Checkpoint 1 - Grupo 01

Análisis Exploratorio

El dataset en crudo que se nos otorga cuenta con más de 450 mil registros y 20 columnas.
Las características más destacables del mismo son:

- **property_price:** El precio de publicación es el dato que buscamos que nuestro modelo pueda predecir por lo que es necesario entrenarlo conociendo este dato. Es de tipo numérico.
- **property_currency:** Sirve para realizar el filtrado inicial. Es de tipo string. Una vez filtrado por este dato, la columna es descartable ya que tendría el mismo valor en todas las filas (“USD”).
- **operation:** Sirve para realizar el filtrado inicial. Es de tipo string. Una vez filtrado por este dato, la columna es descartable ya que tendría el mismo valor en todas las filas (“Venta”).
- **property_type:** Sirve para realizar el filtrado inicial. Determina qué es lo que está siendo publicado (Casa, local comercial, etc). Es de tipo string. Una vez filtrado nos quedan 3 tipos (Casa, PH, Departamento) los cuales pueden llegar a tener correlación con el precio.
- **Property_rooms** y **property_bedrooms:** Indican la cantidad de habitaciones y ambientes (respectivamente). Son de tipo numérico. Puede existir una relación entre estas cantidades y el precio de la publicación.
- **property_surface_total** y **property_surface_covered:** Indican los metros cuadrados del total de la propiedad y los cubiertos (respectivamente). Son de tipo numérico. Puede existir una relación entre estas cantidades y el precio de la publicación.

- **place_l2, place_l3, place_l4...:** Indican la ubicación de la propiedad. Sea provincia, localidad, barrio. Son de tipo string. Estos datos son de alto interés ya que los precios de las propiedades suelen estar altamente influenciados por la ubicación. Existe una grán proporción de datos nulos cuanto más específica es la ubicación.
- **property_title**: Indica el título con el cual se realizó la publicación. Este puede llegar a aportar información útil al momento de tomar decisiones sobre las otras columnas, ya que generalmente en cada registro aporta información sobre la cantidad de ambientes y/o locación. Es de tipo string.

Analizamos las columnas para filtrar según los criterios indicados buscando irregularidades que causen un filtrado menos preciso. Por ejemplo, que la columna **property_currency** (la cual tenemos que filtrar por el valor “USD”) no tenga algunos registros en mayúscula y otros en minúsculas, o no diga “Dólares” en lugar de “USD‘ y por ello no sea captada por el filtro. Ninguna de las columnas que son criterio de filtrado presentó problemas de este tipo.

Al filtrar obtenemos un dataframe de más de 94 mil registros. Este se divide en conjunto de entrenamiento y prueba en una proporción 80/20, teniendo como resultado 75 mil registros de entrenamiento y casi 19 mil de prueba.

Todo el análisis se realiza sobre el conjunto de entrenamiento.

Hipótesis:

- Las casas tienden a tener 1 ambiente más que el total de habitaciones (el living).
- Las coordenadas (latitud y longitud) son mucho más propensas a presentar un grave error de precisión a comparación del barrio.
- La superficie cubierta es menor o igual a la superficie total.
- Las propiedades con un exagerado número de habitaciones (mansiones por ejemplo) son un mercado paralelo y no rigen los mismos criterios que para la propiedad promedio.

Preprocesamiento de Datos

Tras filtrar el dataset según los criterios indicados, nuestro dataset presenta ciertas características:

1. **Columnas eliminadas:**
    a. place_l4: +96% nulls
    b. place_l5: 100% nulls
    c. place_l6: 100% nulls
    d. start_date: No aporta información
    e. end_date: No aporta información

    f. created_on: No aporta información
    g. property_currency: Todos los registros tienen el mismo valor
    h. operation: Todos los registros tienen el mismo valor

2. **Correlaciones**:
    a. Cantidad de habitaciones y cantidad de ambientes: 0.87. Esto confirma nuestra hipótesis sobre la tendencia de las propiedades de tener 1 ambiente más que la cantidad de habitaciones.
    b. Superficie cubierta y superficie total por tipo de propiedad:
    En PH's y casas la correlación es mayor a 0.9. En departamentos ronda el 0.55.
    En casi todos los barrios los departamentos tienen correlación +0.8, lo cual tiene mucho sentido ya que generalmente un departamento es en su mayoría superficie cubierta.
    Este dato nos es útil para imputar las filas donde uno de los datos es nulo y el otro no.

3. **Nuevas features**:
    a. total_m2_price: Precio por metro cuadrado de superficie total.
    b. covered_m2_price: Precio por metro cuadrado de superficie cubierta.
    c. Casa, departament y PH: Aplicando One Hot Encoding reemplazamos la columna "property_type" por estas 3 nuevas columnas dummies
    d. Palermo, Belgrano, Caballito, Otros: Aplicando One Hot Encoding reemplazar la columna "place_l3" por dummies de las ubicaciones con mayor presencia. Útil para el análisis de grupos.

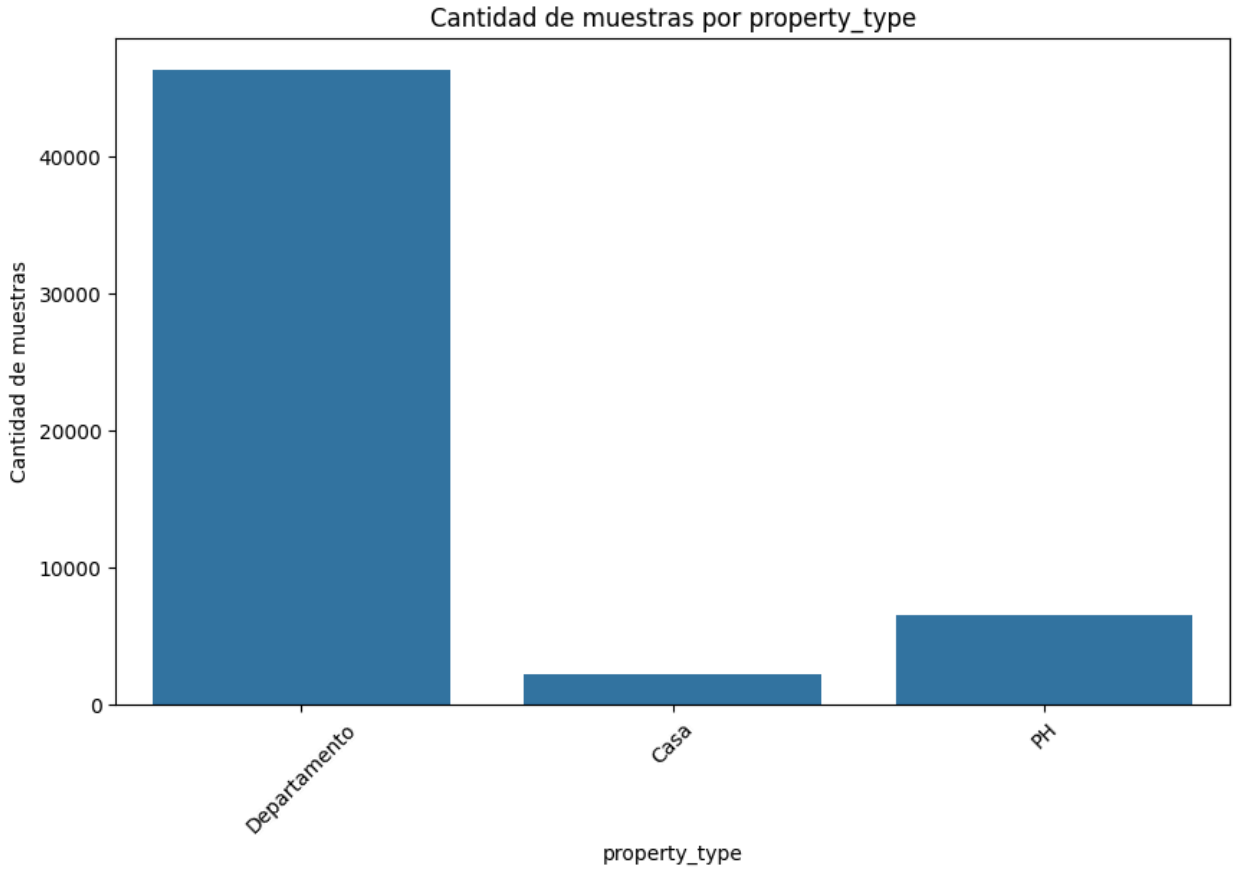
4. **Valores atípicos:**
    a. Superficie total menor a superficie cubierta: Este es un absurdo en el cual optamos por invertir los valores, asignando la superficie cubierta

como superficie total y viceversa. Esto incrementó en grán medida la correlación entre estos datos para propiedades del tipo Casa.

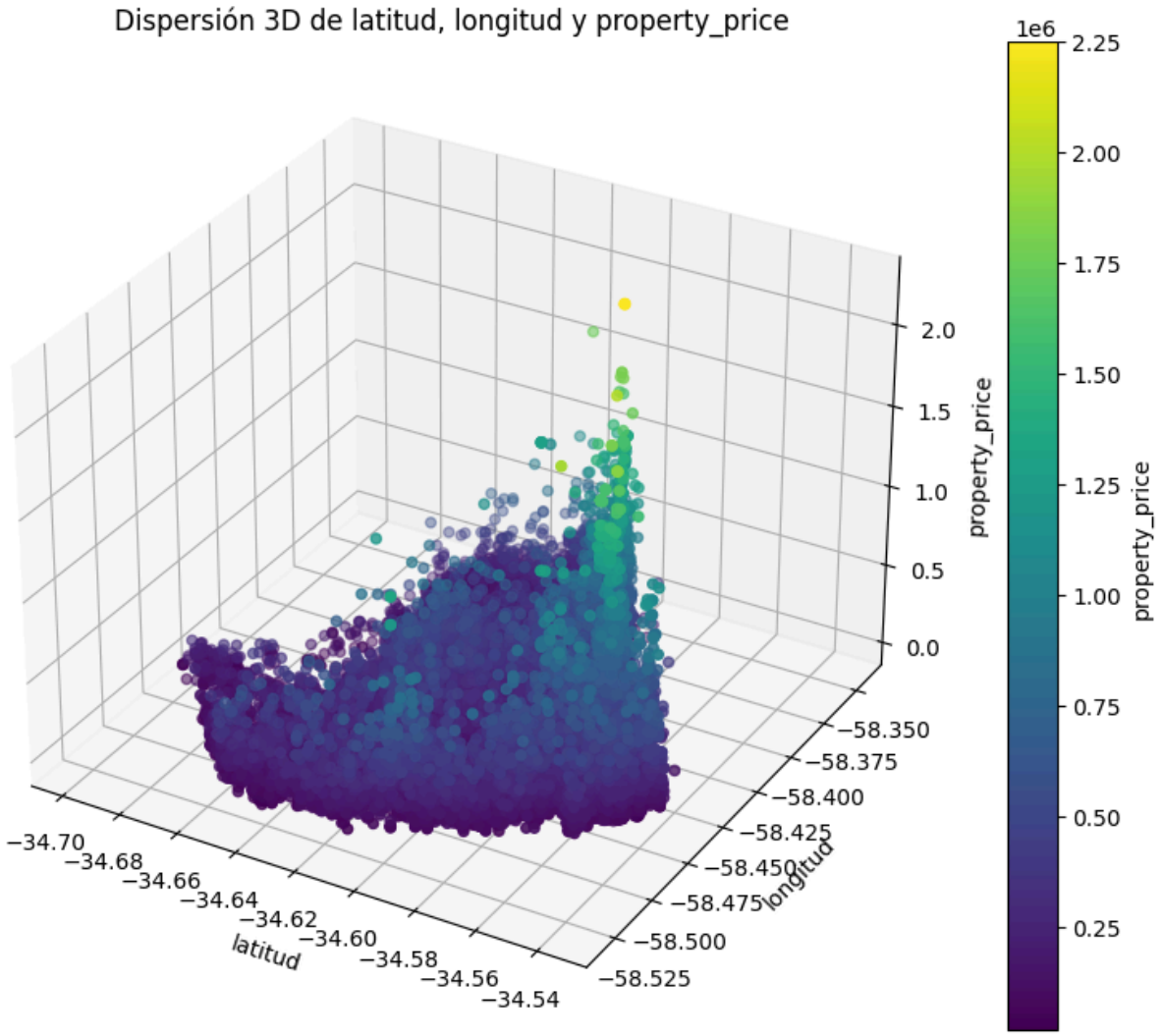
b. Superficie total exageradamente grande: Haciendo un gráfico de dispersión entre la superficie total y la cubierta se ve un claro cúmulo en una esquina del gráfico. Eliminamos las publicaciones cuya superficie total es desde 7 veces mayor a la mediana.
c. Precio por metro cuadrado:
   En algunas propiedades obtuvimos valores por metro cuadrado irrisorios, de más de 5 veces la mediana inclusive. Las eliminamos.

d. Cantidad de habitaciones y de ambientes:
   En aquellos casos donde existen más habitaciones que ambientes asignamos el valor de ambientes al de habitación.
   Eliminamos aquellos registros donde la cantidad de habitaciones es más de 7 veces la mediana.

Visualizaciones



En el gráfico visualizamos la cantidad de muestras por “propety_type” la cual nos permite apreciar que en nuestro dataset poseemos una mayor distribución de Departamentos, lo cual es lo que uno esperaría de Capital Federal. Se eligió mostrar este gráfico ya que nos permite observar si un dato predomina sobre los otros y tenerlo más en cuenta por su peso



El gráfico permite visualizar la dispersión dimensional de las publicaciones gracias a los datos de latitud y longitud, así como también. Vemos un cúmulo importante de de datos con un amplio rango de precios en el sector de color más claro del gráfico. Mientras ciertas zonas mantienen un rango relativamente reducido en sus precios (eje vertical), este cúmulo más claro indica que esa ubicación (Palermo, seguido por Belgrano y Caballito) no tiene una fuerte tendencia del precio basado únicamente en su ubicación ya que el abanico de valores es más amplio.

Clustering

Buscamos con el SSE que cantidad de clusters sería apropiado para el dataframe dado, y hemos llegado a la respuesta de que el mayor silhouette score es para dos grupos. También con los gráficos observamos que el criterio que utilizó el algoritmo de KMeans para separar dichos clusters está relacionado al valor de la vivienda.

Estado de Avance

1. Análisis Exploratorio y Preprocesamiento de Datos

Porcentaje de Avance: 100%/100%

Hemos terminado con esta etapa, lo único que dejamos con cierta duda, es respecto a aquellas filas que tenían un *place_l3* no nulo, pero tanto latitud como longitud sí lo eran. Probaremos cómo funcionan los modelos, y en caso de necesitarlo, buscaremos ubicar a estas propiedades en los centros de los polígonos del barrio correspondiente al *place_l3*.

2. Agrupamiento

Porcentaje de Avance: 100%/100%

Tiempo dedicado

Integrante	Tarea	Prom. Hs Semana
Taiel Molina	Exploración De Datos Visualización de Datos Imputación de Datos Agrupamiento	7
Ignacio Fernández	Análisis de Correlaciones Armado de Reporte	7

	Análisis de outliers Imputación de Datos	
Sebastián Vera	Visualización de Datos Armado de Reporte	4