|  |  |
| --- | --- |
| 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**.

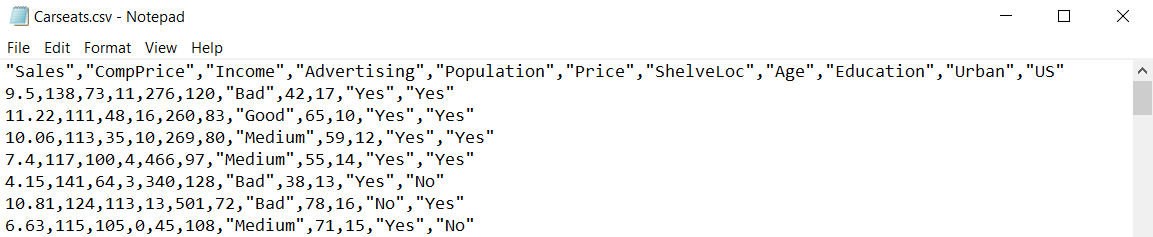
La ubicación en kaggle del dataset utilizado se muestra en el siguiente link:

<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):

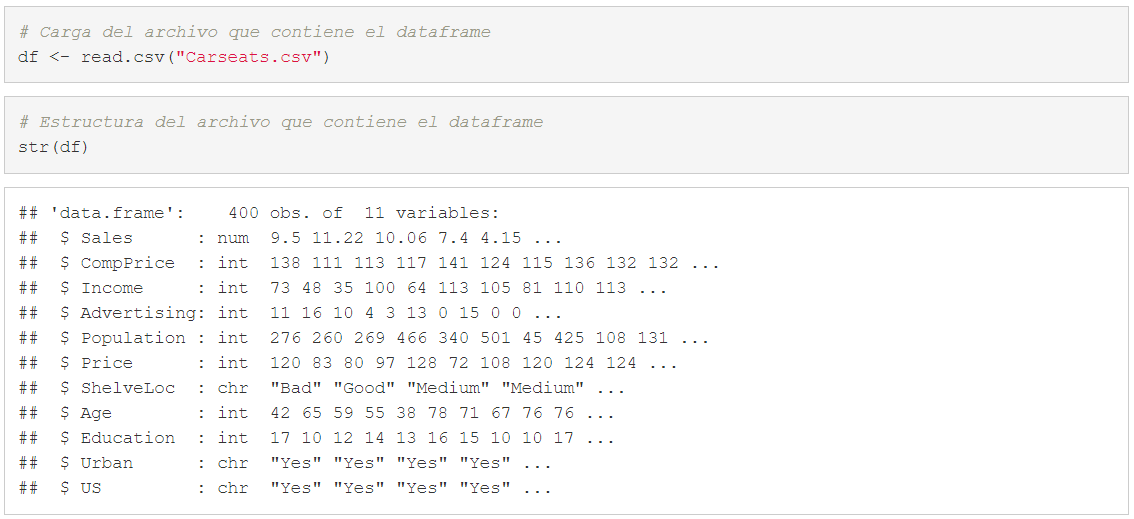
* **CompPrice**: Numérico entero (int)
* **Income**: Numérico entero (int)
* **Advertising**: Numérico entero (int)
* **Population**: Numérico entero (int)
* **Price**: Numérico entero (int)
* **Age**: Numérico entero (int)
* **Education**: Numérico entero (int)

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?

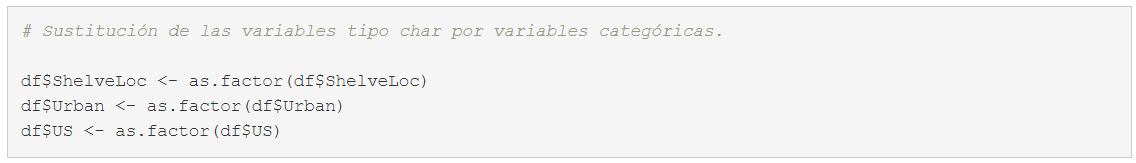
¿Las sillitas se venden igual en ubicaciones de ingresos altos y ingresos bajos?

¿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

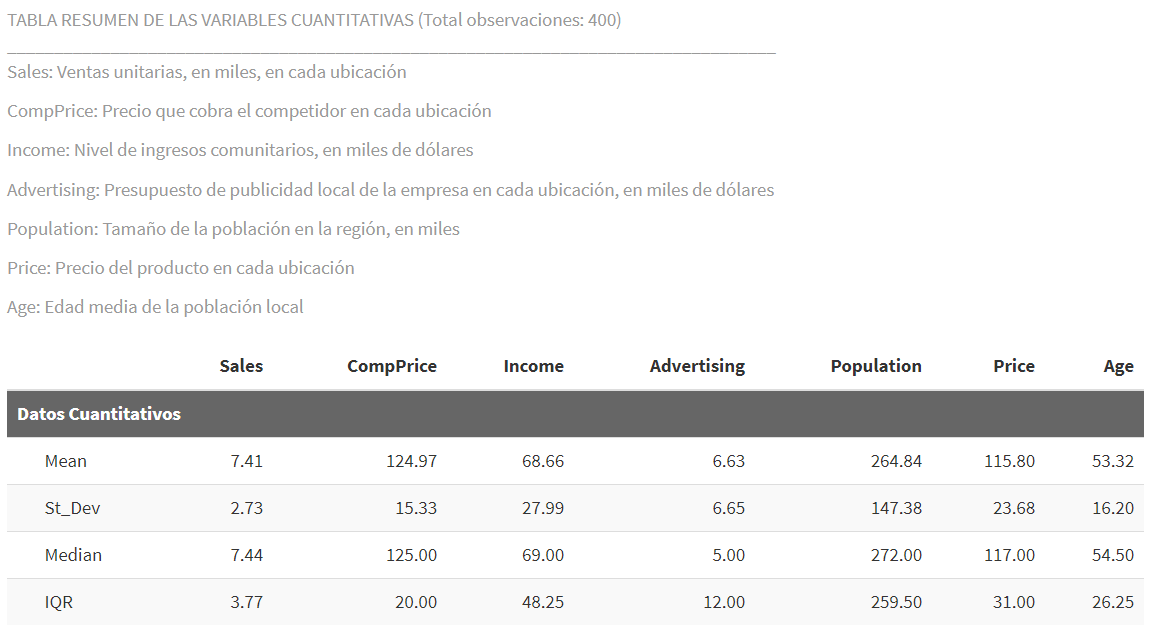
Carlos: OKIS

# 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.

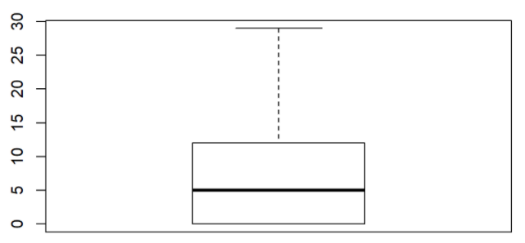
Nota:

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

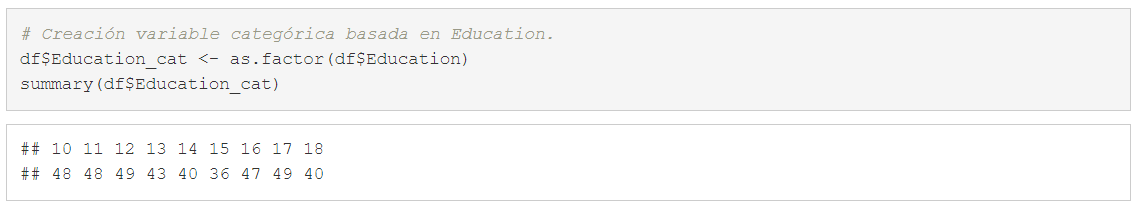
Previo a la exploración de valores cero, nulos o extremos, se realizarán otras tareas de limpieza de datos.

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:

Comprobamos si existen valores que contienen ceros en el juego de datos. Esto lo podremos averiguar aplicando XXX mediante la función XXX.

Comprobamos que el archivo Carseats.csv posee valores iguales a cero en la variable “Advertising”.

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/>