Tipología y ciclo de vida de datos: Práctica 2 - Limpieza y análisis

Autor: Sergio Fernández García

Enero 2020

Contents

Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?

Cada registro del dataset contiene información de un pasajero que viajaba en el Titanic.
En las variables se caracteriza si sobrevivió, sexo, edad, en qué categoría viajaba o si viajaba con familiares, etc ...

Las variables son las siguientes:

Table 1: Dataset Variablesariables

Variable	Definition	Keys
PassengerId	Id del registro	
Survived	Supervivencia (si o no)	0 = No, 1 = Si
Pclass	Clase del pasaje	1 = 1st, 2 = 2nd, 3 = 3rd
Name	Nombre de la persona	
Sex	Sexo de la persona	
Age	Edad, en años	
SibSp	Cantidad Hermanos / cónyuges a bordo	
Parch	Cantidad padres o hijos a bordo	
Ticket	Número de ticket	
Fare	Tarifa	
Cabin	Cabina	
Embarked	Lugar de embarque	C=Cherbourg, Q=Queenstown, S=Southampton

El propósito del presente dataset es tratar de construir un modelo de ML capaz de predecir cuales pasajeros sobreviven o mueren, en función de las varaibles descritas.

Integración y selección de los datos de interés a analizar.

```
df<-read.csv("./titanic.csv", header=T, sep=",")
str(df)
```

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

A priori, vemos que hay varias variables que no aportan información útil en cuanto a supervivencia.
Las variables a descartar, serán:

- PassengerId
- Name
- Ticket
- Fare
- Cabin
- Embarked

Por lo tanto, seleccionamos las siguientes variables:

- Pclass
- Sex
- Age
- SibSp
- Parch

Eliminamos del dataframe las variables que no vamos a usar:

```
df$PassengerId <- NULL
df$Name <- NULL
df$Ticket <- NULL
df$Fare <- NULL
df$Cabin <- NULL
df$Embarked <- NULL

summary(df)
```

```
##     Survived          Pclass          Sex                 Age
##  Min.   :0.0000   Min.   :1.000   Length:891         Min.   : 0.42
##  1st Qu.:0.0000   1st Qu.:2.000   Class :character   1st Qu.:20.12
##  Median :0.0000   Median :3.000   Mode  :character   Median :28.00
##  Mean   :0.3838   Mean   :2.309                      Mean   :29.70
##  3rd Qu.:1.0000   3rd Qu.:3.000                      3rd Qu.:38.00
##  Max.   :1.0000   Max.   :3.000                      Max.   :80.00
##                                                      NA's   :177
##      SibSp           Parch
##  Min.   :0.000   Min.   :0.0000
##  1st Qu.:0.000   1st Qu.:0.0000
##  Median :0.000   Median :0.0000
##  Mean   :0.523   Mean   :0.3816
##  3rd Qu.:1.000   3rd Qu.:0.0000
##  Max.   :8.000   Max.   :6.0000
##
```

Limpieza de los datos

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

Hacemos una exploración visual inicial del csv y vemos que los datos nulos simplemente se han dejado vacíos. Analizamos los valores de las distintas varaibles en busca de elementos vacíos.

```
sapply(df, function(x) sum(is.na(x)))
```

```
## Survived   Pclass      Sex      Age    SibSp    Parch
##        0        0        0      177        0        0
```

La variable `Age` va a suponer un problema. De un total de 890 registros, tenemos 177 vacíos. Esto supone casi un 20% de los registros.
Intuimos que se trata de una variable importante, en lo referente a la supervivencia. Pero aliminar casi un 20% de los registros del dataset es demasiado. Voy a optar por rellenar los datos faltantes de la variable `Age` usando K-nearest neighbor (KNN).

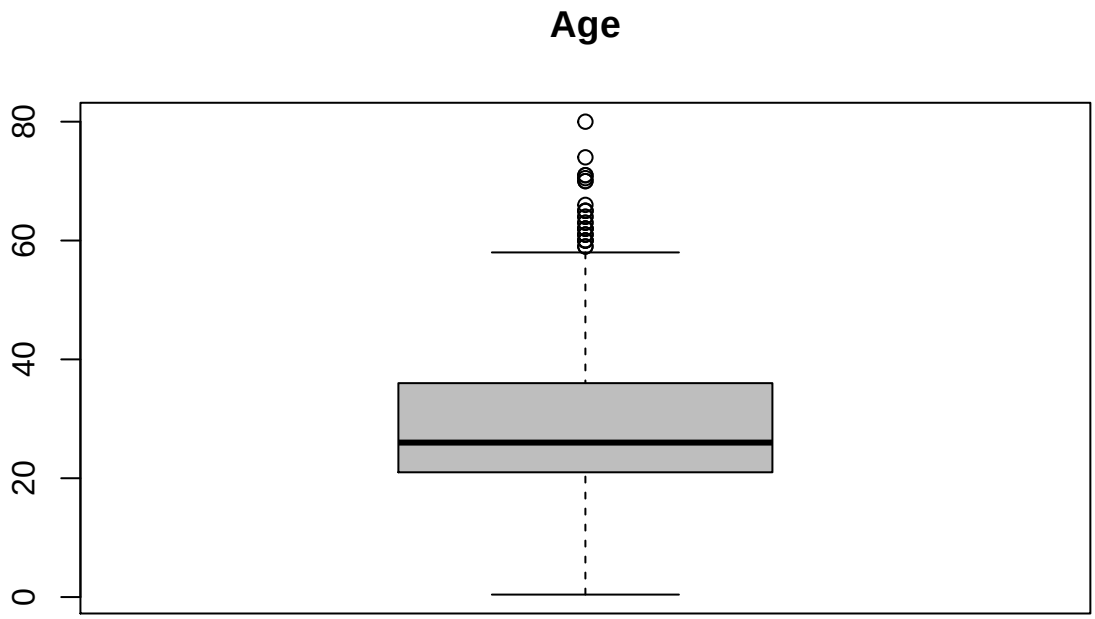
```
require(laeken)
require(VIM)
df$Age <- kNN(df)$Age
sapply(df, function(x) sum(is.na(x)))
```

```
## Survived   Pclass      Sex      Age    SibSp    Parch
##        0        0        0        0        0        0
```

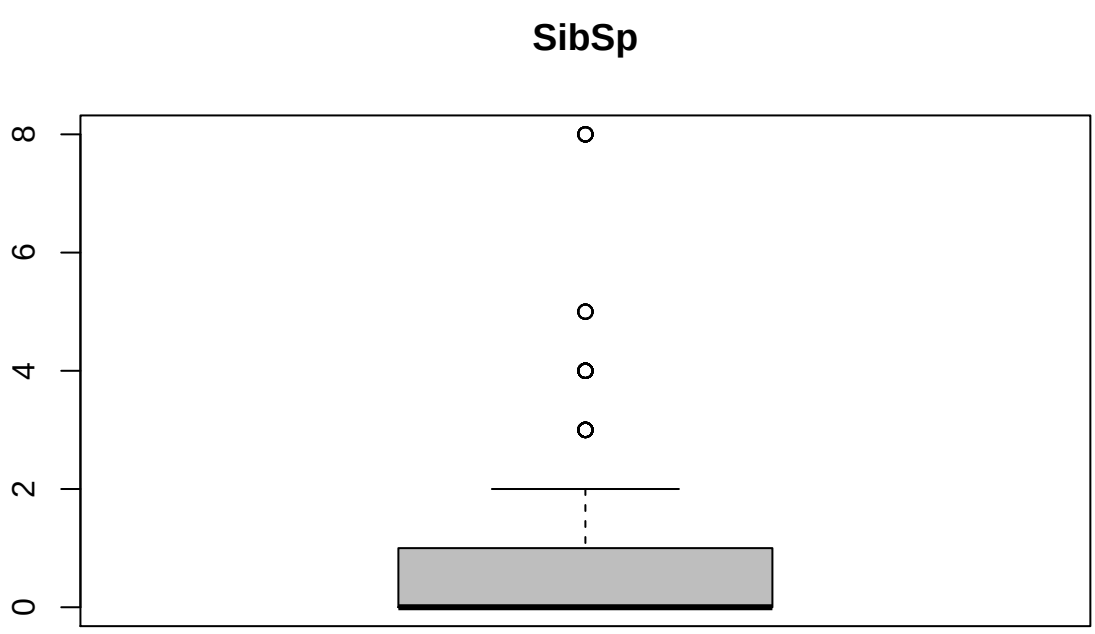
Identificación y tratamiento de valores extremos.

Por el tipo de dato que contienen, los valoers extremos podrián darse en las variables: `Age`, `SibSp` y `Parch`. Las analizamos con un boxplot.

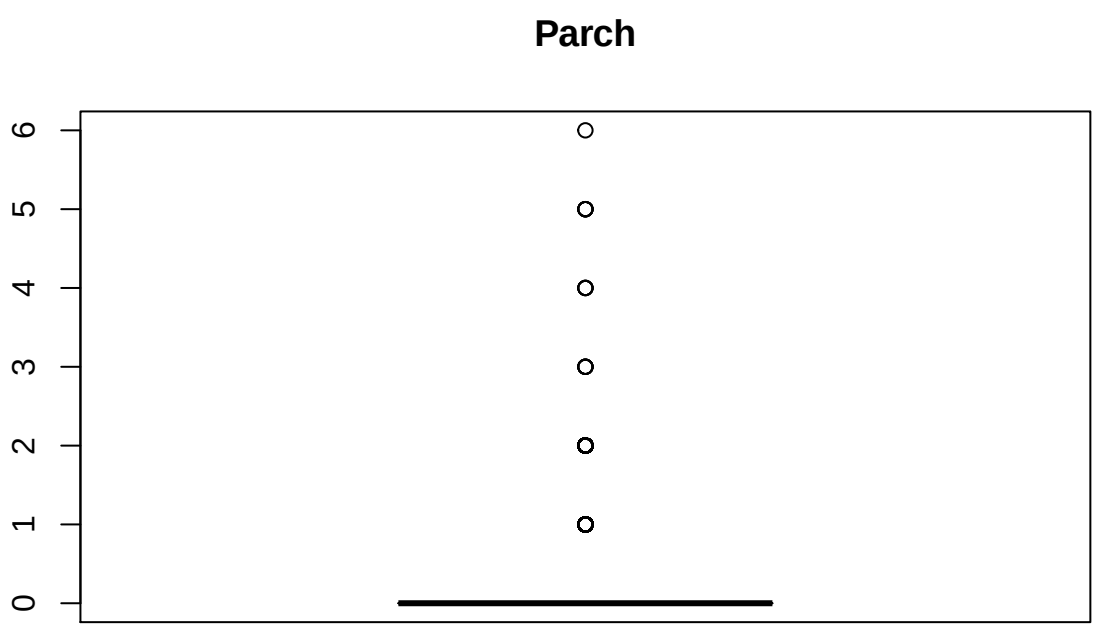
```
boxplot(df$Age, main="Age", col="grey")
```




```
boxplot(df$SibSp, main="SibSp", col="grey")
```




```
boxplot(df$Parch, main="Parch", col="grey")
```



Se puede apreciar que los valores no son *outliers*, simplemente son valores poco frecuentes, pero perfectamente válidos. Voy a optar por mantenerlos sin modificarlos.

Análisis de los datos.

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

Antes de analizar los datos, convertimos los valores lógicos y categóricos en numéricos, para poder calcular correlaciones. Solo será nacesario convertir la variable `Sex`.

```
data <- df
data$Sex <- ifelse(data$Sex == 'female', 0, 1)
```

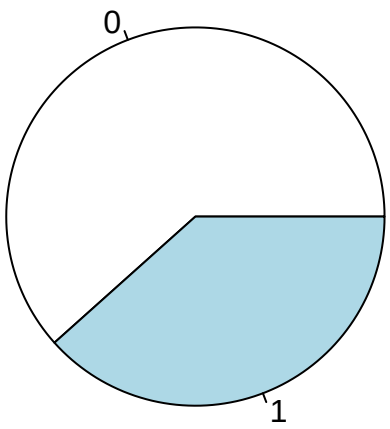
Dada la naturaleza del problema, es fácil suponer que durante el naufragio se pudo dar preferencia a las personas de sexo femenino y a las personas de corta edad. Así mismo, podemos suponer que las personas con más nivel económico (pasajes de 1ª clase) pudieron tener algún tipo de trato de favor.

Es por ello que los grupos que considero interesantes analizar son:

```
menores <- data[data$Age < 12,]  # Pre-adolescentes
mujeres <- data[data$Sex == 0,]
primera_clase <- data[data$Pclass == 1,]
```

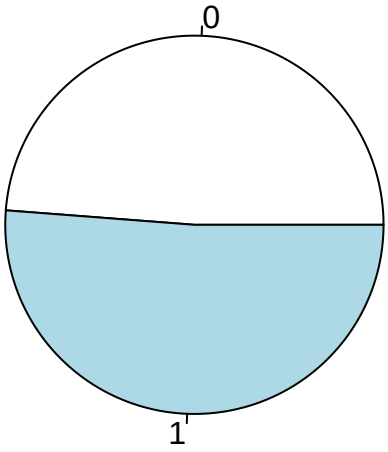
Podemos echar un vistazo, incluso antes de aplicar las pruebas estadísticas.

```
pie(table(data$Survived), main=' Supervivencia (General)')
```



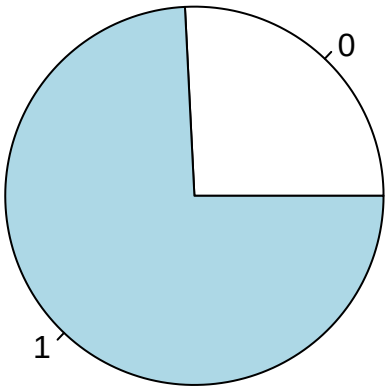

```
pie(table(menores$Survived), main=' Supervivencia entre menores')
```

Supervivencia entre menores

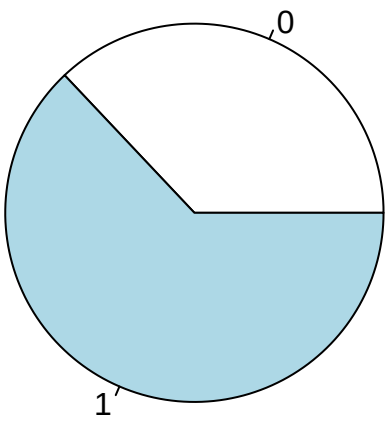



```
pie(table(mujeres$Survived), main=' Supervivencia entre mujeres')
```

Supervivencia entre mujeres




```
pie(table(primera_clase$Survived), main=' Supervivencia pasajeros 1a clase')
```



Como complemento a los datos que tenemos en el dataset, vamos a añadir algunas columnas con datos pre-procesados para tratar de mejorar la predicción.

```
data$IsWoman <- ifelse(data$Sex == 0, 1, 0)
data$IsChild <- ifelse(data$Age < 12, 1, 0)
data$IsRich <- ifelse(data$Pclass == 1, 1, 0)
```

Comprobación de la normalidad y homogeneidad de la varianza.

Comprobamos si las variables, que nos interesan, siguen una distribución normal. Para ello, aplicaremos la prueba de Anderson-Darling.

```
anderson_darling <- function(dataset) {

  alpha = 0.05
  col.names = colnames(dataset)
  cat("Variables que no siguen una distribución normal:\n")

  for (i in 1:ncol(dataset)) {
    if (is.integer(dataset[,i]) | is.numeric(dataset[,i])) {
      p_val = ad.test(dataset[,i])$p.value
      if (p_val < alpha) {
        cat(col.names[i])
        # Format output
        if (i < ncol(dataset) - 1) {
```

```
          cat(", ")
        }
        if (i %% 6 == 0) {
          cat("\n")
        }
      }
    }
  }
}

anderson_darling(data)
```

```
## Variables que no siguen una distribución normal:
## Survived, Pclass, Sex, Age, SibSp, Parch,
## IsWoman, IsChildIsRich
```

```
fligner.test(Survived ~ Pclass, data = data)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Survived by Pclass
## Fligner-Killeen:med chi-squared = 35.766, df = 2, p-value = 1.712e-08
```

```
fligner.test(Survived ~ Age, data = data)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Survived by Age
## Fligner-Killeen:med chi-squared = 102.67, df = 87, p-value = 0.1204
```

```
fligner.test(Survived ~ Sex, data = data)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Survived by Sex
## Fligner-Killeen:med chi-squared = 5.7729, df = 1, p-value = 0.01627
```

```
fligner.test(Survived ~ SibSp, data = data)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Survived by SibSp
## Fligner-Killeen:med chi-squared = 21.832, df = 6, p-value = 0.001298
```

```
fligner.test(Survived ~ Parch, data = data)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Survived by Parch
## Fligner-Killeen:med chi-squared = 17.231, df = 6, p-value = 0.00847
```

```
fligner.test(Survived ~ IsWoman, data = data)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Survived by IsWoman
## Fligner-Killeen:med chi-squared = 5.7729, df = 1, p-value = 0.01627
```

```
fligner.test(Survived ~ IsChild, data = data)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Survived by IsChild
## Fligner-Killeen:med chi-squared = 4.3124, df = 1, p-value = 0.03784
```

```
fligner.test(Survived ~ IsRich, data = data)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Survived by IsRich
## Fligner-Killeen:med chi-squared = 3.1866, df = 1, p-value = 0.07425
```

Consideramos que la varianza de dos variables es homogénea cuando el valor de `p-value > 0.05`.
Parece que las variables, cuya varianza es homogénea con la de la Supervivencia son las de la edad (`Age`) y la de la riquieza (`IsRich`).

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.

Antes de aplicar las pruebas estadíasticas, normalizamos los valores de las diferentes variables para poder compararlas.

```
normalize <- function(x) {
  return ((x - min(x)) / (max(x) - min(x)))
}

normalized_data <- normalize(data)
```

Cálculo de la correlación entre variables

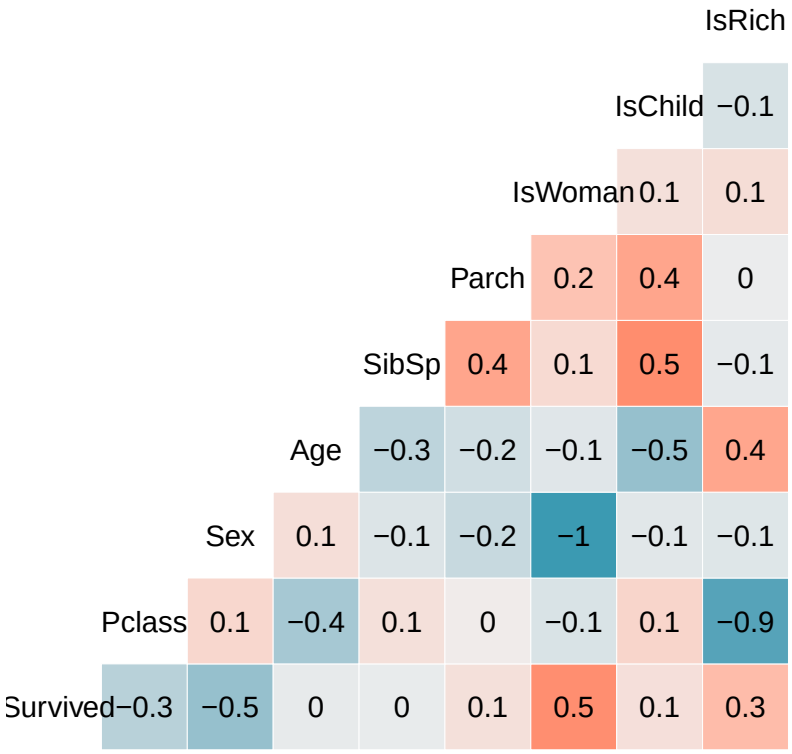
Primero calculamos la matriz de correlación.

```
cor(normalized_data, method = "spearman")
```

```
##             Survived      Pclass         Sex         Age       SibSp
## Survived  1.00000000 -0.33966794 -0.54335138 -0.01021420  0.08887948
## Pclass   -0.33966794  1.00000000  0.13577453 -0.45027709 -0.04301877
## Sex      -0.54335138  0.13577453  1.00000000  0.08069214 -0.19520430
## Age      -0.01021420 -0.45027709  0.08069214  1.00000000 -0.18051049
## SibSp     0.08887948 -0.04301877 -0.19520430 -0.18051049  1.00000000
## Parch     0.13826563 -0.02280134 -0.25451198 -0.25085585  0.45001397
## IsWoman   0.54335138 -0.13577453 -1.00000000 -0.08069214  0.19520430
## IsChild   0.08402912  0.14280822 -0.08210259 -0.50129605  0.39301573
## IsRich    0.28590377 -0.82494696 -0.09801314  0.40074942  0.03387122
##                Parch     IsWoman     IsChild       IsRich
## Survived  0.138265633  0.54335138  0.08402912  0.285903768
## Pclass   -0.022801342 -0.13577453  0.14280822 -0.824946957
## Sex      -0.254511982 -1.00000000 -0.08210259 -0.098013136
## Age      -0.250855849 -0.08069214 -0.50129605  0.400749417
## SibSp     0.450013971  0.19520430  0.39301573  0.033871225
## Parch     1.000000000  0.25451198  0.54376668  0.004300056
## IsWoman   0.254511982  1.00000000  0.08210259  0.098013136
## IsChild   0.543766683  0.08210259  1.00000000 -0.143858359
## IsRich    0.004300056  0.09801314 -0.14385836  1.000000000
```

Los resultados de correlación entre variables son más visibles usando una tabla con colores.

```
ggcorr(normalized_data[,c(1:9)], name = "corr", label = TRUE)+
  theme(legend.position="none")+
labs(title="Correlaciones")+
theme(plot.title=element_text(face='bold', color='black', hjust=0.5, size=12))
```



Parece que las variables más correlacionadas con la `Survived` son `IsWoman` e `IsRich`, lo cual parece que tiene bastante lógica si prestamos atención a las gráficas del punto 4.1.

Contraste de hipótesis de dos muestras sobre la diferencia de medias

Supongamos una hipótesis H_0 y su hipótesis alternativa H_1, con un $\alpha = 0.05$. Siendo μ_1 la media que se extrae de la primera muestra y μ_2 la media de la segunda muestra, describimos el contraste de hipótesis como:

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_1 : \mu_1 - \mu_2 < 0$$

Aplicando el contraste de hipótesis que acabamos de describir, daremos por válida una hipótesis cuando el `p-value` obtenido sea mayor ue el valor de significancia (α).

Contraste de hipótesis: supervivencia segmentando por sexos (hombres *vs* mujeres).

- Aplicamos el Contraste de Hipótersis a dos muestras obtenidas segmentando por sexos. Pretendemos averiguar si el hecho de ser hombre implica una mayor supervivencia en el naufragio:

```
mujeres <- data[data$IsWoman == 1,]
hombres <- data[data$IsWoman == 0,]
t.test(hombres$Survived, mujeres$Survived, alternative="less")
```

```
##
##  Welch Two Sample t-test
##
## data:  hombres$Survived and mujeres$Survived
## t = -18.672, df = 584.43, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.5043259
## sample estimates:
## mean of x mean of y
## 0.1889081 0.7420382
```

Dado que `p-value < 0.05`, **rechazamos la hipótesis** y damos por buena la hipótesis nula, lo cual quiere decir que las mujeres tienen más probabilidades de supervivencia.

Contraste de hipótesis: supervivencia segmentando por edades (niños *vs* mayores).

- Aplicamos el Contraste de Hipótersis a dos muestras obtenidas segmentando por edades. Pretendemos averiguar si el hecho de ser niño/a (menor de 12 años) implica una mayor supervivencia en el naufragio:

```
menores <- data[data$IsChild == 1,]
mayores <- data[data$IsChild == 0,]
t.test(menores$Survived, mayores$Survived, alternative="less")
```

```
##
##  Welch Two Sample t-test
##
## data:  menores$Survived and mayores$Survived
## t = 2.434, df = 96.79, p-value = 0.9916
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf 0.237824
## sample estimates:
## mean of x mean of y
## 0.5121951 0.3708282
```

Dado que `p-value > 0.05`, **aceptamos la hipótesis**, lo cual quiere decir que los menores de 12 años tienen más probabilidades de supervivencia.

Contraste de hipótesis: supervivencia segmentando por clase del pasaje (1ª clase *vs* "el resto").

- Aplicamos el Contraste de Hipótersis a dos muestras obtenidas segmentando por clase de pasaje. Pretendemos averiguar si el hecho de viajar en primera clase implica una mayor supervivencia en el naufragio:

```
first_class <- data[data$IsRich == 1,]
other_classes <- data[data$IsRich == 0,]
t.test(first_class$Survived, other_classes$Survived, alternative="less")
```

```
##
##  Welch Two Sample t-test
```

```
##
## data:  first_class$Survived and other_classes$Survived
## t = 8.6735, df = 348.46, p-value = 1
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf 0.3861366
## sample estimates:
## mean of x mean of y
## 0.6296296 0.3051852
```

Dado que `p-value > 0.05`, **aceptamos la hipótesis**. Lo cual quiere decir que los viajeros de primera clase tienen más probabilidades de supervivencia que el resto del pasaje.

Modelo de regresión Lineal

Vamos a tratar de obtener un modelo de regresión lineal que nos permita, dados unos atributos de pasajero, determinar si hubiese sobrevivido o no, durante el naufragio.

A priori, desconocemos la combinación idónea de regresores. Es por ello que vamos a formular varios modelos de regresión lineal, con distintas combinaciones de regresores.
Después de cada uno de ellos y elegiremos aquel que presente un mayor coeficiente de determinación (Rš).

No vamos a hacer una gran cantidad de modelos con todas las combinaciones de regresores. Elegiremos las variables que mayor grado de correlación hayan presentado con respecto a `Survived`. Basta con echar un vistazo a la tabla coloreada del punto *4.3.1*.

```
first_class = data$IsRich
child = data$IsChild
sex = data$Sex
parent = data$Parch

# Modelos Lineales
alive <- data$Survived
m01 <- lm(alive ~ first_class,                       data=data)
m02 <- lm(alive ~ child,                             data=data)
m03 <- lm(alive ~ sex,                               data=data)
m04 <- lm(alive ~ parent,                            data=data)
m05 <- lm(alive ~ first_class + sex,                 data=data)
m06 <- lm(alive ~ first_class + child,               data=data)
m07 <- lm(alive ~ first_class + parent,              data=data)
m08 <- lm(alive ~ first_class + sex + child,         data=data)
m09 <- lm(alive ~ first_class + sex + child + parent, data=data)
m10 <- lm(alive ~ sex + child,                       data=data)
m11 <- lm(alive ~ sex + parent,                      data=data)
m12 <- lm(alive ~ sex + child + parent,              data=data)
m13 <- lm(alive ~ child + parent,                    data=data)
m14 <- lm(alive ~ child + parent + first_class,      data=data)
m15 <- lm(alive ~ parent + child + first_class,      data=data)
m16 <- lm(alive ~ parent + sex + first_class,        data=data)
m17 <- lm(alive ~ Pclass + SibSp + Parch + Age + Sex, data=data)

coef_table <- matrix(
  c(
    1,  summary(m01)$r.squared,
```

```
    2,  summary(m02)$r.squared,
    3,  summary(m03)$r.squared,
    4,  summary(m04)$r.squared,
    5,  summary(m05)$r.squared,
    6,  summary(m06)$r.squared,
    7,  summary(m07)$r.squared,
    8,  summary(m08)$r.squared,
    9,  summary(m09)$r.squared,
    10, summary(m10)$r.squared,
    11, summary(m11)$r.squared,
    12, summary(m12)$r.squared,
    13, summary(m13)$r.squared,
    14, summary(m14)$r.squared,
    15, summary(m15)$r.squared,
    16, summary(m16)$r.squared,
    17, summary(m17)$r.squared
  ),
  ncol = 2,
  byrow = TRUE
)
colnames(coef_table) <- c("Modelo", "R^2")
coef_table
```

```
##       Modelo         R^2
##  [1,]      1 0.081740964
##  [2,]      2 0.007060893
##  [3,]      3 0.295230723
##  [4,]      4 0.006663360
##  [5,]      5 0.349880907
##  [6,]      6 0.097736717
##  [7,]      7 0.089255144
##  [8,]      8 0.355695883
##  [9,]      9 0.362390041
## [10,]     10 0.296795091
## [11,]     11 0.298081339
## [12,]     12 0.302640185
## [13,]     13 0.009682262
## [14,]     14 0.099106556
## [15,]     15 0.099106556
## [16,]     16 0.351758240
## [17,]     17 0.399366547
```

Los resultados de los *coeficientes de determinación* (Rš) obtenidos para los diferentes modelos, no son demasiado buenos.
El mejor de ellos el el correspondiente al modelo 17. Este modelo está formado por los regresores *Pclass*, *SibSp*, *Parch*, *Age* y *Sex*. Con Rš $= 0.4$ aprox., lo cual no parece un buen resultado.

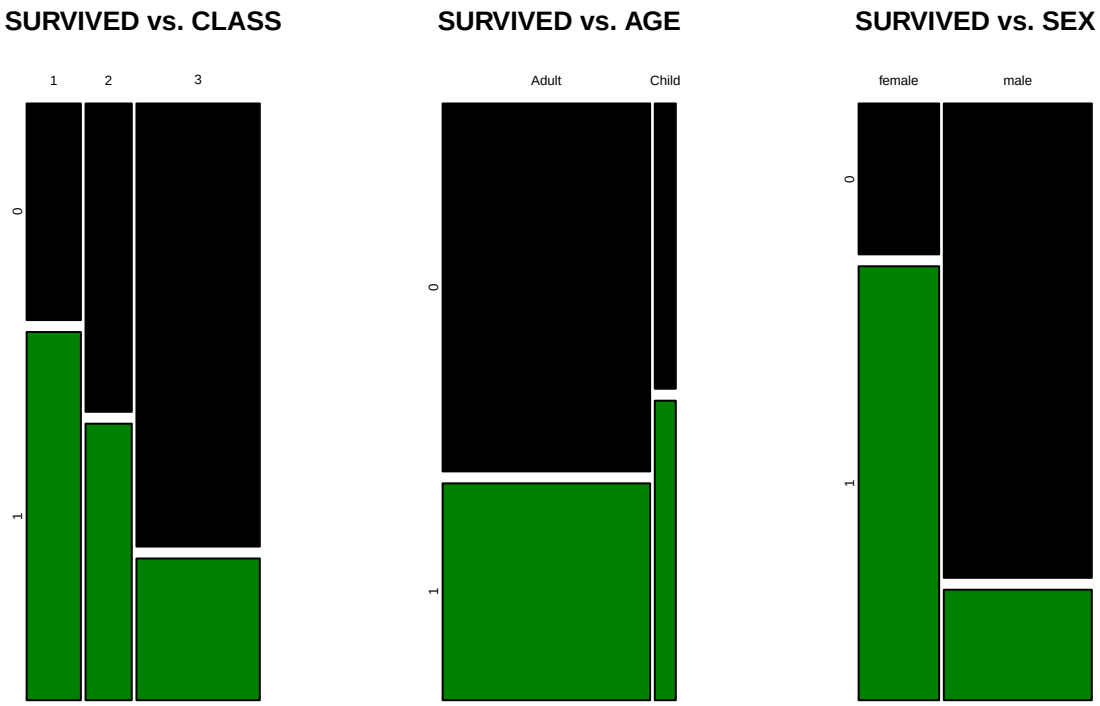
Representación de los resultados a partir de tablas y gráficas.

No hemos podido obtener un modelo de regresión lineal con un buen, *coeficiente de determinación* (Rš). Pero hemos podido ver que algunas variables son más importantes que ortras para predecir la supervivencia

de un pasajero.
Estas variables son el sexo (*Sex*), la clase del pasaje (*Pclass*) y en menor medida, la edad (*Age*).

Veamos, mediante unas gráficas, la relación de estas variables con `Survived`.

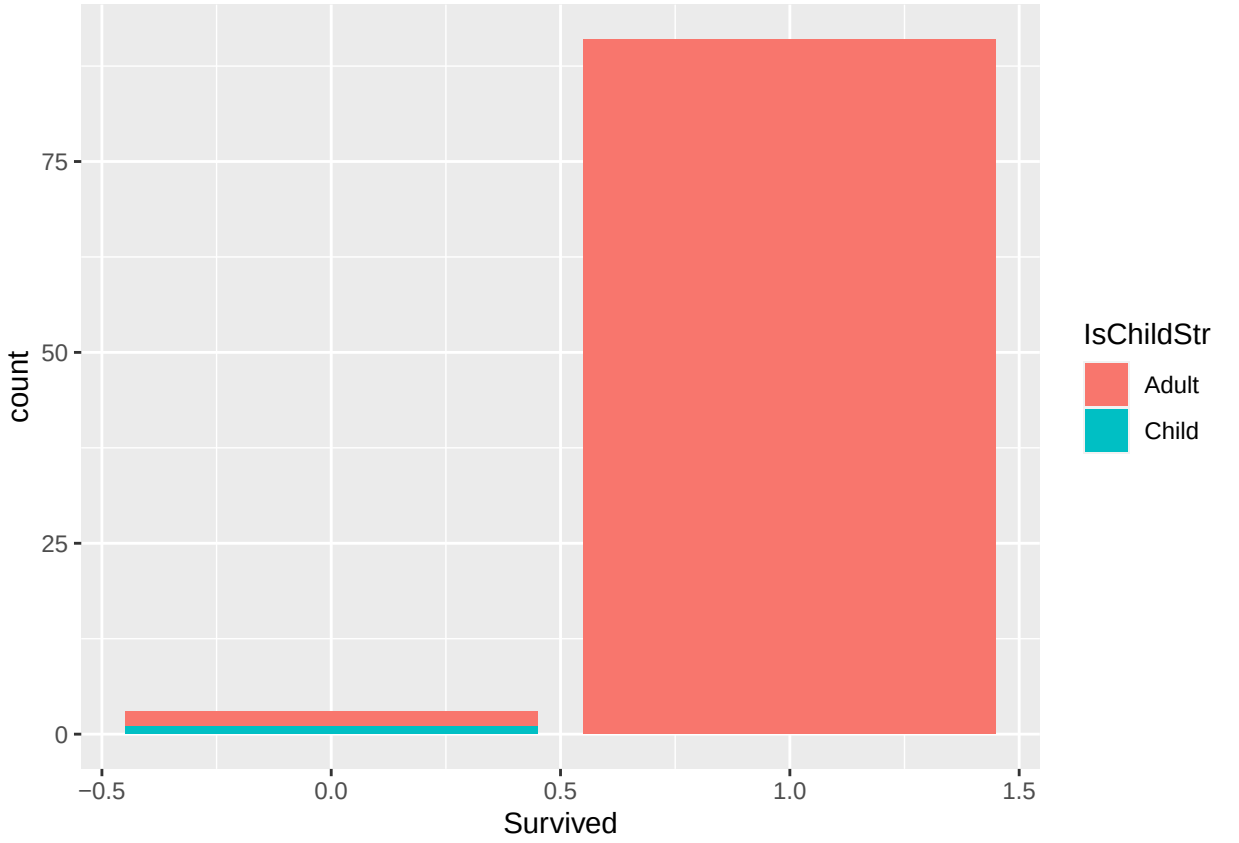
```
# preparamos los datos en forma de tabla para plotear las gráficas
tabla_SST <- table(df$Sex, df$Survived)
tabla_SCT <- table(df$Pclass, df$Survived)
data$IsChildStr <- ifelse(data$IsChild == 1, 'Child', 'Adult')
tabla_SAT <- table(data$IsChildStr, df$Survived)

par(mfrow=c(1, 3))
plot(tabla_SCT, col = c("black","#008000"), main = "SURVIVED vs. CLASS")
plot(tabla_SAT, col = c("black","#008000"), main = "SURVIVED vs. AGE")
plot(tabla_SST, col = c("black","#008000"), main = "SURVIVED vs. SEX")
```



Las gráficas anteriores dan una imagen clara de cuales son los valores de las variables que, en mayor medida, aumentan las probabilidades de supervivencia.

Combinando los valores de las dos variables con más peso, obtendríamos la combinación más favorable (mujeres que viajan en primera clase):

```
first_class_girls <- data %>% filter(IsWoman == 1, IsRich == 1)
ggplot(data=first_class_girls, aes(x=Survived, fill=IsChildStr)) + geom_bar()
```




```
table(first_class_girls$Survived, first_class_girls$IsChildStr)
```

```
##
##     Adult Child
##   0     2     1
##   1    91     0
```

Como podemos ver, la gran mayoría de las mujeres de 1ª clase lograron sobrevivir.

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

A la vista de los resultados obtenidos, las principales conclusiones que podemos sacar son:

- Las variables con más peso a la hora de predecir la supervivencia son: `Pclass`, `Sex` y `Age`.
- Los modelos de regresión lineal no son apropiados para realizar modelos predictivos con este dataset. El mejor de ellos tiene un *coeficiente de determinación* de Rš $\simeq 0.4$.
- A pesar de que no hemos obtenido un modelo de regresión lineal con buena capacidad predictiva, hemos podido determinar algunas combinaciones de variables que arrojan una probabilidad de supervivencia bastante alta.

NOTA:

Para construir un modelo con buenas capacidades de predicción para este dataset, una buena elección podría ser un árbol de decisión. Probablemente tendríamos una mejor tasa de acierto.