En un documento .pdf o .md realizar un reporte de las operaciones que realizaron para obtener el conjunto de datos final. Se debe incluir:

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. Debe ser detallado pero conciso.

Criterios de exclusión

Por columna

Se eliminaron las siguientes columnas:

Suburb: Tiene 314 categorías, por lo que no sirve para agrupar a las propiedades (como sí podremos hacer utilizando CouncilArea ó Regionname). Se la eliminará para un primer análisis y descripción de los datos, pero se la consevará para la caracterización posterior.

Address: Hay tantas direcciones como casas en el dataset, por lo que no aportan información relevante que caracterice a todo un conjunto para un posible modelo de predicción de precio.

Method: Al analizar los precios en función del método de venta (ver gráfico), vemos que SA (vendido en una subasta) posee precios más bajos que el resto de los métodos, pero solamente hay 91 propiedades vendidas de ese modo. El resto de los métodos de venta, no muestran diferencias significativas en los precios.

SellerG: Al no ser un atributo o característica de la propiedad, no debería ayudar a predecir el precio de las mismas.

Lattitude y Longtitude: Al igual que la dirección, son valores demasiado específicos y únicos, por lo que no permiten generalizar y aportar información robusta a un posible modelo predictivo. Además existen otras columnas con información geográfica (CouncilArea, Regionname).

Date: La fecha de venta también es demasiado específica y no es una característica identificatoria de la propiedad.

Observaciones:

Postalcode: Se mantiene esta columna para imputar datos faltantes cruzándolo con otro dataset. Luego de esto, se descarta por no aportar información extra a las características de las propiedades.

Por fila

Se eliminó la única fila cuyo año de construcción (YearBuilt) era anterior a 1800, por contener un dato erróneo.

Se eliminaron las propiedades con más de 6 espacios para guardar autos.

Se eliminan filas con 0 ó más de 6 ambientes.

Se removieron las propiedades cuya ciudad no pertenecía a Melbourne. En estos casos, pertenecían a Macedon Ranges y Moorabool.

Características seleccionadas

Características categóricas

Type: Tipo de propiedad. 3 valores posibles.

CouncilArea: Ciudad. 30 valores posibles.

Regionname: Región geográfica. 8 valores iniciales, luego de la transformación quedaron 6.

Suburb: Barrio. 314 valores posibles.

Todas las características categóricas fueron codificadas con un método OneHotEncoding.

Características numéricas

Rooms: Cantidad de habitaciones.

Price: Precio en dólares.

PrecioPromedioAirbnb: Precio promedio calculado por zona, en base a los precios diarios publicados por Airbnb, agrupando por ciudad.

Distance: Distancia al centro de la ciudad.

Bedroom2: Cantidad de habitaciones, recolectada de otro dataset.

Bathroom: Cantidad de baños.

Car: Cantidad de estacionamientos.

Propertycount: Cantidad de propiedades en el suburbio.

Landsize: Tamaño de la tierra.

BuildingArea: Área construída.

YearBuilt: Año de construcción.

Transformaciones

1. La columna Bathroom fue imputada en aquellas filas con dato erróneo igual a 0, ya que toda propiedad tiene al menos un baño.
2. En la columna CouncilArea habían filas con el valor “Unavailable” (no disponible). Se reemplazó inicialmente este valor por NaN.
3. Se agruparon las regiones Eastern Victoria, Northern Victoria, Western Victoria en una sola llamada Victoria. De esta manera, se pueden diferenciar las distintas zonas geográficas del área metropolitana, de los alrededores de la ciudad.
4. En el dataset complementario de Airbnb se reemplazaron los códigos postales erróneos manualmente, para poder imputar luego las ciudades faltantes.
5. Se imputaron las ciudades faltantes en la columna CouncilArea, a partir de los datos recolectados en la columna city del dataset complementario. Se utilizó el código postal para cruzar los datos.
6. A partir del precio diario de alquiler de Airbnb, se calculó el precio promedio de alquiler por ciudad, y se la agregó al dataset. Se imputaron los precios de las ciudades faltantes con el valor promedio de las restantes.

Encodings

1. Se separaron las columnas del dataset en variables numéricas y categóricas. A las categóricas se les aplicó un encoding de tipo “one-hot”. El resultado final se guardó en un dataframe de nombre **encoded_df.**

2. Las variables YearBuilt y BuildingArea se agregaron como columnas al dataset anterior. A partir de allí, se realizó una imputación de valores mediante un algoritmo de regresión como KNN. Para realizar la imputación previamente se escalaron los datos.
3. Luego estas dos variables que fueron imputadas se realizó un gráfico de distribución para ver los resultados antes y después de realizar la imputación.

Análisis de Componentes Principales

1. La matriz con los variables codificadas como one-hot, junto a las variables cuantitativas se unieron en un dataframe **melb_df_final**
2. A partir de este dataframe se realizó un análisis de componentes principales de las 481 variables.
3. A partir de allí se obtuvo la varianza explicada y se realizó el gráfico correspondiente. Se observa que se llega a un valor constante a partir de aprox 89 variables, por lo que se retuvieron las primeras 88 columnas.
4. Estas componentes se guardaron en un dataframe aparte con nombre `melb_pca.csv`

Composición del resultado

1. La matriz con los variables codificadas como one-hot, junto a las variables cuantitativas se guardaron en un dataframe **melb_df_final en formato csv.**
2. En otra matriz aparte se guardaron los resultados de las 88 primeras componentes principales obtenidas en el anterior punto.