practica2

Alejandro Heredia y Pere Garcia

9/12/2021

Pràctica 2

Enllaços de github i video

- GitHub Repo aherediac/TSVD-PRAC2: Practica 2 (Tipologia i cicle de vida de les dades)
- Video de la pràctica

Llibreries requerides

```
library('nortest')
library('ggplot2')
library('reshape2')
library('corrplot')
```

```
## corrplot 0.92 loaded
```

```
# R version 4.1.2 (2021-11-01) - Bird Hippie
```

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

El dataset que utilitzarem l'hem obtingut del enllaç: Red Wine Quality | Kaggle Aquest dataset conté 12 columnes de dades:

- **fixed acidity**: El sumatori de tots els àcids fixos tant orgànics com minerals de la seva composició. Comprèn valors entre 4.6 i 15.9 g/l. A més alt més acidesa té el vi.
- **volatile acidity**: Acidesa volàtil la qual es pot reduïr utilitzant processos químics. Comprèn valors entre 0.120 i 1.580 g/l.
- **citric acid**: Àcid cítric. Comprèn valors entre 0 i 1 g/l.
- **residual sugar**: Quantitat total de sucre que queda al vi que no s'ha fermentat per les llevadures. Mescla de glucosa i fructosa. És sucre del mosto del raïm. Comprèn valors entre 0.90 i 15.5 g/l.
- **chlorides**: Clorur, quantitat de sal que té el vi. Comprèn valors entre 0.012 i 0.611 g/l.
- **free sulfur dioxide**: Lliure de sulfits. Comprèn valors entre 1 i 72 mg/l.
- **total sulfur dioxide**: Sulfits. Comprèn valors entre 6 i 289 mg/l.
- **density**: Densitat del vi. Comprèn valors entre 0.99007 i 1.00369 g/l. Com menys densitat més alcohol. A més densitat, menys alcohol. Normalment com més alcohol té un raïm, més madur estava en el moment de la collita. Podem assumir aquesta condició en aquesta pràctica.

- **pH**: Representa l'acidesa o alcalinitat. Es medeix en una escala del 0 al 14 i com més proper a 0, més àcid és el vi. Comprèn valors entre 2.74 i 4.01.
- **sulphates**: Sulfats, un aditiu afegit al vi que actua com a antimicrobis i antioxidant. Comprèn valors entre 0.33 i 2 g/l.
- **alcohol**: Graduació d'alcohol del vi. Percentatge que comprèn valors entre 8.4 i 14.9 graus. Normalment els podem catalogar en: Molt baix (< 12.5º), baix (12.5 <= 13.5), alt (13.5 <= 14.5) i molt alt (> 14.5). En aquesta pràctica utilitzarem aquesta classificació.
- **quality**: Puntuació de qualitat otorgada per un sensor extern ja inclòs en el dataset. Comprèn valors entre 0 i 10.

Aquestes dades són estructurades mitjançant un fitxer CSV separat per comes i hi ha un total de 1599 medicions (files).

Les dades que conté són les propietats del vi negre de la varietat portuguesa "Vinho Verde" i hem de tenir en compte que no hi ha molta dispersió de dades, és a dir, no hi ha molts "outliers" (ho diu a la propia documentació del dataset).

```
# Carrega del fitxer
wine_data  <- read.csv("winequality-red.csv", header = TRUE)

# Verifiquem l'estructura del joc de dades
str(wine_data)
```

```
## 'data.frame':    1599 obs. of  12 variables:
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
##  $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
##  $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071 ...
##  $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
##  $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
##  $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
##  $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...
##  $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...
##  $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
##  $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...
```

```
#Estadístiques bàsiques
summary(wine_data)
```

```
##  fixed.acidity   volatile.acidity  citric.acid    residual.sugar
##  Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900
##  1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900
##  Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200
##  Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539
##  3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600
##  Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide    density
##  Min.   :0.01200   Min.   : 1.00       Min.   :  6.00       Min.   :0.9901
##  1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00       1st Qu.:0.9956
##  Median :0.07900   Median :14.00       Median : 38.00       Median :0.9968
##  Mean   :0.08747   Mean   :15.87       Mean   : 46.47       Mean   :0.9967
##  3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00       3rd Qu.:0.9978
```

```
##  Max.   :0.61100   Max.   :72.00       Max.   :289.00       Max.   :1.0037
##        pH          sulphates         alcohol         quality
##  Min.   :2.740   Min.   :0.3300   Min.   : 8.40   Min.   :3.000
##  1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50   1st Qu.:5.000
##  Median :3.310   Median :0.6200   Median :10.20   Median :6.000
##  Mean   :3.311   Mean   :0.6581   Mean   :10.42   Mean   :5.636
##  3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10   3rd Qu.:6.000
##  Max.   :4.010   Max.   :2.0000   Max.   :14.90   Max.   :8.000
```

Ens interessa conèixer els detalls per tal de saber:

1. A partir de les dades fisicoquímiques, volem aclarir si solament amb les dades de l'acidesa, es pot classificar el vi amb la seva puntuació correctament sense gaire marge d'error.
   El dataset conté diverses variables, entre elles diversos àcids. Volem saber si existeix una relació entre l'acidesa i la puntuació que rep el vi.
2. Quina és la variable més relacionada amb la qualitat del vi.
   La qualitat del vi ve delimitada per una puntuació. Ens interesa conèixer quines propietats (variables) fan que el vi tingui una alta puntuació.
3. Poder realitzar models de regressió logística en funció de les variables mes relacionades.
   Tenint en compte les dades del tipus de vi, un cop tractades, podem utilitzar-les per predir futures anyades de vins i poder assignar-les una puntuació.

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

Per tal de saber si amb les dades de l'acidesa podem establir una relació amb les puntuacions de la qualitat, podem eliminar totes les dades que no en formin part, com poden ser el sucre residual, el clorur, els sulfits, el sulfats i l'alcohol. El dataset quedaria així:

```
# Eliminem les primeres
wine_data_acid <- wine_data[, -(4:8)]

#Eliminem els sulfats i l'alcohol
wine_data_acid <- wine_data_acid[, -(5:6)]

colnames(wine_data_acid)
```

```
## [1] "fixed.acidity"    "volatile.acidity" "citric.acid"      "pH"
## [5] "quality"
```

Mentres que per poder trobar la variable que més impacte té en la puntuació, necessitarem totes i cadascuna de les que tenim al dataset, ja que totes formen part de les característiques del vi.

```
colnames(wine_data)
```

```
##  [1] "fixed.acidity"        "volatile.acidity"     "citric.acid"
##  [4] "residual.sugar"       "chlorides"            "free.sulfur.dioxide"
##  [7] "total.sulfur.dioxide" "density"              "pH"
## [10] "sulphates"            "alcohol"              "quality"
```

Per tant, arribats a aquest punt tenim dos datasets, un de complet i un altre amb només dades dels àcids i les puntuacions:

```
# wine_data
# wine_data_acid
```

3. Neteja de les dades.

3.1. Les dades contenen zeros o elements buits? Com gestionaries aquests casos? El primer que podem veure és que els noms de les columnes contenen espais i això ho hem d'eliminar. El mateix R al llegir-ho del csv ja ho tracta, pero es millor refer aquesta notació i utilitzar-ne una que sigui més llegible ja que si fèssim servir Python, la notació amb punts "." podria semblar un pipeline de mètodes. A més aprofitem per treure la majúscula del "pH" per normalitzar tots els noms a minúscules. Per fer-ho:

```
# Renombrem noms de columnes amb espais per subguions "_"

colnames(wine_data_acid)[1] <- "fixed_acidity"
colnames(wine_data_acid)[2] <- "volatile_acidity"
colnames(wine_data_acid)[3] <- "citric_acidity"
colnames(wine_data_acid)[4] <- "ph"

colnames(wine_data)[1] <- "fixed_acidity"
colnames(wine_data)[2] <- "volatile_acidity"
colnames(wine_data)[3] <- "citric_acidity"
colnames(wine_data)[4] <- "residual_sugar"
colnames(wine_data)[6] <- "free_sulfur_dioxide"
colnames(wine_data)[7] <- "total_sulfur_dioxide"
colnames(wine_data)[9] <- "ph"

# Mostreig
colnames(wine_data)
```

```
##  [1] "fixed_acidity"        "volatile_acidity"     "citric_acidity"
##  [4] "residual_sugar"       "chlorides"            "free_sulfur_dioxide"
##  [7] "total_sulfur_dioxide" "density"              "ph"
## [10] "sulphates"            "alcohol"              "quality"
```

Un cop revisades les capçaleres, comprovem que no tinguem cap element NULL, ja que d'entrada sí podem veure que hi han propietats a zero, però pot ser completament correcte, així que d'entrada no veiem que haguem de fer cap operació de neteja d'elements buits sobre el dataset.

```
# Comprovem el total d'elements NULL
colSums(is.na(wine_data))
```

```
##        fixed_acidity     volatile_acidity       citric_acidity
##                    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
```

```
# Comprovem el total d'elements buits
colSums(wine_data=="")
```

```
##        fixed_acidity     volatile_acidity       citric_acidity
##                    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
```

Com podem veure per ambdós casos no tenim valors buits o NULLS, però si zeros que com hem comentat són valors possibles. Si per algún motiu en trobèssim algún de buit o NULL, podriem fer:

1. Eliminar el registre: El més senzill seria eliminar aquest registre i treure problemes si no poguèssim identificar un valor.
2. Intentar trobar-li un valor agrupant les altres dades del registre per exemple utilitzant K-NN, però sempre comprobant el resultat.
3. Assignar un valor dins la mitja de tots els disponibles per no introduïr inconsistències.
4. Identificant-les amb una nova etiqueta per tal de poder separar-les del anàlisis en cas de requerir-ho. Això ajudarà a identificar aquelles dades que no tenen un valor definit i ens permetrà conèixer on són més fàcilment.

En qualsevol cas, totes aquestes mesures que no farà falta aplicar la nostre dataset, s'hauràn de posar en marxa vigilant que no indueixin els resultats a falsos positius o negatius.

3.2. Identificació i tractament de valors extrems. Com deiem al principi de la pràctica, no tenim molts registres amb valors "outliers" molt diferenciats entre altres. Podem veure per exemple que tenim quatre registres amb la dada de "free_sulfur_dioxide" molt diferenciada si les comparem amb les altres, però les altres dades del registre són dins la mitja i també és possible que aquest vi en concret, hagi sortit amb aquestes característiques. Característiques alterades? poder si, però possibles.

```
head(wine_data[order(-wine_data$free_sulfur_dioxide),])
```

```
##      fixed_acidity volatile_acidity citric_acidity residual_sugar chlorides
## 1245           5.9            0.290           0.25           13.4     0.067
## 397            6.6            0.735           0.02            7.9     0.122
## 401            6.6            0.735           0.02            7.9     0.122
## 1559           6.9            0.630           0.33            6.7     0.235
## 1132           5.9            0.190           0.21            1.7     0.045
## 1435          10.2            0.540           0.37           15.4     0.214
##      free_sulfur_dioxide total_sulfur_dioxide density   ph sulphates alcohol
## 1245                  72                  160 0.99721 3.33      0.54    10.3
## 397                   68                  124 0.99940 3.47      0.53     9.9
## 401                   68                  124 0.99940 3.47      0.53     9.9
## 1559                  66                  115 0.99787 3.22      0.56     9.5
## 1132                  57                  135 0.99341 3.32      0.44     9.5
## 1435                  55                   95 1.00369 3.18      0.77     9.0
##      quality
## 1245       6
```

```
## 397        5
## 401        5
## 1559       5
## 1132       5
## 1435       6
```

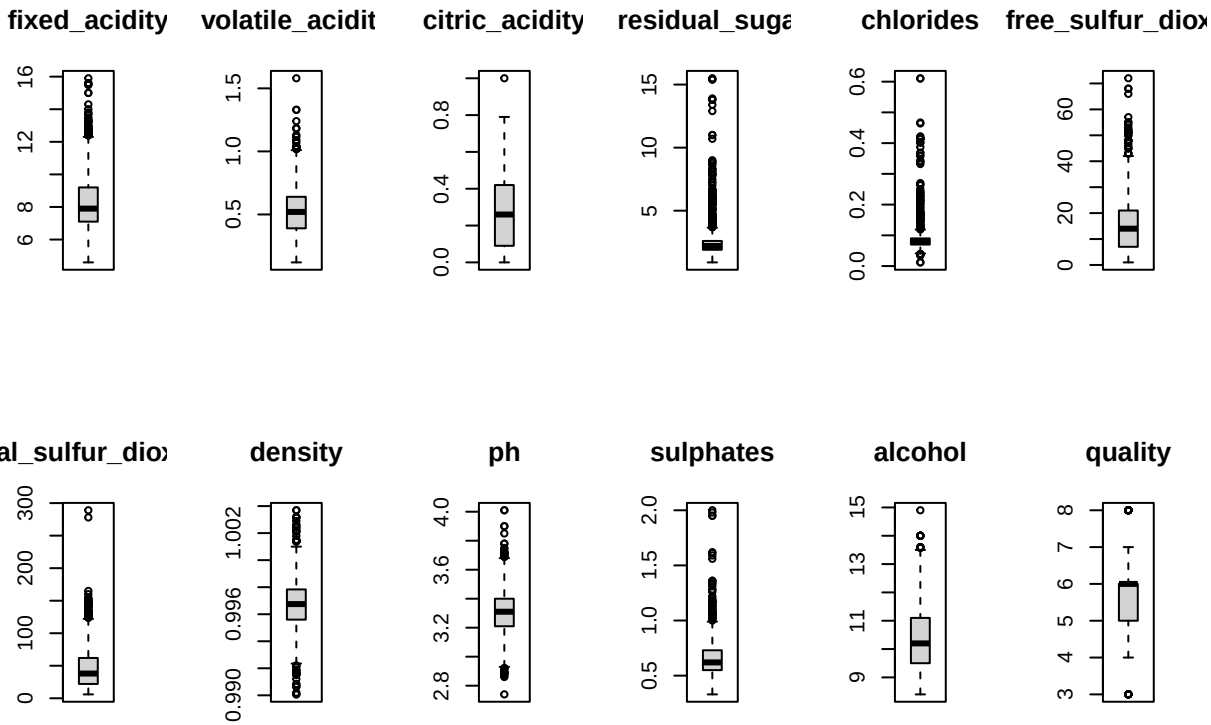
El que sí que veiem són registres duplicats. Si eliminem els duplicats, ens quedem amb 1359 registres, mostra suficient per resoldre la pràctica. De totes maneres en el nostre cas, optem per no eliminar aquests casos, ja que poden ser perfectament reals si provenen de medicions de la mateixa anyada, barrica, zona, etc... En tot cas, deixem el detall trobat:

```
wine_data_unique <- unique(wine_data)
str(wine_data_unique)
```

```
## 'data.frame':    1359 obs. of  12 variables:
##  $ fixed_acidity       : num  7.4 7.8 7.8 11.2 7.4 7.9 7.3 7.8 7.5 6.7 ...
##  $ volatile_acidity    : num  0.7 0.88 0.76 0.28 0.66 0.6 0.65 0.58 0.5 0.58 ...
##  $ citric_acidity      : num  0 0 0.04 0.56 0 0.06 0 0.02 0.36 0.08 ...
##  $ residual_sugar      : num  1.9 2.6 2.3 1.9 1.8 1.6 1.2 2 6.1 1.8 ...
##  $ chlorides           : num  0.076 0.098 0.092 0.075 0.075 0.069 0.065 0.073 0.071 0.097 ...
##  $ free_sulfur_dioxide : num  11 25 15 17 13 15 15 9 17 15 ...
##  $ total_sulfur_dioxide: num  34 67 54 60 40 59 21 18 102 65 ...
##  $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
##  $ ph                  : num  3.51 3.2 3.26 3.16 3.51 3.3 3.39 3.36 3.35 3.28 ...
##  $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.46 0.47 0.57 0.8 0.54 ...
##  $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 10 9.5 10.5 9.2 ...
##  $ quality             : int  5 5 5 6 5 5 7 7 5 5 ...
```

A més, si visualitzem les dades amb el boxplot per facilitar-ne la lectura, veiem que en general, no contenen valors molt dispersos, pero en el cas del sucre residual, el clorur i els sulfits es pot apreciar alguns valors diferenciats. Com en el cas anterior, optem per no modificar-los ja que les condicions ambientals del raïm o mosto, així com de la maduració del vi, les poden introduïr. Veiem-ho a continuació:

```
# Utilitzem invisible per mapejar cada registre, però evitant que printi per pantalla el resultat, ja q
par(mfrow = c(2, ncol(wine_data)/2 ))
invisible(
  lapply(
    1:ncol(wine_data), # per totes les columnes ja que volem veure tots els valors
    function(i)
      boxplot(
        wine_data[, i],
        main = colnames(wine_data[i])
      )
    )
  )
```



Afegim les mitges de cada dada:

```
# Mitjanes
mean(wine_data$fixed_acidity)
```

```
## [1] 8.319637
```

```
mean(wine_data$volatile_acidity)
```

```
## [1] 0.5278205
```

```
mean(wine_data$citric_acidity)
```

```
## [1] 0.2709756
```

```
mean(wine_data$residual_sugar)
```

```
## [1] 2.538806
```

```
mean(wine_data$chlorides)
```

```
## [1] 0.08746654
```

```
mean(wine_data$free_sulfur_dioxide)
```

```
## [1] 15.87492
```

```
mean(wine_data$total_sulfur_dioxide)
```

```
## [1] 46.46779
```

```
mean(wine_data$density)
```

```
## [1] 0.9967467
```

```
mean(wine_data$ph)
```

```
## [1] 3.311113
```

```
mean(wine_data$sulphates)
```

```
## [1] 0.6581488
```

```
mean(wine_data$alcohol)
```

```
## [1] 10.42298
```

```
mean(wine_data$quality)
```

```
## [1] 5.636023
```

3.3. Exportació de les dades netejades. Finalment, exportem les dades a un fitxer. Exportarem les dades de l'acidesa, així com les del dataset sencer i les del dataset sense duplicats.

```
write.csv(wine_data, "wine_data_all.csv")
write.csv(wine_data_acid, "wine_data_acid.csv")
write.csv(wine_data_unique, "wine_data_unique.csv")
```

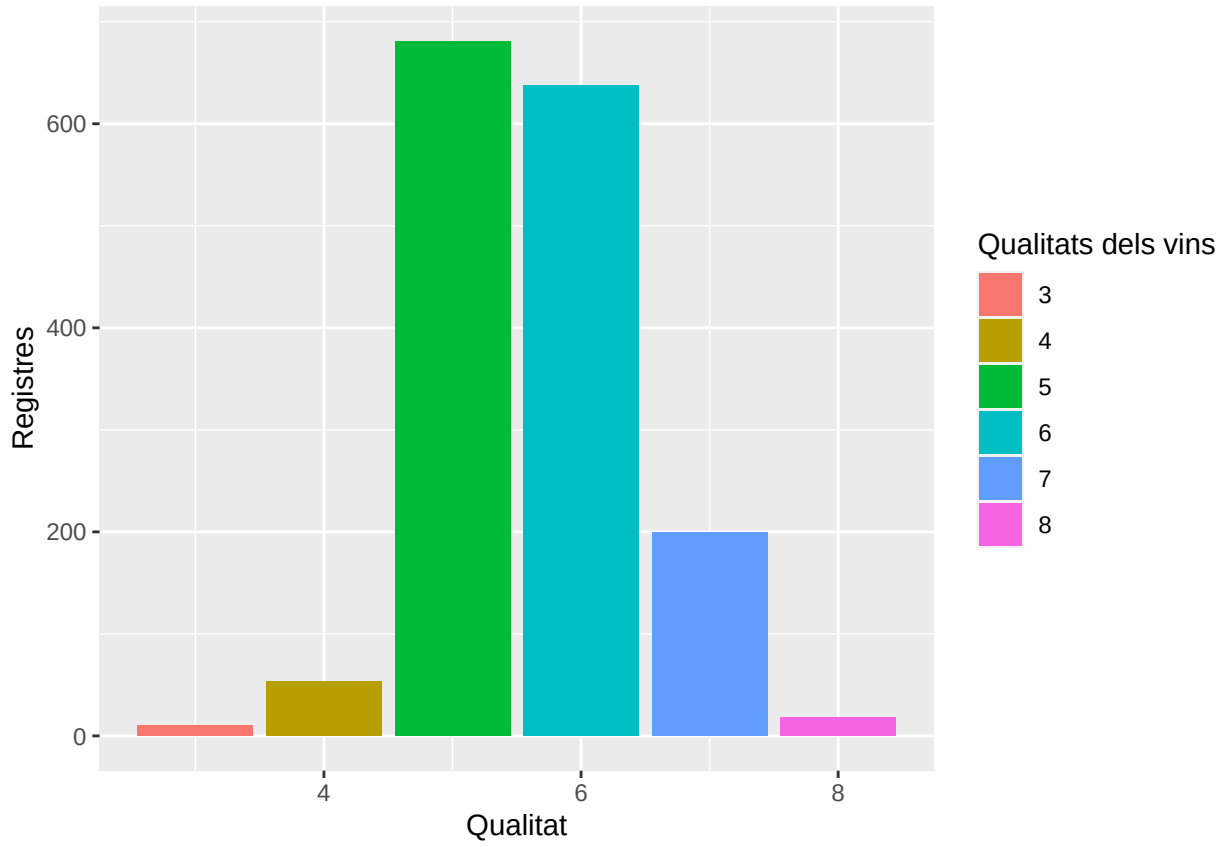
4. Anàlisi de les dades.

4.1. Selecció dels grups de dades que es volen analitzar/comparar (planificació dels anàlisis a aplicar). Anteriorment hem separat les dades que defineixen el component àcid del vi en el data frame "wine_data_acid" ja que hem volgut exportar el dataset per tenir-les netes en un fitxer.
Recordem que primer volem saber si l'acidesa és suficient per donar una puntuació al vi o pel contrari necessitem més dades. També volem saber quines propietats té el vi per rebre una alta puntuació i finalment, a partir d'aquestes dades, poder preveure futurs resultats. Per fer-ho primer haurem de:

1. Tractar l'acidesa individualment (ja hem generat un nou data frame amb les dades)
2. La correlació entre les característiques del vi que li otorguen una alta puntuació
3. Utilitzar models de regressió per predir futures puntuacions

Abans de res, hem de tenir en compte que la variable que ens interessa és la qualitat del vi, per tant, primer mirem quines són les dades que conté aquesta variable:

```
ggplot(wine_data, aes(x=quality, fill=as.factor(quality))) + geom_bar() + xlab("Qualitat") + ylab("Regis
```



Com podem veure la major quantitat de registres es concentren entre la puntuació de qualitat 5 i 6.

D'altra banda, a la definició del dataset hem definit una agrupació pel percentatge d'alcohol i la mateixa informació del dataset ens indica que totes les puntuacions majors que 6.5 són considerades bones.

Posem-ho en pràctica:

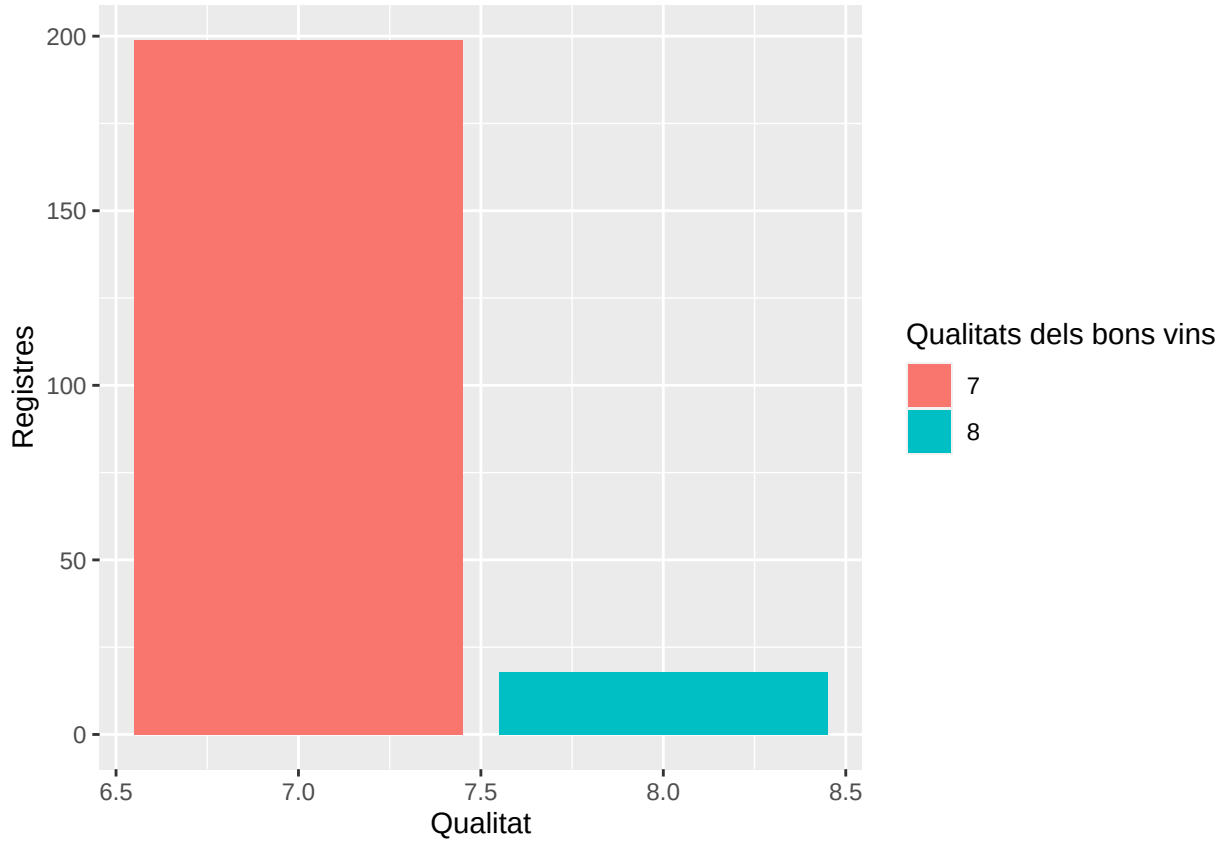
```
# Agrupació pel nivell de qualitat
wine_data.quality_good <- wine_data[wine_data$quality > 6.5, ]
wine_data.quality_bad <- wine_data[wine_data$quality <= 6.5 , ]

# Agrupació segons el tipus d'alcohol
wine_data.alcohol_very_low <- wine_data[wine_data$alcohol < 12.5, ]
wine_data.alcohol_low <- wine_data[wine_data$alcohol >= 12.5 & wine_data$alcohol < 13.5, ]
wine_data.alcohol_high <- wine_data[wine_data$alcohol >= 13.5 & wine_data$alcohol < 14.5, ]
wine_data.alcohol_very_high <- wine_data[wine_data$alcohol > 14.5, ]

# Si revisem les dades dels vins amb bona qualitat, obtenim que es mouen entre el 7 i 8 de la puntuació
nrow(wine_data.quality_good) # Total bons vins
```

```
## [1] 217
```

```
str(unique(wine_data.quality_good$quality)) # Només dos puntuacions
```

```
##  int [1:2] 7 8
```

```
# Mostreig
ggplot(wine_data.quality_good, aes(x=quality, fill=as.factor(quality))) + geom_bar() + xlab("Qualitat")
```



4.2. Comprovació de la normalitat i homogeneïtat de la variància. Abans de dur a terme les proves estadístiques, hem de revisar les dades per saber si tenim normalitat entre aquestes i si la variança és o no igual entre les dades a comparar. En el nostre cas, si la prova supera p-valor > 0.05 significarà que les dades (o la variable tractada), segueix una distribució normal, o en el cas contrari, no la segueix.

Per comprobar-ho, utilitzarem el test de Shapiro-Wilk per les variables que tenim dels vins:

```
par(mfrow=c(3, 4))

shapiro.test(wine_data$fixed_acidity)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  wine_data$fixed_acidity
## W = 0.94203, p-value < 2.2e-16
```

```
qqnorm(wine_data$fixed_acidity)
qqline(wine_data$fixed_acidity, col="green", lwd=2)

shapiro.test(wine_data$volatile_acidity)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  wine_data$volatile_acidity
## W = 0.97434, p-value = 2.693e-16
```

```
qqnorm(wine_data$volatile_acidity)
qqline(wine_data$volatile_acidity, col="green", lwd=2)

shapiro.test(wine_data$citric_acidity)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  wine_data$citric_acidity
## W = 0.95529, p-value < 2.2e-16
```

```
qqnorm(wine_data$citric_acidity)
qqline(wine_data$citric_acidity, col="orange", lwd=2)

shapiro.test(wine_data$residual_sugar)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  wine_data$residual_sugar
## W = 0.56608, p-value < 2.2e-16
```

```
qqnorm(wine_data$residual_sugar)
qqline(wine_data$residual_sugar, col="orange", lwd=2)

shapiro.test(wine_data$chlorides)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  wine_data$chlorides
## W = 0.48425, p-value < 2.2e-16
```

```
qqnorm(wine_data$chlorides)
qqline(wine_data$chlorides, col="green", lwd=2)

shapiro.test(wine_data$free_sulfur_dioxide)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  wine_data$free_sulfur_dioxide
## W = 0.90184, p-value < 2.2e-16
```

```
qqnorm(wine_data$free_sulfur_dioxide)
qqline(wine_data$free_sulfur_dioxide, col="orange", lwd=2)

shapiro.test(wine_data$total_sulfur_dioxide)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  wine_data$total_sulfur_dioxide
## W = 0.87322, p-value < 2.2e-16
```

```
qqnorm(wine_data$total_sulfur_dioxide)
qqline(wine_data$total_sulfur_dioxide, col="orange", lwd=2)

shapiro.test(wine_data$density)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  wine_data$density
## W = 0.99087, p-value = 1.936e-08
```

```
qqnorm(wine_data$density)
qqline(wine_data$density, col="green", lwd=2)

shapiro.test(wine_data$ph)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  wine_data$ph
## W = 0.99349, p-value = 1.712e-06
```

```
qqnorm(wine_data$ph)
qqline(wine_data$ph, col="green", lwd=2)

shapiro.test(wine_data$sulphates)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  wine_data$sulphates
## W = 0.83304, p-value < 2.2e-16
```

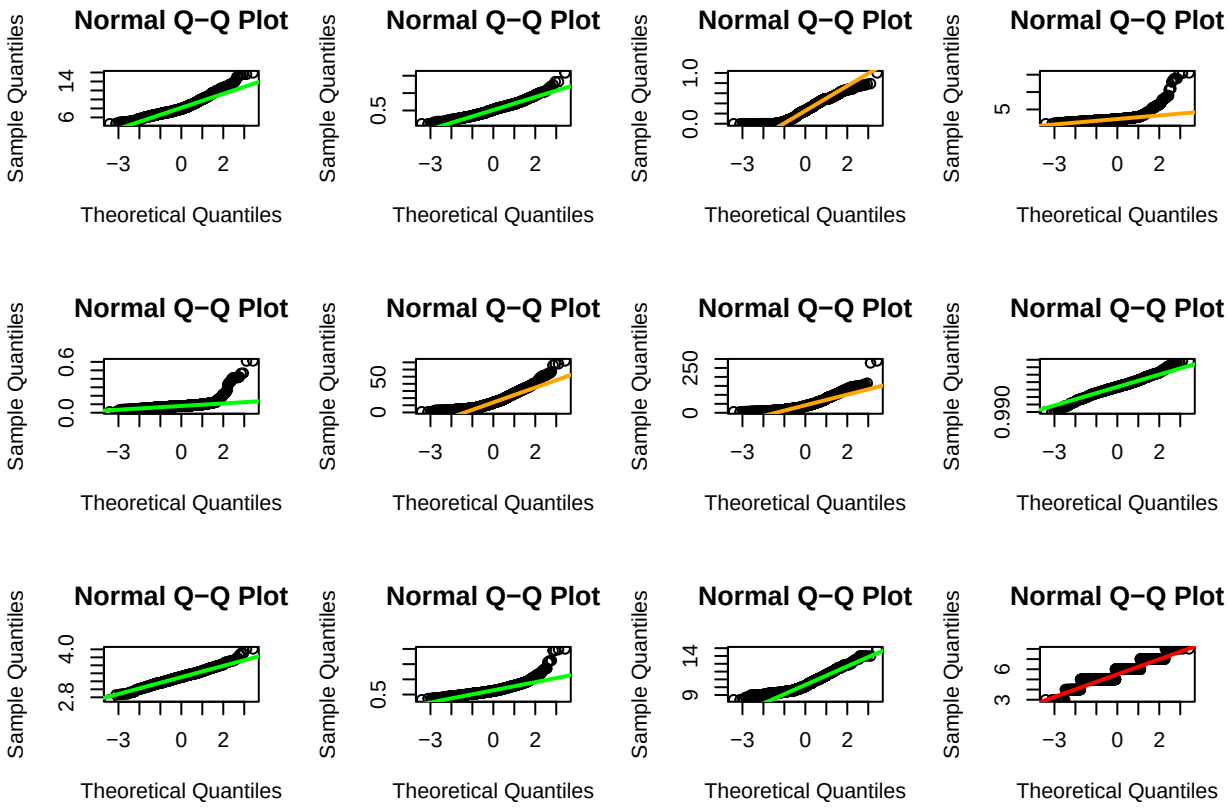
```
qqnorm(wine_data$sulphates)
qqline(wine_data$sulphates, col="green", lwd=2)

shapiro.test(wine_data$alcohol)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  wine_data$alcohol
## W = 0.92884, p-value < 2.2e-16
```

```
qqnorm(wine_data$alcohol)
qqline(wine_data$alcohol, col="green", lwd=2)

shapiro.test(wine_data$quality)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  wine_data$quality
## W = 0.85759, p-value < 2.2e-16
```

```
qqnorm(wine_data$quality)
qqline(wine_data$quality, col="red", lwd=2)
```



Com podem veure amb els resultats del test de Shapiro-Wilk obtenim que cap de les variables segueix una distribució normal, però si fem un qqplot dels quantils veiem que la gran majoria dels valors, menys a la variable "quality", tendeixen a seguir una distribució normal. Aquest comportament és possible en el testing d'hipòtesis (possible cas de teorema central del límit).

Com en principi a nivell de test numèric no es compleix la condició de normalitat de les mostres, a continuació farem el test de Fligner-Killeen (test no paramètric que es basa en la mitjana) per comprobar la homogeneïtat de la variança. La hipòtesis nula la definim com que les dos variances són iguals.

```
fligner.test(quality ~ fixed_acidity, data = wine_data)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  quality by fixed_acidity
## Fligner-Killeen:med chi-squared = 68.457, df = 95, p-value = 0.9818
```

```
fligner.test(quality ~ volatile_acidity, data = wine_data)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  quality by volatile_acidity
## Fligner-Killeen:med chi-squared = 147.35, df = 142, p-value = 0.3621
```

```
fligner.test(quality ~ citric_acidity, data = wine_data)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  quality by citric_acidity
## Fligner-Killeen:med chi-squared = 87.67, df = 79, p-value = 0.2362
```

```
fligner.test(quality ~ ph, data = wine_data)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  quality by ph
## Fligner-Killeen:med chi-squared = 86.558, df = 88, p-value = 0.5235
```

Com obtenim un valor superior al llindar de 0.05, podem considerar que les variables comparades són homogenies (Hem triat les referents a l'acidesa per tenir un conjunt del mateix context).

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. Correlació

La primera prova que farem serà la de correlació amb el test de pearson per tal de obtindre resposta a la primera pregunta que ens fèiem de si l'acidesa és suficient per donar la qualitat d'un vi.

```r
cor(wine_data_acid, method = "pearson")
```

```
##                   fixed_acidity volatile_acidity citric_acidity          ph
## fixed_acidity         1.0000000       -0.2561309      0.6717034 -0.68297819
## volatile_acidity     -0.2561309        1.0000000     -0.5524957  0.23493729
## citric_acidity        0.6717034       -0.5524957      1.0000000 -0.54190414
## ph                   -0.6829782        0.2349373     -0.5419041  1.00000000
## quality               0.1240516       -0.3905578      0.2263725 -0.05773139
##                       quality
## fixed_acidity      0.12405165
## volatile_acidity -0.39055778
## citric_acidity     0.22637251
## ph                -0.05773139
## quality            1.00000000
```

Com podem veure amb el data frame que ens haviem guardat abans, no podem assegurar una correlació entre la qualitat i les variables que representen l'acidesa.

Aprofitem per determinar si hi ha alguna variable que sí que estigui correlacionada amb la qualitat:

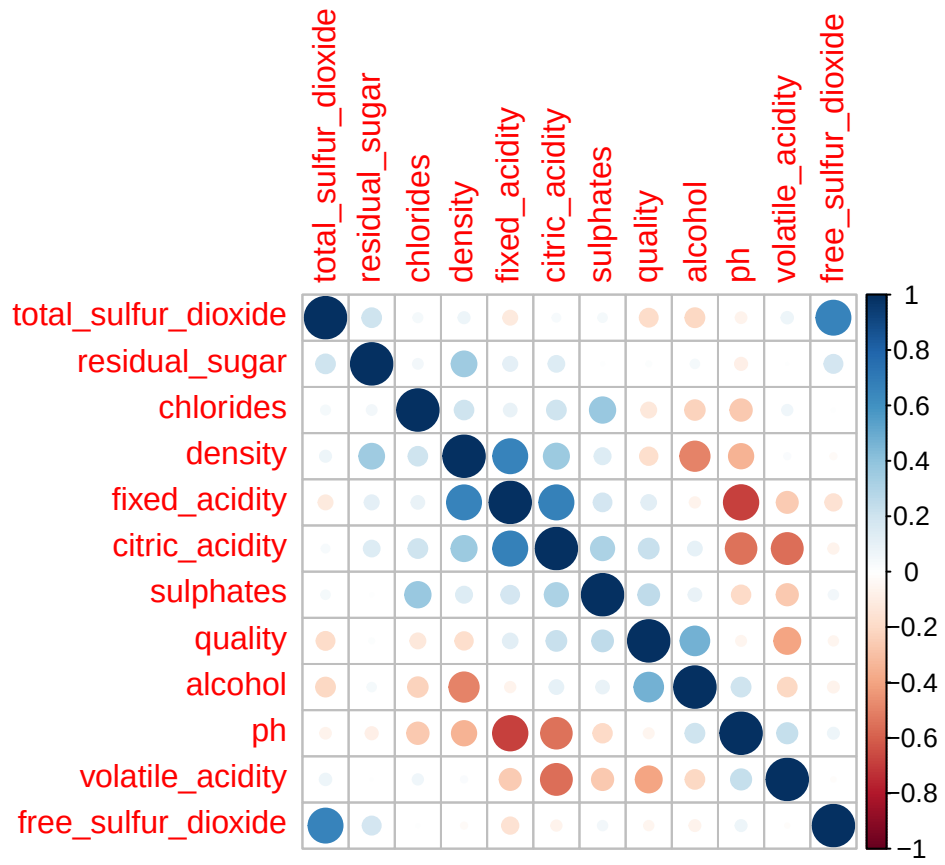
```r
(correlacio <- cor(wine_data, method = "pearson")) # Així la printem en una sola linia
```

```
##                       fixed_acidity volatile_acidity citric_acidity
## fixed_acidity           1.00000000     -0.256130895     0.67170343
## volatile_acidity       -0.25613089      1.000000000    -0.55249568
## citric_acidity          0.67170343     -0.552495685     1.00000000
## residual_sugar          0.11477672      0.001917882     0.14357716
## chlorides               0.09370519      0.061297772     0.20382291
## free_sulfur_dioxide    -0.15379419     -0.010503827    -0.06097813
## total_sulfur_dioxide   -0.11318144      0.076470005     0.03553302
## density                 0.66804729      0.022026232     0.36494718
## ph                     -0.68297819      0.234937294    -0.54190414
## sulphates               0.18300566     -0.260986685     0.31277004
## alcohol                -0.06166827     -0.202288027     0.10990325
## quality                 0.12405165     -0.390557780     0.22637251
##                       residual_sugar    chlorides free_sulfur_dioxide
## fixed_acidity           0.114776724  0.093705186        -0.153794193
## volatile_acidity        0.001917882  0.061297772        -0.010503827
## citric_acidity          0.143577162  0.203822914        -0.060978129
## residual_sugar          1.000000000  0.055609535         0.187048995
## chlorides               0.055609535  1.000000000         0.005562147
## free_sulfur_dioxide     0.187048995  0.005562147         1.000000000
## total_sulfur_dioxide    0.203027882  0.047400468         0.667666450
## density                 0.355283371  0.200632327        -0.021945831
## ph                     -0.085652422 -0.265026131         0.070377499
## sulphates               0.005527121  0.371260481         0.051657572
## alcohol                 0.042075437 -0.221140545        -0.069408354
## quality                 0.013731637 -0.128906560        -0.050656057
##                       total_sulfur_dioxide     density          ph    sulphates
## fixed_acidity                  -0.11318144  0.66804729 -0.68297819  0.183005664
## volatile_acidity                0.07647000  0.02202623  0.23493729 -0.260986685
## citric_acidity                  0.03553302  0.36494718 -0.54190414  0.312770044
## residual_sugar                  0.20302788  0.35528337 -0.08565242  0.005527121
```

```
## chlorides                      0.04740047  0.20063233 -0.26502613  0.371260481
## free_sulfur_dioxide            0.66766645 -0.02194583  0.07037750  0.051657572
## total_sulfur_dioxide           1.00000000  0.07126948 -0.06649456  0.042946836
## density                        0.07126948  1.00000000 -0.34169933  0.148506412
## ph                            -0.06649456 -0.34169933  1.00000000 -0.196647602
## sulphates                      0.04294684  0.14850641 -0.19664760  1.000000000
## alcohol                       -0.20565394 -0.49617977  0.20563251  0.093594750
## quality                       -0.18510029 -0.17491923 -0.05773139  0.251397079
##                          alcohol     quality
## fixed_acidity        -0.06166827  0.12405165
## volatile_acidity     -0.20228803 -0.39055778
## citric_acidity        0.10990325  0.22637251
## residual_sugar        0.04207544  0.01373164
## chlorides            -0.22114054 -0.12890656
## free_sulfur_dioxide  -0.06940835 -0.05065606
## total_sulfur_dioxide -0.20565394 -0.18510029
## density              -0.49617977 -0.17491923
## ph                    0.20563251 -0.05773139
## sulphates             0.09359475  0.25139708
## alcohol               1.00000000  0.47616632
## quality               0.47616632  1.00000000
```

Amb les dades en general, veiem que la variable "alcohol" és la que té més influència en la qualitat del vi, però no és del tot determinant, així que en general, podem afirmar que no hi ha cap variable directament correlacionada amb la qualitat.

```
corrplot(correlacio, method = 'circle', order = 'AOE')
```



Contrast d'hipòtesis

Seguint amb el alcohol, recordem que anteriorment hem fet una separació de les dades dels vins en funció de la qualitat (> 6.5). Bé, ara intentarem esbrinar si el grau d'alcohol és superior en els vins de més qualitat o pel contrari, en els de menys qualitat. Per fer-ho, utilitzarem un contrast d'hipòtesis on tindrem els dos conjunts de dades anteriorment descrits i farem servir la prova t de Student. Aquesta prova assumeix que les mitjanes dels dos grups són les mateixes i com tenim suficient dades per fer-ho en els dos conjunts, no caldrà utilitzar una prova no paramètrica.

Assumim p-valor $= 0.05$ pel contrast de la hipòtesis nula.

```
t.test(wine_data.quality_bad$alcohol, wine_data.quality_good$alcohol)
```

```
##
##  Welch Two Sample t-test
##
## data:  wine_data.quality_bad$alcohol and wine_data.quality_good$alcohol
## t = -17.45, df = 283.78, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.409927 -1.124097
## sample estimates:
## mean of x mean of y
##  10.25104  11.51805
```

Com podem veure doncs, la prova ens demostra que la hipòtesi nula no és certa i per tant podem afirmar que els vins de major qualitat tenen un grau d'alcohol més elevat que els de menys qualitat.

Models de regressió logística

Finalment, generarem models de regressió logística en funció de les variables més relacionades per tal de poder preveure un resultat utilitzant-los.

```
# Variables amb el coeficient de correlació més alt (les més relacionades amb la qualitat)
alcohol = wine_data$alcohol
sulfats = wine_data$sulphates
acidesa_citrica = wine_data$citric_acidity
acidesa_fixa = wine_data$fixed_acidity

# Altres variables
sucre_residual = wine_data$residual_sugar
cloridos = wine_data$chlorides
densidat = wine_data$density
ph = wine_data$ph
sense_sulfits = wine_data$free_sulfur_dioxide
total_sulfits = wine_data$total_sulfur_dioxide
acidesa_volatil = wine_data$volatile_acidity

# Volem predir la qualitat
qualitat = wine_data$quality

# Generació dels models lineals múltiples
model1 <- lm(qualitat ~ alcohol, data = wine_data)
model2 <- lm(qualitat ~ alcohol + sulfats + sucre_residual, data = wine_data)
model3 <- lm(qualitat ~ alcohol + acidesa_citrica + cloridos, data = wine_data)
model4 <- lm(qualitat ~ sulfats + acidesa_citrica + acidesa_fixa + densidat + ph + sense_sulfits, data =
model5 <- lm(qualitat ~ alcohol + sulfats + acidesa_citrica + acidesa_fixa + sucre_residual + cloridos +

# Generem un data frame per poder veure millor les dades
modelsLinealsMultiples <- data.frame(
  c(1,2,3,4,5),
  c(
    summary(model1)$r.squared,
    summary(model2)$r.squared,
    summary(model3)$r.squared,
    summary(model4)$r.squared,
    summary(model5)$r.squared
))
colnames(modelsLinealsMultiples)[1] = "ModelId"
colnames(modelsLinealsMultiples)[2] = "R^2"

modelsLinealsMultiples
```

```
##   ModelId       R^2
## 1       1 0.2267344
## 2       2 0.2699353
## 3       3 0.2619248
## 4       4 0.2240467
## 5       5 0.3589260
```

Com podem veure, no tenim un model amb coeficient de determinació admissible. Si fem una prova utilitzant el model 5 que és el més determinant de tots:

```
predict(model5, data.frame(
  alcohol = 14.0,
  sulfats = 0.39,
  acidesa_citrica = 0.03,
  acidesa_fixa = 5.3,
  sucre_residual = 6.2,
  cloridos = 0.042,
  densidat = 0.99300,
  ph = 3.0,
  total_sulfits = 90,
  acidesa_volatil = 0.25
))
```

```
##        1
## 6.894929
```

Podem extreure un resultat del model, però no és fiable.

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

Hem aplicat ajudes visuals per entendre els resultats mitjançant taules dels data frames i gràfiques en tot moment.

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

Aquesta pràctica ens havia de donar informació a tres preguntes inicials:

1. Aclarir si solament amb les dades de l'acidesa, es pot classificar el vi amb la seva puntuació correctament sense gaire marge d'error.

Comprobant la correlació de les variables hem pogut veure que no és possible otorgar una puntuació de qualitat únicament en funció de les variables relacionades amb l'acidesa.

2. Quina és la variable més relacionada amb la qualitat del vi.

L'alcohol és la variable més relacionada amb la qualitat del vi, no per molt, però ho és amb diferència respecte les demès variables. A més a més, com hem pogut veure, el contrast d'hipòtesis ens ha permès saber que els vins de major qualitat tenen una graduació d'alcohol més elevada que els de menys qualitat.

3. Poder realitzar models de regressió logística en funció de les variables mes relacionades.

Finalment hem generat cinc models per tal de poder preveure la relació entre les variables i la qualitat, però cap dels cinc ens permet donar un resultat determinant.

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.

Contribuciones	Firma
Investigació previa	Alejandro Heredia i Pere Garcia
Redacció de les respostes	Alejandro Heredia i Pere Garcia
Desenvolupament del codi	Alejandro Heredia i Pere Garcia