Actividad: PRA2: Limpieza y análisis de datos

Reison A. Torres Urina

Diciembre 2019

Índice

1. Detalles de la actividad

1.1. Descripción

En esta practica se elabora un caso práctico 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 concretos de esta práctica son:

- Aprender a aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios o multidisciplinares.
- Saber identificar los datos relevantes y los tratamientos necesarios (integración, limpieza y validación) para llevar a cabo un 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

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

- Capacidad de analizar un problema en el nivel de abstracción adecuado a cada situación y aplicar las habilidades y conocimientos adquiridos para abordarlo y resolverlo.
- Capacidad para aplicar las técnicas específicas de tratamiento de datos (integración, transformación, limpieza y validación) para su posterior análisis.

2. Solución

2.1. Descripción del dataset

2.1.1. Carga de los datos

Cargamos el conjunto de datos que se encuentran en los archivos **train.csv**, **test.csv** y **gender_submission.csv** en formato CSV, y representan los datos de los pasajeros que abordaron el Titanic.

Estos datos estarán representados en R por un dataframe para facilitar la manipulación de los mismos en nuestro análisis.

```
# cargamos paquetes R que vamos a utilizar durante nuestro anlisis

if(!require(ggplot2)){
    #install.packages('ggplot2', repos='http://cran.us.r-project.org')
    library(ggplot2)
}

if(!require(ggpubr)){
    #install.packages('ggpubr', repos='http://cran.us.r-project.org')
    library(ggpubr)
}

library(dplyr)
#library(Hmisc)
#library(corrplot)

# Carga del dataset contenido en el archivo train.csv
titanic.train <- read.csv("../datos/train.csv",stringsAsFactors = FALSE, header=T, sep=",")
# Carga del dataset contenido en el archivo test.csv
titanic.test <- read.csv("../datos/test.csv",stringsAsFactors = FALSE, header=T, sep=",")
# Carga del dataset contenido en el archivo gender_submission.csv
titanic.test.survived <- read.csv("../datos/gender_submission.csv",stringsAsFactors = FALSE, header=T, s
```

2.1.2. Descripción

Los datos seleccionados, fueron obtenidos del sitio de data science, **www.Kaggle.com**, en el encontramos una variedad de dataset Open Data. El conjunto de datos seleccionados para desarrollar esta actividad es **Titanic: Machine Learning from Disaster**, en este dataset encontramos, los datos de los pasajeros, que abordaron el Titanic en su viaje inaugural.

Los datos de este dataset se encuentran divididos en dos archivos train.csv con 891 observaciones y test.cvs con 418 observaciones para un total de 1309. El conjunto de datos esta descripto por un conjunto de 12 variables. Las característica presenten en este dataset, nos permitirá cumplir los objetivos propuestos en esta actividad.

Variables contenidas en el dataset **train.csv**:

```
summary(titanic.train)
```

```
##   PassengerId       Survived          Pclass          Name
##  Min.   :  1.0   Min.   :0.0000   Min.   :1.000   Length:891
##  1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000   Class :character
##  Median :446.0   Median :0.0000   Median :3.000   Mode  :character
##  Mean   :446.0   Mean   :0.3838   Mean   :2.309
##  3rd Qu.:668.5   3rd Qu.:1.0000   3rd Qu.:3.000
##  Max.   :891.0   Max.   :1.0000   Max.   :3.000
##
##      Sex                 Age            SibSp           Parch
##  Length:891         Min.   : 0.42   Min.   :0.000   Min.   :0.0000
##  Class :character   1st Qu.:20.12   1st Qu.:0.000   1st Qu.:0.0000
##  Mode  :character   Median :28.00   Median :0.000   Median :0.0000
##                     Mean   :29.70   Mean   :0.523   Mean   :0.3816
##                     3rd Qu.:38.00   3rd Qu.:1.000   3rd Qu.:0.0000
##                     Max.   :80.00   Max.   :8.000   Max.   :6.0000
##                     NA's   :177
##     Ticket               Fare           Cabin             Embarked
##  Length:891         Min.   :  0.00   Length:891         Length:891
##  Class :character   1st Qu.:  7.91   Class :character   Class :character
##  Mode  :character   Median : 14.45   Mode  :character   Mode  :character
##                     Mean   : 32.20
##                     3rd Qu.: 31.00
##                     Max.   :512.33
##
```

Variables contenidas en el dataset **test.cvs**:

```
summary(titanic.test)
```

```
##   PassengerId         Pclass          Name               Sex
##  Min.   : 892.0   Min.   :1.000   Length:418         Length:418
##  1st Qu.: 996.2   1st Qu.:1.000   Class :character   Class :character
##  Median :1100.5   Median :3.000   Mode  :character   Mode  :character
##  Mean   :1100.5   Mean   :2.266
##  3rd Qu.:1204.8   3rd Qu.:3.000
##  Max.   :1309.0   Max.   :3.000
##
##       Age             SibSp           Parch           Ticket
##  Min.   : 0.17   Min.   :0.0000   Min.   :0.0000   Length:418
##  1st Qu.:21.00   1st Qu.:0.0000   1st Qu.:0.0000   Class :character
##  Median :27.00   Median :0.0000   Median :0.0000   Mode  :character
##  Mean   :30.27   Mean   :0.4474   Mean   :0.3923
##  3rd Qu.:39.00   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :76.00   Max.   :8.0000   Max.   :9.0000
##  NA's   :86
##       Fare            Cabin             Embarked
##  Min.   :  0.000   Length:418         Length:418
```

```
##  1st Qu.:  7.896   Class :character   Class :character
##  Median : 14.454   Mode  :character   Mode  :character
##  Mean   : 35.627
##  3rd Qu.: 31.500
##  Max.   :512.329
##  NA's   :1
```

Variables contenidas en el dataset **gender_submission.csv**:

```
summary(titanic.test.survived)
```

```
##   PassengerId        Survived
##  Min.   : 892.0   Min.   :0.0000
##  1st Qu.: 996.2   1st Qu.:0.0000
##  Median :1100.5   Median :0.0000
##  Mean   :1100.5   Mean   :0.3636
##  3rd Qu.:1204.8   3rd Qu.:1.0000
##  Max.   :1309.0   Max.   :1.0000
```

Este dataset **gender_submission.csv** contiene la variable de survived, que luego utilizaremos para agregar al dataset **titanic.test**.

A continuación describimos el conjunto de variables que conforman este dataset:

- **PassengerId:** Número consecutivo que identifica al pasajero.
- **Name:** Nombre del pasajero.
- **Sex:** Define el xeso del pasajero.
- **pclass:** Nivel socioeconómico del pasajero (1st = Upper, 2nd = Middle, 3rd = Lower).
- **age:** Edad del pasajero en años.
- **sibsp: Familiar abordo del Titanic. Define la relación familiar de la siguiente forma:**
  Hermanos => 1 = hermana, 2 = hermano, 3 = hermanastro, 4 = hermanastra
  Esposos => 5 = esposo, 6 = esposa, 7 = amantes y 8 = novio
- **parch: Familiar abordo del Titanic. Define la relación familiar de la siguiente forma:**
  Padres => 1 = madre, 2 = padre
  Hijos => 3 = hija, 4 = hijo, 5 = hijastra, 6 = hijastro
- **ticket:** Número del boleto de abordaje.
- **fare:** Precio del boleto.
- **cabin:** Número de la cabina.
- **embarked:** Puerto de embarque (C = Cherbourg, Q = Queenstown, S = Southampton).
- **survived:** Pasajero superviviente (0 = No, 1 = Yes)

2.1.3. Importancia y objetivos de los análisis

A partir de este conjunto de datos se plantea la problemática de determinar qué variables influyen más en la supervivencia de un pasajero en el naufragio del Titanic. Además, se podrá proceder a crear modelos de regresión que permitan predecir si un pasajero sobrevive o no en función de sus características y contrastes de hipótesis que ayuden a identificar propiedades interesantes en las muestras que puedan ser inferidas con respecto a la población.

Este tipo de análisis pueden ser utilizados por las aseguradoras del sector turístico, para determinar el riesgo que puede tener un turista al viajar en los trasatlánticos. Y asi poder ofreser las cobertura del seguro.

2.2. Integración y selección de los datos de interés a analizar

2.2.1. Integración

Con el fin de tener una estructura de datos coherente y única que contenga mayor cantidad de información, combinaremos los datos procedentes de los dataset train.csv y test.cvs. Luego realizaremos una fusión horizontal para añadir el atributo **survived**, debido a que el dataset **test.cvs** no presenta este atributo. Este valor será extraído del dataset **gender_submission.csv**.

```
# Realizamos una fusión horizontal entre los dataset titanic.test y titanic.test.survived para agregar
titanic.test <- inner_join(titanic.test, titanic.test.survived, by ="PassengerId")

#Creamos el dataset titanic.data con la combinacion de los datos de los dataset  titanic.train y titani
titanic.data <- bind_rows(titanic.train,titanic.test)

# Eliminamos los dataset temporales
rm(titanic.test.survived)
rm(titanic.test)
rm(titanic.train)

# Verificamos la estructura del dataset con los datos combinados
summary(titanic.data)
```

```
##   PassengerId      Survived          Pclass          Name
##  Min.   :   1   Min.   :0.0000   Min.   :1.000   Length:1309
##  1st Qu.: 328   1st Qu.:0.0000   1st Qu.:2.000   Class :character
##  Median : 655   Median :0.0000   Median :3.000   Mode  :character
##  Mean   : 655   Mean   :0.3774   Mean   :2.295
##  3rd Qu.: 982   3rd Qu.:1.0000   3rd Qu.:3.000
##  Max.   :1309   Max.   :1.0000   Max.   :3.000
##
##      Sex                 Age            SibSp            Parch
##  Length:1309        Min.   : 0.17   Min.   :0.0000   Min.   :0.000
##  Class :character   1st Qu.:21.00   1st Qu.:0.0000   1st Qu.:0.000
##  Mode  :character   Median :28.00   Median :0.0000   Median :0.000
##                     Mean   :29.88   Mean   :0.4989   Mean   :0.385
##                     3rd Qu.:39.00   3rd Qu.:1.0000   3rd Qu.:0.000
##                     Max.   :80.00   Max.   :8.0000   Max.   :9.000
##                     NA's   :263
##     Ticket               Fare            Cabin
##  Length:1309        Min.   :  0.000   Length:1309
##  Class :character   1st Qu.:  7.896   Class :character
##  Mode  :character   Median : 14.454   Mode  :character
```

```
##                     Mean   : 33.295
##                     3rd Qu.: 31.275
##                     Max.   :512.329
##                     NA's   :1
##    Embarked
##  Length:1309
##  Class :character
##  Mode  :character
##
##
##
##
```

2.2.2. Selección de los datos

La gran mayoría de las variables contenidas en el conjunto de datos corresponde con características de los pasajeros que abordaron el Titanic, por lo que serán tenidas en cuenta para realizar nuestro análisis. Sin embargo, podremos prescindir de las variables (**PassengerId**,**Name** y **Ticket**) dado que estos atributos no aportan una carasterisitica al pasajero, y no influye en la resolución de nuestro problema.

```
# Eliminamos del dataset las variables "PassengerId" y "Name"
titanic.data <- titanic.data[,!(colnames(titanic.data) %in% c("PassengerId","Name","Ticket"))]

# Verificamos la estructura del dataset
summary(titanic.data)
```

```
##     Survived          Pclass          Sex                 Age
##  Min.   :0.0000   Min.   :1.000   Length:1309        Min.   : 0.17
##  1st Qu.:0.0000   1st Qu.:2.000   Class :character   1st Qu.:21.00
##  Median :0.0000   Median :3.000   Mode  :character   Median :28.00
##  Mean   :0.3774   Mean   :2.295                      Mean   :29.88
##  3rd Qu.:1.0000   3rd Qu.:3.000                      3rd Qu.:39.00
##  Max.   :1.0000   Max.   :3.000                      Max.   :80.00
##                                                      NA's   :263
##      SibSp            Parch            Fare            Cabin
##  Min.   :0.0000   Min.   :0.000   Min.   :  0.000   Length:1309
##  1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:  7.896   Class :character
##  Median :0.0000   Median :0.000   Median : 14.454   Mode  :character
##  Mean   :0.4989   Mean   :0.385   Mean   : 33.295
##  3rd Qu.:1.0000   3rd Qu.:0.000   3rd Qu.: 31.275
##  Max.   :8.0000   Max.   :9.000   Max.   :512.329
##                                   NA's   :1
##    Embarked
##  Length:1309
##  Class :character
##  Mode  :character
##
##
##
##
```

2.3. Limpieza de los datos

2.3.1. Discretización y conversion de tipos de datos

Al cargar los archivos con la función read.csv(), esta de manera automática asigna el tipo de variable en el dataset, en ciertas ocasiones los tipos asignados, no son los correctos. A continuación visualizamos los tipos de variables asignados al dataset, para luego decidir si se requiere una conversión de tipo.

```
# Tipos de variables
titanic.data.ctype <- sapply(titanic.data,class)

titanic.data.ctype <- data.frame(variables = names(titanic.data.ctype),tipo = as.vector(titanic.data.cty

titanic.data.ctype
```

```
##   variables      tipo
## 1  Survived   integer
## 2    Pclass   integer
## 3       Sex character
## 4       Age   numeric
## 5     SibSp   integer
## 6     Parch   integer
## 7      Fare   numeric
## 8     Cabin character
## 9  Embarked character
```

```
rm(titanic.data.ctype)
```

En este paso realizamos un analisis sobre las variables, que en R han sido cargadas como continuas pero en realidad son discretas (factor).Para esto realizamos un análisis de discretizacion sobre los atributos, para identificar que variables tienen sentido discretizar.

```
#summary(titanic.data[,titanic.data.ctype[titanic.data.ctype$tipo == "numeric",]$variables])
# Identificar el número de clases que se encuentra en cada variable del dataset
apply(titanic.data,2, function(x) length(unique(x)))
```

```
## Survived   Pclass      Sex      Age    SibSp    Parch     Fare    Cabin
##        2        3        2       99        7        8      282      187
## Embarked
##        4
```

Con el fin de facilitar la interpretar y comparar los resultados de diferentes grupos de datos, procedemos a discretizar a las variables con pocas clases:

```
cols<-c("Survived","Pclass","Sex","SibSp","Parch","Embarked")
for (i in cols){
  titanic.data[,i] <- as.factor(titanic.data[,i]) # Conversion de variable a tipo factor
}

levels(titanic.data[,"Survived"]) <- c("No","Si")
levels(titanic.data[,"Pclass"]) <- c("Upper","Middle", "Lower")
levels(titanic.data[,"Embarked"]) <- c("?","Cherbourg", "Queenstown", "Southampton")

summary(titanic.data)
```

```
##  Survived    Pclass        Sex           Age        SibSp       Parch
##  No:815   Upper :323   female:466   Min.   : 0.17   0:891   0      :1002
```

```
##  Si:494   Middle:277   male  :843   1st Qu.:21.00   1:319   1      : 170
##           Lower :709                Median :28.00   2: 42   2      : 113
##                                     Mean   :29.88   3: 20   3      :   8
##                                     3rd Qu.:39.00   4: 22   4      :   6
##                                     Max.   :80.00   5:  6   5      :   6
##                                     NA's   :263     8:  9   (Other):   4
##       Fare            Cabin                  Embarked
##  Min.   :  0.000   Length:1309        ?          :  2
##  1st Qu.:  7.896   Class :character   Cherbourg  :270
##  Median : 14.454   Mode  :character   Queenstown :123
##  Mean   : 33.295                      Southampton:914
##  3rd Qu.: 31.275
##  Max.   :512.329
##  NA's   :1
```

2.3.2. Tratamientos de ceros o elementos vacíos

Los datos vacíos o no definidos pueden presentarse en distintos formatos, típicamente “”, ? ,“ “ o NA (Not Available en inglés), pero en algunos contextos pueden incluso tomar valores numéricos como 0 o 999.

A continuación inspeccionaremos, que atributos de nuestro dataset, tienen una cantidad alta de valores no disponibles o valores faltantes en los diferentes formatos (“”,?, o NA):

```r
# Funcion: Explorar atributos con valores faltante
# Parmetros:
# 1. dataset: conjunto de datos con los atributos a explorar
hasValoresFaltantes <- function(dataset){
  # Verificar si existen variables cuantitativas con valores NA
  variablesWithNA <- colSums(is.na(dataset))

  # Verificar si existen variables con cadenas vacias
  variablesWithEmpaty <- colSums(dataset=="")
  variablesWithEmpaty[is.na(variablesWithEmpaty)] <- 0

  # Verificar si existen variables con valores desconocidos ("?").
  variablesWithQuestionMark <- colSums(dataset=="?")
  variablesWithQuestionMark[is.na(variablesWithQuestionMark)] <- 0

  # Verificar si existen variables con valores desconocidos (" ").
  variablesWithSpace <- colSums(dataset==" ")
  variablesWithSpace[is.na(variablesWithSpace)] <- 0

  df <- data.frame(variables = names(variablesWithNA),"NA" = as.vector(variablesWithNA),stringsAsFactors

  df = bind_cols(df,"Empaty" = as.vector(variablesWithEmpaty))
  df = bind_cols(df,"?" = as.vector(variablesWithQuestionMark))
  df = bind_cols(df,"Space" = as.vector(variablesWithSpace))

  df
  #ls <- list(valoresFaltantes = df);
  #ls$totalMuestras <- dim(dataset)[1]
  #ls
}
```

```r
# Verificar si existen variables con valores faltantes
hasValoresFaltantes(titanic.data)
```

```
##   variables NA. Empaty ? Space
## 1  Survived   0      0 0     0
## 2    Pclass   0      0 0     0
## 3       Sex   0      0 0     0
## 4       Age 263      0 0     0
## 5     SibSp   0      0 0     0
## 6     Parch   0      0 0     0
## 7      Fare   1      0 0     0
## 8     Cabin   0   1014 0     0
## 9  Embarked   0      0 2     0
```

Al observar el resultado del análisis anterior, podemos identificar que para las variables Age y Fare presenta valores faltantes (NA). Para la variable Cabin se identifica que presenta una cantidad alta de valores faltantes en el formato vacío (“”). y para la variable Embarked se identifica valores faltantes en el formato “?”.

Llegados a este punto debemos decidir cómo manejar estos registros que contienen valores desconocidos:

Para el atributo **Embarked** realizamos un análisis de proporción de valores faltantes y lo actualizaremos en función del valor mas frecuente. Existen 2 casos con valor faltante con formato “?”, con una proporción del 0.15 %, el valor más frecuentes es “Southampton” con una proporción del 56.98 % .

```r
arrange(data.frame(round(prop.table(table(titanic.data$Embarked)),4)*100),-Freq)
```

```
##          Var1  Freq
## 1 Southampton 69.82
## 2   Cherbourg 20.63
## 3  Queenstown  9.40
## 4           ?  0.15
```

```r
# actualizamos los valores faltantes con el valor más frecuente
titanic.data$Embarked[titanic.data$Embarked=="?"] <- "Southampton"
```

Para el atributo **Cabin** realizamos un análisis de proporción de valores faltantes. Existen 1014 casos con valor faltante con formato vacío (“”), con una proporción del 77.46 %, esto corresponde a más de la mitad de las observaciones. Si intentamos completar los valores faltantes, por alguna de las técnicas de imputación de valores perdidos, debido a la alta cantidad de valores faltantes en este atributo, nos puede generar sesgos en los datos de este atributo. De acuerdo a esto, se decide eliminar el atributo **Cabin** del dataset en estudio.

```r
data.frame(Total=sort(colSums(titanic.data == ""), decreasing = TRUE),
           Porcentaje = sort(round(colMeans(titanic.data == "")*100, digits = 2), decreasing = TRUE))["
```

```
##       Total Porcentaje
## Cabin  1014      77.46
```

```r
# Eliminamos la variable Cabin
titanic.data <- titanic.data[, !(names(titanic.data) %in% c("Cabin"))]
```

Como podemos observar las variables **SibSp**, **Parch** y **Fare**, presenta datos con valores igual a cero, pero para las variables **SibSp**, **Parch** este valor cero significa que no tienen familiares abordo, de acuerdo a esto el valor cero tiene significado para los datos, y no serán gestionados.

Para la variable **Fare** los valores ceros podria significar un error de datos faltantes, ya que tienen un numero de ticket asignado, o tambien podriamos decir que este cero equivale a que estos ticket fueron entregados por un premio. Para esta actividad asumiremos que es un error y lo consideraremos como valores faltantes.

Calculamos la proporciona de valores ceros en la variable **Fare**, y los remplazamos por el formato de valor

faltante (NA), para luego predecir estos valores con el método kNN. Existen 17 casos con valor faltante con formato vacío (0), con una proporción del 1.3 %.

```
# Proporcion en
data.frame(Var = c("Fare"),
           Total = length(titanic.data$Fare[titanic.data$Fare == 0 & !is.na(titanic.data$Fare)]),
           Porcentaje = round((length(titanic.data$Fare[titanic.data$Fare == 0 & !is.na(titanic.data$Fa
```

```
##    Var Total Porcentaje
## 1 Fare    17        1.3
```

```
titanic.data$Fare [titanic.data$Fare == 0 & !is.na(titanic.data$Fare)]<- NA
```

Para los atributo **Fare** y **Age** realizamos un análisis de proporción de valores faltantes. Para el caso del atributo **Fare**, existe 18 caso con valor faltante con formato vacío (NA), con una proporción del 1.38 %; Y para el atributo **Age**, existe 263 casos con valores faltantes con formato vacío (NA), con una proporción del 20.09 %; Debido a que los datos presente en esta variable están un poco dispersos, utilizaremos métodos probalísticos para predecir los valores faltantes.

```
#library(VIM)
if(!require(VIM)){
    #install.packages('VIM', repos='http://cran.us.r-project.org')
    library(VIM)
}
data.frame(Total=sort(colSums(is.na(titanic.data)), decreasing = TRUE),Porcentaje = sort(round(colMeans
```

```
##      Total Porcentaje
## Fare    18       1.38
## Age    263      20.09
```

```
# Para predecir los valores faltantes utilizaremos el metodo kNN
titanic.data.imp <- kNN(titanic.data)

# Imputamos los valores faltantes
titanic.data$Age <- titanic.data.imp$Age # Age
titanic.data$Fare <- titanic.data.imp$Fare #Fare
rm(titanic.data.imp)
```

```
# Verificar si existen variables con valores faltantes
hasValoresFaltantes(titanic.data)
```

```
##   variables NA. Empaty ? Space
## 1  Survived   0      0 0     0
## 2    Pclass   0      0 0     0
## 3       Sex   0      0 0     0
## 4       Age   0      0 0     0
## 5     SibSp   0      0 0     0
## 6     Parch   0      0 0     0
## 7      Fare   0      0 0     0
## 8  Embarked   0      0 0     0
```

2.3.3. Identificación y tratamiento de valores extremos

Los valores extremos (outliers) son aquellos datos que se encuentran muy alejados de la distribución normal de una variable o población. Con este análisis queremos identificar si el dataset contiene observaciones que están alejadas de su distribución normal, con el fin de evitar que estos valores puedan afectar de forma

adversa los resultados de los análisis posteriores, al incrementar el error en la varianza de los datos y sesgar significativamente los cálculos y estimaciones.

Para identificar estos valores en el dataset, realizaremos un análisis por cuartiles, para las variables **Age** y **Fare**. Debido a que el resto de variables pueden ser de tipo categóricas o texto no las incluiremos en este análisis.

Realizaremos un análisis de valores extremos para la variable numérica **Age**, realizando un análisis por quartiles:

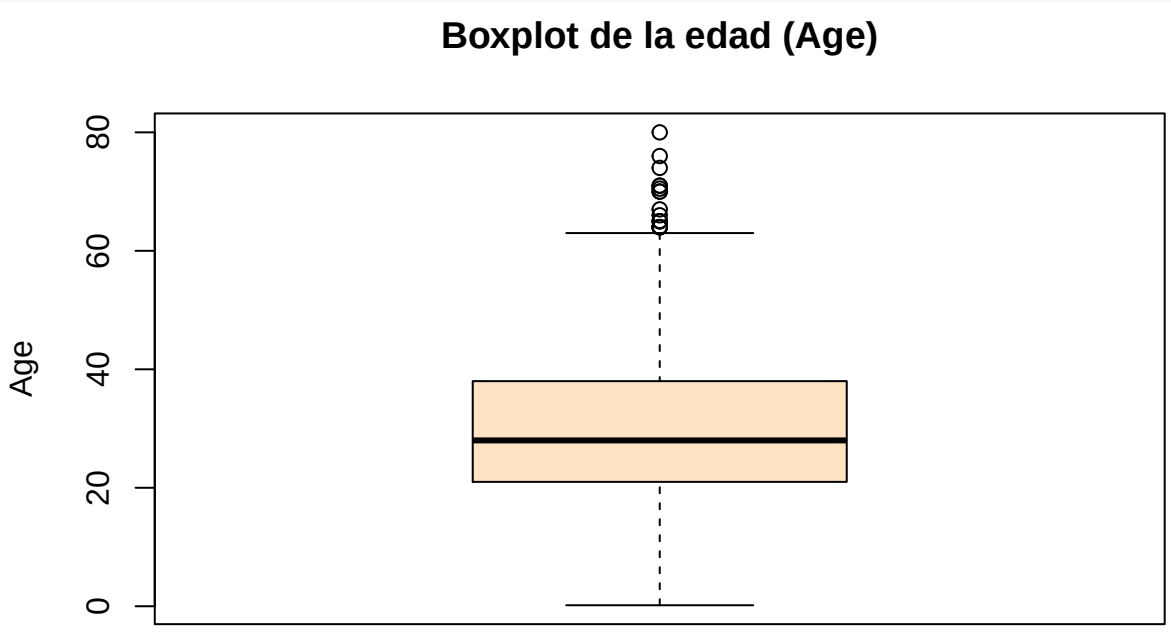
```
# generar los quartiles que representan la distribución del conjunto de datos
summary(titanic.data$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.17   21.00   28.00   29.81   38.00   80.00
```

```
## Calculamos la relación inter quartil (IQR), Q3 - Q1 = IQR()
print(paste("Relación inter quartil (IQR): ",IQR(titanic.data$Age)),quote = FALSE)
```

```
## [1] Relación inter quartil (IQR):  17
```

```
# Grafico de boxplot
gf.boxplot <- boxplot(titanic.data$Age, main="Boxplot de la edad (Age)",
ylab="Age",col = "bisque")
```



Al inspeccionar las estadísticas arrojadas, para la variable **Age**, el valor mínimo es 0.17 y el Máximo es 80. Si analisamos la diferencia entre Q1 y el Mínimo es de 20.83, y la diferencia entre Q3 y el Máximo es 42; cómo podemos ver la diferencia de Q3 y el máximo es mayor que la diferencias entre Q1 y el mínimo. Estos nos indican que el 25 % de los valores superiores es tan más dispersos, que el 75 % restante.

Al analizar el grafico de diagrama de cajas (Boxplot), se observa que no hay valores atípicos en el extremo inferior, y por eso el bigote inferior se extiende hasta el valor mínimo, 0.17. En cambio en el extremo superior vemos varios valores atípicos, representados por unos círculos sobre el bigote superior.

Para detectar los valores atípicos, los bigotes se extendieron hasta un $Mínimo = Q1 - 1{,}5 * IQR$, por debajo de Q1 y hasta un $Máximo = Q3 + 1{,}5 * IQR$, por encima de Q3. Donde **IQR = 17**, **Q1 = 21** y **Q3 = 38**; Entonces el $Mínimo = 21\tilde{}1{,}5 * 17 = -4{,}5$, donde todos los valores menores a este valor son considerados atípicos, en nuestro caso como no hay valores menores que este, por eso el bigote se extiende hasta el mínimo

valor de la variable; Los valores mayores al $Máximo = 38 + 1{,}5 * 17 = 63{,}5$ serán considerados atípicos, que son los valores representados en el grafico por los puntos negros.

Considerando lo anterior, a continuación se muestran los valores atípicos para la variable **Age**. **Donde Age > 63.5**:

```
#Valores extremos encontrados en la variable Age donde Age > 63.5
sort(gf.boxplot$out, decreasing = FALSE)
```

```
##  [1] 64.0 64.0 64.0 64.0 64.0 65.0 65.0 65.0 66.0 67.0 70.0 70.0 70.5 71.0
## [15] 71.0 74.0 76.0 80.0
```

No obstante, si revisamos los anteriores datos, las edades de los pasajeros comprendidas entre 64 y 80, son valores que perfectamente pueden darse. Es por ello que el manejo de estos valores extremos consistirá en simplemente dejarlos como actualmente están recogidos.

Realizaremos un análisis de valores extremos para la variable numérica **Fare**, realizando un análisis por quartiles:

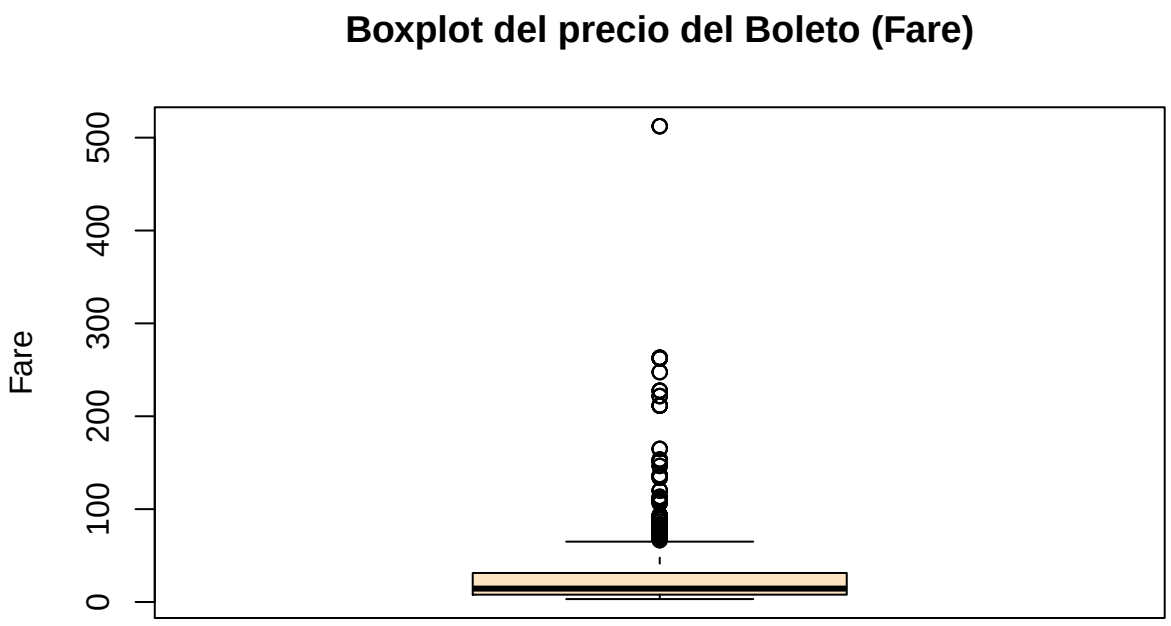
```
# generar los quartiles que representan la distribución del conjunto de datos
summary(titanic.data$Fare)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.171   7.925  14.454  33.456  31.275 512.329
```

```
## Calculamos la relación inter quartil (IQR), Q3 - Q1 = IQR()
print(paste("Relación inter quartil (IQR): ",IQR(titanic.data$Fare)),quote = FALSE)
```

```
## [1] Relación inter quartil (IQR):  23.35
```

```
# Grafico de boxplot
gf.boxplot <- boxplot(titanic.data$Fare, main="Boxplot del precio del Boleto (Fare)",
ylab="Fare",col = "bisque")
```



Al inspeccionar las estadísticas arrojadas, para la variable **Fare**, el valor mínimo es 3.17 y el Máximo es 512.33. Si analisamos la diferencia entre Q1 y el Mínimo es de 4.75, y la diferencia entre Q3 y el Máximo es 481.05; cómo podemos ver la diferencia de Q3 y el máximo es mayor que la diferencias entre Q1 y el mínimo. Estos nos indican que el 25 % de los valores superiores es tan más dispersos, que el 75 % restante.

Al analizar el grafico de diagrama de cajas (Boxplot), se observa que no hay valores atípicos en el extremo inferior, y por eso el bigote inferior se extiende hasta el valor mínimo, 3.17. En cambio en el extremo superior vemos varios valores atípicos, representados por unos círculos sobre el bigote superior.

Para detectar los valores atípicos, los bigotes se extendieron hasta un $Mínimo = Q1 - 1{,}5 * IQR$, por debajo de Q1 y hasta un $Máximo = Q3 + 1{,}5 * IQR$, por encima de Q3. Donde **IQR = 23.35**, **Q1 = 7.93** y **Q3 = 31.28**; Entonces el $Mínimo = 7{,}93\breve{}1{,}5 * 23{,}35 = -27{,}01$, donde todos los valores menores a este valor son considerados atípicos, en nuestro caso como no hay valores menores que este, por eso el bigote se extiende hasta el mínimo valor de la variable; Los valores mayores al $Máximo = 31{,}28 + 1{,}5 * 23{,}35 = 66{,}31$ serán considerados atípicos, que son los valores representados en el grafico por los puntos negros.

Considerando lo anterior, a continuación se muestran los valores atípicos para la variable **Fare**. **Donde Fare > 66.31**:

```
#Valores extremos encontrados en la variable Fare donde Fare > 66.31
sort(gf.boxplot$out, decreasing = FALSE)
```

```
##   [1]  66.6000  66.6000  69.3000  69.3000  69.5500  69.5500  69.5500
##   [8]  69.5500  69.5500  69.5500  69.5500  69.5500  69.5500  69.5500
##  [15]  69.5500  71.0000  71.0000  71.2833  71.2833  73.5000  73.5000
##  [22]  73.5000  73.5000  73.5000  73.5000  73.5000  75.2417  75.2417
##  [29]  75.2500  75.2500  76.2917  76.2917  76.7292  76.7292  76.7292
##  [36]  77.2875  77.2875  77.9583  77.9583  77.9583  78.2667  78.2667
##  [43]  78.8500  78.8500  78.8500  79.2000  79.2000  79.2000  79.2000
##  [50]  79.2000  79.2000  79.6500  79.6500  79.6500  80.0000  80.0000
##  [57]  81.8583  81.8583  81.8583  82.1708  82.1708  82.2667  82.2667
##  [64]  83.1583  83.1583  83.1583  83.1583  83.1583  83.1583  83.4750
##  [71]  83.4750  86.5000  86.5000  86.5000  89.1042  89.1042  90.0000
##  [78]  90.0000  90.0000  90.0000  90.0000  91.0792  91.0792  93.5000
##  [85]  93.5000  93.5000  93.5000 106.4250 106.4250 106.4250 108.9000
##  [92] 108.9000 108.9000 110.8833 110.8833 110.8833 110.8833 113.2750
##  [99] 113.2750 113.2750 120.0000 120.0000 120.0000 120.0000 133.6500
## [106] 133.6500 134.5000 134.5000 134.5000 134.5000 134.5000 135.6333
## [113] 135.6333 135.6333 135.6333 136.7792 136.7792 146.5208 146.5208
## [120] 146.5208 151.5500 151.5500 151.5500 151.5500 151.5500 151.5500
## [127] 153.4625 153.4625 153.4625 164.8667 164.8667 164.8667 164.8667
## [134] 211.3375 211.3375 211.3375 211.3375 211.5000 211.5000 211.5000
## [141] 211.5000 211.5000 221.7792 221.7792 221.7792 221.7792 227.5250
## [148] 227.5250 227.5250 227.5250 227.5250 247.5208 247.5208 247.5208
## [155] 262.3750 262.3750 262.3750 262.3750 262.3750 262.3750 262.3750
## [162] 263.0000 263.0000 263.0000 263.0000 263.0000 263.0000 512.3292
## [169] 512.3292 512.3292 512.3292
```

No obstante, si revisamos los anteriores datos, y miramos de forma aleatoria los precios de los Ticket podemos ver que los precios mas altos corresponde a los pasajeros de clase alta (Pclass = "Upper"), y son valores que perfectamente pueden darse. Es por ello que el manejo de estos valores extremos consistirá en simplemente dejarlos como actualmente están recogidos.

2.4. Análisis de los datos

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

2.6. Código fuente y dataset

3. Recursos