Exploración y Curación de datos

Ejercicio 5: Documentación

Integrantes

Eugenia Primo

Ramiro Jiménez

Agustín Carchano

Mariangel Arias Ferreyra

PARTE 1

Valores extremos

Exploración gráfica de las variables para identificar valores extremos. Se identificaron *outliers* en las siguientes variables: "*Price, Distance, Landsize y BuildingArea".*

Se realizaron dos tipos de tratamientos para determinar los valores extremos:

- *Price*, *Rooms*, *Bathroom* y *Car:* criterio "blando" de 4 desvíos respecto a la media.
- *YearBuilt:* eliminación del año 1196 a partir de análisis gráfico, criterio arbitrario.

Columnas relevantes para la predicción del valor de la propiedad, primera aproximación:

Criterio de selección de variables: matriz de correlación

Variables relevantes:

- *Price*: precio de la propiedad (variable de análisis)
- *Rooms*: número de habitaciones que tiene la propiedad. Se eligió esta variable ya que presenta mayor correlación con el precio que **Bedroom2.**

- *Bathroom*: número de baños que posee la propiedad. Alta correlación.

- *Date*: fecha en que se vendió la propiedad. Datos desde julio de 2017 hasta septiembre de 2017

- *Car*: número de estacionamientos que posee la propiedad. Si bien la variable posee baja correlación con el precio, consideramos oportuno dejarla en el análisis en primera instancia.

- *YearBuilt*: año en que se construyó la propiedad. Exhibe una correlación negativa con el precio, lo que indicaría que mientras más reciente sea construida la propiedad, mayor será su precio.

- *Distance*: distancia de la propiedad a la zona céntrica comercial. Esta variable presentó una correlación baja, pero se optó por dejarla.

Categóricas:

- *Type*: tipo de propiedad: h (house, cottage, villa, semi, terrace), u (unit, duplex) y t (towhhouse).

- *Regionname*: región a la que pertenece la propiedad. Variable relevante, informa el valor de las propiedades en la zona.

- *CouncilArea*: Estado ("provincia") en que se encuentra la propiedad.

- *Postcode*: código postal de la propiedad.

- *Suburb*: barrio o zona en donde se ubica la propiedad.

Criterios de exclusión de variables

- *Bedroom2*: correlación alta con la variable Rooms por lo que aportan información similar.

- *Method*: el método de venta no aporta información relevante para el precio de la propiedad.

- *Seller*: tiene muchos valores categóricos y haciendo un análisis lógico sería muy difícil que la persona encargada de la venta tenga influencia sobre el precio.

- *Address, Lattitude y Longtitude*: una ubicación tan precisa de la propiedad probablemente no aporte demasiado al análisis. Existe otro tipo de referencias más útil para localizar a la propiedad (CouncilArea, Suburb, Regionname, Postcode).

- *BuildingArea y Landsize*: no muestran una correlación alta con el precio, además de exhibir distribuciones con muchos valores extremos. En particular, BuildingArea cuenta con numerosos valores faltantes.

- *PropertyCount*: correlación baja con el precio, no aporta información relevante al análisis.

Información adicional – Datos AirBnB

Unión de los dos dataframes (Melbourne y Airbnb) a partir de la variable "*zipcode*".

Cantidad mínima de registros de Zipcode: > 50

Agrupamiento de cada zipcode

Agregación del precio promedio por cada uno

Método: *Merge, how: left*

Imputación de valores faltantes

Columna *CouncilArea* con datos de *Suburb*

Mapeo de ubicación y reemplazo de valores

PARTE 2

Encodings, pasos:

- Filtramos nuestro dataframe mediante el comando d*f.columns.difference*

- codificación One-hot Encoding para crear una matriz *sickit learn.*

Codificación *OneHotEncoder* - Variables *BuildingArea* y *YearBuilt*

- Se filtraron las variables *BuildingArea* y *YearBuilt* (con el dataframe del práctico 1)
- Codificación *OneHotEncoder* para obtener una *Matriz Esparsa* (matriz_OH), la cual permite codificar las variables categóricas
- Nuevamente se vuelve a traer las variables BuildingArea del data-set original
- Escalado con minMaxScaler
- Finalmente, la imputación por KNN.

Método PCA

A la matriz resultante, se le aplica el método PCA para reducir dimensionalidad en los datos, y se agregan los componentes obtenidos al df.