PRÁCTICA 2: Limpieza y validación de los datos

Autores: Anna Llorens Roig, Carlos Villar Robles

Dataset: Adults

Contents

1. Descripción del dataset

El conjunto de datos objeto de análisis se ha obtenido del repositorio UCI Machine Learning. Se trata del 'Adult data set' el cual consta de 14 atributos ..... Entre los campos del conjunto de datos encontramos los siguientes: -

1.1. Importancia y objetivos de los análisis

You can also embed plots, for example:

2. Inegración y selección de los datos de interés.

Antes de comenzar con la limpieza de los datos, procedemos a realizar la lectura del fichero en formato CSV en el que se encuentran. El resultado devuelto por la llamada a la función read.csv() será un objeto data.frame:

```
# Lectura de datos
data <-read.csv('../data/adults.csv', header = TRUE)
head(data[,1:5])
```

```
##   X age        workclass fnlwgt education
## 1 1  39        State-gov  77516 Bachelors
## 2 2  50 Self-emp-not-inc  83311 Bachelors
## 3 3  38          Private 215646   HS-grad
## 4 4  53          Private 234721      11th
## 5 5  28          Private 338409 Bachelors
## 6 6  37          Private 284582   Masters
```

```
#Eliminamos la primera columna ya que se trata de un índice.
data$X <- NULL

# Vector con los tipos de variables R para cada variable
sapply(data,class)
```

```
##            age      workclass         fnlwgt      education  education_num
##      "integer"       "factor"      "integer"       "factor"      "integer"
##  maritalStatus     occupation   relationship           race            sex
##       "factor"       "factor"       "factor"       "factor"       "factor"
##   capital_gain   capital_loss  hour_per_week native_country         income
```

```
##      "integer"      "integer"      "integer"       "factor"       "factor"
```

```
# Revisión descriptiva de la matriz de datos
summary(data)
```

```
##       age                    workclass         fnlwgt
##  Min.   :17.00   Private         :22696   Min.   :  12285
##  1st Qu.:28.00   Self-emp-not-inc: 2541   1st Qu.: 117827
##  Median :37.00   Local-gov       : 2093   Median : 178356
##  Mean   :38.58   ?               : 1836   Mean   : 189778
##  3rd Qu.:48.00   State-gov       : 1298   3rd Qu.: 237051
##  Max.   :90.00   Self-emp-inc    : 1116   Max.   :1484705
##                  (Other)         :  981
##         education     education_num                  maritalStatus
##  HS-grad     :10501   Min.   : 1.00   Divorced              : 4443
##  Some-college: 7291   1st Qu.: 9.00   Married-AF-spouse     :   23
##  Bachelors   : 5355   Median :10.00   Married-civ-spouse    :14976
##  Masters     : 1723   Mean   :10.08   Married-spouse-absent:  418
##  Assoc-voc   : 1382   3rd Qu.:12.00   Never-married         :10683
##  11th        : 1175   Max.   :16.00   Separated             : 1025
##  (Other)     : 5134                   Widowed               :  993
##            occupation           relationship                    race
##  Prof-specialty :4140   Husband       :13193   Amer-Indian-Eskimo:  311
##  Craft-repair   :4099   Not-in-family : 8305   Asian-Pac-Islander: 1039
##  Exec-managerial:4066   Other-relative:  981   Black             : 3124
##  Adm-clerical   :3770   Own-child     : 5068   Other             :  271
##  Sales          :3650   Unmarried     : 3446   White             :27816
##  Other-service  :3295   Wife          : 1568
##  (Other)        :9541
##      sex          capital_gain    capital_loss    hour_per_week
##  Female:10771   Min.   :    0   Min.   :   0.0   Min.   : 1.00
##  Male  :21790   1st Qu.:    0   1st Qu.:   0.0   1st Qu.:40.00
##                 Median :    0   Median :   0.0   Median :40.00
##                 Mean   : 1078   Mean   :  87.3   Mean   :40.44
##                 3rd Qu.:    0   3rd Qu.:   0.0   3rd Qu.:45.00
##                 Max.   :99999   Max.   :4356.0   Max.   :99.00
##
##        native_country    income
##  United-States:29170   <=50K:24720
##  Mexico       :  643   >50K : 7841
##  ?            :  583
##  Philippines  :  198
##  Germany      :  137
##  Canada       :  121
##  (Other)      : 1709
```

Observamos que tenemos un total de 32561 observaciones en 15 variables.

3. Limpieza de los datos.

4. Análisis de los datos.

5. Representación de los resultados a partir de tablas y gráficas

6. Resolución del problema.

7. Contribuciones

Contribuciones	Firma
Investigación previa	AL, CV
Redacción de las respuestas	centered
Desarrollo código	are neat