Reporte Actividad 6

Se utilizaron dos archivos CSV, listings_amsterdam.csv y listings_porto.csv, que se cargaron en dos DataFrames, LISAM y LISPOR, utilizando la biblioteca Pandas de Python para este análisis de datos. El objetivo inicial fue analizar la estructura de los datos y eliminar cualquier anomalía, como valores nulos o atípicos, que pudiera afectar el análisis posterior.

En el DataFrame LISAM, primero se utilizó el método info() de Pandas para obtener una idea de la estructura de los datos. Este método proporcionó información sobre el número de entradas en cada columna, el tipo de datos y la presencia de valores nulos. A continuación, se utilizó LISAM.isnull().sum(). para conteo de valores nulos en cada columna. Esto permitió identificar columnas que carecían de datos y evaluar la gravedad del problema.

Se creó un nuevo DataFrame llamado LISAMclean para manejar los valores nulos utilizando métodos de interpolación hacia adelante y hacia atrás (ffill y bfill). Este procedimiento garantiza que las brechas en los datos se rellenen utilizando los valores más cercanos a los que se tenían. Adicionalmente, la columna neighbourhood_group fue eliminada porque no contenía datos válidos, por lo que no era necesaria para el análisis.

Después de la limpieza inicial, se revisaron los valores nulos en LISAMclean y se confirmó que las acciones de limpieza habían funcionado. Los valores inusuales en las columnas cuantitativas del DataFrame fueron identificados y gestionados por el análisis. Se crearon dos conjuntos de datos diferentes: cuanti, que incluía columnas de carácter cuantitativo, y cuali, que incluía columnas de carácter cualitativo. Se creó un diagrama de caja, también conocido como diagrama de caja, para mostrar los valores atípicos en las columnas cuantitativas. Esto permitió una comprensión visual de la dispersión de los datos.

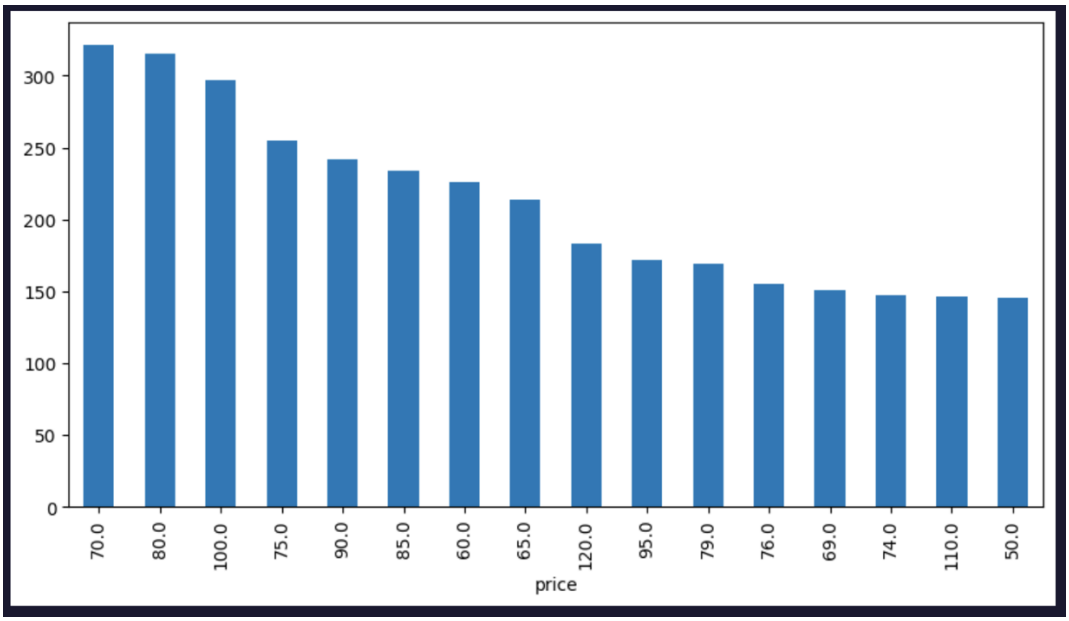
Se utilizaron la media y tres desviaciones estándar de cada columna cuantitativa para calcular los límites superior e inferior permitidos. Los valores fuera de estos límites fueron considerados atípicos y se manejaron en un nuevo DataFrame llamado LAC2. Esto se hizo filtrando los datos para que solo incluyeran valores que estaban dentro de los límites establecidos. Posteriormente, se creó un DataFrame final llamado datac al reemplazar los valores nulos en LAC2 por la media redondeada.

Para crear un DataFrame consolidado, las columnas cualitativas y los datos cuantitativos limpios se combinaron. Se realizó una verificación final de valores nulos en este DataFrame combinado para garantizar la calidad y completitud de los datos para futuros análisis.

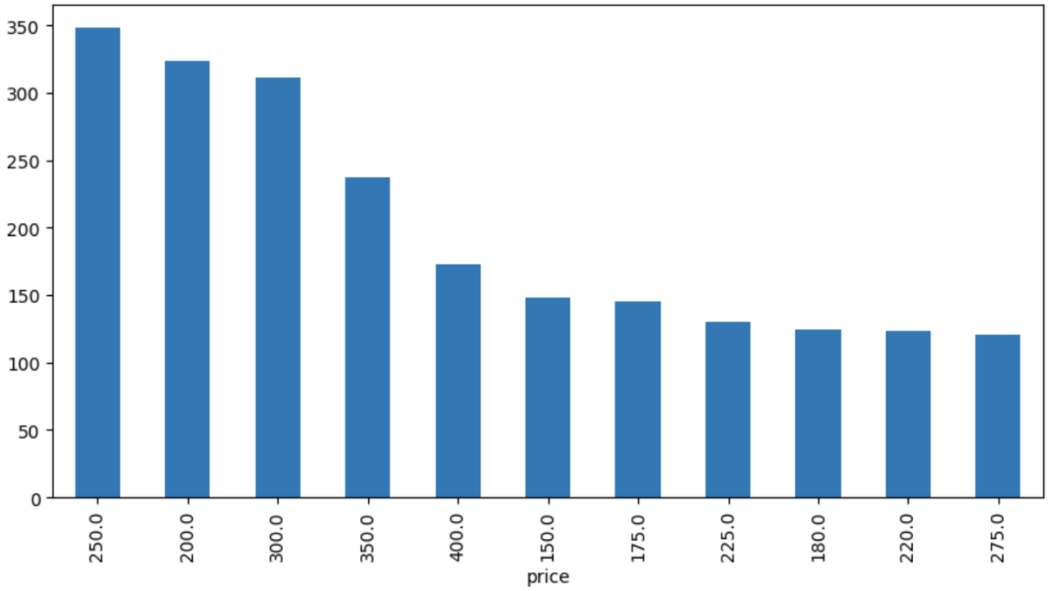
Finalmente se procedió a la creación de gráficas para el análisis de los datos frecuenciales del data set y poder obtener una idea de los datos con los que se contaba:

1) Price

a. Porto

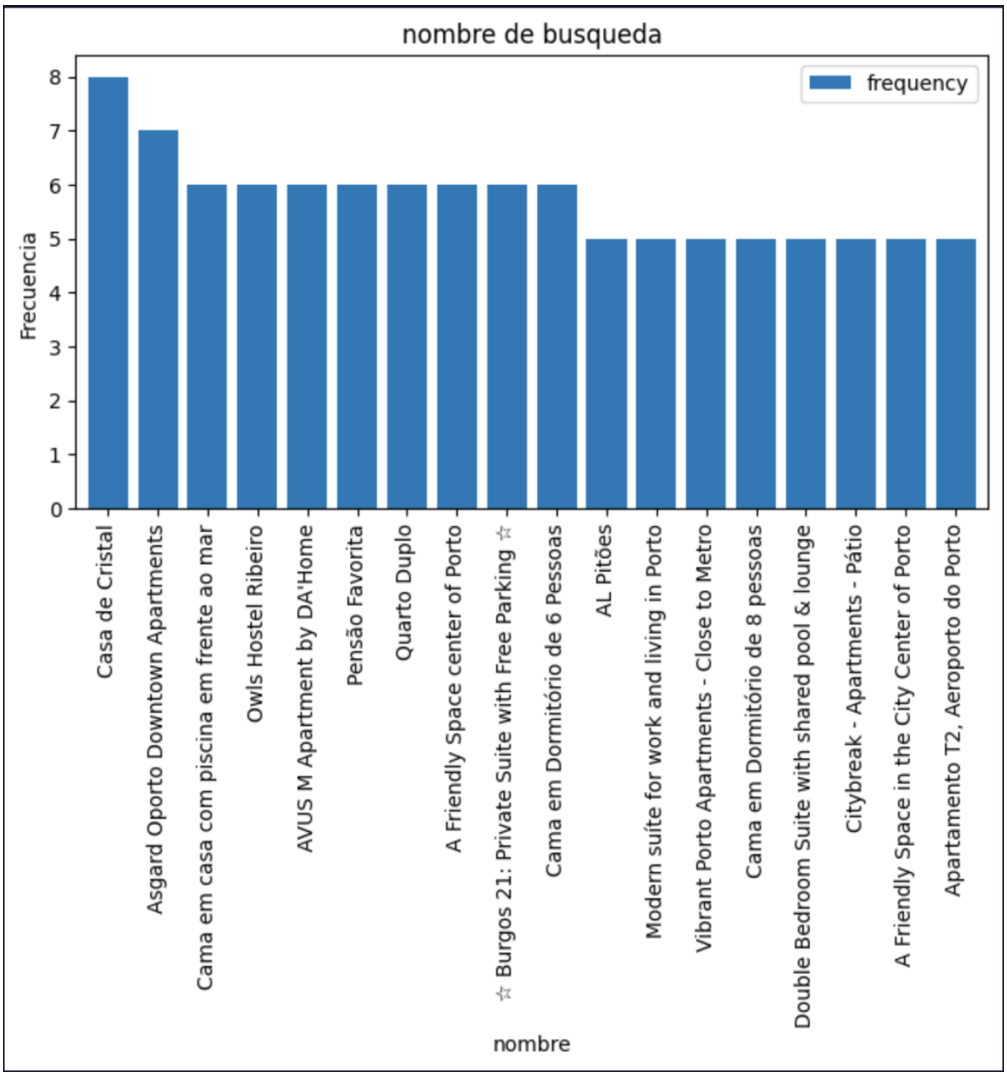


b. Amsterdam

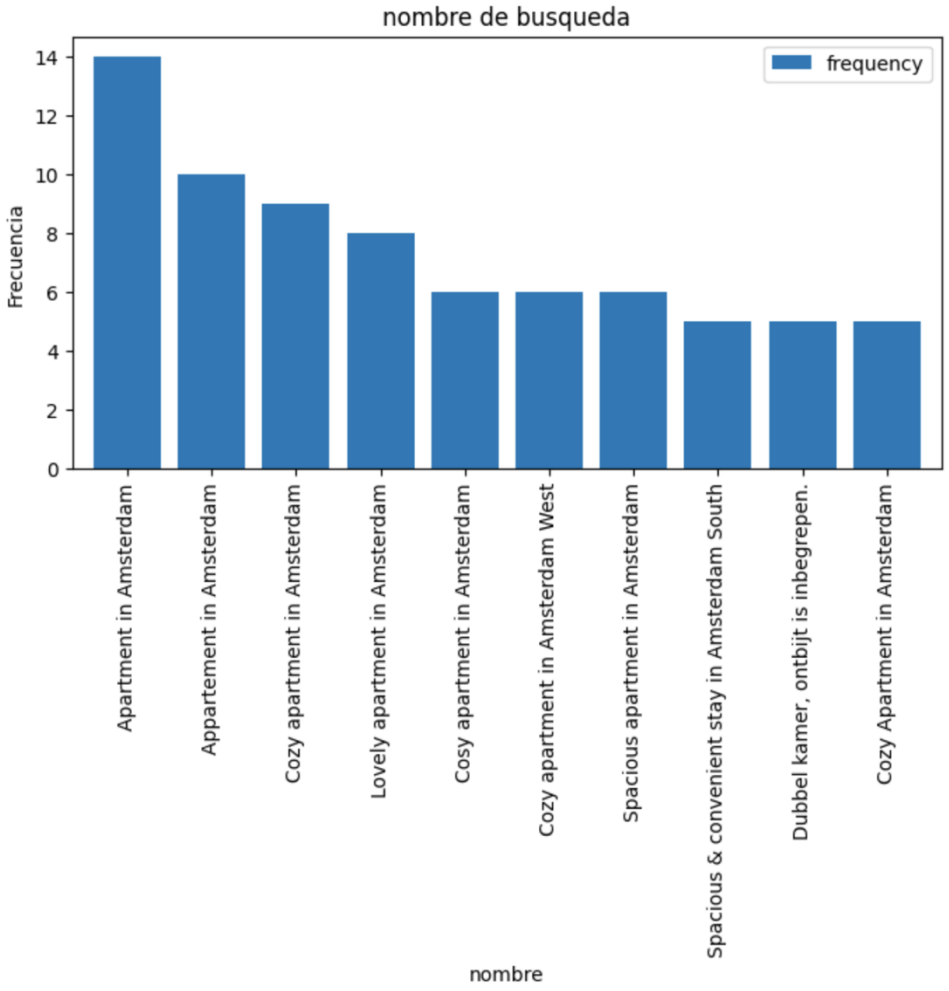


2) Name

a. Porto

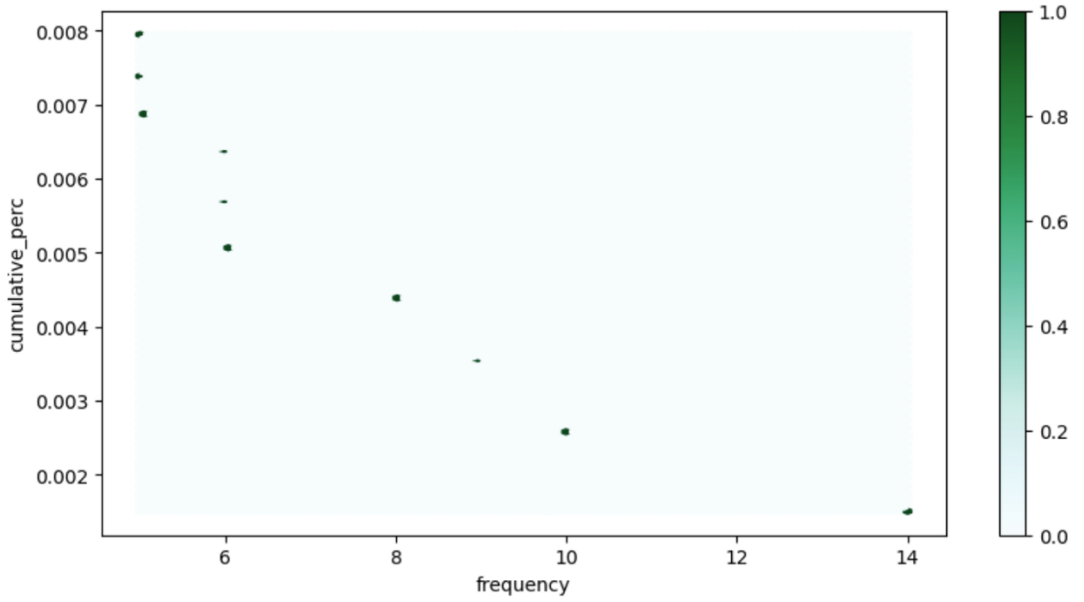

nombre de busqueda
frequency
Frecuencia
0
1
2
3
4
5
6
7
8
Casa de Cristal
Asgard Oporto Downtown Apartments
Cama em casa com piscina em frente ao mar
Owls Hostel Ribeiro
AVUS M Apartment by DA'Home
Pensão Favorita
Quarto Duplo
A Friendly Space center of Porto
☆ Burgos 21: Private Suite with Free Parking ☆
Cama em Dormitório de 6 Pessoas
AL Pitões
Modern suíte for work and living in Porto
Vibrant Porto Apartments - Close to Metro
Cama em Dormitório de 8 pessoas
Double Bedroom Suite with shared pool & lounge
Citybreak - Apartments - Pátio
A Friendly Space in the City Center of Porto
Apartamento T2, Aeroporto do Porto
nombre

b. Amsterdam

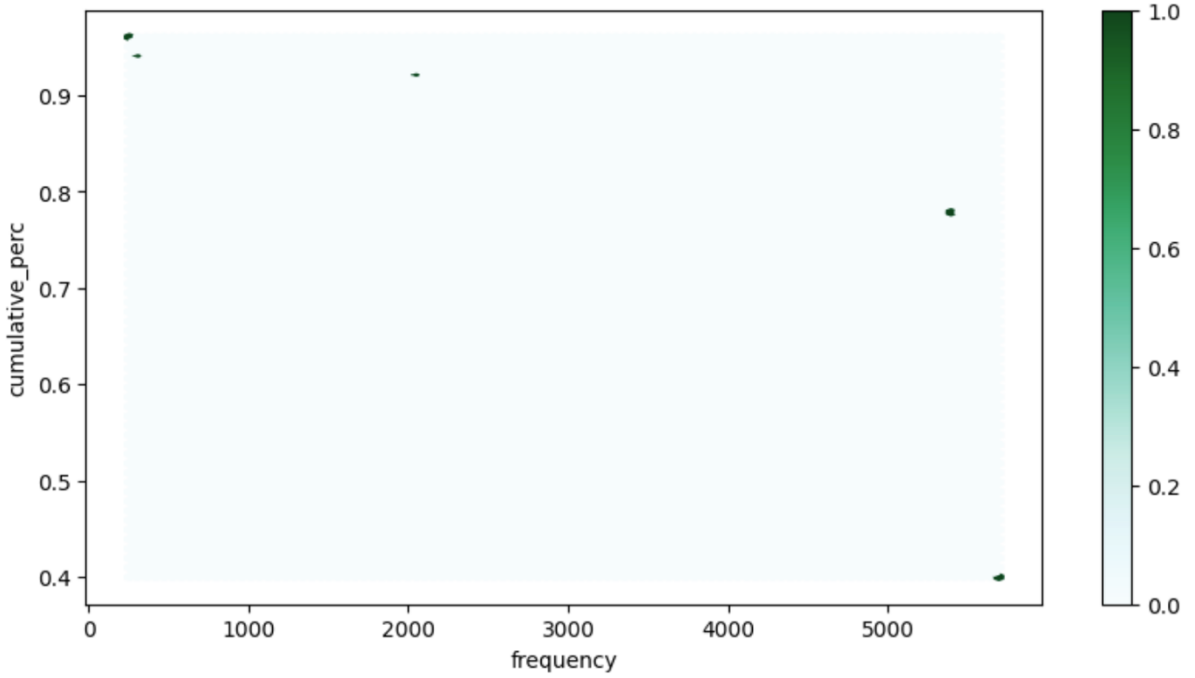


3) Minimum Nights

a. Amsterdam

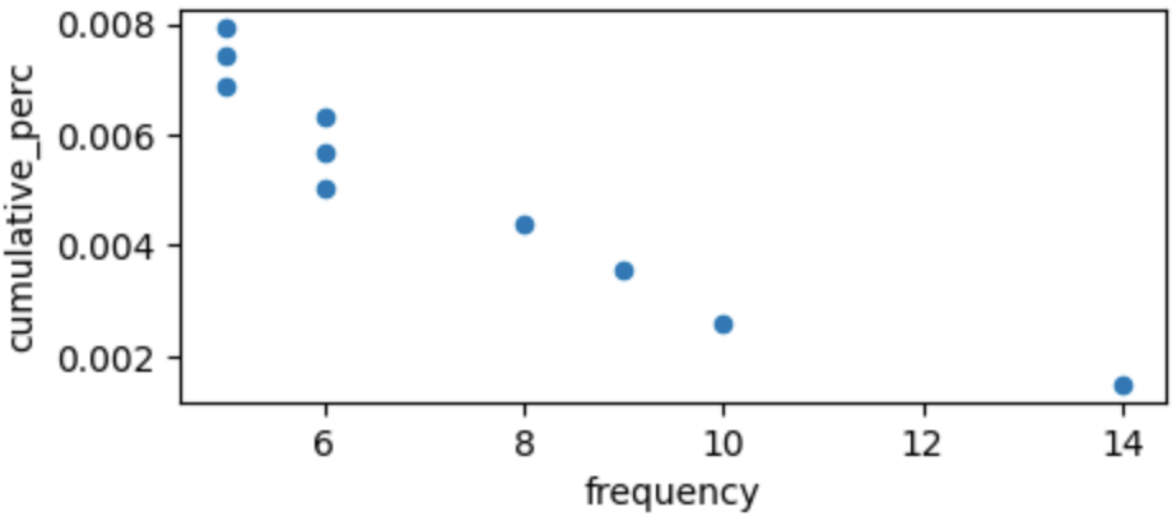


b. Porto

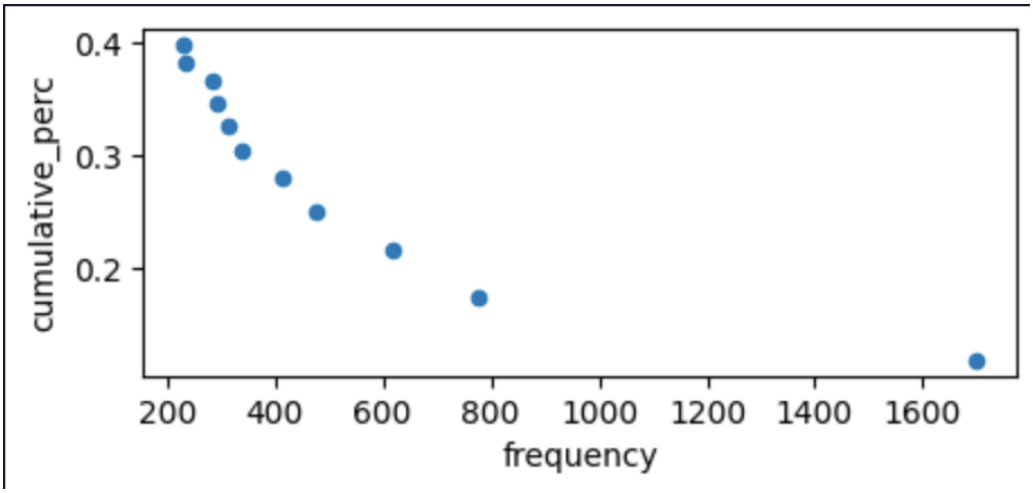


4) Number Of Reviews

a. Amsterdam

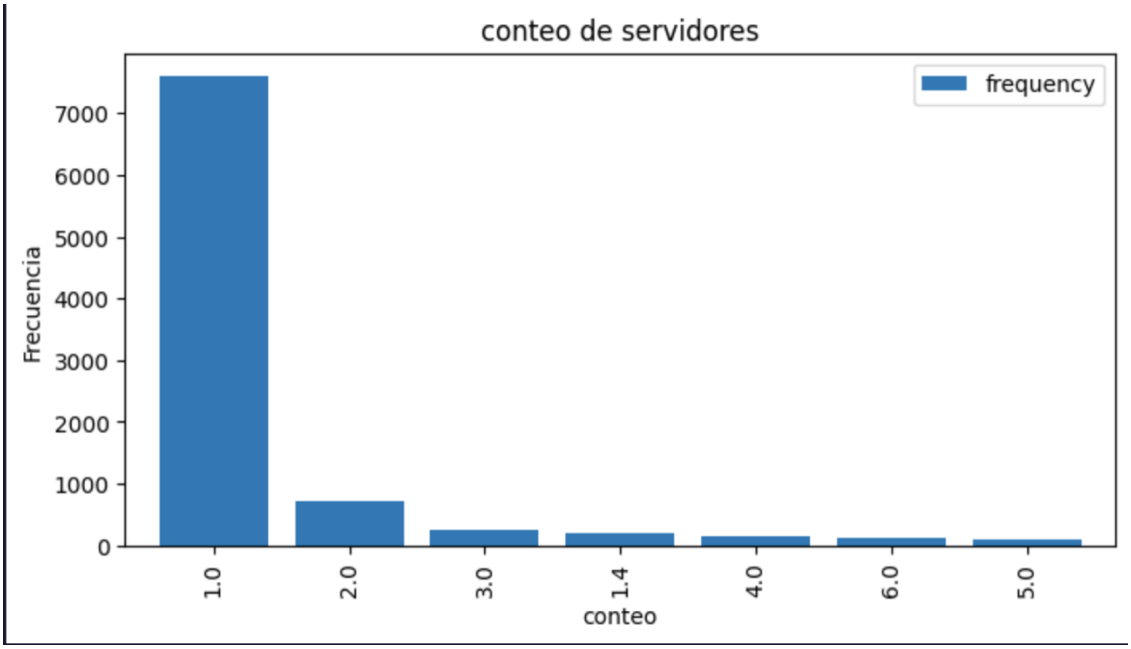


b. Porto

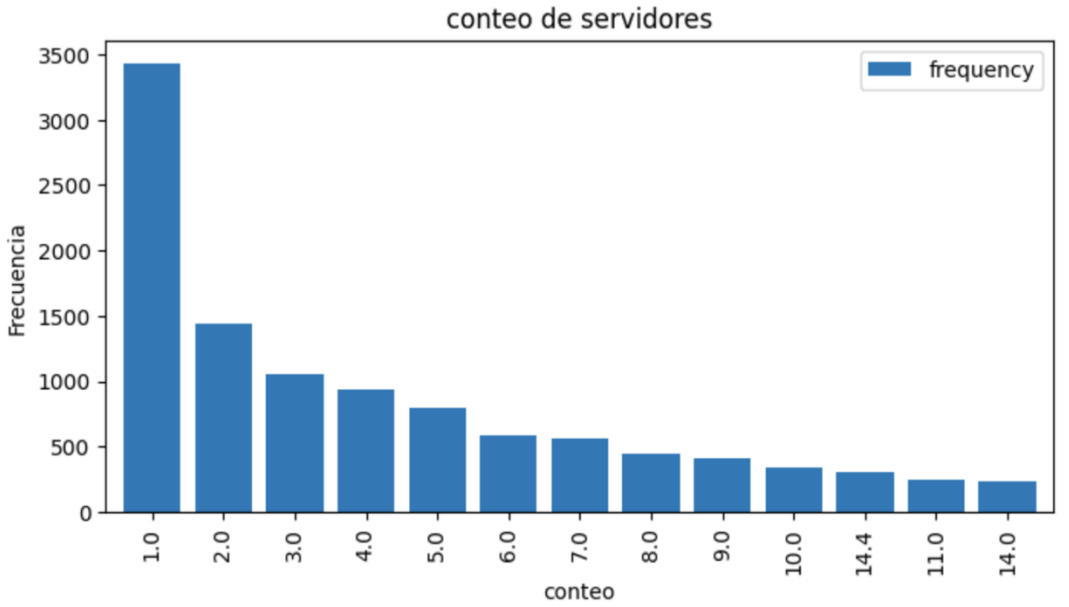


5) Calculated Hos Listed Count

a. Amsterdam

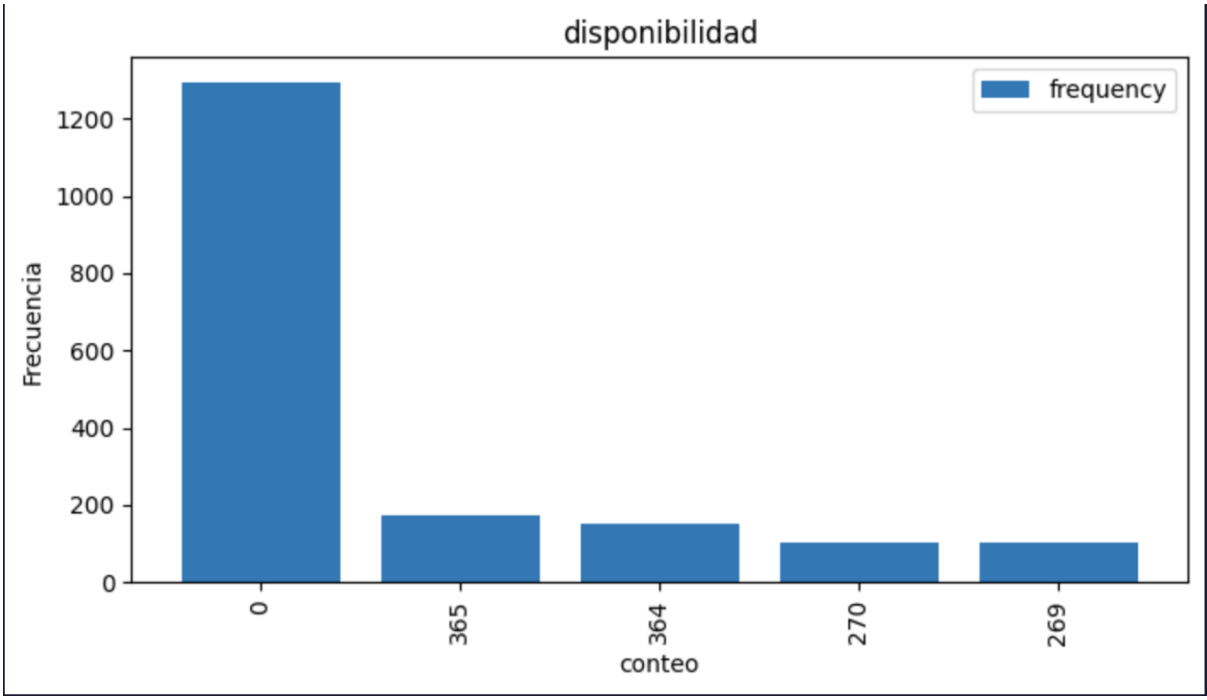


b. Porto

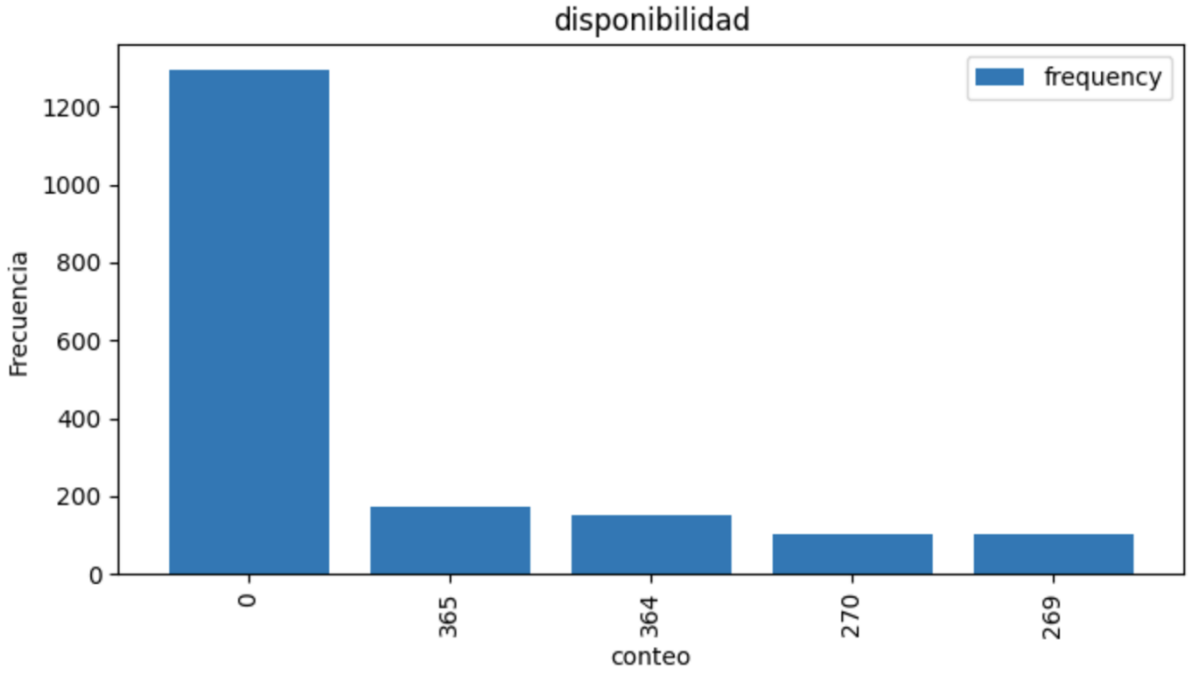


6) Number Of Reviews

a. Amsterdam

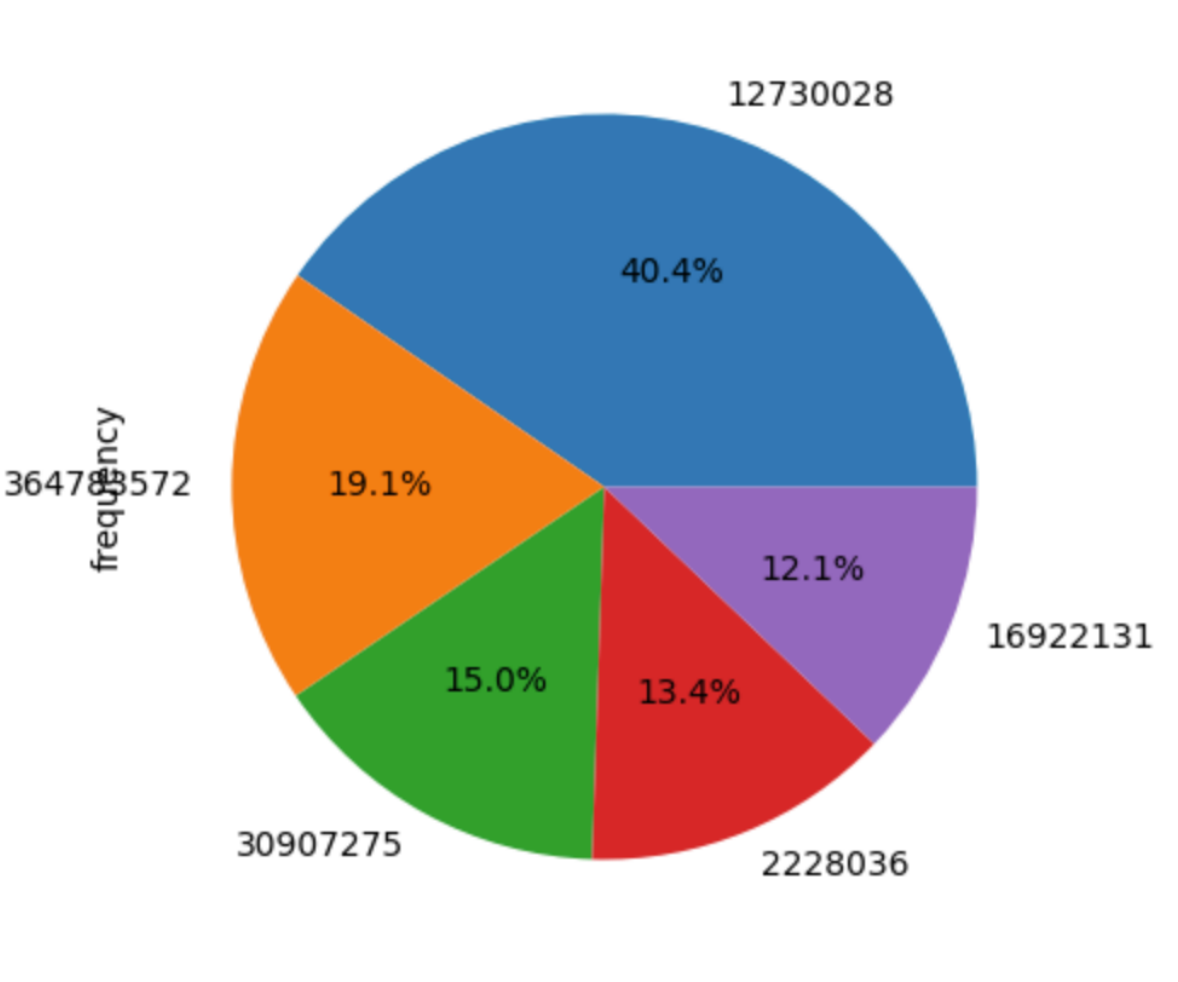


b. Porto

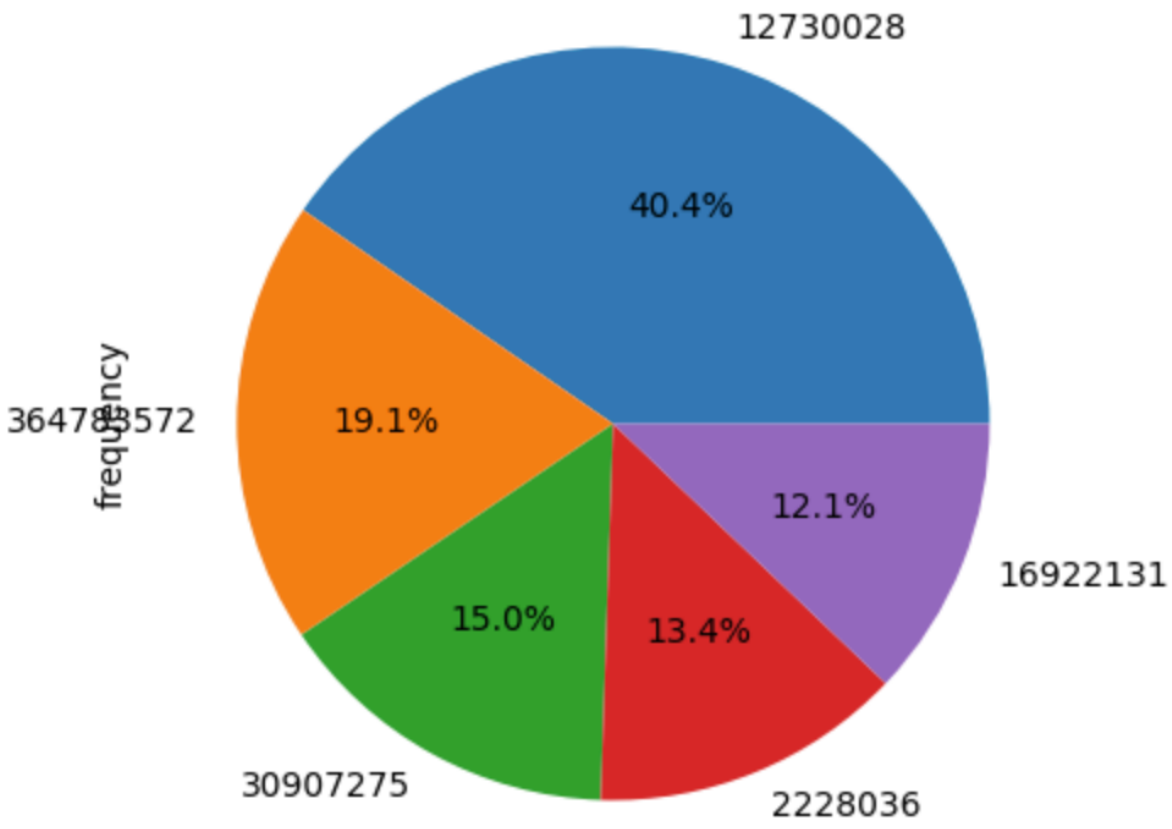


7) Host ID

a. Porto

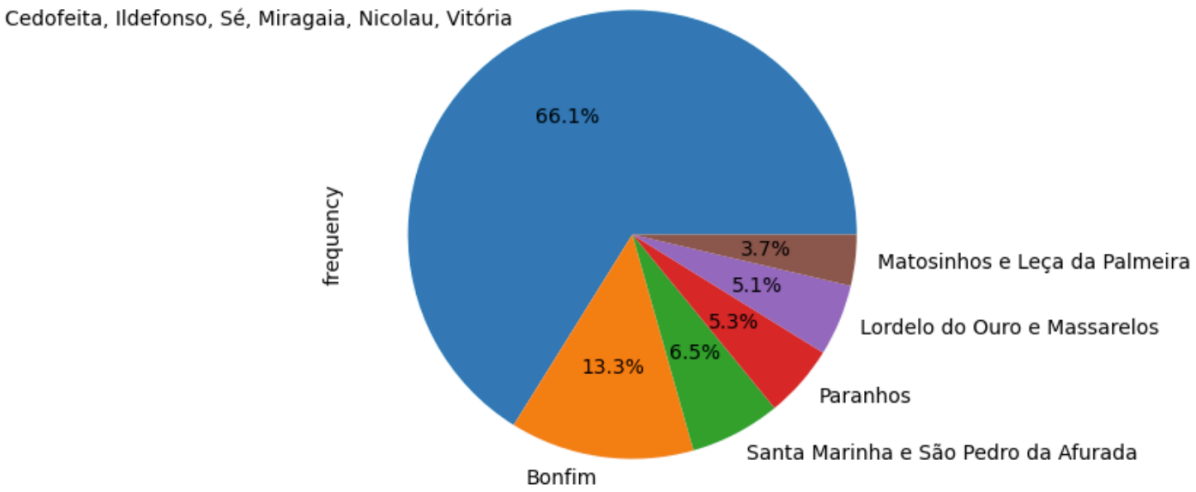


b. Amsterdam

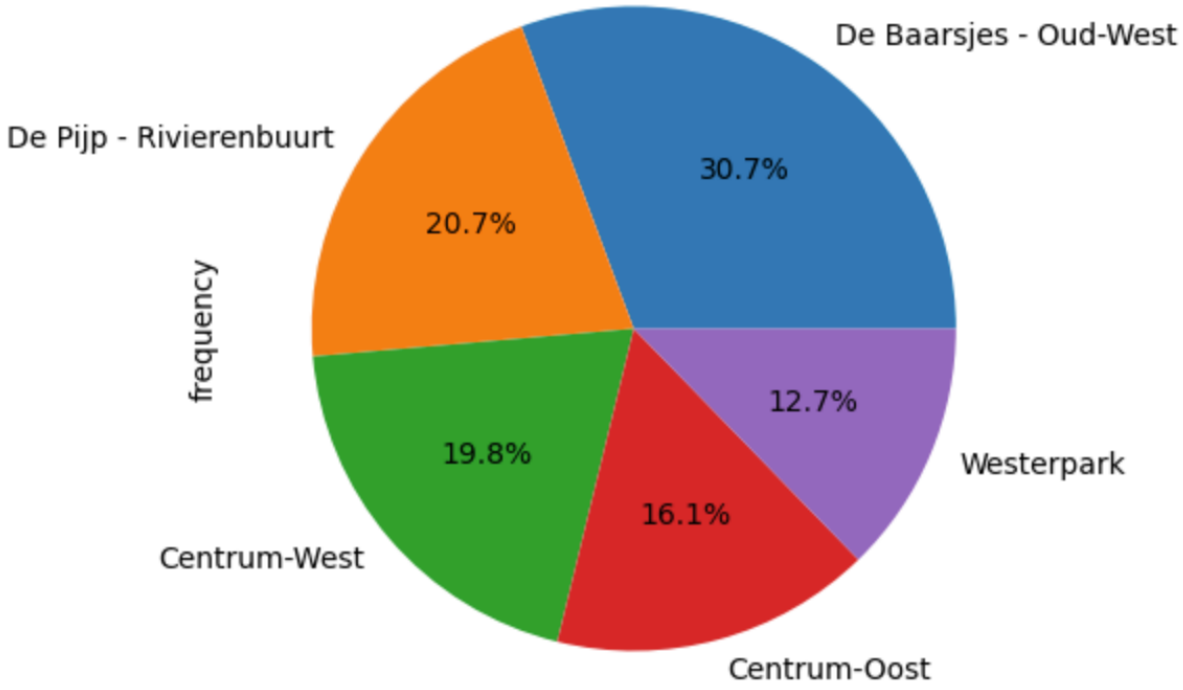


8) Neighbourhood

a. Porto



b. Amsterdam


De Baarsjes - Oud-West
30.7%
De Pijp - Rivierenbuurt
20.7%
frequency
12.7%
Westerpark
19.8%
16.1%
Centrum-West
Centrum-Oost