Pràctica 2: Tractament de les dades ‘Red Wine Quality’

Roger Pera Martin, Jon Iñaki Mujika

1. Descripció del dataset

El conjunt de dades que analitzarem fa referència al vi portugués 'Vinho Verde' i s’ha extret del següent enllaç de Kaggle (https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009). Aquest conjunt de dades està constituït per 12 característiques i 1599 registres de vins. Una de les variables descriu la qualitat del vi, mentre que la resta de variables són característiques fisicoquímiques dels vins. Pel que fa a les característiques del nostre data set podem veure els següents:

- fixed acidity: àcids que es troben en el vi i són fixes o no volàtils.
- volatile acidity: quantitat d’àcid acètic en el vi(quantitats altes poden portar a sabor de vinagre).
- citric acid: quantitat d’àcid cítric.
- residual sugar: quantitat residual de sucre després de la fermentació.
- chlorides: quantitat de sal en el vi.
- free sulfur dioxide: si esta lliure de sulfur de diòxid.
- total sulfur dioxide: quantitat de sulfur de diòxid és superior a 50ppm.
- density: densitat de l’aigua respecte el percentatge d’alcohol i sucre.
- pH: descripció de quant àcid és el vi.
- sulphates: quantitat de sulfats que té el vi.
- alcohol: percentatge d’alcohol contingut.
- quality: variable de sortida.

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

A partir d’aquest dataset el que s’intenta és seleccionar quines poden ser les dades que ens poden ajudar en major mesura a determinar si un vi té una qualitat bona, acceptable o dolenta. Per a arribar a aquest resultat final es poden realitzar models de regressió que en funció de les diferents característiques i contrastos d’hipòtesis extrauran un resultat que nosaltres ens encarregarem de classificar.
Pel que fa a la selecció nosaltres ens quedarem amb totes les característiques recuperades del nostre data set, ja que totes elles poden ser importants per realitzar la nostra classificació.

3. Neteja de les dades

Abans de començar amb la neteja de dades, procedirem a la càrrega del fitxer csv seleccionat:

```
winequality_red <- read.csv('wine.csv')
head(winequality_red)
str(winequality_red)
```

```
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 ...
```

Per tant, les 11 variables independents són numèriques, i la variable dependent ‘quality’ conté números enters.

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

Una vegada realitzada la càrrega podem procedir a mirar si el nostre conjunt de dades conté valors vuits o sense informar, per això ens disposarem a llençar aquesta comanda que ens indicarà si el nostre data set conté valors 0 o NA.

```
sapply(winequality_red, function(x) sum(is.na(x)))
sapply(winequality_red, function(x) sum(0))
```

El resultat és:

```
       fixed.acidity     volatile.acidity          citric.acid       residual.sugar
                   0                   0                     0                     0
           chlorides  free.sulfur.dioxide total.sulfur.dioxide               density
                   0                   0                     0                     0
                  pH           sulphates             alcohol             quality
                   0                   0                     0                     0
       fixed.acidity     volatile.acidity          citric.acid       residual.sugar
                   0                   0                     0                     0
           chlorides  free.sulfur.dioxide total.sulfur.dioxide               density
                   0                   0                     0                     0
                  pH           sulphates             alcohol             quality
                   0                   0                     0                     0
```

En aquest cas podem veure que no trobem valors desconeguts però en el cas de tenir-ne hauríem de plantejar com solucionar aquesta manca de valors. Algun dels mètodes podria ser eliminar els registres(no és molt òptim), omplir els valors amb la mitjana per que afectin en la menor mesura o aplicar un algoritme de k-means que mira els veïns més propers al registre que conté el valor desconegut.

3.2. Identificació i tractament de valors extrems

Els valors extrems vindrien a ser tots aquells que són incongruents si es comparen amb els demés. Farem servir 'boxplot' per a determinar quines observacions són considerades outliers. Per això es

realitzarà una comprovació de tots aquests en cadascuna de les característiques del nostre conjunt de dades:

```
boxplot.stats(winequality_red$`fixed acidity`)$out
boxplot.stats(winequality_red$`volatile acidity`)$out
boxplot.stats(winequality_red$`citric acid`)$out
boxplot.stats(winequality_red$`residual sugar`)$out
boxplot.stats(winequality_red$chlorides)$out
boxplot.stats(winequality_red$`free sulfur dioxide`)$out
boxplot.stats(winequality_red$`total sulfur dioxide`)$out
boxplot.stats(winequality_red$density)$out
boxplot.stats(winequality_red$pH)$out
boxplot.stats(winequality_red$sulphates)$out
boxplot.stats(winequality_red$alcohol)$out
```

El resultat és:

```
 [1] 12.8 12.8 15.0 15.0 12.5 13.3 13.4 12.4 12.5 13.8 13.5 12.6 12.5 12.8 12.8 14.0 13.7
[18] 13.7 12.7 12.5 12.8 12.6 15.6 12.5 13.0 12.5 13.3 12.4 12.5 12.9 14.3 12.4 15.5 15.5
[35] 15.6 13.0 12.7 13.0 12.7 12.4 12.7 13.2 13.2 13.2 15.9 13.3 12.9 12.6 12.6
 [1] 1.130 1.020 1.070 1.330 1.330 1.040 1.090 1.040 1.240 1.185 1.020 1.035 1.025 1.115
[15] 1.020 1.020 1.580 1.180 1.040
[1] 1
  [1]  6.10  6.10  3.80  3.90  4.40 10.70  5.50  5.90  5.90  3.80  5.10  4.65  4.65  5.50
 [15]  5.50  5.50  5.50  7.30  7.20  3.80  5.60  4.00  4.00  4.00  4.00  7.00  4.00  4.00
 [29]  6.40  5.60  5.60 11.00 11.00  4.50  4.80  5.80  5.80  3.80  4.40  6.20  4.20  7.90
 [43]  7.90  3.70  4.50  6.70  6.60  3.70  5.20 15.50  4.10  8.30  6.55  6.55  4.60  6.10
 [57]  4.30  5.80  5.15  6.30  4.20  4.20  4.60  4.20  4.60  4.30  4.30  7.90  4.60  5.10
 [71]  5.60  5.60  6.00  8.60  7.50  4.40  4.25  6.00  3.90  4.20  4.00  4.00  4.00  6.60
 [85]  6.00  6.00  3.80  9.00  4.60  8.80  8.80  5.00  3.80  4.10  5.90  4.10  6.20  8.90
 [99]  4.00  3.90  4.00  8.10  8.10  6.40  6.40  8.30  8.30  4.70  5.50  5.50  4.30  5.50
[113]  3.70  6.20  5.60  7.80  4.60  5.80  4.10 12.90  4.30 13.40  4.80  6.30  4.50  4.50
[127]  4.30  4.30  3.90  3.80  5.40  3.80  6.10  3.90  5.10  5.10  3.90 15.40 15.40  4.80
[141]  5.20  5.20  3.75 13.80 13.80  5.70  4.30  4.10  4.10  4.40  3.70  6.70 13.90  5.10
[155]  7.80
  [1] 0.176 0.170 0.368 0.341 0.172 0.332 0.464 0.401 0.467 0.122 0.178 0.146 0.236 0.610
 [15] 0.360 0.270 0.039 0.337 0.263 0.611 0.358 0.343 0.186 0.213 0.214 0.121 0.122 0.122
 [29] 0.128 0.120 0.159 0.124 0.122 0.122 0.174 0.121 0.127 0.413 0.152 0.152 0.125 0.122
 [43] 0.200 0.171 0.226 0.226 0.250 0.148 0.122 0.124 0.124 0.143 0.222 0.039 0.157 0.422
 [57] 0.034 0.387 0.415 0.157 0.157 0.243 0.241 0.190 0.132 0.126 0.038 0.165 0.145 0.147
 [71] 0.012 0.012 0.039 0.194 0.132 0.161 0.120 0.120 0.123 0.123 0.414 0.216 0.171 0.178
 [85] 0.369 0.166 0.166 0.136 0.132 0.132 0.123 0.123 0.123 0.403 0.137 0.414 0.166 0.168
 [99] 0.415 0.153 0.415 0.267 0.123 0.214 0.214 0.169 0.205 0.205 0.039 0.235 0.230 0.038
 [1] 52 51 50 68 68 43 47 54 46 45 53 52 51 45 57 50 45 48 43 48 72 43 51 51 52 55 55 48 48
[30] 66
 [1] 145 148 136 125 140 136 133 153 134 141 129 128 129 128 143 144 127 126 145 144 135 165
[23] 124 124 134 124 129 151 133 142 149 147 145 148 155 151 152 125 127 139 143 144 130 278
[45] 289 135 160 141 141 133 147 147 131 131 131
 [1] 0.99160 0.99160 1.00140 1.00150 1.00150 1.00180 0.99120 1.00220 1.00220 1.00140 1.00140
[12] 1.00140 1.00140 1.00320 1.00260 1.00140 1.00315 1.00315 1.00315 1.00210 1.00210 0.99170
[23] 0.99220 1.00260 0.99210 0.99154 0.99064 0.99064 1.00289 0.99162 0.99007 0.99007 0.99020
[34] 0.99220 0.99150 0.99157 0.99080 0.99084 0.99191 1.00369 1.00369 1.00242 0.99182 1.00242
[45] 0.99182
 [1] 3.90 3.75 3.85 2.74 3.69 3.69 2.88 2.86 3.74 2.92 2.92 2.92 3.72 2.87 2.89 2.89 2.92
```

```
[18] 3.90 3.71 3.69 3.69 3.71 3.71 2.89 2.89 3.78 3.70 3.78 4.01 2.90 4.01 3.71 2.88 3.72
[35] 3.72
 [1] 1.56 1.28 1.08 1.20 1.12 1.28 1.14 1.95 1.22 1.95 1.98 1.31 2.00 1.08 1.59 1.02 1.03
[18] 1.61 1.09 1.26 1.08 1.00 1.36 1.18 1.13 1.04 1.11 1.13 1.07 1.06 1.06 1.05 1.06 1.04
[35] 1.05 1.02 1.14 1.02 1.36 1.36 1.05 1.17 1.62 1.06 1.18 1.07 1.34 1.16 1.10 1.15 1.17
[52] 1.17 1.33 1.18 1.17 1.03 1.17 1.10 1.01
 [1] 14.00000 14.00000 14.00000 14.00000 14.90000 14.00000 13.60000 13.60000 13.60000
[10] 14.00000 14.00000 13.56667 13.60000
```

Per tant, d'acord amb el criteri de boxplot, hi ha uns quants outliers. Ara, els analitzarem visualment. Representarem cada variable independent amb respecte a la variable 'quality', que és la variable dependent. Però, primer de tot crearem un vector amb els noms de les variables independents (també farem servir aquest vector més endavant):

```
var_indep ← c("fixed.acidity","volatile.acidity","citric.acid","residual.sugar","chlorides","free.sulfur.dioxide","total.sulfu
r.dioxide","density","pH","sulphates","alcohol")

for (i in var_indep){
  boxplot(winequality_red[,i] ~ winequality_red$quality, col='chartreuse3', xlab='Quality', ylab=i, varwidth=T)
}
```

Una vegada vists tots els resultats s'ha decidit esborrar el outlier del àcid cítric y els dos valors més extrems del atribut 7(total sulfur dioxide) ja que creiem que 3 registres no afectaran en el resultat de l'analisis. Tots els demés valors es deixaran igual ja que hem valorat que poden ser valors probables en el conjunt de dades.

```
outliers <- boxplot.stats(winequality_red$citric.acid)$out
winequality_red <- winequality_red[-which(winequality_red$citric.acid %in% outliers),]

outliers2 <- winequality_red$total.sulfur.dioxide[winequality_red$total.sulfur.dioxide >250]
winequality_red <- winequality_red[-which(winequality_red$total.sulfur.dioxide %in%outliers2),]
```

4. Anàlisi de les dades

4.1. Selecció dels grups de dades que es volen analitzar/comparar

En principi farem servir totes les dades per a crear el model final. Una vegada netejades les dades procedim a la seva corresponent exportació a un fitxer CSV:

```
write.csv(winequality_red, "winequality_red_limpio.csv")
```

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

Primer de tot, analitzarem visualment la distribució de les dades, per variables. Primer, els histogrames (utilitzant el vector amb els noms de les variables independents creat amb anterioritat):

```
for (i in var_indep){
  hist(winequality_red_DataFrame[,i],breaks = 50,main=i)
}
```

També visualitzarem les gràfiques Q-Q de les variables independents:

```
for (i in var_indep){
  qqnorm(winequality_red_DataFrame[,i], main=i)
  qqline(winequality_red_DataFrame[,i],col=2)
```

```
}
```

Per tant, sembla que les variables 'density' i 'pH' podrien ser les úniques variables que segueixen una distribució normal, tot i que es necessita una anàlisi més quantitativa.

Per tal de comprovar si les dades efectivament tenen un distribució normal, farem el test de normalitat Shapiro-Wilk:

```
for (i in 1:11) {
  print(shapiro.test(as.numeric( unlist(winequality_red_DataFrame[,i]))))
}
```

```
        Shapiro-Wilk normality test

W = 0.94199, p-value < 2.2e-16
W = 0.97443, p-value = 3.021e-16
W = 0.95459, p-value < 2.2e-16
W = 0.56457, p-value < 2.2e-16
W = 0.49949, p-value < 2.2e-16
W = 0.90135, p-value < 2.2e-16
W = 0.88985, p-value < 2.2e-16
W = 0.99061, p-value = 1.339e-08
W = 0.99307, p-value = 8.078e-07
W = 0.84659, p-value < 2.2e-16
W = 0.92884, p-value < 2.2e-16
```

Per tant, d'acord amb aquest mètode, amb cap de les variables les dades tenen una distribució normal.

4.3. Aplicació de proves estadístiques per comparar els grups de dades

Per tal de determinar quines variables independents tenen més efecte sobre el valor de la variable 'quality', calcularem la correlació entre les variables independents i la dependent. Atès que no tenen una distribució normal, farem servir el mètode 'Spearman'. Els 11 valors obtinguts són:

```
0.1152764
-0.3791709
0.2131077
0.03030028
-0.1856902
-0.05874558
-0.199945
-0.1729519
-0.04247587
0.3833267
0.4762987
```

Per tant, les variables que més influeixen 'quality' són: 'volatile acidity' (correlació de -0.38), 'sulphates' (correlació de 0.38) i 'alcohol' (correlació de 0.48).

Amb aquesta informació, per tal de poder predir si un vi és bo, dolent o acceptable, crearem diferents models de regressió lineal dels quals escollirem el millor per finalment poder fer prediccions òptimes segons els paràmetres d'entrada.

```
#Variables independents
acidity = winequality_red$fixed.acidity
volAcidity = winequality_red$volatile.acidity
citricAcid = winequality_red$citric.acid
resSugar = winequality_red$residual.sugar
chlorides = winequality_red$chlorides
sulfDioxide = winequality_red$free.sulfur.dioxide
totSulfDioxide = winequality_red$total.sulfur.dioxide
density = winequality_red$density
ph = winequality_red$pH
sulphates = winequality_red$sulphates
alcohol = winequality_red$alcohol

# Variable a predir
calidad = winequality_red$quality

# Generació de diversos models
modelo1 <- lm(calidad ~ volAcidity + sulphates + alcohol, data = winequality_red) #Vars que influeixen mes sobre quality nomes
modelo2 <- lm(calidad ~ volAcidity + citricAcid + chlorides + totSulfDioxide + density + sulphates + alcohol, data = winequality_red)
modelo3 <- lm(calidad ~ acidity + volAcidity + citricAcid + resSugar + chlorides, data = winequality_red)
modelo4 <- lm(calidad ~ acidity + volAcidity + citricAcid + resSugar + chlorides + sulfDioxide + totSulfDioxide + density + ph, data = winequality_red)
modelo5 <- lm(calidad ~ acidity + volAcidity + citricAcid + resSugar + chlorides + sulfDioxide + totSulfDioxide + density + ph + sulphates + alcohol, data = winequality_red)
```

Per comparar els diferents models creats, es realitzarà una taula amb el coeficient de determinació dels diferents models:

```
tabla.coeficientes <- matrix(c(1, summary(modelo1)$r.squared,
2, summary(modelo2)$r.squared,
3, summary(modelo3)$r.squared,
4, summary(modelo4)$r.squared,
5, summary(modelo5)$r.squared),
ncol = 2, byrow = TRUE)
colnames(tabla.coeficientes) <- c("Modelo", "R2")
tabla.coeficientes
```

Els valors obtinguts són:

```
     Modelo       R2
[1,]      1 0.3376173
[2,]      2 0.3536505
[3,]      3 0.1625842
[4,]      4 0.2612435
[5,]      5 0.3627040
```

Després de comparar els següents valors, podríem triar el model5, ja que aquest és el que té el coeficient de determinació més alt, per tant, podríem dir que aquest model obtindria una predicció més òptima dels valors qualitat en comparació als altres models segons els seus paràmetres d'entrada. També cal tenir en compte que per aquest model s'han deixat tots els paràmetres d'entrada per a demostrar la importància de tots aquests per a calcular la qualitat d'un vi. En un altre cas, si el que volguéssim fos un model on no tinguéssim que ficar-hi tots els paràmetres d'entrada escolliríem

el segon model, on només s'han afegir part del paràmetres que més influeixen en el resultat de la qualitat del vi.

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

Els resultats més significatius de l'estudi ja han sigut presentats a l'apartat anterior, especialment aquells obtinguts amb els mètodes de regressió i correlació amb la variable dependent.
Per tal de completar l'anàlisi, també analitzarem la correlació entre les variables independents per tal de valorar la similitud entre ells. La correlació es pot calcular amb la següent ordre:

```
cor(winequality_red,method="spearman")
```

Atès que la distribució de les dades per a cada variable no són normals, hem fet servir el mètode Spearman. Amb aquesta comanda es calculen les correlacions entre totes les parelles de variables. Atès que el nombre de parelles és molt gran, seleccionarem les correlacions més significatives, que són les següents:

Parelle de variables	**correlació**
fixed acidity vs citric acid	0.6617
fixed acidity vs density	0.6231
fixed acidity vs pH	-0.7067
volatile acidity vs citric acid	-0.6103
total sulfur dioxide vs free sulfur dioxide	0.7897

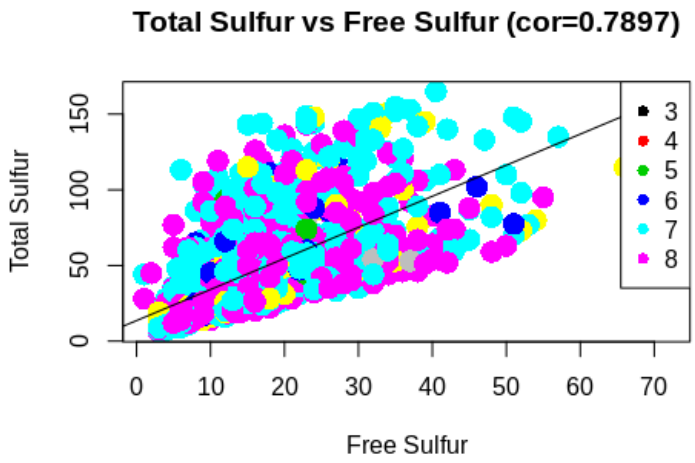
A continuació, representarem gràficament les dades amb respecte aquestes variables. Per fer-ho, també inclourem la classe de la variable 'quality' al qual pertany cada dada (es diferenciaran pel color). Però, primer s'ha de factoritzar la variable 'quality':

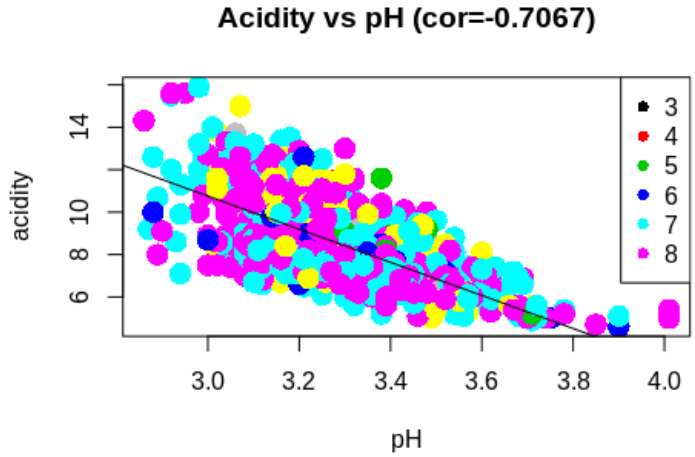
```
ff <- as.factor(winequality_red$quality)
```

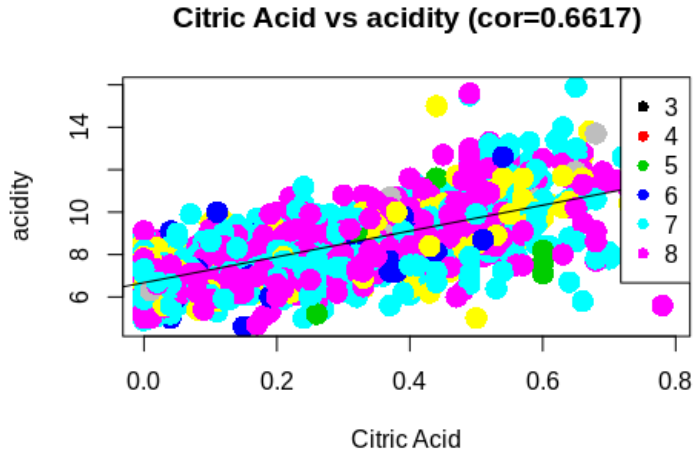
Com a exemple, només inclourem la comanda per a crear la primera gràfica:

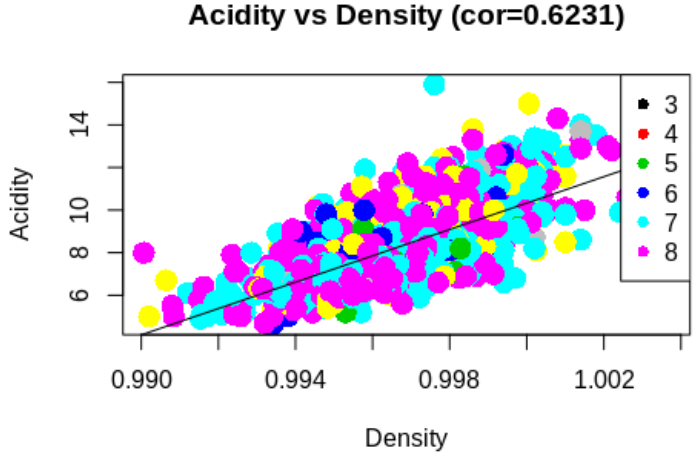
```
fit1 <- lm (totSulfDioxide ~ sulfDioxide, data = winequality_red)
summary(fit1)
plot(sulfDioxide,totSulfDioxide,
    col = winequality_red$quality,
    pch = 16,
    cex = 2,
    ylab = "Total Sulfur",
    xlab = "Free Sulfur",
    main = "Total Sulfur vs Free Sulfur (cor=0.7897)")
abline(fit1)
legend ("topright",legend = levels(ff),col=c(1:6),pch=16)
```

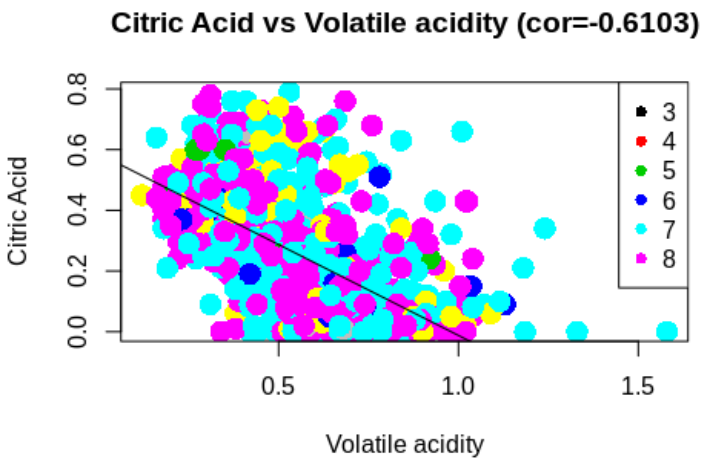
De manera similar es creen la resta de gràfiques. Els resultats són els següents:


Total Sulfur vs Free Sulfur (cor=0.7897)
Total Sulfur
Free Sulfur
0
50
100
150
0
10
20
30
40
50
60
70
3
4
5
6
7
8


Acidity vs pH (cor=-0.7067)
acidity
pH
6
8
10
14
3.0
3.2
3.4
3.6
3.8
4.0
3
4
5
6
7
8


Citric Acid vs acidity (cor=0.6617)
acidity
Citric Acid
6
8
10
14
0.0
0.2
0.4
0.6
0.8
3
4
5
6
7
8


Acidity vs Density (cor=0.6231)
Acidity
Density
6
8
10
14
0.990
0.994
0.998
1.002
3
4
5
6
7
8



on la llegenda indica el color de la classe de la variable ‘quality’. Les gràfiques confirmen que hi ha una certa correlació entre les variables representades, tot i que no massa gran. Aquest fet pot indicar que no hi ha dues variables que donin la mateixa informació i per tant totes poden ser rellevants per a descriure un vi. A més, tampoc s’observa una clara discriminació pels colors, fet que confirma la necessitat de combinar més d’una variable per a generar un model de regressió, tal com es va comprovar a l’apartat anterior.

6. Resolució del problema

Per començar el que s'ha realitzat ha sigut una neteja de les dades les quals han sigut analitzades per poder comprendre quins eren els valors sobrants, erronis o fora de rang. Afortunadament no hi hagut gaires registres que s'hagin tingut que netejar i únicament s'han descartat 3 registres que estaven molt fora de rang i que no influien en gran mesura sobre l'analis a realitzar.

Posteriorment s'ha realitzat l'analisi de les dades ja natejades, entre els quals podem destacar l'estudi de normalitat de les dades, proves estadístiques que miraven la correlació de cadascuna de les variables amb el resultat final, del qual s'ha pogut extreure un ranking influència en el valor final que ha quedat de la següent manera:

1.alcohol

2.Volatil àcid

3.Sulphates

4.citric Acid

5.total Sulfur Dioxide

6.chlorides

7.Altres

I per finalitzar, s'han realitzat diversos models segons les diferent variables en 4 dels casos escollint les variables amb major correlació sobre el resultat final i en un del casos vàries variables aleatòries,

en aquest apartat hem pogut demostrar la certesa de les correlacions calculades en el apartat anterior i s'ha escollit els millors model amb els que es podria realitzar una predicció en diferents casos.

7. Codi

Podeu trobar el codi en aquest fitxer: codi_wine.r

Contribució dels components del grup

Contribucions	Signa
Recerca prèvia	R. P. M., J. I. M. G.
Redacció de les respostes	R. P. M., J. I. M. G.
Desenvolupament codi	R. P. M., J. I. M. G.