Dataset: Titanic

Luis Manuel Camacho Puerma
Carla Comas García
14 de Abril del 2019

Contenido

Luis Camacho Puerma
Carla Comas García

1. Detalles de la actividad

1.1. Descripción

En esta actividad se elabora un caso práctico, consistente en el tratamiento de un conjunto de datos (en inglés, dataset), orientado a aprender a identificar los datos relevantes para un proyecto analítico y usar las herramientas de integración, limpieza, validación y análisis de las mismas.

1.2. Objetivos

Los objetivos que se persiguen mediante el desarrollo de esta actividad práctica son los siguientes:

- 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 proyecto 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.

1.3. Competencias

Así, las competencias del Máster en Data Science que se desarrollan son:

- 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 datos (integración, transformación, limpieza y validación) para su posterior análisis.

2. Resolución

2.1. Descripción del dataset

El conjunto de datos objeto de análisis se ha obtenido a partir del enlace de Kaggle proporcionado en el enunciado de la práctica. El dataset es el referente a los pasajeros del Titanic y está constituido por 12 características (columnas) que presentan 891 pasajeros (filas). Entre los campos de este conjunto de datos, encontramos los siguientes:

- **PassengerId:** (Variable #) número que identifica al pasajero.
- **Survived:** (Variable #) Si sobrevive al accidente o no.
- **Pclass:** (Variable #) clase en la que viaja el pasajero (1ª, 2ª o 3ª).
- **Name:** (Variable Categórica) nombre del pasajero
- **Sex:** (Variable Categórica) sexo del pasajero.
- **Age:** (Variable #) edad del pasajero.
- **SibSp:** (Variable Categórica) número de relaciones familiares en el barco → Hermanos o esposa/amante
- **Parch**: (Variable #) número de relaciones familiares en el barco → Padres o hijos (0 para niños que viajan con niñera).
- **Ticket:** (Variable Categórica) número del ticket
- **Fare:** (Variable #) tarifa del pasajero.
- **Cabin:** (Variable Categórica) número de la cabina donde se aloja.
- **Embarked:** (Variable Categórica) lugar donde embarco el pasajero (C- Cherbourg, S - Southampton, Q = Queenstown).

2.2 Importancia y objetivos del análisis

A partir de este conjunto de datos se plantea saber si existe alguna variable que inluyera en la supervivencia o no del titanic.
Este análisis tendría una gran importancia para una empresa cuyo negocio sea realizar viajes en barco. Esto influiría en el coste de los billetes según el lugar en el que se decida comprar. Para ello nos proponemos a estudiar los datos descritos anteriormente, con el fin de obtener algún resultado remarcable.

2.3 Limpieza de los datos

Antes de comenzar con la limpieza de los datos, procedemos a realizar la lectura del fichero en formato CSV en el que se encuentran. El resultado devuelto por la llamada a la función read.csv() será un objeto data.frame:

```
> df <- read.csv("train.csv", header = TRUE)
> head(df[,1:5])
  PassengerId Survived Pclass                                                Name    Sex
1           1        0      3                             Braund, Mr. Owen Harris   male
2           2        1      1 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female
3           3        1      3                              Heikkinen, Miss. Laina female
4           4        1      1        Futrelle, Mrs. Jacques Heath (Lily May Peel) female
5           5        0      3                            Allen, Mr. William Henry   male
6           6        0      3                                    Moran, Mr. James   male
>
```

Pero para empezar a estudiar el dataset debemos tener una idea clara de que tipo de variables tratamos y que unidades tienen. Una vez sabidos estos nos interesará cuantos datos absentas existen y como están los datos repartidos, para ello ejecutaremos la función summary.

```
> sapply(df, function(x) class(x))
PassengerId    Survived      Pclass        Name         Sex         Age       SibSp       Parch
  "integer"   "integer"   "integer"    "factor"    "factor"   "numeric"   "integer"   "integer"
      Ticket        Fare       Cabin    Embarked
    "factor"   "numeric"    "factor"    "factor"
> summary(df)
  PassengerId       Survived          Pclass                                          Name
 Min.   :  1.0   Min.   :0.0000   Min.   :1.000   Abbing, Mr. Anthony                    :  1
 1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000   Abbott, Mr. Rossmore Edward            :  1
 Median :446.0   Median :0.0000   Median :3.000   Abbott, Mrs. Stanton (Rosa Hunt)       :  1
 Mean   :446.0   Mean   :0.3838   Mean   :2.309   Abelson, Mr. Samuel                    :  1
 3rd Qu.:668.5   3rd Qu.:1.0000   3rd Qu.:3.000   Abelson, Mrs. Samuel (Hannah Wizosky):  1
 Max.   :891.0   Max.   :1.0000   Max.   :3.000   Adahl, Mr. Mauritz Nils Martin         :  1
                                                  (Other)                                :885
     Sex           Age              SibSp           Parch           Ticket         Fare
 female:314   Min.   : 0.42   Min.   :0.000   Min.   :0.0000   1601    :  7   Min.   :  0.00
 male  :577   1st Qu.:20.12   1st Qu.:0.000   1st Qu.:0.0000   347082  :  7   1st Qu.:  7.91
              Median :28.00   Median :0.000   Median :0.0000   CA. 2343:  7   Median : 14.45
              Mean   :29.70   Mean   :0.523   Mean   :0.3816   3101295 :  6   Mean   : 32.20
              3rd Qu.:38.00   3rd Qu.:1.000   3rd Qu.:0.0000   347088  :  6   3rd Qu.: 31.00
              Max.   :80.00   Max.   :8.000   Max.   :6.0000   CA 2144 :  6   Max.   :512.33
              NA's   :177                                      (Other) :852
        Cabin     Embarked
            :687     :  2
 B96 B98    :  4   C:168
 C23 C25 C27:  4   Q: 77
 G6         :  4   S:644
 C22 C26    :  3
 D          :  3
 (Other)    :186
```

2.3.1 Selección de datos de interés

La gran mayoría de los atributos presentes en el conjunto de datos se corresponden con características que reúnen las diversas películas recogidos en forma de registros, por lo que será conveniente tenerlos en consideración durante la realización de los análisis.
En este apartado podemos comprobar todas las variables tienen un significado relativo a la variable endógena es decir que no nos encontramos con ningún casó de único valor o que no nos aporte información sobre el estudio.

2.3.2 Ceros y elementos vacíos

Comúnmente, se utilizan los ceros como centinela para indicar la ausencia de ciertos valores. Sin embargo, no es el caso de este conjunto de datos puesto que, como se comentó, se utilizó NA para denotar un valor desconocido. Así, se procede a conocer a continuación qué campos contienen elementos vacíos:

```
> sapply(df, function(x) sum(is.na(x)))
PassengerId    Survived      Pclass        Name         Sex         Age       SibSp       Parch
          0           0           0           0           0         177           0           0
     Ticket        Fare       Cabin    Embarked
          0           0           0           0
```

Llegados a este punto debemos decidir cómo manejar estos registros que contienen valores desconocidos para algún campo. Una opción podría ser eliminar esos registros que incluyen este tipo de valores, pero ello supondría desaprovechar información.

La función pertinente sería: Df_sinNA <- na.omit(df)

Como alternativa, se empleará un método de imputación de valores basado en la similitud o diferencia entre los registros: la imputación basada en k vecinos más próximos (en inglés, kNN-imputation). La elección de esta alternativa se realiza bajo la hipótesis de que nuestros registros guardan cierta relación. No obstante, es mejor trabajar con datos "aproximados" que con los propios elementos vacíos, ya que obtendremos análisis con menor margen de error.

2.3.3 Valores extremos

Los valores extremos o outliers son aquellos que parecen no ser congruentes sin los comparamos con el resto de los datos. Para identificarlos, podemos hacer uso de dos vías:

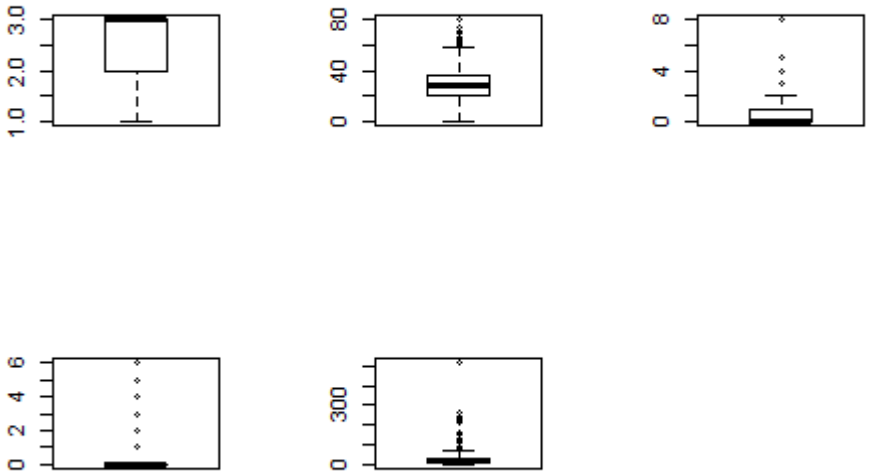
a) representar un diagrama de caja por cada variable y ver qué valores distan mucho del rango intercuartílico (la caja)
b) utilizar la función boxplots.stats() de R, la cual se emplea a continuación. Así, se mostrarán sólo los valores atípicos para aquellas variables que los contienen:

```
> boxplot.stats(df$Age)$out
 [1] 66.0 65.0 59.0 71.0 70.5 61.0 61.0 59.0 62.0 63.0 65.0 61.0 60.0 64.0 65.0 63.0 71.0 64.0 62.0
[20] 62.0 60.0 61.0 61.0 80.0 60.0 70.0 60.0 60.0 70.0 62.0 74.0
> boxplot.stats(df$SibSp)$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 8 4 3 4 8 4 8
> boxplot.stats(df$Age)$out
 [1] 66.0 65.0 59.0 71.0 70.5 61.0 61.0 59.0 62.0 63.0 65.0 61.0 60.0 64.0 65.0 63.0 71.0 64.0 62.0
[20] 62.0 60.0 61.0 61.0 80.0 60.0 70.0 60.0 60.0 70.0 62.0 74.0
> boxplot.stats(df$SibSp)$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 8 4 3 4 8 4 8
> boxplot.stats(df$Parch)$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 2 1 2 1 4 1 1 1 1 2 2
 [48] 1 2 1 1 1 2 1 1 2 2 2 1 1 2 2 1 2 1 1 1 1 1 1 1 2 1 2 2 1 1 2 1 1 2 1 1 1 2 1 1 1 4 1 1 2 2
 [95] 2 2 2 1 1 1 2 2 1 1 2 2 3 4 1 2 1 1 2 1 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
[142] 1 1 2 5 2 1 1 1 2 1 5 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
[189] 1 2 1 2 1 1 1 2 1 1 2 1 2 1 1 1 1 3 2 1 1 1 1 5 2
> boxplot.stats(df$Fare)$out
  [1]  71.2833 263.0000 146.5208  82.1708  76.7292  80.0000  83.4750  73.5000 263.0000  77.2875
 [11] 247.5208  73.5000  77.2875  79.2000  66.6000  69.5500  69.5500 146.5208  69.5500 113.2750
 [21]  76.2917  90.0000  83.4750  90.0000  79.2000  86.5000 512.3292  79.6500 153.4625 135.6333
 [31]  77.9583  78.8500  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 263.0000  75.2500
 [51]  69.3000 135.6333  82.1708 211.5000 227.5250  73.5000 120.0000 113.2750  90.0000 120.0000
 [61] 263.0000  81.8583  89.1042  91.0792  90.0000  78.2667 151.5500  86.5000 108.9000  93.5000
 [71] 221.7792 106.4250  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  73.5000 512.3292
 [91]  76.7292 211.3375 110.8833 227.5250 151.5500 227.5250 211.3375 512.3292  78.8500 262.3750
[101]  71.0000  86.5000 120.0000  77.9583 211.3375  79.2000  69.5500 120.0000  93.5000  80.0000
[111]  83.1583  69.5500  89.1042 164.8667  69.5500  83.1583
> boxplot.stats(df$Pclass)$out
integer(0)
```

Para poder entender y ver gráficamente…

```
## VALORES EXTREMOS
par(mfrow=c(2,3))

boxplot(df$Pclass)
boxplot(df$Age)
boxplot(df$SibSp)
boxplot(df$Parch)
boxplot(df$Fare)
```



No obstante, si revisamos los anteriores datos, comprobamos que son valores que perfectamente pueden darse. Es por ello que el manejo de estos valores extremos consistirá en simplemente dejarlos como actualmente están recogidos.

2.3.4 Valores categóricas

Una vez ya hemos hecho la limpieza de datos se pasa a estudiar nuestro dataset. Normalmente y por lo que nos dice nuestra experiencia los datos acostumbran a trabajar con datos y clientes normales es que no podemos perder ningún tipo de información. Es por este motivo que hemos decidido crear este nuevo apartado para ver como tratamos las variables categóricas.

Para ello hemos identificado cuales son y las hemos convertido a numéricas. Por lo que nuestra experiencia indica, cualquier modelo que intentes aplicar va a dar mejores resultados en caso de tener variables numéricas en vez de categóricas.

Por otra parte eliminamos las dos variables que no nos indica nada sobre la variable que definiremos endógena (el nombre y el id).

Para acabar hacemos una transformación de dos variables que parecen seguir el mismo tipo que el nombre pero que enrealidad nos podrían aportar algo. Estas variables son el número de ticket y el número de cabina. Para el número de cabina vamos a numerar en valor numérico que cabinas existen. Lo que hacemos es coger las últimos dos valores y con esto es suficiente para diferenciarlas entre ellas (no existen más de una cabina que acaben igual).
Por otro lado, el número de ticket veremos si existen supersticiones sobre lo que podría ser el número 13 por ejemplo.

Se ha planteado usar dummies para categorías como el tipo de clase para poder extraer información más concreta sobre que clase pertenece y no sobre que significa este valor. De todas maneras no parece ser relevante para este tipo de análisis de dependencia. En caso de querer ejecutar modelo si seria necesario aplicar dummies y multiplicar las columnas según los factores que tengan.

```
substrRight <- function(x, n){
     sapply(x, function(xx)
          substr(xx, (nchar(xx)-n+1), nchar(xx))
     )
}

df$Sex<-as.numeric(df$Sex)
df$Embarked<-as.numeric(df$Embarked)

#DELETE VAR OF df$Cabin -> NO EXISTEN VALORES REPETIDOS.
#<- cbind(df, dummy(df$Cabin))


df$Cabin<-as.character(df$Cabin)
df$Cabin<-substrRight((df$Cabin),2)
df$Cabin<-as.numeric(df$Cabin)

df$Ticket<-as.character(df$Ticket)
df$Ticket<-substrRight((df$Ticket),2)
df$Ticket<-as.numeric(df$Ticket)


df$PassengerId <- NULL
df$Name <- NULL

sapply(df, function(x) class(x))
```

2.3.5 Guardar preprocesado

Una vez que hemos acometido sobre el conjunto de datos inicial los procedimientos de integración, validación y limpieza anteriores, procedemos a guardar estos en un nuevo fichero denominado Automobile_data_clean.csv:

```
# Exportación de los datos limpios en .csv
write.csv(df, "titanic_clean.csv")
```

2.4 Análisis de los datos

2.4.1 Selección del grupos de datos

A continuación, se seleccionan los grupos dentro de nuestro conjunto de datos que pueden resultar interesantes para analizar y/o comparar. No obstante, como se verá en el apartado consistente en la realización de pruebas estadísticas, no todos se utilizarán.

```
# Agrupación por tipo de pclass
df.pclass1 <- df[df$pclass.type == 1,]
df.pclass2 <- df[df$pclass.type == 2,]
df.pclass3 <- df[df$pclass.type == 3,]

df.male <- df[df$Sex.type == 'male',]
df.famale <- df[df$Sex.type == 'female',]

df.EmbarkedS <- df[df$Embarked.type == 'S',]
df.EmbarkedQ <- df[df$Embarked.type == 'Q',]
df.EmbarkedC <- df[df$Embarked.type == 'C',]
```

2.4.2 Comprovació de la normalitat i homogeneïtat de la variància

Para la comprobación de que los valores que toman nuestras variables cuantitativas provienen de una población distribuida normalmente, utilizaremos la prueba de normalidad de Anderson Darling. Así, se comprueba que para que cada prueba se obtiene un p-valor superior al nivel de significación prefijado α = 0,05. Si esto se cumple, entonces se considera que variable en cuestión sigue una distribución normal.

```
> library(nortest)
> alpha = 0.05
> col.names = colnames(df)
> for (i in 1:ncol(df)) {
+   if (i == 1)
+     cat("Variables que no siguen una distribución normal:\n")
+   if (is.integer(df[,i]) | is.numeric(df[,i])) {
+     p_val = ad.test(df[,i])$p.value
+     if (p_val < alpha) { cat(col.names[i])
+   # Format output
+       if (i < ncol(df) - 1) cat(", ")
+       if (i %% 3 == 0) cat("\n")
+     }
+   }
+ }
Variables que no siguen una distribución normal:
PassengerId, Survived, Pclass,
Age,
SibSp, Parch, Fare,
```

Debido a que nuestras variables no siguen una normal, y completamente esperable debido a que muchas de ellas provienen de factores, pasamos a estudiar la homogeneidad de varianzas mediante la aplicación de un test de Fligner-Killeen. En el siguiente test, la hipótesis nula consiste en que ambas varianzas son iguales.

```
> for(i in 2:ncol(df)){
+   a<-fligner.test(Survived ~ df[,i], data = df)
+   if( (a$p.value) > 0.05){
+     b<-"variancia homogéneas"
+   } else{
+     b<-"variancia heterogenea"
+   }
+   print(c(colnames(df[i]), (a$p.value),b ))
+ }
[1] "Pclass"                "1.71184236923787e-08"  "variancia heterogenea"
[1] "Sex"                   "0.0162747937217524"    "variancia heterogenea"
[1] "Age"                   "0.659325452470611"     "variancia homogéneas"
[1] "SibSp"                 "0.00129846728433547"   "variancia heterogenea"
[1] "Parch"                 "0.00846983277933968"   "variancia heterogenea"
[1] "Ticket"                "0.785212501503716"     "variancia homogéneas"
[1] "Fare"                  "0.299019227419702"     "variancia homogéneas"
[1] "Cabin"                 "0.244495887027863"     "variancia homogéneas"
[1] "Embarked"              "0.0398735792520022"    "variancia heterogenea"
>
```

Puesto que obtenemos un p-valor superior a 0,05, aceptamos la hipótesis de que las varianzas de las muestras serán homogéneas.
En este caso las variables Pclass Sex SibSp Parcg Embarket tendrán variancias heterogéneas.

2.4.3 Aplicació de proves estadístiques

2.4.3.1 Análisis de correlacions

En primer lugar, procedemos a realizar un análisis de correlación entre las distintas variables para determinar cuáles de ellas ejercen una mayor influencia sobre el si la persona va a sobrevivir finalmente. Para ello, se utilizará el coeficiente de correlación de Spearman, puesto que hemos visto que tenemos datos que no siguen una distribución normal.

```
> corr_matrix <- matrix(nc = 2, nr = 0)
> colnames(corr_matrix) <- c("estimate", "p-value")
> for (i in 2:(ncol(df))) {
+   if (is.integer(df[,i]) | is.numeric(df[,i])) {
+     spearman_test = cor.test(df[,i], df[,1], method = "spearman")
+     corr_coef = spearman_test$estimate
+     p_val = spearman_test$p.value
+ # Add row to matrix
+     pair = matrix(ncol = 2, nrow = 1)
+     pair[1][1] = corr_coef
+     pair[2][1] = p_val
+     corr_matrix <- rbind(corr_matrix, pair)
+     rownames(corr_matrix)[nrow(corr_matrix)] <- colnames(df)[i]
+ }
+ }
```

```
> print(corr_matrix)
             estimate      p-value
Pclass    -0.33966794 1.687608e-25
Sex       -0.54335138 1.406066e-69
Age       -0.03930501 2.411787e-01
SibSp      0.08887948 7.941431e-03
Parch      0.13826563 3.453591e-05
Ticket     0.01338751 6.905050e-01
Fare       0.32373614 3.471228e-23
Cabin     -0.08008847 2.906846e-01
Embarked -0.16741298 5.015763e-07
>
```

Así, identificamos cuáles son las variables más correlacionadas con el precio en función de su proximidad con los valores -1 y +1. Teniendo esto en cuenta, queda patente cómo la variable más relevante en la causa de supervivencia es el sexo de la persona (Sex). Nota. Para cada coeficiente de correlación se muestra también su p-valor asociado, puesto que éste puede dar información acerca del peso estadístico de la correlación obtenida.

2.4.3.2 Contraste de hipótesis de dos muestras sobre la diferencia de medias

La segunda prueba estadística que se aplicará consistirá en un contraste de hipótesis sobre dos muestras para determinar si la probabilidad de supervivencia es superior en caso de que el viajero esté en primera clase. Para ello, tendremos dos muestras: la primera de ellas se corresponderá con los a la clase más alta y la segunda a las clases más bajas.

Se debe destacar que un test paramétrico como el que a continuación se utiliza necesita que los datos sean normales, si la muestra es de tamaño inferior a 30. Como en nuestro caso, n > 30, el contraste de hipótesis siguiente es válido.

Así, se plantea el siguiente contraste de hipótesis de dos muestras sobre la diferencia de medias, el cual es unilateral atendiendo a la formulación de la hipótesis alternativa:

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

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

dónde μ1 es la media de la población de la que se extrae la primera muestra y μ2 es la media de la población de la que extrae la segunda. Así, tomaremos α =0,05.

```
> df.1.Survived <-  df[df$Pclass == 1,]$Survived
> df.2.Survived <-  df[df$Pclass == 2 | df$Pclass == 3,]$Survived
> t.test(df.1.Survived, df.2.Survived, alternative = "less")

        Welch Two Sample t-test

data:  df.1.Survived and df.2.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
```

Puesto que obtenemos un p-valor mayor que el valor de significación fijado, no rechazamos la hipótesis nula. Por tanto, podemos concluir que, la clase de categoría no influyó sobre la supervivencia de los viajeros.

2.4.3.3 Modelo de regresión lineal

Tal y como se planteó en los objetivos de la actividad, resultará de mucho interés poder realizar si la persona va a sobrevivir dependiendo de sus características. Así, se calculará un modelo de regresión lineal utilizando regresores tanto cuantitativos con el que poder realizar las predicciones. Para obtener un modelo de regresión lineal considerablemente eficiente, lo que haremos será obtener varios modelos de regresión utilizando las variables que estén más correladas con respecto a la supervivencia debido a que es lo que queremos predecir, según la tabla obtenida anteriormente. Así, de entre todos los modelos que tengamos, escogeremos el mejor utilizando como criterio aquel que presente un mayor coeficiente de determinación (R2).

```
# Generación de varios modelos
modelo1 <- lm(Survived ~ Pclass + Sex , data = df)
modelo2 <- lm(Survived ~ Pclass + Sex + Parch + Fare , data = df)
modelo3 <- lm(Survived ~ Pclass + Sex + Parch + Fare + Embarked, data = df)
modelo4 <- lm(Survived ~ Pclass + Sex + Parch + Fare + Embarked + Parch, data = df)
modelo5 <- lm(Survived ~ ., data = df)

# Tabla con los coeficientes de determinación de cada modelo
tabla.coeficientes <- matrix(c(1, summary(modelo1)$r.squared, 2,
                               summary(modelo2)$r.squared, 3,
                               summary(modelo3)$r.squared, 4,
                               summary(modelo4)$r.squared, 5,
                               summary(modelo5)$r.squared), ncol = 2, byrow = TRUE)
colnames(tabla.coeficientes) <- c("Modelo", "R^2")
tabla.coeficientes

208:22  (Untitled)

Console ~/Documentos/Master/Tipologia i cicle de les dades/titanic/
> colnames(tabla.coeficientes) <- c("Modelo", "R^2")
> tabla.coeficientes
     Modelo       R^2
[1,]      1 0.3676802
[2,]      2 0.3700831
[3,]      3 0.3741774
[4,]      4 0.3741774
[5,]      5 0.3714153
```

Para los anteriores modelos de regresión lineal múltiple obtenidos, podemos utilizar el coeficiente de determinación para medir la bondad de los ajustes y quedarnos con aquel modelo que mejor coeficiente presente.

En este caso, nos quedaremos con el modelo 3, debido a que la diferencia entre el modelo 3 y 4 el coeficiente de determinación es el mismo, pero más óptimo el 3 debido a que hay menos variables exógenas. Lo importante es saber que variables son más relevantes. Que en este caos son: Pclass, Sex, Parch, Fare y Embarked.

Hemos dejado un apartado en R dónde completar las variables en caso de querer sacar la predicción de un cierto viajero. De todas maneras no tiene sentido si no se aplica en el futuro debido a que ya se sabe que viajeros se salvaron y que características cumplían.

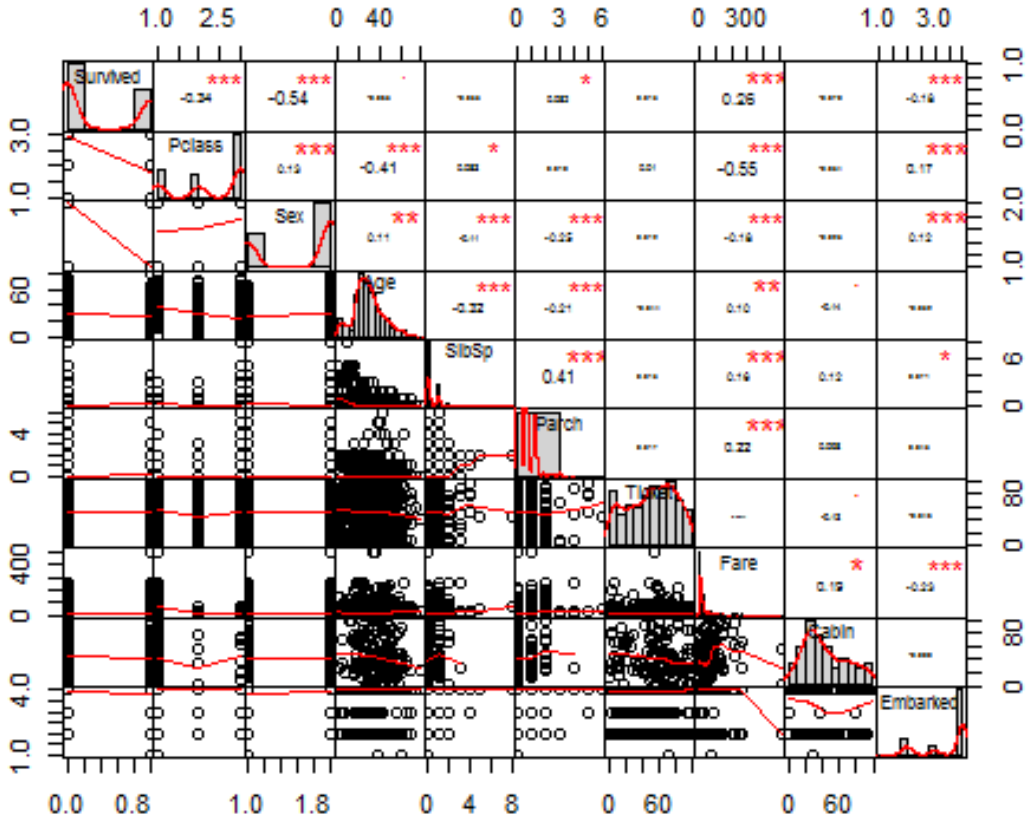
** Se podría haber hecho un análisis de componentes principales que pudiera haber sido más interesante para nuestro tipo de dataset, debido a que probablemente una combinación de las variables expliquen más que ellas individualmente.

2.5 Representación de resultados

2.5.1 Gráfico entre las variables

Al estudiar una variable nos interesa que tenga la máxima correlación cada variable exógena con la variable endógena y además que entre las variables exógenas haya la mínima correlación.
Para ello hemos creado el siguiente gráfico que nos muestra en número a la parte superior derecha la correlación que existe marcando con asteriscos si esta es significativa.

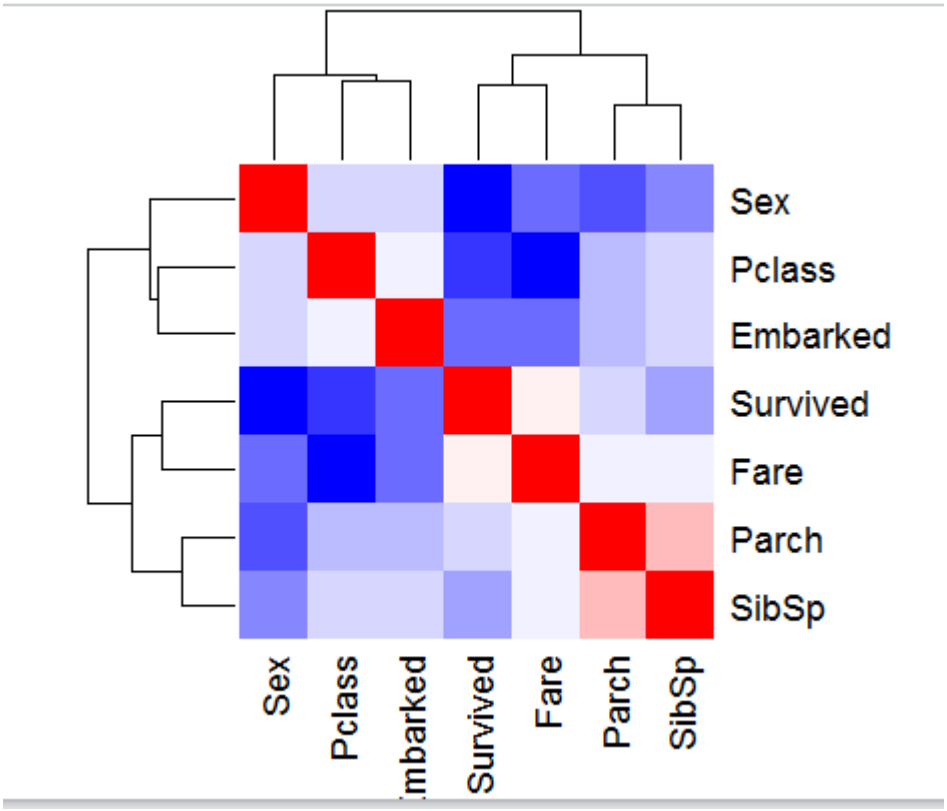
La parte izquierda inferior podemos ver los gráficos de cómo evoluciona cada una de las variables con la otra. Mientras que en la diagonal se puede ver que un estudio de que cantidad de registros (viajeros en nuestro caso) de esa categoría existe.



Confirmamos y justificamos todo lo que se ha comentado anteriormente con este gráfico.

2.5.2 Gráfico de correlación entre las varibales

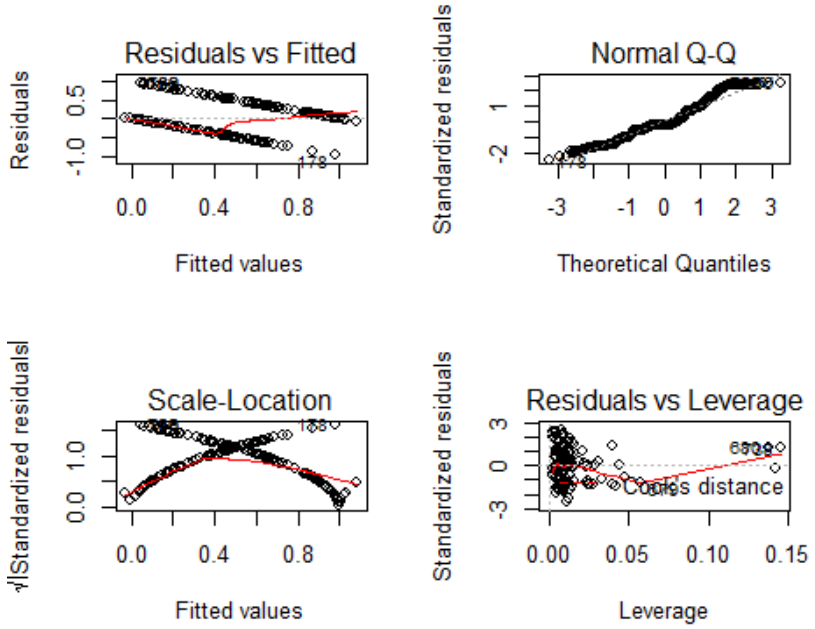
Seguidamente mostramos un gráfico de correlaciones más simple para poder comprobar que no hemos cogido variables muy correlacionadas. Se observa como el color rojo es el más correlacionada mientras que el máximo azul también pero en negativo; a medida que se acerca la correlación a nula, más claro será el color.



Las líneas superiores nos puede servir en caso de hacer un ACP pero no es lo que se ha estudiado y por lo tanto suprimimos esta parte de explicación.

2.5.3 Gráfico de la regresión

Para acabar se muestra el gráfico de la regresión que se ha hecho anteriormente:



2.6 Conclusiones

Se han realizado tres tipos de pruebas estadísticas sobre el conjunto de datos de los pasajeros del titanic con el objetivo de conocer si ciertas variables influían o no en la supervivencia de estos. Para cada una de las pruebas hemos podido comprobar los resultados, en algunos casos mediante tablas y en otros mediante una gráfica.

Mediante los diferentes análisis hemos podido determinar que las variables más significativas para determinar la supervivencia son el sexo, la clase y la tarifa de los pasajeros. Se podría no tener en cuenta la tarifa puesto que esta relacionada con la clase de estos. Estas variables influyen en gran medida en la supervivencia como se puede observar en la grafica entre las variables.

Para este estudio se han sometido los datos a un preprocesamiento para poder manejar los elementos vacíos y los valores extremos (outliers). Para el caso de los valores vacíos se ha hecho uso de un método de imputación de valores de tal forma que no tengamos que eliminar registros. Para el caso de los outliers se ha considerado no eliminarlos puesto que no parecen resultar del todo atípicos.

3. Recursos

- Megan Squire (2015). Clean Data. Packt Publishing Ltd.
- Jason W. Osborne (2010). Data Cleaning Basics: Best Practices in Dealing with Extreme Scores. Newborn and Infant Nursing Reviews; 10 (1): pp. 1527-3369.

4. Tabla contribuciones

Contribuciones	Firma
Investigación Previa	CCG – LMCP
Documento Pdf	CCG – LMCP
Desarrollo Código	CCG – LMCP
GitHub	CCG – LMCP