Creació de la visualització i lliurament del projecte (Pràctica II)

Adrià Vidal de Palol

Semestre Tardor 22/23

Contents

---

0. Introducció

Aquest document presenta la solució a la pràctica 2 de l'assignatura Visualització de dades del Màster de Ciència de Dades de la UOC.

---

1. Descripció del conjunt de dades

Aquest data set es troba en la web de repositoris KaggleI. El data set utilitzat, https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset, recull un conjunt de variables relacionades amb les malalties del cor de pacients de l'hospital de Cleveland. Variables com el colesterol o la quantitat de sucre en sang, son exemples considerats com a factors de risc importants de les malalties coronàries i que podem relacionar-les en aquest data set.

Engloba un total de 14 variables. Els camps són els següents:

- Age: edat dels pacients
- Sex: sexe
- Cp: tipus de dolor al pit
- Trestbps: Pressió arterial en repòs
- Chol :Colesterol sèric en mg / dl
- Fbs: sucre en sang en dejú> 120 mg / dl
- Restecg: resultats electrocardiografies en repòs
- Thalach: freqüència cardíaca màxima aconseguida
- Exang: angina induïda per l'exercici
- Oldpeak: Depressió ST induïda per l'exercici en relació amb el descans
- Slope: el pendent del segment ST d'exercici màxim
- Ca: nombre de vasos sanguinis principals (0-3) acolorits per fluoroscòpia

- Thal: 3 = normal; 6 = defecte solucionat; 7 = defecte reversible
- Num: diagnòstic de malaltia cardíaca (estat de la malaltia angiografia)

En el nostra data set formulem les preguntes següents:

- Quines relació hi ha entre el sexe del pacient i les malalties coronàries ? I per grups de edat?
- Quines relacions podem trobar entre les diferents variables dels pacients? Per exemple pacients amb colesterol alt i amb sucre alt.

Per poder respondre aquestes preguntes, caldrà realitzar un procés simple de integració, selecció, neteja i creació de un nou dataset.

2. Integració i selecció de les dades d'interès

Importarem el fitxer i formatejarem les dades.

```
library(dplyr)
data <- read.csv("./processed.cleveland.data", header=FALSE, sep = ",")
names(data) <- c("Edat", "Sexe", "Tipus", "Pressio Arterial", "Colesterol", "Sucre_alt", "restecg",
                 "Frequencia Cardiaca Maxima", "exang", "oldpeak", "slope", "ca", "thal", "num")
data$Sexe <- as.factor(ifelse(data$Sexe ==0, "Dona", "Home"))
data$Tipus<- as.factor(
  ifelse(data$Tipus==1, "Tipica angina",
         ifelse(data$Tipus==2, "Atipica angina",
                ifelse(data$Tipus==3, "No dolor angina",
                       ifelse(data$Tipus== 4 ,"Asintomatic", NA )))))
data$Sucre_alt <- as.logical(data$Sucre_alt)
data$restecg <- as.factor(data$restecg)
data$exang <- as.logical(data$exang)
data$slope <- as.factor(
  ifelse(data$slope ==1, "upsloping",
         ifelse(data$slope == 2, "flat",
                ifelse(data$slope == 3, "downsloping", NA))))
data$ca[data$ca == "?"] <- NA
data$ca <- as.factor(signif(as.numeric(data$ca)))
data$thal[data$thal =="?"] <-NA
data$thal <- as.numeric(data$thal)
data$thal <- as.factor(
  ifelse(data$thal == 3, "normal",
         ifelse(data$thal== 6, "fixed defect",
                ifelse(data$thal== 7, "reversable defect",NA))))
data$num <- as.factor(
  ifelse(data$num ==  0, "Saludable",  "No Saludable" ))
summary(data)
```

```
##       Edat         Sexe                 Tipus     Pressio Arterial
##  Min.   :29.00   Dona: 97   Asintomatic    :144   Min.   : 94.0
##  1st Qu.:48.00   Home:206   Atipica angina : 50   1st Qu.:120.0
##  Median :56.00              No dolor angina: 86   Median :130.0
##  Mean   :54.44              Tipica angina  : 23   Mean   :131.7
##  3rd Qu.:61.00                                    3rd Qu.:140.0
##  Max.   :77.00                                    Max.   :200.0
##    Colesterol    Sucre_alt       restecg Frequencia Cardiaca Maxima
##  Min.   :126.0   Mode :logical   0:151   Min.   : 71.0
##  1st Qu.:211.0   FALSE:258       1:  4   1st Qu.:133.5
##  Median :241.0   TRUE :45        2:148   Median :153.0
```

```
##  Mean   :246.7                          Mean   :149.6
##  3rd Qu.:275.0                          3rd Qu.:166.0
##  Max.   :564.0                          Max.   :202.0
##    exang            oldpeak             slope        ca
##  Mode :logical   Min.   :0.00   downsloping: 21   0   :176
##  FALSE:204       1st Qu.:0.00   flat       :140   1   : 65
##  TRUE :99        Median :0.80   upsloping  :142   2   : 38
##                  Mean   :1.04                      3   : 20
##                  3rd Qu.:1.60                      NA's:  4
##                  Max.   :6.20
##                 thal               num
##  fixed defect     : 18   No Saludable:139
##  normal           :166   Saludable   :164
##  reversable defect:117
##  NA's             :  2
##
##
```

Fer un cop d'ull al conjunt de les dades

```
str(data)
```

```
## 'data.frame':    303 obs. of  14 variables:
##  $ Edat                       : num  63 67 67 37 41 56 62 57 63 53 ...
##  $ Sexe                       : Factor w/ 2 levels "Dona","Home": 2 2 2 2 1 2 1 1 2 2 ...
##  $ Tipus                      : Factor w/ 4 levels "Asintomatic",..: 4 1 1 3 2 2 1 1 1 1 ...
##  $ Pressio Arterial           : num  145 160 120 130 130 120 140 120 130 140 ...
##  $ Colesterol                 : num  233 286 229 250 204 236 268 354 254 203 ...
##  $ Sucre_alt                  : logi  TRUE FALSE FALSE FALSE FALSE FALSE ...
##  $ restecg                    : Factor w/ 3 levels "0","1","2": 3 3 3 1 3 1 3 1 3 3 ...
##  $ Frequencia Cardiaca Maxima: num  150 108 129 187 172 178 160 163 147 155 ...
##  $ exang                      : logi  FALSE TRUE TRUE FALSE FALSE FALSE ...
##  $ oldpeak                    : num  2.3 1.5 2.6 3.5 1.4 0.8 3.6 0.6 1.4 3.1 ...
##  $ slope                      : Factor w/ 3 levels "downsloping",..: 1 2 2 1 3 3 1 3 2 1 ...
##  $ ca                         : Factor w/ 4 levels "0","1","2","3": 1 4 3 1 1 1 3 1 2 1 ...
##  $ thal                       : Factor w/ 3 levels "fixed defect",..: 1 2 3 2 2 2 2 2 3 3 ...
##  $ num                        : Factor w/ 2 levels "No Saludable",..: 2 1 1 2 2 2 1 2 1 1 ...
```

```
data[data$Edat <= 18, "age_group"] <- "0-18"
data[data$Edat > 18 & data$Edat <= 28, "age_group"] <- "Joves (18-28)"
data[data$Edat > 29 & data$Edat <= 44, "age_group"] <- "Edat Mitjana (29-44)"
data[data$Edat > 44 & data$Edat <= 64, "age_group"] <- "Edat Majors (44-64)"
data[data$Edat > 64, "age_group"] <- "Ancians (> 64)"

data <- data[, c(15,1,3,2,4,5,6,7,8,9,10,11,12,13,14)]
(colSums(is.na(data)))
```

```
##                  age_group                       Edat
##                          1                          0
##                      Tipus                      Sexe
##                          0                          0
##           Pressio Arterial                Colesterol
##                          0                          0
##                  Sucre_alt                   restecg
##                          0                          0
```

```
## Frequencia Cardiaca Maxima                    exang
##                          0                        0
##                    oldpeak                    slope
##                          0                        0
##                         ca                     thal
##                          4                        2
##                        num
##                          0
```

3. Neteja de les dades

Apliquem una funció més específica per mostrar només els elements buits per variable.

```
(colSums(is.na(data)))
```

```
##                   age_group                        Edat
##                           1                           0
##                       Tipus                       Sexe
##                           0                           0
##            Pressio Arterial                  Colesterol
##                           0                           0
##                  Sucre_alt                     restecg
##                           0                           0
## Frequencia Cardiaca Maxima                       exang
##                           0                           0
##                     oldpeak                       slope
##                           0                           0
##                          ca                        thal
##                           4                           2
##                         num
##                           0
```

4. Creació del nou dataset

Exportem el nou dataset per poder analitzar i visualitzar les dades.

```
write.csv(data,file='/Users/vidalot/Desktop/new_data.csv', row.names=FALSE)
```