Pràctica 2: Neteja i anàlisi de les dades

Rigau, Pol. Tienda, Arnau

13/05/2020

Anàlisis de vins

Descripció

En aquesta practica hem escollit el dataset "Red Wine Quality" disponible a la pagina web kraggle i al repositori UCI.

Aquest dataset conté informació sobre diversos paràmetres químics resultants de l'analisis de vins blancs i negres de la regió portugesa 'Vinho verde' i una classificació segons la seva qualitat.

Amb aquestes dades es poden crear algoritmes per a la classificació de vins. Ens permet agrupar vins per les seves semblances i també determinar quins son els factors que afecten més a la seva qualitat.

Neteja de dades

En primer lloc, es carreguen les dades:

```
wine <- read.csv('winequality-red.csv', header = TRUE)
```

Es mostren quins son els atributs

- Àcidesa fixe
- Àcidesa volatil
- Àcid cítric
- Sucre residual
- Clorurs
- Diòxid de sofre lliure
- Diòxid de sofre residual
- Densitat
- pH
- Sulfats
- Alcohol
- Qualitat

Tenim 11 descriptors i un resultat, que es la **qualitat**.

S'observa que no hi ha cap atribut per identificar els registres, pel que s'asigna un número **id** a cada fila.

```
id <- 1:nrow(wine)
wine <- cbind(id=id, wine)
```

A continuació realitzem un primer analisi dels atributs que tenim:

```
str(wine)
```

```
## 'data.frame':    1599 obs. of  13 variables:
##  $ id                  : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ 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 ...
```

Observem que les dades que tenim són númeriques, excepte les corresponents als atributs **id** i **quality**. En el primer cas és correcte, peò les dades de **quality** haurien de ser factors, per tal de tenir una mesura, i per tant les convertim.

```
wine$quality <- as.factor(wine$quality)
```

Ens interesa treballar amb totes les dades que disposem, no descartarem cap dels atributs dels que disposem.

A continuació comprovem si hi ha dades nules.

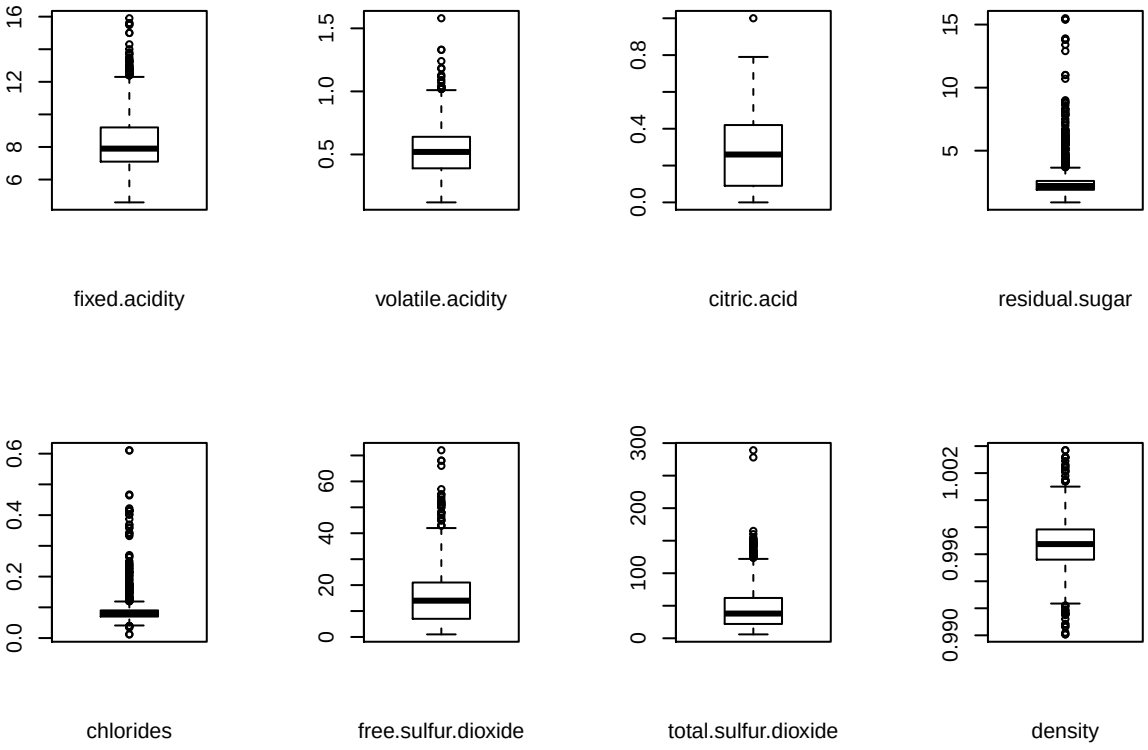
```
sapply(wine, function(x) sum(is.na(x)))
```

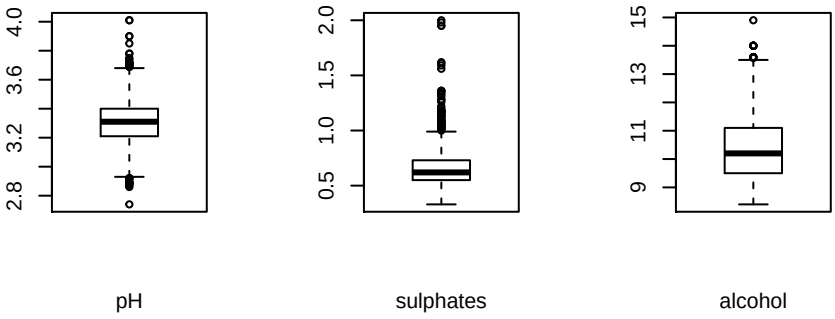
```
##                   id        fixed.acidity     volatile.acidity
##                    0                    0                    0
##          citric.acid       residual.sugar            chlorides
##                    0                    0                    0
##  free.sulfur.dioxide total.sulfur.dioxide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

Veiem que no n'hi ha en cap dels atributs.

A continació realitzarem una exploració dels outliers, pel que graficarem els bloxpot dels atributs que siguin númerics per a tenir una representació visual.

```
graph_wine = par(mfrow = c(2,4))

for (i in colnames(wine)){
  if (!(i %in% c('quality', 'id'))){
    boxplot(wine[[i]], xlab = i)
  }
}
```


6
8
12
16
fixed.acidity
0.5
1.0
1.5
volatile.acidity
0.0
0.4
0.8
citric.acid
5
10
15
residual.sugar
0.0
0.2
0.4
0.6
chlorides
0
20
40
60
free.sulfur.dioxide
0
100
200
300
total.sulfur.dioxide
0.990
0.996
1.002
density



Observem que en la majoria de casos tenim valors extrems en la part de valors més elevats.

Anàlisi de dades

A continuacio es fara un anàlisis de les dades disponibles.

```
summary(wine)
```

```
##        id         fixed.acidity   volatile.acidity  citric.acid
##  Min.   :   1.0   Min.   : 4.60   Min.   :0.1200   Min.   :0.000
##  1st Qu.: 400.5   1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090
##  Median : 800.0   Median : 7.90   Median :0.5200   Median :0.260
##  Mean   : 800.0   Mean   : 8.32   Mean   :0.5278   Mean   :0.271
##  3rd Qu.:1199.5   3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420
##  Max.   :1599.0   Max.   :15.90   Max.   :1.5800   Max.   :1.000
##  residual.sugar     chlorides       free.sulfur.dioxide
##  Min.   : 0.900   Min.   :0.01200   Min.   : 1.00
##  1st Qu.: 1.900   1st Qu.:0.07000   1st Qu.: 7.00
##  Median : 2.200   Median :0.07900   Median :14.00
##  Mean   : 2.539   Mean   :0.08747   Mean   :15.87
##  3rd Qu.: 2.600   3rd Qu.:0.09000   3rd Qu.:21.00
##  Max.   :15.500   Max.   :0.61100   Max.   :72.00
##  total.sulfur.dioxide    density             pH          sulphates
##  Min.   :  6.00       Min.   :0.9901   Min.   :2.740   Min.   :0.3300
##  1st Qu.: 22.00       1st Qu.:0.9956   1st Qu.:3.210   1st Qu.:0.5500
```

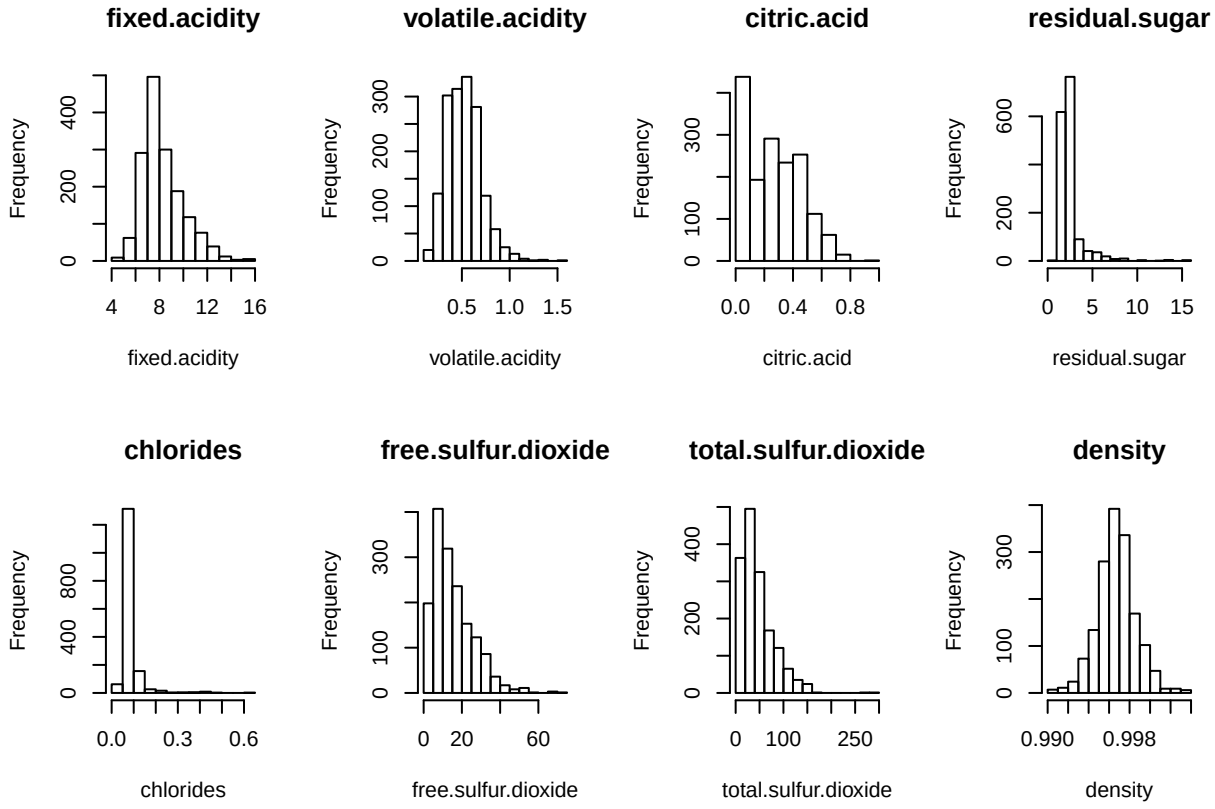
```
##  Median : 38.00       Median :0.9968   Median :3.310   Median :0.6200
##  Mean   : 46.47       Mean   :0.9967   Mean   :3.311   Mean   :0.6581
##  3rd Qu.: 62.00       3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:0.7300
##  Max.   :289.00       Max.   :1.0037   Max.   :4.010   Max.   :2.0000
##     alcohol      quality
##  Min.   : 8.40   3: 10
##  1st Qu.: 9.50   4: 53
##  Median :10.20   5:681
##  Mean   :10.42   6:638
##  3rd Qu.:11.10   7:199
##  Max.   :14.90   8: 18
```

Es realitza un altre tipus de gràfic, en aquest cas un histograma, per a observar la distribució de les dades. Aquest analisi, també ens ajudara a comprobar si segueixen la normalitat.

```
graph_wine = par(mfrow = c(2,4))

for (i in colnames(wine[-1])){
  hist(as.numeric(wine[[i]]), xlab = i, main= i)
  if (!(i == 'quality')){
    results_test = shapiro.test(wine[[i]])
    print(results_test)
  }
}
```

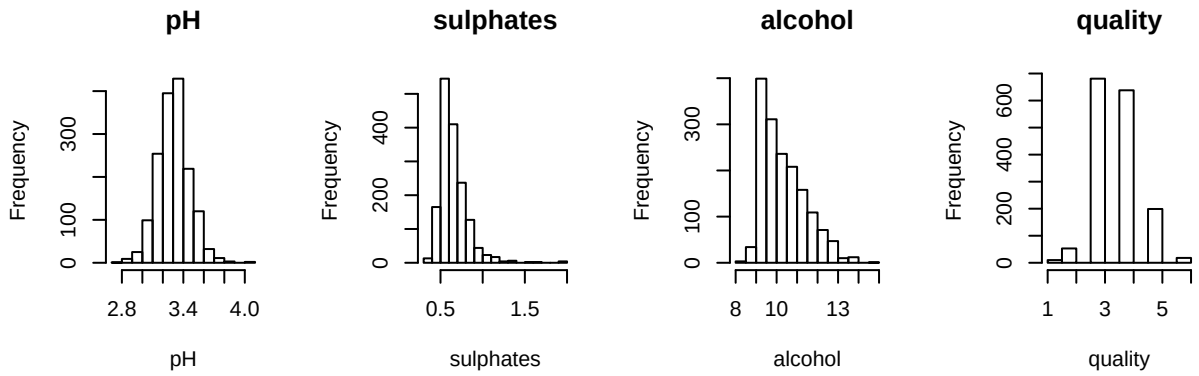
```
##
##  Shapiro-Wilk normality test
##
## data:  wine[[i]]
## W = 0.94203, p-value < 2.2e-16

##
##  Shapiro-Wilk normality test
##
## data:  wine[[i]]
## W = 0.97434, p-value = 2.693e-16

##
##  Shapiro-Wilk normality test
##
## data:  wine[[i]]
## W = 0.95529, p-value < 2.2e-16

##
##  Shapiro-Wilk normality test
##
## data:  wine[[i]]
## W = 0.56608, p-value < 2.2e-16

##
##  Shapiro-Wilk normality test
##
## data:  wine[[i]]
## W = 0.48425, p-value < 2.2e-16
```

```
##
##  Shapiro-Wilk normality test
##
## data:  wine[[i]]
## W = 0.90184, p-value < 2.2e-16

##
##  Shapiro-Wilk normality test
##
## data:  wine[[i]]
## W = 0.87322, p-value < 2.2e-16
```




```
##
##  Shapiro-Wilk normality test
##
## data:  wine[[i]]
## W = 0.99087, p-value = 1.936e-08

##
##  Shapiro-Wilk normality test
##
## data:  wine[[i]]
## W = 0.99349, p-value = 1.712e-06

##
```

```
##  Shapiro-Wilk normality test
##
## data:  wine[[i]]
## W = 0.83304, p-value < 2.2e-16

##
##  Shapiro-Wilk normality test
##
## data:  wine[[i]]
## W = 0.92884, p-value < 2.2e-16
```



S'observa, com ja s'havia vist en el bloxplot, les variables no presenten una distribució normal. Cap d'elles cumpleix la hipotesi nula del test de Shapiro-Wilk, pel que podem determinar que no compleixen la normalitat.

Es realitza aleshores el test per a comprobar la homoesciditat de les dades. Com que aquestes no compleixen la normalitat, es realitza el tes de Fligner-Killeen

```
fligner.test(wine)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  wine
## Fligner-Killeen:med chi-squared = 16958, df = 12, p-value <
## 2.2e-16
```

Els resultats obtinguts són contraris a la hipotesi, pel que les variables no compleixen homodescidat.

A continuació d'estudiara la relació que hi ha entre les dades, si aquestes correlacionen entre si.

```
#pairs(wine[-1])
```

Bibliografía.

Red Wine Quality. https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009/undefined. 2018

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.