Tipología y Ciclo de Vida de los Datos: Práctica 2 - Limpieza y análisis de datos

Autor: Eréndira Teresa Navarro García

Enero 20202

Contents

Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder? **1**

Integración y selección de los datos de interés a analizar. **1**

Limpieza de los datos. **2**
¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos? 3
Identificación y tratamiento de valores extremos. 7
Comprobación de la normalidad y homogeneidad de la varianza. 22
Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes. 22

Representación de los resultados a partir de tablas y gráficas. **30**

Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema? **32**

Código: Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos. Si lo preferís, también podéis trabajar en Python. **32**

En el repositorio https://github.com/ernavaga/AdultIncomeCensus_UOC se encuentra este y el resto de los documentos solicitados

Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?

El conjunto de datos utilizado es Dataset Adult https://archive.ics.uci.edu/ml/datasets/Adult, estos datos provienen del censo de 1994 en Estados Unidos. La extracción fue hecha por Barry Becker, el conjunto de datos ya tiene estos filtros ((AAGE>16) && (AGI>100) && (AFNLWGT>1)&& (HRSWK>0)). Lo que se busca con estos datos es identificar las cractaerísticas que determinan que una persona gane más o menos de 50 mil dólares al año.

Integración y selección de los datos de interés a analizar.

El conjunto de datos contiene los siguientes campos:

- label: >50K, <=50K (etiqueta).
- age: continuous (edad).
- workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked (descripción del trabajo).
- fnlwgt: continuous (ponderador).

- education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool (último nivel de estudios).
- education-num: continuous (número de años de estudio).
- marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse (estatus marital).
- occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces (tipo de ocupación).
- relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried (tipo de relación con las demás personas).
- race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black (raza).
- sex: Female, Male (sexo).
- capital-gain: continuous (ganancia de capital).
- capital-loss: continuous (pérdida de capital).
- hours-per-week: continuous (horas de trabajo por semana).
- native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands (pais de origen).

Partiendo de la descripción de los datos las variables relationship y marital-status son similares, al igual que education y education-num. Se buscaría categorizar la variables continuas y se excluirá el ponderador.

Lo que se bucaría es obtener modelos supervisados con el target de >50K / <=50K.

Limpieza de los datos.

Lectura de dataset

Los datos vienen divididos ya en train y test set, para el tratamiento de los mismos se unirán ambos conjuntos. Se dividirán en variables categóricas y numéricas para su análisis posterior.

```
# Librerías
library(dplyr)
library(ggplot2)
library(gridExtra)
library(leaps)
library(Hmisc)
library(stringr)
library(C50)
library(caret)
library(grid)

# Lectura de datos
adult_train <- read.csv('http://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data',str

adult_test <- read.csv('https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.test',str

# Nombre de los atributos
names(adult_train) <- c("age","workclass","fnlwgt","education","education_num","marital_status","occupa

names(adult_test) <- c("age","workclass","fnlwgt","education","education_num","marital_status","occupat

adult_train["df"] <- "train"
```

```r
adult_test["df"] <- "test"
```

Ambos datasets se unen en uno solo para su tratamiento, se tienen 16 variables con 44,842 registros.

```r
adult <- rbind(adult_train,adult_test)

# Verificamos estructura de los conjuntos de datos
str(adult)
```

```
## 'data.frame':	48842 obs. of  16 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" ...
##  $ target        : chr  " <=50K" " <=50K" " <=50K" " <=50K" ...
##  $ df            : chr  "train" "train" "train" "train" ...
```

```r
catv <- c("workclass", "education", "marital_status", "occupation", "relationship",
          "race", "sex", "native_country", "target")

numv <- c("age", "capital_gain", "capital_loss", "hour_per_week")
```

¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?

Los datos tienen elementos "desconocidos" y están marcados con el símbolo "?", estos datos representan el 6% y están presentes en las variables workclass y occupation, al igual aparece esto en 2% de native-country.

Debido a la naturaleza de los datos, que provienen de un censo, se considera la posibilidad de utilizar la categoria "desconocido" como una categoría en si misma, sobre todo porque se planea agrupar categorias.

Descriptivos variables categóricas:

```r
describe(adult[catv])
```

```
## adult[catv]
##
##  9  Variables      48842  Observations
## --------------------------------------------------------------------------------
## workclass
##        n  missing distinct
##    48842        0        9
##
## lowest :  ?                Federal-gov      Local-gov        Never-worked     Private
## highest:  Private          Self-emp-inc     Self-emp-not-inc State-gov        Without-pay
##
## ? (2799, 0.057), Federal-gov (1432, 0.029), Local-gov (3136, 0.064),
```

```
## Never-worked (10, 0.000), Private (33906, 0.694), Self-emp-inc (1695, 0.035),
## Self-emp-not-inc (3862, 0.079), State-gov (1981, 0.041), Without-pay (21,
## 0.000)
## ---------------------------------------------------------------------------
## education
##        n  missing distinct
##    48842        0       16
##
## lowest :  10th         11th         12th         1st-4th      5th-6th
## highest:  HS-grad      Masters      Preschool    Prof-school  Some-college
##
## 10th (1389, 0.028), 11th (1812, 0.037), 12th (657, 0.013), 1st-4th (247,
## 0.005), 5th-6th (509, 0.010), 7th-8th (955, 0.020), 9th (756, 0.015),
## Assoc-acdm (1601, 0.033), Assoc-voc (2061, 0.042), Bachelors (8025, 0.164),
## Doctorate (594, 0.012), HS-grad (15784, 0.323), Masters (2657, 0.054),
## Preschool (83, 0.002), Prof-school (834, 0.017), Some-college (10878, 0.223)
## ---------------------------------------------------------------------------
## marital_status
##        n  missing distinct
##    48842        0        7
##
## lowest :  Divorced              Married-AF-spouse     Married-civ-spouse    Married-spouse-absent
## highest:  Married-civ-spouse    Married-spouse-absent Never-married         Separated
##
## Divorced (6633, 0.136), Married-AF-spouse (37, 0.001), Married-civ-spouse
## (22379, 0.458), Married-spouse-absent (628, 0.013), Never-married (16117,
## 0.330), Separated (1530, 0.031), Widowed (1518, 0.031)
## ---------------------------------------------------------------------------
## occupation
##        n  missing distinct
##    48842        0       15
##
## lowest :  ?                 Adm-clerical      Armed-Forces      Craft-repair      Exec-managerial
## highest:  Prof-specialty    Protective-serv   Sales             Tech-support      Transport-moving
##
## ? (2809, 0.058), Adm-clerical (5611, 0.115), Armed-Forces (15, 0.000),
## Craft-repair (6112, 0.125), Exec-managerial (6086, 0.125), Farming-fishing
## (1490, 0.031), Handlers-cleaners (2072, 0.042), Machine-op-inspct (3022,
## 0.062), Other-service (4923, 0.101), Priv-house-serv (242, 0.005),
## Prof-specialty (6172, 0.126), Protective-serv (983, 0.020), Sales (5504,
## 0.113), Tech-support (1446, 0.030), Transport-moving (2355, 0.048)
## ---------------------------------------------------------------------------
## relationship
##        n  missing distinct
##    48842        0        6
##
## lowest :  Husband         Not-in-family   Other-relative  Own-child       Unmarried
## highest:  Not-in-family   Other-relative  Own-child       Unmarried       Wife
##
## Value             Husband  Not-in-family Other-relative      Own-child
## Frequency           19716          12583           1506           7581
## Proportion          0.404          0.258          0.031          0.155
##
## Value           Unmarried           Wife
```

```
## Frequency            5125          2331
## Proportion          0.105         0.048
## --------------------------------------------------------------------------------
## race
##        n  missing distinct
##    48842        0        5
##
## lowest :  Amer-Indian-Eskimo  Asian-Pac-Islander  Black               Other               White
## highest:  Amer-Indian-Eskimo  Asian-Pac-Islander  Black               Other               White
##
## Value      Amer-Indian-Eskimo Asian-Pac-Islander              Black
## Frequency                 470               1519               4685
## Proportion              0.010              0.031              0.096
##
## Value                   Other              White
## Frequency                 406              41762
## Proportion              0.008              0.855
## --------------------------------------------------------------------------------
## sex
##        n  missing distinct
##    48842        0        2
##
## Value      Female   Male
## Frequency   16192  32650
## Proportion  0.332  0.668
## --------------------------------------------------------------------------------
## native_country
##        n  missing distinct
##    48842        0       42
##
## lowest :  ?                 Cambodia         Canada           China            Columbia
## highest:  Thailand          Trinadad&Tobago  United-States    Vietnam          Yugoslavia
## --------------------------------------------------------------------------------
## target
##        n  missing distinct
##    48842        0        4
##
## Value       <=50K <=50K.   >50K  >50K.
## Frequency   24720  12435   7841   3846
## Proportion  0.506  0.255  0.161  0.079
## --------------------------------------------------------------------------------
```

```r
dim(adult[adult["native_country"]==" ?",])[1]/dim(adult)[1]
```

```
## [1] 0.01754637
```

Para las variables numéricas no se observan datos nulos explícitos, pero debido a la distribución se identifica que los casos donde capital_gain=99999 son nulos.

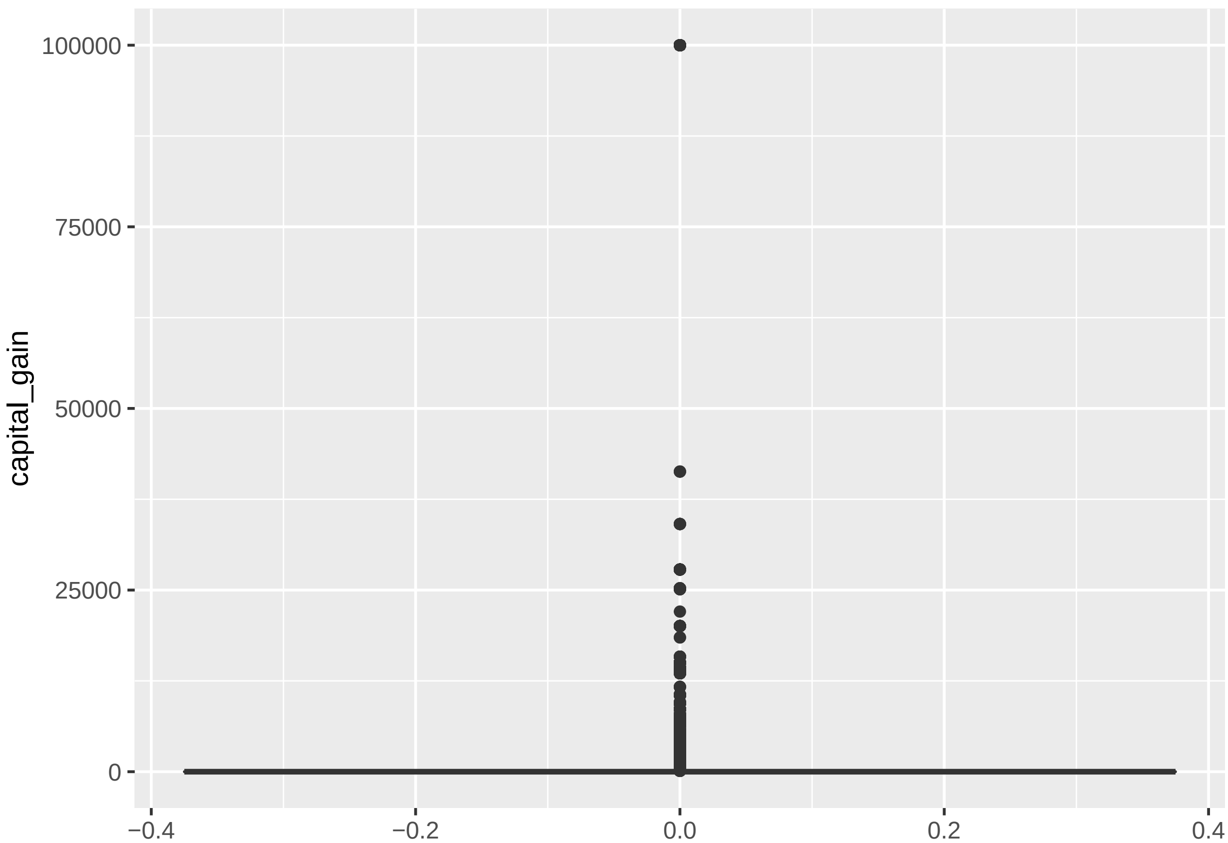
Descriptivos variables numéricas:

```r
########## Descriptivos numéricos
# libreria para descriptivos numéricos
library(psych)

# Descriptivos numéricos
```

```r
describe(adult[numv],quant=c(.25,.75))
```

```
##              vars     n    mean      sd median trimmed   mad min   max range
## age             1 48842   38.64   13.71     37   37.74 14.83  17    90    73
## capital_gain    2 48842 1079.07 7452.02      0    0.00  0.00   0 99999 99999
## capital_loss    3 48842   87.50  403.00      0    0.00  0.00   0  4356  4356
## hour_per_week   4 48842   40.42   12.39     40   40.54  4.45   1    99    98
##                skew kurtosis    se Q0.25 Q0.75
## age            0.56    -0.18  0.06    28    48
## capital_gain  11.89   152.67 33.72     0     0
## capital_loss   4.57    20.01  1.82     0     0
## hour_per_week  0.24     2.95  0.06    40    45
```

```r
# --------------- capital_gain ------------------
# boxplot para verificar el dato 99999
ggplot(data=adult, aes(y=capital_gain)) +
  geom_boxplot()
```




```r
# Se asume 99999 como valor nulo, debido a su distribución
nrow(adult[adult$capital_gain==99999,])
```

```
## [1] 244
```

```r
adult[adult$capital_gain == 99999,'capital_gain'] = NA
```

Identificación y tratamiento de valores extremos.

De acuerdo a la planeación, la variables numéricas se categorizarán, esto nos ayudará con los valores extremos presentes sobre todo en la variables de capital.

Limpieza variables categóricas.

- Eliminar "." de target

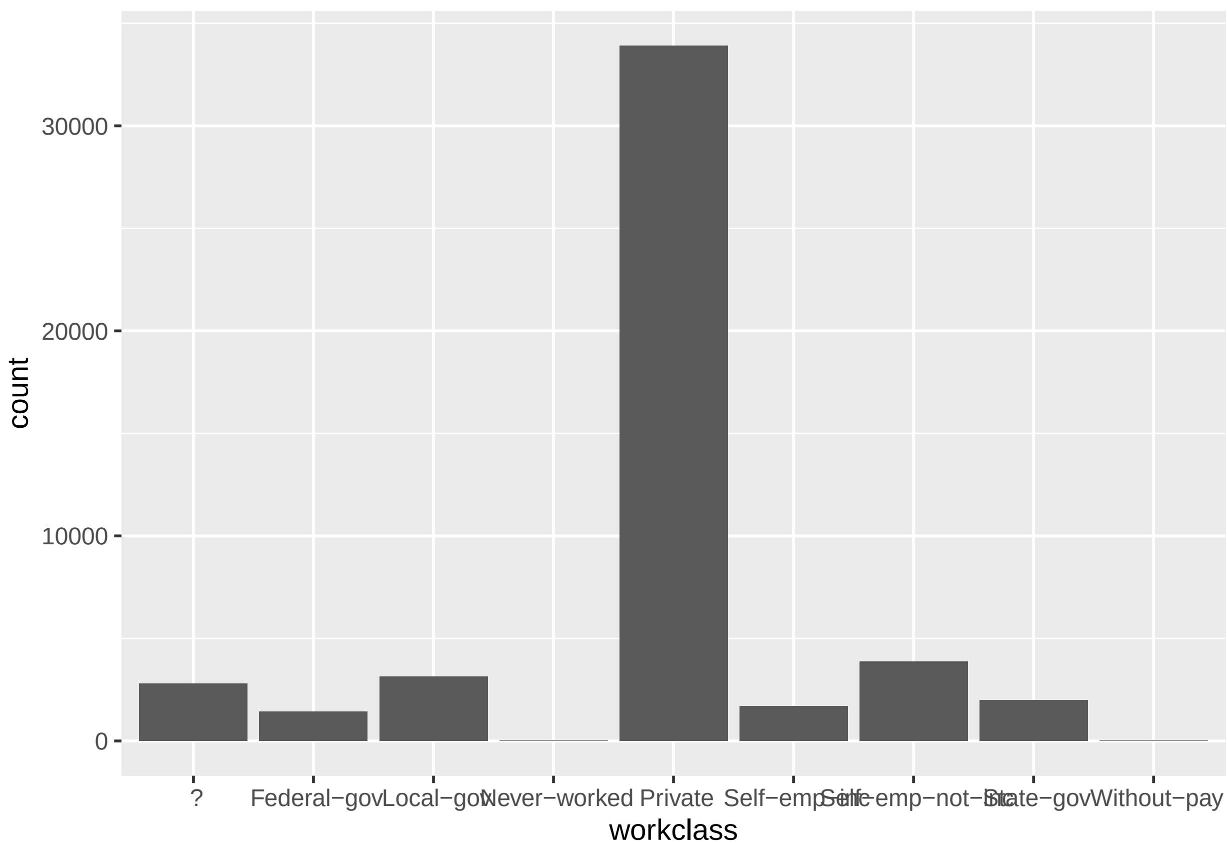
```
# Limpiar texto y recategorizar

# --------------- target ------------------
# Se elimina el "." que está presente en registros del dataset en la variable target
adult$target <- gsub("[.]", "", trimws(tolower(adult$target)))
```

- Recategorización workclass

```
# --------------- workclass ------------------
# muestra diferentes etiquetas -- INICIO
levels(as.factor(adult$workclass))
```

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

```
# Distribución inicial
ggplot(data=adult,aes(x=workclass)) + geom_bar()
```




```
# Estos valores se agrupan en 4 categorias generales: gov, priv, self y other
adult$workclass[grepl("gov",trimws(adult$workclass),ignore.case = T)] <- 'gov'
adult$workclass[grepl("self",trimws(adult$workclass),ignore.case = T)] <- 'self'
```

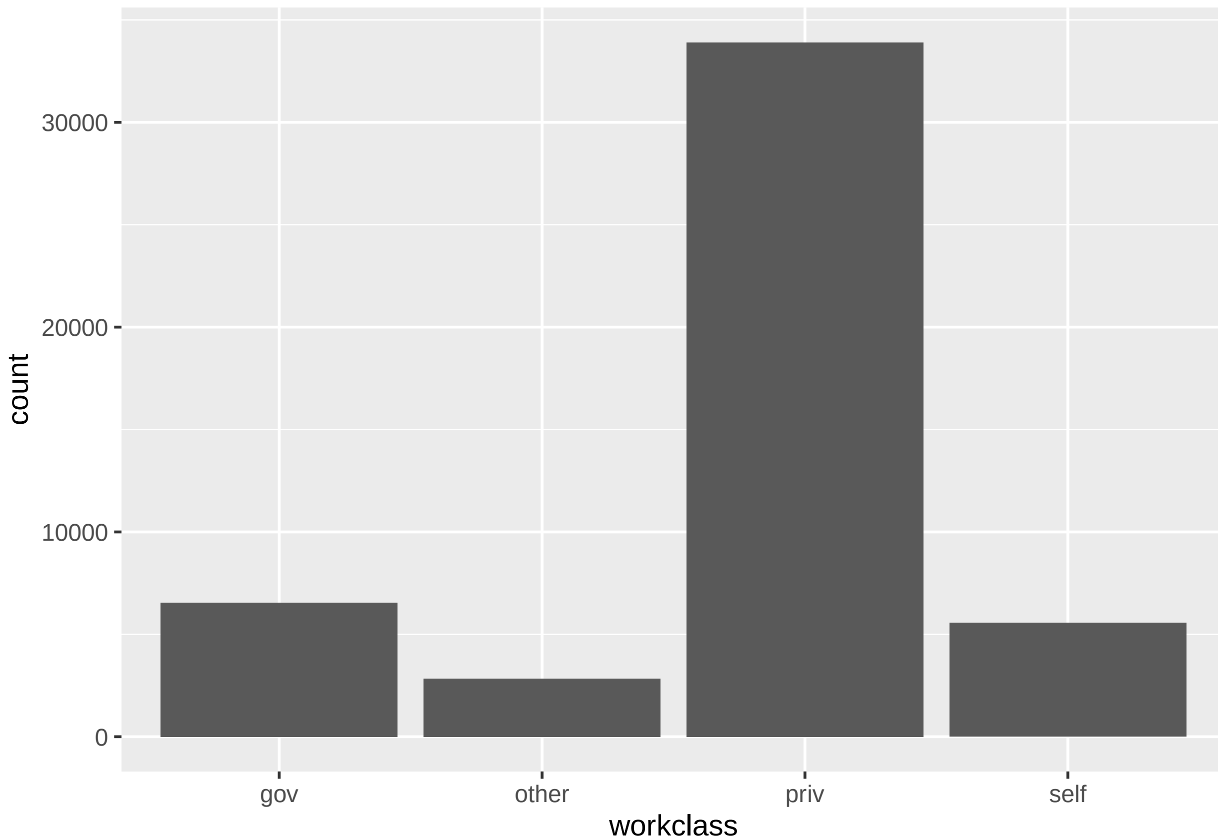
```r
adult$workclass[grepl("(\\?|without|never)",
                     trimws(adult$workclass),ignore.case = T)] <- 'other'
adult$workclass[grepl("priv",trimws(adult$workclass),ignore.case = T)] <- 'priv'

#Verificar missing values
adult[is.na(adult$workclass)==TRUE,'workclass']
```

```
## character(0)
```

```r
# diferentes etiquetas -- FIN
levels(as.factor(adult$workclass))
```

```
## [1] "gov"   "other" "priv"  "self"
```

```r
# Distribución final
ggplot(data=adult,aes(x=workclass)) + geom_bar()
```

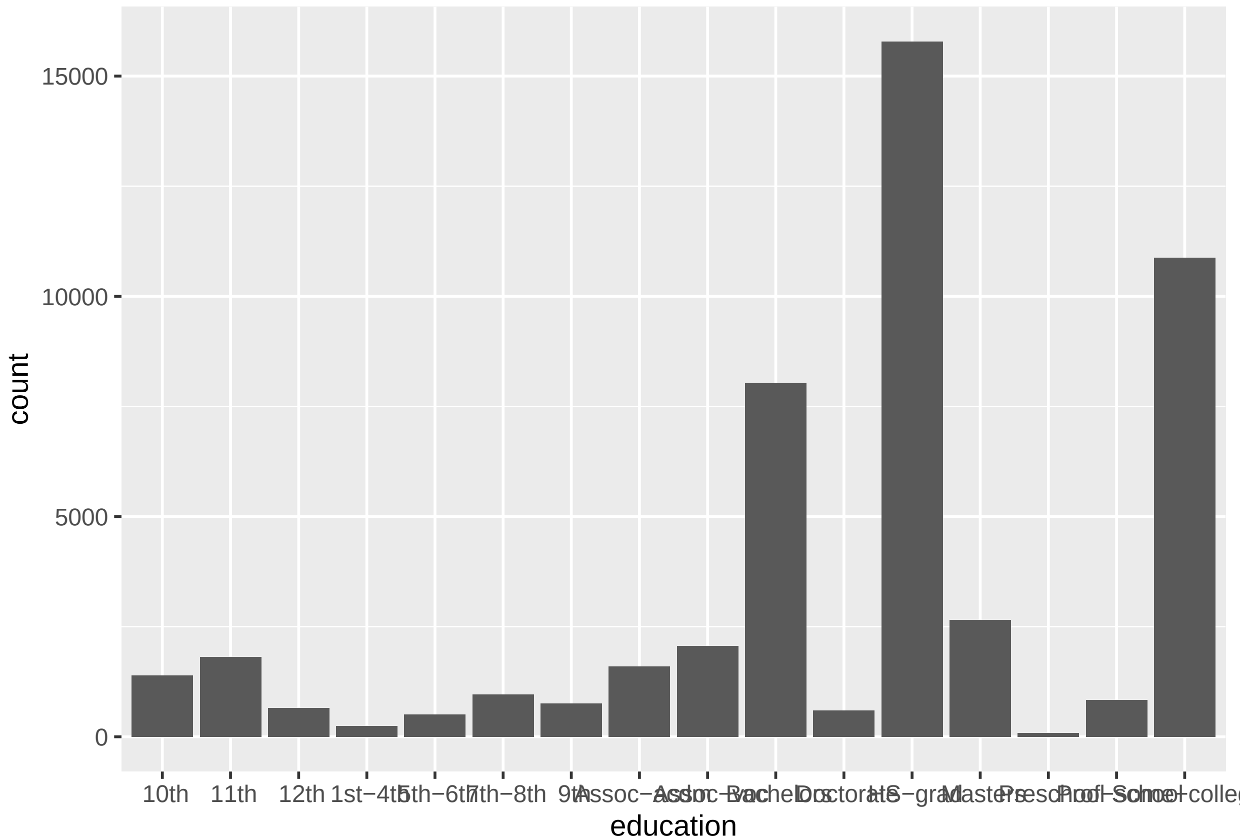


- Recategorización education

```r
# --------------- education ------------------
# diferentes etiquetas INCIO
levels(as.factor(adult$education))
```

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

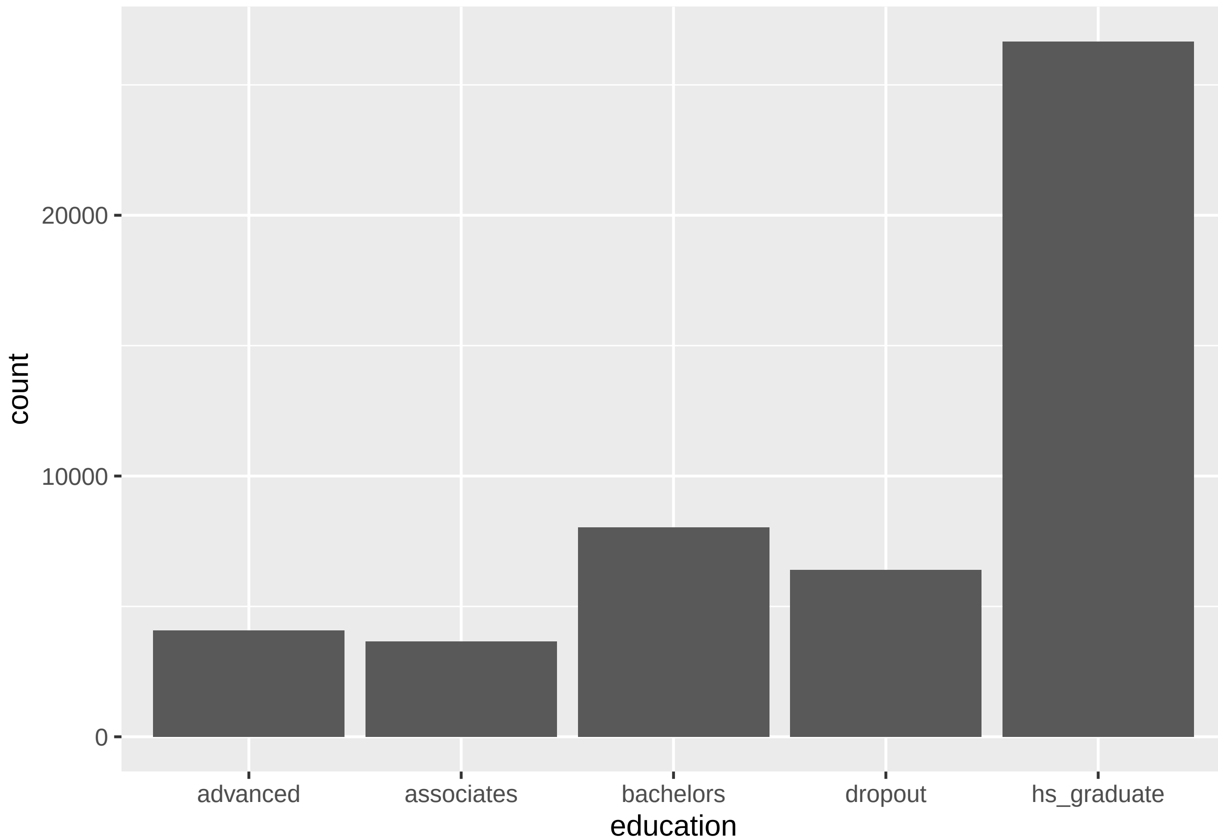
```r
# Distribución inicial
ggplot(data=adult,aes(x=education)) + geom_bar()
```




```r
# Estos valores se pueden agrupar en 5 categorias: no terminado, asociados, high school, bachelor y avar
adult$education[grepl("(th|preschool)",trimws(adult$education),ignore.case = T)] <- 'dropout'
adult$education[grepl("assoc",trimws(adult$education),ignore.case = T)] <- 'associates'
adult$education[grepl("(hs-|college)",trimws(adult$education),ignore.case = T)] <- 'hs_graduate'
adult$education[grepl("(prof|master|docto)",trimws(adult$education),ignore.case = T)] <- 'advanced'
# minúsculas, sin espacios
adult$education <- trimws(tolower(adult$education))

# diferentes etiquetas FINAL
levels(as.factor(adult$education))
```

```
## [1] "advanced"    "associates"  "bachelors"   "dropout"     "hs_graduate"
```

```r
# Distribución final
ggplot(data=adult,aes(x=education)) + geom_bar()
```

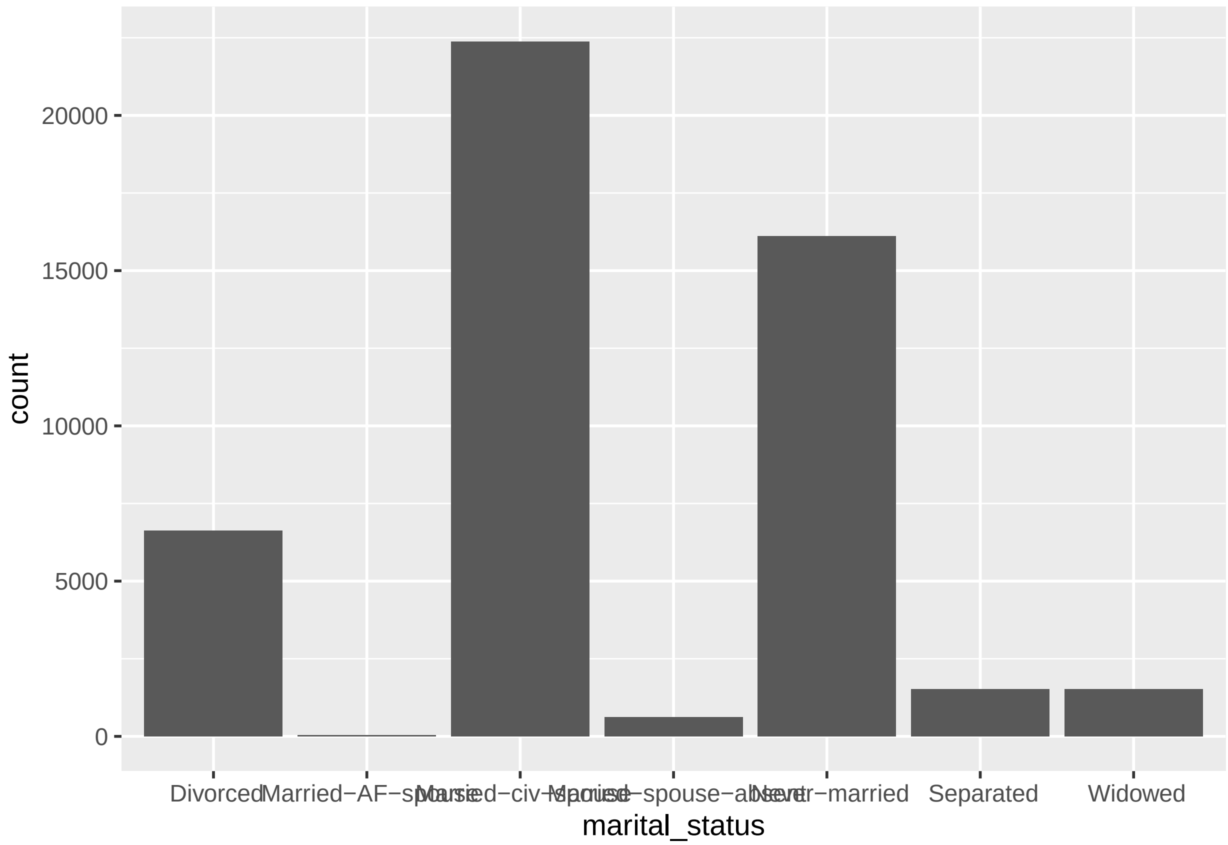


- Recategorización marital status

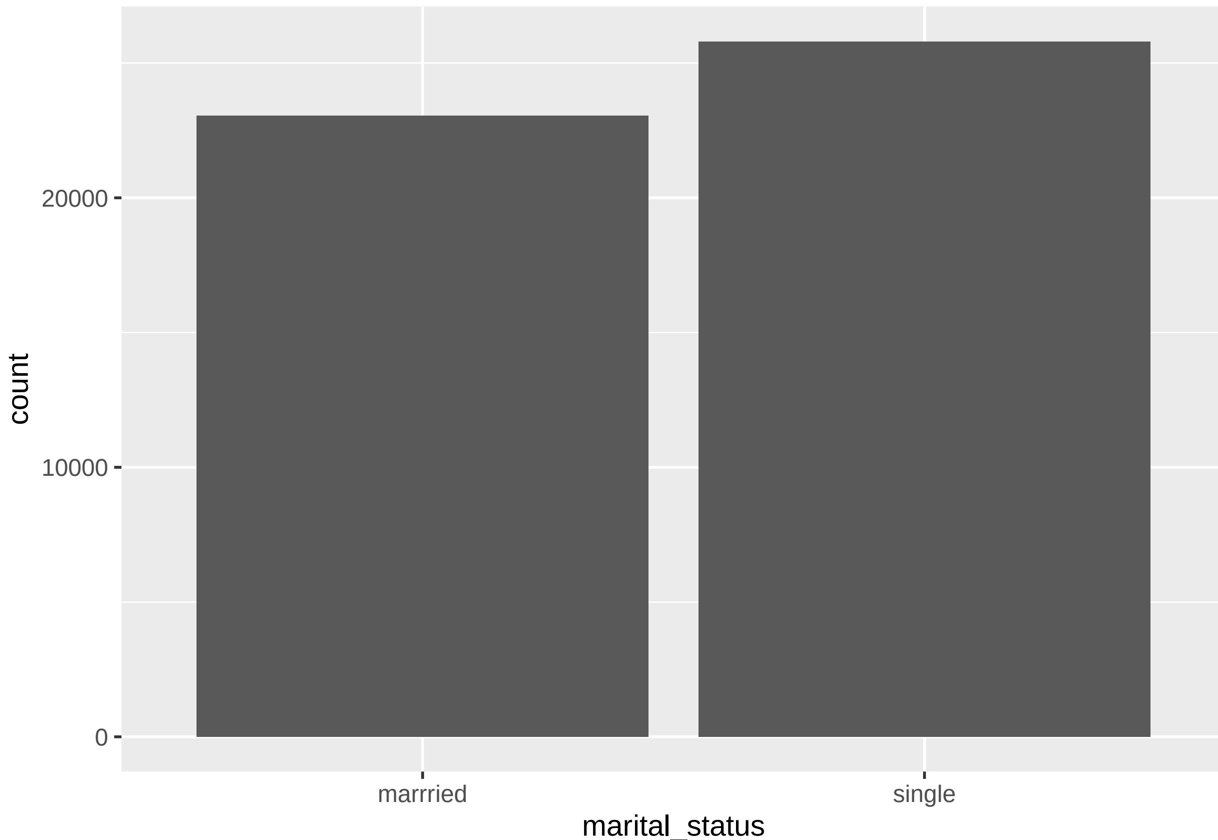
```
# --------------- marital estatus ------------------
# diferentes etiquetas
levels(as.factor(adult$marital_status))
```

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

```
# Distribución inicial
ggplot(data=adult,aes(x=marital_status)) + geom_bar()
```




```
# Estos valores se pueden agrupar en 3 categorias: casados, sin pareja y solteros
adult$marital_status[grepl("^married",trimws(adult$marital_status),
                           ignore.case = T)] <- 'marrried'
adult$marital_status[grepl("(separa|divorced|wido|never)",trimws(adult$marital_status),
                           ignore.case = T)] <- 'single'

# Distribución final
ggplot(data=adult,aes(x=marital_status)) + geom_bar()
```

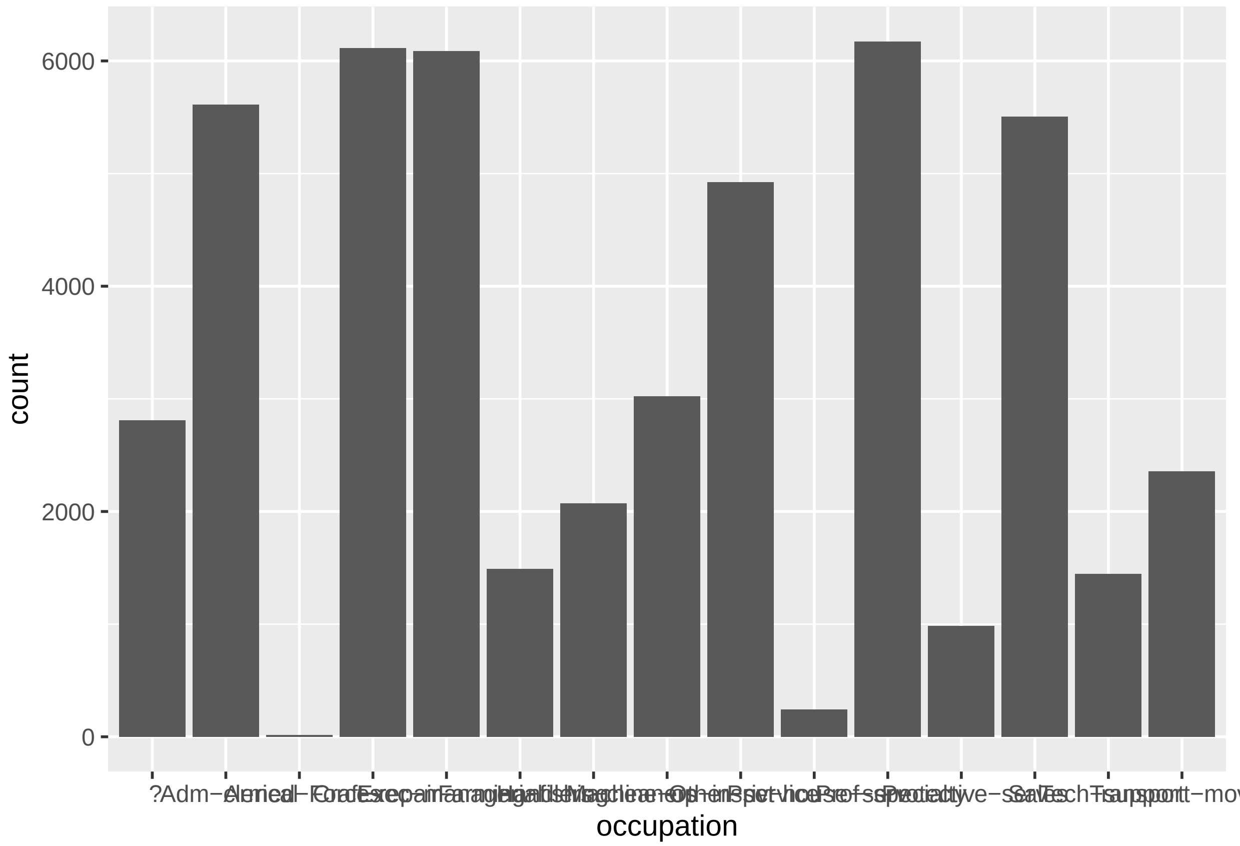


- Recategorización occupation

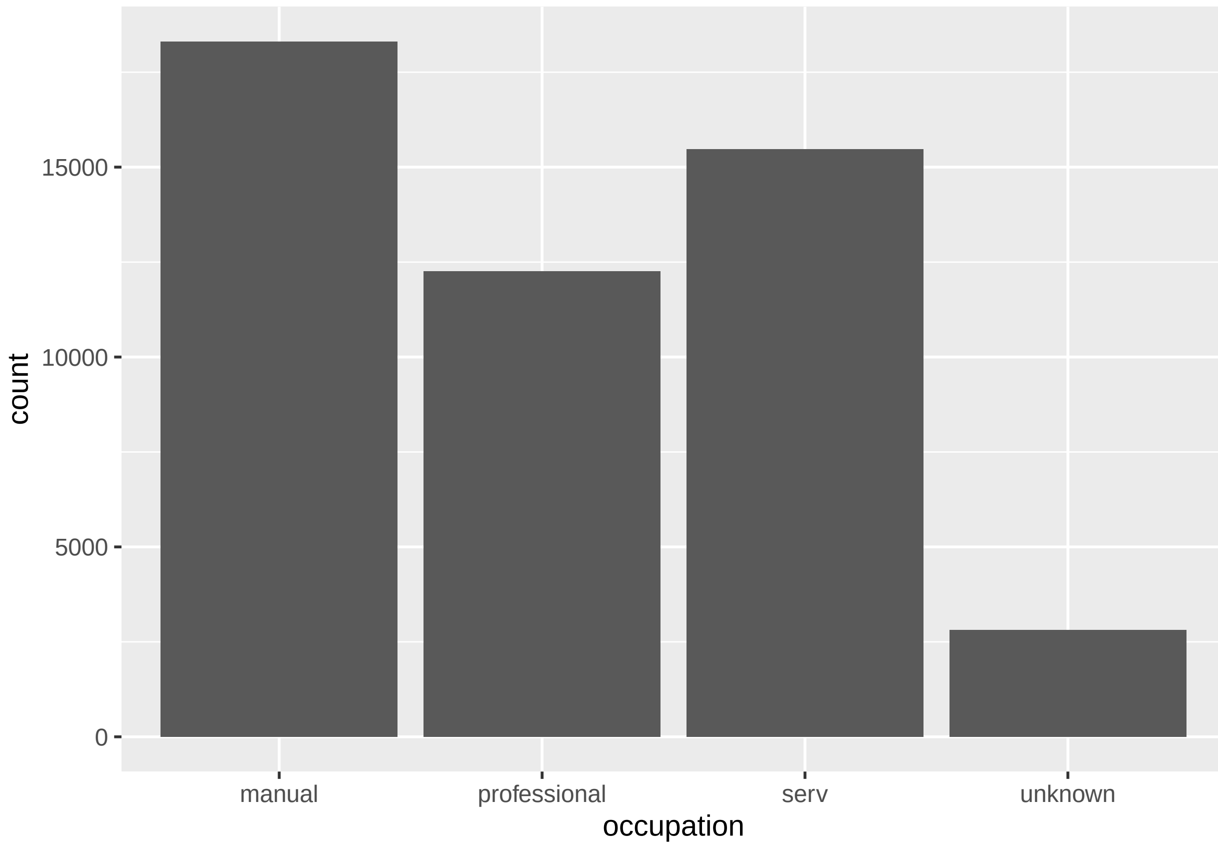
```
# --------------- occupation ------------------
# diferentes etiquetas
levels(as.factor(adult$occupation))
```

```
##  [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"
```

```
# Distribución inicial
ggplot(data=adult,aes(x=occupation)) + geom_bar()
```




```
# Estos valores se pueden agrupar en 6 categorias: servicios, otros servicios, profesionales, ecelsiást
adult$occupation[grepl("(transp|tech|protect|priv|other|armed|sales)",
                        trimws(adult$occupation), ignore.case = T)] <- 'serv'
adult$occupation[grepl("(exec|prof)",trimws(adult$occupation),
                        ignore.case = T)] <- 'professional'
adult$occupation[grepl("(craft|farm|hand|inspct|cleric)",trimws(adult$occupation),
                        ignore.case = T)] <- 'manual'
adult$occupation[grepl("\\?",trimws(adult$occupation),ignore.case = T)] <- 'unknown'
# cambiar guión medio por guión bajo
adult$occupation <- gsub("-", "_", adult$occupation)
# minúsculas, sin espacios
adult$occupation <- trimws(tolower(adult$occupation))

# Distribución final
ggplot(data=adult,aes(x=occupation)) + geom_bar()
```

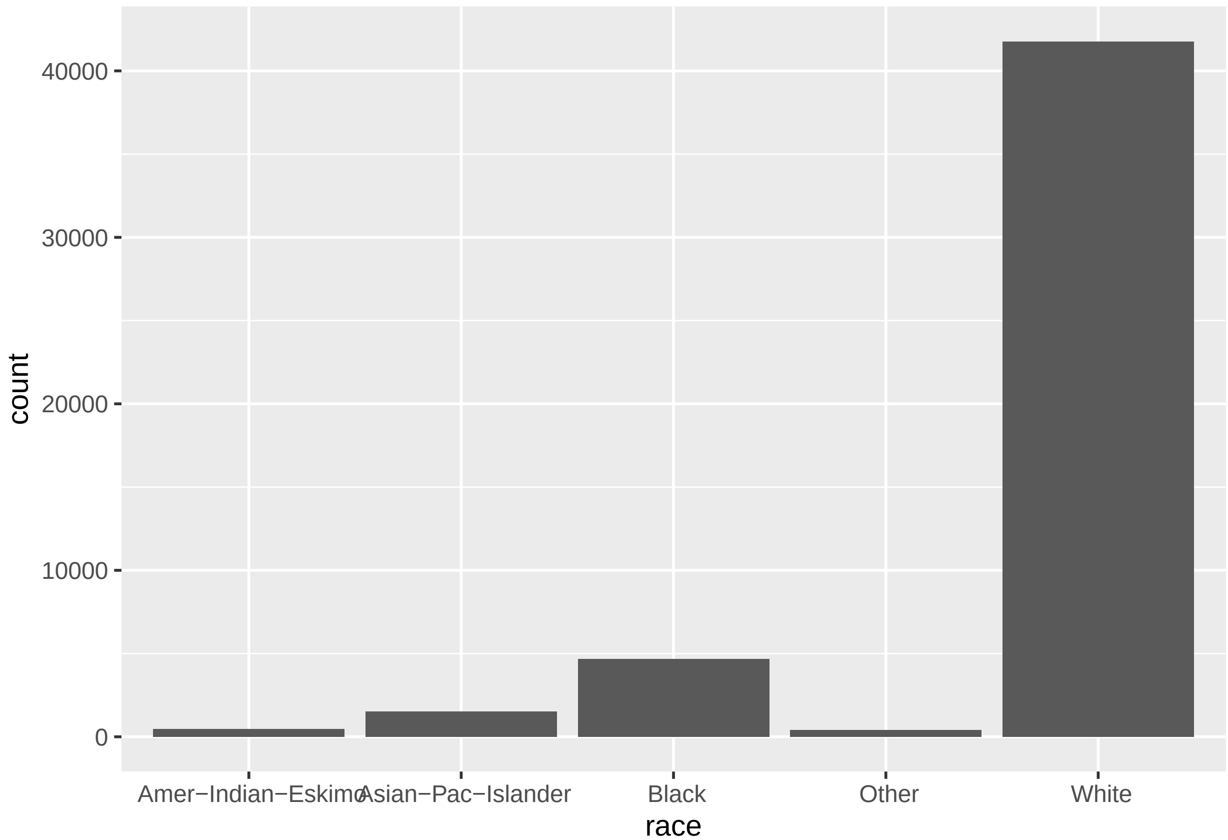


- Recategorización race

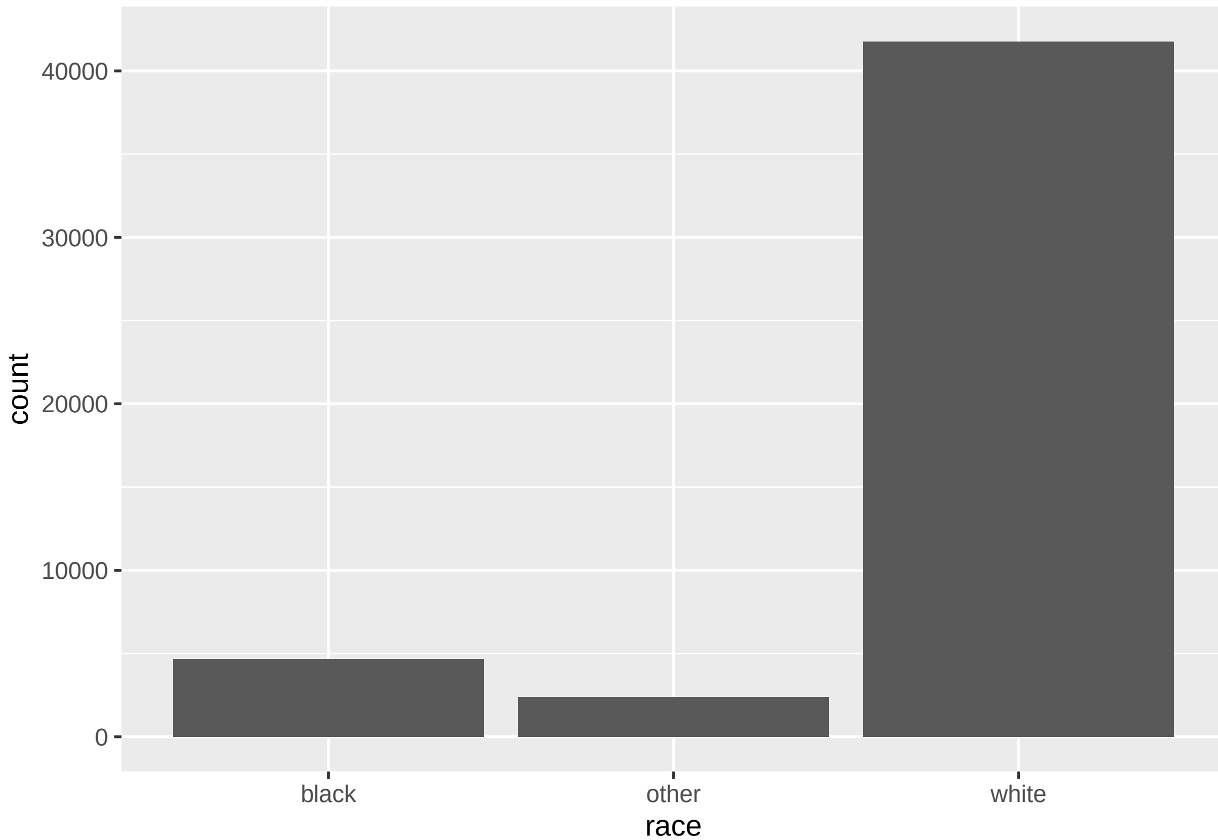
```
# --------------- race ------------------
# diferentes etiquetas
levels(as.factor(adult$race))
```

```
## [1] " Amer-Indian-Eskimo" " Asian-Pac-Islander" " Black"
## [4] " Other"              " White"
```

```
# Distribución inicial
ggplot(data=adult,aes(x=race)) + geom_bar()
```




```r
# Se agrupan otras razas diferentes a white y black en una categoría
adult$race[!grepl("(white|black)",trimws(adult$race),ignore.case = T)] <- 'other'
# minúsculas, sin espacios
adult$race <- trimws(tolower(adult$race))

# Distribución final
ggplot(data=adult,aes(x=race)) + geom_bar()
```

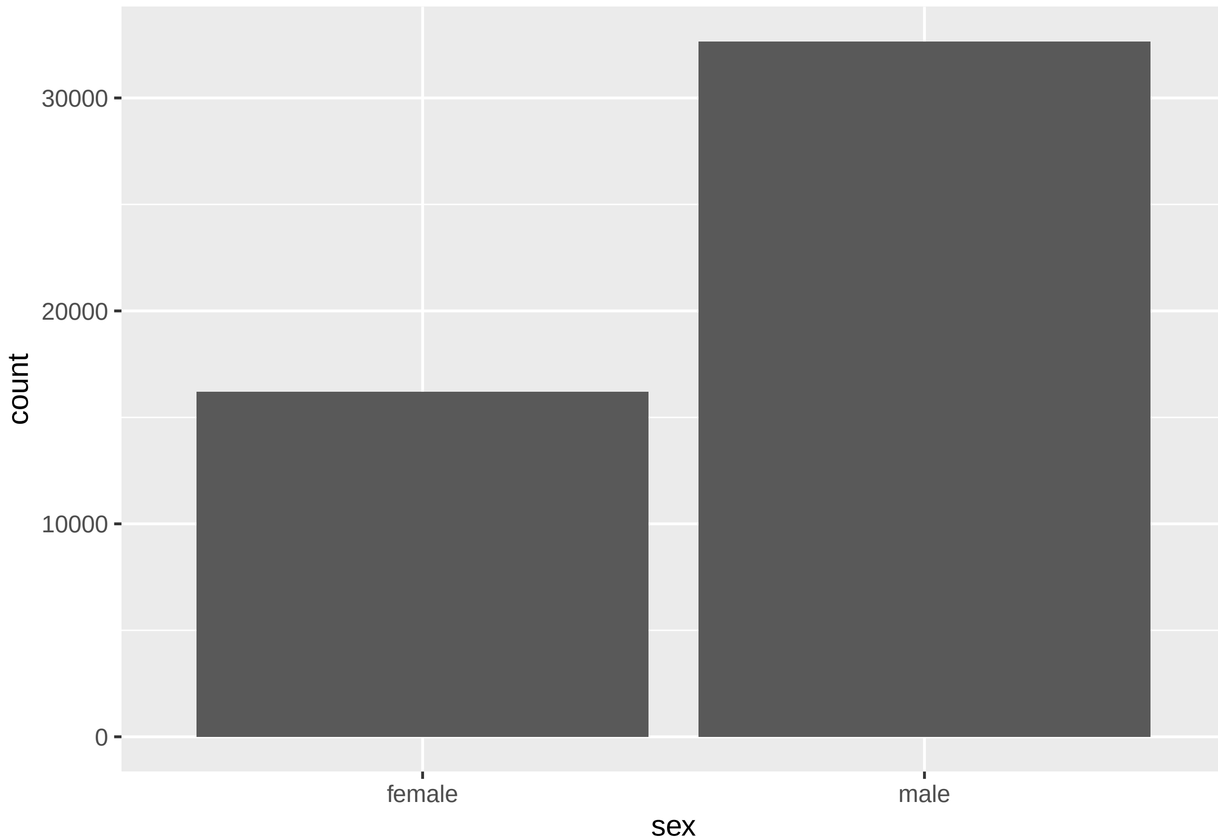


- Sex

```
# --------------- sex ------------------
# diferentes etiquetas
levels(as.factor(adult$sex))
```

```
## [1] " Female" " Male"
```

```
# minúsculas, sin espacios
adult$sex <- trimws(tolower(adult$sex))

# Distribución
ggplot(data=adult,aes(x=sex,)) + geom_bar()
```

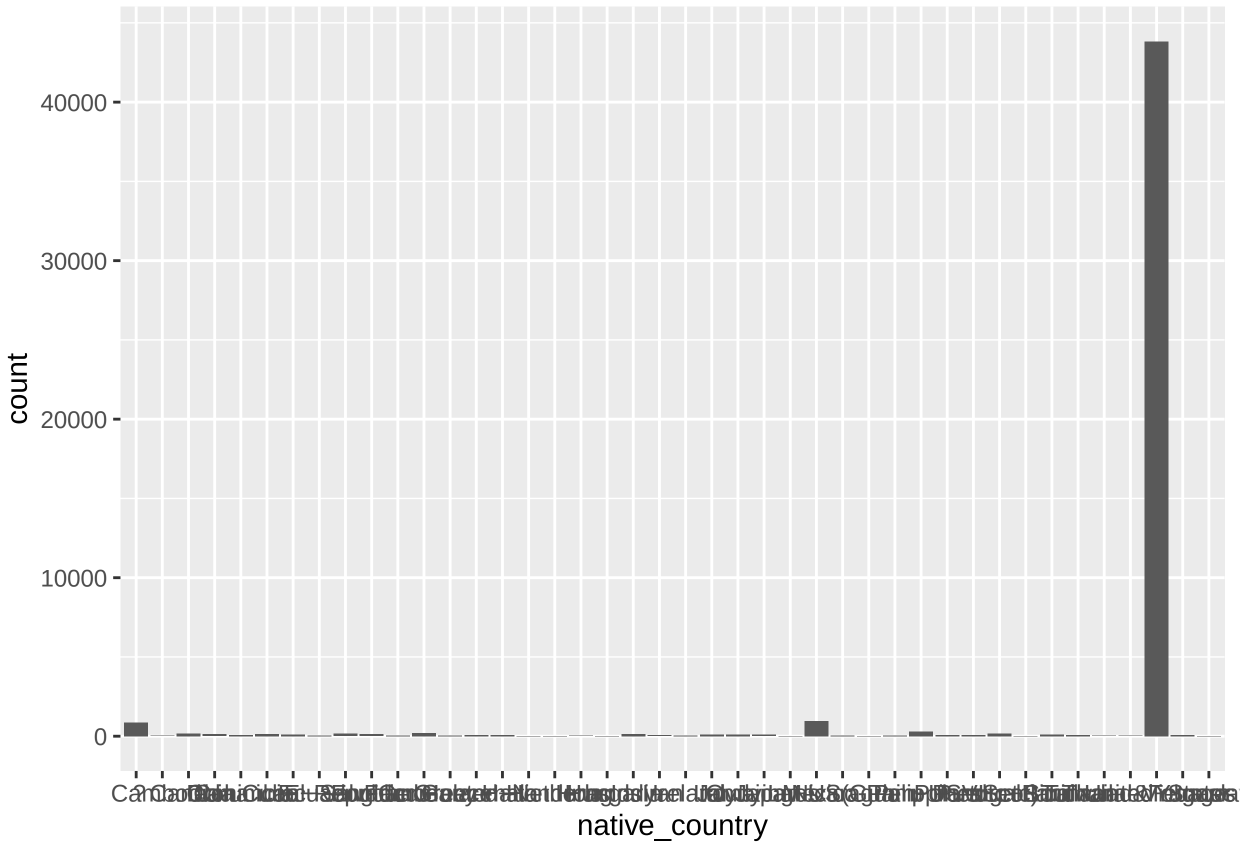


- Recategorización native country

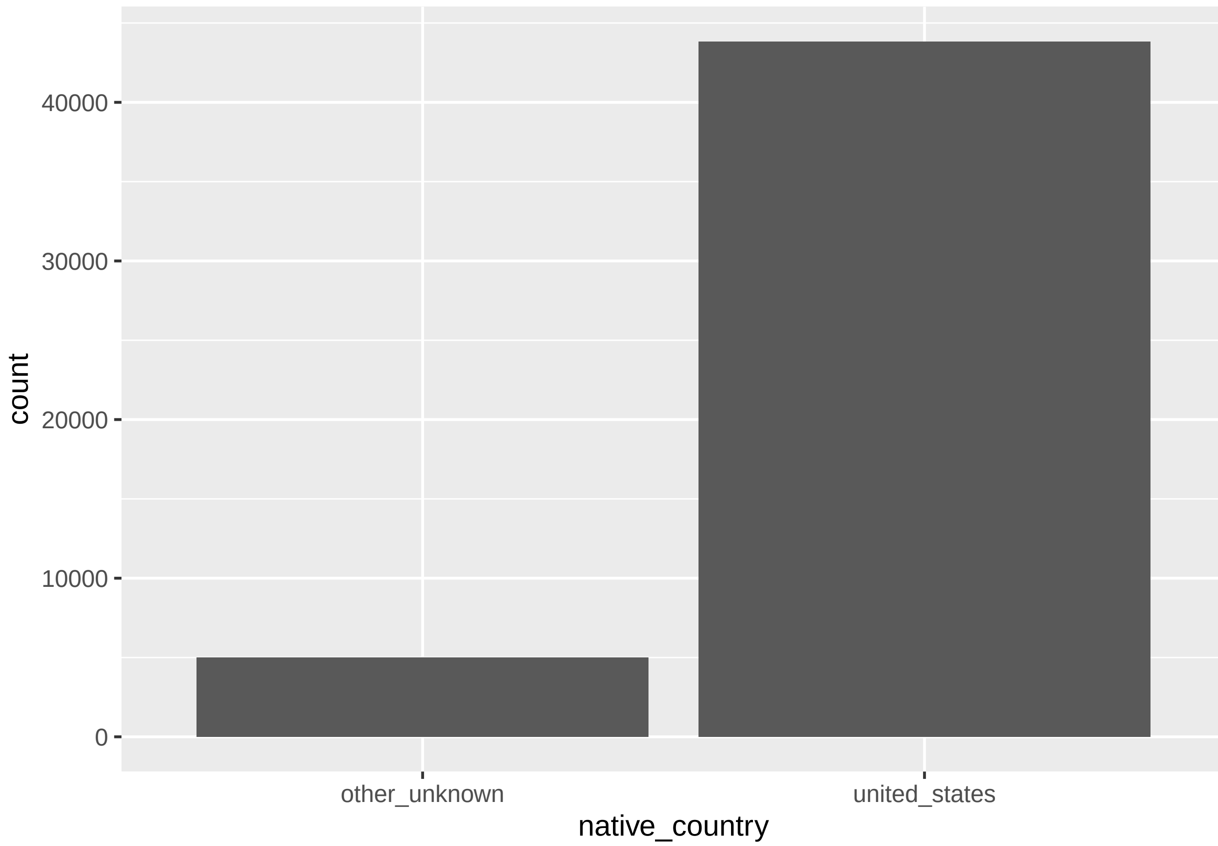
```r
# --------------- native country ------------------
# diferentes etiquetas
levels(as.factor(adult$native_country))
```

```
##  [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"
```

```r
# Distribución inicial
ggplot(data=adult,aes(x=native_country)) + geom_bar()
```

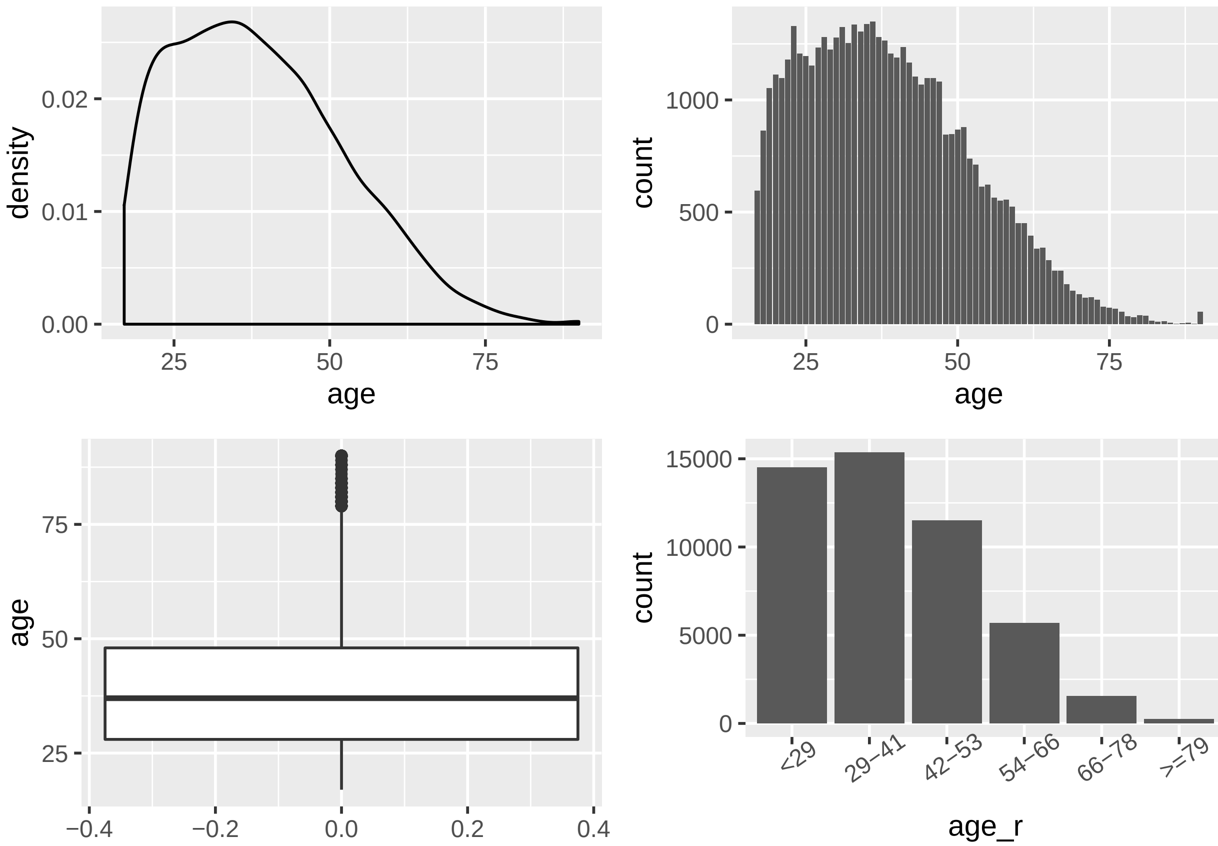



```r
# se agrupan países diferentes a USA en una categoría
adult$native_country[!grepl("(united)",trimws(adult$native_country),
                            ignore.case = T)] <- 'other_unknown'
# guión bajo en lugar de medio y minúsculas, sin espacios
adult$native_country <-trimws(tolower(gsub("-", "_", adult$native_country)))

# Distribución final
ggplot(data=adult,aes(x=native_country)) + geom_bar()
```



- Categorización edad

```
# --------------- age ------------------
# Se agrupa en rangos de edad
library(arules)
adult$age_r <- discretize(adult$age, method = "interval", breaks = 6,
                          labels = c("<29", "29-41", "42-53","54-66","66-78",">=79"))

# Distribución
plot1 <- ggplot(data=adult, aes(x=age)) +
  geom_density(adjust=1.5, alpha=.4)
plot2 <- ggplot(data = adult,aes(x=age)) + geom_bar()
plot3 <- ggplot(data=adult, aes(y=age)) +
  geom_boxplot() + scale_fill_brewer(palette="Set3")
plot4 <- ggplot(data = adult,aes(x=age_r)) + geom_bar() +
  scale_fill_brewer(palette="Set3") +
  theme(axis.text.x = element_text(angle = 35))
grid.arrange(plot1, plot2, plot3, plot4, ncol=2)
```

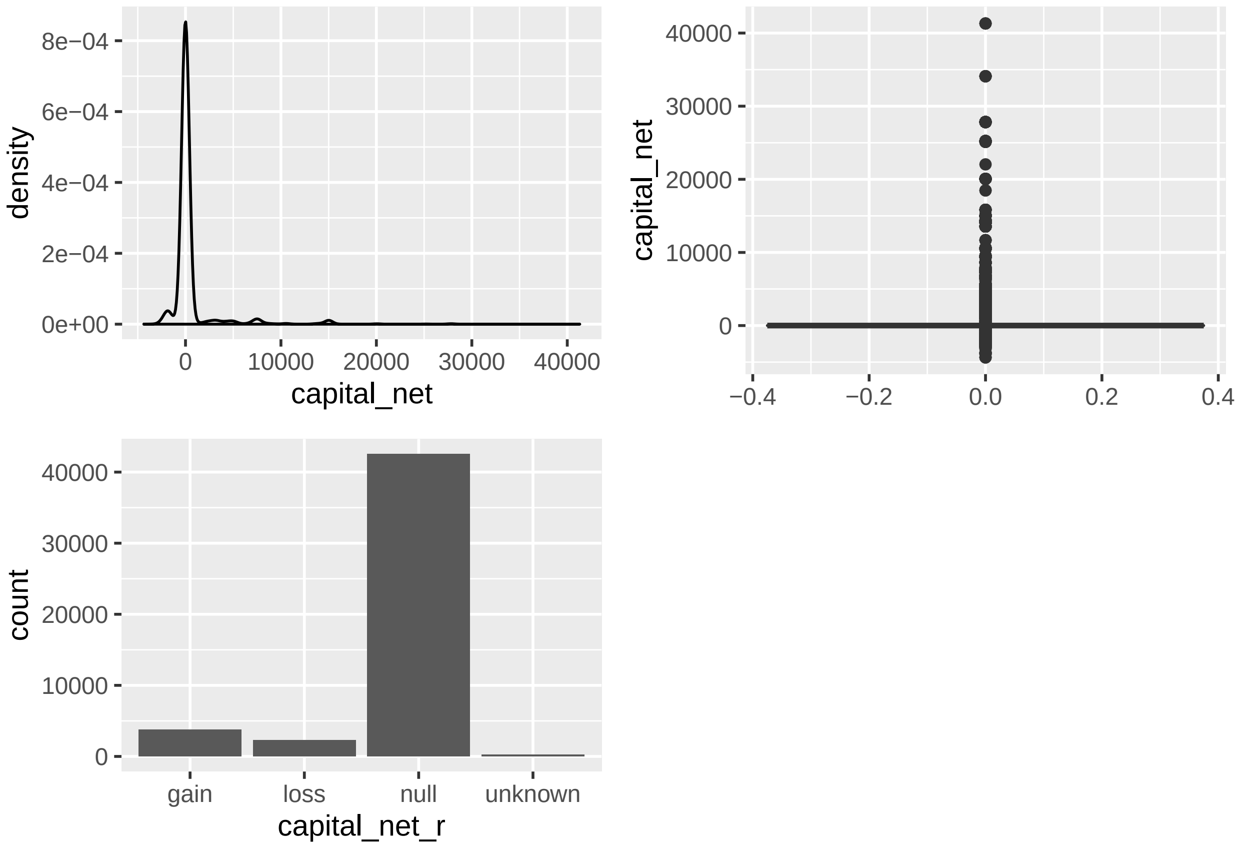


- Para el capital de los censados, se crea una nueva variable para obtener el valor neto (ganancias - pérdidas), este último se categoriza.

```
# se establece una nueva variable que combine capital loss y capital_gain
adult$capital_net <- adult$capital_gain-adult$capital_loss

# resumen estadístico de capital_net
describe(adult$capital_net)
```

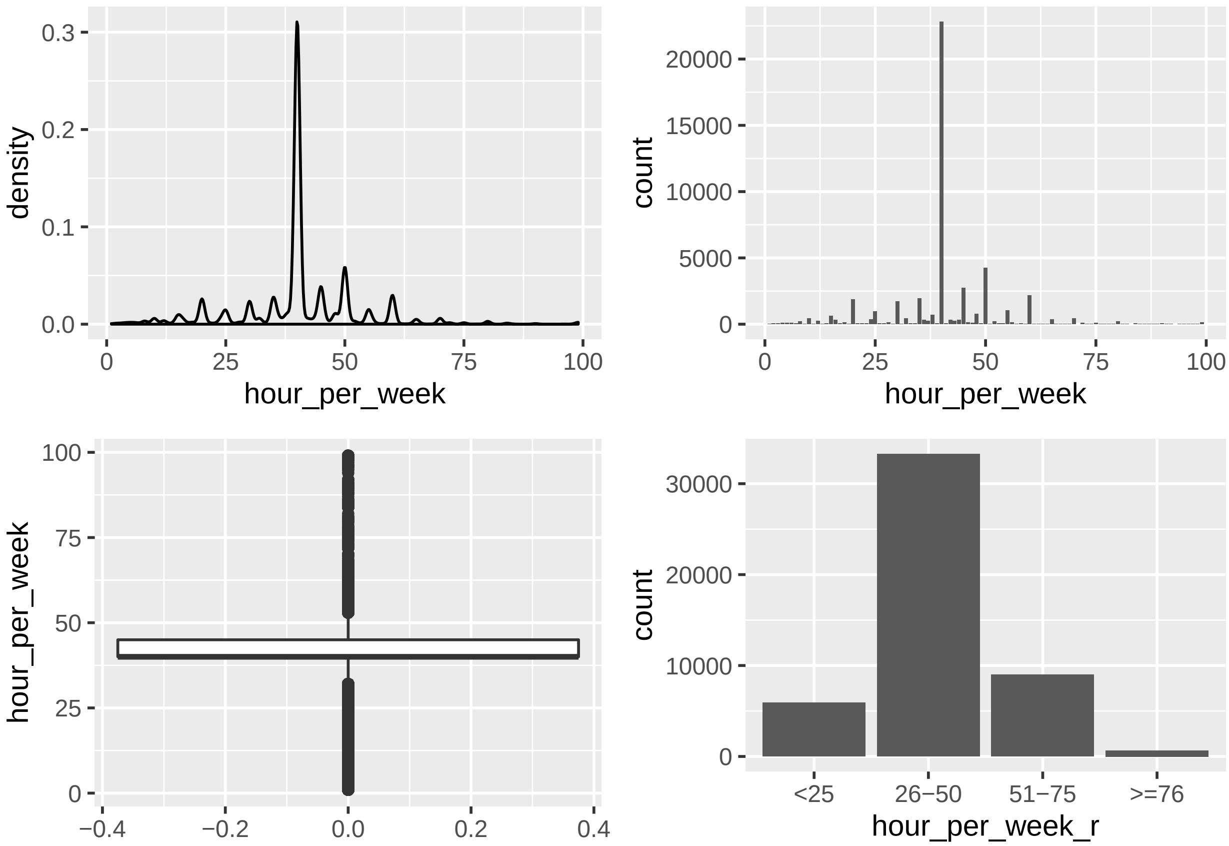
```
##    vars     n   mean      sd median trimmed mad   min   max range skew kurtosis
## X1    1 48598 494.47 2588.48      0       0   0 -4356 41310 45666 5.63    40.69
##       se
## X1 11.74
```

```
# Se categoriza esta nueva variable
adult$capital_net_r <- cut(adult$capital_net, c(-900000000,-0.0001,0.0001,9000000000),
                                labels = c("loss", "null", "gain"), ordered=TRUE)
adult$capital_net_r = as.character(adult$capital_net_r)
adult$capital_net_r[is.na(adult$capital_net)] <- "unknown"

# Distribución final en relación a variable target
p1 <- ggplot(data=adult, aes(x=capital_net)) +
  geom_density(adjust=1.5, alpha=.4)
p3 <- ggplot(data=adult, aes(y=capital_net)) +
  geom_boxplot()
p4 <- ggplot(data = adult,aes(x=capital_net_r)) +
  geom_bar()
grid.arrange(p1, p3, p4, ncol=2)
```



- Categorización horas trabajadas por semana

```
# --------------- hour per week ------------------
# categorización de la variable con base en lo anterior
adult$hour_per_week_r <- discretize(adult$hour_per_week,
                                    method = "interval", breaks = 4,
                                    labels = c("<25","26-50","51-75",">=76"))

# Distribución final en relación a variable target
p1 <- ggplot(data= adult, aes(x=hour_per_week)) +
  geom_density(adjust=1.5, alpha=.4)
p2 <- ggplot(data = adult,aes(x=hour_per_week)) + geom_bar()
p3 <- ggplot(data= adult, aes(y=hour_per_week,)) +
  geom_boxplot()
p4 <- ggplot(data = adult,aes(x=hour_per_week_r)) +
  geom_bar()
grid.arrange(p1, p2, p3, p4, ncol=2)
```



Análisis de los datos. ## Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar). Se utilizarán las variables categóricas que se han formulado y/o recategrizado, con estas variables se estudiará el poder predictivo que tengan, además de la correlación entre variables categóricas. Adicionalmente se observará visualmente el comportamiento de cada una de las variables con la variable target.

Comprobación de la normalidad y homogeneidad de la varianza.

Para este caso en particular, las variables numericas que existian se transofrmaton en categóricas por lo que las pruebas de normalidad y homogeneidad carecen de sentido.

Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.

Por medio de las correlaciones, no se observa relación significativa entre occupation, workclass, race.

```
cat <- c("workclass","education","marital_status","occupation" ,"race","sex","native_country","target",

# Posibles combinaciones de variables categóricas
combinaciones     <- as.data.frame(t(combn(cat, 2)))

# Prueba Chi-Square pata categóricas
chisq_prueba <-
  apply(t(combinaciones), 2, function(x){
                           c(Var1   = x[1],
                             Var2   = x[2],
```

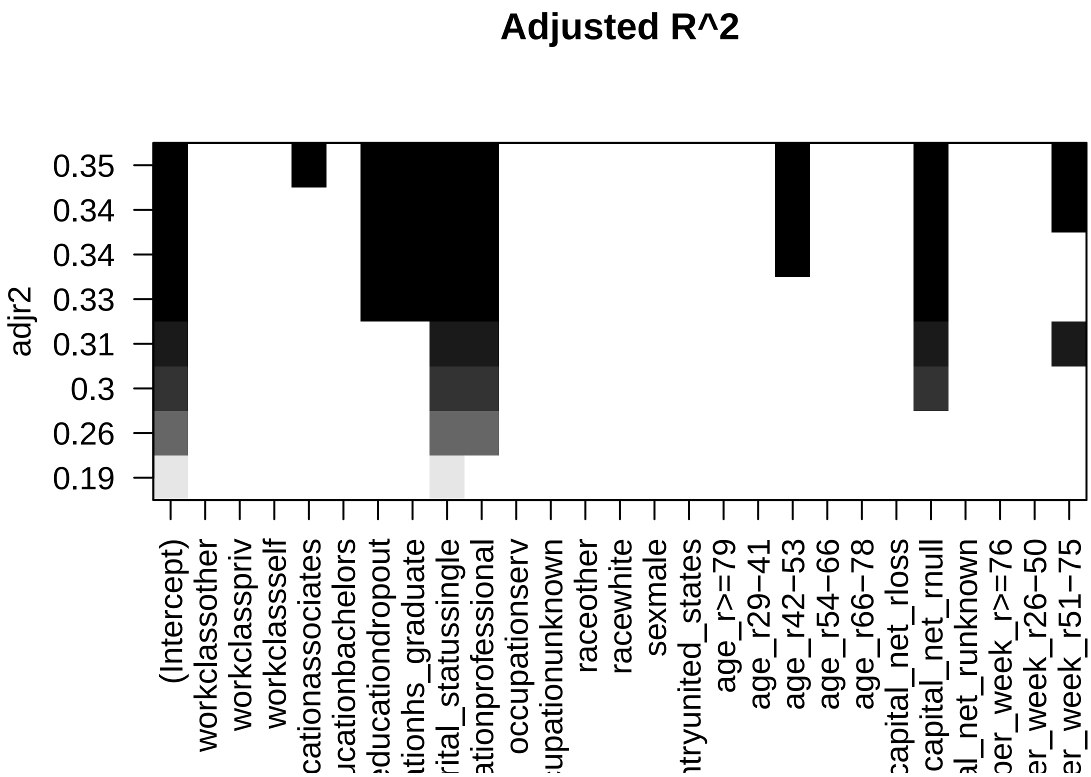
```r
                              pvalue = chisq.test(adult[, x[1]],
                                                  adult[, x[2]])$p.value)
  })

chisq_prueba <- chisq_prueba %>%
  t() %>%
  data.frame(stringsAsFactors=FALSE)

chisq_prueba$pvalue <- as.numeric(chisq_prueba$pvalue)

chisq_prueba<- chisq_prueba%>%
  arrange(pvalue)

knitr::kable(chisq_prueba, caption = "Tabla de correlaciones --categóricas-- ",
             format = "markdown", padding = 0,
             col.names = c("", "", "p-value"))
```

		p-value
workclass	education	0.0000000
workclass	occupation	0.0000000
workclass	age_r	0.0000000
workclass	hour_per_week_r	0.0000000
education	occupation	0.0000000
education	target	0.0000000
education	age_r	0.0000000
education	capital_net_r	0.0000000
education	hour_per_week_r	0.0000000
marital_status	sex	0.0000000
marital_status	target	0.0000000
marital_status	age_r	0.0000000
marital_status	hour_per_week_r	0.0000000
occupation	target	0.0000000
occupation	age_r	0.0000000
occupation	hour_per_week_r	0.0000000
race	native_country	0.0000000
sex	target	0.0000000
sex	hour_per_week_r	0.0000000
target	age_r	0.0000000
target	capital_net_r	0.0000000
target	hour_per_week_r	0.0000000
age_r	hour_per_week_r	0.0000000
workclass	marital_status	0.0000000
education	native_country	0.0000000
workclass	sex	0.0000000
workclass	target	0.0000000
marital_status	capital_net_r	0.0000000
occupation	capital_net_r	0.0000000
age_r	capital_net_r	0.0000000
marital_status	occupation	0.0000000
marital_status	race	0.0000000
education	marital_status	0.0000000
race	sex	0.0000000
capital_net_r	hour_per_week_r	0.0000000

		p-value
sex	age_r	0.0000000
workclass	race	0.0000000
occupation	sex	0.0000000
race	target	0.0000000
race	hour_per_week_r	0.0000000
education	race	0.0000000
workclass	capital_net_r	0.0000000
sex	capital_net_r	0.0000000
occupation	race	0.0000000
education	sex	0.0000000
workclass	native_country	0.0000000
race	age_r	0.0000000
marital_status	native_country	0.0000000
race	capital_net_r	0.0000000
native_country	target	0.0000000
native_country	hour_per_week_r	0.0000000
native_country	age_r	0.0000051
occupation	native_country	0.0001612
native_country	capital_net_r	0.0002307
sex	native_country	0.0141993

Dentro de la prueba para identificar el valor predictivo, se observa que education, marital_status, age, capital neto y hours per week son las variables elegidas.

```
# modelo de búsqueda
adult[cat] = as.data.frame(apply(adult[cat],2,
                                          function(x) as.factor(x)))

search_output<-regsubsets(target~workclass+education+marital_status+occupation+race+sex+native_country+
                          data=adult, method="exhaustive")
plot(search_output, scale = "adjr2", main = "Adjusted R^2")
```




```r
cols_1 <- c("education","marital_status","age_r","hour_per_week_r",
            "capital_net_r","target")
cols_2 <- c("education","marital_status","age_r","hour_per_week_r",
            "capital_net_r")
```

Con el objetivo de tener los datos finales solicitados, se exporta este conjunto de datos.

```r
# Export data a csv
write.csv(adult, "adult_data.csv")
```

Se utilizan los datos finales para elaborar un árbol de decisión que nos ayude en el predicción del ingreso.

```r
# train data set
train <- adult[adult["df"]=="train",cols_1]
trainX <- train[cols_2]
trainy <- train$target

# test data set
test <- adult[adult["df"]=="test",cols_1]
testX <- test[cols_2]
testy <- test$target

#C5.o model
model <- C50::C5.0(trainX, trainy,rules=TRUE )
summary(model)
```

```
##
## Call:
## C5.0.default(x = trainX, y = trainy, rules = TRUE)
##
##
```

```
## C5.0 [Release 2.07 GPL Edition]      Tue Jan  7 01:20:45 2020
## -------------------------------
##
## Class specified by attribute `outcome'
##
## Read 32561 cases (6 attributes) from undefined.data
##
## Rules:
##
## Rule 1: (8972/383, lift 1.3)
##  education in {associates, dropout, hs_graduate}
##  age_r in {<29, >=79, 66-78}
##  ->  class <=50k  [0.957]
##
## Rule 2: (9982/485, lift 1.3)
##  age_r in {<29, >=79, 66-78}
##  capital_net_r = null
##  ->  class <=50k  [0.951]
##
## Rule 3: (3633/181, lift 1.3)
##  hour_per_week_r = <25
##  capital_net_r = null
##  ->  class <=50k  [0.950]
##
## Rule 4: (4253/244, lift 1.2)
##  education = dropout
##  ->  class <=50k  [0.942]
##
## Rule 5: (17144/1105, lift 1.2)
##  marital_status = single
##  ->  class <=50k  [0.935]
##
## Rule 6: (21999/2859, lift 1.1)
##  education in {associates, dropout, hs_graduate}
##  capital_net_r = null
##  ->  class <=50k  [0.870]
##
## Rule 7: (159, lift 4.1)
##  capital_net_r = unknown
##  ->  class >50k  [0.994]
##
## Rule 8: (1239/116, lift 3.8)
##  education in {advanced, bachelors}
##  marital_status = marrried
##  capital_net_r in {gain, loss, unknown}
##  ->  class >50k  [0.906]
##
## Rule 9: (1033/192, lift 3.4)
##  education in {advanced, associates, bachelors}
##  age_r in {29-41, 42-53, 54-66}
##  capital_net_r = gain
##  ->  class >50k  [0.814]
##
## Rule 10: (1229/285, lift 3.2)
```

```
##  education in {advanced, associates, bachelors}
##  capital_net_r = gain
##  ->  class >50k  [0.768]
##
## Rule 11: (2306/580, lift 3.1)
##  marital_status = marrried
##  age_r in {29-41, 42-53, 54-66}
##  capital_net_r in {gain, loss}
##  ->  class >50k  [0.748]
##
## Rule 12: (3848/971, lift 3.1)
##  education in {advanced, bachelors}
##  marital_status = marrried
##  age_r in {29-41, 42-53, 54-66}
##  hour_per_week_r in {>=76, 26-50, 51-75}
##  ->  class >50k  [0.748]
##
## Default class: <=50k
##
##
## Evaluation on training data (32561 cases):
##
##          Rules
##    ----------------
##      No      Errors
##
##      12 5365(16.5%)   <<
##
##
##     (a)   (b)    <-classified as
##    ----  ----
##   23222  1498    (a): class <=50k
##    3867  3974    (b): class >50k
##
##
##  Attribute usage:
##
##   83.90% education
##   82.56% capital_net_r
##   68.86% marital_status
##   49.31% age_r
##   22.98% hour_per_week_r
##
##
## Time: 0.1 secs
```

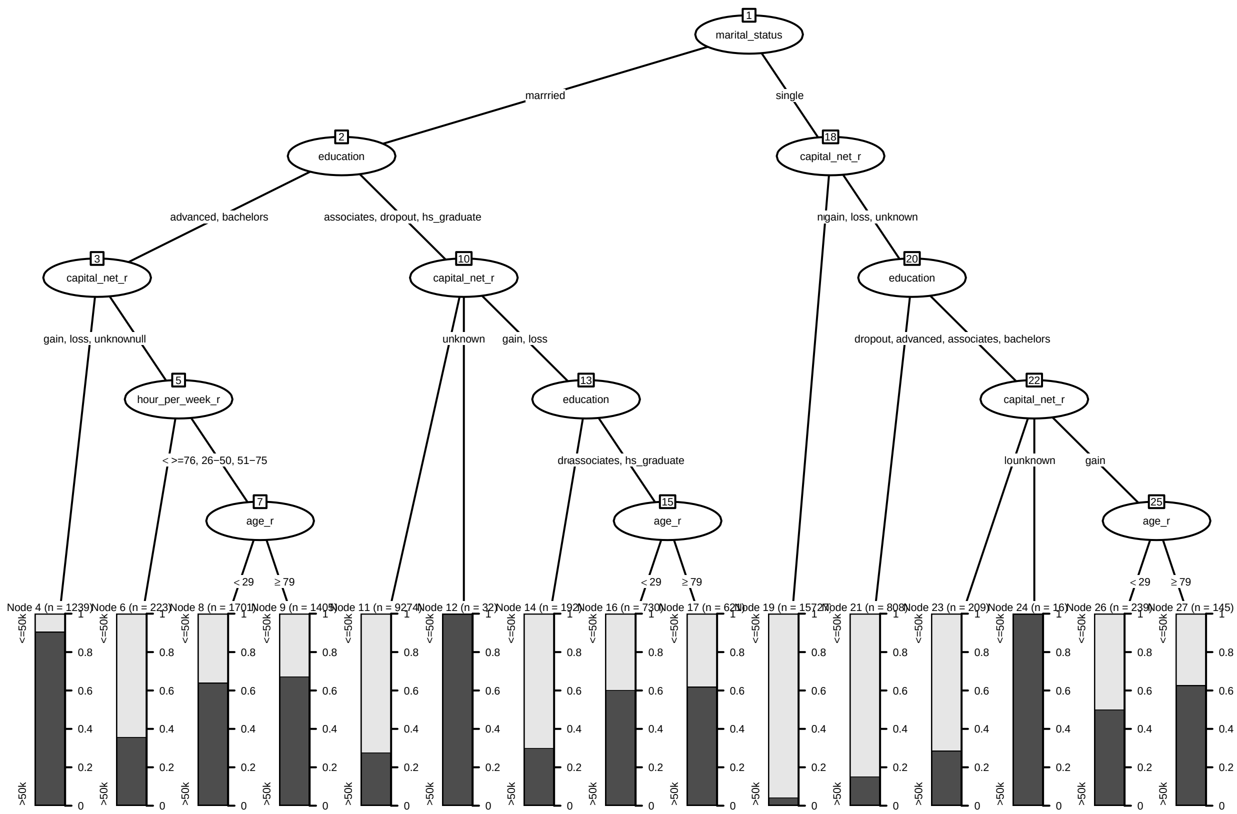
```
model_t <- C50::C5.0(trainX, trainy)
summary(model_t)
```

```
##
## Call:
## C5.0.default(x = trainX, y = trainy)
##
##
## C5.0 [Release 2.07 GPL Edition]      Tue Jan  7 01:20:46 2020
```

```
## -------------------------------
##
## Class specified by attribute `outcome'
##
## Read 32561 cases (6 attributes) from undefined.data
##
## Decision tree:
##
## marital_status = marrried:
## :...education in {advanced,bachelors}:
## :   :...capital_net_r in {gain,loss,unknown}: >50k (1239/116)
## :   :   capital_net_r = null:
## :   :   :...hour_per_week_r = <25: <=50k (223/80)
## :   :       hour_per_week_r in {>=76,26-50,51-75}:
## :   :       :...age_r in {<29,>=79,66-78}: <=50k (348/162)
## :   :           age_r in {29-41,42-53,54-66}: >50k (2758/883)
## :   education in {associates,dropout,hs_graduate}:
## :   :...capital_net_r = null: <=50k (9274/2580)
## :       capital_net_r = unknown: >50k (32)
## :       capital_net_r in {gain,loss}:
## :       :...education = dropout: <=50k (192/58)
## :           education in {associates,hs_graduate}:
## :           :...age_r in {<29,>=79,66-78}: <=50k (206/79)
## :               age_r in {29-41,42-53,54-66}: >50k (1145/398)
## marital_status = single:
## :...capital_net_r = null: <=50k (15727/694)
##     capital_net_r in {gain,loss,unknown}:
##     :...education in {dropout,hs_graduate}: <=50k (808/124)
##         education in {advanced,associates,bachelors}:
##         :...capital_net_r = loss: <=50k (209/60)
##             capital_net_r = unknown: >50k (16)
##             capital_net_r = gain:
##             :...age_r in {<29,>=79,66-78}: <=50k (107/35)
##                 age_r in {29-41,42-53,54-66}: >50k (277/101)
##
##
## Evaluation on training data (32561 cases):
##
##      Decision Tree
##    ----------------
##    Size      Errors
##
##      15 5370(16.5%)   <<
##
##
##     (a)   (b)    <-classified as
##    ----  ----
##   23222  1498    (a): class <=50k
##    3872  3969    (b): class >50k
##
##
##  Attribute usage:
##
##  100.00% marital_status
```

```
##  100.00% capital_net_r
##   51.70% education
##   14.87% age_r
##   10.22% hour_per_week_r
##
##
## Time: 0.0 secs
```

```r
plot(model_t, gp = gpar(fontsize = 4))
```




```r
model_t_pred<-predict(model_t,newdata =testX,type="class")
confusionMatrix(model_t_pred,testy)
```

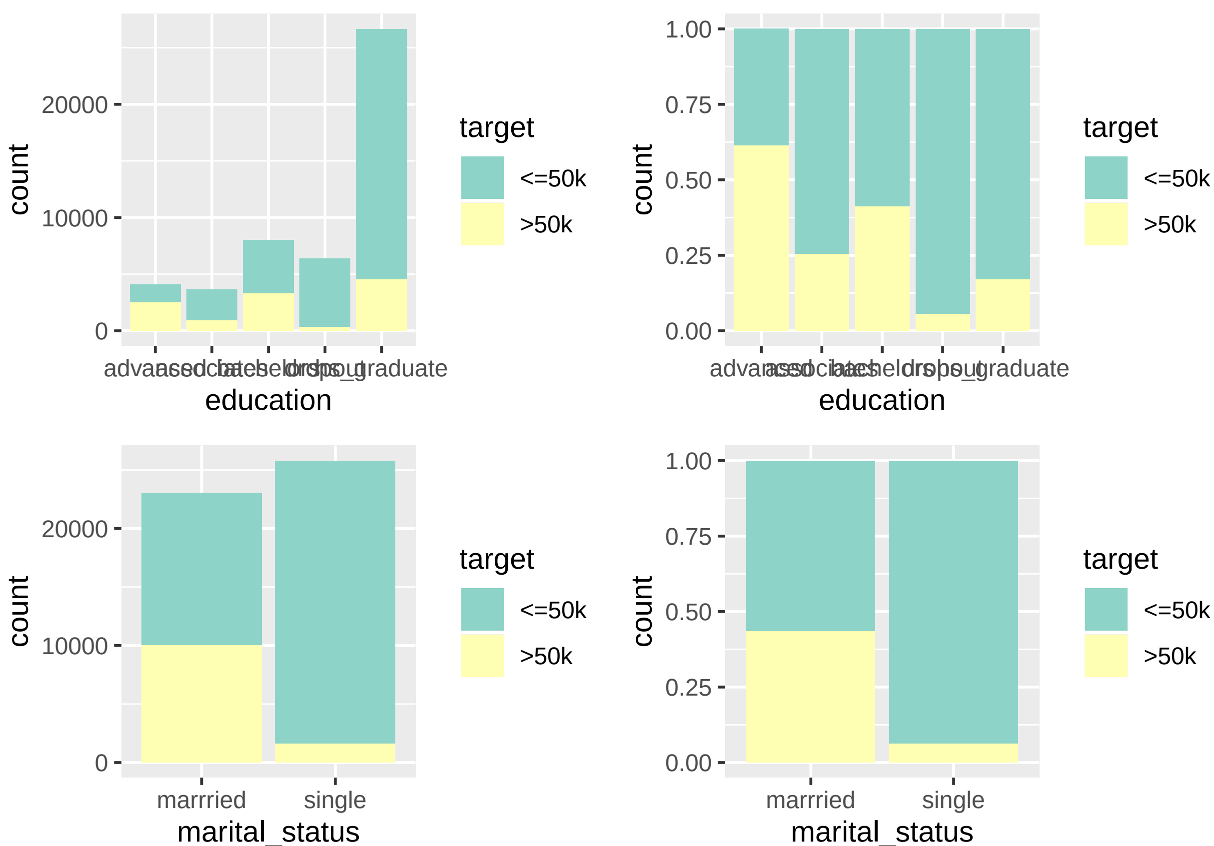
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction <=50k  >50k
##      <=50k 11714  1902
##      >50k    721  1944
##
##                Accuracy : 0.8389
##                  95% CI : (0.8332, 0.8445)
##     No Information Rate : 0.7638
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5006
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
```

```
##             Sensitivity : 0.9420
##             Specificity : 0.5055
##          Pos Pred Value : 0.8603
##          Neg Pred Value : 0.7295
##              Prevalence : 0.7638
##          Detection Rate : 0.7195
##    Detection Prevalence : 0.8363
##       Balanced Accuracy : 0.7237
##
##        'Positive' Class : <=50k
##
```

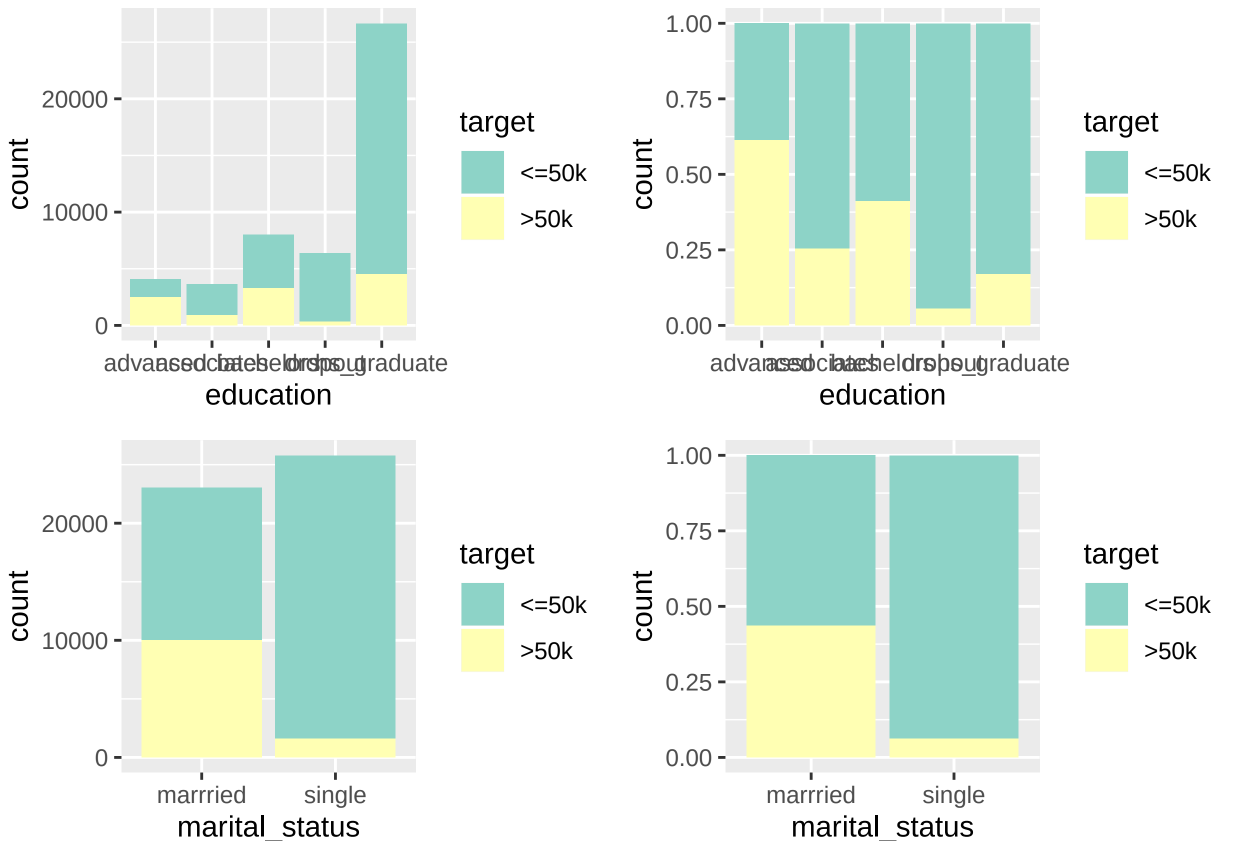
Representación de los resultados a partir de tablas y gráficas.

Transformación y limpieza variables categóricas.

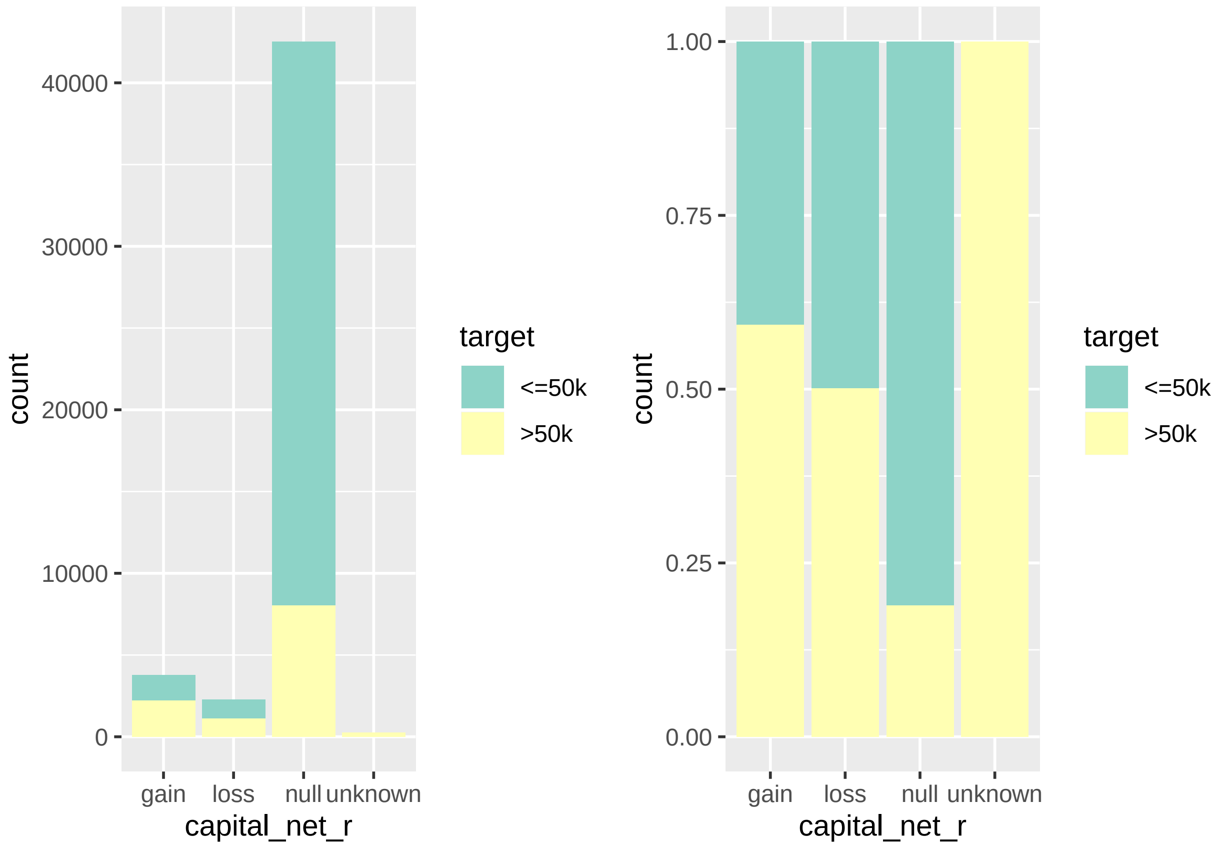
```r
# Distribución final en relación a variable target
p1 = ggplot(data=adult,aes(x=education,fill=target)) + geom_bar() +
  scale_fill_brewer(palette="Set3")
p2 = ggplot(data = adult,aes(x=education,fill=target)) + geom_bar(position="fill") +
  scale_fill_brewer(palette="Set3")
p3 = ggplot(data=adult,aes(x=marital_status,fill=target)) + geom_bar() +
  scale_fill_brewer(palette="Set3")
p4 = ggplot(data = adult,aes(x=marital_status,fill=target)) +
  geom_bar(position="fill") +
  scale_fill_brewer(palette="Set3")
grid.arrange(p1,p2,p3,p4, ncol =2)
```



```r
p5 = ggplot(data=adult,aes(x=age_r,fill=target)) + geom_bar() +
  scale_fill_brewer(palette="Set3")
p6 = ggplot(data = adult,aes(x=age_r,fill=target)) + geom_bar(position="fill") +
  scale_fill_brewer(palette="Set3")
p7 = ggplot(data=adult,aes(x=hour_per_week_r,fill=target)) + geom_bar() +
  scale_fill_brewer(palette="Set3")
p8 = ggplot(data = adult,aes(x=hour_per_week_r,fill=target)) +
  geom_bar(position="fill") +
  scale_fill_brewer(palette="Set3")
grid.arrange(p1,p2,p3,p4, ncol =2)
```




```r
p9 = ggplot(data=adult,aes(x=capital_net_r,fill=target)) + geom_bar() +
  scale_fill_brewer(palette="Set3")
p10 = ggplot(data = adult,aes(x=capital_net_r,fill=target)) +
  geom_bar(position="fill") +
  scale_fill_brewer(palette="Set3")
grid.arrange(p9,p10, ncol =2)
```



Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?

Con base en el análisis previo y el modelo ejectado, se concluye que los factores que intervienen en el ingreso de los individuo son: el nivel educativo, horas trabajadas, estatus civil y edad. Estas variables ofrecen el mejor rendimiento predictivo dentro de las que ofrece este dataset.

Código: Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos. Si lo preferís, también podéis trabajar en Python.

El código se encuentra integrado en el informe.

Este trabajo se realizó de manera individual, por esta razón se omite la tabla solicitada ETNG