Práctica 2. Limpieza y análisis de datos

Asignatura Tipología y ciclo de vida de los datos-Aula 2

Profesor: Diego Pérez

Alumno: Juan Francisco Vallalta Rueda

Contenido

Resolución práctica

Descripción del dataset

Se ha elegido el dataset *Titanic: Machine Learning from Disaster* de la plataforma Kaggle (https://www.kaggle.com/c/titanic).

El objetivo es tratar de predecir a partir de los datos de los pasajeros del Titanic (nombre, edad, precio del ticket, etc.) quien sobrevivirá al naufragio y quien morirá.

El dataset train.csv está formado por 891 observaciones de 12 variables. La estructura del dataset es la siguiente:

- PassengerID: identificador numérico único del pasajero.
- Survived: indica si el pasajero sobrevivió o no al naufragio (0 = No, 1 = Sí).
- Pclass: clase del pasaje (1 = 1ª clase, 2 = 2ª clase, 3 = 3ª clase).
- Name: nombre del pasajero.
- Sex: sexo del pasajero.
- Age: edad del pasajero.
- Sibsp: número de hermanos, cuñados y esposas a bordo (relación familiar con el pasajero).
- Parch: número de padres e hijos a bordo (relación familiar con el pasajero).
- Ticket: número de ticket.
- Fare: importe del pasaje.
- Cabin: número de cabina.
- Embarked: puerto de embarque (C = Cherbourg, Q = Queenstown, S = Southampton).

Integración y selección de datos

En esta práctica se parte de una única fuente de datos ***train.csv*** por lo que no se realizará ninguna actividad de integración o fusión de datos. Procederemos directamente a la carga de los mismos en R.

```
# Cargamos el juego de datos train.csv y visualizamos los primeros registros
titanic <- read.csv('train.csv', stringsAsFactors = FALSE)
head(titanic)
```

	PassengerId <int>	Survived <int>	Pclass <int>	Name <chr>	Sex <chr>	Age <dbl>	SibSp <int>	Parch <int>	Ticket <chr>	Fare <dbl>	Cabin <chr>
1	1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.2500	
2	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85
3	3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.9250	
4	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1000	C123
5	5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.0500	
6	6	0	3	Moran, Mr. James	male	NA	0	0	330877	8.4583	

Seleccionaremos todas las variables del dataset excepto *PassengerID, Name y Ticket* que no son relevantes para el análisis.

```
# Descartamos las variables PassengerId, Name y Ticket
titanic <- titanic %>% select(-PassengerId,-Name, -Ticket)
head(titanic)
```

	Survived <int>	Pclass <int>	Sex <chr>	Age <dbl>	SibSp <int>	Parch <int>	Fare <dbl>	Cabin <chr>	Embarked <chr>
1	0	3	male	22	1	0	7.2500		S
2	1	1	female	38	1	0	71.2833	C85	C
3	1	3	female	26	0	0	7.9250		S
4	1	1	female	35	1	0	53.1000	C123	S
5	0	3	male	35	0	0	8.0500		S
6	0	3	male	NA	0	0	8.4583		Q

Procedemos a la creación de una nueva variable TamFamilia que representa el tamaño de la familia que viaja a bordo.

```
#Creamos una nueva variable tamaño de la familia
titanic <- titanic %>% mutate(TamFamilia = SibSp + Parch +1)
head(titanic)
```

	Survived <int>	Pclass <int>	Sex <chr>	Age <dbl>	SibSp <int>	Parch <int>	Fare <dbl>	Cabin <chr>	Embarked <chr>	TamFamilia <dbl>
1	0	3	male	22	1	0	7.2500		S	2
2	1	1	female	38	1	0	71.2833	C85	C	2
3	1	3	female	26	0	0	7.9250		S	1
4	1	1	female	35	1	0	53.1000	C123	S	2
5	0	3	male	35	0	0	8.0500		S	1
6	0	3	male	NA	0	0	8.4583		Q	1

Limpieza de datos

Procederemos al análisis de datos perdidos y valores extremos.

Datos perdidos

Los datos perdidos representan una pérdida de información por lo que su gestión adecuada es muy relevante para los resultados del análisis.

Los datos perdidos se pueden presentar de diversas formas: espacio en blanco, valor 0, caracteres '?', …

Procederemos a visualizar si existen datos perdidos en cada una de las variables:

```
# Estadísticas de valores vacíos
colSums(is.na(titanic))
```

```
Survived   Pclass      Sex      Age    SibSp    Parch     Fare
       0        0        0      177        0        0        0
   Cabin Embarked TamFamilia
       0        0        0
```

Hay 177 observaciones con el valor perdido para la edad.

```
# Estadísticas de valores vacíos
colSums(titanic=="")
```

```
Survived   Pclass      Sex      Age    SibSp    Parch     Fare
       0        0        0       NA        0        0        0
   Cabin Embarked TamFamilia
     687        2        0
```

Hay 687 observaciones con valor perdido para la variable Cabin y 2 observaciones con valor perdido para la variable Embarked. Estos nulos están codificados con el carácter “”.

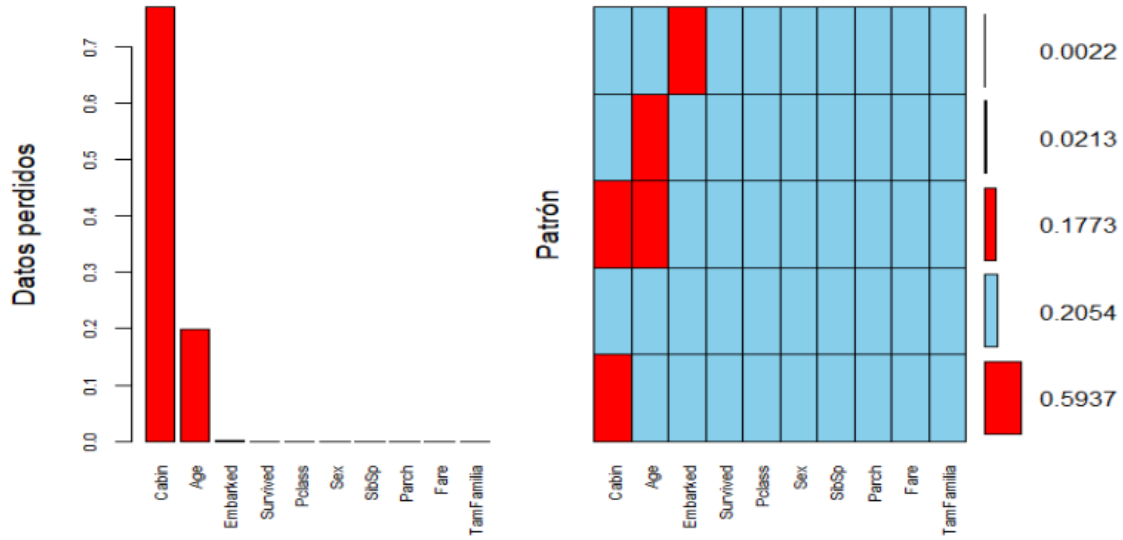
Para una gestión adecuada, en primer lugar, procedemos a representar todos los nulos mediante NA que es el objeto que utiliza R para representar valores perdidos o faltantes.

```
#Asignamos NA a Cabin y Embarked
titanic$Cabin[titanic$Cabin==""]<-NA
titanic$Embarked[titanic$Embarked==""]<-NA
colSums(is.na(titanic))
```

```
Survived   Pclass      Sex      Age    SibSp    Parch     Fare
       0        0        0      177        0        0        0
   Cabin Embarked TamFamilia
     687        2        0
```

Procedemos a la representación gráfica de los valores perdidos mediante la función aggr() del paquete VIM. Esa función nos permite representar el porcentaje de valores perdidos en cada variable así como posibles patrones en el dataset.

```
#Visualización valores perdidos
aggr(titanic, numbers=TRUE, sortVars=TRUE, labels=names(titanic), cex.axis=.7, gap=3,
ylab=c("Datos perdidos","Patrón"))
```



La proporción de valores perdidos en el dataset es la siguiente:

- Cabin: 77,1%.
- Age: 19,9 %
- Embarked: 0,2%.

No se observan patrones no aleatorios en los datos perdidos.

A continuación procedemos a la gestión de los datos nulos. Se optará por las siguientes estrategias:

- Variable Cabin: como el porcentaje de valores nulos en la variable es muy elevado se procederá a descartar la variable del análisis.
- Age: se imputará en función de la similitud de registros basada en los vecinos más próximos (kNN-Imputation)
- Embarked: se imputará en función de la similitud de registros basada en los vecinos más próximos (kNN-Imputation)

```
# Imputación de valores mediante la función kNN() del paquete VIM
titanic<-titanic %>% select(-Cabin)
titanic$Age <- kNN(titanic)$Age
titanic$Embarked <- kNN(titanic)$Embarked
colSums(is.na(titanic))
```

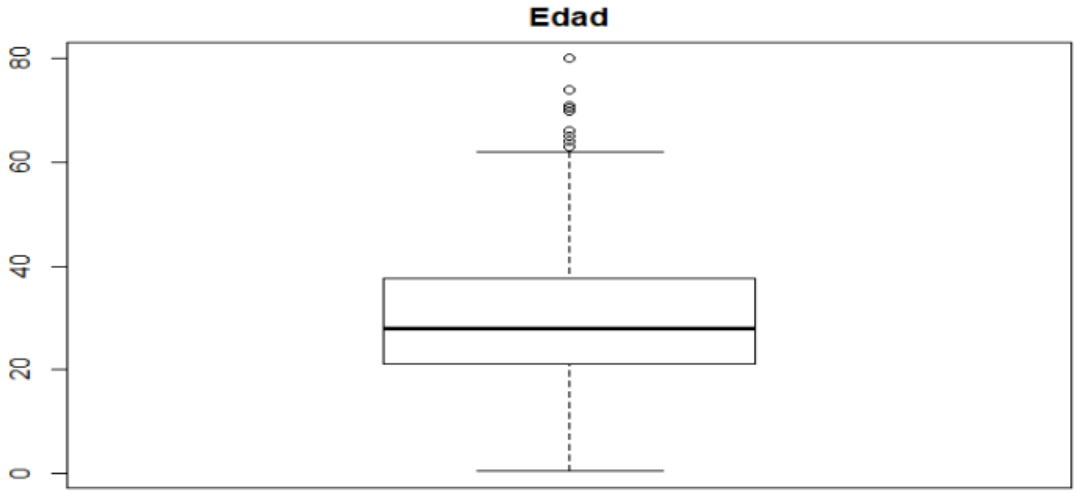
```
  Survived     Pclass        Sex        Age      SibSp      Parch       Fare
         0          0          0          0          0          0          0
  Embarked TamFamilia
         0          0
```

Valores extremos

Utilizares el diagrama de cajas para la visualización de valores extremos de las variables numéricas del dataset:

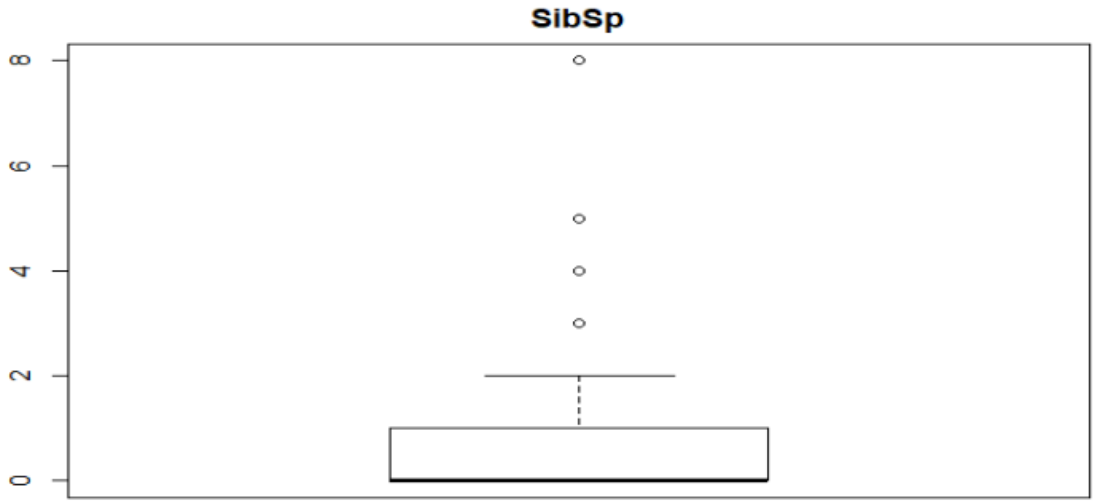
Variable Age

```
#Valores extremos
#Diagrama de cajas Age
Age.bp<-boxplot(titanic$Age, main='Edad')
Age.bp$out
[1] 66.0 65.0 71.0 70.5 63.0 65.0 64.0 65.0 63.0 71.0 64.0 80.0 70.0 70.0
74.0
```



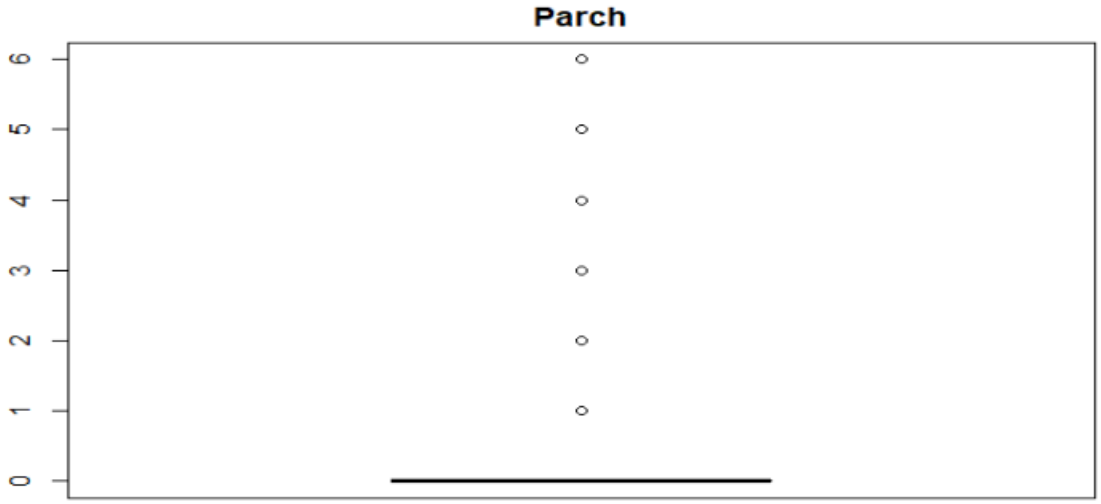
Variable SibSp

```
#Valores extremos
#Diagrama de cajas SibSp
SibSp.bp<-boxplot(titanic$SibSp)
SibSp.bp$out
[1] 3 4 3 3 4 5 3 4 5 3 3 4 8 4 4 3 8 4 8 3 4 4 4 4 8 3 3 5 3 5 3 4 4 3 3 5 4 3 4
[40] 8 4 3 4 8 4 8
```



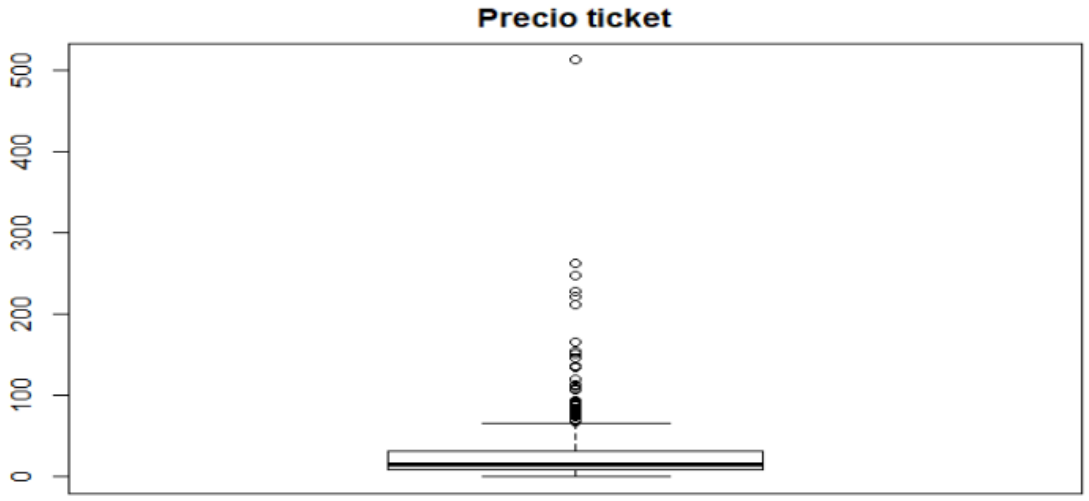
Variable Parch

```
#Valores extremos
#Diagrama de cajas Parch
Parch.bp<-boxplot(titanic$Parch, main='Parch')
Parch.bp$out
[1] 1 2 1 5 1 1 5 2 2 1 1 2 2 2 1 2 2 2 3 2 2 1 1 1 1 2 1 1 2 2 1 2 2 2 1 2 1 1
 [39] 2 1 4 1 1 1 1 2 2 1 2 1 1 1 2 1 1 2 2 2 1 1 2 2 1 2 1 1 1 1 1 1 2 1 2 2 1
 [77] 1 2 1 1 2 1 1 1 1 2 1 1 1 4 1 1 2 2 2 2 2 1 1 1 2 2 1 1 2 2 3 4 1 2 1 1 2 1
[115] 2 1 2 1 1 2 2 1 1 1 1 2 2 2 2 2 2 1 1 2 1 4 1 1 2 1 2 1 1 2 5 2 1 1 1 2 1 5
[153] 2 1 1 1 2 1 6 1 2 1 2 1 1 1 1 1 1 1 3 2 1 1 1 1 2 1 2 3 1 2 1 2 2 1 1 2 1 2
[191] 1 2 1 1 1 2 1 1 2 1 2 1 1 1 1 3 2 1 1 1 1 5 2
```



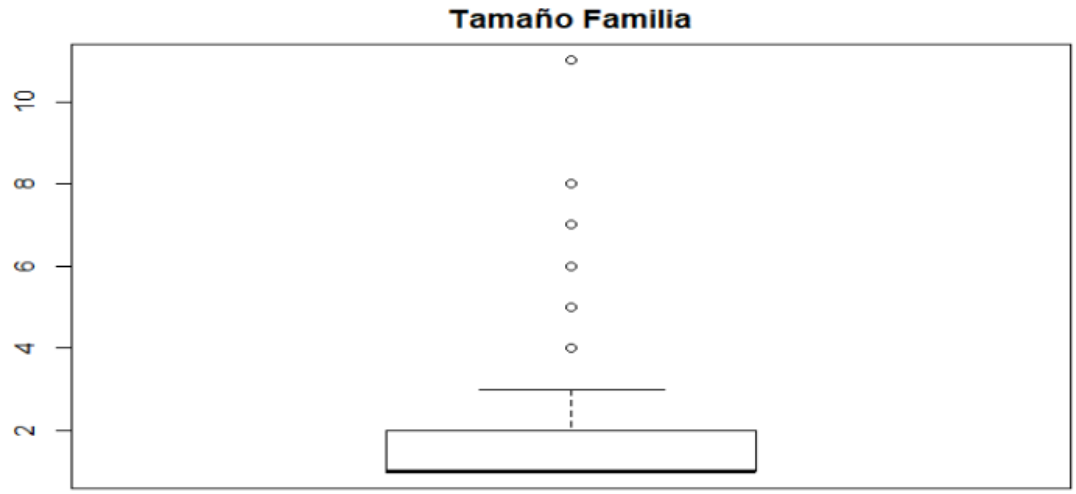
Variable Fare

```
#Valores extremos
#Diagrama de cajas Fare
Fare.bp<-boxplot(titanic$Fare, main='Precio ticket')
Fare.bp$out
[1]  71.2833 263.0000 146.5208  82.1708  76.7292  80.0000  83.4750  73.5000
  [9] 263.0000  77.2875 247.5208  73.5000  77.2875  79.2000  66.6000  69.5500
 [17]  69.5500 146.5208  69.5500 113.2750  76.2917  90.0000  83.4750  90.0000
 [25]  79.2000  86.5000 512.3292  79.6500 153.4625 135.6333  77.9583  78.8500
 [33]  91.0792 151.5500 247.5208 151.5500 110.8833 108.9000  83.1583 262.3750
 [41] 164.8667 134.5000  69.5500 135.6333 153.4625 133.6500  66.6000 134.5000
 [49] 263.0000  75.2500  69.3000 135.6333  82.1708 211.5000 227.5250  73.5000
 [57] 120.0000 113.2750  90.0000 120.0000 263.0000  81.8583  89.1042  91.0792
 [65]  90.0000  78.2667 151.5500  86.5000 108.9000  93.5000 221.7792 106.4250
 [73]  71.0000 106.4250 110.8833 227.5250  79.6500 110.8833  79.6500  79.2000
 [81]  78.2667 153.4625  77.9583  69.3000  76.7292  73.5000 113.2750 133.6500
 [89]  73.5000 512.3292  76.7292 211.3375 110.8833 227.5250 151.5500 227.5250
 [97] 211.3375 512.3292  78.8500 262.3750  71.0000  86.5000 120.0000  77.9583
[105] 211.3375  79.2000  69.5500 120.0000  93.5000  80.0000  83.1583  69.5500
[113]  89.1042 164.8667  69.5500  83.1583
```



Variable TamFamilia

```
#Valores extremos
#Diagrama de cajas Fare
TamFamilia.bp<-boxplot(titanic$TamFamilia, main='Tamaño Familia')
TamFamilia.bp$out
[1]  5  7  6  5  7  6  4  6  4  8  6  7  8  4  5  6  4  7  5 11  6  6  6  5
11  7
[27]  4 11  5  7  7  6  6  4  4  5 11  6  6  5  8  4  5  4  5  6  6  4  4  4
4  8
[53]  5  4  4  7  7  5  4  4  7  4  4  6  6  6  4  8  8  4  6  4  5  5  4  4
5  4
[79]  6  4 11  4  7  6  6 11  7  4 11  6  4
```



En nuestro caso, los valores extremos son legítimos ya que se encuentran dentro del rango de variación posible de la edad, tamaño de familia y precio del ticket.

Hay que considerar que se identifican como extremos según el criterio de normalidad, es decir, estar muy alejados de la media en una distribución normal. Pero la variable no sigue una distribución normal, presentando un elevado grado de asimetría.

Análisis de datos

Procederemos a realizar el análisis de los datos con el objeto de explicar las principales características de los mismos así como responder a las preguntas formuladas en el epígrafe planificación del análisis.

Planificación del análisis

Se seleccionan las siguientes variables para el análisis:

- Survived: indica si el pasajero sobrevivió o no al naufragio (0 = No, 1 = Sí).
- Pclass: clase del pasaje (1 = 1ª clase, 2 = 2ª clase, 3 = 3ª clase).
- Sex: sexo del pasajero.
- Age: edad del pasajero.
- Embarked: puerto de embarque (C = Cherbourg, Q = Queenstown, S = Southampton).
- TamFamilia: número de miembros de la familia.

```
#Selección variables análisis
titanic_ana <- titanic %>% select(Survived, Pclass, Sex, Age, Embarked,
TamFamilia)
head(titanic_ana)
```

	Survived <int>	Pclass <int>	Sex <chr>	Age <dbl>	Embarked <chr>	TamFamilia <dbl>
1	0	3	male	22	S	2
2	1	1	female	38	C	2
3	1	3	female	26	S	1
4	1	1	female	35	S	2
5	0	3	male	35	S	1
6	0	3	male	21	Q	1

Con el análisis se pretende conseguir los siguientes resultados:

- Identificar qué variables son significativas desde el punto de vista estadístico para predecir que el pasajero del Titanic no sobrevivió al hundimiento.
- Construir un modelo de regresión logística que permita predecir si un pasajero en función de su sexo, clase de pasaje, edad, puerto de embarque y tamaño de la familia sobrevivió o no al hundimiento del Titanic.

Análisis estadístico básico

Independientemente del objetivo del análisis a realizar es una buena práctica realizar siempre un análisis estadístico básico para conocer mejor los datos.

En primer lugar, convertiremos a tipo factor las variables categóricas: Pclass, Sex y Embarked.

```
#Convertimos a factor
titanic_ana$Pclass <- as.factor(titanic_ana$Pclass)
titanic_ana$Sex <- as.factor(titanic_ana$Sex)
titanic_ana$Embarked <- as.factor(titanic_ana$Embarked)
head(titanic_ana)
```

	Survived <int>	Pclass <fctr>	Sex <fctr>	Age <dbl>	Embarked <fctr>	TamFamilia <dbl>
1	0	3	male	22	S	2
2	1	1	female	38	C	2
3	1	3	female	26	S	1
4	1	1	female	35	S	2
5	0	3	male	35	S	1
6	0	3	male	21	Q	1

A continuación procedemos a revisar la estructura de los datos:

```
#Estructura de los datos
str(titanic_ana)
 'data.frame':   891 obs. of  6 variables:
  $ Survived  : int  0 1 1 1 0 0 0 0 1 1 ...
  $ Pclass    : Factor w/ 3 levels "1","2","3": 3 1 3 1 3 3 1 3 3 2 ...
  $ Sex       : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
  $ Age       : num  22 38 26 35 35 21 54 2 27 14 ...
  $ Embarked  : Factor w/ 3 levels "C","Q","S": 3 1 3 3 3 2 3 3 3 1 ...
  $ TamFamilia: num  2 2 1 2 1 1 1 5 3 2 ...
```

Procedemos a realizar un resumen estadístico básico:

```
#Estadísticas básicas
summary(titanic_ana)
```

```
    Survived      Pclass      Sex           Age        Embarked   TamFamilia
 Min.   :0.0000   1:216   female:314   Min.   : 0.42   C:168    Min.   : 1.000
 1st Qu.:0.0000   2:184   male  :577   1st Qu.:21.00   Q: 77    1st Qu.: 1.000
 Median :0.0000   3:491                Median :28.00   S:646    Median : 1.000
 Mean   :0.3838                        Mean   :29.57            Mean   : 1.905
 3rd Qu.:1.0000                        3rd Qu.:37.50            3rd Qu.: 2.000
 Max.   :1.0000                        Max.   :80.00            Max.   :11.000
```

Este análisis nos muestra que de los 891 pasajeros que contiene el dataset sobrevivió un 38,4%. La mediana de edad de los pasajeros era de 28 años y la mediana del tamaño familiar era de una

persona. También nos muestran cómo se distribuyen los pasajeros por sexo, clase de pasaje y puerto de embarque.

Análisis estadístico inferencial

Mediante el análisis estadístico inferencial determinaremos que variables tiene un efecto significativo estadísticamente en la supervivencia o no al naufragio.

¿Existen diferencias significativas en la supervivencia por tipo de pasaje?

Aplicaremos un test Chi-cuadrado:

```
#¿Existen diferencias significativas en la supervivencia por tipo de pasaje?´
tabla<-table(titanic_ana$Pclass, titanic_ana$Survived)
tabla
chisq.test(tabla)
```

```
     0   1
1   80 136
2   97  87
3 372 119

        Pearson's Chi-squared test

data:  tabla
X-squared = 102.89, df = 2, p-value < 2.2e-16
```

El p-valor obtenido p<2.2e-16 nos sugiere rechazar la hipótesis nula y afirmar que existen diferencias significativas en la supervivencia de los pasajeros por tipo de pasaje.

.

¿Existen diferencias significativas en la supervivencia por sexo?

Aplicaremos un test Chi-cuadrado:

```
#¿Existen diferencias significativas en la supervivencia por sexo?´
tabla<-table(titanic_ana$Sex, titanic_ana$Survived)
tabla
chisq.test(tabla)

           0   1
  female  81 233
  male   468 109

        Pearson's Chi-squared test with Yates' continuity correction

data:  tabla
X-squared = 260.72, df = 1, p-value < 2.2e-16
```

El p-valor obtenido p<2.2e-16 nos sugiere rechazar la hipótesis nula y afirmar que existen diferencias significativas en la supervivencia de los pasajeros por sexo.

¿Existen diferencias significativas en la supervivencia por puerto de embarque?

Aplicaremos un test Chi-cuadrado:

```
#¿Existen diferencias significativas en la supervivencia por puerto de
embarque?
tabla<-table(titanic_ana$Embarked, titanic_ana$Survived)
tabla
chisq.test(tabla))

      0   1
  C  75  93
  Q  47  30
  S 427 219

        Pearson's Chi-squared test

data:  tabla
X-squared = 25.964, df = 2, p-value = 2.301e-06
```

El p-valor obtenido p<2.3e-6 nos sugiere rechazar la hipótesis nula y afirmar que existen diferencias significativas en la supervivencia de los pasajeros por puerto de embarque.

¿Existen diferencias significativas en la supervivencia por edad de los pasajeros?

En ese caso se quiere analizar si existe alguna diferencia significativa entre la edad media de los pasajero que sobrevivieron de los que no. Al tratarse la edad de una variable numérica hemos de verificar si se cumplen los supuestos de normalidad y homocedasticidad para elegir el contraste de hipótesis adecuado: paramétrico o no paramétrico.

Análisis de normalidad

Se utilizará el test de Shapiro-Wilk para contrastar normalidad.

```
#Test normalidad edad
shapiro.test(titanic_ana$Age)

        Shapiro-Wilk normality test

data:  titanic_ana$Age
W = 0.97898, p-value = 4.948e-10
```

El p-valor obtenido nos sugiere rechazar la hipótesis nula y afirma que los datos no se distribuyen normalmente.

Análisis de homocedasticidad

Se utilizará el test de Fligner-Killeen para contrastar homocestacidad.

```
##Test de homocedasticidad
fligner.test(Age~Survived, data=titanic_ana)

        Fligner-Killeen test of homogeneity of variances

data:  Age by Survived
Fligner-Killeen:med chi-squared = 0.59219, df = 1, p-value = 0.4416
```

El p-valor obtenido nos sugiere aceptar la hipótesis nula y afirma que los datos no presentan diferencias significativas en sus varianzas.

Contraste de hipótesis

Como no se cumplen los requisitos para el contraste paramétrico utilizaremos la prueba de Mann-Whitney.

```
#Constraste de hipótesis
wilcox.test(Age~Survived, data=titanic_ana)
```

```
        Wilcoxon rank sum test with continuity correction

data:  Age by Survived
W = 103370, p-value = 0.01102
alternative hypothesis: true location shift is not equal to 0
```

El p-valor obtenido nos sugiere rechazar la hipótesis nula de igualdad de las medianas y afirmar que existen diferencias significativas en la edad de los pasajeros que sobrevivieron.

¿Existen diferencias significativas en la supervivencia por el número de miembros de la familia?

En ese caso se quiere analizar si existe alguna diferencia significativa entre el número medio de miembro de la familia de los pasajero que sobrevivieron de los que no. Al tratarse el número de miembros de una variable numérica hemos de verificar si se cumplen los supuestos de normalidad y homocedasticidad para elegir el contraste de hipótesis adecuado: paramétrico o no paramétrico.

Se utilizará el test de Shapiro-Wilk para contrastar normalidad y el Fligner-Killeen para contrastar homocestacidad. Como no se cumplen las condiciones de normalidad y homocedasticidad se utilizará el test Mann-Whitney para contrastar la igualdad de las medianas.

```
#Test normalidad edad
shapiro.test(titanic_ana$TamFamilia)

#Test de homocedasticidad
fligner.test(TamFamilia~Survived, data=titanic_ana)

#Constraste de hipótesis
wilcox.test(TamFamilia~Survived, data=titanic_ana)
```

```
        Shapiro-Wilk normality test

data:  titanic_ana$TamFamilia
W = 0.61508, p-value < 2.2e-16


        Fligner-Killeen test of homogeneity of variances

data:  TamFamilia by Survived
Fligner-Killeen:med chi-squared = 19.647, df = 1, p-value = 9.317e-06


        Wilcoxon rank sum test with continuity correction

data:  TamFamilia by Survived
W = 77659, p-value = 7.971e-07
alternative hypothesis: true location shift is not equal to 0
```

El p-valor obtenido nos sugiere rechazar la hipótesis nula de igualdad de las medianas y afirmar que existen diferencias significativas en el número de miembros de la familia de los pasajeros que sobrevivieron.

Modelo de regresión logística

El modelo se construirá para predecir si un pasajero sobrevivió o no en función de las variables que estamos analizando.

```
#Modelo de regresión logística
modelo = glm(Survived~., data=titanic_ana, family="binomial")
summary(modelo)
```

```
Call:
glm(formula = Survived ~ ., family = "binomial", data = titanic_ana)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6594  -0.6209  -0.3792   0.6208   2.5431

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.064060   0.480433  10.541  < 2e-16 ***
Pclass2     -1.236346   0.277999  -4.447 8.70e-06 ***
Pclass3     -2.603841   0.271679  -9.584  < 2e-16 ***
Sexmale     -2.716149   0.202330 -13.424  < 2e-16 ***
Age         -0.049749   0.007913  -6.287 3.23e-10 ***
EmbarkedQ   -0.033917   0.395739  -0.086 0.931700
EmbarkedS   -0.445723   0.239467  -1.861 0.062700 .
TamFamilia  -0.253040   0.067004  -3.776 0.000159 ***
---
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:  769.35  on 883  degrees of freedom
AIC: 785.35

Number of Fisher Scoring iterations: 5
```

El modelo sugiere que los pasajeros de clase, el sexo, la edad y tamaño de familia son factores que influyen significativamente en la predicción de la supervivencia. El valor del criterio de información de Akaike (AIC) para este modelo que incluye todas las variables es AIC = 785.35.

A continuación construiremos una serie de modelos incluyendo variables una a una y lo comparemos entre sí y con el modelo general. El modelo con menor AIC será el modelo seleccionado.

```
#Comparación de modelos
modelo1 = glm(Survived~Pclass, data=titanic_ana, family="binomial")
modelo2 = glm(Survived~Pclass+Sex, data=titanic_ana, family="binomial")
modelo3 = glm(Survived~Pclass+Sex+Age, data=titanic_ana, family="binomial")
modelo4 = glm(Survived~Pclass+Sex+Age+TamFamilia, data=titanic_ana,
family="binomial")
modelo1$aic
modelo2$aic
modelo3$aic
modelo4$aic
```

```
[1] 1089.108
[1] 834.8884
[1] 802.7559
[1] 785.6747
```

El valor de AIC nos indica que seleccionemos el primer modelo con todas las variables para la predicción de la supervivencia.

Conclusiones

Los resultado sugieren que las variables clase, sexo, edad, puerto de embarque y tamaño de familia son factores que influyen significativamente en la supervivencia o no del pasajero en el naufragio del Titanic.

Considerando esas variables se ha construido un modelo de regresión logística capaz de predecir si un pasajero sobrevivirá o no.

Código

A continuación de adjunta el código R que se ha utilizado para la realización de la práctica.

```
# Cargamos los paquetes R que vamos a usar
library(readxl)
library(dplyr)
library(ggplot2)
library(VIM)
library(tidyr)
library(naniar)

# Cargamos el juego de dagos train
titanic <- read.csv('train.csv', stringsAsFactors = FALSE)
head(titanic)

# Descartamos las variables PassengerId, Name y Ticket
titanic <- titanic %>% select(-PassengerId,-Name, -Ticket)
head(titanic)

#Creamos una nueva variable tamaño de la familia
titanic <- titanic %>% mutate(TamFamilia = SibSp + Parch +1)
head(titanic)

# Estadísticas de valores vacíos
colSums(is.na(titanic))
colSums(titanic=="")
colSums(titanic=="?")

#Asignamos NA a Cabin y Embarked
titanic$Cabin[titanic$Cabin==""]<-NA
titanic$Embarked[titanic$Embarked==""]<-NA
colSums(is.na(titanic))

#Visualización valores perdidos
aggr(titanic, numbers=TRUE, sortVars=TRUE, labels=names(titanic), cex.axis=.7, gap=3,
ylab=c("Datos perdidos","Patrón"))

# Imputación de valores mediante la función kNN() del paquete VIM
titanic<-titanic %>% select(-Cabin)
titanic$Age <- kNN(titanic)$Age
titanic$Embarked <- kNN(titanic)$Embarked
colSums(is.na(titanic))
```

```
#Valores extremos
#Diagrama de cajas Age
Age.bp<-boxplot(titanic$Age, main='Edad')
Age.bp$out

#Valores extremos
#Diagrama de cajas SibSp
SibSp.bp<-boxplot(titanic$SibSp, main='SibSp')
SibSp.bp$out

#Valores extremos
#Diagrama de cajas Parch
Parch.bp<-boxplot(titanic$Parch, main='Parch')
Parch.bp$out

#Valores extremos
#Diagrama de cajas Fare
Fare.bp<-boxplot(titanic$Fare, main='Precio ticket')
Fare.bp$out

#Valores extremos
#Diagrama de cajas Fare
TamFamilia.bp<-boxplot(titanic$TamFamilia, main='Tamaño Familia')
TamFamilia.bp$out

#Selección variables análisis
titanic_ana <- titanic %>% select(Survived, Pclass, Sex, Age, Embarked, TamFamilia)
head(titanic_ana)

#Convertimos a factor
titanic_ana$Pclass <- as.factor(titanic_ana$Pclass)
titanic_ana$Sex <- as.factor(titanic_ana$Sex)
titanic_ana$Embarked <- as.factor(titanic_ana$Embarked)
head(titanic_ana)

#Estructura de los datos
str(titanic_ana)

#Estadísticas basicas
summary(titanic_ana)

#Análisis estadistico inferencial
#¿Existen diferencias significativas en la supervivencia por tipo de pasaje?´
```

```
tabla<-table(titanic_ana$Pclass, titanic_ana$Survived)
tabla
chisq.test(tabla)

#¿Existen diferencias significativas en la supervivencia por sexo?´
tabla<-table(titanic_ana$Sex, titanic_ana$Survived)
tabla
chisq.test(tabla)

#¿Existen diferencias significativas en la supervivencia por puerto de embarque?
tabla<-table(titanic_ana$Embarked, titanic_ana$Survived)
tabla
chisq.test(tabla)

#Test normalidad edad
shapiro.test(titanic_ana$Age)

#Test de homocedasticidad
fligner.test(Age~Survived, data=titanic_ana)

#Constraste de hipótesis
wilcox.test(Age~Survived, data=titanic_ana)

#Test normalidad edad
shapiro.test(titanic_ana$TamFamilia)

#Test de homocedasticidad
fligner.test(TamFamilia~Survived, data=titanic_ana)

#Constraste de hipótesis
wilcox.test(TamFamilia~Survived, data=titanic_ana)

#Modelo de regresión logística
modelo = glm(Survived~., data=titanic_ana, family="binomial")
summary(modelo)

#Comparación de modelos
modelo1 = glm(Survived~Pclass, data=titanic_ana, family="binomial")
modelo2 = glm(Survived~Pclass+Sex, data=titanic_ana, family="binomial")
modelo3 = glm(Survived~Pclass+Sex+Age, data=titanic_ana, family="binomial")
modelo4 = glm(Survived~Pclass+Sex+Age+TamFamilia, data=titanic_ana, family="binomial")
modelo5 = glm(Survived~Pclass+Sex+Age+TamFamilia+Embarked, data=titanic_ana,
family="binomial")
modelo1$aic
```

```
modelo2$aic
modelo3$aic
modelo4$aic
modelo5$aic
```

Contribuciones

Contribuciones	Firma
Investigación previa	**Juan Francisco Vallalta Rueda**
Redacción de las respuestas	**Juan Francisco Vallalta Rueda**
Desarrollo código	**Juan Francisco Vallalta Rueda**