Práctica 2:

Limpieza y validación de los datos

Leandro Alonso Von Semasco

11 De junio de 2018

Link_Repositorio

1. Detalles de la actividad.

1.1 Descripción.

En esta actividad vamos a analizar un dataset relativo a los datos de los pasajeros del Titanic cuando este se hundió, teniendo en cuenta tanto datos propios del pasajero, características en el barco y si sobrevivieron al accidente.

La práctica está desarrollada en R cuyo fichero se encuentra en el mismo repositorio. Los pasos detallados desde la carga hasta los análisis estadísticos en detalle se encuentran tanto en el fichero con el código como en el HTML resultante de su ejecución.

(Repositorio: https://github.com/leovs21/Prac-2-Tipologia-y-ciclo-de-vida-de-los-datos)

1.2 Importancia.

Una de los motivos importantes de analizar un conjunto de datos semejante radicaría en intentar comprender el flujo de acciones de la gente ante esta situación en el contexto en el que se encontraban, tanto de la época o con la distinción de clases que había en el momento.

1.3 Problema a resolver.

Como se vio afectado los supervivientes por su condición.

2. Descripción del Dataset.

El dataset utilizado en la activad se encuentra en la url: https://www.kaggle.com/c/titanic/data consta de dos archivos .csv, uno de entrenamiento y otro de test para un predictor sobre el fin que tendría una persona concreta en una situación similar. El archivo de entrenamiento está constituido por 12 variables:

- PassengerId: Identificador del pasajero.
- Survived: un valor binario si sobrevivió o no (0->No, 1->Si).
- Pclass: Clase en la que viajaba (1 = 1st, 2 = 2nd, 3 = 3rd).
- Name: Nombre del pasajero.
- Sex: sexo del pasajero.
- Age: Edad del pasajero.
- SibSp: número de Hermano/a, esposo/a a bordo.
- Parch: número de padres, hijos a bordo del barco.
- Ticket: número de ticket de barco.
- Fare: tarifa de embarque.
- Cabin: número de cabina.
- Embarked: puerta de embarque.

Teniendo un total de 821 registros.

3. Limpieza de los Datos.

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

Si, en el análisis del dataset se vio que en el campo correspondiente había 177 registros de los cuales estaban almacenados con NA, para evitar que el estudio se vea afectado por estos valores, optamos por la eliminación de los registros, pues si pusiéramos cualquier valor estaríamos condicionando el resultado.

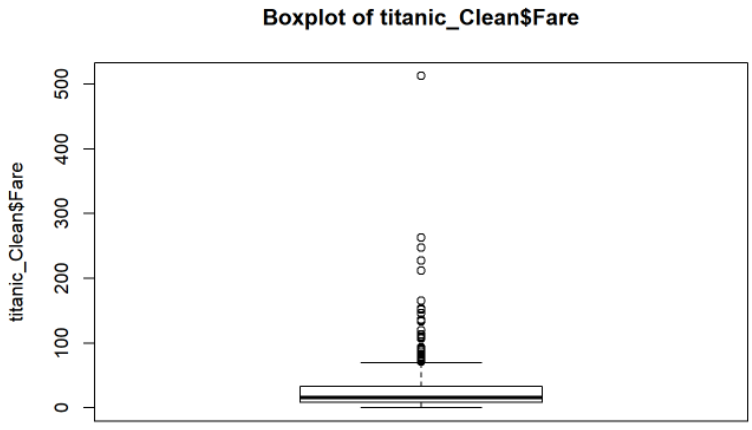
Otro campo con valores ausentes es el de la cabina, siendo una gran proporción de este en esta situación (687) demasiados para realizar la opción anterior, y como no es un dato que aporte una relevancia crucial optamos por suprimir esta columna.

Por último, en el campo de la tarifa se encuentran 15 registros con el valor 0, considerando estos valores debido a que fuesen personal de la embarcación o invitados no se realizo ninguna acción al respecto. Y en el campo embarked después del tratamiento se visualizan dos registros con valores ausentes los cuales procedemos a eliminar.

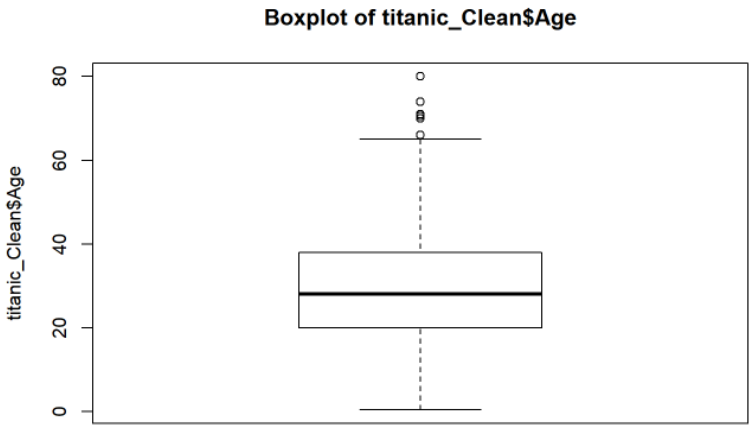
3.2. Identificación y tratamiento de valores extremos.

Una vez realizado el tratamiento de los datos estudiamos estos, en busca de outliers que resalten, los únicos campos que no son de tipo factor y por lo cual pueden destacar son el número de parientes, tanto el correspondiente a hermanos como el de progenitores e hijos; y los de edad y tarifa.

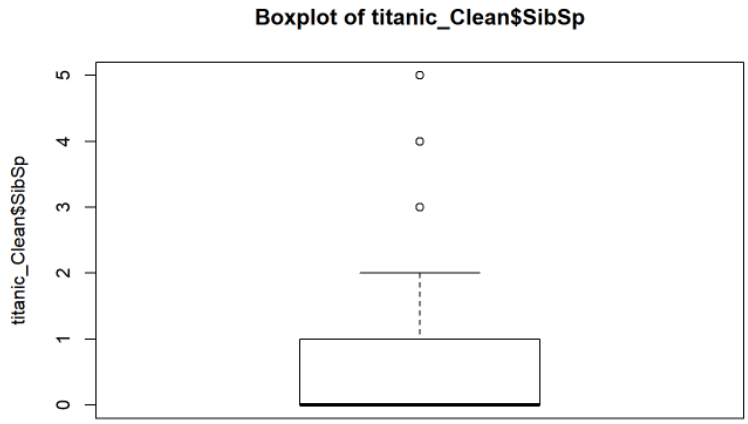
```
boxplot(titanic_Clean$Fare, main="Boxplot of titanic_Clean$Fare",ylab="titanic_Clean$Fare")
```



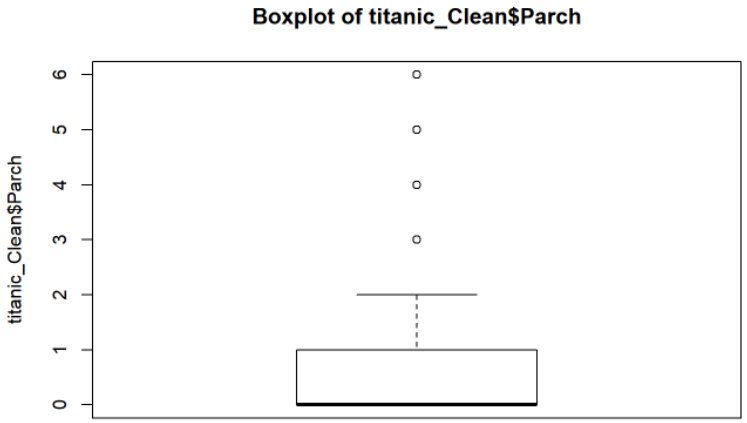
```
boxplot(titanic_Clean$Age, main="Boxplot of titanic_Clean$Age",ylab="titanic_Clean$Age")
```




```
boxplot(titanic_Clean$SibSp, main="Boxplot of titanic_Clean$SibSp",ylab="titanic_Clean$SibSp")
```




```
boxplot(titanic_Clean$Parch, main="Boxplot of titanic_Clean$Parch",ylab="titanic_Clean$Parch")
```



Aquellos ouliers correspondientes a parientes aceptamos los valores designados pues ninguno sobresale fuera de lo razonable. Respecto a la tarifa, algunos valores sobresalen en el costo,

pero viendo estos en detenimiento, se aprecia que todos ellos corresponden al mismo número de tiquet y que tienen asignados varios camarotes.

```
titanic[!titanic$Fare<400.0000,]
```

```
##     PassengerId Survived Pclass                                Name    Sex
## 259         259        1      1                    Ward, Miss. Anna female
## 680         680        1      1 Cardeza, Mr. Thomas Drake Martinez   male
## 738         738        1      1              Lesurer, Mr. Gustave J   male
##     Age SibSp Parch   Ticket     Fare       Cabin Embarked
## 259  35     0     0 PC 17755 512.3292                   C
## 680  36     0     1 PC 17755 512.3292 B51 B53 B55       C
## 738  35     0     0 PC 17755 512.3292        B101       C
```

4. Análisis de los datos.

4.1. Selección de los grupos de datos que se quieren analizar/comparar.

Después del tratamiento de los datos hemos conservado las columnas: Survived, Pclass, Sex, Age, SibSp, Parch, Fare y Embarked.

Convirtiendo la edad en rango de 4, niños, adultos jóvenes, adultos y jóvenes.

4.2. Comprobación de la normalidad y homogeneidad de la varianza.

Agrupación, quartiles y varianza de los datos.

```
titanic[!titanic$Fare<400.0000,]
```

```
##     PassengerId Survived Pclass                                Name    Sex
## 259         259        1      1                    Ward, Miss. Anna female
## 680         680        1      1 Cardeza, Mr. Thomas Drake Martinez   male
## 738         738        1      1              Lesurer, Mr. Gustave J   male
##     Age SibSp Parch   Ticket     Fare       Cabin Embarked
## 259  35     0     0 PC 17755 512.3292                   C
## 680  36     0     1 PC 17755 512.3292 B51 B53 B55        C
## 738  35     0     0 PC 17755 512.3292        B101        C
```

```
#Varianza de los datos.
var(titanic_Clean)
```

```
## Warning in var(titanic_Clean): NAs introducidos por coerción
```

```
##              Survived       Pclass Sex Age        SibSp       Parch
## Survived  0.241217466  -0.14650990  NA  NA -0.007095561 0.03996587
## Pclass   -0.146509901   0.70032515  NA  NA  0.050771188 0.01691714
## Sex                NA           NA  NA  NA           NA         NA
## Age                NA           NA  NA  NA           NA         NA
## SibSp    -0.007095561   0.05077119  NA  NA  0.866187835 0.30474565
## Parch     0.039965865   0.01691714  NA  NA  0.304745650 0.72962594
## Fare      6.918651330 -24.49424021  NA  NA  6.890867913 9.34335917
## Embarked           NA           NA  NA  NA           NA         NA
##                 Fare Embarked
## Survived    6.918651       NA
## Pclass    -24.494240       NA
## Sex              NA       NA
## Age              NA       NA
## SibSp       6.890868       NA
## Parch       9.343359       NA
## Fare     2802.500471       NA
## Embarked          NA       NA
```

- Desviación típica.

```
#Desviacion tipica, Cuanto menor sea mayor homojeneidad habra en los datos
sd(titanic_Clean$Survived)
```

```
## Warning in var(if (is.vector(x) || is.factor(x)) x else as.double(x), na.rm = na.rm): Calling var(x) on a fa
ctor x is deprecated and will become an error.
##   Use something like 'all(duplicated(x)[-1L])' to test for a constant vector.
```

```
## [1] 0.4911389
```

```
sd(titanic_Clean$Pclass)
```

```
## Warning in var(if (is.vector(x) || is.factor(x)) x else as.double(x), na.rm = na.rm): Calling var(x) on a fa
ctor x is deprecated and will become an error.
##   Use something like 'all(duplicated(x)[-1L])' to test for a constant vector.
```

```
## [1] 0.8368543
```

```
sd(titanic$Age)
```

```
## [1] 14.49293
```

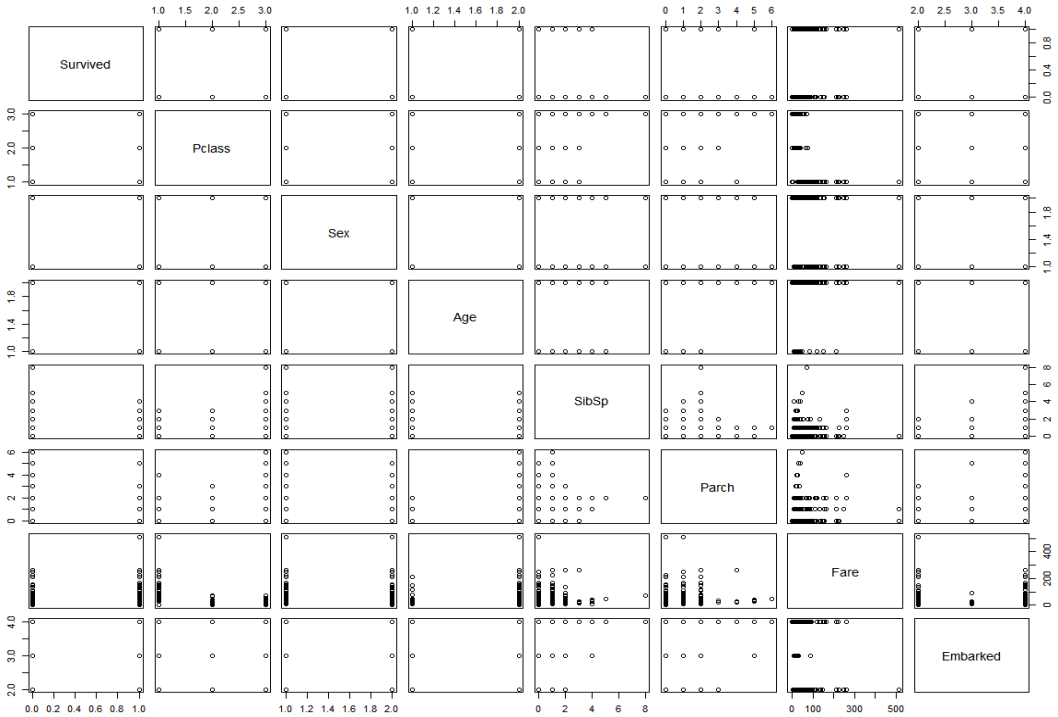
```
sd(titanic_Clean$Fare)
```

```
## [1] 52.93865
```

4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos.

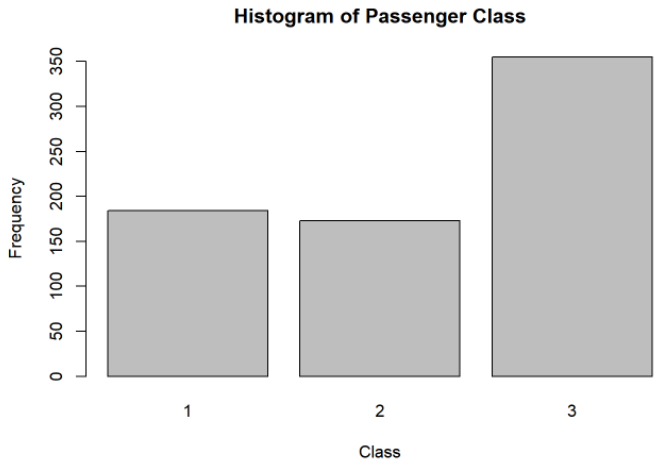
5. Representación de los resultados a partir de tablas y gráficas.

Para este caso la grafica de relación entre las distintas variables no tiene sentido, ya que al tener las agrupaciones, no hace una verdadera revelación su visualización en el análisis de los datos.



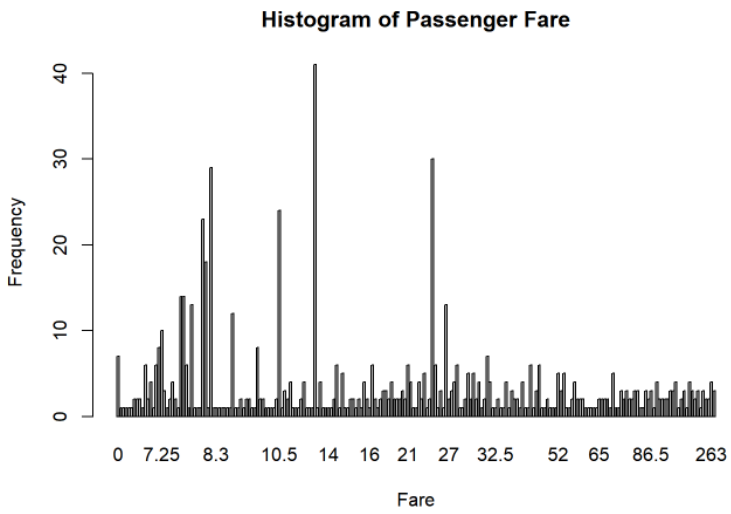
Histograma de la división de las clases a bordo del barco.

```
barplot(table(titanic_Clean$Pclass), xlab="Class", ylab="Frequency", main="Histogram of Passenger Class")
```



Histograma del precio de los billetes del barco.

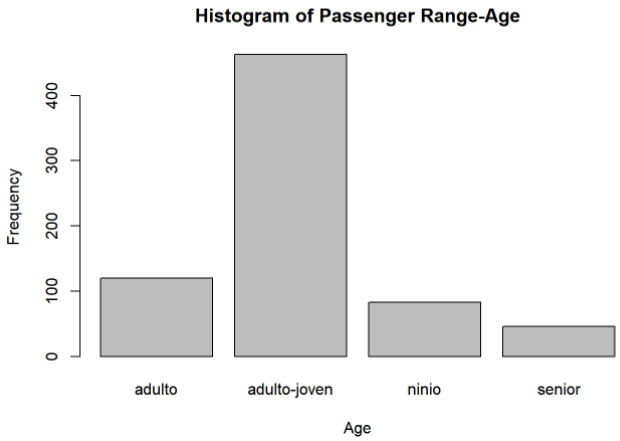
```
barplot(table(titanic_Clean$Fare), xlab="Fare", ylab="Frequency", main="Histogram of Passenger Fare")
```



Histogramas de la división de las edades a bordo del barco.

```
barplot(table(titanic$Age), xlab="Age", ylab="Frequency", main="Histogram of Passenger Age")
```

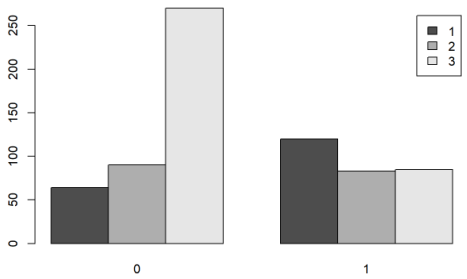
```
barplot(table(titanic_Clean$Age), xlab="Age", ylab="Frequency", main="Histogram of Passenger Range-Age")
```



Comparación de las distintas variables, con el hecho de sobrevivir. (Izquierda no sobrevivieron 0, derecha sobrevivieron 1)

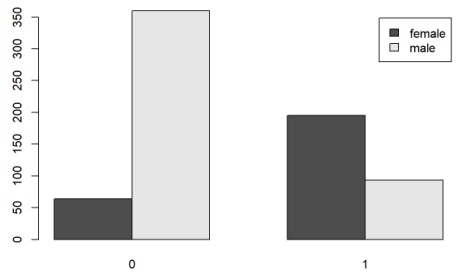
- Relación con las distintas clases abordo.

```
#Relacion de supervivientes en las distintas clases
barplot(table(titanic_Clean$Pclass, titanic_Clean$Survived),legend=TRUE,beside=TRUE )
```



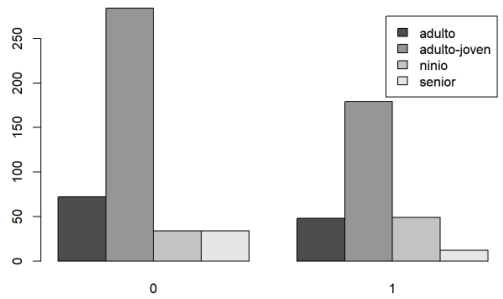
- Relación con los distintos sexos abordo.

```
#Relacion de supervivientes en los distintas sexos
barplot(table(titanic_Clean$Sex, titanic_Clean$Survived),legend=TRUE,beside=TRUE )
```



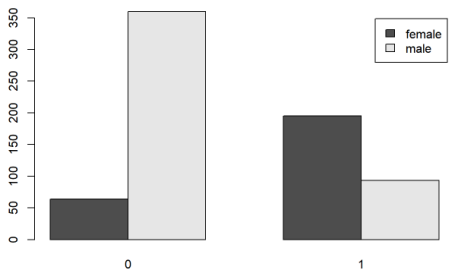
- Relación con las distintas edades abordo.

```
#Relacion de supervivientes en las distintas edades
barplot(table(titanic_Clean$Age, titanic_Clean$Survived),legend=TRUE,beside=TRUE )
```



- Relación con los distintos sexos abordo.

```
#Relacion de supervivientes en los distintas sexos
barplot(table(titanic_Clean$Sex, titanic_Clean$Survived),legend=TRUE,beside=TRUE )
```



El resto de graficas relacionales están en el fichero HTML con el código de R ejecutado, no se adjunta en este documento ya que el resultado no aporto un valor añadido al mismo.

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

Analizando los resultados se puede ver que el nivel económico de la gente afecto en el resultado, teniendo mayor suerte aquellos que viajaban en primera, además los parámetros que más efecto jugaron fueron principalmente el sexo, dando preferencia a las mujeres, y los niños, dejando en último lugar a las personas mayores.

7. Código: Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos. Si lo preferís, también podéis trabajar en Python

El proyecto subido al repositorio consiste del archivo de origen de los datos (train.csv), el archivo resultante de su tratamiento (titanic_Clean.csv), el fichero con el código en R

(PRAC2_LEANDRO_TD.RMD), el archivo HTML con la ejecución del código (PRAC2_LEANDRO_TD.html) y este documento explicativo.