Tipología y ciclo de vida de los datos: Práctica 2 Limpieza y validación de los datos

Autor: Marcos Azuaga Canteras

Diciembre 2018

# Descripción del dataset

Se va a utilizar el dataset de Titanic:Machine Learning form Disaster sacado de <https://www.kaggle.com/c/titanic> , en este data set tenemos una muestra de 891 de 2224 entre pasajeros y tripulación, la tragedia sucedió cuando el barco choco contra un iceberg y al hundirse no habían suficientes botes salvavidas , con este análisis queremos analizar si las características personales podían influir en sobrevivir o morir.

El dataset se divide en los siguientes registros:

* survival: se identifica si el pasajero sobrevive o no: 0 = No, 1 = Si
* pclass: Identifica la clase del pasajero: 1 = 1a, 2 = 2a, 3 = 3a
* sex: Identifica el sexo del pasajero male= Hombre , female=Mujer.
* Age: Edad del pasajero en años.
* sibsp: Número de Hermanos o Esposo a bordo.
* parch: Número de padres o hijos a bordo.
* ticket: Identificador del billete.
* fare: Importe pagado por el pasajero
* cabin: Número de la cabina
* embarked puerto de embarque C = Cherbourg, Q = Queenstown, S = Southampton

# Integración y selección de los datos de interes

Para este caso se realizará la carga completa del fichero train ya que contiene la variable survival que será clave para los análisis que queremos realizar. El data set lo sacamos de la siguiente dirección : <https://www.kaggle.com/c/titanic/download/train.csv>

#ruta:https://www.kaggle.com/c/titanic/download/train.csv  
library(ggplot2)  
library(dplyr)  
# Lectura de fichero  
  
titanic<- read.csv("D:/UOC/Tipologia y ciclo de vida de los datos/Practica2/train.csv", header = TRUE, stringsAsFactors= FALSE)  
# Visualizo la estructura de datos  
str(titanic)

## '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 : chr "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)" "Heikkinen, Miss. Laina" "Futrelle, Mrs. Jacques Heath (Lily May Peel)" ...  
## $ Sex : chr "male" "female" "female" "female" ...  
## $ 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 : chr "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...  
## $ Fare : num 7.25 71.28 7.92 53.1 8.05 ...  
## $ Cabin : chr "" "C85" "" "C123" ...  
## $ Embarked : chr "S" "C" "S" "S" ...

# Limpieza de los datos

## Los datos contienen ceros

#Reviso los valores nulos  
colSums(is.na(titanic))

## PassengerId Survived Pclass Name Sex Age   
## 0 0 0 0 0 177   
## SibSp Parch Ticket Fare Cabin Embarked   
## 0 0 0 0 0 0

#Asigno valores vacíos con la media de edad según sexo  
titanic$Age[is.na(titanic$Age) & titanic$Sex=="female"] <- mean(titanic$Age,na.rm=TRUE)  
titanic$Age[is.na(titanic$Age) & titanic$Sex=="male"] <- mean(titanic$Age,na.rm=TRUE)

Primero comprobamos que hay 177 valore nulos en el campo Age, para intentar minimizar el impacto se le asignará la media de edad según sexo , así pues a un hombre que no tengamos la edad se le asignara la media de edad de los hombres, y lo mismo con las mujeres.

colSums(titanic==0)

## PassengerId Survived Pclass Name Sex Age   
## 0 549 0 0 0 0   
## SibSp Parch Ticket Fare Cabin Embarked   
## 608 678 0 15 0 0

#Calculo la media del gato por categoría  
m1<-mean(titanic$Fare[titanic$Pclass==1],na.rm=T)  
m2<-mean(titanic$Fare[titanic$Pclass==2],na.rm=T)  
m3<-mean(titanic$Fare[titanic$Pclass==3],na.rm=T)  
#modifico 0 con la media de la clase  
titanic$Fare[titanic$Pclass==1 & titanic$Fare==0] <-m1  
titanic$Fare[titanic$Pclass==2 & titanic$Fare==0] <-m2  
titanic$Fare[titanic$Pclass==3 & titanic$Fare==0] <-m3

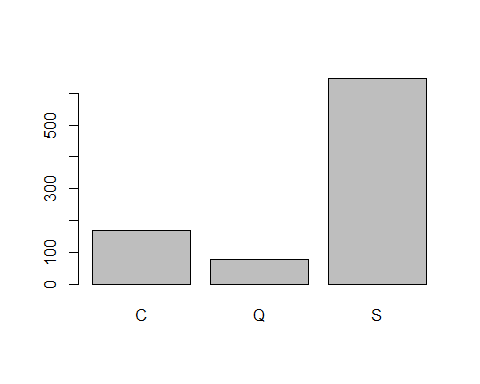
Ahora miramos los ceros, vemos que hay en las variables:

* Survived: Estos valores son correctos ya que son los pasajeros que mueren.
* SibSp: Estos valores ceros son correctos ya que no tienen familiares.
* Parch: Estos valores ceros son correctos ya que no tienen familiares.
* Fare: Se encuentras 15 registros con valor 0 , estos valores pueden variar los datos por eso los corregiremos , usando el valor medio de lo que han pagado cada clase.

colSums(titanic=="")

## PassengerId Survived Pclass Name Sex Age   
## 0 0 0 0 0 0   
## SibSp Parch Ticket Fare Cabin Embarked   
## 0 0 0 0 687 2

#Asigno S a los valores vacios ya que es el más frecuente y no influirán en el resultado  
titanic$Embarked[titanic$Embarked==""]="S"  
titanic$Embarked<-as.factor(titanic$Embarked)  
plot(titanic$Embarked)



#Corregir valor vacio de Cabin  
titanic$Cabin[substr(titanic$Cabin,1,1)==""]<-"Unknow"

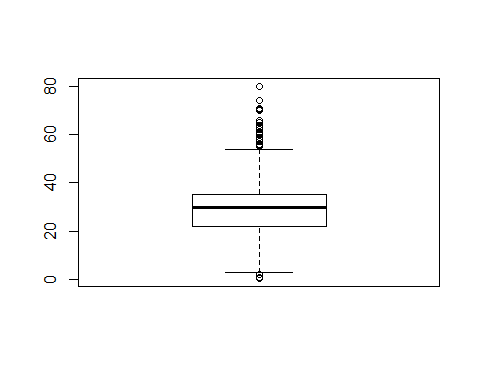
Ahora miramos los valores vacíos, vemos que hay en las variables:

-Cabin: tenemos 687 valores desconocidos, los etiquetáremos como desconocidos. -Embarked : Tenemos dos valores vacíos , les asignaremos el valor más frecuente para disminuir la influencia en el resultado.

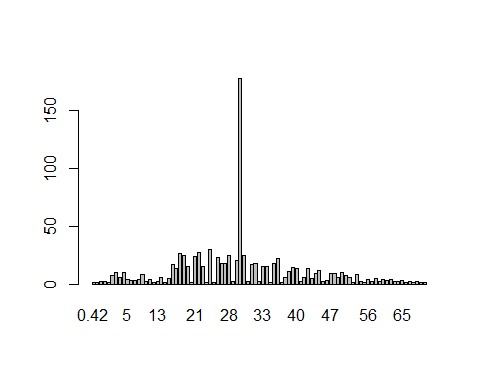
## Identificación y tratamiento de valores extremos

Buscamos valores extremos de Age:

boxplot(titanic$Age)



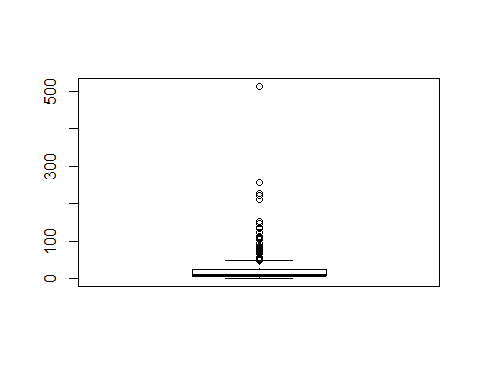
barplot(table(titanic$Age))



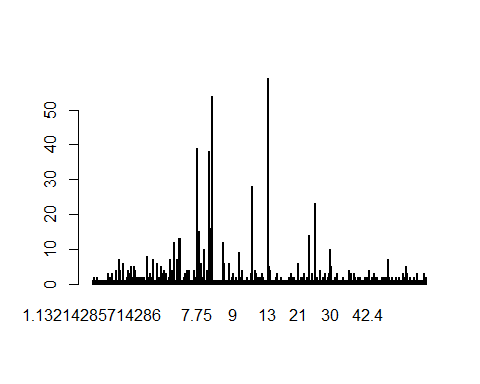
Aunque la variable Age se observan valores extremos, son legítimos.

Ahora, buscamos valores extremos de la variable Fare , asumimos que el valor indicado es por unidad familiar así que dividiremos el importe x el número de componentes de la familia.

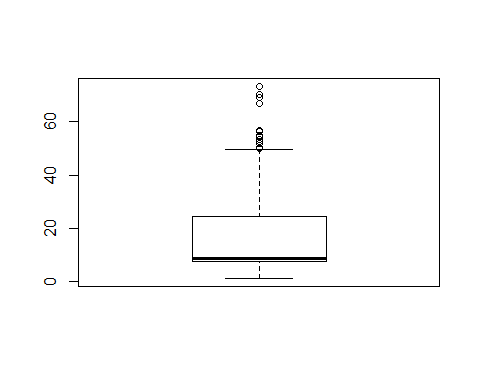
#Sumo la variable sibsp y parch para crear una nueva variable llamada familiares a bordo  
titanic$Familiares<-titanic$SibSp+titanic$Parch  
titanic$SibSp<-NULL  
titanic$Parch<-NULL  
  
#Modifico el valor Fare para refleje lo pagado individualmente y no familiar:  
titanic$Fare<-titanic$Fare/(titanic$Familiares+1)  
  
boxplot(titanic$Fare)



barplot(table(titanic$Fare))

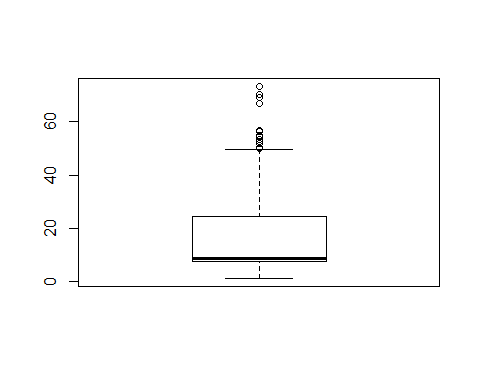


#Modifico los valores extremos al valor más alto  
max<-sort(unique(titanic$Fare), TRUE)[2]  
max<-quantile(titanic$Fare,prob=0.95)  
titanic$Fare[titanic$Fare>max] <-max  
boxplot(titanic$Fare)



En la variable Fare se observan valores extremos que pueden modificar el resultado de posibles cálculos estadísticos por eso se sustituirán los valores extremos por el valor del percentil 95.

boxplot(titanic$Fare)



## Factorizo variables

#Renombro la variable sex a los valores F y M  
titanic$Sex[titanic$Sex=="female"] <-"F"  
titanic$Sex[titanic$Sex=="male"] <-"M"  
titanic$Sex<-as.factor(titanic$Sex)  
  
#Renombro la variable survived a los valores V y M  
Survived<-titanic$Survived  
titanic$Survived[titanic$Survived==0] <-"Muere"  
titanic$Survived[titanic$Survived==1] <-"Vive"  
titanic$Survived<-as.factor(titanic$Survived)  
  
#Renombro la variable pclass a los valores 1a , 2a y 3a  
titanic$Pclass[titanic$Pclass==1] <-"1a"  
titanic$Pclass[titanic$Pclass==2] <-"2a"  
titanic$Pclass[titanic$Pclass==3] <-"3a"  
titanic$Pclass<-as.factor(titanic$Pclass)  
  
#Modifico la variable Cabin para quedarnos con la información de la cubierta.  
titanic$Cabin[substr(titanic$Cabin,1,1)=="A"]<-"A"   
titanic$Cabin[substr(titanic$Cabin,1,1)=="B"]<-"B"   
titanic$Cabin[substr(titanic$Cabin,1,1)=="C"]<-"C"   
titanic$Cabin[substr(titanic$Cabin,1,1)=="D"]<-"D"   
titanic$Cabin[substr(titanic$Cabin,1,1)=="E"]<-"E"  
titanic$Cabin[substr(titanic$Cabin,1,1)=="F"]<-"F"   
titanic$Cabin[substr(titanic$Cabin,1,1)=="G"]<-"G"   
titanic$Cabin[substr(titanic$Cabin,1,1)=="T"]<-"T"   
titanic$Cabin[substr(titanic$Cabin,1,1)==""]<-"Unknow"   
titanic$Cabin<-as.factor(titanic$Cabin)

## Eliminio variables que no usaremos

#Elimino las variables Ticket, Name y PassengerId ya que no aportan valor estadístico  
titanic$Name<-NULL  
titanic$Ticket<-NULL  
titanic$PassengerId <-NULL

# Análisis de datos

## Selección de los grupos de datos

Haremos una prueba estadística para demostrar si influye el importe pagado por el billete con la posibilidad de sobrevivir o no, con un nivel de confianza del 95%. Dividiremos nuestra población de muestra en dos muestras una los pasajeros que Murieron y otra los pasajeros que vivieron

TitanicMuere<-titanic[ which(titanic$Survived=="Muere"),]  
TitanicVive<-titanic[ which(titanic$Survived=="Vive"),]

# Comprobación de la normalidad

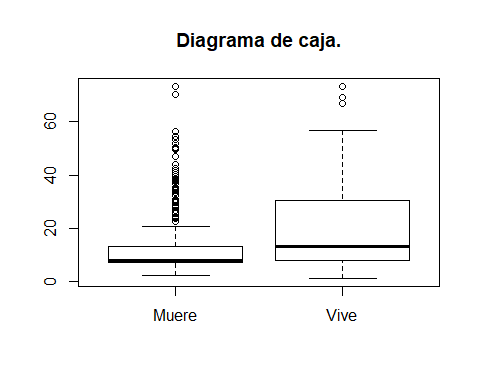
shapiro.test(TitanicMuere$Fare)

##   
## Shapiro-Wilk normality test  
##   
## data: TitanicMuere$Fare  
## W = 0.61169, p-value < 2.2e-16

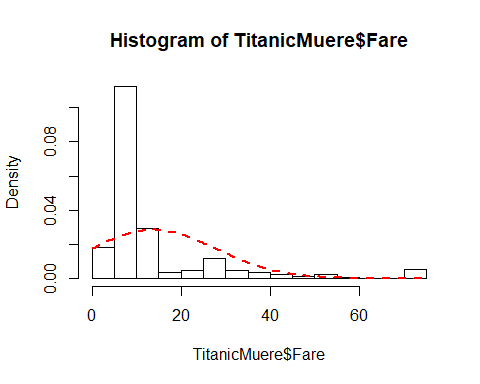
shapiro.test(TitanicVive$Fare)

##   
## Shapiro-Wilk normality test  
##   
## data: TitanicVive$Fare  
## W = 0.78642, p-value < 2.2e-16

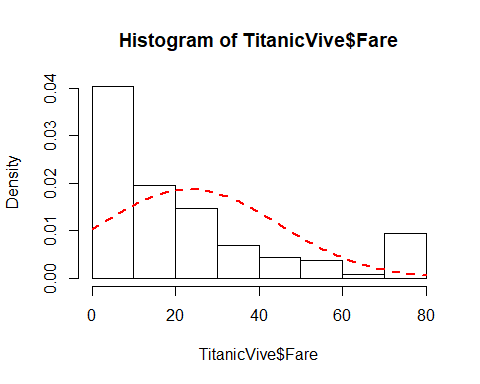
boxplot(TitanicMuere$Fare,TitanicVive$Fare,main="Diagrama de caja.",names=c("Muere","Vive"))



hist(TitanicMuere$Fare, freq = FALSE)  
curve(dnorm(x, mean(TitanicMuere$Fare), sd(TitanicMuere$Fare)), col = 2, lty = 2, lwd = 2, add=T)



hist(TitanicVive$Fare, freq = FALSE)  
curve(dnorm(x, mean(TitanicVive$Fare), sd(TitanicVive$Fare)), col = 2, lty = 2, lwd = 2, add=T)



TitanicMuere$Fare<-log1p(TitanicMuere$Fare)  
TitanicVive$Fare<-log1p(TitanicVive$Fare)

Después de realizar el test Shapiro-Wilk vemos que los dos p-valores son inferiores a 0.05 así pues rechazamos la hipótesis nula, por lo tanto aceptamos la hipótesis alternativa y podemos afirmar que las muestras no se distribuyen siguiendo una normal , así que vamos a normalizar los datos.

## Aplicación de pruebas estadísticas

Ahora vamos a realizar una prueba estadística para demostrar si influye el importe pagado por el billete con la posibilidad de sobrevivir o no, con un nivel de confianza del 95%, para ello usaremos la siguiente prueba de contraste de hipótesis:

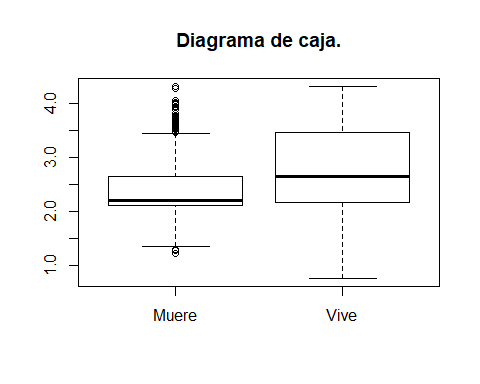
t.test(TitanicVive$Fare,TitanicMuere$Fare,alternative="great",conf.level=0.95, var.equal=FALSE)

##   
## Welch Two Sample t-test  
##   
## data: TitanicVive$Fare and TitanicMuere$Fare  
## t = 7.8205, df = 589.16, p-value = 1.22e-14  
## alternative hypothesis: true difference in means is greater than 0  
## 95 percent confidence interval:  
## 0.320684 Inf  
## sample estimates:  
## mean of x mean of y   
## 2.850189 2.443923

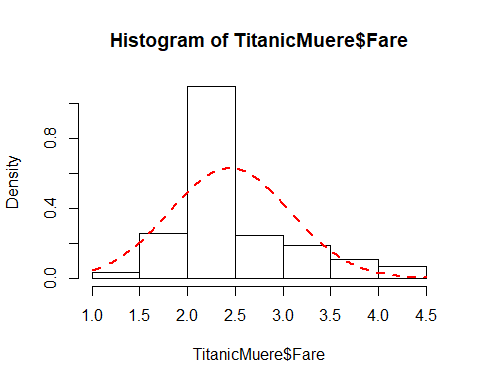
Depués de ejecutar la prueba de hipótesis nos da un p-valor es 0.00.

# Representación de los resultados

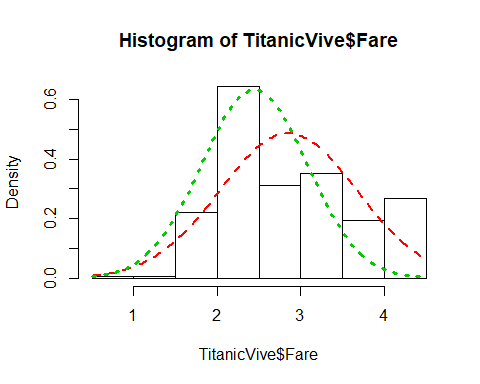
boxplot(TitanicMuere$Fare,TitanicVive$Fare,main="Diagrama de caja.",names=c("Muere","Vive"))



hist(TitanicMuere$Fare, freq = FALSE)  
curve(dnorm(x, mean(TitanicMuere$Fare), sd(TitanicMuere$Fare)), col = 2, lty = 2, lwd = 2, add=T)



hist(TitanicVive$Fare, freq = FALSE)  
curve(dnorm(x, mean(TitanicVive$Fare), sd(TitanicVive$Fare)), col = 2, lty = 2, lwd = 2, add=T)  
curve(dnorm(x, mean(TitanicMuere$Fare), sd(TitanicMuere$Fare)), col = 3, lty = 3, lwd = 3, add=T)



# Resolución del problema

Después de analizar los resultado del contraste de hipóteis, como el p-valor es 0.00 inferior a 0.05 entonces rechazamos la hipótesis nula y aceptamos la hipótesis alternativa con lo que los supervivientes pagarón mas de media que los pasajeros que murieron.

# Regresión lineal

Ahora vamos a generar un modelo predictivo de si un pasajero vive o muere según las siguientes variables: Sexo, Clase , Edad , Precio Billete, Ciudad Origen y Familiares en el barco.

## Modelo

RegModelo <- lm(Survived ~ titanic$Sex+ titanic$Pclass+ titanic$Age + titanic$Fare + titanic$Embarked+ titanic$Familiares )  
RegModelo

##   
## Call:  
## lm(formula = Survived ~ titanic$Sex + titanic$Pclass + titanic$Age +   
## titanic$Fare + titanic$Embarked + titanic$Familiares)  
##   
## Coefficients:  
## (Intercept) titanic$SexM titanic$Pclass2a   
## 1.251677 -0.511341 -0.196230   
## titanic$Pclass3a titanic$Age titanic$Fare   
## -0.396035 -0.005901 -0.001078   
## titanic$EmbarkedQ titanic$EmbarkedS titanic$Familiares   
## -0.012316 -0.075903 -0.030615

## Prueba del modelo

test.titanic<-data.frame(titanic$Pclass,titanic$Sex,titanic$Age,titanic$Fare,titanic$Embarked,titanic$Familiares)  
colnames(test.titanic)<-c("Pclass","Sex","Age","Fare","Embarked","Familiares")  
test.titanic$PredictPre<-predict(RegModelo, newdata =test.titanic, interval = "prediction")  
test.titanic$Predict<-(1/(1+exp(-(test.titanic$PredictPre))))  
  
library(SDMTools)  
c<-confusion.matrix(Survived,test.titanic$Predict[,1],0.65)  
c

## obs  
## pred 0 1  
## 0 514 135  
## 1 35 207  
## attr(,"class")  
## [1] "confusion.matrix"

todo<-c[1,1]+c[1,2]+c[2,1]+c[2,2]  
correcto<-c[1,1]+c[2,2]  
porC<-((correcto)\*100)/todo  
porC

## [1] 80.92031

porF<-100-porC  
porF

## [1] 19.07969

Después de probar el modelo vemos que tienen una tasa de acierto del 80.92% así que es un buen modelo de predicción.