**Dataset: Titanic**

*Luis Manuel Camacho Puerma*

*Carla Comas García*

*14 de Abril del 2019*

Contenido

[1. Detalles de la actividad 3](#_Toc11172769)

[1.1. Descripción 3](#_Toc11172770)

[1.2. Objetivos 3](#_Toc11172771)

[1.3. Competencias 3](#_Toc11172772)

[2. Resolución 4](#_Toc11172773)

[2.1. Descripción del dataset 4](#_Toc11172774)

[2.2 Importancia y objetivos del análisis 4](#_Toc11172775)

[2.3 Limpieza de los datos 4](#_Toc11172776)

[2.3.1 Selección de datos de interés 5](#_Toc11172777)

[2.3.2 Ceros y elementos vacíos 5](#_Toc11172778)

[2.3.3 Valores extremos 6](#_Toc11172779)

[2.3.4 Valores categóricas 7](#_Toc11172780)

[2.3.5 Guardar preprocesado 8](#_Toc11172781)

[2.4 Análisis de los datos 8](#_Toc11172782)

[2.4.1 Selección del grupos de datos 8](#_Toc11172783)

[2.4.2 Comprovació de la normalitat i homogeneïtat de la variància 9](#_Toc11172784)

[2.4.3 Aplicació de proves estadístiques 10](#_Toc11172785)

[2.5 Representación de resultados 13](#_Toc11172786)

[2.5.1 Gráfico entre las variables 13](#_Toc11172787)

[2.5.2 Gráfico de correlación entre las varibales 13](#_Toc11172788)

[2.5.3 Gráfico de la regresión 14](#_Toc11172789)

[2.6 Conclusiones 15](#_Toc11172790)

[3. Recursos 15](#_Toc11172791)

[4. Tabla contribuciones 15](#_Toc11172792)

# Detalles de la actividad

## Descripción

En esta actividad se elabora un caso práctico, consistente en el tratamiento de un conjunto de datos (en inglés, dataset), 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.

## Objetivos

Los objetivos que se persiguen mediante el desarrollo de esta actividad práctica son los

siguientes:

* 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.

## Competencias

Así, las competencias del Máster en Data Science que se desarrollan son:

* 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.

# Resolución

## Descripción del dataset

El conjunto de datos objeto de análisis se ha obtenido a partir del enlace de Kaggle proporcionado en el enunciado de la práctica. El dataset es el referente a los pasajeros del Titanic y está constituido por 12 características (columnas) que presentan 891 pasajeros (filas). Entre los campos de este conjunto de datos, encontramos los siguientes:

* **PassengerId:** (Variable #) número que identifica al pasajero.
* **Survived:** (Variable #) Si sobrevive al accidente o no.
* **Pclass:** (Variable #) clase en la que viaja el pasajero (1ª, 2ª o 3ª).
* **Name:** (Variable Categórica) nombre del pasajero
* **Sex:** (Variable Categórica) sexo del pasajero.
* **Age:** (Variable #) edad del pasajero.
* **SibSp:** (Variable Categórica) número de relaciones familiares en el barco 🡪 Hermanos o esposa/amante
* **Parch**: (Variable #) número de relaciones familiares en el barco 🡪 Padres o hijos (0 para niños que viajan con niñera).
* **Ticket:** (Variable Categórica) número del ticket
* **Fare:** (Variable #) tarifa del pasajero.
* **Cabin:** (Variable Categórica) número de la cabina donde se aloja.
* **Embarked:** (Variable Categórica) lugar donde embarco el pasajero (C- Cherbourg, S - Southampton, Q = Queenstown).

## Importancia y objetivos del análisis

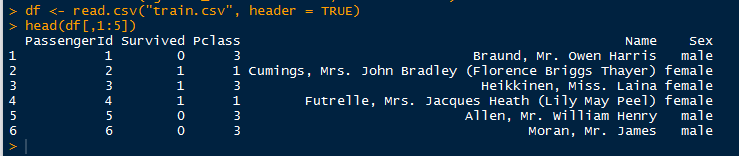
## 

A partir de este conjunto de datos se plantea saber si existe alguna variable que inluyera en la supervivencia o no del titanic.

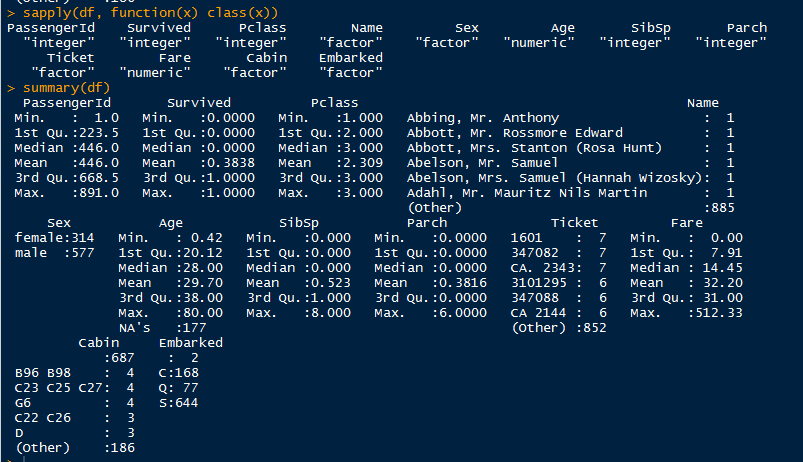
Este análisis tendría una gran importancia para una empresa cuyo negocio sea realizar viajes en barco. Esto influiría en el coste de los billetes según el lugar en el que se decida comprar. Para ello nos proponemos a estudiar los datos descritos anteriormente, con el fin de obtener algún resultado remarcable.

## Limpieza de los datos

Antes de comenzar con la limpieza de los datos, procedemos a realizar la lectura del ﬁchero en formato CSV en el que se encuentran. El resultado devuelto por la llamada a la función read.csv() será un objeto data.frame:



Pero para empezar a estudiar el dataset debemos tener una idea clara de que tipo de variables tratamos y que unidades tienen. Una vez sabidos estos nos interesará cuantos datos absentas existen y como están los datos repartidos, para ello ejecutaremos la función summary.



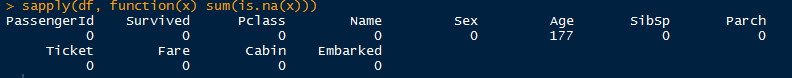
### Selección de datos de interés

La gran mayoría de los atributos presentes en el conjunto de datos se corresponden con características que reúnen las diversas películas recogidos en forma de registros, por lo que será conveniente tenerlos en consideración durante la realización de los análisis.

En este apartado podemos comprobar todas las variables tienen un significado relativo a la variable endógena es decir que no nos encontramos con ningún casó de único valor o que no nos aporte información sobre el estudio.

### Ceros y elementos vacíos

Comúnmente, se utilizan los ceros como centinela para indicar la ausencia de ciertos valores. Sin embargo, no es el caso de este conjunto de datos puesto que, como se comentó, se utilizó NA para denotar un valor desconocido. Así, se procede a conocer a continuación qué campos contienen elementos vacíos:



Llegados a este punto debemos decidir cómo manejar estos registros que contienen valores desconocidos para algún campo. Una opción podría ser eliminar esos registros que incluyen este tipo de valores, pero ello supondría desaprovechar información.

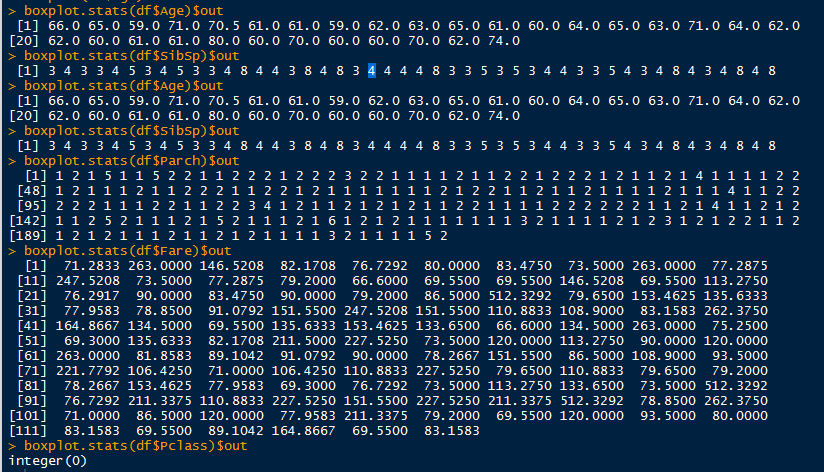
La función pertinente sería: Df\_sinNA <- na.omit(df)

Como alternativa, se empleará un método de imputación de valores basado en la similitud o diferencia entre los registros: la imputación basada en k vecinos más próximos (en inglés, kNN-imputation). La elección de esta alternativa se realiza bajo la hipótesis de que nuestros registros guardan cierta relación. No obstante, es mejor trabajar con datos “aproximados” que con los propios elementos vacíos, ya que obtendremos análisis con menor margen de error.

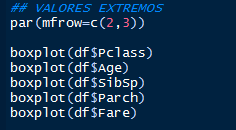
### Valores extremos

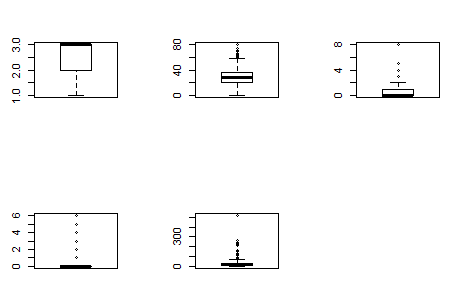
Los valores extremos o outliers son aquellos que parecen no ser congruentes sin los comparamos con el resto de los datos. Para identiﬁcarlos, podemos hacer uso de dos vías:

1. representar un diagrama de caja por cada variable y ver qué valores distan mucho del rango intercuartílico (la caja)
2. utilizar la función boxplots.stats() de R, la cual se emplea a continuación. Así, se mostrarán sólo los valores atípicos para aquellas variables que los contienen:



Para poder entender y ver gráficamente…





No obstante, si revisamos los anteriores datos, comprobamos que son valores que perfectamente pueden darse. Es por ello que el manejo de estos valores extremos consistirá en simplemente dejarlos como actualmente están recogidos.

### Valores categóricas

Una vez ya hemos hecho la limpieza de datos se pasa a estudiar nuestro dataset. Normalmente y por lo que nos dice nuestra experiencia los datos acostumbran a trabajar con datos y clientes normales es que no podemos perder ningún tipo de información. Es por este motivo que hemos decidido crear este nuevo apartado para ver como tratamos las variables categóricas.

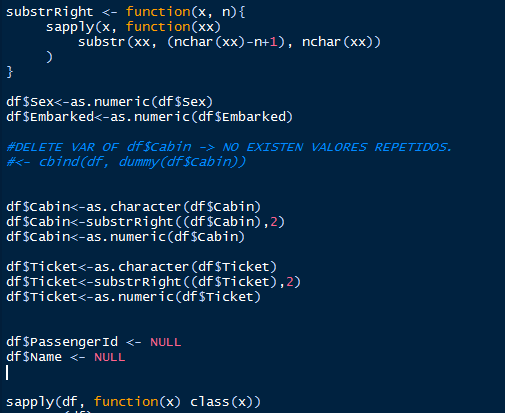
Para ello hemos identificado cuales son y las hemos convertido a numéricas. Por lo que nuestra experiencia indica, cualquier modelo que intentes aplicar va a dar mejores resultados en caso de tener variables numéricas en vez de categóricas.

Por otra parte eliminamos las dos variables que no nos indica nada sobre la variable que definiremos endógena (el nombre y el id).

Para acabar hacemos una transformación de dos variables que parecen seguir el mismo tipo que el nombre pero que enrealidad nos podrían aportar algo. Estas variables son el número de ticket y el número de cabina. Para el número de cabina vamos a numerar en valor numérico que cabinas existen. Lo que hacemos es coger las últimos dos valores y con esto es suficiente para diferenciarlas entre ellas (no existen más de una cabina que acaben igual).

Por otro lado, el número de ticket veremos si existen supersticiones sobre lo que podría ser el número 13 por ejemplo.

Se ha planteado usar dummies para categorías como el tipo de clase para poder extraer información más concreta sobre que clase pertenece y no sobre que significa este valor. De todas maneras no parece ser relevante para este tipo de análisis de dependencia. En caso de querer ejecutar modelo si seria necesario aplicar dummies y multiplicar las columnas según los factores que tengan.



### Guardar preprocesado

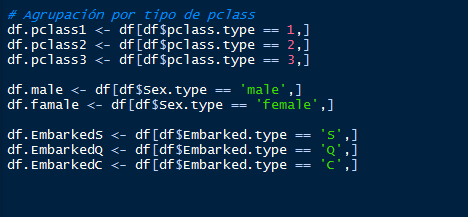
Una vez que hemos acometido sobre el conjunto de datos inicial los procedimientos de integración, validación y limpieza anteriores, procedemos a guardar estos en un nuevo ﬁchero denominado Automobile\_data\_clean.csv:



## Análisis de los datos

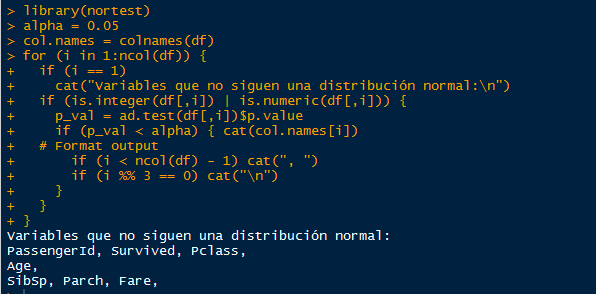
### Selección del grupos de datos

A continuación, se seleccionan los grupos dentro de nuestro conjunto de datos que pueden resultar interesantes para analizar y/o comparar. No obstante, como se verá en el apartado consistente en la realización de pruebas estadísticas, no todos se utilizarán.

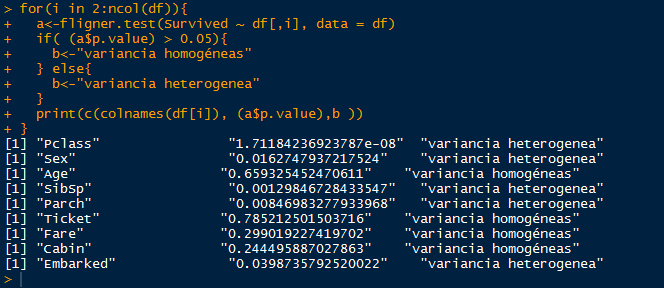


### Comprovació de la normalitat i homogeneïtat de la variància

Para la comprobación de que los valores que toman nuestras variables cuantitativas provienen de una población distribuida normalmente, utilizaremos la prueba de normalidad de Anderson Darling. Así, se comprueba que para que cada prueba se obtiene un p-valor superior al nivel de signiﬁcación preﬁjado α = 0,05. Si esto se cumple, entonces se considera que variable en cuestión sigue una distribución normal.



Debido a que nuestras variables no siguen una normal, y completamente esperable debido a que muchas de ellas provienen de factores, pasamos a estudiar la homogeneidad de varianzas mediante la aplicación de un test de Fligner-Killeen. En el siguiente test, la hipótesis nula consiste en que ambas varianzas son iguales.



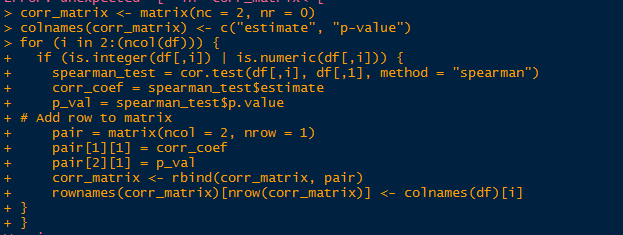
Puesto que obtenemos un p-valor superior a 0,05, aceptamos la hipótesis de que las varianzas de las muestras serán homogéneas.

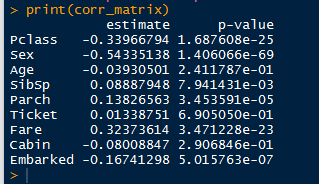
En este caso las variables Pclass Sex SibSp Parcg Embarket tendrán variancias heterogéneas.

### Aplicació de proves estadístiques

#### Análisis de correlacions

En primer lugar, procedemos a realizar un análisis de correlación entre las distintas variables para determinar cuáles de ellas ejercen una mayor inﬂuencia sobre el si la persona va a sobrevivir finalmente. Para ello, se utilizará el coeﬁciente de correlación de Spearman, puesto que hemos visto que tenemos datos que no siguen una distribución normal.





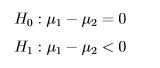
Así, identiﬁcamos cuáles son las variables más correlacionadas con el precio en función de su proximidad con los valores -1 y +1. Teniendo esto en cuenta, queda patente cómo la variable más relevante en la causa de supervivencia es el sexo de la persona (Sex). Nota. Para cada coeﬁciente de correlación se muestra también su p-valor asociado, puesto que éste puede dar información acerca del peso estadístico de la correlación obtenida.

#### Contraste de hipótesis de dos muestras sobre la diferencia de medias

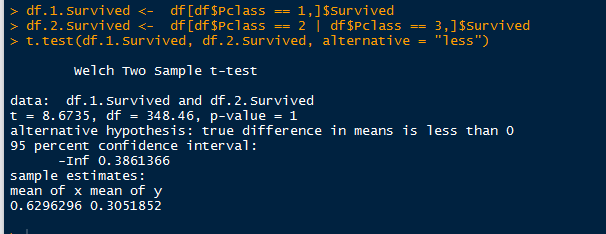
La segunda prueba estadística que se aplicará consistirá en un contraste de hipótesis sobre dos muestras para determinar si la probabilidad de supervivencia es superior en caso de que el viajero esté en primera clase. Para ello, tendremos dos muestras: la primera de ellas se corresponderá con los a la clase más alta y la segunda a las clases más bajas.

Se debe destacar que un test paramétrico como el que a continuación se utiliza necesita que los datos sean normales, si la muestra es de tamaño inferior a 30. Como en nuestro caso, n > 30, el contraste de hipótesis siguiente es válido.

Así, se plantea el siguiente contraste de hipótesis de dos muestras sobre la diferencia de medias, el cual es unilateral atendiendo a la formulación de la hipótesis alternativa:



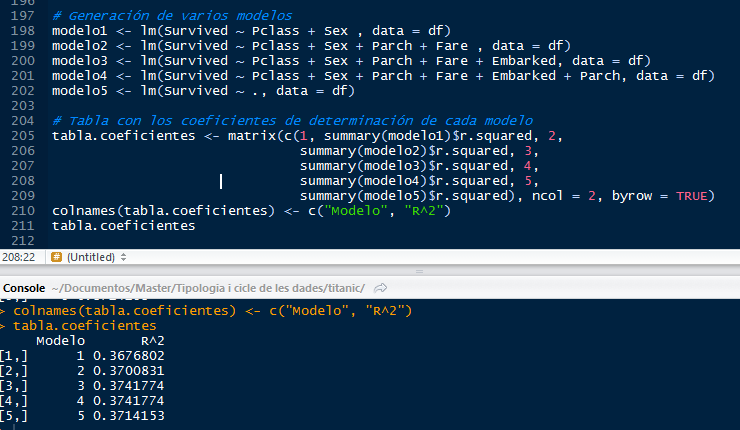
dónde µ1 es la media de la población de la que se extrae la primera muestra y µ2 es la media de la población de la que extrae la segunda. Así, tomaremos α =0,05.



Puesto que obtenemos un p-valor mayor que el valor de signiﬁcación ﬁjado, no rechazamos la hipótesis nula. Por tanto, podemos concluir que, la clase de categoría no influyó sobre la supervivencia de los viajeros.

#### Modelo de regresión lineal

Tal y como se planteó en los objetivos de la actividad, resultará de mucho interés poder realizar si la persona va a sobrevivir dependiendo de sus características. Así, se calculará un modelo de regresión lineal utilizando regresores tanto cuantitativos con el que poder realizar las predicciones. Para obtener un modelo de regresión lineal considerablemente eﬁciente, lo que haremos será obtener varios modelos de regresión utilizando las variables que estén más correladas con respecto a la supervivencia debido a que es lo que queremos predecir, según la tabla obtenida anteriormente. Así, de entre todos los modelos que tengamos, escogeremos el mejor utilizando como criterio aquel que presente un mayor coeﬁciente de determinación (R2).



Para los anteriores modelos de regresión lineal múltiple obtenidos, podemos utilizar el coeﬁciente de determinación para medir la bondad de los ajustes y quedarnos con aquel modelo que mejor coeﬁciente presente.

En este caso, nos quedaremos con el modelo 3, debido a que la diferencia entre el modelo 3 y 4 el coeficiente de determinación es el mismo, pero más óptimo el 3 debido a que hay menos variables exógenas. Lo importante es saber que variables son más relevantes. Que en este caos son: Pclass, Sex, Parch, Fare y Embarked.

Hemos dejado un apartado en R dónde completar las variables en caso de querer sacar la predicción de un cierto viajero. De todas maneras no tiene sentido si no se aplica en el futuro debido a que ya se sabe que viajeros se salvaron y que características cumplían.

\*\* Se podría haber hecho un análisis de componentes principales que pudiera haber sido más interesante para nuesto tipo de dataset, debido a que probablemente una combinación de las variables expliquen más que ellas individualmente.

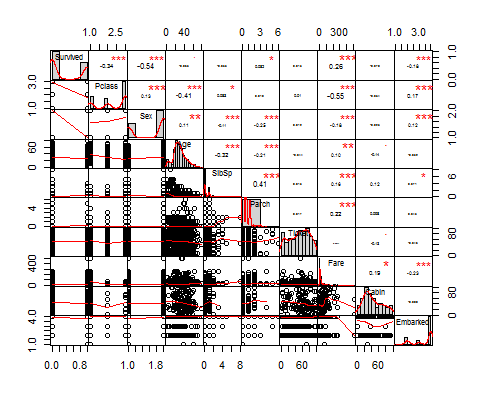
## Representación de resultados

### Gráfico entre las variables

Al estudiar una variable nos interesa que tenga la máxima correlación cada variable exógena con la variable endógena y además que entre las variables exógenas haya la mínima correlación.

Para ello hemos creado el siguiente gráfico que nos muestra en número a la parte superior derecha la correlación que existe marcando con asteriscos si esta es significativa.

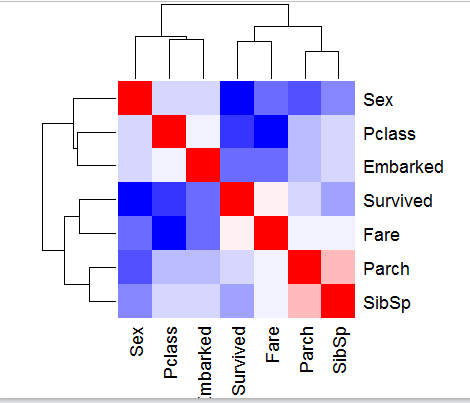
La parte izquierda inferior podemos ver los gráficos de cómo evoluciona cada una de las variables con la otra. Mientras que en la diagonal se puede ver que un estudio de que cantidad de registros (viajeros en nuestro caso) de esa categoría existe.



Confirmamos y justificamos todo lo que se ha comentado anteriormente con este gráfico.

### Gráfico de correlación entre las varibales

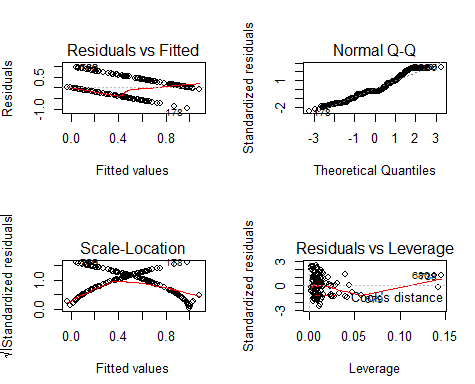
Seguidamente mostramos un gráfico de correlaciones más simple para poder comprobar que no hemos cogido variables muy correlacionadas. Se observa como el color rojo es el más correlacionada mientras que el máximo azul también pero en negativo; a medida que se acerca la correlación a nula, más claro será el color.



Las líneas superiores nos puede servir en caso de hacer un ACP pero no es lo que se ha estudiado y por lo tanto suprimimos esta parte de explicación.

### Gráfico de la regresión

Para acabar se muestra el gráfico de la regresión que se ha hecho anteriormente:



## Conclusiones

Se han realizado tres tipos de pruebas estadísticas sobre el conjunto de datos de los pasajeros del titanic con el objetivo de conocer si ciertas variables influían o no en la supervivencia de estos. Para cada una de las pruebas hemos podido comprobar los resultados, en algunos casos mediante tablas y en otros mediante una gráfica.

Mediante los diferentes análisis hemos podido determinar que las variables más significativas para determinar la supervivencia son el sexo, la clase y la tarifa de los pasajeros. Se podría no tener en cuenta la tarifa puesto que esta relacionada con la clase de estos. Estas variables influyen en gran medida en la supervivencia como se puede observar en la grafica entre las variables.

Para este estudio se han sometido los datos a un preprocesamiento para poder manejar los elementos vacíos y los valores extremos (outliers). Para el caso de los valores vacíos se ha hecho uso de un método de imputación de valores de tal forma que no tengamos que eliminar registros. Para el caso de los outliers se ha considerado no eliminarlos puesto que no parecen resultar del todo atípicos.

# Recursos

* Megan Squire (2015). Clean Data. Packt Publishing Ltd.
* Jason W. Osborne (2010). Data Cleaning Basics: Best Practices in Dealing with Extreme Scores. Newborn and Infant Nursing Reviews; 10 (1): pp. 1527-3369.

# Tabla contribuciones

|  |  |
| --- | --- |
| Contribuciones | Firma |
| Investigación Previa | CCG – LMCP |
| Documento Pdf | CCG – LMCP |
| Desarrollo Código | CCG – LMCP |
| GitHub | CCG – LMCP |