M2.951 - Tipologia i cicle de vida de les dades - Pràctica 2: Neteja i anàlisi de les dades
Serra Plomer, Martí

1. Descripció del dataset. Perquè és important i quina pregunta/problema pretén respondre?
   Per aquesta pràctica, utilitzarem el dataset Red Wine Quality[1], que conté diferents atributs químics sobre el vi negre de la variant portuguesa "Vinho Verde", juntament amb una puntuació de qualitat entre 0 i 10. Amb els atributs químics es pretén poder crear un model de regressió que obtingui la qualitat estimada del vi. També ens permetrà veure quins atributs tenen més influència en aquesta qualitat. Si triem un valor de qualitat que separi un vi bo de un dolent (per exemple, si diem que un vi bo és aquell amb una qualitat superior a 7) també podem enfocar el problema amb un model de classificació.

2. Integració i selecció de les dades d'interès a analitzar. Pot ser el resultat d'addicionar diferents datasets o una subselecció útil de les dades originals, en base a l'objectiu que es vulgui aconseguir.
   Inicialment, usarem tots els atributs disponibles i registres per analitzar la data. El conjunt de dades consta de 1599 registres, amb els següents atributs:
   - ***fixed acidity***: àcids involucrats amb el vi que no s'evaporen ràpidament.
   - ***volatile acidity***: quantitat d'àcid acètic en el vi que pot generar un gust a vinagre si és molt alt.
   - ***citric acid***: àcid cítric per afegir frescor i gust al vi.
   - ***residual sugar***: quantitat de sucre en el vi després de la fermentació.
   - ***chlorides***: quantitat de sal en el vi.
   - ***free sulfur dioxide***: quantitat de SO_2 lliure. Prevén la generació de microorganismes i l'oxidació del vi.
   - ***total sulfur dioxide***: quantitat de SO_2 lliure i molecular.
   - ***density***: densitat del vi.
   - ***pH***: pH del vi, de 0 (àcid) a 14 (base).
   - ***sulphates***: afegits al vi per augmentar els nivells de SO2 i evitar la generació de microorganismes i l'oxidació del vi.
   - ***alcohol***: percentatge d'alcohol del vi.
   - ***quality***: variable de sortida basada en el tast del vi.

3. Neteja de les dades.

3.1. Les dades contenen zeros o elements buits? Gestiona cadascun d'aquests casos.
   Carreguem les dades amb R i n'obtenim el sumari:

```
{r read}
df <- read.csv("winequality-red.csv")
summary(df)
```

```
fixed.acidity   volatile.acidity  citric.acid    residual.sugar     chlorides       free.sulfur.dioxide
Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900   Min.   :0.01200   Min.   : 1.00
1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900   1st Qu.:0.07000   1st Qu.: 7.00
Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200   Median :0.07900   Median :14.00
Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539   Mean   :0.08747   Mean   :15.87
3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600   3rd Qu.:0.09000   3rd Qu.:21.00
Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500   Max.   :0.61100   Max.   :72.00
total.sulfur.dioxide    density             pH          sulphates         alcohol         quality
Min.   :  6.00       Min.   :0.9901   Min.   :2.740   Min.   :0.3300   Min.   : 8.40   Min.   :3.000
1st Qu.: 22.00       1st Qu.:0.9956   1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50   1st Qu.:5.000
Median : 38.00       Median :0.9968   Median :3.310   Median :0.6200   Median :10.20   Median :6.000
Mean   : 46.47       Mean   :0.9967   Mean   :3.311   Mean   :0.6581   Mean   :10.42   Mean   :5.636
3rd Qu.: 62.00       3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10   3rd Qu.:6.000
Max.   :289.00       Max.   :1.0037   Max.   :4.010   Max.   :2.0000   Max.   :14.90   Max.   :8.000
```

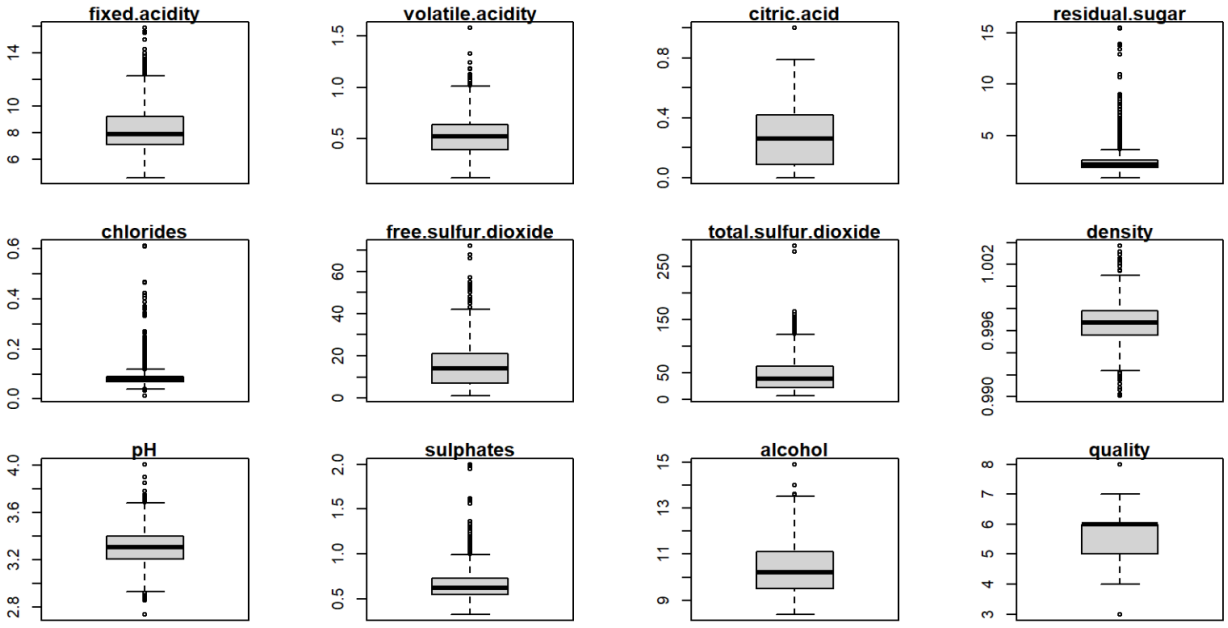
[1] https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009

Podem veure que no tenim elements buits i que l'únic atribut amb zeros és el cítric.acid, però es tracte d'un valor permès i no d'un valor per defecte.

3.2. Identifica i gestiona els valors extrems.

Per visualitzar els valors extrems, mostrarem gràficament la distribució de cadascun dels valors:

```
{r boxplot}
par(mfrow = c(3, 4), mar = c(2, 5, 1, 1))
for (i in 1:length(df)) {
  boxplot(df[, i], main = names(df[i]), type = "l")
}
```



Observem que, basat en la distribució de valors de cada element, tenim valors extrems, com un valor de 15.5 per residual.sugar, o un valor de 2.0 per sulphates. No obstant, tots els valors es troben dins els límits esperats, és a dir, no tenim valors impossibles que es puguin deure a errors o valors que esbiaixin els resultats degut a tenir un ordre de magnitud diferent a la resta.

4. Anàlisi de les dades.

4.1. Selecció dels grups de dades que es volen analitzar/comparar (p. e., si es volen comparar grups de dades, quins són aquests grups i quins tipus d'anàlisi s'aplicaran?).

En aquest cas el que volem fer és comparar cadascun dels atributs amb la variable *quality* per veure com n'afecten el valor, de manera que puguem predir-ne el valor. Per això el que farem serà comprovar la normalitat de cada variable, i l'homoscedasticitat de cada atribut amb *quality*. Les variables que compleixen ambdues propietats són les que podrem usar per aplicar tests estadístics més potents posteriorment.

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

Per comprovar la normalitat de les dades, aplicarem el test de Shapiro sobre cada columna del conjunt:

```
{r shapiro}
sapply(seq(1, length(df)), function(x) {
  paste(colnames(df)[x],
        round(shapiro.test(df[, x])$p.value, 4),
        sep = ": ")
})
```

```
[1] "fixed.acidity: 0"        "volatile.acidity: 0"     "citric.acid: 0"          "residual.sugar: 0"
[5] "chlorides: 0"            "free.sulfur.dioxide: 0"  "total.sulfur.dioxide: 0" "density: 0"
[9] "pH: 0"                   "sulphates: 0"            "alcohol: 0"              "quality: 0"
```

Podem veure que cap dels p-value dels tests realitzats té un valor major a 0.05, de fet, tots els valors són molt propers a 0, per tant, rebutgem la hipòtesi nul·la i no considerem que cap dels atributs segueix una distribució normal. No obstant, com el conjunt de dades es compon d'un nombre de registres prou gran, pel teorema central del límit, considerarem que segueixen una distribució normal per poder aplicar els tests estadístics.

Podem aplicar els tests de homoscedasticitat. Usarem el test de Fligner ja que no podem assumir que les variables segueixen una distribució normal. Compararem cadascuna de les variables amb la variable objectiu (*quality*):

```
{r fligner}
sapply(seq(1, length(df) - 1), function(x) {
  paste(colnames(df)[x],
        round(fligner.test(df[, x] ~ df[, 12])$p.value, 4),
        sep = ": ")
})
```

```
 [1] "fixed.acidity: 0"          "volatile.acidity: 0"       "citric.acid: 0.0531"
 [4] "residual.sugar: 0.1563"    "chlorides: 0.0094"         "free.sulfur.dioxide: 0.0148"
 [7] "total.sulfur.dioxide: 0"   "density: 0"                "pH: 0.9408"
[10] "sulphates: 0.094"          "alcohol: 0"
```

Basant-nos en aquests resultats, podem assumir homoscedasticitat amb *quality* per les variables de *citric.acid*, *residual.sugar*, *pH* i *sulphates*. Aquestes són les variables que usarem per realitzar les proves estadístiques.

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.

Començarem comprovant la correlació. Podem usar la correlació de Pearson amb les variables *citric.acid*, *residual.sugar*, *pH* i *sulphates*. Usarem la correlació de Spearman amb les altres variables.

Comencem creant un dataframe per guardar els resultats:

```
{r corr_df}
df_corr <- data.frame(
  Feature = character(),
  Qlt.Corr = double(),
  Qlt.Corr.p = double(),
  stringsAsFactors = FALSE
)
```

Ara podem calcular les correlacions usant el mètode de Pearson:

```
{r corr_pearson}
for (col in c('citric.acid', 'residual.sugar', 'pH', 'sulphates')) {
  df_corr[nrow(df_corr) + 1,] =
    c(col,
      round(cor(df$quality, df[, col]), 4),
      round(cor.test(df$quality, df[, col])$p.value, 4))
}
```

I usant el mètode de Spearman per les variables que no presentaven homoscedasticitat:

```
{r corr_spearman}
for (col in c(
  'fixed.acidity',
  'volatile.acidity',
  'chlorides',
  'free.sulfur.dioxide',
  'total.sulfur.dioxide',
  'density',
  'alcohol'
)) {
  df_corr[nrow(df_corr) + 1,] =
    c(col,
      round(cor(df$quality, df[, col],
                method = 'spearman'), 4),
      round(
        cor.test(df$quality, df[, col],
                 method = 'spearman', exact = FALSE)$p.value, 4))
}
```

Podem comprovar els resultats de les correlacions:

```
{r corr_results}
df_corr
```

	Feature <chr>	Qlt.Corr <chr>	Qlt.Corr.p <chr>
1	citric.acid	0.2264	0
2	residual.sugar	0.0137	0.5832
3	pH	-0.0577	0.021
4	sulphates	0.2514	0
5	fixed.acidity	0.1141	0
6	volatile.acidity	-0.3806	0
7	chlorides	-0.1899	0
8	free.sulfur.dioxide	-0.0569	0.0229
9	total.sulfur.dioxide	-0.1967	0
10	density	-0.1771	0
11	alcohol	0.4785	0

El p-value de la correlació de residual.sugar és molt superior a 0.05, pel que és possible que la correlació entre residual.sugar i quality sigui 0. Les altres variables tenen valors de p inferior a 0.05, pel que podem afirmar que tenen certa correlació amb la variable quality.
Sembla que els cítrics en el vi influeix positivament en la seva qualitat. Comprovarem si la mitjana de la concentració d'àcid cítric és significativament inferior amb els vins amb menys qualitat que els vins amb més qualitat. Primer separem les dades i obtenim la quantitat de dades a cada grup:

```
{r test_split}
qlt_med <- median(df$quality)
high_qlt_citric <- df[df$quality >= qlt_med, "citric.acid"]
low_qlt_citric <- df[df$quality < qlt_med, "citric.acid"]
c(length(high_qlt_citric), length(low_qlt_citric))
```

```
[1] 855 744
```

Ja hem vist que cítric.acid presentava homoscedasticitat amb quality. Com tenim un nombre elevat de mostres, podem assumir que s'aplica el teorema del límit central i que, per tant, les mitjanes de les mostres segueixen una distribució normal. Ara podem aplicar el test, on la hipòtesi nul·la serà equivalent a assumir que les mitjanes de l'àcid cítric en els vins d'alta i baixa qualitat és la mateixa.

```
{r test}
var.test(high_qlt_citric, low_qlt_citric)
```

```
        F test to compare two variances

data:  high_qlt_citric and low_qlt_citric
F = 1.1883, num df = 854, denom df = 743, p-value = 0.01539
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 1.033583 1.365260
sample estimates:
ratio of variances
          1.188307
```

El p-value és inferior a 0.05, per tant podem rebutjar la hipòtesi nul·la i assumir que les mitjanes són diferents. Com hem vist amb el test de correlació, podem dir que la mitjana d'àcid cítric en els vins de més qualitat és superior que en els vins de menor qualitat.
Finalment, crearem un model de regressió lineal. De moment usarem tots els paràmetres:

```
{r lm}
m1 <- glm(
  formula =
    quality
  ~ citric.acid
  + residual.sugar
  + pH
  + sulphates
  + fixed.acidity
  + volatile.acidity
  + chlorides
  + free.sulfur.dioxide
  + total.sulfur.dioxide
  + density
  + alcohol,
  data = df
)

summary(m1)
```

Aquest és el resultat del model:

```
Call:
glm(formula = quality ~ citric.acid + residual.sugar + pH + sulphates +
    fixed.acidity + volatile.acidity + chlorides + free.sulfur.dioxide +
    total.sulfur.dioxide + density + alcohol, data = df)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.68911  -0.36652  -0.04699   0.45202   2.02498

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           2.197e+01  2.119e+01   1.036   0.3002
citric.acid          -1.826e-01  1.472e-01  -1.240   0.2150
residual.sugar        1.633e-02  1.500e-02   1.089   0.2765
pH                   -4.137e-01  1.916e-01  -2.159   0.0310 *
sulphates             9.163e-01  1.143e-01   8.014 2.13e-15 ***
fixed.acidity         2.499e-02  2.595e-02   0.963   0.3357
volatile.acidity     -1.084e+00  1.211e-01  -8.948  < 2e-16 ***
chlorides            -1.874e+00  4.193e-01  -4.470 8.37e-06 ***
free.sulfur.dioxide   4.361e-03  2.171e-03   2.009   0.0447 *
total.sulfur.dioxide -3.265e-03  7.287e-04  -4.480 8.00e-06 ***
density              -1.788e+01  2.163e+01  -0.827   0.4086
alcohol               2.762e-01  2.648e-02  10.429  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.4199185)

    Null deviance: 1042.17  on 1598  degrees of freedom
Residual deviance:  666.41  on 1587  degrees of freedom
AIC: 3164.3

Number of Fisher Scoring iterations: 2
```

Podem veure que en el model generat, citric.acid, residual.sugar, fixed.acidity i density tenen un nivell de significació superior a 0.05, pel que generarem un segon model per veure si ajusta millor els valors:

```
{r lm2}
m2 <- glm(
  formula =
    quality
  ~ pH
  + sulphates
  + volatile.acidity
  + chlorides
  + free.sulfur.dioxide
  + total.sulfur.dioxide
  + alcohol,
  data = df
)

summary(m2)
```

Aquest és el resultat:

```
Call:
glm(formula = quality ~ pH + sulphates + volatile.acidity + chlorides +
    free.sulfur.dioxide + total.sulfur.dioxide + alcohol, data = df)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.68918  -0.36757  -0.04653   0.46081   2.02954

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           4.4300987  0.4029168  10.995  < 2e-16 ***
pH                   -0.4826614  0.1175581  -4.106 4.23e-05 ***
sulphates             0.8826651  0.1099084   8.031 1.86e-15 ***
volatile.acidity     -1.0127527  0.1008429 -10.043  < 2e-16 ***
chlorides            -2.0178138  0.3975417  -5.076 4.31e-07 ***
free.sulfur.dioxide   0.0050774  0.0021255   2.389    0.017 *
total.sulfur.dioxide -0.0034822  0.0006868  -5.070 4.43e-07 ***
alcohol               0.2893028  0.0167958  17.225  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.4195707)

    Null deviance: 1042.17  on 1598  degrees of freedom
Residual deviance:  667.54  on 1591  degrees of freedom
AIC: 3159

Number of Fisher Scoring iterations: 2
```

Podem veure que el valor de AIC es menor (de 3164.3 a 3159), per tant podem dir que el model és millor. El motiu de que algunes variables no siguin necessàries en el model és perquè es tracten de variables redundants amb altres variables ja existents. Per exemple, density està correlacionat amb alcohol, i cítric.acid està correlacionat amb fixed.acidity, volatile.acidity i pH. Aquestes són les correlacions assumint, incorrectament, homoscedasticitat entre variables:

```
{r corr_full}
round(cor(df), 4)
```

```
                      fixed.acidity volatile.acidity citric.acid residual.sugar chlorides free.sulfur.dioxide
fixed.acidity                1.0000          -0.2561      0.6717         0.1148    0.0937              -0.1538
volatile.acidity            -0.2561           1.0000     -0.5525         0.0019    0.0613              -0.0105
citric.acid                  0.6717          -0.5525      1.0000         0.1436    0.2038              -0.0610
residual.sugar               0.1148           0.0019      0.1436         1.0000    0.0556               0.1870
chlorides                    0.0937           0.0613      0.2038         0.0556    1.0000               0.0056
free.sulfur.dioxide         -0.1538          -0.0105     -0.0610         0.1870    0.0056               1.0000
total.sulfur.dioxide        -0.1132           0.0765      0.0355         0.2030    0.0474               0.6677
density                      0.6680           0.0220      0.3649         0.3553    0.2006              -0.0219
pH                          -0.6830           0.2349     -0.5419        -0.0857   -0.2650               0.0704
sulphates                    0.1830          -0.2610      0.3128         0.0055    0.3713               0.0517
alcohol                     -0.0617          -0.2023      0.1099         0.0421   -0.2211              -0.0694
quality                      0.1241          -0.3906      0.2264         0.0137   -0.1289              -0.0507
                      total.sulfur.dioxide density      pH sulphates alcohol quality
fixed.acidity                  -0.1132  0.6680 -0.6830    0.1830 -0.0617  0.1241
volatile.acidity                0.0765  0.0220  0.2349   -0.2610 -0.2023 -0.3906
citric.acid                     0.0355  0.3649 -0.5419    0.3128  0.1099  0.2264
residual.sugar                  0.2030  0.3553 -0.0857    0.0055  0.0421  0.0137
chlorides                       0.0474  0.2006 -0.2650    0.3713 -0.2211 -0.1289
free.sulfur.dioxide             0.6677 -0.0219  0.0704    0.0517 -0.0694 -0.0507
total.sulfur.dioxide            1.0000  0.0713 -0.0665    0.0429 -0.2057 -0.1851
density                         0.0713  1.0000 -0.3417    0.1485 -0.4962 -0.1749
pH                             -0.0665 -0.3417  1.0000   -0.1966  0.2056 -0.0577
sulphates                       0.0429  0.1485 -0.1966    1.0000  0.0936  0.2514
alcohol                        -0.2057 -0.4962  0.2056    0.0936  1.0000  0.4762
quality                        -0.1851 -0.1749 -0.0577    0.2514  0.4762  1.0000
```

5. Representació dels resultats a partir de taules i gràfiques. Aquest apartat es pot respondre al llarg de la pràctica, sense la necessitat de concentrar totes les representacions en aquest punt de la pràctica.
   Hem presentat tots els resultats obtinguts al llarg de la pràctica.

6. Resolució del problema. A partir dels resultats obtinguts, quines són les conclusions? Els resultats permeten respondre al problema?
   El que volíem respondre era quins atributs afectaven a la qualitat d'un vi i com, per saber si era possible predir-ne el valor. Hem vist, per exemple, que els vins d'alta qualitat presenten una major concentració de cítrics. Amb el model lineal que hem generat observem que la concentració de sucre, sulfats i acidesa en el vi afecten positivament la qualitat del mateix, mentre que una alta concentració de sal i diòxid de sulfur, juntament amb una alta densitat, en disminueixen la qualitat. També podem quantificar com aquestes variacions afecten la qualitat: per exemple, per cada unitat de sulfats que augmentem, esperem augmentar la qualitat en 0.88. No obstant, aquests valors s'han de prendre en cura, ja que en el cas dels sulfats esperem valors entre 0.5 i 2, això no vol dir que si tinguéssim un valor de sulfats de 15, la qualitat del vi seria de 10, ja que hem ajustat els valors usant un model de predicció lineal, quan segurament el comportament de la variable és més complex.
   Podem provar de predir resultats amb el següent codi:

```{r predict}
predict(
  m2,
  data.frame(
    pH = 3.51,
    sulphates = 0.56,
    volatile.acidity = 0.7,
    chlorides = 0.076,
    free.sulfur.dioxide = 11,
    total.sulfur.dioxide = 34,
    alcohol = 9.4
  )
)
```

```
        1
5.024869
```

En aquest cas el valor predit ha sigut 5.02, quan el valor real era 5, pel que sembla força ben ajustat al model.

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.
   S'ha inclòs tot el codi usat en la pràctica, però també es pot trobar el fitxer .Rmd en el següent repositori: https://github.com/rodesdecarro/TipologiaCicleVidaDades-PRA2.