Checkpoint 1 - Grupo 04

Análisis Exploratorio

El dataset con el que vamos a trabajar es de la compañía Properati, una empresa inmobiliaria fundada en Argentina. Utilizaremos los datos públicos que la empresa brinda en cuanto a ventas de inmuebles, más específicamente en Argentina durante el año 2021. Inicialmente el dataset cuenta con 20 columnas y 460154.

El primer filtrado de información lo realizaremos sobre el DS (dataset) en su totalidad, ya que solo nos interesan, para la finalidad de este trabajo, las propiedades que entren dentro de la categoría (departamento, PH o Casa), que se encuentren dentro de la Capital Federal, con un precio en dolares y que estén catalogados como ventas.

Los features más destacables pueden ser, de nuevo, por la importancia que presentan para nuestro análisis, el **tipo de propiedad**, el **place_l2** (representa si la propiedad está en capital federal o no para nosotros), el **tipo de operación** y por último el **property_currency** que nos va a dar la información sobre si la propiedad se vende en dolares o en pesos. A esta altura del análisis de los datos no realizamos ninguna suposición, simplemente hicimos un filtrado inicial de los datos.

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)**

Se eliminaron las columnas '**place_l2**', '**operation**' y '**property_currency**' porque al haber hecho una filtración inicial de propiedades estas quedaron con la misma información en todas las filas por lo que era información redundante.

2. **¿Detectaron correlaciones interesantes (entre qué variables y qué coeficiente)?**

Las correlaciones más interesantes que encontramos fueron:

- '**property_price**' y '**property_surface_total**' con un coeficiente de **0.08523491719899819**
- '**property_price**' y '**property_rooms**' con un coeficiente de **0.488934080178357**
- '**property_price**' y '**property_surface_covered**'' con un coeficiente de **0.05623785863824733**

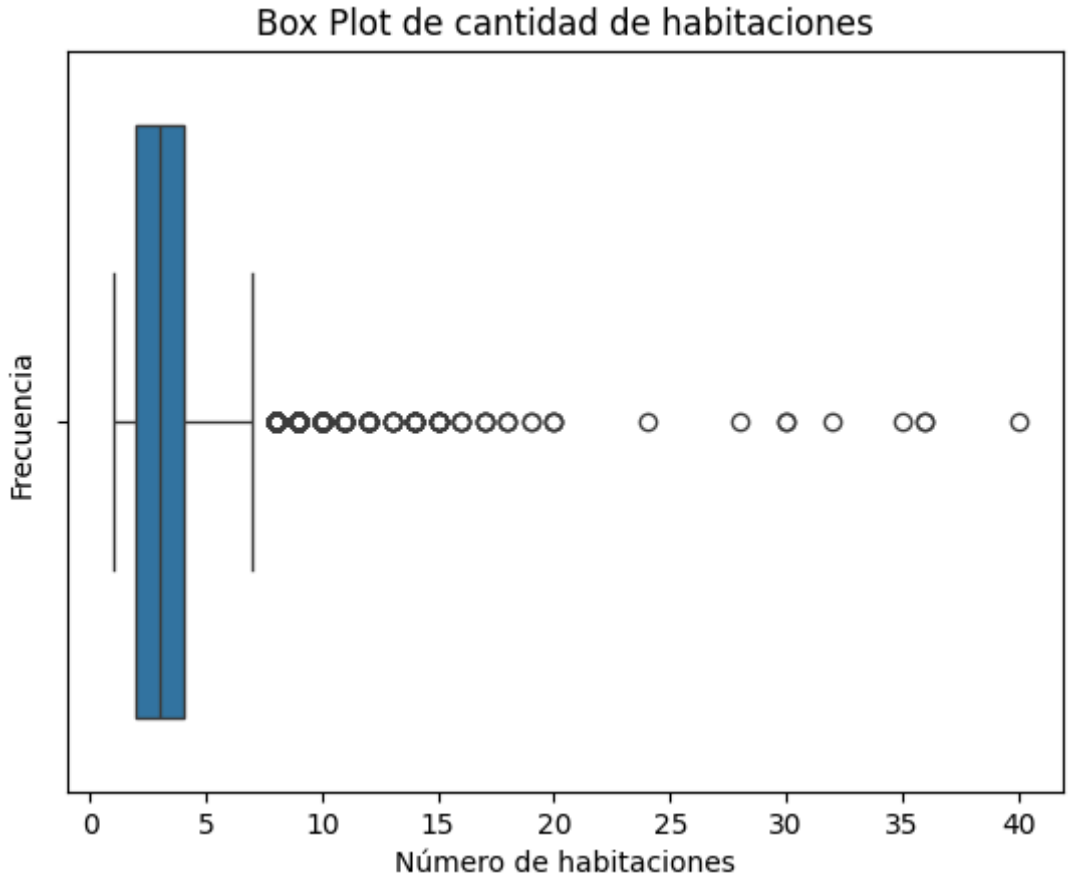
Si bien estas correlaciones podrían indicar una falta de relación entre estas variables, creemos que si hay una relación, y por ende suponemos que es debido a los outliers de los diferentes coeficientes. Una vez hecho el tratado de outliers, volveremos a calcular estas correlaciones.

3. ¿Generaron nuevos features?

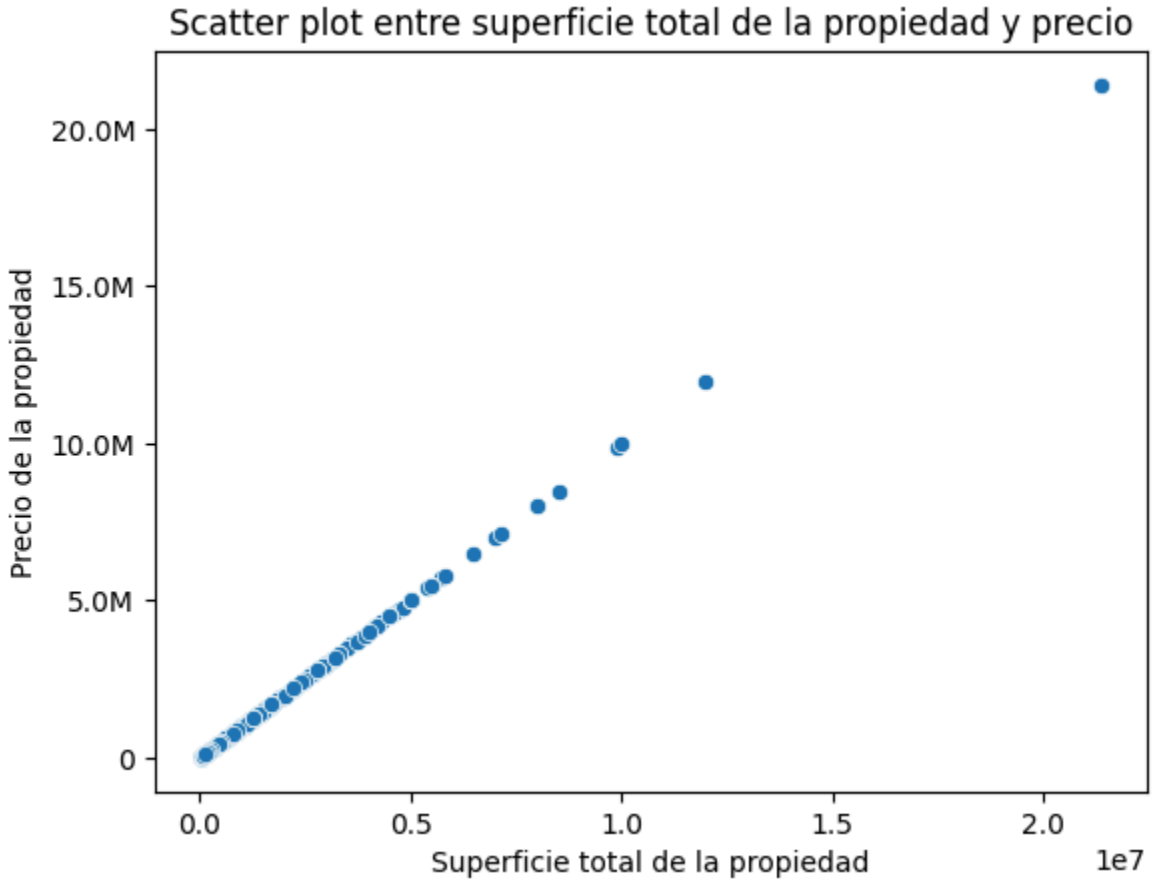
No generamos nuevos features con respecto al DS, al menos no encontramos qué features nuevos inicializar.

4. ¿Encontraron valores atípicos? ¿Cuáles? ¿Qué técnicas utilizaron y qué decisiones tomaron?

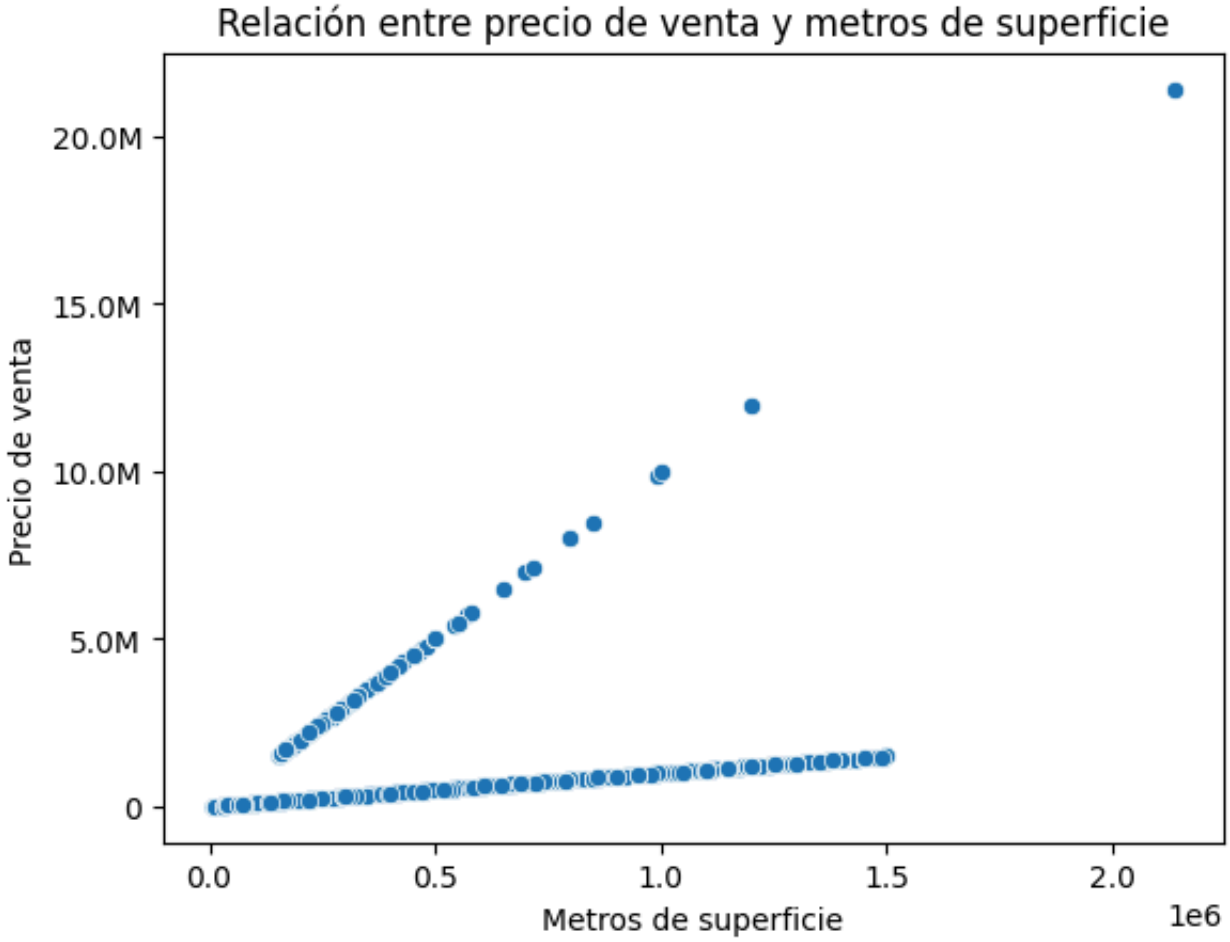
Encontramos valores atípicos tanto para variables univariadas como multivariadas. En el caso de las univariadas, con un gráfico de boxplot podemos observar algunos valores atípicos para la columna de 'property_rooms:



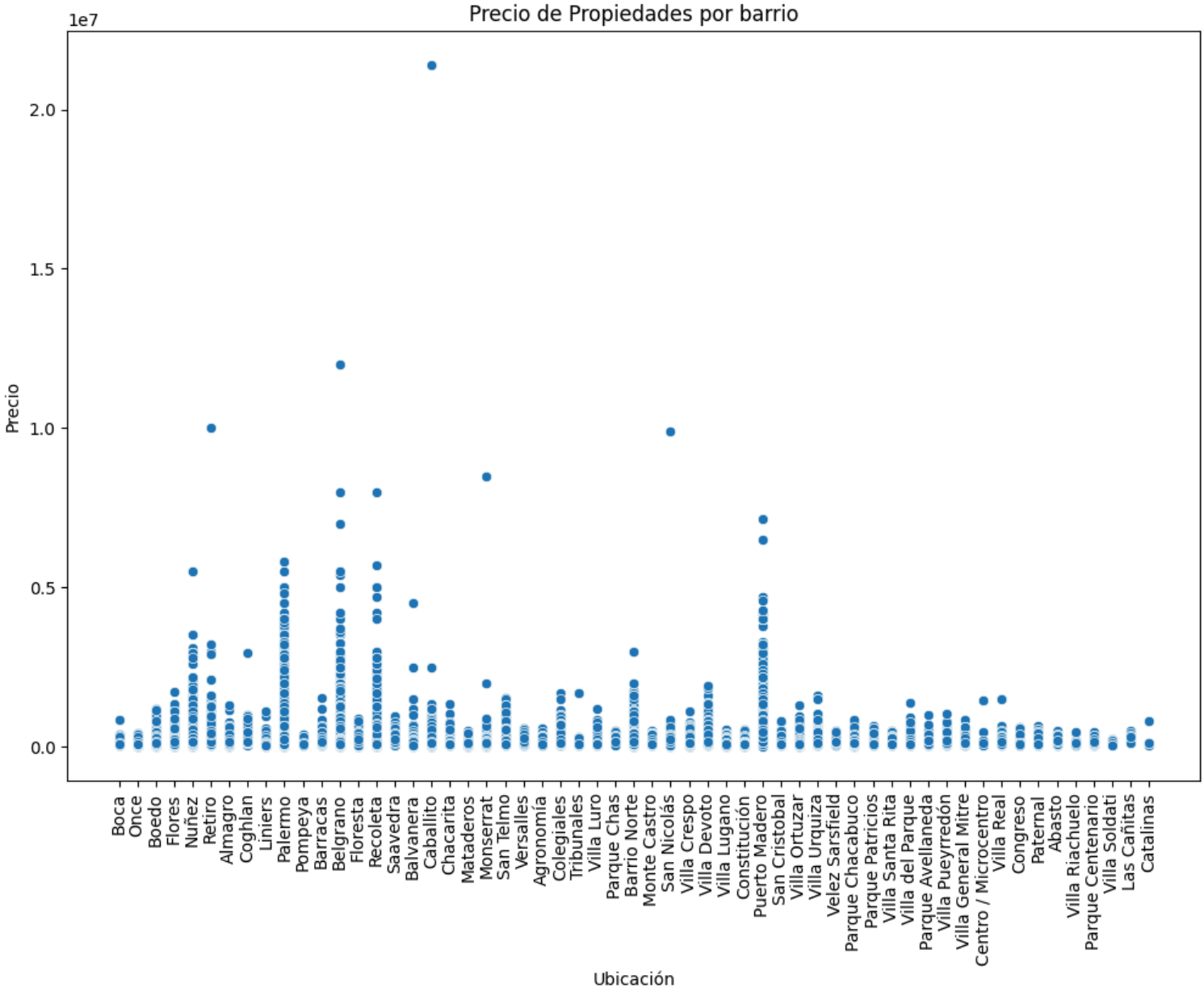
En el caso de las variables multivariadas, un scatterplot nos muestra outliers entre la superficie total de las propiedades y los precios:



Aquí podemos observar lo dicho en la sección de correlaciones. Los outliers quizá causan una disparidad entre las correlaciones de estos 2 features, a pesar de que la lógica nos indica que están altamente relacionados. Una vez tratados los outliers, el gráfico quedó así:

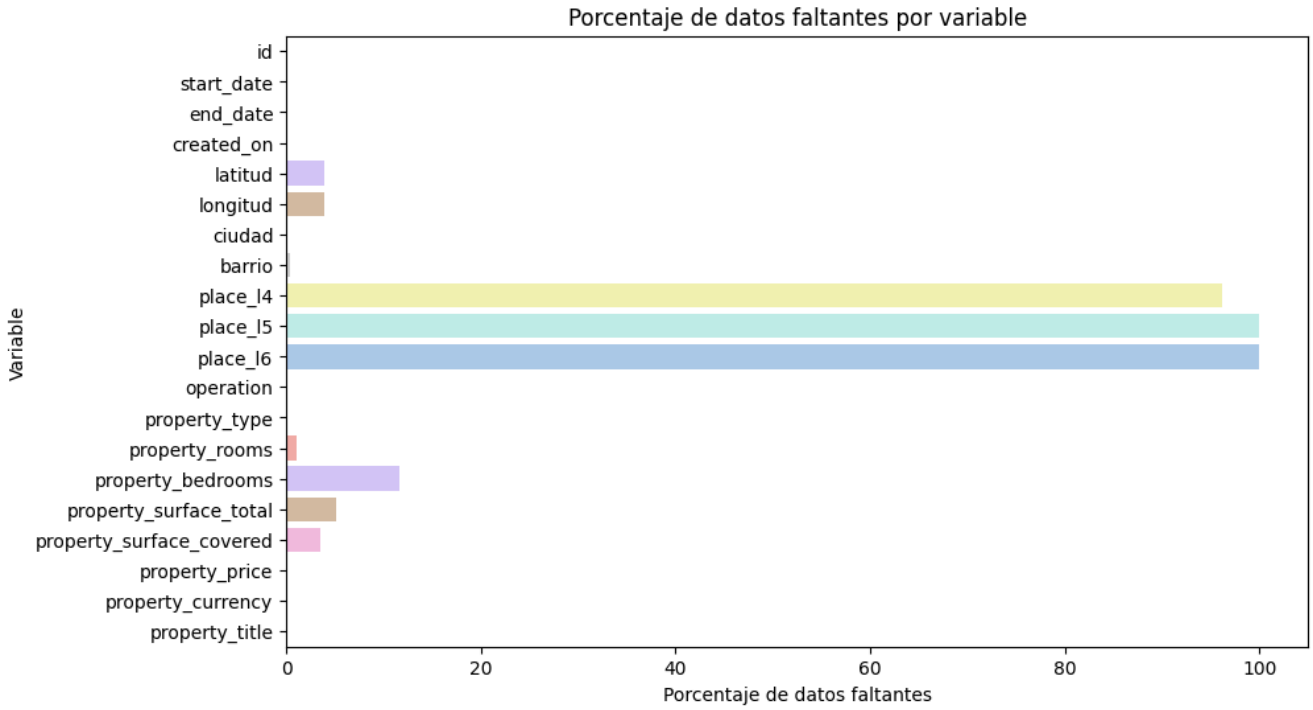

Relación entre precio de venta y metros de superficie
Precio de venta
20.0M
15.0M
10.0M
5.0M
0
0.0
0.5
1.0
1.5
2.0
Metros de superficie
1e6

Si queremos observar los precios de las propiedades según el barrio en el que se encuentran, podemos observar algunos valores atípicos aquí también:



5. ¿Qué columnas tenían datos faltantes? ¿En qué proporción? ¿Qué se hizo con estos registros?

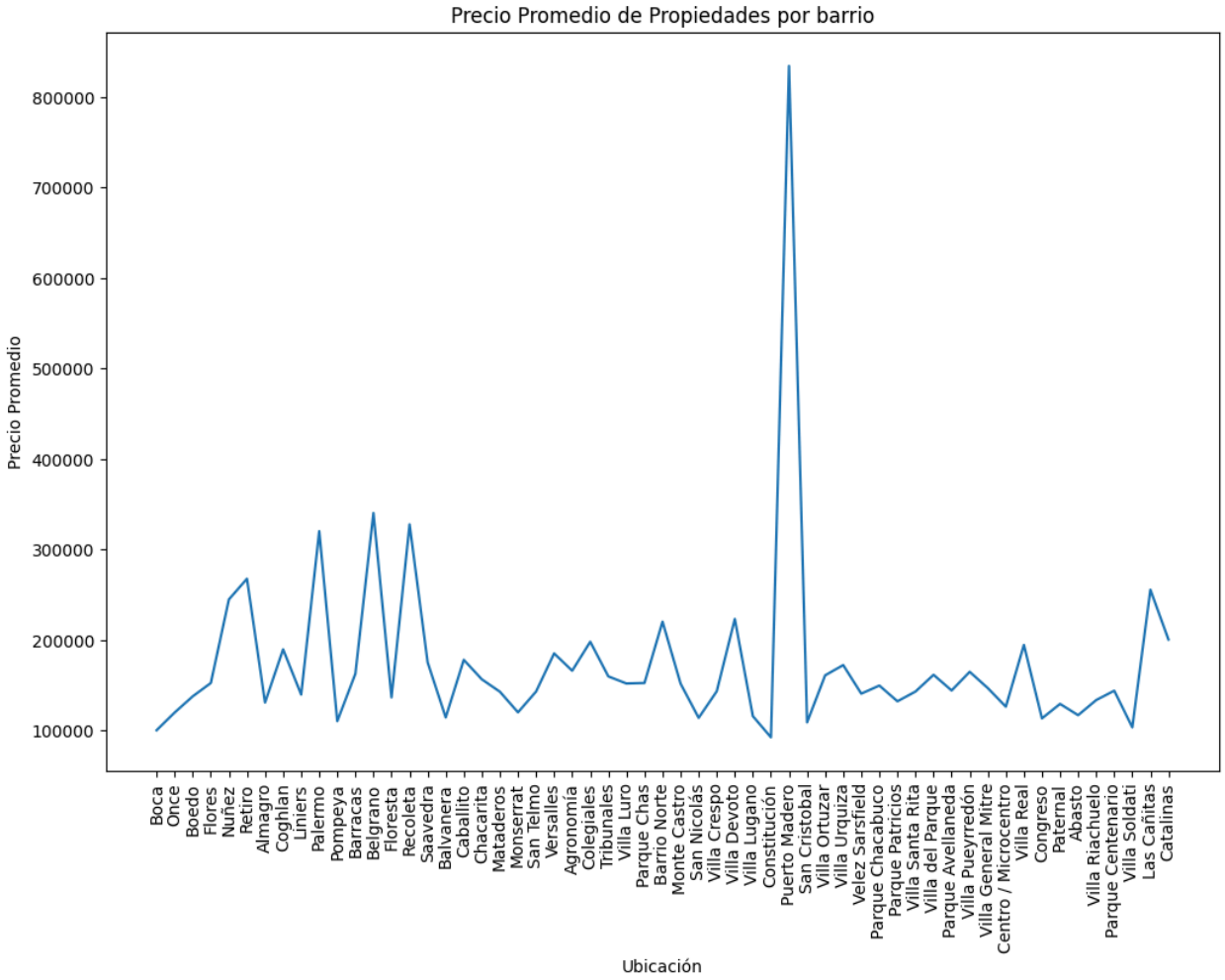
Con este gráfico podemos observar todas las columnas de interés, y sus respectivos datos faltantes:



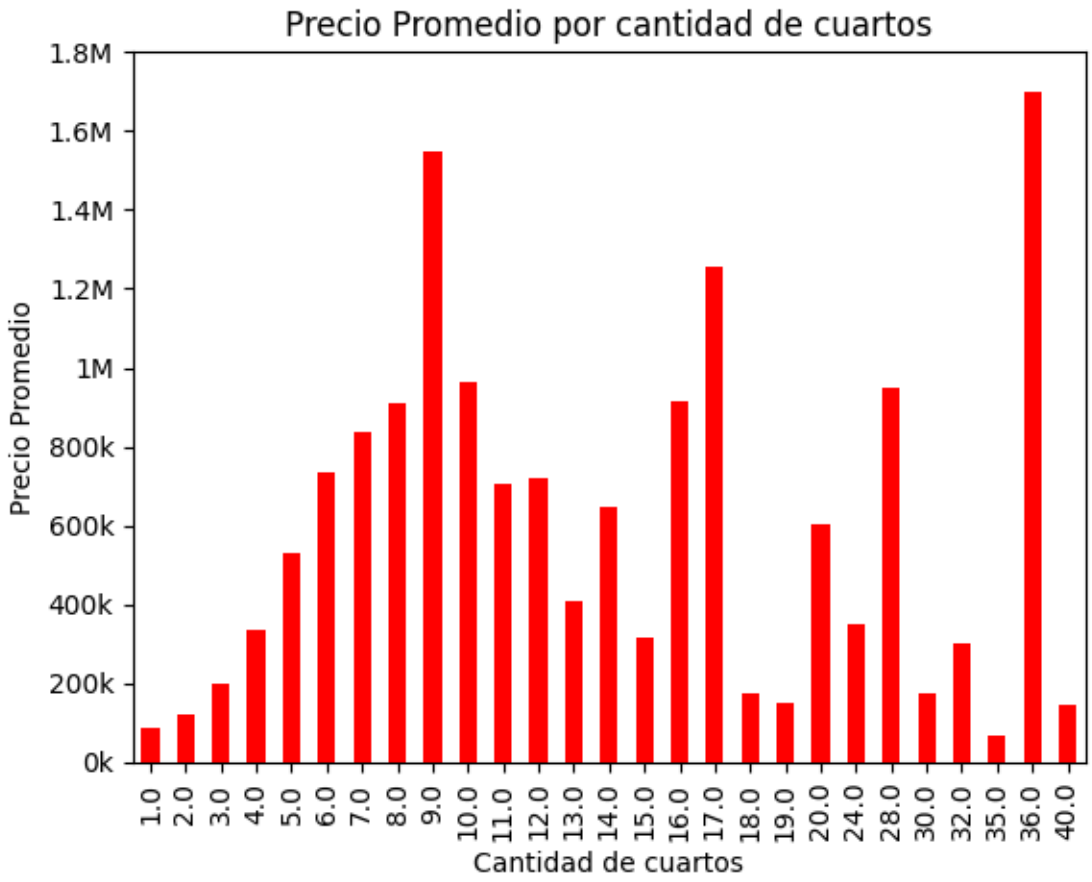
Los features cómo "**place_l4**", "**place_l5**", "**place_l6**" los consideramos con demasiados datos faltantes como para poder ser tratados y utilizados, de esta forma, decidimos removerlos del dataset y trabajar sin ellos, ya que, en caso de solucionar estos problemas de alguna forma, como por ejemplo, promediando con aquellos datos que poseemos, obtendremos resultados que no consideramos reales o demostrativos del dominio con el que estamos trabajando.

Para el caso de la **'longitud'** y la **'latitud'**, decidimos utilizar el dato del barrio en el que se encuentran esas propiedades, y llenar esos datos con la información aproximada de dónde pueden llegar a encontrarse, basándonos en las latitudes y longitudes de otras propiedades que se encuentran en esos mismos barrios.

Visualizaciones



En este gráfico podemos observar el precio de las distintas propiedades según el barrio en donde se ubican, como era de esperarse las propiedades ubicadas más al norte de CABA son más costosas de las que se encuentran en el sur de la capital.



En este otro gráfico observamos la relación precio y cantidad de cuartos. Por un lado, llegamos a la conclusión de que en general cuantos más cuartos tengan más caras serán esas propiedades con algunas excepciones que pueden deberse a la ubicación de dichas propiedades. Por otro lado, también se puede deducir la presencia de algunos datos mal ingresados ya que por ejemplo en nuestro gráfico aparece un promedio de que una propiedad con 40 habitaciones sale menos de 200k USD, en ese caso nos podemos imaginar que a la persona que cargó ese dato se le escapó un 0 y esa propiedad en realidad tiene 4 cuartos.

Estado de Avance

1. Análisis Exploratorio y Preprocesamiento de Datos

Porcentaje de Avance: XX%/100%

Tareas en curso: -.

Tareas planificadas: Analizar outliers en precios de propiedades por barrio.

Impedimentos: Crear buenos índices en los gráficos.

a) Exploración Inicial: -.

b) Visualización de los datos: -.

c) Datos Faltantes: -.

d) Valores atípicos: -.

2. Agrupamiento

Porcentaje de Avance: XX%/100%

Tareas en curso: Clustering.

Tareas planificadas: elegir y determinar cuántos grupos hay que hacer y su respectivo criterio.

Impedimentos: -.

Tiempo dedicado

Integrante	Tarea	Prom. Hs Semana
Catalina Basso	Filtrado inicial. Análisis de variables cualitativas y cuantitativas. Variables Irrelevantes. Visualización de datos. Análisis, decisión y ejecución sobre datos faltantes a nivel columna y fila. Informe.	7
Lucas Ruiz	Gráficos de las distribuciones de las variables más importantes (antes y después de ser tratadas). Correlaciones. Visualizaciones. Visualización de datos faltantes a nivel columna y fila. Análisis y visualización de valores atípicos(univariados y multivariados). Ejecución de la decisión sobre cómo tratar los outliers. Análisis de la relación entre el precio de venta y los metros de superficie.	7
Cristobal Alvarez	Análisis de outliers. Tratado de outliers. Gráficos de variables. Armado de informe.	3
Dalmiro Vilaplana	Análisis de correlaciones. Visualización de datos.	3

	Gráfico de distribución de las variables importantes. Análisis de outliers.	