PRÀCTICA 2
Tipologia i cicle de vida de les dades

L'objectiu d'aquesta activitat serà el tractament d'un dataset, que pot ser el creat a la pràctica 1 o bé qualsevol dataset lliure disponible a Kaggle (https://www.kaggle.com). Alguns exemples de dataset amb els que podeu treballar són: ● Red Wine Quality (https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009). ● Titanic: Machine Learning from Disaster (https://www.kaggle.com/c/titanic). L'últim exemple correspon a una competició activa a Kaggle de manera que, opcionalment, podeu aprofitar el treball realitzat durant la pràctica per entrar en aquesta competició.

Seguint les principals etapes d'un projecte analític, les diferents tasques a realitzar (i justificar) són les següents:

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

Hem escollit dos datasets publicats a Kaggle corresponents a vins negres i blancs. Els dos datasets contenen variables que ens permeten analitzar característiques químiques dels vins, i també trobem una variable que ens qualifica aquests vins del 1 al 10. Aquesta última, a diferencia de les altres, no correspon a una mesura feta per algun instrument, sinó que és la mitja ponderada de les qualificacions fetes per tres experts en degustacions de vi.

La font d'on hem obtingut els datasets és la següent:

https://www.kaggle.com/danielpanizzo/wine-quality

El dataset de vins negres conté 1.599 observacions. El dataset de vins blancs 4.898 observacions.

Data	
red_WINES	1599 obs. of 14 variables
t_wines	6497 obs. of 14 variables
white_WINES	4898 obs. of 14 variables

Les 14 variables que componen els dos datasets són les següents:

```
>  names(red_WINES)
 [1] "id"                "fixed.acidity"     "volatile.acidity"  "citric.acid"
 [5] "residual.sugar"    "chlorides"         "free.sulfur.dioxide" "total.sulfur.dioxide"
 [9] "density"           "pH"                "sulphates"         "alcohol"
[13] "quality"
```

La descripció dels dataset és la següent:

```
>  res <- sapply(red_WINES,class)
>     kable(data.frame(variables=names(res),clase=as.vector(res)))
```

variables	clase
id	integer
fixed.acidity	numeric
volatile.acidity	numeric
citric.acid	numeric
residual.sugar	numeric
chlorides	numeric
free.sulfur.dioxide	numeric
total.sulfur.dioxide	numeric
density	numeric
pH	numeric
sulphates	numeric
alcohol	numeric
quality	integer

On observem que tenim una primera variable que ens identifica el vi (id) i després onze variables quantitatives continues que ens descriuen els atributs químics mesurats per a cada observació. Finalment tenim una variable quantitativa discreta que ens qualifica la puntuació d'aquest vi (quality).

Inicialment havíem plantejat analitzar si existien diferencies de puntuació molt grans entre els dos vins, i si per característiques similars la puntuació entre blancs i negres era molt diferent, però veient la mitjana de les puntuacions d'ambdós datasets on no sembla que existeixi una bretxa molt amplia, hem preferit optar per un altre tipus d'anàlisis que ens permetés treure conclusions més interessants, també enfocant-nos amb la puntuació:

```
>     mean(red_WINES$quality)
[1] 5.6360225
>     mean(white_WINES$quality)
[1] 5.8779094
```

Podem plantejar hipòtesis comparatives (com la que ens havíem plantejat inicialment), hipòtesis relacionals o associatives, o hipòtesis de causalitat. Finalment ens em decantat per plantejar-ne una d'associativa per posteriorment veure les mitjanes de puntuació d'aquestes associacions.

Tan pels vins negres com pels vins blancs, volem veure si existeixen agrupacions de vins amb les mateixes característiques químiques, i quantes agrupacions significativament diferents tenim. Per a cada una d'aquestes agrupacions volem saber la qualificació mitjana que obtenen. Així podrem arribar a la conclusió si en aquest estudi s'han agafat vins de categories molt diferents o s'han escollit vins d'unes categories molt similars que poden portar a qualificacions similars.

La descripció estadística dels dos datasets és la següent:

```
>  summary(red_WINES)
       id         fixed.acidity     volatile.acidity   citric.acid      residual.sugar      chlorides        free.sulfur.dioxide
 Min.   :   1.0   Min.   : 4.6000   Min.   :0.12000   Min.   :0.00000   Min.   : 0.9000   Min.   :0.012000   Min.   : 1.000
 1st Qu.: 400.5   1st Qu.: 7.1000   1st Qu.:0.39000   1st Qu.:0.09000   1st Qu.: 1.9000   1st Qu.:0.070000   1st Qu.: 7.000
 Median : 800.0   Median : 7.9000   Median :0.52000   Median :0.26000   Median : 2.2000   Median :0.079000   Median :14.000
 Mean   : 800.0   Mean   : 8.3196   Mean   :0.52782   Mean   :0.27098   Mean   : 2.5388   Mean   :0.087467   Mean   :15.875
 3rd Qu.:1199.5   3rd Qu.: 9.2000   3rd Qu.:0.64000   3rd Qu.:0.42000   3rd Qu.: 2.6000   3rd Qu.:0.090000   3rd Qu.:21.000
 Max.   :1599.0   Max.   :15.9000   Max.   :1.58000   Max.   :1.00000   Max.   :15.5000   Max.   :0.611000   Max.   :72.000
 total.sulfur.dioxide    density             pH          sulphates         alcohol          quality
 Min.   :  6.000      Min.   :0.99007   Min.   :2.7400   Min.   :0.33000   Min.   : 8.400   Min.   :3.000
 1st Qu.: 22.000      1st Qu.:0.99560   1st Qu.:3.2100   1st Qu.:0.55000   1st Qu.: 9.500   1st Qu.:5.000
 Median : 38.000      Median :0.99675   Median :3.3100   Median :0.62000   Median :10.200   Median :6.000
 Mean   : 46.468      Mean   :0.99675   Mean   :3.3111   Mean   :0.65815   Mean   :10.423   Mean   :5.636
 3rd Qu.: 62.000      3rd Qu.:0.99784   3rd Qu.:3.4000   3rd Qu.:0.73000   3rd Qu.:11.100   3rd Qu.:6.000
 Max.   :289.000      Max.   :1.00369   Max.   :4.0100   Max.   :2.00000   Max.   :14.900   Max.   :8.000
```

```
>     summary(white_WINES)
       id         fixed.acidity     volatile.acidity   citric.acid      residual.sugar     chlorides         free.sulfur.dioxide
 Min.   :   1.0   Min.   : 3.8000   Min.   :0.08000   Min.   :0.00000   Min.   : 0.6000   Min.   :0.009000   Min.   :  2.000
 1st Qu.:1225.2   1st Qu.: 6.3000   1st Qu.:0.21000   1st Qu.:0.27000   1st Qu.: 1.7000   1st Qu.:0.036000   1st Qu.: 23.000
 Median :2449.5   Median : 6.8000   Median :0.26000   Median :0.32000   Median : 5.2000   Median :0.043000   Median : 34.000
 Mean   :2449.5   Mean   : 6.8548   Mean   :0.27824   Mean   :0.33419   Mean   : 6.3914   Mean   :0.045772   Mean   : 35.308
 3rd Qu.:3673.8   3rd Qu.: 7.3000   3rd Qu.:0.32000   3rd Qu.:0.39000   3rd Qu.: 9.9000   3rd Qu.:0.050000   3rd Qu.: 46.000
 Max.   :4898.0   Max.   :14.2000   Max.   :1.10000   Max.   :1.66000   Max.   :65.8000   Max.   :0.346000   Max.   :289.000
 total.sulfur.dioxide    density             pH           sulphates         alcohol         quality
 Min.   :  9.00       Min.   :0.98711   Min.   :2.7200   Min.   :0.22000   Min.   : 8.000   Min.   :3.0000
 1st Qu.:108.00       1st Qu.:0.99172   1st Qu.:3.0900   1st Qu.:0.41000   1st Qu.: 9.500   1st Qu.:5.0000
 Median :134.00       Median :0.99374   Median :3.1800   Median :0.47000   Median :10.400   Median :6.0000
 Mean   :138.36       Mean   :0.99403   Mean   :3.1883   Mean   :0.48985   Mean   :10.514   Mean   :5.8779
 3rd Qu.:167.00       3rd Qu.:0.99610   3rd Qu.:3.2800   3rd Qu.:0.55000   3rd Qu.:11.400   3rd Qu.:6.0000
 Max.   :440.00       Max.   :1.03898   Max.   :3.8200   Max.   :1.08000   Max.   :14.200   Max.   :9.0000
```

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

Com que inicialment analitzarem els dos datasets per separat, ja que el que ens interessa és fer associacions, no integrarem els dos datasets en un de sol de moment.

El que sí que farem inicialment, és separar les variables estadístiques (les que tenen mesures químiques) de les altres, per tal de preparar-les prèviament. És per això que farem una selecció només de les variables quantitatives continues (exclourem el identificador del vi i la seva qualificació).

També farem un escalat d'aquestes variables per normalitzar-les entre si. Escalem les dades per reduir el biaix causat per la combinació de valors mesurats a diferents escales:

```
> red_WINES.scaled <- as.data.frame(scale(red_WINES[,c(2:12)],center=T,scale=T))
> white_WINES.scaled <- as.data.frame(scale(white_WINES[,c(2:12)],center=T,scale=T))
```

Si observem, els valors després d'escalar-los es resumeixen així:

```
>   summary(red_WINES.scaled)
 fixed.acidity      volatile.acidity     citric.acid       residual.sugar       chlorides         free.sulfur.dioxide total.sulfur.dioxide
 Min.   :-2.13638   Min.   :-2.277567   Min.   :-1.391037   Min.   :-1.162333   Min.   :-1.603443   Min.   :-1.42206    Min.   :-1.23020
 1st Qu.:-0.70050   1st Qu.:-0.769690   1st Qu.:-0.929028   1st Qu.:-0.453077   1st Qu.:-0.371113   1st Qu.:-0.84845    1st Qu.:-0.74381
 Median :-0.24102   Median :-0.043675   Median :-0.056343   Median :-0.240300   Median :-0.179889   Median :-0.17924    Median :-0.25742
 Mean   : 0.00000   Mean   : 0.000000   Mean   : 0.000000   Mean   : 0.000000   Mean   : 0.000000   Mean   : 0.00000    Mean   : 0.00000
 3rd Qu.: 0.50564   3rd Qu.: 0.626492   3rd Qu.: 0.765008   3rd Qu.: 0.043403   3rd Qu.: 0.053829   3rd Qu.: 0.48996    3rd Qu.: 0.47217
 Max.   : 4.35379   Max.   : 5.876138   Max.   : 3.742403   Max.   : 9.192806   Max.   :11.123555   Max.   : 5.36561    Max.   : 7.37285
    density              pH            sulphates          alcohol
 Min.   :-3.5376247   Min.   :-3.6992439   Min.   :-1.93590   Min.   :-1.89832
 1st Qu.:-0.6075656   1st Qu.:-0.6549356   1st Qu.:-0.63802   1st Qu.:-0.86611
 Median : 0.0017595   Median :-0.0072104   Median :-0.22506   Median :-0.20924
 Mean   : 0.0000000   Mean   : 0.0000000   Mean   : 0.00000   Mean   : 0.00000
 3rd Qu.: 0.5766445   3rd Qu.: 0.5757422   3rd Qu.: 0.42388   3rd Qu.: 0.63530
 Max.   : 3.6789042   Max.   : 4.5268658   Max.   : 7.91620   Max.   : 4.20114
```

```
> summary(white_WINES.scaled)
 fixed.acidity      volatile.acidity    citric.acid       residual.sugar       chlorides         free.sulfur.dioxide total.sulfur.dioxide
 Min.   :-3.619982   Min.   :-1.96678   Min.   :-2.76146   Min.   :-1.14183   Min.   :-1.68310   Min.   :-1.958477   Min.   :-3.04392
 1st Qu.:-0.657434   1st Qu.:-0.67703   1st Qu.:-0.53042   1st Qu.:-0.92495   1st Qu.:-0.44729   1st Qu.:-0.723701   1st Qu.:-0.71440
 Median :-0.064924   Median :-0.18097   Median :-0.11727   Median :-0.23490   Median :-0.12689   Median :-0.076914   Median :-0.10261
 Mean   : 0.000000   Mean   : 0.00000   Mean   : 0.00000   Mean   : 0.00000   Mean   : 0.00000   Mean   : 0.000000   Mean   : 0.00000
 3rd Qu.: 0.527585   3rd Qu.: 0.41430   3rd Qu.: 0.46115   3rd Qu.: 0.69175   3rd Qu.: 0.19350   3rd Qu.: 0.628672   3rd Qu.: 0.67390
 Max.   : 8.704217   Max.   : 8.15281   Max.   :10.95530   Max.   :11.71292   Max.   :13.74167   Max.   :14.916791   Max.   : 7.09772
    density              pH            sulphates          alcohol
 Min.   :-2.312802   Min.   :-3.101091   Min.   :-2.36447   Min.   :-2.043089
 1st Qu.:-0.770628   1st Qu.:-0.650770   1st Qu.:-0.69964   1st Qu.:-0.824192
 Median :-0.096083   Median :-0.054746   Median :-0.17390   Median :-0.092853
 Mean   : 0.000000   Mean   : 0.000000   Mean   : 0.00000   Mean   : 0.000000
 3rd Qu.: 0.692975   3rd Qu.: 0.607503   3rd Qu.: 0.52708   3rd Qu.: 0.719745
 Max.   :15.029763   Max.   : 4.183648   Max.   : 5.17107   Max.   : 2.995020
```

Guardarem els identificadors i la variable de qualitat, ja que posteriorment ens interessa tornar-ho a relacionar.

3. **Neteja de les dades**

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

En el cas que ens ocupa no tenim dades amb valors zeros o buits (NA). Ho hem comprovat de la següent forma:

-Vins blancs:

```
> sapply(white_WINES.scaled, function(x) sum(is.na(x)))
       fixed.acidity     volatile.acidity          citric.acid       residual.sugar            chlorides  free.sulfur.dioxide
                   0                    0                    0                    0                    0                    0
total.sulfur.dioxide              density                   pH            sulphates              alcohol
                   0                    0                    0                    0                    0
> summarise_all(white_WINES.scaled, funs(sum(is.na(.))))
  fixed.acidity volatile.acidity citric.acid residual.sugar chlorides free.sulfur.dioxide total.sulfur.dioxide density pH sulphates alcohol
1             0                0           0              0         0                   0                     0       0  0         0       0
```

-Vins Negres:

```
> sapply(red_WINES.scaled, function(x) sum(is.na(x)))
       fixed.acidity     volatile.acidity          citric.acid       residual.sugar            chlorides  free.sulfur.dioxide
                   0                    0                    0                    0                    0                    0
total.sulfur.dioxide              density                   pH            sulphates              alcohol
                   0                    0                    0                    0                    0
> summarise_all(red_WINES.scaled, funs(sum(is.na(.))))
  fixed.acidity volatile.acidity citric.acid residual.sugar chlorides free.sulfur.dioxide total.sulfur.dioxide density pH sulphates alcohol
1             0                0           0              0         0                   0                     0       0  0         0       0
```

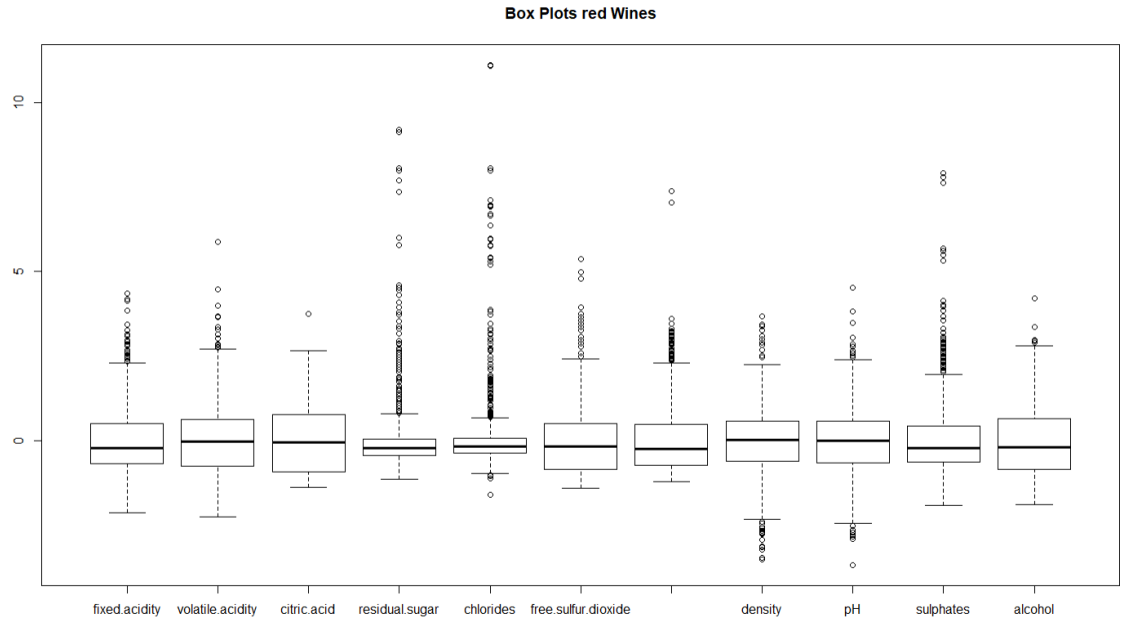
En el cas que n'haguéssim trobat, tindríem dues opcions, mirar de deduir-los per aproximació, amb algoritmes com el KNN() o descartar les observacions. En el cas que ens ocupa, hauríem optar per descartar aquelles observacions, ja que reduïm el risc de que aquestes aproximacions ens emetin soroll o no siguin prou acurades per al anàlisis que volem fer.

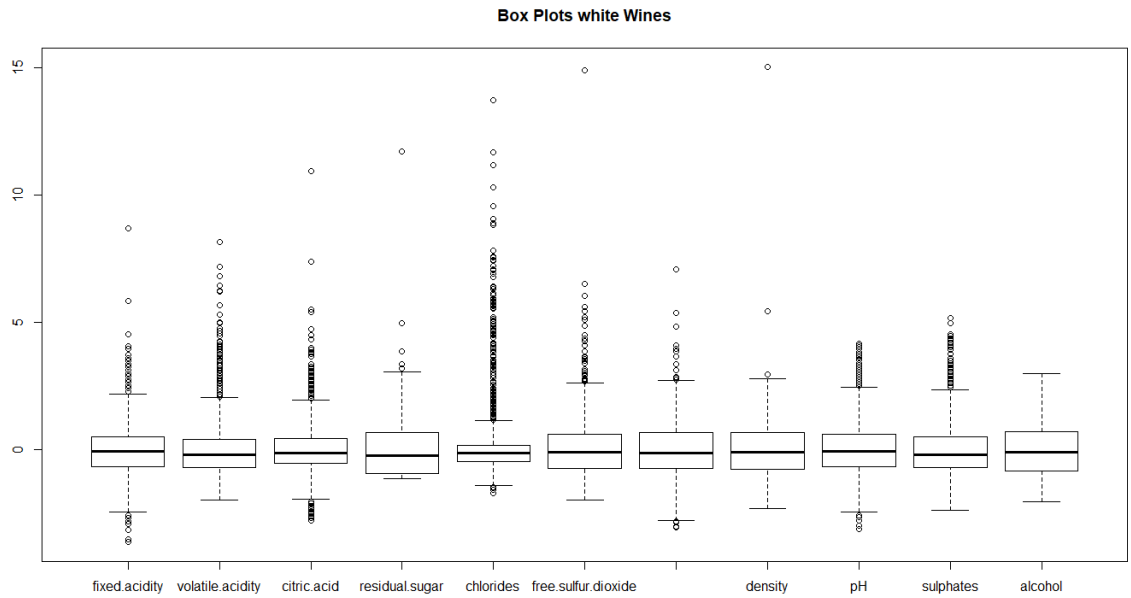
De fet és el que farem més endavant, ja que després de retirar els outliers haurem d'eliminar els valors NA amb els quals s'hauran substituït els outliers, de manera que aquest punt el tractarem en el següent apartat.

3.2 Identificació i tractament de valors extrems.

Es considera un valor outlier o extrem, aquell que s'allunya tres desviacions Standard de la mitjana. Si fem la representació gràfica de com es distribueixen els valors de les nostres variables, veurem que els valors es poden dibuixar en 4 quartils. El segon quartil es el que coneixem com a mitjana i la distància entre el primer quartil i el tercer és el que anomenem IQR i mesura que gràficament veiem com una caixa en el diagrama següent.

Els valors que considerarem outliers en aquesta gràfica, serien els que siguin 1,5 vegades més grans que la distància de la caixa sumada al tercer quartil, i també els que siguin molt petits, és a dir, més petits del primer quartil menys la distància de la caixa. De fet, gràficament de manera automàtica ja ens ho representa així:





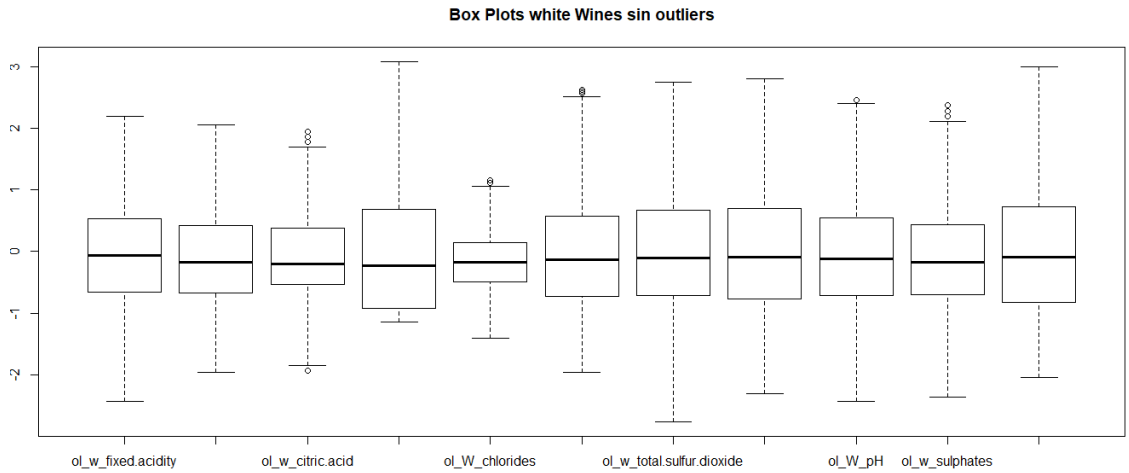
Ara es tracta d'apartar-los de la mostra a tractar. Ho farem de la següent forma:

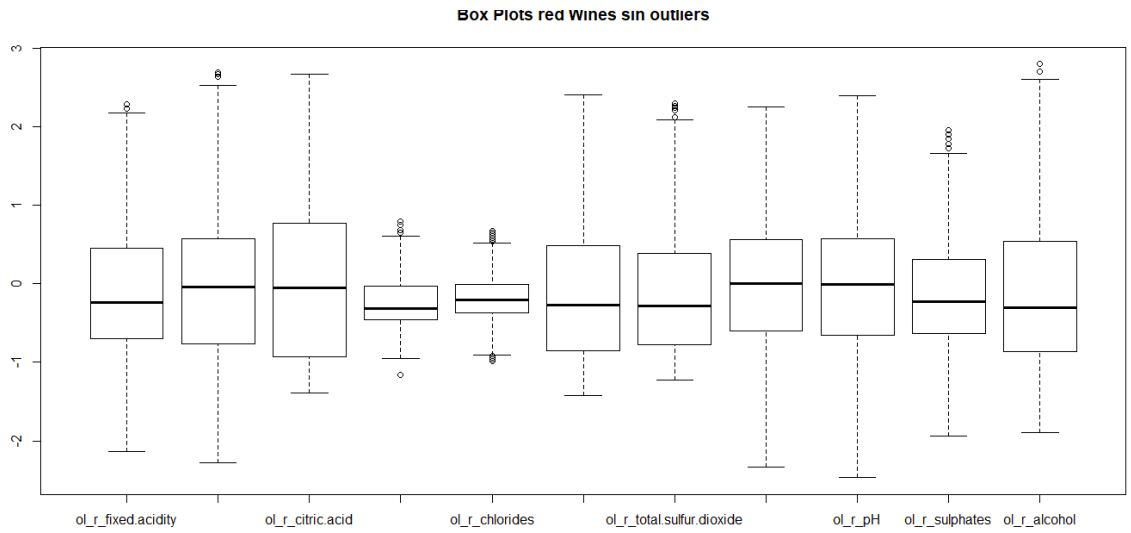
1.-Construïrem en R una funció que ens permeti aplicar la teoria explicada prèviament:

```
>   OutliersReplace <- function(x, na.rm = TRUE, ...)
+   {
+     qnt <- quantile(x, probs=c(.25, .75), na.rm = na.rm, ...)
+     H <- 1.5 * IQR(x, na.rm = na.rm)
+     y <- x
+     y[x < (qnt[1] - H)] <- NA
+     y[x > (qnt[2] + H)] <- NA
+     y
+   }
```

2.-Retirarem per a cada variable els outliers (deixarem el valor NA)

3.-Gráficament el resultat és el següent per vins blancs i vins negres:





Ara sí que si observem les dades resultants, veurem que existeixen valors buits (NA).

Abans de tractar les dades per a fer les associacions, procedirem a tractar-ho, retirant-los, però abans preservant la identificació de cada observació i la seva qualificació:

```
>  sapply(white_WINES.ol, function(x) sum(is.na(x)))
      ol_w_fixed.acidity     ol_w_volatile.acidity          ol_w_citric.acid       ol_w_residual.sugar
                     137                       186                       270                         7
          ol_w_chlorides  ol_w_free.sulfur.dioxide ol_w_total.sulfur.dioxide              ol_w_density
                     208                        50                        19                         5
                ol_w_pH            ol_w_sulphates               ol_w_alcohol
                      75                       124                         0
>   summarise_all(white_WINES.ol, funs(sum(is.na(.))))
  ol_w_fixed.acidity ol_w_volatile.acidity ol_w_citric.acid ol_w_residual.sugar ol_w_chlorides
1                137                  186                270                  7             208
  ol_w_free.sulfur.dioxide ol_w_total.sulfur.dioxide ol_w_density ol_w_pH ol_w_sulphates ol_w_alcohol
1                       50                        19           5      75            124            0

>   sapply(red_WINES.ol, function(x) sum(is.na(x)))
      ol_r_fixed.acidity     ol_r_volatile.acidity          ol_r_citric.acid       ol_r_residual.sugar
                     49                        19                         1                       155
          ol_r_chlorides  ol_r_free.sulfur.dioxide ol_r_total.sulfur.dioxide              ol_r_density
                     112                        33                        55                        45
                ol_r_pH            ol_r_sulphates               ol_r_alcohol
                      35                        59                        14
>   summarise_all(red_WINES.ol, funs(sum(is.na(.))))
  ol_r_fixed.acidity ol_r_volatile.acidity ol_r_citric.acid ol_r_residual.sugar ol_r_chlorides
1                 49                   19                1                155            112
  ol_r_free.sulfur.dioxide ol_r_total.sulfur.dioxide ol_r_density ol_r_pH ol_r_sulphates ol_r_alcohol
1                       33                        55           45      35            59           14
```

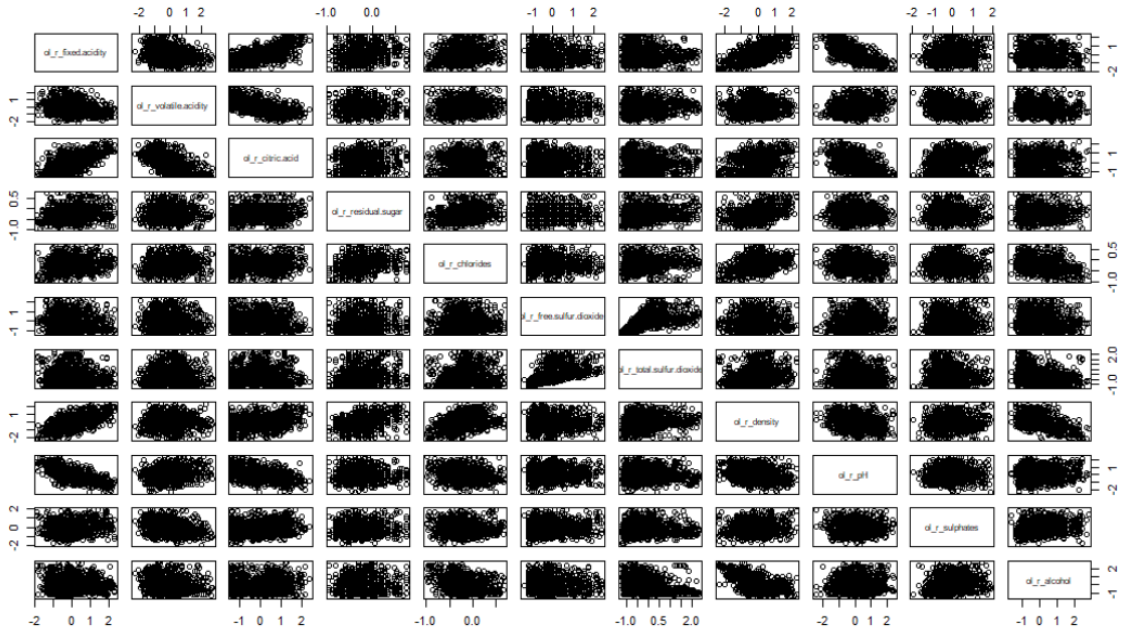
Podem concloure que tots els NA que tenim ara mateix corresponen a outliers que hem retirat. Un total de 1.081 en cas dels vins blancs i un total de 577 en el cas de les variables dels vins negres.

4. Anàlisi de les dades.

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

Abans de plantejar la hipòtesis d'investigació, mirem si algunes variables tenen relació entre elles:

```
>     plot(ds_red[2:12])
```



Veiem que son varies les variables que tenen una correlació forta entre elles:

```
>     cor(ds_red[2:12])
                      ol_r_fixed.acidity ol_r_volatile.acidity ol_r_citric.acid ol_r_residual.sugar ol_r_chlorides ol_r_free.sulfur.dioxide
ol_r_fixed.acidity             1.000000000          -0.274450181     0.6621271586         0.232342956    0.197320230           -0.152920810
ol_r_volatile.acidity         -0.274450181           1.000000000    -0.6282508006         0.026423048    0.113115163           -0.015271188
ol_r_citric.acid               0.662127159          -0.628250801     1.0000000000         0.155314728    0.071410678           -0.074085863
ol_r_residual.sugar            0.232342956           0.026423048     0.1553147281         1.000000000    0.233660421            0.087420735
ol_r_chlorides                 0.197320230           0.113115163     0.0714106783         0.233660421    1.000000000            0.015757606
ol_r_free.sulfur.dioxide      -0.152920810          -0.015271188    -0.0740858634         0.087420735    0.015757606            1.000000000
ol_r_total.sulfur.dioxide     -0.090876941           0.103209390    -0.0024030033         0.193564333    0.174883423            0.620646789
ol_r_density                   0.609222245           0.041950080     0.3047938173         0.393356703    0.412953450           -0.020669890
ol_r_pH                       -0.684399128           0.226737204    -0.4746154253        -0.059887155   -0.174586212            0.150897711
ol_r_sulphates                 0.160878381          -0.317619414     0.2575498706         0.044695479   -0.081529867            0.104767743
ol_r_alcohol                  -0.040956390          -0.220691137     0.1397327293         0.099914160   -0.304717855           -0.024239204
                      ol_r_total.sulfur.dioxide ol_r_density      ol_r_pH ol_r_sulphates ol_r_alcohol
ol_r_fixed.acidity              -0.0908769413   0.609222245 -0.684399128    0.160878381 -0.040956390
ol_r_volatile.acidity            0.1032093899   0.041950080  0.226737204   -0.317619414 -0.220691137
ol_r_citric.acid                -0.0024030033   0.304793817 -0.474615425    0.257549871  0.139732729
ol_r_residual.sugar              0.1935643328   0.393356703 -0.059887155    0.044695479  0.099914160
ol_r_chlorides                   0.1748834232   0.412953450 -0.174586212   -0.081529867 -0.304717855
ol_r_free.sulfur.dioxide         0.6206467893  -0.020669890  0.150897711    0.104767743 -0.024239204
ol_r_total.sulfur.dioxide        1.0000000000   0.148973376  0.015903463   -0.051669363 -0.244727335
ol_r_density                     0.1489733764   1.000000000 -0.226380266    0.070218663 -0.545420325
ol_r_pH                          0.0159034633  -0.226380266  1.000000000    0.012832698  0.122635407
ol_r_sulphates                  -0.0516693634   0.070218663  0.012832698    1.000000000  0.272737352
ol_r_alcohol                    -0.2447273354  -0.545420325  0.122635407    0.272737352  1.000000000
```

Un estudi interessant, seria analitzar les relacions que tenen les diverses variables entre elles. Sabent si podem aplicar proves parmètriques o no paramètriques podríem treure conclusions molt interessants.

Ara mateix, però, el que ens interessa és aplicar tres algoritmes de clustering sobre els vins negres i també sobre els vins blancs, per obtenir-ne una classificació. Usarem posteriorment aquesta classificació per calcular les nota mitja de cada agrupació i veure si veiem diferencies molt significatives entre elles.

Per poder fer això, és important no perdre la traçabilitat del identificador de cada vi i la seva qualificació. És per això que en el procés de retirada dels NA (que hem obtingut després de treure els outliers), siguem curosos en mantenir aquestes dues variables. Ho aconseguirem fent els següents passos (tan pels vins blancs com pels negres):

```
#VINS NEGRES
#Incorporem els id i les puntuacions dels vins en un únic dataframe
df_red <- data.frame(red_WINES$id,red_WINES.ol, red_WINES$quality)

#Preparem el dataset sense outliers
ds_red <- filter(df_red,
                 !is.na(ol_r_fixed.acidity),
                 !is.na(ol_r_volatile.acidity),
                 !is.na(ol_r_citric.acid),
                 !is.na(ol_r_residual.sugar),
                 !is.na(ol_r_chlorides),
                 !is.na(ol_r_free.sulfur.dioxide),
                 !is.na(ol_r_total.sulfur.dioxide),
                 !is.na(ol_r_density),
                 !is.na(ol_r_pH),
                 !is.na(ol_r_sulphates),
                 !is.na(ol_r_alcohol)
                )
```

Finalment obtindrem dos datasets amb menys observacions.

Vins blancs, passem de les 4.898 observacions a 4.002 observacions i en vins negres passem de tenir 1.599 observacions a tenir-ne 1.192 després d'aquesta retirada de NA:

Data	
df_red	1599 obs. of 13 variables
df_white	4898 obs. of 13 variables
ds_red	1192 obs. of 13 variables
ds_white	4002 obs. of 13 variables

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

Si optéssim per plantejar una regressió, una comparació entres dues variables, o una correlació, ens caldria saber si gaudeixen d'una distribució normal o si tenen homogeneïtat en la variància. Aquest tipus d'informació ens ha de permetre escollir millor quines proves estadístiques aplicariem en el següent pas (sobretot si haguéssim d'escollir entre proves paramètriques i proves no paramètriques a l'hora de fer contrastos d'hipòtesis) i quines no ens donarien un resultat fiable en cas d'aplicar-les. Així doncs:

-Comprovem la normalitat:

Per fer-ho existeixen dos mètodes habituals, el de kolmogorov-Smirnov i el de Shapiro-Wilk. Els dos fan una comparativa amb una distribució normal. Com que la hipòtesis nul·la és que la distribució gaudeix de normalitat, en cas d'obtenir un p-valor superior a 0,05 acceptarem que les dades tenen una distribució normal. Així doncs, per a cada una de les nostres variables:

```
>   shapiro.test(ds_white$ol_w_free.sulfur.dioxide)

        Shapiro-Wilk normality test

data:  ds_white$ol_w_free.sulfur.dioxide
W = 0.988314, p-value < 2.22e-16

>   shapiro.test(ds_white$ol_w_total.sulfur.dioxide)

        Shapiro-Wilk normality test

data:  ds_white$ol_w_total.sulfur.dioxide
W = 0.988516, p-value < 2.22e-16

>   shapiro.test(ds_white$ol_w_density)

        Shapiro-Wilk normality test

data:  ds_white$ol_w_density
W = 0.976238, p-value < 2.22e-16

>   shapiro.test(ds_white$ol_w_pH)

        Shapiro-Wilk normality test

data:  ds_white$ol_w_pH
W = 0.994285, p-value = 1.7994e-11

>   shapiro.test(ds_white$ol_w_sulphates)

        Shapiro-Wilk normality test

data:  ds_white$ol_w_sulphates
W = 0.981591, p-value < 2.22e-16

>   shapiro.test(ds_white$ol_w_alcohol)

        Shapiro-Wilk normality test

data:  ds_white$ol_w_alcohol
W = 0.961406, p-value < 2.22e-16
```

```
>   shapiro.test(ds_white$ol_w_fixed.acidity)

        Shapiro-Wilk normality test

data:  ds_white$ol_w_fixed.acidity
W = 0.994367, p-value = 2.3201e-11

>   shapiro.test(ds_white$ol_w_volatile.acidity)

        Shapiro-Wilk normality test

data:  ds_white$ol_w_volatile.acidity
W = 0.983717, p-value < 2.22e-16

>   shapiro.test(ds_white$ol_w_citric.acid)

        Shapiro-Wilk normality test

data:  ds_white$ol_w_citric.acid
W = 0.976361, p-value < 2.22e-16

>   shapiro.test(ds_white$ol_w_residual.sugar)

        Shapiro-Wilk normality test

data:  ds_white$ol_w_residual.sugar
W = 0.897853, p-value < 2.22e-16

>   shapiro.test(ds_white$ol_w_chlorides)

        Shapiro-Wilk normality test

data:  ds_white$ol_w_chlorides
W = 0.996856, p-value = 2.0626e-07
```

Segons la prova de shapiro-wilk, cap de les variables compleix una distribució normal. Tot i que segons el teorema central del límit, al tenir una mostra molt significativa, podríem arribar a concloure que les dades segueixen una distribució normal. Essent conservadors, pensarem que no és així, de manera que la prova d'homogeneïtat la farem emprant mètodes no paramètrics.

-Comprovem la homogeneïtat de la variància:

Per fer-ho usarem Fligner-Killeen, ja que suposem que no tenim distribucions normals. Com que la homogeneïtat de variància es comprova amb agrupacions diferents, encara que no l'emprem en el cas del nostre estudi, ja que no barrejarem els vins negres amb els blancs, si que per veure'n el funcionament compararem diverses variables de les observacions de vi blanc amb les de vins negres. Per fer-ho primer crearem un dataset que uneixi els dos conjunts i posteriorment aplicarem la proba en diverses variables:

```
>   a<- total_wines[total_wines$type==1, "fixed.acidity"]
>    b<- total_wines[total_wines$type==0, "fixed.acidity"]
>    fligner.test(x= list(a,b))

        Fligner-Killeen test of homogeneity of variances

data:  list(a, b)
Fligner-Killeen:med chi-squared = 751.508, df = 1, p-value < 2.22e-16
```

```
>   total_wines <- rbind(aux_red,aux_white)
>   a<- total_wines[total_wines$type==1, "alcohol"]
>   b<- total_wines[total_wines$type==0, "alcohol"]
>   fligner.test(x= list(a,b))

        Fligner-Killeen test of homogeneity of variances

data:  list(a, b)
Fligner-Killeen:med chi-squared = 78.1504, df = 1, p-value < 2.22e-16
```

```
>   a<- total_wines[total_wines$type==1, "pH"]
>   b<- total_wines[total_wines$type==0, "pH"]
>   fligner.test(x= list(a,b))

        Fligner-Killeen test of homogeneity of variances

data:  list(a, b)
Fligner-Killeen:med chi-squared = 0.412195, df = 1, p-value = 0.52086
```

De les tres variables analitzades, veiem que entre els vins blancs i negres, només el PHP presenta homogeneïtat de variàncies entre ambdues poblacions. Les altres dues, alcohol i acidesa fixada no en presenten.

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.

El nostre objectiu és obtenir els grups òptims de vins a partir de les seves característiques químiques. Un cop aconseguit, comparar les notes mitjanes que obtenen aquests grups entre ells per veure si existeixen vins amb categories molt diferents (subjectivament parlant).

Ho farem amb vins negres per comparar les notes mitjanes de les categories entre elles. En funció de les conclusions podríem ampliar l'estudi a la mostra de vins blancs que ja tenim preparada també.

Utilitzarem tres algoritmes de clustering i seleccionarem el que ens agrupi millor els vins.

El primer pas serà escollir el nombre òptim de grups en que volem particionar les observacions:

```
# El nombre òptim de clusters
cluster.optim <- fviz_nbclust(ds_red[2:12], hcut, method = "gap_stat")
cluster.optim.gap <- cluster.optim$data$gap
cluster.optim.gap == max(cluster.optim.gap)
k <- 10 # nombre òptim de clusters
```

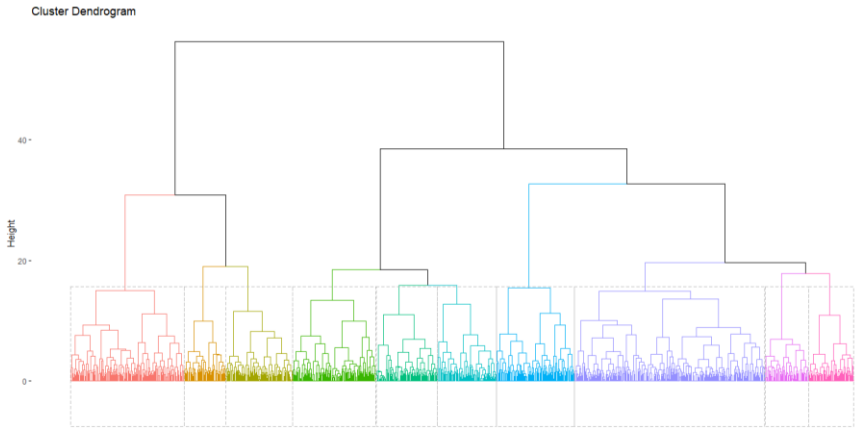
Emprarem els següents tres mètodes de classificació: Jeràrquic, K-means i el GMM

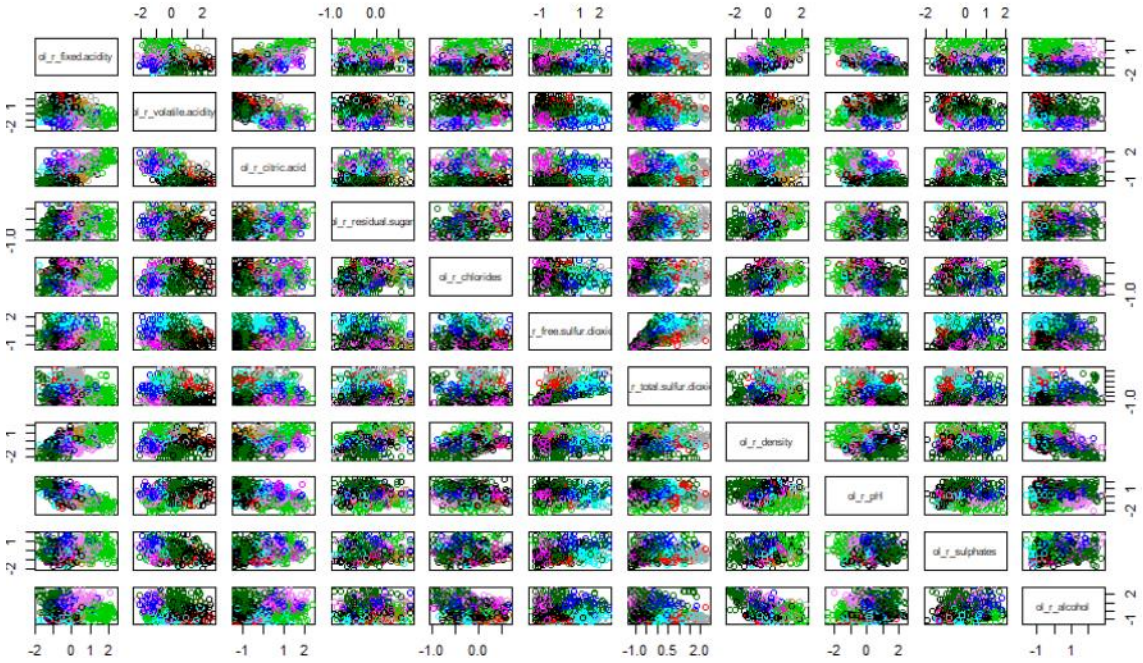
Jeràrquic:

```
# Hierarchical
res.hc <- eclust(ds_red[2:12], "hclust", k = k, graph = FALSE)
fviz_dend(res.hc, rect = TRUE, show_labels = FALSE)

pairs(ds_red[2:12],col=res.hc$cluster,pch = 21)
res.hc.clusters <- res.hc$cluster
eclust(ds_red[2:12], "hclust", k = k, graph = TRUE)
```

Els gràfics resultants per ordre són:





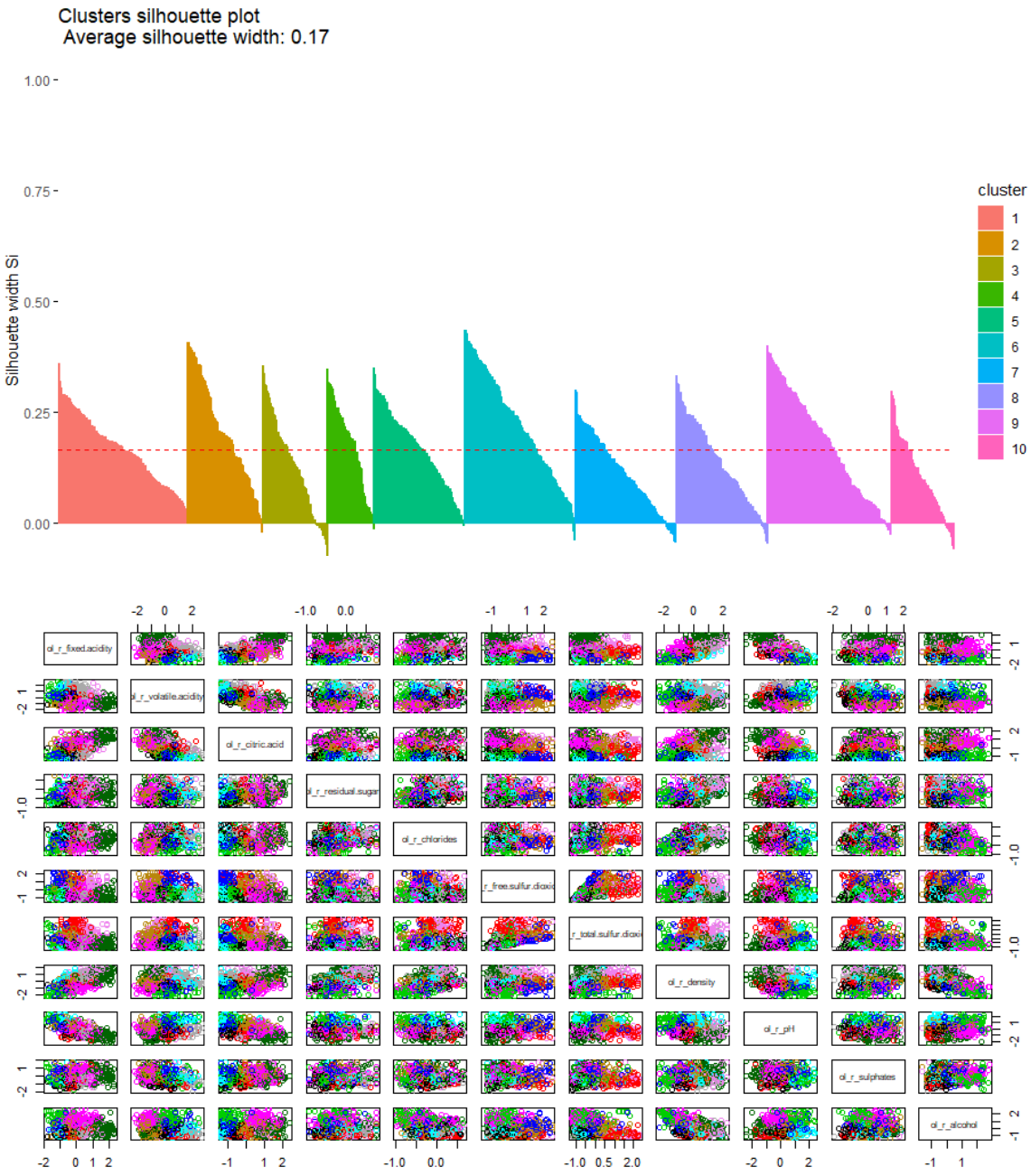
K-Means:

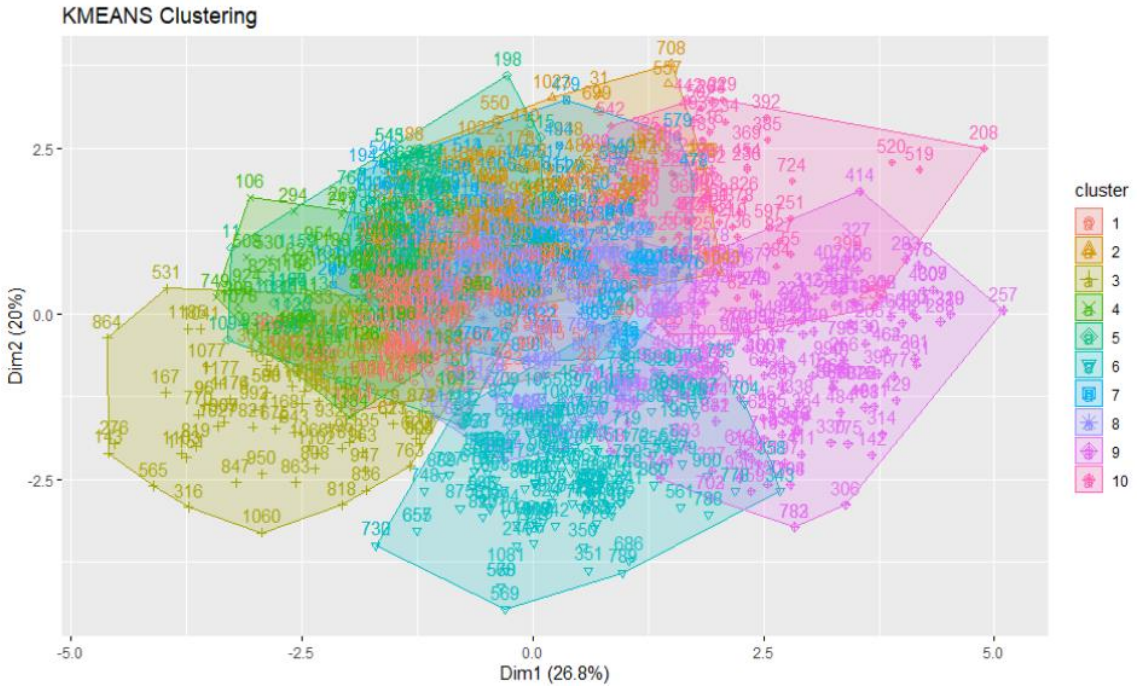
Aplicarem el següent mètode de classificació

```
# k-means
km.res <- eclust(ds_red[2:12], "kmeans", k = k,nstart = 25, graph = FALSE)
fviz_silhouette(km.res)

pairs(ds_red[2:12],col=km.res$cluster,pch = 21)
res.km.clusters <- km.res$cluster
eclust(ds_red[2:12], "kmeans", k = k, nstart = 25, graph = TRUE)
```

Les gràfiques per ordre d'execució són les següents:


Clusters silhouette plot
Average silhouette width: 0.17
Silhouette width Si
1.00
0.75
0.50
0.25
0.00
cluster
1
2
3
4
5
6
7
8
9
10
sl_r_fixed.acidity
sl_r_volatile.acidity
sl_r_citric.acid
sl_r_residual.sugar
sl_r_chlorides
r_free.sulfur.dioxi
r_total.sulfur.dioxi
sl_r_density
sl_r_pH
sl_r_sulphates
sl_r_alcohol



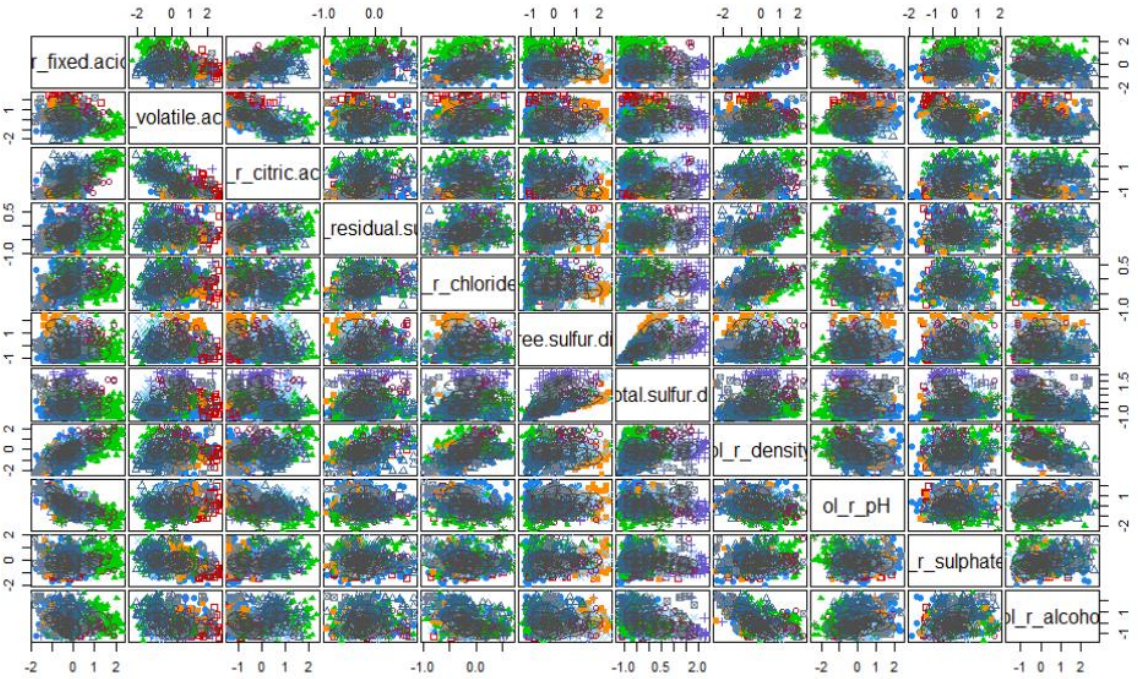
GMM

Finalment aplicarem un tercer mètode:

```
# GMM
gmm.res = Mclust(ds_red[2:12], G = k)
summary(gmm.res, parameters = TRUE)

plot(gmm.res, what = "classification")
res.gmm.clusters <- gmm.res$classification
```

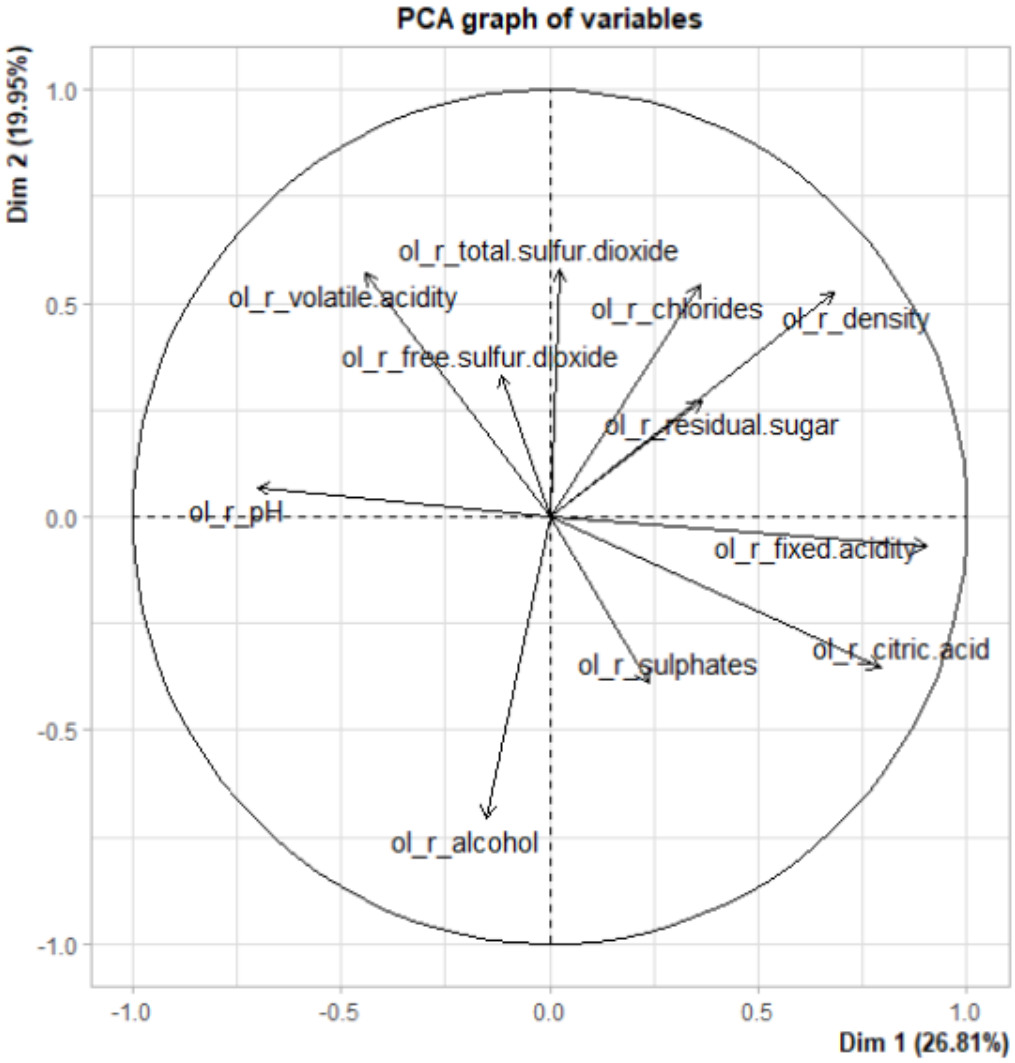
Les gràfiques obtingudes:



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

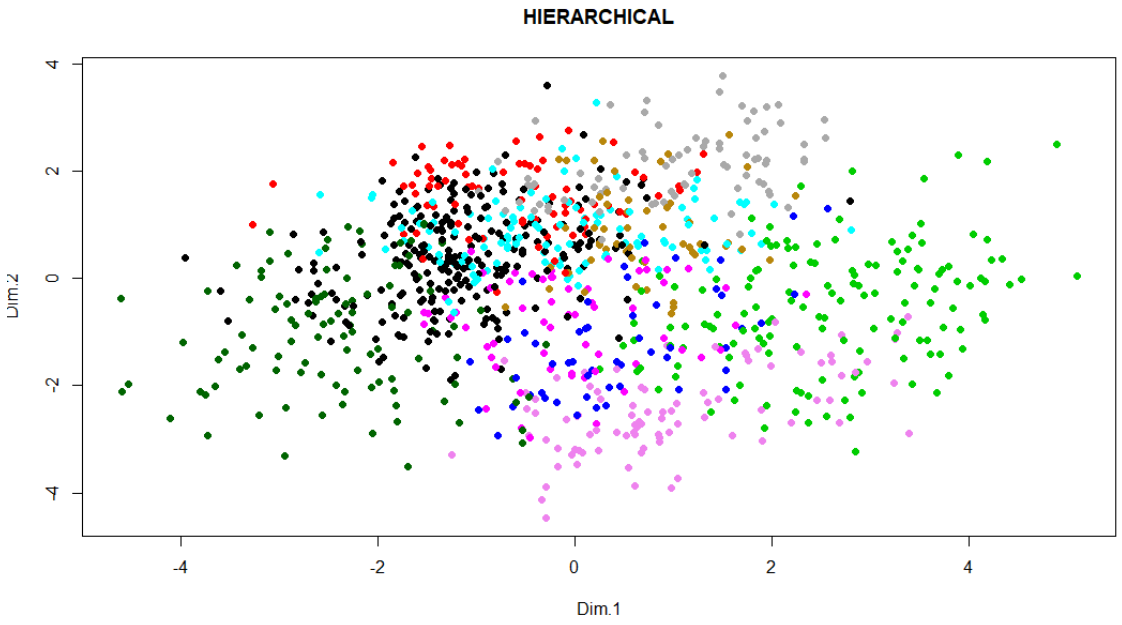
A continuació representem gràficament els resultats obtinguts de classificació a partir dels tres mètodes:

```
#Dedidirem quin mètode ens ha classificat millor les observacions:
pca <- PCA(ds_red[2:12],graph = TRUE)
```



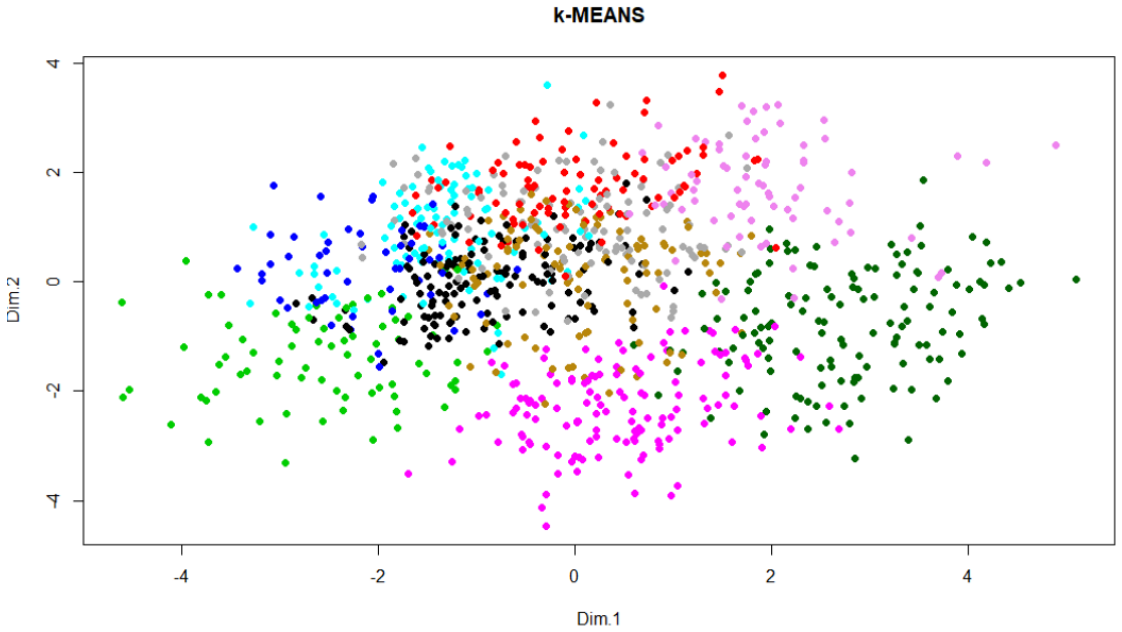
-Jerarquic:

```
plot(pca$ind$coord[,1:2], pch=16, col=res.hc.clusters,
     main="HIERARCHICAL")
```



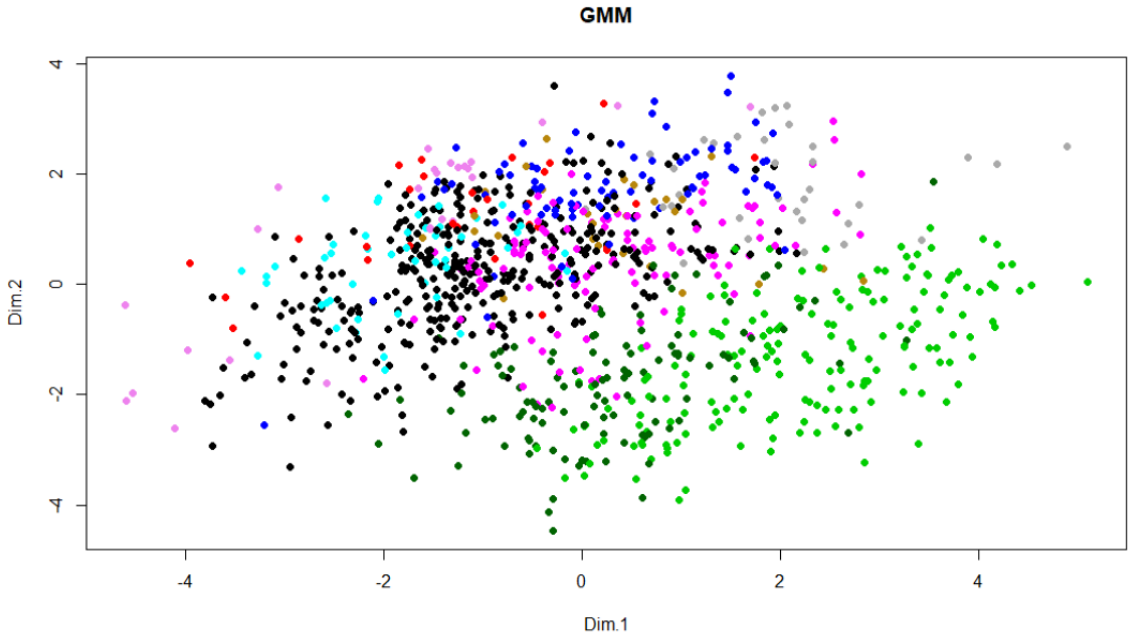
-K-MEANS:

```
plot(pca$ind$coord[,1:2], pch=16, col=res.km.clusters,
     main="k-MEANS")
```

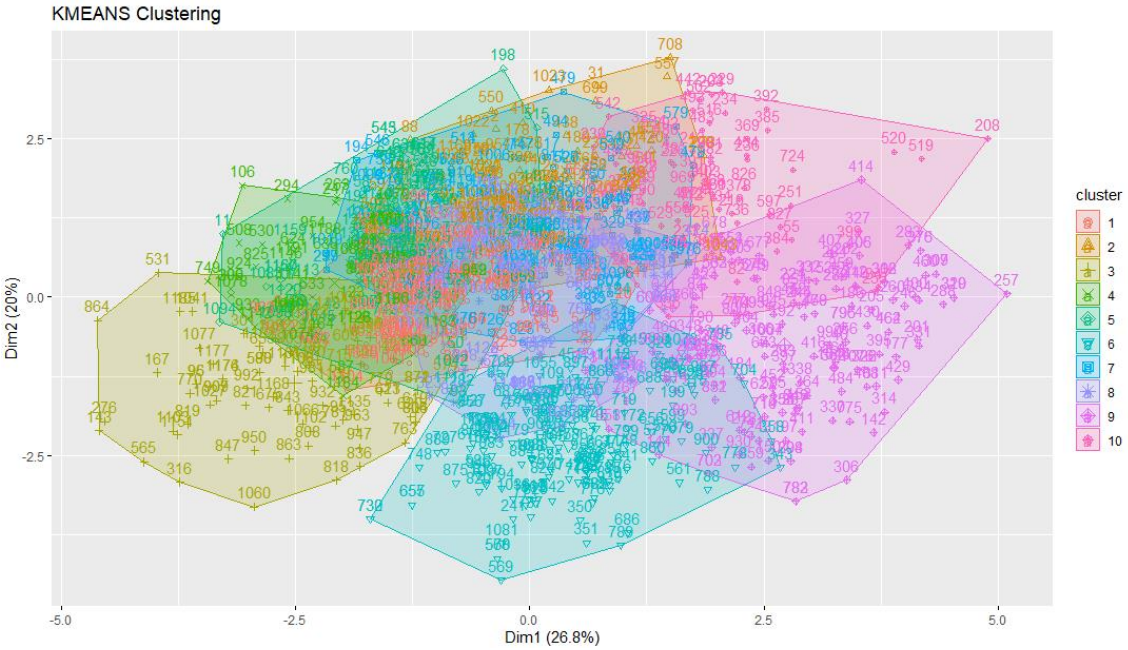


-GMM

```
plot(pca$ind$coord[,1:2], pch=16, col=res.gmm.clusters,
     main="GMM")
```



Conclusió: Veient les gràfiques, totes presenten problemes per mostrar una divisió clara sense solapaments, però malgrat tot, visualment algunes diferencies són evidents, ja que en el cas del K-Means presenta una àrees molt més diferenciades i amb menys solapaments. Així doncs, usarem el algoritme de classificació K-Means per a classificar el nostre dataset de vins negres i poder extraure les conclusions que cerquem:



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

Per tal de poder treure les conclusions, ens cal crear un dataframe amb els següents atributs: identificador del vi, clúster al qual pertany (segons k-means) i nota de qualitat que ha obtingut:

```
#Aqui tenim els resultats de la classificació
res.hc.clusters
res.km.clusters
res.gmm.clusters

#creem el fitxer de sortida
ds_resultat <- data.frame(ds_red$red_WINES.id, res.km.clusters, ds_red$red_WINES.quality)

setnames(ds_resultat, "ds_red.red_WINES.id","id")
setnames(ds_resultat, "res.km.clusters", "Cluster")
setnames(ds_resultat, "ds_red.red_WINES.quality", "Quality")
```

```
>   #Nota mitja total
>   mean(ds_resultat$Quality)
[1] 5.6384228
>
>   #nota mitja Cluster=1
>   mean(filter(ds_resultat, Cluster==1)$Quality)
[1] 5.494186
>   #nota mitja Cluster=2
>   mean(filter(ds_resultat, Cluster==2)$Quality)
[1] 5.23
>   #nota mitja Cluster=3
>   mean(filter(ds_resultat, Cluster==3)$Quality)
[1] 5.8953488
>   #nota mitja Cluster=4
>   mean(filter(ds_resultat, Cluster==4)$Quality)
[1] 5.8225806
>   #nota mitja Cluster=5
>   mean(filter(ds_resultat, Cluster==5)$Quality)
[1] 5.375
>   #nota mitja Cluster=6
>   mean(filter(ds_resultat, Cluster==6)$Quality)
[1] 6.4324324
>   #nota mitja Cluster=7
>   mean(filter(ds_resultat, Cluster==7)$Quality)
[1] 5.1044776
>   #nota mitja Cluster=8
>   mean(filter(ds_resultat, Cluster==8)$Quality)
[1] 5.6803279
>   #nota mitja Cluster=9
>   mean(filter(ds_resultat, Cluster==9)$Quality)
[1] 5.8484848
>   #nota mitja Cluster=10
>   mean(filter(ds_resultat, Cluster==10)$Quality)
[1] 5.373494
```

Veiem que existeixen variacions entre les notes mitjanes dels grups. De fet veiem que existeix un grup on clarament la nota és molt més alta que en la resta. El grup 6 té una nota de 6.4. Els seus 148 vins tenen unes puntuacions molt més altes podem veure, i el grup 7 compta amb les notes més baixes, amb una mitjana de 5.1. Els seus 134 vins compten amb notes molt més baixes si en desglossem el detall:

```
>   #nota mitja Cluster=6
>   mean(filter(ds_resultat, Cluster==6)$Quality)
[1] 6.4324324
>   #nota mitja Cluster=7
>   mean(filter(ds_resultat, Cluster==7)$Quality)
[1] 5.1044776
```

Aquestes dades ens permeten concloure el plantejament que havíem fet inicialment, que vins amb característiques químiques similars, permeten fer agrupacions de les quals algunes tindran qualificacions més altes que les altres.

Ara, gracies a la agrupació obtinguda, podríem analitzar les característiques d'aquests 148 vins que destaquen per sobre dels altres, i intentar aconseguir maximitzar la producció d'aquest tipus de vi que sabem que és molt més apreciat i per tant podrà ser més valorat pel consumidor.

7. **Codi**: Cal adjuntar el codi, preferiblement en R, amb el que s'ha realitzat la neteja, anàlisi i representació de les dades. Si ho preferiu, també podeu treballar en Python.

La ruta de GitHub on es troba la solució amb el codi font és:

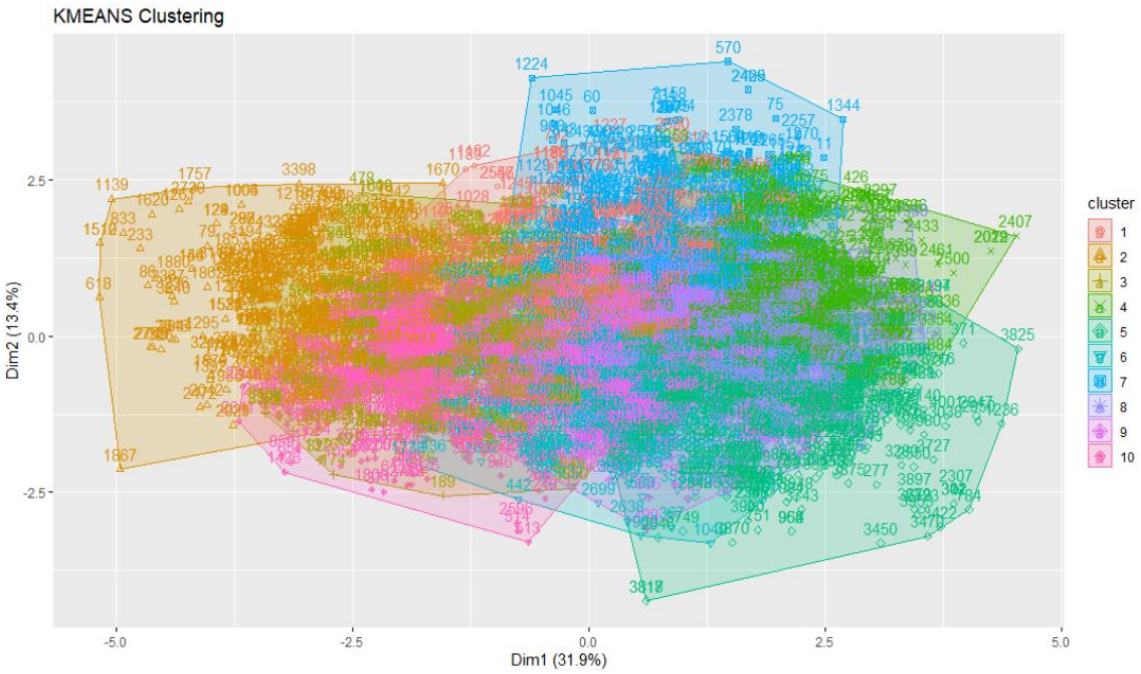
https://github.com/jordipuiggros/WineAnalysis

Hi trobareu adjunts els fitxers de sortida, on hi ha per als vins blancs i negres la classificació que li correspon i la seva nota. Amb aquestes tres columnes és com hem elaborat les conclusions finals del estudi.

AMPLIACIÓ - ANNEX :

Ampliem amb un anàlisis ràpid per als vins blancs amb l'objectiu de trobar les notes mitjanes per a cada clúster:

La gràfica que obtindríem amb la classificació per K-MEANS seria aquesta:



La nota mitja de tot el dataset de vins blancs és: 5.9

I si calculem aquestes mitjanes per grups de classificació obtindrem les següents puntuacions. Veurem que en el cas dels vins blancs, existeixen 3 agrupacions amb característiques químiques similars, però diferents entre elles, que permeten assolir la nota més alta que havíem trobat en una de les agrupacions de vins negres.

```
>  #Nota mitja total
>   mean(ds_resultat2$Quality)
[1] 5.9490255
>
>   #nota mitja Cluster=1
>   mean(filter(ds_resultat2, Cluster==1)$Quality)
[1] 6.1189711
>   #nota mitja Cluster=2
>   mean(filter(ds_resultat2, Cluster==2)$Quality)
[1] 5.8304668
>   #nota mitja Cluster=3
>   mean(filter(ds_resultat2, Cluster==3)$Quality)
[1] 5.7615741
>   #nota mitja Cluster=4
>   mean(filter(ds_resultat2, Cluster==4)$Quality)
[1] 6.4046997
>   #nota mitja Cluster=5
>   mean(filter(ds_resultat2, Cluster==5)$Quality)
[1] 6.4317181
>   #nota mitja Cluster=6
>   mean(filter(ds_resultat2, Cluster==6)$Quality)
[1] 6.0216346
>   #nota mitja Cluster=7
>   mean(filter(ds_resultat2, Cluster==7)$Quality)
[1] 5.6637931
>   #nota mitja Cluster=8
>   mean(filter(ds_resultat2, Cluster==8)$Quality)
[1] 6.438172
>   #nota mitja Cluster=9
>   mean(filter(ds_resultat2, Cluster==9)$Quality)
[1] 5.4839572
>   #nota mitja Cluster=10
>   mean(filter(ds_resultat2, Cluster==10)$Quality)
[1] 5.4415842
```

Així doncs, els clústers 8 amb una nota de 6.438, el clúster 4 amb una nota de 6.404 i el clúster 5 amb una nota de 6.413 serien vins de qualitat alta, igual que el clúster 6 dels vins negres.

Tenim identificades totes les observacions que pertanyen a aquests 4 clústers, de manera que ara podríem ampliar l'estudi analitzant les variables químiques d'aquests clústers i intentar treure conclusions sobre quines són les que tenen una influencia més gran sobre la nota. Per fer-ho ara potser empraríem altres tècniques com la de correlació o potser una ANOVA si volguéssim fer un anàlisis de varies variables simultànies.