Introducción
Se comienza con dos dataset:

- Melbourne data frame
Obtenido de:
'https://cs.famaf.unc.edu.ar/~mteruel/datasets/diplodatos/melb_data.csv')
Consta de 13580 entradas y de las siguientes columnas:
'Suburb', 'Address', 'Rooms', 'Type', 'Price', 'Method', 'SellerG', 'Date', 'Distance', 'Postcode', 'Bedroom2', 'Bathroom', 'Car', 'Landsize', 'BuildingArea', 'YearBuilt', 'CouncilArea', 'Lattitude', 'Longtitude', 'Regionname', 'Propertycount'.

- Airbnb data frame
Obtenido de:
'https://cs.famaf.unc.edu.ar/~mteruel/datasets/diplodatos/cleansed_listings_dec18.csv'
Consta de 22895 entradas y de las siguientes columnas:
'id', 'listing_url', 'scrape_id', 'last_scraped', 'name', 'summary', 'space', 'description', 'neighborhood_overview', 'notes', 'transit', 'access', 'interaction', 'house_rules', 'picture_url', 'host_id', 'host_url', 'host_name', 'host_since', 'host_location', 'host_about', 'host_response_time', 'host_response_rate', 'host_is_superhost', 'host_thumbnail_url', 'host_picture_url', 'host_neighborhood', 'host_verifications', 'host_has_profile_pic', 'host_identity_verified', 'street', 'neighborhood', 'city', 'suburb', 'state', 'zipcode', 'smart_location', 'country_code', 'country', 'latitude', 'longitude', 'is_location_exact', 'property_type', 'room_type', 'accommodates', 'bathrooms', 'bedrooms', 'beds', 'bed_type', 'amenities', 'price', 'weekly_price', 'monthly_price', 'security_deposit', 'cleaning_fee', 'guests_included', 'extra_people', 'minimum_nights', 'maximum_nights', 'calendar_updated', 'has_availability', 'availability_30', 'availability_60', 'availability_90', 'availability_365', 'calendar_last_scraped', 'number_of_reviews', 'first_review', 'last_review', 'review_scores_rating', 'review_scores_accuracy', 'review_scores_cleanliness', 'review_scores_checkin', 'review_scores_communication', 'review_scores_location', 'review_scores_value', 'requires_license', 'license', 'instant_bookable', 'cancellation_policy', 'require_guest_profile_picture', 'require_guest_phone_verification', 'calculated_host_listings_count', 'reviews_per_month'.

Criterios de exclusión de ejemplos en el data frame de Melbourne

Para excluir ejemplos dentro de las características numéricas utilizamos el método de los 3 sigmas: filtramos los valores por encima o por debajo de 3 desvíos estándar de la media de cada columna numérica. Vemos así que el valor de desviación estándar relativa (el desvío dividido la media) disminuye luego de aplicado este método. De esta manera eliminamos un 7.2% de los registros.
Para el caso de las características categóricas ponderamos cada categoría calculando las frecuencias relativas y a las categorías que tenían menor frecuencia relativa que un 0.5% eliminamos los registros
De esta manera eliminamos un 5.6% de los registros.

Características seleccionadas

Características numéricas
Las características numéricas seleccionadas de Melbourne data frame:
- Rooms: Cantidad de habitaciones
- Bedroom2: Cantidad de dormitorios
- Bathrooms: Cantidad de baños
- Distance: Distancia al centro de la ciudad
- Landsize: Superficie del terreno
- BuildingSize: Superficie construida
- Car: Cantidad de cocheras
- YearBuilt: Año de construcción
- Postcode: Código Postal

El Precio para evaluar una eventual predicción.

Algunas características numéricas fueron agregadas de Airbnb df. Estas fueron vinculadas al data frame de Melbourne a través del Zipcode/Postcode. Se unificó el formato de los zipcodes llevando toda la columna a numérica, para solucionar los datos repetidos. Se tomaron aquellos Zipcodes de Airbnb con más de 25 entradas. Sólo un 12% de Postcodes no fueron vinculados. Para vincular a un Zipcode varios registros, las características fueron resumidas en su valor promedio. Luego, la característica numérica agregada de Airbnb data frame es:
- Price_mean: Precio promedio de alquiler por día
- Zipcode: Código postal.

Las columnas de Zipcode y Postcode fueron eliminadas luego del merge de data frames.

Características categóricas
Las características categóricas seleccionadas de Melbourne data frame:
- Type: tipo de propiedad. 3 valores posibles
- CouncilArea: Municipio. 33 valores posibles

Transformaciones
- Todas las características categóricas fueron codificadas con el método getdummies de la librería de Pandas. Este método es equivalente al método OneHotEncoder.
- Todas las características numéricas, menos el precio (la cual es la variable objetivo), fueron estandarizadas utilizando el método MinMaxScaler, de la librería sklearn.preprocessing. Este método escalea los datos al intervalo [0,1].
- Todas las características numéricas con valores nulos (*) fueron imputadas utilizando el algoritmo KNN de la librería sklearn.impute, con n_neighbours = 5.
*Cantidad de valores nulos por columna:

Car	59
BuildingArea	5830
YearBuilt	4818
CouncilArea	1074
airbnb_price_mean	1464

Datos aumentados
- Se agregaron las 5 primeras columnas obtenidas a través del método de PCA, aplicado sobre el conjunto de datos totalmente procesado, menos la columna Price. Estas 5 columnas representan un 50% de la varianza total.
- Se agregó al final del procesamiento la columna Price para una eventual predicción en otro trabajo.

Dataset final
El dataset final queda con 247 columnas numéricas 60 y 12784 filas.