titanic_dataset_project

Manuel Herrera Lara y Anahí Berumen Murillo

15/9/2020

The data domain

Usaremos Machine Learning para crear un modelo que predice cuáles pasajeros sobrevivieron al naufragio del titanic y/o qué tipo de personas tenían mas probabilidades de sobrevivir, usando información de los pasajeros que viajaban en el titanic; como su nombre, edad, sexo, clase socioeconómica, etc. Como breve descripción podemos decir que el hundimiento del titanic fue uno de los naufragios mas infames y recordados de la historia. El RMS Titanic fue un crucero de pasajeros británico que se hundió en el Océano Atlántico Norte y esto sucedió el 15 de abril de 1912, durante su viaje inaugural; y el RMS Titanic, considerado "insumergible", se hundió tras chocar con un iceberg. Desafortunadamente, no había suficientes botes salvavidas para todos a bordo, lo que resultó en la muerte de 1502 de los 2224 pasajeros y la tripulación. Al parecer algunos grupos de personas tenían más probabilidades de sobrevivir que otros. Y por último destacamos que el Titanic era el barco más grande a flote en el momento y fue construido por el astillero Harland and Wolff en Belfast.

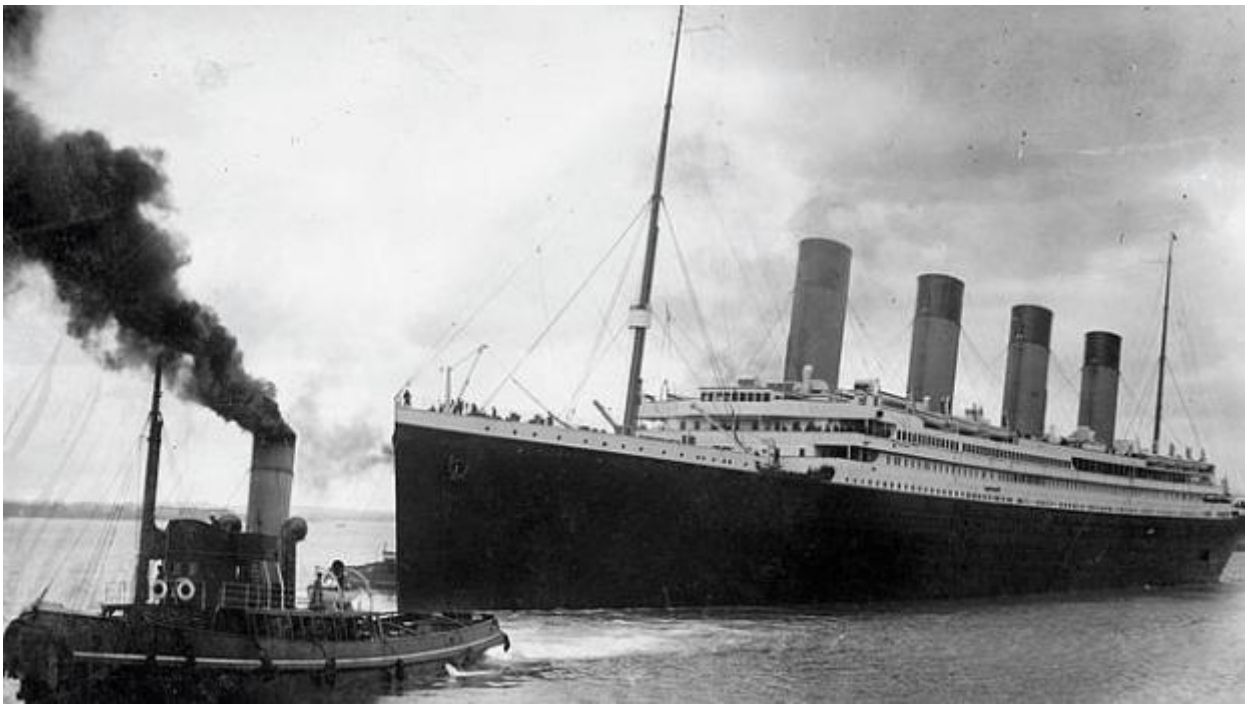


Figure 1: Titanic

Describe each of the variables

» Variable, definición y/o posibles valores

PassengerId Número de identificación del pasajero.
Survived Indica si el pasajero sobrevivió. 0 = No, 1 = Yes.
Pclass Define la clase socioeconómica del pasajero. 1 = Baja, 2 = Media y 3 = Alta.
Name Nombre del pasajero.
Sex Género del pasajero. Masculino y/o Femenino.
Age Edad del pasajero.
SibSp Número de hermanos y/o cónyuges a bordo.
Parch Número de padres y/o niños a bordo.
Ticket Número de boleto del pasajero.
Fare Tarifa de pasajero.
Cabin Número de cabina del pasajero.
Embarked Puerto de embarcación. (C = Cherbourg; Q = Queenstown; S = Southampton).

Notas adicionales para algunas variables

pclass Indica el status o clase socioeconómica del pasajero.
1 = Baja
2 = Media
3 = Alta
sibsp El dataset define las relaciones familiares de esta forma:
sibling = hermano, hermana, hermanastro, hermanastra.
spouse = esposo y/o esposa.
parch El dataset define las relaciones familiares de esta forma:
parent: mamá o papá.
child: hijo, hija, hermanastro y/o hermanastra.

```
knitr::opts_chunk$set(echo = TRUE)
# path of the dataset
setwd("/home/chino/Documentos/17_materias_IS/1_mineria_de_datos/4_semana_miniproyecto1/1_titanic_dataset")

# read the dataset
titanic <- read.csv("titanic.csv", stringsAsFactors = FALSE)
```

Basic summary statics

- mostramos los primeros 10 registros del dataset

```
head(titanic, 10)
```

```
##    PassengerId Survived Pclass
## 1            1        0      3
## 2            2        1      1
## 3            3        1      3
## 4            4        1      1
## 5            5        0      3
## 6            6        0      3
## 7            7        0      1
## 8            8        0      3
## 9            9        1      3
## 10          10        1      2
##                                                  Name    Sex Age SibSp Parch
```

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

- mostramos la estructura de los datos y/o los tipos de datos de los atributos

```
str(titanic)
```

```
## '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 3 1 3 3 2 ...
##  $ Name       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)"
##  $ Sex        : chr  "male" "female" "female" "female" ...
##  $ 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     : chr  "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : chr  "" "C85" "" "C123" ...
##  $ Embarked   : chr  "S" "C" "S" "S" ...
```

- resumen con las medidas estadisticas básicas

```
summary(titanic)
```

```
##   PassengerId       Survived          Pclass          Name
##  Min.   :  1.0   Min.   :0.0000   Min.   :1.000   Length:891
##  1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000   Class :character
##  Median :446.0   Median :0.0000   Median :3.000   Mode  :character
##  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
##
##      Sex                 Age            SibSp           Parch
##  Length:891         Min.   : 0.42   Min.   :0.000   Min.   :0.0000
##  Class :character   1st Qu.:20.12   1st Qu.:0.000   1st Qu.:0.0000
##  Mode  :character   Median :28.00   Median :0.000   Median :0.0000
##                     Mean   :29.70   Mean   :0.523   Mean   :0.3816
##                     3rd Qu.:38.00   3rd Qu.:1.000   3rd Qu.:0.0000
##                     Max.   :80.00   Max.   :8.000   Max.   :6.0000
##                     NA's   :177
##    Ticket               Fare           Cabin             Embarked
##  Length:891         Min.   :  0.00   Length:891         Length:891
##  Class :character   1st Qu.:  7.91   Class :character   Class :character
##  Mode  :character   Median : 14.45   Mode  :character   Mode  :character
##                     Mean   : 32.20
##                     3rd Qu.: 31.00
##                     Max.   :512.33
##
```

Exploring the Categorical Variables

```
table(titanic$Sex)
```

» Gender Grouping

```
##
## female   male
##    314    577
```

```
sex_table <- table(titanic$Sex)
sex_pct <- prop.table(sex_table) * 100
round(sex_pct, digits = 1)
```

Showing Percentages

```
##
## female   male
##   35.2   64.8
```

```
table(titanic$Embarked)
```

» Embarked

```
##
##       C   Q   S
##   2 168  77 644
```

```
embarked_table <- table(titanic$Embarked)
embarked_pct <- prop.table(embarked_table) * 100
round(embarked_pct, digits = 1)
```

Showing Percentages

```
##
##         C    Q    S
##  0.2 18.9  8.6 72.3
```

```r
table(titanic$Pclass)
```

» **PClass**

```
##
##   1   2   3
## 216 184 491
```

```r
pclass_table <- table(titanic$Pclass)
pclass_pct <- prop.table(pclass_table) * 100
round(pclass_pct, digits = 1)
```

Showing Percentages

```
##
##    1    2    3
## 24.2 20.7 55.1
```

```r
table(titanic$Survived)
```

» **Survived**

```
##
##   0   1
## 549 342
```

```r
survived_table <- table(titanic$Survived)
survived_pct <- prop.table(survived_table) * 100
round(survived_pct, digits = 1)
```

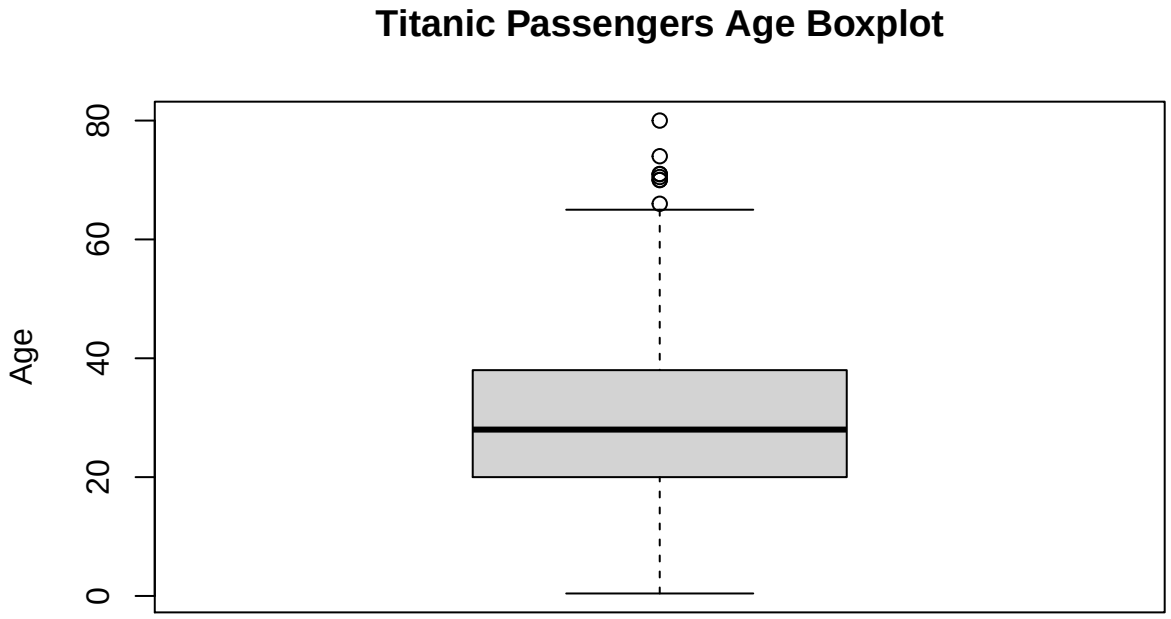
Showing Percentages

```
##
##    0    1
## 61.6 38.4
```

Boxplots - Interpretation

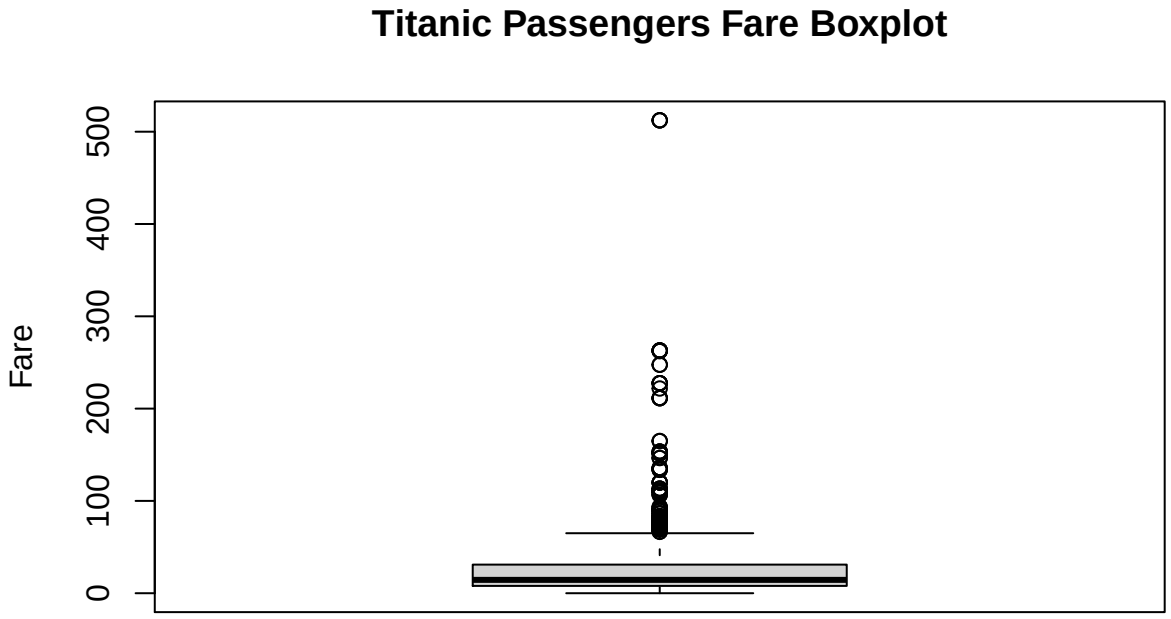
Este boxplot muestra que la edad promedio de los pasajeros que se encontraban en el titanic es de 30 años aproximadamente y la edad media esta en 28 años.
Y también podemos apreciar varios outliers o anomalías, los cuáles son datos que exceden el rango de nuestros valores normales.

```
boxplot(titanic$Age, main = "Titanic Passengers Age Boxplot", ylab ="Age")
```



Este boxplot nos muestra que la tárifa y/o costo promedio de los boletos de los pasajeros es de 32 dolares aproximadamente y la tárifa media es de 14.45
También observamos varios outliers que exceden el rango de valores normales.

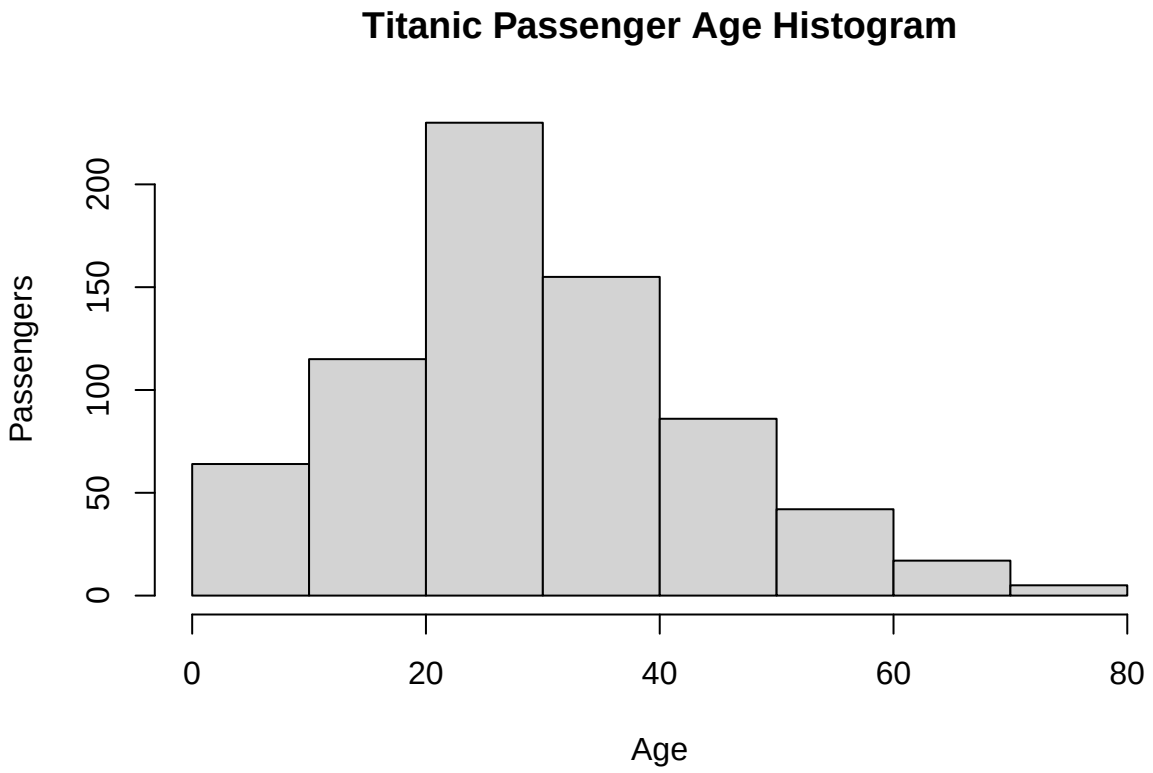
```
boxplot(titanic$Fare, main = "Titanic Passengers Fare Boxplot", ylab = "Fare")
```



Histograms-Interpretation and Skew of the data-Interpretation.

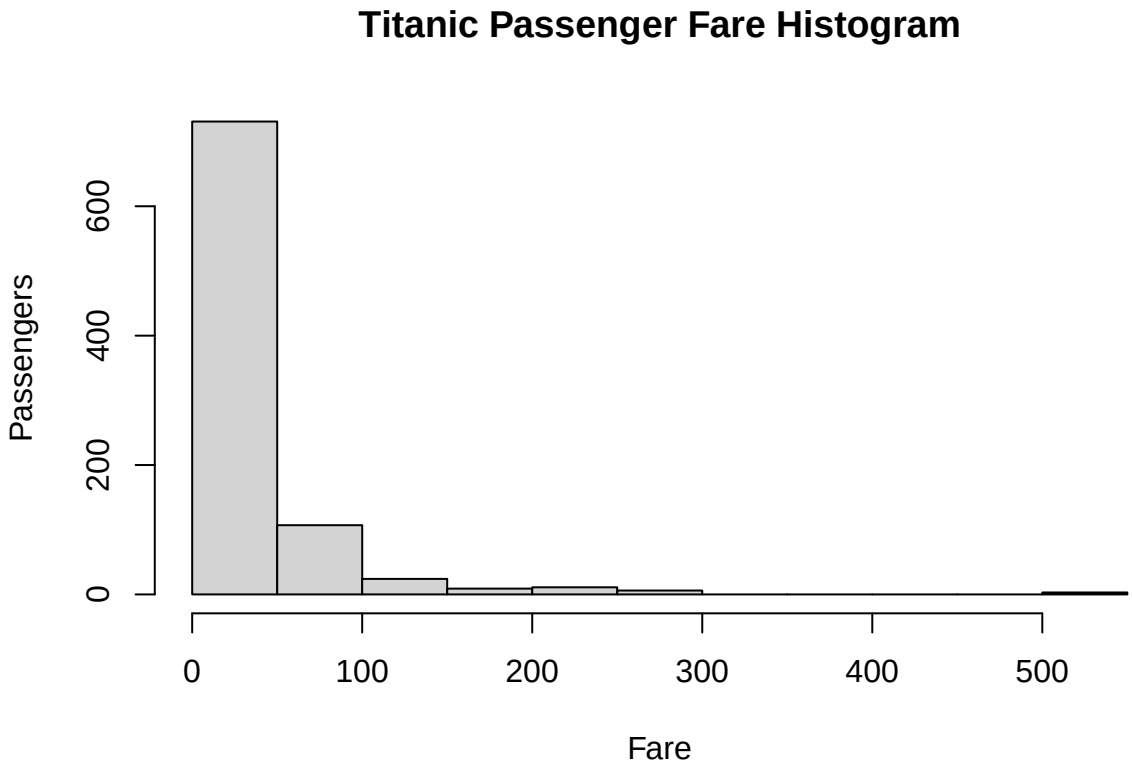
Observamos que la mayoría de los pasajeros era gente joven porque contaba con una edad de 20 a 30 años. Y es una distribución **no simétrica** ya que esta sesgada hacia la derecha, porque la edad promedio es mayor a la mediana.

```
hist(titanic$Age, main = "Titanic Passenger Age Histogram", xlab = "Age", ylab = "Passengers")
```



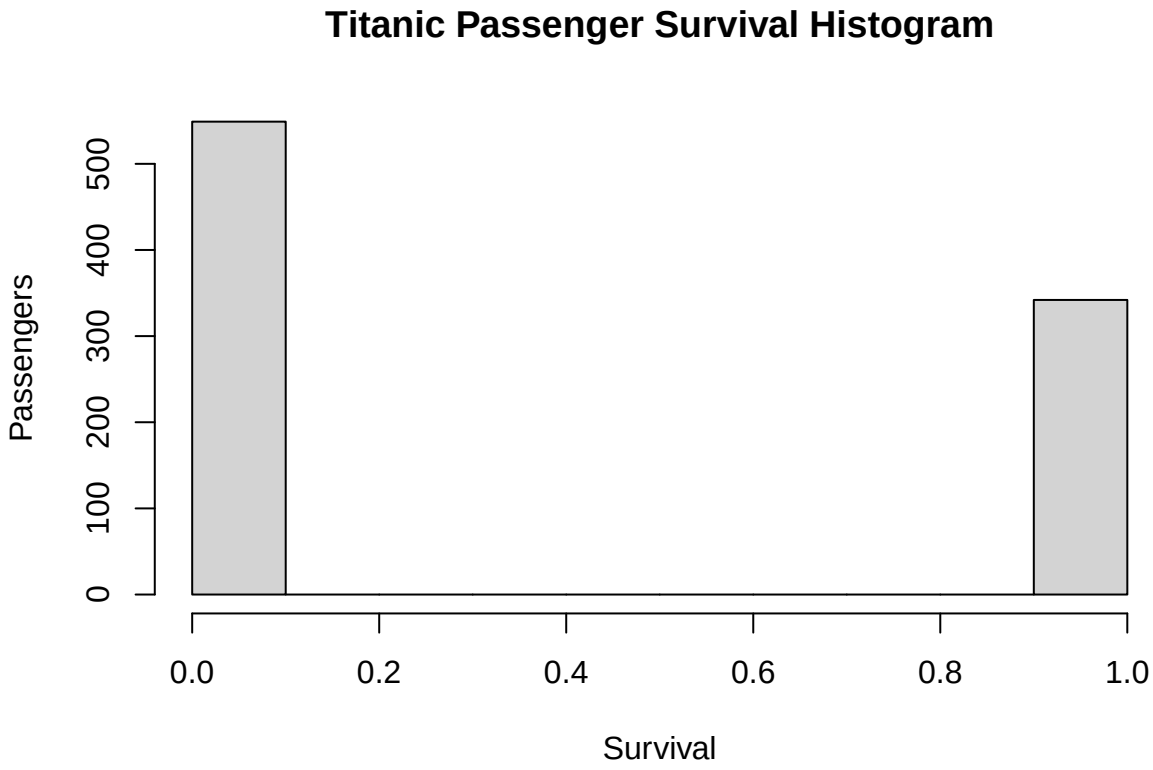
Observamos que la mayoría de los pasajeros pagó un costo menor a 100 dolares en sus boletos de abordar. Y es una distribución **no simétrica** ya que esta sesgada hacia la derecha porque la tárifa promedio es mayor a la tárifa media.

```
hist(titanic$Fare, main = "Titanic Passenger Fare Histogram", xlab = "Fare", ylab = "Passengers")
```



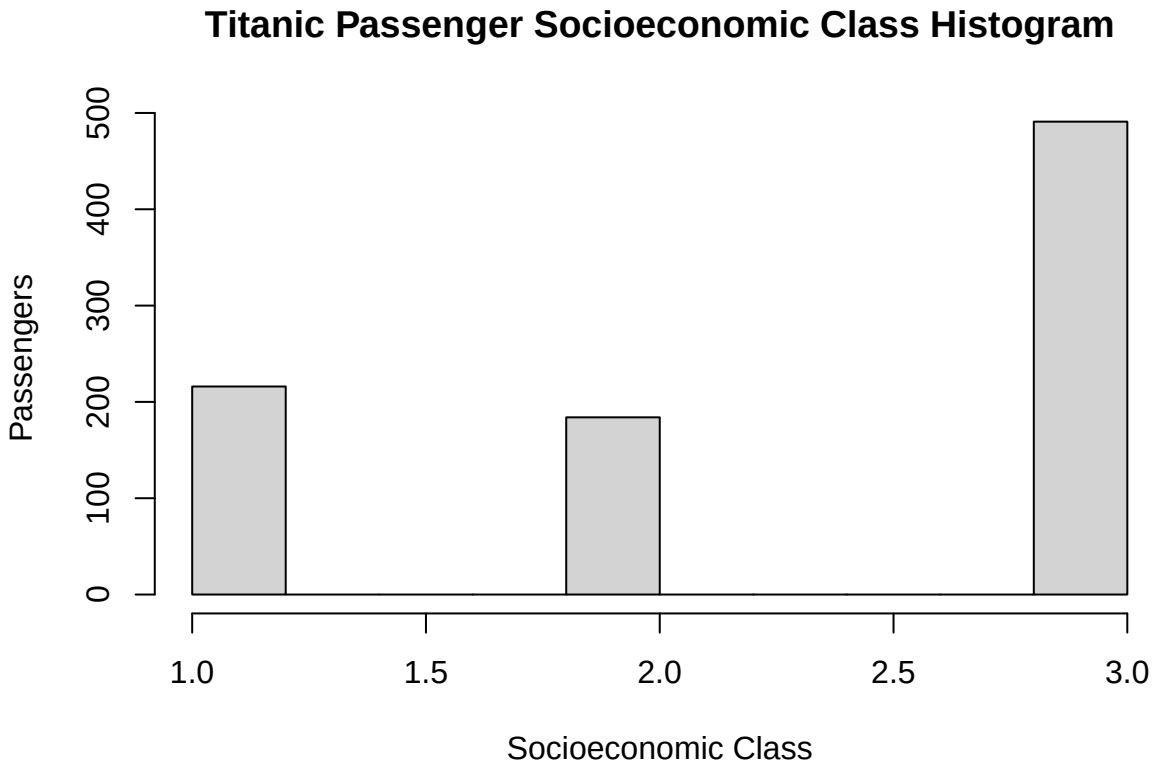
Observamos que la mayoría de los pasajeros que viajaban en el titanic murió y menos de la mitad sobrevivio. Mostrando las cifras 549 pasajeros murieron y 342 sobrevivieron.

```
hist(titanic$Survived, main = "Titanic Passenger Survival Histogram", xlab = "Survival", ylab = "Passeng
```



Observamos que la mayoria de los pasajeros era gente acaudalada o con dinero ya que pertenerían a la clase alta y menos de la mitad de los pasajeros estaban entre la clase media y baja.

```
hist(titanic$Pclass, main = "Titanic Passenger Socioeconomic Class Histogram", xlab = "Socioeconomic Cla
```



Quartiles and interpretation.

Observamos que la mayoría de los pasajeros se encuentra en un **rango de edad de 20 a 38 años.** Y esto hace que las edades máximas sean outliers o anomalías, ya que la mayoría de los datos están entre el 1er y 3er. cuartil; y esto lo dice el IQR.

```
quantile(titanic$Age,na.rm = TRUE)
```

```
##     0%    25%    50%    75%   100%
##  0.420 20.125 28.000 38.000 80.000
```

```
IQR(titanic$Age, na.rm = TRUE)
```

```
## [1] 17.875
```

Observamos que la mayoría de los pasajeros pagó una tárifa y/o costo de boleto de alrededor **de 8 a 31 dolares.** Y esto hace que los costos elevados sean considerados outliers o anomalías.

```
quantile(titanic$Fare, na.rm = TRUE)
```

```
##       0%      25%      50%      75%     100%
##   0.0000   7.9104  14.4542  31.0000 512.3292
```

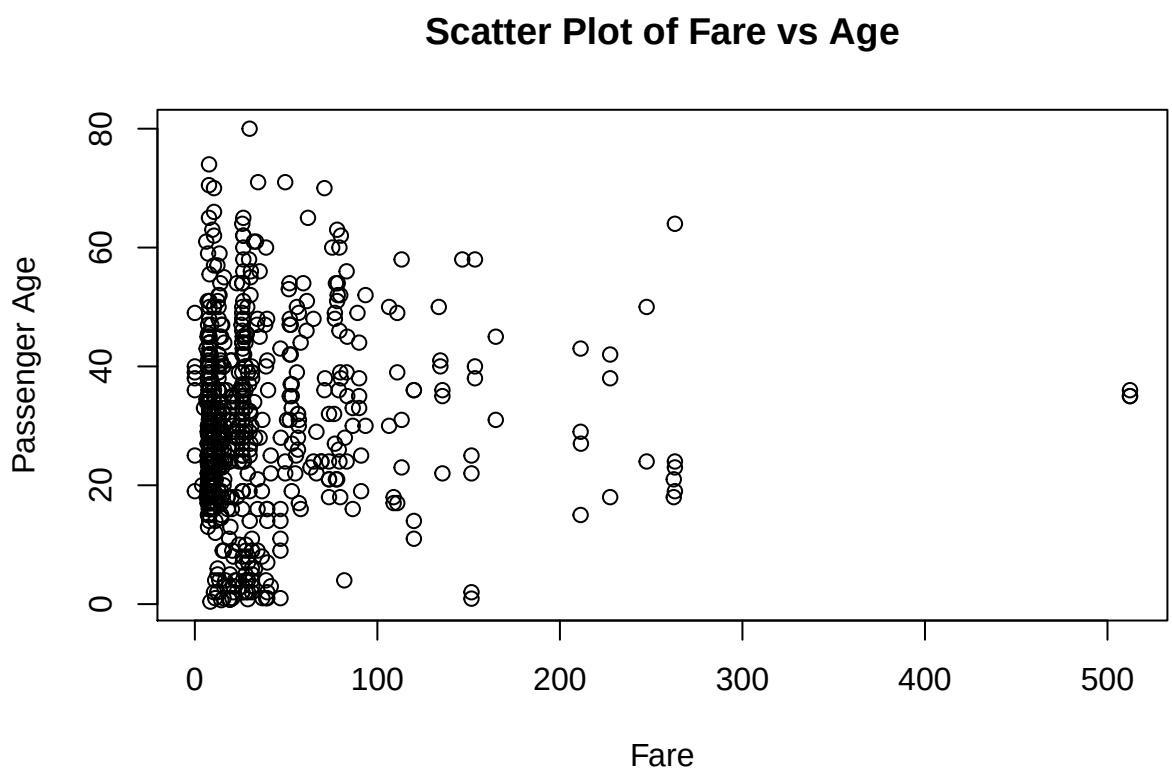
```
IQR(titanic$Fare, na.rm = TRUE)
```

```
## [1] 23.0896
```

Scatterplots. Interpretation.

Observamos que hay muchos pasajeros de 20 a 40 años que compraron un boleto de menos de 100 dolares. Y hay muy pocos pasajeros que compraron boleto con costo mayor a 100 dolares.

```
plot(x=titanic$Fare, titanic$Age, main="Scatter Plot of Fare vs Age", xlab = "Fare", ylab="Passenger Ag
```



#

Observamos que no hay ninguna relación entre el costo del boleto con las probabilidades de supervivencia.

```
plot(x=titanic$Fare, y=titanic$Survived, main = "Scatterplot of Fare vs Survived", xlab = "Fare", ylab =
```