|  |  |
| --- | --- |
| 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](#_Toc71894296)

[1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder? 3](#_Toc71894297)

[1.1 Descripción del dataset. 3](#_Toc71894298)

[1.2 Por qué es importante el dataset? 3](#_Toc71894299)

[1.3 ¿Qué problema pretende responder el dataset?. 3](#_Toc71894300)

[2. Integración y selección de los datos de interés a analiza 4](#_Toc71894301)

[2.1 Integración. 4](#_Toc71894303)

[2.2 Integración. 4](#_Toc71894304)

[3. Limpieza de datos 4](#_Toc71894305)

[3.1 ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos? 4](#_Toc71894307)

[3.2 Identificación y tratamiento de valores extremos. 4](#_Toc71894308)

[4. Análisis de datos 5](#_Toc71894309)

[4.1 Selección de los grupos de datos que se quieren analizar/comparar (planificaciónde los análisis a aplicar). 5](#_Toc71894311)

[4.2 Comprobación de la normalidad y homogeneidad de la varianza. 5](#_Toc71894312)

[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. 5](#_Toc71894313)

[5. Representación de los resultados a partir de tablas y gráficas 6](#_Toc71894314)

[6. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema? 7](#_Toc71894316)

[6.1 Resolución del problema. 7](#_Toc71894318)

[6.2 Conclusiones a partir de los datos obtenidos. 7](#_Toc71894319)

[6.3 ¿Los resultados permiten responder al problema? 7](#_Toc71894320)

[7. Código 8](#_Toc71894321)

[8. Tabla de contribuciones al trabajo 9](#_Toc71894323)

[Referencias / Fuentes de Información 9](#_Toc71894325)

# 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.

XXX

# 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 analiza



# Integración.

XXX

Integración y selección de los datos de interés a analizar.

# Integración.

XXX

Podemos crear una variable nueva con precio de venta vs precio de venta del competidor %

# Limpieza de datos



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

XXX

Para explorar

#install.packages("ISLR")

library("ISLR")

#install.packages("SmartEDA")

library("SmartEDA")

Valores = cero

Valores extremos

Valores centinela

# Identificación y tratamiento de valores extremos.

XXX

# 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.