Proyecto final

Curso de Especialización Mujeres en Tech II

Lina Chaaro
Eva da Costa Moñux
Emma Gallardo Ruiz
Ana Hernández Guardiet
Clara Hernández Sánchez
Natalia Hristova

ÍNDICE

1. Definir Dataset

El dataset sugerido es un extracto del listado de habitaciones disponibles en AirBnB. Se ha tomado un extracto con primeras 21278 filas para tener un volumen de datos manejable con los recursos disponibles. De todos estos datos, la gran mayoría se refiere a la ciudad de Madrid, España.

El dataset consta de los siguientes campos:

Título de campo	**Tipo de dato**	**Descripción**
Location	Cadena de caracteres	Ubicación en formato "País, Ciudad, Barrio"
Coordinates	Cadena de caracteres	Coordenadas Geográficas
Country	Cadena de caracteres	País
City	Cadena de caracteres	Ciudad
Updated.Date	Cadena de caracteres	Fecha de Actualización
Availability	Número (entero)	Disponibilidad del alojamiento
Rooms.rent.by.the.host	Número (entero)	Número total de habitaciones alquiladas por un mismo anfitrión.
Number.of.reviews.per.month	Número (doble)	Número de reseñas por mes
Date.last.review	Cadena de caracteres	Fecha de última reseña
Number.of.reviews	Número (entero)	Número total de reseñas
Minimum.nights	Número (entero)	Número mínimo de noches para reserva
Room.Price	Número (entero)	Precio por habitación
Room.type	Cadena de caracteres	Tipo de habitación
Neighbourhood	Cadena de caracteres	Barrio
Host.ID	Número (entero)	ID de anfitrión
Name	Cadena de caracteres	Nombre/Descripción de la habitación
Room.ID	Número (entero)	ID de habitación

2. Arquitectura y Validación de Datos

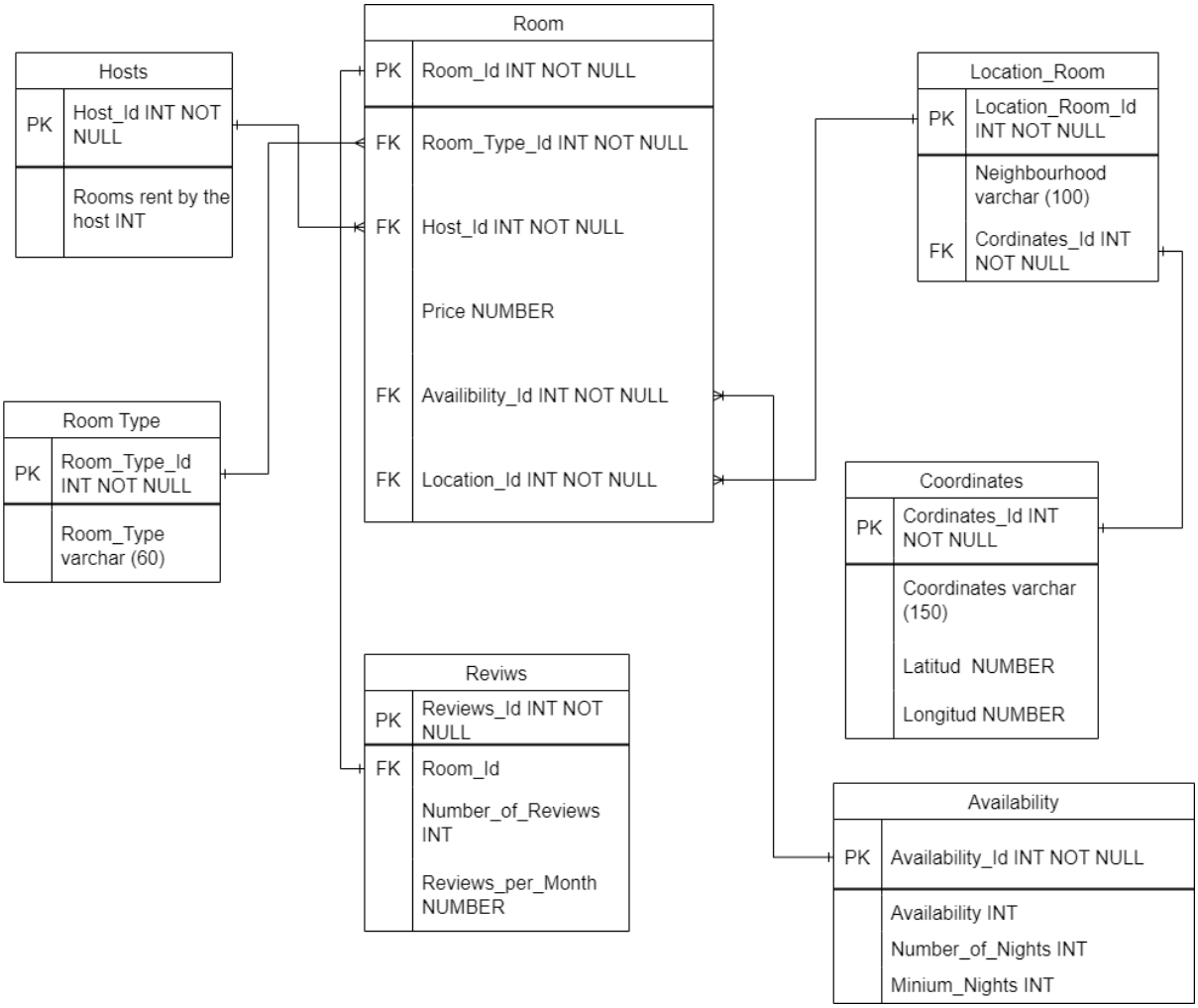
El objetivo principal de este apartado es la normalización, estructuración y organización de los datos de manera eficiente. Al normalizar los datos, buscamos eliminar redundancias, mejorar la integridad de los datos y facilitar futuras consultas y análisis. Este proceso es fundamental para garantizar que el análisis de datos sea preciso, manejable y escalable.

Para comenzar, se han creado varias tablas normalizadas. Todo el proceso se ha llevado a cabo mediante PostgreSQL a través de DBeaver. Cada tabla tiene un propósito específico, diseñado para separar la información en entidades distintas y relacionadas. Las tablas creadas son:

- Hosts. Esta tabla contiene información sobre los anfitriones, específicamente su “Host_Id” y el número de habitaciones que alquilan (“Rooms_Rent_By_The_Host”). Esta separación permite manejar la información de los anfitriones de manera independiente y facilita la consulta sobre cuántas habitaciones tiene cada anfitrión.
- Room_Type. Almacena los diferentes tipos de habitaciones (“Room_Type”), como ”Private room” o ”Entire home/apt”. Esta tabla permite normalizar los tipos de habitaciones y evita la redundancia al tener un listado estándar de tipos.
- Coordinates: En esta tabla, hemos dividido las coordenadas geográficas en “Latitud” y “Longitud”. Esto facilita las consultas geoespaciales y el análisis de localizaciones específicas. Cada conjunto de coordenadas se almacena de manera única para evitar redundancias y asegurar la integridad de los datos geográficos.
- Location_Room. Esta tabla almacena información sobre la localización: “Neighbourhood”, “City”, “Country”. Y una referencia a la tabla de coordenadas: “Coordinates_Id”. Esta división permite una mejor gestión de los datos de localización y facilita consultas basadas en localizaciones específicas.
- Availability. Incluye la disponibilidad de las propiedades (“Availability”), el número de noches mínimas requeridas (“Minimum_Nights”) y el número de noches disponibles (“Number_Of_Nights”). Al separar estos datos, podemos gestionar mejor la disponibilidad y las restricciones de reserva.
- Room. Contiene la información principal de las habitaciones, incluyendo referencias a otras tablas (“Room_Type_Id”, “Host_Id”, “Availability_Id”, “Location_Id”). Esto asegura que cada habitación esté correctamente relacionada con su tipo, anfitrión, disponibilidad y localización.
- Reviews. Esta tabla almacena las reseñas de cada habitación (“Number_Of_Reviews”, “Reviews_Per_Month”). Al separar las reseñas en una

tabla independiente, se facilita el análisis del rendimiento de las habitaciones en función de las evaluaciones de los huéspedes.

La estructura de las tablas se puede observar en el siguiente diagrama:



Para proporcionar un pequeño ejemplo del funcionamiento de las tablas creadas, primero se creó y pobló una tabla temporal, “airbnb_listings”, con una pequeña selección de los datos originales del dataset. A partir de la tabla “airbnb_listings” los datos fueron poblados en las tablas normalizadas.

No todas las columnas del dataset original fueron incluidas en las tablas normalizadas. Columnas como “name”, “date_last_review”, “updated_date” y “location” no fueron normalizadas ya que no son críticas para la estructura relacional que se buscaba. Estos campos pueden mantenerse en la tabla “airbnb_listings” para consultas detalladas, pero no son necesarios para la relación de las entidades principales en las tablas normalizadas.

Por último se realizó una pequeña consulta para obtener qué tipo de habitaciones ofrece cada host, mediante la cual se puede comprobar que las tablas normalizadas funcionan correctamente.

3. Análisis Exploratorio

En este punto queremos realizar un análisis exploratorio, es decir, explorar los datos obtenidos para la extracción de conclusiones sobre los mismos mediante gráficos y estimadores estadísticos.

Este estudio estadístico lo llevaremos a cabo en R.

3.1 Optimización de Datos

Primeramente se ha realizado una revisión de los datos cargados para optimizar el conjunto, descartando información redundante, verificando los formatos numéricos, de fechas y coordenadas geográficas, separando los campos necesarios y normalizando los valores.

Para nuestro caso concreto:

- Extraemos los datos correspondientes únicamente a la ciudad de Madrid.
- Eliminamos la columna "Name", ya que no aporta información nueva y relevante al análisis, sino que cada anuncio se puede identificar con el campo "Room.ID".
- Eliminamos la columna "Updated.Date", debido a que para el dataset actual, contiene la misma información para todas las filas.
- Eliminamos la columna "Location" porque contiene información redundante, ya reflejada en las columnas "Country", "City" y "Neighbourhood".
- Eliminamos las columnas "Country" y "City" porque es información redundante, dado que hemos extraído la información para una ciudad concreta: Madrid y resulta igual para todos nuestros datos. El campo Neighbourhood, sin embargo, se mantiene relevante.
- Convertimos el campo "Date.last.review" a formato fecha.
- Separamos la columna "Coordinates" en dos columnas nuevas llamadas "Latitud" y "Longitud", y las convertimos a formato numérico.
- Comprobamos si hay duplicados en los barrios reflejados en el dataset y comprobar que no haya nombres duplicados o mal escritos, pero como no se han encontrado, no se ha realizado ninguna transformación en este sentido.

A continuación, se han imputado los valores nulos y se han detectado outliers:

- Valores nulos se han encontrado 5.400 casos para:
    - "Number.of.reviews.per.month": en este caso se ha optado por cambiar los valores nulos por 0, ya que cuando no se tiene dato del número del número de reseñas mensuales es porque el número de reseñas es 0.
    - "Date.last.review": es para las mismas filas que en el caso anterior, pero aquí se ha optado por buscar cual es la primera fecha de una reseña

(04-08-2012) y poner una fecha anterior para no dejar los campos vacíos, en este caso, se ha puesto 31-12-2011, de esta forma sabemos que todos los que tengan fecha de 2011, en realidad no tienen ninguna reseña.

3.2 Exploración de variables de interés

En base a los datos proporcionados, se ha decidido estudiar si las variables que tenemos influyen en el precio de las habitaciones de Madrid, es decir, nos interesa estudiar lo siguiente:

- Precio en funcion de barrio
- Precio en función del tipo de alojamiento
- Precio en función del número mínimo de noches
- Precio en función del número de reseñas
- Precio en función del número de reseñas mensuales
- Precio en función de la fecha de última reseña
- Precio en función del número de alojamientos alquilados por un mismo host
- Precio en función de la disponibilidad
- Precio en función de la distancia al centro de la ciudad (Sol)

Para esto último hace falta calcular la distancia al centro de la ciudad en función de las coordenadas que se nos proporcionaban. Durante este proceso se ha obtenido una nueva variable, “Distancia”, representando la distancia que hay desde cada habitación en alquiler hasta el centro. Se han tomado como coordenadas de referencia del centro el promedio de las coordenadas de todas las habitaciones ubicadas en el barrio de Sol.

3.3 Estadística de variables

Aquí, primeramente se ha sacado un resumen para todos nuestros campos:

```
    Room.ID             Host.ID                Neighbourhood              Room.type       Room.Price    Minimum.nights     Number.of.reviews
Min.   :    6369   Min.   :     5154   Embajadores: 2559   Entire home/apt:12704   Min.   :   0   Min.   :   1.000   Min.   :  0.00
1st Qu.:18046636   1st Qu.: 25459146   Universidad: 2059   Hotel room     :  221   1st Qu.:  36   1st Qu.:   1.000   1st Qu.:  0.00
Median :28823836   Median : 95404148   Palacio    : 1490   Private room   : 7998   Median :  60   Median :   2.000   Median :  6.00
Mean   :26948459   Mean   :120532912   Sol        : 1366   Shared room    :  332   Mean   : 164   Mean   :   5.522   Mean   : 34.88
3rd Qu.:37703928   3rd Qu.:209350629   Justicia   : 1116                           3rd Qu.: 100   3rd Qu.:   3.000   3rd Qu.: 38.00
Max.   :44274350   Max.   :356881304   Cortes     :  976                           Max.   :9999   Max.   :1125.000   Max.   :661.00
                                       (Other)    :11689
 Date.last.review     Number.of.reviews.per.month Rooms.rent.by.the.host  Availibility      Latitud          Longitud         Distancia
Min.   :2011-12-31   Min.   : 0.000              Min.   :  1.00         Min.   :  0.0   Min.   :40.33   Min.   :-3.864   Min.   : 0.014
1st Qu.:2011-12-31   1st Qu.: 0.000              1st Qu.:  1.00         1st Qu.:  0.0   1st Qu.:40.41   1st Qu.:-3.708   1st Qu.: 0.786
Median :2019-12-12   Median : 0.390              Median :  2.00         Median :133.0   Median :40.42   Median :-3.701   Median : 1.655
Mean   :2017-10-26   Mean   : 1.065              Mean   : 14.25         Mean   :158.9   Mean   :40.42   Mean   :-3.694   Mean   : 2.606
3rd Qu.:2020-03-08   3rd Qu.: 1.540              3rd Qu.:  7.00         3rd Qu.:335.0   3rd Qu.:40.43   3rd Qu.:-3.687   3rd Qu.: 3.799
Max.   :2020-07-28   Max.   :27.250              Max.   :244.00         Max.   :365.0   Max.   :40.56   Max.   :-3.524   Max.   :21.102
```

Se obtienen algunos valores inesperados:

- “Room.Price”: toma el valor 0 para alguno de los casos, pero no puede ser porque el precio de una habitación siempre tiene que ser superior a 0.
- “Minimum.nights”: toma valores muy altos equivalentes a más de 2 años en alguno de los casos.

Posteriormente, se han obtenido las estadísticas para cada variable numérica, para cada una de ellas se ha calculado:

- Media
- Media truncada
- Mediana
- Desviación típica
- Varianza
- Mínimo
- Máximo
- Boxplot: en este punto se han detectado outliers (rango de variables) para todas las variables numéricas.

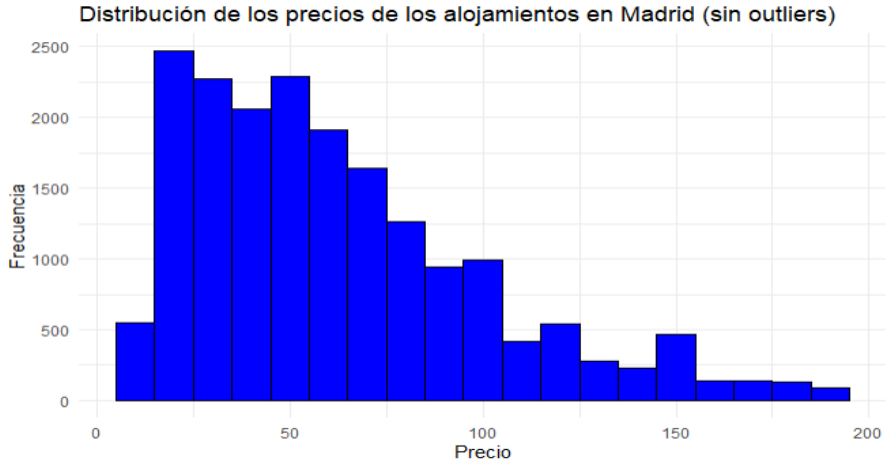
Como resultado, se han eliminado los outliers construyendo nuevos data frames para cada caso para comprobar gráficamente si eliminándolos, se aprecia mejor la relación con otras variables.

Este ha sido el caso de las variables: "Minimum.nights", "Room.Price", "Number.of.reviews", "Number.of.reviews.per.month" y "Rooms.rent.by.the.host". Para "Distancia" no se han eliminado los outliers por considerarse necesarios en la visualización de relaciones con otras variables. Finalmente, para "Disponibilidad", no se detectan outliers.

Además, para las variables "Room.type" y "Neighbourhood", que son de tipo factor, se han creado tablas para ver la distribución de los casos.

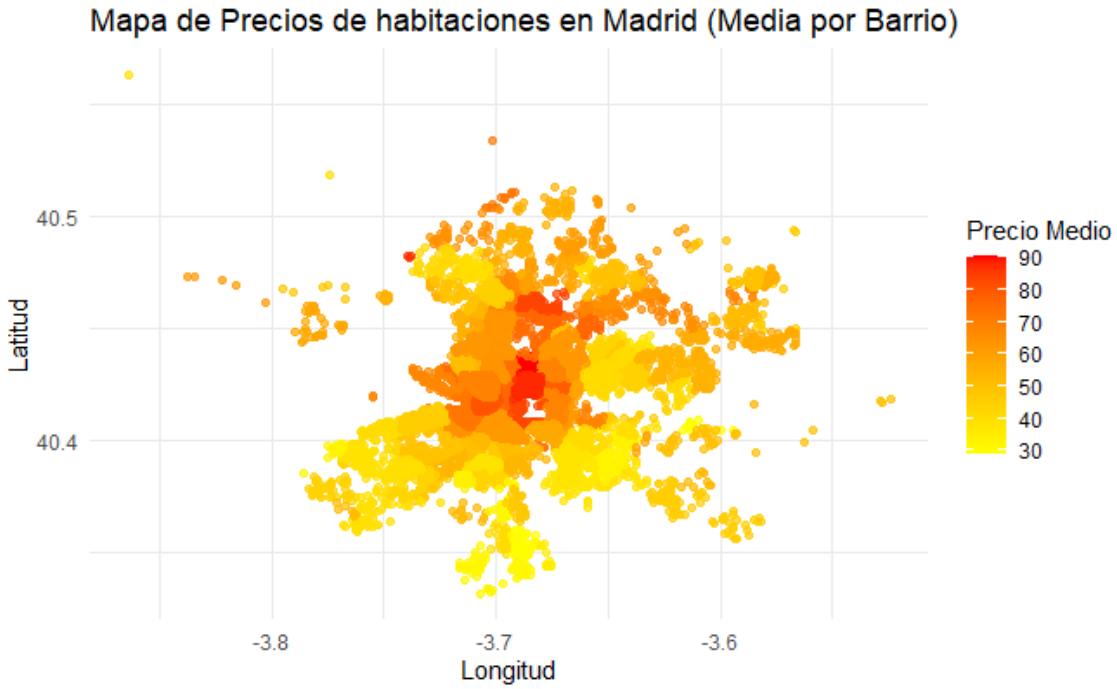
3.4 Gráficas

Lo primero, se ha creado un histograma para ver cómo se distribuyen los precios de los alojamientos, observando que, a medida que va aumentando el precio, disminuye el número de habitaciones que hay, concentrándose la mayoría en los primeros tramos de precios.



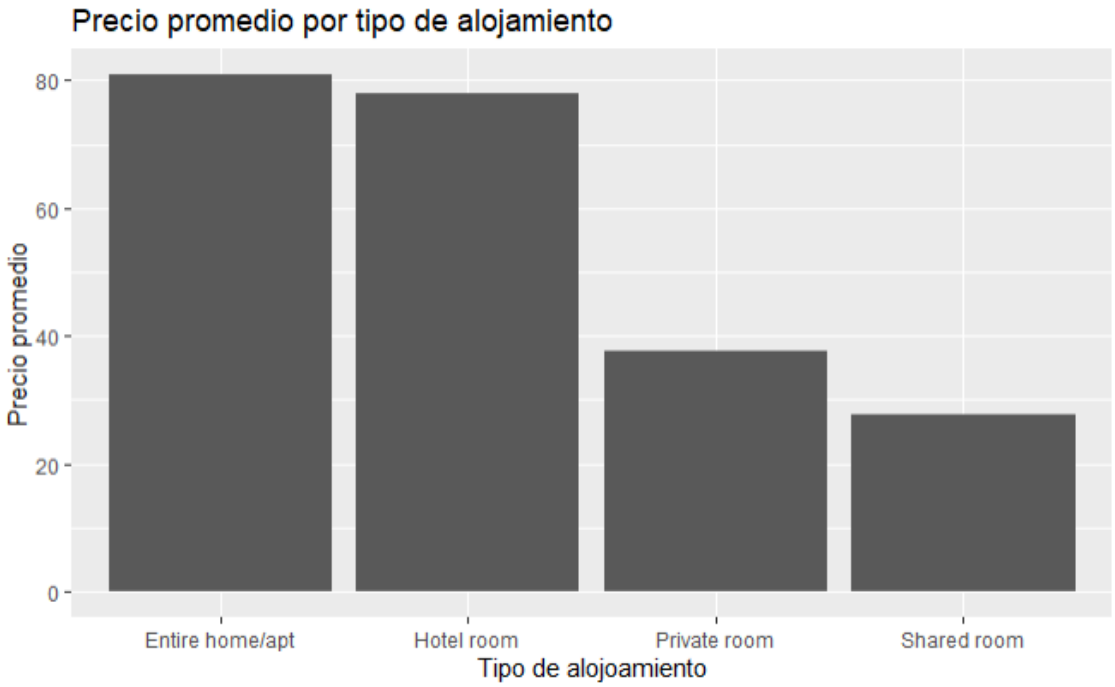
A continuación, se han creado gráficos en los que se representa el precio junto con otra variable para ver gráficamente si hay relación entre ellas.

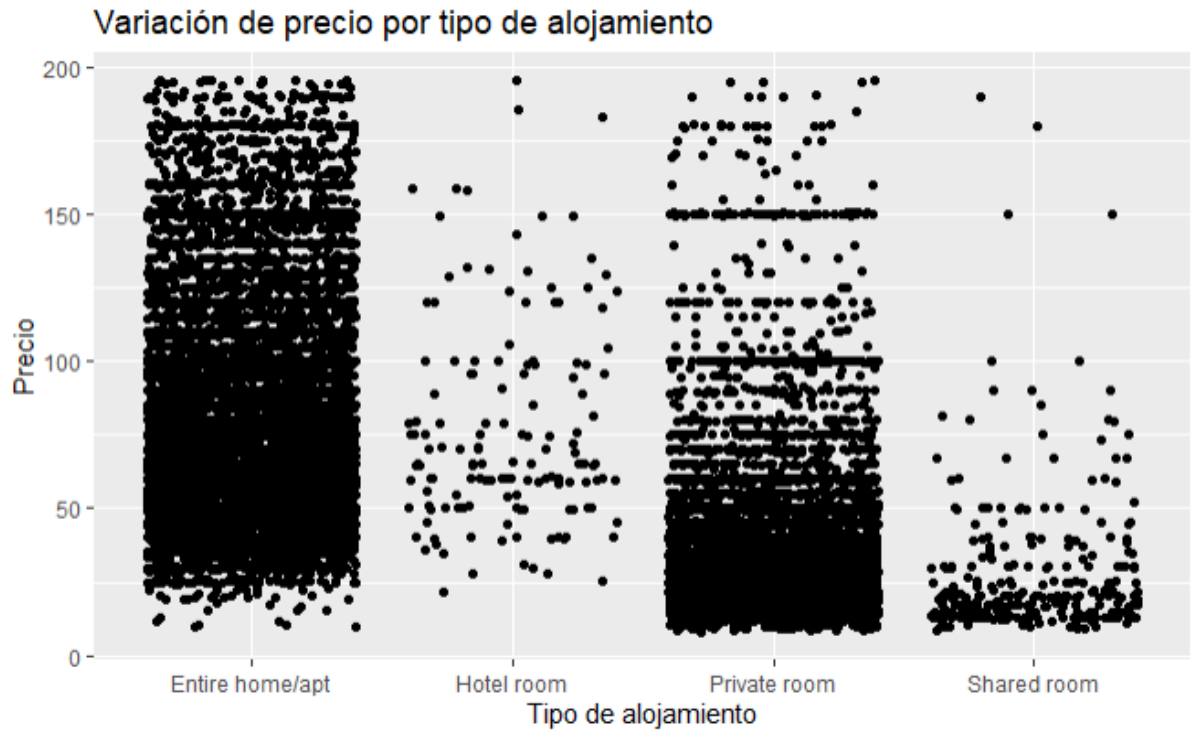
Precio en función del barrio



Los precios de los alojamientos, son mayores cuanto más cerca del centro se encuentran.

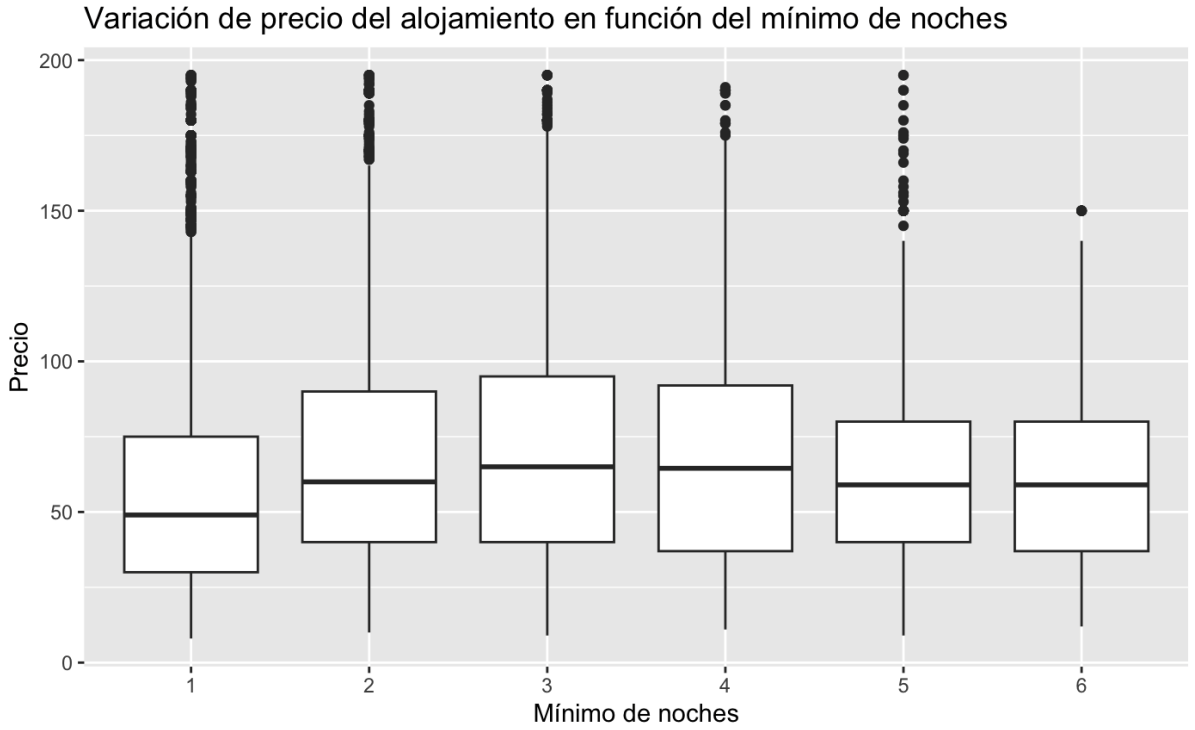
Precio en función del tipo de alojamiento





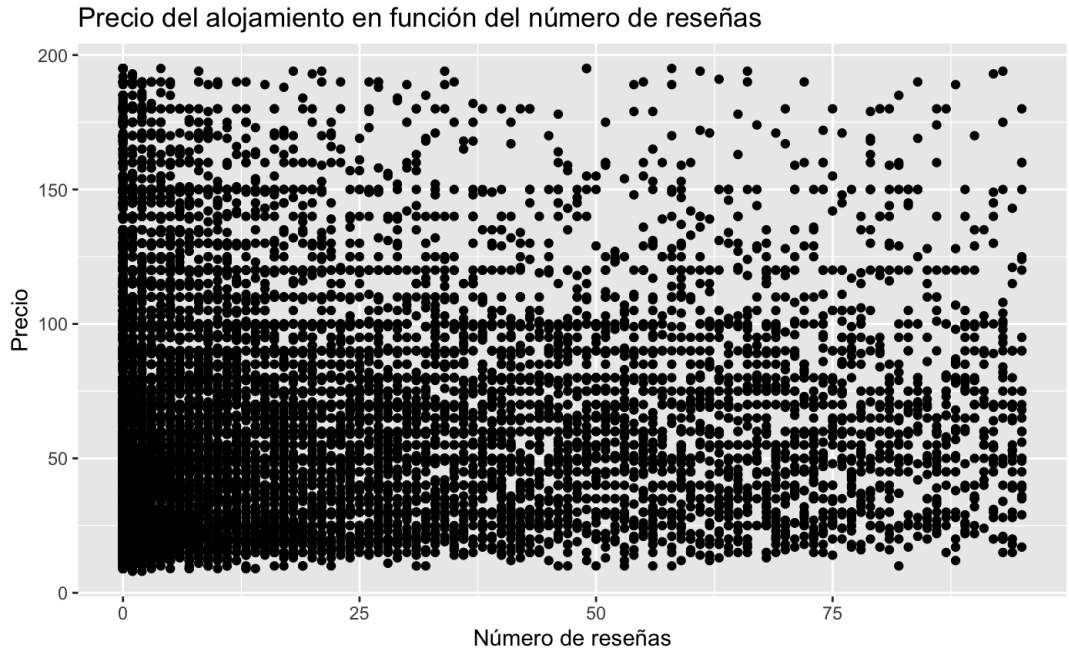
En general, los precios de casas/ apartamentos y habitaciones de hotel son más elevados, además de ser más reducidos para las habitaciones compartidas, estando en medio las habitaciones privadas.

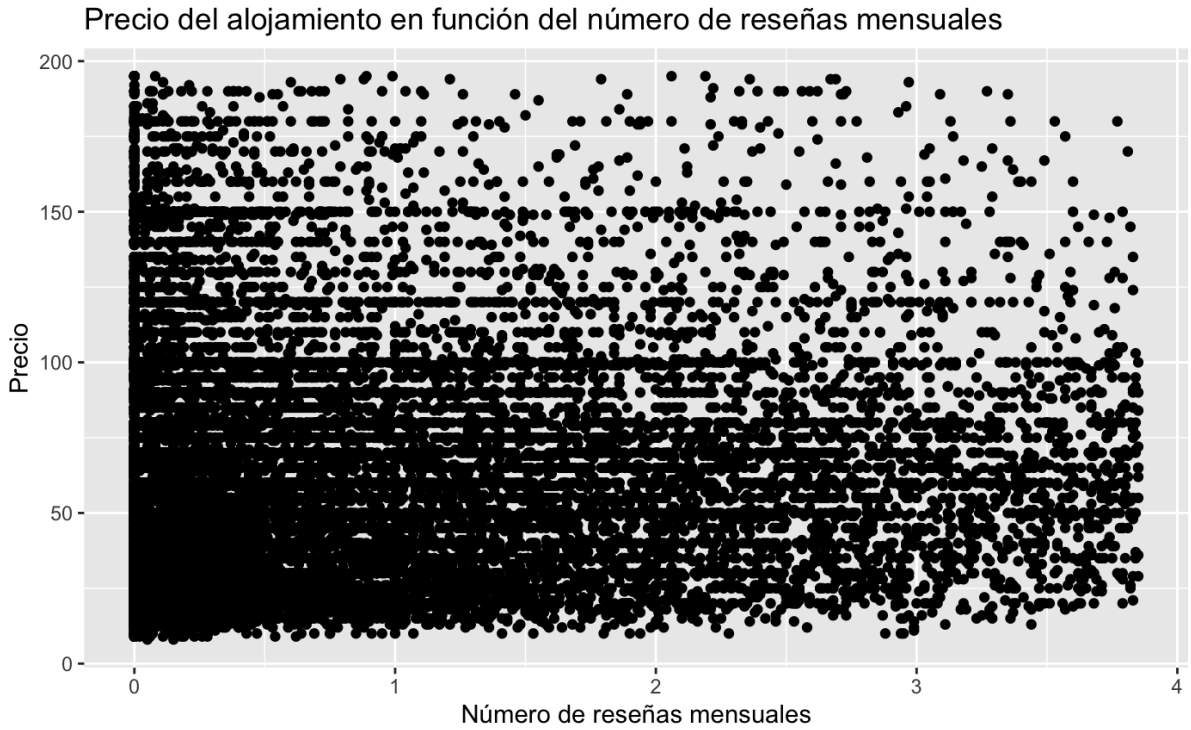
Precio en función del número mínimo de noches



No hay relación entre el precio y el número mínimo de noches, aunque a grandes rasgos, los que piden un número mínimo de noches muy elevado no alcanzan precios tan altos. Solo si nos fijamos sin outliers del mínimo número de noches, se aprecia mejor la relación y cuantas más noches se piden como mínimo, los precios suben menos.

Precio en función del número de reseñas y reseñas mensuales





Este apartado aplica tanto para la variable “Number.of.reviews” como para “Number.of.reviews.per.month”, ya que la segunda variable se obtiene a partir de la primera, llegando a las mismas conclusiones. Dado que ambas variables tienen la misma relevancia con respecto al precio y para evitar multicolinealidad eliminamos una de ellas, “Number.of.reviews.per.month”.

De esta forma, parece que no hay relación entre el precio y el número de reseñas, aunque a grandes rasgos, los que tienen muchas más reseñas, no alcanzan tantos precios tan altos.

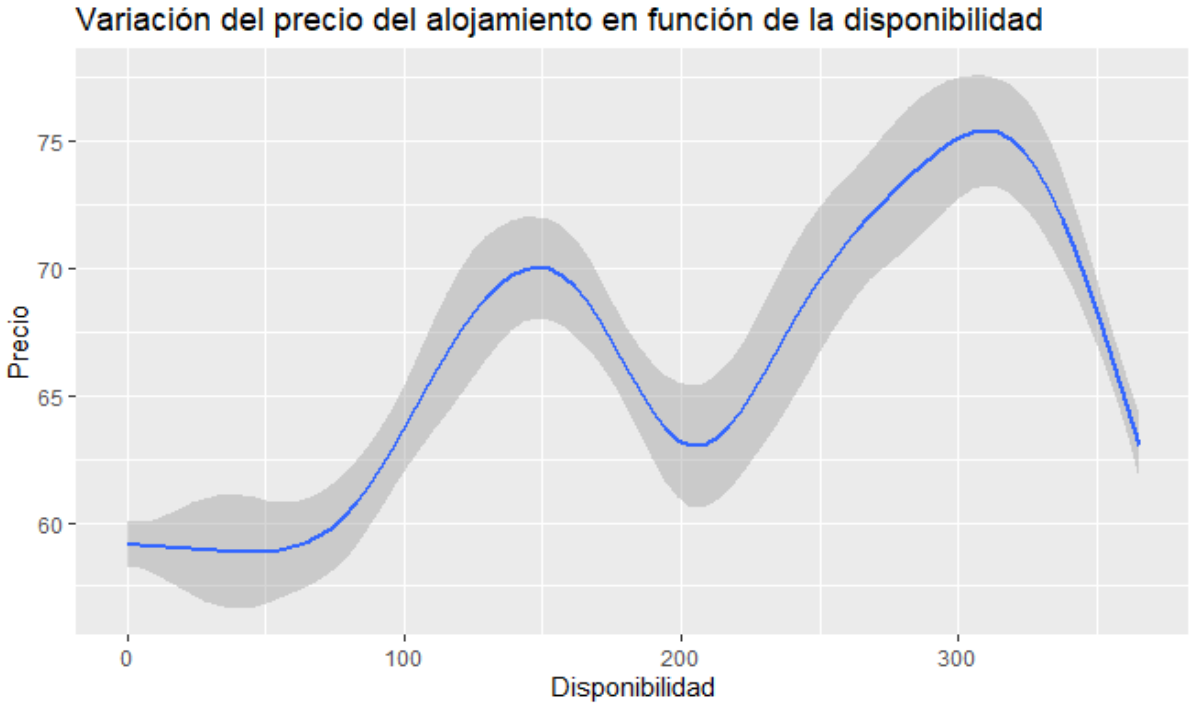
Precio en función del número de reseñas y reseñas mensuales

Al igual que en las variables anteriores, en principio no hay relación entre el precio y la fecha de la última reseña, ya que las que no tienen reseñas tienen precios de todo tipo y las que tienen reseñas, independientemente de la fecha, igualmente tienen toda variedad de precios.

Precio en función del número de alojamientos alquilados por el anfitrión

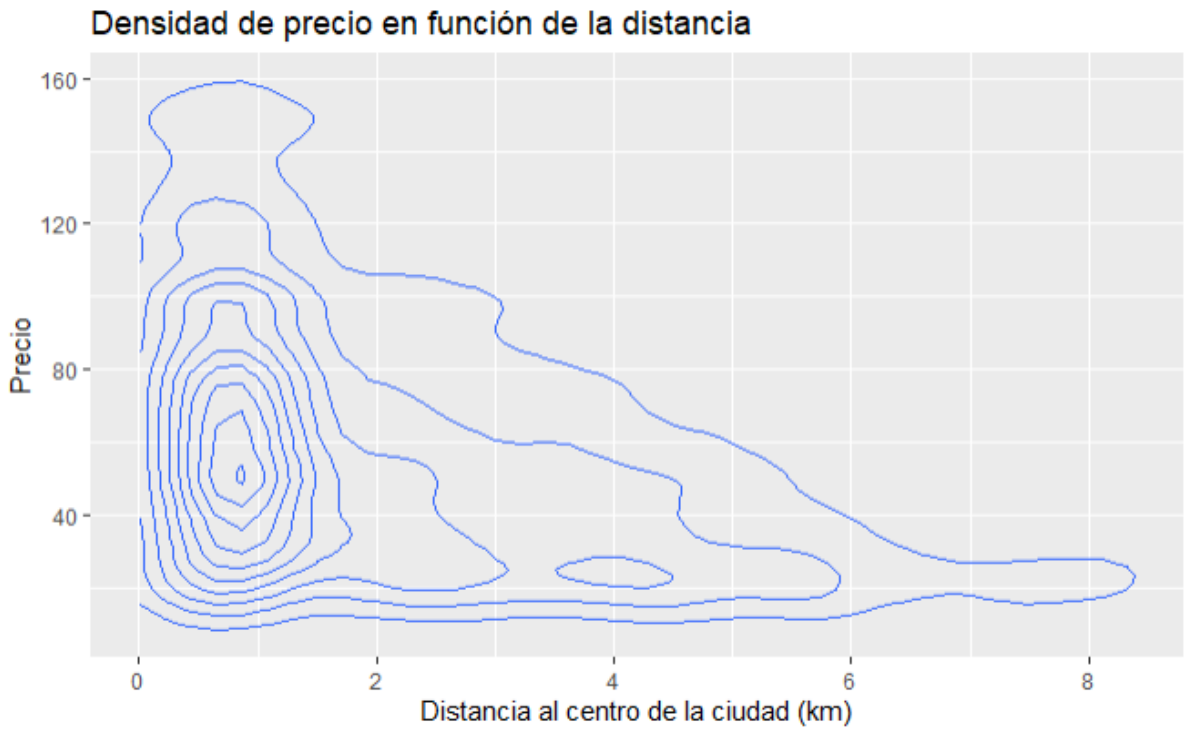
Igualmente, no parece existir una relación entre el precio del alojamiento y el número de alojamientos alquilados por un mismo anfitrión.

Precio en función de la disponibilidad del alojamiento



Cuando el precio de la habitación es bajo, la disponibilidad también lo es. Después, a medida que aumenta el precio, también lo hace la disponibilidad hasta llegar a un punto donde ambas vuelven a caer y esta tendencia se vuelve a repetir. De esta manera, sí que existe relación entre las variables.

Precio en función de la distancia al centro de la ciudad (Sol)



Como era de esperar, al igual que ocurre con el barrio, los alojamientos más céntricos pueden llegar a alcanzar precios muy altos, pero si aumentamos la distancia al centro, en general, cae el precio.

3.5 Conclusiones

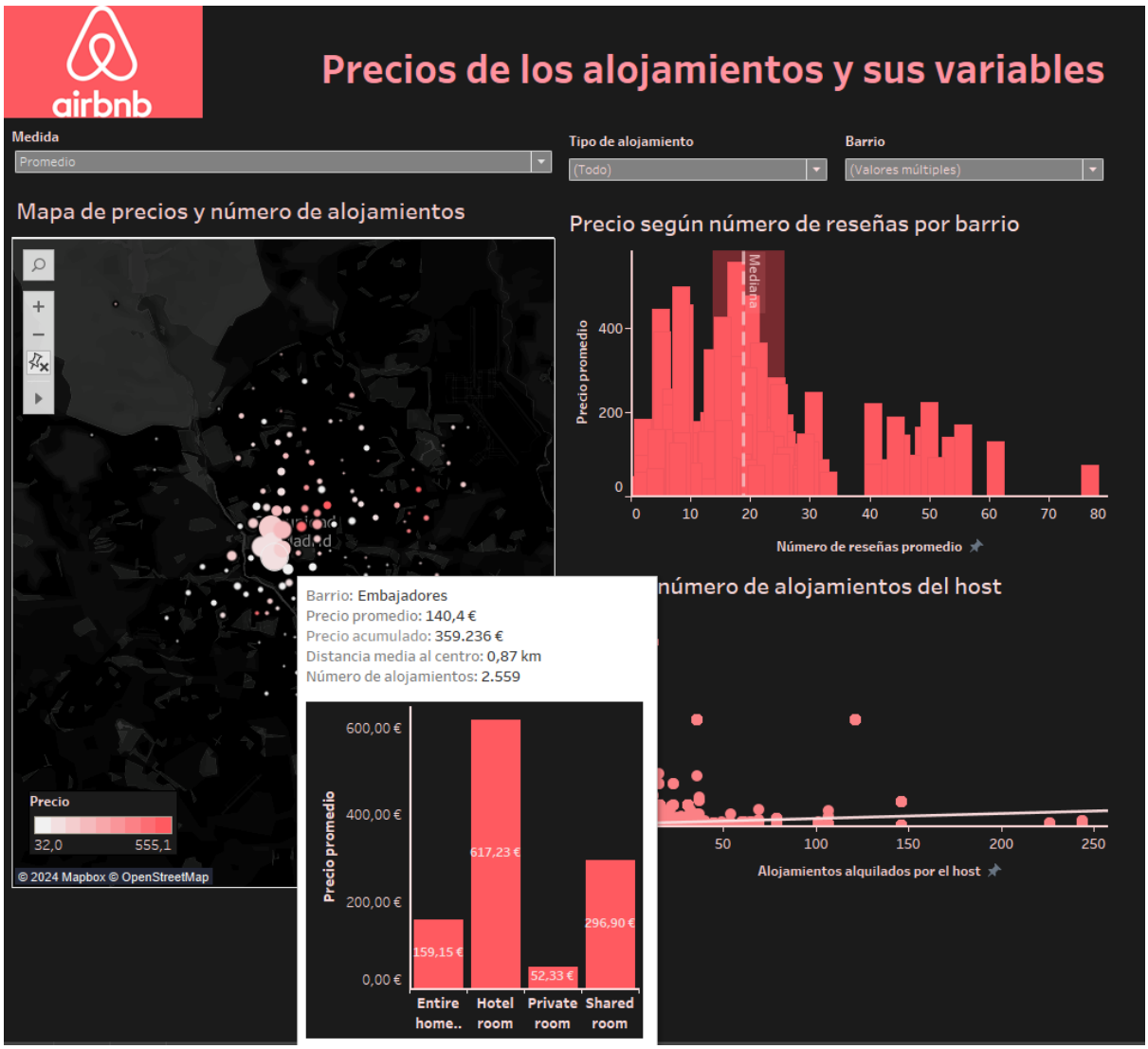
Escogiendo el Room.Price como variable principal, se han estudiado el resto de métricas para encontrar estimadores estadísticos que nos ayuden a determinar los outliers y valores no relevantes a nuestro estudio.
Tras un barrido inicial, se procede a la comparación visual del precio de alojamiento en función de cada una de las métricas que suponemos relevantes.
Finalmente, se ha determinado que los factores que más influyen en el incremento del precio de una habitación son:

- La localización del alojamiento, es decir el Barrio en el que se encuentre y por ende la distancia al barrio más céntrico de Madrid (Sol).

- El tipo de alojamiento, siendo más caros los alojamientos enteros, seguidos de las habitaciones de hotel, a diferencia de las habitaciones privadas y compartidas.
- Por último, la disponibilidad parece mantener relación el precio del alojamiento, ya que una disponibilidad menor indica un precio más reducido.

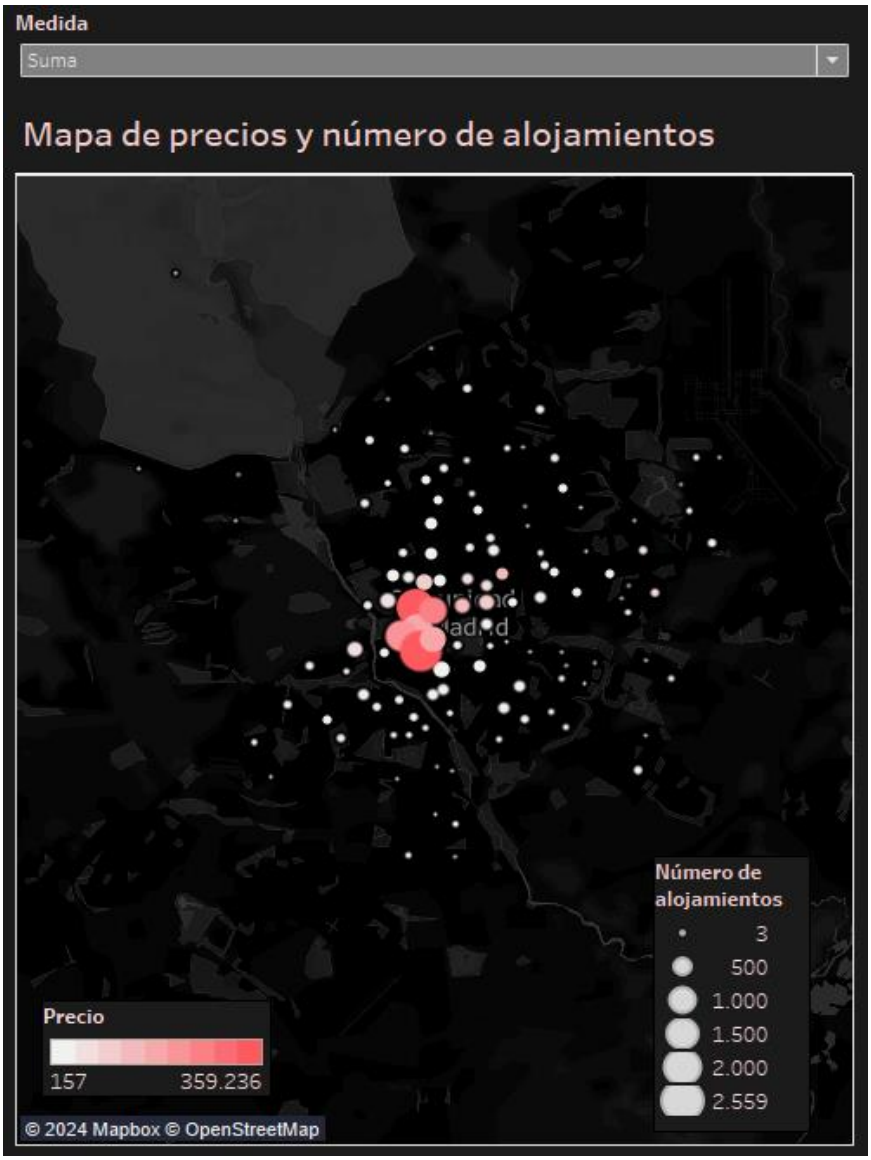
4. Visualización de las métricas

La presentación visual de los datos se realiza mediante un dashboard de Tableau. Se ha elegido un fondo oscuro porque causa menos fatiga visual y una gama de colores que es un gradiente hacia el blanco partiendo del color corporativo de Airbnb.
Este dashboard tiene tres gráficas, que muestran la relación del precio con una variable distinta cada una, y dos filtros que se aplican a todas las gráficas: El tipo de alojamiento (habitación compartida, habitación privada, habitación de hotel o vivienda completa) y el barrio. También es posible elegir el barrio haciendo click sobre el mapa. Las variables a representar se han elegido a partir del heatplot de las correlaciones calculado en el siguiente apartado, tomando las que presentaban una mayor correlación (positiva o negativa) con el precio.



4.1. Mapa

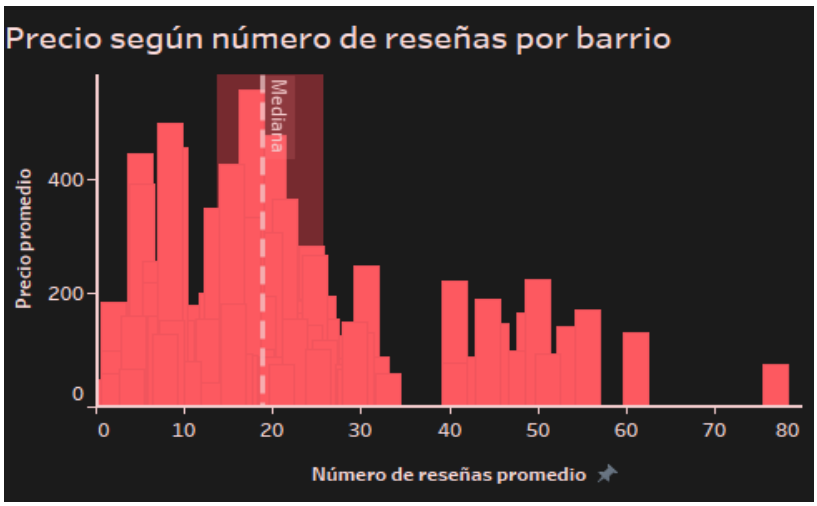
El mapa muestra los datos agregados por barrio y dos métricas.
La primera métrica es el número de alojamientos registrados en ese barrio, codificado en el tamaño del círculo. Se puede ver una clara relación entre el número de alojamientos por barrio y su situación geográfica. Los barrios más cercanos al centro tienen muchos más alojamientos, aunque no todas las zonas centrales son igualmente numerosas.
La segunda métrica es el precio, codificado en el color (cuanto más oscuro, mayor es el precio). Este precio puede visualizarse en promedio o en total para el barrio mediante la elección del parámetro Medida. Eligiendo el promedio se comprueba que hay varios cúmulos de barrios con precios considerablemente más altos. Solo uno de estos barrios es cercano al centro. Por el contrario, los barrios más centrales tienen un precio mucho menor. Por otra parte, usando el precio acumulado de todo el barrio sí que se observa que las zonas más centrales mueven más cantidad de dinero debido al mayor número de alojamientos.



El mapa además cuenta con una pestaña emergente que aparece al colocar el cursor sobre el punto de un barrio. En dicha pestaña aparecen datos estadísticos e identificativos del barrio, como el nombre, el precio promedio, el precio acumulado y la distancia al centro y un pequeño gráfico de barras con la composición de los alojamientos según su tipo con el precio de cada uno.

4.2. Gráfico de barras

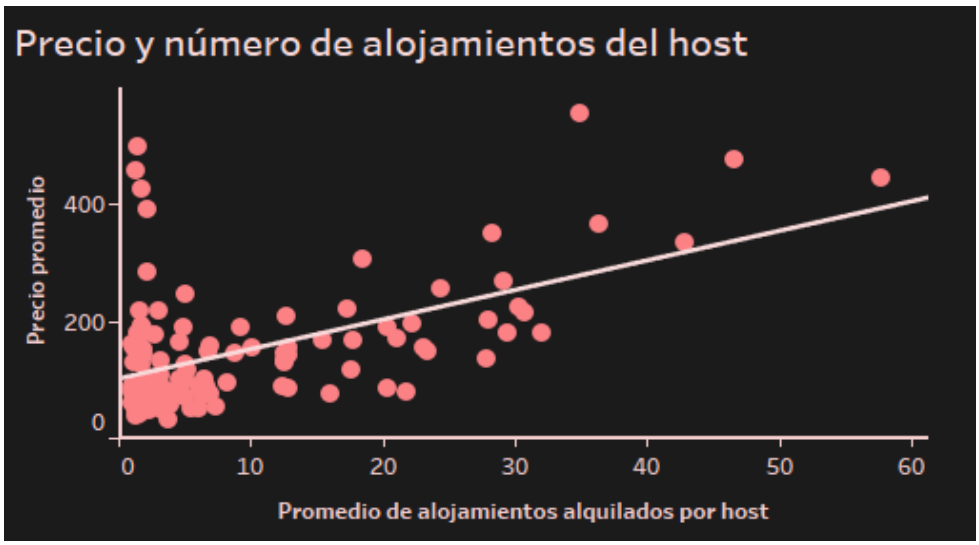
El gráfico de barras representa el precio promedio de cada barrio frente al número promedio de reseñas. El borde de las barras se ha elegido en un color que contraste poco con el relleno para dar más peso a la silueta de la figura que al valor individual de cada barrio, pero sin perder esta información. Esta silueta permite observar rápidamente que un mayorl número de reseñas no determina el precio más alto.



Los precios más altos aparecen concentrados en torno a los mismos valores centrales que los más bajos. Esta conclusión se ha resaltado marcando la mediana y sus cuartiles en la gráfica. Además, se puede ver en el solapamiento de las barras de muchos tamaños distintos.

4.3. Regresión

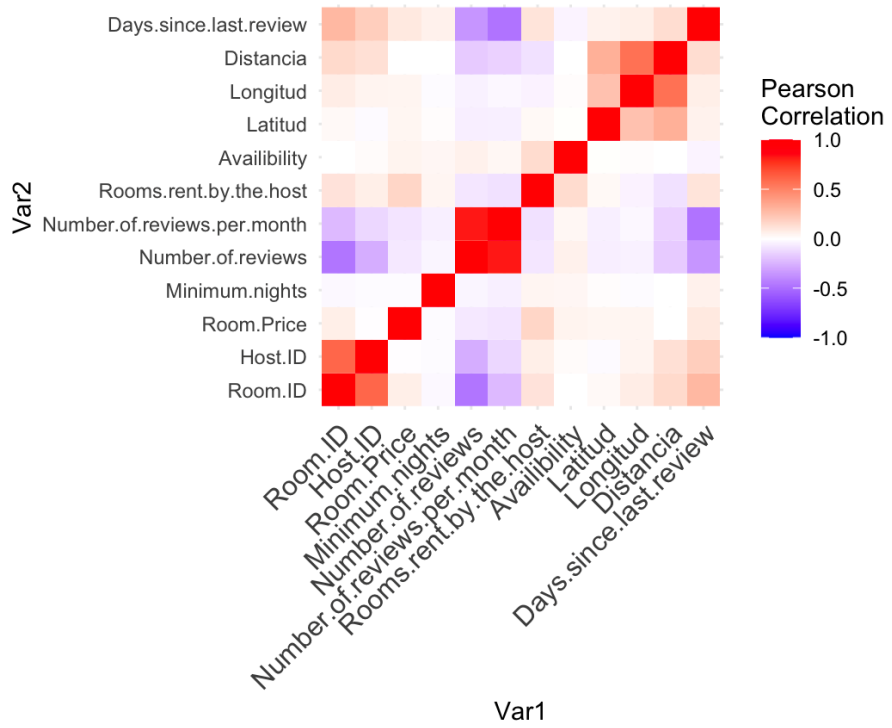
Esta última gráfica presenta la relación entre el precio promedio y el número promedio de alojamientos alquilados por cada anfitrión en cada barrio.
Parece haber una tendencia clara al aumento de precio cuantos más alojamientos alquila cada anfitrión en promedio en el barrio, pero esta tendencia solo se mantiene en los barrios con precios moderados y/o un promedio de alojamientos por anfitrión relativamente alto. Los barrios más caros se desmarcan de la tendencia general, así como los barrios con anfitriones con cinco o menos pisos. Para estos últimos, el rango de valores del precio es más amplio cuantos menos piso por anfitrión. Este resultado podría indicar una diferencia entre los precios de alquileres realizados por particulares y por profesionales.
La diferencia en comportamiento se amplía considerablemente cuando se consideran los anfitriones individuales en lugar de los promedios por barrio. Esa gráfica no está disponible en el Dashboard porque sólo se están considerando resultados a nivel de barrio.



5. Pre-procesamiento y Modelado

5.1 Heatplot

Para entender qué parámetros estaban más correlacionados con el precio, realizamos un heatplot, que nos indica la correlación entre todas las variables numéricas.



Viendo que precio no tenía ninguna correlación +/- 5 0.5 con ninguna variable

---

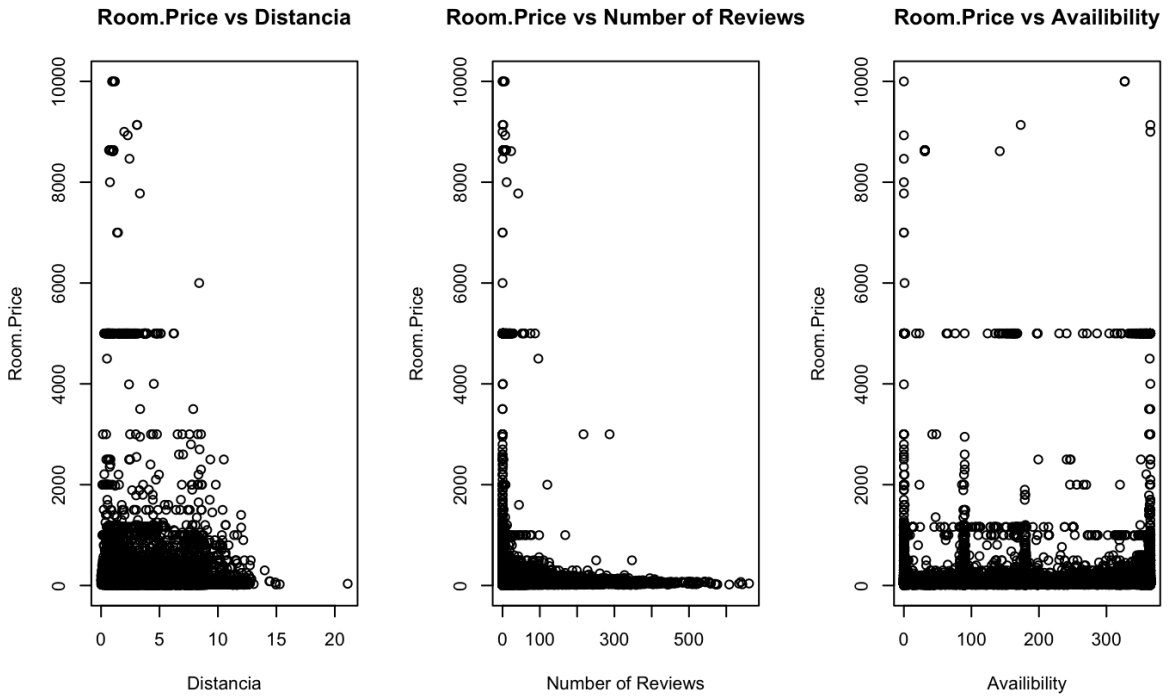
5.2 Pre-Procesamiento

5.2.1. Limpieza de Outliers

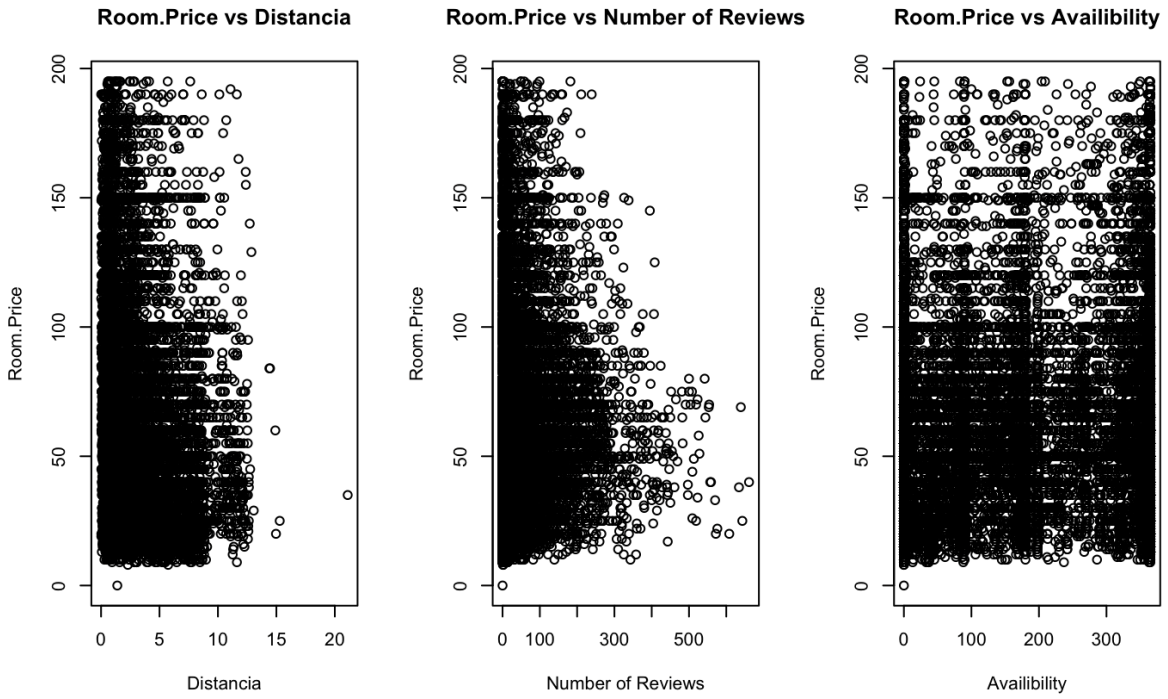
Creamos una gráfica que nos mostraba tanto:

- La relación entre las top variables numéricas con Room.Price
- Los outliers de Room.Price

Debajo vemos cómo hay muchos upper-outliers que no nos permiten apreciar la dispersión del groso de entries de debajo.



Tras eliminar tanto los upper-outliers como aquellas entries que tenían un precio = 0 (lower outliers), nos quedamos con una distribución de Room.Price como la de la imagen inferior.



5.2.2. Feature Engineering

Creación de una nueva variable 'Days since last review', creada a partir de Date.last.review. Esta nueva variable cuenta los días desde la fecha señalada hasta el día de hoy, convirtiéndola así en un integer, más fácil de incorporar al modelo.

5.2.3. Conversión de Variables Categóricas

Conversión de las variables Room.Type y Neighbourhood en factores para poder procesarlas en el modelo.

5.2.4. Elección de Columnas de Interés para el Modelo

Para evitar multicolinealidad, eliminamos columnas innecesarias para el modelo. Debajo un resumen de las variables eliminadas y utilizadas con la explicación.

Título de campo	**Eliminado**	**Utilizado en Modelo**	**¿Por qué?**
Location	X		Convertido en 'Distancia'
Coordinates	X		Convertido en 'Distancia'
Country	X		Todos los datos eran de Madrid
City	X		Todos los datos eran de Madrid
Updated.Date	X		Todos tenían la misma fecha. Varianza = 0.
Availability		X	Relevante
Rooms.rent.by.the.host		X	Relevante
Number.of.reviews.per.month	X		Por evitar multicolinearidad con Number.of.reviews
Date.last.review	X		Convertida a 'Days.since.last.review'
Number.of.reviews		X	Relevante
Minimum.nights		X	Relevante
Room.Price		X	Target Variable

Room.type		X	Relevante
Neighbourhood		X	Relevante
Host.ID	X		Serial ID - sin valor estadístico
Name	X		Difícil de interpretar en modelo.
Room.ID	X		Serial ID - sin valor estadístico
Distancia (creada)		X	Relevante
Days.since.review (creada)		X	Relevante

5.2.5. Train/Test Split

Elegimos una distribución de 0.8 train y 0.2 test.

5.3 Diseño del modelo

Diseñamos dos modelos, uno con el dataset limpio tras estos pasos de procesamiento y otro que además aplicaba normalización a las variables dependientes.

5.3.1 Modelo 1: Sin normalización

- Variables:
    - Distancia
    - Neighbourhood
    - Days.since.review
    - Room.type
    - Minimum.nights
    - Number.of.reviews
    - Rooms.rent.by.host

5.3.2 Modelo 2: Con normalización

- Variables
    - Normalizadas:
        - Distancia
        - Days.since.review
        - Minimum.nights
        - Number.of.reviews

        - Rooms.rent.by.host

    - Demás variables
        - Neighbourhood
        - Room.type

5.4 Evaluación del Modelo

Se nota una ligera mejoría del modelo al normalizar las variables dependientes.

5.4.1 Modelo 1: Sin normalización

- Resultados del modelo:
    - RMSE: 30.75287
    - MAE: 22.5623
    - R^2: 0.3524967

5.4.2 Modelo 2: Con normalización

- Resultados del modelo:
    - RMSE: 30.74166
    - MAE: 22.5367
    - R^2: 0.3528252

6. Informe

6.1 Suposiciones iniciales

La hipótesis inicial que quisimos comprobar fue:

- El precio se puede predecir en función de la distancia al centro

Además, para realizar un análisis más exhaustivo, se plantearon objetivos secundarios, consistiendo en la influencia del resto de variables en el precio del alojamiento.

6.1.1 Cuáles han demostrado ser válidas y cuáles no. ¿Por qué?

Al realizar el estudio, nos dimos cuenta de que no hay ninguna variable que tenga una correlación fuerte con el precio, pero que es mejor utilizar un mayor número de variables para predecir el precio de los alojamientos, ya que el modelo funciona mejor con utilizando una gran parte de las variables (una vez eliminadas las que eran redundantes).

De esta forma, las variables que resultaron más relevantes para predecir el precio de un alojamiento fueron: disponibilidad, número de habitaciones alquiladas por un mismo anfitrión, número de reseñas, número mínimo de noches para alquilar un alojamiento, el tipo de habitación, el barrio y, además, dos variables que creamos a partir de los datos que teníamos disponibles, que fueron: distancia y días desde la última reseña.

6.2 Métricas seleccionadas: ¿han sido las correctas o no? ¿por qué?

Pese a que se ha conseguido encontrar una correlación entre algunas de las variables disponibles en la base de datos con el precio del alojamiento, para optimizar el modelo, nos hubiese gustado tener otro tipo de datos que puedan afectar, como la valoración de cada alojamiento y/o anfitrión, transporte público cercano u otros servicios disponibles, política de cancelación, tamaño del alojamiento, características como si la habitación y/o el baño son privados o compartidos, entre otras.

6.3 Teniendo en cuenta lo aprendido ¿Qué cosas se harían igual y cuales se harían de otra forma? ¿Por qué?

Además del trabajo realizado, para predecir el precio de los alojamientos de manera más precisa, se podría ampliar realizando nuevos análisis, incluyendo nuevas variables o tratando de encontrar otro tipo de relaciones entre ellas.

Una buena práctica, podría haber sido realizar al principio del trabajo un análisis de correlaciones para comprobar desde el principio variables que no nos son útiles para el análisis para así no utilizarlas.

También, se podrían haber tratado de encontrar otro tipo de relaciones entre las variables para comprobar si el modelo se ajusta mejor a otro tipo de distribución y analizarlo con mayor profundidad.

6.4 Conclusiones y “lessons learned”

El precio no parece estar determinado, al menos completamente, a partir de las variables que tenemos disponibles en nuestros datos, aunque sí se ha encontrado relación con algunas de ellas. Por tanto, el precio debe estar fuertemente influenciado por alguna variable adicional que no está presente en nuestro dataset.

En la base de datos, aparecen multitud de outliers, por lo que habría que estudiar en detalle su origen, ya que algunos de ellos pueden ser anomalías en los datos pero otros simplemente errores de registro.

Como lecciones aprendidas, hemos aprendido a utilizar o mejorar el uso de programas muy útiles e interesantes para realizar investigación y análisis de datos. Además, es importante tener en cuenta que no es suficiente con los conocimientos técnicos a la hora de realizar un trabajo en grupo, sino que este aprendizaje debe ir acompañado de una buena coordinación y comunicación entre los miembros del mismo.