



CC216 - FUNDAMENTOS DE DATA SCIENCE

HOJA 2.1 – ADQUISICION, PREPARACION Y VISUALIZACION PRELIMINAR DE DATOS CON R

En esta clase, veremos en la práctica, como un conjunto de datos se convierte en información. A partir de la adquisición del conjunto de datos, se realiza una serie de operaciones sobre ellos, para finalmente visualizar de forma preliminar que información útil nos proporcionan.

OBJETIVO PRINCIPAL

Realizar un análisis exploratorio básico en un conjunto de datos, crear visualizaciones y sacar inferencias.

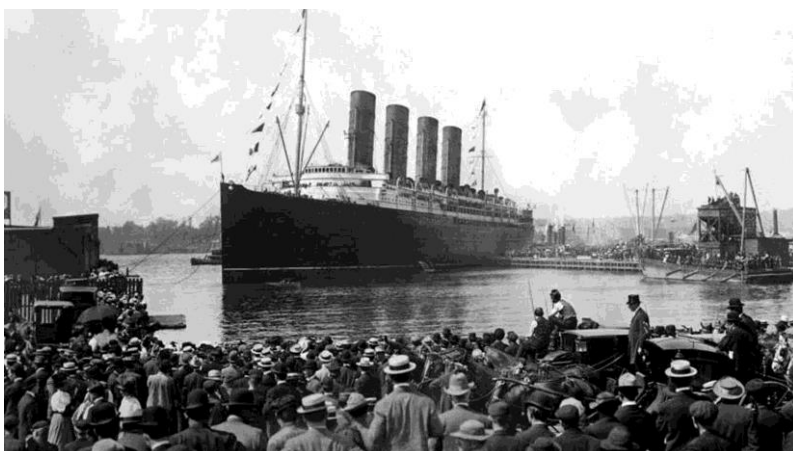
COMPETENCIAS

- Aprender a leer archivos de datos según origen, en este caso, en formato CSV.
- Realizar las primeras operaciones de limpieza de los datos.
- Visualizar información relevante a partir del conjunto de datos procesado.

CASO DE ANALISIS

El hundimiento del RMS Titanic es uno de los naufragios más infames de la historia. El 15 de abril de 1912, durante su viaje inaugural, el Titanic se hundió después de chocar con un iceberg, matando a 1502 de los 2224 pasajeros y tripulación. Esta tragedia conmocionó a la comunidad internacional y dio lugar a mejores normas de seguridad para los buques.

Una de las razones por las que el naufragio provocó tantas pérdidas de vidas fue que no había suficientes botes salvavidas para los pasajeros y la tripulación. Aunque hubo algún elemento de suerte involucrado en sobrevivir al hundimiento, algunos grupos de personas tenían más probabilidades de sobrevivir que otros, como las mujeres, los niños y los pasajeros pertenecientes a la clase alta.



CONJUNTO DE DATOS (DATA SET)

Utilizaremos el conjunto de datos del **Titanic**, el cual contiene registros históricos de todos los pasajeros que lo abordaron.

A continuación, se muestra una breve descripción de las 12 variables del conjunto de datos:

Variable	Descripción
PassengerId	número de serie
Survived	contiene valores binarios de 0 y 1. El pasajero no sobrevivió - 0, El pasajero sobrevivió - 1.
Pclass	Clase de entrada Boleto de primera clase, segunda clase o tercera clase
Name	Nombre del pasajero
Sex	Masculino o Femenino
Age	Edad en años - Entero
SibSp	No. de hermanos / cónyuges - hermanos, hermanas y / o esposo / esposa
Parch	No. de padres / hijos - madre / padre y / o hija, hijo
Ticket	Número de serie
Fare	Tarifa de pasajero
Cabin	Número de cabina
Embarked	Puerto de Embarque C- Cherburgo, Q - Queenstown, S - Southhampton

La fuente de los datos es:

<https://web.stanford.edu/class/archive/cs/cs109/cs109.1166/stuff/titanic.csv>

Pero de forma local, también encontramos el archivo titanic.csv en el aula virtual como material de esta clase.

INSTRUCCIONES EN R / R STUDIO

Se realizarán las siguientes tareas desde la consola en R:

- I. CARGAR DATOS
- II. INSPECCIONAR DATOS
- III. LIMPIAR DATOS
- IV. VISUALIZACION GRAFICA
- V. CONCLUSIONES PRELIMINARES

I. CARGAR DATOS

1. Existen tres funciones en R para poder adquirir/cargar un conjunto de datos desde el tipo de archivo CSV:

read_csv(): para leer archivos con coma (",") como separador

read_csv2(): para leer archivos con punto y coma (";") como separador

read_tsv(): para leer archivos con tabulador ("\t") como separador

read_delim(sep = '|'): para leer archivos con separador distintos como puede ser el símbolo '|'

Para nuestro caso, utilizaremos **read_csv()** desde la línea de comando de la Consola en RStudio.

Podemos cargar los datos desde un sitio en internet:

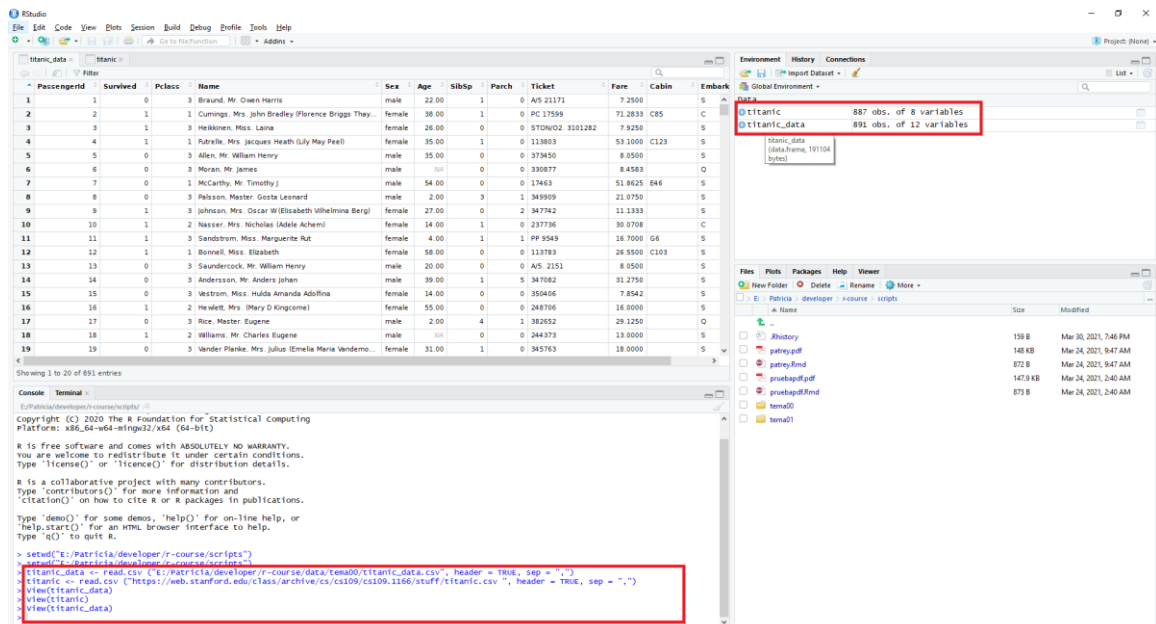
```
titanic <- read.csv  
("https://web.stanford.edu/class/archive/cs/cs109/cs109.1166/stuff/titanic.csv ", header =  
TRUE, sep = ",")
```

O bien, cargar los datos desde un archivo local en nuestra PC:

```
titanic_data <- read.csv ("E:/Patricia/developer/r-course/data/tema00/titanic_data.csv", header  
= TRUE, sep = ",")
```

La instrucción o comando es la misma, pero en este ejemplo, observamos que los conjuntos de datos cargados difieren en el numero de observaciones y variables.

El código anterior lee el archivo titanic.csv en un marco de datos (data frame) llamado titanic . Con Header = TRUE estamos especificando que los datos incluyen un encabezado (nombres de columna) y sep = "" especifica que los valores en los datos están separados por comas.



II. INSPECCIONAR DATOS

Observamos los datos cargados para familiarizarnos con el conjunto de datos

Damos clic sobre el data frame “titanic_data” que se observa en panel de Enviroment. Abriremos el dataframe y veremos el contenido. El mismo resultado se obtiene si se ejecuta en la línea de comando:

> View(titanic_data)

The screenshot shows the RStudio interface. The console window displays the following code and output:

```

# Copyright (C) 2020 The R Foundation for Statistical Computing
# Platform: x86_64-w64-mingw32/x64 (64-bit)

# R is free software and comes with ABSOLUTELY NO WARRANTY.
# You are welcome to redistribute it under certain conditions.
# Type 'license()' or 'licence()' for distribution details.

# R is a collaborative project with many contributors.
# Type 'contributors()' for more information and
# 'citation()' on how to cite R or R packages in publications.

# Type 'demo()' for some demos, 'help()' for on-line help, or
# 'help.start()' for an HTML browser interface to help.
# Type 'q()' to quit R.

> setwd("E:/patricia/developer/r-course/scripts")
> source("E:/patricia/developer/r-course/scripts")
> titanic_data <- read.csv("E:/patricia/developer/r-course/data/titanc_data.csv", header = TRUE, sep = ";")
> titanic <- read.csv("https://web.stanford.edu/class/archive/cs/cs109/cs109.1166/stuff/titanic.csv", header = TRUE, sep = ";")
> view(titanic_data)
> view(titanic)

```

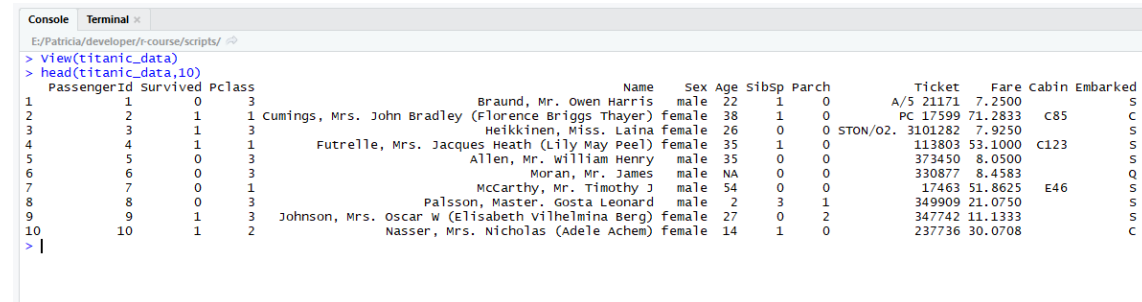
The Environment pane on the right shows the data frame 'titanic_data' with 891 observations and 12 variables. The Files pane on the bottom right shows the project files.

Podemos ver las observaciones (filas) iniciales o finales del dataset con el siguiente comando:

head(titanic_data,n) ó tail((titanic_data,n) y donde “n” es el número de observaciones a visualizar (por defecto n es 5, si se omite el parámetro).

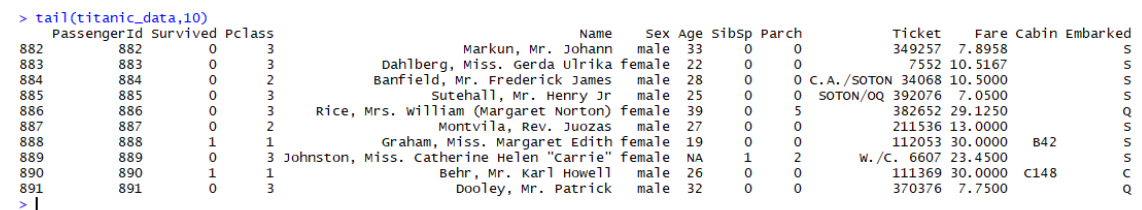
Ejecutando dichas instrucciones, obtenemos las siguientes salidas:

```
> head(titanic_data,10)
```



PassengerId	Survived	Pclass	Name	Sex	Age	Sibsp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.2500		S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.9250		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1000	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.0500		S
6	0	3	Moran, Mr. James	male	NA	0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.0750		S
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708		C

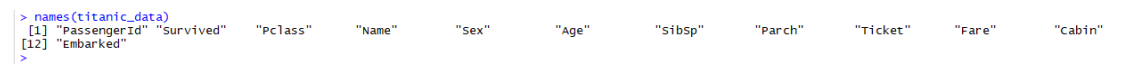
```
> tail(titanic_data,10)
```



PassengerId	Survived	Pclass	Name	Sex	Age	Sibsp	Parch	Ticket	Fare	Cabin	Embarked
882	0	3	Markun, Mr. Johann	male	33	0	0	349257	7.8958		S
883	0	3	Dahlberg, Miss. Gerda Ulrika	female	22	0	0	7552	10.5167		S
884	0	2	Banfield, Mr. Frederick James	male	28	0	0	C.A./SOTON 34068	10.5000		S
885	0	3	Sutehall, Mr. Henry Jr	male	25	0	0	SOTON/OQ 392076	7.0500		S
886	0	3	Rice, Mrs. William (Margaret Norton)	female	39	0	5	382652	29.1250		Q
887	0	2	Montvila, Rev. Juozas	male	27	0	0	211536	13.0000		S
888	1	1	Graham, Miss. Margaret Edith	female	19	0	0	112053	30.0000	B42	S
889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NA	1	2	w./C. 6607	23.4500		S
890	1	1	Behr, Mr. Karl Howell	male	26	0	0	111369	30.0000	C148	C
891	0	3	Dooley, Mr. Patrick	male	32	0	0	370376	7.7500		Q

```
> names(titanic_data)
```

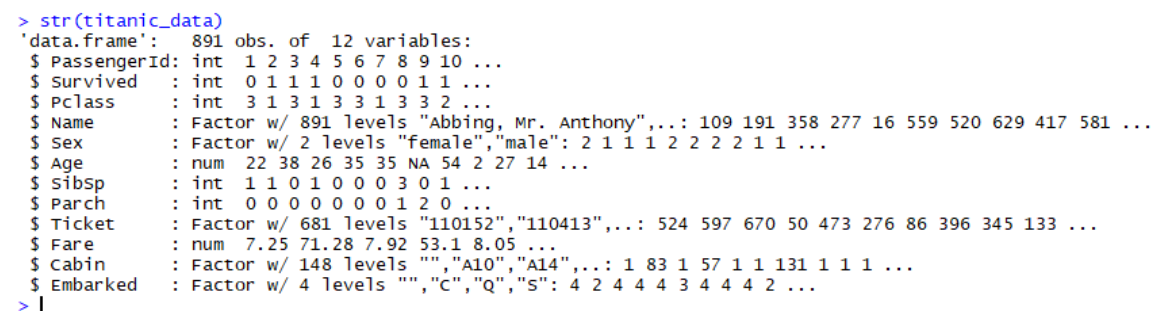
La instrucción **names** nos ayuda a verificar todas las variables del conjunto de datos.



```
[1] "PassengerId" "Survived" "Pclass" "Name" "Sex" "Age" "Sibsp" "Parch" "Ticket" "Fare" "Cabin"
[12] "Embarked"
```

```
> str(titanic_data)
```

La instrucción **str** ayuda a comprender la estructura del conjunto de datos, el tipo de datos de cada atributo y el número de filas y columnas presentes en los datos.



```
'data.frame': 891 obs. of 12 variables:
 $ PassengerId: int 1 2 3 4 5 6 7 8 9 10 ...
 $ Survived : int 0 1 1 1 0 0 0 0 1 1 ...
 $ Pclass : int 3 1 3 1 3 1 3 1 3 2 ...
 $ Name : Factor w/ 891 levels "Abbing, Mr. Anthony",...: 109 191 358 277 16 559 520 629 417 581 ...
 $ Sex : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
 $ Age : num 22 38 26 35 35 NA 54 2 27 14 ...
 $ Sibsp : int 1 1 0 1 0 0 0 3 0 1 ...
 $ Parch : int 0 0 0 0 0 0 0 1 2 0 ...
 $ Ticket : Factor w/ 681 levels "110152","110413",...: 524 597 670 50 473 276 86 396 345 133 ...
 $ Fare : num 7.25 71.28 7.92 53.1 8.05 ...
 $ Cabin : Factor w/ 148 levels "", "A10", "A14",...: 1 83 1 57 1 1 131 1 1 1 ...
 $ Embarked : Factor w/ 4 levels "", "C", "Q", "S": 4 2 4 4 4 3 4 4 4 2 ...
```

```
> summary(titanic_data)
```

Summary () es una de las funciones más importantes que ayudan a resumir cada atributo en el conjunto de datos. Da un conjunto de estadísticas descriptivas, según el tipo de variable:

- En el caso de una variable numérica -> Da media, mediana, moda, rango y cuartiles.
- En caso de una variable de tipo factor -> Da una tabla con las frecuencias.
- En el caso de tipo Factor + Variables numéricas -> Da el número de valores perdidos.

- En caso de variables de tipo carácter -> Da la longitud y la clase.

```
> summary(titanic_data)
  PassengerId  Survived  Pclass
Min.   : 1.0   Min.   :0.0000 Min.   :1.000
1st Qu.:223.5 1st Qu.:0.0000 1st Qu.:2.000
Median :446.0 Median :0.0000 Median :3.000
Mean   :446.0 Mean   :0.3838 Mean   :2.309
3rd Qu.:668.5 3rd Qu.:1.0000 3rd Qu.:3.000
Max.   :891.0 Max.   :1.0000 Max.   :3.000

  Name                               Sex   Age   SibSp   Parch
1 Abbing, Mr. Anthony               : 1 female:314 Min.   : 0.42 Min.   :0.0000 Min.   :0.0000
2 Abbott, Mr. Rossmore Edward       : 1 male :577 1st Qu.:20.12 1st Qu.:0.000 1st Qu.:0.0000
3 Abbott, Mrs. Stanton (Rosa Hunt)  : 1      Median :28.00 Median :0.000 Median :0.0000
4 Abelson, Mr. Samuel               : 1      Mean   :29.70 Mean   :0.523 Mean   :0.3816
5 Abelson, Mrs. Samuel (Hannah Wizosky): 1      3rd Qu.:38.00 3rd Qu.:1.000 3rd Qu.:0.0000
6 Adah!, Mr. Mauritz Nils Martin    : 1      Max.   :80.00 Max.   :8.000 Max.   :6.0000
7 (other)                           :885      NA's :177

  Ticket   Fare   Cabin
1601 : 7   Min.   : 0.00   :687
347082 : 7 1st Qu.: 7.91   B96 B98 : 4
CA. 2343: 7 Median :14.45   C23 C25 C27: 4
3101295 : 6 Mean   :32.20   G6      : 4
347088 : 6 3rd Qu.:31.00   C22 C26 : 3
CA 2144 : 6 Max.   :512.33   D      : 3
(other) :852              (other) :186
> |
```

En caso de que solo necesitemos la estadística de resumen para una variable en particular en el conjunto de datos, podemos usar: `summary(datasetName $ VariableName)`

```
> summary(titanic_data$Age)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
 0.42  20.12   28.00   29.70  38.00   80.00    177
> |
```

as.factor (conjunto de datos \$ ColumnName)

Hay ocasiones en las que algunas de las variables del conjunto de datos son factores, pero pueden interpretarse como numéricas. Por ejemplo, la Pclass (Passenger Class) cuenta los valores 1, 2 y 3, sin embargo, sabemos que estos no deben considerarse como numéricos, ya que son solo niveles. Para que tales variables sean tratadas como factores y no como números, necesitamos convertirlas explícitamente en factores usando la función `as.factor()`

Transformamos a variable de tipo factor los atributos `Survived`, `Pclass`, `Sex` y `Embarked`. Comprobamos luego con la instrucción `str(titanic_data)` que cambiaron dichos atributos a factor.

```
> titanic_data$Survived <- as.factor(titanic_data$Survived)
> titanic_data$Pclass <- as.factor(titanic_data$Pclass)
> titanic_data$Sex <- as.factor(titanic_data$Sex)
> titanic_data$Embarked <- as.factor(titanic_data$Embarked)
> str(titanic_data)
'data.frame': 891 obs. of 12 variables:
 $ PassengerId: int 1 2 3 4 5 6 7 8 9 10 ...
 $ Survived   : Factor w/ 2 levels "0","1": 1 2 2 2 1 1 1 1 2 2 ...
 $ Pclass     : Factor w/ 3 levels "1","2","3": 3 1 3 1 3 3 1 3 3 2 ...
 $ Name       : Factor w/ 891 levels "Abbing, Mr. Anthony",...: 109 191 358 277 16 559 520 629 417 581 ...
 $ Sex        : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
 $ Age        : num 22 38 26 35 35 NA 54 2 27 14 ...
 $ SibSp      : int 1 1 0 1 0 0 0 3 0 1 ...
 $ Parch      : int 0 0 0 0 0 0 0 1 2 0 ...
 $ Ticket     : Factor w/ 681 levels "110152","110413",...: 524 597 670 50 473 276 86 396 345 133 ...
 $ Fare       : num 7.25 71.28 7.92 53.1 8.05 ...
 $ Cabin      : Factor w/ 148 levels "", "A10", "A14",...: 1 83 1 57 1 1 131 1 1 1 ...
 $ Embarked   : Factor w/ 4 levels "", "C", "Q", "S": 4 2 4 4 4 3 4 4 4 2 ...
> |
```

III. LIMPIEZA DE DATOS

Podemos verificar que variables dentro del dataset poseen valores NA (not available) o figuran en blanco.

Para ello ejecutamos las siguientes funciones:

#funcion sin_valor(dataframe) que desliega cuantos valores NA posee cada variable

```
sin_valor <- function(x){
```

```

sum = 0

for(i in 1:ncol(x))
{
  cat("En la columna",colnames(x[i]),"total de valores NA:",colSums(is.na(x[i])),"\n")
}
}

sin_valor(titanic_data)

```

Al ejecutar la función `sin_valor(titanic_data)` observamos que existen 177 observaciones con el valor NA en la variable **Age**.

```

E:/Patricia/developer/r-course/scripts/ ↗
> #función que desliega cuantos valores NA posee cada variable
> sin_valor <- function(x){
+   sum = 0
+   for(i in 1:ncol(x))
+   {
+     cat("En la columna",colnames(x[i]),"total de valores NA:",colSums(is.na(x[i])),"\n")
+   }
+ }
>
> sin_valor(titanic_data)
En la columna PassengerId total de valores NA: 0
En la columna Survived total de valores NA: 0
En la columna Pclass total de valores NA: 0
En la columna Name total de valores NA: 0
En la columna Sex total de valores NA: 0
En la columna Age total de valores NA: 177
En la columna Sibsp total de valores NA: 0
En la columna Parch total de valores NA: 0
En la columna Ticket total de valores NA: 0
En la columna Fare total de valores NA: 0
En la columna Cabin total de valores NA: 0
En la columna Embarked total de valores NA: 0
>

```

#funcion `en_blanco(dataframe)` que desliega cuantos valores en blanco posee cada variable

```

en_blanco <- function(x){

  sum = 0

  for(i in 1:ncol(x))

  {

    cat("En la columna",colnames(x[i]),"total de valores en blanco:",colSums(x[i]==""),"\n")

  }

}

en_blanco(titanic_data)

```

Obtenemos el siguiente resultado:

```

> #funcion en_blanco(dataframe) que desliega cuantos valores en blanco posee cada variable
> en_blanco <- function(x){
+   sum = 0
+   for(i in 1:ncol(x))
+   {
+     cat("En la columna",colnames(x[i]),"total de valores en blanco:",colsums(x[i]==""),"\n")
+   }
+ }
> en_blanco(titanic_data)
En la columna PassengerId total de valores en blanco: 0
En la columna Survived total de valores en blanco: 0
En la columna Pclass total de valores en blanco: 0
En la columna Name total de valores en blanco: 0
En la columna Sex total de valores en blanco: 0
En la columna Age total de valores en blanco: NA
En la columna Sibsp total de valores en blanco: 0
En la columna Parch total de valores en blanco: 0
En la columna Ticket total de valores en blanco: 0
En la columna Fare total de valores en blanco: 0
En la columna Cabin total de valores en blanco: 687
En la columna Embarked total de valores en blanco: 2
> |

```

Entonces, las variables de interés son **Age** y **Embarked** (Cabin no lo consideramos porque no consideramos en nuestro análisis la correlación de cuantas personas sobrevivieron según el tipo de cabina que ocupaba).

a. Investiguemos que pasajeros tienen en blanco el atributo Embarked

```

> titanic_data$PassengerId[titanic_data$Embarked == ""]
[1] 62 830

```

Son dos, los pasajeros con PassengerId 68 y 830 los que no poseen un puerto de embarque.

b. Verifiquemos ahora en que clase viajaban dichos pasajeros y cuanto les costó el boleto.

```

> titanic_data$Pclass[titanic_data$PassengerId == 62]
[1] 1
Levels: 1 2 3
> titanic_data$Fare[titanic_data$PassengerId == 62]
[1] 80

- -
> titanic_data$Pclass[titanic_data$PassengerId == 830]
[1] 1
Levels: 1 2 3
> titanic_data$Fare[titanic_data$PassengerId == 830]
[1] 80

```

Observamos que coincidentemente, ambos pasajeros pertenecían a 1 = Primera clase y pagaron 80 por sus boletos. ¿Entonces, en donde embarcaron?

c. Creamos un nuevo dataset que no contenga a esos dos pasajeros

```

> library(dplyr)

> embark_fare <- titanic_data %>% filter(PassengerId != 62 & PassengerId != 830)

```

Obtendremos un dataset llamado embark_fare conteniendo 889 observaciones y 12 variables

Environment

History

Connections

📁

📄

📊

🔍 Import Dataset

🔗

Global Environment

🔍

Data

▶ embark_fare	889 obs. of 12 variables	📊
▶ titanic	887 obs. of 8 variables	📊
▶ titanic_data	891 obs. of 12 variables	📊

Functions

en_blanco	function (x)	📄
sin_valor	function (x)	📄

- d. Usamos la librería ggplot2 y scales para graficar el precio (Fare) medio que costó un boleto por clase (Pclass) y puerto de embarque (embarked).

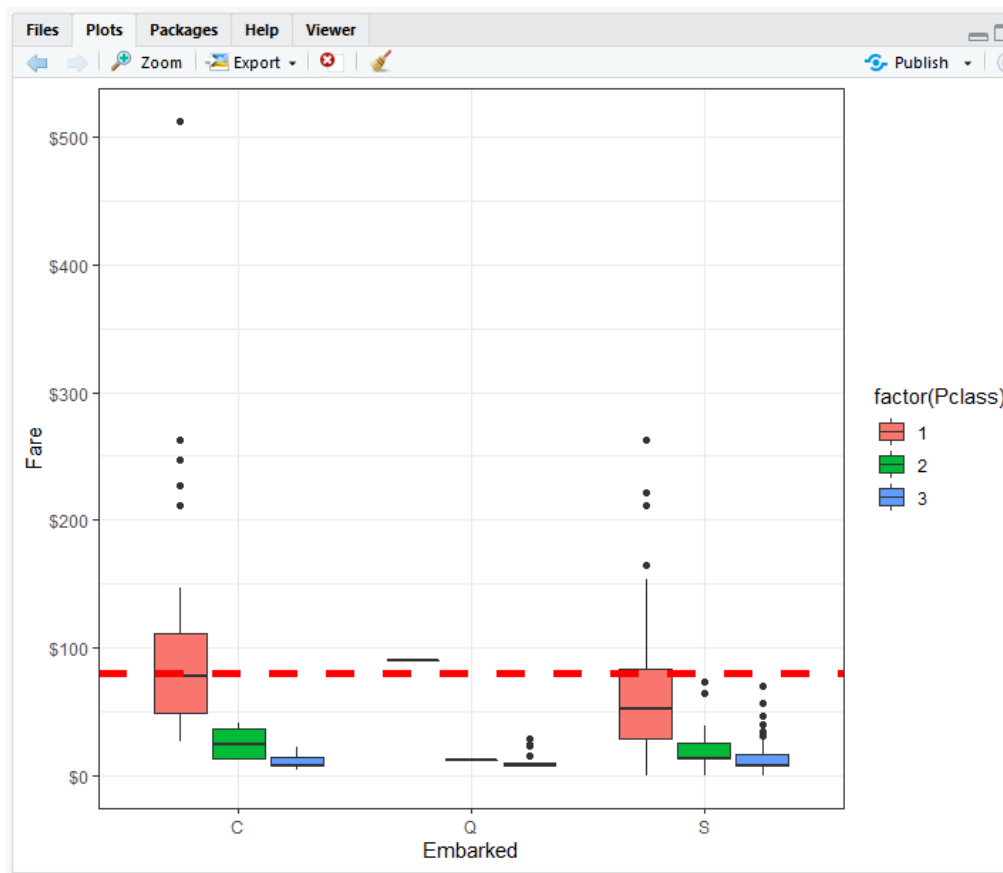
```
library(ggplot2)
```

```
library(scales)
```

```
ggplot(data = embark_fare, aes(x = Embarked, y = Fare, fill = factor(Pclass))) +
  geom_boxplot() +
  geom_hline(aes(yintercept = 80),
    colour = "red", linetype = "dashed", lwd = 2) +
  scale_y_continuous(labels = dollar_format()) +
  theme_bw()
```

```
> library(ggplot2)
> library(scales)
> ggplot(data = embark_fare, aes(x = Embarked, y = Fare, fill = factor(Pclass))) +
+   geom_boxplot() +
+   geom_hline(aes(yintercept = 80),
+               colour = "red", linetype = "dashed", lwd = 2) +
+   scale_y_continuous(labels = dollar_format()) +
+   theme_bw()
>
```

Se obtiene la siguiente gráfica:



De esta gráfica vemos que la tarifa media para el pasajero de primera clase que sale del puerto C (Charbourg) coincide muy bien con los \$ 80 pagados por los pasajeros que no tienen puerto de embarque. Entonces podemos reemplazar con seguridad los datos en blanco de aquellos pasajeros con C.

```
> titanic_data$Embarked[c(62, 830)] <- "C"
```

NOTA: Los datos NA en la variable Age, queda como propuesta de solución para el estudiante.

IV. VISUALIZACION GRAFICA

La visualización de datos es un arte de convertir los datos en conocimientos que se pueden interpretar fácilmente. En esta práctica, analizaremos los patrones de **supervivencia** y buscaremos aquellos factores que lo afectaron.

Ahora que conocemos el conjunto de datos y sus variables, necesitamos identificar las variables de interés. El conocimiento del dominio y la correlación entre variables ayudan a elegirlas. Para simplificar, hemos elegido solo 3 de estas variables: **Edad**, **Sexo**, **Pclass**. Visualicemos cuantas personas sobrevivieron según la edad, genero y clase de viajero.

Preguntas a las que debemos responder:

¿Cuál fue la tasa de supervivencia?

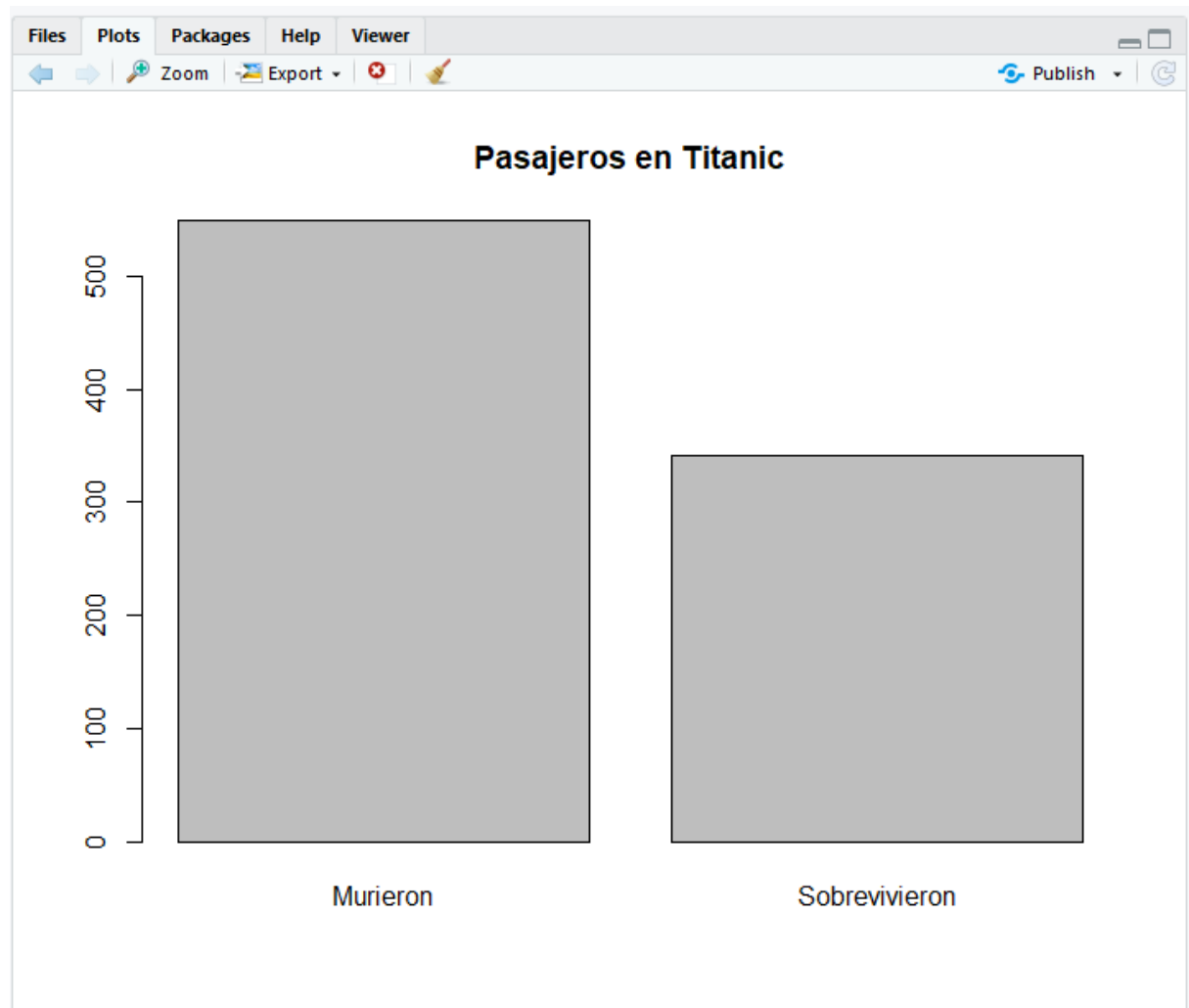
Los datos nos pueden decir cuantas personas sobrevivieron. Utilicemos un gráfico de barras simple para demostrarlo.

a. **Sobrevivencia de Pasajeros del Titanic:** Más pasajeros murieron que los que sobrevivieron

```
> table(titanic_data$Survived)
```

```
 0    1  
549 342
```

```
barplot(table(titanic_data$Survived), main="Pasajeros en Titanic", + na  
mes= c("Murieron", "Sobrevivieron"))
```



En el eje X tenemos la variable Survived, 0 representa a los pasajeros que no sobrevivieron y 1 representa a los pasajeros que sobrevivieron. El eje Y representa el número de pasajeros. Aquí vemos que más de 549 pasajeros no sobrevivieron y 342 pasajeros sobrevivieron.

Dejemos que sea más claro usando la verificación de los porcentajes.

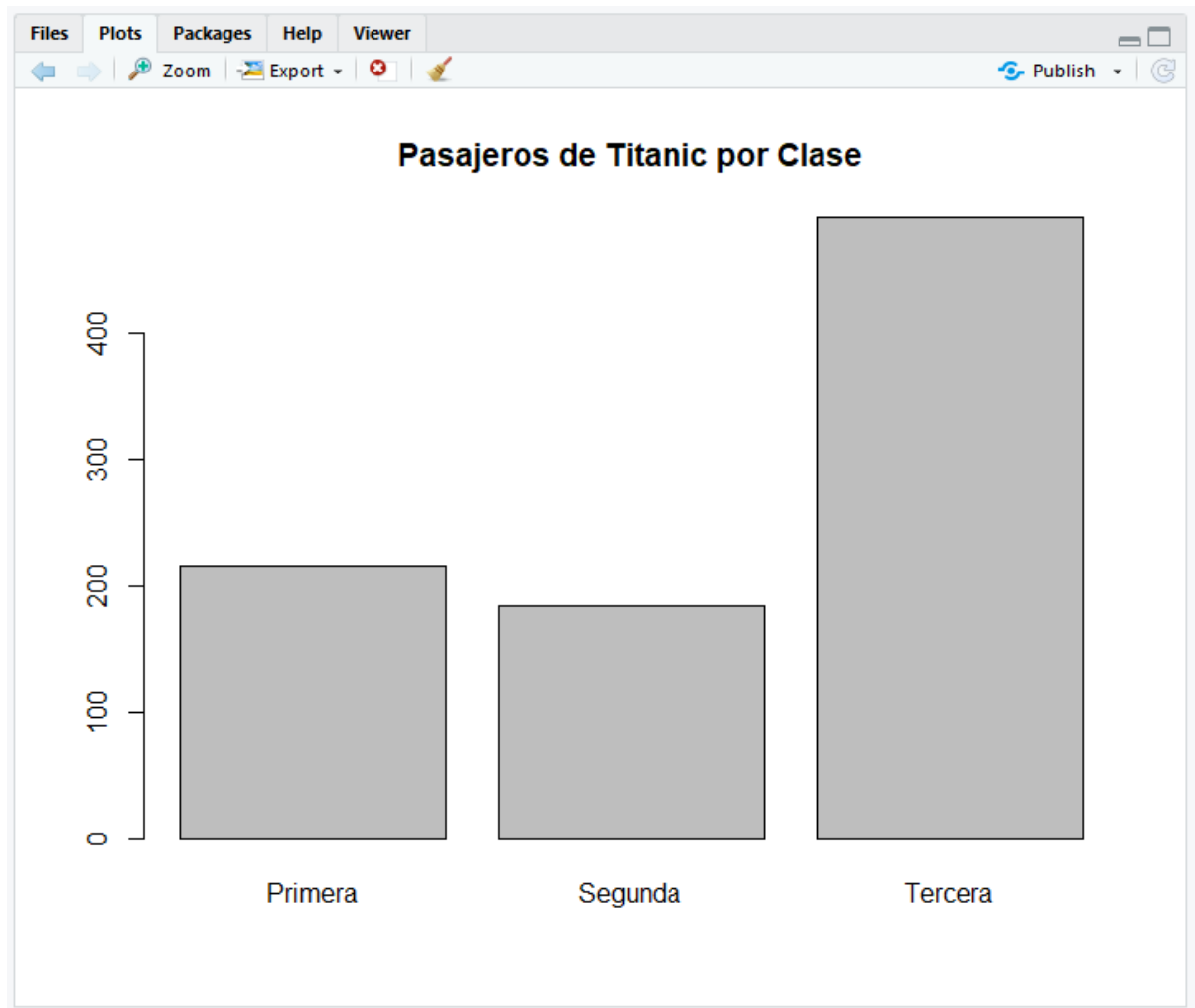
```
> prop.table(table(titanic_data$Survived))
```

```
> prop.table(table(titanic_data$Survived))
```

```
      0      1  
0.6161616 0.3838384
```

- b. **Pasajeros del Titanic por Clase:** la tercera clase de pasajeros fue la más poblada, y por ende, la de costo por boleto más económico

```
> barplot(table(titanic_data$Pclass), main="Pasajeros de Titanic por Clase", names= c("Primera", "Segunda", "Tercera"))
```

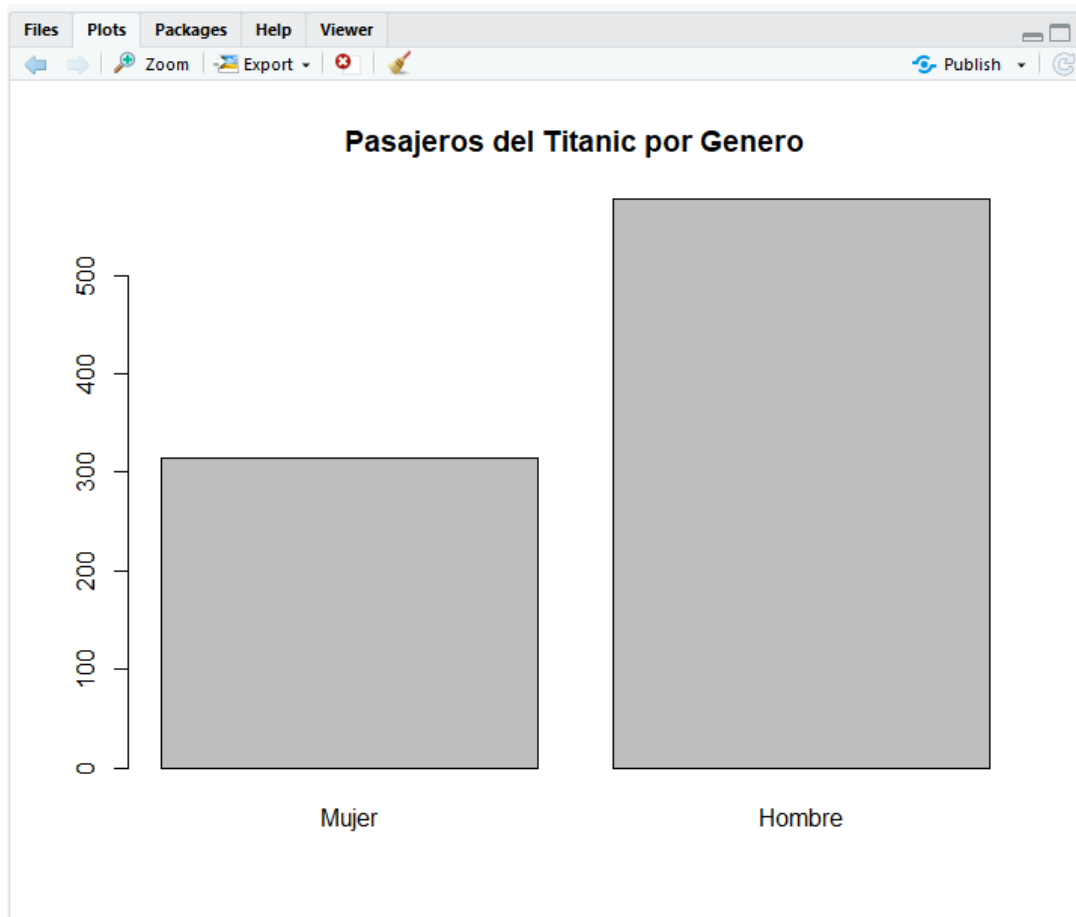


- c. **Pasajeros del Titanic por Genero:** existieron mucho más pasajeros hombres que mujeres.

```
> table(titanic_data$Sex)
```

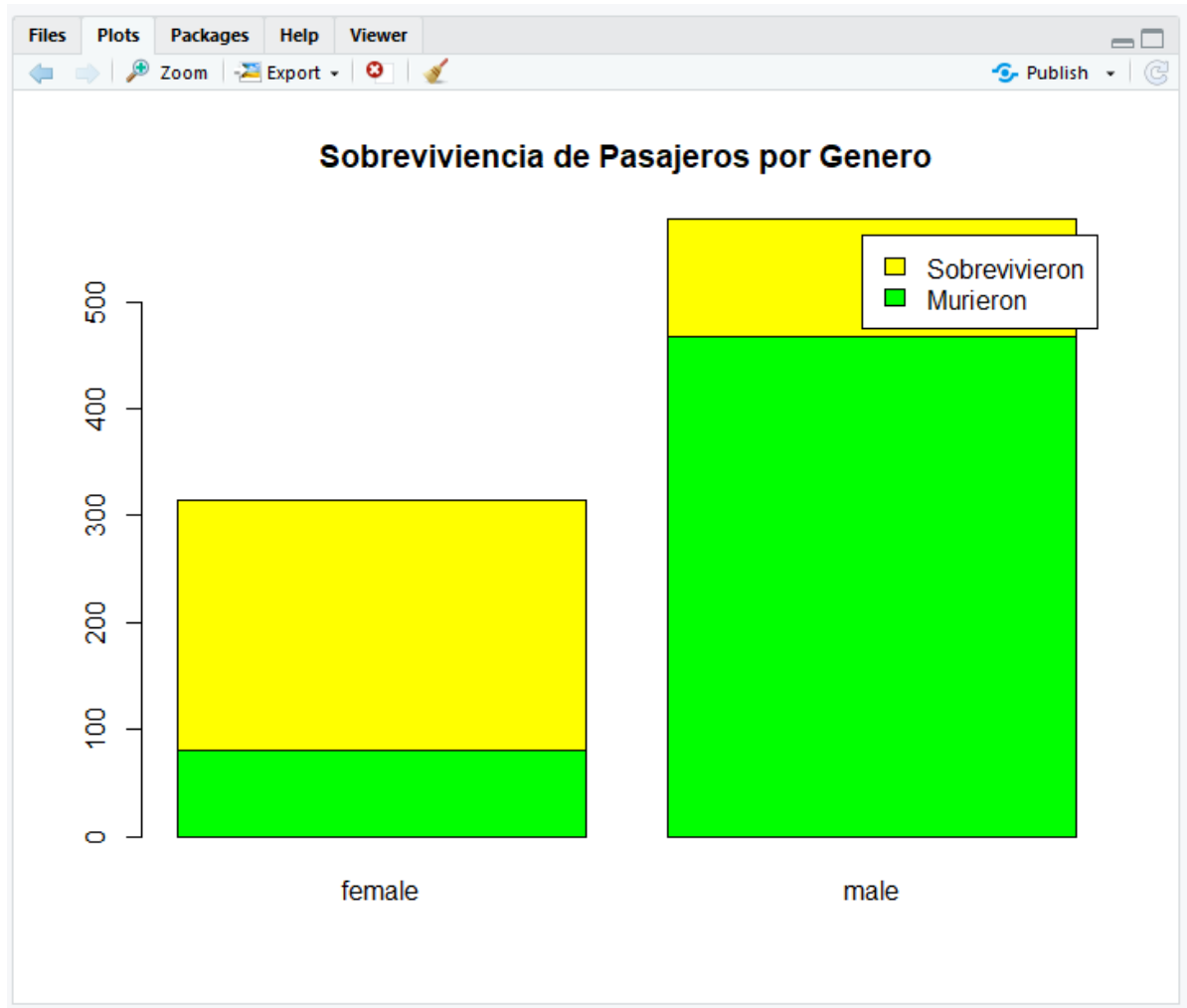
```
female  male  
   314    577
```

```
> barplot(table(titanic_data$Sex), main="Pasajeros del Titanic por Genero", names= c("Mujer", "Hombre"))
```



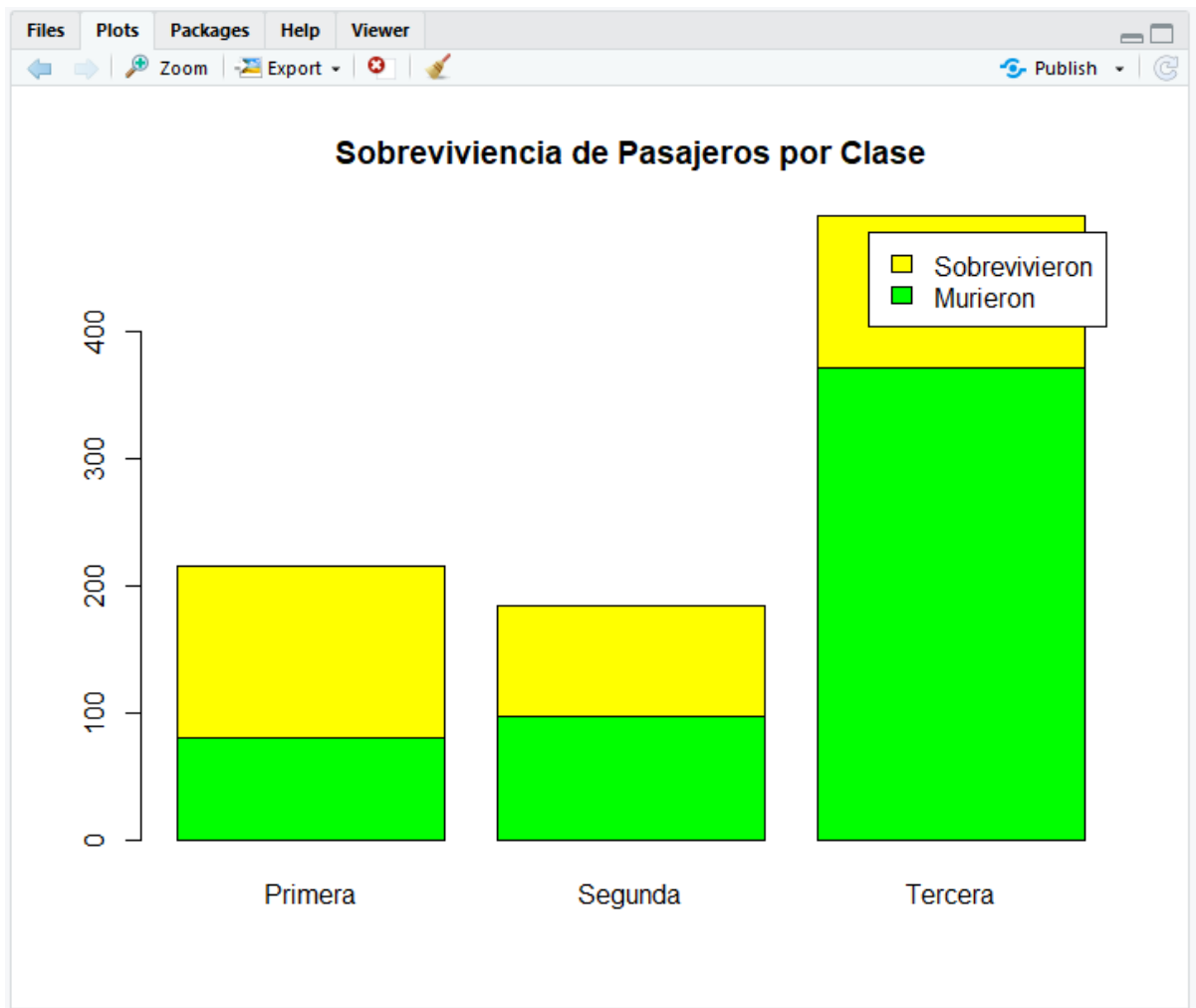
- d. **Sobrevivencia de Pasajeros por Genero:** las mujeres tuvieron una mayor tasa de supervivencia que los hombres durante el naufragio. Explicado en parte por el protocolo de "mujeres y niños primero".

```
> counts = table(titanic_data$Survived, titanic_data$Sex)
> barplot(counts, col=c("green","yellow"), legend = c("Murieron", "Sobrevivieron"), main = "Sobrevivencia de Pasajeros por Genero")
```



- e. **Sobreviviencia de Pasajeros por Clase:** la primera clase tuvo una mayor tasa de supervivencia, acorde a la calidad/costo del pasaje, pero la suposición de personas adineradas que tienden a sobrevivir más puede que no ser tan veraz.

```
> counts1 = table(titanic_data$Survived, titanic_data$Pclass)
> barplot(counts1, col=c("green","yellow"), legend = c("Murieron","Sobrevivieron"), main = "Sobreviviencia de Pasajeros por Clase", names= c("Primera", "Segunda", "Tercera"))
```



V. CONCLUSIONES PRELIMINARES

- Solo el 38,38% de los pasajeros que abordaron el Titanic sobrevivieron.
- Abordaron muchos más pasajeros hombres que mujeres.
- La mayoría de pasajeros pertenecían a la tercera clase.
- Vemos que la tasa de supervivencia entre las mujeres fue significativamente mayor en comparación con los hombres. Explicado en parte por el protocolo de "mujeres y niños primero".