1. Selección del dataset de trabajo: Seleccionar una ciudad de su interés y descargar el dataset detallado de listings correspondiente. Las ciudades disponibles con sus respectivos datasets se encuentran aquí.

Ciudad Escogida: Barcelona.

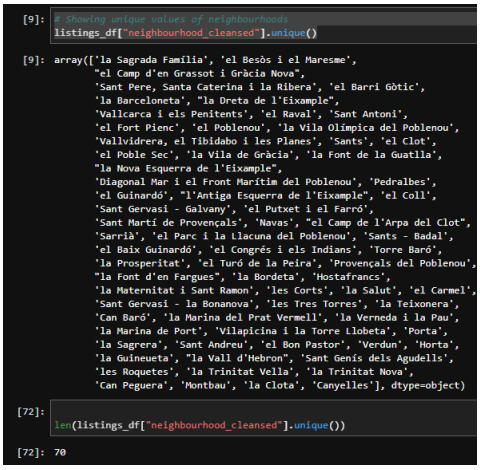
2. [20%] Entendimiento inicial de datos: Generar un breve reporte de entendimiento inicial de datos en donde destaque las dimensiones del dataset, los tipos de datos que contiene y el top 5 de los atributos que considera más importantes para el análisis. Por cada atributo incluya algunos elementos básicos de su comportamiento o distribución (análisis univariado).

Es un dataset de un tamaño de 17230 filas que representan propiedades dentro de la ciudad de Barcelona, con un total de 75 columnas.

Ahora, mi top 6 variables más importantes en este análisis serían las siguientes:

- neighbourhood_cleansed
- room_type
- price_float
- number_of_reviews_ltm
- Availibility in future (availability_30, availability_60, availability_90, availability_365)
- minimum_nights
- last_review

Para la variable neighbourhood_cleansed que es una variable categórica, se encontró lo siguiente:

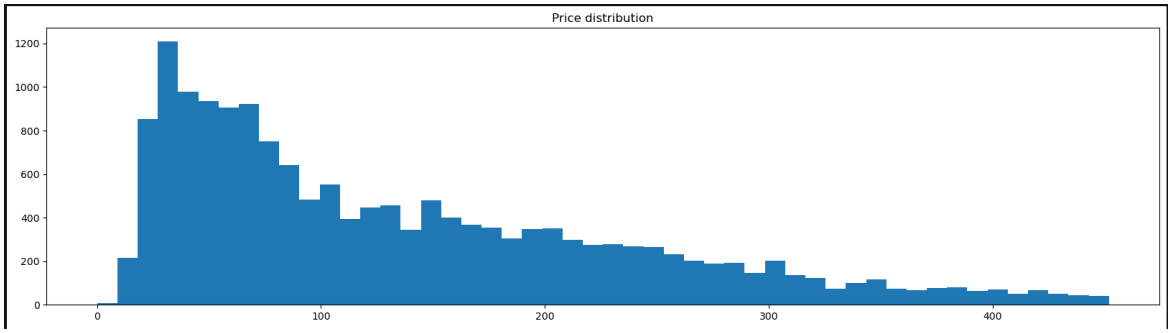


Se encontraron un total de 70 barrios/localidades diferentes.

Para la variable de room_type se encontró que existen 4 tipos diferentes de tipos de habitación y que la que tiene mayor incidencia con un 59% es casa o apartamento entero.

```
[10]: # Calculating the relative frecuency of room types
      listings_df["room_type"].value_counts(dropna=False, normalize=True)

[10]: Entire home/apt    0.595589
      Private room       0.387406
      Shared room        0.008706
      Hotel room         0.008299
      Name: room_type, dtype: float64
```

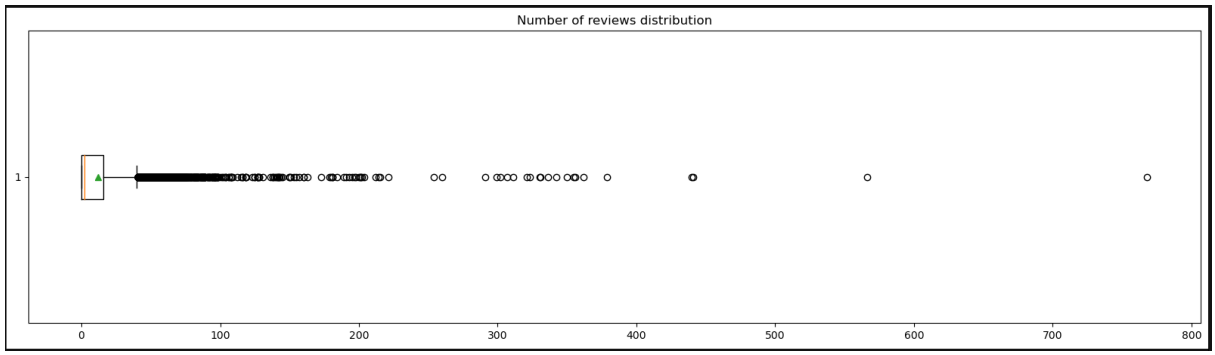
Ahora, con la variable de los precios podemos ver la siguiente distribución (Eje y cantidad de viviendas, eje x precio en euros):



Además, se evaluó la distribución y encontramos una media de 135 euros por día.

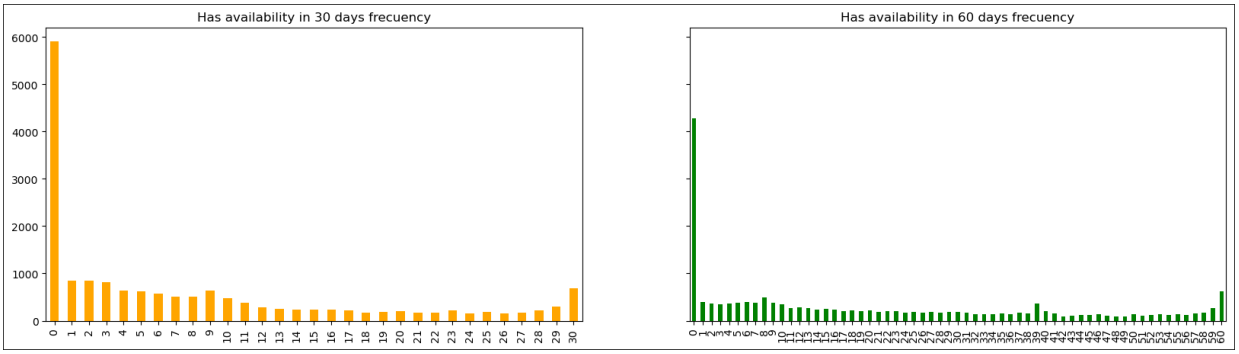
```
listings_df.loc[listings_df["price_float"] <= (q3 + 1.5 * iqr)]["price_float"].describe()

count    16496.000000
mean       135.809772
std        100.904271
min          0.000000
25%         53.000000
50%        104.000000
75%        199.000000
max        452.000000
Name: price_float, dtype: float64
```

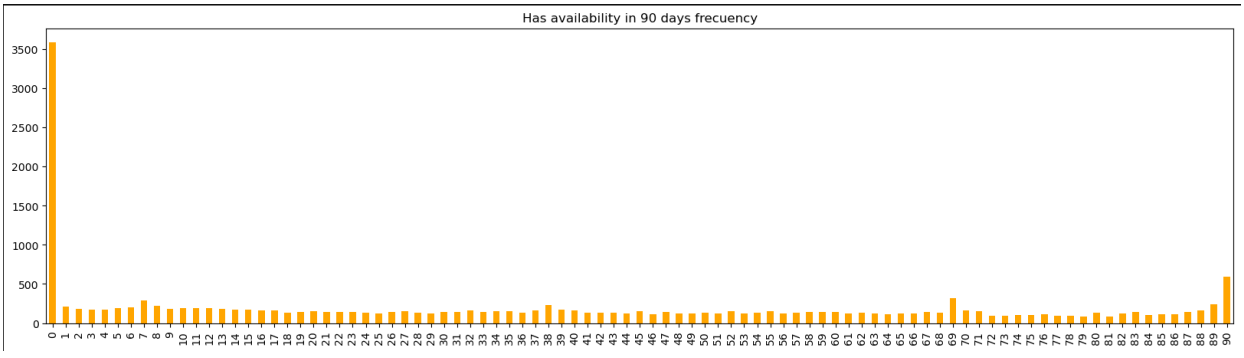
Por otro lado, se encontró con la variable de number_of_reviews_ltm que existe una media de 11.98 reviews por año y una mediana de 2 reviews por año.

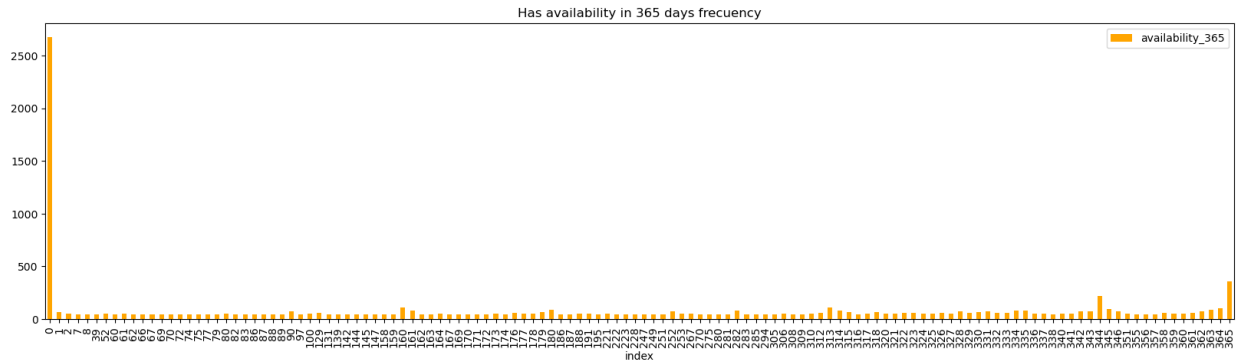


```
[77]: listings_df["number_of_reviews_ltm"].describe()

[77]: count    17230.000000
      mean        11.988799
      std         23.295948
      min          0.000000
      25%          0.000000
      50%          2.000000
      75%         16.000000
      max        768.000000
      Name: number_of_reviews_ltm, dtype: float64

[76]: listings_df["number_of_reviews_ltm"].median()

[76]: 2.0
```

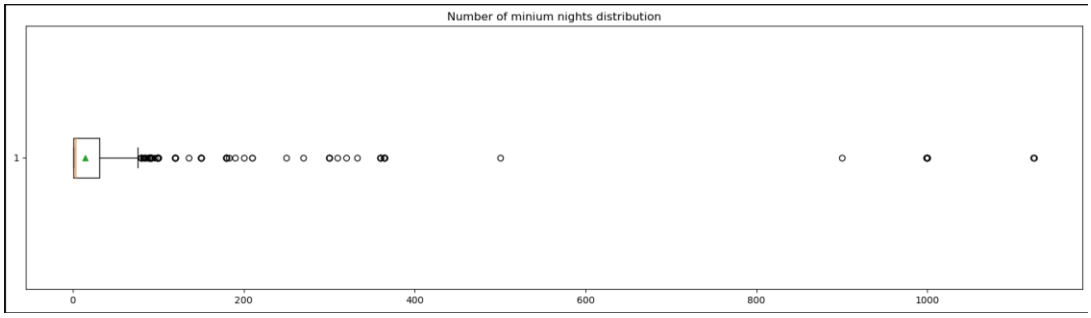
Con las variables de disponibilidad en el futuro tenemos distintos rangos de tiempo (30, 60, 90 y 365 días), con esto presente encontramos que la mayor frecuencia en todos los rangos de tiempo son las propiedades que tienen 0 días disponibles o tienen el máximo de tiempo disponible. Por lo tanto, surgen ciertas dudas, ya que o son propiedades que ya no dan fechas disponibles o tienen noches mínimas muy grandes o tienen precios muy grandes que hacen que no los reserven con tanta antelación en términos de tiempo.





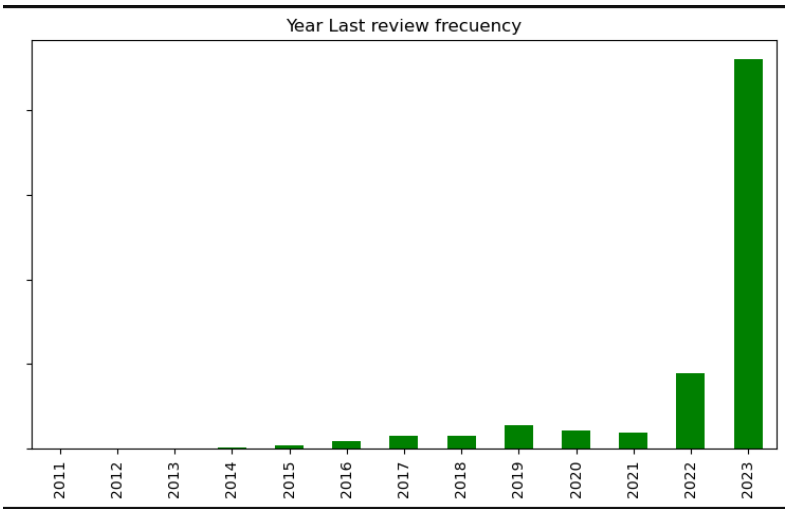


Con la variable de minimum_nights se encontró que existe una media de 14.37 noches minímas por cada propiedad y se tiene una mediana de 3 noches mínimas que es más diciente teniendo en cuenta que contamos con unos outliers muy altos de hasta 1000 días.




```
[78]: listings_df["minimum_nights"].describe()

[78]: count    17230.000000
      mean        14.375218
      std         33.905670
      min          1.000000
      25%          1.000000
      50%          3.000000
      75%         31.000000
      max       1125.000000
      Name: minimum_nights, dtype: float64

[79]: listings_df["minimum_nights"].median()

[79]: 3.0
```

Por último, un detalle muy interesante es que se puede ver la distribución del año de la última reseña (last_review) realizada a una propiedad y se obtuvo lo siguiente:



Se encontró que existen propiedades que ya han dejado de estar activas en cuánto a reseñas se refiere, por lo tanto, se puede inferir que quizás ha estado inactiva en gran medida en cuánto a estar en servicio de arrendamiento.

3. [15%] Estrategia de análisis: Describir de manera concreta en uno o dos párrafos la estrategia de análisis y las técnicas que utilizará, con su respectiva justificación

A partir de las variables explorados en el punto anterior, se tiene como una propuesta de en primera instancia eliminar ciertos datos que quizás no tengan mucho valor y hagan ruido en los análisis que se van a realizar y uno de ellos serían las propiedades que estén inactivas y no sean arrendadas desde el 2021 hacía atrás (Desde la pandemia hacía atrás). Lo anterior, se puede comprobar mediante la columna last_review que nos permite ver el año de la última review recibida por parte de la propiedad. La limpieza de datos propuesta va con el fin de dar unas conclusiones muy cercanas a la actualidad sin tener datos que puedan generar ruido debido a la situación que vivió el mundo con la pandemia del COVID-19. Luego, se tienen que crear dataframes paralelos al dataframe principal para poder eliminar las variables que estén vacías, nulas o que tengan valores inconsistentes de manera controlada y específica sobre ciertos análisis que se quieran realizar.

Los análisis e incidencias que se quieren ver para poder ver las mejores oportunidades de inversión son las siguientes:

- Ver los barrios con mayor cantidad de propiedades y cuales de ellos tienen más relación con atributos de buenas calificaciones y número de reviews.
- Ver si las noches mínimas indican algún tipo de incidencia ya sea con el tipo de sector de la ciudad o los precios establecidos.
- Determinar si el precio puede ser relacionado por ciertos factores como el barrio, número de personas máximas, noches mínimas, etc.

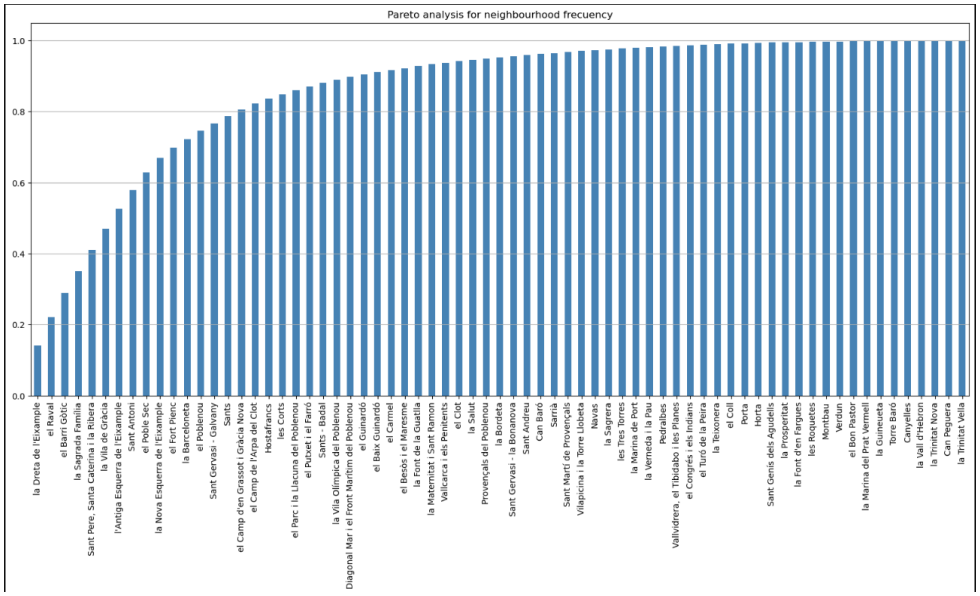
4. [40%] Desarrollo de la estrategia: Implementar la estrategia previamente definida en donde se evidencie claramente cualquier tipo de procesamiento de datos que haya tenido que hacer, la técnica estadística o de visualización de datos utilizada y los insights extraídos tras la interpretación de los resultados.

```
[104]: listings_df["year_last_review"]= listings_df["year_last_review"].astype(float)

[116]: fil_listings_df= listings_df[(listings_df["year_last_review"]>=2022) & ~listings_df["year_last_review"].isnull()].reset_index(drop=True).copy()

[117]: fil_listings_df[["review_scores_rating", "review_scores_accuracy", "review_scores_cleanliness", "review_scores_checkin", "review_scores_communication", "review_scores_location", "review_scores_value"]].info()

       <class 'pandas.core.frame.DataFrame'>
       RangeIndex: 10982 entries, 0 to 10981
       Data columns (total 7 columns):
        #   Column                       Non-Null Count  Dtype
       ---  ------                       --------------  -----
        0   review_scores_rating         10982 non-null  float64
        1   review_scores_accuracy       10982 non-null  float64
        2   review_scores_cleanliness    10982 non-null  float64
        3   review_scores_checkin        10982 non-null  float64
        4   review_scores_communication  10982 non-null  float64
        5   review_scores_location       10982 non-null  float64
        6   review_scores_value          10982 non-null  float64
       dtypes: float64(7)
       memory usage: 600.7 KB

[118]: fil_listings_df[["review_scores_rating", "review_scores_accuracy", "review_scores_cleanliness", "review_scores_checkin", "review_scores_communication", "review_scores_location", "review_scores_value"]].describe()
```

[118]:

	review_scores_rating	review_scores_accuracy	review_scores_cleanliness	review_scores_checkin	review_scores_communication	review_scores_location	review_scores_value
count	10982.000000	10982.000000	10982.000000	10982.000000	10982.000000	10982.000000	10982.000000
mean	4.565583	4.620805	4.592450	4.708247	4.709649	4.747220	4.442202
std	0.471738	0.462206	0.464901	0.414679	0.432853	0.342646	0.515860
min	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
25%	4.400000	4.500000	4.450000	4.620000	4.620000	4.670000	4.270000
50%	4.670000	4.740000	4.710000	4.830000	4.830000	4.830000	4.550000
75%	4.870000	4.910000	4.900000	4.970000	4.980000	4.960000	4.750000
max	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000

Se eliminaron los datos que en el análisis se consideraron obsoletas con respecto al año de la fecha de última revisión recibida de la propiedad y se eliminaron un total de casi 7000 mil datos, que, aunque parezcan muchos, se consideran ruido si quiero hacer una recomendación de compra y se quiere ver como se comportan las propiedades activas después de pandemia (año 2021).



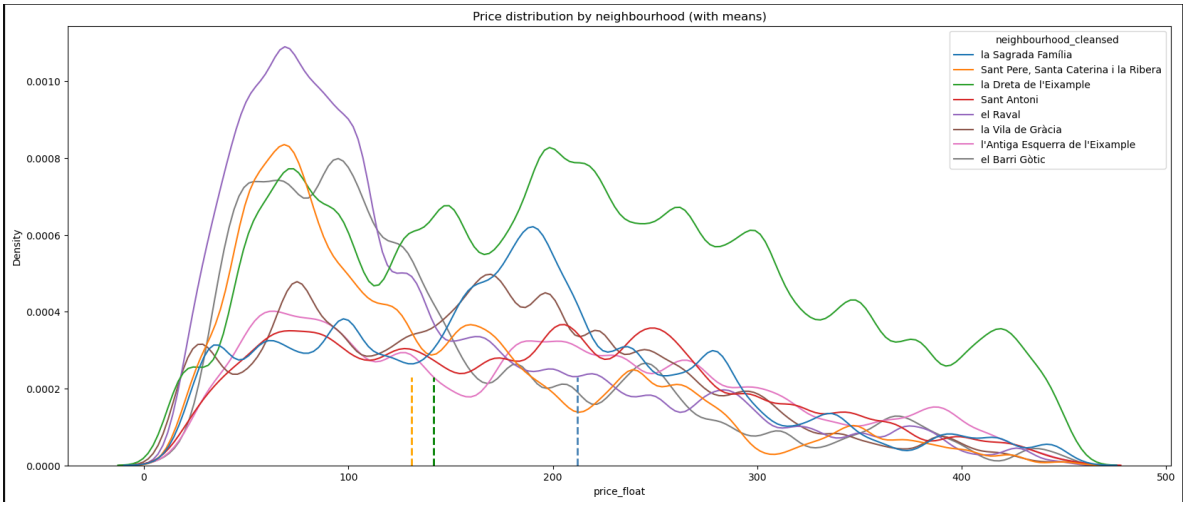

```
[164]: most_representative_neighbourhoods = neighbourhood_frec_cumsum.loc[neighbourhood_frec_cumsum < 0.6].index.tolist()
       most_representative_neighbourhoods

[164]: ["la Dreta de l'Eixample",
        'el Raval',
        'el Barri Gòtic',
        'la Sagrada Família',
        'Sant Pere, Santa Caterina i la Ribera',
        'la Vila de Gràcia',
        "l'Antiga Esquerra de l'Eixample",
        'Sant Antoni']
```

Se realiza un análisis de Pareto sobre la frecuencia relativa de la cantidad de propiedades en ciertos barrios y se encontraron los mejores 8 barrios en términos de cantidad de propiedades. Destacando los tres primeros barrios más incidentes que son:

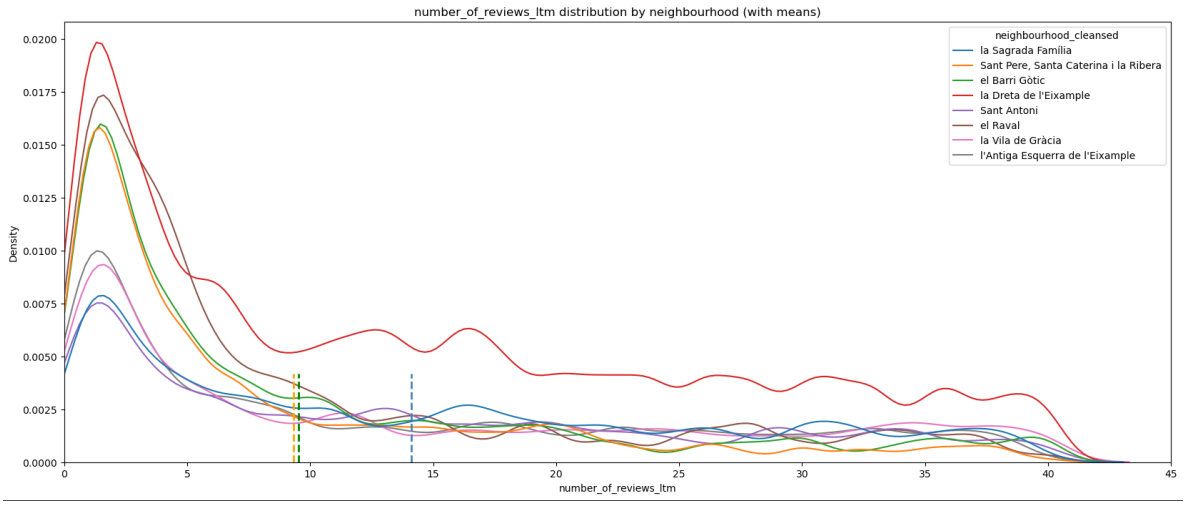
- la Dreta de l'Eixample
- el Raval
- el Barri Gòtic

Ahora, se realizan un análisis de densidad con respecto a los mejores barrios de las siguientes variables:

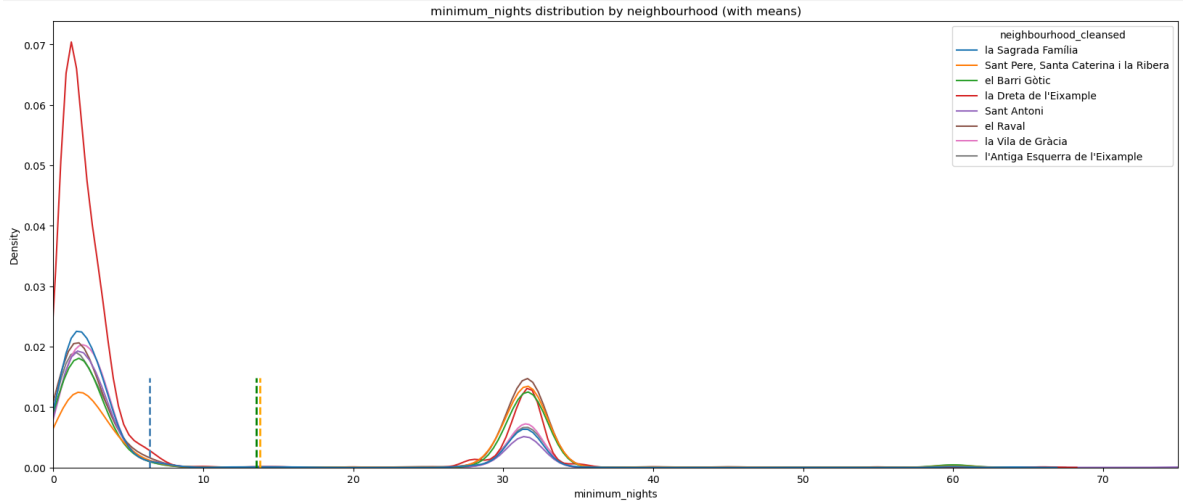
- Price
- number_of_reviews_ltm
- minimum_nights
- review_scores_rating
- availability_30
- availability_60
- availability_90
- availability_365



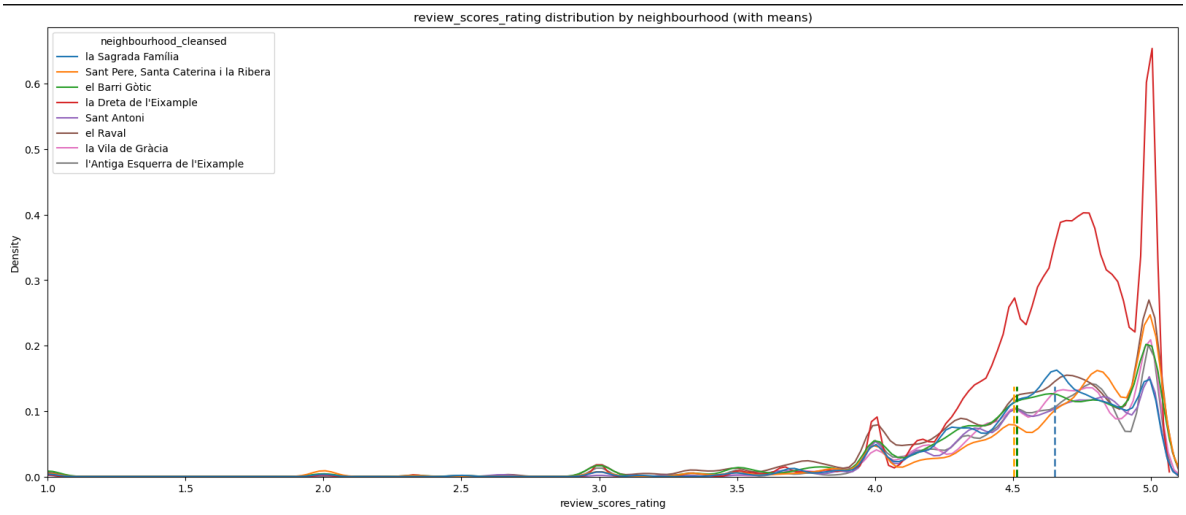
Acá se encuentra que el Raval, el Barri Gótic y el Sant Pere, Santa Caterina i la Ribera tienen precios en promedio mucho más asequibles en su mayoría de propiedades. Por otro lado, un barrio que tiene una variedad de precios tanto asequibles como caros y una densidad proporcionada en este tema es el barrio la Dreta de L'Eixample.



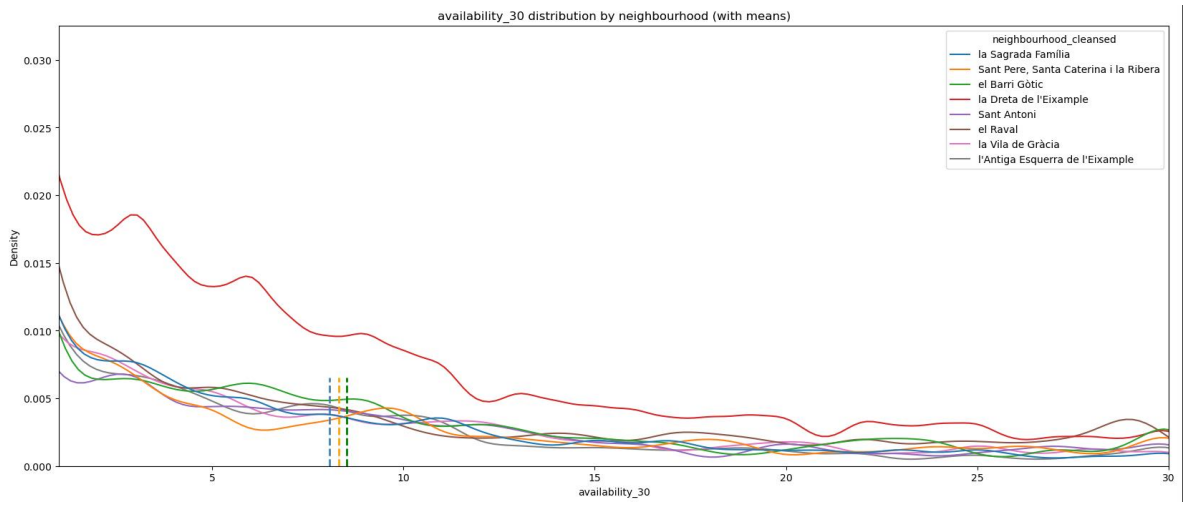
En este apartado se ve un comportamiento muy parecido para todos los barrios y es que la mayoría en el último año obtuvieron alrededor de 1 a 3 reviews. Además, se destaca el barrio la Dreta de L'Eixample por el hecho de que tiene una densidad bastante repartida con las propiedades que tienen 10 reviews o más.

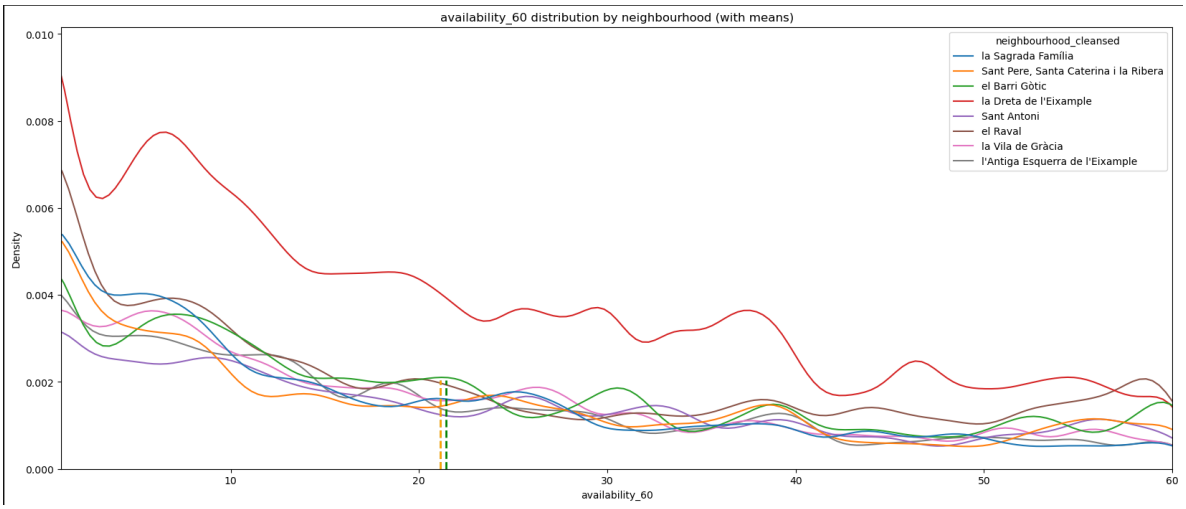


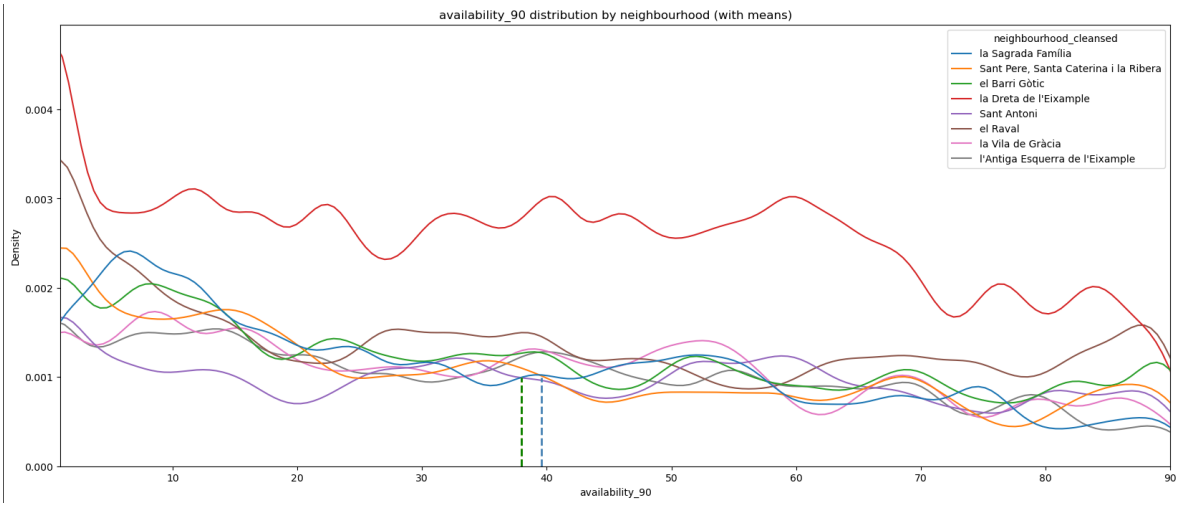
Luego, se observó que existe una gran densidad de los datos en las propiedades que tienen alrededor de 3 noches mínimas en el barrio de la Dreta de L'Eixample y este mismo barrio presenta una densidad mucho menor en noches mínimas más de 32. Con lo anterior dicho se puede afirmar que la Dreta de L'Eixample y la Sagrada Familia son los barrios con más opciones en cuánto a opciones de estadía cortas, mientras que en su forma contraria que los barrios el Raval y Sant Pere, Santa Caterina i la Ribera son los que tienen más opciones para estadías largas mínimas de 32 días en adelante.



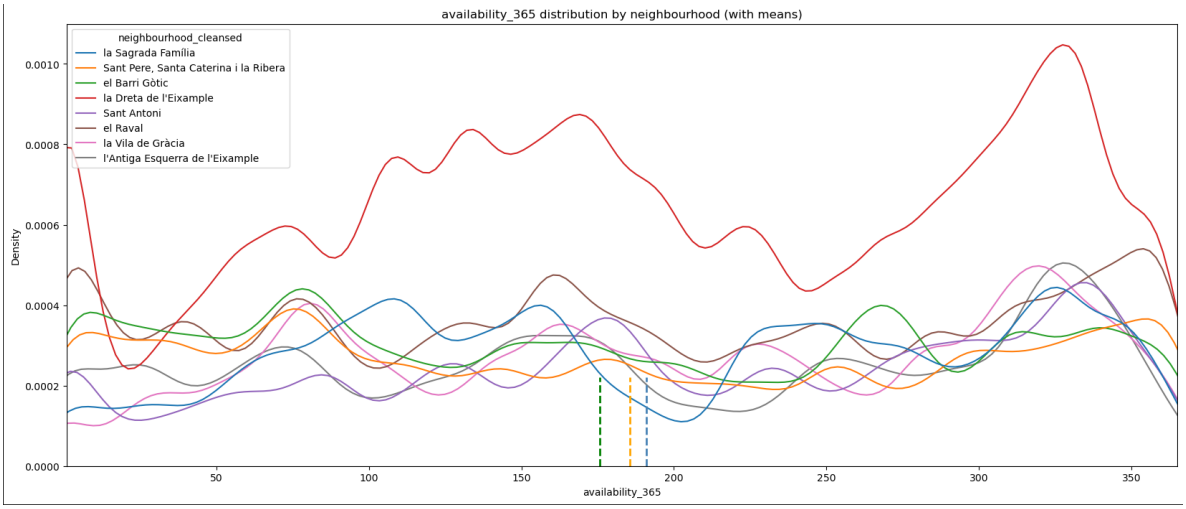
En esta gráfica se puede ver que casi todos los barrios se comportan de una manera parecido, excepto el barrio de la Dreta de l'Eixample que tiene una mayor de densidad de datos desde las calificaciones de 4 hasta la máxima que es 5 por mucha diferencia con respecto a los demás barrios.







En este punto observando las tres gráficas de disponibilidad de 30,60 y 90 días (Se aclara que si una propiedad tiene por ejemplo un valor de 8 quiere decir que tiene 8 días disponibles dentro del tiempo x de días) se determina que tiene comportamientos muy parecidos en cuánto que la mayor densidad de sus datos se presenta en el número de 0 días disponibles, pero se resalta mucho de todas las gráficas el barrio la Dreta de L'Eixample , ya que es el que presenta una densidad mucho más alta que los demás barrios.



En esta gráfica de disponibilidad en un año se puede observar ya un cambio y es que se presentan diferentes picos de densidad en casi todos los barrios, pero se vuelve a resaltar que el barrio la Dreta de L'Eixample tiene una densidad mucho mayor que los demás y se comporta diferente que los demás también.

5. [25%] Generación de resultados: Redactar un informe ejecutivo o una presentación corta en donde recomiende a los inversionistas un pequeño conjunto de sectores, tipos de propiedades, rangos de precios, amenities, entre otros factores de su elección, en los cuales debe invertir. Recuerde que todas estas decisiones deben estar basadas en datos.

Ahora, basándose en todo lo expuesto anteriormente se puede recomendar ciertas opciones de inversión:

- Se podrían comprar propiedades en los barrios el Raval, el Barri Gótic y/o el Sant Pere, Santa Caterina i la Ribera si se quiere entrar en un mercado de precios asequibles y estadías cortas de 1 a 3 días o largas más de 32 días.
- El barrio que genera mucho más expectativa y se destaca en casi todos los aspectos es el barrio la Dreta de L'Eixample que tiene un mercado en todos los tipos de precios, pero estas instancias en su mayoría por densidad apuntan a instancias cortas de 1 a 3 días o hasta un poco más. Lo último mencionado, puede tener sentido si se ve que este barrio se encuentra cerca a la playa de Barcelona y lugares históricos de la ciudad como tal. Además, que este barrio es el que mejores calificaciones recibe en cuánto a densidad por mucha diferencia sobre los demás barrios.
- Es importante resaltar que si se quiere comprar en el barrio la Dreta de L'Eixample puede tener picos de solicitudes debido a la temporada alta que tiene la ciudad de Barcelona que va desde mayo hasta septiembre. Además, este barrio presenta su mayor densidad en no estar disponibles para los próximos 30, 60 y 90 días que, si se sitúa desde la obtención de los datos que fue en el mes de julio, pues tiene sentido con respecto a lo dicho sobre su temporada alta.
- Teniendo en cuenta lo mencionado sobre la temporada alta y la inestabilidad de ocupación en el barrio la Dreta de L'Eixample, se puede considerar la opción de los barrios el Raval, el Barri Gótic y/o el Sant Pere para temporada baja, ya que según la gráfica de disponibilidad en el próximo año presentan una densidad mucho más distribuida sobre los días disponibles que presentan sus propiedades.