Taller 1

John Vicente Moreno Triviño - 202210162
MINE-4101: Ciencia de Datos Aplicada

1. Introducción

El objeto de este ejercicio es generar recomendaciones para un grupo de inversores, respecto a la adquisición de vivienda en una ciudad de libre elección. Estas recomendaciones deben estar basadas en datos y deben permitir la toma de decisiones que mejoren la rentabilidad en el tiempo de las futuras inversiones. Para esto, se cuenta con el dataset de propiedades en alquiler de Airbnb, el cual representa una valiosa fuente de información sobre el marcado inmobiliario.

2. Entendimiento inicial de datos

La ciudad escogida para este ejercicio puntual es el Buenos Aires (CABA: Ciudad Autónoma de Buenos Aires), capital de Argentina. No debe confundirse Buenos Aires (CABA) con el Gran Buenos Aires (AMBA: Área Metropolitana de Buenos Aires). Los datos en este caso corresponden solamente a la ciudad capital, CABA, y por lo tanto, su alcance geográfico se enfoca principalmente en los sectores más tradicionales, famosos y visitados de esta gran urbe.

El dataset contiene 26204 registros de propiedades en alquiler en la plataforma Airbnb. Estos registros son únicos (no hay duplicados según su ID), pero aun así, no es posible descartar que una misma propiedad aparezca en varios registros con ID distintos debido que su propietario la haya inscrito múltiples veces, o haya pasado por diferentes administradores de alquiler, cada uno creando su propio registro en Airbnb.

Por otra parte, y de acuerdo con el diccionario de datos, el dataset tiene 75 columnas. Se ha realizado un proceso de homologación para ajustar los tipos de datos a las necesidades de este ejercicio. Concretamente, todos los tipos de datos como integer, float o texto numérico se han agrupado como tipo “numérico”. Las datos de texto que representan categorías finitas se ha catalogado como “categóricos”, mientras que los datos de texto que representan descripciones o nombres propios se han rotulado como “string”. Lo datos de tipo booleano, fecha y json se han dejado tal cual.

La siguiente tabla muestra el resultado de este ejercicio de homologación de tipos de datos, al tiempo que resalta aquellos atributos que constituyen el Top 5 de relevancia para este ejercicio.

Field	Type	Field	Type
id	numeric	beds	numeric
listing_url	string	amenities	json
scrape_id	numeric	price	numeric
last_scraped	date	minimum_nights	numeric
source	categoric	maximum_nights	numeric
name	string	minimum_minimum_nights	numeric
description	string	maximum_minimum_nights	numeric
neighborhood_overview	string	minimum_maximum_nights	numeric
picture_url	string	maximum_maximum_nights	numeric
host_id	numeric	minimum_nights_avg_ntm	numeric
host_url	string	maximum_nights_avg_ntm	numeric
host_name	string	calendar_updated	date
host_since	date	has_availability	boolean
host_location	categoric	availability_30	numeric
host_about	string	availability_60	numeric
host_response_time	numeric	availability_90	numeric
host_response_rate	numeric	availability_365	numeric
host_acceptance_rate	numeric	calendar_last_scraped	date
host_is_superhost	boolean	number_of_reviews	numeric
host_thumbnail_url	string	number_of_reviews_ltm	numeric
host_picture_url	string	number_of_reviews_l30d	numeric
host_neighbourhood	categoric	first_review	date
host_listings_count	numeric	last_review	date
host_total_listings_count	numeric	review_scores_rating	numeric
host_verifications	numeric	review_scores_accuracy	numeric
host_has_profile_pic	boolean	review_scores_cleanliness	numeric
host_identity_verified	boolean	review_scores_checkin	numeric
neighbourhood	categoric	review_scores_communication	numeric
neighbourhood_cleansed	categoric	review_scores_location	numeric
neighbourhood_group_cleansed	categoric	review_scores_value	numeric
latitude	numeric	license	string
longitude	numeric	instant_bookable	boolean
property_type	categoric	calculated_host_listings_count	numeric
room_type	categoric	calculated_host_listings_count_entire_homes	numeric
accommodates	numeric	calculated_host_listings_count_private_rooms	numeric
bathrooms	numeric	calculated_host_listings_count_shared_rooms	numeric
bathrooms_text	string	reviews_per_month	numeric
bedrooms	numeric		

Tabla 1. Homologación propia de los tipos de datos

1-Top 5: Barrio

Los atributos sombreados en azul constituyen el Top 5 de relevancia. El primero de estos es *"neighbourhood_cleansed"*, el cual representa el barrio de la propiedad, determinado según la longitud y la latitud registradas. Se considera que esta variable es la que mejor describe la ubicación del inmueble puesto que se basa en variables determinísticas como las coordenadas, las cuales contrasta contra un mapa público de los barrios en formato shape. Esto disminuye la posibilidad de errores en la identificación del barrio de cada propiedad.

No se ha considerado conveniente usar el barrio digitado por el anfitrión ya que este podría equivocarse (con o sin intención) en el barrio exacto. Por otra parte, es cierto que las coordenadas (latitud y longitud) podrían dar valiosa información sobre el terreno, sin embargo, dado que el alcance de este ejercicio es el de dar recomendaciones de inversión, se considera que el nivel geográfico de barrio es el idea para generar consejos de compra por sector. La siguiente figura muestra los barrios ordenados por el número de propiedades en alquiler:

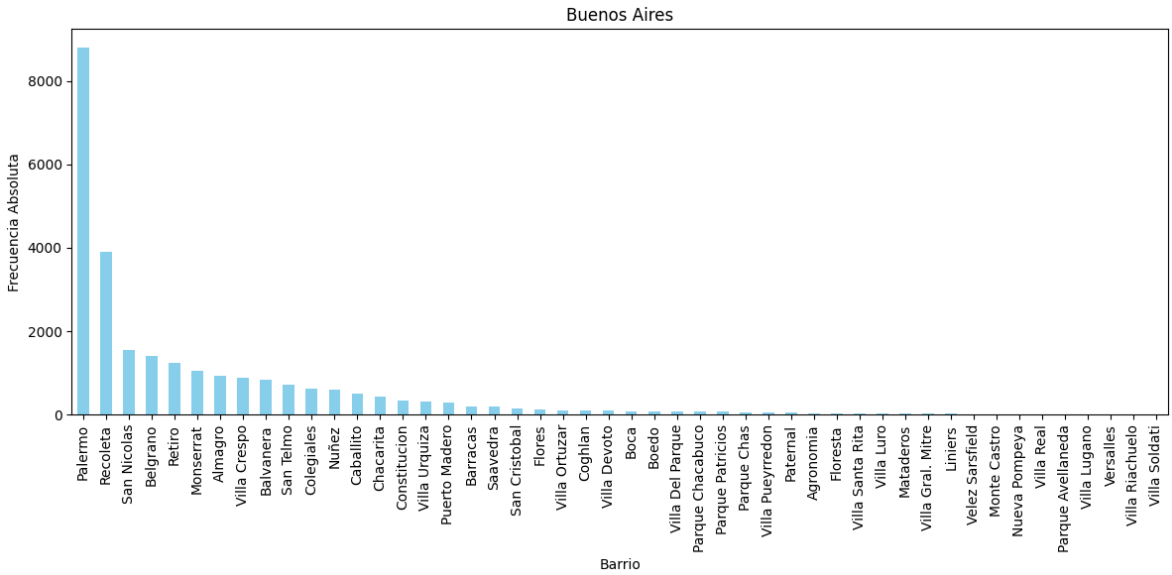

Figura 1. Barrios por número de propiedades en alquiler

Es posible observar que barrios típicamente turísticos como Palermo, San Nicolás (Microcentro), Recoleta y Belgrano comandan la lista. Sin embargo, es notorio que barrios tan populares como San Telmo o Puerto Madero no estén en el top 5. Este último se podría explicar debido a que por sus altísimos precios no es asequible para la mayoría de turistas o huéspedes de corto plazo.

2-Top 5: Tipo de propiedad

El siguiente atributo de interés es “property_type”, el cual indica el tipo de propiedad. Esta variable será fundamental a la hora de generar recomendaciones, debido a que ilustrará a los inversionistas sobre el tipo de inmueble que es más fácil y rentable alquilar. Otra opción hubiera sido escoger la variable “room_type”, pero esta solo posee tres categorías (Propiedad entera, habitación privada y habitación compartida), y por ende, no ofrece una información lo suficientemente desagregada sobre la propiedad. La siguiente figura muestra su distribución por frecuencia. Es posible observar que las propiedades más comunes son unidades que se arriendan enteras (con gran diferencia), bien sean apartamentos, propiedades en condominios o lofts. Siguen en el listado las habitaciones privadas en apartamentos y casas, y luego las habitaciones compartidas. Finalmente, se observa que existen habitaciones en alquiler en hoteles y hostales, sin embargo, estas no

constituyen una parte significativa de la muestra, tal vez porque AirBnB no se enfoca en el negocio de hoteles y hostales, y existen otras plataformas mejor posicionadas en este segmento, tales como Booking o HostelWorld.

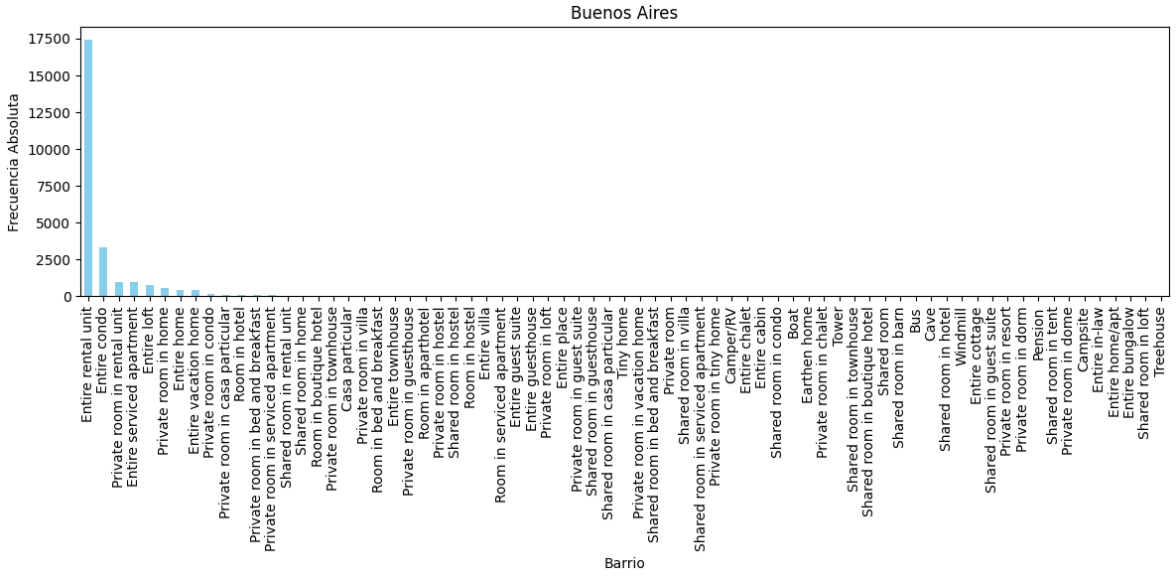

Figura 2. Barrios por número de propiedades en alquiler

3-Top 5: Precio por noche

El tercer atributo seleccionado es “price”, del cual no se considera que sea necesaria una mayor explicación que simplemente decir que el precio de alquiler de una propiedad es uno de los factores más importantes a la hora de calcular la rentabilidad futura de una inversión inmobiliaria. Se muestra un histograma con la distribución de estos precios por noche para Buenos Aires:

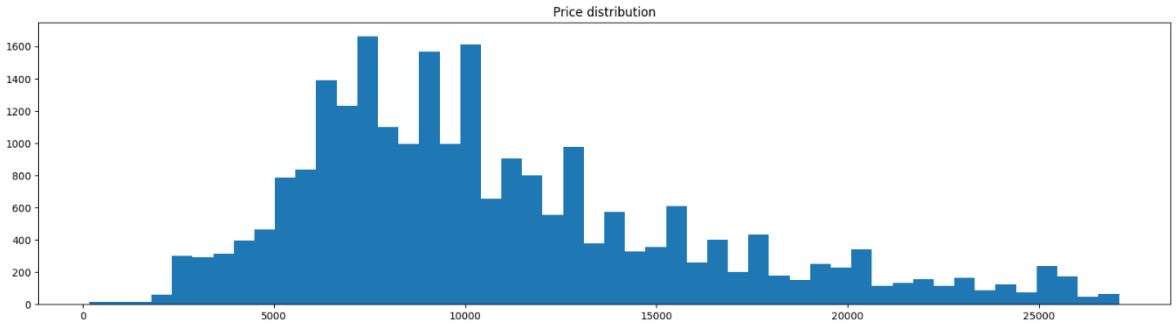

Figura 3. Histograma de precios de alquiler en Buenos Aires

La gráfica, junto con un análisis de percentiles, muestran que la mayoría de los alquileres (percentil 5 a percentil 75) se encuentran en el rango de 4300 a 15300 ARS por noche: entre 17 y 60 USD, usando la tasa de cambio oficial del 28 de junio de 2023, fecha en que fue actualizado este dataset.

4-Top 5: Disponibilidad a 30 días

El cuarto atributo escogido es “availability_30”, entendido como el número de noches en que la propiedad estará disponible en los próximos 30 días. Este atributo permite calcular el número de reservas que la propiedad tiene en el próximo mes (30 días menos “availability_30”). Este número de reservas será muy útil para estimar la ocupación de una propiedad durante un periodo de inversión. Específicamente, se ha escogido el periodo de 30 días en el futuro porque se considera que los huéspedes, en general, suelen hacer sus reservaciones durante el mes anterior a su viaje (esto puede ser refutado, pero se usará como “rule of thumb” para este ejercicio). De este modo, tomar 30 días en el futuro puede dar una visión realista de cómo es la ocupación del inmueble en general, eliminando el ruido producido por reservas en el futuro lejano, las cuales son inciertas.

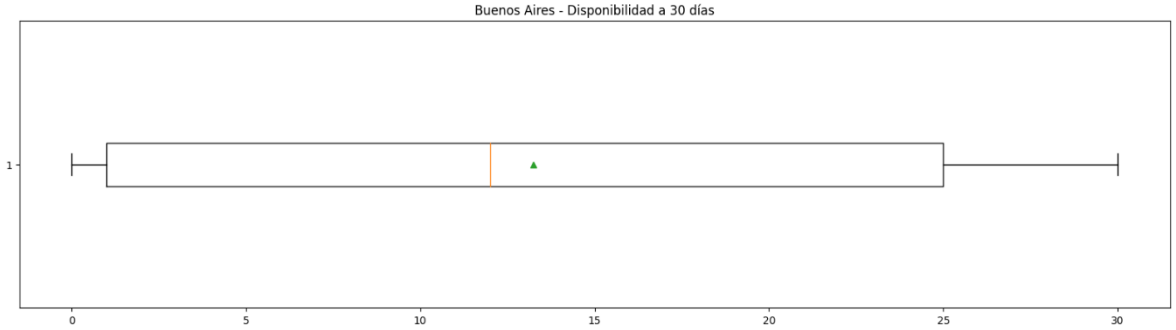

Figura 4. Diagrama de caja con la disponibilidad a 30 días de las propiedades

Sin duda alguna, sería mucho más diciente eficiente tener acceso al historial de alquileres pasados de cada propiedad, teniendo un panorama completo de su ocupación en años y meses anteriores. Lamentablemente, esta información no está disponible en el dataset. En el diagrama de caja es posible observar que las propiedades se distribuyen muy uniformemente en cuanto a su disponibilidad en los próximos 30 días. Principalmente, el grueso de las propiedades está disponible entre 2 y 25 días en el próximo mes.

5-Top 5: Número de opiniones en el último año

La quinta variable permitirá tener una idea de la ocupación en el pasado mediante el análisis del número de opiniones de cada propiedad en el último año “number_of_reviews_ltm”. Se ha escogido el último año para tener una visión de las propiedades alquiladas en tiempos recientes, eliminando el ruido de propiedades alquiladas hace muchos años que tal vez ya no están en el catálogo. Además, se considera que el último año permitirá tener una perspectiva clara de la dinámica actual de alquileres en Buenos Aires: el mercado inmobiliario evoluciona con el tiempo, por lo que es preferible trabajar con datos frescos.

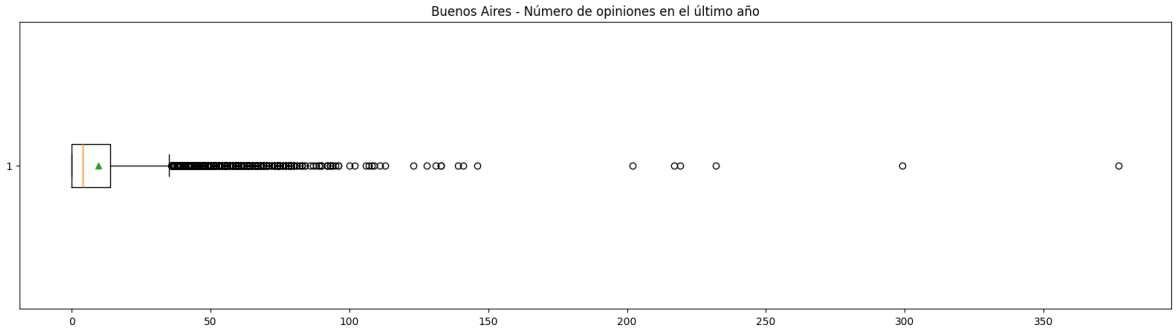

Figura 5. Diagrama de caja con el número de opiniones por propiedad en el último año

El diagrama de caja mostrado permite evidenciar que la gran mayoría (P75 + 1.5IQR) de las propiedades tienen 40 o menos opiniones en el último año. Incluso es posible ver que hasta el P75 solo tienen 75 opiniones. No obstante, no es despreciable el número de propiedades outliers que concentran un gran número de opiniones, y por lo tanto, podrían tener altos niveles de ocupación.

Si bien es cierto que no es posible establecer una relación determinística entre el número de opiniones y la ocupación de una propiedad (esto porque cada opinión puede corresponder a unos pocos días de ocupación o a muchos días de ocupación de diferentes huéspedes, o incluso muchos huéspedes ni siquiera dejan sus opiniones), de todas formas se considera que este atributo sirve como indicador de la ocupación, pensando que entre más opiniones tenga una propiedad, en general es porque ha sido alquilada más veces.

3. Estrategia de análisis

Con base en el análisis estadístico básico que se acaba de presentar, y la combinación de análisis multivariable de los atributos, se procederá a determinar las características de las propiedades en las que más vale la pena invertir. La estrategia básicamente buscará caracterizar a aquellas propiedades que más tiempo pasan ocupadas, con base en su ubicación, el tipo de propiedad y las opiniones de los huéspedes.

Sin embargo, la ocupación por sí sola puede no ser suficiente para estimar la rentabilidad de una propiedad. También es necesario considerar el precio del alquiler según las teorías de mercado: una propiedad muy cara se arrendará muy poco, mientras que una propiedad por debajo de su precio habitual se arrendará más. Por lo tanto, lo interesante aquí será aproximarse al punto de equilibrio que maximiza la rentabilidad, mediante la correcta combinación de precios por noche y días de ocupación.

4. Desarrollo de la estrategia

La estrategia se ha desarrollado en el notebook Colab adjunto en el repositorio, y se presenta aquí el proceso realizado.

4.1 Mejores barrios

Disponibilidad a 30 días vs. Número de opiniones en el último año:

En primer lugar, se intentó verificar si existía una relación entre la disponibilidad a 30 días en el futuro y el número de opiniones en el último año de una propiedad. Para ello, se ejecutó un análisis visual mediante un gráfico de dispersión:

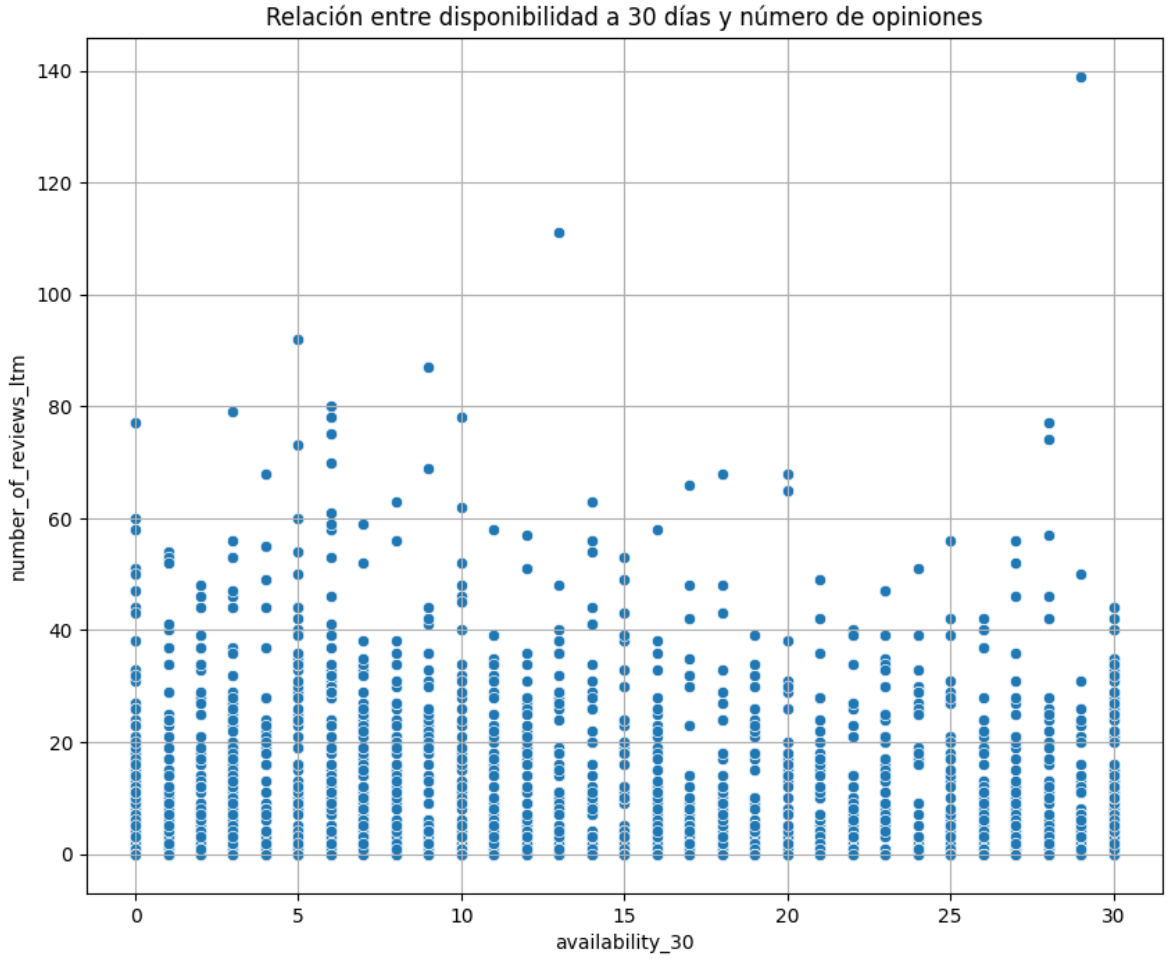

Figura 6. Diagrama de dispersion "number_of_reviews_ltm" vs. "availability_30"

De manera general se puede concluir que parece no haber relación entre estas dos variables. Aterrizando un poco el tema, lo que realmente se busca es un indicador que permita estimar la ocupación real que tendría una propiedad para invertir en ella. Mirando al pasado, el número de opiniones podría revelar cuantos huéspedes han pasado por la propiedad. Mirando al futuro, la disponibilidad del próximo mes podría ser un buen indicador: si la propiedad está disponible muchos días el próximo mes, es porque tal vez no se alquila bien. Esto lleva a pensar que podría haber una relación entre el número de opiniones pasadas y la disponibilidad futura, pero este no parece ser el caso.

Una posible razón para esto es que las opiniones pueden ser tanto buenas como malas. Las buenas favorecen el alquiler, mientras que las malas inciden en que los clientes no opten

por una propiedad. Esto de todas formas es solo una teoría. De momento, se priorizará el análisis con la disponibilidad futura.

Disponibilidad a 30 días vs. Barrio:

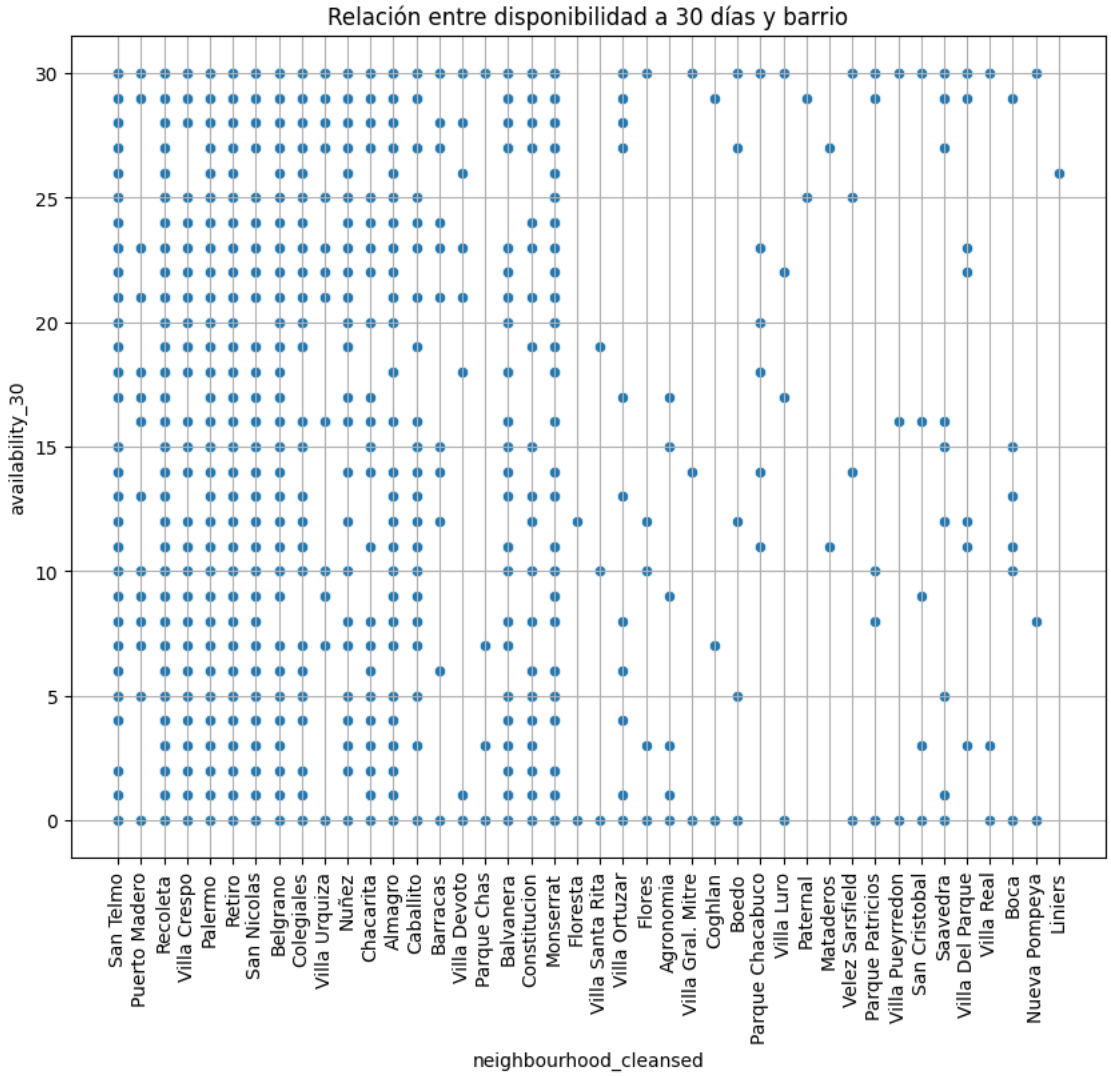

Figura 6. Diagrama de dispersion "neighbourhood_cleansed" vs. "availability_30"

Una primera conclusión de la gráfica anterior es que en una buena parte de los barrios hay propiedades con cero días de disponibilidad en el próximo mes. Sin embargo, esto realmente puede significar que la propiedad ha sido bloqueada por el anfitrión, por lo que estos registros no son útiles para nuestros datos.

Por otra parte, se evidencia que el diagrama de dispersión se divide en dos zonas diferenciadas: a la izquierda, desde San Telmo a Monserrat, se observan propiedades que tienen una distribución de sus días de disponibilidad muy uniforme, siendo las excepciones Villa Devoto y Parque Chas. Por otra parte, a la derecha se observan barrios cuyas

propiedades, en general, suelen estar disponibles entre 10 a 15 días, o más de 25 días en el próximo mes. Lo que lleva a pensar que en estos barrios es más difícil alquilar.

Lo anterior se ratifica por el hecho de que en la parte izquierda de gráfica se encuentran los barrios famosos y turísticos de Buenos Aires, como San Telmo, Puerto Madero, Recoleta, Palermo, San Nicolás (Microcentro), entre otros. De este modo se puede concluir que en los barrios de la parte derecha es difícil alquilar, aunque no se puede dar una conclusión certera sobre los barrios de la parte izquierda. Para ello, se graficará el promedio de la disponibilidad a 30 días vs. el barrio:

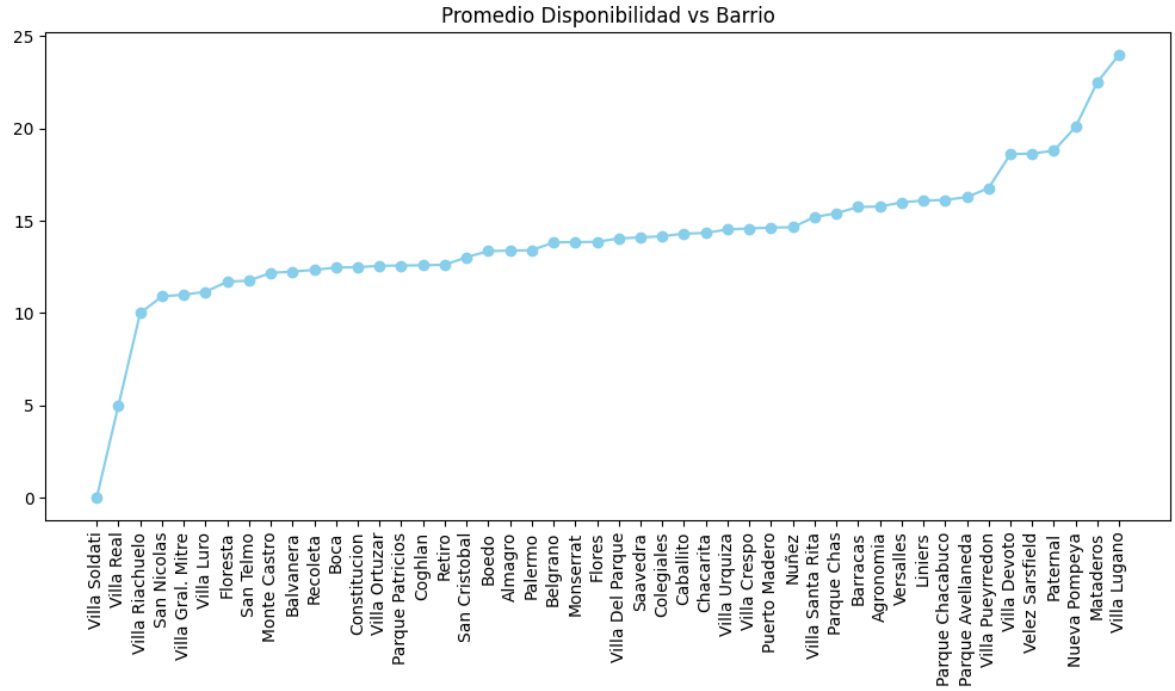

Figura 6. Promedio de la disponibilidad a 30 días por barrio

Aquí el panorama comienza a aclararse. Al tomar el promedio de disponibilidad a 30 días por barrio es más fácil hacerse una imagen general de la situación. Se considerará aquí que cualquier barrio por debajo de la mediana es un barrio con alta ocupación (esto depende de manera más precisa del equilibrio entre oferta y demanda), pero para los objetivos de este ejercicio es útil conocer estos barrios: los da la parte izquierda entre Villa Real y Flores (Se excluye Villa Soldati, porque su promedio de disponibilidad es cero).

Número de opiniones en el último año vs. Barrio:

Cuando se trata de identificar los barrios más populares entre los huéspedes por medio del número de opiniones en el último año, la respuesta es clara. Los barrios a la izquierda de la gráfica, desde San Telmo a Almagro, y añadiendo Balvanera, Constitución y Monserrat son los más comentados:

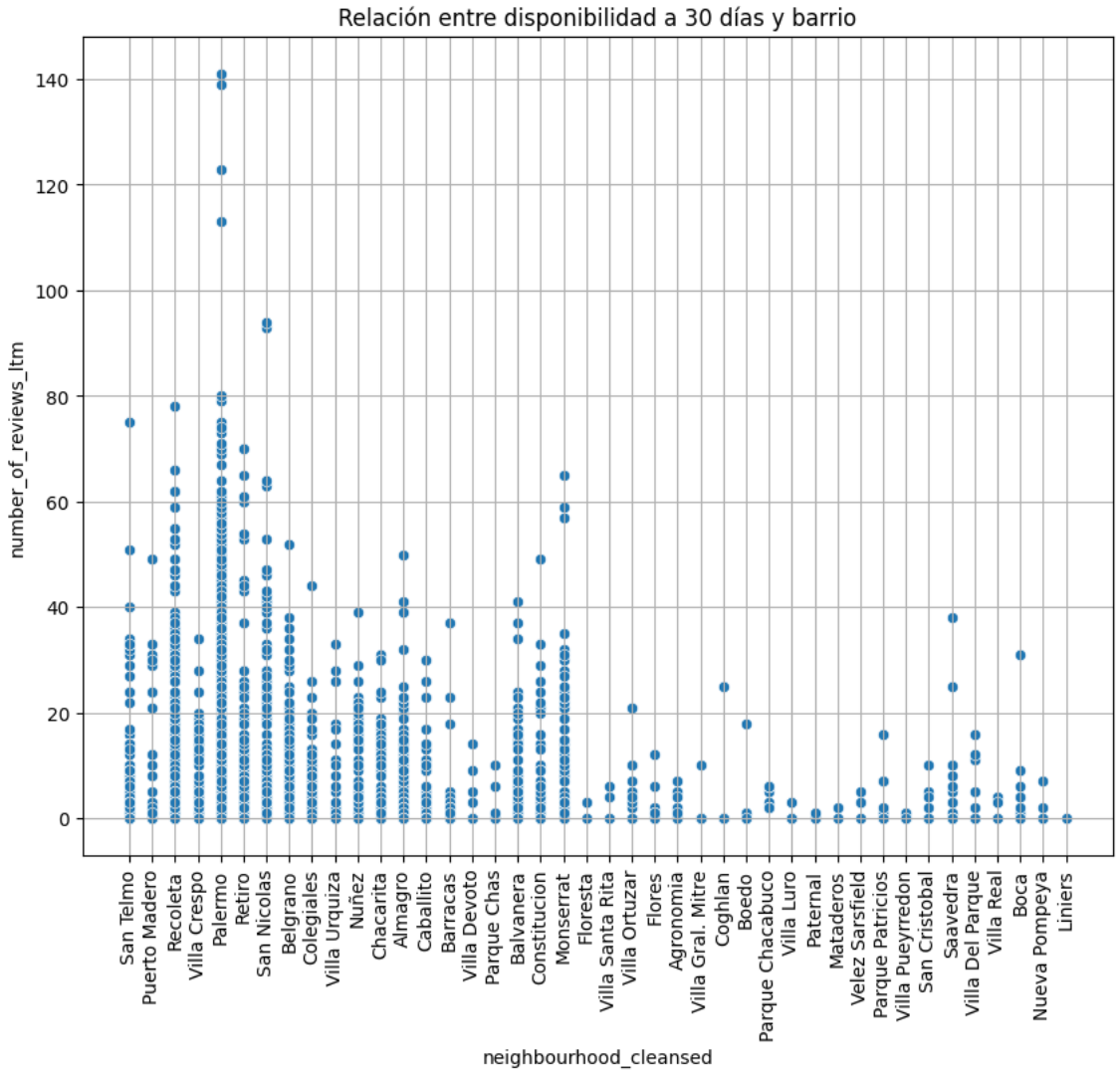

Figura 6. Diagrama de dispersion "neighbourhood_cleansed" vs. "number_of_reviews_ltm"

Se propone entonces el ejercicio de cruzar los tres diagramas de anteriores para generar una lista corta de barrios. Se escogen los barrios han mostrado menor disponibilidad a 30 días y mayor número de opiniones en el último año, simultáneamente:

- San Telmo
- Recoleta
- Palermo
- Retiro
- San Nicolás
- Belgrano
- Almagro
- Balvanera
- Constitución
- Monserrat

4.2 Mejores tipos de propiedades

Para determinar los mejores tipo de propiedades se emulará el análisis presentado para los barrios, pero esta vez enfocado al tipo de propiedad. Las gráficas generadas son las siguientes:

Disponibilidad a 30 días vs. Tipo de propiedad/habitación:

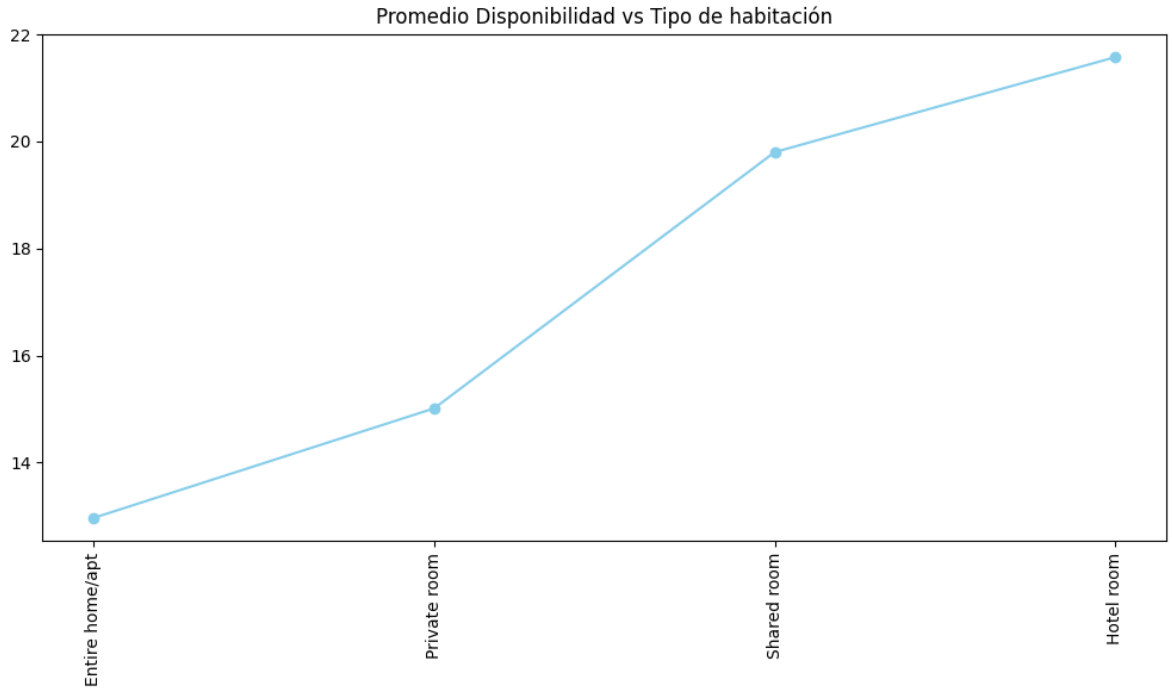

Figura 6. Promedio de la disponibilidad a 30 días por tipo de propiedad

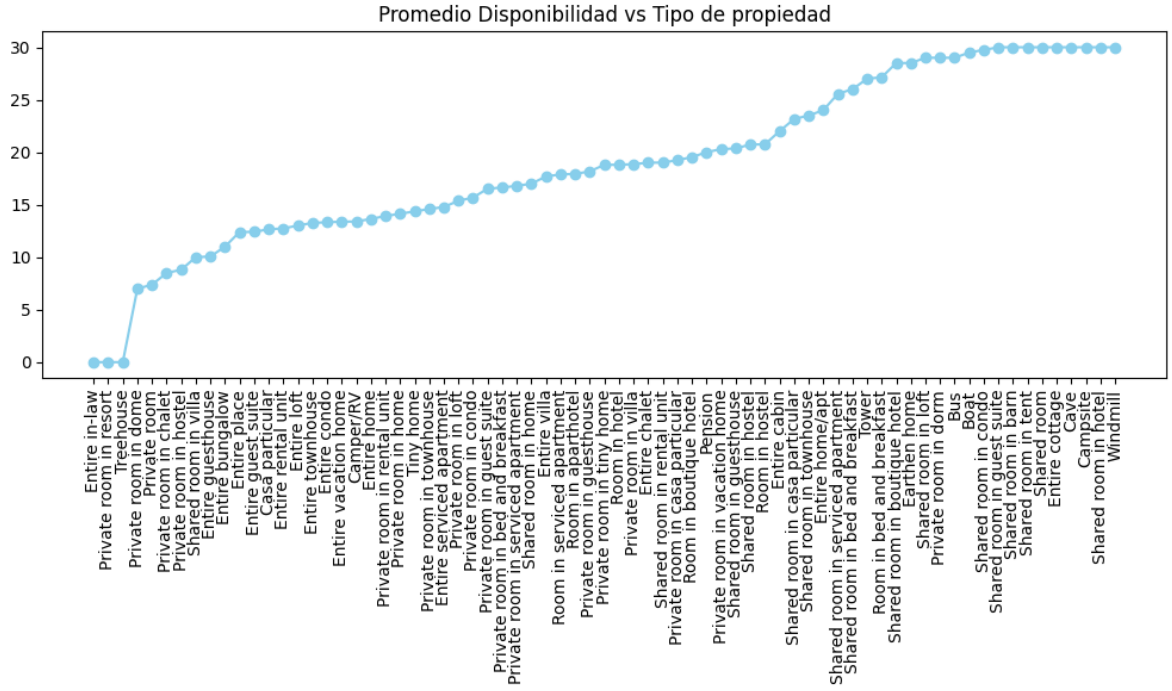

Figura 6. Promedio de la disponibilidad a 30 días por tipo de propiedad

Aquí el análisis se ha realizado con dos niveles de desagregación, el del tipo de habitación y el del tipo de propiedad. De manera conjunta, de estas dos gráficas se puede concluir que

durante el próximo mes las propiedades más ocupadas son aquellas que ofrecen cuartos privados en locaciones como chalets, villas, apartamentos o casas; y las que ofrecen propiedades enteras tales como apartamentos, casas, condominios y lofts.

Número de opiniones en el último año vs. Tipo de propiedad:

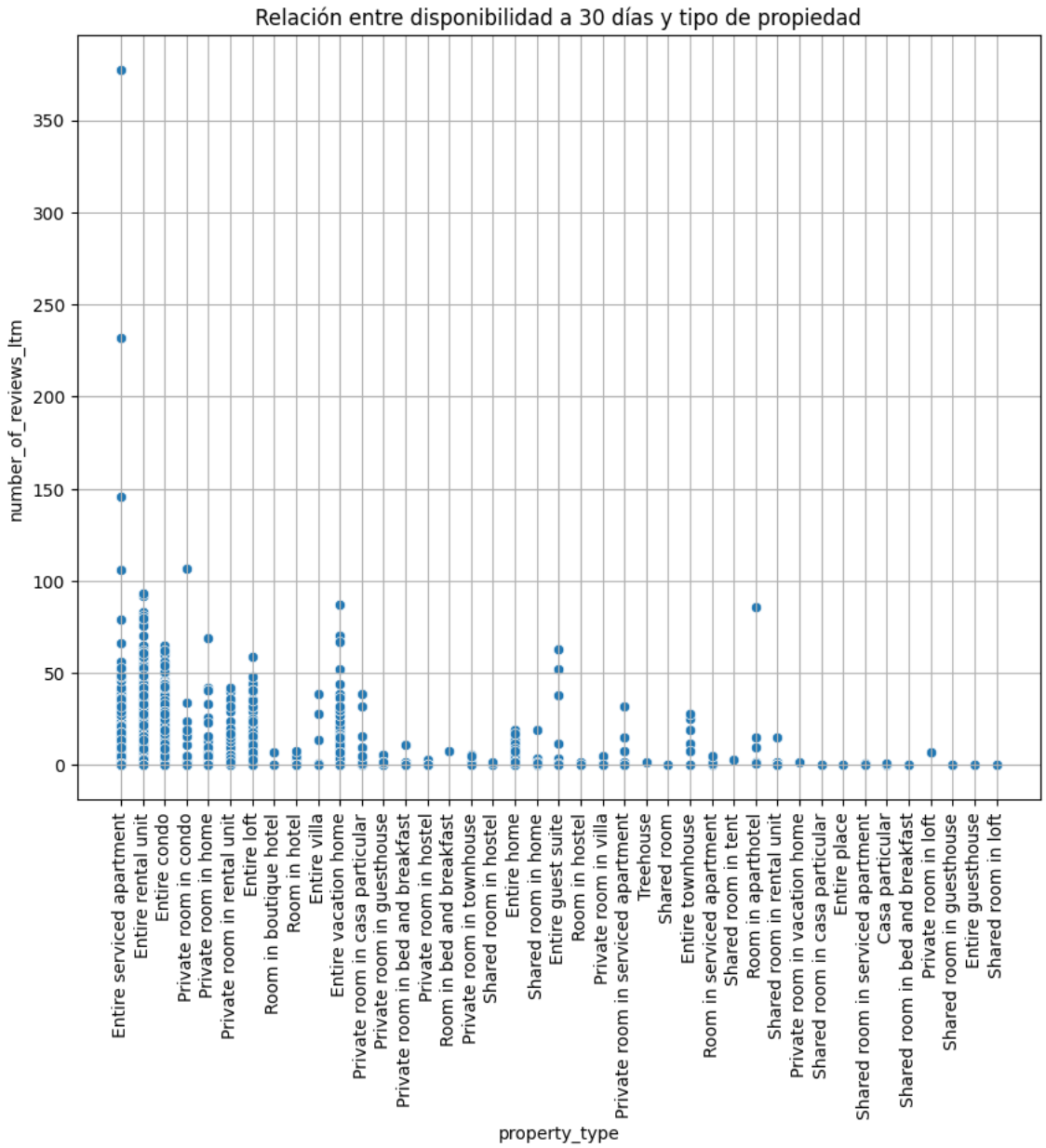

Figura 6. Diagrama de dispersion "neighbourhood_cleansed" vs. "property_type"

Al medir la popularidad de diferentes tipos de propiedad según el número opiniones, es posible ratificar que lo más buscado son las propiedades enteras como apartamentos amoblados y sin amoblar, condominios y lofts. Además, también son muy populares las habitaciones privadas en condominios, casas y apartamentos. En el otro extremo, las

habitaciones compartidas, en general, no resultan ser muy populares y reservadas en Airbnb.

4.3 Rangos de precios

Es necesario determinar ahora los rangos de precios sobre los cuales se deberían alquilar los inmuebles por noche. Estos rangos serán divididos según el tipo de habitación: propiedad entera y habitación privada. (No se considerarán las habitaciones mixtas por no estar dentro de lo más buscado y ocupado por los huéspedes, tal como se vio en la sección anterior).

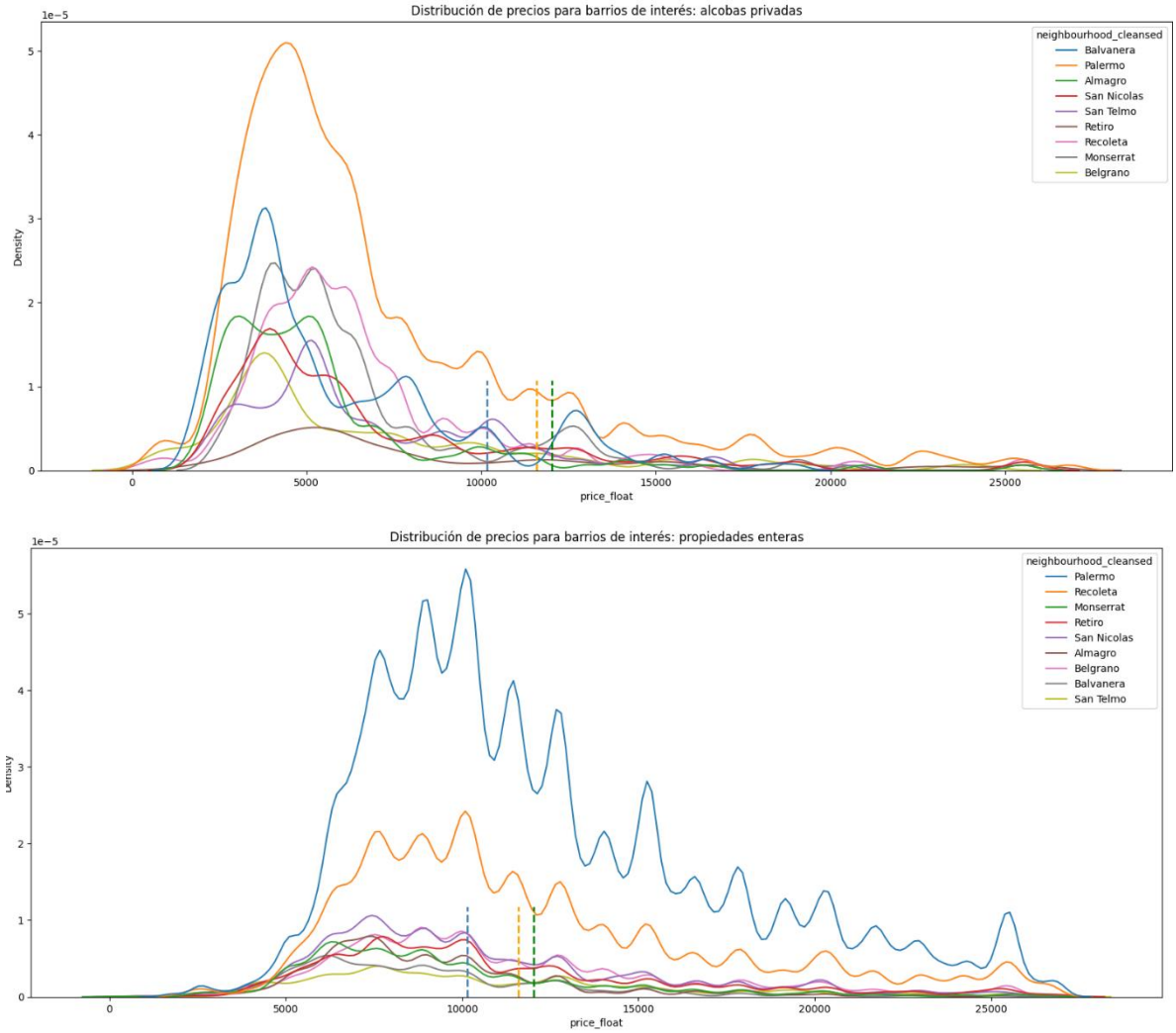

Figura 6. Distribución de precios por tipo de habitación y barrio de interés

Como es de esperar, los precios de habitaciones privadas son menores que los de una propiedad entera. Si se consideran las medias de los principales barrios, es posible decir que una habitación privada vale aproximadamente la mitad que un propiedad entera. Lo que lleva a pensar que si el anfitrión tiene propiedades enteras con 3 habitaciones o más, le convendría más alquilar las habitaciones separadamente. Claro está, esto dependerá

exactamente del número de metros cuadrados y el valor de cada propiedad, pero se puede ir construyendo una idea de la dinámica del negocio.

5. Generación de resultados

El capítulo anterior ha generado valiosos insights para los inversionistas. Aquí se compilarán a modo de conclusiones. Para más detalles, el potencial inversionista puede consultar las correspondientes secciones del capítulo 4.

¿Cuáles son los mejores barrios para invertir en la compra de propiedades de alquiler?

Según la disponibilidad de las propiedades a 30 días en el futuro y la popularidad de los barrios (medido por el número de opiniones en las propiedades) los barrios con mayor prospecto son:

- San Telmo
- Recoleta
- Palermo
- Retiro
- San Nicolás
- Belgrano
- Almagro
- Balvanera
- Constitución
- Monserrat

Todos estos barrios se caracterizan por ser iconos bonaerenses, estar en el centro de la movida cultural y ser los más visitados por los turistas y nómadas digitales.

¿Cuáles son los mejores tipos de propiedades/habitaciones para invertir?

Según el análisis realizado, se concluye que las propiedades enteras como apartamentos amoblados y sin amoblar, condominios y lofts son las más apetecidas. Además, también son muy populares las habitaciones privadas en condominios, casas y apartamentos.

¿Cuáles son las mejores tarifas para alquilar?

Las tarifas dependerán de si se trata de una propiedad entera o una habitación privada. En el primer caso, los rangos de precio por noche habituales están entre 2500 y 7500 ARS. Entretanto, para el segundo caso están entre 7500 y 16000 ARS. Además, tomando la media de las distribuciones de precio, es posible pensar que si un apartamento o casa tiene 3 habitaciones o más, será más rentable arrendar las habitaciones separadas que el apartamento entero.

En conclusión, los inversionistas tienen 9 barrios para escoger. Elegir uno de estos barrios dependerá del costo por metro cuadrado y el retorno de inversión. Con los datos disponibles no es posible calcular el retorno de inversión, pues sería necesario contar con los valores de compra/venta de las propiedades para estimar la inversión inicial. Más allá de eso, si el inversionista posee una cantidad de dinero suficiente, puede optar por comprar

apartamentos o casas grandes, para rentar habitaciones privadas por separado. Por otra parte, si el inversionista tiene una limitación de dinero, podría optar por propiedades pequeñas de una o dos habitaciones para arrendarlas enteras. Los apartamentos tipo loft son una buena opción en estos casos.