Introducción
Se comienza con dos dataset:

- Melbourne df
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 df:
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

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 lo que utilizamos es el criterio de filtrar aquellas categorías cuya participación es menor a un 0.5% del total de registros. De esta manera eliminamos un 5.6% de los registros.

Características seleccionadas

Características numéricas
Las características numéricas seleccionadas de Melbourne df:
- 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 las predicciones.

Algunas características numéricas fueron agregadas de Airbnb df. Estas fueron vinculadas al df de Melbourne a través del Zipcode/Postcode. 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 máximo, mínimo y promedio. Luego, las características numéricas agregadas de Airbnb df son:
- Price_min: Precio mínimo de alquiler por día
- Price_max: Precio máximo de alquiler por día
- Price_mean: Precio promedio de alquiler por día
- Weekly_price_min: Precio mínimo de alquiler por semana
- Weekly_price_max: Precio máximo de alquiler por semana
- Weekly_price_mean: Precio de alquiler por semana
- Monthly_price_min: Precio mínimo de alquiler por mes
- Monthly_price_max: Precio máximo de alquiler por mes

- Monthly_price_mean: Precio promedio de alquiler por mes
- Zipcode: Código postal

Características categóricas
Las características categóricas seleccionadas de Melbourne df:
- Type: tipo de propiedad. 3 valores posibles
- Suburb: Suburbio. 314 valores posibles
- CouncilArea: Municipio. 33 valores posibles

Y las características categóricas agregadas de Airbnb df:
- Neighborhood: Barrio. 59 valores posibles

Transformaciones:
- Todas las características categóricas fueron codificadas con el método getdummies.
- Todas las características numéricas, menos el precio (la cual es la variable objetivo), fueron estandarizadas utilizando el método MinMaxScaler.
- Todas las características numéricas con valores nulos (*) fueron imputadas utilizando el algoritmo KNN, con n_neighbours = 5.
*Cantidad de valores nulos por columna:

Car	111
BuildingArea	8783
YearBuilt	7179
CouncilArea	1485
zipcode	2270
airbnb_price_min	2270
airbnb_price_mean	2270
airbnb_price_max	2270
airbnb_weekly_price_min	10716
airbnb_weekly_price_mean	10716
airbnb_weekly_price_max	10716
airbnb_monthly_price_min	10899
airbnb_monthly_price_mean	10899
airbnb_monthly_price_max	10899

Datos aumentados
- Se agregaron las 20 primeras columnas obtenidas a través del método de PCA, aplicado sobre el conjunto de datos totalmente procesado, menos la columna Price. Estas 20 columnas representan a un 65% de la varianza total.
- Se agregó al final del procesamiento la columna Price.

Dataset final

El dataset final queda con 247 columnas numéricas y 18498 filas.