El RMS Titanic fue en su momento el mayor barco de pasajeros del mundo, hundiéndose en su viaje inaugural de Southampton a Nueva York en el año 1912. En el evento fallecieron 1514 de las 2223 personas que iban a bordo, entre tripulación y pasajeros.

En el presente práctico se trabajará con el conjunto de datos titanic, que figura en el archivo titanic.csv. El conjunto de datos es un clásico de las competencias de "Machine Learning", donde se busca determinar un mecanismo de clasificación que, en función de diversos variables de cada pasajero, prediga si el pasajero sobrevivió o no a la catástrofe. Las variables del conjunto de datos son:

survival: supervivencia (0 No, 1 Sí).

pclass: clase del pasajero (1,2 o 3).

sex: sexo del pasajero ("male", "female").

age: edad del pasajero.

sibsp: cantidad de hermanos y cónyuges (totalizado) embarcados (número entero).

parch: cantidad de padres e hijos (totalizado) embarcados (número entero).

ticket: código del boleto (texto).

fare: tarifa del pasaje (número real).

embarked: puerto de embarque (S= Southampton, Q=Queenstown, C = Cherbourg)

Los datos contienen 1028 pasajeros y algunas variables contienen respuestas faltantes.

1. Borrar todos los objetos existentes en el entorno de trabajo y establecer directorio de trabajo.
2. Leer el conjunto de titanic.csv teniendo en cuenta que en la primera línea del archivo figura el nombre de las variables y el tipo de separación de los datos y asígnelo al data.frame titanic.
3. Inspeccionar los primeros casos del archivo y los últimos.
4. Abrir con el editor al data.frame e inspeccionar el archivo.
5. Establecer el número de variables y de casos.
6. Realizar un attach de titanic.
7. Inspeccionar los nombres de las variables de titanic e identificar de qué tipo de variable se trata cada una de ellas.
8. Calcule la chance de sobrevivir siendo hombre. Calcule la chance de sobrevivir siendo mujer.
9. Cree que el tipo de ticket del pasajero (clase de cabina) esta asociado con su supervivencia?
10. Estudie la distribución de las tarifas. Que observa? Le parece razonable suponer que la variable tarifa tenga distribución normal? Calcule la media y la mediana. Puede decidir de antemano quién es mas grande si la media o la mediana?

11. Estudie la relación entre tarifa y clase y entre edad y clase.

12. Respecto a la relación entre la edad y la tarifa podemos pensar que las personas más jóvenes tenían menos dinero y por ende compraron los tiquetes más baratos. Puede confirmar esto en base a los datos?

La idea de los puntos anteriores era explorar los datos usando boxplots, histogramas, scatterplots. Si le faltó usar alguno de estos gráficos vuelva a empezar.