PRAC2: Neteja i anàlisi de les dades

Enric Pou

2/1/2020

1. Descripció del dataset

El conjunt de dades obtingut es troba disponible a la web de Kaggle en aquest enllaç. El conjunt de dades total està format per 15 variables diferents de 48842 observacions de persones repartatis en 2 arxius CSV.

Les variables d'aquest conjunt són:

- **age**: Edat de la persona. *(enter més gran que 0)*
- **workclass**: Classe del treball que exerceix. *(Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked)*
- **fnlwgt**: Pes final. És el nombre de persones el qual el cens creu que aquesta entrada representa. *(enter més gran que 0)*
- **eduaction**: Nivell d'educació *(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**: Nivell d'educació. *(enter més gran que 0)*
- **marital-status**: Estat civil. *(Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.). civ = civil ; AF = armed forces*
- **occupation**: Sector d'ocupació. *(Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.)*
- **relationship**: Estat sentimental. *(Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.)*
- **race**: Ètnia. *(White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black)*
- **sex**: Sexe. *(Male, Female)*
- **capital-gain**: Guanys provinents de fonts d'inversió diferents del salari. *(enter més gran que 0)*
- **capital-loss**: Pèrdues provinents de fonts d'inversió diferents del salari. *(enter més gran que 0)*
- **hour-per-week**: Hores treballades a la setmana. *(enter)*
- **native-country**: País natal.
- **income**: Ingressos *(<=50k, >50k)*

Aquestes dades formen part del cens d'Estats Units de l'any 1994.

Aquest dataset incorpora tot un recull de variables sociologiques amb les quals es pot predir el nivell d'ingressos (inferior o superior a 50000$). És important remarcar que aquest estudi es pot utiltizar per a poder determinar si existeix un biaix social que afecti a un conjunt o minoria de persones envers el seu nivell d'ingressos. El fet de tenir dades sensibles, les quals des de fa temps s'estan fent estudis per mirar que no hi hagi diferències salarials en igualtats de condicions, tals com el sexe, la raça, etc; es un punt de motivació per a fer un primer estudi per desmentir o reafirmar aquestes suposicions.

Al llarg d'aquest document, es preten donar un enfoc d'aquest caire i intentar respondre aquest tipus de preguntes.

2. Integració i selecció de les dades d'interès.

Primer de tot, veiem que les dades estan repartides en 2 arxius CSV diferents:

- *adult-test.csv*
- *adult-training.csv*

El primer que farem serà carregar ambdos arxius i juntar-los, per obtenir un dataframe amb la totalitat de les dades.

```
# Carreguem l'arxiu de training i de test.
datatrain <- read.csv("adult-training.csv", header = FALSE, stringsAsFactors = FALSE)
datatest <- read.csv("adult-test.csv", skip = 1, header=FALSE, stringsAsFactors =
FALSE)

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

# Assignem els noms a les columnes pertinents.
names(datatrain) <- data_colnames
names(datatest) <- data_colnames

# Ajuntem els dos dataframes.
df <- rbind(datatrain, datatest)
```

Comprovem l'estructura de les dades

```
str(df)
```

```
## 'data.frame':    48842 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" ...
```

Veiem la presència de espais en blanc en totes les cadenes de caràcters. Primer de tot doncs, les eliminarem:

```
# Obtenim quines columnes són de tipus character
truth <- sapply(df,is.character)

# Per a les columnes de tipus charcaters, els hi apliquem la funcio trimws
# i el resultat l'ajuntem (cbind) amb les columnes que no són de tipus character
df <- data.frame(
  cbind(
    sapply(
      df[,truth],
      trimws,
      which="both"
    ),
    df[,!truth]
  )
)

str(df)
```

```
## 'data.frame':    48842 obs. of  15 variables:
##  $ workclass     : Factor w/ 9 levels "?","Federal-gov",..: 8 7 5 5 5 5 5 7 5
5 ...
##  $ education     : Factor w/ 16 levels "10th","11th",..: 10 10 12 2 10 13 7 12 13
10 ...
##  $ marital.status: Factor w/ 7 levels "Divorced","Married-AF-spouse",..: 5 3 1 3 3
3 4 3 5 3 ...
##  $ occupation    : Factor w/ 15 levels "?","Adm-clerical",..: 2 5 7 7 11 5 9 5 11
5 ...
##  $ 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 ...
##  $ native.country: Factor w/ 42 levels "?","Cambodia",..: 40 40 40 40 6 40 24 40
40 40 ...
##  $ income        : Factor w/ 4 levels "<=50K","<=50K.",..: 1 1 1 1 1 1 1 3 3 3 ...
##  $ age           : int  39 50 38 53 28 37 49 52 31 42 ...
##  $ fnlwgt        : int  77516 83311 215646 234721 338409 284582 160187 209642
45781 159449 ...
##  $ education.num : int  13 13 9 7 13 14 5 9 14 13 ...
##  $ 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 ...
```

Veiem ara que totes les columnes són de la classe correcta. Les categòriques són de tipus `Factor`; mentre que les enteres són de tipus `int`.

Anem ara a fer un resum estadístic de les dades:

```
summary(df)
```

```
##             workclass            education
##  Private         :33906   HS-grad     :15784
##  Self-emp-not-inc: 3862   Some-college:10878
##  Local-gov       : 3136   Bachelors   : 8025
##  ?               : 2799   Masters     : 2657
##  State-gov       : 1981   Assoc-voc   : 2061
##  Self-emp-inc    : 1695   11th        : 1812
##  (Other)         : 1463   (Other)     : 7625
##                marital.status           occupation
##  Divorced             : 6633   Prof-specialty : 6172
##  Married-AF-spouse    :   37   Craft-repair   : 6112
##  Married-civ-spouse   :22379   Exec-managerial: 6086
##  Married-spouse-absent:  628   Adm-clerical   : 5611
##  Never-married        :16117   Sales          : 5504
##  Separated            : 1530   Other-service  : 4923
##  Widowed              : 1518   (Other)        :14434
##          relationship                    race           sex
##  Husband       :19716   Amer-Indian-Eskimo:  470   Female:16192
##  Not-in-family :12583   Asian-Pac-Islander: 1519   Male  :32650
##  Other-relative: 1506   Black             : 4685
##  Own-child     : 7581   Other             :  406
##  Unmarried     : 5125   White             :41762
##  Wife          : 2331
##
##        native.country     income           age            fnlwgt
##  United-States:43832   <=50K :24720   Min.   :17.00   Min.   :  12285
##  Mexico       :  951   <=50K.:12435   1st Qu.:28.00   1st Qu.: 117550
##  ?            :  857   >50K  : 7841   Median :37.00   Median : 178144
##  Philippines  :  295   >50K. : 3846   Mean   :38.64   Mean   : 189664
##  Germany      :  206                  3rd Qu.:48.00   3rd Qu.: 237642
##  Puerto-Rico  :  184                  Max.   :90.00   Max.   :1490400
##  (Other)      : 2517
##  education.num    capital.gain    capital.loss    hour.per.week
##  Min.   : 1.00   Min.   :    0   Min.   :   0.0   Min.   : 1.00
##  1st Qu.: 9.00   1st Qu.:    0   1st Qu.:   0.0   1st Qu.:40.00
##  Median :10.00   Median :    0   Median :   0.0   Median :40.00
##  Mean   :10.08   Mean   : 1079   Mean   :  87.5   Mean   :40.42
##  3rd Qu.:12.00   3rd Qu.:    0   3rd Qu.:   0.0   3rd Qu.:45.00
##  Max.   :16.00   Max.   :99999   Max.   :4356.0   Max.   :99.00
##
```

En el resultat anterior podem comprovar que:

- No existeixen inconsistències entre el que en diu la descripció del dataset envers la variable edat (age > 16 && age <= 100).
- Com es distribueixen les variables qualitatives.
- Sospitem la presència de outliers per a la variable `hours-per-week` (99 hores equival a treballar de mijta 19,8h diaries)
- Hi ha més presència de homes que dones al dataset.
- La majoria de registres són de persones natives de EEUU.
- Hi ha més registres amb un `income` superior a 50k.

Respecte a l'objectiu del nostre análisi podem eliminar de la variable `fnlwgt` ja que no aporta "poder" predictiu i, veiem que es pot prescindir de la columna `education` ja que existeix la columna `education.num` que n'és una copia però conté el nombre d'anys estudiats (int) en comptes de paraules (chr).

```
df$fnlwgt <- NULL
df$education <- NULL
```

3. Neteja de les dades

Abans de res, ens hem d'assegurar que els nivell de les columnes que són de tipus `factor` són correctes.

```
factors <- sapply(df, is.factor)
lapply(df[, factors], levels)
```

```
## $workclass
## [1] "?"                "Federal-gov"      "Local-gov"
## [4] "Never-worked"     "Private"          "Self-emp-inc"
## [7] "Self-emp-not-inc" "State-gov"        "Without-pay"
##
## $marital.status
## [1] "Divorced"              "Married-AF-spouse"     "Married-civ-spouse"
## [4] "Married-spouse-absent" "Never-married"         "Separated"
## [7] "Widowed"
##
## $occupation
##  [1] "?"                 "Adm-clerical"      "Armed-Forces"
##  [4] "Craft-repair"      "Exec-managerial"   "Farming-fishing"
##  [7] "Handlers-cleaners" "Machine-op-inspct" "Other-service"
## [10] "Priv-house-serv"   "Prof-specialty"    "Protective-serv"
## [13] "Sales"             "Tech-support"      "Transport-moving"
##
## $relationship
## [1] "Husband"        "Not-in-family"  "Other-relative" "Own-child"
## [5] "Unmarried"      "Wife"
##
## $race
## [1] "Amer-Indian-Eskimo" "Asian-Pac-Islander" "Black"
## [4] "Other"              "White"
##
## $sex
## [1] "Female" "Male"
##
## $native.country
##  [1] "?"                          "Cambodia"
##  [3] "Canada"                     "China"
##  [5] "Columbia"                   "Cuba"
##  [7] "Dominican-Republic"         "Ecuador"
##  [9] "El-Salvador"                "England"
## [11] "France"                     "Germany"
## [13] "Greece"                     "Guatemala"
## [15] "Haiti"                      "Holand-Netherlands"
## [17] "Honduras"                   "Hong"
## [19] "Hungary"                    "India"
## [21] "Iran"                       "Ireland"
## [23] "Italy"                      "Jamaica"
## [25] "Japan"                      "Laos"
## [27] "Mexico"                     "Nicaragua"
## [29] "Outlying-US(Guam-USVI-etc)" "Peru"
## [31] "Philippines"                "Poland"
## [33] "Portugal"                   "Puerto-Rico"
## [35] "Scotland"                   "South"
## [37] "Taiwan"                     "Thailand"
## [39] "Trinadad&Tobago"            "United-States"
## [41] "Vietnam"                    "Yugoslavia"
##
## $income
## [1] "<=50K"  "<=50K." ">50K"   ">50K."
```

Veiem que a la columna `income` trobem els mateixos valors acabats amb o sense punt final, generant així 4 nivell diferents, quan n'hi hauria d'haver només 2. Ho corregim:

```
levels(df$income)[levels(df$income)=="<=50K."] <- "<=50K"
```

```
levels(df$income)[levels(df$income)==">50K."] <- ">50K"
levels(df$income)
```

```
## [1] "<=50K" ">50K"
```

Finalment, anem a ajuntar les columnes `capital.gain` i `capital.loss` en una de sola, a partir de la resta de les pèrdues envers els guanys.

```
df["capital"] = df$capital.gain - df$capital.loss
df$capital.gain <- NULL
df$capital.loss <- NULL
```

3.1 Elements buits

Anem a comprovar si el nostre conjunt de dades conté valors buits.

```
print("Casos amb NA")
```

```
## [1] "Casos amb NA"
```

```
colSums(is.na(df))
```

```
##      workclass marital.status     occupation   relationship           race
##              0              0              0              0              0
##            sex native.country         income            age  education.num
##              0              0              0              0              0
##  hour.per.week        capital
##              0              0
```

```
print("Casos amb cadenes de text buides")
```

```
## [1] "Casos amb cadenes de text buides"
```

```
colSums(df == "")
```

```
##      workclass marital.status     occupation   relationship           race
##              0              0              0              0              0
##            sex native.country         income            age  education.num
##              0              0              0              0              0
##  hour.per.week        capital
##              0              0
```

S’observa doncs que no hi ha presència de valors buits. Ara bé, si ens fixem amb el joc de dades s’observa la presència del caràcter ? en algun de les cel·les. Anem a veure la seva estadística:

```
colSums(df == "?")
```

```
##      workclass marital.status     occupation   relationship           race
##           2799              0           2809              0              0
##            sex native.country         income            age  education.num
##              0            857              0              0              0
##  hour.per.week        capital
##              0              0
```

Per a fer més fàcil les següents tasques, anem a substituïr el valor ‘?’ per a NA (*Not available*), que R l’interpreta com un valor buit.

```
df[df == "?"] <- NA
```

Arribats a aquest punt, hem de decidir què fer amb aquest valors erronis. Una opció sería la de eliminar

aquests registres però això suposaría desaprofitar informació.

És per això doncs, que imputarem aquest valors. Trobem diferents tècniques per a imputar aquests valors: substituïr pel valor més freqüent, per la mitja, etc. El principal tipus de problema amb aquest tipus d'imputació és que es substitueixen tots els valors buits de cada columna per un mateix valor, sense tenir en compte com afecten les altres variables. És per això, que s'ha decidit utilitzar el mètode d'imputació basat en el *K* veïns més pròxims (kNN). D'aquesta manera l'imputació tindrà en compte la relació de la columna a imputar amb les altres, i per a cada cas se li assignarà un valor.

```
df <- kNN(data=df, variable = c("workclass", "occupation", "native.country"),
impNA=TRUE, imp_var=FALSE)
```

Un cop imputats els valors, tornem a revisar si hi ha valors NA (recordem que s'ha substituit els ? per NA.

```
colSums(is.na(df))
```

```
##      workclass marital.status     occupation   relationship           race
##              0              0              0              0              0
##            sex native.country         income            age  education.num
##              0              0              0              0              0
##  hour.per.week        capital
##              0              0
```

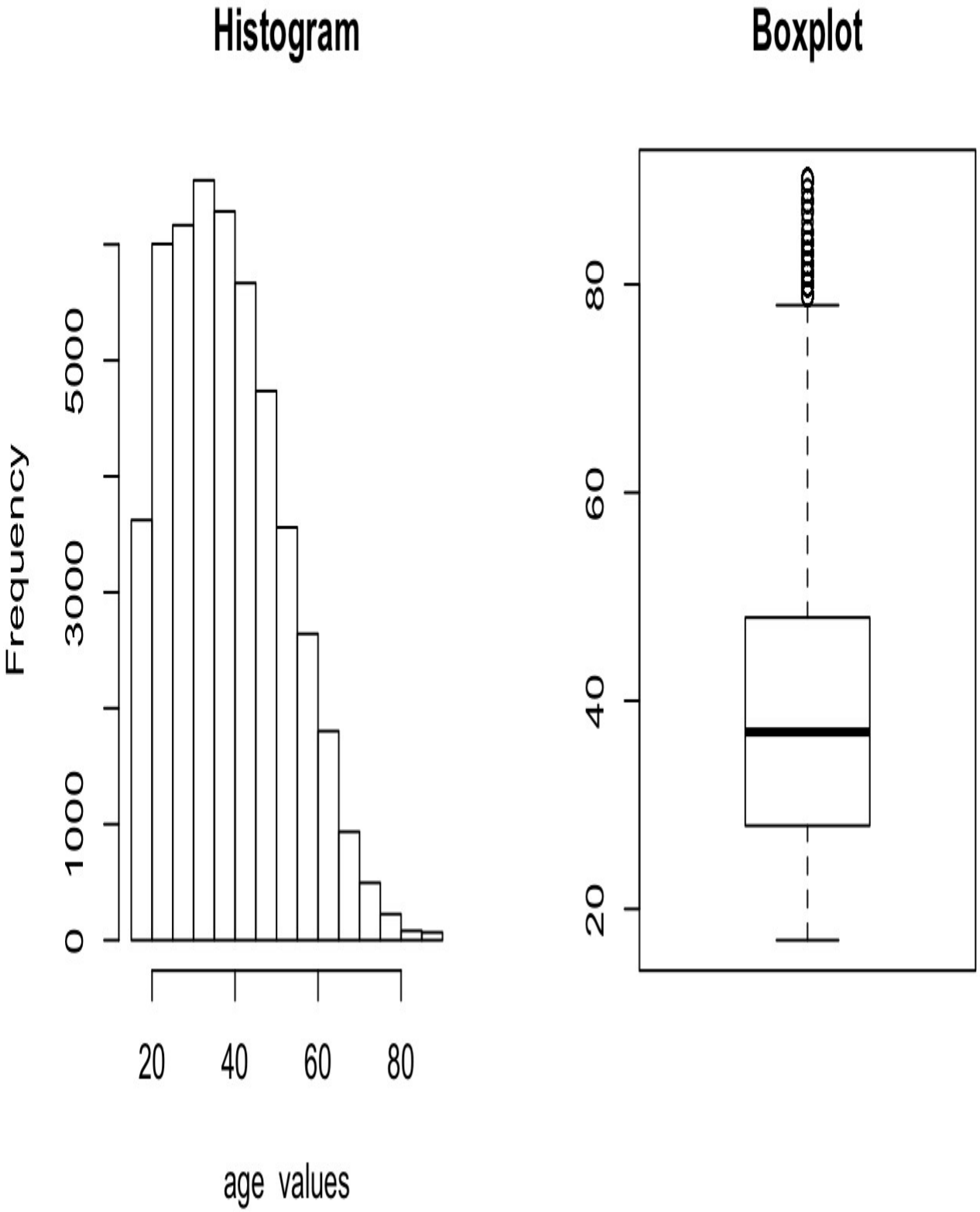
Tal com s'observa, ja no hi ha valors buits.

3.2 Valors extrems

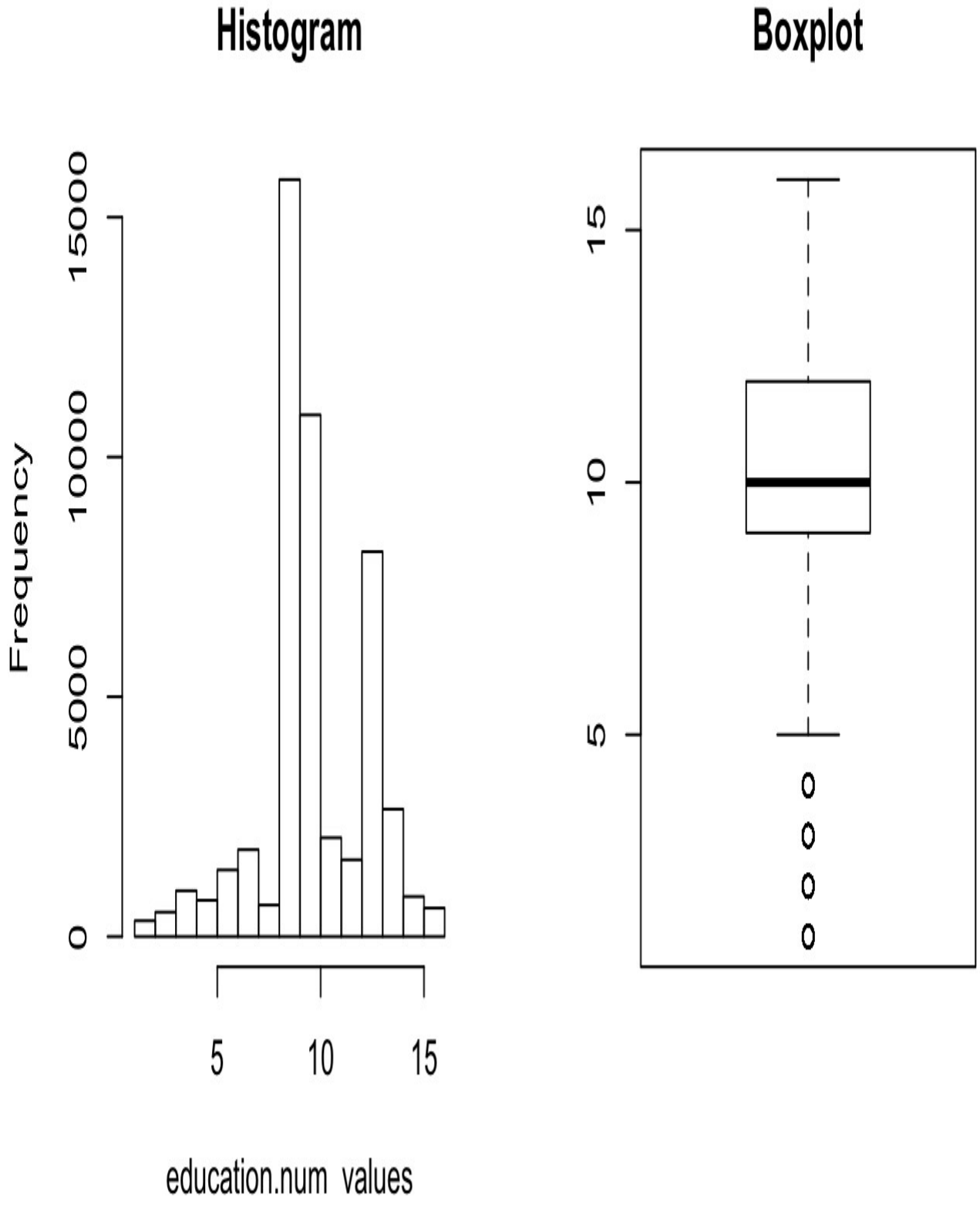
En aquest apartat ens centrarem en detectar i corregir (si escau) els possibles outliers presents en cadascuna de les columnes numèriques. El primer que farem serà buscar valors sentineles de manera visual. Aquest análisi només el durem a terme per a les variables quantitatives.

```
attributs_quantitatius = c('age', 'education.num', 'capital', 'hour.per.week')
for (columnName in attributs_quantitatius){
  # Generem una graella amb 1 fila i 2 columnes
  par(
    mfrow=c(1,2),
    oma = c(0, 0, 2, 0)
  )
  # Mostrem el histograma
  hist(
    df[,columnName],
    main="Histogram",
    xlab=paste(columnName, " values")
  )
  # Mostrem el boxplot
  boxplot(
    df[,columnName],
    main="Boxplot"
  )
  # Posem titol al conjunt de grafiques.
  mtext(
    paste("Analysis for ",columnName),
    outer = TRUE,
    cex = 1.5
  )
}
```

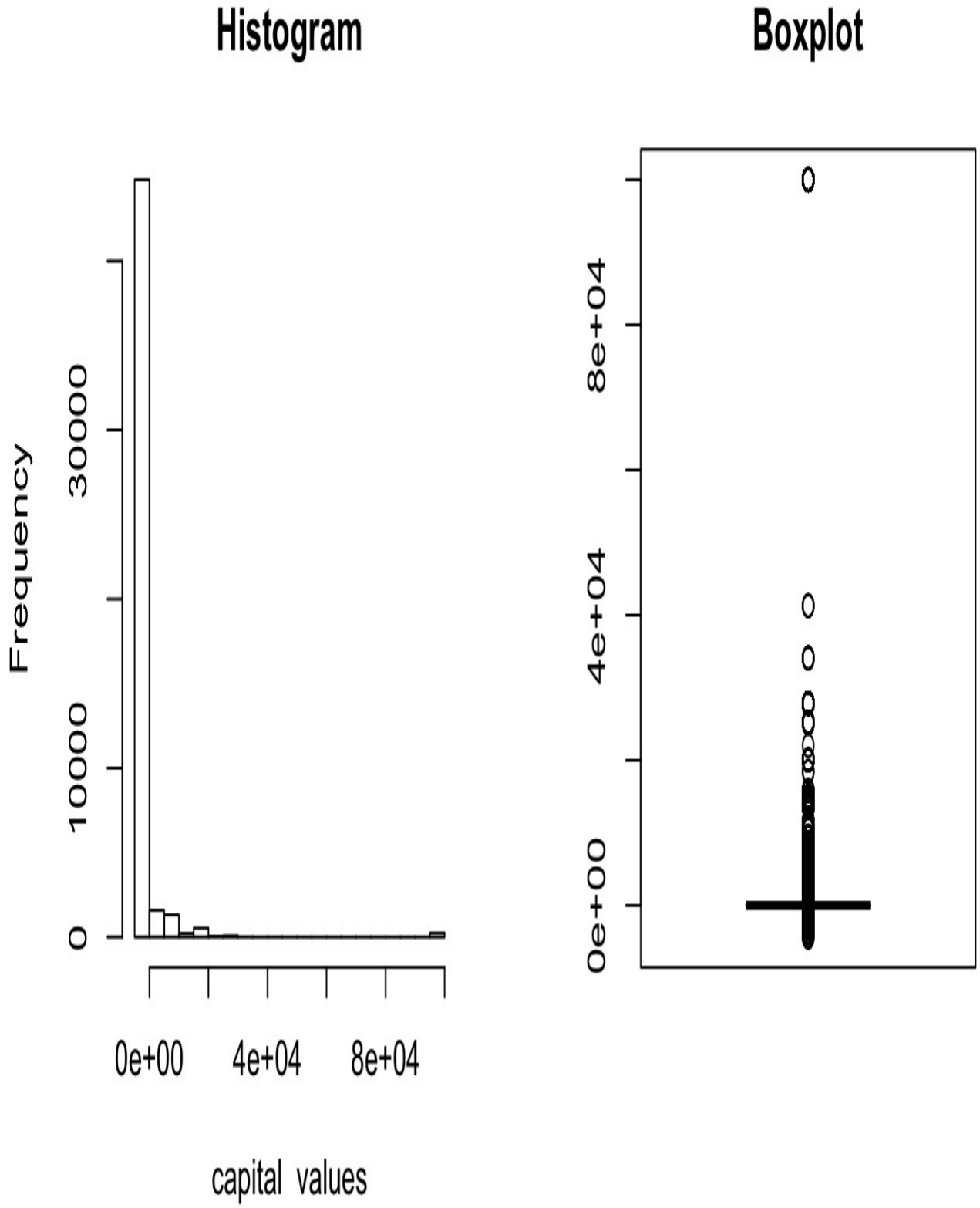
Analysis for age

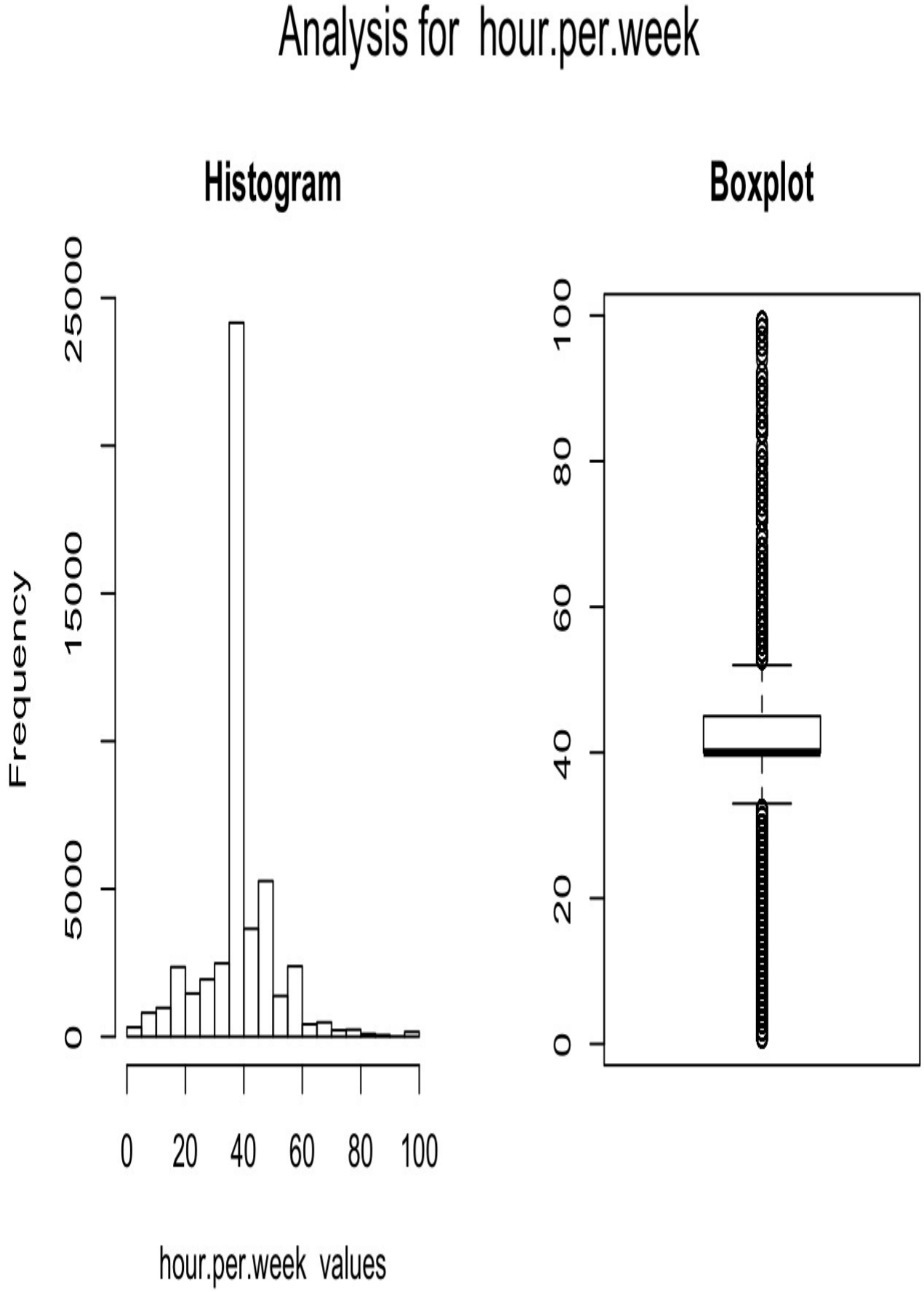


Analysis for education.num



Analysis for capital





Arribats a aquest punt, s'ha d'estudiar cada cas en detall per a decidir què fer amb els possibles outliers. En el cas de corregir els valors, ho farem per la seva mitjana i només ho aplicarem per aquells casos on el Z-score del valor sigui de més de 3 (positiu o negatiu).

Edat:

```
df$age[which(abs(scale(df$age))>3)] <- mean(df$age, na.rm=TRUE)
```

En el cas de `education.num` no podem considerar com a outliers cap valor, ja que aquesta variable és una simple representació numèrica de la columna, previament eliminada, `education`. Per tant, no fa referència a dades que s'han pogut entrar de forma errónia.

El `capital` no l'ajustarem ja que la gran majoria dels valors són 0, per tant tot els valors diferents de 0 el sistema els considera com a outliers.

Finalment, si donem un cop d'ull a la variable d'ocupació, veiem que hi ha persones que han posat xifres molt altes en les hores de la seva joranda laboral (hi ha casos que superen les 90 hores setmanals). Si ens fixem en aquests registres veiem que es dediquen principalment a la ramaderia-pesca, transportistes, i altres feines on es viatja sovint o es treballa per temporades i que implica estar fora de casa durant un jornada llarga. Per altre costat, pel valors extremadament petits trobem gent gran (possible jubilats) o estudiants. Remplaçarem aquest valors per la seva mitja.

```
df$hour.per.week[which(abs(scale(df$hour.per.week))>3)] <- mean(df$hour.per.week,
na.rm=TRUE)
```

4. Anàlisi de les dades

4.1 Selecció dels grups de dades que es volen analitzar/comparar

Degut a que el nostre objectiu era detectar biaixos de caràcter social, agruparem les diferents característiques segons cada variable.

Començarem per a la variable `workclass`.

```
table(df$workclass)
```

```
##
##                 ?      Federal-gov        Local-gov     Never-worked
##                 0             1442             3192               10
##           Private     Self-emp-inc Self-emp-not-inc        State-gov
##             36404             1730             4028             2015
##      Without-pay
##                21
```

Podríem agrupar-los segons si treballen pel govern (acaba en "-gov"), si treballen per ells mateixos (comença en "Self-emp") i si no tenen feina ("Without-pay" i "Never- worked")

```
# Convertim a string per a poder treballar
df$workclass <- as.character(df$workclass)
# Agrupem els valors que comencen per " Self-emp" i li posem "Self"
df$workclass[startsWith(df$workclass, "Self-emp")] = "Self"
# Agrupem els valors que acaben per "-gov" i li posem "Govern"
df$workclass[endsWith(df$workclass, "-gov")] = "Govern"
# Agrupem els valors " Never-worked" i " Without-pay" i li posem "No-work"
df$workclass[df$workclass == "Never-worked"] = "No-work"
df$workclass[df$workclass == "Without-pay"] = "No-work"
# Tornem a transformar a factor
df$workclass <- as.factor(df$workclass)
```

Si tornem a revisar com han quedat distribuïts els valors:

```
table(df$workclass)
```

```
##
##  Govern No-work Private    Self
##    6649      31   36404    5758
```

Passem ara al variable que indica l'estat civil: `marital.status`.

```
table(df$marital.status)
```

```
##
##              Divorced     Married-AF-spouse    Married-civ-spouse
##                  6633                    37                 22379
## Married-spouse-absent         Never-married              Separated
##                   628                 16117                  1530
##               Widowed
##                  1518
```

En aquest cas podem agrupar els valors que comencen per "Married":

```
df$marital.status <- as.character(df$marital.status)
# Agrupem els valors que comencen per " Married-" i li posem "Married"
df$marital.status[startsWith(df$marital.status, "Married-")] = "Married"
df$marital.status <- as.factor(df$marital.status)
```

Si tornem a revisar com han quedat distribuïts els valors:

```
table(df$marital.status)
```

```
##
##      Divorced       Married Never-married     Separated       Widowed
##          6633         23044         16117          1530          1518
```

Ens interesa també agrupar el registres segons la seva nacionalitat, separant per els natius de EEUU i els que no.

```
df$native.country <- as.character(df$native.country)
# Agrupem els valors que no siguin de EEUU
df$native.country[df$native.country != "United-States"] = "Other"
df$native.country <- as.factor(df$native.country)
```

Finalment, podem agrupar les edats segons les etapes de lavida. Veiem que el valor mínim de l'edat són 17. Considerarem les següents etapes:

- Adolescent: 12 < Edat <= 19
- Adult-primerenc: 20 < Edat <= 25
- Adult: 26 < Edat <= 49
- Vellesa: Edat > 50

Per tant, crearem una nova variable anomenada "StageOfLife" on li aplicarem els llindars anteriorment esmentats. Per a fer-ho ens definim primer una funció per calcular l'etapa de la vida segons l'edat i després generem una nova columna a partir de crides a aquesta funció.

```
# Funcio per calcular el stage of life
stageOfLife <- function(age) {
if(12<age&age<=19) { result <- "Adolescent"
  } else if (20 < age & age <= 25) {
    result <- "Adult-primerenc"
  } else if (26 < age & age <= 49) {
```

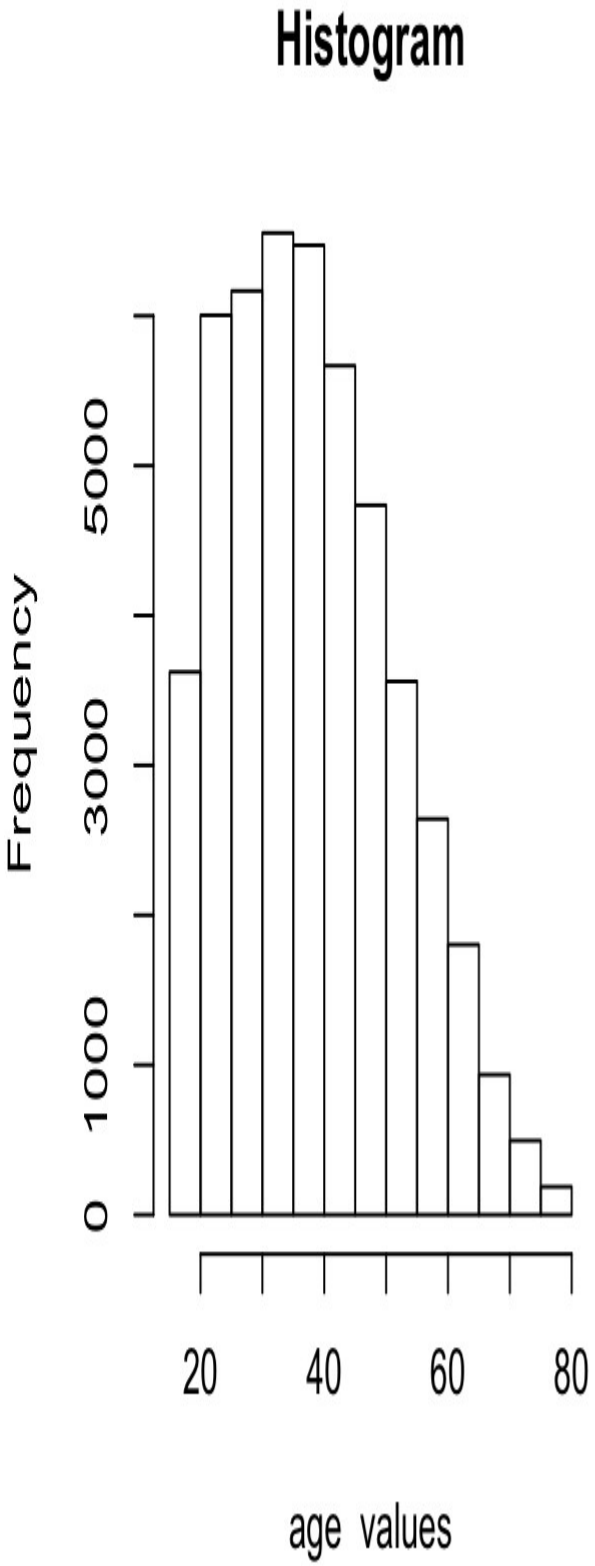
```
    result <- "Adult"
} else {
    result <- "Vellesa"
  }
  return(result)
}

# Afegim una nova variable a partir d'agrupar el camp age.
df$StageOfLife <- sapply(df$age, stageOfLife)
df$StageOfLife <- as.factor(df$StageOfLife)
```

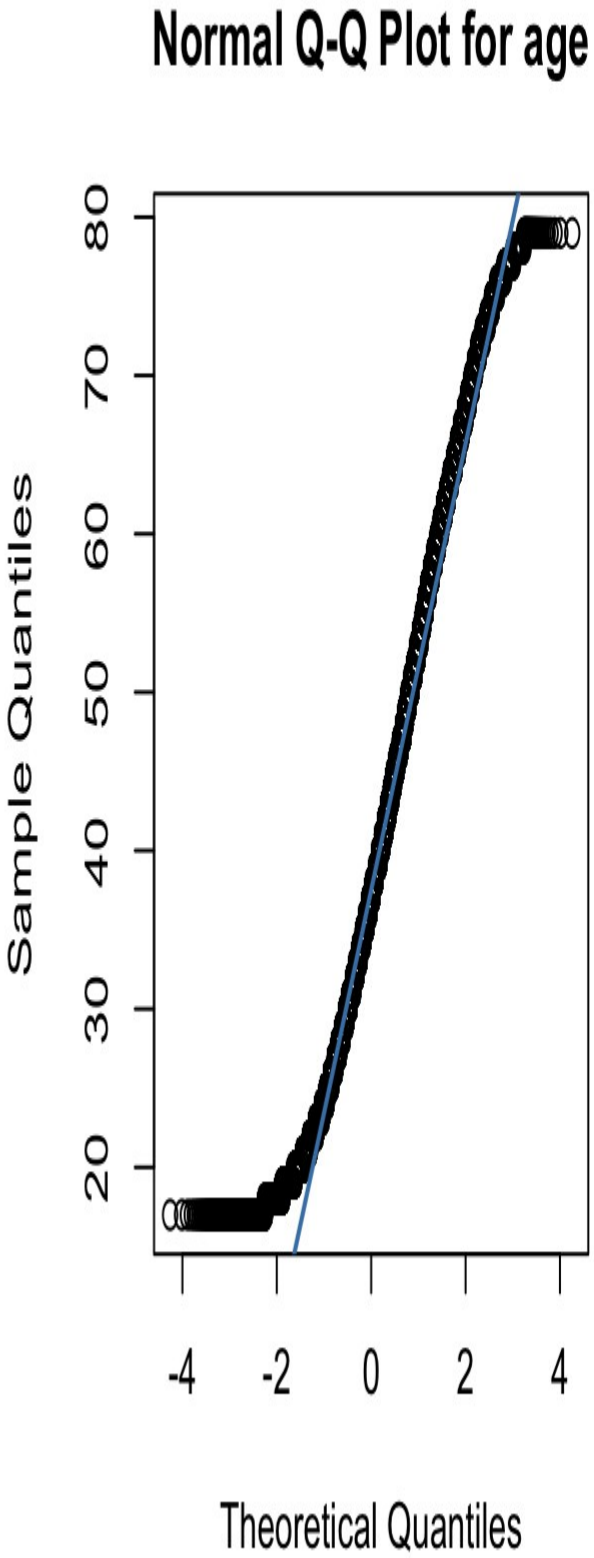
4.2 Comprovació de la normalitat i homogeneïtat de la variància.

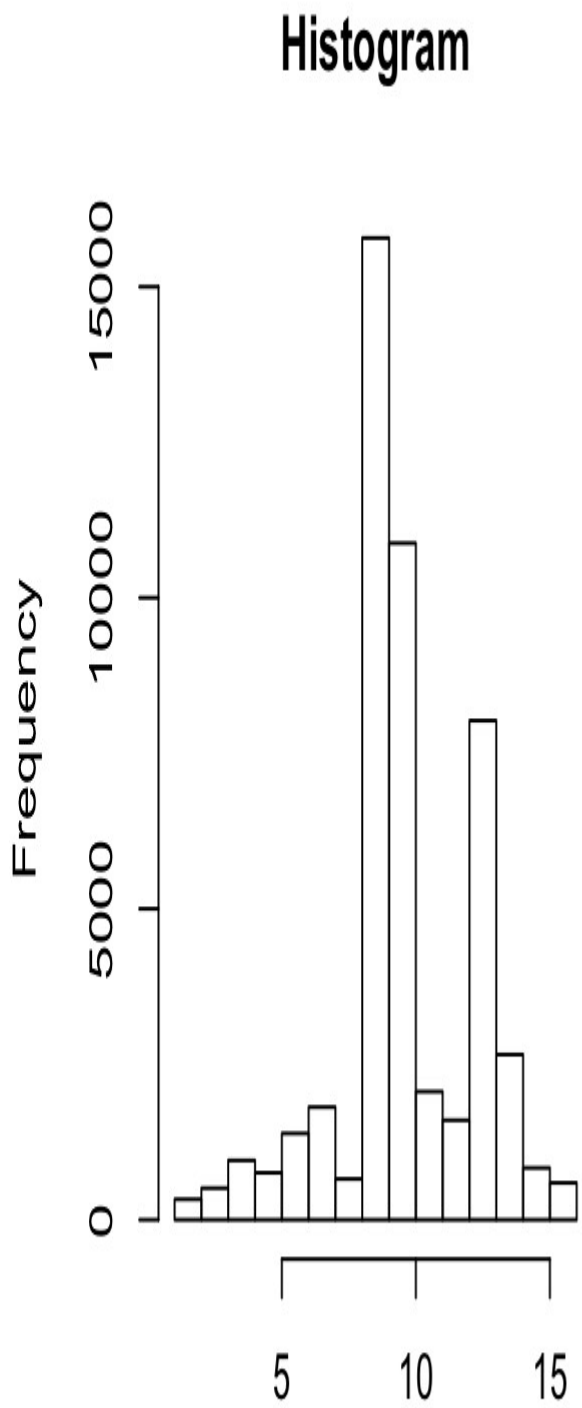
Un cop agrupats els valors anteriors, tenim només tres variables que són numèriques: `age`, `eduaction.num`, `hour.per.week` i `capital`.

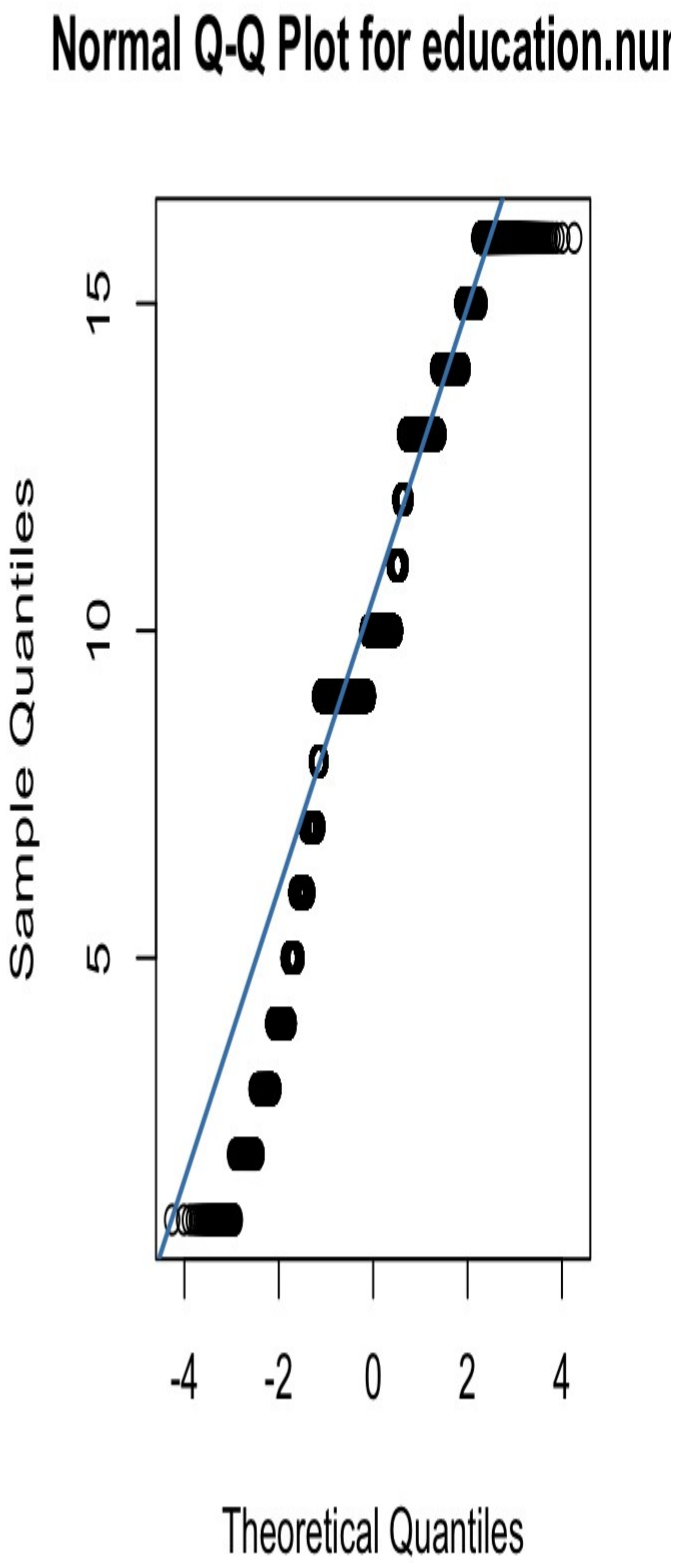
Farem ús de les gràfiques Q-Q y el histograma.

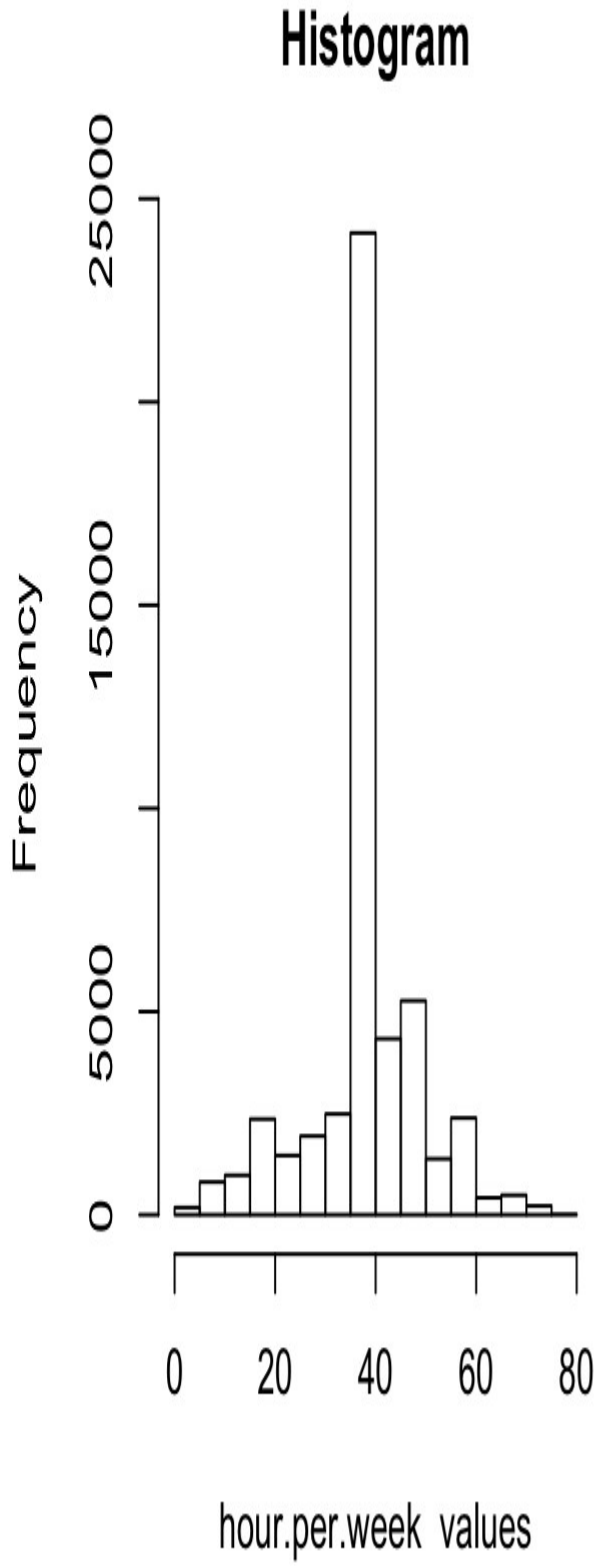
```
columnToAnalyze <- c("age", "education.num", "hour.per.week", "capital")

for (columnName in columnToAnalyze) {
  # Generem una graella amb 1 fila i 2 columnes
  par(
    mfrow=c(1,2),
    oma = c(0, 0, 2, 0)
  )
  # Mostrem el histograma
  hist(
    df[,columnName],
    main="Histogram",
    xlab=paste(columnName, " values")
  )

  # Mostrem el qqplot
  qqnorm(df[,columnName], main = paste("Normal Q-Q Plot for", columnName))
  qqline(df[,columnName], col = "steelblue", lwd = 2)

}
```

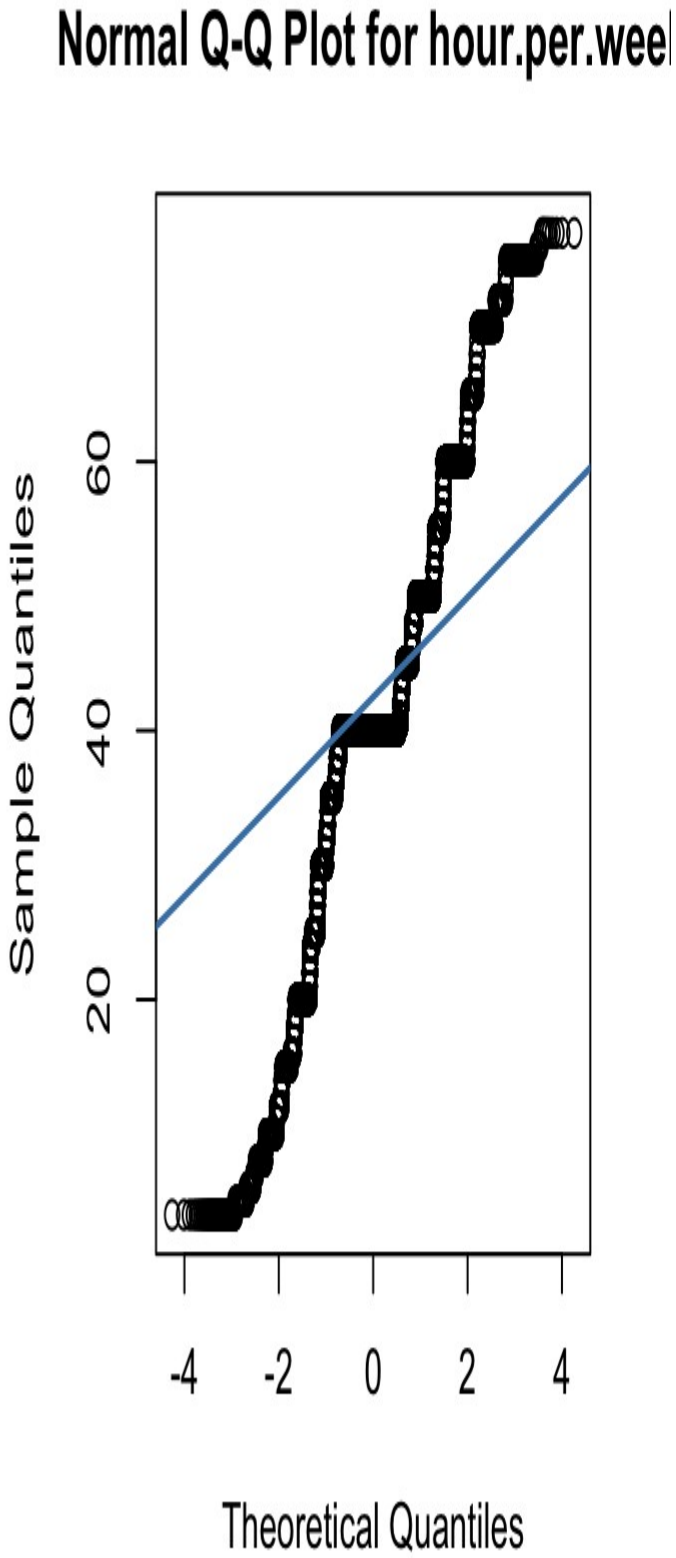

Histogram
Frequency
0
1000
3000
5000
20
40
60
80
age values

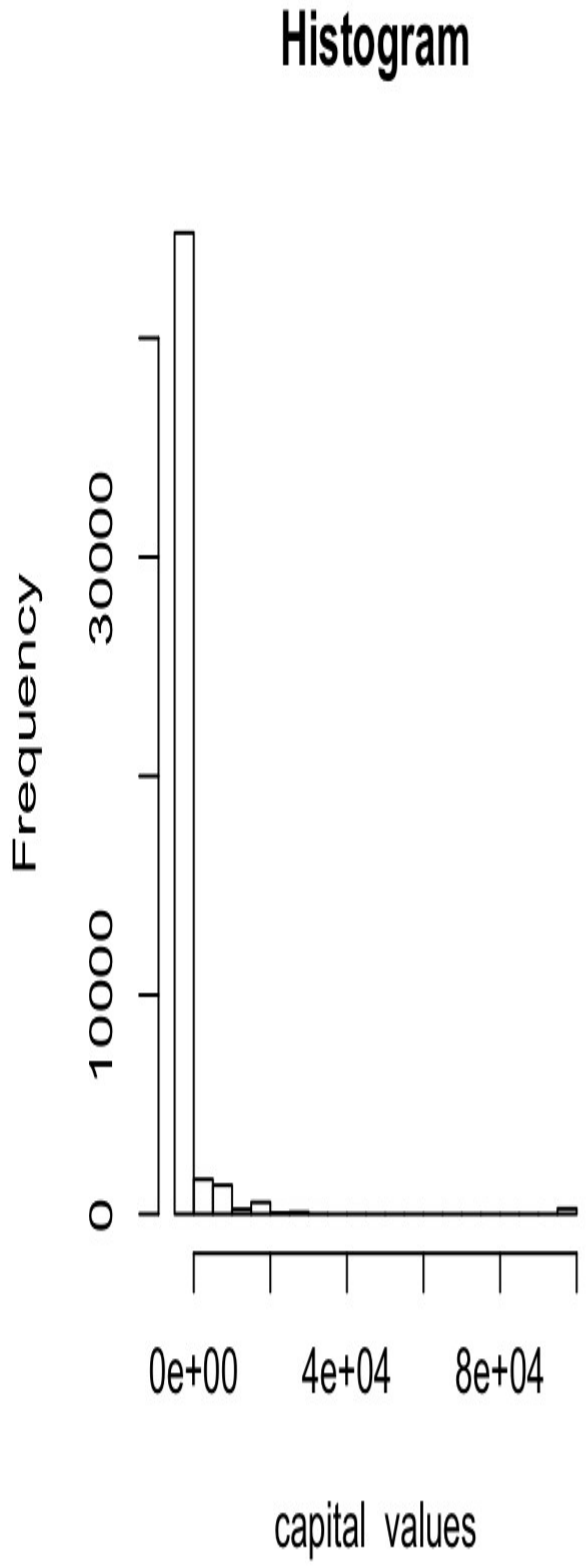

Normal Q-Q Plot for age
Sample Quantiles
20
30
40
50
60
70
80
-4
-2
0
2
4
Theoretical Quantiles

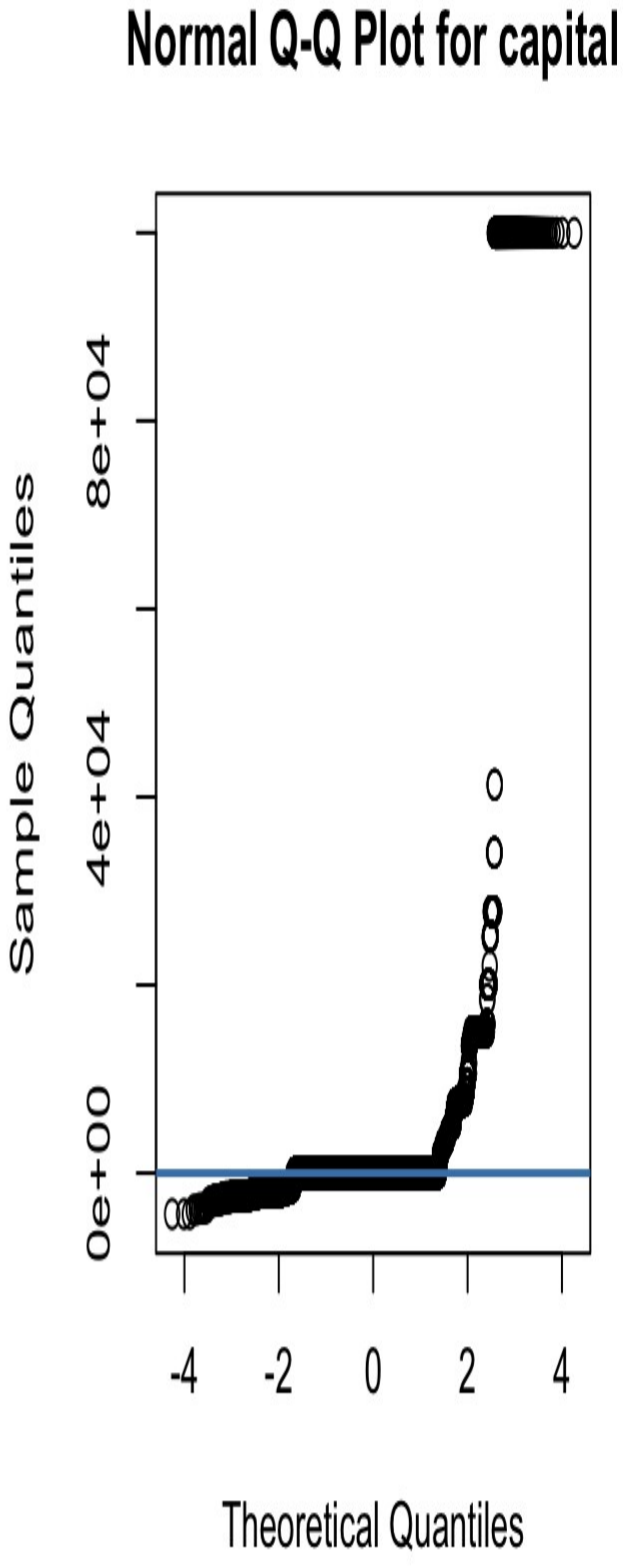

Histogram
Frequency
0
5000
10000
15000
5
10
15
education.num values


Normal Q-Q Plot for education.nur
Sample Quantiles
5
10
15
-4
-2
0
2
4
Theoretical Quantiles


Histogram
Frequency
0
5000
15000
25000
0
20
40
60
80
hour.per.week values


Normal Q-Q Plot for hour.per.wee
Sample Quantiles
20
40
60
-4
-2
0
2
4
Theoretical Quantiles





S’observa dels resultats anteriors, que les úniques variables (considerant només l’inspecció visual) que sembla que podriem considerar com a normal són l’edat i l’educació, ja que són les dues variables la gràfiques Q-Q de les quals encaixa millor amb la linea de quantils.

Per a ser més concisos aplicarem un test de normalitat. Degut a que el test de Shapiro-Wilk accepta només com a màxim 5000 valors d'entrada i nosaltres en tenim més, usarem el test de normalitat d'Anderson-Darling. Aquest test té com a hipòtesi nul·la que la mostra és una distribució normal. Per tant, segons el resultat del p-valor l'acceptarem (cas p−valor>α) o el rebutjarem (cas p−valor<α).

```
library(nortest)
for (columnName in columnToAnalyze) {
  print(ad.test(df[,columnName]))
}
```

```
##
##  Anderson-Darling normality test
##
## data:  df[, columnName]
## A = 348.78, p-value < 2.2e-16
##
##
##  Anderson-Darling normality test
##
## data:  df[, columnName]
## A = 1656.5, p-value < 2.2e-16
##
##
##  Anderson-Darling normality test
##
## data:  df[, columnName]
## A = 2738.1, p-value < 2.2e-16
##
##
##  Anderson-Darling normality test
##
## data:  df[, columnName]
## A = 15000, p-value < 2.2e-16
```

Veiem doncs, que el test ens desmenteix el que creiem. Com que per a tots els casos, el p-valor és inferior al nostre coeficient alpha (0,05), hem de rebutjar l'hipotesi nul·la i entenem que no segueixen una distribució normal.

Tot i això, gràcies el teorema del límit central i al fet de tenir més de 30 mostres, podem aproximar les variables com a una distribució normal de mitja 0 i desviació estandar 1.

4.3 Proves estadístiques

Existeix una diferència de genere respecte els ingressos?

La nostra hipotesi és:

- $H_0 = \mu_{male} - \mu_{female} = 0$
- $H_1 = |\mu_{male} - \mu_{female}| \neq 0$

Anem a mirar com esta distribuit:

```
table(df$sex, df$income)
```

```
##
##          <=50K  >50K
##   Female 14423  1769
##   Male   22732  9918
```

```
prop.table(table(df$sex, df$income), margin = 1)
```

```
##
##              <=50K      >50K
##   Female 0.8907485 0.1092515
##   Male   0.6962328 0.3037672
```

A simple vista podem veure que la proporció de dones qu es troben a la part baixa d'ingressos és més elevada. A més, s'observa que hi ha més homes que dones en el dataset.

Anem doncs a aplicar el test Chi-squared per determianr la independència entre aquestes dues variables.

```
chisq.test(table(df$sex, df$income))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(df$sex, df$income)
## X-squared = 2248.8, df = 1, p-value < 2.2e-16
```

S'observa un p-valor molt petit, per tant rebutgem l'hipotesi nul·la i ens quedem amb l'alternativa; és a dir, les dues variables no són independents. Podem afirmar doncs que existeix un biaix de sexe en el nivell d'ingressos.

Correlacio entre variables

A continuació, anem a veure la correlació entre totes les variables del conjunt de dades. Per a fer-ho ens ajudarem de la llibreria `polycor` que ens permet calcular la correlació entre conjunt de dades numèriques i categòriques.

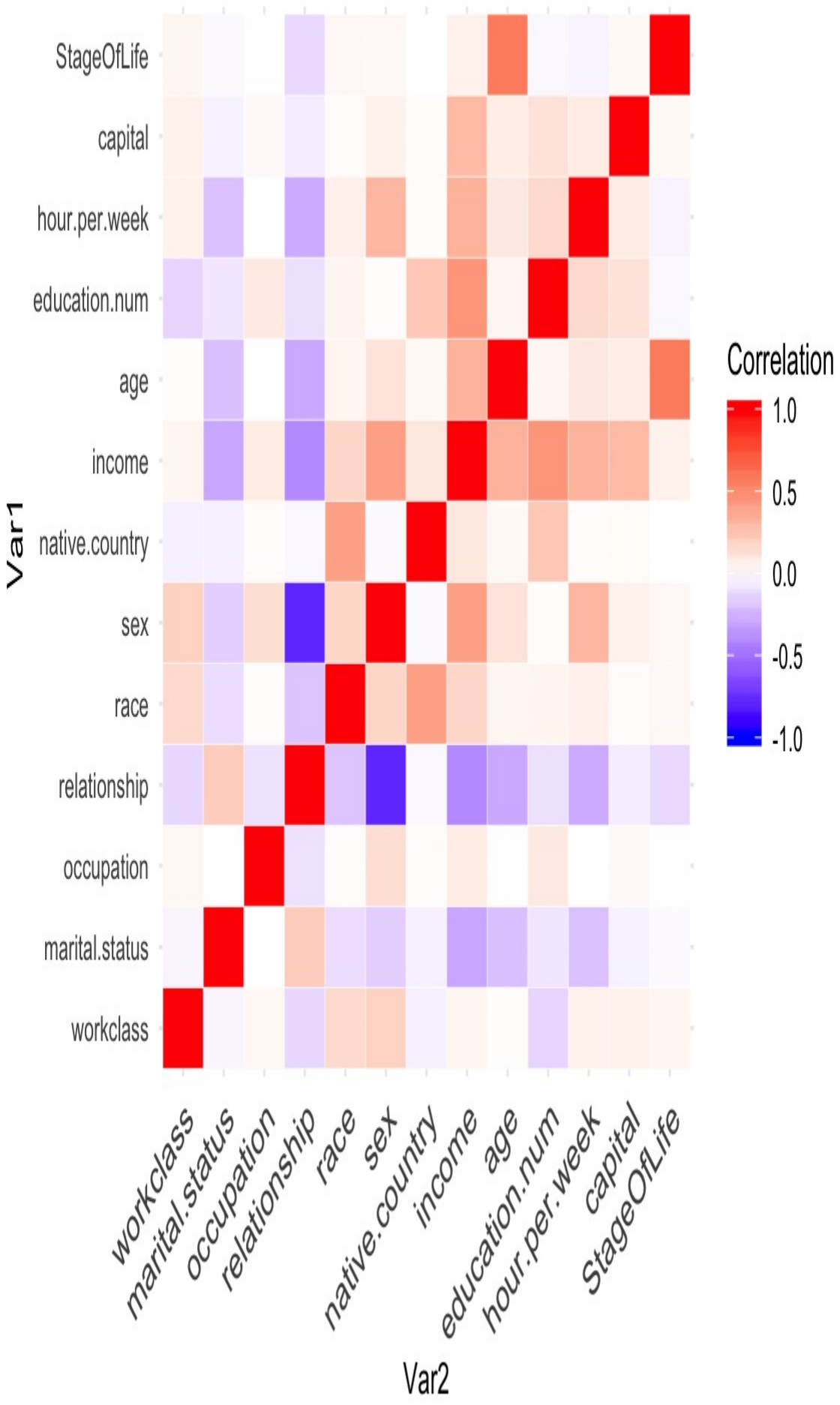
```
library(polycor)

corr_matrix<-hetcor(df, ML=FALSE, std.err=FALSE)
corr_matrix$correlations
```

```
##                  workclass marital.status    occupation relationship
## workclass       1.00000000  -0.0330157751  0.0325338030  -0.13300374
## marital.status -0.03301578   1.0000000000  0.0008640775   0.21461324
## occupation      0.03253380   0.0008640775  1.0000000000  -0.09672095
## relationship   -0.13300374   0.2146132375 -0.0967209455   1.00000000
## race            0.15834633  -0.1087583035  0.0110731271  -0.19428896
## sex             0.18859015  -0.1613443315  0.1373048140  -0.78712749
## native.country -0.04698659  -0.0451286256  0.0150626806  -0.02581032
## income          0.03750627  -0.2994146661  0.0804217471  -0.41341946
## age             0.00944376  -0.2071620407 -0.0023788894  -0.29558877
## education.num  -0.14183127  -0.0853116017  0.0890145032  -0.10071906
## hour.per.week   0.05996798  -0.2043914546 -0.0003748108  -0.28778011
## capital         0.05607929  -0.0429202750  0.0186953917  -0.05923906
## StageOfLife     0.04344715  -0.0164374493 -0.0059391914  -0.12391507
##                       race          sex native.country      income
## workclass       0.15834633  0.18859015   -0.046986595  0.03750627
## marital.status -0.10875830 -0.16134433   -0.045128626 -0.29941467
## occupation      0.01107313  0.13730481    0.015062681  0.08042175
## relationship   -0.19428896 -0.78712749   -0.025810316 -0.41341946
## race            1.00000000  0.16865959    0.404639359  0.16624439
## sex             0.16865959  1.00000000   -0.015879968  0.40652903
## native.country  0.40463936 -0.01587997    1.000000000  0.09981953
## income          0.16624439  0.40652903    0.099819529  1.00000000
## age             0.04259186  0.11598802    0.025548153  0.32703622
## education.num   0.04416966  0.01210332    0.235178802  0.45722847
## hour.per.week   0.06263339  0.30598976    0.011085596  0.33010373
## capital         0.01592477  0.05773198    0.017539434  0.29466624
## StageOfLife     0.02939317  0.03141232   -0.001161956  0.05220074
##                        age education.num hour.per.week     capital
## workclass       0.009443760   -0.14183127  0.0599679820  0.05607929
## marital.status -0.207162041   -0.08531160 -0.2043914546 -0.04292027
```

```
## occupation     -0.002378889    0.08901450 -0.0003748108  0.01869539
## relationship   -0.295588773   -0.10071906 -0.2877801108 -0.05923906
## race            0.042591857    0.04416966  0.0626333899  0.01592477
## sex             0.115988019    0.01210332  0.3059897574  0.05773198
## native.country  0.025548153    0.23517880  0.0110855959  0.01753943
## income          0.327036222    0.45722847  0.3301037291  0.29466624
## age             1.000000000    0.03616739  0.0935512109  0.07477744
## education.num   0.036167391    1.00000000  0.1574493477  0.12038736
## hour.per.week   0.093551211    0.15744935  1.0000000000  0.07885785
## capital         0.074777444    0.12038736  0.0788578494  1.00000000
## StageOfLife     0.571021541   -0.02737359 -0.0363585950  0.02603332
##                 StageOfLife
## workclass       0.043447147
## marital.status -0.016437449
## occupation     -0.005939191
## relationship   -0.123915070
## race            0.029393167
## sex             0.031412320
## native.country -0.001161956
## income          0.052200742
## age             0.571021541
## education.num  -0.027373593
## hour.per.week  -0.036358595
## capital         0.026033322
## StageOfLife     1.000000000
```

Per a una simple interpretació, mostrarem la matriu de correlació resultant mitjançant un mapa de calor.

```
library(reshape2)

ggplot(
  melt(corr_matrix$correlations),
  aes(Var2, Var1, fill = value)
)+
geom_tile(color = "white")+
scale_fill_gradient2(
  low = "blue",
  high = "red",
  mid = "white",
  midpoint = 0,
  limit = c(-1,1),
  space = "Lab",
  name="Correlation") +
theme_minimal()+ # minimal theme
theme(
  axis.text.x = element_text(
    angle = 45, vjust = 1,
    size = 12, hjust = 1))+
 coord_fixed()
```



Dels resultats anteriors s'observa:

- Tal i com era de suposar, existeix una forta relació entre la variable `age` i la variable

`StageOfLife`.

- Existeix un relació entre el sexe i el tipus de relació sentimental.
- Hi ha relació entre el país natiu i el nivell d'estudis.
- Hi ha relació entre el país natiu i la raça.
- Les principals relacions amb els ingressos són: l'estat civil, el tipus de relació, sexe, i educació. En segon pla trobem: raça, edat, hores treballades i capital.

Model de regressió logística

Tal com s'ha comentat previament, sería d'interés poder realitzar prediccions sobre la possibilitat de poder tenir un nivell d'ingresos superior o no a 50000 dolars. Per a fer-ho es calcularà un model de regressió logística utilitzant tant les variables quantitatives com les qualitatives.

Primer ens dividim el conjunt en train/test.

```
library(caTools)
# Fixem un valor per a que es pugui reproduir
set.seed(420)
samples = sample.split(df$age, SplitRatio = 0.75)

# Per a aquest prova usarem la variable de edat que indica l'etapa de la vida. Per
això borrem la variable age
df_temp <- df
df_temp$age <- NULL

df_train <- subset(df_temp, samples == TRUE)
df_test <- subset(df_temp, samples == FALSE)
```

Construirem una regressió logística utilitzant la variable `income` com a sortida i totes les altres variables com a predictors.

```
regLog <- glm(income ~ ., data = df_train, family = binomial('logit'))
summary(regLog)
```

```
##
## Call:
## glm(formula = income ~ ., family = binomial("logit"), data = df_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.4673  -0.5231  -0.1973  -0.0181   3.9169
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -9.669e+00  6.918e-01 -13.976  < 2e-16 ***
## workclassNo-work            -2.170e-01  8.197e-01  -0.265  0.79124
## workclassPrivate             1.146e-02  4.876e-02   0.235  0.81418
## workclassSelf               -1.297e-01  6.251e-02  -2.075  0.03794 *
## marital.statusMarried        7.971e-01  1.515e-01   5.263 1.42e-07 ***
## marital.statusNever-married -4.863e-01  7.906e-02  -6.151 7.71e-10 ***
## marital.statusSeparated     -4.322e-02  1.478e-01  -0.292  0.76999
## marital.statusWidowed        4.198e-01  1.369e-01   3.067  0.00216 **
## occupationArmed-Forces       9.613e-01  1.064e+00   0.903  0.36636
## occupationCraft-repair      -7.777e-02  7.150e-02  -1.088  0.27668
## occupationExec-managerial    7.646e-01  6.849e-02  11.165  < 2e-16 ***
## occupationFarming-fishing   -1.090e+00  1.244e-01  -8.765  < 2e-16 ***
## occupationHandlers-cleaners -7.604e-01  1.279e-01  -5.947 2.73e-09 ***
## occupationMachine-op-inspct -3.827e-01  9.188e-02  -4.165 3.12e-05 ***
## occupationOther-service     -9.106e-01  1.048e-01  -8.691  < 2e-16 ***
## occupationPriv-house-serv   -1.250e+00  7.184e-01  -1.740  0.08192 .
## occupationProf-specialty     5.186e-01  7.068e-02   7.338 2.17e-13 ***
## occupationProtective-serv    2.711e-01  1.118e-01   2.424  0.01537 *
## occupationSales              2.161e-01  7.354e-02   2.938  0.00330 **
```

```
## occupationTech-support       4.658e-01  1.002e-01   4.651 3.31e-06 ***
## occupationTransport-moving  -1.922e-01  8.905e-02  -2.158  0.03093 *
## relationshipNot-in-family   -9.544e-01  1.466e-01  -6.508 7.60e-11 ***
## relationshipOther-relative  -1.294e+00  2.024e-01  -6.393 1.63e-10 ***
## relationshipOwn-child       -1.774e+00  1.817e-01  -9.766  < 2e-16 ***
## relationshipUnmarried       -1.186e+00  1.661e-01  -7.144 9.06e-13 ***
## relationshipWife             1.105e+00  9.374e-02  11.785  < 2e-16 ***
## raceAsian-Pac-Islander       6.180e-01  2.252e-01   2.744  0.00607 **
## raceBlack                    3.601e-01  2.108e-01   1.708  0.08767 .
## raceOther                    4.086e-01  2.997e-01   1.364  0.17270
## raceWhite                    6.125e-01  2.008e-01   3.050  0.00229 **
## sexMale                      7.461e-01  7.174e-02  10.400  < 2e-16 ***
## native.countryUnited-States  2.827e-01  7.176e-02   3.939 8.18e-05 ***
## education.num                2.707e-01  8.621e-03  31.402  < 2e-16 ***
## hour.per.week                3.402e-02  1.693e-03  20.091  < 2e-16 ***
## capital                      2.544e-04  8.201e-06  31.024  < 2e-16 ***
## StageOfLifeAdult             2.529e+00  6.272e-01   4.032 5.54e-05 ***
## StageOfLifeAdult-primerenc   1.007e+00  6.339e-01   1.589  0.11217
## StageOfLifeVellesa           2.646e+00  6.275e-01   4.217 2.48e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 40366  on 36632  degrees of freedom
## Residual deviance: 23968  on 36595  degrees of freedom
## AIC: 24044
##
## Number of Fisher Scoring iterations: 9
```

La regressió logística està modelant la probabilitat que un individu faci més de 50.000 dòlars anuals. Per tant, una resposta més propera a 1 indica una possibilitat més elevada de guanyar més de 50.000 dòlars, mentre que una resposta més propera a 0 indica una possibilitat més gran de guanyar menys de 50.000 dòlars. Així, s'utilitza un llindar de 0,5 per determinar si es preveu que un individu guanyi més de 50.000 dòlars anuals o no.

Anem a representar la matriu de confusió per avaluar el nivell de predicció dels ingressos.

```
prob <- predict(regLog, df_test, type = 'response')
pred <- rep('<=50K', length(prob))
pred[prob >= 0.5] <- '>50K'

# Matriu de confusio
conf_matrix <- table(pred, df_test$income)
conf_matrix
```

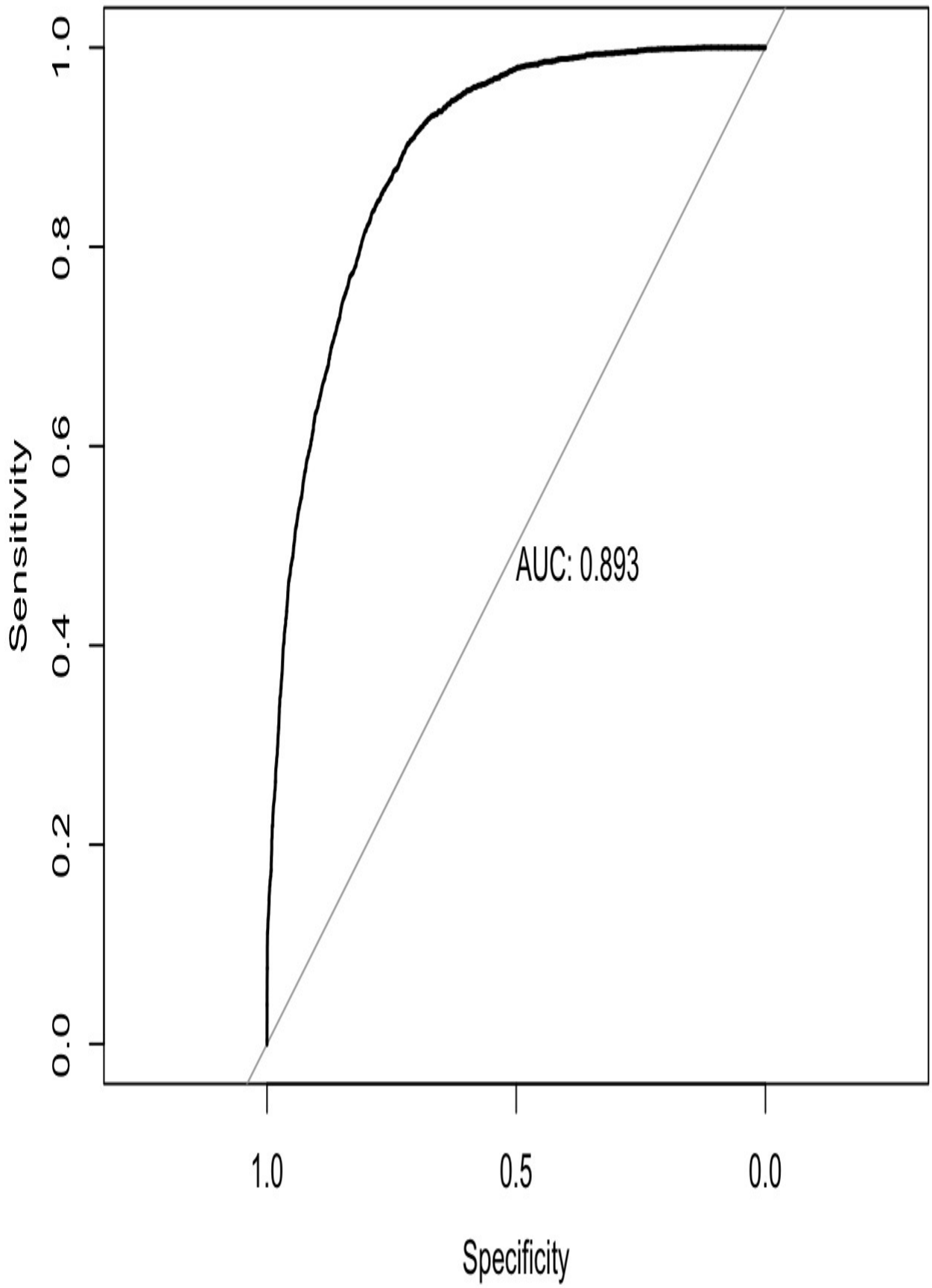
```
##
## pred    <=50K >50K
##   <=50K  8632 1263
##   >50K    678 1636
```

Així doncs obtenim una presició (en percentatge):

```
acc = (conf_matrix[1,1]+conf_matrix[2,2])/sum(conf_matrix)*100
acc
```

```
## [1] 84.10189
```

```
library(pROC)
roc(df_test$income ~ prob, plot = TRUE, print.auc = TRUE)
```




```
##
## Call:
## roc.formula(formula = df_test$income ~ prob, plot = TRUE, print.auc = TRUE)
##
```

```
## Data: prob in 9310 controls (df_test$income <=50K) < 2899 cases (df_test$income
>50K).
## Area under the curve: 0.8926
```

S'obté una area sota la corba (AUC) que ens permet concloure que el model és bo.

Finalment anem a veure els Odd Ratios per a saber l'aportació de cada coeficient:

```
data.frame(V1=sort(exp(coefficients(regLog)), decreasing=TRUE))
```

D'aquest resultat se'n poden extreure més biaxos socials (respecte ingressar més de 50000 dolars).

- De totes les races, la raça negra és la que té el OR més baix.
- El sexe masculí té 2.1 més de OR.
- Ser natiu de EEUU té 1.32 més de OR.

A més, se n'extreuen altres conclusion com:_

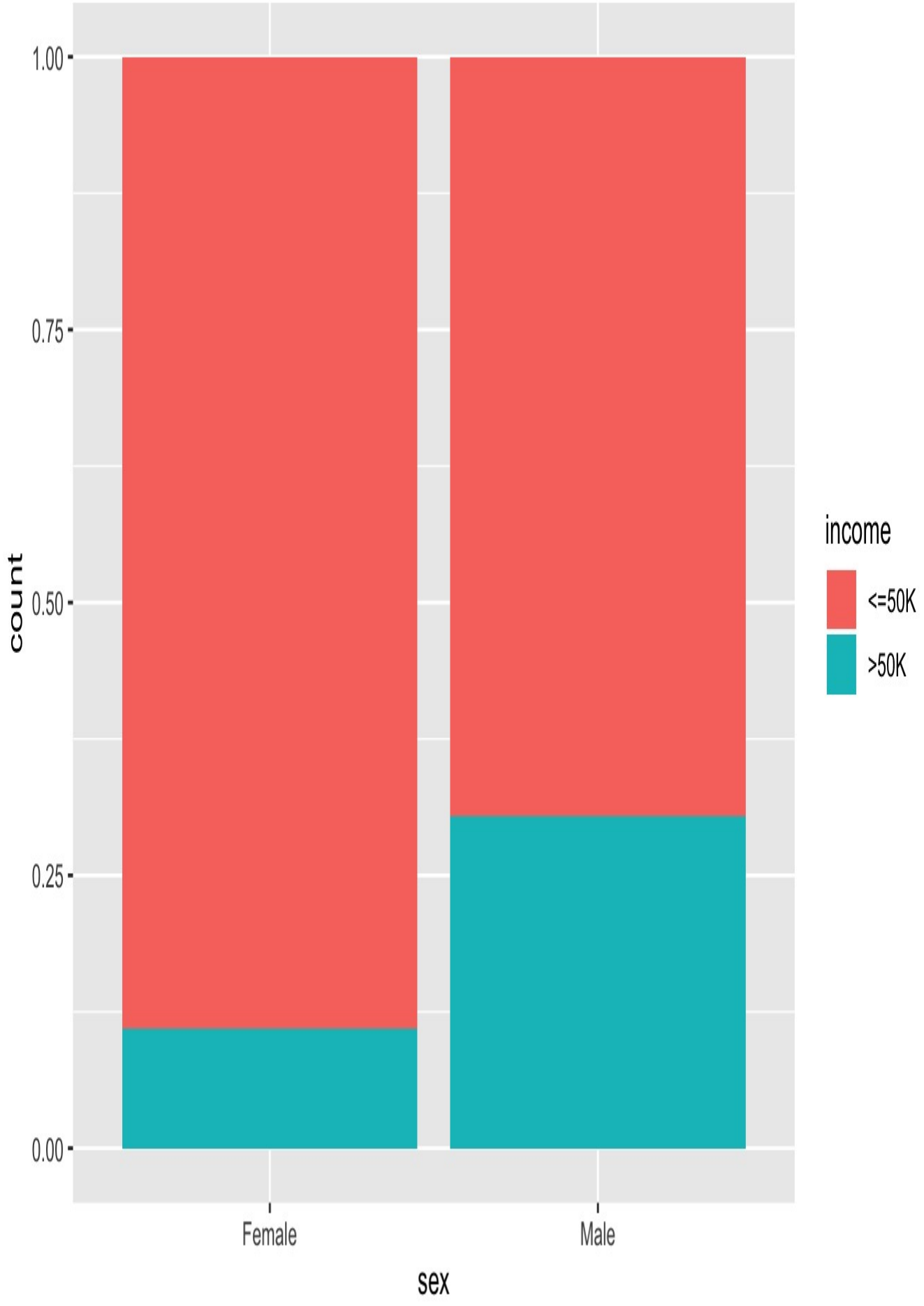
- Per cada any extra d'estudi, augmenta en 1,32 el OR.
- Quan més vell més OR tens.
- Les families (es considera `relationshipWife` i `marital.statusMarried`) tenen més OR que la gent separada, divorciada o viuda.
- Les ocupacions de rol executiu i forces armades tenen més OR que les altres.

5 Gràfiques

A part de les gràfiques anterior, podem afegir-ne d'extres que reforcen les conclusions extretes:

Visualitzarem la relació entre la variable `sex` i `income`.

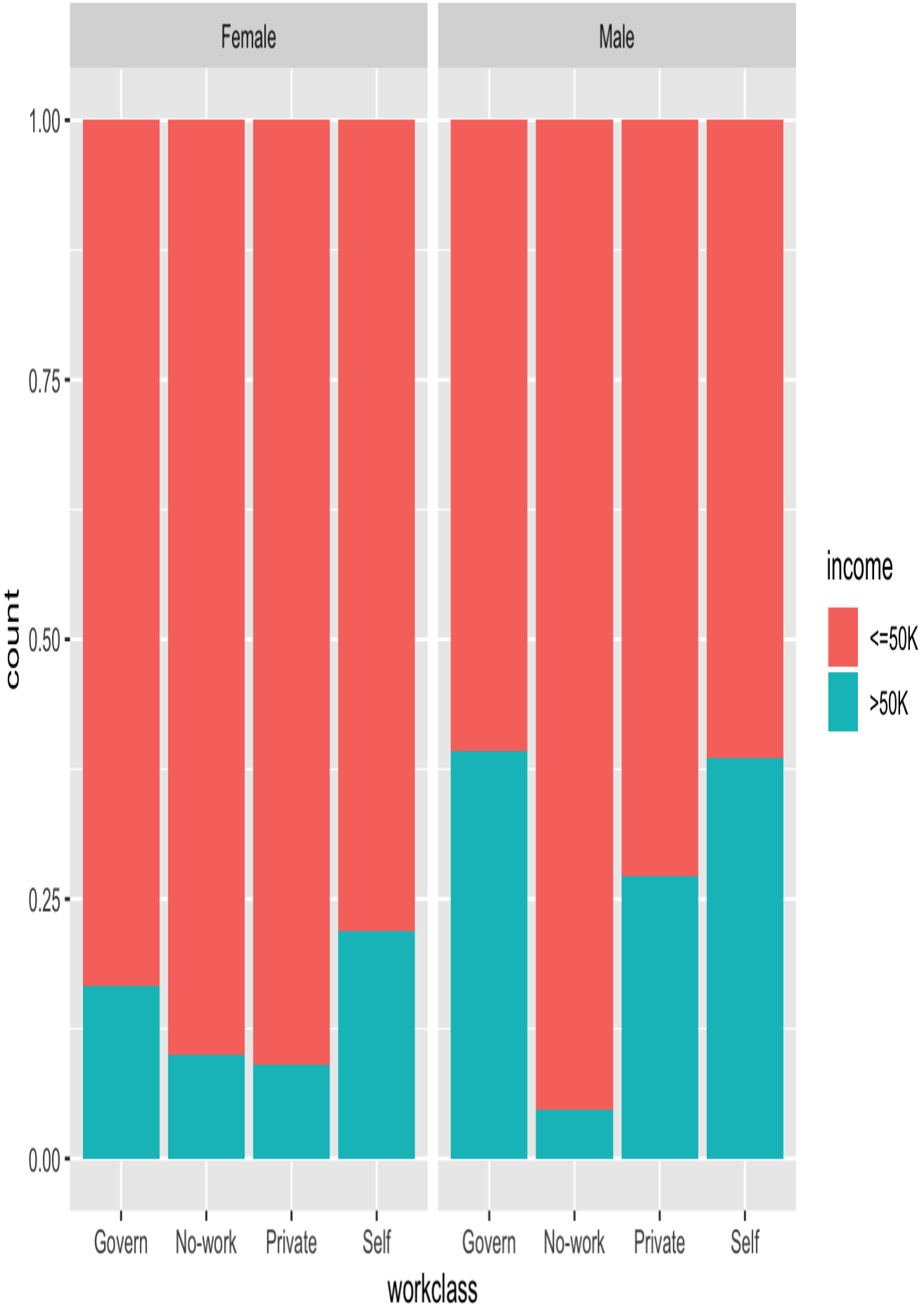
```
ggplot(
    data=df,
aes( x=sex,
        fill=income
    )
)+geom_bar(position="fill")
```



Anem a observar ara la relació `workclass` amb `income` i `sex` a la vegada.

```
# Visualitzem la relació entre les variables "workclass", "income" i "sex".
```

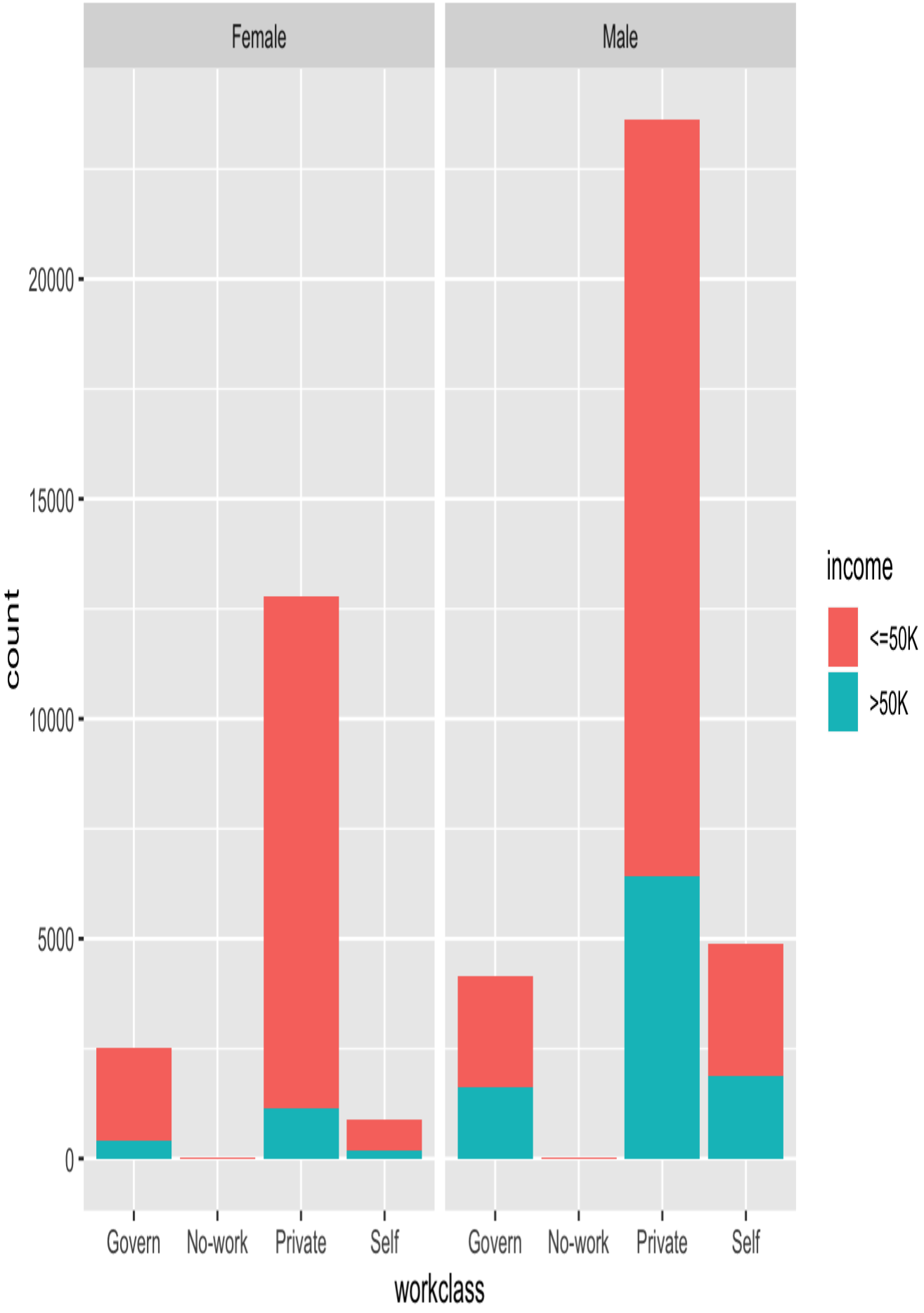
```
ggplot(
    data=df,
    aes(
      x=workclass,
      fill=income
    )
) + geom_bar(position="fill") + facet_wrap(~sex)
```



En aquest cas veiem que, en el cas de les dones, les que treballen per elles mateixes són les que tenen més possibilitats de superar els 50K$, mentre que en els homes està disputat entre els que treballen pel govern i els que treballen per ells mateixos.

Anem a veure ara la quantitat de persones repartides segons la feina:

```
ggplot(
  data=df,
  aes(
    x=workclass,
    fill=income
  )
) + geom_bar() + facet_wrap(~sex)
```



Tot i que hem vist que les dones que treballen per elles mateixes són les que tenen més probabilitats de superar el llindar de 50K$, veiem que és el sector amb menys presència d'aquest sexe. El cas més present en ambdós sexes és treballar a una empresa privada.

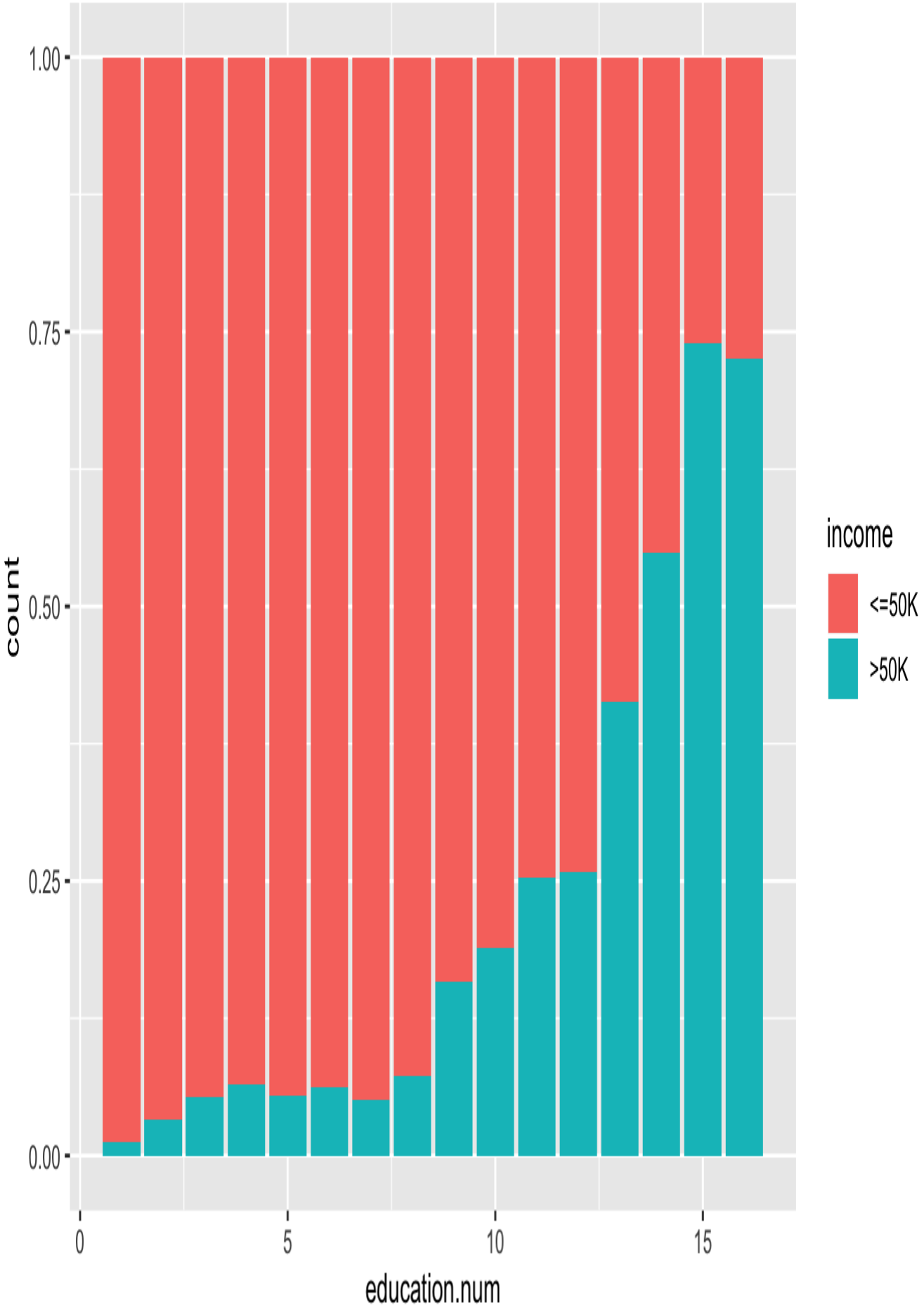
Si ens fixem en el tipus de feina, anem a veure quina feina és més probable segons els estudis d'una persona.

```
# Guardem una variable amb els percentatges que relacionen ocupacio ieducació
pWorkEdu <- prop.table(table(df$occupation, df$education.num), 1) * 100
# Agafem el nom de la fila que té el màxim de probabilitat
apply(pWorkEdu, 2, function(x) rownames(pWorkEdu)[which.max(x)])
```

```
##                   1                   2                   3
##   "Farming-fishing"   "Priv-house-serv"   "Priv-house-serv"
##                   4                   5                   6
##   "Priv-house-serv"   "Priv-house-serv"     "Other-service"
##                   7                   8                   9
## "Handlers-cleaners"      "Armed-Forces"  "Transport-moving"
##                  10                  11                  12
##      "Adm-clerical"      "Tech-support"      "Tech-support"
##                  13                  14                  15
##    "Prof-specialty"    "Prof-specialty"    "Prof-specialty"
##                  16
##    "Prof-specialty"
```

Aquest informació la podem fer servir, a la vegada, per veure la diferència entre el nivell d'ingressos segons l'estudi (i relecionar-ho amb la posició laboral).

```
ggplot(
    data=df,
aes( x=education.num,
        fill=income
    )
)+geom_bar(position="fill")
```

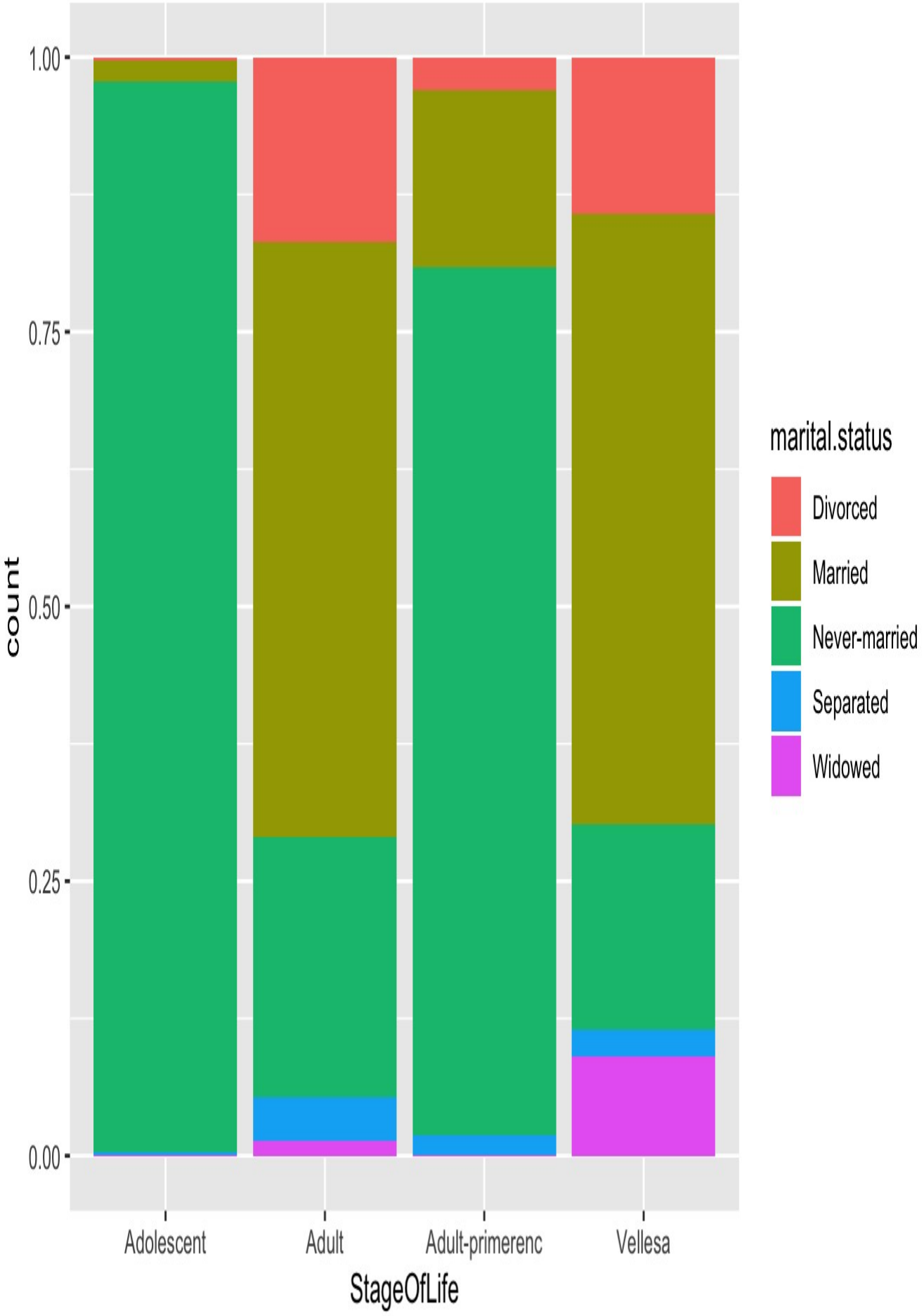


Per tant confirmem que a més educació més possibilitats de superar els 50000 dolars d'ingresos.

A continuació ens proposem observar la relació entre "marital-status" i la nova variable que ens hem creat

StageOfLife.

```
# Relació entre les variables "StageOfLife" i "marital.status"
ggplot(
    data=df,
    aes(
      x=StageOfLife,
      fill=`marital.status`
    )
) + geom_bar(position="fill")
```

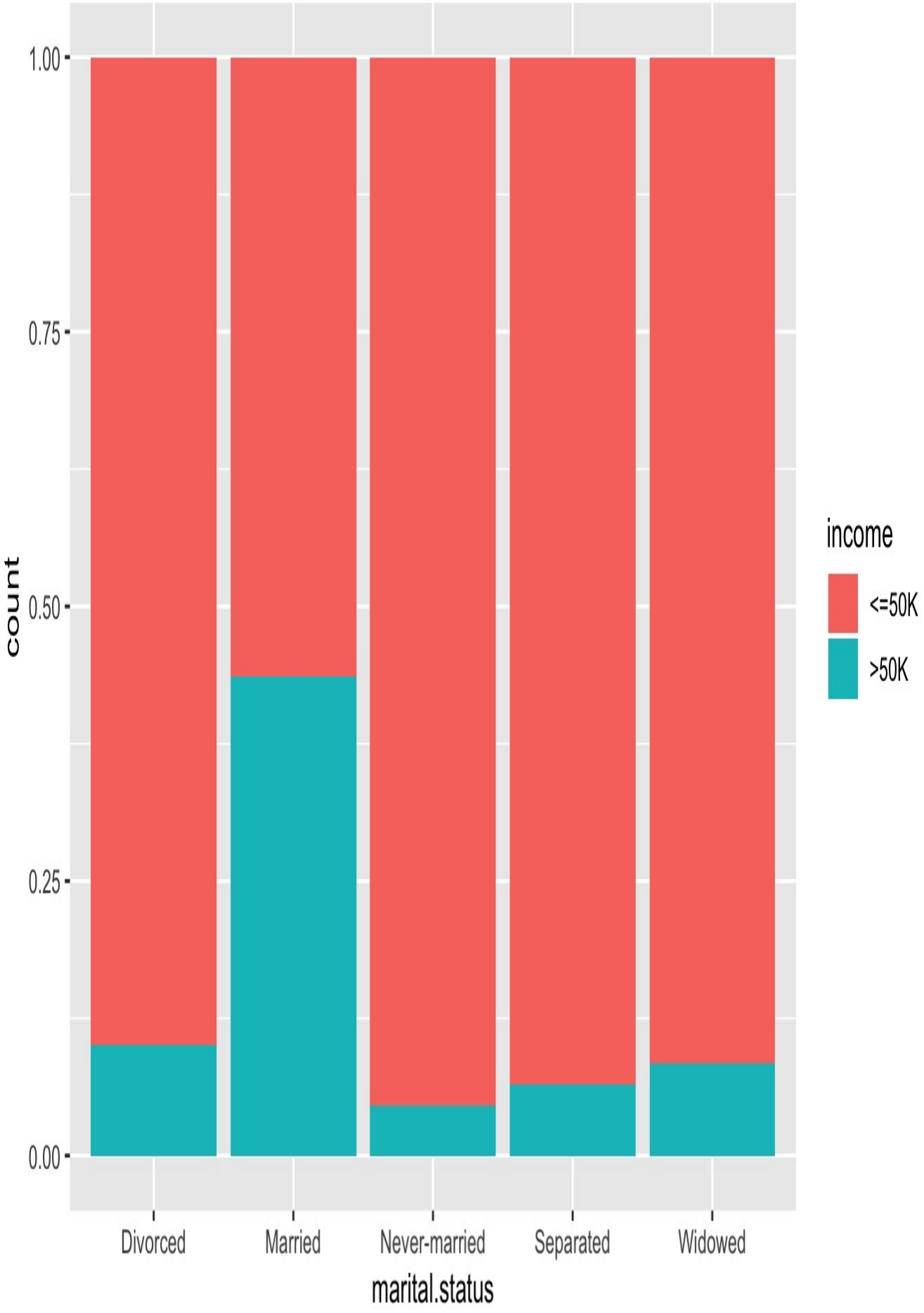


Veiem que es compleix clarament el que suposàvem. A l'adolescència la gran majora mai s'ha casat. Al principi de l'adultesa comencen a incrementar els valors de casaments i, amb menys mesura, de divorcis. A la següent etapa s'incrementen notablement els casaments i també els divorcis. Finalment, a l'última

etapa, els valors anterior es queden bastant estancats però el nombre de persones viudes augmenta.

I si veiem la relació entre `marital-status` i `income`:

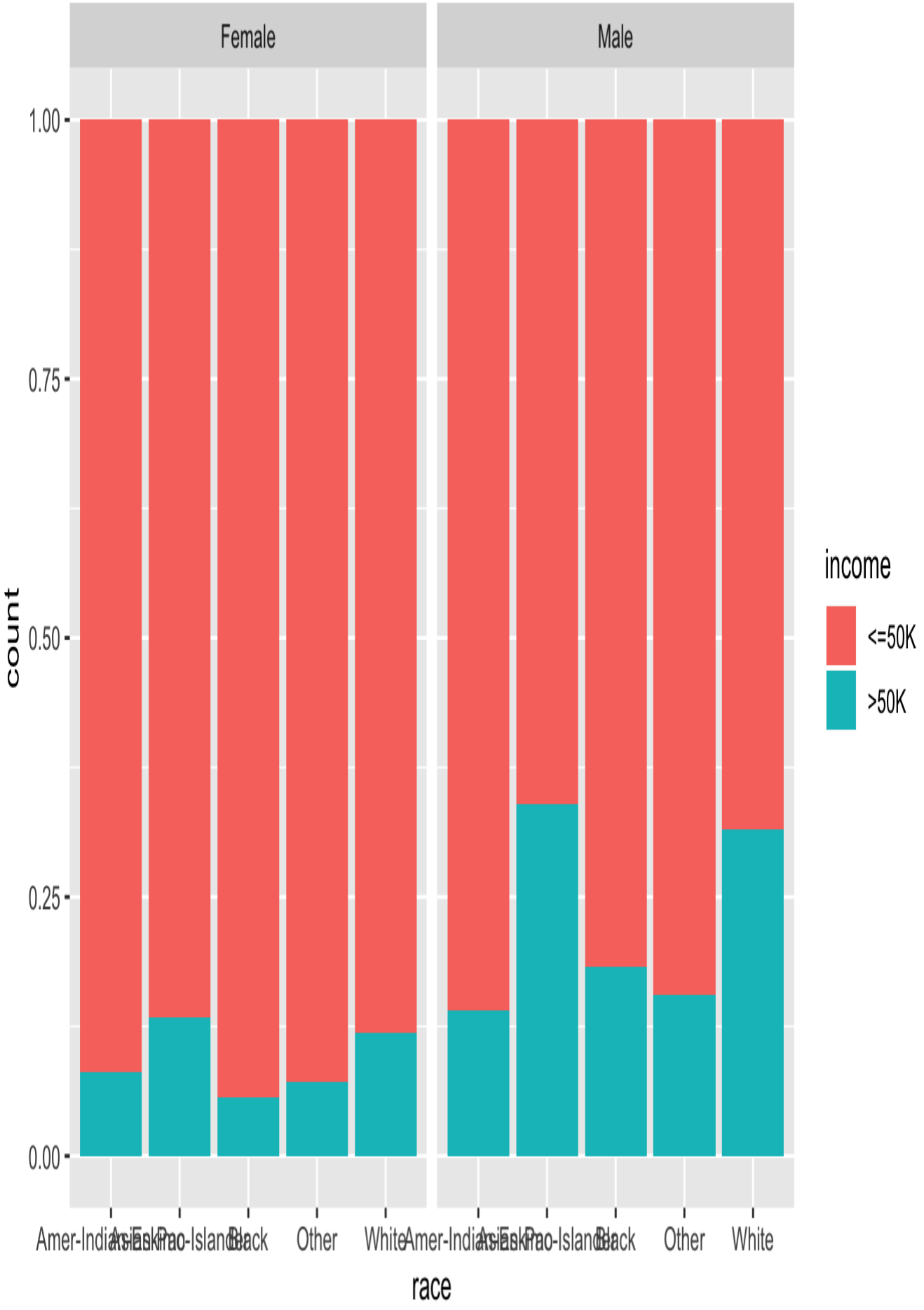
```
ggplot(
  data=df,
  aes(
    x=`marital.status`,
    fill=income
  )
) + geom_bar(position="fill")
```



Podem observar que de les persones que estan casades quasi la meitat del grup passa el llindar de 50K$.

Anem a mirar la comparació segons la raça:

```
ggplot(
  data=df,
  aes(
x=`race`,
    fill=income
  )
) + geom_bar(position="fill") + facet_wrap(~sex)
```



Podem dir que la raça “Blanca” i la “Asian- Pac-Islander” són les que tenen més possibilitats de superar el llindar. Concretament els homes.

6 Conclusions

Al llarg de tota la pràctica s'han dut a terme tota una sèrie de tècniques amb la finalitat principal de netejar les dades i analitzar-les. Primer de tot s'han detectat els valor buits i s'han imputat mitjançant kNN. Més tard, s'han tractat els outliers que s'han cregut necessaris i s'han agrupat i seleccionat les dades en diferents conjunts per al seu posterior analisis.

En el nostre anàlisi, voliem mostrar si hi havia un biaix clar que comprometes les característiques sociològiques dels registres (tals com sexe, raça, nivell d'estudis) envers el seu nivell d'ingressos. En la primera prova de totes, s'ha demostrat que no existeix independència entre el sexe de la persona i el seu nivell d'ingressos, cosa que ens fa pensar en una escletxa salarial basada en el sexe.

En la segona prova, s'ha buscat la correlació entre totes les dades del conjunt i s'ha pogut veure quines eren les seves relacions. Cal destacar que s'han trobat relacions entre el sexe i l'estat cívil; entre la nacionalitat i la raça; i la nacionalitat i el nivell d'estudis. Vull esmentar que no s'ha trobat relació entre la raça i el nivell d'estudis, però si entre la raça i el nivell d'ingressos, fet que ens porta a pensar en una escletxa salarial basada en la raça també. A part, factors que estan principalment relacionats amb els ingressos són: l'estat civil, el tipus de relació, sexe (que ens reafirma la conclusió anterior), i l'educació; i en en menys mesura amb la raça, edat, hores treballades i capital.

Per finalitzar, s'ha fet una regressió logística per a predir si el nivell d'ingressos és inferior o superior als 50000 dolars a partir de la resta de variables (agrupades) del dataset. Els resultats del model obtingut han sigut bons, amb una precisió superior al 84%. A més, quan s'ha observat l'aportació de OR de cada coeficient s'ha detectat que de totes les races, la raça negra és la que té el OR més baix, el sexe masculí té 2.1 més de OR i que ser natiu de EEUU té 1.32 més de OR.

7 Codi i dades

El códi en R es pot trobar a GitHub en el següent enllaç:

https://github.com/epou/adult_income/blob/master/code/census_analysis.R

Les dades de sortida un cop netejades les dades s'han exportat mitjançant la comanda següent:

```
write.csv(df, file="adult-out.csv")
```

Es troben al següent enllaç de GitHub:
https://github.com/epou/adult_income/blob/master/csv/output/adult-out.csv