PRA2 - Tipología y ciclo de vida de los datos

Jonathan Ayuquina y Martha Ayuquina

27 de mayo, 2021

Contents

1 Descripción del dataset. **1**
- 1.1 Importancia . . . 2
- 1.2 Pregunta/problema que se pretende responder . . . 3

2 Integración y selección de los datos de interés a analizar. **3**
- 2.1 Integración . . . 3
- 2.2 Selección . . . 4

3 Limpieza de los datos. **5**
- 3.1 ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos? . 5
- 3.2 Identificación y tratamiento de valores extremos. . . . 7

4 Análisis de los datos. **11**
- 4.1 Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar). . . . 11
- 4.2 Comprobación de la normalidad y homogeneidad de la varianza. . . . 11
- 4.3 Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes. . . . 17

5 Representación de los resultados a partir de tablas y gráficas. **20**

6 Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema? **30**

1 Descripción del dataset.

Desde Kaggle https://www.kaggle.com/c/titanic. obtenemos información sobre pasajeros del Titanic en dos partes, conjunto de entrenamiento y un conjunto de prueba:

- **test.csv** contiene 418 registros.
- **train.csv** contiene 891 registros.

Primero se cargan los ficheros test.csv y train.csv en los objetos correspondientes prueba y entrenamiento, respectivamente e identificando los valores NA.

```
# Cargamos el dataset https://www.kaggle.com/c/titanic/data?select=test.csv
titanic.test <- read.csv("test.csv", na.strings = c("NA", ""))
# Cargamos el dataset https://www.kaggle.com/c/titanic/data?select=train.csv
titanic.train <- read.csv("train.csv", na.strings = c("NA", ""))
```

A continuación, se muestra y examina las variables del objeto prueba, el resultado devuelve 418 observaciones y 11 variables: PassengerId, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked.

```
# Mostramos la estructura interna del objeto test
str(titanic.test)
```

```
## 'data.frame':    418 obs. of  11 variables:
##  $ PassengerId: int  892 893 894 895 896 897 898 899 900 901 ...
##  $ Pclass     : int  3 3 2 3 3 3 3 2 3 3 ...
##  $ Name       : chr  "Kelly, Mr. James" "Wilkes, Mrs. James (Ellen Needs)" "Myles, Mr. Thomas Francis
##  $ Sex        : chr  "male" "female" "male" "male" ...
##  $ Age        : num  34.5 47 62 27 22 14 30 26 18 21 ...
##  $ SibSp      : int  0 1 0 0 1 0 0 1 0 2 ...
##  $ Parch      : int  0 0 0 0 1 0 0 1 0 0 ...
##  $ Ticket     : chr  "330911" "363272" "240276" "315154" ...
##  $ Fare       : num  7.83 7 9.69 8.66 12.29 ...
##  $ Cabin      : chr  NA NA NA NA ...
##  $ Embarked   : chr  "Q" "S" "Q" "S" ...
```

También se muestra y examina las variables del objeto entrenamiento, el resultado muestra 891 observaciones y 12 variables. Contiene los mismos atributos que el objeto de prueba más la variable adicional Survived.

```
# Mostramos la estructura interna del objeto train
str(titanic.train)
```

```
## '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  NA "C85" NA "C123" ...
##  $ Embarked   : chr  "S" "C" "S" "S" ...
```

1.1 Importancia

Como se conoce, el barco RMS Titanic en abril de 1912 durante el viaje inaugural desde Southampton a Nueva York se hundió luego de chocar con un iceberg. Al no contar con suficientes botes salvavidas se produjo la muerte de 1502 de los 2224 pasajeros y la tripulación, es decir, hubo un 68% de personas fallecidas en el accidente.

Basado en lo expuesto, nos interesa efectuar un proyecto de análisis para aprendizaje automático y limpieza de datos por las características que tiene el conjunto de datos, esto es:

- Las observaciones se recogen de pasajeros del Titanic suficiente para realizar un modelo.
- Permite reunir variables continuas y cualitativas para realizar análisis exploratorios.
- Tiene valores ausentes para ser tratados (eliminación o imputación) que influyen en el modelo.
- Requiere de proceso de limpieza y conversión de datos.

Luego de ello, podemos determinar los factores que influyeron en la supervivencia de los pasajeros y así,

mediante modelos predictores y la evaluación de los mismos, estimar las probabilidades de supervivencia de un pasajero.

Este proyecto es relevante para estimar probabilidades de supervivencia de tripulantes ante accidentes marítimos, los cuales en la actualidad no están exentos de darse pese a que existen barcos con infraestructura segura. Entonces, es necesario establecer mecanismos de rescate y salvar vidas en situaciones de emergencia.

1.2 Pregunta/problema que se pretende responder

Es de suponer que ciertos grupos de personas tuvieron más probabilidades de supervivencia siendo necesario un análisis para conocer los factores clave. Al respecto, surgen las siguientes preguntas:

¿Las mujeres tienen más probabilidades de supervivencia que los hombres?

¿Los familiares de los pasajeros tienen mayor probabilidad de supervivencia que los mismos pasajeros?

¿La clase de boleto que adquiere el pasajero influye para sobrevivir?

¿El punto de embarque afectó la probabilidad de supervivencia?

¿Es diferente la edad promedio de los sobrevivientes en comparación a la edad promedio de los fallecidos?

2 Integración y selección de los datos de interés a analizar.

2.1 Integración

En este punto se realiza una exploración de los conjuntos para entender la información que tienen los datos que están divididos en dos conjuntos: entrenamiento y prueba.

A breves rasgos se observa que ambos conjuntos tienen los mismos nombres de atributos y tipo de datos, excepto por la variable "Survived" que se encuentra en el conjunto de entrenamiento mas no en el conjunto de prueba.

La variable "Survived" marca la diferencia entre ambos conjuntos porque en el set de entrenamiento está antedicho de un pasajero sobreviviente y no sobreviviente, mientras en el de pruebas se desconoce si un pasajero sobrevive o no.

Adicionalmente, el atributo "PassengerId" es un código único en ambos conjuntos permitiendo una integración vertical para conformar cantidad mayor de registros y analizarlas globalmente.

Como el atributo "Survived" consta solo en el conjunto de entrenamiento y siendo necesario para nuestro análisis, lo reservaremos en la fusión, para lo cual crearemos la variable "Survived" en el conjunto de prueba con valor vacío. La finalidad es emplear todas las variables cargadas y asignar los tipos que le corresponde almacenar según su naturaleza.

```
# Creamos la variable sobreviviente con el valor NA
titanic.test$Survived <- NA
```

Ahora que tienen las mismas variables se crea la etiqueta "DatasetType" antes de combinar el dataset para determinar la procedencia de las observaciones fácilmente.

```
# Llenar la variable con etiqueta Test y Train
titanic.test$DatasetType <- "Test"
titanic.train$DatasetType <- "Train"
```

Una vez marcado los datos en cada conjunto se fusionan y una sola estructura es presentada.

```
# Se mezcla el datset en uno
titanic.data <- rbind(titanic.test, titanic.train)
# Mostrar la estructura integrada
str(titanic.data)
```

```
## 'data.frame':    1309 obs. of  13 variables:
##  $ PassengerId: int  892 893 894 895 896 897 898 899 900 901 ...
##  $ Pclass     : int  3 3 2 3 3 3 2 3 3 ...
##  $ Name       : chr  "Kelly, Mr. James" "Wilkes, Mrs. James (Ellen Needs)" "Myles, Mr. Thomas Francis
##  $ Sex        : chr  "male" "female" "male" "male" ...
##  $ Age        : num  34.5 47 62 27 22 14 30 26 18 21 ...
##  $ SibSp      : int  0 1 0 0 1 0 0 1 0 2 ...
##  $ Parch      : int  0 0 0 0 1 0 0 1 0 0 ...
##  $ Ticket     : chr  "330911" "363272" "240276" "315154" ...
##  $ Fare       : num  7.83 7 9.69 8.66 12.29 ...
##  $ Cabin      : chr  NA NA NA NA ...
##  $ Embarked   : chr  "Q" "S" "Q" "S" ...
##  $ Survived   : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ DatasetType: chr  "Test" "Test" "Test" "Test" ...
```

La integración de los conjuntos elegidos contiene en total 1309 observaciones y 13 variables que se detallan a continuación:

Variable	Descripción	Tipo
PassengerId	Identificador único del pasajero.	Cuantitativa discreta
Survived	Indicador si el pasajero sobrevivió “1” o no “0”.	Cualitativa nominal
Pclass	Clase de boleto del pasajero “1”, “2” o “3”.	Cualitativa ordinal
Name	Nombre del pasajero.	Cualitativa nominal
Sex	Sexo del pasajero con valores “male” y “female”.	Cualitativa nominal
Age	Edad del pasajero.	Cuantitativa discreta
SibSp	Número de hermanas/os, hermanastras/os, cónyuges en el barco.	Cuantitativa discreta
Parch	Número de padres e hijos que tenían a bordo los pasajeros.	Cuantitativa discreta
Ticket	Identificador del boleto.	Cualitativa nominal
Fare	Precio/tarifa pagado por el boleto.	Cuantitativa continua
Cabin	Identificación de la cabina/camarote asignado al pasajero.	Cualitativa nominal
Embarked	Puerto de embarque (Q=Queenstown, C=Cherburgo y S=Southampton)	Cualitativa nominal
DatasetType	Etiqueta del datset “Test” o “Train”.	Cualitativa nominal

2.2 Selección

Para dar respuesta a las interrogantes del numeral 1.2 y revisando de manera general el contenido de todas las variables puede comprenderse el efecto de la clase, sexo, edad, camarote, etc. en la supervivencia de los pasajeros.

La variable “Survived” es muy importante porque intersa estudiarla respecto a las demás para finalmente predecirla.

Notamos que la variable “Ticket” tiene muchos valores únicos (929) y no es relevante para determinar la supervivencia de un pasajero, entonces se decide ignorar este atributo.

```
# Contar los números de tickets distintos
length(unique(titanic.data$Ticket))
```

```
## [1] 929
```

Más adelante, conforme analicemos cada atributo, determinaremos si una variable es necesaria en el análisis.

Contrario a la selección, se crearán los siguientes atributos:

- Designation: Variable que contiene la designación honorífica de personas, el cual se extrae del contenido del campo “Name” ya que a más del nombre del pasajero, el atributo contiene los tratamientos de

cortesía luego de la coma y este detalle permite generar una nueva distribución de característica de las personas. Ej: Mr, Miss, Mrs, entre otros.

- FamilySize: Contiene el grupo de familia según su número. Para el efecto, empleamos la suma por registro de los atributos "SibSp" y "Parch" más 1 y catalogamos en: Solo si el resultado es 1; Familia pequeña si el resultado es mayor a 2 y menor a 5; Familia numerosa si es mayor a 4.
- AgeForGroup: Contiene la clasificación de las personas por edad. Considerando el atributo "Age", si la edad es menor a 18 años se cataloga en el nuevo atributo como "Menor de edad", caso contrario, si la edad es mayor a 18 se cataloga en el nuevo atributo como "Mayor de edad".

3 Limpieza de los datos.

Con el nuevo conjunto de datos es más fácil limpiar los datos para el análisis y predicción.

3.1 ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?

Se verifica si hay valores perdidos dentro del dataset empleando la función is.na(), después resumimos el número de elementos vacíos en cada variable contando con colSums().

```
# Contar elementos vacíos
colSums(is.na(titanic.data))
```

```
## PassengerId      Pclass        Name         Sex         Age       SibSp
##           0           0           0           0         263           0
##       Parch      Ticket        Fare       Cabin    Embarked    Survived
##           0           0           1        1014           2         418
## DatasetType
##           0
```

El resultado muestra los siguientes casos: - Las variables "Fare" y "Embarked" tienen pocos valores perdidos esto es 1 y 2 elementos. - Hay una cantidad significativa de valores perdidos en las variables "Age" y "Cabin", dado que tienen 263 y 1014 elementos ausentes respectivamente. - La variable Survived denota 418 valores ausentes.

A continuación se examina cada caso para aplicar el tratamiento según amerite.

Como la variable "Embarked" tiene 2 valores perdidos, no preocupa reeemplazarlos con el valor de tendencia, es decir, el mayor valor de la tabla de distribución de los puertos "Q", "C" y "S.

```
# Tabla de distribución de puertos
table(titanic.data$Embarked)
```

```
##
##   C   Q   S
## 270 123 914
```

La mayor cantidad de pasajeros proviene del puerto Southampton "S" con 914 personas y se considerará esta tendencia para reemplazar el valor faltante con "S".

```
# Reemplazar NA con puerto Southampton "S" para mantener la tendencia
titanic.data$Embarked[is.na(titanic.data$Embarked)] <- "S"
```

Así mismo la variable "Fare" tiene 1 valor perdido y puede ser reemplazado por la mediana de valores de "Fare", esto es $14.45.

```
fare.median <- median(titanic.data$Fare, na.rm = TRUE)
fare.median
```

```
## [1] 14.4542
```

```
titanic.data$Fare[is.na(titanic.data$Fare)] <- fare.median
```

Examinando la variable "Cabin" tiene muchos valores diferentes y 1014 están perdidos (1014/1209 = 77% valores perdidos), no se tiene cómo asignarle valor, entonces se decide ignorar este atributo completamente.

```
tail(titanic.data$Cabin)
```

```
## [1] NA     NA     "B42"  NA     "C148" NA
```

La variable "Age" podría tener la misma gestión que "Fare" para llenarse con la mediana, pero tiene 263 valores NA y emplear el método no podría ser preciso. Por lo que utilizaremos un modelo de regresión en base a las variables existentes para predecir los elementos vacíos.

Previamente se revisan las variables para constatar si hay alguna que contribuya al análisis de predicción de la edad faltante y transformarla.

Por mencionar, en la variable "Name" se encuentran los nombres de pasajeros que podemos explorar con la función head().

```
# Devolver una parte del datset
head(titanic.data$Name)
```

```
## [1] "Kelly, Mr. James"
## [2] "Wilkes, Mrs. James (Ellen Needs)"
## [3] "Myles, Mr. Thomas Francis"
## [4] "Wirz, Mr. Albert"
## [5] "Hirvonen, Mrs. Alexander (Helga E Lindqvist)"
## [6] "Svensson, Mr. Johan Cervin"
```

```
# Contar los nombres de pasajeros
length(unique(titanic.data$Name))
```

```
## [1] 1307
```

El resultado muestra el nombre de 1307 personas y es de resaltar que el nombre incluye la designación honorífica del pasajero cuando se registró, como puede ser: "Mr.", "Mrs.", "Miss", etc. Por esto se crea otra variable donde extraer esta clase y observar la frecuencia absoluta de las Variables "Sex" y "Designation".

```
titanic.data$Designation <- gsub("^.*, (.*?)\\..*$", "\\1", titanic.data$Name)

# Tabla de frecuencia absoluta
table(titanic.data$Sex, titanic.data$Designation)
```

```
##
##          Capt Col Don Dona  Dr Jonkheer Lady Major Master Miss Mlle Mme  Mr Mrs
##   female    0   0   0    1   1        0    1     0      0  260    2   1   0 197
##   male      1   4   1    0   7        1    0     2     61    0    0   0 757   0
##
##           Ms Rev Sir the Countess
##   female   2   0   0            1
##   male     0   8   1            0
```

La tabla muestra valores en las clases "Miss", "Mrs", "Mr." y en otras que coinciden con estas designaciones. Siendo así, se podría agruparlas Por ejemplo: la clase "Mlle" y "Ms" son lo mismo que "Miss". La clase "Mme" es lo mismo que "Mrs". Los grupos con menor cantidad de repetición podrían reunirse en la categoría "Other".

```
titanic.data$Designation[titanic.data$Designation %in% c("Mlle", "Ms")] <- "Miss"
titanic.data$Designation[titanic.data$Designation %in% c("Mme")] <- "Mrs"
```

```
titanic.data$Designation[titanic.data$Designation %in% c("Capt", "Col", "Don",
 "Dona", "Dr", "Jonkheer", "Lady", "Major", "Rev", "Sir", "the Countess")] <-
 "Other"
```

Se muestra el resultado de combinar las clases en una tabla que contenga las agrupaciones efectuadas.

```
# Tabla de frecuencia sexo y titulación
table(titanic.data$Sex, titanic.data$Designation)
```

```
##
##          Master Miss  Mr Mrs Other
##   female      0  264   0 198     4
##   male       61    0 757   0    25
```

Por otro lado, la cantidad de las variables "SibSp" y "Parch", corresponden a familiares de personas a bordo del barco y pueden sumarse adicionalmente a 1 pasajero para formar otra que simplifique el conjunto de familia.

```
# Crear variable FamilySize
titanic.data$FamilySize <- titanic.data$SibSp+titanic.data$Parch+1
```

Según esto, con el número de los integrantes de la familia determinamos un tamaño de familia: "Alone" si solo viaja el pasajero (sin familia); "Small" si los familiares y el pasajero conforman un grupo máximo de 4 personas y "Large" si el número de familiares es mayor a 4.

```
titanic.data$FamilyGroup[titanic.data$FamilySize==1] <- "Alone"
titanic.data$FamilyGroup[titanic.data$FamilySize>1 & titanic.data$FamilySize<=4] <- "Small"
titanic.data$FamilyGroup[titanic.data$FamilySize>4] <- "Large"
```

3.2 Identificación y tratamiento de valores extremos.

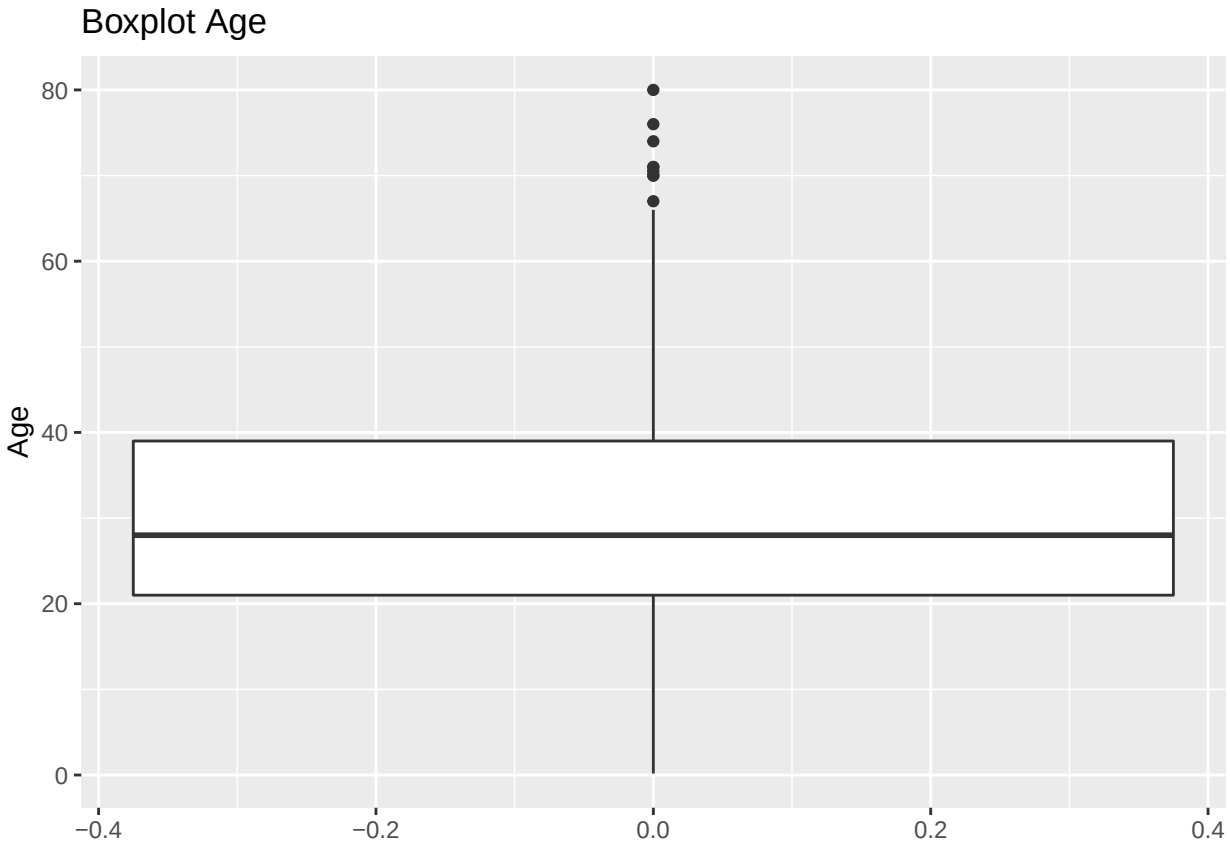
La identificación de valores extremos se realiza a través de diagramas de cajas para interpretar los defectos en los datos de las variables cuantitativas "Fare" y "Age".

Inicialmente se cargan las librerías de R para realizar los gráficos entre variables.

Se muestra el diagrama de la variable "Age" para ver cuántos valores están fuera de la caja, a través de la función ggplot().

```
ggplot(titanic.data, aes(y=Age), na.rm = TRUE)+
  geom_boxplot()+ggtitle("Boxplot Age")
```

```
## Warning: Removed 263 rows containing non-finite values (stat_boxplot).
```



El diagrama de cajas muestra algunas edades atípicas que deben eliminarse porque pueden hacer las predicciones menos precisas. La función boxplot.stats() detalla los valores de los que se trata.

```
# Mostrar edades fuera del extremo
boxplot.stats(titanic.data$Age)$out
```

```
## [1] 67.0 76.0 71.0 70.5 71.0 80.0 70.0 70.0 74.0
```

```
length(boxplot.stats(titanic.data$Age)$out)
```

```
## [1] 9
```

La salida ubica 9 personas del grupo etario vejez fuera de la distribución, esto no significa que sean incoherentes. Y ee estos valores se toma la edad de extremo superior del bigote (66) para filtrar edades hasta el máximo.

```
# Obtener el extremo superior del diagrama
upper.age <- boxplot.stats(titanic.data$Age)$stats[5]
upper.age
```

```
## [1] 66
```

```
# Filtrar valores menores que 66
age.outlier.filter <- titanic.data$Age<upper.age
```

Se rellena los valores perdidos de la variable “Age” utilizando un modelo de regresión con el conjunto de datos de valores atípicos filtrados.

```
age.model <- lm(Age ~ Pclass + Sex + Fare + FamilyGroup + Embarked + Designation, data =
 titanic.data[age.outlier.filter,])
age.na.data <- titanic.data[is.na(titanic.data$Age), c("Pclass", "Sex", "Fare",
```

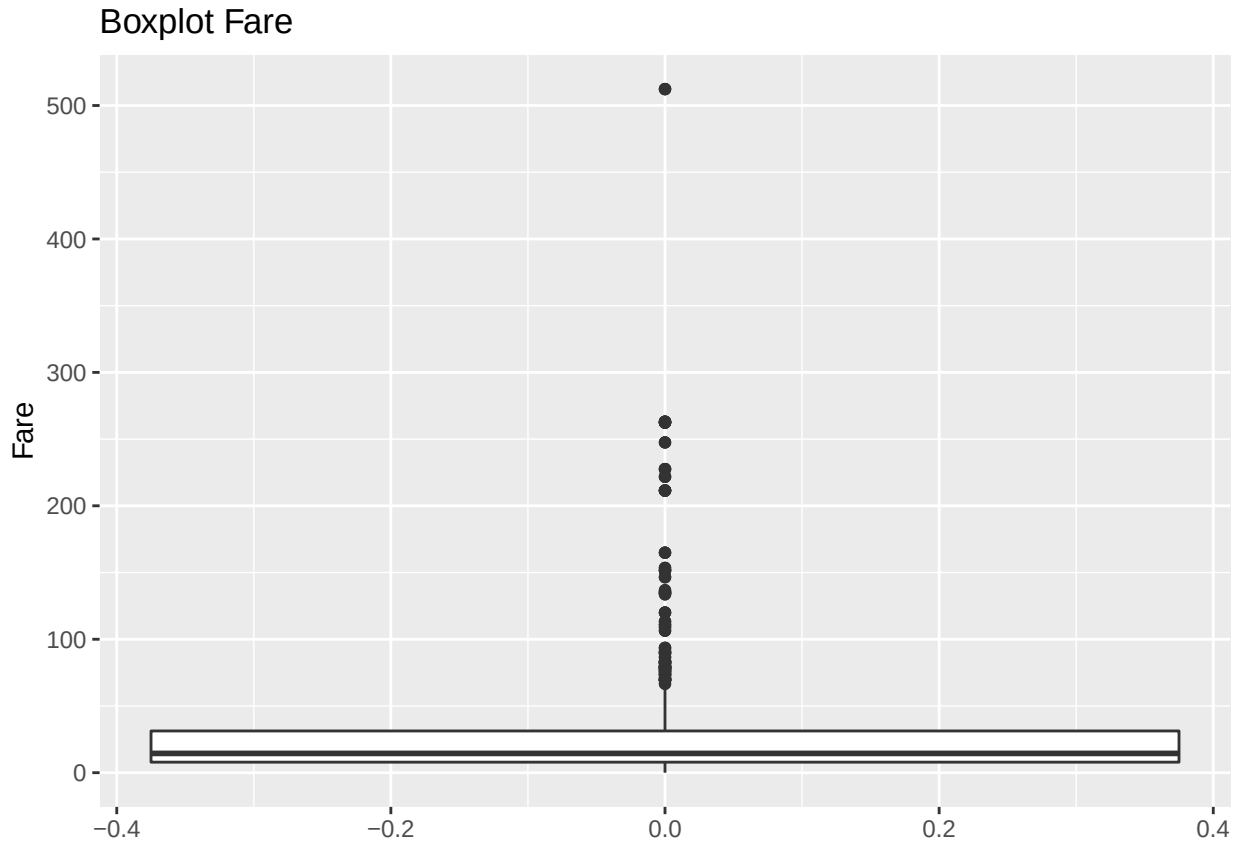
```
 "FamilyGroup", "Embarked", "Designation")]
titanic.data$Age[is.na(titanic.data$Age)] <- predict(age.model, age.na.data)
```

Ahora se puede dividir la variable "Age" en dos grupos de pasajeros: los de menos de 18 años como niños y los mayores como adultos.

```
titanic.data$AgeForGroup[titanic.data$Age<18] <- "Child"
titanic.data$AgeForGroup[titanic.data$Age>=18] <- "Adult"
```

Analizamos también los valores que estén fuera para la variable "Fare" con un diagrama de cajas.

```
ggplot(titanic.data, aes(y=Fare), na.rm = TRUE)+
  geom_boxplot()+ggtitle("Boxplot Fare")
```



Evidenciamos que existen 171 valores extremos considerando el valor extremo superior del bigote (65)

```
# Obtener el extremo superior del diagrama
upper.fare <- boxplot.stats(titanic.data$Fare)$stats[5]
upper.fare
```

```
## [1] 65
```

```
# Filtrar valores menores que $65
length(titanic.data$Fare[titanic.data$Fare>upper.fare])
```

```
## [1] 171
```

Evaluando el valor mínimo y máximo de tarifa, se obtiene 0 y 512.33 respectivamente. Las tarifas con valor cero suponen que los pasajeros no pagaron por su ticket de entrada porque fueran políticos o personas con

importancia económica o también personas que recibieron boletos gratuitos.

```
#Obtengo valor mínimo y máximo
min(titanic.data$Fare)
```

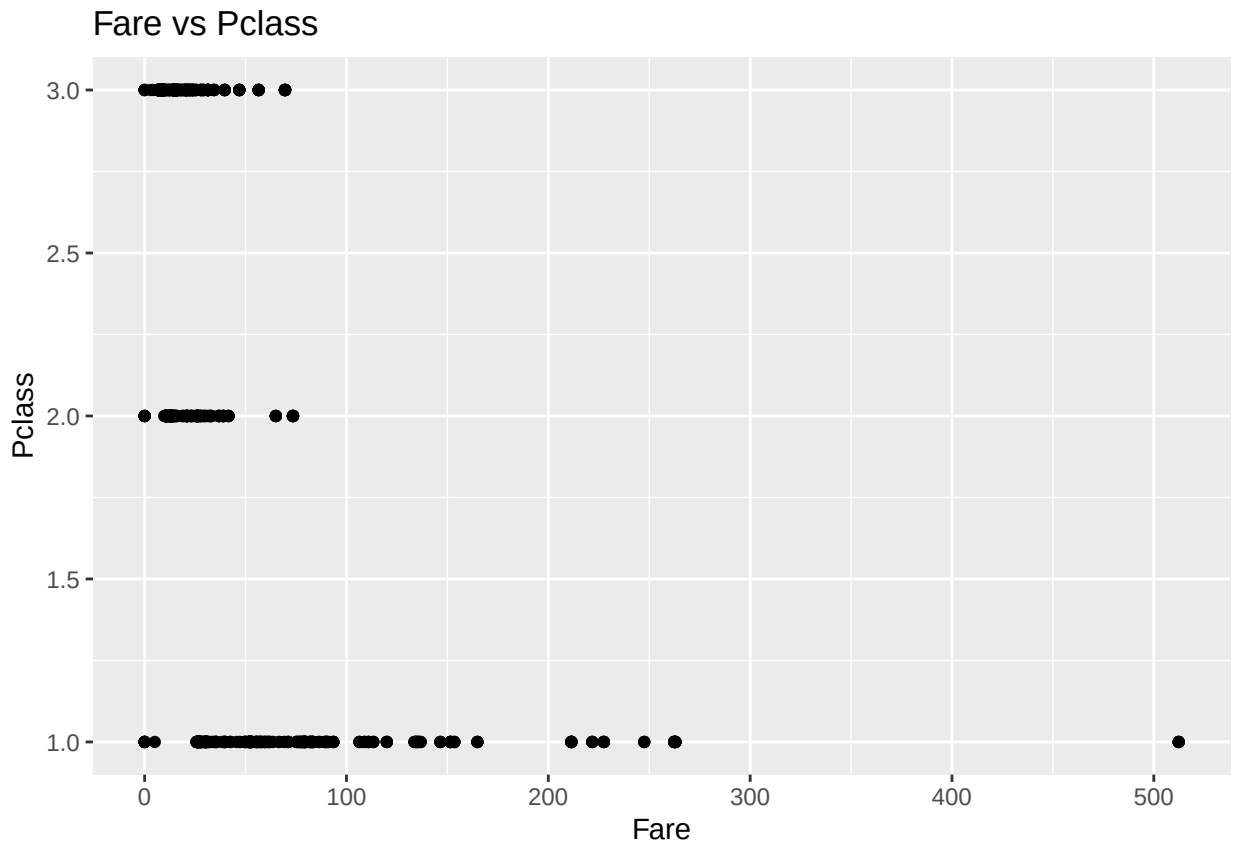
```
## [1] 0
```

```
max(titanic.data$Fare)
```

```
## [1] 512.3292
```

Se realiza una gráfica de comparación entre la tarifa y la clase de boleto para descartar el registro de tarifa cero y se constata que, para las tres clases existen tarifa cero; sin embargo, es posible notar una relación entre ambas variables determinando que no es necesario modificar la tarifa porque son datos legítimos. Entonces, consideramos los outliers de la variable Fare en nuestro análisis.

```
ggplot(titanic.data, aes(x=Fare, y=Pclass))+
  geom_point()+ggtitle("Fare vs Pclass")
```



Realizado el tratamiento de valores extremos se verifica las columnas para asegurarse que no hay más valores perdidos.

```
colSums(is.na(titanic.data))
```

```
## PassengerId      Pclass        Name         Sex         Age       SibSp
##           0           0           0           0           0           0
##       Parch      Ticket        Fare       Cabin    Embarked    Survived
##           0           0           0        1014           0         418
## DatasetType Designation  FamilySize FamilyGroup AgeForGroup
##           0           0           0           0           0
```

Se convierte las variables a los tipos correctos. La variable Survived no conviene almacenarla en formato numérico ya que esto puede llevar a errores como tratar de calcular la media, por lo que se convierte a factor.

```
titanic.data$Pclass <- as.factor(titanic.data$Pclass)
titanic.data$Sex <- as.factor(titanic.data$Sex)
titanic.data$Embarked <- as.factor(titanic.data$Embarked)
titanic.data$Designation <- as.factor(titanic.data$Designation)
titanic.data$FamilySize <- as.factor(titanic.data$FamilySize)
titanic.data$FamilyGroup <- as.factor(titanic.data$FamilyGroup)
titanic.data$AgeForGroup <- as.factor(titanic.data$AgeForGroup)
titanic.data$Survived <- as.factor(titanic.data$Survived)
```

Finalmente hay que dividir el conjunto original con la configuración realizada a las variables.

```
new.titanic.test <- titanic.data[titanic.data$DatasetType=="Test",]
new.titanic.train <- titanic.data[titanic.data$DatasetType=="Train",]
```

4 Análisis de los datos.

4.1 Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).

Como nuestro interés es la respuesta a las preguntas planteadas en el numeral 1 que se enfoca en la explicación de la supervivencia de los pasajeros, determinamos la importancia de las siguientes variables Survived, Sex, Pclass, Embarked, FamilyGroup.

Para seleccionar los grupos de datos a comparar consideraremos los siguientes pasos:

1. Análisis estadístico descriptivo para tener un resumen de cada uno de los atributos del conjunto. Aunque es de aclarar que, durante la limpieza de datos ya realizamos una visión general para estudiar los atributos.
2. Análisis estadístico inferencial, mediante el cual disponiendo de una muestra de datos, procederemos a:

- Conocer que los datos siguen una distribución normal y homocedasticidad.
- Efectuar el análisis de correlación entre pares de variables.
- Crear un modelo de regresión lineal.

4.2 Comprobación de la normalidad y homogeneidad de la varianza.

Para estudiar si la muestra proviene de una población con distribución normal se disponen de tres herramientas: histogramas, gráfico de cuantil cuantil y prueba de hipótesis.

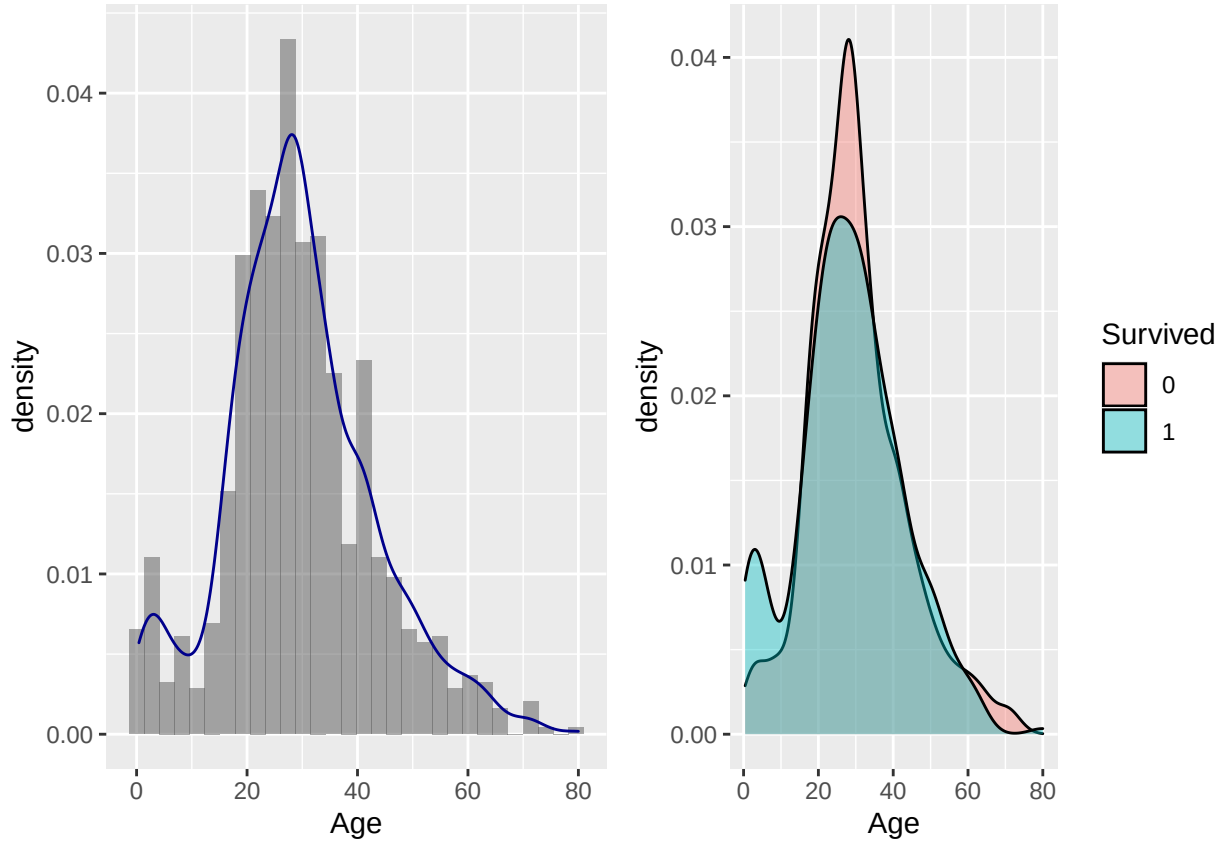
En el histograma o gráfico de densidad se explora la normalidad presente mediante un patrón más o menos simétrico.

A continuación se construye el histograma y gráfico de densidad de los datos considerando que de los pasajeros se desea saber si la variable Age tiene una distribución normal.

Además se hace la relación con la variable Survived para ver la condición de supervivencia.

```
age.density.plot <- ggplot(new.titanic.train, aes(x = Age))+
  geom_histogram(aes(y = ..density..), alpha = 0.5, position = "identity")+
  geom_density(color = "darkblue", alpha = 0.2)

age.survived.density.plot <- ggplot(new.titanic.train, aes(x = Age, fill = Survived))+
  geom_density(alpha = 0.4)

grid.arrange(age.density.plot, age.survived.density.plot, ncol = 2)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



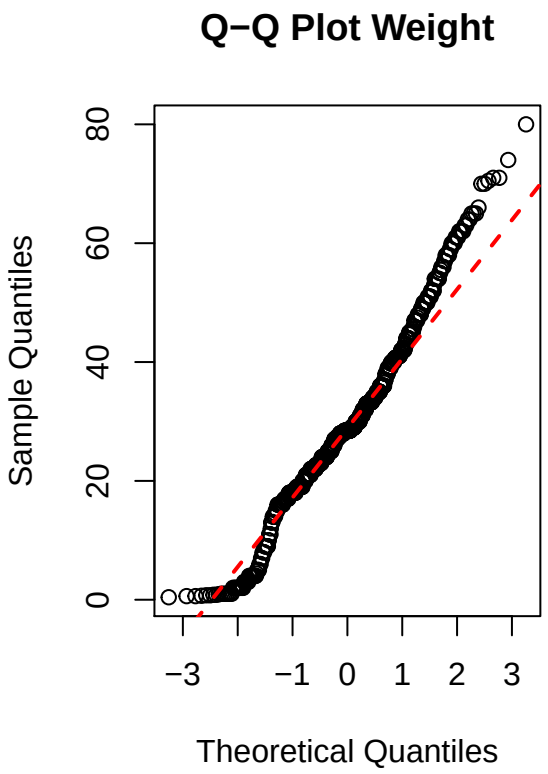
En el gráfico se observa:
- Que las densidades no son perfectamente simétricas.
- La mayor densidad se encuentra en el intervalo de 20 a 30 años. - Hay un sesgo hacia los pasajeros de menor edad, esto hace que se desconfíe de la normalidad de los datos.
- La distribución de edad de los pasajeros es similar entre el grupo de sobrevivientes y fallecidos.

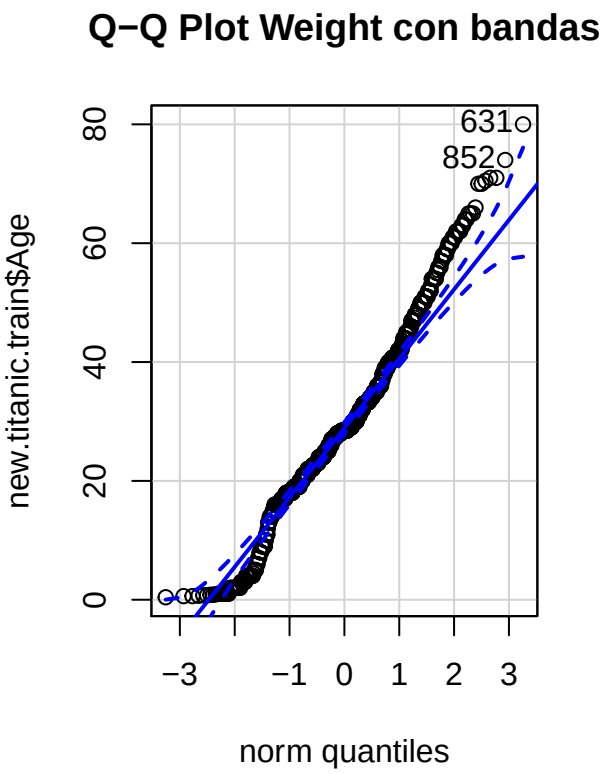
Otra herramienta como el QQplot podría aportar una mejor conclusión de la normalidad si los datos están perfectamente alineados a la línea de referencia , para ello se carga las librerías necesarias.

```
## Loading required package: carData
```

La función qqnorm() sirve para construir el QQplot y la función qqline() agrega una línea de referencia que ayuda a interpretar el gráfico de cierta distribución. La función qqplot() del paquete car permite mostrar bandas para los puntos del gráfico.

```
par(mfrow = c(1,2))
qqnorm(new.titanic.train$Age, main = "Q-Q Plot Weight")
qqline(new.titanic.train$Age, col="red", lwd = 2, lty = 2)
qqPlot(new.titanic.train$Age, main = "Q-Q Plot Weight con bandas")
```






```
## [1] 631 852
```

La figura del QQplot explica: - Los puntos de edad de pasajeros están desalineados.
- Mientras que con las bandas no todos los puntos están dentro.
- Esto lleva a rechazar la hipótesis de normalidad.

Por medio de las pruebas de normalidad se explora la normalidad del conjunto de datos formulando las hipótesis:

H_0: La distribución es normal
H_A: La distribución NO es normal

De los tipos de pruebas existentes se aplica la prueba de normalidad Shapiro Wilks en R con nivel de significancia del 5%.

```
shapiro.test(new.titanic.train$Age)
```

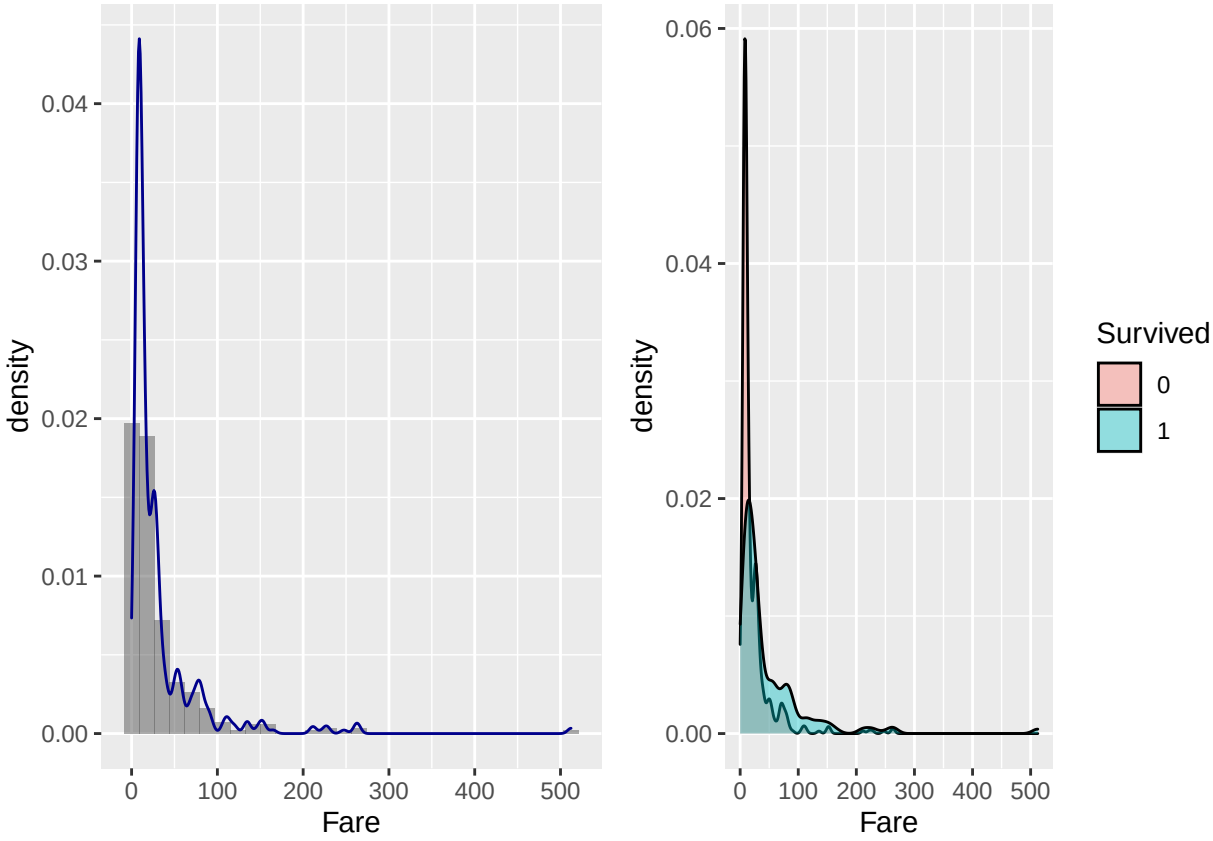
```
##
##  Shapiro-Wilk normality test
##
## data:  new.titanic.train$Age
## W = 0.97959, p-value = 7.753e-10
```

La salida anterior tiene un valor p 7.753e-10 para la prueba, dado que esto es menor que el nivel de significancia de 5% se debe rechazar la hipótesis nula que las edades de los pasajeros en la variable Age se distribuyen normalmente.

Con la variable Fare se grafica el histograma y su densidad para saber si los datos tienen una distribución normal y su relación con la variable Survived para ver la condición de supervivencia.

```r
fare.density.plot <- ggplot(new.titanic.train, aes(x = Fare))+
  geom_histogram(aes(y = ..density..), alpha = 0.5, position = "identity")+
  geom_density(color = "darkblue", alpha = 0.2)

fare.survived.density.plot <- ggplot(new.titanic.train, aes(x = Fare, fill = Survived))+
  geom_density(alpha = 0.4)

grid.arrange(fare.density.plot, fare.survived.density.plot, ncol = 2)
```

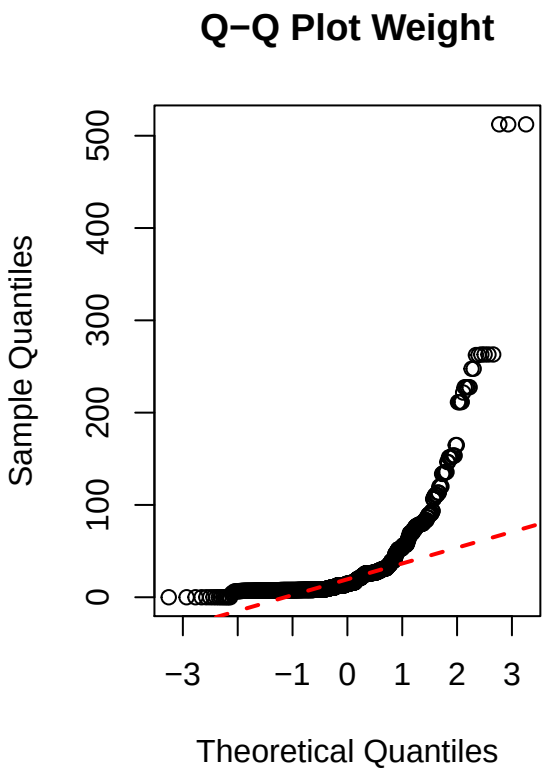
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

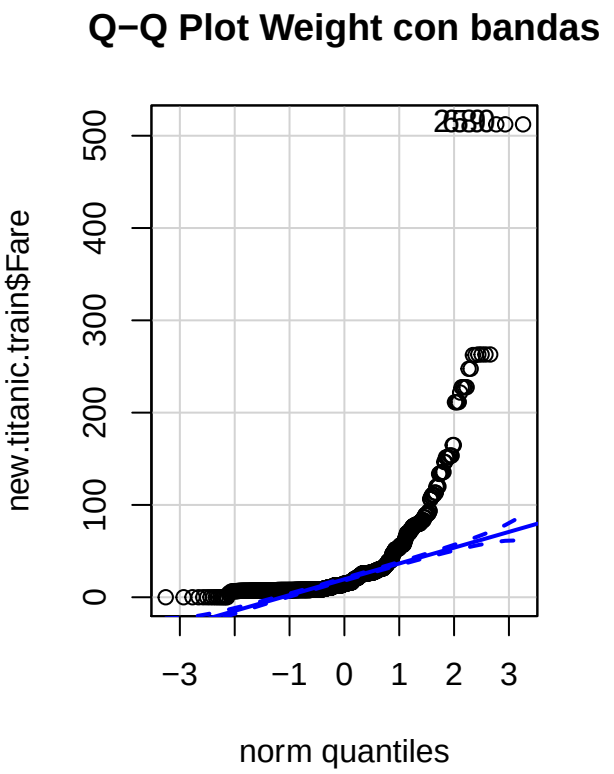


En el gráfico se observa:
- Que las densidades son asimétricas.
- La mayor densidad se encuentra en el intervalo de 0 a 50 (precio de boleto). - Hay un sesgo hacia los billetes de menor valor y unos pocos mayor valor, esto hace que se desconfíe de la normalidad de los datos.

Con la función qqplot() se organiza los puntos respecto a la línea de referencia y bandas para los puntos en la zona de normalidad.

```r
par(mfrow = c(1,2))
qqnorm(new.titanic.train$Fare, main = "Q-Q Plot Weight")
qqline(new.titanic.train$Fare, col="red", lwd = 2, lty = 2)
qqPlot(new.titanic.train$Fare, main = "Q-Q Plot Weight con bandas")
```






```
## [1] 259 680
```

En el gráfico se observa que la densidad del valor del billete no tiene simetría y poseen sesgo a los lados.

```
shapiro.test(new.titanic.train$Fare)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  new.titanic.train$Fare
## W = 0.52189, p-value < 2.2e-16
```

Por medio de las pruebas de normalidad el valor p de ambas muestras es menor al nivel de significancia de 5%, por tanto se puede rechazar la hipótesis nula que el valor del billete distribuye normalmente.

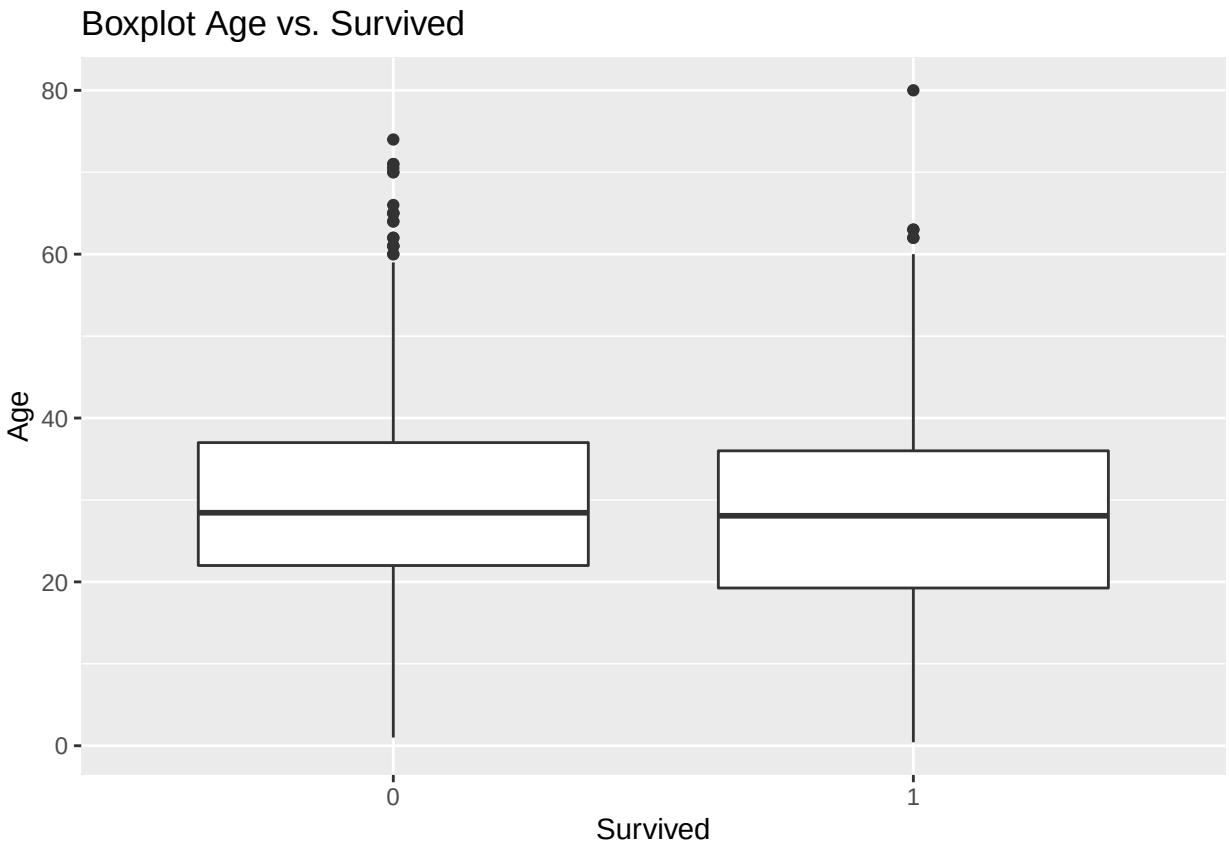
Como no tenemos la certeza de normalidad con las variables antes revisadas, para el análisis de la homgeneidad de varianza emplearemos el test no paramétrico Fligner-Killeen que se basa en la mediana. La hipótesis sería la siguiente:

Hipótesis nula (H_0): Las varianzas son iguales.

Hipótesis alternativa (H_1): Al menos dos de ellos difieren.

Aplicamos el supuesto de homogeneidad para la variable Age.

```
ggplot(new.titanic.train, aes(x=Survived, y=Age))+
  geom_boxplot()+ggtitle("Boxplot Age vs. Survived")
```



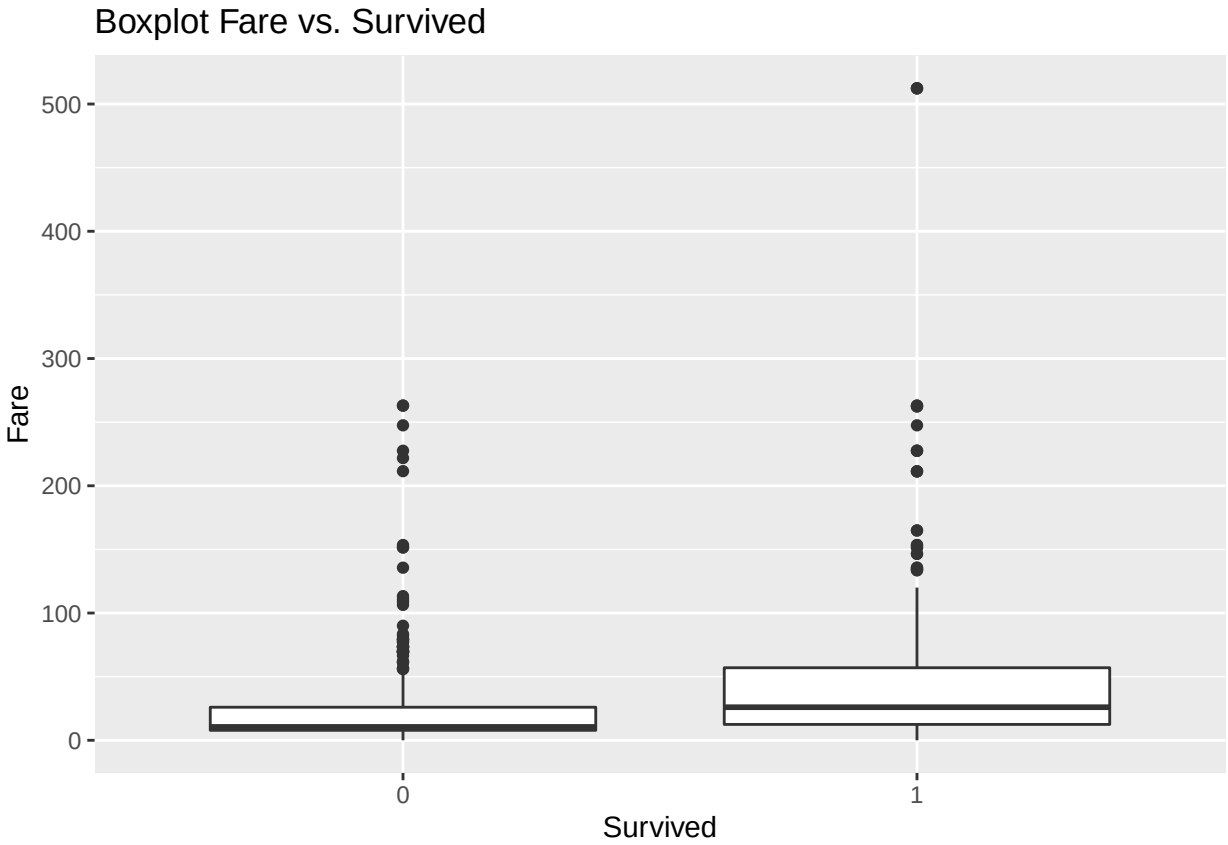

```
# Figner-Killeen Test of Homogeneity of Variances (p<.05 means sig different variances)
fligner.test(Age ~ Survived, data = new.titanic.train)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Age by Survived
## Fligner-Killeen:med chi-squared = 7.1072, df = 1, p-value = 0.007678
```

El resultado de la función genera un valor p = 0.007678, siendo menor que el nivel de significancia de 5% rechazando la hipótesis nula, no existe homogeneidad de varianzas entre Age y Survived.

Ahora revisamos el supuesto de homogeneidad para la variable Fare.

```
ggplot(new.titanic.train, aes(x=Survived, y=Fare))+
  geom_boxplot()+ggtitle("Boxplot Fare vs. Survived")
```




```
# Figner-Killeen Test of Homogeneity of Variances (p<.05 means sig different variances)
fligner.test(Fare ~ Survived, data = new.titanic.train)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Fare by Survived
## Fligner-Killeen:med chi-squared = 96.253, df = 1, p-value < 2.2e-16
```

Se obtiene valor p < 2.2e-16, el cual es menor al nivel de significancia de 5% rechazando la hipótesis nula, es decir, no hay homogeneidad de varianzas entre Fare y Survived.

4.3 Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.

Resumimos una estadística descriptiva básica de las columnas. Para las variables numéricas revisamos el valor mínimo, máximo, la media, mediana y los tres cuartiles y no se observa novedad. Se destaca que el valor de la media en Age es bastante similar a la mediana. 29.58 y 28.44 respectivamente.

```
summary(new.titanic.train)
```

```
##   PassengerId    Pclass      Name              Sex          Age
##  Min.   :  1.0   1:216   Length:891         female:314   Min.   : 0.42
##  1st Qu.:223.5   2:184   Class :character   male  :577   1st Qu.:21.00
##  Median :446.0   3:491   Mode  :character                Median :28.44
```

```
##  Mean   :446.0                                          Mean   :29.58
##  3rd Qu.:668.5                                          3rd Qu.:36.75
##  Max.   :891.0                                          Max.   :80.00
##
##      SibSp           Parch           Ticket               Fare
##  Min.   :0.000   Min.   :0.0000   Length:891         Min.   :  0.00
##  1st Qu.:0.000   1st Qu.:0.0000   Class :character   1st Qu.:  7.91
##  Median :0.000   Median :0.0000   Mode  :character   Median : 14.45
##  Mean   :0.523   Mean   :0.3816                      Mean   : 32.20
##  3rd Qu.:1.000   3rd Qu.:0.0000                      3rd Qu.: 31.00
##  Max.   :8.000   Max.   :6.0000                      Max.   :512.33
##
##     Cabin           Embarked Survived DatasetType        Designation
##  Length:891         C:168    0:549    Length:891         Master: 40
##  Class :character   Q: 77    1:342    Class :character   Miss  :185
##  Mode  :character   S:646             Mode  :character   Mr    :517
##                                                          Mrs   :126
##                                                          Other : 23
##
##
##    FamilySize  FamilyGroup AgeForGroup
##  1      :537   Alone:537   Adult:763
##  2      :161   Large: 62   Child:128
##  3      :102   Small:292
##  4      : 29
##  6      : 22
##  5      : 15
##  (Other): 25
```

Se emplea la función str() para visualizar el tipo de variable de cada columna y una parte de su contenido. Se constatan variables categóricas.

```
str(new.titanic.train)
```

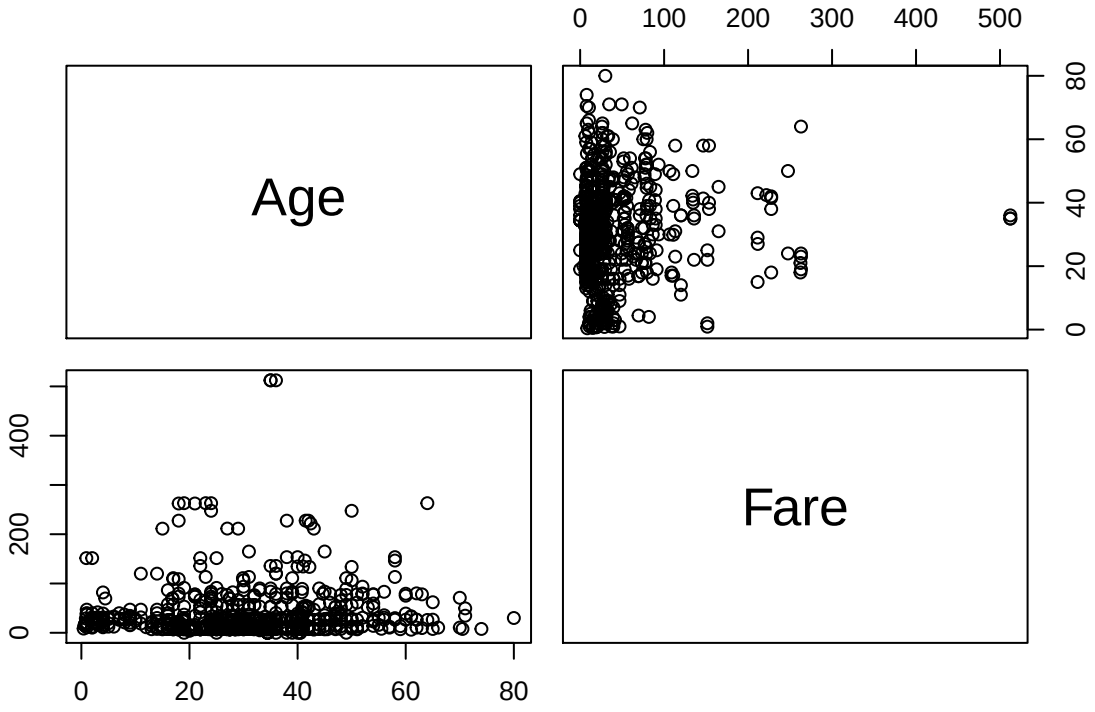
```
## 'data.frame':    891 obs. of  17 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Pclass     : Factor w/ 3 levels "1","2","3": 3 1 3 1 3 3 1 3 3 2 ...
##  $ Name       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)"
##  $ Sex        : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
##  $ Age        : num  22 38 26 35 35 ...
##  $ 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  NA "C85" NA "C123" ...
##  $ Embarked   : Factor w/ 3 levels "C","Q","S": 3 1 3 3 3 2 3 3 3 1 ...
##  $ Survived   : Factor w/ 2 levels "0","1": 1 2 2 2 1 1 1 1 2 2 ...
##  $ DatasetType: chr  "Train" "Train" "Train" "Train" ...
##  $ Designation: Factor w/ 5 levels "Master","Miss",..: 3 4 2 4 3 3 3 1 4 4 ...
##  $ FamilySize : Factor w/ 9 levels "1","2","3","4",..: 2 2 1 2 1 1 1 5 3 2 ...
##  $ FamilyGroup: Factor w/ 3 levels "Alone","Large",..: 3 3 1 3 1 1 1 2 3 3 ...
##  $ AgeForGroup: Factor w/ 2 levels "Adult","Child": 1 1 1 1 1 1 1 2 1 2 ...
```

En este punto se desconoce cuán relacionadas están las variables intervinientes, por lo que es importante descubrir la covarianza y correlación para resumir su relación lineal.

```
titanic.numeric.data <- subset(new.titanic.train, select = c("Age", "Fare"))
```

Se estudia la correlación de cada variable mediante un diagrama de dispersión múltiple, es la mejor manera de evaluar la linealidad entre dos variables.

```
pairs(titanic.numeric.data)
```



En el diagrama se puede encontrar un patrón lineal entre Age y Fare que puede ser interés para las preguntas estadísticas.

La covarianza es el estadístico que indica el sentido de los valores de muestras.
El valor positivo muestra que ambas variables varían en la misma dirección.
El valor negativo muestra que varían en la dirección opuesta.

En R se calcula con la función cov().

```
cov(titanic.numeric.data)
```

```
##             Age       Fare
## Age  181.67654   73.27489
## Fare  73.27489 2469.43685
```

La covarianza sólo informa sobre la dirección. En complemento, el coeficiente de correlación explica sobre el cambio en una variable e indica cuánto cambió de proporción en la segunda variable.
El coeficiente devuelve un valor entre -1 y 1.
La correlación es tanto más fuerte cuanto más se aproxime a 1.
La correlación negativa es tanto más fuerte cuanto más se aproxime a -1.

```
cor(titanic.numeric.data)
```

```
##             Age      Fare
## Age  1.0000000 0.1093973
## Fare 0.1093973 1.0000000
```

Estos resultados detallan la situación que puede resumirse a continuación:

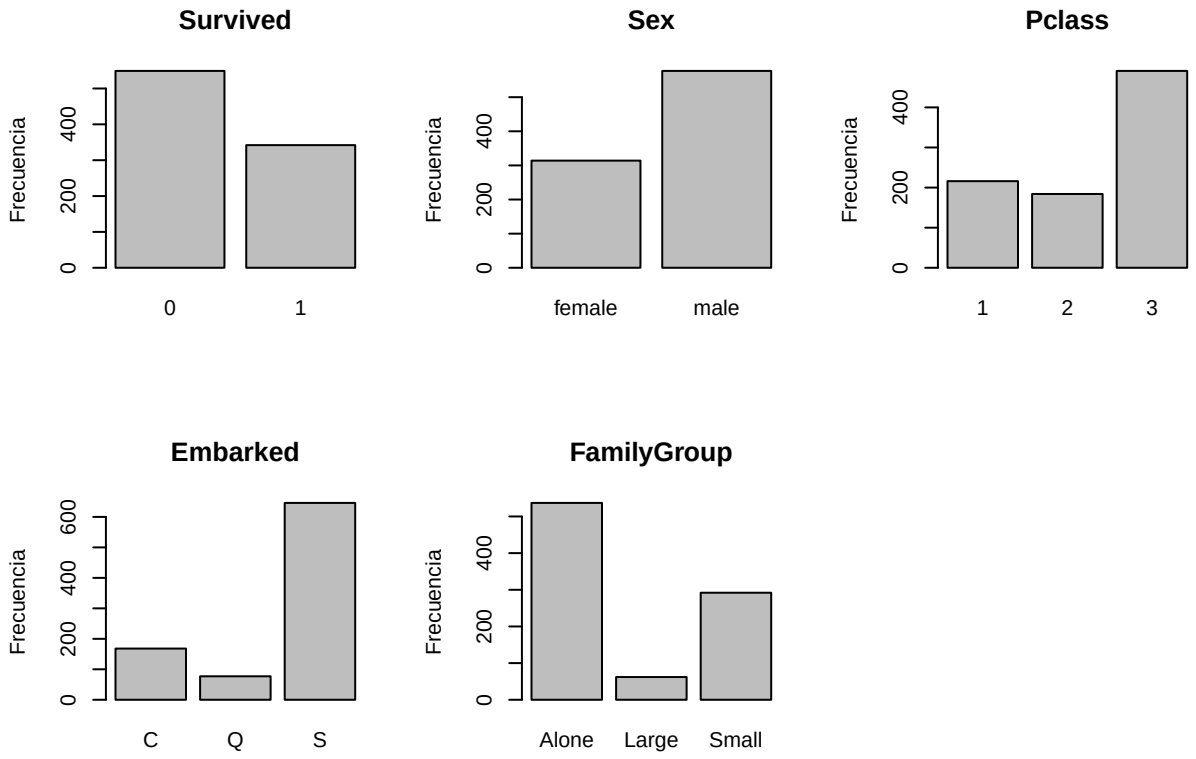
- El diagrama de dispersión tiene cierto patrón lineal, sí se alcanza a ver la dirección de los puntos, pero con fuerza de relación baja, se destacan las variables Age y Fare.
- El valor de covarianza de las variables Age y Fare son positivos y se interpreta que cambian en la misma dirección.
- La cuantificación del coeficiente de correlación establece baja relación entre las variables Age y Fare.

5 Representación de los resultados a partir de tablas y gráficas.

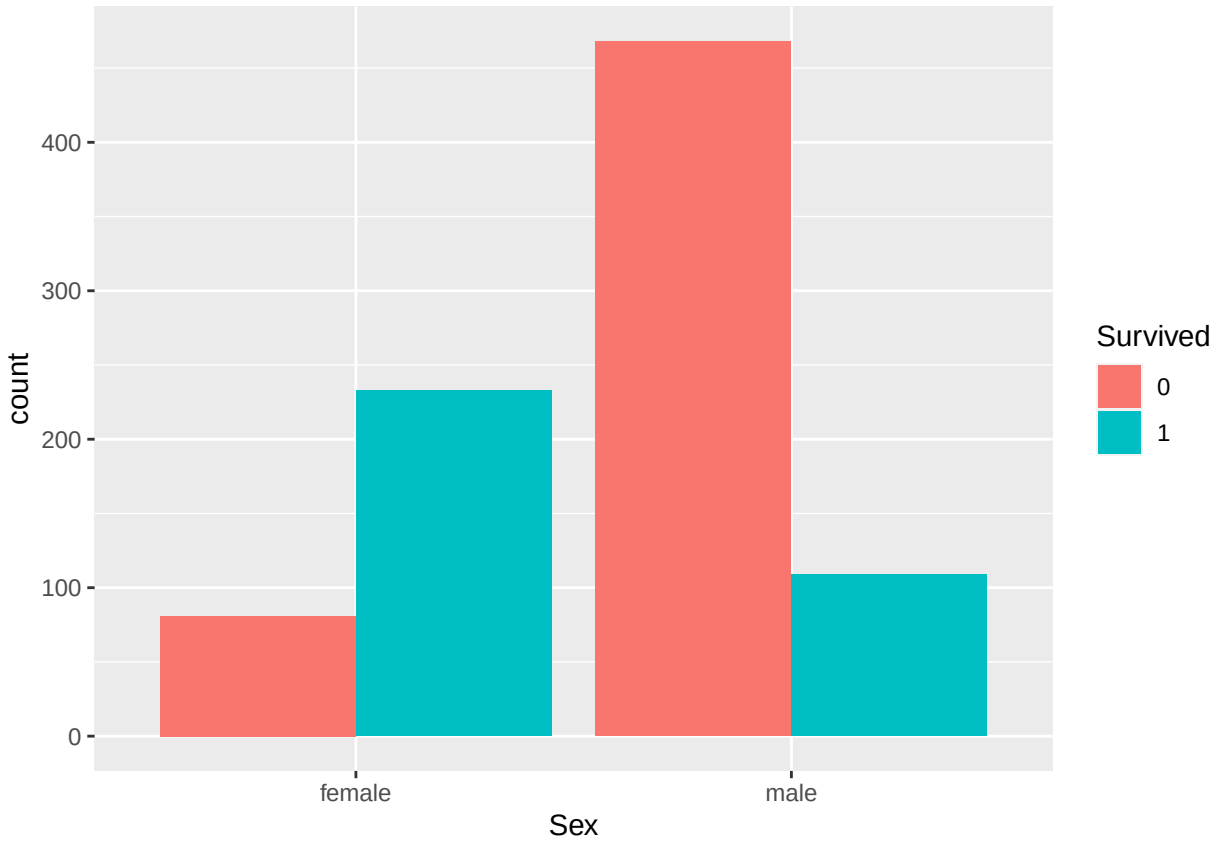
Para las variables Survived, Sex, Pclass, Embarked y FamilyGroup como son factores creo un gráfico de barras.

- En el gráfico de Survived la mayor concentración está en la categoría 0 (no sobrevivieron)
- En el gráfico de Sex se observa que el género se concentra mayormente en hombres.
- En el gráfico de Pclass se evidencia que la clase de boleto se representa significativamente en la clase 3.
- En el gráfico de Embarked se evidencia que el puerto de embarque con mayor registro fue en S=Southampton.
- En el gráfico de FamilyGroup se denota que la mayor concentración está en pasajeros a bordo sin familiares.

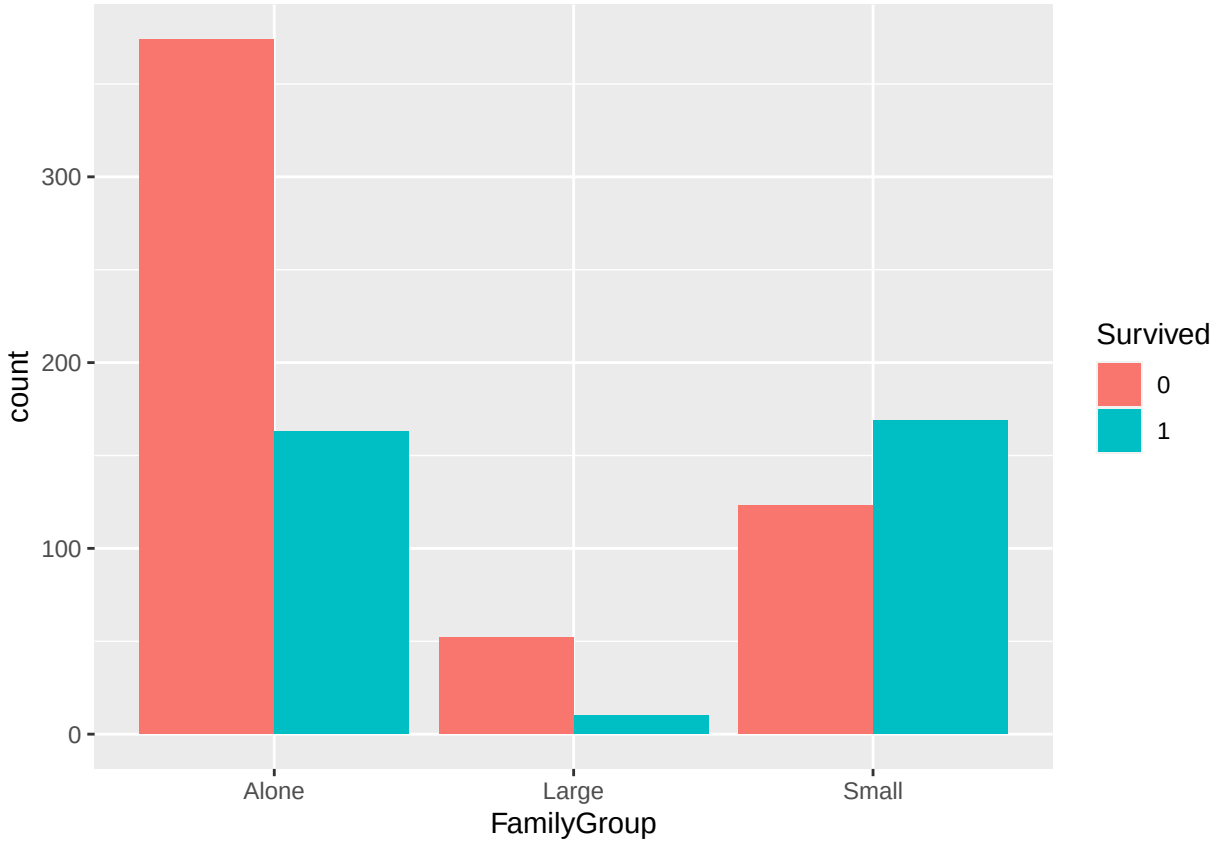
```
par(mfrow=c(2, 3))

Survived1 <- table(new.titanic.train$Survived);
barplot(Survived1, main="Survived", ylab="Frecuencia")

Sex1 <- table(new.titanic.train$Sex);
barplot(Sex1, main="Sex", ylab="Frecuencia")

Pclass1 <- table(new.titanic.train$Pclass);
barplot(Pclass1, main="Pclass", ylab="Frecuencia")

Embarked1 <- table(new.titanic.train$Embarked);
barplot(Embarked1, main="Embarked", ylab="Frecuencia")

FamilyGroup1 <- table(new.titanic.train$FamilyGroup);
barplot(FamilyGroup1, main="FamilyGroup", ylab="Frecuencia")

par(mfrow=c(1, 1))
```



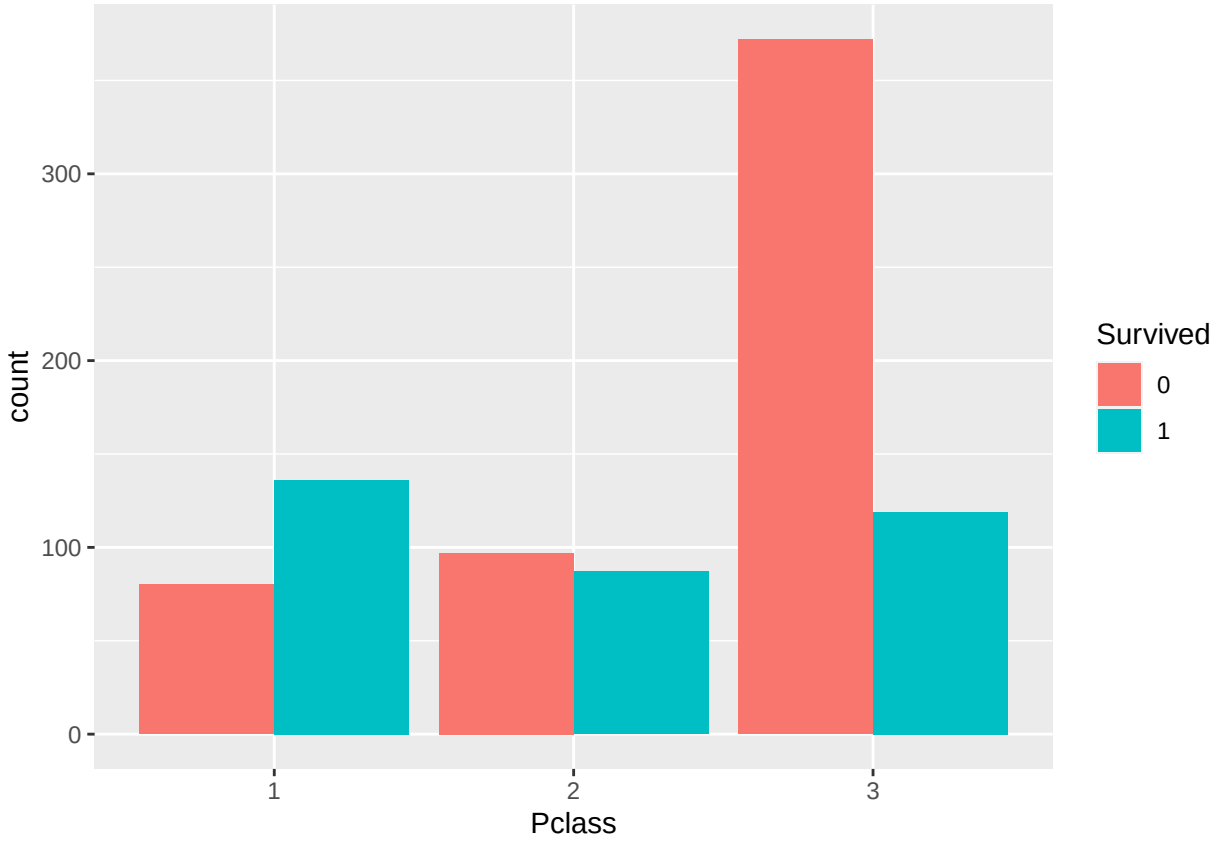
Mediante el siguiente gráfico evaluamos por sexo quiénes sobrevivieron y se denota que son las mujeres.



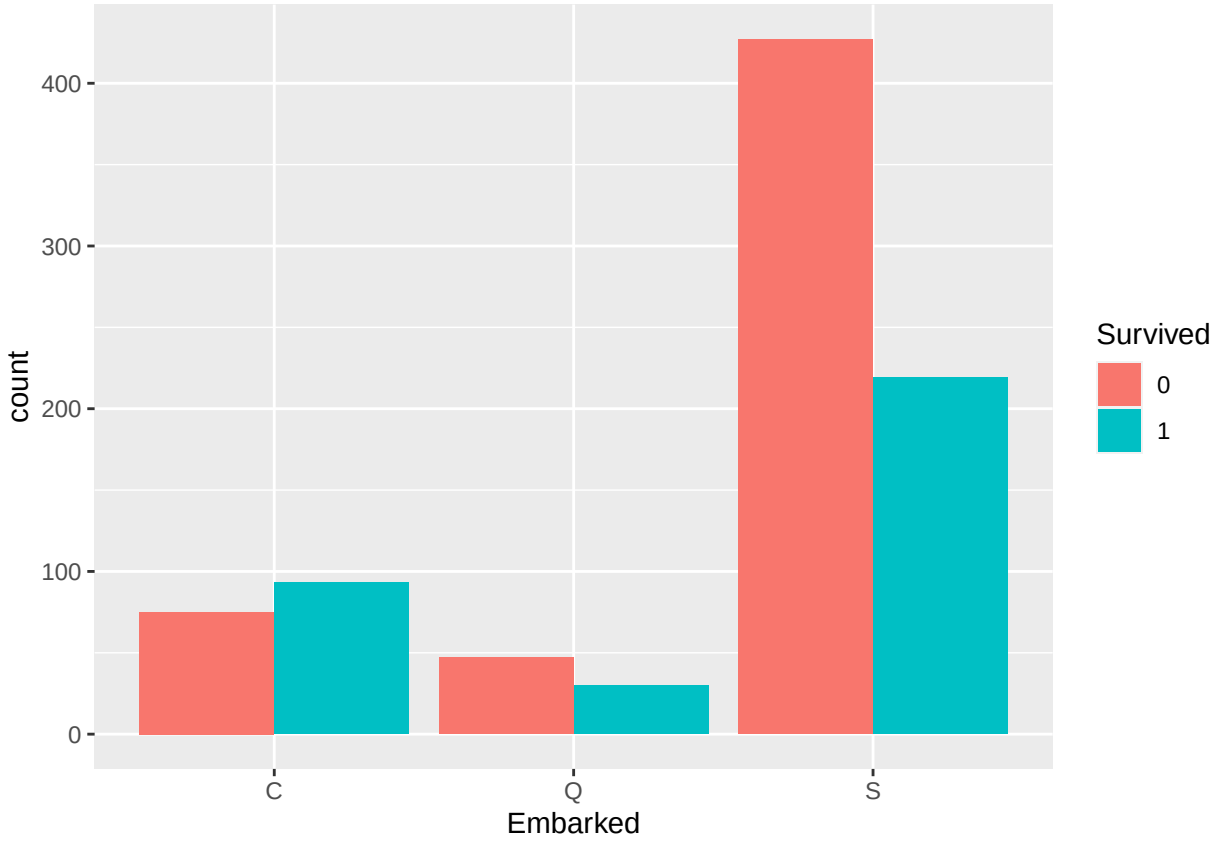
Los pasajeros sin familiares a bordo del barco tienen mayor probabilidad de supervivencia.



Si analizamos la supervivencia en razón de la clase de boleto del pasajero focalizamos que en Pclass = 1 hay mayor número de sobrevivientes.



La supervivencia según el puerto de embarque es mayor en Southampton (S).



El objetivo del análisis es determinar las relaciones enre las variables disponibles y la supervivencia de los pasajeros, entonces se pasará a construir el modelo de predicción.

Se deben cargar las librerías necesarias para el modelo de predicción Random Forest.

```
## Loading required package: lattice
```

```
## Warning: package 'randomForest' was built under R version 4.0.5
```

```
## randomForest 4.6-14
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

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

```
## The following object is masked from 'package:ggplot2':
##
##     margin
```

Luego dividimos más los datos en conjuntos de entrenamiento (70%) y de prueba (30%) de la siguiente manera.

```
set.seed(123)
titanic.data.partition <- createDataPartition(new.titanic.train$Survived,
 p = 0.7, list = FALSE)
```

```
training <- new.titanic.train[titanic.data.partition,]
testing <- new.titanic.train[-titanic.data.partition,]
```

Se construye el modelo para mostrar su precisión.

```
set.seed(123)
random.forest.model <- randomForest(Survived ~ Pclass + Sex + AgeForGroup + Fare + Embarked +
  Designation + FamilyGroup, data = training)
random.forest.predict <- predict(random.forest.model, newdata = testing)
confusion.matrix <- confusionMatrix(random.forest.predict, testing$Survived)
confusion.matrix$table
```

```
##           Reference
## Prediction   0   1
##          0 148  36
##          1  16  66
```

```
confusion.matrix$overall[1]
```

```
##  Accuracy
## 0.8045113
```

El modelo Random Forest devuelve una precisión del 80.45%.

```
prediction <- predict(random.forest.model, newdata = new.titanic.test)
```

Regresión Logística ————————

La variable Survived es una variable dicotómica, que toma el valor 0 cuando el pasajero no sobrevive y 1 cuando el pasajero sobrevive.

Estimo un modelo de regresión logística con la variable dependiente Survived y los regresores Sex, Pclass, Embarked, AgeForGroup y FamilySize.

```
titanic_rl <- titanic.data[,c("Survived","Sex","Pclass","Embarked","AgeForGroup","FamilyGroup")]
```

Genero el modelo predictivo.

```
modelo_log_m <- glm(formula = Survived ~ ., data = titanic_rl,
               family = binomial(link = "logit"))

summary(modelo_log_m)
```

```
##
## Call:
## glm(formula = Survived ~ ., family = binomial(link = "logit"),
##     data = titanic_rl)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6894  -0.6495  -0.4008   0.5894   2.6756
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      2.50811    0.30646   8.184 2.75e-16 ***
## Sexmale         -2.72854    0.20399 -13.376  < 2e-16 ***
## Pclass2         -0.89461    0.26276  -3.405 0.000662 ***
## Pclass3         -1.92688    0.24046  -8.013 1.12e-15 ***
## EmbarkedQ        0.09464    0.38958   0.243 0.808065
```

```
## EmbarkedS        -0.33392    0.24043  -1.389 0.164888
## AgeForGroupChild  1.21413    0.29769   4.079 4.53e-05 ***
## FamilyGroupLarge -2.28409    0.45439  -5.027 4.99e-07 ***
## FamilyGroupSmall  0.20075    0.20251   0.991 0.321532
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1186.66  on 890  degrees of freedom
## Residual deviance:  774.28  on 882  degrees of freedom
##   (418 observations deleted due to missingness)
## AIC: 792.28
##
## Number of Fisher Scoring iterations: 5
```

Con el modelo creado, determinamos:

- Todos los predictores son significativos excepto con la variable Embarked y con el nivel Small de la variable FamilyGroup
- El regresor que tiene mayor impacto en la predicción de supervivencia es Sex porque p-value $<$ 2e-16 siendo menor al nivel de significancia y menor a todos los demás predictores.

Visualizamos las variables más relevantes. En total 4 variables: Sex (con nivel Male), Pclass (2), AgeForGroup (Child) y FamilyGroup (Large).

```
sig_var<- summary(modelo_log_m)$coeff[-1,4] <0.01
names(sig_var)[sig_var == TRUE]
```

```
## [1] "Sexmale"          "Pclass2"          "Pclass3"          "AgeForGroupChild"
## [5] "FamilyGroupLarge"
```

Creamos un tamaño de muestra del 70%.

```
#Creo datos de entrenamiento
input_surv <- titanic_rl[which(titanic_rl$Survived == 1), ]
input_nsurv <- titanic_rl[which(titanic_rl$Survived == 0), ]

set.seed(100) #repetitividad de muestras
input_surv_training_rows <- sample(1:nrow(input_surv), 0.7*nrow(input_surv))
input_nsurv_training_rows <- sample(1:nrow(input_nsurv), 0.7*nrow(input_nsurv))

# Creación de muestras de entrenamiento
training_surv <- input_surv[input_surv_training_rows, ]
training_nsurv <- input_nsurv[input_nsurv_training_rows, ]
trainingData <- rbind(training_surv, training_nsurv) #Unión de muestras de entrenamiento

# Creación de muestra de validación
test_surv <- input_surv[-input_surv_training_rows, ]
test_nsurv <- input_nsurv[-input_nsurv_training_rows, ]
testData <- rbind(test_surv, test_nsurv) #unión de muestras de prueba
```

Se construye modelo

```
logitMod <- glm(Survived ~ . , data=trainingData, family=binomial(link="logit"))
summary(logitMod)
```

```
##
```

```
## Call:
## glm(formula = Survived ~ ., family = binomial(link = "logit"),
##     data = trainingData)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5983  -0.6246  -0.4255   0.5591   2.2123
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)         2.5375     0.3693   6.871 6.39e-12 ***
## Sexmale            -2.8386     0.2459 -11.544  < 2e-16 ***
## Pclass2            -0.8537     0.3267  -2.613  0.00898 **
## Pclass3            -1.6751     0.2860  -5.857 4.72e-09 ***
## EmbarkedQ          -0.1568     0.4708  -0.333  0.73908
## EmbarkedS          -0.3804     0.2957  -1.286  0.19829
## AgeForGroupChild    1.0714     0.3712   2.886  0.00390 **
## FamilyGroupLarge   -2.5365     0.5501  -4.611 4.01e-06 ***
## FamilyGroupSmall    0.1125     0.2464   0.457  0.64802
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 829.60  on 622  degrees of freedom
## Residual deviance: 538.25  on 614  degrees of freedom
## AIC: 556.25
##
## Number of Fisher Scoring iterations: 5
```

Predicción de las probabilidades de la variable Survived.

```
predicted <- predict(logitMod, testData, type="response")
predicted <- ifelse(predicted > 0.5,1,0)
```

Ahora se crea una matriz de confusión para mostrar la tasa de éxito de la predicción del modelo en el conjunto de datos de validación.

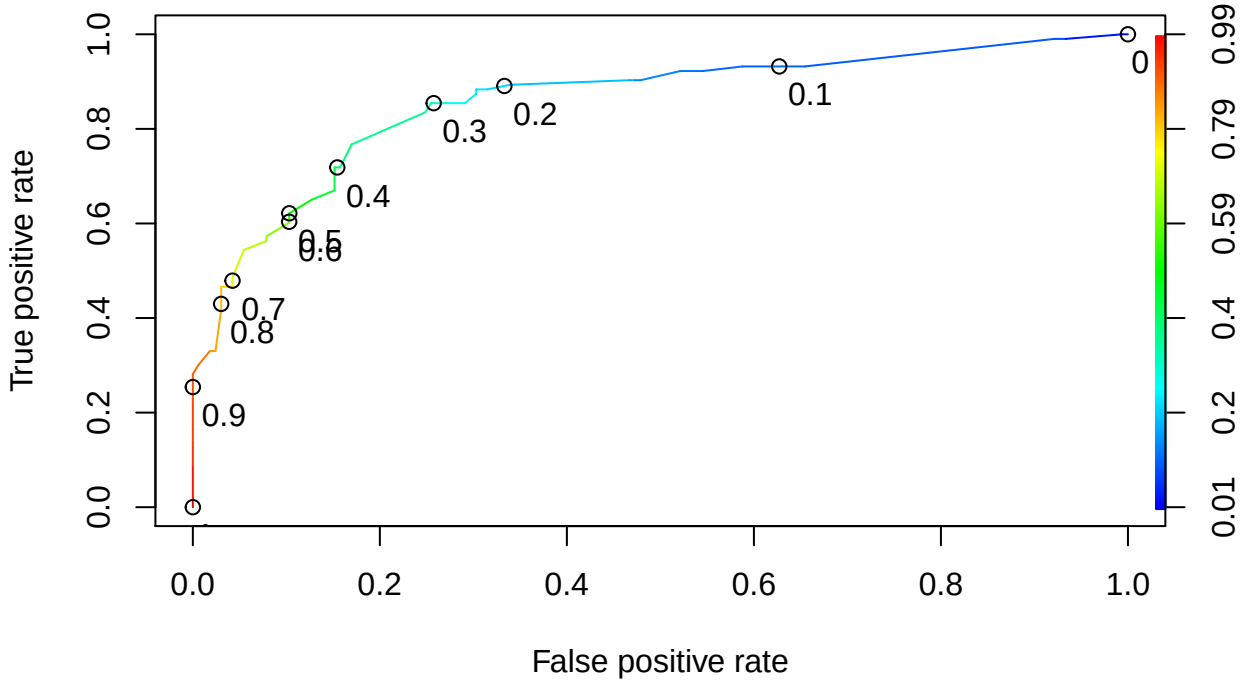
```
library(caret)
confusionMatrix(data=factor(predicted), reference=factor(testData$Survived), positive="1")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 148  39
##          1  17  64
##
##                Accuracy : 0.791
##                  95% CI : (0.7374, 0.8381)
##     No Information Rate : 0.6157
##     P-Value [Acc > NIR] : 5.667e-10
##
##                   Kappa : 0.54
##
##  Mcnemar's Test P-Value : 0.005012
```

```
##
##             Sensitivity : 0.6214
##             Specificity : 0.8970
##          Pos Pred Value : 0.7901
##          Neg Pred Value : 0.7914
##              Prevalence : 0.3843
##          Detection Rate : 0.2388
##    Detection Prevalence : 0.3022
##       Balanced Accuracy : 0.7592
##
##        'Positive' Class : 1
##
```

Graficamos curva ROC (receiver operating characteristic)) para representar la sensibilidad en funciónn de 1 – especificidad. En este caso, el TPR (Verdaderos positivos, eje Y) aumenta más rápido que el FPR (Falsos positivos, eje X) a medida que disminuye la puntación de corte, es decir, existe una muy buena capacidad de predicción del modelo.

```
library(ROCR)
predictions <- predict(logitMod, newdata=testData, type="response")
ROCRpred <- prediction(predictions, testData$Survived)
ROCRperf <- performance(ROCRpred, measure = "tpr", x.measure = "fpr")

plot(ROCRperf, colorize = TRUE, text.adj = c(-0.2,1.7), print.cutoffs.at = seq(0,1,0.1))
```



Respecto al modelo empleado se obtienen las siguientes conclusiones:

- Correcta predicción de 148 personas que sobrevivieron.

- Sensibilidad en la clasificacion de los datos del 62.14%, clasificando los datos de una muy buena manera.
- El número de predicciones correctas son 0 y 0, 1 y 1 (148 + 64 = 212) obteniendo un porcentaje de certeza del modelo de 212 / 268 = 79.10%

6 Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?

- El atributo "PassengerId" es un código único que permitió la integración vertical de los conjuntos de entrenamiento y prueba descargados desde kaggle para conformar una cantidad mayor de registros y analizarlas globalmente.
- Después de comprobar la normalidad de las variables Age y Fare, el análisis de igualdad de varianzas me ha permitido determinar que las muestras no tienen varianzas homogéneas por lo que no puedo aplicar una prueba de hipótesis para estas variables numéricas. Sin embargo, como estrategia a nuestro análisis se emplearon atributos con enfoque similar, así por ejemplo, el atributo AgeForGroup que contiene la clasificación de las personas según su edad y Pclass que contiene la clase de boleto del pasajero.
- Los factores más influyentes en la supervivencia de un pasajero individual son: Sex, PClass, AgeForGroup y FamilySize.
- Los resultados permiten concluir que:
- Las mujeres tienen más probabilidades de supervivencia que los hombres.
- Los pasajeros sin familiares a bordo del barco tienen mayor probabilidad de supervivencia.
- La supervivencia en razón de la clase de boleto del pasajero es mayor en Pclass = 1s.
- Los pasajeros sobrevivientes tuvieron punto de embarque en Southampton (S)

Tips de donde obtener información: – file:///C:/Users/marth/Documents/Maestria%20Ciencia%20de%20Datos/Mineria%20de%20Datos/Asignatura/PEC1/75.584-PEC1.html

https://rpubs.com/camilamila/limpieza https://rstudio-pubs-static.s3.amazonaws.com/421800_30e830cbb8414b6ea8854dd0be118d22.html https://towardsdatascience.com/predicting-the-survival-of-titanic-passengers-30870ccc7e8 https://www.kaggle.com/c/titanic/discussion/28323 https://rstudio-pubs-static.s3.amazonaws.com/555316_3b00cf8efc4c47f4adbc95a4e1f4f1ba.html

https://www.cienciadedatos.net/documentos/41_machine_learning_con_r_y_caret https://rstudio-pubs-static.s3.amazonaws.com/400472_b5699800dc8748608bdef8e555482eaf.html https://jkarakas.github.io/Exploratory-Analysis-of-the-Titanic-Dataset/Titanic_Dataset_Exploratory_Analysis_No_Code.html https://rstudio-pubs-static.s3.amazonaws.com/400472_b5699800dc8748608bdef8e555482eaf.html

https://github.com/emmuzoo/titanicCleaningData/blob/master/muzo_limpieza.Rmd