Data Quality report

Aleix Pérez, Adrián Jaen

Introducció

Disposem d'un document inicial amb 48842 files (45222 si es treuen els valors desconeguts) i 15 columnes, procedent de la base de dades on s'hi emmagatzemen el cens dels Estats Units. Concretament data de l'any 1994, i és extreta per Barry Becker.

El nostre objectiu és, mitjançant una mostra, determinar si una persona guanya més de 50.000 dòlars en un any.

El primer que farem es generar una mostra de 5000 files totalment aleatòria i anar "polint" les dades. A continuació, amb les dades depurades, les analitzarem emprant mètodes estadístics multidimensionals amb el fi de trobar relacions entre diferents variables, i finalment amb la nostra variable objectiu.

Descripció

Breu descripció de cada variable

- *age*: L'edat de la persona entrevistada(variable numèrica)

- *workclass*: Per a quin tipus d'organització treballa:(*categorical*: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked)

- *fnlwgt*: (final weight,variable numèrica)

- *education*: Tenint en compte el sistema educatiu dels EUA, són els diferents tipus d'educació que han cursar els enquestats(*categorical*: 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*: (numèrica)número d'anys que la persona enquestada ha cursat en algun centre educatiu.

- *marital-status:* Estat civil(*categorical*: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse)

- *occupation*: Sector al que treballen.(*categorical*: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Profspecialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farmingfishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces)

- *relationship*: Tipus de relació (*categorical*: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried)

- *race*: "raça" (*categorical*: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried)

- *sex*: (*categorical*: Female,Male)

- *capital-gain*: (numèrica)Guanys en termes de capital.

- *capital-loss*: (numèrica)Pèrdues en termes de capital.

- *hours-per-week*: (numèrica) Quantes hores a la setmana treballa l'enquestat.

- *native-country*: País d'origen. (*categorical*: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, OutlyingUS(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, Holland-Netherlands.)

- *Y.bin*: (numèrica binària) si els guanys en capital en un any igualen o sobrepassen els 50K, serà 1. Altrament, serà 0.

Data Quality

Primer de tot, hem de tenir tot el nostre *workspace* a punt:

```r
# Load libraries - Set global variables

options(contrasts=c("contr.treatment","contr.treatment"))
requiredPackages<-c("car","FactoMineR","missMDA","chemometrics", "AER")
install.packages(requiredPackages)
lapply(requiredPackages, require, character.only = TRUE)
```

Carreguem totes les llibreries que necessitem. Ara toca carregar totes les dades.

```r
setwd("/Users/aleixperezvidal/Desktop/ADEI")

df<-read.table("adult.data",header=F, sep=",", fill=FALSE,strip.white=TRUE,na.string="?")


set.seed(190898)
llista<-sample(1:nrow(df),5000)
llista<-sort(llista)
llista[1:10]

df<-df[llista,]

rm(list=c("llista"))
save.image("mostra.RData")

```

Com podem observar, llegim adult.data i mitjançant set.seed i sample, tindrem el nostre *data frame* preparat.

```r
# Carreguem la mostra i posem noms
filepath<-"/Users/aleixperezvidal/Desktop/ADEI/"
load(paste0(filepath,"mostra.RData"))

names(df)
names(df)[1]<-"age"
names(df)[c(1:2,13)]<-c("age","workclass","hours.per.week")
names(df)[10]<-"sex"
names(df)[4]<-"education"
names(df)[5]<-"education-num"
names(df)[6]<-"marital.status"
names(df)[7]<-"occupation"
names(df)[3]<-"final-weight"
names(df)[8]<-"relationship"
names(df)[9]<-"race"
names(df)[11]<-"capital.gain"
names(df)[12]<-"capital.loss"
names(df)[14]<-"native-country"
names(df)[15]<-"Y.bin"
```

Assignem el respectiu nom a cada variable.

Executem la comanda summary, que serveix per fer un resum del *data frame* o la variable a avaluar:

```
summary(df)
```

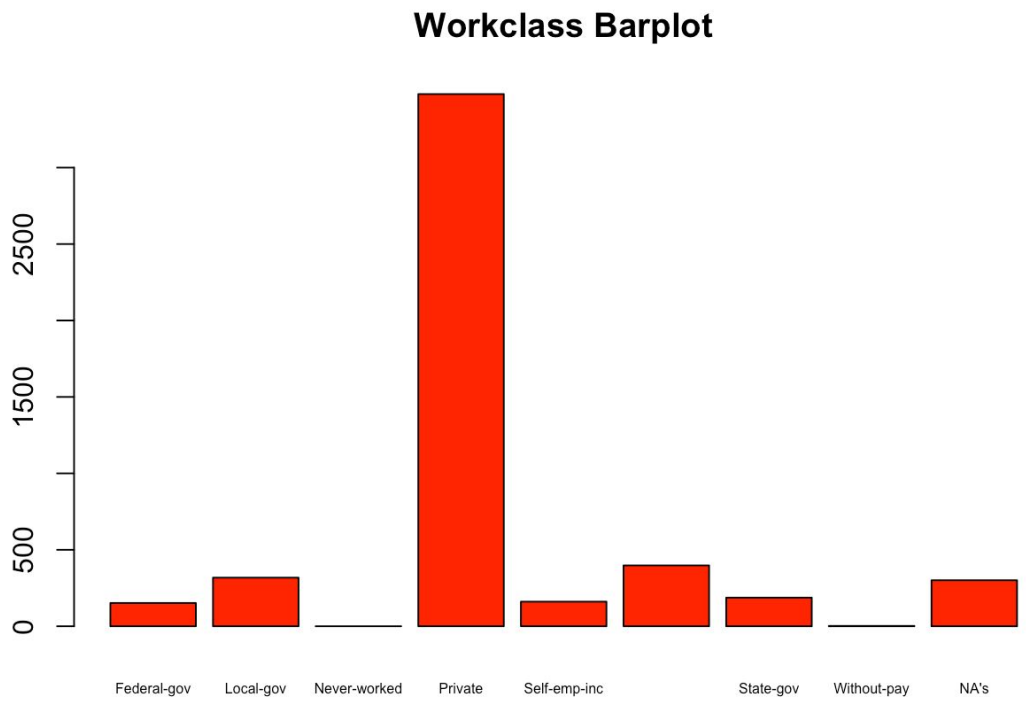
```
##       age                   workclass     final-weight           education
##  Min.   :17.00   Private         :3481   Min.   : 12285   HS-grad     :1547
##  1st Qu.:28.00   Self-emp-not-inc: 398   1st Qu.:117528   Some-college:1131
##  Median :37.00   Local-gov       : 318   Median :178519   Bachelors   : 884
##  Mean   :38.66   State-gov       : 187   Mean   :188622   Masters     : 257
##  3rd Qu.:48.00   Self-emp-inc    : 161   3rd Qu.:238385   Assoc-voc   : 209
##  Max.   :90.00   (Other)         : 154   Max.   :806316   11th        : 166
##                  NA's            : 301                    (Other)     : 806
##  education-num                marital.status           occupation
##  Min.   : 1.00   Divorced             : 689   Prof-specialty : 664
##  1st Qu.: 9.00   Married-AF-spouse    :   4   Exec-managerial: 655
##  Median :10.00   Married-civ-spouse   :2284   Craft-repair   : 584
##  Mean   :10.11   Married-spouse-absent:  66   Adm-clerical   : 580
##  3rd Qu.:13.00   Never-married        :1638   Sales          : 553
##  Max.   :16.00   Separated            : 154   (Other)        :1663
##                  Widowed              : 165   NA's           : 301
##          relationship                 race          sex        capital.gain
##  Husband       :2003   Amer-Indian-Eskimo:  53   Female:1673   Min.   :    0
##  Not-in-family :1270   Asian-Pac-Islander: 162   Male  :3327   1st Qu.:    0
##  Other-relative: 178   Black             : 469                 Median :    0
##  Own-child     : 769   Other             :  42                 Mean   : 1044
##  Unmarried     : 529   White             :4274                 3rd Qu.:    0
##  Wife          : 251                                           Max.   :99999
##
##   capital.loss     hours.per.week        native.country   Y.bin
##  Min.   :   0.00   Min.   : 1.00   United-States:4463   <=50K:3780
##  1st Qu.:   0.00   1st Qu.:39.00   Mexico       :  99   >50K :1220
##  Median :   0.00   Median :40.00   Canada       :  27
##  Mean   :  87.29   Mean   :40.37   Philippines  :  27
##  3rd Qu.:   0.00   3rd Qu.:45.00   Germany      :  25
##  Max.   :4356.00   Max.   :99.00   (Other)      : 266
##                                    NA's         :  93
```

Com es pot observar, necessitem fer un anàlisi variable a variable amb l'objectiu d'identificar missings,errors i outliers. Començarem per les categòriques:

Variables categòriques

Primerament, no considerarem la variable sex perquè no la trobem útil per al nostre estudi.

Workclass



Hi ha masses categories. Les agruparem en 4 grups: Civil, Private, Self-Emp i Other:

```
levels(df$workclass)
hist(as.numeric(df$workclass))
ll<-which(df$workclass=="Private");length(ll)
df$f.type[ll]<-2

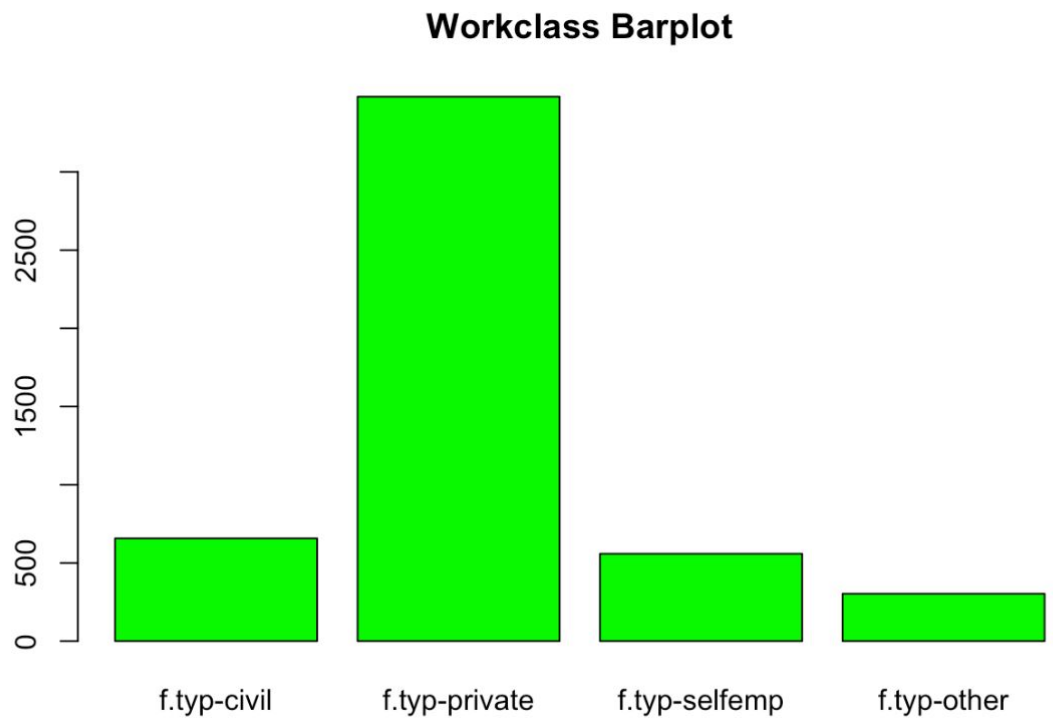
ll<-which((df$workclass=="Self-emp-inc")|((df$workclass=="Self-emp-not-inc")));length(ll)  # Direct

ll<-which(as.numeric(df$workclass)%in%c(5,6));length(ll) # Faster
df$f.type[ll]<-3

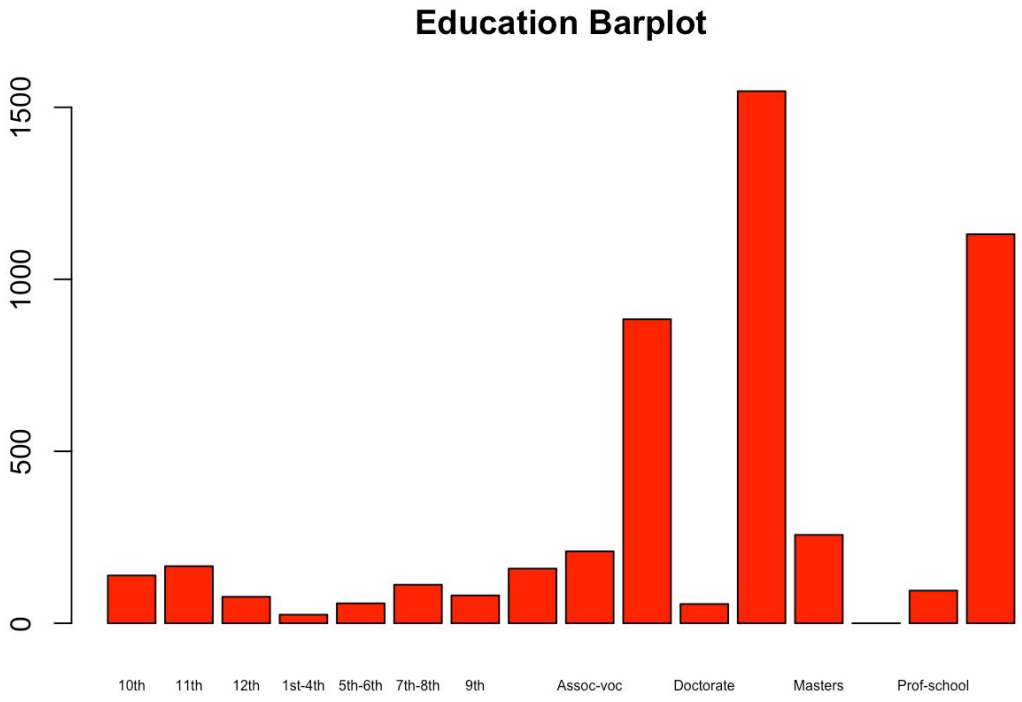
ll<-which(as.numeric(df$workclass)%in%c(1,2,7));length(ll) # Faster
df$f.type[ll]<-1

summary(df$f.type)
df$f.type<-factor(df$f.type,labels=paste0("f.typ-",c("civil","private","selfemp","other")))
```

Hem creat la nova variable f.type per factoritzar correctament les 8 categories anteriors. Queda d'aquesta manera:



Education



Ens trobem amb el mateix problema. I en aquest cas arribem a tenir fins a 16 variables. Procedim de la mateixa manera:

```
#Borramos de la muestra los individius con nivel de educación Preschool ya que práctimente no han empezado la
educación obligatoria
df$f.edu<-4
ll<-which(df$education=="Preschool");length(ll)
if( length(ll)>0) df<-df[-ll,]
levels(df$education)
ll<-which(as.numeric(df$education)%in%c(1,2,3,4,5,6,7));length(ll) # Faster
df$f.edu[ll]<-1
ll<-which(as.numeric(df$education)%in%c(12));length(ll) # Faster
df$f.edu[ll]<-2

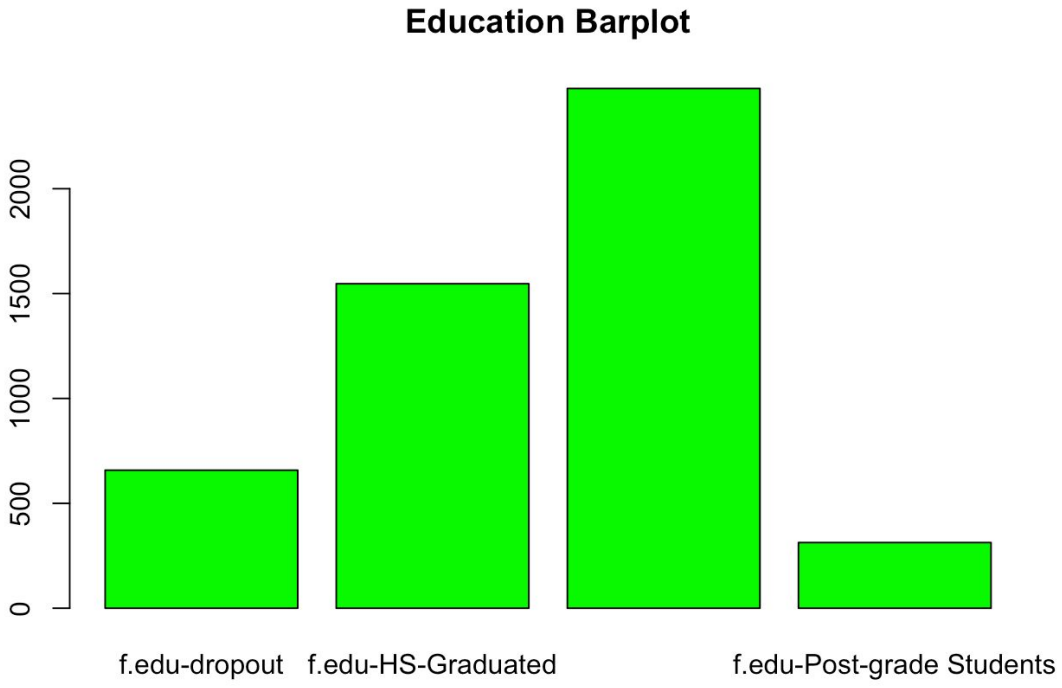
ll<-which(as.numeric(df$education)%in%c(8,9,10,15,16));length(ll) # Faster

df$f.edu[ll]<-3

ll<-which(as.numeric(df$education)%in%c(11,13));length(ll) # Faster
df$f.edu[ll]<-4

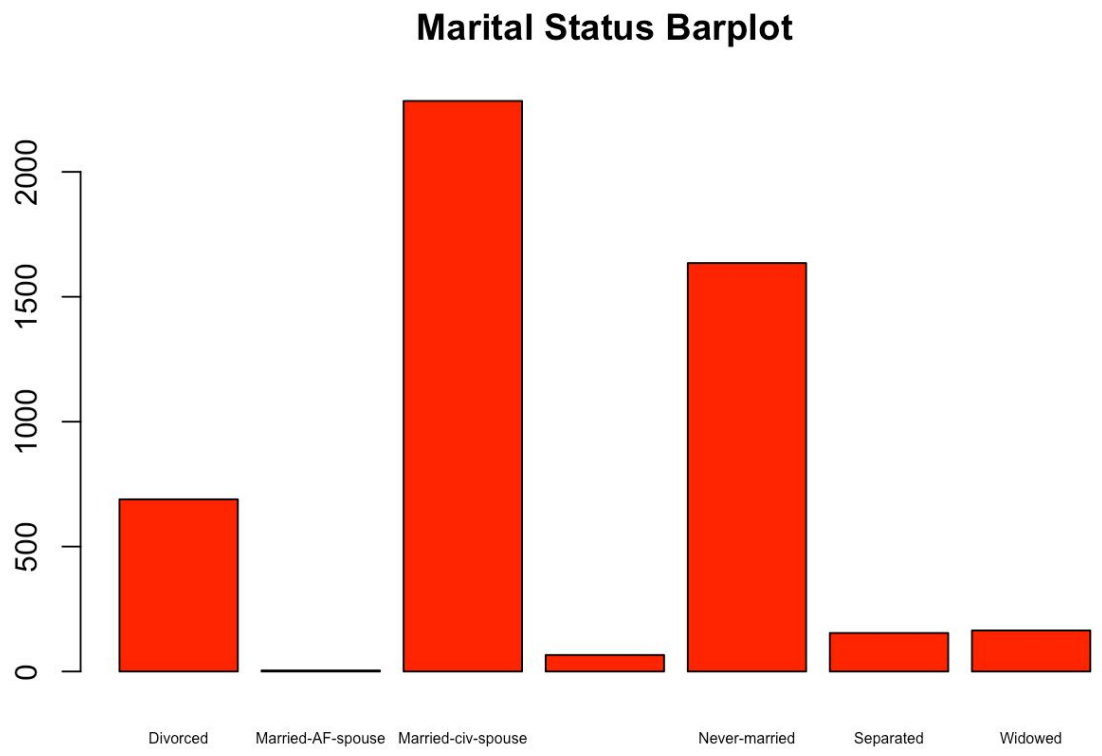
summary(df$f.edu)
df$f.edu<-factor(df$f.edu,labels=paste0("f.edu-",c("dropout","HS-Graduated","Post-school Students","Post-grade
Students")))
```

En aquest cas, eliminarem de la mostra “Preschool”, ja que considerem que pràcticament no han cursat l’ensenyança obligatòria. Com abans, emprarem el prefix f. per denotar la variable factoritzada, concretament en quatre factors.



Marital.status

Procedim a factoritzar la següent variable:




```
df$f.marital<-3
levels(df$marital.status)

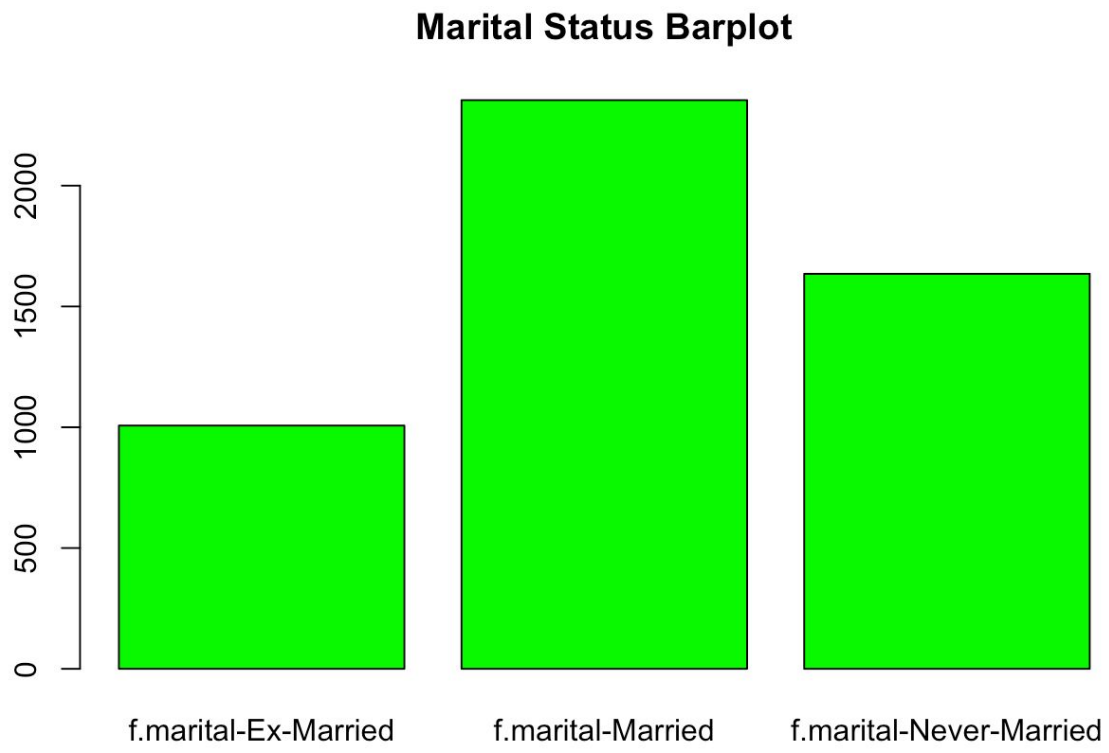
ll<-which(as.numeric(df$marital.status)%in%c(1,6,7));length(ll) # Faster
df$f.marital[ll]<-1

ll<-which(as.numeric(df$marital.status)%in%c(2,3,4));length(ll)
df$f.marital[ll]<-2

ll<-which(as.numeric(df$marital.status)%in%c(5));length(ll)
df$f.marital[ll]<-3

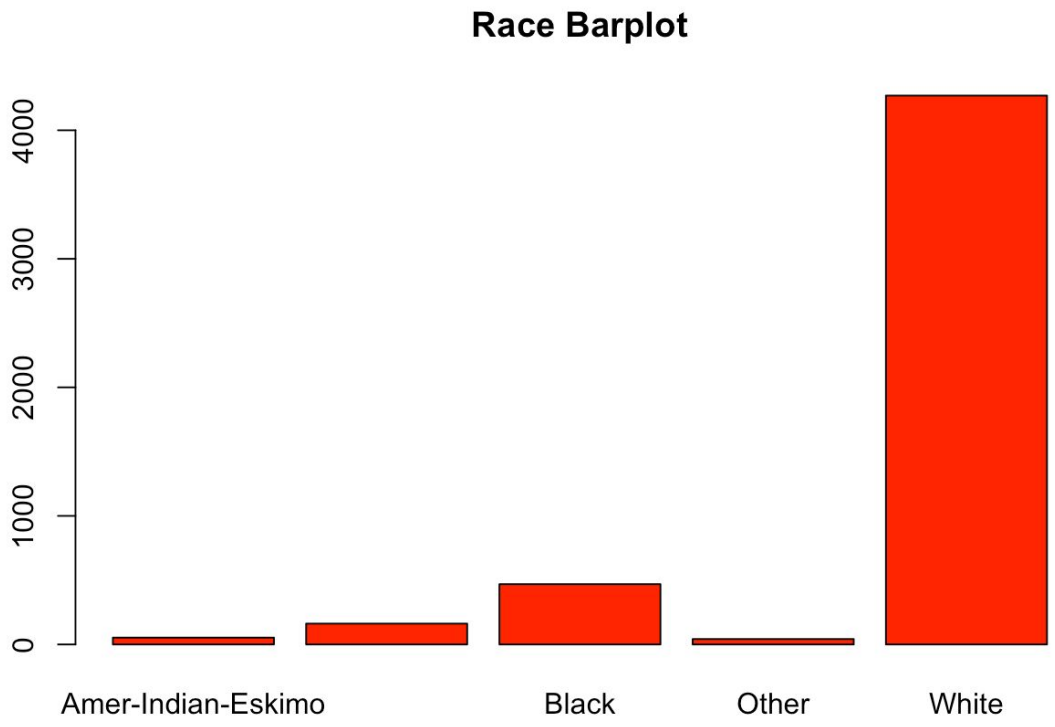
summary(df$f.marital)
df$f.marital<-factor(df$f.marital,labels=paste0("f.marital-",c("Ex-Married","Married","Never-Married")))
```

Ho agruparem de la següent manera: Ex-Married,Married, i Never-Married.



Race

Tenim:



Creiem que, en aquesta mostra, les més importants a tenir en compte són Black i White. La resta es pot factoritzar tot a Others:

```
df$f.race<-3
levels(df$race)

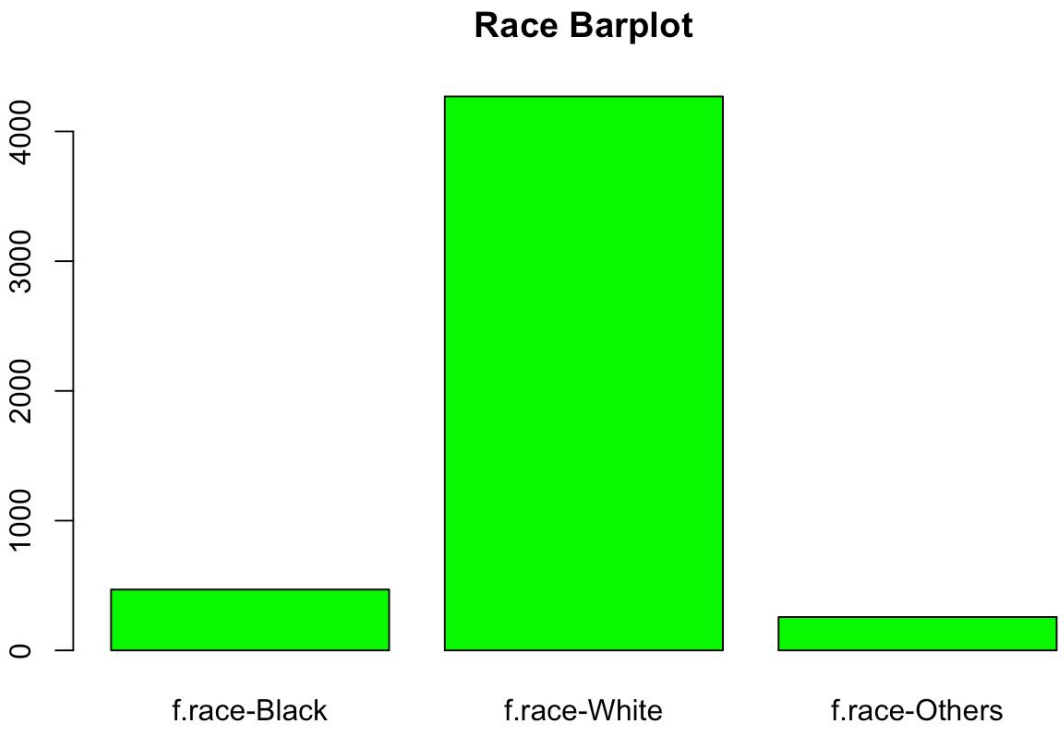
ll<-which(as.numeric(df$race)%in%c(3));length(ll) # Faster
df$f.race[ll]<-1

ll<-which(as.numeric(df$race)%in%c(5));length(ll) # Faster
df$f.race[ll]<-2

ll<-which(as.numeric(df$race)%in%c(1,2,4));length(ll) # Faster
df$f.race[ll]<-3

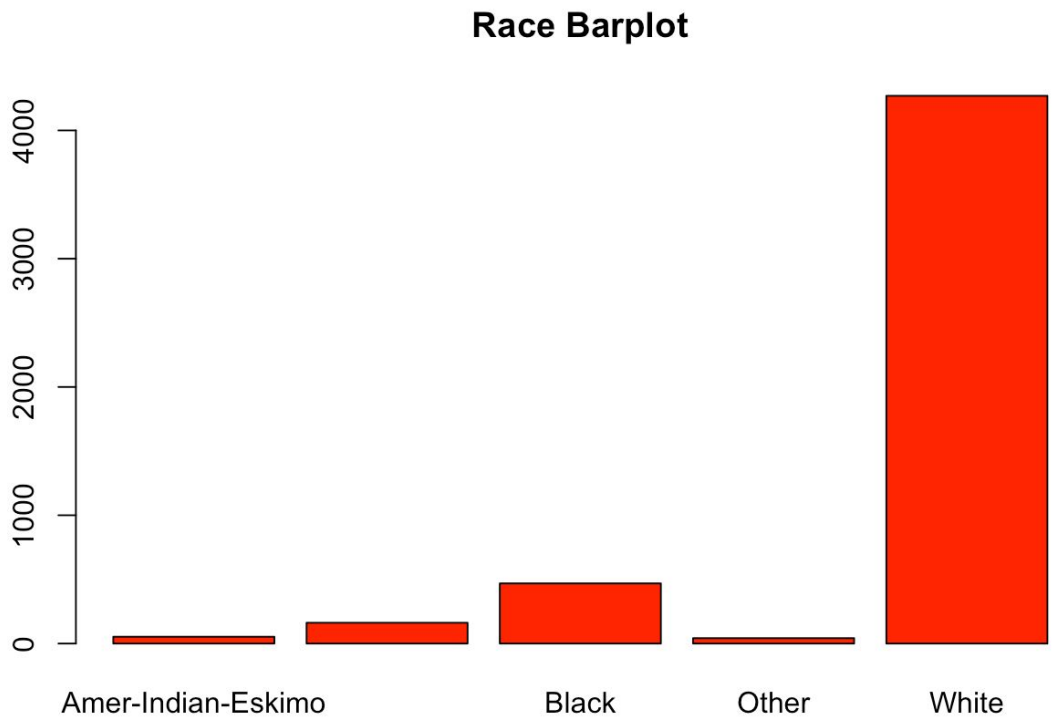
summary(df$f.race)
df$f.race<-factor(df$f.race,labels=paste0("f.race-",c("Black","White","Others")))
```

i tindrem aquest resultat:



Relationship

En aquesta variable tenim grups que podriem tractar-los d'una manera diferent i poder fer ús de la factorització.

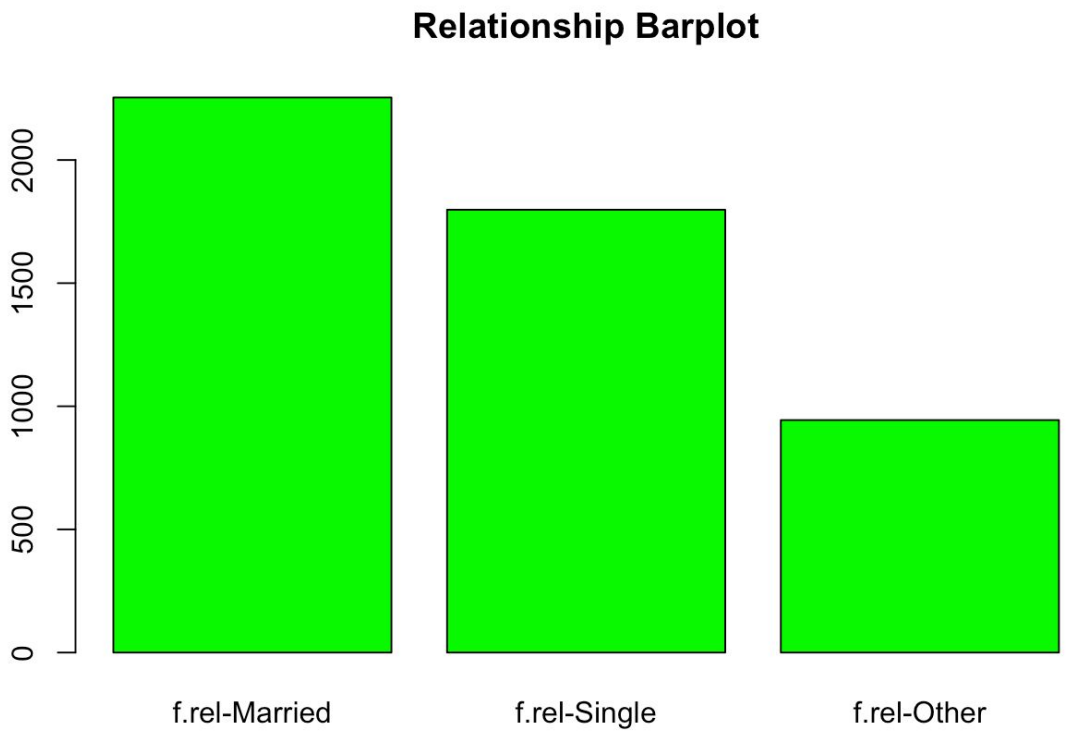


Com hem fet amb les altres, mitjançant el prefix "f." factoritzarem de la manera següent:

```
tapply(df$hours.per.week,df$relationship,mean)
df$f.rel<-3
ll<-which(df$relationship %in% c("Husband","Wife"));length(ll)
df$f.rel[ll]<-1
ll<-which(df$relationship %in% c("Not-in-family","Unmarried"));length(ll)
df$f.rel[ll]<-2
ll<-which(df$relationship %in% c("Other-relative","Own-child"));length(ll)
df$f.rel[ll]<-3

df$f.rel<-factor(df$f.rel,labels=paste0("f.rel-",c("Married","Single","Other")))
```

Entenem com a Married les persones que són Husband i Wife, Single els enquestats que són Other- relative i Own-child, i Other la resta. Queda arreglat d'aquesta manera:



Native country

Al principi crèiem que ordenant els països per continents -ja assumim implícitament que hem de factoritzar una altra vegada una variable daquest tipus- seria el més òptim, però amb les dades a la mà:

(foto barplot paisos)

Veiem que hi ha una sobrecàrrega molt important -i més si les dades són extretes del cens dels EUA- a l'Amèrica del Nord. Per tant, hem decidit agrupar-ho així:

```
df$f.cont<-3
ll<-which(as.numeric(df$native.country)
%in%c(1,9,10,11,12,15,18,19,20,21,22,24,17,25,30,31,32,34,36,37,40,41));length(ll)
df$f.cont[ll]<-1
ll<-which(as.numeric(df$native.country) %in%c(4,5,6,7,8,13,14,16,27,29,33,35,38));length(ll)
df$f.cont[ll]<-2
ll<-which(as.numeric(df$native.country) %in%c(28,2,39));length(ll)
df$f.cont[ll]<-3
df$f.cont<-factor(df$f.cont,labels=paste0("f-count-",c("Eurasia","Center-South-America","NorthAmerica")))

barplot(summary(df$f.cont),main="Native Country Barplot",col = "green")
```

Fixant-nos en les dades, hem fet tres grups diferents: Eurasia,Center-South-America i North- America.

(barplooot)

Occupation

Aquesta és la que ens ha resultat més difícil de factoritzar, ja que costa trobar un factor comú i a la vegada aconseguir un bon balanceig en la mostra. S'havia proposat de agrupar-ho en sectors (primari,secundari, terciari), però tenia certes mancances en termes de balanceig. Disposem d'aquestes dades de moment:

(barplot raw)

Hem arribat a un acord i hem decidit fer-ho de la següent manera:

```
tapply(df$hours.per.week,df$occupation,mean)
summary(df$occupation)
df$f.occ<-4
ll<-which(df$occupation %in% c("Adm-clerical","Armed-Forces","Farming-fishing","Priv-house-serv","Other-service","Pro
tective-serv"));length(ll)
df$f.occ[ll]<-1
ll<-which(df$occupation %in% c("Tech-support","Craft-repair"));length(ll)
df$f.occ[ll]<-2
ll<-which(df$occupation %in% c("Handlers-cleaners","Transport-moving","Machine-op-inspct"));length(ll)
df$f.occ[ll]<-3
ll<-which(df$occupation %in% c("Exec-managerial","Prof-specialty", "Sales"));length(ll)
df$f.occ[ll]<-4
df$f.occ<-factor(df$f.occ,labels=paste0("f.occ-",c("Other","Technics","Basic-Services","Professionals")))
```

Per assolir un correcte balanç, hem agrupat en “Technics” només dues categories, però les dues tenien un número bastant alt; a “Professionals” hem inclòs els càrrecs clàssics d’una empresa. La resta de categories, poc nombroses, les hem agrupat a Other.

Variables contínues

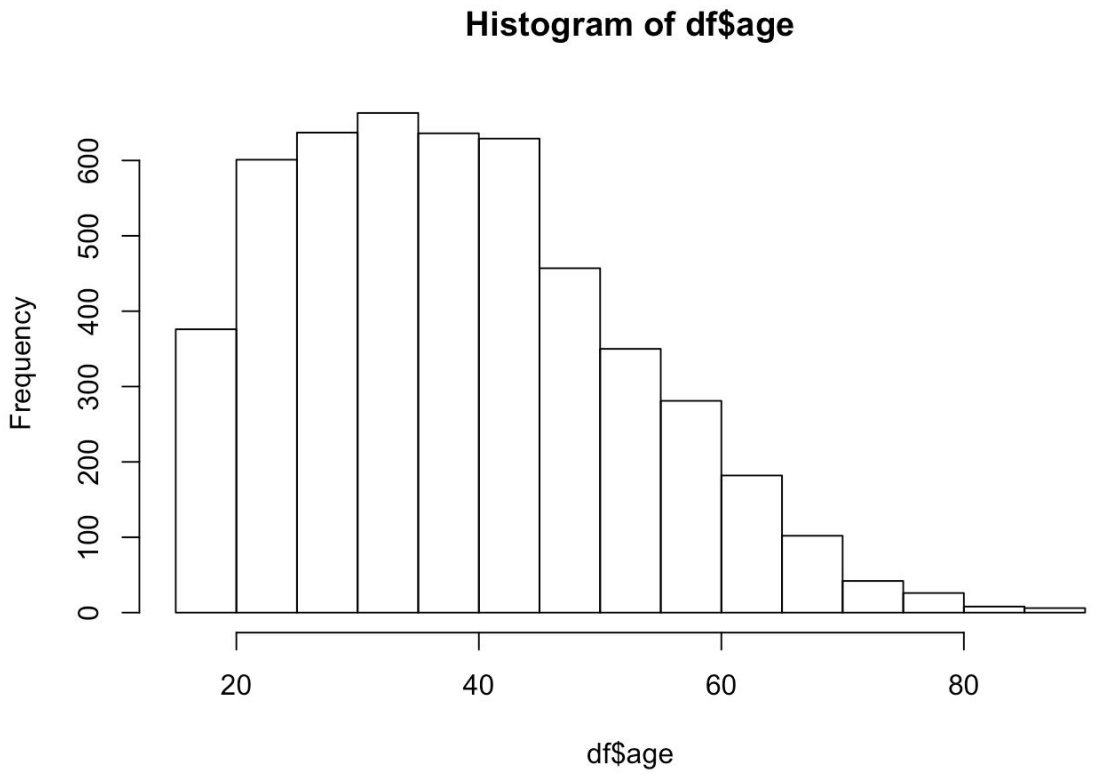
Primerament, no considerarem la variable fnlwgt perquè no la trobem útil per al nostre estudi.

Age

Comencem la primera variable contínua fent un summary;

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.00   28.00   37.00   38.67   48.00   90.00
```

Podem observar que l'edat de la nostra mostra abarca de 17 fins a 90 anys, dividida en quartils. Aquests poden servir-nos per "tallar" la variable en 4 parts. Però abans, mirem com està repartida:



Si la factoritzessim d'aquesta manera, es veuria així:

(barplot quantil)

S'observa que no acaba de reflectir realment la variable; no obstant, hem decidit separar- ho d'una altra forma: "tallant" a 30 40 i 50:

(comandos)

(barplot)

Fent-ho així reflecteix molt millor la variable sense factoritzar.

Capital.gain

Quan fem la descomposició per quartils fins al tercer quartil (75%) les mostres són 0, això significa que més del 50% de les mostres són un 0 en capital.gain.

En aquest cas no es pot descomposar per quartils així que hem tallat per un 1.

Aquí podem veure que la majoria de mostres són 0.

Capital loss

Amb la variable de capital.loss passa exactament el mateix que amb la capital.gain, fins al tercer quartil les mostres són 0.

Quan tallem per l'interval 1 podem veure com en el cas anterior que 4768 individus tenen com a capital.loss 0 o 1.

Aquí podem veure una gràfica amb l'interval a 1.

Hours.per.week

```
summary(df$hours.per.week)
quantile(df$hours.per.week,probs=seq(0,1,0.1),na.rm=T)
hist(df$hours.per.week,50)

ll<-which(df$hours.per.week > 80);length(ll)
#Borramos directamente de la muestra los valores de hours.per.week que consideramos como Outliers
#ya que se trata de el target numerico
if( length(ll)>0) df<-df[-ll,]

df$f.hours<-factor(cut(df$hours.per.week,breaks=c(0,35,40,80),include.lowest = T))

tapply(df$hours.per.week,df$f.hours,median)

df$f.hours<-factor(df$f.hours,labels=paste0("f.age-",levels(df$f.hours)))
summary(df$f.hours)
```

Imputació

En aquest pas realitzarem la imputació de les dades per a donar valor a totes aquelles dades de la mostra que o bé no teniem, o bé hem acabat marcant com a NA degut a que es tractava d'un outlier.

El primer que farem es realitzar la imputació del nostre target numèric, la variable "hours.per.week". Al tractar-se del nostre target numèric, en comptes de donar valor a totes aquelles dades que hem considerat que eren outliers, les eliminarem de la mostra.

```
> quantile(df$hours.per.week,probs=seq(0,1,0.1),na.rm=T)
  0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
   1   24   35   40   40   40   40   40   50   55   99
> #Borramos directamente de la muestra los valores de hours.per.week que consideramos como outliers
> #ya que se trata de el target numerico
> if( length(ll)>0) df<-df[-ll,]
> quantile(df$hours.per.week,probs=seq(0,1,0.1),na.rm=T)
  0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
   1   24   35   40   40   40   40   40   48   55   80
```

Aquí podem veure els quartils de "hours.per.week" abans i després de la imputació, com es pot comprovar, nosaltres hem interpretat que realitzar més de 80 hores a la setmana es podia considerar un outlier ja que estem parlant de realitzar més de dues jornades laborals completes.

Imputació variables categòriques

En el cas de les variables categòriques, observarem aquelles variables que tenien NA's, ja siguin atribuïts per nosaltres degut al tractament d'outliers o no, i cridarem a la funció imputePCA per a que realitzi la imputació.

```
#Imputació variables categòriques
```{r}
vars_dis<-names(df)[c(2,4,6:10,14,15)]
res.input<-imputeMCA(df[,vars_dis],ncp=30)
summary(res.input$completeObs)
```
```

Un cop realitzada la imputació, comprovarem la diferència resultant d'aquesta en aquelles variables que tenien NA's.

```
         workclass          education                  marital.status          occupation          relationship
 Federal-gov     : 152   HS-grad     :1535   Divorced              : 684   Adm-clerical   : 683   Husband       :1981
 Local-gov       : 317   Some-college:1125   Married-AF-spouse     :   4   Prof-specialty : 676   Not-in-family :1263
 Private         :3759   Bachelors   : 880   Married-civ-spouse    :2260   Exec-managerial: 675   Other-relative: 178
 Self-emp-inc    : 158   Masters     : 255   Married-spouse-absent:   66   Craft-repair   : 648   Own-child     : 764
 Self-emp-not-inc: 385   Assoc-voc   : 207   Never-married         :1627   Sales          : 574   Unmarried     : 524
 State-gov       : 186   11th        : 164   Separated             : 154   Other-service  : 541   Wife          : 249
 Without-pay     :   2   (Other)     : 793   Widowed               : 164   (Other)        :1162
                race          sex            native.country   Y.bin
 Amer-Indian-Eskimo:  53   Female:1659   United-States:4520   <=50K:3750
 Asian-Pac-Islander: 159   Male  :3300   Mexico       :  99   >50K :1209
 Black             : 466                 Canada       :  27
 Other             :  41                 Philippines  :  27
 White             :4240                 Germany      :  25
                                         Cuba         :  20
                                         (Other)      : 241
```

Aquí podem veure el resultat de la imputació, i podem apreciar que, tal i com era d'esperar, no tenim cap NA, al contrari del que teníem abans per exemple a Occupation, on en la següent imatge podrem comprovar com hi havien 300 NA.

```
> summary(df$occupation)
     Adm-clerical      Armed-Forces      Craft-repair   Exec-managerial   Farming-fishing Handlers-cleaners
              580                 3               584               653               139               215
Machine-op-inspct     Other-service   Priv-house-serv    Prof-specialty   Protective-serv             Sales
              301               487                20               657                66               546
     Tech-support  Transport-moving              NA's
              150               258               300
```

El mateix passa amb "Native-Country" i "Workclass", variables a les que també teníem un número considerable de NA.

```
> summary(df$native.country)
                  Cambodia                    Canada                     China                  Columbia
                         5                        27                        10                        10
                      Cuba        Dominican-Republic                   Ecuador                El-Salvador
                        20                        14                         3                        12
                   England                    France                   Germany                    Greece
                        12                         5                        25                         2
                 Guatemala                     Haiti      Holand-Netherlands                  Honduras
                        10                         4                         0                         1
                      Hong                   Hungary                     India                      Iran
                         4                         3                        13                         6
                   Ireland                     Italy                   Jamaica                     Japan
                         3                        17                        12                        12
                      Laos                    Mexico                 Nicaragua Outlying-US(Guam-USVI-etc)
                         4                        99                         4                         1
                      Peru                Philippines                    Poland                  Portugal
                         3                        27                         9                         9
                Puerto-Rico                  Scotland                     South                    Taiwan
                        11                         3                        14                         6
                  Thailand            Trinadad&Tobago             United-States                   Vietnam
                         5                         4                      4428                         8
                Yugoslavia                     NA's
                         2                        92
```

```
> summary(df$workclass)
    Federal-gov       Local-gov     Never-worked          Private     Self-emp-inc Self-emp-not-inc        State-gov
            152             318               0            3481              161              398              187
    Without-pay            NA's
              2             301
```

Imputació variables numèriques

A les variables numèriques realitzarem un procediment gairebé calcat al que hem realitzat amb les variables categòriques. Primer els hi aplicarem la funció imputePCA i posteriorment comprovarem els resultats amb les dades que teníem abans de invocar aquesta funció.

```
> summary(df[,vars_con])
      age          final-weight     education-num    capital.gain      capital.loss     hours.per.week
 Min.   :17.00   Min.   : 12285   Min.   : 2.00   Min.   :    0.0   Min.   :   0.00   Min.   : 1.00
 1st Qu.:28.00   1st Qu.:117679   1st Qu.: 9.00   1st Qu.:    0.0   1st Qu.:   0.00   1st Qu.:39.00
 Median :37.00   Median :178644   Median :10.00   Median :    0.0   Median :   0.00   Median :40.00
 Mean   :38.65   Mean   :188774   Mean   :10.11   Mean   :  566.3   Mean   :  86.27   Mean   :39.99
 3rd Qu.:48.00   3rd Qu.:238389   3rd Qu.:13.00   3rd Qu.:    0.0   3rd Qu.:   0.00   3rd Qu.:45.00
 Max.   :90.00   Max.   :806316   Max.   :16.00   Max.   :27828.0   Max.   :4356.00   Max.   :80.00
  .                                                NA's   :24
```

En aquest cas podem comprovar que la única variable en la que tenim NA, és a Capital Gain, a més, aquests NA han sigut atribuits per nosaltres ja que durant l'observació de la mostra hem vist que teníem 24 outliers.
A continuació podem veure el resultat de la imputació.

```
> summary(res_num$completeObs)
      age          final-weight     education-num    capital.gain      capital.loss     hours.per.week
 Min.   :17.00   Min.   : 12285   Min.   : 2.00   Min.   :    0.0   Min.   :   0.00   Min.   : 1.00
 1st Qu.:28.00   1st Qu.:117679   1st Qu.: 9.00   1st Qu.:    0.0   1st Qu.:   0.00   1st Qu.:39.00
 Median :37.00   Median :178644   Median :10.00   Median :    0.0   Median :   0.00   Median :40.00
 Mean   :38.65   Mean   :188774   Mean   :10.11   Mean   :  568.4   Mean   :  86.27   Mean   :39.99
 3rd Qu.:48.00   3rd Qu.:238389   3rd Qu.:13.00   3rd Qu.:    0.0   3rd Qu.:   0.00   3rd Qu.:45.00
 Max.   :90.00   Max.   :806316   Max.   :16.00   Max.   :27828.0   Max.   :4356.00   Max.   :80.00
> quantile(res_num$completeObs[,"capital.gain"])
   0%   25%   50%   75%  100%
    0     0     0     0 27828
```

En aquest cas veiem que els quartils de capital gain continuen absolutament igual i que els NA han desaparegut.

Profiling

En aquest el nostre objectiu és realitzar un perfil del tipus de persona que és més probable que compleixi el nostre target, o millor dit, mirar quines son les variables que estan relacionades amb el nostre target, i a la vegada veure quins valors d'aquestes variables fan que una persona tingui més probabilitats de guanyar més de 50k $ a l'any.

El primer que realitzarem és el profiling del nostre numèric target "hours.per.week" i així veurem quin és el perfil de persona que treballa més hores (o menys).

```
# Profiling

## Target numeric: hours per week

```{r}
library(FactoMineR)
names(df)
condes(df,num.var=13) # Dificult interpretation

vars_condes<-names(df)[c(1,5,10:13,15:28)]
vars_condes  # Check number position
condes(df[,vars_condes],num.var=6,proba=0.01)

# Manual check
tapply(df$hr.per.week,df$f.relship,mean)

```
```

```
$quanti
              correlation      p.value
education-num  0.15932697 1.484034e-29
capital.gain   0.08652662 1.040102e-09
capital.loss   0.06437605 5.707741e-06
age            0.06435106 5.755680e-06

$quali
                     R2      p.value
f.hours      0.739555057 0.000000e+00
f.rel        0.079808687 3.090319e-90
f.age        0.067777043 4.482888e-75
sex          0.059268148 7.973067e-68
Y.bin        0.057649063 5.735712e-66
f.type       0.056911962 1.207088e-62
f.marital    0.050999331 4.639093e-57
f.educn      0.029279569 1.026817e-31
f.occ        0.025715590 8.450099e-28
f.education  0.020089803 1.172813e-21
f.cgain      0.007419390 9.674324e-09
f.closs      0.003258003 3.076615e-04
f.race       0.002253501 3.733247e-03
```

En la imatge anterior podem veure com la educació està molt relacionada amb el nostre target numèric, ja que, si agaféssim com a H0 la hipòtesi de que no estan relacionats, com el p-valor es tant petit, es rebutja aquesta hipòtesi. El mateix passa amb el capital gain, el capital loss i l'edat, encara que no de forma tan pronunciada.

```
$category
                                          Estimate      p.value
f.hours=f.age-(40,80]                   13.8030109 0.000000e+00
sex=Male                                 3.0614414 7.973067e-68
Y.bin=>50K                               3.3178961 5.735712e-66
f.rel=f.rel-Married                      4.0128653 3.223498e-51
f.marital=f.marital-Married              3.1056176 1.696789e-50
f.age=f.age-(30,40]                      3.2168231 2.386461e-34
f.type=f.typ-selfemp                     6.4230031 2.849249e-34
f.educn=(13,16]                          4.1079281 1.524231e-19
f.age=f.age-(40,50]                      2.5369729 2.585203e-18
f.education=f.education-Post-grade Students   3.8199580 5.806469e-10
f.educn=(10,13]                          0.7614448 1.067088e-09
f.occ=f.occ-Professionals                0.8128819 8.316376e-07
f.occ=f.occ-Technics                     1.6924728 6.219748e-06
f.closs=f.closs-Yes                      2.0901236 5.776728e-05
f.cgain=NA                               5.1937815 1.790988e-04
f.race=f.race-Black                     -1.4141564 1.057953e-03
f.closs=f.closs-No                      -1.1239702 5.981734e-05
f.cgain=f.cgain-Yes                     -1.0646874 2.594737e-06
f.age=f.age-(50,90]                     -1.9054967 8.165682e-07
f.educn=[2,9]                           -2.0017976 1.655246e-07
f.cgain=f.cgain-No                      -4.1290941 3.164727e-08
f.educn=(9,10]                          -2.8675753 2.218833e-09
f.education=f.education-dropout         -3.8633541 3.891339e-17
f.occ=f.occ-Other                       -3.2925335 3.460184e-29
f.type=f.typ-other                      -7.7405170 5.752358e-37
f.marital=f.marital-Never-Married       -2.9507124 2.727840e-46
f.age=f.age-[0,30]                      -3.8482994 9.724403e-51
Y.bin=<=50K                             -3.3178961 5.735712e-66
sex=Female                              -3.0614414 7.973067e-68
f.rel=f.rel-Other                       -5.1588578 4.789879e-78
f.hours=f.age-[0,35]                   -14.8841607 0.000000e+00
```

En aquesta imatge podem observar realment quin és el perfil de persona en relació al nostre target numèric. El primer que veiem és que les persones d'entre 40 i 80 anys treballen una mitjana de 13.8 hores més que la resta, però també veiem que el p-valor es 0, per tant si agafem la hipòtesi nula de que això no és real, acceptem aquesta hipòtesi.

El següent que podem veure és que els homes treballen de mitjana 3 hores més que les dones, també veiem que els autònoms són els que més treballen i que els casats treballen més que els no casats. Si continuem llegint els resultats podem veure que quan més alt és el nivell d'educació, major és la quantitat d'hores treballades, i que les persones de entre 40 i 50 anys treballen més hores.

A continuació realitzarem el profiling per al target categòric:

```
## Profile categorical target: y.bin

```{r}
names(df)
# vars_catdes or vars_condes is already valid
vars_condes
catdes(df[,vars_condes],num.var=which(vars_condes=="Y.bin"),proba=0.01)
```
```

```
Link between the cluster variable and the categorical variables (chi-square test)
================================================================================
                  p.value df
f.rel       8.951000e-221  2
f.marital   7.635094e-203  2
f.educn     2.405842e-125  3
f.age        1.468562e-90  3
f.hours      6.752123e-85  2
f.education  1.121939e-83  3
f.cgain      4.274204e-82  2
f.occ        2.911233e-69  3
sex          7.182485e-54  1
f.type       5.345429e-27  3
f.closs      5.131073e-26  2
f.race       1.536459e-09  2
f.fnlwgt     7.434736e-04  4
f.cont       9.412567e-04  2
```

Podem observar totes les variables estan relacionades amb el target, ja que totes tenen un p-valor molt petit, i per tant descartaria la hipòtesi nula de que no estan relacionades.

$`>50K`

	Cla/Mod	Mod/Cla	Global	p.value	v.test
f.rel=f.rel-Married	45.605381	84.1191067	44.9687437	5.293941e-229	32.308616
f.marital=f.marital-Married	43.991416	84.7808106	46.9852793	1.307426e-214	31.267100
f.hours=f.age-(40,80]	41.453287	49.5450786	29.1389393	4.596981e-68	17.433470
f.educn=(13,16]	61.346633	20.3473945	8.0863077	3.336141e-61	16.505741
sex=Male	31.060606	84.7808106	66.5456745	4.060586e-59	16.213316
f.occ=f.occ-Professionals	35.575139	63.4408602	43.4765074	3.908414e-58	16.073581
f.cgain=f.cgain-Yes	61.246612	18.6931348	7.4410163	1.077846e-55	15.721475
f.education=f.education-Post-grade Students	56.818182	14.4747725	6.2109296	2.276463e-36	12.594063
f.educn=(10,13]	37.560193	38.7096774	25.1260335	5.811873e-34	12.148934
f.age=f.age-(40,50]	37.627433	33.5814723	21.7584190	1.638763e-28	11.076089
f.closs=f.closs-Yes	54.112554	10.3391232	4.6581972	3.037678e-23	9.931462
f.education=f.education-Post-school Students	29.788961	60.7113317	49.6874370	9.891084e-19	8.836334
f.type=f.typ-selfemp	39.226519	17.6178660	10.9497883	3.663237e-16	8.149205
f.cgain=NA	100.000000	1.9851117	0.4839685	1.633764e-15	7.966390
f.age=f.age-(50,90]	33.973711	27.7915633	19.9435370	2.184247e-14	7.639289
f.race=f.race-White	25.872642	90.7361456	85.5011091	6.342122e-10	6.181684
f.type=f.typ-civil	30.886850	16.7080232	13.1881428	4.744357e-05	4.067876
f.fnlwgt=(1.18e+05,1.79e+05]	28.376206	29.1976840	25.0857028	1.762023e-04	3.750898
f.age=f.age-(30,40]	27.492212	29.1976840	25.8923170	2.772469e-03	2.991900
f.fnlwgt=(2.38e+05,8.06e+05]	20.967742	21.5053763	25.0050413	1.103506e-03	-3.262714
f.cont=f-count-Center-South-America	10.909091	0.9925558	2.2181892	3.684772e-04	-3.561689
f.hours=f.age-(35,40]	21.330561	42.4317618	48.4976810	1.181189e-06	-4.858767
f.type=f.typ-private	22.109827	63.2754342	69.7721315	2.290256e-08	-5.588510
f.educn=(9,10]	17.333333	16.1290323	22.6860254	1.320656e-10	-6.424774
f.race=f.race-Black	12.875536	4.9627792	9.3970559	1.168527e-10	-6.443366
f.type=f.typ-other	9.602649	2.3986766	6.0899375	1.762151e-11	-6.724485
f.occ=f.occ-Basic-Services	14.341085	9.1811414	15.6079855	1.236549e-13	-7.412806
f.education=f.education-HS-Graduated	17.394137	22.0843672	30.9538213	4.707505e-15	-7.834490
f.closs=f.closs-No	22.932092	89.6608768	95.3216374	5.102382e-23	-9.879624
f.marital=f.marital-Ex-Married	10.978044	9.0984285	20.2056866	4.164107e-32	-11.794552
f.education=f.education-dropout	5.061350	2.7295285	13.1478121	1.352856e-44	-14.010063
f.occ=f.occ-Other	10.733591	11.4971050	26.1141359	2.819780e-45	-14.120996
f.hours=f.age-[0,35]	8.746619	8.0231596	22.3633797	4.069469e-50	-14.885904
f.educn=[2,9]	13.717421	24.8138958	44.1016334	9.965060e-57	-15.871611
sex=Female	11.091019	15.2191894	33.4543255	4.060586e-59	-16.213316
f.cgain=f.cgain-No	21.003066	79.3217535	92.0750151	5.681838e-67	-17.289115
f.rel=f.rel-Single	9.904868	14.6401985	36.0354910	1.828519e-78	-18.753035
f.age=f.age-[0,30]	7.093964	9.4292804	32.4057270	1.077351e-99	-21.194324
f.rel=f.rel-Other	1.592357	1.2406948	18.9957653	7.625570e-103	-21.533101
f.marital=f.marital-Never-Married	4.548248	6.1207610	32.8090341	3.613838e-139	-25.112699

Aquí podem observar quin és el perfil de les persones que guanyen més de 50k Dollars a l'any segons les categories.

```
Link between the cluster variable and the quantitative variables
================================================================
                      Eta2       P-value
education-num  0.11197502 5.021953e-130
capital.gain   0.09579819 1.467342e-110
age            0.05891218  2.042458e-67
hours.per.week 0.05764906  5.735712e-66
capital.loss   0.02697061  2.546516e-31

Description of each cluster by quantitative variables
=====================================================
$`<=50K`
                 v.test Mean in category Overall mean sd in category  Overall sd       p.value
capital.loss   -11.56375         48.89547     86.27062     300.219646  400.811404  6.290153e-31
hours.per.week -16.90633         38.37413     39.99193      11.868368   11.866737  4.040915e-64
age            -17.09054         36.76613     38.65316      14.011061   13.692320  1.745479e-65
capital.gain   -21.79375        128.78987    568.43185     715.927181 2501.629504 2.659695e-105
education-num  -23.56209          9.62400     10.11393       2.455896    2.578575 9.437307e-123

$`>50K`
                 v.test Mean in category Overall mean sd in category  Overall sd       p.value
education-num  23.56209         11.63358     10.11393        2.34753    2.578575 9.437307e-123
capital.gain   21.79375       1932.08565    568.43185     4649.77571 2501.629504 2.659695e-105
age            17.09054         44.50620     38.65316       10.71390   13.692320  1.745479e-65
hours.per.week 16.90633         45.00993     39.99193       10.36345   11.866737  4.040915e-64
capital.loss   11.56375        202.19851     86.27062      601.33668  400.811404  6.290153e-31
```

Per últim podem veure que les variables quantitatives estan molt relacionades amb aquest target categòric. També podem veure quina és la mitjana de les variables quantitatives d'una persona que guanya més de 50 mil dòlars a l'any, en aquest cas estaríem parlant d'una persona de 44 anys, que treballa unes 45 hores a la setmana, amb uns estudis superiors als obligatoris i que rep més ingressos de forma alterna al treball dels que perd.