Capstone Project

Presentation on

Airbnb Booking Analysis

Presented by :
1. Gyana Ranjan Dash
2. Rahul Sharma
3. Anis Bagwan
4. Priyabrata Mohanty

Agenda

- Problem statement
- Objective
- Data summary
- Data cleaning
- Key Findings
- Conclusion
- Recommendations
- Appendix
    - Data Sources
    - Data Assumptions

Problem statement

To analyze the past data of Airbnb and deliver valuable insights to the Company.

Objective

- Improve our shared understanding about the market condition
- Improve our shared understanding about the customer preferences
- Provide early recommendation to the Airbnb company.

Data summary

The dataset span over 2 years - 2018 and 2019

The dataset contains 16 features and 48895 observations

Important features:-

Host_id - unique id given to every host by Airbnb

neighbourhood _group and neighbourhood - Place where the property is listed

Price - price of the listed room

Room_type - The type of room listed i.e Entire home/apt , private room, shared room

Minimum_nights - It is the minimum number of nights the property has to be booked for.

Number of reviews - It indicates the total number of reviews for a particular property.

Data cleaning

The columns like **last_review** and **reviews_per_month** contains more than 20% of null values which needs to treated.

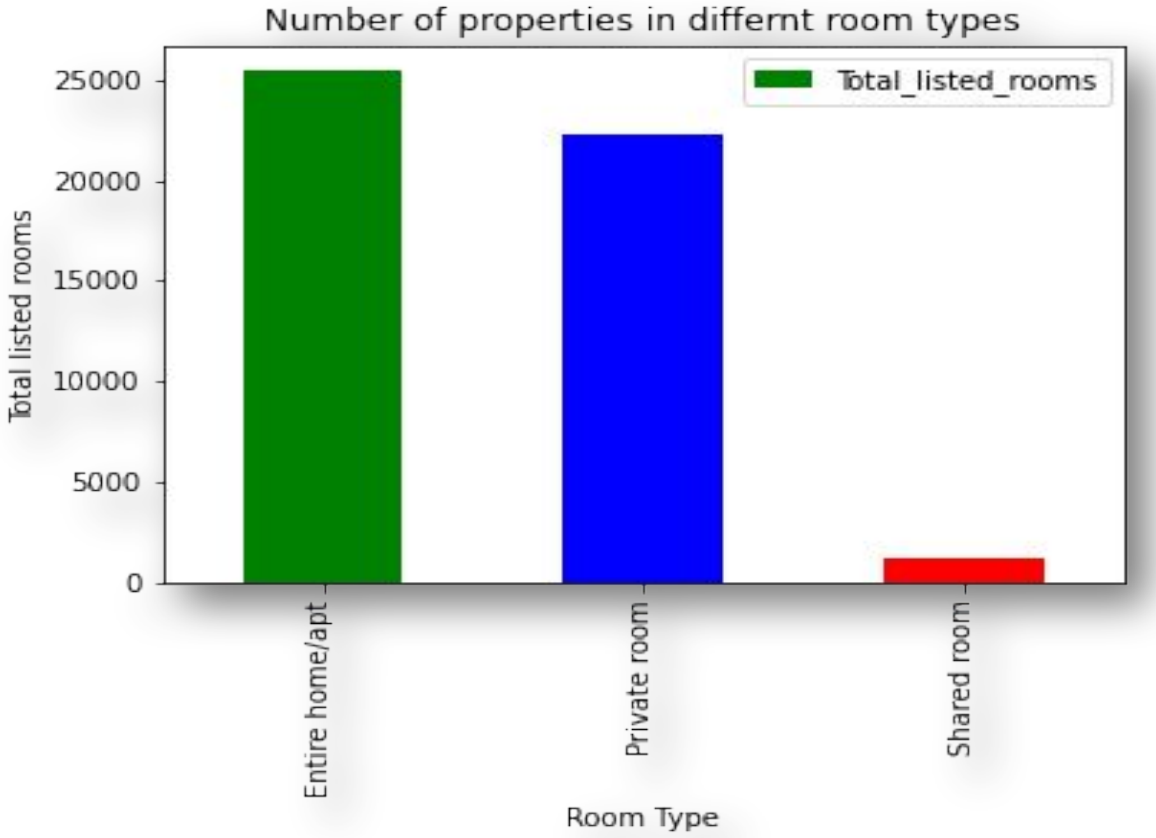
For project the few columns are irrelevant such as **name**, **host_name** and **last_review.**

For the column **reviews_per_month** we can impute the null values with 0.

After looking into **price** column we can see 95% of data lies below $400 but the max in $10000. So we removed the top 5% of the data for better visualization.

Number of properties

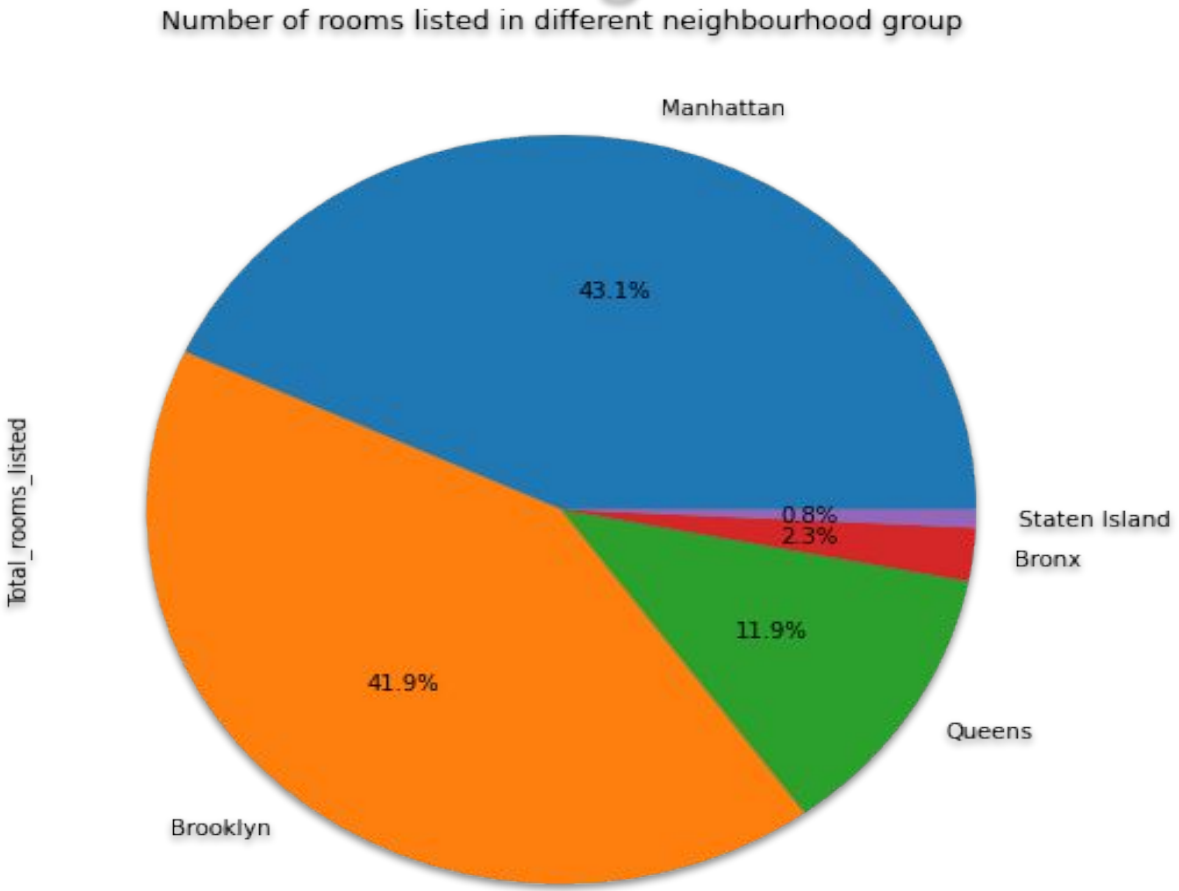
From the above bar plot we can clearly see that **entire home or apartment** is more highest in number where as the number of **shared room** is very less.



From the above pie chart the following instances we can conclude that :

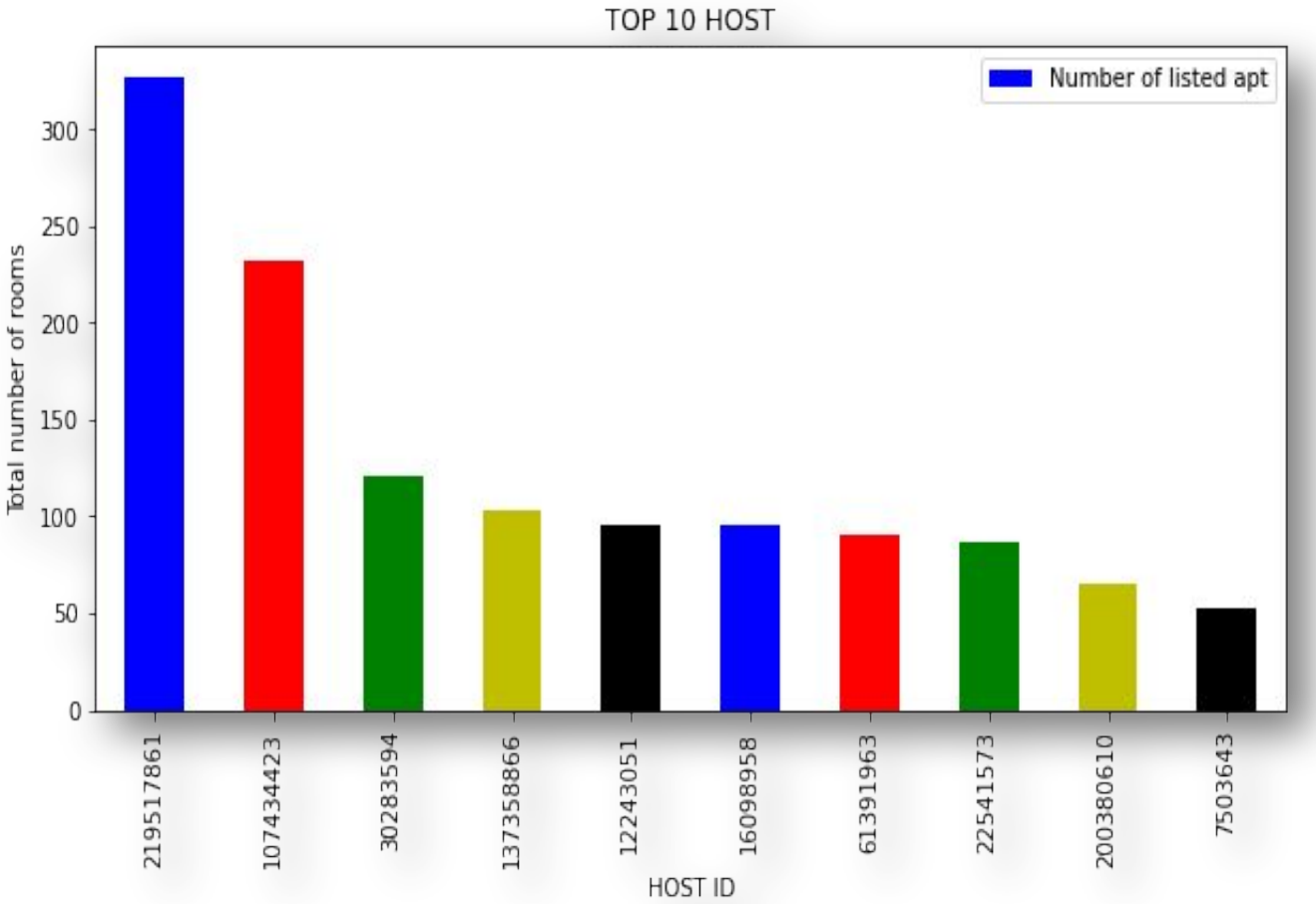
- **Manhattan** and **Brooklyn** have most number of rooms, together have **more than 85%** of total rooms available.
- **Manhattan** has highest number of rooms which is **more than 44%.**
- **Staten Island** has lowest number of rooms which is **less than 10%.**

Rooms in Neighbourhood



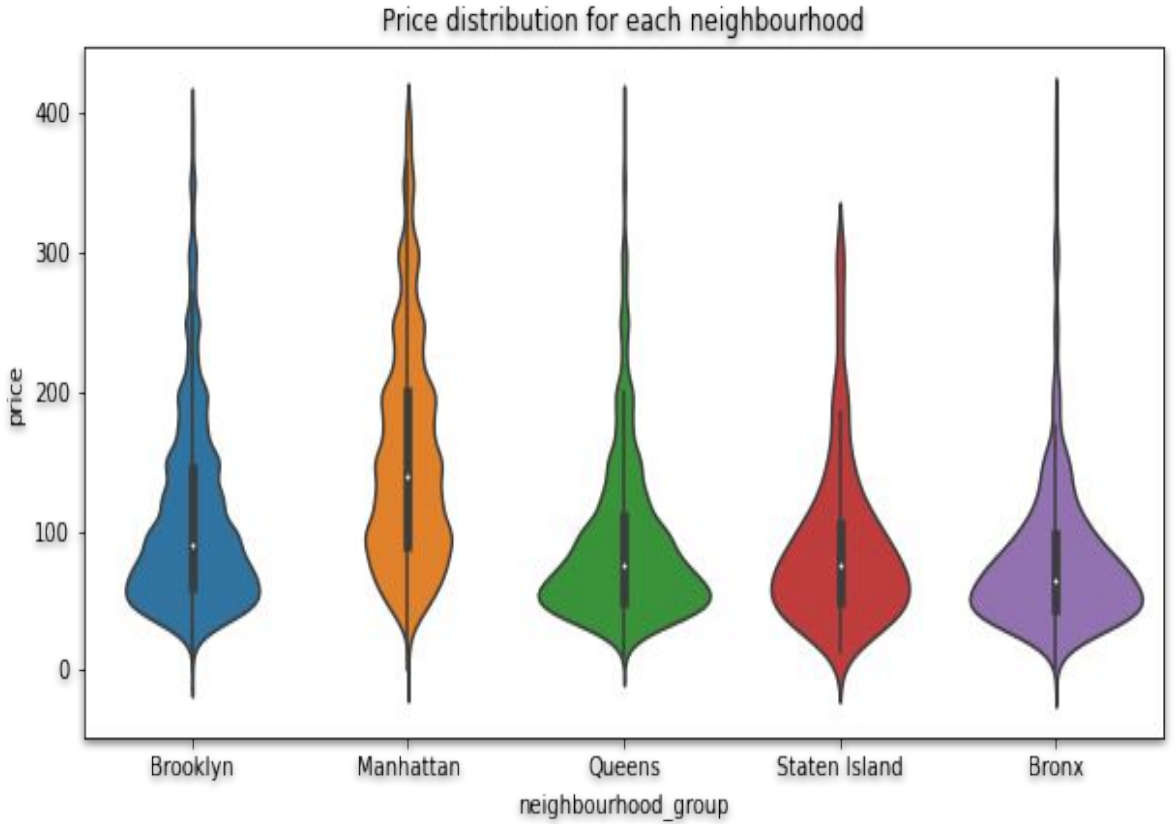
Top 10 Hosts

- Following bar plot shows the **hosts** with **maximum number of rooms** listed on their name.
- **X-axis** shows the **individual ID's of hosts.**
- **Y-axis** shows the **number of rooms listed** for corresponding **host.**



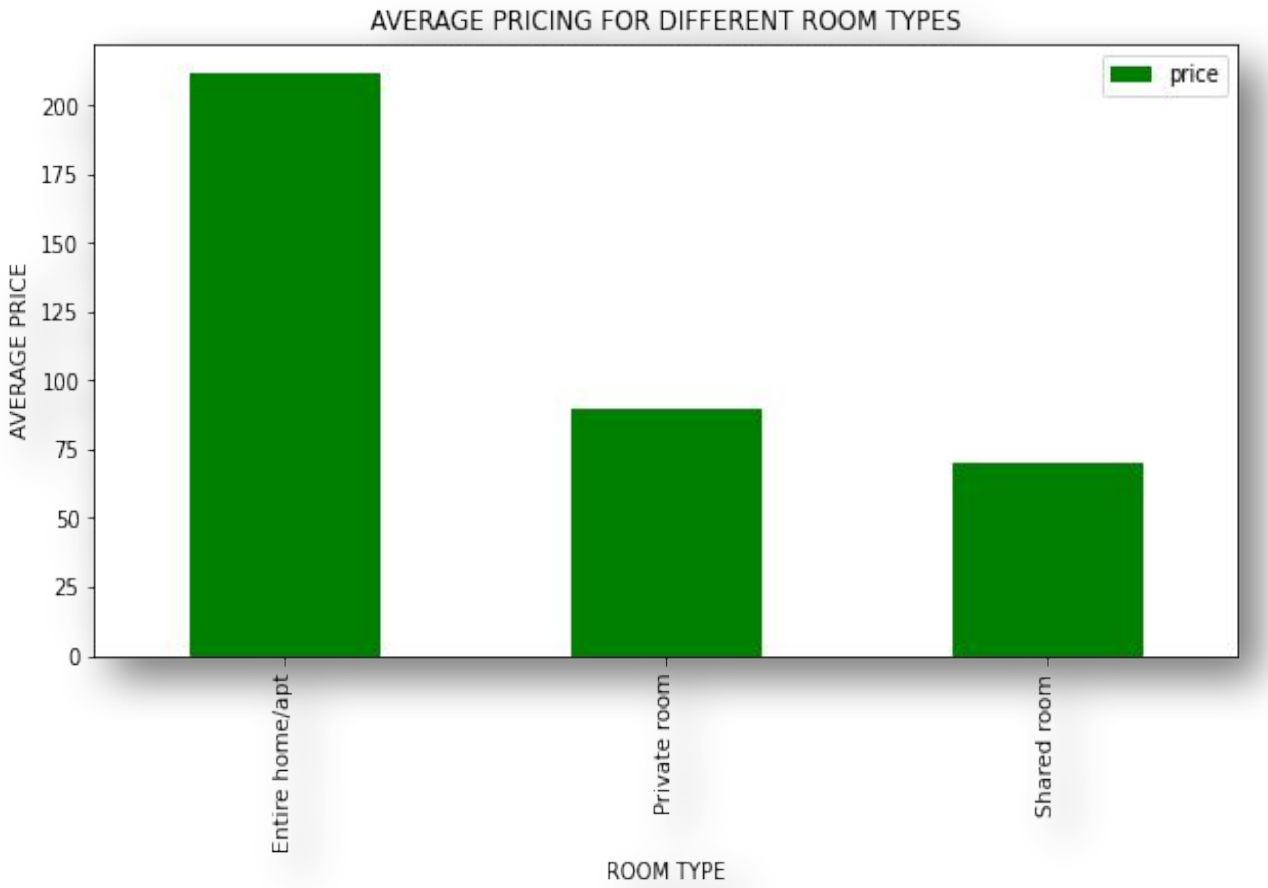
Price Distribution

- In the above figure we can see the **distribution of prices of properties** in different neighbourhood group.
- We can clearly see that **manhattan have more premium properties.**
- **Bronx , staten island and queens** have **reasonable prices** as compared to the manhattan and brooklyn.



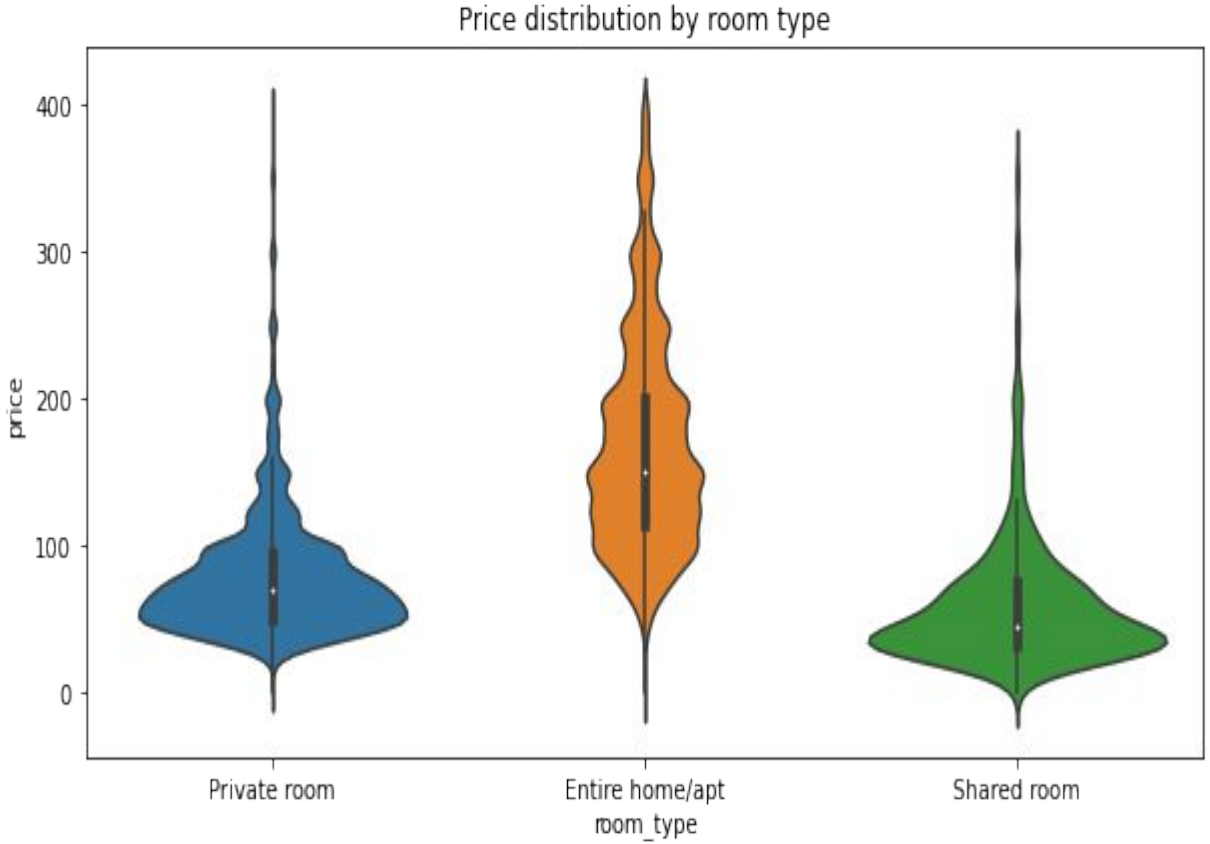
Average prices of rooms

- We can clearly see the average price of **entire home or apartment** is much **highe**r than any other room types.
- The average price of **Shared room** is very **low.**
- It seems obvious as the entire home or apartment tend to have higher pricing than a shared room.



Price Distribution by room type

- As expected, **shared rooms** have **the lowest mean price,** while **entire homes** have the **highest.**
- All room types seem to have a similar spread, however **private rooms** and **shared rooms** seemed to be more **centered around their mean.**

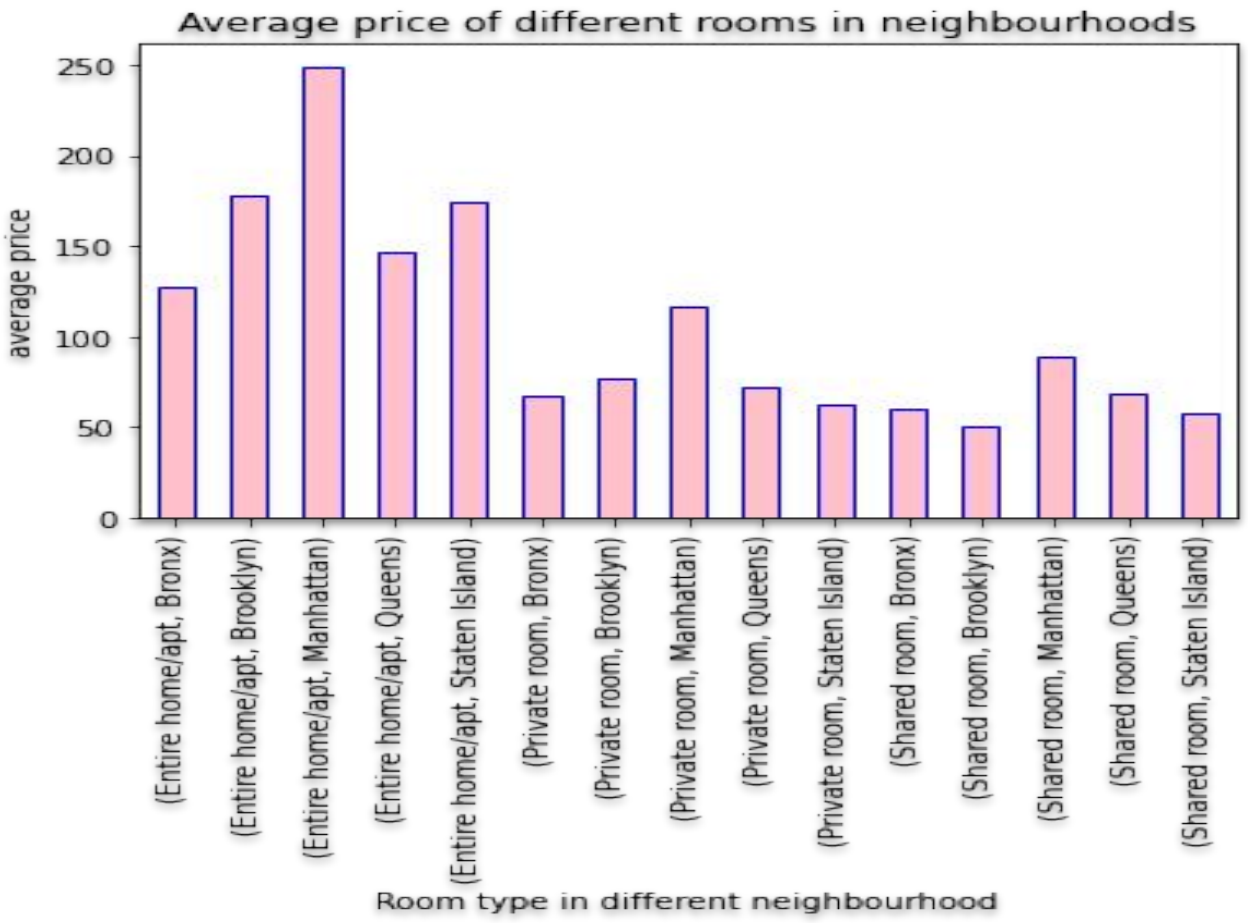


Expensive place to rent

From the lateral plot we can say :

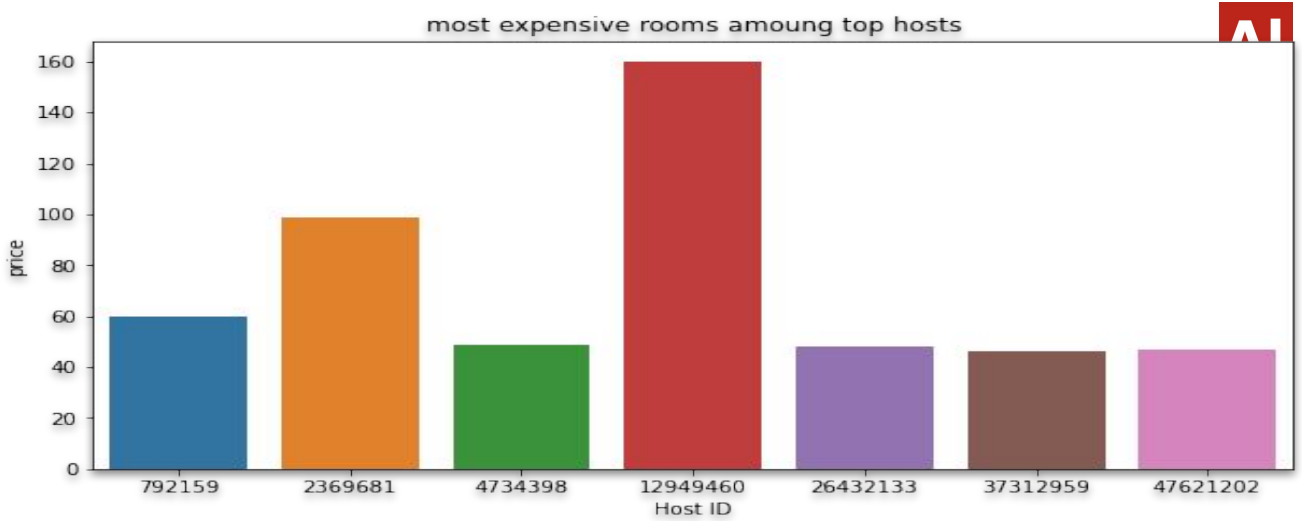
- The average price of Entire room/apt in Manhattan is higher than any other in neighborhood
- The average price of shared room in Manhattan is higher than any other neighborhoods.
- The average price of shared room in Manhattan is higher than any other neighborhood.

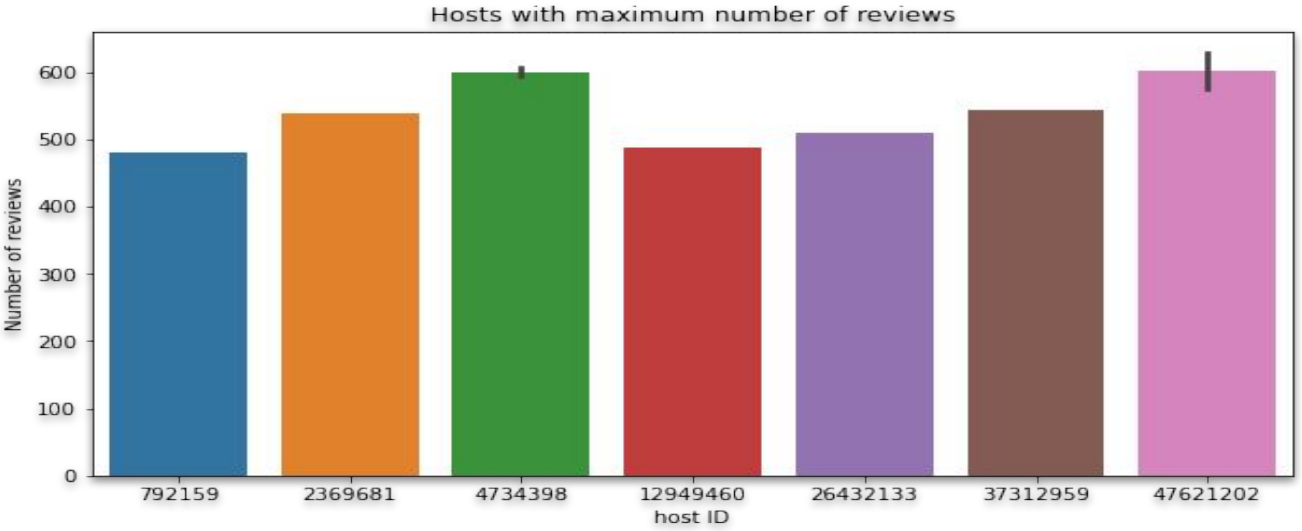
From the above insights we can conclude that the **Manhattan is a expensive place to rent property** in any category.



Rooms and Reviews

- The host with **Host ID 12949460** list's **most expensive** room but the same ID have **less no of reviews** compared to other reviews.
- The hosts which are **not that much expensive** have maximum number of reviews.

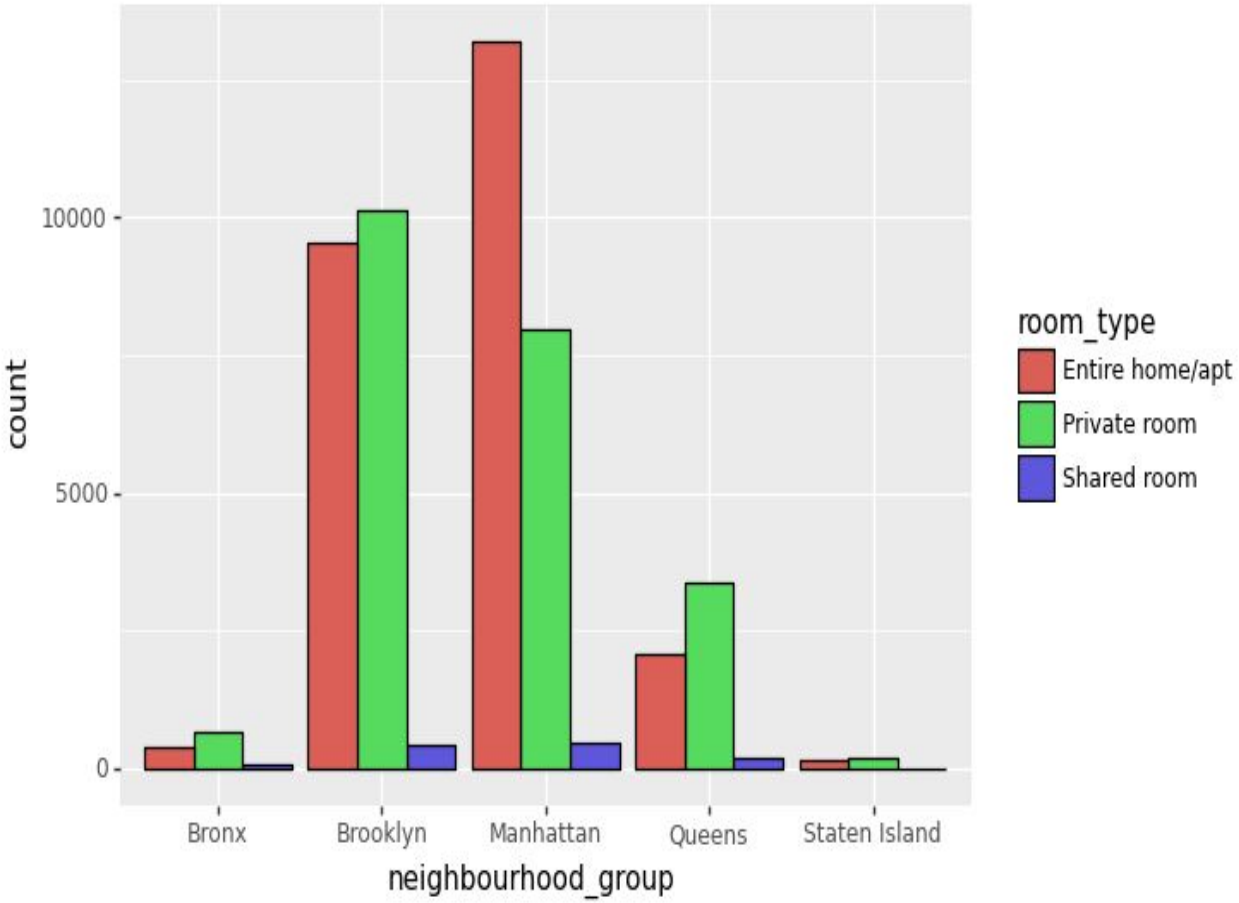




Most preferable room types

From the above plot we can draw following insights :

- In Manhattan the number of listed Entire home/apt is much higher than any other room type.
- In Brooklyn the number of Private Room is slightly higher than Entire home/apt.
- In Queens the total number of listed Private Rooms is much higher than the any other room type.
- In Bronx and Staten Island the number of rooms in all the category are almost same.

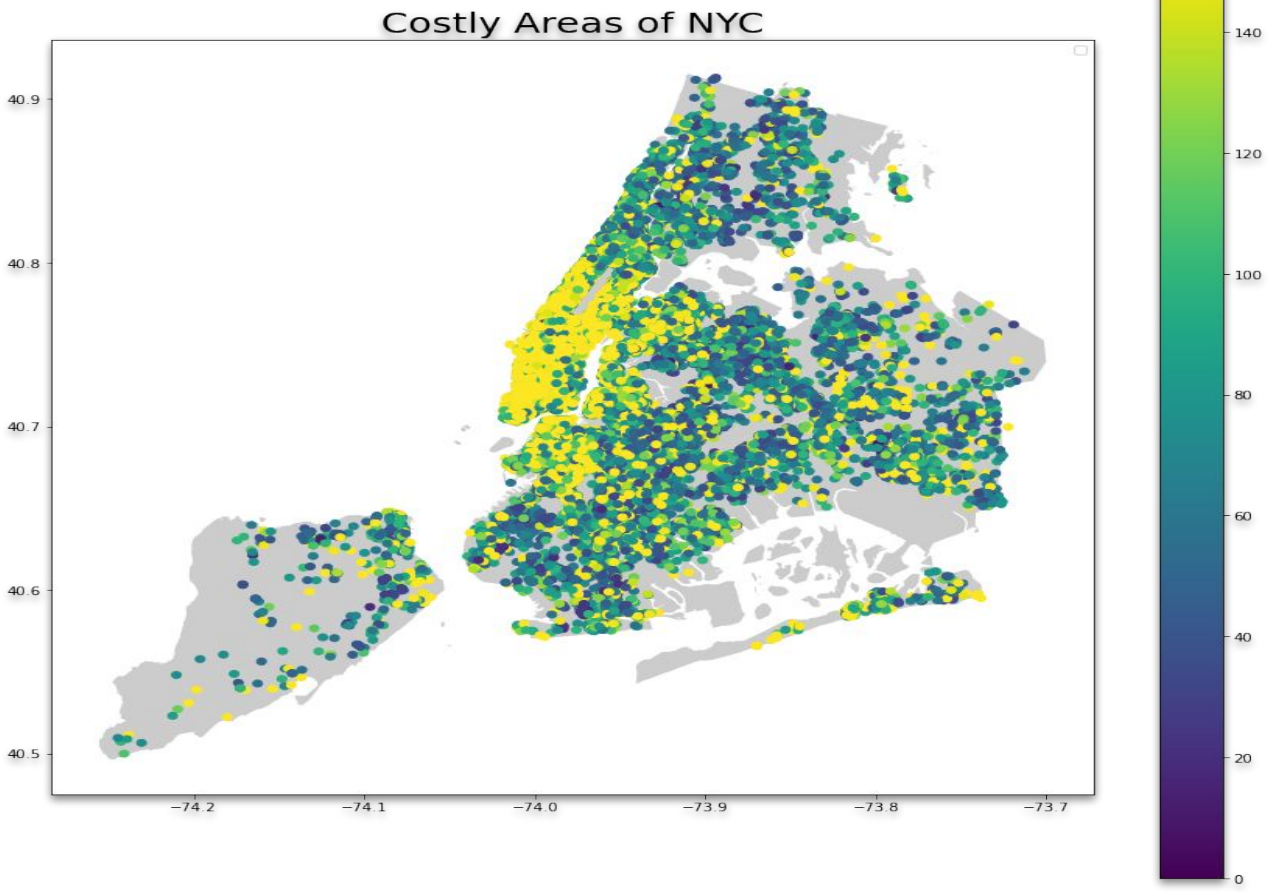


From the above insights we can conclude that the people prefer to live in Entire home/apt in Manhattan.

Costly areas

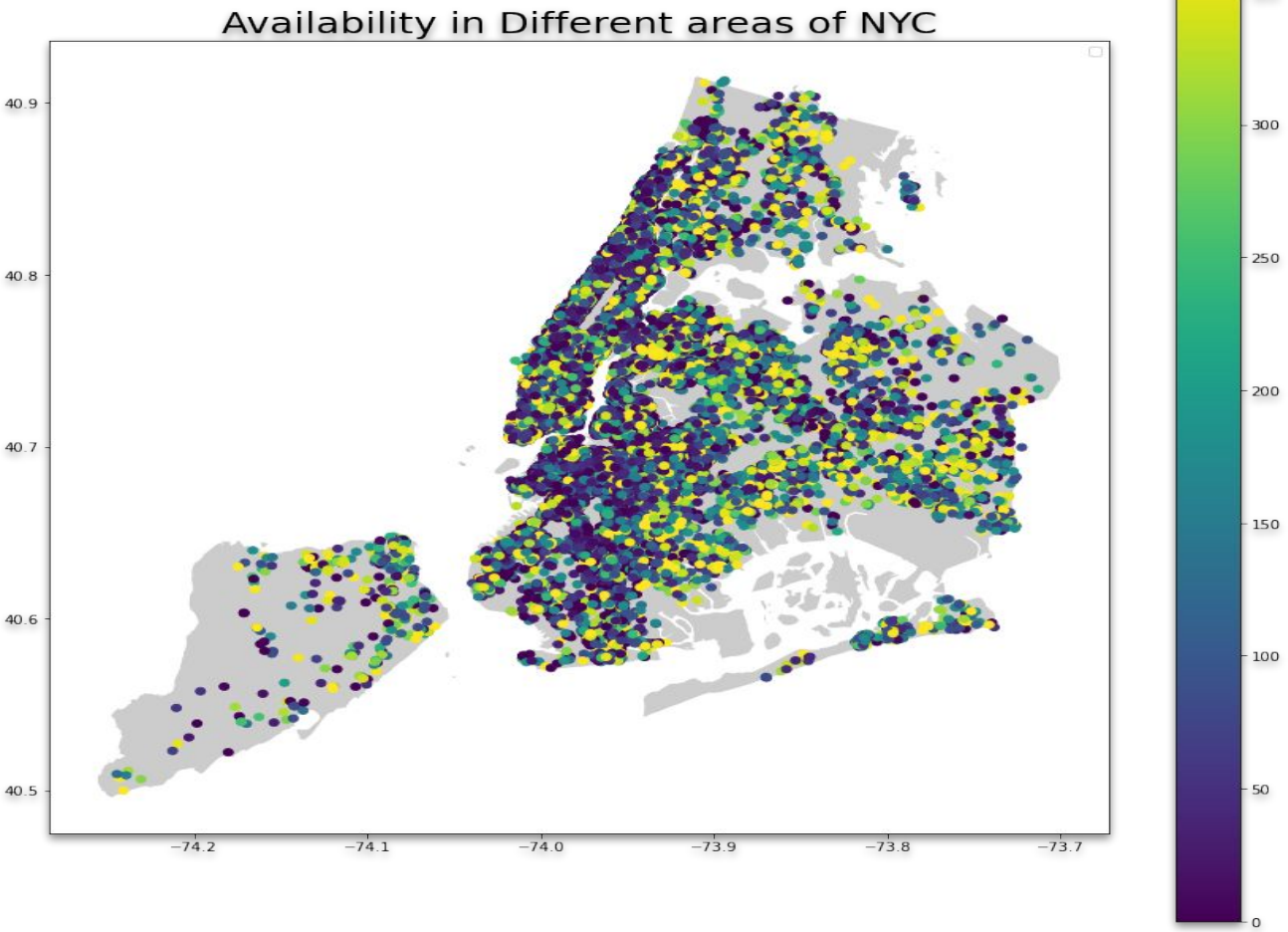
- **Yellow dots** represents the places where the **properties are costly.**
- **Blue dots** represents the places where the **properties are comparatively cheaper.**

The place where we can see high concentration of yellow dots is the most costly region in NYC



Availability in areas of NYC

- When we compare two graphs we can say that most costly places are also a cluster of considerable number of rooms availabilities as well.
- The empty spots on the maps are the opportunities to expand and enroll new hosts from there.



Conclusion

- Manhattan and Brooklyn are the prime locations with highest number of rooms
- Manhattan is the most expensive place to rent a room
- Customers prefer Entire home/ apt more
- Properties which are expensive tend to attract less number of customers.

Recommendations

- Increase the listing of private rooms in Queens and Manhattan.
- Try to keep Minimum number of nights below 7 days.
- Keep avg price of the private rooms in Queens and Manhattan low.

Appendix - Data sources

- Here is a snapshot of data dictionary.
  - Property information such as
    - Name, neighbourhood, price, availability and reviews
  - Host information such as
    - Host ID, host name

- We used the past dataset from 2018 and 2019

Appendix - Data Methodology

- We conducted a thorough analysis of the Airbnb-NYC dataset. The process includes
  - **Data cleaning** - Discard irrelevant columns and null imputation
  - **Correlation** -Finding correlation between the different parameters
  - **EDA** - Understanding data using different methods