TIPOLOGÍA I CICLE DE VIDA DE LES DADES

PRÀCTICA 2

Tractament del Dataset “Red Wine Quality”

Extret de la Web https://www.kaggle.com

Noms dels estudiants: Xavier Badia Mulero i Josep Pou Mas

Resum

L'objectiu d'aquest treball és establir a través de tècniques de l'anàlisi de dades les variables més influents en la qualitat sensorial del vi. S'analitza una variable dependent (qualitat) i 10 variables independents: acidesa volàtil, diòxid de sofre total, densitat, pH, sulfats i alcohol. Els resultats mostren les variables més influents en la qualitat del vi: alcohol, pH, sulfats, acidesa cítrica i la relació alcohol i sulfats. Es conclou l'estudi que el control d'aquestes quatre variables és suficient per millorar la qualitat del vi. No obstant això, és necessari ampliar aquests estudis amb un espectre mostral més ampli.

Introducció

La denominació d'origen **Vinho Verde** va néixer el 1959 i es troba en el nord-est de Portugal, en una zona tradicionalment coneguda com a **Entre-Douro-e-Minho** (entre el Duero i el Miño). El territori del **Vinho Verde** s'estén cap al sud fins al riu Duero i les muntanyes de Freita, Arada i Montemuro; a l'est limita amb les muntanyes de Peneda Gerês, Cabreira i Marão; i a l'oest arriba fins a l'oceà Atlàntic. Així, la DO Viño Verde és la regió demarcada més gran de Portugal i una de les més grans de tota Europa.

Actualment s'han establert molts mecanismes per a determinar la qualitat sensorial del vi, sent en molts casos subjectius, amb el recolzament de persones expertes o proves molt costoses. La predicció de la qualitat del vi, de forma subjetiva es un problema que implica temps i personal especialitzat en tot allò que fa referència al coneixement del vi.

En aquest pràctica volem determinar la qualitat dels vins, utilitzant tècniques de mineria de dades. L'objectiu és utilitzar una estratègia intel·ligent per a resoldre aquest problema, que ha de permetre a les empreses del sector, una millor innovació, competitivitat i qualitat del vi. Aspecte, que pot afectar l'estat d'ànims i emocions dels consumidors.

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

Aquest dataset està format per una sèrie de valors que ens indiquen les característiques fisicoquímiques (com pot ser l'acidesa, el sucre, el pH, l'alcohol) després d'haver fet diferents anàlisis a les variants negres del vi portuguès "Vinho Verde".

Les dades del dataset tenen com a objectiu determinar qu característiques són les indicadores de vi negra de millor qualitat, i generar informació sobre cada un d'aquests factors per la qualitat del vi del nostre model.

Aquestes dades ens permetran crear diferents models per determinar com diferents variables independents ajuden a predir la nostra variable dependent, la qualitat. Saber com afectarà cada variable a la qualitat del vi negre ajudarà als productors, distribuïdors i empreses de la indústria del vi a avaluar millor la seva estratègia de producció, distribució i preus.

El dataset està format per 11 variables independents i una variable dependent anomenada quality, la qual és mesura en una escala del 0 al 10, on el 0 representa la qualitat més baixa i el 10 la qualitat més alta. La següent taula descriu les variables.

Variables independents	
Nom	**Descripció**
fixed acidity	Acidesa fixa garanteix una preservació, conservació i estabilitat del color, aromes i sabors del vi. Impedeix l'aparició d'agents microbians malignes pel motiu que és un àcid.
volatile acidity	Permet que els vins tinguin un aroma afruitat.
citric acid	En poques quantitats aporta una sensació de frescor i amargor. Té una funció estabilitzadora en el vi.
residual sugar	Sucre residual provinent dels sucres naturals del raïm que queda en el vi una vegada finalitzada la fermentació alcohòlica. Es mesura en grams per litre.
chlorides	Concentració de clorurs en vins vermells.
free sulfur dioxide	Diòxid de sofre lliure, és l'agent principal en la conservació de vins.
total sulfur dioxide	Diòxid de sofre total.
density	La densitat relativa és la relació entre la massa volúmica del vi i la massa volúmica de l'aigua. Aquesta determinació ens dóna idea del contingut en sucres i a més és un paràmetre que permet fer un seguiment de la fermentació alcohòlica.
pH	El pH és una mesura d'acidesa d'una solució. En el món del vi, el pH pot variar depenent de la maduració del raïm, de la varietat, de la concentració dels àcids orgànics en el moment de la collita, de la temperatura de fermentació, etc.
sulphates	Es generen de forma natural en el procés de fermentació dels llevats del vi. Tenen funcions de conservació, són antioxidants, antimicrobians.
alcohol	El sucre que conté el most de forma natural es converteix en alcohol etílic gràcies a l'acció dels llevats.
Variables dependents	

Nom	Descripció
quality (score between 0 and 10)	Variable de sortida que volem calcular en l'estudi, si és 7 o més gran (la qualitat del vi és 'bona') i si és menor de 7 (la qualitat és 'no bona').

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

Detallem a continuació la realització de la integració i l'exploració de les dades, per tal de crear una estructura de dades única i coherent, amb l'objectiu de presentar una millor informació pel nostre estudi. Analitzem, si cal reduir el nombre de variables o bé, crear de noves a través d'unes altres ja existents.

Integració → En el nostre cas, per dur a terme l'anàlisi de dades cal accedir al dataset ***winequality-red***, que conté les dades dels vins negres. No és necessari integrar cap informació més procedent d'altres datasets.

Selecció de dades d'interès a analitzar → El mesurament de cada una de les variables independents juga un paper molt important en la qualitat del vi, tal com es comenta en els punts següents.

- L'acidesa volàtil influeix de forma directa en la qualitat del vi, ja que com més petit és el seu valor, major és la qualitat del vi.
- En el cas de l'àcid cítric, produeix lleugerament el sabor amarg que caracteritza els bons vins.
- El sucre residual té una gran influència en el seu tast, ja que segons la quantitat de sucre residual en el vi pot ser sec, semisec, semidolç i dolç.
- L'elevació dels clorurs afecta la qualitat del vi.
- El diòxid de sofre lliure i total generen sulfats a conseqüència de la seva oxidació, això fa que la proporció de sulfats tingui una gran influència en la qualitat del vi.
- La densitat ens indica el gruix que es percep a la boca quan és prova un vi.
- El pH és un dels factors predominats a l'acidesa del vi i és un factor important en la qualitat del vi.
- Els graus de l'alcohol és un factor rellevant sent ideal que aquest es trobi entre els 10 i 14 graus.

Tal com s'ha comentat, la importància que té cadascuna envers la qualitat del vi fa que sigui important tractar totes aquestes variables com a dades d'interès a analitzar.

3. Neteja de les dades.

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

Dades que contenen zeros → En aquest cas trobem la variable de l'acidesa cítrica que presenta valors 0, en aquest cas hi ha **132** registres d'aquesta variable que presenten un 0.

```
sapply(data_wine, function(x) sum(x == 0))
```

```
##        fixed.acidity     volatile.acidity          citric.acid
##                    0                    0                  132
##       residual.sugar            chlorides  free.sulfur.dioxide
##                    0                    0                    0
## total.sulfur.dioxide              density                   pH
##                    0                    0                    0
##            sulphates              alcohol              quality
##                    0                    0                    0
```

Ara és important saber, si aquest valor 0 es fa servir per indicar o no l'absència de valors en aquesta variable. Analitzarem millor aquesta variable, per poder tenir un millor coneixement dels diferents valors que presenta.

L'acidesa cítrica es troba en petites quantitats (0 a 0,5 g/L-1). Els vins vermells generalment estan desproveïts d'aquesta acidesa, ja que els bacteris que originen la fermentació malolàctica també metabolitzen l'acidesa cítrica.

Per tant, aquesta variable pot presentar valors 0 que fan referència a valors del nivell d'acidesa que podem trobar al raïm.

Dades que contenen elements buits → Les variables del dataset d'entrada no presenta cap element buit. A la següent imatge podem veure el resultat 0 que ens indica que no ha trobat cap element buit.

```
# Nombre de valors desconeguts de les variable
sapply(data_wine, function(x) sum(is.na(x)))
```

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

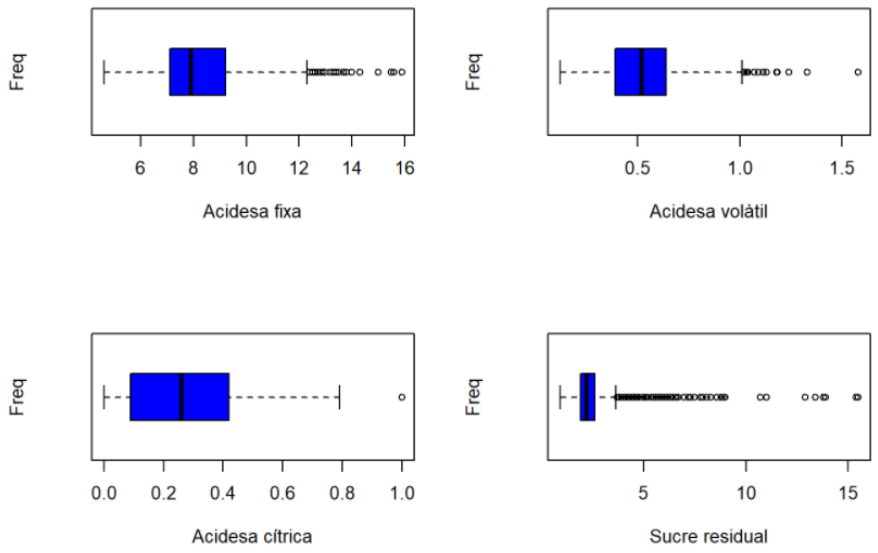
3.2. Identificació i tractament de valors extrems.

Per identificar els valors extrems de les variables utilitzem la representació de la distribució de les dades, hem optat per un diagrama de caixes per cada una de les variables del dataset d'entrada.

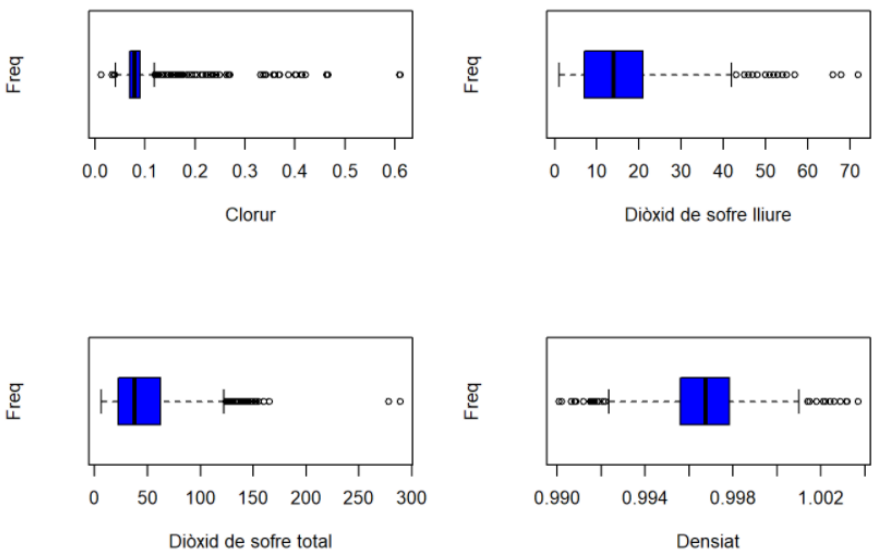
Hem de buscar aquelles dades que es troben molt allunyades de la distribució normal de les variables.

Les següents imatges podem veure la representació gràfica dels diagrames de caixes. El conjunt de dades de les variables presenten valors atípics per excés majors al tercer quartil (Q3) i menors al primer quartil (Q1).

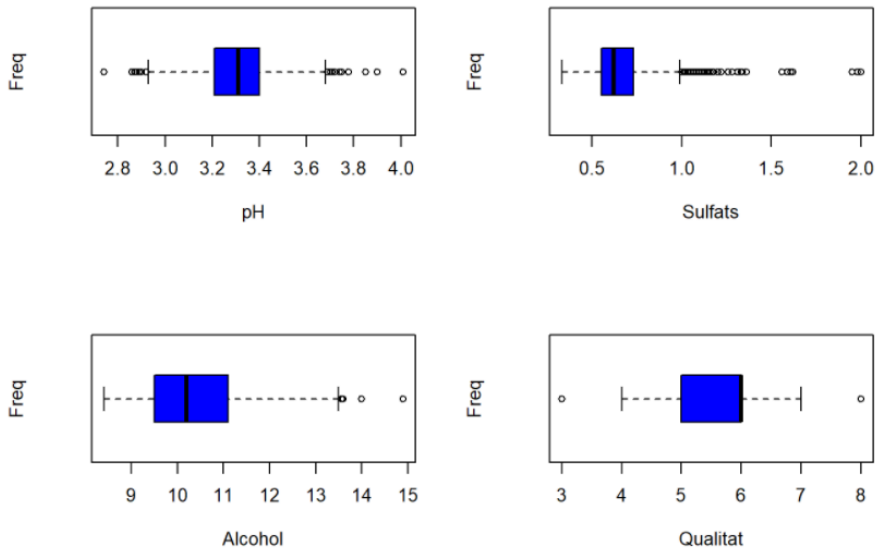
```
par(mfrow=c(2,2))

boxplot(data_wine$fixed.acidity, xlab="Acidesa fixa", col="blue", horizontal = TRUE, ylab = "Freq")
boxplot(data_wine$volatile.acidity, xlab="Acidesa volàtil", col="blue", horizontal = TRUE, ylab = "Freq")
boxplot(data_wine$citric.acid, xlab="Acidesa cítrica", col="blue", horizontal = TRUE, ylab = "Freq")
boxplot(data_wine$residual.sugar, xlab="Sucre residual", col="blue", horizontal = TRUE, ylab = "Freq")
```




```
boxplot(data_wine$chlorides, xlab="Clorur", col="blue", horizontal = TRUE, ylab = "Freq")
boxplot(data_wine$free.sulfur.dioxide, xlab="Diòxid de sofre lliure", col="blue", horizontal=TRUE, ylab="Freq")
boxplot(data_wine$total.sulfur.dioxide, xlab="Diòxid de sofre total", col="blue", horizontal=TRUE, ylab="Freq")
boxplot(data_wine$density, xlab="Densiat", col="blue", horizontal = TRUE, ylab = "Freq")
```



```
boxplot(data_wine$pH, xlab="pH" ,col="blue", horizontal = TRUE, ylab = "Freq")
boxplot(data_wine$sulphates, xlab="Sulfats", col="blue", horizontal = TRUE, ylab = "Freq")
boxplot(data_wine$alcohol, xlab="Alcohol", col="blue", horizontal = TRUE, ylab = "Freq")
boxplot(data_wine$quality, xlab="Qualitat", col="blue", horizontal = TRUE, ylab = "Freq")
```



Analitzem de forma detallada les variables més representatives a l'hora de valorar la qualitat del vi. Podem visualitzar de forma més clara els valors atípics si utilitzem la funció Boxplot.Stats. A continuació presentem dos casos, l'acidesa cítrica i el sulfat.

Acidesa cítrica

```
summary(data_wine$citric.acid)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.090   0.260   0.271   0.420   1.000
```

```
boxplot.stats(data_wine$citric.acid)$out
```

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

Aquesta variable presenta un valor extrem. El valor màxim respecte a la variable és 1, la mitja és 0,271. Els valors extrems són aquells valors que disten 3 cops el rang IQR per sobre de Q3. Aceptem aquest valor, ja que aquest tipus d'acidesa està present en vins i raïms en concentració entre 0,1 i 1 g/l. Per tant, aquest valor extrem es pot assumir com a correcte.

Sulfat

```
summary(data_wine$sulphates)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3300  0.5500  0.6200  0.6581  0.7300  2.0000
```

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

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

Aquesta variable presenta un total de 59 valors extrems. El valor màxim respecte a la variable és 2, la mitja és 0,6581. El valor extrem són aquells valors que disten 3 cops el rang IQR per sobre de Q3.

El sulfat s'utilitza com a conservant i antioxidant i és una forma d'assegurar la conservació del vi i l'eliminació de bacteris. El límit de sulfat en el vi és de 200 mg/l (2 g/l). Per tant, acceptem els valors extrems.

Els diferents elements tenen un impacte directe sobre la qualitat del vi, això ens indica que hem de tenir en compte els valors extrems de cadascuna de les variables que formen part del dataset. Per aquest motiu, decidim no menysprear cap valor.

4. Anàlisi de les dades.

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

El nostre dataset no presenta un gran nombre d'atributs, i no caldria aplicar cap dels mètodes de reducció de la dimensionalitat per obtenir una representació reduïda de les dades, com és el cas de l'anàlisi de **components principals (ACP)** que s'utilitzen per reduir el nombre d'atributs del nostre dataset de vins. També es podria fer servir mètodes de **clustering** o **sampling** per reduir el nombre de registres.

Construïm un joc de dades d'entrenament amb el 70% de registres per construir els nostres models i un joc de dades de proves o validacions amb el 30% de registres restants per validar els models. Aquesta separació d'ambdós conjunts és aleatòria.

```
# Dividim el fitxer en 70% entrenament i 30% validació

# Seed inicialitza el generador de nombres aleatoris que utilitzarem per separar lles dades en train i test.
# Utilitzant un seed fixe, ens assegurem de generar el mateix conjunt de dades i els resultats són reproduibles.

set.seed(1234)
index <- sample(2, nrow(data_wine), replace=TRUE, prob=c(0.7, 0.3))
train_data_wine <- data_wine[index==1,]
test_data_wine <- data_wine[index==2,]
```

Per poder realitzar l'**Homogeneïtat de la variància** i el **Contrast d'hipòtesis,** encarregats de validar la mitja de la qualitat del vi envers a la graduació de l'alcohol, crearem dos grups entorn

a la variable alcohol format pels vins de qualitat menys bona (menors del valor 7), i els vins de qualitat més bona (igual o major de 7).
Cal comentar que a part de l'aigua, el component més important de qualsevol vi és l'alcohol etílic o etanol, produït principalment pel llevat al transformar el sucre durant la fermentació. L'alcohol ha d'estar equilibrat amb la resta dels components del vi.

```
vi_qualitat_alta <- data_wine$alcohol[data_wine$quality >= 7]
vi_qualitat_baixa <- data_wine$alcohol[data_wine$quality < 7]
```

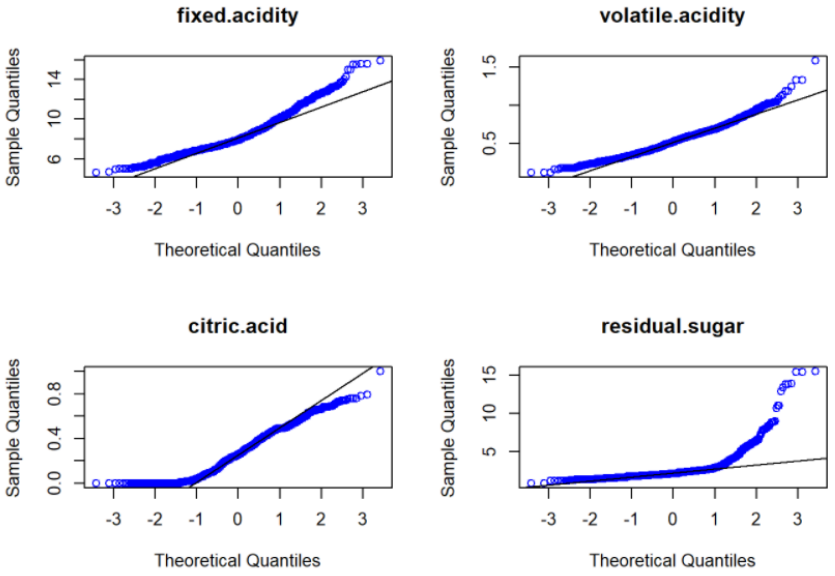
4.2. Comprovació de la normalitat i homogeneïtat de la variància.

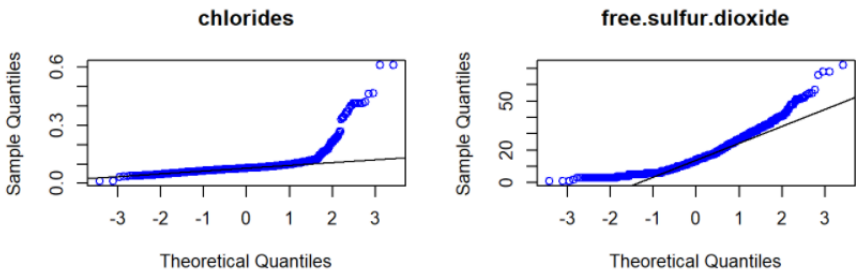
Comprovació de la normalitat:

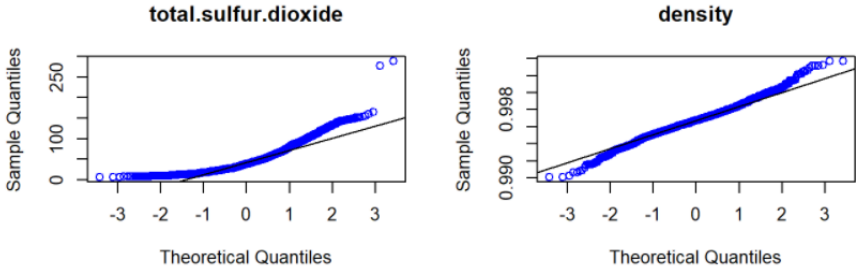
Farem la comprovació de la normalitat de dues maneres, mitjançant inspecció visual i aplicant el test de **Shapiro-Wilk** a cada variable.

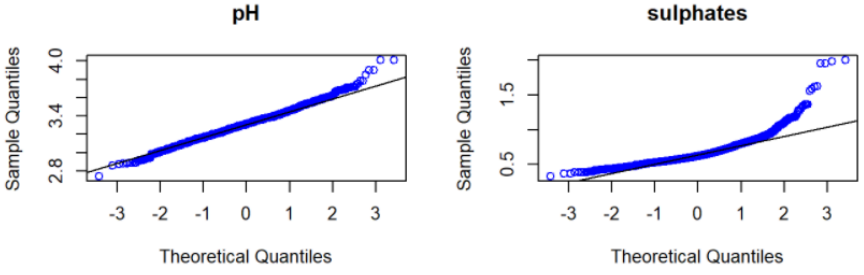
Inspecció visual: Utilitzarem el gràfic Q-Q o gràfic de quantils teòrics. El que fa és comparar els quantils de la distribució observada amb els quantils teòrics d'una distribució normal, de manera que com més s'aproximen les dades a una normal, més alineats es mostren els seus punts a la recta.

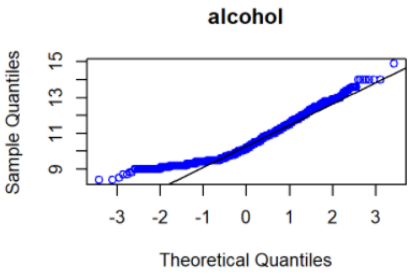
```
for(variable in variables){
  qqnorm(data_wine[,variable], main = paste0(variable), col = "blue")
  qqline(data_wine[,variable])
}
```











Test Shapiro-Wilk: Aplicarem aquest test com a segona comprovació, en el que si el p-valor resultant és més petit que el nivell de significació, generalment 0,05, llavors la hipòtesi nul·la és rebutjada i es conclou que les dades no segueixen una distribució normal.

```
variables <- colnames(data_wine)[1:11]

par(mfrow = c(2,2))

for(variable in variables){
  print(paste0("Nom de la variable: ", variable))
  print(paste0("P-valor: ", shapiro.test(data_wine[,variable])$p.value))
}
```

```
## [1] "Nom de la variable: fixed.acidity"
## [1] "P-valor: 1.52501179295091e-24"
## [1] "Nom de la variable: volatile.acidity"
## [1] "P-valor: 2.69293489456032e-16"
## [1] "Nom de la variable: citric.acid"
## [1] "P-valor: 1.02193162131975e-21"
## [1] "Nom de la variable: residual.sugar"
## [1] "P-valor: 1.02016171149076e-52"
## [1] "Nom de la variable: chlorides"
## [1] "P-valor: 1.17905575371677e-55"
## [1] "Nom de la variable: free.sulfur.dioxide"
## [1] "P-valor: 7.69459692029225e-31"
## [1] "Nom de la variable: total.sulfur.dioxide"
## [1] "P-valor: 3.57345139578549e-34"
## [1] "Nom de la variable: density"
## [1] "P-valor: 1.93605282884883e-08"
## [1] "Nom de la variable: pH"
## [1] "P-valor: 1.71223728301906e-06"
## [1] "Nom de la variable: sulphates"
## [1] "P-valor: 5.82314039765996e-38"
## [1] "Nom de la variable: alcohol"
## [1] "P-valor: 6.64405672007326e-27"
```

A partir dels resultats de les gràfiques Q-Q podem veure que la distribució de les dades de la majoria de les variables s'aproxima a la recta de quantils teòrica, tot i que visualment no podem prendre una decisió amb claredat si segueixen una distribució normal.

Amb la segona comprovació, aplicant el test Shapiro-Wilk, podem veure que el resultat del p-valor de cada variable és inferior al nivell de significació, generalment 0,05, per tant, podem rebutjar la hipòtesi nul·la del test, entenent que les variables del nostre dataset no segueixen una distribució normal.

Tanmateix, ja que tenim una mostra prou gran (1.599 observacions), aplicant el teorema central del límit, la mitjana d'una mostra és cada vegada més normal a mesura que augmenta la quantitat d'observacions.

Homogeneïtat de la variància:

Per comprovar l'homogeneïtat de la variància entre dues mostres, aplicarem la funció que en ens ve donada en R **var.test**

```
var.test(x = vi_qualitat_alta, y = vi_qualitat_baixa)
```

```
##
##  F test to compare two variances
##
## data:  vi_qualitat_alta and vi_qualitat_baixa
## F = 1.0596, num df = 216, denom df = 1381, p-value = 0.5562
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.8713587 1.3083664
## sample estimates:
## ratio of variances
##           1.059625
```

Podem observar que el **ratio of variances** té el valor 1,059625 i cau dintre de l'interval de confiança del 95% (rang entre 0,871 i 1,3083). També es pot observar que el **p-value** té el valor 0,5562 i és més gran de 0,05, per tant, no hi ha diferència significativa entre les variàncies dels dos grups.

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.

Contrast d'hipòtesis:
A continuació, fem un contrast d'hipòtesis a partir de la pregunta següent:

Podem acceptar que els vins de bona qualitat (Quality >= 7) tenen en mitjana una graduació d'alcohol superior que els de menys qualitat (Quality <7)?

H_0: Els vins de bona qualitat tenen en mitjana una graduació d'alcohol igual als vins de menys qualitat.

H_1: Els vins de bona qualitat tenen en mitjana una graduació d'alcohol superior als vins de menys qualitat.

H_0: $\mu 0 - \mu 1 = 0$, o també es pot expressar com $\mu 0 = \mu 1$
H_1: $\mu 0 - \mu 1 > 0$, o també es pot expressar com $\mu 0 > \mu 1$

On $\mu 0$ és la mitjana de graduació d'alcohol dels vins de bona qualitat i $\mu 1$ és la mitjana de graduació d'alcohol dels vins de menys qualitat.

En funció dels resultats obtinguts anteriorment i la pregunta que ens plantegem, aplicarem un contrast d'hipòtesis de dues mostres independents sobre la mitjana amb igualtat de variàncies amb un nivell de confiança del 95%

```
# t de Student
t.test(vi_qualitat_alta, vi_qualitat_baixa, # les dos mostres
       alternative = "greater",  # contrast per resta de mitjanes
       paired = FALSE,  # mostres independents
       var.equal = TRUE, #
       conf.level = 0.95)
```

```
##
##  Two Sample t-test
##
## data:  vi_qualitat_alta and vi_qualitat_baixa
## t = 17.823, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  1.150012      Inf
## sample estimates:
## mean of x mean of y
##  11.51805  10.25104
```

El resultat del contrast d'hipòtesis dóna un **p-valo**r més petit que 0.05, que és el nivell d'acceptació, i també veiem que no cau dintre de l'interval d'acceptació de la hipòtesi nul·la, llavors ens porta a rebutjar la hipòtesi nul·la, per tant, acceptem la hipòtesi alternativa i podem dir que els vins de bona qualitat tenen en mitjana una graduació superior que els vins de més baixa qualitat.

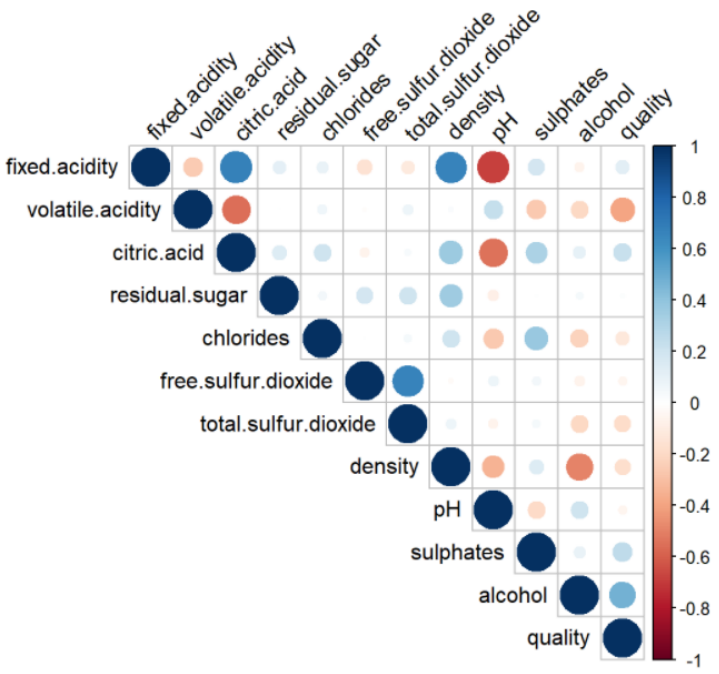
Correlacions

Per calcular les correlacions entre les variables utilitzem la funció anomenada **cor()**, i ho representem de forma gràfica amb la funció **corrplot()**.

```
rcor <- cor(data_wine, use="pairwise.complete.obs", method = "pearson")
rcor
```

```
                      fixed.acidity volatile.acidity citric.acid residual.sugar    chlorides free.sulfur.dioxide total.sulfur.dioxide     density          pH    sulphates
fixed.acidity           1.00000000     -0.256130895  0.67170343     0.114776724  0.093705186        -0.153794193          -0.11318144  0.66804729 -0.68297819  0.183005664
volatile.acidity       -0.25613089      1.000000000 -0.55249568     0.001917882  0.061297772        -0.010503827           0.07647000  0.02202623  0.23493729 -0.260986685
citric.acid             0.67170343     -0.552495685  1.00000000     0.143577162  0.203822914        -0.060978129           0.03553302  0.36494718 -0.54190414  0.312770044
residual.sugar          0.11477672      0.001917882  0.14357716     1.000000000  0.055609535         0.187048995           0.20302788  0.35528337 -0.08565242  0.005527121
chlorides               0.09370519      0.061297772  0.20382291     0.055609535  1.000000000         0.005562147           0.04740047  0.20063233 -0.26502613  0.371260481
free.sulfur.dioxide    -0.15379419     -0.010503827 -0.06097813     0.187048995  0.005562147         1.000000000           0.66766645 -0.02194583  0.07037750  0.051657572
total.sulfur.dioxide   -0.11318144      0.076470005  0.03553302     0.203027882  0.047400468         0.667666450           1.00000000  0.07126948 -0.06649456  0.042946836
density                 0.66804729      0.022026232  0.36494718     0.355283371  0.200632327        -0.021945831           0.07126948  1.00000000 -0.34169933  0.148506412
pH                     -0.68297819      0.234937294 -0.54190414    -0.085652422 -0.265026131         0.070377499          -0.06649456 -0.34169933  1.00000000 -0.196647602
sulphates               0.18300566     -0.260986685  0.31277004     0.005527121  0.371260481         0.051657572           0.04294684  0.14850641 -0.19664760  1.000000000
alcohol                -0.06166827     -0.202288027  0.10990325     0.042075437 -0.221140545        -0.069408354          -0.20565394 -0.49617977  0.20563251  0.093594750
quality                 0.12405165     -0.390557780  0.22637251     0.013731637 -0.128906560        -0.050656057          -0.18510029 -0.17491923 -0.05773139  0.251397079
                         alcohol      quality
fixed.acidity         -0.06166827  0.12405165
volatile.acidity      -0.20228803 -0.39055778
citric.acid            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
```

```
corrplot(rcor, type = "upper", tl.col = "black", method = "circle", tl.srt = 45)
```



En el gràfic de correlacions resultant, els cercles més grans i de més intensitat ens indiquen un grau de correlació més alt entre les dues variables i els més petits i de menys intensitat un grau de correlació més baix.

Es pot observar en aquest gràfic que la correlació més gran que hi ha en referència a la variable Quality és amb les variables alcohol i acidesa volàtil.

Regressió lineal múltiple

Seguidament aplicarem diversos models de regressió lineal, en concret en farem cinc, que aproximi la relació de dependència lineal entre la variable Quality i les variables que anirem introduint en els models que seran l'alcohol, l'acidesa volàtil, l'àcid cítric, el pH i els sulfats que són les variables que tenen una correlació més forta amb la variable Quality.

```
rl1 <- lm(quality ~ alcohol, data = data_wine)
rl2 <- lm(quality ~ alcohol + volatile.acidity, data = data_wine)
rl3 <- lm(quality ~ alcohol + volatile.acidity + citric.acid, data = data_wine)
rl4 <- lm(quality ~ alcohol + volatile.acidity + citric.acid + pH, data = data_wine)
rl5 <- lm(quality ~ alcohol + volatile.acidity + citric.acid + pH + sulphates, data = data_wine)
```

Pels models de regressió lineal múltiple obtinguts, podem utilitzar el coeficient de determinació per a mesurar la bondat dels ajustos i quedar-nos amb aquell model que millor coeficient presenta.

```
# Taula dels coeficients de determinació de cada model de regressió
coeficients <- matrix(c(1, summary(rl1)$r.squared,
                        2, summary(rl2)$r.squared,
                        3, summary(rl3)$r.squared,
                        4, summary(rl4)$r.squared,
                        5, summary(rl5)$r.squared),

ncol = 2, byrow = TRUE)
colnames(coeficients) <- c("model regresió lineal", "Coeficient de determinació")
coeficients
```

```
##      model regresió lineal Coeficient de determinació
## [1,]                     1                  0.2267344
## [2,]                     2                  0.3170024
## [3,]                     3                  0.3171882
## [4,]                     4                  0.3236633
## [5,]                     5                  0.3413432
```

Veient el resultat dels models de regressió lineal, en els que hem anat afegint una variable cada vegada, no s'aprecia un increment significatiu en el **coeficient de determinació (R-squared)** resultant. Es pot veure també que els coeficients de determinació mantenen uns valors baixos en tots els models realitzats, es conclou que les variables es correlacionen amb poca força amb l'atribut Quality.

Random forest

Passem a crear una nova variable data_wine$quality que ha d'emmagatzemar l'agrupació dels vins segons la seva qualitat: Bo, Normal i Dolent. A continuació, dividim les dades en dos dataframe nous: un amb les dades d'entrenament (70% de les dades del dataset), i les dades de test (30% de les dades del dataset).

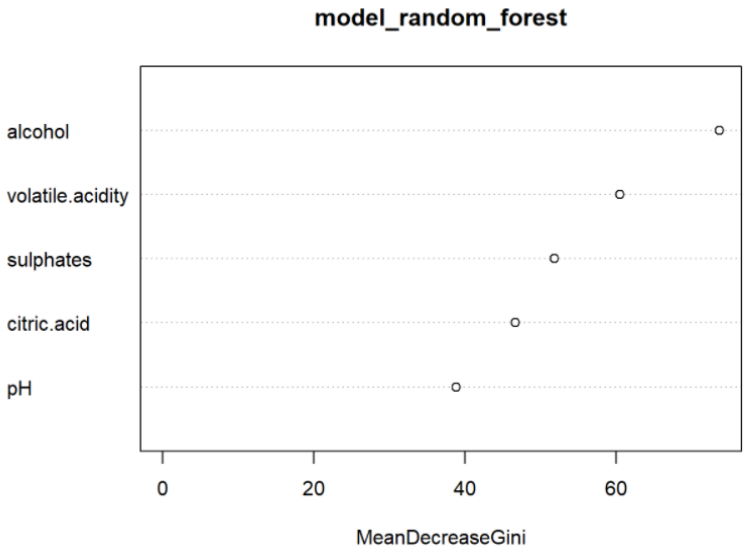
```
# Agrupem la qualitat dels vins a una nova variable
data_wine$nivell.qualitat <- ifelse(data_wine$quality < 5, "Dolent", "Bo")
data_wine$nivell.qualitat[data_wine$quality == 5 | data_wine$quality == 6] <- "Normal"
data_wine$nivell.qualitat <- as.factor(data_wine$nivell.qualitat)

# Convertim en factor la variable creada
data_wine$nivell.qualitat <- as.factor(data_wine$nivell.qualitat)

# Seed inicialitza el generador de nombres aleatoris que utilitzarem per separar les dades en train i t
est.
# Utilitzant un seed fixe, ens assegurem de generar el mateix conjunt de dades i els resultats són repr
oduïbles
set.seed(1234)

# Dividim el dataframe en 70% entrenament i 30% validació
index <- sample(2, nrow(data_wine), replace=TRUE, prob=c(0.7, 0.3))
train_data_wine <- data_wine[index==1,]
test_data_wine <- data_wine[index==2,]
```

Seguidament, utilitzem el mètode supervisat de classificació randomForest que a partir dels resultats obtinguts mitjançant el càlcul de n arbres, construeix el resultat de l'estimació que volem obtenir.

```
set.seed(12)

#Declaram la funció de l'arbre
ArbrelRF <- nivell.qualitat ~ alcohol + volatile.acidity + citric.acid + pH + sulphates

# #Apliquem l'algoritme
model_random_forest <- randomForest(ArbrelRF, data = train_data_wine, ntree=100, proximity= T, nodesize
=5) #indiquem el nombre d'arbres mitjançant ntree = 500

# #Obtenim la importancia de cada variable en el procés de clasificació
importance(model_random_forest)      #Importancia de las variables en formato text
```

```
##                 MeanDecreaseGini
## alcohol                 73.61480
## volatile.acidity        60.50129
## citric.acid             46.63158
## pH                      38.84138
## sulphates               51.83553
```

```
varImpPlot(model_random_forest) #Importancia de las variables en formato gráfico
```




```
# Validem la capacitat de predicció de l'arbre amb el fitxer de validació.
preds_random_forest <- predict(model_random_forest, newdata = test_data_wine)
table(preds_random_forest, test_data_wine$nivell.qualitat)
```

```
##
## preds_random_forest  Bo Dolent Normal
##              Bo      27      0     13
##              Dolent   0      1      0
##              Normal  33     18    366
```

```
# Calculem el % d'encerts
sum(preds_random_forest == test_data_wine$nivell.qualitat)/ length(test_data_wine$nivell.qualitat) * 100
```

```
## [1] 86.0262
```

L'arbre de decisió obtingut mitjançant el paquet **randomForest** classifica correctament un 86,02% dels registres. Un resultat no massa alt però acceptable.

Arbre de decisions

Per a construir un arbre de decisions és necessari definir una funció que relacioni una variable categòrica dependent (factor) amb n variables que poden ser categòriques o numèriques. En el nostre cas treballarem amb:

1 variable factor dependent -> nivell.qualitat

5 variables independents -> alcohol, volatile.acidity, citric.acid, pH, sulphates

L'algoritme de classificació busca quina és la variable que permet obtenir una submostra més diferenciada per la variable dependent (nivell.qualitat), i identificar també quins intervals (si la variable és quantitativa) o agrupació de categories de les variables independents permetran maximitzar aquesta divisió.

```r
# Declarem la funció de l'arbre
ArbreRpart = nivell.qualitat ~ alcohol + volatile.acidity + citric.acid + pH + sulphates
#Apliquem l'algoritme
model_tree = rpart(ArbreRpart, method="class", data=train_data_wine)
# Validem la capacitat de predicció de l'arbre amb el fitxer de validació
preds_tree <- predict(model_tree, newdata = test_data_wine, type = "class")
# Visualitzem una matriu de confusió
table(preds_tree, test_data_wine$nivell.qualitat)
```

```
##
## preds_tree  Bo Dolent Normal
##     Bo      28      0     16
##     Dolent   0      0      0
##     Normal  32     19    363
```

```r
# Calcula el % d'encerts
sum(preds_tree == test_data_wine$nivell.qualitat)/ length(test_data_wine$nivell.qualitat)*100
```

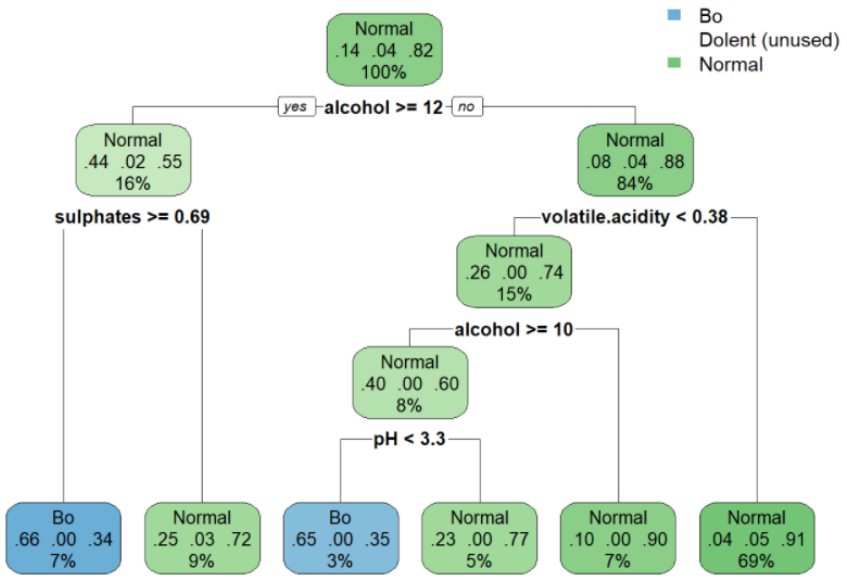
```
## [1] 85.37118
```

El model obtingut amb el paquet rpart presenta una classificació d'un 85,37% dels registres, un resultat alt i acceptable. Ha realitzat una predicció de 28 vins com a bons, i 32 com a vins normals.

5. Representació dels resultats a partir de taules i gràfiques

Arbre de decisions

```
rpart.plot(model_tree)
```



Podem observar que la quantitat d'alcohol és el primer node que realitza la partició, això indica que aquesta variable té més importància a l'hora de predir els vins de més qualitat. El primer node ens indica que el 14% són dades amb el valor Bo, el 4% dades amb valor Dolent i el 82% dades que presenten el valor Normal. També ens indica que hi ha el 100% de les dades.

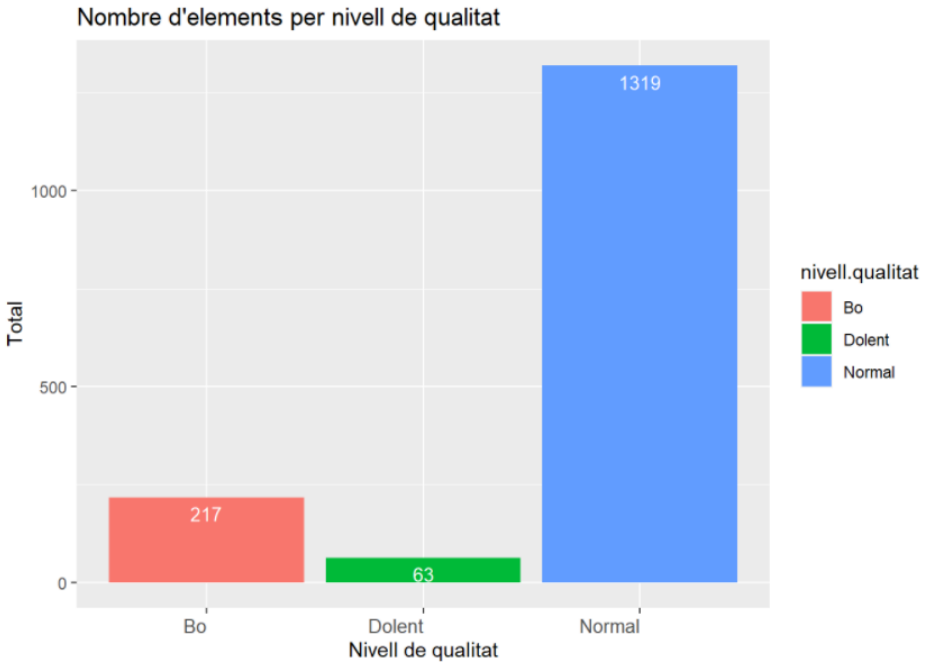
Altres variables importants són els sulfats, on el node de la seva partició presenta una totalitat del 16% de les dades del dataset, d'aquestes el 44% tenen un valor Bo, un 2% valor Dolent i un 55% com a Normals.

En total tenim un resultat d'un 10% de vins catalogats com a Bo, dels quals el 7% presenten un alcohol més gran de 12 i un sulfat més gran o igual a 0,69. El 3% restant, presenten un alcohol entre 10 i 12, una acidesa volàtil més petita de 0,38 i el valor del pH és més petit de 3,3.

Gràfiques

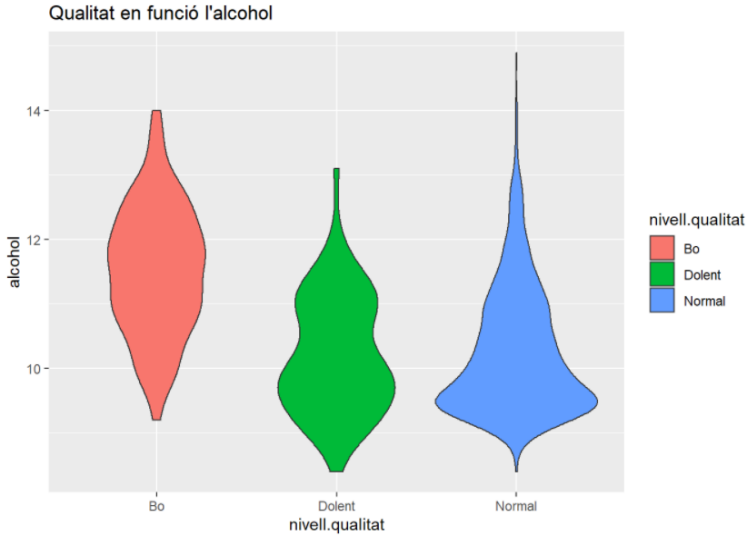
Nombre de registres per nivell de qualitat

```
n_nivell.qualitat <- dplyr::count(data_wine, nivell.qualitat)

ggplot(data = n_nivell.qualitat, aes(nivell.qualitat , n, fill = nivell.qualitat)) +
  geom_bar(stat="identity") +
  geom_text(aes(label=n), vjust=1.6, color="white",position = position_dodge(0.9), size=3.5) +
  xlab("Nivell de qualitat") + ylab("Total") + ggtitle("Nombre d'elements per nivell de qualitat") +
  theme(axis.text.x = element_text(angle = 0, size = 10, hjust = 1, vjust = 1))
```



Qualitat en funció de l'alcohol

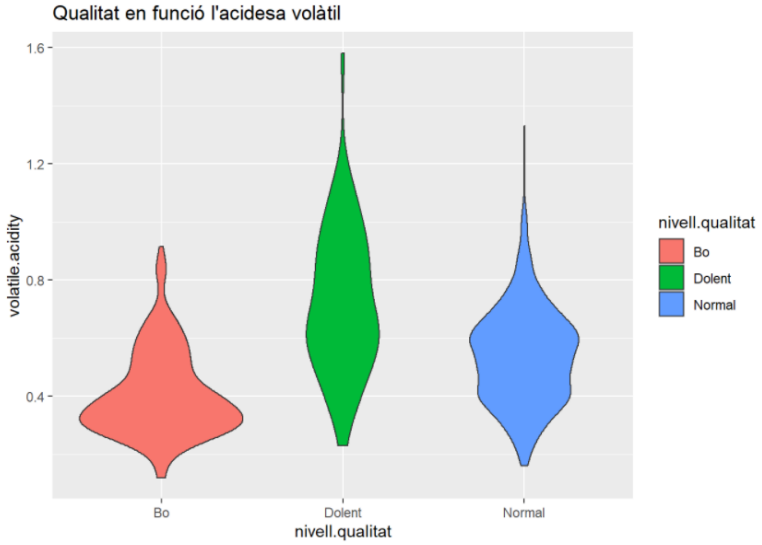
```
ggplot(data_wine, aes(x=nivell.qualitat, y=alcohol, fill=nivell.qualitat)) +
  ggtitle("Qualitat en funció l'alcohol") +
  geom_violin()
```



Segons els experts, el grau d'alcohol no és un condicionant de la qualitat del vi, no obstant això, en aquesta ocasió podem veure que els vins de qualitat alta tendeixen a tenir un grau d'alcohol major, però estan distribuïts en un rang aproximadament entre 9º i 14º.

Qualitat en funció de l'acidesa volàtil

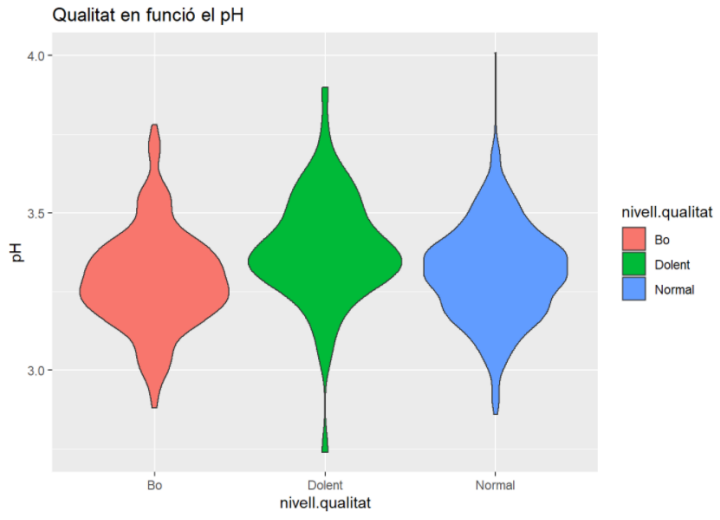
```
ggplot(data_wine, aes(x=nivell.qualitat, y=volatile.acidity, fill=nivell.qualitat)) +
  ggtitle("Qualitat en funció l'acidesa volàtil") +
  geom_violin()
```



L'acidesa volàtil afecta el sabor del vi, per aquest motiu és important que el seu valor sigui baix. Podem veure que els vins de qualitat més alta (Bo) tendeixen a tenir una acidesa volàtil més baixa.

Qualitat en funció del pH

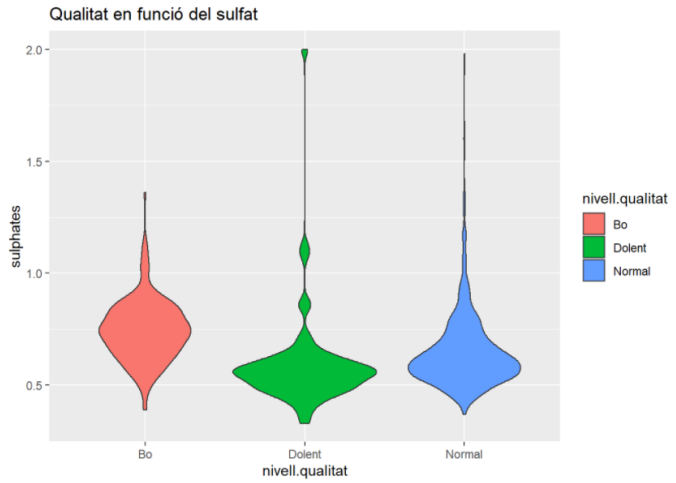
```
ggplot(data_wine, aes(x=nivell.qualitat, y=pH, fill=nivell.qualitat)) +
  ggtitle("Qualitat en funció el pH") +
  geom_violin()
```



En aquest gràfic podem veure com els vins etiquetats com a Bo tendeixen a presentar més dades al voltant del valor de pH 3,25.

Qualitat en funció del sulfat

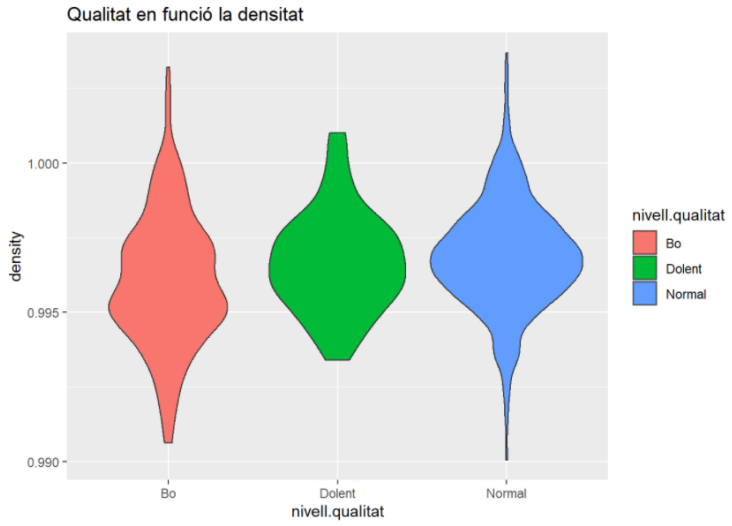
```
ggplot(data_wine, aes(x=nivell.qualitat, y=sulphates, fill=nivell.qualitat)) +
  ggtitle("Qualitat en funció del sulfat") +
  geom_violin()
```



Els vins de qualitat més bona, en general presenten una quantitat de sulfat més gran que els vins de menys qualitat. Aquest element s'utilitza com a conservant i antioxidant.

Qualitat en funció de la densitat

```
ggplot(data_wine, aes(x=nivell.qualitat, y=density, fill=nivell.qualitat)) +
  ggtitle("Qualitat en funció la densitat") +
  geom_violin()
```



Els vins etiquetats com a bons presenten una densitat més baixa que els etiquetats com a dolents o normal.

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

El problema que es vol resoldre és l'estudi de la qualitat d'un vi a partir de l'anàlisi fisicoquímica dels seus components (alcohol, pH, sulfats, densitat, acidesa cítrica, acidesa volàtil, etc.). Es tracta d'identificar quins components afecten més a l'hora de determinar la qualitat final d'un vi en una escala de l'1 al 10.

Hem fet tres proves estadístiques (contrast d'hipòtesis, correlacions i regressió lineal múltiple) i dos models supervisats (random forest i arbre de decisions).

Hem pogut observar que no existeix una correlació gaire forta entre les variables que hem seleccionat com a més importants per a l'estudi.

Segons els resultats obtinguts de l'estudi estadístic fet anteriorment, podem concloure que els vins de major qualitat (Quality) presenten uns nivells alts d'alcohol, superior a 10, sent aquesta variable la que més han influït en l'estudi. Altres variables que han influït força per arribar a aquestes conclusions són el pH, els sulfats, l'acidesa cítrica i l'acidesa volàtil.

Cal comentar, que és important ampliar l'estudi amb un nombre més gran de variables que no apareixen en el dataset i que presenten una gran importància a l'hora de decidir quin vi és de millor qualitat.

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 font per a realitzar aquesta pràctica està escrit en llenguatge R i és accesible mitjançant el següent enllaç al repositori de GitHub:

https://github.com/jpoumas/Practica2-Tipologia-i-cicle-de-vida-de-les-dades

Contribucions	Signatura
Investigació prèvia	**X.B.M. - J.P.M.**
Redacció de les respostes	**X.B.M. - J.P.M.**
Desenvolupament del codi	**X.B.M. - J.P.M.**