Tipología y ciclo de vida de los datos: Práctica2 - Limpieza y validación de los datos

Autor: Paula León Gil-Gibernau (aula 1) y Silvina Guijarro Domingo (aula 2)

Junio 2019

Contents

Presentación

En esta práctica se debe elaborar 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. El objetivo de esta actividad será el tratamiento, limpieza y validación de los datos, del dataset del Titanic: Machine Learning from Disaster (https://www.kaggle.com/c/titanic).

Descripción del dataset

Carga de los datos

Cargamos los archivos de datos de train, test y gender_submission:

```
data_train<-read.csv("../csv_original/train.csv", header=T, sep=",")
data_test_original<-read.csv("../csv_original/test.csv", header=T, sep=",")
data_real<-read.csv("../csv_original/gender_submission.csv", header=T, sep=",")
str(data_train)
```

```
## '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 58
##  $ 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 ...
```

```
str(data_test_original)
```

```
## 'data.frame':    418 obs. of  11 variables:
##  $ PassengerId: int  892 893 894 895 896 897 898 899 900 901 ...
##  $ Pclass     : int  3 3 2 3 3 3 3 2 3 3 ...
##  $ Name       : Factor w/ 418 levels "Abbott, Master. Eugene Joseph",..: 210 409 273 414 182 370 85 5
##  $ Sex        : Factor w/ 2 levels "female","male": 2 1 2 2 1 2 1 2 1 2 ...
##  $ Age        : num  34.5 47 62 27 22 14 30 26 18 21 ...
##  $ SibSp      : int  0 1 0 0 1 0 0 1 0 2 ...
##  $ Parch      : int  0 0 0 0 1 0 0 1 0 0 ...
##  $ Ticket     : Factor w/ 363 levels "110469","110489",..: 153 222 74 148 139 262 159 85 101 270 ...
```

```
##  $ Fare       : num  7.83 7 9.69 8.66 12.29 ...
##  $ Cabin      : Factor w/ 77 levels "","A11","A18",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Embarked   : Factor w/ 3 levels "C","Q","S": 2 3 2 3 3 3 2 3 1 3 ...
```

```
str(data_real)
```

```
## 'data.frame':    418 obs. of  2 variables:
##  $ PassengerId: int  892 893 894 895 896 897 898 899 900 901 ...
##  $ Survived   : int  0 1 0 0 1 0 1 0 1 0 ...
```

El dataset está compuesto de dos conjuntos uno de entrenamiento y otro de test- El conjunto de entrenamiento, tiene un total de 891 observaciones/muestras y 12 variables/atributos, de los cuales uno es nuestro atributo objetivo (Survived) que será 0 si no sobrevivió o 1 si sobrevivió. El conjunto de test, está compuesto por 418 observaciones/muestras y un total de 11 variables/atributos, puesto que el atributo objetivo es el que tenemos que predecir. La predicción real de estas observaciones está en el fichero gender_submission.csv. En esta tabla se muestran las variables/atributos:

1. PassengerId: Id del pasajero. Tipo: integer.
2. Survived: Si sobrevivieron, 0=no, 1=sí. Tipo: Factor.
3. Pclass: Clase, 1=primera, 2=segunda, 3=tercera. Tipo: Factor.
4. Name: Nombre. Tipo: Factor.
5. Sex: Sexo. Tipo: Factor.
6. Age: Años de edad. Tipo: Num.
7. SibSp: Número de parientes como hermanos, hermanas, esposo o esposa.. Tipo: Integer.
8. Parch: Número de parientes como padre, madre, hijo, hija.. Tipo: Integer.
9. Ticket: Número de ticket. Tipo: Factor.
10. Fare: Tarifa de pasajero. Tipo: Num.
11. Cabin: Número de cabina. Tipo: Factor.
12. Embarked Puerto de embarque, C=Cherbourg, Q=Queenstown, S=Southampton. Tipo: Factor.

¿Por qué es importante y qué pregunta/problema pretende responder?

La pregunta/problema que se pretende responder es qué grupos de personas sobrevivieron a la tragedia del Titanic. Es decir qué clase de personas tenían más números de sobrevivir, ¿niños, mujeres, clase alta? Se trata por tanto de crear un modelo a partir de los datos de entrenamiento que dada una observación del conjunto de test sea capaz de predecir si el pasajero sobrevivió o no. Esta pregunta es binaria, es decir que la clasificación que se pretende hacer con el dataset es superviviente y no supervivientes.

Limpieza de los datos

Integración

Como los datos de test los tenemos en dos archivos/datasets distintos, los uniremos usando la función de r merge, por la variable PassangerId, obteniendo así un único dataset de test para saber el accuracy de nuestro modelo, creado con el conjunto de entrenamiento.

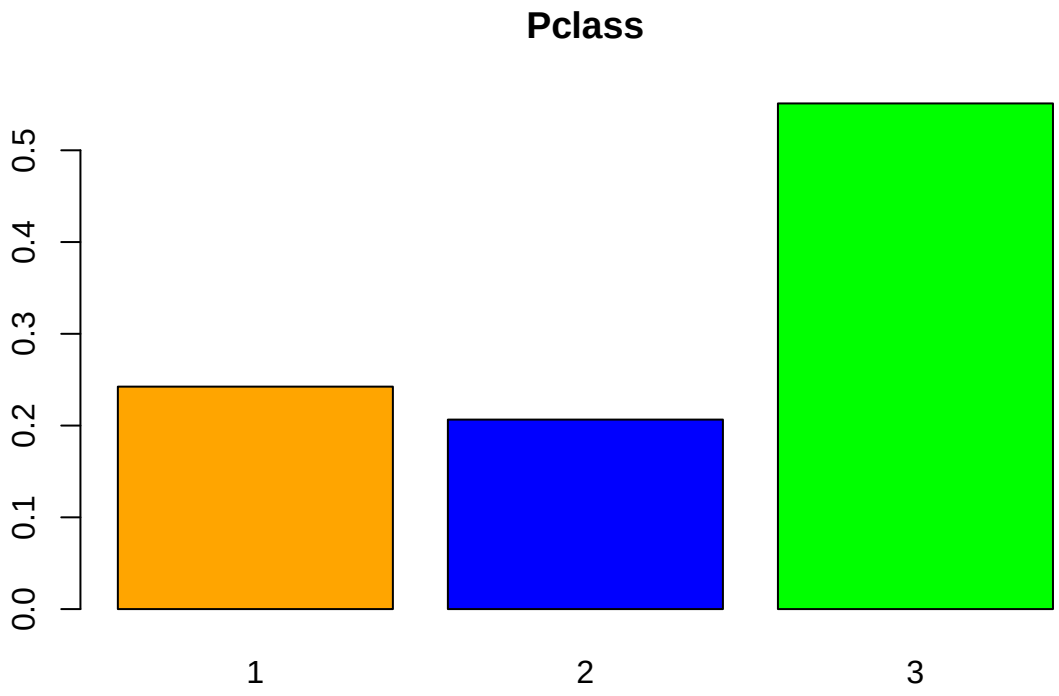
```
data_test<-merge(data_test_original, data_real, by.x="PassengerId", by.y="PassengerId")
str(data_test)
```

```
## 'data.frame':    418 obs. of  12 variables:
##  $ PassengerId: int  892 893 894 895 896 897 898 899 900 901 ...
##  $ Pclass     : int  3 3 2 3 3 3 3 2 3 3 ...
##  $ Name       : Factor w/ 418 levels "Abbott, Master. Eugene Joseph",..: 210 409 273 414 182 370 85 5
##  $ Sex        : Factor w/ 2 levels "female","male": 2 1 2 2 1 2 1 2 1 2 ...
##  $ Age        : num  34.5 47 62 27 22 14 30 26 18 21 ...
##  $ SibSp      : int  0 1 0 0 1 0 0 1 0 2 ...
##  $ Parch      : int  0 0 0 0 1 0 0 1 0 0 ...
##  $ Ticket     : Factor w/ 363 levels "110469","110489",..: 153 222 74 148 139 262 159 85 101 270 ...
##  $ Fare       : num  7.83 7 9.69 8.66 12.29 ...
##  $ Cabin      : Factor w/ 77 levels "","A11","A18",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Embarked   : Factor w/ 3 levels "C","Q","S": 2 3 2 3 3 3 2 3 1 3 ...
##  $ Survived   : int  0 1 0 0 1 0 1 0 1 0 ...
```

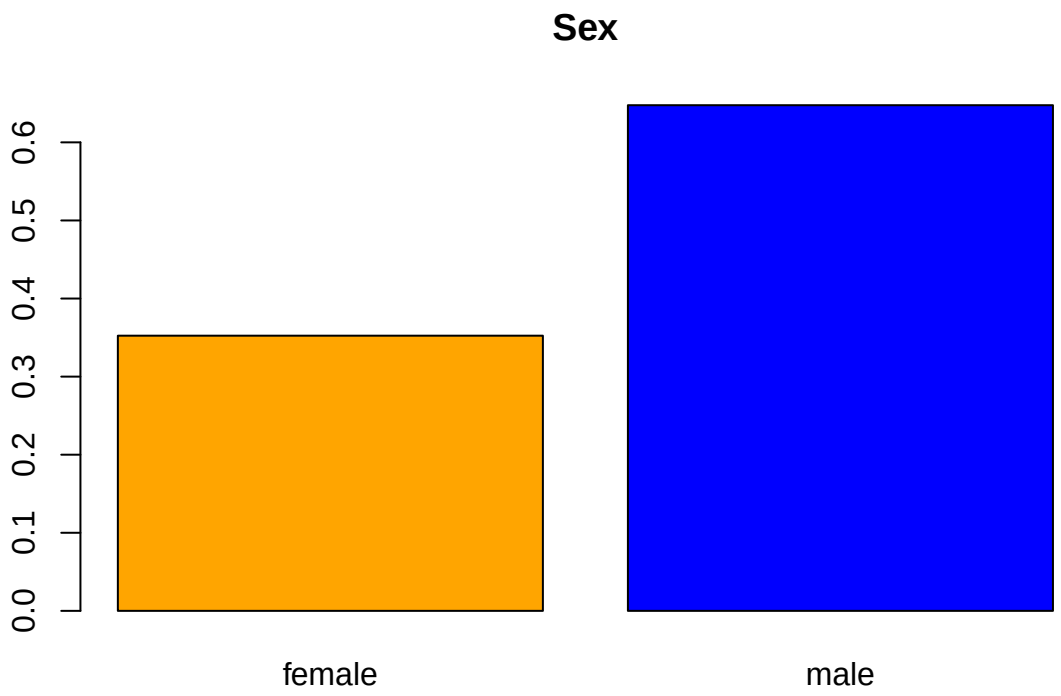
Análisis visual de los datos

Las variables de tipo factor, que son variables cualitativas que pueden tomar pocos valores, se representarán visualmente mediante diagramas de barras:

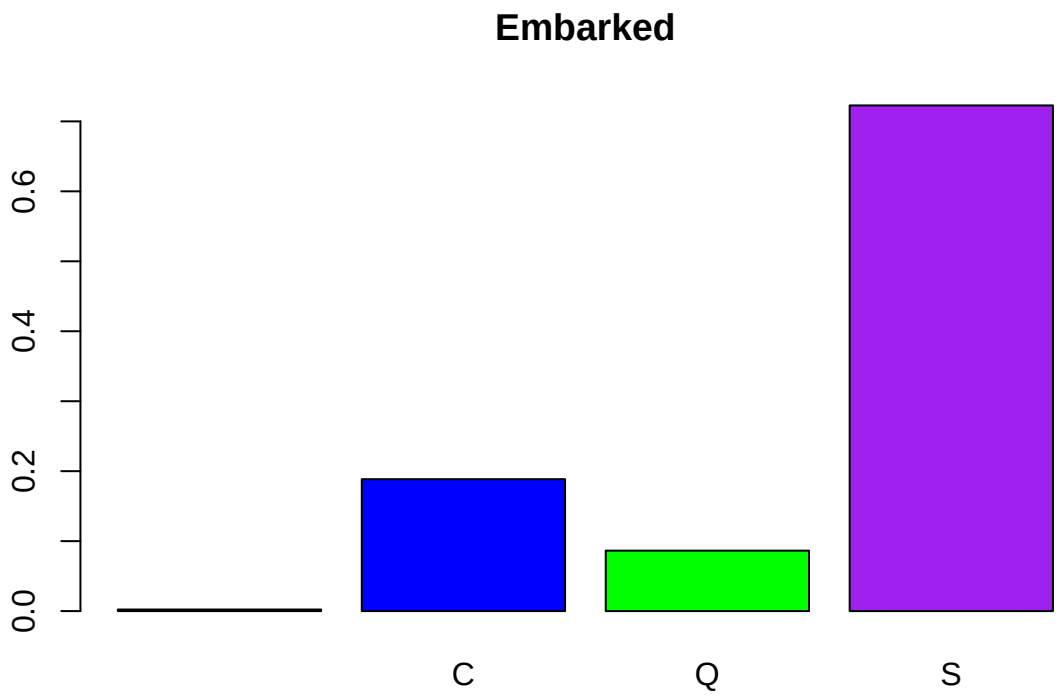
```
colors=c("orange","blue","green", "purple")
barplot(prop.table(table(data_train$Pclass)),col=colors,main="Pclass")
```



```
barplot(prop.table(table(data_train$Sex)),col=colors,main="Sex")
```

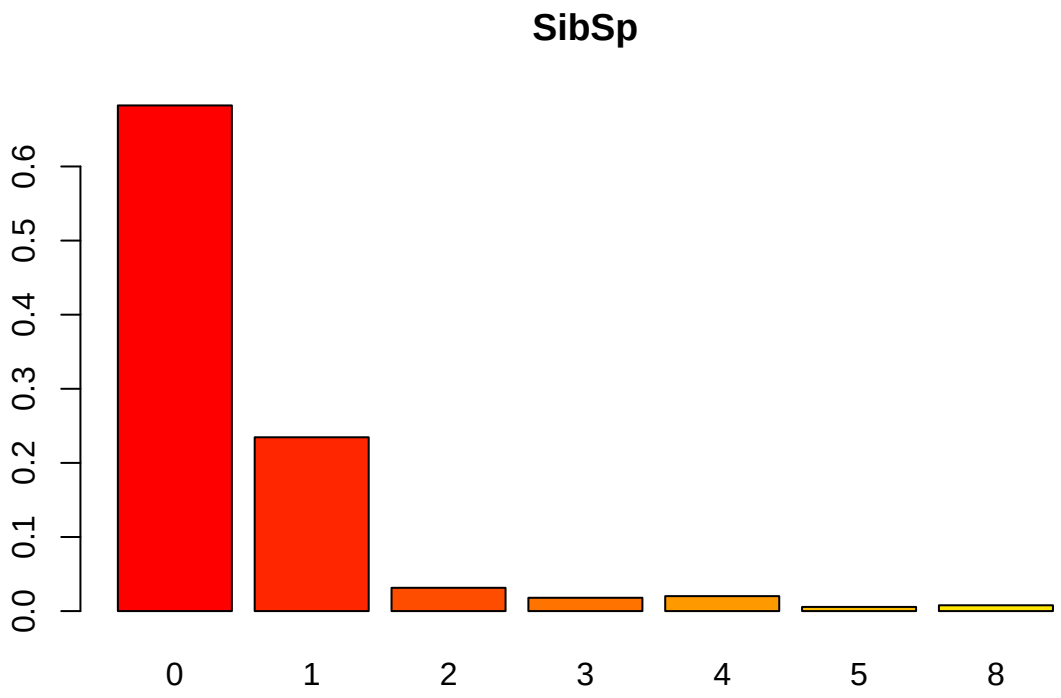



```
barplot(prop.table(table(data_train$Embarked)),col=colors,main="Embarked")
```

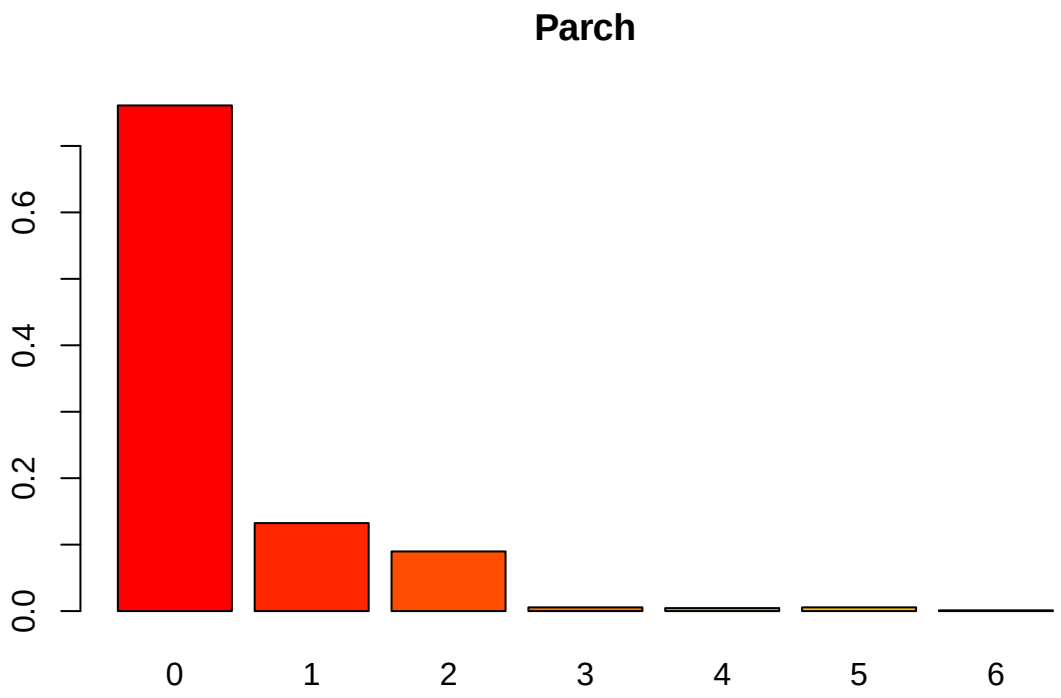


Las variables de tipo integer, que son variables cuantitativas discretas que no toman demasiados valores, se representarán visualmente mediante diagramas de barras:

```
#install.packages("rainbow")#paleta de colores
library("rainbow")
colors=rainbow(40)
barplot(prop.table(table(data_train$SibSp)),col=colors,main="SibSp")
```

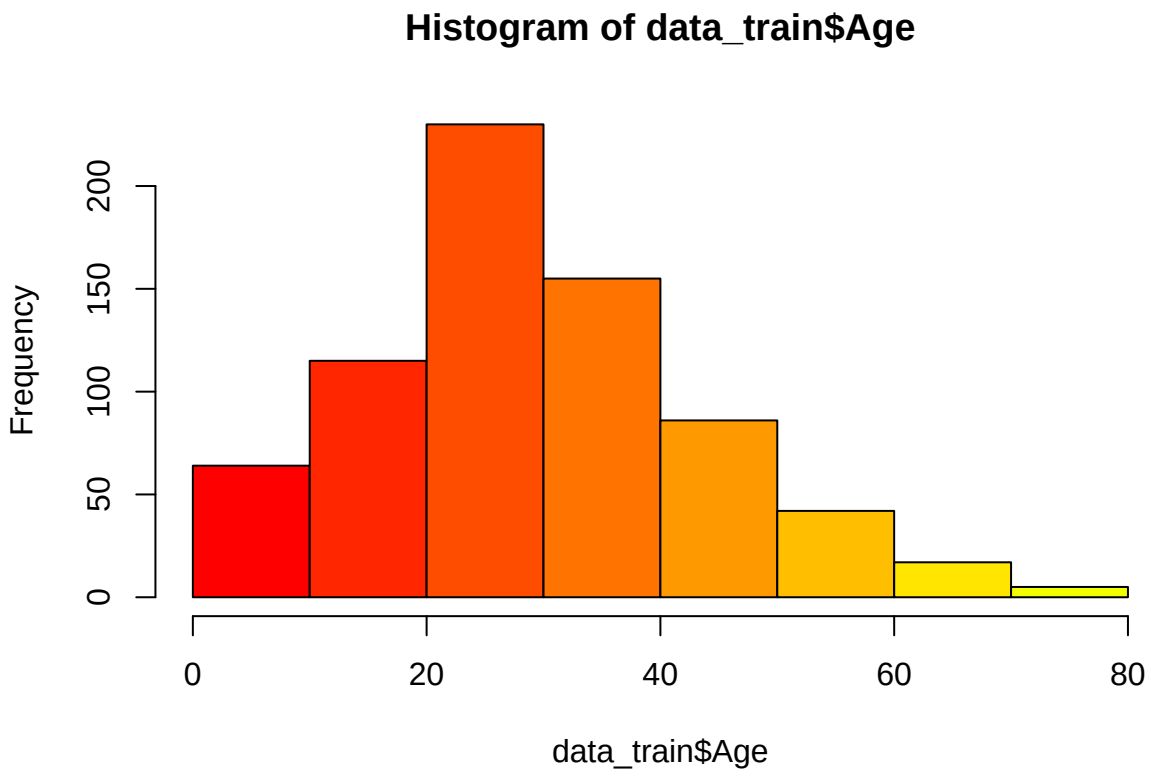



```
barplot(prop.table(table(data_train$Parch)),col=colors,main="Parch")
```

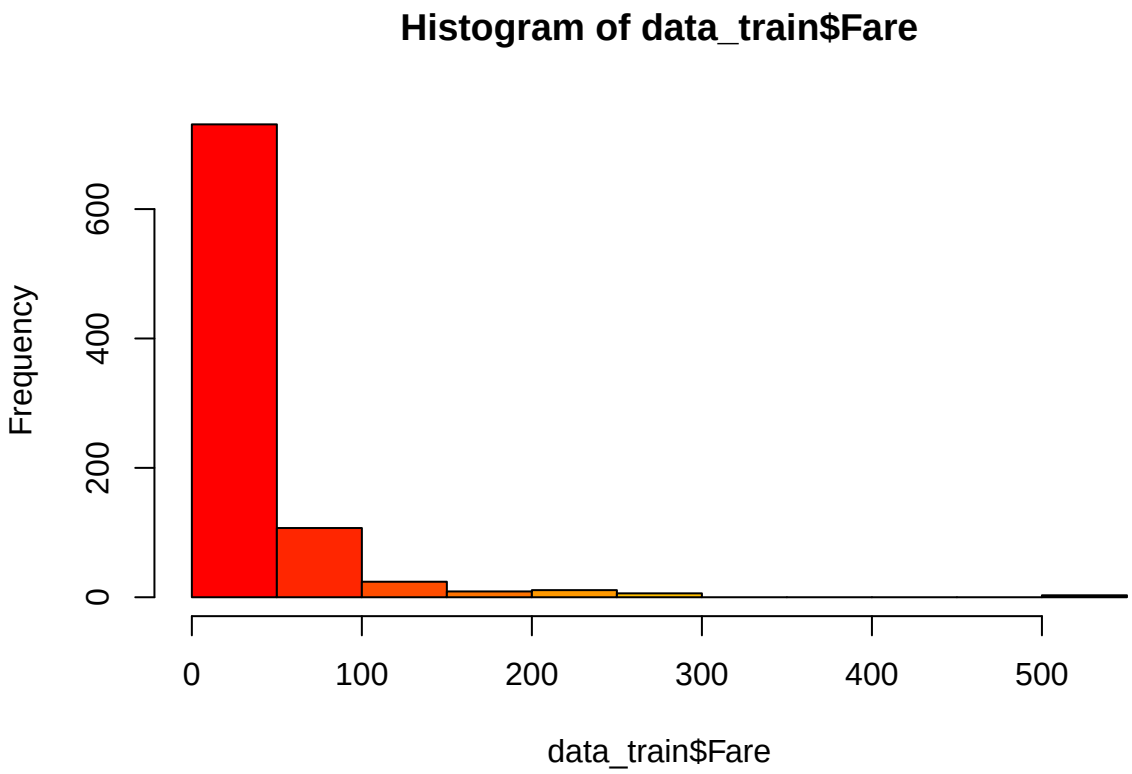


Las variables de tipo numeric, que son variables cuantitativas continuas, se representarán visualmente mediante histogramas:

```
hist(data_train$Age,col=colors)
```




```
hist(data_train$Fare,col=colors)
```



Selección de los datos de interés a analizar

Existen variables/atributos que no aportan informacion relevante a la supervivencia de los pasajeros. Estas variables son: PassengerId, Name, Ticket y Cabin.

Datos perdidos

```
colSums(is.na(data_train))
```

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

```
colSums(data_train=="")
```

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

Valores extremos

Análisis de los datos

Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar)

Comprobación de la normalidad y homogeneidad de la varianza

Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes

---

Representación de los resultados

Resolución del problema

A partir de los resultados obtenidos, ¿cuáles son las conclusiones?

¿Los resultados permiten responder al problema?