

Tipología i cicle de vida de les dades - Pràctica 2: neteja i validació de les dades

Jordi Boldú Millà

5 de enero, 2019

Contents

1. Descripció del dataset	1
2. Integració i selecció de les dades d'interès a analitzar.	3
3. Neteja de les dades.	3
Estadístics bàsics del data set de treball	3
Tractament dels valors buits, nuls i zeros	4
Identificació i tractament de valors extrems	6
4. Anàlisi de les dades.	9
Selecció dels grups de dades que es volen analitzar/comparar. Planificació dels anàlisis a aplicar	9
Comprovació de la normalitat i homogeneïtat de la variància	9
Aplicació de proves estadístiques per comparar els grups de dades.	10
5. Representació dels resultats a partir de taules i gràfiques.	19
Boxplot de cada atribut	19
Histogrames de cada atribut	20
Alcohol vs Qualitat	21
Volatile acidity vs Qualitat	22
Normalitat : gràfics Q-Q (normalitat) de cada atribut	23
Variància : gràfic comparant la variància segons la classe de vi	24
PCA : gràfic que mostra la relació entre els diferents components	25
PCA : gràfic que mostra el % (parcial i acumulat) de la variància segons component	26
6. Resolució del problema.	27
A partir dels resultats obtinguts, quines són les conclusions? Els resultats permeten respondre al problema?	27
Bibliografia i referències	28

1. Descripció del dataset

Font de les dades

El conjunt de dades a analitzar

Red Wine Quality

Simple and clean practice dataset for regression or classification modelling

ha estat descarregat del lloc web *Kaggle*

<https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009>

La mostra es composa de 1599 files i 12 columnes

Cliqueu [aquí](#) per anar a la pàgina d'origen

Descripció

Aquest data set recull informació de diferents variants de vi negre del vi portuguès “Vinho Verde”. Per temes de privacitat, només hi ha disponibles variables fisicoquímiques (entrades) i sensorials (la sortida). No hi ha dades sobre altres tipus d'informació com els tipus de raïm, la marca del vi, el preu de venda del vi, etc.

Llistat d'atributs del data set

- **fixed acidity**: concentració d'àcid tartàric dins del vi. Es mesura en g/dm3
- **volatile acidity**: concentració d'àcid acètic dins del vi. En altes concentracions provoca gust de vinagre. Es mesura en g/dm3
- **citric acid**: concentració d'àcid cítric. En petites quantitats dona sensació de frescor i gust al vi. Es mesura en g/dm3
- **residual sugar**: concentració de sucre remanent després dels processos de fermentació. És estrany trobar vins amb concentracions inferiors a 1 g/l. Els vins amb concentracions de 45 g/l o superiors són considerats dolços. Es mesura en g/dm3
- **chlorides**: concentració de sal. Es mesura en g/dm3
- **free sulfur dioxide**: concentració de diòxid de sofre en forma de gas. Ajuda a prevenir el creixement de microbis i la oxidació del vi. Es mesura en mg/dm3
- **total sulfur dioxide**: concentració total de diòxid de sofre. En baixes concentracions és gairebé indetectable però en concentracions superiors a 50 ppm, té efectes sobre l'olfacte i el gust del vi. Es mesura en mg/dm3
- **density**: densitat de l'aigua en funció del percentatge d'alcohol i el contingut de sucre. Es mesura en g/cm3
- **pH**: indicador que descriu l'acidesa o basicitat d'un vi en una escala de 0 (molt àcid) a 14 (molt bàsic). La majoria dels vins tenen un pH entre 3 i 4.
- **sulphates**: additiu del vi (sulfat de potassi) que pot contribuir als nivells de diòxid de sofre (recordem que aquest actua com agent anti microbial i anti oxidant). Es mesura en g/dm3
- **alcohol**: percentatge d'alcohol que conté el vi. Es mesura en % en volum
- **quality**: indicador de qualitat del vi en una escala de 1 (pitjor qualitat) a 10 (millor qualitat)

Mostra de les 5 primeres files del data set

```
head(dfInputOriginal)
```

```
##   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            0.9978 3.51     0.56     9.4
## 2                 25            0.9968 3.20     0.68     9.8
## 3                 15            0.9970 3.26     0.65     9.8
## 4                 17            0.9980 3.16     0.58     9.8
## 5                 11            0.9978 3.51     0.56     9.4
```

```

## 6          13        40  0.9978 3.51      0.56     9.4
##   quality
## 1      5
## 2      5
## 3      5
## 4      6
## 5      5
## 6      5

```

Perquè és important i quina pregunta/problema pretén respondre?

Per mitjà de l'anàlisi i exploració d'aquest conjunt de dades intentarem esbrinar si es pot establir alguna relació entre les propietats fisicoquímiques dels vins analitzats (dades objectives) i la seva qualitat (valoració subjectiva) i determinar-ne la importància de cadascuna d'elles.

Les conclusions de l'anàlisi, més enllà del propòsit didàctic d'aquesta pràctica, podrien servir per ajudar a determinar quines característiques cal potenciar i quins processos de producció es poden optimitzar per a l'elaboració de nous vins, ja siguin de bona qualitat o no. Entenem que aquesta decisió dependria d'altres factors o criteris de decisió empresarial (quin és el target objectiu, percentatge d'ingressos que proporciona cada tipus de vi, etc.) i que s'escapen de l'objectiu d'aquesta pràctica.

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

Les dades que farem servir en el següent estudi provenen totes del mateix conjunt donat, és a dir, no hem hagut de realitzar cap integració amb dades externes.

Després de llegir amb deteniment la descripció de cadascún dels atributs del data set sembla que n'hi ha alguns que probablement estiguin relacionats (els relatius a la **acidesa** i els que tenen a veure amb el **diòxid de sofre**). En cas d'existir, aquesta relació s'haurà de veure reflectida per exemple, per l'existència d'un **coeficient de correlació** amb un cert grau de significació, ja sigui positiu o negatiu.

Malgrat aquesta observació, a priori, no podem descartar cap dels atributs del conjunt inicial de dades doncs encara no disposem de cap indici que permeti fonamentar l'eliminació de cap d'ells.

3. Neteja de les dades.

Estadístics bàsics del data set de treball

Abans de tractar el punt referent a la neteja de les dades, a títol introductori, adjuntem un breu resum descriptiu dels estadístics bàsics del nostre data set.

```

# noms original de les variables
names(dfInput)

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

# Estadístics bàsics
summary(dfInput)

## 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 : 7.40  Median : 0.4280  Median : 0.540  Median : 1.000
## Mean   : 7.78  Mean   : 0.5237  Mean   : 0.671  Mean   : 1.078
## 3rd Qu.: 10.00 3rd Qu.: 0.5880 3rd Qu.: 0.760 3rd Qu.: 1.725
## Max.  : 16.00  Max.  : 1.0000  Max.  : 2.760  Max.  : 3.995

```

```

## 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

# Printem el nom i el tipus de variable
nom_i_tipus <- sapply(dfInput,class)
data.frame(Variables=names(nom_i_tipus),Classe=as.vector(nom_i_tipus))

```

```

##          Variables Classe
## 1      fixed.acidity numeric
## 2      volatile.acidity numeric
## 3      citric.acid numeric
## 4      residual.sugar numeric
## 5      chlorides numeric
## 6      free.sulfur.dioxide numeric
## 7      total.sulfur.dioxide numeric
## 8      density numeric
## 9      pH numeric
## 10     sulphates numeric
## 11     alcohol numeric
## 12     quality integer

```

NOTA : A l'apartat 5 d'aquest document s'adjunten gràfiques relatives a la distribució dels valors dels diferents atributs en format boxplot i histograma

Tractament dels valors buits, nuls i zeros

Tal i com s'indica en la descripció detallada que accompanya al data set original, les dades **no contenen elements buïts o nuls (NA)**

```
sapply(dfInput, function(x) sum(is.na(x)))
```

```
##      fixed.acidity      volatile.acidity      citric.acid
```

```

##          0          0          0
## residual.sugar chlorides free.sulfur.dioxide
##          0          0          0
## total.sulfur.dioxide density pH
##          0          0          0
## sulphates alcohol quality
##          0          0          0

```

En canvi, veiem que hi ha files on l'atribut *citric.acid* **conté zeros**

```
lapply(dfInput, function(x) sum(x==0))
```

```

## $fixed.acidity
## [1] 0
##
## $volatile.acidity
## [1] 0
##
## $citric.acid
## [1] 132
##
## $residual.sugar
## [1] 0
##
## $chlorides
## [1] 0
##
## $free.sulfur.dioxide
## [1] 0
##
## $total.sulfur.dioxide
## [1] 0
##
## $density
## [1] 0
##
## $pH
## [1] 0
##
## $sulphates
## [1] 0
##
## $alcohol
## [1] 0
##
## $quality
## [1] 0

```

Si ens fixem en la distribució de valors que pot prendre aquest atribut juntament amb la gran quantitat de zeros, podem conoure que **no es tracta de cap error** sinó que són valors perfectament vàlids dins la mostra i, per tant, **no els descartem**

Identificació i tractament de valors extrems

Segons Jason W. Osborne (*Data Cleaning Basics: Best Practices in Dealing with Extreme Scores. Newborn and Infant Nursing Reviews*) un valor extrem es pot descriure com una observació que es desvia tant d'altres observacions com per despertar sospites que va ser generat per un mecanisme diferent.

Com a punt de partida, podem considerar que aquelles observacions allunyades 3 o més desviacions estàndard de la mitjana de la mostra són susceptibles de ser valors extrems, tret que la mostra sigui particularment petita.

Aquesta valors extrems són un cas particular dels valors què, convencionalment i de manera generalitzada, es coneixen com a **outliers** i que per definició són aquells valors que estan situats a:

- una distància superior a 1,5 vegades el rang interquartílic per sobre el 3er quartil
- una distància inferior a 1,5 vegades el rang interquartílic per sota del 1er quartil

Farem servir aquesta darrera definició (configuració per defecte del paràmetre *coef* de la funció *boxplot.stats*) per al càcul dels valors aïpics

```
c <- names(dfInput)
total_outliers <- 0
for (i in 1:ncol(dfInput))
{
  cat("  \n")
  a <- boxplot.stats(dfInput[, i],)$out
  cat("**Atribut '", c[i], "**  \n", sep = '')
  cat("*Num. outliers : ", length(a), "*  \n", sep = '')
  cat(sort(a), "  \n")
  total_outliers <- total_outliers + length(a)
}
```

Atribut ‘fixed.acidity’

Num. outliers : 49

12.4 12.4 12.4 12.5 12.5 12.5 12.5 12.5 12.5 12.5 12.6 12.6 12.6 12.6 12.7 12.7 12.7 12.7 12.7 12.8 12.8 12.8 12.8 12.9 12.9 13 13 13 13.2 13.2 13.2 13.3 13.3 13.3 13.4 13.5 13.7 13.7 13.7 13.8 14 14.3 15 15 15.5 15.5 15.6 15.6 15.9

Atribut ‘volatile.acidity’

Num. outliers : 19

1.02 1.02 1.02 1.02 1.025 1.035 1.04 1.04 1.04 1.07 1.09 1.115 1.13 1.18 1.185 1.24 1.33 1.33 1.58

Atribut ‘citric.acid’

Num. outliers : 1

1

Atribut ‘residual.sugar’

Num. outliers : 155

3.7 3.7 3.7 3.7 3.75 3.8 3.8 3.8 3.8 3.8 3.8 3.8 3.9 3.9 3.9 3.9 3.9 3.9 3.9 3.9 3.9 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4.1 4.1 4.1 4.1 4.2 4.2 4.2 4.2 4.25 4.3 4.3 4.3 4.3 4.3 4.3 4.3 4.3 4.3 4.4 4.4 4.4 4.4 4.4 4.4 4.5 4.5 4.5 4.5 4.5 4.6 4.6 4.6 4.6 4.6 4.6 4.6 4.6 4.6 4.65 4.65 4.7 4.7 4.8 4.8 4.8 5 5.1 5.1 5.1 5.1 5.1 5.15 5.2 5.2 5.2 5.4 5.5 5.5 5.5 5.5 5.5 5.5 5.5 5.5 5.5 5.5 5.6 5.6 5.6 5.6 5.6 5.6 5.6 5.7 5.7 5.8 5.8 5.8 5.9 5.9 5.9 6.7 6.7 7 7.2 7.3 7.5 7.5 7.8 7.8 7.9 7.9 7.9 8.1 8.1 8.3 8.3 8.3 8.6 8.8 8.8 8.9 9 10.7 11 11 12.9 13.4 13.8 13.8 13.9 15.4 15.4 15.5

Atribut ‘chlorides’

Num. outliers : 112

0.012 0.012 0.034 0.038 0.038 0.039 0.039 0.039 0.039 0.12 0.12 0.12 0.121 0.121 0.122 0.122 0.122 0.122 0.122 0.122 0.122 0.123 0.123 0.123 0.123 0.123 0.123 0.124 0.124 0.124 0.125 0.125 0.126 0.127 0.128 0.132 0.132 0.132 0.136 0.137 0.143 0.145 0.146 0.147 0.148 0.152 0.152 0.153 0.157 0.157 0.157 0.159 0.161 0.165 0.166

0.166 0.166 0.168 0.169 0.17 0.171 0.171 0.172 0.174 0.176 0.178 0.186 0.19 0.194 0.2 0.205 0.205 0.213
0.214 0.214 0.214 0.216 0.222 0.226 0.226 0.23 0.235 0.236 0.241 0.243 0.25 0.263 0.267 0.27 0.332 0.337 0.341
0.343 0.358 0.36 0.368 0.369 0.387 0.401 0.403 0.413 0.414 0.414 0.415 0.415 0.415 0.422 0.464 0.467 0.61 0.611

Atribut ‘free.sulfur.dioxide’

Num. outliers : 30

43 43 43 45 45 45 46 47 48 48 48 48 50 50 51 51 51 51 52 52 52 53 54 55 55 57 66 68 68 72

Atribut ‘total.sulfur.dioxide’

Num. outliers : 55

124 124 124 125 125 126 127 127 128 128 129 129 129 130 131 131 131 131 133 133 133 134 134 135 135 136 136
139 140 141 141 141 142 143 143 144 144 144 145 145 145 147 147 147 148 148 149 151 151 152 153 155 160
165 278 289

Atribut ‘density’

Num. outliers : 45

0.99007 0.99007 0.9902 0.99064 0.99064 0.9908 0.99084 0.9912 0.9915 0.99154 0.99157 0.9916 0.9916 0.99162
0.9917 0.99182 0.99182 0.99191 0.9921 0.9922 0.9922 1.0014 1.0014 1.0014 1.0014 1.0014 1.0014 1.0015 1.0015
1.0018 1.0021 1.0021 1.0022 1.0022 1.00242 1.00242 1.0026 1.0026 1.00289 1.00315 1.00315 1.00315 1.0032
1.00369 1.00369

Atribut ‘pH’

Num. outliers : 35

2.74 2.86 2.87 2.88 2.88 2.89 2.89 2.89 2.89 2.9 2.92 2.92 2.92 2.92 3.69 3.69 3.69 3.69 3.69 3.7 3.71 3.71 3.71
3.72 3.72 3.72 3.74 3.75 3.78 3.78 3.85 3.9 3.9 4.01 4.01

Atribut ‘sulphates’

Num. outliers : 59

1 1.01 1.02 1.02 1.02 1.03 1.03 1.04 1.04 1.05 1.05 1.05 1.06 1.06 1.06 1.06 1.06 1.07 1.07 1.07 1.08 1.08 1.08 1.09 1.1 1.1
1.11 1.12 1.13 1.13 1.14 1.14 1.15 1.16 1.17 1.17 1.17 1.17 1.17 1.17 1.18 1.18 1.18 1.2 1.22 1.26 1.28 1.28 1.31 1.33
1.34 1.36 1.36 1.36 1.56 1.59 1.61 1.62 1.95 1.95 1.98 2

Atribut ‘alcohol’

Num. outliers : 13

13.56667 13.6 13.6 13.6 13.6 14 14 14 14 14 14 14 14 14.9

Atribut ‘quality’

Num. outliers : 28

3 3 3 3 3 3 3 3 8

Total outliers : 601

A priori apareixen 601 valors susceptibles de ser considerats ‘atípics’ i distribuïts de forma heterogènia entre els diferents atributs però, donat que no tenim prou coneixement sobre el món dels vins per saber si es tracta d’errors de medició o són valors possibles, optarem per a deixar-los, sense fer-ne cap tractament, tot i que ja podem anticipar que la seva dispersió pot impactar en l’anàlisi de les dades

Una altra tècnica que sovint s’utilitza per a trobar valors atípics és utilitzant la **distància de Mahalanobis** que per mitja del concepte de similitud, permet identificar aquelles observacions que *més s’allunyen* (o són menys similars) respecte la resta de valors.

A continuació, a tall d’exemple pràctic, realitzarem les següents operacions

- Calcularem la **distància de Mahalanobis** per a cadascuna de les files del data set
- Calcularem els estadístics bàsics de la **distància de Mahalanobis** i els **outliers**
- Visualitzarem gràficament quins elements del conjunts de dades són *menys similars* (aqueells que estiguin dispersos i més allunyats)

Distancia de mahalanobis

```

# Settings del gràfic
options(repr.plot.width=6, repr.plot.height=4)

# Càlcul de la distància de Mahalanobis
md <- mahalanobis(dfInput[, c(1:11)], colMeans(dfInput[, c(1:11)]), cov(dfInput[, c(1:11)]))

# Estadístics bàsics de la distància de Mahalanobis
summary(md)

##      Min. 1st Qu. Median    Mean 3rd Qu.    Max.
## 1.092   5.120   7.738 10.993 12.172 155.546

# Rang interquartílic
iqr <- IQR(md)
iqr

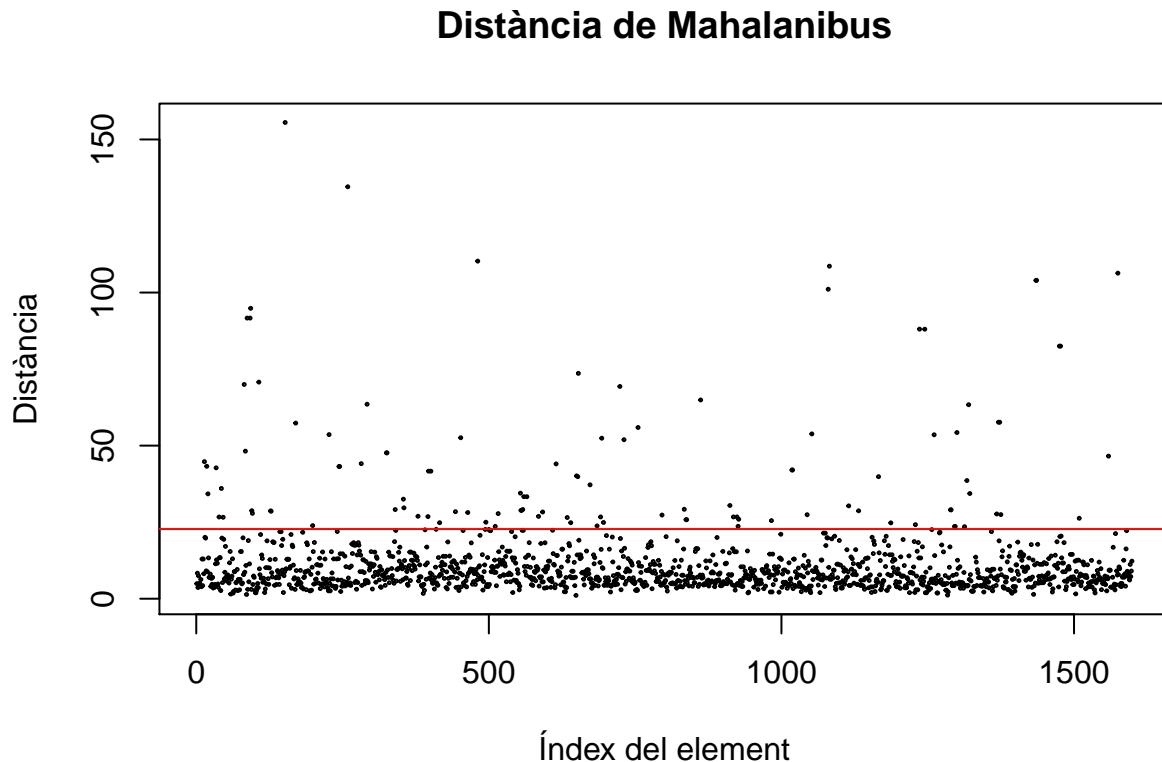
## [1] 7.052017

# Llindar
max_threshold <- IQR(md)*(1.5) + summary(md)[[5]] # 3er quartil
max_threshold

## [1] 22.74971

# Representació gràfica
plot(md, pch=1, cex=.2, main="Distància de Mahalanibus",
      xlab = "Índex del element", ylab="Distància")
abline(h = max_threshold, col="red") # add cutoff line

```



```

# Nombre d'outliers
1 <- length(boxplot.stats(md)$out)
1

## [1] 110
dfInput$md <- NULL

```

4. Anàlisi de les dades.

Selecció dels grups de dades que es volen analitzar/comparar. Planificació dels anàlisis a aplicar

En principi treballarem amb tots els atributs del conjunt de dades orginals i, per a l'estudi de la **homogeneïtat de la variància** (apartat següent), **discretitzarem** la variable *quality* en una nova variable *classe*, qualitativa, que contindrà 3 valors possibles (*Dolent, Normal o Bo*) relatives a la valoració del vi

Comprovació de la normalitat i homogeneïtat de la variància

NOTA : A l'apartat 5 d'aquest document s'adjunten gràfiques relatives normalitat i la variància

Estudi de la normalitat

Per saber si cadascun dels atributs segueix una distribució normal plantejarem el següent contrast d'hipòtesi per a cadascun dels atributs amb un nivell de significació α del 0.05

- **H0** : La mostra de tamany n segueix una distribució Normal
- **H1** : La mostra de tamany n NO segueix una distribució Normal

Aquest contrast el durem a terme amb el **test de normalitat de Shapiro-Wilk** per a *cadascun dels atributs*. Si el *p-valor* obtingut per a cada atribut és menor al nivell de significació α (< 0.05), rebutjarem la hipòtesi nul.la (H0) i afirmarem que la mostra NO segueix una distribució normal

```

# Normalitat
for (i in 1:ncol(dfInput))
{
  cat("Atribut '",c[i],"'", sep = '')
  pvalor <- shapiro.test(dfInput[, i])[[["p.value"]]]
  cat("p-valor '", pvalor,"'\n", sep = '')
}

## Atribut 'fixed.acidity', p-valor '1.525012e-24'
## Atribut 'volatile.acidity', p-valor '2.692935e-16'
## Atribut 'citric.acid', p-valor '1.021932e-21'
## Atribut 'residual.sugar', p-valor '1.020162e-52'
## Atribut 'chlorides', p-valor '1.179056e-55'
## Atribut 'free.sulfur.dioxide', p-valor '7.694597e-31'
## Atribut 'total.sulfur.dioxide', p-valor '3.573451e-34'
## Atribut 'density', p-valor '1.936053e-08'
## Atribut 'pH', p-valor '1.712237e-06'
## Atribut 'sulphates', p-valor '5.82314e-38'
## Atribut 'alcohol', p-valor '6.644057e-27'
## Atribut 'quality', p-valor '9.515085e-36'

```

Podem observar que els valors de cadascun dels atribut del data set **NO segueixen una distribució Normal**

Estudi de la homogeneïtat de la variància

Estudiarem la homogeneïtat de la variància aplicant també un contrast d'hipòtesi amb un nivell de significació α del 0.05.

Per a fer-ho, tal i com s'indicava en l'enunciat **hem discretitzat** la variable *quality* en una nova variable qualitativa *classe*, qualitativa, que contindrà 3 valors possibles (*Dolent*, *Normal* o *Bo*) relatives a la valoració del vi.

Estudiarem la homogeneïtat comparant les variàncies de les mostres de vins agrupats per aquest nou atribut *classe*

- **H0** : Les variàncies poblacionals són iguals (Homoscedasticitat)
- **H1** : Les variàncies poblacionals són diferents (Heteroscedasticitat)

Aquest contrast el durem a terme per mitjà del **test de Levene**.

Si el *p*-valor obtingut és menor que el nivell de significació α (< 0.05) rebutjarem la hipòtesi nul.la (H0) i afirmarem que les variàncies poblacionals són diferents (Heteroscedasticitat). En cas contrari, no podrem rebutjar la hipòtesi nul.la H0

```
# Homogeneitat de les variances
dfInput2 <- dfInput
dfInput2$classe <- 0
idx <- which(dfInput2$quality < 5)
dfInput2$classe[idx] <- "Dolent"
idx <- which(dfInput2$quality >= 5 & dfInput2$quality <= 6)
dfInput2$classe[idx] <- "Normal"
idx <- which(dfInput2$quality >= 7)
dfInput2$classe[idx] <- "Bo"
dfInput2$quality <- NULL
dfInput2$classe <- as.factor(dfInput2$classe)
table(dfInput2$classe)
```

```
##
##      Bo Dolent Normal
##    217     63   1319
# Apliquem el test de Levene
leveneTest(dfInput2$alcohol,dfInput2$classe)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value Pr(>F)
## group     2  0.7375 0.4785
##           1596
```

Com que el p-valor (0.4785) és més gran que el nivell de significació α (0.05), no podrem rebutjar la hipòtesi nul.la H0 que les variàncies poblacionals són iguals i per tant, podem afirmar que **hi ha homoscedasticitat**

Aplicació de proves estadístiques per comparar els grups de dades.

Tal i com s'ha comentat, volem analitzar i explorar el conjunt de dades per intentar esbrinar si es pot establir alguna relació entre les propietats fisicoquímiques dels vins analitzats (dades objectives) i la seva qualitat (valoració subjectiva) i determinar-ne la importància de cadascuna d'elles.

Per a dur-ho a terme, realitzarem 3 proves estadístiques

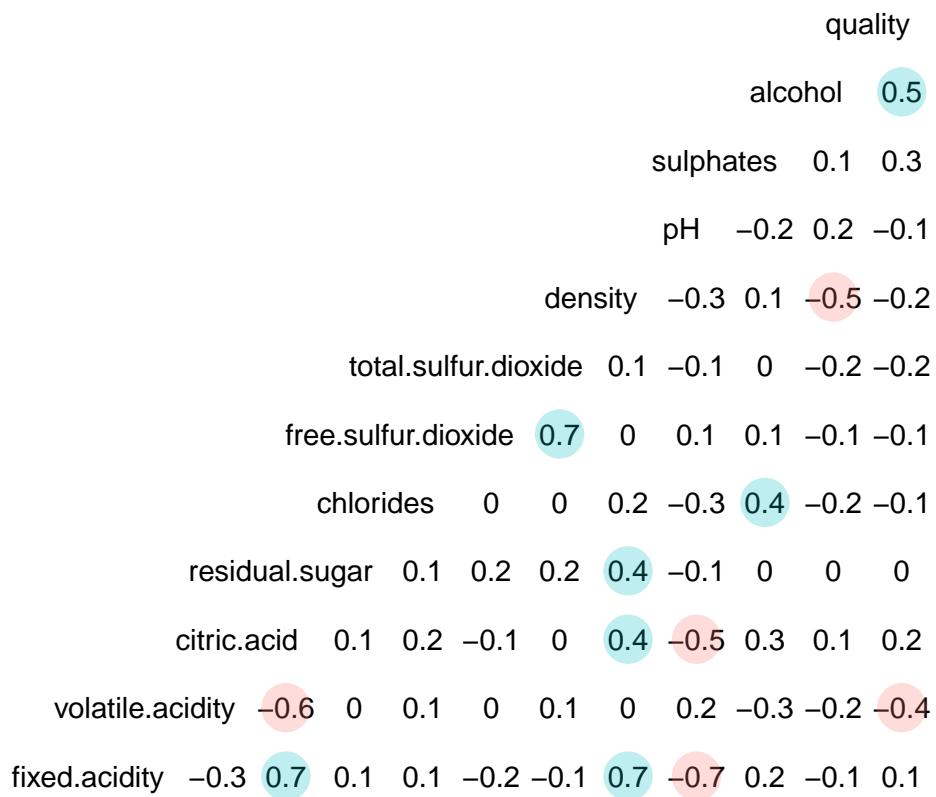
- Buscarem la matriu de correlació entre els diferents atributs del data set
- Mirarem de reduir la cardinalitat del conjunt de dades aplicant un algorisme de PCA
- Finalment, aplicarem models de regressió lineal per veure si aquests expliquen o no la qualitat del vi en funció de la resta d'atributs i si el model resulta prou acurat

Matrius de correlació

Presentem de manera gràfica, dues gràfiques on queden representades de manera molt visible, els coeficients de correlació, i per tant, la relació directa o inversa entre els diferents atributs del data set

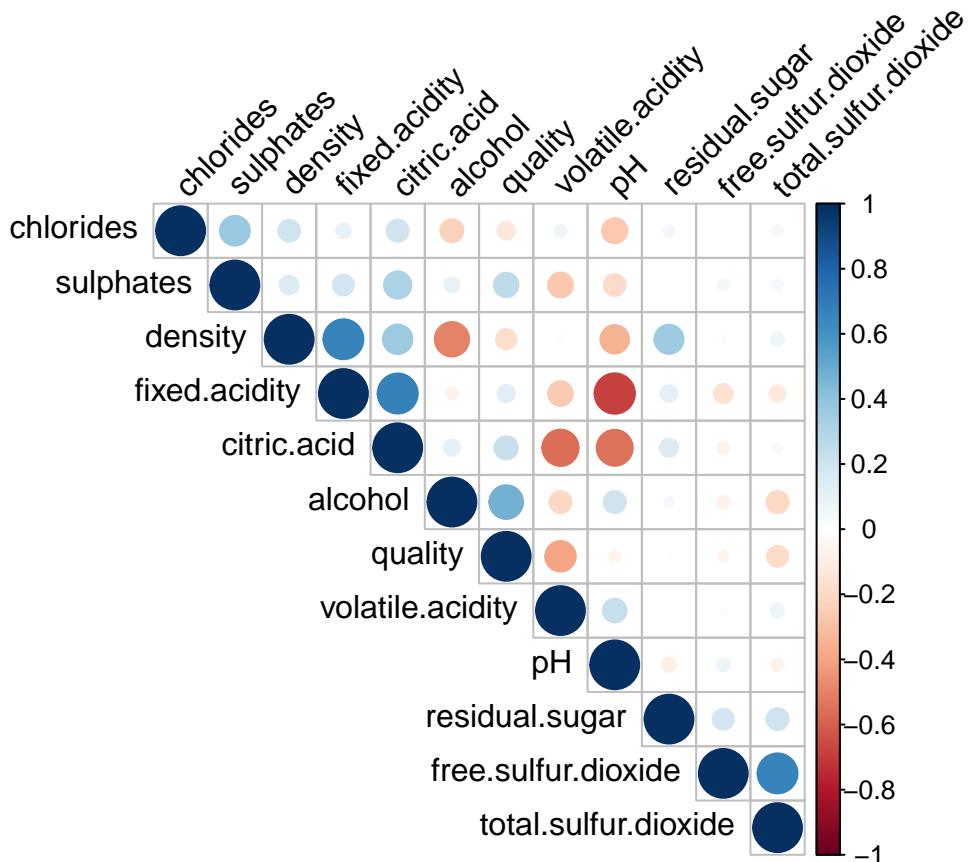
```
# Opcions pel grafic
options(repr.plot.width=4, repr.plot.height=3)

ggcorr(dfInput, geom = "blank", label = TRUE,
       hjust = 0.9, layout.exp = 2) +
  geom_point(size = 8, aes(color = coefficient > 0,
                            alpha = abs(coefficient) > 0.35)) +
  scale_alpha_manual(values = c("TRUE" = 0.25, "FALSE" = 0)) +
  guides(color = FALSE, alpha = FALSE)
```



```
# gràfica matriu correlació

corrplot(cor(dfInput), type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45)
```



Observacions

- La qualitat (**quality**) del vi sembla estar principalment relacionada amb els nivells d'alcohol (**alcohol**), de manera *directa* i amb la concentració d'àcid acètic (**volatile.acidity**), de manera *inversa*.
- El % d'alcohol d'un vi (**alcohol**) està *inversament relacionat* amb la seva densitat (**density**)
- El valor d'el pH (**pH**) està *inversa i significativament* relacionat amb atributs que tenen a veure amb l'acidesa (**citric.acid** i **fixed.acidity**)
- La densitat (**density**) és la **proprietat amb un major nombre de correl.lacions** amb d'altres atributs (*directa o inversament*) : **alcohol**, **residual.sugar**, **citric.acid** i **fixed.acidity**
- Els atributs que mesuren substàncies similars (**total.sulfur.dioxide** i **free.sulfur.dioxide**, per un costat i **citric.acid**, **volatile.acidity** i **fixed.acidity**, per l'altre) estan fortament correlacionades (*directa o inversament*), com era d'esperar i ja havíem enunciat.
- Finalment, l'atribut **fixed.acidity** és la que està més fortament correlacionada (*directa o inversament*) amb d'altres atributs: **citric.acid**, **density** i **pH**

Com a conclusions de les observacions adjuntes podem intuir que hi ha atributs **que semblen millors candidats** a ser explicats amb mètodes de regressió lineal que no pas la qualitat (**quality**)

PCA

NOTA : A l'apartat 5 d'aquest document s'adjunten diverses gràfiques relatives a aquesta secció (aportació de l'explicació de la variància per cada component, acumulats, scatter plots per components, ...)

Després d'analitzar la correl.lació que hi ha entre els diferents atributs, tot i que la seva cardinalitat NO és molt alta, semblaia que NO són necessaris tants atributs per determinar la qualitat del vi. De fet, veiem que l'atribut **quality** només està fortament correl.lacionat amb els atributs **alcohol** i **volatile.acidity**

Farem servir la tècnica del PCA, més que amb l'objectiu de reduir la cardinalitat del conjunt, com a eïna per detectar quins atributs són rellevants.

```
# Treiem l'atribut 'quality' i realitzem el PCA
dfInput.pca <- prcomp(dfInput[,1:11], scale. = TRUE)

# Examinem el model ...
summary(dfInput.pca)

## Importance of components:
##                               PC1        PC2        PC3        PC4        PC5        PC6        PC7
## Standard deviation    1.7604   1.3878   1.2452   1.1015   0.97943   0.81216   0.76406
## Proportion of Variance 0.2817   0.1751   0.1410   0.1103   0.08721   0.05996   0.05307
## Cumulative Proportion  0.2817   0.4568   0.5978   0.7081   0.79528   0.85525   0.90832
##                               PC8        PC9        PC10       PC11
## Standard deviation     0.65035  0.58706  0.42583  0.24405
## Proportion of Variance 0.03845  0.03133  0.01648  0.00541
## Cumulative Proportion  0.94677  0.97810  0.99459  1.00000
```

Veiem que l'aportació de cada component per explicar la variància del conjunt de dades és molt baixa i que es necessiten 7 components per explicar el 90% d'aquesta variància

```
# Aportació de cada atribut del dataset x component (-1 a +1)
dfInput.pca$rotation
```

```
##                               PC1        PC2        PC3        PC4
## fixed.acidity      0.48931422 -0.110502738  0.12330157 -0.229617370
## volatile.acidity   -0.23858436  0.274930480  0.44996253  0.078959783
## citric.acid        0.46363166 -0.151791356 -0.23824707 -0.079418256
## residual.sugar     0.14610715  0.272080238 -0.10128338 -0.372792562
## chlorides          0.21224658  0.148051555  0.09261383  0.666194756
## free.sulfur.dioxide -0.03615752  0.513566812 -0.42879287 -0.043537818
## total.sulfur.dioxide  0.02357485  0.569486959 -0.32241450 -0.034577115
## density            0.39535301  0.233575490  0.33887135 -0.174499758
## pH                 -0.43851962  0.006710793 -0.05769735 -0.003787746
## sulphates          0.24292133 -0.037553916 -0.27978615  0.550872362
## alcohol             -0.11323206 -0.386180959 -0.47167322 -0.122181088
##                               PC5        PC6        PC7        PC8
## fixed.acidity      0.08261366 -0.10147858  0.35022736 -0.17759545
## volatile.acidity   -0.21873452 -0.41144893  0.53373510 -0.07877531
## citric.acid        0.05857268 -0.06959338 -0.10549701 -0.37751558
## residual.sugar     -0.73214429 -0.04915555 -0.29066341  0.29984469
## chlorides          -0.24650090 -0.30433857 -0.37041337 -0.35700936
## free.sulfur.dioxide  0.15915198  0.01400021  0.11659611 -0.20478050
## total.sulfur.dioxide  0.22246456 -0.13630755  0.09366237  0.01903597
## density            -0.15707671  0.39115230  0.17048116 -0.23922267
## pH                 -0.26752977  0.52211645  0.02513762 -0.56139075
## sulphates          -0.22596222  0.38126343  0.44746911  0.37460432
```

```

## alcohol           -0.35068141 -0.36164504  0.32765090 -0.21762556
##                  PC9          PC10         PC11
## fixed.acidity    0.194020908  0.24952314  0.639691452
## volatile.acidity -0.129110301 -0.36592473  0.002388597
## citric.acid     -0.381449669 -0.62167708 -0.070910304
## residual.sugar   0.007522949 -0.09287208  0.184029964
## chlorides        0.111338666  0.21767112  0.053065322
## free.sulfur.dioxide 0.635405218 -0.24848326 -0.051420865
## total.sulfur.dioxide -0.592115893  0.37075027  0.068701598
## density          0.020718675  0.23999012 -0.567331898
## pH                -0.167745886  0.01096960  0.340710903
## sulphates        -0.058367062 -0.11232046  0.069555381
## alcohol           0.037603106  0.30301450 -0.314525906

```

Ara utilitzarem el nostre model obtingut via PCA amb els 7 components per veure si podem establir una regressió lineal entre la qualitat del vi i la resta de components i quina precisió té (paràmetre *R-squared*)

```

fitPCA <- lm(dfInput$quality ~ dfInput.pca$x[,1:7])
summary(fitPCA)

```

```

##
## Call:
## lm(formula = dfInput$quality ~ dfInput.pca$x[, 1:7])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.70114 -0.37034 -0.06334  0.49300  1.94724
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)               5.636023  0.016415 343.343 < 2e-16 ***
## dfInput.pca$x[, 1:7]PC1  0.050621  0.009327  5.427 6.61e-08 ***
## dfInput.pca$x[, 1:7]PC2 -0.225087  0.011832 -19.023 < 2e-16 ***
## dfInput.pca$x[, 1:7]PC3 -0.258946  0.013187 -19.637 < 2e-16 ***
## dfInput.pca$x[, 1:7]PC4 -0.032376  0.014908 -2.172  0.030 *
## dfInput.pca$x[, 1:7]PC5 -0.083721  0.016765 -4.994 6.57e-07 ***
## dfInput.pca$x[, 1:7]PC6  0.025114  0.020218  1.242  0.214
## dfInput.pca$x[, 1:7]PC7  0.095257  0.021491  4.432 9.95e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6564 on 1591 degrees of freedom
## Multiple R-squared:  0.3422, Adjusted R-squared:  0.3393
## F-statistic: 118.3 on 7 and 1591 DF,  p-value: < 2.2e-16

```

Observacions

El valor del paràmetre *R-squared* és molt baix cosa que ens indica que la precisió del model **no és bona**.

Regressió lineal

Anem a calcular un model de regressió lineal que tingui en compte tots els atributs del data set i que, en principi, hauria de ser el millor model de regressió que podem crear

```
fit0 <- lm(quality ~ ., data=dfInput)
summary(fit0)

##
## Call:
## lm(formula = quality ~ ., data = dfInput)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -2.68911 -0.36652 -0.04699  0.45202  2.02498 
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)            2.197e+01  2.119e+01   1.036   0.3002    
## fixed.acidity         2.499e-02  2.595e-02   0.963   0.3357    
## volatile.acidity      -1.084e+00 1.211e-01  -8.948  < 2e-16 ***  
## citric.acid          -1.826e-01 1.472e-01  -1.240   0.2150    
## residual.sugar        1.633e-02  1.500e-02   1.089   0.2765    
## chlorides             -1.874e+00 4.193e-01  -4.470  8.37e-06 ***  
## free.sulfur.dioxide  4.361e-03  2.171e-03   2.009   0.0447 *   
## total.sulfur.dioxide -3.265e-03 7.287e-04  -4.480  8.00e-06 ***  
## density               -1.788e+01 2.163e+01  -0.827   0.4086    
## pH                    -4.137e-01 1.916e-01  -2.159   0.0310 *   
## sulphates             9.163e-01 1.143e-01   8.014  2.13e-15 ***  
## alcohol               2.762e-01 2.648e-02  10.429  < 2e-16 ***  
## ---                
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
##
## Residual standard error: 0.648 on 1587 degrees of freedom
## Multiple R-squared:  0.3606, Adjusted R-squared:  0.3561 
## F-statistic: 81.35 on 11 and 1587 DF,  p-value: < 2.2e-16
```

Observem que la precisió/bondat del model (paràmetre *R-squared*), tenint en compte tots els atributs, **és molt baixa**

Atenent els codis de significació que acompanyen a cada atribut, crearem un model nou on només apareguin els atributs rellevants, a veure com difereix del model complert

```
fit1 <- lm(quality ~ volatile.acidity+chlorides+free.sulfur.dioxide+total.sulfur.dioxide+pH+alcohol+sul
summary(fit1)

##
## Call:
## lm(formula = quality ~ volatile.acidity + chlorides + free.sulfur.dioxide + 
##     total.sulfur.dioxide + pH + alcohol + sulphates, data = dfInput)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -2.68918 -0.36757 -0.04653  0.46081  2.02954 
##
## Coefficients:
```

```

##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)           4.4300987  0.4029168 10.995 < 2e-16 ***
## volatile.acidity     -1.0127527  0.1008429 -10.043 < 2e-16 ***
## chlorides            -2.0178138  0.3975417 -5.076 4.31e-07 ***
## free.sulfur.dioxide  0.0050774  0.0021255  2.389  0.017 *
## total.sulfur.dioxide -0.0034822  0.0006868 -5.070 4.43e-07 ***
## pH                   -0.4826614  0.1175581 -4.106 4.23e-05 ***
## alcohol              0.2893028  0.0167958 17.225 < 2e-16 ***
## sulphates            0.8826651  0.1099084  8.031 1.86e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6477 on 1591 degrees of freedom
## Multiple R-squared:  0.3595, Adjusted R-squared:  0.3567
## F-statistic: 127.6 on 7 and 1591 DF,  p-value: < 2.2e-16

```

Hem obtingut un model amb pràcticament la mateixa precisió que el total.

```
# Comparem les bondats dels 2 models
summary(fit0)$r.squared
```

```
## [1] 0.3605517
```

```
summary(fit1)$r.squared
```

```
## [1] 0.3594709
```

Ja per acabar, procedirem a la predicción de la categoria d'un vi amb dades de mostra inventades

```
# Predim la qualitat d'un nou vi inventat
```

```
set.seed(123)
```

```
fixed.acidity = sample(dfInput$fixed.acidity,1)
volatile.acidity = sample(dfInput$volatile.acidity,1)
citric.acid = sample(dfInput$citric.acid,1)
residual.sugar = sample(dfInput$residual.sugar,1)
chlorides = sample(dfInput$chlorides,1)
free.sulfur.dioxide = sample(dfInput$free.sulfur.dioxide,1)
total.sulfur.dioxide = sample(dfInput$total.sulfur.dioxide,1)
density = sample(dfInput$density,1)
pH = sample(dfInput$pH,1)
sulphates = sample(dfInput$sulphates,1)
alcohol = sample(dfInput$alcohol,1)
```

```
dadesMostra <- data.frame(fixed.acidity, volatile.acidity,citric.acid, residual.sugar,chlorides,free.su
```

```
str(dadesMostra)
```

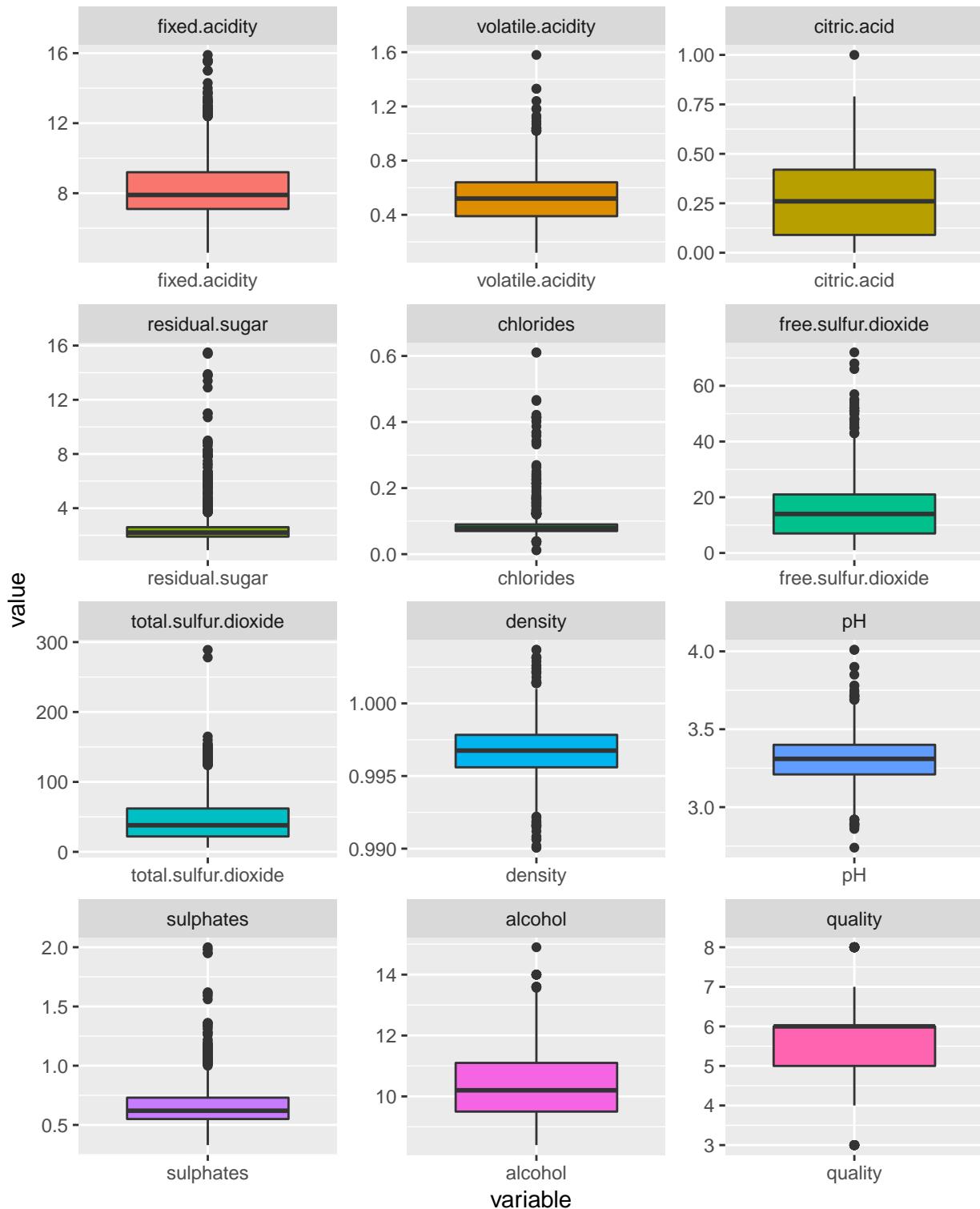
```
## 'data.frame':   1 obs. of  11 variables:
## $ fixed.acidity      : num 11.6
## $ volatile.acidity    : num 0.635
## $ citric.acid        : num 0.59
## $ residual.sugar      : num 2.4
## $ chlorides          : num 0.077
## $ free.sulfur.dioxide : num 18
## $ total.sulfur.dioxide: num 42
## $ density             : num 0.997
```

```
## $ pH : num 3.48
## $ sulphates : num 0.67
## $ alcohol : num 10.2
predict(fit0, dadesMostra)
```

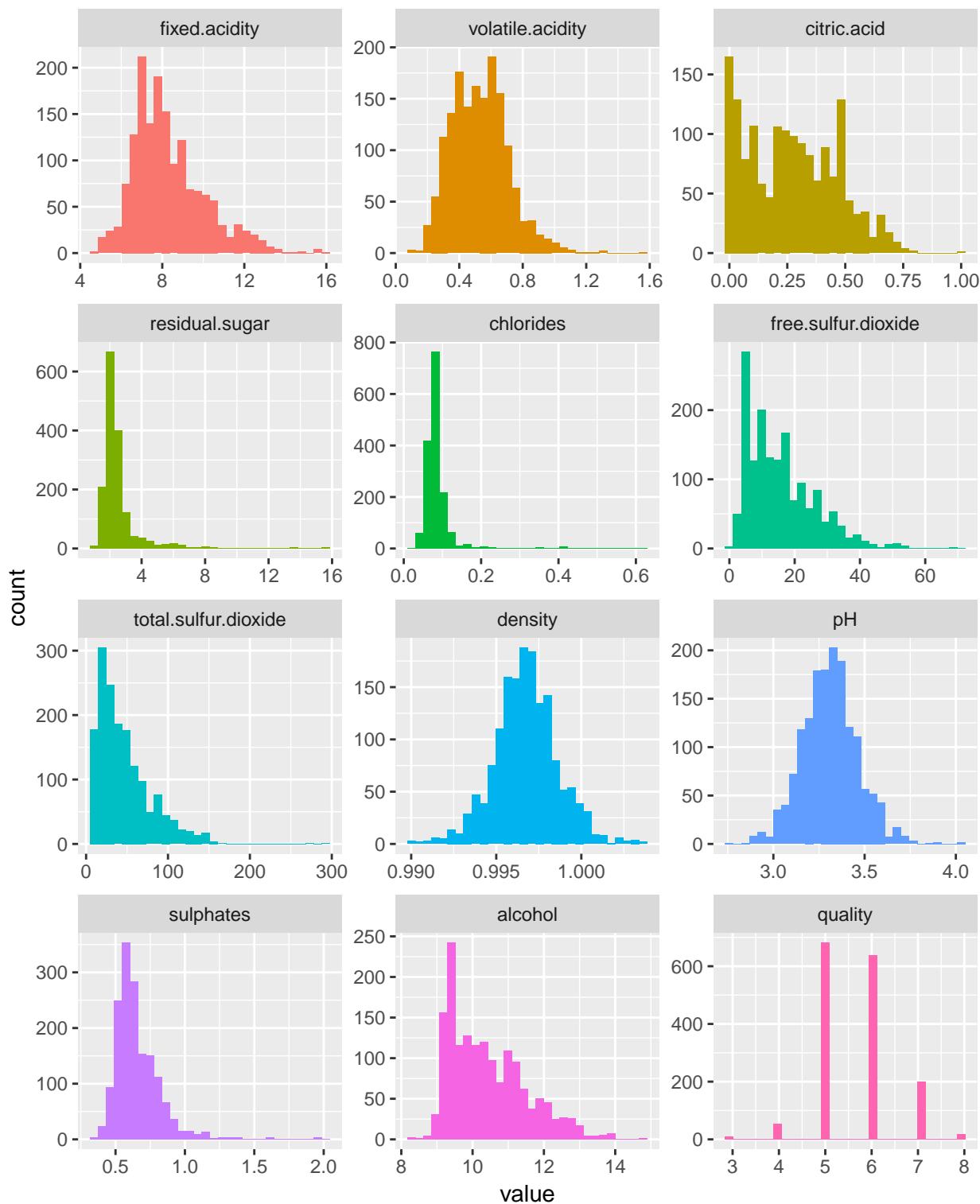
```
##      1
## 5.457917
```

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

Boxplot de cada atribut

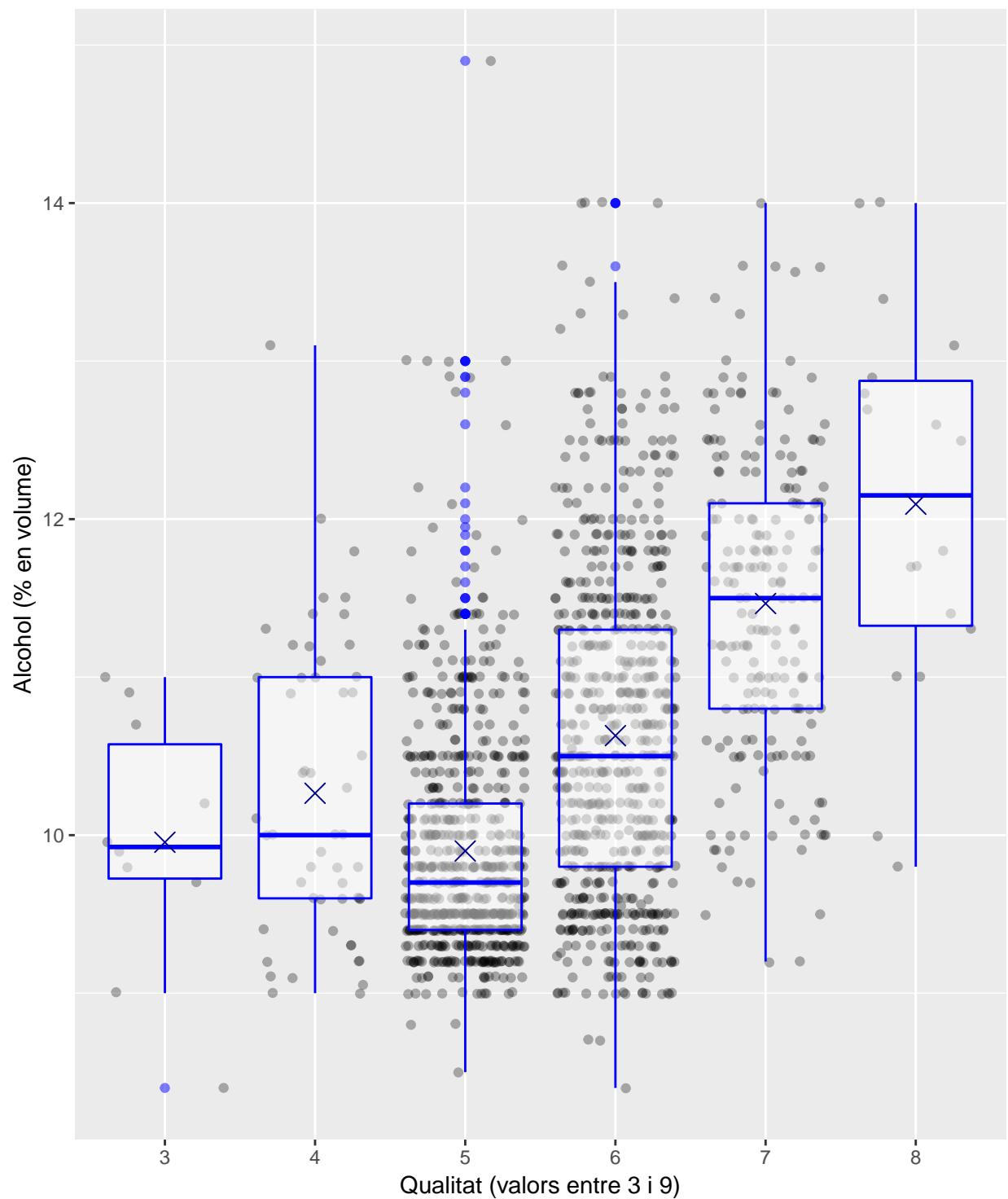


Histogrammes de cada atribut

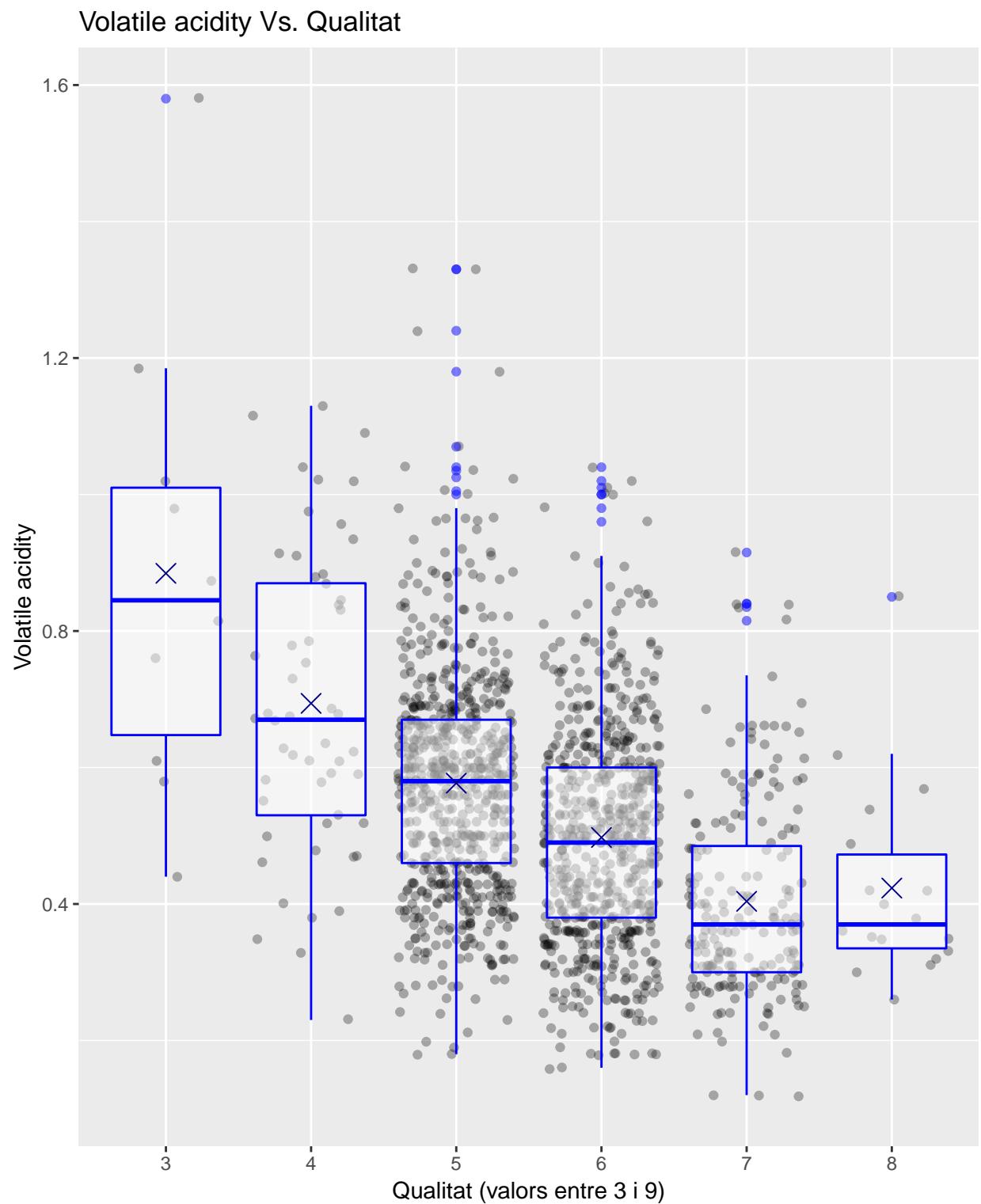


Alcohol vs Qualitat

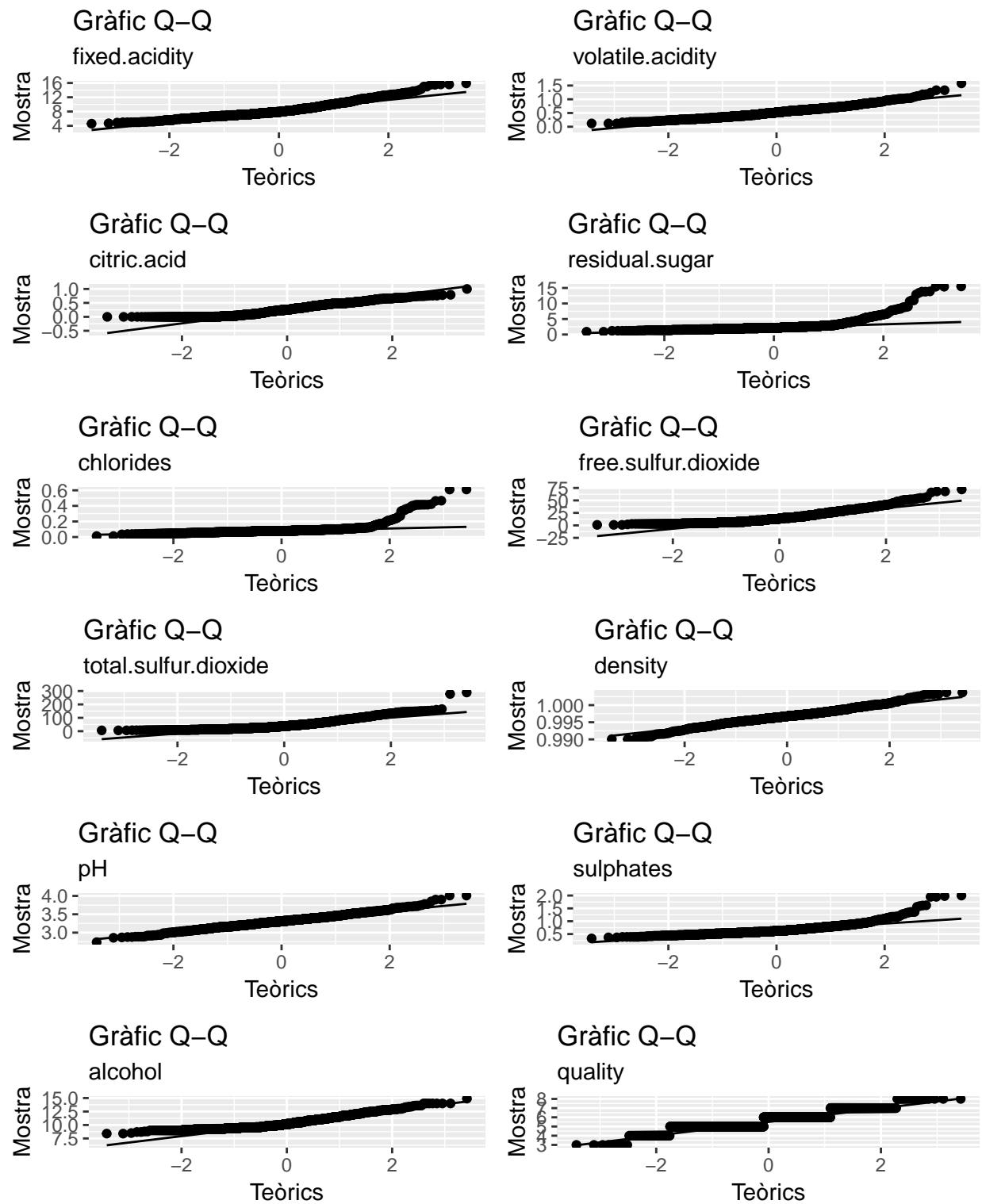
Alcohol Vs. Qualitat



Volatile acidity vs Qualitat

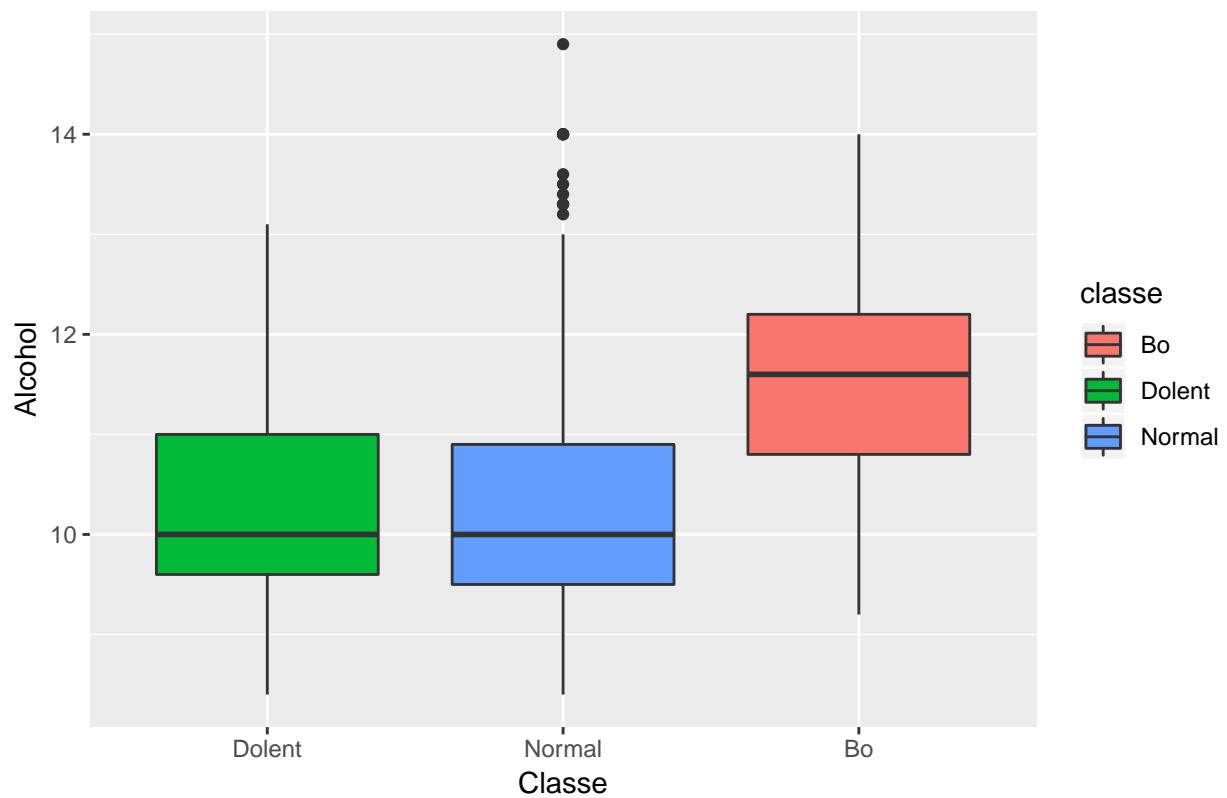


Normalitat : gràfics Q-Q (normalitat) de cada atribut

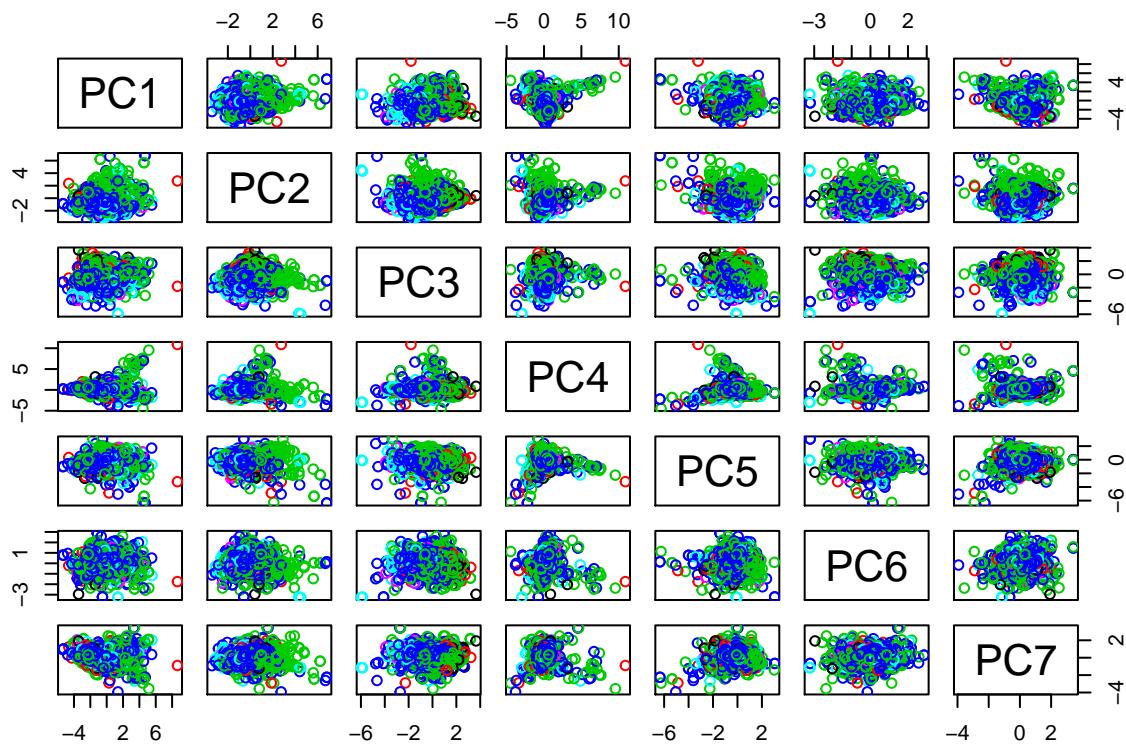


Variància : gràfic comparant la variància segons la classe de vi

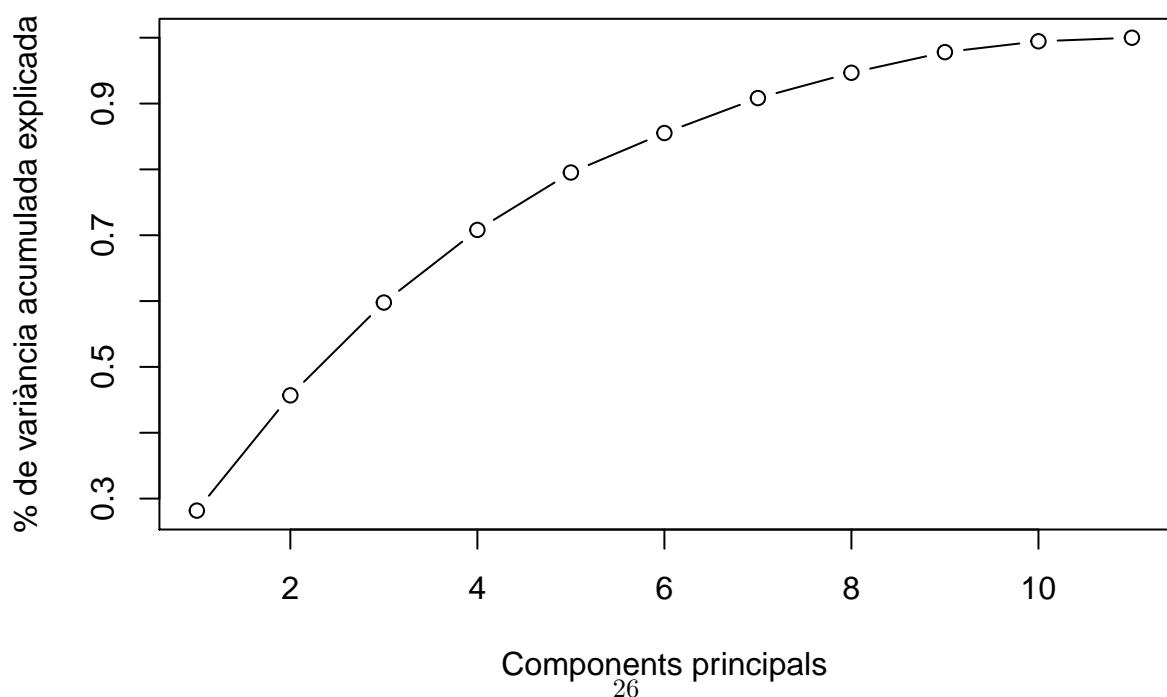
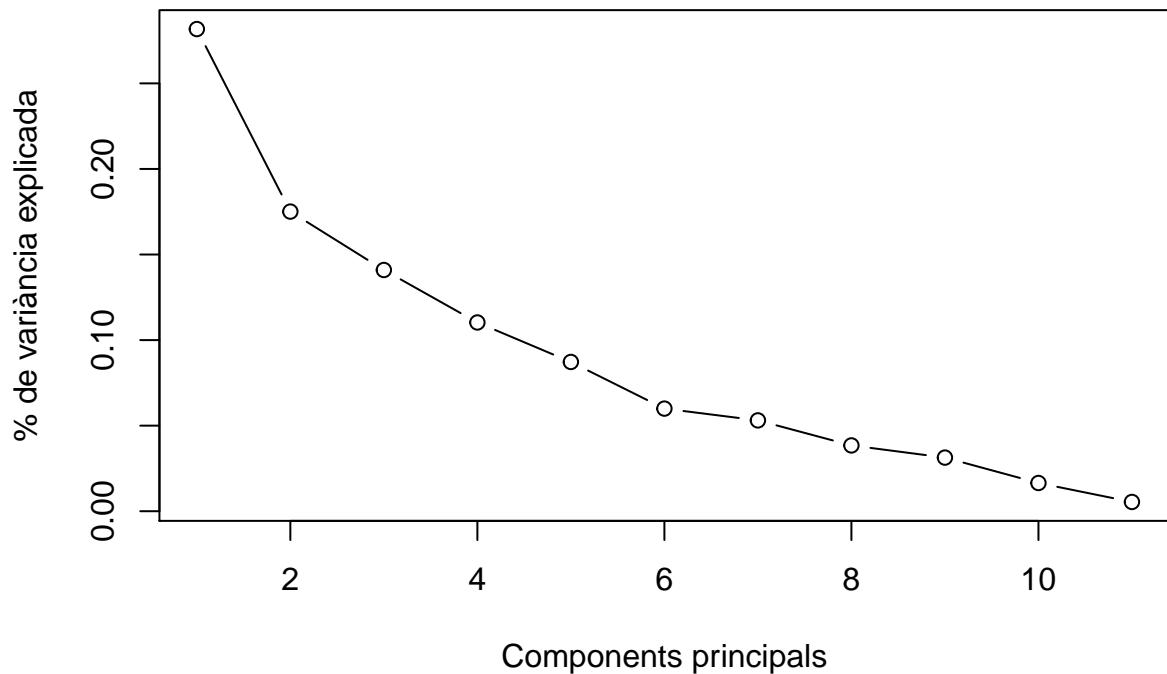
Gràfica Alcohol (asc) agrupant per Classe



PCA : gràfic que mostra la relació entre els diferents components



PCA : gràfic que mostra el % (parcial i acumulat) de la variància segons component



6. Resolució del problema.

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

Després de la realització de l'anàlisi i a la vista dels resultats clarament podem conoure que:

- la *regressió lineal* NO és un bon model que expliqui la qualitat d'un vi en funció dels atributs existents. Segurament, això és degut, en part, a que estem intentant predir *criteris subjectius (quality)* a partir de *valors objectius* (propietats mesurables d'un vi). En aquest cas, potser funcionaria millor *discretitzar* la qualitat del vi en 2 possibles valors binaris ($bo=1$, dolent=0) i aplicar un model de *regressió logística*
- Per altra banda, el model de regressió lineal probablement serviria per explicar, amb un cert grau de bondat, altres atributs, sense tenir en compte la qualitat.

Bibliografia i referències

- Materials de l'assignatura ‘Tipología i cicle de vida de les dades’, UOC
- Dataset de mostra : <https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009>
- <https://stackoverflow.com/questions/49044753/scale-kable-table-to-fit-page-width>
- <https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf>
- <https://owi.usgs.gov/blog/boxplots/>
- <http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html#Box%20Plot>
- <https://www.cyclismo.org/tutorial/R/pValues.html>
- <http://www.sthda.com>
- <http://r-statistics.co/Outlier-Treatment-With-R.html>
- <https://www.kaggle.com/tsilveira/wine-r/comments>
- <https://owi.usgs.gov/blog/boxplots/>
- <https://stackoverflow.com/questions/7196450/create-a-data-frame-of-unequal-lengths>
- <https://stackoverflow.com/questions/34004008/transposing-in-dplyr>
- <https://briatte.github.io/ggcorr/#controlling-the-main-geometry>
- https://rpubs.com/Joaquin_AR/287787
- <https://towardsdatascience.com>