Juan David Ayala Nariño

Ciencia de Datos Aplicada

Taller 1

Análisis de oportunidades en Airbnb: Berlín, Alemania.

Se realizó un Análisis Exploratorio de Datos para la ciudad de Berlín, Alemania, para descubrir posibles oportunidades que se puedan presentar en el mercado de esta ciudad.

Para esto se utilizó el dataset correspondiente a esta ciudad, el cuál se encuentra en la página insideairbnb y adicionalmente se utilizó el diccionario de datos suministrado.

1. Entendimiento inicial de los datos.

El dataset seleccionado posee las siguientes características:

- Cantidad de registros: 12 472
- Cantidad de Columnas: 75
- Tipos de Datos: Números enteros (int64), números con decimales (float64) y "object" que puede hacer referencia a cadenas de texto u otro tipo de datos mixtos.

Las distintas variables fueron revisadas y se tomó la conclusión de excluir totalmente algunas de ellas como el id de Airbnb que no aportará información adicional, urls que indican la dirección web del anuncio, url de fotos, el campo "bathrooms" y "calendar updated" que no contienen información, el campo "neighbourhood" que en la mayoría de los casos incluye información del país y la ciudad entre otros.

El campo "**description**" fue eliminado, pero sería interesante realizar un análisis particular sobre este para identificar si la descripción puede afectar el valor y el ingreso que percibe una propiedad.

Una vez se realizó esta operación el dataset queda reducido a 51 variables. Es posible que más adelante algunas de estas variables no sean necesarias o que tengan una gran correlación con otras, pero esto se tratará más adelante.

Inicialmente se consideraron los siguientes atributos como los más importantes del dataset:

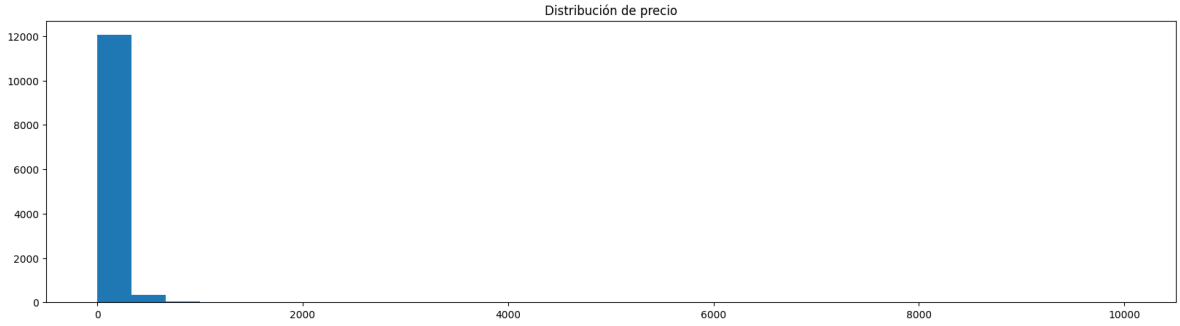
a) **price**: El precio permitirá identificar variaciones que se presenten y con esta información maximizar la posibilidad de alquiler. Este campo es necesario hacerle ajustes ya que se reconoció como una cadena de texto.
b) **room_type**: se asume que el tipo de cuarto puede dar alguna información extra sobre la ocupación y el precio.
c) **Accomodates**: puede ser un buen indicador del tamaño del inmueble a arrendar.
d) **Mínimum_nights**: Se espera que los inmuebles con una menor cantidad de días mínimos tengan una mayor ocupación a lo largo del mes maximizando los ingresos.

e) **Availability (30-60-90-360):** Ayudará a identificar los inmuebles que tienen una mayor o menor disponibilidad, aumentando así el ingreso sobre la propiedad. Es importante tener en cuenta que puede estar bloqueado por el dueño sin necesidad de estar arrendado.

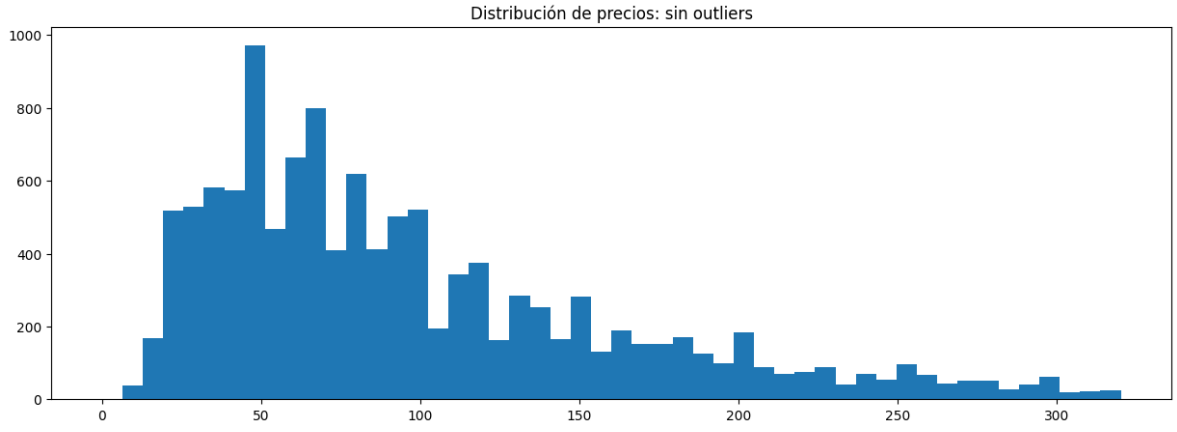
Adicionalmente, el campo "**neighbourhood_cleansed**" puede brindar información adicional si se llega a encontrar un patrón por ubicación y el campo "**amenities**" puede brindar información sobre los servicios que buscan los usuarios en un inmueble.

Los campos mencionados anteriormente fueron sometidos a ajustes para validar que puedan ser utilizados de la manera requerida.

a) Price: Fue necesario eliminar el signo "$" que se encontraba presente en la columna para poder utilizarlo como un número.
Una vez se realizó esto se realizó un histograma para observar la distribución del precio, encontrando que existen valores muy altos que no permiten apreciar la distribución real con valores de hasta 1000 euros la noche.

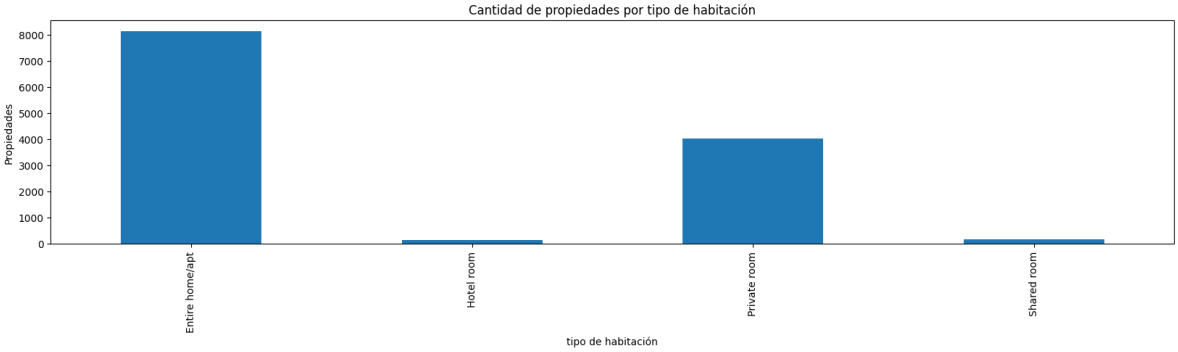


En este caso, con ayuda de los rangos interquartil se calcula sin los outliers:

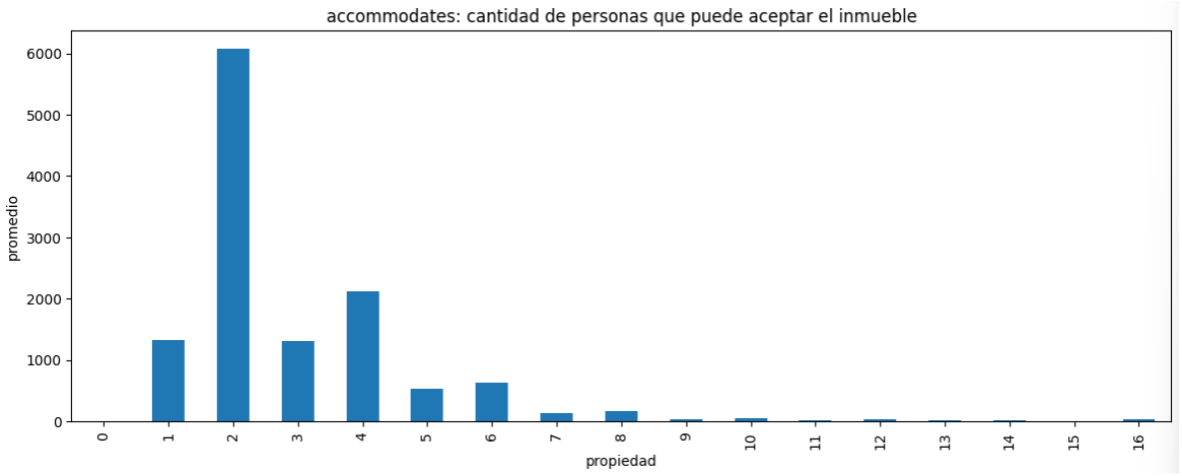


Esta distribución es sesgada a la izquierda, donde el 50% de los valores son menores a 81 euros/noche.

b) Room_type: Se verifica la cantidad de habitaciones por tipo disponible. El 65% corresponde a un apartamento/casa completa, el 32% a una habitación privada y el porcentaje restante a habitación de hotel y habitaciones compartidas.



c) Accomodates: Se realiza el cálculo para entender la cantidad de gente que puede hospedar cada propiedad. Para la ciudad, el 48.6% de las propiedades puede hospedar hasta 2 personas.



d) Minimum_nights: Curiosamente, el 50% de los inmuebles tiene una tasa de noches mínima mayor a 7 días, es decir, el 50% de los inmuebles renta para periodos mayores a una semana.

e) Availability: Este campo indica que existen 4517 inmuebles que no tienen disponibilidad en los próximos 365 días. Al no tener un experto de negocio para preguntar si este comportamiento es normal y válido, se realizará el análisis manteniendo estos usuarios, sin embargo, es importante validar si es normal este comportamiento y si se deben tener en cuenta estos usuarios.

El campo “amenities” requirió un procesamiento adicional: para identificar si una propiedad tiene o no un servicio se realizó un one_hot de los distintos amenities limpiando los valores que estuvieran repetidos por cosas como el tamaño del televisor o la marca de un shampoo.

2. Estrategia:

Se buscará identificar los factores comunes que tengan los mejores inmuebles, ya sea por sus características o por sus ubicaciones geográficas.

Inicialmente se revisarán las demás unidades de que poseen los distintos atributos y se realizarán ajustes de acuerdo con lo encontrado como se realizó con el precio (cadena a número, eliminando caracteres innecesarios). También se validarán posibles valores que se encuentren con errores de acuerdo con el diccionario de datos.

Es importante agregar que el diccionario de datos es una herramienta fundamental porque permite identificar las unidades y los valores que deben de tener los distintos campos. Un pequeño preprocesamiento ya fue realizado para reducir la cantidad de variables a solamente aquellas que puedan aportar información valiosa al análisis.

También se buscarán correlaciones entre las variables cuantitativas y gráficos para análisis univariados/multivariados que permitan entender mejor el comportamiento de las variables presentes. Igualmente se realizará este análisis progresando desde un punto de vista general a uno particular donde se busca descubrir patrones ocultos que permitan obtener el mejor desempeño de una propiedad.

por último. se podrán generar campos calculados que brinden una mayor información sobre algunas variables del negocio.

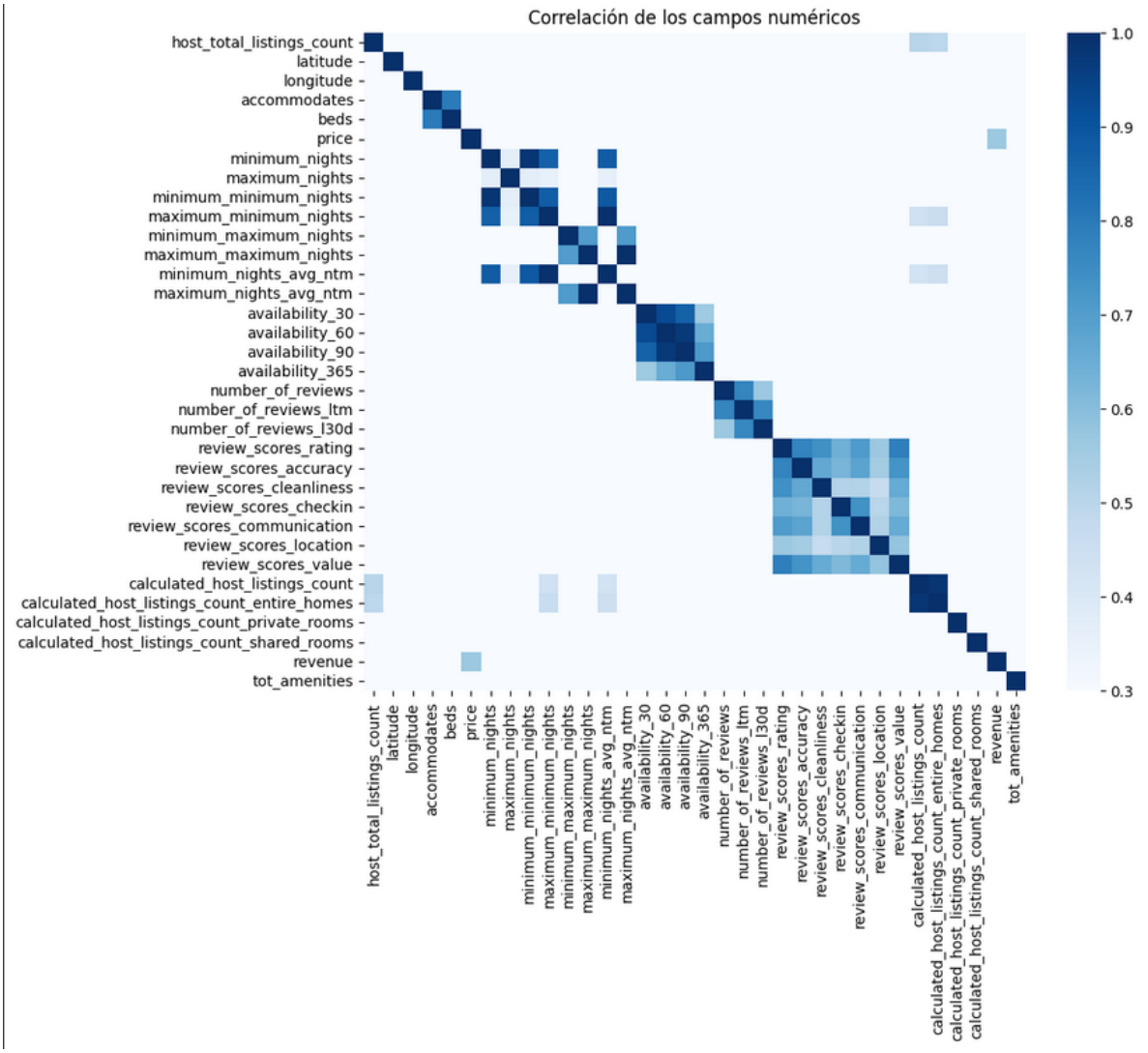
3. Desarrollo:

Correlación de variables:

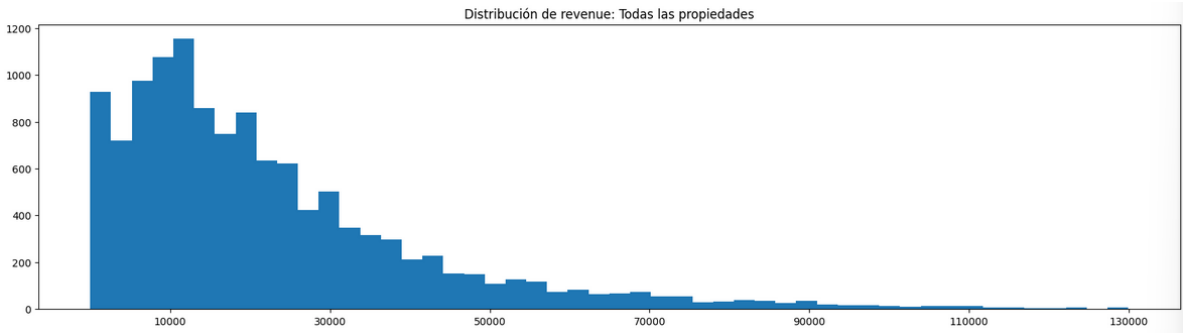
Se tomaron las variables iniciales del dataset y comprobó su correlación mediante un mapa de calor.
Existen correlaciones altas entre la cantidad mínima de noches y su mínimo al igual que la cantidad máxima de noches y su máximo, adicionalmente se encuentra una correlación entra las disponibilidades de días. Teniendo en cuenta esto, se pueden ignorar algunas variables. Otra correlación encontrada es entre la cantidad de camas con la que cuenta l a propiedad y la cantidad de gente que puede albergar. Esta, sin embargo, es esperada.
El número de reviews también tiene una correlación alta si se compara con los del último mes y con los últimos 30 días.
En general este gráfico no ofrece información relevante que pueda ser utilizada.

Se crea un nuevo campo llamado “revenue”. Este calcula el ingreso de la propiedad al multiplicar la cantidad de días no disponibles en el siguiente año por el precio de una noche en la propiedad.

También se crea un campo con la cantidad de servicios totales ofrecidos pare ver si se correlaciona con el precio, pero no se encuentra una correlación entre estos campos.



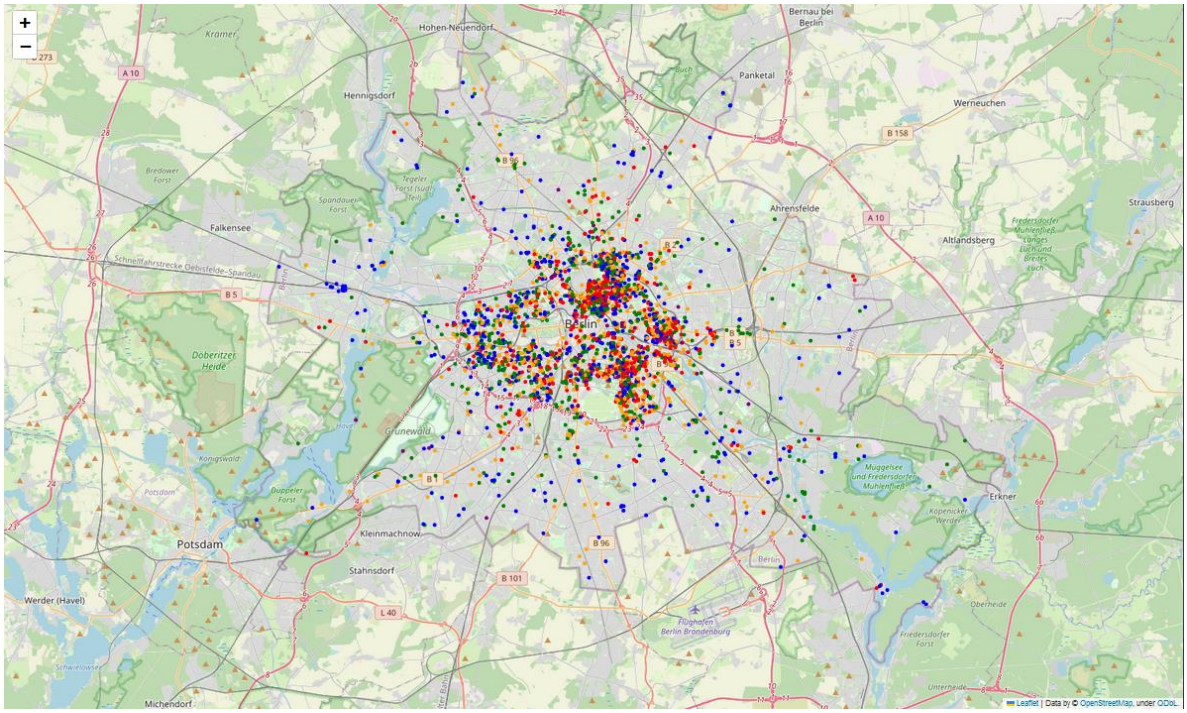
Este tiene una distribución similar a la del precio por noche de las propiedades.



También se realiza un mapa de calor con los servicios principales, sin embargo, a simple vista, no indica que exista una correlación importante entre estos.

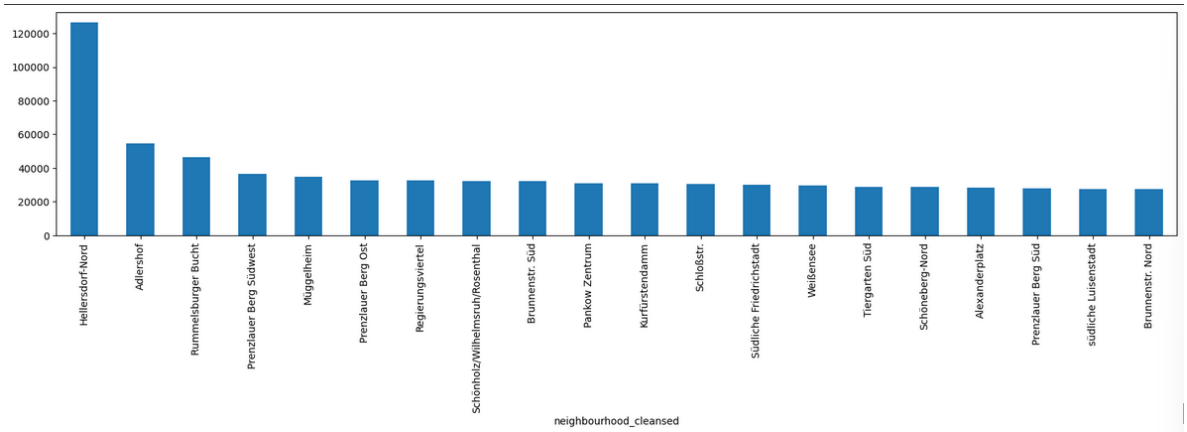
Luego se realiza un gráfico de las propiedades por su latitud y longitud con su revenue para identificar si hay sectores de la ciudad con precios más altos que otros.

Se toma una muestra de 3000 propiedades y se codifican. El 25% con el revenue más bajo tienen el color azul, del 25% al 50% verde, del 50% al 75% naranja, del 75% al 95% rojo y el 5% más alto púrpura.



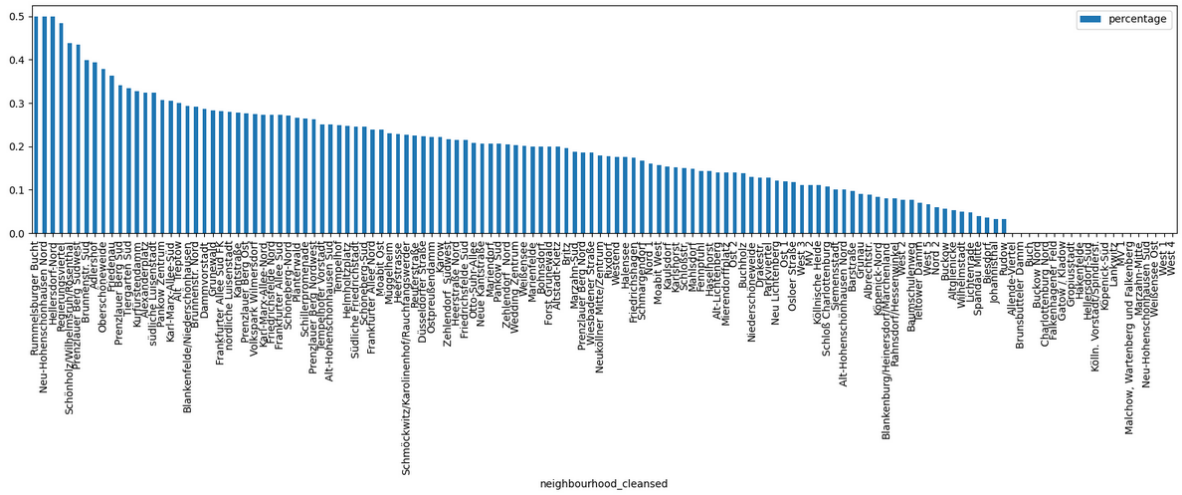
Si bien existen propiedades de todos los precios en toda la ciudad, existen agrupaciones de color que dan a entender que existen áreas que cuentan con un revenue mayor que otras.

Se realiza ahora un análisis por el barrio donde se encuentran algunos barrios con ingresos superiores a los demás:

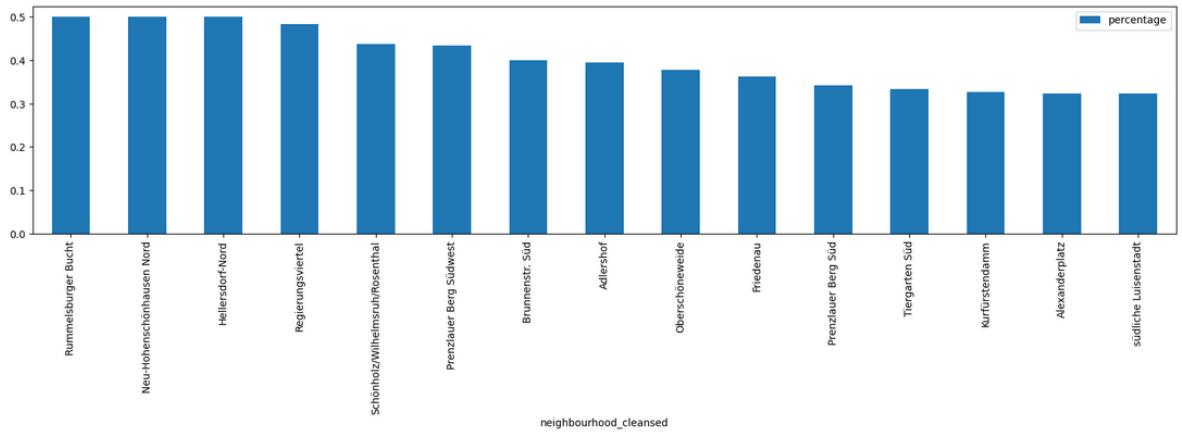


Sin embargo, esto se puede deber a una sola propiedad con un valor muy alto que esté “arrastrando” a las demás. Para solucionar esto se realiza un cálculo de los porcentajes de propiedades con revenue mayor al 75% que se encuentran en cada barrio.

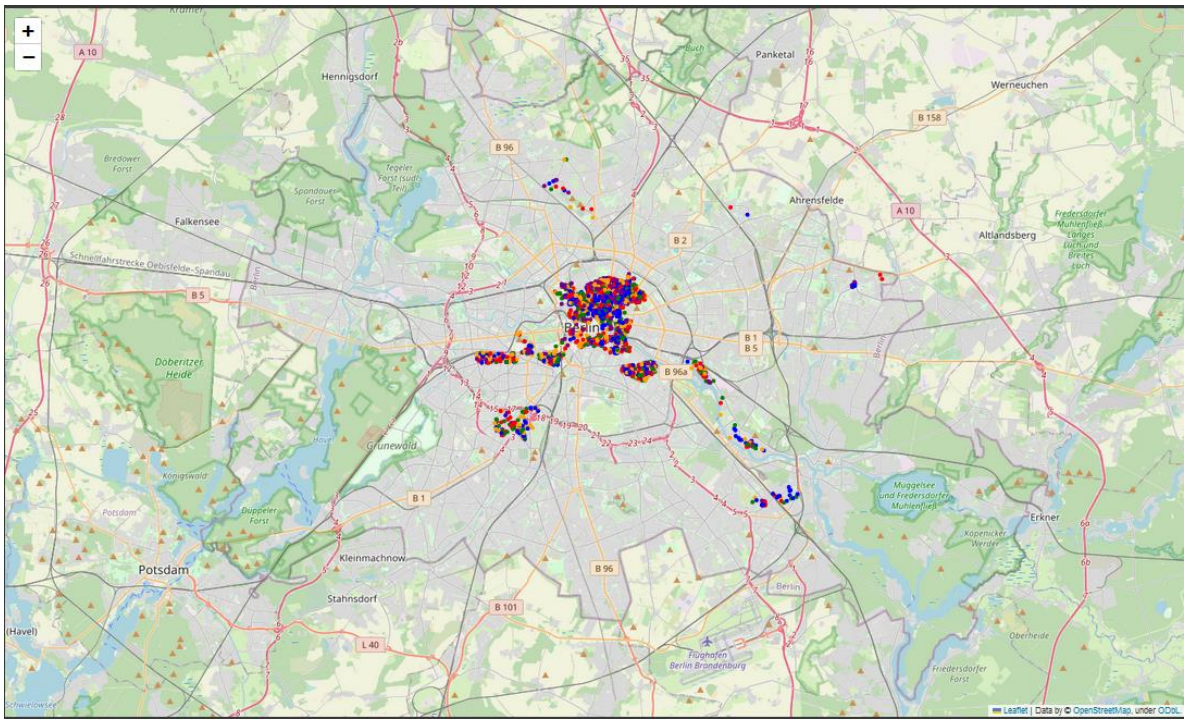
El resultado indica estos barrios con su respectivo porcentaje de propiedades por encima del 75%.



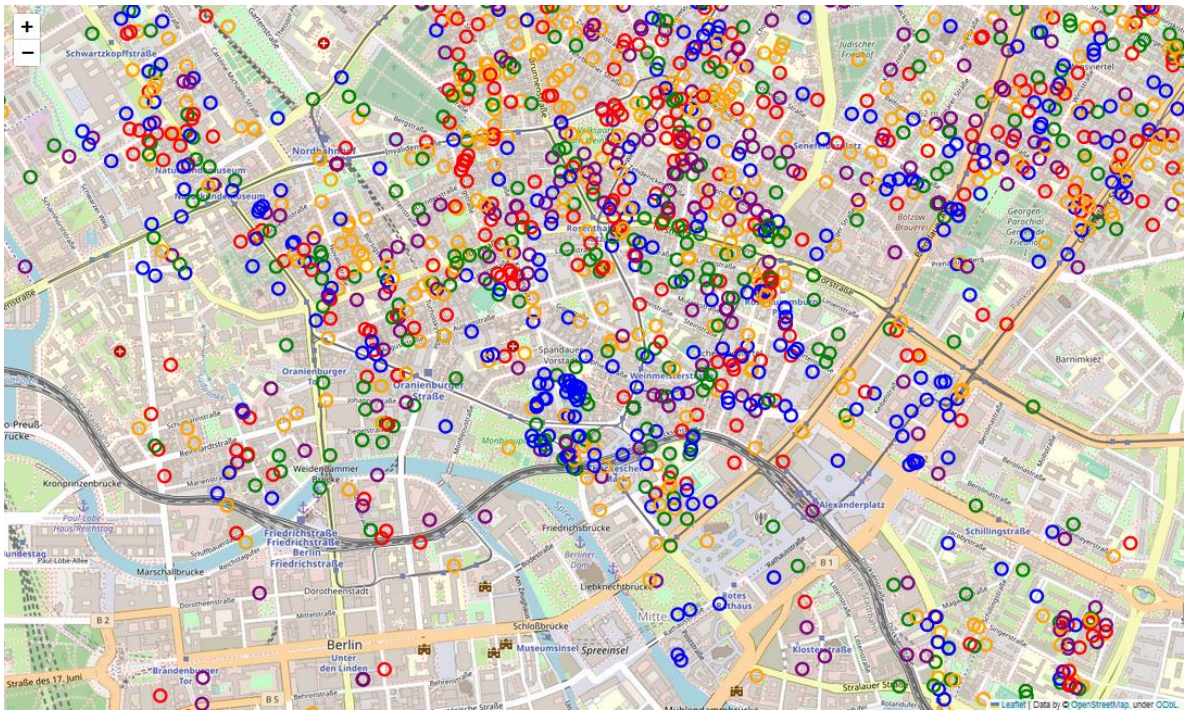
Los 15 barrios principales se muestran a continuación:



Y con ayuda del mapa se puede incluso determinar las manzanas donde el revenue no es tan bueno dentro de esos barrios, por ejemplo,

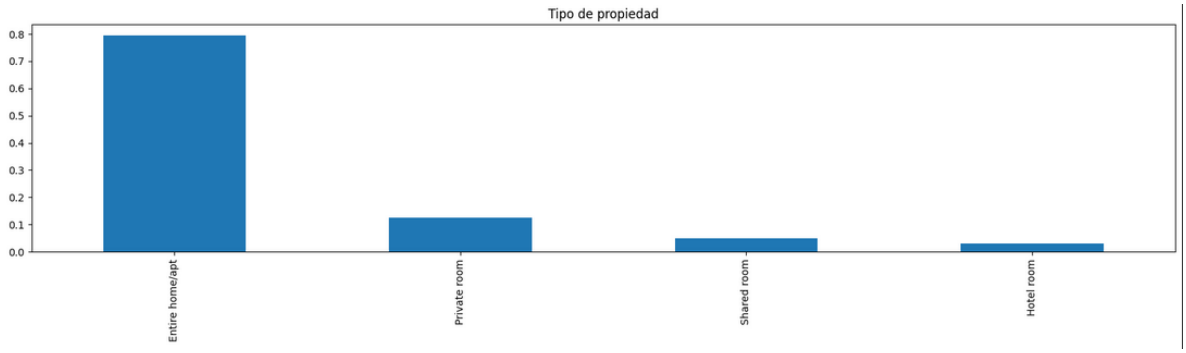


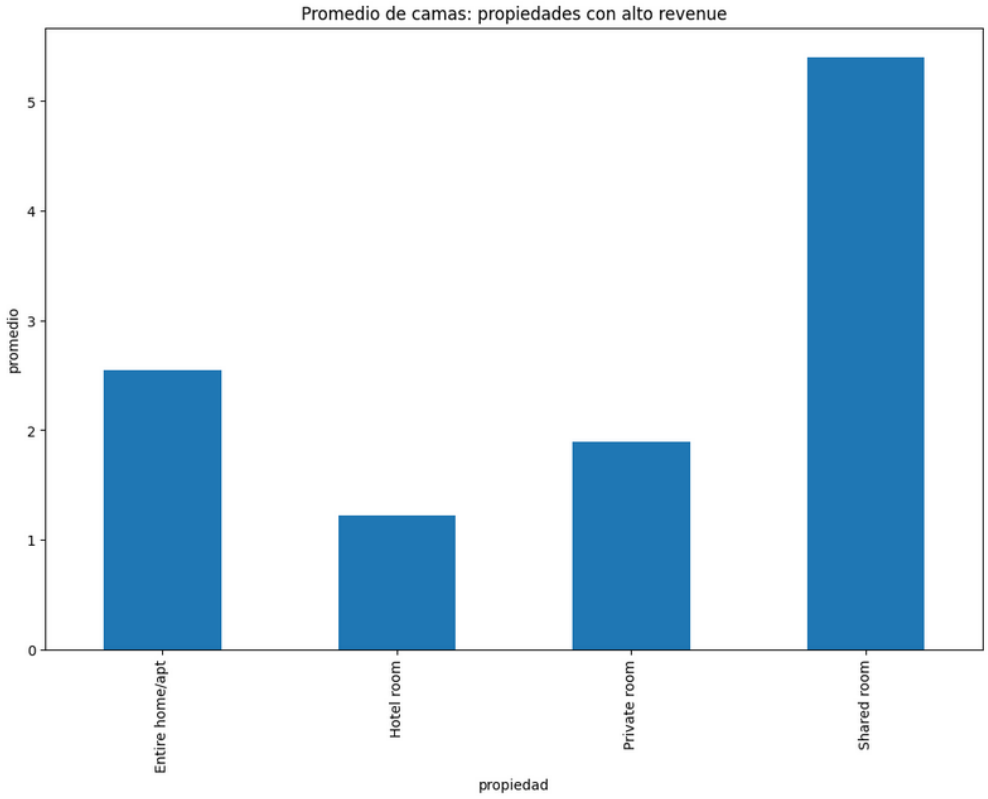
Vista general de los 15 barrios con mejores ingresos.



Zoom a uno de los barrios donde se pueden ignorar manzanas determinadas (predominantemente color azul).

Por último, se validan las características de los inmuebles con un revenue alto en estos sectores:





80% son casa o apartamentos completos, con un promedio de 2.5 camas

4. Resultados:
    - Se recomienda adquirir un inmueble en los siguientes barrios:
        - 'Rummelsburger Bucht',
        - 'Neu-Hohenschönhausen Nord',
        - 'Hellersdorf-Nord',
        - 'Regierungsviertel',
        - 'Schönholz/Wilhelmsruh/Rosenthal',
        - 'Prenzlauer Berg Südwest',
        - 'Brunnenstr. Süd',
        - 'Adlershof',
        - 'Oberschöneweide',
        - 'Friedenau',
        - 'Prenzlauer Berg Süd',

    - 'Tiergarten Süd',
    - 'Kurfürstendamm',
    - 'Alexanderplatz',
    - 'südliche Luisenstadt'

- El inmueble debe ser arrendado preferiblemente completo.
- El inmueble debe tener entre 2 y 3 camas
- Los 20 servicios más comunes que ofrecen los inmuebles con un alto revenue en este sector son:
    - "bed linens'",
    - 'shampoo'",
    - ' "washer'",
    - ' "room-darkening shades'",
    - ' "wine glasses'",
    - ' "heating'",
    - ' "essentials'",
    - ' "baking sheet'",
    - ' "oven'",
    - ' "paid parking off premises'",
    - ' "iron'",
    - ' "blender'",
    - ' "high chair'",
    - ' "elevator'",
    - ' "hair dryer'",
    - 'refrigerator'",
    - ' "kitchen'",
    - ' "board games'",
    - ' "wifi'",
    - ' "pack "

Es importante tener en cuenta las suposiciones que se realizaron en este análisis y agregar información relevante de contar con esta como lo es el valor del predio y si el retorno es el esperado.

También es importante validar cuadras específicas dentro del sector que pueden no ser tan atractivas.