Capstone Project – 1

EDA - Airbnb Booking Analysis

Team Members

Darshan Chotalia
Ratnendra Chauhan
Riddhish Joshi
Sonica Sinha
Sourabh Umarani

CONTENT

- Introduction
- Problem Statement
- Understanding the Dataset
- Agenda
- Challenges Faced
- Conclusion

Introduction

- Airbnb's New York City(NYC) Booking dataset of 2019 helps us in exploring the insights about host listing and booking patterns in the New York City area.
- We are also analyzing the relationship between - price and areas, property type in particular neighborhood group, price by property type, minimum nights listed in a particular area, etc.
- For the purpose of analysis, we are using Python and its different libraries including Pandas, Matplotlib, Seaborn, and others.
- The key tasks in performing exploratory data analysis include - data loading, data understanding, data cleansing, and data exploration with visualization.

Problem Statement

- The dataset provided is particularly from New York City, USA. The dataset has 49000 entries with 16 variables by which we can assess the data.
- Data analysis on thousands of listings provided through Airbnb is a crucial factor for the company.
- Some of the important columns may include neighborhood groups, price, minimum nights, etc.
- We will try to find some important insights based on the questions.

Understanding the Dataset

List of columns

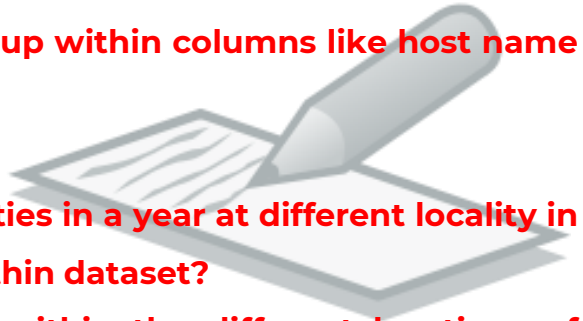
- **id:** listing ID
- **name:** name of the listing
- **host_id:** host ID
- **host_name:** name of the host
- **neighbourhood_group:** location
- **neighbourhood:** area
- **latitude:** latitude coordinates
- **longitude:** longitude coordinates
- **room_type:** listing space type
- **price:** price in dollars
- **minimum_nights:** amount of nights minimum
- **number_of_reviews:** number of reviews
- **last_review:** latest review
- **reviews_per_month:** number of reviews per month
- **calculated_host_listings_count:** amount of listing per host
- **availability_365:** number of days when listing is available for booking

- **There are 49,000 observations with various types of field in our dataset.**

Agenda

→ **We try to answer following questions for Airbnb:**

- **What is the different types of room types and neighbourhood group available in the dataset?**
- **What is the current status of unique neighbourhood group within columns like host name and host listing?**
- **What is the distribution of price in Airbnb dataset?**
- **What is the average room rent and availability of properties in a year at different locality in accordance with minimum nights and price provided within dataset?**
- **What is the overall density and distribution of price within the different locations of neighbourhood group and also showing the price available under $500?**
- **How is neighbourhood, neighbourhood group and room types related to each other in top locality and does it get affected by the booking of guest?**
- **What is the trend of minimum nights within the area in respect of average price range?**



Unique values of the different variables identified during data cleaning

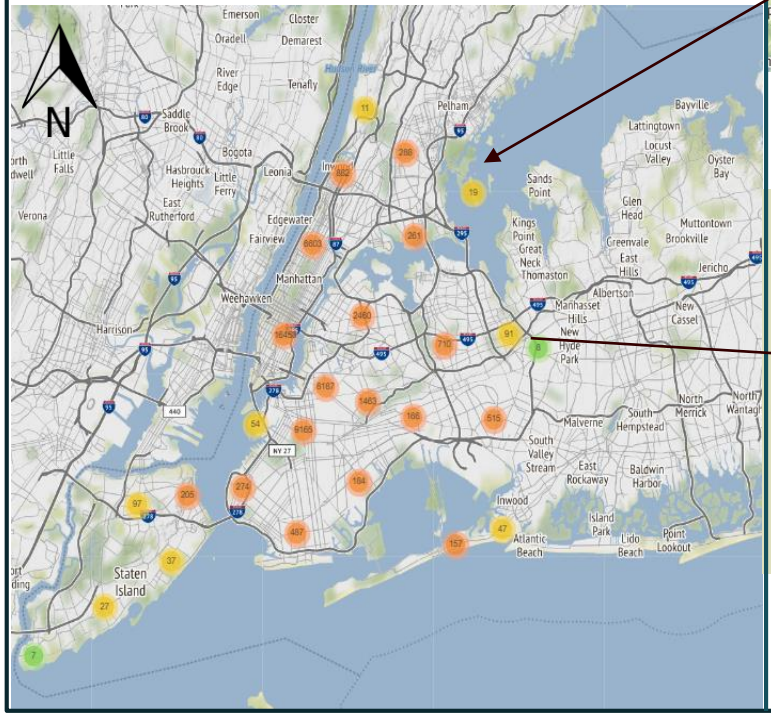
```
Total Unique Values in id - 48895
Total Unique Values in name - 47906
Total Unique Values in host_id - 37457
Total Unique Values in host_name - 11453
Total Unique Values in neighbourhood_group - 5
Total Unique Values in neighbourhood - 221
Total Unique Values in latitude - 19048
Total Unique Values in longitude - 14718
Total Unique Values in room_type - 3
Total Unique Values in price - 674
Total Unique Values in minimum_nights - 109
Total Unique Values in number_of_reviews - 394
Total Unique Values in last_review - 1765
Total Unique Values in reviews_per_month - 938
Total Unique Values in calculated_host_listings_count - 47
Total Unique Values in availability_365 - 366
```

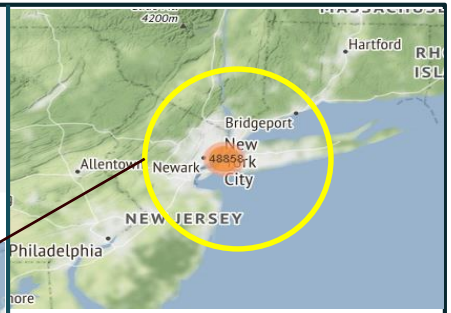
Map of New York City(NYC)

Along with pointing out the locations hosts at different neighbourhood areas.

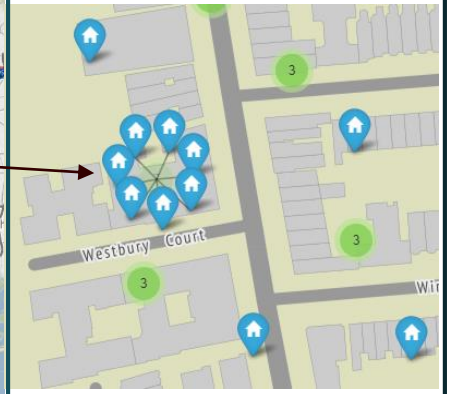
```
# Getting location address
address = 'New York City, NY'

# Getting the latitude and longitude of New York City, USA
geolocator = Nominatim(user_agent="Location_NY")
location = geolocator.geocode(address)
latitude = location.latitude
longitude = location.longitude
print('The geograpical coordinate of New York City are {}, {}.'.format(latitude, longitude))
```

```
The geograpical coordinate of New York City are 40.7127281, -74.0060152.
```





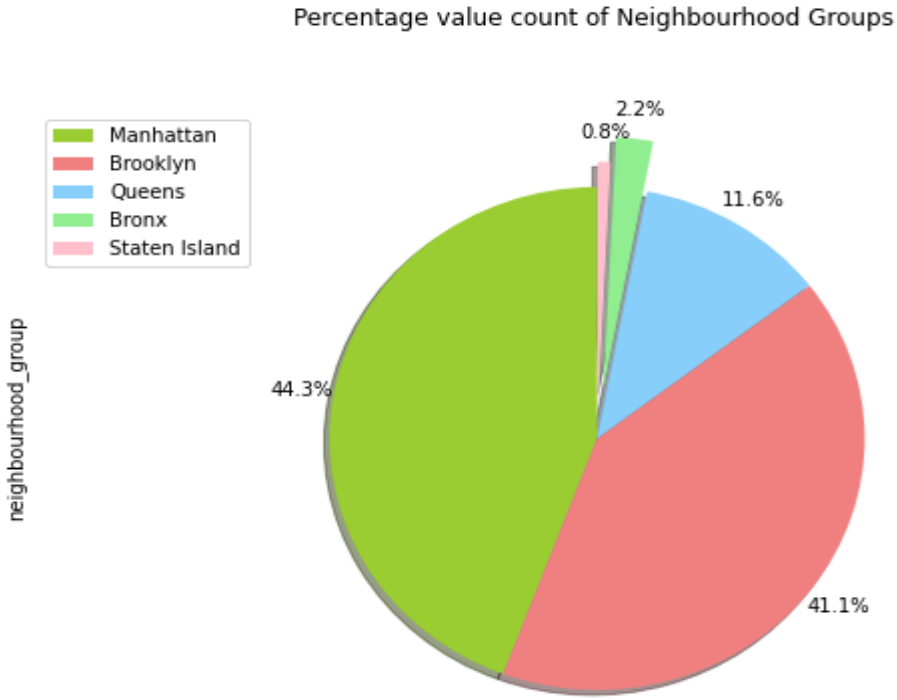
Location of New York City (NYC), USA



❖ What is the different types of room types and neighbourhood group available in the dataset?

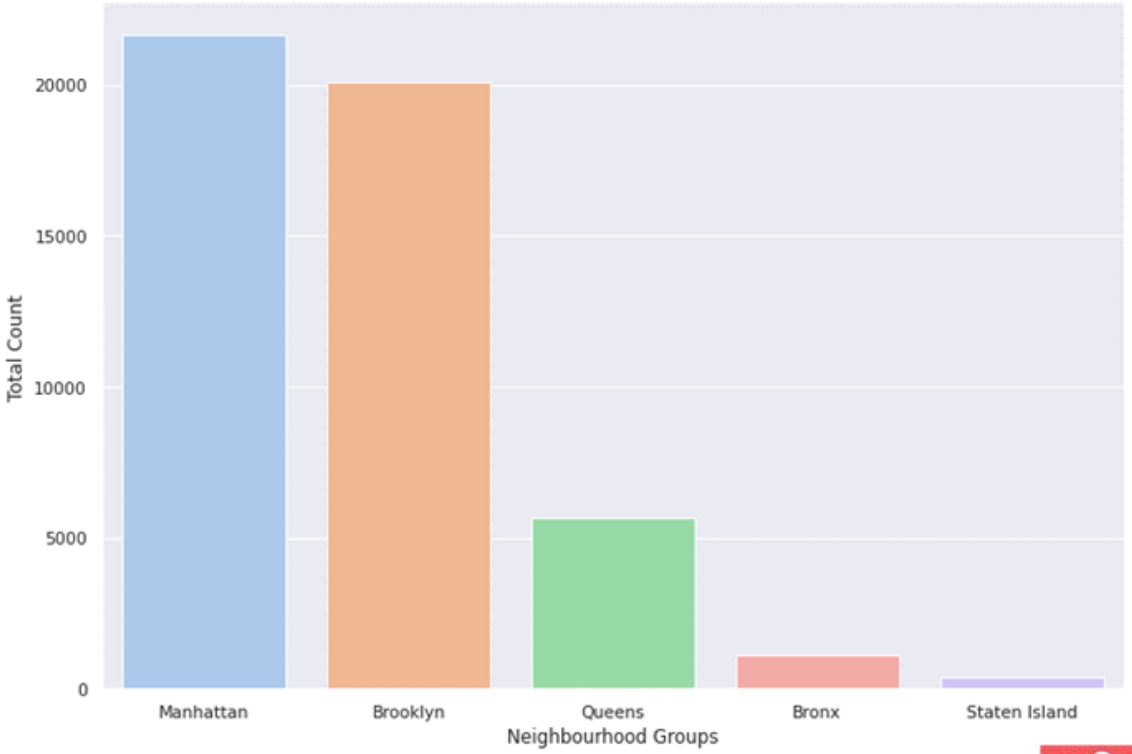
As per analysis :

- **Maximum percentage of neighbourhood group is – Manhattan with 44.3%.**
- **Second share is – Brooklyn with 41.1%.**
- **Queens shares 11.6% while least share is with Bronx with 2.2% and Staten Island with 0.8% only.**



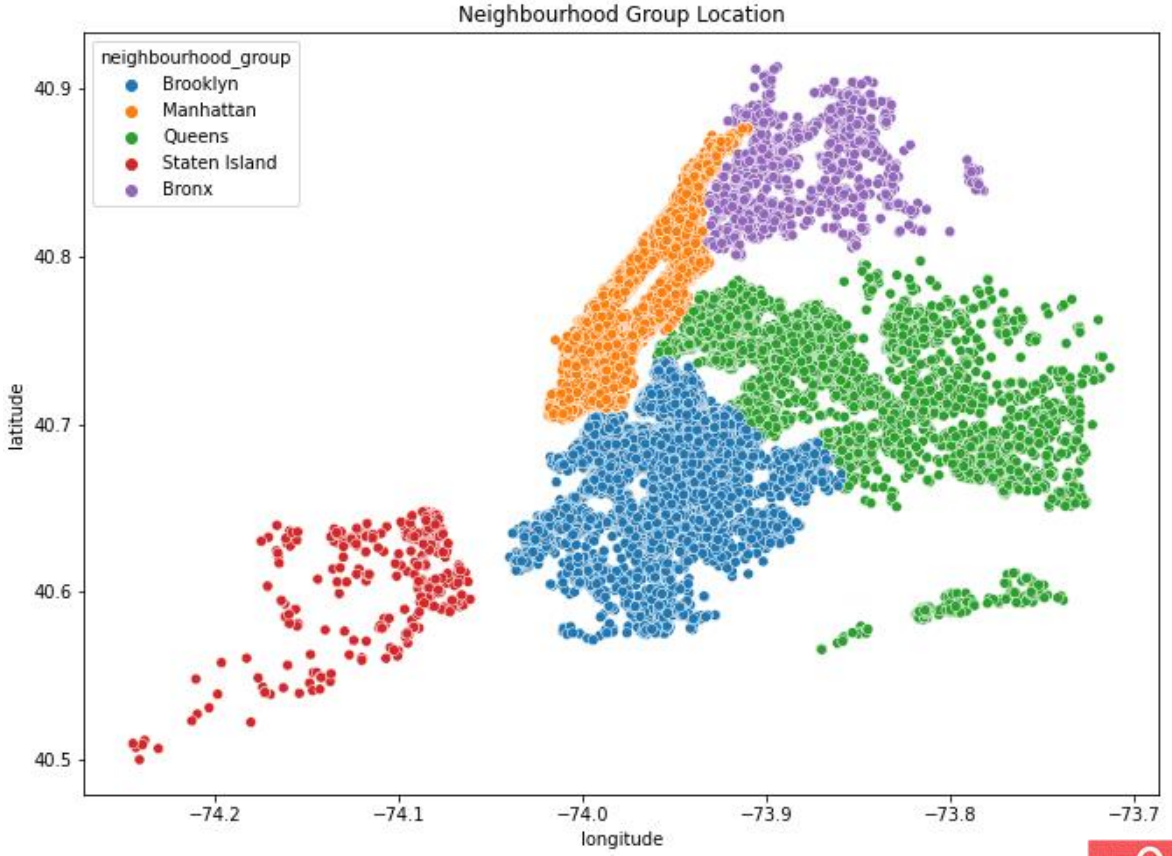
❖ (Cont....) Count based share of neighbourhood groups as per listings

- **Here we used barplot() to plot our graph.**
- **We can see clearly in - count based share the maximum value is again shared by Manhattan.**



❖ (Cont....) Location based neighbourhood groups as per listings

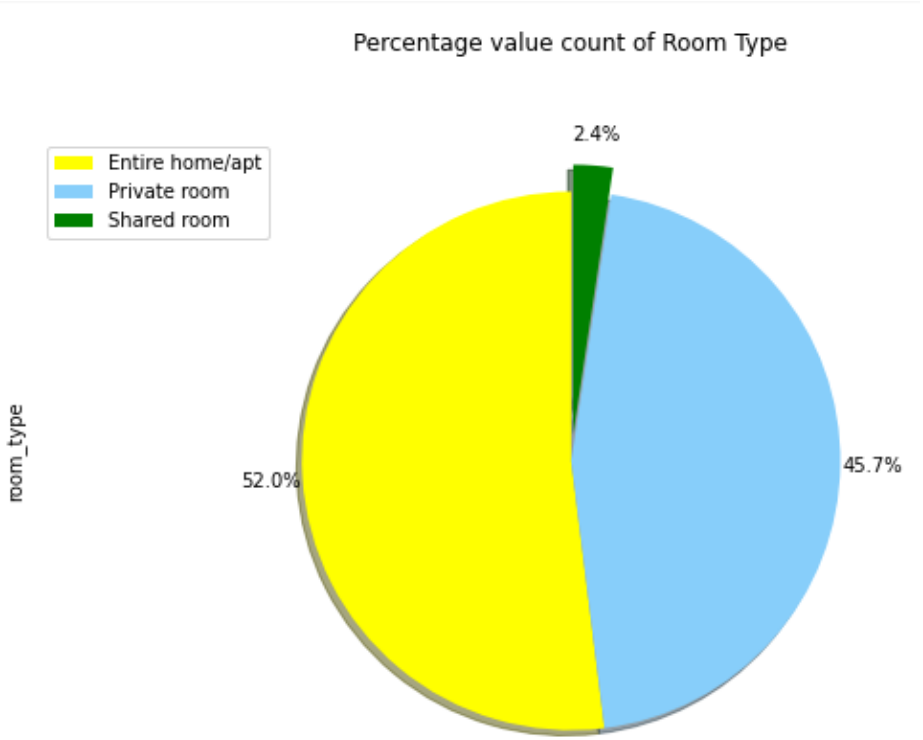
Locating the positions of different “neighbourhood groups” using geospatial coordinates.



❖ (Cont....) Percentage share of different Room types as per listings

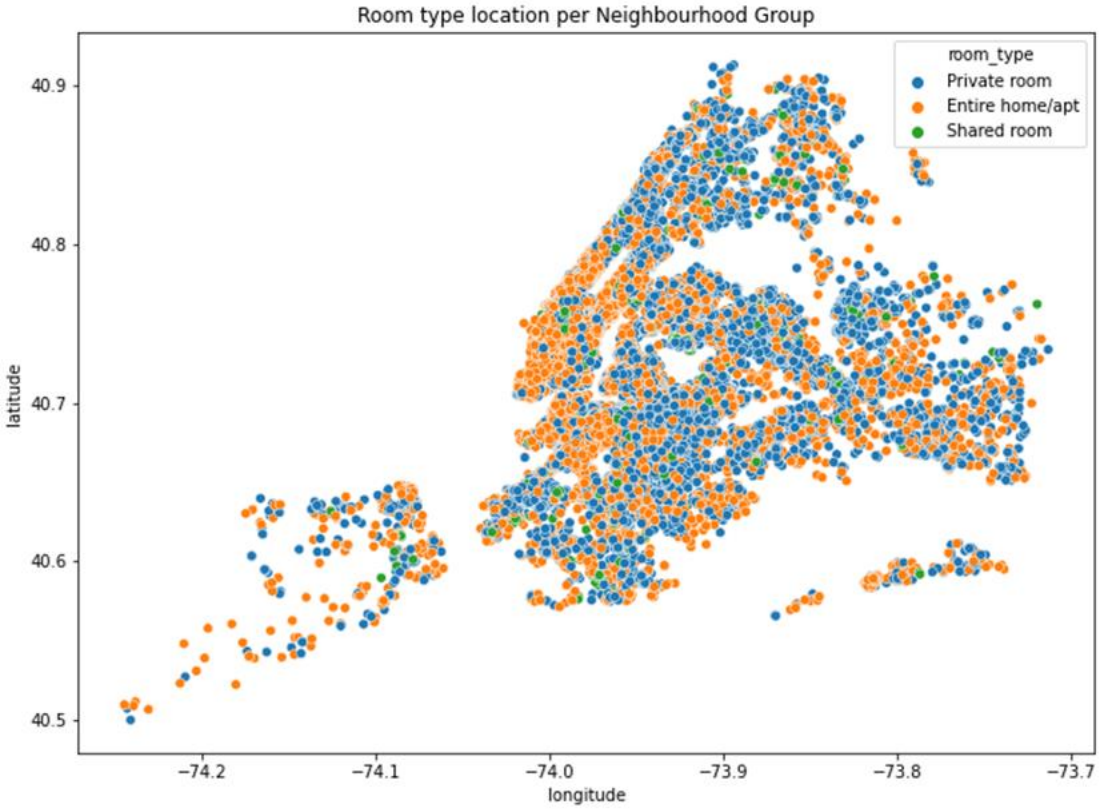
As per analysis :

- **Maximum percentage share of room type is – Entire home/apt with 52.0%.**
- **Second share is – Private room type with 45.7%.**
- **Least share is - Shared room type with 2.4%**



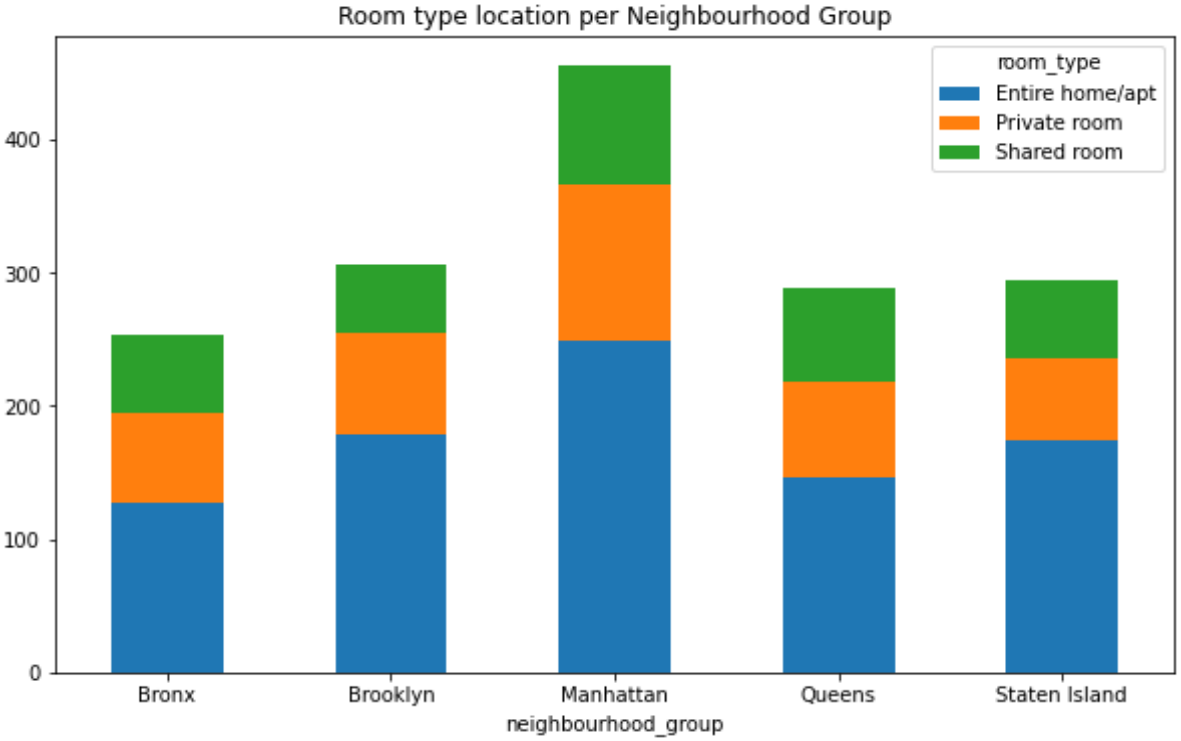
❖ (Cont....) Location based room types as per listings

Locating the positions of different “room types” using geospatial coordinates.



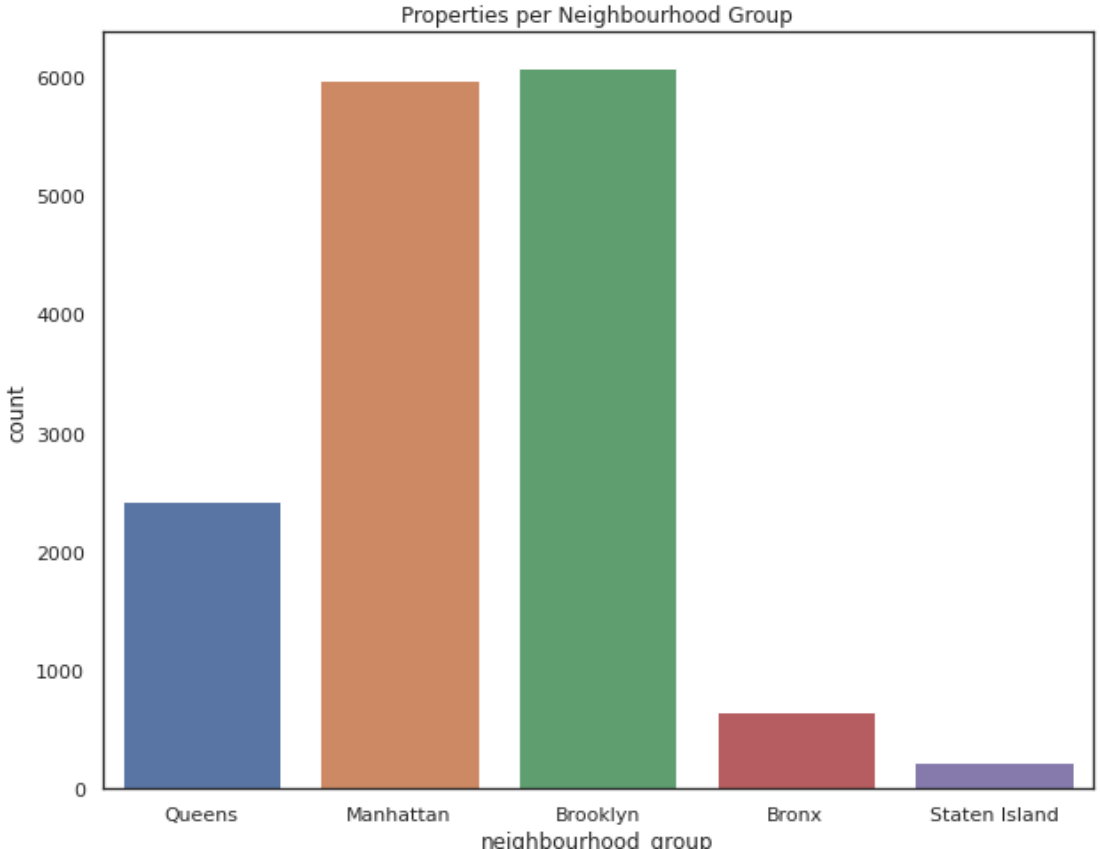
❖ (Cont....) Share among Neighbourhoods and Room types

- **Majority share of Entire home/apt (room type) are located in Manhattan – 249.23, which is further followed by Brooklyn.**
- **Again majority of private rooms are located in Manhattan- 116.80 and in shared room – 88.97.**



room_type	Entire home/apt	Private room	Shared room
neighbourhood_group			
Bronx	127.645503	66.788344	58.610169
Brooklyn	178.362609	76.510619	50.527845
Manhattan	249.238211	116.805594	88.977083
Queens	147.050573	71.776855	69.020202
Staten Island	173.846591	62.292553	57.444444

❖ What is the current status of unique neighbourhood group within columns like host name and host listing?

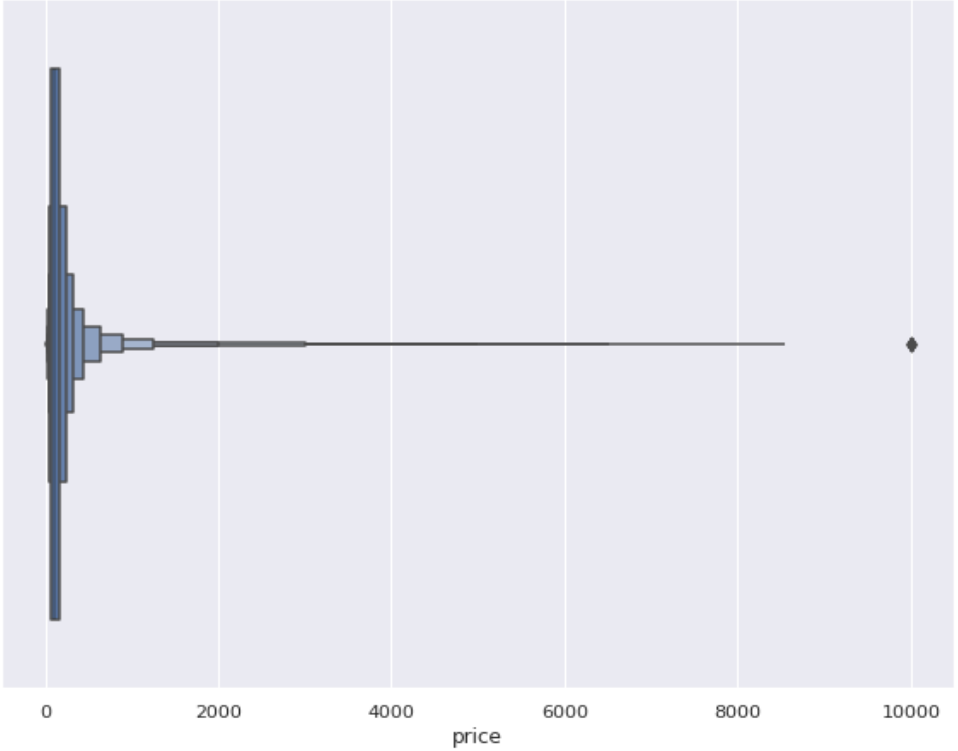


	host_name	neighbourhood_group	calculated_host_listings_count
13214	Sonder (NYC)	Manhattan	327
1833	Blueground	Manhattan	230
9740	Michael	Manhattan	212
3249	David	Manhattan	202
9739	Michael	Brooklyn	159

- **Maximum host listing has been seen at Manhattan which is of host named Sonder (NYC) of total 327.**
- **Overall maximum host listing by hosts seen in Brooklyn. Followed by Queens, Bronx and Staten Island.**

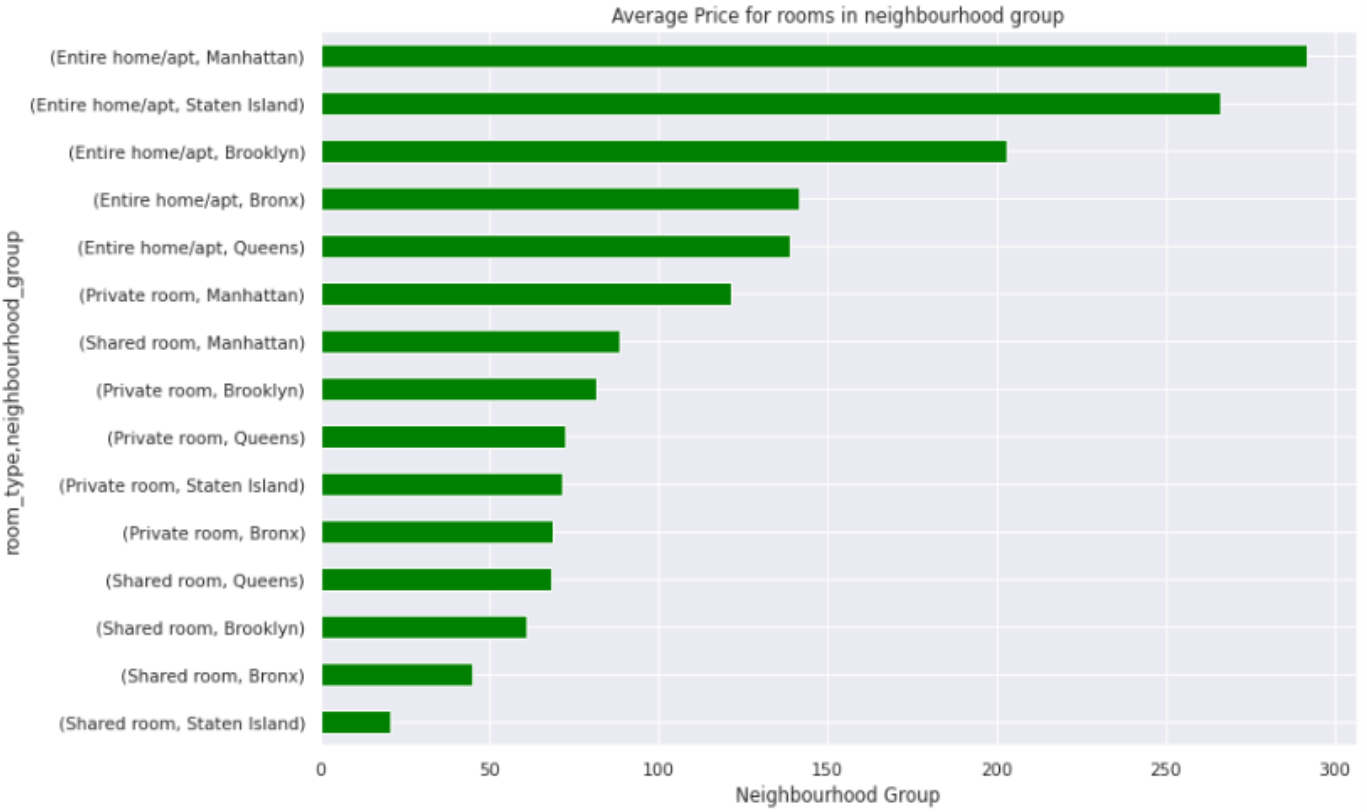
❖ What is the distribution of price in Airbnb dataset?

```
count    48858.000000
mean       152.740309
std        240.232386
min          0.000000
25%         69.000000
50%        106.000000
75%        175.000000
max      10000.000000
Name: price, dtype: float64
```



- The summary statistics clearly shows that the Price ranges from 0– 180. But there also exists price which has a maximum of $10000.
- In our main dataset we have also found some values are 0, which might be due to dynamic pricing or willingness of not to share price with the Airbnb and will be sharing to guest directly during booking.

❖ What is the average room rent and availability of properties in a year at different locality in accordance with minimum nights and price provided within dataset ?



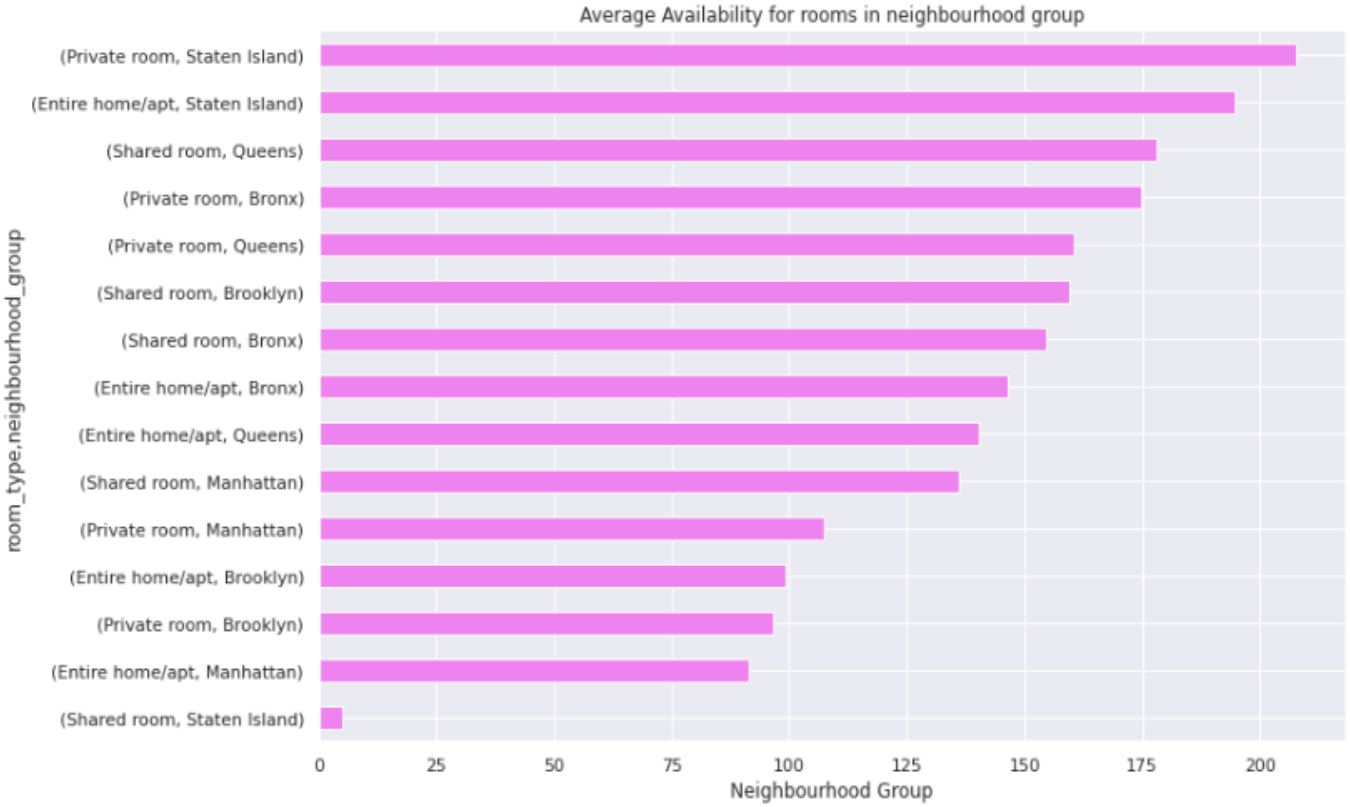
Neighbour

List of Average Price per night based on the neighbourhood group

room_type	neighbourhood_group	price
Entire home/apt	Queens	139.036260
	Bronx	141.541176
	Brooklyn	202.944196
	Staten Island	266.205128
	Manhattan	291.840807
Private room	Bronx	69.025862
	Staten Island	71.394366
	Queens	72.487346
	Brooklyn	81.731334
	Manhattan	121.497409
Shared room	Staten Island	21.000000
	Bronx	44.818182
	Brooklyn	60.921212
	Queens	68.459459
	Manhattan	88.462898

❖ (Cont....)

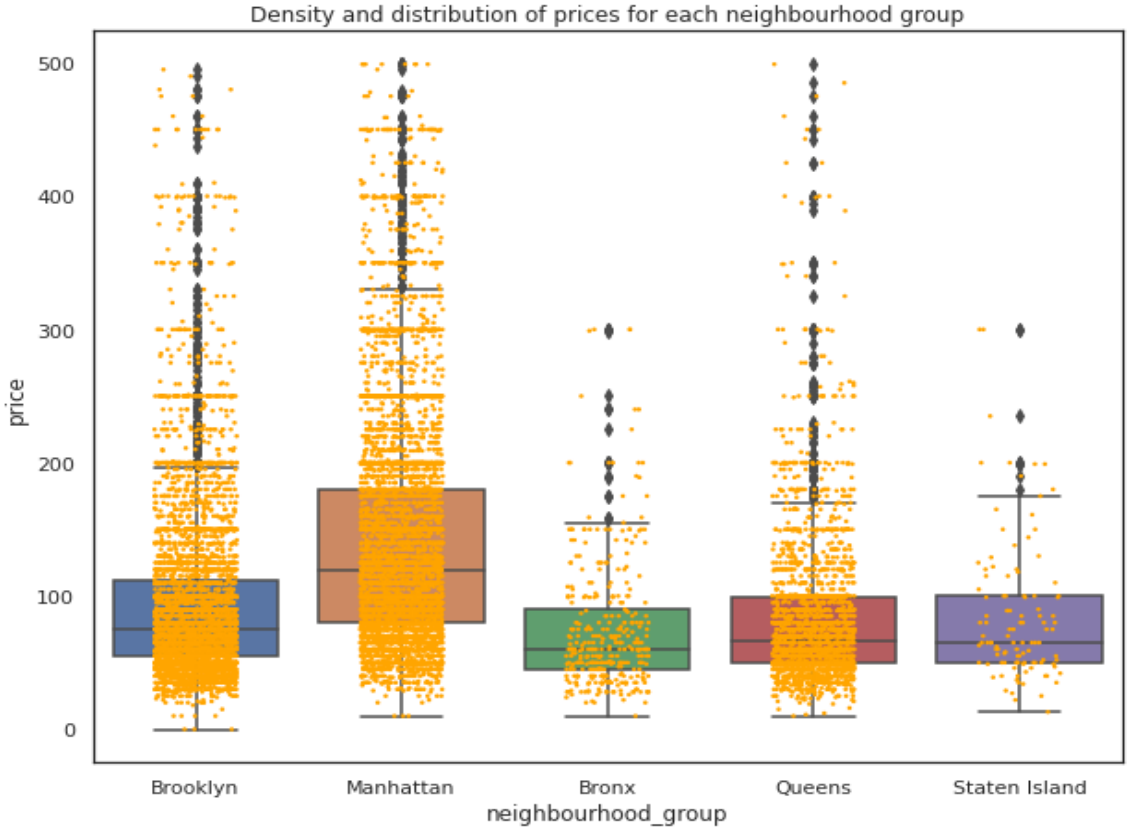
- **Average availability of properties, Staten Island has the least availability of rooms in shared room category among all the neighbourhood in terms of price.**
- **our result also points out that there is less flow of guests in Staten Island in compare to other neighourhood groups.**



List of Average Availability of 365 days of room based on the neig

room_type	neighbourhood_group	availability_365
Entire home/apt	Manhattan	91.535426
	Brooklyn	99.080357
	Queens	140.221374
	Bronx	146.317647
	Staten Island	194.589744
Private room	Brooklyn	96.512763
	Manhattan	107.503282
	Queens	160.376379
	Bronx	174.909483
	Staten Island	207.802817
Shared room	Staten Island	5.000000
	Manhattan	136.159011
	Bronx	154.590909
	Brooklyn	159.612121
	Queens	177.981982

❖ What is the overall density and distribution of price within the different locations of neighborhood group and also showing the price available under $500 ?



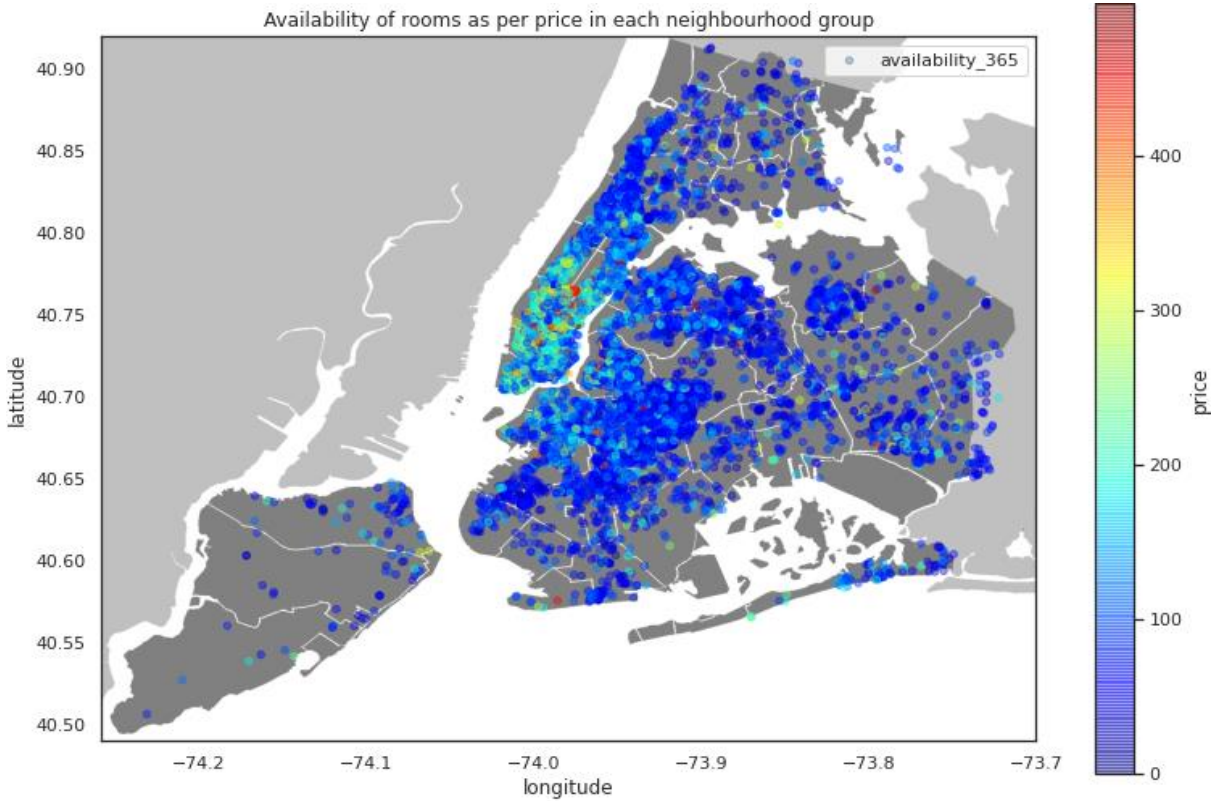
- **Manhattan has the highest range of prices for the listings with $150 price as average observation, followed by Brooklyn with $90 per night.**
- **Queens and Staten Island appear to have very similar distributions, Bronx is the cheapest of them all.**

Stats	Brooklyn	Manhattan	Queens	Staten Island	Bronx
min	0.0	10.0	10.0	13.00	10.0
25%	55.0	81.0	50.0	50.00	45.0
50%	75.0	120.0	68.0	67.50	60.0
75%	119.0	199.0	99.0	106.25	95.0
max	8000.0	7703.0	2000.0	5000.00	1000.0

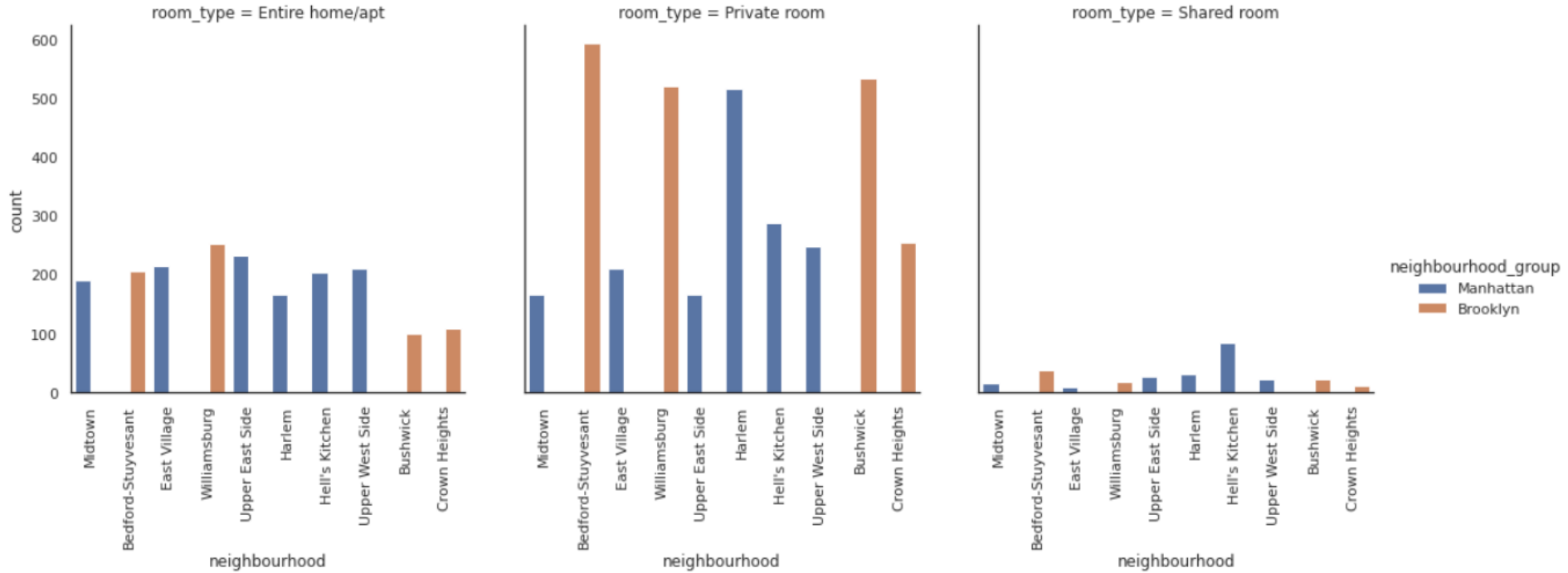
Here we have used Jitter along with our boxplot() and striplot(), which is simply the addition of a small amount of horizontal (or vertical) variability to the data in order to ensure all data points are visible.

❖ Location based price under $500 against the availability of properties in a year

- The properties under $500 can be detected in all 5 neighbourhoods for 365 days of availability.
- But least number of available properties can be found in Staten Island for 365 days.



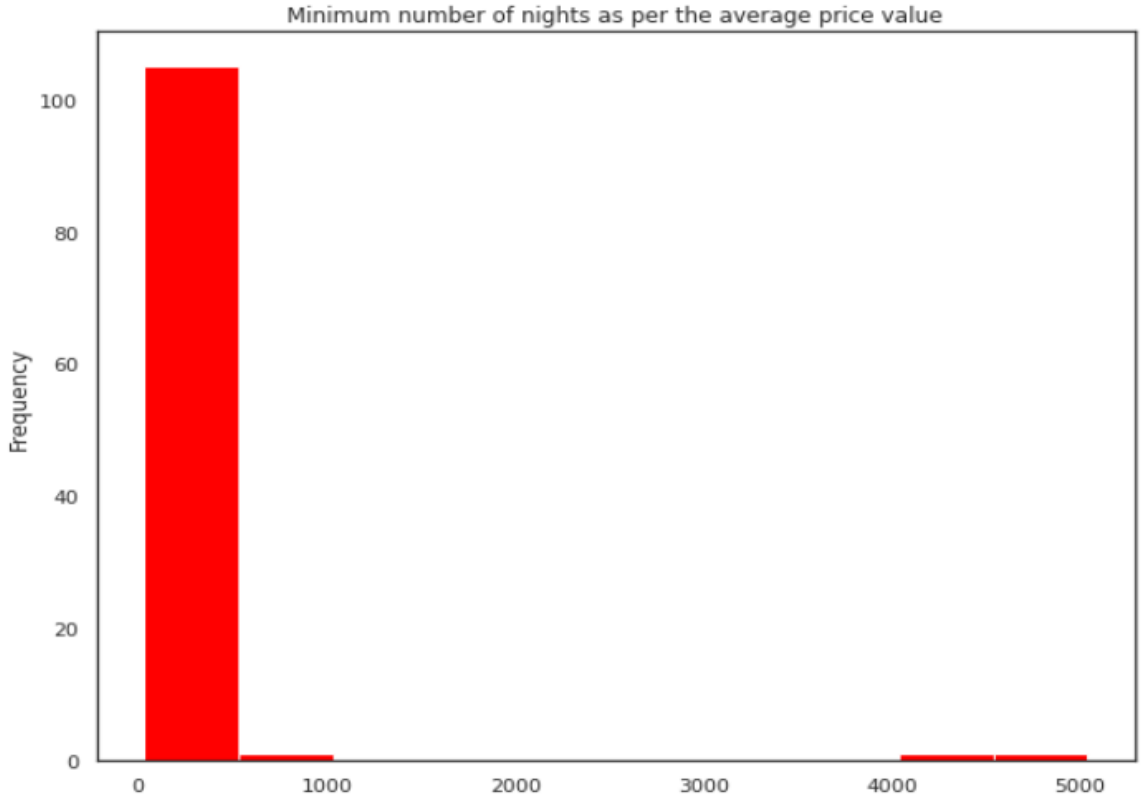
❖ How is neighbourhood, neighbourhood group and room types related to eachother in top locality and can it affect the booking of guest ?



- **Here we took, top 10 neighbourhoods around the major neighbourhood group of Manhattan and Brooklyn.**
- **The catplot() clearly shows that majority of the room booked in the neighbourhood is Private room (short term booking) while for long term booking Entire room/apt is chosen the most while the least is Shared room.**

❖ **What is the trend of minimum nights within the area in respect of average price range ?**

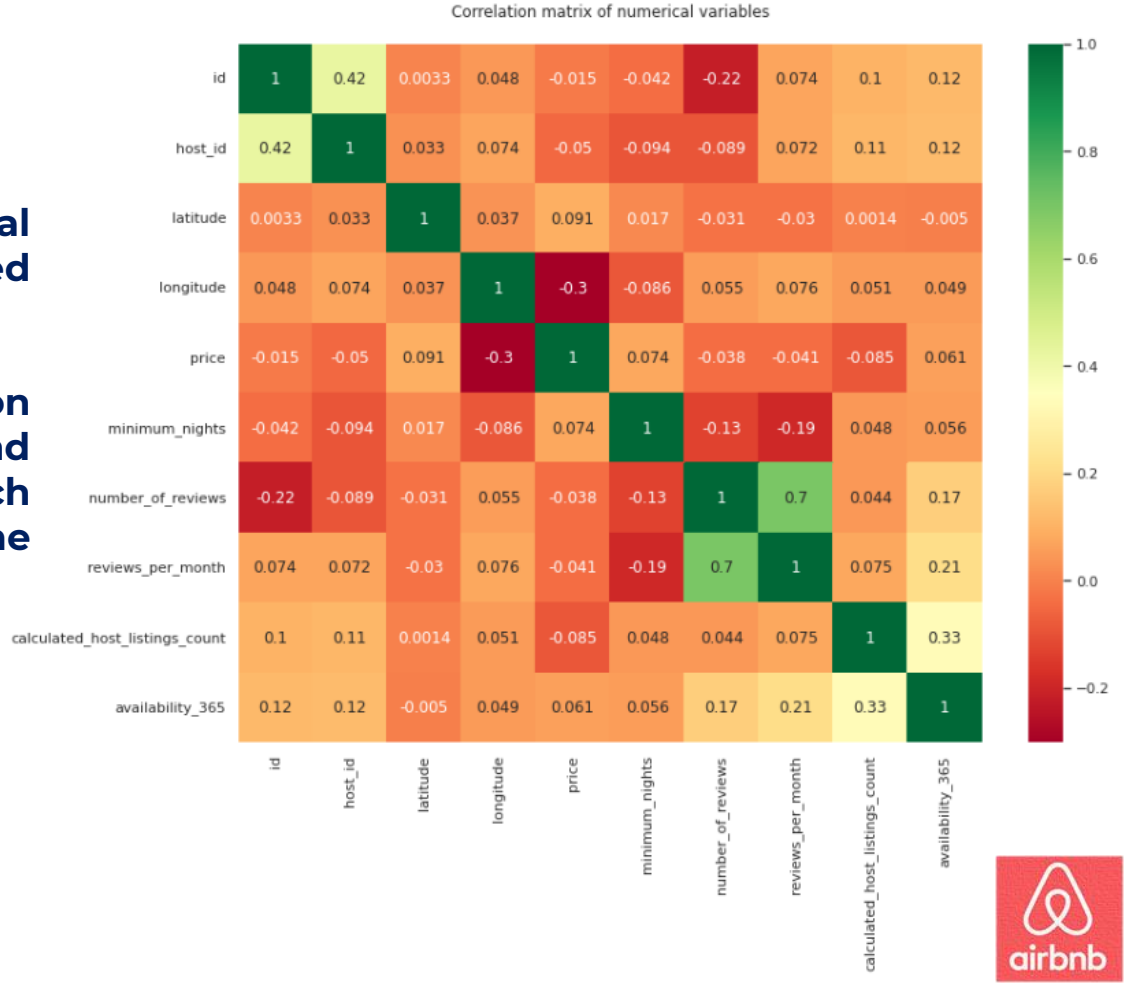
Here we seen that maximum, minimum nights listing found within the range of under 7 days of booking and again values goes up within 30 days of listings.



❖ Correlation

Here the correlation among numerical values of different columns showed us-

- **There is not a strong correlation except review_per_month and number_of_review. Else not much correlation we can get from the present dataset.**

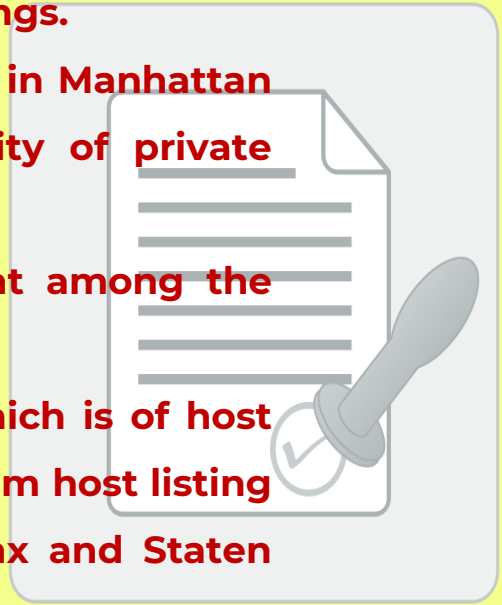


Challenges Faced

- Reading the dataset and understanding of some columns like, calculated_host_listing_count.
- Handling the NaN values, along with some missing values in dataset.
- Understanding the business model and working style of Airbnb from their website.
- Extracting out the latitude and longitude of New York City(NYC), USA using the open street map or open source mapping.
- Putting number of plots as well as location based maps (using the geographical coordinates) to make it more interactive and informative for easily summarizing our outputs to the reader.

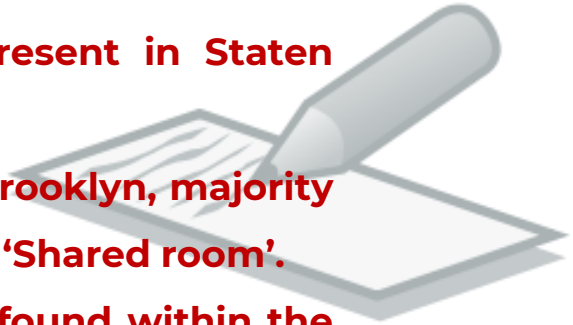
Conclusion

- Manhattan is the most demanding and expensive place for the business according to the Airbnb 2019 dataset for bookings.
- We found that majority of Entire home/apt are located in Manhattan which is further followed by Brooklyn where majority of private rooms are located.
- There were 52.0% coverage of entire home/apartment among the room types.
- Maximum host listing has been seen at Manhattan which is of host named Sonder (NYC) of total 327 but for overall maximum host listing by hosts seen in Brooklyn. Followed by Queens, Bronx and Staten Island.
- The summary statistics clearly shows that the Price ranges from 0–180. But there also exists price which has a maximum of $10000.



Conclusion (Cont...)

- Majority of the price variation among room type in neighbourhoods depends upon number of bookings and engagement of guests for the purpose they are visiting.
- Least and cheapest availability of bookings is present in Staten Island.
- Among top 10 neighbourhoods of Manhattan and Brooklyn, majority of the booked rooms are ‘Entire room/apt’ and least ‘Shared room’.
- There is maximum minimum nights listing can be found within the range of under 7 days of booking and again minimum nights for 30 days.
- There is not a strong correlation except ‘review_per_month’ and ‘number_of_review’. Else not much correlation we can get from the present dataset.




Thank
You!