Reporte de las operaciones que se realizaron para obtener el conjunto de datos final:

Criterios de exclusión e inclusión de columnas:

Columnas a eliminar de la base de datos “melb_df”

- columns_to_drop = ['Address', 'Method', 'SellerG', 'Date', 'Bedroom2, 'Regionname"]

Address: No es relevante saber la dirección de cada propiedad, son strings distintos todos y no aportan tanta información. Se elimina.

Análisis de variables “Rooms vs Bedrooms2”: La tabla de contingencia muestra la frecuencia conjunta de los valores de las variables en análisis. La variable Rooms aporta más información que Bedroom2, ya que la distribución de los valores es mejor en los datos de la variable Rooms.

Rooms / Bedroom2	1	2	3	4	5	6	7	8	10
0.0	0	5	8	3	0	0	0	0	0
1.0	663	21	5	2	0	0	0	0	0
2.0	16	3539	162	19	1	0	0	0	0
3.0	2	74	5626	175	18	1	0	0	0
4.0	0	8	73	2473	42	4	0	1	0
5.0	0	1	5	15	531	2	2	0	0
6.0	0	0	0	0	2	59	0	2	0
7.0	0	0	0	0	1	1	8	0	0
8.0	0	0	0	0	1	0	0	4	0
9.0	0	0	1	1	0	0	0	1	0
10.0	0	0	0	0	0	0	0	0	1
20.0	0	0	1	0	0	0	0	0	0

Method: No tiene relevancia para el análisis la forma en la que se vendió la propiedad. Se elimina.

SellerG: No interesa saber el nombre del vendedor. Se elimina.

Date: Indica la fecha en la que se vendió la propiedad, no nos interesa. Se elimina.

Regionname: Solamente indica la región norte, sur, este, oeste y combinaciones. No aporta información muy relevante al análisis.

Columnas a incluir en el df final de la base de datos “melb_df”

Deberíamos droppear también: **'YearBuilt'**, **'BuildingArea'**. No las sacamos porque en el entregable 2 se piden.

Dejamos:**'Type'**, **'Distance'**, **'Propertycount'**. Son relevantes para el análisis, se mantienen.

'Lattitude' y **'Longtitude'** se dejan debido a que puede ser utilizadas para análisis geoespacial cómo por ejemplo Interpolación con la Distancia Inversa Ponderada e interpolación con Kriging.

Columnas a eliminar de la base de datos “airbnb_df”

- columns_to_drop=['airbnb_record_count','airbnb_weekly_price_mean','airbnb_monthly_price_mean']

'Postcode' Se usa para hacer el merge con la df de airbnb, se mantiene. Una vez hecho el merge, se elimina.

'airbnb_record_count': No relevante, sólo indica la cantidad de datos que se usaron para sacar las medias del df de aribnb.

'airbnb_weekly_price_mean' y 'airbnb_monthly_price_mean': Puede estar calculada con muy pocos valores, no es confiable. Por ejemplo, un cálculo de media con un solo valor.

Columnas a eliminar de la base de datos "airbnb_df"

airbnb_price_mean: Media de precios de alquileres de propiedades agrupadas por zipcode. Se mantiene para enriquecer el dataset.

Criterios de exclusión e inclusión de filas:

En la variable de Council Area, encontramos que tenía una pérdida sistemática de los datos. Al ser una variable categórica teníamos 2 opciones: eliminar la variable o imputarla. Optamos por imputar la variable en sus filas con pérdidas con el valor de la moda de dicha variable ya que la misma era de tipo categórica.

Pensamos que podría ser una variable de interés más adelante, dependiendo del modelo de aprendizaje automático para el que esté destinado el data frame. Eliminar esas filas con pérdida sistemática sería equivalente a perder el 10% de los datos del data frame.

En caso de no ser necesaria la información aportada por la variable para el modelo de aprendizaje inteligente, se eliminaría la variable entera.

En las variables, BuildingArea y YearBuilt se encontró pérdida de datos aleatoria. Por lo que se mantuvieron todas las filas y las filas que tenían pérdida de datos fueron importadas con Regresión.

Características seleccionadas para el Data Frame analizado:

Características categóricas:

- 'Suburb'
- 'Type'
- 'CouncilArea'
- 'Regionname'

Características numéricas:

- 'Rooms'
- 'Price'
- 'Distance'
- 'Bathroom'
- 'Car'

- 'Landsize'
- 'BuildingArea'
- 'YearBuilt'
- 'Lattitude'
- 'Longtitude'
- 'Propertycount'
- 'zipcode'
- 'airbnb_price_mean'

Interpretación de columnas presentes:

- Rooms: Number of rooms
- Price: Price in dollars
- Type: **br** - *bedroom(s)/*habitaciones; **h** - *house, cottage, villa, semi, terrace*/casa, cabaña, casa de campo, edificio; **u** - *unit, duplex*/unitario, duplex; **t** - *townhouse; dev site - development site; o res - other residential*/casa urbana, sitio en desarrollo, otro tipo de residencial.
- Distance: Distancia al distrito financiero o corazón financiero (en inglés, CBD - Central Business District)
- Propertycount: Número de propiedades que existen en el barrio.
- Bathroom: Número de baños
- Car: Número de estacionamientos
- Landsize: Área del lote
- BuildingArea: Área cubierta
- CouncilArea: Consejo de gobierno de la zona
- airbnb_price_mean: Media de precios de alquileres de propiedades agrupadas por zipcode.

Transformaciones:

- ENCODING

Se dividió a las variables en numéricas y categóricas para poder realizar un chequeo de nulos extra y ver cuántas filas agregaría al data frame al realizar un encoding de tipo one hot. La variable que más columnas agregó al data frame fue la de Suburb con 248 valores únicos en sus datos. CouncilArea agregó 27 columnas y Type solo 3.

- IMPUTACIONES

Se realizó una imputación por KNRegressor como se pedía en la consigna de las variables YearBuilt y BuildingArea. Se analizó si deberíamos hacer un escalamiento de los datos, dado que los mismos podrían diferir entre las dos variables. El análisis concluyó que es necesario para realizar una imputación más acorde a los datos que se dan como input al algoritmo. El error en no escalar los datos consiste en que los datos de una observación en BuildingArea puede ser 4000 y e YearBuilt ronda siempre los 1900 a 2018. Entonces este diferencia entre los datos afecta al algoritmo, es mejor escalarlo para una optimización del funcionamiento del KNregressor.

Para las variables que tenían pérdida de datos: Car y airbnb_price_mean se realizó una imputación simple debido a que eran pocos datos faltantes y podían solventarse con una imputación por sus medianas (ya que la distribución de sus datos es asimétrica).

- PCA

Se realizó un PCA, o Análisis de Componentes Principales, es una técnica utilizada para comprender y visualizar patrones en datos que tienen muchas características. Los componentes etiquetados como PC1, PC2, PC3, PC4 y PC5, representan los patrones o variaciones más importantes en los datos y se ordenan según su importancia. Los datos originales tienen muchas características diferentes, como precio, distancia, año de construcción, etc. Estos componentes no se corresponden directamente con esas características originales, en cambio, son combinaciones de las características originales que capturan la máxima cantidad de variación en los datos. El Análisis de Componentes Principales permite descubrir patrones y reducir la dimensionalidad de datos complejos permitiendo visualizar los datos de una manera simplificada e identificar los factores importantes que contribuyen a la variación general en el conjunto de datos.