE)
dim(white)
```

```
## [1] 4898   12
```

```
str(white)
```

```
## 'data.frame':    4898 obs. of  12 variables:
##  $ 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 ...
```

```
white$color <-1
#Assigning 1 to white

wine <- rbind(red,white)
wine$color <- factor(wine$color)
dim(wine)
```

```
## [1] 6497   13
```

```
str(wine)
```

```
## 'data.frame':    6497 obs. of  13 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 ...
##  $ color               : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
```

Each dataset has 12 features such as quality of wine, alcohol level,sugar and so on. All the features have continues values except the quality of wine which is discrete variable bounded in the range 3 to 9(more the better). The red wine dataset has 1599 samples and white wine dataset has 4899 samples.

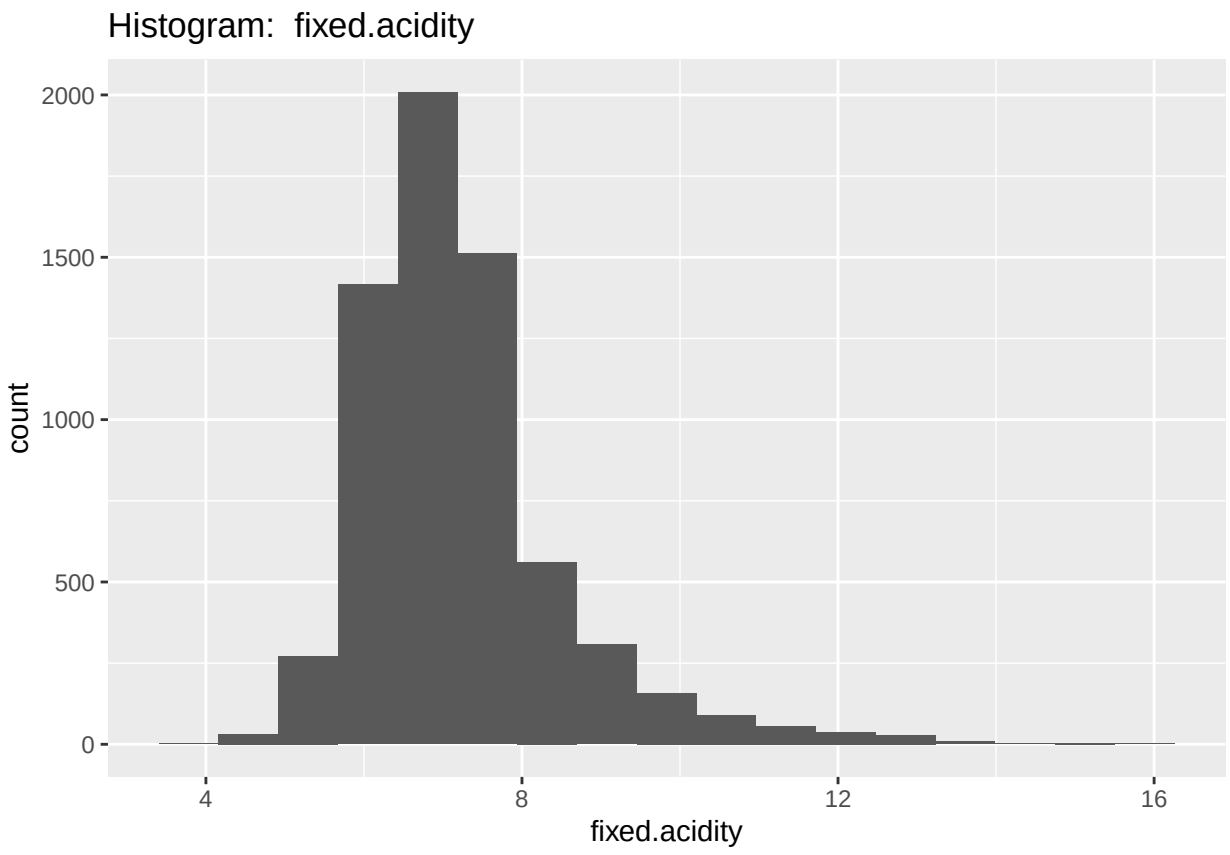
Analysing the features

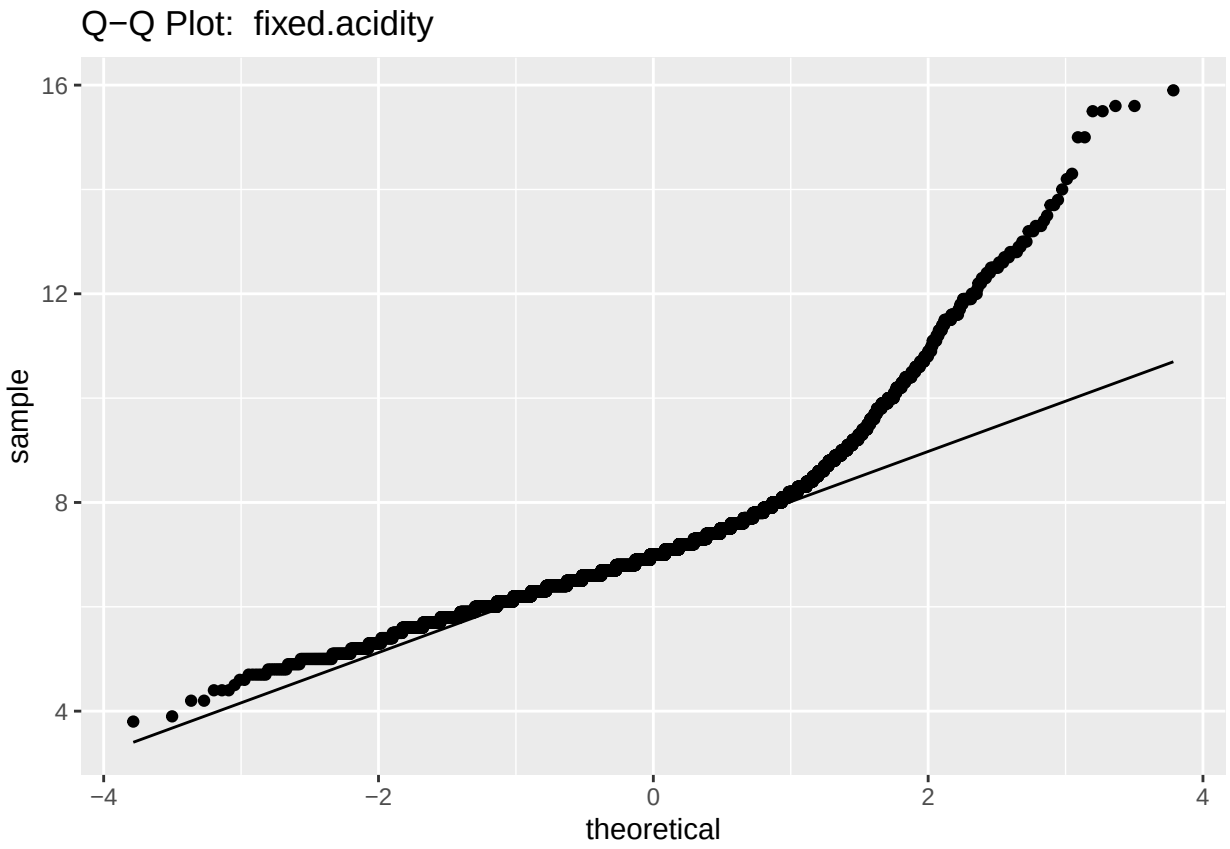
Attribute info function helps to analyze the individual feature's properties such as min, max, mean, histogram, Q-Q plot, Box-plot etc and this will help to see if the feature has normal distribution or not.

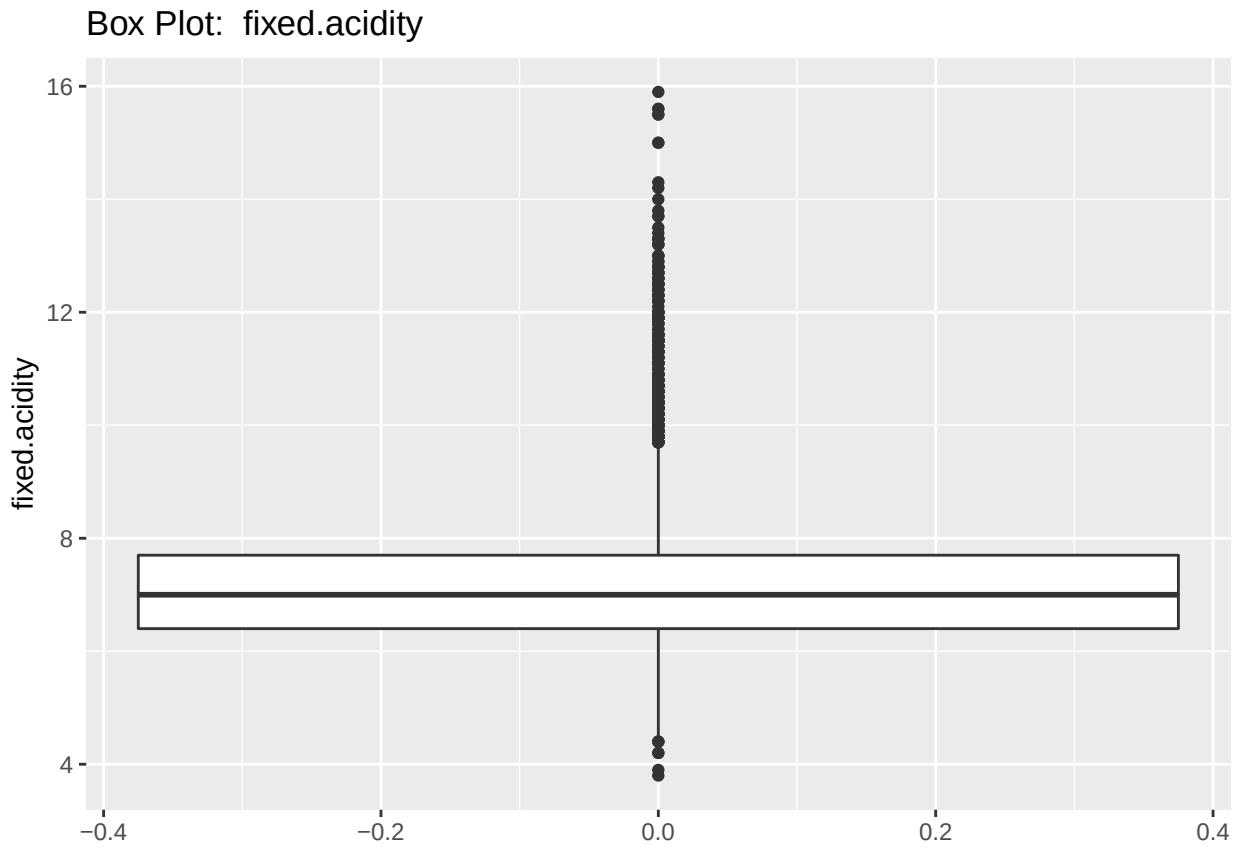
```r
attribute_info <- function(cl,bin_size=30)
{
print(ggplot(data = wine, aes(x=wine[[cl]])) + geom_histogram(bins=bin_size) +
  labs(title = paste("Histogram: ",cl))+xlab(cl))

qqplot <- ggplot(wine, aes(sample=wine[[cl]]))
print(qqplot +geom_qq() +geom_qq_line() + labs(title = paste("Q-Q Plot: ",cl)))
print(ggplot(data = wine, aes(y=wine[[cl]])) + geom_boxplot() +
  labs(title = paste("Box Plot: ",cl))+ylab(cl))

print(describe(wine[[cl]]))
m <- mean(wine[[cl]])
s <- sd(wine[[cl]])
z_score <- (wine[[cl]] - m) / s

print(paste0("Percent of values outside 2% standard deviation: ",
      round(length(z_score[abs(z_score) > 2])/length(z_score)*100,4)))

}
```

1.1.fixed.acidity

```r
attribute_info("fixed.acidity",17)
```




Q−Q Plot: fixed.acidity
16
12
8
4
sample
−4
−2
0
2
4
theoretical



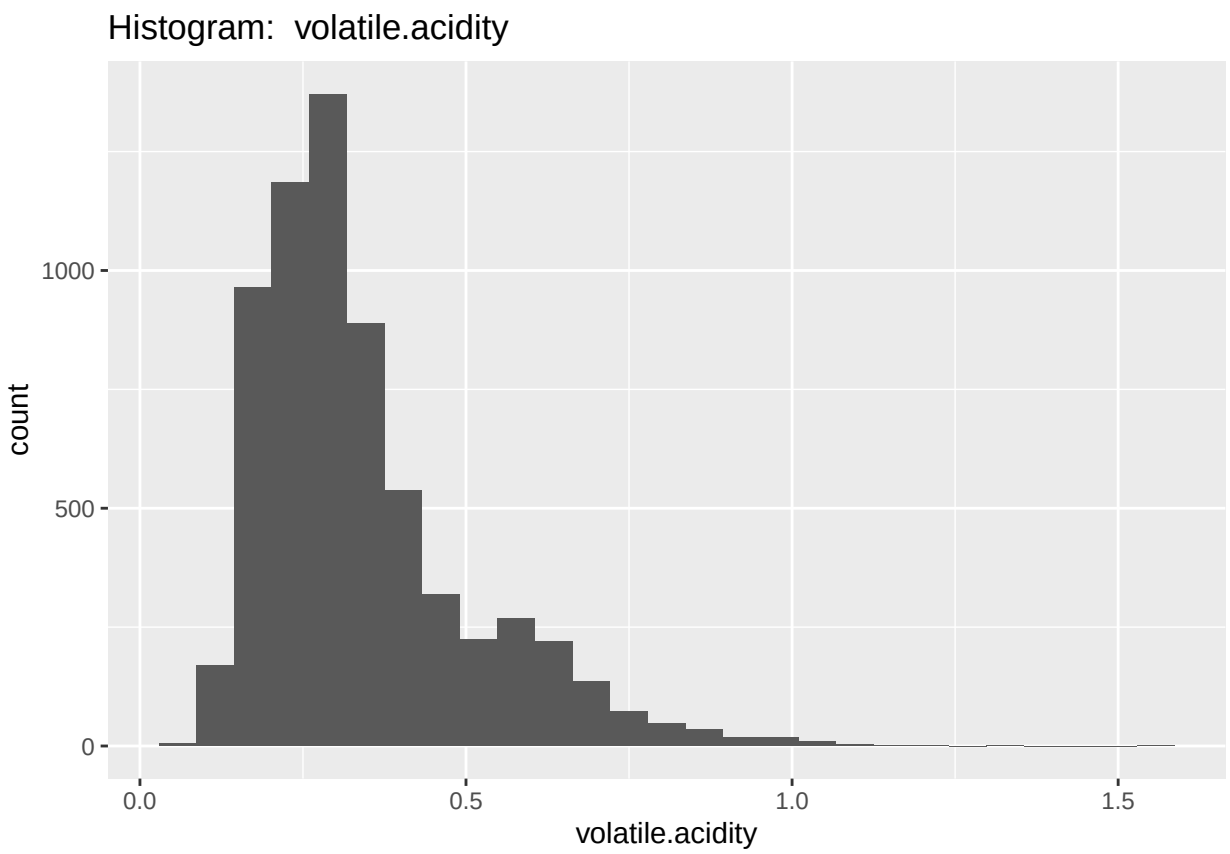

```
##    vars    n mean  sd median trimmed  mad min  max range skew kurtosis
## X1    1 6497 7.22 1.3      7    7.06 0.89 3.8 15.9  12.1 1.72     5.05
##      se
## X1 0.02
## [1] "Percent of values outside 2% standard deviation: 4.9869"
```

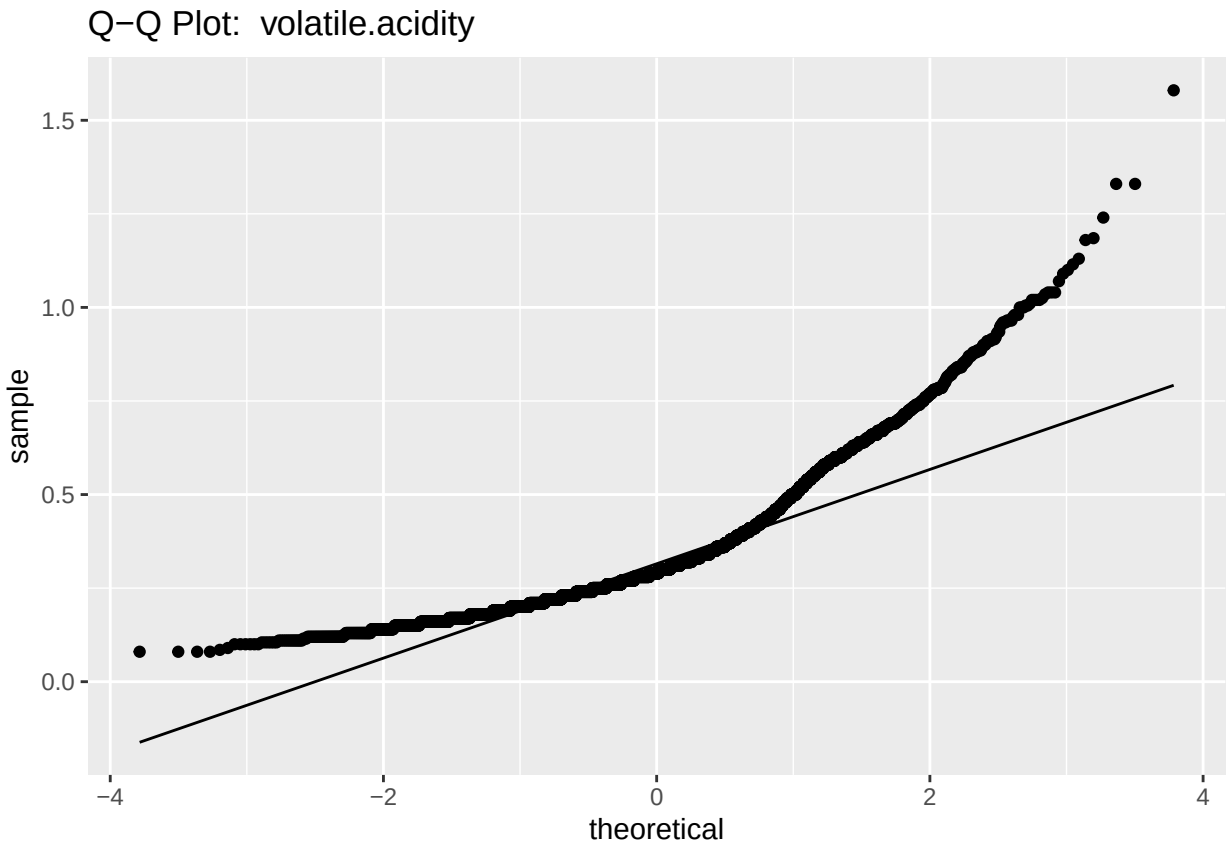
fixed acidity is in the range of 3.8 to 15.9. The Q-Q plot show that the line sags consistently consistently rises above it, then this shows that the kurtosis differs from a normal distribution.

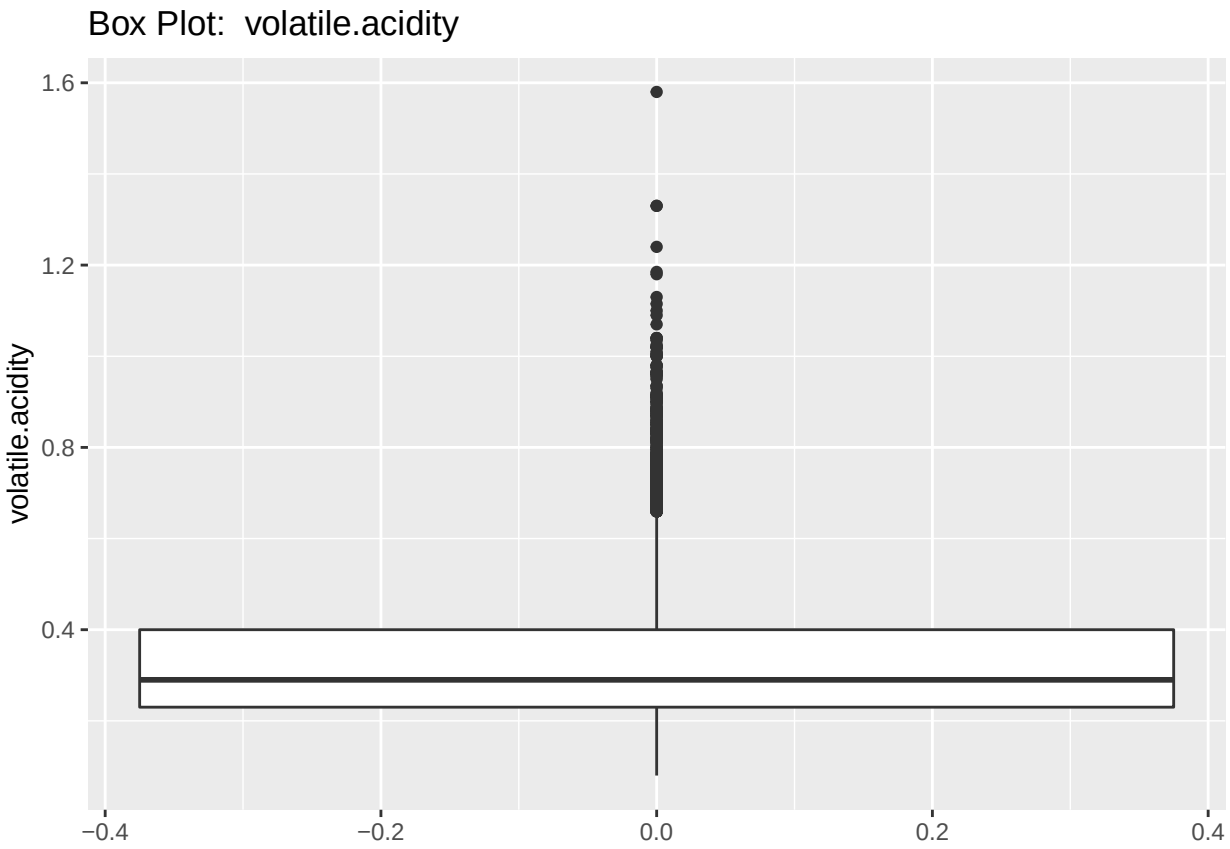
The fixed acidity has 4.99% values outside 2 standard deviation which is acceptable (we expect 4.6% of values to be outside 2 standard deviation). Hence, the data don't have any outliers.

1.2 volatile.acidity

```
attribute_info("volatile.acidity",27)
```




Q−Q Plot: volatile.acidity
sample
theoretical
1.5
1.0
0.5
0.0
−4
−2
0
2
4

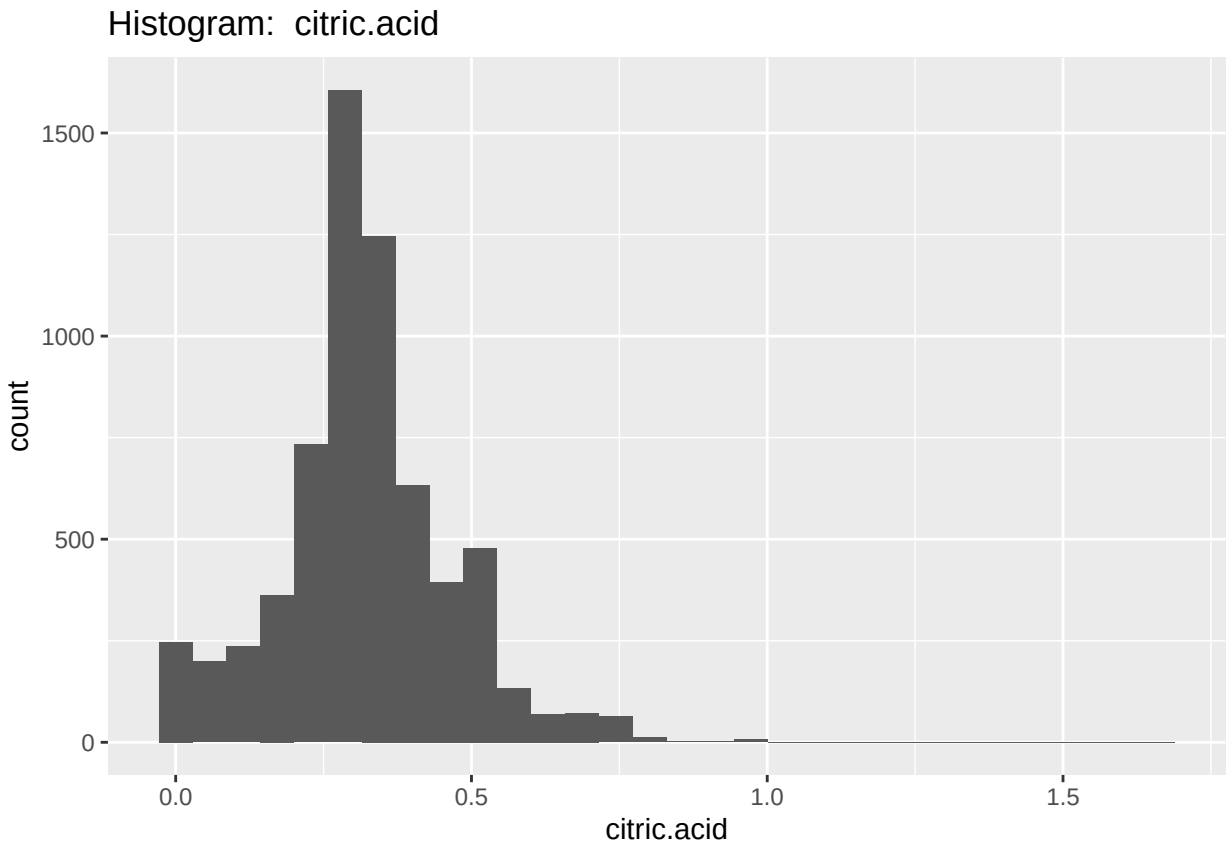


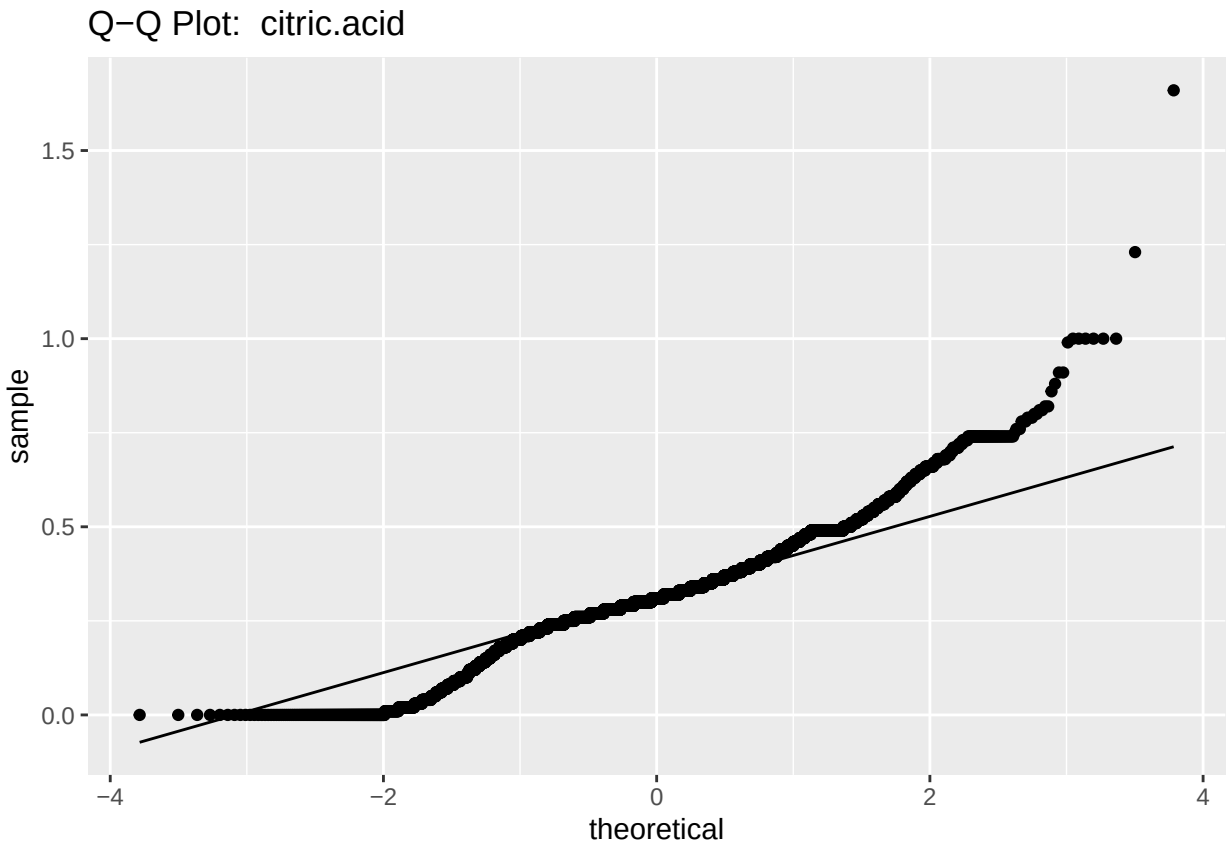

```
##    vars    n mean   sd median trimmed  mad  min  max range skew kurtosis
## X1    1 6497 0.34 0.16   0.29    0.32 0.12 0.08 1.58   1.5 1.49     2.82
##    se
## X1  0
## [1] "Percent of values outside 2% standard deviation: 5.2948"
```

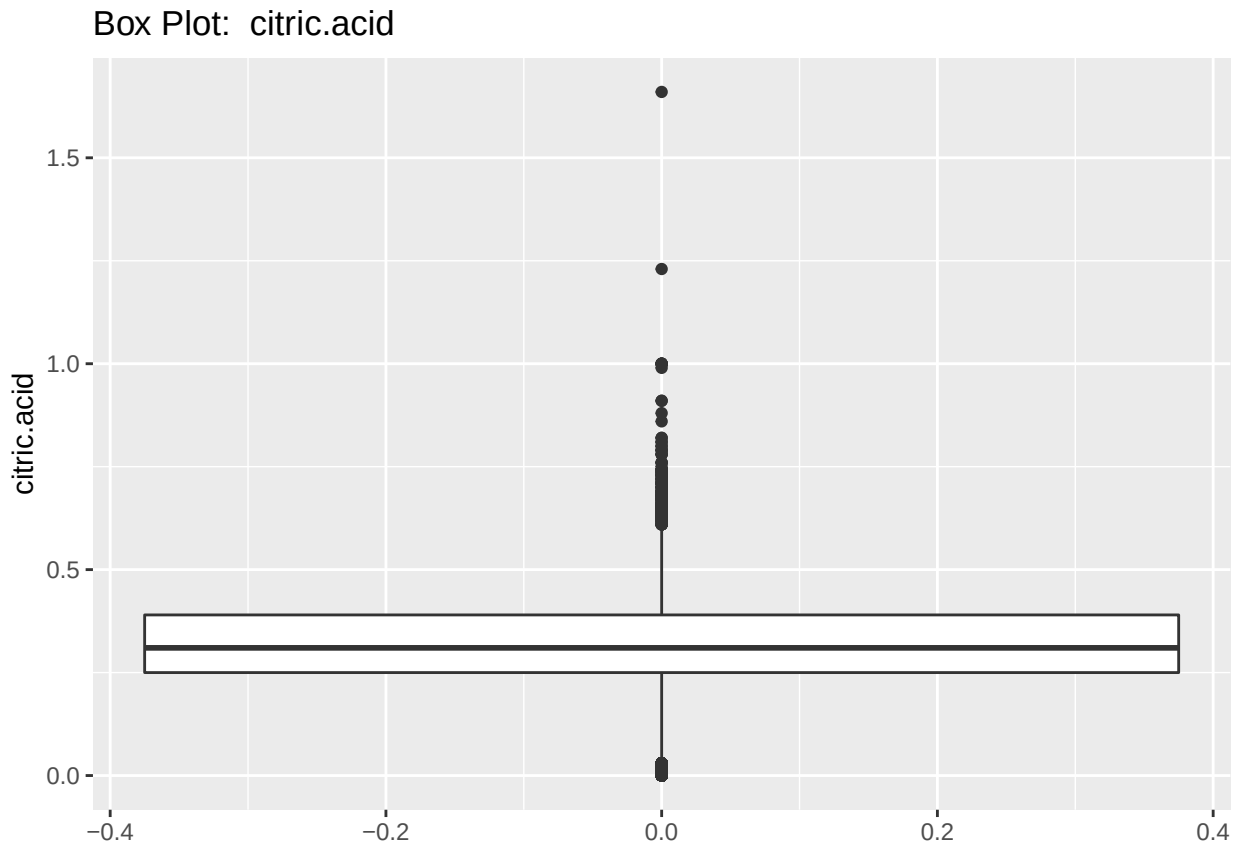
The volatile acidity is in the range of 0.08 to 1.59. The Q-Q plot show that the data doesn't follow the diagonal line. Hence, it differs from a normal distribution.

Volatile acidity has 5.30% values outside 2 standard deviation which is acceptable (we expect 4.6% of values to be outside 2 standard deviation). Hence, the data don't have any outliers. # 1.3 citric.acid

```
attribute_info("citric.acid")
```


Histogram: citric.acid
1500
1000
500
0
count
0.0
0.5
1.0
1.5
citric.acid


Q−Q Plot: citric.acid
sample
theoretical
1.5
1.0
0.5
0.0
−4
−2
0
2
4




```
##    vars    n mean   sd median trimmed mad min  max range skew kurtosis se
## X1    1 6497 0.32 0.15   0.31    0.32 0.1   0 1.66  1.66 0.47     2.39  0
## [1] "Percent of values outside 2% standard deviation: 7.3419"
```

```
cl <- "citric.acid"
print(describe(wine[[cl]]))
```

```
##    vars    n mean   sd median trimmed mad min  max range skew kurtosis se
## X1    1 6497 0.32 0.15   0.31    0.32 0.1   0 1.66  1.66 0.47     2.39  0
```

```
m <- mean(wine[[cl]])
s <- sd(wine[[cl]])
z_score <- (wine[[cl]] - m) / s

print(paste0("Percent of values outside 3% standard deviation: ",
      round(length(z_score[abs(z_score) > 3])/length(z_score)*100)))
```

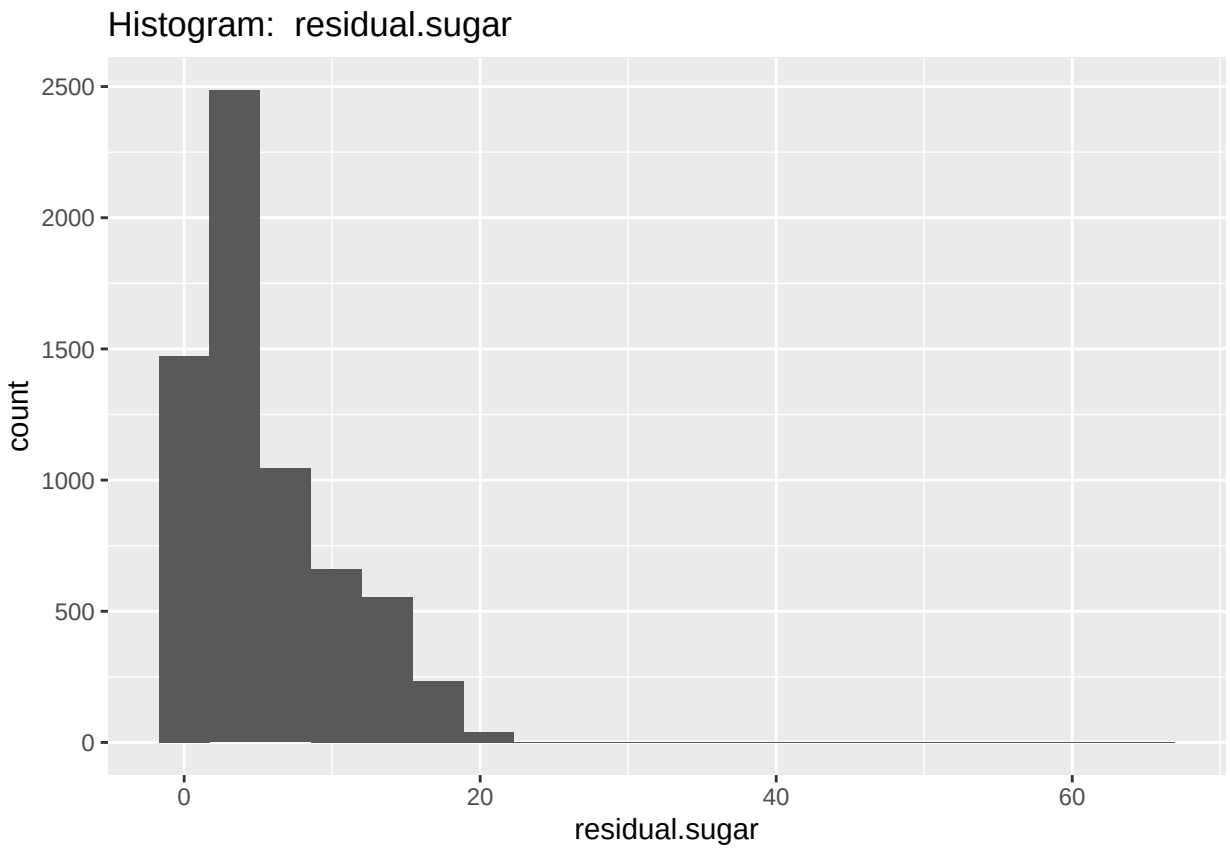
```
## [1] "Percent of values outside 3% standard deviation: 0"
```

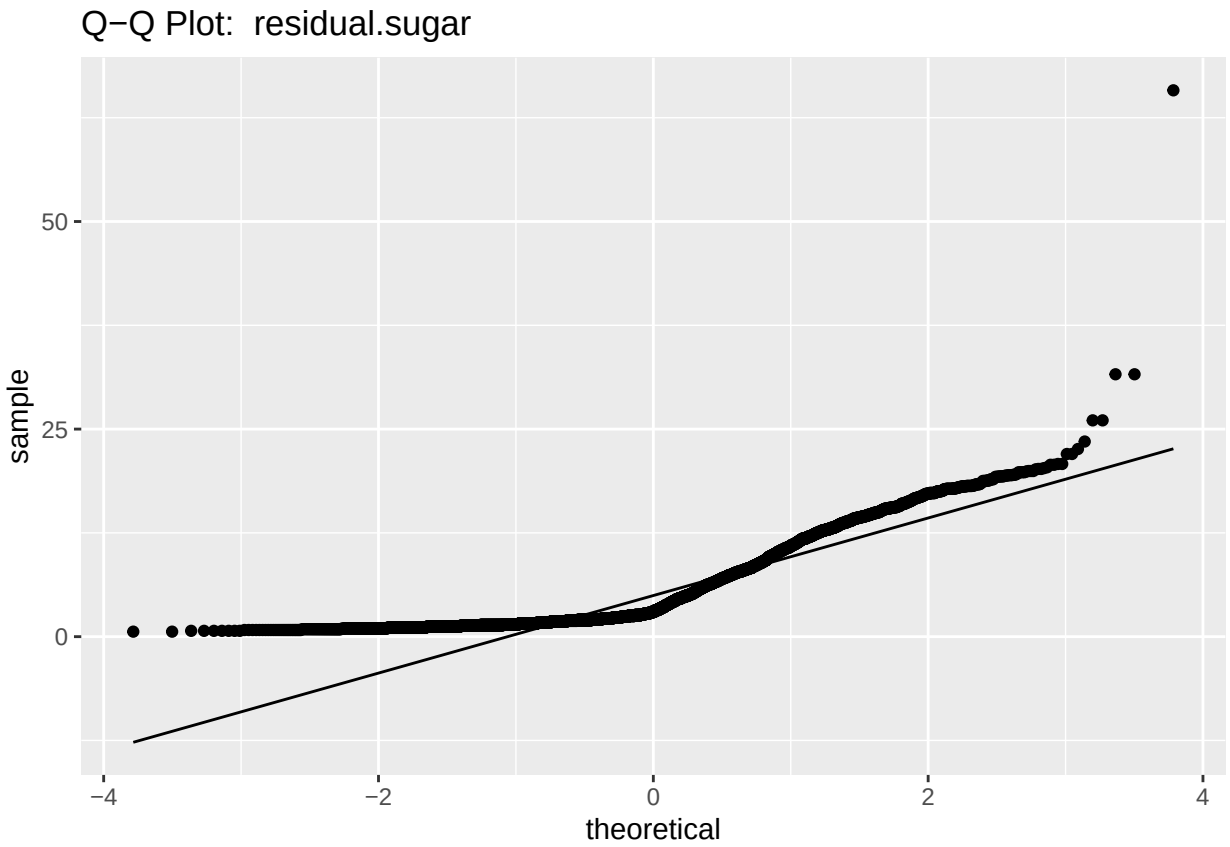
The citric acid is in the range of 0 to 1.66. The Q-Q plot show that the data is not normal.

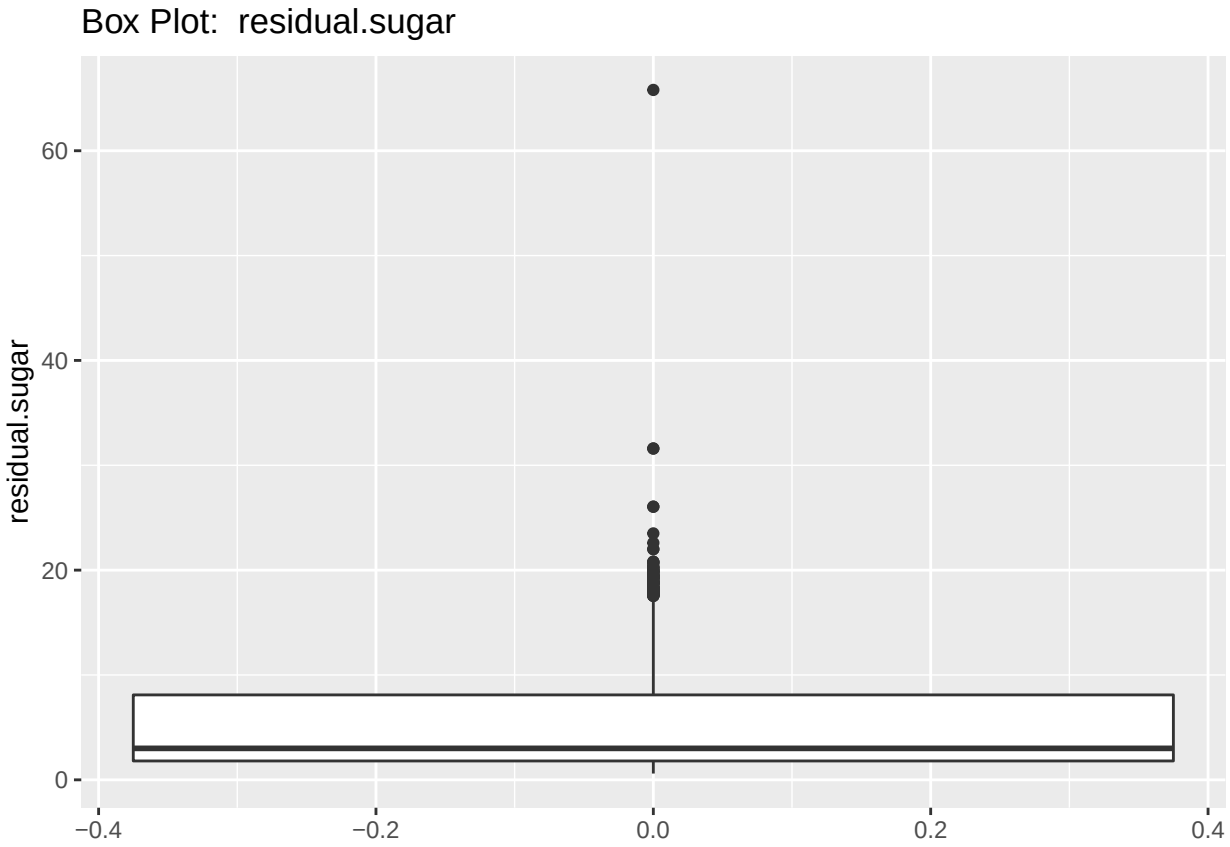
The citric acidity has 7.29% values outside 2 standard deviation and 0% values outside 3 standard deviation. By looking at the histogram it is evident that majority of citric acid's value are clustered in small range. All the values are valid values and hence, we accept all the data.

1.4 residual.sugar

```
attribute_info("residual.sugar",20)
```




Q−Q Plot: residual.sugar
sample
50
25
0
−4
−2
0
2
4
theoretical




```
##    vars    n mean   sd median trimmed  mad min  max range skew kurtosis
## X1    1 6497 5.44 4.76      3     4.7 2.52 0.6 65.8  65.2 1.43     4.35
##      se
## X1 0.06
## [1] "Percent of values outside 2% standard deviation: 5.187"
```

```
wine$residual.sugar[wine$residual.sugar>=40]
```

```
## [1] 65.8
```

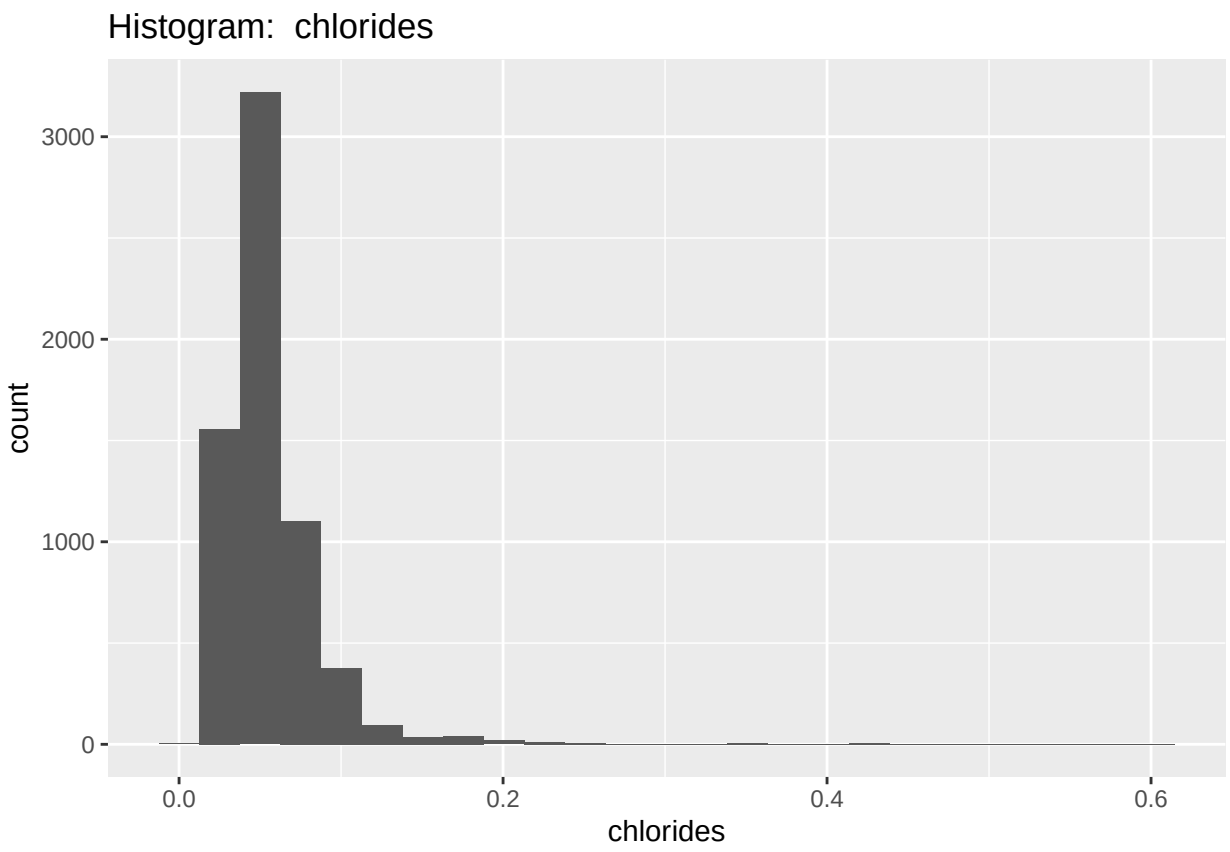
```
wine <- wine[!(wine$residual.sugar>=40),]
```

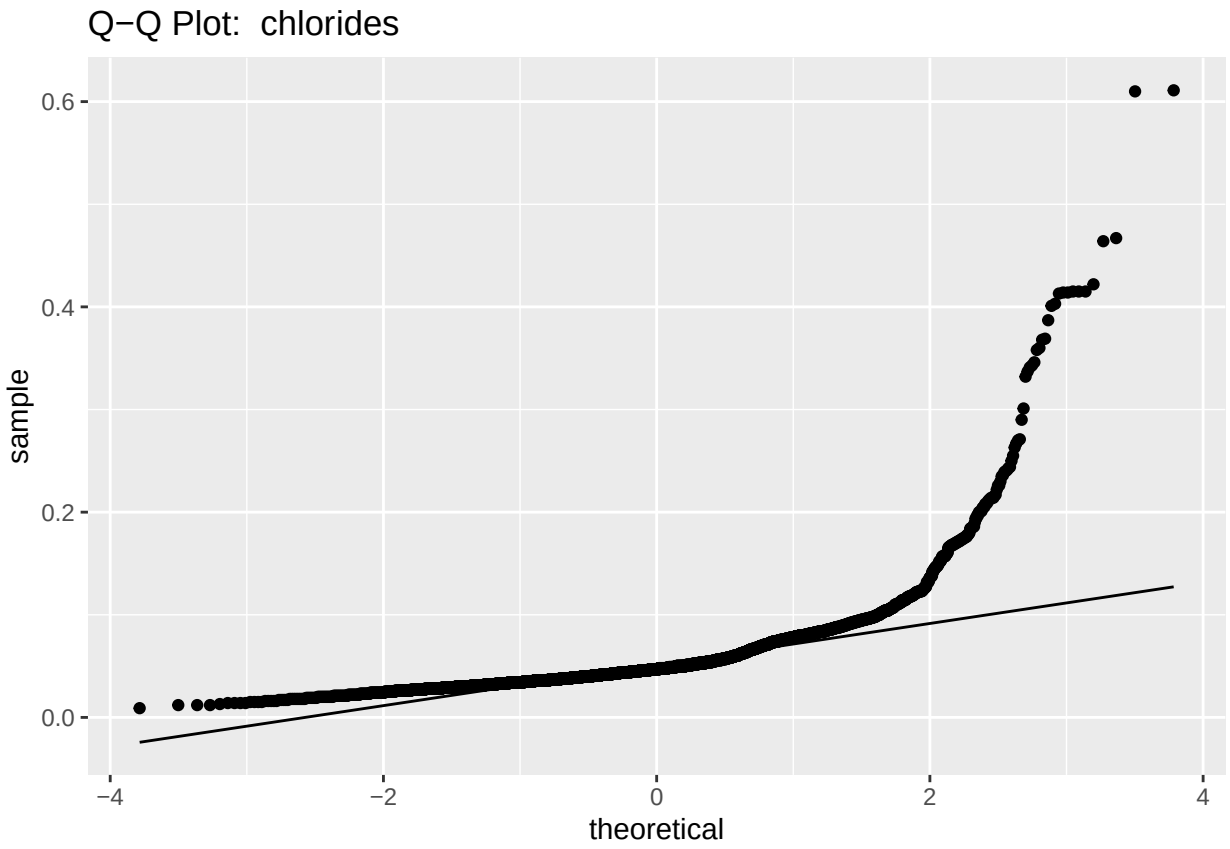
residual sugar seems to vary between 0.6 to 65.8. The Q-Q plot show that the line sags consistently consistently rises above it, this shows that the kurtosis differs from a normal distribution.

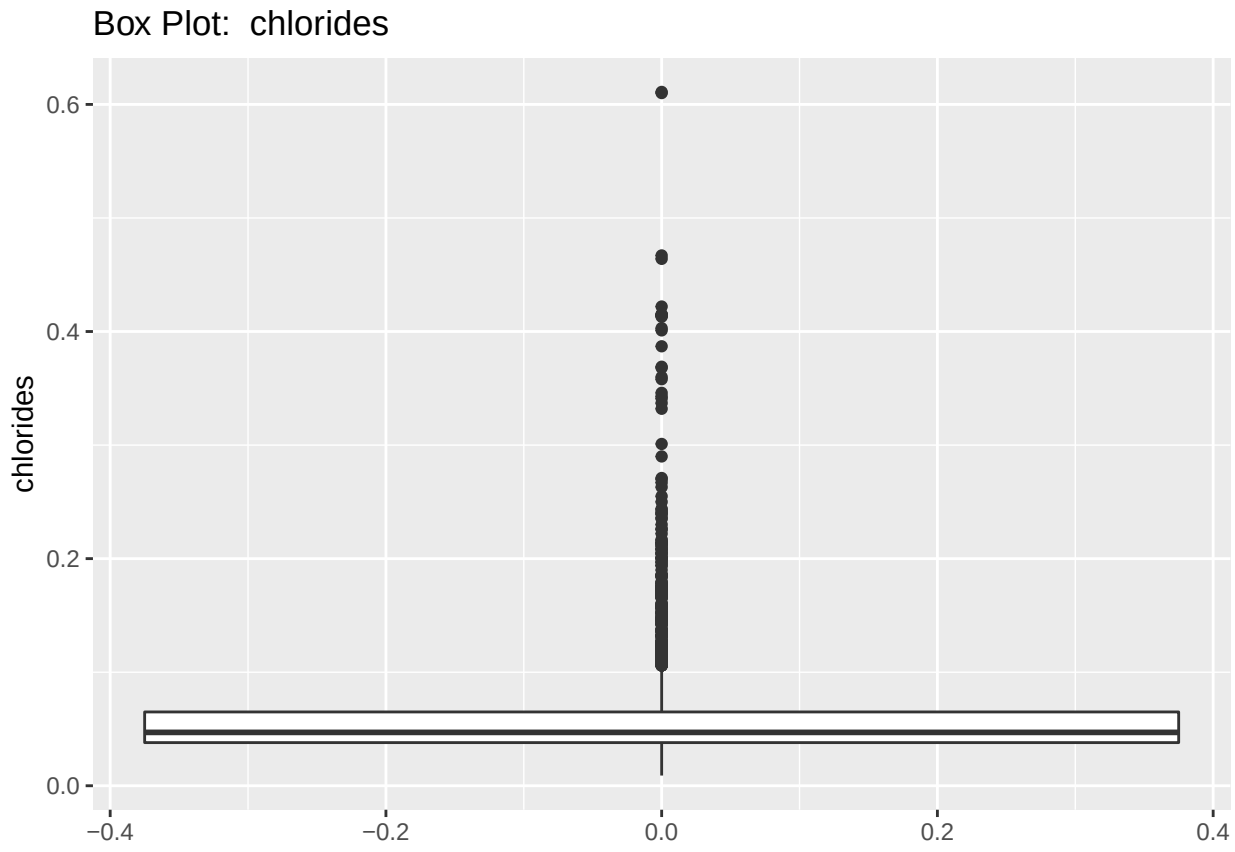
The residual sugar has 5.18% values outside 2 standard deviation which is acceptable (we expect 4.6% of values to be outside 2 standard deviation). However, with mean of 5.44 and standard deviation of 4.76 one point has maximum value of sugar at 65.8. we removed that point.

1.5 chlorides

```
attribute_info("chlorides",25)
```




Q−Q Plot: chlorides
0.6
0.4
0.2
0.0
sample
−4
−2
0
2
4
theoretical



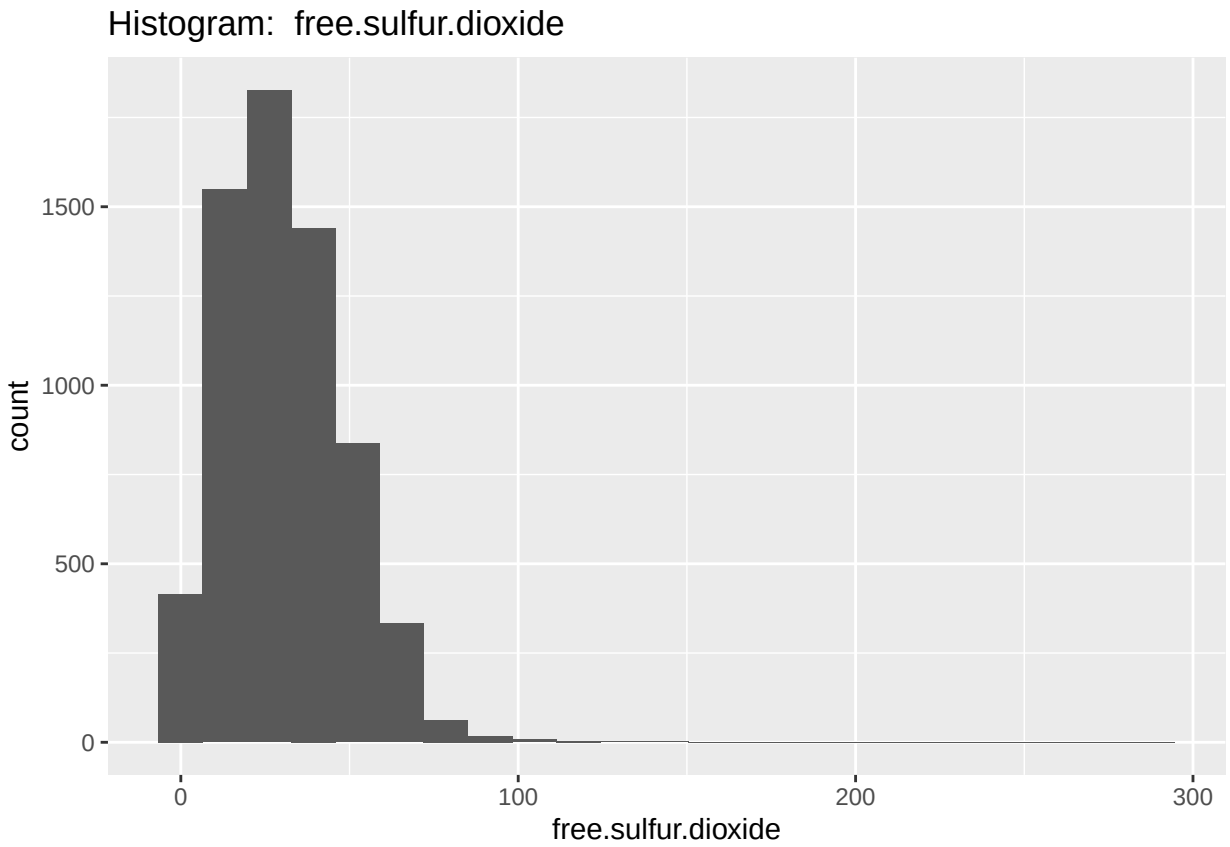

```
##    vars    n mean   sd median trimmed  mad  min  max range skew kurtosis
## X1    1 6496 0.06 0.04   0.05    0.05 0.02 0.01 0.61   0.6  5.4    50.84
##    se
## X1  0
## [1] "Percent of values outside 2% standard deviation: 2.4784"
```

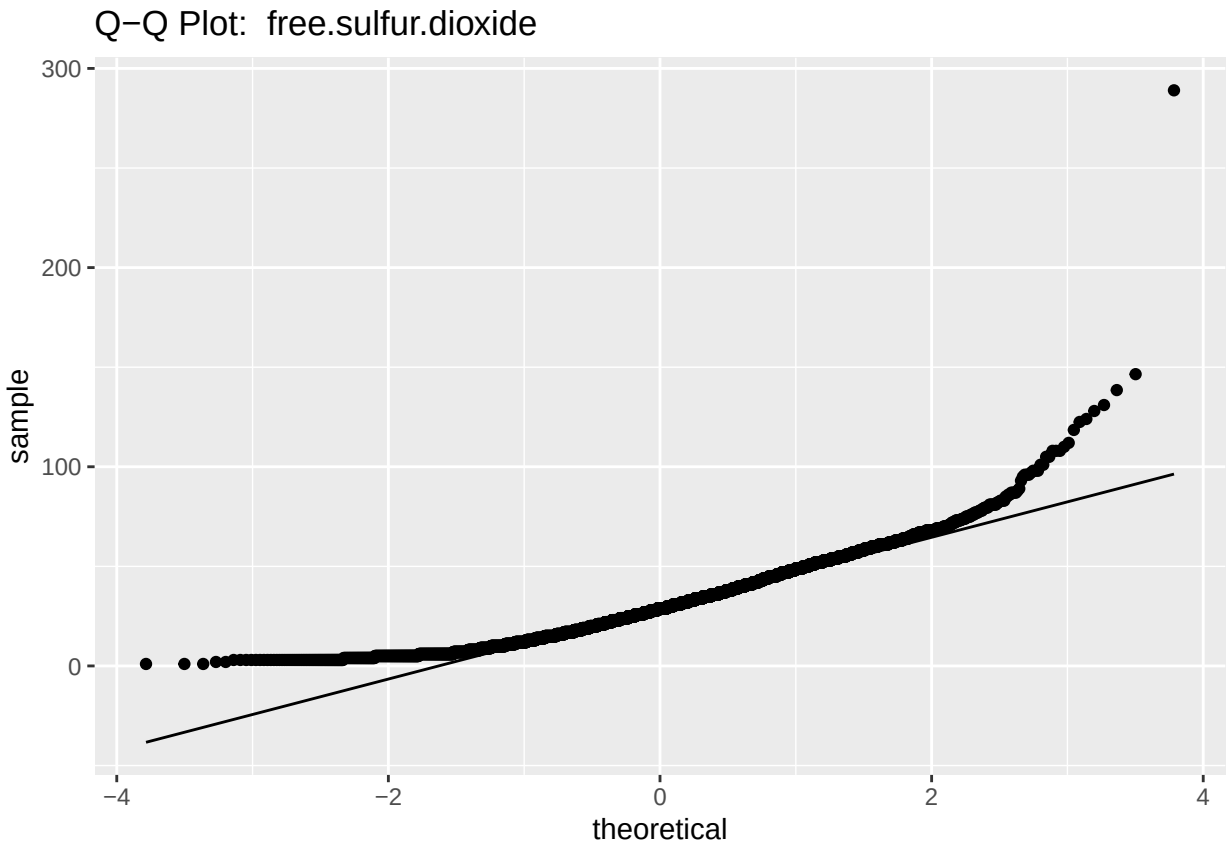
The chlorides is in the range of 0.01 to 0.61 The Q-Q plot show that the data doesn't follow the diagonal line. Hence, it differs from a normal distribution.

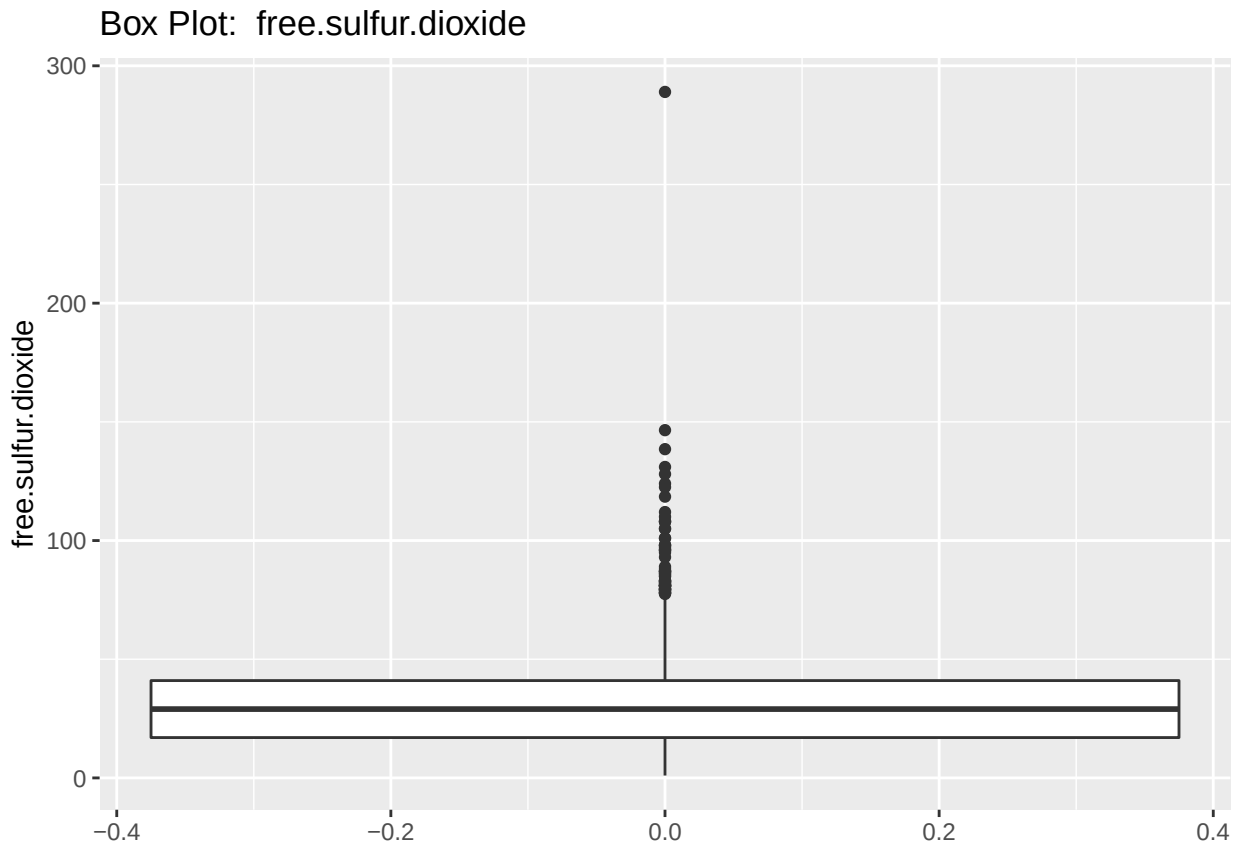
Chlorides has 2.4784% values outside 2 standard deviation which is acceptable (we expect 4.6% of values to be outside 2 standard deviation). Hence, the data don't have any outliers.

1.6 free.sulfur.dioxide

```
attribute_info("free.sulfur.dioxide",23)
```


Histogram: free.sulfur.dioxide
count
1500
1000
500
0
0
100
200
300
free.sulfur.dioxide


Q−Q Plot: free.sulfur.dioxide
300
200
100
0
sample
−4
−2
0
2
4
theoretical




```
##    vars    n  mean    sd median trimmed   mad min max range skew kurtosis
## X1    1 6496 30.53 17.75     29   29.32 17.79   1 289   288 1.22      7.9
##      se
## X1 0.22
## [1] "Percent of values outside 2% standard deviation: 2.8479"
```

```
wine$free.sulfur.dioxide[wine$free.sulfur.dioxide>=150]
```

```
## [1] 289
```

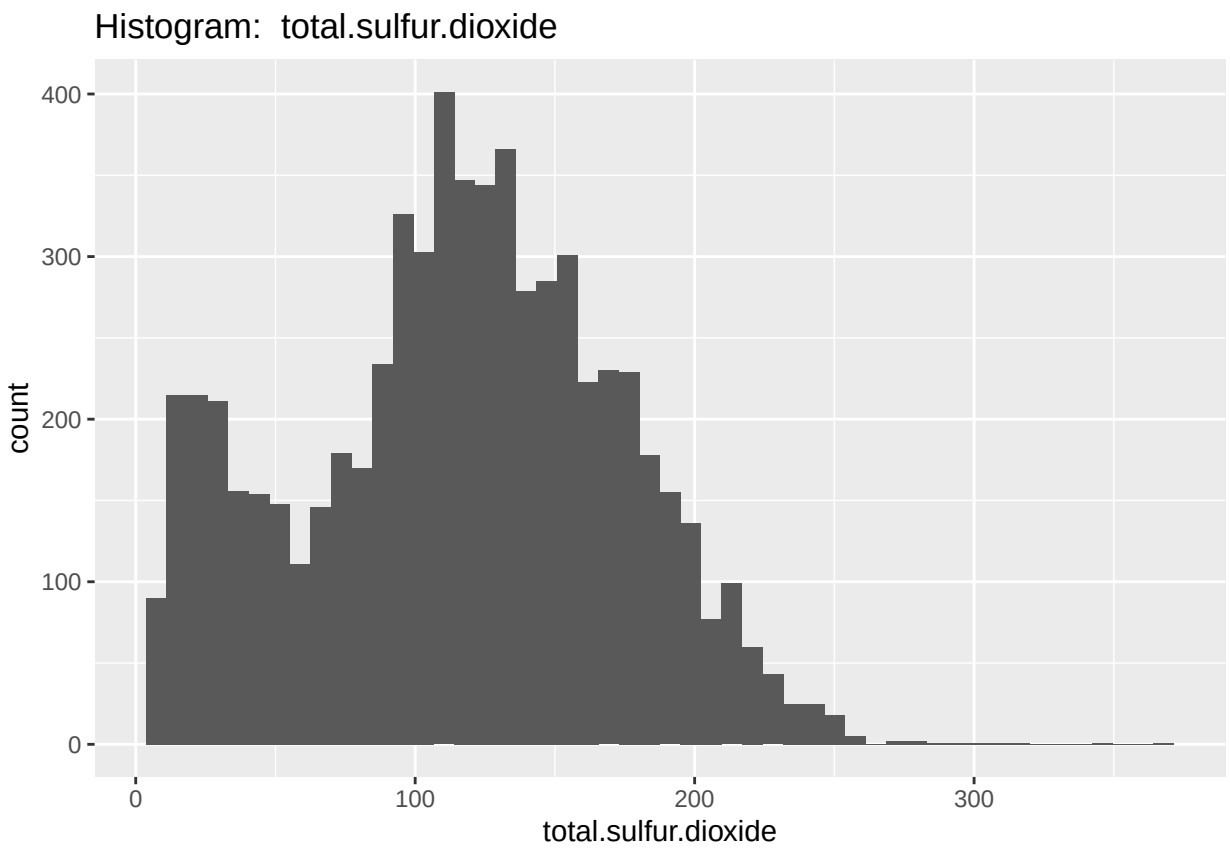
```
wine <- wine[!(wine$free.sulfur.dioxide>=150),]
```

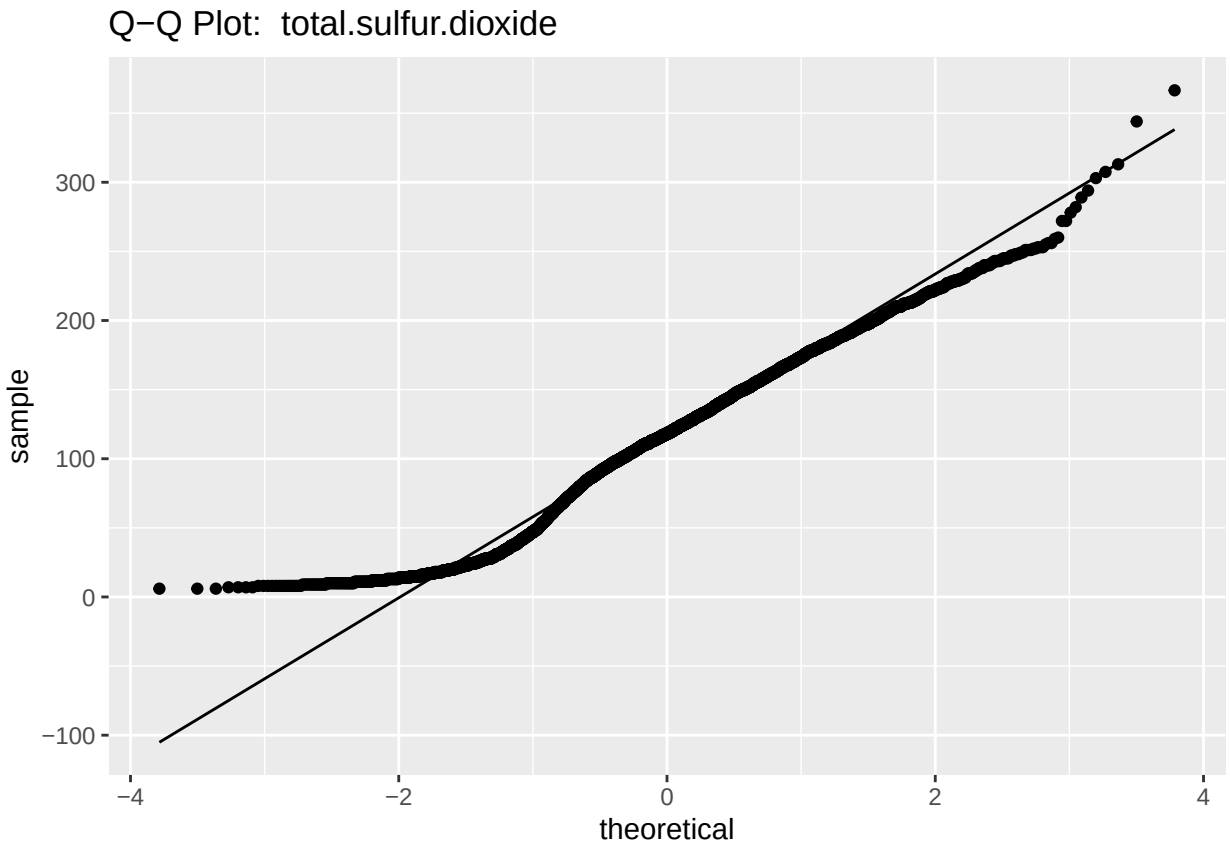
The free sulfur dioxide seems to very between the 1-100 count with peaking around 50 mark. From the Q-Q plot and histogram it looks like normal-distribution.However, one point was around 300. hence, removed that point.

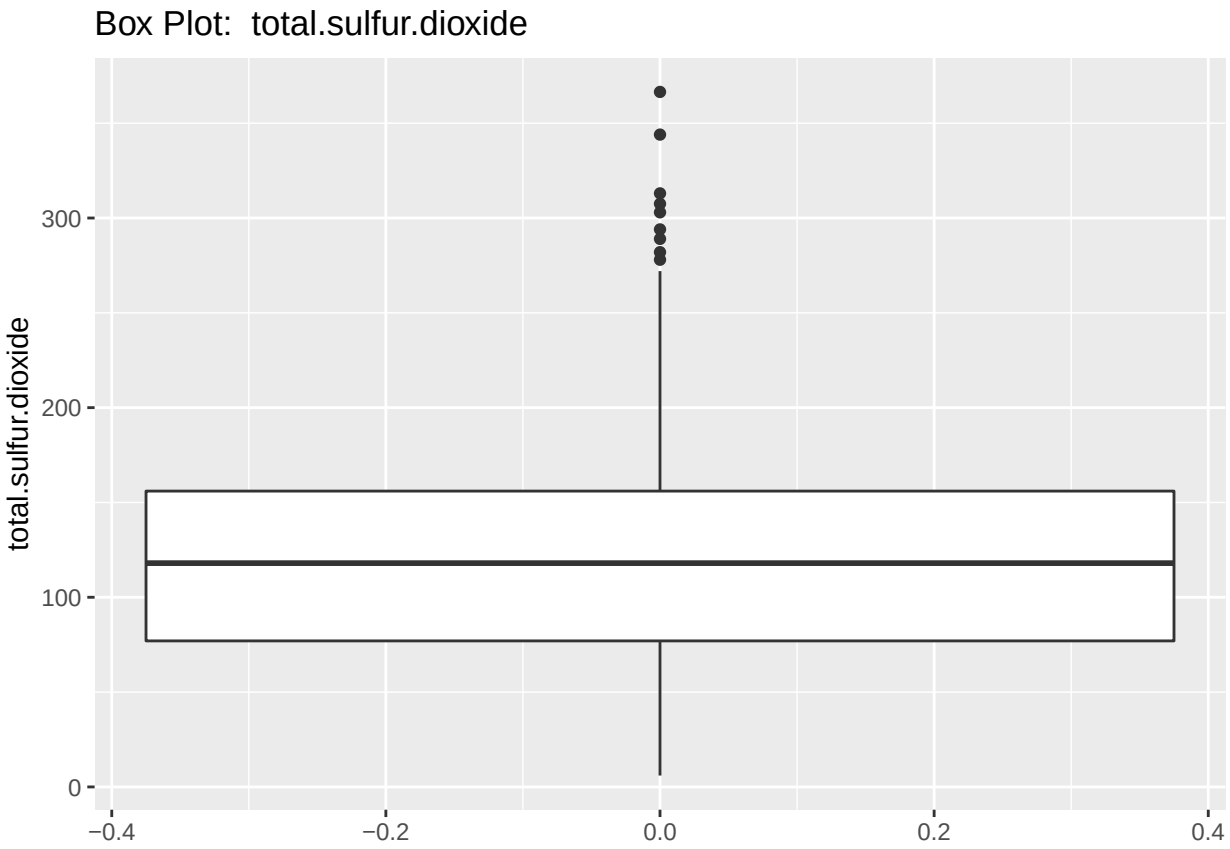
free.sulfur.dioxide has 2.8479% values outside 2 standard deviation which is acceptable (we expect 4.6% of values to be outside 2 standard deviation). Hence, the data don't have any outliers.

1.7 total.sulfur.dioxide

```
attribute_info("total.sulfur.dioxide",50)
```




Q−Q Plot: total.sulfur.dioxide
sample
theoretical
300
200
100
0
−100
−4
−2
0
2
4



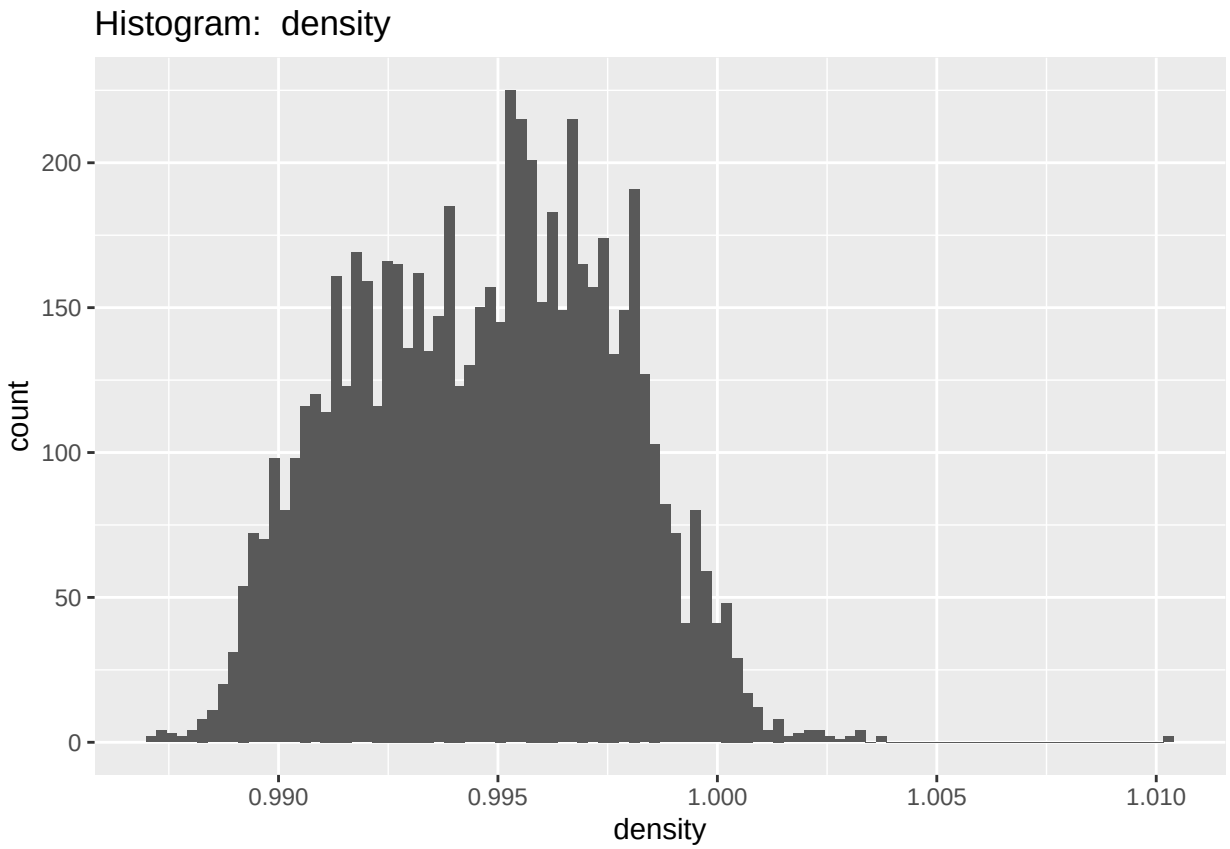

```
##    vars    n   mean    sd median trimmed   mad min   max range  skew
## X1    1 6495 115.69 56.38    118   115.9 57.82   6 366.5 360.5 -0.03
##    kurtosis  se
## X1    -0.52 0.7
## [1] "Percent of values outside 2% standard deviation: 1.6012"
```

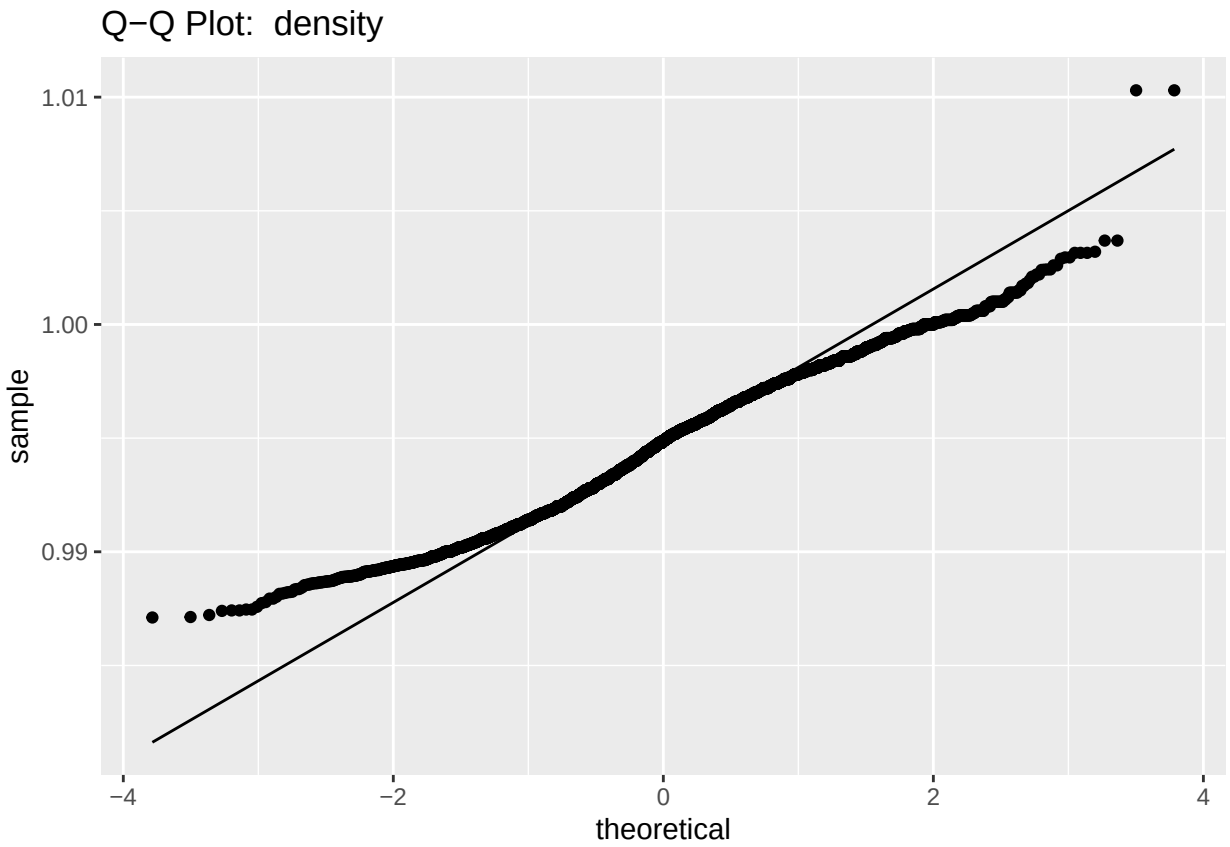
The total sulfur dioxide seems to very between 0 and 300 and exhibiting peak around 150.Moreover, There seems to be another lower peak around the 50 mark. From the Q-Q plot it is evident that it follows the normal distribution.

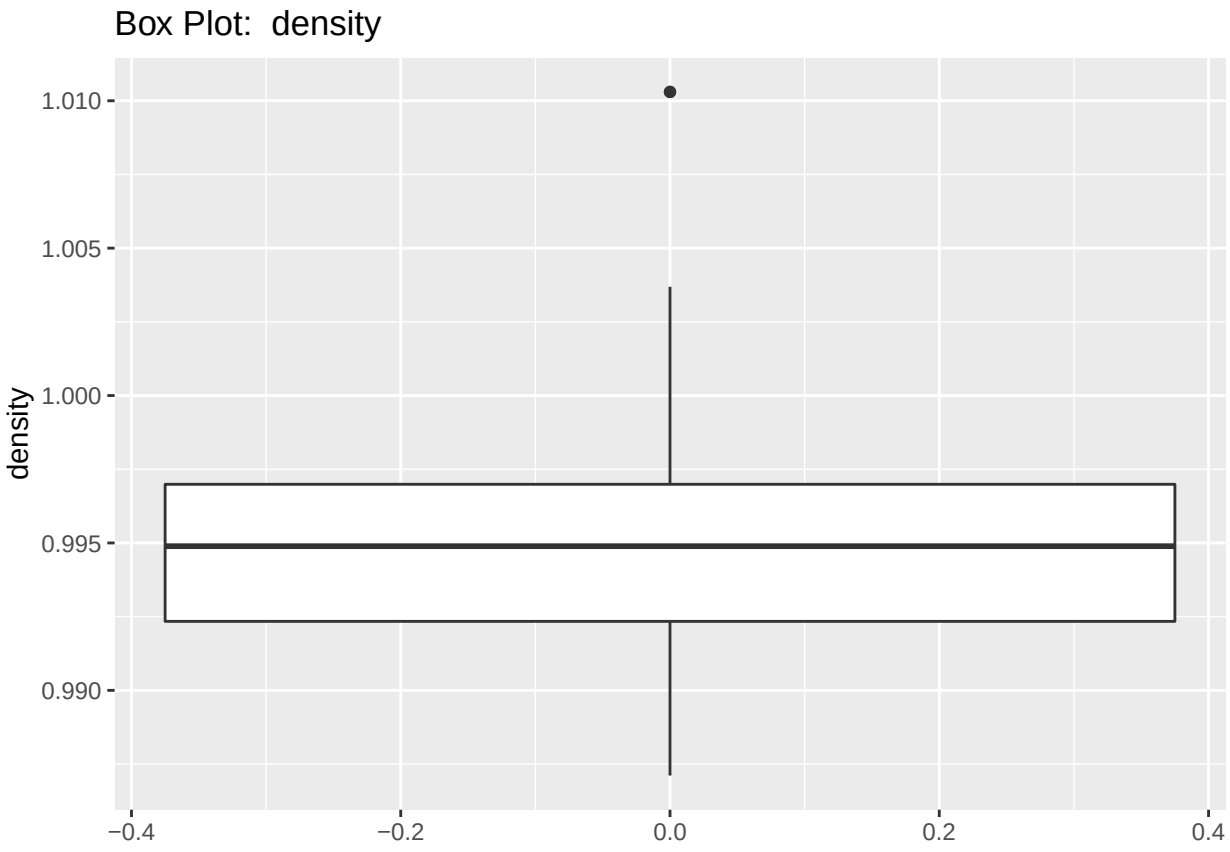
total.sulfur.dioxide has 1.60% values outside 2 standard deviation which is acceptable (we expect 4.6% of values to be outside 2 standard deviation). Hence, the data don't have any outliers.

1.8 density

```
attribute_info("density",100)
```


Histogram: density
count
200
150
100
50
0
0.990
0.995
1.000
1.005
1.010
density


Q−Q Plot: density
1.01
1.00
0.99
sample
−4
−2
0
2
4
theoretical



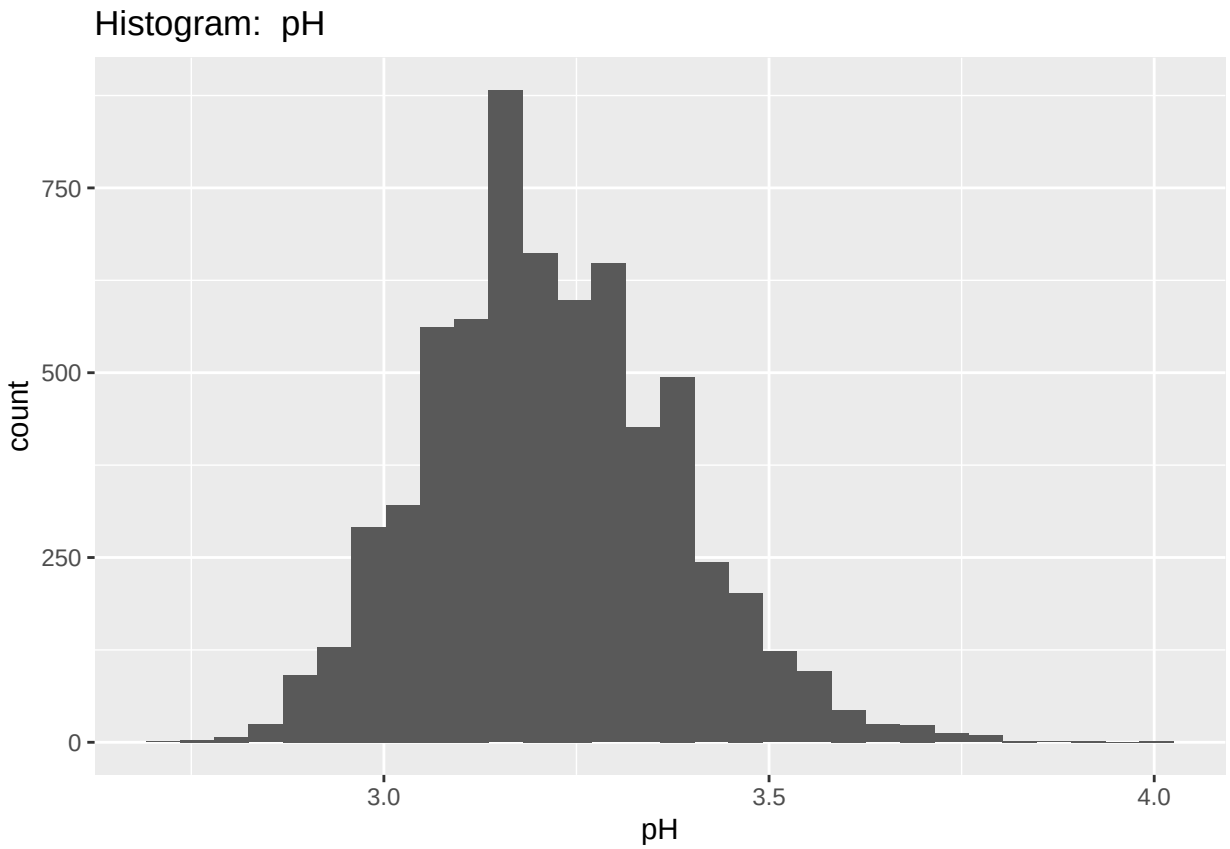

```
##    vars    n mean sd median trimmed mad  min  max range skew kurtosis se
## X1    1 6495 0.99  0   0.99    0.99   0 0.99 1.01  0.02 0.01    -0.56  0
## [1] "Percent of values outside 2% standard deviation: 1.8014"
```

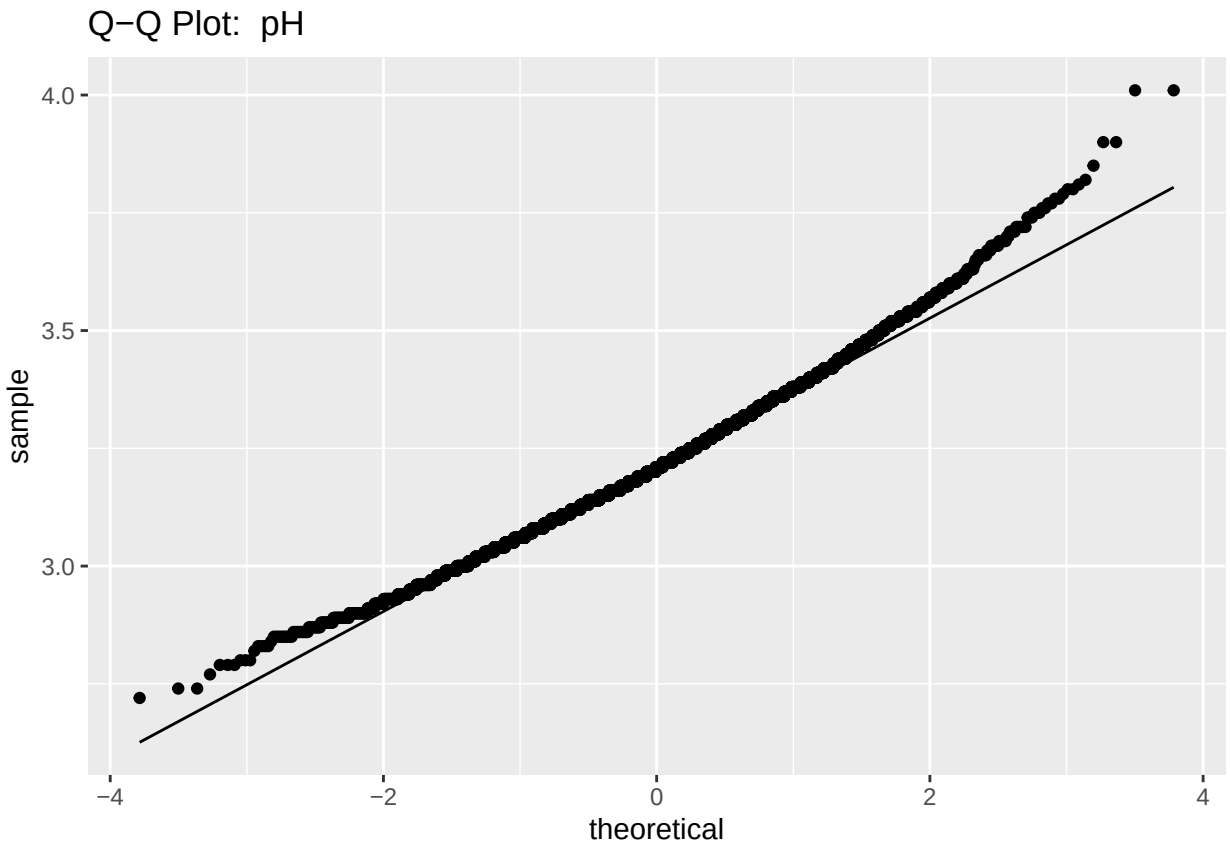
The density seems to very between the 0.99-1.04 range. From the Q-Q plot and histogram it looks like normal-distribution.

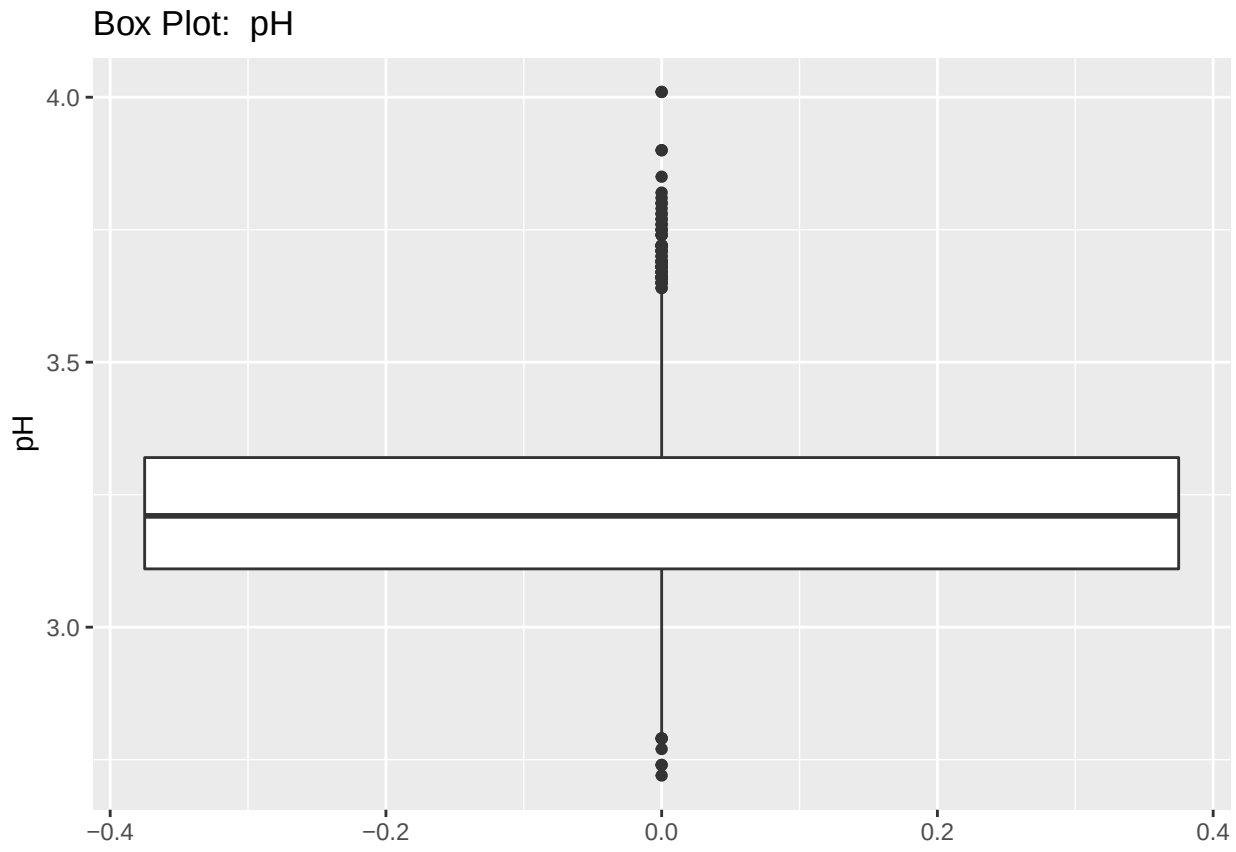
density has 1.80% values outside 2 standard deviation which is acceptable (we expect 4.6% of values to be outside 2 standard deviation). Hence, the data don't have any outliers.

1.9 pH

```
attribute_info("pH")
```


Histogram: pH
750
500
250
0
count
3.0
3.5
4.0
pH


Q−Q Plot: pH
4.0
3.5
3.0
sample
−4
−2
0
2
4
theoretical



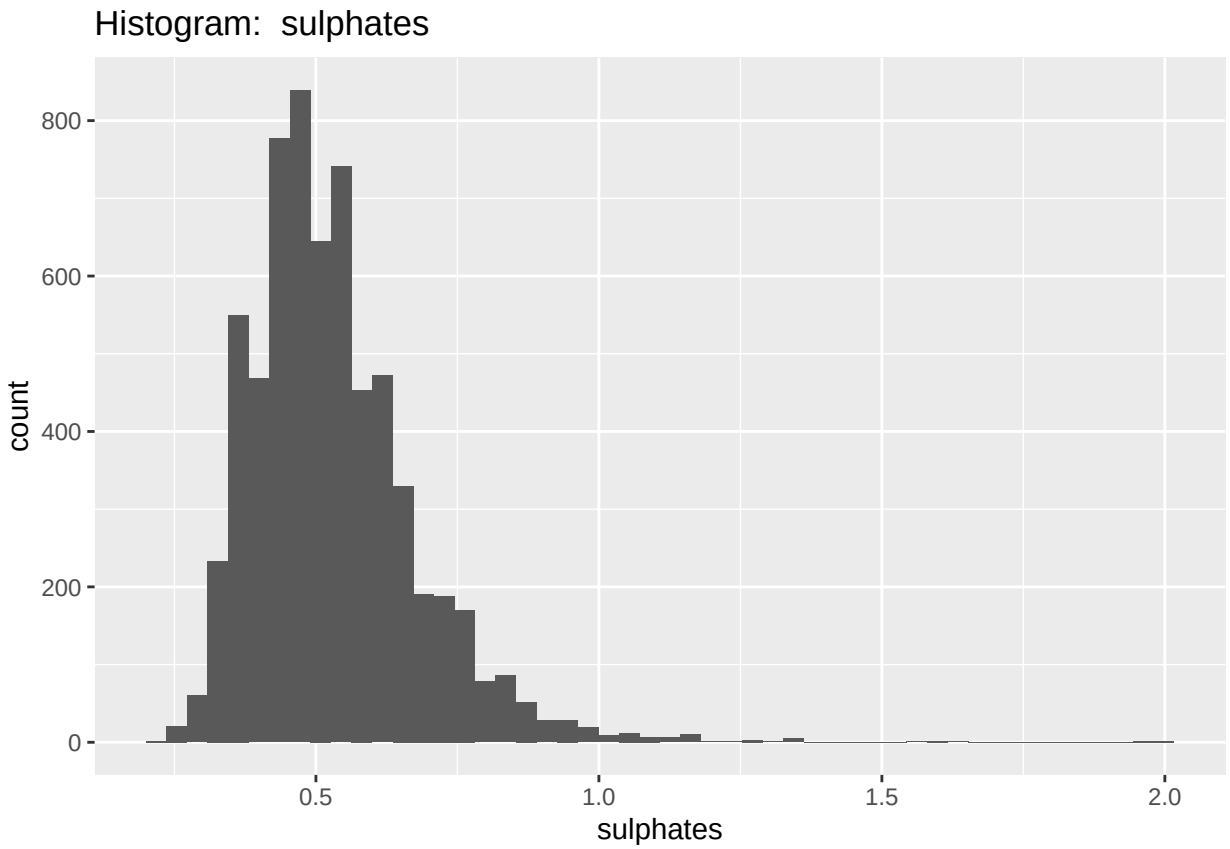

```
##    vars    n mean   sd median trimmed  mad  min  max range skew kurtosis
## X1    1 6495 3.22 0.16   3.21    3.21 0.16 2.72 4.01  1.29 0.39     0.37
##    se
## X1  0
## [1] "Percent of values outside 2% standard deviation: 4.5266"
```

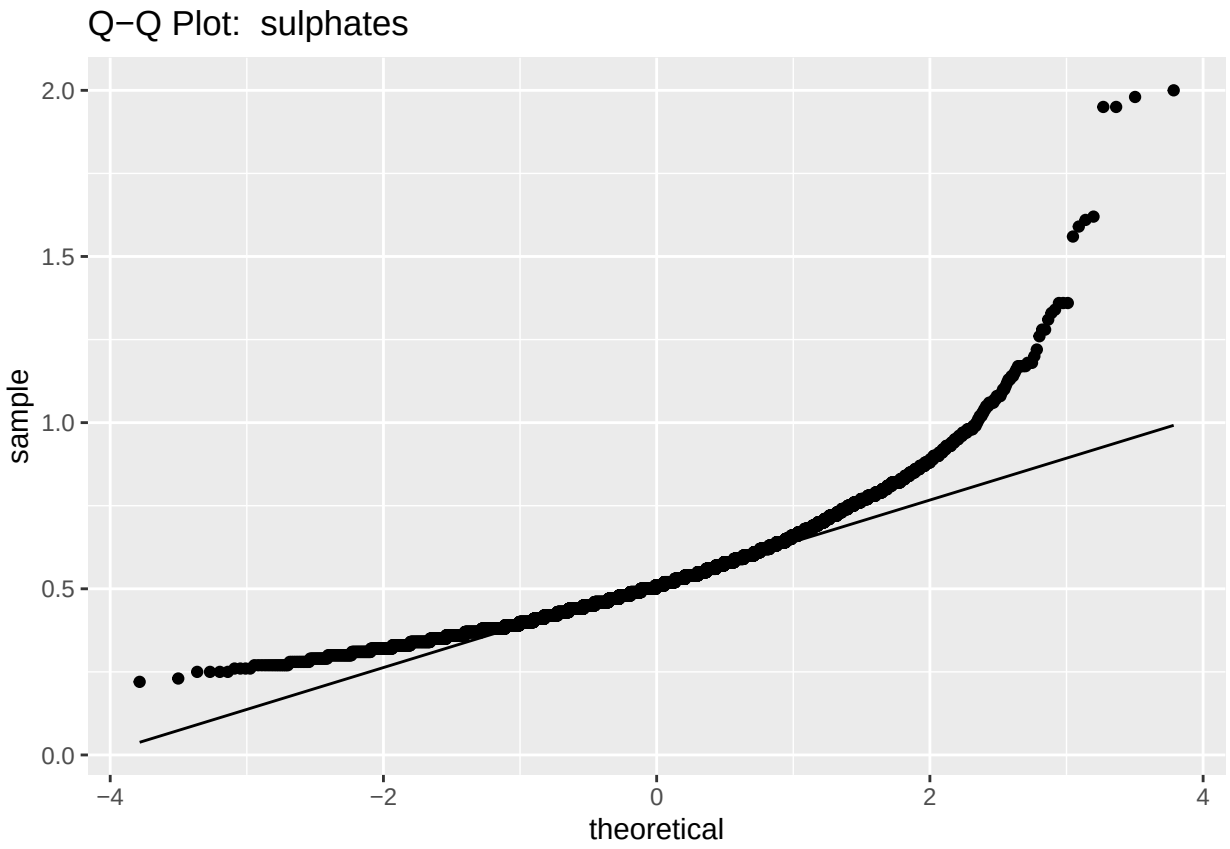
- The pH seems to display a normal distribution with major samples exhibiting values between 3.0 and 3.5.

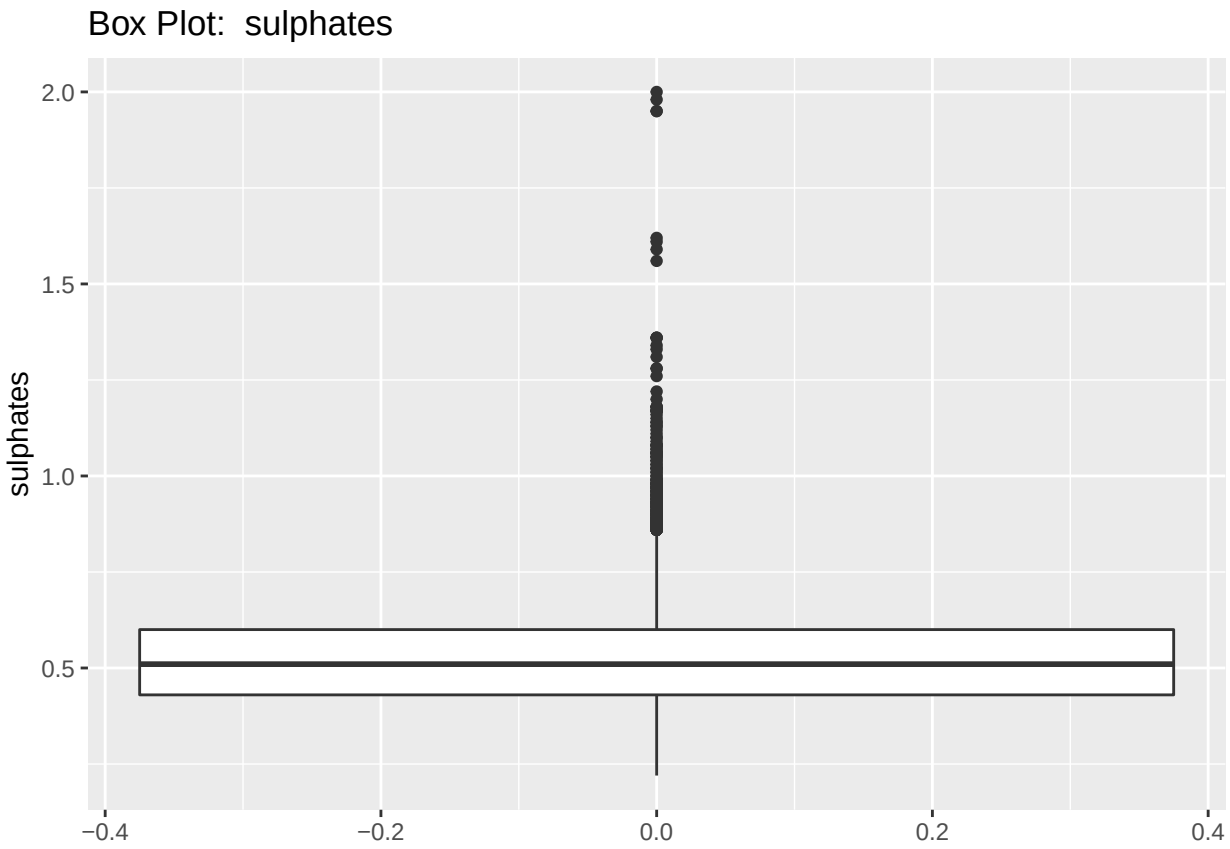
pH has 4.5266% values outside 2 standard deviation which is acceptable (we expect 4.6% of values to be outside 2 standard deviation). Hence, the data don't have any outliers.

1.10 sulphates

```
attribute_info("sulphates",50)
```


Histogram: sulphates
800
600
400
200
0
count
0.5
1.0
1.5
2.0
sulphates


Q−Q Plot: sulphates
sample
theoretical
2.0
1.5
1.0
0.5
0.0
−4
−2
0
2
4



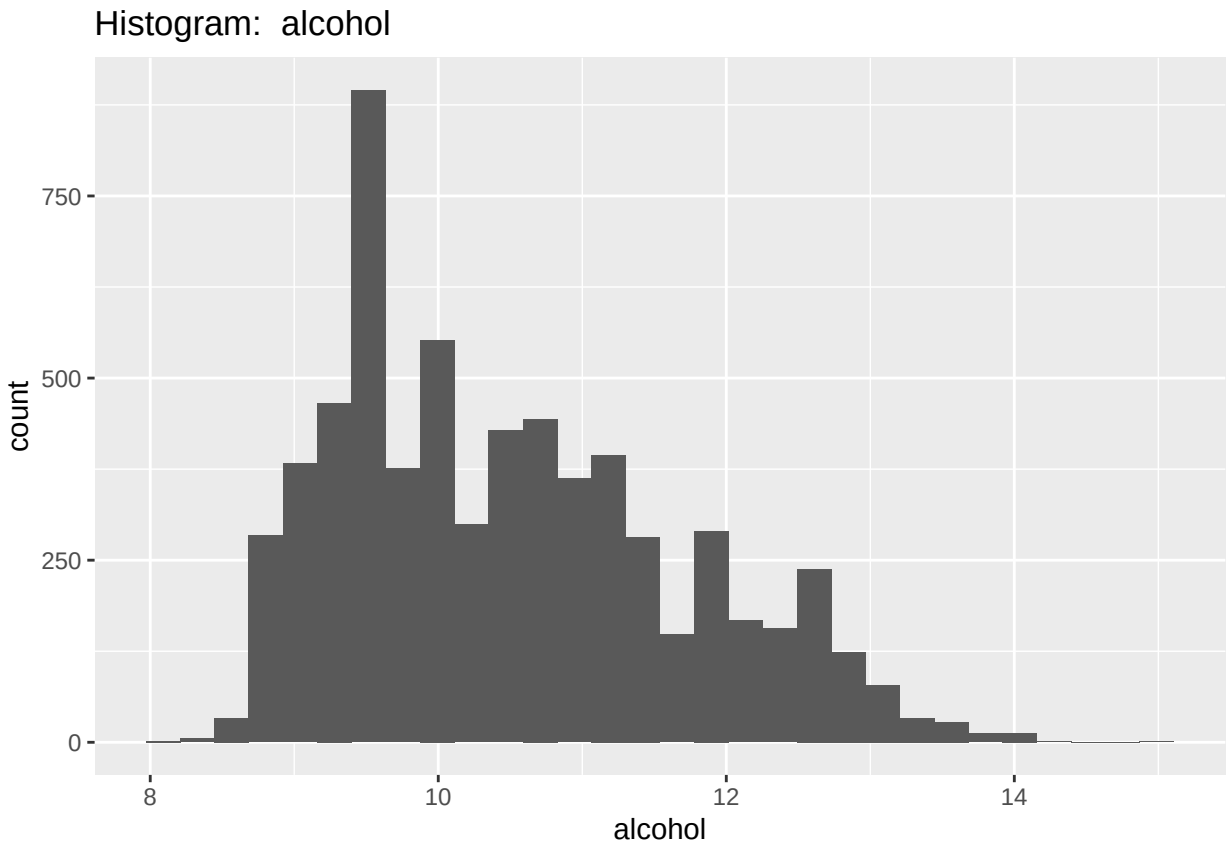

```
##    vars    n mean   sd median trimmed  mad  min max range skew kurtosis se
## X1    1 6495 0.53 0.15   0.51    0.52 0.12 0.22   2  1.78  1.8     8.65  0
## [1] "Percent of values outside 2% standard deviation: 3.7567"
```

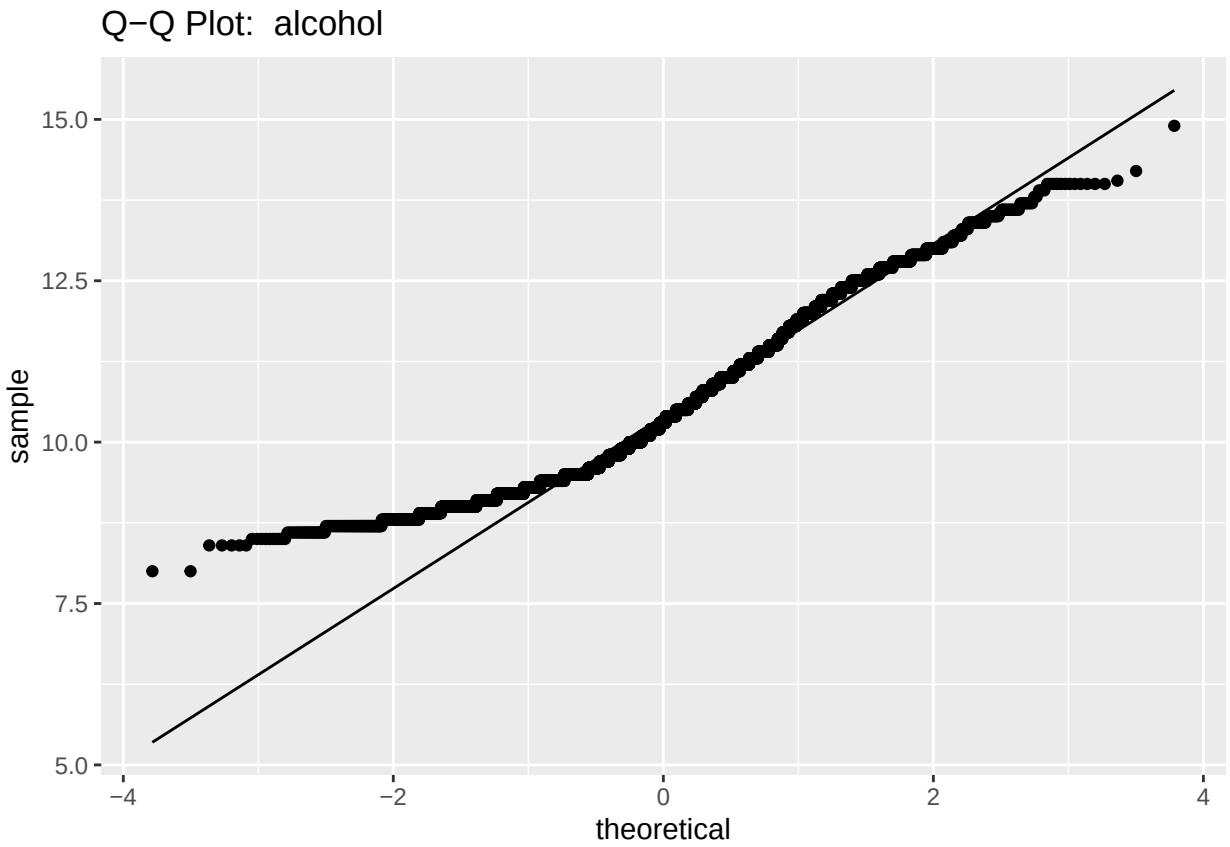
The sulphates seems to very between the 0.22 to 2. From the Q-Q plot and histogram it looks like sulphates doesn't follow the normal-distribution.

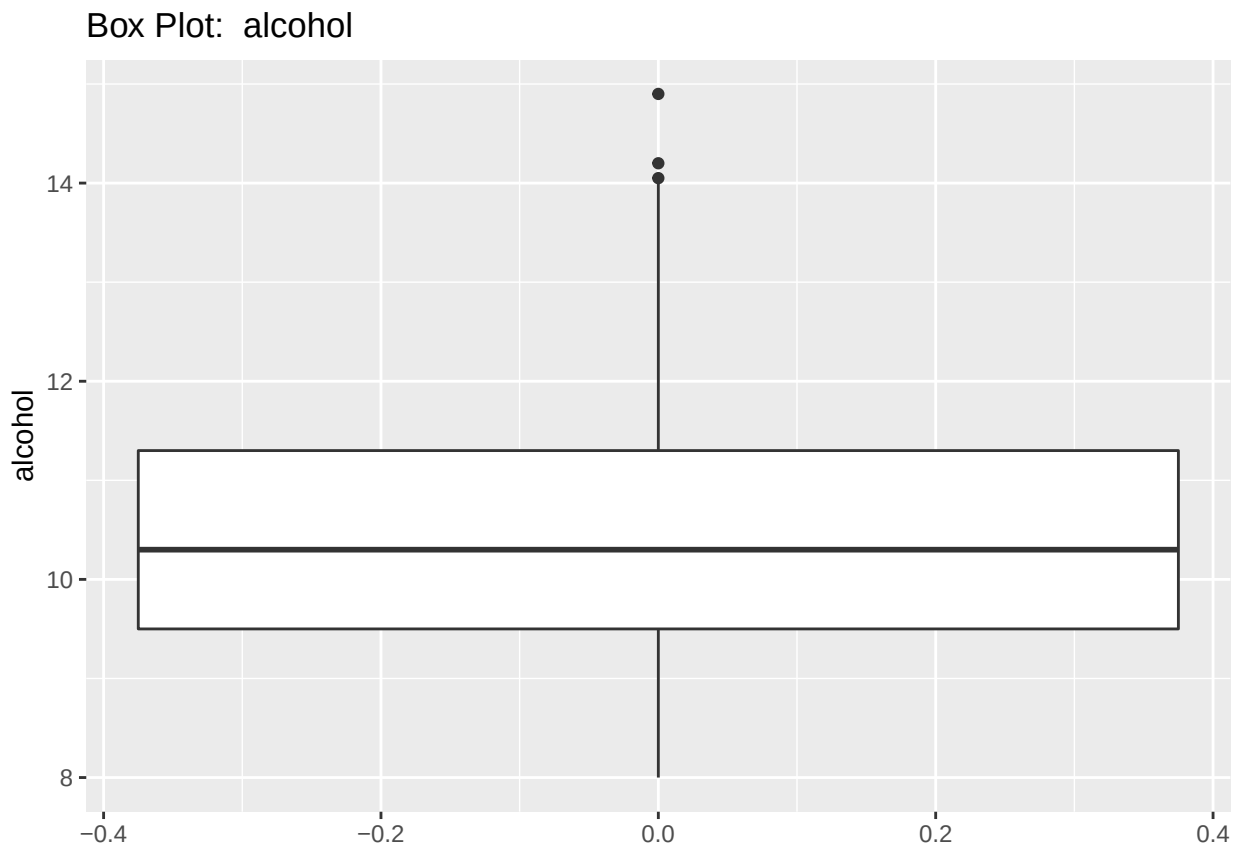
sulphates has 3.7567% values outside 2 standard deviation which is acceptable (we expect 4.6% of values to be outside 2 standard deviation). Hence, the data don't have any outliers.

1.11 alcohol

```
attribute_info("alcohol")
```


Histogram: alcohol
750
500
250
0
count
8
10
12
14
alcohol


Q−Q Plot: alcohol
15.0
12.5
10.0
7.5
5.0
sample
−4
−2
0
2
4
theoretical



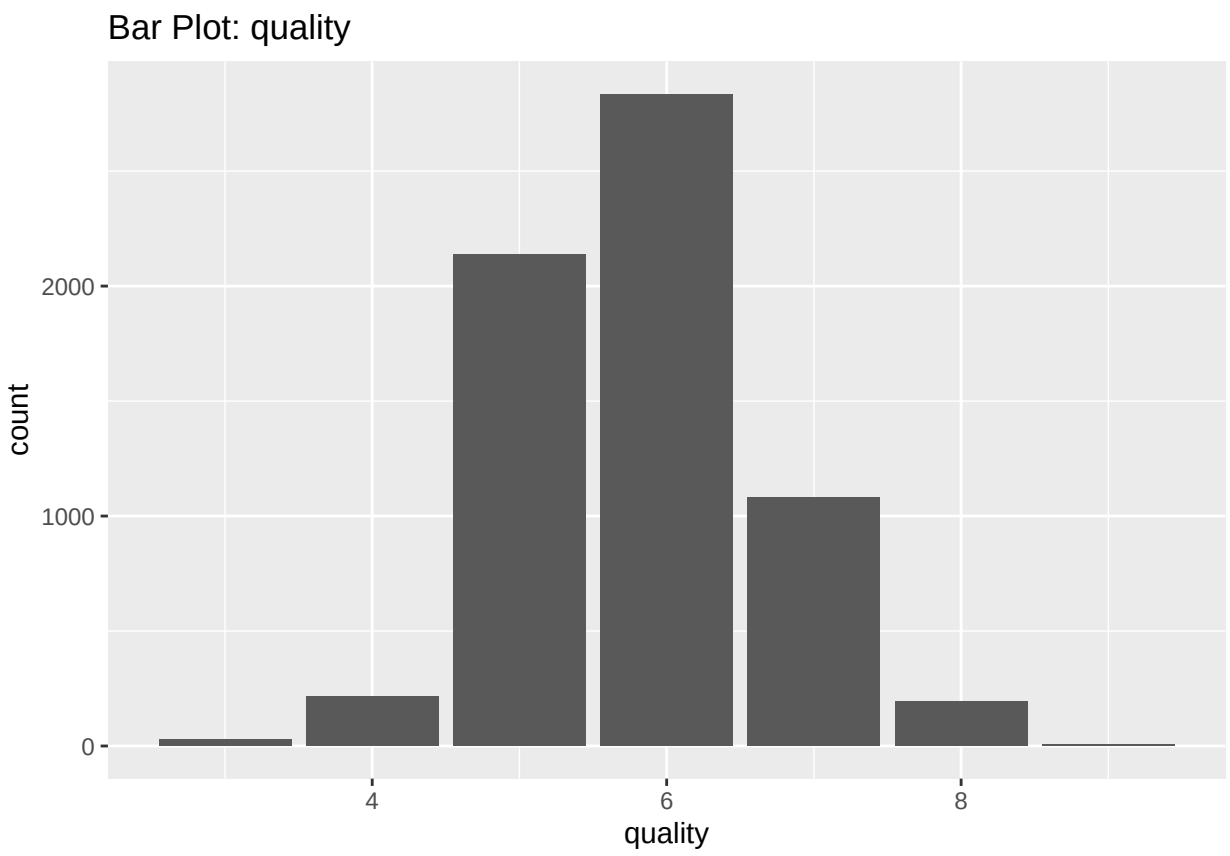

```
##    vars    n  mean   sd median trimmed  mad min  max range skew kurtosis
## X1    1 6495 10.49 1.19   10.3    10.4 1.33   8 14.9   6.9 0.57    -0.53
##      se
## X1 0.01
## [1] "Percent of values outside 2% standard deviation: 3.3718"
```

The alcohol seems to vary from 8 to 14 with major peaks around 10 with a lower count between 13 and 14 and has no outliers. But doesn't follow the normal distribution as it is seen from the Q-Q plot and histogram.

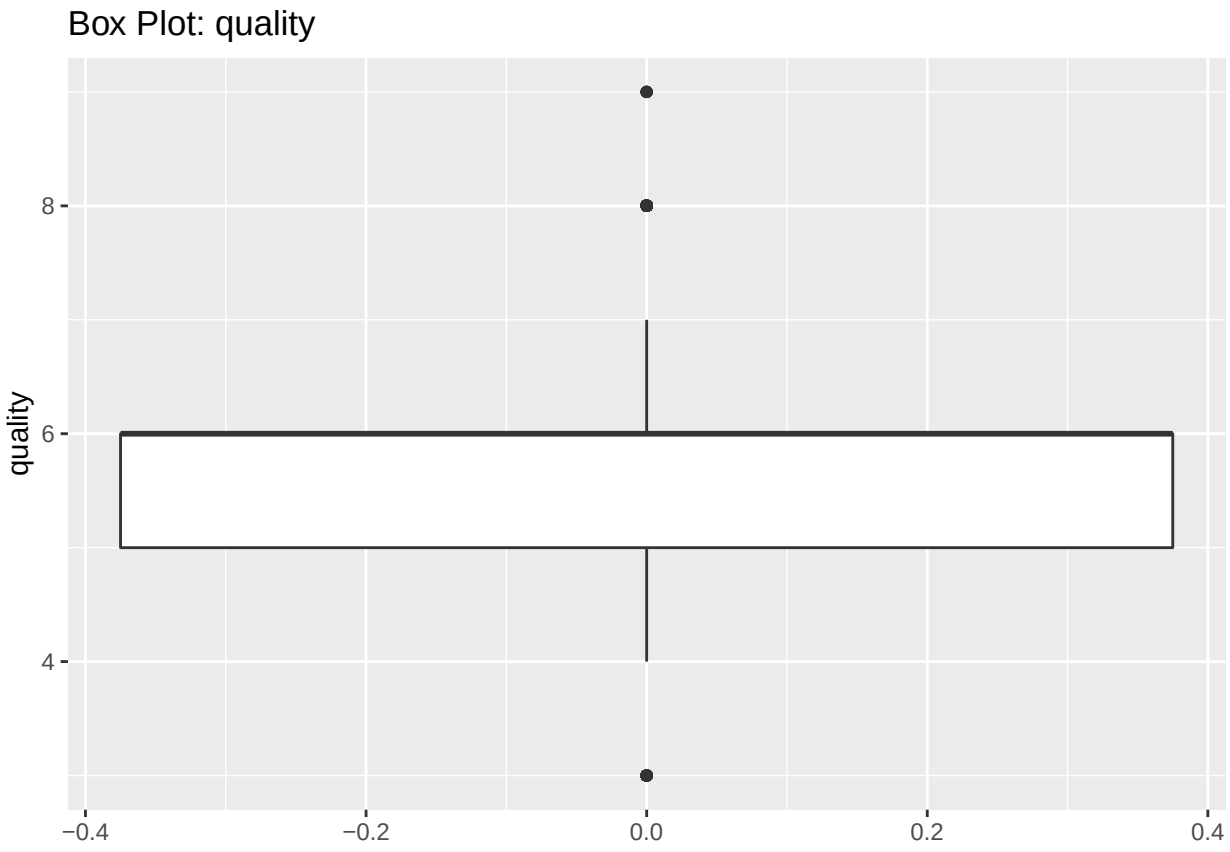
alcohol has 3.3718% values outside 2 standard deviation which is acceptable (we expect 4.6% of values to be outside 2 standard deviation). Hence, the data don't have any outliers.

1.12 quality

```
ggplot(data = wine, aes(x=quality)) + geom_bar() + labs(title = "Bar Plot: quality")
```

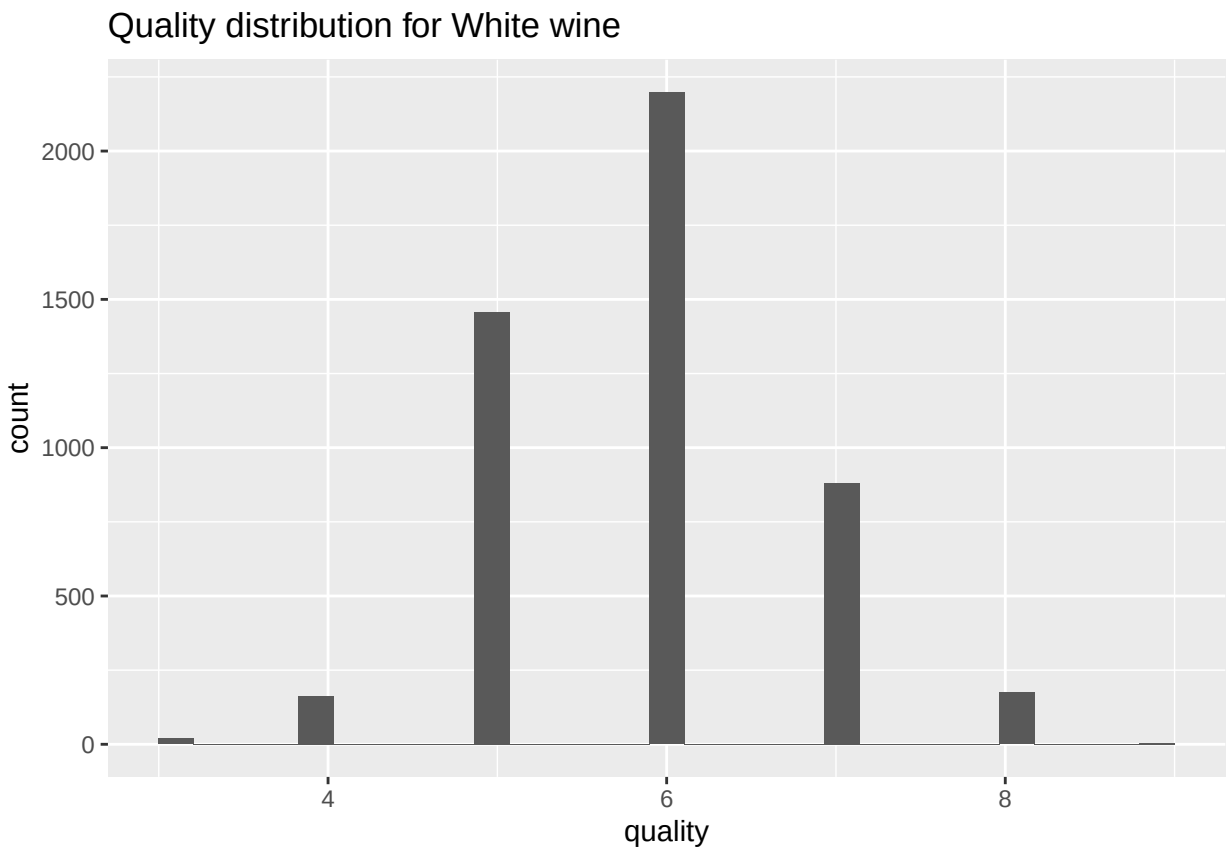



```
ggplot(data = wine, aes(y=quality)) + geom_boxplot() + labs(title = "Box Plot: quality")
```



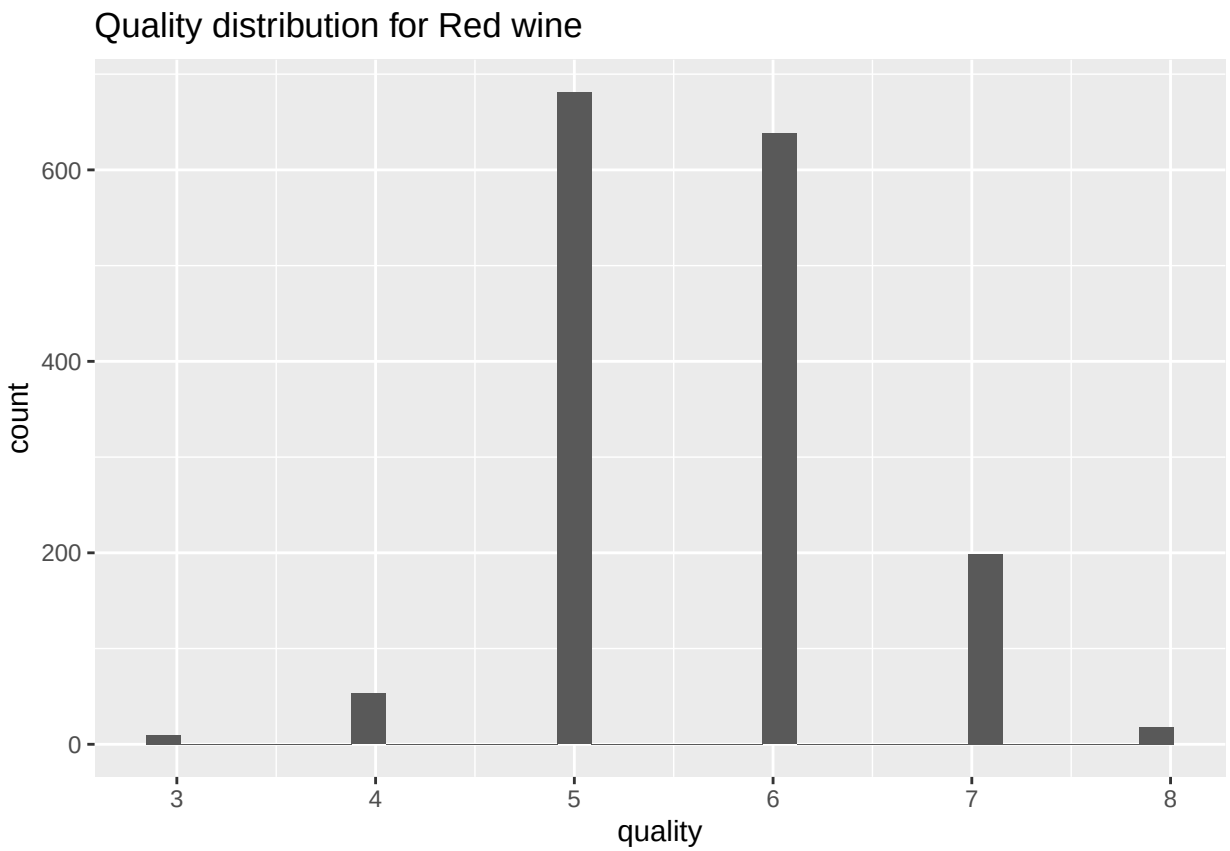

```
ggplot(data = white, aes(x=quality)) + geom_histogram() +
  ggtitle('Quality distribution for White wine')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(data = red, aes(x=quality)) + geom_histogram() +
  ggtitle('Quality distribution for Red wine')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

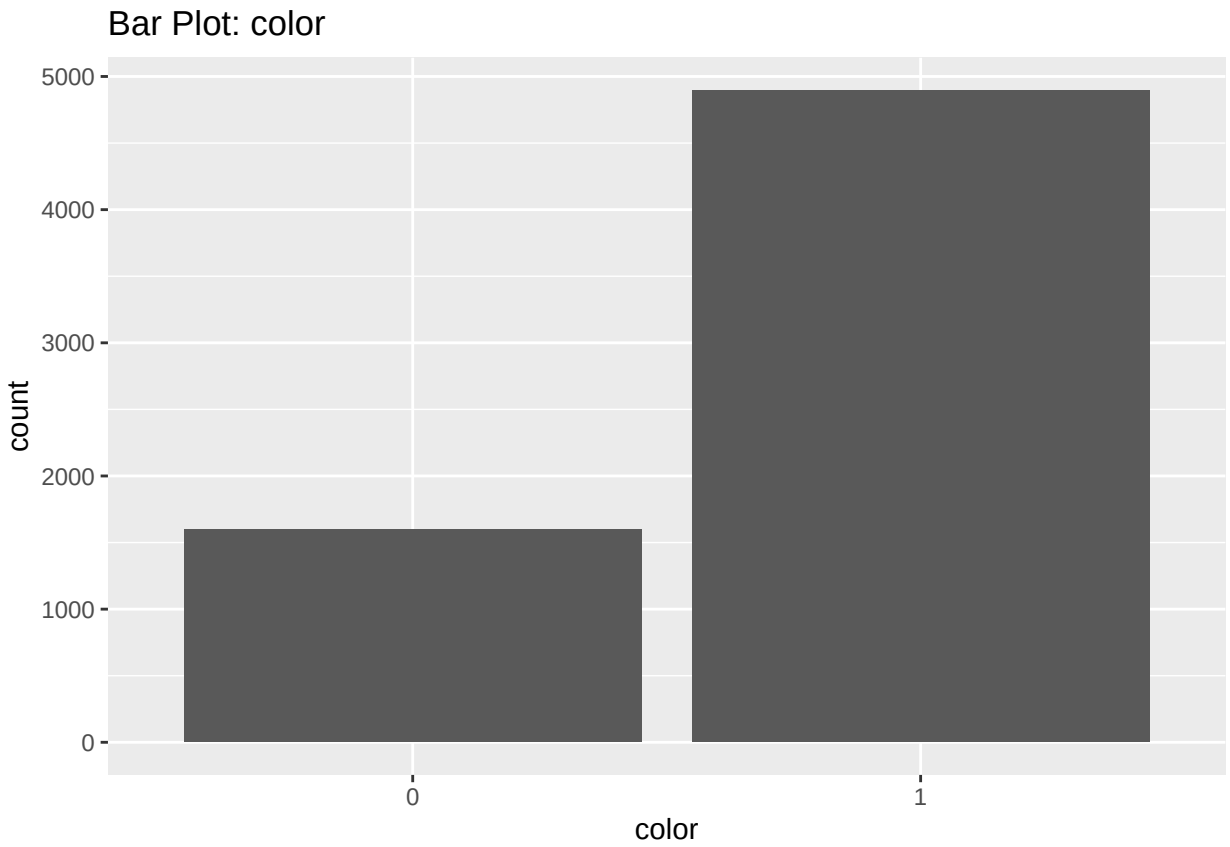


The first thing to notice is that there are few really bad wines, i.e. those whose quality is below 4. In the same way, there are few really good wines, ie, those whose quality is above 8. It's also effort to notice this distribution seems like a normal one.

The spread for the quality for both Red and White seems to exhibit similar normal distribution except for the fact that White wine distribution exhibit a peak quality around quality rating of 6 while Red wine exhibit a peak quality rating of approx 5.

1.13 color

```
ggplot(data = wine, aes(x=color)) + geom_bar() + labs(title = "Bar Plot: color")
```



* The white wine has 3 times more samples than red wine.

First Analysis

Step 1: Plan

I would like to analyse whether red or white wine hold higher quality. Basically, i want to compare means of the wine quality grouped by color.

Assumption check for independent t-test: 1) The sampling distribution is normally distributed.
-> The Quality of wine is normally distributed as we have seen in feature analyse.
2) Data are measured at least at the interval level
-> Wine Quality is ordinal categorical variable in the range 1-10. However, for the purpose of analysis we can assume that the Quality of wine is an continues interval variable.
3) Scores in different treatment conditions are independent
-> We assume that Different scores of wine quality are independent
4) Homogeneity of variance.

-> We check the homogeneity of variance using levene's test.

```
leveneTest(wine$quality,wine$color, center = median)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value Pr(>F)
## group    1  2.3982 0.1215
```

```
##        6493
```

The result F(1,6493) = 2.39, p>0.05 is non-significant for the wine quality (the value in the Pr (>F) column is more than .05). This indicates that the variances are similar between groups and the homogeneity of variance assumption is tenable.

Step 2: Test

```
# Wine Quality boxplot
colorBoxplot <- ggplot(wine, aes(group=color, y=quality, x=color))
colorBoxplot + geom_boxplot() + geom_jitter(width=0.1, height=0.5, aes(color="red")) +
  labs(title = "Wine Quality boxplot grouped by color")
```


```
wine$color<-factor(wine$color,labels = c("red","white"))

# Independent T-test
ind.t.test <- t.test(quality~color,data=wine,paired = F)
ind.t.test
```

```
##
##  Welch Two Sample t-test
##
```

```
## data:  quality by color
## t = -10.175, df = 2948.8, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2891713 -0.1957281
## sample estimates:
##   mean in group red mean in group white
##            5.636023            5.878472
```

```
#Calculating Effect size
t <- ind.t.test$statistic[[1]]
df <- ind.t.test$parameter[[1]]
r <- sqrt(t^2/(t^2+df))
print(paste0("Effect Size: ",round(r,3)))
```

```
## [1] "Effect Size: 0.184"
```

Step 3: Conclude

As we can see from the independent t-test, On average, The quality of white wine is(M=5.878, SE=0.025) greater than quality of red wine (M=5.636, SE=0.022). The difference is significant (t(2948.8) = -10.275, p < 0.001). Moreover, it represented a small sized effect (r=.18)

From the above mentioned results, we conclude that Quality of white wine is slightly greater than red wine.

2nd Analyses

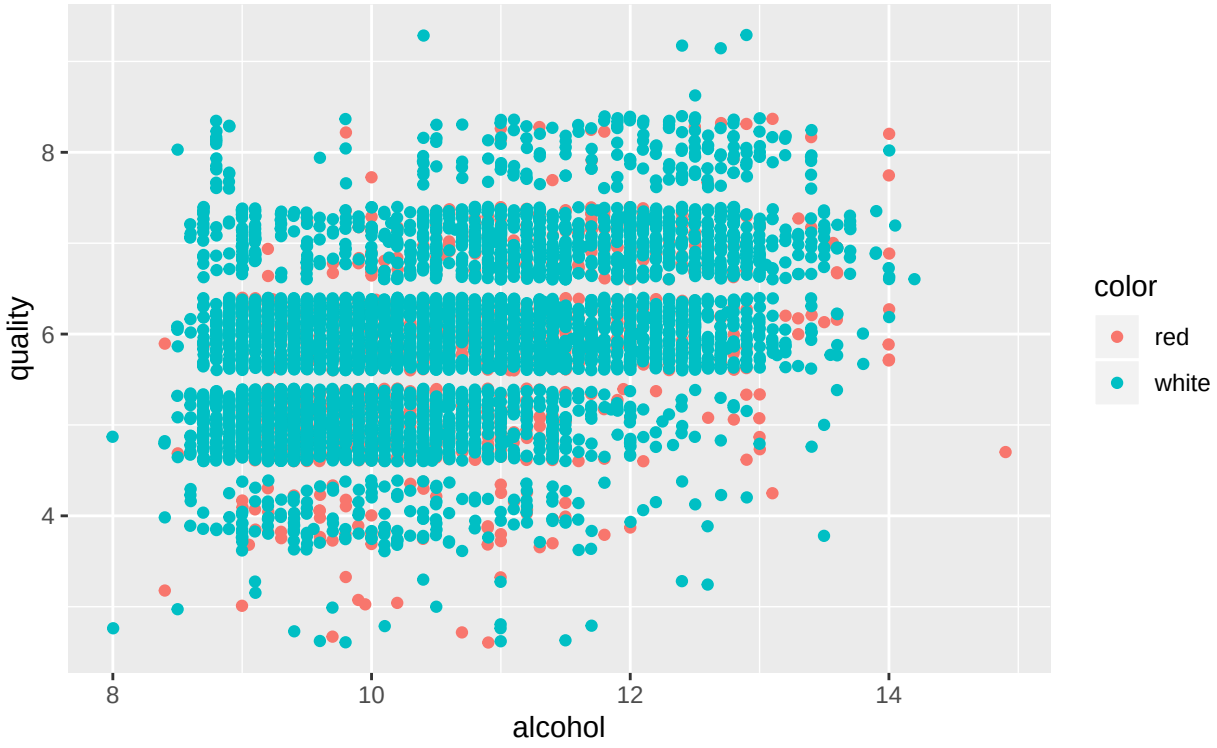
Step 1: Plan

We would like to analyse the influence of the various variables such as alcohol, sugar, sulphates etc. have on wine quality.

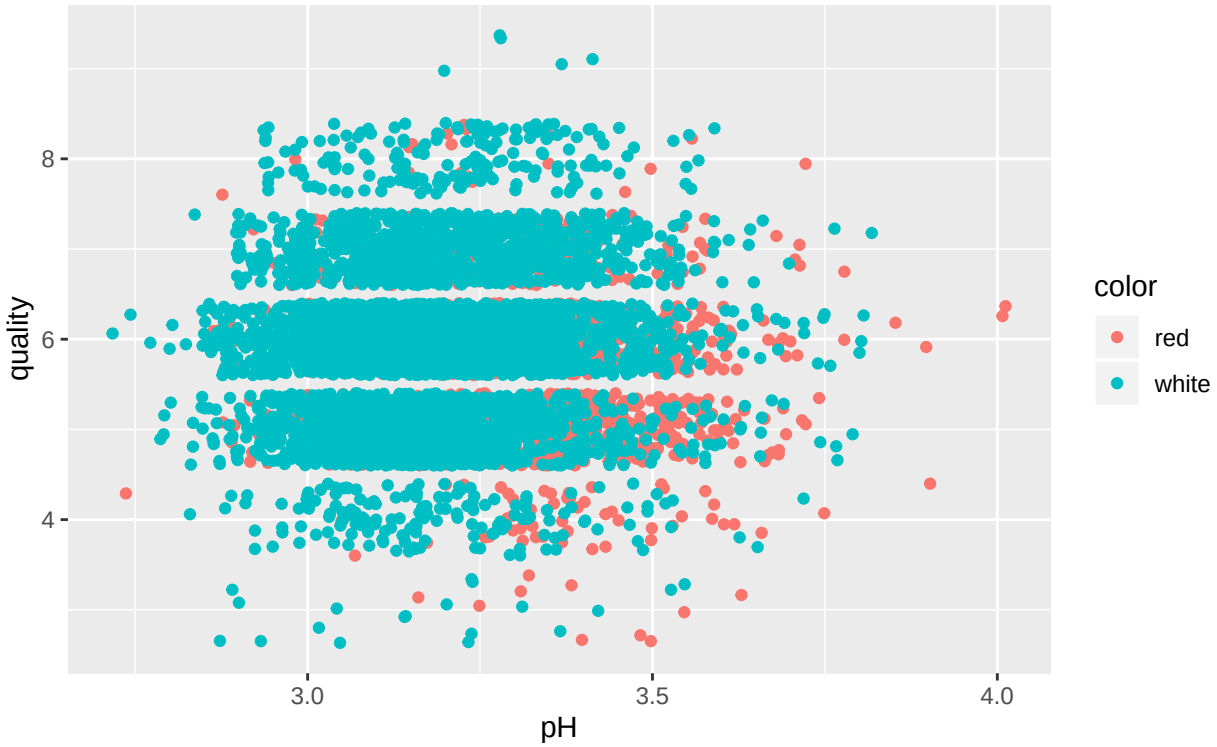
Outcome variable: Quality
Predictor Variables: Other variables except quality. We need to check for assumption like multicolinearity between predictors before finally using the variable as the predictor variable.

2. Try multiple scatter plots with respect to alcohol.

```
ggplot(data = wine,aes(x=alcohol,y=quality,color=color)) +geom_jitter()
```



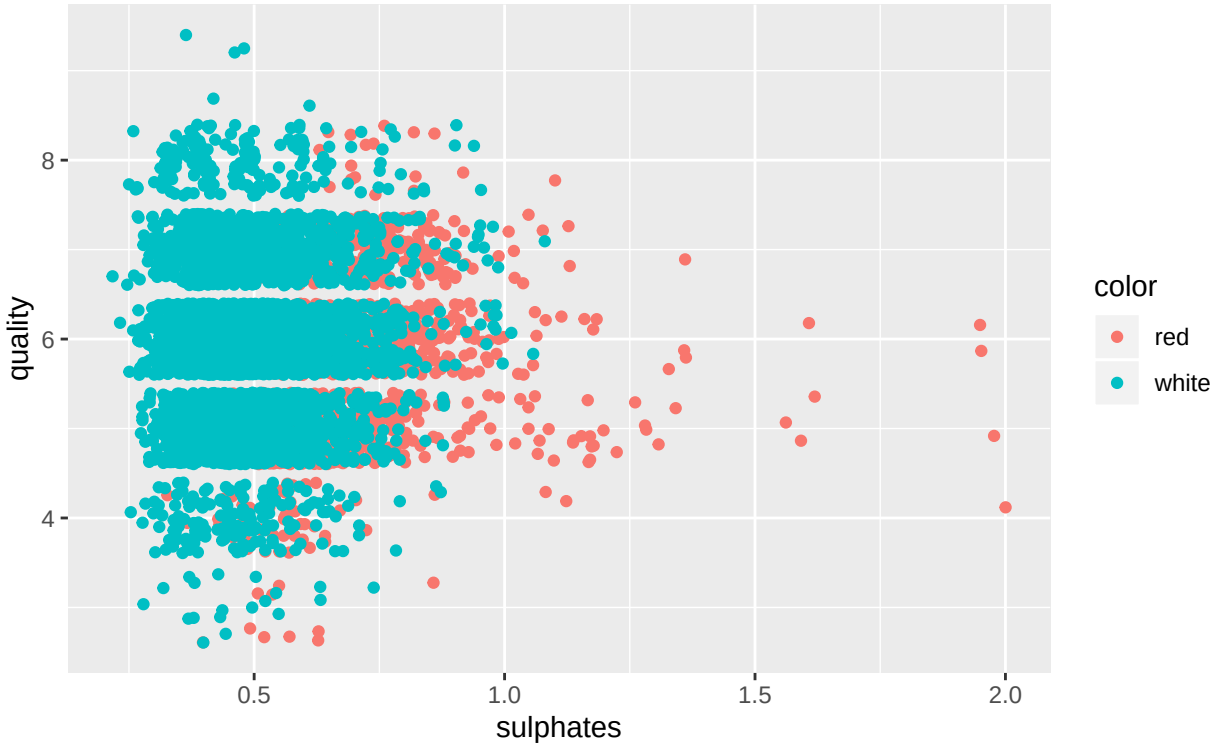

```
ggplot(data = wine,aes(x=pH,y=quality,color=color)) +geom_jitter()
```



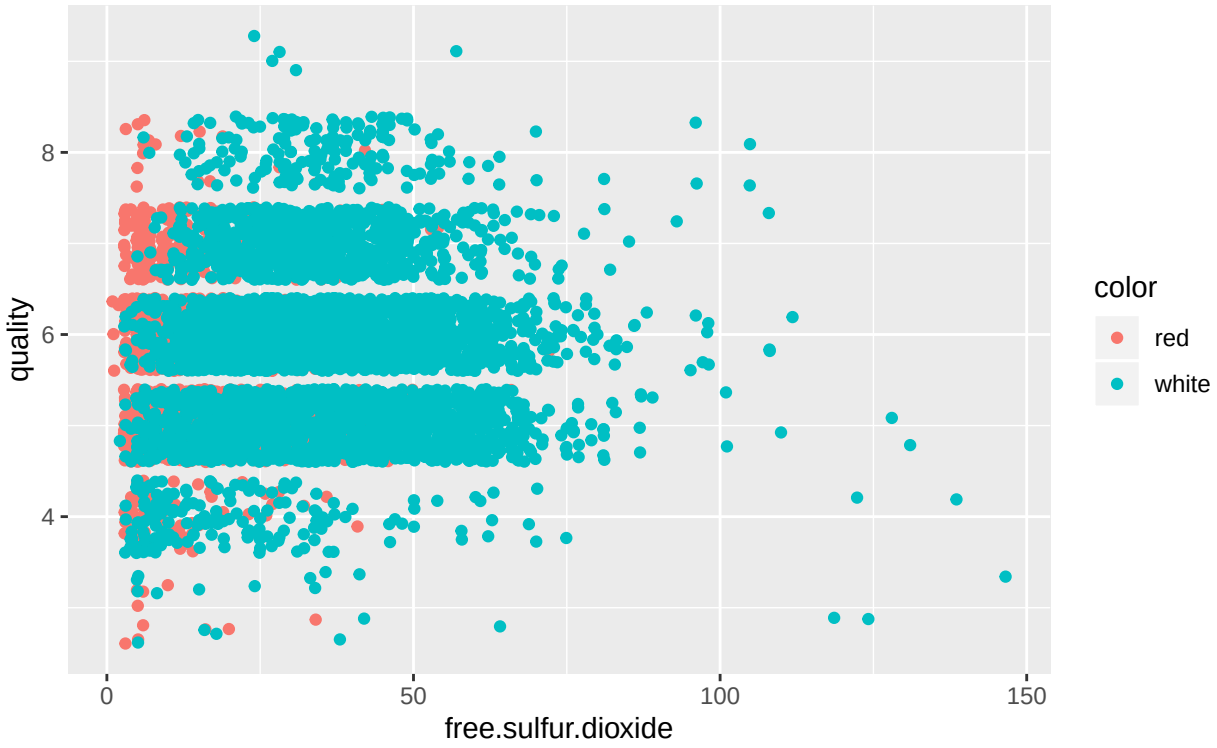
```
ggplot(data = wine,aes(x=volatile.acidity,y=quality,color=color)) +geom_jitter()
```


```
ggplot(data = wine,aes(x=residual.sugar,y=quality,color=color)) +geom_jitter()
```

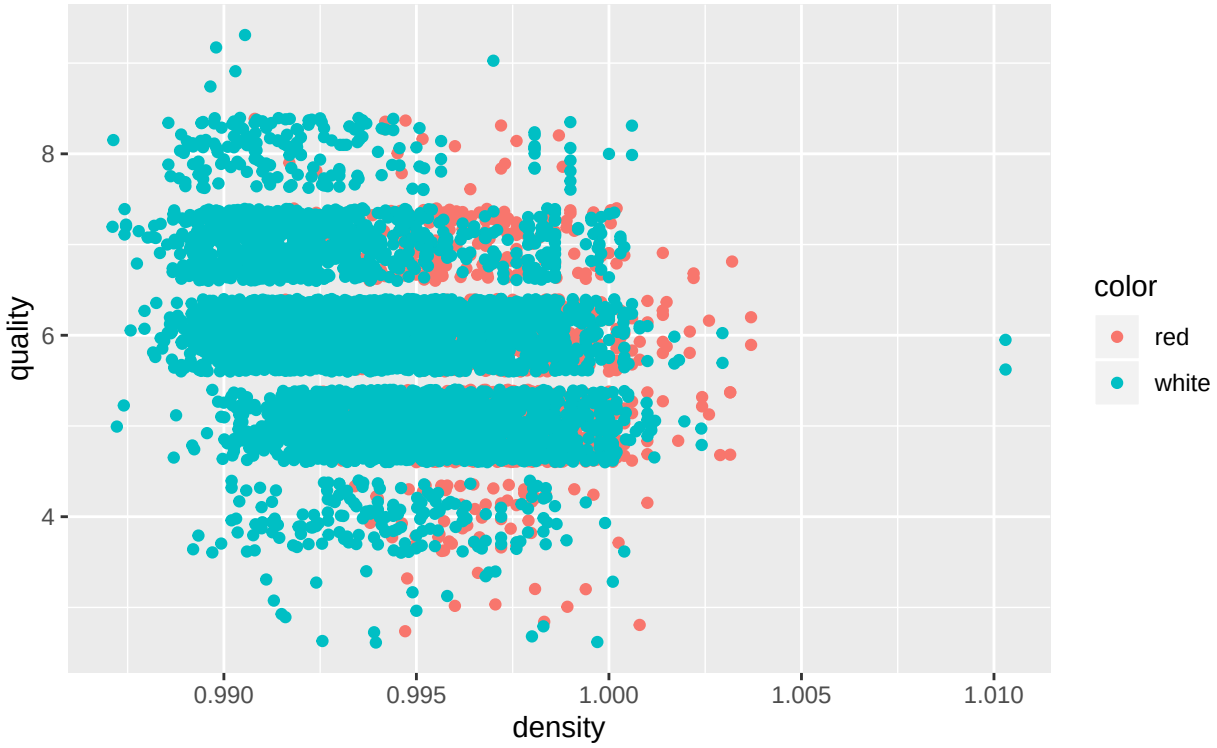
```
ggplot(data = wine,aes(x=sulphates,y=quality,color=color)) +geom_jitter()
```



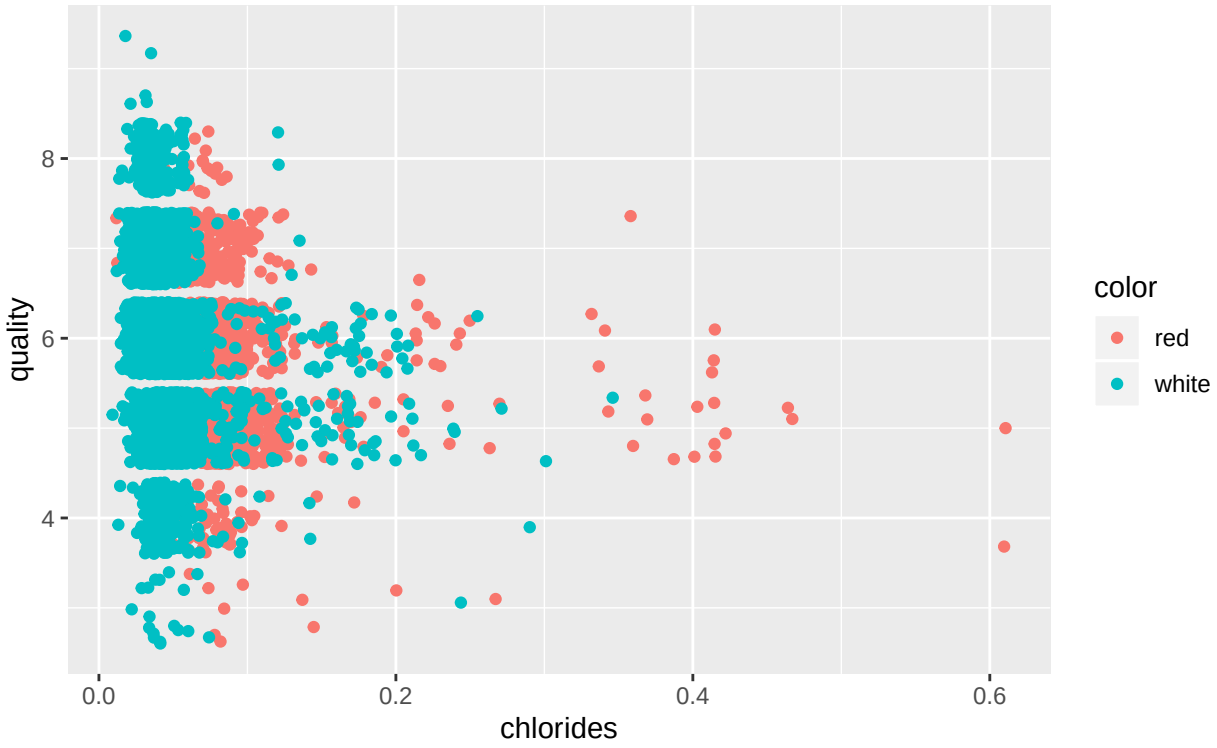

```
ggplot(data = wine,aes(x=free.sulfur.dioxide,y=quality,color=color)) +geom_jitter()
```



```
ggplot(data = wine,aes(x=density,y=quality,color=color)) +geom_jitter()
```



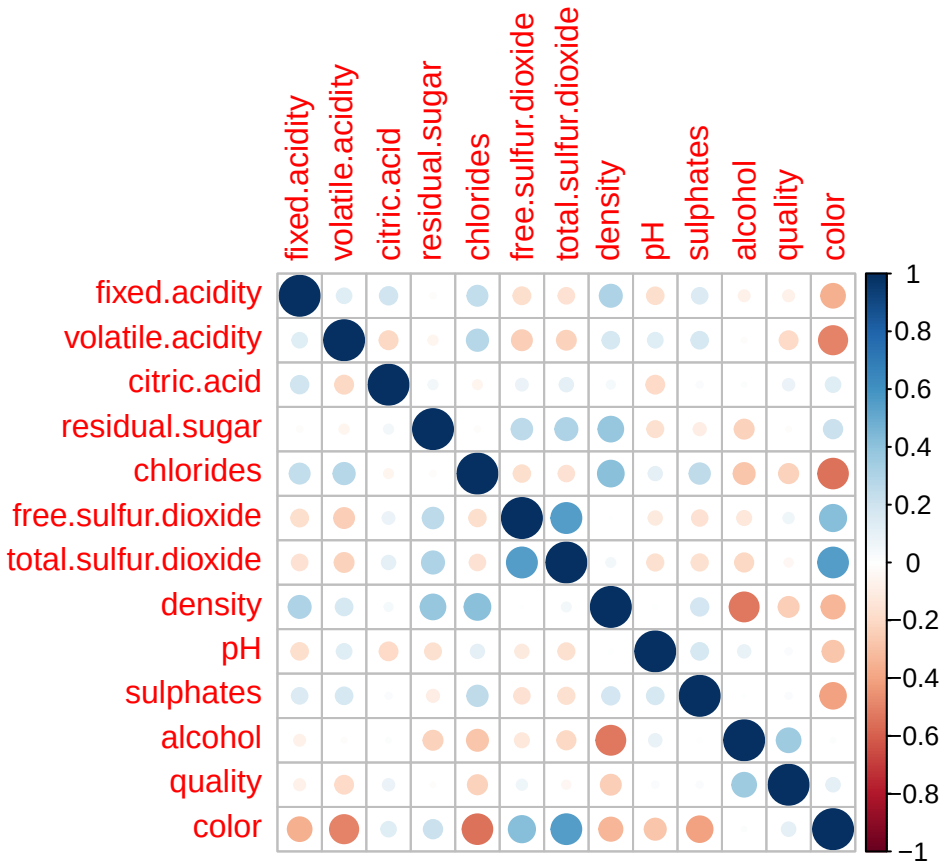

```
ggplot(data = wine,aes(x=chlorides,y=quality,color=color)) +geom_jitter()
```



* From the above plots, it looks like none of the features have clearly visible linear relationship with wine quality.

- Here, most of the feature's have significant kurtosis and some variables are categorical. Hence, I've used Kendall's correlation coefficient for all the pairs of the features.

```
wine$color<-factor(wine$color,labels = c(0,1))
wine$color <- as.numeric(levels(wine$color))[wine$color]
corrplot(cor(wine,method = "kendall"))
```




```
cor(wine,method ="kendall")
```

```
##                       fixed.acidity volatile.acidity citric.acid
## fixed.acidity            1.00000000       0.13393019  0.19010488
## volatile.acidity         0.13393019       1.00000000 -0.20187379
## citric.acid              0.19010488      -0.20187379  1.00000000
## residual.sugar          -0.01900448      -0.05334075  0.05251913
## chlorides                0.24906235       0.28171228 -0.05072527
## free.sulfur.dioxide     -0.17763903      -0.24783906  0.08791880
## total.sulfur.dioxide    -0.15365861      -0.22346954  0.11264642
## density                  0.30372036       0.17402664  0.04074751
## pH                      -0.17249731       0.13562171 -0.19880507
## sulphates                0.15252793       0.17325435  0.02353995
## alcohol                 -0.07439802      -0.01368673  0.01336671
## quality                 -0.07633470      -0.19938844  0.08197708
```

```
## color                  -0.35104953      -0.49501590  0.13329496
##                      residual.sugar   chlorides free.sulfur.dioxide
## fixed.acidity           -0.01900448  0.24906235        -0.177639035
## volatile.acidity        -0.05334075  0.28171228        -0.247839056
## citric.acid              0.05251913 -0.05072527         0.087918801
## residual.sugar           1.00000000 -0.01801775         0.264246328
## chlorides               -0.01801775  1.00000000        -0.170065709
## free.sulfur.dioxide      0.26424633 -0.17006571         1.000000000
## total.sulfur.dioxide     0.30541129 -0.15184457         0.559632347
## density                  0.38206628  0.41403136         0.001496176
## pH                      -0.16021903  0.11067341        -0.111900815
## sulphates               -0.09252653  0.25941757        -0.150807597
## alcohol                 -0.22629229 -0.27591145        -0.123330735
## quality                 -0.01317613 -0.22902821         0.067177911
## color                    0.21087097 -0.54791671         0.428474986
##                      total.sulfur.dioxide      density           pH
## fixed.acidity                -0.15365861  0.303720357 -0.172497311
## volatile.acidity             -0.22346954  0.174026642  0.135621713
## citric.acid                   0.11264642  0.040747515 -0.198805068
## residual.sugar                0.30541129  0.382066275 -0.160219033
## chlorides                    -0.15184457  0.414031361  0.110673412
## free.sulfur.dioxide           0.55963235  0.001496176 -0.111900815
## total.sulfur.dioxide          1.00000000  0.050902398 -0.163711344
## density                       0.05090240  1.000000000  0.008284505
## pH                           -0.16371134  0.008284505  1.000000000
## sulphates                    -0.16519641  0.188421050  0.174191163
## alcohol                      -0.20654980 -0.521842295  0.096802372
## quality                      -0.04193812 -0.248333793  0.025503082
## color                         0.55364064 -0.334187598 -0.277527604
##                        sulphates      alcohol     quality       color
## fixed.acidity         0.152527930 -0.074398022 -0.07633470 -0.35104953
## volatile.acidity      0.173254353 -0.013686732 -0.19938844 -0.49501590
## citric.acid           0.023539949  0.013366710  0.08197708  0.13329496
## residual.sugar       -0.092526532 -0.226292292 -0.01317613  0.21087097
## chlorides             0.259417569 -0.275911446 -0.22902821 -0.54791671
## free.sulfur.dioxide  -0.150807597 -0.123330735  0.06717791  0.42847499
## total.sulfur.dioxide -0.165196413 -0.206549798 -0.04193812  0.55364064
## density               0.188421050 -0.521842295 -0.24833379 -0.33418760
## pH                    0.174191163  0.096802372  0.02550308 -0.27752760
## sulphates             1.000000000  0.008397288  0.02386683 -0.40857451
## alcohol               0.008397288  1.000000000  0.35258274  0.01359452
## quality               0.023866834  0.352582741  1.00000000  0.11490170
## color                -0.408574509  0.013594517  0.11490170  1.00000000
```

- The features such as alcohol and density, free sulfur dioxide and total sulfur dioxide are strongly correlated.
- The color of the wine is also strongly correlated with volatile acidity chlorides and total sulfur dioxide.
- We have used this correlation matrix while choosing the predictor variables that can have influence on wine quality.

Linear Regression

Assumptions for Linear model:

We will need to test residuals for linearity and normality, etc. once we construct the model.
1) Predictor variables is either continues or categorical and We are assuming that quality is continues and unbounded.
2) Predictor variables is not highly correlated with any other variable in the dataset. However, we are assuming that it is also uncorrelated with any other external variable that is not in the dataset.

Assumptions needed to test after linear model is created:
3) Multicolinearity between the predictor variables
4) independence of residuals
5) normality of residuals

- Let's try to predict the wine quality.

```
m1 <- lm(quality ~ alcohol, data = wine)
m2 <- update(m1, ~ . + volatile.acidity)
m3 <- update(m2, ~ . + residual.sugar)
m4 <- update(m3, ~ . + free.sulfur.dioxide)
m5 <- update(m4, ~ . + sulphates)
mtable(m1,m2,m3,m4,m5)
```

```
##
## Calls:
## m1: lm(formula = quality ~ alcohol, data = wine)
## m2: lm(formula = quality ~ alcohol + volatile.acidity, data = wine)
## m3: lm(formula = quality ~ alcohol + volatile.acidity + residual.sugar,
##     data = wine)
## m4: lm(formula = quality ~ alcohol + volatile.acidity + residual.sugar +
##     free.sulfur.dioxide, data = wine)
## m5: lm(formula = quality ~ alcohol + volatile.acidity + residual.sugar +
##     free.sulfur.dioxide + sulphates, data = wine)
##
## ============================================================================================
##                             m1           m2           m3           m4           m5
## --------------------------------------------------------------------------------------------
##   (Intercept)             2.405***     2.929***     2.565***     2.470***     2.033***
##                          (0.086)      (0.085)      (0.099)      (0.103)      (0.109)
##   alcohol                 0.325***     0.318***     0.342***     0.344***     0.350***
##                          (0.008)      (0.008)      (0.008)      (0.008)      (0.008)
##   volatile.acidity                    -1.326***    -1.226***    -1.162***    -1.266***
##                                       (0.057)      (0.058)      (0.061)      (0.061)
##   residual.sugar                                    0.016***     0.013***     0.016***
##                                                    (0.002)      (0.002)      (0.002)
##   free.sulfur.dioxide                                            0.002***     0.003***
##                                                                 (0.001)      (0.001)
##   sulphates                                                                   0.711***
##                                                                              (0.064)
## --------------------------------------------------------------------------------------------
##   R-squared               0.198        0.260        0.266        0.267        0.281
##   N                    6495         6495         6495         6495         6495
## ============================================================================================
##   Significance: *** = p < 0.001; ** = p < 0.01; * = p < 0.05
```

- I kept adding the variables to the model which shows significant influence to the wine quality at P<.05 and not correlated with the variables already added.

```
summary(m5)
```

```
##
## Call:
## lm(formula = quality ~ alcohol + volatile.acidity + residual.sugar +
##     free.sulfur.dioxide + sulphates, data = wine)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3097 -0.4818 -0.0344  0.4585  3.0971
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          2.0333033  0.1091855  18.622  < 2e-16 ***
## alcohol              0.3496846  0.0083835  41.711  < 2e-16 ***
## volatile.acidity    -1.2664868  0.0612866 -20.665  < 2e-16 ***
## residual.sugar       0.0164830  0.0023086   7.140 1.04e-12 ***
## free.sulfur.dioxide  0.0026095  0.0006121   4.263 2.04e-05 ***
## sulphates            0.7109326  0.0644062  11.038  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7404 on 6489 degrees of freedom
## Multiple R-squared:  0.2808, Adjusted R-squared:  0.2803
## F-statistic: 506.8 on 5 and 6489 DF,  p-value: < 2.2e-16
```

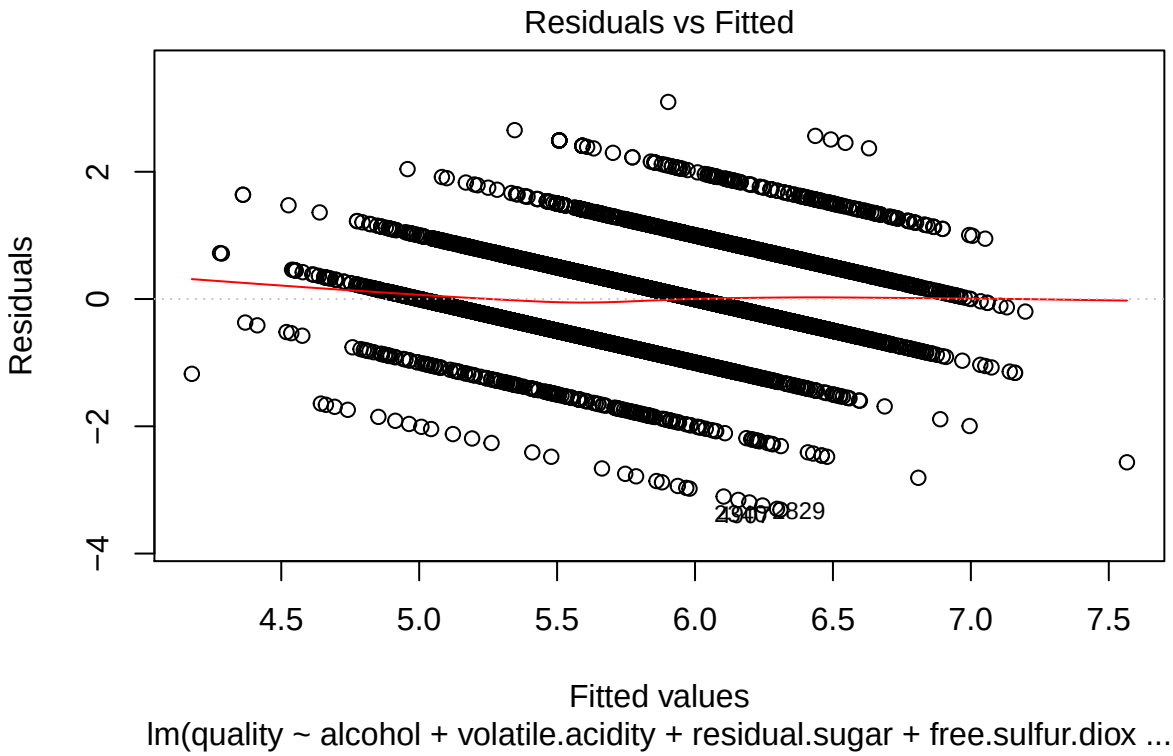
```
confint(m5)
```

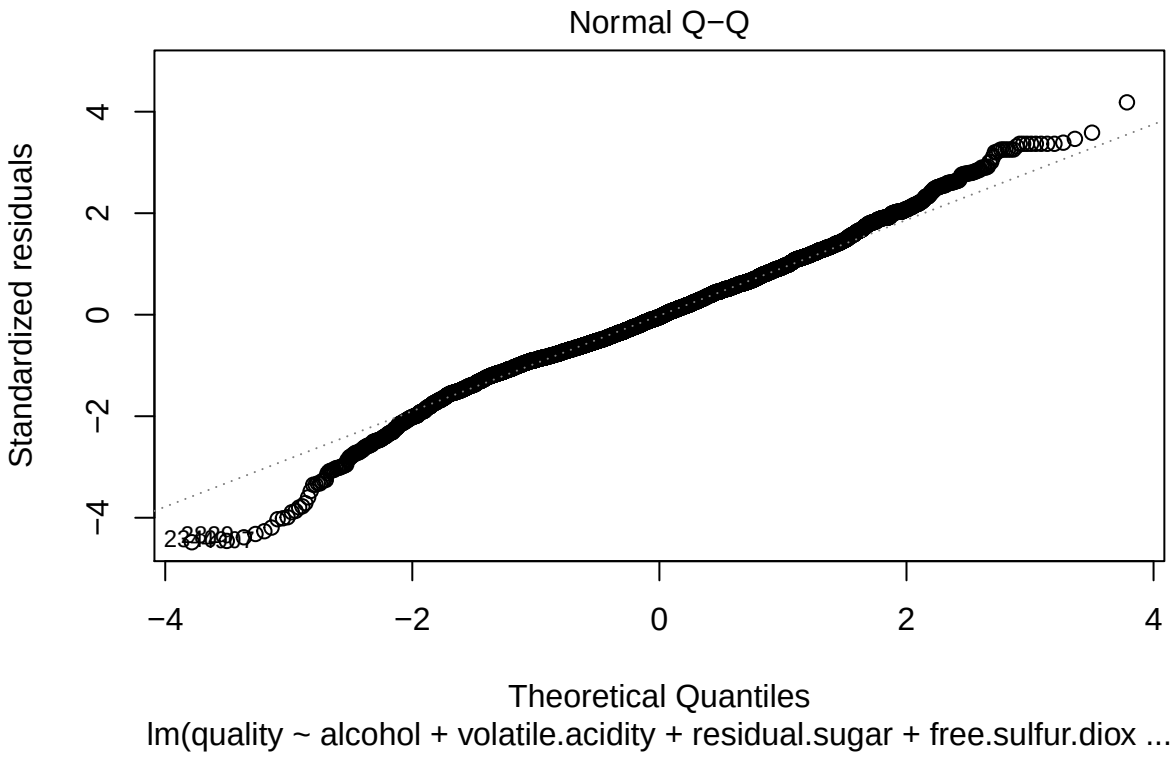
```
##                           2.5 %       97.5 %
## (Intercept)          1.819263810  2.247342777
## alcohol              0.333250195  0.366119040
## volatile.acidity    -1.386628785 -1.146344793
## residual.sugar       0.011957314  0.021008714
## free.sulfur.dioxide  0.001409611  0.003809307
## sulphates            0.584675209  0.837189895
```

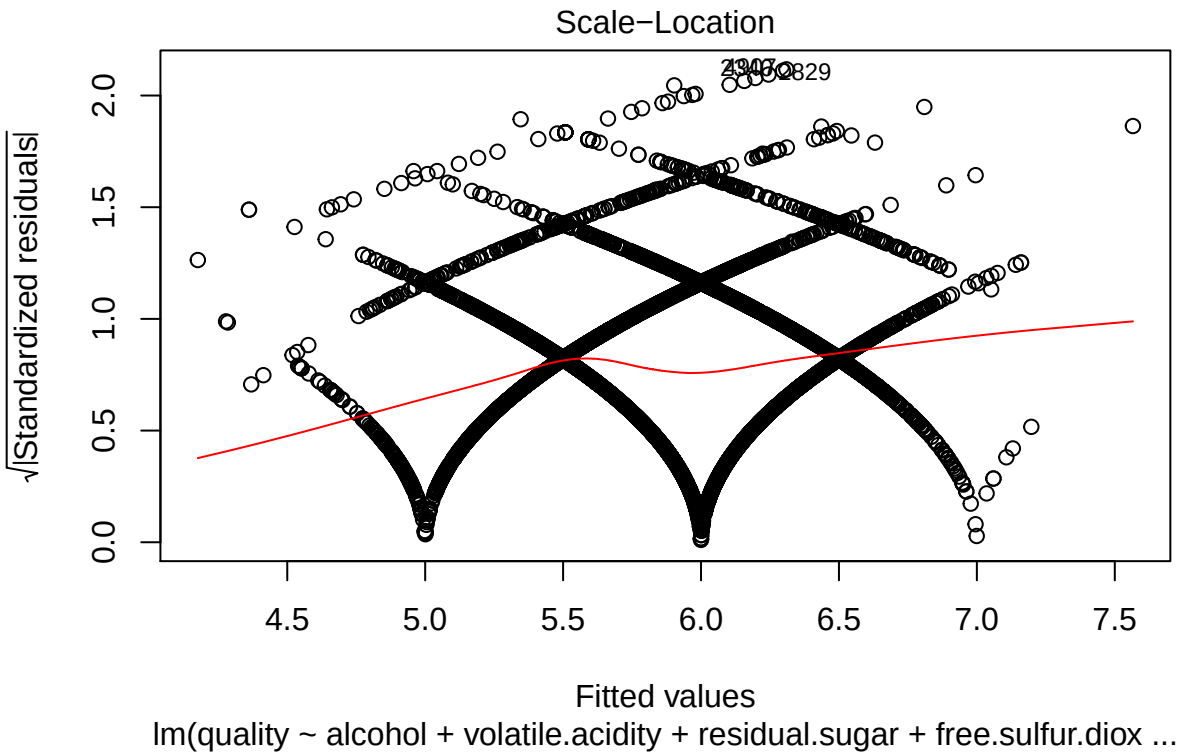
```
lm.beta(m5)
```

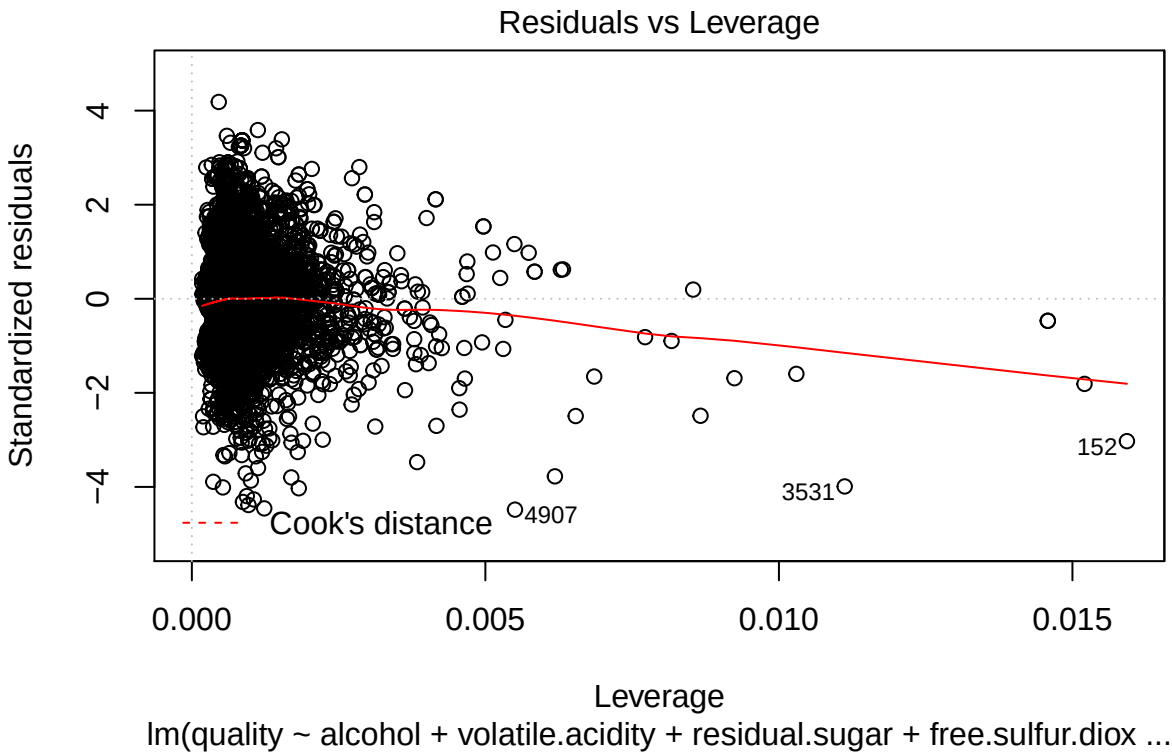
```
##             alcohol    volatile.acidity      residual.sugar
##          0.47795441         -0.23869580          0.08875524
## free.sulfur.dioxide           sulphates
##          0.05220104          0.12122755
```

```
plot(m5)
```


Residuals vs Fitted
Residuals
2
0
−2
−4
4.5
5.0
5.5
6.0
6.5
7.0
7.5
Fitted values
lm(quality ~ alcohol + volatile.acidity + residual.sugar + free.sulfur.diox ...


Normal Q−Q
Standardized residuals
4
2
0
−2
−4
−4
−2
0
2
4
Theoretical Quantiles
lm(quality ~ alcohol + volatile.acidity + residual.sugar + free.sulfur.diox ...

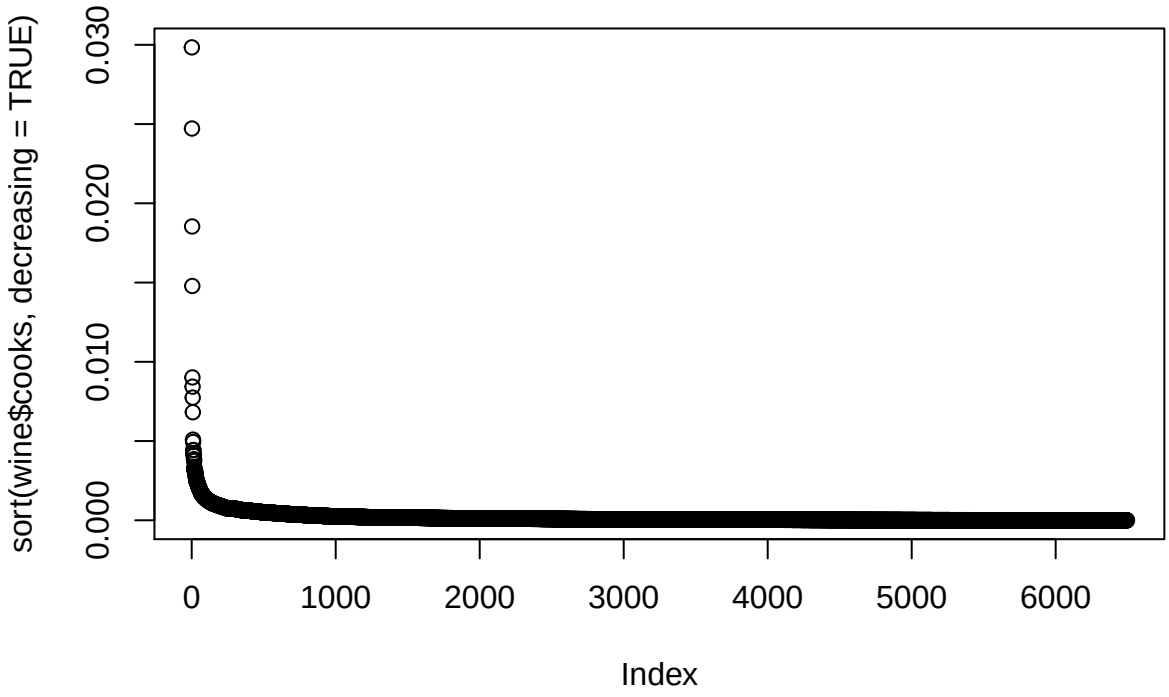

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
Fitted values
lm(quality ~ alcohol + volatile.acidity + residual.sugar + free.sulfur.diox ...



* Interpret the summary of the model: **All the predictor variables (mentioned below) have an influence on the wine quality at 5% level of significance: alcohol (t(6489)=41.711, p < 2e-16), volatile.acidity (t(6489)=-20.655, p < 2e-16), residual.sugar (t(6489)= 7.140, p < 1.04e-12), free.sulfur.dioxide (t(6489)=4.263, p < 2.04e-05),and sulphates (t(6489)=11.038, p < 2e-16). The intercept is significantly different from 0 (t(6489) = 18.622, p=2e-16).

- Multiple R-squared value is 0.281 and Adjusted R- squared value is .2803. Which shows our model does well to generalize on the population. Moreover, it also exhibits that our model explains the 28.08% variance in the Wine Quality feature.
- F is 506.8, which is significant at p < .001. This result tells us that there is less than a 0.1% chance that an F-ratio this large would happen if the null hypothesis (Null Hypothesis: the coefficients associated with the variables is equal to zero) were true. Moreover, The confidence Interval's of the variables coefficient doesn't cross zero which shows that null hypothesis is false.
- Therefore, we can conclude that our regression model results in significantly better prediction of Wine Quality than if we used the mean value of wine quality. In short, the regression model overall predicts Wine Quality significantly well.
- However, The predictor variables accounts for only 28% of variance in the wine quality. It means that there there is still 72% of variance unexplained.
- The residuals are also normally distributed as it is evident from the Q-Q plot of the residuals.
- The features such as alcohol,volatile.acidity,sulphates,residual.sugar and free.sulfur.dioxide impact the wine quality. The alcohol has highest impact on quality of wine such as 1 standard deviation change in the value of alcohol brings 0.477 standard deviation of change in wine quality.

```
wine$fitted <- m5$fitted
wine$residuals <- m5$residuals
wine$standardized.residuals <- rstandard(m5)
wine$cooks <- cooks.distance(m5)
plot(sort(wine$cooks, decreasing=TRUE))
```




```
max(wine$cooks)
```

```
## [1] 0.02983844
```

```
durbinWatsonTest(m5)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.1779979      1.643991       0
##  Alternative hypothesis: rho != 0
```

To find out the influential cases, we calculated the cook's distance on the model. Maximum Cook's distance was around 0.0298, which is way below of threshold value of 1. Thus, we conclude that there are no influential cases in the data frame.

The Durbin-Watson test for independent errors was significant at the 5% level of significance (d=1.64,p=0.0).Even though d=1.64 which doesn't impliy autocorelation, p value is 0. so we reject the null hypothesis that the errors are independent, and Thus, we conclude that errors are Dependent. The implication is that the model has not accounted for all the signal and therefore the residuals consist of signal plus noise.[2]

Outliers:

```
possible.outliers <- subset(wine, standardized.residuals < -2 | standardized.residuals > 2)
dim(possible.outliers)[1]/dim(wine)[1]*100
```

```
## [1] 5.296382
```

From the above calculations, we found around 5.296% of sample data points(344 residuals) are above or below the 2% standard deviation. We do expect 4.56% of data to be outside of 2 standard deviation range. Hence, we do not remove any data points from the data frame.

Multicolinearity

```
vif(m5)
```

```
##             alcohol    volatile.acidity      residual.sugar
##            1.184714            1.203824            1.394348
## free.sulfur.dioxide           sulphates
##            1.352679            1.088289
```

```
mean(vif(m5))
```

```
## [1] 1.244771
```

```
1/vif(m5)
```

```
##             alcohol    volatile.acidity      residual.sugar
##           0.8440852           0.8306865           0.7171810
## free.sulfur.dioxide           sulphates
##           0.7392738           0.9188739
```

The largest VIF was 1.352679, less than 10; the average vif was 1.244771. The lowest tolerance (1/VIF) was 0.7171810, much greater than .1 (which would indicate a serious problem) and .2 (which indicates a potential problem). Thus, we conclude that there is no collinearity in our data.

What follow up do you have (e.g., changes to the model, followup questions)? **From the developed model, it is evident that alcohol levels, residual sugar, volatile acidity, free.sulfur.dioxide and sulphates explains 28% of variance in quality of wine. The predictor variables impact the wine quality. However, there is still 72% of variance left unexplained. Hence, Possible followup includes collecting more features for the data to explain more varinace.**

Step 4: Conclusion

We built a **linear model to assess the impact of alcohol levels, residual sugar, volatile acidity, free.sulfur.dioxide and sulphates on the wine quality All the mentioned features explains 28% of variance in quality of wine. The Residuals of the model are not independent. It means that residuals are auto correlated. Thus, Assumptions of linear model were not met as residuals**

contained the signals in addition to noise. However, it still explains 28% of signals in the quality.

Produce a table of the model parameters (Beta 0, 1, ...):

```
summary(m5)$coefficients
```

```
##                         Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept)          2.033303293 0.1091854552  18.622474 1.902884e-75
## alcohol              0.349684618 0.0083834993  41.711057 0.000000e+00
## volatile.acidity    -1.266486789 0.0612866295 -20.664977 6.200104e-92
## residual.sugar       0.016483014 0.0023086424   7.139700 1.037313e-12
## free.sulfur.dioxide  0.002609459 0.0006120644   4.263374 2.042180e-05
## sulphates            0.710932552 0.0644061798  11.038266 4.435691e-28
```

```
paste("beta values")
```

```
## [1] "beta values"
```

```
t(t(lm.beta(m5)))
```

```
##                            [,1]
## alcohol              0.47795441
## volatile.acidity    -0.23869580
## residual.sugar       0.08875524
## free.sulfur.dioxide  0.05220104
## sulphates            0.12122755
```

Apart from this, we can futher investigate the impact of the variables on Wine Quality. Intuitively, the alcohol has the highest impact on the quality of wine

This sums up the analysis for given dataset.

Further work can be done by finding out the quality and type of grape that was used to make wine and the storing conditions as well.

A good reference book to understand statistics in-depth from basic using R can be found here - https://www.amazon.ca/Discovering-Statistics-Using-Andy-Field/dp/1446200469/ref=sr_1_1?gclid=CjwKCAjw8ZHsBRA6EiwA7hw_saOIYjWaxz0qOkQ9rnIYWfdZBwE&hvadid=208318866958&hvdev=c&hvlocphy=9001027&hvnetw=g&hvpos=1t1&hvqmt=e&hvrand=18277962018890616395&hvtargid=aud-749198100220%3Akwd-364808766957&hydadcr=22486_9261677&keywords=discovering+statistics+using+r&qid=1569021131&s=gateway&sr=8-1