Ejercicio 5: Documentación entregable EyC

Grupo 6

Integrantes:

- Marilina Trevisan (marilinatrevisan@gmail.com).
- Gustavo Ariel Venchiarutti (gustavo.venchiarutti@gmail.com).
- Ani Salama (anisalama@gmail.com).
- Anahí Sulca (nanisulca@gmail.com).
- Agustín Trulli (agustintrulli@gmail.com).

Criterios de exclusión de ejemplos

1. Se eliminan las observaciones con valores extremos de la variable Price usando el metodo de rango interquartilico.

Combinación de Datasets

1. Combinamos los datasets de Melburne y airbnb con un LEFT JOIN para mantener todas las observaciones de la primera tabla, usando como campo para el join el zipcode.
De la tabla de Melbourne mantenemos todas las columnas, de la tabla de airbnb traemos solamente las siguientes:

```
airbnb_cols = ['description', 'neighborhood_overview',
   'street', 'neighborhood', 'city', 'suburb', 'state', 'zipcode',
   'price', 'weekly_price', 'monthly_price',
   'latitude', 'longitude']
```

Características seleccionadas

Características categóricas

1. Type: tipo de propiedad. 3 valores posibles

2. Regionname: Región general. 8 valores posibles

Estas dos variables categóricas fueron codificadas con un método OneHotEncoding utilizando todos sus valores posibles (3 para Type y 8 para Regionname).

Luego se concatenaron los arrays obtenidos a el dataset y se reordenaron las columnas para que cada array quede al lado de la variable categórica que le dio origen.

Transformaciones:

1. Las columnas “YearBuilt” y “BuildingArea” fueron imputadas utilizando el algoritmo IterativeImputer con un estimador KNeighborsRegressor.

Previamente se excluyeron de la variable “YearBuilt” se excluye una observación del año 1196.

De la variable “BuildingArea” se consideraron solo aquellas con una superficie de 275 o más, siedo 274 el percentil 95 de la distribución.

Datos aumentados

1. Se agregan las 4 primeras columnas obtenidas a través del método de PCA, aplicado sobre el conjunto de datos. Elegimos 4 columnas porque con esa cantidad se logra un 93% de explicabilidad de los datos.