

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 negra 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 venda, etc.

Comptem amb dos datasets (un per la variant blanca i un per la negra) 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?**

Ahora 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.

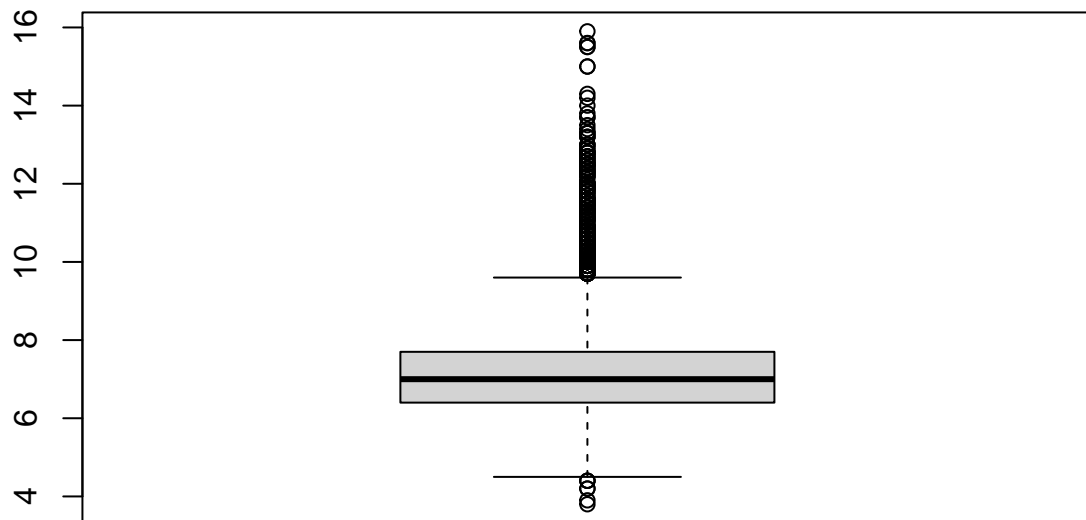
3.2. Valors extrems.

Identifica i gestiona els valors extrems.

Observem primer la distribució de la “fixed acidity”.

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

Distribució fixed acidity

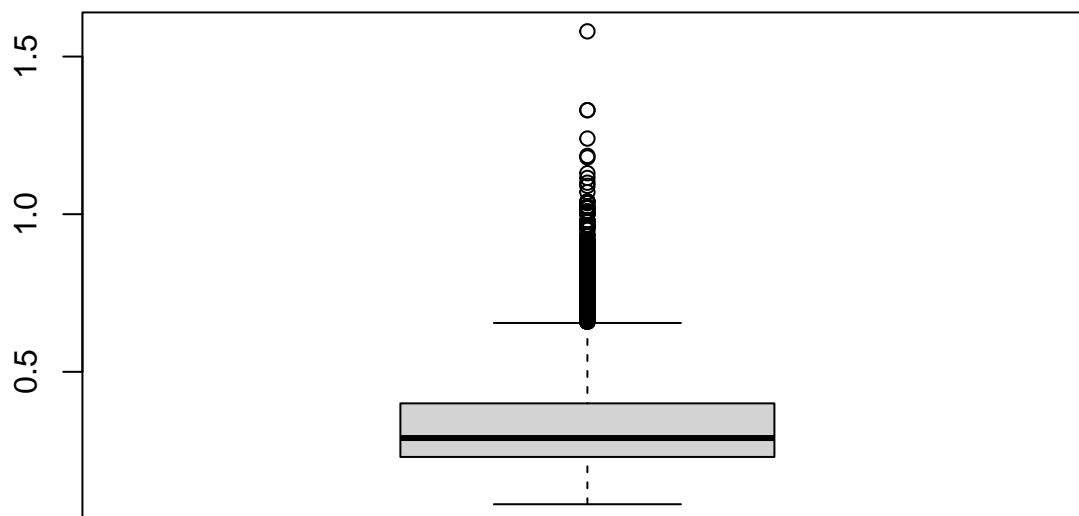


Malgrat hi ha molts valors atípics, no es consideren anòmals i per tant no els tractem.

Observem la distribució de la “volatile acidity”.

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

Distribució volatile acidity

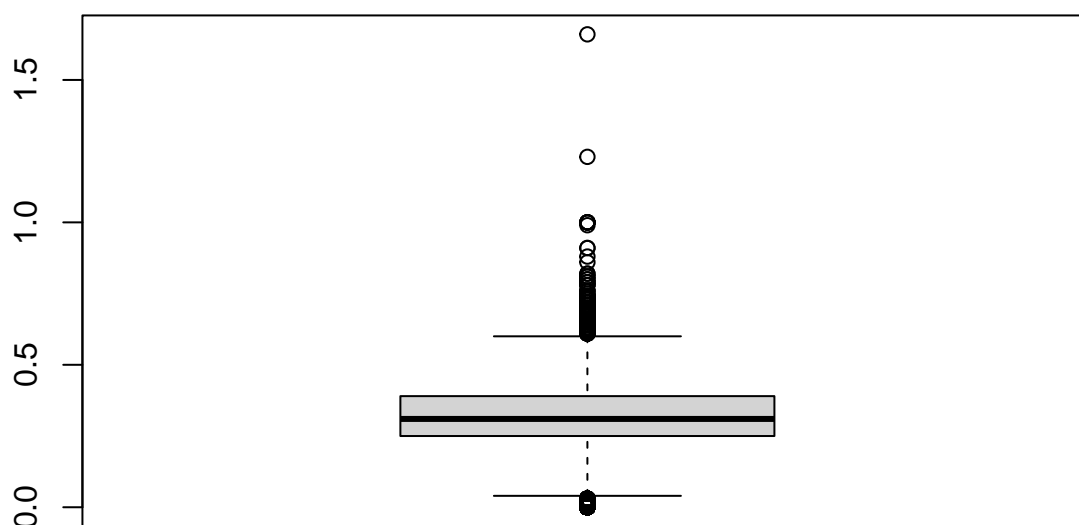


Malgrat hi ha molts valors atípics, no es consideren anòmals i per tant no els tractem.

Observem la distribució de l'àcid cítric.

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

Distribució cítric 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
```

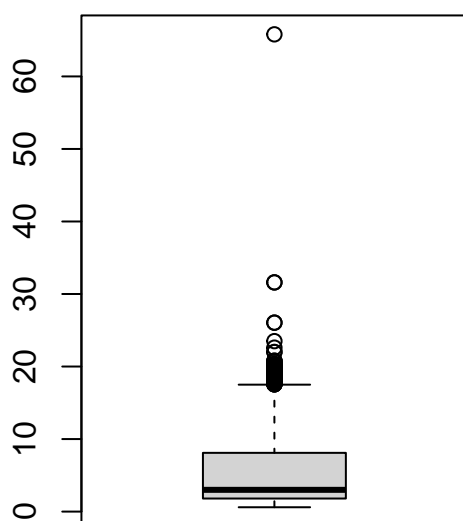
Veiem que només tenim dues files i és un percentatge ínfim del total i decidim eliminar les files.

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

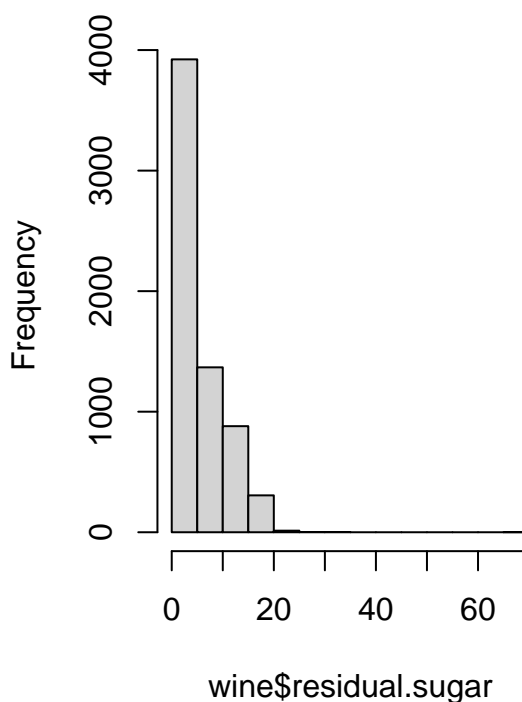
Observem ara la distribució del sucre residual.

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


Distribució residual sugar



Histogram of 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
```

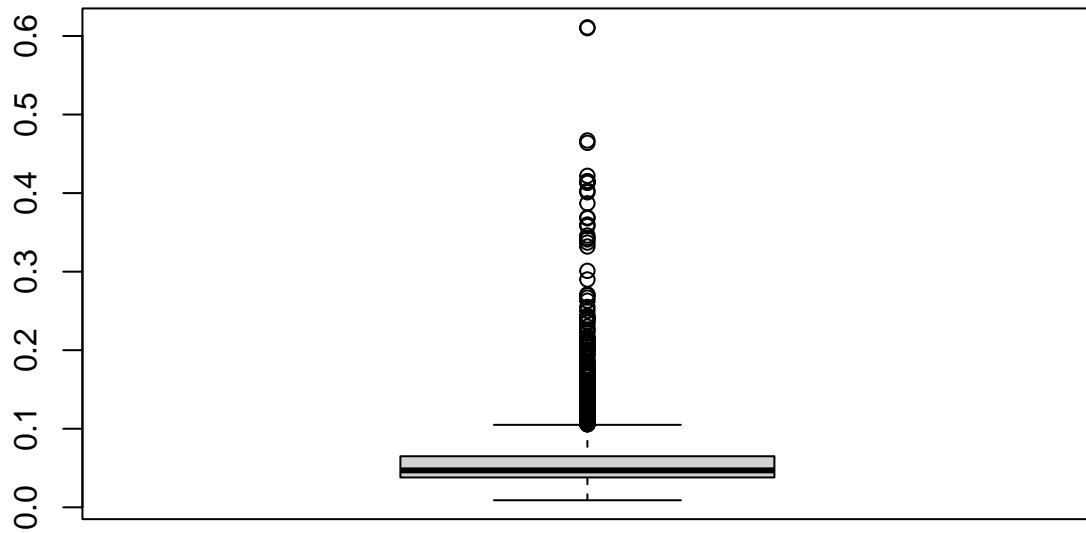
Veiem que només tenim tres files i és un percentatge ínfim del total i decidim eliminar les files.

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

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

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

Distribució chlorides

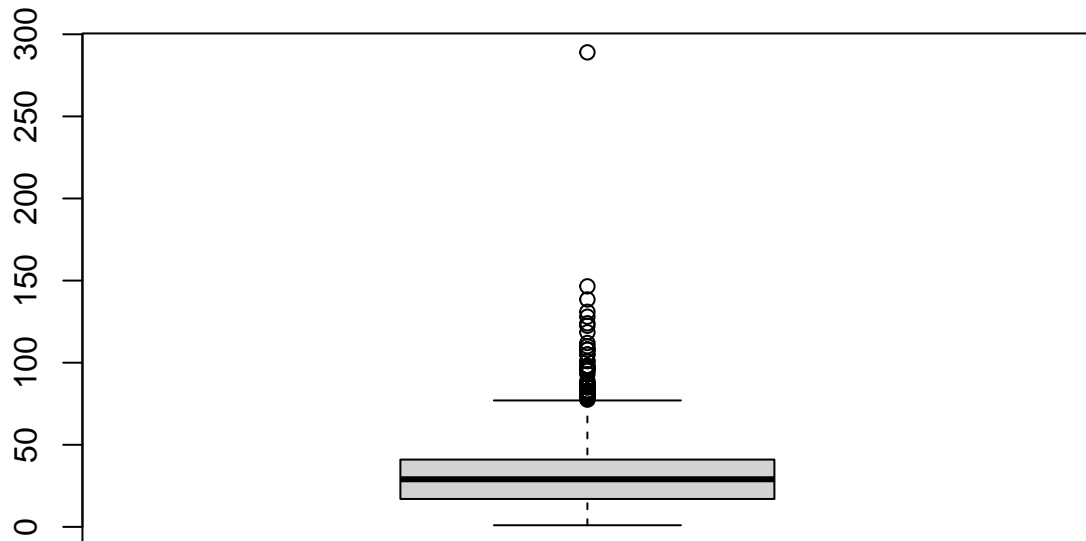


Malgrat hi ha molts valors atípics, no es consideren anòmals i per tant no els tractem.

Observem ara la distribució del diòxid de sofre lliure.

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

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
```

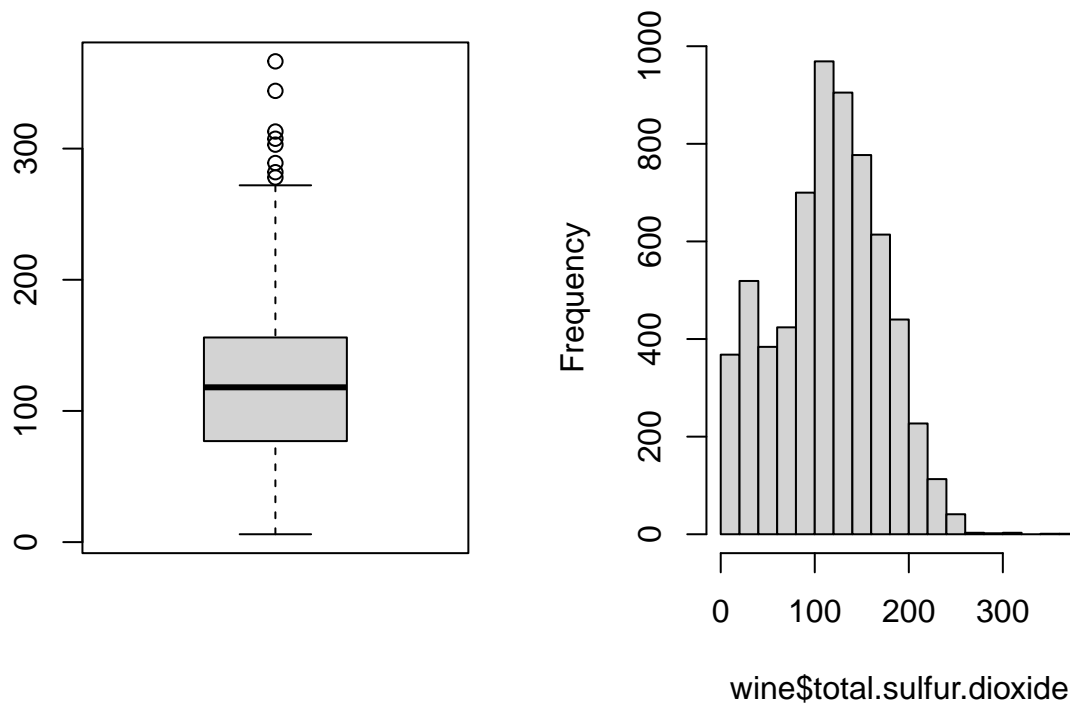
Veiem que només tenim una fila i decidim eliminar-la.

```
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)
```

Distribució total sulfur dioxide Histogram of wine\$total.sulfur.dio



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

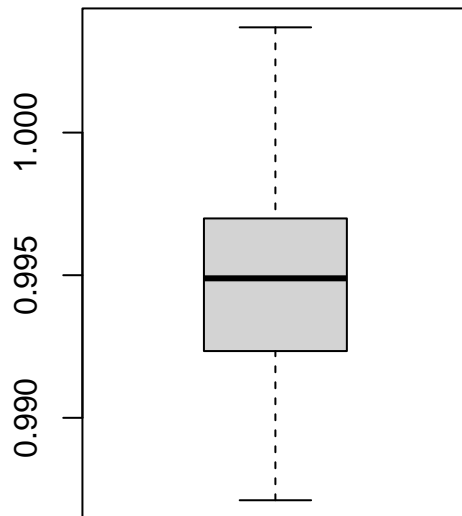
```
## [1] 278.0 289.0 313.0 366.5 307.5 344.0 282.0 303.0
```

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

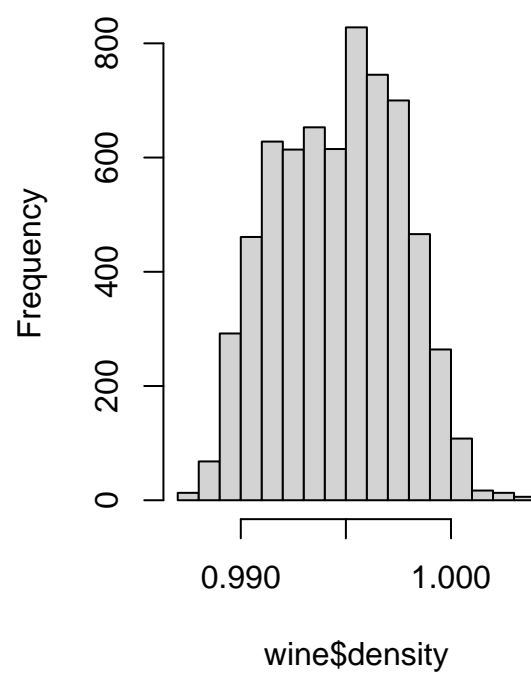
Observem ara la distribució de la densitat.

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

Distribució density



Histogram of wine\$density



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

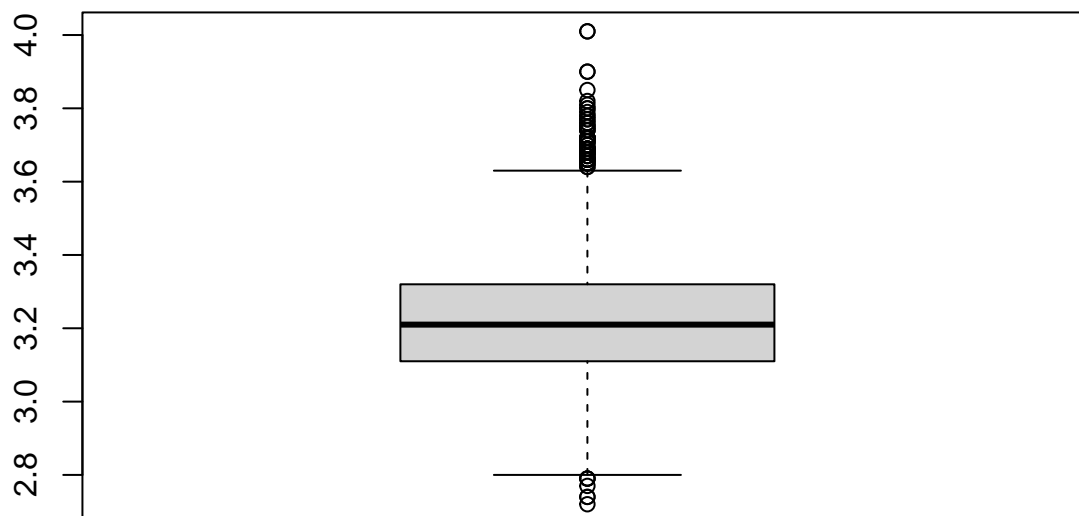
```
## numeric(0)
```

No tenim cap outlier en la variable density.

Observem ara la distribució del pH.

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

Distribució pH

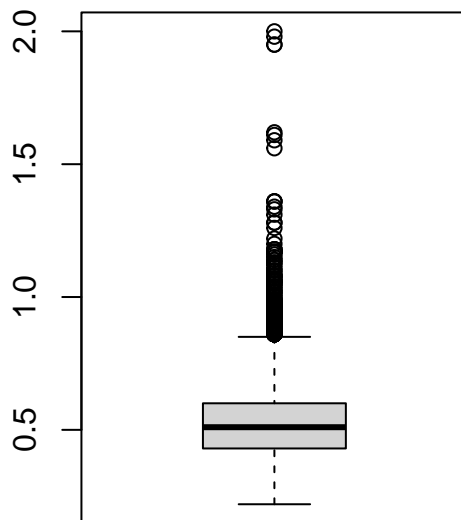


Observem valors atípics en la distribució del pH, però són valors lògics dins l'escala del pH, per tant els donem per vàlids.

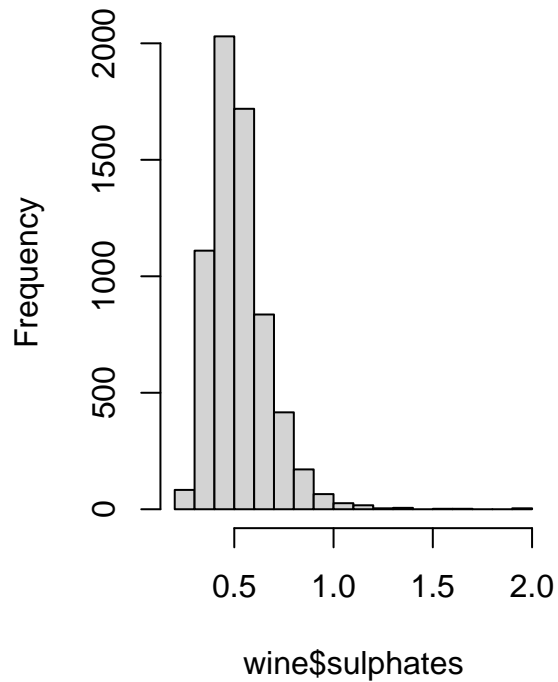
Mirem ara la distribució dels sulfats.

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

Distribució sulphates



Histogram of wine\$ sulphates



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

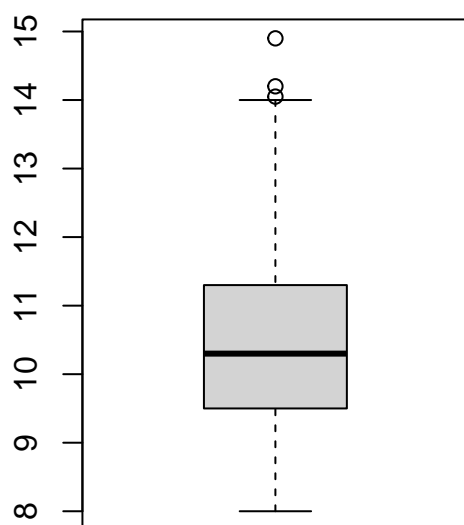
```
## [1] 1.56 0.88 0.93 1.28 1.08 0.91 0.91 0.90 1.20 0.95 1.12 1.28 1.14 1.95 1.22
## [16] 1.95 1.98 1.31 0.93 0.93 0.92 2.00 1.08 1.59 1.02 0.97 1.03 0.88 0.86 1.61
## [31] 1.09 0.96 0.96 1.26 0.87 0.86 0.91 0.97 0.97 0.91 0.97 1.08 0.86 0.95 0.86
## [46] 1.00 1.36 1.18 0.87 0.89 0.93 0.92 0.86 0.98 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.

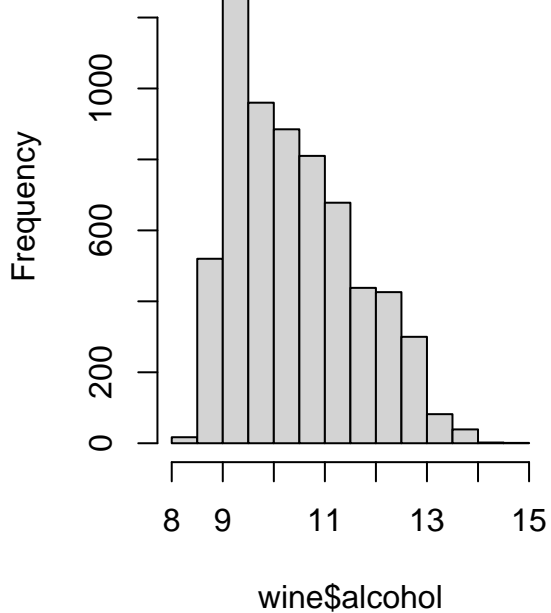
Observem ara la distribució del percentatge d'alcohol.

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

Distribució alcohol



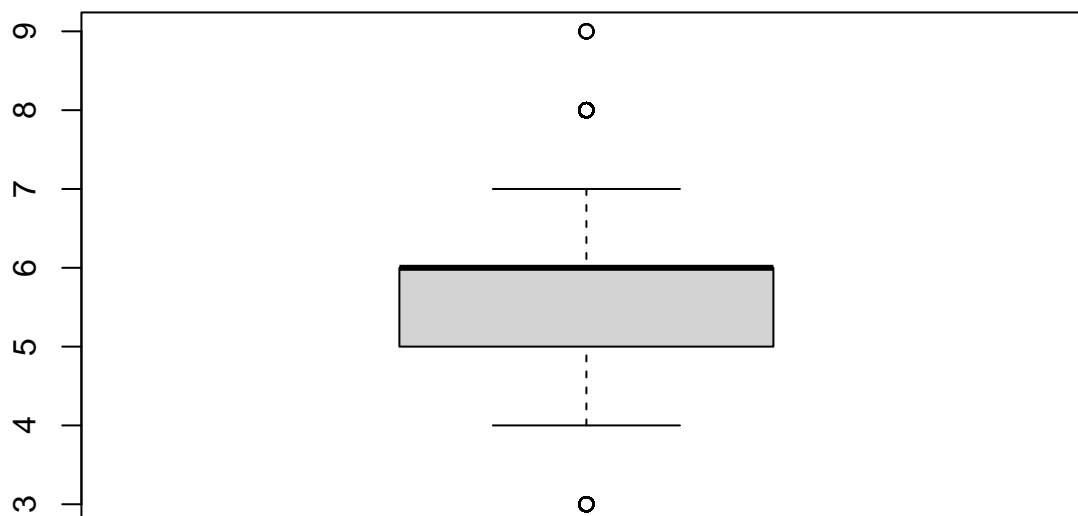
Histograma alcohol



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")
```


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
```

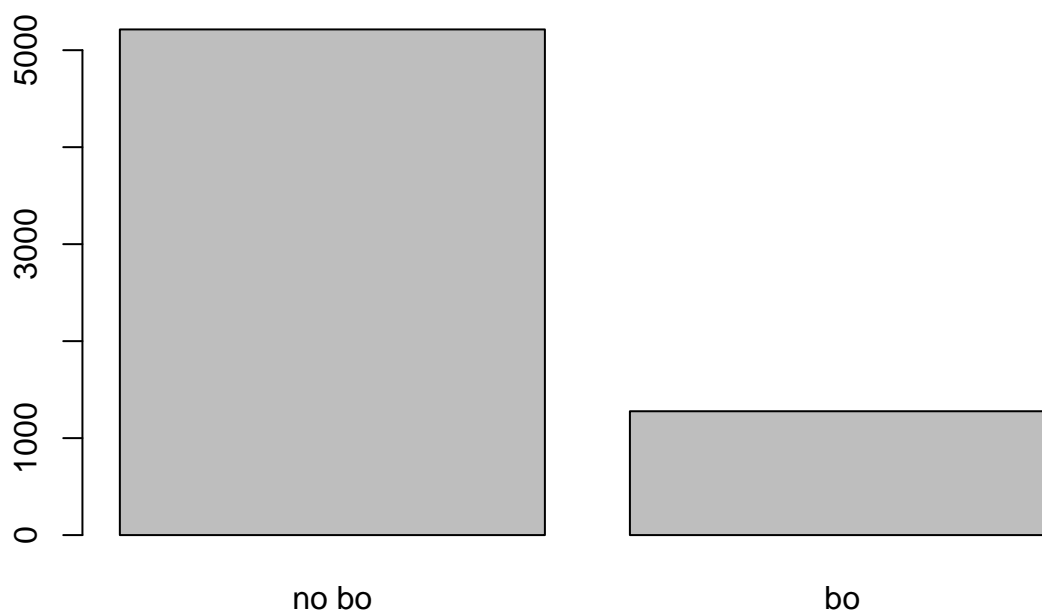
El dataset final de treball té 6491 files i 13 variables.

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 dicotomitització. 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")
```

Variable quality discretitzada



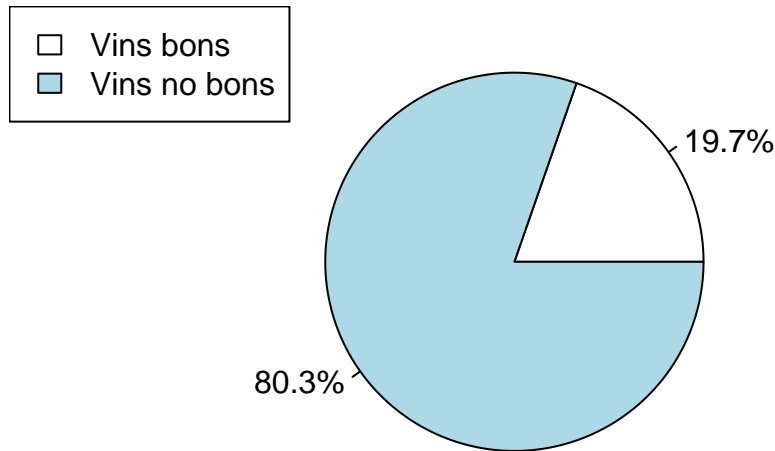
Observem ara el percentatge de vins bons amb un pie chart.

```
# 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"))
```

Percentatge de vins bons



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 necessitem saber posteriorment.

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

```
## Warning: package 'car' was built under R version 4.1.3
```

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

```
## Warning: package 'carData' was built under R version 4.1.3
```

```
# 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 variàncies 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 hipòtesi nul·la d'igualtat de variàncies 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)
```

```
## Warning: package 'corrplot' was built under R version 4.1.3
```

```
## corrplot 0.92 loaded
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.1.3
```

```
##
```

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

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

```
##
```

```
##      recode
```

```
## The following objects are masked from 'package:stats':
```

```
##
```

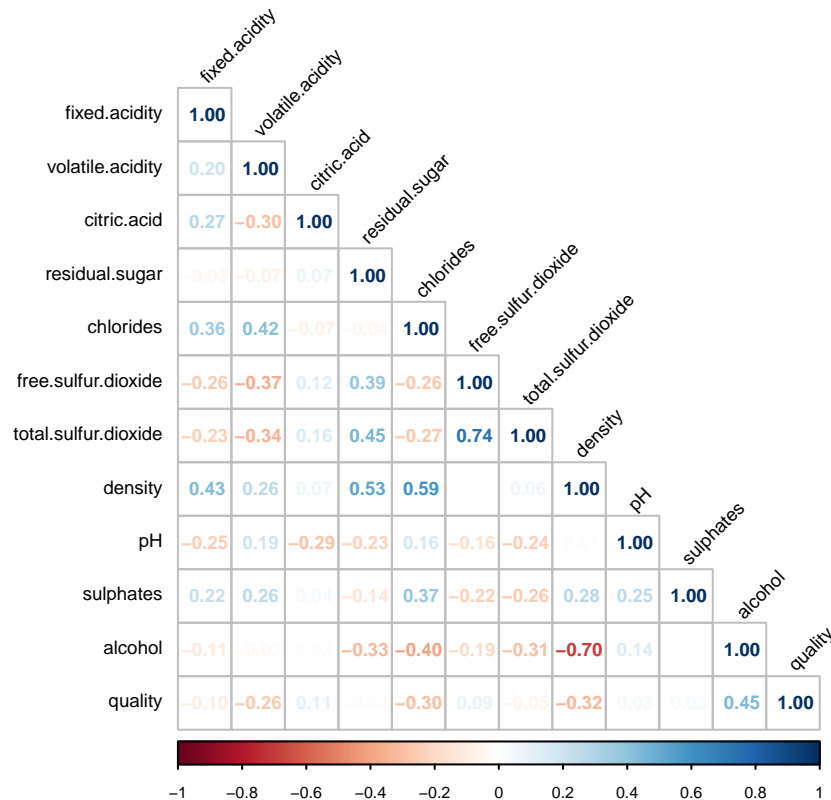
```
##      filter, lag
```

```
## 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, chlorides,
                                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.

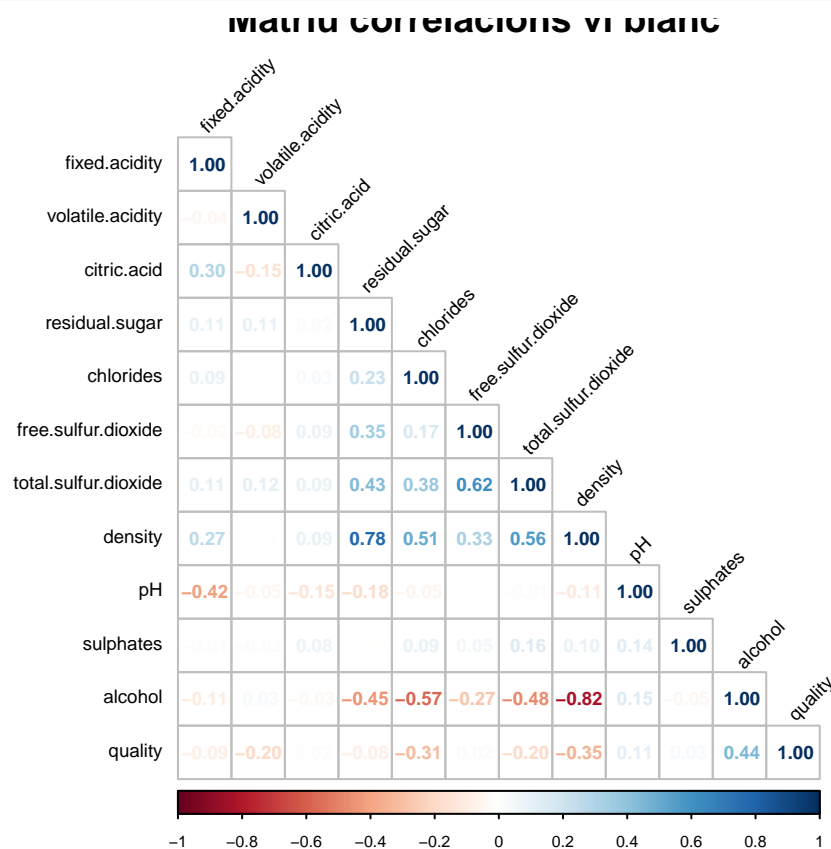
Cal destacar també la relació entre les dues variables següents: la densitat (*density*) i l'alcohol (*alcohol*), amb un coeficient de -0.70. Aquesta forta correlació negativa ens fa pensar que degut a la química, la quantitat d'alcohol redueix la densitat, i d'aquí que la quantitat d'alcohol serà la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

El SO₂ lliure i el SO₂ total estan altament correlacionats entre si, com podem 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")
```

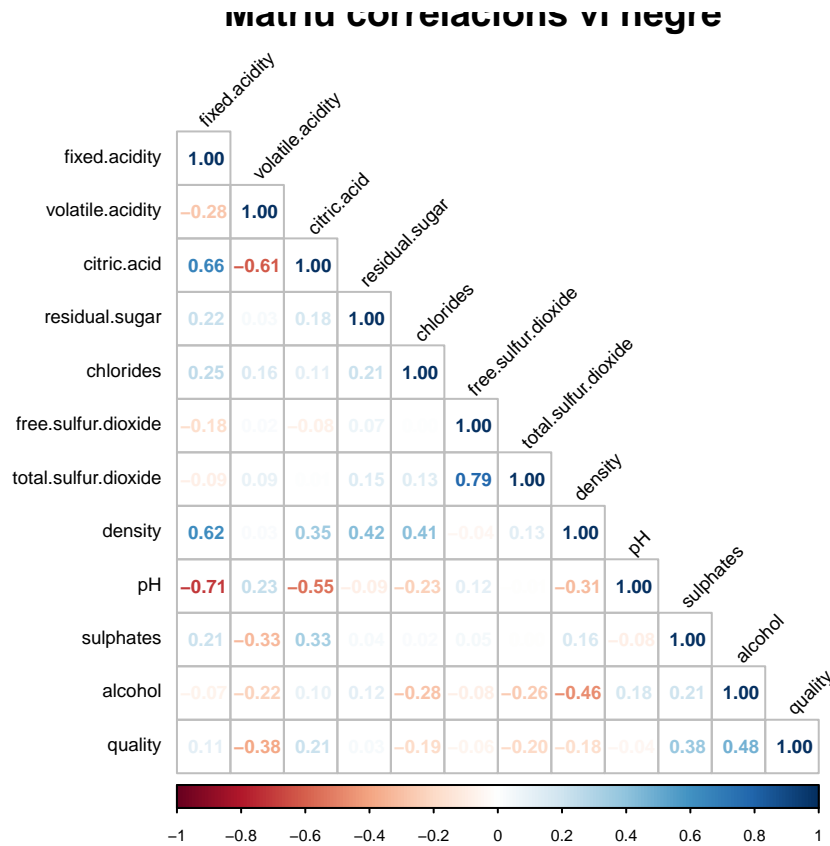
```
corrplot(corr = cor(x = select(white_wine, fixed.acidity, volatile.acidity, citric.acid, residual.sugar,
                              density, pH, sulphates, alcohol, quality), method = "spearman"),
          method = "number", type = "lower", cl.cex = 0.5, number.cex = 0.6,
          tl.col = "black",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



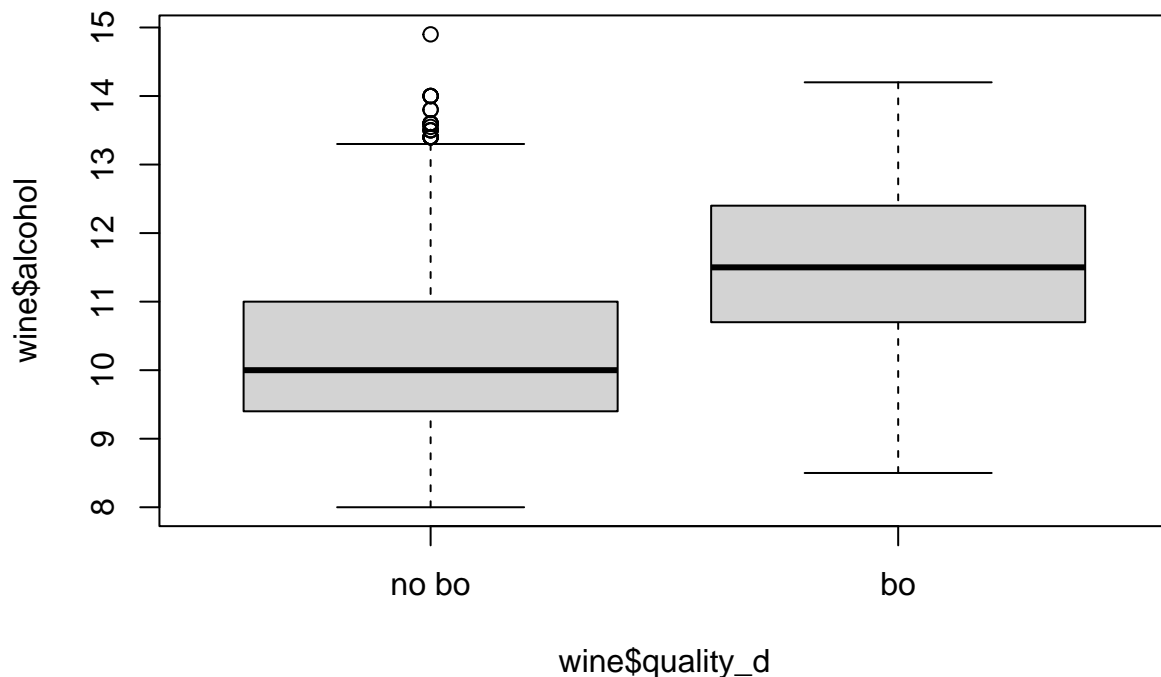
Veiem com en el cas dels vins blancs les correlacions relacionades amb la qualitat són molt similars.

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")
```

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* (H_0) és que la mitjana d'alcohol és iguals entre vins bons i dolents.

La *hipòtesi alternativa* (H_1) 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 CONTRAST 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 percebut 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

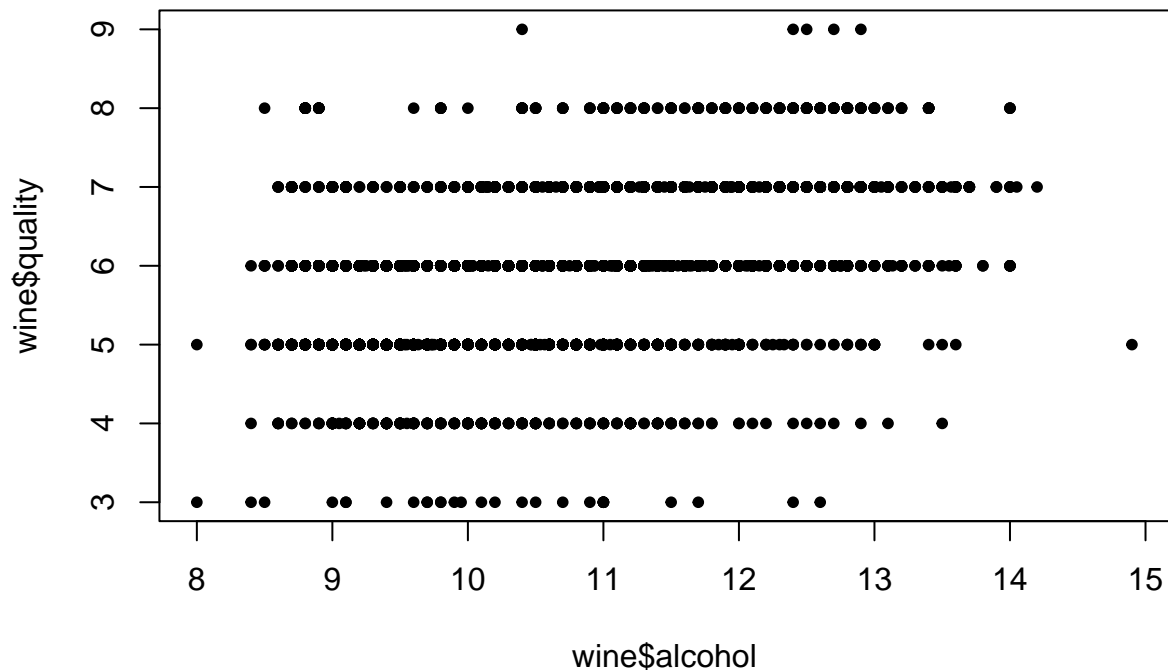
Utilitzem la funció lm per ajustar un model de regressió. Mostrem el gràfic amb la recta de regressió entre quality i les variables amb més correlació amb aquesta variable objectiu.

```
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
```

Recta de regressió amb núvol de punts



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 R^2 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.

```
# La variable "target" serà la quality
reg2 <- lm(quality ~ alcohol + volatile.acidity + sulphates + residual.sugar +
  citric.acid + type + free.sulfur.dioxide + density + total.sulfur.dioxide +
  chlorides + pH + fixed.acidity, data = wine)
summary(reg2)
```

```
##
## Call:
## lm(formula = quality ~ alcohol + volatile.acidity + sulphates +
##   residual.sugar + citric.acid + type + free.sulfur.dioxide +
##   density + total.sulfur.dioxide + chlorides + pH + fixed.acidity,
##   data = wine)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.6196 -0.4691 -0.0411  0.4576  3.0245
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)          1.257e+02  1.570e+01   8.008 1.37e-15 ***
## alcohol              1.985e-01  1.983e-02  10.010 < 2e-16 ***
## volatile.acidity     -1.484e+00  8.127e-02 -18.263 < 2e-16 ***
## sulphates            7.465e-01  7.645e-02   9.765 < 2e-16 ***
## residual.sugar       6.788e-02  6.274e-03  10.819 < 2e-16 ***
## citric.acid          -5.501e-02  8.062e-02  -0.682   0.495
## typeWhite            -4.177e-01  5.915e-02  -7.061 1.83e-12 ***
## free.sulfur.dioxide   5.691e-03  7.785e-04   7.309 3.01e-13 ***
## density              -1.251e+02  1.591e+01  -7.866 4.27e-15 ***
## total.sulfur.dioxide -1.333e-03  3.251e-04  -4.101 4.16e-05 ***
## chlorides            -7.433e-01  3.339e-01  -2.226   0.026 *
## pH                   5.710e-01  9.310e-02   6.133 9.14e-10 ***
## fixed.acidity        1.027e-01  1.668e-02   6.156 7.91e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7312 on 6478 degrees of freedom
## Multiple R-squared:  0.2996, Adjusted R-squared:  0.2983
## F-statistic: 231 on 12 and 6478 DF, p-value: < 2.2e-16
```

Amb totes les variables introduïdes com predictors l'ajust és del 0.2996. És a dir, un el model és capaç d'explicar el 29.96% de la variabilitat observada en la qualitat del vi. A més a més, el valor *p-value* del model és significatiu ($2.2e-16$) i, per tant, es pot acceptar que el model no és per l'atzar.

Per millorar el model es seleccionen els millors predictors per eliminar interferències provocades entre ells. És realitzarà amb les funcionabilitats de la llibreria *MASS* per observar el valor *Akaike (AIC)*. Es realitza la comparació amb la metodologia *backward*.

```
regBack <- stepAIC(reg2, trace=FALSE, direction="backward")
summary(regBack)
```

```
##
## Call:
## lm(formula = quality ~ alcohol + volatile.acidity + sulphates +
##     residual.sugar + type + free.sulfur.dioxide + density + total.sulfur.dioxide +
##     chlorides + pH + fixed.acidity, data = wine)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.6082 -0.4721 -0.0415  0.4565  3.0237
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.263e+02  1.567e+01   8.059 9.11e-16 ***
## alcohol        1.970e-01  1.971e-02   9.995 < 2e-16 ***
## volatile.acidity -1.466e+00  7.654e-02 -19.147 < 2e-16 ***
## sulphates       7.438e-01  7.634e-02   9.743 < 2e-16 ***
## residual.sugar   6.799e-02  6.272e-03  10.841 < 2e-16 ***
## typeWhite       -4.214e-01  5.889e-02  -7.155 9.29e-13 ***
## free.sulfur.dioxide 5.692e-03  7.785e-04   7.312 2.96e-13 ***
## density         -1.257e+02  1.588e+01  -7.916 2.88e-15 ***
## total.sulfur.dioxide -1.353e-03  3.238e-04  -4.179 2.96e-05 ***
## chlorides       -7.816e-01  3.291e-01  -2.375   0.0176 *
## pH              5.760e-01  9.281e-02   6.206 5.77e-10 ***
## fixed.acidity    1.004e-01  1.634e-02   6.145 8.50e-10 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7312 on 6479 degrees of freedom
## Multiple R-squared:  0.2996, Adjusted R-squared:  0.2984
## F-statistic: 251.9 on 11 and 6479 DF,  p-value: < 2.2e-16

regBack$anova

## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## quality ~ alcohol + volatile.acidity + sulphates + residual.sugar +
##      citric.acid + type + free.sulfur.dioxide + density + total.sulfur.dioxide +
##      chlorides + pH + fixed.acidity
##
## Final Model:
## quality ~ alcohol + volatile.acidity + sulphates + residual.sugar +
##      type + free.sulfur.dioxide + density + total.sulfur.dioxide +
##      chlorides + pH + fixed.acidity
##
##
##              Step Df Deviance Resid. Df Resid. Dev      AIC
## 1                      6478   3463.667 -4050.923
## 2 - citric.acid   1  0.24896    6479   3463.916 -4052.457

Amb el model obtingut l'ajust millora fins el valor de El model obtingut elimina la variable citric.acid.

reg3 <- lm(quality ~ alcohol + volatile.acidity + sulphates + residual.sugar +
  type + free.sulfur.dioxide + density + total.sulfur.dioxide +
  chlorides + pH + fixed.acidity, data = wine)
summary(reg3)

##
## Call:
## lm(formula = quality ~ alcohol + volatile.acidity + sulphates +
##      residual.sugar + type + free.sulfur.dioxide + density + total.sulfur.dioxide +
##      chlorides + pH + fixed.acidity, data = wine)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.6082 -0.4721 -0.0415  0.4565  3.0237
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.263e+02  1.567e+01   8.059 9.11e-16 ***
## alcohol         1.970e-01  1.971e-02   9.995 < 2e-16 ***
## volatile.acidity -1.466e+00  7.654e-02 -19.147 < 2e-16 ***
## sulphates       7.438e-01  7.634e-02   9.743 < 2e-16 ***
## residual.sugar   6.799e-02  6.272e-03  10.841 < 2e-16 ***
## typeWhite      -4.214e-01  5.889e-02  -7.155 9.29e-13 ***
## free.sulfur.dioxide 5.692e-03  7.785e-04   7.312 2.96e-13 ***
## density        -1.257e+02  1.588e+01  -7.916 2.88e-15 ***
## total.sulfur.dioxide -1.353e-03  3.238e-04  -4.179 2.96e-05 ***
## chlorides      -7.816e-01  3.291e-01  -2.375  0.0176 *
## pH              5.760e-01  9.281e-02   6.206 5.77e-10 ***
```

```
## fixed.acidity          1.004e-01  1.634e-02   6.145 8.50e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7312 on 6479 degrees of freedom
## Multiple R-squared:  0.2996, Adjusted R-squared:  0.2984
## F-statistic: 251.9 on 11 and 6479 DF,  p-value: < 2.2e-16
```

Es proven altres metodologies per obtenir els millors predictors.

```
empty.model <- lm(quality ~ 1, data=wine)
horizonte <- formula(quality ~ alcohol + volatile.acidity + sulphates + residual.sugar +
  citric.acid + type + free.sulfur.dioxide + density + total.sulfur.dioxide +
  chlorides + pH + fixed.acidity)

sumBoth <- stepAIC(empty.model, trace=FALSE, direction="both", scope=horizonte)
sumBoth$anova
```

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## quality ~ 1
##
## Final Model:
## quality ~ alcohol + volatile.acidity + sulphates + residual.sugar +
##      type + free.sulfur.dioxide + density + total.sulfur.dioxide +
##      chlorides + pH + fixed.acidity
##
##
##
```

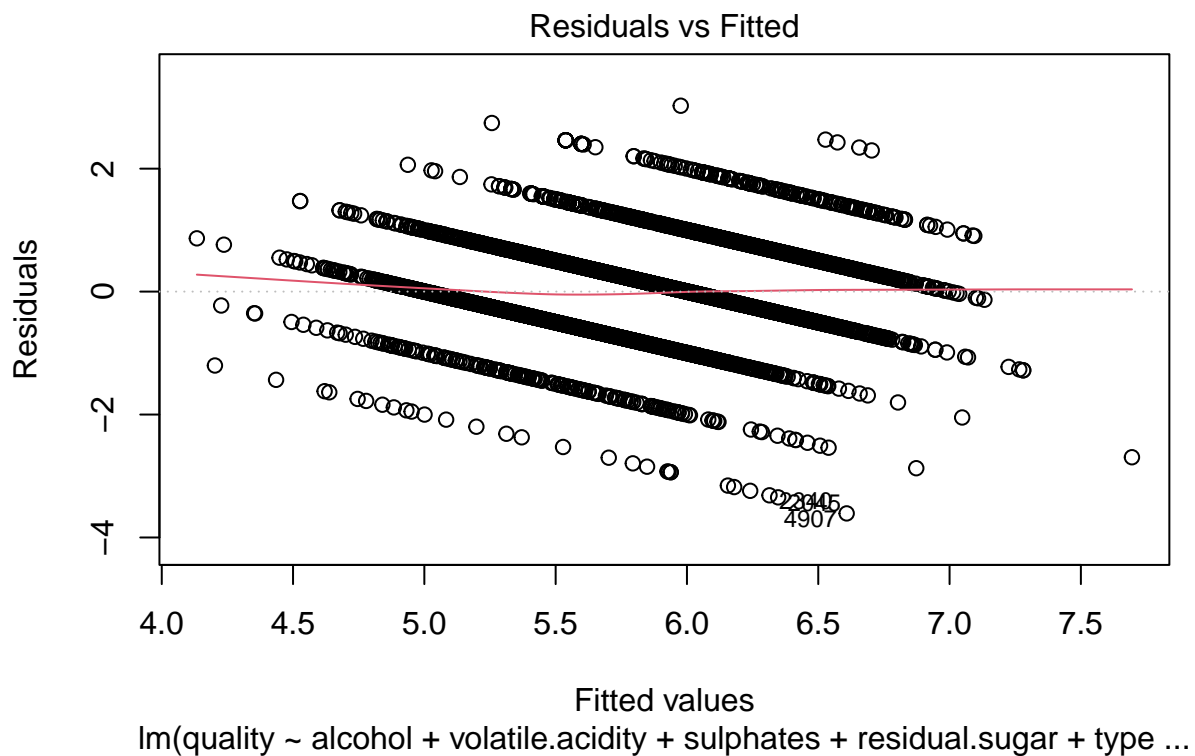
	Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
## 1				6490	4945.577	-1763.052
## 2	+ alcohol	1	979.505803	6489	3966.071	-3193.730
## 3	+ volatile.acidity	1	308.693188	6488	3657.378	-3717.693
## 4	+ sulphates	1	48.900920	6487	3608.477	-3803.066
## 5	+ residual.sugar	1	42.800663	6486	3565.676	-3878.517
## 6	+ type	1	20.306379	6485	3545.370	-3913.589
## 7	+ free.sulfur.dioxide	1	21.223263	6484	3524.147	-3950.562
## 8	+ density	1	18.088009	6483	3506.059	-3981.963
## 9	+ total.sulfur.dioxide	1	12.141006	6482	3493.918	-4002.480
## 10	+ chlorides	1	6.573556	6481	3487.344	-4012.703
## 11	+ pH	1	3.242561	6480	3484.102	-4016.742
## 12	+ fixed.acidity	1	20.185266	6479	3463.916	-4052.457

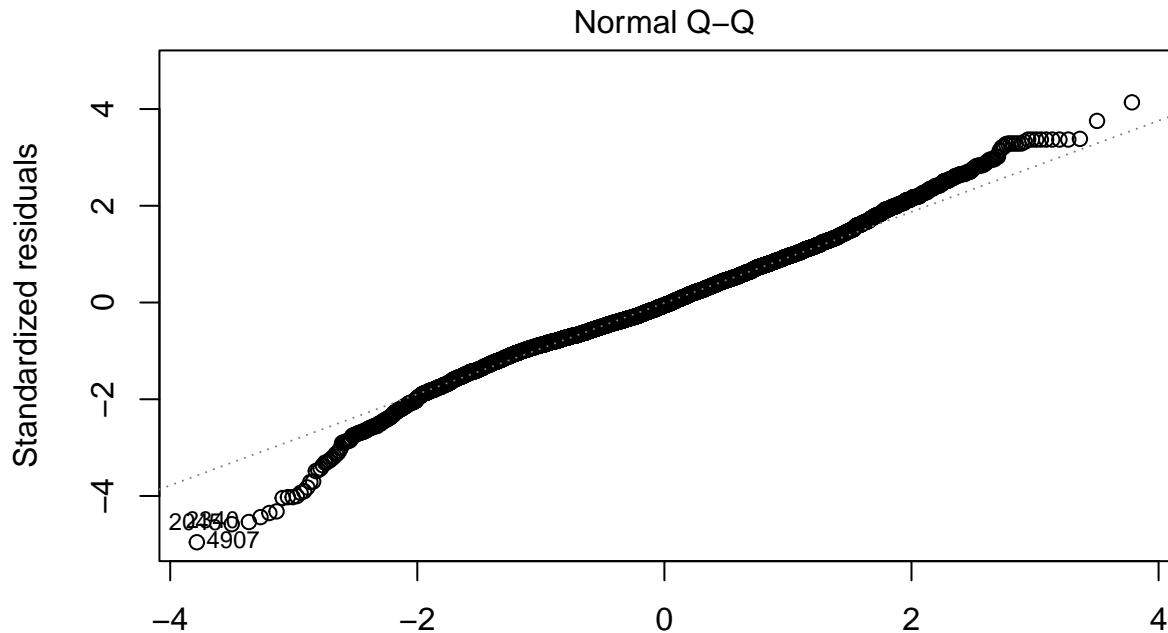
```
reg4 <- lm(quality ~ alcohol + volatile.acidity + sulphates + residual.sugar +
  type + free.sulfur.dioxide + density + total.sulfur.dioxide +
  chlorides + pH + fixed.acidity, data = wine)
summary(reg4)
```

```
##
## Call:
## lm(formula = quality ~ alcohol + volatile.acidity + sulphates +
##      residual.sugar + type + free.sulfur.dioxide + density + total.sulfur.dioxide +
##      chlorides + pH + fixed.acidity, data = wine)
##
## Residuals:
```



```
##      Min      1Q  Median      3Q      Max
## -3.6082 -0.4721 -0.0415  0.4565  3.0237
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.263e+02  1.567e+01   8.059 9.11e-16 ***
## alcohol        1.970e-01  1.971e-02   9.995 < 2e-16 ***
## volatile.acidity -1.466e+00  7.654e-02 -19.147 < 2e-16 ***
## sulphates       7.438e-01  7.634e-02   9.743 < 2e-16 ***
## residual.sugar  6.799e-02  6.272e-03  10.841 < 2e-16 ***
## typeWhite      -4.214e-01  5.889e-02  -7.155 9.29e-13 ***
## free.sulfur.dioxide 5.692e-03  7.785e-04   7.312 2.96e-13 ***
## density        -1.257e+02  1.588e+01  -7.916 2.88e-15 ***
## total.sulfur.dioxide -1.353e-03  3.238e-04  -4.179 2.96e-05 ***
## chlorides       -7.816e-01  3.291e-01  -2.375  0.0176 *
## pH              5.760e-01  9.281e-02   6.206 5.77e-10 ***
## fixed.acidity    1.004e-01  1.634e-02   6.145 8.50e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7312 on 6479 degrees of freedom
## Multiple R-squared:  0.2996, Adjusted R-squared:  0.2984
## F-statistic: 251.9 on 11 and 6479 DF,  p-value: < 2.2e-16
plot(reg4, which=c(2,1))
```





Theoretical Quantiles
lm(quality ~ alcohol + volatile.acidity + sulphates + residual.sugar + type ...

La distribució dels valors ajustats enfront dels residus estan prou centrats entorn del valor 0. Per tant, hi ha una correcta distribució de l'error en el model. Per altra banda, la funció QQ mostra desviacions en els valors extrems de la gràfica, però els punts centrals es mantenen propers a la línia.

A més a més, es generen mètodes per validar si algun es comporta millor.

```
# Generacio de models
model0 <- lm(quality ~ alcohol + volatile.acidity + sulphates + residual.sugar + citric.acid +
  type + free.sulfur.dioxide + density + total.sulfur.dioxide +
  chlorides + pH + fixed.acidity, data = wine)

model1 <- lm(quality ~ alcohol + volatile.acidity + sulphates + residual.sugar +
  type + free.sulfur.dioxide + density + total.sulfur.dioxide +
  chlorides + pH, data = wine)

model2 <- lm(quality ~ alcohol + volatile.acidity + sulphates + residual.sugar +
  type + free.sulfur.dioxide + density + total.sulfur.dioxide +
  chlorides + pH + fixed.acidity, data = wine)

model3 <- lm(quality ~ alcohol + sulphates + residual.sugar + citric.acid +
  free.sulfur.dioxide + density + total.sulfur.dioxide +
  chlorides + pH + fixed.acidity, data = wine)

model4 <- lm(quality ~ alcohol + sulphates + residual.sugar + citric.acid +
  type + free.sulfur.dioxide + density + total.sulfur.dioxide +
  chlorides + pH + fixed.acidity, data = wine)
```

```

model5 <- lm(quality ~ alcohol + sulphates + residual.sugar +
  type + free.sulfur.dioxide + density +
  chlorides + pH + fixed.acidity, data = wine)

model6 <- lm(quality ~ alcohol + density + total.sulfur.dioxide +
  chlorides + pH, data = wine)

model7 <- lm(quality ~ alcohol + volatile.acidity + sulphates + residual.sugar +
  type + free.sulfur.dioxide + density + total.sulfur.dioxide +
  chlorides + pH + fixed.acidity, data = wine)

model8 <- lm(quality ~ sulphates + residual.sugar + type + free.sulfur.dioxide +
  chlorides + pH + fixed.acidity, data = wine)

model9 <- lm(quality ~ volatile.acidity + sulphates +
  type + free.sulfur.dioxide + density +
  pH + fixed.acidity, data = wine)

# Tabla con los coeficientes de determinación de cada modelo
tabla.coeficientes <- matrix(c(0, summary(model0)$r.squared,
  1, summary(model1)$r.squared,
  2, summary(model2)$r.squared,
  3, summary(model3)$r.squared,
  4, summary(model4)$r.squared,
  5, summary(model5)$r.squared,
  7, summary(model6)$r.squared,
  8, summary(model7)$r.squared,
  9, summary(model8)$r.squared,
  10, summary(model9)$r.squared),
  ncol = 2, byrow = TRUE)

colnames(tabla.coeficientes) <- c("Model n°", "R^2")
tabla.coeficientes

```

```

##      Model n°      R^2
## [1,]      0 0.29964342
## [2,]      1 0.29551160
## [3,]      2 0.29959308
## [4,]      3 0.26332681
## [5,]      4 0.26358242
## [6,]      5 0.25449569
## [7,]      7 0.21031429
## [8,]      8 0.29959308
## [9,]      9 0.06867407
## [10,]     10 0.18181245

```

Modelo

LES MILLORS VARIABLES PER PREDIR LA QUALITAT DEL VI SÓN.

El model 0 és el que millors resultats d'ajust aporta. Per tant, si es reduïx més el nombre de variables es perd capacitat de predicció en model. Les variables significatives són:

alcohol + volatile.acidity + sulphates + residual.sugar + citric.acid + type + free.sulfur.dioxide + density + total.sulfur.dioxide + chlorides + pH + fixed.acidity

Després d'haver extret les variables del conjunt s'observa si aquestes variables es mantenen en el cas d'analitzar

el vi negre i el vi blanc per serpat.

```
viBlanc <- wine[wine['type'] == 'White',]  
viNegre <- wine[wine['type'] == 'Red',]
```

S'analitza les variables significatives del vi blanc.

```
empty.model <- lm(quality ~ 1, data=viBlanc)  
horizonte <- formula(quality ~ alcohol + volatile.acidity + sulphates + residual.sugar +  
  citric.acid + free.sulfur.dioxide + density + total.sulfur.dioxide +  
  chlorides + pH + fixed.acidity)  
  
modelViBlanc <- stepAIC(empty.model, trace=FALSE, direction="both", scope=horizonte)  
modelViBlanc$anova
```

```
## Stepwise Model Path  
## Analysis of Deviance Table  
##  
## Initial Model:  
## quality ~ 1  
##  
## Final Model:  
## quality ~ alcohol + volatile.acidity + residual.sugar + free.sulfur.dioxide +  
## density + pH + sulphates + fixed.acidity  
##  
##  
##
```

		Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
##	1				4891	3832.632	-1191.890
##	2	+ alcohol	1	730.20799	4890	3102.424	-2223.903
##	3	+ volatile.acidity	1	196.00140	4889	2906.422	-2541.158
##	4	+ residual.sugar	1	67.71880	4888	2838.704	-2654.489
##	5	+ free.sulfur.dioxide	1	29.00377	4887	2809.700	-2702.729
##	6	+ density	1	22.10980	4886	2787.590	-2739.377
##	7	+ pH	1	25.38397	4885	2762.206	-2782.128
##	8	+ sulphates	1	20.90722	4884	2741.299	-2817.297
##	9	+ fixed.acidity	1	16.21458	4883	2725.084	-2844.318

S'analitzen les variables del vi negre

```
empty.model <- lm(quality ~ 1, data=viNegre)  
horizonte <- formula(quality ~ alcohol + volatile.acidity + sulphates + residual.sugar +  
  citric.acid + free.sulfur.dioxide + density + total.sulfur.dioxide +  
  chlorides + pH + fixed.acidity)  
  
modelViNegre <- stepAIC(empty.model, trace=FALSE, direction="both", scope=horizonte)  
modelViNegre$anova
```

```
## Stepwise Model Path  
## Analysis of Deviance Table  
##  
## Initial Model:  
## quality ~ 1  
##  
## Final Model:  
## quality ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide +  
## chlorides + pH + free.sulfur.dioxide  
##
```

```
##
##           Step Df   Deviance Resid. Df Resid. Dev       AIC
## 1
## 2      + alcohol  1 236.294646    1597   805.8705 -1091.6519
## 3    + volatile.acidity  1  94.074227    1596   711.7962 -1288.1379
## 4      + sulphates  1  19.691588    1595   692.1046 -1330.9971
## 5 + total.sulfur.dioxide  1   8.217587    1594   683.8871 -1348.0961
## 6      + chlorides  1   8.036982    1593   675.8501 -1364.9988
## 7      + pH       1   5.918882    1592   669.9312 -1377.0640
## 8 + free.sulfur.dioxide  1   2.394132    1591   667.5371 -1380.7885
```

S'observa que les variables principals del vi blanc són:

- alcohol
- volatile.acidity
- residual.sugar
- free.sulfur.dioxide
- density
- pH
- sulphates
- fixed.acidity

S'observa que les variables principals del vi negre són:

- alcohol
- volatile.acidity
- free.sulfur.dioxide
- pH
- sulphates
- total.sulfur.dioxide
- chlorides

S'observen que les variables *residual.sugar*, *density* i *fixed.acidity* són úniques pel vi blanc. Tanmateix, les variables *total.sulfur.dioxide* i *chlorides* són úniques pel vi negre.

Es generen els models de regressió lineal pel vi blanc i pel vi negre amb les variables significatives per cada un.

```
lmViBlanc <- lm(quality ~ alcohol + volatile.acidity + residual.sugar + free.sulfur.dioxide +
  density + pH + sulphates + fixed.acidity, data = viBlanc)
summary(lmViBlanc)
```

```
##
## Call:
## lm(formula = quality ~ alcohol + volatile.acidity + residual.sugar +
##     free.sulfur.dioxide + density + pH + sulphates + fixed.acidity,
##     data = viBlanc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.5962 -0.4979 -0.0412  0.4652  3.1172
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.219e+02  2.189e+01  10.139 < 2e-16 ***
## alcohol         1.073e-01  2.886e-02   3.718 0.000203 ***
## volatile.acidity -1.860e+00  1.090e-01 -17.066 < 2e-16 ***
## residual.sugar    1.038e-01  8.286e-03  12.525 < 2e-16 ***
## free.sulfur.dioxide 4.421e-03  6.924e-04   6.385 1.87e-10 ***
```

```
## density          -2.228e+02  2.216e+01 -10.057 < 2e-16 ***
## pH               9.040e-01  1.095e-01  8.254 < 2e-16 ***
## sulphates        7.249e-01  1.009e-01  7.182 7.92e-13 ***
## fixed.acidity     1.208e-01  2.241e-02  5.390 7.37e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.747 on 4883 degrees of freedom
## Multiple R-squared:  0.289, Adjusted R-squared:  0.2878
## F-statistic: 248.1 on 8 and 4883 DF,  p-value: < 2.2e-16
```

```
lmViNegre <- lm(quality ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide +
  chlorides + pH + free.sulfur.dioxide, data = viNegre)
summary(lmViNegre)
```

```
##
## Call:
## lm(formula = quality ~ alcohol + volatile.acidity + sulphates +
##     total.sulfur.dioxide + chlorides + pH + free.sulfur.dioxide,
##     data = viNegre)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.68918 -0.36757 -0.04653  0.46081  2.02954
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.4300987   0.4029168   10.995 < 2e-16 ***
## alcohol         0.2893028   0.0167958   17.225 < 2e-16 ***
## volatile.acidity -1.0127527   0.1008429  -10.043 < 2e-16 ***
## sulphates       0.8826651   0.1099084    8.031 1.86e-15 ***
## total.sulfur.dioxide -0.0034822  0.0006868  -5.070 4.43e-07 ***
## chlorides      -2.0178138   0.3975417  -5.076 4.31e-07 ***
## pH             -0.4826614   0.1175581   -4.106 4.23e-05 ***
## free.sulfur.dioxide  0.0050774   0.0021255    2.389  0.017 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6477 on 1591 degrees of freedom
## Multiple R-squared:  0.3595, Adjusted R-squared:  0.3567
## F-statistic: 127.6 on 7 and 1591 DF,  p-value: < 2.2e-16
```

S'observa que l'ajust del vi blanc és de 0.289 i l'ajust del vi negre és de 0.3595. És a dir, en el conjunt del vi negre mostra una millor adaptabilitat a la regressió lineal múltiple.

Finalment, es genera una regressió lineal de tot el conjunt de dades amb les variables fisicoquímiques rellevants per ambdues tipologies de vi.

```
reg5 <- lm(quality ~ alcohol + volatile.acidity + sulphates + pH + free.sulfur.dioxide, data = wine)
summary(reg5)
```

```
##
## Call:
## lm(formula = quality ~ alcohol + volatile.acidity + sulphates +
##     pH + free.sulfur.dioxide, data = wine)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.3589 -0.4686 -0.0398  0.4690  3.1761
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.9703154  0.2021574   9.746 < 2e-16 ***
## alcohol        0.3279212  0.0079892  41.045 < 2e-16 ***
## volatile.acidity -1.3331537  0.0627409 -21.249 < 2e-16 ***
## sulphates      0.6354432  0.0647409   9.815 < 2e-16 ***
## pH             0.1244165  0.0606014   2.053  0.0401 *
## free.sulfur.dioxide 0.0040255  0.0005832   6.903 5.58e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.743 on 6485 degrees of freedom
## Multiple R-squared:  0.2761, Adjusted R-squared:  0.2755
## F-statistic: 494.7 on 5 and 6485 DF,  p-value: < 2.2e-16
```

S'observa que si sols es recullen les variables significatives que comparteixen els conjunts de vi blanc i vi negre s'obté un resultat pitjor que en el cas de fer l'anàlisi amb el conjunt unit en primera instància.

Una altra metodologia per obtenir les variables més rellevants d'un conjunt és realitzar un arbre de decisions. A continuació s'explora aquesta altra metodologia. S

```
# Separem el conjunt
y <- wine$quality
x <- wine
x$quality <- NULL
x$quality_d <- NULL

# Arbre
y = as.factor(y)
arbre1 <- C50::C5.0(x, y, rules=FALSE )
```

S'observa que les variables més usades són *volatile.acidity*, *alcohol*, *fixed.acidity*, *free.sulfur.dioxide* i *residual.sugar*

```
predicted_model <- predict(arbre1, x, type="class" )
print(sprintf("La precisió de l'arbre en el conjunt d'entrenament és: %.4f %%", 100*sum(predicted_model == y)))
```

```
## [1] "La precisió de l'arbre en el conjunt d'entrenament és: 82.6375 %"
```

S'observa una bona precisió en el del model pel conjunt d'entrenament. Ara es diferenciarà els models pel tipus de vi.

Es genera l'arbre de decisió pel vi blanc

```
# Separem el conjunt
y <- viBlanc$quality
x <- viBlanc
x$quality <- NULL
x$quality_d <- NULL
x$type <- NULL

# Arbre
y = as.factor(y)
arbreViBlanc <- C50::C5.0(x, y, rules=FALSE )
```

S'observa que les variables més usades són *alcohol*, *volatile.acidity*, *free.sulfur.dioxide* i *fixed.acidity*

```
predicted_model <- predict(arbreViBlanc, x, type="class" )
print(sprintf("La precisió de l'arbre en el conjunt d'entrenament és: %.4f %%",100*sum(predicted_model
```

```
## [1] "La precisió de l'arbre en el conjunt d'entrenament és: 88.9207 %"
```

Es realitza l'arbre de decisió pel vi negre.

```
# Separem el conjunt
y <- viNegre$quality
x <- viNegre
x$quality <- NULL
x$quality_d <- NULL
x$type <- NULL

# Arbre
y = as.factor(y)
arbreViNegre <- C5.0(x, y, rules=FALSE )
```

S'observa que les variables més usades són *alcohol*, *total.sulfur.dioxide*, *sulphates*, *pH* i *volatile.acidity*.

```
predicted_model <- predict(arbreViNegre, x, type="class" )
print(sprintf("La precisió de l'arbre en el conjunt d'entrenament és: %.4f %%",100*sum(predicted_model
```

```
## [1] "La precisió de l'arbre en el conjunt d'entrenament és: 90.8068 %"
```

Les variables rellevants pel vi blanc i el vi negre són significativament diferents. Per últim, es realitza un arbre de decisions amb les variables rellevants de vi blanc i vi negre.

```
# Separem el conjunt
y <- wine$quality
x <- wine[,c('alcohol','volatile.acidity','sulphates','free.sulfur.dioxide','pH','fixed.acidity')]

# Arbre
y = as.factor(y)
arbre2 <- C5.0(x, y)
```

```
predicted_model <- predict(arbre2, x, type="class" )
print(sprintf("La precisió de l'arbre en el conjunt d'entrenament és: %.4f %%",100*sum(predicted_model
```

```
## [1] "La precisió de l'arbre en el conjunt d'entrenament és: 81.5745 %"
```

A continuació es realitza el model logístic per predir si és bo o no bo un vi.

```
m1 <- glm(quality_d ~ alcohol + volatile.acidity + sulphates + residual.sugar +
  citric.acid + type + free.sulfur.dioxide + density + total.sulfur.dioxide +
  chlorides + pH + fixed.acidity, family = binomial, data= wine)
summary(m1)
```

```
##
## Call:
## glm(formula = quality_d ~ alcohol + volatile.acidity + sulphates +
##     residual.sugar + citric.acid + type + free.sulfur.dioxide +
##     density + total.sulfur.dioxide + chlorides + pH + fixed.acidity,
##     family = binomial, data = wine)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
```



```
## -2.8167 -0.6280 -0.3683 -0.1761 3.0593
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    4.036e+02  6.600e+01  6.115 9.68e-10 ***
## alcohol        4.546e-01  8.096e-02  5.615 1.97e-08 ***
## volatile.acidity -3.617e+00  3.888e-01 -9.303 < 2e-16 ***
## sulphates      2.449e+00  2.855e-01  8.578 < 2e-16 ***
## residual.sugar  2.188e-01  2.631e-02  8.317 < 2e-16 ***
## citric.acid    -1.228e-01  3.566e-01 -0.344 0.730584
## typeWhite      -7.967e-01  2.454e-01 -3.246 0.001170 **
## free.sulfur.dioxide 1.165e-02  3.048e-03  3.821 0.000133 ***
## density        -4.245e+02  6.689e+01 -6.347 2.19e-10 ***
## total.sulfur.dioxide -3.641e-03  1.341e-03 -2.714 0.006644 **
## chlorides      -7.730e+00  2.503e+00 -3.088 0.002012 **
## pH             2.601e+00  3.616e-01  7.192 6.37e-13 ***
## fixed.acidity   4.906e-01  6.724e-02  7.295 2.98e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 6437.0 on 6490 degrees of freedom
## Residual deviance: 5073.9 on 6478 degrees of freedom
## AIC: 5099.9
##
## Number of Fisher Scoring iterations: 6
```

S'eliminen les variables no significatives,

```
sig.var<- summary(m1)$coeff[-1,4] >0.05
names(sig.var)[sig.var == TRUE]
```

```
## [1] "citric.acid"
```

Es torna a generar el model sense la varianle *crtric.acid* perquè no és significativa.

```
m2 <- glm(quality_d ~ alcohol + volatile.acidity + sulphates + residual.sugar +
  type + free.sulfur.dioxide + density + total.sulfur.dioxide +
  chlorides + pH + fixed.acidity, family = binomial, data= wine)
summary(m2)
```

```
##
## Call:
## glm(formula = quality_d ~ alcohol + volatile.acidity + sulphates +
##     residual.sugar + type + free.sulfur.dioxide + density + total.sulfur.dioxide +
##     chlorides + pH + fixed.acidity, family = binomial, data = wine)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.8093  -0.6284  -0.3690  -0.1760   3.0597
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    4.050e+02  6.585e+01  6.151 7.72e-10 ***
## alcohol        4.506e-01  8.013e-02  5.624 1.87e-08 ***
```

```
## volatile.acidity      -3.573e+00  3.670e-01  -9.737  < 2e-16 ***
## sulphates            2.445e+00  2.852e-01   8.571  < 2e-16 ***
## residual.sugar       2.190e-01  2.630e-02   8.328  < 2e-16 ***
## typeWhite            -8.026e-01  2.448e-01  -3.278  0.001044 **
## free.sulfur.dioxide   1.168e-02  3.047e-03   3.834  0.000126 ***
## density              -4.260e+02  6.674e+01  -6.383  1.74e-10 ***
## total.sulfur.dioxide -3.680e-03  1.337e-03  -2.752  0.005920 **
## chlorides            -7.795e+00  2.494e+00  -3.126  0.001774 **
## pH                   2.606e+00  3.612e-01   7.214  5.43e-13 ***
## fixed.acidity        4.856e-01  6.568e-02   7.395  1.42e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 6437.0  on 6490  degrees of freedom
## Residual deviance: 5074.1  on 6479  degrees of freedom
## AIC: 5098.1
##
## Number of Fisher Scoring iterations: 6
```

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

5. Resultats

Al llarg del projecte s'han mostrat les gràfiques i els resultats obtinguts dels diversos tractaments de joc de dades.

6. Conclusions

En aquest treball es pretén respondre a les següents preguntes. Quines són les propietats fisicoquímiques que fan que un vi sigui bo? Les propietats que fan que un vi sigui bo són les mateixes per vins negres i blancs?

Les propietats de fisicoquímiques que fan que un bo sigui bo o no, és a dir, que tingui una alta qualificació, són diferents el vi blanc i el vi negre. No obstant això, hi ha propietats comunes que són rellevants per ambdues tipologies de vins. Les propietats fisicoquímiques significatives són: l'alcohol, la volatilitat de l'àcid, el sulfur de diòxid lliure, el pH i els sulfats. Aquestes variables significatives s'han obtingut independentment per metodologies de regressions múltiples lineals i per metodologies d'arbres de decisions i ambdues han aportat les mateixes variables rellevants del conjunt.