Checkpoint 2 - Grupo 12

Análisis Exploratorio

El dataset sobre el que vamos a trabajar es un registro de anuncios de propiedades en venta en Argentina durante el año 2021. Para empezar, tenemos un total de 460.154 registros, es decir, anuncios. Para cada anuncio contamos con 20 columnas.

En esta primera etapa vamos a trabajar sobre qué columnas son importantes y de cuáles podemos prescindir. Para empezar, vamos a reducir nuestro estudio a las propiedades de tipo PH, departamento o casa dentro de Capital Federal cuyo tipo de operación sea venta en moneda USD. De esta forma la cantidad de registros válidos con los que nos quedamos son 89.709.

Preprocesamiento de Datos

Detallar las tareas más importantes que realizaron sobre el dataset, les dejamos algunas preguntas cómo guía:

1. ¿Se eliminaron columnas (Nombre de la columna y motivo de eliminación?
2. ¿Detectaron correlaciones interesantes (entre qué variables y qué coeficiente)?
3. ¿Generaron nuevos features?
4. ¿Encontraron valores atípicos?¿Cuáles?¿Qué técnicas utilizaron y qué decisiones tomaron?
5. ¿Qué columnas tenían datos faltantes?

   ¿En qué proporción? ¿Qué se hizo con estos registros?

Luego de la filtración según los parámetros establecidos, vamos a ver los datos nulos dentro del dataset bajo estudio. Cuando realizamos el porcentaje de celdas vacías dentro de todas las columnas, observamos que las columnas *place_l5 y place_l6* tienen un 100% de celdas nulas. Sin duda, eliminamos dichas columnas. Ahora, *place_l4* tiene un porcentaje de nulos del 96%, valor muy elevado como para hacer algún tipo de corrección o relleno, por lo tanto, eliminamos la columna. De esta forma eliminamos directamente tres columnas.

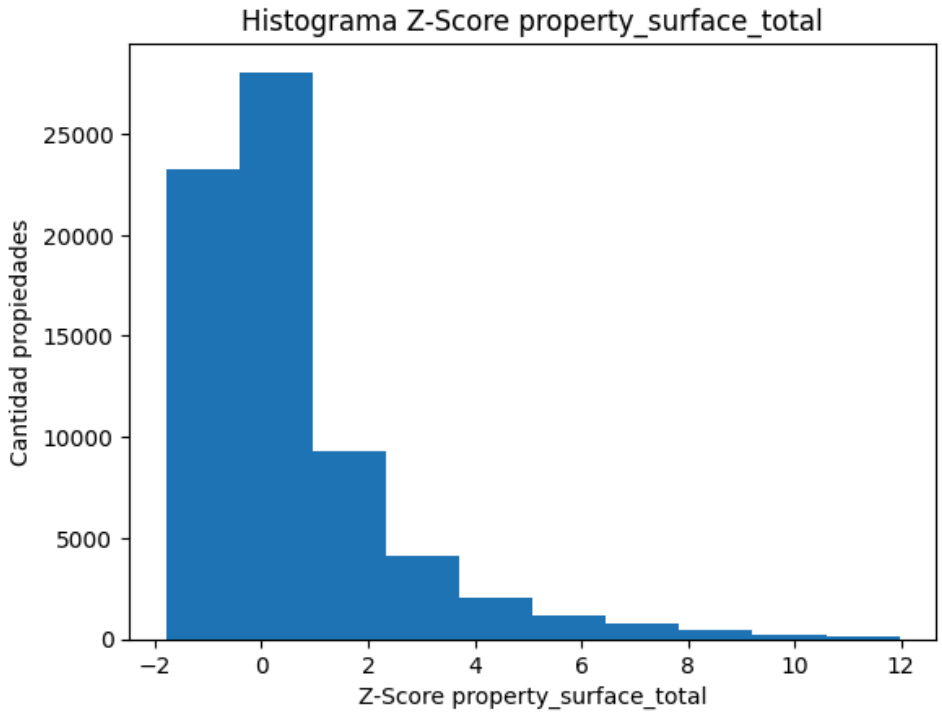
Luego nos encontramos con un porcentaje mucho más cercano a cero, pero de variables que están relacionadas entre sí y que además son muy importantes para el dataset. Estas columnas son *property_surface_total* y *property_surface_covered* que tienen un porcentaje de nulos de 5% y 3.5% respectivamente. El método que vamos a utilizar para rellenar las celdas vacías va a ser el de rellenar los datos a partir del **cociente** de la media de las 2 columnas.

$$\frac{media(property\_surface\_total)}{media(property\_surface\_covered)}$$

Antes de seguir, nos aseguramos que ninguna propiedad tenga ambos valores nulos y para cada celda vacía de la columna *property_surface_total* vamos a multiplicar su valor de *property_surface_covered* por el cociente. De forma análoga, hacemos lo mismo para los valores nulos de *property_surface_covered*, pero esta vez, dividiremos por el cociente el valor de *property_surface_total*.

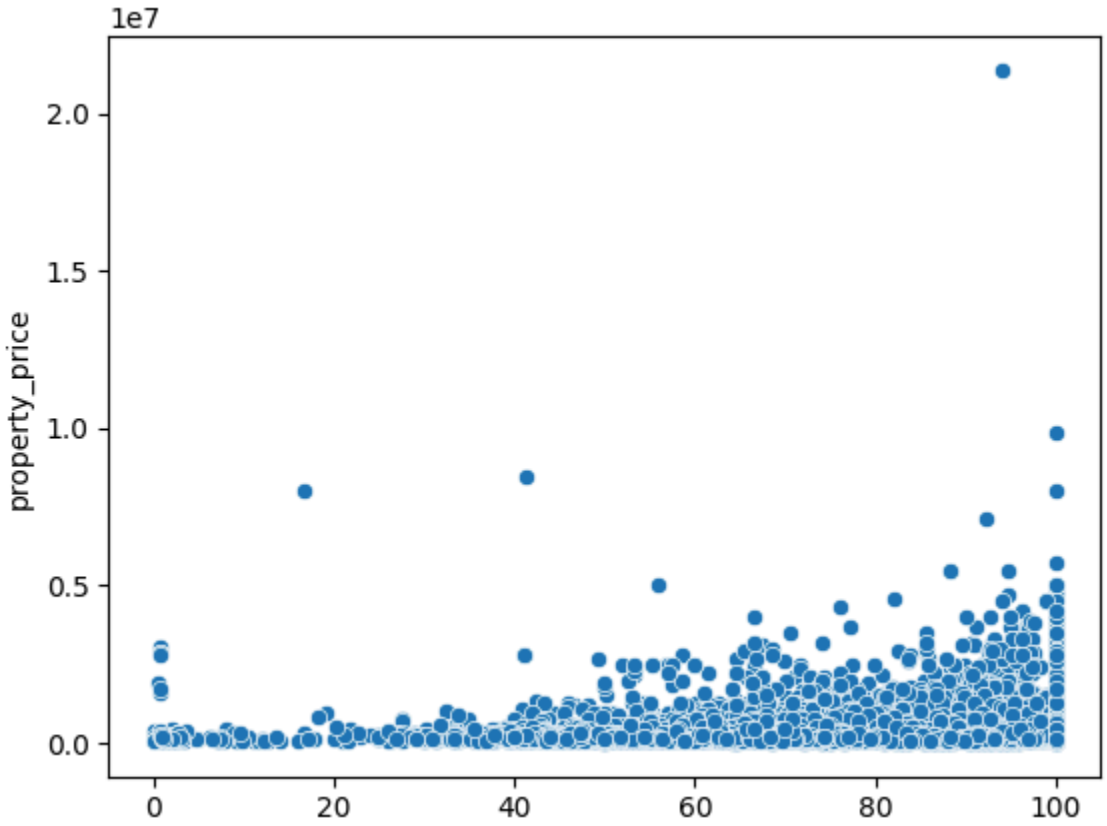
Más adelante en nuestro análisis, vemos que la columna de *id* no va a ser nunca utilizada y que la columna *created_on* es redundante ya que la columna *start_date* brinda la misma información.

A la hora de las visualizaciones fuimos encontrando valores que carecen de sentido, como propiedades con decenas de miles o centenas de miles de metros cuadrados de superficie. Dichas anomalías se detectaron fácilmente con un dispersograma para luego confirmar de manera más eficiente con el método z-score que observamos en el siguiente gráfico:



Visualizaciones

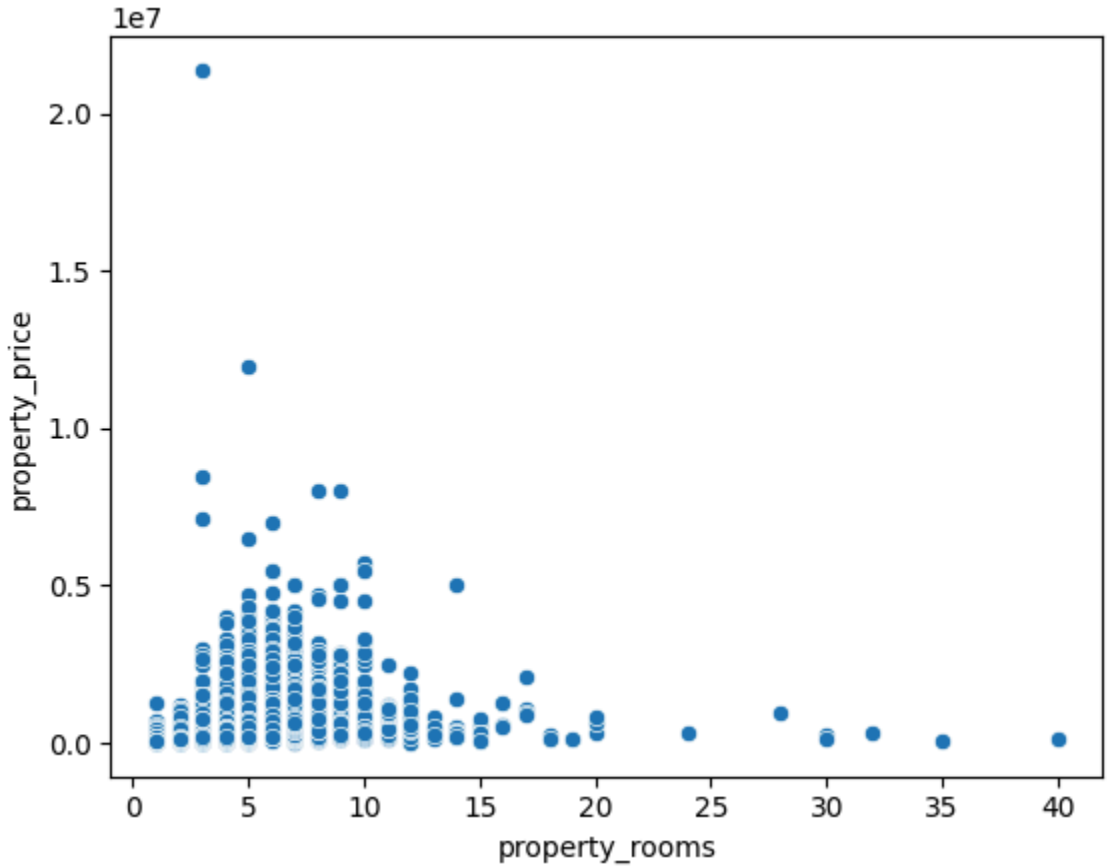
1) Precio de propiedad VS Porcentaje de superficie cubierta



En este gráfico se puede ver el precio de propiedad (en el eje Y) comparado con el porcentaje de superficie cubierta (en el eje X). Este porcentaje se calcula con la siguiente fórmula: $\frac{Superficie\ Cubierta}{Superficie\ Total} \cdot 100\%$

Nos pareció interesante ya que la relación entre superficie cubierta y no cubierta parece tener un muy ligero crecimiento cuadrático, lo cual puede ser útil para predecir precios.

2) Precio de propiedad VS Porcentaje de superficie cubierta



En este gráfico se puede ver el precio de una propiedad (en el eje Y) comparado con la cantidad de ambientes de la misma (en el eje X)
Nos pareció interesante ya que, si bien uno pensaría que cuanto más ambientes tiene una propiedad mayor el precio, este gráfico sugiere que la relación tiene una forma de campana de Gauss. Por lo tanto, este gráfico deja en evidencia que la cantidad de ambientes quizá no sea la mejor variable para predecir precios cuando la misma tiene valores altos.

Clustering

Para determinar la cantidad de cantidad apropiada de grupos se usa el método de codo.
Para esto se usa el algoritmo de Kmeans con un número de clusters (k) entre 2 y 10.
Se graficó el SSE en función de k y se buscó el punto de inflexión, dicho punto se da en k=2, anteriormente con el dataset parcialmente preprocesado nos daba 3.
También se graficaron la longitud y latitud en un mapa de la CABA, coloreados por cluster.

Clasificación

La alternativa elegida para construir la variable *tipo_precio* será la de tener en cuenta el tipo de propiedad para entender mejor la distribución de precios relativos. Hicimos un gráfico de la distribución de la variable en el mapa y notamos que, objetivamente, lo obtenido es bastante fiel a la realidad ya que los precios más elevados para las propiedades se encuentran desde el norte hacia el noreste en toda la periferia, algunos puntos del noroeste y cerca del centro geográfico de la ciudad.
En cuanto a la comparación con el gráfico obtenido a través de K-means, no podemos sacar una conclusión ya que la variable *tipo_precio* depende pura y exclusivamente de *pxm2*, haciendo que la clasificación esté "sesgada". A la hora de aplicar K-means se trabajó con más variables en vez de la relación directa entre superficie total y precio.
No llegamos a entrenar ninguno de los modelos.

Estado de Avance

1. Análisis Exploratorio y Preprocesamiento de Datos

Porcentaje de Avance: 100%/100%

Tareas en curso: -

Tareas planificadas: Ninguna, ya que creemos que esta parte está completa, pero esperamos posible correcciones de los profesores

Impedimentos: -

a) Exploración Inicial: -
b) Visualización de los datos: -
c) Datos Faltantes: -

d) Valores atípicos: -
e) Opcional: -

2. Agrupamiento

Porcentaje de Avance: 100%/100%

Tareas en curso: -

Tareas planificadas: Ninguna, ya que creemos que esta parte está completa, pero esperamos posible correcciones de los profesores

Impedimentos: -

3. Clasificación

Porcentaje de Avance: 10%/100%

Tareas en curso: Estamos iniciando los entrenamientos de modelo, definiendo los datasets de entrenamiento y prueba. No logramos avanzar en la lectura de métricas ya que en el punto anterior nos fue evidente que faltaba completar la parte de preprocesamiento para llegar a mejores resultados.

Tareas planificadas: optimización de hiperparámetros, profundización en el análisis de métricas y avance en general de las consignas

Impedimentos: Hasta ahora, el único impedimento fue la falta de un dataset limpio, con la menor cantidad de outliers para entender mejor los gráficos y variables.

4. Regresión

Porcentaje de Avance: 0%/100%

Tareas en curso: -

Tareas planificadas: Avanzar en general

Impedimentos: -

Tiempo dedicado

Indicar brevemente en qué tarea trabajó cada integrante del equipo durante estas semanas.Si trabajaron en las mismas tareas lo detallan en cada caso (como en el ejemplo el armado de reporte). Deben indicar el promedio de horas semanales que

dedicaron al trabajo práctico. En esta tabla solo deben incluir las tareas que realizaron luego de entregar el CHP1.

Integrante	Tarea	Prom. Hs Semana
Elvis Claros	Separación de datos, agrupamiento.	4
Ramiro Gareis	Exploración Inicial y Imputación de Datos	4
Gonzalo Olmos	Avance clasificación y regresión	5
Matías Venglar	Detección de Outliers Visualizaciones	5