RM— title: 'Pràctica 2 - Tipologia i cicle de vida de les dades' author: "Marta Puigdemasa i Antoni Garcia" date: "Gener 2021" output: word_document: default pdf_document: latex_engine: xelatex —

1.Descripció del dataset. Perquè és important i quina pregunta/problema pretén respondre?

Per a la **Pràctica 1** *Comparativa de l'evolució de la incidència de la Covid19-Sars2 a l'Estat Espanyol respecte del conjunt d'Europa entre el 20 d'octubre i el 4 de novembre de 2020* vam dur a terme un programa de *web scrapping* que ens permetia recollir dades sanitàries de forma diària. Tanmateix, com que aquestes dades han resultat ser molt estables pel que fa a errors o valors nuls ja que són definides i curades (*curated*) per ens governamentals. Conseqüentment, el data set que vam elaborar no és del tot adient per a fer anàlisis relacionades amb tasques rutinàries de processament, neteja o comparació de dades.

Conseqüentment, hem decidit seleccionar el conjunt de dades **Winequality-red.csv** degut a què: (i) presenta unes dimensions considerables que permeten fer una aproximació a casos reals de *Big Data*, (ii) és un dels exemples per antonomàsia d'estudis de mercat en què es correlaciona paràmetres quantitatius descriptius amb resultats qualitatius comercials, en aquest cas *propietats organolèptiques* amb *avaluació qualitativa*, (iii) conté valors problemàtics que cal tractar i (iv) permet plantejar contrasts d'hipòtesis per a tests estadístics suficientment robustos.

Un cop decidit el conjunt de dades i en havent analitzat i corregit els atributs **les dues principals preguntes a contestar seran:**

1. Quins dels atributs tenen més influència amb la qualitat del vi?

2, A partir dels valors dels atributs, és possible predir la qualitat del vi?

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

En primer lloc, carregarem el fitxer *winequality-red.csv* com al dataframe **vins_rojos** i en farem una anàlisi estructural amb les funcions *dim()* i *str()*.

Com es pot comprovar a la **Taula 1.1**, el fitxer seleccionat consta de 1599 entrades per a 12 variables (matriu 1599 x 12), que corresponen a 12 tipus numèrics (*num* o *int*) i representen les següents propietats organolèptiques (**Taula 1.2**):

1 - fixed acidity => acidesa deguda a àcids no-volàtils.

2 - volatile acidity=> contingut en àcids volàtils com l'àcid acètic (incida **amargor**)

3 - citric acid => contingut en àcids volàtils com l'àcid acètic (incida **frescor**)

4 - residual sugar => contingut residual en sucres.

5 - chlorides => contingut residual en clòrid.

6 - free sulfur dioxide => contingut en SO2.

7 - total sulfur dioxide => contingut total en SO2.

8 - density => densitat del vi (depèn dels sucres i del grau d'alcohol).

9 - pH => mesura de l'acidesa.

10 - sulphates => contingut en sulfats, emprats per a aturar les fermentacions

11 - alcohol=> contingut en etanol.

12 - quality: nota rebuda pels someliers en base a la qualitat del vi.

```
# IMPORTACIÓ DE DADES
vins_rojos=read.csv("winequality-red.csv", sep=",", na.strings = NA)
print("Taula 1.1.Les dimensions inicials de la matriu vins_rojos són:")

## [1] "Taula 1.1.Les dimensions inicials de la matriu vins_rojos són:"

dim(vins_rojos)

## [1] 1599   12

print("Taula 1.2.Les variables inicials de la matrius vins_rojos són:")

## [1] "Taula 1.2.Les variables inicials de la matrius vins_rojos són:"

str(vins_rojos)

## '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 0.5
8 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 0.06
9 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 ...
```

Òbviament, *quality* no és una propietat, sinó un resultat atribuït, i per tant caldrà factoritzar-lo mitjançant la funció *as.factor()*.

```
vins_rojos$quality=as.factor(vins_rojos$quality)
```

Per a fer una primera aproximació al tipus de dades incloses, farem un resum amb al funció *summary()*, que ens permet recuperar els valors mínims/màxims, el quartils i paràmetres de centralitat (mitjana i mediana) (**Taula 2.1**). Aquesta funció és molt útil per a detectar **valors nuls** ja que també recompta **NAs**. Tanmateix, tot i que aquest no és el nostre cas, sí que podem observar que els mínim per la mesura d'*àcid cítric* és 0. Açò és un poc estrany, ja que bioquímicament la fermentació alcohòlica del llevat involucra reaccions del cicle de l'àcid cítric i aquest n'esdevé un subproducte.

A més, podem fer-nos una idea de valors que, per a determinades variables, mostren una gran desviació respecte de la mitjana. Per exemple, per a *sucre residual* la **mitjana = 2.5** però els **valors mínims=0.9** i el **màxim=15.5**.

Finalment, veiem que en el dataframe predominen vins *mediocres* amb notes de 5 i 6.

```
# ANÀLISI DE VALORS
print("Taula 2.1.Estadístiques descriptives dels valors de cada variable
al dataset vins_rojos")
```

```
## [1] "Taula 2.1.Estadístiques descriptives dels valors de cada variable
al dataset vins_rojos"
```

```
summary(vins_rojos)
```

```
##  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   3: 10
##  1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50   4: 53
##  Median :3.310   Median :0.6200   Median :10.20   5:681
##  Mean   :3.311   Mean   :0.6581   Mean   :10.42   6:638
```

```
##  3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10   7:199
##  Max.   :4.010   Max.   :2.0000   Max.   :14.90   8: 18

vins_rojos$quality=as.factor(vins_rojos$quality)
```

3. Neteja de les dades.

En base a allò que hem comentat en el punt anterior caldrà fer un procés de neteja de dades acumulades per a tractar **valors nuls i extrems** abans de poder fer-ne una anàlisi més exhaustiva.

3.1. Les dades contenen zeros o elements buits? Com gestionaries aquests casos?

En primer lloc verifiquem amb *is.na()* que no existeixen valors *NAs* en tot el joc de dades (**Taula 3.1**).

A continuació iterem cada columna del data set mitjançant un *for loop()* que recorre els valors, quantifica en quants casos es dona un valor zero (0) o buit (nul) i guarda el resultat dintre del dataframe **avaluacio_nuls**.

Com a resultat podem apreciar a la **Taula 3.2** que tan sols per al paràmetre *àcid cítric* hem trobat valors nuls, concretament **132 valors 0**, que haurem de corregir.

```
# ANÀLISI DE VALORS NULS
# Comprovació absència NAs
print("Taula 3.1. Nombre de valors NA en el joc de dades incial:")

## [1] "Taula 3.1. Nombre de valors NA en el joc de dades incial:"

table(is.na(vins_rojos))

##
## FALSE
## 19188

# Quantificació valors 0 i buits:
# Dataframe ontenidor:
avaluacio_nuls=NULL
# Iteració sobre totes les columnes:
for (i in seq(1:length(vins_rojos))){
  # comprovació quants valor 0 hi ha:
  zero=nrow(subset(vins_rojos, vins_rojos[,i]==0))
  # comprovació quants valor nuls hi ha:
  null=nrow(subset(vins_rojos, vins_rojos[,i]==""))
  # guardem resultat
  avaluacio_nuls=cbind(avaluacio_nuls,c(zero,null))
}
# Definició nom files i columnes
colnames(avaluacio_nuls)=colnames(vins_rojos)
row.names(avaluacio_nuls)=c("zeros","buits")
```

```
# Visualització resultats
print("Taula 3.2. Nombre de valors nuls (0) i buits (' ') en el joc de da
des incial:")
```

```
## [1] "Taula 3.2. Nombre de valors nuls (0) i buits (' ') en el joc de d
ades incial:"
```

```
avaluacio_nuls
```

```
##       fixed.acidity volatile.acidity citric.acid residual.sugar chlori
des
## zeros             0                0         132              0
0
## buits             0                0           0              0
0
##       free.sulfur.dioxide total.sulfur.dioxide density pH sulphates al
cohol
## zeros                   0                    0       0  0         0
0
## buits                   0                    0       0  0         0
0
##       quality
## zeros       0
## buits       0
```

L'algorisme dels **k-veïns més pròxims** (*k-nearest neighbours*) és el mecanisme més hegemònic en Ciència de Dades per a imputar valors perduts. Nosaltres emprarem la funció *kNN()* del paquet *VIM*, tot i que en el nostre cas, per a guanyar precisió, mirarem la distribució dels valors 0 per *quality*. Amb aquesta finalitat representarem mitjançant un histograma la freqüència de 0 per *quality* i en calcularem el percentatge.

La **Figura 1** i la **Taula 4** mostren que els vins de categoria 3 i 4 contenen un 30.0% i un 18.9% de casos de valors nuls, per tant, enlloc d'imputar els valors perduts utilitzant tot el dataset, convé imputar per tipus de *quality* ja que així es consideraran aquells veïns més "similars" i no el total, fet que podria introduir alguna mena de biaix.

Amb aquesta aproximació aconseguim un nou dataset, **vins**, que conté els valors nuls imputats.

```
# Creem una matriu on hi ha el valors 0
citric_zeros = subset(vins_rojos, vins_rojos$citric.acid==0)
# En representem la distribució:
plot(citric_zeros$quality, main="Figura 1 Distribució valors 0 per a àcid
cítric per categoria de vi", xlab="categoria", ylab= "freqüència")
# Càlcul percentate valors nuls per categoria:
print("Taula 4. Percentatge de valors 0 per a àcid cítric en funció de la
qualitat del vi")
```

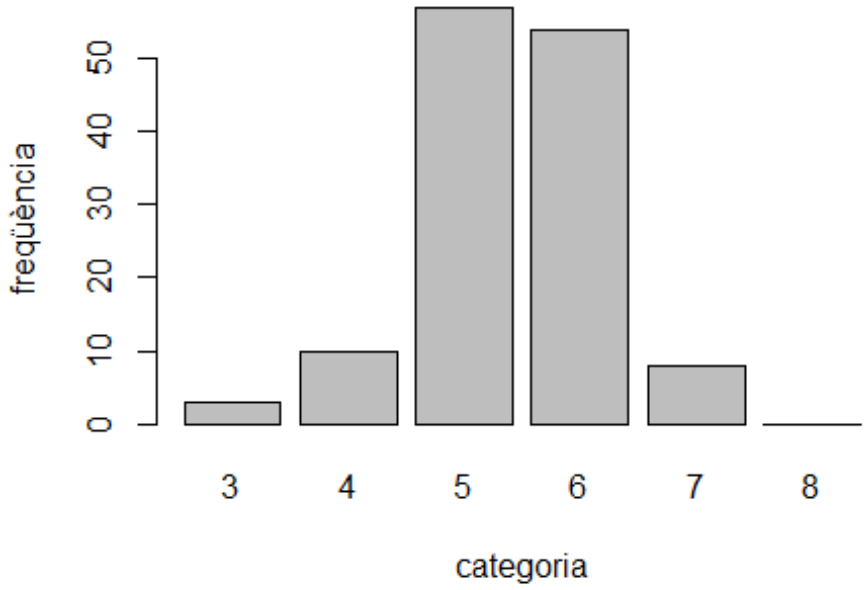
```
## [1] "Taula 4. Percentatge de valors 0 per a àcid cítric en funció de l
a qualitat del vi"
```

```
perc_citrics_zeros=round(100*table(citric_zeros$quality)/table(vins_rojos
$quality), 1)
perc_citrics_zeros

##
##    3    4    5    6    7    8
## 30.0 18.9  8.4  8.5  4.0  0.0

#  Carreguem llibreries:
if(!require(VIM)) install.packages('VIM'); library(VIM)
```

ura 1 Distribució valors 0 per a àcid cítric per catego




```
# Guardem en quines posicions són els valors 0:
valors_remplacar = which(vins_rojos$citric.acid==0)
# Convertim els 0 en NAs
for (i in valors_remplacar){
  vins_rojos[i, 3]=NA
}
# Creem un nou dataframe on guardar les dades amb imputacions:
vins=NULL
# Iterem per a cada quality:
for (i in (3:8)){
  # Delimitem cada grup:
  grup=subset(vins_rojos, vins_rojos$quality==i)
  # Fem les imputacions amb kNN():
  grup_imputat = kNN(grup)
  # Eliminem del columnes informatives
  grup_final= grup_imputat[,1:12]
```

```
  # Afegim el resultat al nou dataframe.
  vins=rbind(vins, grup_final)
}
# Comprovació
prova_zeros= subset(vins, vins$citric.acid==0)
prova_zeros
```

```
##  [1] fixed.acidity        volatile.acidity     citric.acid
##  [4] residual.sugar       chlorides            free.sulfur.dioxide
##  [7] total.sulfur.dioxide density              pH
## [10] sulphates            alcohol              quality
## <0 rows> (or 0-length row.names)
```

3.2.Identificació i tractament de valors extrems.

Per altra banda, hem vist a la **Taula 2.1** que hi ha valors màxims que poden desviar-se més de tres mitjanes dels valors centrals. Sabem que la representació dels valors mitjançant diagrames de caixes amb la funció *boxplot()* ens dóna accés no només a aquest tipus de gràfic, sinó a explorar els valors extrems (*outliers*) amb l'afegit *$out*.

Prepararem un *for loop()* que investiga per a cada tipus de qualitat de vi quins són els valors extrems a cada columna i els substitueix, en aquest cas, per la mediana de cada grup. Aquest procediment també és una **mesura robusta** de centralitat.

Podem comprovar que la primera ronda de diagrames de caixes contenen valors extrems (**Figures 2.1 a 7.11**), mentre que la versió corregida **vins2** origina diagrames molt més nets (**Figures 8.1 a 13.11**).

```
# ANÀLISI DE VALORS EXTREMS
# Variable on guardar els nous valors
vins2=NULL
# Fixem quants gràfics per pantalla volem obtindre (3x2)
par(mfrow=c(3,2))
# Iterem per a cadascuna de les qualitats
for (i in (3:8)){
  # Separem per qualitat
  grup=subset(vins, vins$quality==i)
  # Dins de cada qualitat recorrem totes les columnes amb variables:
  for (j in (1:11)){
    # Representem els diagrames de caixes:
    box_plot=boxplot(grup[,j], main=paste("Figura", i-1, ".", j, ".", col
names(grup)[j], " vi qualitat", i), col="lightblue")
    # Guardem els outliers en una llista:
    outliers=box_plot$out
    # Calculem la mediana per a cada columna:
    med=median(grup[,j])
    for (k in outliers){
      # Cerquem per a cada columna els valors extrems i els substituïm pe
r la mediana:
      grup[j][grup[j]==k]=med
```

```
    }
  }
  # Actualitzem el dataframe vins2:
  vins2=rbind(vins2, grup)
}
```

Figura 2 . 1 . fixed.acidity vi qualitat

Figura 2 . 2 . volatile.acidity vi qualit

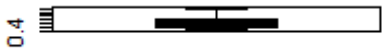


Figura 2 . 3 . citric.acid vi qualitat

Figura 2 . 4 . residual.sugar vi qualit

Figura 2 . 5 . chlorides vi qualitat 3

gura 2 . 6 . free.sulfur.dioxide vi qual

igura 2 . 7 . total.sulfur.dioxide vi qua

Figura 2 . 8 . density vi qualitat 3

Figura 2 . 9 . pH vi qualitat 3

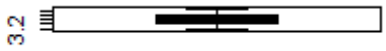


Figura 2 . 10 . sulphates vi qualitat

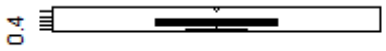


Figura 2 . 11 . alcohol vi qualitat 3

Figura 3 . 1 . fixed.acidity vi qualitat

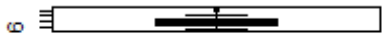


Figura 3 . 2 . volatile.acidity vi qualit

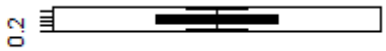


Figura 3 . 3 . citric.acid vi qualitat

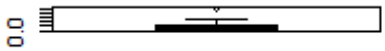


Figura 3 . 4 . residual.sugar vi qualit

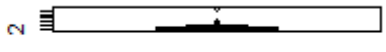
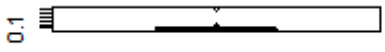


Figura 3 . 5 . chlorides vi qualitat



igura 3 . 6 . free.sulfur.dioxide vi qua

Figura 3 . 7 . total.sulfur.dioxide vi qua

Figura 3 . 8 . density vi qualitat 4

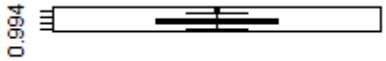


Figura 3 . 9 . pH vi qualitat 4

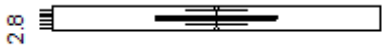


Figura 3 . 10 . sulphates vi qualitat

Figura 3 . 11 . alcohol vi qualitat 4

Figura 4 . 1 . fixed.acidity vi qualitat

Figura 4 . 2 . volatile.acidity vi qualit

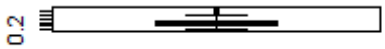


Figura 4 . 3 . citric.acid vi qualitat 5

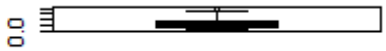


Figura 4 . 4 . residual.sugar vi qualita

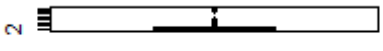
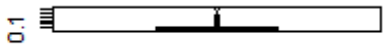
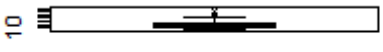


Figura 4 . 5 . chlorides vi qualitat 5



igura 4 . 6 . free.sulfur.dioxide vi qual



igura 4 . 7 . total.sulfur.dioxide vi qua

Figura 4 . 8 . density vi qualitat 5

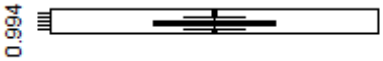


Figura 4 . 9 . pH vi qualitat 5

Figura 4 . 10 . sulphates vi qualitat

Figura 4 . 11 . alcohol vi qualitat 5

Figura 5 . 1 . fixed.acidity vi qualitat

Figura 5 . 2 . volatile.acidity vi qualit

Figura 5 . 3 . citric.acid vi qualitat

Figura 5 . 4 . residual.sugar vi qualita

Figura 5 . 5 . chlorides vi qualitat 6

igura 5 . 6 . free.sulfur.dioxide vi qua

igura 5 . 7 . total.sulfur.dioxide vi qua

Figura 5 . 8 . density vi qualitat 6

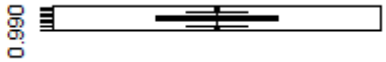


Figura 5 . 9 . pH vi qualitat 6

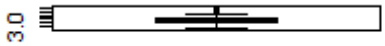


Figura 5 . 10 . sulphates vi qualitat

Figura 5 . 11 . alcohol vi qualitat 6

Figura 6 . 1 . fixed.acidity vi qualitat

Figura 6 . 2 . volatile.acidity vi qualita

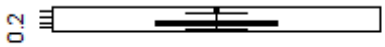


Figura 6 . 3 . citric.acid vi qualitat

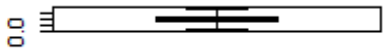


Figura 6 . 4 . residual.sugar vi qualita

Figura 6 . 5 . chlorides vi qualitat 7

Figura 6 . 6 . free.sulfur.dioxide vi qual

igura 6 . 7 . total.sulfur.dioxide vi qua

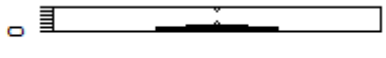


Figura 6 . 8 . density vi qualitat 7

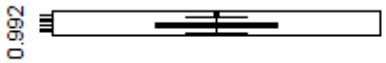


Figura 6 . 9 . pH vi qualitat 7

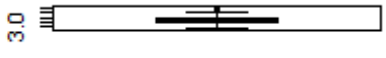


Figura 6 . 10 . sulphates vi qualitat

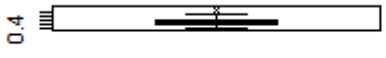


Figura 6 . 11 . alcohol vi qualitat 7

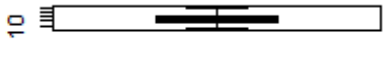


Figura 7 . 1 . fixed.acidity vi qualitat

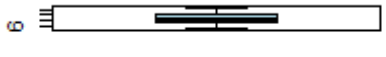


Figura 7 . 2 . volatile.acidity vi qualit

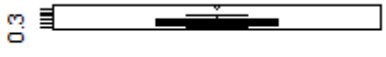


Figura 7 . 3 . citric.acid vi qualitat

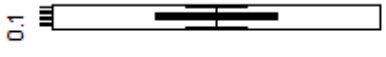
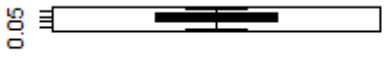
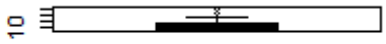


Figura 7 . 4 . residual.sugar vi qualit

Figura 7 . 5 . chlorides vi qualitat 8



igura 7 . 6 . free.sulfur.dioxide vi qua



igura 7 . 7 . total.sulfur.dioxide vi qua

Figura 7 . 8 . density vi qualitat 8

Figura 7 . 9 . pH vi qualitat 8

Figura 7 . 10 . sulphates vi qualitat

Figura 7 . 11 . alcohol vi qualitat 8


```
# Comprovació de què no apareixen valors nuls al nou dataset:
par(mfrow=c(3,2))
for (i in (3:8)){
  grup_categoria=subset(vins2, vins2$quality==i)
  for (j in (1:11)){
    box_plot=boxplot(grup[,j], main=paste("Figura", i+5, ".", j, ".", col
names(grup)[j], " vi qualitat", i), col="red")
  }
}
```

Figura 8 . 1 . fixed.acidity vi qualitat

Figura 8 . 2 . volatile.acidity vi qualit

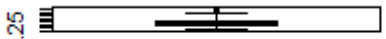


Figura 8 . 3 . citric.acid vi qualitat

Figura 8 . 4 . residual.sugar vi qualit

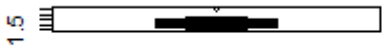


Figura 8 . 5 . chlorides vi qualitat 3

Figura 8 . 6 . free.sulfur.dioxide vi qual

igura 8 . 7 . total.sulfur.dioxide vi qua

Figura 8 . 8 . density vi qualitat 3

Figura 8 . 9 . pH vi qualitat 3

Figura 8 . 10 . sulphates vi qualitat

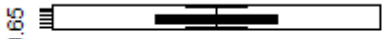


Figura 8 . 11 . alcohol vi qualitat 3

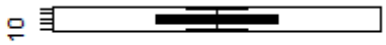


Figura 9 . 1 . fixed.acidity vi qualitat

Figura 9 . 2 . volatile.acidity vi qualit

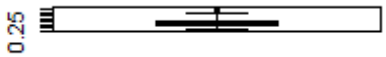


Figura 9 . 3 . citric.acid vi qualitat

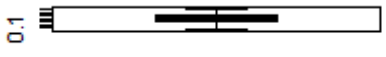


Figura 9 . 4 . residual.sugar vi qualit

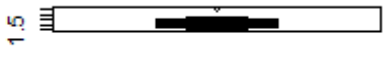
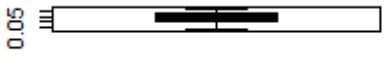


Figura 9 . 5 . chlorides vi qualitat



igura 9 . 6 . free.sulfur.dioxide vi qua

Figura 9 . 7 . total.sulfur.dioxide vi qua

Figura 9 . 8 . density vi qualitat 4

Figura 9 . 9 . pH vi qualitat 4

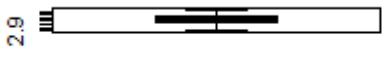


Figura 9 . 10 . sulphates vi qualitat

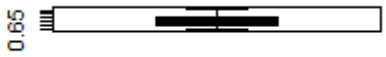


Figura 9 . 11 . alcohol vi qualitat 4

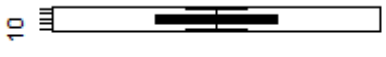


Figura 10 . 1 . fixed.acidity vi qualita

Figura 10 . 2 . volatile.acidity vi qualit

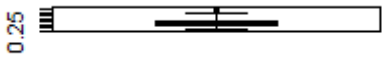


Figura 10 . 3 . citric.acid vi qualitat

Figura 10 . 4 . residual.sugar vi qualit

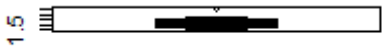
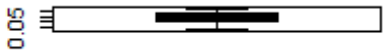


Figura 10 . 5 . chlorides vi qualitat



gura 10 . 6 . free.sulfur.dioxide vi qua

gura 10 . 7 . total.sulfur.dioxide vi qua

Figura 10 . 8 . density vi qualitat 5

Figura 10 . 9 . pH vi qualitat 5

Figura 10 . 10 . sulphates vi qualitat

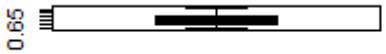


Figura 10 . 11 . alcohol vi qualitat

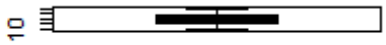


Figura 11 . 1 . fixed.acidity vi qualita

Figura 11 . 2 . volatile.acidity vi qualit

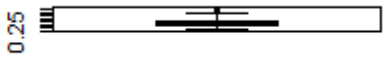


Figura 11 . 3 . citric.acid vi qualitat

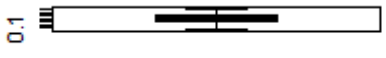


Figura 11 . 4 . residual.sugar vi qualit

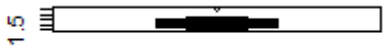
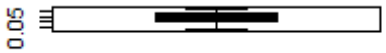


Figura 11 . 5 . chlorides vi qualitat



gura 11 . 6 . free.sulfur.dioxide vi qua

gura 11 . 7 . total.sulfur.dioxide vi qua

Figura 11 . 8 . density vi qualitat 6

Figura 11 . 9 . pH vi qualitat 6

Figura 11 . 10 . sulphates vi qualitat

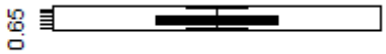


Figura 11 . 11 . alcohol vi qualitat

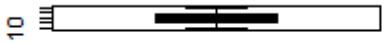


Figura 12 . 1 . fixed.acidity vi qualita

Figura 12 . 2 . volatile.acidity vi qualit

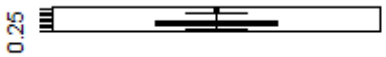


Figura 12 . 3 . citric.acid vi qualitat

Figura 12 . 4 . residual.sugar vi qualit

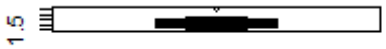
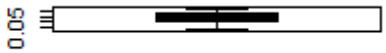


Figura 12 . 5 . chlorides vi qualitat



gura 12 . 6 . free.sulfur.dioxide vi qua

gura 12 . 7 . total.sulfur.dioxide vi qua

Figura 12 . 8 . density vi qualitat 7

Figura 12 . 9 . pH vi qualitat 7

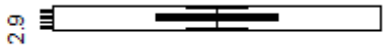


Figura 12 . 10 . sulphates vi qualitat

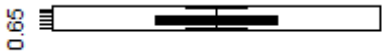


Figura 12 . 11 . alcohol vi qualitat

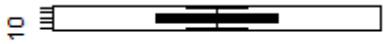


Figura 13 . 1 . fixed.acidity vi qualita

Figura 13 . 2 . volatile.acidity vi qualit

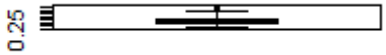


Figura 13 . 3 . citric.acid vi qualitat

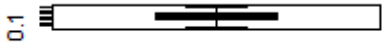


Figura 13 . 4 . residual.sugar vi qualit

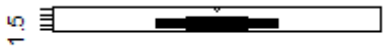
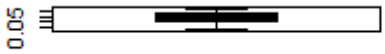


Figura 13 . 5 . chlorides vi qualitat



gura 13 . 6 . free.sulfur.dioxide vi qua gura 13 . 7 . total.sulfur.dioxide vi qua

Figura 13 . 8 . density vi qualitat 8

Figura 13 . 9 . pH vi qualitat 8

Figura 13 . 10 . sulphates vi qualitat

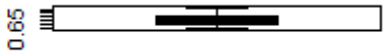
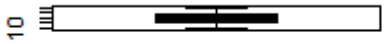


Figura 13 . 11 . alcohol vi qualitat



Per a què consten les correccions introduídes al joc de dades inicials, procedirem a exportar el fitxer **vins2**.

```
# Exportació del fitxer amb dades corregides.
write.csv(vins2, "red_wine_clean.csv", sep=",", quote =F)
```

4.Anàlisi de les dades.

4.1.Selecció dels grups de dades que es volen analitzar/comparar (planificació dels anàlisis a aplicar).

La variable dependent a investigar és la *qualitat del vi*, per tant, necessitarem fer una primera anàlisi estadística descriptiva dels valors mitjans de cada paràmetre per a cada tipus de vi (**Taula 5**).

```
# Preparació d'un data set amb les mitjanes per a cada tipus de qualitat:
vins_qualitat=split(vins2, vins2$quality)
vins_matriu=NULL
for (i in seq(3:8)){
    vi= data.frame(vins_qualitat[i])
    mitjanes=colMeans(vi[,1:11])
    vins_matriu=rbind(vins_matriu,mitjanes)
}
colnames(vins_matriu)=colnames(vins2[1:11])
row.names(vins_matriu)=c("Qualitat 3","Qualitat 4","Qualitat 5","Qualitat
6","Qualitat 7", "Qualitat 8")
```

```
print("Taula 5. Mitjana aritmètica dels paràmetre per a cada qualitat de
vi")

## [1] "Taula 5. Mitjana aritmètica dels paràmetre per a cada qualitat de
vi"

vins_matriu

##            fixed.acidity volatile.acidity citric.acid residual.sugar
chlorides
## Qualitat 3      8.360000        0.8845000   0.1860000       2.635000 0
.10485000
## Qualitat 4      7.522642        0.6939623   0.1839623       2.171698 0
.07686792
## Qualitat 5      7.970778        0.5690308   0.2516446       2.178928 0
.08225698
## Qualitat 6      8.266458        0.4918730   0.2830094       2.174060 0
.07716144
## Qualitat 7      8.775879        0.3918844   0.4029648       2.252261 0
.07442211
## Qualitat 8      8.566667        0.3966667   0.3911111       2.166667 0
.06844444
##            free.sulfur.dioxide total.sulfur.dioxide   density       pH
## Qualitat 3            8.200000             24.90000 0.9974640 3.398000
## Qualitat 4           10.830189             32.84906 0.9963868 3.383396
## Qualitat 5           16.091043             56.51395 0.9970688 3.304890
## Qualitat 6           14.963950             38.78370 0.9966448 3.316897
## Qualitat 7           12.608040             28.69347 0.9960256 3.281558
## Qualitat 8            9.888889             33.44444 0.9952122 3.240000
##            sulphates   alcohol
## Qualitat 3 0.5385000  9.955000
## Qualitat 4 0.5432075 10.265094
## Qualitat 5 0.5838032  9.795081
## Qualitat 6 0.6592163 10.602717
## Qualitat 7 0.7342211 11.465913
## Qualitat 8 0.7477778 12.094444
```

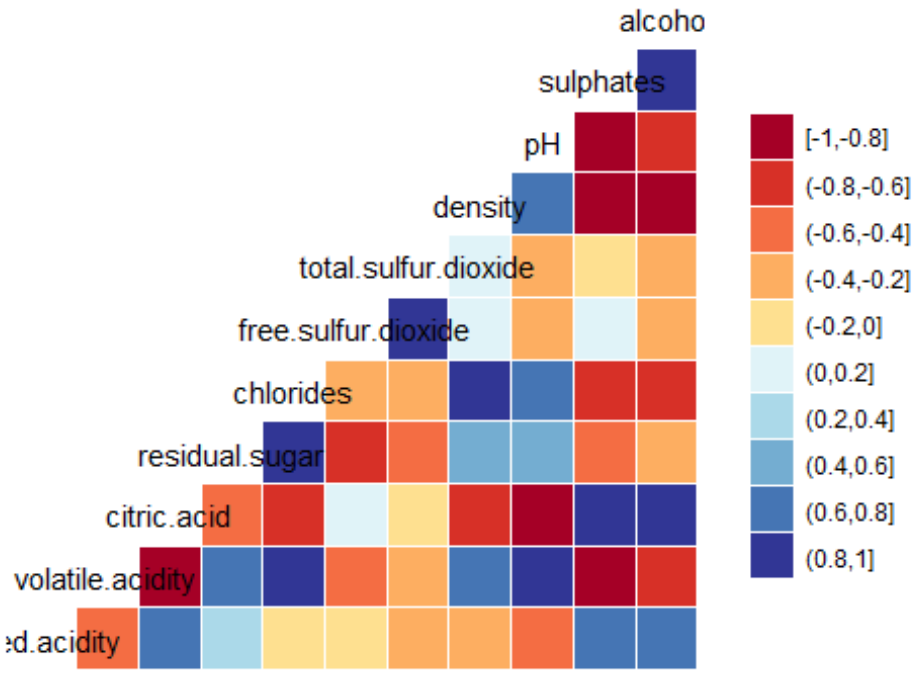
Amb aquestes dades podem explorar com es correlacionen les variables entre sí mitjançant la representació gràfica disponible amb la funció *ggcorr()* del paquet *GGaly*. Tal i com reflecteix la **Figura 14**, es donen una sèrie de correlacions positives i negatives altes (coeficient de Pearson r >= 0.6 en valor absolut), les més rellevants les descriurem a continuació:

- A mesura que el contingut en *alcohol* creix, també ho fa el nivell de *sulfats*, però en disminueix la *densitat*.
- Els *sulfats* s'anticorrelacionen amb el *pH* del vi i l' *acidesa volàtil* però presenta la mateixa tendència que el contingut d'*àcid cítric*.
- La *densitat* creix de la mateixa manera que creix el contingut en *clorits*.

Aquest gràfic de correlacions indica que si som capaços de poder modificar algun dels paràmetres per a millorar/empitjorar-ne la qualitat, açò tindrà repercussions en altres propietats organolèptiques.

```
# Importe la llibreria GGally:
if (!require('GGally')) install.packages('GGally'); library('GGally')
# Correlació per parells de variables mitjançant la representació gràfica
de la funció ggcorr():
ggcorr(vins_matriu, nbreaks = 10, palette = "RdYlBu", geom = "tile") +
  labs(title = "Figura 14. Correlació per parells de variables")
```

Figura 14. Correlació per parells de variables



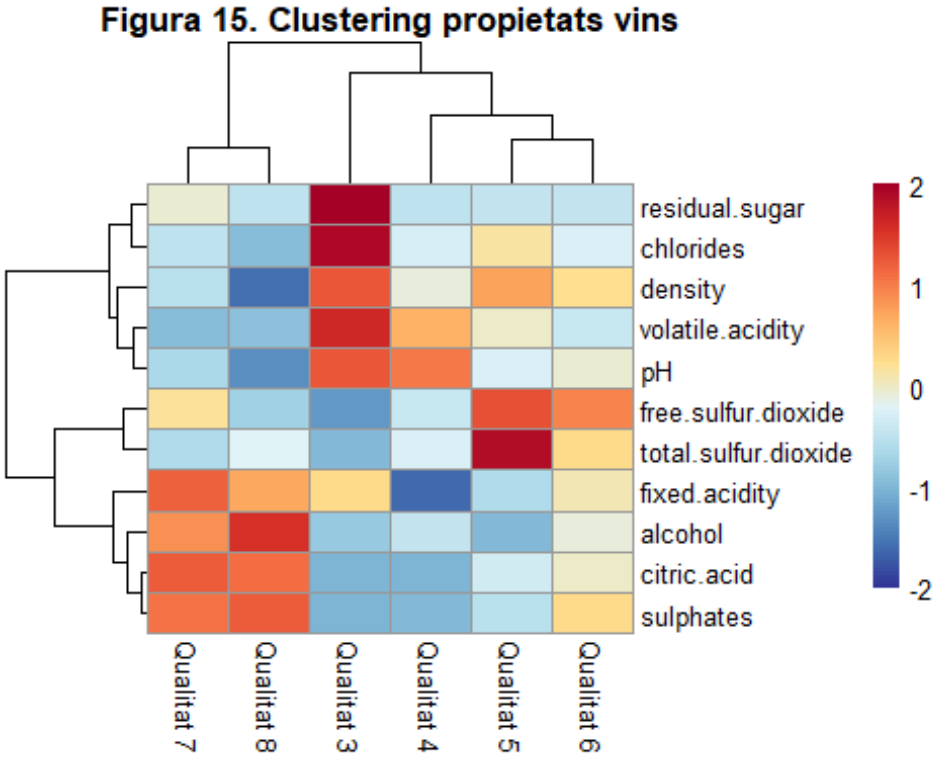
Un comp vistes les correlacions, ens interessa saber quins patrons defineixen cadascuna de les *qualitats de vi* que disposem al joc de dades. Un model **no-supervisat** basat en la distància que permet definir clústers i visualitzar tendències és el basat en **mapes de calor amb classificacio jeràrquica**. Aquest algorisme elabora dendogrames que descriuen la proximitat entre les mostres en base al càlcul de la *distància Euclídea* i dibuixa els patrons en un sistema bi-color. La funció *pheatmap()* és excel·lent per a executar aquest propòsit. A més, conté paràmetres com *cluster_cols* i *cluster_rows* que agrupa sobre el mapa de calor variables organolèptiques i vins en base a la variable independent,*qualitat*. A més, amb el paràmetre *scale="rows"* indiquem que les dades s'han de presentar centrades, és a dir, es farà un *z-score* dels valors de totes les variables per a què totes tinguin una mitjana +- error estàndard de 0 +- 1, fet que facilita enormement la comparació entre variables.

Tal i com reflecteix la **Figura 15**, en el nostre cas, els vins de qualitat superior (7-8) s'agrupen en un mateix cluster, mentre que la resta defineixen un altre grup. Aquesta

observació és molt important, ja que indica que, efectivament, **existeixen patrons de propietats organolèptiques** que defineixen la qualitat del vi, no és un factor aleatori.

Per a desxifrar **què fa bo un vi** mirarem els patrons de color. Pareix ser que **els bons vins** tenen, principalment, valors alts d'*acidesa* (segurament degut un alt contingut en *àcid cítric*), *sulfats* i alcohol. Per altra banda, l'acumulació de *sucres*, *clorids*i *densitat* **penalitzen en la qualitat del vi** (qualitat 3-4). Els vins de qualitat intermèdia (5-6) presenten un patró elevat de *sulfits* i uns nivells intermedis per a la resta de paràmetres.

```
#Importe la llibreria pheatmap:
if (!require('pheatmap')) install.packages('pheatmap'); library('pheatmap
')
if (!require('RColorBrewer')) install.packages('RColorBrewer'); library('
RColorBrewer')
# Representació:
heatmap <- pheatmap(as.matrix(t(vins_matriu)), color = colorRampPalette(r
ev(brewer.pal(n = 10, name ="RdYlBu")))(100), scale = "row", cluster_rows
= TRUE, cluster_cols = T, clustering_distance_cols = "euclidean", cluster
ing_method = "ward.D2", fontsize = 10 , main = "Figura 15. Clustering pro
pietats vins", show_rownames = T, cellwidth = 30)
```



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

Ara que ja coneixem que existeixen patrons que defineixen la qualitat del vi, convindria aplicar contrastos d'hipòtesis per a validar si aquestes diferències entre

grups són estadísticament significatives. Les anàlisis estadístiques supervisades s'han de dur a terme mitjançant **tests paramètrics** o **no-paramètrics** en funció de la normalitat que presenten i la igualtat de variàncies (*homoscedasticitat*) .

En primer lloc comprovarem si la distribució de freqüències segueix un comportament gaussià mitjançant histogrames de freqüència i compararem la distribució dels valors amb els d'una població normal (gràfics quantil-quantil o Q-Q). A més, aportarem poder estadístic a la nostra conclusió amb el test de Shapiro–Wilk (*shapiro.test()*), que proporciona p-valors que sustenten si hi ha o no **normalitat** en base al següent contrast d'hipòtesis:

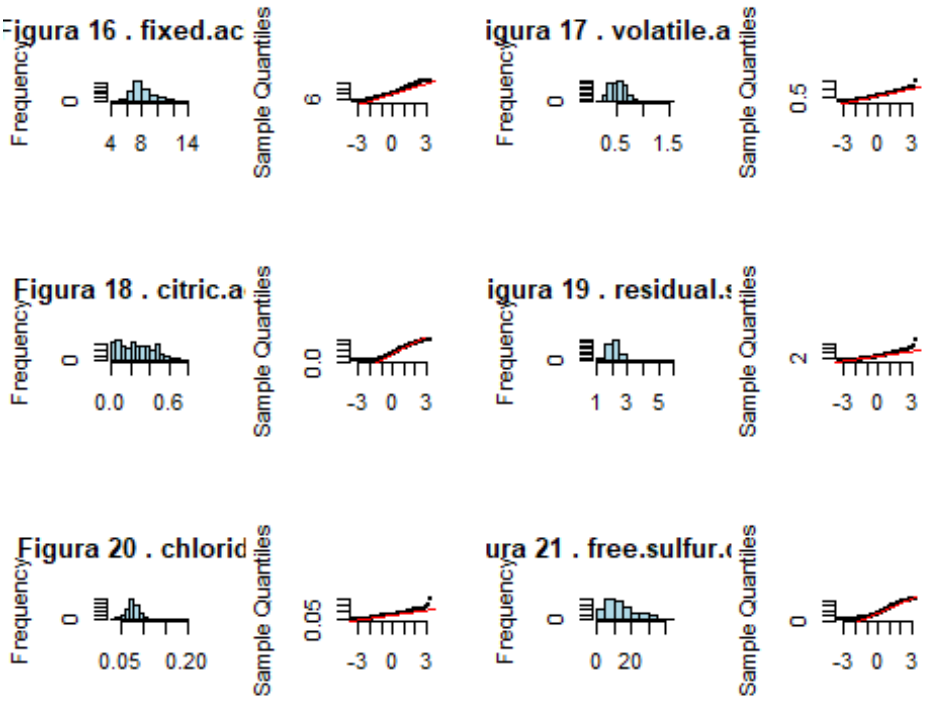
Ho: la mostra es distribueix de forma normal.

Ha: la mostra no es distribueix de forma normal.

Especialment, els gràfics Q-Q (**Figures 16-26**) mostren que les dades es desvien considerablement de la línea que representa la normalitat, i en tots els casos el test de Shapiro-Wilk ha proporcionat **p-valor < 0.05** (**Taula 6**), de manera que considerem que els dades **no es distribueixen de forma normal**.

En conseqüència, **no es compleix el primer criteri de parametricitat**.

```
# Estudi de la distribució normal
# Dataframe on guardarem els resultats:
p_val_normalitat=NULL
# Fixem el format dels gràfics:
par(mfrow=c(3,4))
# Iterem sobre les columnes del dataset vins2:
for (i in (1:11)){
  # Construïm histogrames de freqüència:
  hist(vins2[,i], main=paste("Figura", i+15, ".", colnames(vins2)[i]), xl
ab= "", col="lightblue")
  # Test de normalitat i p-valor:
  p_normalitat=shapiro.test(vins2[,i])$p.value
  # Actualitzem els p-valors i determinem si són significants per a alpha
=0.05:
  variable=colnames(vins2[i])
  p_val_normalitat=rbind(p_val_normalitat, c(variable, p_normalitat, c(p_
normalitat<=0.05)))
  # Preparem gràfics de quantils amb les dades de cada variable:
  qqnorm(vins2[,i], pch = 19, frame = FALSE, cex=0.3, main="", xlab= "")
  # Representem la líniea que adoptaria una població completament normal:
  qqline(vins2[,i], col = "red", lwd = 1)
}
```

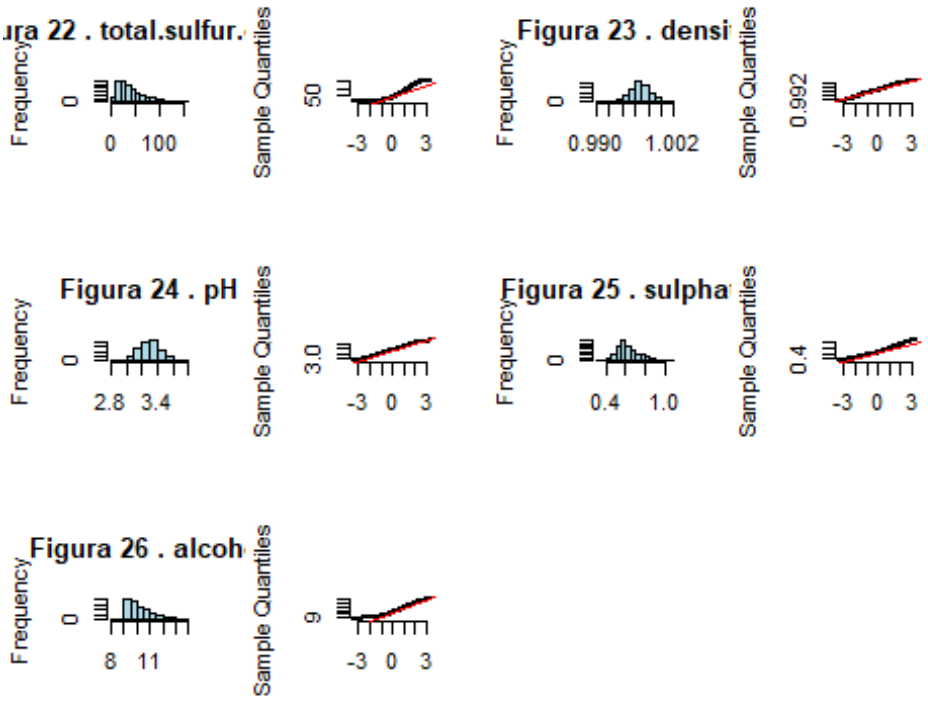



```
# Arreglem el dataframe amb els resultats i els visualitzem:
colnames(p_val_normalitat)=c("propietat", "normalitat(p-val)", "significa
nt")
print("Taula 6. Anàlisi estadística de la distribució normal dels paràmet
re per a cada qualitat de vi")
```

```
## [1] "Taula 6. Anàlisi estadística de la distribució normal dels paràme
tre per a cada qualitat de vi"
```

```
p_val_normalitat
```

```
##       propietat              normalitat(p-val)      significant
##  [1,] "fixed.acidity"        "3.27800084803818e-20" "TRUE"
##  [2,] "volatile.acidity"     "7.09697301369467e-13" "TRUE"
##  [3,] "citric.acid"          "2.42181439580563e-20" "TRUE"
##  [4,] "residual.sugar"       "3.71703116598443e-22" "TRUE"
##  [5,] "chlorides"            "5.9535961597275e-17"  "TRUE"
##  [6,] "free.sulfur.dioxide"  "2.25062237592829e-26" "TRUE"
##  [7,] "total.sulfur.dioxide" "1.10328342623071e-32" "TRUE"
##  [8,] "density"              "0.000424702941059658" "TRUE"
##  [9,] "pH"                   "0.0361483000069471"   "TRUE"
## [10,] "sulphates"            "9.50085301528725e-19" "TRUE"
## [11,] "alcohol"              "3.02969604123927e-27" "TRUE"
```



Seguidament hauríem d'analitzar si hi ha **igualtat de variàncies** entre les mostres, és a dir, *homoscedasticitat*. La ferramenta *var.test()* ens ajuda a aplicar un *test de la F* per al següent contrast d'hipòtesis:

Ho: variància mostra A = variaància mostra B.

Ha: variaància mostra A != variaància mostra B.

Si prenem la diagonal superior de la matriu resultant de p-valors a la **Taula 7**, observarem que en tots el casos els **p-valors < 0.05** i, per tant, **no podem assumir homoscedasticitat**.

En conseqüència caldrà emprar **tests no-paramètrics** per a l'estadística descriptiva.

```
# Igualtat de variàncies: homoscedasticitat
# Dataframe on guardarem els resultats:
homoscedasticitat=NULL
# For loop que recorre totes les columnes dues vegades:
for (i in (1:11)){
  for (j in (1:11)){
  # var.test() en forma de vis-a-vis (pairwise):
  homo=var.test(vins2[,i], vins2[,j], alternative = c("two.sided"))$p.val
ue
  homoscedasticitat=cbind(homoscedasticitat, c(homo))
  }
}
# Arreglem el dataframe:
```

```
homoscedasticitat=matrix(unlist(homoscedasticitat), ncol = 11, byrow=T)
colnames(homoscedasticitat)=colnames(vins2[1:11])
row.names(homoscedasticitat)=colnames(vins2[1:11])
# Considerem la matriu superior:
homoscedasticitat[lower.tri(homoscedasticitat)]=NA
print("Taula 7. Matriu diagonal de p-valors per al test d'homoscedasticit
at aplicat a les propietats del vi")
```

```
## [1] "Taula 7. Matriu diagonal de p-valors per al test d'homoscedastici
tat aplicat a les propietats del vi"
```

```
homoscedasticitat
```

```
##                       fixed.acidity volatile.acidity citric.acid residu
al.sugar
## fixed.acidity                     1                0 0.000000000   0.00
0000e+00
## volatile.acidity                 NA                1 0.002826195  9.814
152e-281
## citric.acid                      NA               NA 1.000000000  4.388
512e-243
## residual.sugar                   NA               NA          NA   1.00
0000e+00
## chlorides                        NA               NA          NA
NA
## free.sulfur.dioxide              NA               NA          NA
NA
## total.sulfur.dioxide             NA               NA          NA
NA
## density                          NA               NA          NA
NA
## pH                               NA               NA          NA
NA
## sulphates                        NA               NA          NA
NA
## alcohol                          NA               NA          NA
NA
##                       chlorides free.sulfur.dioxide total.sulfur.dioxid
e density
## fixed.acidity                 0                   0
0       0
## volatile.acidity              0                   0
0       0
## citric.acid                   0                   0
0       0
## residual.sugar                0                   0
0       0
## chlorides                     1                   0
0       0
## free.sulfur.dioxide          NA                   1
0       0
```

```
## total.sulfur.dioxide        NA                  NA
1       0
## density                     NA                  NA                    N
A       1
## pH                          NA                  NA                    N
A      NA
## sulphates                   NA                  NA                    N
A      NA
## alcohol                     NA                  NA                    N
A      NA
##                                pH   sulphates       alcohol
## fixed.acidity        0.00000e+00 0.00000e+00  0.000000e+00
## volatile.acidity     1.67466e-12 0.00000e+00  0.000000e+00
## citric.acid          0.00000e+00 0.00000e+00  0.000000e+00
## residual.sugar       0.00000e+00 0.00000e+00 1.402835e-224
## chlorides            0.00000e+00 0.00000e+00  0.000000e+00
## free.sulfur.dioxide  0.00000e+00 0.00000e+00  0.000000e+00
## total.sulfur.dioxide 0.00000e+00 0.00000e+00  0.000000e+00
## density              0.00000e+00 0.00000e+00  0.000000e+00
## pH                   1.00000e+00 4.28324e-13  0.000000e+00
## sulphates                    NA 1.00000e+00  0.000000e+00
## alcohol                      NA          NA  1.000000e+00
```

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.

Abans d'aplicar els tests, hem d'investigar amb un poc més de detall sobre quins paràmetres volem centrat l'anàlisi estadística. Una bona pràctica seria realitzar una nova correlació per saber quines són les variables que tenen major influència amb la *qualitat del vi*. Amb aquest propòsit tornem a definir *qualitat* com una variable numèrica amb *as.numeric()* i calculem les correlacions amb segons el *coeficient d'Spearman*, l'alternativa no-paramètrica per a correlacions.

Tal i com es resumeix a la **Taula 8**, després d'obtenir la correlació de les 11 variables amb la qualitat, la primera conclusió que extraiem és que **els atributs no tenen molta correlació tot i presentar una alta significança (p-valors < 0.05)**, ja que no n'hi ha cap que arribi a 0.6 en valor absolut. Si prenen com a criteri un coeficient en valor absolut superior o igual a 0.4, **concloem que les propietats organolèptiques** més influents són l'*acidesa volàtil* (rho=-0.4), el nivell de *sulfats* (rho=0.46) i el contingut en *alcohol* (rho=0.52). Com ja hem vist al mapa de calor de la **Figura 15**, a més *sulfats* i *alcohol*, més alta serà a *qualitat del vi*, mentre que un increment en *acidesa volàtil* tindrà l'efecte oposat.

De totes aquestes, farem un seguiment dels nivells de *sulfats* i del contingut en *alcohol* ja que presenten **els dos coeficients més alts** i a mes són **bioquímicament fàcils de**

manipular tot afegint-ne més derivats (cas dels *sulfats*) o allargant la fermentació (*cas alcohol*).

```
#Qualitat no es numèric, el convertim a numèric:
cvins=vins2
cvins$quality = as.numeric(cvins$quality)
#Calculem la correlació de les 12 variables amb la qualitat
Taula <- matrix(nc = 2, nr = 0)
colnames(Taula) <-c("CORRELACIÓ", "p-value")
for (i in 1:(ncol(cvins) - 1)) {
  if (is.integer(cvins[,i]) | is.numeric(cvins[,i])) {
        spearman_test = cor.test(cvins[,i], cvins[,length(cvins)],method
= "spearman")
        Correlacio = spearman_test$estimate
        p_value = spearman_test$p.value
          pair = matrix(ncol = 2, nrow = 1)
        pair[1][1] = Correlacio
        pair[2][1] = p_value
        Taula <- rbind(Taula, pair)
        rownames(Taula)[nrow(Taula)] <- colnames(cvins)[i]
     }
}
# Imprimim la taula
print("Taula 8. Llistat de les propietats organolèptiques més influents e
n la qualitat del vi segons el coeficient d'Spearman")
```

```
## [1] "Taula 8. Llistat de les propietats organolèptiques més influents
en la qualitat del vi segons el coeficient d'Spearman"
```

```
Taula
```

```
##                       CORRELACIÓ       p-value
## fixed.acidity          0.14308304  9.102895e-09
## volatile.acidity      -0.40095916  8.230325e-63
## citric.acid            0.24864535  5.880994e-24
## residual.sugar         0.04802267  5.486898e-02
## chlorides             -0.20244587  2.989487e-16
## free.sulfur.dioxide   -0.06861717  6.052589e-03
## total.sulfur.dioxide -0.22944463  1.514523e-20
## density               -0.18725166  4.395100e-14
## pH                    -0.05681817  2.308197e-02
## sulphates              0.46447445  2.291809e-86
## alcohol                0.51692086 5.936751e-110
```

Realitzem per tant els contrasts d'hipòtesis per a ambdós paràmetres (*sulfats* i contingut en *alcohol*) entre la gamma de qualitat baixa (*qualitat vi* = 3 o 4) i la gamma alta (*qualitat vi* = 7 a 8). Cal recordar que com que les variables mostres un comportament no-normal hem d'emprar testos no-paramètrics. Aleshores, enlloc del **test de la t d'Student**, farem ús del seu equivalent no-paramètric, **el test de Wilcox**. Amb els paràmetres *alternative = "two.sided", mu = 0, paired = F* i *conf.int = 0.95*

indique que el test és bidireccional (*two-sided*), que la hipòtesi nul·la per a la mitjana és 0, que el test no és emparellat i que cerquem una confiança del 95% per a no rebutjar al hipòtesi nul·la.

Plantegem el primer **contrast d'hipòtesi**:

Ho: La mediana de las diferències en *alcohol* per cada parell de dades és zero.

Ha: La mediana de las diferències en *alcohol* per cada parell de dades no és zero.

El test indica un **p-valor = 1.508e-15** per al contrast proposat (**Taula 9**), de manera que podem rebutjar la hipòtesi nul·la amb molta confiança i declarar que efectivament, **les diferències en contingut d'alcohol són molt significatives**.

```
# Extraiem alcohol, sulfats i qualitat
tests=subset(vins2, select=c(alcohol, sulphates, quality))
# Filtrem la variable alcohol per a gamma alta i baixa
alcohol_baixa=subset(tests, (quality==3 | quality== 4))
alcohol_alta=subset(tests, (quality==7 | quality== 8))
# Executem el test de Wilcox per a alcohol
print("Taula 9. Anàlisi estadístic de la diferència en contingut d'alcoho
l entre mostres de difernt qualitat")
```

```
## [1] "Taula 9. Anàlisi estadístic de la diferència en contingut d'alcoh
ol entre mostres de difernt qualitat"
```

```
wilcox.test(x=alcohol_baixa$alcohol, y=alcohol_alta$alcohol, alternative
= "two.sided", mu = 0, paired = F, conf.int = 0.95)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  alcohol_baixa$alcohol and alcohol_alta$alcohol
## W = 2324.5, p-value = 1.508e-15
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##  -1.600032 -1.000058
## sample estimates:
## difference in location
##              -1.300014
```

Plantegem el segon **contrast d'hipòtesi**:

Ho: La mediana de las diferències en *sulfats* per cada parell de dades és zero.

Ha: La mediana de las diferències en *sulfats* per cada parell de dades no és zero.

El test indica un **p-valor = 2.2e-16** per al contrast proposat (**Taula 10**), de manera que podem rebutjar la hipòtesi nul·la amb molta confiança i declarar que efectivament, **les diferències en contingut de sulfats són molt significatives**.

```
# Filtrem la variable alcohol per a gamma alta i baixa
sulphats_baixa=subset(tests, (quality==3 | quality== 4))
sulphats_alta=subset(tests, (quality==7 | quality== 8))
# Executem el test de Wilcox per a alcohol
print("Taula 10. Anàlisi estadístic de la diferència en contingut d'alcoh
ol entre mostres de difernt qualitat")

## [1] "Taula 10. Anàlisi estadístic de la diferència en contingut d'alco
hol entre mostres de difernt qualitat"

wilcox.test(x=alcohol_baixa$sulphates, y=alcohol_alta$sulphates, alternat
ive = "two.sided", mu = 0, paired = F, conf.int = 0.95)

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  alcohol_baixa$sulphates and alcohol_alta$sulphates
## W = 1105.5, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##  -0.2199800 -0.1699204
## sample estimates:
## difference in location
##             -0.1900411
```

Finalment, tot i que les nostres variables no presenten una distribució normal, intentarem construir un **model de regressió lineal** per veure si podem predir la *qualitat del vi* en funció de les variables que hem recollit. Aquest sistema de predicció basat en un model **supervisat** pot ser-nos útil de cara a anticipar-nos durant les proves d'elaboració de vi i de millora, avaluacions de si val la pena llançar al mercat una nova varietat, o **sobretot en l'automatització del procés de control de qualitat**.

L'estratègia que hem seguit ha sigut dividir el conjunt de dades **vins2** aleatòriament en un 66% per a entrenar el model (train-testX) i un 33% per a provar-lo (train-testY). La variable *prova* per a classificar serà, com no, *vins_quality*. Generem diferents models de regressió lineal, aleatòriament, amb diferents conjunts de variables. A la **Taula 11** recollim quins són els millors models en base al paràmetre de qualitat **R2**. Tal i com es pot apreciar a la **Taula 11**, ens quedem amb el **model 3**: lm(trainy ~ alcohol + sulphates + volatile.acidity + citric.acid, pH, data = trainX), ja que té el valor més alt, **R2=0.5**.

Tot seguit avaluem amb la funció *predict()* a quina categoria pertanyen els vins del set *testX* segons el **model 3** i tot seguit comparem quantes vegades hi ha hagut una coincidència entre allò predit (*perdict*) vs. les dades reals (*testY*) (**Taula 12**) i en calculem el percentatge. Malauradament, el model ha classificat totes les mostres a la categoria 4. El percentatge d'encert del model representa la **precisió del model 3**, que és del **40.15%**. Açò significa que és capaç de predir correctament 4 de cada 10 de les instància a partir de les variables que li proporcionem. Tanmateix, hem de ser molt

curosos i hauríem de testar el model amb dades provinents d'altres datasets o refer-lo de forma més precisa o utilitzant altres algorismes perquè aquest resultat pot estar influït per una saturació del model o **overfitting**.

```
# Preparem el conjunt d'entrenament i prova
# Delimite el comptador aleatori set.seed()
set.seed(1234)
# Desordene les fileres del fitxer:
data_random = vins2[sample(nrow(vins2)),]
# Destrie com a y la variables pregunta de la resta, X.
y = as.numeric(data_random[,length(vins2)]) # Cal convertir la qualitat a
numèric
X = data_random[, 1:length(vins2)-1]
# Definesc quant és un terç del total de fileres:
terc= round(2/3*nrow(data_random))
# Definesc els conjunts d'entrenament i test en base al terç:
trainX <- X[1:terc,]
trainy <- y[1:terc]
testX  <- X[(terc+1):nrow(data_random),]
testy  <- y[(terc+1):nrow(data_random)]
#Definim diferents models amb regressors quantitaus basant-no en les vari
ables que més coeficient de correlació tenen amb la qualitat
model1 <- lm(trainy ~ density + pH + sulphates + citric.acid , data = tra
inX)
model2 <- lm(trainy ~ alcohol + sulphates + volatile.acidity + citric.aci
d , data = trainX)
model3 <- lm(trainy ~ alcohol + sulphates + volatile.acidity + citric.aci
d, pH, data = trainX)
model4 <- lm(trainy ~ density + pH,  data = trainX)
# Avaluació de la qualitat dels models en base a R2:
taula <- matrix(c(1, summary(model1)$r.squared,
2, summary(model2)$r.squared,
3, summary(model3)$r.squared,
4, summary(model4)$r.squared),
ncol = 2, byrow = TRUE)
colnames(taula) <- c("Model", "R^2")
print("Taula 11. Avaluació models línials")

## [1] "Taula 11. Avaluació models línials"

taula

##      Model        R^2
## [1,]     1 0.32351067
## [2,]     2 0.43452873
## [3,]     3 0.49997491
## [4,]     4 0.07072654

# Predicció i avaluació de la precisió del model:
dades <- data.frame(alcohol=testX$alcohol, sulphates=testX$sulphates, vol
atile.acidity=testX$volatile.acidity, citric.acid=testX$citric.acid, pH=t
```

```
estX$pH)
print("Taula 12. Predicció de qualitat amb el model 3")
```

```
## [1] "Taula 12. Predicció de qualitat amb el model 3"
```

```
prediccio=unname(predict(model3, dades))
prediccio
```

```
##   [1] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
##  [38] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
##  [75] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
## [112] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
## [149] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
## [186] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
## [223] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
## [260] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
## [297] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
## [334] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
## [371] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
## [408] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
## [445] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
## [482] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
## [519] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
```

```
# Recompte quantes equivalència entre les prediccion si els valors origni
nals per a qualitat de vi hi ha.
# Càlculem la diferència entre els valors de qualitat i recompte quants s
ón encerts (valor 0) pe a estimar la precisió:
rr=round(prediccio-testy)
print("Taula 13. Precisió del model 3")
```

```
## [1] "Taula 13. Precisió del model 3"
```

```
100*table(rr==0)/length(testy)
```

```
##
##    FALSE     TRUE
## 59.84991 40.15009
```

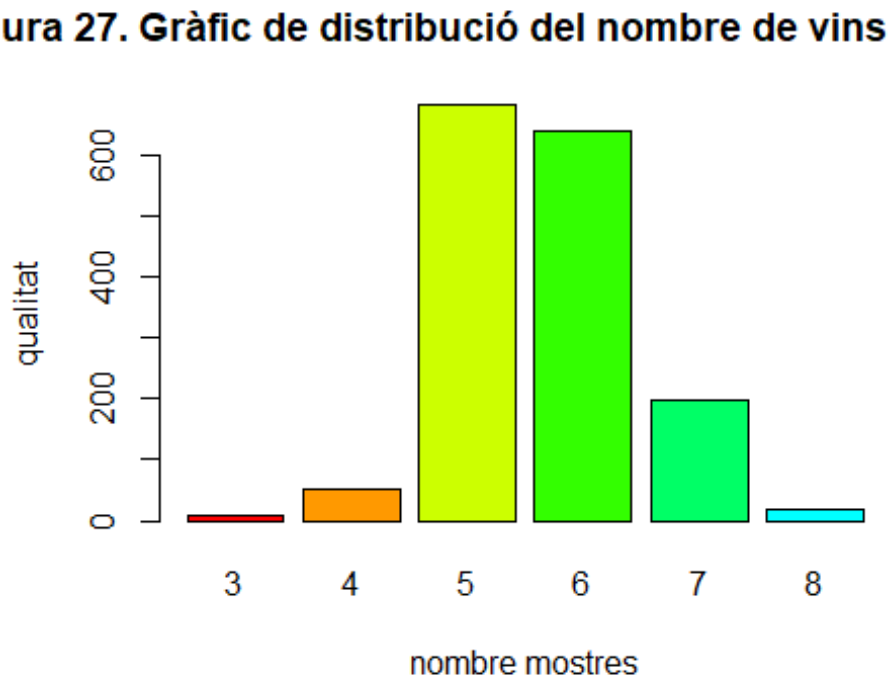
5.Representació dels resultats a partir de taules i gràfiques.

Al llarg d'aquest estudi ja hem presentat diferents representacions gràfiques dels resultats que hem anat obtenint (correlacions, mapes de calor, distribució de valors, etc). Tanmateix en aquest apartat procedirem a elaborar alguns gràfics resum addicionals.

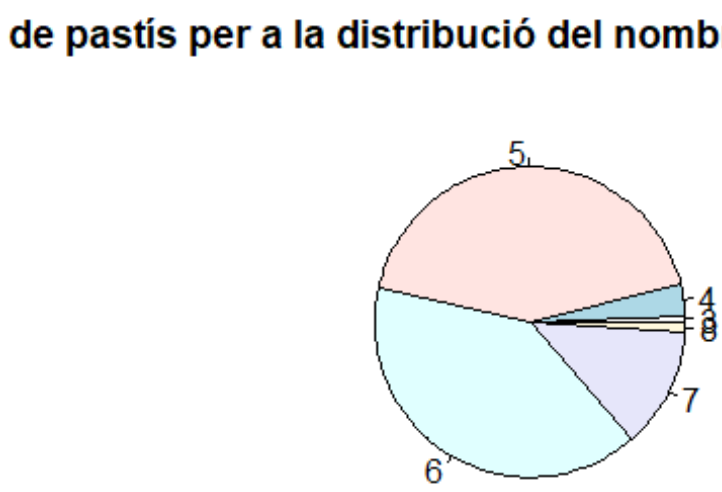
En primer lloc, les **Figures 27 i 28** mostren amb dues modalitats diferents (recompte i gràfic de pastís) el nombre d'entrades recollides per a cada vi en funció de la *qualitat*. Podem comprovar que les dades per a la qualitat 5 i 6 són les més abundants amb diferència amb vora 600 entrades.

A continuació hem elaborat un recompte de la distribució de valors per a la distribució dels valors per a *alcohol* (**Figura 29**) i *sulfats* (**Figura 30**) en funció de la categoria. Podem comprovar que els vins amb qualitat alta (6-8) s'acumulen cap als valors d'*alcohol* i per a *sulfats* alts.

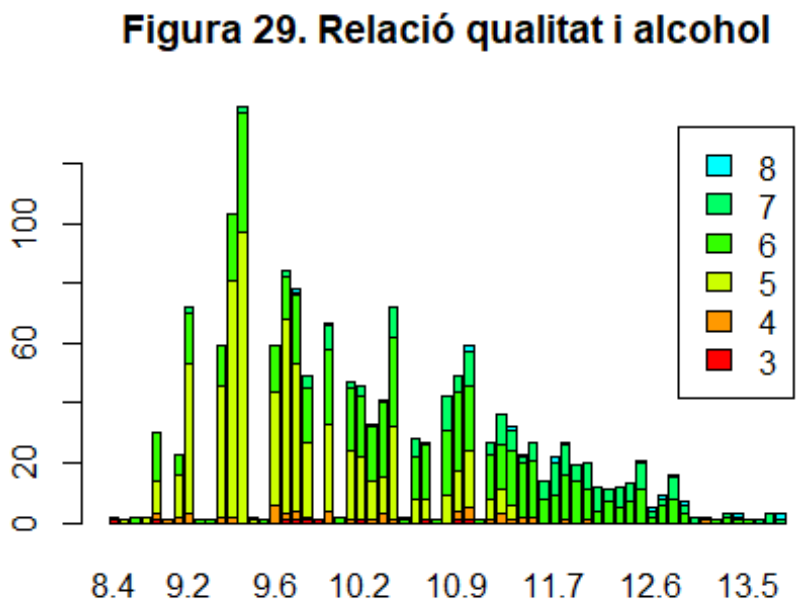
```
# Carreguem llibreries d'interés
library(ggplot2)
library(scales)
# Gràfic de barres per lo veure la població que hi ha en cada qualitat. E
s veu que la 3 i 4 es la més poblada
plot(x =vins2$quality, main = "Figura 27. Gràfic de distribució del nombr
e de vins per qualitat", xlab = "nombre mostres", ylab = "qualitat" ,  co
l=rainbow(10) )
```



```
# Gràfic de pastis
subQualitat <- table(vins2$quality)
pie(subQualitat, main = "Figura 28. Gràfic de pastís per a la distribució
del nombre de vins per qualitat de la qualitat")
```

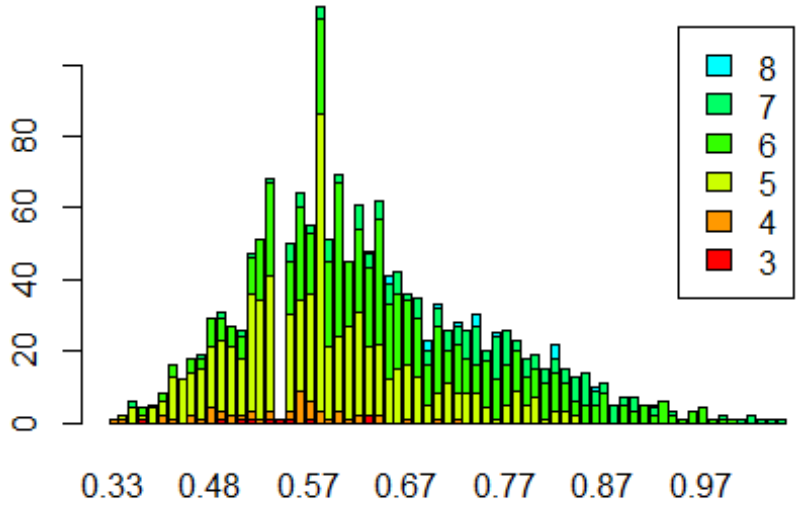



```
# Gràfica de barres relacionant la qualitat i alcohol
subtaula <- table(vins2$quality, vins2$alcohol)
barplot(subtaula, col=rainbow(10),cylindrical=TRUE,shadow=TRUE, main = "F
igura 29. Relació qualitat i alcohol", legend.text=c("3","4","5","6","7",
"8"))
```




```
# Gràfica de barres relacionant la qualitat i els sulfats
subtaula2 <- table(vins2$quality, vins2$sulphates)
barplot(subtaula2, col=rainbow(10),cylindrical=TRUE,shadow=TRUE, main = "
Figura 30. Relació qualitat i sulfats", legend.text=c("3","4","5","6","7"
,"8"))
```

Figura 30. Relació qualitat i sulfats



6.Resolució del problema. A partir dels resultats obtinguts, quines són les conclusions? Els resultats permeten respondre al problema?

El joc de dades que hem carregat ha sigut seleccionat per a investigar quines propietats organolèptiques tenen més pes a l'hora de determinar la qualitat d'un vi. El nostre objectiu ha sigut identificar-les i elaborar un model de regressió lineal capaç d'anticipar en base a les propietats del vi quina qualitat presenta. Com ja hem esmentat aquest tipus d'estudi és de gran ajuda per a tasques d'automatització de control de qualitat en base a algorismes d'intel·ligència artificial.

Fent ús de mètodes de correlació, hem comprovat que el contingut en alcohol i en sulfats són dues propietats que contribueixen de forma positiva a la qualitat del vi, tot i que de forma moderada (coeficient d'Spearman al voltant 0.5). De fet, veiem en els mapes de calor que els vins amb les dues qualificacions més altes acumules els valors més alts per a ambdós paràmetres. Els mètodes d'estadística analítica basats en tests no-paramètrics han contribuït a confirmar que les diferències entre els grups d'alta i baixa qualitat pel que fa a ambdós paràmetres són molt significatives (p-valor < 0.05, Wilxon test).

Finalment hem aconseguit un model no-supervisat de regressió lineal amb 5 variables capaç de preveure la qualitat del vi amb una precisió del 40%. Tanmateix, hem de ser curosos perquè potser degut a la natura de les dades o al fet que hi haja una sobrerepresentació de la categoria 5 i 6, el model pot sofrir *overfitting*.

En definitiva, hem sigut capaços d'adequar i analitzar un joc de dades quantitatiu per a obtenir patrons de comportament i anticipar tendències que és molt extrapolable a qualsevol tipus d'estudi de mercat que pretengui investigar l'acceptació o qualitat d'un producte en base a paràmetres fàcils de monitoritzar.

7.Codi: Cal adjuntar el codi, preferiblement en R, amb el que s'ha realitzat la neteja, anàlisi i representació de les dades. Si ho preferiu, també podeu treballar en Python.

El codi es presenta en format de fitxer *.Rmd*, fet que permet veure el codi elaborat junt amb els resultats en format fàcil de lectura, tipus *.pdf*, *.html* o *.doc*.

Contribució	Signatura
Investigacio previa	Antoni Garcia i Marta Puigdemasa
Redaccio de les respostes	Antoni Garcia i Marta Puigdemasa
Desenvolupament codi	Antoni Garcia i Marta Puigdemasa