Prac2 - Wine quality

Ferran Pintó Haro i Arnau Santos Ribelles

9/5/2022

R Markdown

0. Lectura del fitxer i preparació de les dades

Llegeix el fitxer CensusIncome_clean.csv i guarda les dades en un objecte amb identificador denominat cens. Verifica que les dades s'han carregat correctament.

Carreguem els fitxers de dades i els guardem en dos objectes denominats *red_wine_df* i *white_wine_df*.

```
red_wine_df <- read.csv('winequality-red.csv', sep= ",")
white_wine_df <- read.csv('winequality-white.csv', sep= ";")
```

Examinem els valors resum de cada tipus de variable.

```
summary(red_wine_df)
```

```
##  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
```

```
str(red_wine_df)
```

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

```
summary(white_wine_df)
```

```
##  fixed.acidity    volatile.acidity  citric.acid     residual.sugar
##  Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600
##  1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700   1st Qu.: 1.700
##  Median : 6.800   Median :0.2600   Median :0.3200   Median : 5.200
##  Mean   : 6.855   Mean   :0.2782   Mean   :0.3342   Mean   : 6.391
##  3rd Qu.: 7.300   3rd Qu.:0.3200   3rd Qu.:0.3900   3rd Qu.: 9.900
##  Max.   :14.200   Max.   :1.1000   Max.   :1.6600   Max.   :65.800
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide    density
##  Min.   :0.00900   Min.   :  2.00      Min.   :  9.0        Min.   :0.9871
##  1st Qu.:0.03600   1st Qu.: 23.00      1st Qu.:108.0        1st Qu.:0.9917
##  Median :0.04300   Median : 34.00      Median :134.0        Median :0.9937
##  Mean   :0.04577   Mean   : 35.31      Mean   :138.4        Mean   :0.9940
##  3rd Qu.:0.05000   3rd Qu.: 46.00      3rd Qu.:167.0        3rd Qu.:0.9961
##  Max.   :0.34600   Max.   :289.00      Max.   :440.0        Max.   :1.0390
##        pH          sulphates         alcohol         quality
##  Min.   :2.720   Min.   :0.2200   Min.   : 8.00   Min.   :3.000
##  1st Qu.:3.090   1st Qu.:0.4100   1st Qu.: 9.50   1st Qu.:5.000
##  Median :3.180   Median :0.4700   Median :10.40   Median :6.000
##  Mean   :3.188   Mean   :0.4898   Mean   :10.51   Mean   :5.878
##  3rd Qu.:3.280   3rd Qu.:0.5500   3rd Qu.:11.40   3rd Qu.:6.000
##  Max.   :3.820   Max.   :1.0800   Max.   :14.20   Max.   :9.000
```

```
str(white_wine_df)
```

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

Veiem com tenim les mateixes variables en els dos datasets que sumen 12 variables. En el de vi negre tenim 1599 observacions i en el de vi blanc 4898, sent totes les variables numèriques i la de qualitat sent de tipus integer (sense decimals). Podem veure també els valors entre els que es troba cada variable.

1. Descripció del dataset

Perquè és important i quina pregunta/problema pretén respondre?

El dataset conté informació relacionada amb les variants negre i blanca de la varietat de vi portuguès "Vinho Verde". Segons els autors del conjunt de dades, per motius de privacitat i logística només hi ha variables fisicoquímiques (entrades) i sensorials (sortida), i no hi ha, per exemple, dades sobre tipus de raïm, marca de vi, preu de venta, etc.

Comptem amb dos datasets (un per la variant blanca i un per la negre) que ajuntem posteriorment. Les 14 variables que conté el dataset final són:

"type": valor categòric corresponent al tipus de varietat (white / red).

"fixed acidity": valor numèric corresponent al la quantitat d'àcids implicats (que no s'evaporen fàcilment).

"volatile acidity": valor numèric corresponent a la quantitat d'àcid acètic en el vi.

"citric acid": valor numèric corresponent a la quantitat d'àcid cítric. Entre 0 i 1.

"residual sugar": valor numèric corresponent a la quantitat (grams/litre) de sucre que queda després de la parada de la fermentació.

"chlorides": valor numèric corresponent a la quantitat de sal al vi.

"free sulfur dioxide": valor numèric corresponent a la quantitat de la forma lliure del SO2.

"total sulfur dioxide": valor numèric corresponent a la quantitat de formes lliures i lligades de SO2.

"density": valor numèric corresponent a la densitat del vi.

"pH": valor numèric corresponent al PH del vi. Escala de 0 (molt àcid) al 14 (molt bàsic).

"sulphates": valor numèric corresponent a la quantitat de sulfats.

"alcohol": valor numèric corresponent als graus d'alcohol que té el vi (percentatge d'alcohol).

"quality": qualitat percebuda del vi. Valor numèric entre 0 i 10.

"quality_d": variable categòrica creada segons la qualitat del vi ("bo" si la qualitat és 7 o superior i "no bo" si és inferior a 7)

És un dataset rellevant per poder fer tasques de classificació i regressió, al tenir diferents variables numèriques corresponents a les propietats del vi i una variable corresponent a la qualitat percebuda. Així podrem observar quines característiques del vi estan més relacionades amb la qualitat.

És pretén respondre la pregunta següent: **Quines són les propietats fisicoquímiques que fan que un vi sigui bo?**

Alhora que observar: **Les propietats que fan que un vi sigui bo són les mateixes per vins negres i blancs?**

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

Pot ser el resultat d'addicionar diferents datasets o una subselecció útil de les dades originals, en base a l'objectiu que es vulgui aconseguir.

Integrarem els dos datasets per tenir un únic dataset corresponent a vins. Abans de fer-ho, com que ens interessarà saber la varietat de vi que correspon a cada registre, creem una columna anomenada "type" en els dos datasets que indiqui el tipus de vi que és: "White" si és vi blanc i "Red" si és negre.

```
red_wine_df["type"] <- "Red"
white_wine_df["type"] <- "White"
```

Ara integrem els dos datasets en un de sol, anomenat "wine". Ho fem mitjançant una fusió vertical, per incloure nous registres a un dataset. Per fer-ho és important que el format de les bases de dades a integrar sigui el mateix, com hem comprovat.

```
wine <- rbind(red_wine_df,white_wine_df)
```

Comprovem que té el número de files i columnes que hauria de tenir i veiem els 3 primers i últims registres.

```
dim(wine)
```

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

```
head(wine,3)
```

```
##   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1           7.4             0.70        0.00            1.9     0.076
## 2           7.8             0.88        0.00            2.6     0.098
## 3           7.8             0.76        0.04            2.3     0.092
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  11                   34  0.9978 3.51      0.56     9.4
## 2                  25                   67  0.9968 3.20      0.68     9.8
## 3                  15                   54  0.9970 3.26      0.65     9.8
##   quality type
## 1       5  Red
## 2       5  Red
## 3       5  Red
```

```
tail(wine,3)
```

```
##      fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 6495           6.5             0.24        0.19            1.2     0.041
## 6496           5.5             0.29        0.30            1.1     0.022
## 6497           6.0             0.21        0.38            0.8     0.020
##      free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 6495                  30                  111 0.99254 2.99      0.46     9.4
## 6496                  20                  110 0.98869 3.34      0.38    12.8
## 6497                  22                   98 0.98941 3.26      0.32    11.8
##      quality  type
## 6495       6 White
## 6496       7 White
## 6497       6 White
```

Ara tenim en un dataset les 6497 files corresponents als vins blanc i negre, amb les mateixes 13 files. Com esperavem, les tres primeres files corresponen a registres de vi negre (Red) i les tres últimes a vi blanc (White). No eliminarem files ja que les farem servir totes per observar la seva relació amb la qualitat.

3. Neteja de les dades.

3.1. Zeros o elements buits

Les dades contenen zeros o elements buits? Gestiona cadascun d'aquests casos.

Comprovem si el dataset té valors absents (NA).

```
colSums(is.na(wine))
```

```
##        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
##                 type
##                    0
```

Veiem com no tenim cap element buit en el dataset, cap valor nul en cap columna.

Observem ara els zeros.

```
colSums(wine==0)
```

```
##        fixed.acidity     volatile.acidity          citric.acid
##                    0                    0                  151
##       residual.sugar            chlorides  free.sulfur.dioxide
##                    0                    0                    0
## total.sulfur.dioxide              density                   pH
##                    0                    0                    0
##            sulphates              alcohol              quality
##                    0                    0                    0
##                 type
##                    0
```

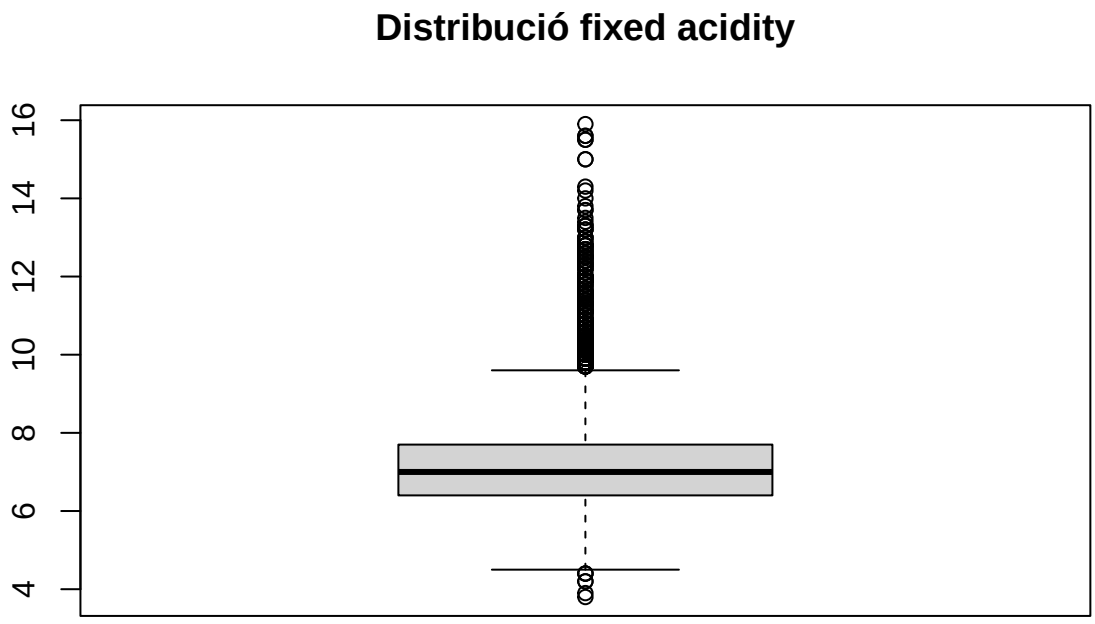
Veiem que la columna "citric.acid" té 151 zeros, però són vàlids ja que està al rang de valors possibles de la variable. Pot no haver-hi àcid cítric als vins.

3.2. Valors extrems.

Identifica i gestiona els valors extrems.

Observem primer la distribució de la "fixed acidity".

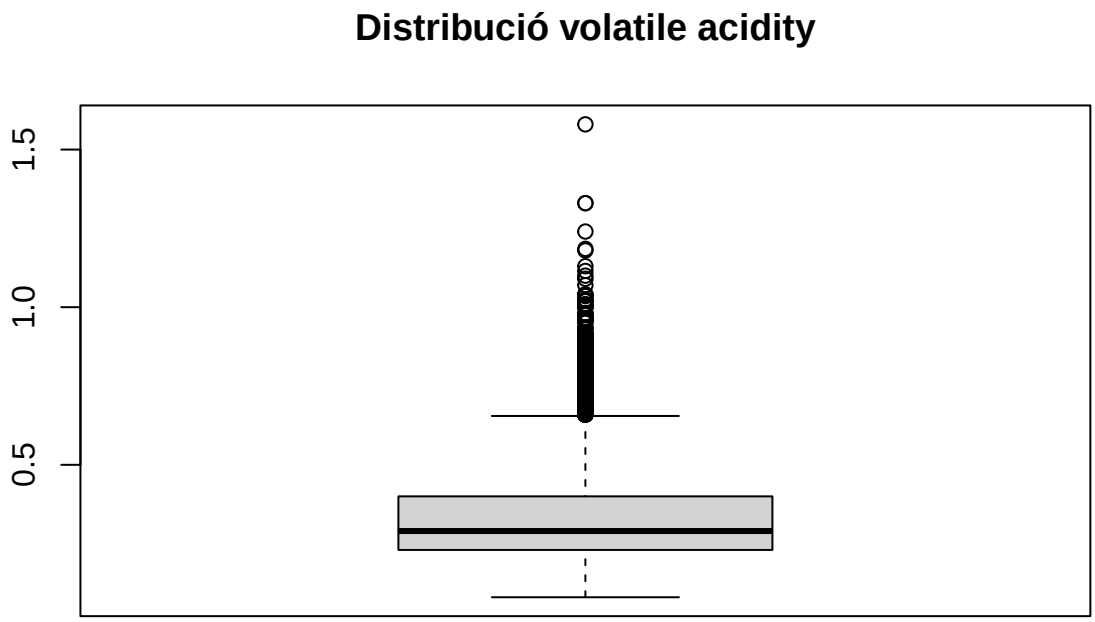
```
boxplot(wine$fixed.acidity, main="Distribució fixed acidity")
```



Malgrat hi ha molts valors atípics, no es consideren anòmals i per tant no els tractem.

Observem la distribució de la "volatile acidity".

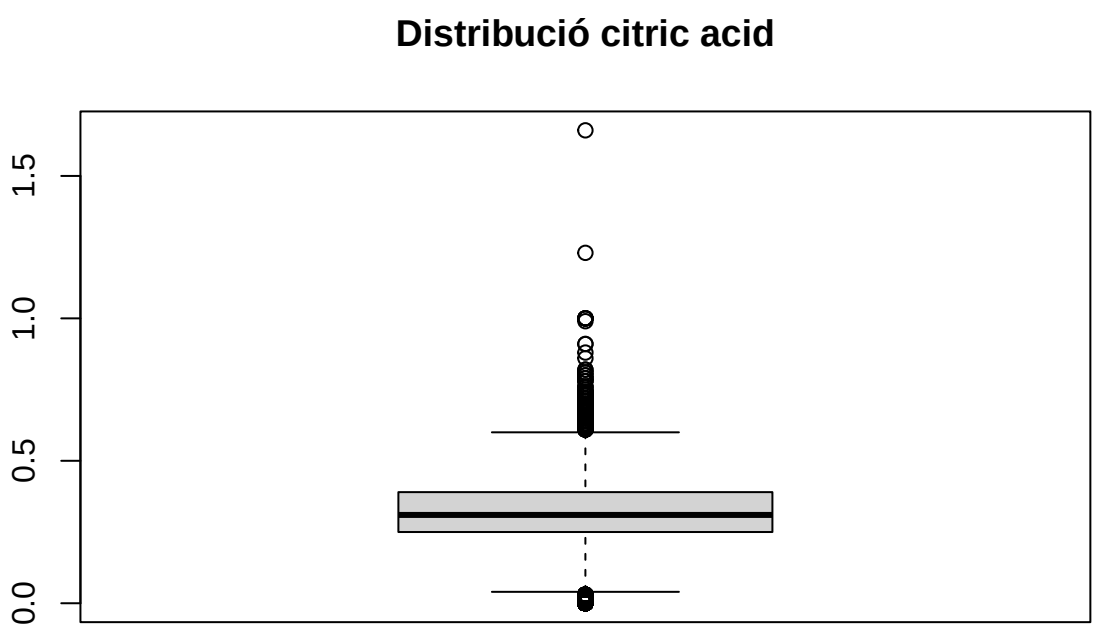
```
boxplot(wine$volatile.acidity, main="Distribució volatile acidity")
```



Malgrat hi ha molts valors atípics, no es consideren anòmals i per tant no els tractem.

Observem la distribució de l'àcid cítric.

```
boxplot(wine$citric.acid, main="Distribució citric acid")
```



En una cerca, s'ha trobat que la quantitat legal màxima d'àcid cítric en el vi és de 1 gram per litre, per tant, considerem anòmals els valors atípics a partir d'aquest valor.

Contem el número de files amb més d'1 gram d'àcid cítric per litre.

```
nrow(wine[(wine$citric.acid > 1),])
```

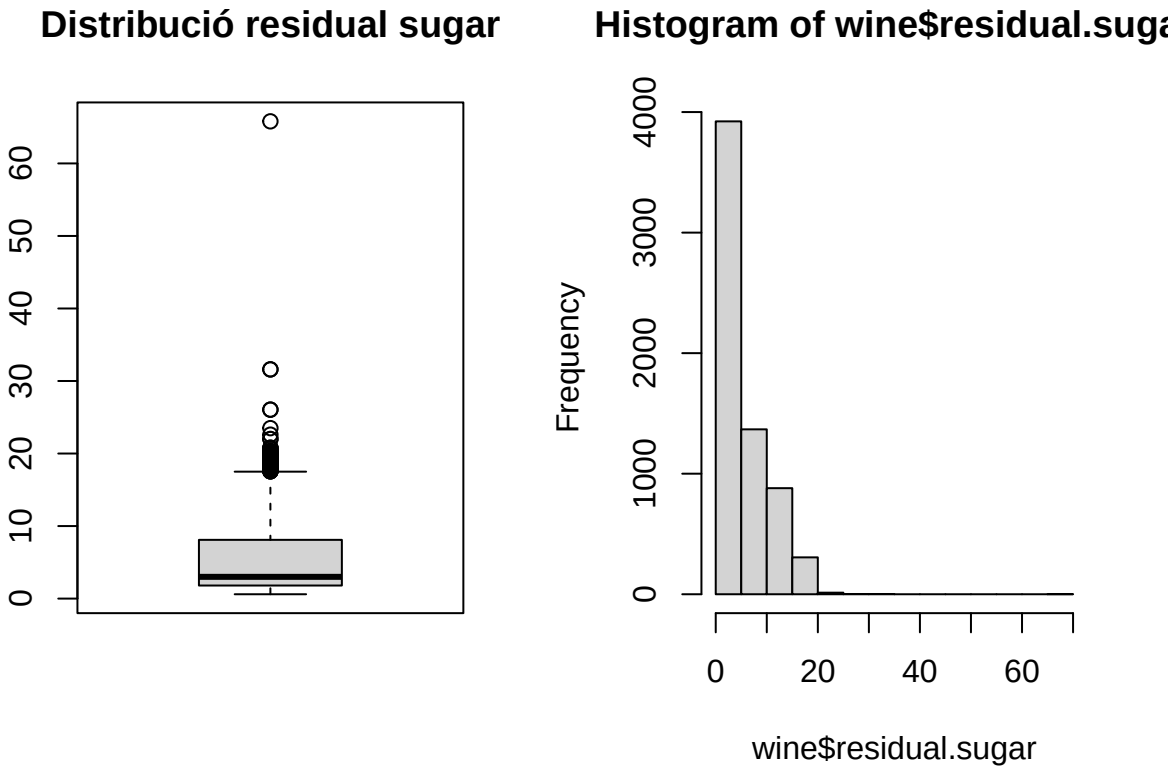
```
## [1] 2
```

Veiem que només tenim dues files i és un percentatge ínfim del total i decidim eliminar les files.

```
wine <- wine[!(wine$citric.acid > 1),]
```

Observem ara la distribució del sucre residual.

```
par(mfrow=c(1,2))
boxplot(wine$residual.sugar, main="Distribució residual sugar")
hist(wine$residual.sugar)
```



Considerem anòmals els valors per sobre de 30.

Contem el número de files amb valors per sobre dels 30 en sucre residual.

```
nrow(wine[(wine$residual.sugar > 30),])
```

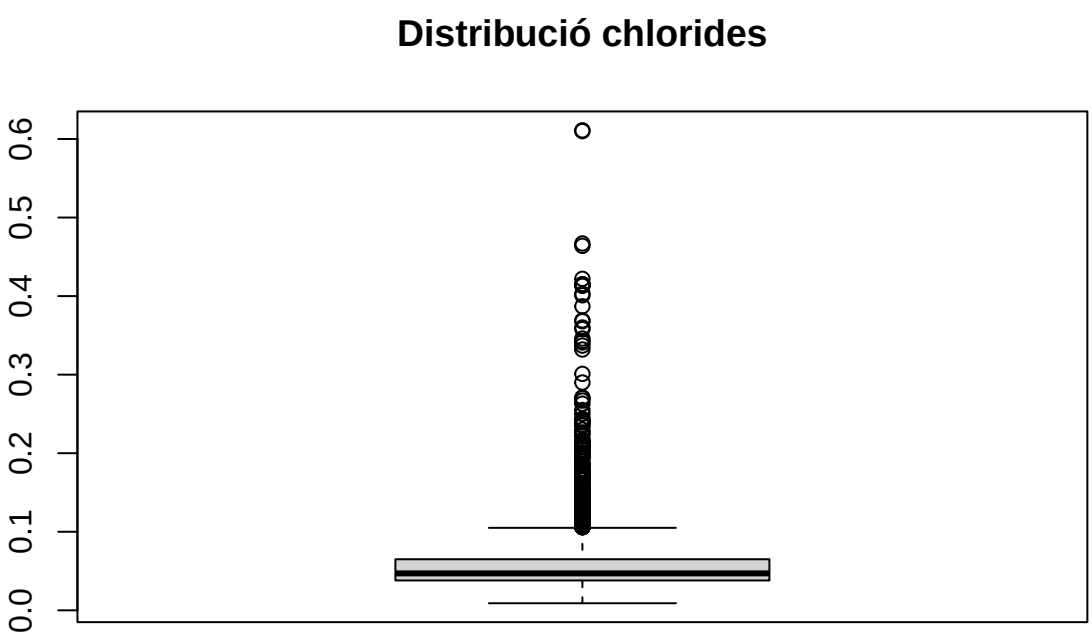
```
## [1] 3
```

Veiem que només tenim tres files i és un percentatge ínfim del total i decidim eliminar les files.

```
wine <- wine[!(wine$residual.sugar > 30),]
```

Observem ara la distribució de chlorides (quantitat de sal).

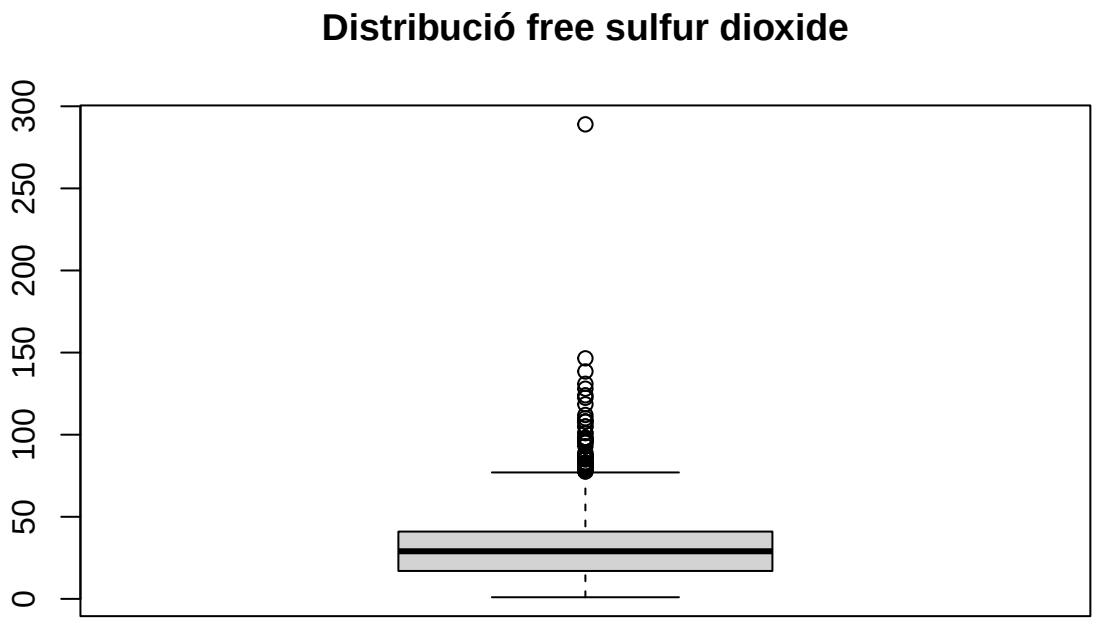
```
boxplot(wine$chlorides, main="Distribució chlorides")
```



Malgrat hi ha molts valors atípics, no es consideren anòmals i per tant no els tractem.

Observem ara la distribució del diòxid de sofre lliure.

```
boxplot(wine$free.sulfur.dioxide, main="Distribució free sulfur dioxide")
```



Observem, entre d'altres, un valor atípic molt llunyà a la resta, que considerem anòmal. Considerem anòmals els valors superiors a 150, deixant la resta igual.

Veiem el número de files amb aquests valors.

```
nrow(wine[(wine$free.sulfur.dioxide > 150),])
```

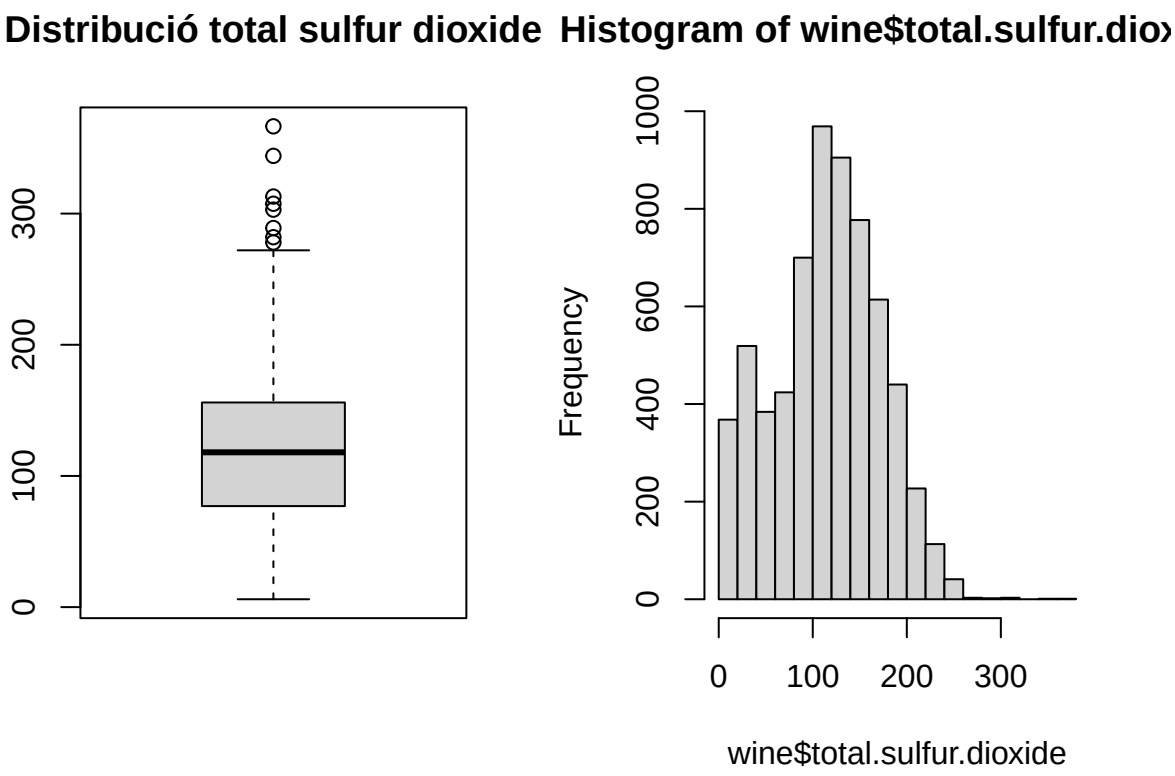
```
## [1] 1
```

Veiem que només tenim una fila i decidim eliminar-la.

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

Observem ara la distribució del "total sulfur dioxide".

```
par(mfrow=c(1,2))
boxplot(wine$total.sulfur.dioxide, main="Distribució total sulfur dioxide")
hist(wine$total.sulfur.dioxide)
```



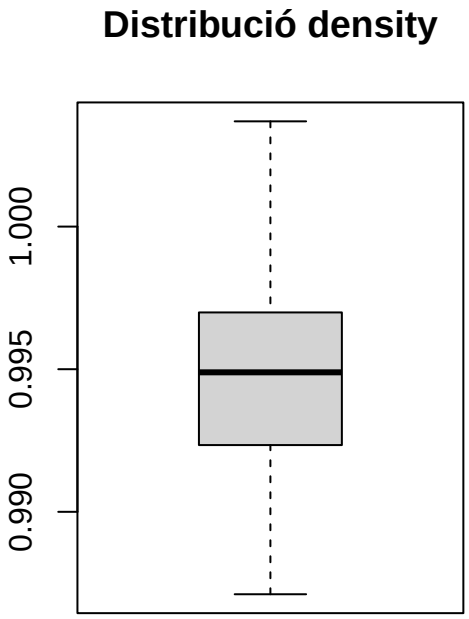

```
boxplot.stats(wine$total.sulfur.dioxide)$out
```

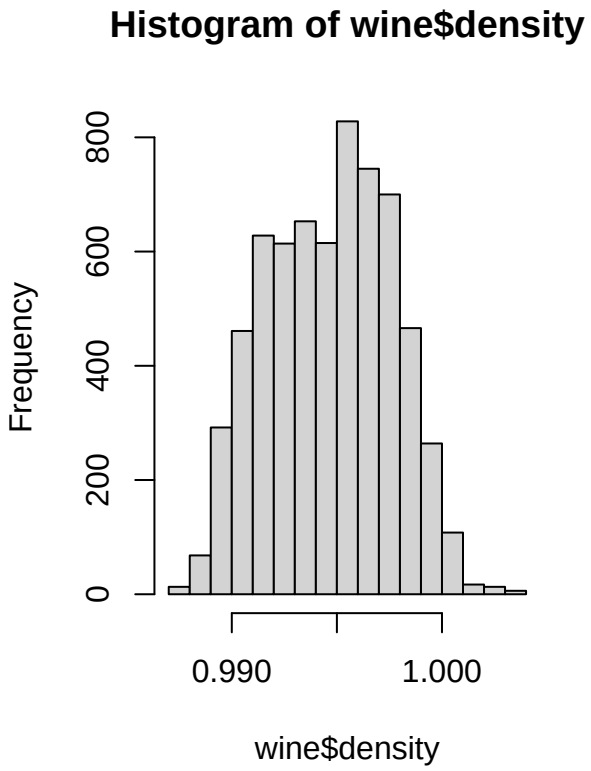
```
## [1] 278.0 289.0 313.0 366.5 307.5 344.0 282.0 303.0
```

Veiem com tenim varis valors extrems, però no els considerem anòmals.

Observem ara la distribució de la densitat.

```
par(mfrow=c(1,2))
boxplot(wine$density, main="Distribució density")
hist(wine$density)
```





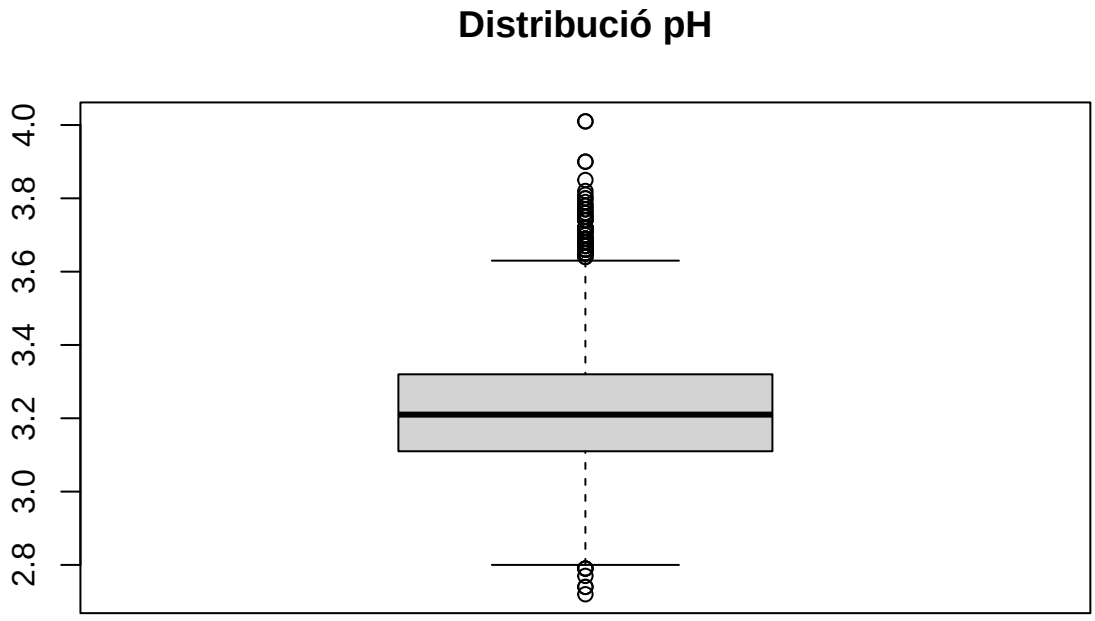

```
boxplot.stats(wine$density)$out
```

```
## numeric(0)
```

No tenim cap outlier en la variable density.

Observem ara la distribució del pH.

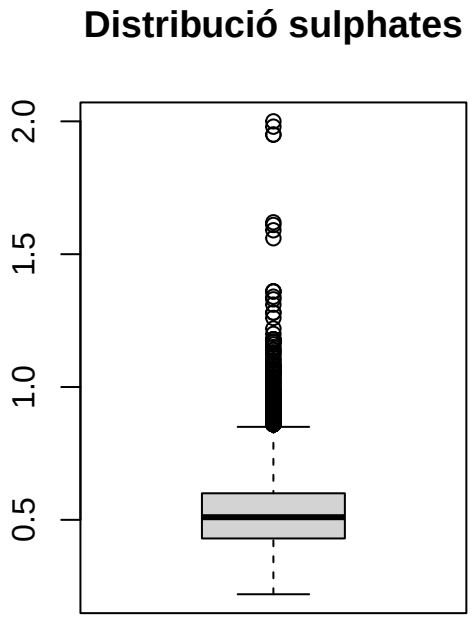
```
boxplot(wine$pH, main="Distribució pH")
```

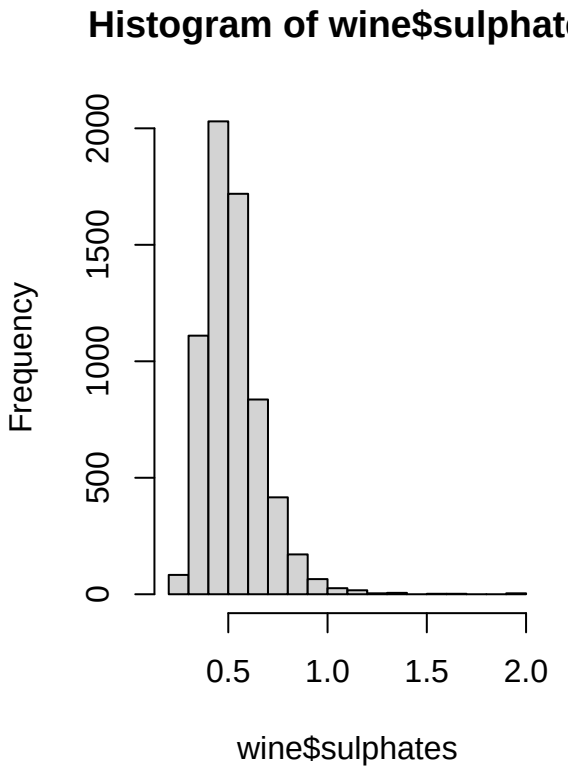


Observem valors atípics en la distribució del pH, però són valors lògics dins l'escala del pH, per tant els donem per vàlids.

Mirem ara la distribució dels sulfats.

```
par(mfrow=c(1,2))
boxplot(wine$sulphates, main="Distribució sulphates")
hist(wine$sulphates)
```





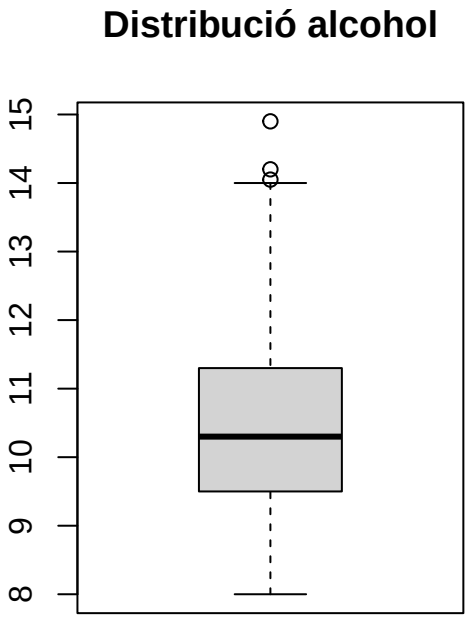

```
boxplot.stats(wine$sulphates)$out
```

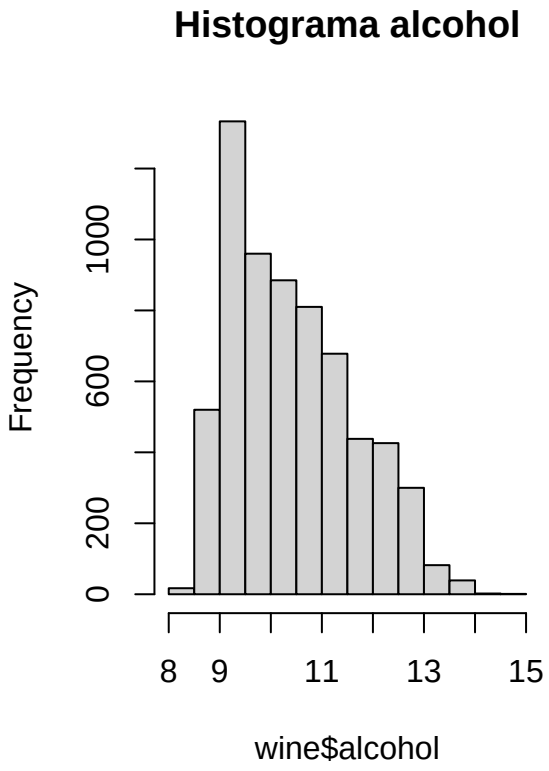
```
##   [1] 1.56 0.88 0.93 1.28 1.08 0.91 0.91 0.90 1.20 0.95 1.12 1.28 1.14 1.95 1.22
##  [16] 1.95 1.98 1.31 0.93 0.93 0.92 2.00 1.08 1.59 1.02 0.97 1.03 0.88 0.86 1.61
##  [31] 1.09 0.96 0.96 1.26 0.87 0.86 0.91 0.97 0.97 0.91 0.97 1.08 0.86 0.95 0.86
##  [46] 1.00 1.36 1.18 0.87 0.89 0.93 0.92 0.86 0.98 0.88 0.91 0.87 0.93 1.13 0.87
##  [61] 1.04 1.11 1.13 0.99 1.07 0.90 0.90 0.89 0.89 1.06 0.91 0.89 1.06 0.92 1.05
##  [76] 1.06 0.92 0.90 1.04 1.05 1.02 1.14 0.90 0.99 0.87 0.87 0.86 0.91 1.02 1.36
##  [91] 0.93 0.96 1.36 1.05 1.17 1.62 1.06 0.92 0.91 1.18 0.94 0.86 0.86 0.86 1.07
## [106] 0.89 0.89 0.87 0.90 0.99 0.86 0.87 0.87 1.34 0.89 0.86 0.86 0.88 0.87 0.87
## [121] 1.16 1.10 0.98 0.88 0.86 0.94 0.87 1.15 0.87 1.17 1.17 1.33 1.18 1.17 1.03
## [136] 1.17 1.10 0.90 0.94 0.93 1.01 0.93 0.94 0.90 0.93 0.88 0.88 0.97 0.97 0.93
## [151] 0.96 0.97 0.95 0.95 0.95 0.90 0.88 0.88 0.87 0.86 0.90 0.90 0.92 0.98 1.06
## [166] 0.88 0.88 0.88 1.00 0.90 0.90 0.89 0.94 0.99 0.86 0.95 0.87 0.88 0.88 0.98
## [181] 0.98 0.98 0.98 0.98 0.96 1.01 0.96 0.92 0.94 0.95 1.08
```

Malgrat hi ha molts valors atípics, no es consideren anòmals i per tant no els tractem. Pel que s'ha vist cercant a Internet, és possible tenir fins a 2 grams per litre de sulfats.

Observem ara la distribució del percentatge d'alcohol.

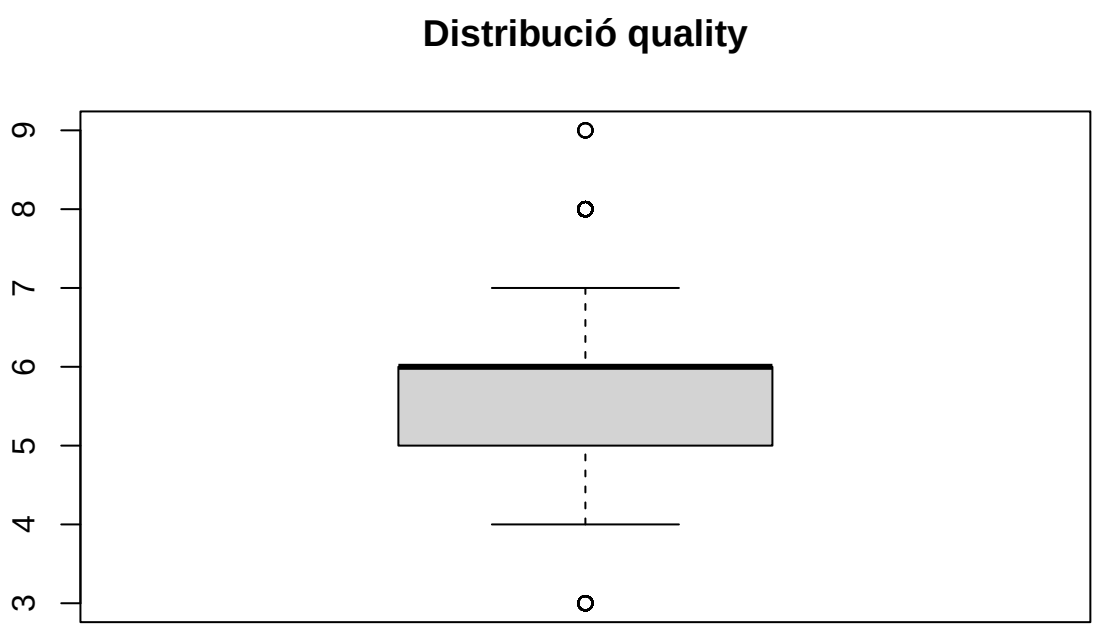
```
par(mfrow=c(1,2))
boxplot(wine$alcohol, main="Distribució alcohol")
hist(wine$alcohol, main="Histograma alcohol")
```





Malgrat hi ha valors atípics, no es consideren anòmals ja que pot haver-hi una graduació de 15% d'alcohol en un vi. Per tant no els tractem.

```
boxplot(wine$quality, main="Distribució quality")
```



Apareixen com a valors extrems de la variable quality els valors 3, 8 i 9. Com que l'escala (el rang) de la variable va del 0 al 10, els donarem com a valors vàlids.

Cal dir que a l'eliminar les files en haver-hi valors considerats extrems, pot ser que en alguna variable no hagin aparagut valors extrems o valors que haguessim considerat extrems perquè en eliminar files amb el criteri d'una altra variable, hagin estat eliminades files que tenien valors extrems en altres columnes.

Tornem a observar la dimensió del dataset.

```
dim(wine)
```

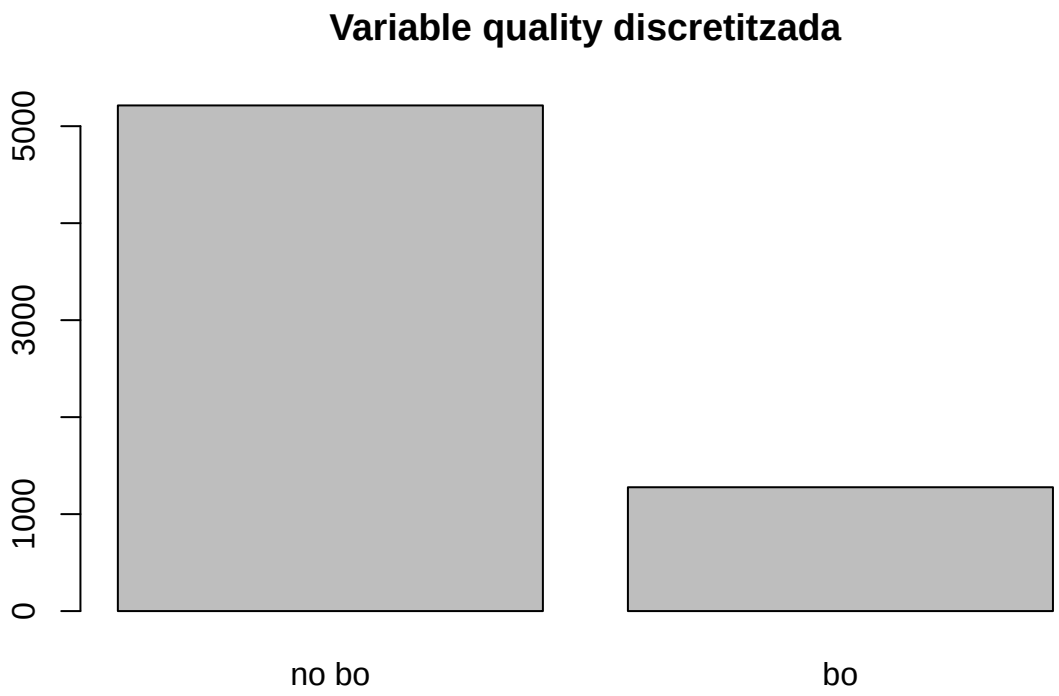
```
## [1] 6491   13
```

El dataset final de treball té 6491 files i 13 variables.

3. Altres. Distretització de la variable qualitat del vi.

Discretitzem la variable "quality", substituint els valors numèrics per etiquetes (bo/no bo). Per tant, es tracta de dicotomització. Amb aquesta variable nova podrem interpretar i comparar resultats. Classifiquem els 7 o superior com a "bo" i la resta com a "no bo".

```
wine["quality_d"] <- cut(wine$quality, breaks = c(0,6.5,10), labels = c("no bo", "bo"))
plot(wine$quality_d, main="Variable quality discretitzada")
```

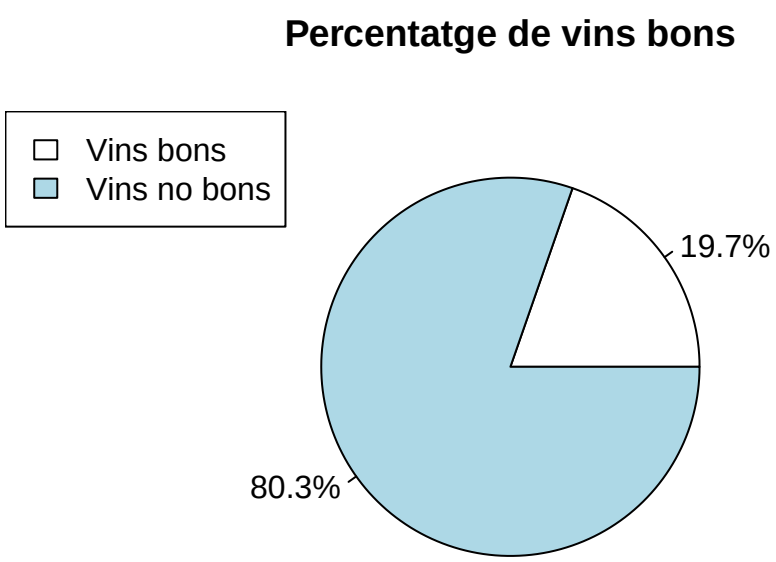


Observem ara el percentatge de vins bons amb un pie chart.

```
# Número de vins bons i la resta
total_bo <- sum(wine$quality_d == "bo")
total_nobo <- sum(wine$quality_d == "no bo")

# Percentatges
perc_bo <- (total_bo*100)/nrow(wine)
perc_nobo <- (total_nobo*100)/nrow(wine)

# Gràfic de la proporció de vins bons i no bons
pie(c(perc_bo, perc_nobo), labels=paste0(c(round(perc_bo,1), round(perc_nobo,1)), "%"),
main = "Percentatge de vins bons")
legend("topleft", legend = c("Vins bons", "Vins no bons"), fill = c("white", "lightblue"))
```



4. Anàlisi de les dades.

4.1. Selecció dels grups de dades

Selecció dels grups de dades que es volen analitzar/comparar (p. e., si es volen comparar grups de dades, quins són aquests grups i quins tipus d'anàlisi s'aplicaran?).

Un cop integrades i netejades les dades, és el moment de l'anàlisi de les dades. Preparem els grups de dades que volem analitzar o comparar. En aquest cas només tenim dues variables categòriques: quality_d, que indica si el vi és bo o no, i "type" que indica si el vi es negre o blanc.

Compararem si la qualitat del vi blanc i del vi negre són percebudes diferents, i farem alguns anàlisis diferenciats segons el tipus de vi. D'altra banda, utilitzarem els vins "bons" i els que no, per veure les diferències en les característiques.

ESPECIFICAR ANALISIS A FER

4.2. Comprovació de la normalitat i homogeneïtat de la variància.

Per comprovar la normalitat de cada variable farem servir el test de Kolmogorov-Smirnov, ja que la prova de Shapiro-Wilk accepta fins a 5000 registres i en tenim més.

```
ks.test(wine$fixed.acidity, pnorm, mean(wine$fixed.acidity), sd(wine$fixed.acidity))
```

```
## Warning in ks.test(wine$fixed.acidity, pnorm, mean(wine$fixed.acidity), : ties
## should not be present for the Kolmogorov-Smirnov test

##
##  One-sample Kolmogorov-Smirnov test
##
## data:  wine$fixed.acidity
## D = 0.13042, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
ks.test(wine$volatile.acidity, pnorm, mean(wine$volatile.acidity), sd(wine$volatile.acidity))
```

```
## Warning in ks.test(wine$volatile.acidity, pnorm, mean(wine$volatile.acidity), :
## ties should not be present for the Kolmogorov-Smirnov test

##
##  One-sample Kolmogorov-Smirnov test
##
## data:  wine$volatile.acidity
## D = 0.14952, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
ks.test(wine$citric.acid, pnorm, mean(wine$citric.acid), sd(wine$citric.acid))
```

```
## Warning in ks.test(wine$citric.acid, pnorm, mean(wine$citric.acid),
## sd(wine$citric.acid)): ties should not be present for the Kolmogorov-Smirnov
## test

##
##  One-sample Kolmogorov-Smirnov test
##
## data:  wine$citric.acid
## D = 0.080399, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
ks.test(wine$residual.sugar, pnorm, mean(wine$residual.sugar), sd(wine$residual.sugar))
```

```
## Warning in ks.test(wine$residual.sugar, pnorm, mean(wine$residual.sugar), : ties
## should not be present for the Kolmogorov-Smirnov test

##
##  One-sample Kolmogorov-Smirnov test
##
## data:  wine$residual.sugar
## D = 0.20215, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
ks.test(wine$chlorides, pnorm, mean(wine$chlorides), sd(wine$chlorides))
```

```
## Warning in ks.test(wine$chlorides, pnorm, mean(wine$chlorides),
## sd(wine$chlorides)): ties should not be present for the Kolmogorov-Smirnov test
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  wine$chlorides
## D = 0.18437, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
ks.test(wine$free.sulfur.dioxide, pnorm, mean(wine$free.sulfur.dioxide), sd(wine$free.sulfur.dioxide))
```

```
## Warning in ks.test(wine$free.sulfur.dioxide, pnorm,
## mean(wine$free.sulfur.dioxide), : ties should not be present for the Kolmogorov-
## Smirnov test
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  wine$free.sulfur.dioxide
## D = 0.056971, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
ks.test(wine$total.sulfur.dioxide, pnorm, mean(wine$total.sulfur.dioxide), sd(wine$total.sulfur.dioxide
```

```
## Warning in ks.test(wine$total.sulfur.dioxide, pnorm,
## mean(wine$total.sulfur.dioxide), : ties should not be present for the
## Kolmogorov-Smirnov test
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  wine$total.sulfur.dioxide
## D = 0.049149, p-value = 4.807e-14
## alternative hypothesis: two-sided
```

```
ks.test(wine$density, pnorm, mean(wine$density), sd(wine$density))
```

```
## Warning in ks.test(wine$density, pnorm, mean(wine$density), sd(wine$density)):
## ties should not be present for the Kolmogorov-Smirnov test
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  wine$density
## D = 0.045164, p-value = 6.32e-12
## alternative hypothesis: two-sided
```

```
ks.test(wine$pH, pnorm, mean(wine$pH), sd(wine$pH))
```

```
## Warning in ks.test(wine$pH, pnorm, mean(wine$pH), sd(wine$pH)): ties should not
## be present for the Kolmogorov-Smirnov test
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  wine$pH
## D = 0.042907, p-value = 8.341e-11
## alternative hypothesis: two-sided
```

```
ks.test(wine$sulphates, pnorm, mean(wine$sulphates), sd(wine$sulphates))
```

```
## Warning in ks.test(wine$sulphates, pnorm, mean(wine$sulphates),
## sd(wine$sulphates)): ties should not be present for the Kolmogorov-Smirnov test
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  wine$sulphates
## D = 0.094896, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
ks.test(wine$alcohol, pnorm, mean(wine$alcohol), sd(wine$alcohol))
```

```
## Warning in ks.test(wine$alcohol, pnorm, mean(wine$alcohol), sd(wine$alcohol)):
## ties should not be present for the Kolmogorov-Smirnov test
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  wine$alcohol
## D = 0.095909, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
ks.test(wine$quality, pnorm, mean(wine$quality), sd(wine$quality))
```

```
## Warning in ks.test(wine$quality, pnorm, mean(wine$quality), sd(wine$quality)):
## ties should not be present for the Kolmogorov-Smirnov test
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  wine$quality
## D = 0.22099, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

Veiem com la normalitat de totes les variables (fixed.acidity, volatile.acidity, citric.acid, residual.sugar, chlorides, free.sulfur.dioxide, total.sulfur.dioxide, density, pH, sulphates, alcohol i quality), tenen un valor p inferior a 0.05, i acceptem que les dades no provenen d'una distribució normal ja que rebutgem la hipòtesi nul · la, en totes les variables.

Pel que fa a la homoscedasticitat, contrastem amb la prova de Levene la igualtat de variàncies entre grups que necessitarem saber posteriorment.

```
library(car) # Llibreria pel test de homoscedasticitat (levene)
```

```
## Loading required package: carData
```

```
# Comprovant la homoscedasticitat entre tipus de vi en la variable qualitat.
leveneTest(quality ~ type, data = wine)
```

```
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.

## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value Pr(>F)
## group    1  2.3097 0.1286
##       6489
```

Observem que les variancies entre el vi blanc i negre són iguals pel que fa a la qualitat ja que el p-valor superior a 0.05 ens porta a acceptar la hipòtesis nul · la d'igualtat de variancies per als diferents tipus de vi.

***FALTA HOMOGENEÏTAT DE LA VARIÀNCIA, HA DE SER NOMÉS COMPROVADA LA QUE FAREM SERVIR?

4.3. Aplicació de proves estadístiques per comparar els grups de dades.

En funció de les dades i de l'objectiu de l'estudi, aplicar proves de contrast d'hipòtesis, correlacions, regressions, etc. Aplicar almenys tres mètodes d'anàlisi diferents.

Fem una matriu de correlació entre les variables corresponents a les característiques i la variable quality. Com que a l'apartat anterior s'ha vist que les dades (cap variable) no segueix una distribució normal, la correlació serà fet amb Spearman, una alternativa no paramètrica que mesura el grau de dependència entre dues variables i no comporta cap suposició sobre la distribució de les dades.

```
library(corrplot)
```

```
## corrplot 0.90 loaded
```

```
library(dplyr)
```

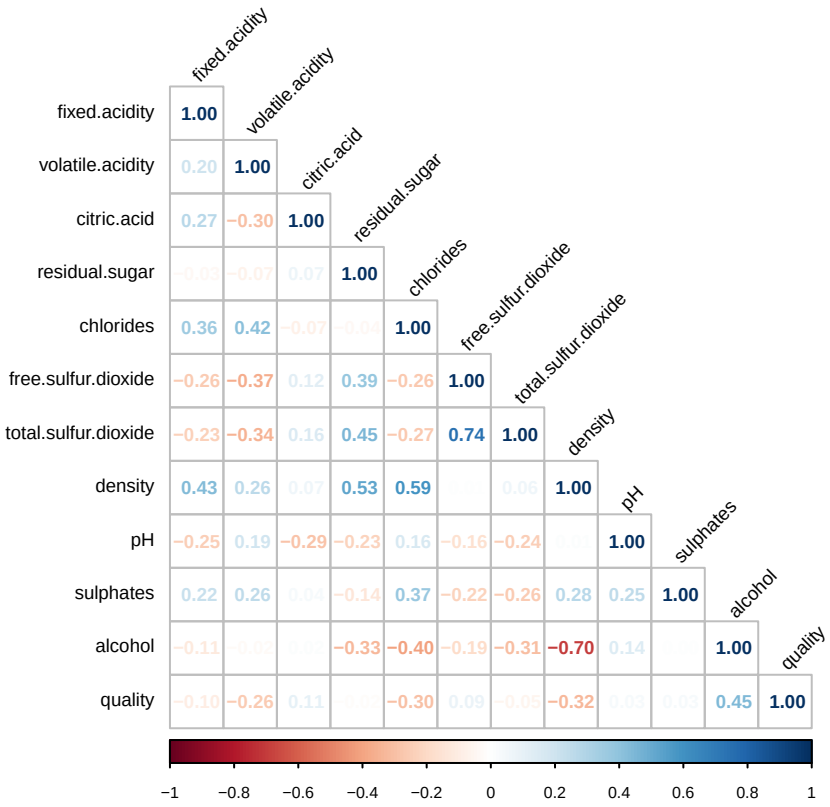
```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:car':
##
##     recode

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
corrplot(corr = cor(x = select(wine, fixed.acidity, volatile.acidity, citric.acid, residual.sugar, chlo
                               density, pH, sulphates, alcohol, quality), method = "spearman"),
         method = "number", type = "lower", cl.cex = 0.5, number.cex = 0.6,
         tl.col = "black", tl.srt = 45, tl.cex = 0.6)
```



Atenent als coeficients de correlació *(method = "spearman")*, observem que les relacions més destacables respecte quality es donen amb alcohol (0.45) i density (-0.32). També veiem una relació negativa moderada entre chlorides i la qualitat (-0.30). La resta són relacions dèbils/baixes. Les dues correlacions anomenades són correlacions moderades: la primera implica que a mesura que en certa mesura, a mesura que augmenta l'alcohol, augmenta la qualitat percebuda del vi; i la segona implica que a mesura que augmenta la densitat, disminueix la qualitat. A mesura que augmenten els chlorides, disminueix en certa manera la qualitat.

Cal destacar també la relació entre les dues variables següents: la densitat (density) i l'alcohol (alcohol), amb un coeficient de -0.70. Aquesta forta correlació negativa ens fa pensar que degut a la química, la quantitat d'alcohol redueix la densitat, i d'aquí que la quantitat d'alcohol serà la millor opció com a predictor de la qualitat del vi.

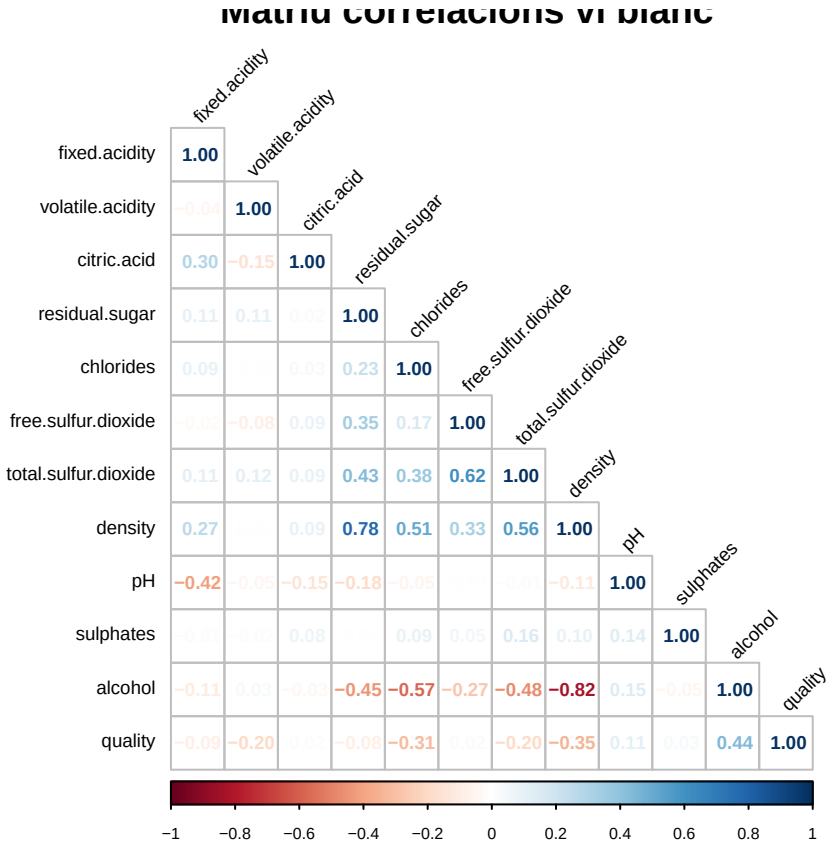
El SO2 lliure i el SO2 total estan altament correlacionats entre si, com podiem esperar.

Podem fer la matriu de correlacions pels vins blancs i pels vins negres, per observar si aquestes correlacions es veuen accentuades en algun dels dos tipus de vins.

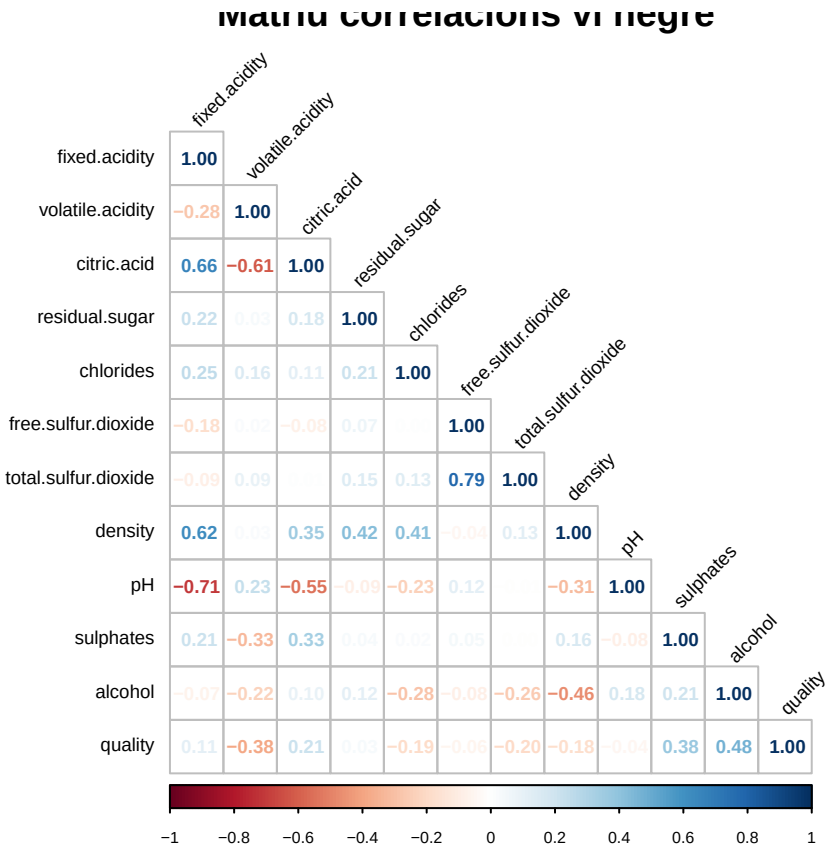
```
white_wine <- subset(wine, type=="White")
red_wine <- subset(wine, type=="Red")


corrplot(corr = cor(x = select(white_wine, fixed.acidity, volatile.acidity, citric.acid, residual.sugar
```

```
                                density, pH, sulphates, alcohol, quality), method = "spearman"),
         method = "number", type = "lower", cl.cex = 0.5, number.cex = 0.6,
         tl.col = "black", tl.srt = 45, tl.cex = 0.6, main = "Matriu correlacions vi blanc")
```




```
corrplot(corr = cor(x = select(red_wine, fixed.acidity, volatile.acidity, citric.acid, residual.sugar,
                                density, pH, sulphates, alcohol, quality), method = "spearman"),
         method = "number", type = "lower", cl.cex = 0.5, number.cex = 0.6,
         tl.col = "black", tl.srt = 45, tl.cex = 0.6, main = "Matriu correlacions vi negre")
```



Veiem com en el cas dels vins blancs les correlacions relacionades amb la qualitat són molt similars.

En el cas dels vins negres, la relació de densitat amb la qualitat ara és de -0.18 (molt dèbil), la relació de l'alcohol amb la qualitat augmenta una mica respecte el total (0.48) i apareix una altra correlació mitjana, entre la volatilitat de l'acidesa (volatile acidity) i la qualitat (-0.38), en el sentit que a més acidesa volàtil, menys qualitat. Tampoc té gaire rellevància la correlació entre els chlorides i la qualitat (coeficient de -0.19)

Com que hem vist que l'alcohol està correlacionat amb la qualitat del vi, observem la distribució de la variable alcohol segons si el vi és bo o no:

```
boxplot(wine$alcohol ~ wine$quality_d, main="Distribució alcohol segons bo/no bo")
```

Ara cal analitzar estadísticament si la mitjana d'alcohol és diferent pels vins bons i els vins no bons. Per fer-ho, farem un contrast d'hipòtesi per la diferència de mitjanes de alcohol.

La pregunta de recerca és:

"La quantitat d'alcohol és diferent en els vins bons i els vins no bons?"

La *hipòtesi nul · la (H0)* és que la mitjana d'alcohol és iguals entre vins bons i dolents.

La *hipòtesi alternativa (H1)* és que la mitjana d'alcohol és diferents entre vins bons i dolents.

Apliquem un test de dues mostres sobre la mitjana amb variàncies desconegudes. Pel teorema del límit central podem assumir normalitat.

Per utilitzar l'estadístic adequat cal comprovar la igualtat de variàncies de les dues poblacions:

```
vibo <- wine[wine$quality_d == "bo",]
vinobo <- wine[wine$quality_d == "no bo",]

var.test(vibo$alcohol, vinobo$alcohol)
```

```
##
##  F test to compare two variances
##
## data:  vibo$alcohol and vinobo$alcohol
## F = 1.2995, num df = 1276, denom df = 5213, p-value = 1.099e-09
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.193046 1.418665
```

```
## sample estimates:
## ratio of variances
##           1.299475
```

El resultat del test és un valor p<0.05. Rebutgem la hipòtesi nul · la d'igualtat de variàncies: assumim que les variàncies són diferents amb un nivell de confiança del 95%. En conseqüència, el test correspondrà a un test de dues mostres independents sobre la mitjana amb variàncies desconegudes diferents.

```
t.test(vibo$alcohol, vinobo$alcohol, alternative = "two.sided", var.equal=FALSE, conf.level=0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  vibo$alcohol and vinobo$alcohol
## t = 31.619, df = 1786.8, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.099841 1.245309
## sample estimates:
## mean of x mean of y
##  11.43336  10.26078
```

El valor p del test és inferior a 0.05, per tant amb un nivell de confiança del 95% podem rebutjar la hipòtesi nul · la d'igualtat de mitjanes, i podem afirmar que la mitjana de alcohol és estadísticament diferent entre els vins bons i els no bons. Si veiem les mitjanes, veiem que la mitjana d'alcohol és més alta en els vins bons (11.43 graus) que en els vins no bons (10.26 graus).

***MÉS CONSTRAST D'HIPÒTESI?

.....

QUALITAT DE VINS BLANCS I NEGRES

Volem observar si les mitjanes de qualitat entre vins blancs i negres són les mateixes.

Malgrat no es compleix la normalitat, apliquem el teorema central del límit (que s'aplica a la mitjana de la mostra d'un conjunt de dades), i considerem que les dades segueixen una distribució normal, al tenir mides de les mostres grans. Pel que fa a la homoscedasticitat, en l'apartat 4.2 hem vist que les variàncies entre grups (tipus de vi) són iguals pel que fa a la variable qualitat.

Per tant, com que es compleixen els supòsits, podem aplicar la prova t de Student pel contrast d'hipòtesis.

```
# Comparem mitjanes de la qualitat entre el tipus de vi
t.test(quality ~ type, data = wine)
```

```
##
##  Welch Two Sample t-test
##
## data:  quality by type
## t = -10.168, df = 2951, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group Red and group White is not equal to 0
## 95 percent confidence interval:
##  -0.2890826 -0.1956181
## sample estimates:
##   mean in group Red mean in group White
##            5.636023            5.878373
```

Veiem com el p-valor de la prova és menor al nivell de significació (0.05), i rebutgem la hipòtesi nula; concloent que existeixen diferències estadísticament significatives entre el vi blanc i el negre en la qualitat percebuda. El vi blanc és percep amb una millor qualitat.

. . . . . (Si no es compleixen les condicions s'han d'aplicar podem aplicar una prova no paramètrica com Mann-Whitney (els grups de dades són independents).) (Prova per comparar mitjanes de la qualitat entre tipus de vi pq no es compleixen les condicions) (wilcox.test(quality ~ type, data = wine))

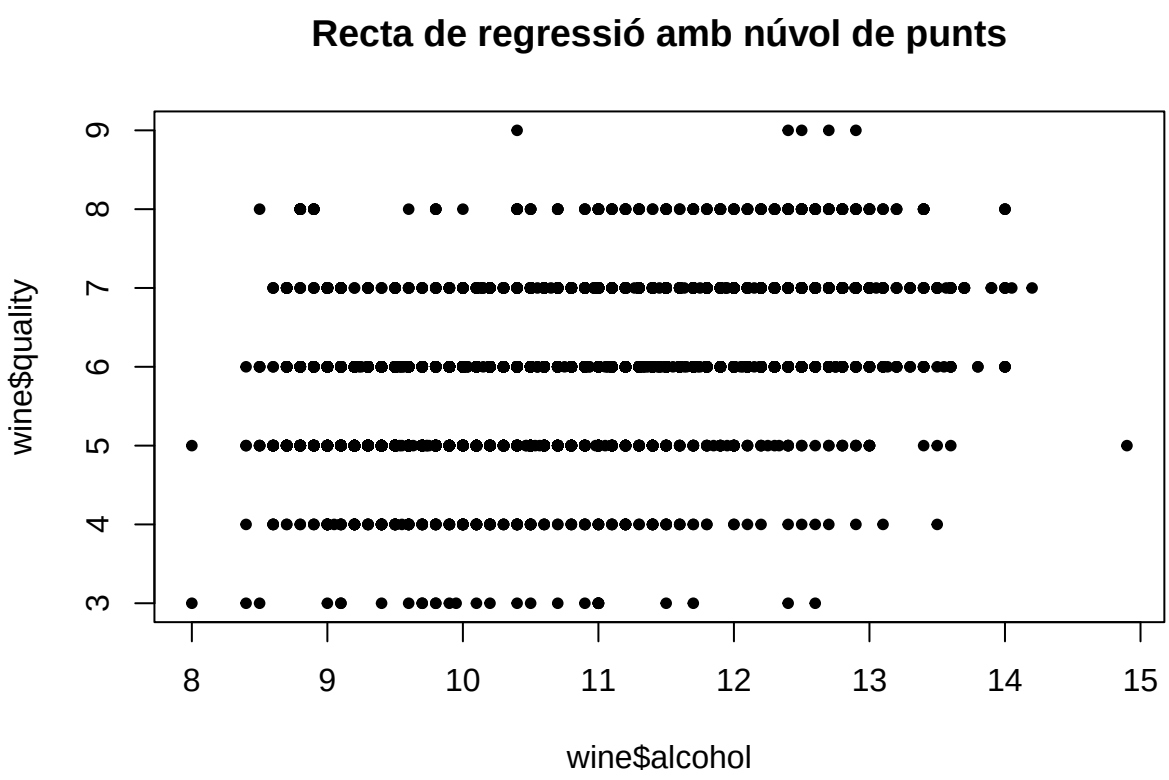
REGRESSIÓ

Utilitzem la funció lm per ajustar un model de regressió. Mostrem el gràfic amb la recta de regressió entre quality i les variables amb més correlació amb aquesta variable objectiu.

```
regr1 <- lm(alcohol ~ quality, data = wine)
summary(regr1)
```

```
##
## Call:
## lm(formula = alcohol ~ quality, data = wine)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3172 -0.7939 -0.1939  0.7061  4.9061
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.95499    0.08933   77.86   <2e-16 ***
## quality      0.60778    0.01518   40.03   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.068 on 6489 degrees of freedom
## Multiple R-squared:  0.1981, Adjusted R-squared:  0.1979
## F-statistic:  1603 on 1 and 6489 DF,  p-value: < 2.2e-16
```

```
plot(wine$alcohol, wine$quality, main = "Recta de regressió amb núvol de punts", pch=20)
abline(regr1, col="red") # FALTA QUE SURTI BÉ LA LINIA DE REGRESSIÓ, NO SE PQ NO SURT
```



Pel que fa a aquest model, és significatiu (p-valor menor que 0.05). Pel que fa a la bondat de l'ajust, el coeficient de determinació R2 és capaç d'explicar el 19.81% de la variabilitat present en la variable de resposta (quality) mitjançant la variable independent (alcohol).

**Si afegim la resta de variables al model, veiem que..... (fer un model de regressió lineal múltiple amb totes les variables i anar-les reduint fins a tenir el millor model de regressió per predir la qualitat). Intentem construir el millor model per predir la qualitat del vi.

LES MILLORS VARIABLES PER PREDIR LA QUALITAT DEL VI SÓN...

*** PODEM FER REGRESSIÓ LOGÍSTICA PER PREDIR EL RESULTAT DE QUALITAT DEL VI BO/NO BO, ja que és una variable dicotòmica dependent.