08
22

INFORME PROPERATI

por DSConsultores

RESUMEN
de los objetivos

ANALISIS
del dataset

CONCLUSION
del trabajo realizado

RESUMEN

El presente informe tiene por objetivo realizar un análisis del dataset provisto por la inmobiliaria Properati, quien publica de manera periodica informacion sobre ofertas de distintos tipos de propiedades. Todo el trabajo sera realizado en Python en un entorno de Jupyter Lab.

Este trabajo esta compuesto por dos etapas:

En la **primer etapa** tendremos como objetivos:

- Efectuar una limpieza del dataset provisto, diseñando estrategias para lidiar con los datos perdidos en ciertas variable.
- Realizar un análisis descriptivo de las principales variables
- Crear nuevas columnas a partir de las características dadas que puedan tener valor predictivo
- Arribar a un dataset depurado, ordenado y completo con datos certeros y robustos que puedan utilizarse en la segunda etapa.

En la **segunda etapa**, a partir del dataset depurado del paso 1, se desarrollara un modelo de regresión que permita predecir el precio por metro cuadrado de una propiedad, para que la empresa pueda utilzarlo como tasador automático en las próximas propiedades que sean comercializadas.

A continuacion se muestran todo el trabajo realizado de la primer etapa y se adjunta el notebook con el codigo correspondiente.

INDICE

1

INTRODUCCION

En este primer workshop del curso de Data Science de Digital House, nos enfocaremos en hacer un análisis exploratorio del dataset provisto, realizar todo el proceso de ETL necesario para que el mismo quede lo más íntegro posible y así poder descubrir *insights* que le permitan a la empresa tomar decisiones de manera segura. De igual manera, sentar las bases para luego desarrollar un modelo de regresión que permita predecir el precio por metro cuadrado de una propiedad.

Se plantearán diversas estrategias para abordar fallas y/o faltantes de información, en vistas de lograr el objetivo mencionado anteriormente, mediante la aplicación de los conocimientos adquiridos hasta el momento a lo largo del cursado.

2

DATASET

El dataset contiene información sobre todas las propiedades georeferenciadas de la base de datos de la empresa.
La información que cada propiedad incluye es la siguiente:

operation: sell, rent, **property_type:** house, apartment, ph, **place_name**, **place_with_parent_names**, **country_name**, **state_name**, **geonames_id** (si está disponible), **lat-lon**, **price** (precio original del aviso), **currency:** ARS, USD, **price_aprox_local_currency:** ARS, **price_aprox_usd**, **surface_total_in_m2**, **surface_covered_in_m2**, **price_usd_per_m2**, **price_per_m2**, **floor:** (si corresponde), **rooms**, **expenses**, **properati_url**, **description**, **title**, **image_thumbnail**

Por ser información introducida por seres humanos a través de una plataforma, presenta muchas inconsistencias de todo tipo: valores faltantes, erróneos, duplicados, vacios, etc., que deberán resolverse antes de sacar conclusiones.

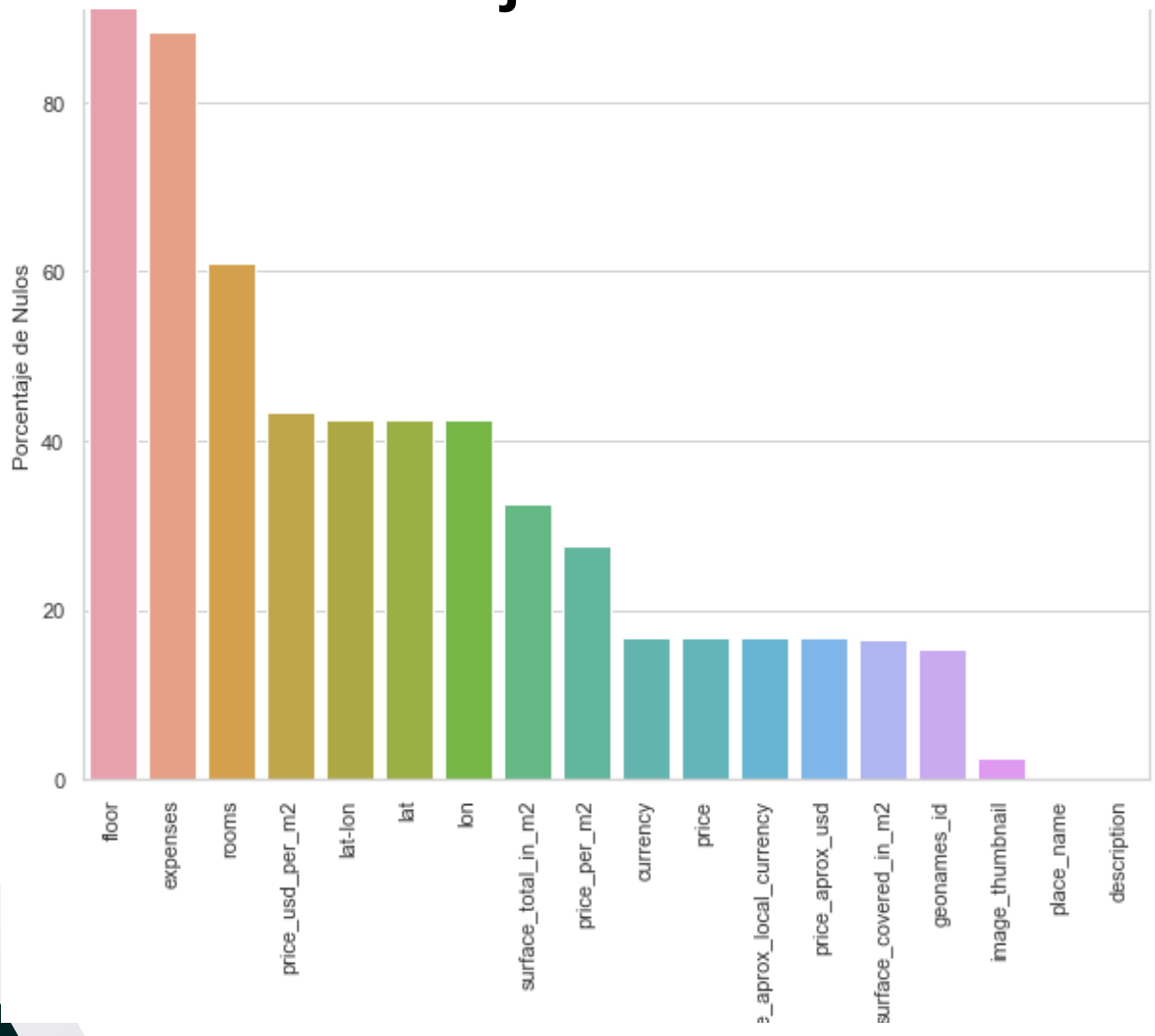
3 ANÁLISIS PRELIMINAR DEL DATASET

En esta primer instancia realizamos un analisis general de los datos que contiene el dataset y obtenemos como mas relevante lo siguiente:

121.220
FILAS

26
COLUMNAS

Porcentajes de Nulos



4 ANÁLISIS DE NULOS, VACÍOS Y DUPLICADOS

Vemos que hay varias columnas con valores nulos o vacios:

- Tenemos las columnas de floor y expenses con un 94% y 89% respectivamente
- Notamos que más del 43% de los registros no tienen el price_usd_per_m2
- Aprox el 18 % de los registros tienen nulls en el campo price_aprox_usd
- También vemos que alrededor del 18% de los registros no cuentan con el campo Superficie Cubierta en m2
- Casi un 35% no cuenta con el campo Superficie Total en m2
- En específico, nos va a interesar completar los campos de precio, cantidad de ambientes y superficie.

Tambien realizamos una busqueda de duplicados tomando en consideración las columnas de 'place_name', 'geonames_id','price','rooms', 'description','title', 'image_thumbnail' y encontramos:

240
DUPLICADOS

	Nulos Totales	Porcentaje de Nulos
floor	113321	93.48
expenses	106958	88.23
rooms	73830	60.91
price_usd_per_m2	52603	43.39
lat-lon	51550	42.53
lat	51550	42.53
lon	51550	42.53
surface_total_in_m2	39328	32.44
price_per_m2	33562	27.69
currency	20411	16.84
price_aprox_local_currency	20410	16.84
price	20410	16.84
price_aprox_usd	20410	16.84
surface_covered_in_m2	19907	16.42
geonames_id	18717	15.44
image_thumbnail	3112	2.57
place_name	23	0.02

5 IMPUTACIÓN DE DATOS FALTANTES

Para poder completar los datos faltantes detectados, luego del proceso de analisis del dataset, se aplicaron los siguientes pasos:

1. **Utilización de expresiones regulares:** Con la aplicacion de distintas regex logramos completar datos faltantes y obtener nuevos datos que consideramos seran importantes para tener en cuenta en el modelo.

Primero analizamos las columnas *Description* y *Title* para completar la mayor cantidad de datos faltantes en las columnas de:

- AMBIENTES: rooms
- UBICACIÓN: place_name
- SUPERFICIE TOTAL: surface_total_in_m2
- SUPERFICIE CUBIERTA: surface_covered_in_m2
- PRECIO: price_aprox_usd
- EXPENSAS : expenses

y agregamos nuevas columnas de tipo "binario", para obtener otros datos que puedan resultar de utilidad,ya que seguramente estos aspectos influyan en el precio de las propiedades, las mismas son:

- Pileta
- Cochera
- Balcón
- Terraza-Patio
- A estrenar

2. Para los campos de Precio total en USD, Precio por metro cuadrado en USD, Superficie Total en m2, rellenamos los valores faltantes utilizando la fórmula Precio Total (USD) = Precio (USD/m2), Superficie Total (m2)

3. Se eliminaron los nulos de precio.

Criterio de prioridad de la información

Luego de los analisis realizados y a partir de los datos obtenidos se acordaron aplicar distintos criterios de prioridad a la informacion, a continuacion detallamos los criterios aplicados en cada caso:

1.Ambientes:

Vamos a dar más importancia a los datos cargados en el sistema en _rooms_, luego los datos de _title_, ya que la información es más precisa, y por último el de _description_. Para eso, almacenamos los datos en una nueva columna llamada _ambientes.

Luego de las imputaciones realizadas Observamos que bajamos los datos faltantes en _rooms_ de 60% a 38%

2.Price:

Vamos a crear una nueva columna llamada ARS_to_USD para recolectar todos los datos de precios en USD,\

los precios que estan en ARS los pasamos a USD dividiendo por el tipo de cambio que observamos en en dataset

6 ANÁLISIS DE OUTLIERS

En esta etapa nos centramos en el analisis de outliers:

6.1) Eliminación de *outliers* evidentes

Lo primero que eliminamos fueron aquellos *outliers* más evidentes dentro de la base de datos, estos eran:

- Precios por metro cuadrados muy elevados (mayor a USD50.000) o muy pequeños (menor a USD100)

- Propiedades especiales, es decir aquellas que no parecieran ser una vivienda finalizada. Con ese criterio eliminamos 'hoteles', 'hostel' y 'desarrollo inmobiliarios'

6. 2) Análisis de outliers según precio m2 CABA y GBA

En segundo lugar, centramos el análisis de *outliers* y limpieza de datos en las propiedades de Buenos Aires, con foco en Capital Federal (CABA) y Gran Buenos Aires (GBA).

Nos centramos en Buenos Aires, al representar dicha provincia más del 80% de la muestra y contar con la mayor cantidad de registros por barrio.

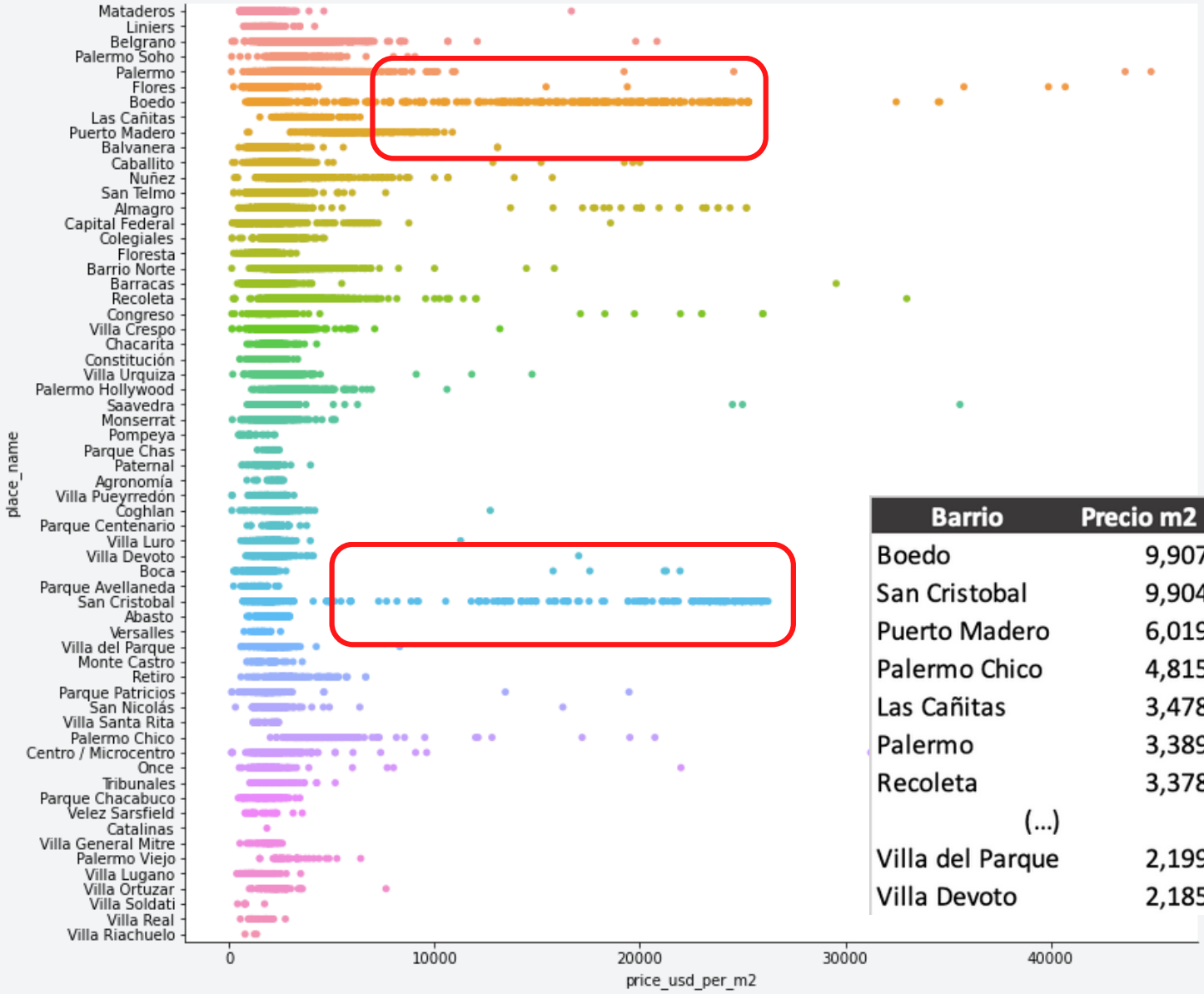
6.2.1. Análisis de Outliers CABA

6.2.1.A) Análisis por barrio

Primero hicimos un análisis de precio por metro cuadraro por barrio a través de cálculo de precio por metros cuadrado promedio en dólares por barrio y gráficos de dispersión, como un análisis inicial del *dataset*.

Precio por metro cuadrado por barrio

Para la variable precio m2 se observa que tiene outliers más marcados en ciertos barrios, como es el caso de Boedo y San Cristóbal:

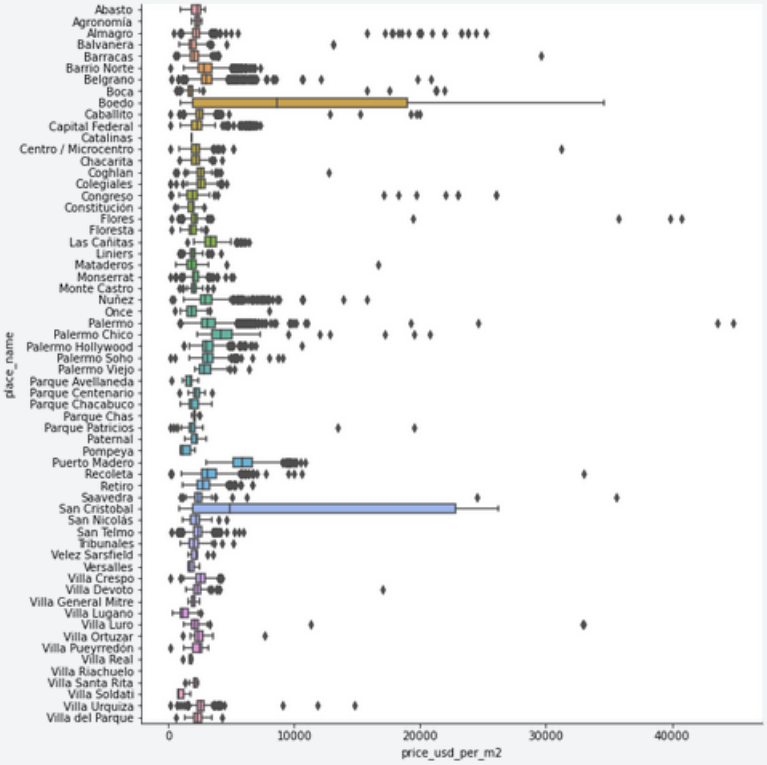


Barrio	Precio m2
Boedo	9,907
San Cristobal	9,904
Puerto Madero	6,019
Palermo Chico	4,815
Las Cañitas	3,478
Palermo	3,389
Recoleta	3,378
(...)	
Villa del Parque	2,199
Villa Devoto	2,185

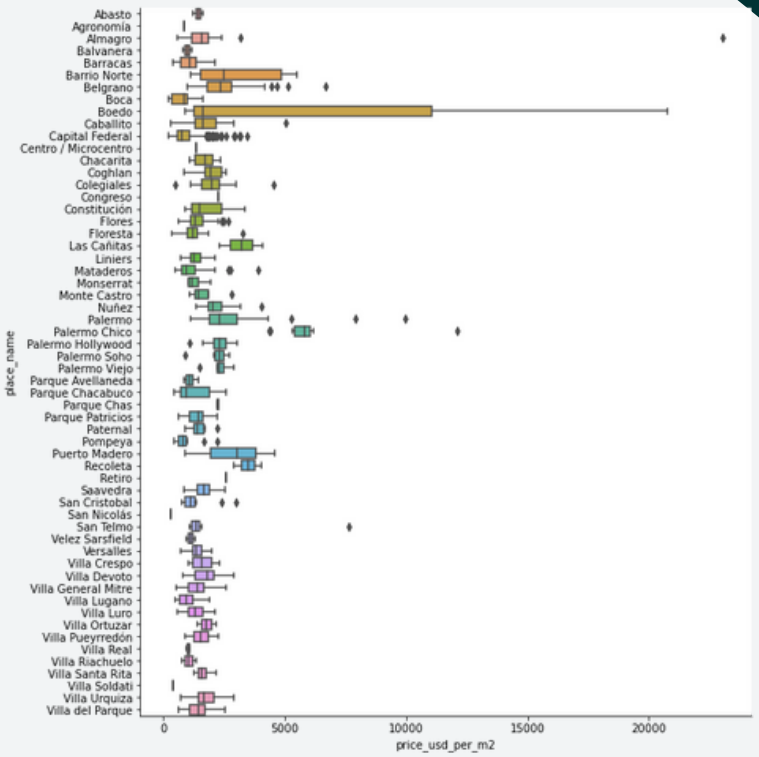
6.2.1.B) Análisis de outliers por tipo de propiedad y barrio

En segundo lugar, agregamos un filtro adicional, por tipo de propiedad.

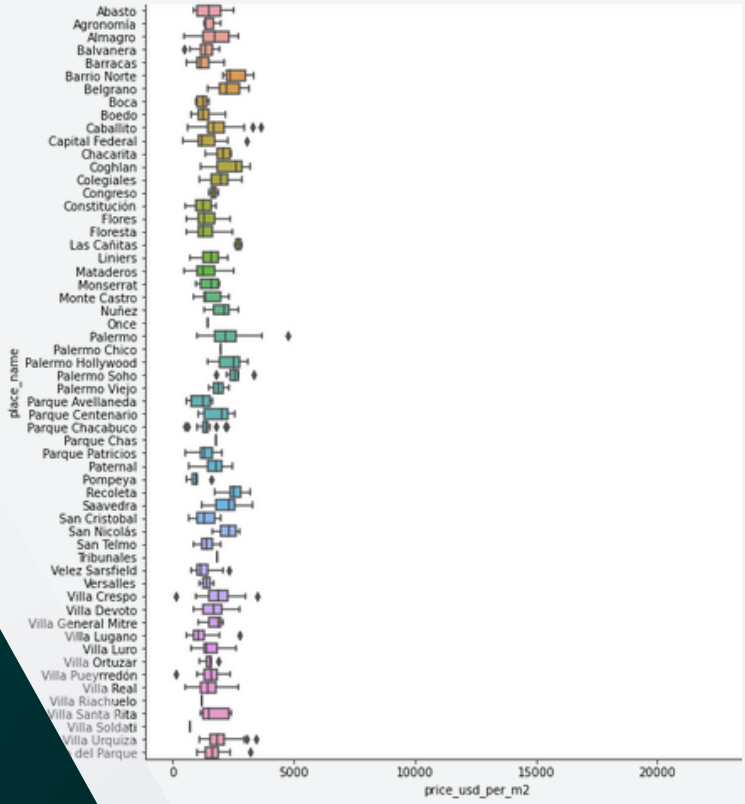
Departamentos



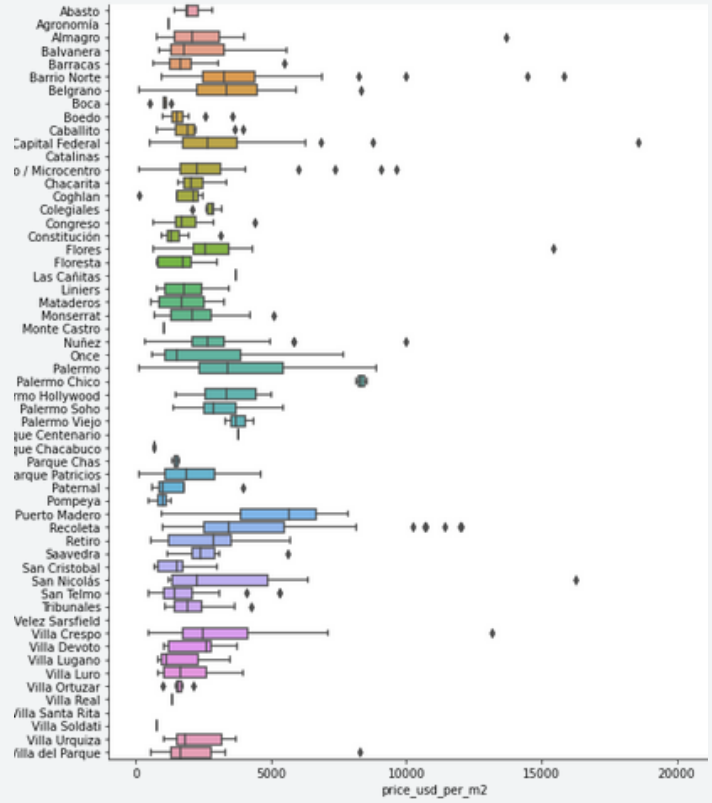
Casas



Propiedad Horizontal - PH



Almacenes



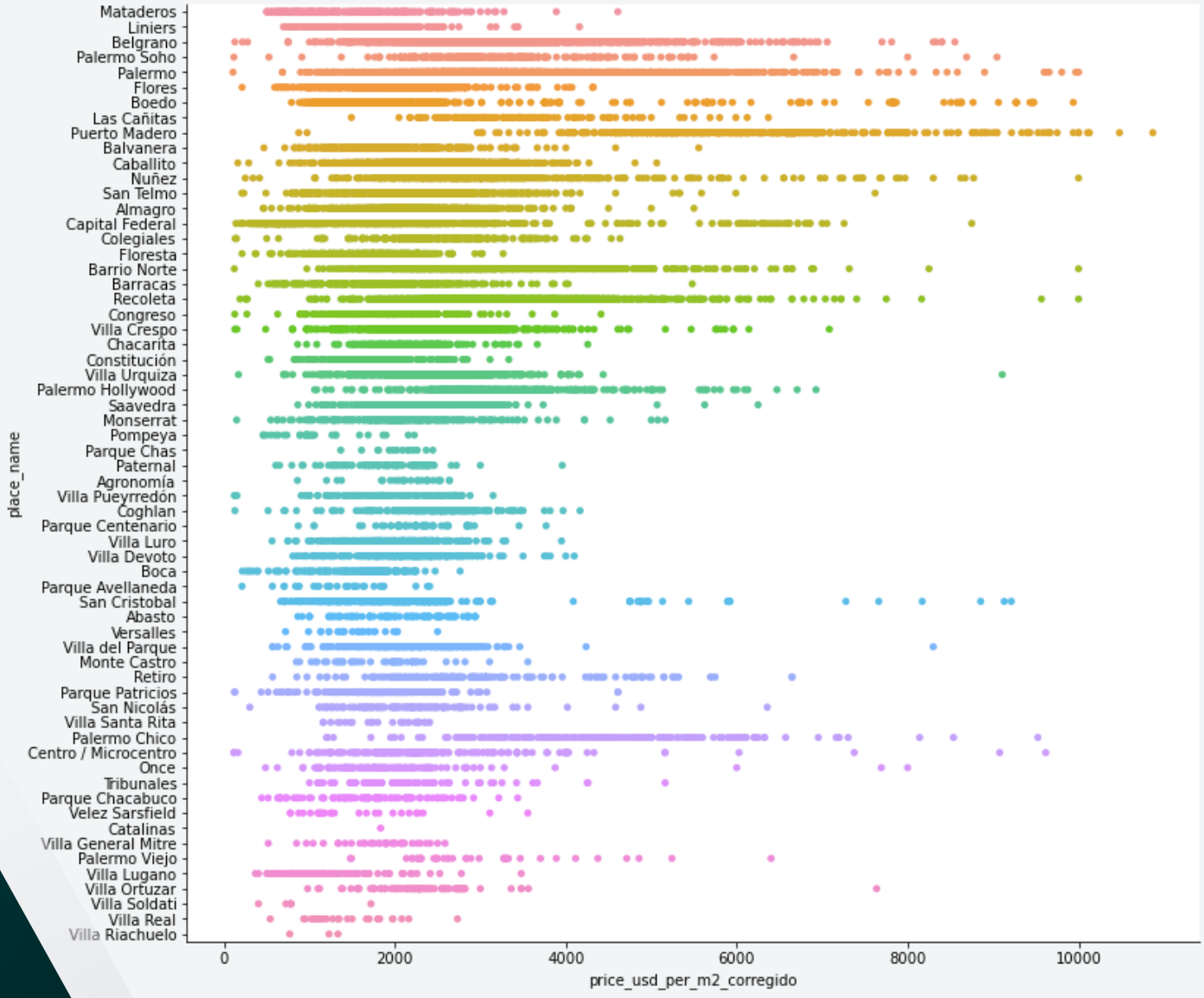
por DSConsultores

En los gráficos anteriores podemos observar que Boedo y San Cristobal (y otros barrios en menor medida) tienen un error en la impuación de datos.

En particular observamos que el precio podría tener un error en el primer decimal . Por lo tanto, procedimos a corregir aquellos datos a los cuales el precio por metro cuadrado en dólares sea mayor a USD10.000.

Al corregir esos valores, vemos como mejora la dispersión.

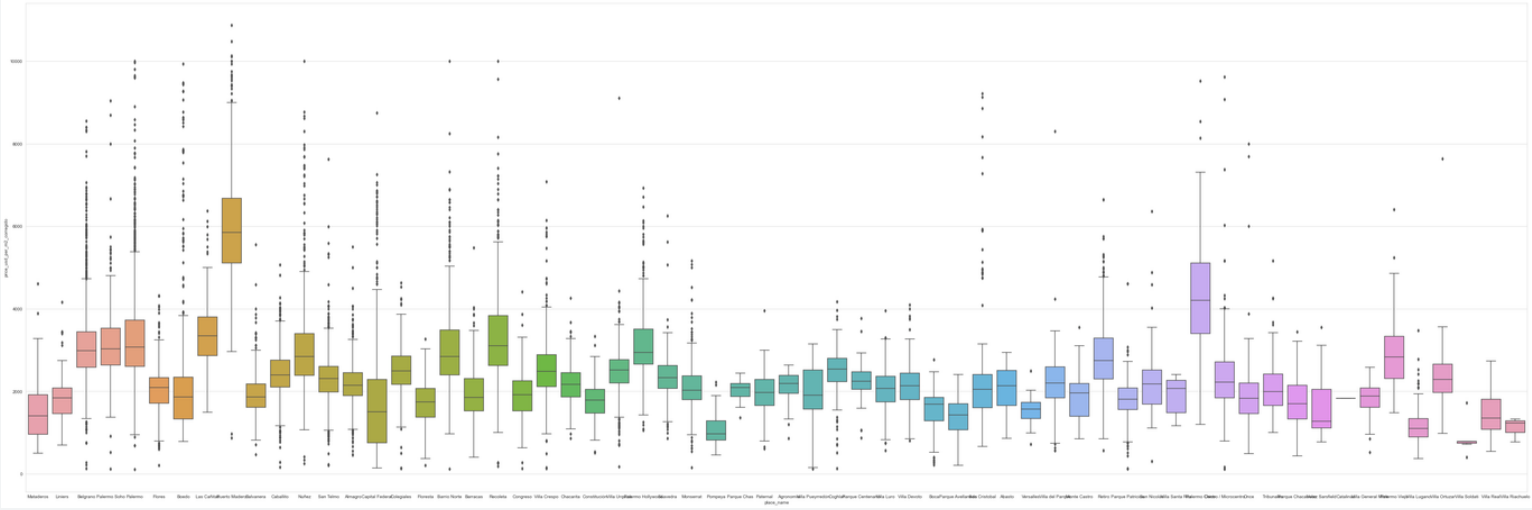
Precio por metro cuadrado corregido por barrio



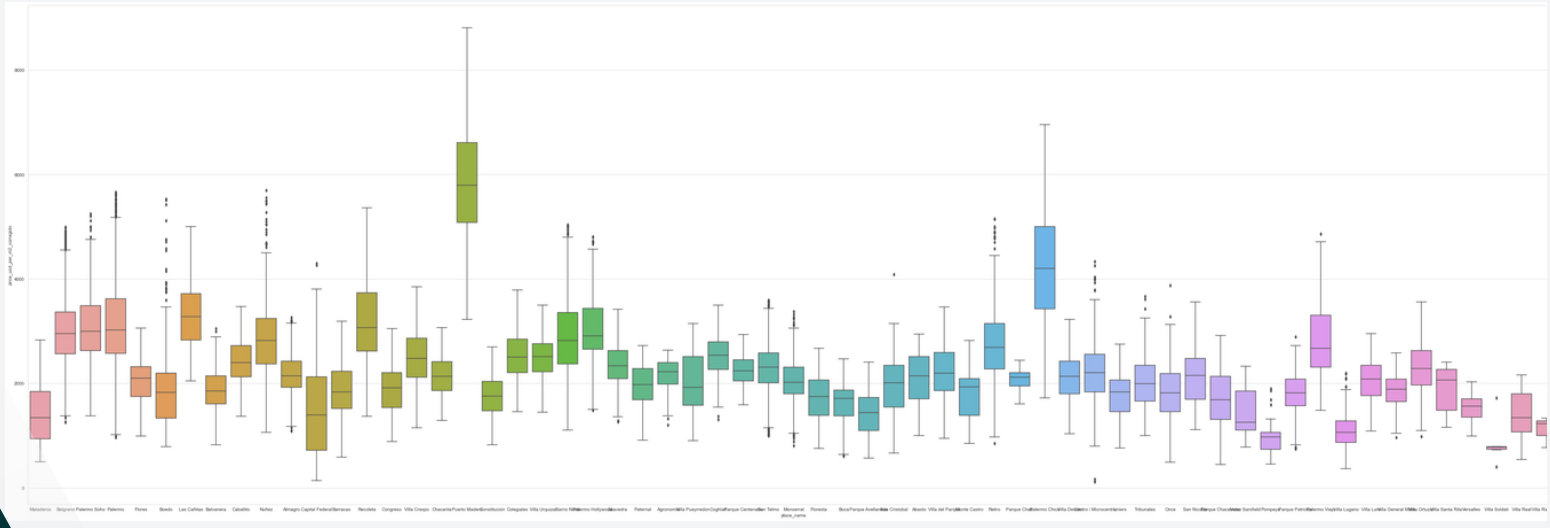
Una vez con los datos sin errores, procedimos a eliminar *outliers* en cada barrio, eliminando aquellos registros que estuvieran a dos desvio de su grupo (barrio).

En el segundo gráfico, vemos como se eliminan los outliers (sin los puntos negros)

Distribución de datos antes de eliminación de outliers



Distribución de datos después de eliminación de outliers

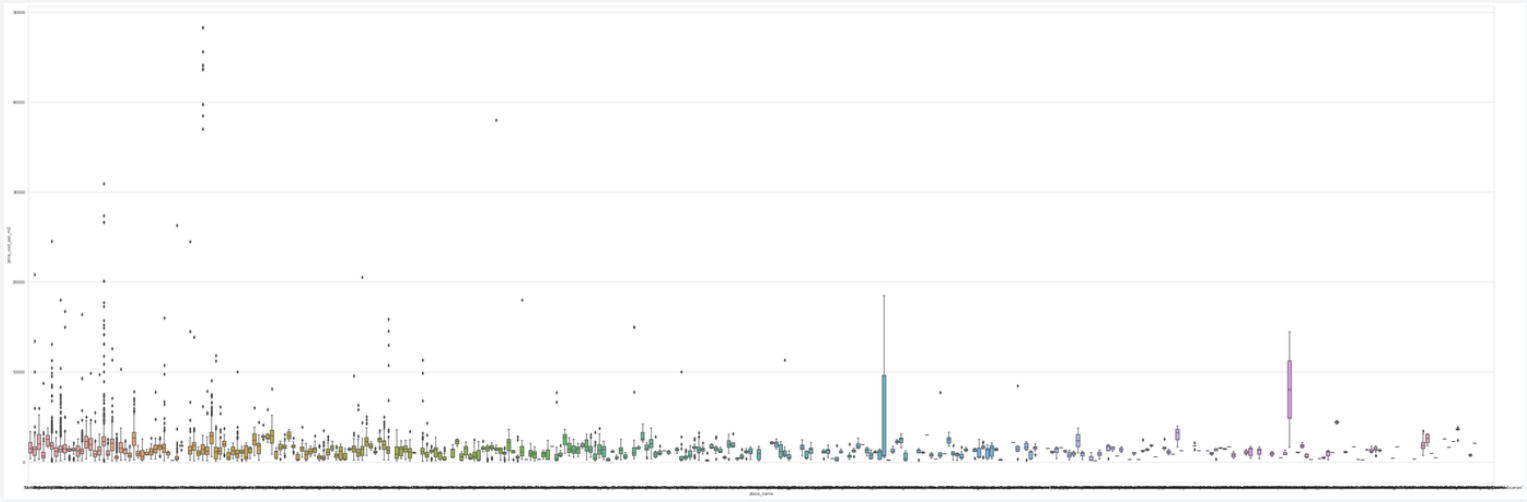


6.2.2. Análisis de Outliers GBA

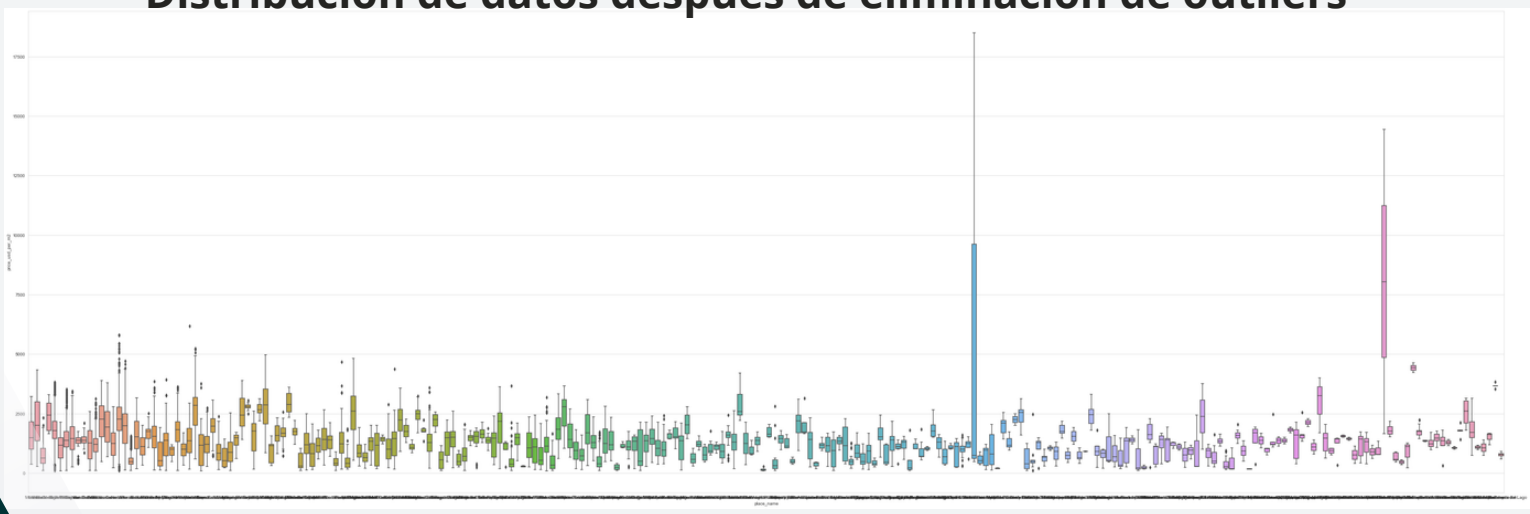
En GBA, un análisis preliminar de datos no indicaba ningún error en la imputación de datos.

Al no identificar datos erróneos, aplicamos el mismo criterio para eliminar aquello outliers, por barrio, dentro de los datos de GBA.

Distribución de datos antes de eliminación de outliers



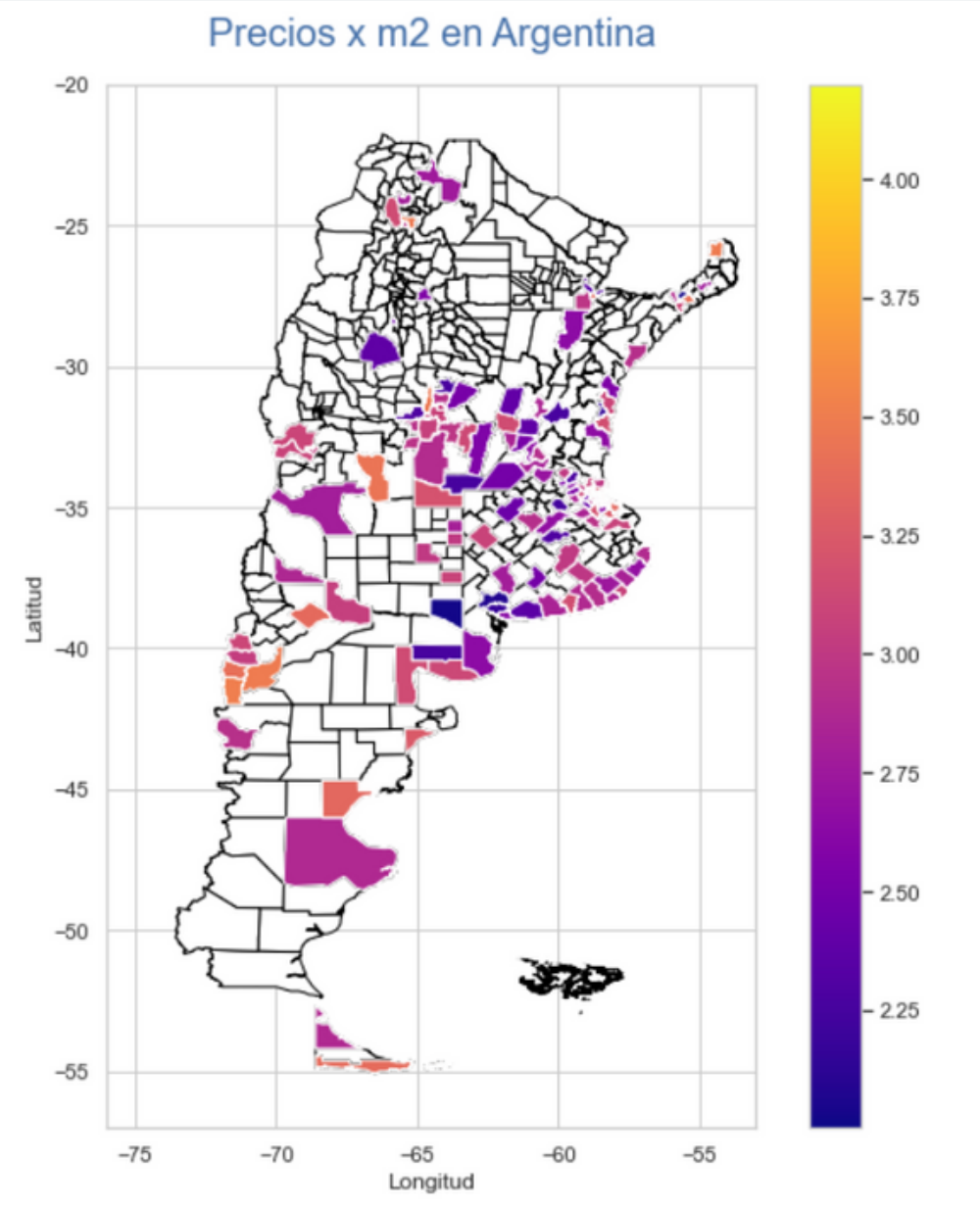
Distribución de datos después de eliminación de outliers



6.3. Limpieza de Outliers

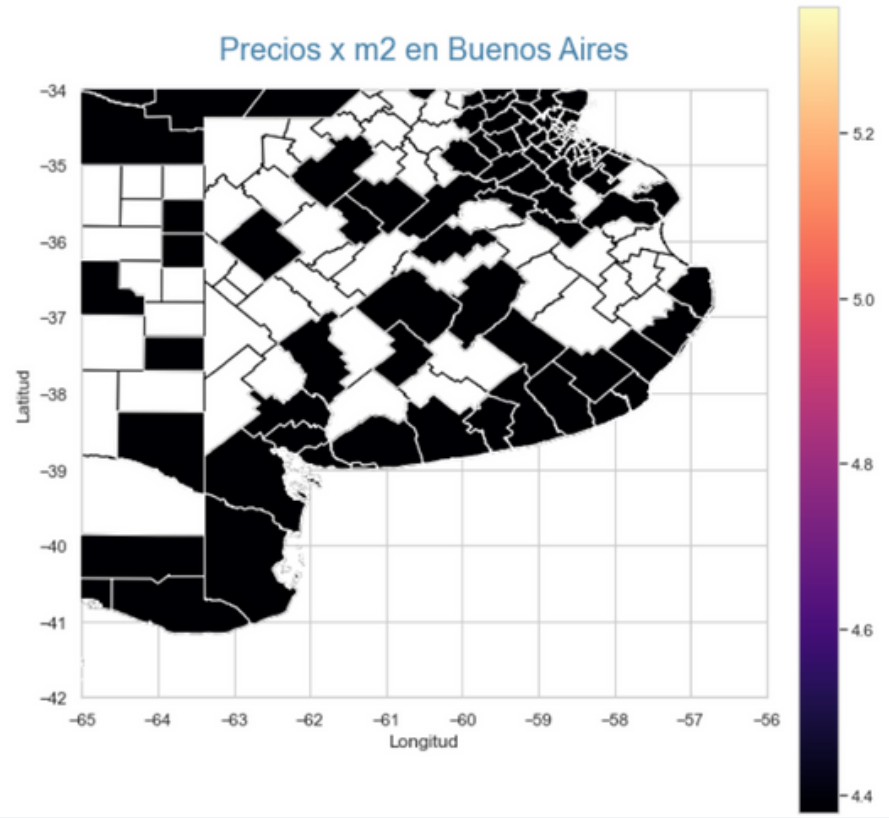
7 GEOLOCALIZACION (GEOPANDAS)

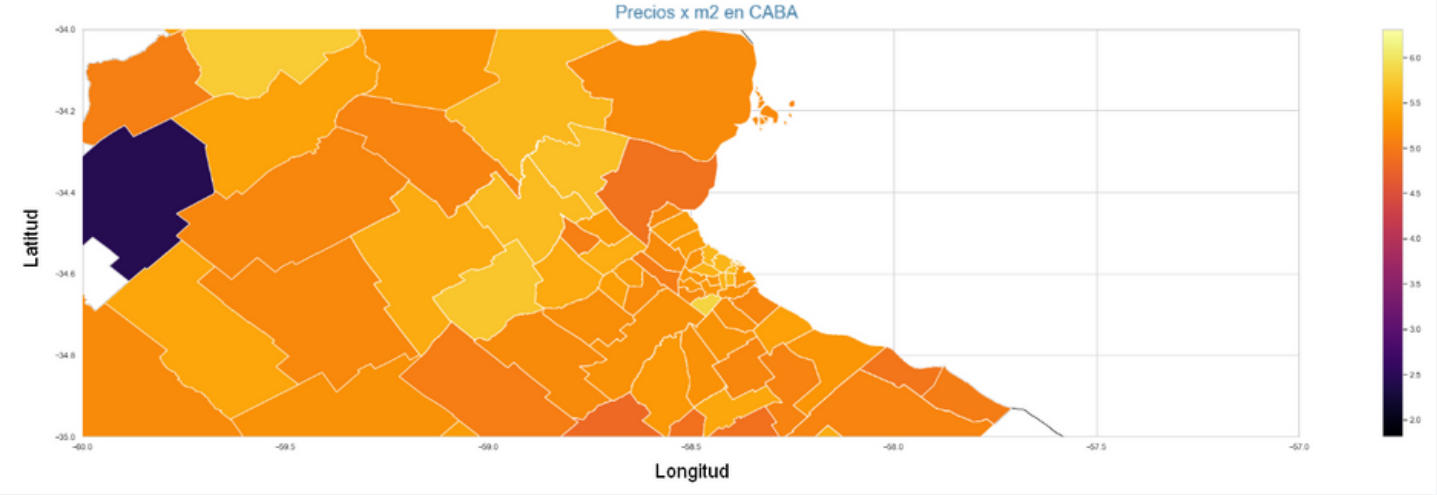
En este punto vamos a trabajar por departamentos, buscamos las latitudes y long de todos los lugares, y a travez de la media de los datos rellenamos los vacios. Una vez que completamos los vacios, eliminamos los datos sin geoposicionamiento. Luego con un archivo json agregamos una columna que representa los departamentos y asociamos el lugar al departamento para graficar con poligonos en vez de puntos.



7

GEOLOCALIZACION (GEOPANDAS)

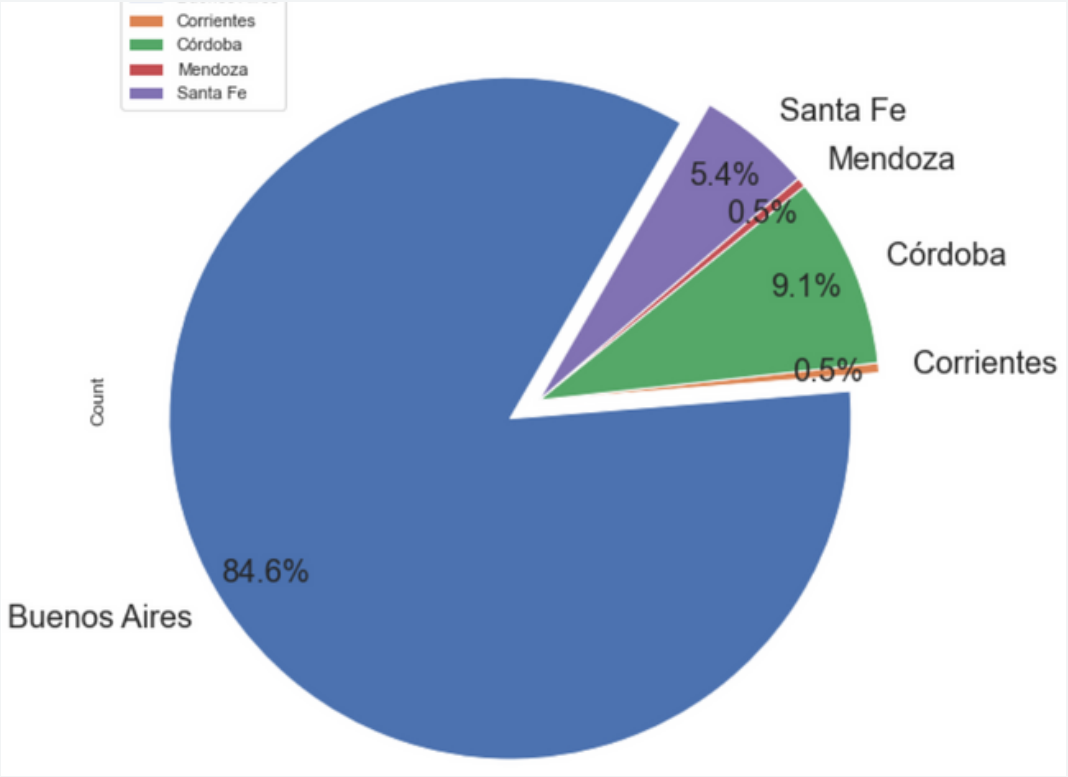




8

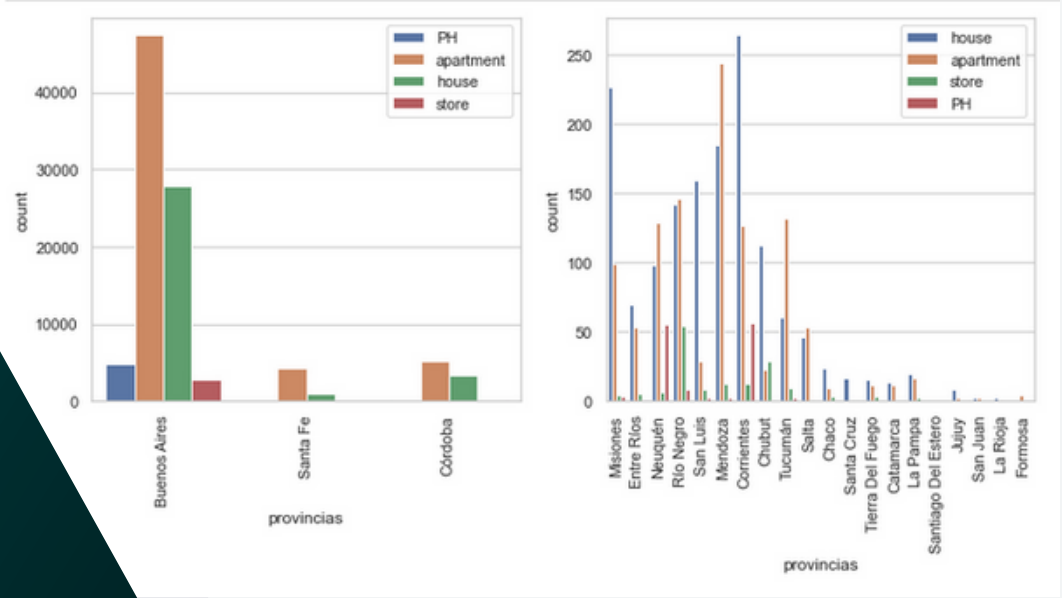
CONCLUSIONES

ANÁLISIS POR PROVINCIA: Identificamos qué provincias son las que tienen mayor cantidad de registros en nuestro Dataset. Focalizamos nuestro análisis en las 3 Provincias que más datos tienen:

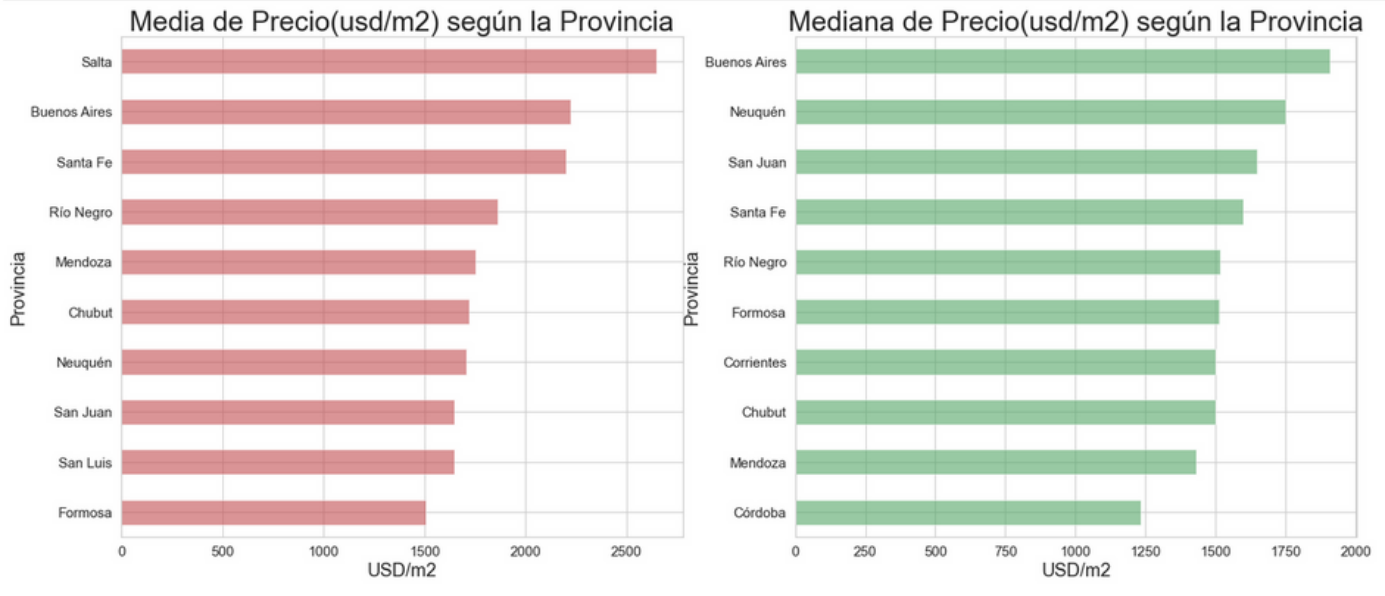


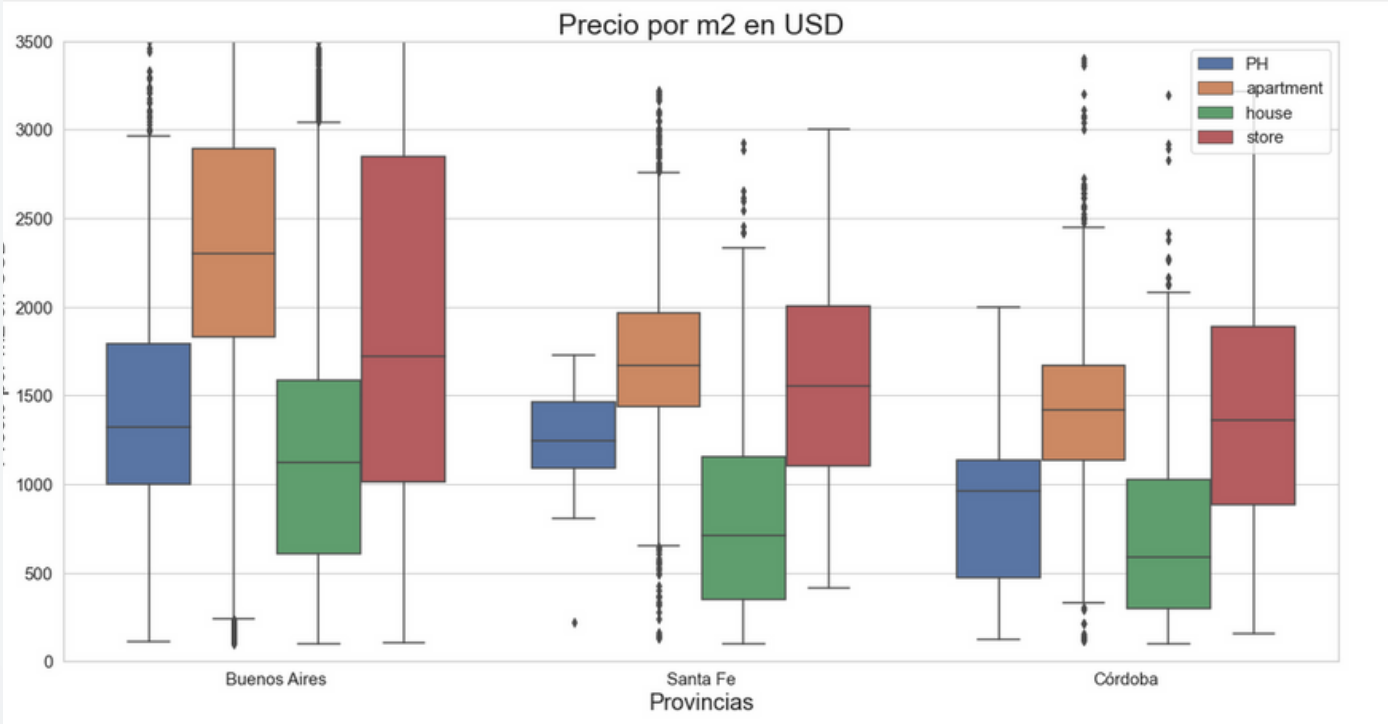
provincias	Count
Buenos Aires	82838
Córdoba	8866
Santa Fe	5294
Corrientes	460
Mendoza	444
Río Negro	350
Misiones	333
Neuquén	288
Tucumán	203
San Luis	198
Chubut	165
Entre Ríos	129
Salta	100
La Pampa	39
Chaco	36
Tierra Del Fuego	31
Catamarca	26
Santa Cruz	19
Jujuy	10
San Juan	4
Formosa	4
La Rioja	3
Santiago Del Estero	2

CANTIDAD DE REGISTROS POR TIPO DE PROPIEDAD Y PROVINCIA:



ANÁLISIS POR PROVINCIA: Utilizamos Boxplots en nuestros análisis, ya que las variables que nos muestran (Mediana, Rango Intercuartil) no son tan sensibles a la presencia de Outliers como si lo es la Media.

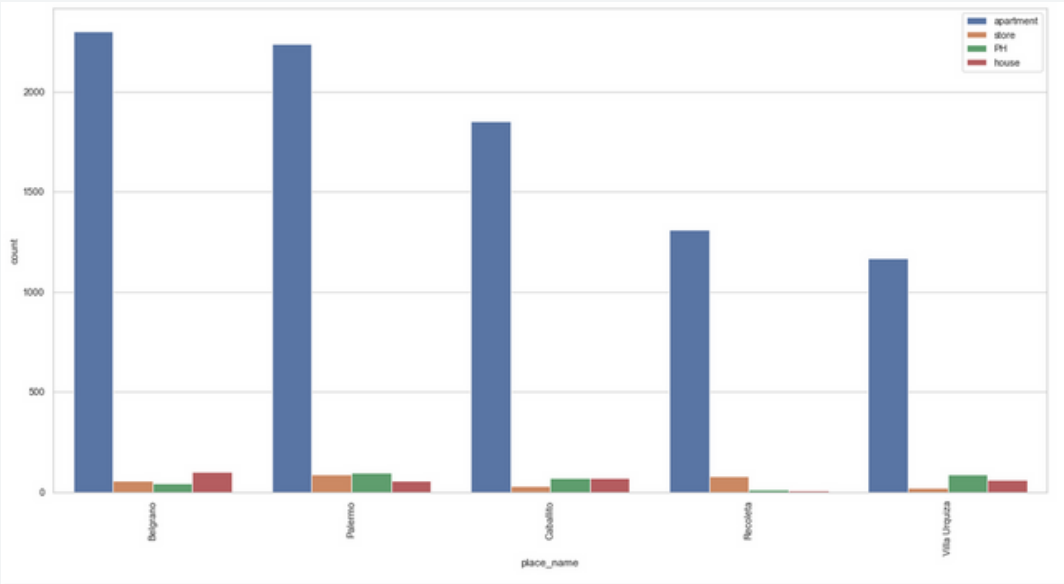




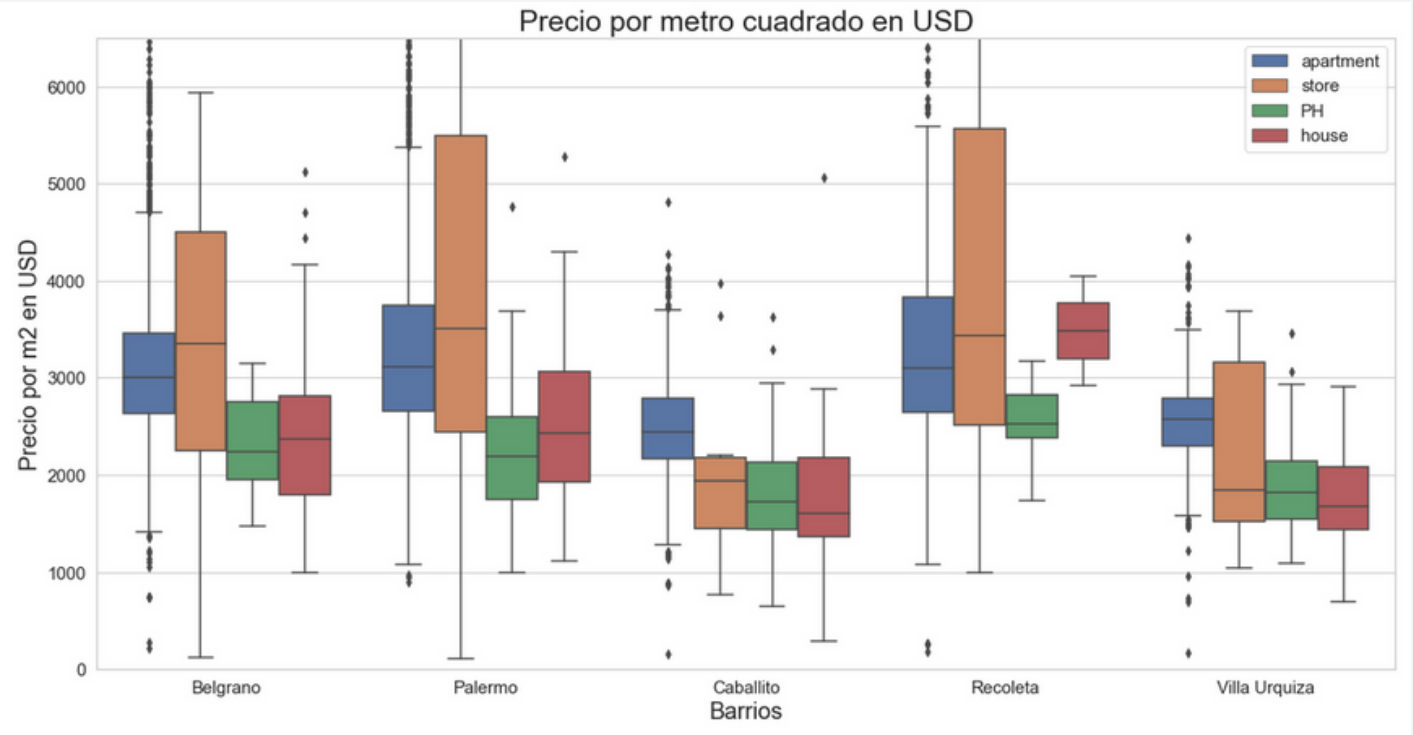
Vemos que en la provincia Buenos Aires tenemos una mediana y cuartiles más elevados que las demás Provincias respecto del Precio por metro cuadrado en USD. Se supone que en esa Provincia debe haber propiedades más caras por m2 que en las otras Provincias.

ANÁLISIS POR PROVINCIA DE BUENOS AIRES (CAPITAL FEDERAL + GBA): Representan casi el 85% de los registros del dataset

- **CAPITAL FEDERAL**: Identificamos los 5 barrios con más registros para mejor visualización en los Boxplot:



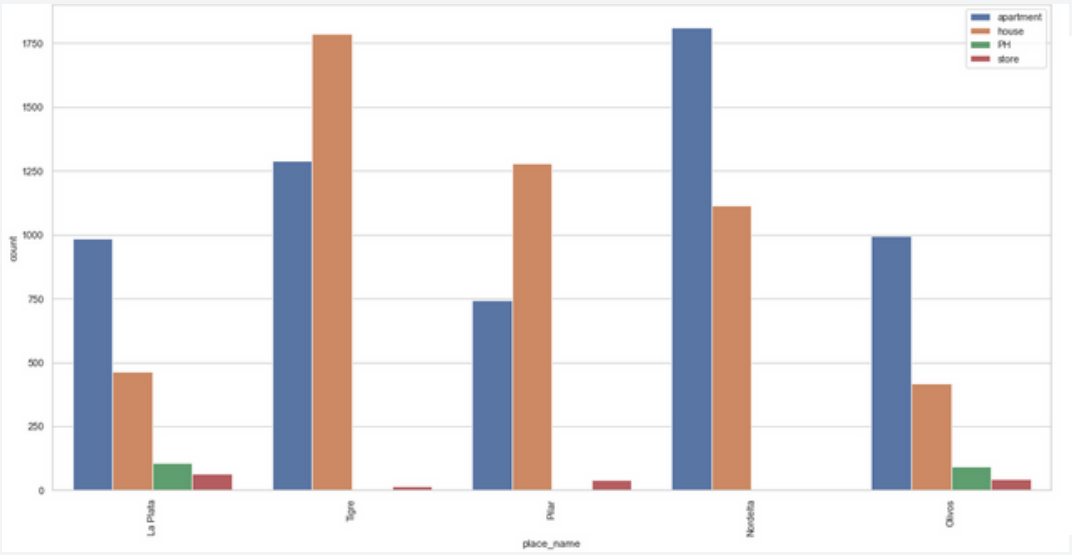
place_name	Count
Belgrano	2499
Palermo	2480
Caballito	2016
Recoleta	1401
Villa Urquiza	1334



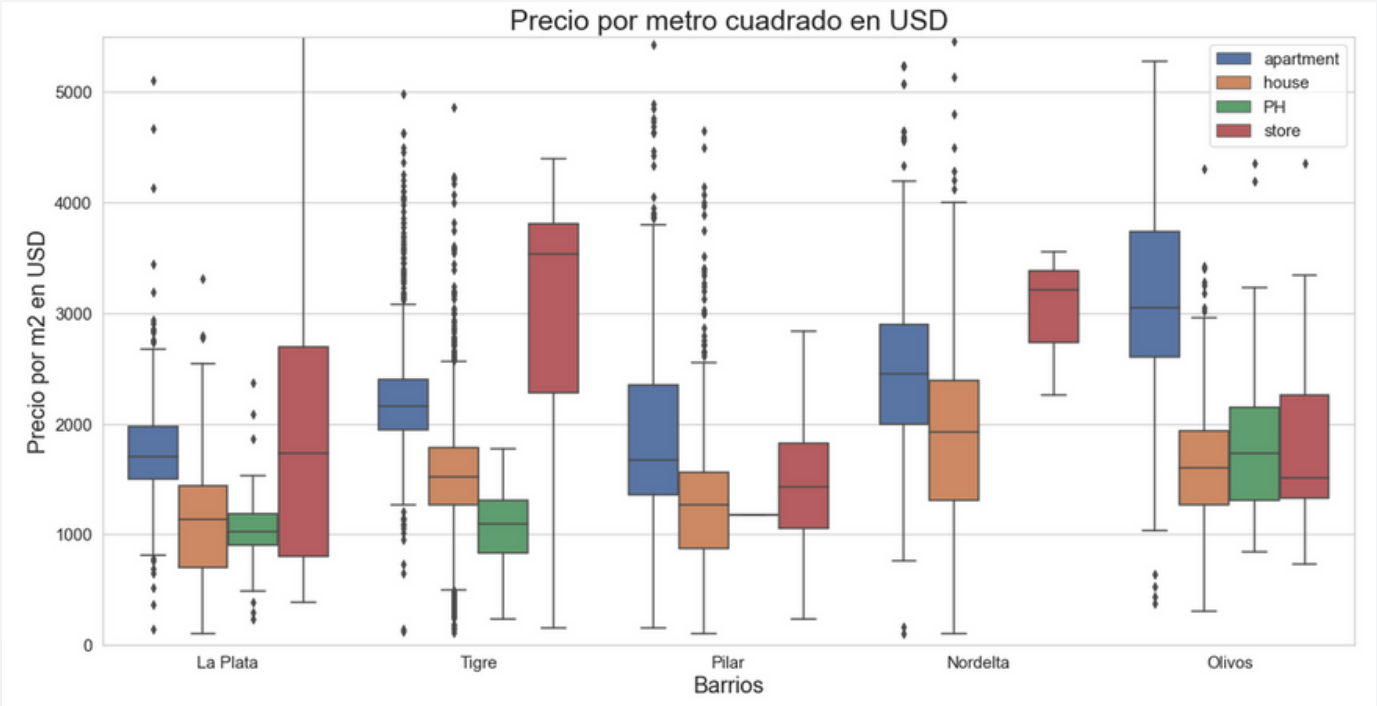
Vemos que en Recoleta y Palermo tenemos una mediana y cuartiles más elevados que los demás barrios respecto del Precio por m2 en USD. Se supone que en esos barrios debe haber propiedades más caras que en los otros barrios.

ANÁLISIS POR PROVINCIA DE BUENOS AIRES (CAPITAL FEDERAL + GBA): Representan casi el 85% de los registros del dataset

- **GBA**: Identificamos los 5 barrios con más registros para mejor visualización en los Boxplot:



place_name	Count
Tigre	3097
Nordelta	2936
Pilar	2069
La Plata	1618
Olivos	1553



Se destacan con mayor precio los stores de Tigre, las casas de Nordelta, y los departamentos y PH de Olivos.

ANÁLISIS PROVINCIA DE BUENOS AIRES (CABA + GBA)

BARRIOS MÁS EXCLUSIVOS

Detalle de los barrios más exclusivos, por registrar el mayor precio por m2

CABA

place_name	price_usd_per_m2
Puerto Madero	5,792
Palermo Chico	4,213
Boedo	3,927
Las Cañitas	3,282
Recoleta	3,077
Palermo	3,028
Palermo Soho	2,991
Belgrano	2,954
Palermo Hollywood	2,909
Nuñez	2,829

GBA

place_name	price_usd_per_m2
St. Patrick Country	8,048
San Isidro Chico	4,425
Manzone	3,657
Bahía del Sol	3,247
Islas del Canal	2,881
La Lucila	2,867
Vicente López	2,840
Enyoi	2,788
Albanueva Barrio Cerrado	2,786
QBay Yacht	2,653

ANÁLISIS PROVINCIA DE BUENOS AIRES (CABA + GBA)

BARRIOS MÁS ECONÓMICOS

Detalle de los barrios más económicos, por registrar el menor precio por m2

CABA

place_name	price_usd_per_m2
Villa Soldati	780
Pompeya	972
Villa Lugano	1,064
Villa Riachuelo	1,230
Velez Sarsfield	1,254
Villa Real	1,340
Mataderos	1,349
Capital Federal	1,402
Parque Avellaneda	1,443
Versalles	1,567

GBA

place_name	price_usd_per_m2
Virrey del Pino	122
Pontevedra	172
Zelaya	186
Fátima	200
Dique Luján	214
Francisco Alvarez	221
Cuartel V	224
González Catán	244
El Canton Barrio Puerto	270
Marcos Paz	286

DISCRETIZACION VARIABLES

Por ultimo y como un extra tambien clasificamos a las propiedades en "pequeñas", "medias" y "grandes" dependiendo la cantidad de ambientes, y en una Pivot Table mostramos cuánto cuesta de Mediana una propiedad en los distintos barrios de Cap. Fed. y GBA, y con la Mediana de cantidad de Ambientes que buscamos:

```
#Clasificamos a las propiedades bajo 3 categorías (pequeño, medio o grande) dependiendo de la cantidad de ambientes:
#1-2 ambientes: "pequeño"
#2-4 ambientes: "medio"
#4-10 ambientes: "grande"
q_rooms=pd.cut(BsAs_dataframe['rooms'],[1,2,4,10],labels=["pequeño","medio","grande"])
q_rooms.head()
```

```
#Tabla Resumen:

result_2 = BsAs_dataframe.pivot_table(['ARS_to_USD','price_usd_per_m2','surface_total_in_m2','rooms'],index=['C.A.B.A./GBA?','place_name',q_roo
           , aggfunc='median').sort_values(by=['C.A.B.A./GBA?','place_name']).round(0)
result_2
```

			ARS_to_USD	price_usd_per_m2	rooms	surface_total_in_m2
C.A.B.A./GBA?	place_name	rooms				
Capital Federal	Abasto	pequeño	84,000	1,953	2	45
		medio	145,000	1,904	3	73
		grande	244,000	1,708	5	120
	Agronomía	pequeño	126,480	2,368	2	56
		medio	140,000	2,000	3	76
		grande	277,500	1,324	6	157
	Almagro	pequeño	95,000	2,256	2	45
		medio	170,000	2,150	3	77
		grande	290,000	1,446	5	199
	Balvanera	pequeño	82,000	1,980	2	41
		medio	139,000	1,750	3	80
		grande	255,000	1,377	5	160
	Barracas	pequeño	103,500	1,954	2	52
		medio	180,000	1,941	3	100
		grande	249,900	1,222	5	180
	Barrio Norte	pequeño	152,000	3,171	2	50
		medio	269,500	2,833	3	96
		grande	500,000	2,768	5	187
	Belgrano	pequeño	150,000	2,833	2	51
		medio	285,000	2,857	3	100
		grande	690,000	2,913	5	254
	Boca	pequeño	70,000	1,742	2	47
		medio	117,000	1,419	3	72

Observamos que a medida que aumenta el tamaño del departamento, el precio total aumenta, y cada vez se vuelve más difícil encontrar compradores que puedan pagar ese precio, por lo que el precio por m2 baja.

No se muestra toda la Pivot Table en la imagen ya que cuenta con aprox 1000 registros.

por DSConsultores