Ejercicio 5: Documentación

Dataset Original: Melboune (melb_data.csv)

Columnas

- Cantidad: 21
- Nombres: Suburb, Address, Rooms, Type, Price, Method, SellerG, Date, Distance, Postcode, Bedroom2, Bathroom, Car, Landsize, BuildingArea, YearBuilt, CouncilArea, Lattitude, Longtitude, Regionname, Propertycount

Filas

- Cantidad: 13580

1 - Criterios de exclusión e inclusión:

- Se eliminan filas, consideradas outliers de la columna **Price**, a través del método de caja o el método que utiliza el boxplot para marcar los outliers. La cantidad de filas eliminadas es de 612.
- Se elimina la columna **Method**, por la falta de información que aporta y considerada poco relevante para nuestro análisis.
- Se elimina la columna **SellerG**, variable categórica de poca relevancia a nuestro criterio. También consideramos la cantidad de valores distintos que tenía la columna. Se consideró agruparla y así conservarla, pero no teníamos las herramientas suficientes, ni el conocimiento de dominio para efectuar esta acción.
- Se elimina la columna **Bedroom2**, la cual consideramos que anunciando la cantidad de habitaciones es suficiente para el análisis.
- Se elimina la columna **Propertycount**, la cual se descarta porque no consideramos que influya la cantidad de propiedades en el precio final de la propiedad a analizar.
- Las columnas **Lattitude** y **Longtitude**, se conservan hasta el guardado del csv del primer entregable, ya que son consideradas importantes para muchos análisis de proximidad si así se requiere. En la segunda parte, se termina descartando por falta de uso.
- La columna **Address** fue considerada como una columna categórica que podía ser de utilidad, pero al final del análisis del entregable 1, fue descartada antes de ser guarda, ya que era una columna con mucha variación y se nos hacía difícil de explicar a la hora de cotizar una propiedad.

2 - Características Categóricas

- **Suburb**: Suburbio en Melbourne. Columna con valores categóricos. 314 valores únicos.
- Date: Fecha de venta de la propiedad. Columna categórica ordinal, con 58 valores únicos.
- **CouncilArea**: Se la consideró como las ciudades dentro de la metrópolis de Melbourne. Variable categórica con 33 valores únicos.
- **Postcode**: Códigos postales dentro de Melbourne. 198 valores únicos.
- **Regionname**: Regiones dentro de Melbourne. 8 valores únicos iniciales.
- **AgeRange**: Rango de edad de la propiedad respecto a la fecha de la venta.

Todas las columnas categóricas fueron transformadas a columnas con el proceso del algoritmo de OneHotEncoding para el tratamiento de datos. No se descartaron ninguna, ya que no se lo consideró necesario.

3 – Características Numéricas

- **Rooms**: Cantidad de habitaciones de la propiedad. 9 valores únicos.
- **Price**: Precio de venta de la propiedad.
- **Distance**: Distancia de la propiedad hacia el centro ciudad.
- **Bathroom**: Cantidad de baños que tiene la propiedad.
- **Car**: Cantidad de autos que puede almacenar o que se pueden guardar dentro del garage de la propiedad.
- **Landsize**: Área de la propiedad.
- **BuildingArea**: Área de construcción de la propiedad. Con algunos valores nulos, que luego se imputaron.
- **YearBuilt**: Año de construcción de la casa. Con algunos valores nulos, que luego se imputaron.
- **price_mean_city**: Precio medio diario de las publicaciones de la plataforma AirBnB en el mismo código postal. Se la consideró como un precio por día o como un precio de alquiler de la propiedad.

4 - Transformaciones

- Las columnas categóricas fueron codificadas y convertidas en columnas para un mejor manejo de los datos.
- La columna **Regioname** fue agrupada para convertirla en una variable categórica de 4 valores únicos.
- La columna Date se convierte en **Date_Quarter**, la cual maneja el cuatrimestre en donde fue vendida la propiedad.
- La columna AgeRange, es una variable categórica que almacena la diferencia de la fecha de venta y el año de construcción.
- Se imputa la moda a los datos faltantes de la columna **Car**.
- Se imputa la columna **CouncilArea** por código postal, utilizando el dataset de AirBnB. Se usa el merge para este proceso.
- Se utiliza el método de imputación por vecinos más cercanos (KNN) para imputar sobre las columnas **YearBuilt** y **BuildingArea**.
- Todas las variables numéricas fueron escaladas por el algoritmo StandarScaler de Sklearn.

Datos Aumentados

- Se agregan las 20 primeras componentes principales obtenidas por el método de PCA, aplicando sobre el conjunto de datos final.
- Se agrega la columna price_mean_city, que como explicamos, se la considera como el precio diario de alquiler según el dataset de AirBnB.