Taller 1 – Análisis Exploratorio de Datos

Kevin Yesid Infante Hernández
Ciencia de Datos Aplicada
Universidad de los Andes, Bogotá, Colombia
{ky.infante3022}@uniandes.edu.co
Fecha de presentación: agosto 29 de 2023

Tabla de contenido

1 Introducción

En este trabajo se realizará un análisis exploratorio de un dataset de propiedades listadas en Airbnb para la ciudad de Buenos Aires, de modo que se puedan descubrir insights, tendencias y oportunidades que le permitan a las personas interesadas en invertir en propiedades para alquiler a través de la plataforma Airbnb en Buenos Aires tomas decisiones basadas en datos sobre cuáles tipos de propiedades son las adecuadas para lograr una mayor rentabilidad, entendida como una alta tasa de ocupación y un precio por noche que los usuarios estén dispuestos a pagar.

2 Entendimiento inicial de datos

El Dataset de propiedades de Airbnb en la ciudad de Buenos Aires cuenta con 26,204 registros y 75 atributos, los cuales están ordenados de la siguiente manera: 9 atributos correspondientes a la identificación de la propiedad en Airbnb, tales como id, listing_url, name, description, neighborhood_overview), 17 atributos correspondientes a los datos del huésped (host_id, host_url, host_name, host_acceptance_rate, entre otros), 5 atributos de la ubicación del inmueble (neighbourhood, neighbourhood_cleansed, neighbourhood_group_cleansed, latitude, longitude), 9 atributos de las condiciones del inmueble, entre los cuales destacan para este análisis property_type, room_type, accommodates, amenities y price, 15 atributos correspondientes a la reserva, tales como minimum_nights, maximum_nights, availability_365, entre otros), 12 atributos de las calificaciones del inmueble, como first_review, last_review, review_scores_rating, review_scores_communication y review_scores_value, 2 atributos adicionales del inmueble (license, instant_bookable), 4 atributos de información calculada automáticamente, iniciando por calculated_host_listings_count y terminando en calculated_host_listings_count_shared_rooms, y 1 atributo adicional de calificaciones del inmueble (reviews_per_month).

El dataset no tiene inmuebles duplicados, y cuenta principalmente con variables cualitativas nominales como id (identificación del inmueble), neighbourhood_cleansed (barrio), property_type (tipo de propiedad), entre otras, sin embargo, cuenta también con variables cuantitativas continuas como price (precio por noche) y aquellas correspondientes a la calificación del inmueble, y variables cuantitativas discretas como accommodates (capacidad de huéspedes del inmueble), host_listings_count (número de inmuebles del huésped), entre otras. No se identificaron variables cualitativas nominales.

Dado que este análisis está orientado a encontrar aquellos atributos de los inmuebles que sean más relevantes para obtener mejores ganancias y lograr una tasa de ocupación tan alta como sea posible, los atributos principales que se deben tomar para este análisis son el precio por noche (price) y la disponibilidad del inmueble, ya que conociendo su disponibilidad se puede calcular su tasa de ocupación. Se tienen 4 atributos de disponibilidad en días del inmueble, que corresponden a su disponibilidad en días en los siguientes 30, 60, 90 y 365 días. Para este trabajo, se tomará la disponibilidad en días en los siguientes 365 días (availability_365), con el fin de tener presentes tantos días como sea posible al momento de calcular la tasa de ocupación y no dejar por fuera en el cálculo reservas que se hayan realizado con una gran anticipación.

Por otra parte, con el fin de determinar los atributos más relevantes para obtener mejores ganancias y lograr una mayor tasa de ocupación, se analizarán los cambios en el precio por noche y disponibilidad con respecto al barrio (neighbourhood_cleansed), tipo de habitación (room_type), tipo de propiedad (property_type) y cantidad de huéspedes (accommodates), a la vez que se analizará la influencia de las calificaciones de los inmuebles sobre el precio.

A continuación se muestra el resultado del análisis univariado sobre cada una de las variables seleccionadas para este análisis:

2.1 Atributos principales

Precio (price)

Es una variable cuantitativa continua. Se entiende que corresponde al precio por noche y por persona por alquiler de una propiedad. En la siguiente tabla se muestran sus estadísticas básicas.

Métrica	Valor
Count	26,204.00
Mean	17,529.33
Std	175,401.77
Min	175.00
5%	4,331.00
25%	7,406.50
50%	10,190.00
75%	15,286.00
95%	33,976.75
Max	25,295,088.00

Tabla 1. Estadísticas básicas de precio

La alta desviación estándar que tiene esta variable se ve explicada por la presencia de valores de precio muy grandes con respecto a la media, que de hecho son máximo el 5% de la muestra. A continuación se muestra el diagrama de cajas y bigotes correspondiente a esta variable.

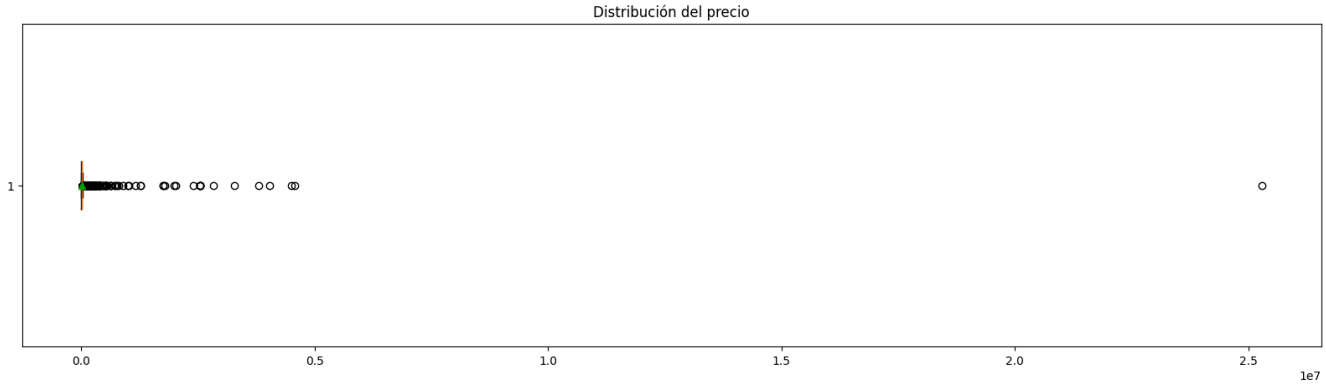

Figura 1. Distribución de precio

Se puede ver que existen propiedades con un valor mucho mayor al de las demás: mientras la gran mayoría no sobrepasa los 100,000, estas propiedades tienen un valor del orden de 25,000,000, por lo cual podrían tratarse como un outlier. También se encontró que todas las propiedades tienen precio asignado. Descartando los 76 outliers, se tiene el siguiente diagrama de bigotes. De este punto en adelante se descartan los 76 outliers.

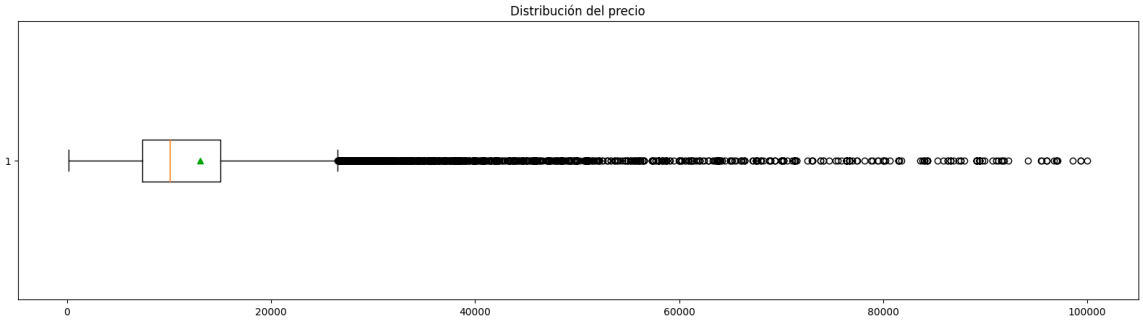

Figura 2. Distribución de precio sin el outlier

Disponibilidad 365 (availability_365)

Es una variable cuantitativa discreta, que corresponde a la disponibilidad de la habitación en días en los siguientes 365 días. En la siguiente tabla se muestran sus estadísticas básicas.

Métrica	Valor
count	26,008.00
mean	213.88
std	127.33
min	0
5%	0
25%	89.00
50%	226.00
75%	341.00
95%	365.00
max	365.00

Tabla 2. Estadísticas básicas de Disponibilidad 365

Tal como se espera, los valores mínimos y máximos de esta variable son 0 y 365, por lo que no parecen haber valores fuera de los rangos intercuartílicos, lo cual se confirma con el diagrama de cajas y bigotes.

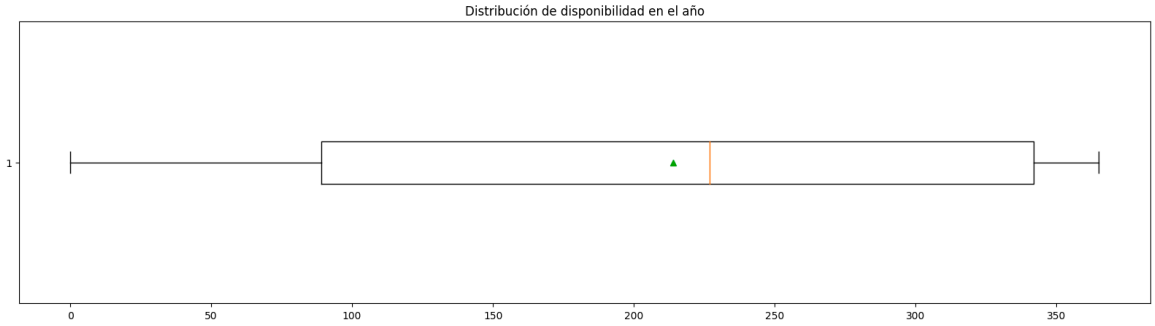

Figura 3. Distribución de disponibilidad 365

2.2 Atributos dependientes

Barrio (neighborhood_cleansed)

Es una variable categórica nominal, que corresponde al barrio donde se ubica la propiedad. A continuación se muestra el gráfico de frecuencias absolutas y de Pareto.

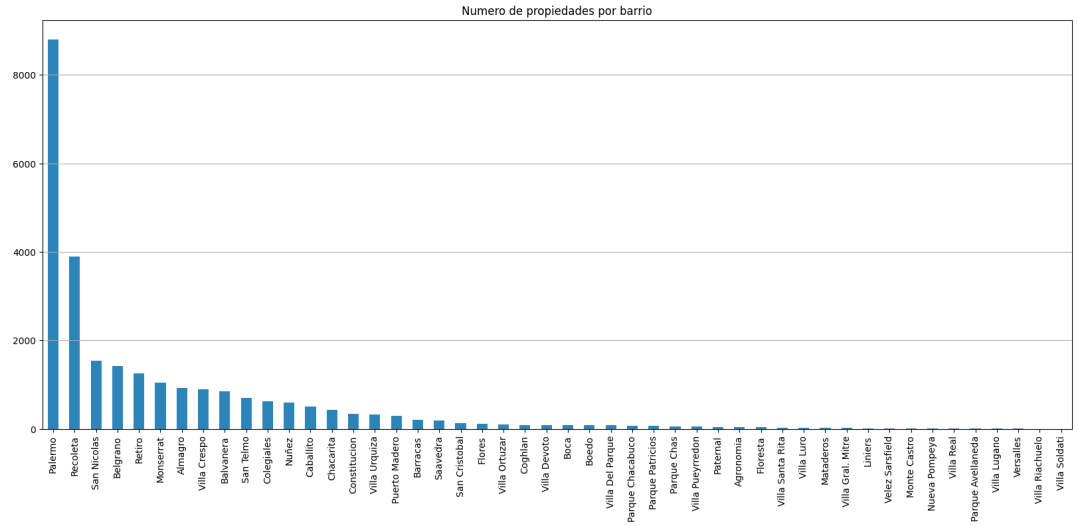

Figura 4. Número de propiedades por barrio

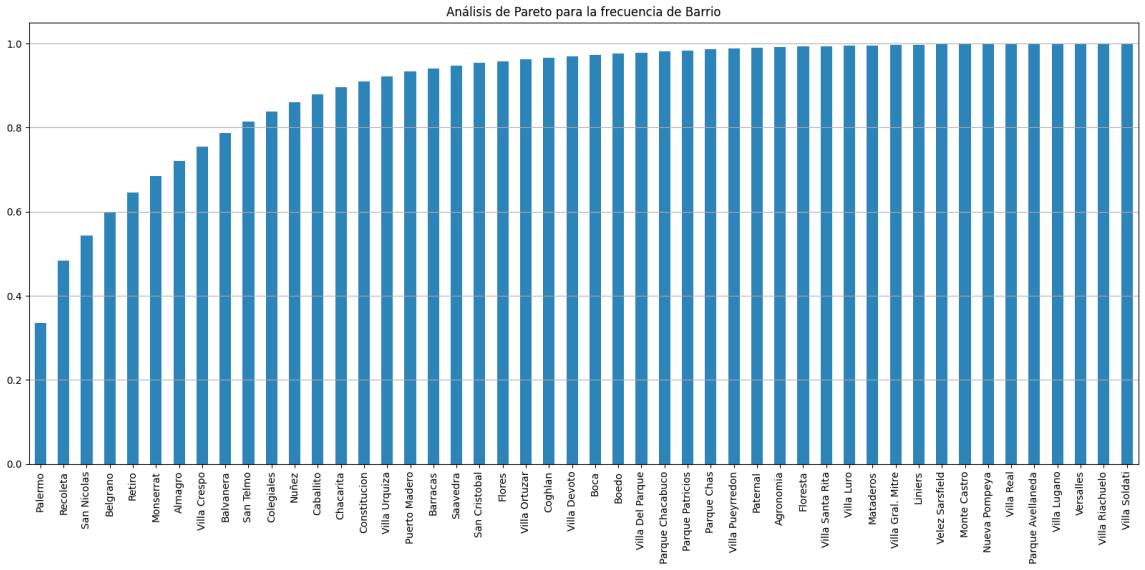

Figura 5. Análisis de Pareto

Se ve que la gran mayoría de propiedades, alrededor del 33.6%, se ubican en Palermo, duplicando las que se encuentran en el siguiente barrio, Recoleta, que son alrededor del 14.8%. Es probable que esta distribución de propiedades obedezca a la densidad de población de cada barrio. Se tiene un total de 48 barrios, y todas las propiedades tienen un barrio asignado.

Tipo de habitación (neighborhood_cleansed)

Es una variable categórica nominal. A continuación se muestra el gráfico de frecuencias absolutas de esta variable.

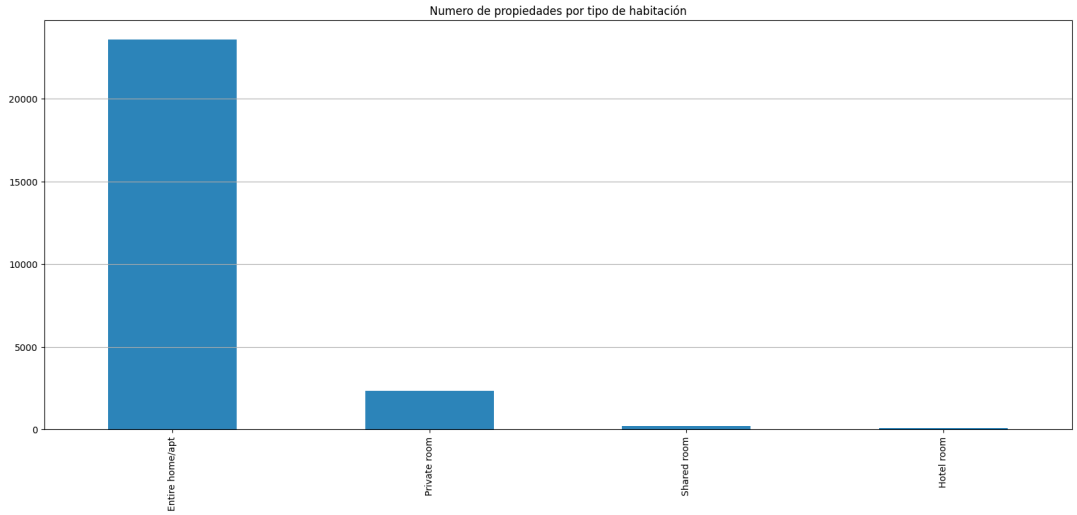

Figura 6. Número de propiedades por tipo de habitación

Se ve que la gran mayoría de habitaciones, alrededor del 90%, son de tipo “apartamento o casa completos”, mientras que el 9% son de tipo “habitación privada”, y las restantes son habitación compartida (0.7%) y habitación de hotel (0.3%). Todas las propiedades tienen un tipo de habitación asignado.

Tipo de propiedad (property_type)

Es una variable categórica nominal. A continuación se muestra el gráfico de frecuencias absolutas.

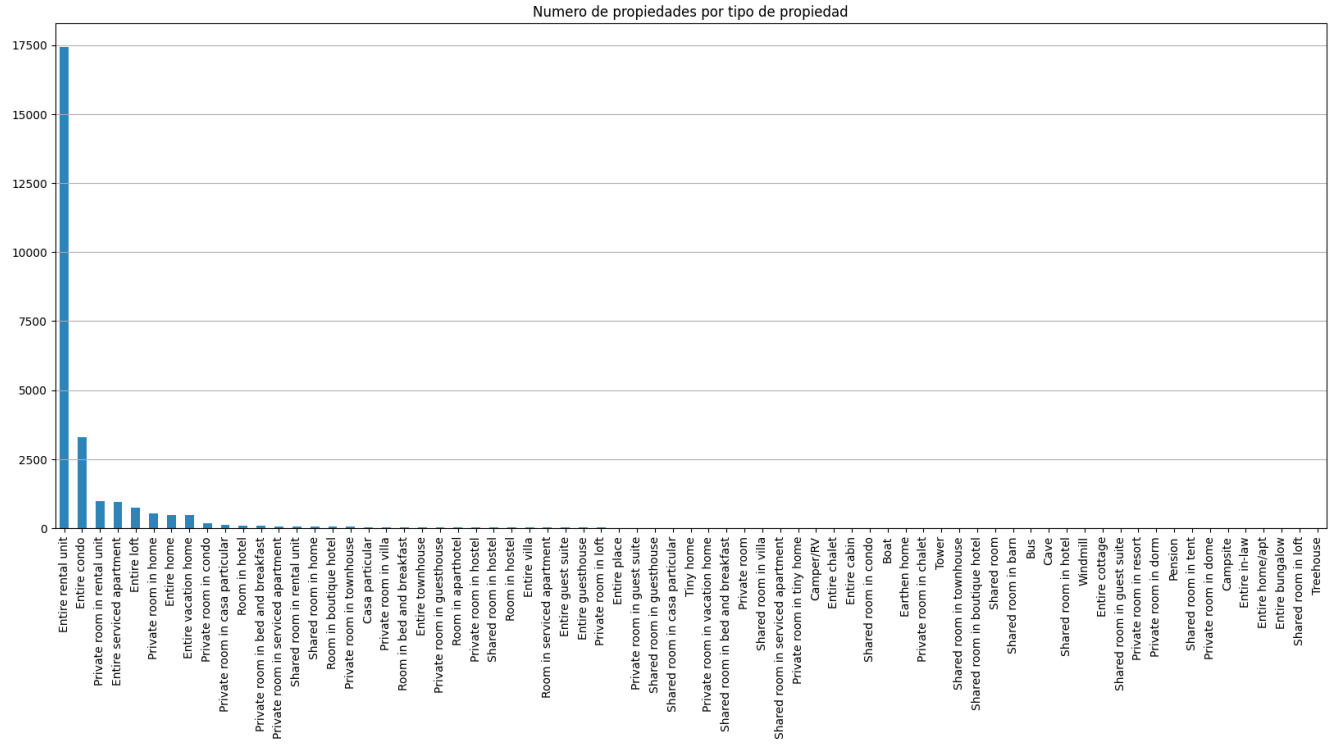

Figura 7. Número de propiedades por tipo de propiedad

Al igual que se observó en “tipo de habitación”, se puede observar que en la gran mayoría de servicios (66.4%) se alquila la unidad completa, seguida de “condo completo” (12.6%) y habitación privada en una unidad rentada (3.7%). Estas tres categorías ya representan más del 80% del total de propiedades.

Capacidad de huéspedes (accommodates)

Es una variable cuantitativa discreta, que corresponde a la capacidad máxima de huéspedes en una propiedad. En la siguiente tabla se muestran sus estadísticas básicas

Métrica	Valor
count	26,008.00
mean	2.85
std	1.41
min	1
25%	2
50%	2.00
75%	4.00
max	16.00

Tabla 3. Estadísticas básicas de capacidad de huéspedes

Según indican las métricas, todas las propiedades tienen registrada una capacidad de al menos una persona y máximo 16, lo cual indica que esta variable tiene datos completos y consistentes. Ahora bien, al ser una variable discreta con valores completos y consistentes, se puede revisar la cantidad de propiedades por cada valor de capacidad de huéspedes, gráfica que se muestra a continuación.

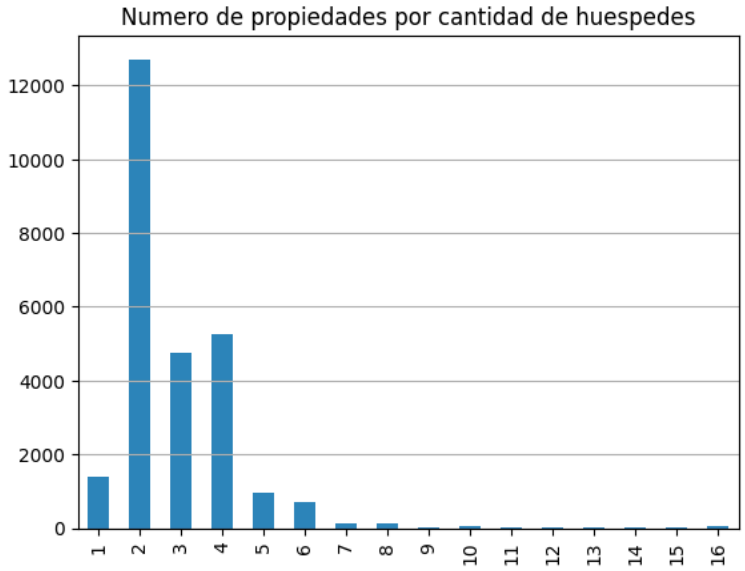

Figura 8. Número de propiedades por cantidad de huéspedes

Se ve que la moda de esta variable es 2 por un gran margen, es decir, la mayoría de propiedades tiene una capacidad de 2 huéspedes.

3 Estrategia de análisis

Dado que el objetivo de negocio del análisis exploratorio de datos es descubrir insights que le permitan a los inversionistas tomar decisiones basadas en datos sobre en cuál tipo de propiedades deberían invertir, se toman como variables principales de análisis el precio (price) y la disponibilidad del inmueble en los siguientes 365 días (availability_365), gracias a la cual, tal como se mencionó previamente, se puede calcular la ocupación de la habitación. La estrategia a seguir en este análisis es encontrar los valores de barrio (neighbourhood_cleansed), tipo de habitación (room_type), capacidad de huéspedes (accomodates) y tipo de propiedad (property_type) que tengan viviendas con los precios más altos y el índice de ocupación más alto, de modo que los inversionistas sepan en qué tipo de propiedades invertir para obtener mayores ganancias. Además, se analizará si existe una correlación entre la calificación del inmueble y su precio, de modo que se pueda determinar si esta variable es relevante para obtener mayores ganancias.

Para esto, se realizará un análisis bivariado en el cual se hallen los barrios, tipos de habitación, capacidades de huéspedes y tipos de propiedades cuyo promedio de precio sea el mayor y cuyo promedio de disponibilidad sea el menor. Se elige el promedio como medida ya que es de interés para este análisis tener en cuenta la influencia tanto de las propiedades con un precio muy alto como con un precio muy bajo. Posteriormente, se realizará un análisis gráfico multivariado usando un gráfico de burbujas cuyas dimensiones son el precio, el índice de ocupación entendido como 365 – availability_365, y la cantidad de propiedades por barrio y tipo de propiedad, de modo que el inversionista disponga de una herramienta gráfica para encontrar en un vistazo aquellas propiedades cuyo precio e índice de ocupación sean mayores, a la vez que pueda analizar otras variables que sean de su interés.

4 Desarrollo de la estrategia

4.1 Correlación entre calificación del inmueble y precio - Scatterplot

A continuación se muestra la correlación usando un gráfico scatterplot:

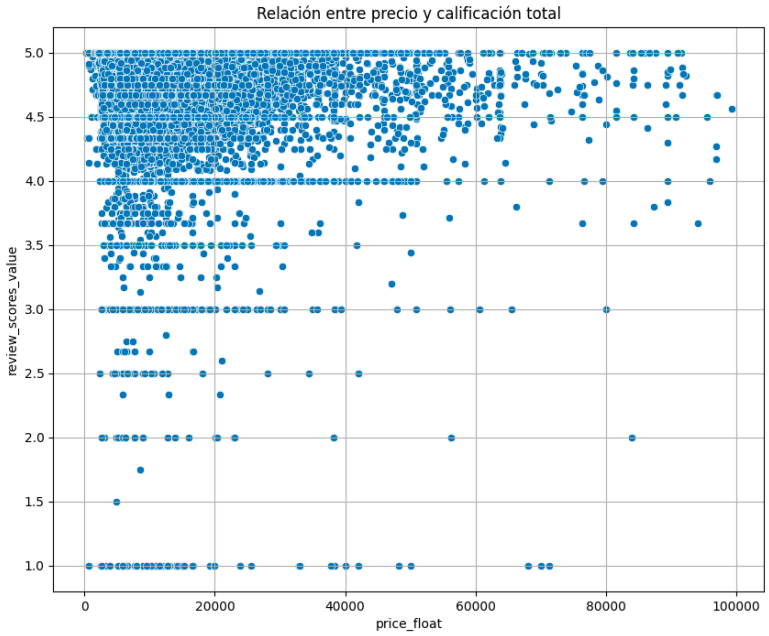

Figura 9. Scatterplot precio vs calificación del inmueble

Se puede observar que no se puede deducir una correlación entre el precio y la calificación del inmueble, por lo cual se descarta tener en cuenta la calificación del inmueble como un posible criterio que pueda permitir obtener mayores ganancias.

4.2 Precio y disponibilidad por barrio

A continuación se muestra el top 5 de barrios con un promedio mayor de precio junto con la disponibilidad 365, y el top 6 de barrios con una menor disponibilidad (mayor ocupación) promedio junto con su precio.

Barrio	Precio Promedio	Disponibilidad	Frecuencia relativa
Puerto Madero	28,845.26	220.05	1.13%
Villa Real	17,647.38	127.38	0.03%
Retiro	14,985.57	219.80	4.79%
Floresta	14,775.78	194.78	0.14%
Palermo	14,650.66	212.89	33.53%

Tabla 4. Top 5 barrios con mayor promedio de precio

Barrio	Precio Promedio	Disponibilidad	Frecuencia relativa
Villa Soldiati	4,000.00	-	0.38%
Villa Riachuelo	3,826.67	30.00	1.15%
Versalles	8,348.20	123.00	1.92%
Villa Real	17,647.38	127.37	3.08%
Parque Avellaneda	4,511.86	168.14	2.69%
Villa Gral. Mitre	10,391.80	182.60	7.69%

Tabla 5. Top 6 barrios con menor disponibilidad 365

En la siguiente gráfica de burbujas, la cual es interactiva y se encuentra disponible en el notebook, el inversionista puede, en un vistazo, encontrar los barrios con un mayor precio por noche e índice de ocupación promedio, a su vez que puede comparar cuantas propiedades hay en un barrio con respecto a los demás. La ocupación en el año se calcula como 365 – availability_365, y se muestra en esta gráfica para dejar las propiedades con mayor ocupación, que en esta dimensión son las de más interés para los inversionistas, en la parte superior de la gráfica, mientras que las propiedades con mayores precios se ubican en la parte derecha.

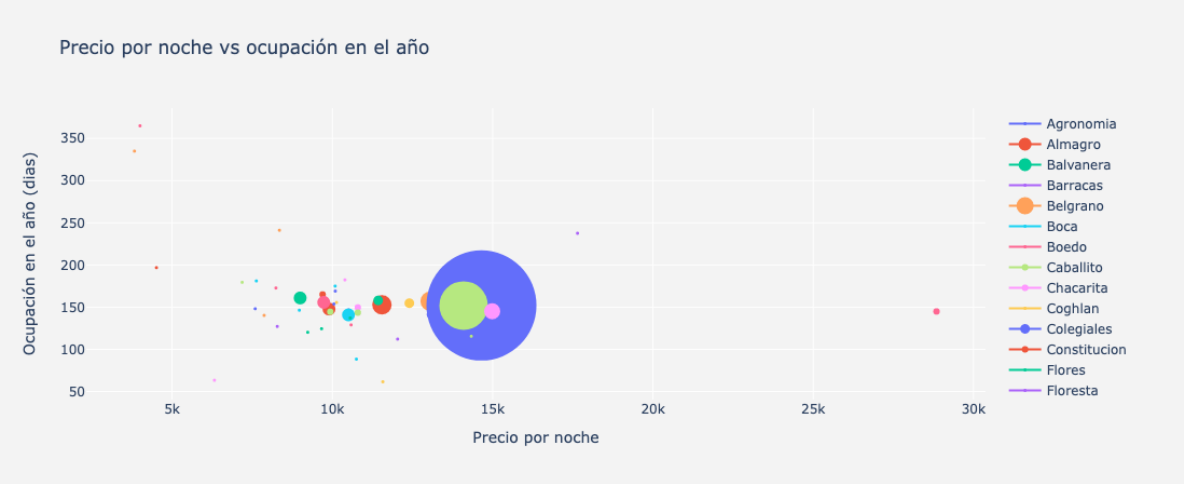

Figura 10. Gráfico de burbujas precio vs ocupación en el año vs frecuencia relativa

Se observa en las tablas 4 y 5 que los barrios con un mejor promedio por precio son Puerto Madero, Villa Real y Retiro, de los cuales el único que se encuentra en el top 6 de barrios con menor disponibilidad (y por ende mayor ocupación) es Villa Real, por lo cual este barrio resulta en una alternativa interesante de inversión.

4.3 Precio y disponibilidad por tipo de habitación

A continuación se muestra el promedio de precios y disponibilidad 365 por tipo de habitación, ordenado de mayor a menor por promedio de precio. La disponibilidad 365 no varía en gran medida de un tipo de habitación a otra.

Tipo de habitación	Precio Promedio	Frecuencia relativa	Disponibilidad
Hotel room	14,768.51	0.3768	276.02
Entire home/apt	13,568.79	89.9723	214.24
Private room	8,522.60	8.8819	204.47
Shared room	5,204.41	0.7689	250.18

Tabla 6. Promedio de precios y disponibilidad 365 por tipo de habitación

El precio por noche y persona de la renta de habitación de una unidad completa, sea un apartamento o casa, prácticamente dobla el precio por noche y persona de renta de una habitación privada o compartida, y el precio por noche y persona más alto, tal como es de esperarse, es el de una habitación de hotel.

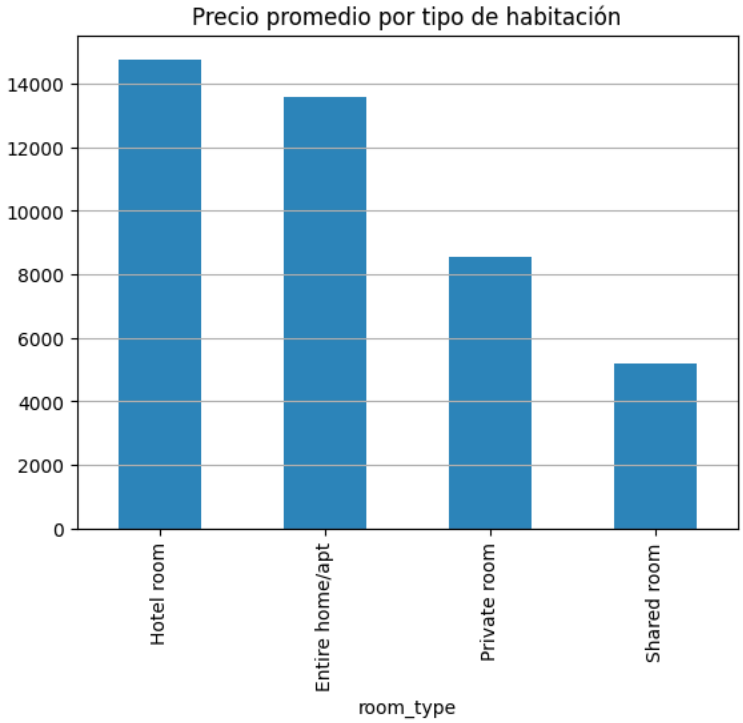

Figura 11. Gráfico de barras precio vs tipo de habitación

4.4 Precio y disponibilidad por capacidad de huéspedes

En las siguientes gráficas de barras se muestra el promedio de precio y disponibilidad por capacidad de huéspedes, ordenados de mayor a menor (precio) y menor a mayor (disponibilidad).

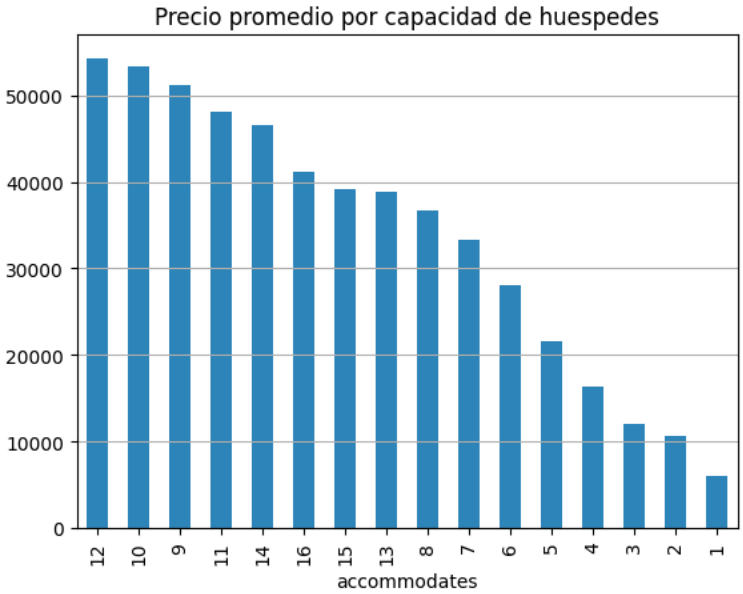

Figura 12. Gráfico de barras precio vs capacidad de huéspedes

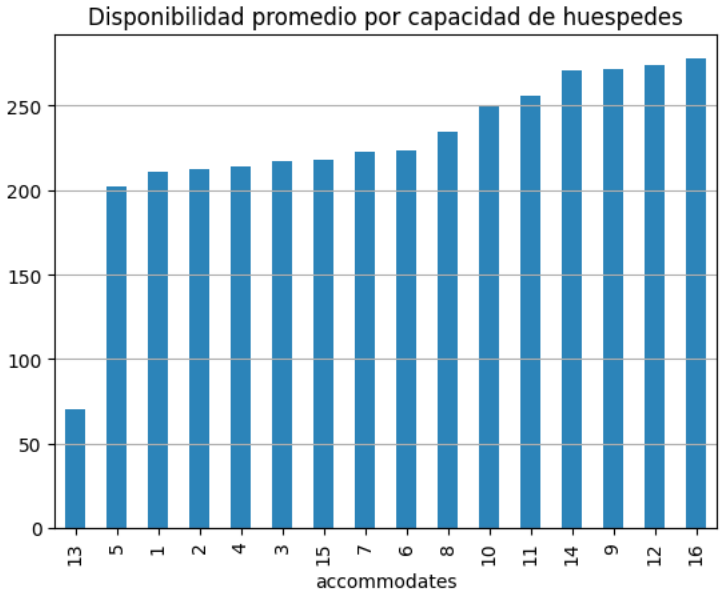

Figura 13. Gráfico de barras disponibilidad promedio vs capacidad de huéspedes

El comportamiento de esta variable tanto para el precio como para la disponibilidad es el esperado: las propiedades con mayor capacidad de huéspedes tienen en promedio precios por noche mayores y mayor disponibilidad, por lo cual el análisis de esta variable no aporta insights al objetivo de negocio.

4.5 Precio y disponibilidad por tipo de propiedad

Los siguientes gráficos de barras muestran el precio promedio por tipo de propiedad ordenado de mayor a menor y la disponibilidad por tipo de propiedad ordenada de menor a mayor.

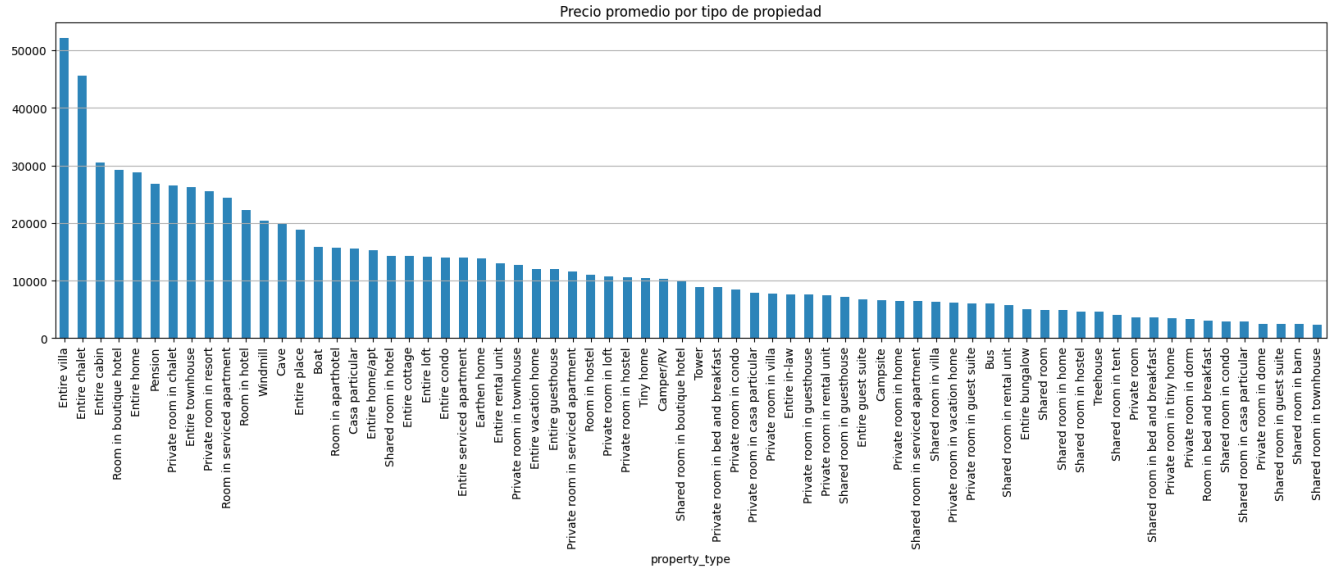

Figura 14. Gráfico de barras precio promedio vs tipo de propiedad

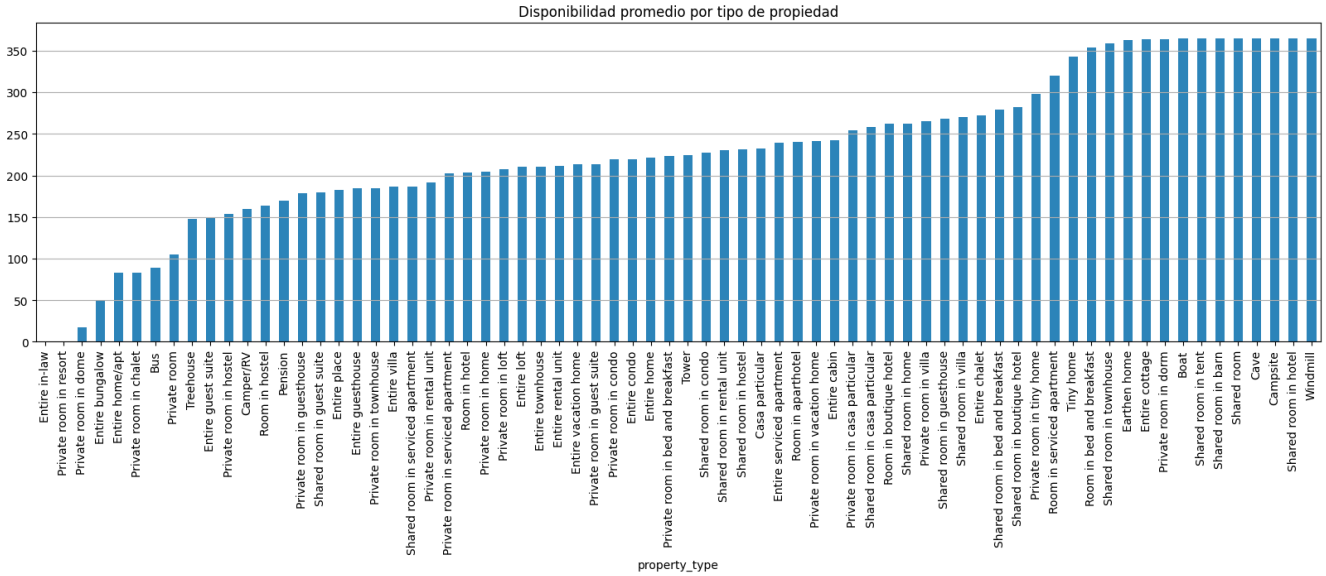

Figura 15. Gráfico de barras disponibilidad promedio vs tipo de propiedad

Con el fin de potenciar la toma de decisiones por parte de los inversionistas y tener una dimensión más clara de cuantas propiedades hay por tipo de propiedad, se preparó el siguiente gráfico interactivo de burbujas. La ocupación en el año se calcula como 365 – availability_365, y se muestra en esta gráfica para dejar las propiedades con mayor ocupación, que en esta dimensión son las de más interés para los inversionistas, en la parte superior de la gráfica, mientras que las propiedades con mayores precios se ubican en la parte derecha.

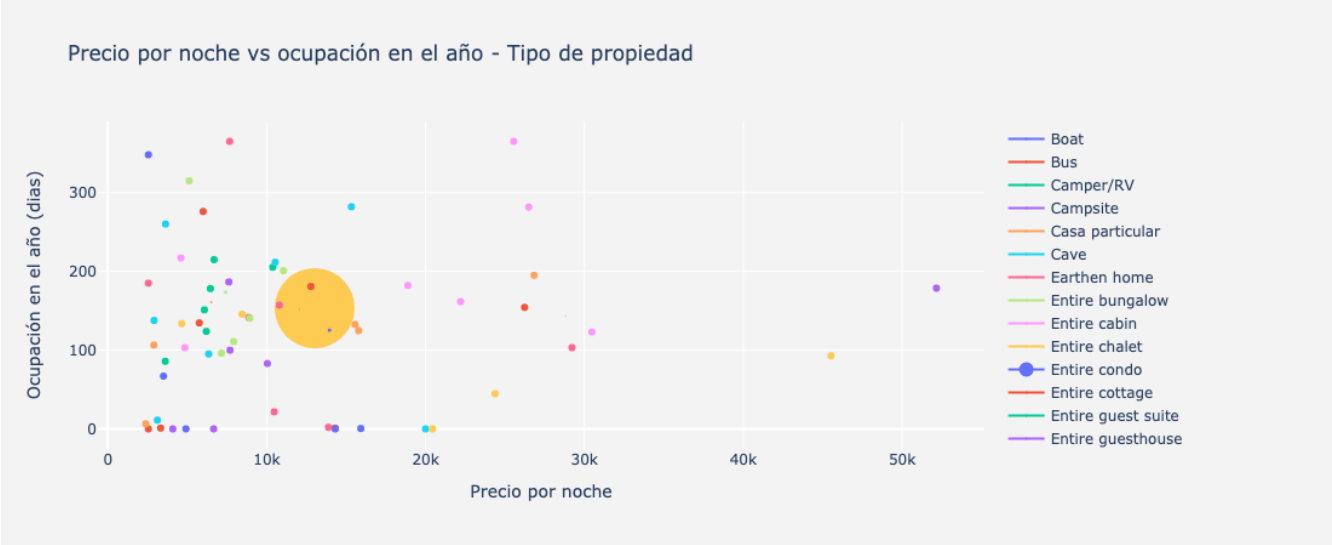

Figura 16. Gráfico de burbuja precio x noche por tipo de propiedad, ocupación en el año y frecuencia relativa

Esta variable muestra que en general, la renta de propiedades enteras tiene un precio por noche mayor al precio por alquiler de habitaciones compartidas, que equivale a lo mostrado por la variable “tipo de habitación”, y además muestra que la renta de propiedades privadas tiene un mayor índice de ocupación que la renta de propiedades compartidas.

5 Generación de resultados

De acuerdo con el análisis realizado, las variables más influyentes sobre el precio por noche y la ocupación promedio de los inmuebles son el barrio, el tipo de habitación y el tipo de propiedad, dado

que las otras dos variables analizadas, calificación y capacidad de huéspedes, o bien no están correlacionadas con el precio (calificación), o bien presentan comportamientos predecibles (capacidad de huéspedes), ya que una propiedad con una alta capacidad de huéspedes se espera que tenga un mayor precio por noche y una mayor disponibilidad.

En ese orden de ideas, analizando la información de las tablas 4 y 5 se ve que el barrio Villa Real es el segundo con un mayor promedio de precio por noche y el cuarto con un mayor índice de ocupación (menor disponibilidad), lo cual lo convierte en una alternativa altamente recomendable para invertir. Por otra parte, los barrios Puerto Madero y Retiro (1 y 3 con mayor precio por noche) tienen índices de ocupación por debajo de la media, lo cual los hace menos favorables que el barrio Villa Real. Otros barrios que pueden ser una alternativa interesante son Floresta y Palermo, ya que son el 4 y 5 con mayor precio por noche y su ocupación está por debajo de la media.

Por otra parte, se ve que la renta de una unidad completa (apartamento o casa) tiene un precio promedio por noche y persona superior a la renta de una habitación compartida o privada, lo cual lo hace la alternativa más rentable si se piensa en invertir en una propiedad particular. En caso contrario, en el cual se piense invertir en habitaciones de hotel, este es el tipo de habitación con un mayor precio promedio, sin embargo, no supera en gran medida el precio promedio por noche y persona de una unidad particular, tal como lo muestra la Figura 11. Además, la renta de propiedades privadas tiene en general mayores índices de ocupación, lo cual confirma que la renta de una unidad completa es la mejor opción en términos de rentabilidad e índice de ocupación.