# Documentación

En el presente documento se indicarán, a modo de corolario, las operaciones que hemos realizado para obtener el conjunto de datos final para completar los entregables de la materia "Exploración y Curación de Datos". Se explicarán los siguientes aspectos:

- Criterios de exclusión (o inclusión) de filas
- Interpretación de las columnas presentes
- Todas las transformaciones realizadas

## Criterios de selección de filas

Comenzamos haciendo una copia del df, para mantener un original tal como está para poder volver a acceder al mismo en caso de ser necesario.

En el proceso de limpieza de la base de datos, se eliminaron los siguientes valores en los campos referidos al *Precio de Venta* y al *Código Postal*.

- Campo "Price": se eliminaron los outliers de la cola derecha, solo el 1% ya que el resto no está tan distante de la mediana de la distribución.
- Campo "Zipcode": nos quedamos con los registros que tengan una cantidad mínima de 2 registros ya que 8,5 es el primer cuartil.

## Interpretación de las columnas presentes

Para la creación del dataset final, uniendo ambas bases, hemos eliminado los siguientes campos de la base de Melbourne:

- Address: indica la dirección de la propiedad, la eliminamos porque es un dato exclusivo de cada casa y no nos ayudará a realizar estimaciones generales.
- Method: indica el método de venta, la eliminamos porque consideramos que no nos aporta valor para la estimación de valores de venta.
- SellerG: indica el Vendedor de la propiedad, la eliminamos porque consideramos que no nos aporta valor para la estimación de valores de venta.
- Date: indica la fecha de venta de la propiedad, la eliminamos porque consideramos que no nos aporta valor para la estimación de valores de venta.
- Bedroom2, indica la cantidad de habitaciones, la eliminamos porque tiene muchos faltantes y alta correlación con Rooms.
- BuildingArea: indica los metros construidos, la eliminamos porque tiene muchos valores faltantes. Luego la volveremos a llamar y le asignaremos valores por tratarse de un factor clave en la determinación del precio de una vivienda.
- YearBuilt: indica el año en que fue construida la casa, la eliminamos porque tiene muchos valores faltantes. Luego la volveremos a llamar y le asignaremos valores por tratarse de un factor clave en la determinación del precio de una vivienda.
- CouncilArea: indica el condado/municipio en donde está situada la propiedad, la eliminamos por ser un dato redundante teniendo el Código Postal y el barrio.
- Lattitude: indica la latitud donde está situada la casa, la eliminamos porque es un dato exclusivo de cada casa y no nos ayudará a realizar estimaciones generales.

- Longtitude: indica la longitud donde está situada la casa, la eliminamos porque es un dato exclusivo de cada casa y no nos ayudará a realizar estimaciones generales.

Por otro lado, los campos seleccionados de la base de Melbourne para confeccionar el dataset final son los siguientes:

- Suburb: indica el barrio de la vivienda, la elegimos porque consideramos que puede haber semejanza en el valor de las propiedades de un mismo barrio
- Rooms: indica la cantidad de habitaciones, la elegimos porque la cantidad de habitaciones influye en el valor de una propiedad. En general a mayor habitaciones, mayor precio.
- Type: indica el tipo de vivienda, la elegimos porque cada tipo de vivienda tiene cotizaciones diferentes de otro tipo.
- Price: indica el precio de venta de la propiedad, la elegimos porque es el dato que necesitamos evaluar.
- Distance: indica la distancia al centro de la ciudad, la elegimos porque la podemos utilizar para validar el dato del Código Postal.
- Postcode: indica el Código Postal, la elegimos porque nos ayudará a unir las bases y da buena información de geolocalización.
- Bathroom: indica la cantidad de baños, la elegimos porque el costo de construir un baño es significativo respecto del resto de las habitaciones ergo, le agrega valor a la propiedad.
- Car: indica la cantidad de espacios para autos, la elegimos porque le da nivel a las viviendas y en caso de departamentos influye significativamente en su valor.
- Landsize: indica los metros cuadrados del terreno, la elegimos porque aporta valor en el caso de las casas.
- Regionname: indica el nombre de la región donde está ubicada la propiedad, la elegimos porque agrega información que puede resultar de utilidad.
- Propertycount: indica la cantidad de propiedades en el barrio, la elegimos porque da idea de escasez y eso puede influir en el precio de venta.

Asimismo, los datos utilizados de la base de airbnb para generar el dataset final, además del Código Postal, son los siguientes:

- Price: indica el precio del alquiler diario de cada vivienda.
- Review_scores_value: indica la puntuación que le asignaron los inquilinos a la propiedad.
- Review_scores_location: indica la puntuación que le asignaron los inquilinos a la ubicación de la propiedad.

Luego creamos dos “data frame”, uno con las variables categóricas y otro con las numéricas:

1. Variables categóricas:
    a. Suburb (barrio), 314 valores posibles: a los fines de simplificar la resolución del ejercicio no la consideraremos para no generar más de 300 columnas con el método get.dummies.
    b. Type (tipo de propiedad), 3 valores posibles:

        i. h: house, cottage, villa, semi, terrace
        ii. u: unit
        iii. t: townhouse
    c. Regionname (nombre de la región), 8 valores posibles:
        i. Northern Metropolitan
        ii. Western Metropolitan
        iii. Southern Metropolitan
        iv. Eastern Metropolitan
        v. South-Eastern Metropolitan
        vi. Eastern Victoria
        vii. Northern Victoria
        viii. Western Victoria

2. Variables numéricas
    a. Rooms (cantidad de habitaciones)
    b. Price (precio de venta)
    c. Distance (distancia al centro)
    d. Postcode (código postal, de la base de Melbourne)
    e. Car (cantidad de espacio para guardar autos)
    f. Bathroom (cantidad de baños)
    g. Landasize (cantidad de metros cuadrados del terreno)
    h. zipcode (código postal, de la base de airbnb)
    i. airbnb_price_mean (valor del alquiler diario promedio por código postal)
    j. airbnb_record_count (cantidad de filas con valores no nulos en el precio de alquiler diario agrupados por código postal)
    k. airbnb_scores_value_mean (valor promedio de las evaluaciones de los inquilinos agrupadas por código postal)
    l. airbnb_scores_location_mean (valor promedio de las evaluaciones de los inquilinos agrupadas por código postal)

Las variables categóricas “Type” y Regionname” han sido codificadas mediante el método de pandas “get_dummies”, para lograr pasar los datos categóricos a numéricos, eliminando la primera columna para evitar problemas de colinealidad, de este modo se elimina el campo original y se crean dos columnas de unos y ceros para “Type” y siete columnas para “Regionname”. No se efectuaron codificaciones con la variable Suburb por tener una elevada cardinalidad.

## Transformaciones realizadas

Lo primero que hicimos fue normalizar todos los valores de cada una de las variables, obteniendo valores entre 0 y 1, mediante el método MinMaxScaler.

Luego realizamos la imputación de los valores faltantes de “YearBuilt” y “BuildingArea” mediante el método de vecinos cercanos (KNN), con 5 vecinos. Este modelo lo que hace es para cada valor faltante de la variable, imputarle un valor promedio entre los cinco vecinos más cercanos. Éstos se consiguen estudiando el resto de las variables del dataset que no cuentan con valores faltantes.

Para el resto de las variables, imputamos los datos faltantes utilizando los datos más frecuentes, debido a que los faltantes en estos campos son muy pocos.

A continuación, y para finalizar, se aplicó el método PCA, para realizar reducción de dimensionalidad al data frame procesado, es decir con las imputaciones mencionadas y con los valores normalizados. Hemos observado que más del 75% de la variabilidad de los datos se explica en los primeros 3 componentes principales, valor que consideramos aceptable para el análisis exploratorio.

Con el objetivo de mostrar una aplicación de análisis exploratorio, graficamos en 2 y en 3 dimensiones (dos  y tres componentes principales respectivamente) dos variables categóricas: por un lado el tipo de propiedad y además la región donde están situadas.