Minería de datos: Exploratory Data Analysis

Autor: Nombre estudiante

Octubre 2020

Contents

Introducción

Presentación

Esta prueba de evaluación continuada cubre el módulo 1,2 y 8 del programa de la asignatura.

Competencias

Las competencias que se trabajan en esta prueba son:

- Uso y aplicación de las TIC en el ámbito académico y profesional
- Capacidad para innovar y generar nuevas ideas.
- Capacidad para evaluar soluciones tecnológicas y elaborar propuestas de proyectos teniendo en cuenta los recursos, las alternativas disponibles y las condiciones de mercado.
- Conocer las tecnologías de comunicaciones actuales y emergentes, así como saberlas aplicar convenientemente para diseñar y desarrollar soluciones basadas en sistemas y tecnologías de la información.
- Aplicación de las técnicas específicas de ingeniería del software en las diferentes etapas del ciclo de vida de un proyecto.
- Capacidad para aplicar las técnicas específicas de tratamiento, almacenamiento y administración de datos.
- Capacidad para proponer y evaluar diferentes alternativas tecnológicas para resolver un problema concreto.
- Capacidad de utilizar un lenguaje de programación.
- Capacidad para desarrollar en una herramienta IDE.
- Capacidad de plantear un proyecto de minería de datos.

Objetivos

- Asimilar correctamente el módulo 1 y 2.
- Qué es y qué no es MD.
- Ciclo de vida de los proyectos de MD.
- Diferentes tipologías de MD.
- Conocer las técnicas propias de una fase de preparación de datos y objetivos a alcanzar.

Descripción de la PEC a realizar

La prueba está estructurada en 1 ejercicio teórico/práctico y 1 ejercicio práctico que pide que se desarrolle la fase de preparación en un juego de datos.
Deben responderse todos los ejercicios para poder superar la PEC.

Recursos

Para realizar esta práctica recomendamos la lectura de los siguientes documentos:

- Módulo 1, 2 y 8 del material didáctico.
- RStudio Cheat Sheet: Disponible en el aula Laboratorio de Minería de datos.
- R Base Cheat Sheet: Disponible en el aula Laboratorio de Minería de datos.

Criterios de evaluación

Ejercicios teóricos
Todos los ejercicios deben ser presentados de forma razonada y clara, especificando todos y cada uno de

los pasos que se hayan llevado a cabo para su resolución. No se aceptará ninguna respuesta que no esté claramente justificada.

Ejercicios prácticos
Para todas las PEC es necesario documentar en cada apartado del ejercicio práctico qué se ha hecho y cómo se ha hecho.

Formato y fecha de entrega

El formato de entrega es: usernameestudiant-PECn.html y rmd
Fecha de Entrega: 28/10/2020
Se debe entregar la PEC en el buzón de entregas del aula

Nota: Propiedad intelectual

> A menudo es inevitable, al producir una obra multimedia, hacer uso de recursos creados por terceras personas. Es por lo tanto comprensible hacerlo en el marco de una práctica de los estudios de Informática, Multimedia y Telecomunicación de la UOC, siempre y cuando esto se documente claramente y no suponga plagio en la práctica.
>
> Por lo tanto, al presentar una práctica que haga uso de recursos ajenos, se debe presentar junto con ella un documento en qué se detallen todos ellos, especificando el nombre de cada recurso, su autor, el lugar dónde se obtuvo y su estatus legal: si la obra está protegida por el copyright o se acoge a alguna otra licencia de uso (Creative Commons, licencia GNU, GPL ...). El estudiante deberá asegurarse de que la licencia no impide específicamente su uso en el marco de la práctica. En caso de no encontrar la información correspondiente tendrá que asumir que la obra está protegida por copyright.
>
> Deberéis, además, adjuntar los ficheros originales cuando las obras utilizadas sean digitales, y su código fuente si corresponde.

Enunciado

Como ejemplo, trabajaremos con el conjunto de datos "Titanic" que recoge datos sobre el famoso crucero y sobre el que es fácil realizar tareas de clasificación predictiva sobre la variable "Survived".

De momento dejaremos para las siguientes prácticas el estudio de algoritmos predictivos y nos centraremos por ahora en el estudio de las variables de una muestra de datos, es decir, haremos un trabajo descriptivo del mismo.

Las actividades que llevaremos a cabo en esta práctica suelen enmarcarse en las fases iniciales de un proyecto de minería de datos y consisten en la selección de características o variables y la preparación del los datos para posteriormente ser consumido por un algoritmo.

Las técnicas que trabajaremos son las siguientes:

1. Normalización

2. Discretización
3. Gestión de valores nulos
4. Estudio de correlaciones
5. Reducción de la dimensionalidad
6. Análisis visual del conjunto de datos

Ejemplo de estudio visual con el juego de datos Titanic

Procesos de limpieza del conjunto de datos

Primer contacto con el conjunto de datos, visualizamos su estructura.

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


# Cargamos el fichero de datos
totalData <- read.csv('titanic.csv',stringsAsFactors = FALSE)
filas=dim(totalData)[1]

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

```
## 'data.frame':    2207 obs. of  11 variables:
##  $ name    : chr  "Abbing, Mr. Anthony" "Abbott, Mr. Eugene Joseph" "Abbott, Mr. Rossmore Edward" "Al
##  $ gender  : chr  "male" "male" "male" "female" ...
##  $ age     : num  42 13 16 39 16 25 30 28 27 20 ...
##  $ class   : chr  "3rd" "3rd" "3rd" "3rd" ...
##  $ embarked: chr  "S" "S" "S" "S" ...
##  $ country : chr  "United States" "United States" "United States" "England" ...
##  $ ticketno: int  5547 2673 2673 2673 348125 348122 3381 3381 2699 3101284 ...
##  $ fare    : num  7.11 20.05 20.05 20.05 7.13 ...
##  $ sibsp   : int  0 0 1 1 0 0 1 1 0 0 ...
##  $ parch   : int  0 2 1 1 0 0 0 0 0 0 ...
##  $ survived: chr  "no" "no" "no" "yes" ...
```

Descripción de las variables contenidas en el fichero:

name a string with the name of the passenger.

gender a factor with levels male and female.

age a numeric value with the persons age on the day of the sinking. The age of babies (under 12 months) is given as a fraction of one year (1/month).

class a factor specifying the class for passengers or the type of service aboard for crew members.

embarked a factor with the persons place of of embarkment.

country a factor with the persons home country.

ticketno a numeric value specifying the persons ticket number (NA for crew members).

fare a numeric value with the ticket price (NA for crew members, musicians and employees of the shipyard company).

sibsp an ordered factor specifying the number if siblings/spouses aboard; adopted from Vanderbild data set.

parch an ordered factor specifying the number of parents/children aboard; adopted from Vanderbild data set.

survived a factor with two levels (no and yes) specifying whether the person has survived the sinking.

Mostramos estadísticas bàsicas y después trabajamos los atributos con valores vacíos.

```
#Estadísticas básicas
summary(totalData)
```

```
##      name              gender              age             class
##  Length:2207        Length:2207        Min.   : 0.1667   Length:2207
##  Class :character   Class :character   1st Qu.:22.0000   Class :character
##  Mode  :character   Mode  :character   Median :29.0000   Mode  :character
##                                        Mean   :30.4367
##                                        3rd Qu.:38.0000
##                                        Max.   :74.0000
##                                        NA's   :2
##    embarked           country              ticketno            fare
##  Length:2207        Length:2207        Min.   :      2   Min.   :  3.030
##  Class :character   Class :character   1st Qu.:  14262   1st Qu.:  7.181
##  Mode  :character   Mode  :character   Median : 111427   Median : 14.090
##                                        Mean   : 284216   Mean   : 33.405
##                                        3rd Qu.: 347077   3rd Qu.: 31.061
##                                        Max.   :3101317   Max.   :512.061
##                                        NA's   :891       NA's   :916
##      sibsp            parch          survived
##  Min.   :0.0000   Min.   :0.0000   Length:2207
##  1st Qu.:0.0000   1st Qu.:0.0000   Class :character
##  Median :0.0000   Median :0.0000   Mode  :character
##  Mean   :0.4996   Mean   :0.3856
##  3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :8.0000   Max.   :9.0000
##  NA's   :900      NA's   :900
```

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

```
##     name   gender      age    class embarked  country ticketno     fare
##        0        0        2        0        0       81      891      916
##    sibsp    parch survived
##      900      900        0
```

```
colSums(totalData=="")
```

```
##     name   gender      age    class embarked  country ticketno     fare
##        0        0       NA        0        0       NA       NA       NA
##    sibsp    parch survived
##       NA       NA        0
```

```
# Tomamos valor "Desconocido" para los valores vacíos de la variable "country"
totalData$Embarked[totalData$country==""]="Desconocido"

# Tomamos la media para valores vacíos de la variable "Age"
totalData$Age[is.na(totalData$age)] <- mean(totalData$age,na.rm=T)
```

Discretizamos cuando tiene sentido y en función de las capacidades de cada variable.

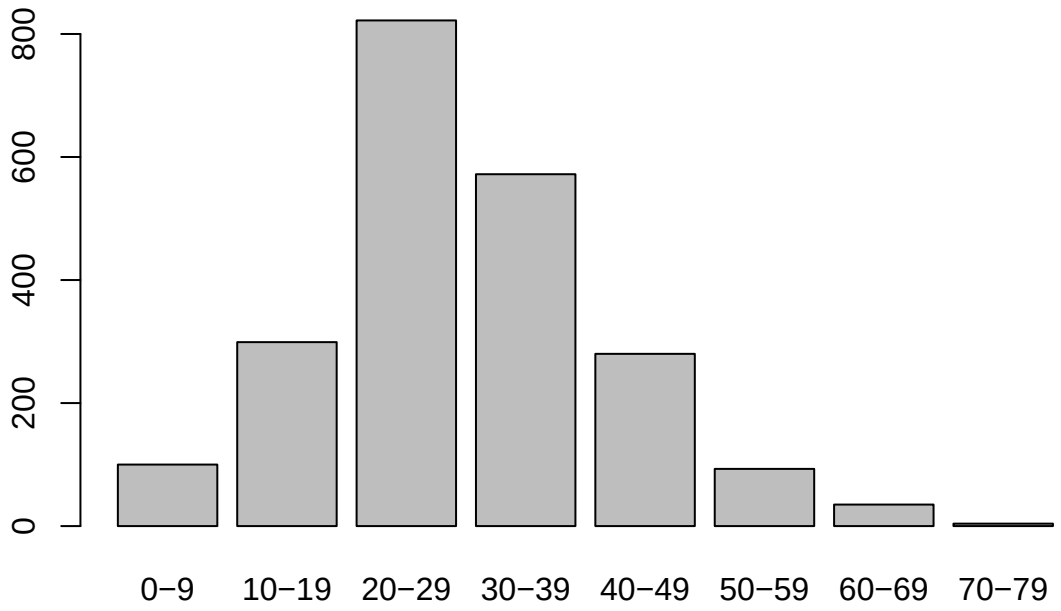
```
# Añadimos una variable nueva a los datos. Este valor es la edad discretizada con un método simple de in
# Vemos cómo se distribuyen los valore
summary(totalData[,"age"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.1667 22.0000 29.0000 30.4367 38.0000 74.0000       2
```

```
# Discretizamos
totalData["segmento_edad"] <- cut(totalData$age, breaks = c(0,10,20,30,40,50,60,70,100), labels = c("0-9
# Observamos los datos discretizados.
head(totalData)
```

```
##                             name gender age class embarked       country
## 1            Abbing, Mr. Anthony   male  42   3rd        S United States
## 2      Abbott, Mr. Eugene Joseph   male  13   3rd        S United States
## 3    Abbott, Mr. Rossmore Edward   male  16   3rd        S United States
## 4 Abbott, Mrs. Rhoda Mary 'Rosa' female  39   3rd        S       England
## 5    Abelseth, Miss. Karen Marie female  16   3rd        S        Norway
## 6 Abelseth, Mr. Olaus JÃ¸rgensen   male  25   3rd        S United States
##   ticketno  fare sibsp parch survived Embarked Age segmento_edad
## 1     5547  7.11     0     0       no     <NA>  NA         40-49
## 2     2673 20.05     0     2       no     <NA>  NA         10-19
## 3     2673 20.05     1     1       no     <NA>  NA         10-19
## 4     2673 20.05     1     1      yes     <NA>  NA         30-39
## 5   348125  7.13     0     0      yes     <NA>  NA         10-19
## 6   348122  7.13     0     0      yes     <NA>  NA         20-29
```

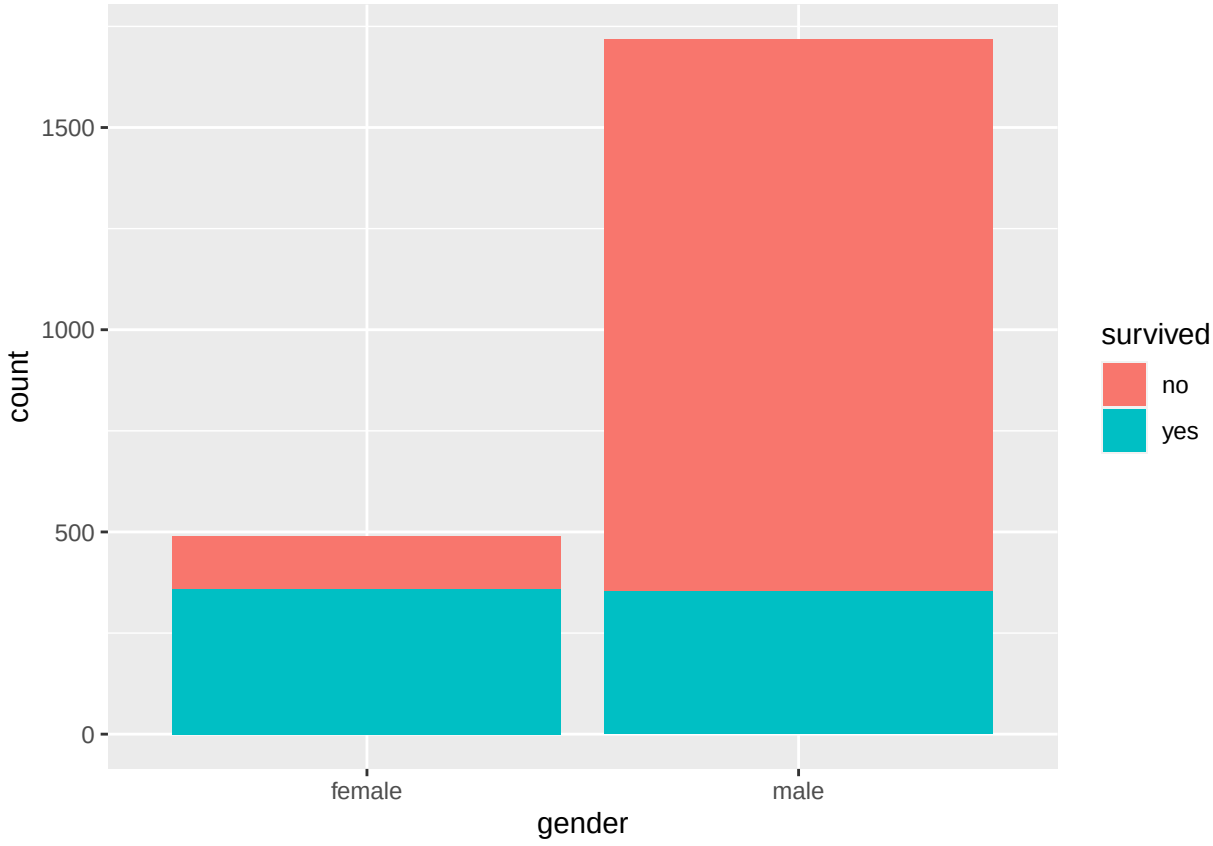
```
# Vemos como se agrupan los datos.
plot(totalData$segmento_edad)
```



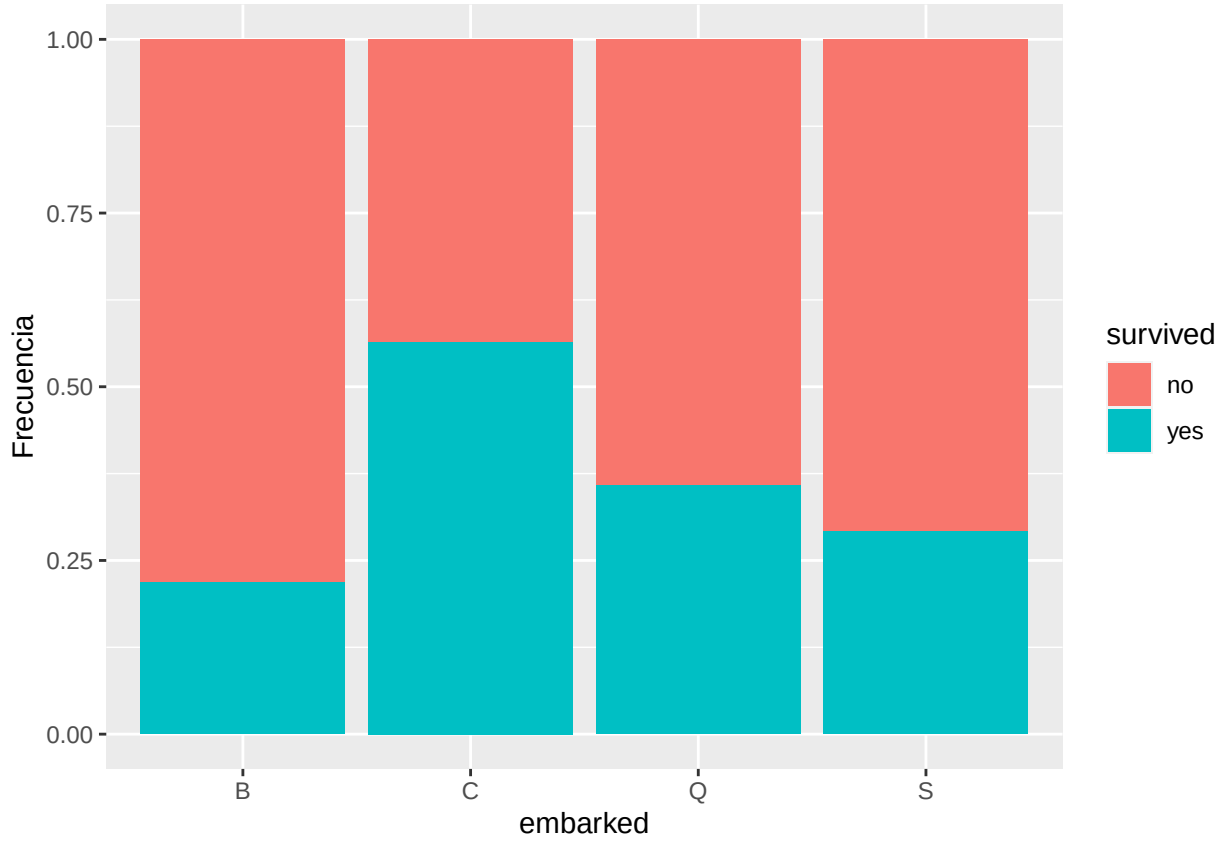
Procesos de análisis del conjunto de datos

Nos proponemos analizar las relaciones entre las diferentes variables del conjunto de datos para ver si se relacionan y como.

```
# Visualizamos la relación entre las variables "sex" y "survival":
ggplot(data=totalData[1:filas,],aes(x=gender,fill=survived))+geom_bar()
```




```
# Otro punto de vista. Survival como función de Embarked:
ggplot(data = totalData[1:filas,],aes(x=embarked,fill=survived))+geom_bar(position="fill")+ylab("Frecue
```



En la primera gráfica podemos observar fácilmente la cantidad de mujeres que viajaban respecto hombres y observar los que no sobrevivieron. Numéricamente el número de hombres y mujeres supervivientes es similar.

En la segunda gráfica de forma porcentual observamos los puertos de embarque y los porcentajes de supervivencia en función del puerto. Se podría trabajar el puerto C (Cherburgo) para ver de explicar la diferencia en los datos. Quizás porcentualmente embarcaron más mujeres o niños... O gente de primera clase?

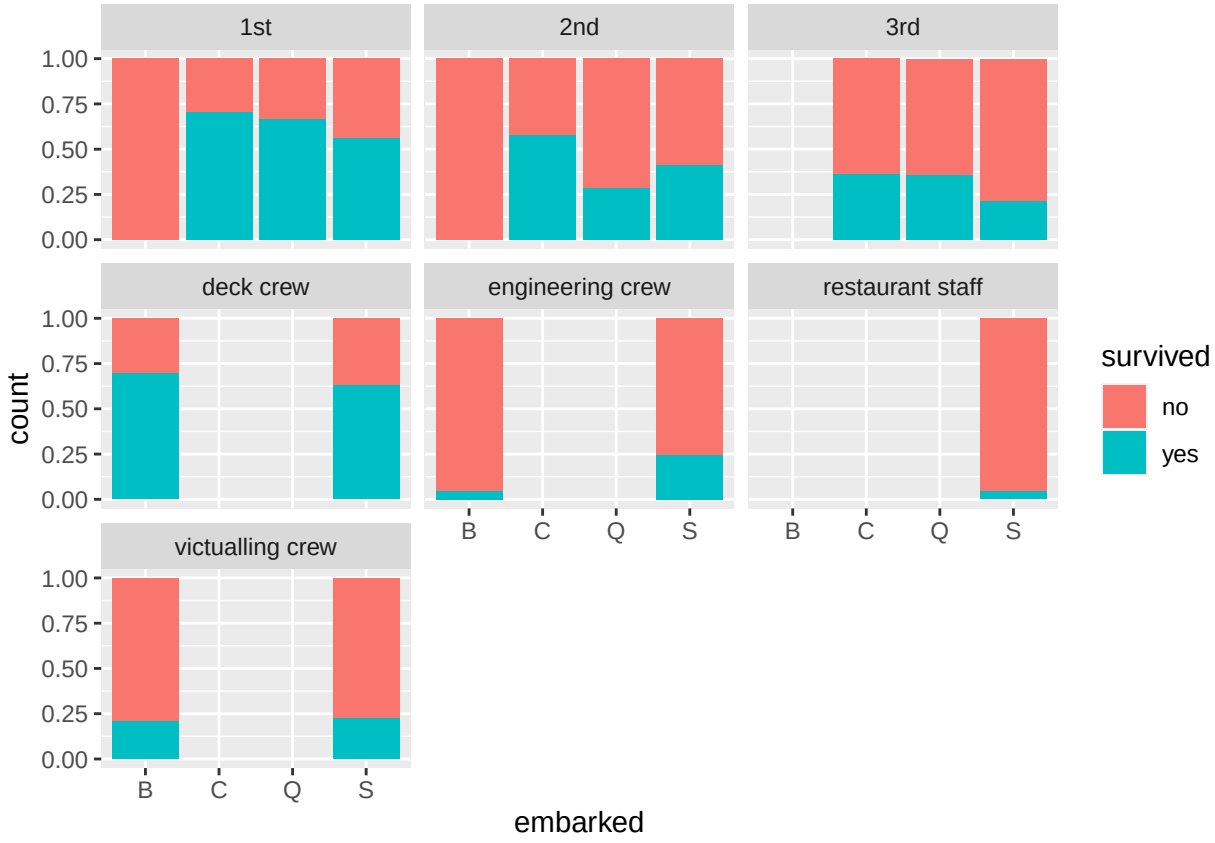
Obtenemos ahora una matriz de porcentajes de frecuencia. Vemos, por ejemplo que la probabilidad de sobrevivir si se embarcó en "C" es de un 56.45%

```
t<-table(totalData[1:filas,]$embarked, totalData[1:filas,]$survived )
for (i in 1:dim(t)[1]){
    t[i,]<-t[i,]/sum(t[i,])*100
}
t
```

```
##
##          no      yes
##   B 78.17259 21.82741
##   C 43.54244 56.45756
##   Q 64.22764 35.77236
##   S 70.85396 29.14604
```

Veamos ahora como en un mismo gráfico de frecuencias podemos trabajar con 3 variables: Embarked, Survived y Pclass.

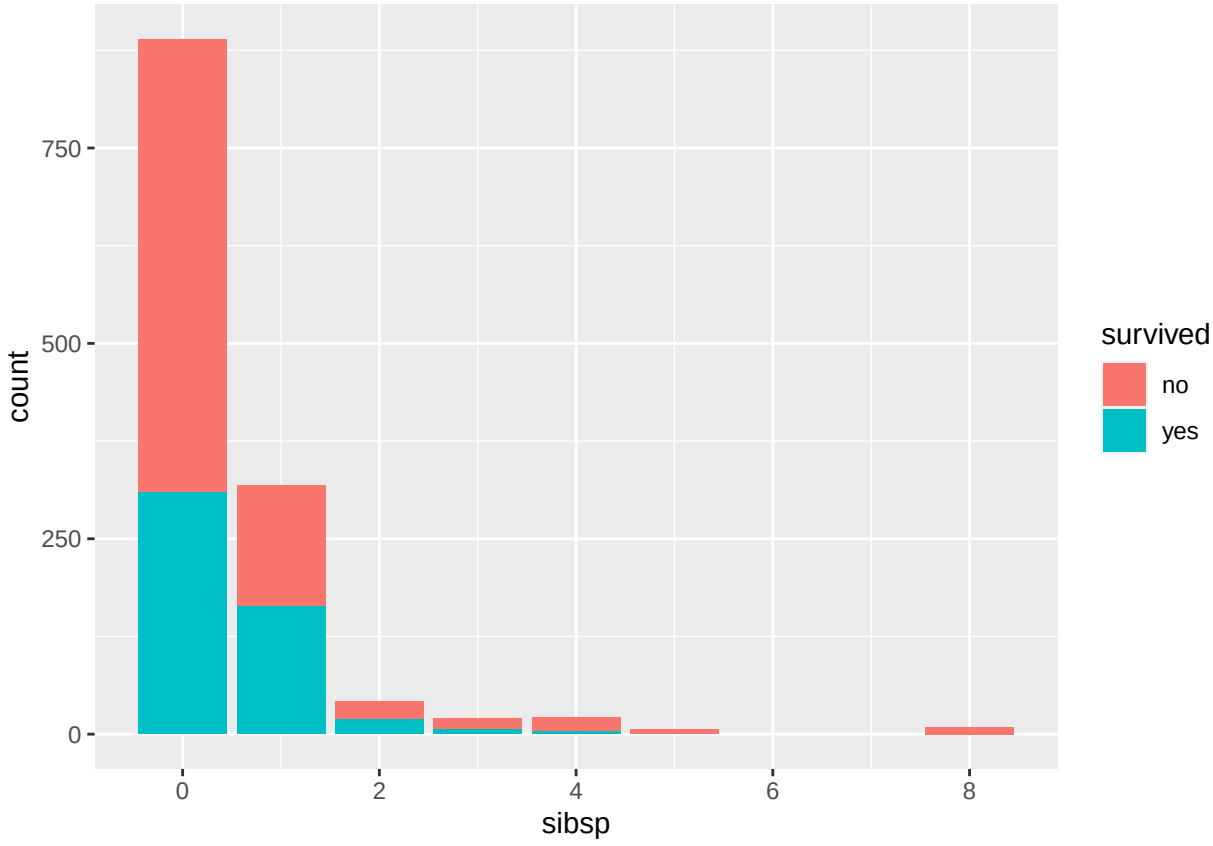
```
# Ahora, podemos dividir el gráfico de Embarked por Pclass:
ggplot(data = totalData[1:filas,],aes(x=embarked,fill=survived))+geom_bar(position="fill")+facet_wrap(~
```



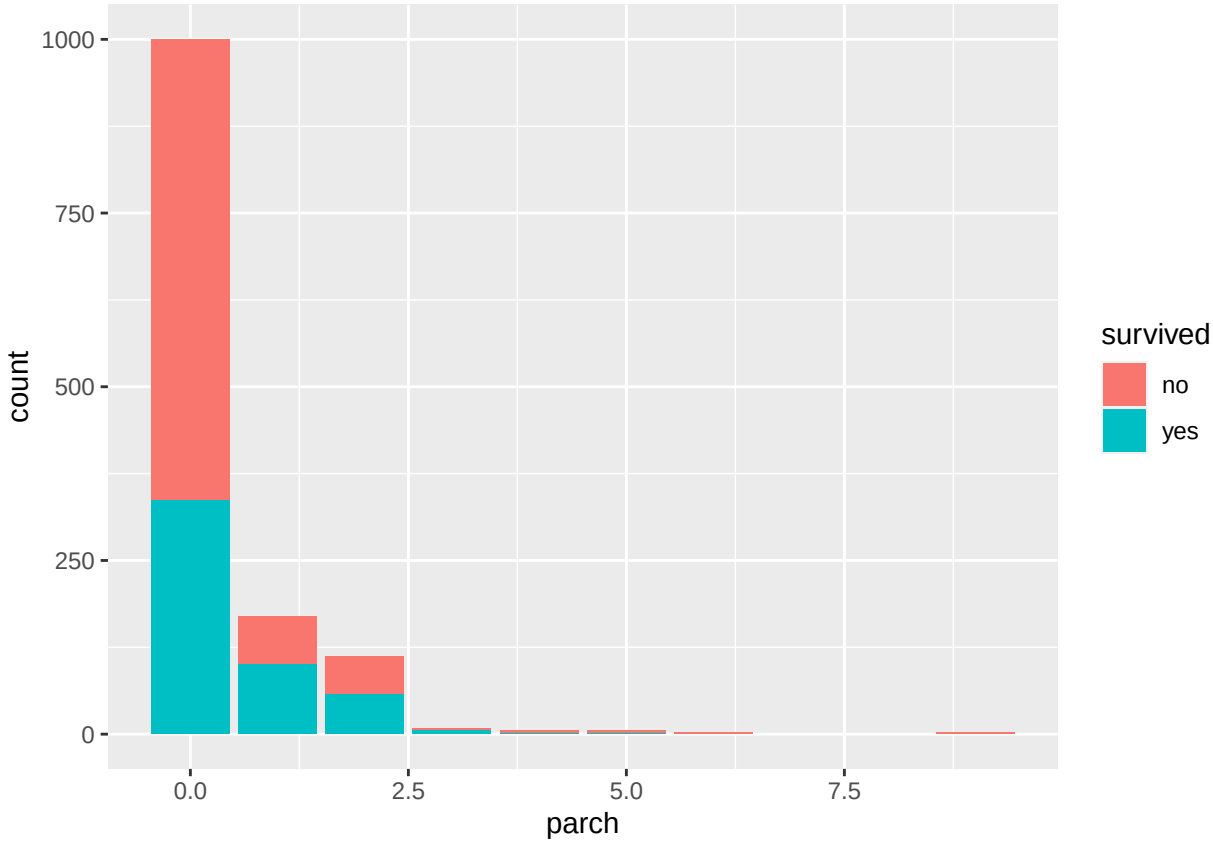
Aquí ya podemos extraer mucha información. Como propuesta de mejora se podría hacer un gráfico similar trabajando solo la clase. Habría que unificar toda la tripulación a una única categoría.

Comparemos ahora dos gráficos de frecuencias: Survived-SibSp y Survived-Parch

```
# Survivial como función de SibSp y Parch
ggplot(data = totalData[1:filas,],aes(x=sibsp,fill=survived))+geom_bar()
```



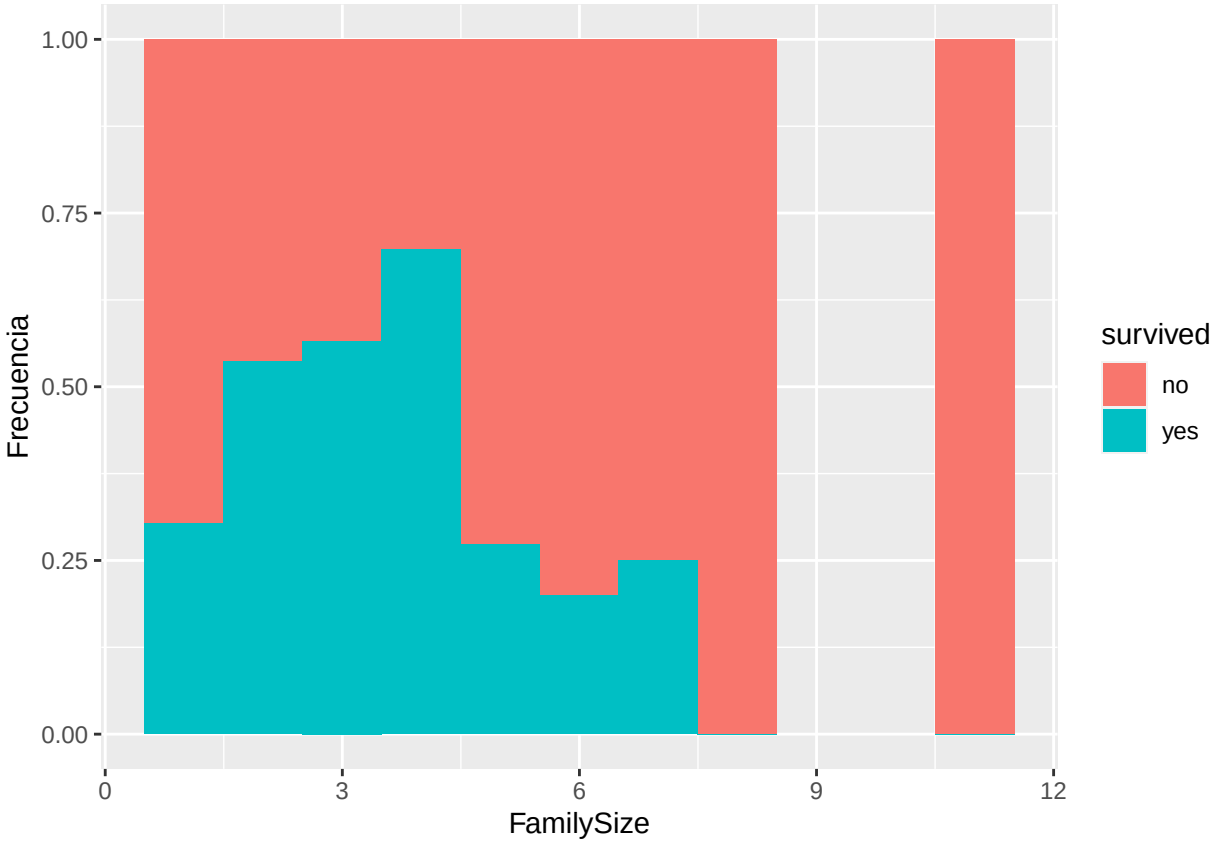

```
ggplot(data = totalData[1:filas,],aes(x=parch,fill=survived))+geom_bar()
```




```
# Vemos como las forma de estos dos gráficos es similar. Este hecho nos puede indicar presencia de corre
```

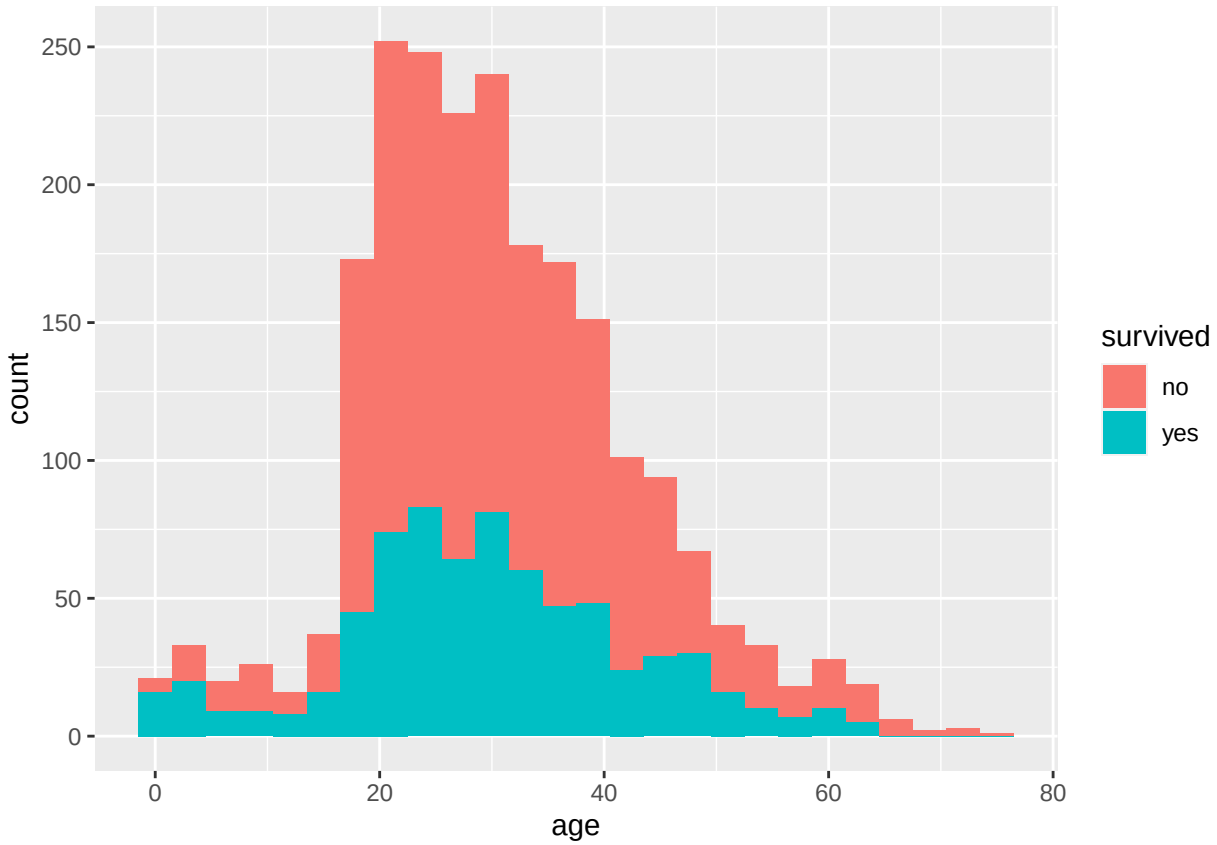
Veamos un ejemplo de construcción de una variable nueva: Tamaño de familia

```
# Construimos un atributo nuevo: family size.
totalData$FamilySize <- totalData$sibsp + totalData$parch +1;
totalData1<-totalData[1:filas,]
ggplot(data = totalData1[!is.na(totalData[1:filas,]$FamilySize),],aes(x=FamilySize,fill=survived))+geom_
```

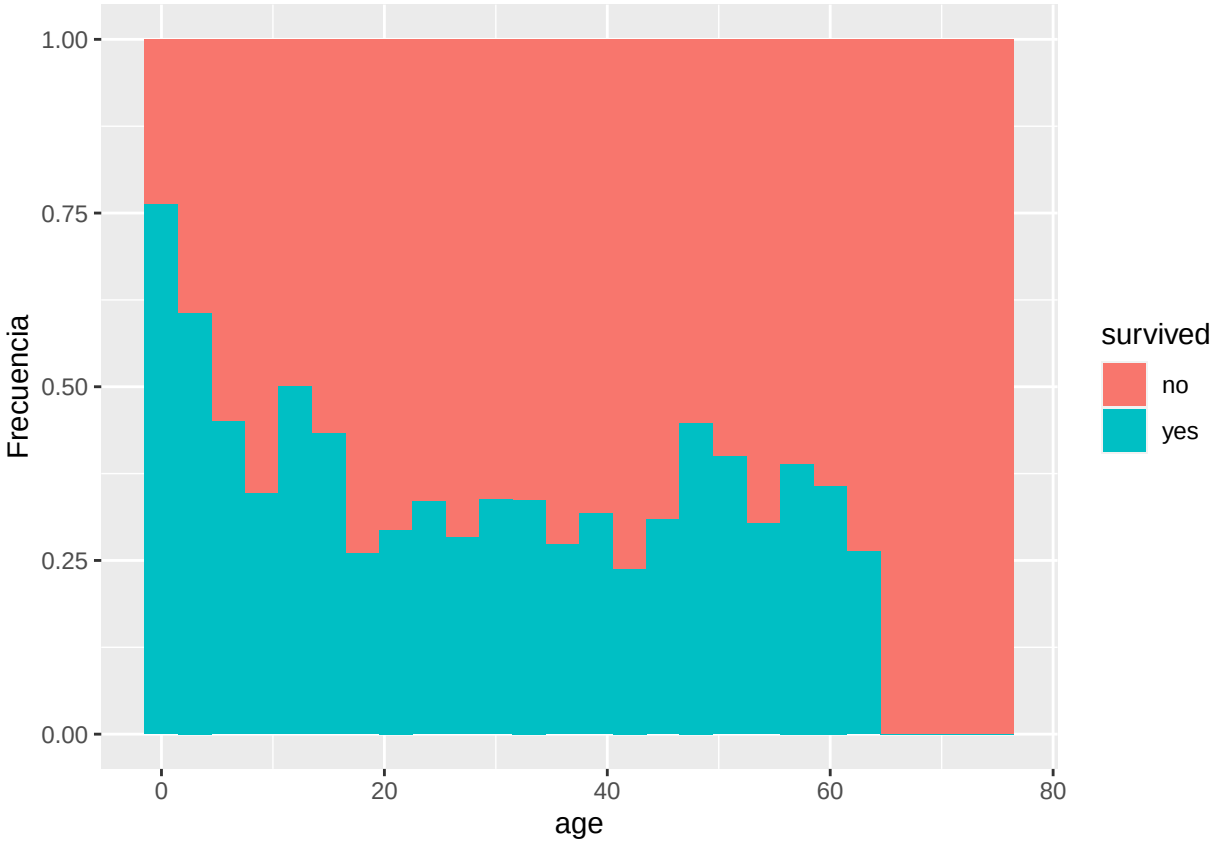


Veamos ahora dos gráficos que nos compara los atributos Age y Survived.
Observamos como el parámetro position=“fill” nos da la proporción acumulada de un atributo dentro de otro

```
# Survival como función de age:
ggplot(data = totalData1[!(is.na(totalData[1:filas,]$age)),],aes(x=age,fill=survived))+geom_histogram(bi
```




```
ggplot(data = totalData1[!is.na(totalData[1:filas,]$age),],aes(x=age,fill=survived))+geom_histogram(bin
```



Ejercicios

Ejercicio 1:

Estudia los tres casos siguientes y contesta, de forma razonada la pregunta que se realiza:

1. Disponemos de un conjunto de variables referentes a vehículos, tales como la marca, modelo, año de matriculación, etc. También se dispone del precio al que se vendieron. Al poner a la venta a un nuevo vehículo, se dispone de las variables que lo describen, pero se desconoce el precio. ¿Qué tipo de algoritmo se debería aplicar para predecir de forma automática el precio?

2. En un almacén de naranjas se tiene una máquina, que de forma automática obtiene un conjunto de variables de cada naranja, como su tamaño, acidez, grado maduración, etc. Si se desea estudiar las naranjas por tipos, según las variables obtenidas, ¿qué tipo de algoritmo es el más adecuado?

3. Un servicio de música por internet dispone de los historiales de audición de sus clientes: Qué canciones y qué grupos eligen los clientes a lo largo del tiempo de sus escuchas. La empresa desea crear un sistema que proponga la siguiente canción y grupo en función de la canción que se ha escuchado antes. ¿Qué tipo de algoritmo es el más adecuado?

Respuesta 1:

Dado que es un dominio conocido, como lo es la venta de automoviles y de que contamos con toda la informacion (atributos) correspondiente a cada uno de ellos, ademas del historial de ventas, y siendo que, lo que se intenta aqui es predecir cual sera el valor de venta actual, creo que el mejor algoritmo a aplicar es un modelo **predictivo** clasico como lo es la **regresion lineal**, ya que permite la prediccion de valores numericos (no concretos como un SI o un NO, 0 o 1, etc..) a partir una o mas variables, en este caso los distintos atributos o variables que describen a los automoviles y sus precios de venta anteriores. Tal como se describe en distintos sitios web, como ser en wikipedia: Regresion Lineal o hasta KDNuggets: Linear Regression Model, en este caso el precio de venta a predecir seria la variable dependiente (target), mientras que los atributos de los coches y precios de ventas historicos sus variables independientes (inputs). Conformando asi una regresion lineal multiple que permite explorar y cuantificar la relacion entre las variables independientes y la dependiente.

Respuesta 2:

Clustering. Ya que necesito agrupar las naranjas por sus caracteristicas sin tener informacion previa de como agruparlas segun algun criterio predefinido. Si hubiesemos querido usar clasificacion, deberiamos tener alguna informacion previa acerca de como agruparlas para luego analizar cada grupo y entender mejor que las diferencia, que caracteriza cada grupo, etc... Pero como no es el caso, la mejor alternativa son los modelos de agregacion. Como ejemplo podemos tomar lo explicado en este post acerca de lo que es clustering y algunos ejemplos de tecnicas existentes: comparing clustering techniques concise technical overview

Respuesta 3:

Este ultimo caso, tambien es una tarea de **prediccion** de la proxima accion del usuario, en particular la seleccion de la proxima cancion. Para ello lo mejor sera usar algun **sistema de recomendacion**. Tal como se describe a lo largo cientos de sitios web tenemos varias alternativas de sistemas de recomendacion, pero si nos enfocamos por ejemplo en el siguiente post de KDNuggets el algoritmo que mejor aplica a este caso es un "**Collaborative Filtering**", ya que con la informacion que contamos son las canciones que cada usuario ha escuchado historicamente y por ende tambien que canciones o bandas en comun tiene cada usuario, y no un "**Content-based Systems**" ya que no contamos por ejemplo con un ranking de popularidad de canciones calificadas por el usuario, ni ningun otro dato propio del usuario o de la cancion en si. O sea no tenemos datos propios del usuario ni de la cancion o banda. Solamente lo que cada usuario escuchó. Por lo tanto si nos basamos solo en esto: las canciones o bandas que mas han escuchado usuarios similares, seguramente tambien le gustaran a un nuevo usuario que hasta ahora ha escuchado canciones similares a los que tiene en comun.

Ejercicio 2:

A partir del conjunto de datos disponible en el siguiente enlace http://archive.ics.uci.edu/ml/datasets/Adult , realiza un estudio tomando como propuesta inicial al que se ha realizado con el conjunto de datos "Titanic". Amplia la propuesta generando nuevos indicadores o solucionando otros problemas expuestos en el módulo 2. Explica el proceso que has seguido, qué conocimiento obtienes de los datos, qué objetivo te has fijado y detalla los pasos, técnicas usadas y los problemas resueltos.

Nota: Si lo deseas puedes utilizar otro conjunto de datos propio o de algún repositorio open data siempre que sea similar en diversidad de tipos de variables al propuesto.

Respuesta 2:

```
library(ggplot2)
library(dplyr)
library(scales)

# Cargamos el juego de datos
datosAdult <- read.csv('http://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data',stri

# Asignamos nombres a las columnas a partir de lo informado en http://archive.ics.uci.edu/ml/datasets/A
names(datosAdult) <- c("age","workclass","fnlwgt","education","education-num","marital-status","occupat

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

Procesos de limpieza del conjunto de datos

```
## 'data.frame':    32561 obs. of  15 variables:
##  $ age           : int  39 50 38 53 28 37 49 52 31 42 ...
##  $ workclass     : chr  " State-gov" " Self-emp-not-inc" " Private" " Private" ...
##  $ fnlwgt        : int  77516 83311 215646 234721 338409 284582 160187 209642 45781 159449 ...
##  $ education     : chr  " Bachelors" " Bachelors" " HS-grad" " 11th" ...
##  $ education-num : int  13 13 9 7 13 14 5 9 14 13 ...
##  $ marital-status: chr  " Never-married" " Married-civ-spouse" " Divorced" " Married-civ-spouse" ...
##  $ occupation    : chr  " Adm-clerical" " Exec-managerial" " Handlers-cleaners" " Handlers-cleaners"
##  $ relationship  : chr  " Not-in-family" " Husband" " Not-in-family" " Husband" ...
##  $ race          : chr  " White" " White" " White" " Black" ...
##  $ sex           : chr  " Male" " Male" " Male" " Male" ...
##  $ capital-gain  : int  2174 0 0 0 0 0 0 0 14084 5178 ...
##  $ capital-loss  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ hour-per-week : int  40 13 40 40 40 40 16 45 50 40 ...
##  $ native-country: chr  " United-States" " United-States" " United-States" " United-States" ...
##  $ income        : chr  " <=50K" " <=50K" " <=50K" " <=50K" ...
```

Hasta aqui los valores que toman cada variable del dataset parecen ser los descriptos en la seccion "Attribute Information" de **http://archive.ics.uci.edu/ml/datasets/Adult**

Ahondemos mas en el contenido del dataset y hagamos un muestreo de 10 filas del dataset y luego estadisticas basicas:

```
#  Previsualicemos 30 registros del dataset para entender mejor el contenido
head(datosAdult,30)
```

```
##    age         workclass fnlwgt     education education-num
## 1   39         State-gov  77516     Bachelors            13
## 2   50  Self-emp-not-inc  83311     Bachelors            13
## 3   38           Private 215646       HS-grad             9
## 4   53           Private 234721          11th             7
## 5   28           Private 338409     Bachelors            13
## 6   37           Private 284582       Masters            14
## 7   49           Private 160187           9th             5
```

```
## 8   52  Self-emp-not-inc 209642       HS-grad             9
## 9   31           Private  45781       Masters            14
## 10  42           Private 159449     Bachelors            13
## 11  37           Private 280464  Some-college            10
## 12  30         State-gov 141297     Bachelors            13
## 13  23           Private 122272     Bachelors            13
## 14  32           Private 205019    Assoc-acdm            12
## 15  40           Private 121772     Assoc-voc            11
## 16  34           Private 245487       7th-8th             4
## 17  25  Self-emp-not-inc 176756       HS-grad             9
## 18  32           Private 186824       HS-grad             9
## 19  38           Private  28887          11th             7
## 20  43  Self-emp-not-inc 292175       Masters            14
## 21  40           Private 193524     Doctorate            16
## 22  54           Private 302146       HS-grad             9
## 23  35       Federal-gov  76845           9th             5
## 24  43           Private 117037          11th             7
## 25  59           Private 109015       HS-grad             9
## 26  56         Local-gov 216851     Bachelors            13
## 27  19           Private 168294       HS-grad             9
## 28  54                 ? 180211  Some-college            10
## 29  39           Private 367260       HS-grad             9
## 30  49           Private 193366       HS-grad             9
##            marital-status         occupation   relationship                race
## 1           Never-married       Adm-clerical  Not-in-family               White
## 2      Married-civ-spouse    Exec-managerial        Husband               White
## 3                Divorced  Handlers-cleaners  Not-in-family               White
## 4      Married-civ-spouse  Handlers-cleaners        Husband               Black
## 5      Married-civ-spouse     Prof-specialty           Wife               Black
## 6      Married-civ-spouse    Exec-managerial           Wife               White
## 7   Married-spouse-absent      Other-service  Not-in-family               Black
## 8      Married-civ-spouse    Exec-managerial        Husband               White
## 9           Never-married     Prof-specialty  Not-in-family               White
## 10     Married-civ-spouse    Exec-managerial        Husband               White
## 11     Married-civ-spouse    Exec-managerial        Husband               Black
## 12     Married-civ-spouse     Prof-specialty        Husband  Asian-Pac-Islander
## 13          Never-married       Adm-clerical      Own-child               White
## 14          Never-married              Sales  Not-in-family               Black
## 15     Married-civ-spouse       Craft-repair        Husband  Asian-Pac-Islander
## 16     Married-civ-spouse   Transport-moving        Husband  Amer-Indian-Eskimo
## 17          Never-married    Farming-fishing      Own-child               White
## 18          Never-married  Machine-op-inspct      Unmarried               White
## 19     Married-civ-spouse              Sales        Husband               White
## 20               Divorced    Exec-managerial      Unmarried               White
## 21     Married-civ-spouse     Prof-specialty        Husband               White
## 22              Separated      Other-service      Unmarried               Black
## 23     Married-civ-spouse    Farming-fishing        Husband               Black
## 24     Married-civ-spouse   Transport-moving        Husband               White
## 25               Divorced       Tech-support      Unmarried               White
## 26     Married-civ-spouse       Tech-support        Husband               White
## 27          Never-married       Craft-repair      Own-child               White
## 28     Married-civ-spouse                  ?        Husband  Asian-Pac-Islander
## 29               Divorced    Exec-managerial  Not-in-family               White
## 30     Married-civ-spouse       Craft-repair        Husband               White
```

```
##        sex capital-gain capital-loss hour-per-week native-country income
## 1     Male         2174            0            40  United-States  <=50K
## 2     Male            0            0            13  United-States  <=50K
## 3     Male            0            0            40  United-States  <=50K
## 4     Male            0            0            40  United-States  <=50K
## 5   Female            0            0            40           Cuba  <=50K
## 6   Female            0            0            40  United-States  <=50K
## 7   Female            0            0            16        Jamaica  <=50K
## 8     Male            0            0            45  United-States   >50K
## 9   Female        14084            0            50  United-States   >50K
## 10    Male         5178            0            40  United-States   >50K
## 11    Male            0            0            80  United-States   >50K
## 12    Male            0            0            40          India   >50K
## 13  Female            0            0            30  United-States  <=50K
## 14    Male            0            0            50  United-States  <=50K
## 15    Male            0            0            40              ?   >50K
## 16    Male            0            0            45         Mexico  <=50K
## 17    Male            0            0            35  United-States  <=50K
## 18    Male            0            0            40  United-States  <=50K
## 19    Male            0            0            50  United-States  <=50K
## 20  Female            0            0            45  United-States   >50K
## 21    Male            0            0            60  United-States   >50K
## 22  Female            0            0            20  United-States  <=50K
## 23    Male            0            0            40  United-States  <=50K
## 24    Male            0         2042            40  United-States  <=50K
## 25  Female            0            0            40  United-States  <=50K
## 26    Male            0            0            40  United-States   >50K
## 27    Male            0            0            40  United-States  <=50K
## 28    Male            0            0            60          South   >50K
## 29    Male            0            0            80  United-States  <=50K
## 30    Male            0            0            40  United-States  <=50K
```

Ahora veamos como dijimos estadisticas basicas del dataset

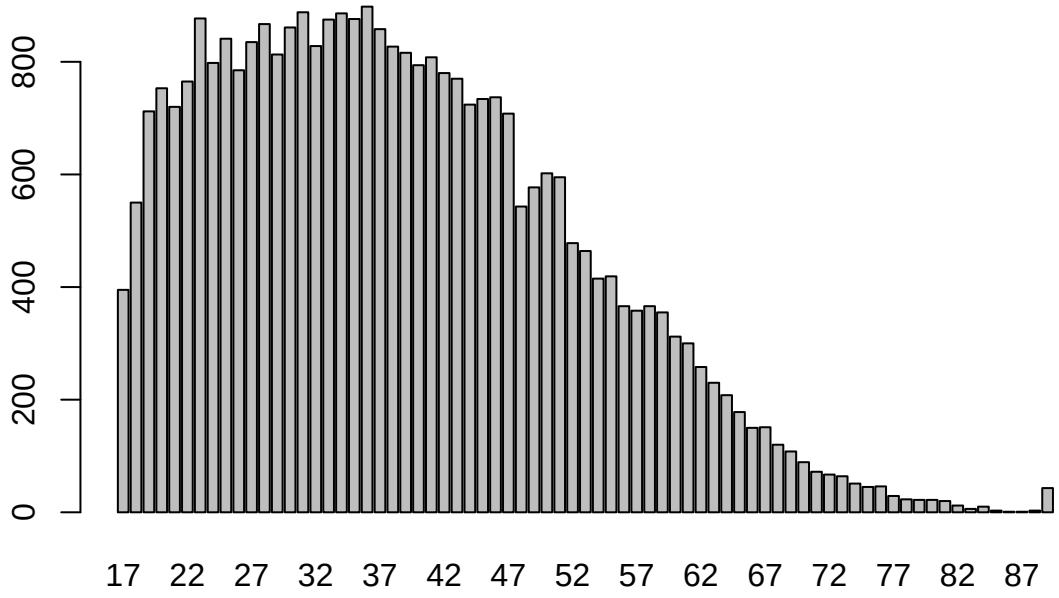
```
# Estadisticas Basicas
summary(datosAdult)
```

```
##       age         workclass            fnlwgt         education
##  Min.   :17.00   Length:32561       Min.   :  12285   Length:32561
##  1st Qu.:28.00   Class :character   1st Qu.: 117827   Class :character
##  Median :37.00   Mode  :character   Median : 178356   Mode  :character
##  Mean   :38.58                      Mean   : 189778
##  3rd Qu.:48.00                      3rd Qu.: 237051
##  Max.   :90.00                      Max.   :1484705
##  education-num   marital-status      occupation        relationship
##  Min.   : 1.00   Length:32561       Length:32561       Length:32561
##  1st Qu.: 9.00   Class :character   Class :character   Class :character
##  Median :10.00   Mode  :character   Mode  :character   Mode  :character
##  Mean   :10.08
##  3rd Qu.:12.00
##  Max.   :16.00
##      race               sex             capital-gain    capital-loss
##  Length:32561       Length:32561       Min.   :    0   Min.   :   0.0
```

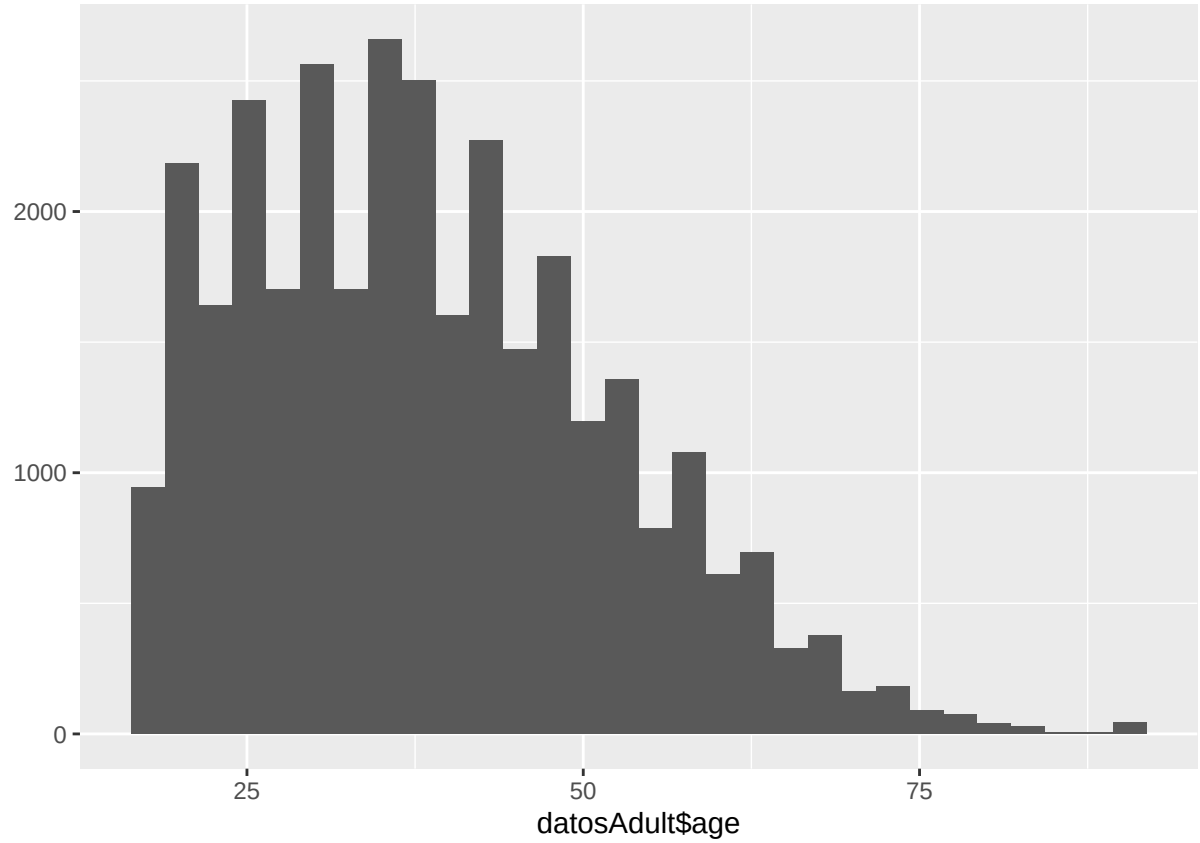
```
##  Class :character   Class :character   1st Qu.:    0   1st Qu.:   0.0
##  Mode  :character   Mode  :character   Median :    0   Median :   0.0
##                                        Mean   : 1078   Mean   :  87.3
##                                        3rd Qu.:    0   3rd Qu.:   0.0
##                                        Max.   :99999   Max.   :4356.0
##  hour-per-week   native-country        income
##  Min.   : 1.00   Length:32561       Length:32561
##  1st Qu.:40.00   Class :character   Class :character
##  Median :40.00   Mode  :character   Mode  :character
##  Mean   :40.44
##  3rd Qu.:45.00
##  Max.   :99.00
```

Y finalmente observemos todos los posibles valores de cada variable para comprender totalmente el contenido de cada variable del fichero, como tambien la existencia de valores nulos, missings o caracteres extraños.

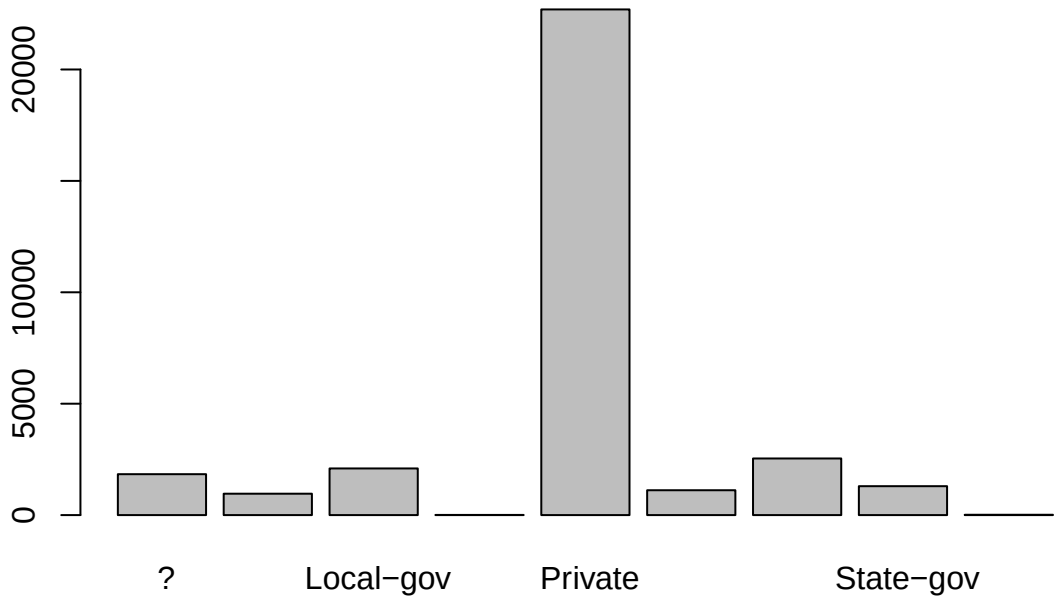
```
# Creamos barplots por cada variable
barplot(table(datosAdult$age))
```



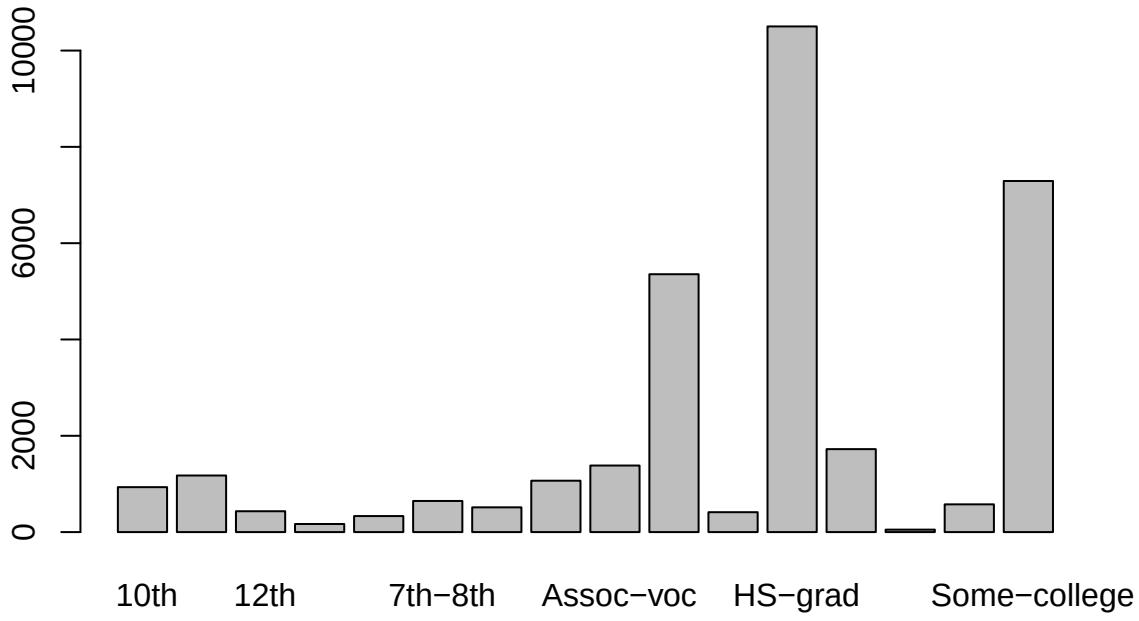
```
qplot(datosAdult$age)
```



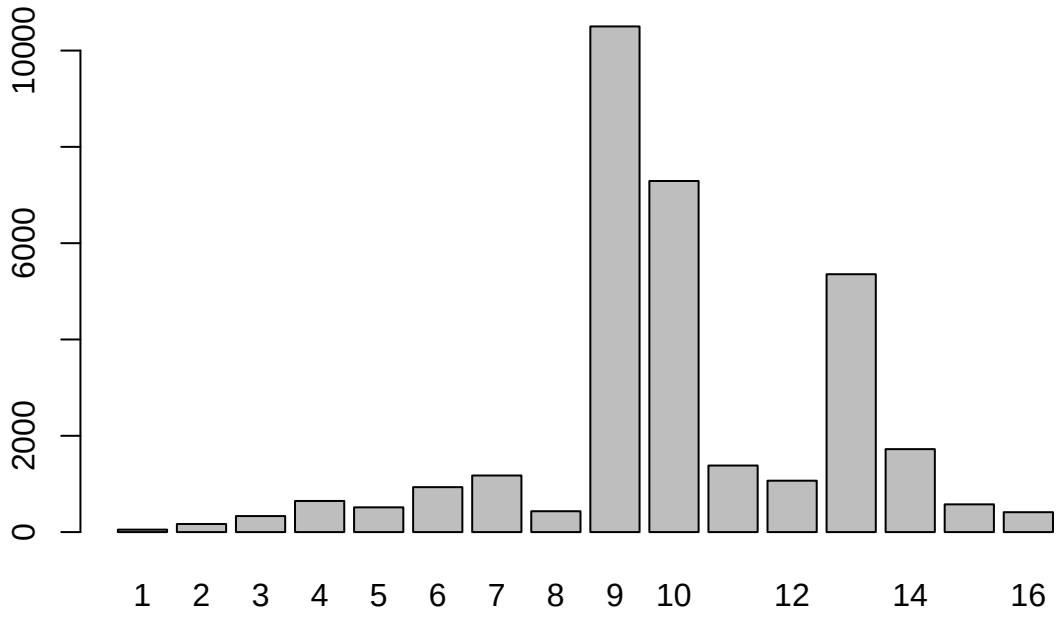

```
barplot(table(datosAdult$workclass))
```




```
barplot(table(datosAdult$education))
```

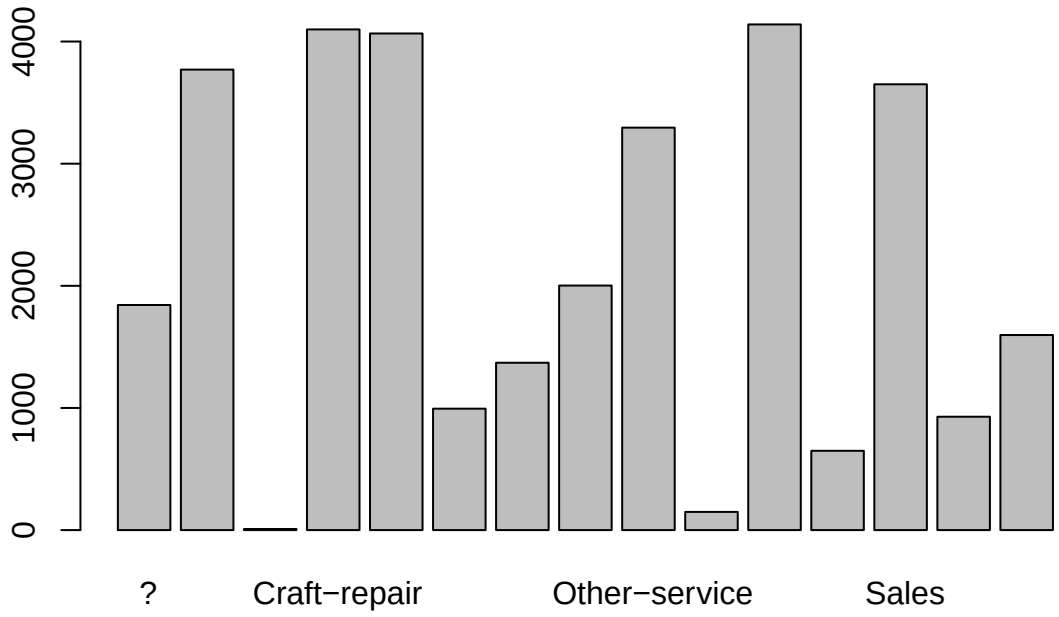



```
barplot(table(datosAdult$`education-num`))
```

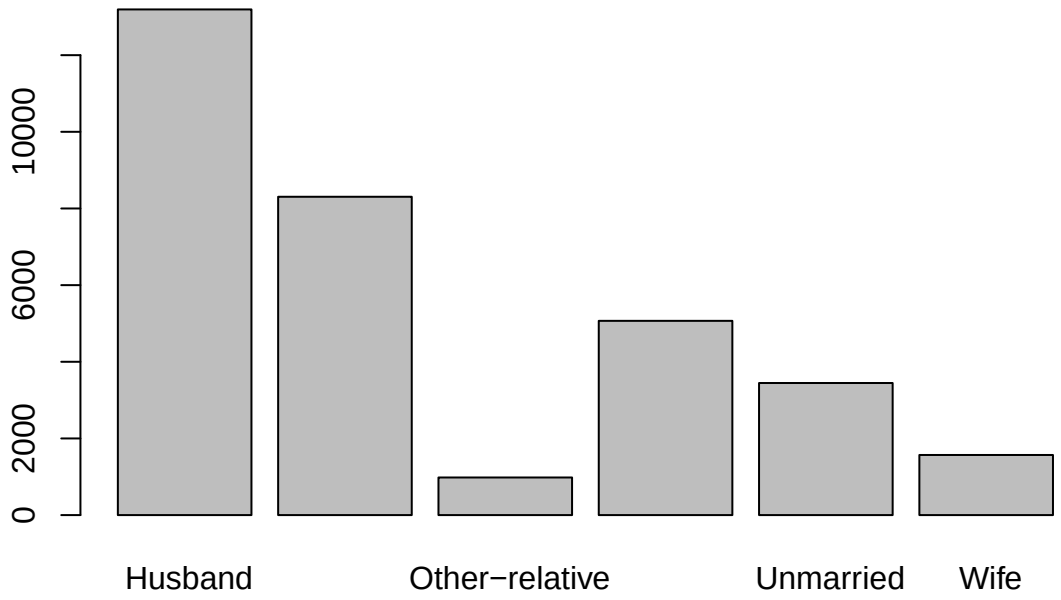



```
barplot(table(datosAdult$`marital-status`))
```

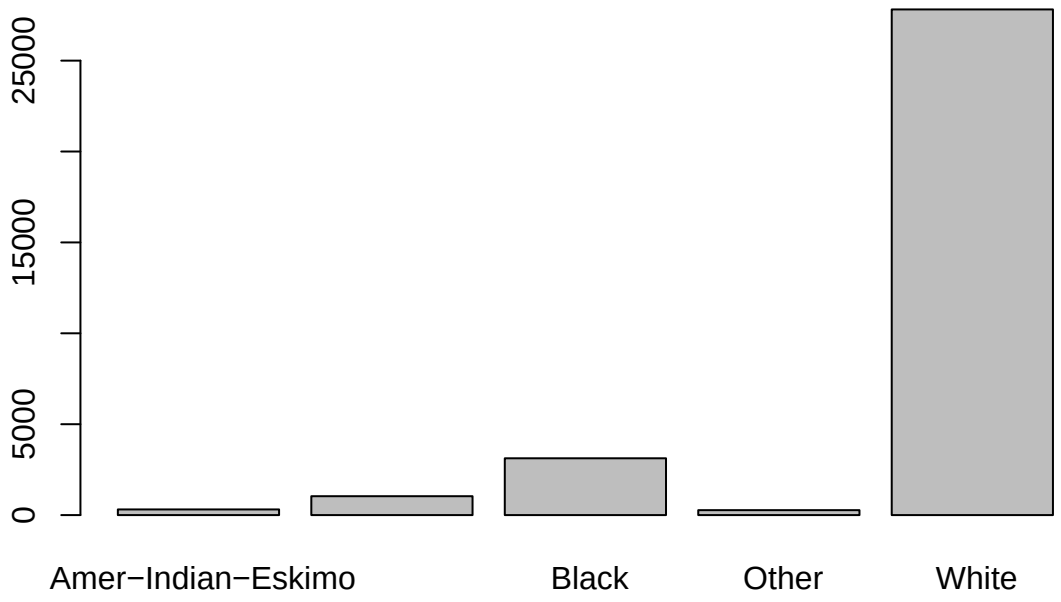

```
barplot(table(datosAdult$occupation))
```



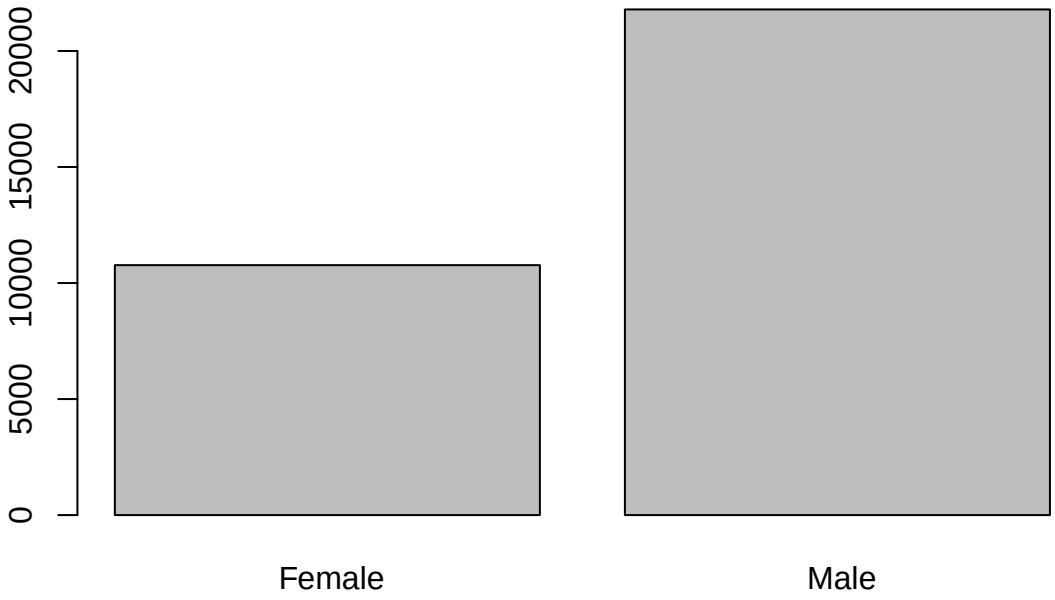

```
barplot(table(datosAdult$relationship))
```



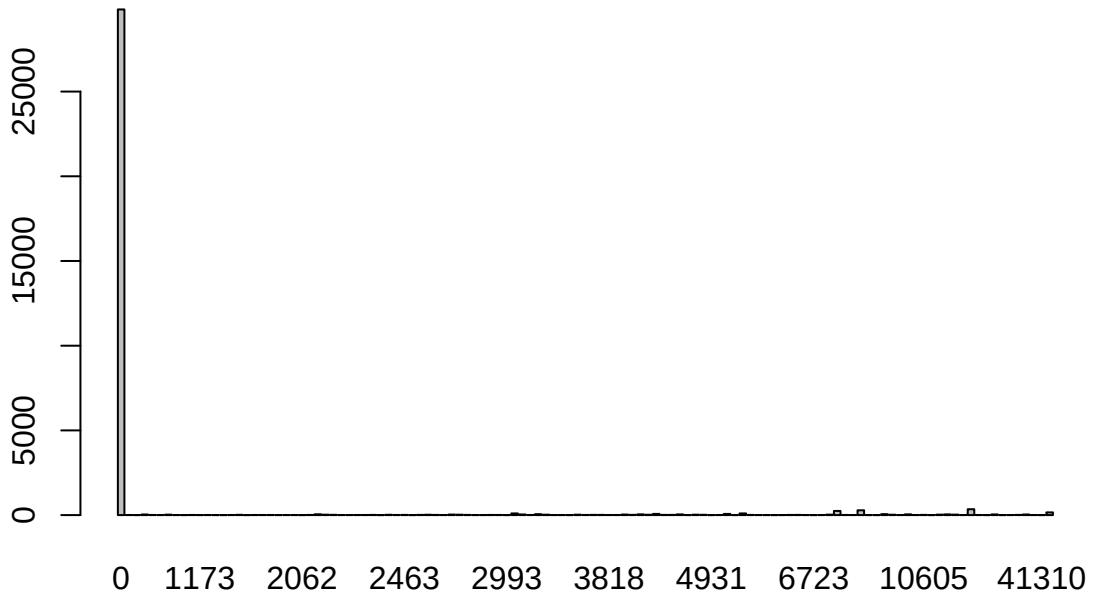

```
barplot(table(datosAdult$race))
```




```
barplot(table(datosAdult$sex))
```

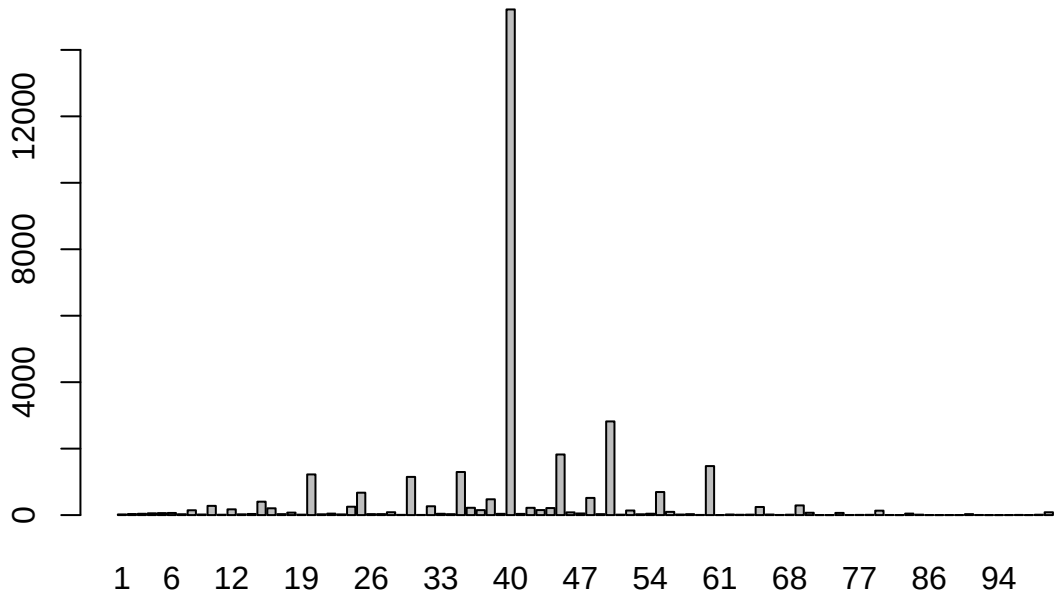



```
barplot(table(datosAdult$`capital-gain`))
```

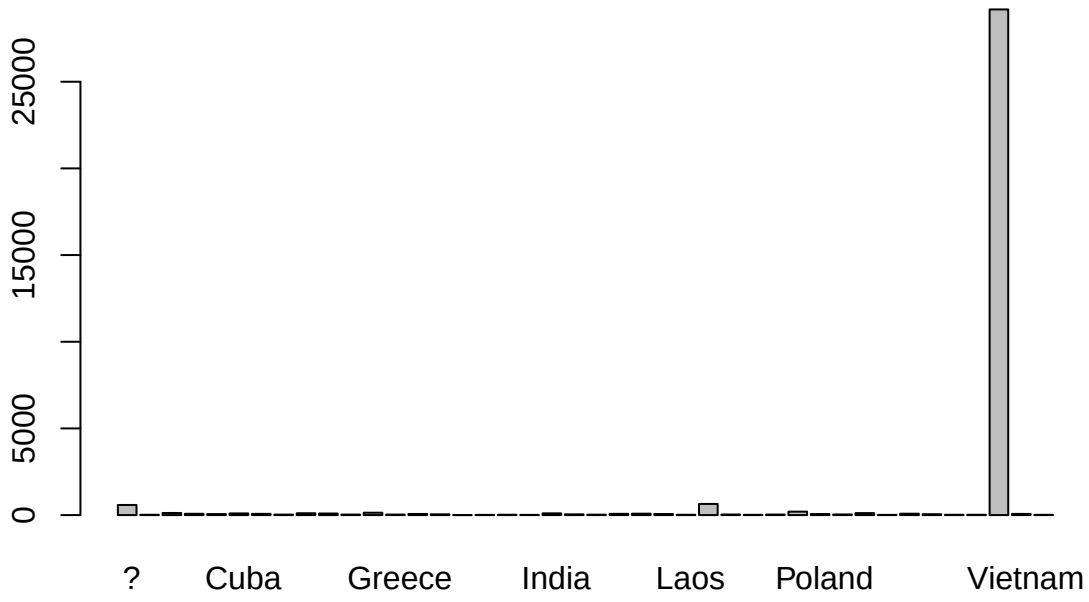



```
barplot(table(datosAdult$`capital-loss`))
```

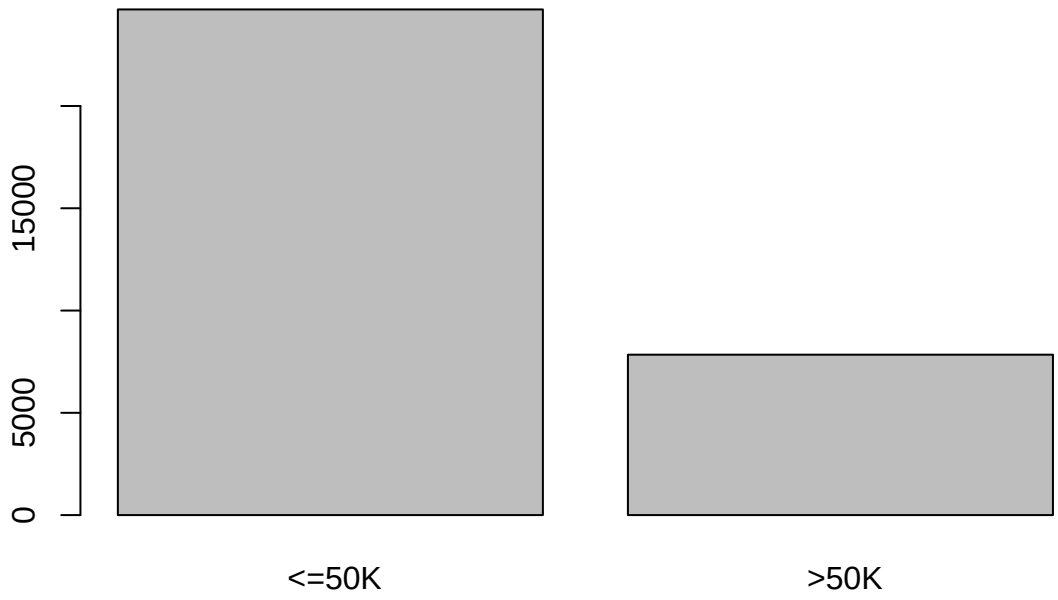

```
barplot(table(datosAdult$`hour-per-week`))
```




```
barplot(table(datosAdult$`native-country`))
```




```
barplot(table(datosAdult$income))
```




```
# 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  hour-per-week native-country         income
##              0              0              0              0              0
```

```
# y ahora los missing
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  hour-per-week native-country         income
##              0              0              0              0              0
```

Como vimos arriba no hay missings values ni nulos, pero algunas variables tiene un '?' en lugar de un valor real.

```r
# Obtenemos estadisticas de ese caracter
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  hour-per-week native-country         income
##              0              0              0              0              0
```

```r
# con colSums, seguimos sin ver ese signo de pregutna, tendra espacios?
# probemos cada variable los valores distintos para tener mas pistas:
unique(datosAdult$workclass)
```

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

```r
unique(datosAdult$occupation)
```

```
##  [1] " Adm-clerical"      " Exec-managerial"   " Handlers-cleaners"
##  [4] " Prof-specialty"    " Other-service"     " Sales"
##  [7] " Craft-repair"      " Transport-moving"  " Farming-fishing"
## [10] " Machine-op-inspct" " Tech-support"      " ?"
## [13] " Protective-serv"   " Armed-Forces"      " Priv-house-serv"
```

```r
unique(datosAdult$`native-country`)
```

```
##  [1] " United-States"              " Cuba"
##  [3] " Jamaica"                    " India"
##  [5] " ?"                          " Mexico"
##  [7] " South"                      " Puerto-Rico"
##  [9] " Honduras"                   " England"
## [11] " Canada"                     " Germany"
## [13] " Iran"                       " Philippines"
## [15] " Italy"                      " Poland"
## [17] " Columbia"                   " Cambodia"
## [19] " Thailand"                   " Ecuador"
## [21] " Laos"                       " Taiwan"
## [23] " Haiti"                      " Portugal"
## [25] " Dominican-Republic"         " El-Salvador"
## [27] " France"                     " Guatemala"
## [29] " China"                      " Japan"
## [31] " Yugoslavia"                 " Peru"
## [33] " Outlying-US(Guam-USVI-etc)" " Scotland"
## [35] " Trinadad&Tobago"            " Greece"
## [37] " Nicaragua"                  " Vietnam"
## [39] " Hong"                       " Ireland"
## [41] " Hungary"                    " Holand-Netherlands"
```

```
## como vimos arriba todos los valores incluso los que tienen valores correctos
## tiene espacios por delante o incluso por detras, esto puede verse
## incluso usando la funcion factor
levels(factor(datosAdult$workclass))
```

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

```
## una opcion para salvar esta situacion es aplicar trimws a los campos de texto, por ejemplo:
sum(trimws(datosAdult$workclass)=="?")
```

```
## [1] 1836
```

```
## entonces ahora si sacaramos estadisticas,  pero agregando manualmente el espacios, comprobaremos esta
colSums(datosAdult==" ?")
```

```
##            age      workclass         fnlwgt      education  education-num
##              0           1836              0              0              0
## marital-status     occupation   relationship           race            sex
##              0           1843              0              0              0
##   capital-gain   capital-loss  hour-per-week native-country         income
##              0              0              0            583              0
```

Existes variables missing o nulas? que podemos haces para solucionarlo?

Dado el punto inmediato anterior donde en campos de texto se encontraron espacios de mas, realicemos algunas transformaciones para mejorar los datos para posteriores analisis. En este caso convertiremos el "?" en "Not Informed", quitando los espacios ademas de todos los campos de texto

```
# primero, quitamos todos los espacios de los campos de texto
datosAdult$workclass <- trimws(datosAdult$workclass)
datosAdult$occupation <- trimws(datosAdult$occupation)
datosAdult$`native-country` <- trimws(datosAdult$`native-country`)
datosAdult$education <- trimws(datosAdult$education)
datosAdult$`marital-status` <- trimws(datosAdult$`marital-status`)
datosAdult$relationship <- trimws(datosAdult$relationship)
datosAdult$race <- trimws(datosAdult$race)

# y lo volvemos a chequear luego del cambio, al menos para las variables que poseia el signo de interro
unique(datosAdult$workclass)
```

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

```
unique(datosAdult$occupation)
```

```
##  [1] "Adm-clerical"      "Exec-managerial"   "Handlers-cleaners"
##  [4] "Prof-specialty"    "Other-service"     "Sales"
##  [7] "Craft-repair"      "Transport-moving"  "Farming-fishing"
## [10] "Machine-op-inspct" "Tech-support"      "?"
## [13] "Protective-serv"   "Armed-Forces"      "Priv-house-serv"
```

```
unique(datosAdult$`native-country`)
```

```
##  [1] "United-States"              "Cuba"
##  [3] "Jamaica"                    "India"
##  [5] "?"                          "Mexico"
##  [7] "South"                      "Puerto-Rico"
##  [9] "Honduras"                   "England"
## [11] "Canada"                     "Germany"
## [13] "Iran"                       "Philippines"
## [15] "Italy"                      "Poland"
## [17] "Columbia"                   "Cambodia"
## [19] "Thailand"                   "Ecuador"
## [21] "Laos"                       "Taiwan"
## [23] "Haiti"                      "Portugal"
## [25] "Dominican-Republic"         "El-Salvador"
## [27] "France"                     "Guatemala"
## [29] "China"                      "Japan"
## [31] "Yugoslavia"                 "Peru"
## [33] "Outlying-US(Guam-USVI-etc)" "Scotland"
## [35] "Trinadad&Tobago"            "Greece"
## [37] "Nicaragua"                  "Vietnam"
## [39] "Hong"                       "Ireland"
## [41] "Hungary"                    "Holand-Netherlands"
```

```
# volvemos a chequear los signos
colSums(datosAdult=="?")
```

```
##            age      workclass         fnlwgt      education  education-num
##              0           1836              0              0              0
## marital-status     occupation   relationship           race            sex
##              0           1843              0              0              0
##   capital-gain   capital-loss  hour-per-week native-country         income
##              0              0              0            583              0
```

```
# Ahora, reemplazamos el "?" por "Unknown"
datosAdult$workclass[datosAdult$workclass=="?"]="Unknown"
datosAdult$occupation[datosAdult$occupation=="?"]="Unknown"
datosAdult$`native-country`[datosAdult$`native-country`=="?"]="Unknown"

# volvemos a chequear los signos
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  hour-per-week native-country         income
##              0              0              0              0              0
```

```
colSums(datosAdult=="Unknown")
```

```
##            age      workclass         fnlwgt      education  education-num
##              0           1836              0              0              0
## marital-status     occupation   relationship           race            sex
##              0           1843              0              0              0
##   capital-gain   capital-loss  hour-per-week native-country         income
##              0              0              0            583              0
```

Transformemos algunos atributos para aprovechamiento posterior

```
## Intentaremos reagrupar algunos atributos, porque quizas tenga sentido para posteriores analisis, com

## Ya que si lo vimos graficamente antes, si hacemos un count de cada valor vemos que estan distribuido
table(datosAdult$workclass)
```

```
##
##      Federal-gov        Local-gov     Never-worked          Private
##              960             2093                7            22696
##     Self-emp-inc Self-emp-not-inc        State-gov          Unknown
##             1116             2541             1298             1836
##      Without-pay
##               14
```

```
## Esto podria quedar asi:
# Goverment: Federal-gov    Local-gov State-gov
# NoPay/Others: Never-worked  Without-pay Unknown
# Private
# Self-Employed: Self-emp-inc  Self-emp-not-inc

# Lo hago de esta manera, si existe alguna forma mas perfomante en R, bievenida la sugerencia :)

datosAdult$workclass[datosAdult$workclass %in% c('Federal-gov','Local-gov','State-gov')]="Goverment"

datosAdult$workclass[datosAdult$workclass%in%c('Never-worked', 'Without-pay','Unknown','')]="NoPay/Othe

datosAdult$workclass[datosAdult$workclass%in%c('Self-emp-inc','Self-emp-not-inc')]="Self-Employed"

table(datosAdult$workclass)
```

```
##
##     Goverment  NoPay/Others       Private Self-Employed
##          4351          1857         22696          3657
```

```
## Lo mismo sucede con ocupacion, me parece que podria agruparse por tipo de empleo
table(datosAdult$occupation)
```

```
##
##      Adm-clerical      Armed-Forces      Craft-repair   Exec-managerial
##              3770                 9              4099              4066
##   Farming-fishing Handlers-cleaners Machine-op-inspct     Other-service
##               994              1370              2002              3295
##   Priv-house-serv    Prof-specialty   Protective-serv             Sales
##               149              4140               649              3650
##      Tech-support  Transport-moving           Unknown
##               928              1597              1843
```

```
## Esto podria quedar asi:
# Manual_Works: Craft-repair Farming-fishing Handlers-cleaners  Transport-moving
# Admin/Proffesional: Adm-clerical Exec-managerial Machine-op-inspct Prof-specialty
# Sales: Sales
# Services: Priv-house-serv Protective-serv Other-service Tech-support
# Others: Unknown  Armed-Forces

datosAdult$occupation[datosAdult$occupation %in% c('Craft-repair','Farming-fishing','Handlers-cleaners'

datosAdult$occupation[datosAdult$occupation %in% c('Adm-clerical','Exec-managerial','Machine-op-inspct'

datosAdult$occupation[datosAdult$occupation %in% c('Priv-house-serv','Protective-serv','Other-service',

datosAdult$occupation[datosAdult$occupation %in% c('Unknown', 'Armed-Forces')]="Others"

table(datosAdult$occupation)
```

```
##
## Admin/Proffesional       Manual_Works             Others              Sales
##              13978               8060               1852               3650
##           Services
##               5021
```

Discretizemos algunas variables

```
## Como se pudo ver en las estadisticas e historgramas anteriormenete, tnto CapitalGain como CapitalLoss
datosAdult$capital_gain_flag<-ifelse(datosAdult$`capital-gain`>0,1,0)
datosAdult$capital_loss_flag<-ifelse(datosAdult$`capital-loss`>0,1,0)

# sin embargo viendo esto, parecen no aportar mucho, ya que esta todo mayormente en 0
table(datosAdult$capital_gain_flag)
```

```
##
##     0     1
## 29849  2712
```

```
table(datosAdult$capital_loss_flag)
```

```
##
##     0     1
## 31042  1519
```

```
## Para native-country haremos algo similar, debido que casi todos lo valores corresponden a USA. Separa
table(datosAdult$`native-country`)
```

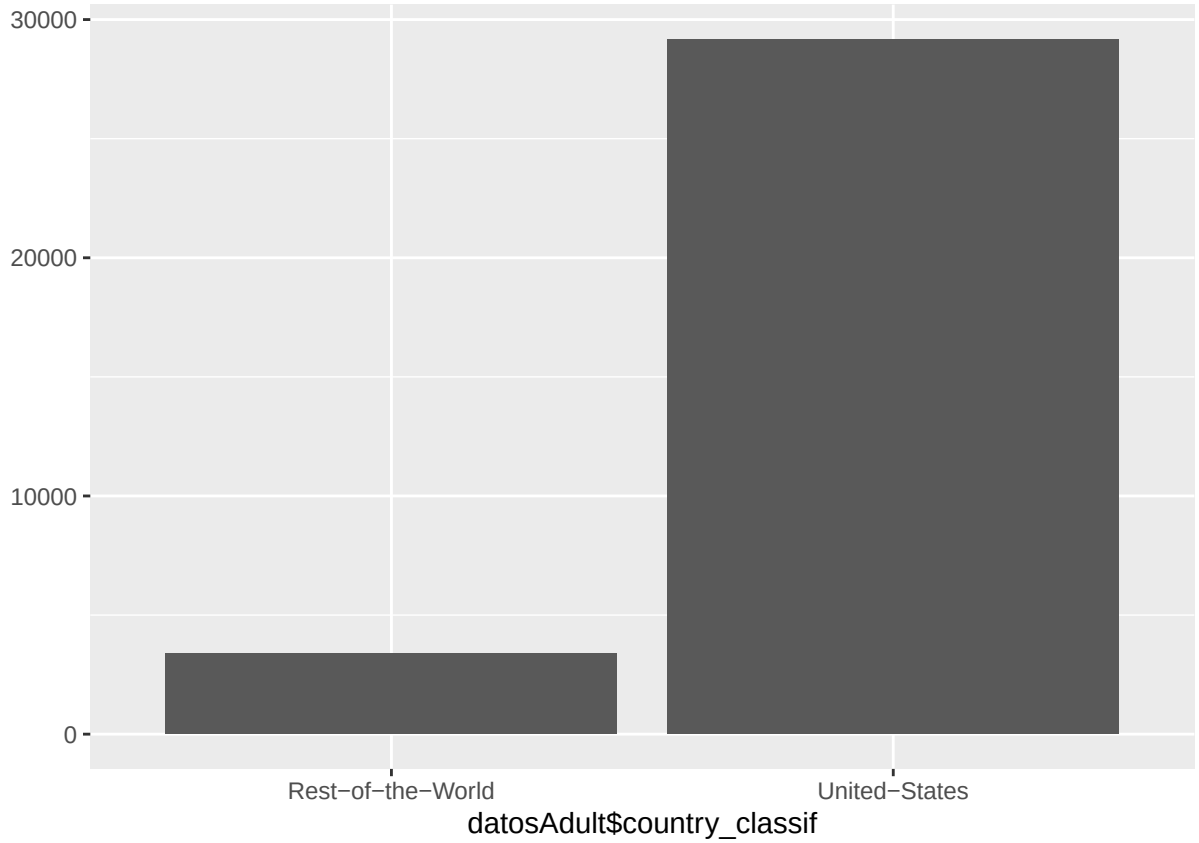
```
##
##                   Cambodia                     Canada
##                         19                        121
##                      China                   Columbia
##                         75                         59
##                       Cuba         Dominican-Republic
```

```
##                         95                         70
##                    Ecuador                El-Salvador
##                         28                        106
##                    England                     France
##                         90                         29
##                    Germany                     Greece
##                        137                         29
##                  Guatemala                      Haiti
##                         64                         44
##         Holand-Netherlands                   Honduras
##                          1                         13
##                       Hong                    Hungary
##                         20                         13
##                      India                       Iran
##                        100                         43
##                    Ireland                      Italy
##                         24                         73
##                    Jamaica                      Japan
##                         81                         62
##                       Laos                     Mexico
##                         18                        643
##                  Nicaragua Outlying-US(Guam-USVI-etc)
##                         34                         14
##                       Peru                Philippines
##                         31                        198
##                     Poland                   Portugal
##                         60                         37
##                Puerto-Rico                   Scotland
##                        114                         12
##                      South                     Taiwan
##                         80                         51
##                   Thailand            Trinadad&Tobago
##                         18                         19
##               United-States                    Unknown
##                      29170                        583
##                    Vietnam                  Yugoslavia
##                         67                         16
```

```
datosAdult$country_classif<-ifelse(datosAdult$`native-country`=='United-States','United-States','Rest-o

# sin embargo viendo esto, parecen no apotar mucho
table(datosAdult$country_classif)
```

```
##
## Rest-of-the-World     United-States
##              3391             29170
```

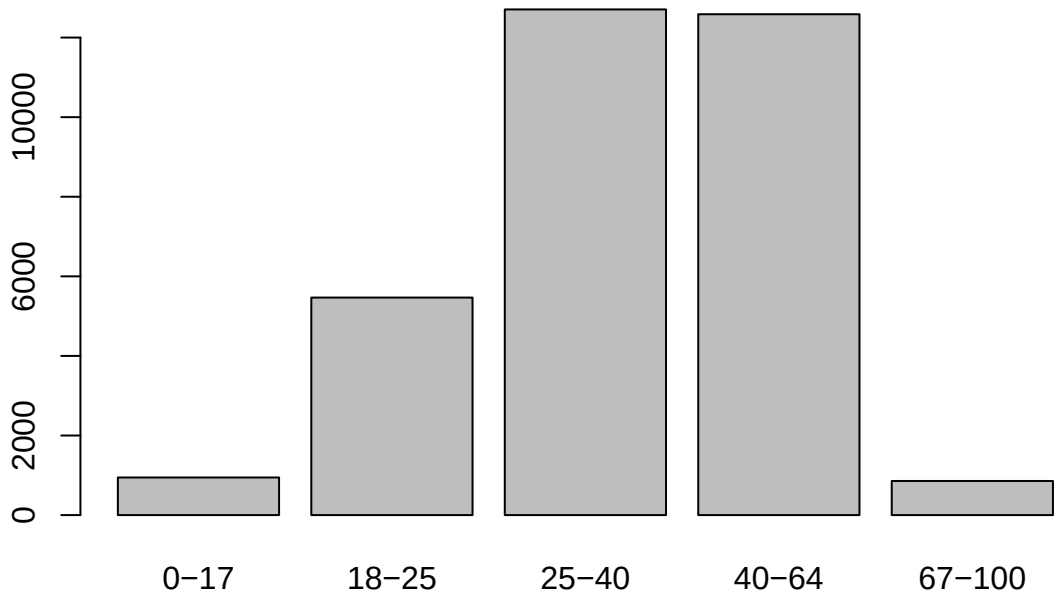
```
qplot(datosAdult$country_classif)
```




```
# Parece aportar poco esta distribucion

## Ahora discretizamos age, por los clasico rangos de edades:
## 0-17 / 18-25 / 25-40 / 40-64 / +67
datosAdult["grouped_age"] <- cut(datosAdult$age, breaks = c(0,18,25,40,67,100), labels = c("0-17", "18-
# tabularmente
table(datosAdult$grouped_age)
```

```
##
##   0-17  18-25  25-40  40-64 67-100
##    945   5466  12707  12586    857
```

```
# graficamente:
plot(datosAdult$grouped_age)
```



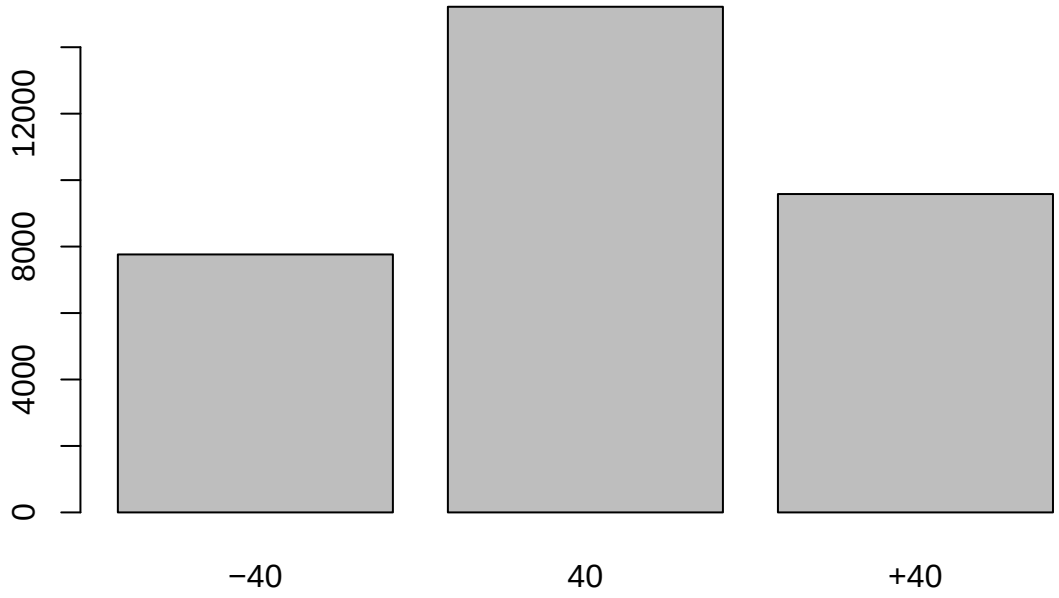

```
# Tambien se podria discretiar las horas trabajadas, para los que trabajan las clasicas 40hs semanales,
table(datosAdult$`hour-per-week`)
```

```
##
##     1     2     3     4     5     6     7     8     9    10    11    12    13
##    20    32    39    54    60    64    26   145    18   278    11   173    23
##    14    15    16    17    18    19    20    21    22    23    24    25    26
##    34   404   205    29    75    14  1224    24    44    21   252   674    30
##    27    28    29    30    31    32    33    34    35    36    37    38    39
##    30    86     7  1149     5   266    39    28  1297   220   149   476    38
##    40    41    42    43    44    45    46    47    48    49    50    51    52
## 15217    36   219   151   212  1824    82    49   517    29  2819    13   138
##    53    54    55    56    57    58    59    60    61    62    63    64    65
##    25    41   694    97    17    28     5  1475     2    18    10    14   244
##    66    67    68    70    72    73    74    75    76    77    78    80    81
##    17     4    12   291    71     2     1    66     3     6     8   133     3
##    82    84    85    86    87    88    89    90    91    92    94    95    96
##     1    45    13     2     1     2     2    29     3     1     1     2     5
##    97    98    99
##     2    11    85
```

```
datosAdult["hours_per_week_group"] <- cut(datosAdult$`hour-per-week`, breaks = c(0,39,40,100), labels =

# tabularmente
table(datosAdult$hours_per_week_group)
```

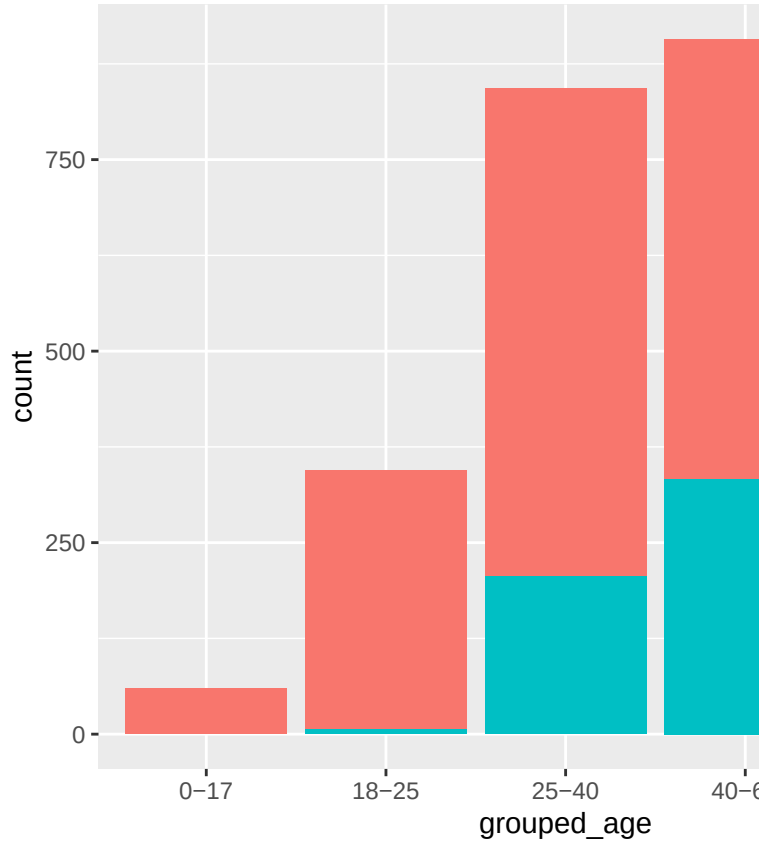
```
##
##   -40    40   +40
##  7763 15217  9581
```

```
# graficamente:
plot(datosAdult$hours_per_week_group)
```



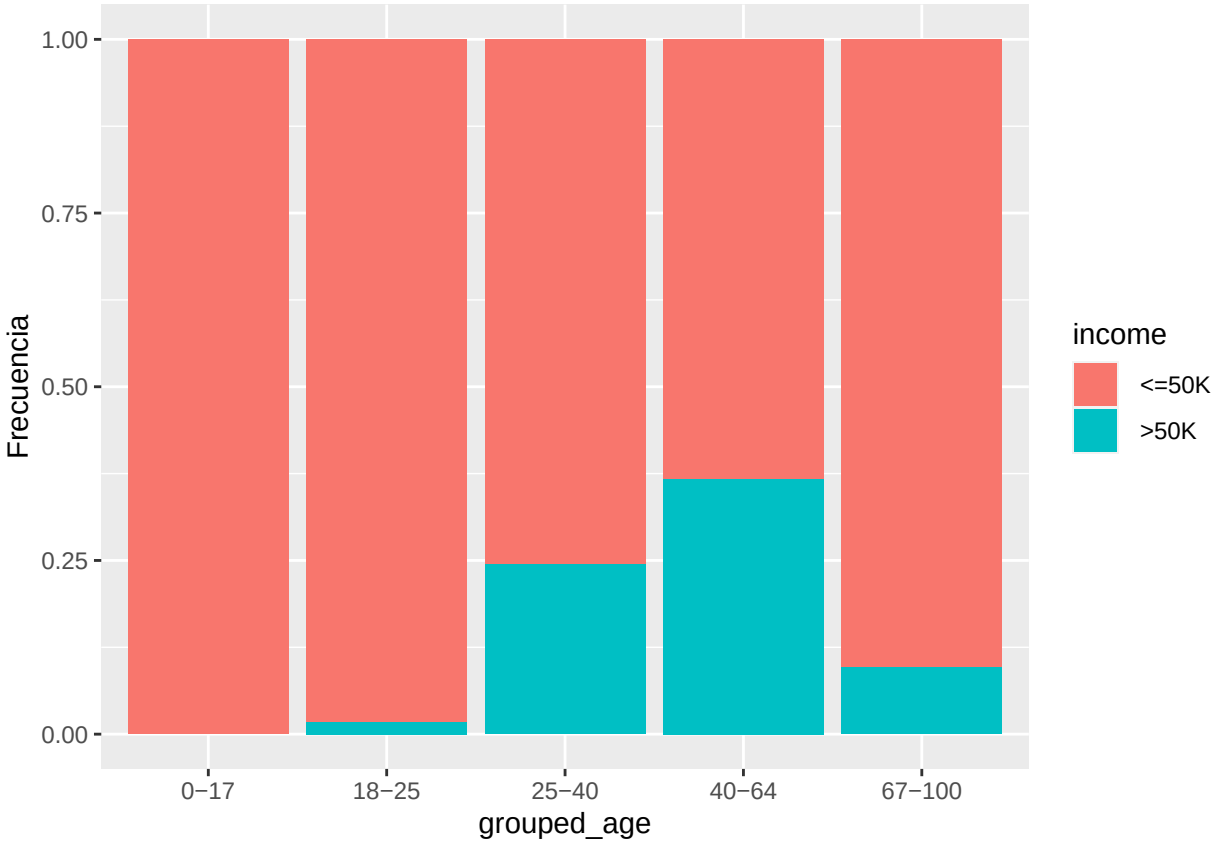

```
# Por ultimo discretizamos el nivel de educacion
datosAdult$education[datosAdult$education %in% c('10th','11th','12th','1st-4th','5th-6th','7th-8th','9t
datosAdult$education[datosAdult$education %in% c('Assoc-acdm','Assoc-voc')]="Associate"
```

```


# Comencemos a analizar el income por las distintas variables

#-------------------------------------------------------------------------
# Como es la distribucion del ingreso segun el rango de edad, en cantidades y como frecuencia/porcentaj
ggplot(data=datosAdult[1:filas,],aes(x=grouped_age,fill=income))+geom_bar()
```

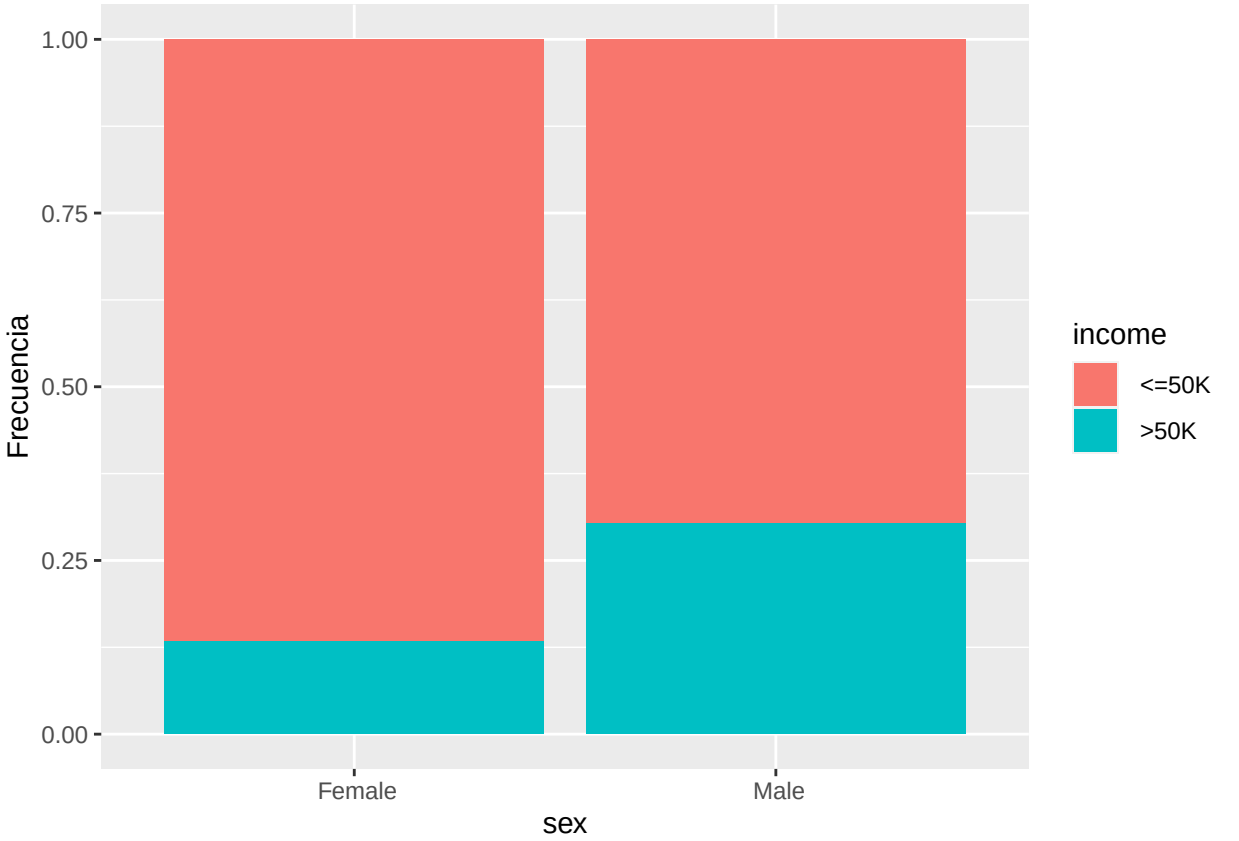


Proceso de analisis visual de variables y relaciones

```
ggplot(data=datosAdult[1:filas,],aes(x=grouped_age,fill=income))+geom_bar(position="fill")+ylab("Frecuer
```

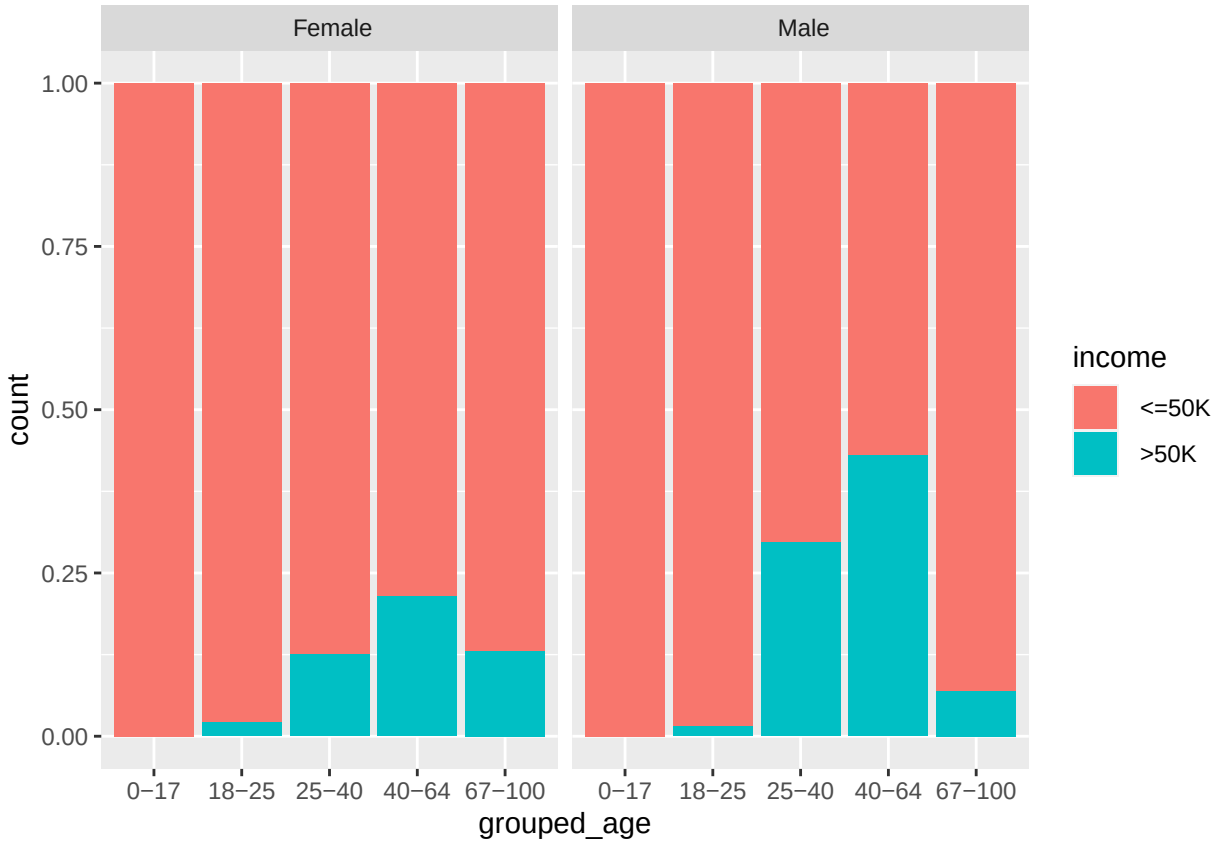



```
# Aca claramente podemos determinar que la mayor cantidad de ingresos se da en las personas mayores de 1
#-------------------------------------------------------------------------
# Veamos el income por sexo
ggplot(data=datosAdult[1:filas,],aes(x=sex,fill=income))+geom_bar(position="fill")+ylab("Frecuencia")
```




```
## El grafico indica que los hombres ganan mas que las mujeres
```

```
# Ahora, combinemos algunas variables para sacar alguna otra conclusion
ggplot(data = datosAdult[1:filas,],aes(x=grouped_age,fill=income))+geom_bar(position="fill")+facet_wrap
```



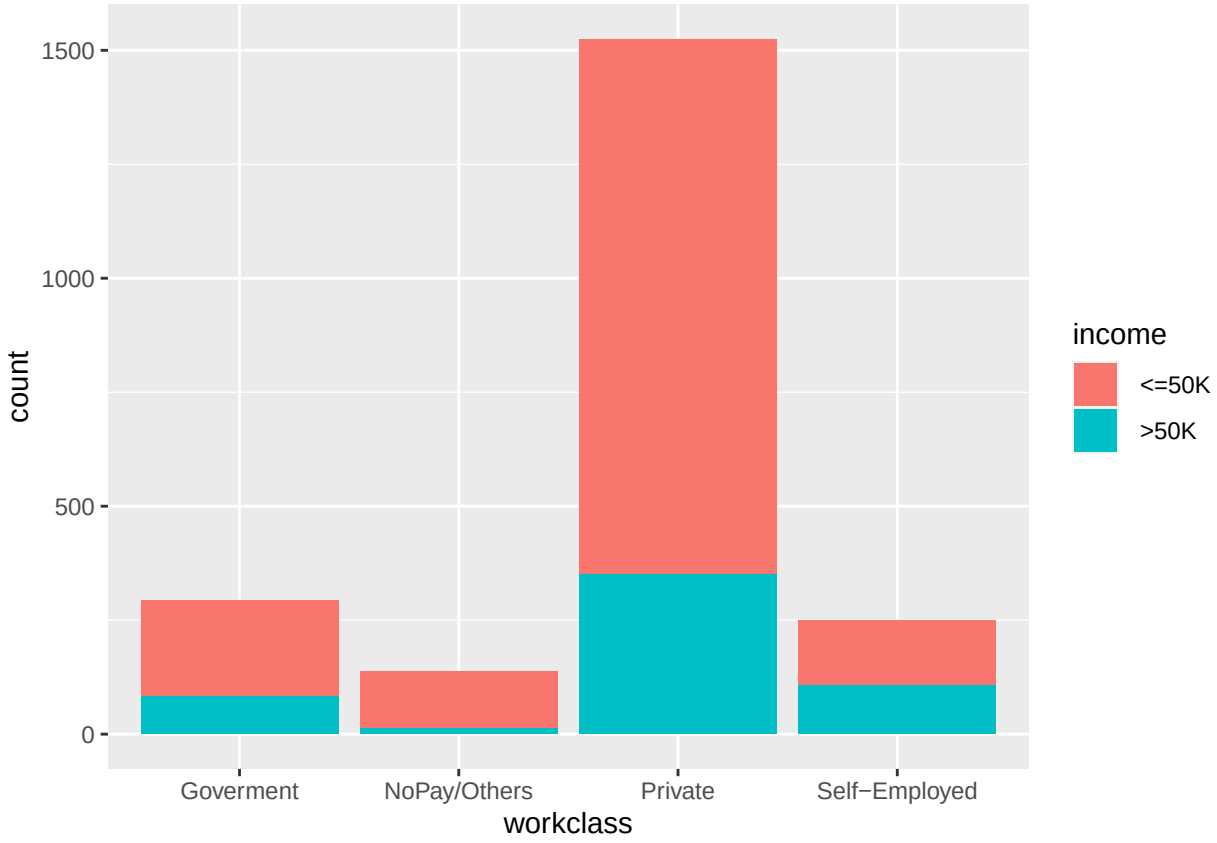

```
# Aqui vemos que sean hombres o mujeres, el rango etario de mayor ingreso sigue siendo el de 25 a 65 añ

#Sabemos por esto que son mas hombres que mujeres
table(datosAdult$sex[datosAdult$grouped_age=='67-100'])
```

```
##
##  Female    Male
##     284     573
```

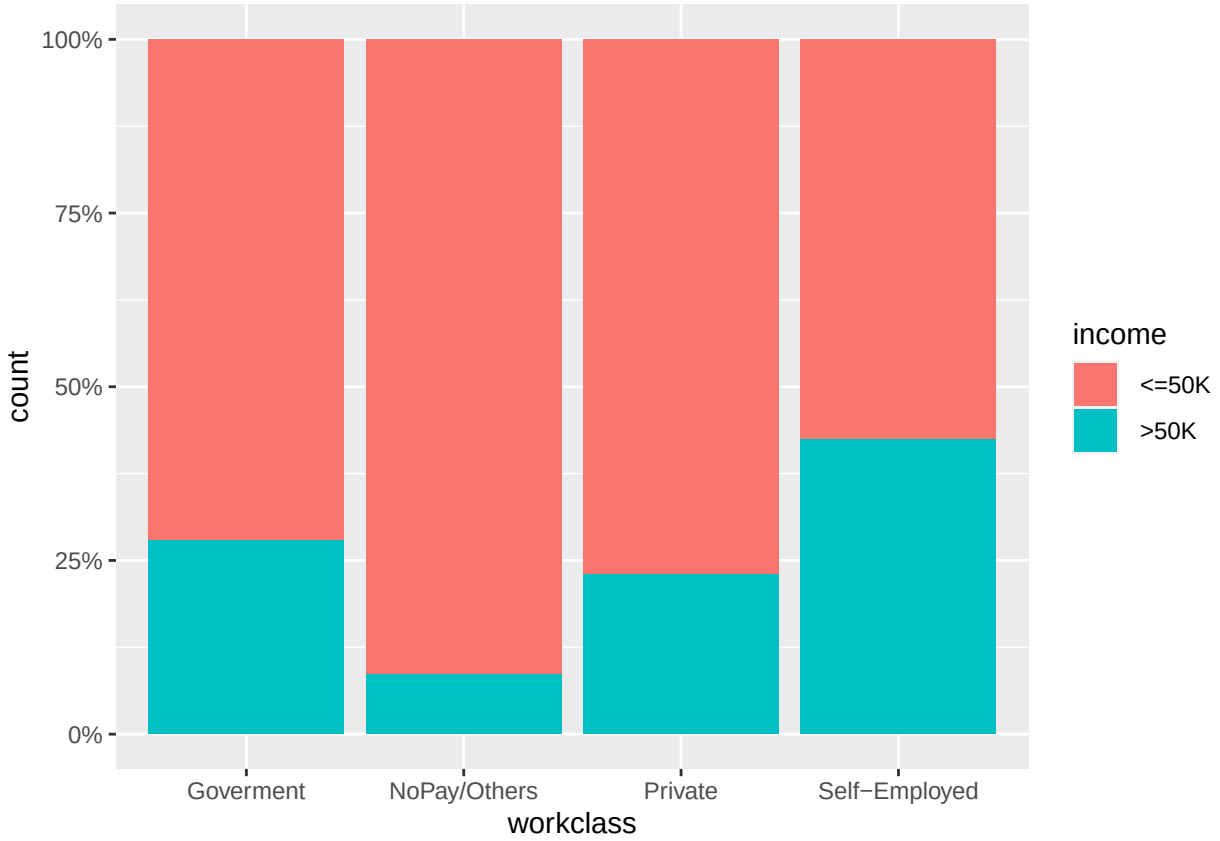
```
# Asi todo siendo menor cantidad, podemos confirmar, que las mujeres luego de su jubilacion tiene mayor

#-------------------------------------------------------------------------
```

Analicemos ahora temas como cantidad de horas trabajas, workclass, education y ocupacion, siguiendo el estado civil y las variables que teniamos pendiente validar como capital gain y loss. Para terminar analizando las razas y paises.

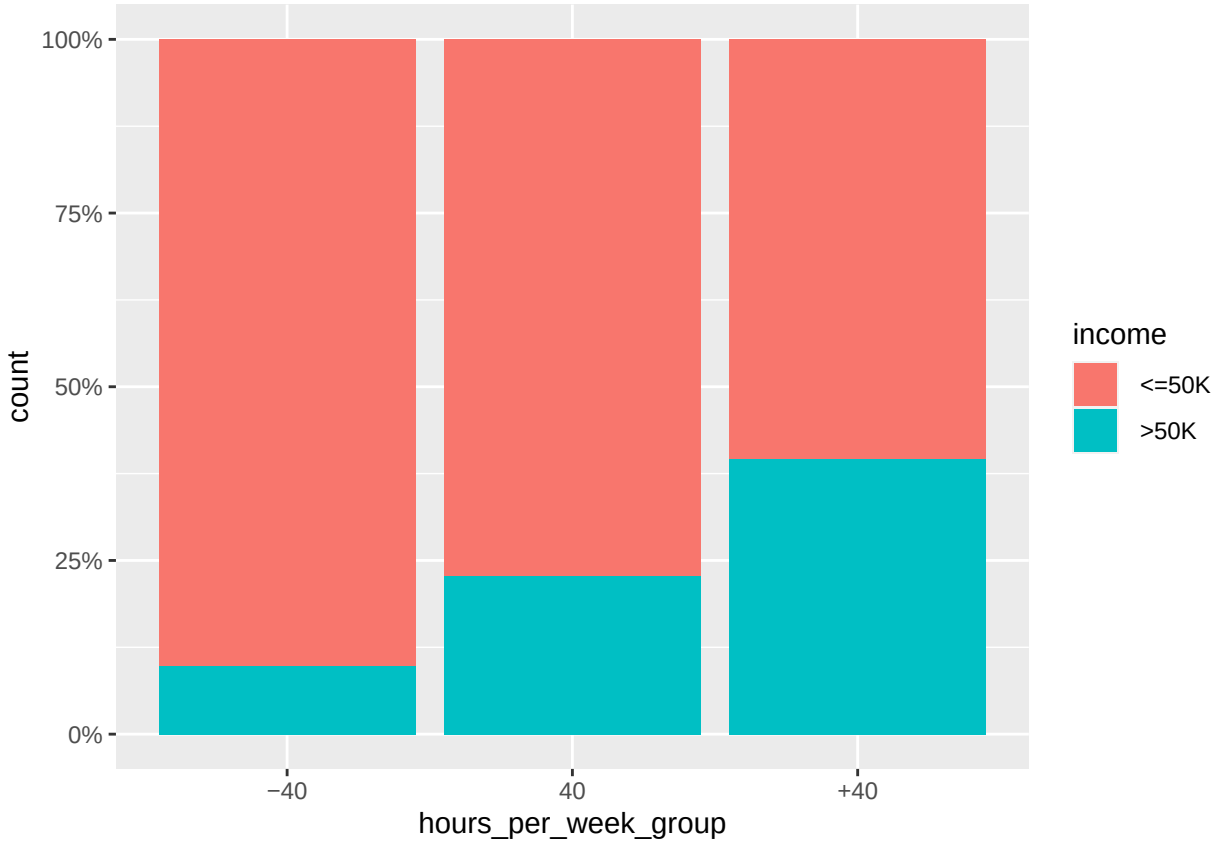
```
# Primero vamos por workclass, cantidad de horas trabajadas, education y ocupacion
#---------------WORKCLASS-----------------------------------------------------
ggplot(data=datosAdult[1:filas,],aes(x=workclass,fill=income))+geom_bar()
```



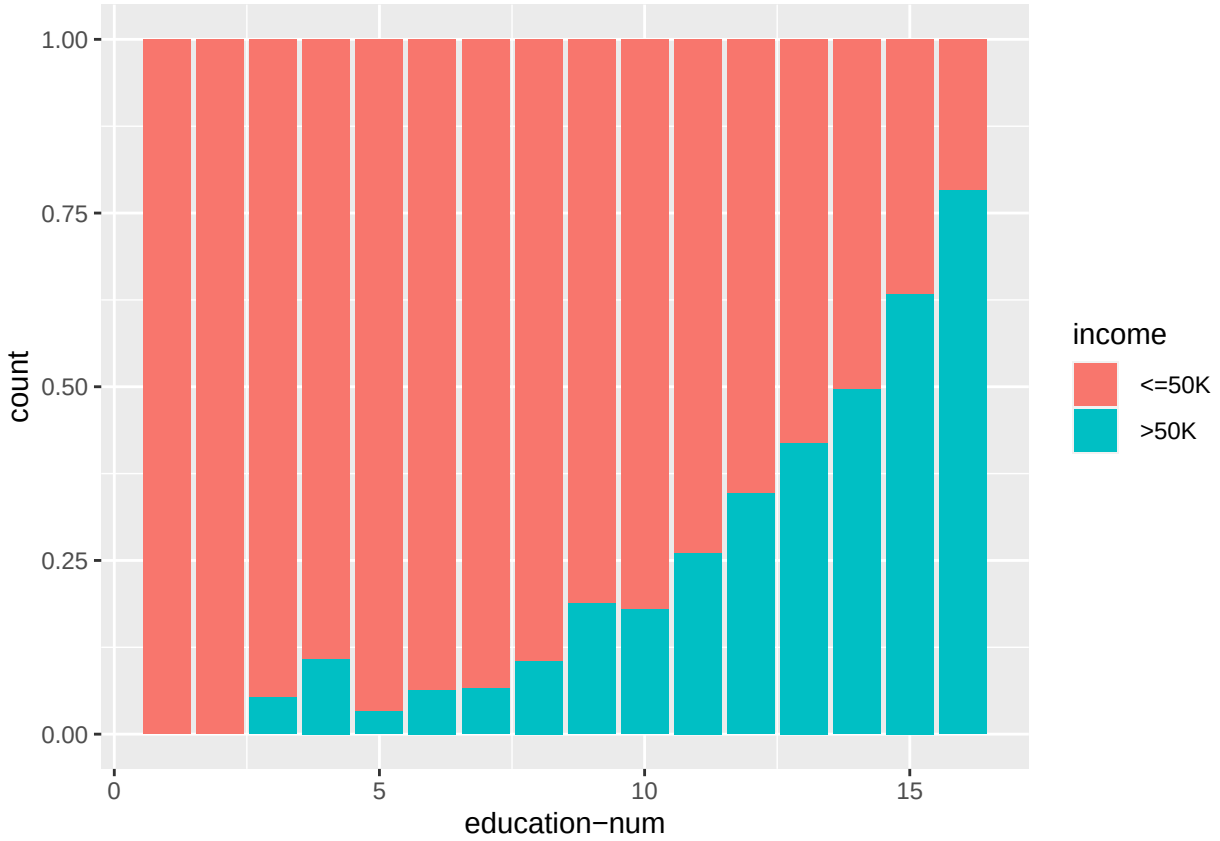

```
# Como vemos el sector privado es es donde se generan mayor ingreso, pero a la vez es donde mayor cantid
# Al verlo en porcentajes es mucho mas claro:
ggplot(data=datosAdult[1:filas,],aes(x=workclass,fill=income))+geom_bar(position="fill")+scale_y_continu
```



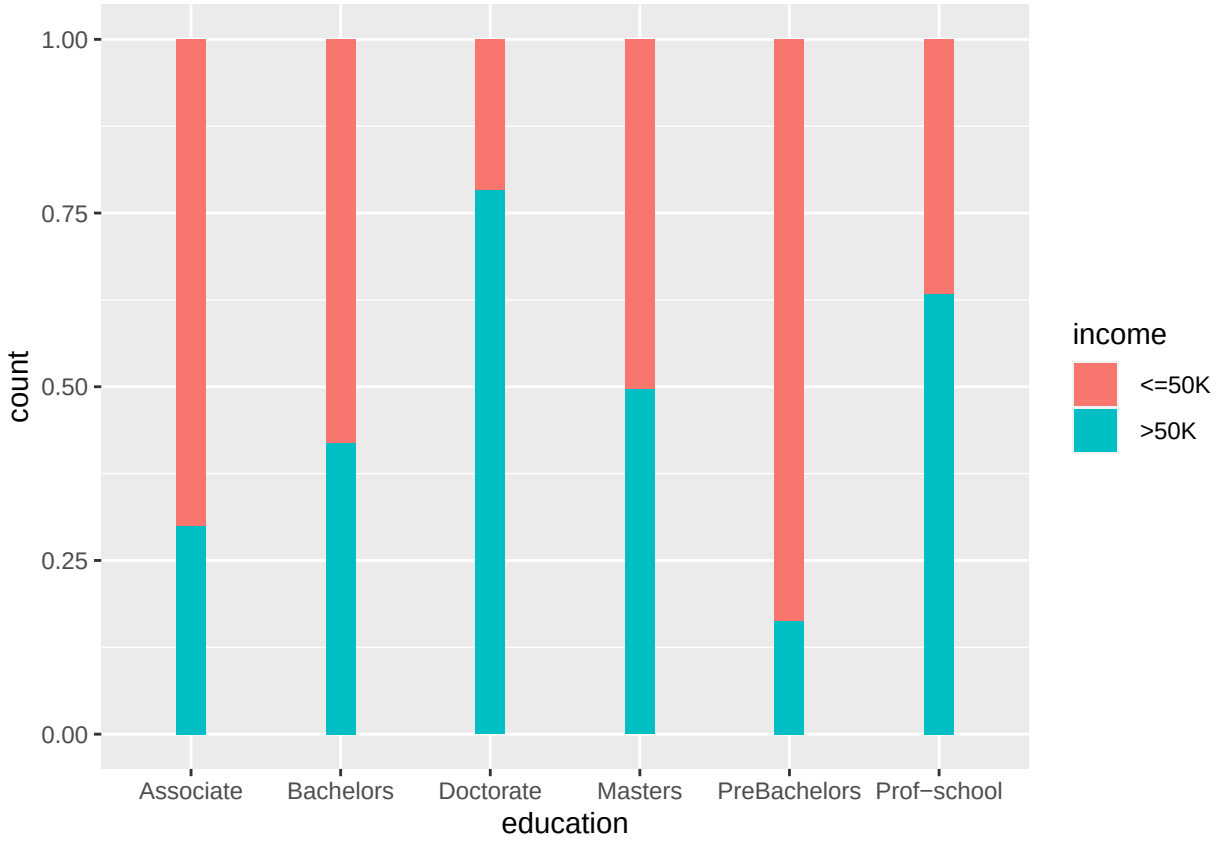

```
# Por lo que si bien la mayor masa salarial es generada en ese sector, eso no garantiza que pertenecer a
#---------------HORAS POR SEMANA ----------------------------------------------
# Veamos que pasa con las horas trabajadas:
ggplot(data=datosAdult[1:filas,],aes(x=hours_per_week_group,fill=income))+geom_bar(position="fill")+scal
```



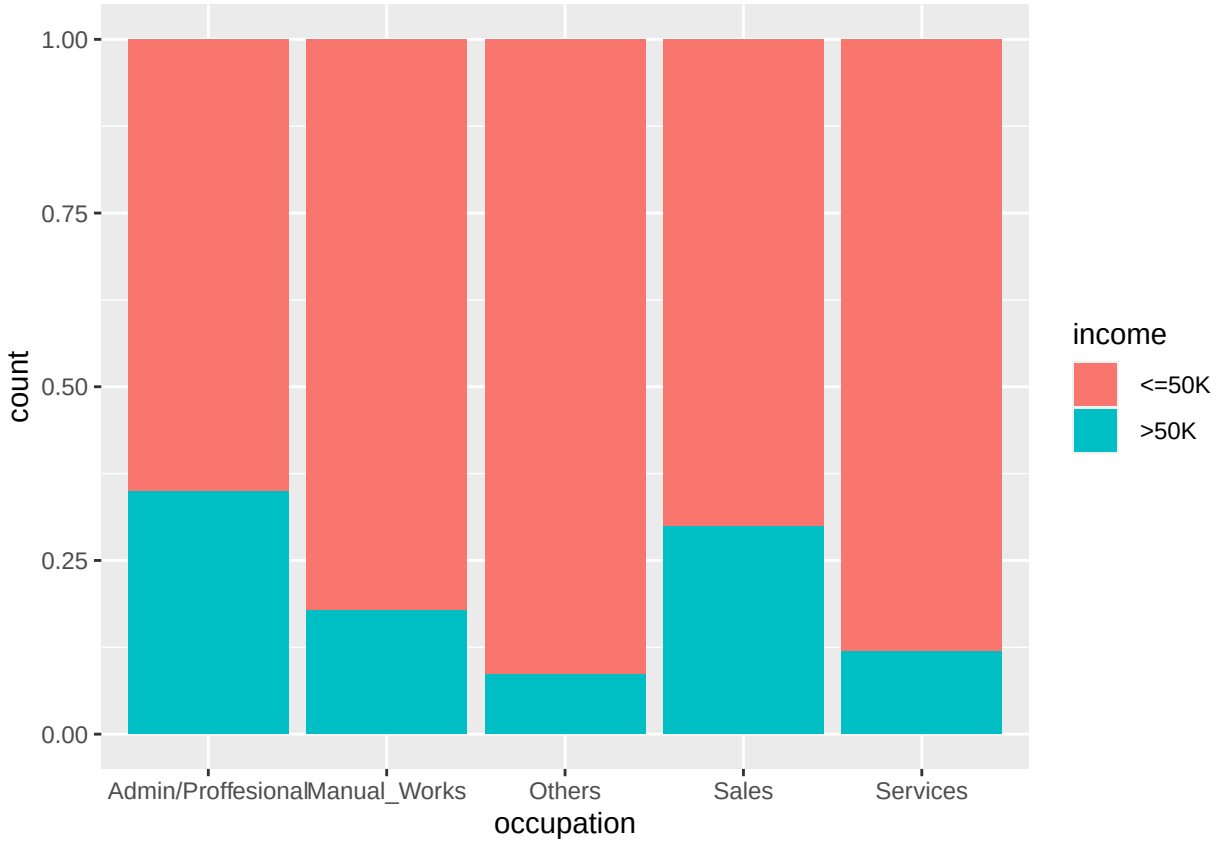

```
# Lamentablemente para todos, esto representa que cuanto mas horas dediquemos por semana mayor sera el i


#----------------EDUCACION-----------------------------------------------------
# Rapidamente vemos que cuanto mas años se dediquen a educacion en general mayores ingresos se tendran
ggplot(data=datosAdult[1:filas,],aes(x=`education-num`,fill=income))+geom_bar(position="fill")
```



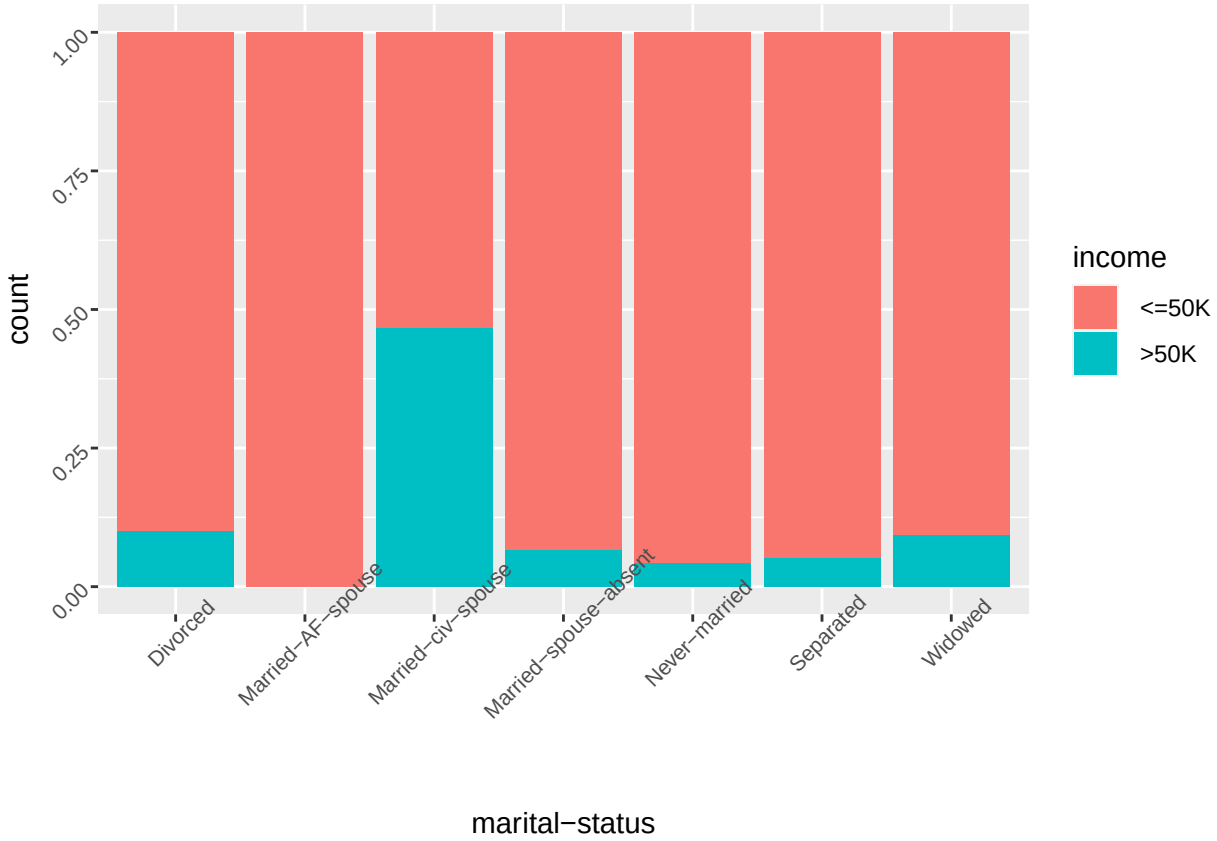

```
# Mientra que si lo vemos por los estudio obtenidos
ggplot(data=datosAdult[1:filas,],aes(x=education,fill=income))+geom_bar(position="fill",width = 0.2)
```




```
# Evidentemente hay una relaion entre el nivel de estudios y el nivel de ingreso, y obviamente el nivel


#---------------OCUPACION-----------------------------------------------------
# Como podemos imaginarnos los profesionales, personas en rangos de administracion y areas de ventas pu
ggplot(data=datosAdult[1:filas,],aes(x=occupation,fill=income))+geom_bar(position="fill")
```

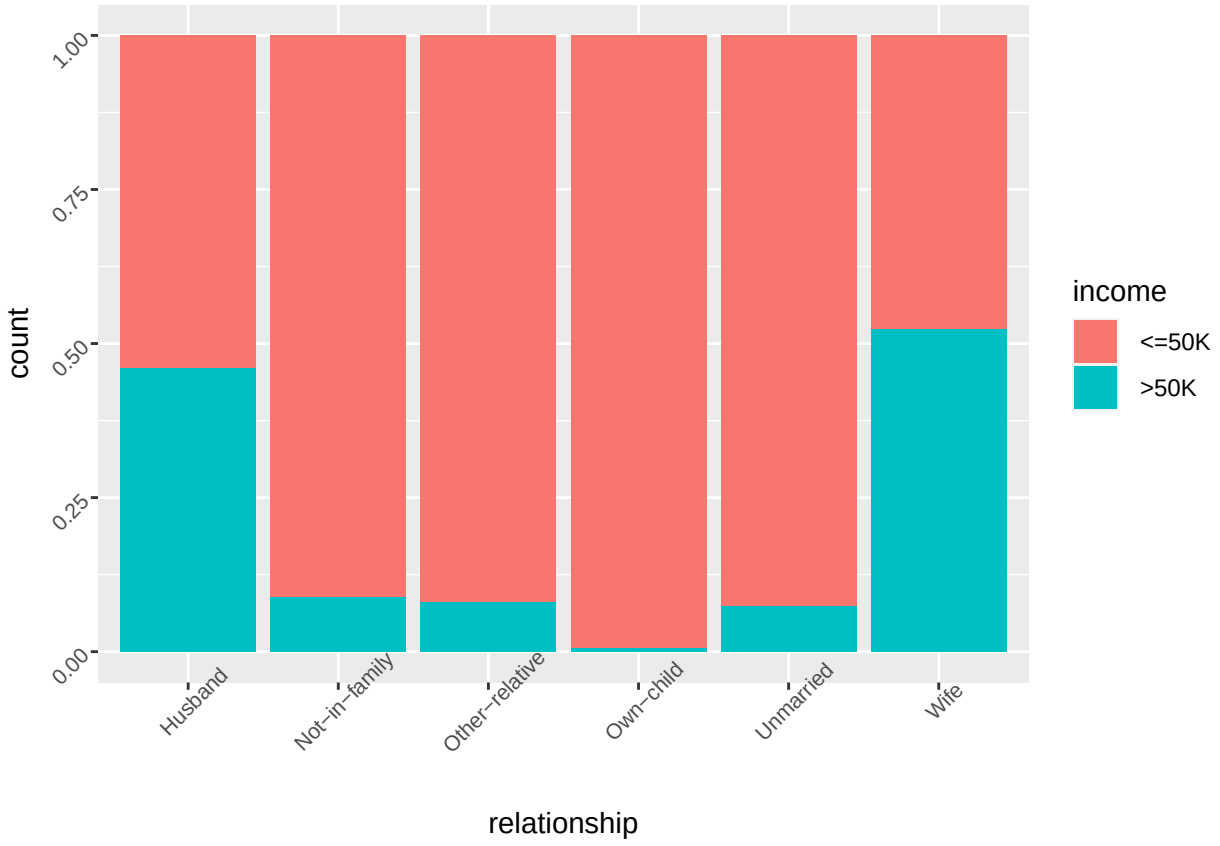



```
# Por ultimo estado civil y las variables que teniamos pendiente validar como capital gain y loss
#----------------ESTADO CIVIL-------------------------------------------------------
# Como vemos marcadamente aqui, el estado civil si que importa a la hora de obtener mayores ingresos. La
ggplot(data=datosAdult[1:filas,],aes(x=`marital-status`,fill=income))+geom_bar(position="fill")+theme(ax
```

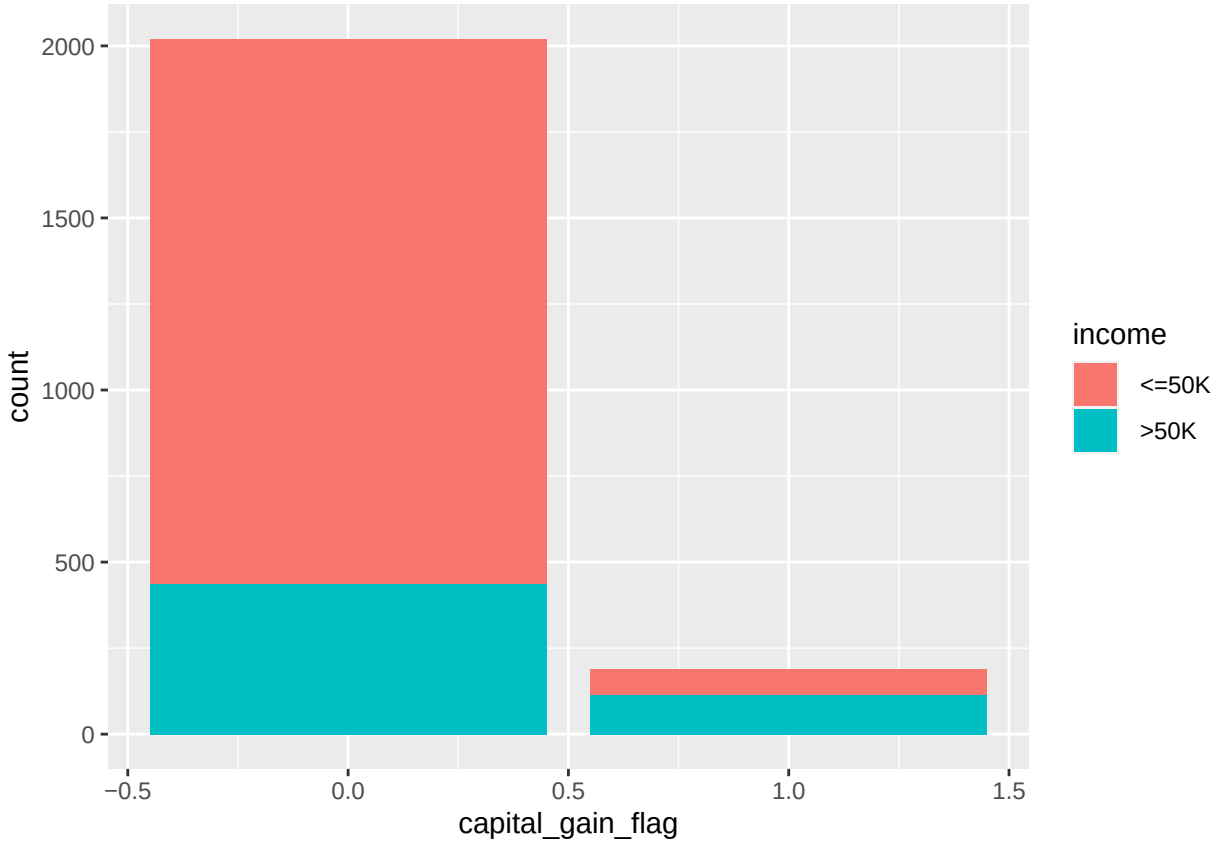



```
#Nota: He girado 45 grados y achicado el texto de las variables para que sea mas claro el grafico. He bu

# Incluso es mas notorio si lo contrastamos con alguna otra variable, como es la
# ocupacion, donde vimos anteriormente que invididualmente se destacaban mas los
# puestos de administracion/profesionales o ventas, pero si lo contrastamos contra
# el estado civil vemos que sea la profesion que sea te ira mejor si estas casado.
ggplot(data=datosAdult[1:filas,],aes(x=occupation,fill=income))+geom_bar(position="fill")+theme(axis.te
```

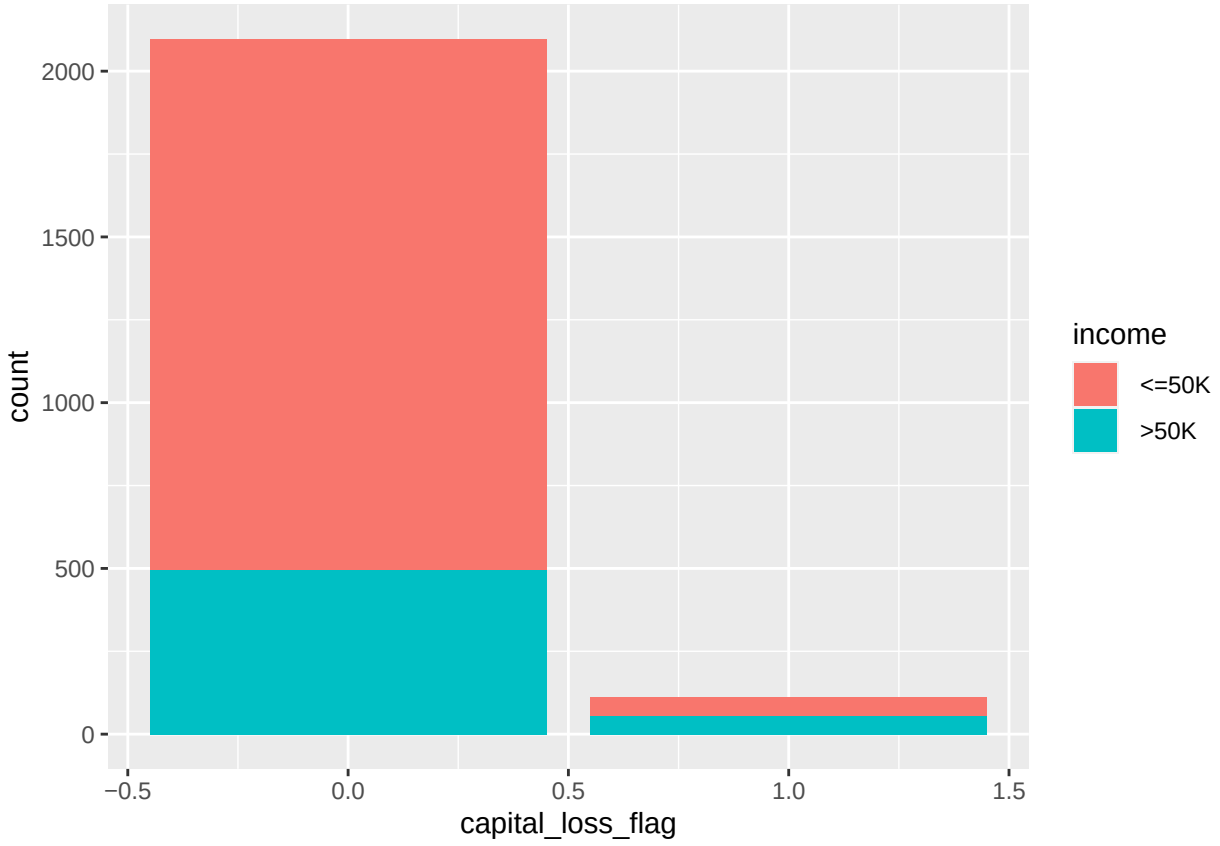

```
# Esto confirma un poco mas lo dicho antes, el hecho de estar casado implica mayor
# ingreso, y analizar si eres esposo o esposa no cambia la distribucion, lo que
# importa es el estado civil, es decir estar casado. Es un factor determinante.
ggplot(data=datosAdult[1:filas,],aes(x=relationship,fill=income))+geom_bar(position="fill")+theme(axis.
```




```
#---------------CAPITALs-------------------------------------------------------
# Como dijimos antes, la hipotesis es que las variabels de capital no aportarian
# nada paara el analisis, pero hagamos un analisis real y comprobemoslo para
# entender si deberan ser excluidos para sigueitnes etapas.

# Vemos que graficamente son lo mismo, y no tiene sentido que haya misma
# proporcion de que ganen y pierdan, parece no estar bien los datos
# recolectados de estas variables
ggplot(data=datosAdult[1:filas,],aes(x=capital_gain_flag,fill=income))+geom_bar()
```




```
ggplot(data=datosAdult[1:filas,],aes(x=capital_loss_flag,fill=income))+geom_bar()
```



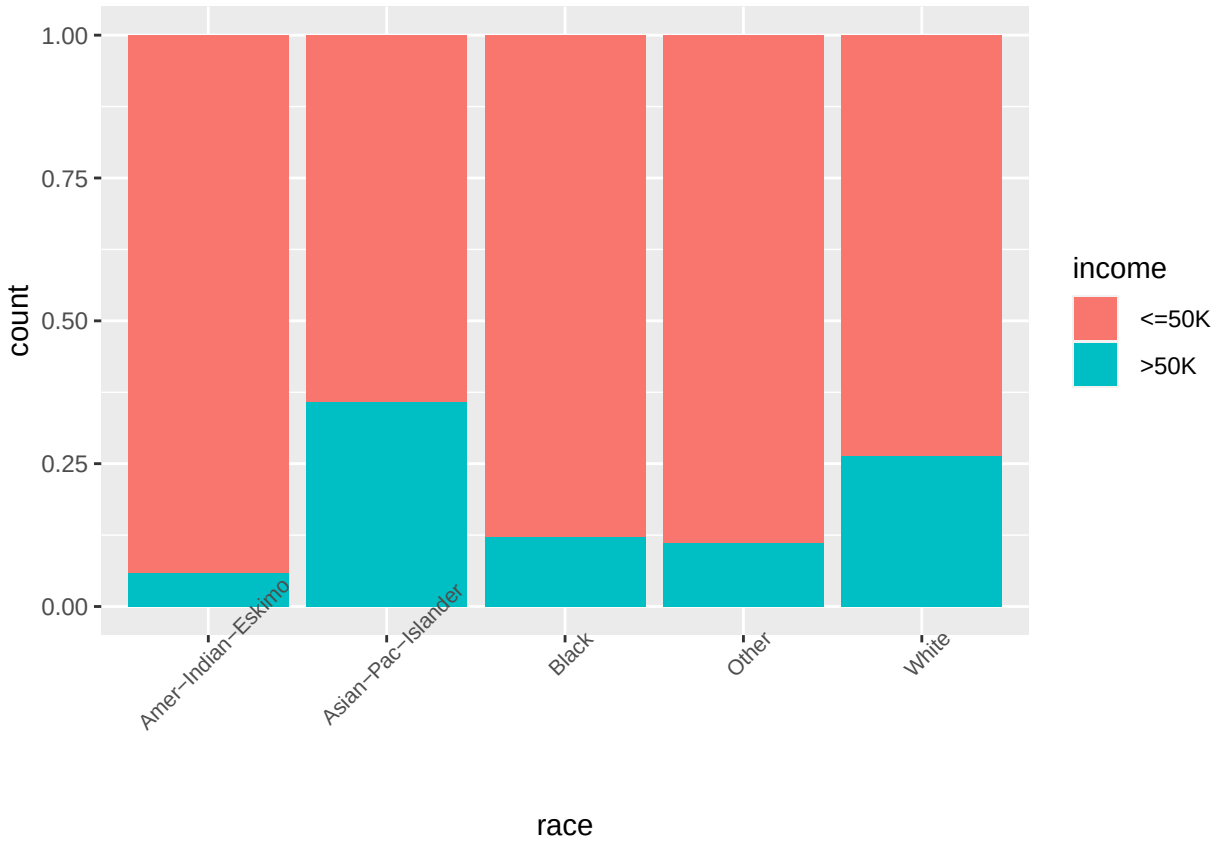

```
# incluso en porcentages se ve mas claro que la mayoria se lo llevan los valores en 0.
sum(datosAdult$capital_gain_flag==0)/length(datosAdult$capital_gain_flag)*100
```

```
## [1] 91.67102
```

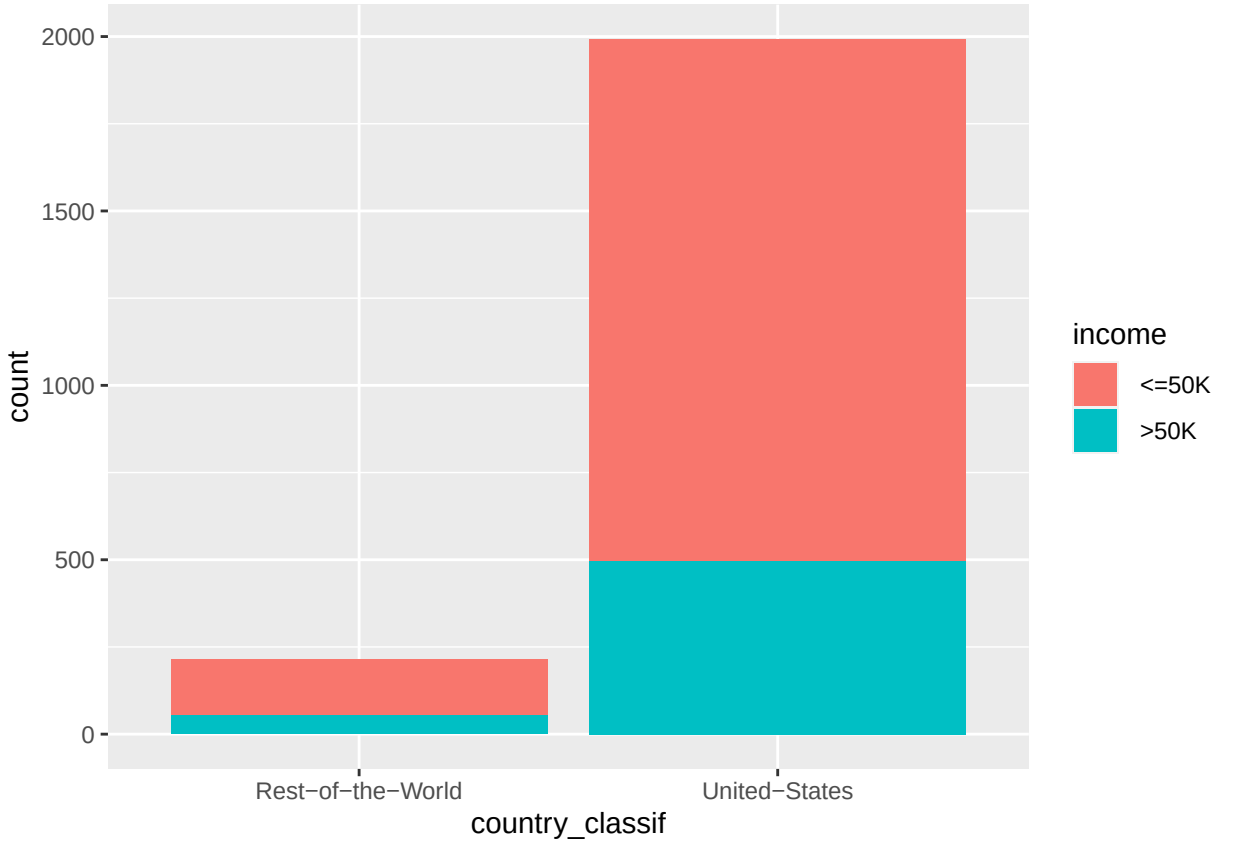
```
sum(datosAdult$capital_loss_flag==0)/length(datosAdult$capital_loss_flag)*100
```

```
## [1] 95.33491
```

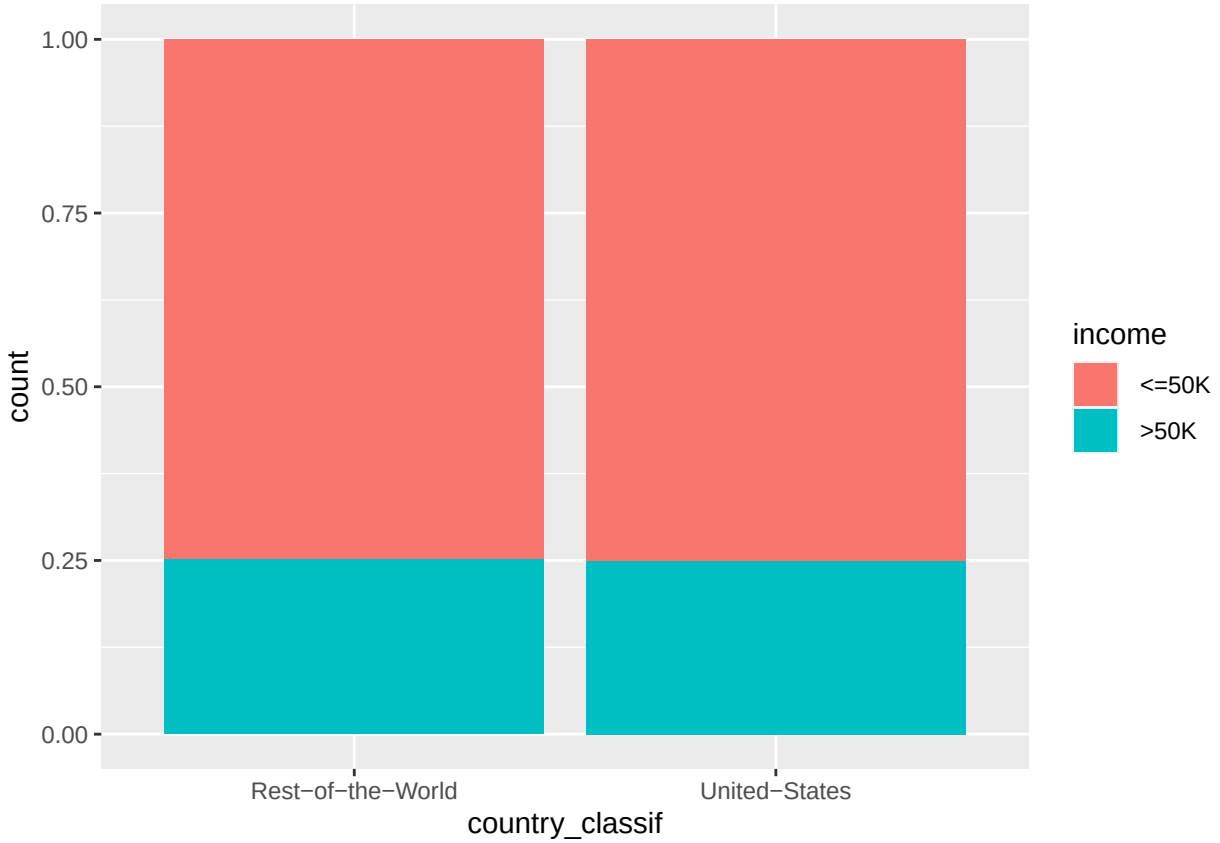
```
# Para terminar de analizar las razas y paises
#---------------RAZAS-----------------------------------------------------
# Esto confirma que el prejuicio de que los blancos ganan mejores sueldo parece
# ser verdad, pero se suma aqui que los asiaticos tambien.
ggplot(data=datosAdult[1:filas,],aes(x=race,fill=income))+geom_bar(position="fill")+theme(axis.text.x=el
```



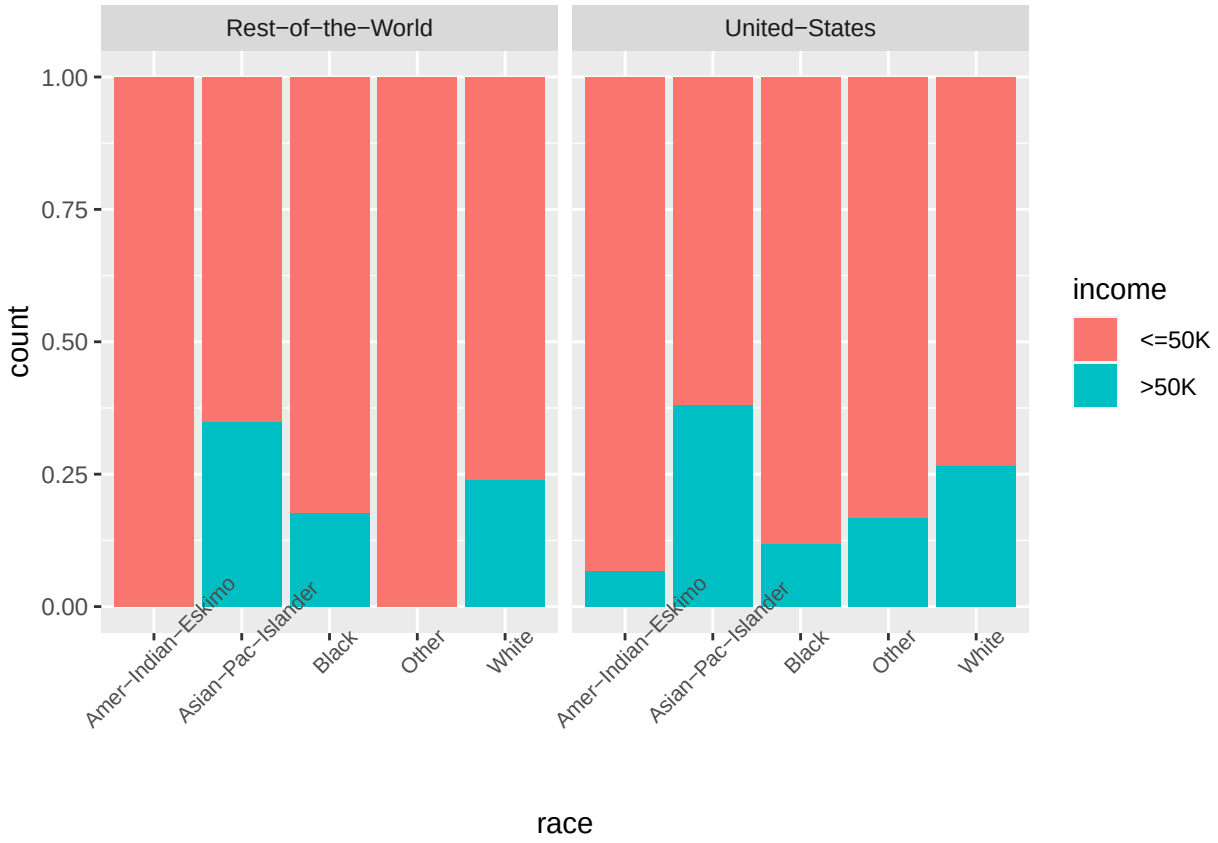

```
# Pero que pasa por paises?. Usemos aqui la discretizacion que hicimos antes donde
# vimos que la mayorias de las observacioens correspondian a Estados unidos, pero
# veremos que pasa para el resto del mundo para cada raza.

# Como vemos aqui, claramente tenemos casi todas las observaciones para USA:
ggplot(data=datosAdult[1:filas,],aes(x=country_classif,fill=income))+geom_bar()
```




```
# Pero si lo vemos en porcentaje, entre estados unidos y el resto del mundo como
# conjunto no hay ninguna diferencia que marque que segun de donde seas nativo te
# garantize un sueldo mayor a 50k, es indistinto, claro esto puede llevar a error,
# si consideramos a el resto del mundo como un todo, claramente habria que abrir
# por pais. Dicho esto, con los datos que tenemos excluiria el country de este
# analisis porque no aporta nada.
ggplot(data=datosAdult[1:filas,],aes(x=country_classif,fill=income))+geom_bar(position="fill")
```




```
# Como vemos, sea el pais que sea, la distribucion de ingresos de las razas no es afectada por el pais.
ggplot(data=datosAdult[1:filas,],aes(x=race,fill=income))+geom_bar(position="fill")+facet_wrap(~country
```



Lectura recomendada:

https://arxiv.org/ftp/arxiv/papers/1810/1810.10076.pdf Nivel de analisis al que aspiro llegar.