En este documento se realiza el reporte de las operaciones que realizaron para obtener el conjunto de datos final, incluyendo:

1. Criterios de exclusión (o inclusión) de filas
2. Interpretación de las columnas presentes
3. Todas las transofrmaciones realizadas

Este documento es de uso técnico exclusivamente, y su objetivo es permitir que otres desarrolladores puedan reproducir los mismos pasos y obtener el mismo resultado.

Criterios de exclusión de filas y columnas

Columnas del dataset que a priori no vamos a tener en cuenta para los análisis ya que no intervendrían en el valor de la propiedad

Address: se indica una dirección para cada propiedad, hay otras columnas que aportan información sobre la ubicación de la propiedad que permitirían un mejor agrupamiento por precio, por ejemplo, Regionname y CouncilArea.
SellerG: el nombre del vendedor no debería interferir con el valor de la propiedad
Date: la fecha de venta no sería indicativo del valor de la propiedad
Lattitude y Longtitud: podría pensarse a la longitud y latitud de la misma manera que a dirección (Address). Cabe aclarar que estas dos variables combinadas podrían ser útiles en cuanto a la ubicación específica de las propiedades en relación a otras características importantes de la región/barrio donde se encuentra (por ejemplo, cercanía a centros educativos, comercios etc).
Propertycount: es una variable relacionada a Suburb, se encuentra repetida con cada nombre de barrio diferente
Method: El método de venta no parece influir sobre el precio de las propiedades. El método SA parece tener valores más bajos pero representa un porcentaje de datos bajo en comparación al total de transacciones. según lo observado con anterioridad. Por otro lado el tipo de propiedad (Type) si parece influir en el precio.

Importante: La variable Postalcode se puede utilizar como punto de unión entre datasets. Será utilizada posteriormente para incluir datos desde el dataset Airbnb.

A partir de los análisis descriptivos realizados, los criterios para la eliminación de filas del dataset fueron los siguientes:

.- Se descartaron aquellas filas para las cuales Bedroom2 tenía un valor 0 (16) y valores mayores a 7 (20).
.- Se eliminaron las filas para las cuales Bathroom tenía un valor 0 (34) y valores mayores a 5 (9).
.- Se eliminaron las filas para las cuales Car tenía un valor mayor a 6 (21).
.- Se eliminaron las filas para las cuales Buildingarea era igual a 0 (no consideramos propiedades vendidas al "pozo").
.- Se eliminaron las filas para las cuales YearBuilt era menor al año 1850.
.- Se eliminaron las filas para las cuales Councilarea era igual a “Macedon Ranges” y “Moorabool” (no pertenecen a Melbourne) y Unavailable.

Selección de variables:

Variables categóricas:
.- Suburb: Barrio donde se encentra la propiedad
.- Type: tipo de propiedad, 3 valores posibles,
.- CouncilArea: nombre de la ciudad, 30 valores posibles.
.- Regionname: Región geográfica donde se encuentra la propiedad, 6 valores posibles.

Todas las características categóricas fueron codificadas con un método OneHotEncoding

Variables numéricas:
.- Rooms: Número de ambientes de la propiedad
.- Price: Precio de la propiedad
.- Distance: Distancia de la propiedad al centro
.- Postcode: Código postal
.- Bedroom2: Número de habitaciones recolectadas de un dataset secundario

.- Bathroom: Número de baños

. - Car: Número de cocheras

.- Landsize: Metros cuadrados de parque

.- BuildingArea: Metros cuadrados construidos

.- YearBuilt: Año de construcción.

Se seleccionaron las columnas “Price”, “city” y “zipcode” del dataset Airbnb. City es equivalente a la columna CouncilArea y es utilizada para imputar a la misma. A su vez, Zipcode es el equivalente a Postcode y estas columnas son utilizadas para “unir” ambos datasets. El precio promedio de alquiler diario (Price) es la variable que se agrega al dataset, si bien es el que menos información contiene es el que menos valores faltantes presenta.

Transformaciones realizadas:

.- Se agrupan los valores correspondientes a Victoria de la variable Regioname para que esta región geográfica sea representativa en el dataset.

.- En el dataset Airbnb se reemplazaron los códigos postales con valores erróneos o mal anotados

.- Los valores faltantes de la columna CouncilArea fueron imputados a partir de los presentes en la variable City de Airbnb, utilizando Postocode y Zipcode como variables “clave”. Los valores faltantes restantes fueron imputados utilizando la columna Suburb.

.- A partir del El precio promedio de alquiler diario (Price), se calculó el precio promedio de alquiler por ciudad, columna que fue agregada al dataset.

.- Las variables YearBuilt y BuildingArea fueron imputadas utilizando IterativeImputer con estimador KNeighborsRegressor. Para aplicar este método es necesario escalar los datos, dado que KNN utiliza como distancia entre valores la distancia euclidia, la cual necesita que los valores se encuentren en rangos equivalentes (se utiliza el método (PowerTransformer).

Componentes Principales (datos aumentados)

Se genero un dataframe (df_melb_final) con el resultado de la codificación de las variables categóricas y el escalado e imputación de las variables numéricas. Luego, se procedió a realizar un análisis de componentes principales utilizando todas las variables disponibles (489).

Como resultado, se obtuvo que con los primeros 21 PCAs recuperamos aproximadamente el 90% de la variancia contenida en los datos. Estos PCAs fueron incluidos en un nuevo dataframe df_melb_pca_final.to_csv