Pràctica2: Neteja i Validació de les Dades

Joaquim Dalmases i Juanjo Díez

9 de junio, 2019

Contents

1 Introducció.

1.1 Presentació.

En aquesta pràctica s'elabora un cas pràctic orientat a aprendre a identificar les dades rellevants per un projecte analític i usar les eines d'integració, neteja, validació i anàlisi de les mateixes. Per fer aquesta pràctica haureu de treballar en grups de 2 persones. Haureu de lliurar un sol fitxer amb l'enllaç Github on es trobin les solucions incloent els noms dels components de l'equip. Podeu utilitzar la Wiki de Github per descriure el vostre equip i els diferents arxius que corresponen a la vostra entrega. Cada membre de l'equip haurà de contribuir amb el seu usuari Github.

1.2 Competències.

En aquesta pràctica es desenvolupen les següents competències del Màster de Data Science:

- Capacitat d'analitzar un problema en el nivell d'abstracció adequat a cada situació i aplicar les habilitats i coneixements adquirits per abordar-lo i resoldre'l.
- Capacitat per aplicar les tècniques específiques de tractament de dades (integració, transformació, neteja i validació) per al seu posterior anàlisi. Desenvolupar la capacitat de cerca, gestió i ús d'informació i recursos en l'àmbit de la ciència de dades.

1.3 Objectius.

Els objectius concrets d'aquesta pràctica són:

- Aprendre a aplicar els coneixements adquirits i la seva capacitat de resolució de problemes en entorns nous o poc coneguts dintre de contextos més amplis o multidisciplinaris.
- Saber identificar les dades rellevants i els tractaments necessaris (integració, neteja i validació) per dur a terme un projecte analític.
- Aprendre a analitzar les dades adequadament per abordar la informació continguda en les dades.
- Identificar la millor representació dels resultats per tal d'aportar conclusions sobre el problema plantejat en el procés analític.
- Actuar amb els principis ètics i legals relacionats amb la manipulació de dades en funció de l'àmbit d'aplicació.
- Desenvolupar les habilitats d'aprenentatge que els permetin continuar estudiant d'una manera que haurà de ser en gran manera autodirigida o autònoma.

2 Resolució.

Aquesta pràctica s'ha desenvolupat seguintla bibliografia recomanada: (Calvo M 2019; Squire 2015; Jiawei Han 2012; Dalgaard 2008)

2.1 Descripció del dataset.

Perquè és important i quina pregunta/problema pretén respondre?

Per l'elaboració de la pràctica s'ha triat:

- el repositori de *Kaggle* **Red Wine Quality**
- que correspon amb el repositori de *UCI* **Wine Quality Data Set** i
- l'accés a les dades completes es pot trobar a *aquest enllaç*.

2.1.1 Càrrega de dades

```
# Fixem el directori de treball:
setwd("C:/Users/juanj/OneDrive/Documentos/GitHub/Practica2")

# Llegim els fitxers amb les dades de vins blancs i negres
# Ho ubiquem a dos datasets dsRed i dsWhite.
redFile <-"winequality-red.csv"
whiteFile <-"winequality-white.csv"
dsRed <-read.csv(file.path(getwd(), redFile), sep=";", encoding="UTF-8")
dsWhite <-read.csv(file.path(getwd(), whiteFile), sep=";", encoding="UTF-8")
# Observem que els fitxers originals tenen iguals capçaleres.

# Comprobació de la bona lectura/transferència de dades, mirem les dues primeres fileres
# de cada dataset i vegem la composició .
head(dsRed,2)
```

```
##   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1           7.4             0.70           0            1.9     0.076
## 2           7.8             0.88           0            2.6     0.098
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  11                   34  0.9978 3.51      0.56     9.4
## 2                  25                   67  0.9968 3.20      0.68     9.8
##   quality
## 1       5
## 2       5
```

```
head(dsWhite,2)
```

```
##   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1           7.0             0.27        0.36           20.7     0.045
## 2           6.3             0.30        0.34            1.6     0.049
##   free.sulfur.dioxide total.sulfur.dioxide density  pH sulphates alcohol
## 1                  45                  170   1.001 3.0      0.45     8.8
```

```
## 2                14                  132   0.994 3.3      0.49     9.5
##   quality
## 1       6
## 2       6
```

```
summary(dsRed)
```

```
##  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
##  Min.   :0.01200   Min.   : 1.00       Min.   :  6.00
##  1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00
##  Median :0.07900   Median :14.00       Median : 38.00
##  Mean   :0.08747   Mean   :15.87       Mean   : 46.47
##  3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00
##  Max.   :0.61100   Max.   :72.00       Max.   :289.00
##     density             pH          sulphates         alcohol
##  Min.   :0.9901   Min.   :2.740   Min.   :0.3300   Min.   : 8.40
##  1st Qu.:0.9956   1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50
##  Median :0.9968   Median :3.310   Median :0.6200   Median :10.20
##  Mean   :0.9967   Mean   :3.311   Mean   :0.6581   Mean   :10.42
##  3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10
##  Max.   :1.0037   Max.   :4.010   Max.   :2.0000   Max.   :14.90
##     quality
##  Min.   :3.000
##  1st Qu.:5.000
##  Median :6.000
##  Mean   :5.636
##  3rd Qu.:6.000
##  Max.   :8.000
```

```
summary(dsWhite)
```

```
##  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
```

Perquè és important i quina pregunta/problema pretén respondre?

El dataset '**Red Wine**' emmagatzema les característiques fisico-químiques de les mostres de vi blanc i negre junt amb el ratio de la qualitat otorgada, en una escala de 0 a 10. Conté 1599 mostres de vi "Vinho Verde", negre i 4898 de blanc, de la zona nord de Portugal.

Cada mostra de vi té assignada un valor de qualitat resultats de proves realitzades en la seva composició (tests de quantitat d'alcohol, nivell d'acidesa, contingut residual de sucres etc...). En total són 12 atributs descrivint característiques entre fisicoquímiques i la classificació de qualitat de la mostra.

Emprearem aquest dataset per respondre a la pregunta de quínes característiques principals defineixen un vi de qualitat?. Cercarem si aquestes canvien segons el color del vi, negre o blanc.

Informació a tenir en compte: 'Vinho verde' és un producte únic de la regió de Minho (nord-oest) de Portugal. Es apreciat per les característiques: Alcohol grau mig, i especialment per la seva frescor.

Descripció dels atributs o camps del datatset:

Atribut	Traducció	Descripció
fixed.acidity	*Acidesa fixe*	És la quantitat d'àcidesa que no s'evapora i per tant resta fixe al vi.
volatile.acidity	*Acidesa volàtil*	La quantitat en excès d'àcid acètic en vi, pot afegir sabor amarg o avinagrat, si les quantitats són altes.
citric.acid	*Àcid cítric*	Trobat en petites quantitas, l'àcid cítric pot afegir frescor i sabor als vins.
residual.sugar	*Sucre residual*	Quantitat de sucre derivada del procés de fermentació (normalment trobem més de 1 gr/litre i si supera els 45grm./litre considerem el vi dolç.
chlorides	*Clorurs*	La quantitat de sal del vi.
free.sulfur.dioxide	*Diòxid de sofre*	Prevé el creixement microbià i l'oxidació del vi (anti-oxidant). Els vins blancs mantenen millor l'aspecte de vi jove. La normativa de la Comunitat Europea obliga des de l'any 2005 que qualsevol aliment o beguda que contingui més de 10 mg/l de sulfits ha de portar-ho en l'etiqueta com advertència. El motiu és que aquest additiu té capacitat al · lèrgena, és a dir, un petit percentatge de la població pot ser sensible o al · lèrgic als sulfits.
total.sulfur.dioxide	*diòxide de sofre total*	Concentracions per sobre de 50 ppm (tant lliure com unit), Es detecta per olfacte i tast. Les quantitats excessives de SO2 poden inhibir la fermentació i causar efectes sensorials indesitjables.

Atribut	Traducció	Descripció
density	*Densitat*	Serà propera a la de l'aigua (997 kg/m³) i variaria segons les quantitats de sucre i alcohol, segons la qualitat de la fermentació.
pH	*pH*	Descriu com un vi àcid o bàsic és a una escala de 0 (molt àcida) a 14 (molt bàsica); la majoria dels vins tenen entre 3-4 a l'escala de pH.
sulphates	*Sulfats*	Un additiu de vi que pot contribuir als nivells de diòxid de sofre (S02), que actuen com a antimicrobians i antioxidants
alcohol	*Alcohol*	El percentatge de contingut alcohòlic del vi, és una variable de sortida (basada en dades sensorials)
quality	*Qualitat*	(escala 0-10) És la qualitat atorgada la vi.
color	*Color*	Determina si el vi és blanc o negre. Afegida per nosaltres a efectes de integrar les dades.

2.2 Integració i selecció de les dades d'interés a analitzar.

Disposem de dos fitxers de dades un que conté les característiques del vins blancs i l'altre dels vins negres, per tant ens interesera comprovar que tenen les mateixes capçalers i que els podem integrar en un sol dataset. A més per tal de no perdre informació en la integració afegirem una columna 'color' que identificara la font de les files o mostres emmagatzemant el color del vi amb valors (blanc)

```
# Volem analitzar el dataset de Red Wine tenint en compte el color del vi,
# afegim una columna 'color' i fusionem les dades tant dels vins blancs com
# dels negres, diferentciant-los per el camp color.

# Afegim el camp 'color' a cada datatset
dsRed["color"]<-"negre"
dsWhite["color"]<-"blanc"

# Tenim capçaleres iguals, si la suma de noms iguals és la suma total de camps.
a<-colnames(dsRed)
b<-colnames(dsWhite)
cat(paste0("El nombre de camps (",ncol(dsRed),") és igual al nombre de camps iguals ",
           sum(a==b),"\n"))
```

```
## El nombre de camps (13) és igual al nombre de camps iguals 13
```

```
cat("Files - Instàncies de vi negre:",nrow(dsRed),"\nColumnes-Atributs-Variables:",
    ncol(dsRed),"\n")
```

```
## Files - Instàncies de vi negre: 1599
## Columnes-Atributs-Variables: 13
```

```
cat("Files - Instàncies de vi blanc:",nrow(dsWhite),"\nColumnes-Atributs-Variables:",
    ncol(dsWhite),"\n")
```

```
## Files - Instàncies de vi blanc: 4898
## Columnes-Atributs-Variables: 13
```

```
# Combinem les mostres dels dos fitxers i factoritzem el camp color per determinar
# els valors que pren: 'blanc i 'negre'
d<-rbind(dsRed,dsWhite)
d$color<-factor(d$color)

# Dimensions del dataset:
cat("Files - Instàncies:",nrow(d),"\nColumnes-Atributs-Variables:",ncol(d),"\n")
```

```
## Files - Instàncies: 6497
## Columnes-Atributs-Variables: 13
```

```
# Revisem l'estructura de camps del datatset:
summary(d)
```

```
##  fixed.acidity    volatile.acidity  citric.acid     residual.sugar
##  Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600
```

```
##  1st Qu.: 6.400   1st Qu.:0.2300   1st Qu.:0.2500   1st Qu.: 1.800
##  Median : 7.000   Median :0.2900   Median :0.3100   Median : 3.000
##  Mean   : 7.215   Mean   :0.3397   Mean   :0.3186   Mean   : 5.443
##  3rd Qu.: 7.700   3rd Qu.:0.4000   3rd Qu.:0.3900   3rd Qu.: 8.100
##  Max.   :15.900   Max.   :1.5800   Max.   :1.6600   Max.   :65.800
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide
##  Min.   :0.00900   Min.   :  1.00      Min.   :  6.0
##  1st Qu.:0.03800   1st Qu.: 17.00      1st Qu.: 77.0
##  Median :0.04700   Median : 29.00      Median :118.0
##  Mean   :0.05603   Mean   : 30.53      Mean   :115.7
##  3rd Qu.:0.06500   3rd Qu.: 41.00      3rd Qu.:156.0
##  Max.   :0.61100   Max.   :289.00      Max.   :440.0
##     density             pH          sulphates         alcohol
##  Min.   :0.9871   Min.   :2.720   Min.   :0.2200   Min.   : 8.00
##  1st Qu.:0.9923   1st Qu.:3.110   1st Qu.:0.4300   1st Qu.: 9.50
##  Median :0.9949   Median :3.210   Median :0.5100   Median :10.30
##  Mean   :0.9947   Mean   :3.219   Mean   :0.5313   Mean   :10.49
##  3rd Qu.:0.9970   3rd Qu.:3.320   3rd Qu.:0.6000   3rd Qu.:11.30
##  Max.   :1.0390   Max.   :4.010   Max.   :2.0000   Max.   :14.90
##     quality        color
##  Min.   :3.000   blanc:4898
##  1st Qu.:5.000   negre:1599
##  Median :6.000
##  Mean   :5.818
##  3rd Qu.:6.000
##  Max.   :9.000
```

```
# write.csv(d,"Dataset_inicial.csv",row.names = FALSE)
```

Com es pot veure ens quedem amb tots els atributs i més tard en la fase d'analisi determinarem si es possible una reducció de camps. Ara per ara podem comptar amb tots els camps disponibles al dataset, per esbrinar quíns ens determinaran la qualitat del vi. Cada característica és candidata per formar part de la composició del vi, i el color ens permerà comparar el seu grau d'importància segons tipus de vi.

A més a més, de la recerca que hem fet sobre el vi 'Vinho verde', sabem que tant el camp 'alcohol', com el camp 'citric.acid' seràn importants, ja que són característiques importants del vi (graduació mitjana en alcohol i sabor fresc).

```
cat("Atributs considerats:\n ")
```

```
## Atributs considerats:
##
```

```
colnames(d)
```

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

2.3 Neteja de les dades.

2.3.1 Zeros i elements buits.

Cerquem elements buits i Na:

```
cat("Valors ´na´: \n")
```

```
## Valors ´na´:
```

```
colSums(is.na(d))
```

```
##        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
##                color
##                    0
```

```
cat("Valors buits: \n")
```

```
## Valors buits:
```

```
colSums(d=="")
```

```
##        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
##                color
##                    0
```

```
cat("Zeros: \n")
```

```
## Zeros:
```

```
colSums(d==0)
```

```
##        fixed.acidity     volatile.acidity          citric.acid
##                    0                    0                  151
##       residual.sugar            chlorides  free.sulfur.dioxide
##                    0                    0                    0
```

```
## total.sulfur.dioxide              density                   pH
##                    0                    0                    0
##            sulphates              alcohol              quality
##                    0                    0                    0
##                color
##                    0
```

```
cat("Vegem si els valors son numérics: \n")
```

```
## Vegem si els valors son numérics:
```

```
str(d)
```

```
## 'data.frame':    6497 obs. of  13 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 ...
##  $ color               : Factor w/ 2 levels "blanc","negre": 2 2 2 2 2 2 2 2 2 2 ...
```

Podem veure que el dataset no presenta valors nulls (´na´) ni buids (“”) i que els zeros s'identifiquen bé: només el camp citric.acid conté valors 0 i que tots els camps son numérics llevat de quality que es sencer i color que és un factor amb dos valors (blanc i negre).

Amb lo que es pot concloure que hem verificat que **no cal tractar buids i Na**.

La gestió d'aquests cassos, si n'haguès aparegut algún, per aquest dataset podria ser omplir el valor faltant amb el resultat d'usar la imputació per kNN, ('k-Nearest Neighbors', es a dir els k-veïns més propers) per omplir els valors perduts ('NA').

Per defecte s'utilitza un valor de **k=5** que representa els 5 registres veïns més propers. La mètrica usada podría ser la distància de Gower, usant la resta de variables. El valor usat en la substitució és la mediana d'aquests 5 valors.

Aquesta tècnica és robusta quan existeixen valors extrems (en anglès 'outliers'). Si sabem que els valors de la mitjana i la mediana són molt propers, podem substituïr els valors faltants per la mitjana o mediana de l'atribut en qüestió.

2.3.2 Identificació i tractament de valors extrems.

Per identificar els valors extrems, podem utilitzar l'ordre 'boxplot.stats' i el paràmetre de sortida 'out' que determina els valors extrems. Es consideren normalment valors extrems o atípics en una mostra, aquells que superen els límits de 3 vegades la desviació típica a banda i banda de la mitjana de la mostra.

Per descriure els valors extrems i la distribució de cada variable hem definit una funció personalitzada on podem observar el histograma, el boxplot (amb valors descriptius com bigotis i mediana) en la mateixa escala i els valors extrems cercats:

```
bxplot_hist<- function(camp,lbcamp){
  par(mfrow=c(2,1))
  boxplot(camp,border="dodgerblue", ylim=c(min(camp,na.rm = TRUE),
                                           max(camp,na.rm = TRUE)), horizontal = TRUE)
  hist(camp,xlim = c(min(camp,na.rm = TRUE),max(camp,na.rm = TRUE)),
       main="",xlab = lbcamp, ylab="Freqüència", border="blue")
  # Descripció dels valors dels límits d'interpretació del boxplot
  # Vector $Stats indica consecutivament els valors de:
  #    Bigoti inferior, Q1(25%), Mediana, Q3(75%), Bigoti superior.
  # Vector $out  ens proporciona els valors atípics ordenats ascendentment.
  r<-boxplot.stats(camp)
  print("Valors atípics:  ")
  print(r$out)
  cat("\nValors del boxplot:\n1) Bigoti inf.:",r$stats[1],"\n2) Q1(25%)    :",
      r$stats[2],"\n3) Mediana    :",r$stats[3],"\n4) Q3(75%)    :",
      r$stats[4],"\n5) Bigoti Sup.:", r$stats[5],"\n")
  }
```

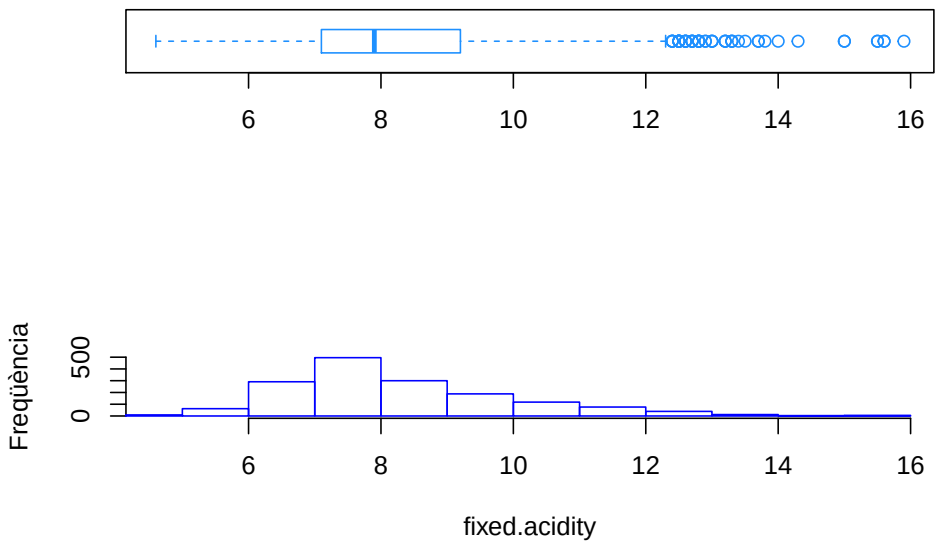
Per cada atribut mirem de trobar els valors extrems i decidim quína estratègia seguir.

2.3.2.1 Valors extrems de l'atribut: fixed.acidity

```
# Valors extrems de 'fixed.acidity'
cat("Vi negre \n")
```

```
## Vi negre
```

```
bxplot_hist(dsRed$fixed.acidity,'fixed.acidity')
```

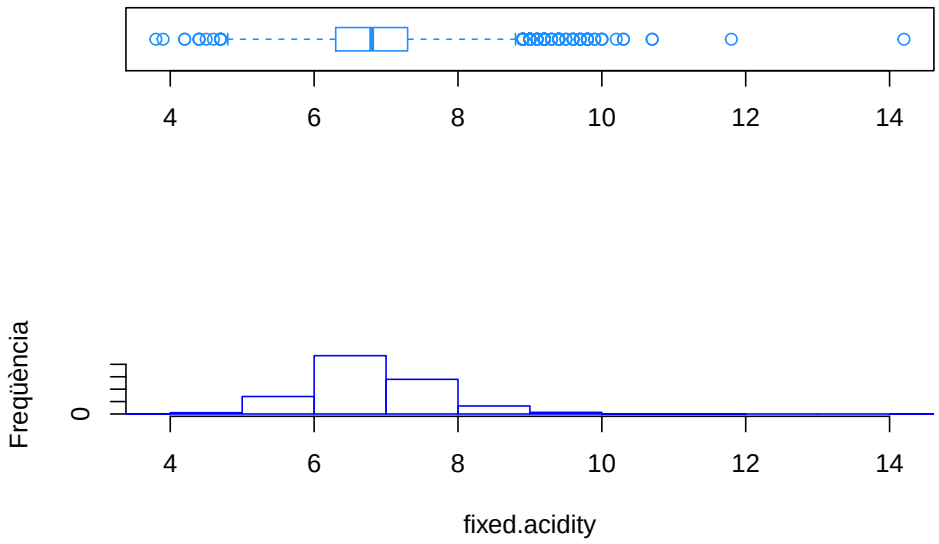


```
## [1] "Valors atípics:  "
##  [1] 12.8 12.8 15.0 15.0 12.5 13.3 13.4 12.4 12.5 13.8 13.5 12.6 12.5 12.8
## [15] 12.8 14.0 13.7 13.7 12.7 12.5 12.8 12.6 15.6 12.5 13.0 12.5 13.3 12.4
## [29] 12.5 12.9 14.3 12.4 15.5 15.5 15.6 13.0 12.7 13.0 12.7 12.4 12.7 13.2
## [43] 13.2 13.2 15.9 13.3 12.9 12.6 12.6
##
## Valors del boxplot:
## 1) Bigoti inf.: 4.6
## 2) Q1(25%)    : 7.1
## 3) Mediana    : 7.9
## 4) Q3(75%)    : 9.2
## 5) Bigoti Sup.: 12.3
```

```
cat("Vi blanc \n")
```

```
## Vi blanc
```

```
bxplot_hist(dsWhite$fixed.acidity,'fixed.acidity')
```



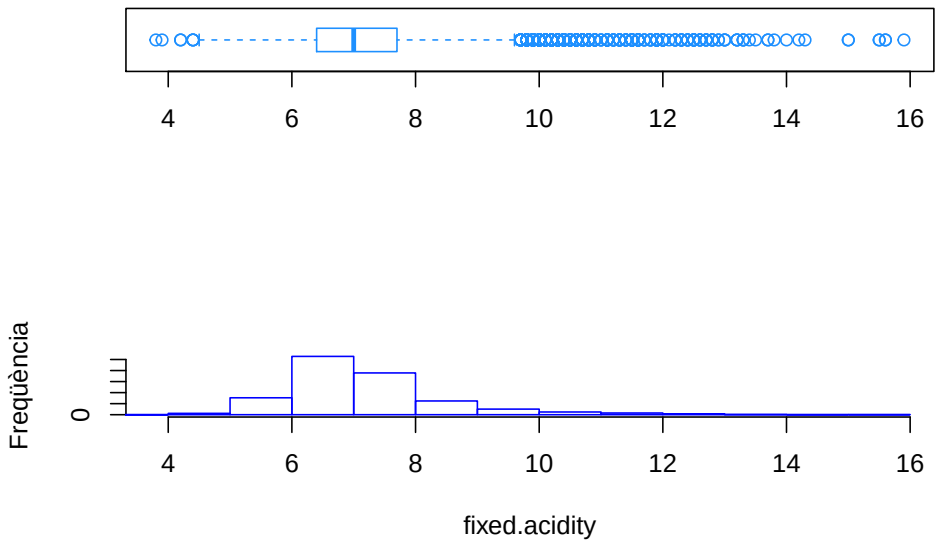

```
## [1] "Valors atípics:  "
##   [1]  9.8  9.8 10.2  9.1 10.0  9.2  9.2  9.0  9.1  9.2 10.3  9.4  9.2  9.8
##  [15]  9.6  9.2  9.0  9.3  9.2  9.1  8.9  9.8  8.9  9.2  9.7  9.4 10.3  9.6
##  [29]  9.0  9.7  9.2  9.4  9.6  9.2  9.0  9.2 10.7 10.7  9.0  9.2  9.8  9.2
##  [43] 14.2  8.9  8.9  9.1  9.1  9.8  9.0  9.3  8.9  9.0  9.0  8.9  9.0  9.3
##  [57]  9.2  9.6  9.4  9.4 10.0  8.9  8.9 10.0  9.2  9.2  9.2  9.9  9.5  9.0
##  [71]  9.0  8.9  9.5 11.8  9.4  9.1  9.8  9.9  9.2  8.9  9.2  9.4  9.4  9.4
##  [85]  4.6  8.9  9.4  9.2  9.2  9.8  9.0  9.0  9.0  8.9  8.9  4.5  9.2  9.6
##  [99]  4.2  9.7  9.7  9.0  4.2  9.4  8.9  8.9  8.9  4.7  4.7  3.8  4.4  4.7
## [113]  9.0  9.0  4.7  4.4  3.9  4.7  4.4
##
```

```
## Valors del boxplot:
## 1) Bigoti inf.: 4.8
## 2) Q1(25%)    : 6.3
## 3) Mediana    : 6.8
## 4) Q3(75%)    : 7.3
## 5) Bigoti Sup.: 8.8
```

```
cat("Vinho verde sense diferenciar color \n")
```

```
## Vinho verde sense diferenciar color
```

```
bxplot_hist(d$fixed.acidity,'fixed.acidity')
```




```
## [1] "Valors atípics:  "
##   [1] 11.2 10.2  9.7 10.1 11.5 12.8 12.8 11.0  9.7 11.6 12.0 15.0 15.0 10.8
##  [15] 11.1 10.0 12.5 11.8 11.5 11.5 10.9 11.5 10.3 11.4  9.9  9.9 12.0 11.6
##  [29] 11.0 10.4 13.3 10.8 10.6 11.1 10.3 10.3 10.3 10.3  9.8  9.8 10.3 10.0
##  [43] 10.0 11.6 10.3 13.4 10.7 10.2 10.2 11.9 12.4 12.5 12.2 10.6 10.9 10.9
##  [57] 11.9 13.8 10.7 13.5 11.5 10.5 11.9 12.6 11.9 12.5 12.8 10.0 12.8 10.4
##  [71] 10.3 14.0 11.5 11.5 11.4 13.7 13.7 12.7 12.0 11.5 11.5 12.2 11.4  9.8
##  [85] 12.0 10.4 12.5  9.9 10.6 11.9 10.5 12.8 10.5 11.9 12.3 10.4 12.3 11.1
##  [99] 10.4 12.6 11.9 15.6 10.0 12.5 11.9 11.9 10.4 11.3 10.4 11.6 11.0 11.5
## [113] 10.0 10.3 11.4 13.0 12.5  9.9 10.5 10.4 10.6 10.6 10.6 10.6 10.2 10.2
## [127] 10.2 11.6 10.7 10.7 10.4 10.4 10.5 10.5 10.2 10.4 11.2 10.0 13.3 12.4
## [141] 10.0 10.7 10.5 10.5 12.5 10.4 10.9  9.8 10.4  9.9 11.9 11.9 10.3 10.0
## [155]  9.9 12.9 11.2 11.2 14.3 10.6 12.4 15.5 15.5 10.9 15.6 10.9 13.0 12.7
## [169] 13.0 12.7  9.8 11.5 10.2 10.5 10.6 12.3  9.9 10.6 12.3 12.3 11.7 12.0
## [183] 11.8 11.1 10.2  9.9 12.4 11.9 12.7 13.2 13.2 10.1 13.2 11.5 11.4 11.3
## [197] 10.0 10.4 10.1  9.9  9.9  9.9 10.7  9.8 15.9  9.7 10.7 12.0 10.1 12.1
## [211] 11.3 10.0 11.3  9.8 10.8 10.8 10.8 13.3  9.8 11.8 10.6  9.7 10.6  9.9
```

```
## [225] 11.6 11.1  9.9  9.9 10.0 10.0 10.1 10.8 12.9 10.8 12.6 10.8  9.8 10.8
## [239] 10.4 11.6 10.1 11.1 10.6  9.9 11.7 10.4 10.7 10.7 10.1 10.0 12.0  9.9
## [253] 10.1  9.8 10.2 10.2 10.2 10.4 10.4 10.1 12.2 12.2  9.8  9.7 10.0  9.9
## [267] 10.5 11.3 11.3 10.1  9.9 11.6 10.2 11.1 11.1  9.9 10.3 11.6 11.6 10.0
## [281] 10.8 10.7 10.0 10.5 10.4 10.4 10.0 10.0 10.2 10.6  9.9  9.7  9.8 10.2
## [295]  9.9  9.9 10.2 10.9 10.9 10.5 12.6 10.2  9.8  9.8 10.4  9.8 11.3  9.7
## [309]  9.7  9.9  9.7 11.5 11.6 11.6  9.9 10.0 10.0 10.2 10.2 10.0 11.7 10.0
## [323]  9.9  9.9 11.1 11.2  9.8  9.8 10.2 10.0 10.3  9.8  9.8  9.7 10.3  9.7
## [337] 10.7 10.7  9.8 14.2  9.8 10.0 10.0  9.9 11.8  9.8  9.9  9.8  4.2  9.7
## [351]  9.7  4.2  3.8  4.4  4.4  3.9  4.4
##
## Valors del boxplot:
## 1) Bigoti inf.: 4.5
## 2) Q1(25%)    : 6.4
## 3) Mediana    : 7
## 4) Q3(75%)    : 7.7
## 5) Bigoti Sup.: 9.6
```

No tenim constància de que les dades tinguin errors de captura. Els valors poden ser correctes. Els llindars per els valors extrems serien:

```
# Cas Vi negre
# Llindar inferior per els valors atípics de 'fixed.acidity'.
cat("Vi negre\nLlindar inferior:",mean(dsRed$fixed.acidity)-3*sd(dsRed$fixed.acidity))
```

```
## Vi negre
## Llindar inferior: 3.096348
```

```
# Llindar superior per el valors atípics de 'fixed.acidity'.
cat("Vi negre\nLlindar superior:",mean(dsRed$fixed.acidity)+3*sd(dsRed$fixed.acidity))
```

```
## Vi negre
## Llindar superior: 13.54293
```

```
# Cas Vi blanc
# Llindar inferior per els valors atípics de 'fixed.acidity'.
cat("Vi blanc\nLlindar inferior:",mean(dsWhite$fixed.acidity)-3*sd(dsWhite$fixed.acidity))
```

```
## Vi blanc
## Llindar inferior: 4.323183
```

```
# Llindar superior per el valors atípics de 'fixed.acidity'.
cat("Vi negre\nLlindar superior:",mean(dsWhite$fixed.acidity)+3*sd(dsWhite$fixed.acidity))
```

```
## Vi negre
## Llindar superior: 9.386392
```

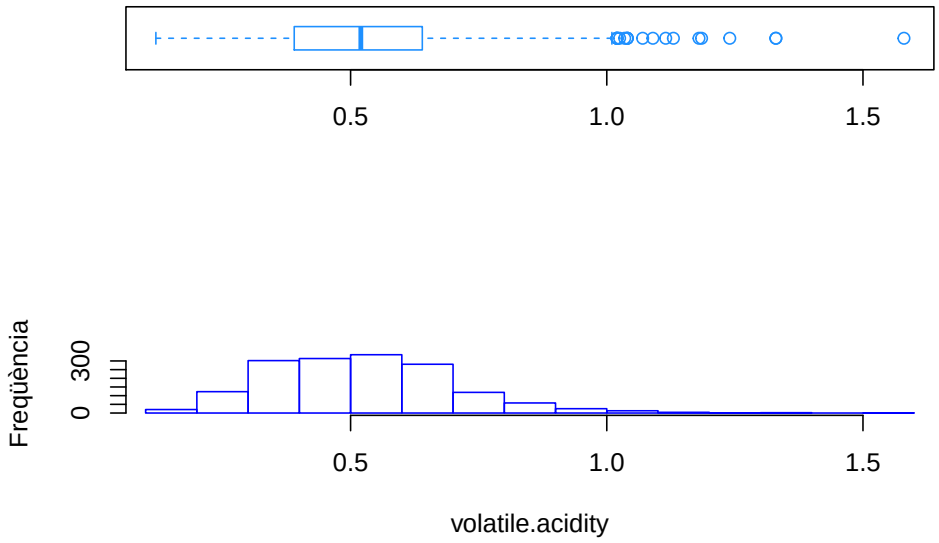
En aquest cas no modifiquem els valors extrems detectats pel camp '*fixed.acidity*'.

2.3.2.2 Valors extrems de l'atribut: volatile.acidity

```r
# Valors extrems de 'volatile.acidity'
cat("Vi negre")
```

```
## Vi negre
```

```r
bxplot_hist(dsRed$volatile.acidity,'volatile.acidity')
```

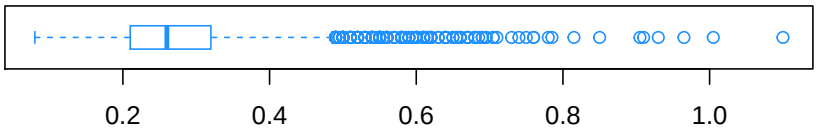


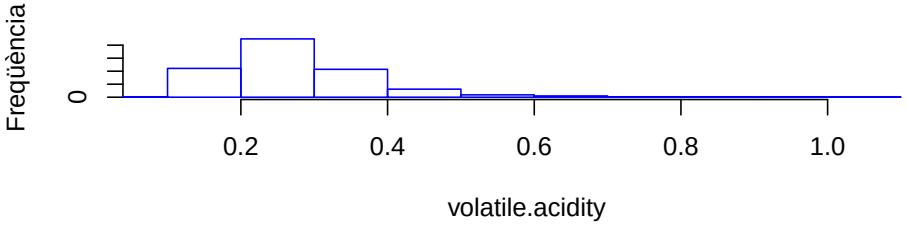

```
## [1] "Valors atípics:  "
##  [1] 1.130 1.020 1.070 1.330 1.330 1.040 1.090 1.040 1.240 1.185 1.020
## [12] 1.035 1.025 1.115 1.020 1.020 1.580 1.180 1.040
##
## Valors del boxplot:
## 1) Bigoti inf.: 0.12
## 2) Q1(25%)    : 0.39
## 3) Mediana    : 0.52
## 4) Q3(75%)    : 0.64
## 5) Bigoti Sup.: 1.01
```

```r
cat("Vi blanc")
```

```
## Vi blanc
```

```r
bxplot_hist(dsWhite$volatile.acidity,'volatile.acidity')
```






```
## [1] "Valors atípics:  "
##   [1] 0.660 0.660 0.670 0.540 0.595 0.670 0.530 0.540 0.570 0.685 0.495
##  [12] 0.640 0.520 0.580 0.585 0.590 0.600 0.580 0.590 0.550 0.905 0.550
##  [23] 0.490 0.550 0.520 0.600 0.550 0.510 0.620 0.510 0.560 0.570 0.670
##  [34] 0.500 0.560 0.560 0.655 0.595 0.705 0.520 0.550 0.600 0.640 0.680
##  [45] 0.490 0.510 0.550 0.520 0.500 0.550 0.600 0.610 0.610 0.610 0.660
##  [56] 0.570 0.500 0.500 0.590 0.580 0.540 0.580 0.570 0.640 0.560 0.490
##  [67] 0.490 0.670 0.550 0.560 0.520 0.520 0.850 0.510 0.620 0.510 0.530
##  [78] 0.640 0.550 0.490 0.490 0.610 0.545 0.620 0.490 0.500 0.490 0.490
##  [89] 0.550 0.490 0.910 0.530 0.490 0.710 1.005 0.490 0.550 0.550 0.760
## [100] 0.500 0.930 0.490 0.495 0.695 0.705 0.815 0.560 0.560 0.560 0.510
## [111] 0.540 0.540 0.500 0.615 0.500 0.520 0.600 0.680 0.655 0.510 0.510
## [122] 0.615 0.615 0.965 0.740 0.530 0.780 0.680 0.640 0.540 0.750 0.640
## [133] 0.640 0.655 0.580 0.520 0.530 0.600 0.530 0.580 0.670 0.610 0.730
## [144] 0.650 0.580 1.100 0.500 0.500 0.500 0.650 0.520 0.550 0.585 0.560
## [155] 0.555 0.555 0.540 0.610 0.550 0.530 0.660 0.615 0.500 0.620 0.500
## [166] 0.490 0.510 0.510 0.540 0.610 0.695 0.695 0.630 0.630 0.690 0.690
## [177] 0.590 0.620 0.785 0.760 0.500 0.540 0.520 0.600 0.540 0.530
##
## Valors del boxplot:
## 1) Bigoti inf.: 0.08
## 2) Q1(25%)    : 0.21
## 3) Mediana    : 0.26
## 4) Q3(75%)    : 0.32
## 5) Bigoti Sup.: 0.485
```

```
# Cas Vi negre
# Llindar inferior per els valors atípics de 'volatile.acidity'.
cat("Vi negre\nLlindar inferior:",mean(dsRed$volatile.acidity)-3*sd(dsRed$volatile.acidity))
```

```
## Vi negre
```

```
## Llindar inferior: -0.0093586
```

```
# Llindar superior per el valors atípics de 'volatile.acidity'.
cat("Vi negre\nLlindar superior:",mean(dsRed$volatile.acidity)+3*sd(dsRed$volatile.acidity))
```

```
## Vi negre
## Llindar superior: 1.065
```

```
# Cas Vi blanc
# Llindar inferior per els valors atípics de 'volatile.acidity'.
cat("Vi blanc\nLlindar inferior:",mean(dsWhite$volatile.acidity)-3*sd(dsWhite$volatile.acidity))
```

```
## Vi blanc
## Llindar inferior: -0.02414253
```

```
# Llindar superior per el valors atípics de 'volatile.acidity'.
cat("Vi negre\nLlindar superior:",mean(dsWhite$volatile.acidity)+3*sd(dsWhite$volatile.acidity))
```

```
## Vi negre
## Llindar superior: 0.5806248
```

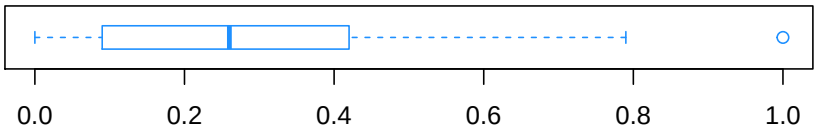
La variació d'aquests valors no afecta sensiblement als paràmetres representatius de la seva distribució, però si ho necessitem més endavant podriem substituïr el seu valor per el de la mediana 0.52, com a valor representatiu de la seva distribució. Mètodes més robustos com la imputació de dades amb kNN, package R "DMwR" mitjançant la función **knnImputation()**, o el package R "VIM" usant la funció **kNN()** també són una opció.

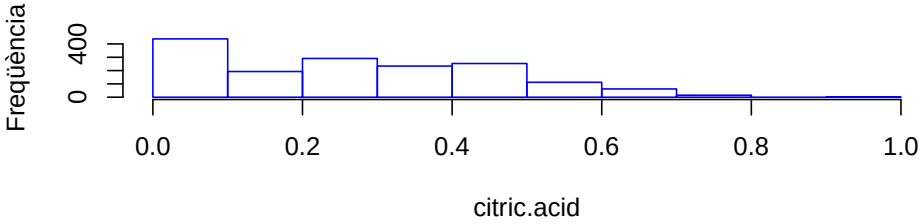
2.3.2.3 Valors extrems de l'atribut: citric.acid

```
cat("Vi negre")
```

```
## Vi negre
```

```
bxplot_hist(dsRed$citric.acid,'citric.acid')
```



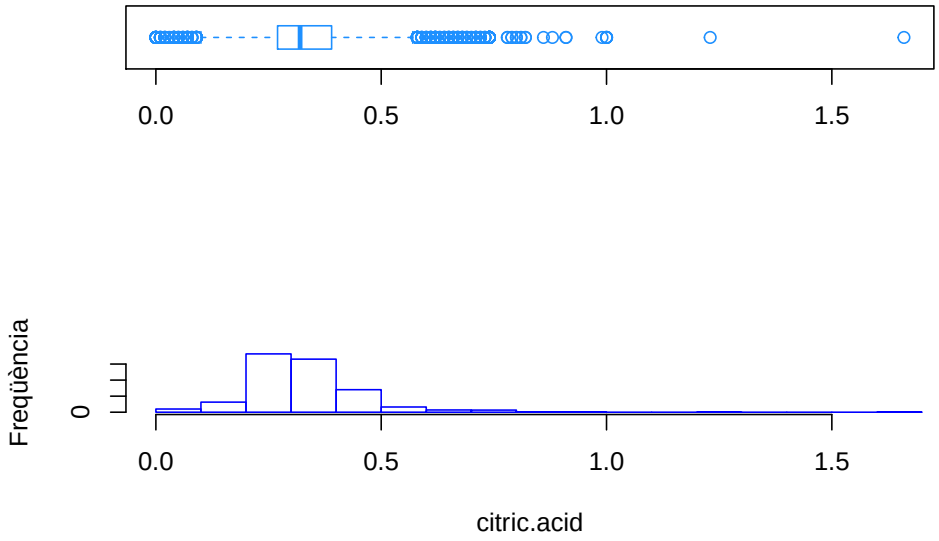



```
## [1] "Valors atípics:  "
## [1] 1
##
## Valors del boxplot:
## 1) Bigoti inf.: 0
## 2) Q1(25%)    : 0.09
## 3) Mediana    : 0.26
## 4) Q3(75%)    : 0.42
## 5) Bigoti Sup.: 0.79
```

```
cat("Vi blanc")
```

```
## Vi blanc
```

```
bxplot_hist(dsWhite$citric.acid,'citric.acid')
```




```
## [1] "Valors atípics:  "
##   [1] 0.62 0.04 0.59 0.07 0.03 0.61 0.62 0.63 0.61 0.62 0.63 0.66 0.66 0.00
##  [15] 0.04 0.67 0.67 0.04 0.04 0.07 0.88 0.08 0.59 0.07 0.07 0.07 0.07 0.58
##  [29] 0.70 0.00 0.00 0.60 0.07 0.09 0.04 0.62 0.58 0.62 0.70 0.62 0.62 0.58
##  [43] 0.02 0.65 0.65 0.71 0.66 0.66 0.07 0.06 0.07 0.06 0.68 0.68 0.68 0.68
##  [57] 0.06 0.72 0.69 0.58 0.70 1.66 0.04 0.63 0.60 0.00 0.08 0.58 0.58 0.05
##  [71] 0.58 0.00 0.00 0.65 0.58 0.00 0.05 0.05 0.62 0.62 0.58 0.58 1.00 0.09
##  [85] 0.01 0.71 0.71 0.60 0.06 0.74 0.81 0.69 0.58 0.69 0.00 0.07 0.64 0.72
##  [99] 0.73 0.65 0.68 0.65 0.74 0.71 0.59 0.68 0.08 0.72 0.64 0.02 0.74 0.74
## [113] 0.74 0.74 0.74 0.74 0.74 0.74 0.74 0.74 0.74 0.74 0.74 0.74 0.74 0.74
## [127] 0.74 0.74 0.74 0.74 0.74 0.99 0.74 0.74 0.74 0.74 0.74 0.74 0.74 0.74
## [141] 0.74 0.74 0.74 0.74 0.74 0.74 0.01 0.74 0.01 0.74 0.74 1.00 0.04 0.58
## [155] 0.07 1.00 0.00 0.58 0.61 0.61 0.61 0.02 0.67 0.67 0.67 0.58 0.65 0.58
## [169] 0.09 0.08 0.71 0.04 0.03 0.05 0.64 0.64 0.58 0.58 0.81 0.58 0.61 0.62
## [183] 0.59 0.00 0.04 0.63 0.73 0.68 0.09 0.78 0.79 0.09 0.64 0.65 0.65 0.00
## [197] 0.73 0.73 0.64 0.60 0.71 0.72 0.82 0.07 0.58 0.58 1.00 0.66 0.80 0.80
## [211] 1.23 0.59 0.02 0.00 1.00 0.62 0.00 0.71 0.71 0.71 0.61 0.61 0.00 0.60
## [225] 0.58 0.09 0.09 0.72 0.62 0.62 0.79 0.82 0.67 0.01 0.01 0.86 0.61 0.02
## [239] 0.05 0.00 0.69 0.69 0.59 0.01 0.66 0.66 0.78 0.00 0.04 0.91 0.91 0.06
## [253] 0.06 0.04 0.04 0.74 0.09 0.09 0.60 0.62 0.73 0.00 0.09 0.00 0.09 0.67
## [267] 0.01 0.09 0.00 0.02
##
## Valors del boxplot:
## 1) Bigoti inf.: 0.1
## 2) Q1(25%)    : 0.27
## 3) Mediana    : 0.32
## 4) Q3(75%)    : 0.39
## 5) Bigoti Sup.: 0.57
```

```r
# Cas Vi negre
# Llindar inferior per els valors atípics de 'citric.acid'.
cat("Vi negre\nLlindar inferior:",mean(dsRed$citric.acid)-3*sd(dsRed$citric.acid))
```

```
## Vi negre
## Llindar inferior: -0.3134278
```

```r
# Llindar superior per el valors atípics de 'citric.acid'.
cat("Vi negre\nLlindar superior:",mean(dsRed$citric.acid)+3*sd(dsRed$citric.acid))
```

```
## Vi negre
## Llindar superior: 0.855379
```

```r
# Cas Vi blanc
# Llindar inferior per els valors atípics de 'citric.acid'.
cat("Vi blanc\nLlindar inferior:",mean(dsWhite$citric.acid)-3*sd(dsWhite$citric.acid))
```

```
## Vi blanc
## Llindar inferior: -0.02886791
```

```r
# Llindar superior per el valors atípics de 'citric.acid'.
cat("Vi negre\nLlindar superior:",mean(dsWhite$citric.acid)+3*sd(dsWhite$citric.acid))
```

```
## Vi negre
## Llindar superior: 0.6972509
```

En aquest cas, suposem que tot i que hem trobat valors atípics per el vi negre, amb valor de citric.acid=1.0, podrien ser correctes en cassos on es busca donar frescor al vi, (característica d'aquests vins) i tampoc modifiquem el valor. Els valors atípics per el vi blanc estàn dintre dels valors correctes del vi negre, i poden ser correctes per tant els mantenim.

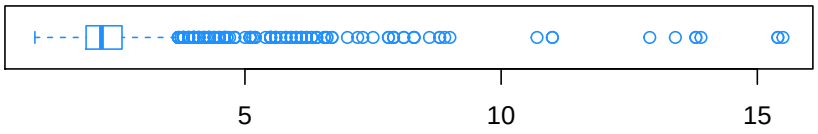
Prenem nota de que el nombre de zeros, és alt (132), respecte el total de instàncies 1599 en el cas del vi negre. Podria ser que la característica de 'frescor' fos més freqüent en aquest tipus de vi.

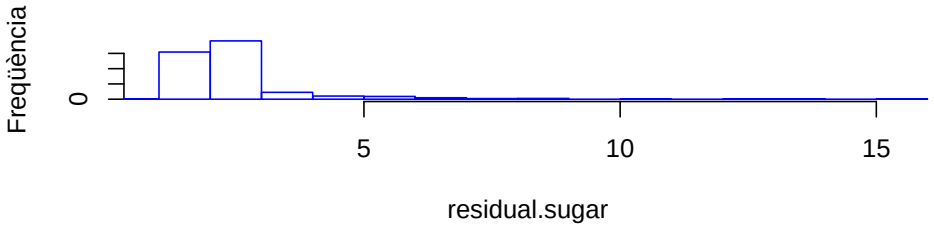
2.3.2.4 Valors extrems de l'atribut: residual.sugar

```r
# Valors extrems de 'residual.sugar'
cat("Vi negre")
```

```
## Vi negre
```

```r
bxplot_hist(dsRed$residual.sugar,'residual.sugar')
```



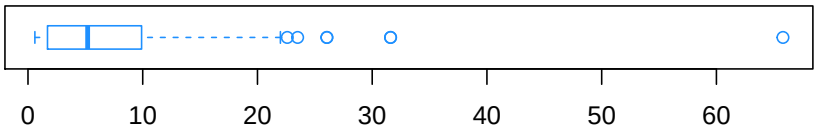


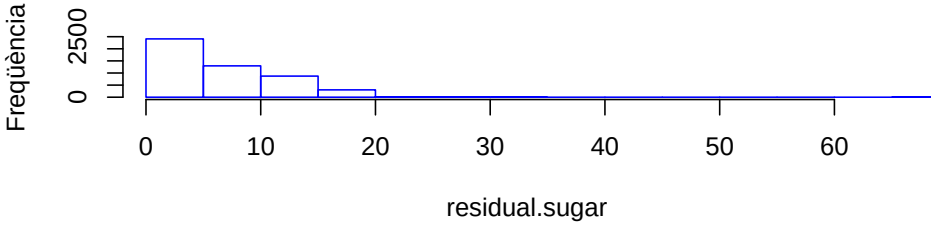

```
## [1] "Valors atípics:  "
##   [1]  6.10  6.10  3.80  3.90  4.40 10.70  5.50  5.90  5.90  3.80  5.10
##  [12]  4.65  4.65  5.50  5.50  5.50  5.50  7.30  7.20  3.80  5.60  4.00
##  [23]  4.00  4.00  4.00  7.00  4.00  4.00  6.40  5.60  5.60 11.00 11.00
##  [34]  4.50  4.80  5.80  5.80  3.80  4.40  6.20  4.20  7.90  7.90  3.70
##  [45]  4.50  6.70  6.60  3.70  5.20 15.50  4.10  8.30  6.55  6.55  4.60
##  [56]  6.10  4.30  5.80  5.15  6.30  4.20  4.20  4.60  4.20  4.60  4.30
##  [67]  4.30  7.90  4.60  5.10  5.60  5.60  6.00  8.60  7.50  4.40  4.25
##  [78]  6.00  3.90  4.20  4.00  4.00  4.00  6.60  6.00  6.00  3.80  9.00
##  [89]  4.60  8.80  8.80  5.00  3.80  4.10  5.90  4.10  6.20  8.90  4.00
## [100]  3.90  4.00  8.10  8.10  6.40  6.40  8.30  8.30  4.70  5.50  5.50
## [111]  4.30  5.50  3.70  6.20  5.60  7.80  4.60  5.80  4.10 12.90  4.30
## [122] 13.40  4.80  6.30  4.50  4.50  4.30  4.30  3.90  3.80  5.40  3.80
## [133]  6.10  3.90  5.10  5.10  3.90 15.40 15.40  4.80  5.20  5.20  3.75
## [144] 13.80 13.80  5.70  4.30  4.10  4.10  4.40  3.70  6.70 13.90  5.10
## [155]  7.80
##
## Valors del boxplot:
## 1) Bigoti inf.: 0.9
## 2) Q1(25%)    : 1.9
## 3) Mediana    : 2.2
## 4) Q3(75%)    : 2.6
## 5) Bigoti Sup.: 3.65
```

```
cat("Vi blanc")
```

```
## Vi blanc
```

```
bxplot_hist(dsWhite$residual.sugar,'residual.sugar')
```






```
## [1] "Valors atípics:  "
## [1] 23.50 31.60 31.60 65.80 26.05 26.05 22.60
##
## Valors del boxplot:
## 1) Bigoti inf.: 0.6
## 2) Q1(25%)    : 1.7
## 3) Mediana    : 5.2
## 4) Q3(75%)    : 9.9
## 5) Bigoti Sup.: 22
```

```
# Cas Vi negre
# Llindar inferior per els valors atípics de 'residual.sugar'.
cat("Vi negre\nLlindar inferior:",mean(dsRed$residual.sugar)-3*sd(dsRed$residual.sugar))
```

```
## Vi negre
## Llindar inferior: -1.690979
```

```
# Llindar superior per el valors atípics de 'residual.sugar'.
cat("Vi negre\nLlindar superior:",mean(dsRed$residual.sugar)+3*sd(dsRed$residual.sugar))
```

```
## Vi negre
## Llindar superior: 6.76859
```

```
# Cas Vi blanc
# Llindar inferior per els valors atípics de 'residual.sugar'.
cat("Vi blanc\nLlindar inferior:",mean(dsWhite$residual.sugar)-3*sd(dsWhite$residual.sugar))
```

```
## Vi blanc
## Llindar inferior: -8.824758
```

```r
# Llindar superior per el valors atípics de 'residual.sugar'.
cat("Vi negre\nLlindar superior:",mean(dsWhite$residual.sugar)+3*sd(dsWhite$residual.sugar))
```

```
## Vi negre
## Llindar superior: 21.60759
```

En aquest cas hem trobat valors atípics molt extrems.

Si tenim en compte l'escala de dolçor dels vins:

- 0-9 g/l Sucre residual (SR): Sec
- 9-18 g/l: No Sec
- 18-50 g/l: Mig sec-Semi dolç
- 50-120 g/l: mig-dolç
- 120 g/l Dolç-Molt dolç

Deduïm que el 'Vinho verde' és un vi que pot ser 'Sec', 'No sec', 'Mig sec' però no 'Semi dolç', 'Mig dolç' o 'Dolç-Molt dolç'. Per tant establim un llindar lògic de residual.sugar=18. **Decidim eliminar aquests valors atípics** que hem trobat en el vi blanc, per considerar que són poques instàncies, que representen un 2% del total d'instàncies de vi blanc. Hem observat que la majoria de valors eliminats tenia un valor de qualitat 5,6,7,8 i això implicaria que un valor incorrecte del 'sucre residual' no fos una característica que ens ajudes a discriminar la qualitat del vi.

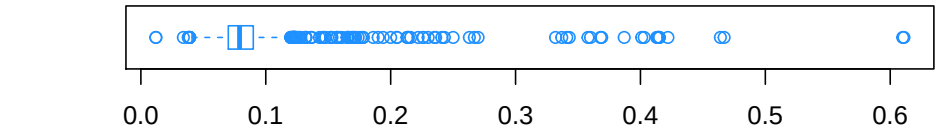
```r
# Eliminem les instàncies amb valors atípics incorrectes.
dsWhite<-dsWhite[dsWhite$residual.sugar<=18,]
```

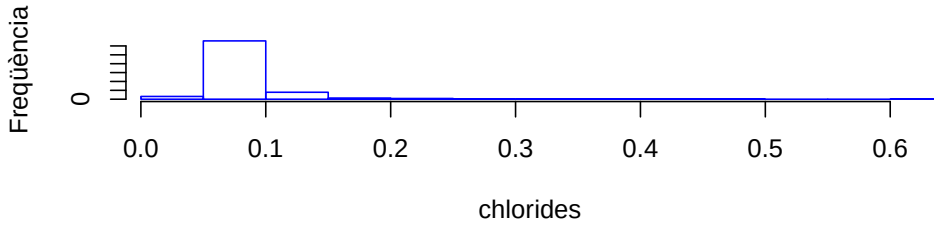
2.3.2.5 Valors extrems de l'atribut: chlorides

```r
# Valors extrems de 'chlorides'
cat("Vi negre")
```

```
## Vi negre
```

```r
bxplot_hist(dsRed$chlorides,'chlorides')
```



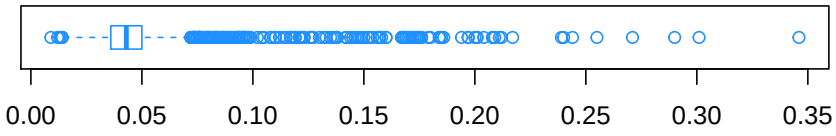


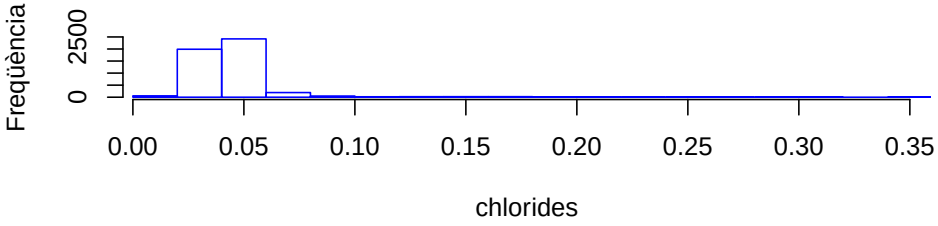

```
## [1] "Valors atípics:  "
##   [1] 0.176 0.170 0.368 0.341 0.172 0.332 0.464 0.401 0.467 0.122 0.178
##  [12] 0.146 0.236 0.610 0.360 0.270 0.039 0.337 0.263 0.611 0.358 0.343
##  [23] 0.186 0.213 0.214 0.121 0.122 0.122 0.128 0.120 0.159 0.124 0.122
##  [34] 0.122 0.174 0.121 0.127 0.413 0.152 0.152 0.125 0.122 0.200 0.171
##  [45] 0.226 0.226 0.250 0.148 0.122 0.124 0.124 0.143 0.222 0.039 0.157
##  [56] 0.422 0.034 0.387 0.415 0.157 0.157 0.243 0.241 0.190 0.132 0.126
##  [67] 0.038 0.165 0.145 0.147 0.012 0.012 0.039 0.194 0.132 0.161 0.120
##  [78] 0.120 0.123 0.123 0.414 0.216 0.171 0.178 0.369 0.166 0.166 0.136
##  [89] 0.132 0.132 0.123 0.123 0.123 0.403 0.137 0.414 0.166 0.168 0.415
## [100] 0.153 0.415 0.267 0.123 0.214 0.214 0.169 0.205 0.205 0.039 0.235
## [111] 0.230 0.038
##
## Valors del boxplot:
## 1) Bigoti inf.: 0.041
## 2) Q1(25%)    : 0.07
## 3) Mediana    : 0.079
## 4) Q3(75%)    : 0.09
## 5) Bigoti Sup.: 0.119
```

```
cat("Vi blanc")
```

```
## Vi blanc
```

```
bxplot_hist(dsWhite$chlorides,'chlorides')
```






```
## [1] "Valors atípics:  "
##   [1] 0.074 0.080 0.172 0.173 0.147 0.092 0.082 0.092 0.200 0.197 0.197
##  [12] 0.074 0.132 0.089 0.108 0.081 0.073 0.346 0.090 0.114 0.186 0.180
##  [23] 0.084 0.083 0.096 0.094 0.240 0.290 0.185 0.110 0.078 0.130 0.135
##  [34] 0.115 0.072 0.170 0.080 0.119 0.126 0.150 0.152 0.088 0.244 0.137
##  [45] 0.093 0.077 0.079 0.073 0.072 0.076 0.201 0.201 0.074 0.074 0.301
##  [56] 0.138 0.169 0.083 0.093 0.168 0.122 0.172 0.167 0.239 0.076 0.138
##  [67] 0.137 0.123 0.123 0.133 0.073 0.073 0.211 0.123 0.123 0.255 0.204
##  [78] 0.208 0.083 0.080 0.076 0.086 0.168 0.160 0.179 0.076 0.076 0.087
##  [89] 0.217 0.094 0.157 0.157 0.148 0.158 0.157 0.168 0.157 0.092 0.099
## [100] 0.084 0.085 0.091 0.093 0.080 0.095 0.096 0.096 0.147 0.142 0.079
## [111] 0.075 0.121 0.121 0.079 0.079 0.014 0.156 0.012 0.119 0.119 0.081
## [122] 0.170 0.171 0.082 0.152 0.169 0.073 0.014 0.078 0.112 0.154 0.126
## [133] 0.126 0.104 0.142 0.102 0.184 0.184 0.096 0.076 0.146 0.117 0.117
## [144] 0.118 0.014 0.087 0.087 0.076 0.088 0.160 0.167 0.014 0.009 0.098
## [155] 0.098 0.086 0.086 0.194 0.094 0.013 0.144 0.149 0.185 0.084 0.175
## [166] 0.090 0.098 0.110 0.110 0.095 0.174 0.097 0.142 0.145 0.208 0.209
## [177] 0.105 0.086 0.176 0.176 0.108 0.096 0.271 0.212 0.094 0.094 0.117
## [188] 0.173 0.074 0.076 0.076 0.175 0.174 0.075 0.127 0.127 0.096 0.136
##
## Valors del boxplot:
## 1) Bigoti inf.: 0.015
## 2) Q1(25%)    : 0.036
## 3) Mediana    : 0.043
## 4) Q3(75%)    : 0.05
## 5) Bigoti Sup.: 0.071
```

```r
# Cas Vi negre
# Llindar inferior per els valors atípics de 'chlorides'.
cat("Vi negre\nLlindar inferior:",mean(dsRed$chlorides)-3*sd(dsRed$chlorides))
```

```
## Vi negre
## Llindar inferior: -0.05372936
```

```
# Llindar superior per el valors atípics de 'chlorides'.
cat("Vi negre\nLlindar superior:",mean(dsRed$chlorides)+3*sd(dsRed$chlorides))
```

```
## Vi negre
## Llindar superior: 0.2286624
```

```
# Cas Vi blanc
# Llindar inferior per els valors atípics de 'chlorides'.
cat("Vi blanc\nLlindar inferior:",mean(dsWhite$chlorides)-3*sd(dsWhite$chlorides))
```

```
## Vi blanc
## Llindar inferior: -0.02010456
```

```
# Llindar superior per el valors atípics de 'chlorides'.
cat("Vi negre\nLlindar superior:",mean(dsWhite$chlorides)+3*sd(dsWhite$chlorides))
```

```
## Vi negre
## Llindar superior: 0.1114649
```

En el cas dels clorurs també trobem valors extrems molt alts. Per tant actuem de la mateixa manera que amb el sucre residual. **Eliminem aquells valors que no estan en un rang de valors adient per Europa**, ja que sabem que els vins són de Portugal i aquest valor de clorurs és força atípic.

El percentatge d'instàncies eliminades no és gaire gran:

- **2%** (vi negre).
- **2%** (vi blanc).

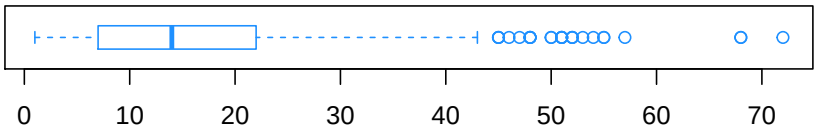
```
# Eliminem les instàncies amb valors atípics incorrectes.
dsRed<-dsRed[dsRed$chlorides>0 & dsRed$chlorides<=0.22,]
dsWhite<-dsWhite[dsWhite$chlorides>0 & dsWhite$chlorides<=0.11,]
```

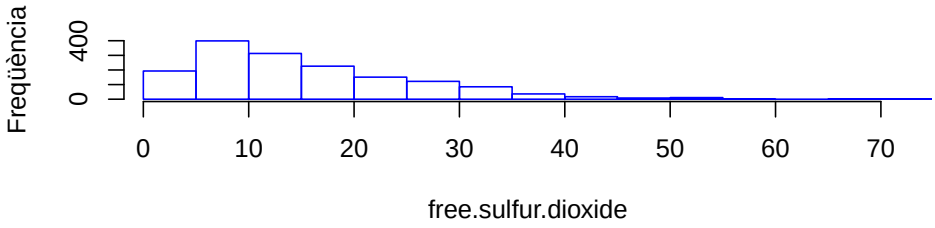
2.3.2.6 Valors extrems de l'atribut: free.sulfur.dioxide

```
# Valors extrems de 'free.sulfur.dioxide'
cat("Vi negre")
```

```
## Vi negre
```

```
bxplot_hist(dsRed$free.sulfur.dioxide,'free.sulfur.dioxide')
```



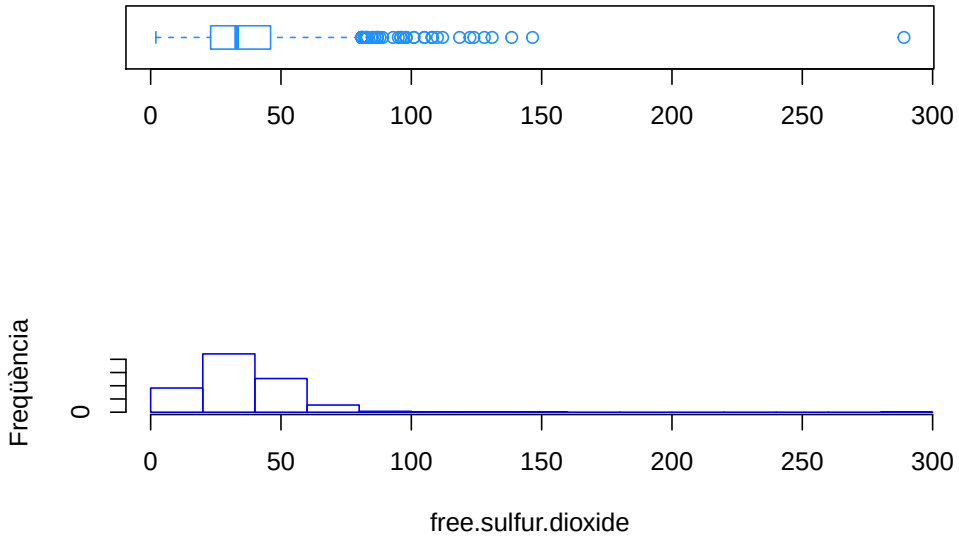



```
## [1] "Valors atípics:  "
##  [1] 52 51 50 68 68 47 54 46 45 53 52 51 45 57 50 45 48 48 72 51 51 52 55
## [24] 55 48 48
##
## Valors del boxplot:
## 1) Bigoti inf.: 1
## 2) Q1(25%)    : 7
## 3) Mediana    : 14
## 4) Q3(75%)    : 22
## 5) Bigoti Sup.: 43
```

```
cat("Vi blanc")
```

```
## Vi blanc
```

```
bxplot_hist(dsWhite$free.sulfur.dioxide,'free.sulfur.dioxide')
```




```
## [1] "Valors atípics:  "
##  [1]  81.0  82.0 131.0  82.5  87.0  87.0  83.0 122.5  83.0  81.0  88.0
## [12]  82.0 118.5  81.0  96.0  83.0  83.0 146.5 128.0 110.0  85.0  89.0
## [23]  86.0  86.0  96.0  96.0  93.0  85.0  81.0 138.5  95.0 124.0  87.0
## [34]  87.0 105.0 105.0 101.0 101.0 108.0 108.0  98.0  98.0 112.0 108.0
## [45]  98.0  81.0  81.0  81.0 289.0  97.0
##
## Valors del boxplot:
## 1) Bigoti inf.: 2
## 2) Q1(25%)    : 23
## 3) Mediana    : 33
## 4) Q3(75%)    : 46
## 5) Bigoti Sup.: 80
```

```r
# Cas Vi negre
# Llindar inferior per els valors atípics de 'free.sulfur.dioxide'.
cat("Vi negre\nLlindar inferior:",
    mean(dsRed$free.sulfur.dioxide)-3*sd(dsRed$free.sulfur.dioxide))
```

```
## Vi negre
## Llindar inferior: -15.43324
```

```r
# Llindar superior per el valors atípics de 'free.sulfur.dioxide'.
cat("Vi negre\nLlindar superior:",
    mean(dsRed$free.sulfur.dioxide)+3*sd(dsRed$free.sulfur.dioxide))
```

```
## Vi negre
## Llindar superior: 47.22813
```

```r
# Cas Vi blanc
# Llindar inferior per els valors atípics de 'free.sulfur.dioxide'.
cat("Vi blanc\nLlindar inferior:",
    mean(dsWhite$free.sulfur.dioxide)-3*sd(dsWhite$free.sulfur.dioxide))
```

```
## Vi blanc
## Llindar inferior: -15.91874
```

```r
# Llindar superior per el valors atípics de 'free.sulfur.dioxide'.
cat("Vi negre\nLlindar superior:",
    mean(dsWhite$free.sulfur.dioxide)+3*sd(dsWhite$free.sulfur.dioxide))
```

```
## Vi negre
## Llindar superior: 86.13236
```

En el cas del '*free.sulfur.dioxide*', sabem que la normativa admet fins valors de **300 g/l** tot i que a Europa no es solen observar valors superiors a **50 g/l**.

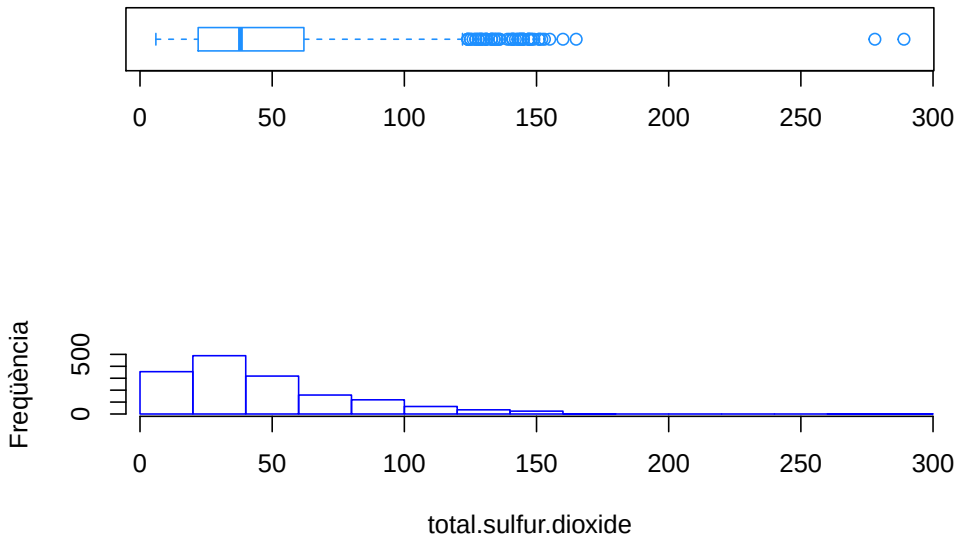
Observem que els valors atípics tenen valors alts sobretot en el vi blanc, però com no sobrepasen la normativa i poden ser correctes i els valors estan representats en totes les categories de qualitat del vi, **decidim no modificar les dades**. L'ús del '*free.sulfur.dioxide*' en el vi és com a conservant i volem mantenir l'efecte de que és més usat en el vi blanc.

2.3.2.7 Valors extrems de l'atribut: total.sulfur.dioxide

```r
# Valors extrems de 'total.sulfur.dioxide'
cat("Vi negre")
```

```
## Vi negre
```

```r
bxplot_hist(dsRed$total.sulfur.dioxide,'total.sulfur.dioxide')
```

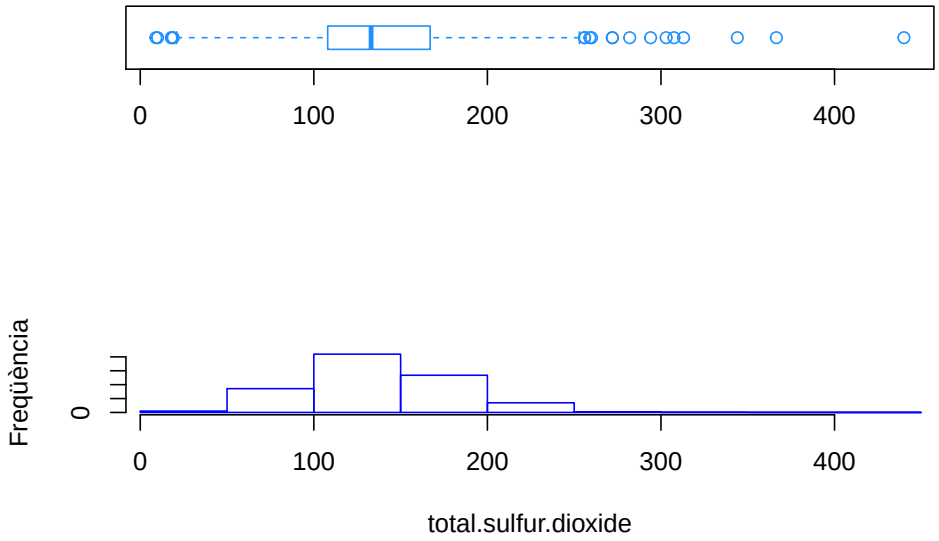


```
## [1] "Valors atípics:  "
##  [1] 145 148 136 125 140 136 133 153 134 141 129 128 129 128 143 144 127
## [18] 126 145 144 135 165 124 124 134 124 129 151 133 142 149 147 145 148
## [35] 155 151 152 125 127 139 143 144 130 278 289 135 160 141 141 133 147
## [52] 147 131 131 131
##
## Valors del boxplot:
## 1) Bigoti inf.: 6
## 2) Q1(25%)    : 22
## 3) Mediana    : 38
## 4) Q3(75%)    : 62
## 5) Bigoti Sup.: 122
```

```
cat("Vi blanc")
```

```
## Vi blanc
```

```
bxplot_hist(dsWhite$total.sulfur.dioxide,'total.sulfur.dioxide')
```




```
## [1] "Valors atípics:  "
##  [1] 272.0 313.0 260.0  19.0 366.5 307.5 256.0 256.0 344.0 282.0 303.0
## [12] 272.0  18.0  18.0 294.0   9.0  10.0 259.0 440.0
##
## Valors del boxplot:
## 1) Bigoti inf.: 21
## 2) Q1(25%)    : 108
## 3) Mediana    : 133
## 4) Q3(75%)    : 167
## 5) Bigoti Sup.: 253
```

```r
# Cas Vi negre
# Llindar inferior per els valors atípics de 'total.sulfur.dioxide'.
cat("Vi negre\nLlindar inferior:",
    mean(dsRed$total.sulfur.dioxide)-3*sd(dsRed$total.sulfur.dioxide))
```

```
## Vi negre
## Llindar inferior: -52.64119
```

```r
# Llindar superior per el valors atípics de 'total.sulfur.dioxide'.
cat("Vi negre\nLlindar superior:",
    mean(dsRed$total.sulfur.dioxide)+3*sd(dsRed$total.sulfur.dioxide))
```

```
## Vi negre
## Llindar superior: 145.4719
```

```r
# Cas Vi blanc
# Llindar inferior per els valors atípics de 'total.sulfur.dioxide'.
cat("Vi blanc\nLlindar inferior:",
    mean(dsWhite$total.sulfur.dioxide)-3*sd(dsWhite$total.sulfur.dioxide))
```

```
## Vi blanc
## Llindar inferior: 10.43815
```

```r
# Llindar superior per el valors atípics de 'total.sulfur.dioxide'.
cat("Vi negre\nLlindar superior:",
    mean(dsWhite$total.sulfur.dioxide)+3*sd(dsWhite$total.sulfur.dioxide))
```

```
## Vi negre
## Llindar superior: 265.0597
```

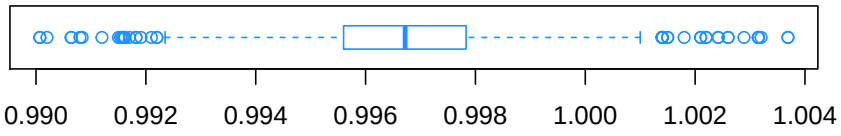
En aquest obtenim valors força alts però sabem que aquest camp s'obté de la suma del diòxid de sofre lliure i el fixat, per tant **preveiem que aquest atribut no l'utilitzem en els anàlisis**, per ser combinació lineal de '*free.sulfur.dioxide*'.

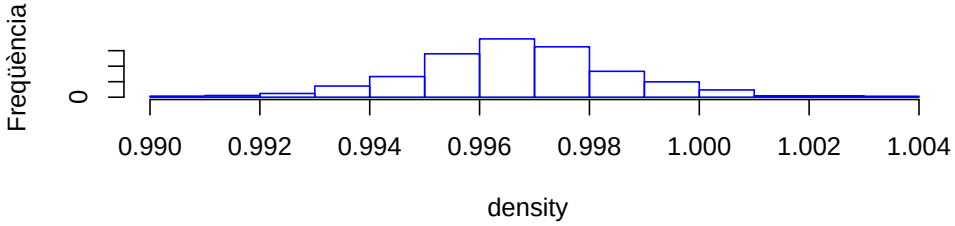
2.3.2.8 Valors extrems de l'atribut: density

```r
# Valors extrems de 'density'
cat("Vi negre")
```

```
## Vi negre
```

```r
bxplot_hist(dsRed$density,'density')
```



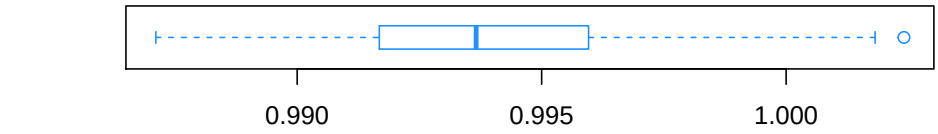


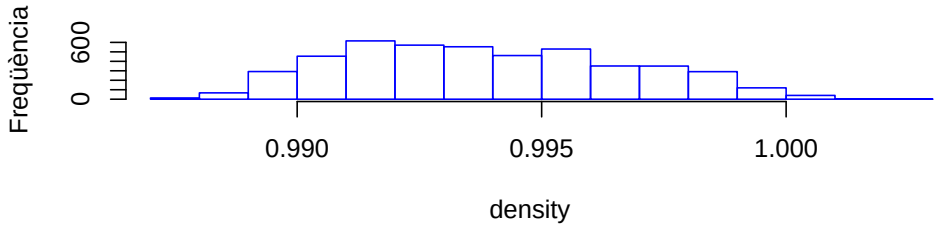

```
## [1] "Valors atípics:  "
##  [1] 0.99160 0.99160 1.00140 1.00150 1.00150 1.00180 0.99120 1.00220
##  [9] 1.00220 1.00140 1.00140 1.00140 1.00140 1.00320 1.00260 1.00140
## [17] 1.00315 1.00315 1.00315 1.00210 1.00210 0.99170 0.99220 1.00260
## [25] 0.99210 0.99154 0.99064 0.99064 1.00289 0.99162 0.99007 0.99007
## [33] 0.99020 0.99220 0.99150 0.99157 0.99080 0.99084 0.99191 1.00369
## [41] 1.00369 1.00242 0.99182 1.00242 0.99182
##
## Valors del boxplot:
## 1) Bigoti inf.: 0.99235
## 2) Q1(25%)    : 0.9956
## 3) Mediana    : 0.99672
## 4) Q3(75%)    : 0.99783
## 5) Bigoti Sup.: 1.001
```

```
cat("Vi blanc")
```

```
## Vi blanc
```

```
bxplot_hist(dsWhite$density,'density')
```






```
## [1] "Valors atípics:  "
## [1] 1.00241
##
## Valors del boxplot:
## 1) Bigoti inf.: 0.98711
## 2) Q1(25%)    : 0.99168
## 3) Mediana    : 0.99366
## 4) Q3(75%)    : 0.99596
## 5) Bigoti Sup.: 1.00182
```

```
# Cas Vi negre
# Llindar inferior per els valors atípics de 'density'.
cat("Vi negre\nLlindar inferior:",mean(dsRed$density)-3*sd(dsRed$density))
```

```
## Vi negre
## Llindar inferior: 0.9910428
```

```
# Llindar superior per el valors atípics de 'density'.
cat("Vi negre\nLlindar superior:",mean(dsRed$density)+3*sd(dsRed$density))
```

```
## Vi negre
## Llindar superior: 1.002437
```

```
# Cas Vi blanc
# Llindar inferior per els valors atípics de 'density'.
cat("Vi blanc\nLlindar inferior:",mean(dsWhite$density)-3*sd(dsWhite$density))
```

```
## Vi blanc
## Llindar inferior: 0.9853813
```

```r
# Llindar superior per el valors atípics de 'density'.
cat("Vi negre\nLlindar superior:",mean(dsWhite$density)+3*sd(dsWhite$density))
```

```
## Vi negre
## Llindar superior: 1.002433
```

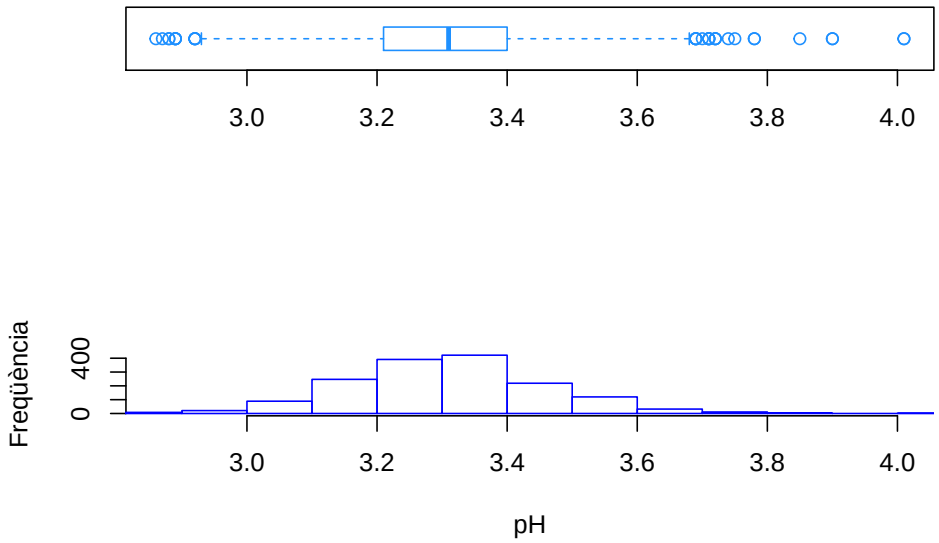
En el cas de la '*densitat*', obtenim valors atípics, però no estem segurs de que siguin incorrectes. Les distribucions queden ben descrites per els paràmetres estadístics com la mediana. No volem perdre informació del dataset i per tant **no eliminem ni modifiquem cap valor**.

2.3.2.9 Valors extrems de l'atribut: pH

```r
# Valors extrems de 'pH'
cat("Vi negre")
```

```
## Vi negre
```

```r
bxplot_hist(dsRed$pH,'pH')
```

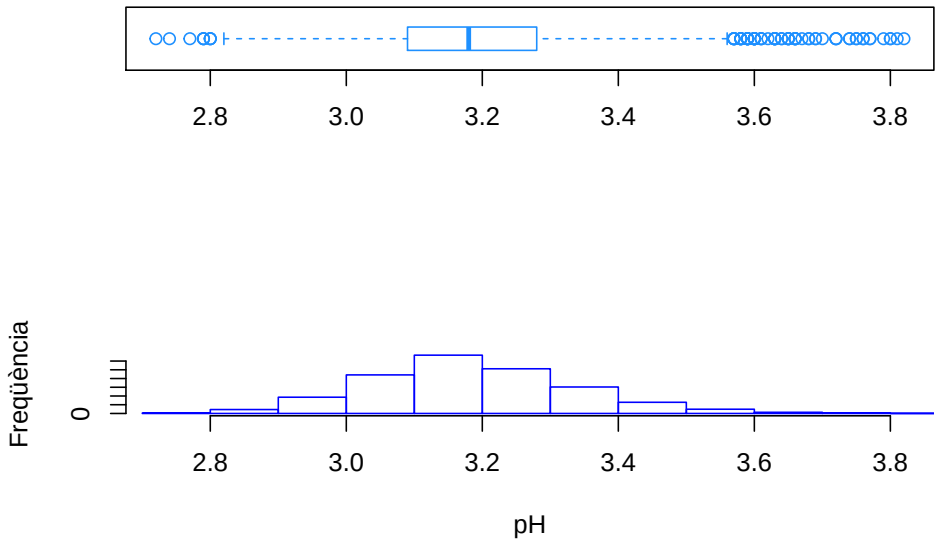



```
## [1] "Valors atípics:  "
##  [1] 3.90 3.75 3.85 3.69 3.69 2.88 2.86 3.74 2.92 2.92 2.92 3.72 2.87 2.89
## [15] 2.89 2.92 3.90 3.71 3.69 3.69 3.71 3.71 2.89 2.89 3.78 3.70 3.78 4.01
## [29] 4.01 3.71 2.88 3.72 3.72
##
## Valors del boxplot:
## 1) Bigoti inf.: 2.93
## 2) Q1(25%)    : 3.21
## 3) Mediana    : 3.31
## 4) Q3(75%)    : 3.4
## 5) Bigoti Sup.: 3.68
```

```r
cat("Vi blanc")
```

```
## Vi blanc
```

```r
bxplot_hist(dsWhite$pH,'pH')
```




```
## [1] "Valors atípics:  "
##  [1] 3.69 3.63 3.72 3.61 3.64 3.64 3.72 3.72 3.58 3.58 3.66 3.59 2.74 3.82
## [15] 3.81 3.65 3.65 3.59 3.77 3.62 3.63 3.58 3.58 3.65 3.74 2.80 3.60 3.60
## [29] 2.72 3.60 2.79 2.79 3.57 3.80 3.60 3.60 3.68 3.63 3.63 2.77 3.63 3.60
## [43] 3.60 3.61 3.61 3.59 3.79 3.59 3.68 3.59 3.66 3.70 3.74 3.80 3.57 3.57
## [57] 3.57 3.65 3.58 2.80 3.77 3.76 3.69 3.66 3.59 2.79 3.75 3.63 3.75 3.76
## [71] 3.66 3.66 2.80 3.67 3.57
##
## Valors del boxplot:
## 1) Bigoti inf.: 2.82
## 2) Q1(25%)    : 3.09
## 3) Mediana    : 3.18
## 4) Q3(75%)    : 3.28
## 5) Bigoti Sup.: 3.56
```

```r
# Cas Vi negre
# Llindar inferior per els valors atípics de 'pH'.
cat("Vi negre\nLlindar inferior:",mean(dsRed$pH)-3*sd(dsRed$pH))
```

```
## Vi negre
## Llindar inferior: 2.858448
```

```r
# Llindar superior per el valors atípics de 'pH'.
cat("Vi negre\nLlindar superior:",mean(dsRed$pH)+3*sd(dsRed$pH))
```

```
## Vi negre
## Llindar superior: 3.771546
```

```r
# Cas Vi blanc
# Llindar inferior per els valors atípics de 'pH'.
cat("Vi blanc\nLlindar inferior:",mean(dsWhite$pH)-3*sd(dsWhite$pH))
```

```
## Vi blanc
## Llindar inferior: 2.737161
```

```r
# Llindar superior per el valors atípics de 'pH'.
cat("Vi negre\nLlindar superior:",mean(dsWhite$pH)+3*sd(dsWhite$pH))
```

```
## Vi negre
## Llindar superior: 3.645711
```

En el cas del '*pH*', és igual que el cas anterior, **podem mantenir els valors atípics**, ja que poden ser valors correctes i no tenim seguretat d'error. En general abans de modificar les dades cal estar segur de que realment és necessari, perque podriem estar eliminant patrons en les dades que per no semblar-nos 'coherents' d'entrada, són valors que indiquen una variació en una tendència incipient.

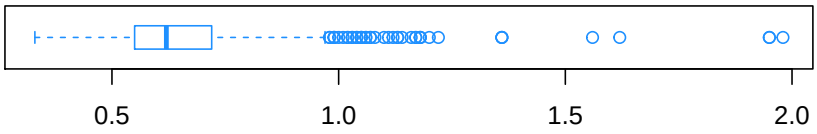
Sabem que el 'pH' marca la acidesa del vi, per tant és probable que el 'pH' tingui alguna correlació amb els atributs 'fixed.acidity' o 'citric.acid'.

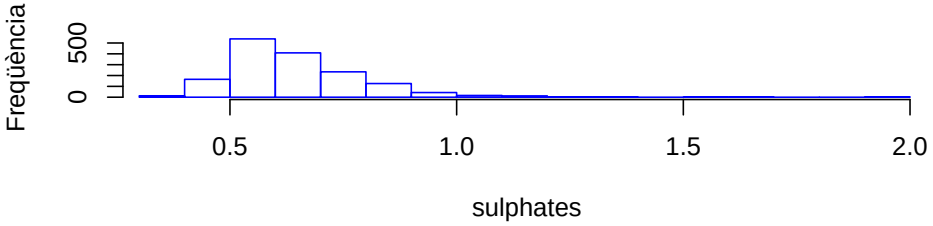
2.3.2.10 Valors extrems de l'atribut: sulphates

```r
# Valors extrems de 'sulphates'
cat("Vi negre")
```

```
## Vi negre
```

```r
bxplot_hist(dsRed$sulphates,'sulphates')
```



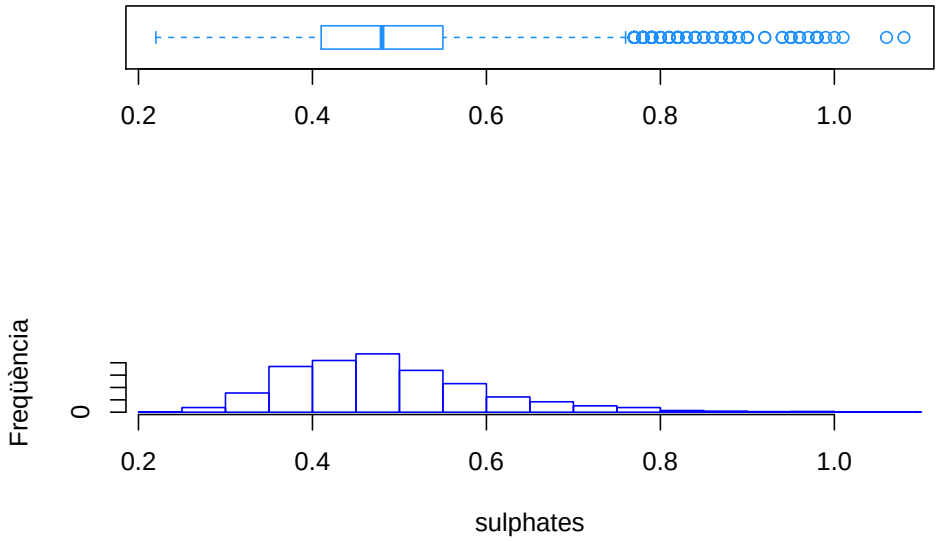



```
## [1] "Valors atípics:  "
##  [1] 1.56 1.20 1.12 1.95 1.22 1.95 1.98 1.08 1.03 1.00 1.36 1.18 0.98 1.13
## [15] 1.04 1.11 1.13 0.99 1.07 1.06 1.05 1.06 1.04 1.05 1.02 1.14 0.99 1.02
## [29] 1.36 1.36 1.05 1.62 1.18 1.07 0.99 1.16 0.98 1.17 1.17 1.18 1.03 1.10
## [43] 1.01
##
## Valors del boxplot:
## 1) Bigoti inf.: 0.33
## 2) Q1(25%)    : 0.55
## 3) Mediana    : 0.62
## 4) Q3(75%)    : 0.72
## 5) Bigoti Sup.: 0.97
```

```
cat("Vi blanc")
```

```
## Vi blanc
```

```
bxplot_hist(dsWhite$sulphates,'sulphates')
```




```
## [1] "Valors atípics:  "
##   [1] 0.77 0.84 0.77 0.79 0.85 0.78 0.79 0.79 0.79 0.77 0.78 0.85 0.96 0.97
##  [15] 0.82 0.82 0.77 0.95 0.95 0.77 0.95 0.82 0.82 0.90 0.88 0.88 0.79 0.80
##  [29] 0.80 0.78 0.78 0.87 0.86 0.90 0.90 0.78 0.79 0.81 0.81 0.77 0.82 0.79
##  [43] 0.79 0.77 0.82 0.92 0.79 0.79 0.82 0.82 0.82 0.82 0.82 0.79 0.78 0.79
##  [57] 0.77 0.77 0.77 0.98 1.06 0.88 0.88 0.88 0.80 0.78 1.00 0.80 0.90 0.90
##  [71] 0.89 0.94 0.99 0.86 0.84 0.95 0.84 0.84 0.81 0.80 0.87 0.82 0.78 0.78
##  [85] 0.78 0.78 0.78 0.77 0.85 0.78 0.78 0.88 0.88 0.78 0.78 0.78 0.78 0.79
##  [99] 0.77 0.77 0.83 0.83 0.81 0.81 0.98 0.98 0.98 0.98 0.79 0.79 0.78 0.82
## [113] 0.98 0.77 0.96 1.01 0.77 0.96 0.77 0.92 0.94 0.95 1.08 0.79
##
## Valors del boxplot:
## 1) Bigoti inf.: 0.22
## 2) Q1(25%)    : 0.41
## 3) Mediana    : 0.48
## 4) Q3(75%)    : 0.55
## 5) Bigoti Sup.: 0.76
```

```r
# Cas Vi negre
# Llindar inferior per els valors atípics de 'sulphates'.
cat("Vi negre\nLlindar inferior:",mean(dsRed$sulphates)-3*sd(dsRed$sulphates))
```

```
## Vi negre
## Llindar inferior: 0.1880322
```

```r
# Llindar superior per el valors atípics de 'sulphates'.
cat("Vi negre\nLlindar superior:",mean(dsRed$sulphates)+3*sd(dsRed$sulphates))
```

```
## Vi negre
## Llindar superior: 1.111942
```

```r
# Cas Vi blanc
# Llindar inferior per els valors atípics de 'sulphates'.
cat("Vi blanc\nLlindar inferior:",mean(dsWhite$sulphates)-3*sd(dsWhite$sulphates))
```

```
## Vi blanc
## Llindar inferior: 0.1453524
```

```r
# Llindar superior per el valors atípics de 'sulphates'.
cat("Vi negre\nLlindar superior:",mean(dsWhite$sulphates)+3*sd(dsWhite$sulphates))
```

```
## Vi negre
## Llindar superior: 0.836141
```

Els 'sulphates' i calci apareixen en l'aigua i per tant el raïm i el vi poden contenir-los. Una alta concentració dels mateixos podria fer que una aigua no fos de qualitat alimentària. Una aigua amb una quantitat de sulfats inferior a 250mg /l es considera en aquest aspecte una aigua de qualitat i amb valors superiors a 400 mg / l insalubre.

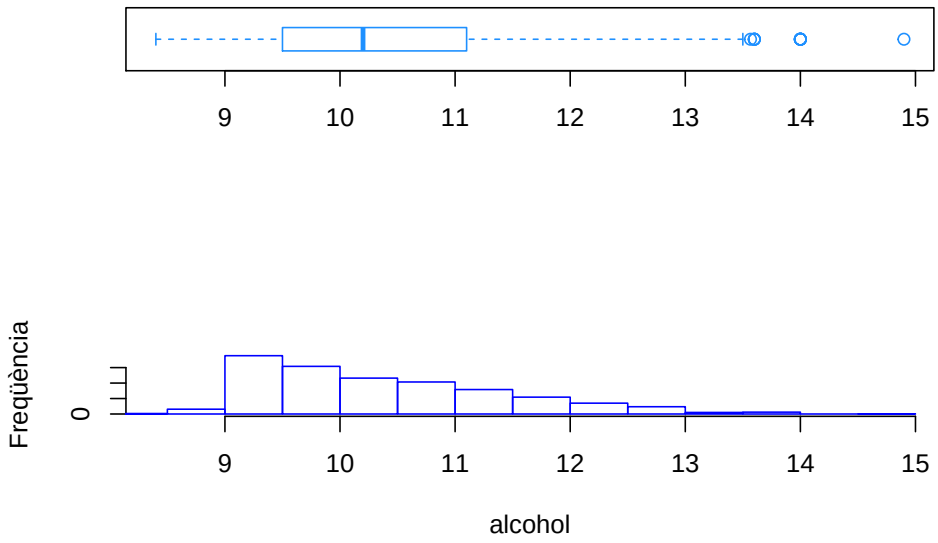
Per tant suposem aquest camp important per determinar la qualitat del vi. Igual que en els dos darrers cassos no trobem valors atípics que no puguin ser correctes, i per tant **conservem tots els valors**.

2.3.2.11 Valors extrems de l'atribut: alcohol

```r
# Valors extrems de 'alcohol'
cat("Vi negre")
```

```
## Vi negre
```

```r
bxplot_hist(dsRed$alcohol,'alcohol')
```

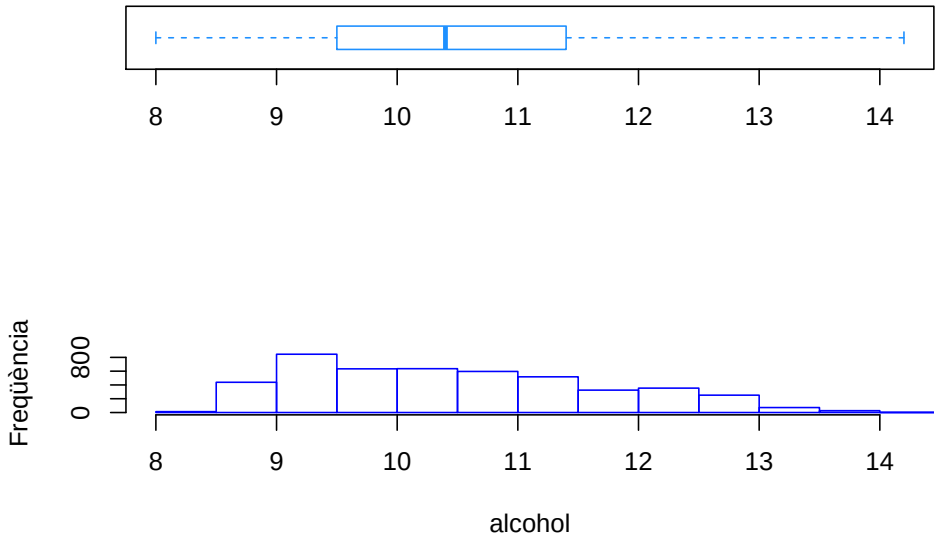


```
## [1] "Valors atípics:  "
##  [1] 14.00000 14.00000 14.00000 14.00000 14.90000 14.00000 13.60000
##  [8] 13.60000 13.60000 14.00000 14.00000 13.56667 13.60000
##
## Valors del boxplot:
## 1) Bigoti inf.: 8.4
## 2) Q1(25%)    : 9.5
## 3) Mediana    : 10.2
## 4) Q3(75%)    : 11.1
## 5) Bigoti Sup.: 13.5
```

```
cat("Vi blanc")
```

```
## Vi blanc
```

```
bxplot_hist(dsWhite$alcohol,'alcohol')
```




```
## [1] "Valors atípics:  "
## numeric(0)
##
## Valors del boxplot:
## 1) Bigoti inf.: 8
## 2) Q1(25%)    : 9.5
## 3) Mediana    : 10.4
## 4) Q3(75%)    : 11.4
## 5) Bigoti Sup.: 14.2
```

```
# Cas Vi negre
# Llindar inferior per els valors atípics de 'alcohol'.
cat("Vi negre\nLlindar inferior:",mean(dsRed$alcohol)-3*sd(dsRed$alcohol))
```

```
## Vi negre
## Llindar inferior: 7.244888
```

```
# Llindar superior per el valors atípics de 'alcohol'.
cat("Vi negre\nLlindar superior:",mean(dsRed$alcohol)+3*sd(dsRed$alcohol))
```

```
## Vi negre
## Llindar superior: 13.63748
```

```
# Cas Vi blanc
# Llindar inferior per els valors atípics de 'alcohol'.
cat("Vi blanc\nLlindar inferior:",mean(dsWhite$alcohol)-3*sd(dsWhite$alcohol))
```

```
## Vi blanc
## Llindar inferior: 6.875105
```

```
# Llindar superior per el valors atípics de 'alcohol'.
cat("Vi negre\nLlindar superior:",mean(dsWhite$alcohol)+3*sd(dsWhite$alcohol))
```

```
## Vi negre
## Llindar superior: 14.23365
```

En de l'alcohol, tampoc trobem valors atípics que no puguin ser correctes, i per tant **conservem tots els valors**. En aquest cas observem l'antiguitat del dataset perquè actualment la tendència en els darrers anys en la gradució dels vins és tà per sobre de 13º, (sobretot en el vi negre).

2.3.3 Defenció del dataset definitiu.

Un cop tractats els valors buits, zeros i identificats i tractats els valors atípics de tots els atributs, definim una nova versió del dataset de vins.

```
# Tenim capçaleres iguals, si la suma de noms iguals és la suma total de camps.
a<-colnames(dsRed)
b<-colnames(dsWhite)
cat(paste0("El nombre de camps (",ncol(dsRed),") és igual al nombre de camps iguals ",
           sum(a==b)))
```

```
## El nombre de camps (13) és igual al nombre de camps iguals 13
```

```
# Combinem les mostres dels dos fitxers i factoritzem el camp color per determinar
# els valors que pren: 'blanc i 'negre'
d<-rbind(dsRed,dsWhite)
d$color<-factor(d$color)

# Dimensions del dataset:
cat("Files - Instàncies:",nrow(d),"\nColumnes-Atributs-Variables:",ncol(d),"\n")
```

```
## Files - Instàncies: 6279
## Columnes-Atributs-Variables: 13
```

```r
# Comprovem els valors del dataset:
head(d)
```

```
##   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1           7.4             0.70        0.00            1.9     0.076
## 2           7.8             0.88        0.00            2.6     0.098
## 3           7.8             0.76        0.04            2.3     0.092
## 4          11.2             0.28        0.56            1.9     0.075
## 5           7.4             0.70        0.00            1.9     0.076
## 6           7.4             0.66        0.00            1.8     0.075
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  11                   34  0.9978 3.51      0.56     9.4
## 2                  25                   67  0.9968 3.20      0.68     9.8
## 3                  15                   54  0.9970 3.26      0.65     9.8
## 4                  17                   60  0.9980 3.16      0.58     9.8
## 5                  11                   34  0.9978 3.51      0.56     9.4
## 6                  13                   40  0.9978 3.51      0.56     9.4
##   quality color
## 1       5 negre
## 2       5 negre
## 3       5 negre
## 4       6 negre
## 5       5 negre
## 6       5 negre
```

```r
tail(d)
```

```
##      fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 4893           6.5             0.23        0.38            1.3     0.032
## 4894           6.2             0.21        0.29            1.6     0.039
## 4895           6.6             0.32        0.36            8.0     0.047
## 4896           6.5             0.24        0.19            1.2     0.041
## 4897           5.5             0.29        0.30            1.1     0.022
## 4898           6.0             0.21        0.38            0.8     0.020
##      free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates
## 4893                  29                  112 0.99298 3.29      0.54
## 4894                  24                   92 0.99114 3.27      0.50
## 4895                  57                  168 0.99490 3.15      0.46
## 4896                  30                  111 0.99254 2.99      0.46
## 4897                  20                  110 0.98869 3.34      0.38
## 4898                  22                   98 0.98941 3.26      0.32
##      alcohol quality color
## 4893     9.7       5 blanc
## 4894    11.2       6 blanc
## 4895     9.6       5 blanc
## 4896     9.4       6 blanc
## 4897    12.8       7 blanc
## 4898    11.8       6 blanc
```

2.4 Anàlisi de les dades.

2.4.1 Selecció dels grups de dades i planificació dels anàlisis.

En primer lloc, comprovarem si el color del vi influeix en la seva qualitat.

Després volem extreure patrons de correlació entre atributs, i per això calculem una matriu de correlació usant l'index de correlació de Pearson. L'ordre en que les característiques són llistades en la matriu, és correspón amb una ordenació de component principal (paràmetre **order**).

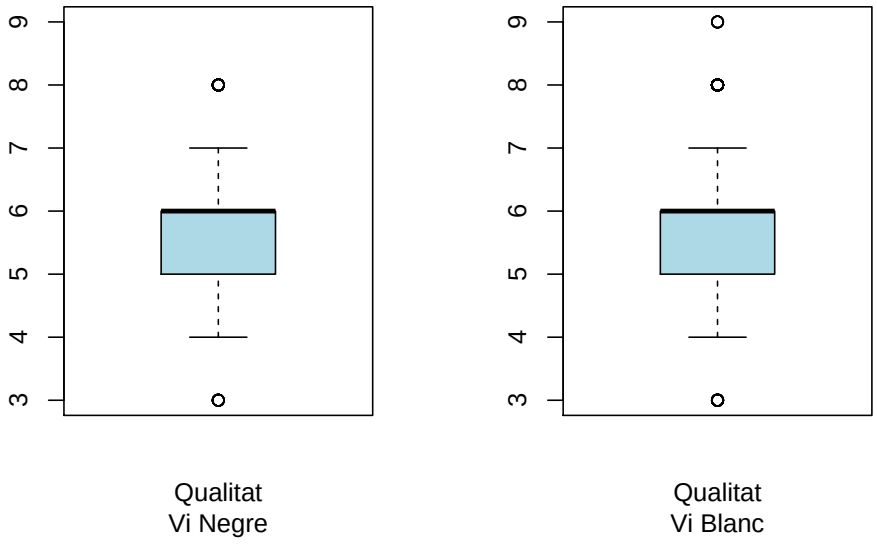
També definirem grups en l'atribut més correlacionat amb la qualitat del vi i compararem estadísticament la qualitat del vi per cadascuna de les categories definides.

Finalment crearem un model de regressió lineal multiple utilitzant 5 regressors per tal de poder realitzar prediccions de la qualitat del vi.

Primer de tot analitzem la variable de qualitat del vi per saber si depen del color:

```
# Visualitzem el boxplot de la qualitat segons color del vi.
par(mfrow=c(1,2))

# Vi negre
boxplot(d[d$color=='negre',"quality"], ylim=c(3,9), col="lightblue", xlab="Qualitat\nVi Negre")
# Vi blanc
boxplot(d[d$color=='blanc',"quality"], col="lightblue", xlab="Qualitat\nVi Blanc")
```



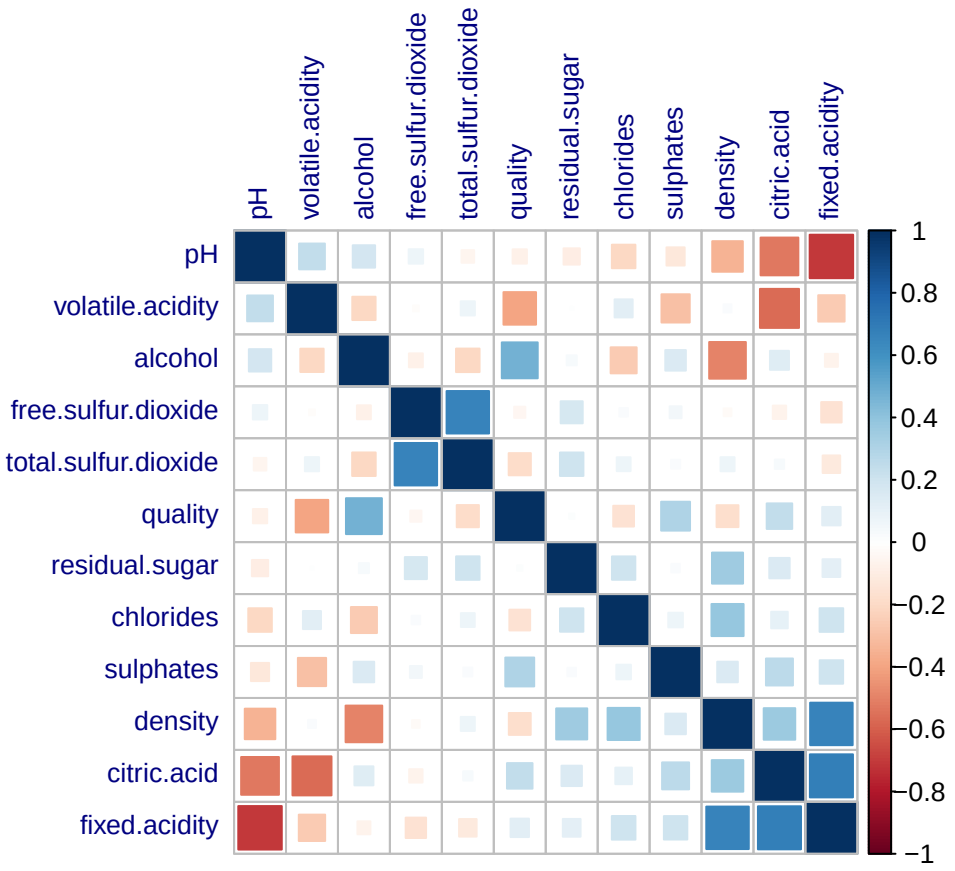

```
par(mfrow=c(1,1))
```

Hem elaborat els boxplot de cada tipus de vi en paral · lel per poder-los comparar bé. Observem que **amb les dades que disposem**, no hin han diferències, es a dir coneixent el color del vi no podem determinar la seva qualitat. Necessitem cercar informació de la qualitat en els elements de la seva composició.

Calculem ara les matrius de correlació del vi negre i el vi blanc, usant l'index de correlació de Pearson.

2.4.1.1 Selecció dels grups de dades del Vi Negre

```
# Vi negre
nc=ncol(dsRed)
df <- dsRed[,1:11]
df$quality <- as.integer(dsRed[,12])
correlacio <- cor(df,method="pearson")
corrplot(correlacio, number.cex = 2, method = "square", hclust.method = "ward", order = "FPC",
         type = "full", tl.cex=0.8,tl.col = "navy")
```



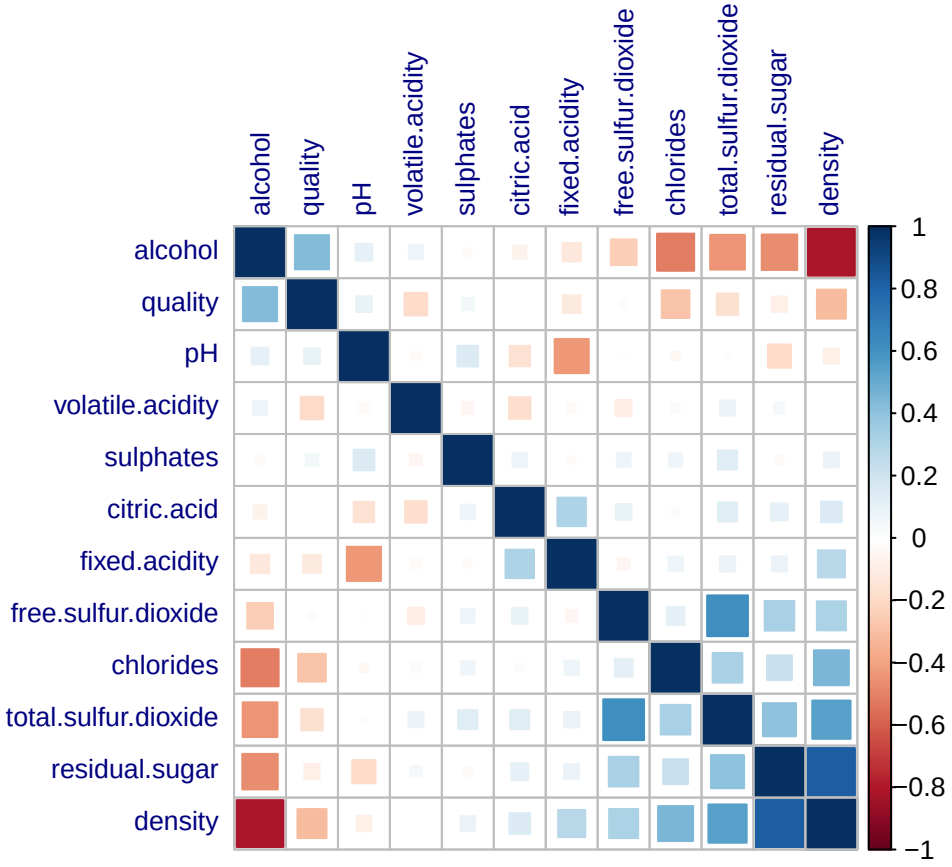
De la matriu anterior podem establir les següents correlacions entre atributs per el vi negre:

Tipus de correlació	Atributs que correlacionen
Alta correlació	total.sulfur.dioxide amb free.sulfur.dioxide
Alta correlació	fixed.acidity amb density i citric.acid
Correlació relativa	alcohol amb qualitat
Alta correlació inversa	fixed.acidity amb pH
Correlació relativa inversa	citric.acid amb pH i volatile.acidity

2.4.1.2 Selecció dels grups de dades del Vi Blanc

```
# Vi blanc
nc=ncol(dsWhite)
df <- dsWhite[,1:11]
```

```r
df$quality <- as.integer(dsWhite[,12])
correlacio <- cor(df,method="pearson")
corrplot(correlacio, number.cex = 2, method = "square", hclust.method = "ward", order = "FPC",
         type = "full", tl.cex=0.8,tl.col = "navy")
```



De la matriu anterior podem establir les següents correlacions entre atributs per el vi blanc:

Tipus de correlació	Atributs que correlacionen
Alta correlació	total.sulfur.dioxide amb free.sulfur.dioxide
Alta correlació	densitat amb residual.sugar
Correlació relativa	alcohol amb qualitat
Alta correlació inversa	densitat amb alcohol
Correlació relativa inversa	alcohol amb residual.sugar
Correlació relativa inversa	total.sulfur.dioxide i fixed.acidity amb pH

2.4.1.3 Planificació de l'anàlisi

Amb els resultats dels dos apartats anteriors decidim seleccionar els atributs com característiques representatives de la qualitat del vi:

- **alcohol**
- **volatile.acidity** i
- **sulfats**

Table 4: Taula 1: Classificació de la graduació dels vins.

Molt.Baixa	Moderadament.baixa	Alta	Molt.Alta
< 12.5%	[12.5%, 13.5%)	[13.5%, 14.5%]	> 14.5%

Són les característiques que més correlació o correlació inversa tenen amb els dos tipus de vins.

En les següents proves estadístiques es compararà el valor de qualitat del vi blanc i negre amb l'**alcohol** que és la característica amb la que té més correlació.

Establim les següents categories:

```
ah_Class<-data.frame("Molt Baixa"="< 12.5%",
                     "Moderadament baixa"="[12.5%, 13.5%)",
                     "Alta"="[13.5%, 14.5%]",
                     "Molt Alta"=" > 14.5%")
kable(ah_Class,caption=paste0("Taula ",1,": Classificació de la graduació dels vins."),
      align=c('c','c','c','c')) %>%
  column_spec(1,  color='black', bold = T, background = "orange") %>%
  column_spec(2,  color='white', bold = T, background = "blue") %>%
  column_spec(3,  color='white', bold = T, background = "blue") %>%
  column_spec(4,  color='black', bold = T, background = "orange") %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed", "responsive"), full_width = F)
```

```
# Definició de grups de graduació d'alcohol en el vi.
d1<-d
d1["alcohol_class"]<-""
d1[d1$alcohol<9.5,'alcohol_class']<-"Molt_Baixa"
d1[d1$alcohol>=9.5 & d1$alcohol<12.5,'alcohol_class']<-"Moderadament_Baixa"
d1[d1$alcohol>=12.5 & d1$alcohol<=14,'alcohol_class']<-"Alta"
d1[d1$alcohol>14,'alcohol_class']<-"Molt_Alta"

summary(d[,c("alcohol","color","quality")])
```

```
##     alcohol        color         quality
##  Min.   : 8.00   blanc:4714   Min.   :3.000
##  1st Qu.: 9.50   negre:1565   1st Qu.:5.000
##  Median :10.40                Median :6.000
##  Mean   :10.53                Mean   :5.831
##  3rd Qu.:11.30                3rd Qu.:6.000
##  Max.   :14.90                Max.   :9.000
```

2.4.2 Comprovació de la normalitat i homogeneïtat de la variància.

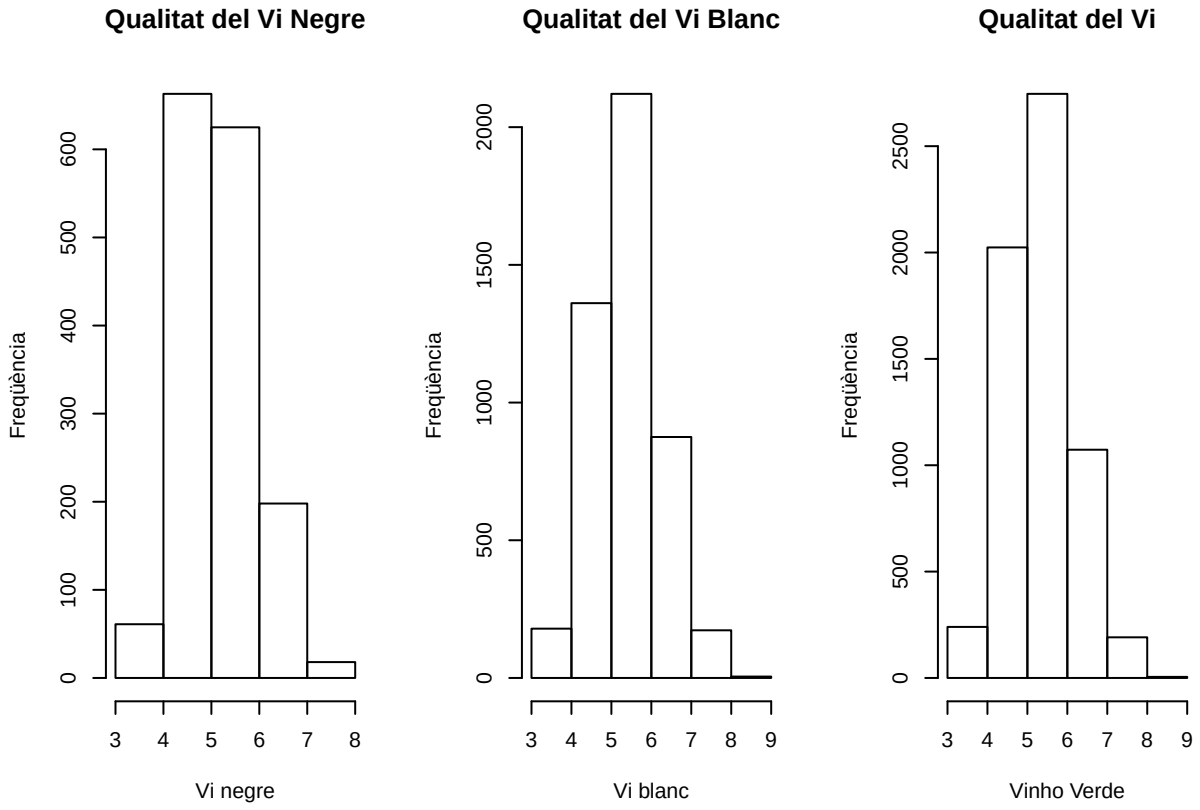
Mirem la normalitat de la 'qualitat' del vi.

Normalitat de la Qualitat

Primer observem la distribució de la qualitat segons cada tipus de vi:

```
par(mfrow=c(1,3))
# Visualització de l'histograma
hist(d[d$color=='negre',"quality"],
```

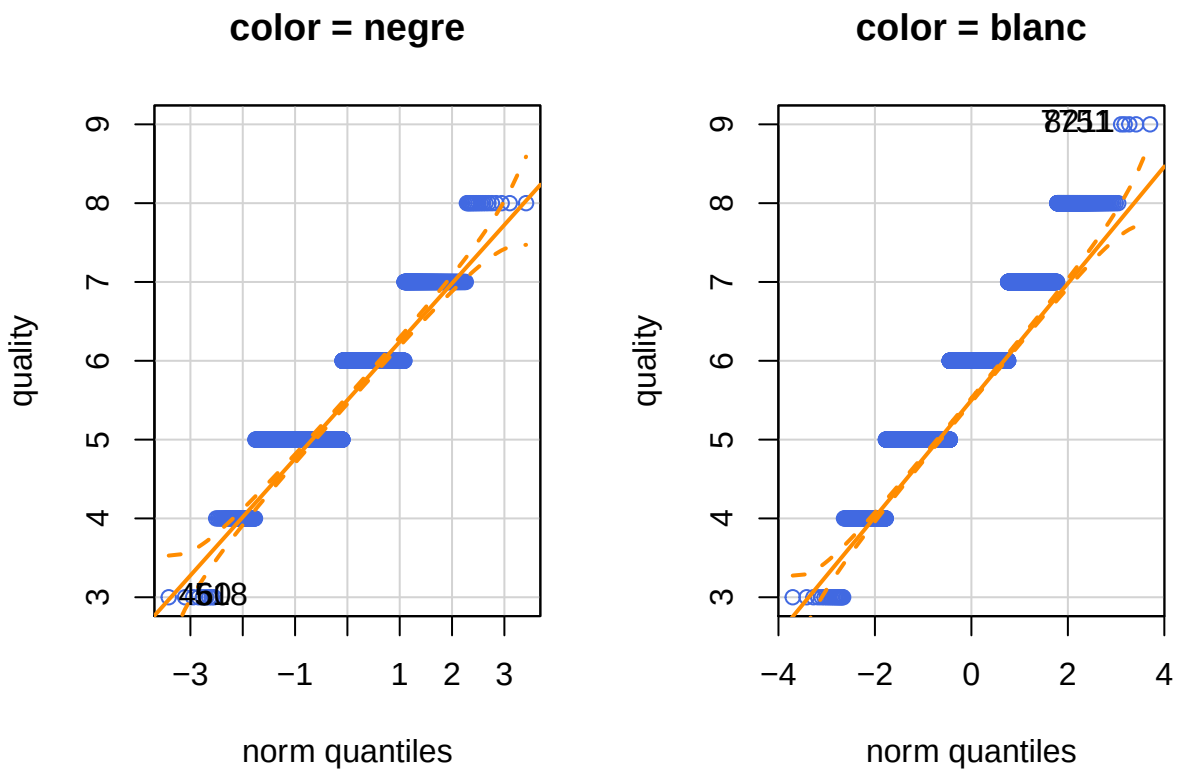
```
     breaks=length(unique(d[d$color=='negre',"quality"])),
     main="Qualitat del Vi Negre",
     xlab="Vi negre", ylab="Freqüència")
hist(d[d$color=='blanc',"quality"],
     breaks=length(unique(d[d$color=='blanc',"quality"])),
     main="Qualitat del Vi Blanc",
     xlab="Vi blanc", ylab="Freqüència")
hist(d[,"quality"],
     breaks=length(unique(d[,"quality"])),
     main="Qualitat del Vi",
     xlab="Vinho Verde", ylab="Freqüència")
```




```
par(mfrow=c(1,1))
```

Ara cerquem si la seva distribució és **normal** per cada tipus de vi:

```
# Gràfic QQ-plot de la variable 'quality' segons tipus de color ('negre','blanc').
# Utilitzem la funció qqPlot de la llibreria "car".
qqPlot(quality~color, data=d, envelope=.95, col='royalblue',col.lines='darkorange',lwd=1)
```



Si observem les dues gràfiques Q-Q plots **no indiquen normalitat de manera estricta**, doncs en les cues, ens allunyem de la recta de referència. En general, la forma de la mostra no es d'una recta. La gràfica Q-Q plot és molt intuitiva però no és definitiva, així que podem confirmar aquest resultat amb el test de normalitat de **Shapiro-Wilk**:

```
shw_Negre<-shapiro.test(d[d$color=="negre",'quality'])
shw_Blanc<-shapiro.test(d[d$color=="blanc",'quality'])
shw_Negre
```

```
##
##  Shapiro-Wilk normality test
##
## data:  d[d$color == "negre", "quality"]
## W = 0.85816, p-value < 2.2e-16
```

```
print(shw_Negre$p.value)
```

```
## [1] 2.266894e-35
```

```
shw_Blanc
```

```
##
##  Shapiro-Wilk normality test
##
## data:  d[d$color == "blanc", "quality"]
## W = 0.89089, p-value < 2.2e-16
```

```
print(shw_Blanc$p.value)
```

```
## [1] 1.275358e-49
```

En quan al resultat del test de **Shapiro-Wilk**, obtenim el valor de l'estadistic **W** i la probabilitat **p-valor**:

- per "negre": 0.85816 i p-valor = **2.266894^{-35}**
- i per "blanc": 0.89089 i p-valor = **1.275358^{-49}**

En els dos cassos **rebutgem la hipotesis H0 de normalitat**, perque el **p-value es inferior a 0.05**.

El valor de l'estadístic W, és recomenable però, que sigui proper a W=0.99 per obtenir bons resultats, i que el p-valor sigui superior a 0.05 per afirmar que la mostra pot estar originada en una població normal.

Dels resultats del test de Shapiro-Wilk **deduïm que la 'qualitat' del vi no segueix una ditribució normal per cap tipus de vi**.

Normalitat del Alcohol

Estudiem ara la normalitat de la distribució d'alcohol del vi, aplicant directament el test de Shapiro-Wilk:

```
shw1_Negre<-shapiro.test(d[d$color=="negre",'alcohol'])
shw1_Blanc<-shapiro.test(d[d$color=="blanc",'alcohol'])
shw1_Negre
```

```
##
##  Shapiro-Wilk normality test
##
## data:  d[d$color == "negre", "alcohol"]
## W = 0.93121, p-value < 2.2e-16
```

```
print(shw1_Negre$p.value)
```

```
## [1] 3.082983e-26
```

```
shw1_Blanc
```

```
##
##  Shapiro-Wilk normality test
##
## data:  d[d$color == "blanc", "alcohol"]
## W = 0.95923, p-value < 2.2e-16
```

```
print(shw1_Blanc$p.value)
```

```
## [1] 1.880557e-34
```

Per el alcohol, el resultat del test de **Shapiro-Wilk**, obtenim el valor de l'estadistic **W** i la probabilitat **p-valor**:

- per "negre": 0.93121 i p-valor = **3.082983e^{-26}**

- i per "blanc": 0.95923 i p-valor = **1.880557^{-34}**

En els dos cassos **rebutgem la hipotesis H0 de normalitat**, perque el **p-value es inferior a 0.05**.

El valor de l'estadístic W, és recomenable però, que sigui proper a W=0.99 per obtenir bons resultats, i que el p-valor sigui superior a 0.05 per afirmar que la mostra pot estar originada en una població normal.

Concloem del test de Shapiro-Wilk que **la graduació d'alcohol del vi no segueix una ditribució normal per cap tipus de vi**.

Al obtenir resultats de no normalitat, comprovar la homogeneïtat de la variancia no és necessàri.

2.4.3 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.

Com els tests de no normalitat, són afirmatius haurem de comparar dades amb proves no paramètriques. Per l'atribut alcohol disposem d'una categorització de 4 valors. Volem comprovar si els grups creats són estadísticament significatius. Per fer-ho aplicarem el **test de kruskal-Wallis** que ens permet veure si la 'qualitat' del vi és estadisticament diferent per cada grup de graduació d'alcohol definit en el cas de disposar de 3 o més grups:

```
# Aplicació del test de Kruskal-Wallis
d1$alcohol<-d$alcohol
d1$alcohol_class<-factor(d1$alcohol_class,
                         labels=c("Molt_Baixa","Moderadament_Baixa","Alta","Molt_Alta"))
kr1<-kruskal.test(d1$quality~d1$alcohol_class)
```

Amb un nivell de significancia de 0.05, arribem a la conclusió de que **la qualitat del vi per els grups de graduació considerats (Molt_Baixa, Moderadament_Baixa, Alta i Molt_Alta) és diferent**.

Finalment generem un model de regressió lineal múltiple per els atributs: **alcohol, citric.acid, sulphates, pH i color** amb variable depenent **quality**: La variable **color**, és categòrica per tant definirem un variable dummy, indicant quína serà la categoria de referència ('negre'):

```
# Definim la categoria de referència de colorR:
d2<-d
d2$colorR<-relevel(d2$color,ref='negre')
head(d2,1)
```

```
##   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1           7.4              0.7           0            1.9     0.076
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  11                   34  0.9978 3.51      0.56     9.4
##   quality color colorR
## 1       5 negre  negre
```

```
# Comprovem que ho hem definit bé:
contrasts(d2$colorR)
```

```
##       blanc
## negre     0
## blanc     1
```

Busquem explicar Y (quality) en funció de x1: alcohol, X2: citric.acid, x3: sulphates, x4: pH, x5: colorR

$$Y = \beta 0 + \beta 1\chi 1 + \beta 2\chi 2 + \beta 3\chi 3 + \beta 4\chi 4 + \beta 5\chi 5 + e$$

on:

- βj per j=0,1..5 són els coeficients desconeguts que volem estimar.
- **e** és l'error o residu que representa l'efecte de totes les variables que poden afectar a Y, però que no es contemplen.

Càlcul del model:

Executem la regressió lineal múltiple amb 5 regressors (alcohol, citric.acid, sulphates, pH, colorR) i una variables 'dummy' (color):

```
m1<-lm(quality~alcohol+citric.acid+sulphates+pH+colorR, data=d2)
summary(m1)
```

```
##
## Call:
## lm(formula = quality ~ alcohol + citric.acid + sulphates + pH +
##     colorR, data = d2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4937 -0.4931 -0.0501  0.4995  3.1311
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.557939   0.237016   6.573 5.32e-11 ***
## alcohol     0.320396   0.008275  38.717  < 2e-16 ***
## citric.acid 0.393903   0.074307   5.301 1.19e-07 ***
## sulphates   0.729550   0.079029   9.231  < 2e-16 ***
## pH          0.048158   0.068054   0.708    0.479
## colorRblanc 0.311871   0.027127  11.497  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7727 on 6273 degrees of freedom
## Multiple R-squared:  0.224,  Adjusted R-squared:  0.2234
## F-statistic: 362.1 on 5 and 6273 DF,  p-value: < 2.2e-16
```

De la sortida resum del model podem observar que el **p-value** associat al contrast és inferior a 0.05>. Per tant **com a mínim una de les variables explicatives contribueix de manera significativa a explicar la variable Y** de la qualitat del vi.

Bondat de l'ajust

De la sortida del model, obtenim **R2 ajustat = 0.2233556**. Les variables utilitzades expliquen en un **22.34%** el model, però no és un valor proper a 1 com nosaltres voldriem. Cal utilitzar el **coeficient de determinació ajustat** perque a diferència del coeficient de determinació no ajustat, no incrementa sempre quan augmentem el nombre de variables en el model.

Els coeficients estimats i els seus p-valor són:

Table 5: Taula: Paràmetres estimats i p-valor.

	Coeficients_Beta	P_Valor
(Intercept)	1.5579391	0.0000000
alcohol	0.3203958	0.0000000
citric.acid	0.3939028	0.0000001
sulphates	0.7295503	0.0000000
pH	0.0481581	0.4791920
colorRblanc	0.3118714	0.0000000

```
p_valor<-summary(m1)[["coefficients"]][, "Pr(>|t|)"]
coefs<-m1$coefficients
taula<-data.frame("Coeficients_Beta"=coefs,"P_Valor"=p_valor)
kable(taula,caption="Taula: Paràmetres estimats i p-valor.") %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed", "responsive"),
                full_width = F)
```

```
coefsr<-round(coefs,2)
```

i el model queda definit per la següent equació de regressió d'ajust:

$$\textbf{quality}_i = \textbf{1.56} + \textbf{0.32} \cdot \chi 1 + \textbf{0.39} \cdot \chi 2 + \textbf{0.73} \cdot \chi 3 + \textbf{0.05} \cdot \chi 4 + \textbf{0.31} \cdot \chi 5 + \textbf{e}$$

En un model de regressió multilineal, els coeficients de regressió associats a cadascuna de les **variables fictícies** (en anglès **dummy**) són interpretades com la diferència esperada entre la mitjana de la sortida de la propia variable en comparació amb la del seu valor de referència, quan la resta de variables són constants. I si volem comparar dues categories qualsevols d'una variable 'dummy' aleshores podem fer la diferència entre els seus coeficients estimats.

En la següent taula podem observar les equacions del model que obtenim segons els diferents valors de la variable qualitativa **colorR**, considerant la resta de variables constants:

```
# Construïm la taula resum dels models:
taula1=data.frame("colorRBlanc"=c(0,1),
                  "Significat"=c("Vi negre.","Vi blanc"),
                  "Model"=c(paste0("Y'=",coefsr[1]," + ",coefsr[2],"·alcohol",
                                   " + ",coefsr[3],"·citric.acid"," + ",
                                   coefsr[4],"·sulphates"," + ",coefsr[5],
                                   "·pH"," + ",coefsr[6],"·1")))

print("Tenint en compte la equació anterior:")
```

```
## [1] "Tenint en compte la equació anterior:"
```

```
kable(taula1,
      caption="Taula: Equacions del model segons valors de les variables dummy.") %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed", "responsive"),
                full_width = F)
```

Table 6: Taula: Equacions del model segons valors de les variables dummy.

colorRBlanc	Significat	Model
0	Vi negre.	Y'=1.56 + 0.32 · alcohol + 0.39 · citric.acid + 0.73 · sulphates + 0.05 · pH + 0.31 · 1
1	Vi blanc	Y'=1.56 + 0.32 · alcohol + 0.39 · citric.acid + 0.73 · sulphates + 0.05 · pH + 0.31 · 1

Interpretació dels paràmetres

- **(Intercept)** *B0 = 1.56* Valor que pren la variable Y (quality) quan tots els regressors qualitatius són zero. La interpretació no la fem per no tenir sentit composició del vi igual a 0.
- **Alcohol** *B1 = 0.32* El coeficient estimat és positiu. La variable depenent 'quality' CREIX una proporció de 0.32 quan aumenta una unitat (1 grau) la graduació d'alcohol, i la resta de variables es mantenen constants. El p-valor de la variable 'alcohol' és inferior a 0.05 per tant rebutgem la hipotesi de que la variable no contribueix al model.
- **citric.acid** *B2 = 0.39* El coeficient estimat és positiu. La variable depenent 'quality' s'INCREMENTA una proporció de 0.39 quan aumenta una unitat de citric.acid, i la resta de variables es mantenen constants. El p-valor de la variable 'citric.acid' és inferior a 0.05 per tant rebutgem la hipotesi de que la variable no contribueix al model.
- **sulphates** *B3 = 0.73* El coeficient estimat és positiu. La variable depenent 'quality' CREIX una proporció de 0.73 quan aumenta una unitat el component de sulphates, i la resta de variables es mantenen constants. El p-valor de la variable 'sulphates' és inferior a 0.05 per tant rebutgem la hipotesi de que la variable no contribueix al model.
- **pH** *B4 = 0.05* El coeficient estimat és positiu. La variable depenent 'quality' CREIX una proporció de 0.05 quan aumenta una unitat el component de pH, i la resta de variables es mantenen constants. El p-valor de la variable 'pH' és inferior a 0.05 per tant rebutgem la hipotesi de que la variable no contribueix al model.
- **colorRblanc** *B5 = 0.31* El coeficient estimat és positiu. Al ser positiu representa un increment de qualitat 0.31 (31%) del vi blanc respecte el vi negre (que és la categoria de referència), quan la resta de variables resta constant. El seu p-valor es major a 0.05 per tant el coeficient no és significatiu i podríem eliminar la variable del model.

Podem comprovar que si elimienm la variable **pH** del model, obtenim resultats semblants:

```
# regressió lineal multiple sense el pH:
m2<-lm(quality~alcohol+citric.acid+sulphates+colorR, data=d2)
summary(m2)
```

```
##
## Call:
## lm(formula = quality ~ alcohol + citric.acid + sulphates + colorR,
##     data = d2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4876 -0.4932 -0.0493  0.4971  3.1337
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.70972    0.10084  16.956   <2e-16 ***
## alcohol      0.32113    0.00821  39.116   <2e-16 ***
## citric.acid  0.37826    0.07094   5.332    1e-07 ***
## sulphates    0.73627    0.07845   9.385   <2e-16 ***
## colorRblanc  0.30791    0.02654  11.600   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7727 on 6274 degrees of freedom
## Multiple R-squared:  0.2239, Adjusted R-squared:  0.2234
## F-statistic: 452.5 on 4 and 6274 DF,  p-value: < 2.2e-16
```

Finalment també comprovem el model resultant si escollim els components del vi més correlacionats amb la qualitat del vi, però observem que els resultats **no milloren** molt:

```
# regressió lineal amb els components més correlacionats amb la qualitat del vi:
m3<-lm(quality~alcohol+volatile.acidity+sulphates+colorR, data=d2)
summary(m3)
```

```
##
## Call:
## lm(formula = quality ~ alcohol + volatile.acidity + sulphates +
##     colorR, data = d2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3026 -0.4719 -0.0413  0.4874  3.1613
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.784860   0.109016  25.545  < 2e-16 ***
## alcohol           0.318462   0.007965  39.983  < 2e-16 ***
## volatile.acidity -1.577037   0.076694 -20.563  < 2e-16 ***
## sulphates         0.559814   0.076070   7.359 2.09e-13 ***
## colorRblanc      -0.089928   0.032644  -2.755  0.00589 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7496 on 6274 degrees of freedom
## Multiple R-squared:  0.2696, Adjusted R-squared:  0.2692
## F-statistic:   579 on 4 and 6274 DF,  p-value: < 2.2e-16
```

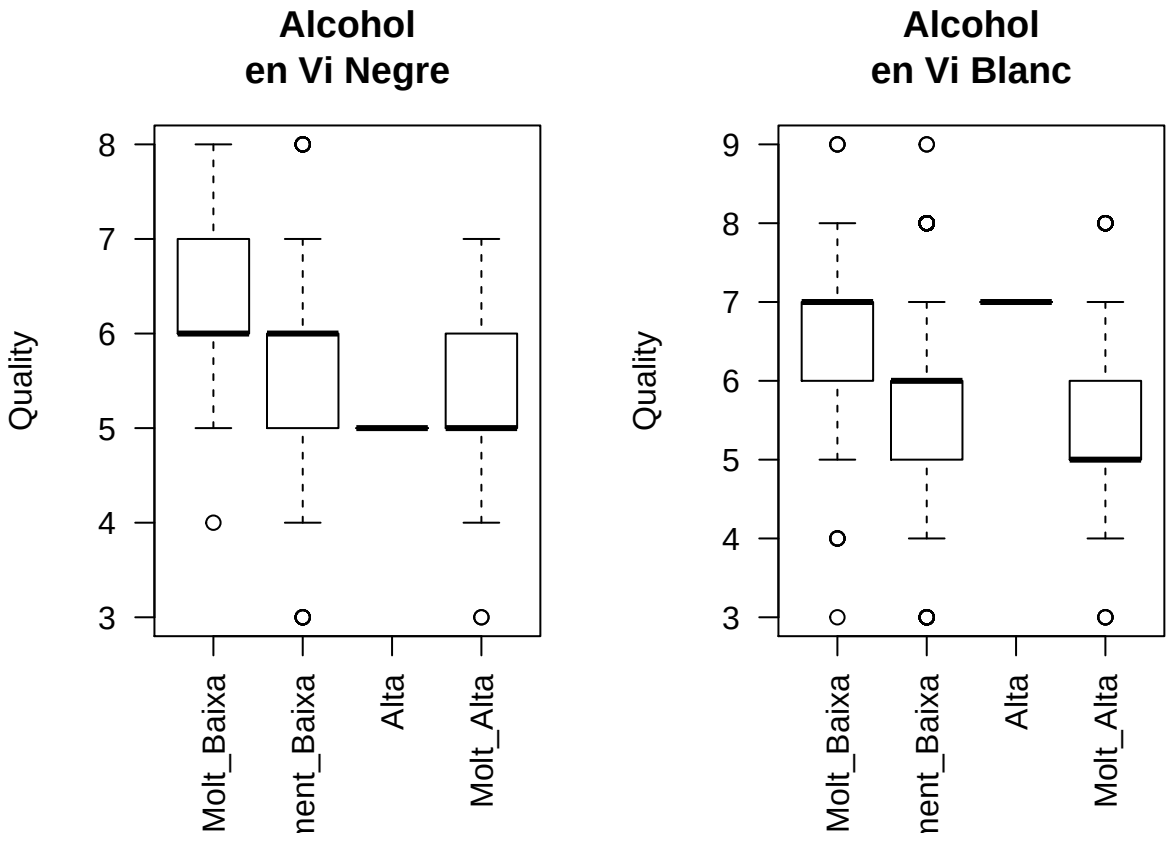
2.5 Representació dels resultats a partir de taules i gràfiques.

Al llarg de la pràctica hem anat combinant les operacions i anàlisis realitzats amb visualitzacions gràfiques dels resultats (matrius de correlació, gràfics Q-Q Plot, histogrames, taules en els models de regressió).

En aquest cas visualitzarem el comportament de la categorització de la graduació d'alcohol, que com a resultat dels nostres anàlisis, pot estar categoritzada i ens pot permetre determinar la qualitat del vi.

En l'apartat d'anàlisi hem obtingut que els grups definits eren estadísticament diferents a partir de les dades disponibles. Utilitzant un gràfic de boxplot, podem veure que:

```
# Gràfics boxplot de la qualitat del vi per cada tipus de vi respecte cada categoria definida en la gra
par(mfrow=c(1,2))
boxplot(d1[d1$color=='negre','quality']~d1[d1$color=='negre','alcohol_class'],
        ylab='Quality', main='Alcohol\nen Vi Negre', las=2)
boxplot(d1[d1$color=='blanc','quality']~d1[d1$color=='blanc','alcohol_class'],
        ylab='Quality', main='Alcohol\nen Vi Blanc', las=2)
```




```
par(mfrow=c(1,1))
```

Apliquem un gràfic per cada tipus de vi, ja que el color en si no ens determina la qualitat del vi.

Per el vi negre:

- Quan la graduació és inferior a 9.5º: la qualitat com a minim és de 5, el valor més freqüent és 6, i és la categoria on trobem les qualitats més altes fins 8. Ens trobem amb vins de qualitat bona.

- Quan la graducio es troba entre 9.5º i és inferior a 12.5º: la qualitat no superarà el 7, i la pitjor no és inferior a 4, per tant ens trobem majoritàriament en vins de qualitat acceptable-bona
- Si la graduació es troba entre 12.5º i 14.5º: la qualitat és 5. Ens trobem amb vins de qualitat correcta però no bona.
- Si la graduació es major que 14.5º: Podem trobar algun vi de qualitat notable, però majoritariament la qualitat és acceptable.

Per el vi blanc:

- Quan la graduació és inferior a 9.5º: la qualitat com a minim és de 5 i el valor més freqüent és 5 ó 6, i és la categoria on trobem les qualitats més altes fins 9. Ens trobem amb vins de qualitat bona.
- Quan la graducio es troba entre 9.5º i és inferior a 12.5º: el vin blanc es comporta com el negre.
- Si la graduació es troba entre 12.5º i 14.5º: A diferència del negre, és sempre de qualitat notable.
- Si la graduació es major que 14.5º: Podem trobar vins de molta qualitat però majoritariament la qualitat és acceptable.

2.6. Resolució del problema i conclusions.

> A partir dels resultats obtinguts, quines són les conclusions? Els resultats permeten respondre al problema?

En les conclusions finals, podem resumir que tenim la seguretat de que **la qualitat del vi no depen del seu color**, i que ens **hem de centrar en la seva composició per fer pronòstics de qualitat** per cada tipus de vi per separat.

També podem concloure que **la graduació d'alcohol és el element que més ens pot determinar més la qualitat del vi** d'entre els que disposem però **tampoc amb un grau de correlació molt alt (0.44)**. Segons aquests resultats hem definit varies categories de graduació. A més a més hem analitzat si les podiem considerar diferents amb resultats satisfactoris.

Recordem que per les categories fixades: Molt_Baixa, Moderadament_Baixa, Alta, Molt_Alta hem aplicat el Test de Kruskal-Wallis que ens ha determinat amb un nivell de significació de 0,05 que **la qualitat del vi per els grups de graduació considerats és diferent**.

També hem pogut determinar la qualitat del vi segons la seva graduació, sempre tenint en compte el tipus de vi (negre o blanc).

Els resultats obtinguts si que ens permeten respondre al problema però no amb resultats gaire satisfactoris. La conclusió és que **cal revisar el dataset i potser equilibrar-lo millor** en quan a la quantitat de mostres de cada qualitat de vi per mirar de millorar els resultats del vi. Vegem que hi ha una sobre representació de vins de qualitat mitja i molts pocs vins de molt alta qualitat o qualitat baixa, i sabem que no respon a una distribució donat els resultats dels QQPlots i els Test de Shapiro-Wilk, per lo que pensem que per ventura la mostra estigui esbiaixada cap a vins de qualitat mitjana.

3 Recursos

Calvo M, Pérez D, Subirats L. 2019. *Introducción a La Limpieza Y Análisis de Los Datos*. Editorial UOC.

Dalgaard, Peter. 2008. *Introductory Statistics with R*. Springer Science & Business Media.

Jiawei Han, Jian Pei, Micheine Kamber. 2012. *Data Mining: Concepts and Techniques*. Morgan Kaufmann.

Squire, Megan. 2015. *Clean Data*. Packt Publishing Ltd.