PRÀCTICA 2

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

El dataset utilitzat per a la realització d'aquesta pràctica és "Wine Quality Data Set" de l'UCI Machine Learning Repository. Aquest està compost de 2 datasets diferenciats amb observacions de vins rosats i blancs de la denominació d'origen vinho verde del nord de Portugal.

Enllaç: https://archive.ics.uci.edu/ml/datasets/wine+quality

Les variables d'aquests datasets es mostren a continuació:

Variables d'entrada (basades en tests fisicoquímics):

1 - fixed acidity

2 - volatile acidity

3 - citric acid

4 - residual sugar

5 - chlorides

6 - free sulfur dioxide

7 - total sulfur dioxide

8 - density

9 - pH

10 - sulphates

11 - alcohol

Variable de sortida (basada en dades sensorials):

12 - quality (puntuació entre 0 i 10)

Tot i que en un principi el dataset està pensat per predir la qualitat d'un vi a partir de les seves característiques, en aquesta pràctica ens centrarem en identificar el tipus de vi (rosat o blanc) a partir de les seves característiques fisicoquímiques, observar quines d'aquestes contribueixen més a l'hora de classificar els vins i si hi ha diferències significatives entre les mitjanes de les variables dels dos grups.

El dataset dels vins rosats consta de 1.599 observacions i 12 variables.

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

El dataset dels vins blancs consta de 4.898 observacions i 12 variables.

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

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

En un principi volem fer servir totes les variables disponibles dels datasets, ja que no sabem a priori quines poden ser més útils i quines no.

Així doncs, afegim una altra columna als dos datasets anomenada type, que contindrà el valor "red" per als vins rosats i el valor "white" per als vins blancs.

Un cop fet això, s'integraran els dos conjunts de dades en un de sol, que passarà a tenir 6.497 observacions i 13 variables.

```
## '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 ...
##  $ type                : Factor w/ 2 levels "red","white": 1 1 1 1 1 1 1 1 1 1 ...
```

3. Neteja de les dades.

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

Després de comprovar si hi ha valors nuls o buits, observem que no n'hi ha cap, així que no s'ha de fer cap tractament addicional en aquest cas.

Però al comprovar si les dades contenen 0, veiem que la variable cítric.acid sí que en conté en un total de 151 files.

```
##        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 
##                 type 
##                    0
```

Tot i això, tenint en compte la resta de valors per a aquesta variable, sembla que aquest valor forma part del domini i, per tant, és vàlid en aquest context.

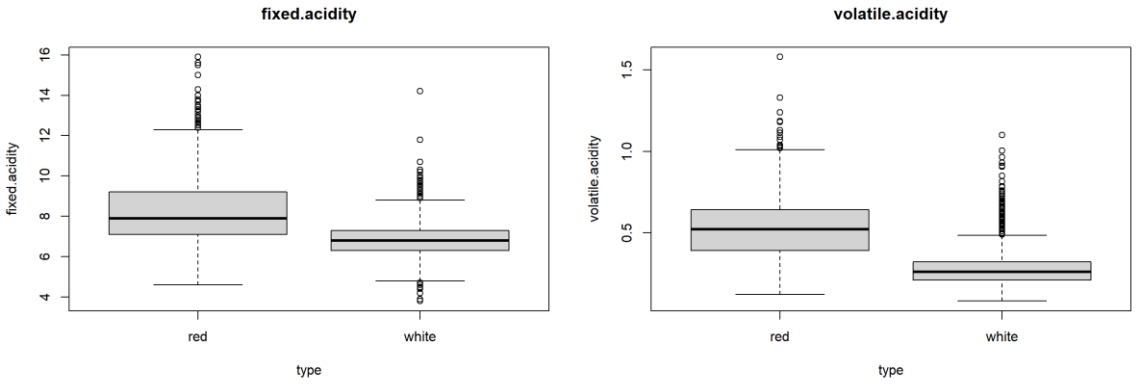
En cas d'haver trobat valors nuls o buits, com que no tenim accés a la font original de les dades, s'hagués pogut substituir el valor per la mitjana de la columna agrupada per la variable type. En el cas que en una mateixa observació li faltessin més de la meitat dels valors, s'eliminaria del data.frame, ja que seria una observació poc representativa de la realitat.

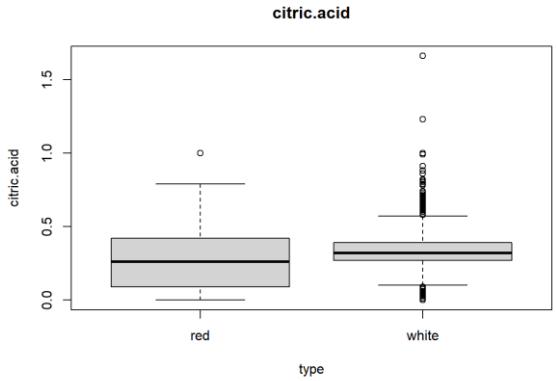
3.2. Identificació i tractament de valors extrems.

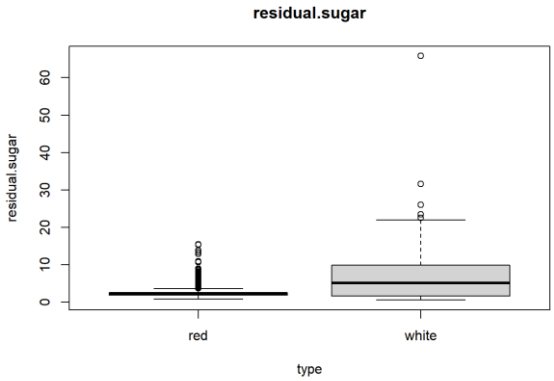
Per a la identificació de valors extrems s'ha fet servir el mètode de diagrama de caixa o boxplot.

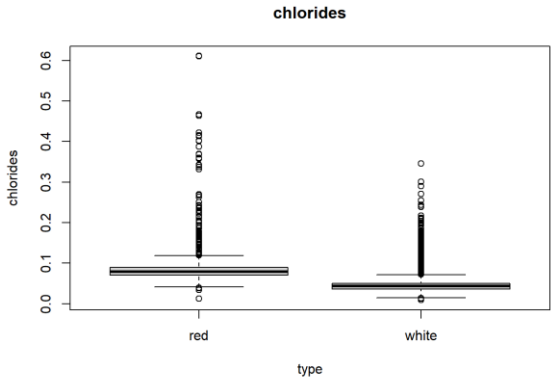
Els diagrames de caixa ajuden a visualitzar una variable quantitativa mostrant cinc valors de resum comuns (mínim, mediana, primer i tercer quartils i màxim) i qualsevol observació que s'hagi classificat com a valor atípic sospitós mitjançant el criteri del rang interquartil (IQR, interquartile range, en anglès). El criteri IQR significa que totes les observacions anteriors a $q_{0,75} + 1{,}5 \cdot IQR$ o per sota de $q_{0,25} - 1{,}5 \cdot IQR$ es consideren possibles valors atípics per R (on $q_{0,25}$ i $q_{0,75}$ corresponen al primer i tercer quartils respectivament, i IQR és la diferència entre el tercer i el primer quartil).

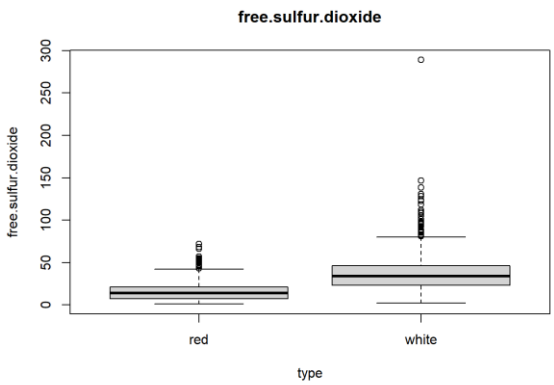
Els valors atípics es mostren en el gràfic com a punts o cercles.

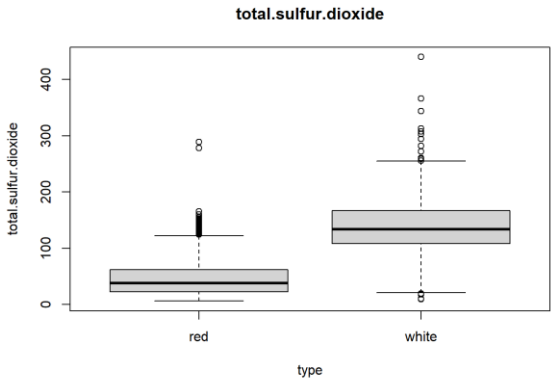


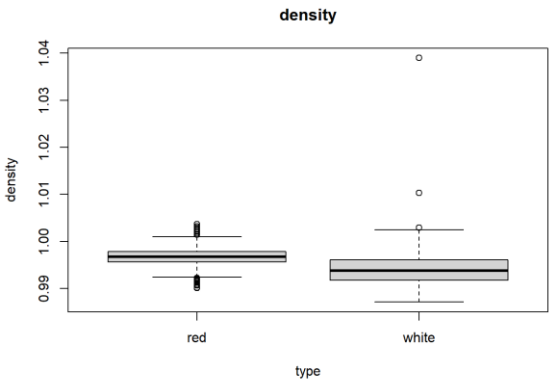

citric.acid
citric.acid
0.0
0.5
1.0
1.5
red
white
type

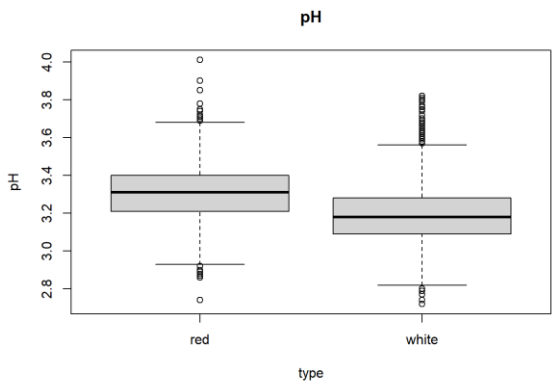

residual.sugar
residual.sugar
0
10
20
30
40
50
60
red
white
type

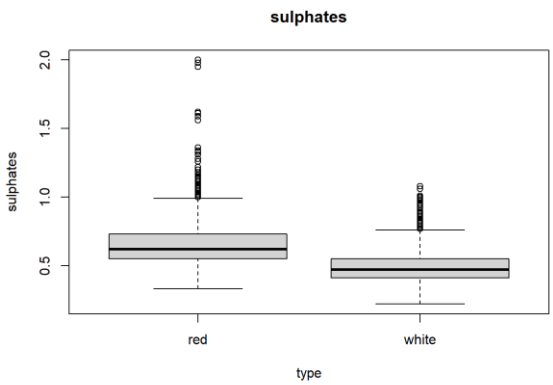

chlorides
chlorides
0.0
0.1
0.2
0.3
0.4
0.5
0.6
red
white
type

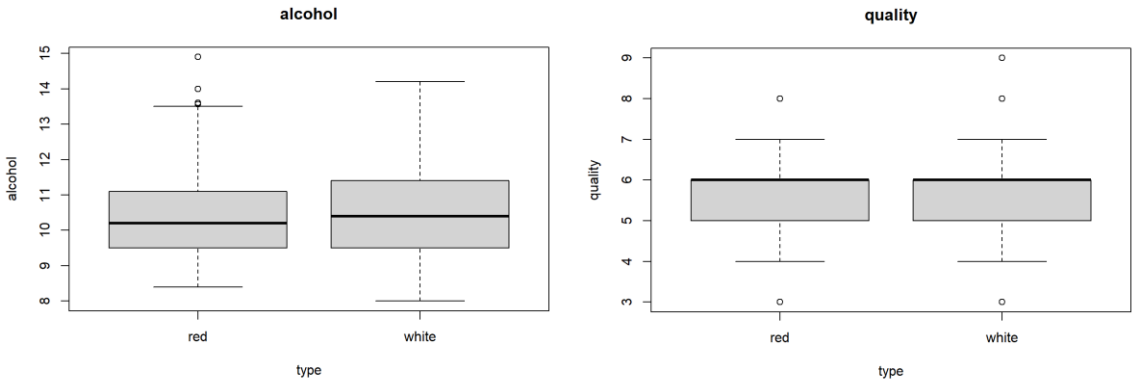

free.sulfur.dioxide
free.sulfur.dioxide
0
50
100
150
200
250
300
red
white
type


total.sulfur.dioxide
total.sulfur.dioxide
0
100
200
300
400
red
white
type


density
density
0.99
1.00
1.01
1.02
1.03
1.04
red
white
type


pH
pH
2.8
3.0
3.2
3.4
3.6
3.8
4.0
red
white
type


sulphates
sulphates
0.5
1.0
1.5
2.0
red
white
type



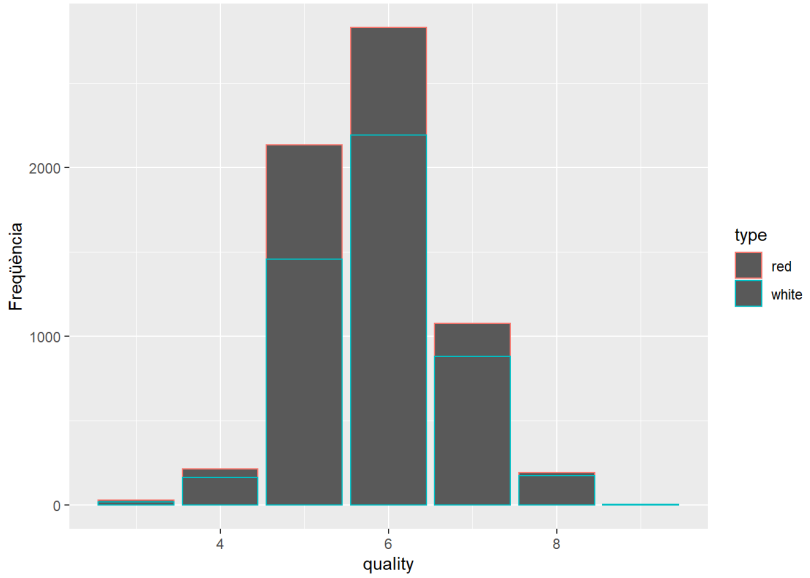
Es pot observar que hi ha punts detectats com a possibles outliers en totes les variables, tot i això, la majoria no estan excessivament allunyats, cosa que fa pensar que no es tracta d'errors sinó de valors vàlids. Per aquest motiu, s'ha decidit seleccionar aquelles observacions amb alguna de les característiques més allunyades visualment respecte a la resta de les del seu grup, i eliminar-les del dataset, ja que es tracta de vins amb unes característiques bastant particulars i diferents de la resta.

En total s'han eliminat 12 vins.

4. Anàlisi de les dades.

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

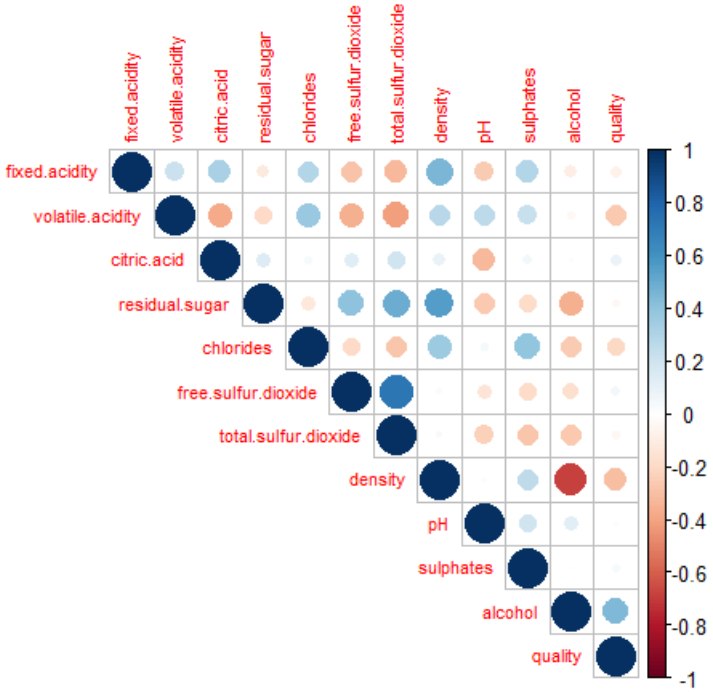
Observant la distribució de la variable quality es veu que no està igualment distribuïda, amb molts vins de qualitat intermèdia, però molt pocs d'excel·lents o horribles.



Per aquest motiu, també s'eliminaran del data.frame aquells vins amb una puntuació de 3 o de 9, quedant-nos amb un grup de vins molt més homogenis entre sí.

En total hi havia 5 vins amb puntuació de 9 i 29 vins amb puntuació de 3, per tant el data.frame queda amb 6.451 observacions.

Per saber quines de les variables és interessant incloure a l'anàlisi i si n'hi ha alguna de no rellevant, es farà un anàlisi de correlacions.



Com es pot comprovar en el gràfic de correlacions, les variables free.sulfur.dioxide i total.sulfur.dioxide estan molt relacionades, així que no aporten gaire informació extra i se'n pot eliminar una. En aquest cas s'ha decidit prescindir de la variable free.sulfur.dioxide.

Totes les altres variables independents es conservaran per als anàlisis posteriors.

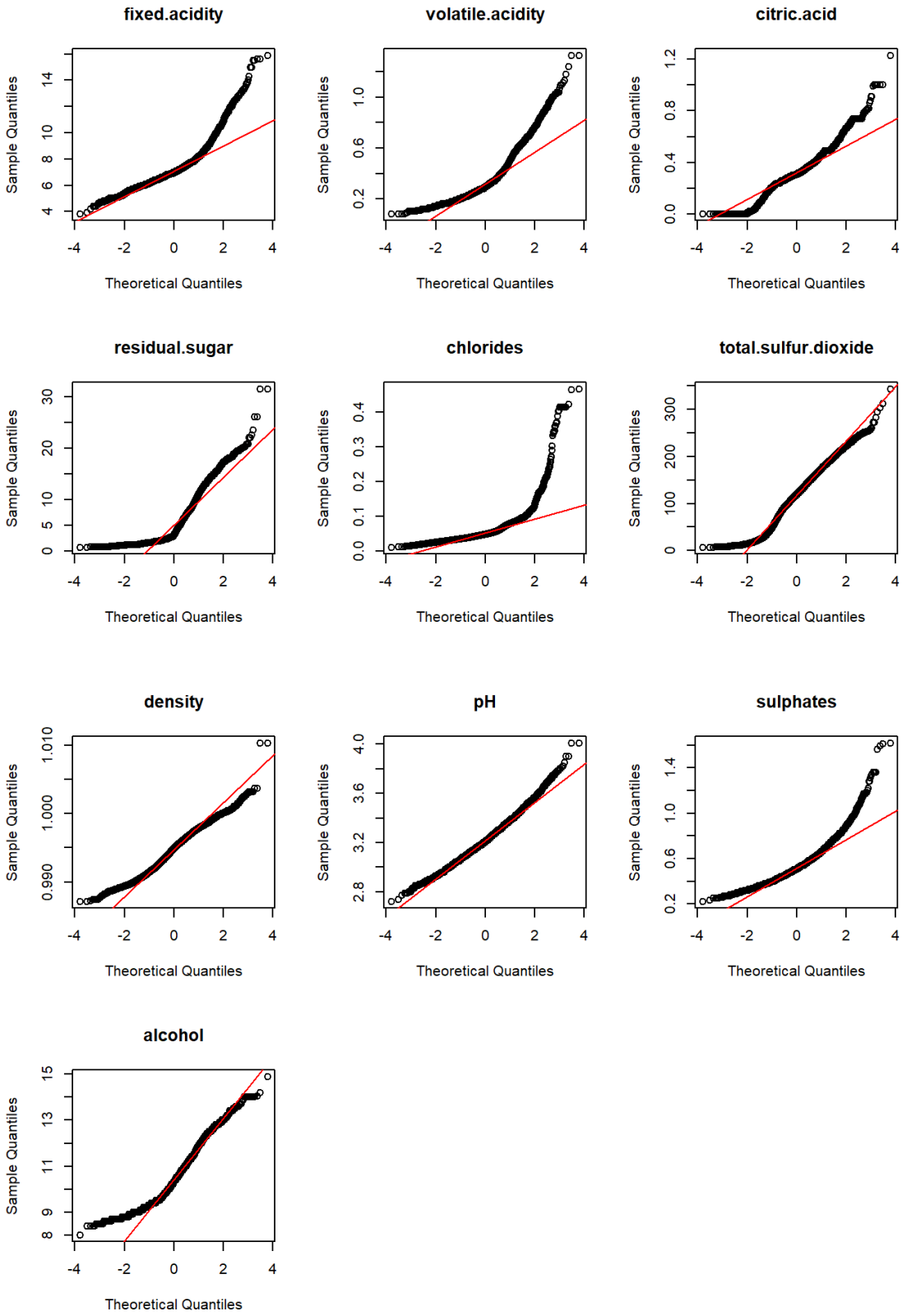
Com em planejat a l'inici del projecte, es vol aconseguir comparar els grups de vins de tipus rosat i blanc, així com entendre quines variables són més importants.

Per això, els anàlisis proposats són:

- Contrast d'hipòtesi per a cada variable independent
- Clustering
- Regressió logística

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

Per comprovar la normalitat de les diferents variables es poden fer servir els gràfics Q-Q.



Es pot observar que els punts no coincideixen perfectament amb les línies, això significa que les variables no segueixen uns distribució normal.

També es pot observar visualment gràcies a gràfics de densitat, en els quals es pot dir que una variable segueix una distribució normal si el gràfic de forma de campana simètrica.

Finalment, es pot fer el test de normalitat de Shapiro-Wilk, que es mostra resumit en la següent taula:

data	W	p-value
wine_sample$fixed.acidity	0.87574	< 2.2e-16
wine_sample$volatile.acidity	0.87481	< 2.2e-16
wine_sample$citric.acid	0.97341	< 2.2e-16
wine_sample$residual.sugar	0.83572	< 2.2e-16
wine_sample$chlorides	0.62952	< 2.2e-16
wine_sample$total.sulfur.dioxide	0.98154	< 2.2e-16
wine_sample$density	0.98563	< 2.2e-16
wine_sample$pH	0.99135	< 2.2e-16
wine_sample$sulphates	0.9201	< 2.2e-16
wine_sample$alcohol	0.95254	< 2.2e-16

El p-value en el test de normalitat de Shapiro-Wilk és inferior a 0.05 en totes les variables cosa que implica que la distribució d'aquestes és significativament diferent de la distribució normal, i per tant no podem assumir-ne la normalitat.

Tot i això, el Teorema del Límit Central diu que el contrast d'hipòtesis sobre la mitjana d'una mostra s'aproxima a una distribució normal encara que la població original no segueixi una distribució normal, sempre que la mida de la mostra sigui suficientment gran (n>30). Com que la mostra té més de 30 observacions podem aplicar el TLC.

Per comprovar l'homogeneïtat de la variància es fa servir la funció var.test() de R, que es mostra resumit a la següent taula:

variable	p-value
fixed.acidity	< 2.2e-16
volatile.acidity	< 2.2e-16
citric.acid	< 2.2e-16
residual.sugar	< 2.2e-16
chlorides	< 2.2e-16
total.sulfur.dioxide	< 2.2e-16
density	< 2.2e-16
pH	0.536
sulphates	< 2.2e-16
alcohol	1.017e-11

Pel que fa al test per comprovar si les variàncies de les variables són iguals o diferents entre els vins rosat i els blancs, la majoria dels p-valors són inferiors a 0.05 i, per tant, es refusa la hipòtesis nul·la de que la variable per als vins rosats i els blancs prové de distribucions amb la mateixa variància, amb l'excepció de la variable pH, que té un p-valor > 0.05.

4.3. Aplicació de proves estadístiques per comparar els grups de dades. En funció de les dades i de l'objectiu de l'estudi, aplicar proves de contrast d'hipòtesis, correlacions, regressions, etc. Aplicar almenys tres mètodes d'anàlisi diferents.

- Contrast d'hipòtesis

Es farà un test sobre la mitjana de dues mostres de poblacions independents amb distribucions normals considerant que les variàncies són desconegudes i diferents (menys a la variable pH), que es realitza amb un estadístic de contrast que segueix una distribució *t* d'Student amb *v* graus de llibertat.

<u>fixed.acidity</u>

Hipòtesi nul·la, H0: μ(red)=μ(white) (la fixed.acidity és igual entre els vins rosats que els blancs)

Hipòtesi alternativa, H1: μ(red)>μ(white) (la fixed.acidity és superior entre els vins rosats)

```
##
##  Welch Two Sample t-test
##
## data:  wine.red$fixed.acidity and wine.white$fixed.acidity
## t = 32.332, df = 1819.1, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  1.39496     Inf
## sample estimates:
## mean of x mean of y
##  8.319216  6.849445
```

El p-valor és més petit que 0.05, per tant, refusem la hipòtesis nul·la, i podem concloure que amb un nivell de confiança del 95% que la fixed.acidity entre els vins rosats és superior a la dels vins blancs.

<u>volatile.acidity</u>

Hipòtesi nul·la, H0: μ(red)=μ(white) (la volatile.acidity és igual entre els vins rosats que els blancs)

Hipòtesi alternativa, H1: μ(red)>μ(white) (la volatile.acidity és superior entre els vins rosats)

```
##
##  Welch Two Sample t-test
##
## data:  wine.red$volatile.acidity and wine.white$volatile.acidity
## t = 53.4, df = 1925.1, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.240477      Inf
## sample estimates:
## mean of x mean of y
## 0.5260051 0.2778815
```

El p-valor és més petit que 0.05, per tant, refusem la hipòtesis nul·la, i podem concloure que amb un nivell de confiança del 95% que la volatile.acidity entre els vins rosats és superior a la dels vins blancs.

citric.acid

Hipòtesi nul·la, H0: μ(red)=μ(white) (la citric.acid és igual entre els vins rosats que els blancs)

Hipòtesi alternativa, H1: μ(red)<μ(white) (la citric.acid és inferior entre els vins rosats)

```
##
##  Welch Two Sample t-test
##
## data:  wine.red$citric.acid and wine.white$citric.acid
## t = -12.329, df = 1990.5, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##         -Inf -0.05497333
## sample estimates:
## mean of x mean of y
## 0.2702971 0.3337379
```

El p-valor és més petit que 0.05, per tant, refusem la hipòtesis nul·la, i podem concloure que amb un nivell de confiança del 95% que la citric.acid entre els vins rosats és inferior a la dels vins blancs.

residual.sugar

Hipòtesi nul·la, H0: μ(red)=μ(white) (el residual.sugar és igual entre els vins rosats que els blancs)

Hipòtesi alternativa, H1: μ(red)<μ(white) (el residual.sugar és inferior entre els vins rosats)

```
##
##  Welch Two Sample t-test
##
## data:  wine.red$residual.sugar and wine.white$residual.sugar
## t = -48.236, df = 6358.6, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -3.719374
## sample estimates:
## mean of x mean of y
##  2.531985  6.382686
```

El p-valor és més petit que 0.05, per tant, refusem la hipòtesis nul·la, i podem concloure que amb un nivell de confiança del 95% que el residual.sugar entre els vins rosats és inferior a la dels vins blancs.

chlorides

Hipòtesi nul·la, H0: μ(red)=μ(white) (el chlorides és igual entre els vins rosats que els blancs)

Hipòtesi alternativa, H1: μ(red)>μ(white) (el chlorides és superior entre els vins rosats)

```
##
##  Welch Two Sample t-test
##
## data:  wine.red$chlorides and wine.white$chlorides
## t = 36.227, df = 1848.2, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.03897605        Inf
## sample estimates:
##  mean of x  mean of y
## 0.08658850 0.04575765
```

El p-valor és més petit que 0.05, per tant, refusem la hipòtesis nul·la, i podem concloure que amb un nivell de confiança del 95% que el chlorides entre els vins rosats és superior a la dels vins blancs.

total.sulfur.dioxide

Hipòtesi nul·la, H0: μ(red)=μ(white) (el total.sulfur.dioxide és igual entre els vins rosats que els blancs)

Hipòtesi alternativa, H1: μ(red)<μ(white) (el total.sulfur.dioxide és inferior entre els vins rosats)

```
##
##  Welch Two Sample t-test
##
## data:  wine.red$residual.sugar and wine.white$residual.sugar
## t = -48.236, df = 6358.6, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -3.719374
## sample estimates:
## mean of x mean of y
##  2.531985  6.382686
```

El p-valor és més petit que 0.05, per tant, refusem la hipòtesis nul·la, i podem concloure que amb un nivell de confiança del 95% que el total.sulfur.dioxide entre els vins rosats és inferior a la dels vins blancs.

density

Hipòtesi nul·la, H0: μ(red)=μ(white) (la density és igual entre els vins rosats que els blancs)

Hipòtesi alternativa, H1: μ(red)>μ(white) (la density és superior entre els vins rosats)

```
##
##  Welch Two Sample t-test
##
## data:  wine.red$density and wine.white$density
## t = 43.115, df = 4183.8, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.002622077         Inf
## sample estimates:
## mean of x mean of y
## 0.9967440 0.9940179
```

El p-valor és més petit que 0.05, per tant, refusem la hipòtesis nul·la, i podem concloure que amb un nivell de confiança del 95% que la density entre els vins rosats és superior a la dels vins blancs.

pH

Hipòtesi nul·la, H0: μ(red)=μ(white) (el pH és igual entre els vins rosats que els blancs)

Hipòtesi alternativa, H1: μ(red)>μ(white) (el pH és superior entre els vins rosats)

```
##
##  Welch Two Sample t-test
##
## data:  wine.red$pH and wine.white$pH
## t = 28.168, df = 2655.2, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.1167765       Inf
## sample estimates:
## mean of x mean of y
##  3.312187  3.188166
```

El p-valor és més petit que 0.05, per tant, refusem la hipòtesis nul·la, i podem concloure que amb un nivell de confiança del 95% que el pH entre els vins rosats és superior a la dels vins blancs.

sulphates

Hipòtesi nul·la, H0: μ(red)=μ(white) (el sulphates és igual entre els vins rosats que els blancs)

Hipòtesi alternativa, H1: μ(red)>μ(white) (el sulphates és superior entre els vins rosats)

```
##
##  Welch Two Sample t-test
##
## data:  wine.red$sulphates and wine.white$sulphates
## t = 38.936, df = 2158.8, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.1583323       Inf
## sample estimates:
## mean of x mean of y
## 0.6551960 0.4898768
```

El p-valor és més petit que 0.05, per tant, refusem la hipòtesis nul·la, i podem concloure que amb un nivell de confiança del 95% que el sulphates entre els vins rosats és superior a la dels vins blancs.

alcohol

Hipòtesi nul·la, H0: μ(red)=μ(white) (el alcohol és igual entre els vins rosats que els blancs)

Hipòtesi alternativa, H1: μ(red)<μ(white) (el alcohol és inferior entre els vins rosats)

```
##
##  Welch Two Sample t-test
##
## data:  wine.red$alcohol and wine.white$alcohol
## t = -2.7097, df = 3058.4, p-value = 0.003386
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -0.034139
## sample estimates:
## mean of x mean of y
##  10.42592  10.51283
```

El p-valor és més petit que 0.05, per tant, refusem la hipòtesis nul·la, i podem concloure que amb un nivell de confiança del 95% que el alcohol entre els vins rosats és inferior a la dels vins blancs.

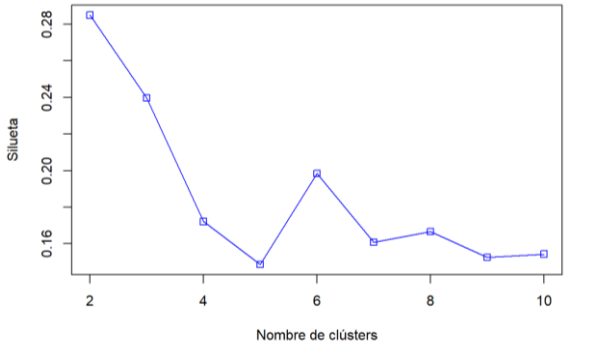
Com es pot observar després de fer el contrast d'hipòtesis per a cada variable, hi ha diferències significatives entre les mitjanes de totes les variables independents per als grups de vins rosats i blancs.

- Clustering

Es farà un anàlisi de clustering per veure com s'agrupen les dades sense la etiqueta de classe (no supervisat).

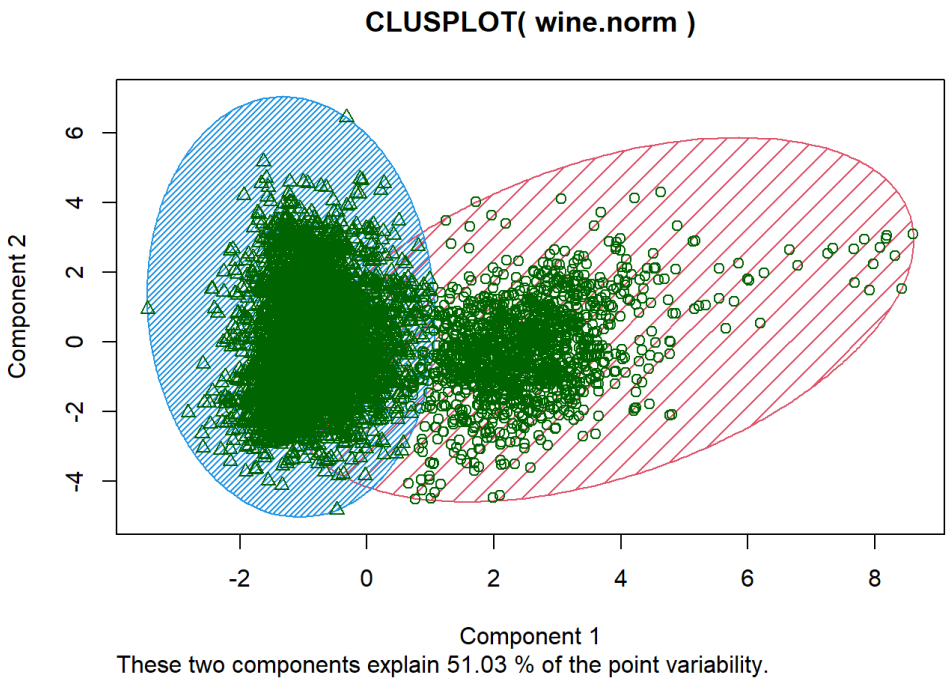
En primer lloc es normalitza el data.frame ja que està en diferents escales.

Per saber el nombre ideal de clústers es pot comprovar el valor de la silueta.



Sembla que el nombre de clústers que dona una silueta més gran com es pot apreciar al gràfic és 2. És el valor esperat, ja que se sap que hi ha dos tipus de vins, els rosats i els blancs.

Després d'aplicar l'algorisme kmeans per a 2 clústers les dades s'agrupen de la següent forma:



Els dos grups estan prou ben diferenciats, així que sembla que les característiques fisicoquímiques dels vins serveixen per distingir si aquests són rosats o blancs.

- Regressió logística

Per últim s'aplicarà una regressió logística per veure si totes les variables serveixen per explicar la variable dependent i quines contribueixen més.

```
##
## Call:
## glm(formula = type ~ fixed.acidity + volatile.acidity + citric.acid +
##     residual.sugar + chlorides + total.sulfur.dioxide + density +
##     pH + sulphates + alcohol, family = binomial(), data = wine.norm)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.1743   0.0010   0.0118   0.0345   4.5851
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)             4.4927     0.2772  16.209  < 2e-16 ***
## fixed.acidity           1.1759     0.3234   3.636 0.000277 ***
## volatile.acidity       -0.8163     0.1662  -4.913 8.99e-07 ***
## citric.acid             0.5375     0.1938   2.773 0.005558 **
## residual.sugar          4.7503     0.4966   9.567  < 2e-16 ***
## chlorides              -0.8580     0.1427  -6.011 1.84e-09 ***
## total.sulfur.dioxide    2.8418     0.2617  10.858  < 2e-16 ***
## density                -7.3712     0.6560 -11.236  < 2e-16 ***
## pH                      0.5555     0.2439   2.277 0.022766 *
## sulphates              -0.2784     0.2046  -1.361 0.173660
## alcohol                -3.2183     0.4152  -7.751 9.15e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7186.90  on 6450  degrees of freedom
## Residual deviance:  316.46  on 6440  degrees of freedom
## AIC: 338.46
##
## Number of Fisher Scoring iterations: 10
```

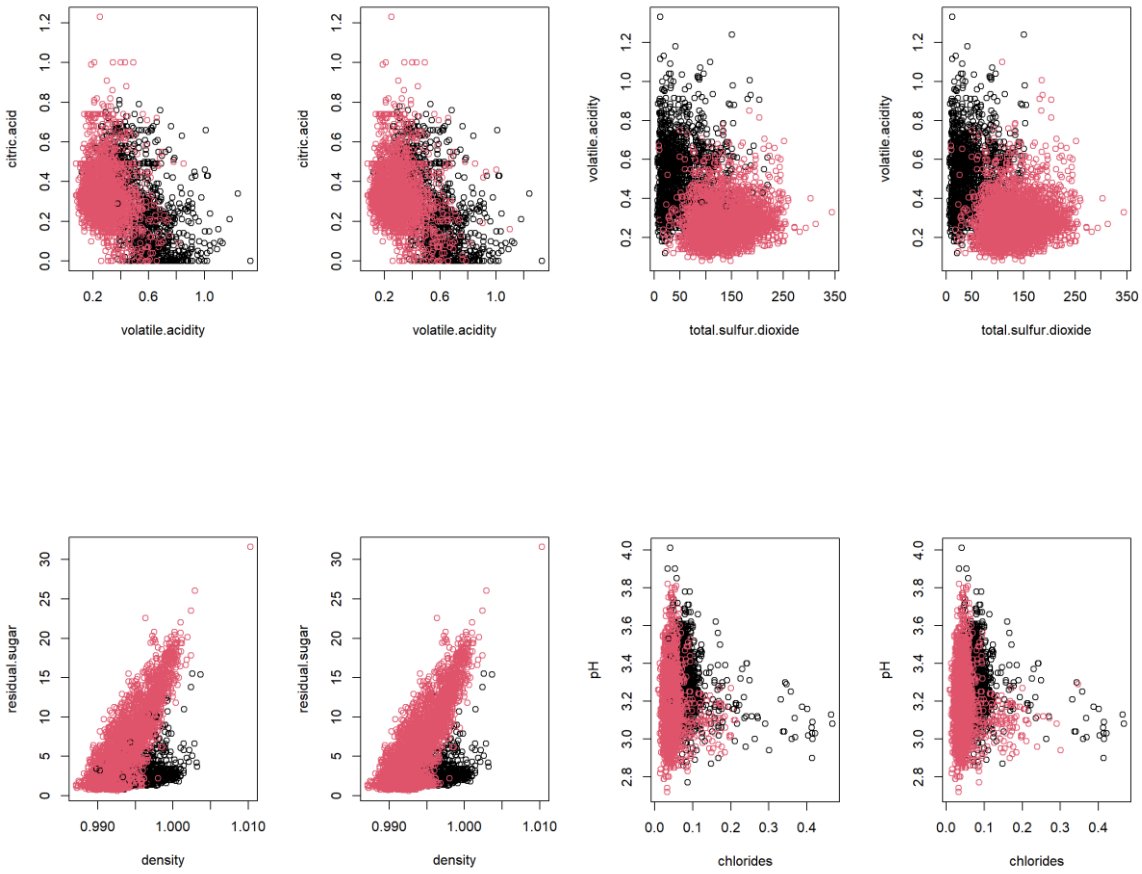
Segons la regressió logística, totes les variables seleccionades són significatives per a explicar la variable resposta type (p-valor < 0.05), amb l'excepció de la variable sulphates (p-valor = 0.17366).

Si observem els coeficients més detingudament podem veure que es variables que més contribueixen són density (-7.37) i residual.sugar (4.75).

```
##          (Intercept)        fixed.acidity     volatile.acidity
##            4.4926562            1.1759120           -0.8163403
##          citric.acid       residual.sugar            chlorides
##            0.5374747            4.7502932           -0.8580403
## total.sulfur.dioxide              density                   pH
##            2.8418168           -7.3712380            0.5554830
##            sulphates              alcohol
##           -0.2784166           -3.2183069
```

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

Es pot comparar si els grups trobats per l'algorisme kmeans corresponen amb els reals per a diferents combinacions de variables.



Visualment s'observa que la majoria dels punts s'ajusten amb la realitat, tot i que n'hi ha uns pocs que no estan al grup corresponent.

Amb la matriu de confusió s'observa com s'han agrupat les observacions.

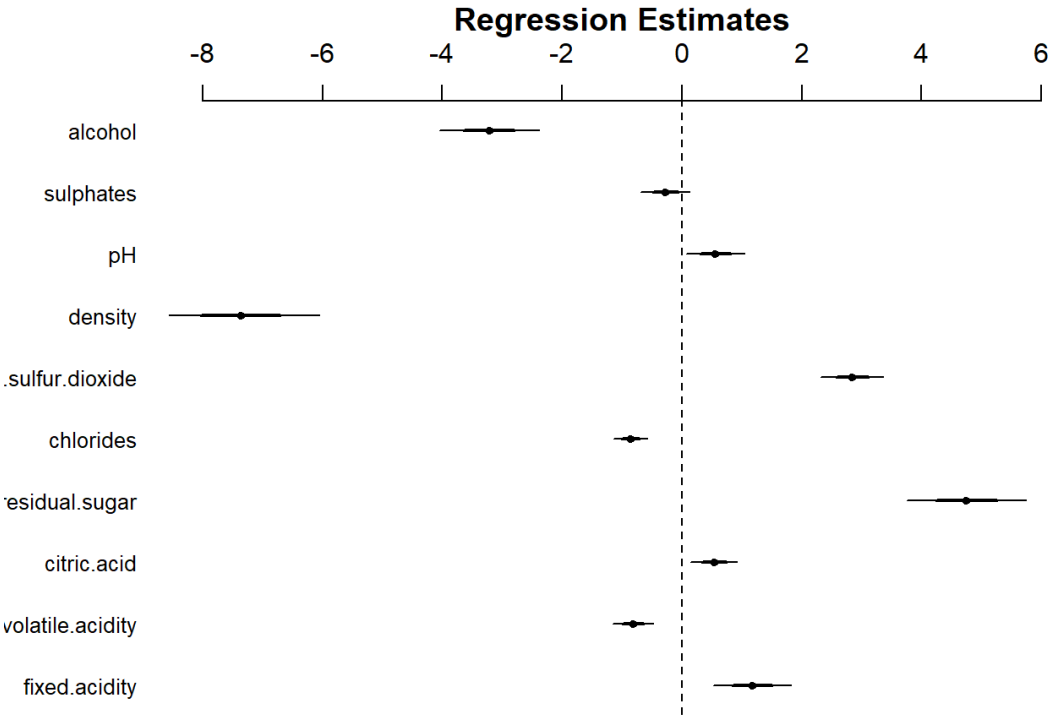
```
   red white
1 1559    60
2   23  4809
```

Això implica una precisió del 98,71% per al model.

Taula de resultats:

```
## Confusion Matrix and Statistics
##
##
##          red white
##   red   1559    60
##   white   23  4809
##
##                Accuracy : 0.9871
##                  95% CI : (0.9841, 0.9897)
##     No Information Rate : 0.7548
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9655
##
##  Mcnemar's Test P-Value : 7.766e-05
##
##             Sensitivity : 0.9855
##             Specificity : 0.9877
##          Pos Pred Value : 0.9629
##          Neg Pred Value : 0.9952
##              Prevalence : 0.2452
##          Detection Rate : 0.2417
##    Detection Prevalence : 0.2510
##       Balanced Accuracy : 0.9866
##
##        'Positive' Class : red
##
```

Pel que fa a la regressió logística els resultats dels coeficients han sigut els següents:



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

Amb l'anàlisi de contrast d'hipòtesis s'ha pogut comprovar com totes les variables tenien mitjanes significativament diferents entre els vins rosats i els vins blancs.

Amb l'anàlisi de clustering s'ha comprovat que el conjunt de dades era ideal per a realitzar dos grups i aquests es podien diferenciar prou clarament, l'algorisme kmeans ha agrupat correctament el 98,71% de les observacions.

Amb la regressió logística s'ha comprovat que totes les variables eren significatives per a explicar el tipus de vi amb l'excepció dels sulphates.

Per tant, amb tot aquest coneixement extret de les dades es pot concloure que les característiques fisicoquímiques usades en aquest dataset són significativament diferents per als vins rosats i els blancs, i aquestes serveixen per poder identificar-los com a un dels dos tipus.

A més, s'ha vist que les variables que més impacte tenen són la density i el residual.sugar.

Així doncs, els resultats han permès resoldre el problema que s'havia plantejat al principi.

7. Codi: Cal adjuntar el codi, preferiblement en R, amb el que s'ha realitzat la neteja, anàlisi i representació de les dades.

La pràctica s'ha realitzat completament amb R, el codi es pot trobar als arxius prac2.html i prac2.rmd dins de la carpeta codi del repositori de github.

Referències:

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

Taula de contribucions:

Contribucions	Firma
Investigació prèvia	A.B.
Redacció de les respostes	A.B.
Desenvolupament codi	A.B.