Adults

Autors: Josep Alòs Pascual i Daniel Galan Vilella

Gener 2020

Contents

Descripció del dataset

El dataset que utilitzarem per aquesta pràctica és el Adult Data Set que trobem en el següent enllaç: https://archive.ics.uci.edu/ml/datasets/Adult

Informació sobre els atributs:

Llista d'atributs i el seu tipus:

- age: Variable continua.
- workclass: Categòrica. Possibles valors: *Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.*
- fnlwgt: Variable continua.
- education: Categòrica. Possibles valors: *Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.*
- education-num: Variable numèrica ordinal.
- marital-status: Categòrica. Possibles valors: *Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.*
- occupation: Categòrica. Possibles valors: *Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.*
- relationship: Categòrica. Possibles valors: *Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.*
- race: Categòrica. Possibles valors: *White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.*
- sex: Categòrica: *Female, Male.*

- capital-gain: Variable continua.
- capital-loss: Variable continua.
- hours-per-week: Variable continua.
- native-country: Categòrica. Possibles valors: *United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.*
- Income: Categòrica. Possibles valors: $> 50\ K$, $<= 50\ K$. Aquesta variable és la que s'intenta predir en l'article original del conjunt de dades.

L'objectiu d'aquest projecte és estudiar la relació que hi ha entre els diferents atributs censals de la població d'Estats Units, i addicionalment com es relacionen amb si tenen uns ingressos superiors o inferiors a 50.000 dòlars anuals.

El principal estudi que es vol fer és trobar quina variable és més significativa a l'hora d'explicar si els ingressos superen aquest llindar o no. Addicionalment, es buscaràn correlacions entre els diferents atributs, com per exemple les hores treballades a la setmana i els guanys de capital. Per últim, s'intentarà trobar regles que ens intentin explicar si una persona guanyarà més o menys de llindar de 50.000 dòlars anuals.

Itegració i selecció de les dades d'interès a analitzar

Primer de tot carreguem les dades del conjunt de dades i n'anomenem les columnes:

```
# Carreguem el joc de dades
dadesAdult <- read.csv('adult.data',stringsAsFactors = FALSE,
                       header = FALSE, strip.white = TRUE)

# Noms dels atributs
names(dadesAdult) <- c("age", "workclass", "fnlwgt", "education", "education-num",
                       "marital-status", "occupation", "relationship", "race",
                       "sex", "capital-gain", "capital-loss", "hour-per-week",
                       "native-country", "income")

numericalCols <- c("age", "fnlwgt", "education-num", "capital-gain", "capital-loss",
                   "hour-per-week")
categoricalCols <- c("workclass", "education", "marital-status",
                     "occupation", "relationship", "race", "sex",
                     "native-country", "income")
```

Un cop carregades les dades, assignem el tipus correcte a les columnes:

```
# Convertim a factors les variables categòriques
for (i in categoricalCols){
  dadesAdult[,i] <- as.factor(dadesAdult[,i])
}

summary(dadesAdult)
```

```
##       age                    workclass         fnlwgt
##  Min.   :17.00   Private         :22696   Min.   :  12285
##  1st Qu.:28.00   Self-emp-not-inc: 2541   1st Qu.: 117827
##  Median :37.00   Local-gov       : 2093   Median : 178356
##  Mean   :38.58   ?               : 1836   Mean   : 189778
```

```
##  3rd Qu.:48.00   State-gov       : 1298   3rd Qu.: 237051
##  Max.   :90.00   Self-emp-inc    : 1116   Max.   :1484705
##                  (Other)         :  981
##         education     education-num                  marital-status
##  HS-grad     :10501   Min.   : 1.00   Divorced             : 4443
##  Some-college: 7291   1st Qu.: 9.00   Married-AF-spouse    :   23
##  Bachelors   : 5355   Median :10.00   Married-civ-spouse   :14976
##  Masters     : 1723   Mean   :10.08   Married-spouse-absent:  418
##  Assoc-voc   : 1382   3rd Qu.:12.00   Never-married        :10683
##  11th        : 1175   Max.   :16.00   Separated            : 1025
##  (Other)     : 5134                   Widowed              :  993
##            occupation           relationship                   race
##  Prof-specialty :4140   Husband       :13193   Amer-Indian-Eskimo:  311
##  Craft-repair   :4099   Not-in-family : 8305   Asian-Pac-Islander: 1039
##  Exec-managerial:4066   Other-relative:  981   Black             : 3124
##  Adm-clerical   :3770   Own-child     : 5068   Other             :  271
##  Sales          :3650   Unmarried     : 3446   White             :27816
##  Other-service  :3295   Wife          : 1568
##  (Other)        :9541
##      sex          capital-gain    capital-loss    hour-per-week
##  Female:10771   Min.   :    0   Min.   :   0.0   Min.   : 1.00
##  Male  :21790   1st Qu.:    0   1st Qu.:   0.0   1st Qu.:40.00
##                 Median :    0   Median :   0.0   Median :40.00
##                 Mean   : 1078   Mean   :  87.3   Mean   :40.44
##                 3rd Qu.:    0   3rd Qu.:   0.0   3rd Qu.:45.00
##                 Max.   :99999   Max.   :4356.0   Max.   :99.00
##
##        native-country    income
##  United-States:29170   <=50K:24720
##  Mexico       :  643   >50K : 7841
##  ?            :  583
##  Philippines  :  198
##  Germany      :  137
##  Canada       :  121
##  (Other)      : 1709
```

Aquest dataset té una peculiaritat, i és que ens proporciona la variable *fnlwgt*, que ens indica el valor estimat de persones en el cens que són similars al registre actual. Per tant, s'hauria de tenir en compte aquest valor quan es fan estudis de la distribució de les dades, per exemple. Una forma de tenir en compte aquesta dada és repetir cada registre aquest nombre de vegades, potser afegint una mica de soroll per evitar tenir molts valors idèntics. Tot i això, de cara a aquest estudi, no es tindrà en compte aquesta variable. Tampoc es tindran en compte les variables *relationship* (que es pot deduïr de *marital status*), ni *education*, equivalent a la variable *education_num*. S'ha optat per mantenir la variable numèrica sobre l'educació ja que, al estar expressada de forma numèrica, ens permet mantenir l'ordre dels nivells d'estudis. Per tant, eliminem les columnes esmentades:

```
numericalCols <- c("age", "education-num", "capital-gain", "capital-loss",
                   "hour-per-week")
categoricalCols <- c("workclass", "marital-status", "occupation", "race",
                     "sex", "native-country", "income")
dadesAdult$relationship <- NULL
dadesAdult$education <- NULL
dadesAdult$fnlwgt <- NULL
```

Neteja de les dades

Identificació d'inconsistències

Busquem persones que diuen que no han treballat mai i han reportat N hores per setmana:

```
indices <- which(dadesAdult$workclass == "Never-worked" &
                   dadesAdult$`hour-per-week` > 0)
dadesAdult[indices, c(2, 10)]
```

```
##          workclass hour-per-week
## 5362  Never-worked            40
## 10846 Never-worked            35
## 14773 Never-worked            30
## 20338 Never-worked            10
## 23233 Never-worked            40
## 32305 Never-worked            40
## 32315 Never-worked             4
```

En aquests casos, assignem a 0 el valor de les hores treballades per setmana:

```
dadesAdult$`hour-per-week`[indices] <- 0
```

Identificació de valors buits

Per tal de fer una neteja de les dades i comprovar si existeixen valors buits, comencem mirant aquells que són nulls.

```
colSums(is.na(dadesAdult))
```

```
##            age      workclass  education-num marital-status     occupation
##              0              0              0              0              0
##           race            sex   capital-gain   capital-loss  hour-per-week
##              0              0              0              0              0
## native-country         income
##              0              0
```

Seguim comprovant si existeixen columnes amb una cadena de text buida.

```
colSums(dadesAdult == "", na.rm=TRUE)
```

```
##            age      workclass  education-num marital-status     occupation
##              0              0              0              0              0
##           race            sex   capital-gain   capital-loss  hour-per-week
##              0              0              0              0              0
## native-country         income
##              0              0
```

Finalment, busquem columnes que continguin valors buits indicats amb el valor '?':

```
colSums(dadesAdult == "?", na.rm=TRUE)
```

```
##            age      workclass  education-num marital-status     occupation
##              0           1836              0              0           1843
##           race            sex   capital-gain   capital-loss  hour-per-week
##              0              0              0              0              0
## native-country         income
```

```
##             583               0
```

Veiem que a *workclass*, *occupation*, i a *native-country* ens aparexien valors buits. Eliminem aquests atributs i els assignem com a buits:

```
dadesAdult$workclass[which(dadesAdult$workclass == "?")] <- NA
dadesAdult$occupation[which(dadesAdult$occupation == "?")] <- NA
dadesAdult$`native-country`[which(dadesAdult$`native-country` == "?")] <- NA
```

Comprovem que, efectivament, hem transformat els valors "?" a nuls.

```
colSums(is.na(dadesAdult))
```

```
##             age       workclass  education-num marital-status      occupation
##               0            1836              0              0            1843
##            race             sex   capital-gain   capital-loss  hour-per-week
##               0               0              0              0               0
## native-country          income
##             583               0
```

```
colSums(dadesAdult == "?", na.rm=TRUE)
```

```
##             age       workclass  education-num marital-status      occupation
##               0               0              0              0               0
##            race             sex   capital-gain   capital-loss  hour-per-week
##               0               0              0              0               0
## native-country          income
##               0               0
```

Imputació dels valors buits

Utilitzem el mètode kNN per assignar valors als camps buits, utilitzant els 5 registres més propers:

```
dadesAdult <- kNN(dadesAdult,
                  variable = c("workclass", "occupation", "native-country"),
                  k = 5)

summary(dadesAdult)
```

```
##       age                     workclass     education-num
##  Min.   :17.00   Private         :24355   Min.   : 1.00
##  1st Qu.:28.00   Self-emp-not-inc: 2655   1st Qu.: 9.00
##  Median :37.00   Local-gov       : 2120   Median :10.00
##  Mean   :38.58   State-gov       : 1310   Mean   :10.08
##  3rd Qu.:48.00   Self-emp-inc    : 1137   3rd Qu.:12.00
##  Max.   :90.00   Federal-gov     :  963   Max.   :16.00
##                  (Other)         :   21
##                marital-status           occupation
##  Divorced             : 4443   Prof-specialty :4278
##  Married-AF-spouse    :   23   Craft-repair   :4269
##  Married-civ-spouse   :14976   Exec-managerial:4199
##  Married-spouse-absent:  418   Adm-clerical   :4055
##  Never-married        :10683   Sales          :3937
##  Separated            : 1025   Other-service  :3756
##  Widowed              :  993   (Other)        :8067
##                  race           sex         capital-gain
##  Amer-Indian-Eskimo:  311   Female:10771   Min.   :    0
##  Asian-Pac-Islander: 1039   Male  :21790   1st Qu.:    0
```

```
##  Black              : 3124                  Median :    0
##  Other              :  271                  Mean   : 1078
##  White              :27816                  3rd Qu.:    0
##                                             Max.   :99999
##
##   capital-loss    hour-per-week         native-country    income
##  Min.   :   0.0   Min.   : 0.00   United-States:29736   <=50K:24720
##  1st Qu.:   0.0   1st Qu.:40.00   Mexico       :  658   >50K : 7841
##  Median :   0.0   Median :40.00   Philippines  :  198
##  Mean   :  87.3   Mean   :40.43   Germany      :  137
##  3rd Qu.:   0.0   3rd Qu.:45.00   Canada       :  121
##  Max.   :4356.0   Max.   :99.00   Puerto-Rico  :  114
##                                   (Other)      : 1597
##  workclass_imp   occupation_imp  native-country_imp
##  Mode :logical   Mode :logical   Mode :logical
##  FALSE:30725     FALSE:30718     FALSE:31978
##  TRUE :1836      TRUE :1843      TRUE :583
##
##
##
##
```

Per assegurar la consistència, mirem si les dades imputades incompleixen la condició que hem imposat anteriorment, on es comprova si la gent que mai ha treballat havia imputat hores setmanals.

```
which(dadesAdult$workclass == "Never-worked" & dadesAdult$`hour-per-week` > 0)
```

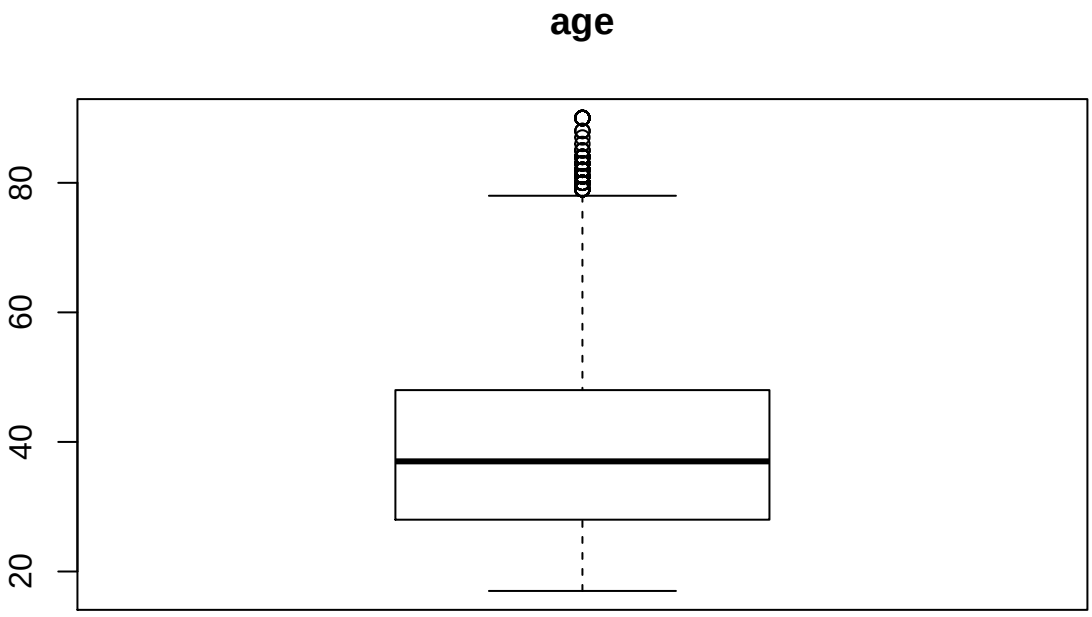
```
## integer(0)
```

Veiem que no hem afegit cap inconsistència a les dades en fer la imputació.

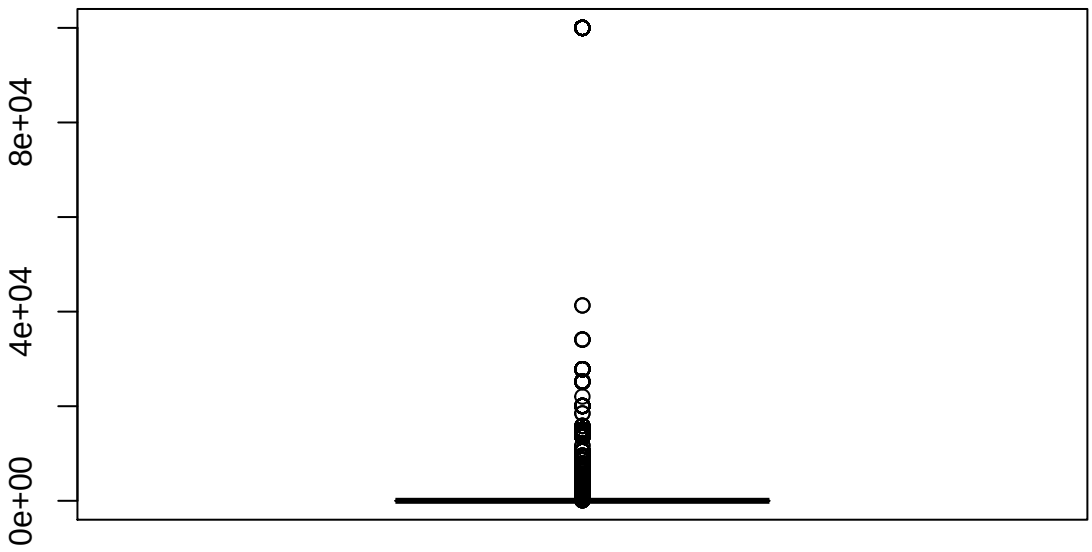
Detecció d'outliers

Busquem valors atípics en el dataset. Per fer-ho, es mostraran les variables continues utilitzant gràfics de caixes. Els valors que siguin menors que $1,5 * Q_1$ o majors que $1,5 * Q_4$ (amb Q_n sent el quartil N dels valors) es consideren outliers.

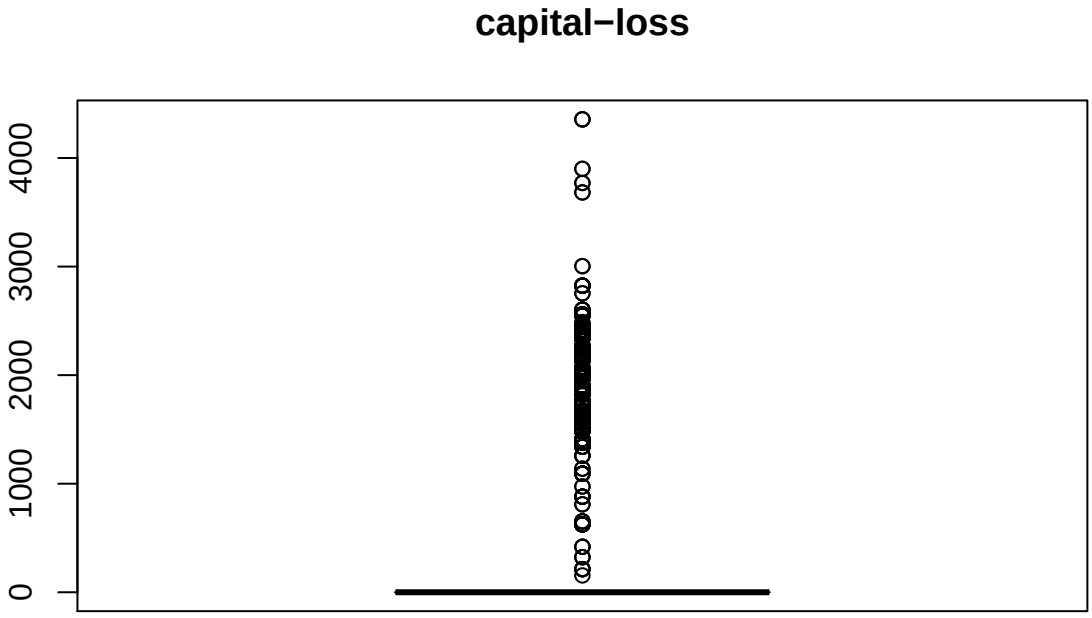
```
continuousAttrs <- c("age", "capital-gain", "capital-loss", "hour-per-week")
for (i in continuousAttrs){
  boxplot(dadesAdult[,i], main=i)
  outliers <- boxplot.stats(dadesAdult[,i])$out
  ran <- range(outliers)
  print(ran)
  print(sprintf("%d outliers", length(outliers)))

  # Uncomment to print the indices of the outliers
  # indices <- which(dadesAdult[,i] %in% outliers)
  # print(paste(c("Their indices are ", indices)))
}
```



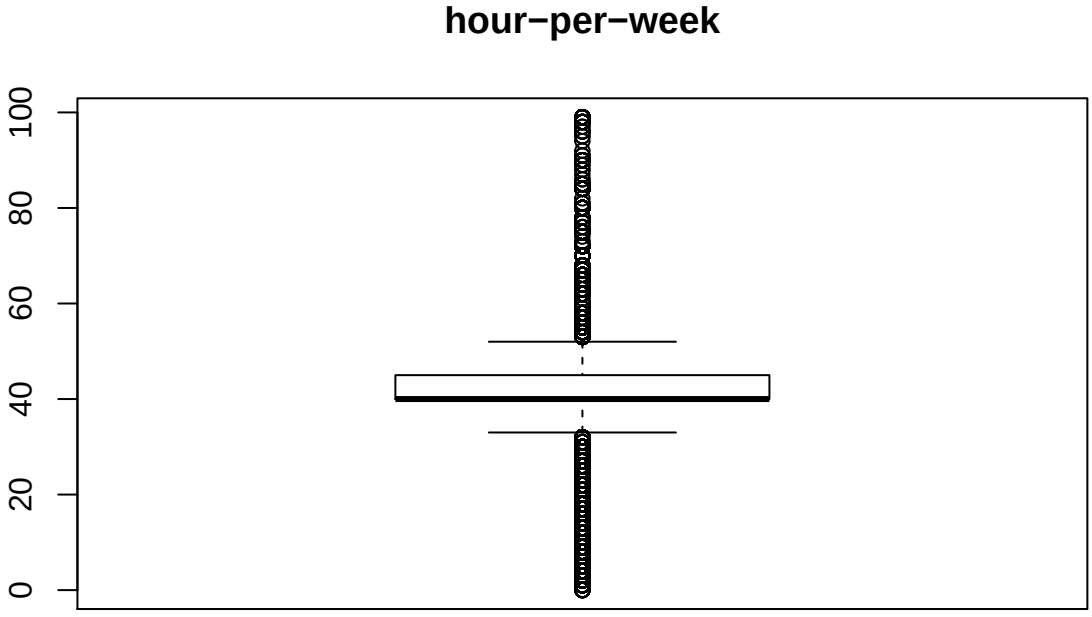

```
## [1] 79 90
## [1] "143 outliers"
```




```
## [1]   114 99999
## [1] "2712 outliers"
```




```
## [1]  155 4356
## [1] "1519 outliers"
```




```
## [1]  0 99
## [1] "9012 outliers"
```

Veiem que hi ha molts valors considerats oultiers amb la condició que prèviament hem especificat. En aquest dataset, però, no els eliminarem ja que esperem trobar unes dades amb unes distribucions amb una variància elevada.

Exportació de les dades preprocessades

```
write.csv(dadesAdult, "Adults_data_clean.csv")
```

Anàlisi de les dades

Selecció dels grups de dades que es volen analitzar/comparar

En aquest estudi s'analitzaran diferents atributs del cens per tal de trobar una explicació a si una persona guanya més o menys de 50.000 dòlars anuals. Per fer aquest estudi, primer de tot s'estudiarà la normalitat i homogeneïtat dels diferents atributs i després s'estudiaran les dades censals a partir d'anàlisis estadístics.

En l'apartat de càrrega de dades, hem s'han eliminat les variables *fnlwgt*, *relationship* i *education*. Totes les altres variables seran utilitzades en els estudis estadístics.

Comprovació de la normalitat i homogeneïtat de la variància

Per tal de comprovar la normalitat en les variables quantitatives, utilitzarem la prova Anderson-Darling. per acabar determinant aquelles variables que no segueixen una distribució normal.

```
alpha = 0.05
col.names = colnames(dadesAdult)

for (i in 1:ncol(dadesAdult)) {
  if (i == 1) cat("Variables que no segueixen una distribució normal:\n")
  if (is.integer(dadesAdult[,i]) | is.numeric(dadesAdult[,i])) {
    p_val = ad.test(dadesAdult[,i])$p.value
    if (p_val < alpha) {
      cat(col.names[i])

      # Format output
      if (i < ncol(dadesAdult) - 1) cat(", ")
      if (i %% 3 == 0) cat("\n")
    }
  }
}
```

```
## Variables que no segueixen una distribució normal:
## age, education-num,
## capital-gain, capital-loss,
## hour-per-week,
```

Una vegada comprovada la normalitat, utilitzarem el test de Fligner-Killeen per comprovar la homogeneïtat de les variances mitjançant la mitjana. Comprovem la homogeneïtat de *capital-gain* amb el *sexe*. Considerem com a hipòtesi nul · la que les variàncies són homogènies, i utilitzem una confiança del 95%.

```
fligner.test(`capital-gain` ~ `sex`, data=dadesAdult)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  capital-gain by sex
## Fligner-Killeen:med chi-squared = 161.04, df = 1, p-value <
## 2.2e-16
```

No podem acceptar la hipotesis nula ja que p és inferior a 0,05 i per tant, les variàncies no són homogènies.

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.

Estudi de l'impacte dels atributs en la variança d'*income*

Primer de tot, intentem trobar quina variable explica més la variança en la variable *income*. Per fer-ho, utilitzarem l'anàlisi PCA (*Principal Component Analysis*) amb les dades numèriques, i l'anàlisi MCA (*Multiple Correspondence Analysis*) per les variables categòriques.

```
dades.pca <- prcomp(dadesAdult[,numericalCols], center=TRUE, scale=TRUE)
summary(dades.pca)
```

```
## Importance of components:
##                           PC1    PC2    PC3    PC4    PC5
## Standard deviation     1.1389 1.0150 0.9868 0.9421 0.9007
## Proportion of Variance 0.2594 0.2061 0.1948 0.1775 0.1623
## Cumulative Proportion  0.2594 0.4655 0.6602 0.8377 1.0000
```

Com que s'han fet servir 5 variables en l'anàlisi PCA, i el resultat són 5 components que expliquen un 90% de la variança, veiem que no ens ha servit per reduïr la dimensionalitat del dataset, però aquestes 5 variables tenen un gran impacte en la variança.

```
dades.mca <- MCA(dadesAdult[,categoricalCols], graph=FALSE)
head(dades.mca$eig)
```

```
##       eigenvalue percentage of variance cumulative percentage of variance
## dim 1  0.3090645              3.004794                           3.004794
## dim 2  0.2525976              2.455810                           5.460604
## dim 3  0.2178067              2.117565                           7.578169
## dim 4  0.1867715              1.815834                           9.394004
## dim 5  0.1794026              1.744192                          11.138195
## dim 6  0.1773855              1.724581                          12.862777
```

En l'anàlisi MCA, en canvi, amb les 5 components principals només podem explicar un 10% de la variança.

Estudi de correlació

Volem estudiar si la variable *capital-gain* i la variable *sex* són independents o no. Per fer-ho, es farà un estudi amb la següent hipòtesi: "els guanys de capital per persones de sexe masculí segueix la mateixa distribució per les persones de sexe femení".

Formalitzem l'estudi estadístic:

$$H_0 : \mu_{masculi} = \mu_{femeni}$$

$$H_1 : \mu_{masculi} \neq \mu_{femeni}$$

Segons els tests de normalitat realitzats anteriorment, no podem assumir normalitat en la variable *capital-gain*. Per tant, utilitzarem el test de suma de rangs de Wilcoxon (o test U de Mann-Whitney) amb una confiança del 95%.

```
gainsHomes <- dadesAdult$`capital-gain`[which(dadesAdult$sex == "Male")]
gainsDones <- dadesAdult$`capital-gain`[which(dadesAdult$sex == "Female")]

wilcox.test(gainsHomes, gainsDones, correct=FALSE)
```

```
##
##  Wilcoxon rank sum test
```

```
##
## data:  gainsHomes and gainsDones
## W = 121950000, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

El valor p és més petit de 0,05; i per tant rebutgem H_0.

Regles del dataset

L'últim estudi que es farà ens servirà per extreure regles explicatives de la variable *income* a partir dels altres atributs. Per fer-ho, utilitzarem el model per crear àrbres de decisió C5.0.

```
y <- dadesAdult[,12]
X <- dadesAdult[,1:11]

model <- C50::C5.0(X, y, rules=TRUE, control=C50::C5.0Control(
  seed=555,
  CF=.01,
  noGlobalPruning=FALSE
))
summary(model)
```

```
##
## Call:
## C5.0.default(x = X, y = y, rules = TRUE, control = C50::C5.0Control(seed
##  = 555, CF = 0.01, noGlobalPruning = FALSE))
##
##
## C5.0 [Release 2.07 GPL Edition]      Tue Jan  7 23:03:28 2020
## -------------------------------
##
## Class specified by attribute `outcome'
##
## Read 32561 cases (12 attributes) from undefined.data
##
## Rules:
##
## Rule 1: (464, lift 1.3)
##  capital-gain > 401
##  capital-gain <= 2993
##  ->  class <=50K  [0.998]
##
## Rule 2: (309, lift 1.3)
##  capital-gain > 3103
##  capital-gain <= 4101
##  ->  class <=50K  [0.997]
##
## Rule 3: (131, lift 1.3)
##  marital-status = Married-civ-spouse
##  capital-loss > 1504
##  capital-loss <= 1762
##  ->  class <=50K  [0.992]
##
## Rule 4: (88, lift 1.3)
##  capital-loss > 1980
```

```
##  capital-loss <= 2163
##  ->  class <=50K  [0.989]
##
## Rule 5: (82, lift 1.3)
##  marital-status = Married-civ-spouse
##  capital-gain > 4416
##  capital-gain <= 5060
##  ->  class <=50K  [0.988]
##
## Rule 6: (503/60, lift 1.2)
##  capital-loss > 625
##  capital-loss <= 1762
##  ->  class <=50K  [0.879]
##
## Rule 7: (31162/6462, lift 1.0)
##  capital-gain <= 6849
##  ->  class <=50K  [0.793]
##
## Rule 8: (1399/20, lift 4.1)
##  capital-gain > 6849
##  ->  class >50K  [0.985]
##
## Rule 9: (585/14, lift 4.0)
##  marital-status = Married-civ-spouse
##  capital-loss > 1762
##  capital-loss <= 1980
##  ->  class >50K  [0.974]
##
## Rule 10: (398/13, lift 4.0)
##  education-num > 12
##  marital-status = Married-civ-spouse
##  capital-loss > 1762
##  ->  class >50K  [0.965]
##
## Rule 11: (26, lift 4.0)
##  capital-gain > 4650
##  capital-gain <= 4787
##  ->  class >50K  [0.964]
##
## Rule 12: (1362/108, lift 3.8)
##  marital-status in {Married-AF-spouse, Married-civ-spouse}
##  capital-gain > 4101
##  ->  class >50K  [0.920]
##
## Rule 13: (97/7, lift 3.8)
##  capital-gain > 2993
##  capital-gain <= 3103
##  ->  class >50K  [0.919]
##
## Rule 14: (53/10, lift 3.3)
##  marital-status in {Divorced, Married-spouse-absent, Never-married,
##                     Separated, Widowed}
##  capital-loss > 2352
##  ->  class >50K  [0.800]
```

```
##
## Rule 15: (3350/688, lift 3.3)
##  age > 28
##  education-num > 12
##  marital-status in {Married-AF-spouse, Married-civ-spouse}
##  occupation in {Exec-managerial, Prof-specialty, Protective-serv, Sales,
##                         Tech-support}
##  hour-per-week > 31
##  ->  class >50K  [0.794]
##
## Rule 16: (3747/924, lift 3.1)
##  age > 33
##  age <= 63
##  workclass in {Federal-gov, Local-gov, Private, Self-emp-inc}
##  education-num > 9
##  marital-status in {Married-AF-spouse, Married-civ-spouse}
##  occupation in {Adm-clerical, Exec-managerial, Prof-specialty,
##                         Protective-serv, Sales, Tech-support}
##  hour-per-week > 34
##  ->  class >50K  [0.753]
##
## Rule 17: (4975/1528, lift 2.9)
##  age > 33
##  workclass in {Federal-gov, Local-gov, Private, Self-emp-inc}
##  marital-status in {Married-AF-spouse, Married-civ-spouse}
##  occupation in {Adm-clerical, Exec-managerial, Prof-specialty,
##                         Protective-serv, Sales, Tech-support}
##  hour-per-week > 34
##  ->  class >50K  [0.693]
##
## Rule 18: (199/65, lift 2.8)
##  capital-loss > 2206
##  ->  class >50K  [0.672]
##
## Rule 19: (41/15, lift 2.6)
##  education-num <= 12
##  marital-status = Married-civ-spouse
##  capital-loss > 1340
##  capital-loss <= 1504
##  ->  class >50K  [0.628]
##
## Default class: <=50K
##
##
## Evaluation on training data (32561 cases):
##
##          Rules
##    ----------------
##      No      Errors
##
##      19 4263(13.1%)   <<
##
##
##     (a)   (b)    <-classified as
```

```
##    ----  ----
##   23523  1197    (a): class <=50K
##    3066  4775    (b): class >50K
##
##
##  Attribute usage:
##
##  100.00% capital-gain
##   20.92% marital-status
##   18.24% age
##   18.24% occupation
##   18.24% hour-per-week
##   15.28% workclass
##   15.03% education-num
##    4.23% capital-loss
##
##
## Time: 0.3 secs
```

El model ens ha generat multiples regles que descriuen el dataset. Per exemple, la regla:

$$4650 < capitalGain <= 3103 \implies income > 50K$$

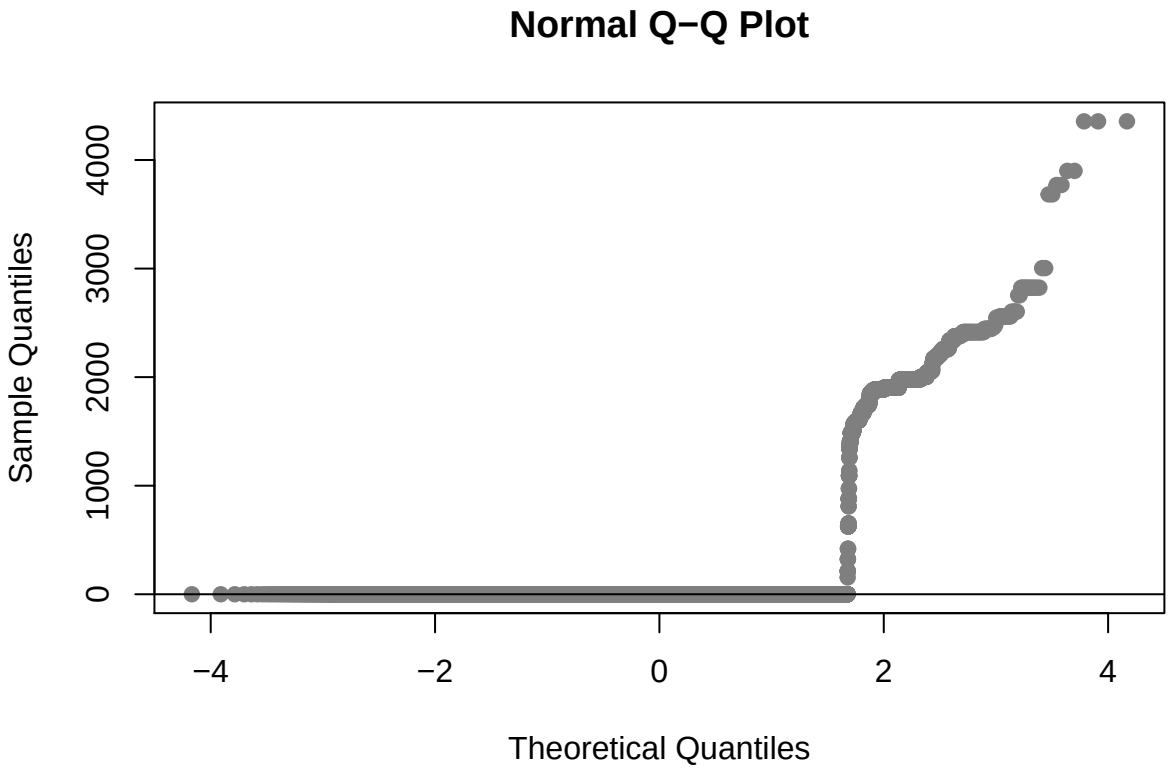
ens indica que quan el capital es troba en aquest rang, amb un 96,4% de probabilitat la persona tindrà uns ingressos anuals superiors a 50.000 dòlars.

Veiem també que l'atribut que s'ha fet servir més vegades per definir el model és el nivell d'educació.

Representació dels resultats a partir de taules i gràfiques

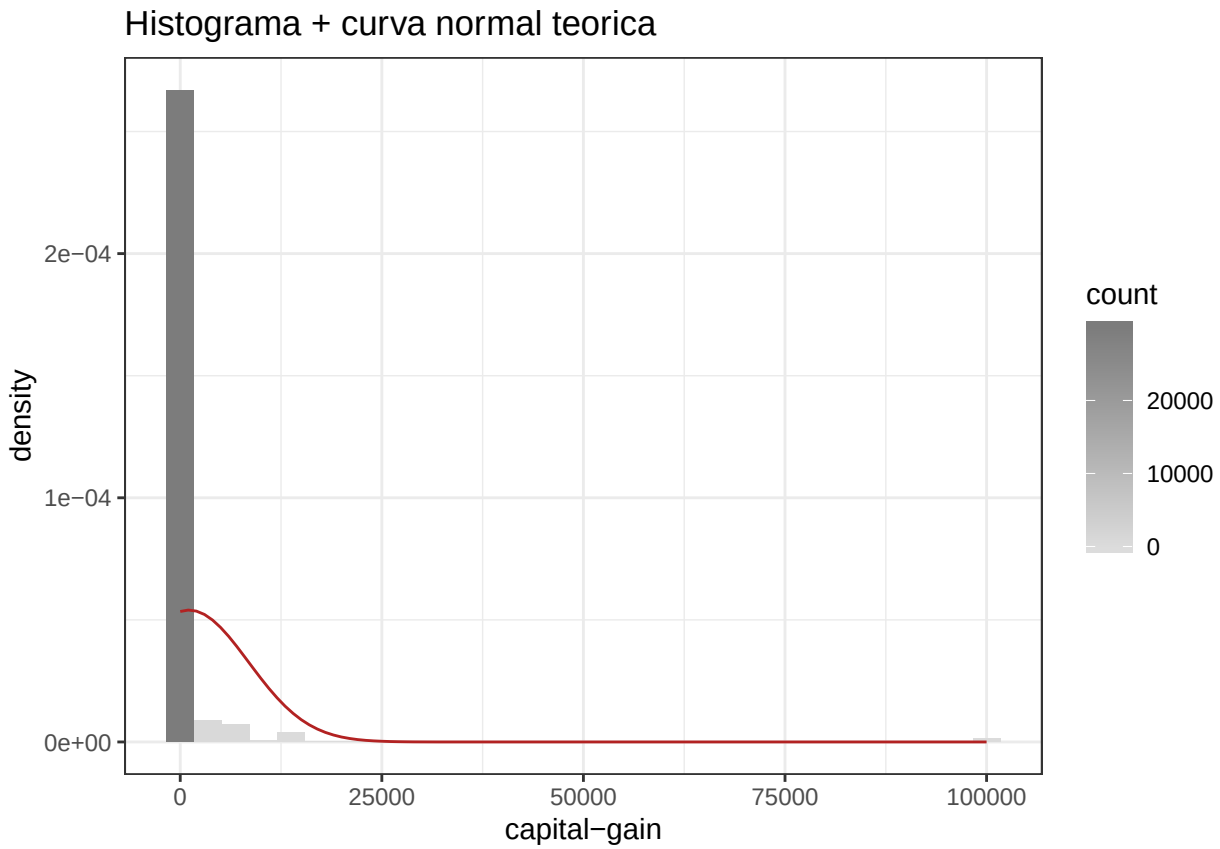
Podem veure en el següent gràfic com *capital-loss* no compleix el principi de normalitat.

```
qqnorm(dadesAdult$`capital-loss`, pch = 19, col = "gray50")
qqline(dadesAdult$`capital-loss`)
```



En el següent gràfic, podem veure com *capital-gain* no es distribueix normalment.

```
library(ggplot2)
ggplot(data = dadesAdult, aes(x = `capital-gain`)) +
  geom_histogram(aes(y = ..density.., fill = ..count..)) +
  scale_fill_gradient(low = "#DCDCDC", high = "#7C7C7C") +
  stat_function(fun = dnorm, colour = "firebrick",
                args = list(mean = mean(dadesAdult$`capital-gain`),
                            sd = sd(dadesAdult$`capital-gain`))) +
  ggtitle("Histograma + curva normal teorica") +
  theme_bw()
```




```
table(dadesAdult$`occupation`)
```

```
##
##                 ?      Adm-clerical      Armed-Forces      Craft-repair
##                 0              4055                 9              4269
##   Exec-managerial   Farming-fishing Handlers-cleaners Machine-op-inspct
##              4199              1053              1453              2116
##     Other-service   Priv-house-serv    Prof-specialty   Protective-serv
##              3756               155              4278               659
##             Sales      Tech-support  Transport-moving
##              3937               960              1662
```

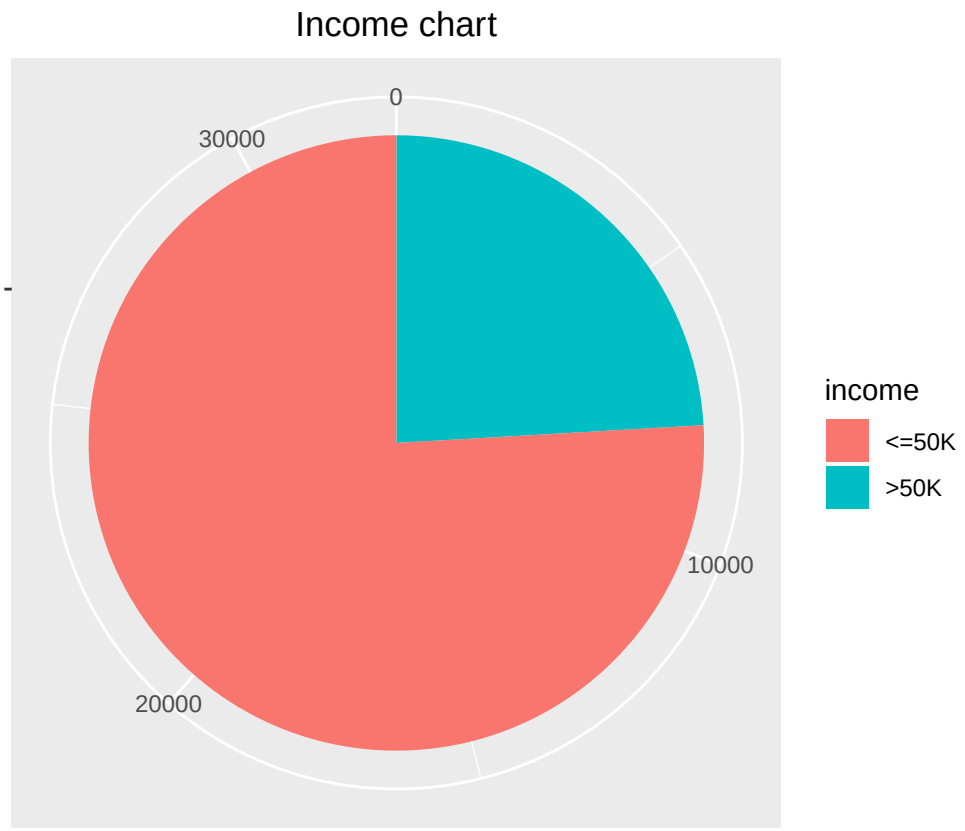
```
table(dadesAdult$`marital-status`)
```

```
##
##              Divorced     Married-AF-spouse    Married-civ-spouse
##                  4443                    23                 14976
## Married-spouse-absent         Never-married             Separated
##                   418                 10683                  1025
##               Widowed
##                   993
```

Igualment, podem veure la freqüència d'*income*:

```
df <- as.data.frame(table(dadesAdult$income))
colnames(df) <- c("income", "freq")
pie <- ggplot(df, aes(x = "", y=freq, fill = factor(income))) +
  geom_bar(width = 1, stat = "identity") +
  theme(axis.line = element_blank(),
```

```
        plot.title = element_text(hjust=0.5)) +
  labs(fill="income",
       x=NULL,
       y=NULL,
       title="Income chart")
pie + coord_polar(theta = "y", start=0)
```



Conclusions

En aquest dataset hem pogut veure com a partir de les variables censals, podem discriminar una població i predir si tenen uns ingressos superiors a 50.000 dòlars anuals o no. Per fer-ho, hem vist que podem utilitzar les variables numèriques (p.e. *capital-loss*) per explicar la variança de les dades. A més, considerant que no hem trobat que les dades segueixen una distribució normal, seria interessant repetir l'estudi tenint en compte la variable *fnlwgt* i comprovar si afecta o no en els resultats. S'ha comprovat també que la majoria de la població representada en aquest cens no arriba a aquest llindar d'ingressos, fet que concorda amb el que s'esperava. Finalment, s'han trobat unes regles explicatives per determinar la relació entre certs atributs i la variable objectiu, com per exemple que el 75% de les persones que ingressen més d'aquest llindar han reportat treballar més de 34 hores setmanals.