REPORTE DE OPERACIONES

Criterios de exclusión

1. En función de lo visto en la visualización se quitarán los siguientes valores extremos:
    - Outliers en columna precio (2.5 q3)
    - Superficie de terreno mayor a 10000
    - Area construida mayor a 600
    - Se eliminan ejemplos donde el año de construcción es previo a 1900

2. A continuación se listan las columnas que no se consideran relevantes para definir el precio:
    - 'Address'
    - 'SellerG'
    - 'Date'
    - 'Distance'
    - 'Bedroom2'
    - 'Lattitude'
    - 'Longtitude'
    - 'Regionname'
    - 'Propertycount'

Características seleccionadas

Características categóricas

1. Type: tipo de propiedad. 3 valores posibles en el dataset, dentro de las siguientes opciones:
    - **br** - bedroom(s)
    - **h** - house,cottage,villa, semi,terrace
    - **u** - unit, dúplex
    - **t** – townhouse
    - **dev site** - development site
    - **res** - other residential.
2. Suburb: barrio, 312 barrios distintos
3. Method: método de venta, 5 valores posibles en el dataset, dentro de las siguientes opciones:
    - **S** - property sold
    - **SP** - property sold prior
    - **PI** - property passed in
    - **PN** - sold prior not disclosed
    - **SN** - sold not disclosed
    - **NB** - no bid
    - **VB** - vendor bid

- **W** - withdrawn prior to auction
- **SA** - sold after auction
- **SS** - sold after auction price not disclosed.
- **N/A** - price or highest bid not available.

La característica categórica Method fue codificada con un método OneHotEncoding utilizando los valores que detecta por defecto: por medio de la identificación de valores únicos (unique values) de la característica seleccionada.

Características numéricas

1. Rooms: cantidad de habitaciones
2. Price: precio de las propiedades.
3. Postcode: código postal
4. Bathroom: cantidad de baños
5. Car: cantidad de cocheras
6. Landsize: Tamaño del terreno
7. BuildingArea: área construida
8. YearBuilt: año de construcción
9. airbnb_mean_price: Se agrega el precio promedio diario de publicaciones de la plataforma AirBnB en el mismo código postal. [Link al repositorio con datos externos].
10. airbnb_record_count: Se agrega conteo de publicaciones de la plataforma AirBnB con el mismo código postal. [Link al repositorio con datos externos].
11. airbnb_weekly_price_mean: Se agrega el precio promedio semanal de publicaciones de la plataforma AirBnB en el mismo código postal. [Link al repositorio con datos externos].
12. airbnb_monthly_price_mean: Se agrega el precio promedio semanal de publicaciones de la plataforma AirBnB en el mismo código postal. [Link al repositorio con datos externos].

Transformaciones:

1. Todas las características numéricas fueron estandarizadas.
2. A la columna 'Bathroom' se le asignó el valor 1 a las filas que tenían valor 0 (34 filas).
3. A la columna 'Landsize' se le asignó la media a las filas que tenían valor 0 (1901 filas).
4. A la columna 'BuildingArea' se le asignó la media a las filas que tenían valor 0 (16 filas).
5. La columna 'YearBuilt' fue imputada utilizando el método 'IterativeImputer' con un estimador 'KNeighborsRegressor'
6. La columna 'BuildingArea' fue imputada utilizando el método 'IterativeImputer' con un estimador 'KNeighborsRegressor'

Datos aumentados

1. Se agregan las 2 últimas columnas obtenidas a través del método de PCA, aplicado sobre el conjunto de datos totalmente procesado.
2. Se agregan al final 5 columnas resultantes de la transformación One-hot encoding de la variable Method.