Grupo 31

Integrantes:

- Victor Nicolas Rocco
- Maria Mercedes Silva
- Santiago Franco
- Williams Gremoliche

Criterios de exclusión (o inclusión) de filas

Se eliminan los outliers de las variables con los siguientes criterios:

1. **YearBuilt:** se eliminan las construidas antes de 1900
2. **BuildingArea:** se eliminan las áreas construidas iguales a 0 o mayores a 250 m^2
3. **Price:** se eliminan las que cuestan menos de $250.000 o más de $2.000.000
4. **Rooms:** se eliminan las que tienen más de 4 habitaciones
5. **Landsize:** se eliminan los terrenos iguales a 0 o mayores a 1200 m^2
6. **Distance**: se eliminan las distancias al centro iguales a 0 o mayores a 22km

Se eliminaron variables que no tuvieran relación con el precio de las viviendas o estuvieran correlacionadas con otras variables ya incluidas. Ejemplos:

1. **SellerG** fue eliminada ya que el cambio de valor de esta variable no tiene relación con el precio final.
2. **Bedroom2** fue excluida ya que se verificó en la parte 1 del practico que tenía una fuerte correlación con **Rooms**

Características seleccionadas

Características categóricas

Se redujo el Dataframe y se trató las variables categóricas con OneHotEncoding.

1. Type: tipo de propiedad. 3 valores posibles. h=house,cottage,villa, semi,terrace; u= unit, duplex; t=townhouse
2. Suburb 314 valores posibles de suburbios
3. Regionname 8 valores posibles de regiones (West, North West, North, North east ...etc)

Características numéricas

1. **Rooms:** cantidad de habitaciones.
2. **Distance:** distancia al centro de la ciudad.
3. **Price:** precio de compra/venta de la propiedad.

4. **Landsize:** tamaño del terreno.
5. **BuildingArea:** metros cuadrados construidos.
6. **YearBuilt:** año de construcción.
7. **avg_price:** se agrega el precio promedio diario de publicaciones de la plataforma AirBnB en el mismo código postal. [Link al repositorio con datos externos].
8. **avg_weekly_price**: se agrega el precio semanal diario de publicaciones de la plataforma AirBnB en el mismo código postal. [Link al repositorio con datos externos].
9. **avg_monthly_price:** se agrega el precio promedio mensual de publicaciones de la plataforma AirBnB en el mismo código postal. [Link al repositorio con datos externos].

avg_price, **avg_monthly_price** y **avg_weekly_price** obtenidos de AirBnB y procesados en 20220607_entregable_parte_1_22.ipynb

Transformaciones:

1. Todas las características numéricas fueron escaladas usando MinMaxScaler previo a ser imputadas, solo los valores distintos a NaN. Al aplicar MinMaxScaler, reemplazaba los NaN por 0, así que solo se les aplicó a los valores no nulos, manteniendo los NaN para luego ser imputados.
2. Las columnas 'YearBuilt' y 'BuildingArea' fueron imputadas utilizando el algoritmo IterativeImputer, con el estimador KNeighborsRegressor. Se optó por un K=2 ya que imputaba más próximo a los datos originales.
3. Se escalaron los datos con MinMaxScaler entre -1 y 1
4. Se aplicó el método PCA para reducir la dimensionalidad y se calculó la varianza de las primeras 20 componentes.

Datos aumentados

1. Se agregan las 5 primeras columnas obtenidas a través del método de PCA, ya que estas representan prácticamente el 80% de las componentes principales, para luego ser aplicado sobre el conjunto de datos totalmente procesado.