Project Summary

Exploratory Data Analysis of Airbnb: The exploratory data analysis (EDA) of Airbnb listings offers valuable insights into the trends and patterns within the platform's ecosystem. By analyzing various attributes such as pricing, location, amenities, and host characteristics, several significant findings emerge.

Key Insights:

Price Distribution: The analysis reveals a wide range of prices across different listings, with certain locations and property types commanding higher prices. Understanding the factors influencing pricing can help both hosts and guests make informed decisions.

Location Trends: Geographic analysis highlights popular neighborhoods and regions favored by Airbnb users. Identifying these hotspots can aid in strategic decision-making for hosts and provide valuable information for travelers seeking desirable locations.

Property Characteristics: Analysis of property attributes, including size, type, and amenities, sheds light on the preferences of Airbnb guests. Hosts can leverage this information to optimize their listings and enhance guest satisfaction.

Host Profiles: Exploring host demographics and behaviors offers insights into the diversity of individuals participating in the Airbnb marketplace. Understanding host characteristics can inform strategies for attracting and retaining hosts, thereby enriching the platform's offerings.

Seasonal Variations: Examination of booking patterns over time reveals seasonal fluctuations in demand, which can influence pricing strategies and resource allocation for hosts and Airbnb administrators.

Problem Statements :

1. the distribution of airbnb price range.
2. Total Listing/Property count in Each Neighborhood Group.
3. Average Price Of listings/property in each Neighborhood Groups.
4. Top Neighborhoods and Hosts by Listing/property in entire NYC.
5. Active Hosts Per Location by Each Neighborhood Groups.
6. room type in entire NYC.
7. which type of rooms people like most.
8. Most reviewed room type in Neighborhood groups.
9. top 5 night stayed in neighbourhood.
10. top 5 host name in minimum night.

there is a lot of problem statements and we have to finds information and insights through different different problem statements so now lets start...

Importing the necessary libraries

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

Load Airbnb NYC Dataset

In [2]:
```
df=pd.read_csv('E:/New folder/Airbnb NYC 2019.csv')
df.head()
```

Out[2]:

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitu
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.647
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.753
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.809
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.685
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.798

UNDERSTAND THE GIVEN VARIABLES

id: The unique identifier for each Airbnb listing.

name: The title or name of the Airbnb listing.

host_id: The unique identifier for the host of the listing.

host_name: The name of the host who manages the listing.

neighbourhood_group: The broader geographical area or district within which the listing is located (e.g., borough, district).

neighbourhood: The specific neighborhood or area within the neighbourhood_group where the listing is situated.

latitude: The latitude coordinate of the listing's location.

longitude: The longitude coordinate of the listing's location.

room_type: The type of room offered in the listing (e.g., entire home/apartment, private room, shared room).

price: The nightly price for renting the listing.

minimum_nights: The minimum number of nights required for booking the listing.

number_of_reviews: The total number of reviews left for the listing.

last_review: The date of the most recent review left for the listing.

reviews_per_month: The average number of reviews the listing receives per month.

calculated_host_listings_count: The total number of listings managed by the host.

availability_365: The number of days the listing is available for booking within a 365-day period.

Data Exploration and Data Cleaning

In [34]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 48895 entries, 0 to 48894
Data columns (total 16 columns):
 #   Column                          Non-Null Count  Dtype  
---  ------                          --------------  -----  
 0   id                              48895 non-null  int64  
 1   name                            48895 non-null  object 
 2   host_id                         48895 non-null  int64  
 3   host_name                       48895 non-null  object 
 4   neighbourhood_group             48895 non-null  object 
 5   neighbourhood                   48895 non-null  object 
 6   latitude                        48895 non-null  float64
 7   longitude                       48895 non-null  float64
 8   room_type                       48895 non-null  object 
 9   price                           48895 non-null  int64  
 10  minimum_nights                  48895 non-null  int64  
 11  number_of_reviews               48895 non-null  int64  
 12  last_review                     48895 non-null  object 
 13  reviews_per_month               48895 non-null  float64
 14  calculated_host_listings_count  48895 non-null  int64
```

So, host_name, neighbourhood_group, neighbourhood and room_type fall into categorical variable category.

While host_id, latitude, longitude, price, minimum_nights, number_of_reviews, last_review, reviews_per_month, host_listings_count, availability_365 are numerical variables

In [4]:
```
df.isnull().sum()
```

Out[4]:
```
id                                  0
name                               16
host_id                             0
host_name                          21
neighbourhood_group                 0
neighbourhood                       0
latitude                            0
longitude                           0
room_type                           0
price                               0
minimum_nights                      0
number_of_reviews                   0
last_review                     10052
reviews_per_month               10052
calculated_host_listings_count      0
availability_365                    0
dtype: int64
```

In summary, encountering null values in the dataset's "name," "host_name," "last_review," and "reviews_per_month" columns suggests that there is missing information in these fields. This could be due to various reasons such as incomplete data entry, data collection issues, or other factors. Handling null values involves imputation depending on the analysis requirements and dataset characteristics.

fill the null value

In [5]:
```
df.name.fillna("unknown",inplace=True)
```

In [6]:
```
df.host_name.fillna("unknown",inplace=True)
```

In [7]:
```
df.last_review.fillna("unknown",inplace=True)
```

In [8]:
```
df.reviews_per_month.fillna(0,inplace=True)
```

In [9]:
```
df.isnull().sum()
```

Out[9]:
```
id                                0
name                              0
host_id                           0
host_name                         0
neighbourhood_group               0
neighbourhood                     0
latitude                          0
longitude                         0
room_type                         0
price                             0
minimum_nights                    0
number_of_reviews                 0
last_review                       0
reviews_per_month                 0
calculated_host_listings_count    0
availability_365                  0
dtype: int64
```

In [10]:
```
df.columns
```

Out[10]:
```
Index(['id', 'name', 'host_id', 'host_name', 'neighbourhood_group',
       'neighbourhood', 'latitude', 'longitude', 'room_type', 'price',
       'minimum_nights', 'number_of_reviews', 'last_review',
       'reviews_per_month', 'calculated_host_listings_count',
       'availability_365'],
      dtype='object')
```

Data Visualization

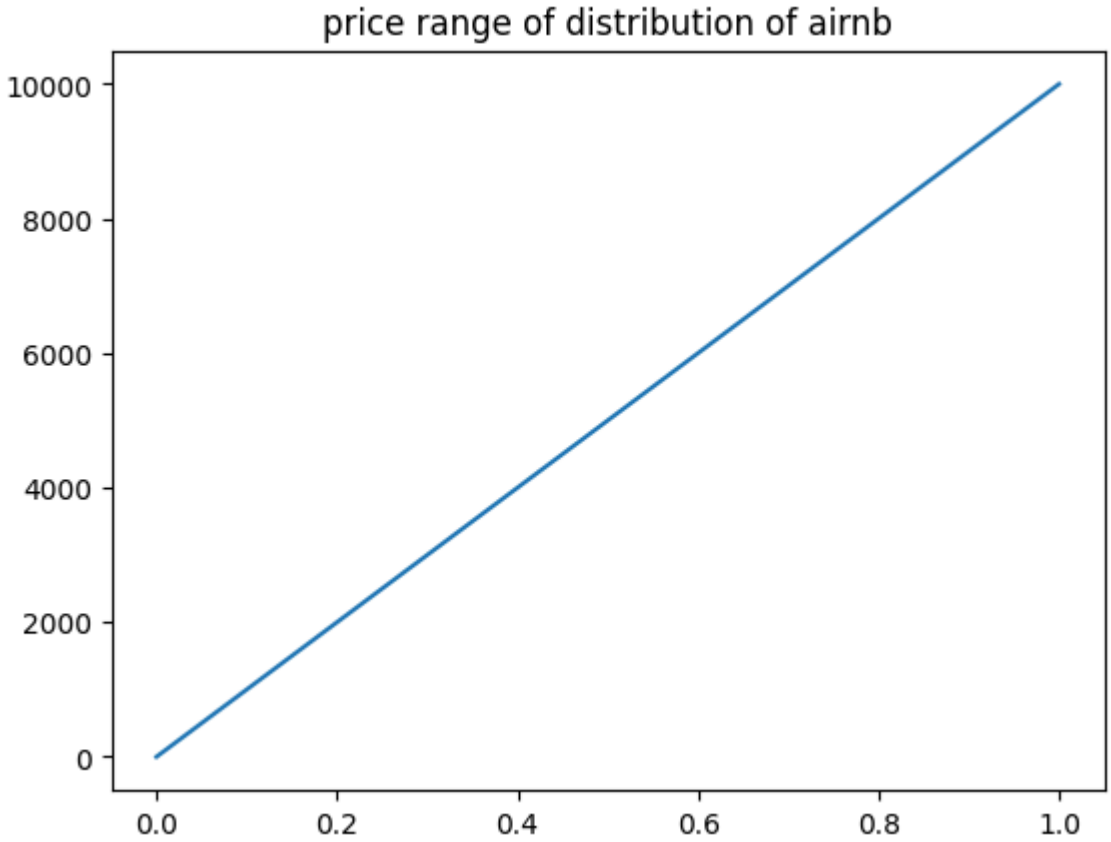
the distribution of airbnb price range

In [11]:
```
price_range=df['price'].min(),df['price'].max()
print("the distribution of airbnb price range is : ",price_range)
```

```
the distribution of airbnb price range is :  (0, 10000)
```

The distribution of Airbnb price range is described as being between $0 and 10,000. This indicates that prices can vary widely, ranging from no cost up to a maximum of $10,000. This range encompasses a diverse set of listings, from budget accommodations to luxury rentals. Understanding this distribution is important for analyzing pricing trends, identifying outliers, and making informed decisions about accommodation options on Airbnb.

In [12]:
```
plt.plot(price_range)
plt.title('price range of distribution of airnb')
plt.show()
```



Total Listing/Property count in Each Neighborhood Group

In [13]:
```
nei = df.groupby(['neighbourhood_group'])
total_count = nei[['calculated_host_listings_count']].sum()
total_count = total_count.sort_values(['calculated_host_listings_count'],as
total_count
```

Out[13]:

neighbourhood_group	calculated_host_listings_count
Manhattan	277073
Brooklyn	45925
Queens	23005
Bronx	2437
Staten Island	865

This information gives an overview of the distribution of Airbnb listings across the various boroughs of New York City, indicating that Manhattan has the highest number of listings, followed by Brooklyn, Queens, the Bronx, and Staten Island.

In [2]:
```
total_count.plot(kind='bar',figsize=(6,5),color='y',fontsize=10,alpha=0.4)
plt.xlabel('neighbourhood_group',fontsize=20,style='oblique')
plt.ylabel('calculated_host_listings_count',fontsize=10,style='oblique')
plt.title('Total Listing/Property count in Each Neighborhood Group',fontsiz
plt.show()
```

```
---------------------------------------------------------------------------
-
NameError                                 Traceback (most recent call las
t)
Cell In[2], line 1
----> 1 total_count.plot(kind='bar',figsize=(6,5),color='y',fontsize=10,al
pha=0.4)
      2 plt.xlabel('neighbourhood_group',fontsize=20,style='oblique')
      3 plt.ylabel('calculated_host_listings_count',fontsize=10,style='obl
ique')

NameError: name 'total_count' is not defined
```

Average Price Of listings/property in each Neighborhood Groups

In [36]:
```
neighborhood_group=df.groupby(['neighbourhood_group'])
neighborhood_group=neighborhood_group[['price']].mean()
neighborhood_group
```

Out[36]:

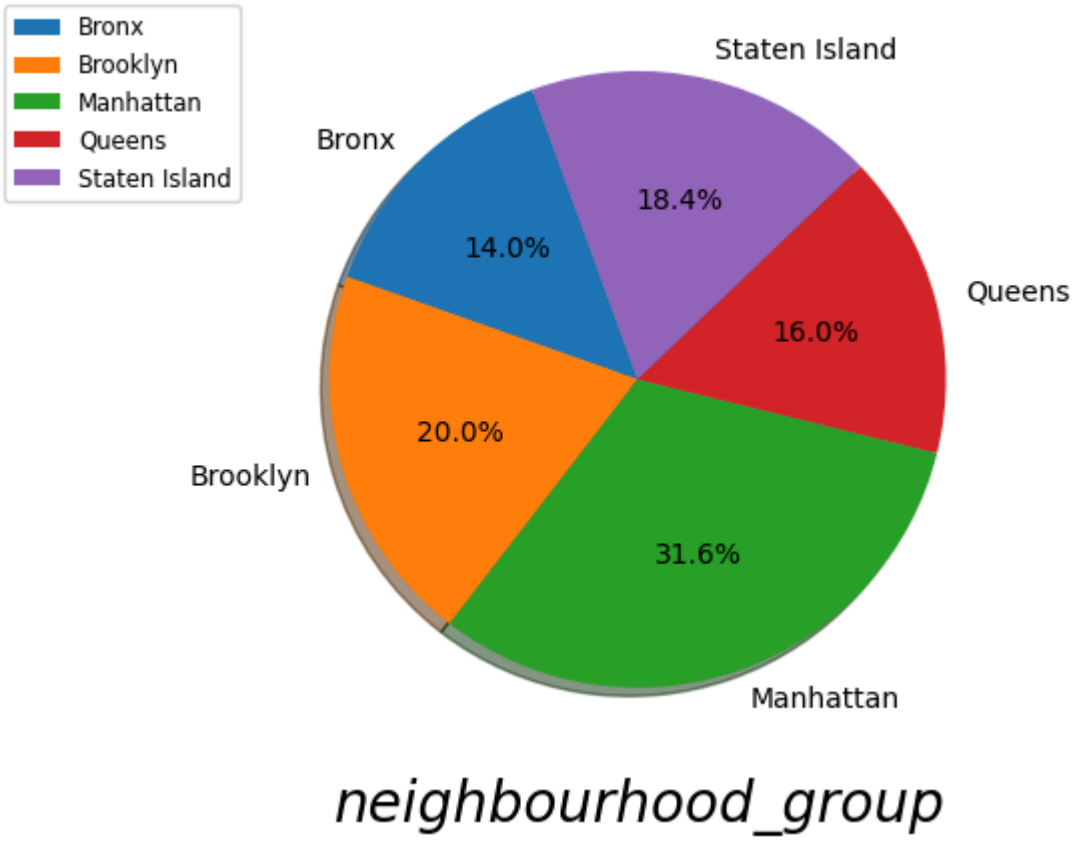
neighbourhood_group	price
Bronx	87.496792
Brooklyn	124.383207
Manhattan	196.875814
Queens	99.517649
Staten Island	114.812332

Bronx:-The average price for Airbnb listings in the Bronx is 87.50 dollers per night. This suggests that accommodations in the Bronx are relatively affordable compared to other boroughs in the city. Brooklyn:-With an average price of 124.38 dollers per night, Brooklyn offers a diverse range of accommodations at various price points.

Manhattan:-As one of the most sought-after areas for accommodations, Manhattan has the highest average price among the boroughs at 196.88 dollers per night. Queens:- Airbnb listings in Queens have an average price of 99.52 dollers per night, making it a relatively affordable option compared to Manhattan. Staten Island:With an average price of 114.81 dollers per night, Staten Island offers moderately priced accommodations compared to other boroughs.

This comparison provides a comprehensive overview of the average Airbnb prices in each neighborhood group, highlighting the affordability and diversity of accommodations across different areas of New York City.

In [16]:
```
neighborhood_group.plot(kind='pie',subplots=True,figsize=(5,5),autopct='%1.
plt.xlabel('neighbourhood_group',fontsize=20,style='italic')
plt.ylabel('',fontsize=20,style='italic')
plt.legend(bbox_to_anchor=(0,1), fontsize='small')
plt.show()
```



Top Neighborhoods and Hosts by Listing/property in entire NYC

In [17]:
```
top=df.groupby(['host_name'])
top=top[['calculated_host_listings_count']].sum()
top=top.sort_values(['calculated_host_listings_count'],ascending=False)
top=top.head()
top
```

Out[17]:

host_name	calculated_host_listings_count
Sonder (NYC)	106929
Blueground	53824
Kara	14679
Kazuya	10609
Sonder	9216

Sonder (NYC): This host has the highest calculated host listings count, with 106,929 listings attributed to them.

Blueground: Blueground follows with 53,824 listings, making them one of the prominent hosts in the city.

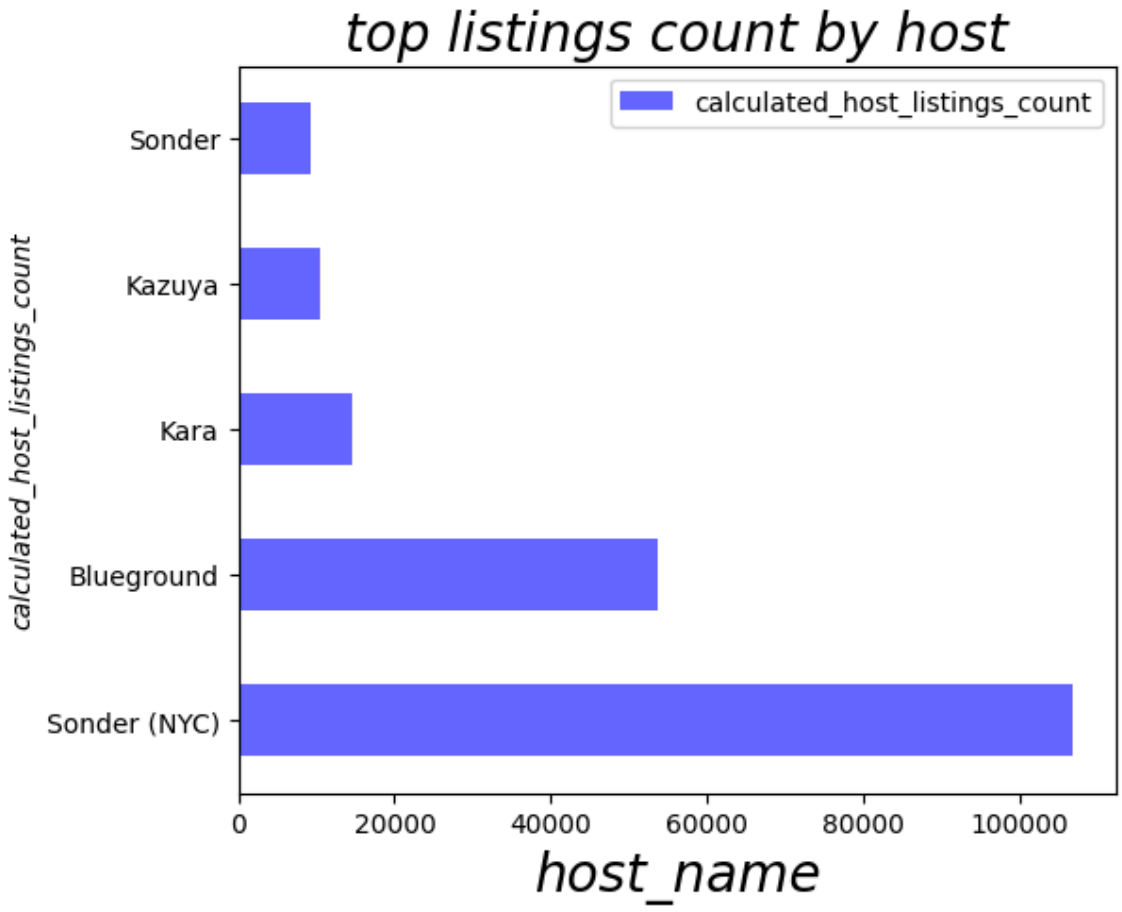
Kara: Kara has 14,679 listings, indicating a significant presence in the Airbnb market in New York City.

Kazuya: Kazuya is another major host with 10,609 listings, contributing to the diversity of available accommodations.

Sonder: Apart from Sonder (NYC), another host named Sonder also appears on the list with 9,216 listings, suggesting a notable presence in the market.

These hosts play a significant role in providing accommodations to guests visiting New York City through Airbnb, with a wide range of listings under their management.

In [18]:
```
top.plot(kind='barh',figsize=(6,5),color='blue',fontsize=10,alpha=0.6)
plt.xlabel('host_name',fontsize=20,style='oblique')
plt.ylabel('calculated_host_listings_count',fontsize=10,style='oblique')
plt.title('top listings count by host',fontsize=20,style='oblique')
plt.show()
```



the Number Of Active Hosts Per Location by Each Neighborhood Groups

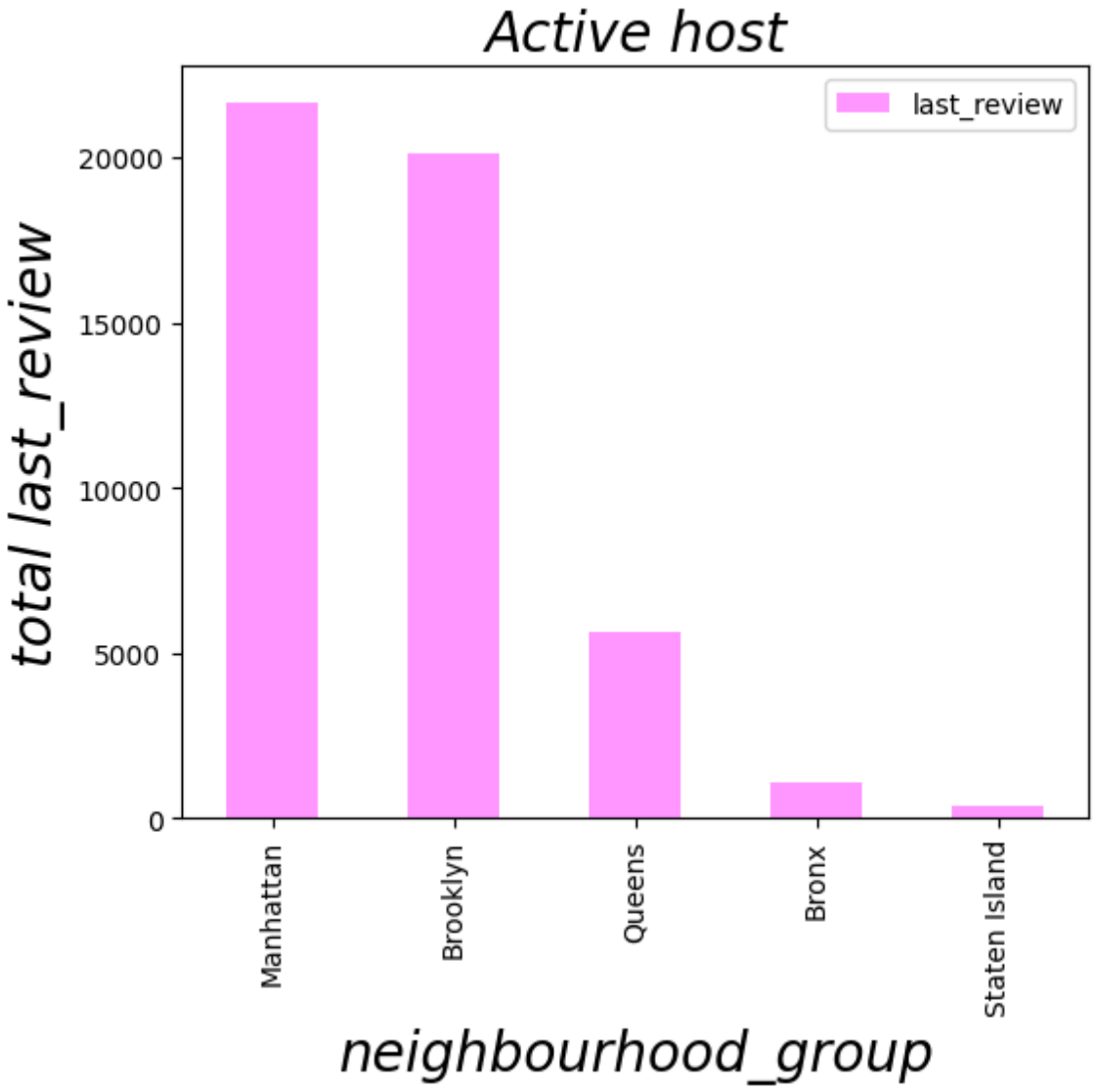
In [20]:
```
cl=df.groupby('neighbourhood_group')['last_review'].count().reset_index().s
cl=cl.sort_values(by='last_review',ascending=False)
cl
```

Out[20]:

neighbourhood_group	last_review
Manhattan	21661
Brooklyn	20104
Queens	5666
Bronx	1091
Staten Island	373

These figures indicate the number of reviews that were most recently made for listings in each borough. It gives insight into the recent activity and popularity of Airbnb accommodations in different parts of New York City.

In [21]:
```
colors=['Magenta']
cl.plot(kind='bar',figsize=(6,5),color=colors,fontsize=10,alpha=0.4)
plt.xlabel('neighbourhood_group',fontsize=20,style='oblique')
plt.ylabel('total last_review',fontsize=20,style='oblique')
plt.title('Active host',fontsize=20,style='oblique')
plt.show()
plt.show()
```



room type

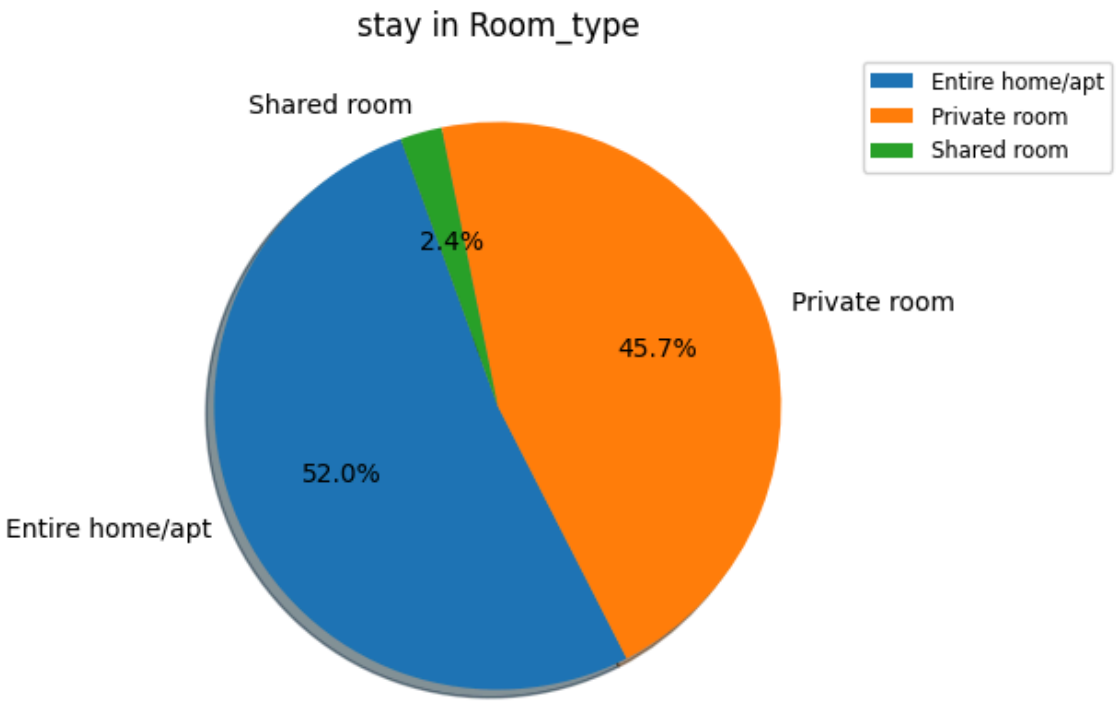
In [22]:
```
room=df.groupby(['room_type'])
room=room[['room_type']].count()
room
```

Out[22]:

	room_type
room_type	
Entire home/apt	25409
Private room	22326
Shared room	1160

In this case, the most common room type is Entire home/apt, with 25,409 . This suggests that the majority of Airbnb accommodations in the dataset offer the entire home or apartment for rent rather than just a private room or a shared room.

In [23]:
```
room.plot(kind='pie',subplots=True,figsize=(5,5),autopct='%1.1f%%',shadow =
plt.legend(bbox_to_anchor=(1,1), fontsize='small')
plt.ylabel('')
plt.title('stay in Room_type')
plt.show()
```



room type vs minimum nights

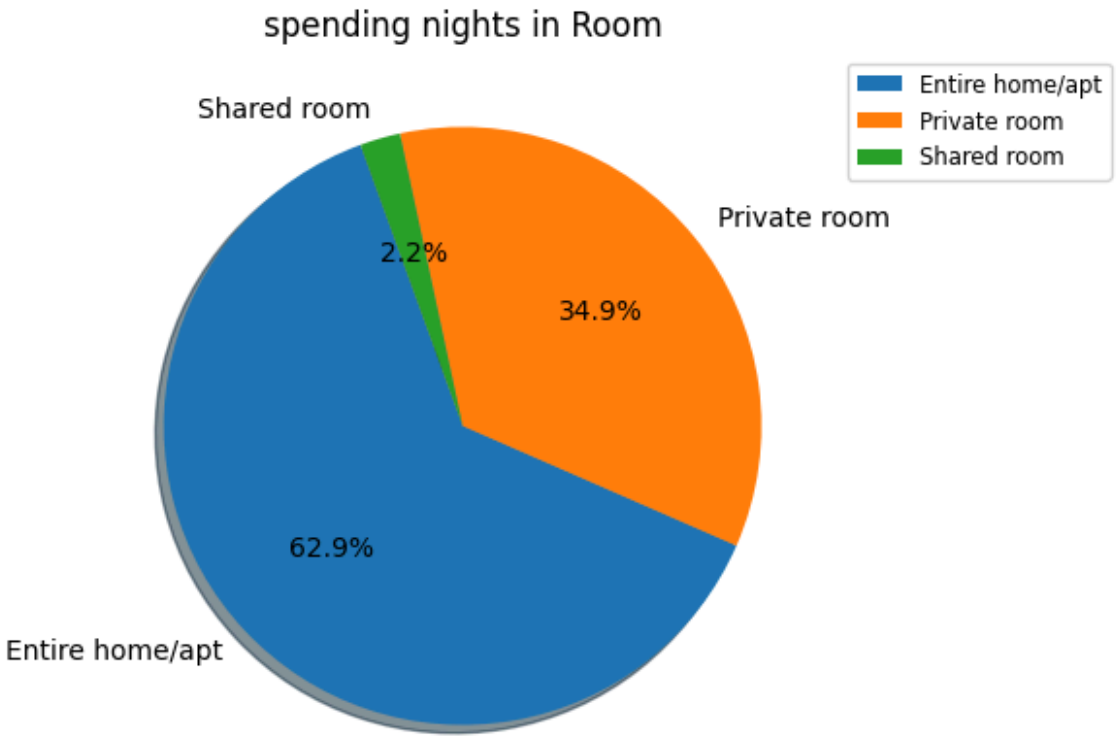
In [24]:
```
stay=df.groupby(['room_type'])
stay=stay[['minimum_nights']].sum()
stay=stay.sort_values(['minimum_nights'],ascending=False)
stay
```

Out[24]:

room_type	minimum_nights
Entire home/apt	216152
Private room	120067
Shared room	7511

the room type "Entire home/apt" has the highest minimum nights requirement, while "Shared room" has the lowest minimum nights requirement.

In [25]:
```
stay.plot(kind='pie',subplots=True,figsize=(5,5),autopct='%1.1f%%',shadow =
plt.legend(bbox_to_anchor=(1,1), fontsize='small')
plt.ylabel('')
plt.title('spending nights in Room')
plt.show()
```



top reviews in neighbourhood_group

In [26]:
```
review=df.groupby(['neighbourhood_group'])
review=review[['number_of_reviews']].sum()
review=review.sort_values(['number_of_reviews'],ascending=False)
review
```

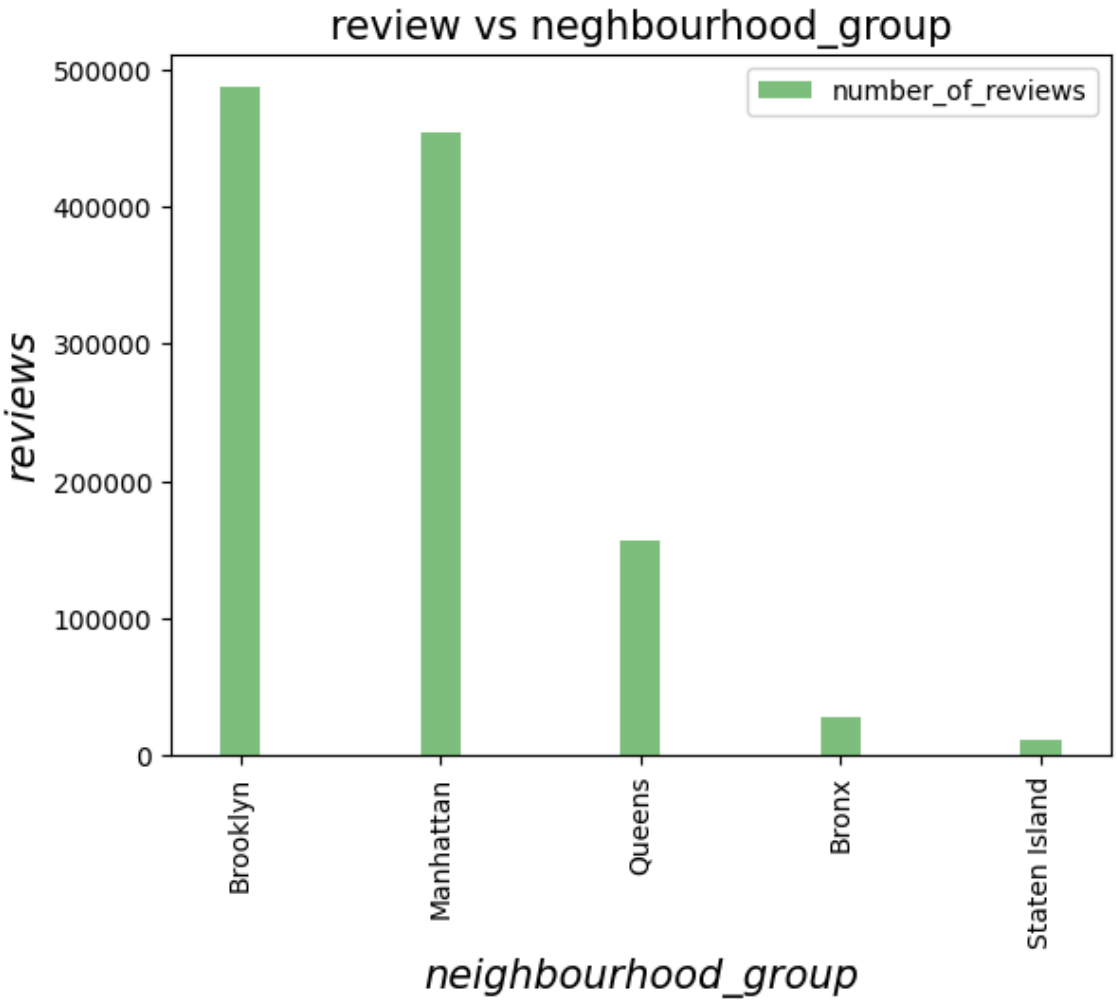
Out[26]:

neighbourhood_group	number_of_reviews
Brooklyn	486574
Manhattan	454569
Queens	156950
Bronx	28371
Staten Island	11541

The total number of reviews for Airbnb listings varies across the different boroughs of New York City. At the lowest end, Staten Island has received 11,541 reviews, reflecting a relatively lower level of review activity compared to other areas. Moving up, the Bronx has garnered 28,371 reviews, indicating a moderate level of guest engagement and feedback. Queens follows with 156,950 reviews, suggesting a substantial number of guests have stayed in accommodations within this borough. Manhattan boasts the second-highest

number of reviews, totaling 454,569, reflecting its popularity as a prime destination for Airbnb stays. Finally, Brooklyn leads with the highest number of reviews at 486,574, showcasing its prominence as a top choice among Airbnb guests for accommodations in New York City.

In [27]:
```
review.plot(kind='bar',width=0.2,color='g',alpha=0.5)
plt.xlabel('neighbourhood_group',style='italic',fontsize=15)
plt.ylabel('reviews',style='italic',fontsize=15)
plt.title('review vs neghbourhood_group',fontsize=15)
plt.show()
```



Most reviewed room type in Neighborhood groups per month.

In [49]:
```
gd=df.groupby(['neighbourhood_group','room_type'])
gd=gd[['reviews_per_month']].sum()
gd=gd.sort_values(by='reviews_per_month',ascending=False)
gd
```

Out[49]:

		reviews_per_month
neighbourhood_group	**room_type**	
Manhattan	**Entire home/apt**	11378.43
Brooklyn	**Entire home/apt**	10757.13
	Private room	10037.27
Manhattan	**Private room**	9141.42
Queens	**Private room**	5220.53
	Entire home/apt	3427.04
Bronx	**Private room**	860.23
	Entire home/apt	691.58
Manhattan	**Shared room**	638.23
Staten Island	**Entire home/apt**	311.16
Brooklyn	**Shared room**	310.58
Staten Island	**Private room**	270.17
Queens	**Shared room**	231.48
Bronx	**Shared room**	58.13
Staten Island	**Shared room**	6.66

Manhattan:

Entire home/apt: 11,378.43 reviews per month

Private room: 9,141.42 reviews per month

Shared room: 638.23 reviews per month

Brooklyn:

Entire home/apt: 10,757.13 reviews per month

Private room: 10,037.27 reviews per month

Shared room: 310.58 reviews per month

Queens:

Entire home/apt: 3,427.04 reviews per month

Private room: 5,220.53 reviews per month

Shared room: 231.48 reviews per month

Bronx:

Entire home/apt: 691.58 reviews per month

Private room: 860.23 reviews per month

Shared room: 58.13 reviews per month

Staten Island:

Entire home/apt: 311.16 reviews per month

Private room: 270.17 reviews per month

Shared room: 6.66 reviews per month

top 5 night stayed in neighbourhood

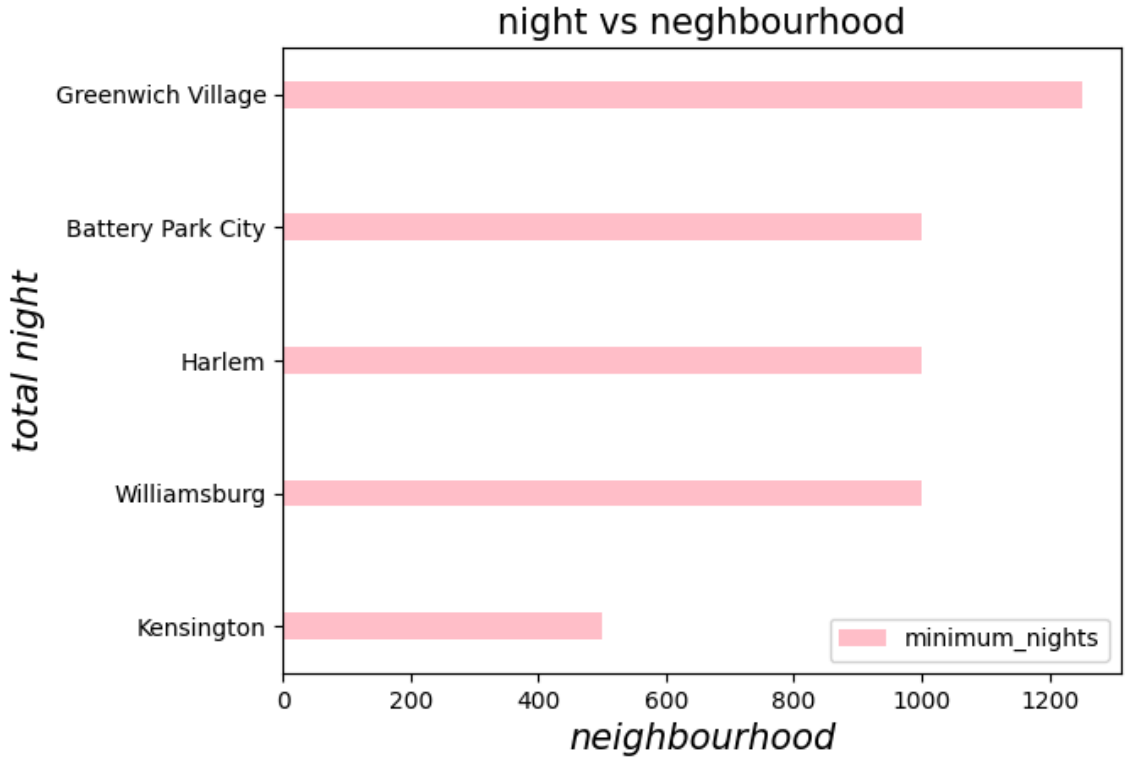
In [29]:
```
ga=df.groupby(['neighbourhood'])
ga=ga[['minimum_nights']].max()
ga=ga.sort_values(['minimum_nights'],ascending=True)
ga=ga.tail(5)
ga
```

Out[29]:

neighbourhood	minimum_nights
Kensington	500
Williamsburg	999
Harlem	999
Battery Park City	1000
Greenwich Village	1250

In New York City, Airbnb listings in various neighborhoods come with different minimum night requirements, reflecting the diverse preferences and policies set by hosts. Kensington boasts a relatively lower minimum night requirement of 500, offering flexibility for shorter stays. Conversely, neighborhoods like Williamsburg and Harlem set the bar higher with a requirement of 999 nights, suggesting a preference for longer-term bookings. Battery Park City takes it a step further with a minimum night requirement of 1000, indicating a focus on extended stays. At the top end of the spectrum, Greenwich Village stands out with a substantial minimum night requirement of 1250, signaling a preference for guests committed to longer durations. These minimum night requirements reflect the unique characteristics and offerings of each neighborhood, catering to a range of guest needs and preferences in the vibrant landscape of New York City.

In [30]:
```
ga.plot(kind='barh',width=0.2,color='pink')
plt.xlabel('neighbourhood',style='italic',fontsize=15)
plt.ylabel('total night',style='italic',fontsize=15)
plt.title('night vs neghbourhood',fontsize=15)
plt.show()
```



top 5 host name in minimum night

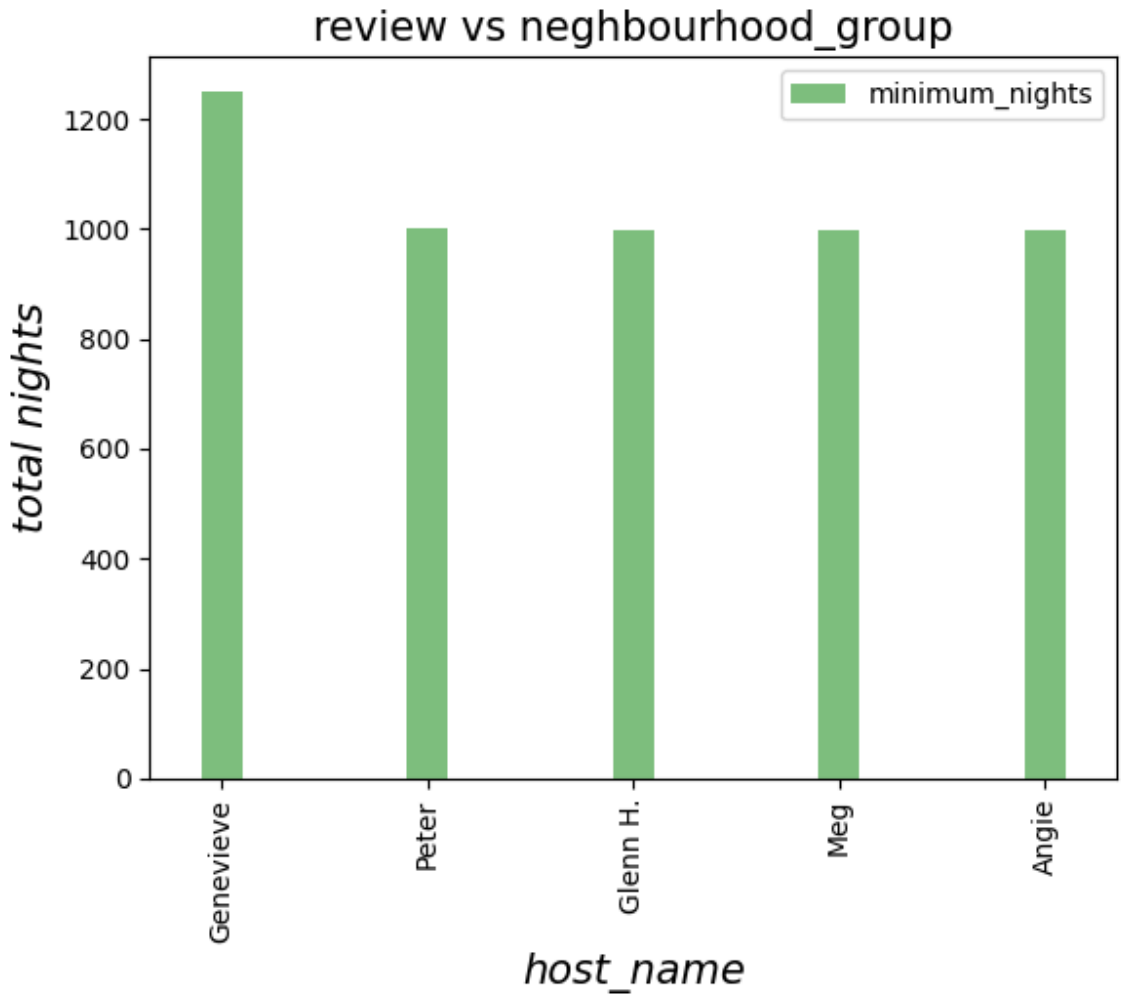
In [31]:
```
gb=df.groupby(['host_name'])
gb=gb[['minimum_nights']].max()
gb=gb.sort_values(['minimum_nights'],ascending=False)
gb=gb.head(5)
gb
```

Out[31]:

host_name	minimum_nights
Genevieve	1250
Peter	1000
Glenn H.	999
Meg	999
Angie	999

Genevieve has the highest minimum night requirement, followed by Peter, Glenn H., Meg, and Angie, all with a minimum requirement of 999 nights.

In [32]:
```
gb.plot(kind='bar',width=0.2,color='g',alpha=0.5)
plt.xlabel('host_name',style='italic',fontsize=15)
plt.ylabel('total nights',style='italic',fontsize=15)
plt.title('review vs neghbourhood_group',fontsize=15)
plt.show()
```



In [ ]: