PRACTICA 2: LIMPIEZA Y VALIDACIÓN DE LOS DATOS

Jose Rodriguez

December 29, 2019

1 INTRODUCCION

Esta actividad se elabora como parte del curso Tipologia y cliclo de vida de los datos, el cual, forma parte de la Maestria de Ciencia de Datos de la UOC.

Esta practica consistente en el tratamiento de un conjunto de datos, 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.

2 EL DATASET

2.1 DESCRIPCION

El dataset seleccionado para la ejecucion de esta practica, fue obtenidos mediante el siguiente enlace de kaggle: https://www.kaggle.com/c/titanic/data. En este, se encuentran tres archivos, pero debido a que solo uno indica que personas sobrevieron al accidente del Titanic, unicamente se utilizara ese.

El archivo es llamado “train.csv” y tiene 12 columnas y 891 filas (registros de personas). A continuacion, se presentaran las variables incluidas en el dataset:

- PassengerID: ID unico para identificar a cada uno de los pasajeros
- Name: nombre del pasajero
- Sex: sexo
- Age: edad en annos
- Pclass: clase (1era - Superior, 2do = Medio, 3ro = inferior)
- Sibsp: numero de hermanos/pareja a bordo
- Parch: numero de padres/hijos a bordo
- Ticket: numero de tiquete
- Fare: monto pagado por el tiquete
- Cabin: numero de cabina
- Embarked: puerto de embarcacion (C = Cherbourg, Q = Queenstown, S = Southampton)
- Survival: indica si la persona sobrevivio o no. * Solo se encuentra en el archivo train.csv

2.2 IMPORTANCIA

El 15 de abril de 1912, durante el viaje inaugural, el ampliamente considerado “insumergible” RMS Titanic se hundió después de chocar con un iceberg. Hasta hoy en dia, este hecho es el naufragio más conocido en la historia.

En ese entonce, desafortunadamente no hubo suficientes botes salvavidas para todos los pasajeros a bordo, lo que resultó en la muerte de

1502 de 2224 pasajeros y tripulantes.

Si bien hubo algún elemento de suerte involucrado en la supervivencia, parece que algunos grupos de personas tenían más probabilidades de sobrevivir que otros.

Por ende, este trabajo ira enfocado no solo a hacer una limpieza y tratamiento de los datos, sino tambien, a encontrar los factores/variables que puedan dar respuesta a la pregunta: “¿qué tipo de personas tenían más probabilidades de sobrevivir?”.

3 INTEGRACION Y SELECCION DE DATOS

Como se menciono anteriormente, se van a unir dos archivos, por lo que primero se hara la lectura de ellos, para luego, poder unir estos.

```
#Lectura archivo 1 (train.csv)
Titanic_1 <- read.csv('C:/Users/rodriguezjos/OneDrive - VMware, Inc/VMwareCorp/Desktop/Master Data Sciences/3- II Semestre 2019/Tipologia y cliclo de vida de los datos/PRAC 2/Titanic_train.csv', sep = ',')

#Se muestran los datos para asegurarnos una correcta lectura de ellos
head(Titanic_1)
```

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

Para comprender mejor nuestros datos, se analizara su estructura y se hara un resumen de los mismos.

```
#Estructura de los datos
str(Titanic_1)
```

```
## 'data.frame':    891 obs. of  12 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : int  0 1 1 1 0 0 0 0 1 1 ...
##  $ Pclass     : int  3 1 3 1 3 3 1 3 3 2 ...
##  $ Name       : Factor w/ 891 levels "Abbing, Mr. Anthony",..: 109 191 358 277 16 559 520 629 417 581 ...
##  $ Sex        : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
##  $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
##  $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Ticket     : Factor w/ 681 levels "110152","110413",..: 524 597 670 50 473 276 86 396 345 133 ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : Factor w/ 148 levels "","A10","A14",..: 1 83 1 57 1 1 131 1 1 1 ...
##  $ Embarked   : Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...
```

Este analisis nos permite comprobar que efectivamente nuestro archivo tiene 891 registros y 12 variables. Asi mismo, podemos ver que tipo de variable es cada una.

Con base en esto, ya se puede hacer la seleccion correcta de los datos y eliminar aquellas variables que no se consideran impresndibles para el analisis.

Las siguientes son las variables que se eliminaran debido a que considero que no son factores que nos ayuden a entender que

caracteristicas tenian las personas que sobrevivieron.

- PassengerId: es un consecutivo de numeros.
- Name: el nombre no considero que sea determinante
- Cabin: es solo una numeracion dentro del crucero
- Ticket: es una numeracion con el que se identifico a un pasajero dentro del crucero

```
library(dplyr)
#Seleccion de variables
Titanic_1 <- Titanic_1 %>% select(Pclass, Sex, Age, SibSp, Parch, Fare, Survived, Embarked)
```

4 LIMPIEZA DE DATOS

4.1 VALORES NULOS

Uno de los primeros pasos cuando se trabaja con datos, es observar si las variables tienen valores perdidos o nulos. Esto significa, que por alguna razon no se pude obtener o registrar el valor.

Si bien, lo ideal es que nunca falten ningun dato, lo cierto es que en la realidad eso pocas veces pasa, por lo que se deben de aplicar tecnicas para poder estimar esos valores que no estan disponibles.

Lo primero que se hara, es verificar que no existan valores perdidos en el conjunto de datos a analizar.

```
## Verificar en que columnashay valores perdidos
colSums(is.na(Titanic_1))
```

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

Al parecer, unicamente la variable edad tiene valores perdidos, sin embargo, para estar 100% seguros, se realizara un resumen de los datos, ya que muchas veces los valores perdidos pueden tomar algun valor y no ser percibidos como un NA (Not Available)

```
#Resumen de los datos
summary(Titanic_1)
```

```
##      Pclass          Sex           Age            SibSp
##  Min.   :1.000   female:314   Min.   : 0.42   Min.   :0.000
##  1st Qu.:2.000   male  :577   1st Qu.:20.12   1st Qu.:0.000
##  Median :3.000                Median :28.00   Median :0.000
##  Mean   :2.309                Mean   :29.70   Mean   :0.523
##  3rd Qu.:3.000                3rd Qu.:38.00   3rd Qu.:1.000
##  Max.   :3.000                Max.   :80.00   Max.   :8.000
##                               NA's   :177
##      Parch             Fare           Survived      Embarked
##  Min.   :0.0000   Min.   :  0.00   Min.   :0.0000    :  2
##  1st Qu.:0.0000   1st Qu.:  7.91   1st Qu.:0.0000   C:168
##  Median :0.0000   Median : 14.45   Median :0.0000   Q: 77
##  Mean   :0.3816   Mean   : 32.20   Mean   :0.3838   S:644
##  3rd Qu.:0.0000   3rd Qu.: 31.00   3rd Qu.:1.0000
##  Max.   :6.0000   Max.   :512.33   Max.   :1.0000
##
```

El resumen nos confirma que efectivamente existen 177 casos en los que no existe el registro de edad en los pasajeros, pero, al mismo tiempo, nos indica que la embarcacion tiene dos valores en blanco.

Debido a que el margen de error al trabajar con datos aproximados es menor que con datos perdidos, se realizara un método de imputacion de valores basado en la similitud de los datos, esto bajo la hipotesis de que los registros tienen cierta relacion. Para ello, en el caso de las edades, se utilizara la tecnica de imputacion basada en k vecinos más próximos (en inglés, kNN-imputation), mientras que para el caso de la embarcacion se hara con la moda, la cual, en este caso es “S” (Southampton).

Como la idea es tratar de ser lo mas precisos posible, las imputaciones de edad se haran segun el sexo de las personas.

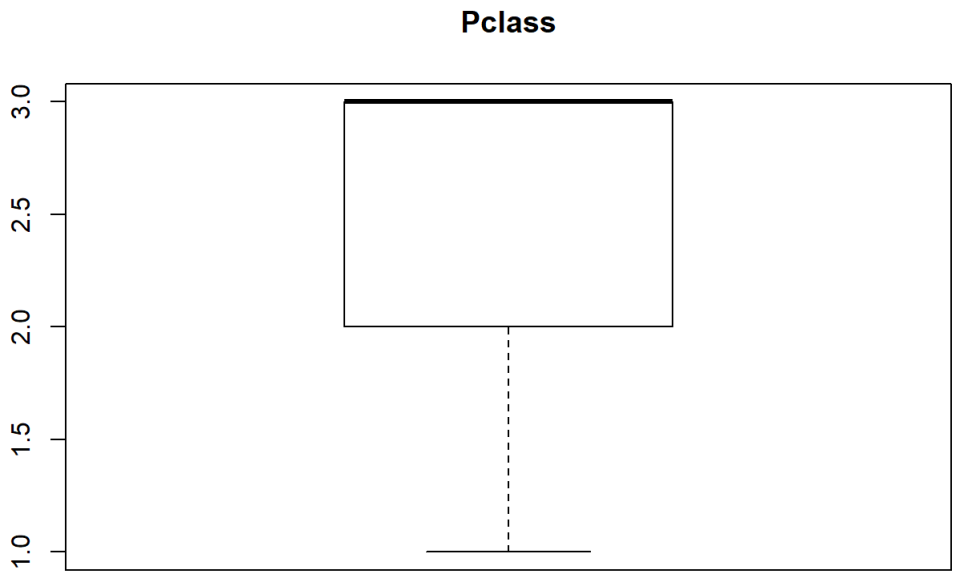
```
library(VIM)

#Seleccion de variables cuantitativas para la imputacion
cuant.names <- (Titanic_1%>% select(which(sapply(.,is.numeric))) %>% colnames())

## Imputacion de edad para Hombres
Titanic_1[Titanic_1$Sex == 'male',] <- kNN(Titanic_1[Titanic_1$Sex == 'male',], "Age", k=5, dist_var = cuant
.names, impNA=TRUE)

## Imputacion de edad para Mujeres
Titanic_1[Titanic_1$Sex == 'female',] <- kNN(Titanic_1[Titanic_1$Sex == 'female',], "Age", k=5, dist_var = c
uant.names, impNA=TRUE)

## Imputacion embarcacion
Titanic_1$Embarked[Titanic_1$Embarked==""] <- "S"
Titanic_1$Embarked <- factor(Titanic_1$Embarked)

#Resumen de los datos para verificar que los datos se hayan imputado correctamente
#summary(Titanic_1)
```

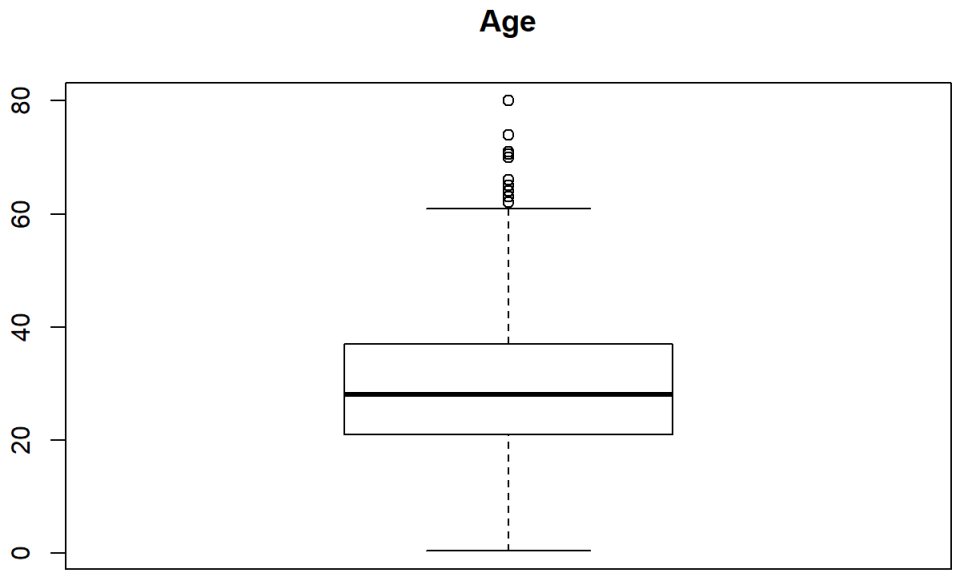
4.2 VALORES EXTREMOS

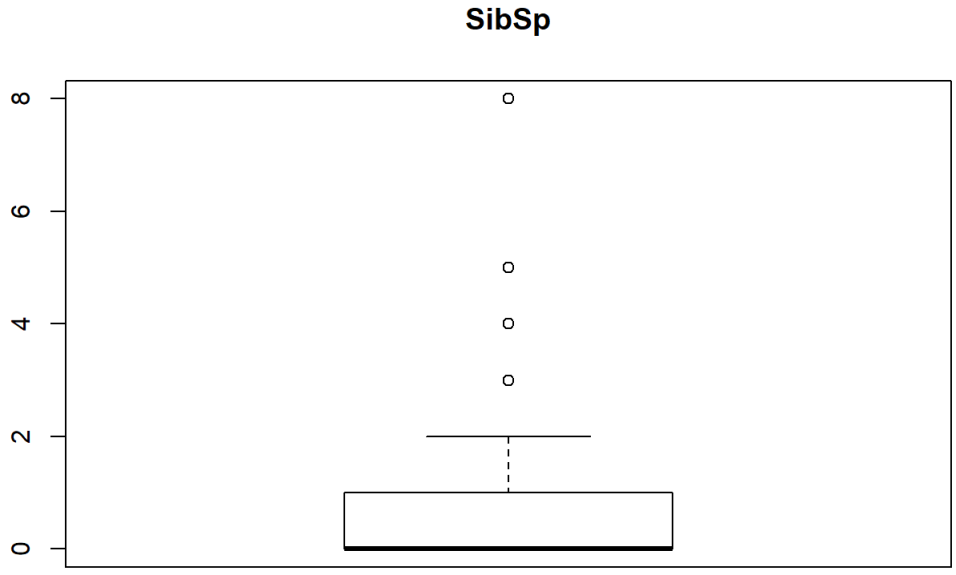
Los valores extremos son aquellos que parecen no ser congruentes con el comportamiento “normal” de los datos, es decir, valores que parecieran ser muy altos o muy bajos a compararse con el retos.

Para identificar si existen valores extremos, se realizaran boxplot para cada una de las variables numericas.

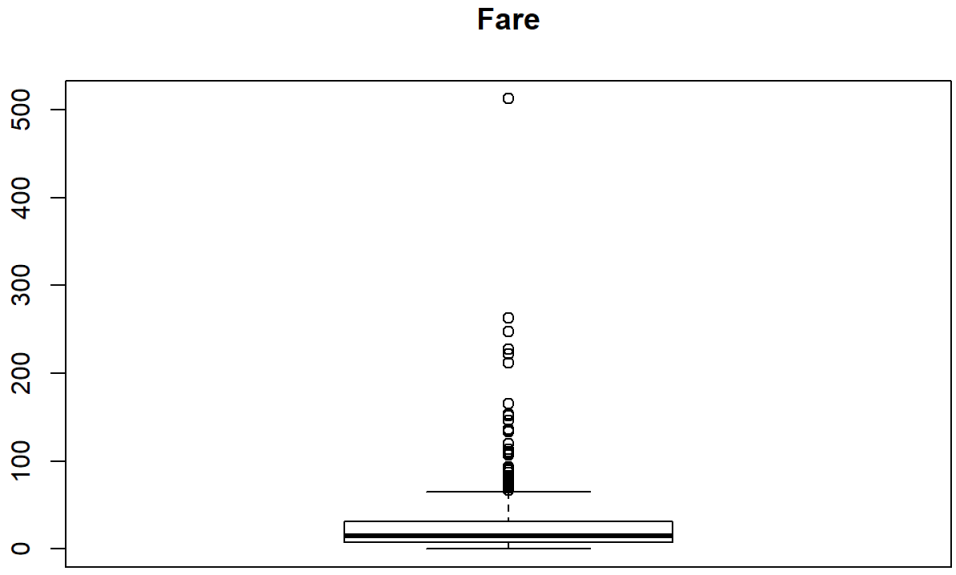
```
#Se grafican todas las variables cuantitativas
b <- Titanic_1[,cuant.names]

for (i in 1:length(b)) {
        boxplot(b[,i], main=names(b[i]), type="l", id.method="y")
}
```

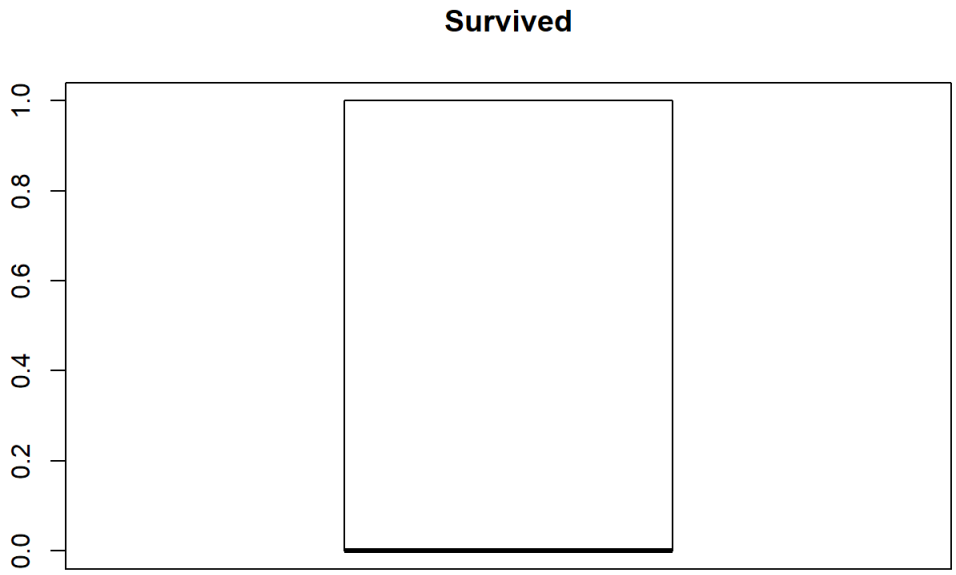



Age
0
20
40
60
80


SibSp
0
2
4
6
8


Parch
0
1
2
3
4
5
6


Fare
0
100
200
300
400
500



Graficamente, pareciera que la edad, el costo del pasaje, el numero de hermanos/pareja a bordo y el numero de padres/hijos a bordo tienen valores extremos, no obstante, todos los valores tienen sentido y estan dentro de las posibilidades que se pueden dar. Son menos normales, pero son valores totalmente validos, por lo que no se hara ningun tratamiento de ouliers.

4.3 Exportacion de los datos

Posterior a la seleccion de las variables y la imputacion de datos, se guardaran los datos limpios en otro archivo csv demoninado Titanic_clean.csv

```
#Se guardan los datos en un nuevo archivo csv
#write.csv(Titanic_1, "Titanic_clean.csv")
```

5 ANALISIS DE DATOS

5.1 SELECCION DE GRUPOS A ANALIZAR

A continuacion se crearan grupos que podrian resultar interesantes para analizar y/o comparar. El proposito sera conocer si realmente existen diferencias en estos grupos, los cuales, nos puedan ayudar a determinar que caracteristicas se debian tener para tener una mayor probabilidad de sobrevivir en el naufragio.

```
#Agrupacion por sexo
Hombres <- Titanic_1 %>% filter(Sex=="male")
Mujeres <- Titanic_1 %>% filter(Sex=="female")

#Agrupacion por clase
Primera <- Titanic_1 %>% filter(Pclass==1)
Segunda <- Titanic_1 %>% filter(Pclass==2)
Tercera <- Titanic_1 %>% filter(Pclass==3)
```

5.2 COMPROBACION DE NORMALIDAD Y HOMOGENEIDAD DE LA VARIANZA

Para comprobar si las variables cuantitativas siguen una distribucion normal, se utilizara el test de Shapiro-Wilk, donde las hipotesis a probar son las siguientes:

H0: La muestra proviene de una distribución normal.

H1: La muestra no proviene de una distribución normal.

Para ello, se utilizara un Alpha = 0.01.

```
#Test de Shapiro para Age
shapiro.test(Titanic_1$Age)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Titanic_1$Age
## W = 0.97778, p-value = 2.074e-10
```

```
#Test de Shapiro para SibSp
shapiro.test(Titanic_1$SibSp)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Titanic_1$SibSp
## W = 0.51297, p-value < 2.2e-16
```

```
#Test de Shapiro para Parch
shapiro.test(Titanic_1$Parch)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Titanic_1$Parch
## W = 0.53281, p-value < 2.2e-16
```

```
#Test de Shapiro para Fare
shapiro.test(Titanic_1$Fare)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Titanic_1$Fare
## W = 0.52189, p-value < 2.2e-16
```

Para evaluar el test anterior, Royston (1995) indico que es adecuado comparar los resultados del test de Shapiro contra una p de 0.1, por lo que si comparamos los resultados del test contra este p-value, podemos concluir que existen evidencias significativas para concluir que ninguna de las 4 variables analizadas tienen una distribucion normal.

Posteriormente, se analizara la homogeneidad de varianzas utilizando el test no parametrico de Fligner-Killeen. Este test, compara las varianzas basándose en la mediana.

Para este test, se constrastaran las varianzas de las edades de los hombres contra las de las mujeres.

H0: Varianza de edades entre hombres y mujeres es igual

H1: Varianza de edades entre hombres y mujeres es diferente

```
#Test Fligner-Killeen
fligner.test(Age ~ Sex, data = Titanic_1)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Age by Sex
## Fligner-Killeen:med chi-squared = 0.2814, df = 1, p-value = 0.5958
```

Con un nivel de significancia del 5%, se puede concluir que las varianzas de edades para ambos sexos son homogeneas. Esto, debido a que el resultado del test indica un p-value de 0.5958, el cual, es mayor al 0.05 contrastado.

5.3 PRUEBAS ESTADISTICAS

¿Que variables cuantitativas estan mas relacionadas con la probabilidad de haber sobrevivir del Titanic?

A pesar de que existen varias variables en la base de datos, no quiere decir que todas realmente influyan o esten correlacionadas con el

hecho de que una persona haya sobrevivido o no al naufragio del Titanic, por ende, este analisis intentara identificar, cuales de esas caracteristicas si estan ligadadas.

```
#Correlacion
round(cor(Titanic_1[names(Titanic_1) %in% cuant.names], method = "spearman"),2)
```

```
##          Pclass   Age SibSp Parch  Fare Survived
## Pclass     1.00 -0.37 -0.04 -0.02 -0.69    -0.34
## Age       -0.37  1.00 -0.21 -0.28  0.10    -0.07
## SibSp     -0.04 -0.21  1.00  0.45  0.45     0.09
## Parch     -0.02 -0.28  0.45  1.00  0.41     0.14
## Fare      -0.69  0.10  0.45  0.41  1.00     0.32
## Survived  -0.34 -0.07  0.09  0.14  0.32     1.00
```

El analisis de correlaciones indica que no existe ninguna variable que este fuertemente relacionada con la sobrevivencia de las personas, sin embargo, si muestra algunas relaciones que, aunque no son tan fuertes, pueden ayudar a darnos una idea de que caracteristicas pudieron tener esas personas.

Por ejemplo, las variables con mayor correlacion son la clase y el monto pagado, que de hecho, estas dos estan relacionadas entre si. La correlacion negativa lo que nos muestra es que, entre mejor sea la clase, existe mayor probabilidad de sobrevivir.

¿El sexo de las personas influye en la probabilidad de haber sobrevivir del Titanic?

Como el objetivo es identificar caracteristicas que pudieron haber hecho que una persona sobreviviera al naufragio, se evaluará la probabilidad de que un paciente pueda o no haber sobrevivido dependiendo de si era hombre o mujer. Para evaluar esta probabilidad, primero se realizará el test chi-cuadrado, para valorar la relación entre las variables y luego, se calculara el OR (odds-ratio).

```
chisq.test(Titanic_1$Survived, Titanic_1$Sex)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  Titanic_1$Survived and Titanic_1$Sex
## X-squared = 260.72, df = 1, p-value < 2.2e-16
```

Bajo la hipotesis:

H0: las dos variables son independientes

H1: las dos variables estan relacionadas

Analizando la prueba de chi cuadrado y tomando un nivel de significancia del 5%, se puede concluir que el sexo si esta relacioada con la variable de sobrevivencia. Esto, debido a que el p-value es menor a 0.05.

Como se sabe que si estan relacionadas, se calculara el OR para conocer el efecto que el sexo tiene sobre la probabilidad de haber sobrevivido

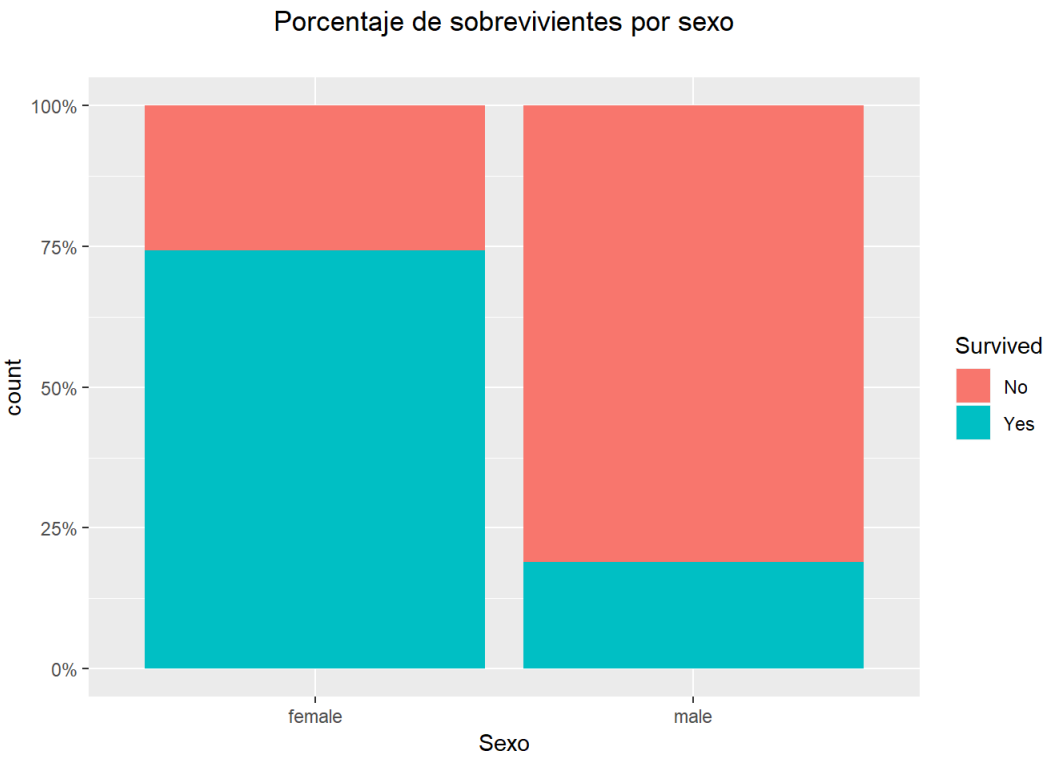
```
library(DescTools)
#Se crea un modelo de regresion logistica
rl_sex <- glm(Survived ~ Sex, data = Titanic_1, family = "binomial")

#Se visualizan los odds ratio
OddsRatio(rl_sex)
```

```
##
## Call:
## glm(formula = Survived ~ Sex, family = "binomial", data = Titanic_1)
##
## Odds Ratios:
##                or or.lci or.uci  Pr(>|z|)
## (Intercept) 2.877  2.245  3.725  2.58e-16 ***
## Sexmale     0.081  0.058  0.112 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Brier Score: 0.167     Nagelkerke R2: 0.354
```

Al realizar un modelo de regresion logistica, este nos confirma que el sexo si esta relacionado con la variable sobrevivencia ya que su p-value es menor al 5%. Asi mismo, al observar el OR, este indica que las personas de sexo masculino, tenian menor probabilidad de haber sobrevivido; y este se debe, a que su valor es menor a 1.

De hecho, si observamos el porcentaje de personas que sobrevivieron segun el sexo, se puede observar que definitivamente las mayores sobrevivientes eran del sexo femenino. Se podria decir que 3 de cada 4 mujeres sobrevivian, mientras que en los hombres, fue 1 de cada 5 aproximadamente.

```
library(ggplot2)
brks <- c(0, 0.25, 0.5, 0.75, 1)
ggplot(Titanic_1, aes(Sex)) + geom_bar(aes(fill=factor(Survived)), position = "fill") + ggtitle("Porcentaje de sobrevivientes por sexo\n") + theme(plot.title = element_text(hjust = 0.5)) + scale_fill_discrete(name = "Survived", labels = c("No", "Yes")) + xlab("Sexo") + scale_y_continuous(breaks = brks, labels = scales::percent(brks))
```



Modelo de Bosques Aleatorios

Los Bosques Aleatorios es un algoritmo de Machine Learning flexible y facil de usar que produce buenos resultados la mayor parte del tiempo. Por eso mismo, es uno de los algoritmos más utilizados, debido a su simplicidad y al hecho de que se puede usar tanto para tareas de clasificación como de regresión.

En este caso particular, el problema que se deriva es de clasificacion, debido a que lo que se desea predecir es si una persona sobreviviria o no al naufragio del Titanic, segun algunas caracteristicas.

El modelo de bosques aleatorios no solo nos permite hacer la prediccion, sino tambien, nos indica cuales fueron las variables mas importantes que tomo en cuenta para realizar la prediccion.

Para crear una buena prediccion, se crearan varios modelos de random forest y luego se compararan entre ellos para elegir el que logro hacer la mayor cantidad de predicciones.

```
library(caTools)
library(randomForest)
library(caret)
set.seed(123)
#Se hace una particion de los datos con el 80%
sample = sample.split(Titanic_1$Survived, SplitRatio = 0.8)

#Se crean dos data set, una para entrenar el modelo y el otro para testearlo
train_titanic = subset(Titanic_1, sample==TRUE)
test_titanic = subset(Titanic_1, sample==FALSE)

#Tabla para conocer cuantas personas sobrevivieron o no al naufragio del Titanix
table(train_titanic$Survived)
```

```
##
##   0   1
## 439 274
```

```
#MOdelo 1
modelo1 <- randomForest(x=train_titanic[,-7],y=as.factor(train_titanic[,7]))
pred1 <- predict(modelo1, newdata = test_titanic[,-7])
confusionMatrix(as.factor(test_titanic[,7]), pred1)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 100  10
##          1  20  48
##
##                Accuracy : 0.8315
##                  95% CI : (0.7682, 0.8833)
##     No Information Rate : 0.6742
##     P-Value [Acc > NIR] : 1.724e-06
##
##                   Kappa : 0.6327
##  Mcnemar's Test P-Value : 0.1003
##
##             Sensitivity : 0.8333
##             Specificity : 0.8276
##          Pos Pred Value : 0.9091
##          Neg Pred Value : 0.7059
##              Prevalence : 0.6742
##          Detection Rate : 0.5618
##    Detection Prevalence : 0.6180
##       Balanced Accuracy : 0.8305
##
##        'Positive' Class : 0
##
```

Este primer modelo tiene precision del 83%, por lo que se buscara crear otros modelos que logren ser mejores que este inicial sin hacerle algun ajuste.

```
modelo2 <- randomForest(x=train_titanic[,-7],y=as.factor(train_titanic[,7]), ntree = 1000, mtry= 2)
pred2 <- predict(modelo2, newdata = test_titanic[,-7])
confusionMatrix(as.factor(test_titanic[,7]), pred2)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 99 11
##          1 21 47
##
##                Accuracy : 0.8202
##                  95% CI : (0.7558, 0.8737)
##     No Information Rate : 0.6742
##     P-Value [Acc > NIR] : 9.336e-06
##
##                   Kappa : 0.6083
##  Mcnemar's Test P-Value : 0.1116
##
##             Sensitivity : 0.8250
##             Specificity : 0.8103
##          Pos Pred Value : 0.9000
##          Neg Pred Value : 0.6912
##              Prevalence : 0.6742
##          Detection Rate : 0.5562
##    Detection Prevalence : 0.6180
##       Balanced Accuracy : 0.8177
##
##        'Positive' Class : 0
##
```

El segundo modelo, mas bien fue peor que el primero, por lo que se buscaran mas opciones.

```
library(doParallel)
```

```
## Loading required package: foreach
```

```
## Loading required package: iterators
```

```
## Loading required package: parallel
```

```
registerDoParallel(cores=6)

modelo3 <- foreach(ntree=rep (550,25), .combine=randomForest::combine, .packages= 'randomForest') %dopar%
randomForest (x=train_titanic[,-7],y=as.factor(train_titanic[,7]),
              sampsize = c(150,150), ntree = ntree, mtry= 2, nodesize = 1, do.trace = 100)
pred3 <- predict(modelo3, newdata = test_titanic[,-7])
confusionMatrix(as.factor(test_titanic[,7]), pred3)
```

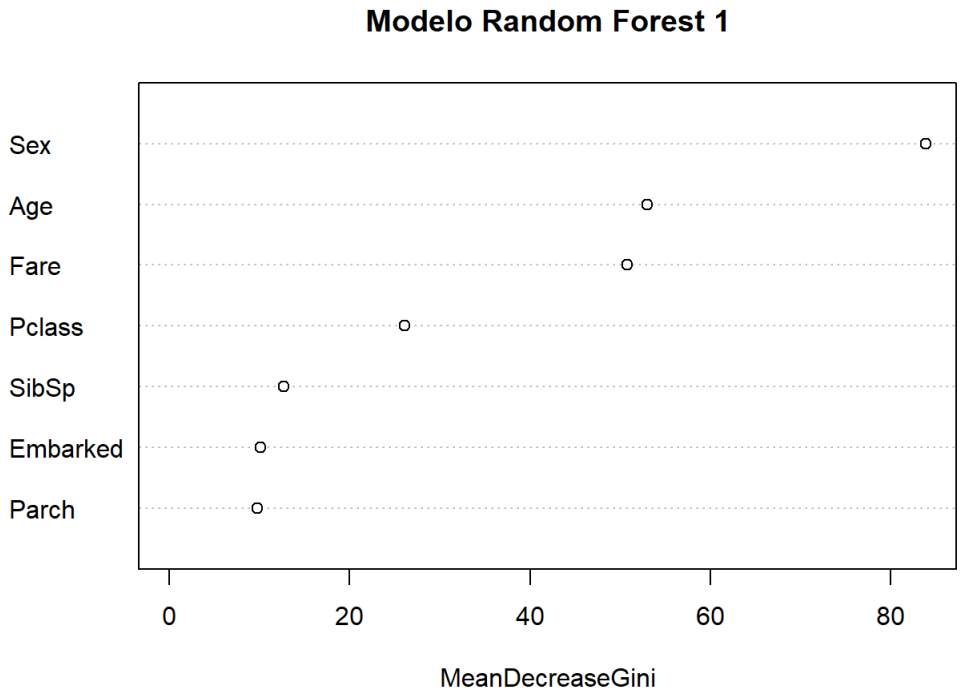
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 91 19
##          1 12 56
##
##                Accuracy : 0.8258
##                  95% CI : (0.762, 0.8785)
##     No Information Rate : 0.5787
##     P-Value [Acc > NIR] : 1.684e-12
##
##                   Kappa : 0.6383
##  Mcnemar's Test P-Value : 0.2812
##
##             Sensitivity : 0.8835
##             Specificity : 0.7467
##          Pos Pred Value : 0.8273
##          Neg Pred Value : 0.8235
##              Prevalence : 0.5787
##          Detection Rate : 0.5112
##    Detection Prevalence : 0.6180
##       Balanced Accuracy : 0.8151
##
##        'Positive' Class : 0
##
```

Al parecer, entre mas complejo intento hacer el modelo, la precision baja mas. Probe varios ajustes mas, sin embargo, no los dejo debido a que los resultados fueron peores a los mostrados.

De acuerdo con lo anterior, el modelo elegido sera el primero, que a pesar de que la precision no es tan alta como quisiera, al menos es la mejor entre los otros que se probaron.

El algoritmo de Bosques Aleatorios no solo nos ayuda a predecir a nuevos sujetos, sino tambien, nos indica cuales son las variables mas importantes que tomo en cuenta al momento de crear la prediccion.

```
varImpPlot(modelo1,main="Modelo Random Forest 1")
```



En el grafico se logra apreciar que el sexo de la persona definitivamente es el aspecto mas importante que utiliza el modelo para predecir, seguido de la edad y el monto pagado.

Finalmente, como ejercicio, usare los datos mios y los de mi novia, como si los dos hubiesemos viajado juntos, para conocer si alguno de los dos hubiese sobrevivido

```
#Creacion de data frame con datos reales
nuevosdatos <- data.frame("Pclass" = c(2,2), "Sex" = c("male", "female"), "Age" = c(32,35), "SibSp" = c(0,0
), "Parch" = c(0,0), "Fare" = c(20,20), "Embarked" = c("C","C"))

#Reclasificacion de la clase de alguna variables
nuevosdatos$Pclass = as.integer(nuevosdatos$Pclass)
nuevosdatos$SibSp = as.integer(nuevosdatos$SibSp)
nuevosdatos$Parch = as.integer(nuevosdatos$Parch)

#Asignacion de los mismo niveles
levels(nuevosdatos$Embarked) = levels(Titanic_1$Embarked)

#Prediccion
predict(modelo1, newdata = nuevosdatos)
```

```
## 1 2
## 0 1
## Levels: 0 1
```

A pesar de que ambos tenemos caracteristicas muy similares, al paracer ella si sobreviviria por el hecho de ser mujer.

6 CONCLUSIONES

A pequena escala, esta practica es basicamente un estracto de lo que normalmente puede llegar a ser un proyecto de Ciencia de Datos.

Inicialmente, se realizaron ciertas pruebas para detectar valores peridos y extremos, con el proposito de tener nuestros datos completos y de forma correcta, tratando de evitar caer en conclusiones errones por culpa de ello.

Con los valores perdidos, se hizo una imputacion de acuerdo al sexo y demas caracteristicas de las personas para tratar depredecir los valores faltantes de una manera mas precisa. En el caso de la embarcacion, se utiizo la moda, es decir, se les asigno el puerto mas comun dentro de la muestra.

Posteriormente, se analizo la distribucion de las variables y se encontro que ninguna de ellas sigue una distribucion normal. Esto es importante, ya que segun el analisis que se quiera hacer,se debia considerar que tenian que hacerse pruebas no parametricas.

Seguidamente,se crearon grupos de interes, como por ejemplo, separa a mujeres y hombres, asi como a las personas que compraron boletos en las diferentes clases del Titanis (1ra, 2da o era clase). Cabe mencionar, que este ultimo grupo no se analizo en el proyecto, pero se considera importante indagar mas en el tema a futuro para ver la relacion real que tiene esta sobre las personas que sobrevivieron.

Con respecto al sexo de las personas, se encontro que esta variable esta algo relacionada con la variable de si sobrevive o no. Tambien se logro identificar que las mujeres tenian mayor probabilidad de haber sobrevivido y finalmente, el modelo de prediccion indico que esta fue la variable mas importante al momento de predecir.

De hecho, utilizando el modelo de prediccion, al introducir datos de un hombre y una mujer, el modelo indico que la mujer si hubiese sobrevivido, mientras que el hombre no.

La edad, a pesar de que parece no estar tan correlacionada con si se sobrevive o no, el modelo predictivo la toma como la segunda de mayor importancia. Asi misma, se identifico que la varianza de edades entre hombres y mujeres es estadisticamente homogenea.

Finalmente, con base en los resultados obtenidos, se podria decir que el sexo, la edad, el monto pagado y la clase, tienen cierta relevancia para identificar si una persona iba a sobrevivir o no.

Como proyecto personal a futuro, me queda seguir trabajando el dataset con el fin de mejorar la prediccion. Considero que seria bueno identificar el monto pagado por persona y no por familia o grupo de personas. Conocer si las varianzas de las edades en las diferentes clases son homogeneas o no. Tal vez, si no son homogeneas, la imputacion podria hacerse de otra forma.

Tampoco se logro identificar en este proceso, si las personas que eran madres o padres de familia, se salvaron junto con sus hijos o si eso no tuvo ningun peso.

Al final, el proyecto es muy interesante, pero uno queda con muchas interrogantes que por tiempo, no se pudieron desarrollar, sin embargo, las trabajare para subir las prediccion en Kaggle.