titanic_dataset_project

Manuel Herrera Lara y Anahí Berumen Murillo

15/9/2020

1.- The data domain

It will be performed an analysis on the titanic dataset to identify which passengers survived the wreck or what kind of people were most likely to survive taking into account their characteristics such as name, age, sex, socioeconomic class, etc. As a brief description we can say that the sinking of the Titanic was one of the most infamous and remembered shipwrecks in history. The RMS Titanic was a British passenger cruise ship that sank in the North Atlantic Ocean and this happened on April 15, 1912, during its maiden voyage; and the RMS Titanic, considered "unsinkable", sank after hitting an iceberg. Unfortunately, there were not enough lifeboats for everyone on board, resulting in the deaths of 1,502 of the 2,224 passengers and crew. Apparently some groups of people were more likely to survive than others. And finally we highlight that the Titanic was the largest ship afloat at the time and was built by the Harland and Wolff shipyard in Belfast.

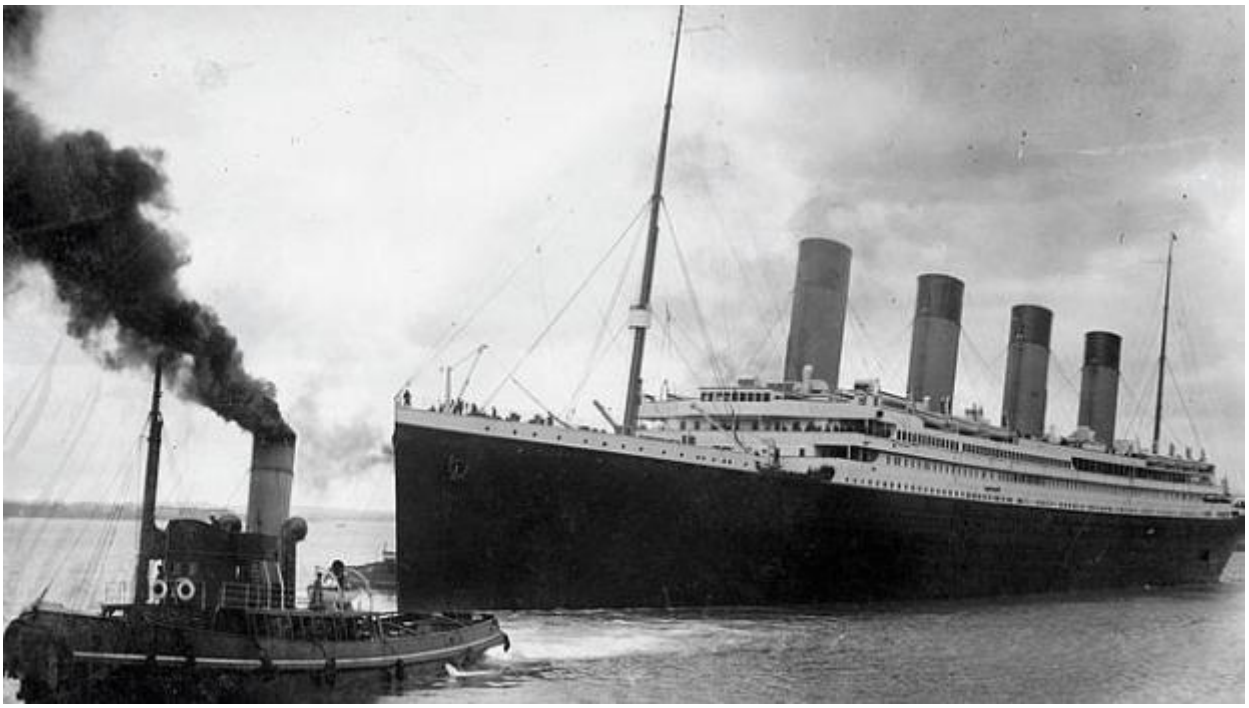


Figure 1: Titanic

2.- Describe each of the variables

Variable, definition and / or possible values

PassengerId Passenger identification number, numeric.
Survived Indicates if the passenger survived. (0 = No, 1 = Yes), categoric.
Pclass Define the socioeconomic class of the passenger. (1 = Lower class, 2 = Middle class, 3 = Upper class), ordinal.
Name Name of the passenger.
Sex Gender of the passenger. (Male and/or Female), categoric.
Age Passenger age, numeric.
Sib/Sp Number of Siblings/Spouses Aboard, discrete.
Parch Number of Parents/Children Aboard, discrete.
Ticket Passenger's ticket number.
Fare Passenger fare.
Cabin Passenger cabin number.
Embarked Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton), categoric.

Additional notes for some variables

Sibsp The dataset defines family relationships like this:
Sibling - Brother, Sister, Stepbrother, or Stepsister of Passenger Aboard Titanic.
Spouse - Husband or Wife of Passenger Aboard Titanic.
Parch Parent - Mother or Father of Passenger Aboard Titanic.
Child - Son, Daughter, Stepson, or Stepdaughter of Passenger Aboard Titanic.

```
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)
```

» **(dataset reading)**

3.- Basic summary statics

- It shows the first 10 records of the 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
```

- It shows the structure of the data and/or the data types of the attributes.

```
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" ...
```

- summary with basic statistical measures.

```
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)
```

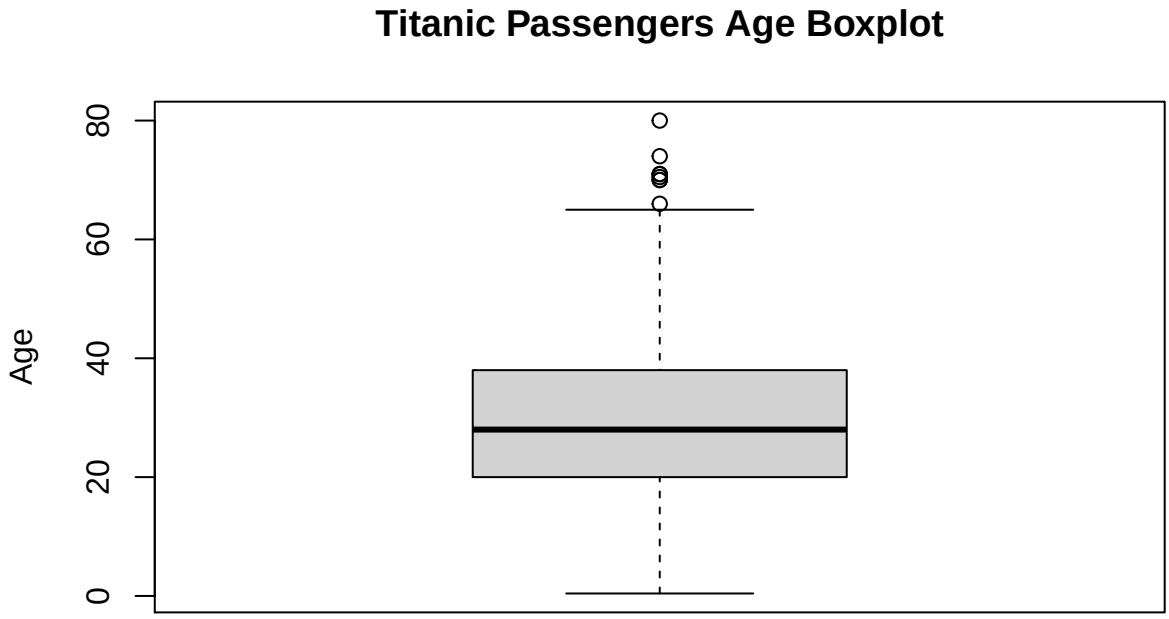
Showing Percentages

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

4.- Boxplots - Interpretation

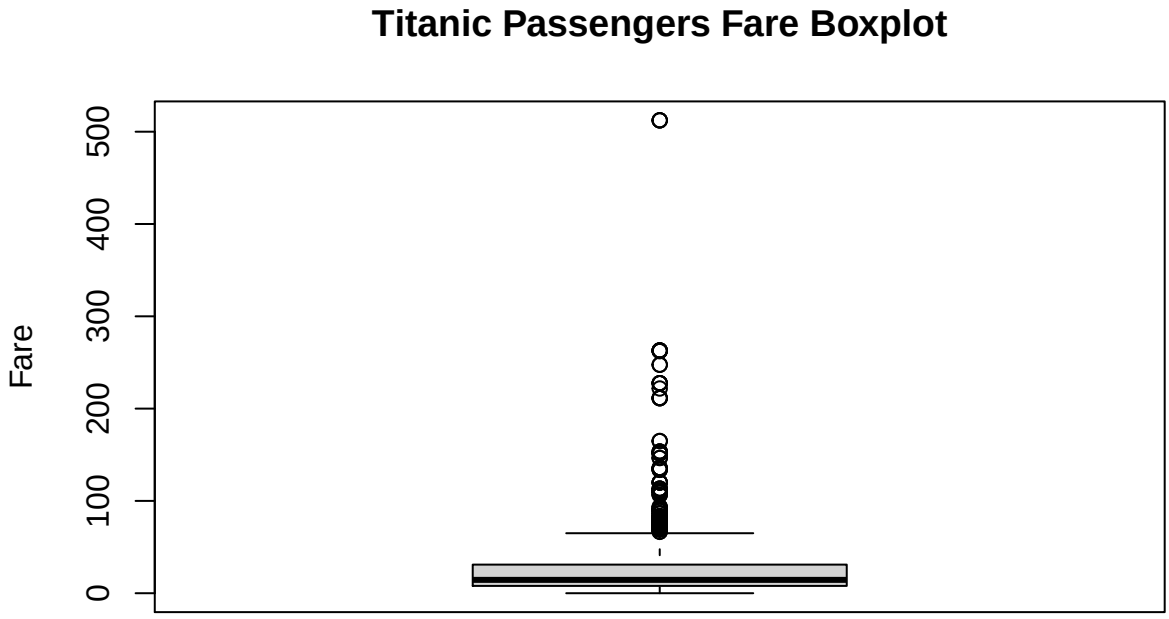
This boxplot shows that **the average age of the passengers who were on the titanic is approximately 30 years and the average age is 28 years.** And we can also appreciate several outliers or anomalies, which are data that exceed the range of our normal values.

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



This boxplot shows us that the average rate and / or cost of passenger tickets is approximately $ 32 and the average rate is 14.45
We also observed several outliers that exceed the range of normal values.

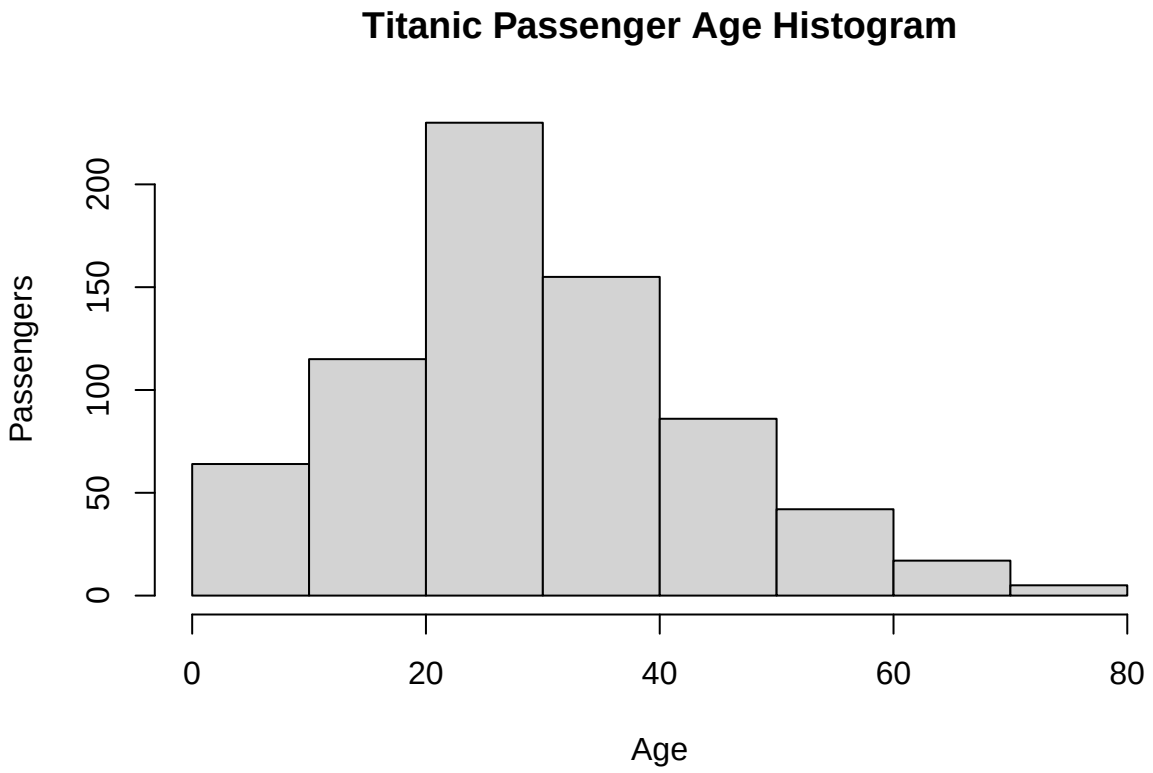
```
boxplot(titanic$Fare, main = "Titanic Passengers Fare Boxplot", ylab = "Fare")
```



5 y 6.- Histograms-Interpretation and Skew of the data-Interpretation.

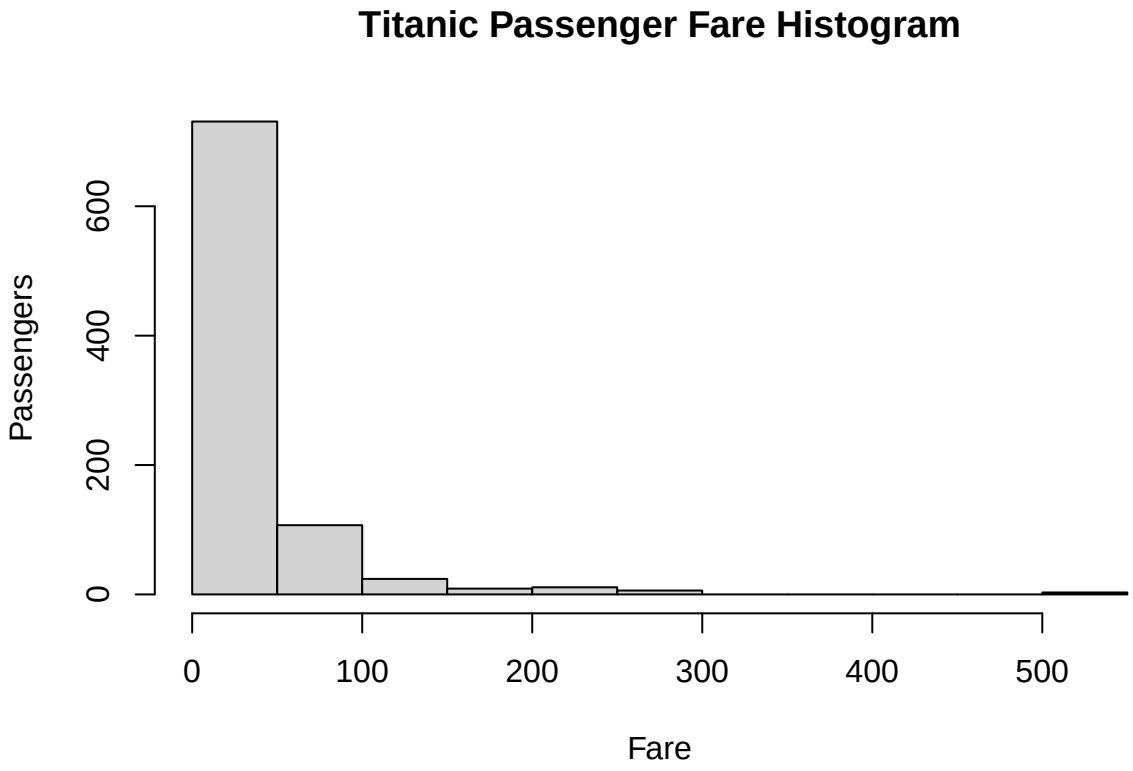
As seen in the graph, the majority of the passengers were young people because they were between 20 and 30 years old. And it is a **non-symmetric distribution** since it is skewed to the right, because the mean age is greater than the median.

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



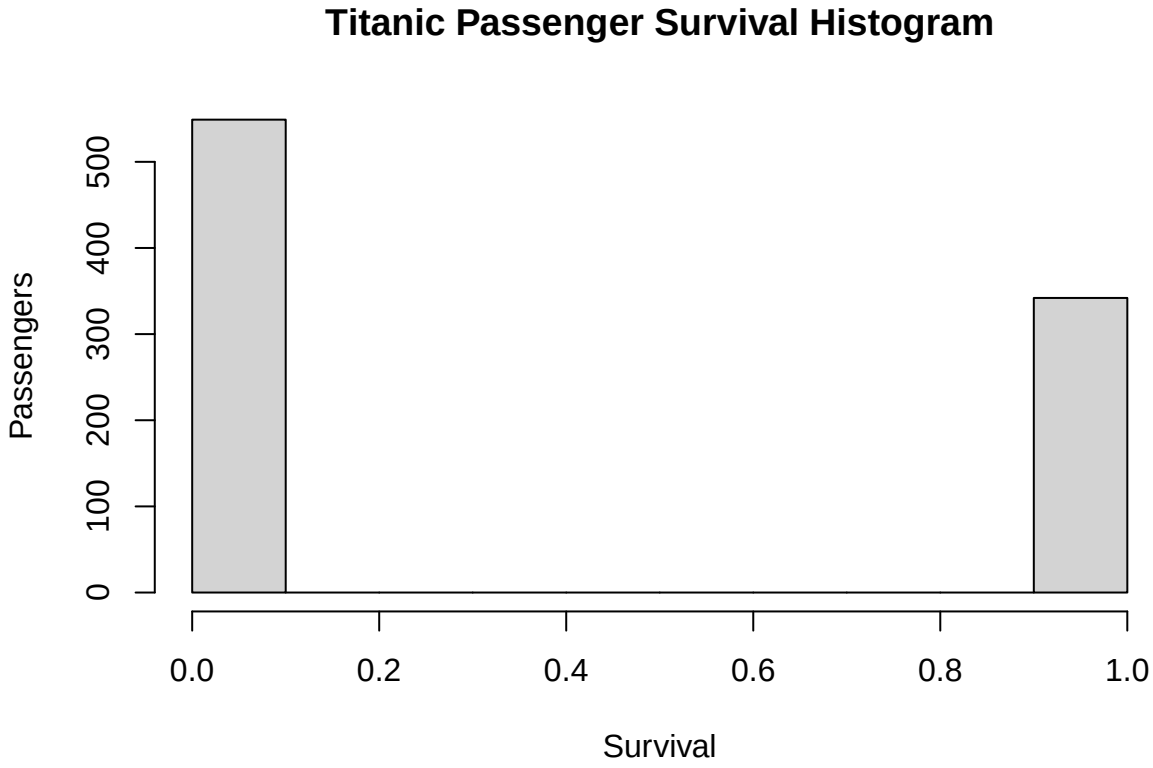
As seen in the graph, the majority of passengers paid less than $ 100 on their boarding tickets. And it is a **non-symmetric distribution** since it is skewed to the right, because the mean fare is greater than the median.

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



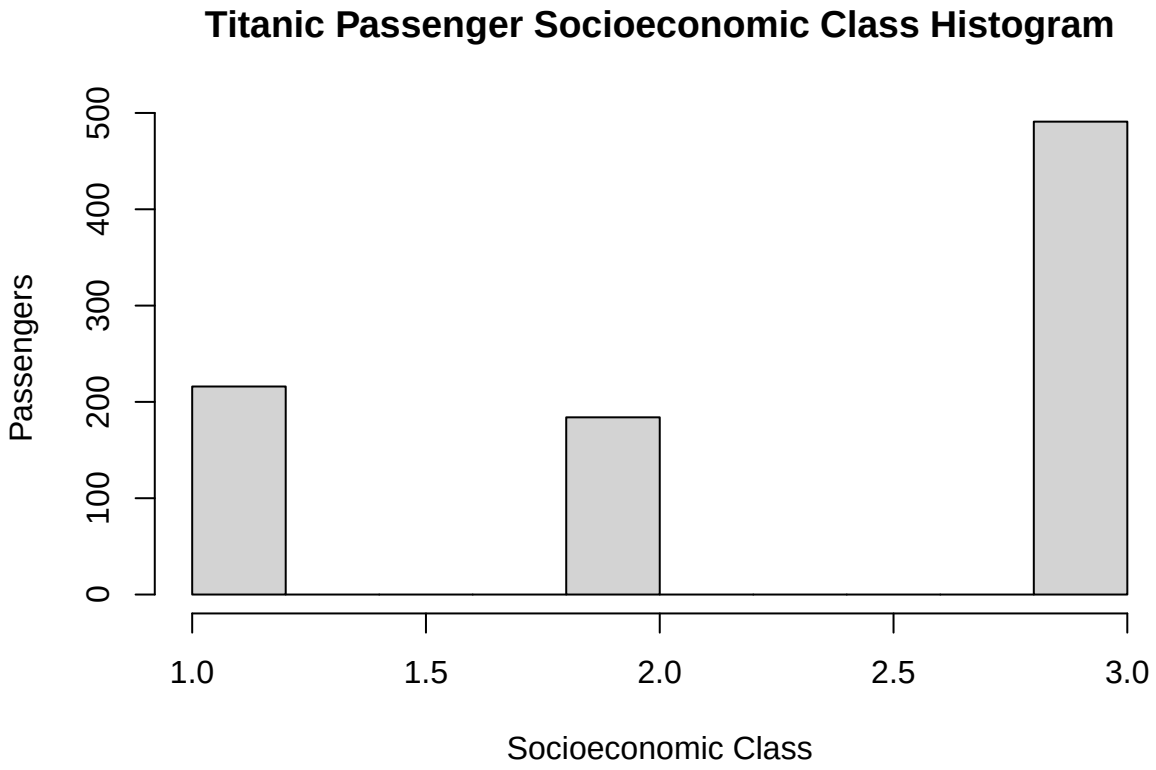
As seen in the graph, the majority of the passengers traveling on the titanic died and less than half survived. Showing the figures 549 passengers died and 342 survived.

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



It's observed that the majority of the passengers were wealthy people or with money since they belong to the upper class and less than half of the passengers were distributed between the middle and lower class.

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



7 y 8.- Quartiles and interpretation.

It's observed that the majority of passengers are in **the age range of 20 to 38 years.** And this makes the maximum ages outliers or anomalies, since most of the data is between the 1st and 3rd. quartile; and this is said by the 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
```

It's observed that the majority of passengers paid a fare and / or ticket cost of around ** 8 to 31 dollars. ** And this makes the high costs considered outliers or anomalies.

```
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
```

9.- Scatterplots. Interpretation.

As seen in the graph, there are many passengers between the ages of 20 and 40 who bought a ticket for less than $ 100. And there are very few passengers who bought a ticket with a cost greater than 100 dollars, with this plot we can see that the fare had not relation with the age of the passenger, there is no correlation.

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

As seen in the graph, there is no relationship between the cost of the ticket and the chances of survival.

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