# 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**

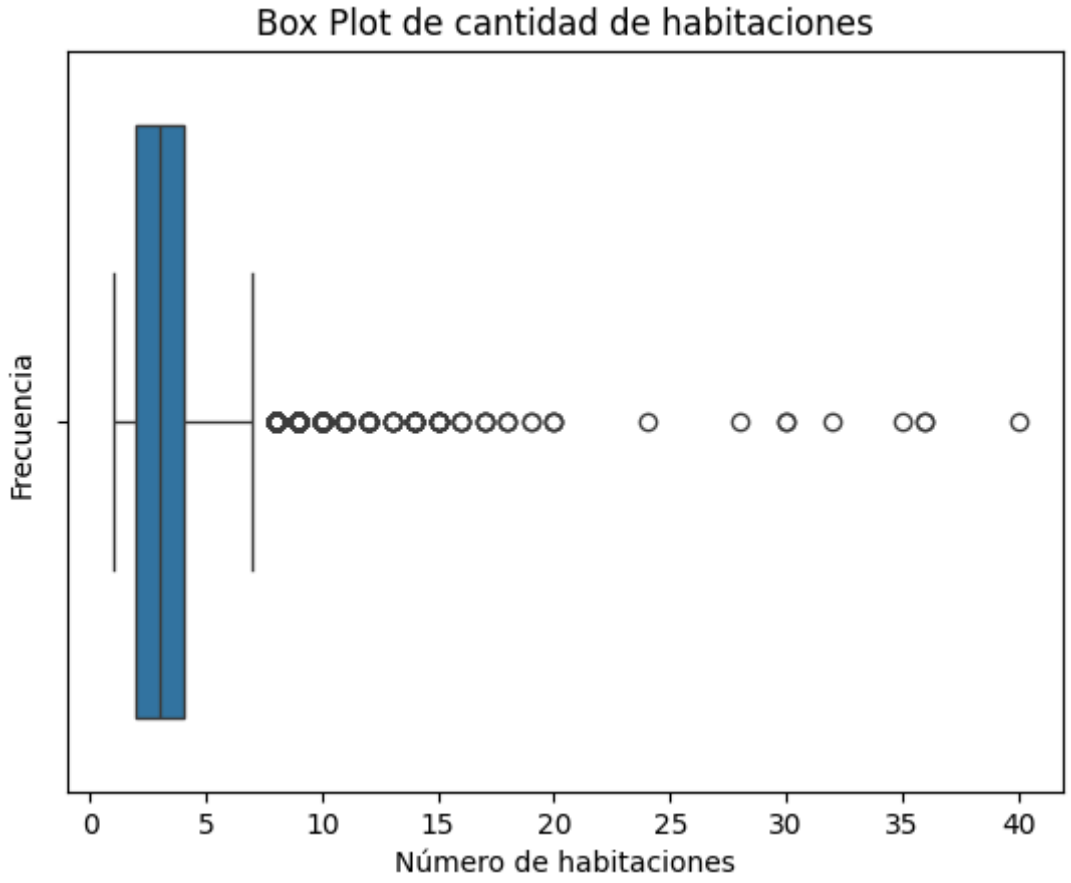
Si bien estas correlaciones podrían indicar una falta de relación entre estas variables, creemos que si hay una relación, y por ende suponemos que es debido a los outliers de los diferentes coeficientes. Una vez hecho el tratado de outliers, volveremos a calcular estas correlaciones.

**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:

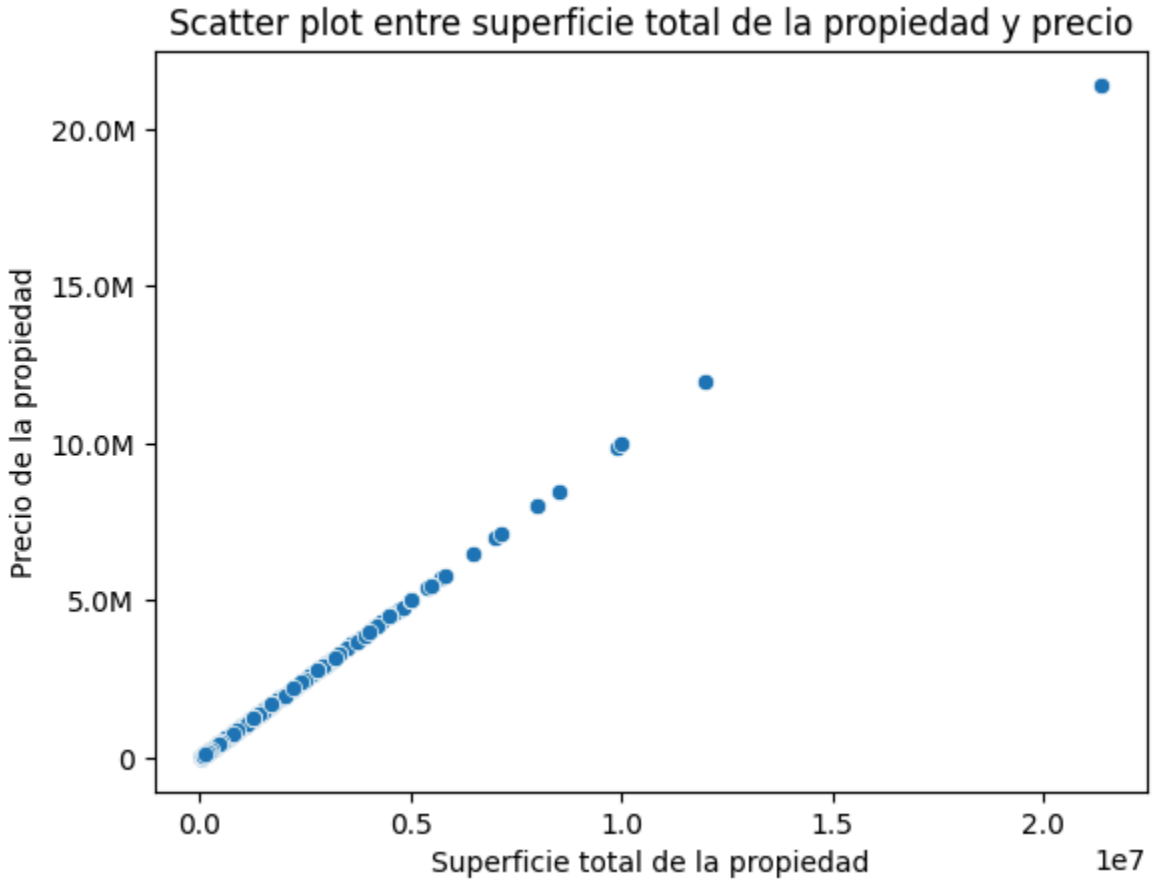


También utilizamos el método Z-score para detectar outliers en las columnas "`property_rooms`" y "`property_bedrooms`". En ambos casos se probó con distintos valores para el umbral, hasta llegar a uno que nos devolvió valores realmente alejados de la media y que no son coherentes con lo que representan.

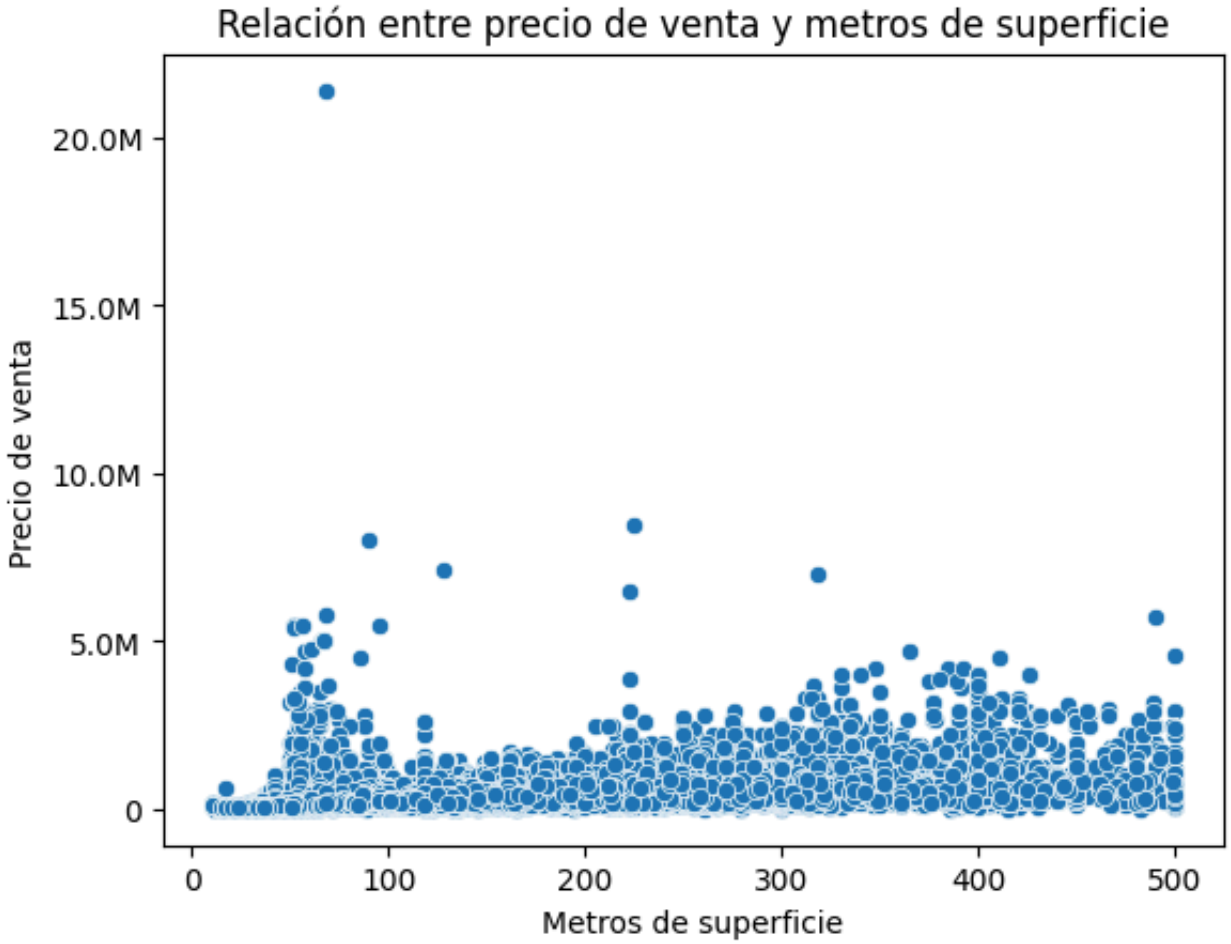
En el caso de la columna "`property_rooms`" se utilizó un umbral de 9, que nos devolvió 22 valores que rondan entre 16 y 40. Los cuales son demasiado altos para representar la cantidad de habitaciones en una propiedad.

En el caso de la columna "`property_bedrooms`" se utilizó un umbral de 5, que nos devolvió 224 valores que rondan entre 9 y 39. En este caso nos devolvió una cantidad de valores mayor que en la columna anterior y con un rango mayor. Pero consideramos que estos valores son atípicos ya que es raro encontrar propiedades con 9 o más dormitorios.

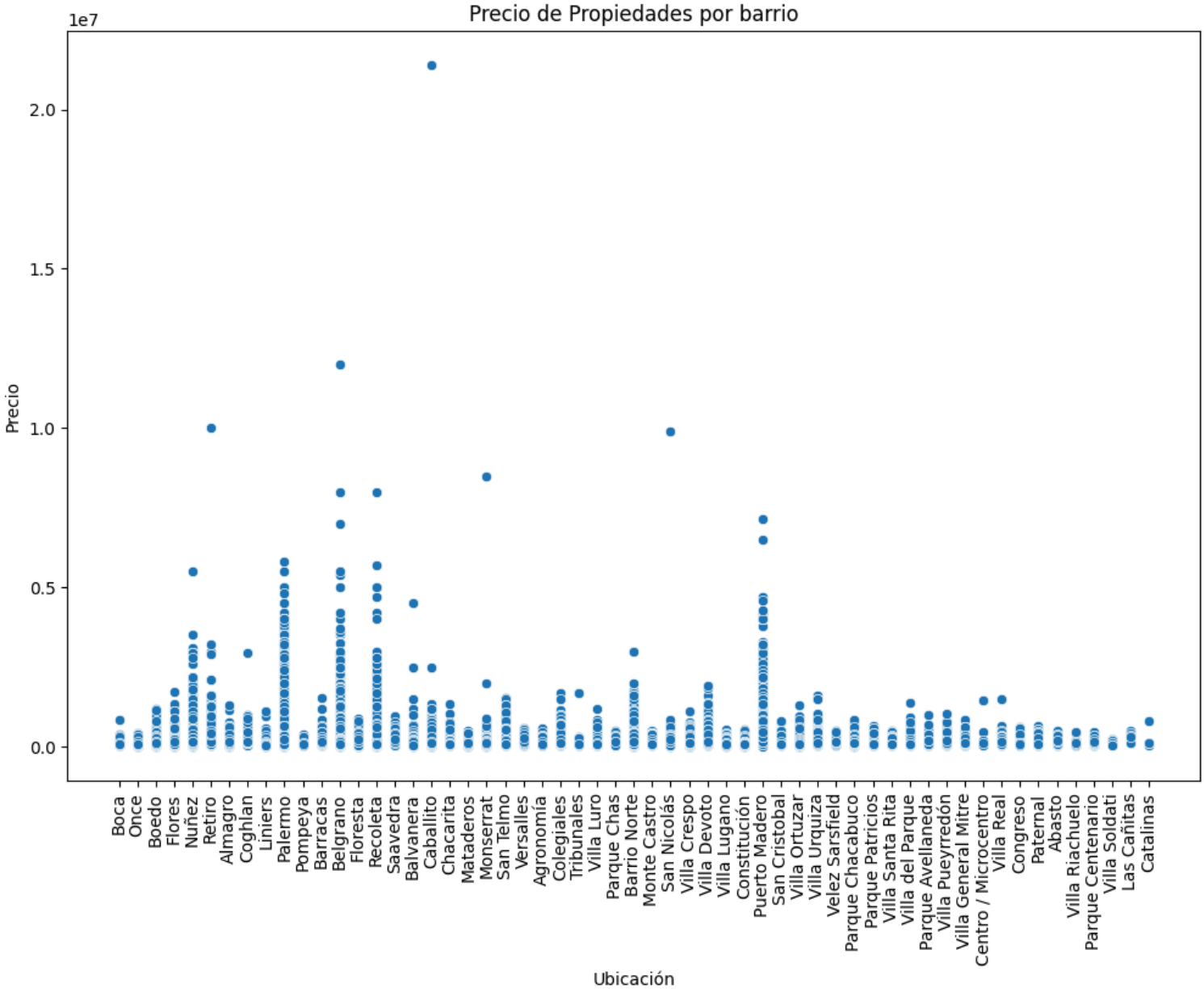
En el caso de las variables multivariadas, un scatterplot nos muestra outliers entre la superficie total de las propiedades y los precios:




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:

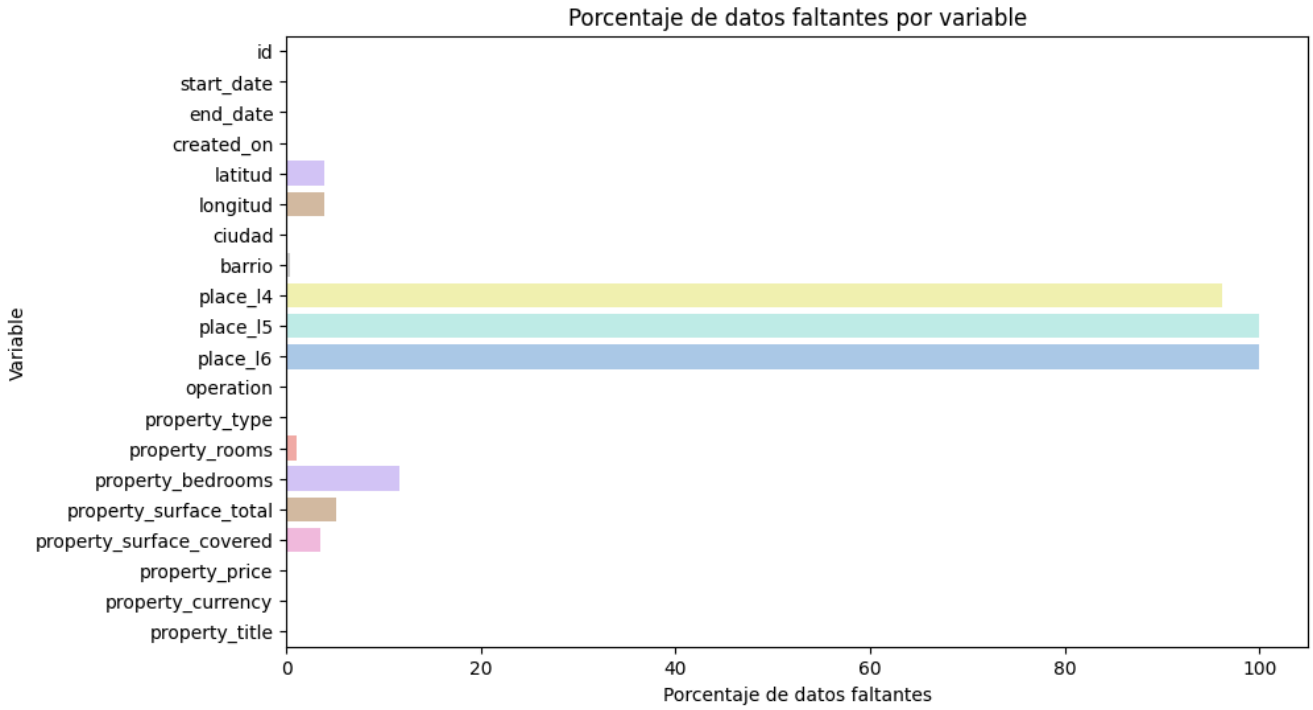

Relación entre precio de venta y metros de superficie
Precio de venta
20.0M
15.0M
10.0M
5.0M
0
0
100
200
300
400
500
Metros de superficie

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.

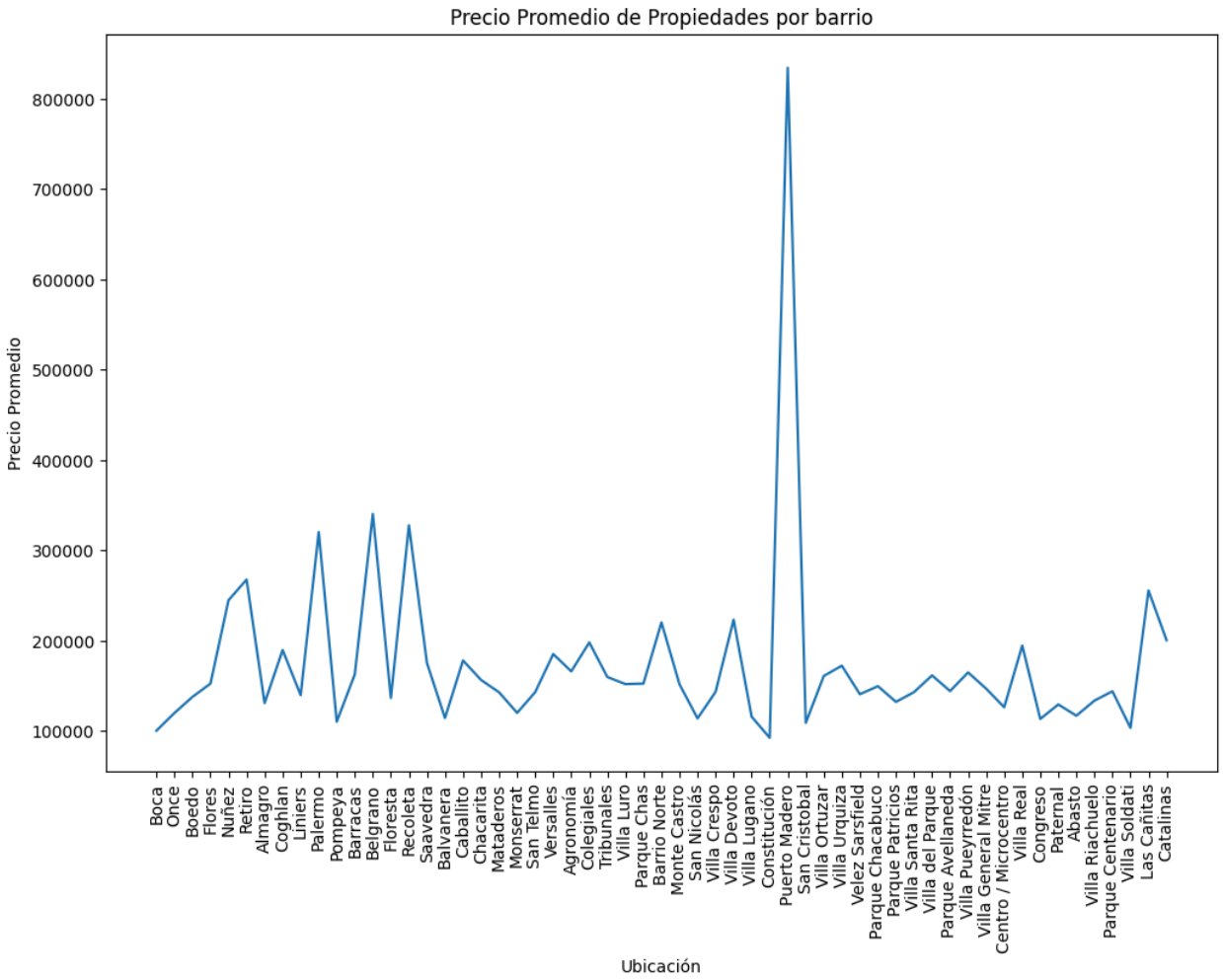
Para las propiedades que tienen la coordenada **'barrio'** vacía, buscamos propiedades cercanas y les copiamos el barrio.

En cuanto a los valores faltantes en la columna **‘rooms’**, completamos con el promedio de cantidad de habitaciones en las propiedades que sí tienen esta información.

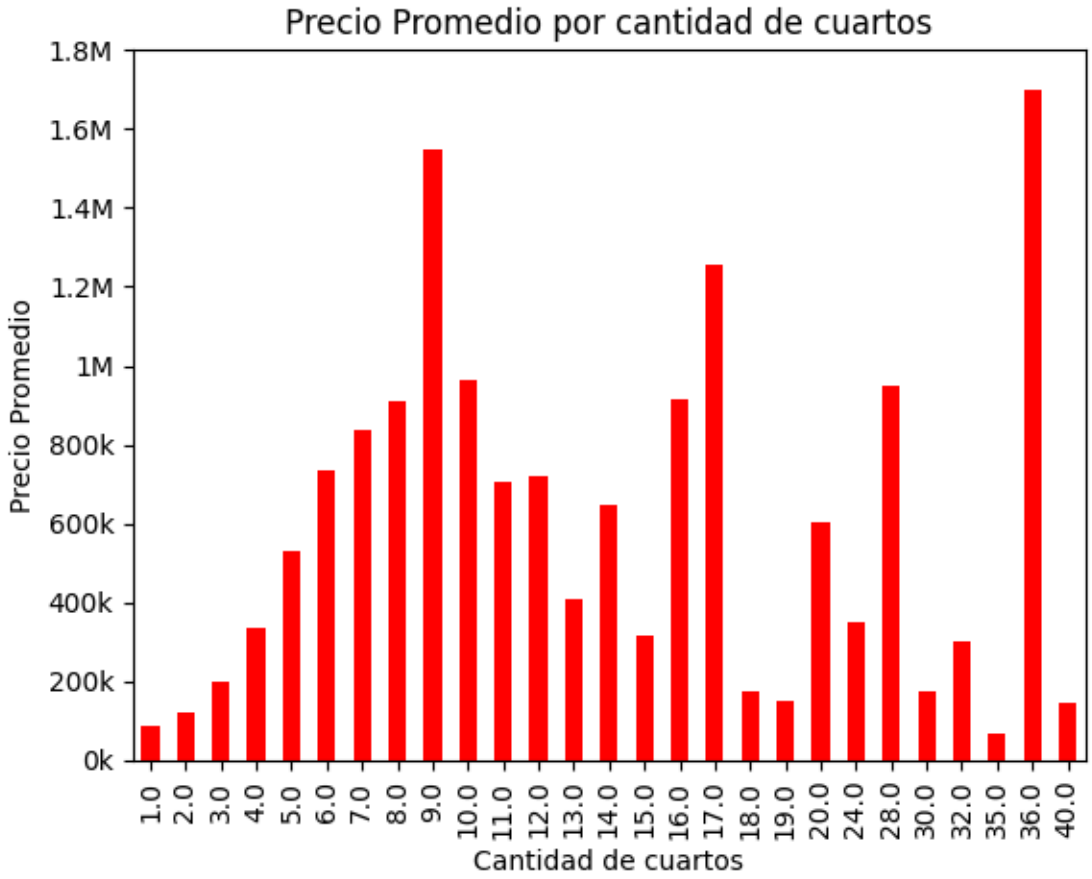
Para la columna de **‘property_bedrooms’**, completamos los valores faltantes con la cantidad de cuartos que tiene esa propiedad - 1. Ya que, en general, en una propiedad se cuentan como ambientes todas las habitaciones más el living.

Por último, completamos los valores faltantes de los atributos **'surface_total'** y **'surface_covered'** usando la técnica de imputación por vecinos más cercanos(KNN). Usamos el valor de vecinos por defecto (n = 5).

## 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.

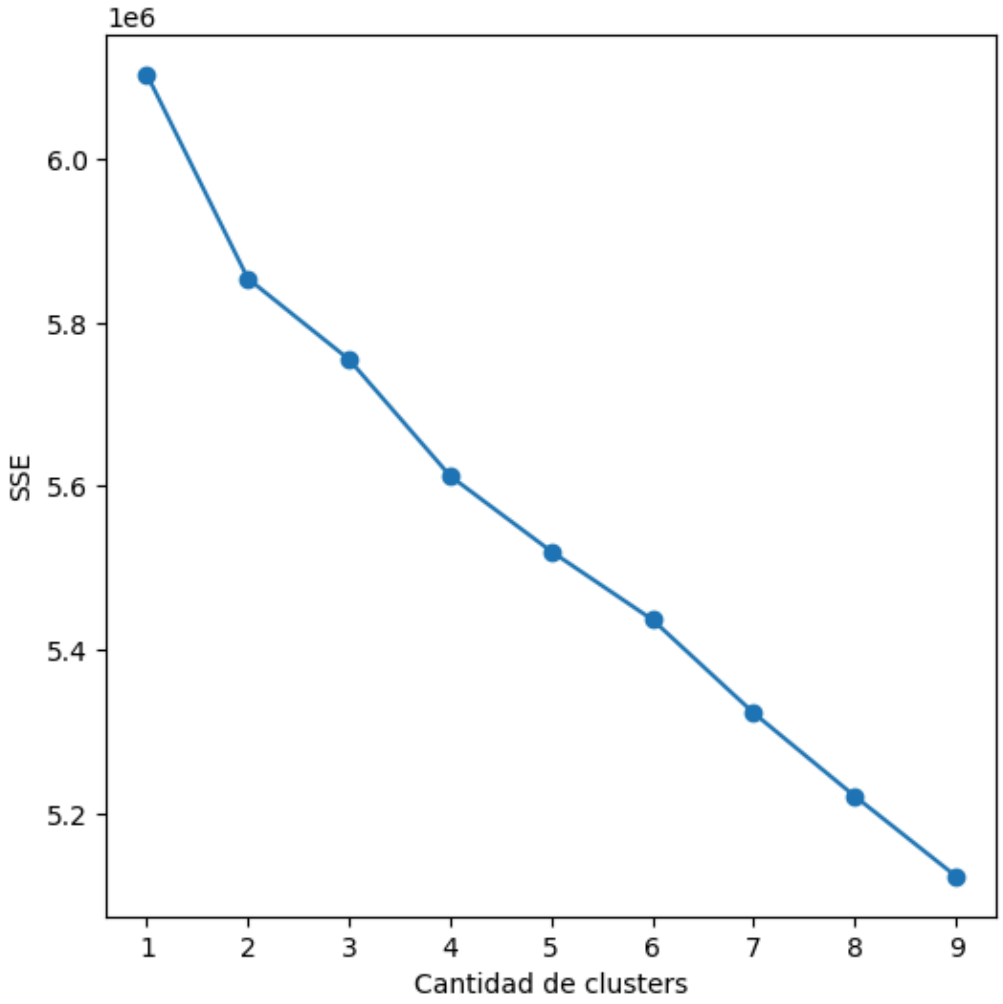



En este otro gráfico observamos la relación precio y cantidad de cuartos. Por un lado, llegamos a la conclusión de que en general cuantos más cuartos tengan más caras serán esas propiedades con algunas excepciones que pueden deberse a la ubicación de dichas propiedades. Por otro lado, también se puede 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.

## Clustering

Antes de empezar nuestro proceso de clusterización lo que hicimos fue transformar las columnas con valores no numéricos a valores numéricos para poder utilizarlos. Para hacer esto, usamos el método del 'One Hot Encoding' y transformamos los barrios y los tipos de propiedad en booleanos para poder utilizarlos. Además, eliminamos algunas columnas que no nos son utiles para este analisis: 'start_date', 'end_date', 'id', 'created_on', 'property_title' y filtramos nuestras propiedades en base a las cuales, según su latitud y longitud solo se encuentran dentro de CABA. Como última cosa antes de agrupar nuestro dataset, estandarizamos sus valores con un StandarScaler.

Luego, para decidir cuántos clusters usar, recurrimos al método del codo. Este nos ayuda a decidir viendo en qué número "se quiebra" el gráfico:



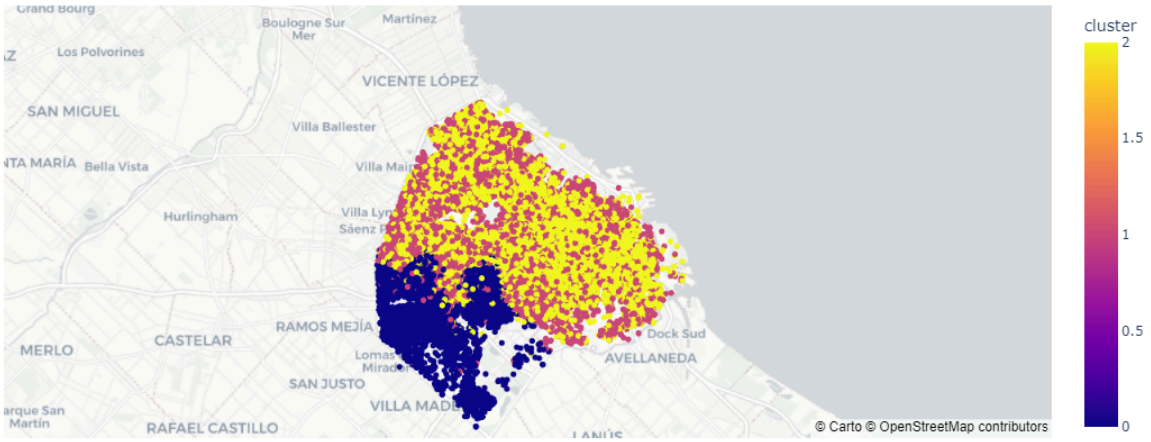
Para tener una “segunda opinión” de cuantos clusters utilizar, también analizamos el número de silhouette para cada cluster. Finalmente, con ambos métodos llegamos a la decisión de usar 2 clusters.

Ahora, con nuestro dataset agrupado en 2 clusters analizamos cuál es el “criterio” que utilizan para agruparse según este mapa de CABA.


Luego de realizar varios gráficos según cantidad de habitaciones, superficie cubierta, en que barrio se ubican cada propiedad y a cluster pertenece llegamos a la conclusión que tenemos una clusterización donde aquellos elementos que pertenecen al grupo azul(0) suelen encontrarse en la zona céntrica, mientras que aquellos que pertenecen al grupo amarillo(1) generalmente se van a encontrar mas cerca de la periferia.

Otra manera de agrupacion que podemos notar es que los elementos del grupo amarillo(1) tienen una mayor superficie total y una mayor cantidad de habitaciones.

Luego, repetimos el análisis anteriormente detallado pero para 3 clusters como pide el enunciado. Veamos como queda distribuido el mapa:



Luego de analizar el mapa y otros gráficos de barras (los mismos vistos para 2 clusters), vemos que el grupo de color azul(0), se concentra en a perfieria de la Ciudad de Buenos Aires, tambien vemos que esta zona “nueva” posee propiedades de precios mas bajos y con una mayor cantidad de casas. Los otros dos grupos parecen respetar la separación que vimos cuando utilizamos únicamente 2 clusters. Vemos una clara tendencia del grupo formado sobre la zona céntrica y cercana al río(amarillo) con un aumento en los precios de dichas propiedades, como es el caso del cluster que tiene la mayor cantidad de datos sobre zonas como 'Recoleta' o 'Puerto Madero'.

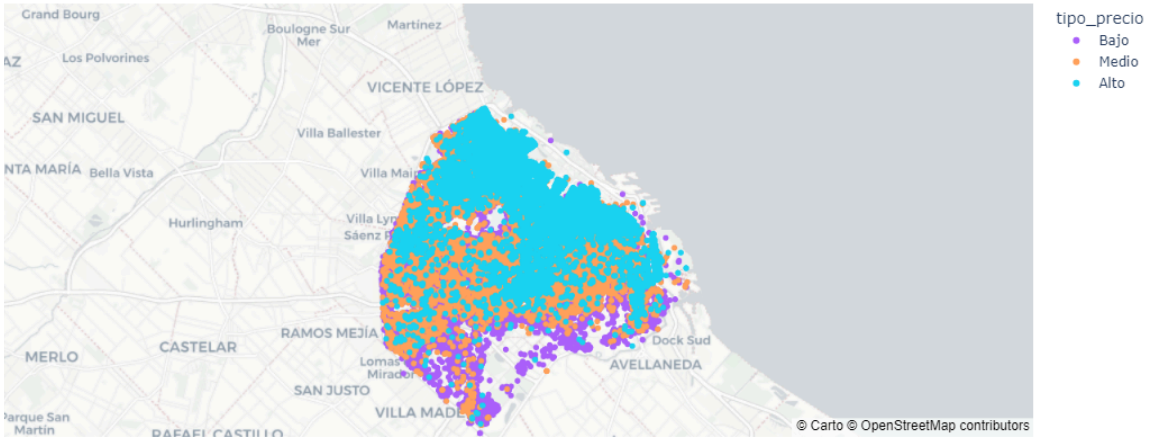
Como conclusión, con 3 grupos vemos una agrupación más clara y distintiva entre las distintas propiedades del dataset.

## Clasificación

Mencionar cuál es la alternativa seleccionada para construir la variable “tipo_precio” justificando su elección. Mostrar en un mapa de CABA los avisos coloreados por tipo_precio ¿Qué diferencias o similitudes encuentran con el agrupamiento realizado por K-Means con 3 grupos?

Para construir la variable ‘tipo_precio’, consideramos que es mejor utilizar la división de cuantiles de la forma 25/50/25 sin diferenciar las propiedades por tipo. Llegamos a esta decisión debido a que cuando separamos nuestro análisis por tipo de propiedad, los casos de *departamento* y *PH* tienden a tomar valores de precios altos. Esto puede ser debido a un mal tratado de outliers de nuestra parte, o mismo porque la distribución no es tan efectiva como la elegimos.

A continuacion tenemos un mapa de CABA con las propiedades coloreadas segun su ‘tipo_precio’:



Esta distribución tiene una similitud con el análisis hecho por K-means para 3 grupos ya que, podemos ver que los avisos de precio 'bajo' se encuentran, casi todos, al sur de la ciudad. Mientras que los de precio 'medio' y 'alto' están ubicadas en la zona más céntrica y cercana al río. Además, podemos ver que cuanto más al norte está la propiedad va aumentando el valor de la misma.

## Estado de Avance

### 1. Análisis Exploratorio y Preprocesamiento de Datos

**Porcentaje de Avance:** 100%/100%

**Tareas en curso:** -..

**Tareas planificadas:** -.

**Impedimentos:** -.

Esta parte del TP la consideramos terminada y corregida.

### 2. Agrupamiento

**Porcentaje de Avance:** 100%/100%

**Tareas en curso:** -.

**Tareas planificadas:** -.

**Impedimentos:** Lo que más nos costó en esta etapa fue poder imprimir los clusters en el mapa de CABA. .

Esta parte del TP la consideramos terminada.

### 3. Clasificación

**Porcentaje de Avance:** 50%/100%

**Tareas en curso:** estamos empezando el punto b) Entrenamiento y predicción, a. Modelo 1: Árbol de decisión.

**Tareas planificadas:** terminar de crear el modelo de árbol de decisión y proceder con los próximos
**Impedimentos:** Nos costó preparar los datos para poder aplicarlos a nuestro cross validation.

## Tiempo dedicado

| Integrante | Tarea | Prom. Hs Semana |
|---|---|---|
| Catalina Basso | Filtrado inicial.<br>Análisis de variables cualitativas y cuantitativas.<br>Variables Irrelevantes.<br>Visualización de datos.<br>Análisis, decisión y ejecución sobre datos faltantes a nivel columna y fila.<br>Informe chp1.<br>Análisis de en cuántos clusters agrupar.<br>Análisis de calidad de los clusters.<br>Ordenar colab.<br>Informe chp2. | 7 |
| Lucas Ruiz | Gráficos de las distribuciones de las variables más importantes (antes y después de ser tratadas).<br>Correlaciones.<br>Visualizaciones.<br>Visualización de datos faltantes a nivel columna y fila.<br>Análisis y visualización de valores atípicos(univariados y multivariados).<br>Ejecución de la decisión sobre cómo tratar los | 8 |

| | | |
|---|---|---|
| | outliers.<br>Análisis de la relación entre el precio de venta y los metros de superficie.<br>Análisis de cada grupo intentando entender en función de qué características fueron formados<br>Gráfico de los grupos sobre un mapa de CABA.<br>Colaboración al crear la variable tipo_precio según el análisis del precio por metro cuadrado de las propiedades.<br>Colaboración al comparar la variable tipo_precio y la agrupación por K-means. | |
| Cristobal Alvarez | Análisis de outliers.<br>Tratado de outliers.<br>Gráficos de variables.<br>Armado de informe.<br>Análisis de tendencia al clustering de nuestro dataset.<br>Estimación de la cantidad apropiada de grupos a utilizar.<br>Evaluación de la calidad de los grupos formados.<br>Gráfico de los grupos sobre un mapa de CABA.<br>Creación de la variable tipo_precio según el análisis del precio por metro cuadrado de las propiedades.<br>Comparación entre la variable tipo_precio y la agrupación por K-means.<br>Obtención de HiperParametros para la | 9 |

| | | |
|---|---|---|
| | creación del árbol de decisión. | |
| Dalmiro Vilaplana | Análisis de correlaciones.<br>Visualización de datos.<br>Gráfico de distribución de las variables importantes.<br>Análisis de outliers.<br>Analisis de clusters<br>correcciones chp 1<br>ejecucion del Z-score para el nuevo analisis de outliers | 3 |