|  |  |
| --- | --- |
| 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.

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

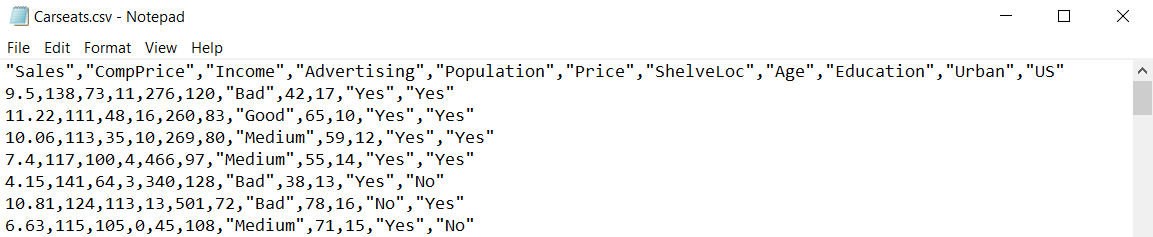
La ubicación en kaggle del dataset utilizado se muestra en el siguiente link:

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

Los datos originales sin modificar corresponden a estas variables:

* **Sales** (Ventas unitarias, 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):

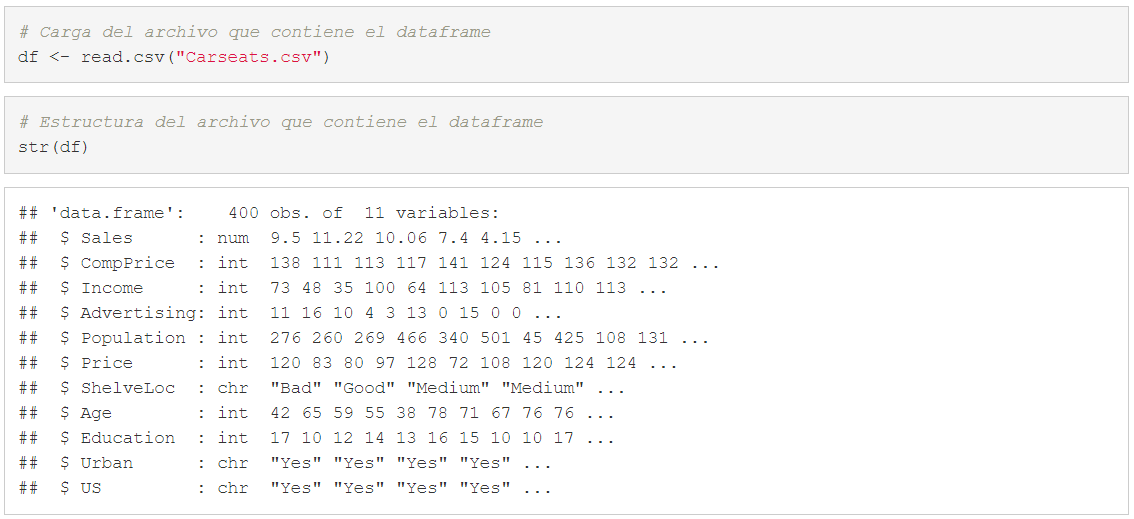
* **CompPrice**: Numérico entero (int)
* **Income**: Numérico entero (int)
* **Advertising**: Numérico entero (int)
* **Population**: Numérico entero (int)
* **Price**: Numérico entero (int)
* **Age**: Numérico entero (int)
* **Education**: Numérico entero (int)

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"



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

# ¿Qué problema pretende responder el dataset?.

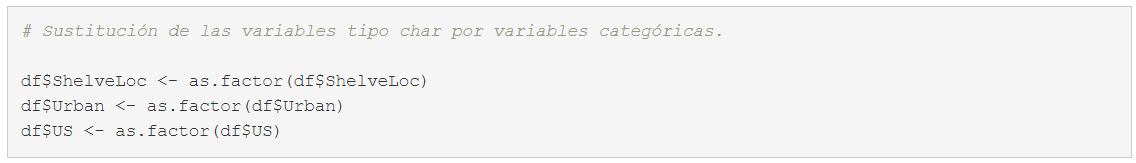
XXX

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

# 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**



# Integración.

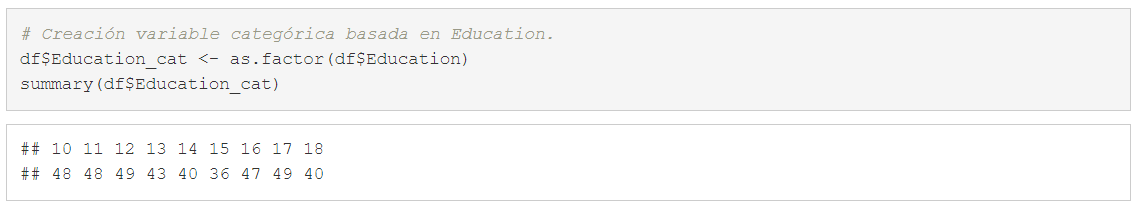
En primer lugar XXX

Integración y selección de los datos de interés a analizar.

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

# Selección.

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.



# Limpieza de datos



# ¿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.



Gestión de los valores nulos:

**Carlos:**

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

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).

La estrategia 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”.

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.

# Análisis de datos



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

XXX

# 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.

# 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/>