Tipologia i Cicle de Vida de les Dades

PRAC2 - Neteja i anàlisi de les dades

Òscar del Álamo i Guaus

1/06/2020

Contents

---

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

El dataset triat és "Wine Quality Data Set". El dataset està disponible a kaggle, en aquest enllaç o al repositori "UCI machine learning", en aquest enllaç. S'hi inclouen dos conjunts de dades relacionats amb mostres de les variants negres i blanques del vi "Vinho Verde", del nord de Portugal. Per a més detalls, consulteu: Enllaç web o la referència [Cortez et al., 2009]. Per motius de de privacitat i logística, només hi ha variables fisicoquímiques (d'entrada) i sensorials (de sortida) (per exemple, no hi ha dades sobre tipus de raïm, marca de vi, preu de venda de vi, etc.).

Aquests conjunts de dades es poden veure com a tasques de classificació o de regressió. Les classes estan ordenades i no equilibrades (per exemple, hi ha molts més vins normals que excel · lents o pobres). Es poden utilitzar algorismes de detecció de valors atípics per detectar els pocs vins excel · lents o pobres. A més, no estem segurs de si totes les variables d'entrada són rellevants. Així que podria ser interessant provar mètodes de selecció d'atributs.

Les variables disponibles són, variables d'entrada o atributs: 1 - fixed acidity 2 - volatile acidity 3 - citric acid 4 - residual sugar 5 - chlorides 6 - free sulfur dioxide 7 - total sulfur dioxide 8 - density 9 - pH 10 - sulphates 11 - alcohol Variable de sortida o classe: 12 - quality

---

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

```
# Carreguem el conjunt de dades.
data <- read.csv(
  "winequality-white.csv",
  sep = ';',
```

```
  header = TRUE,
  stringsAsFactors = FALSE
)

# Mostrem l'estructura d'aquest.
str(data)
```

```
## 'data.frame':    4898 obs. of  12 variables:
##  $ fixed.acidity       : num  7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...
##  $ volatile.acidity    : num  0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 ...
##  $ citric.acid         : num  0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34 0.43 ...
##  $ residual.sugar      : num  20.7 1.6 6.9 8.5 8.5 6.9 7 20.7 1.6 1.5 ...
##  $ chlorides           : num  0.045 0.049 0.05 0.058 0.058 0.05 0.045 0.045 0.049 0.044 ...
##  $ free.sulfur.dioxide : num  45 14 30 47 47 30 30 45 14 28 ...
##  $ total.sulfur.dioxide: num  170 132 97 186 186 97 136 170 132 129 ...
##  $ density             : num  1.001 0.994 0.995 0.996 0.996 ...
##  $ pH                  : num  3 3.3 3.26 3.19 3.19 3.26 3.18 3 3.3 3.22 ...
##  $ sulphates           : num  0.45 0.49 0.44 0.4 0.4 0.44 0.47 0.45 0.49 0.45 ...
##  $ alcohol             : num  8.8 9.5 10.1 9.9 9.9 10.1 9.6 8.8 9.5 11 ...
##  $ quality             : int  6 6 6 6 6 6 6 6 6 6 ...
```

Per aquesta pac, ens centrarem en l'estudi del dataset de vins blancs. Com veiem el dataset té un total de 4898 registres amb les 12 variables, comptant la qualitat, esmentades anteriorment.

La classe, qualitat, és una variable quantitativa discreta, ja que només pot prendre valors enters entre 0 i 10. Els atributs, en canvi, com són el resultat de fer mesures, són variables quantitatives contínues. Veiem que R ja ha detectat aquesta diferència pel tipus de dades que hi havia a la columna del fitxer csv, int o num. Així doncs, pel tipus de dades que tenim, ens haurem de centrar en l'anàlisi quantitatiu.

També està bé comentar que en aquest pas, segons el dataset que analitzem, ens podríem trobar casos amb presència d'espais en blanc en variables de tipus caràcter o trobar categories etiquetades de manera lleugerament diferent. També ens podríem trobar casos de variables numèriques on per indicar els valors decimals hi ha casos en què s'utilitza la coma i hi ha casos en què s'utilitza el punt, depenent d'on s'han pres la mesura si a Europa o a Amèrica. Tots, aquests casos els hauríem de corregir segons el que creiéssim més convenient.

```
# Resum descriptiu de les dades.
summary(data)
```

```
##  fixed.acidity    volatile.acidity  citric.acid     residual.sugar
##  Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600
##  1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700   1st Qu.: 1.700
##  Median : 6.800   Median :0.2600   Median :0.3200   Median : 5.200
##  Mean   : 6.855   Mean   :0.2782   Mean   :0.3342   Mean   : 6.391
##  3rd Qu.: 7.300   3rd Qu.:0.3200   3rd Qu.:0.3900   3rd Qu.: 9.900
##  Max.   :14.200   Max.   :1.1000   Max.   :1.6600   Max.   :65.800
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide
##  Min.   :0.00900   Min.   :  2.00      Min.   :  9.0
##  1st Qu.:0.03600   1st Qu.: 23.00      1st Qu.:108.0
##  Median :0.04300   Median : 34.00      Median :134.0
##  Mean   :0.04577   Mean   : 35.31      Mean   :138.4
##  3rd Qu.:0.05000   3rd Qu.: 46.00      3rd Qu.:167.0
##  Max.   :0.34600   Max.   :289.00      Max.   :440.0
##     density             pH          sulphates         alcohol
##  Min.   :0.9871   Min.   :2.720   Min.   :0.2200   Min.   : 8.00
##  1st Qu.:0.9917   1st Qu.:3.090   1st Qu.:0.4100   1st Qu.: 9.50
```

```
##  Median :0.9937   Median :3.180   Median :0.4700   Median :10.40
##  Mean   :0.9940   Mean   :3.188   Mean   :0.4898   Mean   :10.51
##  3rd Qu.:0.9961   3rd Qu.:3.280   3rd Qu.:0.5500   3rd Qu.:11.40
##  Max.   :1.0390   Max.   :3.820   Max.   :1.0800   Max.   :14.20
##     quality
##  Min.   :3.000
##  1st Qu.:5.000
##  Median :6.000
##  Mean   :5.878
##  3rd Qu.:6.000
##  Max.   :9.000
```

En aquest resum descriptiu podem alguns descriptors estadístics per tenir més informació sobre com es distribueixen les nostres dades.

---

3. Neteja de les dades

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

A la descripció del dataset del repositori UCI ens informen que no hi ha valors nuls, però mai està de més comprovar-ho.

```
colSums(is.na(data))
```

```
##        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
```

Com veiem no tenim elements buits. S'ha de dir que el concepte element buit pot variar en cada dataset i hauríem d'analitzar-ho cada cop tenint en compte que pot ser que els elements buits estiguin etiquetats amb un caràcter especial i llavors hauríem de buscar casos que coincideixin amb aquell caràcter enlloc de buscar elements buits.

Per gestionar aquests casos hi ha diverses aproximacions. Una opció seria eliminar els registres que contenen elements buits, però amb això estaríem perdent informació. Una tècnica més freqüent és fer una imputació. Per fer-la, podem substituir els elements buits pel valor més freqüent d'aquell atribut, per la mitja, etc. El problema de només tenir en compte el mateix atribut, a part de substituir tots els elements buits per un mateix valor, fa que la substitució pugui tenir poca concordància amb la resta d'atributs.

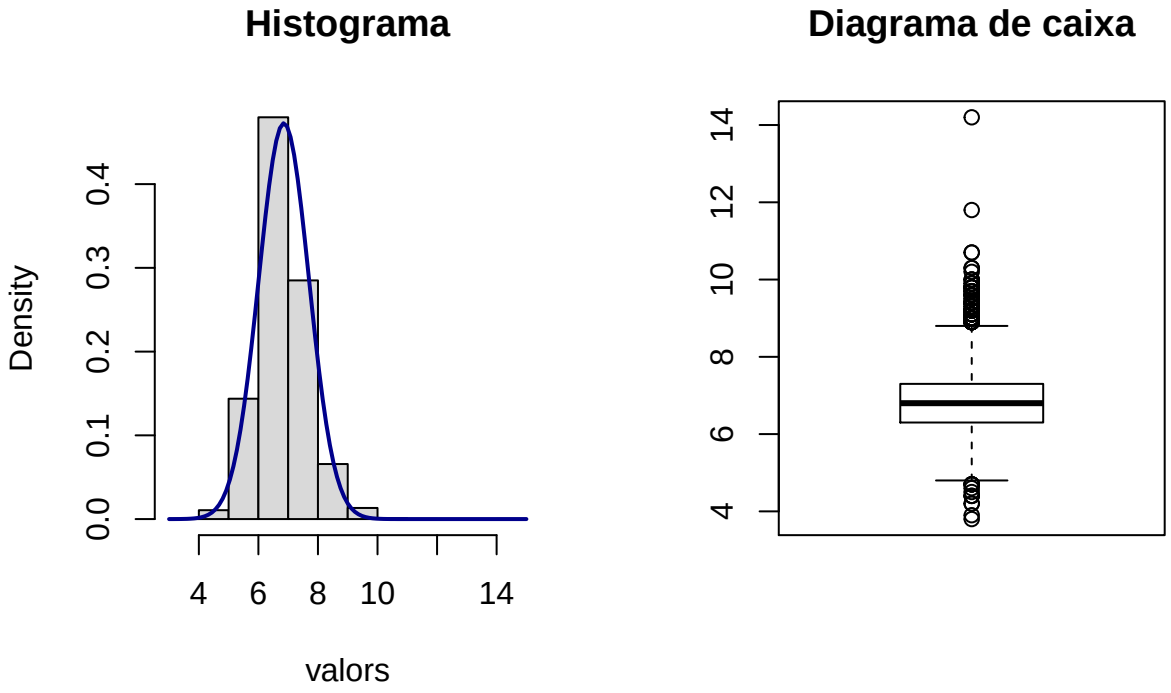
Per tal de solucionar aquest problema, s'utilitzen mètodes d'imputació basades en els veïns més pròxims, com el kNN. Per aquesta via, la imputació té en compte les relacions que hi ha entre els atributs i, en cada cas, s'assignarà un valor diferent per substituir l'element buit.

3.2. Identificació i tractament de valors extrems.

Teòricament considerem els valors extrems aquells que tenen un z-score > 3. El z-score ens indica a quantes desviacions estàndard està un valor respecte la mitja del conjunt. Per tal d'identificar els valors extrems, el diagrama de caixa és el què ens permet detectar-los de manera més fàcil. També utilitzem l'histograma per veure més clarament amb quina freqüència es distribueixen els diferents valors atributs. Per cada atribut, també veiem quina quantitat de registres tenen valors considerats extrems seguint la teoria esmentada abans.

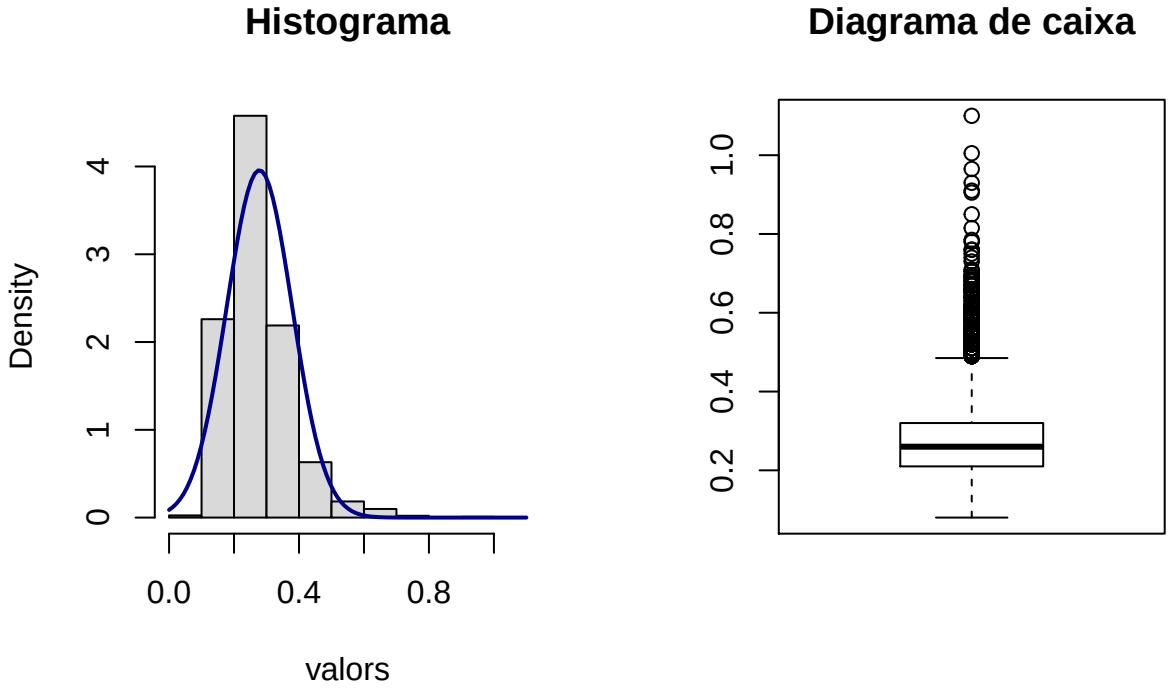
```
atributs = colnames(data)[1:11]

for (atribut in colnames(data)){
  par(
    mfrow=c(1,2),
    oma = c(0, 0, 2, 0)
  )
  hist(
    data[,atribut],
    xlab=paste("valors"),
    main="Histograma",
    prob=T,
    col="grey85"
  )
  curve(
    dnorm(
      x,
      mean=mean(data[,atribut]),
      sd=sd(data[,atribut])
    ),
    col="darkblue",
    lwd=2,
    add=TRUE
  )
  boxplot(
    data[,atribut],
    main="Diagrama de caixa"
  )
  mtext(
    paste("Atribut analitzat: ",atribut),
    outer = TRUE,
    cex = 1.5
  )
  teorics <- length(data[,atribut][which(abs(scale(data[,atribut]))>3)])
  box_out <- length(boxplot.stats(data[, atribut])$out)
  print(
    paste(
      "Nombre de valors extrems: ",
      teorics,
      " de teòrics i ",
      box_out,
      " segons la funció boxplot.stats."
    )
  )
}
```

Atribut analitzat: fixed.acidity



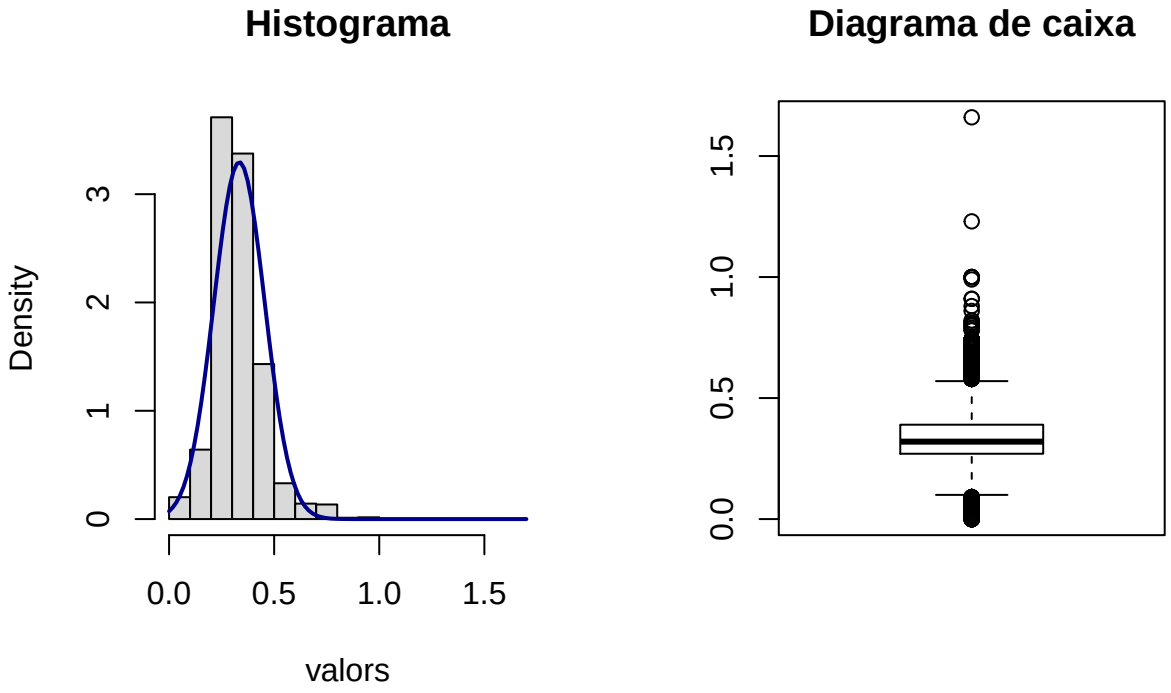

```
## [1] "Nombre de valors extrems:  46  de teòrics i  119  segons la funció boxplot.stats."
```

Atribut analitzat: volatile.acidity



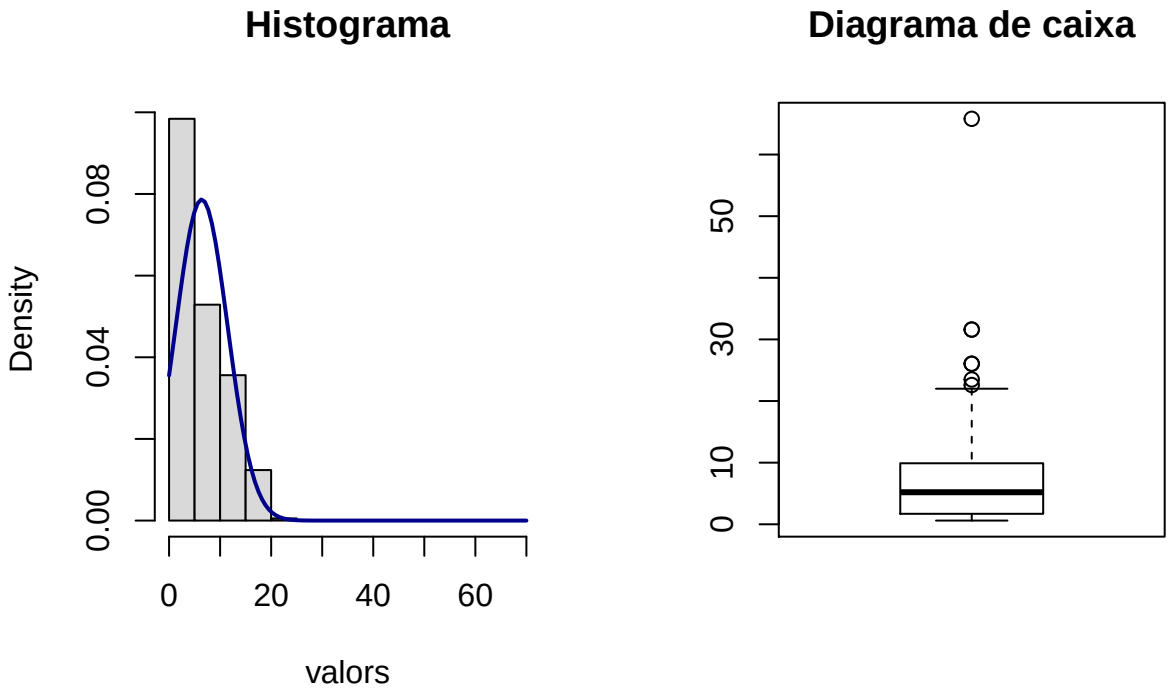

```
## [1] "Nombre de valors extrems:  81  de teòrics i  186  segons la funció boxplot.stats."
```

Atribut analitzat: citric.acid



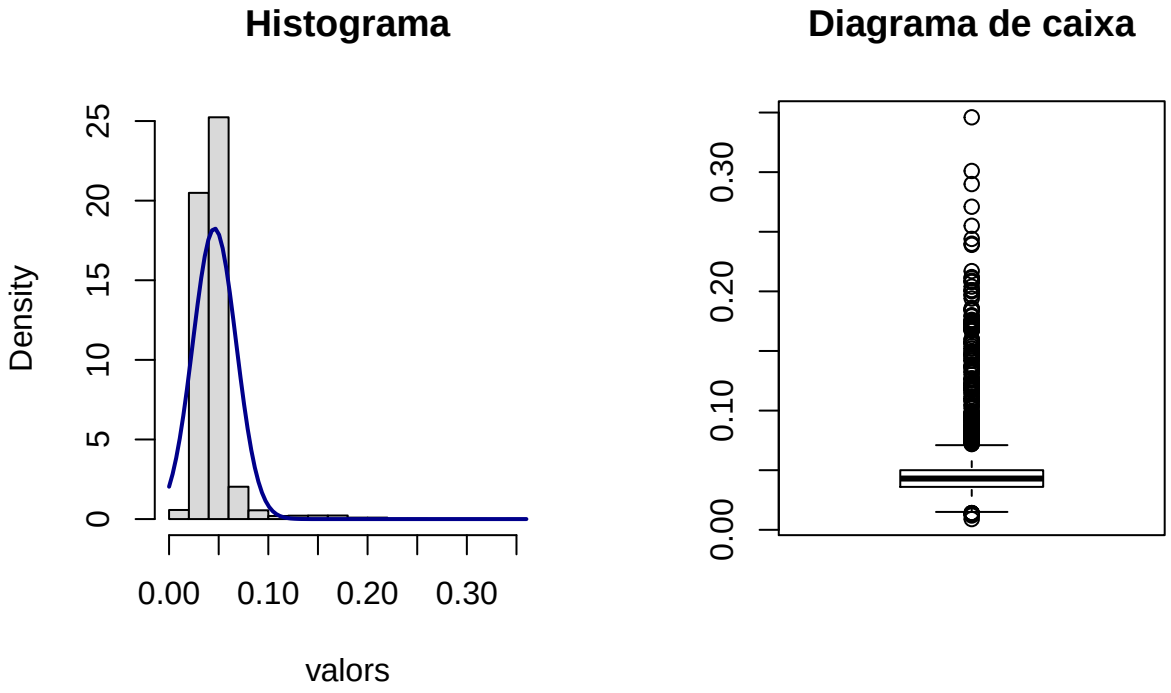

```
## [1] "Nombre de valors extrems:  85  de teòrics i  270  segons la funció boxplot.stats."
```

Atribut analitzat: residual.sugar



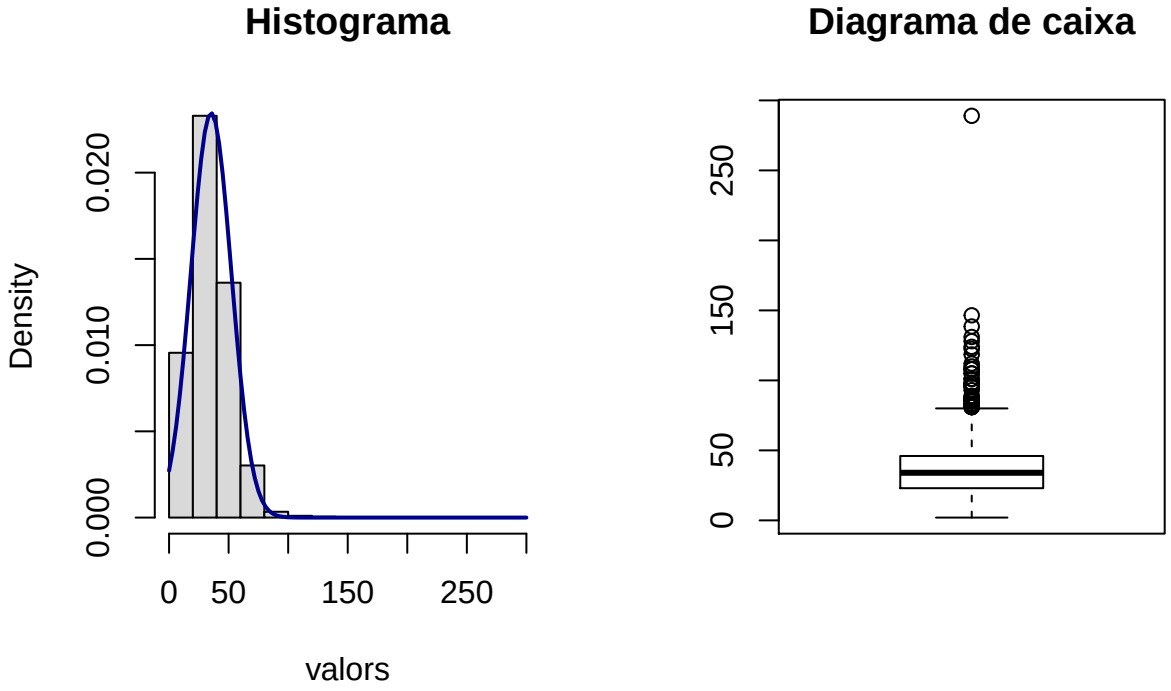

```
## [1] "Nombre de valors extrems:  9  de teòrics i  7  segons la funció boxplot.stats."
```

Atribut analitzat: chlorides



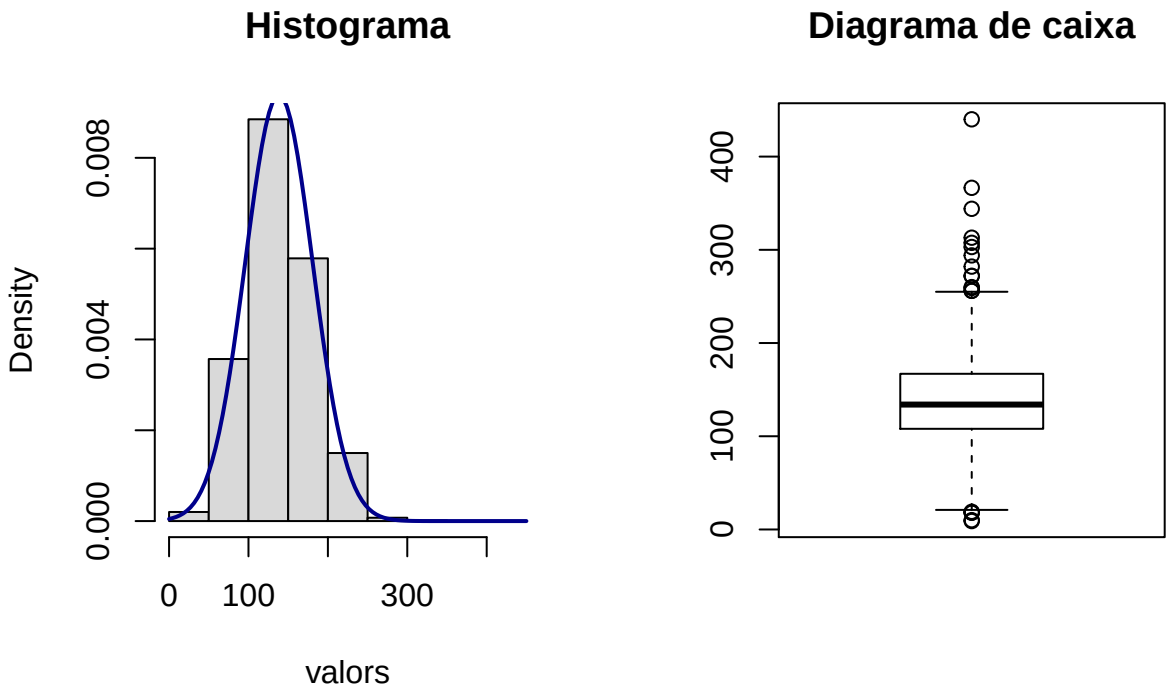

```
## [1] "Nombre de valors extrems:  102  de teòrics i  208  segons la funció boxplot.stats."
```

Atribut analitzat: free.sulfur.dioxide



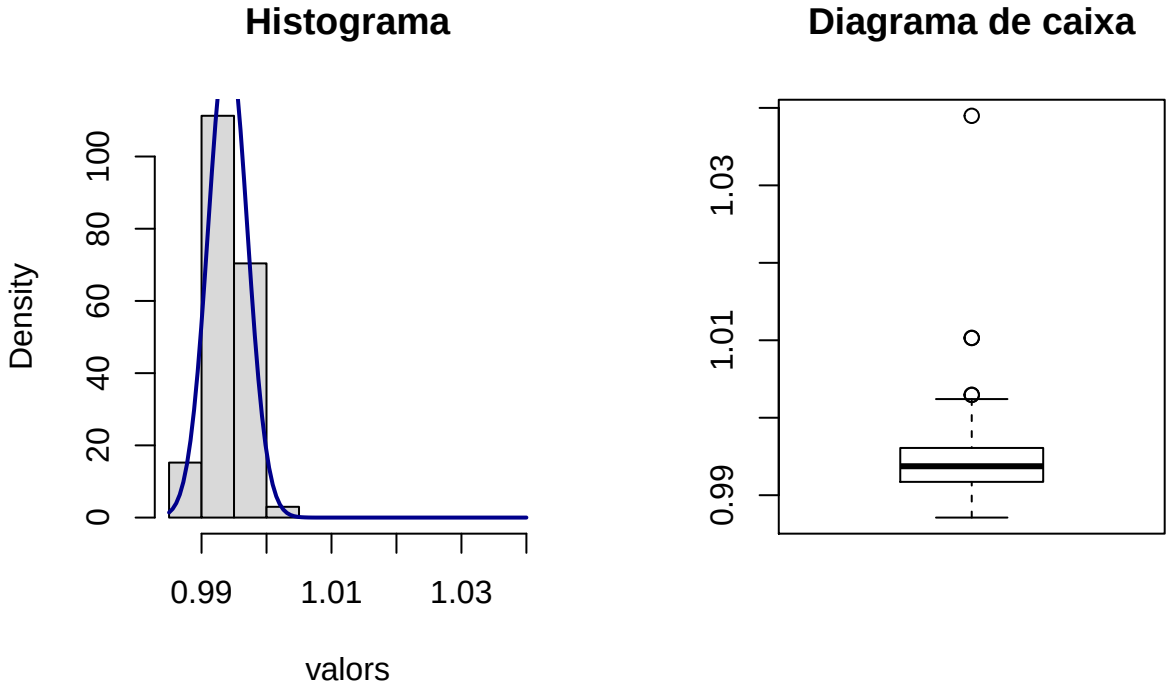

```
## [1] "Nombre de valors extrems:  32  de teòrics i  50  segons la funció boxplot.stats."
```

Atribut analitzat: total.sulfur.dioxide



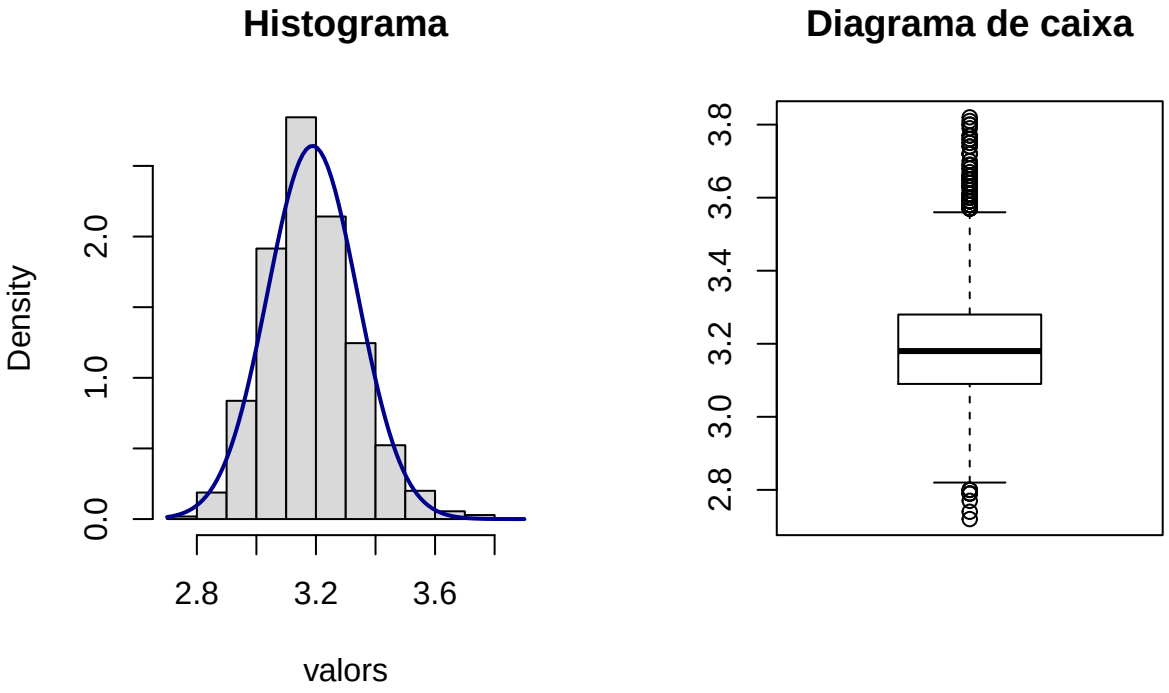

```
## [1] "Nombre de valors extrems:  12  de teòrics i  19  segons la funció boxplot.stats."
```

Atribut analitzat: density



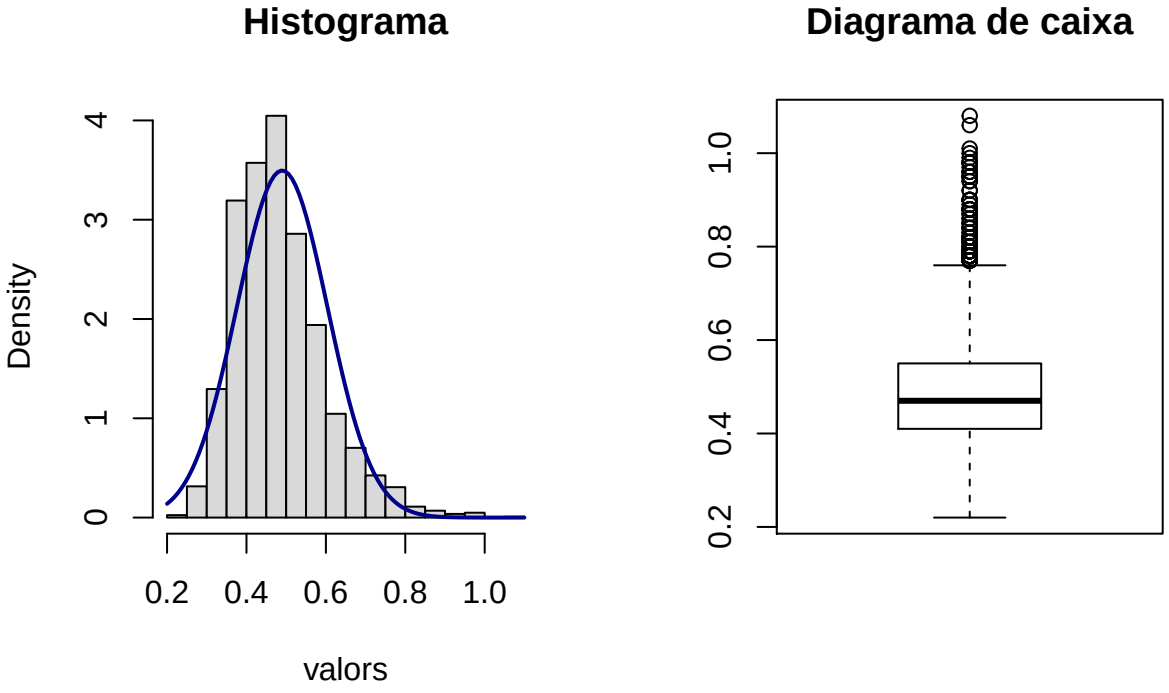

```
## [1] "Nombre de valors extrems:  3  de teòrics i  5  segons la funció boxplot.stats."
```

Atribut analitzat: pH



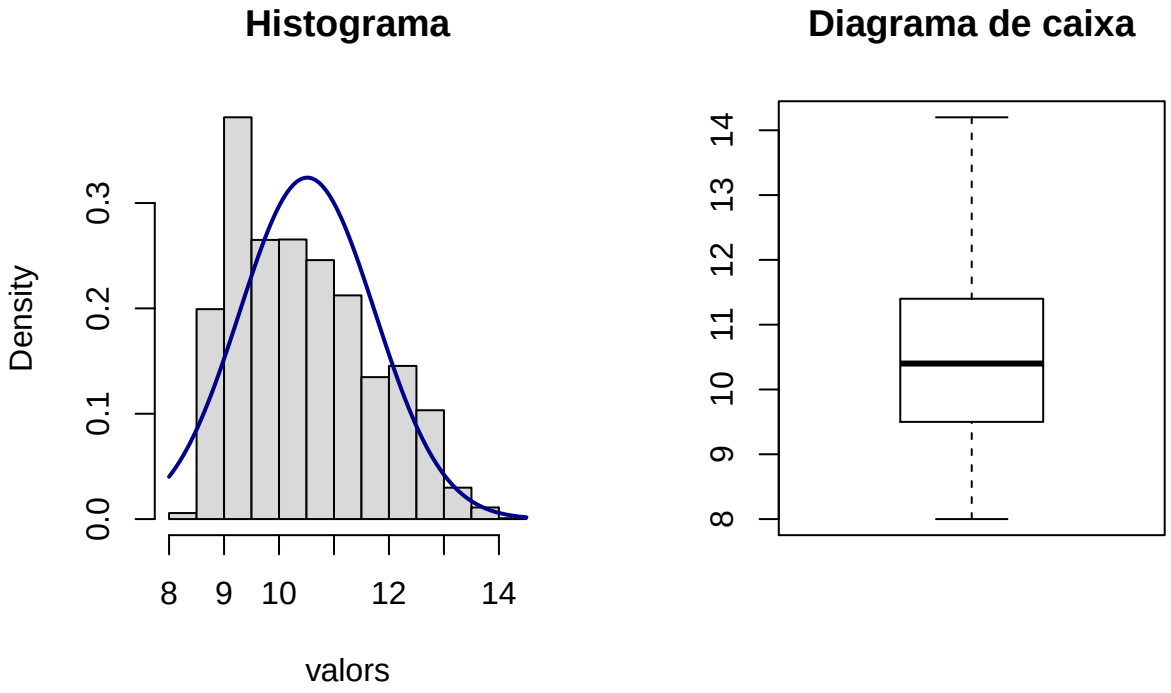

```
## [1] "Nombre de valors extrems:  32  de teòrics i  75  segons la funció boxplot.stats."
```

Atribut analitzat: sulphates



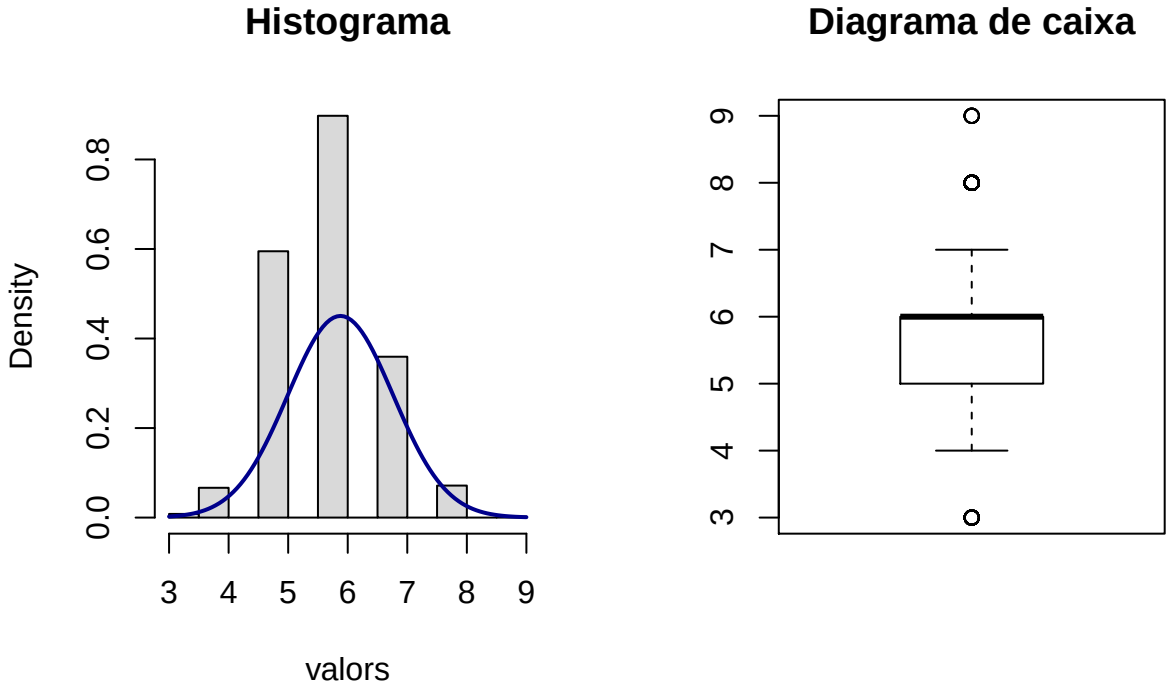

```
## [1] "Nombre de valors extrems:  48  de teòrics i  124  segons la funció boxplot.stats."
```

Atribut analitzat: alcohol




```
## [1] "Nombre de valors extrems:  0  de teòrics i  0  segons la funció boxplot.stats."
```

Atribut analitzat: quality

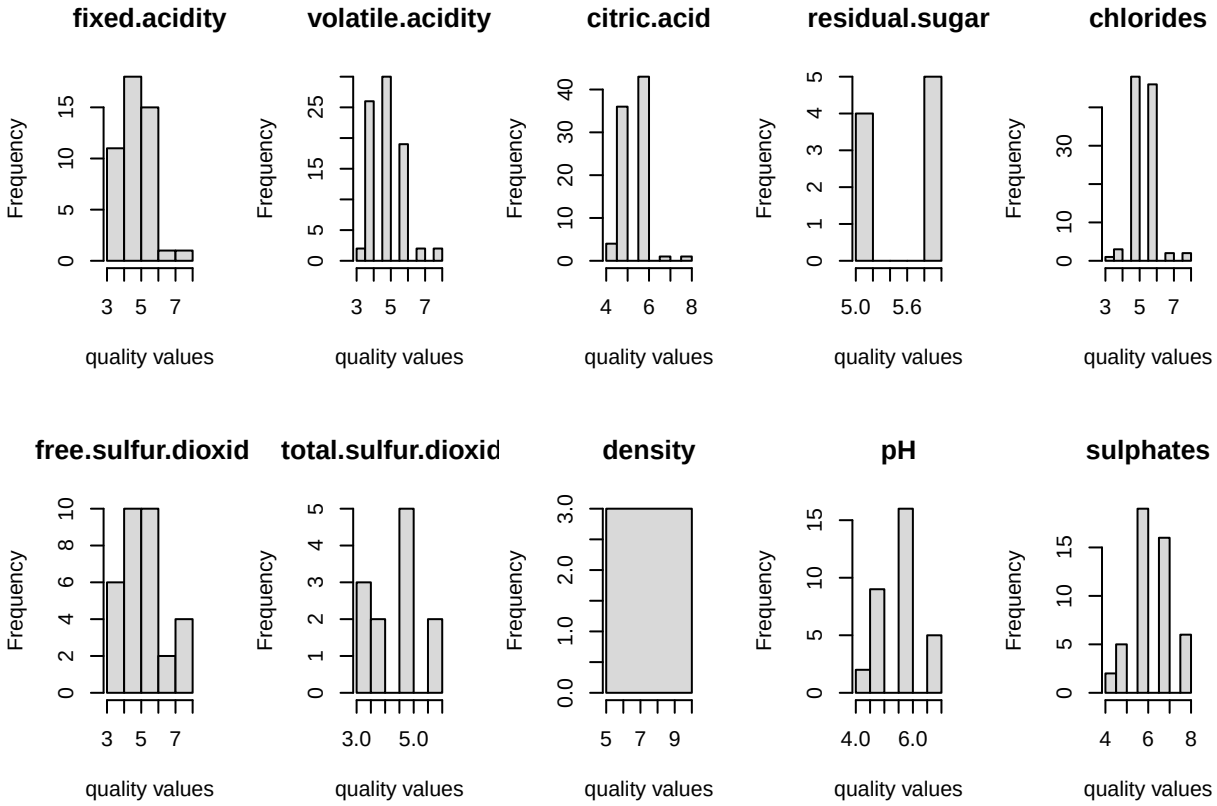



```
## [1] "Nombre de valors extrems:  25  de teòrics i  200  segons la funció boxplot.stats."
```

Veient els resultats de les gràfiques, hem de considerar per cada tipus atribut com tractem els valors extrems.L'atribut alcohol no mostra valors extrems. Pel que fa a l'atribut pH, veiem que tots els valors estan dintre del rang (0, 7) que ens indica acidesa, de manera que podria considerar-se la opció de manternir els valors extrems si ho consideressim necessari. Per acabar de decidir, també podem veure si realment hi ha algun atribut que tingui valors extrems on la distribució de la qualitat del vi sigui molt clara. Per exemple, que els valors extrems del pH facin qe un vi sigui de millor qualitat.

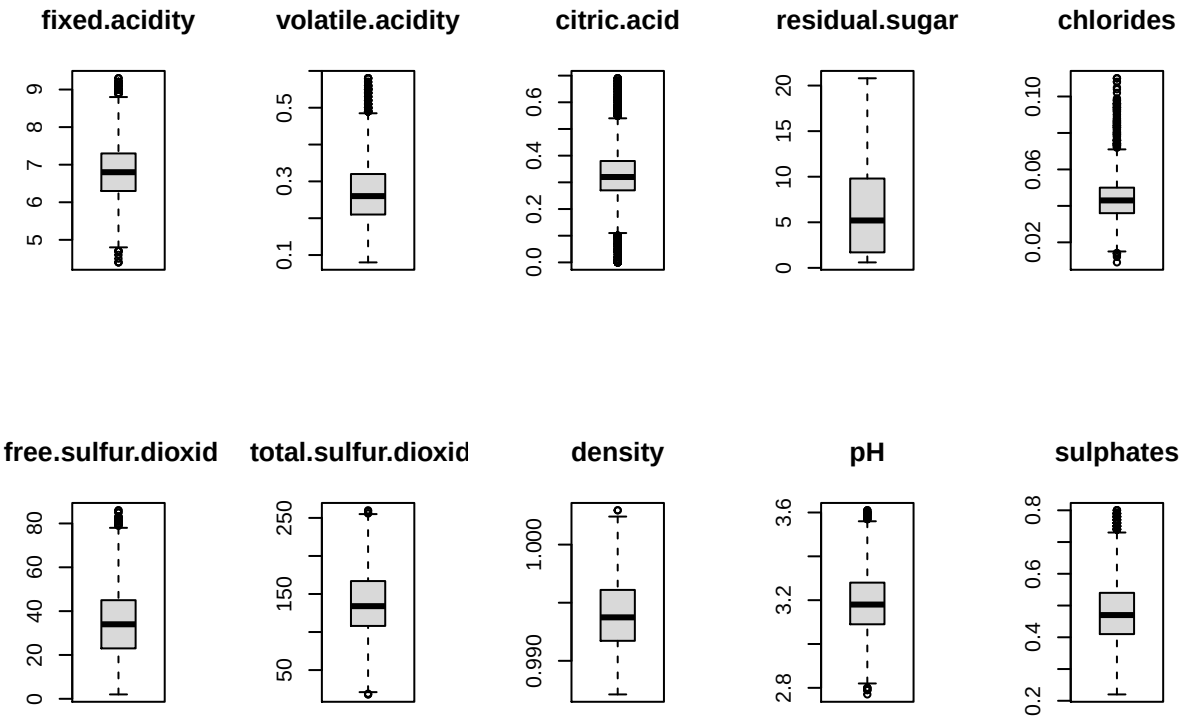
```
## [1] "Histogrames de distribució de la qualitat del vi"
```

```
## [1] "En funció de les mostres amb valors extrems de cada atribut"
```



Com s'aprecia a les gràfiques els valors extrems no són determinants pel que fa a la qualitat del vi, de manera que es poden corregir utilitzant la seva mitjana i només ho aplicarem en els casos con el z-score sigui un valor més gran que 3, tant el cas de desviacions positives com negatives.

```
atributs = colnames(data)[1:10]
par(mfrow=c(2,5))
data$pH[which(abs(scale(data$pH))>3)] <- mean(data$pH)
data$sulphates[which(abs(scale(data$sulphates))>3)] <- mean(data$sulphates)
for (atribut in atributs){
  data[,atribut][which(abs(scale(data[,atribut]))>3)] <- mean(data[,atribut])
  boxplot(
    data[,atribut],
    main=paste(atribut),
    col="grey85"
  )
}
```



4. Anàlisi de les dades

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

En aquesta secció, depenent dels anàlisis a realitzar, es poden utilitzar mètodes de reducció de dimensionalitat com el PCA o el t-SNE i també mètodes de de feature engineering, per a partir dels atributs inicials, calcular-ne d'altres aprofitant el coneixament del camp que s'està estudiant. També es poden aplicar mètodes de clustering per reduir el nombre de registres a entrenar.

En el nostre cas, com no tenim un nombre d'atributs molt gran i el nombre de mostres és de 5k aproximadament, d'entrada no ens caldria aplicar aquests mètodes. Si més no, si que separarem el conjunt de dades en 2 datasets, un de train i un de test per tal de mirar d'entrenar un model de regressió lineal i validar-ne els resultats. I també crearem una variable nova agrupant els vins segons la puntuació. Per una qualitat >5 l'etiqueta d'aquesta nova variable serà "Good Quality" i en cas contrari "Poor Quality". Això ho farem servir per fer un contrast d'hipòtesis a l'apartat 4.3.

```
# Fixem el random seed perquè la partició sigui reproduible.
set.seed(340)
mostres = sample.split(data$quality, SplitRatio = 0.75)
data_train = subset(data, mostres == TRUE)
data_test = subset(data, mostres == FALSE)
table(data_train$quality)
```

```
##
##    3    4    5    6    7    8    9
##   15  122 1093 1648  660  131    4
```

```r
table(data_test$quality)
```

```
##
##   3   4   5   6   7   8   9
##   5  41 364 550 220  44   1
```

```r
# Creem la nova variable segons la puntuació de la qualitat.
data$qualitat_categorica <- ifelse(data$quality > 5, "Good Quality", "Poor Quality")
table(data$qualitat_categorica)
```

```
##
## Good Quality Poor Quality
##         3258         1640
```

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

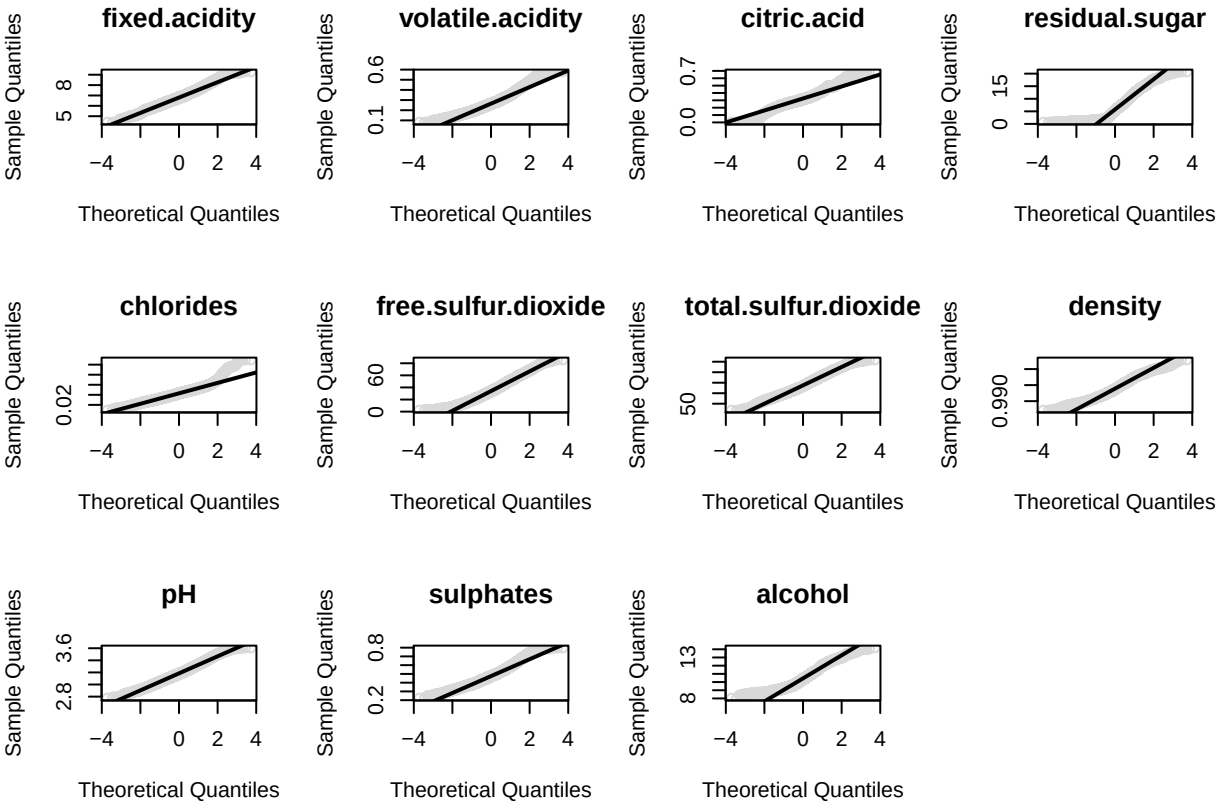
Podem comprovar la normalitat de manera visual i amb tests de normalitat. Per fer-ho de manera visual es poden utilitzar histogrames, com hem fet a la secció anterior, i també les gràfiques Q-Q.

Per realitzar els tests de normalitat, utilitzarem Shapiro-Wilk i Anderson-Darling, per veure si hi ha alguna diferència entre els resultats. Ambdós tenen com a hipòtesis nul·la, 0, que el conjunt té una distribució normal, de manera que si el p-value resultant és: $p - value > \alpha = 0.5$ no podrem rebutjar 0 i assumirem normalitat. En cas contrari, $p - value < \alpha = 0.5$ rebutjarem 0, per tan, rebutjarem normalitat.

```r
atributs = colnames(data)[1:11]
par(mfrow=c(3,4))
for (atribut in atributs){
  qqnorm(data[,atribut], main=paste(atribut), col="grey85")
  qqline(data[,atribut], lwd=2)
}

for (atribut in atributs){
  print(paste(atribut))
  print(ad.test(data[,atribut])$p.value)
  print(shapiro.test(data[,atribut])$p.value)
}
```

```
## [1] "fixed.acidity"
## [1] 3.7e-24
## [1] 2.185962e-16
## [1] "volatile.acidity"
## [1] 3.7e-24
## [1] 8.350407e-33
## [1] "citric.acid"
## [1] 3.7e-24
## [1] 2.577955e-29
## [1] "residual.sugar"
## [1] 3.7e-24
## [1] 1.711981e-49
## [1] "chlorides"
## [1] 3.7e-24
## [1] 1.017629e-38
## [1] "free.sulfur.dioxide"
## [1] 3.7e-24
## [1] 1.289726e-21
```

```
## [1] "total.sulfur.dioxide"
## [1] 3.7e-24
## [1] 4.548457e-15
## [1] "density"
## [1] 3.7e-24
## [1] 6.555373e-25
## [1] "pH"
## [1] 8.002872e-18
## [1] 3.741145e-13
## [1] "sulphates"
## [1] 3.7e-24
## [1] 1.530909e-28
## [1] "alcohol"
## [1] 3.7e-24
## [1] 2.569014e-36
```



Mirant els resultats de les gràfiques Q-Q, veiem que les corbes de la majoria de les variables s'aproximen força a la recta de quantils teòrica, de totes maneres a partir de la inspecció visual no podem acabar de determinar normalitat.

A partir dels resultats del test veiem que el p-value és inferior al nostre coeficient alfa, que és 0.05. Així doncs podem rebutjar la hipòtesi nuła i entenem que no segueixen una distribució normal. No obstant, com tenim més de 30 mostres podem aproximar les variables com una distribució normal de mitja 0 i desviació estàndar 1 pel teorema del límit central.

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.

Test de comparació de dues mitjanes

Es vol investigar si els vins amb una qualitat del tipus "Good Quality" tenen en mitjana un graduació d'alcohol que supera amb més de 0.9 graus la graduació dels vins del tipus "Bad Quality". El nivell de confiança establert per la hipòtesis és del 98%.

Es tract d'un contrast de dues mostres independents, ja que es vol comparar la mitjana segons el tipus de qualitat i les dades s'han obtingut de manera independent per cada vi.

Tenint en compte que μ_1 representa la mitjana de la distribució normal dels vins que tenen "Good Quality" i μ_2 dels que tenen "Poor Quality": - La hipòtesi nulla serà: $H_0 : \mu_1 = \mu_2$ - La hipòtesi alternativa serà bilateral: $H_1 : \mu_1 - \mu_2 > 0.9$

Podem aplicar un test paramètric:

```
good_wines <- data[which(data$qualitat_categorica=="Good Quality"), c("alcohol")]
summary(good_wines)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.50    9.80   10.80   10.85   11.90   14.20
```

```
bad_wines <- data[which(data$qualitat_categorica=="Poor Quality"), c("alcohol")]
summary(bad_wines)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00    9.20    9.60    9.85   10.40   13.60
```

```
wilcox.test(good_wines, bad_wines,  mu=0.9, alternative='greater', paired=FALSE, conf.int=0.98)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  good_wines and bad_wines
## W = 2782714, p-value = 0.00864
## alternative hypothesis: true location shift is greater than 0.9
## 95 percent confidence interval:
##  0.9000624       Inf
## sample estimates:
## difference in location
##              0.9999469
```

El test no paramètric ens confirma la hipòtesi alternativa, on la mitja de la graduació d'alcohol dels vins amb "Good Quality" és més gran de 0.9 graus que la dels vins amb "Poor Quality".

Correlacions entre els atributs

Calculem les correlacions entre totes les nostres columnes quantitatives:

```
matriu_correlacions <- hetcor(data[1:12], ML=FALSE, std.err=FALSE)
matriu_correlacions$correlations
```

```
##                      fixed.acidity volatile.acidity citric.acid
## fixed.acidity          1.00000000     -0.029044239  0.28360266
## volatile.acidity      -0.02904424      1.000000000 -0.14202646
## citric.acid            0.28360266     -0.142026462  1.00000000
```

```
## residual.sugar          0.10373303      0.066601455  0.07796296
## chlorides               0.08042274      0.018392792  0.02707775
## free.sulfur.dioxide    -0.02460436     -0.077926012  0.11677108
## total.sulfur.dioxide    0.09189515      0.103756953  0.11056076
## density                 0.26607140      0.003883785  0.13745993
## pH                     -0.39203821     -0.038495044 -0.15952902
## sulphates              -0.01223633     -0.036692812  0.06335540
## alcohol                -0.11687661      0.076664478 -0.06644367
## quality                -0.08937982     -0.150818890  0.01726977
##                      residual.sugar   chlorides free.sulfur.dioxide
## fixed.acidity            0.10373303  0.08042274         -0.02460436
## volatile.acidity         0.06660146  0.01839279         -0.07792601
## citric.acid              0.07796296  0.02707775          0.11677108
## residual.sugar           1.00000000  0.23280805          0.34126871
## chlorides                0.23280805  1.00000000          0.12147653
## free.sulfur.dioxide      0.34126871  0.12147653          1.00000000
## total.sulfur.dioxide     0.41528163  0.32042547          0.60819356
## density                  0.82771247  0.44984152          0.33467177
## pH                      -0.19136274 -0.04633979         -0.01011839
## sulphates               -0.01738548  0.07727344          0.06050526
## alcohol                 -0.46078622 -0.51010340         -0.26480759
## quality                 -0.09845284 -0.28578648          0.03399732
##                      total.sulfur.dioxide      density           pH
## fixed.acidity                 0.091895153  0.266071401 -0.392038212
## volatile.acidity              0.103756953  0.003883785 -0.038495044
## citric.acid                   0.110560758  0.137459931 -0.159529015
## residual.sugar                0.415281628  0.827712472 -0.191362743
## chlorides                     0.320425473  0.449841525 -0.046339793
## free.sulfur.dioxide           0.608193557  0.334671770 -0.010118393
## total.sulfur.dioxide          1.000000000  0.548249039 -0.003972054
## density                       0.548249039  1.000000000 -0.107466473
## pH                           -0.003972054 -0.107466473  1.000000000
## sulphates                     0.152127449  0.095926871  0.141944258
## alcohol                      -0.454737508 -0.803758749  0.126068762
## quality                      -0.164935981 -0.317230117  0.109538922
##                        sulphates     alcohol     quality
## fixed.acidity        -0.01223633 -0.11687661 -0.08937982
## volatile.acidity     -0.03669281  0.07666448 -0.15081889
## citric.acid           0.06335540 -0.06644367  0.01726977
## residual.sugar       -0.01738548 -0.46078622 -0.09845284
## chlorides             0.07727344 -0.51010340 -0.28578648
## free.sulfur.dioxide   0.06050526 -0.26480759  0.03399732
## total.sulfur.dioxide  0.15212745 -0.45473751 -0.16493598
## density               0.09592687 -0.80375875 -0.31723012
## pH                    0.14194426  0.12606876  0.10953892
## sulphates             1.00000000 -0.04950849  0.02664095
## alcohol              -0.04950849  1.00000000  0.43557472
## quality               0.02664095  0.43557472  1.00000000
```

Aquí podem observar que hi ha atributs que tenen una correlació molt alta, com la "density" amb el "residual.sugar" o el "total.sulfur.dioxide" amb el "free.sulfur.dioxide". Si necessitem reduir atributs, podríem eliminar-ne un de cada parella.

També s'observa que la qualitat té una forta correlació amb l'alcohol i una mica menys forta amb el pH.

Model de regressió lineal múltiple

Volem crear un model lineal que expliqui la variable quality en funció de l'alcohol i el pH del vi.

Per fer-ho veure'm tres models amb diferents atributs que tenen una correlació alta amb la puntuació de la qualitat.

```
model1 = lm(quality ~ alcohol + sulphates + pH + free.sulfur.dioxide + citric.acid, data=data_train)
summary(model1)
```

```
##
## Call:
## lm(formula = quality ~ alcohol + sulphates + pH + free.sulfur.dioxide +
##     citric.acid, data = data_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3506 -0.5228 -0.0208  0.4821  3.1771
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         0.5633460  0.3183754   1.769  0.07690 .
## alcohol             0.3332615  0.0110886  30.054  < 2e-16 ***
## sulphates           0.2738127  0.1275472   2.147  0.03188 *
## pH                  0.3920891  0.0944413   4.152 3.38e-05 ***
## free.sulfur.dioxide 0.0089798  0.0008682  10.343  < 2e-16 ***
## citric.acid         0.3591185  0.1270646   2.826  0.00473 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7876 on 3667 degrees of freedom
## Multiple R-squared:  0.2106, Adjusted R-squared:  0.2095
## F-statistic: 195.7 on 5 and 3667 DF,  p-value: < 2.2e-16
```

```
model2 = lm(quality ~ alcohol + sulphates + pH + citric.acid, data=data_train)
summary(model2)
```

```
##
## Call:
## lm(formula = quality ~ alcohol + sulphates + pH + citric.acid,
##     data = data_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5664 -0.5242 -0.0211  0.4930  3.0948
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.00599    0.32001   3.144 0.001682 **
## alcohol      0.30267    0.01084  27.921  < 2e-16 ***
## sulphates    0.30853    0.12933   2.386 0.017102 *
## pH           0.43372    0.09571   4.532 6.04e-06 ***
## citric.acid  0.49255    0.12822   3.841 0.000124 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.7989 on 3668 degrees of freedom
## Multiple R-squared:  0.1876, Adjusted R-squared:  0.1867
## F-statistic: 211.7 on 4 and 3668 DF,  p-value: < 2.2e-16
```

```
model3 = lm(quality ~ alcohol + pH + citric.acid, data=data_train)
summary(model3)
```

```
##
## Call:
## lm(formula = quality ~ alcohol + pH + citric.acid, data = data_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6046 -0.5255 -0.0145  0.4907  3.0834
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.04958    0.31969   3.283  0.00104 **
## alcohol      0.30092    0.01082  27.806  < 2e-16 ***
## pH           0.47046    0.09452   4.977 6.74e-07 ***
## citric.acid  0.51552    0.12794   4.029 5.71e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7994 on 3669 degrees of freedom
## Multiple R-squared:  0.1863, Adjusted R-squared:  0.1857
## F-statistic:   280 on 3 and 3669 DF,  p-value: < 2.2e-16
```

```
model4 = lm(quality ~ alcohol + pH, data=data_train)
summary(model4)
```

```
##
## Call:
## lm(formula = quality ~ alcohol + pH, data = data_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5544 -0.5217 -0.0108  0.4968  3.1479
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.42985    0.30608   4.672 3.10e-06 ***
## alcohol      0.29863    0.01083  27.575  < 2e-16 ***
## pH           0.41142    0.09357   4.397 1.13e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.801 on 3670 degrees of freedom
## Multiple R-squared:  0.1827, Adjusted R-squared:  0.1823
## F-statistic: 410.2 on 2 and 3670 DF,  p-value: < 2.2e-16
```

El model que ens ha donat millors resultats és el model4, on només fem servir els dos atributs amb una correlació més alta amb la qualitat, l'alchohol i el pH.

```
coef(model4)
```

```
## (Intercept)     alcohol          pH
```

```
##   1.4298524   0.2986307   0.4114162
```

Com es veu al resum del model, la nostra recta per predir la qualitat quedaria així: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ amb $\beta_0 = 1.4298524$, $\beta_1 = 0.2986307$ i $\beta_2 = 0.4114162$

El coeficient $R^2 = 0.801$. És a dir, en un principi el model explica el 80,1% de la variancia de les mostres.

Quan els valors de les variables explicatives valen zero, el valor de la qualitat predit és 1.4298524. L'efecte que tenen les variables explicatives es dedueix dels valors que les acompanyen. L'alcohol augmenta el valor predit en β_1 per cada unitat. El pH té una rellevància una mica superior i, com que $beta_2$ és positiva, contribueix a augmentar el valor predit.

```
pred = predict(model3, newdata=data_test)

actuals_preds <- data.frame(cbind(actuals=data_test$quality, predicteds=pred))
head(actuals_preds)
```

```
##    actuals predicteds
## 7        6   5.516928
## 11       5   6.278620
## 15       5   5.690066
## 17       6   5.483294
## 19       6   6.134293
## 28       6   6.000415
```

```
# min_max precisió
mean(apply(actuals_preds, 1, min) / apply(actuals_preds, 1, max))
```

```
## [1] 0.9023464
```

```
# desviació percentual mitjana absoluta
mean(abs((actuals_preds$predicteds - actuals_preds$actuals))/actuals_preds$actuals)
```

```
## [1] 0.1097475
```

---

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

Durant tota la pràctica els anàlisis han anat acompanyats de gràfiques, però apart també en podem afegir algunes de noves.

Per exemple, els resultats dels tests de normalitat en format taula:

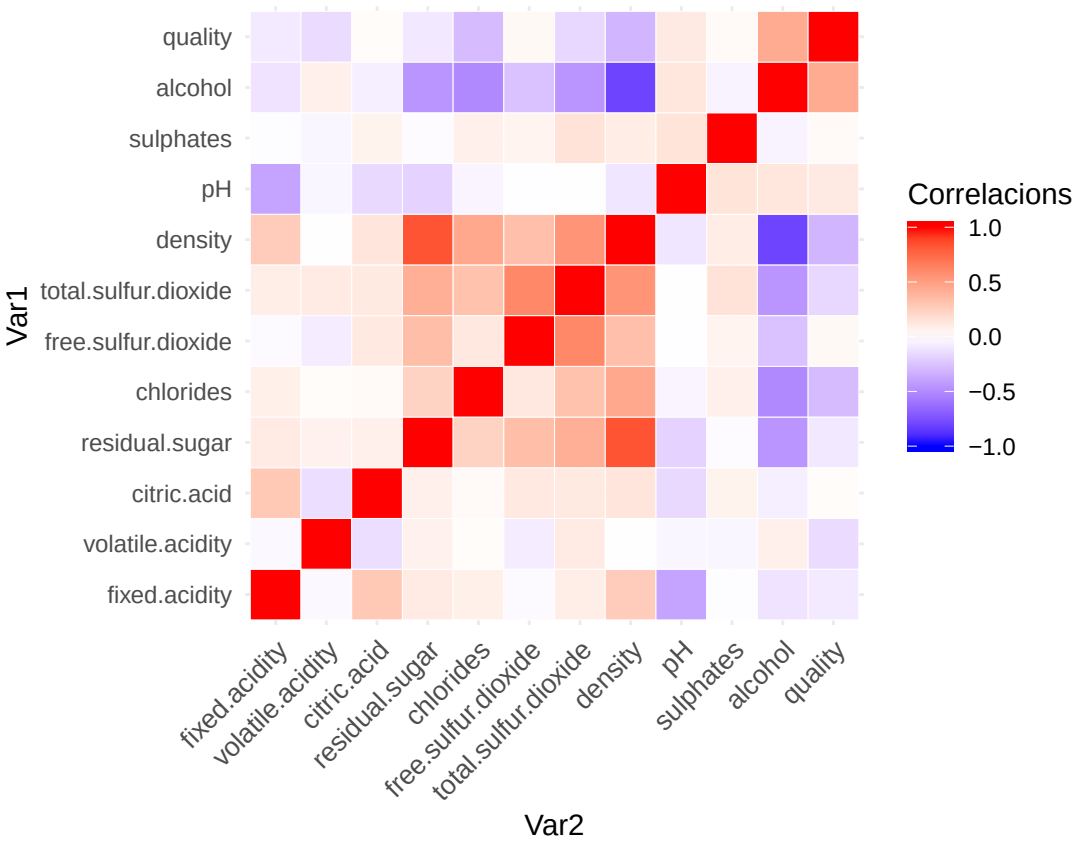
```
sw_p_value <- function(aux){
  return(shapiro.test(aux)$p.value)
}
ad_p_value <- function(aux){
  return(ad.test(aux)$p.value)
}
Shapiro_Wilk_results <- c(as.vector(sapply(data[1:12], sw_p_value)))
Anderson_Darling_results <- c(as.vector(sapply(data[1:12], ad_p_value)))

data.frame(
  variables=colnames(data)[1:12],
  Shapiro_Wilk_results,
```

```
  Anderson_Darling_results
  )
```

```
##               variables Shapiro_Wilk_results Anderson_Darling_results
## 1         fixed.acidity         2.185962e-16             3.700000e-24
## 2      volatile.acidity         8.350407e-33             3.700000e-24
## 3           citric.acid         2.577955e-29             3.700000e-24
## 4        residual.sugar         1.711981e-49             3.700000e-24
## 5             chlorides         1.017629e-38             3.700000e-24
## 6   free.sulfur.dioxide         1.289726e-21             3.700000e-24
## 7  total.sulfur.dioxide         4.548457e-15             3.700000e-24
## 8               density         6.555373e-25             3.700000e-24
## 9                    pH         3.741145e-13             8.002872e-18
## 10            sulphates         1.530909e-28             3.700000e-24
## 11              alcohol         2.569014e-36             3.700000e-24
## 12              quality         1.340111e-50             3.700000e-24
```

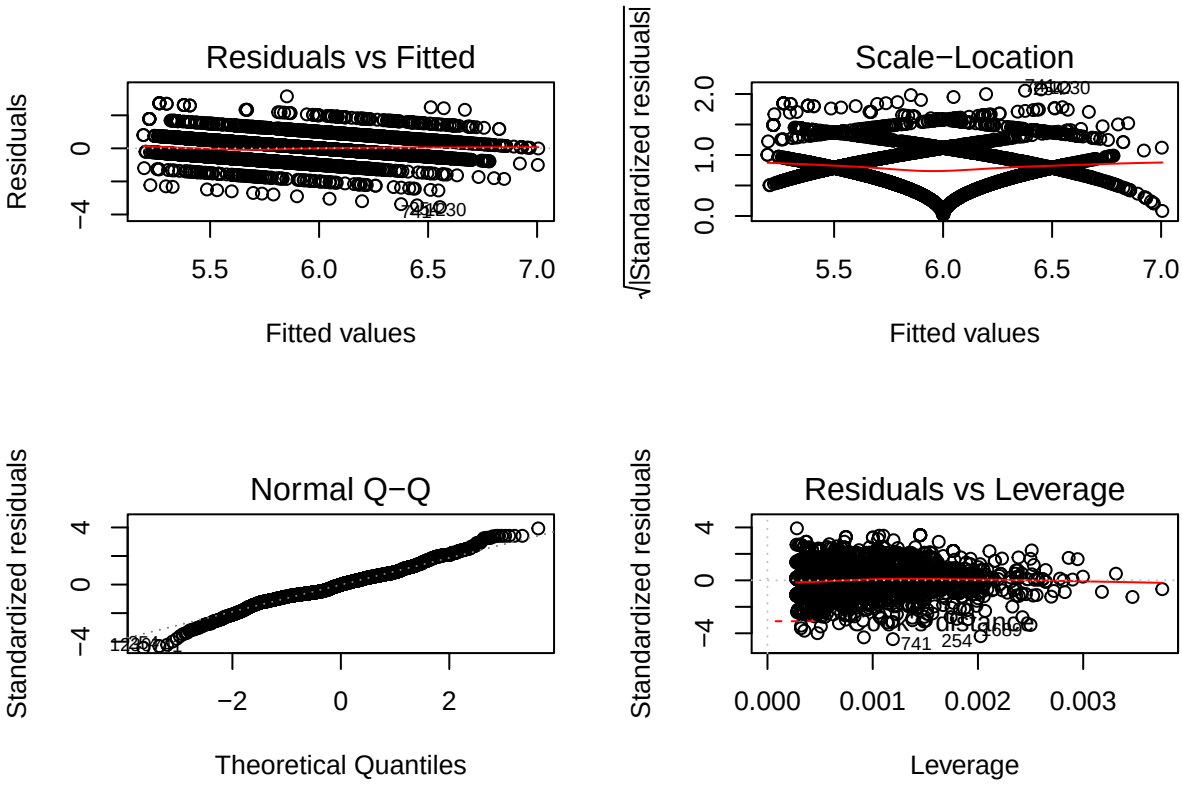
També podem veure,a través d'un heatmap, la taula de correlacions de les columnes quantitaves del nostre dataset.

```
ggplot(
  melt(matriu_correlacions$correlations),
  aes(Var2, Var1, fill=value)
)+
geom_tile(color="white")+
scale_fill_gradient2(
  low = "blue",
  mid = "white",
  high = "red",
  midpoint = 0,
  limit = c(-1, 1),
  name = "Correlacions"
)+
theme_minimal()+
theme(
  axis.text.x = element_text(
    angle=45, vjust=1, size=10, hjust=1
  )
)+
coord_fixed()
```



La visualització dels resultats del model4 en funció de diferents paràmetres:

```
layout(matrix(c(1,2,3,4),2,2))
plot(model4)
```



---

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

Al llarg d'aquest pràctica s'ha analitzat el conjunt de dades de vins blancs del dataset "Wine Quality Data Set". S'ha comprovat que no hi ha errors de format en el dataset ni elements buits i s'han vist descriptors estadístics genèrics. També s'ha explicat com es gestionarien els elements buits o els possibles errors de format.

A més, a la secció de selecció de dades s'ha explicat com es procediria, en cas que fos necessari, a reduir el nombre d'atributs o el nombre de mostres del conjunt.

Durant l'anàlisi hem pogut comprovar que no hi ha un patró clar que indiqui que la qualitat del vi depèn del valors extrems dels diferents atributs fisicoquímics obtinguts de cada vi.

A partir del contrast de la mitjana de l'alcohol entre les mostres de vins amb una qualitat superior a 5 i una qualitat inferior a 5, s'ha pogut comprovar que, amb una confiança del 98%, els vins de millor qualitat tenen una graduació d'alcohol 0.9 graus superior de mitja respecte els vins de pitjor qualitat.

A través de les correlacions entre les diferents variables hem vist que n'hi ha dues parelles que podrien ser considerades suplementàries de manera que se'n podria eliminar una variable de cada parell.

També s'han provat de realitzar diferents models de regressió lineal múltiple amb diferents atributs, i s'ha vist que el model que funció millor per explicar la qualitat és el que depèn de la graduació d'alcohol i del pH del vi.

---

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 pot trobar a GitHub al següent enllaç: https://github.com/poskinx/wine_quality.