Proyecto Titanic

autor: Luis Garcia

Aquistapace, Galarraga, Palma, Pedrosa, Portabella, Ruoti, Sanchez

Introducción

Nuestra muestra son todos los pasajeros del fatal viaje inaugural del transatlántico "Titanic". La muestra objetivo es N= 891. En nuestro caso, excluimos los datos en los que el usuario no respondió a las preguntas en algunas variables y causó la pérdida de datos. Eliminamos aquellas observaciones que eran datos perdidos, dejando una muestra de N= 714. Utilizaremos R Studio para el proceso de cálculos como medidas de tendencia central, dispersión, posiciones, tablas y gráficos. Finalmente, se realizará un análisis de los resultados más relevantes del análisis.

OBJETIVO GENERAL

Aplicar las técnicas estadísticas de descriptiva a la base de datos de Titanic, usando la asistencia del programa R Studio

DATOS

Este conjunto de datos proporciona información sobre el destino de los pasajeros en el "Titanic". Sus variables son: situación económica (clase), sexo, edad, supervivencia, tarifa edad, numero de parientes hombre y mujer.

```
  library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.6
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
## -- Attaching packages ---------------------------------------------------------
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.2     v dplyr   1.0.0
## v tidyr   1.1.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
## -- Conflicts ------------------------------------------------------------------ t
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(moments)
library(titanic)
## Warning: package 'titanic' was built under R version 4.0.3
library(agricolae)
```

```
##
## Attaching package: 'agricolae'
```

```
## The following objects are masked from 'package:moments':
##
##     kurtosis, skewness
```

```
##
## Attaching package: 'agricolae'
## The following objects are masked from 'package:moments':
##
##     kurtosis, skewness
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
```

```
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
##
## Attaching package: 'reshape2'
## The following object is masked from 'package:tidyr':
##
##     smiths
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
##
## Attaching package: 'gridExtra'
## The following object is masked from 'package:dplyr':
##
##     combine
data=titanic_train
str(data)
```

```
## '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" ...
```

```
## '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" ...
data<-na.omit(data)
```

Separamos la data en variables cualitativas y cuantitativas. La cual la renombramos con el nombre num y chr.

```
  #-------------- seleccion de muestra --------
names(data)
```

```
##  [1] "PassengerId" "Survived"    "Pclass"      "Name"        "Sex"
##  [6] "Age"         "SibSp"       "Parch"       "Ticket"      "Fare"
## [11] "Cabin"       "Embarked"
```

```
##  [1] "PassengerId" "Survived"    "Pclass"      "Name"        "Sex"
##  [6] "Age"         "SibSp"       "Parch"       "Ticket"      "Fare"
## [11] "Cabin"       "Embarked"
#variables cuantitativa
num<-data[,c(3,6,7,8,10)]

#variables cualitativa
chr<-data[,c(2,5,12)]

#renombrar variables
colnames(num)<-c("Clase_Pasajero","Edad","Pariente_Masculino","Pariente_Femenino","Tarifa_Pasaje")
names(num)
```

```
## [1] "Clase_Pasajero"     "Edad"               "Pariente_Masculino"
## [4] "Pariente_Femenino"  "Tarifa_Pasaje"
```

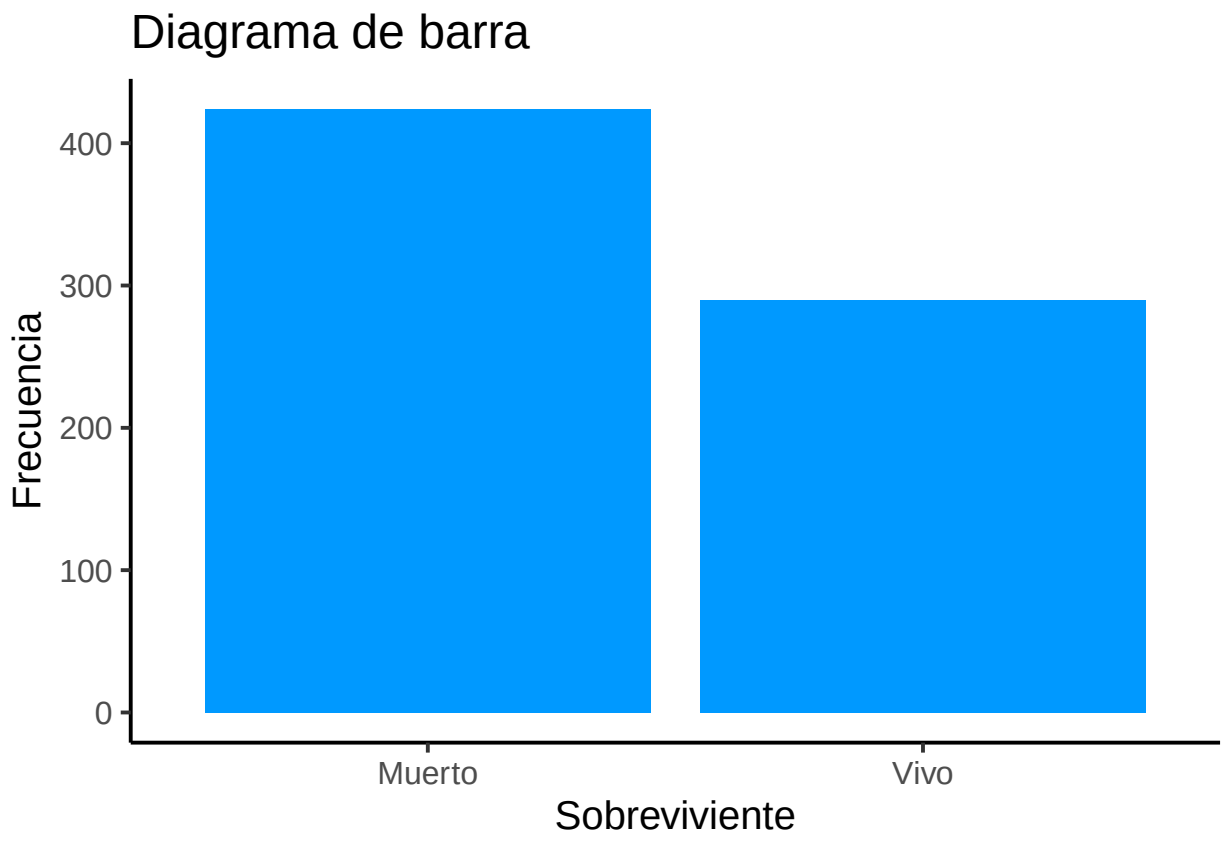
```
## [1] "Clase_Pasajero"     "Edad"               "Pariente_Masculino"
## [4] "Pariente_Femenino"  "Tarifa_Pasaje"
colnames(chr)<-c("Sobreviviente","Sexo","Puerto_Embarcadero")
names(chr)
```

```
## [1] "Sobreviviente"      "Sexo"               "Puerto_Embarcadero"
```

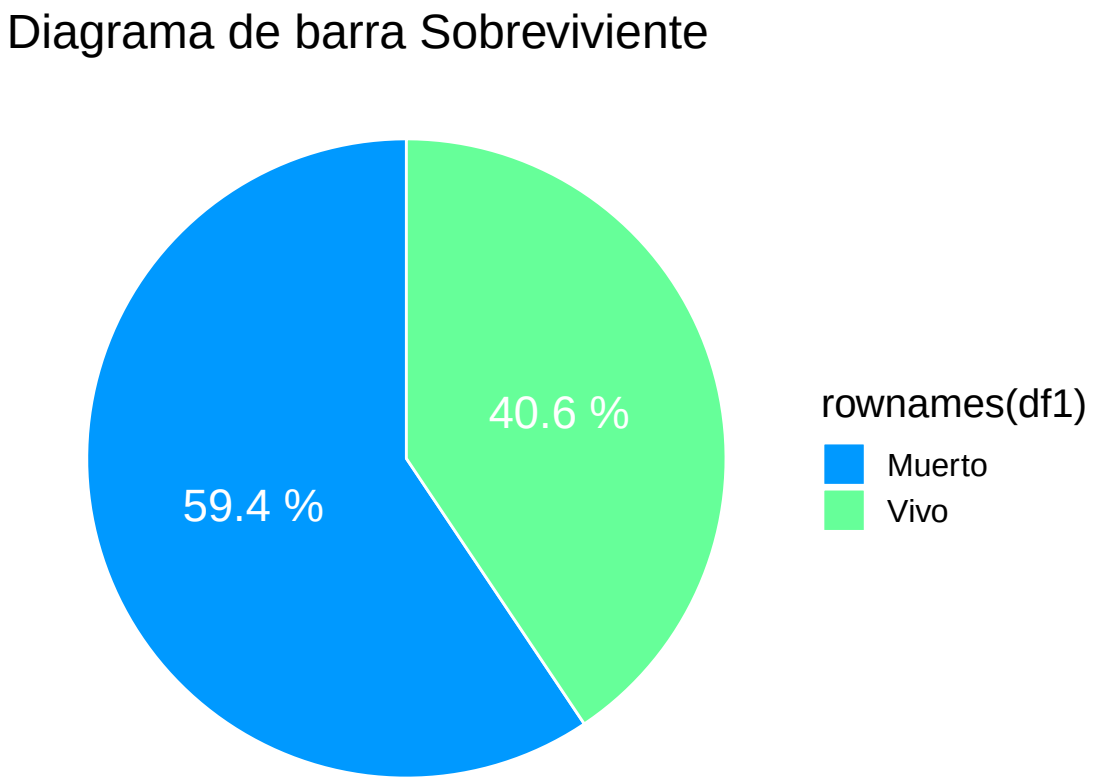
```
## [1] "Sobreviviente"      "Sexo"               "Puerto_Embarcadero"
#-------------- renombrar en chr ------------
chr[chr$Sobreviviente==0,"Sobreviviente"]<-"Muerto"
chr[chr$Sobreviviente==1,"Sobreviviente"]<-"Vivo"
chr[chr$Sexo=="female","Sexo"]<-"Mujer"
chr[chr$Sexo=="male","Sexo"]<-"Hombre"
chr[chr$Puerto_Embarcadero=="C","Puerto_Embarcadero"]<-"Cherbourg"
chr[chr$Puerto_Embarcadero=="Q","Puerto_Embarcadero"]<-"Queenston"
chr[chr$Puerto_Embarcadero=="S","Puerto_Embarcadero"]<-"Southapmtpon"
chr[chr$Puerto_Embarcadero=="","Puerto_Embarcadero"]<-"Southapmtpon"

#----------- estadistica descriptiva cualitativa-----------
resumen1<-function(x){
  round(cbind(frecuencia =table(x),relativo=prop.table(table(x))),3)}
#sobreviviente
df1<-data.frame(resumen1(chr$Sobreviviente))
df1
```

```
##        frecuencia relativo
## Muerto        424    0.594
## Vivo          290    0.406
```

```
##        frecuencia relativo
## Muerto        424    0.594
## Vivo          290    0.406
ggplot(chr,aes(Sobreviviente)) +geom_bar( fill="#0099ff" )+
  labs(title="Diagrama de barra", y="Frecuencia", x="Sobreviviente") + theme_classic(base_size=15)
```




```
ggplot(df1,aes(x="",y=relativo, fill=rownames(df1)))+geom_bar(stat = "identity",color="white")+
theme_void(base_size=15)+coord_polar(theta="y")+labs(title="Diagrama de barra Sobreviviente")+
scale_fill_manual(values = c("#0099ff","#66ff99"))+
geom_text(aes(label=paste(relativo*100,"%")),position=position_stack(vjust=0.5),color="white",size=6)
```

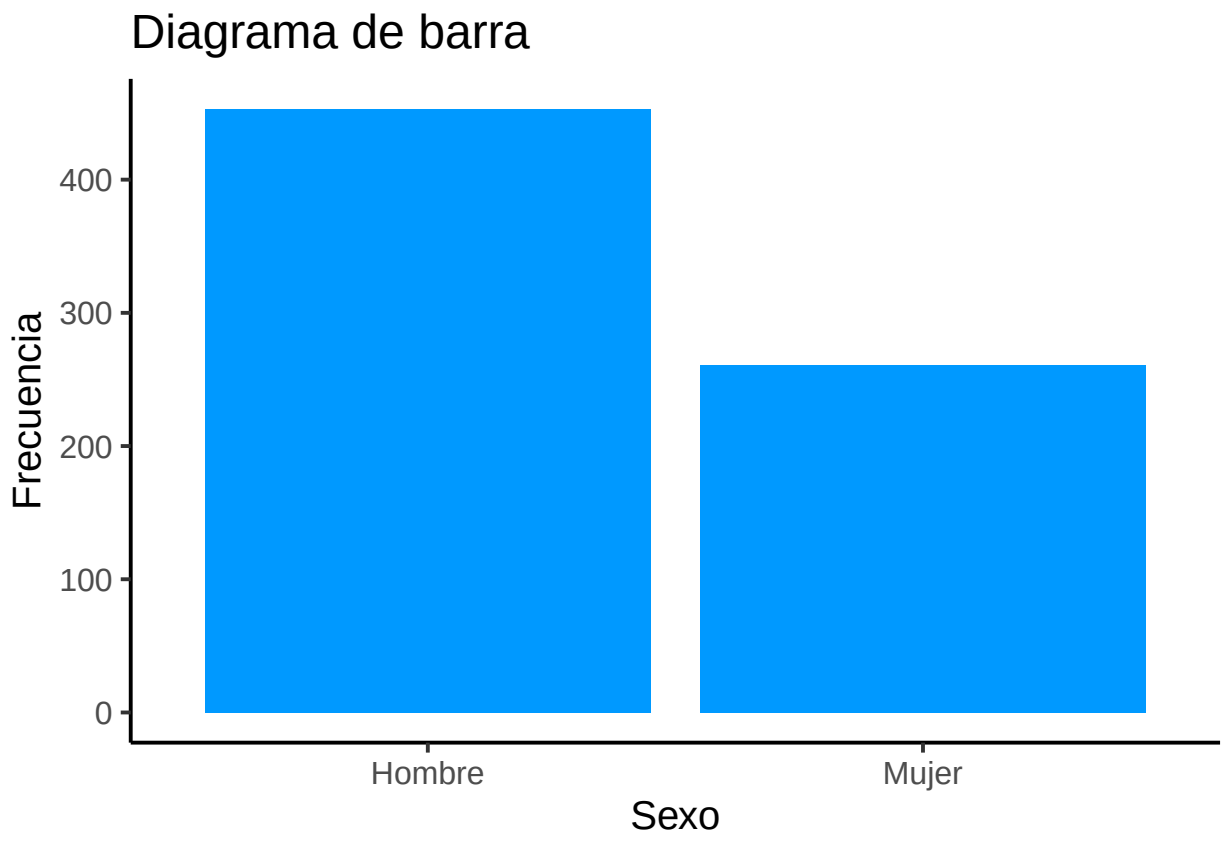


Se tiene la variable sobreviviente donde de los 714 personas el 59.4% fallecio en el accidente y solo 40.6% lograron sobrevivir.

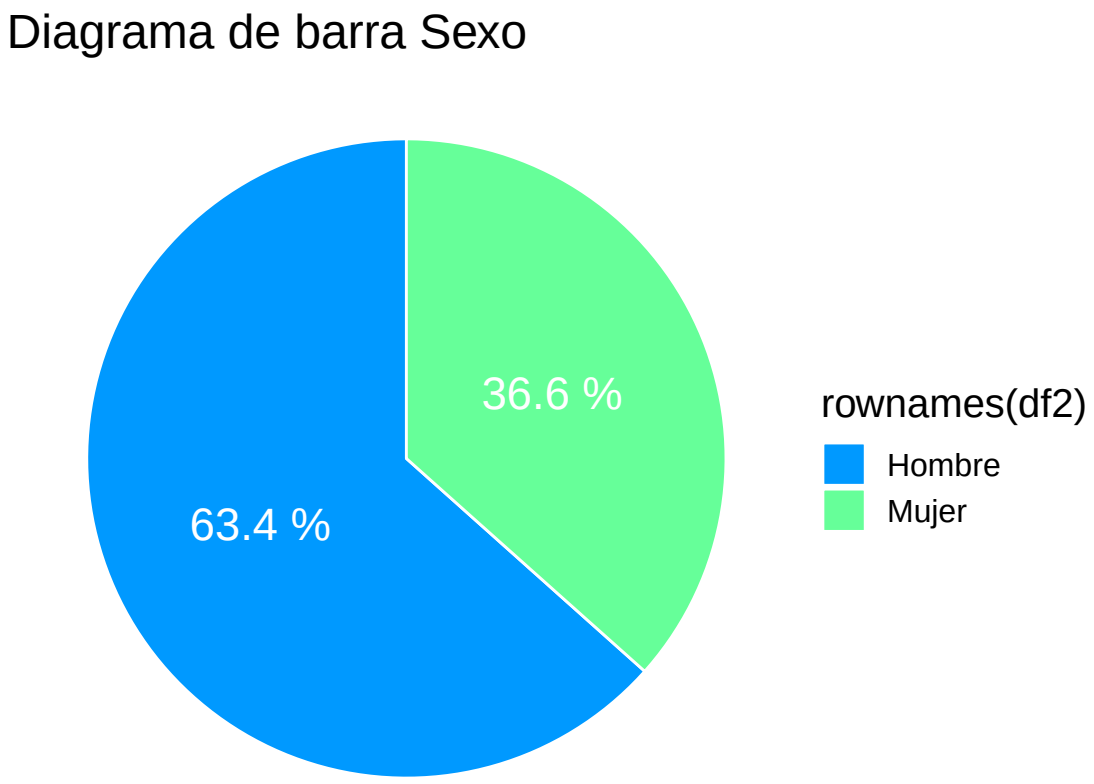
```
  #sexo
df2<-data.frame(resumen1(chr$Sexo))
df2
```

```
##        frecuencia relativo
## Hombre        453    0.634
## Mujer         261    0.366
```

```
##        frecuencia relativo
## Hombre        453    0.634
## Mujer         261    0.366
ggplot(chr,aes(Sexo)) +geom_bar( fill="#0099ff" )+
  labs(title="Diagrama de barra", y="Frecuencia", x="Sexo") + theme_classic(base_size=15)
```




```
ggplot(df2,aes(x="",y=relativo, fill=rownames(df2)))+geom_bar(stat = "identity",color="white")+
  theme_void(base_size=15)+coord_polar(theta="y")+labs(title="Diagrama de barra Sexo")+
  scale_fill_manual(values = c("#0099ff","#66ff99"))+
  geom_text(aes(label=paste(relativo*100,"%")),position=position_stack(vjust=0.5),color="white",size=6)
```

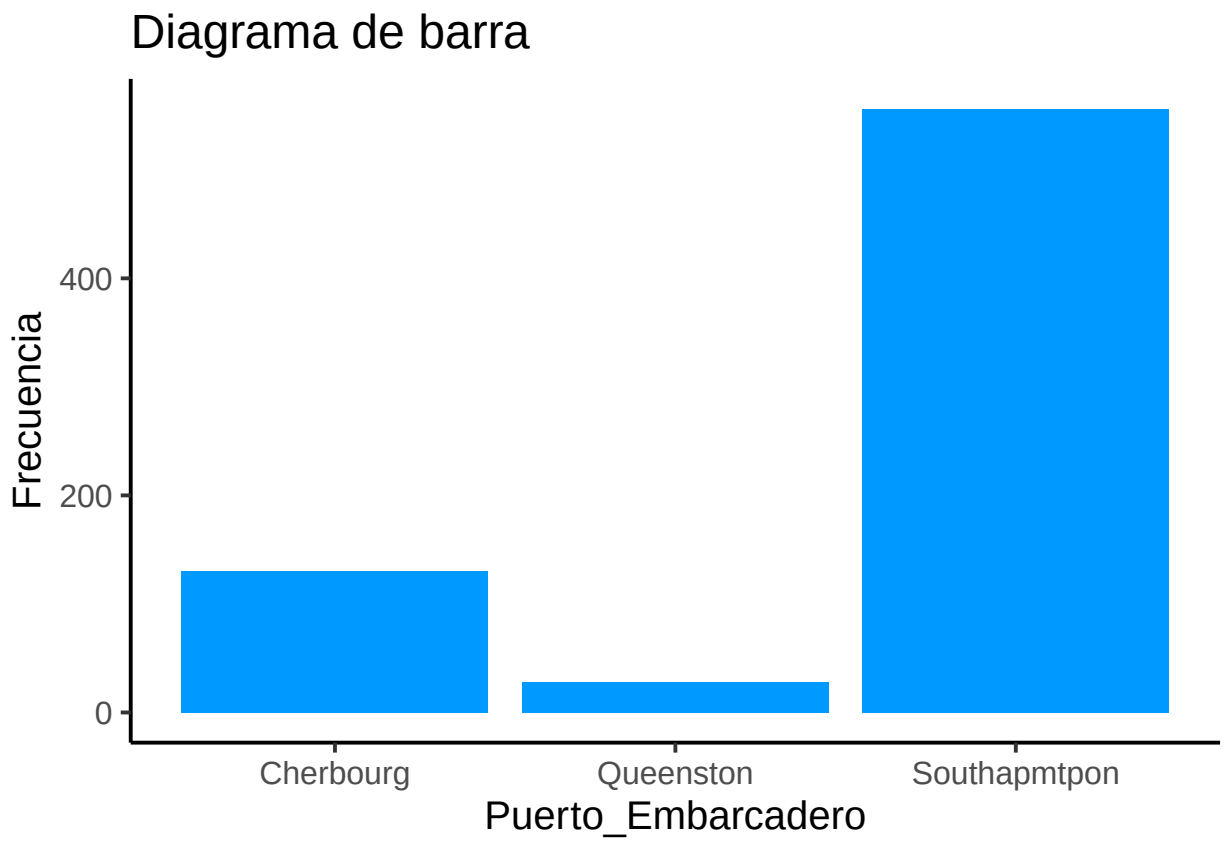


La variable sexo los 714 personas corresponde el 63,4% son hombre y el 36.6% son mujeres

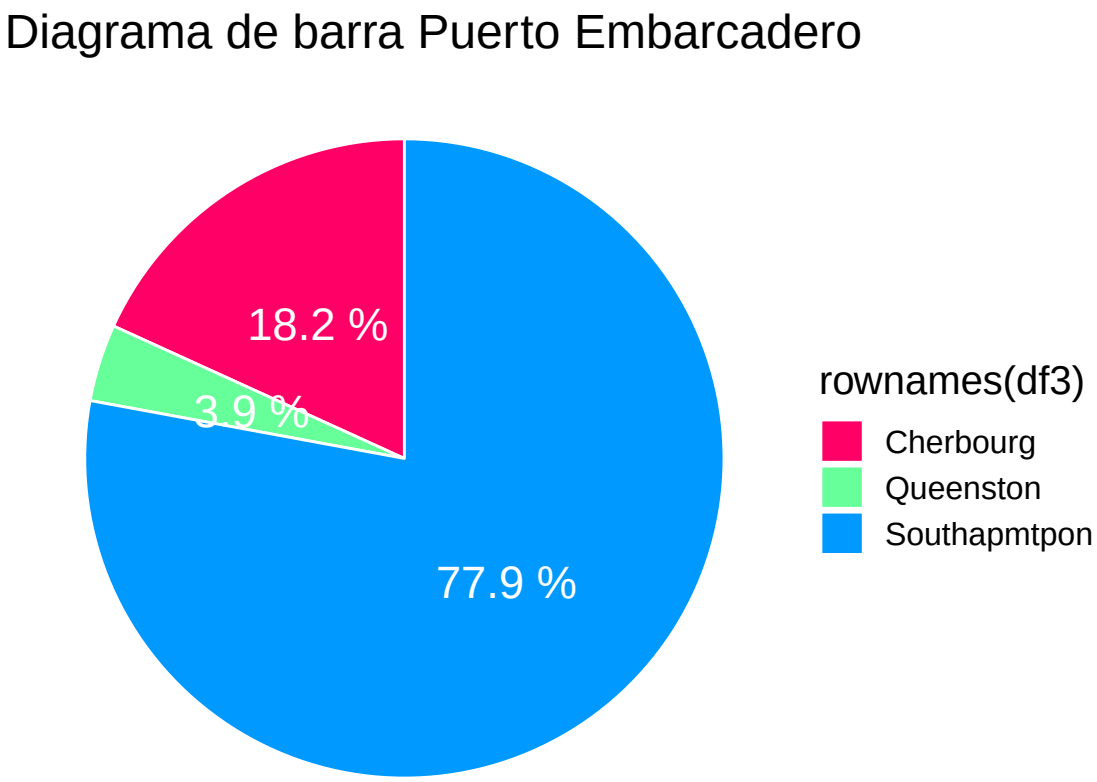
```
  #puerto embarquero
df3<-data.frame(resumen1(chr$Puerto_Embarcadero))
df3
```

```
##              frecuencia relativo
## Cherbourg           130    0.182
## Queenston            28    0.039
## Southapmtpon        556    0.779
```

```
##              frecuencia relativo
## Cherbourg           130    0.182
## Queenston            28    0.039
## Southapmtpon        556    0.779
ggplot(chr,aes(Puerto_Embarcadero)) +geom_bar( fill="#0099ff" )+
  labs(title="Diagrama de barra", y="Frecuencia", x="Puerto_Embarcadero") + theme_classic(base_size=15)
```




```
ggplot(df3,aes(x="",y=relativo, fill=rownames(df3)))+geom_bar(stat = "identity",color="white")+
theme_void(base_size=15)+coord_polar(theta="y")+labs(title="Diagrama de barra Puerto Embarcadero")+
scale_fill_manual(values = c("#ff0066","#66ff99","#0099ff"))+
geom_text(aes(label=paste(relativo*100,"%")),position=position_stack(vjust=0.5),color="white",size=6)
```



De los 714 pasajeros, se realize la parada en los puerto embarcadero Southapmtpon en el 77.9%, 18.2% Cherbourg y por ultimo con 3.9% Queenston.

```
  #----------- estadistica descriptiva cuantitativa -----------
resumen2<-function(y){
  q<-quantile(y, prob=c(0.25,0.5,0.75))
  nombre<-c("min","cuart 1","media","cuart 2","cuart 3","max","sd","asimetria","kurtorsi")
  valor<-round(c(min(y),q[1],mean(y),q[2],q[3],max(y),sd(y),skewness(y),kurtosis(y)),3)
  data.frame(nombre,valor)}

tabla<-function(x){
  r<-range(x)
  amp<-(r[2]-r[1])/nclass.Sturges(x)
  tab<-table.freq(hist(x, breaks=seq(r[1],r[2],amp) ,include.lowest=TRUE, right=FALSE, plot=F))
  tab
}


#clase de pasajero
resumen2(num$Clase_Pasajero)
```

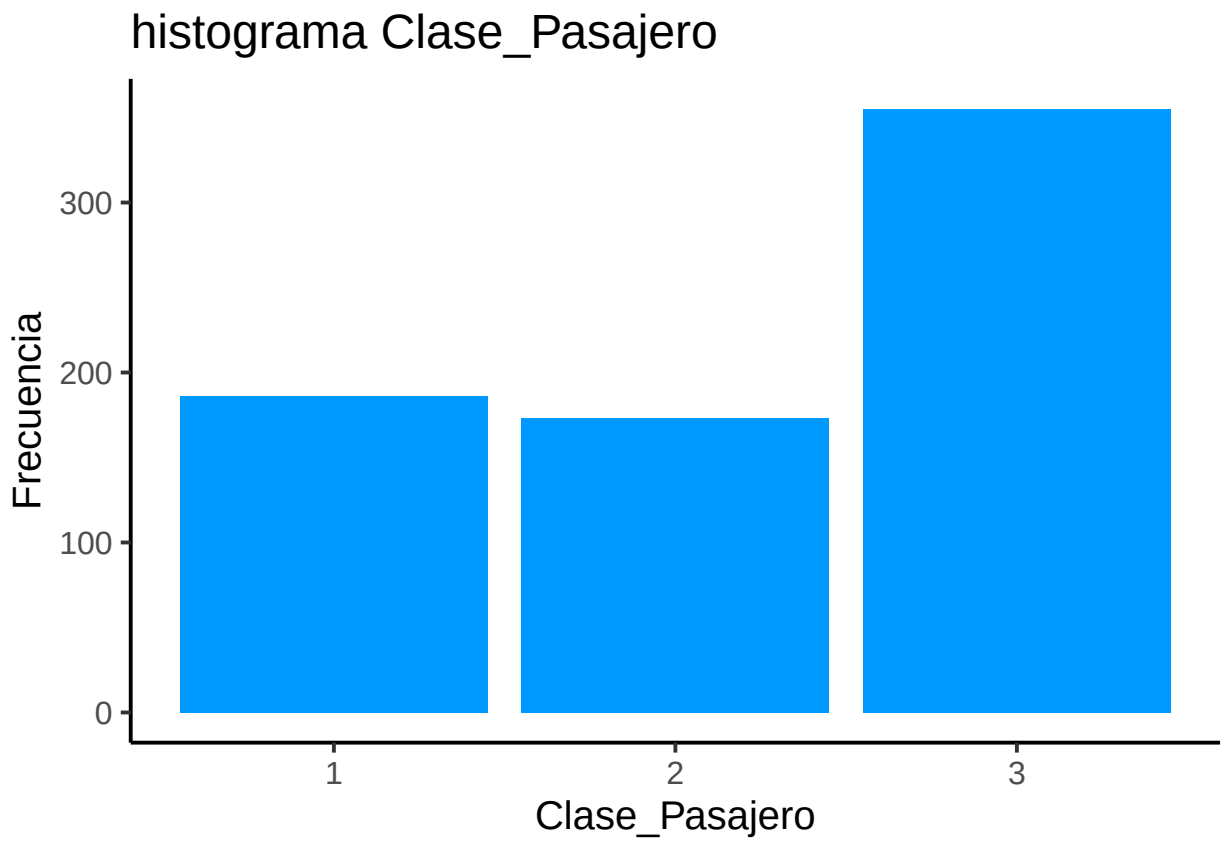
```
##      nombre  valor
## 1       min  1.000
## 2   cuart 1  1.000
## 3     media  2.237
## 4   cuart 2  2.000
```

```
## 5   cuart 3  3.000
## 6       max  3.000
## 7        sd  0.838
## 8 asimetria -0.469
## 9  kurtorsi -1.420
```

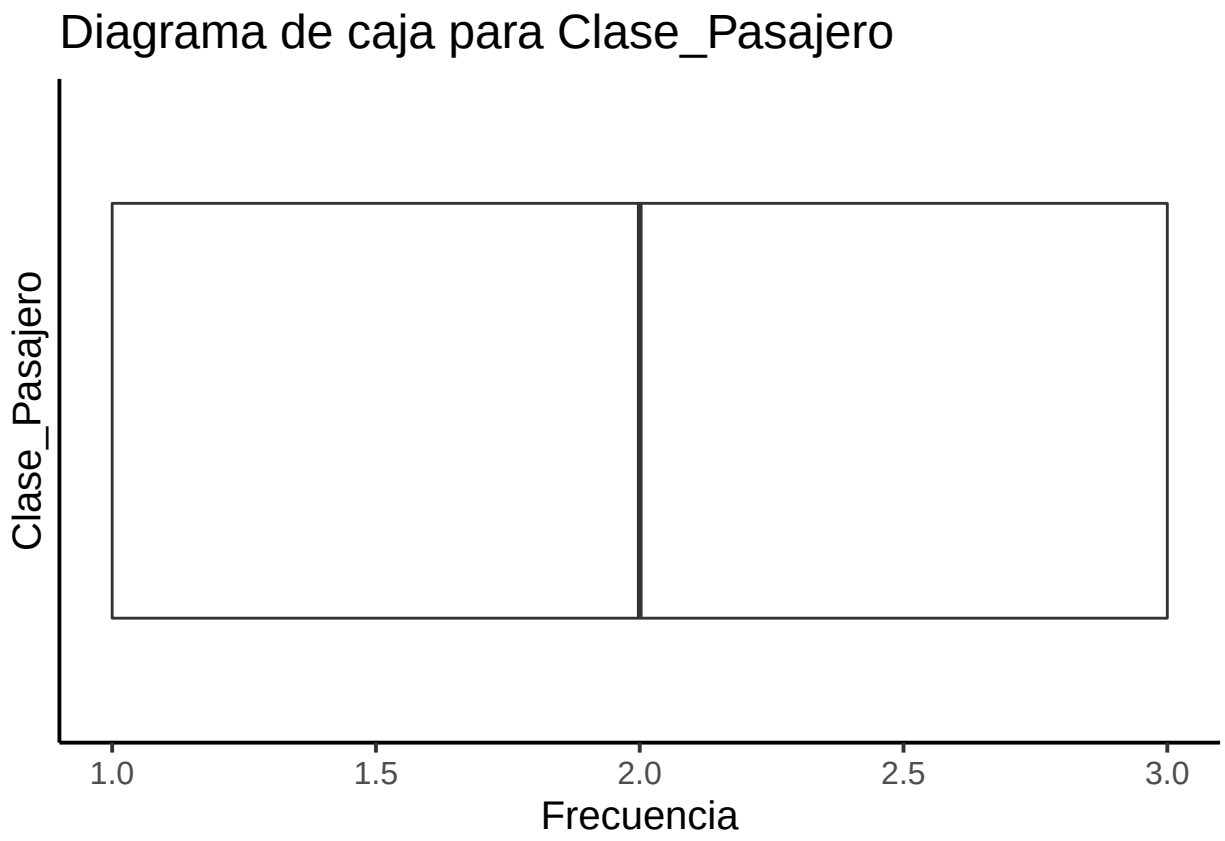
```
##      nombre  valor
## 1       min  1.000
## 2   cuart 1  1.000
## 3     media  2.237
## 4   cuart 2  2.000
## 5   cuart 3  3.000
## 6       max  3.000
## 7        sd  0.838
## 8 asimetria -0.469
## 9  kurtorsi -1.420
resumen1(num$Clase_Pasajero)
```

```
##   frecuencia relativo
## 1        186    0.261
## 2        173    0.242
## 3        355    0.497
```

```
##   frecuencia relativo
## 1        186    0.261
## 2        173    0.242
## 3        355    0.497
ggplot(num,aes(Clase_Pasajero)) +geom_bar( fill="#0099ff" )+
  labs(title="histograma Clase_Pasajero", y="Frecuencia", x="Clase_Pasajero") + theme_classic(base_size=
```




```
ggplot(num, aes(factor(0),y=Clase_Pasajero)) +
geom_boxplot()+ scale_x_discrete(breaks = NULL)+
labs(title="Diagrama de caja para Clase_Pasajero", y="Frecuencia", x="Clase_Pasajero",color=NULL)  +
coord_flip()+theme_classic(base_size=15)
```



La media de los boletos de clase para los pasajeros es un poco más de segunda clase 2.24 con una desviación de 0.838. Su mediana es de segunda clase 2. La tabla de frecuencias la concentración más alta la tuvo la clase 3 que tiene 355, con el 49.7%. Seguido por la primera clase con el 26.1% con 186 pasajeros. El histograma muestra una asimetría es -0.469 que refleja en la distribución un ligero sesgo hacia la derecha. El diagrama de caja no muestra valores atípicos, entonces no hay aberrancias en los datos.

```
 #edad
resumen2(num$Edad)
```

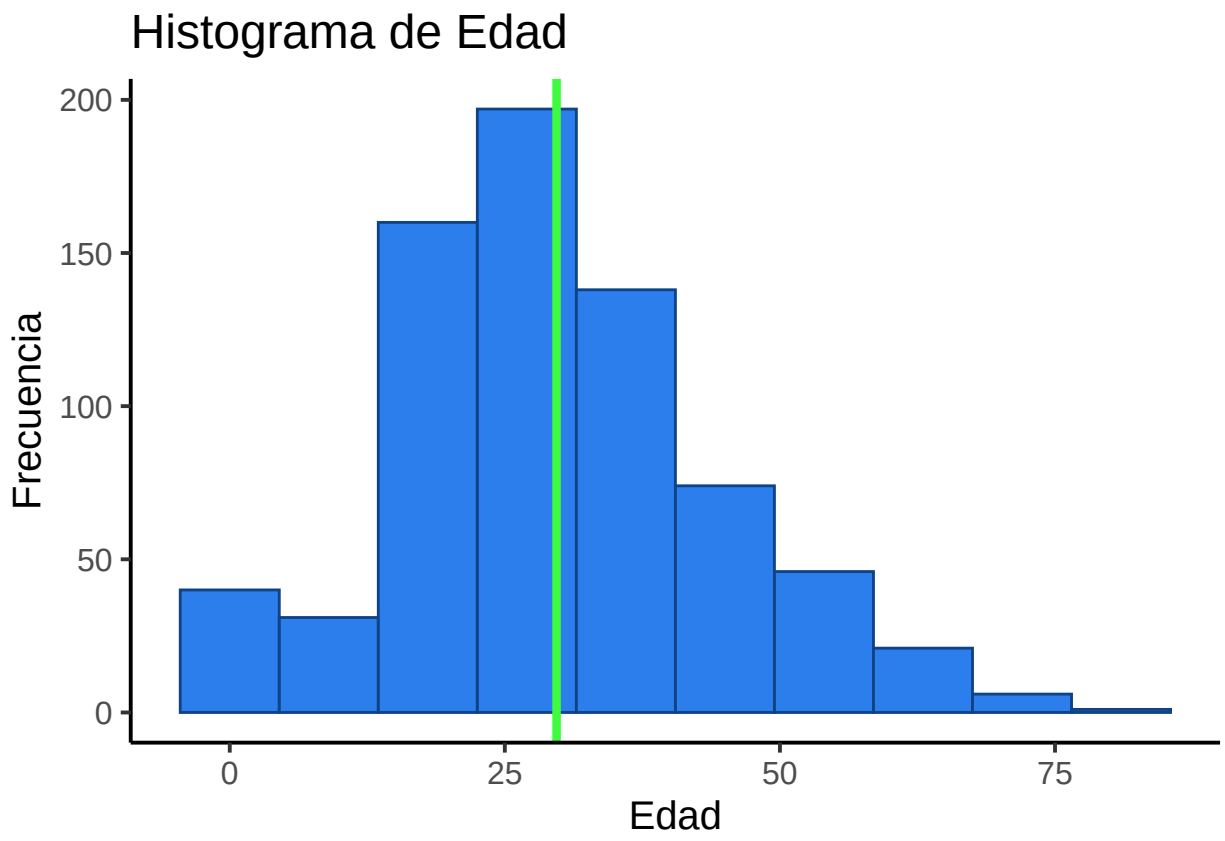
```
##      nombre  valor
## 1       min  0.420
## 2   cuart 1 20.125
## 3     media 29.699
## 4   cuart 2 28.000
## 5   cuart 3 38.000
## 6       max 80.000
## 7        sd 14.526
## 8 asimetria  0.389
## 9  kurtorsi  0.178
```

```
##      nombre  valor
## 1       min  0.420
## 2   cuart 1 20.125
## 3     media 29.699
## 4   cuart 2 28.000
## 5   cuart 3 38.000
```

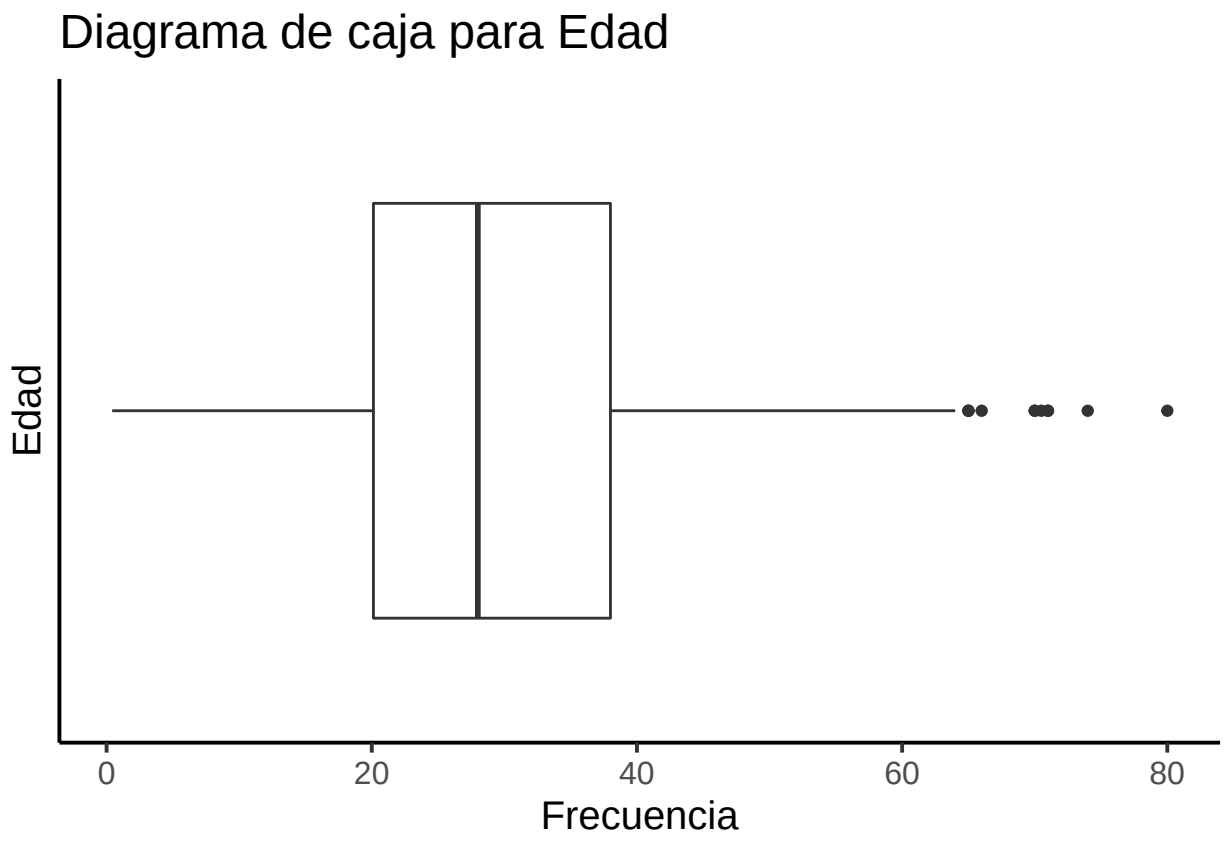
```
## 6       max 80.000
## 7        sd 14.526
## 8 asimetria  0.389
## 9  kurtorsi  0.178
tabla(num$Edad)
```

```
##        Lower     Upper      Main Frequency Percentage  CF   CPF
## 1   0.420000  7.654545  4.037273        50        7.0  50   7.0
## 2   7.654545 14.889091 11.271818        28        3.9  78  10.9
## 3  14.889091 22.123636 18.506364       153       21.4 231  32.4
## 4  22.123636 29.358182 25.740909       153       21.4 384  53.8
## 5  29.358182 36.592727 32.975455       136       19.0 520  72.8
## 6  36.592727 43.827273 40.210000        70        9.8 590  82.6
## 7  43.827273 51.061818 47.444545        67        9.4 657  92.0
## 8  51.061818 58.296364 54.679091        29        4.1 686  96.1
## 9  58.296364 65.530909 61.913636        20        2.8 706  98.9
## 10 65.530909 72.765455 69.148182         6        0.8 712  99.7
## 11 72.765455 80.000000 76.382727         2        0.3 714 100.0
```

```
##        Lower     Upper      Main Frequency Percentage  CF   CPF
## 1   0.420000  7.654545  4.037273        50        7.0  50   7.0
## 2   7.654545 14.889091 11.271818        28        3.9  78  10.9
## 3  14.889091 22.123636 18.506364       153       21.4 231  32.4
## 4  22.123636 29.358182 25.740909       153       21.4 384  53.8
## 5  29.358182 36.592727 32.975455       136       19.0 520  72.8
## 6  36.592727 43.827273 40.210000        70        9.8 590  82.6
## 7  43.827273 51.061818 47.444545        67        9.4 657  92.0
## 8  51.061818 58.296364 54.679091        29        4.1 686  96.1
## 9  58.296364 65.530909 61.913636        20        2.8 706  98.9
## 10 65.530909 72.765455 69.148182         6        0.8 712  99.7
## 11 72.765455 80.000000 76.382727         2        0.3 714 100.0
ggplot(num,aes(Edad))+geom_histogram( binwidth=9, fill="#1571EA", color="#104385", alpha=0.9) +
  labs(title="Histograma de Edad", y="Frecuencia", x="Edad",color=NULL) +
  geom_vline(xintercept = mean(num$Edad), color="#3EFB3F",size=1.5)+
  theme_classic(base_size=15)
```




```
ggplot(num, aes(factor(0),y=Edad)) +
geom_boxplot()+ scale_x_discrete(breaks = NULL)+
labs(title="Diagrama de caja para Edad", y="Frecuencia", x="Edad",color=NULL)  +
coord_flip()+theme_classic(base_size=15)
```



La edad media de los pasajeros a bordo era de 29.7 años y una desviación de 14.5 años. Su mediana es de 28 años. La tabla de frecuencias muestra la concentración más alta entre 14 a 22 años, y 22 a 29 años, ambos con la misma proporción de 21,4%. Después con el 19,0% entre los 29-36 años. El histograma muestra la distribución con un sesgo hacia la izquierda, donde el valor de la asimetría es 0.389 que nos indica la ligera concentración de lado izquierdo de los datos. El diagrama de caja muestra valores atípicos debido al sesgo de la izquierda de la distribución, donde las aberrancias aparecen pasado los 60 años.

```
#pariente masculino
resumen2(num$Pariente_Masculino)
```

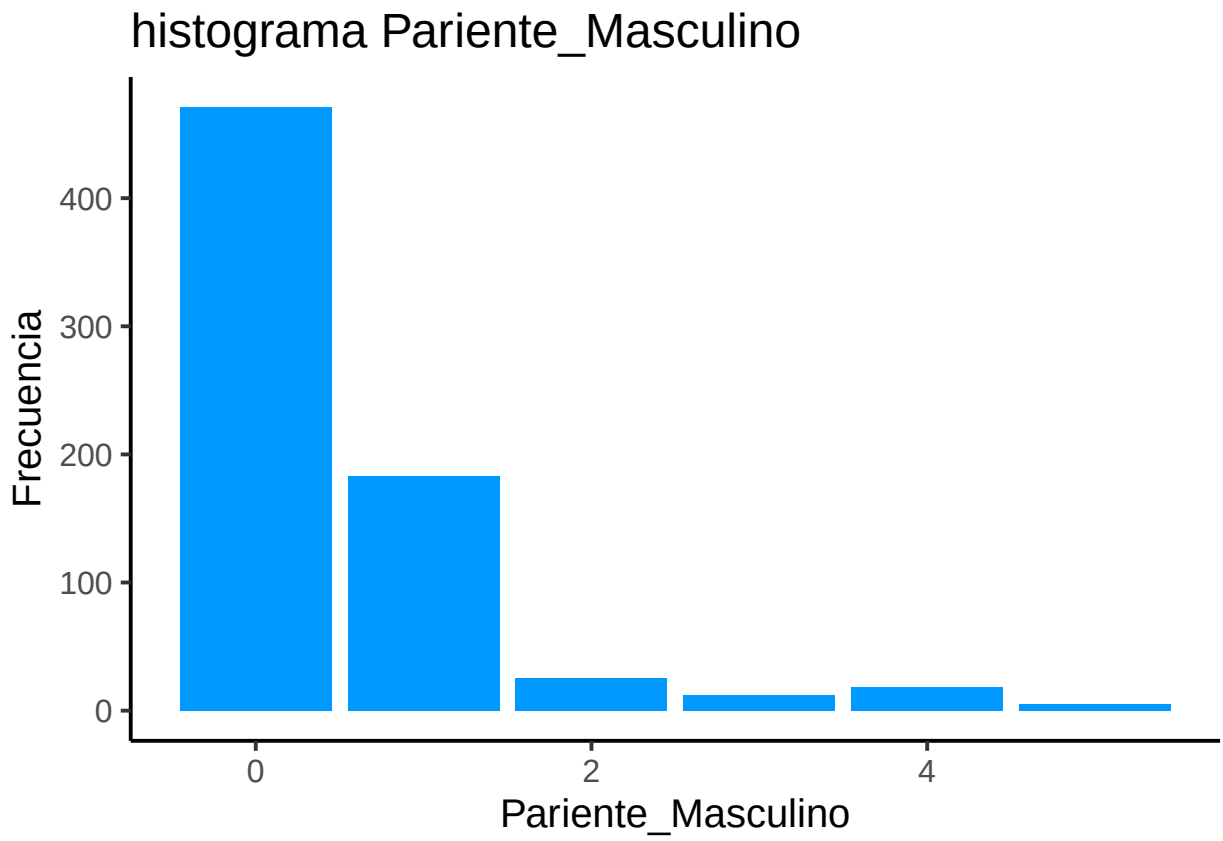
```
##      nombre valor
## 1       min 0.000
## 2   cuart 1 0.000
## 3     media 0.513
## 4   cuart 2 0.000
## 5   cuart 3 1.000
## 6       max 5.000
## 7        sd 0.930
## 8 asimetria 2.520
## 9  kurtorsi 7.045
```

```
##      nombre valor
## 1       min 0.000
## 2   cuart 1 0.000
## 3     media 0.513
## 4   cuart 2 0.000
```

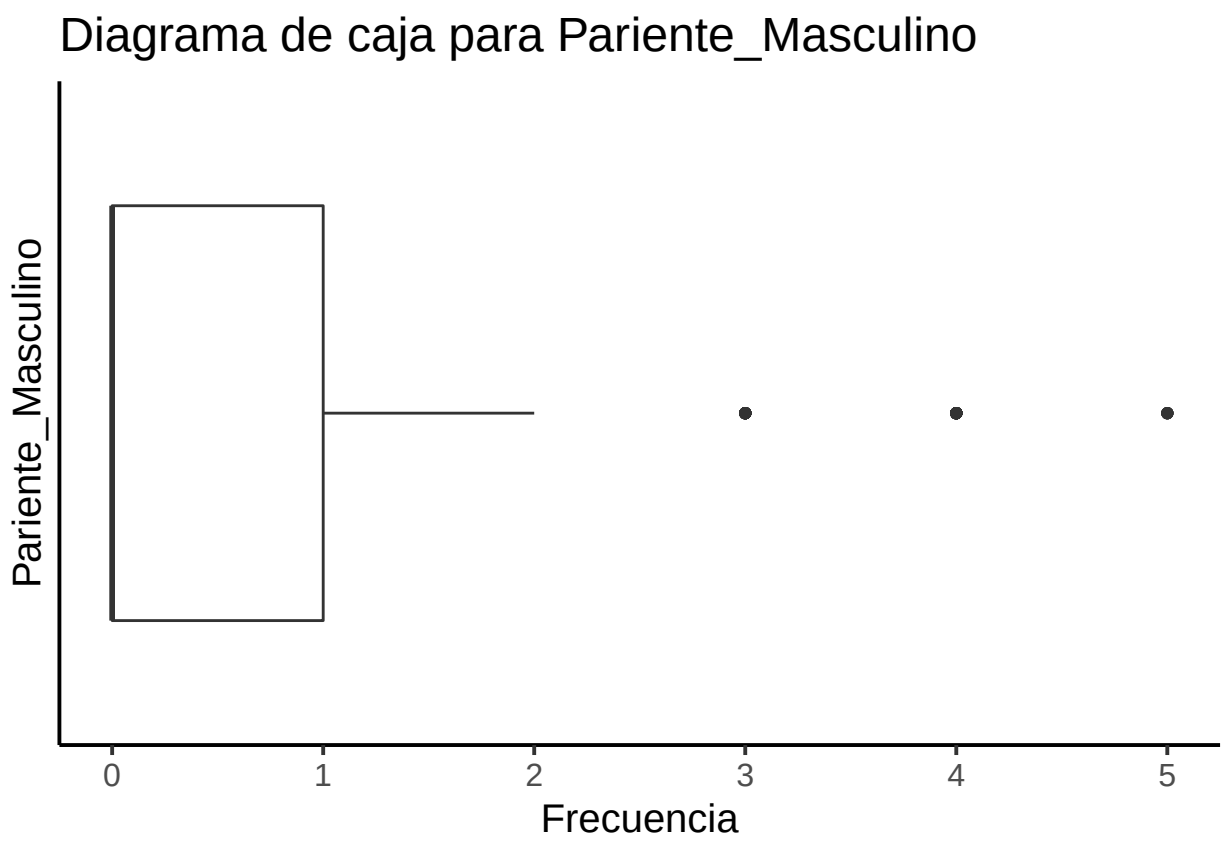
```
## 5   cuart 3 1.000
## 6       max 5.000
## 7        sd 0.930
## 8 asimetria 2.520
## 9  kurtorsi 7.045
resumen1(num$Pariente_Masculino)
```

```
##   frecuencia relativo
## 0        471    0.660
## 1        183    0.256
## 2         25    0.035
## 3         12    0.017
## 4         18    0.025
## 5          5    0.007
```

```
##   frecuencia relativo
## 0        471    0.660
## 1        183    0.256
## 2         25    0.035
## 3         12    0.017
## 4         18    0.025
## 5          5    0.007
ggplot(num,aes(Pariente_Masculino)) +geom_bar( fill="#0099ff" )+
  labs(title="histograma Pariente_Masculino", y="Frecuencia", x="Pariente_Masculino") + theme_classic(ba
```



```
  ggplot(num, aes(factor(0),y=Pariente_Masculino)) +
  geom_boxplot()+ scale_x_discrete(breaks = NULL)+
  labs(title="Diagrama de caja para Pariente_Masculino", y="Frecuencia", x="Pariente_Masculino",color=NU
  coord_flip()+theme_classic(base_size=15)
```



La media del número de parientes masculinos que eran pasajeros es 0.51 con una desviación de 0.93. Su mediana es de segunda clase 0. La tabla de frecuencias la concentración más alta la tuvo con cero parientes masculino que era 471, con el 66.0%. Seguido de un pariente masculino con el 25.6% de 183 pasajeros. El histograma muestra una asimetría es 2.52 que refleja en la distribución tiene un gran sesgo hacia la izquierda. El diagrama de caja muestra valores atípicos, pasado los 2 familiares masculino

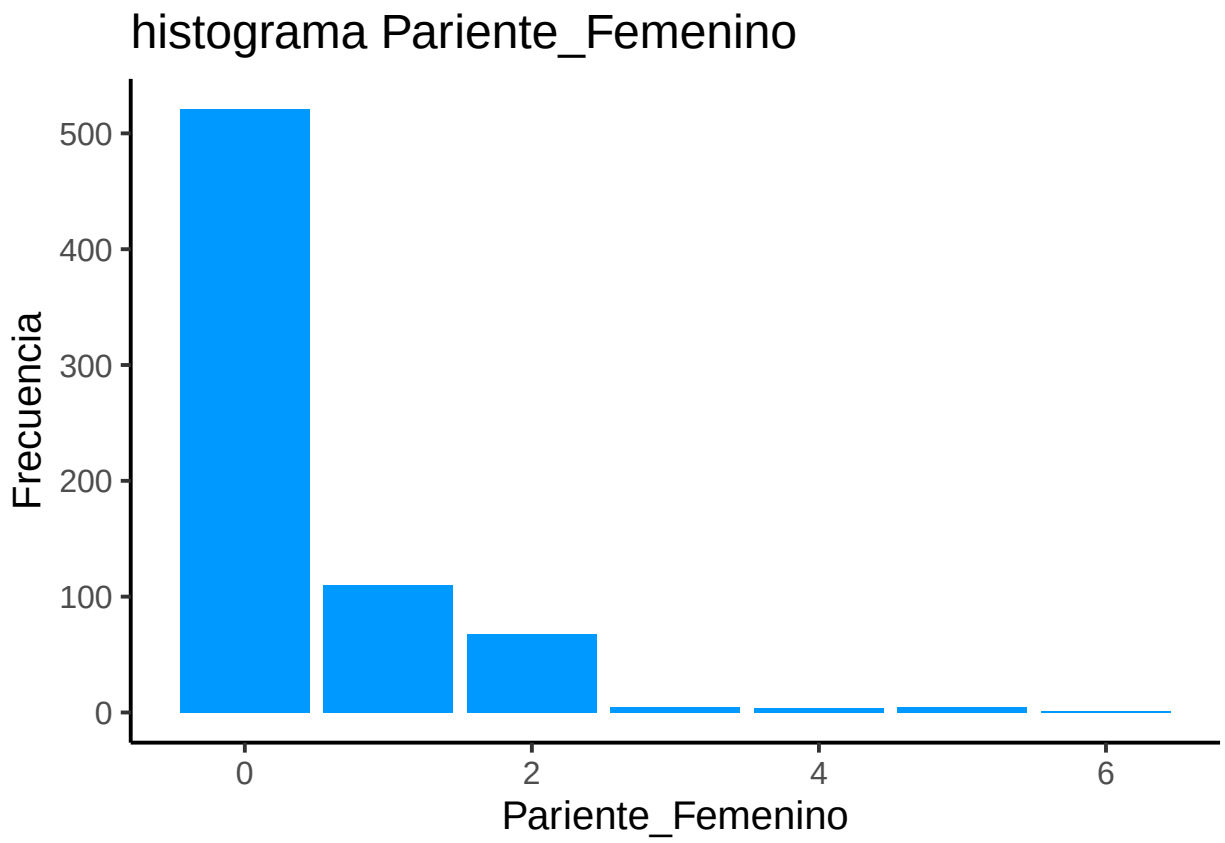
```
  #pariente femenino
resumen2(num$Pariente_Femenino)
```

```
##      nombre valor
## 1       min 0.000
## 2   cuart 1 0.000
## 3     media 0.431
## 4   cuart 2 0.000
## 5   cuart 3 1.000
## 6       max 6.000
## 7        sd 0.853
## 8 asimetria 2.619
## 9  kurtorsi 8.853
```

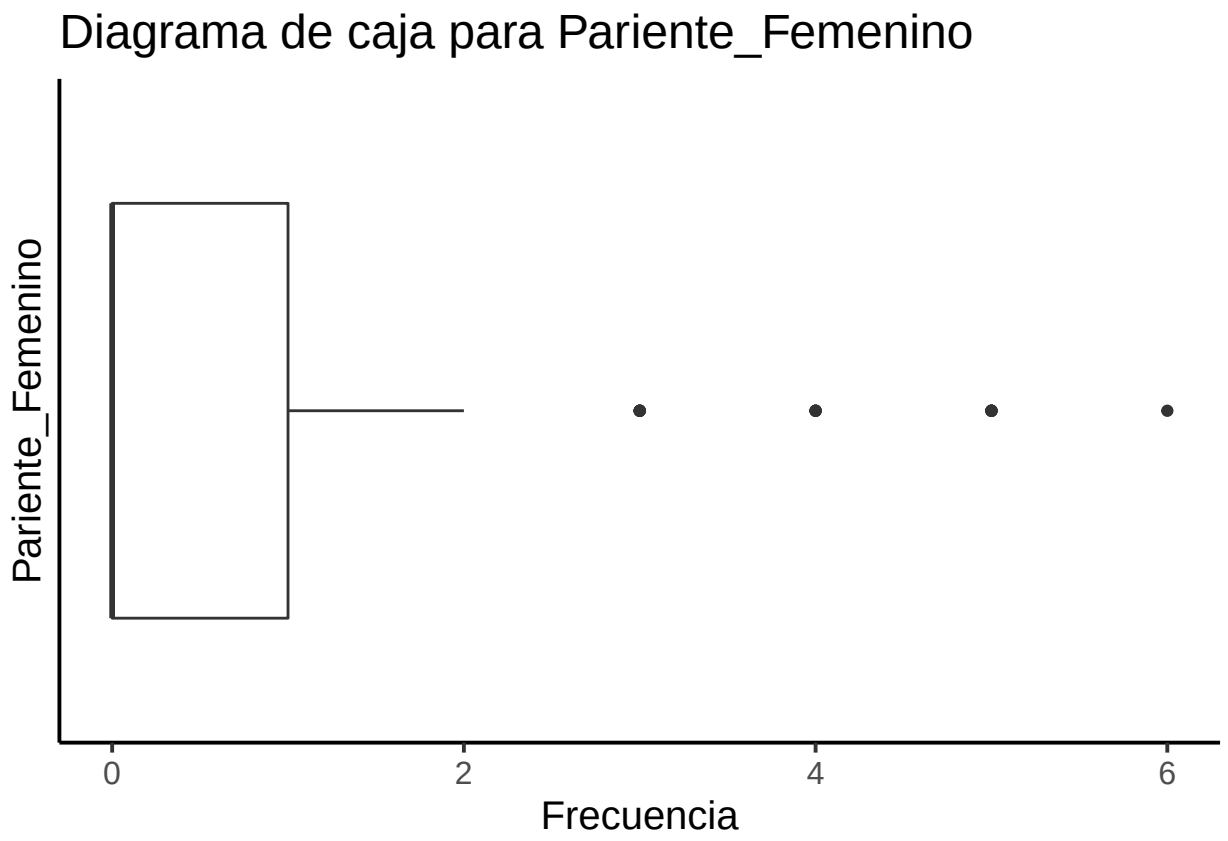
```
##      nombre valor
## 1       min 0.000
## 2   cuart 1 0.000
## 3     media 0.431
## 4   cuart 2 0.000
## 5   cuart 3 1.000
## 6       max 6.000
## 7        sd 0.853
## 8 asimetria 2.619
## 9  kurtorsi 8.853
resumen1(num$Pariente_Femenino)
```

```
##   frecuencia relativo
## 0        521    0.730
## 1        110    0.154
## 2         68    0.095
## 3          5    0.007
## 4          4    0.006
## 5          5    0.007
## 6          1    0.001
```

```
##   frecuencia relativo
## 0        521    0.730
## 1        110    0.154
## 2         68    0.095
## 3          5    0.007
## 4          4    0.006
## 5          5    0.007
## 6          1    0.001
ggplot(num,aes(Pariente_Femenino)) +geom_bar( fill="#0099ff" )+
  labs(title="histograma Pariente_Femenino", y="Frecuencia", x="Pariente_Femenino") + theme_classic(bas
```




```
ggplot(num, aes(factor(0),y=Pariente_Femenino)) +
geom_boxplot()+ scale_x_discrete(breaks = NULL)+
labs(title="Diagrama de caja para Pariente_Femenino", y="Frecuencia", x="Pariente_Femenino",color=NULL
coord_flip()+theme_classic(base_size=15)
```



desviación de 0.85. Su mediana es de segunda clase 0. La tabla de frecuencias la concentración más alta la tuvo con cero parientes femeninos que era 521, con el 73.0%. Seguido de un pariente femeninos con el 15.4% de 110 pasajeros. El histograma muestra una asimetría es 2.62 que refleja en la distribución tiene un gran sesgo hacia la izquierda. El diagrama de caja muestra valores atípicos, pasado los 2 familiares masculino

```
  #tarifa de pasaje
resumen2(num$Tarifa_Pasaje)
```

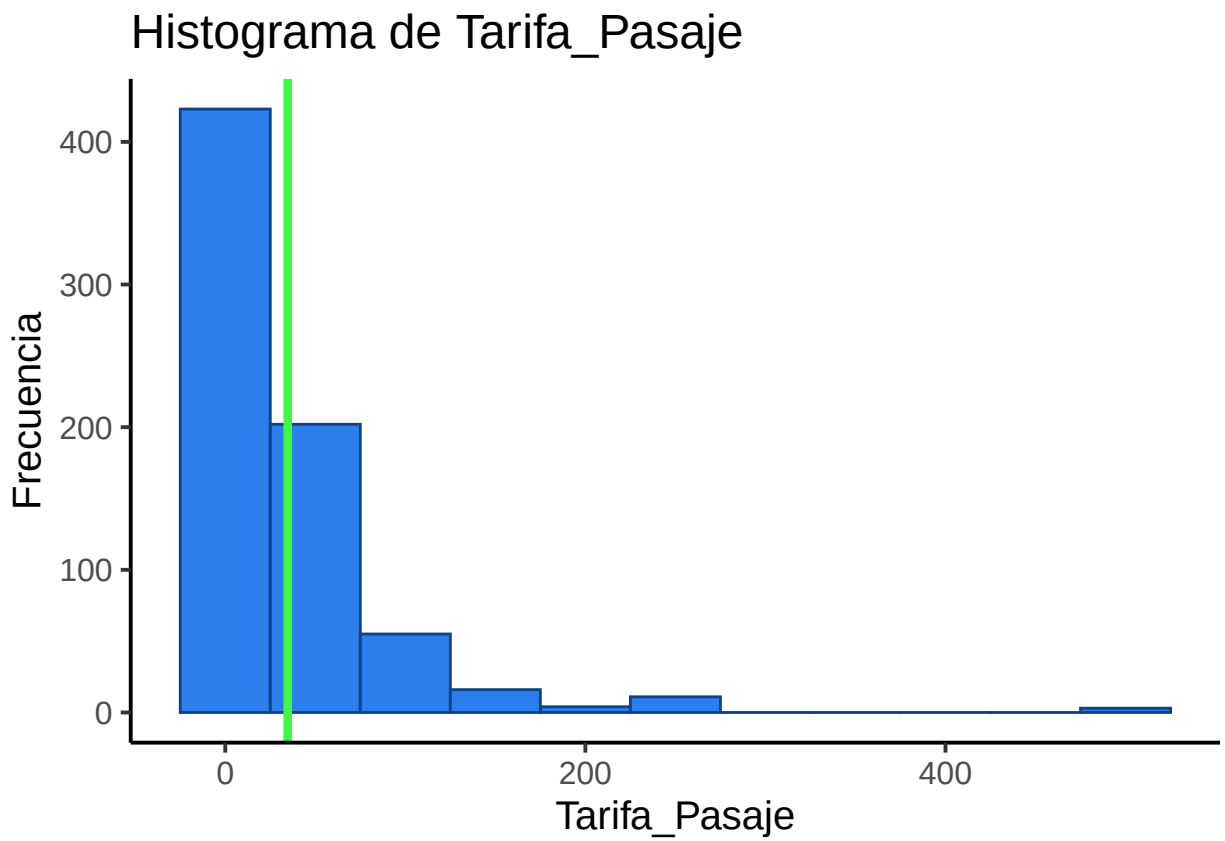
```
##      nombre   valor
## 1       min   0.000
## 2   cuart 1   8.050
## 3     media  34.695
## 4   cuart 2  15.742
## 5   cuart 3  33.375
## 6       max 512.329
## 7        sd  52.919
## 8 asimetria   4.654
## 9  kurtorsi  30.924
```

```
##      nombre   valor
## 1       min   0.000
## 2   cuart 1   8.050
## 3     media  34.695
## 4   cuart 2  15.742
## 5   cuart 3  33.375
## 6       max 512.329
```

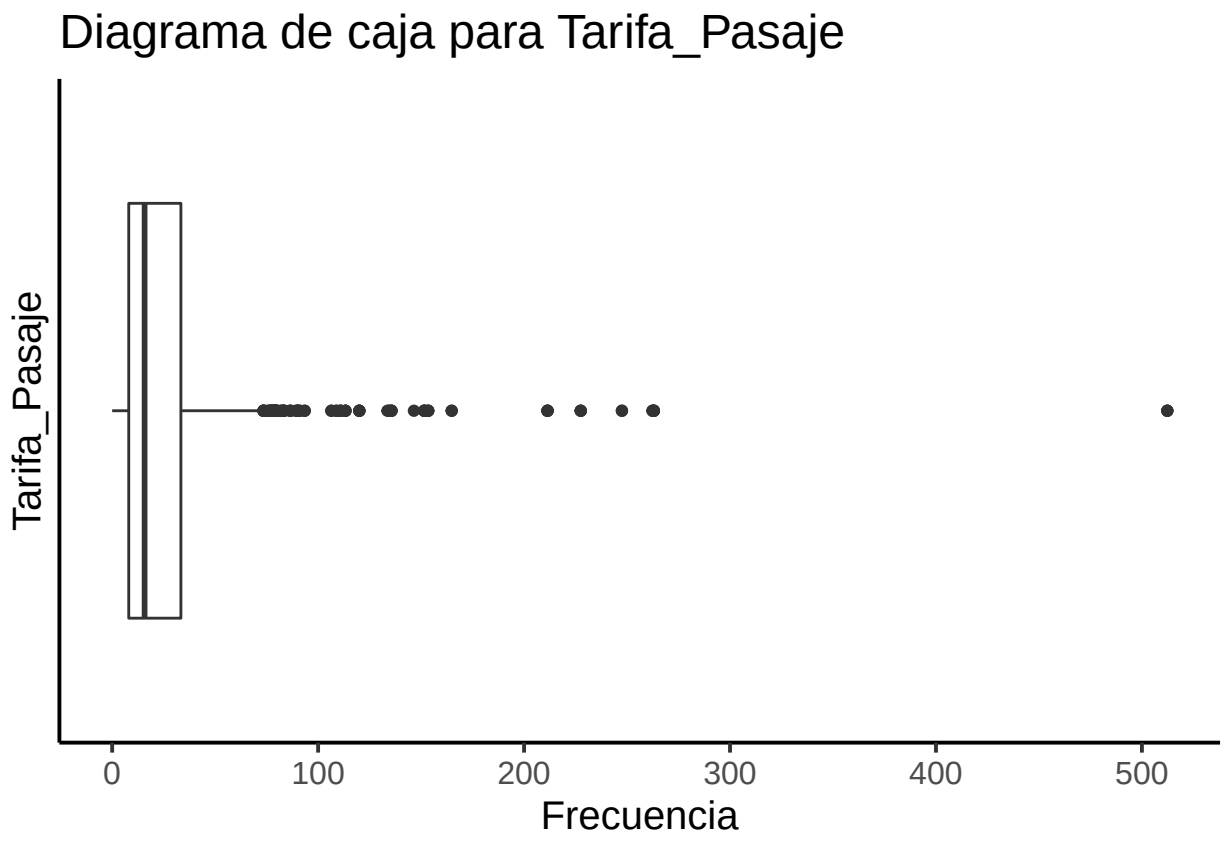
```
## 7        sd  52.919
## 8 asimetria   4.654
## 9  kurtorsi  30.924
tabla(num$Tarifa_Pasaje)
```

```
##        Lower     Upper      Main Frequency Percentage  CF   CPF
## 1    0.00000  46.57538  23.28769       566       79.3 566  79.3
## 2   46.57538  93.15076  69.86307        98       13.7 664  93.0
## 3   93.15076 139.72615 116.43845        22        3.1 686  96.1
## 4  139.72615 186.30153 163.01384        10        1.4 696  97.5
## 5  186.30153 232.87691 209.58922         7        1.0 703  98.5
## 6  232.87691 279.45229 256.16460         8        1.1 711  99.6
## 7  279.45229 326.02767 302.73998         0        0.0 711  99.6
## 8  326.02767 372.60305 349.31536         0        0.0 711  99.6
## 9  372.60305 419.17844 395.89075         0        0.0 711  99.6
## 10 419.17844 465.75382 442.46613         0        0.0 711  99.6
## 11 465.75382 512.32920 489.04151         3        0.4 714 100.0
```

```
##        Lower     Upper      Main Frequency Percentage  CF   CPF
## 1    0.00000  46.57538  23.28769       566       79.3 566  79.3
## 2   46.57538  93.15076  69.86307        98       13.7 664  93.0
## 3   93.15076 139.72615 116.43845        22        3.1 686  96.1
## 4  139.72615 186.30153 163.01384        10        1.4 696  97.5
## 5  186.30153 232.87691 209.58922         7        1.0 703  98.5
## 6  232.87691 279.45229 256.16460         8        1.1 711  99.6
## 7  279.45229 326.02767 302.73998         0        0.0 711  99.6
## 8  326.02767 372.60305 349.31536         0        0.0 711  99.6
## 9  372.60305 419.17844 395.89075         0        0.0 711  99.6
## 10 419.17844 465.75382 442.46613         0        0.0 711  99.6
## 11 465.75382 512.32920 489.04151         3        0.4 714 100.0
ggplot(num,aes(Tarifa_Pasaje))+geom_histogram( binwidth=50, fill="#1571EA", color="#104385", alpha=0.9)
  labs(title="Histograma de Tarifa_Pasaje", y="Frecuencia", x="Tarifa_Pasaje",color=NULL) +
  geom_vline(xintercept = mean(num$Tarifa_Pasaje), color="#3EFB3F",size=1.5)+
  theme_classic(base_size=15)
```




```
ggplot(num, aes(factor(0),y=Tarifa_Pasaje)) +
geom_boxplot()+ scale_x_discrete(breaks = NULL)+
labs(title="Diagrama de caja para Tarifa_Pasaje", y="Frecuencia", x="Tarifa_Pasaje")  +
coord_flip()+theme_classic(base_size=15)
```

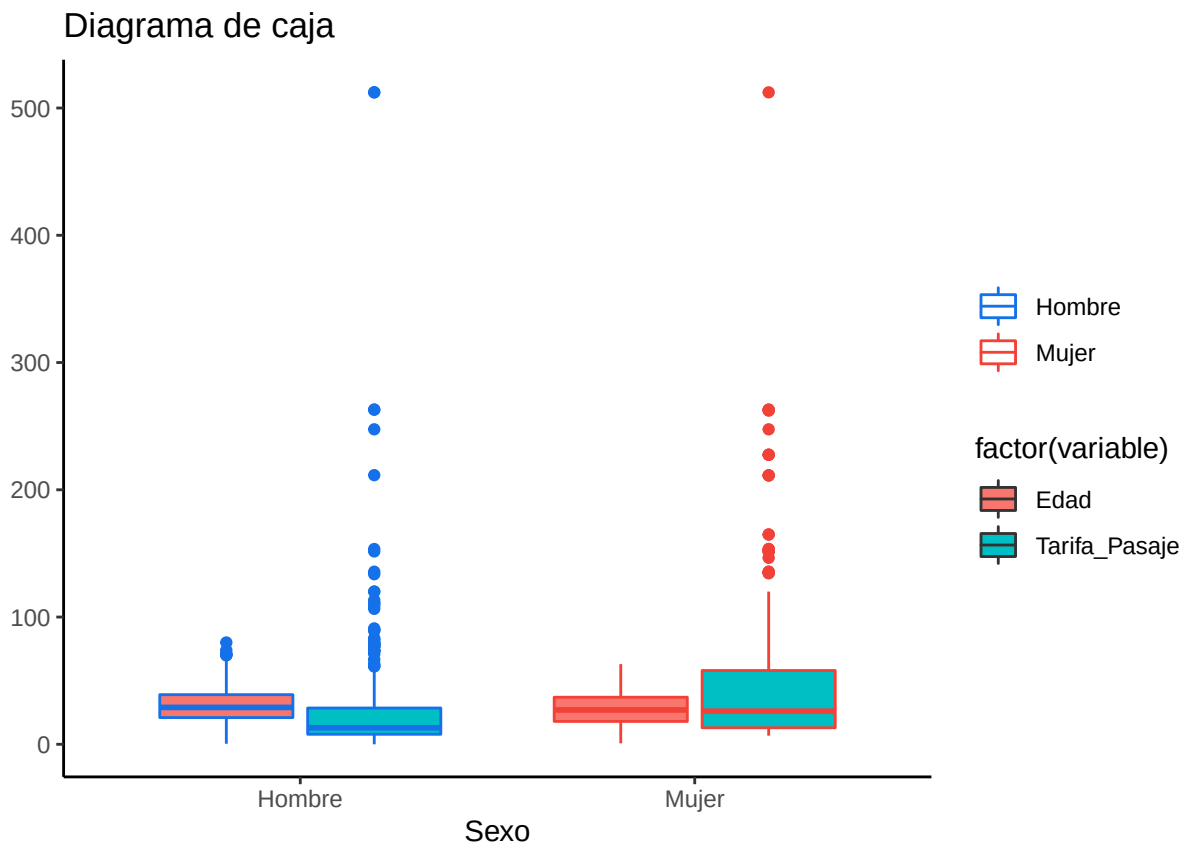


La tarifa media de los precios es de 34.7 con una desviación de 52.9. Su mediana es de 15.7. La tabla de frecuencias muestra la concentración más alta entre 0 a 46.6, y con un 79,3%. Después con el 13,7% entre los 47 a 93. El histograma muestra la distribución con un sesgo hacia la izquierda, donde el valor de la asimetría es 4.654 que nos indica una gran concentración de lado izquierdo de los datos. El diagrama de caja muestra valores atípicos debido al sesgo de la izquierda de la distribución, donde las aberrancias aparecen pasado los 80.

```
 #------------- estadistica multivariante----------
dt<-cbind(num[,c(2,5)],chr)
dt1<- melt(dt, measure.vars=1:2)
names(dt1)
```

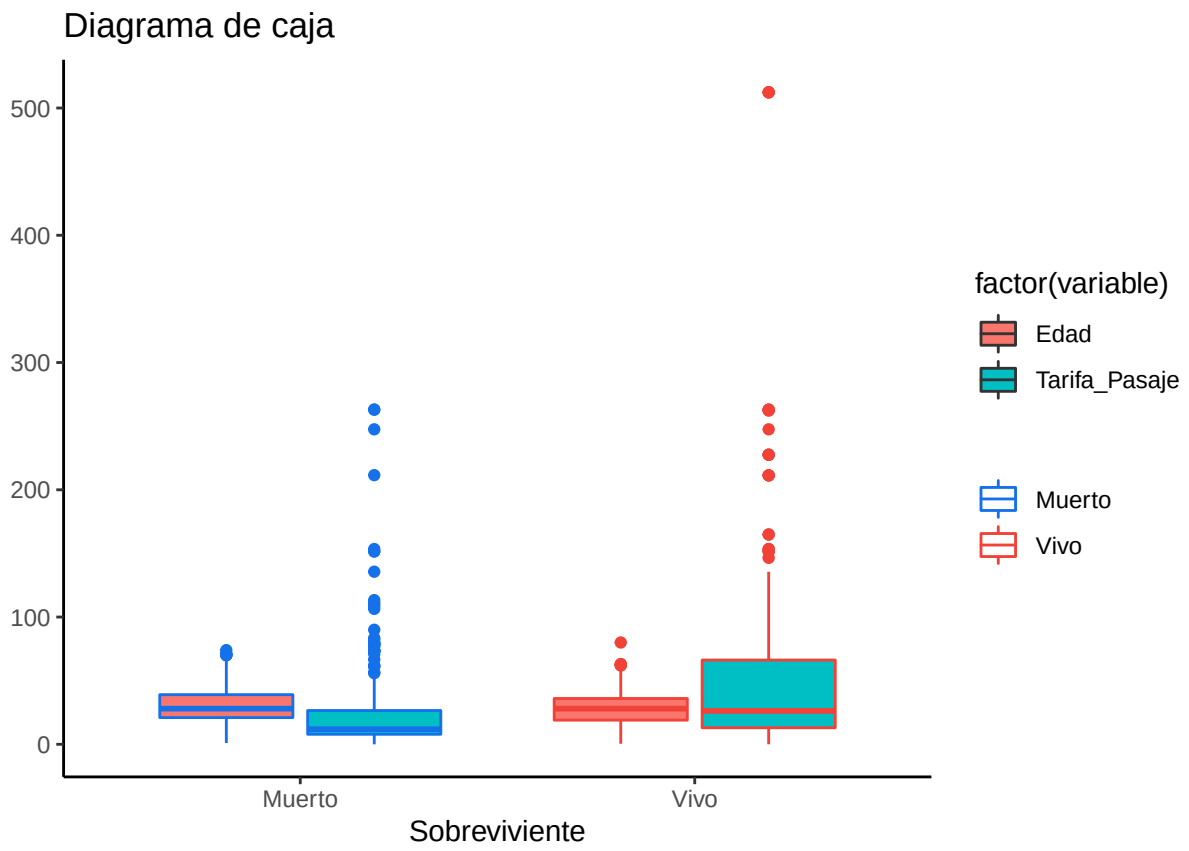
```
## [1] "Sobreviviente"      "Sexo"               "Puerto_Embarcadero"
## [4] "variable"           "value"
```

```
## [1] "Sobreviviente"      "Sexo"               "Puerto_Embarcadero"
## [4] "variable"           "value"
ggplot(dt1, aes(x=factor(Sexo), y=value,colour=Sexo)) +
  geom_boxplot(aes(fill=factor(variable)))+scale_color_manual(values=c("#1571EA", "#F24237"))+
  labs(title="Diagrama de caja", y="", x="Sexo",color=NULL) +
  theme_classic()
```



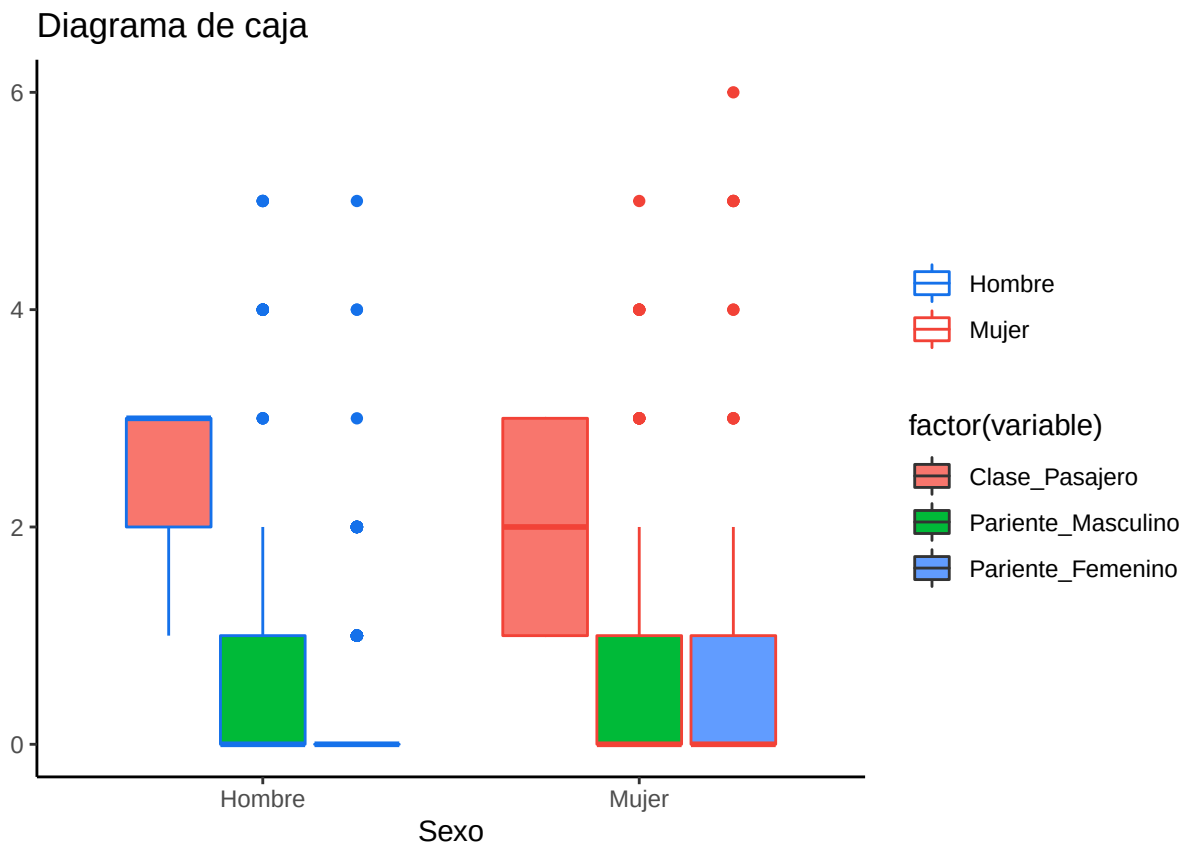
Se tiene las variables edad y tarifa, separada tanto por hombre y mujer. Donde observamos que en edad la mediana de la edad de los hombres es mayor que de mujeres, donde en los hombres se encuentra que hay valores atípicos y en mujeres no. Para las tarifas en hombres y mujeres se tiene que la mediana de los precios para mujeres es mayor al de los hombres. Donde los hombres tienen mayores aberrancias que las mujeres. También se puede observar que el diagrama de caja de las mujeres tienen un bigote superior mas grande debido que un gran número de ellas paga tarifas han comprado boletos mas caros.

```
ggplot(dt1, aes(x=factor(Sobreviviente), y=value,colour=Sobreviviente)) +
  geom_boxplot(aes(fill=factor(variable)))+scale_color_manual(values=c("#1571EA", "#F24237"))+
  labs(title="Diagrama de caja", y="", x="Sobreviviente",color=NULL) +
  theme_classic()
```



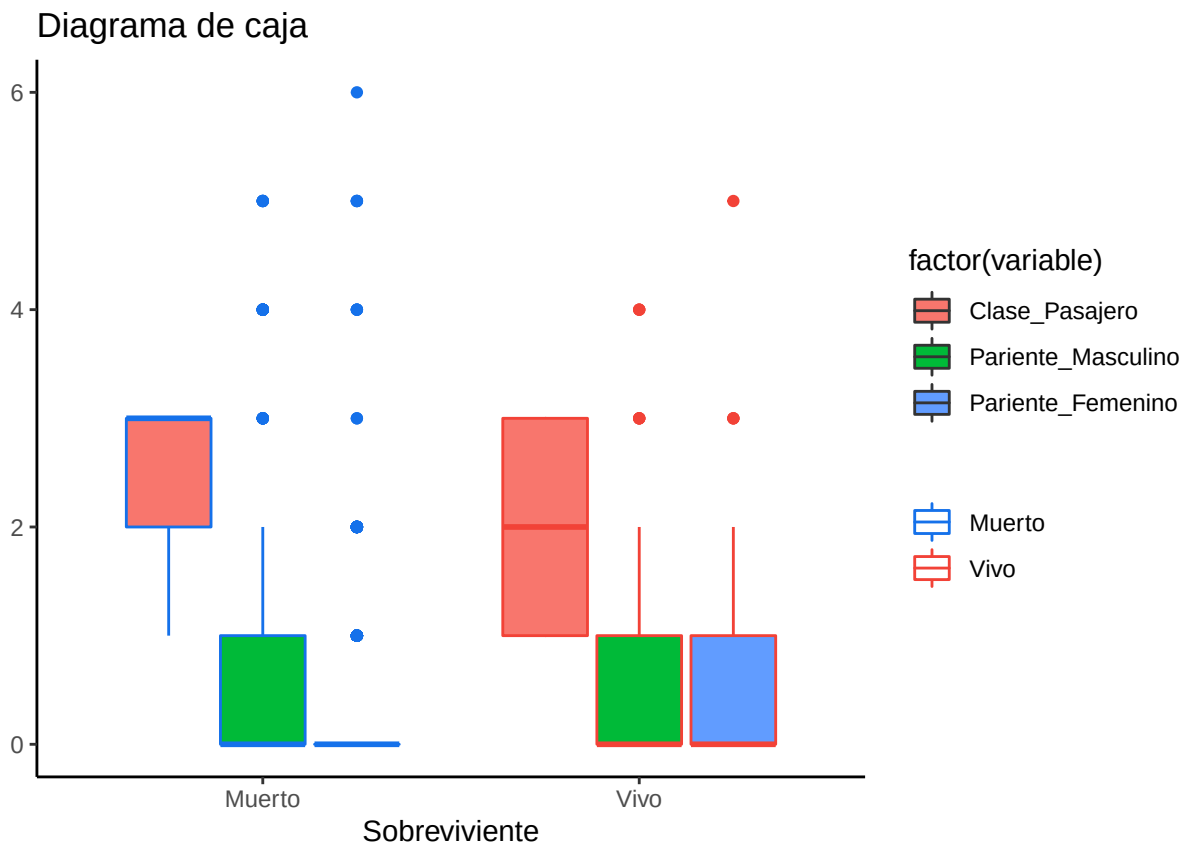
Se tiene las variables edad y tarifa, separada tanto por estado vivo o muerto. Donde observamos que la mediana de la edad de personas muertas es ligeramente mayor en hombres que en mujeres. Para las tarifa de los precios de las personas vivas o muertas, se puede ver que hay mas personas que han vivido comparado las personas muertas, donde en al grafica anterior se había observado que la mayores personas en los precios eran las mujeres con respecto al hombre.

```
  dt<-cbind(num[,c(1,3,4)],chr)
dt1<- melt(dt, measure.vars=1:3)

ggplot(dt1, aes(x=factor(Sexo), y=value,colour=Sexo)) +
  geom_boxplot(aes(fill=factor(variable)))+scale_color_manual(values=c("#1571EA", "#F24237"))+
  labs(title="Diagrama de caja", y="", x="Sexo",color=NULL) +
  theme_classic()
```



Separamos por hombre y mujer, los tipos de clases de pasajeros parientes masculinos y parientes femeninos. Para los tipos de clases de pasajeros se observa que la mediana de hombres son de la clase más baja que es tercera clase, mientras que en mujeres su mediana son de clase media que es segunda clase. Para los parientes masculinos que observa que tanto hombres y mujeres tiene una similitud con una mediana de cero en ambos casos. Para los parientes femeninos también se observa que su mediana en ambos casos es igual a cero, pero con la diferencia que en mujeres su cuartil 3 ha llevado un pariente femenino.

```
ggplot(dt1, aes(x=factor(Sobreviviente), y=value,colour=Sobreviviente)) +
geom_boxplot(aes(fill=factor(variable)))+scale_color_manual(values=c("#1571EA", "#F24237"))+
labs(title="Diagrama de caja", y="", x="Sobreviviente",color=NULL) +
theme_classic()
```



Separamos por vivo o muerto los tipos de clases de pasajeros parientes masculinos y parientes femeninos. Para los tipos de clases de pasajeros se observa que la mediana de personas muertas son de la clase más baja que es tercera clase, mientras que en estado vivo su mediana son de clase media que es segunda clase, donde se recuerda que había una concentración de mujeres vivas. Para los parientes masculinos que observa que tanto en vivo y muerto tiene una similitud con una mediana de cero en ambos casos. Para los parientes femeninos también se observa que su mediana en ambos casos es igual a cero, pero con la diferencia que en vivo su cuartil 3 ha llevado un pariente femenino.

```
  #------------ estadistica de correlacion ---------------------------
#matriz de correlacion y covarianza
covar<-round(var(num),2)
covar
```

```
##                   Clase_Pasajero   Edad Pariente_Masculino Pariente_Femenino
## Clase_Pasajero              0.70  -4.50               0.05              0.02
## Edad                       -4.50 211.02              -4.16             -2.34
## Pariente_Masculino          0.05  -4.16               0.86              0.30
## Pariente_Femenino           0.02  -2.34               0.30              0.73
## Tarifa_Pasaje             -24.58  73.85               6.81              9.26
##                   Tarifa_Pasaje
## Clase_Pasajero          -24.58
## Edad                     73.85
## Pariente_Masculino        6.81
## Pariente_Femenino         9.26
## Tarifa_Pasaje          2800.41
```

```
##                    Clase_Pasajero   Edad Pariente_Masculino Pariente_Femenino
## Clase_Pasajero               0.70  -4.50               0.05              0.02
## Edad                        -4.50 211.02              -4.16             -2.34
## Pariente_Masculino           0.05  -4.16               0.86              0.30
## Pariente_Femenino            0.02  -2.34               0.30              0.73
## Tarifa_Pasaje              -24.58  73.85               6.81              9.26
##                    Tarifa_Pasaje
## Clase_Pasajero           -24.58
## Edad                      73.85
## Pariente_Masculino         6.81
## Pariente_Femenino          9.26
## Tarifa_Pasaje           2800.41
corr<-round(cor(num),2)
corr
```

```
##                    Clase_Pasajero  Edad Pariente_Masculino Pariente_Femenino
## Clase_Pasajero               1.00 -0.37               0.07              0.03
## Edad                        -0.37  1.00              -0.31             -0.19
## Pariente_Masculino           0.07 -0.31               1.00              0.38
## Pariente_Femenino            0.03 -0.19               0.38              1.00
## Tarifa_Pasaje               -0.55  0.10               0.14              0.21
##                    Tarifa_Pasaje
## Clase_Pasajero             -0.55
## Edad                        0.10
## Pariente_Masculino          0.14
## Pariente_Femenino           0.21
## Tarifa_Pasaje               1.00
```

```
##                    Clase_Pasajero  Edad Pariente_Masculino Pariente_Femenino
## Clase_Pasajero               1.00 -0.37               0.07              0.03
## Edad                        -0.37  1.00              -0.31             -0.19
## Pariente_Masculino           0.07 -0.31               1.00              0.38
## Pariente_Femenino            0.03 -0.19               0.38              1.00
## Tarifa_Pasaje               -0.55  0.10               0.14              0.21
##                    Tarifa_Pasaje
## Clase_Pasajero             -0.55
## Edad                        0.10
## Pariente_Masculino          0.14
## Pariente_Femenino           0.21
## Tarifa_Pasaje               1.00
```

Observando los diferentes cruces entre las variables cuantitativas se observa que ninguna variable tiene correlaciones fuertes superior a 0.7. Donde la correlación más alta es el precio de la tarifa y el nivel de clase de pasajero con un -0.55, que representa una ligera correlación inversa fuerte, es decir, entre mayor sea la clase (primera clase) mayor es la tarifa y mientras menor sea la clase (tercera clase) menor será la tarifa. Las demás variables tienen una correlación muy débil el cual se podría decir que son independiente entre sí.

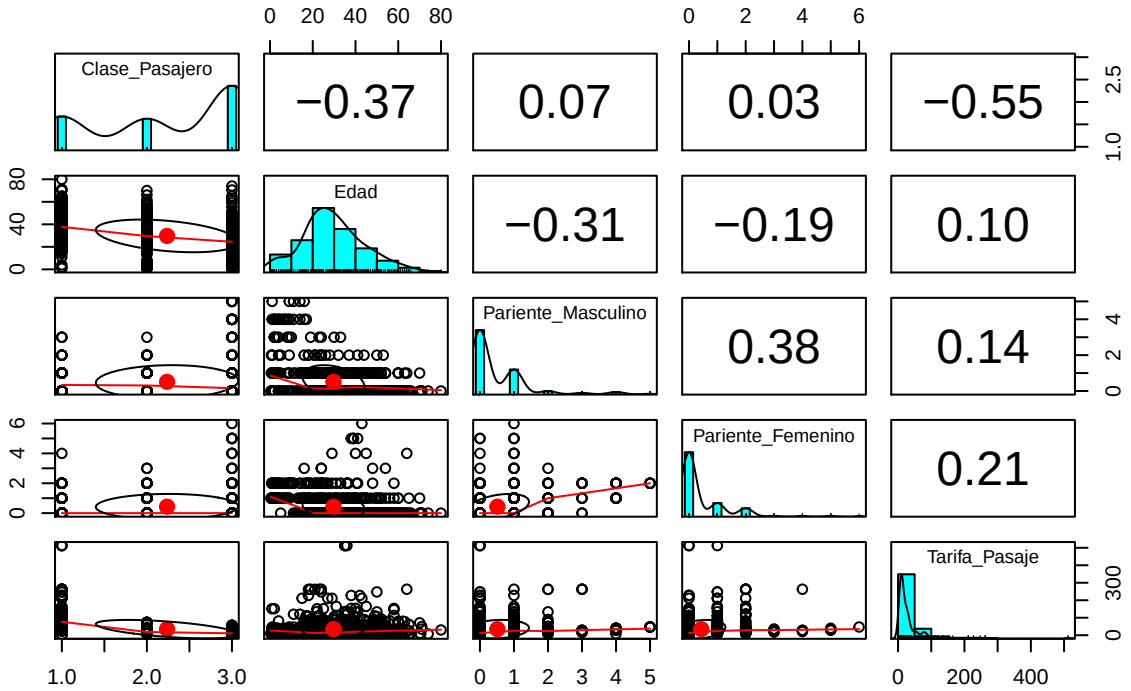
```
  library(psych)
```

```
##
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

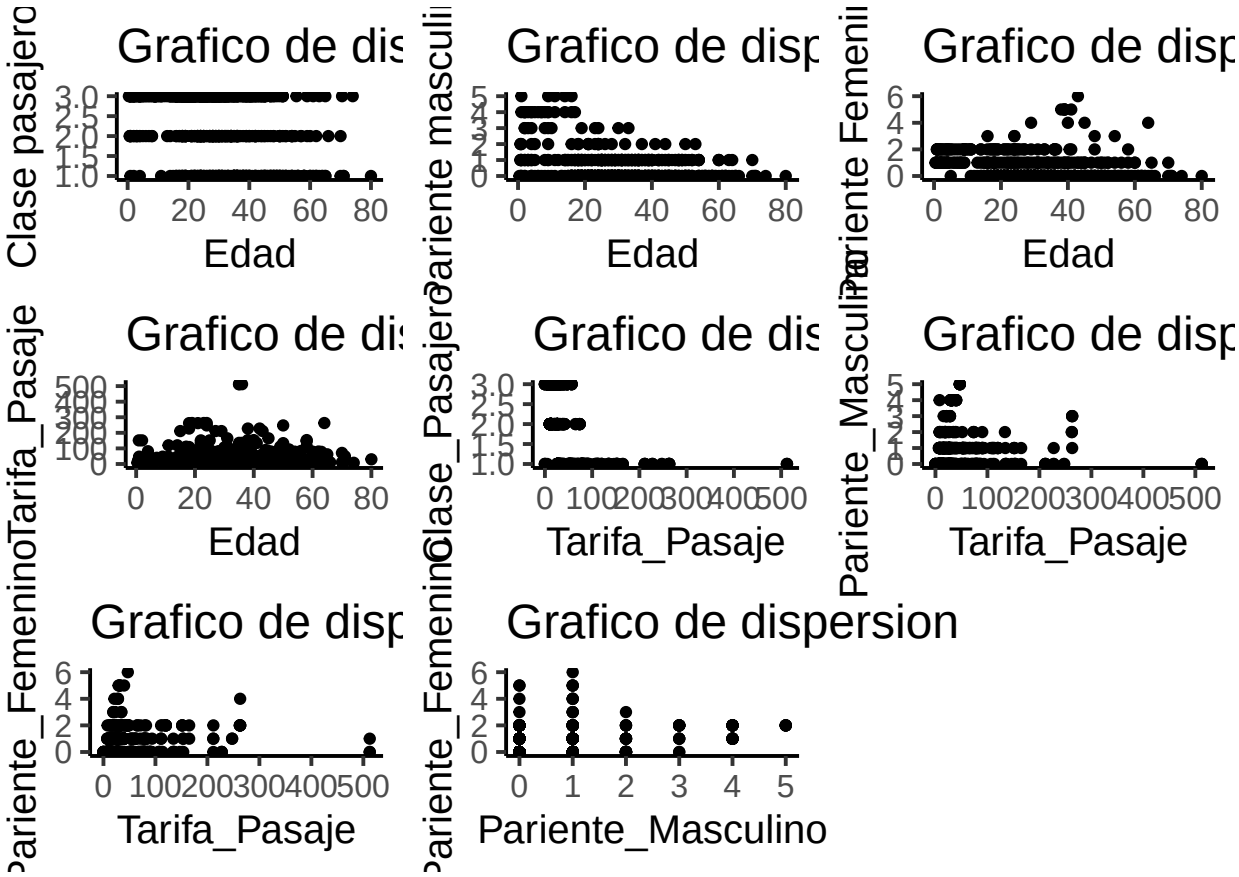
```
##
## Attaching package: 'psych'
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
pairs.panels(num, pch=21,main="Gráfico 01.6: Matriz de Dispersión, Histograma y Correlación")
```

Gráfico 01.6: Matriz de Dispersión, Histograma y Correlación




```
Plot1 <- ggplot(num, aes(Edad,Clase_Pasajero)) +geom_point()+
  labs(title="Grafico de dispersion", y="Clase pasajero", x="Edad",color=NULL) +
  theme_classic(base_size=15)
Plot2 <- ggplot(num, aes(Edad,Pariente_Masculino)) +geom_point()+
  labs(title="Grafico de dispersion", y="Pariente masculino", x="Edad",color=NULL) +
  theme_classic(base_size=15)
Plot3 <- ggplot(num, aes(Edad,Pariente_Femenino)) +geom_point()+
  labs(title="Grafico de dispersion", y="Pariente Femenino", x="Edad",color=NULL) +
  theme_classic(base_size=15)
Plot4 <- ggplot(num, aes(Edad,Tarifa_Pasaje)) +geom_point()+
  labs(title="Grafico de dispersion", y="Tarifa_Pasaje", x="Edad",color=NULL) +
  theme_classic(base_size=15)
Plot5 <- ggplot(num, aes(Tarifa_Pasaje,Clase_Pasajero)) +geom_point()+
  labs(title="Grafico de dispersion", y="Clase_Pasajero", x="Tarifa_Pasaje",color=NULL) +
```

```
  theme_classic(base_size=15)
Plot6 <- ggplot(num, aes(Tarifa_Pasaje,Pariente_Masculino)) +geom_point()+
  labs(title="Grafico de dispersion", y="Pariente_Masculino", x="Tarifa_Pasaje",color=NULL) +
  theme_classic(base_size=15)
Plot7 <- ggplot(num, aes(Tarifa_Pasaje,Pariente_Femenino)) +geom_point()+
  labs(title="Grafico de dispersion", y="Pariente_Femenino", x="Tarifa_Pasaje",color=NULL) +
  theme_classic(base_size=15)
Plot8 <- ggplot(num, aes(Pariente_Masculino,Pariente_Femenino)) +geom_point()+
  labs(title="Grafico de dispersion", y="Pariente_Femenino", x="Pariente_Masculino",color=NULL) +
  theme_classic(base_size=15)


grid.arrange(Plot1, Plot2, Plot3, Plot4, Plot5,Plot6, Plot7,Plot8, ncol = 3)
```



Para el grafico de dispersión se observa que en ninguna de las distribuciones se encuentra alguna relacion que diga que hay dependencia entre ambas variables. Donde en al única que podría decirse que hay cierta tendencia a un patron es en clase de pasajero y tarifa del pasajero.

CONCLUSIONES

- se obtuvieron los siguientes resultados que el 60% de los tripulantes murieron. Había más hombre que mujeres a bordo, donde había un 63% de hombre. la media de los pasajeros era de segunda clase. Y la media de la edad eran de alrededor de los 29 años.

- Con respecto a la tarifa están tienen un gran sesgo, donde su media y mediana son de 34 y 16, el cual es resultado de las grandes aberrancias que había debido a los precios.
- El número de parientes hombres y mujeres son casi similares con una media y mediana igual a cero.
- En el estudio bivariado se encuentra que la mayoría de los vivos fueron las mujeres.
- No existieron correlaciones o dependencia entre variables, donde la única que tenia la una ligera correlación fuerte es tarifa del precio y nivel de clase de pasajero.

Autor: Luis Garcia, 23/11/2020