Tipologia i cicle de vide de les dades : PAC3

Descripció del dataset:

Com a estudi he escollit una base de dades de vins vermells per tal de predir si són bons o no. N'hi ha una columna que ens indica la seva qualitat basada en un sensor; aquest son els diferents atributs:

1 - fixed acidity
2 - volatile acidity
3 - citric acid
4 - residual sugar
5 - chlorides
6 - free sulfur dioxide [1-72]
7 - total sulfur dioxide [6-289]
8 - density
9 - pH [0-14]
10 - sulphates
11 - alcohol

Output variable (based on sensory data):
12 - quality (score between 0 and 10)

Les dades estan penjades a la web: https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009

Objectius de l'anàlisis

A partir d'aquestes dades es vol saber quines variables contribueixen més sobre la qualitat del vi. També es vol crear models que permetin predir la qualitat del vi; però això faré servir un model de regressió amb "Xarxa Neuronal". També faré servir proves de contrast d'hipòtesi per a identificar propietats interessants.

Integració i selecció de les dades d'interès a analitzar.

De cara a les proves faré servir tots els atributs; l'atribut "quality" em permetrà la classificació de vins i serà el que faré servir per al model d'aprenentatge supervisat.

El primer que faig és carregar les dades:

```
library(ggplot2)
library(corrplot)
library(dplyr)
library(GGally)
library(lattice)
library(caret)
library(neuralnet)
library("kableExtra")

wine_data <- read.csv('winequality-red.csv', header=T, sep="," ,
                      fileEncoding = "UTF-8-BOM",
                      na.strings = "NA", stringsAsFactors = FALSE)
```

```
attach(wine_data)
colnames(wine_data) <- c("fixed_acidity","volatile_acidity","citric_acid",
                         "residual_sugar","chlorides","free_sulfur_dioxide",
                         "total_sulfur_dioxide", "density","pH","sulphates",
                         "alcohol","quality")
```

#Descripció del dataset.

L'estructura de les dades és la que es veu al següent diagrama:

```
str(wine_data)
```

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

I els valors que poden prendre es pot veure a la següent taula:

```
summary(wine_data)
```

```
##  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
```

Veig que totes les dades són de tipus numèric i l'atribut quality agafa només un rang de valor de sencers; aquesta es pot considerar de tipus categòric.

```
summary(wine_data)
```

```
##  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
```

Selecció de les dades d'interes

Miro quins atributs es poden discretitzar i en trobo que quality és idoni.

```
  wine_data$quality <- as.factor(wine_data$quality)
  str(wine_data)
```

```
## '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             : Factor w/ 6 levels "3","4","5","6",..: 3 3 3 4 3 3 3 5 5 3 ...
```

```
  levels(wine_data$quality)
```

```
## [1] "3" "4" "5" "6" "7" "8"
```

Neteja de dades

Abans de començar a netejar vaig a visualitzar les dades per veure que contenen valors correctes:

```
head(wine_data[,1:4])
```

```
##   fixed_acidity volatile_acidity citric_acid residual_sugar
## 1           7.4             0.70        0.00            1.9
## 2           7.8             0.88        0.00            2.6
## 3           7.8             0.76        0.04            2.3
```

```
## 4          11.2             0.28        0.56            1.9
## 5           7.4             0.70        0.00            1.9
## 6           7.4             0.66        0.00            1.8
```

Dades amb zeros o elements buits

Tenim moltes dades numèriques amb valor zero; aquests valors en aquest cas són normals, no es corresponen a cap valor desconegut. Per altra banda no n'hi han valors buits:

```
colSums(is.na(wine_data))
```

```
##        fixed_acidity     volatile_acidity          citric_acid
##                    0                    0                    0
##       residual_sugar            chlorides  free_sulfur_dioxide
##                    0                    0                    0
## total_sulfur_dioxide              density                   pH
##                    0                    0                    0
##            sulphates              alcohol              quality
##                    0                    0                    0
```

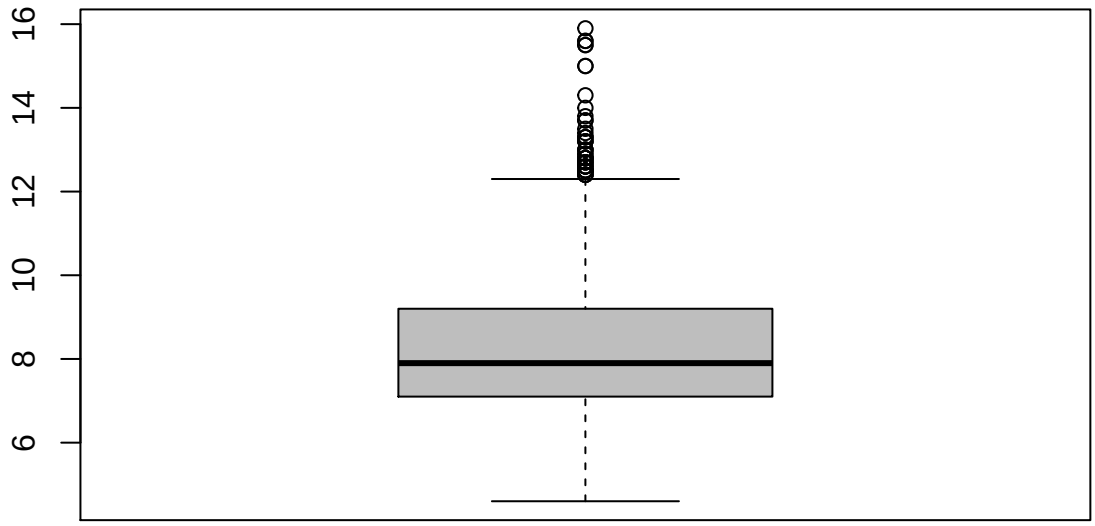
Identificació i tractament de valors extrems

Els valors extrems o outliers són aquells que no semblen normals dintre de les mostres. Per tal d'identificar-los faré servir la funció boxplot de R per tal de mostrar els valors numèrics d'aquest outliers.

```
boxplot.stats(wine_data$fixed_acidity)$out
```

```
##  [1] 12.8 12.8 15.0 15.0 12.5 13.3 13.4 12.4 12.5 13.8 13.5 12.6 12.5 12.8 12.8
## [16] 14.0 13.7 13.7 12.7 12.5 12.8 12.6 15.6 12.5 13.0 12.5 13.3 12.4 12.5 12.9
## [31] 14.3 12.4 15.5 15.5 15.6 13.0 12.7 13.0 12.7 12.4 12.7 13.2 13.2 13.2 15.9
## [46] 13.3 12.9 12.6 12.6
```

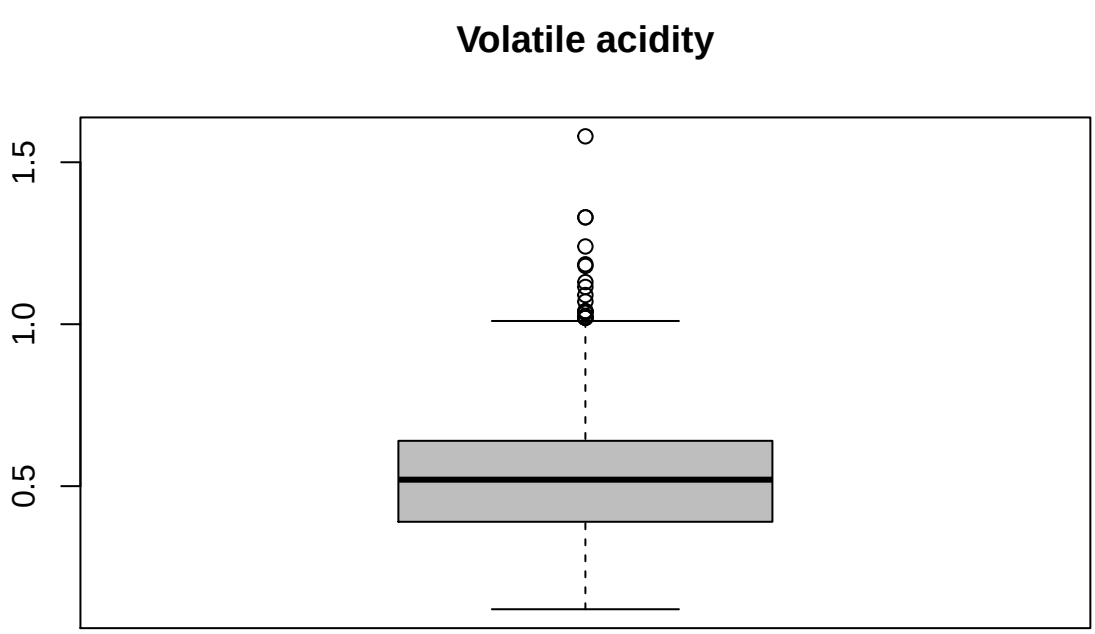
```
boxplot(wine_data$fixed_acidity,main="Fixed acidity Weight",col="gray")
```




```
boxplot.stats(wine_data$volatile_acidity)$out
```

```
##  [1] 1.130 1.020 1.070 1.330 1.330 1.040 1.090 1.040 1.240 1.185 1.020 1.035
## [13] 1.025 1.115 1.020 1.020 1.580 1.180 1.040
```

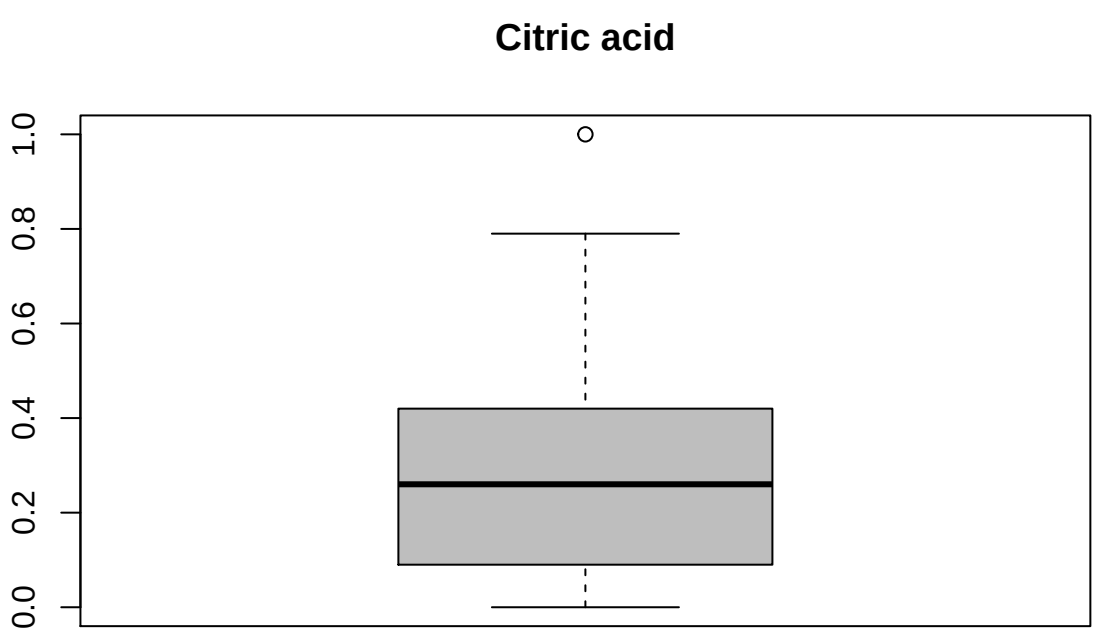
```
boxplot(wine_data$volatile_acidity,main="Volatile acidity",col="gray")
```




```
boxplot.stats(wine_data$citric_acid)$out
```

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

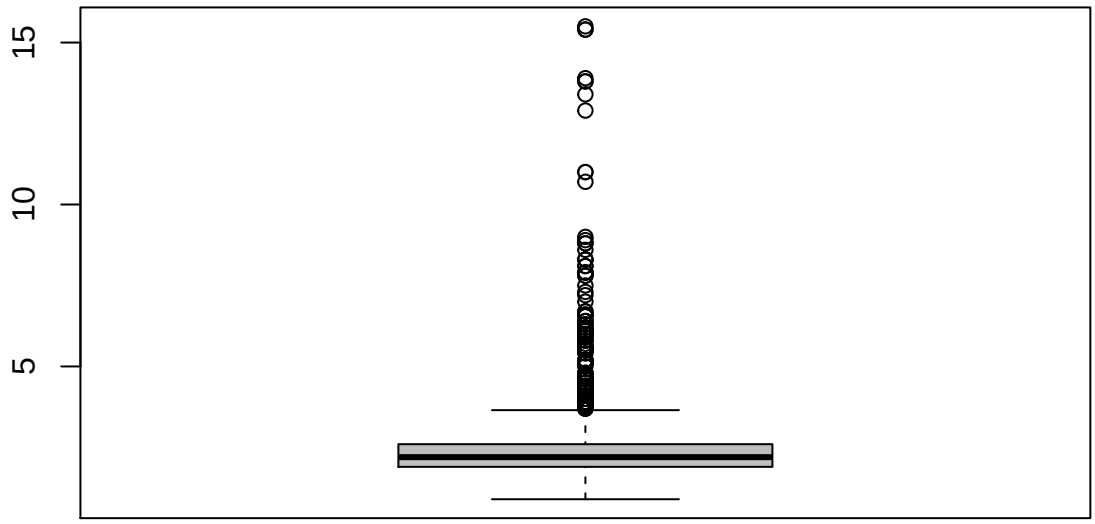
```
boxplot(wine_data$citric_acid,main="Citric acid",col="gray")
```




```
boxplot.stats(wine_data$residual_sugar)$out
```

```
##   [1]  6.10  6.10  3.80  3.90  4.40 10.70  5.50  5.90  5.90  3.80  5.10  4.65
##  [13]  4.65  5.50  5.50  5.50  5.50  7.30  7.20  3.80  5.60  4.00  4.00  4.00
##  [25]  4.00  7.00  4.00  4.00  6.40  5.60  5.60 11.00 11.00  4.50  4.80  5.80
##  [37]  5.80  3.80  4.40  6.20  4.20  7.90  7.90  3.70  4.50  6.70  6.60  3.70
##  [49]  5.20 15.50  4.10  8.30  6.55  6.55  4.60  6.10  4.30  5.80  5.15  6.30
##  [61]  4.20  4.20  4.60  4.20  4.60  4.30  4.30  7.90  4.60  5.10  5.60  5.60
##  [73]  6.00  8.60  7.50  4.40  4.25  6.00  3.90  4.20  4.00  4.00  4.00  6.60
##  [85]  6.00  6.00  3.80  9.00  4.60  8.80  8.80  5.00  3.80  4.10  5.90  4.10
##  [97]  6.20  8.90  4.00  3.90  4.00  8.10  8.10  6.40  6.40  8.30  8.30  4.70
## [109]  5.50  5.50  4.30  5.50  3.70  6.20  5.60  7.80  4.60  5.80  4.10 12.90
## [121]  4.30 13.40  4.80  6.30  4.50  4.50  4.30  4.30  3.90  3.80  5.40  3.80
## [133]  6.10  3.90  5.10  5.10  3.90 15.40 15.40  4.80  5.20  5.20  3.75 13.80
## [145] 13.80  5.70  4.30  4.10  4.10  4.40  3.70  6.70 13.90  5.10  7.80
```

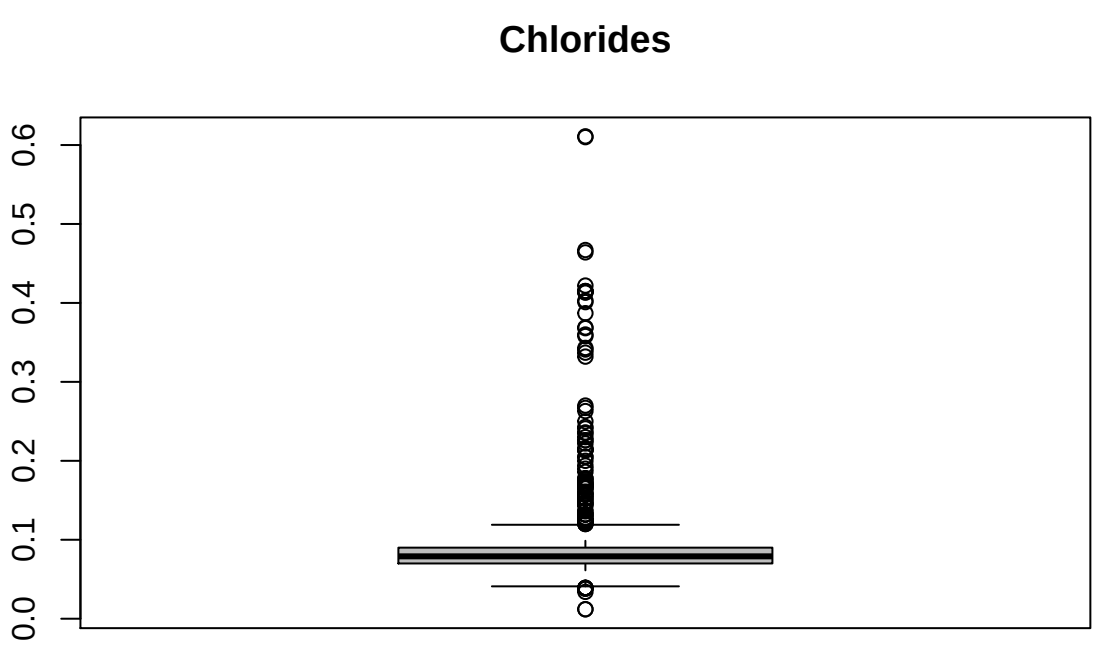
```
boxplot(wine_data$residual_sugar,main="Residual sugar",col="gray")
```




```
boxplot.stats(wine_data$chlorides)$out
```

```
##   [1] 0.176 0.170 0.368 0.341 0.172 0.332 0.464 0.401 0.467 0.122 0.178 0.146
##  [13] 0.236 0.610 0.360 0.270 0.039 0.337 0.263 0.611 0.358 0.343 0.186 0.213
##  [25] 0.214 0.121 0.122 0.122 0.128 0.120 0.159 0.124 0.122 0.122 0.174 0.121
##  [37] 0.127 0.413 0.152 0.152 0.125 0.122 0.200 0.171 0.226 0.226 0.250 0.148
##  [49] 0.122 0.124 0.124 0.143 0.222 0.039 0.157 0.422 0.034 0.387 0.415 0.157
##  [61] 0.157 0.243 0.241 0.190 0.132 0.126 0.038 0.165 0.145 0.147 0.012 0.012
##  [73] 0.039 0.194 0.132 0.161 0.120 0.120 0.123 0.123 0.414 0.216 0.171 0.178
##  [85] 0.369 0.166 0.166 0.136 0.132 0.132 0.123 0.123 0.123 0.403 0.137 0.414
##  [97] 0.166 0.168 0.415 0.153 0.415 0.267 0.123 0.214 0.214 0.169 0.205 0.205
## [109] 0.039 0.235 0.230 0.038
```

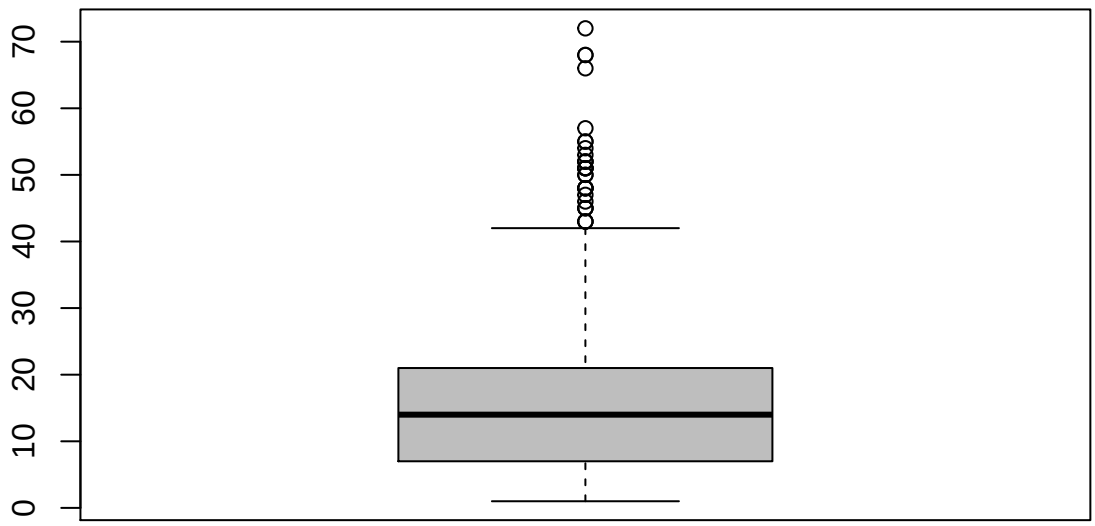
```
boxplot(wine_data$chlorides,main="Chlorides",col="gray")
```




```
boxplot.stats(wine_data$free_sulfur_dioxide)$out
```

```
##  [1] 52 51 50 68 68 43 47 54 46 45 53 52 51 45 57 50 45 48 43 48 72 43 51 51 52
## [26] 55 55 48 48 66
```

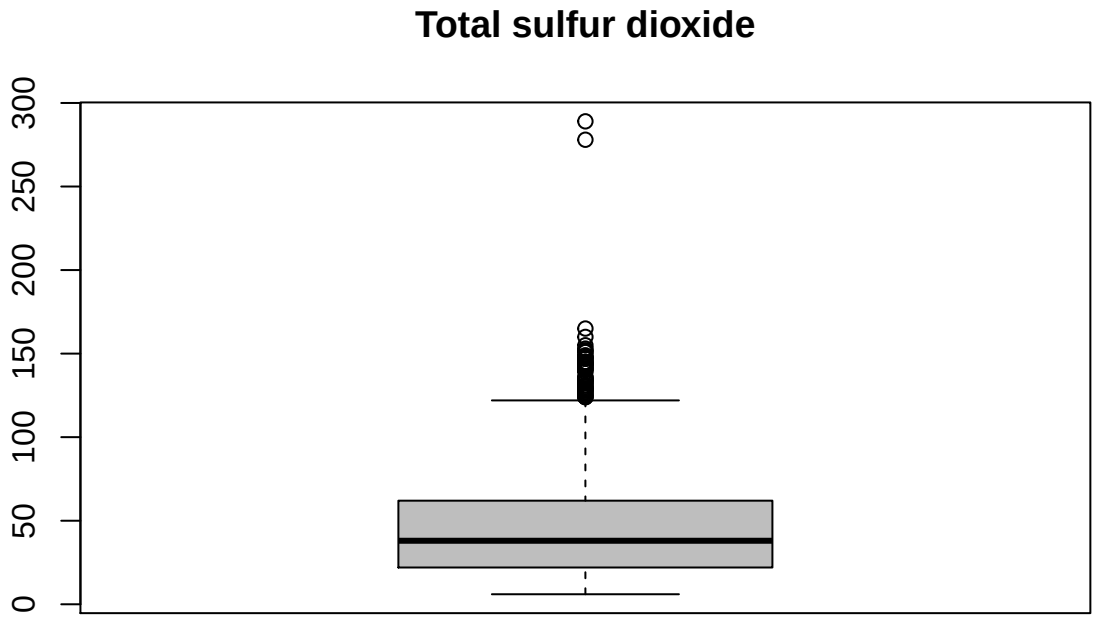
```
boxplot(wine_data$free_sulfur_dioxide,main="Free sulfur dioxide",col="gray")
```




```
boxplot.stats(wine_data$total_sulfur_dioxide)$out
```

```
##  [1] 145 148 136 125 140 136 133 153 134 141 129 128 129 128 143 144 127 126 145
## [20] 144 135 165 124 124 134 124 129 151 133 142 149 147 145 148 155 151 152 125
## [39] 127 139 143 144 130 278 289 135 160 141 141 133 147 147 131 131 131
```

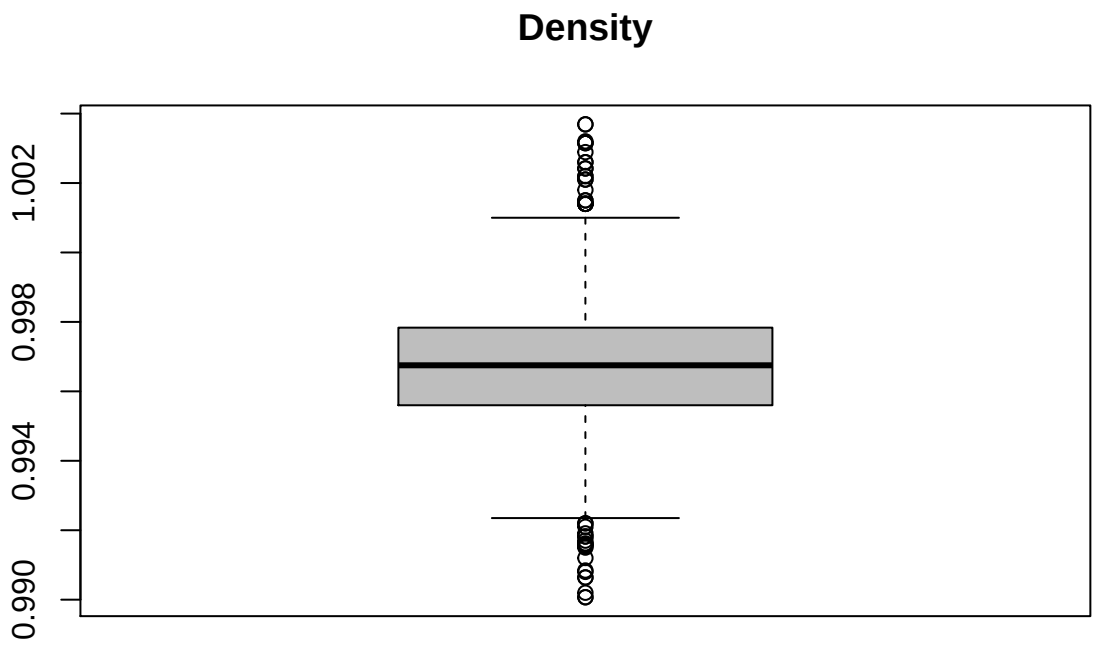
```
boxplot(wine_data$total_sulfur_dioxide,main="Total sulfur dioxide",col="gray")
```




```
boxplot.stats(wine_data$density)$out
```

```
##  [1] 0.99160 0.99160 1.00140 1.00150 1.00150 1.00180 0.99120 1.00220 1.00220
## [10] 1.00140 1.00140 1.00140 1.00140 1.00320 1.00260 1.00140 1.00315 1.00315
## [19] 1.00315 1.00210 1.00210 0.99170 0.99220 1.00260 0.99210 0.99154 0.99064
## [28] 0.99064 1.00289 0.99162 0.99007 0.99007 0.99020 0.99220 0.99150 0.99157
## [37] 0.99080 0.99084 0.99191 1.00369 1.00369 1.00242 0.99182 1.00242 0.99182
```

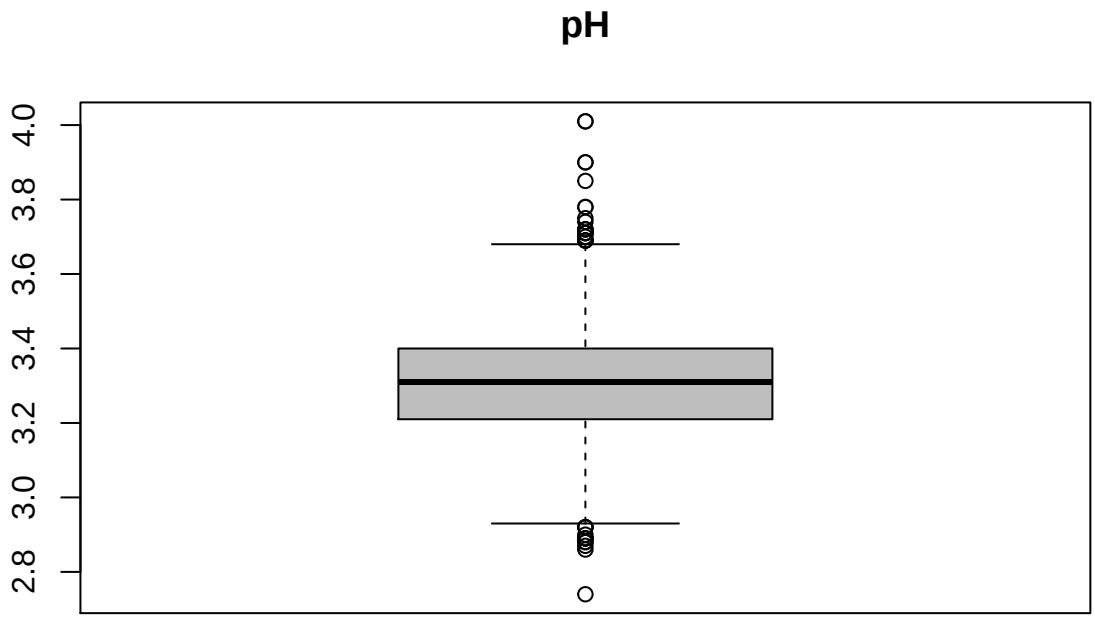
```
boxplot(wine_data$density,main="Density",col="gray")
```




```
boxplot.stats(wine_data$pH)$out
```

```
##  [1] 3.90 3.75 3.85 2.74 3.69 3.69 2.88 2.86 3.74 2.92 2.92 2.92 3.72 2.87 2.89
## [16] 2.89 2.92 3.90 3.71 3.69 3.69 3.71 3.71 2.89 2.89 3.78 3.70 3.78 4.01 2.90
## [31] 4.01 3.71 2.88 3.72 3.72
```

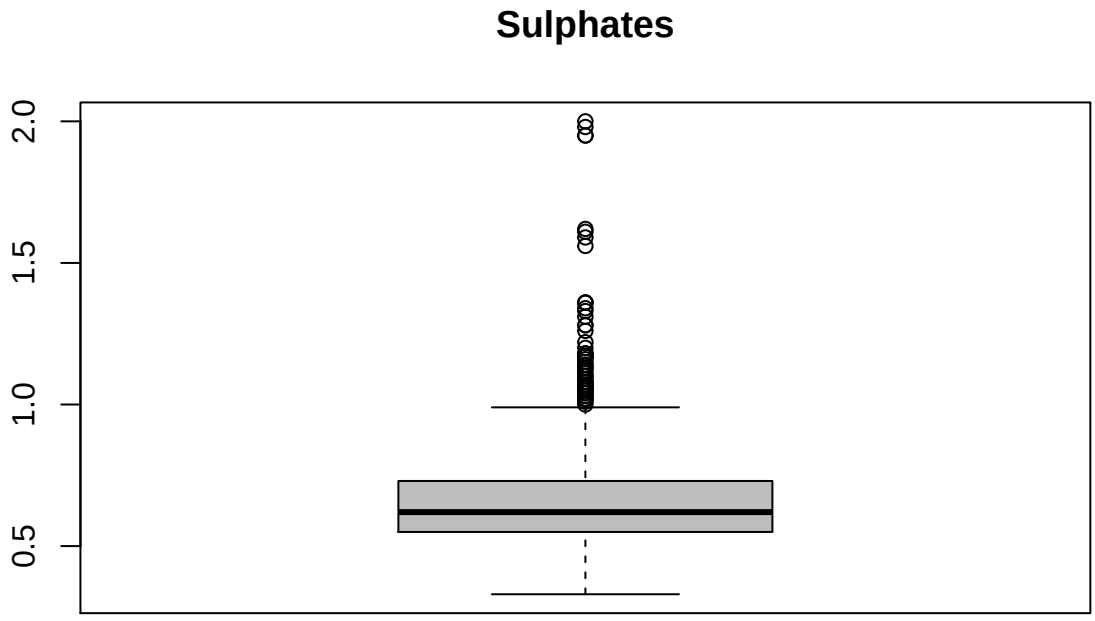
```
boxplot(wine_data$pH,main="pH",col="gray")
```




```
boxplot.stats(wine_data$sulphates)$out
```

```
##  [1] 1.56 1.28 1.08 1.20 1.12 1.28 1.14 1.95 1.22 1.95 1.98 1.31 2.00 1.08 1.59
## [16] 1.02 1.03 1.61 1.09 1.26 1.08 1.00 1.36 1.18 1.13 1.04 1.11 1.13 1.07 1.06
## [31] 1.06 1.05 1.06 1.04 1.05 1.02 1.14 1.02 1.36 1.36 1.05 1.17 1.62 1.06 1.18
## [46] 1.07 1.34 1.16 1.10 1.15 1.17 1.17 1.33 1.18 1.17 1.03 1.17 1.10 1.01
```

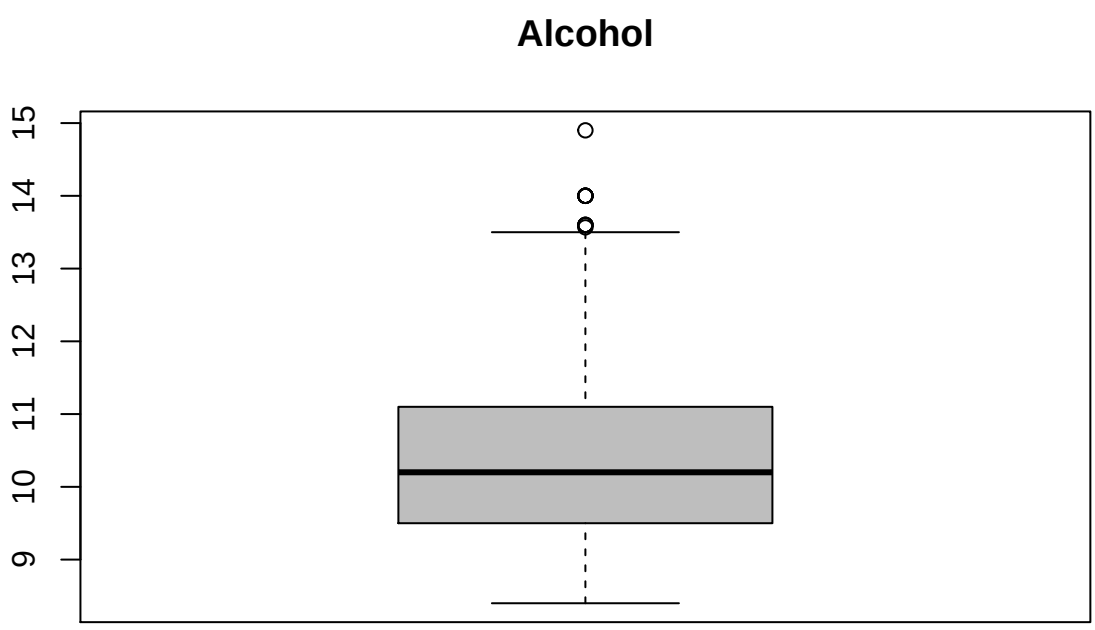
```
boxplot(wine_data$sulphates,main="Sulphates",col="gray")
```




```
boxplot.stats(wine_data$alcohol)$out
```

```
##  [1] 14.00000 14.00000 14.00000 14.00000 14.90000 14.00000 13.60000 13.60000
##  [9] 13.60000 14.00000 14.00000 13.56667 13.60000
```

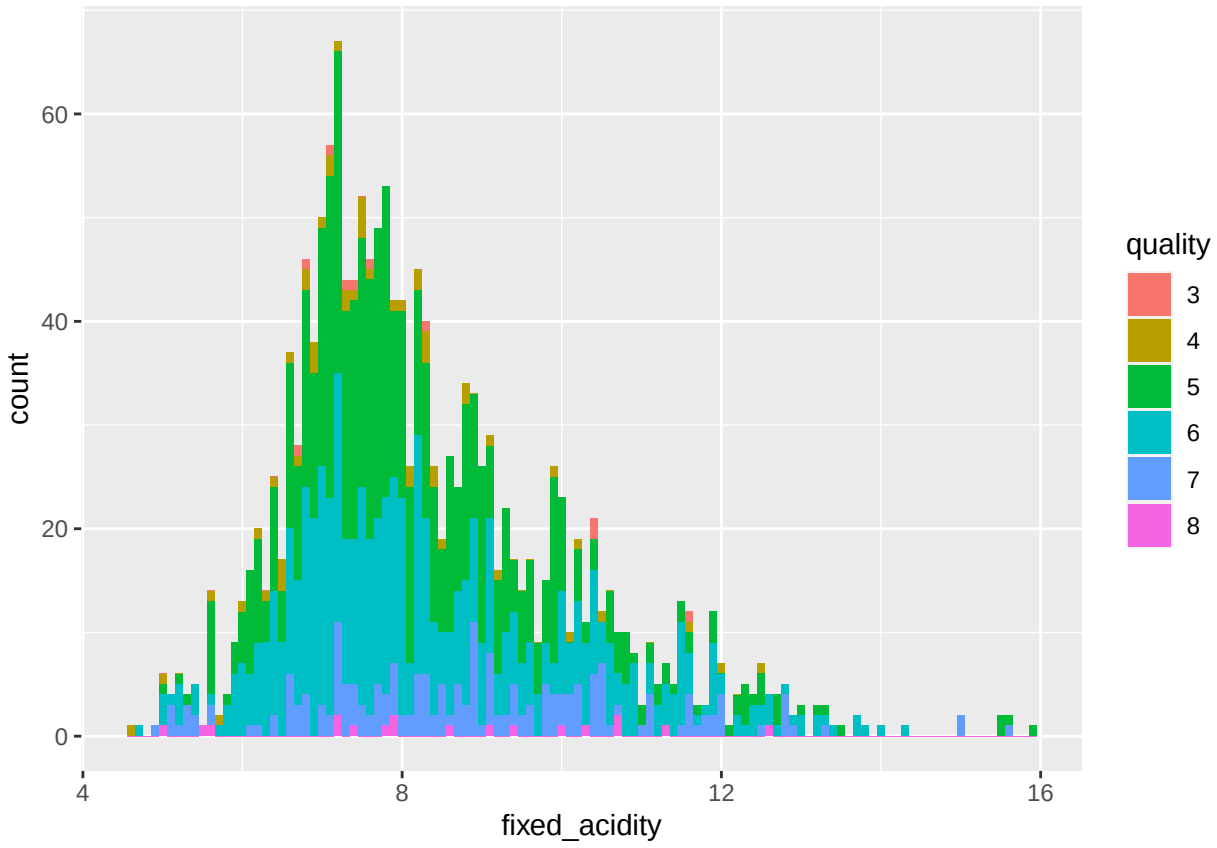
```
boxplot(wine_data$alcohol,main="Alcohol",col="gray")
```



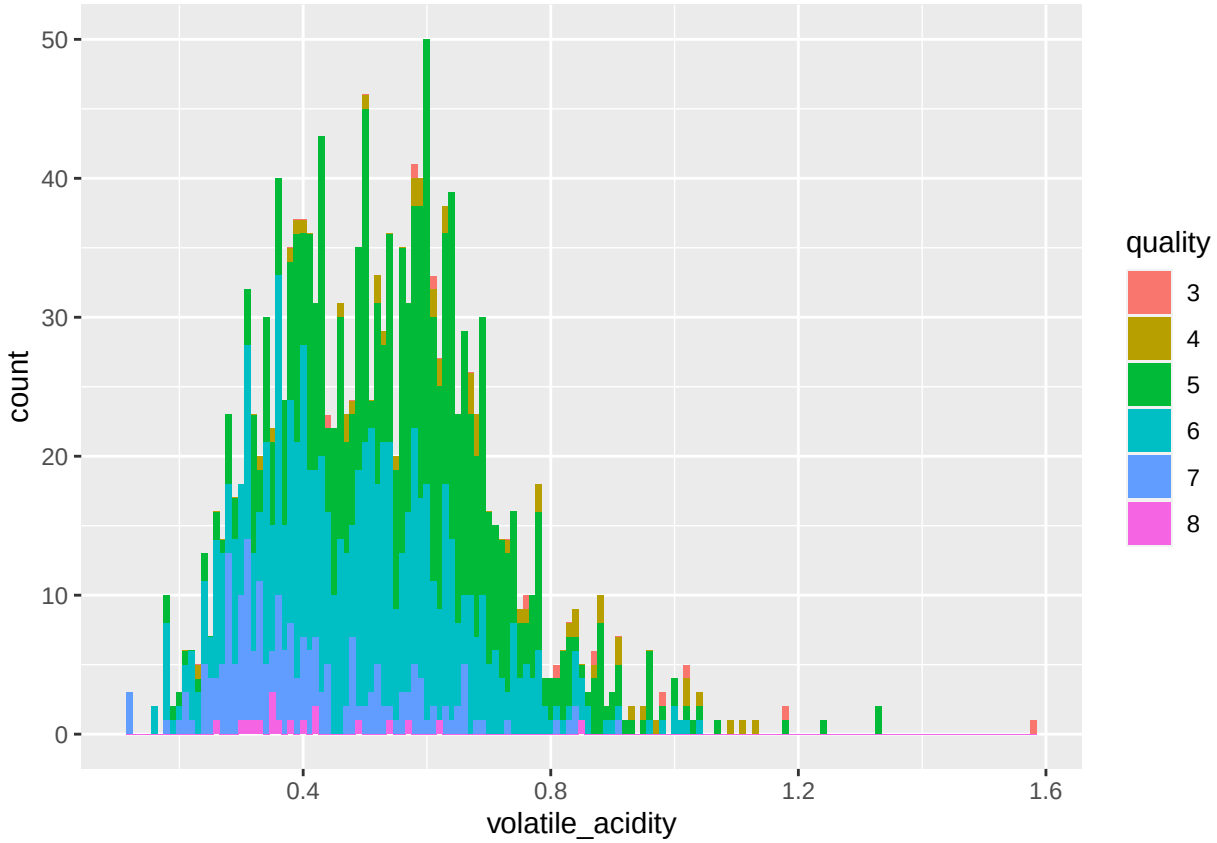
Aquests valors d'outliers són valors vàlids; estan dintre del rang de valors possibles a nivell químic. Això es veu més endavant en comprovar que les funcions de distribució de probabilitat d'aquests atributs no corresponen exactament a una distribució normal.

Faig una reprepresentació del l'histograma de les variables per tal de veure com es la funció de distribució d'aquestes i la contribució d'aquestes a la qualitat del vi:

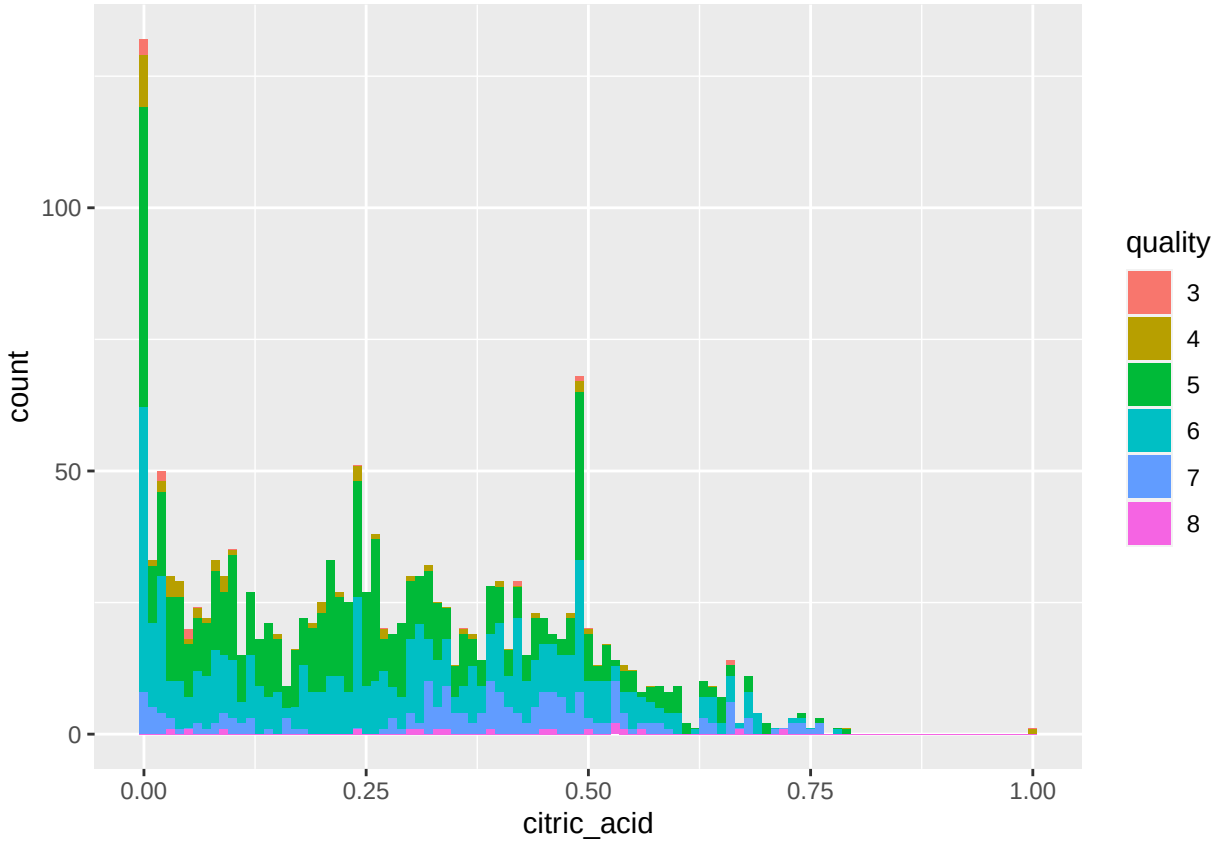
```
## Analitzem

filas=dim(wine_data)[1]

ggplot(data = wine_data[1:filas,],aes(x=fixed_acidity,fill=quality))+ geom_histogram(binwidth = 0.1)
```



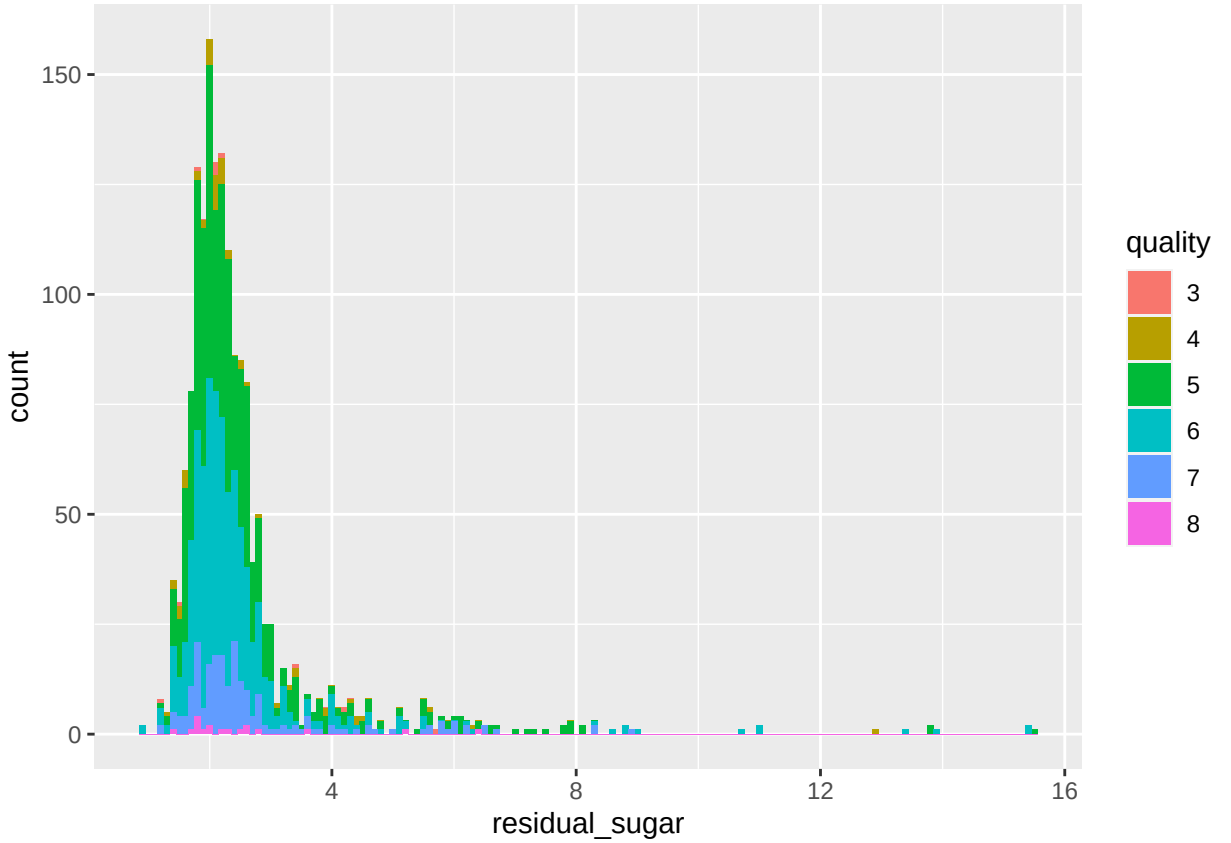

```
ggplot(data = wine_data[1:filas,],aes(x=volatile_acidity,fill=quality))+ geom_histogram(binwidth = 0.01)
```



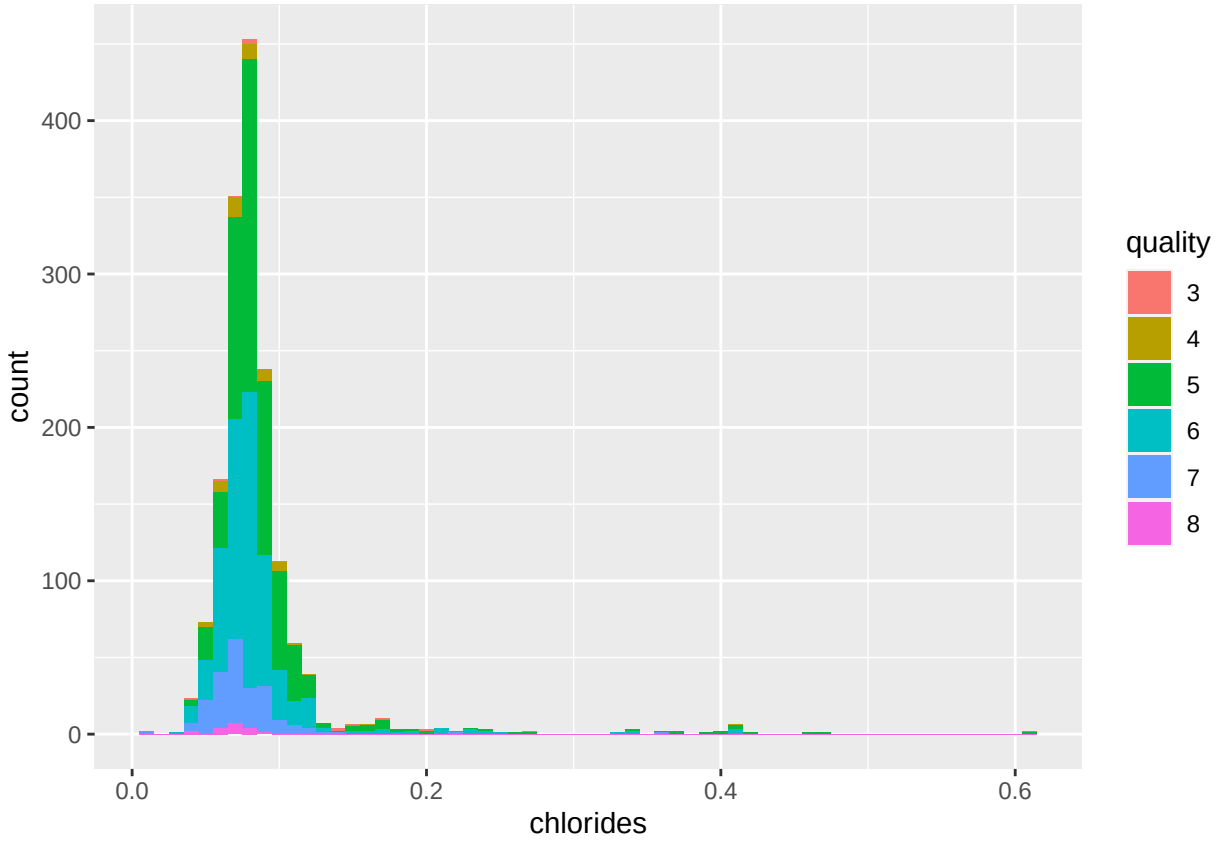

```
ggplot(data = wine_data[1:filas,],aes(x=citric_acid,fill=quality))+ geom_histogram(binwidth = 0.01)
```



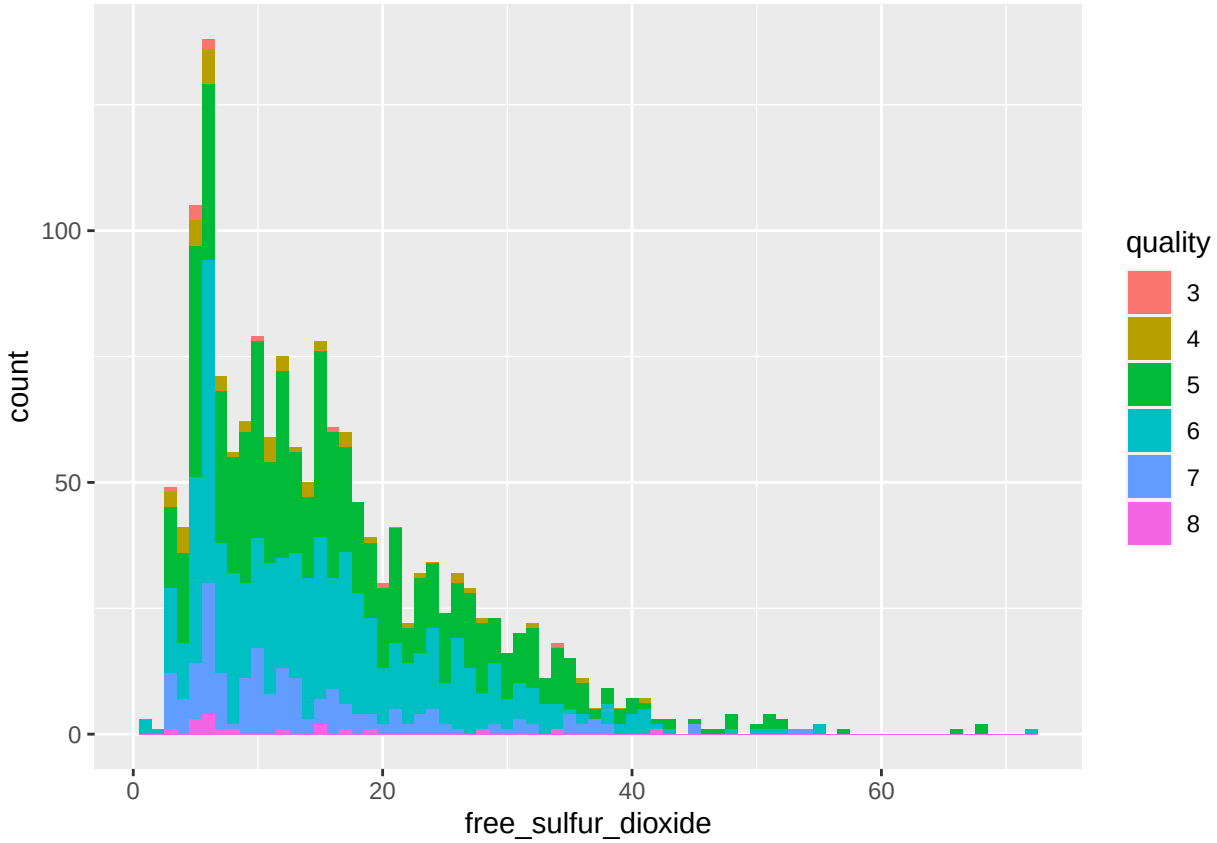

```
ggplot(data = wine_data[1:filas,],aes(x=residual_sugar,fill=quality))+geom_histogram(binwidth = 0.1)
```



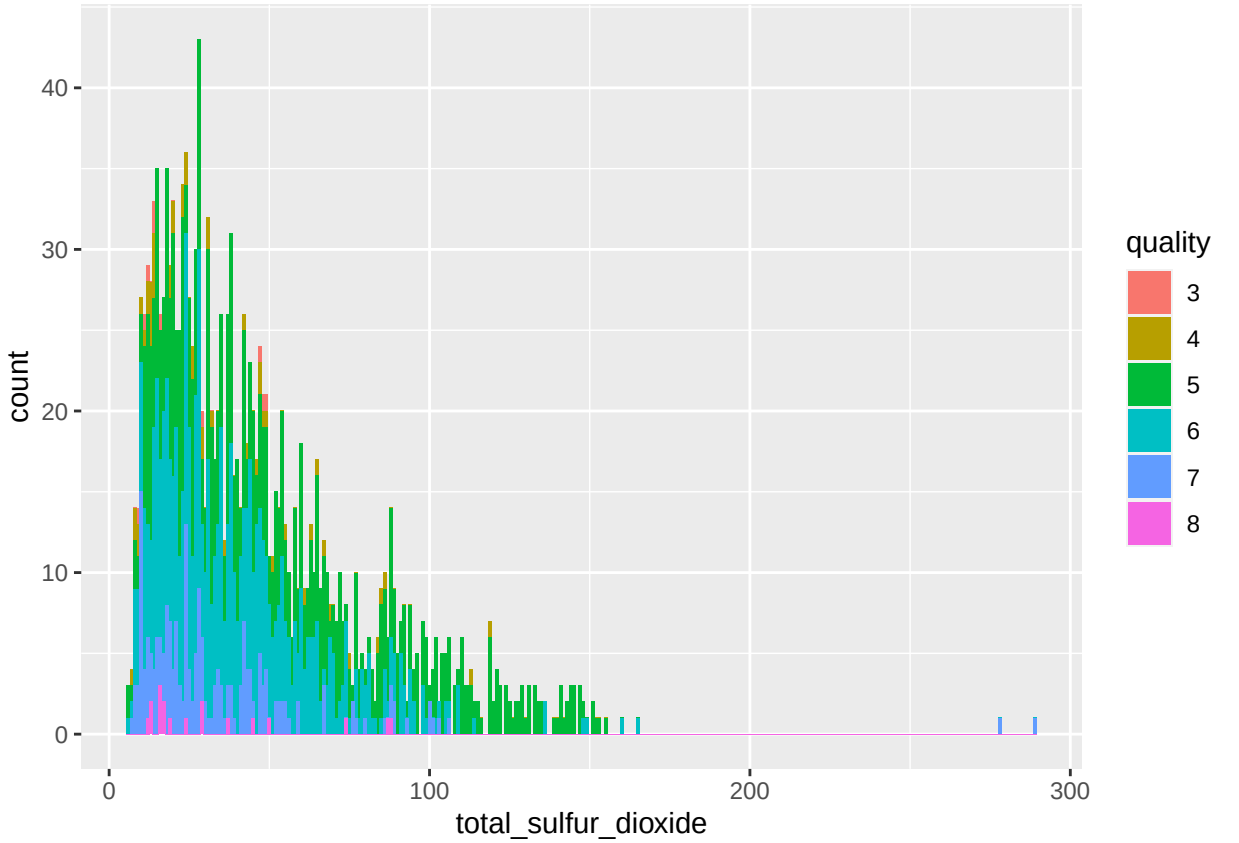

```
ggplot(data = wine_data[1:filas,],aes(x=chlorides,fill=quality))+geom_histogram(binwidth = 0.01)
```



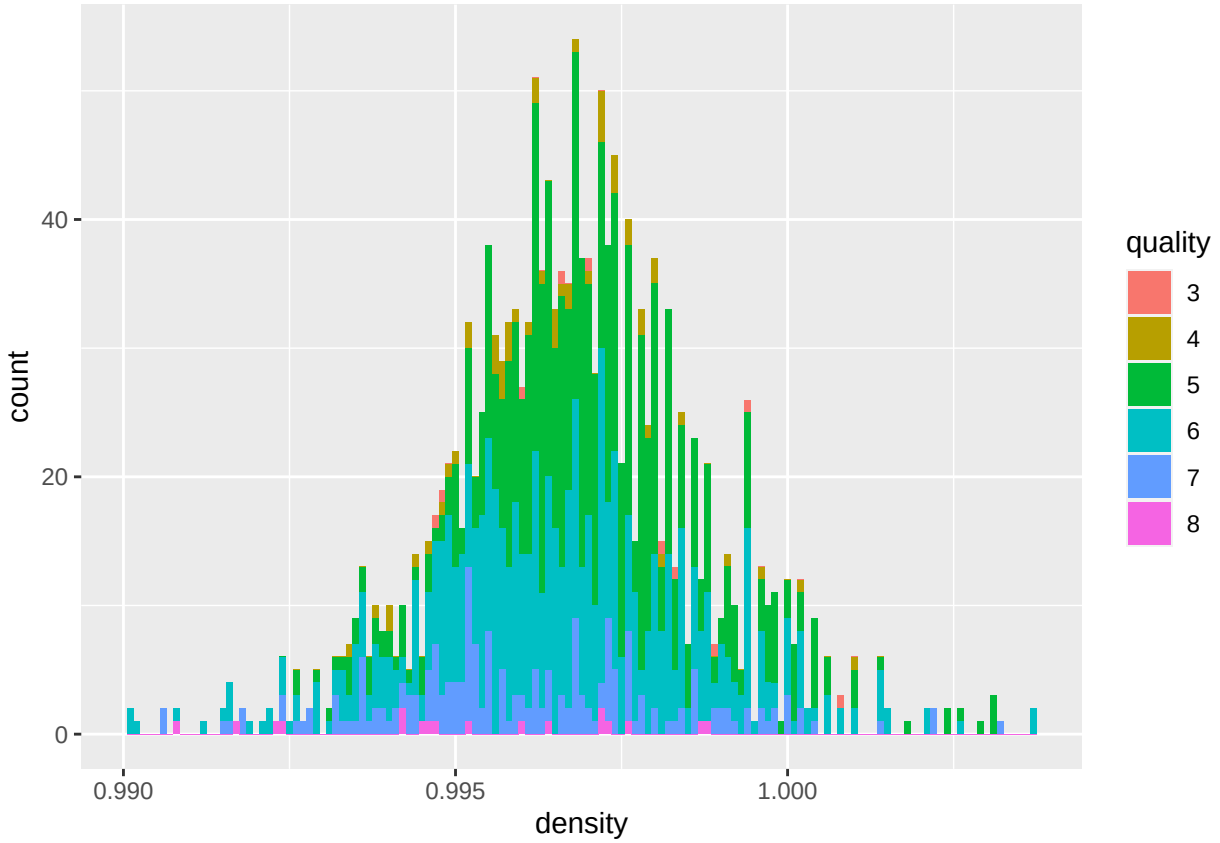

```
ggplot(data = wine_data[1:filas,],aes(x=free_sulfur_dioxide,fill=quality))+geom_histogram(binwidth = 1)
```



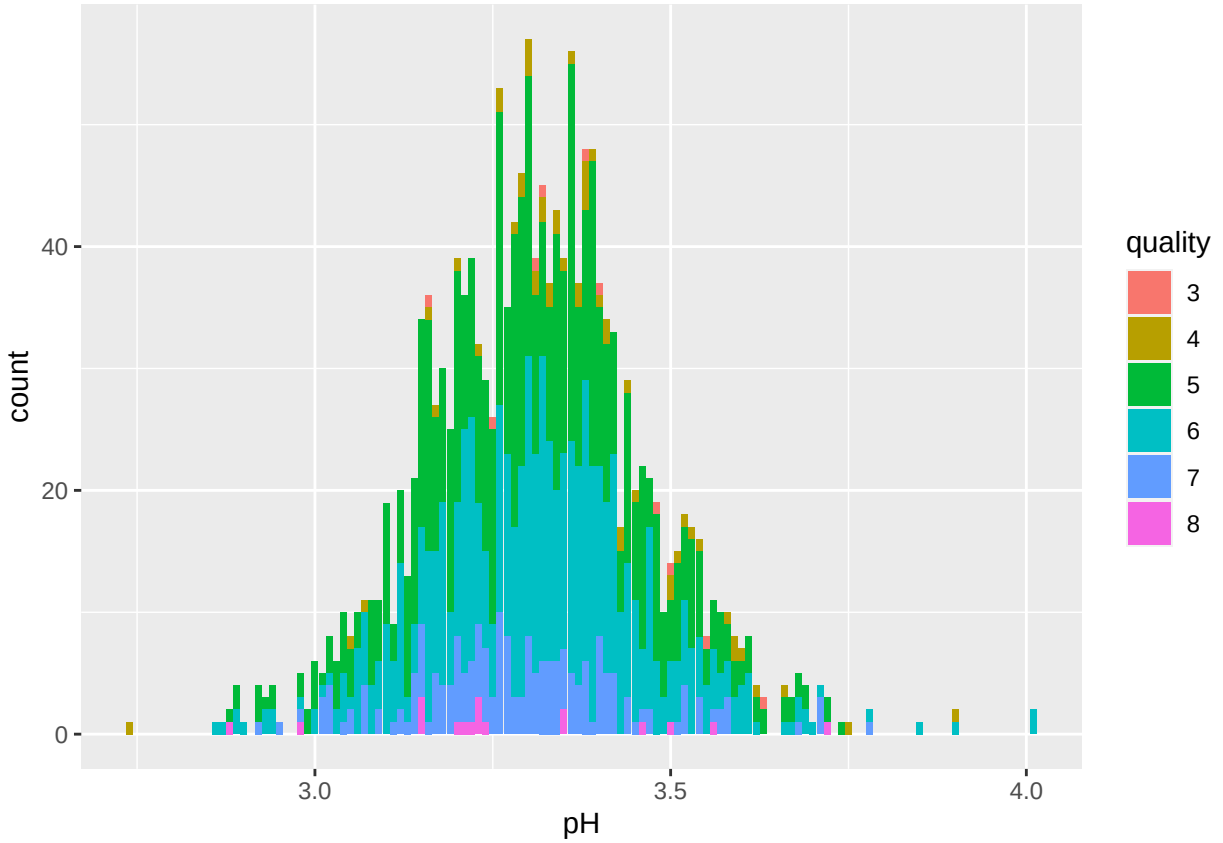

```
ggplot(data = wine_data[1:filas,],aes(x=total_sulfur_dioxide,fill=quality))+geom_histogram(binwidth = 1
```



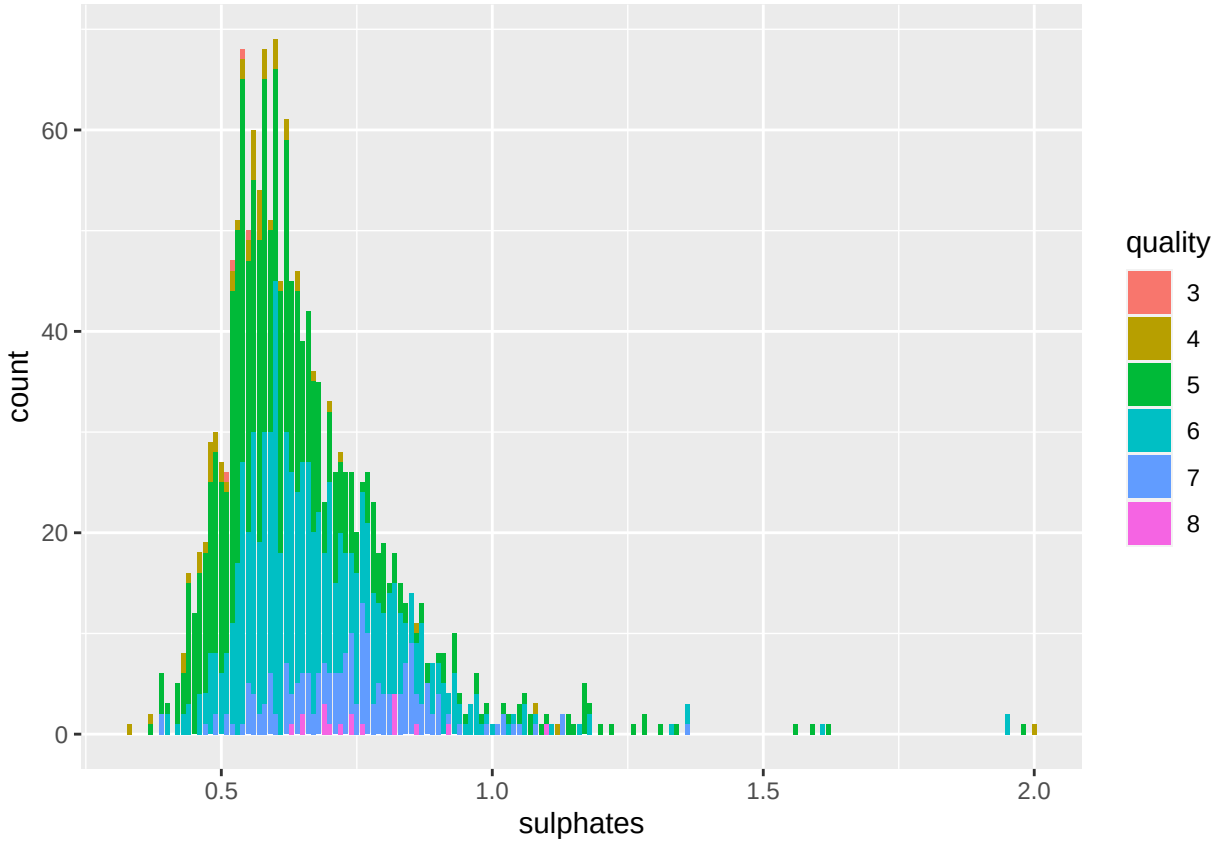

```
ggplot(data = wine_data[1:filas,],aes(x=density,fill=quality))+ geom_histogram(binwidth = 0.0001)
```



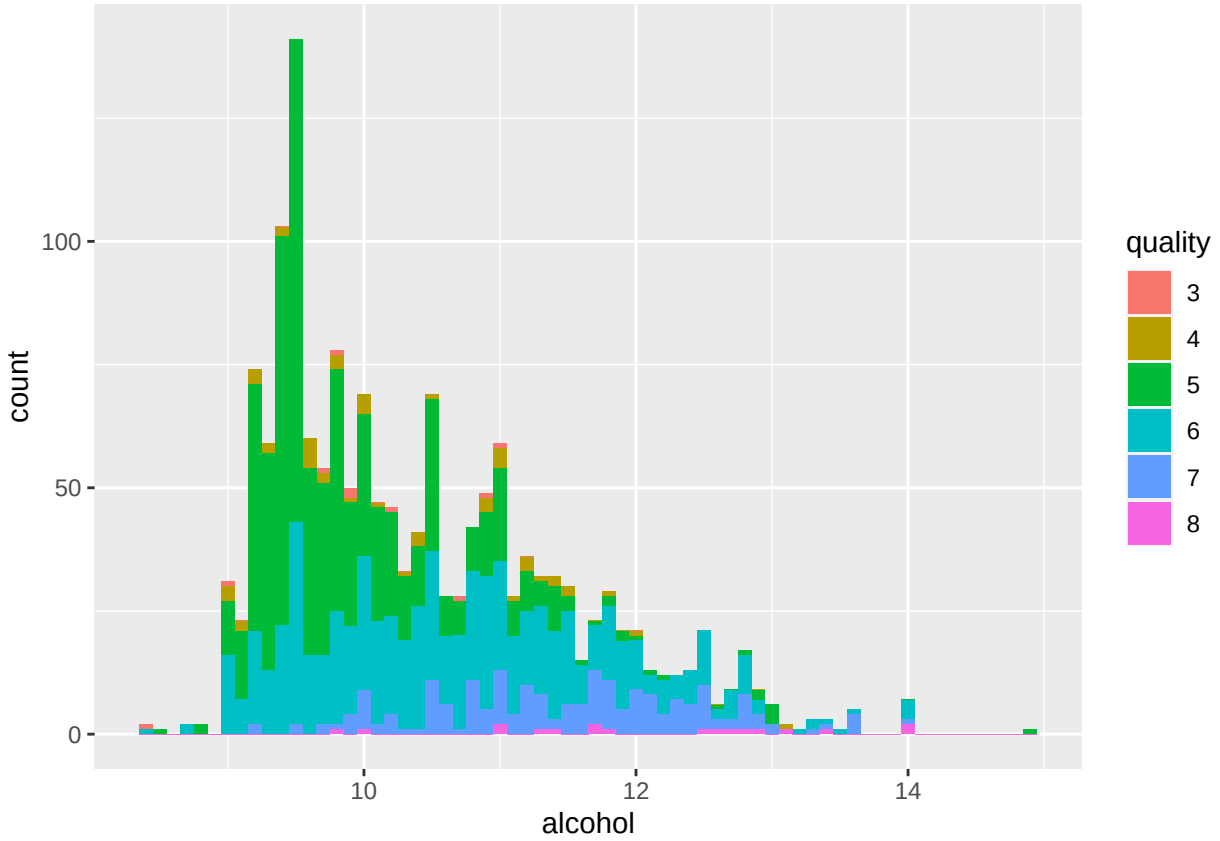

```
ggplot(data = wine_data[1:filas,],aes(x=pH,fill=quality))+geom_bar()
```



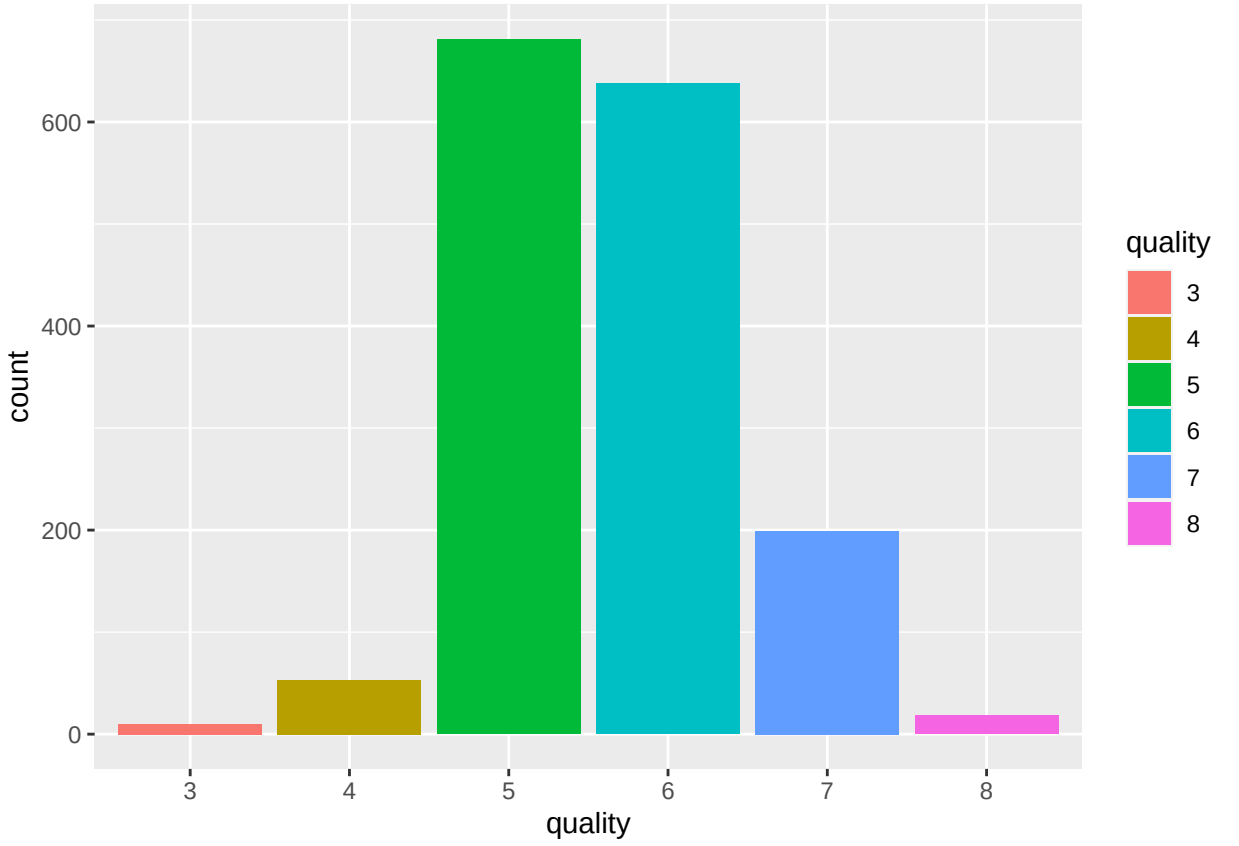

```
ggplot(data = wine_data[1:filas,],aes(x=sulphates,fill=quality))+geom_bar()
```




```
ggplot(data = wine_data[1:filas,],aes(x=alcohol,fill=quality))+geom_histogram(binwidth = 0.1)
```



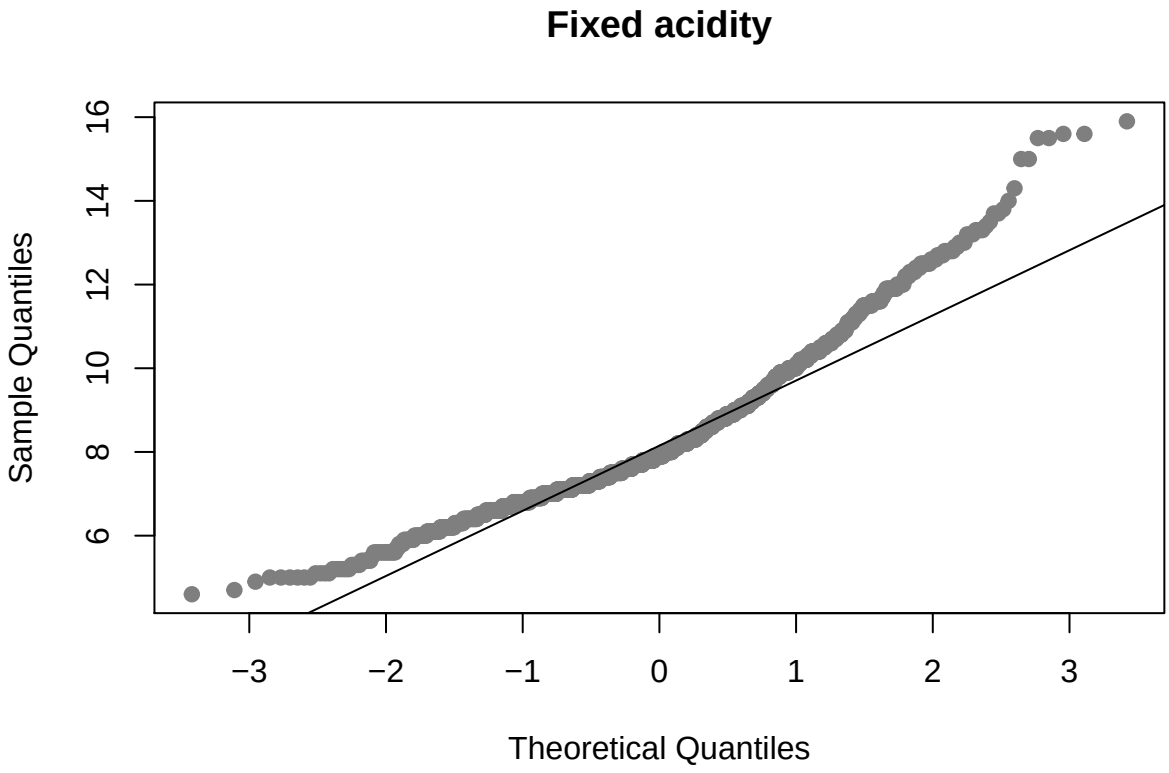

```
ggplot(data = wine_data[1:filas,],aes(x=quality,fill=quality))+geom_bar()
```



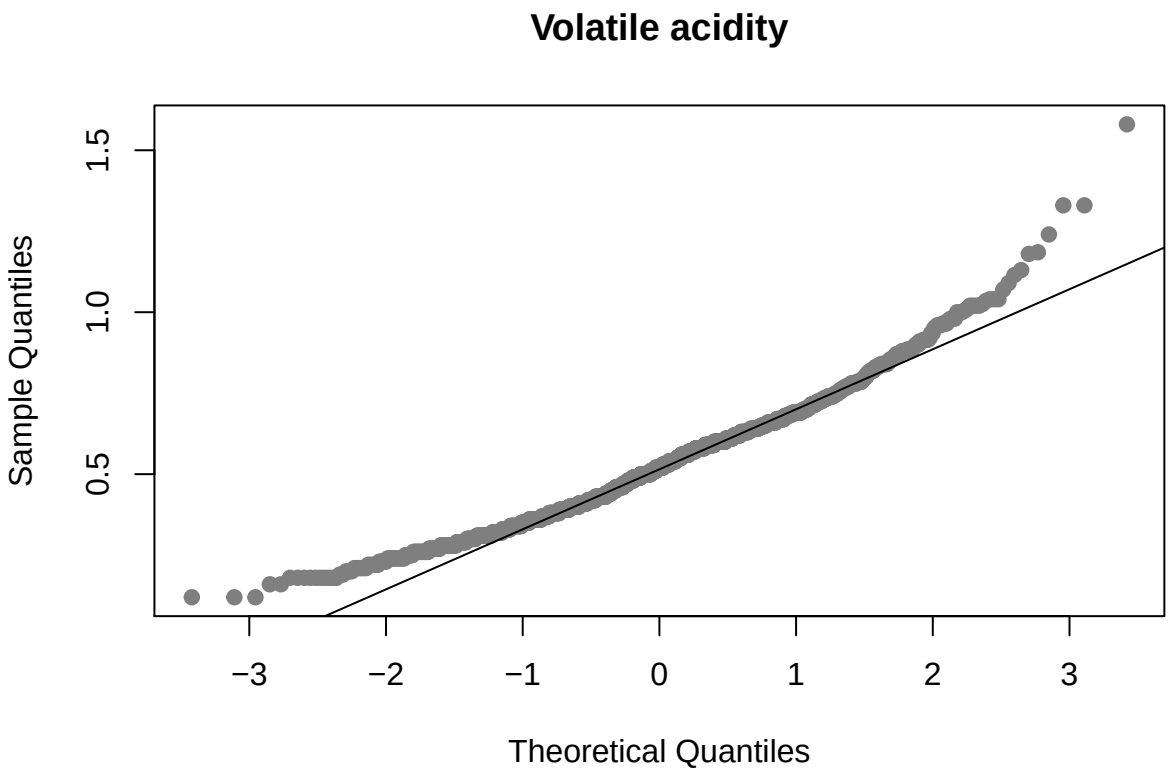
Clàrament es veu que la major part dels valors dels paràmetres que contribueixen a la qualitat del ví estan acotats en una serie de valors i no n'hi han valors gaire predominants (una excepció serien els clorures o "chlorides"); aquestes distribucions no son del tot normals.

Ara vaig a mostrar gràfics QQ-plots per veure l'aproximació que tenen a la normalitat

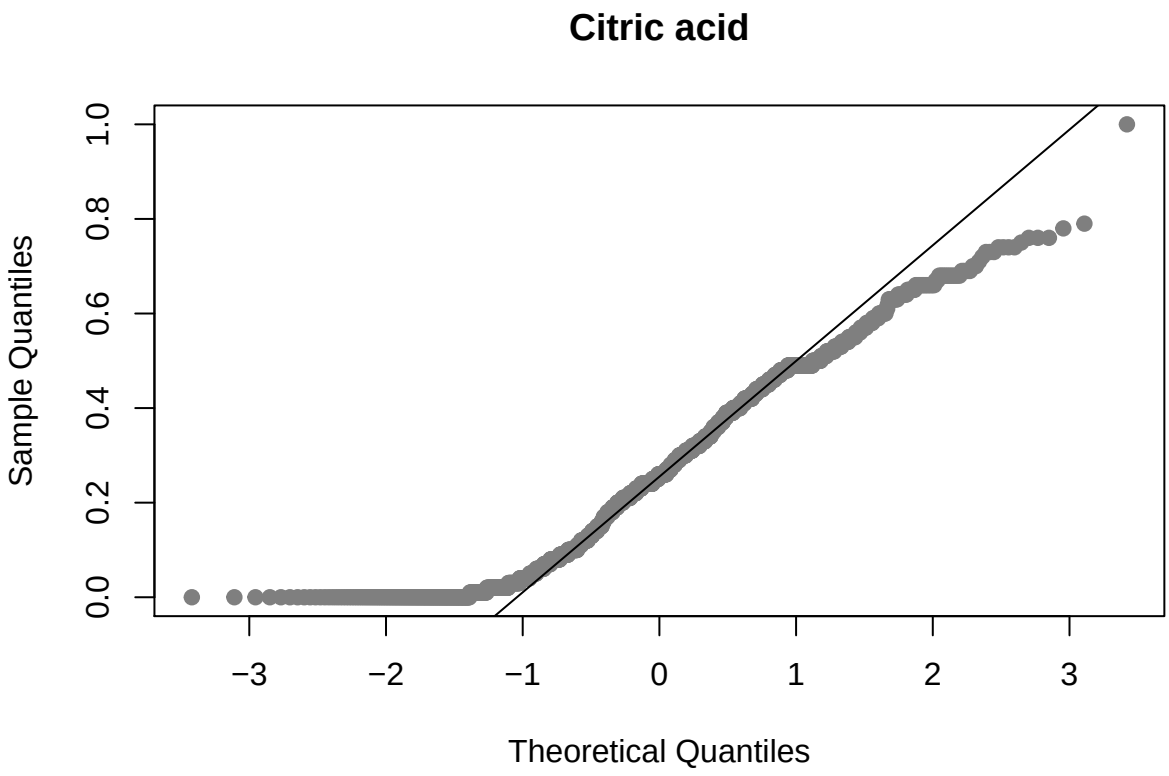
```
qqnorm(wine_data$fixed_acidity, main = "Fixed acidity",pch = 19, col = "gray50")
qqline(wine_data$fixed_acidity)
```



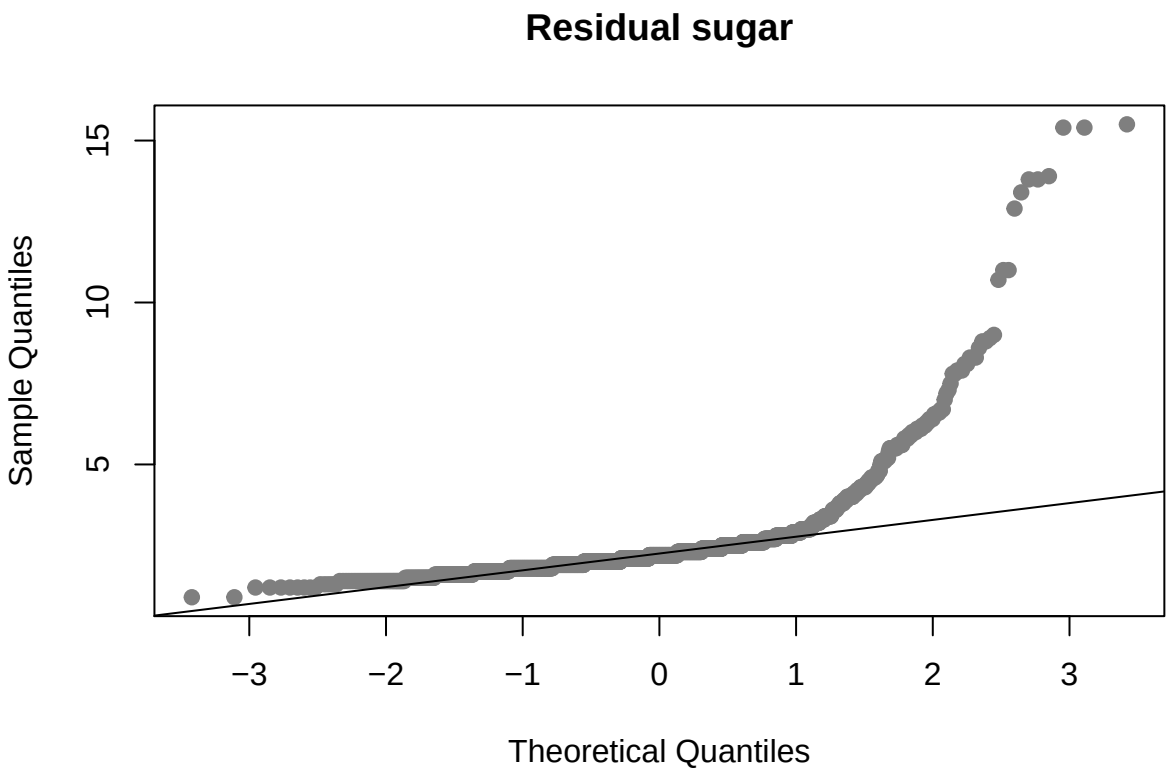

```
qqnorm(wine_data$volatile_acidity, main = "Volatile acidity", pch = 19, col = "gray50")
qqline(wine_data$volatile_acidity)
```



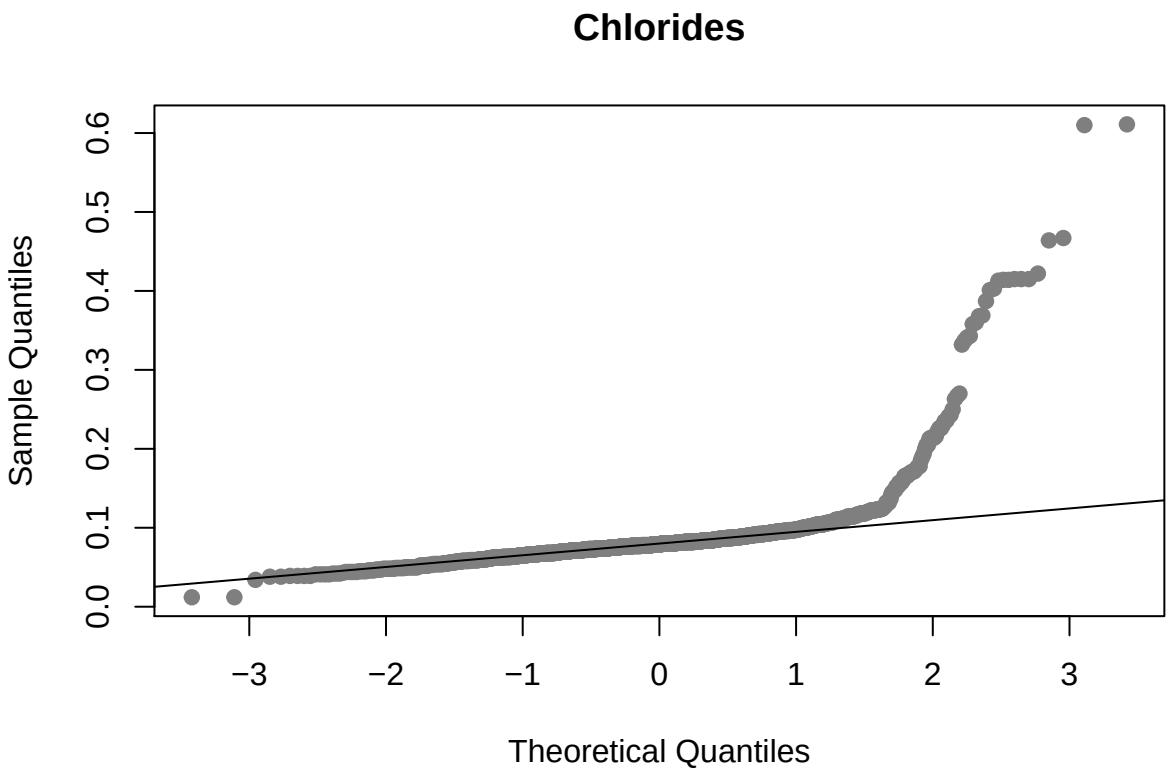

```
qqnorm(wine_data$citric_acid, main = "Citric acid", pch = 19, col = "gray50")
qqline(wine_data$citric_acid)
```



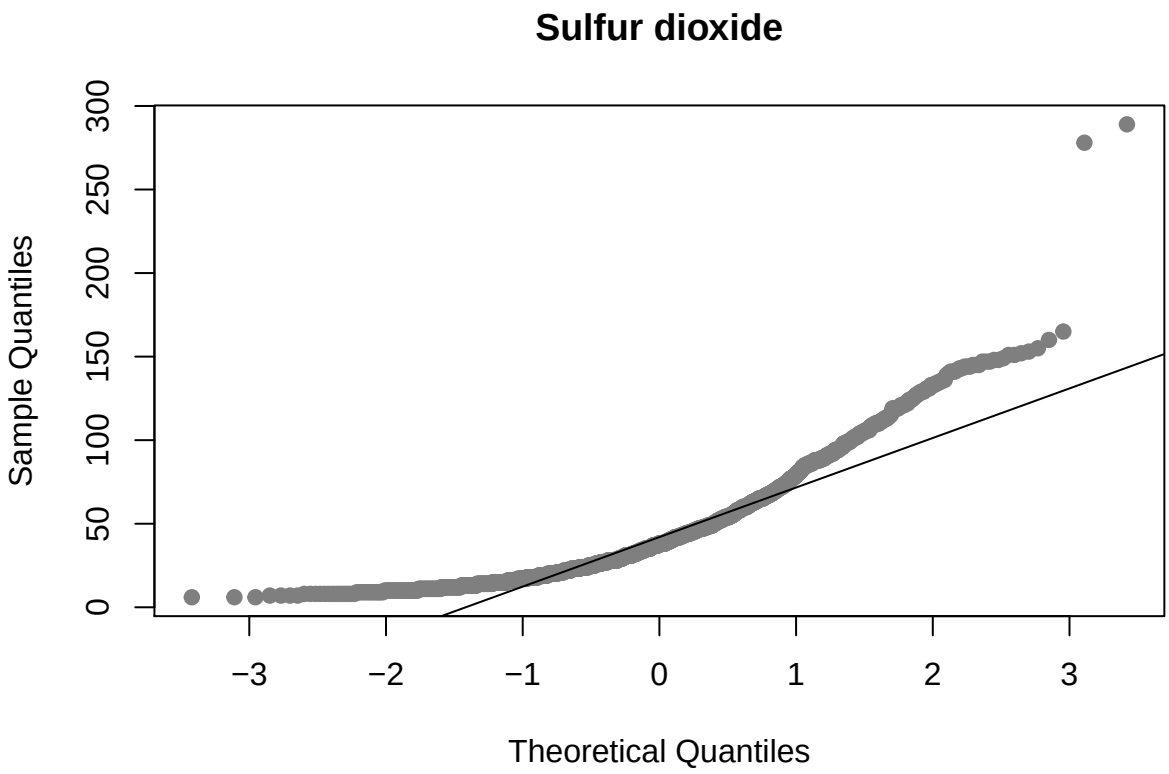

```
qqnorm(wine_data$residual_sugar,  main = "Residual sugar", pch = 19, col = "gray50")
qqline(wine_data$residual_sugar)
```



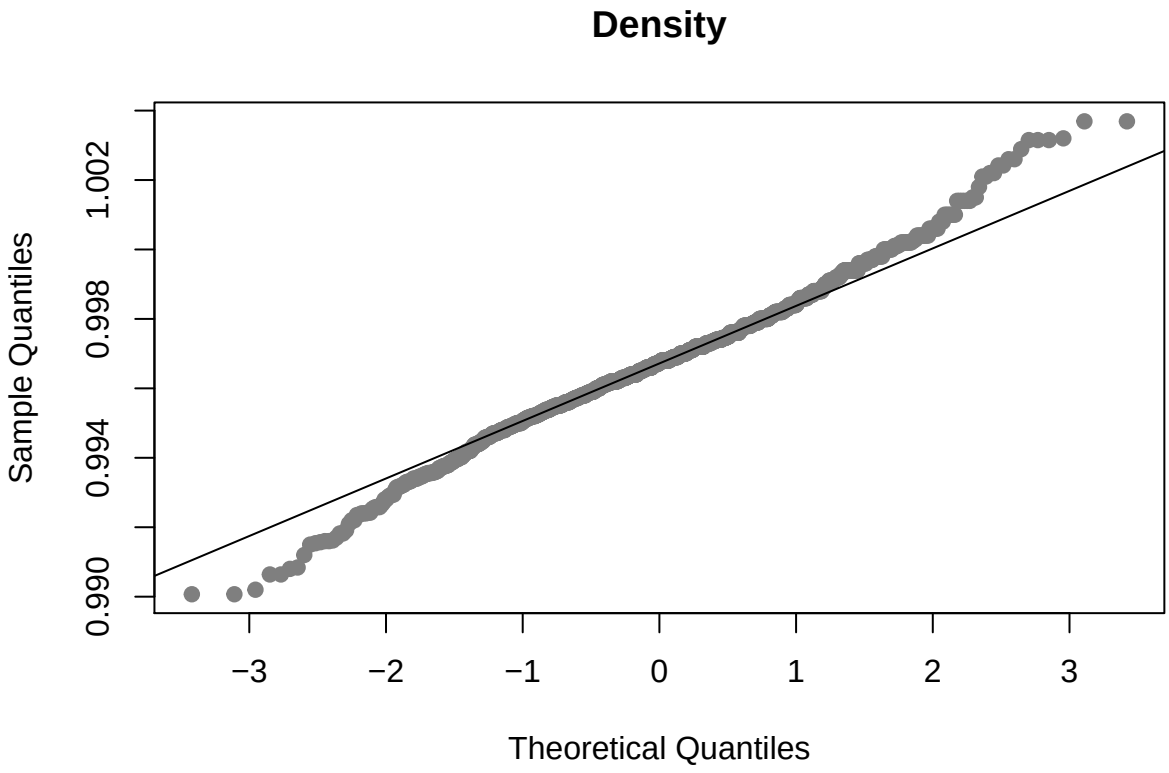

```
qqnorm(wine_data$chlorides, main = "Chlorides", pch = 19, col = "gray50")
qqline(wine_data$chlorides)
```



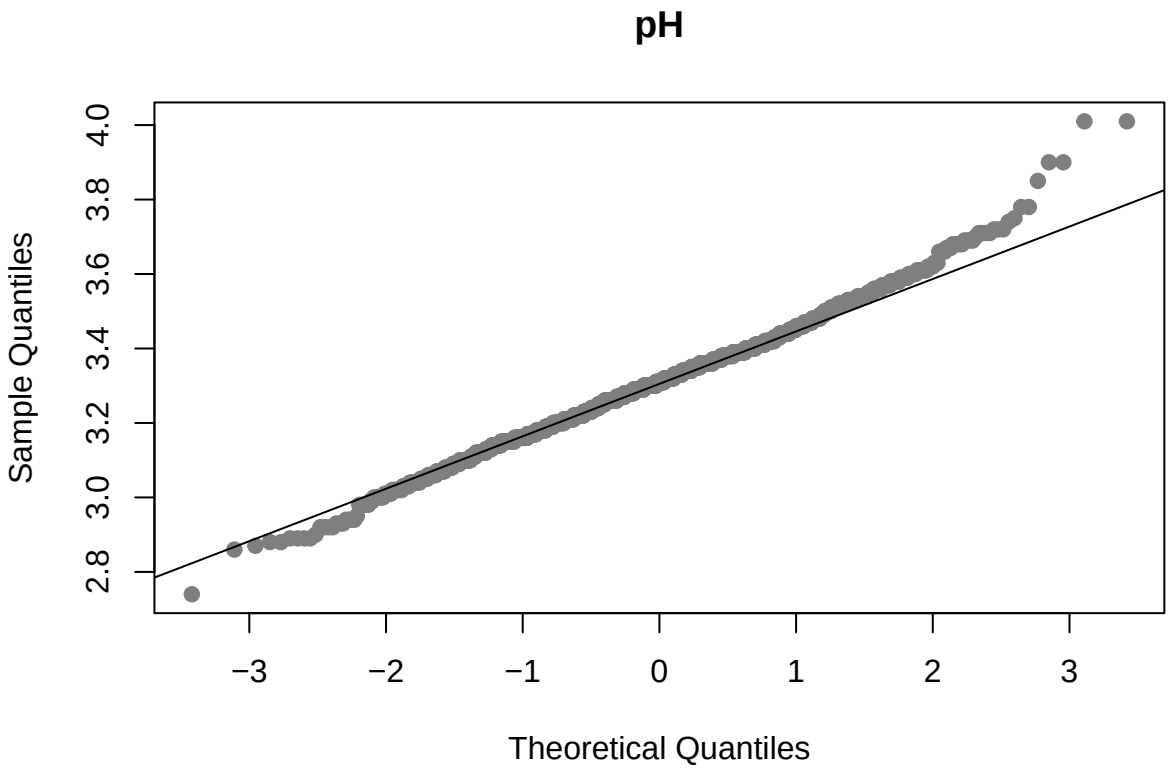

```
qqnorm(wine_data$total_sulfur_dioxide,  main = "Sulfur dioxide",pch = 19, col = "gray50")
qqline(wine_data$total_sulfur_dioxide)
```



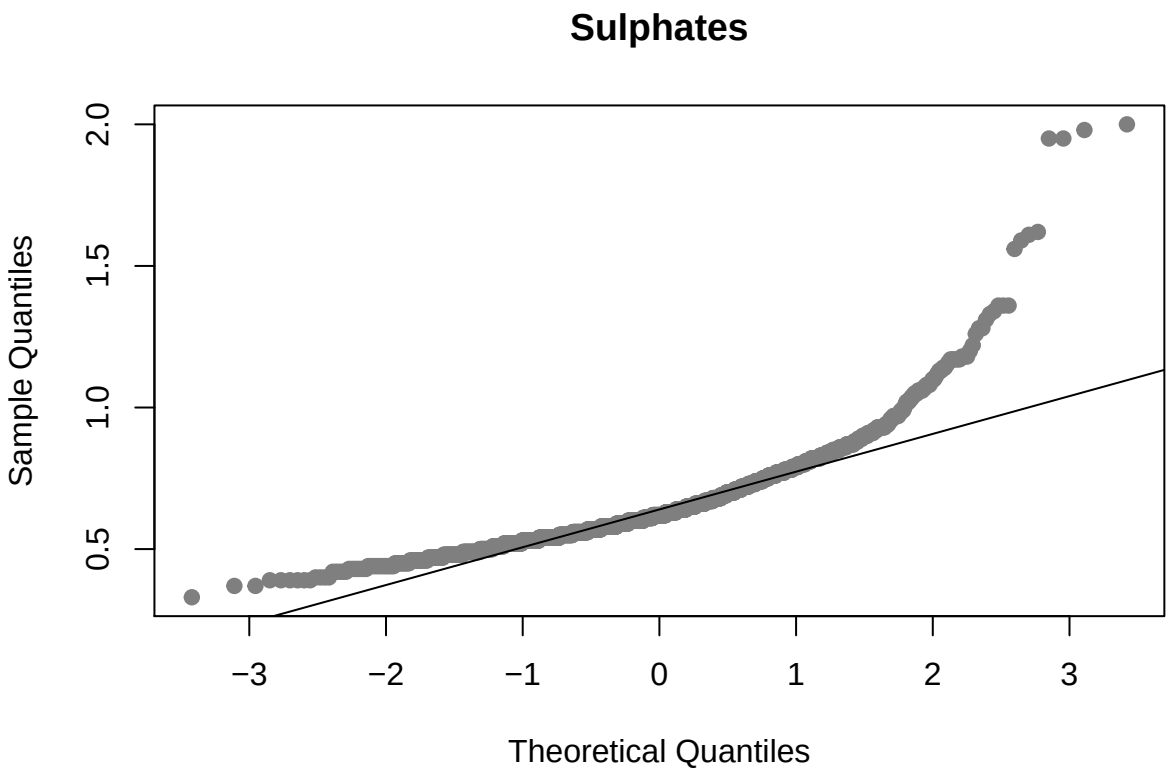

```
qqnorm(wine_data$density, pch = 19, main = "Density",col = "gray50")
qqline(wine_data$density)
```



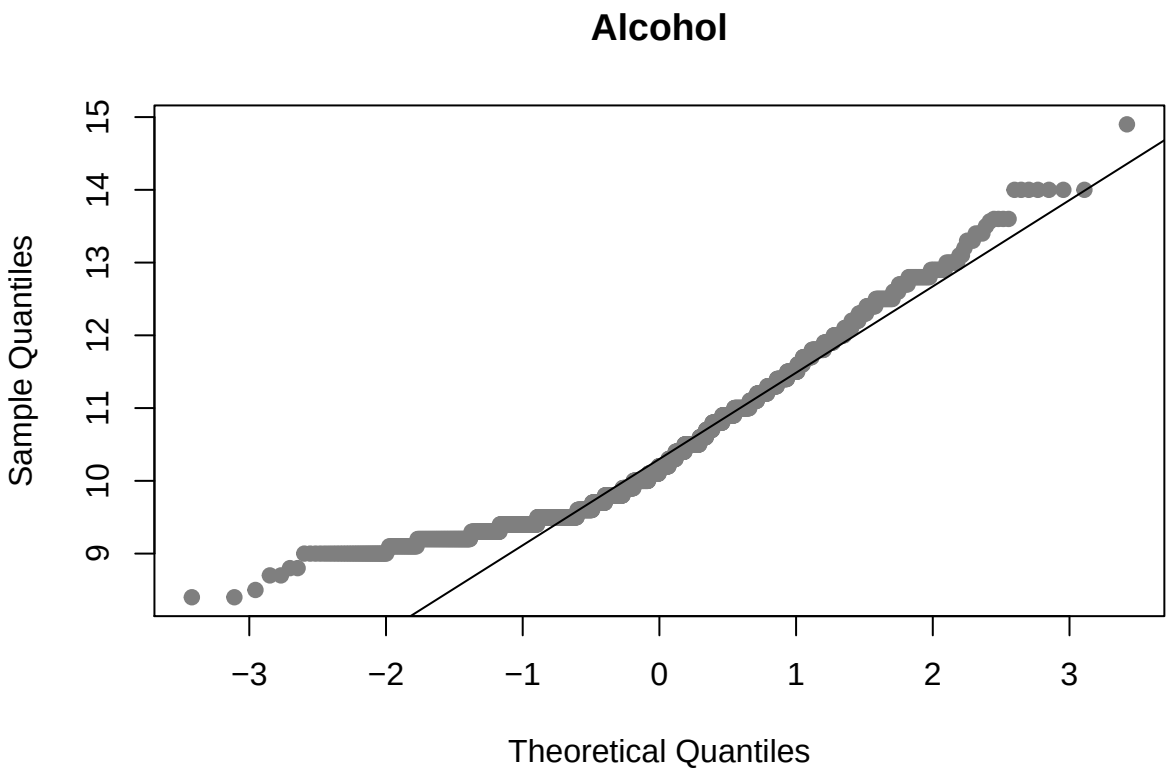

```
qqnorm(wine_data$pH, pch = 19, main = "pH", col = "gray50")
qqline(wine_data$pH)
```




```
qqnorm(wine_data$sulphates, main = "Sulphates", pch = 19, col = "gray50")
qqline(wine_data$sulphates)
```



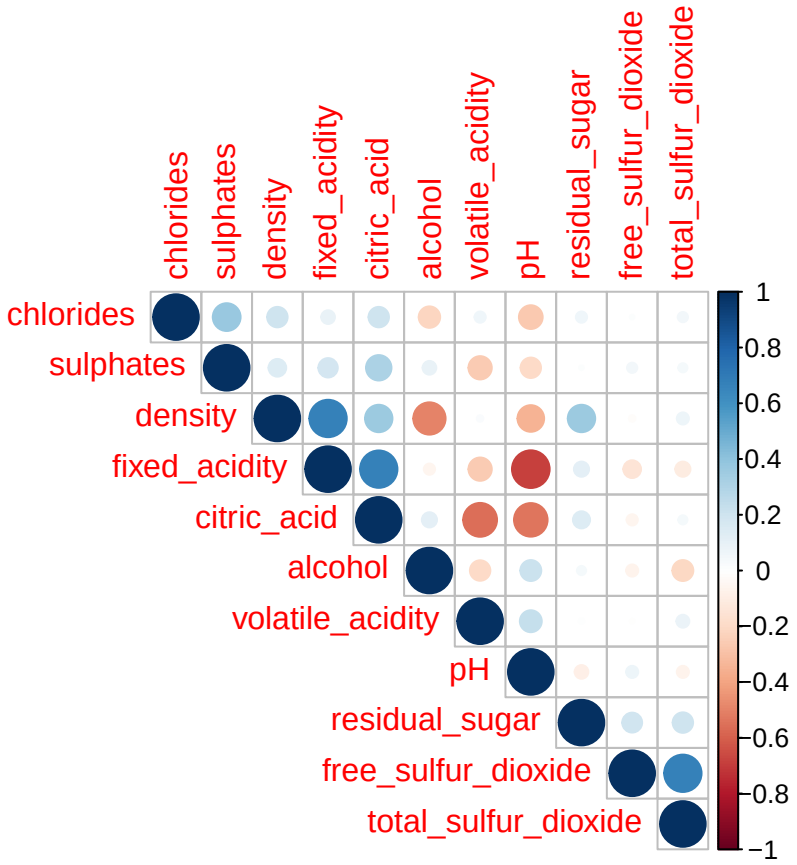

```
qqnorm(wine_data$alcohol, main = "Alcohol", pch = 19, col = "gray50")
qqline(wine_data$alcohol)
```



La variable que mes s'aproxima a la normalitat es el pH.

Un aspecte molt importat a l'hora de seleccionar els atributs estudiar la correlació de les variables no depenents; a la següent gràfica es poden veure molt clarament les correlacions entre variables; els colors que són molt intens estan molt correlacionats.

```
# Visualitzo la correlació
mcor<-round(cor(wine_data[,-12]),2)
corrplot(mcor, type= "upper", order ="hclust", t1.col="black", t1.srt=45)
```




```
mcor2 <-round(cor(wine_data[,c(1,3,6,7,8,9)],method = "spearman"),2)
mcor2
```

```
##                       fixed_acidity citric_acid free_sulfur_dioxide
## fixed_acidity                  1.00        0.66               -0.18
## citric_acid                    0.66        1.00               -0.08
## free_sulfur_dioxide           -0.18       -0.08                1.00
## total_sulfur_dioxide          -0.09        0.01                0.79
## density                        0.62        0.35               -0.04
## pH                            -0.71       -0.55                0.12
##                       total_sulfur_dioxide density    pH
## fixed_acidity                        -0.09    0.62 -0.71
## citric_acid                           0.01    0.35 -0.55
## free_sulfur_dioxide                   0.79   -0.04  0.12
## total_sulfur_dioxide                  1.00    0.13 -0.01
## density                               0.13    1.00 -0.31
## pH                                   -0.01   -0.31  1.00
```

En aquest cas tenim: a) El "PH" i "fixed_acidity" ho estan molt correlacionats (un augmenta en decréixer l'altre). b) "Density" i "fixed_acidity" ho estan també però en menor grau. c) "citric_acid"" i Fixed_acidity" ho estan també però en menor grau. d) "Free_sulfur_dioxid" i "total_sulfur_dioxid" ho estan també però en menor grau.

Els vins poden contenir diferents àcids tals com el tartàric, el màlic, el cítric i el succínic; per tant l'àcid cítric s'ha de considerar com a part dels àcids que pot contenir el vi.

El sulfur d'òxid pot estar lliure dintre del vi o afegit a altres substàncies químiques; és normal que la proporció de lliure estigui relacionada amb la quantitat no lliure; es tracta d'una dissolució. La proporció d'aquests

estats pot variar, per tant s'han de considerar.

Exportació de les dades netejades

Una vegada s'han agafat les dades amb el format correcte es procedeix a guardar-los en un fitxer amb nom "winequality-red-clean.csv":

```
write.csv(wine_data, "winequality-red-clean.csv")
```

Anàlisis de les dades

Selecció dels grups de dades que es volen analitzar

Ara es seleccionaran els grups de dades que poden ser interessants per analitzar o comparar; en aquest cas puc agrupar els vins per diferents tipus de qualitat. Tenim 6 nivells de qualitat de vins que van del 3 al 8; s'agafen 3 grups diferents agrupant-los en mitjans, bons i dolents tal com es mostra a la següent selecció:

```
# Agrupació per categoria
wine_data_bad  <- wine_data[ ((wine_data$quality == 3) | (wine_data$quality == 4)) ,]
wine_data_medium <- wine_data[ ((wine_data$quality == 5) | (wine_data$quality == 6)) ,]
wine_data_good <- wine_data[ ((wine_data$quality == 7) | (wine_data$quality == 8)) ,]
```

Comprobació de la normalidad i homogeneitat de la variància

Per comprovar que les variables quantitatives segueixen una distribució normal faré servir el test de Shapiro. Si a les proves s'obté un p-valor superior al nivell de significació prefixat de 0,05 llavors es considera que la variable segueix una distribució normal.

```
alpha = 0.05
col.names = colnames(wine_data)
for (i in 1:ncol(wine_data)) {
   if (i == 1) cat("Variables que no tenen distribució normal:\n")
   if (is.integer(wine_data[,i]) | is.numeric(wine_data[,i])) {
      p_val = shapiro.test(wine_data[,i])$p.value
      if (p_val < alpha) {
         cat(col.names[i])
         # Format output
         if (i < ncol(wine_data) - 1) cat(", ")
         if (i %% 3 == 0) cat("\n")
      }
   }
}
```

```
## Variables que no tenen distribució normal:
## fixed_acidity, volatile_acidity, citric_acid,
## residual_sugar, chlorides, free_sulfur_dioxide,
## total_sulfur_dioxide, density, pH,
## sulphates, alcohol
```

Per tant tenim que cap variable segueix una distribució normal.

Ara per mirar l'homogeneïtat de les variàncies dels diferents grups de vins existents faré servir el test de Fligner-Killeen. És un test no paramètric que compara les variàncies considerant les mitjanes.

En aquest cas faré l'estudi de l'homogeneïtat del pH dels grups de vins segons el tipus de qualitat; haig de fer totes les combinacions:

```
   a <- wine_data[wine_data$quality == 3, "sulphates"]
   b <- wine_data[wine_data$quality == 4, "sulphates"]
   c <- wine_data[wine_data$quality == 5, "sulphates"]
   d <- wine_data[wine_data$quality == 6, "sulphates"]
   e <- wine_data[wine_data$quality == 7, "sulphates"]
   f <- wine_data[wine_data$quality == 8, "sulphates"]

   fligner.test(x = list(a,b), data = wine_data)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  list(a, b)
## Fligner-Killeen:med chi-squared = 0.0096875, df = 1, p-value = 0.9216
```

```
   fligner.test(x = list(a,c), data = wine_data)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  list(a, c)
## Fligner-Killeen:med chi-squared = 0.59102, df = 1, p-value = 0.442
```

```
   fligner.test(x = list(a,d), data = wine_data)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  list(a, d)
## Fligner-Killeen:med chi-squared = 1.0665, df = 1, p-value = 0.3017
```

```
   fligner.test(x = list(a,e), data = wine_data)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  list(a, e)
## Fligner-Killeen:med chi-squared = 1.1042, df = 1, p-value = 0.2933
```

```
   fligner.test(x = list(a,f), data = wine_data)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  list(a, f)
## Fligner-Killeen:med chi-squared = 0.1242, df = 1, p-value = 0.7245
```

```
   fligner.test(x = list(b,c), data = wine_data)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  list(b, c)
## Fligner-Killeen:med chi-squared = 0.96113, df = 1, p-value = 0.3269
```

```
   fligner.test(x = list(b,d), data = wine_data)
```

```
##
```

```
##  Fligner-Killeen test of homogeneity of variances
##
## data:  list(b, d)
## Fligner-Killeen:med chi-squared = 2.8385, df = 1, p-value = 0.09203
```

```
   fligner.test(x = list(b,e), data = wine_data)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  list(b, e)
## Fligner-Killeen:med chi-squared = 3.2684, df = 1, p-value = 0.07063
```

```
   fligner.test(x = list(b,f), data = wine_data)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  list(b, f)
## Fligner-Killeen:med chi-squared = 0.0669, df = 1, p-value = 0.7959
```

```
   fligner.test(x = list(c,d), data = wine_data)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  list(c, d)
## Fligner-Killeen:med chi-squared = 5.13, df = 1, p-value = 0.02352
```

```
   fligner.test(x = list(c,e), data = wine_data)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  list(c, e)
## Fligner-Killeen:med chi-squared = 3.3481, df = 1, p-value = 0.06728
```

```
   fligner.test(x = list(c,f), data = wine_data)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  list(c, f)
## Fligner-Killeen:med chi-squared = 0.13482, df = 1, p-value = 0.7135
```

```
   fligner.test(x = list(d,e), data = wine_data)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  list(d, e)
## Fligner-Killeen:med chi-squared = 0.041597, df = 1, p-value = 0.8384
```

```
   fligner.test(x = list(d,f), data = wine_data)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  list(d, f)
```

```
## Fligner-Killeen:med chi-squared = 0.735, df = 1, p-value = 0.3913
```

```
   fligner.test(x = list(e,f), data = wine_data)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  list(e, f)
## Fligner-Killeen:med chi-squared = 1.1293, df = 1, p-value = 0.2879
```

Es veu clarament que els p-valors son superior a 0,05; per tant s'accepta l'hipòtesi que les variàncies de les mostres són homogènies. N'hi ha un cas, el grup "c" i "d", que tenen variàncies diferents; però aquest fet es pot considerar fortuït i en general es pot considerar que les variàncies són iguals en general.

Aplicació de proves estadístiques

Quines variables quantitatives influeixen a la qualitat del vi?. Per això es mira el coeficient de correlació de Spearman perquè tenim que les dades no segueixen una distribució normal.

```
wine_data$quality <- as.numeric(wine_data$quality)

corr_matrix <- matrix(nc = 2, nr = 0)
colnames(corr_matrix) <- c("estimate", "p-value")
# Calcular el coeficiente de correlación para cada variable cuantitativa
# con respecto al campo "precio"
for (i in 1:(ncol(wine_data) - 1)) {
  if (is.integer(wine_data[,i]) | is.numeric(wine_data[,i])) {
    spearman_test = cor.test(wine_data[,i], wine_data[,length(wine_data)],
    method = "spearman")
    corr_coef = spearman_test$estimate
    p_val = spearman_test$p.value
    # Add row to matrix
    pair = matrix(ncol = 2, nrow = 1)
    pair[1][1] = corr_coef
    pair[2][1] = p_val
    corr_matrix <- rbind(corr_matrix, pair)
    rownames(corr_matrix)[nrow(corr_matrix)] <- colnames(wine_data)[i]
   }
}
wine_data$quality <- as.factor(wine_data$quality)
print(corr_matrix)
```

```
##                       estimate      p-value
## fixed_acidity         0.11408367 4.801220e-06
## volatile_acidity     -0.38064651 2.734944e-56
## citric_acid           0.21348091 6.158952e-18
## residual_sugar        0.03204817 2.002454e-01
## chlorides            -0.18992234 1.882858e-14
## free_sulfur_dioxide  -0.05690065 2.288322e-02
## total_sulfur_dioxide -0.19673508 2.046488e-15
## density              -0.17707407 9.918139e-13
## pH                   -0.04367193 8.084594e-02
## sulphates             0.37706020 3.477695e-55
## alcohol               0.47853169 2.726838e-92
```

Les variables que més influeixen en la qualitat del vi en ordre d'importància son: "alcohol", "volatile_acidity" i "sulphates".

El valor mitg dels "sulfats" dels bons vins es mes gran que la resta de vins?

Aquesta prova consisteix a comparar el valor mig dels sulfats en vins de qualitat baixa, mitja i alta.

Per tractar aquesta situació on tenim que les variables no són normals treballarem amb els valors mitjans; segons el teorema central del límit la mitja dels valors de la mostra es comporten com una distribució normal per mostres superiors a 30. Es planteja el següent contrast d'hipòtesis de dues mostres sobre la diferència de les mitjanes; aquest contrast és unilateral.

H0 : $\ 1 - \ 2 = 0$ H1 : $\ 1 - \ 2 > 0$

On $\ 1$ és la mitjana de sulfats de vins bons i $\ 2$ és la mitjana de sulfats de vins moderats; prenem alfa = 0, 05.

```
t.test(wine_data_good$sulphates, wine_data_medium$sulphates,alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  wine_data_good$sulphates and wine_data_medium$sulphates
## t = 9.4315, df = 337.33, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.07937086        Inf
## sample estimates:
## mean of x mean of y
## 0.7434562 0.6472631
```

Veiem que obtenim un p-value $<$ 2.2e-16 $<$ 0.05; per tant rebutgem l'hipòtesi nul · la. Els sulfats són més altes en vins bons.

Si agafo $\ 1$ com la mitjana de sulfats de vins moderats i $\ 2$ és la mitjana de sulfats de vins dolents amb alfa = 0, 05 veig el següent:

```
t.test(wine_data_medium$sulphates, wine_data_bad$sulphates,alternative = "greater")
```

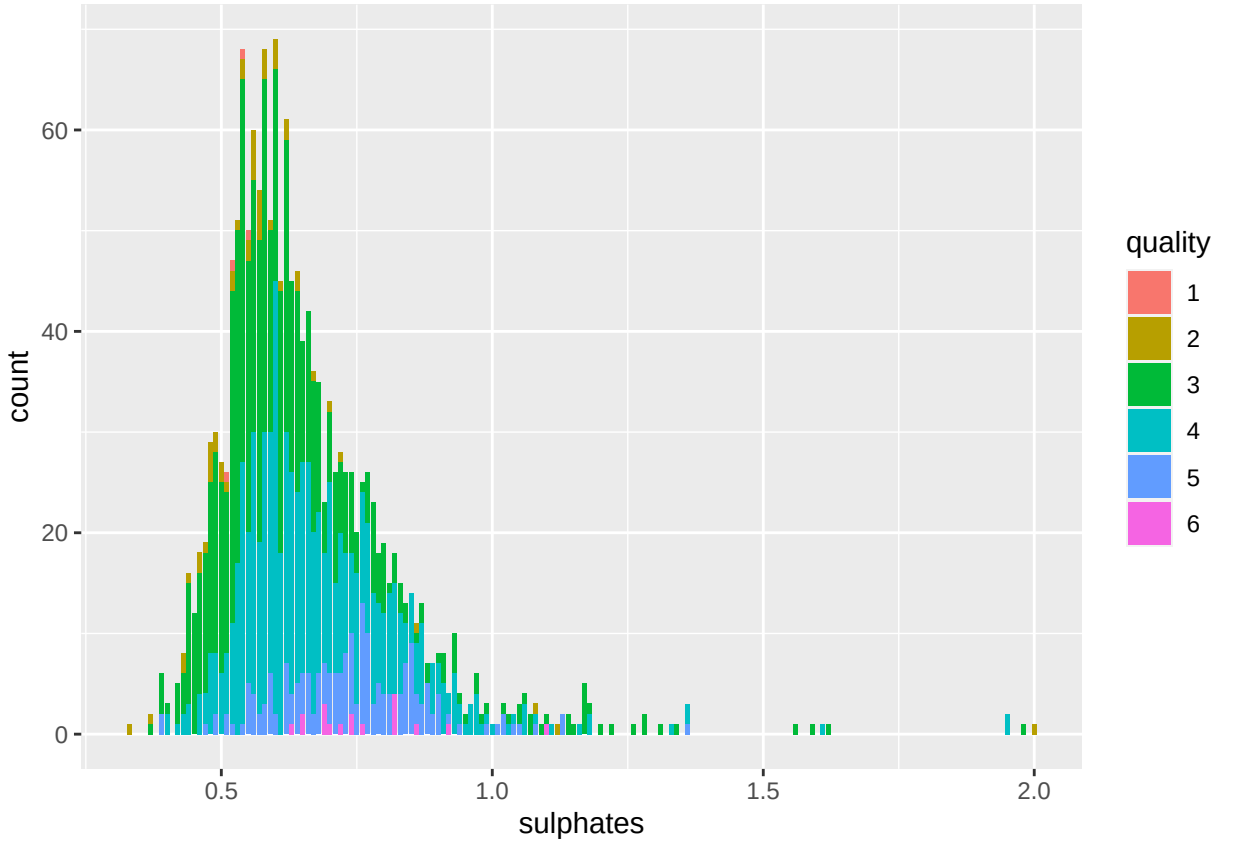
```
##
##  Welch Two Sample t-test
##
## data:  wine_data_medium$sulphates and wine_data_bad$sulphates
## t = 1.9221, df = 65.337, p-value = 0.02947
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.007262794         Inf
## sample estimates:
## mean of x mean of y
## 0.6472631 0.5922222
```

Veig un altre cop que es rebutja l'hipòtesi nul · la (p-value = 0.02947 $<$ 0.05); és a dir els sulfats milloren la qualitat del vi.

Representació dels resultats a partir de taules i gràfiques.

El resultats de la correlació de les variables en format de taula ja s'han comentat a l'apartat anterior. Per altre banda per veure d'una manera visual el que he descobert sobre els sulfats a la qualitat del vins faig una representació gràfica de variació la distribució de la variable "sulphat" en funció de la qualitat del vi :

```
ggplot(data = wine_data[1:filas,],aes(x=sulphates,fill=quality))+geom_bar()
```



Visualment es pot veure com la mitja dels valors dels sulfats dels vins de millor qualitat son superior a les dels vins de qualitat inferior; la distribució de sulfats depen de la qualitat del vi de manera clara.

Model de xarxes neuronals

Les xarxes neuronals es caracteritzen per donar bons resultats en models que poden no tindre linealitat com és aquest cas. Per aquest model es faran servir totes les variables quantitatives. Per trobar el model que dóna més eficiència s'agafaran les variables que més estiguin correlacionades.

Ara vaig a crear un grup de mostres agafades aleatòriament sense repetició per fer l'entrenament i el test; també escalo les variables abans d'aplicar l'algorisme doncs diferents escales de variables poden afectar de manera desigual el pes dels paràmetres de l'algorisme:

```
# Spliting training and testing dataset

index = sample( 1:nrow( wine_data ), nrow( wine_data ) * 0.6, replace = FALSE )

#Train
train = wine_data[ index, ]
trainset = subset( train, select = -quality )
trainset.scaled <- scale(trainset)


# Test
test = wine_data[ -index, ]
testset = subset( test, select = -quality )
testset.scaled <- scale(testset);
```

Ara faig la preparació pel model de xarxes neuronals dels valors "quality" que es volen predir; aquest han de separar-se en columnes i agafar valors "TRUE" i "FALSE":

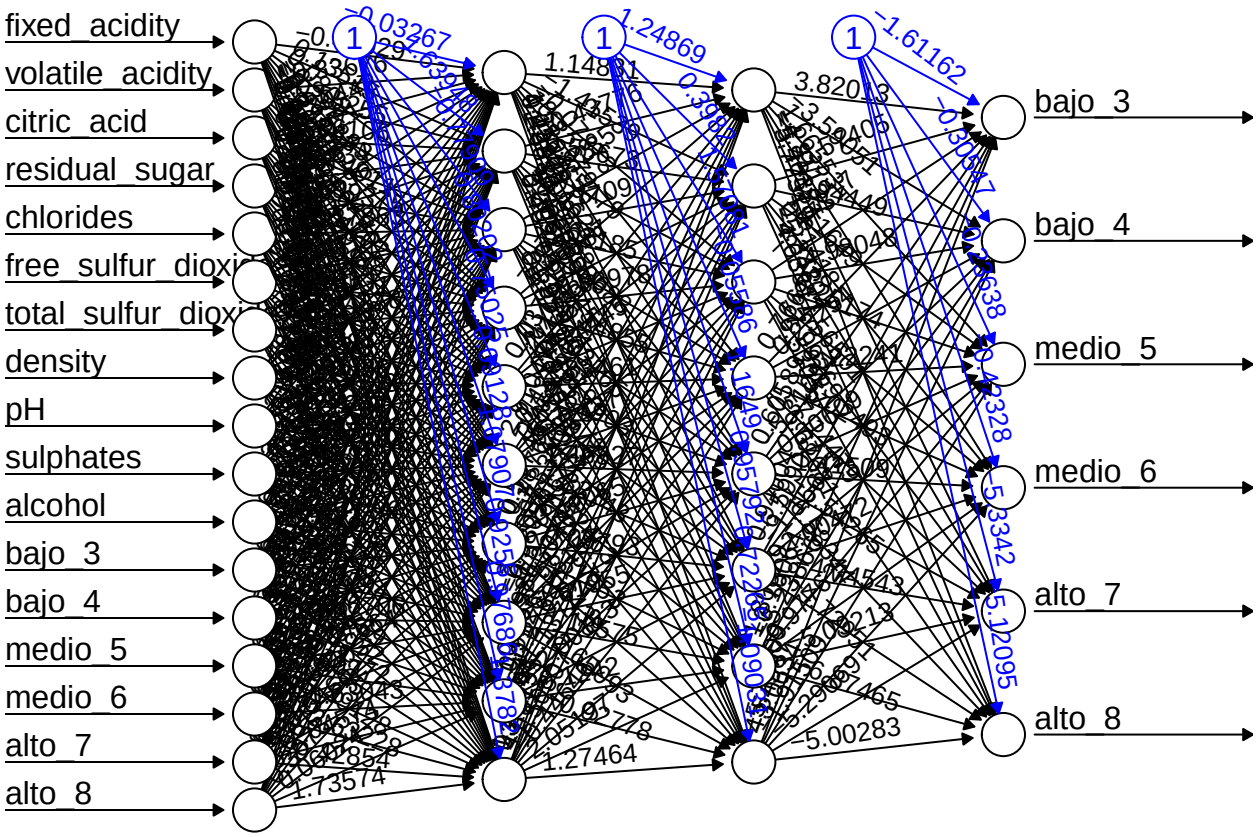
```
trainset.final <- data.frame(trainset.scaled,quality=train$quality)
trainset.final$quality <- as.factor(trainset.final$quality)

testset.final <- data.frame(testset.scaled,quality=test$quality)
testset.final$quality <- as.factor(testset.final$quality)

# Preparem les dades per l'algotime de xarxes neuronals
trainset.final <- cbind(trainset.final, trainset.final$quality == 3)
trainset.final <- cbind(trainset.final, trainset.final$quality == 4)
trainset.final <- cbind(trainset.final, trainset.final$quality == 5)
trainset.final <- cbind(trainset.final, trainset.final$quality == 6)
trainset.final <- cbind(trainset.final, trainset.final$quality == 7)
trainset.final <- cbind(trainset.final, trainset.final$quality == 8)
names(trainset.final)[13:18] <- c('bajo_3', 'bajo_4','medio_5',
                                  'medio_6', 'alto_7','alto_8')
trainset.final = subset( trainset.final, select = -quality )

testset.final <- cbind(testset.final, testset.final$quality == 3)
testset.final <- cbind(testset.final, testset.final$quality == 4)
testset.final <- cbind(testset.final, testset.final$quality == 5)
testset.final <- cbind(testset.final, testset.final$quality == 6)
testset.final <- cbind(testset.final, testset.final$quality == 7)
testset.final <- cbind(testset.final, testset.final$quality == 8)
names(testset.final)[13:18] <- c('bajo_3', 'bajo_4','medio_5',
                                 'medio_6','alto_7','alto_8')
testset.final = subset( testset.final, select = -quality )
```

Es crea el model de xarxes neuronals:

```
# Build the neural network (NN)

nam = colnames( trainset.final )
fn = as.formula( paste( "bajo_3+bajo_4+medio_5+medio_6+alto_7+alto_8~",
                        paste( nam[!nam %in% "quality" ], collapse = "+" ) ) )
nn = neuralnet( fn, trainset.final, hidden = c(10,8), linear.output = FALSE)
```

Seguidament es procedeix a la representació gràfica de la xarxa neuronal generada:

```
# Plot the NN
plot( nn, rep = "best" )
```



El següent pas es testejar el resultat de l'algorisme; primer genero les dades que prediu el model pel joc de dades de test:

```
# Test the resulting output
nn.results = predict(nn, testset.final,type="class" )
```

El resultat de la predicció feta amb el model de xarxa neuronal dóna sis paràmetres que indiquen la probabilitat que cadascun dels diferents nivells de qualitat sigui factible. Per mirar els casos que s'encerten considerem que els valors que tenen una probabilitat més gran de 0,5 donen lloc a una resposta positiva; per tant es correspon al nivell que estem tractant.

Faig diversos "dataframes" que contenen els valors originals del test i els valors predits per cadascuna de les categories de vins existents:

```
# Function to always round 0.5 down
round2 <- function(x) {
  ret <- round(x)
  if (x==0.5) { ret <- 0 }
  else {ret <- round(x)}
  return(ret)
}


dfResult <- as.data.frame(nn.results)
results3 <- data.frame(actual = testset.final$bajo_3*1 ,
                       prediction = round2(dfResult$V1))
results4 <- data.frame(actual = testset.final$bajo_4*1,
                       prediction = round2(dfResult$V2))
```

```
results5 <- data.frame(actual = testset.final$medio_5*1,
                       prediction = round2(dfResult$V3))
results6 <- data.frame(actual = testset.final$medio_6*1,
                       prediction = round2(dfResult$V4))
results7 <- data.frame(actual = testset.final$medio_6*1,
                       prediction = round2(dfResult$V5))
results8 <- data.frame(actual = testset.final$medio_6*1,
                       prediction = round2(dfResult$V6))
```

Un exemple sería el següent:

```
head(results3)
```

```
##   actual prediction
## 1      1          1
## 2      1          1
## 3      1          1
## 4      1          1
## 5      1          1
## 6      1          1
```

Ara miro la precisió del model calculat comparant aquells casos en què s'ha encertat respecte a els casos totals:

```
total_ok <-  sum(results3$actual==results3$prediction) +
             sum(results4$actual==results4$prediction) +
             sum(results5$actual==results5$prediction) +
             sum(results6$actual==results6$prediction) +
             sum(results7$actual==results7$prediction) +
             sum(results8$actual==results8$prediction);
accuracy <- 100 * total_ok/(nrow(results3)+nrow(results4)+nrow(results5)+
                      nrow(results6)+nrow(results7)+nrow(results8))
sprintf("La precisió de la xarxa neuronal es: %s",accuracy)
```

```
## [1] "La precisió de la xarxa neuronal es: 99.6354166666667"
```

Tenim doncs una precisió molt elevada.

Conclusions

S'ha vist que s'han realitzat tres tipus de proves estadístiques sobre el conjunt de dades que es corresponen amb variables relatives a les qualitats del vi. L'anàlisi de correlació i el contrast d'hipòtesi han permès veure quines d'aquestes variables són més importants a la qualitat del vi i com intervenen els sulfits a la qualitat del vi. Per altra banda el model de xarxa neuronal obtingut és de molta utilitat a l'hora de fer prediccions, doncs té una precisió per sobre del 99,5%.

Contribucions

Han contribuit a aquesta pràctica:

```
contributions =data.frame(stringsAsFactors=FALSE,
                          Contribuciones = c("Investigació prèvia", "Redacció de les respostes",
                          Firma = c("Sergio García","Sergio García","Sergio García"))

contributions %>% kable() %>% kable_styling()
```

Contribuciones	Firma
Investigació prèvia	Sergio García
Redacció de les respostes	Sergio García
Desenvolupament codi	Sergio García