Unificar modificaciones para lograr un primer dataset

Notebook Roberto

Posición [2] importa el dataset indicando que la columna 0 es el index para no duplicarlo

Posición [8] [9] [10] y [11] Cuenta las palabras del campo [‘description’] y las agrupa

Se exportó a un csv. Desde un Excel se tomaron las más significativas

Posición [13] genera la columna [‘ambientes’] 10292 registros no Nulos cuando [‘rooms’] es nulo

La columna [‘rooms’] tiene 73830 nulos originales

Posición [15] genera la columna [‘dormitorios’] que genera 34366 valores no-nulos

Posición [17] genera la columna[‘cochera’] que genera 41977 registros True, resto False

Segmentando se ve impacto positivo en los **apartment** que publican ‘cochera’

Las pruebas de impacto con la segmentación están en la posición [32]

Posición [20] genera columna [‘parrilla’] que genera 29778 registros True, resto False

Segmentando se ve impacto positivo en las **house** que publican ‘parrilla’

Posición [21] genera columna [‘patio’] que genera 17269 registros True

Esta es diferente, porque los **apartment** que publican ‘patio’ tienen menor precio que los que no. Esto puede ser xq están en planta baja o bien xq la superficie total incluye más descubierto. Lo dejaría para ver qué hacer

Posición [22] genera columna [‘pileta’] que genera 28821 registros True

Segmentando se ve impacto positivo en **apartment** y **house** de los que publican pileta

Posición [24] genera columna [‘vista’] que genera 33376 registros True

Segmentando se ve impacto positivo en **apartment** de los que publican vista o frente

Posiciones [25] y [26] generan la columna [‘usd\_per\_m2\_2’]

Sale de la columna [‘price\_per\_m2’] limitando entre 500 y 6000 donde claramente corresponde a usd x m2

Posición [30] genera la columna [‘usd\_m2\_final’] que toma:

[‘price\_usd\_per\_m2’] original si no es nulo

[‘usd\_per\_m2\_2’] si la anterior es nula

La columna original [‘price\_usd\_per\_m2’] tiene 52603 nulos

La columna nueva [‘usd\_per\_m2\_2’] tiene 32639 nulos (19964 menos)

Posición [31] chequeo de distribuciones entre [‘price\_usd\_per\_m2’] original y [usd\_per\_m2\_2]