Link a la base de datos

https://drive.google.com/file/d/1XD4ClHOeHYxJDNbwU8UwZYn4os9bseew/view?usp=sharing

Criterio de eliminación de casos

- Landsize menor a 10 y superior a 5000.
- Price mayor a 6.000.000
- Datos nulos en Airbnb_Mean

Criterio de selección de las variables categóricas

Se eligen las variables categóricas que más relevancia tienen a la hora de predecir el precio de una propiedad. Estas son: Type, Method, y CouncilArea. La variable Suburb es descartada debido a que ya se encuentra representada, de manera más general, por CouncilArea.

Descripción de las variables categóricas seleccionadas

- Type: Tipo de propiedad: Tres valores categóricos son posibles (h, u, t).
- Method. Cinco valores categóricos posibles (S, SP, PI, VB, SA).
- CouncilArea: Nombre del condado donde se encuentra.

Criterio de selección de las variables numéricas

Al igual que en el caso de las variables categóricas, a la hora de elegir las variables numéricas se analiza cuáles de ellas pueden incorporar información valiosa para predecir el precio de la propiedad. En este sentido, variables como la longitud y latitud son descartadas ya que la ubicación exacta de la propiedad no es estrictamente necesaria. Más convendría saber, por ejemplo, la zona donde se encuentra (representada por la variable CouncilArea). Por otro lado, variables como Bedroom2 y Bathroom están incluidas es Rooms, por lo que también son descatadas. Del dataset de Airbnb se mantienen las variables Airbnb_Mean y Airbnb_DevStd y se descartan Airbnb_Qtty y Airbnb_Median.

Descripción de las variables numéricas seleccionadas

- Rooms: Cantidad de habitaciones de la propiedad.
- Price: Precio de la propiedad (es lo que se quiere predecir).
- Distance: Distancia al centro de la ciudad.

- Landsize: Tamaño de la propiedad.
- Airbnb_Mean: Precio medio de la propiedad (extraído del dataset de Airbnb).
- Airbnb_DevStd: Desviación estándar del precio de la propiedad (extraído del dataset de Airbnb).

Transformaciones

1. Luego de aplicar el método OneHotEncoder y de agregar a la matriz resultante las variables numéricas, se incorporan las columnas BuildingArea y YearBuilt.
2. Se escalan los datos con el método Standard Scaler
3. Paralelamente, se escalan de manera aislada los datos de las dos columnas BuildingArea y YearBuilt.
4. Se imputan los datos faltantes teniendo en cuenta toda la matriz y, por otro lado, solo las dos columnas. Se usa el método KNN con n_neighbors = 2, random_state = 0 y max_iter = 10.
5. Matrices escaladas: se grafica la densidad vs BuildingArea y YearBuilt según cada método de imputación (matriz completa y columnas aisladas). También se incluye el dataset original con los datos nulos descartados.
6. Matrices no escaladas: también se grafica la densidad vs BuildingArea y YearBuilt según cada método de imputación (matriz completa y columnas aisladas). Se incluye el dataset original con los datos nulos descartados. El objetivo es evaluar la importancia del escalado de los datos a la hora de imputar datos faltantes.

Datos aumentados

Se incorporan las dataset las 5 primeras columnas de la matriz PCA ya que cubren aproximadamente un 50% de la varianza explicada.