

Visualización de datos: PRA2

Autor: Waziri Ajibola Lawal Mohammed

MAY 2021

Contents

Introducción	1
Realización de la tarea	1
Archivo final	9

Introducción

En esta tarea se realizará la transformación del conjunto de datos escogido para la realización de la PRA2 con el objetivo de generar un juego de datos final que nos permita responder a las preguntas planteadas en la PRA1 mediante visualizaciones interactivas.

Realización de la tarea

A partir del conjunto de datos disponible en el siguiente enlace <http://archive.ics.uci.edu/ml/datasets/Adult>, se generará un nuevo juego de datos para implementar la visualización interactiva.

```
# Cargamos el juego de datos
datosAdult <- read.csv('http://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data',
                      stringsAsFactors = FALSE, header = FALSE, na.strings = "?",
                      strip.white = TRUE, fill = F)

# Nombres de los atributos
names(datosAdult) <- c("age", "workclass", "fnlwgt", "education", "education_num", "marital_status",
                      "occupation", "relationship", "race", "sex", "capital_gain", "capital_loss",
                      "hour_per_week", "native_country", "income")
```

```
# Verificamos la estructura del conjunto de datos
str(datosAdult)
```

```
## 'data.frame':    32561 obs. of  15 variables:
## $ age           : int  39 50 38 53 28 37 49 52 31 42 ...
## $ workclass     : chr  "State-gov" "Self-emp-not-inc" "Private" "Private" ...
## $ fnlwgt        : int  77516 83311 215646 234721 338409 284582 160187 209642 45781 159449 ...
## $ education     : chr  "Bachelors" "Bachelors" "HS-grad" "11th" ...
## $ education_num : int  13 13 9 7 13 14 5 9 14 13 ...
## $ marital_status: chr  "Never-married" "Married-civ-spouse" "Divorced" "Married-civ-spouse" ...
## $ occupation    : chr  "Adm-clerical" "Exec-managerial" "Handlers-cleaners" "Handlers-cleaners" ...
## $ relationship  : chr  "Not-in-family" "Husband" "Not-in-family" "Husband" ...
## $ race          : chr  "White" "White" "White" "Black" ...
## $ sex           : chr  "Male" "Male" "Male" "Male" ...
## $ 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 ...
## $ hour_per_week  : int  40 13 40 40 40 40 16 45 50 40 ...
## $ native_country: chr  "United-States" "United-States" "United-States" "United-States" ...
## $ income        : chr  "<=50K" "<=50K" "<=50K" "<=50K" ...
```

Descripción de las variables contenidas en el fichero:

- age: valor numérico con la edad de la unidad que responde.
- work-class: factor que especifica en qué entidad o sector trabaja la unidad que responde.
- fnlwgt: peso final que describe el número de individuos de la población objetiva que representa el colectivo representado en el conjunto de datos.
- education: especifica el nivel académico más alto completado del individuo.
- education_num: valor numérico que especifica el número de años de educación del individuo.
- marital_status: especifica la relación del individuo con otra persona.
- occupation: especifica el trabajo del individuo.
- relationship: especifica el papel de la familia del individuo.
- race: describe la característica física del individuo.
- sex: factor con dos niveles (masculino y femenino) que especifica el género del individuo.
- capital_gain: valor numérico que especifica las ganancias en los activos de capital del individuo.
- capital_loss: valor numérico que especifica la pérdida incurrida en los activos de capital del individuo.
- hour_per_week: valor numérico que especifica las horas de trabajo semanales del individuo.
- native_country: factor que especifica el país nacimiento del individuo.
- income: atributo objetivo que especifica los ingresos del individuo con nivel ≤ 50 y > 50 .

```
#Estadísticas básicas
summary(datosAdult)
```

```
##      age      workclass      fnlwgt      education
## Min.   :17.00  Length:32561  Min.    : 12285  Length:32561
## 1st Qu.:28.00  Class :character  1st Qu.: 117827  Class :character
## Median :37.00  Mode  :character  Median : 178356  Mode  :character
## Mean   :38.58                      Mean    : 189778
## 3rd Qu.:48.00                      3rd Qu.: 237051
## Max.    :90.00                      Max.    :1484705
## education_num marital_status occupation relationship
## Min.      : 1.00  Length:32561  Length:32561  Length:32561
## 1st Qu.: 9.00  Class :character  Class :character  Class :character
```

```
## Median :10.00 Mode :character Mode :character Mode :character
## Mean :10.08
## 3rd Qu.:12.00
## Max. :16.00
## race sex capital_gain capital_loss
## Length:32561 Length:32561 Min. : 0 Min. : 0.0
## Class :character Class :character 1st Qu.: 0 1st Qu.: 0.0
## Mode :character Mode :character Median : 0 Median : 0.0
## Mean : 1078 Mean : 87.3
## 3rd Qu.: 0 3rd Qu.: 0.0
## Max. :99999 Max. :4356.0
## hour_per_week native_country income
## Min. : 1.00 Length:32561 Length:32561
## 1st Qu.:40.00 Class :character Class :character
## Median :40.00 Mode :character Mode :character
## Mean :40.44
## 3rd Qu.:45.00
## Max. :99.00
```

A continuación, procedemos a contar cuantos nulos (y '?' convertidos a NA) hay en el dataset escogido para la práctica.

```
# Estadística para estudiar si existen valores vacíos
colSums(is.na(datosAdult))
```

```
## age workclass fnlwgt education education_num
## 0 1836 0 0 0
## marital_status occupation relationship race sex
## 0 1843 0 0 0
## capital_gain capital_loss hour_per_week native_country income
## 0 0 0 583 0
```

Vemos que los atributos workclass, occupation y native_country contienen nulos. Debido a que son atributos categóricos no se puede aplicar ningún tipo de media u operación estadística, por lo tanto, procedemos a eliminar los valores nulos.

```
datosAdult <- na.omit(datosAdult)
```

Observamos si es necesario discretizar alguna variable del conjunto de datos.

```
apply(datosAdult,2, function(x) length(unique(x)))
```

```
## age workclass fnlwgt education education_num
## 72 7 20263 16 16
## marital_status occupation relationship race sex
## 7 14 6 5 2
## capital_gain capital_loss hour_per_week native_country income
## 118 90 94 41 2
```

```
# Discretizamos las variables con pocas clases que interesan para este estudio
cols<-c("income","occupation","sex")
```

```

for (i in cols){
  datosAdult[,i] <- as.factor(datosAdult[,i])
}

# Después de los cambios, analizamos la nueva estructura del conjunto de datos
str(datosAdult)

## 'data.frame':    30162 obs. of  15 variables:
## $ age           : int  39 50 38 53 28 37 49 52 31 42 ...
## $ workclass      : chr  "State-gov" "Self-emp-not-inc" "Private" "Private" ...
## $ fnlwgt         : int  77516 83311 215646 234721 338409 284582 160187 209642 45781 159449 ...
## $ education      : chr  "Bachelors" "Bachelors" "HS-grad" "11th" ...
## $ education_num  : int  13 13 9 7 13 14 5 9 14 13 ...
## $ marital_status : chr  "Never-married" "Married-civ-spouse" "Divorced" "Married-civ-spouse" ...
## $ occupation     : Factor w/ 14 levels "Adm-clerical",...: 1 4 6 6 10 4 8 4 10 4 ...
## $ relationship   : chr  "Not-in-family" "Husband" "Not-in-family" "Husband" ...
## $ race           : chr  "White" "White" "White" "Black" ...
## $ 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 ...
## $ hour_per_week  : int  40 13 40 40 40 40 16 45 50 40 ...
## $ native_country : chr  "United-States" "United-States" "United-States" "United-States" ...
## $ income         : Factor w/ 2 levels "<=50K", ">50K": 1 1 1 1 1 1 1 2 2 2 ...
## - attr(*, "na.action")= 'omit' Named int [1:2399] 15 28 39 52 62 70 78 94 107 129 ...
## ..- attr(*, "names")= chr [1:2399] "15" "28" "39" "52" ...

```

Existen variables que se pueden combinar o eliminar ya que añaden información innecesaria al conjunto de datos o se pueden representar mediante otras variables:

- Con el nivel de educación completado (“education”) podemos conocer o representar el número total de años de educación (“education_num”).
- Mediante las variables “sex” que nos indica el género y ‘marital_status’ que nos indica la situación sentimental podemos conocer o representar la variable “relationship”.
- La variable “fnlwgt” representa el peso final que describe el número de individuos de la población y por lo tanto se puede considerar irrelevante.
- Si observamos la variable “native_country”, vemos que la mayoría de observaciones son Estados Unidos, por lo tanto también podemos excluirla.

```

# Reducción de la dimensionalidad
datosAdult$education <- NULL
datosAdult$relationship <- NULL
datosAdult$fnlwgt <- NULL
datosAdult$native_country <- NULL

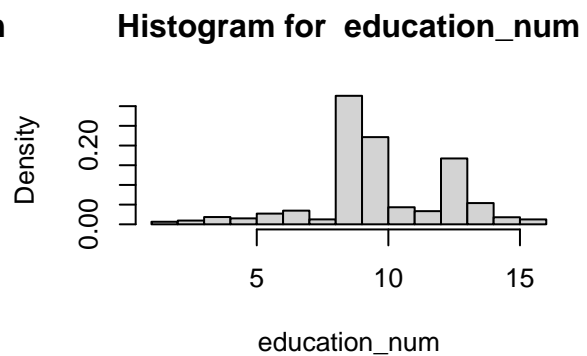
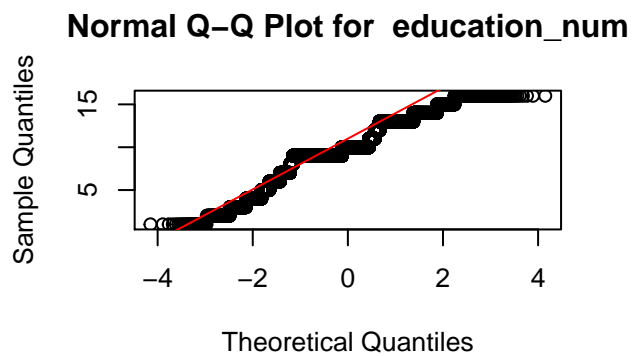
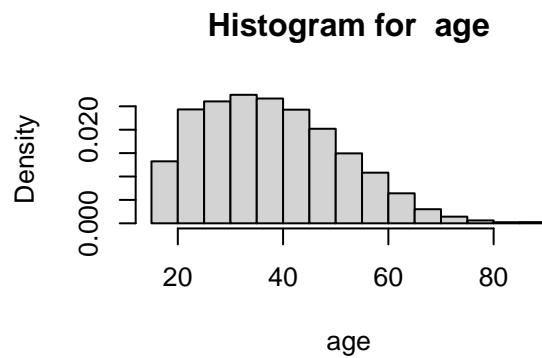
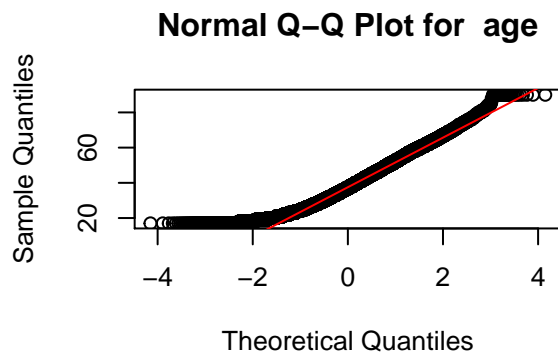
```

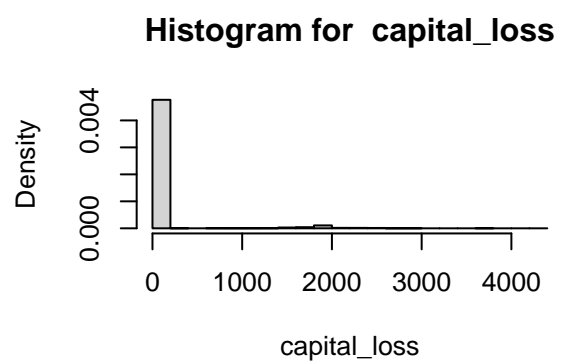
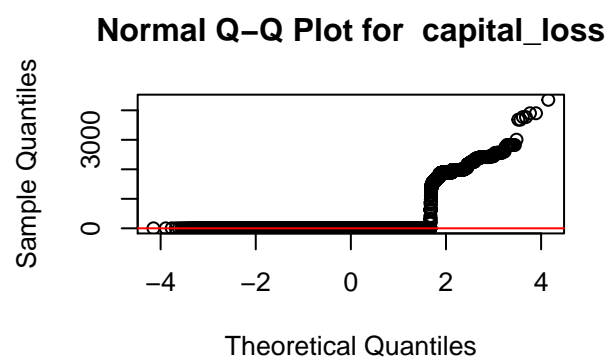
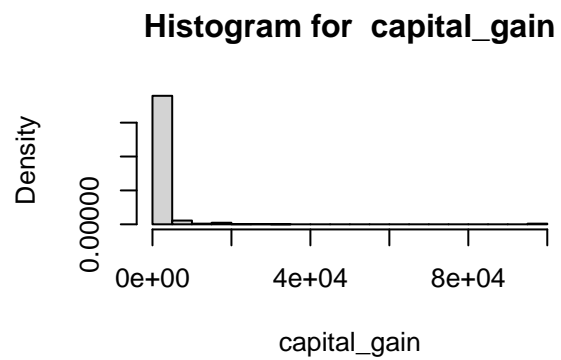
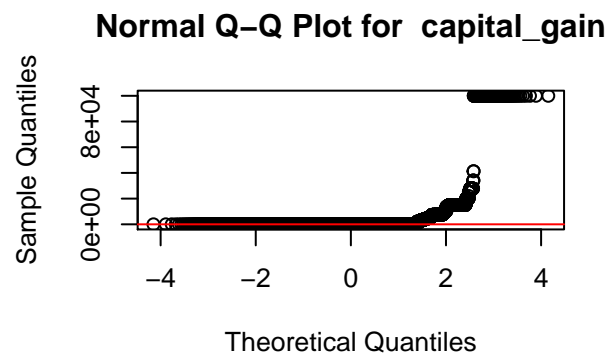
Procedemos a la generación de histogramas y de las gráficas quantile-quantile para entender la distribución de cada variable numérica para decidir si existe más variables a eliminar para reducir nuestro conjunto de datos.

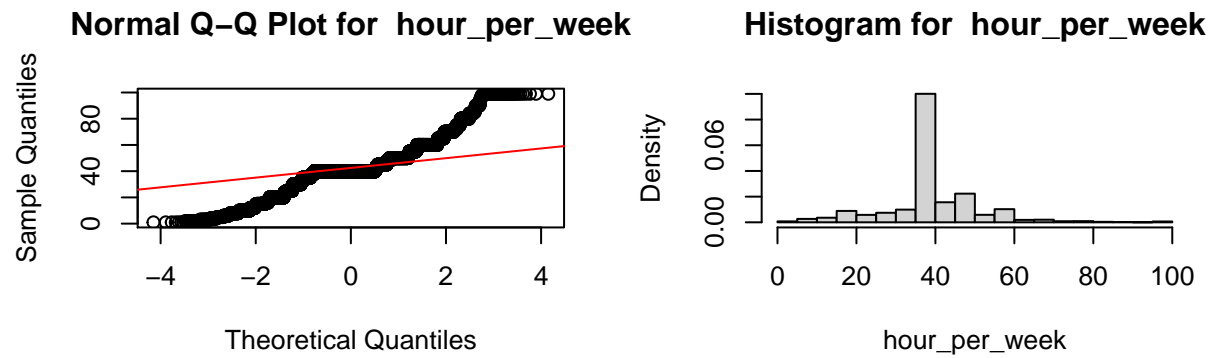
```

par(mfrow=c(2,2))
for(i in 1:ncol(datosAdult)) {
  if (is.numeric(datosAdult[,i])){
    qqnorm(datosAdult[,i],main = paste("Normal Q-Q Plot for ",colnames(datosAdult)[i]))
    qqline(datosAdult[,i],col="red")
    hist(datosAdult[,i],
         main=paste("Histogram for ", colnames(datosAdult)[i]),
         xlab=colnames(datosAdult)[i], freq = FALSE)
  }
}

```







Observamos que casi todas las distribuciones están sesgadas positivamente. Además, los histogramas muestran que las distribuciones de las variables “capital_gain” y “capital_loss” están muy confusas ya que la mayoría de sus observaciones son 0. Por lo tanto podemos eliminar estas variables.

```
# Reducción de la dimensionalidad
datosAdult$capital_gain <- NULL
datosAdult$capital_loss <- NULL
```

Procedemos a analizar las variables del conjunto de datos para sacar una conclusión.

Como se puede observar en el conjunto de datos, podemos agrupar los atributos de las variables “workclass”, “occupation” en subgrupos.

```
# Empezamos por la variable "workclass" que se dividirá en 4 grupos: Government (Federal-gov, Local-gov)
# summary(datosAdult$workclass)
table(datosAdult$workclass)
```

```
##
##      Federal-gov      Local-gov      Private      Self-emp-inc
##           943           2067          22286           1074
## Self-emp-not-inc      State-gov      Without-pay
##           2499           1279             14
```

```

datosAdult$workclass <- gsub("Federal-gov", "Government", datosAdult$workclass)
datosAdult$workclass <- gsub("Local-gov", "Government", datosAdult$workclass)
datosAdult$workclass <- gsub("State-gov", "Government", datosAdult$workclass)
datosAdult$workclass <- gsub("Self-emp-inc", "SelfEmployed", datosAdult$workclass)
datosAdult$workclass <- gsub("Self-emp-not-inc", "SelfEmployed", datosAdult$workclass)
datosAdult$workclass <- gsub("Never-worked", "Unknown", datosAdult$workclass)
datosAdult$workclass <- gsub("Without-pay", "Unknown", datosAdult$workclass)

table(datosAdult$workclass)

```

```

##
##   Government   Private SelfEmployed   Unknown
##         4289         22286         3573         14

```

Podemos combinar los trabajos y separarlos en varios grupos.

```

# summary(datosAdult$occupation)
table(datosAdult$occupation)

```

```

##
##   Adm-clerical   Armed-Forces   Craft-repair   Exec-managerial
##           3721           9           4030           3992
##   Farming-fishing Handlers-cleaners Machine-op-inspct   Other-service
##           989           1350           1966           3212
##   Priv-house-serv   Prof-specialty   Protective-serv   Sales
##           143           4038           644           3584
##   Tech-support   Transport-moving
##           912           1572

```

```

datosAdult$occupation <- gsub("Adm-clerical", "OfficeLabour", datosAdult$occupation)
datosAdult$occupation <- gsub("Exec-managerial", "OfficeLabour", datosAdult$occupation)
datosAdult$occupation <- gsub("Craft-repair", "ManualLabour", datosAdult$occupation)
datosAdult$occupation <- gsub("Farming-fishing", "ManualLabour", datosAdult$occupation)
datosAdult$occupation <- gsub("Handlers-cleaners", "ManualLabour", datosAdult$occupation)
datosAdult$occupation <- gsub("Machine-op-inspct", "ManualLabour", datosAdult$occupation)
datosAdult$occupation <- gsub("Transport-moving", "ManualLabour", datosAdult$occupation)
datosAdult$occupation <- gsub("Other-service", "Service", datosAdult$occupation)
datosAdult$occupation <- gsub("Priv-house-serv", "Service", datosAdult$occupation)
datosAdult$occupation <- gsub("Protective-serv", "Service", datosAdult$occupation)
datosAdult$occupation <- gsub("Tech-support", "Service", datosAdult$occupation)
datosAdult$occupation <- gsub("Prof-specialty", "Professional", datosAdult$occupation)

```

Se podria agrupar Armed-Forces con professional pero ya que son pocos datos lo agrupo con Unknown.

```

datosAdult$occupation <- gsub("Armed-Forces", "Unknown", datosAdult$occupation)
table(datosAdult$occupation)

```

```

##
## ManualLabour OfficeLabour Professional   Sales   Service   Unknown
##         9907         7713         4038         3584         4911         9

```



```
# summary(datosAdult$marital_status)
table(datosAdult$occupation)
```

```
##
## ManualLabour OfficeLabour Professional Sales Service Unknown
## 9907 7713 4038 3584 4911 9
```

```
datosAdult$marital_status <- gsub("Married-AF-spouse", "Married", datosAdult$marital_status)
datosAdult$marital_status <- gsub("Married-civ-spouse", "Married", datosAdult$marital_status)
datosAdult$marital_status <- gsub("Married-spouse-absent", "Married", datosAdult$marital_status)
datosAdult$marital_status <- gsub("Never-married", "Single", datosAdult$marital_status)
```

```
table(datosAdult$occupation)
```

```
##
## ManualLabour OfficeLabour Professional Sales Service Unknown
## 9907 7713 4038 3584 4911 9
```

Archivo final

Una vez realizado el preprocesamiento sobre los datos, guardaremos el nuevo juego de datos.

```
write.csv(datosAdult, "../data/adult_clean.csv", row.names = FALSE)
```