|  |  |
| --- | --- |
| 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 con el cual realizaremos la limpieza y el análisis de datos hace referencia a las ventas de 400 sillas de coche infantiles en función de un conjunto de parámetros de las regiones de venta del producto.

OLGA: El dataset con el cual realizaremos la limpieza y el análisis de datos hace referencia a las ventas de sillas de coche infantiles en 400 puntos de venta diferentes y en función de un conjunto de parámetros de las regiones de venta del producto.

El archivo que contiene los datos está en formato csv y su nombre es **carseats.csv**.

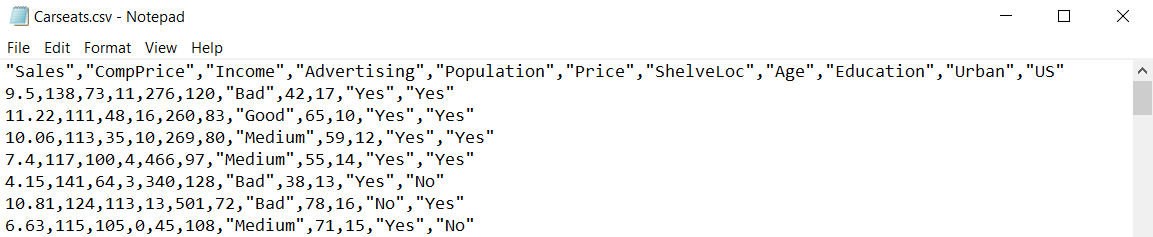
La ubicación en kaggle del dataset utilizado se muestra en el siguiente link:

<https://www.kaggle.com/huhao05133/carseats>

Los datos originales sin modificar corresponden a estas variables:

* **Sales** (Ventas unitarias de sillas, en miles de dólares, en cada ubicación)
* **CompPrice** (Precio del producto en dólares que cobra el competidor en cada ubicación)
* **Income** (Nivel de ingresos comunitarios, en miles de dólares)
* **Advertising** (Presupuesto de publicidad local de la empresa en cada ubicación, en miles de dólares)
* **Population** (Tamaño de la población en la región, en miles)
* **Price** (Precio del producto en dólares en cada ubicación)
* **ShelveLoc** (Una clasificación clasificando el dato entre Bad, Good y Medium que indica la calidad de la ubicación del producto en los puntos de venta)
* **Age** (Edad media de la población local)
* **Education** (Nivel educativo)
* **Urban** (Una clasificación clasificando el dato entre los niveles Yes y No, que indica si la tienda está en una ubicación rural o urbana)
* **US** (Una clasificación clasificando el dato entre Yes y No para indicar si la tienda se encuentra en USA o no).

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 “,”.



Cargamos el dataset y mostramos su estructura y tipo de datos:

400 instancias

11 variables

Tipo de dato integer (numérico entero):

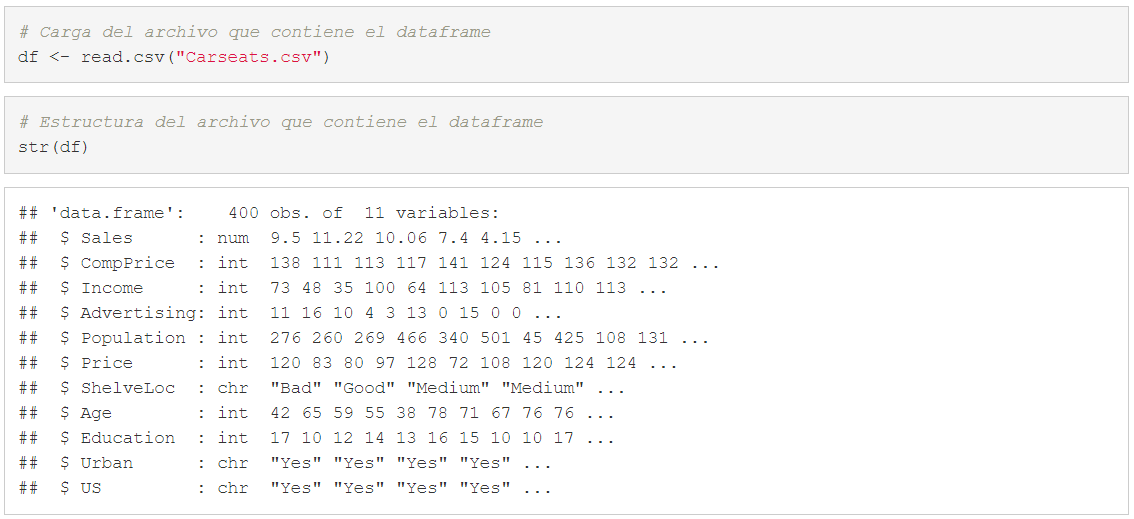
* **CompPrice**: Numérico entero (int)
* **Income**: Numérico entero (int)
* **Advertising**: Numérico entero (int)
* **Population**: Numérico entero (int)
* **Price**: Numérico entero (int)
* **Age**: Numérico entero (int)
* **Education**: Numérico entero (int)

Tipo de dato numérico (número real):

* **Sales**: Numérico real (num)

Tipo de dato character:

* **ShelveLoc**: Categórico (factor) con 3 niveles "Bad","Good" y "Medium"
* **Urban**: Categórico (factor) con 2 niveles "No" y "Yes"
* **US**: Categórico (factor) con 2 niveles "No" y "Yes"



OLGA: Observamos que la variable Education sido interpretadas como numéricas pero el tipo de dato corresponde al tipo cualitativo ordinal. Esta variable representa los niveles de educación y un mayor valor indica un nivel de educación superior. No podemos considerar esta variable como numérica dado que no tendría sentido realizar operaciones matemáticas con la misma.

Según se observa a primera vista mediante la lectura del dataset en el bloc de notas y mediante el output de la estructura del archivo, el dataset estaría preprocesado y libre de valores nulos o errores de inconsistencia en los strings de las variables cualitativas. Esto se comprobará en los siguientes apartados.

***DESCRIPCION SOLUCION PRA1 COMO EJEMPLO:***

*Tal como expresa el título, el dataset está basado sobre los datos más importantes a tomar en cuenta para la evaluación de un cierto canal de Twitch.*

*En este dataset, se presentan dichos datos para cada uno de los top streamers dentro de la plataforma en un periodo de 365 días.*

*Las unidades o magnitudes de las características extraídas son en horas y cantidad según el caso.*

*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. Por ejemplo, hay campos donde los valores enteros se presentan como un string del tipo “XXX horas”, en vez de ser simplemente el valor XXX.*

*En este caso, se extrajo la información para el top 50 de los canales de la plataforma.*

*La descripción de las características extraídas son descritas en las siguientes preguntas. El formato del dataset es un fichero CSV que facilita su visualización y tratamiento.*

# Por qué es importante el dataset?

XXX

La empresa que se dedica a la fabricación de sillitas de coche para niños dispone de un conjunto de datos sobre los distintos establecimientos donde es posible adquirir su producto.

# ¿Qué problema pretende responder el dataset?.

XXX

Pronosticar las ventas en función de las variables descriptivas de los puntos de venta.

El dataset está compuesto por datos propios del punto de venta (posicionamiento en la tienda, publicidad, unidades vendidas, precio unitario, etc) así como de datos que representan las características de la población objetivo (edad, nivel de educación, tipo de área: rural o urbana). Con estos datos podemos realizar varios tipos de análisis, como por ejemplo:

Estudiar si las ventas de las sillitas de coche guarda una relación con el precio, el gasto en publicidad u otras características del punto de venta o bien está más influenciada por características de la población.

El dataset puede responder a varias preguntas relacionadas con las ventas:

¿Son mayores las ventas en zonas urbanas o en zonas rurales?

¿Las zonas con población envejecida tienen menos ventas?

¿El nivel de educación de la población tiene algunarrelación con las ventas?

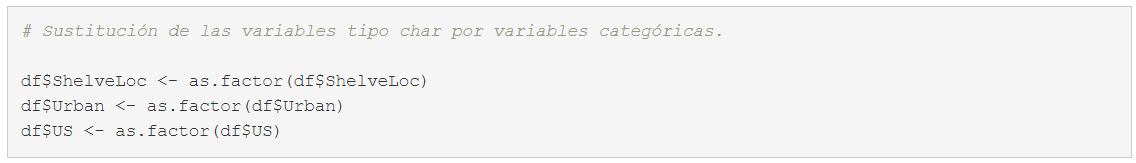
¿Las sillitas se venden igual en ubicaciones de ingresos altos y ingresos bajos?

¿Los precios de productos análogos de la competencia influyen sobre las ventas?

# Integración y selección de los datos de interés a analizar

En primer lugar, modificaremos el tipo de datos de algunas variables para facilitar su análisis posteriormente.

Cambiaremos las variables ShelveLoc, Urban y US, de tipo carácter a tipo factor.



Carlos:

**CAMBIAMOS TODOS LOS PRECIOS A DOLARES? ACTUALMENTE LAS UNIDADES SON MILES DE DOLARES Y DOLARES**

**OLGA: cambiamos.. Lo hago yo**

OLGA: VENTAS - gasto en publicidad

Dicotómica competencia gasta más que empresa o no (0,1)



# 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).

**~~CARLOS: CREAMOS UNA VARIABLE YES/NO PARA DECIR SI TIENEN PRESUPUESTO (VALOR ¡=0) O NO (VALOR = 0)?~~**

OLGA: Presupuesto en publicidad? Podemos probar pero me parece que no.. Sin embargo, he pensado que tal vez un menor presupuesto se puede compensar con una mejor colocación en la tienda.. Habría que verlo

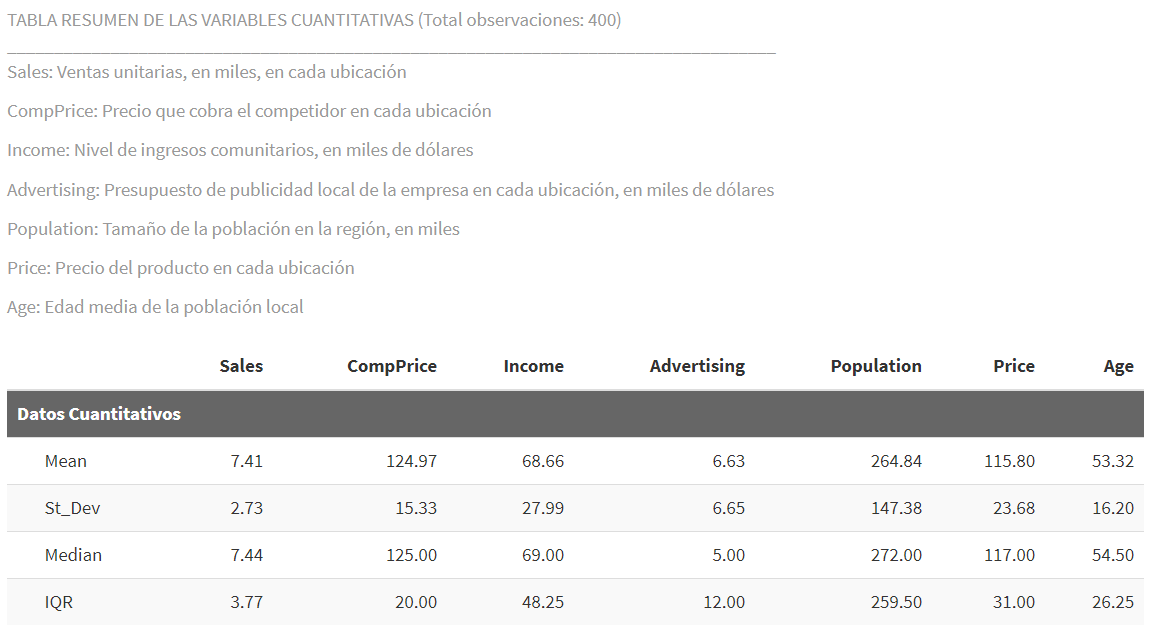
Carlos: OKIS

# Selección.

Normalmente se suele aprovechar en esta primera exploración de los datos (o screening) para **analizar las suposiciones en los datos** requeridas por las pruebas estadísticas que se aplicarán posteriormente. Por tanto, realizaremos una tabla de los datos cuantitativos donde aparezca la media, la mediana, la desviación estándar y el rango intercuartílico y se comentarán los resultados.

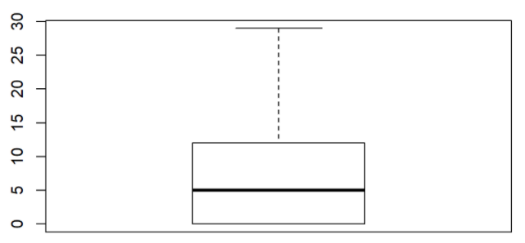
Nota:

Hemos observado que la variable Education tiene valores mínimo y máximo 10 y 18 respectivamente, lo que desvirtuaría los valores estadísticos que se pudieran mostrar en la tabla. Decidimos no incluir la variable Education en la tabla. OLGA: ya hemos dicho que es una variable culitativa ordinal y pertenece con las variables cualitativas. -> OKIS, borra lo que consideres oportuno 😉



Comparando la Desviación estandar con la media en todas las variables, podemos hacernos una idea de la “forma” de la distribución (suponiéndola normal) en cuando a su amplitud. Observamos que todas las variables excepto Advertising tienen desviaciones estandard en relación a la media aceptables (St\_Dev/Mean aprox 0.5 o menos), pero para Advertising este ratio es mayor de uno. Si la desviación estándar es más grande que la media, esto probablemente indica un sesgo, es decir, la presencia de valores extremos u otra peculiaridad, así que habría que estudiar en detalle esta situación.

Observamos también que el IQR de Advertising también es el doble que la media, indicando también que la distribución no es simétrica. Realizaremos un boxplot y comprobaremos que efectivamente la distribución parece que no es normal.



En todas las variables excepto Advertising y Population, media y mediana son aproximadamente las mismas relativamente, indicando que se podría cumplir una de las hipótesis de tener una distribución normal para dichas variables.

**Nota en Price - CompPrice:**

Podemos observar que tanto la media como la mediana en el precio de los competidores son más altos que los precios de la empresa en estudio, esta situación, a igualdad de producto (mismo modelo de silla, misma calidad, etc..) sería ventajosa para la empresa. Habría que estudiar si en el precio de los competidores se añade algún servicio como el de entrega a domicilio o extensión de garantía, para estar realmene seguros que estamos comparando exactamente los mismos conceptos, puesto que en caso contrario, el cliente podría elegir la competencia si nuestro precio global “combinado” fuera mayor.

**Nota en Age**:

Podemos observar que la edad media local es alta si nuestro cliente objetivo son padres/madres con niños pequeños. Normalmente la edad con la que se tienen niños es menor que 53 años y por tanto, deberíamos analizar la ampliación del negocio a zonas con menor edad, es decir, con mayor probabilidad de tener clientes potenciales.

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.

CARLOS:

FILTRAR LOS DATOS EN FUNCION DE ALGUN VALOR DE ALGUNA VARIABLE:

OLGA: he puesto una función que hace 4 plots de una columna: histograma, qqplot, boxplot normal y según los valores que tome una variable que indicas. En R he hecho boxplots para varias.

OLGA: en esta fase es cuando se crean variables nuevas. Hemos visto las posibilidades de precio de empresa mayor que el de la competencia, variable beneficios que es ventas unitarias por precio unidad – publicidad.. O bien simplemente ventas unitarias por precio de unidad. Pongo todas estas variables en el dataset por si las queremos usar.

OTRA VARIABLE que se me acaba de ocurrir es ingresos por persona! Como no lo hemos visto antes? - creada

He generado plots en bucle para luego poder elegir los que más nos gusten

# Limpieza de datos

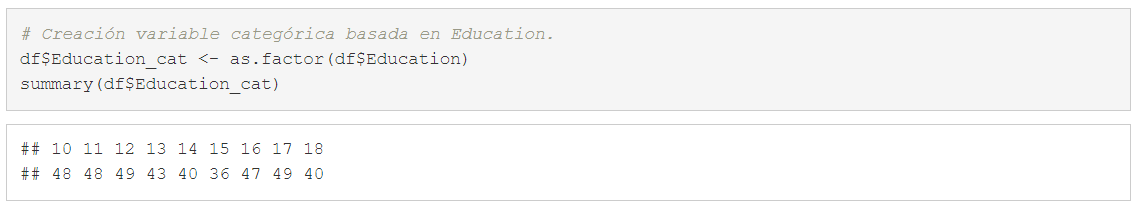
Previo a la exploración de valores cero, nulos o extremos, se realizarán otras tareas de limpieza de datos.

Observamos que la variable **Education** tiene asignados unos valores númericos tipo entero. Se asumirá que cada número indica una tipo de estudio, por tanto podríamos considerar esta variable como una variable categórica. Crearemos una nueva variable **Education\_cat** con tipo de dato categórico.

OLGA. Tenemos que crear la variable? La variable ya la tenemos, solamente la tenemos que transformar en factor y sugiero que sea un factor ordenado, con niveles que tienen significado numérico en el sentido que son ordenables.

He puesto esto en la línea 98 o más o menos por ahí

df$Education <- factor(df$Education, levels = c(10, 11, 12, 13, 14, 15, 16, 17, 18))-> OK, puedes borrar lo que consideres 😊





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

Comprobaremos si existen valores nulos en el juego de datos. Esto lo podremos averiguar aplicando la suma de los valores nulos si existen mediante la función is.na().

Comprobamos que el archivo Carseats.csv no posee valores nulos.



OLGA

Gestión de los valores nulos:

**~~Carlos:~~**

**~~PONGO DISTANCIA DE GOWER, SABES ALGUNA OTRA TECNICA?~~**

**~~Olga: regresión lineal.. Etc~~**

En el supuesto de que existieran valores que realmente tienen un significado que no encaja con el contexto general de los datos, tendríamos a nuestra disposición una serie de técnicas para solventarlo.

En el caso de detectar algún valor perdido en las variables cuantitativas se podría realizar una imputación de valores en estas variables mediante el algoritmo kNN, eligiendo un valor k (vecinos más cercanos) y usando una métrica para la distancia, por ejemplo la distancia de Gower (ya que según la documentación oficial se puede aplicar tanto a variables numéricas como categóricas).

OLGA: Para las variables cuantitativas podríamos optar también por una imputación de valores usando kNN o a través de una regresión. Para ello tendríamos que comprobar si existe relación lineal entre las distintas variables o si sería más adecuado usar otro tipo de regresor (gradient boosting, vector support machines, etc).

La estrategia para imputar valores con kNN sería la siguiente:

1. Buscar las variables con instancias de valores nulos.
2. Comprobar y mostrar las filas del dataframe donde existen valores nulos.
3. Crear un dataframe auxiliar con las columnas seleccionadas para imputar los valores nulos.
4. Aplicar la imputación de valores nulos con los k vecinos más cercanos usando la distancia de Gower mediante función kNN() (en R, bajo la libreria VIM, por ejemplo).
5. Asignar como nuevos registros de los campos con valores nulos los registros obtenidos de las imputaciones del dataframe auxiliar.

Nota:

Se podrían hacer iteraciones con el valor de k para optimizar el resultado mediante el “método del codo”.

OLGA:

Valores extremos:

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/>

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.

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

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:

|  |  |  |
| --- | --- | --- |
| Variable | Descripción | Tipo de dato |
| PassengerId | Identificador único del pasajero | Entero |
| Survived | Etiqueta de clase, Indica si el pasajero ha sobrevivido la catástrofe (0- no sobrevive, 1 - sobrevive | Entero |
| Pclass | Clase en la que viajaba el pasajero, valores 1, 2, 3 | Categórica, factor |
| Name | Nombre del pasajero | String |
| Sex | Sexo del pasajero (M, F) | String |
| Age | Edad del pasajero | Entero |
| SibSp | Indica si el pasajero tenía hermanos o pareja a bordo | Entero |
| Parch | Indica si el pasajero tenía padres o hijos a bordo | Enterro |
| Ticket | Número de billete | String |
| Fare | Precio del billete sin especificar si es un billete individual o grupal | Número real |
| Cabin | Número de camarote | String |
| Embarked | Indica el lugar de embarque del pasajero, C = Cherbourg, Q = Queenstown, S = Southampton | String |

¿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.

2. Integración y selección de los datos de interés a analizar.

En esta sección integraremos los datos de los dataset de train y test en uno solo para poder así procesarlos conjuntamente. Una vez integrados los datos, buscaremos registros duplicados, crearemos las variables relevantes para nuestro estudio y descartaremos algunas de las variables a las que no les podemos dar uso.

Integración de los Datos

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).

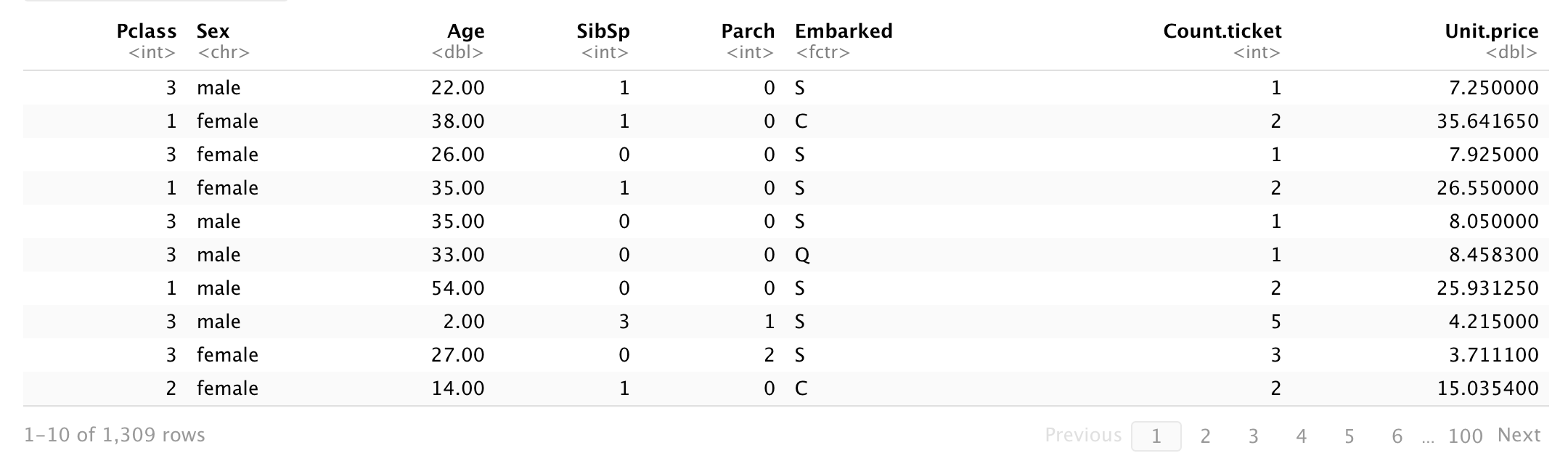
En nuestro caso, tenemos dos archivos train.csv y test.csv, dónde la diferencia entre ambos es que el test no tiene las etiquetas de la variable “Survived”.

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).



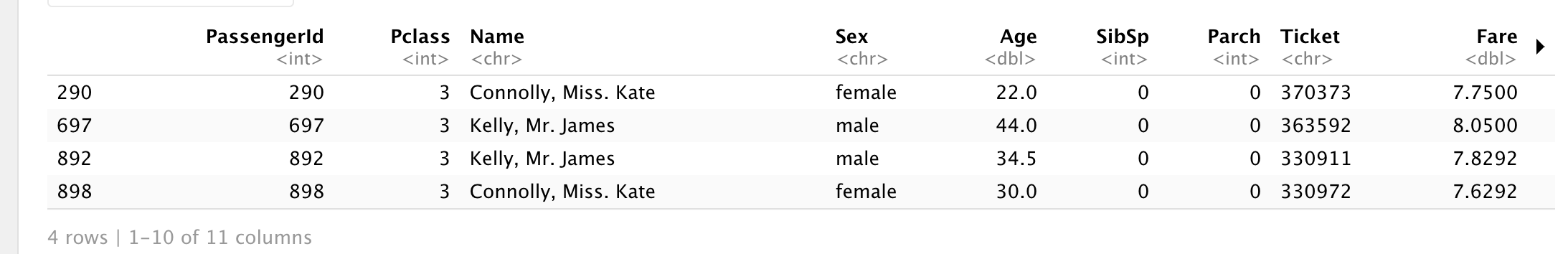
Integración Horizontal:

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.

Comprobación de líneas duplicadas:

En la búsqueda de registros duplicados tomando en consideración las variables Name y Sex encontramos que existen dos pares de personas con el mismo nombre en el dataset. Nos aseguramos que realmente se trata de personas diferentes.

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 de los Datos

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 no es de mucha utilidad para nuestros análisis ya que el nombre no debería influir a priori en la supervivencia de los viajeros y también la variable “PassengerId” puesto que simplemente es un identificador.

Por lo tanto, además de esta primera selección realizada, 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 al ser independiente al estudio |
| PassengerId | Eliminación | 2.2 | Variable no útil al ser un simple identificador |
| Ticket | Eliminación | 2.3 | Usada para crear nueva variable y ya no es útil |
| Fase | Eliminación | 2.3 | Usada para crear nueva variable y ya no es útil |
| Cabin | Eliminación | 3.1 | Existencia masiva de valores nulos |

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.

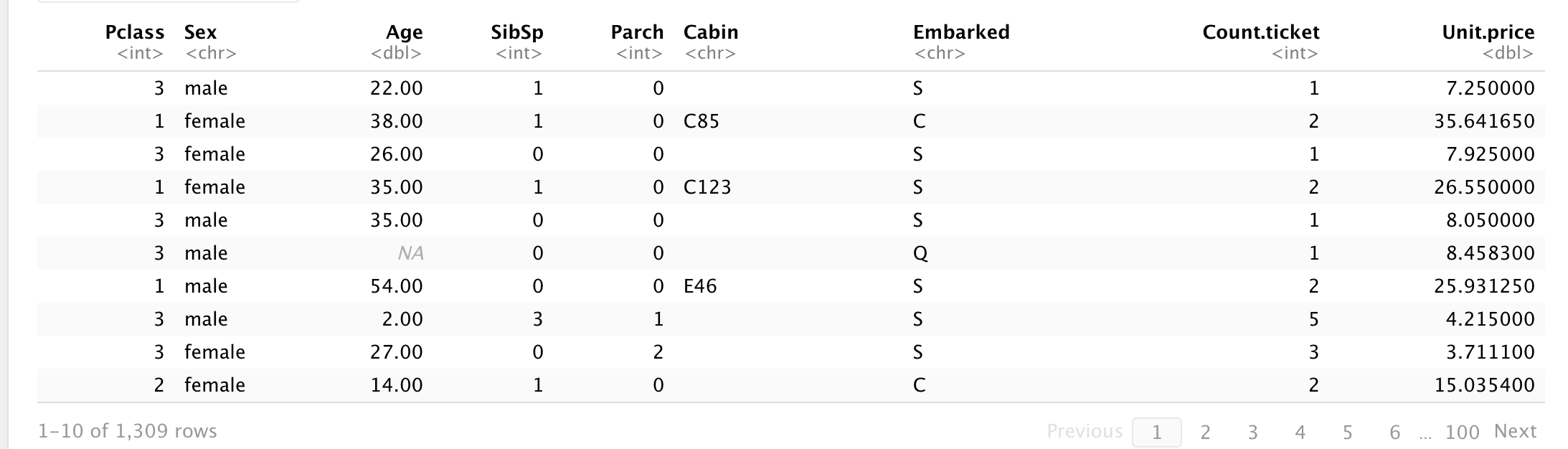
Para reflejar el número de personas que viajan con un mismo billete, creamos una variable que contiene el recuento de ocurrencia del billete del pasajero en todo el dataset.

Una vez usada la variable Ticket, ya no nos resulta útil, por lo tanto la excluimos del dataset.

Las variables que usaremos para los análisis posteriores son:

“ Pclass", "Sex", "Age", "SibSp", "Parch", "Cabin", "Embarked", "Count.ticket", "Unit.price"

El resultado de las transformaciones que acabamos de describir se muestra a continuación:



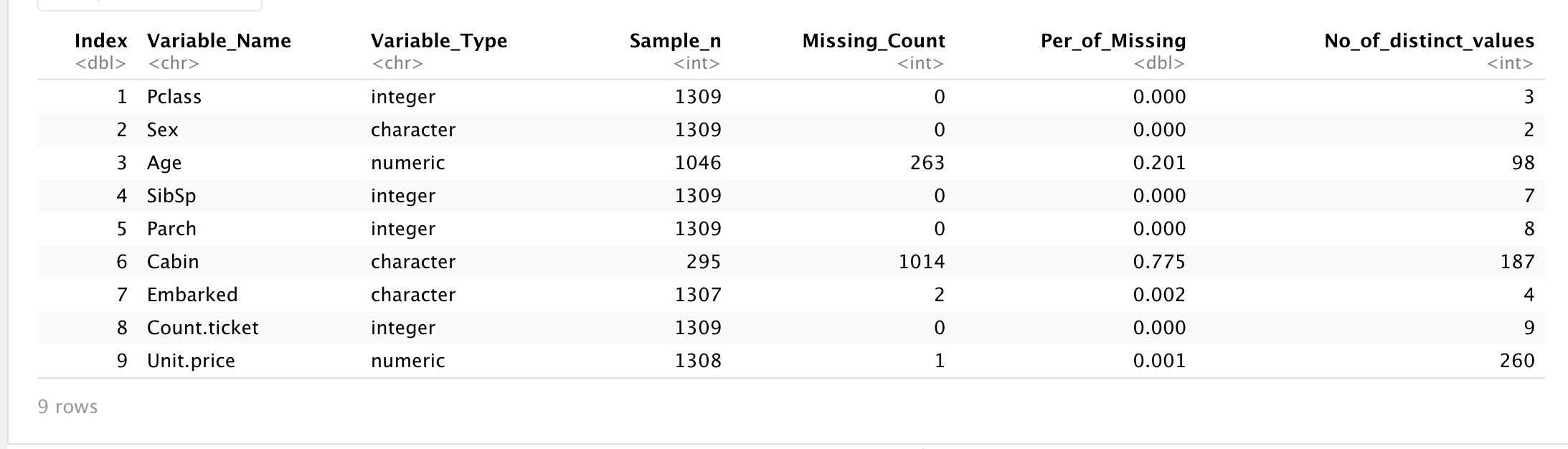
3. Limpieza de los datos.

Recordemos que en uno de los apartados anteriores hemos integrado los datos de entrenamiento y de test en un único dataset. Las tareas de limpieza de datos las realizaremos sobre el dataset integrado.

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

Elementos vacíos en el dataset

Hacemos uso de la función ExpData para comprobar 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`. Aplicamos la función e imputamos los valores NA usando todos los demás campos del dataset y con un valor de k 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.

Variable Embarked:

Se observa que la mayoría de las instancias pertenecen a la categoría S, por tanto, las instancias con valores nulos en esta variable, las imputaremos a S.

Variable Unit.price:

Actuaremos de igual forma que con la variable Age e imputaremos a través del uso del kNN

Gestión de los valores iguales a “cero” en el dataset

Comprobamos las variables que toman valores igual a cero sin que tenga sentido que tomen este tipo de valor.

La variable que representa la "clase" toma valores iguales a cero y consideramos que es correcto, lo mismo ocurre con las variables SibSp, Parch, donde consideramos normal que existan valores iguales a cero, significa que los pasajeros viajaban solos.

En cambio, los valores iguales a cero para la variable Unit.price son algo más extraños. Entre los pasajeros que tienen un Unit.price igual a cero hay personas que viajaban en primera, segunda y tecera clase.

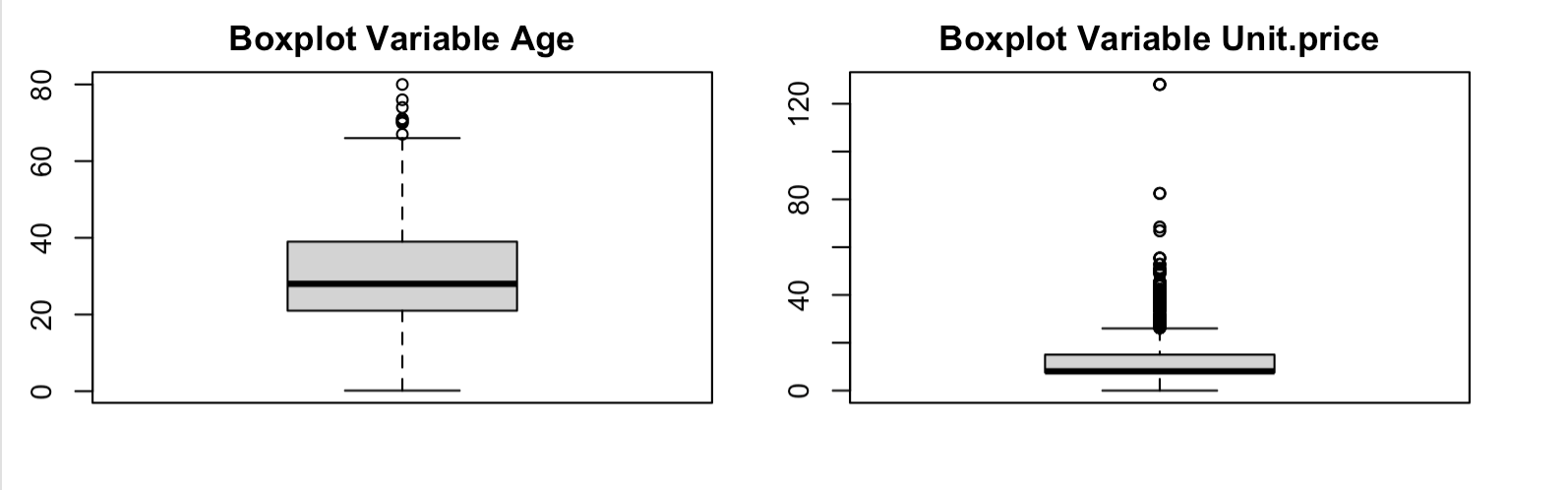
La idea que un ticket sea gratuito no sería posible, por tanto, volveremos a aplicar el método kNN para imputar estos valores.

Primero cambiaremos el valor de cero a NA y después actuaremos como en el apartado anterior.

3.2. Identificación y tratamiento de valores extremos.

No hemos encontrado valores que estén fuera de un rango razonable. Las comprobaciones las hemos hecho anteriormente con `sapply(df, summary)`.

Volvemos a visualizar boxplots para las variables numéricas que tenemos: Age y Fare.



El boxplot muestra que hay outliers en estas dos variables. La mayoría de los pasajeros eran jóvenes, aunque también encontramos pasajeros de más de 65 años. En cuanto a la variable Unit.price, se comprueba que a medida que el precio sube, la clase va bajando de 3 a 2 y de 2 a 1, con lo cual no hay razón porqué pensar que los precios no son reales. En la tabla siguiente podemos visualizar algunos de los pasajeros que pagaron un precio de billete alto. Notamos que todos son de primera clase.

Aunque se acepte que los precios son reales, hay uno que es extremadamente alto y, aunque cierto, podría desvirtuar posibles futuras predicciones, por tanto se estima que se podría cambiar por la media de Unit.price agrupado por la primera clase, para tener un valor imputado más real.

4. Análisis de los datos.

Una vez limpiado el archivo que contenía las líneas de los conjuntos train y test, deberemos separar otra vez los conjuntos ya que únicamente tenemos datos de la etiqueta para el conjunto de entrenamiento. Una vez separados los datos, volvemos a añadir al conjunto de entrenamiento sus etiquetas de clase.

Screening

Antes de crear las visualizaciones determinamos que las variables numéricas son: Age, SibSp, Parch y Unit.price que corresponden a la edad de los pasajeros y el precio unitario del billete.

Survived, Sex, Embarked son variables categóricas y Pclass es una variable categórica ordinal (existen rangos en los valores de las variables).

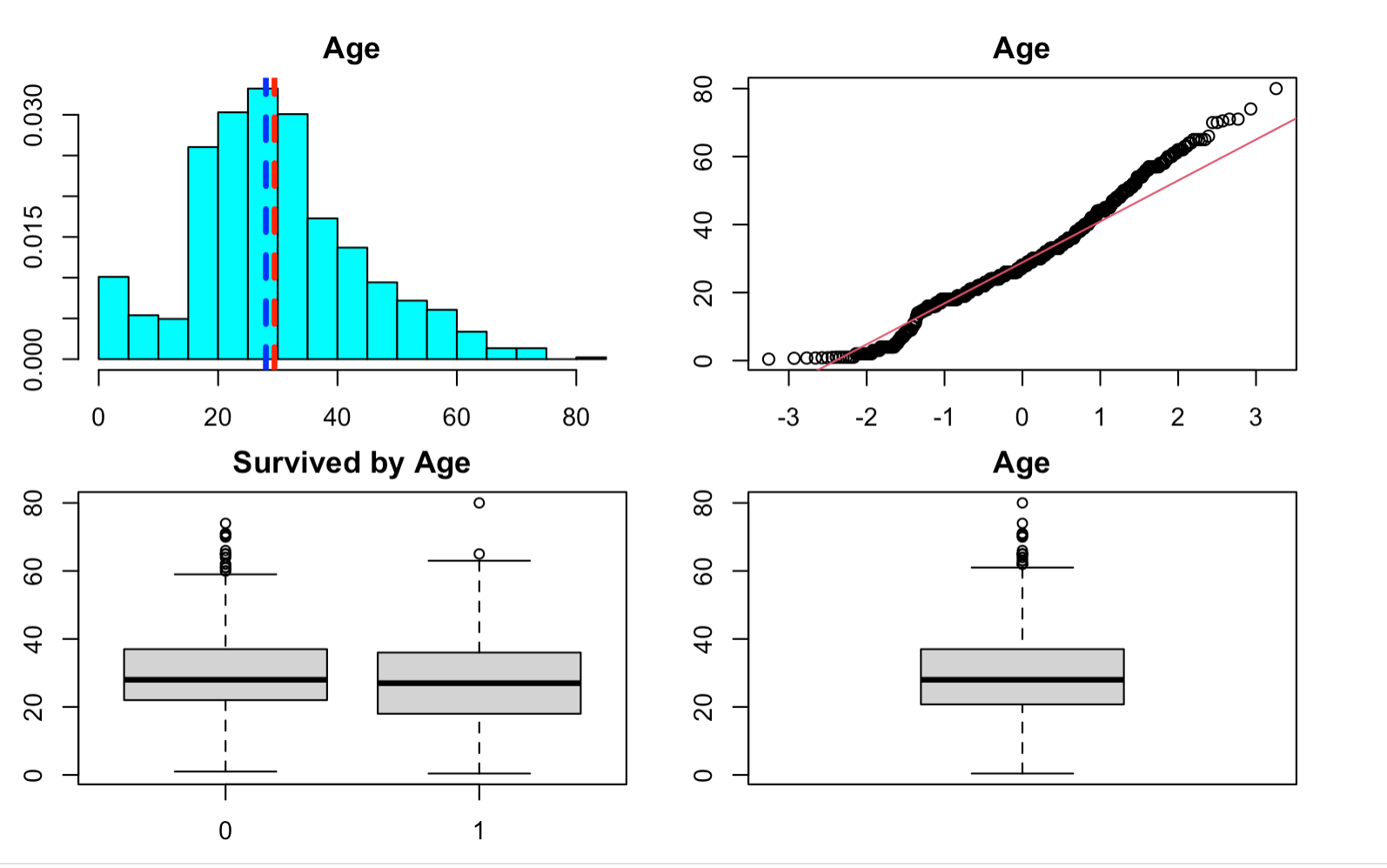
Realizamos las transformaciones oportunas para guardar las variables con sus tipos correspondientes.

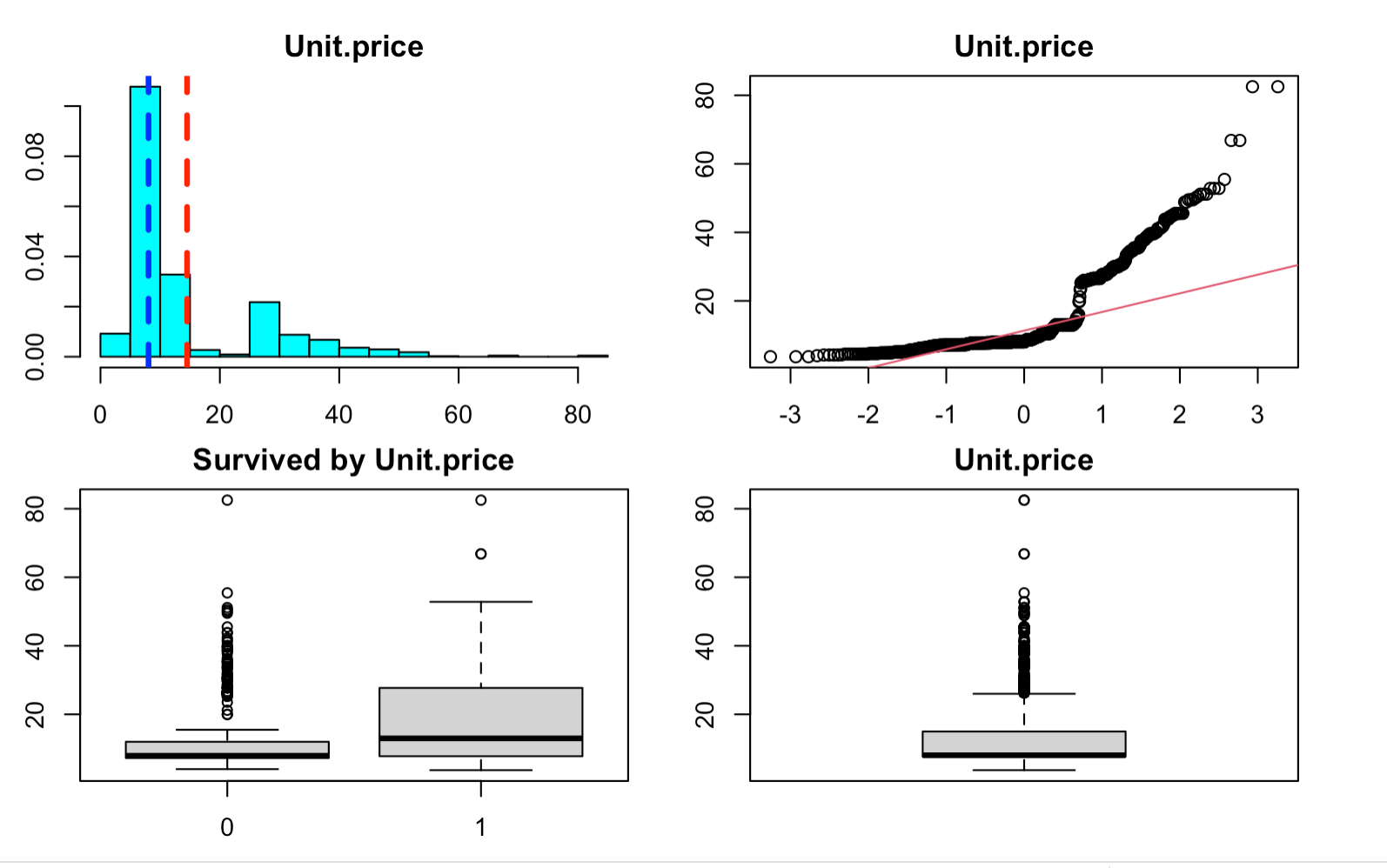
Visualización de las variables numéricas

Los pasajeros que tienen edades entre los cuantiles 25 y 75 tienen entre 20 y 40 años. La edad mediana para los pasajeros que sobrevivieron y los que no es muy similar.

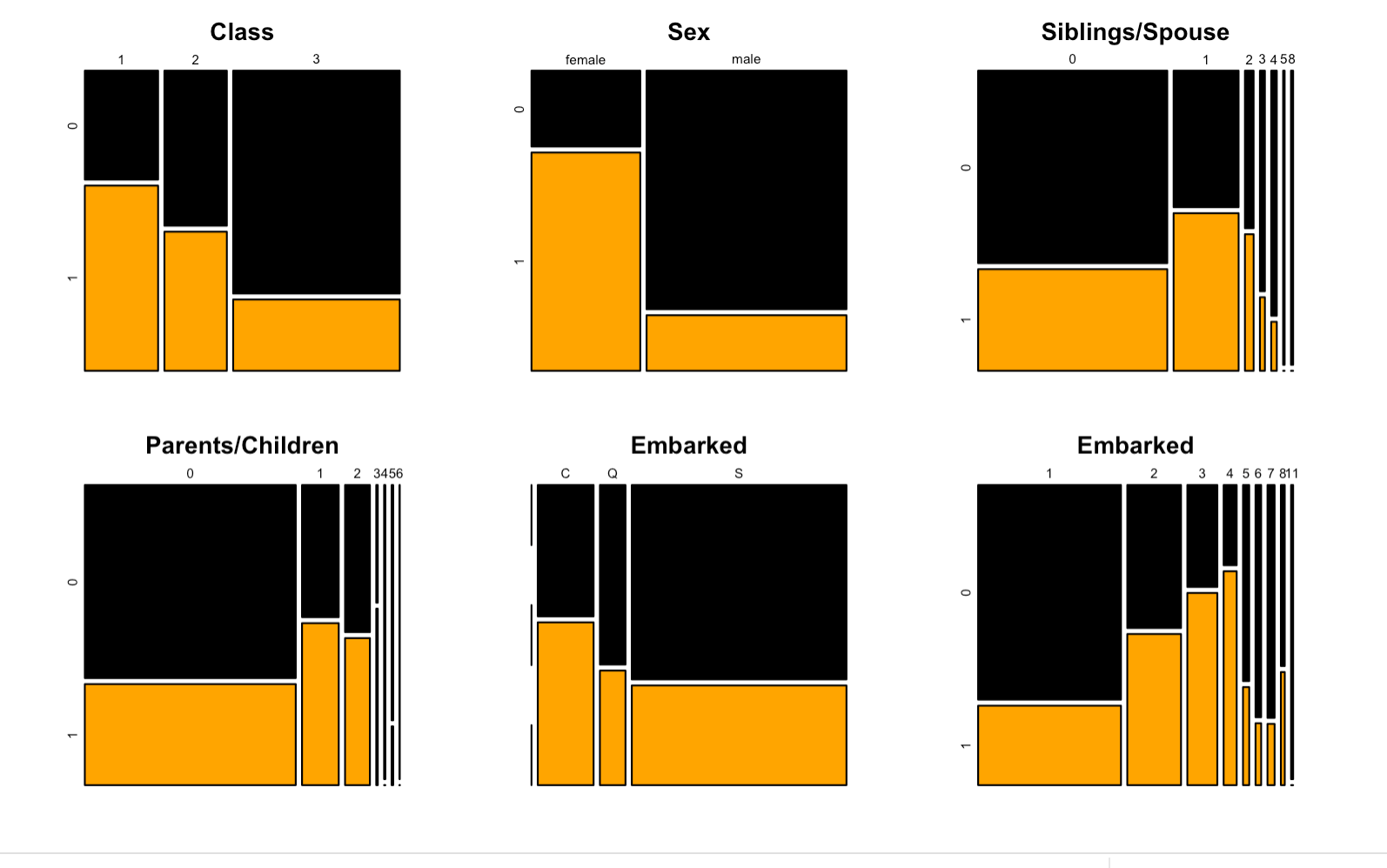
Si vemos la variable Unit.price (precio unitario del billete), la mayoría de los pasajeros pagaron muy poco por el billete. Aquí en el boxplot encontramos outliers pero consideramos que no deberíamos descartar estas entradas ya que hay pocos pasajeros que pagaron un importe alto por el billete y esta información puede ser muy interesante para predecir la posibilidad de supervivencia: ¿Viajar en una clase privilegiada aumenta la posibilidad de sobrevivir?

En el boxplot además podemos ver que las personas que sobreviven tienen una mediana más alta en el precio del billete.





A continuación visualizamos los datos para las variables Class, Sex, Siblings/Spouse, Parents/Children



Las personas que viajaban en tercera clase tienen la menor proporción de supervivencia comparando con las personas que viajaban en primera clase. Las mujeres que viajaban en el titanic sobrevivieron en su mayoría. Los hombres sobrevivieron en mucho menor proporción.

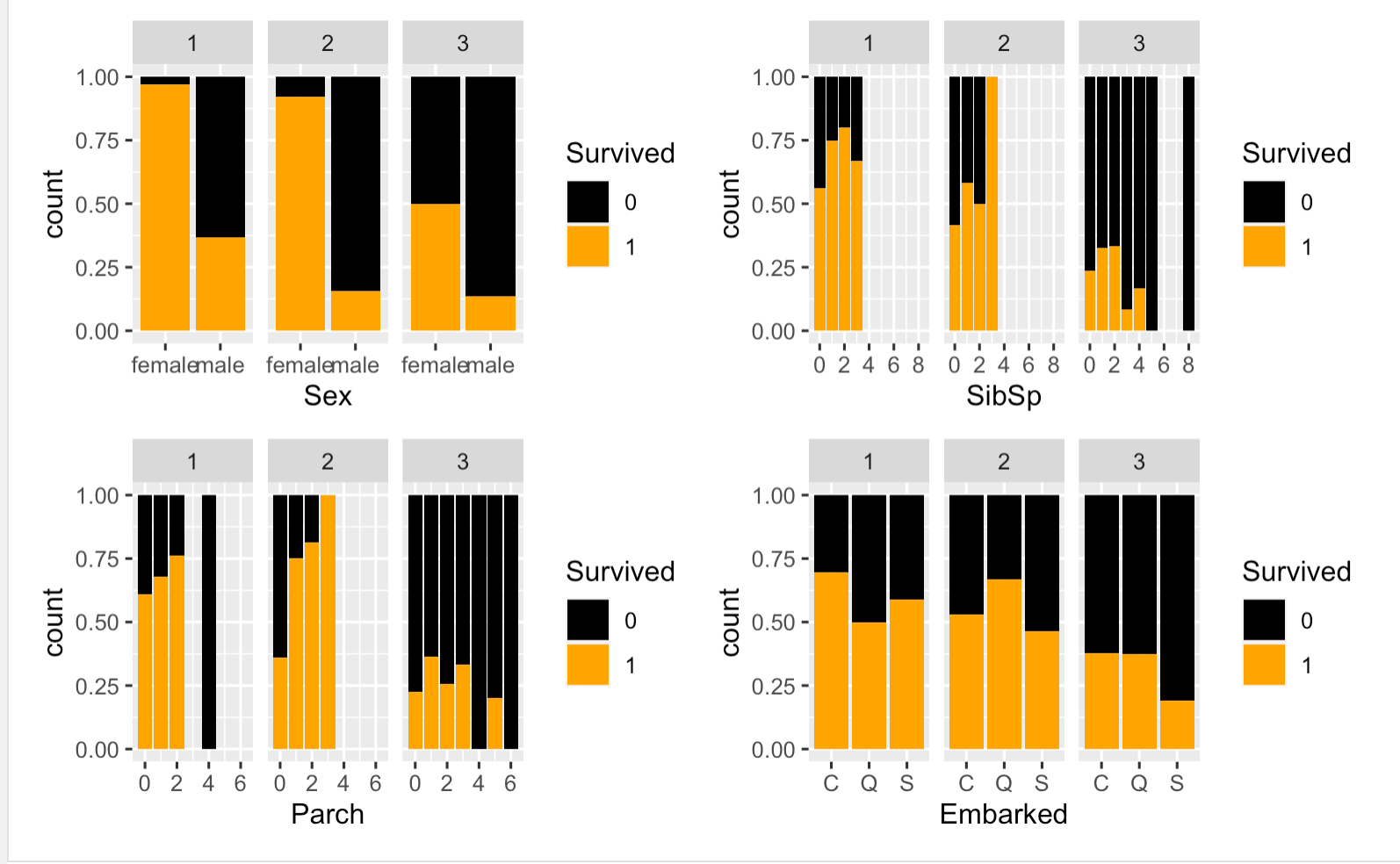
La mayoría de personas viajaba sin familiares (hermanos, pareja, padres o hijos) y parece ser que el porcentaje de supervivencia es algo más alto en personas que tenían familiares a bordo.

La mayoría de personas embarcaron en el punto S, pero el mayor porcentaje de supervivencia lo tienen las personas que embarcaron en C.

Los pasajeros que viajaban varias personas con el mismo billete también tienen una mayor supervivencia ( hasta 4 pasajeros ). 1 o más de 4 pasajeros con el mismo billete tienen la misma proporción de supervivencia.

Análisis bivariable

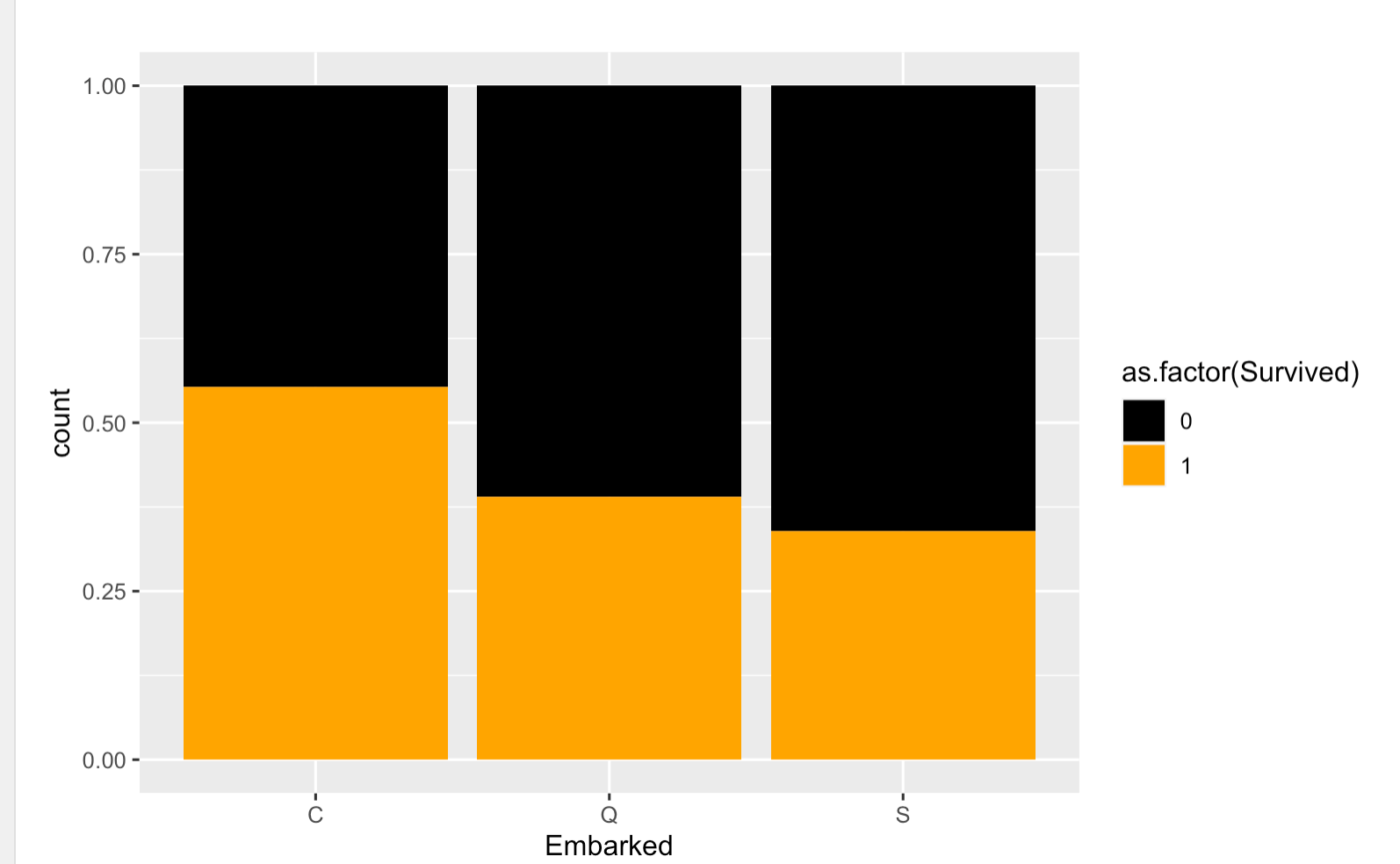
Vamos a visualizar algunos plots que nos permiten ver la supervivencia de pasajeros combinando dos variables.



Las mujeres que viajaban en primera y segunda clase sobrevivieron casi todas. Los hombres sobrevivieron en mucho menor medida, incluso los hombres que viajaron en primera clase.

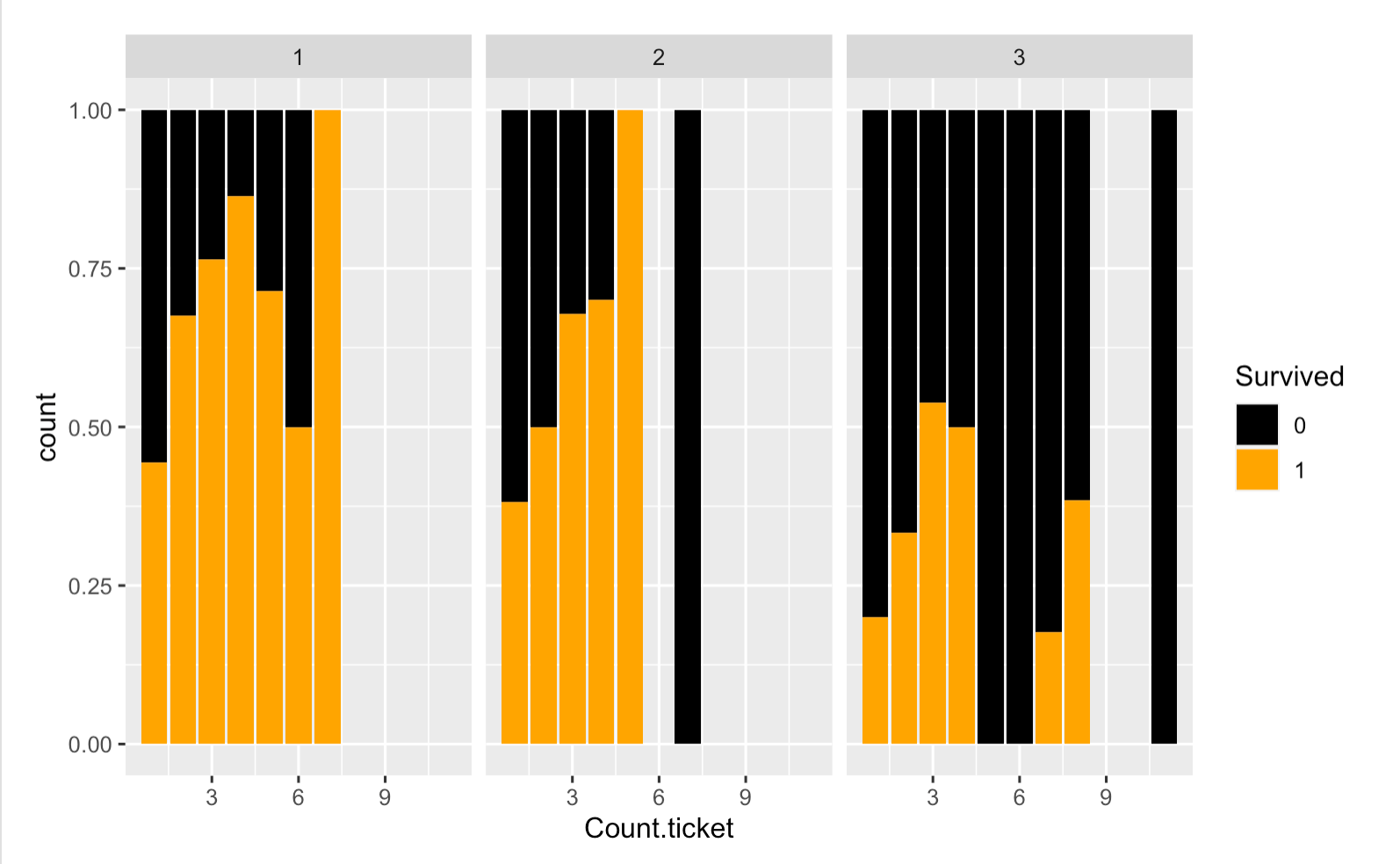
Las personas que viajaban con hermanos o pareja en primera clase sobrevivieron en mayor medida que las personas que viajaron solas. En primera y segunda clase no hay personas que viajasen con más de 3 hermanos. En tercera clase hay personas que viajaron con hasta 8 hermanos en este caso un menor número de hermanos (excepto cero) parece indicar una mayor supervivencia.

También hemos visualizado la supervivencia en función del lugar donde los pasajeros embarcaron en el Titanic. Los pasajeros de tercera clase que embarcaron en el punto S han tenido menos proporción de supervivencia que los demás. Los pasajeros de segunda clase que embarcaron en Q sobrevivieron en mayor proporción que los pasajeros de segunda clase que embarcaron en otros puntos. Para los pasajeros de primera clase el punto de embarque con mayor porcentaje de superviviencia es C.



Visualizamos la supervivencia usando las variables de la clase en la que viaja el pasajero y el número de personas que viajan con el mismo billete.

En el Titanic viajaron familias enteras de hasta 7 personas en primera clase y incluso de 11 personas en tercera clase. La supervivencia de personas que viajaban sobre el mismo billete en tercera clase es comparable con el mismo dato en segunda clase.



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

OLGA: Si seguimos con lo que hemos hablado, de centrarnos en si los pasajeros viajan solos o no, debemos separar estos 3 pares de datasets.\*\*las variables sibsp, parch están como factor, voy a comentar la parte donde se les transforma en factor..\*\*

```{r echo=TRUE, message=FALSE, warning=FALSE}

# Dataframe por pasajeros que viajaban solos o acompañados

train.single <- train[train$Count.ticket=="1",]

train.many <- train[train$Count.ticket>1,]

```

```{r echo=TRUE, message=FALSE, warning=FALSE}

# Dataframe por pasajeros que viajaban solos o acompañados

train.spouse <- train[train$SibSp>0,]

train.not.spouse <- train[train$sibsp==0,]

```

```{r echo=TRUE, message=FALSE, warning=FALSE}

# Dataframe por pasajeros que viajaban solos o acompañados

train.children <- train[train$Parch>10,]

train.no.children <- train[train$Parch==0,]

```

```{r echo=TRUE, message=FALSE, warning=FALSE}

# Dataframes por clase de pasajero

train.first <- train[train$Pclass==1,]

train.second <- train[train$Pclass==2,]

train.first.second <- train[train$Pclass==1 | train$Pclass==2,]

train.first.second$Pclass <- factor(train.first.second$Pclass, levels = c(1,2))

```

```{r echo=TRUE, message=FALSE, warning=FALSE}

# Dataframe por sexo del pasajero

train.male <- train[train$Sex=="male", ]

train.female <- train[train$Sex=="female",]

```

```{r echo=TRUE, message=FALSE, warning=FALSE}

# Dataframe por edad del pasajero

train.young <- train[train$Age<18,]

train.older <- train[train$Age>=18,]

```

```{r echo=TRUE, message=FALSE, warning=FALSE}

train.survived <- train[train$Survived==1,]

train.not.survived <- train[train$Survived==0,]

```

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

Visualizamos los datos de la edad para personas que sobrevivieron en el hundimiento frente a personas que no sobrevivieron.

La media y la mediana de los pasajeros que sobrevivieron están muy cerca. Las personas que no sobrevivieron tienen una edad mediana más baja que la edad media. Los valores de estos dos indicadores de tendencia central son más altos para personas que no sobrevivieron: las personas que muerieron tenían una media de edad algo más alta que los que sobrevivieron.

Según el qqplot las dos distribuciones son cercanas a la normal y aplicando el teorema del límite central podemos asumir la distribución normal de la media de estas dos muestras ya que sus tamaños son mayores que 30 observaciones.

Comprobaremos de la igualdad de las varianzas entre los conjuntos de datos de pasajeros que sobrevivieron y no. Para ello usamos la función `var.test`.

El contraste de varianzas se realiza mediante un contraste de hipótesis, donde aceptar H0 significaría que las varianzas son iguales.

El valor p-value que obtenemos es de 0.15. Este valor indica que rechazando la hipótesis nula de igualdad de varianzas probablemente estaríamos comentiendo un error. Las varianzas en la edad de los pasajeros que sobrevivieron y los que no sobrevivieron son iguales y podemos hacer esta afirmación con un nivel de confianza del 95%.

Variable Unit price: precio del billete unitario

Visualizamos el precio del billete unitario para los pasajeros que sobrevivieron frente a los que no. Podemos ver que la media y la mediana del precio de billete para los pasajeros que sobrevivieron están bastante separadas. El precio que corresponde a la mediana es alrededor de 13 y la media es más cercana a 20. Existe esta diferencia porque la distribución tiene una cola larga a la derecha, tenemos pasajeros que pagaron un precio muy alto por sus billetes.

La mayoría de los pasajeros que no sobrevivieron pagaron un precio bajo por el billete, hay pocas personas que pagaron precios más altos y no sobrevivieron.

Ninguna de las dos distribuciones es normal, aunque para el fin de hacer un contraste de hipótesos sobre la media podemos asumir que la media poblacional se distribuye normalmente (teorema del límite central).

Realizamos el test de igualdad de las varianzas para el precio del billete para pasajeros que sobrevivieron frente a los que no sobrevivieron. El valor p nos indica que podemos rechazar H0, es decir, las varianzas entre estos dos grupos de pasajeros son diferentes.

En los histogramas podemos ver que los pasajeros que sobrevivieron de media pagaron más por sus billetes. Tanto en el grupo de supervivientes como de no supervivientes la media es mayor que la mediana, dado que tenemos outliers de precio de billete unitario muy alto.

Variable SibSp: pareja o hermanos

Variable Parch: padres o hijos

El test de igualdad de las varianzas indica que no podemos rechazar la hipótesis nula. Entre los grupos que sobrevivieron y los que no sobrevivieron las varianzas de la variable Parch son iguales.

Variable Count.ticket: número de personas que viajaban con un mismo billete

Visualizamos un histograma que muestra la frecuencia de recuento de billetes: pasajeros que viajaban solos hasta pasajeros que viajaban con muchos acompañantes.

La media para el recuento de billetes es de 2 mientras que la mediana es 1.

En este caso tenemos un valor p muy bajo y por lo tanto podemos aceptar H1 que indica que las varianzas de la variable Count.ticket son diferentes en el grupo de supervivientes y no supervivientes.

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.

Contraste de hipótesis sobre el sexo y nivel de supervivencia

¿La proporción de supervivencia en hombres es inferior a la de las mujeres?

Acorde con la pregunta planteada, realizaremos un contraste de hipótesis con las siguientes hipótesis de partida:

Hipótesis nula - H0: La proporción de hombres que sobreviven es igual o mayor a 50% -> H0: p >= p0, siendo p0 = 0.5

Hipótesis alternativa - H1: La proporción de hombres es menor a 50% -> H1: p < p0, siendo po = 0.5

Este contraste de hipótesis representa un test de la proporción de la población por la cola izquierda, por tanto, el valor p0 representará, bajo la hipótesis nula, el límite inferior de la supuesta verdadera proporción de la población.

Este contraste rechazará la HO si el estadístico calculado es menor o igual al valor crítico (con signo negativo) al nivel de significación estimado.

El hecho que la proporción de hombres que sobreviven sea menor que 0.5, implicaría que la proporción de hombres que sobreviven es inferior a la de las mujeres.

Debido a que es un contraste de hipótesis sobre proporción de una muestra, y esta muestra es considerada grande (n > 30), podremos definir un estadístico de contraste como una observación de una variable aleatoria que se distribuye aproximadamente como una N(0,1).

Debido al planteamiento de la hipótesis nula y su alternativa y por cómo se han descrito las hipótesis H0 y H1, se observa que la hipótesis alternativa es unilateral, puesto que se plantea como un límite a un solo valor dado.

Determinamos el estadístico de contraste, en este caso, la muestra es grande y proviene de una distribución de Bernoulli de parámetro p, con lo cual, según el Teorema del Límite Central podremos utilizar el estadístico de contraste mostrado anteriormente.

Tal y como se ha planteado H0 y H1, siendo H1 p < 0.5, rechazaremos H0 si el valor del estadístico z\_p es menor que el valor crítico (en signo negativo), siendo este el caso -6.70 < -1.64.

Según el estadístico de contraste utilizado y el valor crítico calculado, al ser el primero (-6.70) menor que el segundo (-1.64), al 0.05 de nivel de significancia, podemos rechazar la hipótesis nula de que la proporción de supervivencia en hombres es mayor o igual a 50% respecto al de las mujeres.

De igual forma, al ser valor p casi nulo y por tanto menor que nuestro nivel de significación (0.05), el valor p es significativo y podemos rechazar, confirmando el contraste anterior, la hipótesis nula.

Siguiendo la misma metodología, responderemos a las siguientes preguntas:

¿La proporción de supervivencia en menores es inferior a la de los mayores?

Hipótesis nula - H0: La proporción de menores que sobreviven es igual o mayor a 50% -> H0: p >= p0, siendo po = 0.5

Hipótesis alternativa - H1: La proporción de menores es menor a 50% -> H1: p < p0, siendo po = 0.5

\*\*OLGA\*\* algo no va bien aquí, este -4.43 de donde sale? lo de hombres y mujeres tampoco es correcto

Según el estadístico de contraste utilizado y el valor crítico calculado, al ser el primero (-4.43) menor que el segundo (-1.64), al 0.05 de nivel de significancia, podemos rechazar la hipótesis nula de que la proporción de supervivencia en hombres es mayor o igual a 50% respecto al de las mujeres.

De igual forma, al ser valor p casi nulo y por tanto menor que nuestro nivel de significación (0.05), el valor p es significativo y podemos rechazar, confirmando el contraste anterior, la hipótesis nula.

¿La proporción de supervivencia en viajeros con billete único es inferior a la de los viajeros con billete en conjunto?

Hipótesis nula - H0: La proporción de viajeros con billete único que sobreviven es igual o mayor a 50% -> H0: p >= p0, siendo po = 0.5

Hipótesis alternativa - H1: La proporción de viajeros con billete único es menor a 50% -> H1: p < p0, siendo po = 0.5

Según el estadístico de contraste utilizado y el valor crítico calculado, al ser el primero (-10.92) menor que el segundo (-1.64), al 0.05 de nivel de significancia, podemos rechazar la hipótesis nula de que la proporción de supervivencia en hombres es mayor o igual a 50% respecto al de las mujeres.

De igual forma, al ser valor p casi nulo y por tanto menor que nuestro nivel de significación (0.05), el valor p es significativo y podemos rechazar, confirmando el contraste anterior, la hipótesis nula.

Relación entre supervivencia y billete único o grupal

Utilizamos el test Chi cuadrado para comprobar si los pasajeros de primera clase sobrevivieron en mayor medida que los pasajeros de segunda clase.

El test se aplica en R con un nivel de confianza por defecto del 95%.

Crearemos dos tablas de contingencia entre las variables que indican supervencia y la que indica el tipo de billete. Una primera en valores absolutos, y una segunda mostrando las proporciones.

Dos variables categóricas que forman parte de una tabla de contingencia pueden ser sujetas a un test de independencia. Este test puede ser representado for el ChiSquare Test mediante un contraste de hipótesis.

El contraste se realizará para observar si las dos variables son independientes o no, por tanto, podemos plantear las hipótesis como se indica a continuación:

Hipótesis nula - H0: Las variables Survived y Tipo Billete son independientes.

Hipótesis alternativa - H1: Las variables Survived y Tipo Billete están relacionadas.

El cálculo de los grados de libertad para una distribución Chi Square se calcula como df = (c - 1)(r - 1) donde c es el número´de columnas y r el número de filas. En nuestro caso c = r = 2, por lo tanto df = (2 - 1)(2 - 1) = 1.

Debido a que el estadístico Chi Square es mayor al valor crítico calculado de la distribución Chi Square (con un grado de libertad y con un nivel de significación de 0.05), podemos rechazar la hipótesis nula de que no hay relación entre Survived y Tipo de Billete, es decir, rechazamos la hipótesis al 95 % de nivel de confianza de que tales variables sean independientes.

Como el valor de p es menor que el nivel de significancia, podemos rechazar la hipótesis nula de que las variables Survived y Tipo de Billete son independientes, cuadrando este resultado con los cálculos anteriores.

Por tanto, se puede afirmar al 95% de nivel de confianza que las variables Survived y Tipo de Billete \*\*están relacionadas\*\*.

Contraste de hipótesis sobre la edad de supervivencia

Planteamos la pregunta de investigación: Queremos saber si las personas que sobrevivieron eran más jóvenes que las personas que murieron en el accidente o, por el contrario, la edad para las personas que muerieron y sobrevivieron es similar.

Nos encontramos ante el caso en el que podemos asumir la normalidad de las distribuciones de la media de las dos muestras pero sabemos que las varianzas de las muestras \*\*(EL RESULTADO HA CAMBIADO)\*\* son diferentes. Se trata de un test paramétrico de dos muestras.

\*\*OLGA: ok, entonces var.equal = TRUE\*\*

H0: La media de edad de las personas supervivientes es la misma que de las personas que muerieron

H1: La media de edad de las personas que murieron es mayor que la de las personas supervivientes.

Usamos un test paramétrico definido por `t.test`

Con un valor p-value igual por debajo de 0.05 podemos aceptar la hipótesis H1. Con un nivel de confianza del 95% afirmamos que las personas que murieron en el Titanic tienen una media de edad más alta que las personas que sobrevivieron.

Correlaciones

Partimos el conjunto train en subconjuntos de train y validación. De esta manera dispondremos de datos para comprobar el funcionamiento de los modelos que construyamos.

\*\*Regresión\*\*

Aplicamos una regresión logística para predecir la probabilidad de supervivencia usando como variable predictora la variable sexo.

Según los datos que nos proporciona la función `summary` la variable sexo es significativa para predecir la supervivencia y ser de sexo masculino reduce la probabilidad de sobrevivir.

Calculamos los Odds Ratio. Los odds ratio indican que ser hombre es factor de protección para la clase 1 (sobrevivir). Los hombres están "protegidos" de la supervivencia en comparación con las mujeres.

Usamos los datos de validación para realizar una predicción de supervivencia. Tenemos que 75 instancias de supervivencia y 157 instancias de no supervivencia han sido predichas correctamente. 39 instancias de supervivencia real han sido predichas como no supervivencia. 26 instancias de no supervivencia fueron predichas incorrectamente como supervivencia.

El modelo de regresión que utiliza únicamente la variable Sexo para predecir la supervivencia tiene una precisión del 78%.

Aplicamos una regresión logística para comprobar si las variables Sexo y Clase son significativas para la supervivencia en el accidente.

Ambas variables son significativas y así lo indican los asteriscos junto a los valores Pr(>|z|).

Ser hombre, viajar en clase 2 o clase 3 reduce la posibilidad de supervivencia con respecto a ser mujer y viajar en primera clase. Esto viene indicado por el signo negativo que acompaña el valor del coeficiente para Sexmale, Pclass2 y Pclass3.

Calculamos los odds ratio. Igual que en el caso anterior vemos que la probabilidad ser hombre y sobrevivir es mucho menor que la de ser mujer y sobrevivir. Por clases el OR indica que estar en segunda o tercera clase es factor de protección (pocas probable sobrevivir) respecto a la primera clase.

Realizamos la predicción sobre el conjunto de validación que obtuvimos y con el resultado creamos una matriz de confusión y calculamos la precisión. El resultado es igual que en la regresión anterior.

Por último generamos un modelo que incluye todas las variables disponibles

En los datos hemos detectado que había hasta 8 pasajeros con el mismo billete. Posiblemente se tratara de familia o amigos. Usaremos esta variable para crear un modelo.

El resultado del modelo indica que cuando tres o cuatro personas viajaban juntas, la probabilidad de sobrevivir aumentaba (signo positivo del coeficiente).

Calculamos los ORS. Para las personas que viajaban con otra persona, otras dos o tres personas el "riesgo" de sobrevivir es mucho mayor.

\*\*Árboles de decisión CART\*\*

En primer lugar creamos un árbol de decisión para predecir la supervivencia de los pasajeros del Titanic tomando en cuanta las variables explicatovas del conjunto de entrenamiento y la etiqueta de clase (Supervivencia o no)

Usamos los datos del conjunto train para entrenar el modelo y visualizamos el árbol y los datos del modelo.

El árbol ha realizado las particiones de los datos en base a la variable Sex. Los pasajeros de sexo masculino, mayores de 14 años tienen como clase por defecto la no supervivencia. Los pasajeros que pagaron entre 26 y 31 dólares y menores de 52 años no sobreviven. Los pasajeros menores de 14 años si viajaban con menos de 3 hermanos casi todos sobreviven.

Por otra parte, los pasajeros de sexo Femenino en que viajaban en tercera clase sobrevivieron si viajaban con menos de 5 personas (count ticket), si pagaron menos de 8.1 dólares y si tenían edad menor de 28 años. Todos los demás se clasifican como no supervivientes. Las mujeres en otras clases distintas de la tercera sobrevivieron en su mayoría.

En el árbol podemos ver que se ha considerado variables bastante diferentes para clasificar a los hombres y a las mujeres. Para los hombres la edad ha sido un factor importante para la supervivencia, principalmente sobrevivieron los hombres menores de 14 años. En el caso de las mujeres el factor determinante ha sido la clase y las personas con las que viajaban.

Predecimos usando los datos de validación.

Con la matriz de confusión y el valor de Accuracy podemos ver que la predicción ha mejorado con respecto a la regresión logística.

\*\*Árbol de decisión usando cross validation y búsqueda de mejor parámetro\*\*

Podemos entrenar un árbol de decisión usando cross validation y la búsqueda en rejilla.

Haremos la validación cruzada o crossvalidation con 10 folds y busqueda de los mejores parámetros usando el expand.grid. Utilizaremos la métrica Accuracy que mide el porcentaje de instancias correctas sobre total. Entrenamos el modelo con los datos trainX y trainy, establecemos el método que es el mismo que aplicamos en el primer árbol de decisión de la pec (C5.0) como parámetro de la función.

5. Representación de los resultados a partir de tablas y gráficas.

6. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las

conclusiones? ¿Los resultados permiten responder al problema?

7. Código: Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos. Si lo preferís, también podéis trabajar en Python.

## Including Plots

Representación de resultados

Tabla resumen de las variables cualitativas (datos completos)

Realizaremos una tabla resumen con las frecuencias relativas y las frecuencias absolutas de las variables cualitativas.

Creamos un dataframe auxiliar para generar nuestra tabla.

Calculamos, para todos los campos, la frecuencia relativa y absoluta a través del contaje dividido por el número total de filas del dataframe.

Tabla resumen de las variables cuantitativas (datos completos)

Realizaremos una tabla resumen con los estadísticos principales de tendencia central y dispersión, con medidas robustas y no robustas.

Para ello, utilizaremos tres funciones que nos aportarán diferentes estadísticos a utilizar:

describe()

winsor.mean() (aplicaremos unos límites del 5 %).

stat.desc()

Estas tres funciones nos darán diversos estadísticos que uniremos y ordenaremos en una única tabla para mostrar un completo resumen estadístico de las variables cuantitativas.

Finalmente, creamos una tabla mediante la función kable().