PROYECTO FINAL KEEPCODING: AIRBNB

Anastasia Prischep Chulannikova
Elvira Beatriz Méndez Sánchez

1. Definición, validación de los datos y arquitectura

En el presente documento trabajaremos con datos de Airbnb sacados del portal OpenDataSoft, que a su vez lo ha sacado del portal Inside Airbnb. Inside Airbnb es una iniciativa que se dedica a proporcionar información y datos relevantes sobre el impacto de Airbnb en las comunidades residenciales. En este caso la zona de estudio se trata de la capital de España, Madrid.

a) Muestreo y exploración inicial de los datos

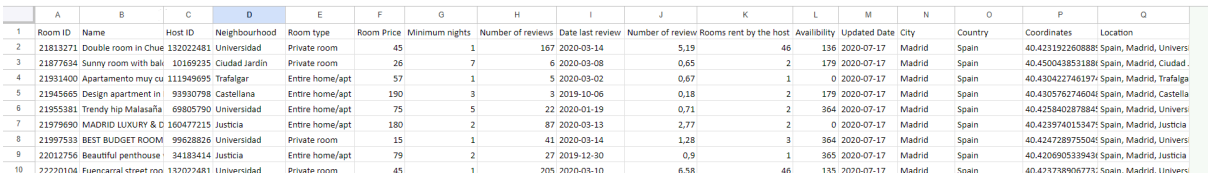
En el Dataset de entrada tenemos las siguientes 16 columnas con un total de 21.278 registros, el cual tienen la siguiente tipología de datos:

Room ID	Name	Host ID	Neighbourhood	Room type	Room Price	Minimum nights
Double	*Character*	*Double*	*Character*	*Character*	*Double*	*Double*

Number of reviews	Date last review	Number of reviews per month	Rooms rent by the host
Double	*Double*	*Double*	*Double*

Availibility	Updated Date	City	Country	Coordinates	Location
Character	*Character*	*Character*	*Character*	*Character*	*Character*

A continuación se enseña una pequeña muestra de los datos del fichero:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Room ID	Name	Host ID	Neighbourhood	Room type	Room Price	Minimum nights	Number of reviews	Date last review	Number of review	Rooms rent by the host	Availibility	Updated Date	City	Country	Coordinates	Location
2	21813271	Double room in Chue	132022481	Universidad	Private room	45	1	167	2020-03-14	5,19	46	136	2020-07-17	Madrid	Spain	40.4231922608888	Spain, Madrid, Universi
3	21877634	Sunny room with bal	10169235	Ciudad Jardín	Private room	26	7	6	2020-03-08	0,65	2	179	2020-07-17	Madrid	Spain	40.4500438531886	Spain, Madrid, Ciudad J
4	21931400	Apartamento muy cu	111949695	Trafalgar	Entire home/apt	57	1	5	2020-03-02	0,67	1	0	2020-07-17	Madrid	Spain	40.4304227461974	Spain, Madrid, Trafalga
5	21945665	Design apartment in	93930798	Castellana	Entire home/apt	190	3	3	2019-10-06	0,18	2	179	2020-07-17	Madrid	Spain	40.4305762746044	Spain, Madrid, Castella
6	21955381	Trendy hip Malasaña	69805790	Universidad	Entire home/apt	75	5	22	2020-01-19	0,71	2	364	2020-07-17	Madrid	Spain	40.4258402878844	Spain, Madrid, Universi
7	21979690	MADRID LUXURY & D	160477215	Justicia	Entire home/apt	180	2	87	2020-03-13	2,77	2	0	2020-07-17	Madrid	Spain	40.4239740153475	Spain, Madrid, Justicia
8	21997533	BEST BUDGET ROOM	99628826	Universidad	Private room	15	1	41	2020-03-14	1,28	3	364	2020-07-17	Madrid	Spain	40.4247289755044	Spain, Madrid, Universi
9	22012756	Beautiful penthouse	34183414	Justicia	Entire home/apt	79	2	27	2019-12-30	0,9	1	365	2020-07-17	Madrid	Spain	40.4206905339434	Spain, Madrid, Justicia
10	22220104	Fuencarral street roo	132022481	Universidad	Private room	45	1	205	2020-03-10	6,58	46	135	2020-07-17	Madrid	Spain	40.4237389067734	Spain, Madrid, Universi

b) Definir y implementar el Datawarehouse

Nuestro trabajo se basa principalmente en hacer un análisis para ayudar a un usuario final a tomar mejores decisiones a la hora de escoger un apartamento para alquilar de Airbnb, dependiendo de la zona, el rango de precio y su tipología, así que hemos hecho las siguientes correcciones al dataset.

Hemos procedido a hacer una primera selección de los datos que hemos querido utilizar, siendo estos los siguientes (10 columnas con 21.256 registros):

	A	B	C	D	E	F	G	H	I	J
1	Room ID	Host ID	Neighbourhood	Room type	Room Price	Minimum nights	Rooms rent by the host	Availibility	Coordinates	Location
2	21859113	157114944	Argüelles	Entire home/apt	147	1	36	0	40.421911823829916, -3.716298875039167	Spain, Madrid, Argüelles
3	21862103	8851341	Recoletos	Entire home/apt	625	5	1	177	40.42205487717857, -3.6889613202974916	Spain, Madrid, Recoletos
4	21875158	159570292	Sol	Private room	500	1	2	347	40.41573794398538, -3.7044166450981693	Spain, Madrid, Sol
5	21932504	160055902	Palacio	Entire home/apt	62	2	1	286	40.421747567976034, -3.710994512030235	Spain, Madrid, Palacio
6	22042300	160981040	Portazgo	Private room	12	15	3	106	40.391493766002014, -3.6450831761057922	Spain, Madrid, Portazgo
7	22043848	160967778	Sol	Entire home/apt	70	4	6	125	40.418435611771486, -3.7013169106966464	Spain, Madrid, Sol
8	22048206	105797031	Cortes	Entire home/apt	69	10	2	47	40.4137713650942, -3.6978832654430414	Spain, Madrid, Cortes
9	22066174	54343289	Imperial	Entire home/apt	81	1	6	0	40.41094353528144, -3.7216939632075747	Spain, Madrid, Imperial
10	22079865	34563914	Castillejos	Entire home/apt	50	7	1	301	40.45956949608513, -3.6960111111683593	Spain, Madrid, Castillejos

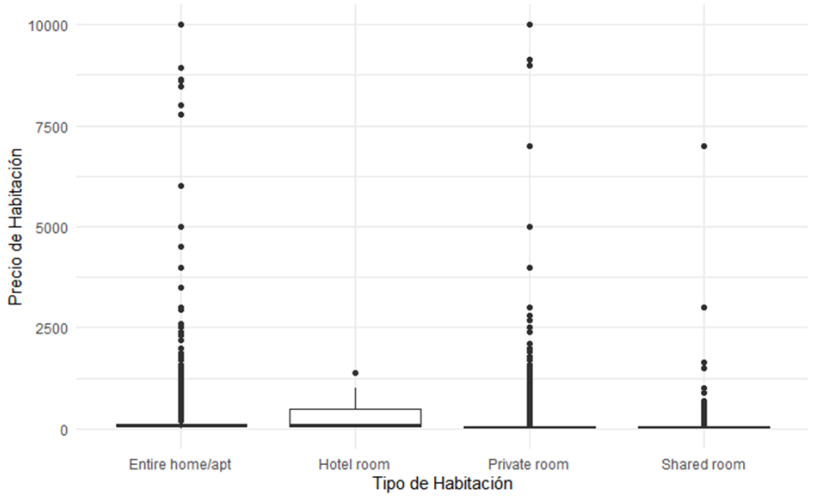
Room ID	Host ID	Neighbourhood	Room type	Room Price	Minimum nights	Rooms rent by the host	Availibility	Coordinates	Location
Double	*Double*	*Character*	*Character*	*Double*	*Double*	*Double*	*Character*	*Character*	*Character*

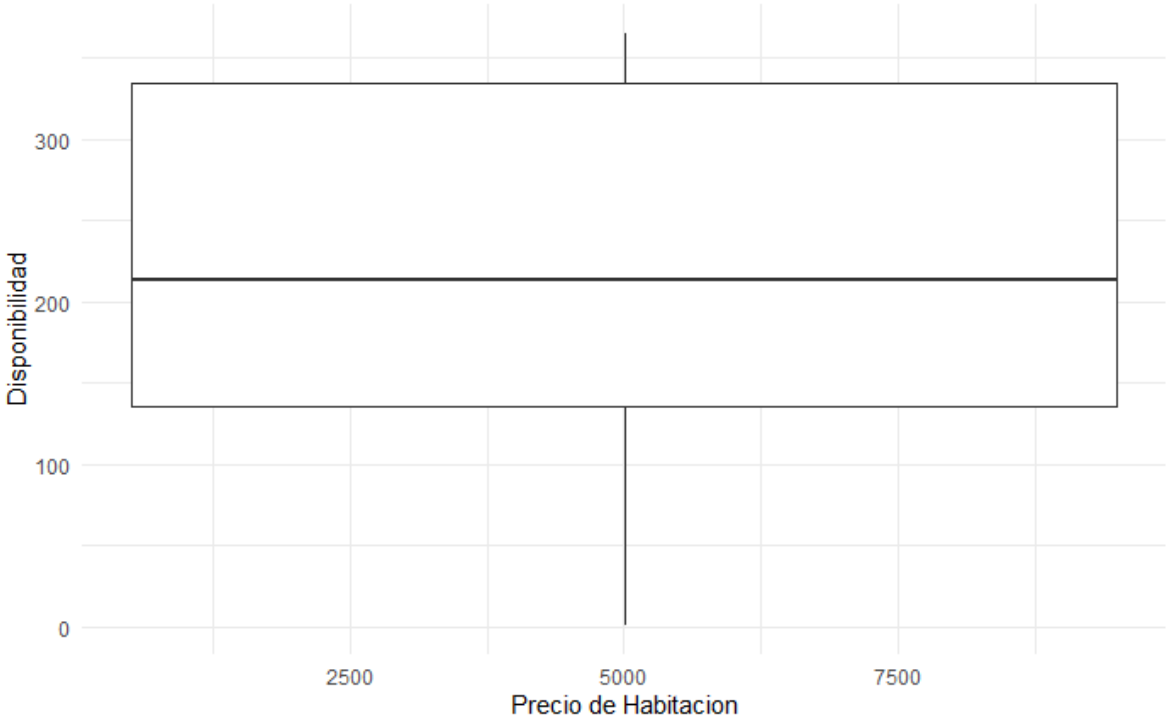
Los demás los hemos eliminado debido a que los consideramos irelevantes para nuestro análisis.
Después de este primer contacto con el Dataset, hemos procedido a ver que los datos que tenemos son los necesarios y a proceder a su normalización y revisión de su calidad, y que por lo tanto no contengan errores, tales como que sean los precisos de la zona de estudio, como que no tengan duplicidades, que estén en el formato correcto que no tenga valores nulos, entre otros.

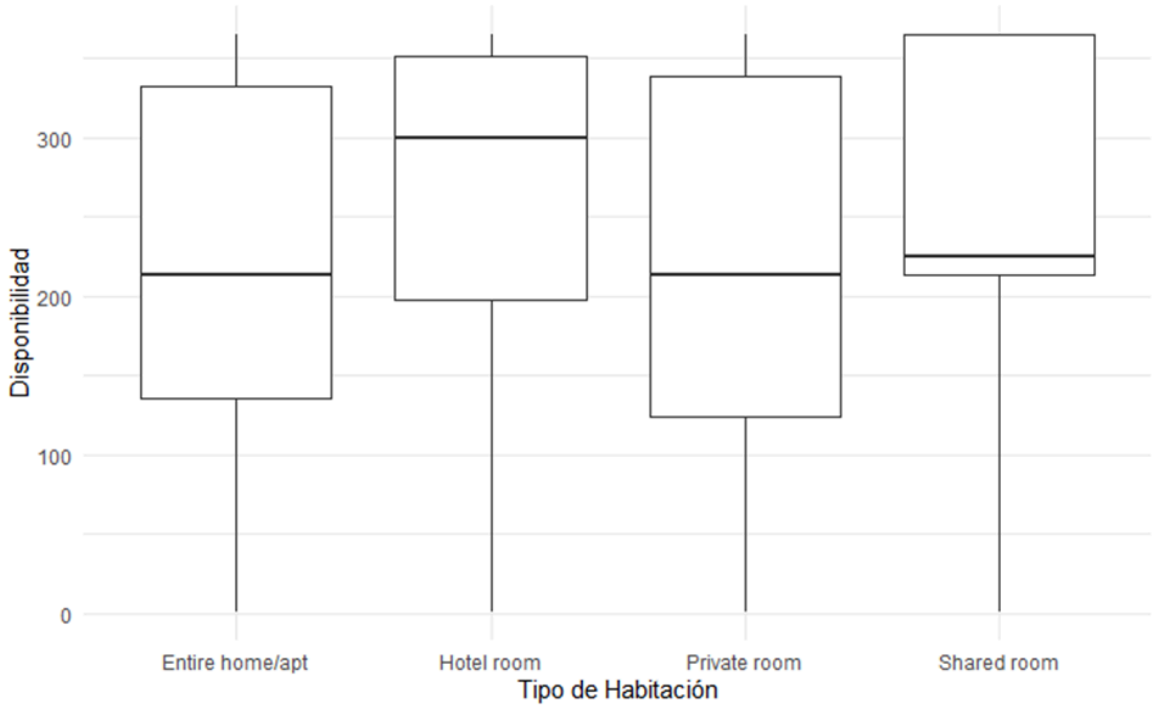
2. Análisis exploratorio

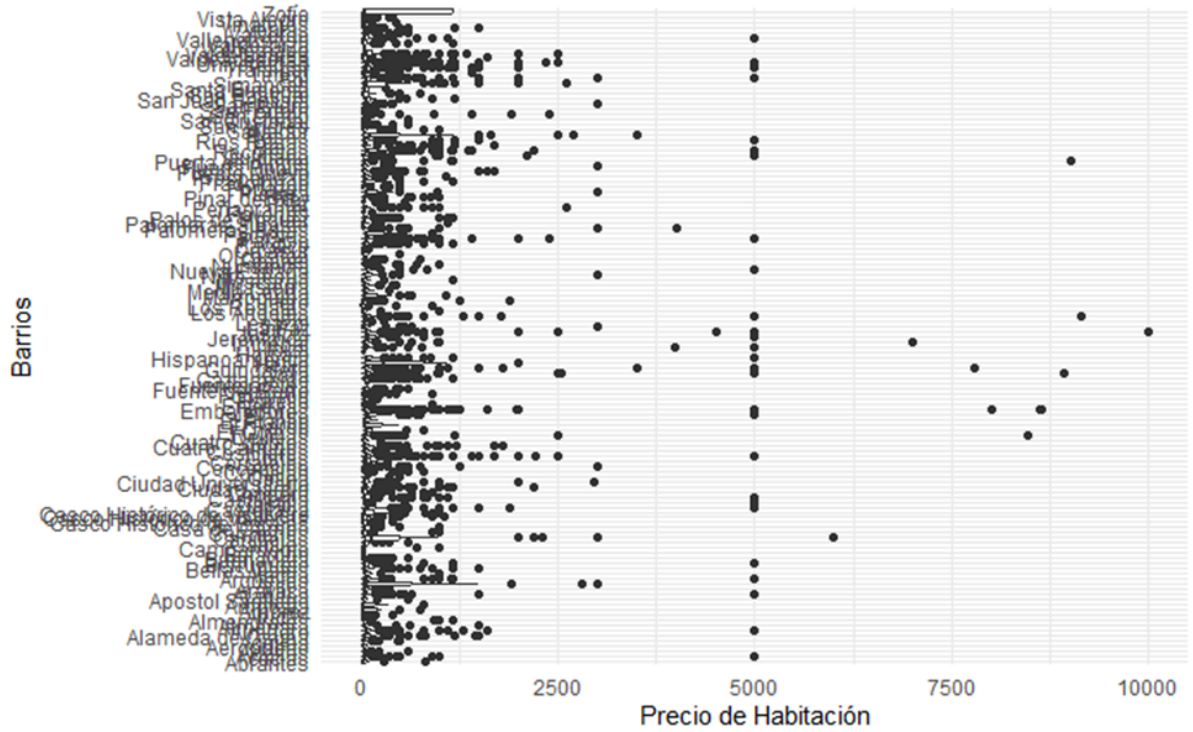
A priori hemos utilizado una limpieza de datos para poder expresar toda la Información que queríamos poner, hemos realizado tanto el precio de habitaciones como en los lugares los barrios el tipo de habitación así como la disponibilidad de las mismas.

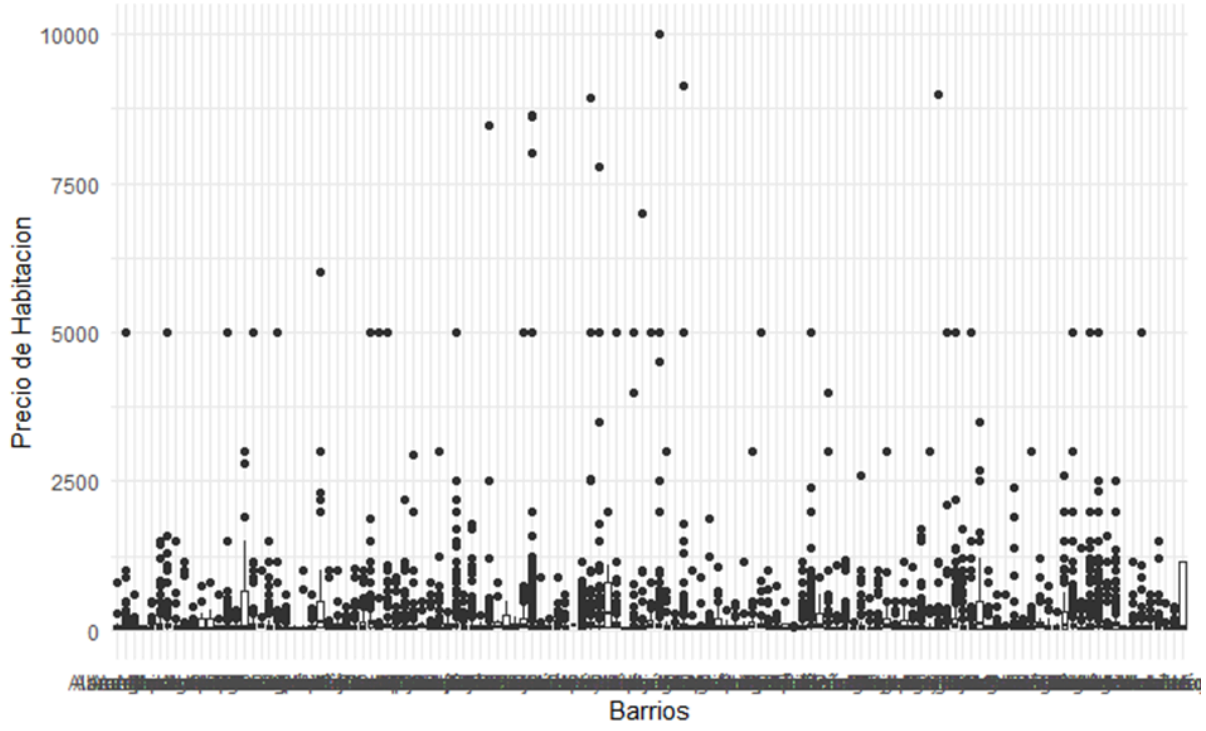
Los Boxplots utilizados actualmente podemos observar que los precios de la habitación y el tipo de habitación tienen una amplia gama de precios, así como el tipo de habitación y la disponibilidad también son variables. En el caso de los precios de la habitación y de los barrios podemos observar que no están claros los campos concretos entonces aquí tendríamos que haber hecho una limpieza de los barrios conforme a barrios más destacados con precios más grandes y los barrios con los precios más pequeños.



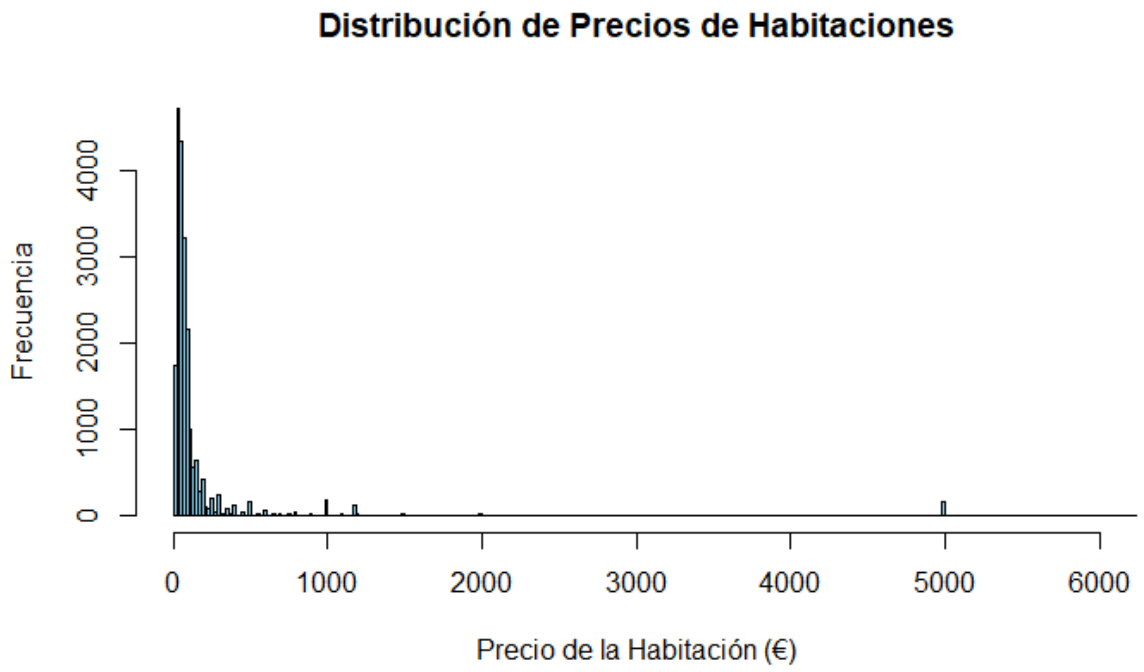

Disponibilidad
300
200
100
0
2500
5000
7500
Precio de Habitacion

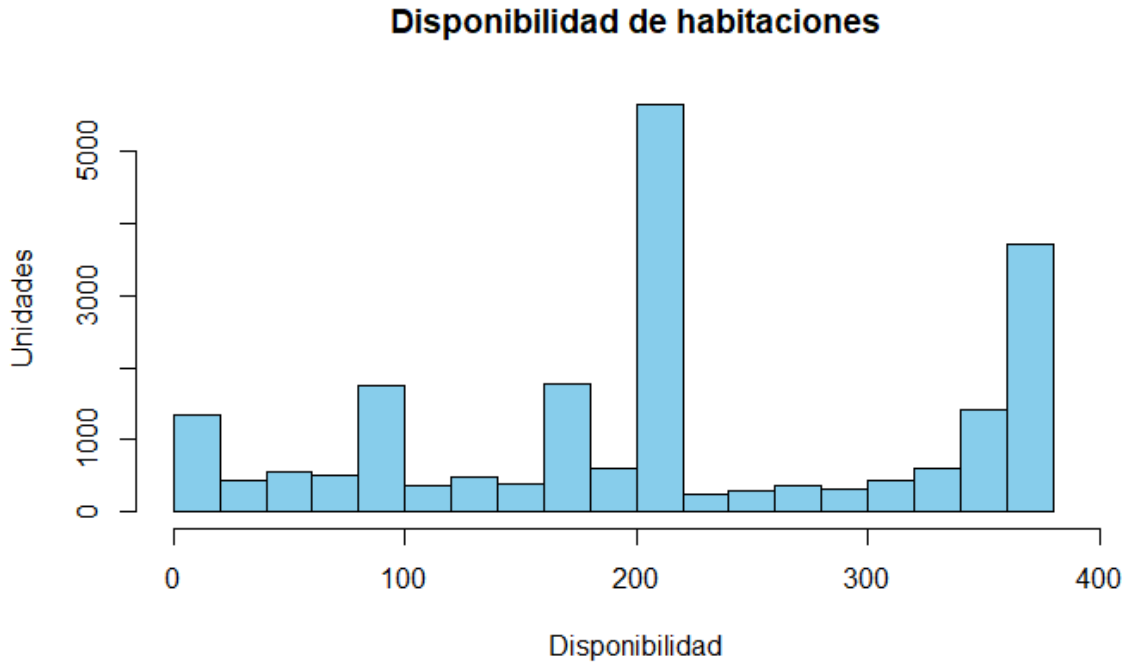

Disponibilidad
300
200
100
0
Entire home/apt
Hotel room
Private room
Shared room
Tipo de Habitación





También sacamos a destacar el histograma sobre los precios de las habitaciones y su Disponibilidad:





3. Visualización de las métricas en Tableau

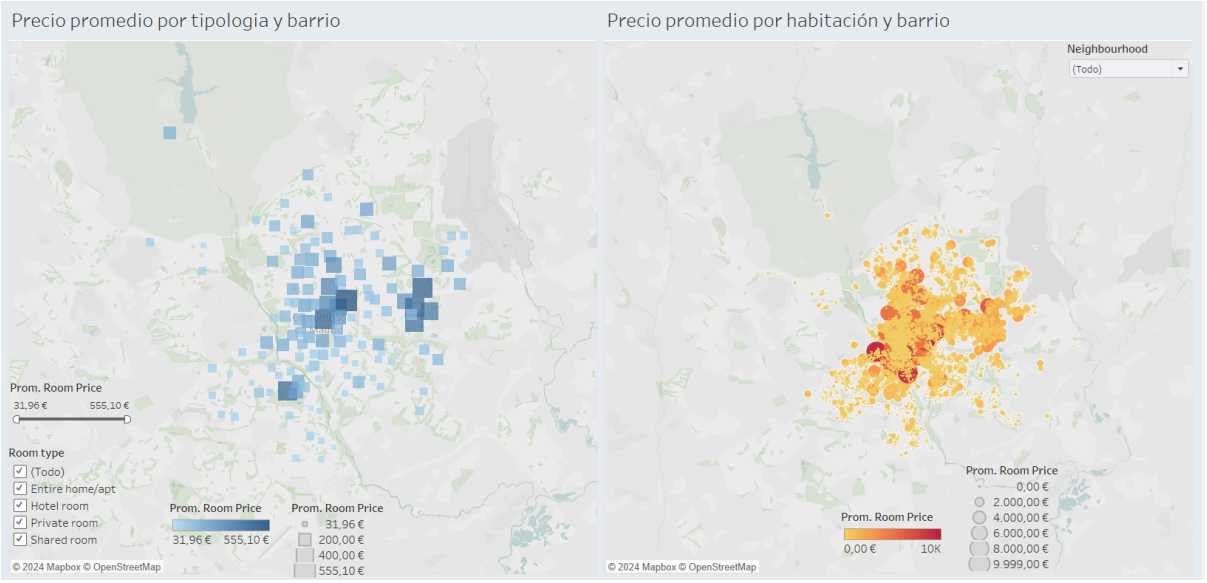
a) Definición del KPI

Para la visualización de las métricas, se ha planteado la pregunta siguiente:

- En qué zonas de Madrid y qué tipologías són las más caras para alquilar un apartamento dependiendo de su tipología.

b) Elaboración y diseño final del dashboard

Primeramente se han recalculado los campos de Latitud y Longitud creados anteriormente en el Dataset, seguidamente se ha procedido a hacer un dashboard con dos visualizaciones interactivas.
En el primer mapa se puede interactuar con el precio promedio que un usuario esté buscando según la tipología de habitación que necesite. Cómo más grande sea un cuadrado más caro será el precio, así cómo su tamaño
En el segundo caso se puede interactuar por el barrio y se muestra el promedio del precio de las habitaciones, siendo las zonas más caras las que tengan los círculos más grandes y más oscuros.



4. Resumen y conclusiones

a) Conclusiones finales

Tras analizar los datos de Airbnb en Madrid, hemos identificado varias tendencias y hallazgos clave que proporcionan una visión detallada del mercado de alquileres a corto plazo en la ciudad madrileña.

- Ocupación y Precios: Se ha observado una alta tasa de ocupación en los barrios céntricos y parte norte, con precios más altos en comparación con zonas más periféricas. Los alojamientos en estos barrios también han mostrado una mayor variabilidad en los precios..

- Tipos de Propiedades: Los apartamentos completos son el tipo de alojamiento más común. Los alquileres de habitaciones privadas también son populares. También sorprende la cantidad de habitaciones compartidas que hay

- Tendencias de Demanda: La mayor demanda en los barrios céntricos sugiere una preferencia por la proximidad a los principales atractivos turísticos y la vida nocturna. Así cómo la zona norte, que se trata de una zona más adinerada y con mayores cuidados. Esto podría estar influenciado por la facilidad de acceso y la riqueza de actividades disponibles en estas áreas.

- Impacto en la Comunidad: La concentración de alquileres a corto plazo en ciertas zonas puede estar contribuyendo al aumento de los precios de la vivienda, lo que afecta a los residentes locales (gentrificación). Además, la alta rotación de huéspedes podría estar generando preocupaciones sobre el ruido y la seguridad en algunos vecindarios.

Hay que reservar con anticipación y explorar opciones en barrios menos céntricos que puedan resultar más económicos y auténticos.

b) Lecciones aprendidas

Durante este trabajo y a lo largo del curso, hemos adquirido una serie de habilidades esenciales para el análisis de datos. Inicialmente, aprendimos a definir un dataset adecuado, en este caso, datos obtenidos de Airbnb en Madrid. Este proceso incluyó familiarizarnos con la estructura y contenido del dataset, comprendiendo la información recogida y su organización.

Posteriormente, nos sumergimos en la arquitectura y validación de los datos. Aprendimos a extraer muestras representativas y realizar exploraciones iniciales para identificar patrones y anomalías. Diseñamos e implementamos un Datawarehouse eficiente y comprendimos el proceso de ETL (Extract, Transform, Load) para asegurar la integridad y consistencia de los datos cargados.

En el análisis exploratorio, utilizamos R para realizar estudios estadísticos y determinar las métricas adecuadas para el dataset. Evaluamos y mejoramos la calidad de los datos, detectando y tratando valores atípicos y nulos. También utilizamos visualizaciones como boxplots y normalizamos los datos, corrigiendo inconsistencias que podrían afectar el análisis.

Para la visualización de métricas, identificamos y calculamos un KPI que nos ha proporcionado información valiosa para la toma de decisiones. Diseñamos un dashboard efectivo en Tableau, incorporando buenas prácticas de visualización, interactividad y campos calculados avanzados para mejorar la comprensión y funcionalidad.

Finalmente, desarrollamos un modelo de regresión lineal para predecir el precio de un inmueble en función de sus características, seleccionando variables relevantes y evaluando el rendimiento del modelo.

Este proyecto nos ha permitido obtener una comprensión profunda y práctica de todo el ciclo de vida del análisis de datos y aprender a utilizar herramientas y técnicas clave para el análisis y modelado de estos.