|  |  |
| --- | --- |
| M2.851  **-TIPOLOGíA Y CICLO DE VIDA DE LOS DATOS-** | **PRA 2**  **Limpieza y análisis de datos**  Fecha límite de entrega: 8 junio 2021  Autores:  **Olga Garcés Ciemerozum**  **Carlos Acosta Quintas**  Máster Universitario en Ciencia de Datos  Universitat Oberta de Catalunya |

INDICE DE CONTENIDOS

[Introducción 2](#_Toc72313457)

[1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder? 3](#_Toc72313458)

[1.1 Descripción del dataset. 3](#_Toc72313459)

[1.2 Por qué es importante el dataset? 5](#_Toc72313460)

[1.3 ¿Qué problema pretende responder el dataset?. 5](#_Toc72313461)

[2. Integración y selección de los datos de interés a analizar 6](#_Toc72313462)

[2.1 Integración. 6](#_Toc72313464)

[2.2 Selección. 6](#_Toc72313465)

[3. Limpieza de datos 6](#_Toc72313466)

[3.1 ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos? 6](#_Toc72313468)

[3.2 Identificación y tratamiento de valores extremos. 7](#_Toc72313469)

[4. Análisis de datos 8](#_Toc72313470)

[4.1 Selección de los grupos de datos que se quieren analizar/comparar (planificaciónde los análisis a aplicar). 8](#_Toc72313472)

[4.2 Comprobación de la normalidad y homogeneidad de la varianza. 8](#_Toc72313473)

[4.3 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.Comprobación de la normalidad y homogeneidad de la varianza. 8](#_Toc72313474)

[5. Representación de los resultados a partir de tablas y gráficas 9](#_Toc72313475)

[6. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema? 10](#_Toc72313477)

[6.1 Resolución del problema. 10](#_Toc72313479)

[6.2 Conclusiones a partir de los datos obtenidos. 10](#_Toc72313480)

[6.3 ¿Los resultados permiten responder al problema? 10](#_Toc72313481)

[7. Código 11](#_Toc72313482)

[8. Tabla de contribuciones al trabajo 12](#_Toc72313484)

[Referencias / Fuentes de Información 12](#_Toc72313486)

# Introducción

El presente informe forma parte de la segunda práctica de la asignatura M2.851 - Tipología y ciclo de vida de los datos del Máster Universitario en Ciencia de Datos impartido por la Universitat Oberta de Catalunya.

En esta práctica se realizarán técnicas de limpieza de datos aplicadas a un juego de datos determinado y también se analizarán dichos datos para extraer información relevante y útil.

A su vez, se entregará, junto con la presente memoria, una serie de archivos con el código necesario para la realización de la limpieza y análisis con el que el usuario podrá realizar diferentes estudios analíticos a posteriori si lo desease.

# Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?

# Descripción del dataset.

El dataset Titanic reune los datos sobre los pasajeros que viajaban a bordo del Titanic y registra para cada persona su supervivencia o no en el accidente. El Titanic transportaba a pasajeros con gran diversidad en sus niveles de renta y edad y a bordo se encontraban familias enteras.

La etiqueta (variable a predecir) es la variable dicotómica que indica si el viajero ha sobrevivido o no.

La ubicación en kaggle del dataset utilizado se muestra en el siguiente link:

<https://www.kaggle.com/c/titanic/data>

Los archivos disponibles son 3 y están en formato csv. Sus nombres son:

* **train.csv**
* **test.csv**
* **gender\_submission.csv**: Ejemplo a seguir en la entrega de la competición Kaggle (no útil).

Según los registros, en el Titanic viajaban 2229 personas, de las cuales 913 formaban parte de la tripulación del barco. El dataset que obtenemos de Kaggle tiene un total de 1309 registros, por lo tanto, no todos los pasajeros que viajaban a bordo están incluidos en el dataset y podemos asumir que el juego de datos es una muestra de toda la población a analizar.

El dataset original está compuesto por dos ficheros: el fichero pensado para realizar el entrenamiento de un modelo (**train.csv**) y el fichero con los datos destinados a testear la calidad del modelo (**test.csv**). El fichero de entrenamiento contiene una columna más que el fichero de prueba. Esta columna corresponde a la columna de la clase "**Survived**".

El fichero de entrenamiento tiene 891 registros mientras que el fichero de test contiene 418 instancias.

Las variables de las que se compone el dataset son y sus unidades o magnitudes de las características son:

**PassengerId:** Identificador del pasajero

*Tipo:* Entero indicando un identificador único de casa instancia.

**Survived:** Indica si el pasajero ha sobrevivido la catástrofe

*Tipo:* Entero (categórica) 0 = No ha sobrevivido; 1 = Ha sobrevivido

**Pclass:** Clase en la que viajaba el pasajero

*Tipo:* String (categórica) 1 = 1 a clase; 2 = 2 a clase; 3 = 3a clase

**Name:** Nombre del pasajero

*Tipo:* String

**Sex:** Sexo del pasajero

*Tipo:* String (categórica) female = Mujer; male = hombre

**Age:** Edad del pasajero

*Tipo:* Entero

**SibSp:** Indica si el pasajero tenía hermanos o pareja a bordo

*Tipo:* Entero

**Parch:** Indica si el pasajero tenía padres o hijos a bordo

*Tipo:* Entero

**Ticket:** Número del billete

*Tipo:* String alfanumérico

**Fare**: Precio del billete sin especificar si es un billete individual o grupal

*Tipo:* Número Real

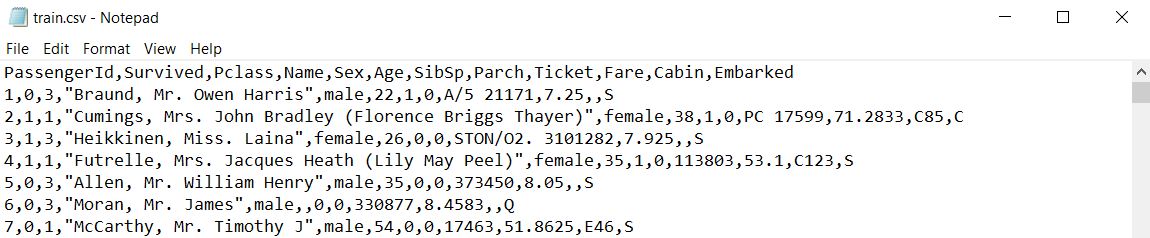
**Cabin:** Número de camarote

*Tipo:* String

**Embarked:** Indica si el pasajero ha embarcado o no y donde:

*Tipo:* String (categórica) C = Cherbourg, Q = Queenstown, S = Southampton

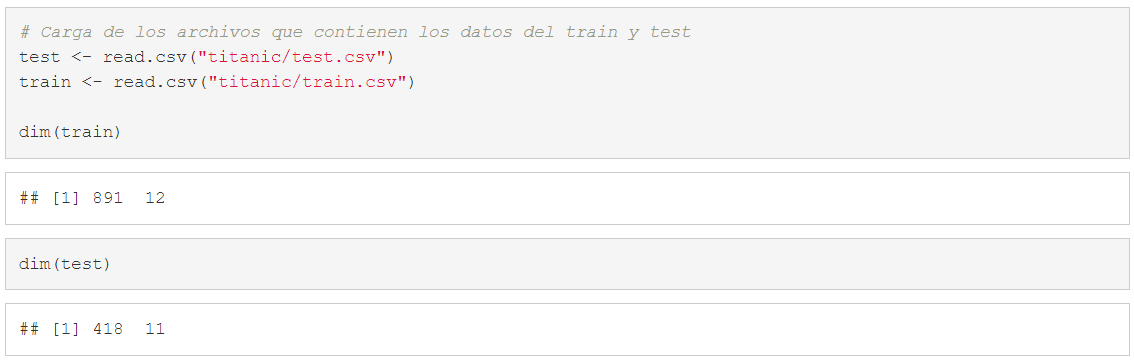
Observamos en primera instancia el juego de datos en Bloc de Notas. Observamos su estructura y vemos que la separación de los campos es a través de la “,”.

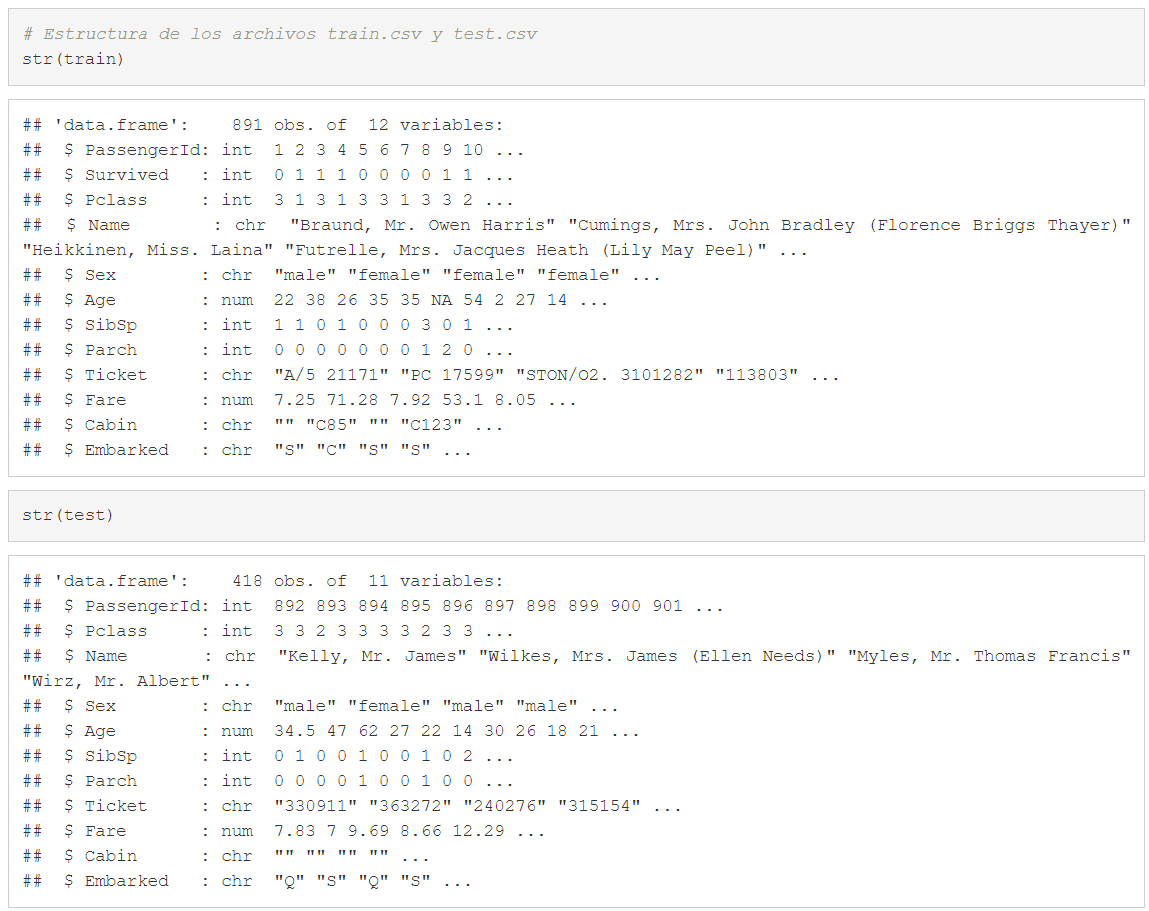


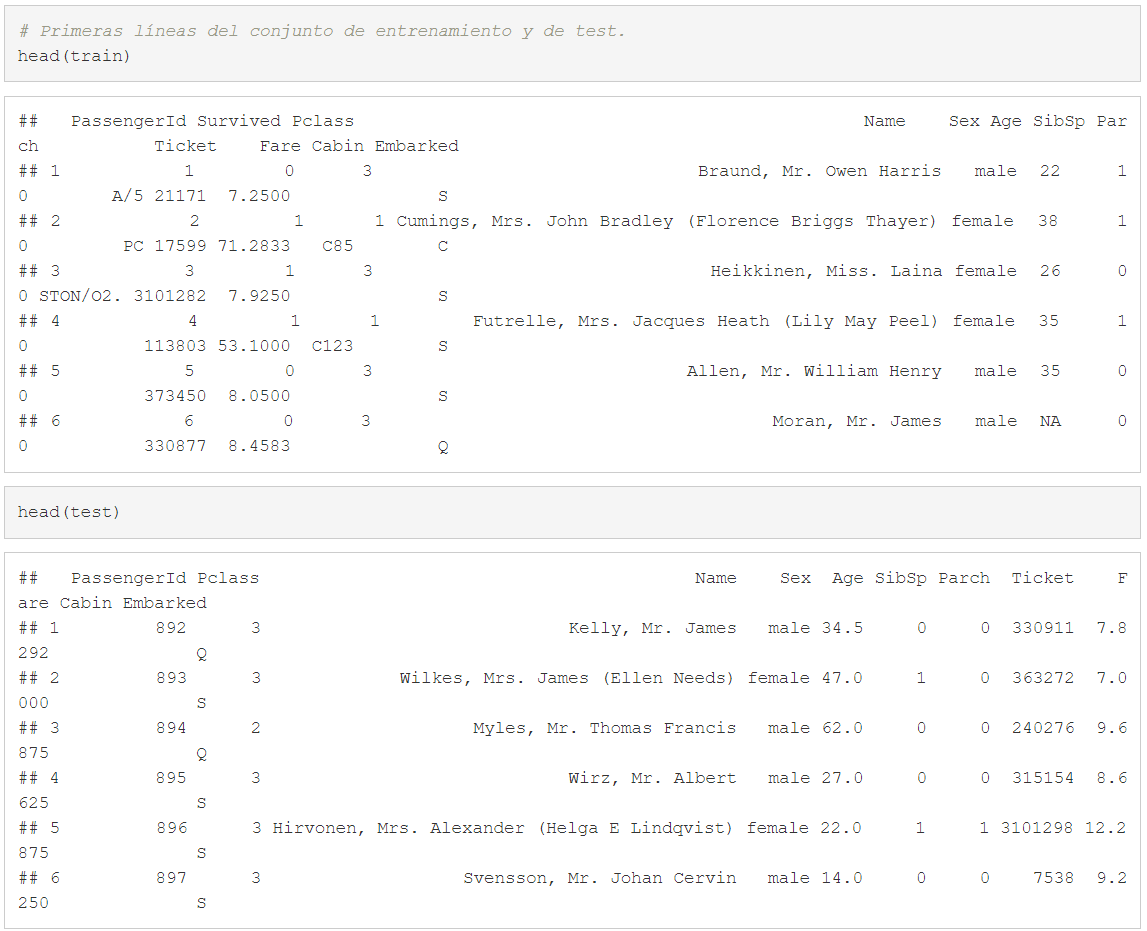
Los datos no han pasado por un proceso de preprocesado o limpieza, por lo que aún pueden existir inconsistencias y el formato no es necesariamente el más adecuado para un análisis directo.

**Carga del dataset:**

Cargamos el dataset y mostramos sus dimensiones, estructura y tipo de datos:







# ¿Por qué es importante el dataset?

Este dataset es importante porque nos permite esclarecer qué factores pudieron influir en la supervivencia de viajeros del Titanic y obtener el conocimiento necesario para poder hacer predicciones con nuevas instancias.

Estos factores intuimos que pueden ser el estatus social, el sexo, la edad y también tener familiares cerca.

Asimismo, podemos ver si las pautas marcadas por la sociedad de "mujeres y niños primero" se cumplen cuando las personas se encuentran en situaciones de estres extremo.

De igual forma, y en el ámbito de la ciencia de datos, este dataset es importante porque es considerado un clásico y ha ayudado a muchos estudiantes a enfrentarse por primera vez a un problema de limpieza de datos, análisis estadísticos e incluso a técnicas de machine learning.

# ¿Qué problema pretende responder el dataset?

Este dataset pretende responder a cuáles son los diferentes factores que afectaron a la posibilidad de supervivencia de personas en el accidente del Titanic.

# Integración y selección de los datos de interés a analizar



# Integración.

La integración es un proceso que forma parte de la fase de limpieza de datos y se entiende como la fusión de datos para crear una estructura única que tenga la información necesaria para el posterior análisis de datos.

Existe la integración horizontal, que básicamente se compone de la adición de nuevos atributos a partir de otras fuentes mediante sus relaciones usando claves primarias y la integración vertical, que se basaría en añadir más instancias al juego de datos (siempre manteniendo la integridad de los atributos).

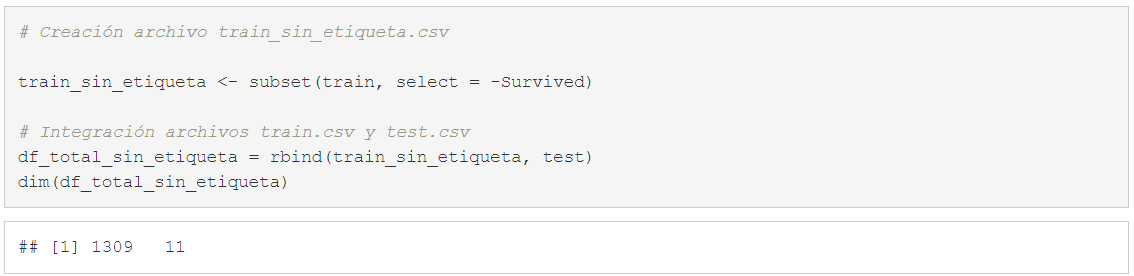
En nuestro caso, tenemos dos archivos **train.csv** y **test.csv**, dónde la diferencia entre ambos es que el test no tiene las etiquetas de la variable “Survived”.

Integración Vertical:

**Con la finalidad de observar las distribuciones de las variables que serán base del estudio en la predicción de “Survived”** integraremos verticalmente los dos archivos y así obtendremos un mayor número de datos para ver sus medidas de tendencia central y dispersión.

Para que la integración vertical sea satisfactoria, las variables y estructura de ambos archivos debe coincidir, por tanto, crearemos un dataframe **train\_sin\_etiqueta** que se integrará con las instancias de **test.csv** al cual llamaremos df\_total\_sin\_etiqueta.

Observamos que la integración es satisfactoria puesto que las instancias ahora son 1309 (891 + 418).

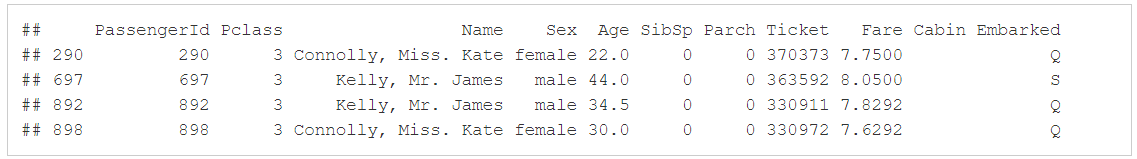


Integración Horizontal:

Los archivos en la plataforma Kaggle no exponen ni fuentes externas ni csv adicionales que definan nuevas variables que se puedan integrar horizontalmente a nuestro juego de datos.

Comprobamos si hay líneas duplicadas en el dataframe usando `duplicated`. No existen registros duplicados, pero sí detectamos dos pares de personas con el mismo nombre. Para asegurarnos que se trata de personas diferentes, buscamos los registros que tengan los nombres Connolly, Miss. Kate o Kelly, Mr. James.

Podría tratarse de la misma persona que ha comprado dos billetes, pero en estos registros vemos que las personas tienen edades diferentes y no hay motivo para pensar que se trata de duplicados.



# Selección.

La selección se puede entender como un **primer filtro de los datos**, no solamente a través de poner límites a los valores de algunas instancias o elegir algún valor cualitativo específico, sino también a través de la inspección de las correlaciones entre los atributos y la posterior eliminación del dataset de aquellos que sean redundantes.

Debido a que el problema planteado es interpretar qué factores influyen en la supervivencia, a priori, no sabríamos si debemos descartar alguna variable o no (eliminación de la variable del estudio) o si deberíamos filtrar los datos, ya sean numérica o categóricamente.

No obstante, en esta sección eliminaremos la variable “Name” porque es de mucha utilidad para nuestros análisis ya que el nombre no debería influir a priori en la supervivencia de los viajeros.

Mantendremos el PassengerId hasta que finalicemos la limpieza de datos con el archivo global para tener indexadas las instancias y después poder dividir otra vez los datos en train y test.

Por lo tanto, además de esta primera selección relizada, esta fase del proceso la dejaremos abierta en este punto y retomaremos una vez la exploración y análisis nos vaya indicando qué debemos seleccionar y/o filtrar. A continuación, se hace una lista de las selecciones realizadas en este apartado y a posteriori.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable Modificada** | **Tipo de selección** | **Apartado realizado** | **Motivo** |
| Name | Eliminación | 2.2 | Variable no útil |
| Ticket | Eliminación | 2.3 | Usada para crear nueva variable y ya no es útil |
| Fare | Eliminación | 2.3 | Usada para crear nueva variable y ya no es útil |
| Cabin | Eliminación | 3.1 | Existencia masiva de valores nulos |

**Se ha eliminado XXX en el apartado YYY por la razón ZZZ**

# Creación de nuevas variables.

Se ha detectado que hay números de billetes duplicados. Esto indica que hay dos tipos de tickets:

* Individuales
* Grupales

Se observa que la variable “Fare” muestra el mismo precio para los tickets grupales, por tanto, para saber realmente el precio del ticket por viajero **y también para poder usar correctamente la variable “Fare”**, deberíamos saber de cuántas personas es el ticket grupal y después dividir la variable “Fare” for dicha cantidad.

Crearemos una columna con el recuento de billetes con el mismo id para cada pasajero y otra con el precio unitario.

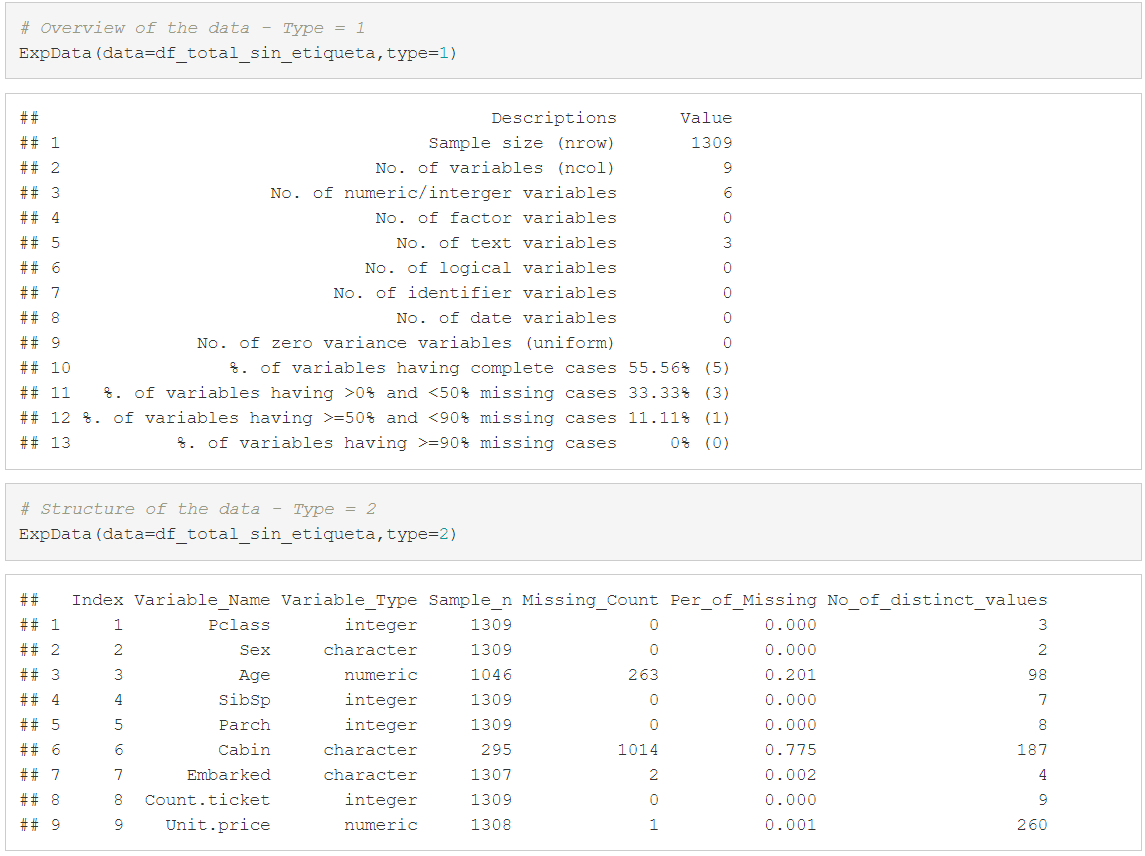
# Limpieza de datos

Hay que mencionar que se la limpieza de datos en este proyecto en particular debe afectar tanto al archivo train.csv como al test.csv, por tanto limpiaremos los datos en base al dataframe global creado anteriormente (df\_total\_sin\_etiqueta).



# ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?

Comprobaremos si existen valores nulos o inexistentes en el juego de datos.



Una vez que sabemos que tenemos valores nulos, cuántos tenemos y sabemos las variables afectadas, se decide la estrategia para imputar dichos valores.

**Variable Cabin:**

Observamos que la variable “Cabin” tiene 1014 valores nulos de 1309, por tanto se decide eliminar dicha variable por la imposibilidad de realizar una imputación generalizada.

**Variable Age:**

El número de registros de Age que son NA representan aproximadamente el 20% de los registros totales.

Este dataset contiene variables categóricas y numéricas y para imputar los valores nulos de la variable Age podemos usar el método `kNN` del paquete VIM. Aplicamos la función e imputamos los valores NA usando todos los demás campos del dataset y con un valor de k estándar igual a 3. El algoritmo busca los registros de los 3 pasajeros más parecidos (cercanos según la distancia Gower) al que contiene un valor nulo y usa los datos de edades de estos pasajeros para imputar el valor faltante.

Una vez ejecutado el algoritmo para imputar los valores, volvemos a comprobar si existen valores NA y podemos confirmar que todos los NA para la variable edad han sido imputados.

PONER SCREENSHOT IMPLEMENTACION VALORES CEROS

Gestión de los valores iguales a “cero”:

Primero se debería analizar si estos valores ceros son válidos o no en el conjunto y el rango de valores de la variable (ya sea cuantitativa o cualitativa), o si por el contrario el dataset generó los valores cero por defecto por alguna razón desconocida.

En nuestro caso en particular, la variable “Advertising” posee instancias con valor numérico entero cero. Al tratarse de una variable que indica el presupuesto de publicidad local de la empresa en cada ubicación, en miles de dólares, este valor cero puede indicar dos cosas:

* No hay presupuesto de publicidad
* Hay presupuesto, pero es menor de 500 dólares y el dato ha sido redondeado a cero ya que se muestra en enteros.

Se decide no modificar los ceros y asumir que dichas instancias indican que no hay prepuesto.

# Identificación y tratamiento de valores extremos.

En casos de identificar valores extremos, su tratamiento se podría dividir en 3 acciones:

* Eliminar los valores extremos
* Realizar imputaciones sobre los valores extremos
* Mantener los valores extremos por su valor explicativo en el conjunto de datos

Identificación:

Para cada variable numérica, dibujamos sus diagramas de caja, su función boxplot.stats() que nos indicará los outliers basados en el criterio IQR (Interquartile Range Criterion) y algunas gráficas adicionales que nos servirán para tomar decisiones sobre si existen o no realmente valores extremos y qué estrategia seguir.

Olga plots por pares identificando valores extremos. Pairplot??

# Análisis de datos



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

XXX

En este momento podemos empezar a mirar la relación entre las variables explicativas y la variable objetivo que puede ser ventas \* precio unitario o ser ventas \* precio unitario - advertisement

Cuando se pide “grupos de datos”, yo entiendo que debemos crear sub-datasets y realizar pruebas sobres los estadísticos para demostrar que se pueden aplicar luego otras técnicas… lo hablamos?

# Comprobación de la normalidad y homogeneidad de la varianza.

XXX

Histogramas, boxplots para todas las variables

Comprobación de la homocedasticidad

leveneTest(count ~ spray, data = InsectSprays)

Comprobación de la normalidad

ks.test(iris$Sepal.Length, pnorm, mean(iris$Sepal.Length), sd(iris$Sepal.Length))

# 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.Comprobación de la normalidad y homogeneidad de la varianza.

XXX

Regresión lineal para estimar ventas a partir de precio, precio competidor, gasto en publicidad, etc. Ir probando varias combinaciones hasta encontrar la que de mejor R2 – regresores cuantitativos

Regresión lineal con regresores cuantitativos y cualitativos.

CARLOS: PODEMOS HACER UN CONTRASTE DE HIPOTESIS PARA VER SI LAS MEDIAS DE LOS SUBGROUPS (US Yes/NO) es la misma o no.

Esto iría relacionado con el problema a resolver que describimos al principio de la práctica.

# Representación de los resultados a partir de tablas y gráficas



# Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?



# Resolución del problema.

XXX

# Conclusiones a partir de los datos obtenidos.

XXX

# ¿Los resultados permiten responder al problema?

XXX

# Código



**ADJUNTAR SCREENSHOTS???**

# Tabla de contribuciones al trabajo



Mediante la siguiente tabla, los estudiantes Olga Garcés Ciemerozum y Carlos Acosta Quintas certifican que ambos han colaborado y elaborado conjuntamente tanto en la Investigación previa del proyecto, como en la redacción de las respuestas y el desarrollo del código.

|  |  |
| --- | --- |
| Contribuciones | Firma |
| *Investigación previa* | *O. G. / C. A.* |
| *Redacción de las respuestas.* | *O. G. / C. A.* |
| *Desarrollo código* | *O. G. / C. A.* |

# **Referencias / Fuentes de Información**

* Calvo M, Subirats L, Pérez D (2019). Introducción a la limpieza y análisis de los datos. Editorial UOC.
* Squire, Megan (2015). Clean Data. Packt Publishing Ltd.
* Jiawei Han, Micheine Kamber, Jian Pei (2012). Data mining: concepts and techniques. Morgan Kaufmann.
* Jason W. Osborne (2010). Data Cleaning Basics: Best Practices in Dealing with Extreme Scores. Newborn and Infant Nursing Reviews; 10 (1): pp. 1527-3369.
* Peter Dalgaard (2008). Introductory statistics with R. Springer Science & Business Media.
* Wes McKinney (2012). Python for Data Analysis. O’Reilley Media, Inc.

kNN function - RDocumentation

<https://www.rdocumentation.org/packages/VIM/versions/6.1.0/topics/kNN>

Outliers detection in R - Stats and R

<https://statsandr.com/blog/outliers-detection-in-r/>