Tipología y ciclo de vida de los datos: PRAC2

Daniel Gerald Orbegoso Barrantes

junio 2021

Contents

Introducción

Presentación

Vamos a realizar un caso práctico mediante el cual vamos a identificar los datos relevantes para un proyecto analítico. Para ello usaremos herramientas de integración, limpieza, validación y análisis de las mismas.

Competencias

En esta práctica se desarrollan las siguientes competencias del Máster de Data Science:

* Capacidad de analizar un problema en el nivel de abstracción adecuado a cada situación y aplicar las habilidades y conocimientos adquiridos para abordarlo y resolverlo.

* Capacidad para aplicar las técnicas específicas de tratamiento de dato s (integración, transformación, limpieza y validación) para su posteri or análisis.

Objetivos

Los objetivos concretos de esta práctica son:

* Aprender a aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios o multidisciplinares.

* Saber identificar los datos relevantes y los tratamientos necesarios (integración, limpieza y validación) para llevar a cabo un proyect o analítico.

* Aprender a analizar los datos adecuadamente para abordar la información contenida en los datos.

* Identificar la mejor representación de los resultados para aportar conclusiones sobre el problema planteado en el proceso analítico.

* Actuar con los principios éticos y legales relacionados con la manipulación de datos en función del ámbito de aplicación.

* Desarrollar las habilidades de aprendizaje que les permitan continuar estudiando de un modo que tendrá que ser en gran medida autodirigido o autónomo.

* Desarrollar la capacidad de búsqueda, gestión y uso de información y recursos en el ámbito de la ciencia de datos.

Descripción de la PRA a realizar

El objetivo de esta actividad será el tratamiento de un dataset, que puede ser el creado en la práctica 1 o bien cualquier dataset libre disponible en Kaggle (https://www.kaggle.com). Algunos ejemplos de dataset con los que podéis trabajar son:

* Red Wine Quality (https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009)
* Titanic: Machine Learning from Disaster (https://www.kaggle.com/c/titanic)

El último ejemplo corresponde a una competición activa de Kaggle de manera que, opcionalmente, podéis aprovechar el trabajo realizado durante la práctica para entrar en esta competición.

Siguiendo las principales etapas de un proyecto analítico, las diferentes tareas a realizar (y justificar) son las siguientes:

1. Descripción del dataset. ¿Por qué es importante y qué pregunta/proble
ma pretende responder?
2. Integración y selección de los datos de interés a analizar.
3. Limpieza de los datos.
3.1. ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías ca
da uno de estos casos?
3.2. Identificación y tratamiento de valores extremos.
4. Análisis de los datos.
4.1. Selección de los grupos de datos que se quieren analizar/comparar
(planificación de los análisis a aplicar).
4.2. Comprobación de la normalidad y homogeneidad de la varianza.
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.
5. Representación de los resultados a partir de tablas y gráficas.
6. Resolución del problema. A partir de los resultados obtenidos, ¿cuále
s son las conclusiones? ¿Los resultados permiten responder al problema?
7. Código: Hay que adjuntar el código, preferiblemente en R, con el que s
e ha realizado la limpieza, análisis y representación de los datos. Si l
o preferís, también podéis trabajar en Python.

Recursos

Los siguientes recursos son de utilidad para la realización de la práctica:

* Calvo M., Subirats L., Pérez D. (2019). Introducción a la limpieza
y análisis de los datos. Editorial UOC.
* Megan Squire (2015). Clean Data. Packt Publishing Ltd.
* Jiawei Han, Micheine Kamber, Jian Pei (2012). Data mining: concepts and
techniques. Morgan Kaufmann.
* Jason W. Osborne (2010). Data Cleaning Basics: Best Practices in Dealing
with Extreme Scores. Newborn and Infant Nursing Reviews; 10 (1): pp. 1527-3369.
* Peter Dalgaard (2008). Introductory statistics with R. Springer Science
& Business Media.
* Wes McKinney (2012). Python for Data Analysis. O'Reilley Media, Inc.
* Tutorial de Github https://guides.github.com/activities/hello-world.

Para la realización de este trabajo vamos a usar el dataset que se nos va a proveer en Kaggle:

* Titanic: Machine Learning from Disaster (https://www.kaggle.com/c/titanic)

En éste vamos a encontrar tres archivos, de los cuales tenemos un set para entrenamiento y otro para test. Usaremos el set de entrenamiento para nuestra práctica.

Procesos iniciales con los datos

Primer contacto con el juego de datos

Instalamos y cargamos las librerías ggplot2 y dplry

```
# https://cran.r-project.org/web/packages/ggplot2/index.html
if (!require('ggplot2')) install.packages('ggplot2'); library('ggplot2')
# https://cran.r-project.org/web/packages/dplyr/index.html
if (!require('dplyr')) install.packages('dplyr'); library('dplyr')
```

Necesitamos isntalar estas librerías para poder imprimir los elementos en pdf.

```
+ install.packages("tinytex")
+ tinytex::install_tinytex()
```

colección de los datos

Vamos a cargar los datos

```
totalData <- read.csv('titanic/train.csv',stringsAsFactors = FALSE)
filas=dim(totalData)[1]
```

Comprensión y exploración del dataset

El dataset usado para esta competición es el famoso dataset del *Titanic*, con la información de los pasajeros que iban a bordo del barco. El Dataset busca responder preguntas del tipo "que tipo de personas tenían mayores probabilidades de sobrevivir?" Este es un dataset muy importante, puesto que ha sido usado para dar los primeros pasos a todos los cientificos de datos en sus primeros análisis. Podemos quizás pensar que las personas con familias eran las que tenían mayores posibilidades de sobrevivir.

Verificamos la estructura del juego de datos principal

```
str(totalData)
```

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

Vemos que tenemos 891 registros que se corresponden a los viajeros y tripulación del Titánic y 12 variables que los caracterizan.

Revisamos la descripción de las variables contenidas en el fichero y si los tipos de variable se corresponde al que hemos cargado:

PassengerId

```
int valor para identificar a los pasajeros.
```

Survived

```
int variables con dos valores (0 y 1) indicando si el pasajero a
sobrevivido al hudimiento
```

Pclass

```
int columna con la información sobre el estado socio-económico del
pasajero -> 1=1st-2=2nd-3=3rd
```

Name

```
string con el nommbre del pasajero.
```

Sex

```
factor con el sexo del pasajero.
```

Age

```
numeric valor con la edad de las personas en el dia que se hundió el barco.
En el caso que la edad sea menor de un anho, será una fracción de 1. En el
caso de que la edad sea estimada, estará en el formato de xx.5
```

Sibsp

```
int el número de hermanos o conyuges en el Titanic. En el caso de hermanos
se tiene en cuenta hermano, hermana, hermanastro o hermanastra. En el caso
de esposos se tiene en cuenta marido o mujer (amantes y novias son ignoradas)
```

Parch

```
int contiene el nombre de padres e hijos en el Titanic.En el caso de padres
se tiene en cuenta madres o padre. En el caso de hijos, se tiene en cuenta
hijo, hija, hijastro,hijastra (algunos ninhos viajaban con su nihera, por
lo que se les da un valor de 0)
```

Ticket

`string valor con el valor del ticket para el pasajero.`

Fare

`numeric valor con el precio que pagó el pasajero.`

Cabin

`string columna con el número de la cabina donde viajaba el pasajero.`

Embarked

`string puerta por la que embarcó el pasajero.`

Para llevar a cabo este estudio vamos a usar algunso de los datos que se nos va a probeer dentro del dataset de entrenamiento. Puede que creemos alguna nueva variable más adelante, en el caso de que sea necesario.

A continuación vamos a sacar estadísticas básicas y después trabajamos los atributos con valores vacíos.

```
summary(totalData)
```

```
##   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
##
```

Para cada valor podemos comprobar los siguiente:

PassengerId

contiene 891 valores.

Survived

podemos comprobar que la media es de 0,38, con una mediana de 0, lo cual nos indica que hubo más muertos que supervivientes.

Pclass

la media se encuentra en 2,3 y la mediana en 3. Podemos extraer de esta información que la gran mayoría de los pasajeros se encontraban entre persona de segunda y tercera clase. Esto lo podemos comprobar más adelante mediante algunas gráficas.

Name

contiene 891 valores.

Sex

contiene 891 valores.

Age

la edad media se encuentra en 29,70 mientras la mediana se encuentra en 28. Esto nos indica que la mayoría de los pasajeros que viajaban eran relativamente jóvenes, teniendo la persona más joven con 0,42 anhos y la mayor con 80.

Sibsp

la media aquí nos indica que una de cada dos personas solía tener hermanos o cónyuges, con un máximo de 8 hermanos o cónyuges. Podemos comprobar que hay un gran número de familias, más que personas sola viajando.

Parch

la media aquí nos indica que una de cada tres persona tenía hijos o padres abordo. Como máximo podemos encontrar alguien con 6 hijos/as.

Ticket

string valor con el valor del ticket para el pasajero.

Fare

numeric valor con el precio que pagó el pasajero.

Cabin

string columna con el número de la cabina donde viajaba el pasajero.

Embarked

string puerta por la que embarcó el pasajero.

Limpieza y preparación de los datos

Vamos a explorar los datos que me faltan. Para ello mostramos las siguientes estadísticas

Estadísticas de valores vacíos o con NA

```
colSums(is.na(totalData))
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0           0           0           0           0         177
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           0           0           0           0           0           0
```

```
colSums(totalData=="")
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0           0           0           0           0          NA
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           0           0           0           0         687           2
```

Podemos comprobar como el campo Age tiene 177 valores na y luego podemos comprobar como cabin o embarked tienen también valores en blanco

Podemos comprobar que el valor Cabin tiene un gran número de missing data. Vamos a comprobar que porcentaje de data falta y vamos a decidir si lo descartamos.

```
colMeans(totalData=="")
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
## 0.000000000 0.000000000 0.000000000 0.000000000 0.000000000          NA
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
## 0.000000000 0.000000000 0.000000000 0.000000000 0.771043771 0.002244669
```

Comprobamos que el 77% de los datos faltan en esta columna, por lo que la vamos a descartar más adelante.

Para "Embarked" puesto que tenemos solo dos elementos que faltan lo vamos a rellena con el elemento más común.

```
tail(names(sort(table(totalData$Embarked))), 1)
```

```
## [1] "S"
```

Como podemos comprobar que el elemento más común es S vamos a rellenarlo con éste elemento.

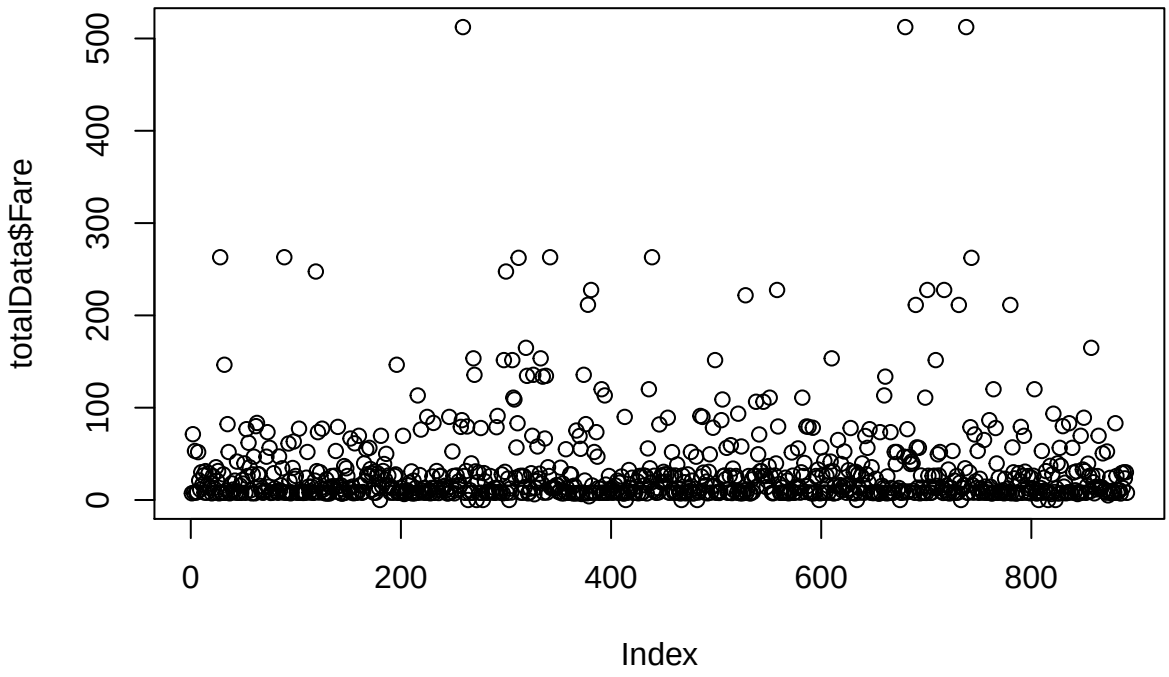
```
totalData$Embarked[is.na(totalData$Embarked)] <- "S"
totalData$Embarked[totalData$Embarked==""] <- "S"
```

Asignamos la media para valores vacíos de la variable "Age"

```
totalData$Age[is.na(totalData$Age)] <- mean(totalData$Age,na.rm=T)
```

Viendo la información que nos dan las estadísticas, podemos comprobar que no tenemos outliers. Quizás en "Fare" por lo que para ello vamos a plasmar los datos, para comprobar si tenemos algún outlier

```
plot(totalData$Fare)
```



Podemos comprobar que hay tres elementos que sobresalen del resto.

```
sd(totalData$Fare)
```

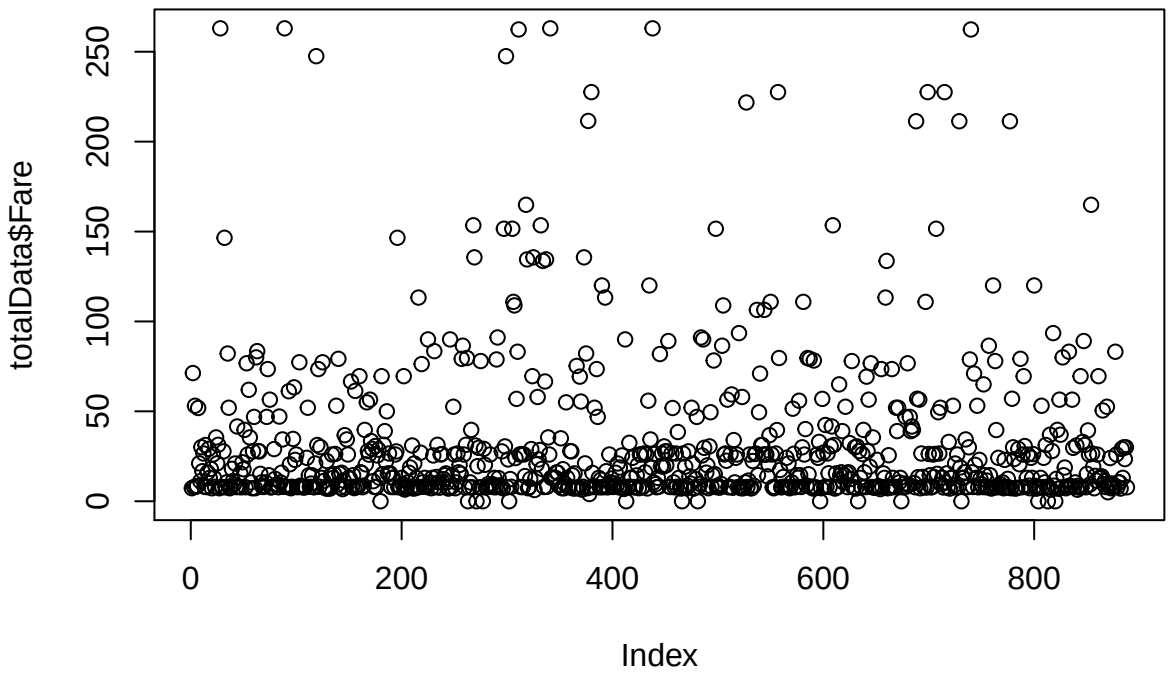
```
## [1] 49.69343
```

Podemos comprobar que debido a la desviación estandar de “49,69”, teniendo en cuenta el tercer percentil que es de “31” y los outliers que son de al rededor de “512”, podriamos descartar estos valores.

```
 totalData <- totalData[totalData$Fare<=500,]
```

Tras esto hemos eliminado las columnas con los outliers y podemos plasmarlo de nuevo para comprobar que es así

```
plot(totalData$Fare)
```



Vamos a comprobar algo que me ha llamado la atención y es la probabilidad de que relacionar la mortalidad del accidente con las personas que viajaban en familia o solas. Puede que la probabilidad de que uno muriese fuera mayor si uno viajaba sólo o en familia? También puesto que cuanto más joven, mejor se pude nadar, quizás sería interesante tener en cuenta esta varibale a la hora de hacer nuestro estudios.

Esta pregunta me ha venido a la cabeza debido a los valores que hemos obtenido de Parch y Sibsp.

Puesto que para nuestro objetivo de estudiar la mortalidad, no necesitamos la columna de Id del pasajero, el Name, el Ticket, ni Cabin. Borraremos esta columnas para no crear ruido.

```
totalData <- subset( totalData, select = -PassengerId )
totalData <- subset( totalData, select = -Name )
totalData <- subset( totalData, select = -Ticket )
totalData <- subset( totalData, select = -Cabin )
summary(totalData)
```

```
##     Survived          Pclass          Sex                 Age
##  Min.   :0.0000   Min.   :1.000   Length:888         Min.   : 0.42
##  1st Qu.:0.0000   1st Qu.:2.000   Class :character   1st Qu.:22.00
##  Median :0.0000   Median :3.000   Mode  :character   Median :29.70
##  Mean   :0.3818   Mean   :2.313                      Mean   :29.68
##  3rd Qu.:1.0000   3rd Qu.:3.000                      3rd Qu.:35.00
##  Max.   :1.0000   Max.   :3.000                      Max.   :80.00
##      SibSp            Parch             Fare          Embarked
##  Min.   :0.0000   Min.   :0.0000   Min.   :  0.000   Length:888
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:  7.896   Class :character
##  Median :0.0000   Median :0.0000   Median : 14.454   Mode  :character
```

```
##  Mean   :0.5248   Mean   :0.3818   Mean   : 30.582
##  3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.: 30.772
##  Max.   :8.0000   Max.   :6.0000   Max.   :263.000
```

Como vamos a modificar éste dataset debido a nuestro estudio, me gustaría crear un dataset con la información en éste punto para poder usarla después en el momento de crear algunos gráficos, puesto que los datos han sido entendidos y limpiados.

```
totalDataOriginal <- totalData
```

Análisis de los datos

Estudio de la normalidad

Para llevar a cabo nuestro estudio he escogido los siguientes datos que quiero analizar = Sex,Age,Survived, SibSp y Parch

He escogido estos datos puesto que pienso que tienen importancia a la hora de poder predecir si una persona va a sobrevivir o no. La pregunta que queriamos responder era si es cierto que las personas con familia tenían mayores posibiliades de sobrevivir o no.

Vamos a estudiar a continuación la normalidad de nuestros datos. Para ello vamos a realizar distintas pruebas de normalidad, para comprobar si los datos se ajustan o no a una normal estandar. Aplicamos Shapiro-Wilk y Kolmogorov-Smirnoff. Antes de estos vamos a instalar la libreria de nortest.

Para llevar a cabo la comprobación de la normalidad, necesitamos que las columnas sobre las cuales vamos a estudiar la normalidad sean de tipo entero o cuantitativas, por lo que probablemente tengamos que hacer conversiones de tipo string/int a numeric.

Vamos a comprobar la normalidad en la columna edad, puesto que va a tener relevancia en nuestro estudio. Empezamos con Shapiro-Wilk, que tiene una sensibilidad muy alta con pequenhas cantidades de datos.

```
shapiro.test(totalData$Age)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  totalData$Age
## W = 0.95851, p-value = 3.678e-15
```

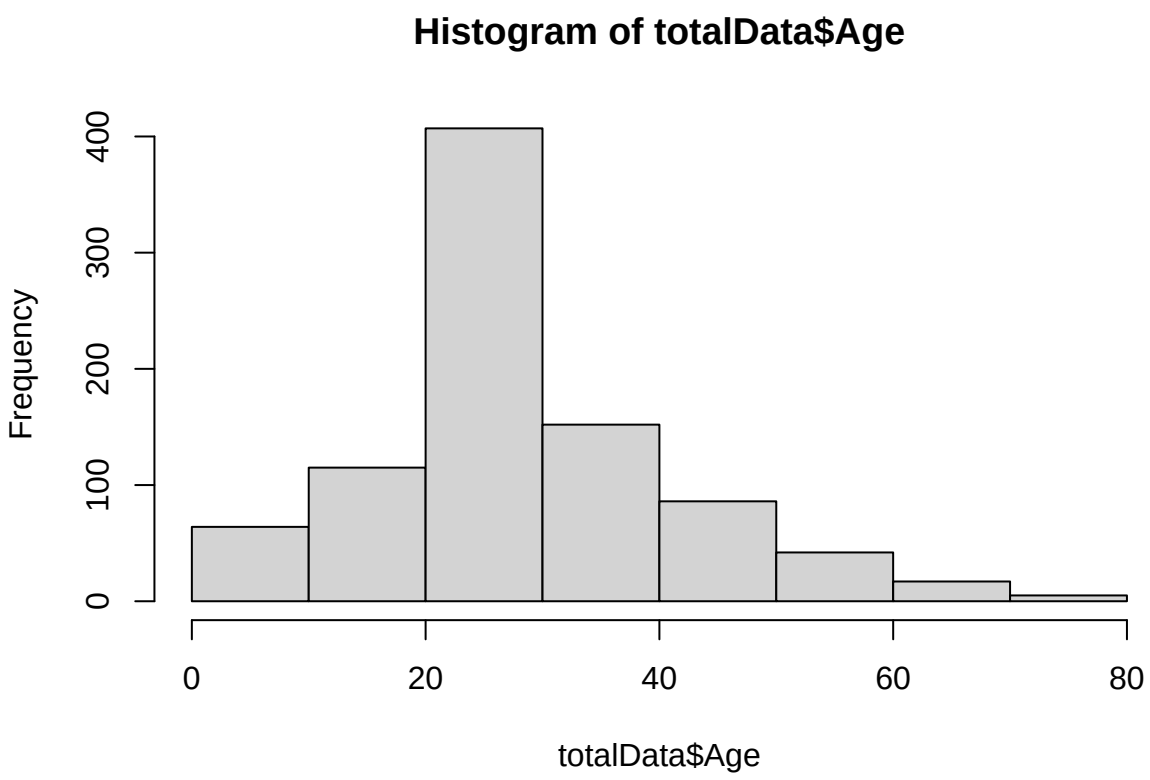
Podemos comprobar que nuestro p valor es extremadamente pequenho, algo que nos indica que estamos alejados de la aceptación de la hipotesis de normalidad para esta variable.

Entonces cuanto más pequenho sea nuestro valor p, más se va a rechazar nuestra hipotesis de normalidad, o sea la hipotesis nula.

Habitualmente si tenemos un valor de $p > 0.05$ es cuando nosotros vamos a poder decir que los valores apoyan la hipótesis de normalidad.

Vamos a revisar el historigrama y podemos comprobar que no se parece a una campana de Gauss.

```
hist(totalData$Age)
```



Vamos a aplicar también Kolgomorov-Smirnoff, que tiene mayor sensibilidad a mayores cantidades de datos para comparar si tenemos resultados similares.

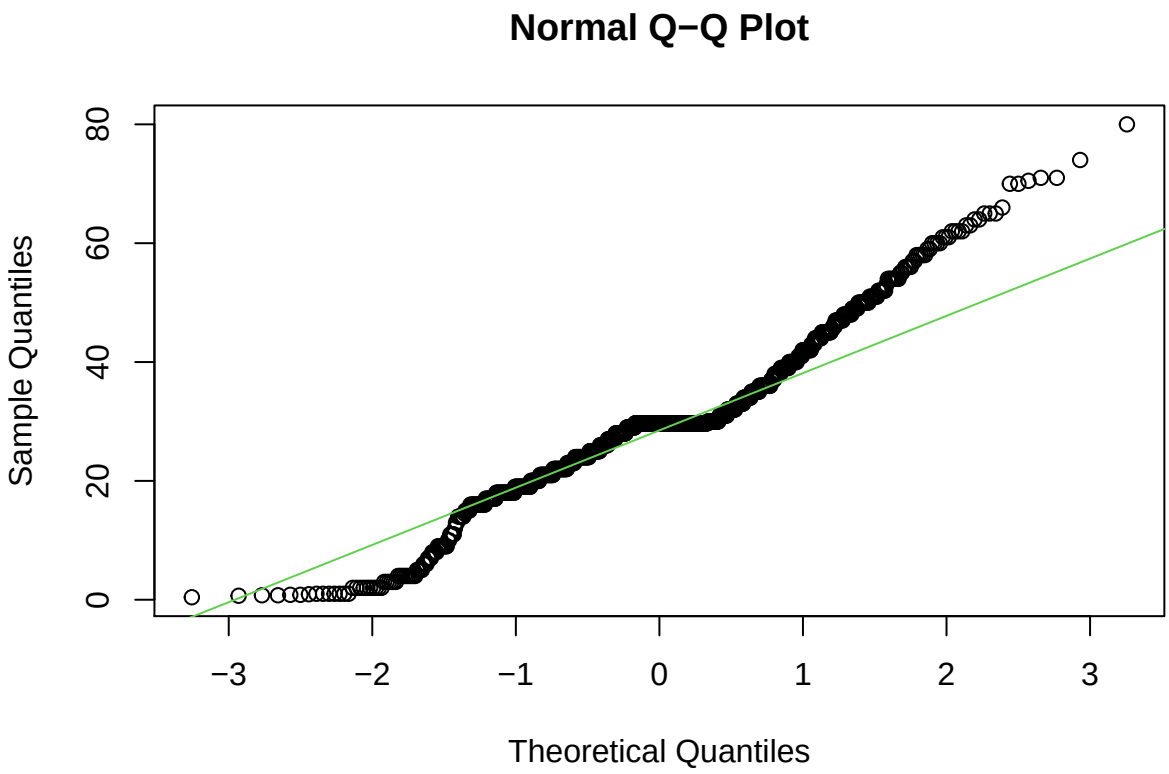
```
library(nortest)
lillie.test(totalData$Age)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  totalData$Age
## D = 0.15011, p-value < 2.2e-16
```

Podemos comprobar aquí que los valores continuan siendo pequenhos y se encuentra muy lejano al valor 1.

Pero para hacer una comprobación más vamos a mirar los quantiles, para ello vamos mirar la gráfica qqnorm. Aquí podremos comparar los datos de la muestra vs los datos teóricos de la muestra normal.

```
qqnorm(totalData$Age)
qqline(totalData$Age,col=3)
```



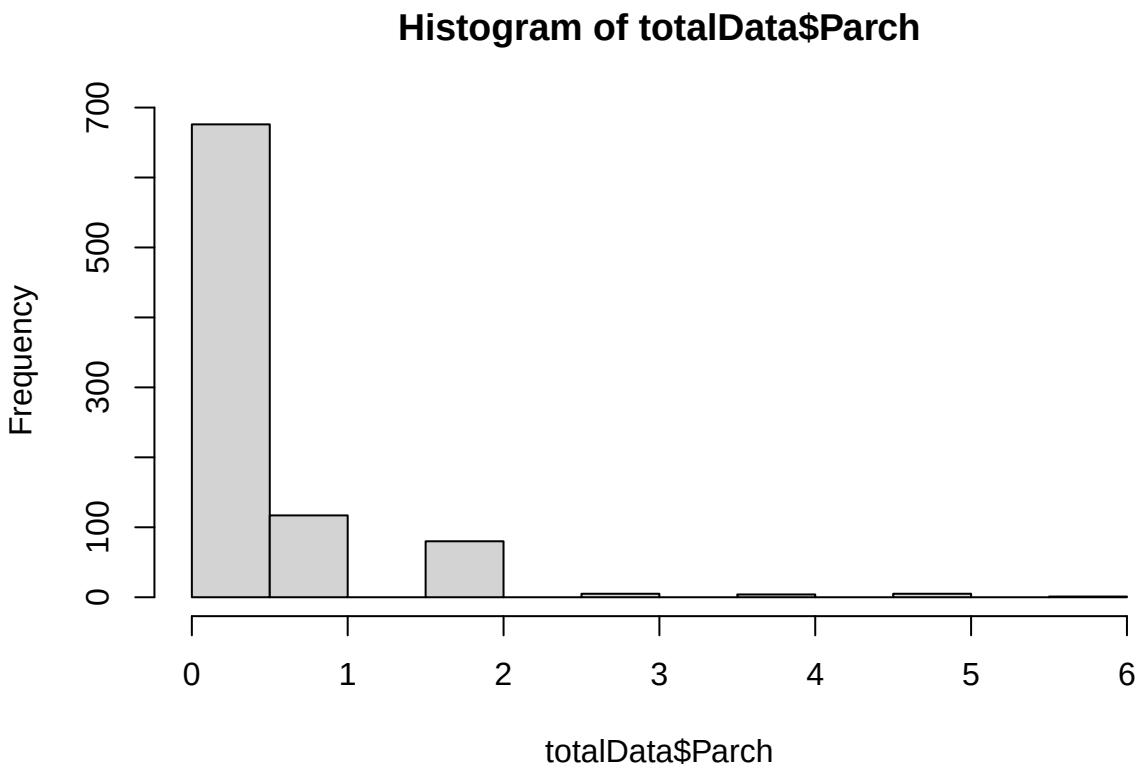
Podemos comprobar que si tuviesemos hipótesis de normalidad debería seguir aquí la linea de normalidad, mostrada en verde. Pero en cambio nuestra gráfica tanto a partir del primer quantil en ambas direcciones muestra desviaciones que se salen de la linea de normalidad.

Vamos a comprobar la normalidad para SibSp, aunque podemos comprobar mediante el histograma que se encuentra muy lejos de la campana de Gauss, por consiguiente va a estar muy lejos de la normalidad. Aun así convertimos la columna a tipo numerico

```
totalData$SibSp <- as.numeric(as.character(totalData$SibSp))
```

Miramos el histograma de la columna

```
hist(totalData$Parch)
```



Aplicamos Shapiro

```
shapiro.test(totalData$Parch)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  totalData$Parch
## W = 0.53254, p-value < 2.2e-16
```

Como deciamos anteriormente, el valor es también muy pequenho, por lo que rechazamos la hipótesis nula. Al igual viendo el histograma podemos comprobar que la figura no es la esperada.

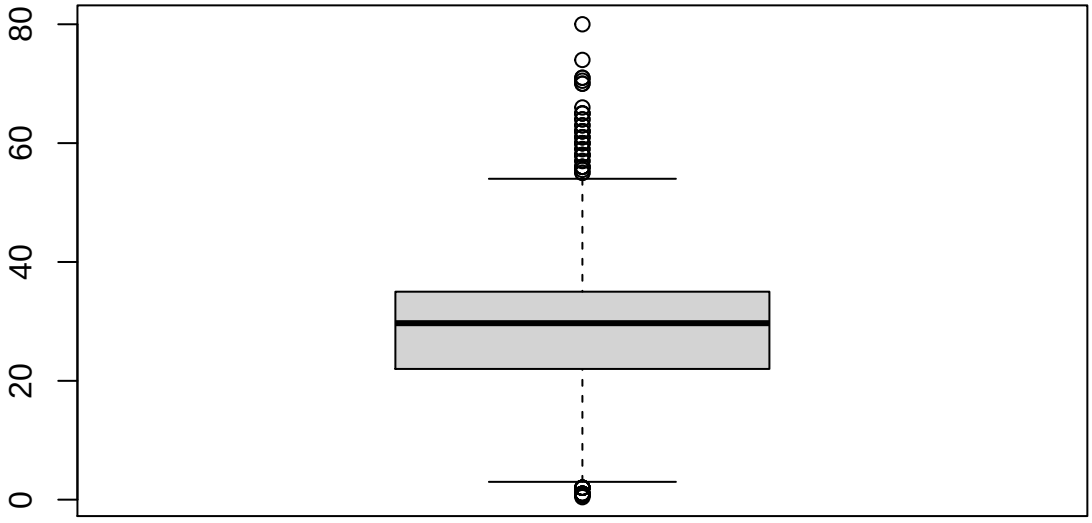
Podriamos aplicar los test de normalidad para el resto de variables, pero puestos que el p-value de Age y SibSp es menor que 0.05 rechazamos la hipótesis nula y determinamos que los datos de esta muestra no tienen una distribución normal.

Estudio de la homogeneidad de la varianza

Vamos a llevar a cabo ahora el estudio de la homogeneidad de la varianza

La hipotesis nula es que las varianzas de las poblaciones son iguales. Los gráficos de cajas y bigotes nos dan una idea de la distribución de los datos.

```
boxplot(totalData$Age)
```



Vamos a aplicar dos modelos aquí por un lado Barlett y Leven. Barlett será aconsejado unsar cuando sabemos que las variables usadas proceden de poblaciones con distribución normal, mientras que por otro lado usando la mediana, podremos usar Leven par distribuciones que no son normales. Vamos a aplicar barlett y comprobaremos con la variable factor sex.

```
bartlett.test(totalData$Age, totalData$Sex)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  totalData$Age and totalData$Sex
## Bartlett's K-squared = 0.044572, df = 1, p-value = 0.8328
```

Podemos comprobar que directamente aquí no se rechaza la hipotesis nula porque el valor es mayor de 0.05, por lo que la homogeneidad se cumple para ambos sexos.

Para hacer una prueba más de varianza vamos a usar esta vez levene, el cual se encuentra en la libreria lawstat. Levene se encuentra relacionada con la mediana.

```
library(lawstat)
levene.test(totalData$Age, totalData$Sex)
```

```
##
```

```
##  Modified robust Brown-Forsythe Levene-type test based on the absolute
##  deviations from the median
##
## data:  totalData$Age
## Test Statistic = 1.0365, p-value = 0.3089
```

Al igual que en el caso anterior podemos comprobar que la distribución que tenemos es mayor de 0.05, aceptando la hipótesis nula, puesto que podemos comprobar que la variación entre estas muestras es constante.

Vamos a hacer la misma prueba ahora pero cogeremos como variable factor Survived y Sibsp.

```
bartlett.test(totalData$Age, totalData$Survived)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  totalData$Age and totalData$Survived
## Bartlett's K-squared = 4.5571, df = 1, p-value = 0.03278
```

```
bartlett.test(totalData$Age, totalData$SibSp)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  totalData$Age and totalData$SibSp
## Bartlett's K-squared = Inf, df = 6, p-value < 2.2e-16
```

Podemos comprobar que en estos casos para ambas variables el valor de p es menor a 0.05, lo que me indica que rechazamos la hipótesis nula. Puesto que con estas dos variables no se cumple la homogeneidad de la varianza, vamos a rechazar la hipotesis nula.

Puesto que tanto la normalidad como la homocedasticidad se va a rechazar, vamos a aplicar pruebas no paramétricas como Wilcoxon para probar que existen diferencias estadísticamente significativas entres los grupos de datos que estamos analizando.

```
wilcox.test(totalData$Age, totalData$SibSp)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  totalData$Age and totalData$SibSp
## W = 784182, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

En este caso podemos comprobar diferencias estadisticamente significativas para la edad en relación al número de parientes con un valor muy pequenho para p.

Correlación

Vamos a estudiar la correlación sobre las variables que hemos escogido, para comprobar la asociación que tenemos entre dos variables. Vamos a usar las variables Age y SibSP, los cuales no cumplen con las distribución de normalidad, por lo que vamos a usar Spearman para analizarlas.

```
cor.test(totalData$Age, totalData$SibSp, method = "spearman",exact=FALSE)
```

```
##
##  Spearman's rank correlation rho
##
## data:  totalData$Age and totalData$SibSp
## S = 133731149, p-value = 1.272e-05
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.1458967
```

Podemos comprobar como el valor nos da negativo indicando que los valores elevados de una varibales se asocian con valores pequenhos de la otra. Esto tiene mucho sentido, puesto que cuanto más mayor sea más probabilidad de tener mas hermanos y esposos/as.

Vamos a comprobar la correlación para otro valor, en este caso Sex. Para ello vamos a cambiar el valor sex a tipo int

```
library(plyr)
```

```
## ------------------------------------------------------------------------------
```

```
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)
```

```
## ------------------------------------------------------------------------------
```

```
##
## Attaching package: 'plyr'
```

```
## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
totalData$Sex = revalue(totalData$Sex,c("male"=0, "female"=1))
totalData$Sex <-as.integer(totalData$Sex)
```

```
cor.test(totalData$Survived, totalData$Sex, method = "spearman",exact=FALSE)
```

```
##
##  Spearman's rank correlation rho
##
## data:  totalData$Survived and totalData$Sex
## S = 52995539, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.5458993
```

En este otro caso podemos comprobar como hay una gran correlación entre el valor sobrevivir y el sexo de la persona. Puesto que en este caso los valores son positivos nos indica que están creciendo simultaneamente.Esto concuerda con los resultados que nos indica que la gran mayoría de personas que sobrevivían eran mujeres

Puesto que una de las preguntas que queriamos responder era si tenía relación el hecho de pertenecer a una familia para tener mayores posibilidades de sobrevivir, podemos responderlo parcialmente a traves de la correlación de dichas variables.

```
cor.test(totalData$Survived, totalData$Parch, method = "spearman",exact=FALSE)
```

```
##
##  Spearman's rank correlation rho
##
## data:  totalData$Survived and totalData$Parch
## S = 100566124, p-value = 3.552e-05
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.1382831
```

```
cor.test(totalData$Survived, totalData$SibSp, method = "spearman",exact=FALSE)
```

```
##
##  Spearman's rank correlation rho
##
## data:  totalData$Survived and totalData$SibSp
## S = 105969518, p-value = 0.006088
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## 0.09198325
```

Podemos comprobar que de una manera también positiva cuando uno tenía familia ligeramente tenía más posibilidades de sobrevivir. (me gustaría remarcar que ligeramente las posibilidades aumentaban, puesto que hay otros factores que tienen más relevancia)

Por consiguiente si eras una mujer y tenías hijos o padres tus posivilidades aumentaban considerablemente.

Regresión

Mediante una regresión vamos a predecir el valor de una variable dependiente en función de un valor conocido de la variable independiente. Va a describirnos como una varibale independiente está relacionada numéricamente con una varibale dependiente.

A continuación vamos a implementar una regresión, para comprobar la relación de dependencia lineal entre una variable dependiente y una de variables independientes.

Mediante la función lm() vamos a implementar la regresión y sta vez vamos a realizar una de tipo simple y daremos como variables Survived es la variable dependiente y Sex la varibale independiente o predictora.

```
ml = lm(Survived ~ Sex , data = totalData)
```

```
summary(ml)
```

```
##
## Call:
## lm(formula = Survived ~ Sex, data = totalData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7412 -0.1861 -0.1861  0.2588  0.8139
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.18609    0.01699   10.95   <2e-16 ***
## Sex          0.55513    0.02862   19.39   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4075 on 886 degrees of freedom
## Multiple R-squared:  0.298,  Adjusted R-squared:  0.2972
## F-statistic: 376.1 on 1 and 886 DF,  p-value: < 2.2e-16
```

Podemos comprobar observando el coeficiente de determinación R-squared como las variables se correlacionan, con un valor de 0.298. Ahora la pregunta es como interpretramos estos valores, y la respuesta como muchas veces es depende. Si seguimos lo que decía Cohen(1992) un valor mayor de 0.26 muestra una gran correlación: Pero esto depende del campo en el cual estamos realizando este estudio. Mientras que en el campo de las ciencias puras necesitamos un valor mayor al 60%, en campos como las artes o humanidades un valor de 10% sería aceptado como adecuado. Por consiguiente el sexo de una persona podía influir bastante a la hora de predecir si alguien va a sobrevivir, probando el hecho de que las mujeres tenían una gran probabilidad de sobrevivir

Vamos comprobar otras variables que creemos podría ayudar a predecir si alguien sobrevivia o no como deciamos en un principio, el hecho de tener familia.

```
ml = lm(Survived ~ SibSp, data = totalData)
```

```
summary(ml)
```

```
##
## Call:
## lm(formula = Survived ~ SibSp, data = totalData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3895 -0.3895 -0.3748  0.6105  0.6693
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.38947    0.01806  21.562   <2e-16 ***
## SibSp       -0.01470    0.01478  -0.995     0.32
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.4861 on 886 degrees of freedom
## Multiple R-squared:  0.001115,   Adjusted R-squared:  -1.221e-05
## F-statistic: 0.9892 on 1 and 886 DF,  p-value: 0.3202
```

```
ml = lm(Survived ~ Parch, data = totalData)
```

```
summary(ml)
```

```
##
## Call:
## lm(formula = Survived ~ Parch, data = totalData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6598 -0.3629 -0.3629  0.6371  0.6371
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.36286    0.01800  20.163   <2e-16 ***
## Parch        0.04949    0.02017   2.454   0.0143 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4847 on 886 degrees of freedom
## Multiple R-squared:  0.00675,    Adjusted R-squared:  0.005629
## F-statistic: 6.021 on 1 and 886 DF,  p-value: 0.01433
```

Podemos comprobar como para nuestra predicción, el hecho de tener influencia no va a influir prácticamente, prácticamente confirmando que nuestra hipotesis en un principio planteada no se cumple, o como informábamos antes, va a afectar ligeramente, pero sólo ligeramente, sin tener gran relevancia.

Voy a introducir en la ecuación una variable que me tiene algo intrigrado y es el hecho de que si se pertenecía a una clase superior, las probabilidades de sobrevivir se incrementaban. Para ello vamos a contar con la clase Pclass.

```
ml = lm(Survived ~ Pclass, data = totalData)
```

```
summary(ml)
```

```
##
## Call:
## lm(formula = Survived ~ Pclass, data = totalData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6374 -0.2480 -0.2480  0.3626  0.7520
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.83213    0.04538   18.34   <2e-16 ***
## Pclass      -0.19471    0.01846  -10.55   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4584 on 886 degrees of freedom
## Multiple R-squared:  0.1116, Adjusted R-squared:  0.1106
## F-statistic: 111.3 on 1 and 886 DF,  p-value: < 2.2e-16
```

Para sorpresa mia, he podido comprobar como la variable Pclass nos da una correlación de un 11%, lo que nos muestra que el hecho de tener una clase superior tenía mucho más peso a la hora de tener probabilidades de sobrevivir que el tener famiia a bordo del barco.

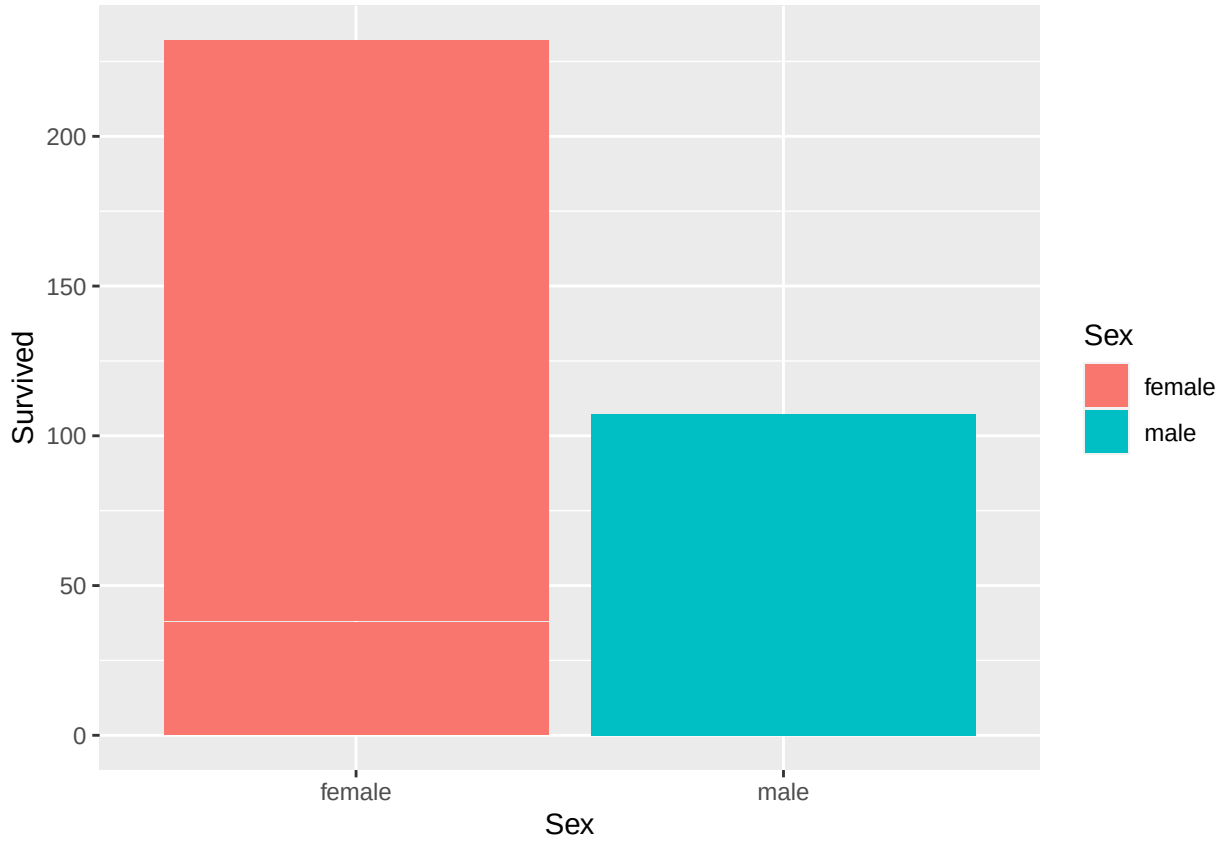
Para finalizar vamos a mostrar algunas de las cosas que hemos ido comprendiendo a lo largo de este proyecto

Representación de resultados

Nos proponemos analizar las relaciones entre las diferentes variables del juego de datos para mostrar como se relacionan

Visualizamos la relación entre las variables "sex" y "survived":

```
library(ggplot2)
ggplot(totalDataOriginal,aes(x=Sex,y=Survived, fill = Sex))+geom_bar(stat="identity")
```



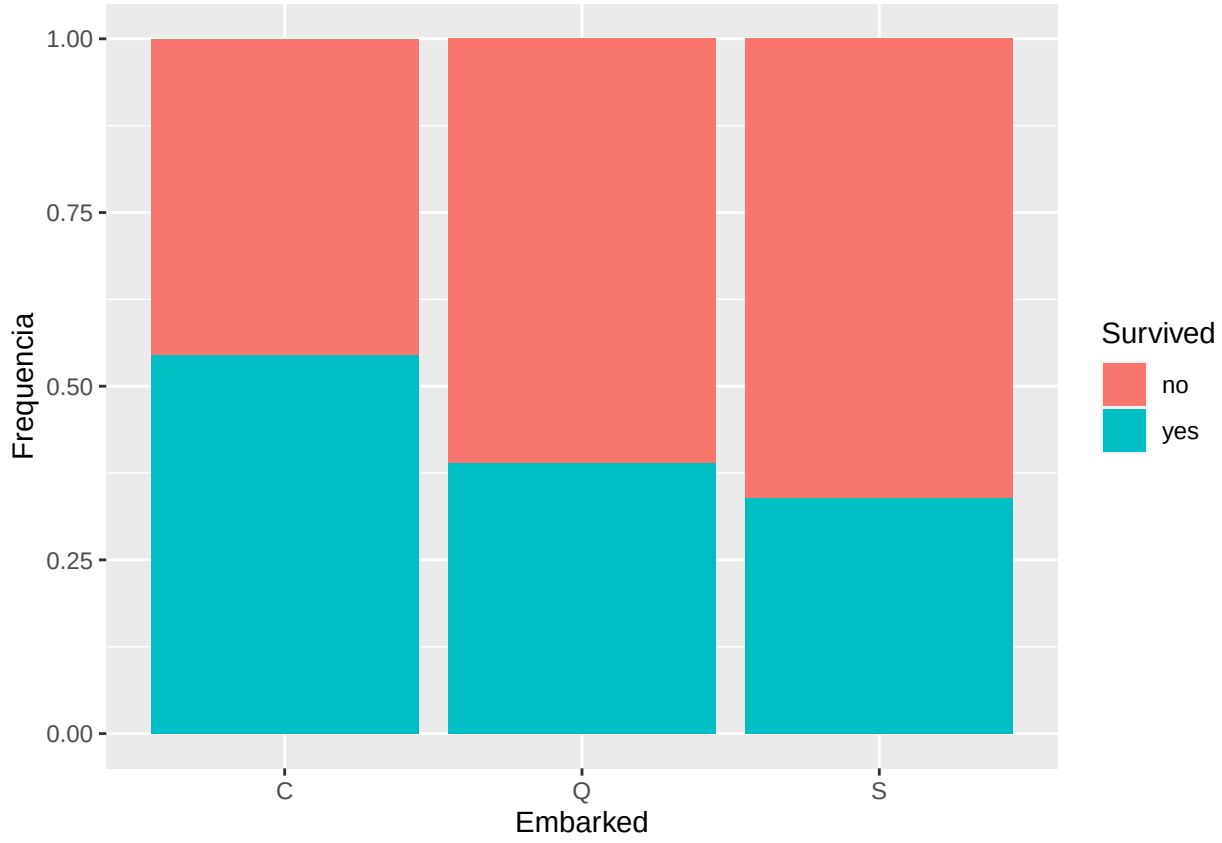
En esta gráfica podemos observar fácilmente separado por sexo la cantidad de hombres y mujeres que sobrevivieron al accidente del Titanic. Podemos comprobar como

Otro punto de vista sería comprobar la gente que sobrevivió dependiendo de donde embarcaron, una variable que hasta este momento no hemos tocado. Para ello necesitamos hacer una conversión en nuestro dataset y cambiar survived por un tipo chr

```
totalDataOriginal$Survived <-as.character(totalDataOriginal$Survived)
totalDataOriginal$Survived = revalue(totalDataOriginal$Survived,c("0"='no', "1"='yes'))
```

Tras esta conversión podemos plasmar la gráfica

```
ggplot(data = totalDataOriginal,aes(x=Embarked,fill=Survived))+geom_bar(position="fill")+ylab("Frequenc
```



Podemo comprobar que la probabilidad de sobrevivir cuando se accedía por la puerta C era mayor que las otras puertas. Esto se puede deber que esta entrada estuviese más frecuentada por mujeres o gente de primera clase o con familiares.

Vamos a ver ahora una matriz de porcentajes de frecuencia.

```
t<-table(totalDataOriginal[1:filas,]$Embarked,totalDataOriginal[1:filas,]$Survived)
for (i in 1:dim(t)[1]){
    t[i,]<-t[i,]/sum(t[i,])*100
    }
```

```
t
```

```
##
```

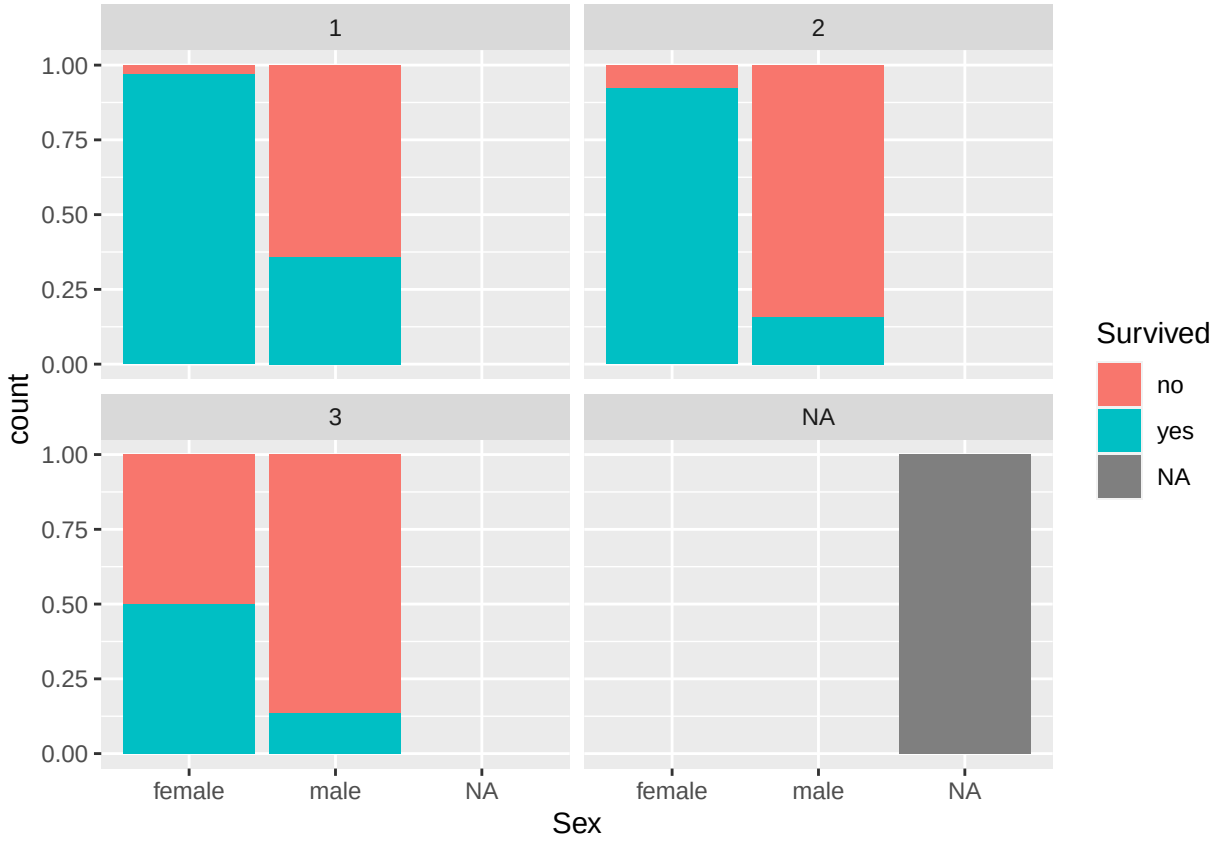
```
##           no      yes
##   C 45.45455 54.54545
##   Q 61.03896 38.96104
##   S 66.09907 33.90093
```

Vemos, por ejemplo, que la probabilidad de sobrevivir si se embarcó en "C" es de un 54%

Podemos poner en un mismo gráfico de frecuencias distintas variables para comprobar la frecuencia de sobrevivir despendiendo de la clase: Embarked, Survived y Pclass.

Mostramos el gráfico de embarcados por Pclass:

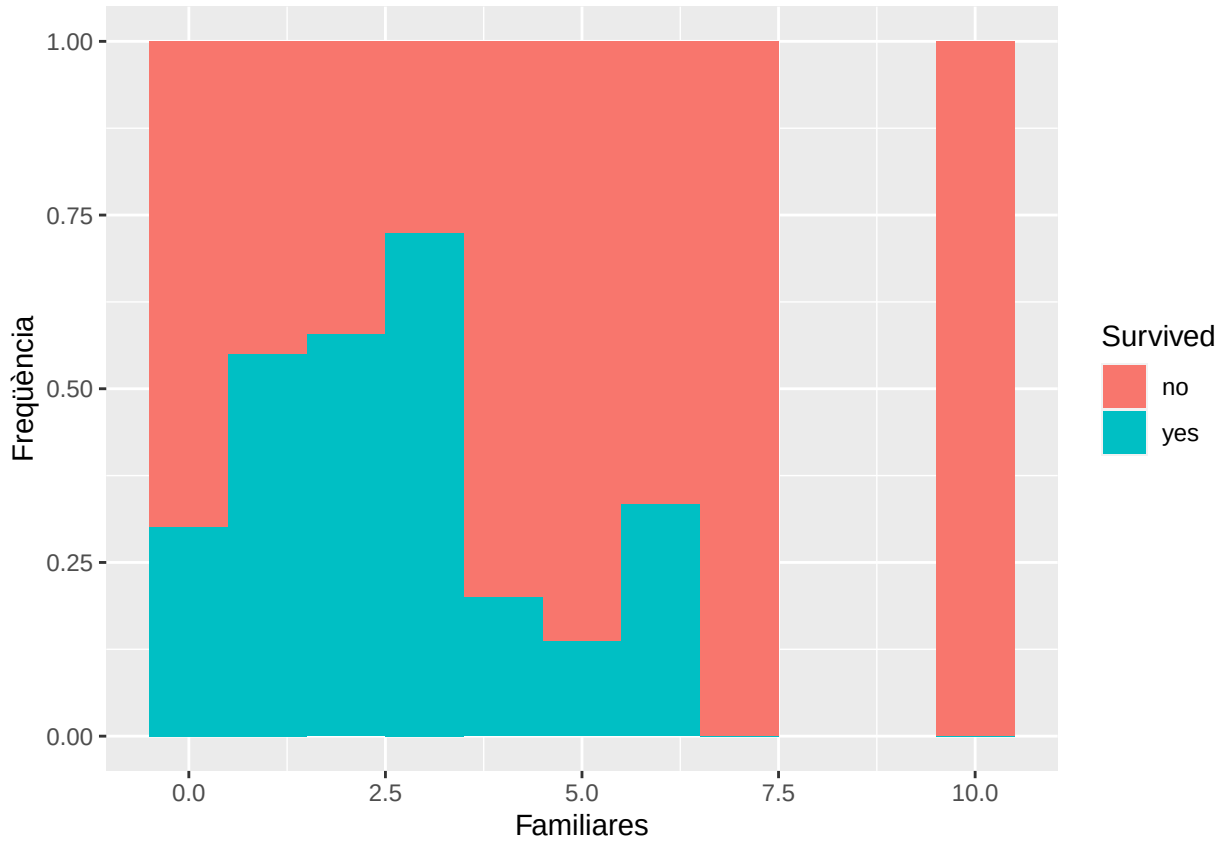
```
ggplot(data = totalDataOriginal[1:filas,],aes(x=Sex,fill=Survived))+geom_bar(position="fill")+facet_wrap
```



Podemos comprobar aquí de manera visual lo que hemos corroborado e ignorabamos en este estudio, y es que la influencia de la clase y el sexo es muy grande y rompe con nuestra hipótesis una vez más, donde nos preguntabamos que si los pasajeros con familia tenían más posibilidades de sobrevivir. Las personas que pertenecían a primera clase tenían grandes posibilidades de sobrevivir, al igual que las de segunda, mientras que las de tercera descendían drásticamente, y en todos los casos podemos ver como aumenta drásticamente si eras una mujer.

Vamos crear una variable nueva, familiares, y con ella vamos a comprobar que probabilidad había de sobrevivir si teniamos más o menos familiares para poder responder de manera visual a la pregunta que planteamos en un principio.

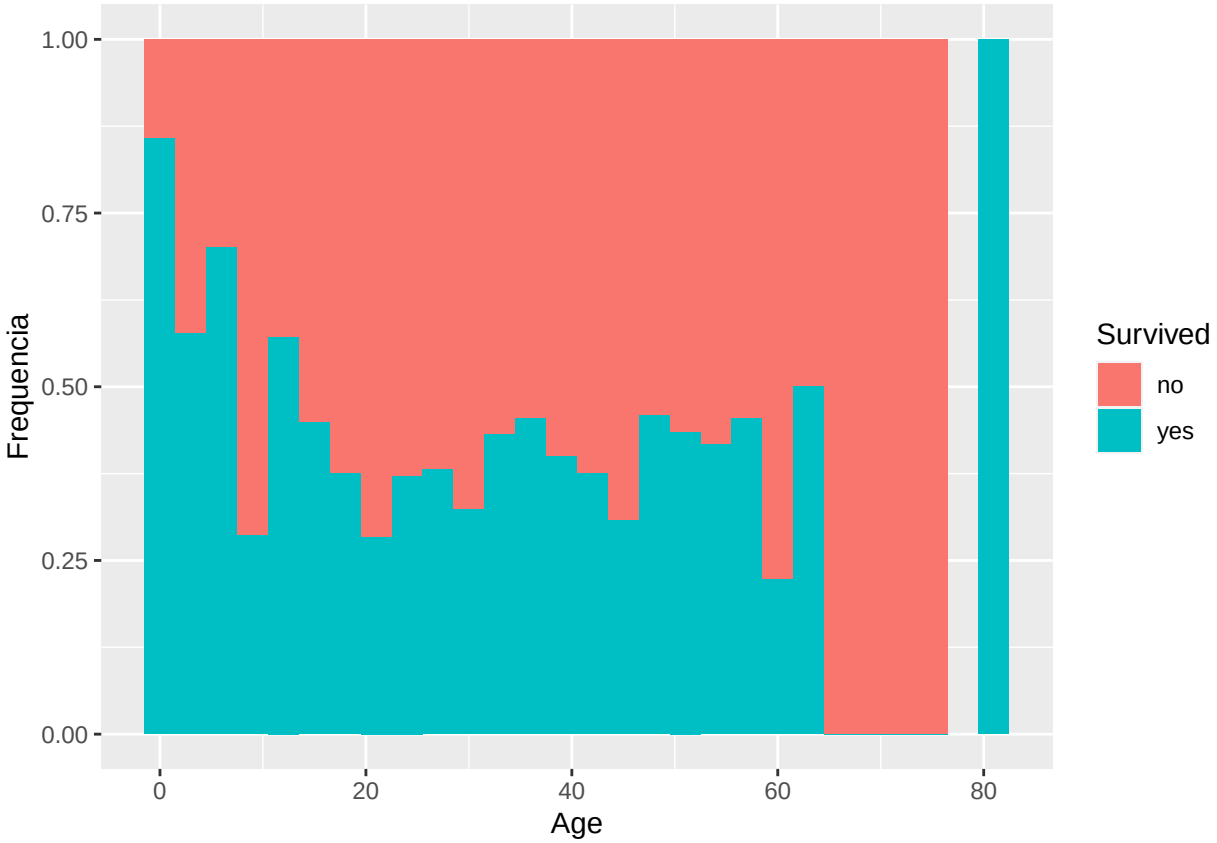
```
totalDataOriginal$Familiares <- totalDataOriginal$SibSp + totalDataOriginal$Parch;
totalData1<-totalDataOriginal[1:filas,]
ggplot(data = totalData1[!is.na(totalDataOriginal[1:filas,]$Familiares),],aes(x=Familiares,fill=Survive
```



Podemos comprobar como el tema de viajar con familiares en sí no incrementaba drásticamente las posibilidades de sobrevivir. Es cierto que se incrementa cuando se tenía entre uno y tres familiares pero el hecho de tener más familiares o viajar solo tienen probabilidades de sobrevivir similares o incluso mejor, para el caso de tener cinco familiares.

Para finializar vamos a mostrar una gráfica en la cual podemos ver las probabiliades de sobrevivir en relación a la edad

```
ggplot(data = totalData1[!is.na(totalDataOriginal[1:filas,]$Age),],aes(x=Age,fill=Survived))+geom_histo
```



Podemos ver que en general la edad no afectaba en si la posibilidad de sobrevivir. Podemos ver alguna excepción para los ninhos o menores de 10 anhos. Por consiguiente rompe nuestra hipótesis que planteabamos al principio, el hecho de que la gente más joven tenía mayores posibilidades de sobrevivir en comparación las más mayor.

Conclusiones finales

Podemos decir que el estudio muestra que los datos en si tienen una calidad adecuada. Tenemos buena información sobre los atributos y está bien documentado. Tenemos una variable "survive" que podemos usar para realizar estudios mediante metodos de clasificiación. Me hubiese gustado aplicar algún metodo de clasificación pero no me ha sido posible debido a la falta de tiempo.

Hemos podido comprobar a traves del estudio que unas de las variables con más peso era Pclass y Sex. El hecho de tener familia no afectaba o influía en las probabilidades de sobrevivir, al igual que tampoco la edad era un factor que influyese de manera importante en la probabilidad de sobrevivir.