Tipología y ciclo de vida de los datos - Práctica 2

Mike Findlay

7 Enero 2019

Contents

1 Descripción del Dataset

El conjunto que utilizamos es el "Adult dataset" que esta disponible desde Kaggle y del UCI Machine Learning Repository. Consiste de aproximademente 32000 observaciones, y 15 variables.

El objetivo es ver lo bien que podemos predecir si los ingresos anuales (income) de una persona superior a $50000 utilizando el conjunto de variables en este conjunto de datos.

Aquí esta la descripción de las variables:

- age – The age of the individual
- workclass – The type of employer the individual has. Whether they are government, military, private, and so on.
- fnlwgt – The # of people the census takers believe that observation represents. We will be ignoring this variable
- education – The highest level of education achieved for that individual
- education.num – Highest level of education in numerical form

- marital.status – Marital status of the individual
- occupation – The occupation of the individual
- relationship – Contains family relationship values like husband, father, and so on, but only contains one per observation.
- race – descriptions of the individuals race. Black, White, etc
- sex – Male or Female
- capital.gain – Capital gains recorded
- capital.loss – Capital Losses recorded
- hours.per.week – Hours worked per week
- native.country – Country of origin for person
- income – Boolean Variable. Whether or not the person makes more than $50,000 per annum income.

2 Integración y selección de los datos de interés a analizar

2.1 Primer contacto con el juego de datos, visualizamos su estructura.

```
# Cargamos los paquetes R que vamos a usar
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(knitr)
library(nortest)
```

```
## Warning: package 'nortest' was built under R version 3.5.2
```

```
# Cargamos el juego de datos
datosAdult <- read.csv('adult.csv',stringsAsFactors = TRUE, header = TRUE)

# Nombres de los atributos
#names(datosAdult) <- c("age","workclass","fnlwgt","education","education-num","marital-status","occupa
```

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

```
## 'data.frame':    32561 obs. of  15 variables:
##  $ age           : int  90 82 66 54 41 34 38 74 68 41 ...
##  $ workclass     : Factor w/ 9 levels "?","Federal-gov",..: 1 5 1 5 5 5 5 8 2 5 ...
##  $ fnlwgt        : int  77053 132870 186061 140359 264663 216864 150601 88638 422013 70037 ...
##  $ education     : Factor w/ 16 levels "10th","11th",..: 12 12 16 6 16 12 1 11 12 16 ...
##  $ education.num : int  9 9 10 4 10 9 6 16 9 10 ...
##  $ marital.status: Factor w/ 7 levels "Divorced","Married-AF-spouse",..: 7 7 7 1 6 1 6 5 1 5 ...
##  $ occupation    : Factor w/ 15 levels "?","Adm-clerical",..: 1 5 1 8 11 9 2 11 11 4 ...
##  $ relationship  : Factor w/ 6 levels "Husband","Not-in-family",..: 2 2 5 5 4 5 5 3 2 5 ...
##  $ race          : Factor w/ 5 levels "Amer-Indian-Eskimo",..: 5 5 3 5 5 5 5 5 5 5 ...
```

```
##  $ sex           : Factor w/ 2 levels "Female","Male": 1 1 1 1 1 1 2 1 1 2 ...
##  $ capital.gain  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ capital.loss  : int  4356 4356 4356 3900 3900 3770 3770 3683 3683 3004 ...
##  $ hours.per.week: int  40 18 40 40 40 45 40 20 40 60 ...
##  $ native.country: Factor w/ 42 levels "?","Cambodia",..: 40 40 40 40 40 40 40 40 40 1 ...
##  $ income        : Factor w/ 2 levels "<=50K",">50K": 1 1 1 1 1 1 1 2 1 2 ...
```

Tenemos 32561 observaciones en 15 variables

```r
kable(head(datosAdult))
```

age	workclass	fnlwgt	education	education.num	marital.status	occupation	relationship	race	s
90	?	77053	HS-grad	9	Widowed	?	Not-in-family	White	F
82	Private	132870	HS-grad	9	Widowed	Exec-managerial	Not-in-family	White	F
66	?	186061	Some-college	10	Widowed	?	Unmarried	Black	F
54	Private	140359	7th-8th	4	Divorced	Machine-op-inspct	Unmarried	White	F
41	Private	264663	Some-college	10	Separated	Prof-specialty	Own-child	White	F
34	Private	216864	HS-grad	9	Divorced	Other-service	Unmarried	White	F

3 Limpieza de los datos

3.1 Trabajamos en los atributos con valores vacíos

```r
# Estadísticas de valores vacíos
colSums(is.na(datosAdult))
```

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

```r
colSums(datosAdult=="")
```

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

Parece que no existen valores vacios. Sin embargo, en el fichero adult.names que describe los datos dice:

Conversion of original data as follows:

1. Discretized agrossincome into two ranges with threshold 50,000.
2. Convert U.S. to US to avoid periods.
3. Convert Unknown to "?"
4. Run MLC++ GenCVFiles to generate data,test.

O sea que han cambiado los Unknowns (vacios) a "?".

Reimportamos los datos para poner cambiar los "?" a valores vacios (NA)s.

```r
# utilizamos na.strings para definir los caracteres de NA.
datosAdult <- read.csv('adult.csv',stringsAsFactors = TRUE, header = TRUE, na.strings="?")
```

Ahora comprobamos los valores vacios de nuevo.

```r
# Estadísticas de 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 hours.per.week native.country         income
##              0              0              0            583              0
```

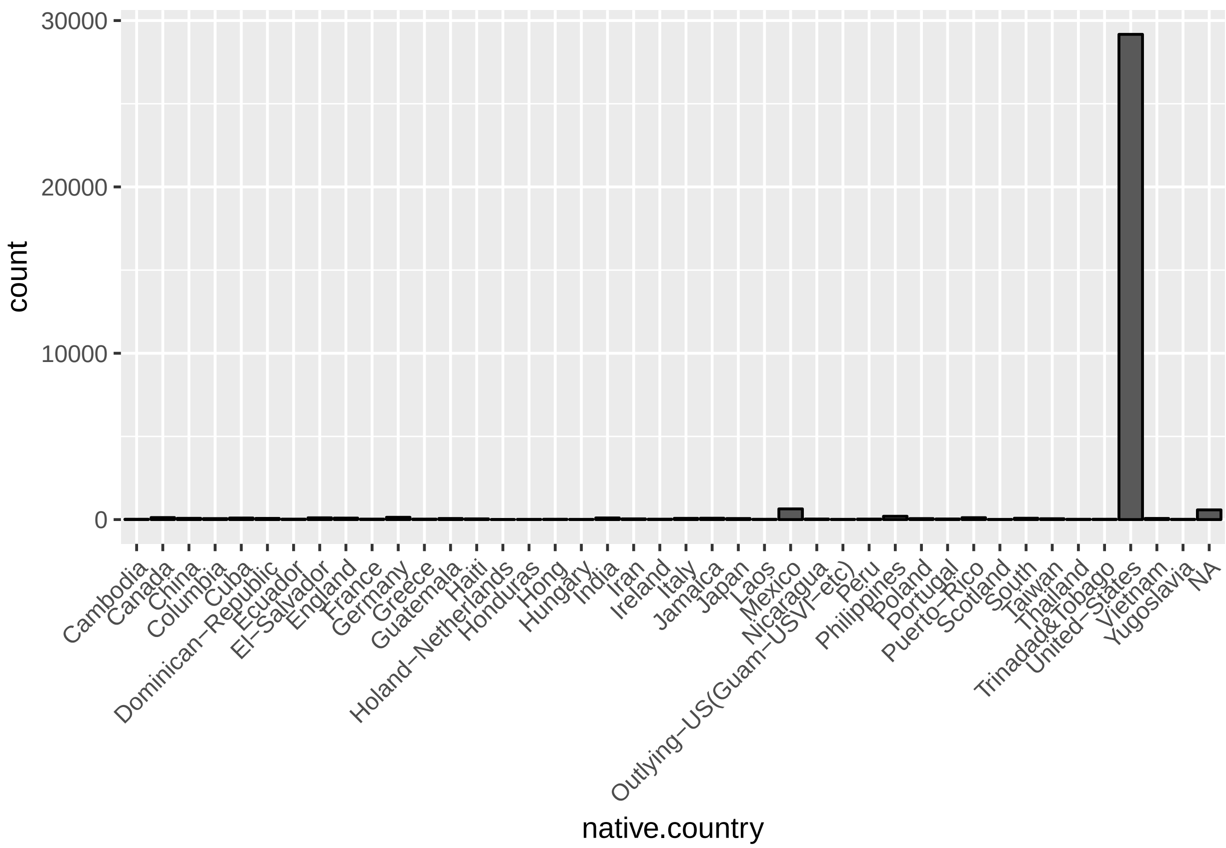
```r
colSums(datosAdult=="")
```

```
##            age      workclass         fnlwgt      education  education.num
##              0             NA              0              0              0
## marital.status     occupation   relationship           race            sex
##              0             NA              0              0              0
##   capital.gain   capital.loss hours.per.week native.country         income
##              0              0              0             NA              0
```

Ahora vemos que el atributo workclass tiene 1836 valores vacios, occupation tiene 1843 valores vacios, y native.country tiene 583 valores vacios.

Podemos intentar predecir los valores que son vacios. Primero miramos a **native.country** .

```r
# Visualizamos la distribución de la variable "native.country":
ggplot(data=datosAdult,aes(x=native.country))+geom_bar(color='black')+theme(axis.text.x=element_text(ang
```



Podemos asumir que los valores nulos de **native.country** son probablemente de origen EEUU. Ya que la mayoria de los datos son de ahí y normalmente la razón de no especificar el pais nativo es que se trata de una persona que ya es un nativo de EEUU.

Podemos borrar los casos que tienen **occupation** de valor vacío - estaría dificil hacer una estimación de los valores correctos.

Podemos borrar los casos que tienen **workclass** de valor vacío - estaría dificil hacer una estimaci?ó de los valores correctos.

Además no necesitamos el atributo **fnlwgt** para tratar los datos y se puede eliminarlo.

```
# Aplicamos el valor "United-States" para los valores vacíos de la variable "native.country"
datosAdult$native.country[is.na(datosAdult$native.country)]="United-States"
```

```
#Borramos los casos con valores vacios de occupation y workclass
#Before the delete we check
summary(datosAdult)
```

```
##       age                    workclass         fnlwgt
##  Min.   :17.00   Private         :22696   Min.   :  12285
##  1st Qu.:28.00   Self-emp-not-inc: 2541   1st Qu.: 117827
##  Median :37.00   Local-gov       : 2093   Median : 178356
##  Mean   :38.58   State-gov       : 1298   Mean   : 189778
##  3rd Qu.:48.00   Self-emp-inc    : 1116   3rd Qu.: 237051
##  Max.   :90.00   (Other)         :  981   Max.   :1484705
##                  NA's            : 1836
##         education     education.num                 marital.status
##  HS-grad     :10501   Min.   : 1.00   Divorced             : 4443
```

```
##  Some-college: 7291   1st Qu.: 9.00   Married-AF-spouse    :   23
##  Bachelors   : 5355   Median :10.00   Married-civ-spouse   :14976
##  Masters     : 1723   Mean   :10.08   Married-spouse-absent:  418
##  Assoc-voc   : 1382   3rd Qu.:12.00   Never-married        :10683
##  11th        : 1175   Max.   :16.00   Separated            : 1025
##  (Other)     : 5134                   Widowed              :  993
##           occupation           relationship                   race
##  Prof-specialty : 4140   Husband       :13193   Amer-Indian-Eskimo:  311
##  Craft-repair   : 4099   Not-in-family : 8305   Asian-Pac-Islander: 1039
##  Exec-managerial: 4066   Other-relative:  981   Black             : 3124
##  Adm-clerical   : 3770   Own-child     : 5068   Other             :  271
##  Sales          : 3650   Unmarried     : 3446   White             :27816
##  (Other)        :10993   Wife          : 1568
##  NA's           : 1843
##      sex          capital.gain    capital.loss    hours.per.week
##  Female:10771   Min.   :    0   Min.   :   0.0   Min.   : 1.00
##  Male  :21790   1st Qu.:    0   1st Qu.:   0.0   1st Qu.:40.00
##                 Median :    0   Median :   0.0   Median :40.00
##                 Mean   : 1078   Mean   :  87.3   Mean   :40.44
##                 3rd Qu.:    0   3rd Qu.:   0.0   3rd Qu.:45.00
##                 Max.   :99999   Max.   :4356.0   Max.   :99.00
##
##        native.country    income
##  United-States:29753   <=50K:24720
##  Mexico       :  643   >50K : 7841
##  Philippines  :  198
##  Germany      :  137
##  Canada       :  121
##  Puerto-Rico  :  114
##  (Other)      : 1595
```

```r
#delete the NAs

datosAdult <- na.omit(datosAdult)

# remember to re-factor the effected categories

datosAdult$workclass <- factor(datosAdult$workclass)
datosAdult$occupation <- factor(datosAdult$occupation)
datosAdult$native.country <- factor(datosAdult$native.country)
```

```r
#after the delete we check again
summary(datosAdult)
```

```
##       age                     workclass          fnlwgt
##  Min.   :17.00   Federal-gov     :  960   Min.   :  13769
##  1st Qu.:28.00   Local-gov       : 2093   1st Qu.: 117829
##  Median :37.00   Private         :22696   Median : 178517
##  Mean   :38.44   Self-emp-inc    : 1116   Mean   : 189846
##  3rd Qu.:47.00   Self-emp-not-inc: 2541   3rd Qu.: 237317
##  Max.   :90.00   State-gov       : 1298   Max.   :1484705
##                  Without-pay     :   14
##         education     education.num                  marital.status
##  HS-grad     :9968   Min.   : 1.00   Divorced             : 4258
##  Some-college:6775   1st Qu.: 9.00   Married-AF-spouse    :   21
```

```
##  Bachelors   :5182   Median :10.00   Married-civ-spouse   :14339
##  Masters     :1675   Mean   :10.13   Married-spouse-absent:  389
##  Assoc-voc   :1321   3rd Qu.:13.00   Never-married        : 9912
##  11th        :1056   Max.   :16.00   Separated            :  959
##  (Other)     :4741                   Widowed              :  840
##            occupation          relationship                   race
##  Prof-specialty :4140   Husband       :12704   Amer-Indian-Eskimo:  286
##  Craft-repair   :4099   Not-in-family : 7865   Asian-Pac-Islander:  974
##  Exec-managerial:4066   Other-relative:  918   Black             : 2909
##  Adm-clerical   :3770   Own-child     : 4525   Other             :  248
##  Sales          :3650   Unmarried     : 3271   White             :26301
##  Other-service  :3295   Wife          : 1435
##  (Other)        :7698
##      sex          capital.gain    capital.loss     hours.per.week
##  Female: 9930   Min.   :    0   Min.   :   0.00   Min.   : 1.00
##  Male  :20788   1st Qu.:    0   1st Qu.:   0.00   1st Qu.:40.00
##                 Median :    0   Median :   0.00   Median :40.00
##                 Mean   : 1106   Mean   :  88.91   Mean   :40.95
##                 3rd Qu.:    0   3rd Qu.:   0.00   3rd Qu.:45.00
##                 Max.   :99999   Max.   :4356.00   Max.   :99.00
##
##        native.country    income
##  United-States:28060   <=50K:23068
##  Mexico       :  610   >50K : 7650
##  Philippines  :  188
##  Germany      :  128
##  Puerto-Rico  :  109
##  Canada       :  107
##  (Other)      : 1516
```

We now have reduced the observations in the dataset datosAdult from 32561 to 30718 records.

Ahora verificamos que no tenemos los valores vacíos.

```r
# Estadísticas de valores vacíos
colSums(is.na(datosAdult))
```

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

```r
colSums(datosAdult=="")
```

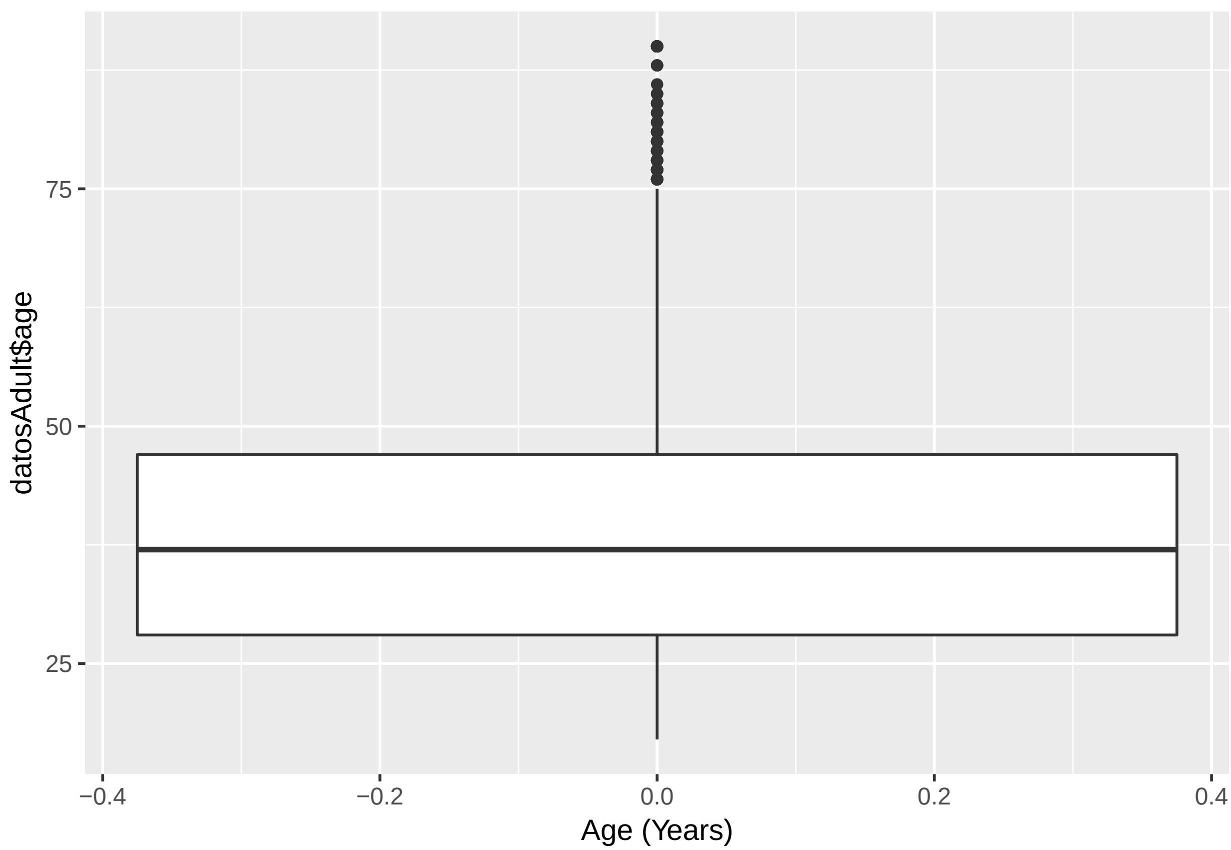
```
##            age      workclass         fnlwgt      education  education.num
##              0              0              0              0              0
## marital.status     occupation   relationship           race            sex
##              0              0              0              0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

3.2 Identificación y tratamiento de valores extremos

We will check the numeric variables (age, education.num, capital.gain, capital.loss, hours.per.week) for outliers.

3.2.1 Variable age

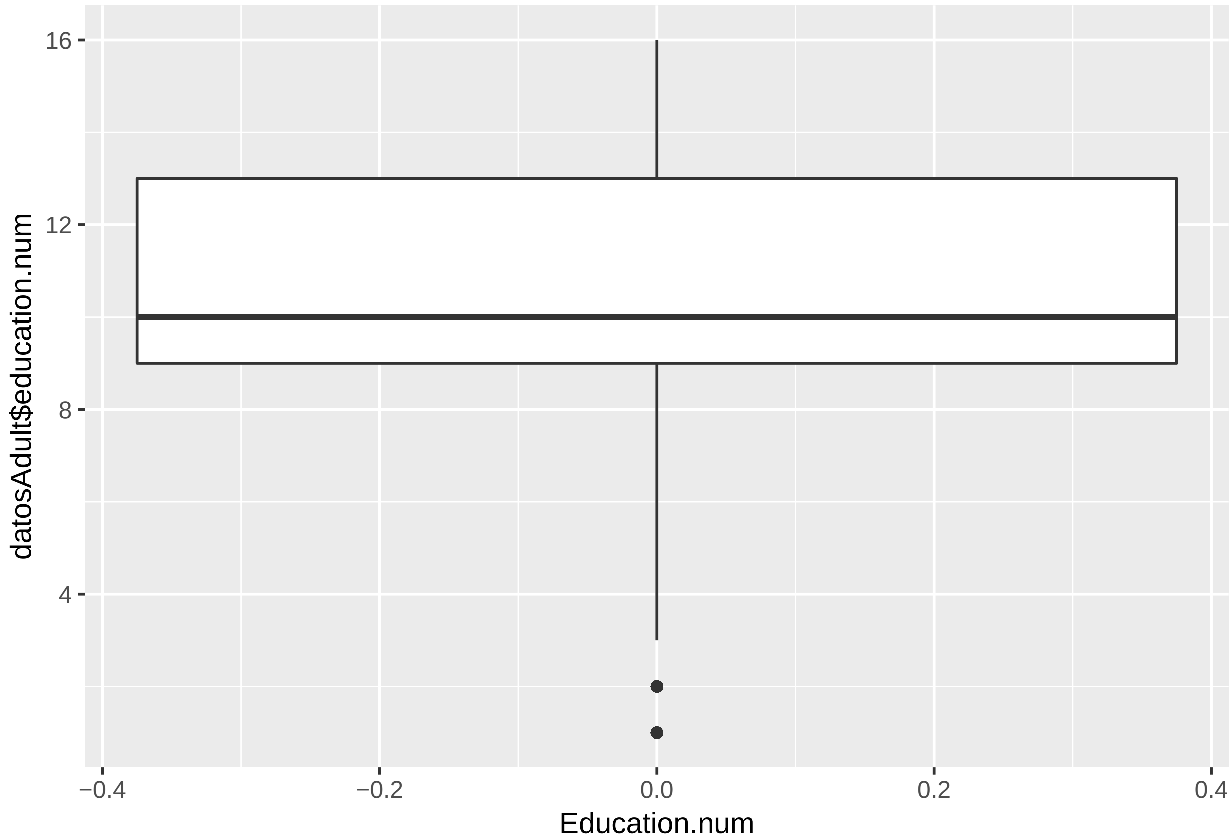
```
AgeBoxplot<-ggplot(datosAdult,aes(y=datosAdult$age))+geom_boxplot() +labs(x="Age (Years)")+ guides(fill=

AgeBoxplot
```



In this case we see that although there are outliers with age above 75, these are not unreasonable data values, so we will keep these values.

3.2.2 Variable Education.num

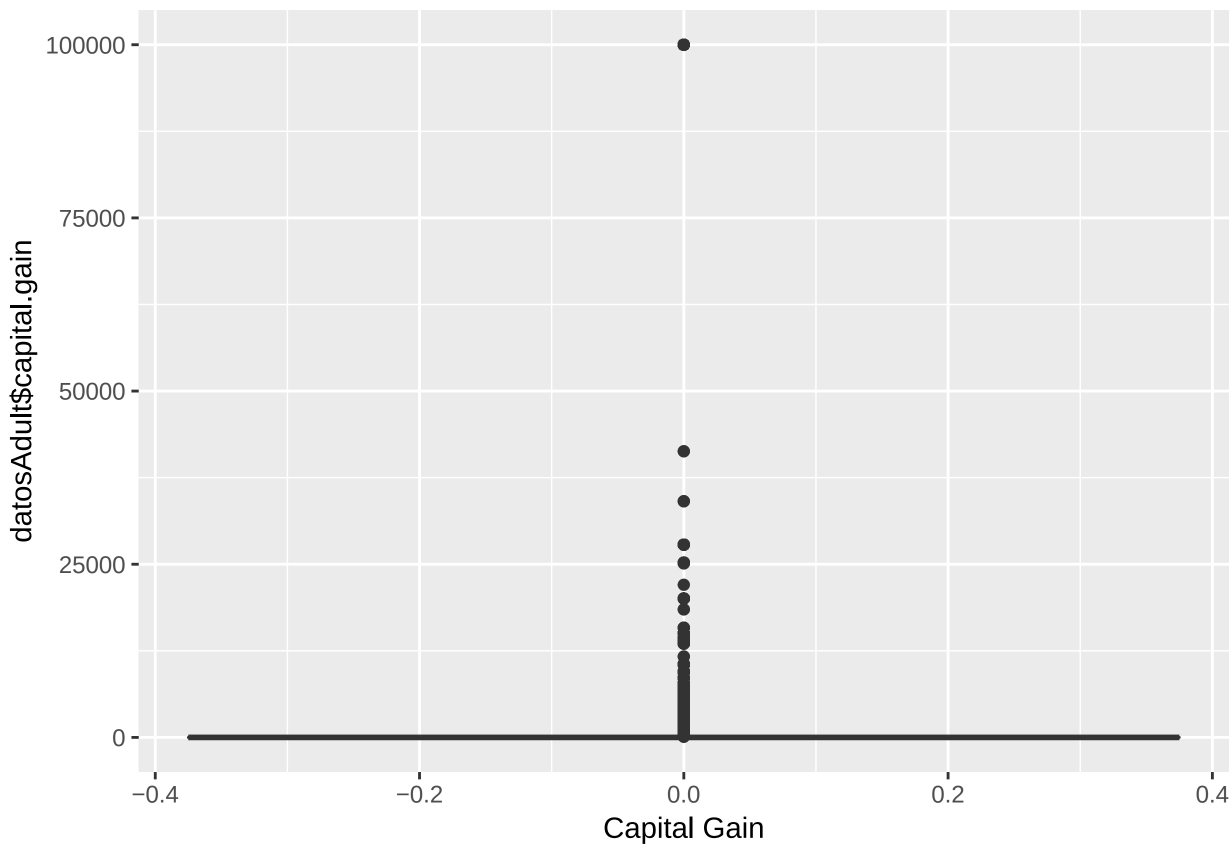
```
Education.numBoxplot<-ggplot(datosAdult,aes(y=datosAdult$education.num))+geom_boxplot() +labs(x="Educati

Education.numBoxplot
```



In this case we see that the outliers with values 1 and 2 are not unreasonable values and correspond to an educational level of the person.

3.2.3 Variable Capital.gain

```
CapGainBoxplot<-ggplot(datosAdult,aes(y=datosAdult$capital.gain))+geom_boxplot() +labs(x="Capital Gain")

CapGainBoxplot
```



Here we see that capital gain values are mostly zeros. The main outlier of concern is the one around 100000. Let´s check the variation of capital.gain.

```
summary(datosAdult$capital.gain)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##       0       0       0    1106       0   99999
```

So we have a median value of zero and a mean of 1106. The maximum value of 99999 is an outlier probably caused by data entry field size limitations.

The percentage of zero values is very high for this variable.

```
(nrow(subset(datosAdult, datosAdult$capital.gain == 0))/nrow(datosAdult))*100
```

```
## [1] 91.57172
```

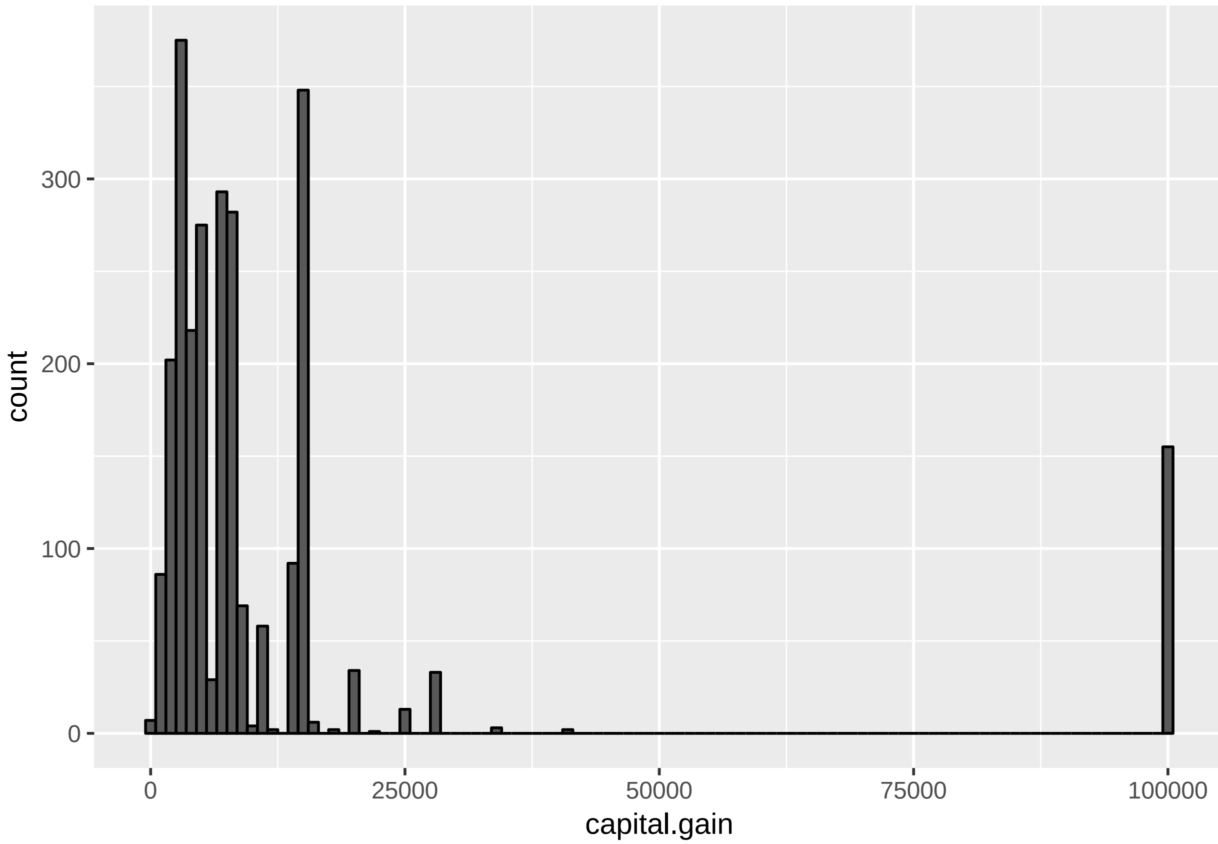
Let's look at the distribution of the non-zero values of **capital.gain**

```
summary(datosAdult$capital.gain[datosAdult$capital.gain !=0])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     114    3464    7298   13123   14084   99999
```

And let's see a barchart of the non-zero values.

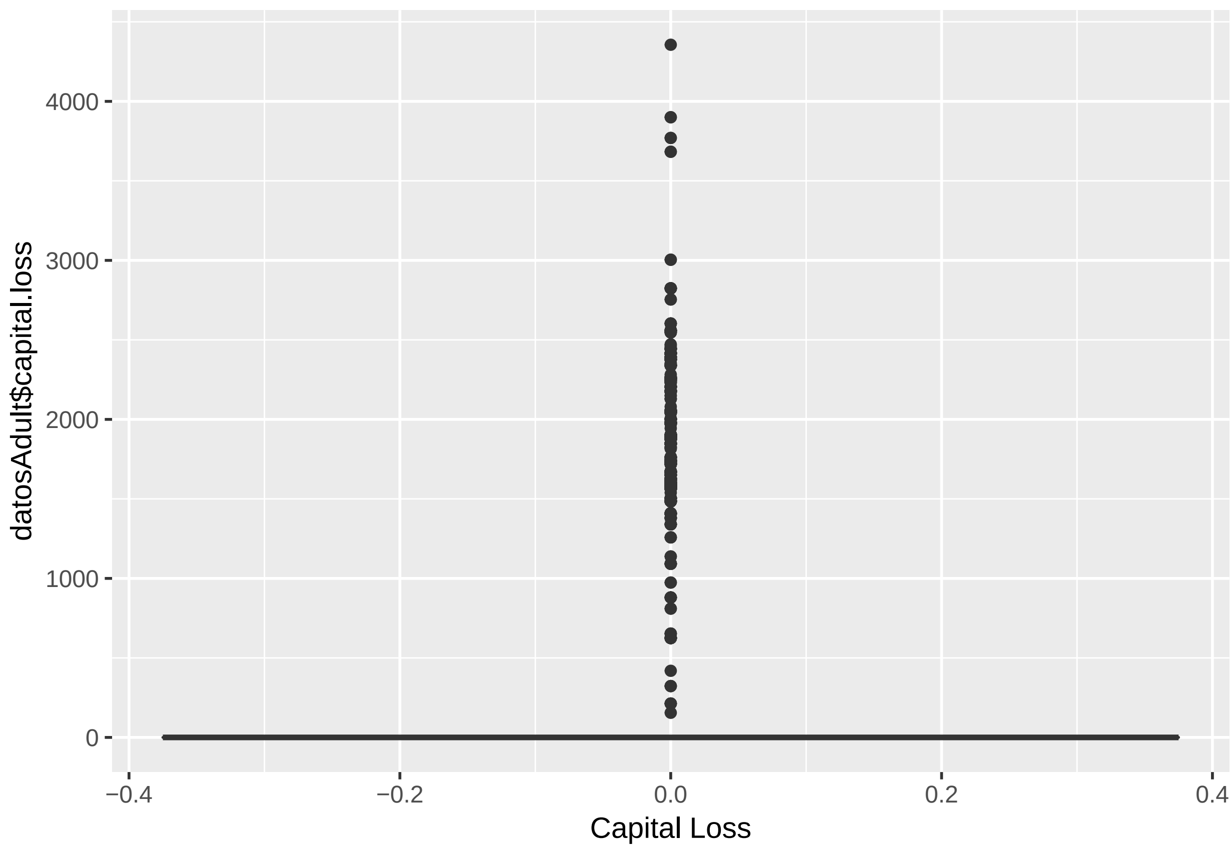
```
# Miramos a capital.gain en bins de tamaño 1000.


ggplot(datosAdult[which(datosAdult$capital.gain !=0),]) + aes(x=capital.gain) +
  geom_histogram(binwidth=1000, color='black')
```



Así que los valores de 99999 representan a gente con capital.gain muy alta y no podemos descartar el valor. Miramos a la otra variable capital.loss

3.2.4 Variable Capital.loss

```
CapLossBoxplot<-ggplot(datosAdult,aes(y=datosAdult$capital.loss))+geom_boxplot() +labs(x="Capital Loss")
CapLossBoxplot
```



Here we see that capital loss values are mostly zeros. Let´s check the variation of capital.loss.

```
summary(datosAdult$capital.loss)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    0.00   88.91    0.00 4356.00
```

So we have a median value of zero and a mean of 88.91. The maximum value of 4356 is not really an outlier as we have many zero values for capital.loss.

The percentage of zero values is very high for this variable.

```
(nrow(subset(datosAdult, datosAdult$capital.loss == 0))/nrow(datosAdult))*100
```

```
## [1] 95.24383
```

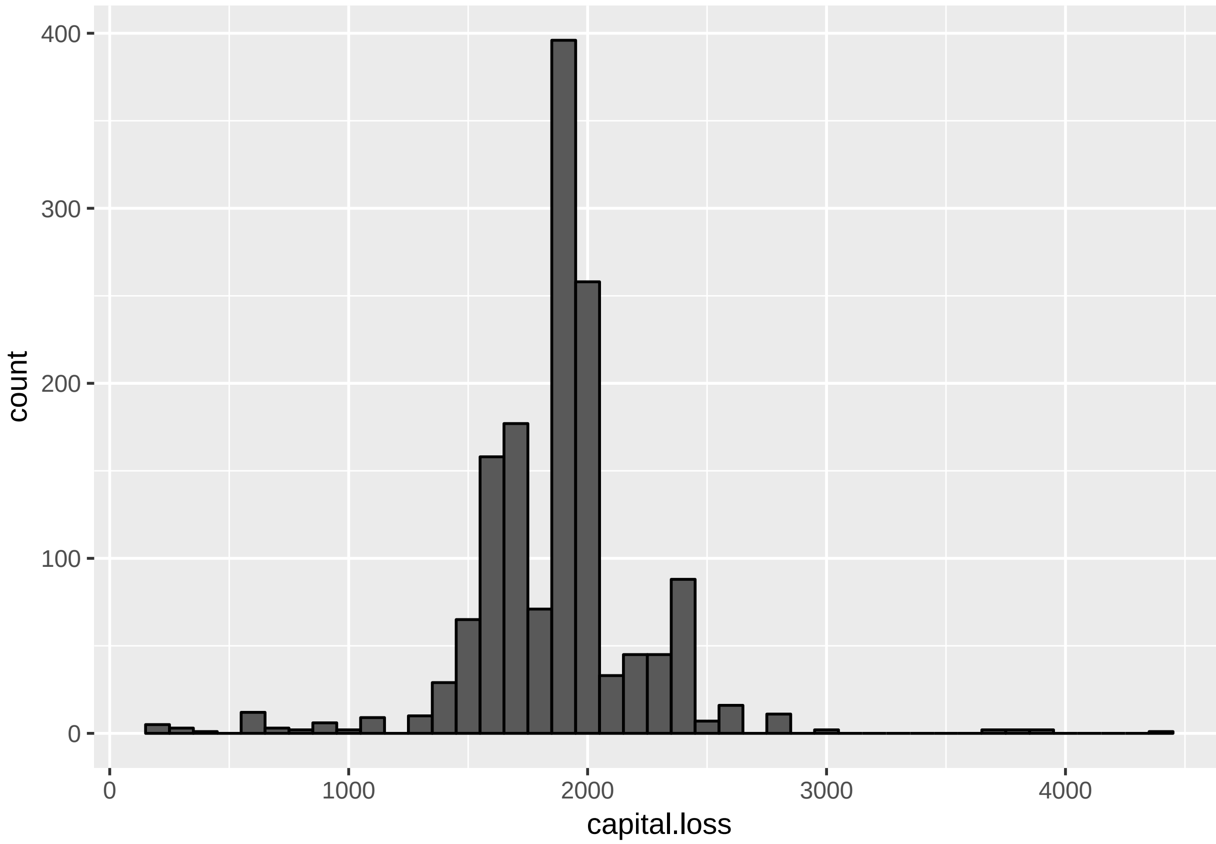
Let's look at the distribution of the non-zero values of **capital.loss**

```
summary(datosAdult$capital.loss[datosAdult$capital.loss !=0])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     155    1672    1887    1869    1977    4356
```

And let's see a barchart of the non-zero values.

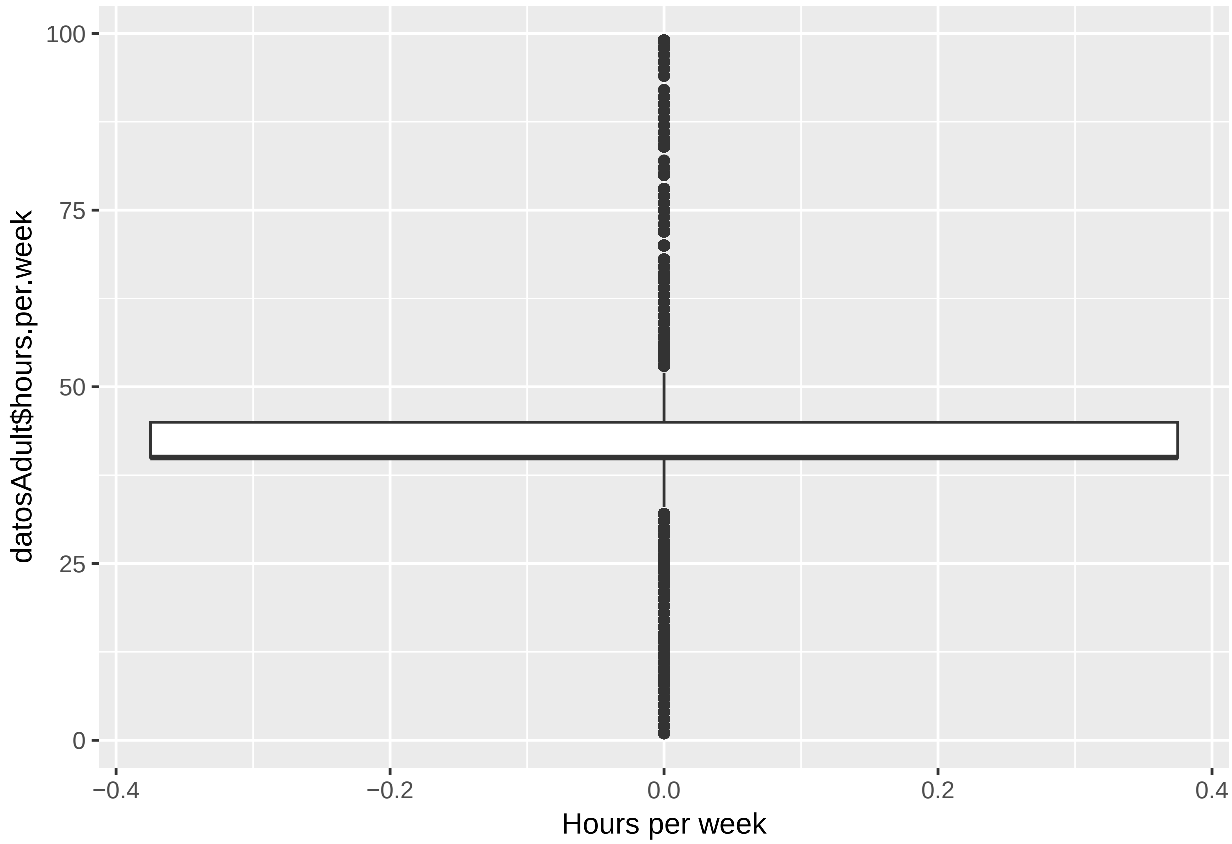
```
# Miramos a capital.loss en bins de tamaño 100.


ggplot(datosAdult[which(datosAdult$capital.loss !=0),]) + aes(x=capital.loss) +
  geom_histogram(binwidth=100, color='black')
```



Así que los valores de de capital.loss parecen razonables.

3.2.5 Variable hours.per.week

```
HoursBoxplot<-ggplot(datosAdult,aes(y=datosAdult$hours.per.week))+geom_boxplot() +labs(x="Hours per wee

HoursBoxplot
```



Here we see that the boxplot highlights significant numbers of outliers. Let´s check the variation of hours.per.week.

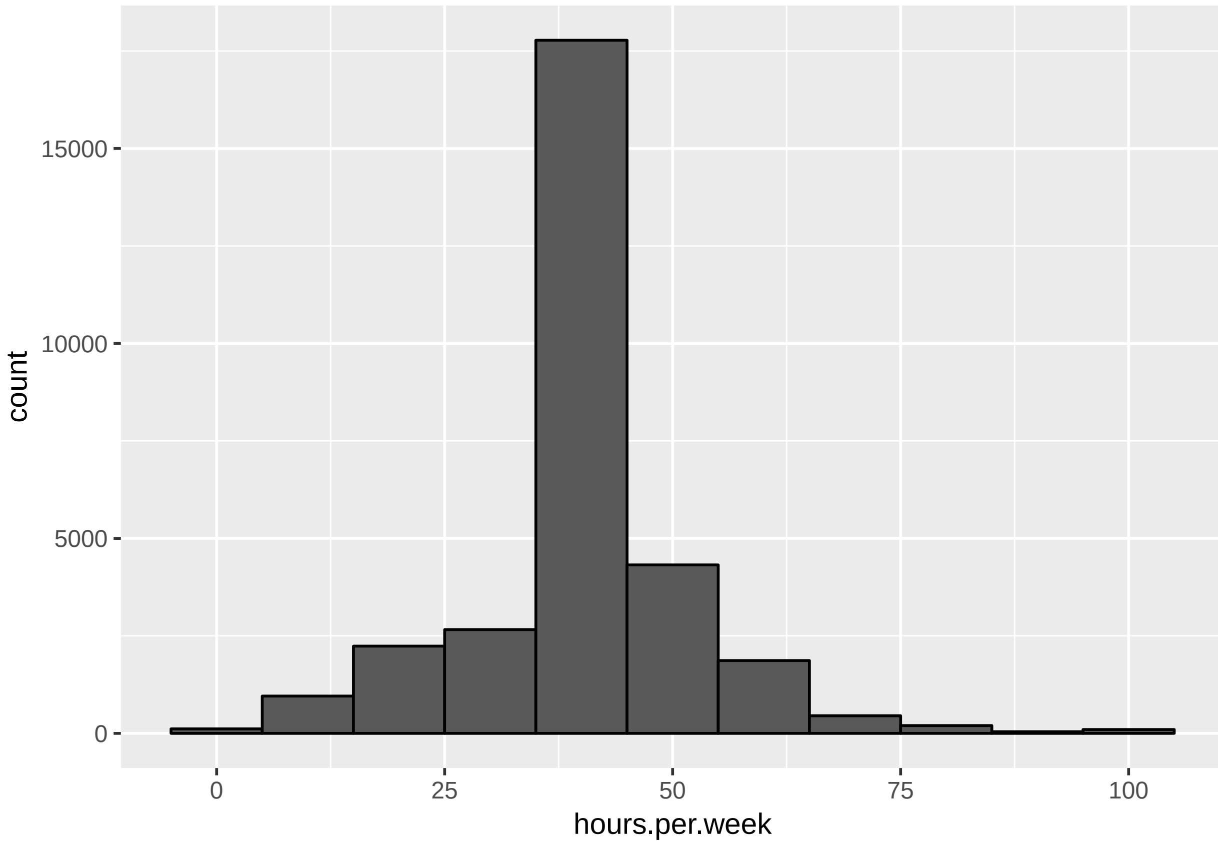
```
summary(datosAdult$hours.per.week)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    1.00   40.00   40.00   40.95   45.00   99.00
```

Let's see a histogram of the distribution of values.

Miramos a hours.per.week en bins de tamaño 10 horas.

```
ggplot(datosAdult) + aes(x=hours.per.week) +
  geom_histogram(binwidth=10, color='black')
```



Aunque hay algunos persona que traban más de 100 horas por semana, la distribución no indica que deberíamos descartar outliers.

4 Análisis de los datos

4.1 Selección de los grupos de datos

We will start by considering which variables may have an important correlation with income bracket (two values <50k and >50k). To begin with, we will remove the variable fnlwgt which assigns a weighting related on the population size of the US State in which the person lives.

```
datosAdult$fnlwgt<-NULL
```

Now we will change the *income* factor variable to have the values 0 or 1 to represent <50k and >50k

```
datosAdult$income <- as.numeric(datosAdult$income)-1
```

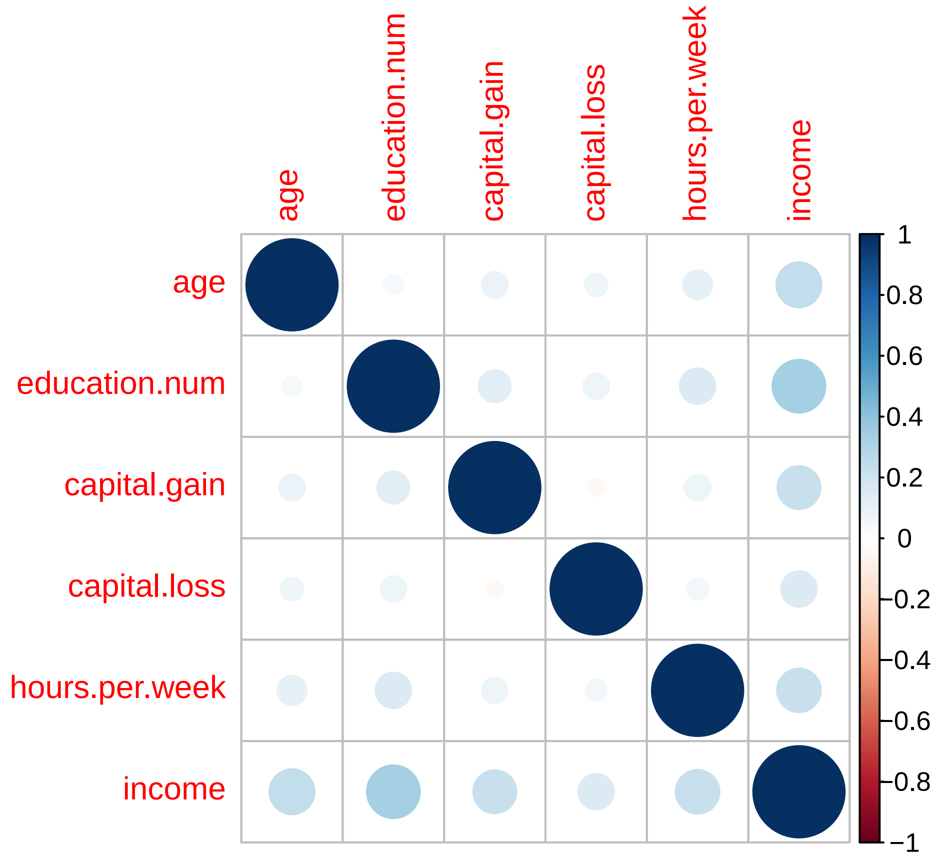
4.1.1 Correlation of numeric variables

Now we will correlate the numeric variables (age, education.num, capital.gain, capital.loss, hours.per.week and the class income) to see what shows up!

```
#Correlation plot
num.var <- c(1,4,10:12, 14)
library("corrplot")
```

```
## corrplot 0.84 loaded
```

```r
corrplot(cor(datosAdult[,num.var]))
```



So we see a positive correlation with all numeric variables, but especially with **education.num**, **age** and **hours.per.week**

4.1.2 Category variables

Let's look at the category variables workclass, education, marital.status, occupation, relationship, race, sex, native.country.

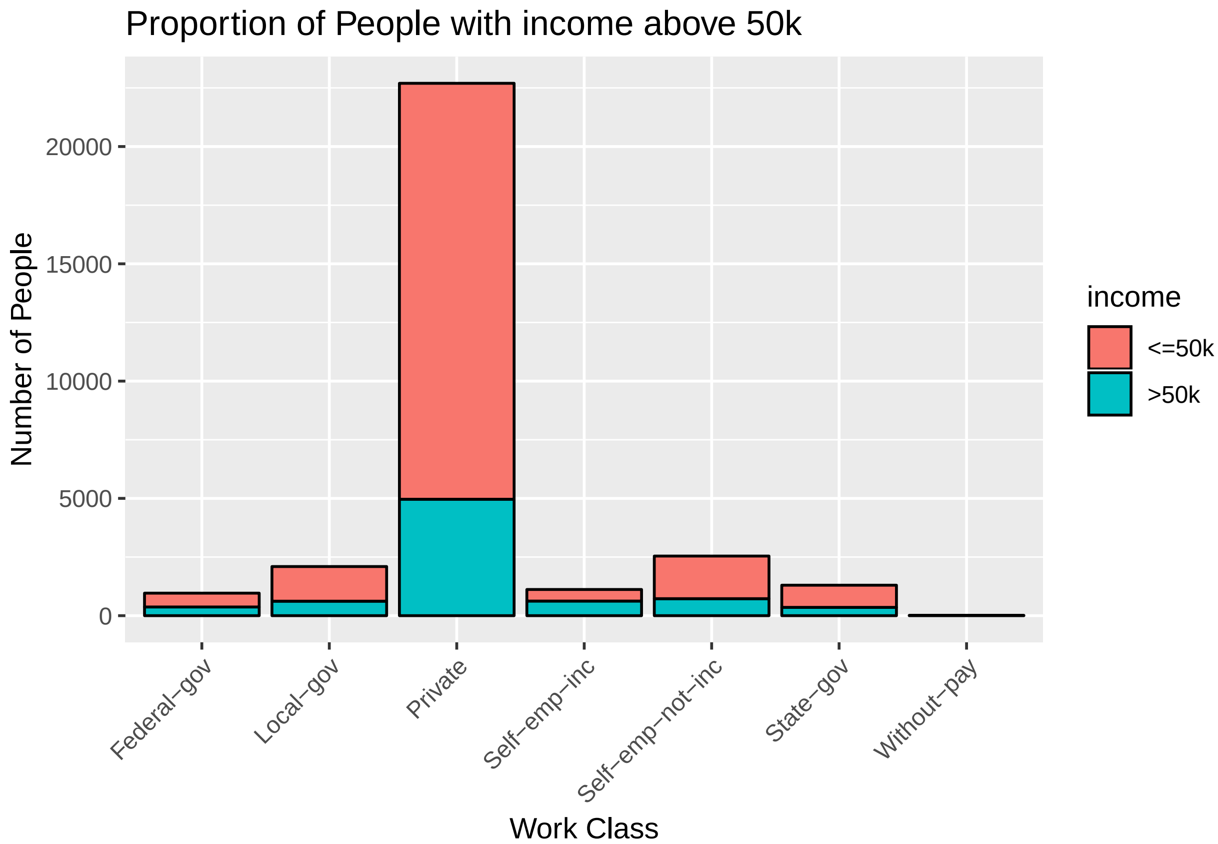
First let's re-factor the income class.

```r
datosAdult$income <- factor(datosAdult$income, labels=c("<=50k", ">50k"))
#Checking the levels
levels(datosAdult$income)
```

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

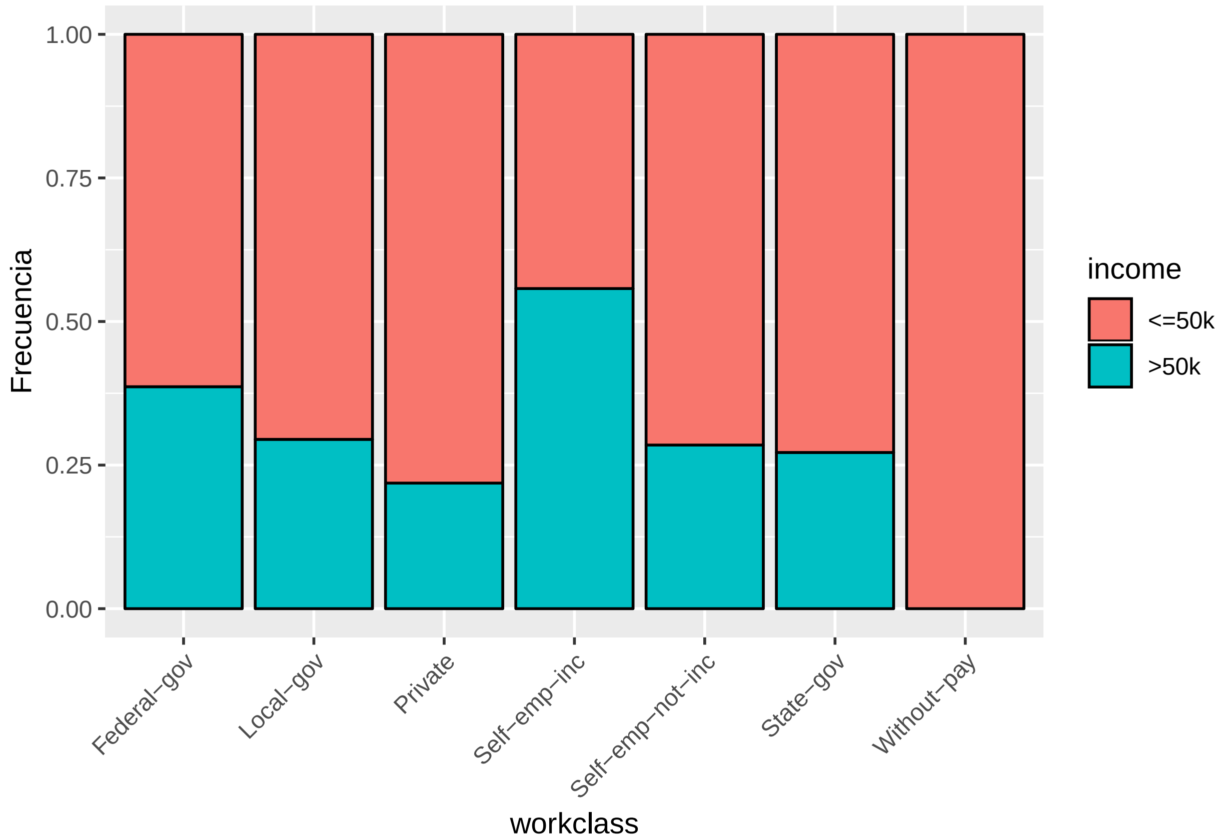
4.1.2.1 workclass

```r
ggplot(datosAdult,aes(x=workclass,fill=income))+
  geom_bar(color='black')+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))+
  ggtitle('Proportion of People with income above 50k')+
  xlab("Work Class")+ylab("Number of People")
```



Let's display as a frequency plot too.

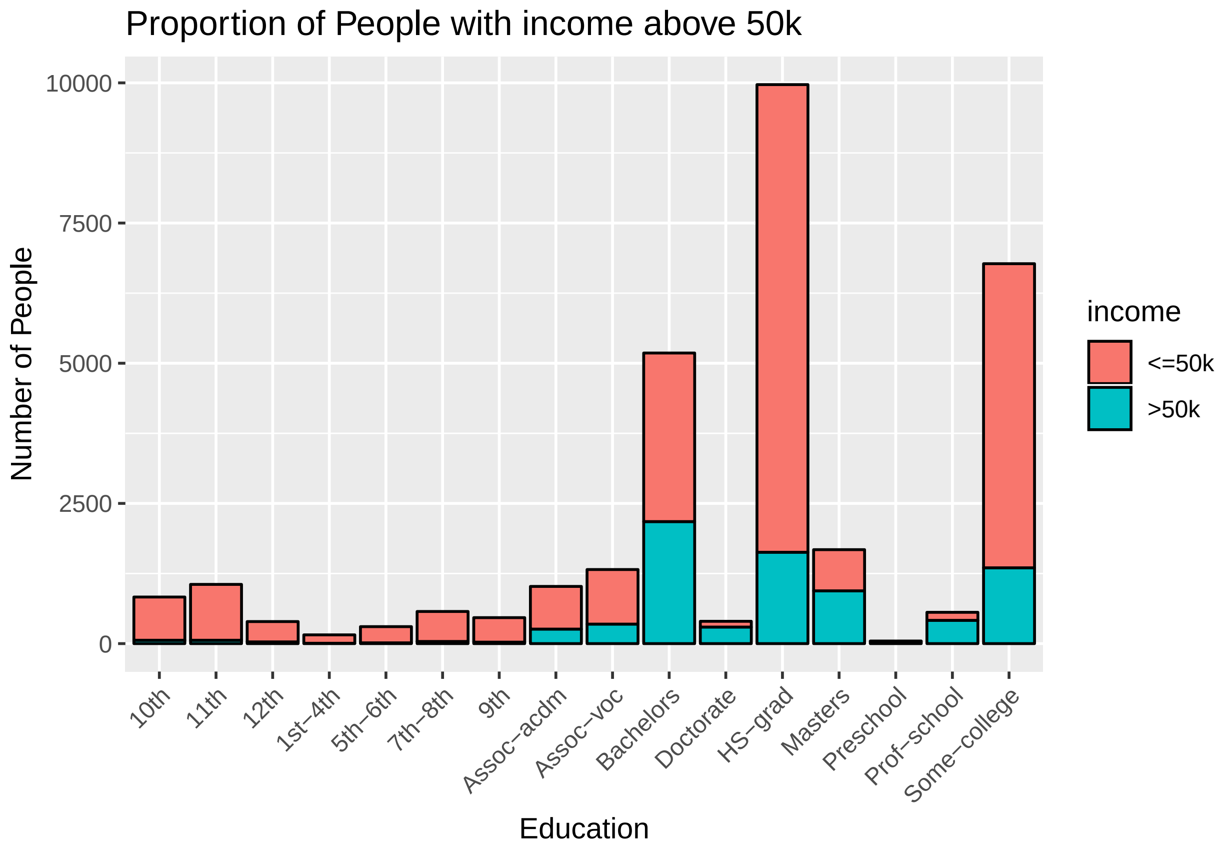
```
ggplot(data = datosAdult,aes(x=workclass,fill=income))+geom_bar(color='black',position="fill")+theme(ax
```



So the self-employed-inc and federal-gov employees are most likely to have high salaries. Those Without-pay do not have high salaries.

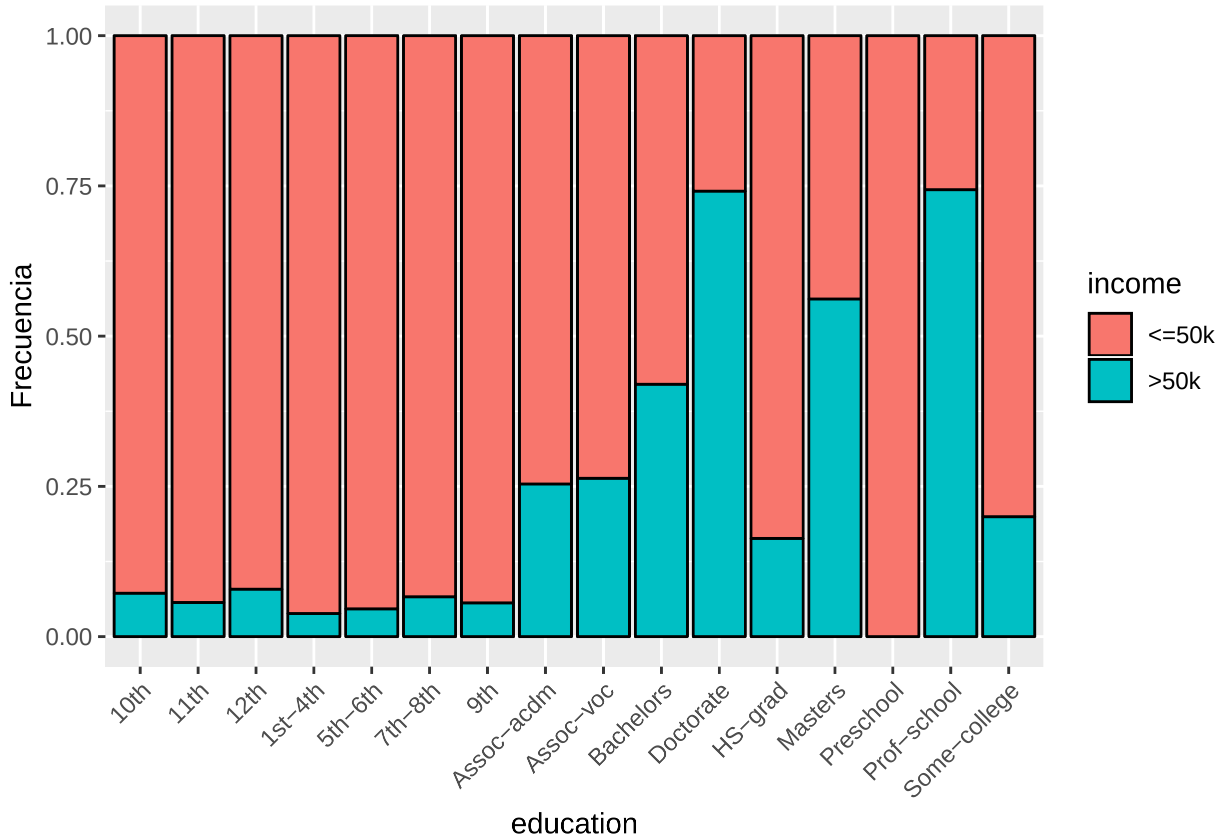
4.1.2.2 education

```
ggplot(datosAdult,aes(x=education,fill=income))+
  geom_bar(color='black')+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))+
  ggtitle('Proportion of People with income above 50k')+
  xlab("Education")+ylab("Number of People")
```



Let's display as a frequency plot too.

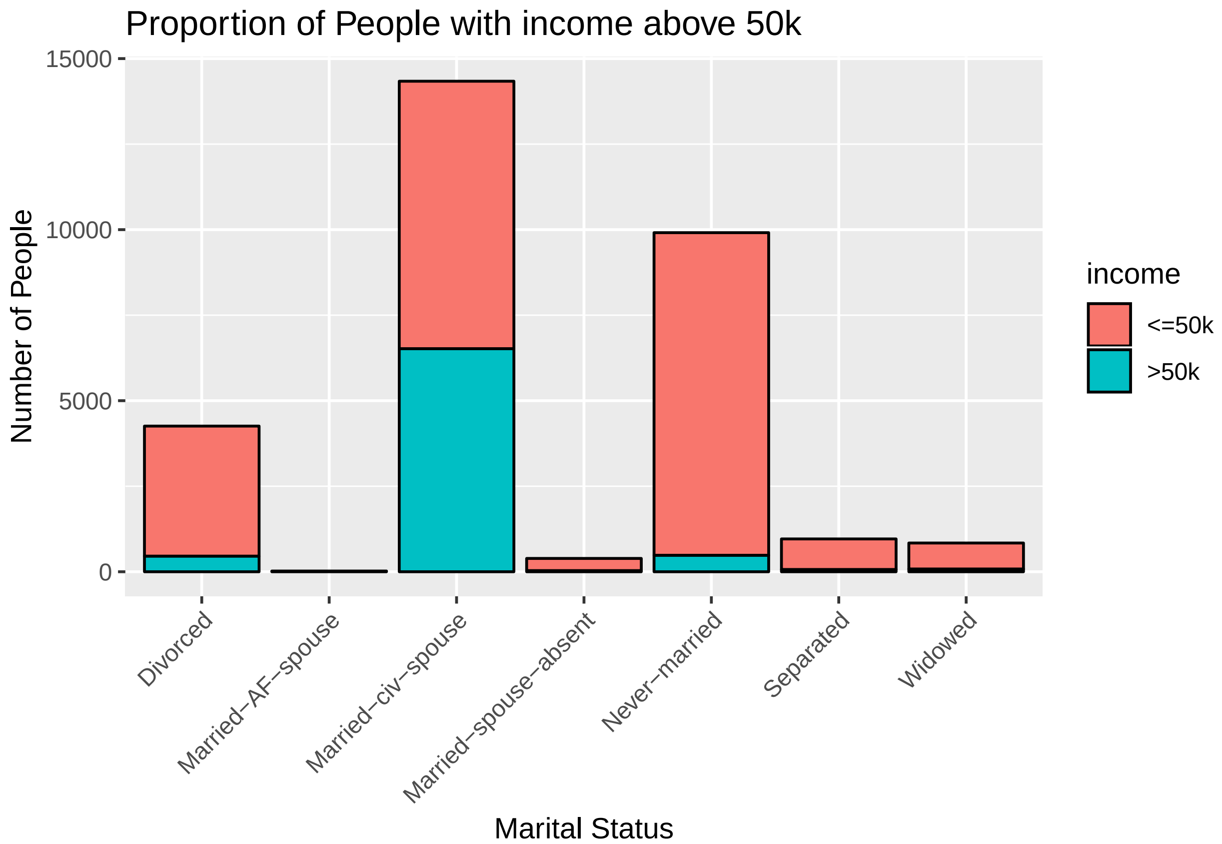
```
ggplot(data = datosAdult,aes(x=education,fill=income))+geom_bar(color='black',position="fill")+theme(ax
```



So as we might expect the people with degrees and professional schooling are more likely to have high salaries.

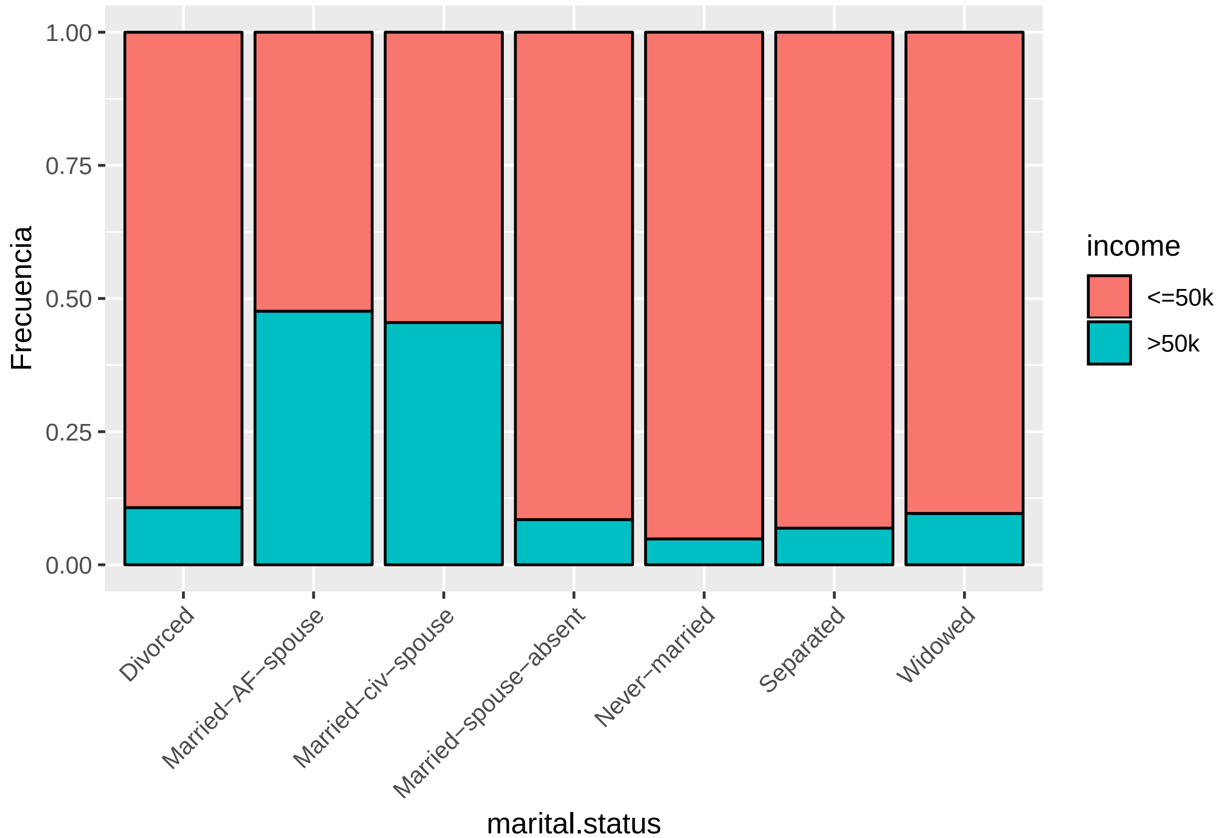
4.1.2.3 marital.status

```
ggplot(datosAdult,aes(x=marital.status,fill=income))+
  geom_bar(color='black')+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))+
  ggtitle('Proportion of People with income above 50k')+
  xlab("Marital Status")+ylab("Number of People")
```



Let's display as a frequency plot too.

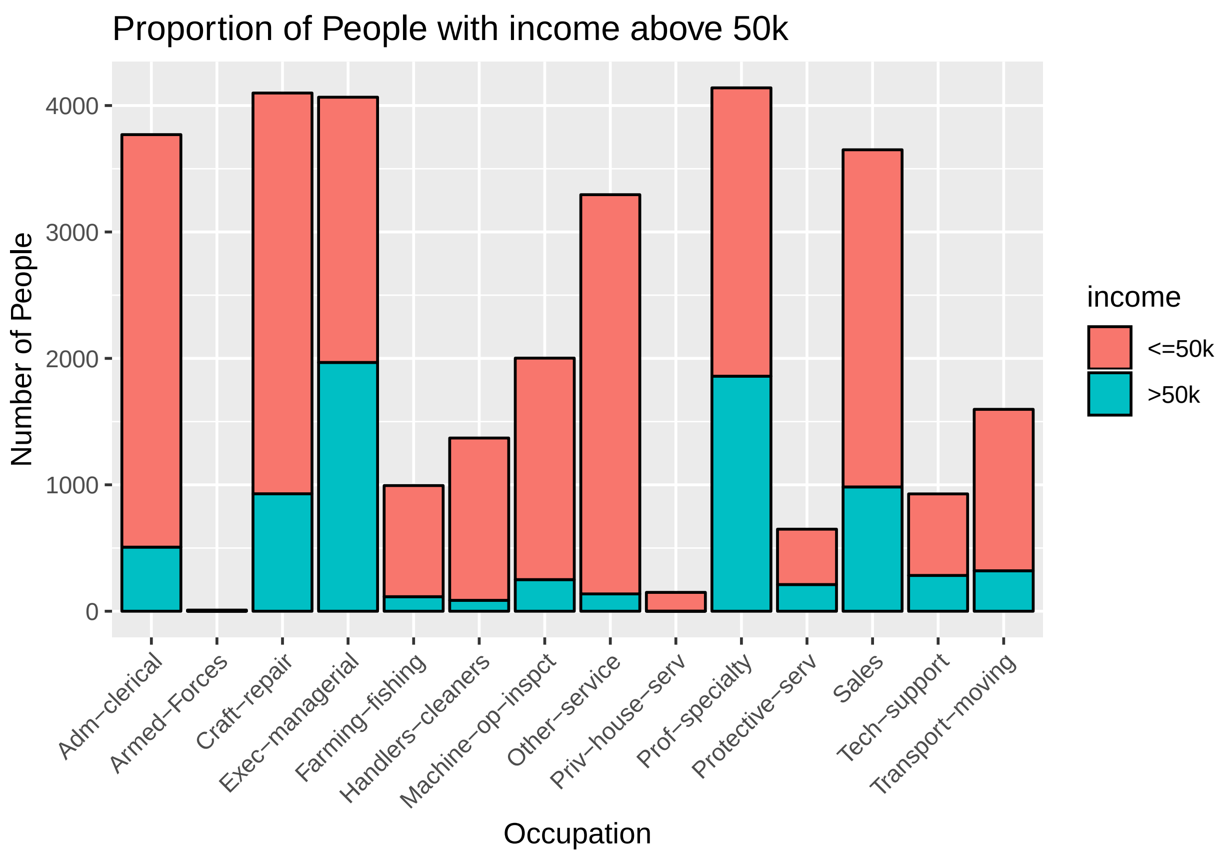
```
ggplot(data = datosAdult,aes(x=marital.status,fill=income))+geom_bar(color='black',position="fill")+ther
```



So we see that people who are married are more likely to have high salaries.

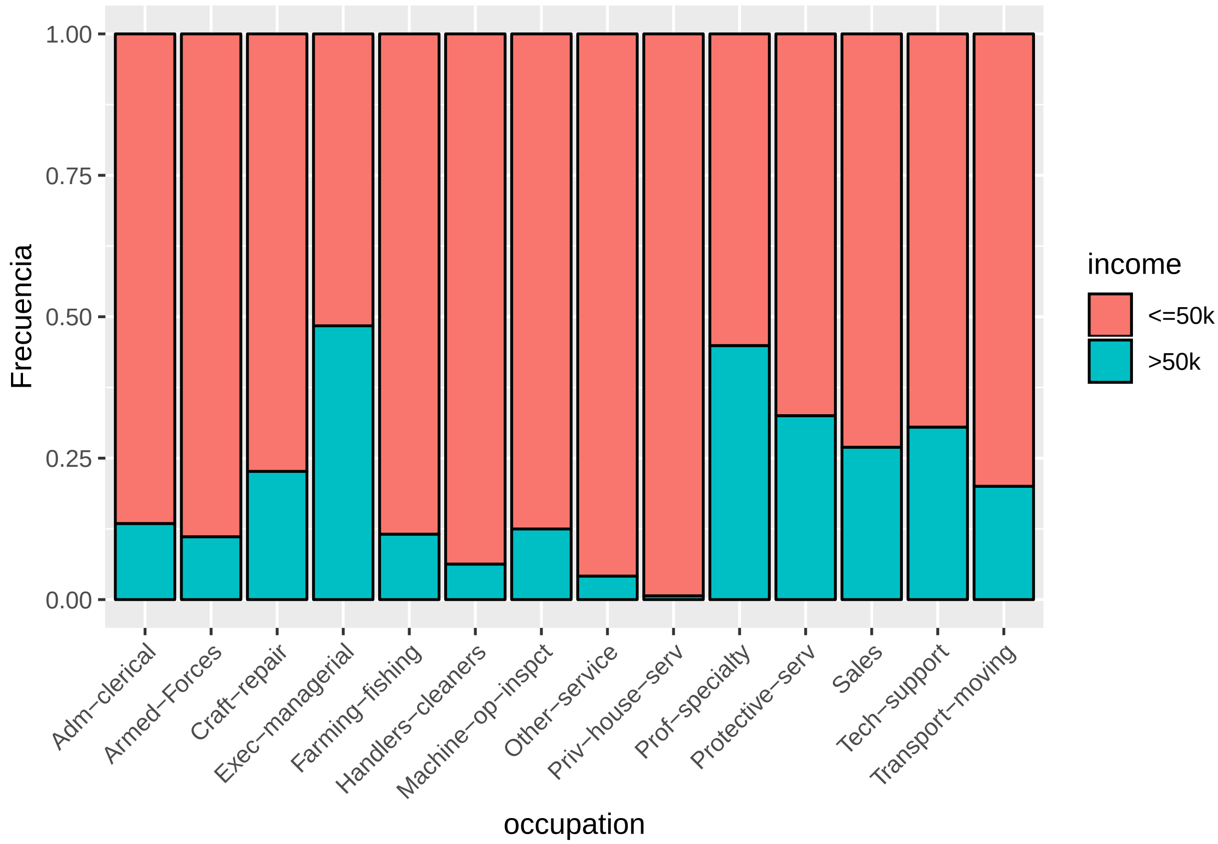
4.1.2.4 occupation

```
ggplot(datosAdult,aes(x=occupation,fill=income))+
  geom_bar(color='black')+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))+
  ggtitle('Proportion of People with income above 50k')+
  xlab("Occupation")+ylab("Number of People")
```



Let's display as a frequency plot too.

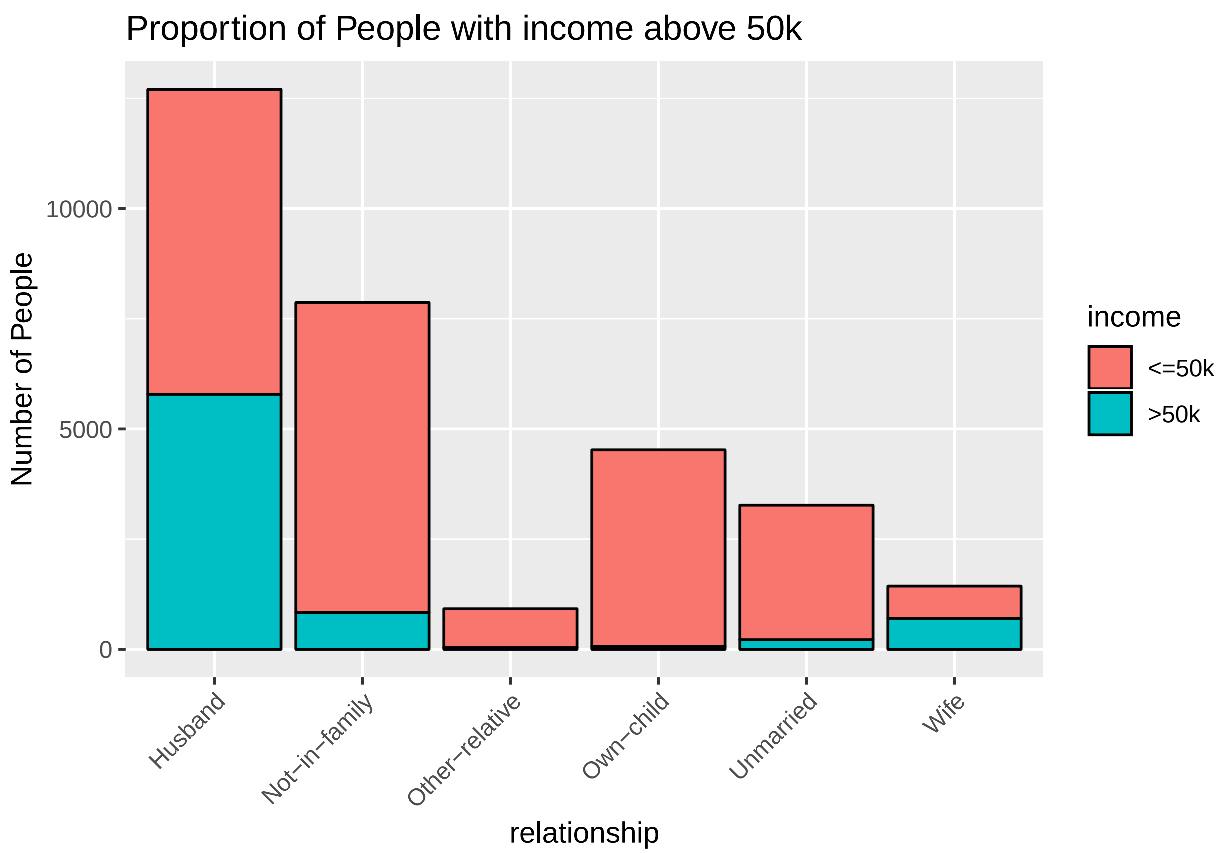
```
ggplot(data = datosAdult,aes(x=occupation,fill=income))+geom_bar(color='black',position="fill")+theme(a
```



So as we might expect the people who are executive managers or have aq p'rofessional speciality are more likely to have high salaries.

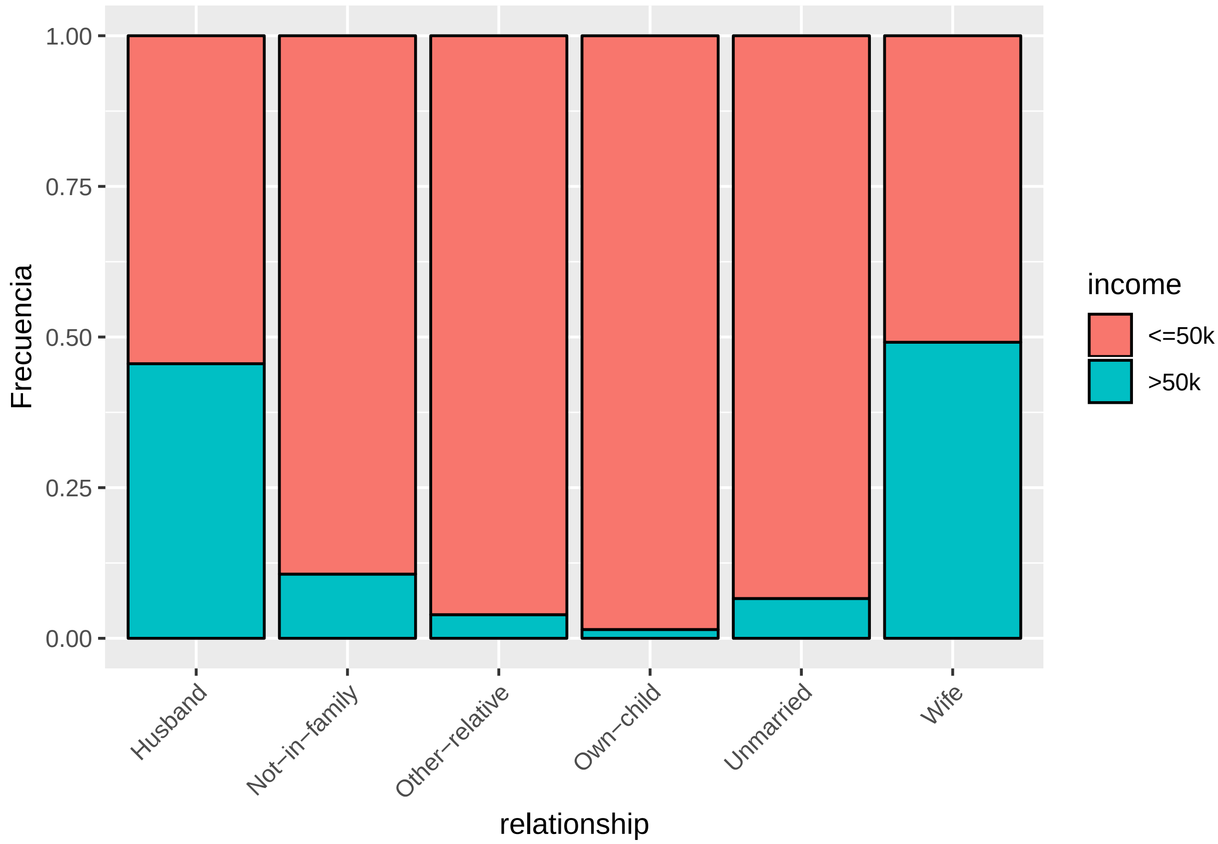
4.1.2.5 relationship

```
ggplot(datosAdult,aes(x=relationship,fill=income))+
  geom_bar(color='black')+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))+
  ggtitle('Proportion of People with income above 50k')+
  xlab("relationship")+ylab("Number of People")
```



Let's display as a frequency plot too.

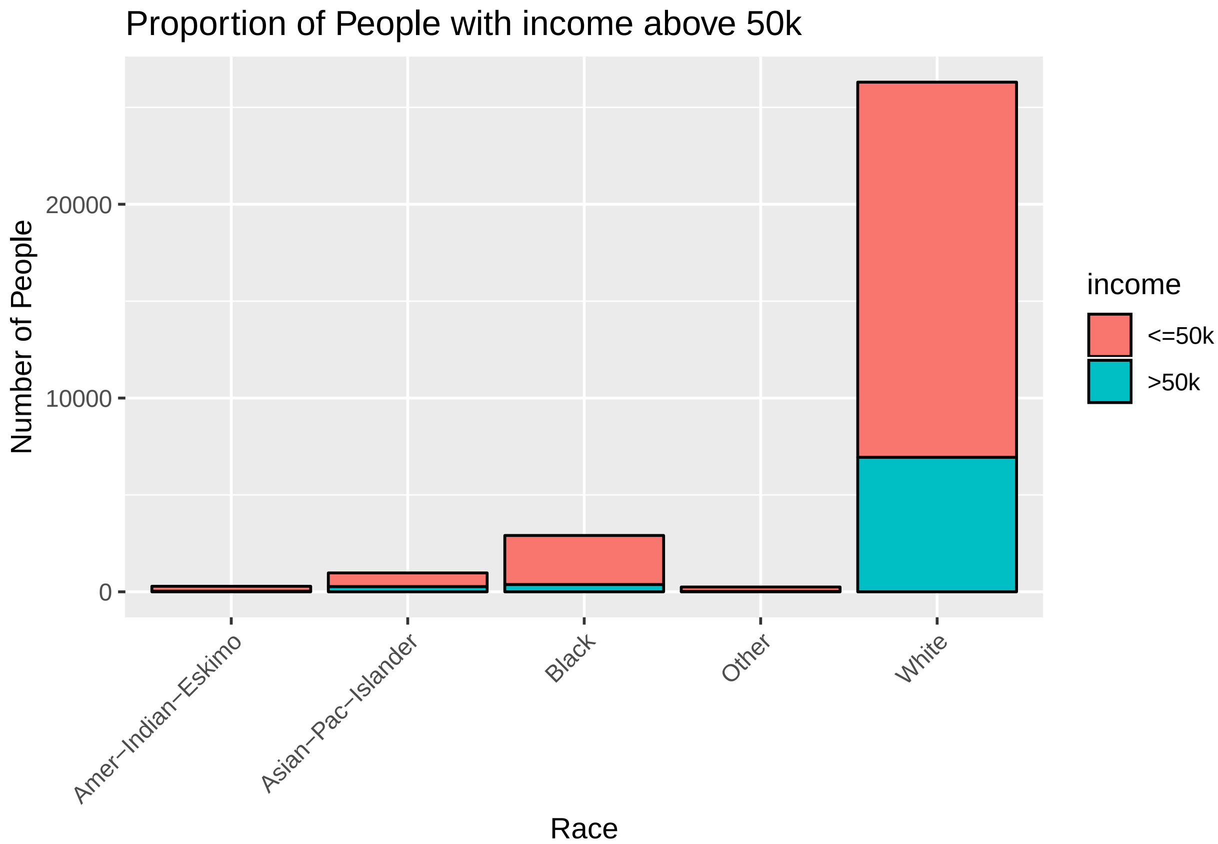
```
ggplot(data = datosAdult,aes(x=relationship,fill=income))+geom_bar(color='black',position="fill")+theme
```



As with the marital.status variable, people who are married are more likely to have high salaries.

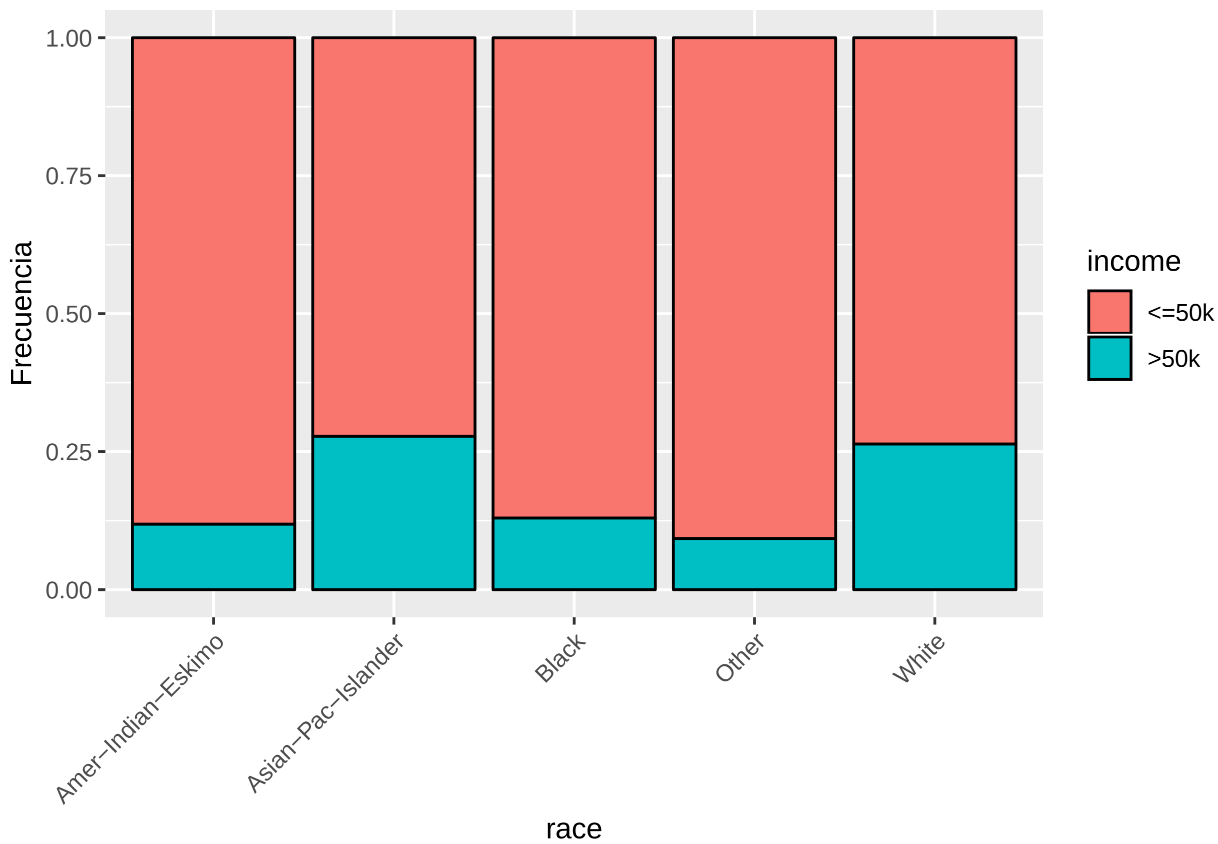
4.1.2.6 race

```
ggplot(datosAdult,aes(x=race,fill=income))+
  geom_bar(color='black')+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))+
  ggtitle('Proportion of People with income above 50k')+
  xlab("Race")+ylab("Number of People")
```



Let's display as a frequency plot too.

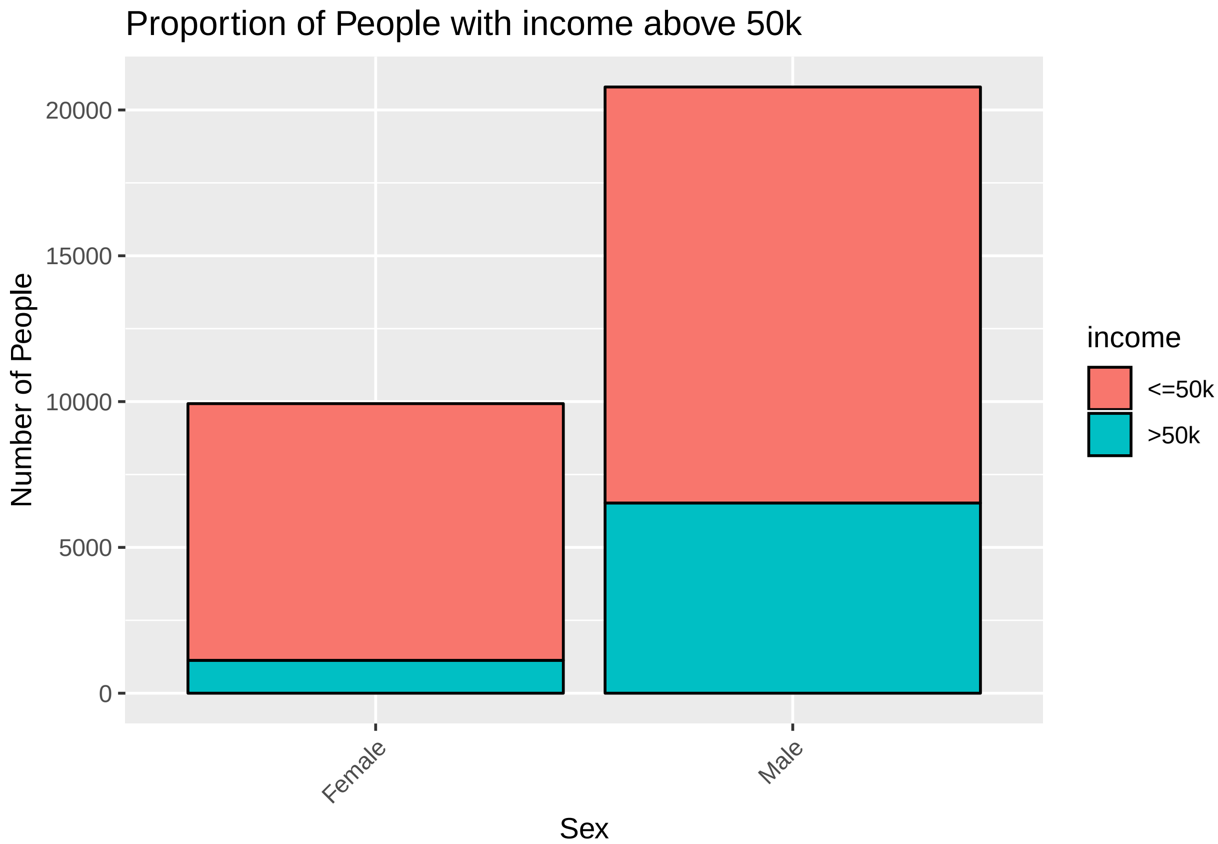
```
ggplot(data = datosAdult,aes(x=race,fill=income))+geom_bar(color='black',position="fill")+theme(axis.te
```



So we see that White and Asian-Pac-Islander ethnic groups are more likely to have high salaries.

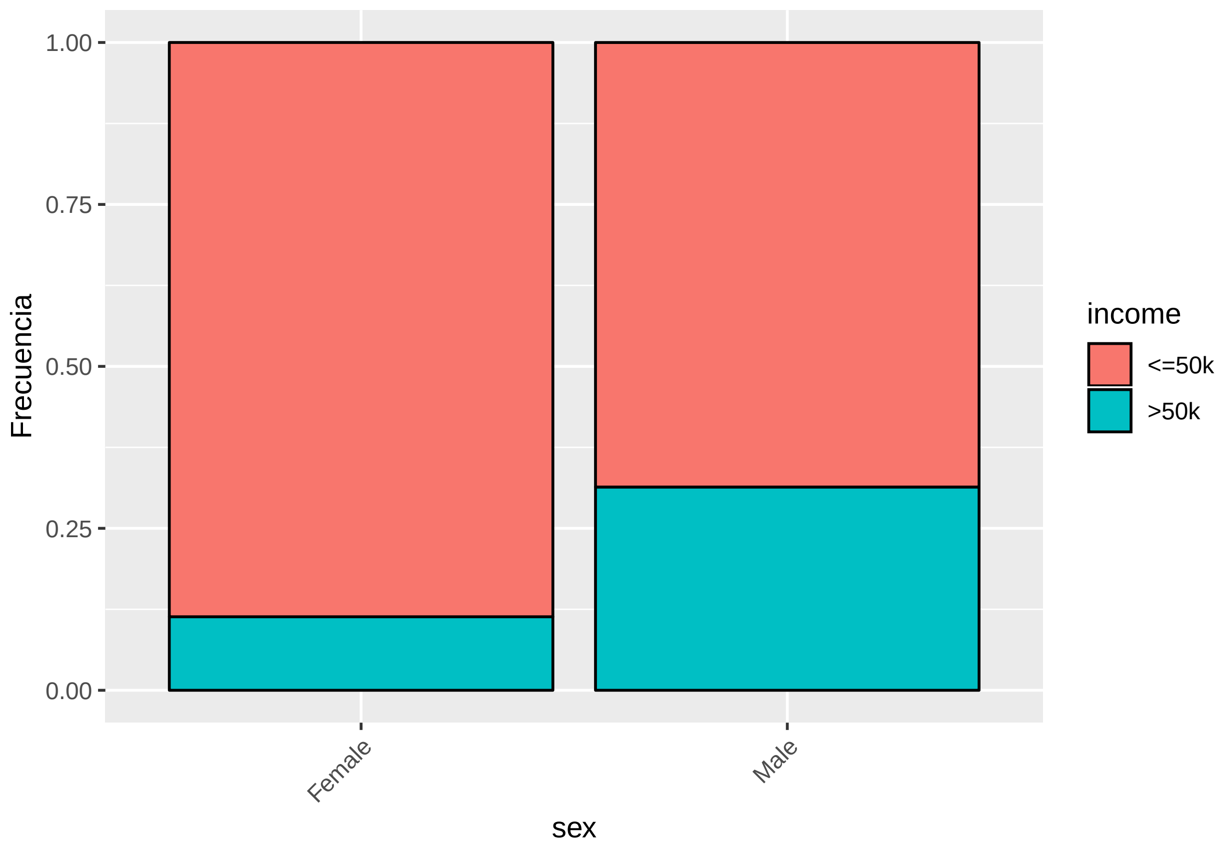
4.1.2.7 sex

```
ggplot(datosAdult,aes(x=sex,fill=income))+
  geom_bar(color='black')+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))+
  ggtitle('Proportion of People with income above 50k')+
  xlab("Sex")+ylab("Number of People")
```



Let's display as a frequency plot too.

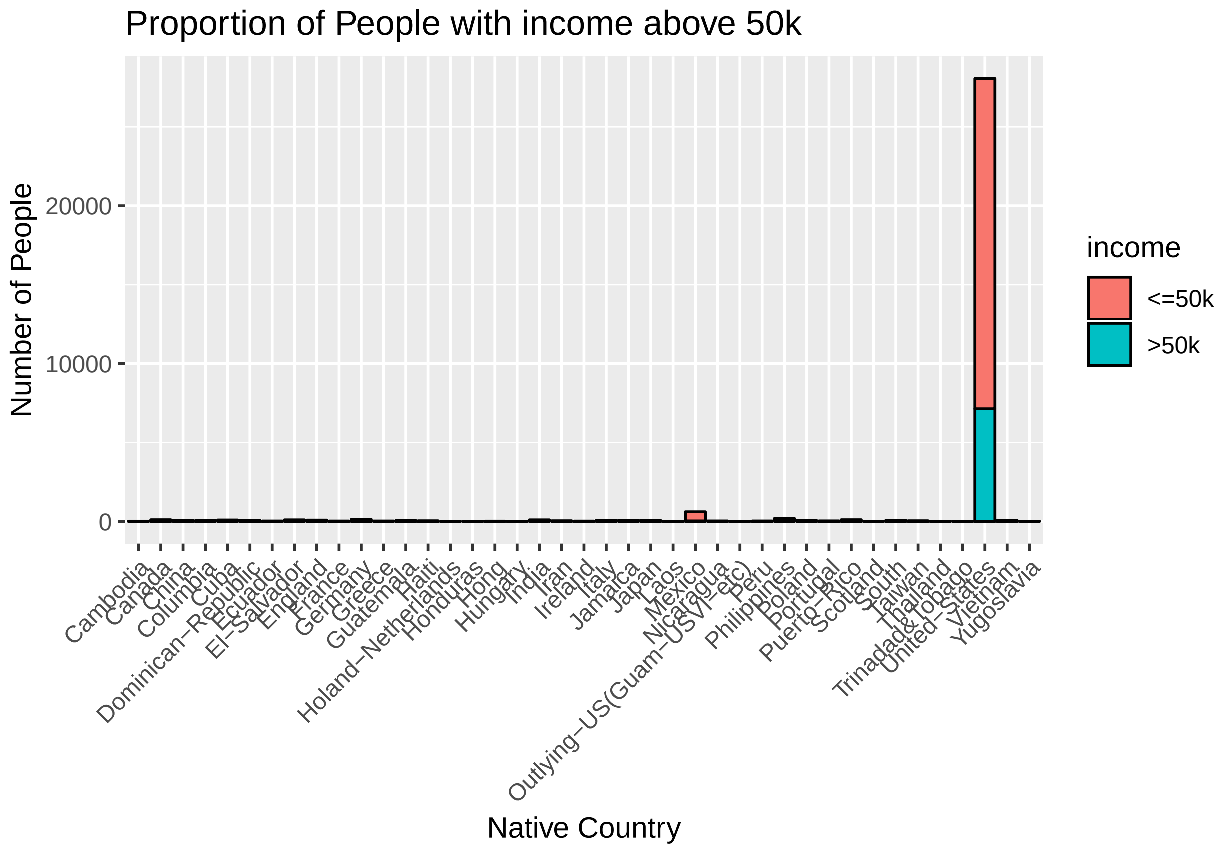
```
ggplot(data = datosAdult,aes(x=sex,fill=income))+geom_bar(color='black',position="fill")+theme(axis.tex
```



So we find that Males are more likely to have high salaries than Females.

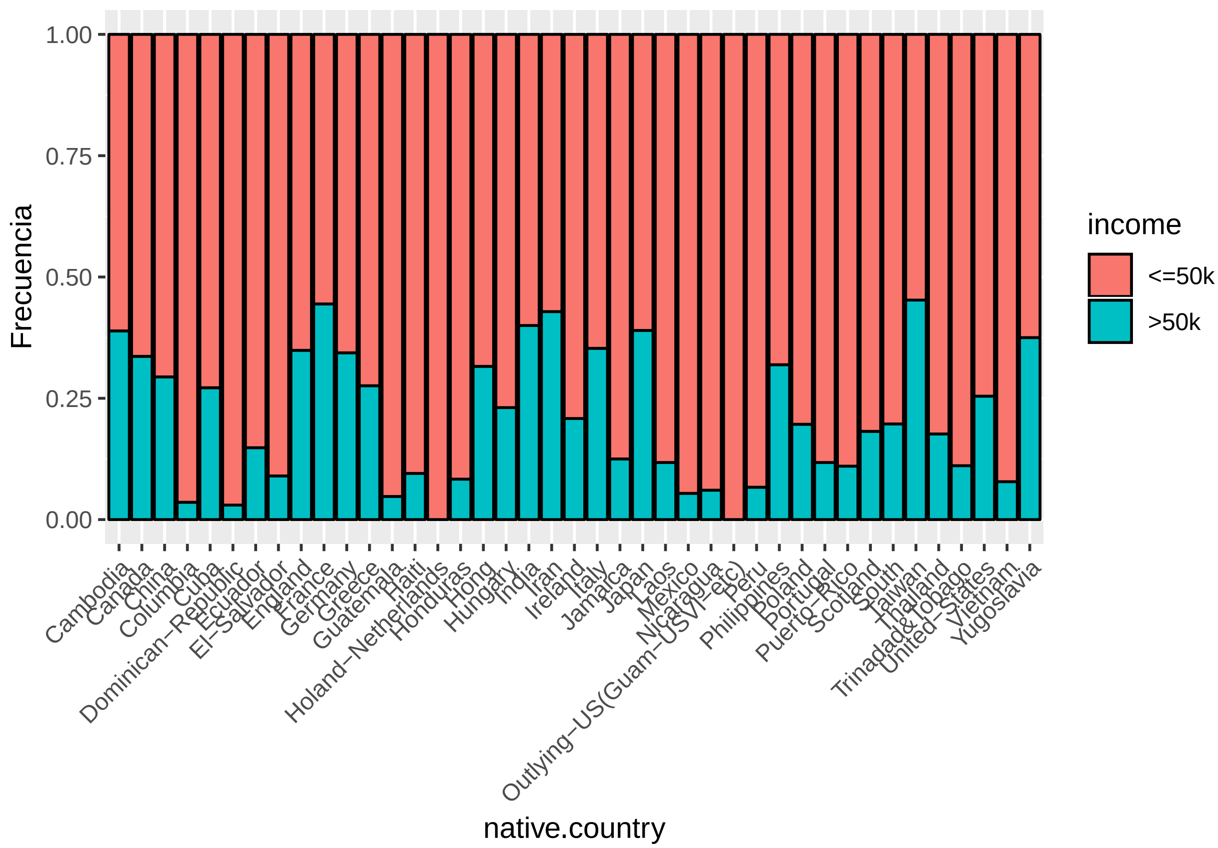
4.1.2.8 native.country

```
ggplot(datosAdult,aes(x=native.country,fill=income))+
  geom_bar(color='black')+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))+
  ggtitle('Proportion of People with income above 50k')+
  xlab("Native Country")+ylab("Number of People")
```



Let's display as a frequency plot too.

```
ggplot(data = datosAdult,aes(x=native.country,fill=income))+geom_bar(color='black',position="fill")+ther
```



There are no particularly strong variations for native.country, so it looks like we should remove the variable from the analysis dataset, but we will check with a Chi-squared test for variable dependency first of all.

4.2 Comprobación de la normalidad y homogeneidad de la varianza

Utilizamos la prueba Anderson-Darling normality test for each of the variables age, education.num, capital.gain, capital.loss, hours.per.week.

```
pvalage=ad.test(datosAdult$age)$p.value
pvaledu=ad.test(datosAdult$education.num)$p.value
pvalcapg=ad.test(datosAdult$capital.gain)$p.value
pvalcapl=ad.test(datosAdult$capital.loss)$p.value
pvalhours=ad.test(datosAdult$hours.per.week)$p.value

pvals<-matrix(c(pvalage,pvaledu,pvalcapg,pvalcapl,pvalhours),ncol=1, byrow=TRUE)
colnames(pvals)<-"pvalue"
rownames(pvals)<-c("age","education.num","capital.gain","capital.loss","hours.per.week")

as.table(pvals)
```

```
##                 pvalue
## age            3.7e-24
## education.num  3.7e-24
## capital.gain   3.7e-24
## capital.loss   3.7e-24
## hours.per.week 3.7e-24
```

El p.value para todas las variables es menor que 0.05 así que niguna de las variables tiene una distribución normal.

Utilizamos la prueba Fligner-Killeen paracomprobar la homogeneity de las variables.

```
fligner.test(education.num~education, data=datosAdult)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  education.num by education
## Fligner-Killeen:med chi-squared = -Inf, df = 15, p-value = 1
```

Así que veremos que education y education.num con un p-value de 1.0 (>0.05) indica que tienen variances que son homogeneas.

4.3 Aplicación de pruebas estadísticas

4.3.1 Pruebas de contraste de hipótesis

Here we will look at tests for independence of the categorical variables using Pearson's Chi-Squared test. The null hypothesis is: $H_0 : thetwovariablesareindependentinthesample$ The alternative hypothesis is: $H_A : thetwovariablesaredependentwithinthesample$

4.3.1.1 Workclass and Income

```
chisq.test(table(datosAdult$workclass, datosAdult$income))
```

```
## Warning in chisq.test(table(datosAdult$workclass, datosAdult$income)): Chi-
## squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(datosAdult$workclass, datosAdult$income)
## X-squared = 825.27, df = 6, p-value < 2.2e-16
```

The p-value is very small, so we reject the null hypothesis at the 0.05 significance level and we expect that the two variable are dependent.

4.3.1.2 Education and Income

```
chisq.test(table(datosAdult$education, datosAdult$income))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(datosAdult$education, datosAdult$income)
## X-squared = 4133.4, df = 15, p-value < 2.2e-16
```

The p-value is very small, so we reject the null hypothesis at the 0.05 significance level and we expect that the two variable are dependent.

4.3.1.3 Marital.status and Income

```r
chisq.test(table(datosAdult$marital.status, datosAdult$income))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(datosAdult$marital.status, datosAdult$income)
## X-squared = 6164.2, df = 6, p-value < 2.2e-16
```

The p-value is very small, so we reject the null hypothesis at the 0.05 significance level and we expect that the two variable are dependent.

4.3.1.4 Occupation and Income

```r
chisq.test(table(datosAdult$occupation, datosAdult$income))
```

```
## Warning in chisq.test(table(datosAdult$occupation, datosAdult$income)):
## Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(datosAdult$occupation, datosAdult$income)
## X-squared = 3744.9, df = 13, p-value < 2.2e-16
```

The p-value is very small, so we reject the null hypothesis at the 0.05 significance level and we expect that the two variable are dependent.

4.3.1.5 Relationship and Income

```r
chisq.test(table(datosAdult$relationship, datosAdult$income))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(datosAdult$relationship, datosAdult$income)
## X-squared = 6336.7, df = 5, p-value < 2.2e-16
```

The p-value is very small, so we reject the null hypothesis at the 0.05 significance level and we expect that the two variable are dependent.

4.3.1.6 Race and Income

```r
chisq.test(table(datosAdult$race, datosAdult$income))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(datosAdult$race, datosAdult$income)
## X-squared = 314.93, df = 4, p-value < 2.2e-16
```

The p-value is very small, so we reject the null hypothesis at the 0.05 significance level and we expect that the two variable are dependent.

4.3.1.7 Sex and Income

```r
chisq.test(table(datosAdult$sex, datosAdult$income))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(datosAdult$sex, datosAdult$income)
## X-squared = 1440.4, df = 1, p-value < 2.2e-16
```

The p-value is very small, so we reject the null hypothesis at the 0.05 significance level and we expect that the two variable are dependent.

4.3.1.8 Native.country and Income

```r
chisq.test(table(datosAdult$native.country, datosAdult$income))
```

```
## Warning in chisq.test(table(datosAdult$native.country, datosAdult$income)):
## Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(datosAdult$native.country, datosAdult$income)
## X-squared = 317.76, df = 40, p-value < 2.2e-16
```

The p-value is very small, so we reject the null hypothesis at the 0.05 significance level and we expect that the two variable are dependent.

As a consequence, we see that the **income** class is dependent on all the category variables, including **native.country**, so we won't remove this from the result set.

4.3.2 Correlaciones

We will get values for the correlations of the 3 strongest variables that we plotted earlier (i.e. Age, Education.num and Hours.per.week) Now we will change the *income* factor variable to have the values 0 or 1 to represent <50k and >50k

```r
datosAdult$income <- as.numeric(datosAdult$income)-1
```

```r
corAge=cor(datosAdult$age,datosAdult$income)
corEducation.num=cor(datosAdult$education.num,datosAdult$income)
corHours=cor(datosAdult$hours.per.week,datosAdult$income)

corAge
```

```
## [1] 0.2424308
```

```r
corEducation.num
```

```
## [1] 0.3346403
```

```r
corHours
```

```
## [1] 0.2285466
```

4.3.3 Regresiones

```r
# Regresores cuantitativos con mayor coeficiente
# de correlación con respecto al income

ageV = datosAdult$age
eduV = datosAdult$education.num
hoursV = datosAdult$hours.per.week

# Regresores cualitativos
occupationV = datosAdult$occupation
maritalV = datosAdult$marital.status
nativeV = datosAdult$native.country
sexV = datosAdult$sex

# Variable a predecir
incomeV = datosAdult$income
# Generación de varios modelos
modelo1 <- lm(incomeV ~ ageV + eduV + hoursV + occupationV + maritalV, data = datosAdult)
modelo2 <- lm(incomeV ~ ageV + eduV + hoursV + occupationV + maritalV + nativeV , data = datosAdult)
modelo3 <- lm(incomeV ~ ageV + eduV + hoursV + occupationV + maritalV + nativeV + sexV , data = datosAdu

tabla.coeficientes <- matrix(c(1, summary(modelo1)$r.squared,
2, summary(modelo2)$r.squared,
3, summary(modelo3)$r.squared),
ncol = 2, byrow = TRUE)

colnames(tabla.coeficientes) <- c("Modelo", "R^2")
tabla.coeficientes
```

```
##      Modelo       R^2
## [1,]      1 0.3274863
## [2,]      2 0.3287867
## [3,]      3 0.3305613
```

En este caso e modelo3 es el mejor fit porque tiene un mayor coeficiente de determinación.

Generamos el conjunto de datos para hacer más modelos.

```r
write.csv(datosAdult, file = "datosAdult_out.csv", row.names=FALSE)
```

5 Representación de los resultados

As presented graphically and in tables above, we have seen that several variable factors strongly influence the model generation. We have seen the correlations between numerical and categorical variables and the class of income. It seems likely that this dataset can be used to predict income class given the set of variables available.

Although the normality of the variable distributions is proven to be not normal, we can still apply statistical analysis because the sample size is so large (much greater than 30 records).

6 Resolucíon del problema

We have created a simple regression model that allows for fitting the data and making income predictions for a new set of data. In practice this would be a starting point in order to obtain a much better model using other techniques. In this case, because of the large number of categorical variables as well as numerical variables, the best modelling option may be to use decision trees.