

Entrega1

Loading data and Sample selection

Pol Renau
Miguel Angel Merino

Carregar les dades

Les dades es diuen `adult.data` i es troben en el directori actual.

```
df<-read.table("adult.data",header=F, sep=",",fill=FALSE,
strip.white=TRUE,na.string="?")
names(df)<-c("age", "type.employer", "fnlwgt", "education",
"education.num","marital", "occupation",
"relationship", "race","sex", "capital.gain",
"capital.loss",
"hr.per.week", "country", "y.bin")
```

Selecció de la mostra

Inicialitzem un generador aleatori, amb una llavor que es igual a la data de neixament d'un dels integrants del grup, i agafem 5000 observacions de les dades totals

```
set.seed(14121997)
sam<-sort(sample(1:nrow(df),5000))
```

```
str(df)
```

```
## 'data.frame':    32561 obs. of  15 variables:
## $ age           : int  39 50 38 53 28 37 49 52 31 42 ...
## $ type.employer: Factor w/ 8 levels "Federal-gov",...: 7 6 4 4
4 4 4 6 4 4 ...
## $ fnlwgt        : int  77516 83311 215646 234721 338409 284582
160187 209642 45781 159449 ...
## $ education     : Factor w/ 16 levels "10th","11th",...: 10 10
12 2 10 13 7 12 13 10 ...
## $ education.num: int  13 13 9 7 13 14 5 9 14 13 ...
## $ marital       : Factor w/ 7 levels "Divorced","Married-AF-
spouse",...: 5 3 1 3 3 3 4 3 5 3 ...
## $ occupation    : Factor w/ 14 levels "Adm-clerical",...: 1 4 6
6 10 4 8 4 10 4 ...
## $ relationship  : Factor w/ 6 levels "Husband","Not-in-
family",...: 2 1 2 1 6 6 2 1 2 1 ...
## $ race          : Factor w/ 5 levels "Amer-Indian-Eskimo",...:
5 5 5 3 3 5 3 5 5 5 ...
## $ sex           : Factor w/ 2 levels "Female","Male": 2 2 2 2
1 1 1 2 1 2 ...
## $ capital.gain  : int  2174 0 0 0 0 0 0 0 14084 5178 ...
## $ capital.loss  : int  0 0 0 0 0 0 0 0 0 0 ...
## $ hr.per.week   : int  40 13 40 40 40 40 16 45 50 40 ...
## $ country       : Factor w/ 41 levels "Cambodia","Canada",...:
39 39 39 39 5 39 23 39 39 39 ...
```

```
## $ y.bin : Factor w/ 2 levels "<=50K",">50K": 1 1 1 1 1  
1 1 2 2 2 ...  
  
# Select sample  
df<-df[sam,]
```

Guardar la mostra

Guardarem la mostra com a mostra.RData, en el directori actua, aquest pas el podriem evitar, no obstant el fem perquè creiem que es important saber guardar les dades.

```
save(list="df",file="mostra.RData")
```

Fitxa de dades del cens

Descripció

variables d'entrada:

1. age: continuous.
2. workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
3. fnlwgt: continuous.
4. education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
5. education-num: continuous.
6. marital.status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
7. occupation: 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.
8. relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
9. race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
10. sex: Female, Male.
11. capital.gain: continuous.
12. capital.loss: continuous.
13. hours.per.week: continuous. Numeric target.
14. native.country: 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, Trinidad&Tobago, Peru, Hong, Holand-Netherlands.

15. y.bin: Making more than \$50K per year. Binary target.

Carrega de paquets

Carregarem tots els paquets necessaris per utilitzar al llarg de la pràctica.

```
options(contrasts=c("contr.treatment", "contr.treatment"))

requiredPackages <- c("effects", "FactoMineR", "car",
"factoextra", "ggplot2", "dplyr", "ggmap", "ggthemes", "knitr")
missingPackages <- requiredPackages[!(requiredPackages %in%
installed.packages()[, "Package"])]
if(length(missingPackages)) install.packages(missingPackages)

## Installing package into '/usr/local/lib/R/site-library'
## (as 'lib' is unspecified)

## also installing the dependency 'jpeg'

## Warning in install.packages(missingPackages): installation of
package
## 'jpeg' had non-zero exit status

## Warning in install.packages(missingPackages): installation of
package
## 'ggmap' had non-zero exit status

lapply(requiredPackages, require, character.only = TRUE)

## Loading required package: effects

## Loading required package: carData

## lattice theme set by effectsTheme()
## See ?effectsTheme for details.

## Loading required package: FactoMineR

## Loading required package: car

## Registered S3 methods overwritten by 'car':
##   method                               from
##   influence.merMod                      lme4
##   cooks.distance.influence.merMod      lme4
##   dfbeta.influence.merMod              lme4
##   dfbetas.influence.merMod             lme4
```

```
## Loading required package: factoextra
## Loading required package: ggplot2
## Welcome! Related Books: `Practical Guide To Cluster Analysis
in R` at https://goo.gl/13EFCZ
## Loading required package: dplyr
##
## Attaching package: 'dplyr'
## The following object is masked from 'package:car':
##
##      recode
## The following objects are masked from 'package:stats':
##
##      filter, lag
## The following objects are masked from 'package:base':
##
##      intersect, setdiff, setequal, union
## Loading required package: ggmap
## Warning in library(package, lib.loc = lib.loc, character.only
= TRUE,
## logical.return = TRUE, : there is no package called 'ggmap'
## Loading required package: ggthemes
## Loading required package: knitr
## [[1]]
## [1] TRUE
##
## [[2]]
## [1] TRUE
##
## [[3]]
## [1] TRUE
##
## [[4]]
## [1] TRUE
##
## [[5]]
## [1] TRUE
##
## [[6]]
## [1] TRUE
##
## [[7]]
```

```
## [1] FALSE
##
## [[8]]
## [1] TRUE
##
## [[9]]
## [1] TRUE
```

Carregar mostra

Carreguem el model previament creat.

```
# Clear objects
rm(list=ls())
# Clear plots
if(!is.null(dev.list())) dev.off()

## null device
##          1

# Command or Windows-like method
load("mostra.RData")
summary(df)
```

| ## | age | type.employer | fnlwgt |
|-------------------|------------|-----------------------|------------------------|
| ## Min. | :17.0 | Private :3468 | Min. : 18827 |
| ## 1st Qu.: | 27.0 | Self-emp-not-inc: 376 | 1st Qu.: 118008 |
| ## Median : | 37.0 | Local-gov : 327 | Median : 178950 |
| ## Mean : | 38.7 | State-gov : 205 | Mean : 192215 |
| ## 3rd Qu.: | 48.0 | Self-emp-inc : 175 | 3rd Qu.: 241215 |
| ## Max. : | 90.0 | (Other) : 141 | Max. : 1268339 |
| ## | | NA's : 308 | |
| ## | education | education.num | marital |
| ## HS-grad | :1621 | Min. : 1.00 | Divorced : |
| 701 | | | |
| ## Some-college: | 1096 | 1st Qu.: 9.00 | Married-AF-spouse : |
| 1 | | | |
| ## Bachelors | : 793 | Median :10.00 | Married-civ- |
| spouse | :2283 | | |
| ## Masters | : 271 | Mean :10.04 | Married-spouse-absent: |
| 72 | | | |
| ## Assoc-voc | : 228 | 3rd Qu.:12.00 | Never- |
| married | :1606 | | |
| ## 11th | : 185 | Max. :16.00 | Separated : |
| 167 | | | |
| ## (Other) | : 806 | | Widowed : |
| 170 | | | |
| ## | occupation | relationship | |
| ## | race | | |
| ## Prof-specialty | : 635 | Husband :1987 | Amer-Indian- |

```

Eskimo: 45
## Exec-managerial: 624 Not-in-family :1315 Asian-Pac-
Islander: 154
## Craft-repair : 595 Other-relative: 169 Black
: 507
## Adm-clerical : 591 Own-child : 758 Other
: 34
## Sales : 565 Unmarried : 505 White
:4260
## (Other) :1682 Wife : 266

## NA's : 308

## sex capital.gain capital.loss hr.per.week
## Female:1681 Min. : 0 Min. : 0.00 Min. : 1.00
## Male :3319 1st Qu.: 0 1st Qu.: 0.00 1st Qu.:40.00
## Median : 0 Median : 0.00 Median :40.00
## Mean : 1073 Mean : 94.46 Mean :40.43
## 3rd Qu.: 0 3rd Qu.: 0.00 3rd Qu.:45.00
## Max. :99999 Max. :3900.00 Max. :99.00

##
## country y.bin
## United-States:4488 <=50K:3800
## Mexico : 94 >50K :1200
## Germany : 27
## Philippines : 26
## Canada : 24
## (Other) : 253
## NA's : 88

```

Algunes funcions útils

Definim totes les funcions que ens podran ser utils al llarg de la pràctica.

```

calcQ <- function(x) {
  s.x <- summary(x)
  iqr<-s.x[5]-s.x[2]
  list(souti=s.x[2]-3*iqr, mouti=s.x[2]-1.5*iqr, min=s.x[1],
  q1=s.x[2], q2=s.x[3],

```

```

        q3=s.x[5], max=s.x[6], mouts=s.x[5]+1.5*iqr,
souts=s.x[5]+3*iqr ) }

countNA <- function(x) {
  mis_x <- NULL
  for (j in 1:ncol(x)) {mis_x[j] <- sum(is.na(x[,j])) }
  mis_x <- as.data.frame(mis_x)
  rownames(mis_x) <- names(x)
  mis_i <- rep(0,nrow(x))
  for (j in 1:ncol(x)) {mis_i <- mis_i + as.numeric(is.na(x[,j]))}
}
  list(mis_col=mis_x,mis_ind=mis_i) }

```

Preparació de les dades

Preparació de les dades, separem entre aquelles variables que tenen un valor numéric i aquelles que són descriptives.

```

vars_con<-names(df)[c(1,3,5,11:13)];
vars_dis<-names(df)[c(2,4,6:10,14:15)];

```

`summary(df[,vars_con])` *# Example of descriptive for numeric variables*

```

##      age          fnlwgt      education.num
capital.gain
##  Min.   : 0 Min.   : 18827 Min.   : 1.00 Min.   :
0
##  1st Qu.:27.0 1st Qu.: 118008 1st Qu.: 9.00 1st Qu.:
0
##  Median :37.0 Median : 178950 Median :10.00 Median :
0
##  Mean   :38.7 Mean   : 192215 Mean   :10.04 Mean   :
1073
##  3rd Qu.:48.0 3rd Qu.: 241215 3rd Qu.:12.00 3rd Qu.:
0
##  Max.   :90.0 Max.   :1268339 Max.   :16.00
Max.   :99999
##  capital.loss      hr.per.week
##  Min.   : 0.00 Min.   : 1.00
##  1st Qu.: 0.00 1st Qu.:40.00
##  Median : 0.00 Median :40.00
##  Mean   : 94.46 Mean   :40.43
##  3rd Qu.: 0.00 3rd Qu.:45.00
##  Max.   :3900.00 Max.   :99.00

```

`summary(df[,vars_dis])`

```

##      type.employer      education
marital

```



```

## Private      :3468  HS-grad      :1621  Divorced
: 701
## Self-emp-not-inc: 376  Some-college:1096  Married-AF-spouse
: 1
## Local-gov    : 327  Bachelors    : 793  Married-civ-
spouse :2283
## State-gov    : 205  Masters      : 271  Married-spouse-
absent: 72
## Self-emp-inc  : 175  Assoc-voc    : 228  Never-married
:1606
## (Other)      : 141  11th         : 185  Separated
: 167
## NA's         : 308  (Other)      : 806  Widowed
: 170
##              occupation      relationship
race
## Prof-specialty : 635  Husband      :1987  Amer-Indian-
Eskimo: 45
## Exec-managerial: 624  Not-in-family :1315  Asian-Pac-
Islander: 154
## Craft-repair   : 595  Other-relative: 169  Black
: 507
## Adm-clerical   : 591  Own-child    : 758  Other
: 34
## Sales          : 565  Unmarried    : 505  White
:4260
## (Other)        :1682  Wife         : 266

## NA's          : 308

##              sex              country      y.bin
## Female:1681  United-States:4488  <=50K:3800
## Male :3319   Mexico          : 94  >50K :1200
##              Germany          : 27
##              Philippines       : 26
##              Canada           : 24
##              (Other)          : 253
##              NA's            : 88

```

Preparació dels factors

En aquest apartat realitzarem la reagrupació d'aquells factors en classes més generals, això només ho farem per aquelles variables que hem cregut necessaries, de reagrupar en altres classes.

type.employer

Desició conceptual:

1. Civil => Federal, local and state gov

2. Private
3. SelfEm => Treballadors autònoms amb ingressos.
4. Other => Self-emp-not-inc, Never-worked, Without-pay

```
par(mfrow=c(1,2))
```

```
levels(df$type.employer)
```

```
## [1] "Federal-gov"      "Local-gov"        "Never-worked"
## [4] "Private"          "Self-emp-inc"     "Self-emp-not-inc"
## [7] "State-gov"        "Without-pay"
```

```
barplot(table(df$type.employer),main="Original",col=rainbow(12))
table(df$type.employer)
```

```
##
##      Federal-gov      Local-gov      Never-worked
Private
##           138           327           0
3468
##      Self-emp-inc Self-emp-not-inc      State-gov
Without-pay
##           175           376           205
3
```

```
tapply(df$hr.per.week,df$type.employer,mean)
```

```
##      Federal-gov      Local-gov      Never-worked
Private
##      39.83333      40.99694      NA
40.30421
##      Self-emp-inc Self-emp-not-inc      State-gov
Without-pay
##      48.96000      44.19149      40.10244
25.66667
```

```
df$f.type<-1
```

```
ll<-which(df$type.employer == "Private");length(ll)
```

```
## [1] 3468
```

```
df$f.type[ll]<-2
```

```
ll<-which(df$type.employer == "Self-emp-inc");length(ll)
```

```
## [1] 175
```

```
df$f.type[ll]<-3
```

```
ll<-which(df$type.employer %in% c("Self-emp-not-inc", "Never-
worked", "Without-pay"));length(ll)
```

```
## [1] 379
```

```
df$f.type[ll]<-4
```

```
df$f.type<-
factor(df$f.type,levels=1:4,labels=paste0("f.typ-",c("Civil","Pri
vate","SelfEm","Other")))
```

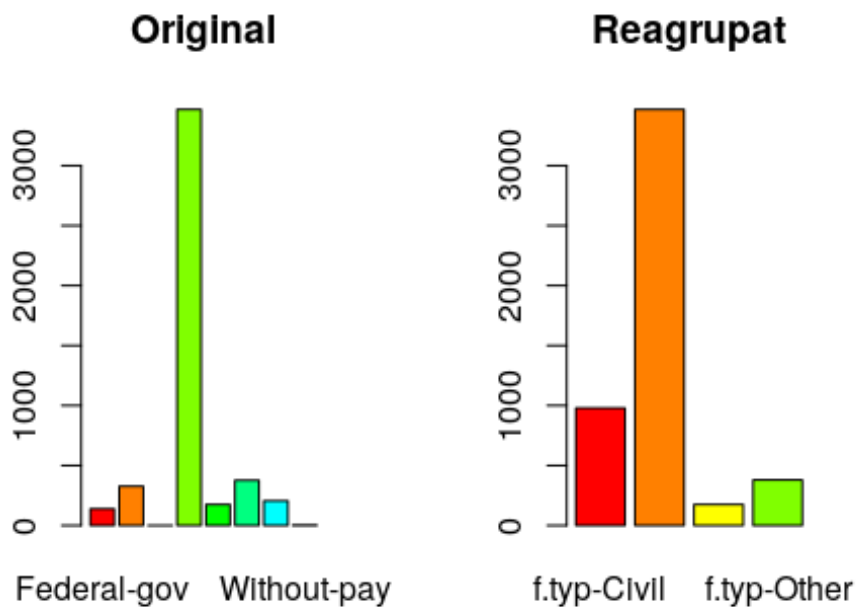
```
summary(df$f.type)
```

```
##    f.typ-Civil f.typ-Private f.typ-SelfEm f.typ-Other
##              978          3468          175          379
```

```
summary(df$type.employer)
```

```
##      Federal-gov      Local-gov      Never-worked
Private
##              138              327              0
3468
##      Self-emp-inc Self-emp-not-inc      State-gov
Without-pay
##              175              376              205
3
##      NA's
##      308
```

```
barplot(table(df$f.type),main="Reagrupat",col=rainbow(12))
```



marital

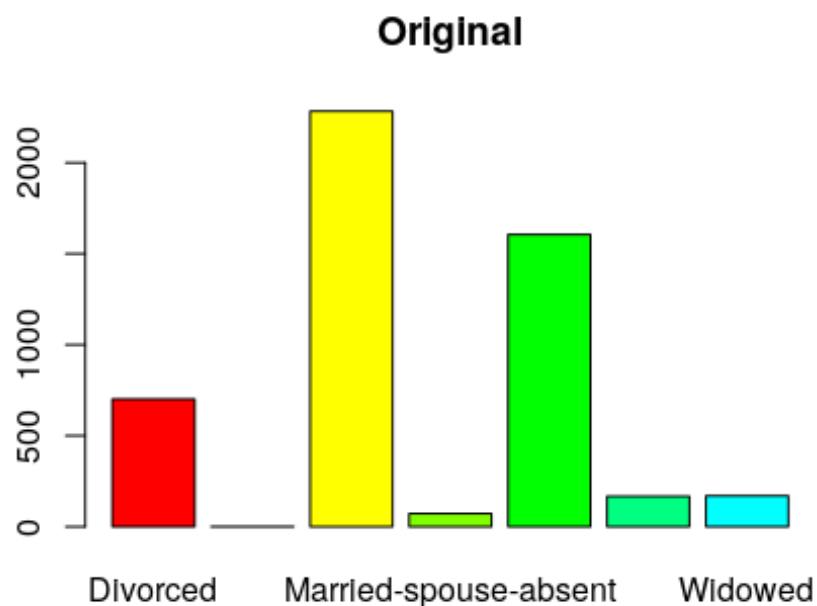
Desició conceptual:

1. Married => tots aquells que esan casats
2. No-married=> Divorced i Separated
3. Never-Married
4. Widowed

```
levels(df$marital)
```

```
## [1] "Divorced"          "Married-AF-spouse"  "Married-  
civ-spouse"  
## [4] "Married-spouse-absent" "Never-married"  
"Separated"  
## [7] "Widowed"
```

```
barplot(table(df$marital),main="Original",col=rainbow(12))
```



```
table(df$marital)
```

```
##  
##           Divorced  Married-AF-spouse  Married-civ-  
spouse  
##           701          1  
2283  
## Married-spouse-absent  Never-married  
Separated
```

```
##              72              1606
167
##              Widowed
##              170

tapply(df$hr.per.week,df$marital,mean)

##              Divorced      Married-AF-spouse      Married-civ-
spouse
##              40.99572              44.00000
43.41086
## Married-spouse-absent      Never-married
Separated
##              38.40278              37.18991
38.13174
##              Widowed
##              31.65294

df$f.marital<-1
ll<-which(df$marital %in% c ("Divorced","Separated")); length(ll)
## [1] 868

df$f.marital[ll]<-2
ll<-which(df$marital == "Never-married"); length(ll)
## [1] 1606

df$f.marital[ll]<-3
ll<-which(df$marital == "Widowed"); length(ll)
## [1] 170

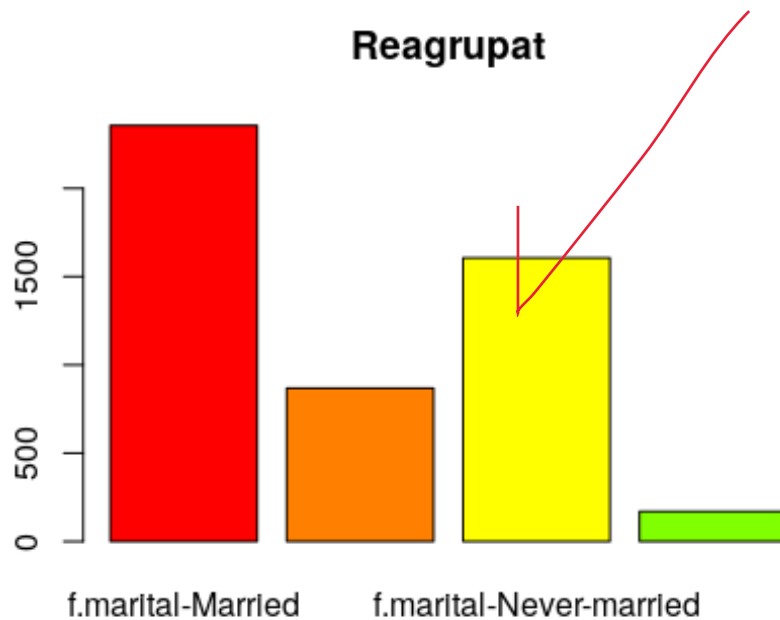
df$f.marital[ll]<-4

df$f.marital<-
factor(df$f.marital,levels=1:4,labels=paste0("f.marital-",c("Married",
"No- Married","Never-married","Widowed")))

summary(df$f.marital)

##          f.marital-Married      f.marital-No- Married f.marital-
Never-married
##              2356              868
1606
##          f.marital-Widowed
##              170

barplot(table(df$f.marital),main="Reagrupat",col=rainbow(12))
```



education

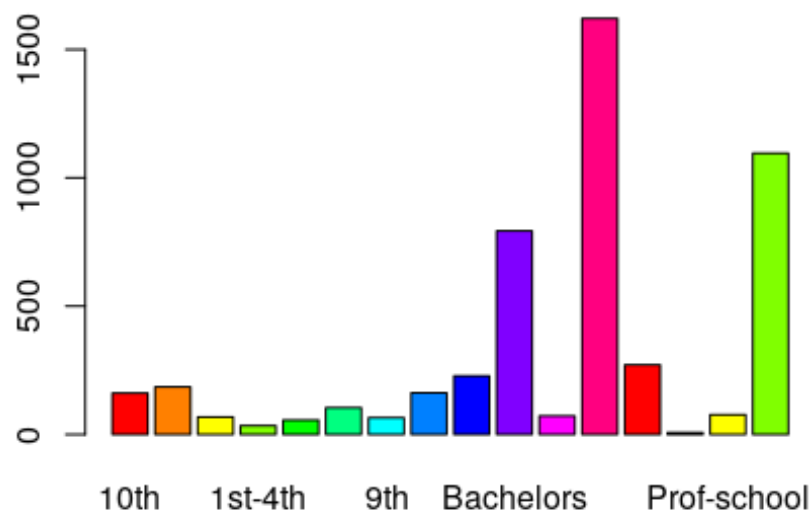
Desició conceptual:

1. Non-Graduatee => tots aquells que no han superat res més que els estudis obligatoris, o bé que no ho han fet
2. Some-college
3. University-Or-More => Doctorate, Bachelors, HS-grad, Masters
4. Assoc => Assoc-acdm, Assoc-voc
5. Prof-school

```
levels(df$education)
```

```
## [1] "10th"      "11th"      "12th"      "1st-4th"
## [5] "5th-6th"   "7th-8th"   "9th"       "Assoc-acdm"
## [9] "Assoc-voc" "Bachelors" "Doctorate" "HS-grad"
## [13] "Masters"   "Preschool" "Prof-school" "Some-
college"
```

```
barplot(table(df$education),col=rainbow(12))
```



```
table(df$education)
```

```
##
##      10th      11th      12th      1st-4th      5th-
6th
##      161      185        68        34
56
##      7th-8th      9th  Assoc-acdm  Assoc-voc
Bachelors
##      104        65        162        228
793
##  Doctorate  HS-grad  Masters  Preschool  Prof-
school
##      72      1621      271        7
77
## Some-college
##      1096
```

```
tapply(df$hr.per.week,df$education,mean)
```

```
##      10th      11th      12th      1st-4th      5th-
6th
##      37.32919      33.69189      38.10294      33.11765
36.80357
##      7th-8th      9th  Assoc-acdm  Assoc-voc
Bachelors
##      39.76923      37.24615      40.93827      42.50439
```

```

42.46784
##      Doctorate      HS-grad      Masters      Preschool      Prof-
school
##      49.41667      40.55706      43.71587      32.42857
47.11688
## Some-college
##      38.82208

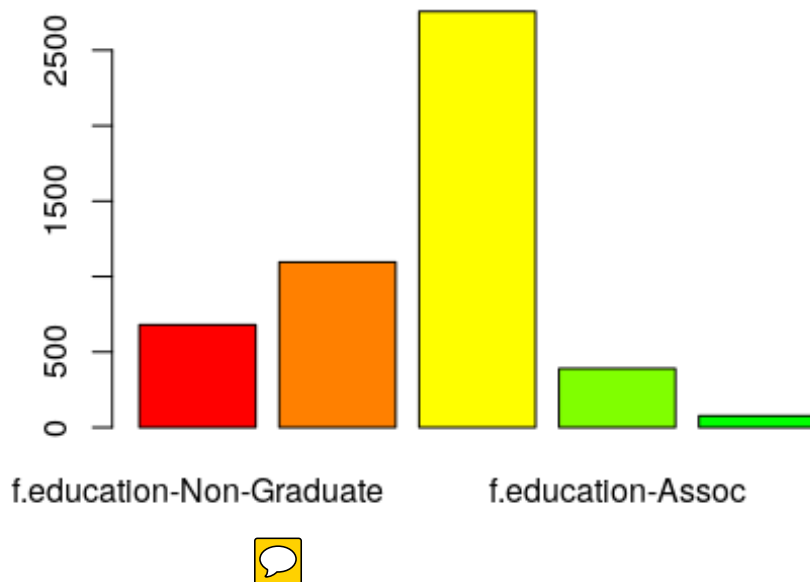
df$f.education<-1
ll<-which(df$education == "Some-college")
df$f.education[ll]<-2
ll<-which(df$education %in% c("Doctorate","Bachelors","HS-
grad","Masters"))
df$f.education[ll]<-3
ll<-which(df$education %in% c("Assoc-acdm","Assoc-voc"))
df$f.education[ll]<-4
ll<-which(df$education == "Prof-school")
df$f.education[ll]<-5

df$f.education<-
factor(df$f.education,levels=1:5,labels=paste0("f.education-",c("
Non-Graduate","Some-college","University-Or-More","Assoc","Proof-
school")))

summary(df$f.education)

##      f.education-Non-Graduate      f.education-Some-college
##      680      1096
## f.education-University-Or-More      f.education-Assoc
##      2757      390
##      f.education-Proof-school
##      77
barplot(table(df$f.education),col=rainbow(12))

```

Discretització de variables numèriques

En aquest apartat reagruparem totes aquelles variables numèriques en categories més generals, segons el nostre criteri pròpi.

Age

Agruparem el terme edat en els valors de tall que ens donen els quartils de la mostra.

Desició conceptual:

1. [17-29]
2. [30,39]
3. [40,49]
4. [50,90]

```
summary(df$age)
```

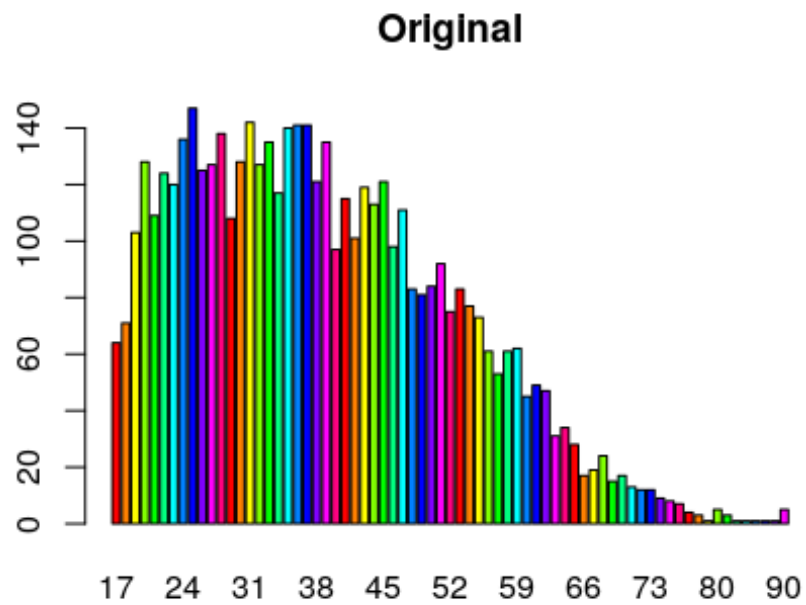
```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      17.0   27.0   37.0   38.7   48.0   90.0
```

```
df$f.age<- factor(cut(df$age,c(17,29,39,49,90),include.lowest =
T))
```

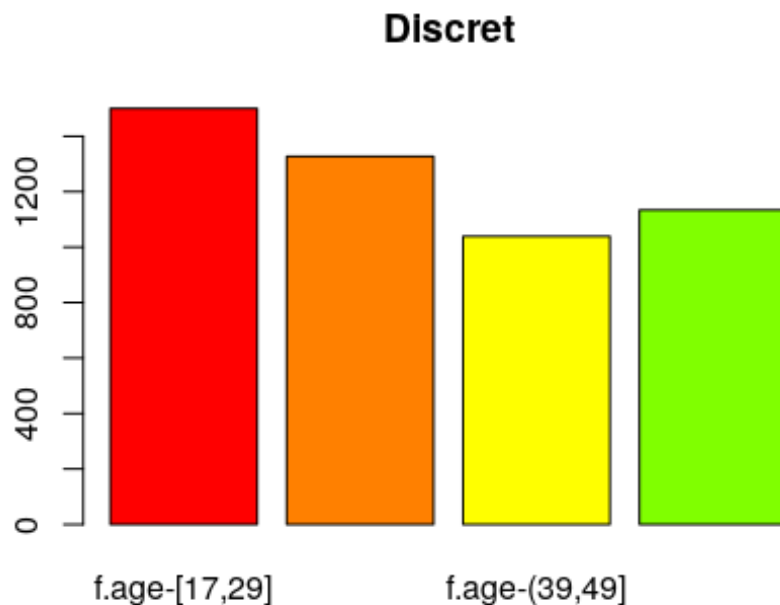
```
summary(df$f.age)
```

```
## [17,29] (29,39] (39,49] (49,90]
##      1500      1327      1039      1134
```

```
levels(df$f.age)<-paste0("f.age-",levels(df$f.age))  
barplot(table(df$age),main="Original",col=rainbow(12))
```



```
barplot(table(df$f.age),main="Discret",col=rainbow(12))
```



capital.gain & capital.loss

Hem cregut que aquestes variables tenen una relació gran que seria el benefici, es a dir capital.gain - capital.loss.

Desició conceptual:

1. Neutre
2. Positiu
3. Negatiu

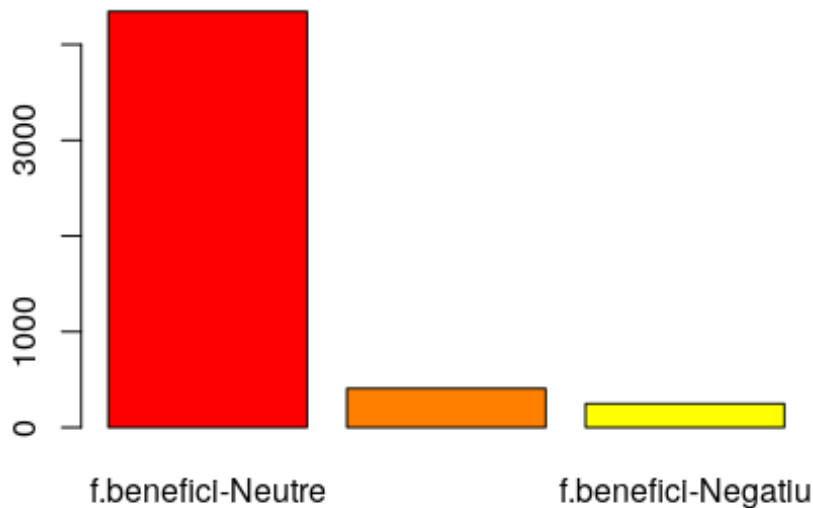
```
df$f.benefici<-1 #Neutre
ll<-which((df$capital.gain - df$capital.loss) > 0)
df$f.benefici[ll]<-2 #positiu
ll<-which((df$capital.gain-df$capital.loss) < 0)
df$f.benefici[ll]<-3

df$f.benefici<-
factor(df$f.benefici,levels=1:3,labels=paste0("f.benefici-",c("Neutre",
"Positiu","Negatiu")))

summary(df$f.benefici)

## f.benefici-Neutre f.benefici-Positiu f.benefici-Negatiu
##                4345                407                248
```

```
par(mfrow=c(1,1))
barplot(table(df$f.benefici),col=rainbow(12))
```



```
par(mfrow=c(1,2))
```



Qualitat de les dades

En aquest apartat, per cada variable contarem el nombre d'errors, missings i outliers. Per definir els outliers i errors, en cada categoria s'establiran valors límits en els que considerarem que a partir d'allà ja són valors que poden comprometre la qualitat de les dades. I per cada individu calcularem el nombre total d'errors + missings + outliers i s'afegirà com una variable extra del dataframe.

No obstant, per les variables discretes només es calcula el nombre de missings i errors (considerar outliers no té sentit). Considerarem que un error en una variable discreta és tot aquell valor que pren i que no es considera com a possible valor a prendre (tal i com queda explicat a la definició de les dades).

```
iout<- rep(0,nrow(df))
jout<- rep(0,length(vars_con))
```

```
ierr<- rep(0,nrow(df))
jerr<- rep(0,ncol(df))
```

```
imiss<-rep(0,nrow(df))
jmiss<-rep(0,ncol(df))
```

```
dfaux<-df
```

Hem creat un dataframe auxiliar que serà una copia del dataframe original, per poder fer en tot moment la comparativa del que són les dades reals i les dades que anem tractant.

age

Per a la variable "age", establim que tota edat que sigui 0 o bé sigui negativa serà considerada com a error.

```
#Calcul missing data
```

```
missingData<-which(is.na(dfaux$age)); length(missingData) #no missing data
```

```
## [1] 0
```

```
#Calcul errors (que assignem com NA per a la imputation)
```

```
sel<-which(df$age <= 0); length(sel) # errors
```

```
## [1] 0
```

```
if(length(sel)>0){
  dfaux[sel,"age"]<-NA
}
```

```
outers <- calcQ(dfaux$age)
```

```
outlier<-which(dfaux$age > outers$souts);length(outlier)
```

```
## [1] 0
```

```
dfaux[outlier ,"age"]<-NA
```

```
outlier<-which(dfaux$age < outers$souti);length(outlier)
```

```
## [1] 0
```

```
dfaux[outlier ,"age"]<-NA
```

```
# 0 outliers, es adir que per la variable age no tenim ni errors ni miss
```

```
outlier<-which(dfaux$age > outers$mouts);length(outlier)
```

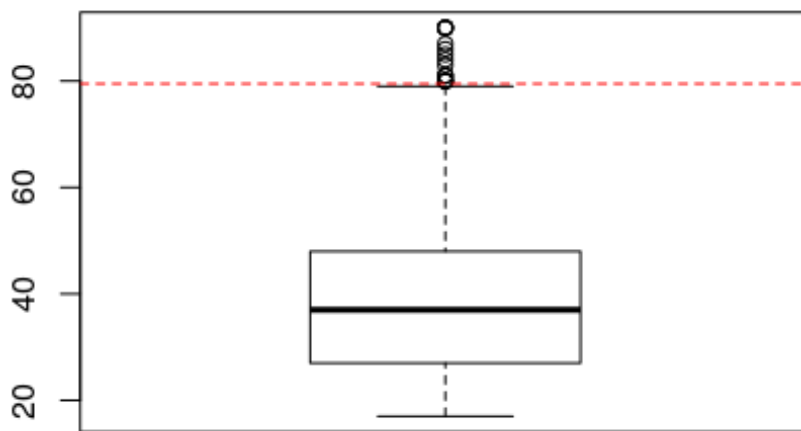
```
## [1] 18
```

```
outlier<-which(dfaux$age < outers$mouti);length(outlier)
```

```
## [1] 0

#tenim 0 outliers inferiors.

par(mfrow=c(1,1))
boxplot(df$age)
# A continuació veiem per on tallarien els outliers la mostra
d'entrada.
abline(h= outliers$mouts,col="red",lty=2)
```



Per aquesta variable, hem decidit que no hi hauran outliers, ja que els outliers que ens dona la teoria de quartils, creiem que no representen la diversitat d'aquest cens. Per tant mostrem un boxplot on es veu per on hauriem de tallar segons els valors teòrics, no obstant per desició pròpia decidim no fer-ho.

workclass

Per aquesta variable, hem establert que com a errors tractarem com a errors a tots aquells valors que no formin part de les categories d'entrada definides al inici.

```
missingData<-which(is.na(df$workclass));
length(missingData)
```

```
## [1] 308
```

```

imiss[missingData]<- imiss[missingData] +1
jmiss[2] <- jmiss[2]+ length(missingData)

#Tractarem com a error tot allo que no pertanyi al rang de valors
que contemplem
sel<-which(df$type.employer != 'Private' & df$type.employer !=
'Self-emp-not-inc' &
df$type.employer != 'Self-emp-inc' &
df$type.employer != 'Federal-gov' &
df$type.employer != 'Local-gov' & df$type.employer !=
= 'State-gov' &
df$type.employer != 'Without-pay' & df$type.employer
!= 'Never-worked'); length(sel) # errors

## [1] 0

if(length(sel)>0){
  dfaux[sel,"type.employer"]<-NA
}

#Tenim 0 errors

```

Com podem observar no hi han ni errors, no obstant si que tenim algun NA.

fnlwgt

En aquest cas, considerem errors aquells valors iguals o menor a 0. Amb aquesta variable no te sentit calcular els outliers perque ens és inútil.

```

missingData<-which(is.na(dfaux$fnlwgt)); length(missingData) #no
missing data

## [1] 0

#no tenim missing data

sel<-which(dfaux$fnlwgt <= 0); length(sel) # errors

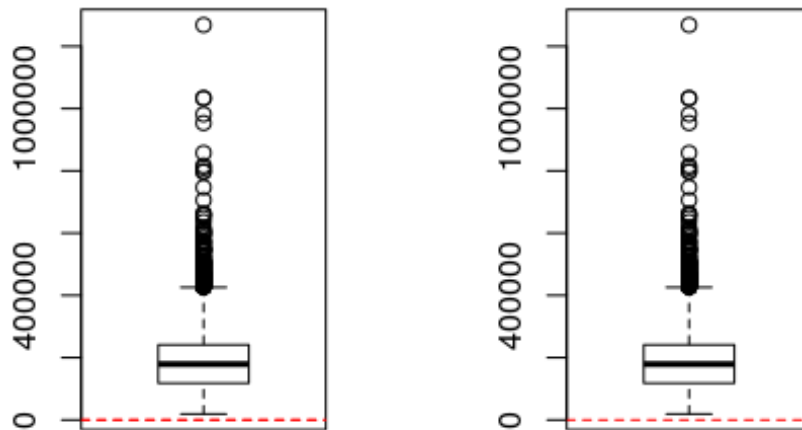
## [1] 0

if(length(sel)>0){
  dfaux[sel,"fnlwgt"]<-NA
}

par(mfrow=c(1,2))
boxplot(dfaux$fnlwgt)
abline( h= outliers$mouts, col="red", lty= 2)
abline( h= outliers$souts, col="red", lty= 2)

```

```
boxplot(df$fnlwgt)
abline( h= outers$mouts, col="red", lty= 2)
```



Reportem que no tenim cap missing value ni errors.

education

Seguim l'esquema inicial pel càlcul de missings i errors a variables discretes.

```
missingData<-which(is.na(dfaux$education)); length(missingData)
## [1] 0


#no tenim missing data
sel<-which(df$education != 'Bachelors' & df$education != 'Some-
college' &
          df$education != '11th' & df$education != 'HS-grad' &
          df$education != 'Prof-school' & df$education !=
'Assoc-acdm' &
          df$education != 'Assoc-voc' & df$education != '9th'
&
          df$education != '7th-8th' & df$education != '12th' &
          df$education != 'Masters' & df$education != '1st-
4th' &
          df$education != '10th' & df$education != 'Doctorate'
&
```



```

df$education != '5th-6th' & df$education !=
'Preschool'); length(sel) # errors
## [1] 0
if(length(sel)>0){
  df$education[sel]<-NA
}
#no tenim errors

```



No tenim ni missing data ni errors.

education.num

Veiem que aquesta variable sembla ser una discretització de la variable “education” (o que estan bastant lligades).

```

df %>% slice (1:20) %>% select(education,education.num)
##      education education.num
## 1      11th           7
## 2 Some-college       10
## 3      HS-grad        9
## 4      Assoc-voc      11
## 5      Prof-school    15
## 6      HS-grad        9
## 7 Some-college       10
## 8      7th-8th        4
## 9      HS-grad        9
## 10     Doctorate      16
## 11     Assoc-voc      11
## 12 Some-college       10
## 13     HS-grad        9
## 14     HS-grad        9
## 15     Assoc-voc      11
## 16     Bachelors      13
## 17 Some-college       10
## 18     HS-grad        9
## 19     HS-grad        9
## 20     Masters       14

summary(df$education.num)
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      1.00   9.00   10.00  10.04  12.00   16.00

missingData<-which(is.na(df$education.num));length(missingData)
## [1] 0

```

```
# no hi ha errors
```

```
sel<- which(dfaux$education.num < 1 | dfaux$education.num > 99); length(sel)
```

```
## [1] 0
```

```
#no hi ha errors
```

No tenim ni missing data ni errors. Com veiem al summary, no hi ha valors extrems i per tant considerem que no hi ha outliers.

marital status

Repetim càlcul de missing i errors per variables discretes.

```
missingData<-which(is.na(dfaux$marital)); length(missingData)
```

```
## [1] 0
```

```
sel<-which(df$marital != 'Married-civ-spouse' & df$marital !=  
'Divorced' &  
          df$marital != 'Never-married' & df$marital !=  
'Separated' &  
          df$marital != 'Widowed' & df$marital != 'Married-  
spouse-absent' &  
          df$marital != 'Married-AF-spouse'); length(sel) #
```

```
errors
```

```
## [1] 0
```

```
if(length(sel)>0){  
  dfaux[sel,"marital"]<-NA  
}
```

No tenim ni missing data ni errors.

occupation

Repetim càlcul de missing i errors per variables discretes.

```
missingData<-which(is.na(dfaux$occupation)); length(missingData)
```

```
## [1] 308
```

```
imiss[missingData]<- imiss[missingData] +1  
jmiss[7]<- jmiss[7] + length(missingData)
```

```
sel<-which(df$occupation != 'Tech-support' & df$occupation !=  
'Craft-repair' &  
          df$occupation != 'Other-service' & df$occupation !=  
'Sales' &  
          df$occupation != 'Exec-managerial' & df$occupation !=
```

```

= 'Prof-specialty' &
      df$occupation != 'Handlers-cleaners' & df$occupation
!= 'Machine-op-inspct' &
      df$occupation != 'Adm-clerical' & df$occupation !=
'Farming-fishing' &
      df$occupation != 'Transport-moving' & df$occupation
!= 'Priv-house-serv' &
      df$occupation != 'Protective-serv' & df$occupation !=
= 'Armed-Forces'); length(sel) # errors
## [1] 0
if(length(sel)>0){
  dfaux[sel,"occupation"]<-NA
}

```

En aquest cas tenim 308 missing values i cap error.

relationship

Repetim càlcul de missing i errors per variables discretes.

```

missingData<-which(is.na(dfaux$relationship));
length(missingData)
## [1] 0
sel<-which(df$relationship != 'Wife' & df$relationship != 'Own-
child' &
      df$relationship != 'Husband' & df$relationship !=
'Not-in-family' &
      df$relationship != 'Other-relative' &
df$relationship != 'Unmarried'); length(sel) # errors
## [1] 0
if(length(sel)>0){
  dfaux[sel,"relationship"]<-NA
}

```

No tenim ni missing data ni errors.

race

Repetim càlcul de missing i errors per variables discretes.

```

missingData<-which(is.na(dfaux$race)); length(missingData)
## [1] 0
sel<-which(df$race != 'White' & df$race != 'Asian-Pac-Islander' &
      df$race != 'Amer-Indian-Eskimo' & df$race != 'Other'

```

```

&
df$race != 'Black'); length(sel) # errors
## [1] 0
if(length(sel)>0){
  dfaux[sel,"race"]<-NA
}

```

No tenim ni missing data ni errors.

sex

Repetim càlcul de missing i errors per variables discretes.

```

missingData<-which(is.na(dfaux$sex)); length(missingData)
## [1] 0
sel<-which(df$sex != 'Female' & df$sex != 'Male'); length(sel) #
errors
## [1] 0
if(length(sel)>0){
  dfaux[sel,"race"]<-NA
}

```

No tenim ni missing data ni errors.

capital.gain

Considerem errors aquells valors inferiors a 0 o iguals a 99999.
D'altra banda, calculem els outliers i per aquells valors considerats com a "several outlier" els posem a NA per a que posteriorment siguin inputats.

```

summary(dfaux$capital.gain)
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##         0         0         0   1073     0   99999
#Calcul missing data
missingData<-which(is.na(dfaux$capital.gain));
length(missingData) #no missing data
## [1] 0
sel<-which(dfaux$capital.gain < 0 | dfaux$capital.gain == 99999);
length(sel) # errors
## [1] 25

```

```

ierr[sel]<-ierr[sel] +1
jerr[11]<- jerr[11]+length(sel)

if(length(sel)>0){
  dfaux[sel,"capital.gain"]<-NA
}

aux<- sort(dfaux[dfaux$capital.gain >
0,"capital.gain"],decreasing=TRUE); aux[1:30]

## [1] 34095 27828 27828 27828 27828 27828 27828 27828 27828 27828
25124 25124
## [12] 22040 20051 20051 20051 15024 15024 15024 15024 15024
15024 15024
## [23] 15024 15024 15024 15024 15024 15024 15024 15024 15024

```

Decidim per el criteri propi establir que tot capital gain superior a 20000 serà considerat outlier. No considerem outlier inferior, perquè les dades que siguin negatives(si hi ha), hauran estat tractades com a errors.

```

outlimit <- 20000

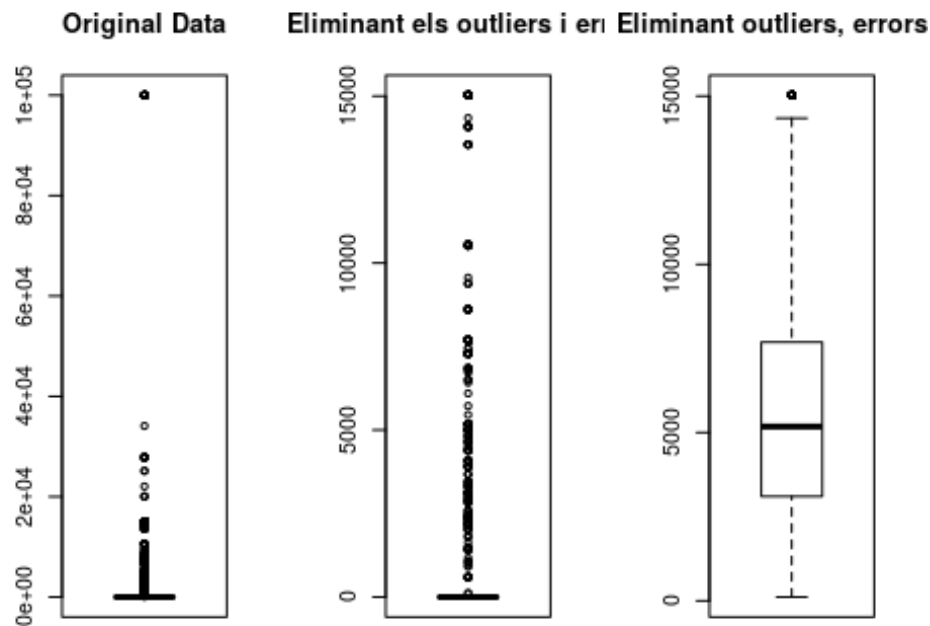
outlier<-which(dfaux$capital.gain > outlimit);length(outlier)

## [1] 15

ierr[outlier] <- ierr[outlier]+1
jerr[11]<- jerr[11] + length(outlier)
dfaux[outlier ,"capital.gain"]<-NA

par(mfrow=c(1,3))
boxplot(df$capital.gain,main="Original Data")
boxplot(dfaux$capital.gain, main= "Eliminant els outliers i errors")
boxplot(dfaux[dfaux$capital.gain>0,"capital.gain"], main=
"Eliminant outliers, errors i 0")

```



En el primer boxplot no veiem res al respecte, ja que la majoria de dades són 0, per tant mostrem que si treiem les que són 0 del segon boxplot, on hem posat els outliers a NA, ens queda un boxplot bastant bonic.

No tenim errors pero tenim 407 missing data. Els several outliers han estat posats a NA. ✓

capital.loss

Calculem errors, missings i outliers de manera anàloga a com s'ha fet amb la variable capital.loss

```
summary(dfaux$capital.loss)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      0.00    0.00    0.00   94.46    0.00 3900.00
```

```
missingData<-which(is.na(dfaux$capital.loss));
length(missingData) #no missing data
```

```
## [1] 0
```

```
sel<-which(df$capital.loss < 0 | df$capital.loss == 9999);
length(sel) # errors
```

```
## [1] 0
```

```

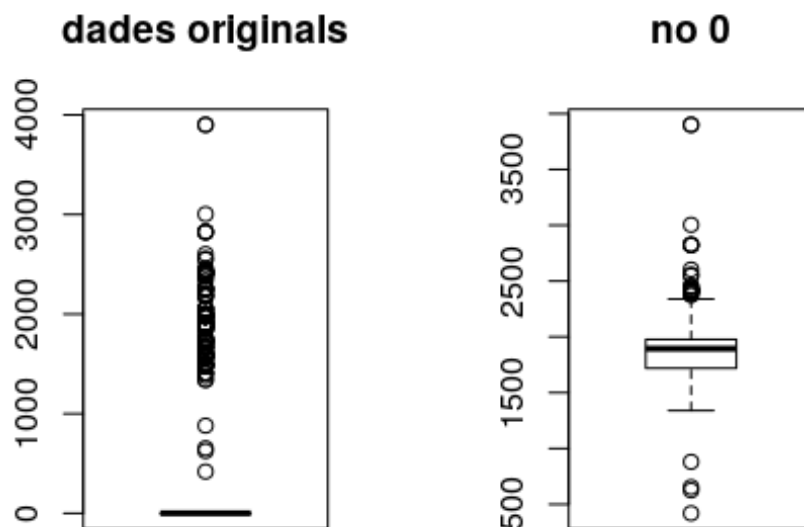
if(length(sel)>0){
  dfaux[sel,"capital.loss"]<-NA
}
#no hi han errors

aux<- sort(dfaux[dfaux$capital.loss >
0,"capital.loss"],decreasing=TRUE); aux[1:30]

## [1] 3900 3900 3004 2824 2824 2824 2824 2603 2559 2547 2457
2444 2444 2415
## [15] 2415 2415 2415 2415 2415 2415 2415 2415 2415 2415 2415 2415
2415 2392 2377
## [29] 2377 2339

par(mfrow=c(1,2))
boxplot(df$capital.loss,main="dades originals")
boxplot(dfaux[dfaux$capital.loss>0,"capital.loss"],main="no 0")

```



No tenim errors però tenim 248 missing data. En aquesta variable no tenim en compte outliers, ja que després d'analitzar les dades hem vist que no hi ha cap valor tant extrem com per a considerar-lo outlier.

hr.per.week

Considerem errors aquells valors que siguin menor o igual a 0 o iguals a 99.

```
summary(dfaux$hr.per.week)

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      1.00   40.00   40.00   40.43   45.00   99.00

ll<-which(is.na(dfaux$hr.per.week));ll

## integer(0)

#no tenim na
sel<-which(dfaux$hr.per.week <= 0 | dfaux$hr.per.week ==99);
length(sel) # errors

## [1] 15

ierr[sel]<- ierr[sel]+1
jerr[13]<- jerr[13]+length(sel)
dfaux[sel,"hr.per.week"]<-NA
```

Tenint en compte que la jornada labroal màxima es de 40 hores setmanals, establim el limit a un 150% d'aquesta, es a dir 60 hores. Establim un limit inferior també, ja que considerarem que treballar menys de 10 hores serà outlier

```
outlimit<- 60
outlier<-which(dfaux$hr.per.week > outlimit );length(outlier)
#outliers superiors critics

## [1] 153

ierr[outlier]<- ierr[outlier] +1
jerr[13]<-jerr[13]+length(outlier)
dfaux[outlier,"hr.per.week"]<-NA

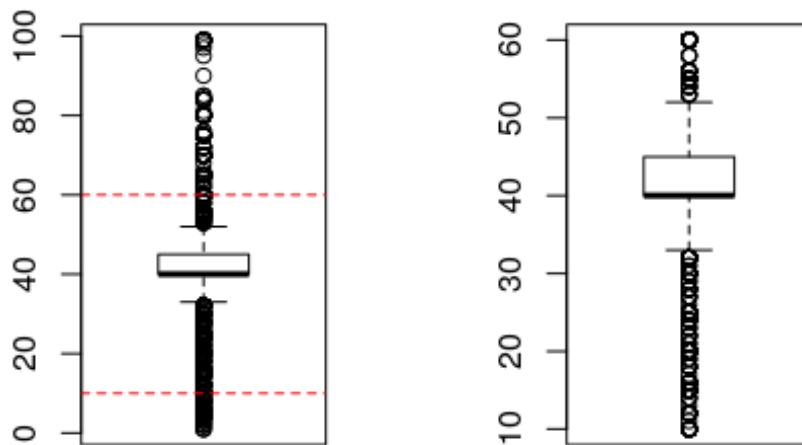
outlimit<- 10
outlier<-which(dfaux$hr.per.week < outlimit );length(outlier)
#outliers superiors critics

## [1] 80

ierr[outlier]<- ierr[outlier] +1
jerr[13]<-jerr[13]+length(outlier)
dfaux[outlier,"hr.per.week"]<-NA

par(mfrow=c(1,2))
boxplot(df$hr.per.week)
abline(h= 60,col="red",lty=2)
abline(h= 10,col="red",lty=2)

boxplot(dfaux$hr.per.week)
```

No tenim errors ni missing values. ~~Els several outliers han estat posats a NA.~~

country

```
missingData<-which(is.na(df$country)); length(missingData)
```

```
## [1] 88
```

```
imiss[missingData]<- imiss[missingData] + 1
```

```
jmiss[14]<-jmiss[14]+length(missingData)
```

```
sel<-which(df$country != 'United-States' & df$country !=
'Cambodia' &
df$country != 'England' & df$country != 'Puerto-
Rico' &
df$country != 'Canada' & df$country != 'Germany' &
df$country != 'Outlying-US(Guam-USVI-etc)' &
df$country != 'India' &
df$country != 'Japan' & df$country != 'Greece' &
df$country != 'South' & df$country != 'China' &
df$country != 'Cuba' & df$country != 'Iran' &
df$country != 'Honduras' & df$country !=
'Philippines' &
df$country != 'Italy' & df$country != 'Poland' &
df$country != 'Jamaica' & df$country != 'Vietnam' &
df$country != 'Mexico' & df$country != 'Portugal' &
df$country != 'Ireland' & df$country != 'France' &
```

```

df$country != 'Dominican-Republic' & df$country !=
'Laos' &
df$country != 'Ecuador' & df$country != 'Taiwan' &
df$country != 'Haiti' & df$country != 'Columbia' &
df$country != 'Hungary' & df$country != 'Guatemala'
&
df$country != 'Nicaragua' & df$country != 'Scotland'
&
df$country != 'Thailand' & df$country !=
'Yugoslavia' &
df$country != 'El-Salvador' & df$country !=
'Trinidad&Tobago' &
df$country != 'Peru' & df$country != 'Hong' &
df$country != 'Holand-Netherlands'); length(sel) #
errors
## [1] 0
if(length(sel)>0){
  dfaux[sel,"country"]<-NA
}

```

Tenim 88 missing values i no tenim cap error.

y.bin

En aquest cas estem tractant una variable binària, per tant només té sentit analitzar el nombre de missing values.

```

missingData<-which(is.na(dfaux$y.bin)); length(missingData)
## [1] 0

```

No tenim missing values. Tot aquest procés es podria fer per les variables reagrupades i discretitzades. No obstant, considerem que no té sentit ja que previament ja estem tractant tots els casos.

Recompte d'errors, per individu

A continuació veurem quants errors té cada individu, i també veurem la mitjana d'error per cada classe. Realitzarem la mitjana fent la suma de error, outliers i missing dividit entre 3, de tal manera que veurem per cada classe quina és la mitjana d'errors outliers i missings.

```

#afegim la variable que és la suma dels errors missings i
outliers al df
dfaux$i.rank<- ierr + imiss + iout

#realitzar la mitjana de tot per variable.
aux<-(countNA(dfaux)$mis_col)/3

```

```

install.packages("corrplot")

## Installing package into '/usr/local/lib/R/site-library'
## (as 'lib' is unspecified)

library(corrplot)

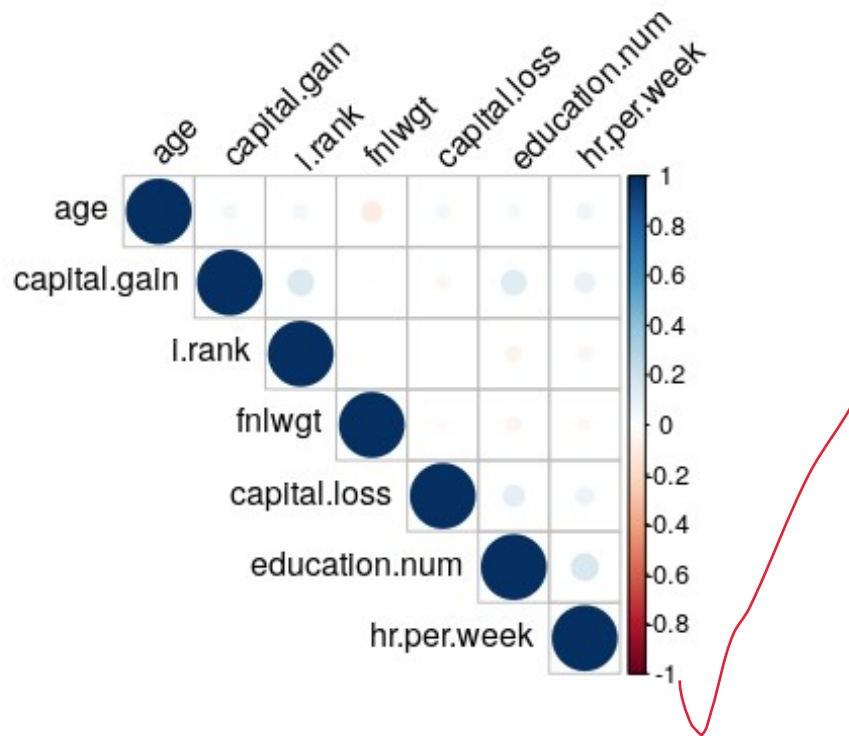
## corrplot 0.84 loaded

t<- df[,vars_con]
df$i.rank <- dfaux$i.rank
t$i.rank <- df[, "i.rank"]
corMatrix<-cor(t); corMatrix

##
##          age          fnlwgt education.num
capital.gain
## age          1.00000000 -0.0938874683    0.03748753
0.048825838
## fnlwgt       -0.09388747  1.0000000000    -0.04762524
0.008635798
## education.num 0.03748753 -0.0476252438    1.00000000
0.139627539
## capital.gain  0.04882584  0.0086357980    0.13962754
1.000000000
## capital.loss  0.04621191 -0.0225491721    0.10466473 -
0.032178822
## hr.per.week   0.05832764 -0.0306001824    0.16637797
0.089539722
## i.rank        0.04750264 -0.0003940381    -0.05039291
0.151180385
##
##          capital.loss hr.per.week          i.rank
## age          0.046211911  0.05832764  0.0475026386
## fnlwgt       -0.022549172 -0.03060018 -0.0003940381
## education.num 0.104664731  0.16637797 -0.0503929118
## capital.gain  -0.032178822  0.08953972  0.1511803854
## capital.loss   1.000000000  0.06633876  0.0029919329
## hr.per.week   0.066338758  1.000000000 -0.0419182111
## i.rank        0.002991933 -0.04191821  1.0000000000

corrplot(corMatrix, type = "upper", order = "hclust",
          tl.col = "black", tl.srt = 45)

```



Veiem que la variable i.rank, aquella que resumeix quants errors/missings/outliers hi ha per individu, no te gaire correlació amb les altres variables (numèriques).

Imputació de variables

En aquest apartat el que realitzarem és per totes aquelles variables que hem categoritzat com preilloses, és a dir que estan en la categoria err/miss/out, farem una aproximació del valor o categoria, mitjançant imputePCA o imputeMCA, respectivament.

```
install.packages("missMDA")

## Installing package into '/usr/local/lib/R/site-library'
## (as 'lib' is unspecified)


library(missMDA)

# numericas
res.num<-imputePCA(dfauX[,vars_con])
summary(res.num$completeObs)

##      age      fnlwgt      education.num
capital.gain
## Min.   :17.0   Min.    : 18827   Min.    : 1.00   Min.    :
```



0.0

```
## 1st Qu.:27.0    1st Qu.: 118008    1st Qu.: 9.00    1st Qu.:
0.0
## Median :37.0    Median : 178950    Median :10.00    Median :
0.0
## Mean :38.7     Mean : 192215     Mean :10.04     Mean :
502.9
## 3rd Qu.:48.0    3rd Qu.: 241215    3rd Qu.:12.00    3rd Qu.:
0.0
## Max. :90.0     Max. : 1268339    Max. :16.00
Max. :15024.0
## capital.loss  hr.per.week
## Min. : 0.00    Min. :10.00
## 1st Qu.: 0.00    1st Qu.:40.00
## Median : 0.00    Median :40.00
## Mean : 94.46    Mean :39.84
## 3rd Qu.: 0.00    3rd Qu.:43.03
## Max. :3900.00    Max. :60.00
```

```
summary(dfaux[,vars_con])
```

```
##      age      fnlwgt      education.num
capital.gain
## Min. :17.0    Min. : 18827    Min. : 1.00    Min. :
0.0
## 1st Qu.:27.0    1st Qu.: 118008    1st Qu.: 9.00    1st Qu.:
0.0
## Median :37.0    Median : 178950    Median :10.00    Median :
0.0
## Mean :38.7     Mean : 192215     Mean :10.04     Mean :
499.4
## 3rd Qu.:48.0    3rd Qu.: 241215    3rd Qu.:12.00    3rd Qu.:
0.0
## Max. :90.0     Max. : 1268339    Max. :16.00
Max. :15024.0
##                                     NA's :40



##      capital.loss      hr.per.week
## Min. : 0.00    Min. :10.00
## 1st Qu.: 0.00    1st Qu.:40.00
## Median : 0.00    Median :40.00
## Mean : 94.46    Mean :39.81
## 3rd Qu.: 0.00    3rd Qu.:45.00
## Max. :3900.00    Max. :60.00
##                                     NA's :248
```

Per a les variables numèriques, veiem que no hi ha gran diferència entre el summary de la original, es a dir amb els NA, que a la actual, els valors mean i quartils, no es veuen afectats de gran manera, així que acceptem aquesta imputació.

```
# descriptivas
```

```
res.des<-imputeMCA(dfaux[,vars_dis])
```

```
summary(res.des$completeObs)
```

```
##                type.employer                education
marital
## Federal-gov      : 138    HS-grad      :1621    Divorced
: 701
## Local-gov        : 327    Some-college:1096    Married-AF-spouse
: 1
## Private          :3776    Bachelors   : 793    Married-civ-
spouse :2283
## Self-emp-inc      : 175    Mas s      : 271    Married-spouse-
absent: 72
## Self-emp-not-inc: 376  Assoc-voc : 228    Never-married
:1606
## State-gov        : 205    11th      : 185    Separated
: 167
## Without-pay      : 3      (Other)     : 806    Widowed
: 170
```

```
##                occupation                relationship
```

```
race
## Adm-clerical      : 716    Husband      :1987    Amer-Indian-
Eskimo: 45
## Prof-specialty    : 681    Not-in-family :1315    Asian-Pac-
Islander: 154
## Craft-repair      : 643    Other-relative: 169    Black
: 507
## Exec-managerial: 638    Own-child      : 758    Other
: 34
## Sales             : 565    Unmarried      : 505    White
:4260
## Other-service     : 562    Wife           : 266
## (Other)           :1195
```

```
##                sex                country                y.bin
## Female:1681    United-States:4576    <=50K:3800
## Male :3319    Mexico      : 94    >50K :1200
##                Germany      : 27
##                Philippines   : 26
##                Canada        : 24
##                Puerto-Rico   : 22
##                (Other)       : 231
```

```
summary(dfaux[,vars_dis])
```

```
##                type.employer                education
marital
## Private          :3468    HS-grad      :1621    Divorced
: 701
```

```
## Self-emp-not-inc: 376 Some-college:1096 Married-AF-spouse
: 1
## Local-gov : 327 Bachelors : 793 Married-civ-
spouse :2283
## State-gov : 205 Masters : 271 Married-spouse-
absent: 72
## Self-emp-inc : 175 Assoc-voc : 228 Never-married
:1606
## (Other) : 141 11th : 185 Separated
: 167
## NA's : 308 (Other) : 806 Widowed
: 170
## occupation relationship
race
## Prof-specialty : 635 Husband :1987 Amer-Indian-
Eskimo: 45
## Exec-managerial: 624 Not-in-family :1315 Asian-Pac-
Islander: 154
## Craft-repair : 595 Other-relative: 169 Black
: 507
## Adm-clerical : 591 Own-child : 758 Other
: 34
## Sales : 565 Unmarried : 505 White
:4260
## (Other) :1682 Wife : 266

## NA's : 308

## sex country y.bin
## Female:1681 United-States:4488 <=50K:3800
## Male :3319 Mexico : 94 >50K :1200
## Germany : 27
## Philippines : 26
## Canada : 24
## (Other) : 253
## NA's : 88
```

El mateix que per les variables numèriques, veiem ara amb les descriptives. No hi han grans alteracions de les dades, que ens facin rebutjar la imputació d'aquestes.

Un cop tenim les dades correctes, procedim a modificarles directament en el data frame.

```
dfaux[,vars_con]<- res.num$completeObs
dfaux[,vars_dis]<- res.des$completeObs
```

Profiling

En aquest apartat veurem la rellevància de cada variable, respecte els nostres targets (hr.per.week, i y.bin).

En primer lloc realitzarem el profiling per el target numéric (hr.per.week)

hr.per.week

```
vars<-names(dfaux)[c(13,1,3,5:12,14:21)]
```

```
condes(dfaux[,vars],1,prob=0.01)
```

```
## $quanti
##               correlation      p.value
## education.num  0.19651751 1.029850e-44
## age            0.09969124 1.605256e-12
## capital.gain   0.08784074 4.916267e-10
## capital.loss   0.05439922 1.188463e-04
## i.rank         -0.09451823 2.130714e-11
```

```
## $quali
##               R2      p.value
## relationship 0.110607964 2.669338e-124
## occupation   0.077085618 1.833097e-77
## marital      0.069073334 3.757814e-74
## f.marital    0.065725288 2.552400e-73
## sex          0.060532871 7.785357e-70
## y.bin        0.052782197 6.885118e-61
## f.age        0.048346203 2.166215e-53
## f.education  0.045695922 2.132574e-49
## f.type       0.021378895 2.988465e-23
## f.benefici   0.007718743 3.908575e-09
## race         0.005762989 8.287345e-06
```

```
## $category
```

| | Estimate |
|--------------------------------|-----------|
| p.value | |
| ## relationship=Husband | 4.7763436 |
| ## sex=Male | 2.4804003 |
| ## y.bin=50K | 2.5619753 |
| ## marital=Married-civ-spouse | 3.0210095 |
| ## f.marital=f.marital-Married | 3.8513373 |
| ## occupation=Exec-managerial | 4.0086635 |


| | |
|---|------------|
| 8.527091e-25 | |
| ## f.education=f.education-University-Or-More | 0.6826905 |
| 6.183215e-24 | |
| ## f.age=f.age-(39,49] | 2.0493493 |
| 6.515364e-18 | |
| ## f.type=f.typ-SelfEm | 4.2681764 |
| 9.710092e-17 | |
| ## f.age=f.age-(29,39] | 1.4820344 |
| 3.633890e-14 | |
| ## occupation=Transport-moving | 3.6494035 |
| 5.763996e-08 | |
| ## f.education=f.education-Proof-school | 5.0342136 |
| 2.239735e-07 | |
| ## race=White | 0.8335159 |
| 8.332228e-07 | |
| ## f.benefici=f.benefici-Positui | 0.9069227 |
| 9.040816e-07 | |
| ## occupation=Prof-specialty | 1.9391430 |
| 6.048272e-06 | |
| ## f.type=f.typ-Other | 0.4693562 |
| 1.183258e-05 | |
| ## relationship=Unmarried | 0.1918107 |
| 4.264665e-04 | |
| ## f.benefici=f.benefici-Negatiu | 0.7312630 |
| 5.225828e-04 | |
| ## f.education=f.education-Assoc | 0.8552218 |
| 2.636731e-03 | |
| ## occupation=Farming-fishing | 2.5938377 |
| 3.037365e-03 | |
| ## country=Japan | 12.6081484 |
| 7.812425e-03 | |
| ## country=Philippines | -6.1995439 |
| 9.695281e-04 | |
| ## relationship=Other-relative | -0.9440270 |
| 3.946402e-04 | |
| ## occupation=Priv-house-serv | -7.2369833 |
| 5.722396e-05 | |
| ## occupation=Handlers-cleaners | -2.1882230 |
| 2.888862e-05 | |
| ## race=Black | -1.5315953 |
| 1.843173e-07 | |
| ## f.type=f.typ-Civil | -3.0347255 |
| 1.184541e-07 | |
| ## relationship=Wife | -1.4878873 |
| 4.948176e-08 | |
| ## f.education=f.education-Some-college | -1.9722400 |
| 1.585957e-08 | |
| ## f.benefici=f.benefici-Neutre | -1.6381857 |
| 5.015363e-10 | |
| ## occupation=Adm-clerical | -2.3764595 |

```

2.437808e-17
## f.marital=f.marital-Widowed -4.6553225
1.329083e-18
## marital=Widowed -5.6088598
1.329083e-18
## f.education=f.education-Non-Graduate -4.5998860
1.949181e-33
## occupation=Other-service -4.2514764
2.829742e-35
## f.marital=f.marital-Never-married -1.0385328
1.938486e-43
## marital=Never-married -1.9920701
1.938486e-43
## f.age=f.age-[17,29] -3.1297624
1.464304e-46
## y.bin=<=50K -2.5619753
6.885118e-61
## sex=Female -2.4804003
7.785357e-70
## relationship=Own-child -4.1993299
1.172618e-76

```

Veiem que les variables que tenen major correlació amb el target quantitat d'hores treballades, són education.num i relationship. amb correlacions superiors al 0.1.

També observem que hi han variables que tenen importància, no obstant no tanta com les que hem esmentat anteriorment, i per últim tenim aquelles variables que realment no tenen molta relevància, com seria la raça. Aquest  particularment ens ha sobtat, ja que a priori creiem que anava a ser una de les que anava a tenir major relevància ja que habitualment creiem que la raça ens limita al moment de establir un sou.

No ens ha sorprès que la variable i.rank, la que defineix el nombre de missings i d'errors sigui inversament proporcional al número d'hores treballades, ja a major nombre d'hores treballades indica que hi ha major nombre d'hores d'estudi, amb el que podem concloure que aquelles persones que més anys han estudiat, generen menys errors o no es deixen les dades per completar, en enquestes del tipus que es planteja en aquest informe.

y.bin

```
vars<-names(dfaux)[c(15,1,3,7:10,13:14,16:21)]
```

```
catdes(dfaux[,vars],1,prob=0.01)
```

```
##
```

```
## Link between the cluster variable and the categorical
```

variables (chi-square test)

##

=====

| ## | | p.value | df |
|-----------------|---------------|---------|----|
| ## relationship | 4.196083e-224 | 5 | |
| ## f.marital | 3.478295e-203 | 3 | |
| ## occupation | 1.431812e-134 | 13 | |
| ## f.benefici | 1.777278e-106 | 2 | |
| ## f.age | 2.305331e-85 | 3 | |
| ## f.education | 2.568685e-66 | 4 | |
| ## sex | 2.684671e-42 | 1 | |
| ## f.type | 6.855541e-28 | 3 | |
| ## race | 1.366388e-13 | 4 | |

##

Description of each cluster by the categories

=====

\$`<=50K`

##

Cla/Mod

Mod/Cla Global

f.marital=f.marital-Never-married 94.95641

40.1315789 32.12

f.age=f.age-[17,29] 93.53333

36.9210526 30.00

f.benefici=f.benefici-Neutre 81.17376

92.8157895 86.90

relationship=Own-child 99.34037

19.8157895 15.16

occupation=Other-service 97.50890

14.4210526 11.24

sex=Female 87.56692

38.7368421 33.62

f.education=f.education-Non-Graduate 94.26471

16.8684211 13.60

relationship=Not-in-family 88.89734

30.7631579 26.30

f.marital=f.marital-No- Married 91.12903

20.8157895 17.36

relationship=Unmarried 94.65347

12.5789474 10.10

occupation=Adm-clerical 88.12849

16.6052632 14.32

race=Black 88.95464

11.8684211 10.14

relationship=Other-relative 97.04142

4.3157895 3.38

occupation=Handlers-cleaners 94.22222

5.5789474 4.50

country=Mexico 95.74468

2.3684211 1.88

| | |
|---|---------------|
| ## f.marital=f.marital-Widowed | 90.58824 |
| 4.0526316 3.40 | |
| ## f.type=f.typ-Private | 77.99885 |
| 71.1842105 69.36 | |
| ## occupation=Machine-op-inspct | 86.46865 |
| 6.8947368 6.06 | |
| ## f.education=f.education-Some-college | 80.29197 |
| 23.1578947 21.92 | |
| ## occupation=Priv-house-serv | 100.00000 |
| 0.6578947 0.50 | |
| ## race=Amer-Indian-Eskimo | 93.33333 |
| 1.1052632 0.90 | |
| ## f.age=f.age-(29,39] | 73.17257 |
| 25.5526316 26.54 | |
| ## country=United-States | 75.21853 |
| 90.5789474 91.52 | |
| ## f.age=f.age-(49,90] | 68.16578 |
| 20.3421053 22.68 | |
| ## race=White | 74.15493 |
| 83.1315789 85.20 | |
| ## f.education=f.education-University-Or-More | 71.63584 |
| 51.9736842 55.14 | |
| ## f.type=f.typ-SelfEm | 40.57143 |
| 1.8684211 3.50 | |
| ## relationship=Wife | 46.61654 |
| 3.2631579 5.32 | |
| ## f.age=f.age-(39,49] | 62.84889 |
| 17.1842105 20.78 | |
| ## f.education=f.education-Proof-school | 16.88312 |
| 0.3421053 1.54 | |
| ## f.benefici=f.benefici-Negatiu | 42.74194 |
| 2.7894737 4.96 | |
| ## occupation=Prof-specialty | 54.03818 |
| 9.6842105 13.62 | |
| ## sex=Male | 70.14161 |
| 61.2631579 66.38 | |
| ## occupation=Exec-managerial | 51.88088 |
| 8.7105263 12.76 | |
| ## f.benefici=f.benefici-Positui | 41.03194 |
| 4.3947368 8.14 | |
| ## relationship=Husband | 55.96376 |
| 29.2631579 39.74 | |
| ## f.marital=f.marital-Married | 56.45161 |
| 35.0000000 47.12 | |
| ## | p.value |
| v.test | |
| ## f.marital=f.marital-Never-married | 9.012756e-125 |
| 23.758325 | |
| ## f.age=f.age-[17,29] | 2.817808e-95 |
| 20.709945 | |

| | | |
|---|--------------|---|
| ## f.benefici=f.benefici-Neutre | 1.023816e-92 | |
| 20.423976 | | |
| ## relationship=Own-child | 1.064373e-89 | |
| 20.081814 | | |
| ## occupation=Other-service | 5.499322e-51 | |
| 15.019169 | | |
| ## sex=Female | 1.146411e-45 | |
| 14.184277 | | |
| ## f.education=f.education-Non-Graduate | 1.326698e-41 | |
| 13.512107 | | |
| ## relationship=Not-in-family | 1.722531e-41 | |
| 13.492874 | | |
| ## f.marital=f.marital-No- Married | 1.046142e-35 | |
| 12.473145 | | |
| ## relationship=Unmarried | 6.071266e-32 | |
| 11.762765 | | |
| ## occupation=Adm-clerical | 2.420233e-18 | |
| 8.735760 | | |
| ## race=Black | 1.022517e-14 | |
| 7.736424 | | |
| ## relationship=Other-relative | 1.448749e-14 | |
| 7.691980 | | |
| ## occupation=Handlers-cleaners | 1.132376e-13 | |
| 7.424464 | | |
| ## country=Mexico | 2.048243e-07 | |
| 5.194905 | | |
| ## f.marital=f.marital-Widowed | 8.112097e-07 | |
| 4.932651 | | |
| ## f.type=f.typ-Private | 8.381738e-07 | |
| 4.926262 | | |
| ## occupation=Machine-op-inspct | 3.447791e-06 | |
| 4.642164 | | |
| ## f.education=f.education-Some-college | 1.338104e-04 | |
| 3.819338 | | |
| ## occupation=Priv-house-serv | 1.028205e-03 | |
| 3.282693 | | |
| ## race=Amer-Indian-Eskimo | 2.821914e-03 | |
| 2.986499 | | |
| ## f.age=f.age- (29,39] | 5.232697e-03 | - |
| 2.792348 | | |
| ## country=United-States | 9.802713e-06 | - |
| 4.421480 | | |
| ## f.age=f.age- (49,90] | 6.429593e-12 | - |
| 6.869791 | | |
| ## race=White | 1.197214e-14 | - |
| 7.716336 | | |
| ## f.education=f.education-University-0r-More | 7.217966e-16 | - |
| 8.066775 | | |
| ## f.type=f.typ-SelfEm | 2.616307e-24 | - |
| 10.173034 | | |

| | | |
|---|---------------|---|
| ## relationship=Wife | 2.464475e-26 | - |
| 10.617743 | | |
| ## f.age=f.age-(39,49] | 5.134904e-27 | - |
| 10.763211 | | |
| ## f.education=f.education-Proof-school | 4.016679e-28 | - |
| 10.995503 | | |
| ## f.benefici=f.benefici-Negatiu | 1.139926e-30 | - |
| 11.512597 | | |
| ## occupation=Prof-specialty | 4.052161e-42 | - |
| 13.599132 | | |
| ## sex=Male | 1.146411e-45 | - |
| 14.184277 | | |
| ## occupation=Exec-managerial | 1.721208e-46 | - |
| 14.316693 | | |
| ## f.benefici=f.benefici-Positiu | 2.478837e-56 | - |
| 15.814317 | | |
| ## relationship=Husband | 3.419847e-159 | - |
| 26.883554 | | |
| ## f.marital=f.marital-Married | 2.216248e-219 | - |
| 31.616131 | | |
| ## | | |
| ## \$`>50K` | | |
| ## | | |
| | Cla/Mod | |
| Mod/Cla Global | | |
| ## f.marital=f.marital-Married | 43.5483871 | |
| 85.5000000 47.12 | | |
| ## relationship=Husband | 44.0362355 | |
| 72.9166667 39.74 | | |
| ## f.benefici=f.benefici-Positiu | 58.9680590 | |
| 20.0000000 8.14 | | |
| ## occupation=Exec-managerial | 48.1191223 | |
| 25.5833333 12.76 | | |
| ## sex=Male | 29.8583911 | |
| 82.5833333 66.38 | | |
| ## occupation=Prof-specialty | 45.9618209 | |
| 26.0833333 13.62 | | |
| ## f.benefici=f.benefici-Negatiu | 57.2580645 | |
| 11.8333333 4.96 | | |
| ## f.education=f.education-Proof-school | 83.1168831 | |
| 5.3333333 1.54 | | |
| ## f.age=f.age-(39,49] | 37.1511068 | |
| 32.1666667 20.78 | | |
| ## relationship=Wife | 53.3834586 | |
| 11.8333333 5.32 | | |
| ## f.type=f.typ-SelfEm | 59.4285714 | |
| 8.6666667 3.50 | | |
| ## f.education=f.education-University-Or-More | 28.3641639 | |
| 65.1666667 55.14 | | |
| ## race=White | 25.8450704 | |
| 91.7500000 85.20 | | |

| | |
|---|------------|
| ## f.age=f.age-(49,90] | 31.8342152 |
| 30.0833333 22.68 | |
| ## country=United-States | 24.7814685 |
| 94.5000000 91.52 | |
| ## f.age=f.age-(29,39] | 26.8274303 |
| 29.6666667 26.54 | |
| ## race=Amer-Indian-Eskimo | 6.6666667 |
| 0.2500000 0.90 | |
| ## occupation=Priv-house-serv | 0.0000000 |
| 0.0000000 0.50 | |
| ## f.education=f.education-Some-college | 19.7080292 |
| 18.0000000 21.92 | |
| ## occupation=Machine-op-inspct | 13.5313531 |
| 3.4166667 6.06 | |
| ## f.type=f.typ-Private | 22.0011534 |
| 63.5833333 69.36 | |
| ## f.marital=f.marital-Widowed | 9.4117647 |
| 1.3333333 3.40 | |
| ## country=Mexico | 4.2553191 |
| 0.3333333 1.88 | |
| ## occupation=Handlers-cleaners | 5.7777778 |
| 1.0833333 4.50 | |
| ## relationship=Other-relative | 2.9585799 |
| 0.4166667 3.38 | |
| ## race=Black | 11.0453649 |
| 4.6666667 10.14 | |
| ## occupation=Adm-clerical | 11.8715084 |
| 7.0833333 14.32 | |
| ## relationship=Unmarried | 5.3465347 |
| 2.2500000 10.10 | |
| ## f.marital=f.marital-No- Married | 8.8709677 |
| 6.4166667 17.36 | |
| ## relationship=Not-in-family | 11.1026616 |
| 12.1666667 26.30 | |
| ## f.education=f.education-Non-Graduate | 5.7352941 |
| 3.2500000 13.60 | |
| ## sex=Female | 12.4330756 |
| 17.4166667 33.62 | |
| ## occupation=Other-service | 2.4911032 |
| 1.1666667 11.24 | |
| ## relationship=Own-child | 0.6596306 |
| 0.4166667 15.16 | |
| ## f.benefici=f.benefici-Neutre | 18.8262371 |
| 68.1666667 86.90 | |
| ## f.age=f.age-[17,29] | 6.4666667 |
| 8.0833333 30.00 | |
| ## f.marital=f.marital-Never-married | 5.0435866 |
| 6.7500000 32.12 | |
| ## | p.value |
| v.test | |

| | | |
|---|---------------|---|
| ## f.marital=f.marital-Married | 2.216248e-219 | |
| 31.616131 | | |
| ## relationship=Husband | 3.419847e-159 | |
| 26.883554 | | |
| ## f.benefici=f.benefici-Positiu | 2.478837e-56 | |
| 15.814317 | | |
| ## occupation=Exec-managerial | 1.721208e-46 | |
| 14.316693 | | |
| ## sex=Male | 1.146411e-45 | |
| 14.184277 | | |
| ## occupation=Prof-specialty | 4.052161e-42 | |
| 13.599132 | | |
| ## f.benefici=f.benefici-Negatiu | 1.139926e-30 | |
| 11.512597 | | |
| ## f.education=f.education-Proof-school | 4.016679e-28 | |
| 10.995503 | | |
| ## f.age=f.age-(39,49] | 5.134904e-27 | |
| 10.763211 | | |
| ## relationship=Wife | 2.464475e-26 | |
| 10.617743 | | |
| ## f.type=f.typ-SelfEm | 2.616307e-24 | |
| 10.173034 | | |
| ## f.education=f.education-University-Or-More | 7.217966e-16 | |
| 8.066775 | | |
| ## race=White | 1.197214e-14 | |
| 7.716336 | | |
| ## f.age=f.age-(49,90] | 6.429593e-12 | |
| 6.869791 | | |
| ## country=United-States | 9.802713e-06 | |
| 4.421480 | | |
| ## f.age=f.age-(29,39] | 5.232697e-03 | |
| 2.792348 | | |
| ## race=Amer-Indian-Eskimo | 2.821914e-03 | - |
| 2.986499 | | |
| ## occupation=Priv-house-serv | 1.028205e-03 | - |
| 3.282693 | | |
| ## f.education=f.education-Some-college | 1.338104e-04 | - |
| 3.819338 | | |
| ## occupation=Machine-op-inspct | 3.447791e-06 | - |
| 4.642164 | | |
| ## f.type=f.typ-Private | 8.381738e-07 | - |
| 4.926262 | | |
| ## f.marital=f.marital-Widowed | 8.112097e-07 | - |
| 4.932651 | | |
| ## country=Mexico | 2.048243e-07 | - |
| 5.194905 | | |
| ## occupation=Handlers-cleaners | 1.132376e-13 | - |
| 7.424464 | | |
| ## relationship=Other-relative | 1.448749e-14 | - |
| 7.691980 | | |


```
## race=Black 1.022517e-14 -
7.736424
## occupation=Adm-clerical 2.420233e-18 -
8.735760
## relationship=Unmarried 6.071266e-32 -
11.762765
## f.marital=f.marital-No- Married 1.046142e-35 -
12.473145
## relationship=Not-in-family 1.722531e-41 -
13.492874
## f.education=f.education-Non-Graduate 1.326698e-41 -
13.512107
## sex=Female 1.146411e-45 -
14.184277
## occupation=Other-service 5.499322e-51 -
15.019169
## relationship=Own-child 1.064373e-89 -
20.081814
## f.benefici=f.benefici-Neutre 1.023816e-92 -
20.423976
## f.age=f.age-[17,29] 2.817808e-95 -
20.709945
## f.marital=f.marital-Never-married 9.012756e-125 -
23.758325
```

```
##
```

```
##
```

```
## Link between the cluster variable and the quantitative
variables
```

```
##
```

```
=====
```

```
## Eta2 P-value
```

```
## hr.per.week 0.05278220 6.885118e-61
```

```
## age 0.04774912 4.083489e-55
```

```
##
```

```
## Description of each cluster by quantitative variables
```

```
## =====
```

```
## $`<=50K`
```

```
## v.test Mean in category Overall mean sd in
```

```
category
```

```
## age -15.44985 37.01421 38.70380
```

```
14.164738
```

```
## hr.per.week -16.24371 38.60780 39.83755
```

```
9.657673
```

```
## Overall sd p.value
```

```
## age 13.759406 7.562039e-54
```

```
## hr.per.week 9.525189 2.474817e-59
```

```
##
```

```
## $`>50K`
```


```
## v.test Mean in category Overall mean sd in
```

```
category
```

| | | | |
|----------------|------------|--------------|----------|
| ## hr.per.week | 16.24371 | 43.73175 | 39.83755 |
| | 7.920081 | | |
| ## age | 15.44985 | 44.05417 | 38.70380 |
| | 10.761640 | | |
| ## | Overall sd | p.value | |
| ## hr.per.week | 9.525189 | 2.474817e-59 | |
| ## age | 13.759406 | 7.562039e-54 | |

Per a la variable si cobren més de 50 mil anuals o no, veiem que per exemple la mitjana d'edat que tenen un sou inferior a 50 mil es menor a la que els tenen major, de 37 anys de mitjana a 44. és a dir que l'edat té una gran importància en el que ve a ser el fet de tenir un sou més elevat, probablement això es degut a que una persona d'edat major té més experiència, i per tant té millor remuneració en el seu àmbit de treball.

També tenim una dada important que són les hores de treball de mitjana, la majoria de persones que cobren més de 50mil són aquelles que setmanalment excedeixen el límit de 40 hores establert a Espanya. Amb el que majoritariamente podem dir que aquelles persones que treballen més hores acostumen a tenir un sou més elevat.

També podem veure  variables importants com els estudis, la gran part de les persones que no tenen estudis, o que tenen uns estudis baixos, acostumen a tenir un sou menor a 50mil anuals. No obstant les persones més preparades, si que tenen un percentatge més alt de cobrar un sou més elevat no obstant no son la majoria que tenen un sou elevat.