

# **Universidad Tecnológica Nacional**

**Facultad Regional Córdoba**

**Ingeniería en Sistema de información**



**Ciencia de Datos**

**Trabajo Práctico Integral**

**Informe Final**

**Grupo n°: 10**

**Alumnos:**

- Benjamin, Ian Nicolas, 80738
- Pedernera, Nicolas, 83201
- Ricse, Javier, 80122
- Charra Marquez, Giuliano, 63162

**Curso: 5K4**

**Fecha de presentación: 03/07/2024**

# Índice

[Visualización de Datos y Análisis Estadístico](#)

[Análisis y Comprensión de los Datos](#)

[Proceso ETL](#)

[Extracción](#)

[Transformación](#)

[Carga](#)

[Problemas identificados al analizar los diversos gráficos](#)

[Selección de Variables](#)

[Variables Complementarias](#)

[Variables a Descartar](#)

[Análisis de Modelos](#)

[Modelos Utilizados](#)

[Acciones para mejorar el modelo:](#)

[Mapa de Correlación Final](#)

[Resultados Obtenidos](#)

[Conclusion](#)

# Introducción

En este trabajo práctico integral, nuestro grupo analizará a fondo un dataset de Airbnb en Nueva York, que abarca tres etapas donde se aplicarán técnicas de Ciencia de Datos para obtener valiosos insights y realizar predicciones relacionadas a los alquileres de la plataforma.

## **Primera Etapa: Investigación y Adquisición de datos**

Nuestra primera tarea se enfoca en investigar y obtener un dataset de base que sirva como fundamento para nuestro análisis. Siguiendo esta tarea hemos seleccionado un conjunto de datos de Airbnb correspondiente a Nueva York. Este dataset contiene información sobre las propiedades y sus

comodidades, sus precios, sus reseñas, etc, que se convertirán en la materia prima de nuestra investigación.

## **Segunda Etapa: Análisis, Preparación y Exploración**

En la segunda etapa de nuestro trabajo, nos sumergimos en un análisis detallado y una comprensión profunda de los datos. Nuestro objetivo primordial es evaluar la calidad de los datos, identificando posibles errores, valores faltantes e inconsistencias. Posteriormente, determinamos estrategias adecuadas para abordar estos problemas, que pueden incluir la eliminación de datos defectuosos, la creación de nuevas columnas en base a otras existentes, etc. También seleccionamos las columnas y filas relevantes para nuestro análisis y realizamos el análisis de los datos, buscando relaciones entre variables y aplicando diferentes visualizaciones como Mapas de Calor, Matrices de Correlación, Gráficos de Torta, etc, lo que nos permitió entender y observar patrones.

## **Tercera Etapa: Modelado y Extracción de Conocimiento**

En la tercera y última etapa de nuestro proyecto, nos sumergimos en el proceso de modelado de datos según los requerimientos previamente establecidos. Utilizando el dataset objetivo y el análisis de datos realizado, seleccionamos los algoritmos adecuados para la extracción de patrones y definimos los valores de los parámetros correspondientes a cada técnica. En esta etapa utilizamos entre otros estos modelos: Random Forest y XGBoost. Recopilamos los resultados y, finalmente, nos adentramos en la etapa de interpretación del conocimiento obtenido, analizando los resultados de los distintos modelos.

# **Presentación de los Datasets**

En nuestro proceso de selección del dataset con el cual trabajaríamos tuvimos opciones para seleccionar:

- **Cosecha de vinos:** Un conjunto de datos clásico sobre las características de diferentes tipos de vino. Interesante para modelar la calidad del vino, pero no tenía las suficientes variables para analizar.
- **Fraude de Tarjetas de Crédito:** Un dataset sobre transacciones fraudulentas con tarjetas de crédito. Muy interesante para aplicar técnicas de detección de anomalías, pero no queríamos predecir un booleano.
- **Airbnb a nivel mundial:** Este dataset abarca información de anuncios de Airbnb a nivel mundial. Permitirá analizar tendencias globales en la industria de alojamiento, pero hay muchos lugares que carecen de datos.
- **Airbnb en Nueva York:** Este dataset contiene información detallada sobre anuncios de Airbnb en la ciudad de Nueva York. Es ideal para analizar tendencias de la industria de alojamiento. Este conjunto de datos es el que finalmente seleccionamos para nuestro trabajo.

### Información relevante sobre el dataset:

El dataset originalmente posee 103004 filas y 26 columnas que se describirán a continuación:

- **Id:** Identificación de publicación.
- **Name:** Nombre de publicación.
- **Host id:** Identificación del anfitrión.
- **Host Identity:** Verificación del anfitrión.

- **Neighbourhood Group:** Grupo de barrios al que pertenece la ubicación.
- **Neighbourhood:** Barrio al que pertenece la ubicación.
- **Lat y Long:** Coordenadas de la ubicación.
- **Country:** País al que pertenece.
- **Country Code:** Código del país.
- **Instant Bookable:** Reservable al momento.
- **Cancellation Policy:** Política de Cancelación.
- **Construction year:** Año de construcción.
- **Service Fee:** Tarifa a cobrar sobre el precio del alquiler.
- **Price:** Precio del alquiler.
- **Minimum Nights:** Noches mínimas para alquilar.
- **Number of Reviews:** Nro. en total de reseñas realizadas a la publicación.
- **Last Review:** Fecha de última reseña.
- **reviews per month:** Reseñas al mes.
- **review rate number:** Puntuación de la publicación.
- **calculated host listings count:** Publicaciones por anfitrión.
- **availability 365:** Disponibilidad de días en el año.
- **House rules:** Reglas del alquiler.
- **License:** Licencia del anfitrión.

## Captura del Dataset:

id	NAME	host id	host_identity	host name	neighbourhc	neighbourhc	lat	long	country	country code	instant_book	cancellation	room type	Construction	price	service fee	minimum ni	number of re	last review	reviews per r
1001254	Clean & quiet	8.0014E+10	unconfirmed	Madaline	Brooklyn	Kensington	4.064.749	-7.397.237	United States	US	FALSE	strict	Private room	2020	\$ 966	\$ 193	10	9	10/19/2021	0.21
1002102	Skylin Midtow	5.2335E+10	verified	Jenna	Manhattan	Midtown	4.075.362	-7.398.377	United States	US	FALSE	moderate	Entire home/	2007	\$ 142	\$ 28	30	45	5/21/2022	0.38
1002403	THE VILLAGE	7.8829E+10	verified	Elise	Manhattan	Harlem	4.080.902	-739.419	United States	US	TRUE	flexible	Private room	2005	\$ 620	\$ 124	3	0		
1002755		8.5098E+10	unconfirmed	Garry	Brooklyn	Clinton Hill	4.068.514	-7.395.976	United States	US	TRUE	moderate	Entire home/	2005	\$ 368	\$ 74	30	270	7/5/2019	4.64
1003689	Entire Apt. Sc	9.2038E+10	verified	Lyndon	Manhattan	East Harlem	4.079.851	-7.394.399	United States	US	FALSE	moderate	Entire home/	2009	\$ 204	\$ 41	10	9	11/19/2018	0.1
1004098	Large Cozy 1	4.5499E+10	verified	Michelle	Manhattan	Murray Hill	4.074.767	-73.975	United States	US	TRUE	flexible	Entire home/	2013	\$ 577	\$ 115	3	74	6/22/2019	0.59
1004650	BlissArtsSpa	6.1301E+10	verified	Alberta	Brooklyn	Bedford-Stuy	4.068.688	-7.395.596	United States	US	FALSE	moderate	Private room	2015	\$ 71	\$ 14	45	49	10/5/2017	0.4
1005202	BlissArtsSpa	9.0822E+10	unconfirmed	Emma	Brooklyn	Bedford-Stuy	4.068.688	-7.395.596	United States	US	FALSE	moderate	Private room	2009	\$ 1.06	\$ 212	45	49	10/5/2017	0.4
1005754	Large Furnis	7.9384E+10	verified	Evelyn	Manhattan	Hell's Kitch	4.076.489	-7.398.493	United States	US	TRUE	strict	Private room	2005	\$ 1.02	\$ 204	2	430	6/24/2019	3.47
1006307	Cozy Clean G	7.5528E+10	unconfirmed	Carl	Manhattan	Upper West S	4.080.178	-7.396.723	United States	US	FALSE	strict	Private room	2015	\$ 291	\$ 58	2	118	7/21/2017	0.99
1006859	Cute & Cozy i	1.280143094	verified	Miranda	Manhattan	Chinatown	4.071.344	-7.399.037	United States	US	FALSE	flexible	Entire home/	2004	\$ 319	\$ 64	1	160	6/9/2019	1.33
1007411	Beautiful Libr	1.8825E+10	verified	Alan	Manhattan	Upper West S	4.080.316	-7.396.545	United States	US	TRUE	flexible	Entire home/	2008	\$ 606	\$ 121	5	53	6/22/2019	0.43
1007964	Central Manl	8.8136E+10	verified		Manhattan	Hell's Kitch	4.076.076	-7.398.867	United States	US	FALSE	strict	Private room	2008	\$ 714	\$ 143	2	188	6/23/2019	1.5
1008516	Lovely Room	2.6802E+10	verified	Darcy	brooklin	South Slope	4.066.829	-7.398.779	United States	US	TRUE	moderate	Private room	2010	\$ 580	\$ 116	4	167	6/24/2019	1.34
1009068	Wonderful G	8.892E+10	verified	Leonardo	Manhattan	Upper West S	4.079.826	-7.396.113	United States	US	FALSE	flexible	Private room	2019	\$ 149	\$ 30	2	113	7/5/2019	0.91
1009621	West Village	4.6552E+10	verified	Daniel	Manhattan	West Village	407.353	-7.400.525	United States	US	TRUE	flexible	Entire home/	2018	\$ 578		90	27	10/31/2018	0.22
1010173	Only 2 stops	6.2566E+10	unconfirmed	Heather	Brooklyn	Williamsbur	4.070.837	-7.395.352	United States	US		moderate	Entire home/	2009	\$ 778		2	148	6/29/2019	1.2
1010725	Perfect for Yc	8.038E+10	verified	Ryan	Brooklyn	Fort Greene	4.069.169	-7.397.185	United States	US		flexible	Entire home/	2006	\$ 656		2	198	6/28/2019	1.72
1011277	Chelsea Perfi	7.3863E+10	verified	Alberta	manhattan	Chelsea	4.074.192	-7.398.501	United States	US		moderate	Private room	2008	\$ 460		1	260	7/1/2019	2.12
1011830	Hip Historic i	7.2145E+10	verified	Martin	Brooklyn	Crown Heigh	4.067.592	-7.394.694	United States	US		moderate	Entire home/	2004	\$ 1.10		3	53	6/22/2019	4.44
1012382	Huge 2 BR Ug	7.9805E+10	verified	Audrey	Manhattan	East Harlem	4.079.685	-7.394.872	United States	US		moderate	Entire home/	2013	\$ 281	\$ 56	7	0		
1012934	Sweet and Sc	8.6555E+10	verified	Alissa	Brooklyn	Williamsbur	4.071.842	-7.395.718	United States	US		flexible	Entire home/	2016	\$ 477	\$ 95	3	9	12/28/2021	0.07
1013487	CBG CypBd i	5.3754E+10	verified	Mary	Brooklyn	Park Slope	4.068.069	-7.397.706	United States	US		moderate	Private room	2013	\$ 133	\$ 27	2	130	7/1/2019	1.09
1014039	CBG Helos Hi	8.7669E+10	verified	William	Brooklyn	Park Slope	4.067.989	-7.397.798	United States	US		moderate	Private room	2017	\$ 1.05	\$ 210	1	39	1/1/2019	0.37

Fuente: [AirBnB Dataset price prediction \(kaggle.com\)](https://www.kaggle.com/datasets/airbnb-price-prediction)

## Justificación de la elección del dataset:

- **Relevancia:** Elegimos este dataset debido a su relevancia en el contexto actual. El mercado de alquiler a corto plazo en Nueva York siempre tiene una alta demanda debido al gran turismo que genera en sí la ciudad como también por su calidad de centro político y económico mundial , y entender los patrones y tendencias en Airbnb puede ser útil para los anfitriones, los turistas y las autoridades locales.
- **Disponibilidad:** El dataset de Airbnb en Nueva York es accesible, se encuentra en la pagina de Kaggle y presenta una actualización constante.
- **Interés académico:** Nos interesa profundizar en el análisis de datos y aprender más sobre cómo los factores como la ubicación, las características de las propiedades y las revisiones de los huéspedes pueden influir en los precios y la satisfacción.

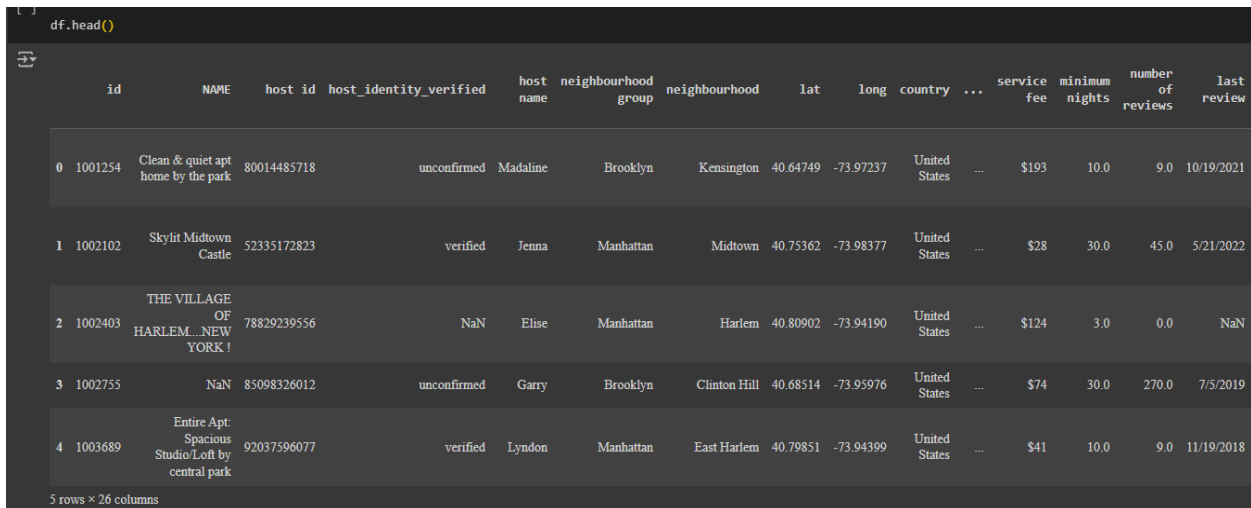
# Proceso de Extracción, Transformación y Carga

## Contexto Inicial:

En el transcurso del Trabajo Práctico Integrador de Ciencia de Datos, investigamos a través de los datos de Airbnb, específicamente en las ciudades de Buenos Aires y Nueva York. La fase inicial se centró en la visualización y exploración de datos para desentrañar patrones y peculiaridades que podrían influir en la predicción de precios de alquiler.

## Visualización de Datos y Análisis Estadístico

Para comenzar, utilizamos el método `head()` para visualizar las primeras filas del DataFrame, lo que nos permite comprender mejor su estructura.



	id	NAME	host_id	host_identity_verified	host_name	neighbourhood_group	neighbourhood	lat	long	country	...	service_fee	minimum_nights	number_of_reviews	last_review
0	1001254	Clean & quiet apt home by the park	80014485718	unconfirmed	Madaline	Brooklyn	Kensington	40.64749	-73.97237	United States	...	\$193	10.0	9.0	10/19/2021
1	1002102	Skylit Midtown Castle	52335172823	verified	Jenna	Manhattan	Midtown	40.75362	-73.98377	United States	...	\$28	30.0	45.0	5/21/2022
2	1002403	THE VILLAGE OF HARLEM...NEW YORK !	78829239556	NaN	Elise	Manhattan	Harlem	40.80902	-73.94190	United States	...	\$124	3.0	0.0	NaN
3	1002755	NaN	85098326012	unconfirmed	Garry	Brooklyn	Clinton Hill	40.68514	-73.95976	United States	...	\$74	30.0	270.0	7/5/2019
4	1003689	Entire Apt. Spacious Studio/Loft by central park	92037596077	verified	Lyndon	Manhattan	East Harlem	40.79851	-73.94399	United States	...	\$41	10.0	9.0	11/19/2018

5 rows x 26 columns

El DataFrame consta de 26 columnas, aunque solo mostramos algunas de ellas para poder apreciar mejor los datos que contiene.



Para obtener una idea general de cómo se distribuyen los datos en cada columna numérica, realizamos un análisis estadístico utilizando el método `describe()`.

```
[ ] df.describe()
```

	id	host_id	lat	long	Construction year	price	service fee	minimum nights	number of reviews	reviews per month	review rate number	calculated host listings count	availability 365
count	9.927600e+04	9.927600e+04	99267.000000	99267.000000	99073.000000	99034.000000	99022.000000	98875.000000	99095.000000	83879.000000	98972.000000	98990.000000	98827.000000
mean	2.841613e+07	4.925738e+10	40.728121	-73.949615	2012.487822	624.985409	124.965715	8.158139	27.497240	1.376668	3.284090	8.078129	142.332359
std	1.582818e+07	2.854869e+10	0.055845	0.049594	5.764952	331.724845	66.336253	30.791336	49.713378	1.750687	1.281015	32.670293	135.502384
min	1.001254e+06	1.236005e+08	40.499790	-74.249840	2003.000000	50.000000	10.000000	-1223.000000	0.000000	0.010000	1.000000	1.000000	-10.000000
25%	1.470873e+07	2.455438e+10	40.688720	-73.982590	2007.000000	339.000000	68.000000	2.000000	1.000000	0.220000	2.000000	1.000000	4.000000
50%	2.841613e+07	4.912444e+10	40.722350	-73.954470	2012.000000	624.000000	125.000000	3.000000	7.000000	0.750000	3.000000	1.000000	99.000000
75%	4.212352e+07	7.402010e+10	40.762770	-73.932270	2017.000000	913.000000	182.000000	5.000000	30.000000	2.010000	4.000000	2.000000	270.000000
max	5.583092e+07	9.876313e+10	40.916970	-73.705220	2022.000000	1200.000000	240.000000	5645.000000	1024.000000	90.000000	5.000000	332.000000	3677.000000

Este análisis estadístico nos brinda una visión general sobre la distribución, centralización y variabilidad de los datos, permitiéndonos identificar rápidamente posibles valores atípicos y comprender mejor la naturaleza de las variables que estamos analizando.

## Visualización de Datos y Análisis Estadístico

Para comenzar, utilizamos el método `head()` para visualizar las primeras filas del DataFrame, lo que nos permite comprender mejor su estructura.

```
df.head(5)
```

	id	NAME	host id	host_identity_verified	host name	neighbourhood group	neighbourhood	lat	long	country	...	service fee	minimum nights	number of reviews	last review
0	1001254	Clean & quiet apt home by the park	80014485718	unconfirmed	Madaline	Brooklyn	Kensington	40.64749	-73.97237	United States	...	\$193	10.0	9.0	10/19/2021
1	1002102	Skiyllt Midtown Castle	52335172823	verified	Jenna	Manhattan	Midtown	40.75362	-73.98377	United States	...	\$28	30.0	45.0	5/21/2022
2	1002403	THE VILLAGE OF HARLEM...NEW YORK!	78829239556	NaN	Elise	Manhattan	Harlem	40.80902	-73.94190	United States	...	\$124	3.0	0.0	NaN
3	1002755	NaN	85098326012	unconfirmed	Garry	Brooklyn	Clinton Hill	40.68514	-73.95976	United States	...	\$74	30.0	270.0	7/5/2019
4	1003689	Entire Apt: Spacious Studio/Loft by central park	92037596077	verified	Lyndon	Manhattan	East Harlem	40.79851	-73.94399	United States	...	\$41	10.0	9.0	11/19/2018

5 rows x 26 columns

El DataFrame consta de 26 columnas, aunque solo mostramos algunas de ellas para poder apreciar mejor los datos que contiene.

Para obtener una idea general de cómo se distribuyen los datos en cada columna numérica, realizamos un análisis estadístico utilizando el método `describe()`.

```
df.describe()
```

✓ 0.1s

	id	host id	lat	long	Construction year	minimum nights	number of reviews	reviews per month	review rate number	calculated host listings count
count	1.025990e+05	1.025990e+05	102591.000000	102591.000000	102385.000000	102190.000000	102416.000000	86720.000000	102273.000000	102280.000000
mean	2.914623e+07	4.925411e+10	40.728094	-73.949644	2012.487464	8.135845	27.483743	1.374022	3.279106	7.936605
std	1.625751e+07	2.853900e+10	0.055857	0.049521	5.765556	30.553781	49.508954	1.746621	1.284657	32.218780
min	1.001254e+06	1.236005e+08	40.499790	-74.249840	2003.000000	-1223.000000	0.000000	0.010000	1.000000	1.000000
25%	1.508581e+07	2.458333e+10	40.688740	-73.982580	2007.000000	2.000000	1.000000	0.220000	2.000000	1.000000
50%	2.913660e+07	4.911774e+10	40.722290	-73.954440	2012.000000	3.000000	7.000000	0.740000	3.000000	1.000000
75%	4.320120e+07	7.399650e+10	40.762760	-73.932350	2017.000000	5.000000	30.000000	2.000000	4.000000	2.000000
max	5.736742e+07	9.876313e+10	40.916970	-73.705220	2022.000000	5645.000000	1024.000000	90.000000	5.000000	332.000000

Este análisis estadístico nos brinda una visión general sobre la distribución, centralización y variabilidad de los datos, permitiéndonos identificar rápidamente posibles valores atípicos y comprender mejor la naturaleza de las variables que estamos analizando.

## Análisis y Comprensión de los Datos

Se llevó a cabo un análisis exhaustivo de los datos con el objetivo de identificar posibles problemas de calidad, como datos erróneos, faltantes e inconsistentes.

Para obtener una idea de la cantidad de filas que contiene cada columna, utilizamos el método `info()`. Este método nos muestra la cantidad de valores no nulos presentes en cada columna, así como el tipo de dato de cada una. Esta información es crucial para determinar cómo tratar cada columna en el análisis posterior.

```
Data columns (total 26 columns):
```

#	Column	Non-Null Count	Dtype
0	id	102599 non-null	int64
1	NAME	102349 non-null	object
2	host id	102599 non-null	int64
3	host_identity_verified	102310 non-null	object
4	host name	102193 non-null	object
5	neighbourhood group	102570 non-null	object
6	neighbourhood	102583 non-null	object
7	lat	102591 non-null	float64
8	long	102591 non-null	float64
9	country	102067 non-null	object
10	country code	102468 non-null	object
11	instant_bookable	102494 non-null	object
12	cancellation_policy	102523 non-null	object
13	room type	102599 non-null	object
14	Construction year	102385 non-null	float64
15	price	102352 non-null	object
16	service fee	102326 non-null	object
17	minimum nights	102190 non-null	float64
18	number of reviews	102416 non-null	float64
19	last review	86706 non-null	object
...			
24	house_rules	50468 non-null	object
25	license	2 non-null	object

# Proceso ETL

## Extracción

Al principio, exploramos diversas fuentes como CONICET, NASA, y el Banco Mundial de Datos en busca de un dataset adecuado. Después de una extensa búsqueda, encontramos el dataset de Airbnb en Kaggle. Este dataset fue ideal debido a su gran cantidad de datos y la variedad de columnas disponibles, lo que nos permitió realizar análisis detallados y predicciones precisas en nuestro trabajo final.

## Transformación

Examinamos las diferentes categorías del dataset utilizando gráficos y análisis de tendencias de valores. Este enfoque nos permitió identificar varios problemas que debían corregirse, como valores faltantes, formatos incorrectos e inconsistencias en los datos. Estos hallazgos fueron esenciales para limpiar y preparar el dataset para un análisis más profundo y preciso.

## Carga

Después de completar el proceso de transformación, obtuvimos un nuevo dataset limpio y estructurado, que está listo para ser utilizado en nuestras predicciones finales. Este dataset refinado nos permitirá realizar análisis más precisos y confiables.

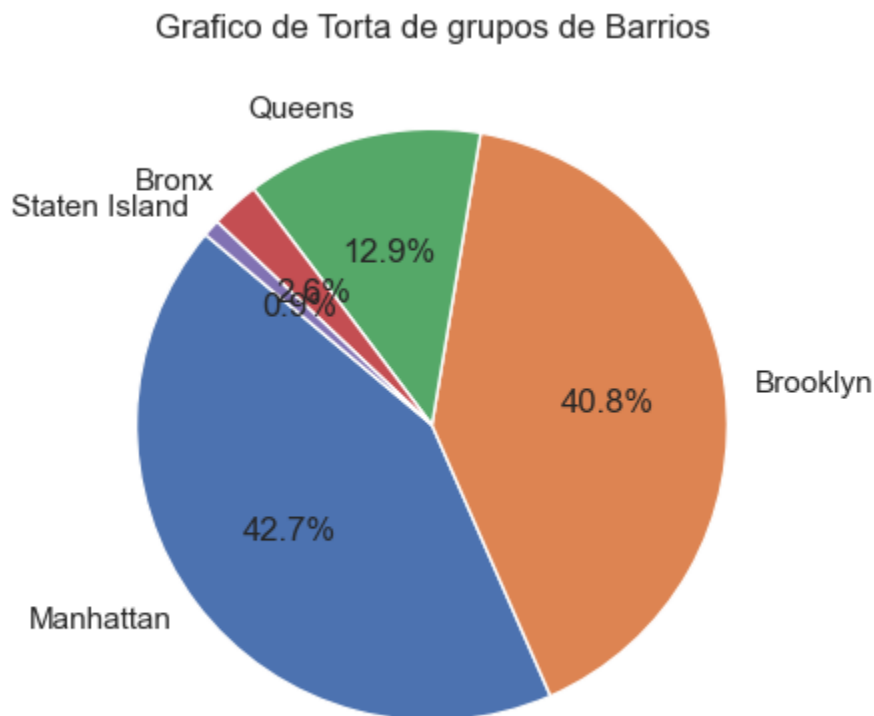
## Problemas identificados al analizar los diversos gráficos

### Categorías Repetidas:

En la categoría "neighbourhood group", observamos que algunos distritos están escritos de manera inconsistente.

neighbourhood group	
Bronx	2712
Brooklyn	41842
Manhattan	43792
Queens	13267
Staten Island	955
brookln	1
manhatan	1

Para un análisis más preciso, estandarizamos el formato de todos los distritos escritos de manera inconsistente.



## Valores faltantes:

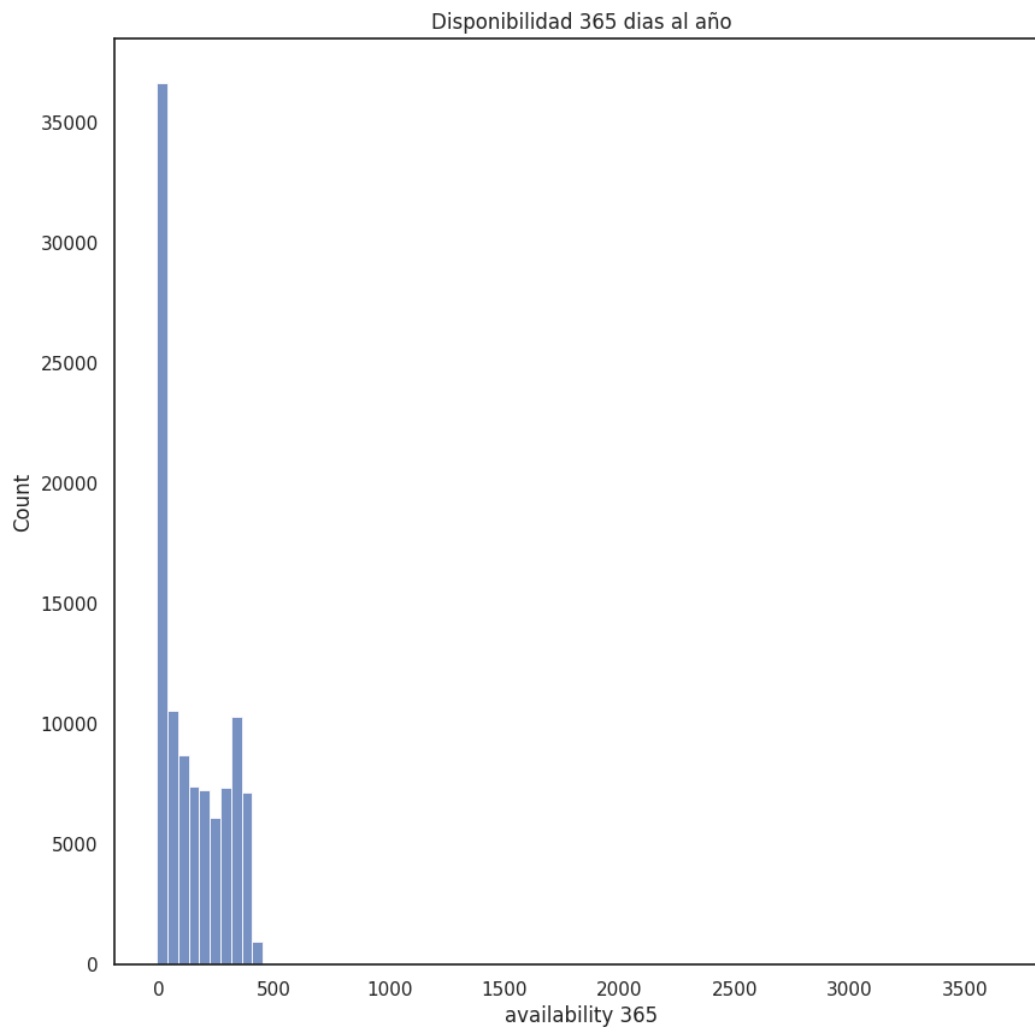
Como se puede observar, encontramos una gran cantidad de valores nulos. En algunas categorías, la cantidad de valores nulos es baja, por lo que decidimos eliminarlos, ya que no representan un problema significativo. En cambio, para aquellas categorías con un número mayor de valores nulos, optamos por usar la moda para variables categóricas y la mediana para variables numéricas, con el fin de reemplazar los valores faltantes.

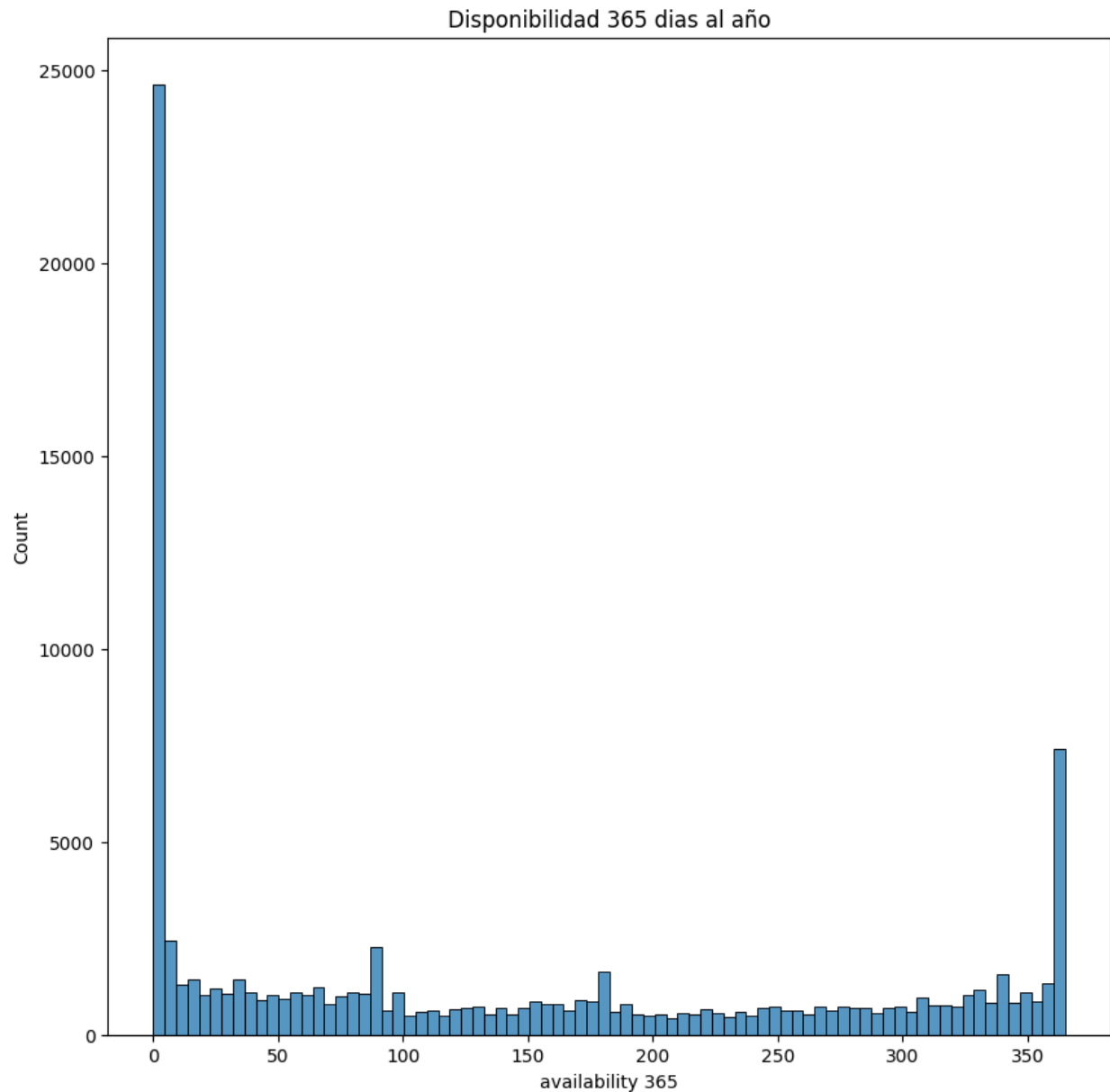
neighbourhood group	28
neighbourhood	15
lat	8
long	8
instant_bookable	96
cancellation_policy	75
room type	0
Construction year	190
price	241
service fee	253
minimum nights	0
number of reviews	178
reviews per month	15324
review rate number	282
calculated host listings count	285
availability 365	424

## Gran cantidad de valores de columnas muy altos en 0 y valores inconsistentes:

Al analizar la variable "disponibilidad 365 días", notamos que contiene valores fuera de los límites esperados (valores entre 0 y 365). Para corregir esto, transformamos los valores negativos a sus equivalentes absolutos. Además, cualquier valor superior a 365 fue ajustado al límite de 365, considerando que podría tratarse de un error de entrada de datos.

Además, observamos un gran número de valores igual a 0. Para abordar esto, consideramos que estos valores representan que la publicación o alquiler no está disponible por lo que creamos una nueva columna llamada "esta disponible", que utiliza valores booleanos: 0 para "no disponible" y 1 para "disponible" para representar esta situación.



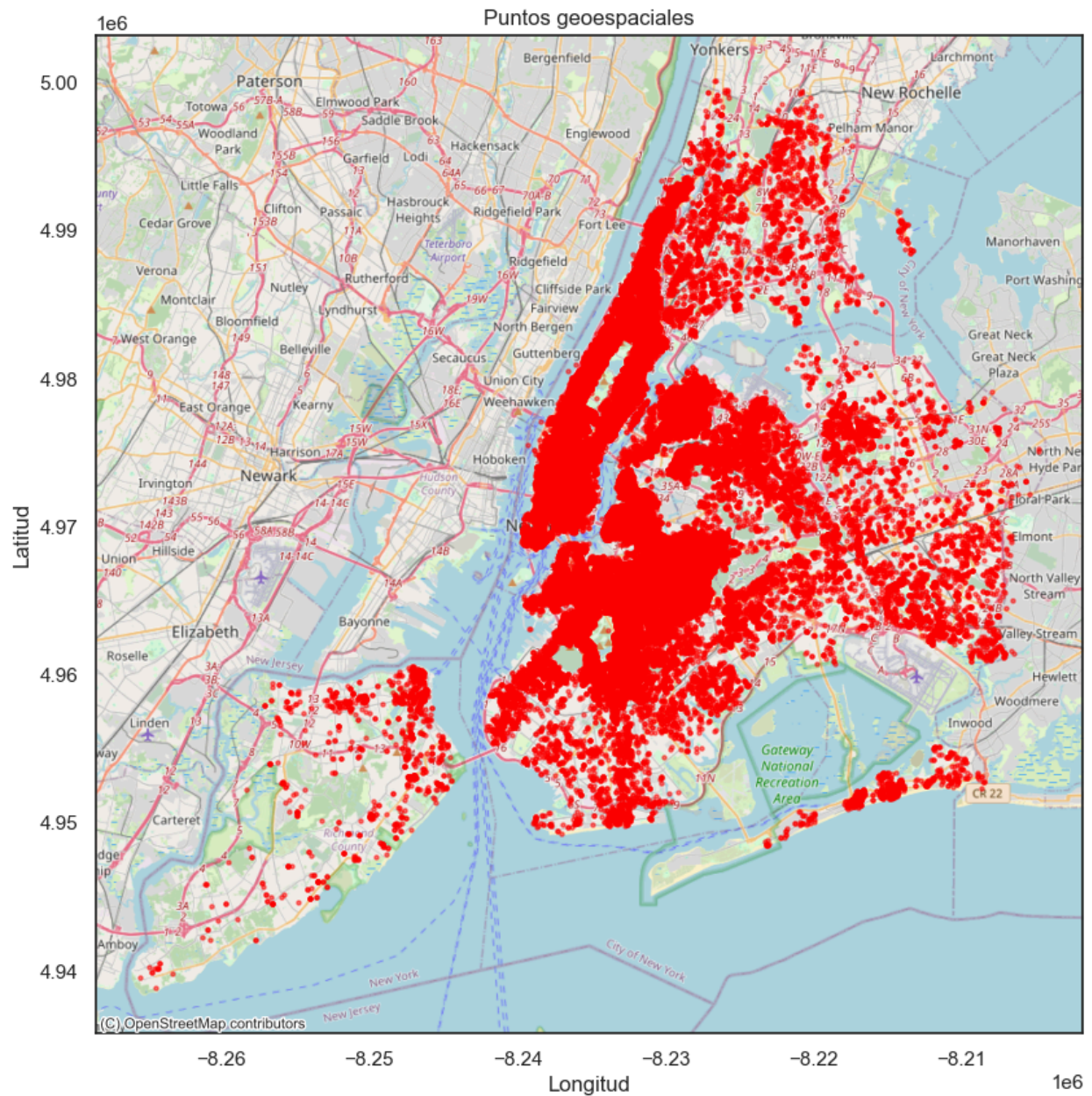


Aquí se puede apreciar mejor el gráfico después de haber transformado los valores fuera de los límites establecidos los valores están comprendidos entre 0 y 365 días.

### **Reincorporación de las variables de latitud y longitud**

Inicialmente decidimos eliminar las variables de latitud y longitud, ya que pensamos que no serían útiles para nuestro análisis y predicción. Sin embargo, luego decidimos reincorporarlas.



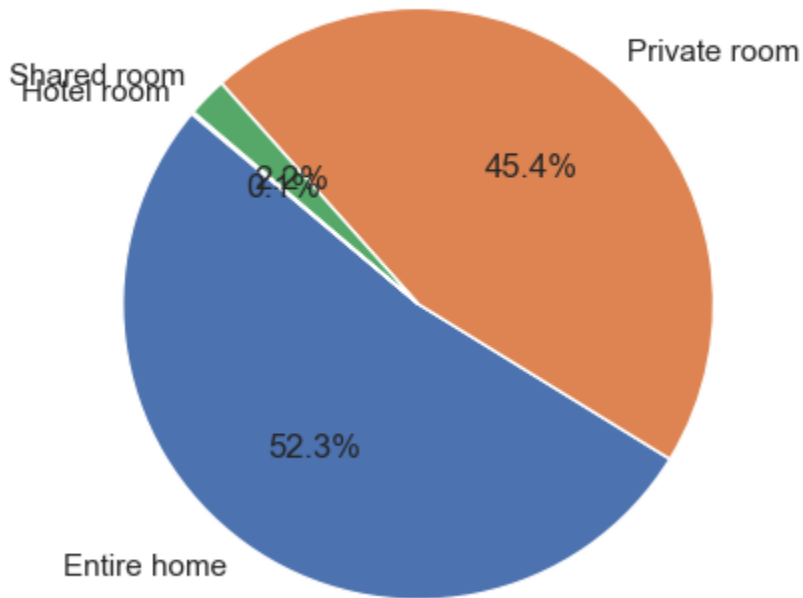


Podemos observar una distribución más clara de los distritos que ofrecen Airbnb mediante el gráfico geográfico. Esto nos ayudará a predecir la variable de precio más adelante, dado que el precio varía según el distrito en el que se encuentre la propiedad.

## Análisis de variables

## Tipo de habitación

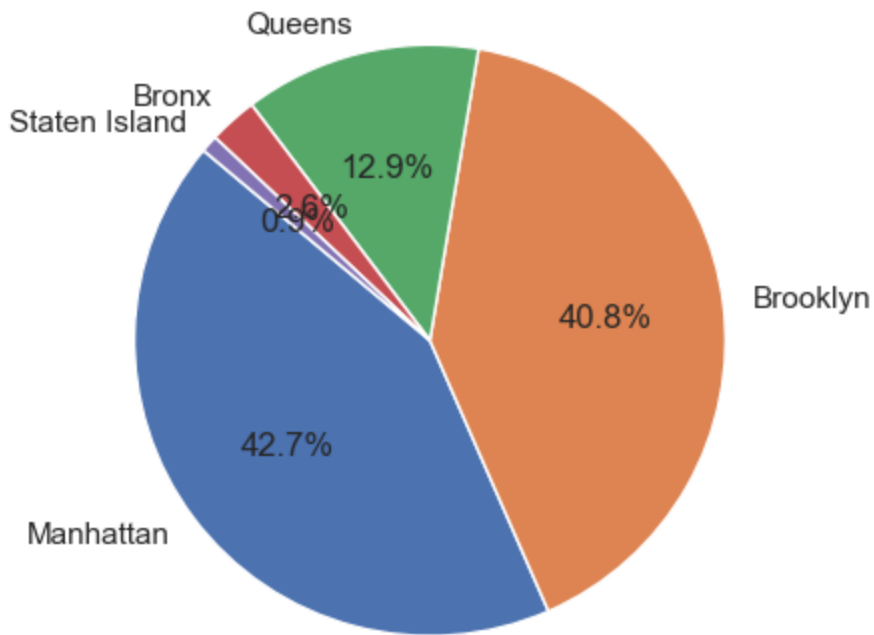
Grafico de Torta de Tipos de Habitaciones



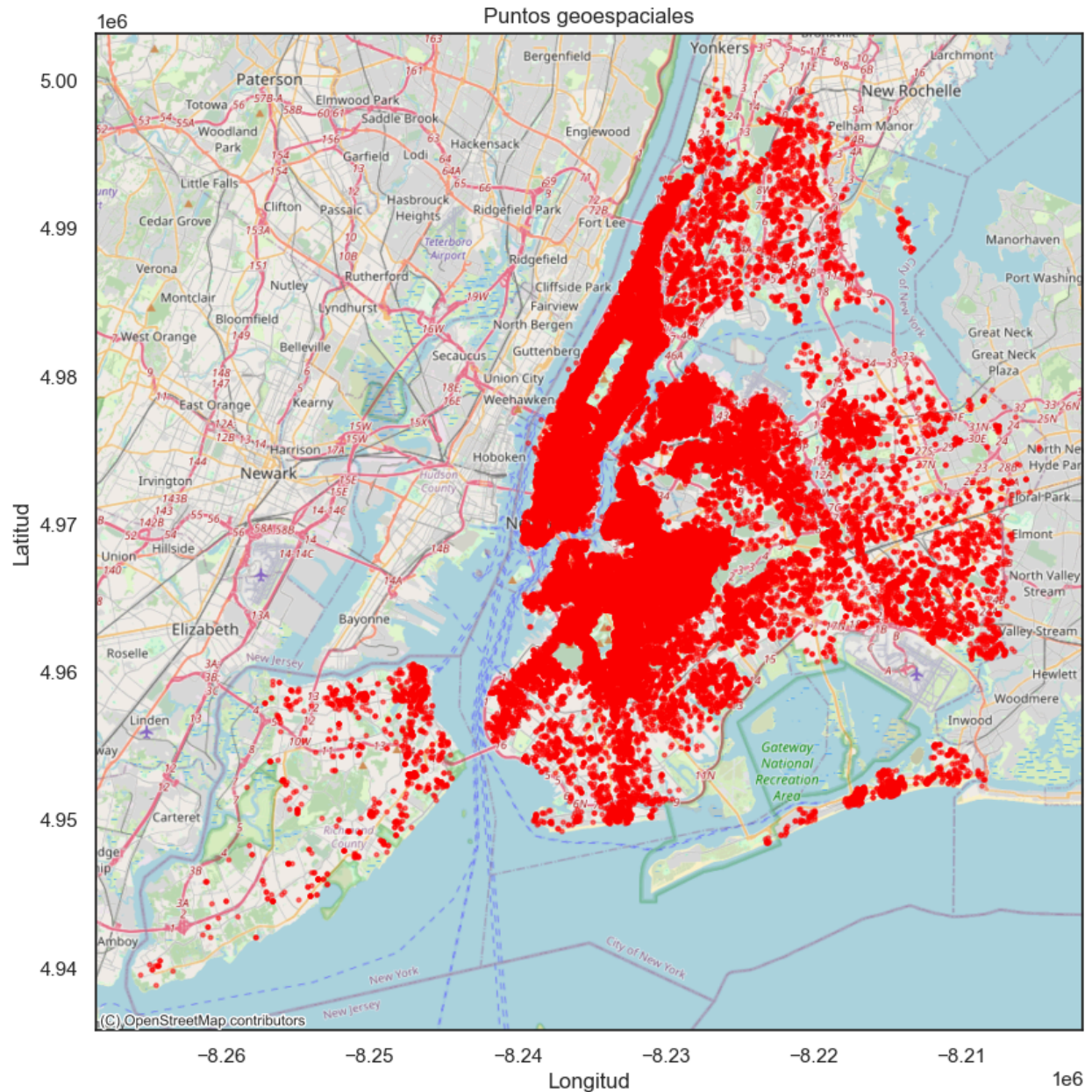
Podemos observar que la mayor parte de las publicaciones realizadas son habitaciones privadas o casa completa. Lo que es consistente con el modelo de negocio de la plataforma, donde particulares publican sus propiedades y no hoteles y moteles.

## Grupos de Barrios o Distritos

Grafico de Torta de grupos de Barrios

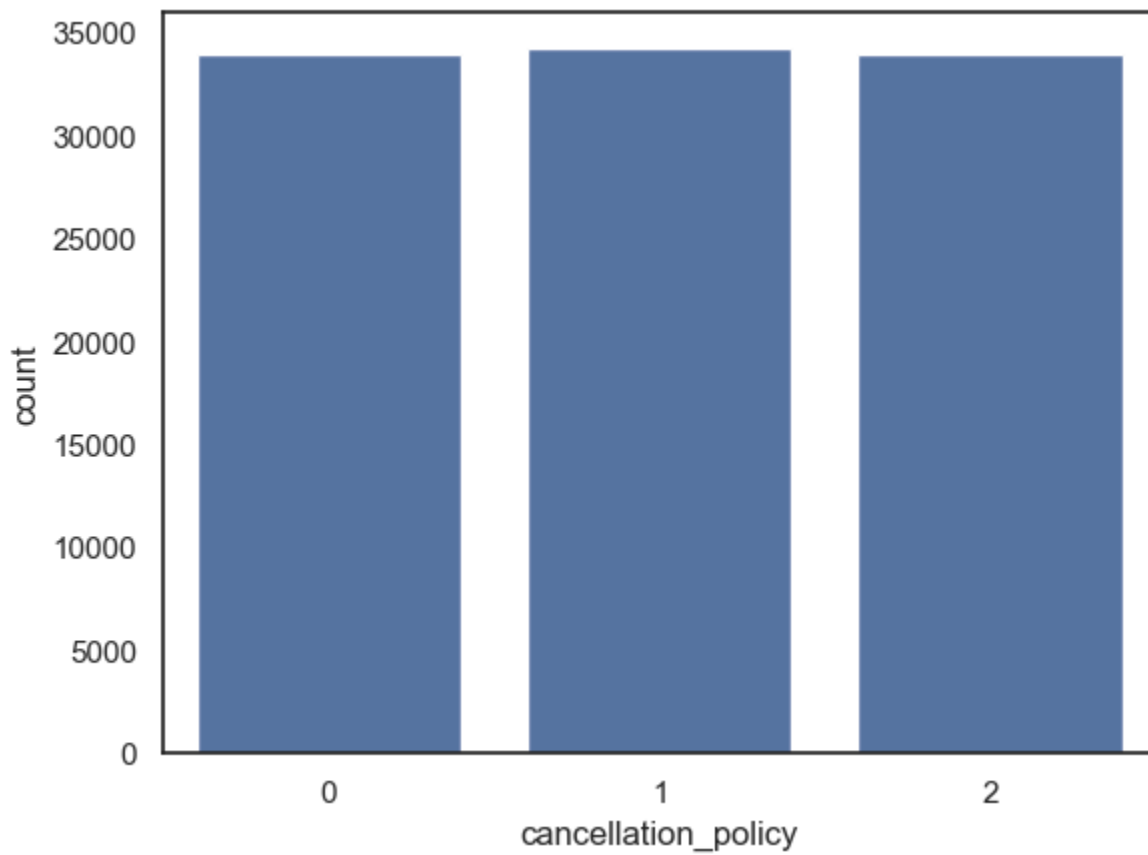


Podemos analizar a través del gráfico que la mayor parte de los alquileres se ubican en los distritos de Brooklyn y de Manhattan, los cuales son los mas famosos y mas visitados por el turismo. En un menor porcentaje se ven representados los distritos de Bronx y Staten Island, que por cuestiones de inseguridad y lejanía del centro de la ciudad no son muy solicitadas estas ubicaciones.



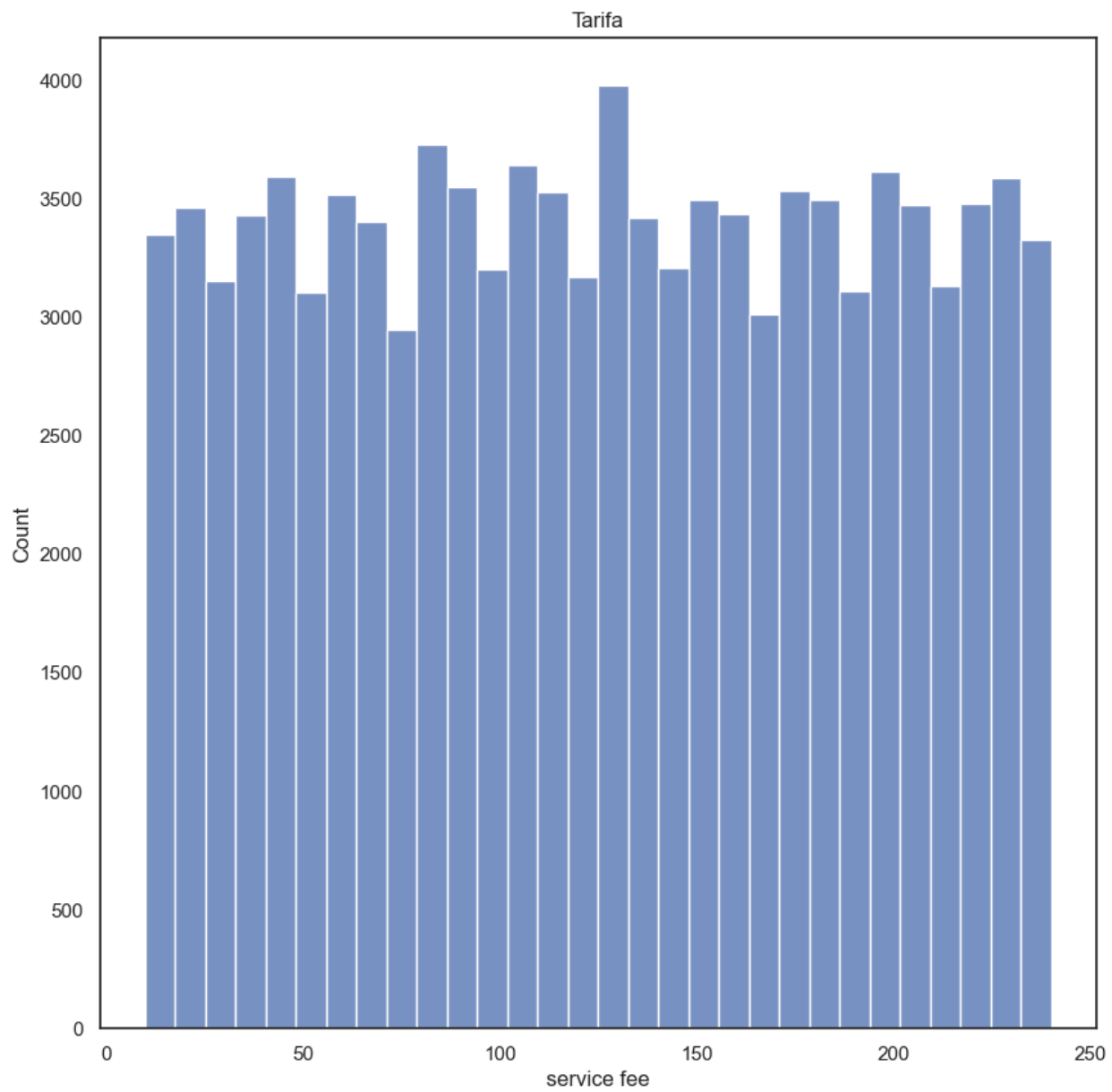
En el mapa si visualiza lo mencionado, hay una mayor densidad de propiedades en el centro del mapa, donde se ubica manhattan y Brooklyn, mientras que en la esquina inferior izquierda se ubica Staten Island, en el sector derecho Quenns y en la parte superior el Bronx. En estos se ve una marcada disminuci3n de la densidad de publicaciones.

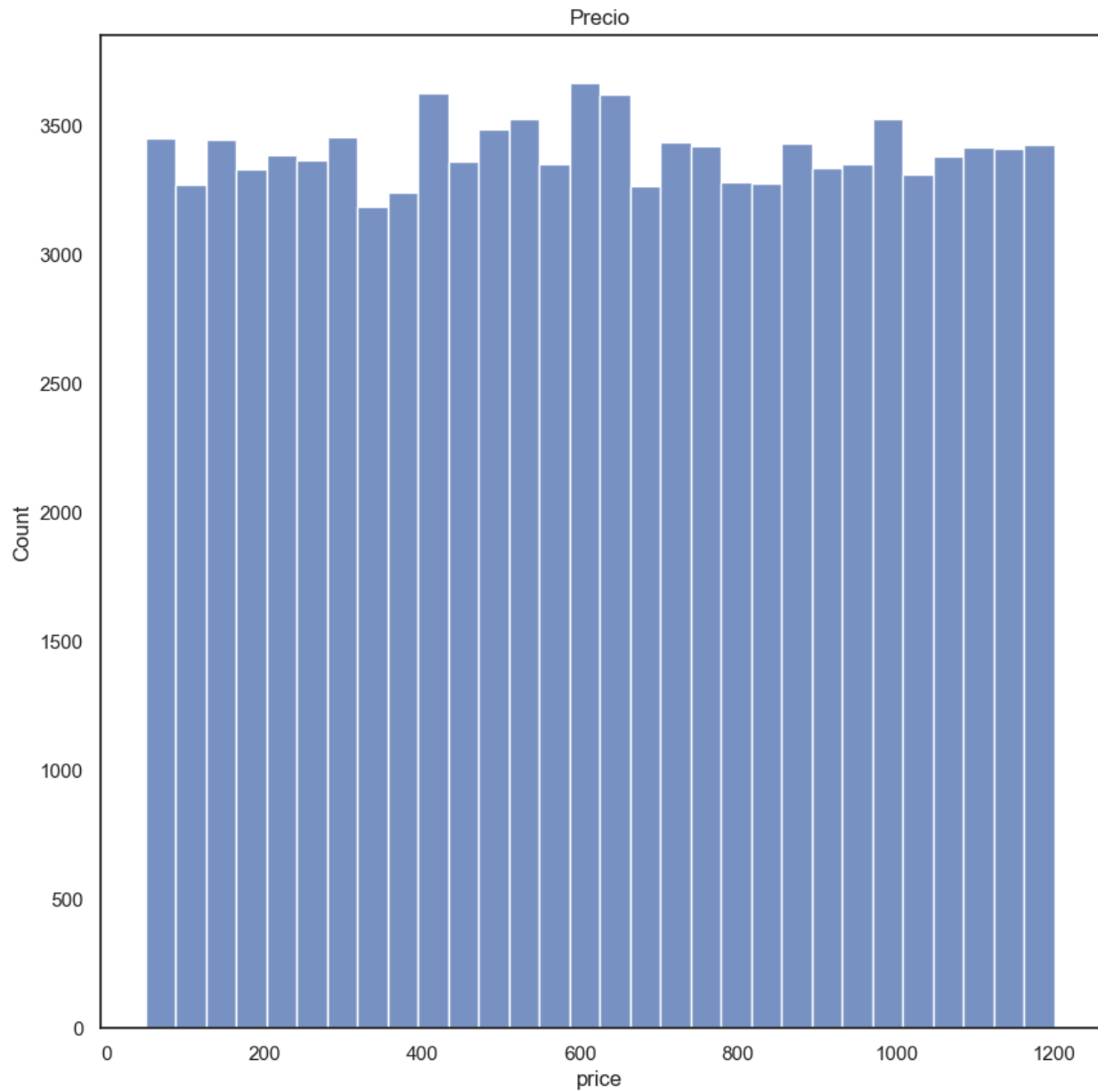
## Política de Cancelación



Se visualiza que en la columna cada categoría (flexible, moderado y estricto, en ese orden del gráfico) presenta un número similar al resto.

## Precio y Tarifa

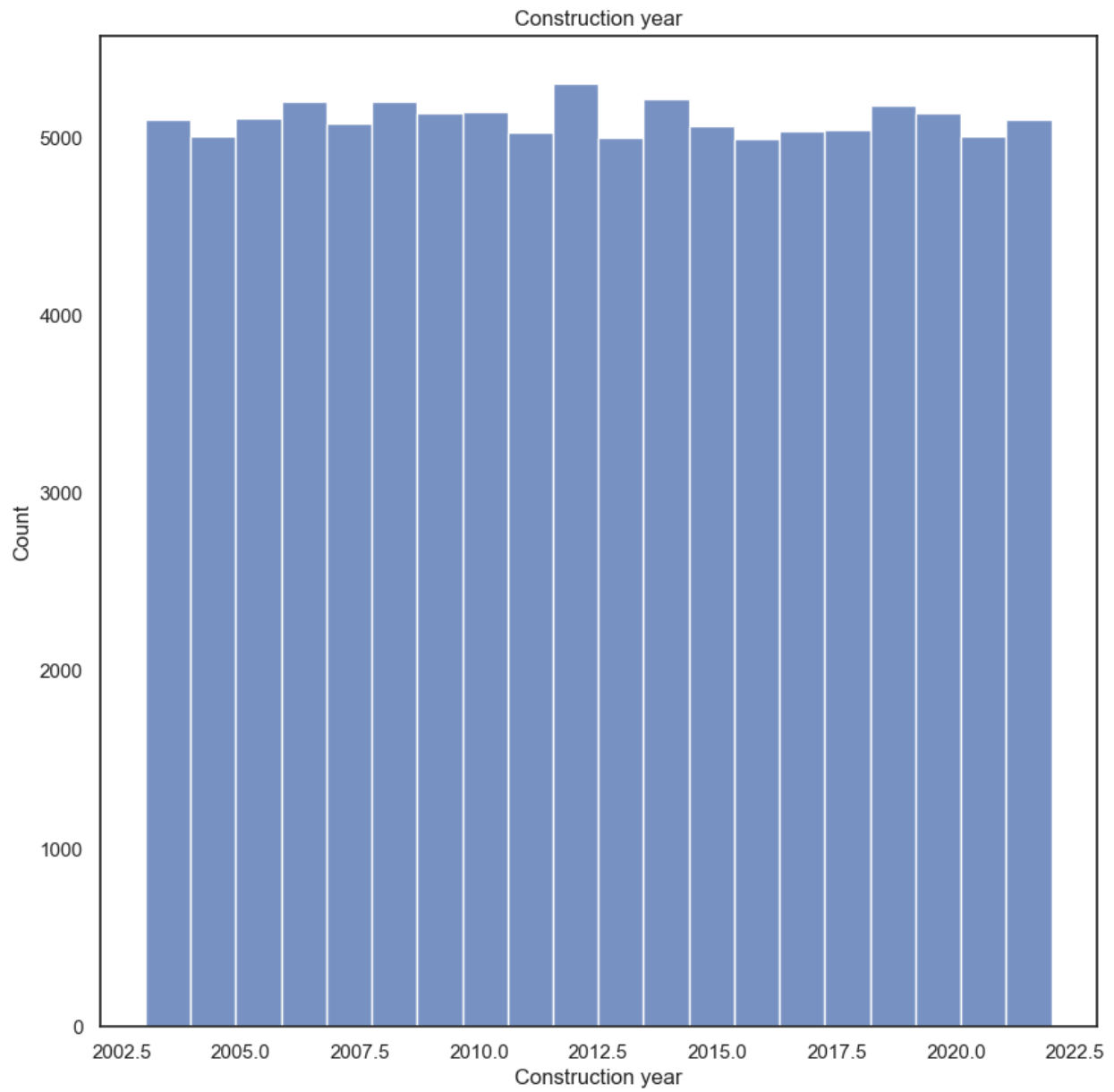




Podemos visualizar en ambos histogramas que los precios y las tarifas presentan una distribución similar y uniforme. Lo que indica que hay un numero similar de propiedades para cada precio y tarifa.

### **Año de Construcción**

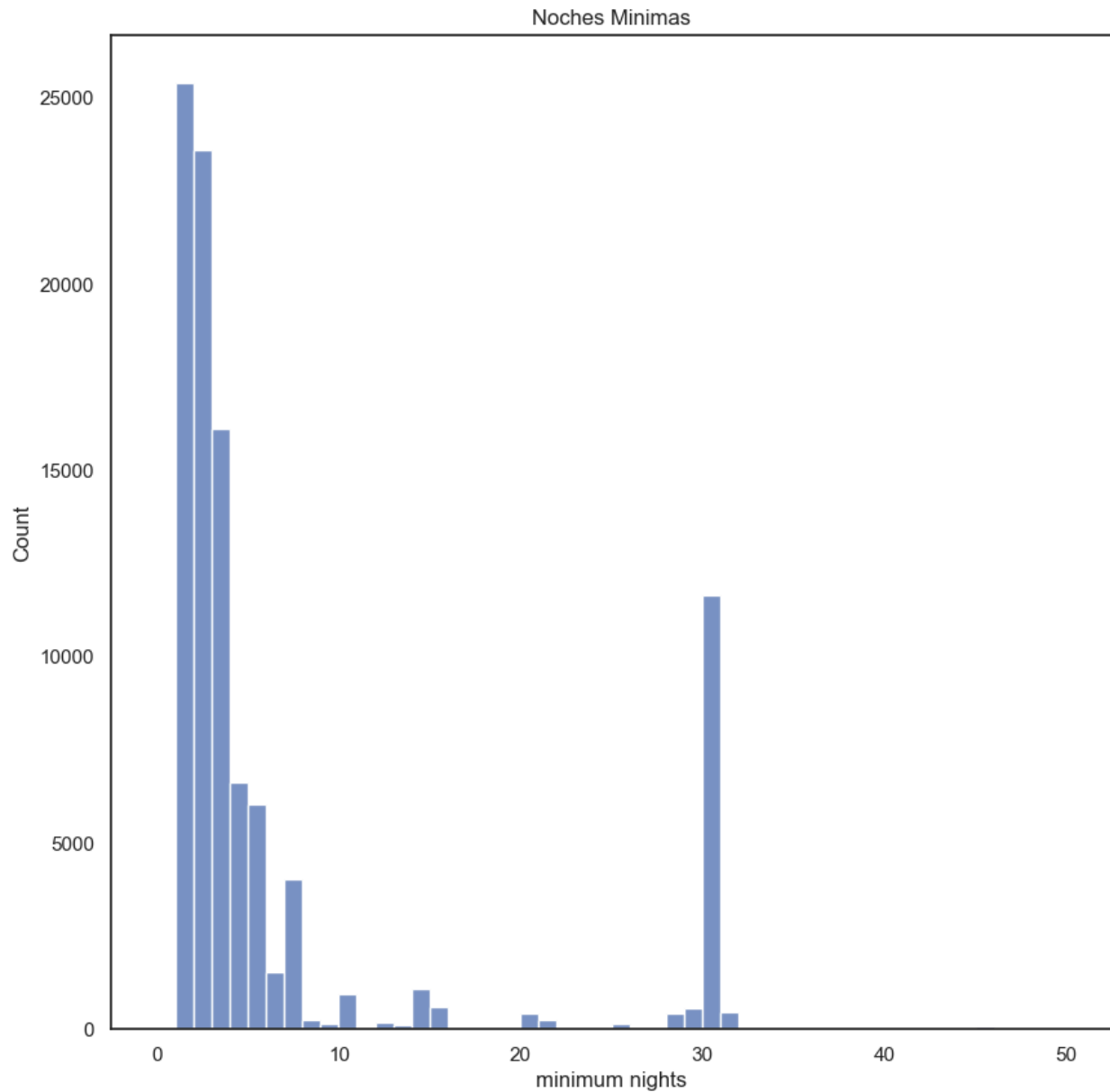




Presenta una distribución uniforme.

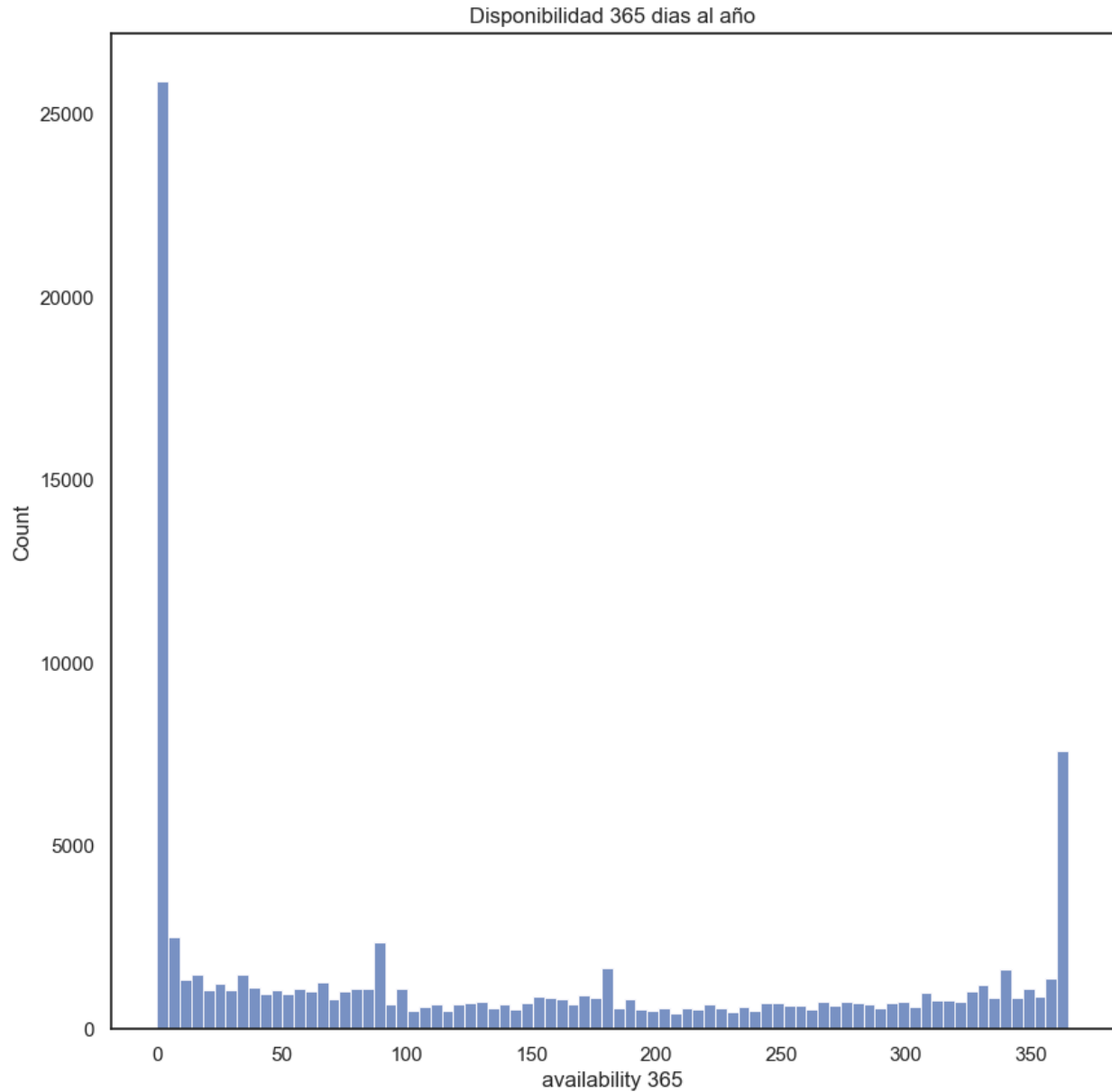
**Noches Mínimas**





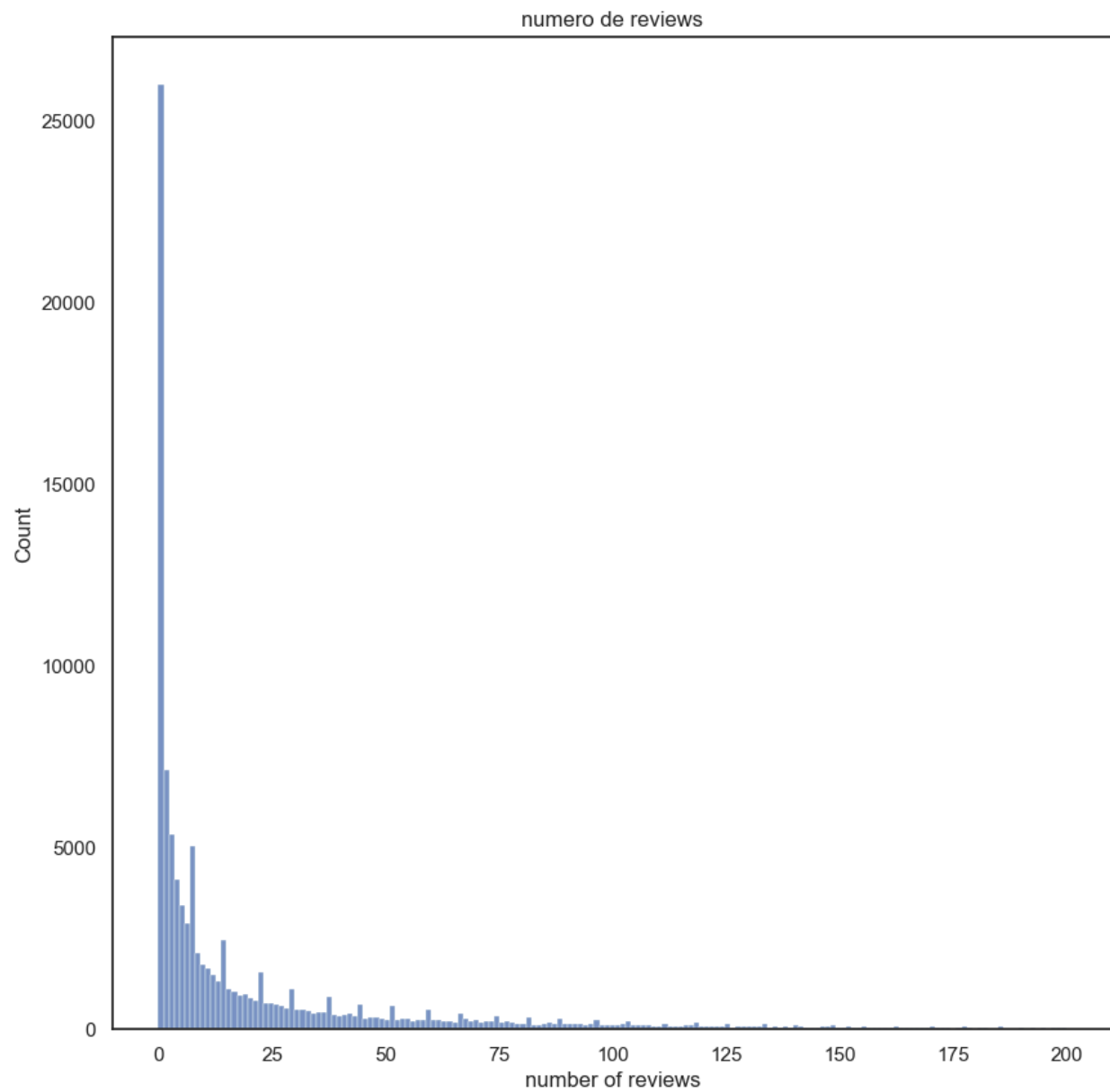
Observamos una distribucion exponencial negativa donde la mayor parte de las propiedades solamente requieren entre 1 y 5 días minimos para reservar.

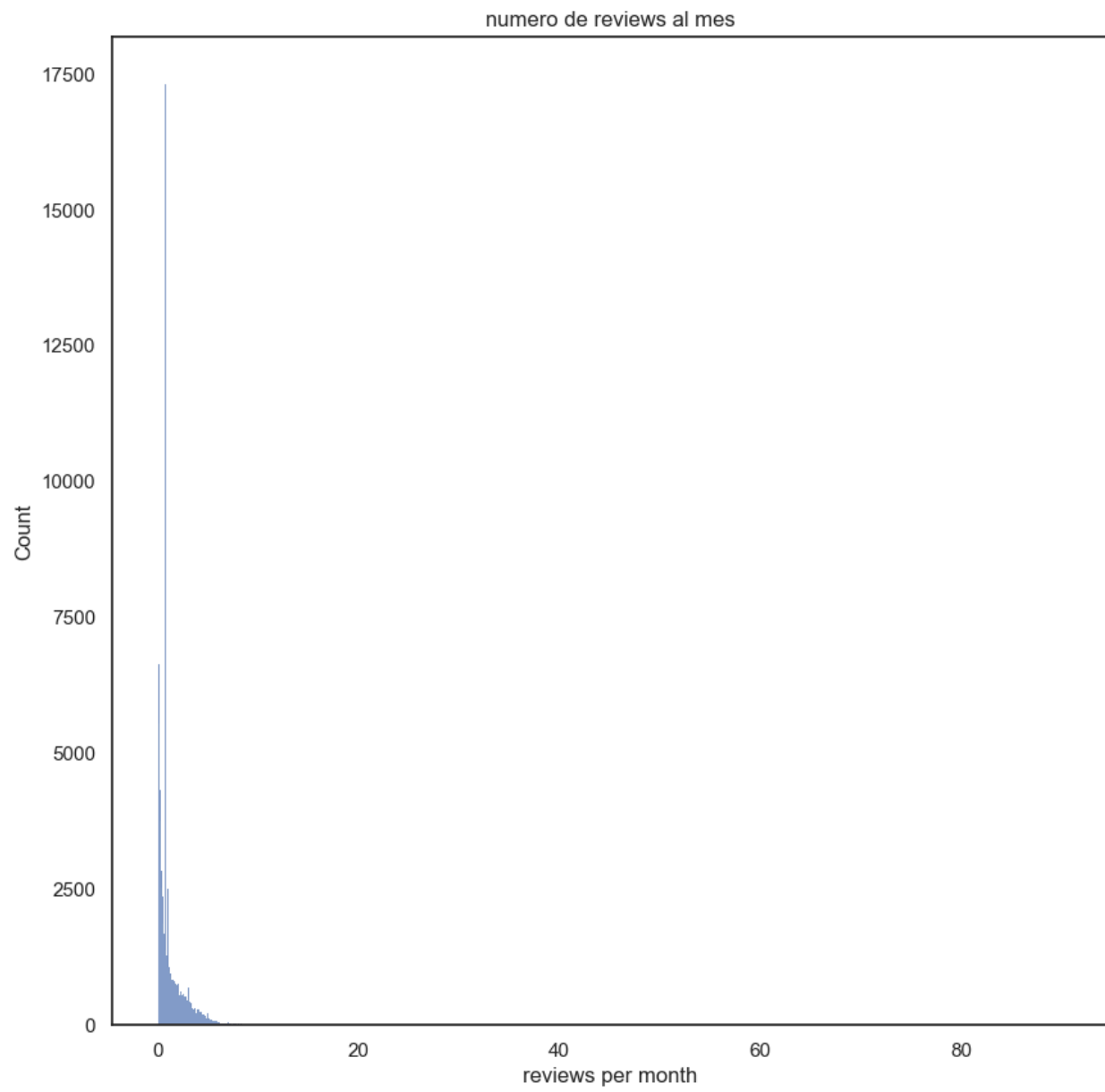
### Disponibilidad 365

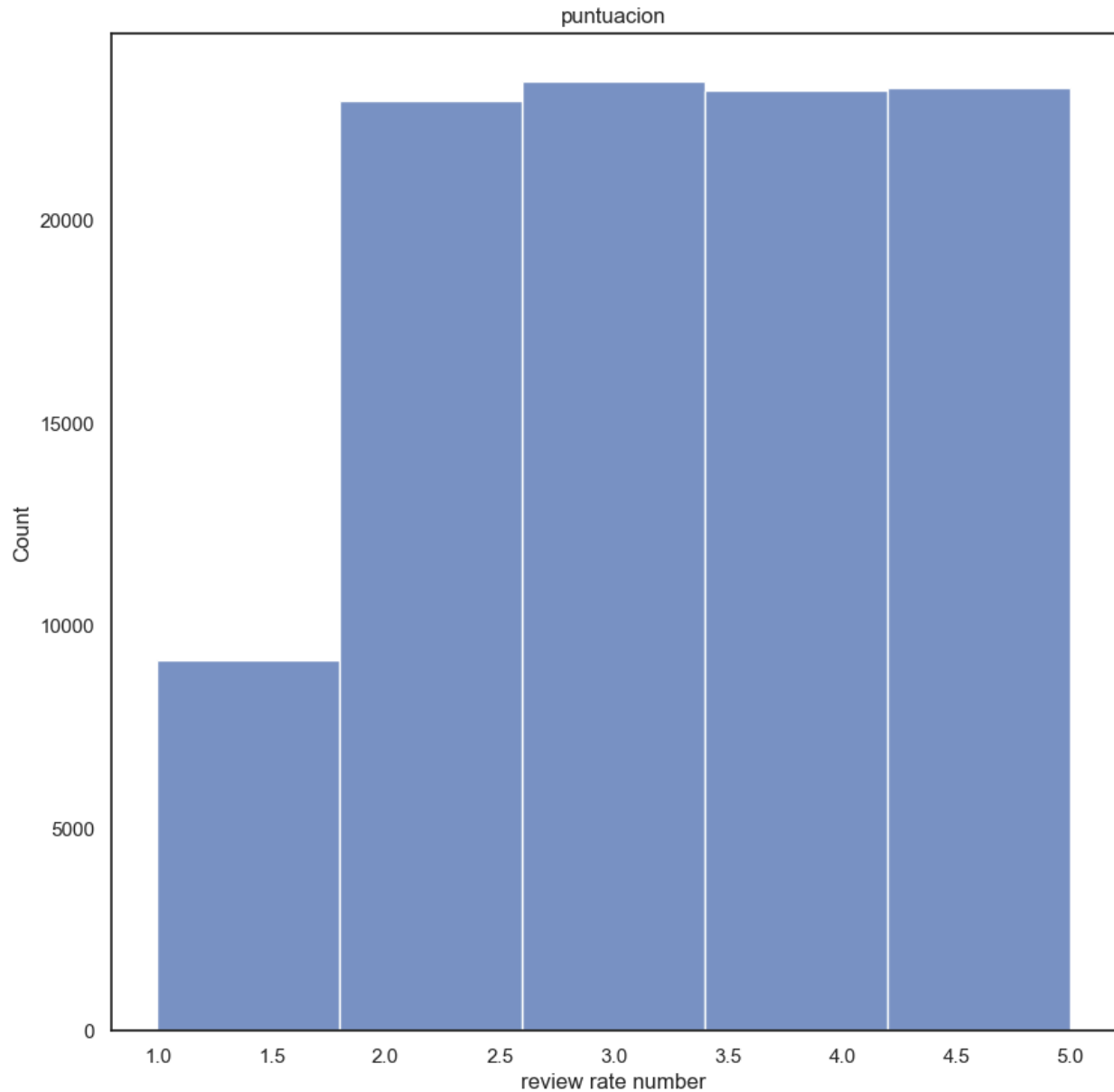


Visualizamos que una gran parte de las propiedades tienen disponibilidad 0, lo que indica que estos alquileres no están disponibles en el momento.

### **Número de reviews, Reviews al mes y Puntuación**







En Reviews por mes y Numero de reviews podemos analizar que presentan una distribucion exponencial negativa donde la mayor parte de las publicaciones tienen pocas reseñas o reviews. En cambio la columna puntuacion presenta que hay una cantidad pareja en todas opciones que presenta dicha categoria.

# Selección de Variables

## **Variables Claves**

Después de un análisis exhaustivo, llegamos a la conclusión de las variables clave que consideramos fundamentales para la predicción del precio, tales como el barrio, grupo de barrios, tipo de alquiler, tipo de habitación, número de reseñas, latitud, longitud, etc.

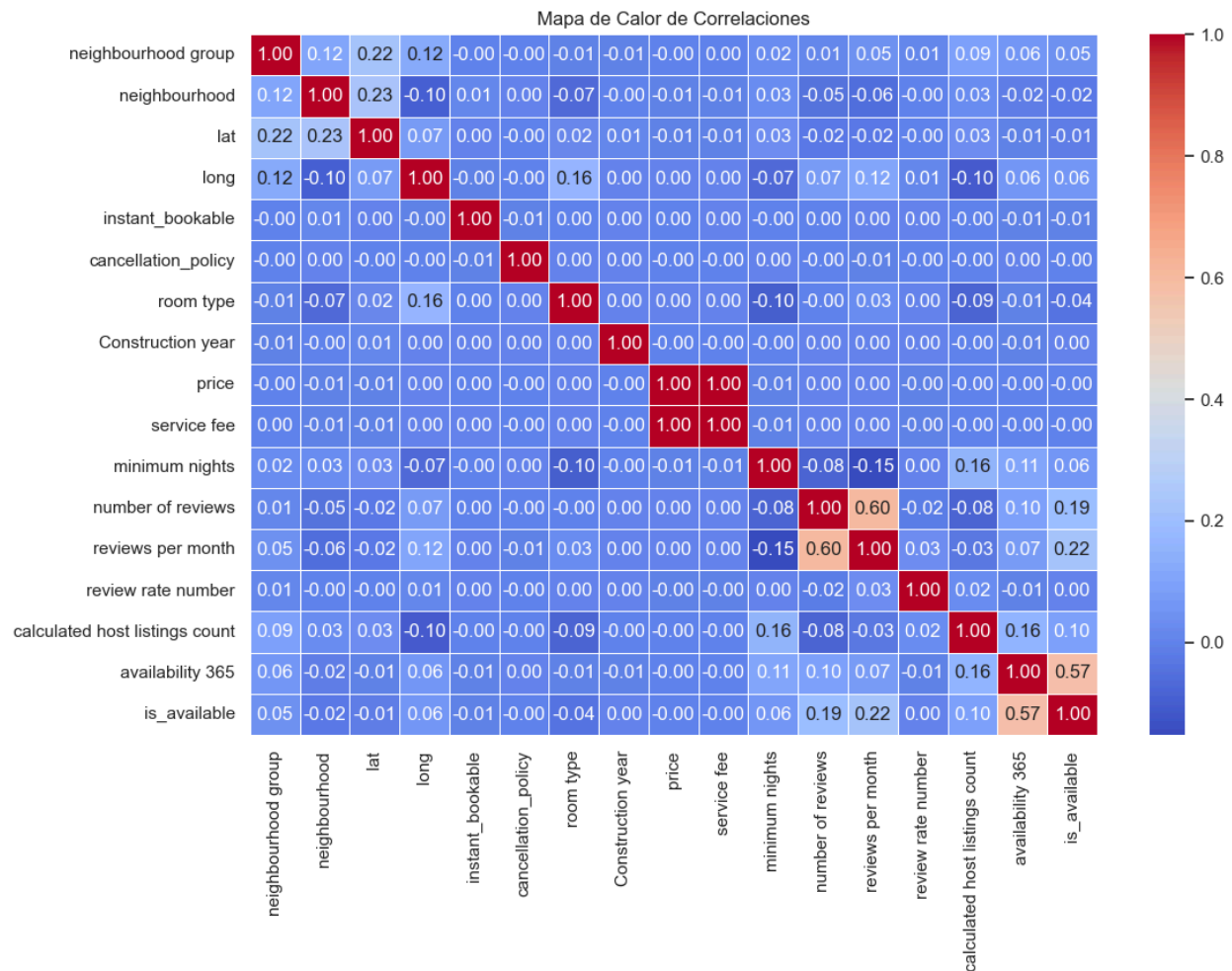
## **Variables Complementarias**

Consideramos variables adicionales que podrían aportar información relevante para la predicción, como las reseñas, la disponibilidad anual y las noches mínimas.

## **Variables a Descartar**

Decidimos eliminar variables con poca relevancia como: año de construcción, reserva instantánea, id, host id, licencia, reglas de casa, etc.

## Gráfico de correlación



Podemos observar en el gráfico de correlación que la mayoría de las variables tienen una relación muy débil con la variable de precio. Esto complica la predicción del precio, ya que la mayoría de las variables tienen una correlación baja.

## Analisis de Modelos

### Métricas de Desempeño

Las métricas que emplearemos para evaluar la eficacia del modelo de predicción son las siguientes:

- **R-cuadrado ( $R^2$ ):** Esta métrica indica la proporción de la variabilidad en la variable dependiente que es explicada por el modelo. Un valor más cercano a 1 indica que el modelo ajusta bien los datos, mientras que valores bajos sugieren que el modelo no explica bien la variabilidad.
- **Error Cuadrático Medio (ECM):** El ECM mide el promedio de los errores al cuadrado entre los valores predichos por el modelo y los valores reales. Es una medida de la precisión del modelo, donde valores más bajos indican una menor dispersión de errores y, por lo tanto, un mejor ajuste del modelo.

Estas métricas se utilizan conjuntamente para evaluar la precisión y el ajuste de los modelos predictivos, proporcionando una comprensión clara de cómo se comportan los modelos en términos de predicción y explicación de la variabilidad en los datos.

### **Importancia de las variables**

Analizamos qué variables tienen mayor impacto en las predicciones de cada modelo

### **Ajuste de Hiperparametros**

Optimizamos los hiperparámetros de los modelos para mejorar su rendimiento.

## **Modelos Utilizados**

Dividimos el conjunto de datos en dos partes: un conjunto de entrenamiento, que representa el 80% del total para entrenar el modelo, y un conjunto de prueba, que comprende el restante 20%, utilizado para evaluar el rendimiento del modelo.

- **Regresión lineal:** Utiliza vectores de soporte para encontrar la mejor línea de ajuste (o hiperplano en espacios de mayor dimensión) que maximice el margen entre los puntos de datos y la línea de ajuste.



```
Error cuadrático medio: 109526.63102742819
Error absoluto medio (MAE): 285.7515307393088
Coeficiente de determinación (R-cuadrado): -0.0002097744538702706
```

Los valores obtenidos de las métricas de desempeño indican que el modelo no explica adecuadamente la variabilidad de los datos y carece de una capacidad predictiva satisfactoria.

- **Random forest:** Es un conjunto de árboles de decisión, donde cada árbol es entrenado de forma independiente y el resultado final se promedia o se combina.

```
Error cuadrático medio (Random Forest): 70733.88745632286
Coeficiente de determinación (R-cuadrado, Random Forest): 0.3540500154595563
```

Los valores de las métricas de desempeño indican que el modelo explica alrededor del 35.41% de la variabilidad de los datos, lo que muestra una capacidad predictiva moderada.

- **SVR:** Utiliza vectores de soporte para encontrar la mejor línea de ajuste (o hiperplano en espacios de mayor dimensión) que maximice el margen entre los puntos de datos y la línea de ajuste.

```
Error cuadrático medio (SVR): 110005.5324868445
Coeficiente de determinación (R-cuadrado, SVR): -0.004583157586493636
```

Los valores obtenidos de las métricas de desempeño indican que el modelo no explica adecuadamente la variabilidad de los datos y carece de una capacidad predictiva satisfactoria.

- **Redes Neuronales:** Modelos inspirados en el funcionamiento del cerebro humano, compuestos por capas de neuronas interconectadas que aprenden a partir de datos.

- **Error cuadrático medio:** 109527.6
- **R-cuadrado:** -0.0002

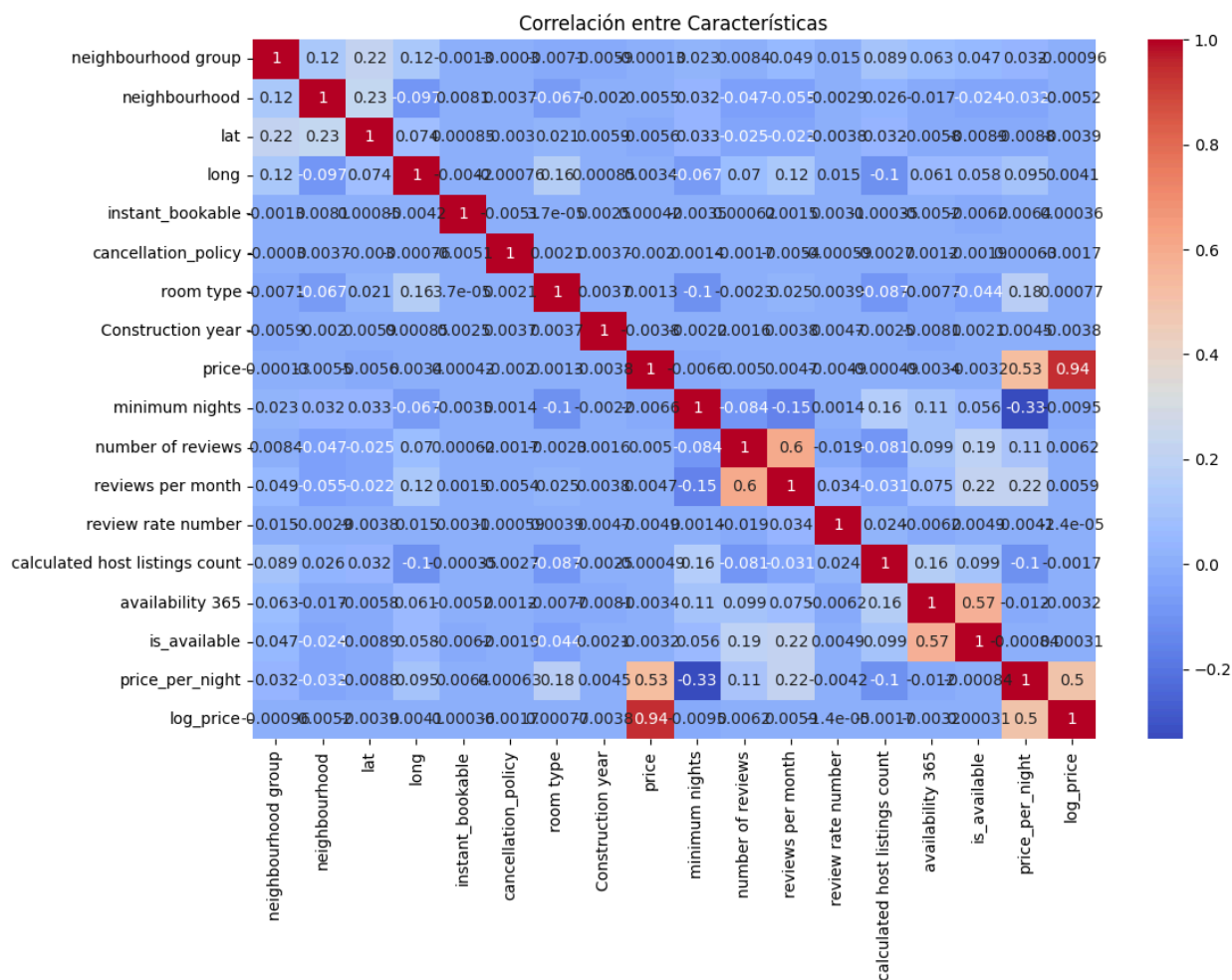
Al igual que los modelos anteriores, podemos observar que este modelo no predice adecuadamente la variabilidad de los datos y carece de una capacidad predictiva satisfactoria.

## Acciones para mejorar el modelo:

1. Eliminar columnas irrelevantes como barrios, grupos de barrios, política de cancelación, reserva instantánea y año de construcción.
2. Evitar el sobreajuste del modelo.
3. Aplicar validación cruzada y ajustar hiperparámetros para optimizar el rendimiento.
4. Realizar transformación logarítmica de la columna de precio.
5. Introducir una nueva variable para el análisis del precio por noche.

Número final de columnas después de las mejoras: 13.

# Mapa de Correlación Final



Después de reintegrar algunas columnas que inicialmente consideramos innecesarias, observamos una correlación más fuerte con la variable de precio que estamos prediciendo.

## Resultados Obtenidos

Analizamos la columna de precios y aplicamos una transformación logarítmica para mitigar posibles sesgos en su distribución. Además, realizamos ingeniería de características al crear la columna de precio por noche, calculada como el cociente entre el precio y el número mínimo de noches.

```
Resultados en el Conjunto de Prueba:  
MSE: 143.9813189263409  
MAE: 4.996829779014178  
R2: 0.9986851460582876
```

Podemos observar que el modelo explica aproximadamente el 99.87% de la variabilidad de los datos, lo cual sugiere un ajuste muy cercano a los datos observados y una capacidad predictiva excelente. Sin embargo, hemos notado la presencia de valores extremos que pueden estar afectando los resultados. Por esta razón, hemos decidido cambiar la variable de predicción.

- **Error cuadrático medio:** 14750.7
- **R-cuadrado:** 0.72

Al utilizar el modelo de Random Forest con la variable precio por noche, observamos que el modelo explica aproximadamente el 72% de la variabilidad de los datos. Esto indica un ajuste razonablemente bueno del modelo a los datos observados. Obtenemos un resultado mejorado en comparación con el uso de la variable precio.

## Conclusion

Para concluir nuestro trabajo destacamos la importancia de la limpieza y preparación de datos, ya que la calidad de los datos es fundamental para lograr predicciones precisas. Hemos explorado y aplicado técnicas de ingeniería de características para identificar las variables más relevantes que influyen en los precios de alquiler. Además, hemos utilizado algoritmos de aprendizaje automático y técnicas de evaluación de modelos para crear un modelo de predicción robusto. Durante el proceso, también hemos enfrentado desafíos, como la gestión de valores atípicos, la selección de variables adecuadas y la optimización de hiper parámetros. Sin embargo,

estos desafíos nos han brindado la oportunidad de mejorar nuestras habilidades y adoptar un enfoque más riguroso en nuestro análisis. En última instancia, hemos obtenido un modelo que puede proporcionar estimaciones razonablemente precisas de los precios de alquiler en función de las características específicas de las propiedades y del mercado local. Esto podría ser de gran utilidad tanto para los propietarios que desean establecer precios competitivos como para los inquilinos que buscan alojamiento asequible. Nuestro trabajo no solo ha mejorado nuestras habilidades técnicas, sino que también nos ha permitido comprender la importancia de la colaboración y la comunicación en proyectos de análisis de datos. Cada miembro del equipo ha aportado su experiencia y conocimientos únicos, lo que ha enriquecido nuestro enfoque y resultados. Destacamos también la importancia de la comprensión contextual de los datos con los que trabajaremos, hay cuestiones fundamentales que no pueden cuantificarse y por lo tanto debemos analizar desde una perspectiva más social para hacer del proceso de análisis de datos más eficiente. Luego de realizado este trabajo práctico integrador de Ciencia de Datos sobre Airbnb podemos concluir que ha sido una experiencia valiosa y enriquecedora. Hemos adquirido habilidades técnicas, desarrollado un modelo de predicción útil y fortalecido nuestra capacidad para trabajar en equipo. Este trabajo es un testimonio de nuestro compromiso con la excelencia en el campo de las Ciencias de Datos y demuestra cómo podemos aplicar nuestras habilidades para abordar desafíos del mundo real con éxito.