Capstone Project

Airbnb Bookings Analysis

By

Team :- Team Denver

Team Member :- Sumit , Bhoomika

Data Science Student

Alma Better

What we are discuss :-

- About the Airbnb
- Problem statement
- Data Exploration
- Data Cleaning
- Host and neighborhood group
- Location of neighborhood group
- Price distribution across neighborhood
- Popular neighborhood by review
- Preferred room types
- Limitation
- Scope of Improvement
- Conclusion

1. Airbnb, Inc. based in San Francisco, California, operates an online marketplace focused on short-term homestays and experiences. The company acts as a broker and charges a commission from each booking. The company was founded in 2008 by Brian Chesky, Nathan Blecharczyk, and Joe Gebbia. Airbnb is a shortened version of its original name, AirBedandBreakfast.com. The company has been the subject of criticism for lack of regulations and enabling increases in home rents.

2. The booking information on Airbnb from 2008 till 2019

Problem Statement :

With help of python data visualization , libraries . We will try to solve the answer of the following questions

- What can we learn about the different hosts and areas?
- What can we learn from prediction’s (prices , reviews etc.)
- Which type of rooms are customer demands in most popular neighborhood, neighborhood_group
- Why type of reviews are made by the most of costumer’s
- Limitation in Airbnb’s data
- Scope of Improvement (How we help to resolve the problem)

Data Exploration (variable name):

The data and its features are given below :-

- **Id**
  It is a particular identity number of property which is given to customer(host)
- **Name**
  It gives the name of property given to customer
- **Host_id**
  It is the identity number of host who have register on airbnb
- **Host_name**
  It is the customer name who registered their property on Airbnb
- **Neighbourhood_group**
  It tell the neighborhood group present in the particular city (ex :- NYK , San Francisco etc.)
- **Neighborhood**
  It tells the neighborhood present in neighbourhood_group in the city.

Data Exploration (variable name):

- **Neighbourhood_group**
  it gives a group of area
- **Latitude**
  it gives the coordinate of listing
- **Longitude**
  it gives the coordinate of listing
- **Room_type**
  it tells the type to categorize the rooms
- **Price**
  it gives the price of rooms according to room_type
- **Minimum_nights**
  It gives the info about minimum nights required to stay in a single visit.

Data Exploration (variable name):

- **Number_of_reviews**
  total count of reviews given by visitors.
- **Last_review**
  date of last review given.
- **Reviews_per_month**

  it gives rate of reviews given per month
- **Calculated_host_listings_count**
  it gives total no of listing registered under the host
- **Availability_365**
  it gives the number of days for which a host is available in a year.

.

Data Cleaning :-

Fixing the null values

We have drop the unnecessary null values like numbers_of_reviews, last review ,longitude, latitude and reviews_per_month (because it has not much meaning full values) .

```
[22]  1 df.isna().sum() # checking null values in data
```

```
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
```

Before Cleaning the Data

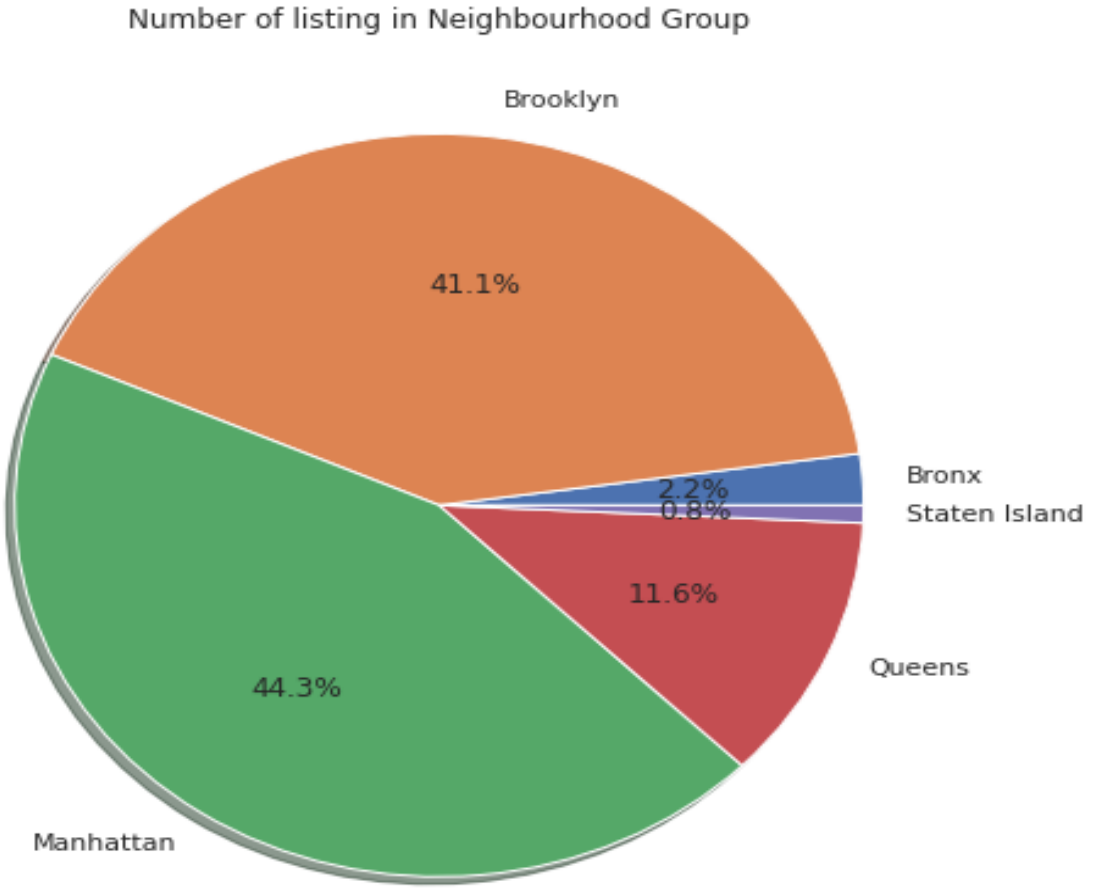
```
1 df.isna().sum() # again check total number
```

```
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
calculated_host_listings_count      0
availability_365                    0
dtype: int64
```

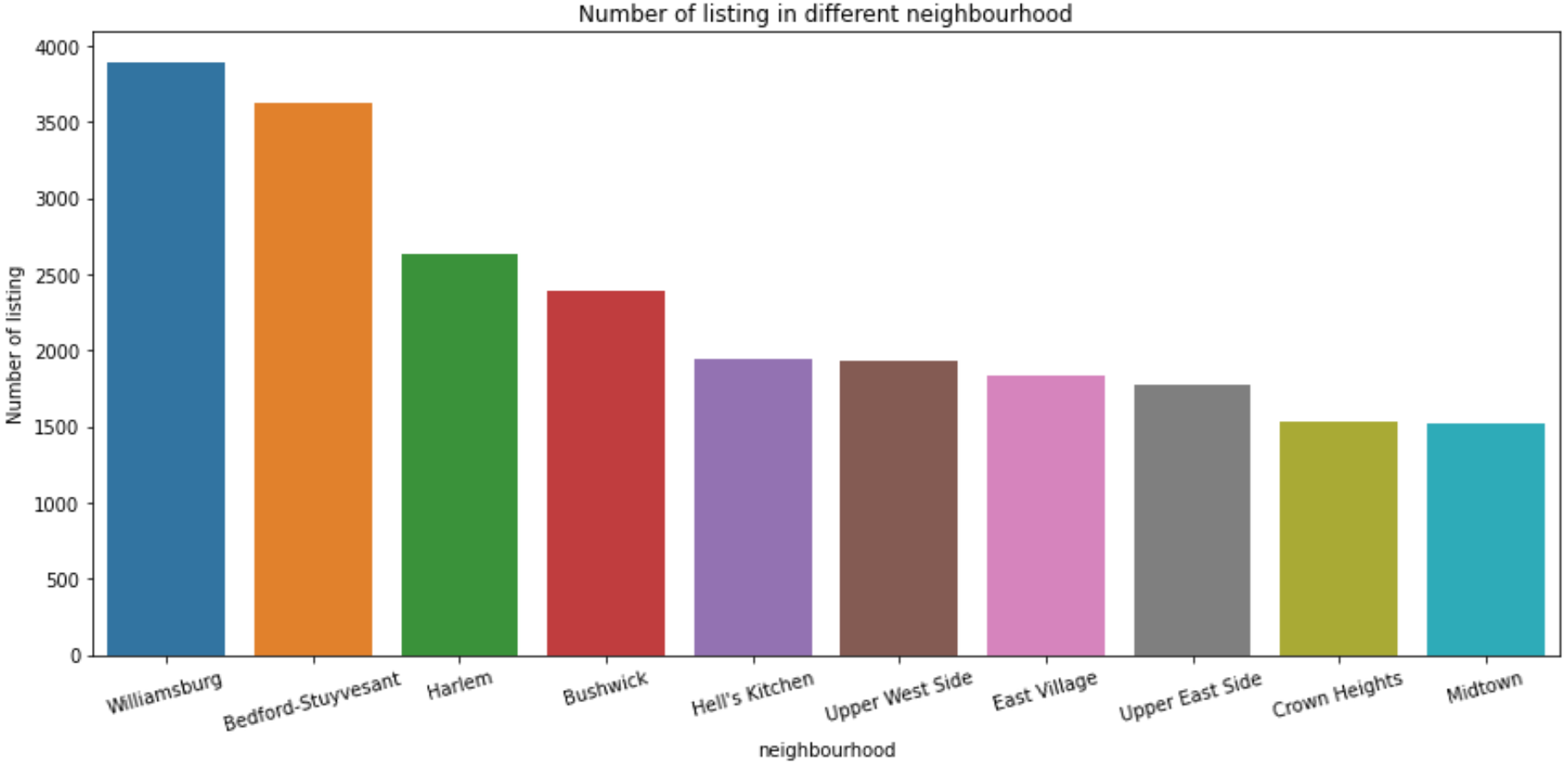
After Cleaning the Data

No of list made by host across Neighborhood Group

- It is observed that Manhattan has highest number of listing of 21661 which is 44.3% of total listing.
- Brooklyn has second highest number of listing 20104 which is 41.1% of total listing.
- Queens are at third place with 5666 listing and Bronx and Staten have least number of listing.

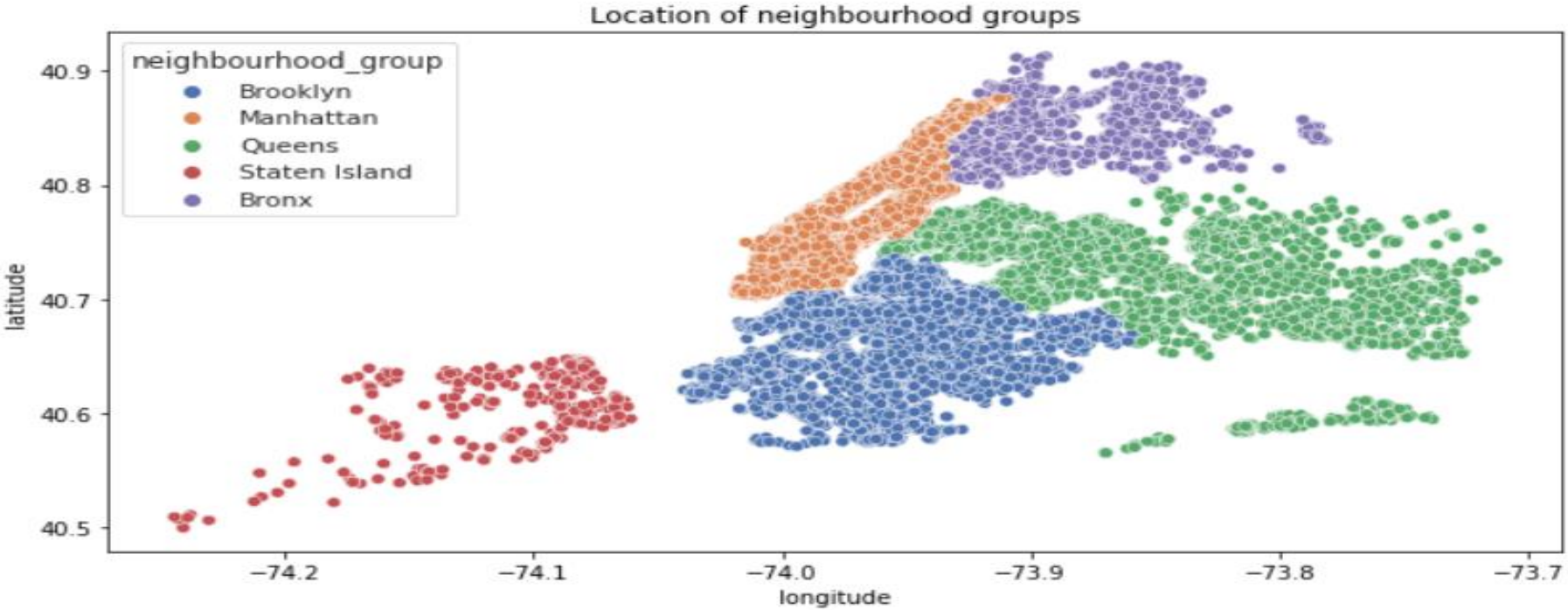


No of list made by host across Neighborhood



- **Most of peoples wants to live in Williamsburg and Bedford-Stuyvesant.**

Location of Neighborhood Groups




```
1 # it tells the unique value in neighbou
2 len(df['neighbourhood'].unique())
```

```
221
```

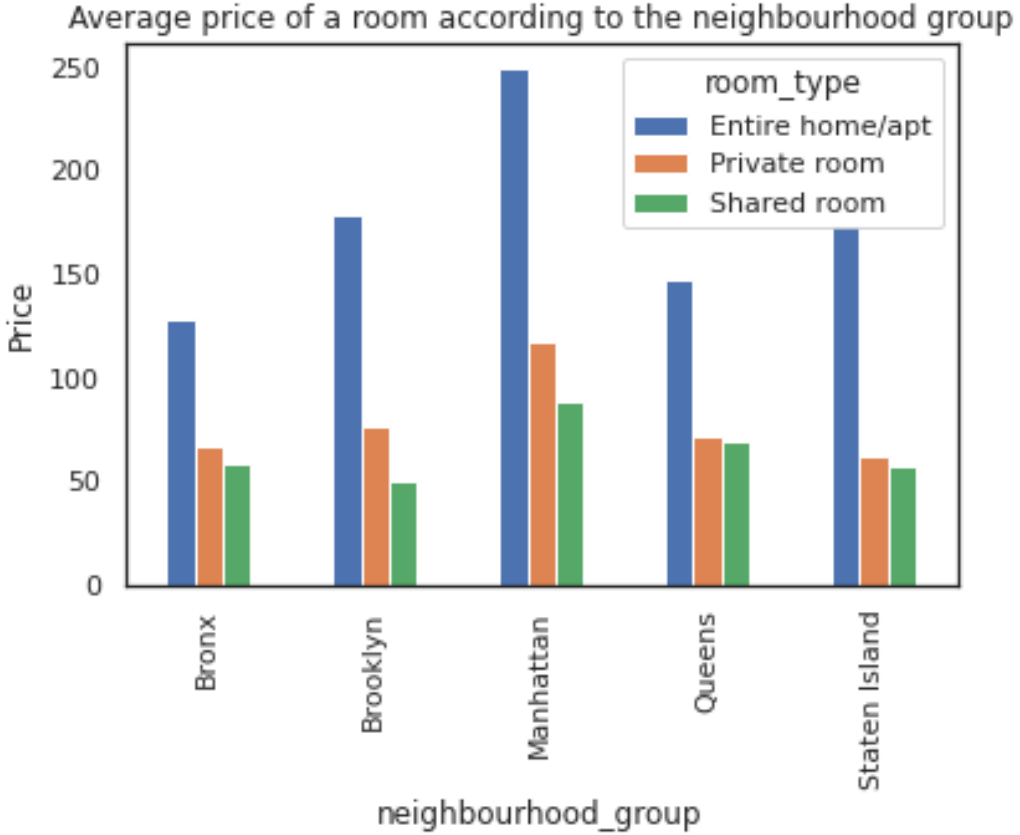
There are 221 unique neighbourhoods falls under 5 groups

Price distribution across neighbourhood_group

(which room type are most expensive and where it is located)

- Manhattan has the highest price for room types with Entire home/apt ranging to nearly 222$/night, followed by Private room with 109$/night . And it's obvious being the most expensive place to live in!

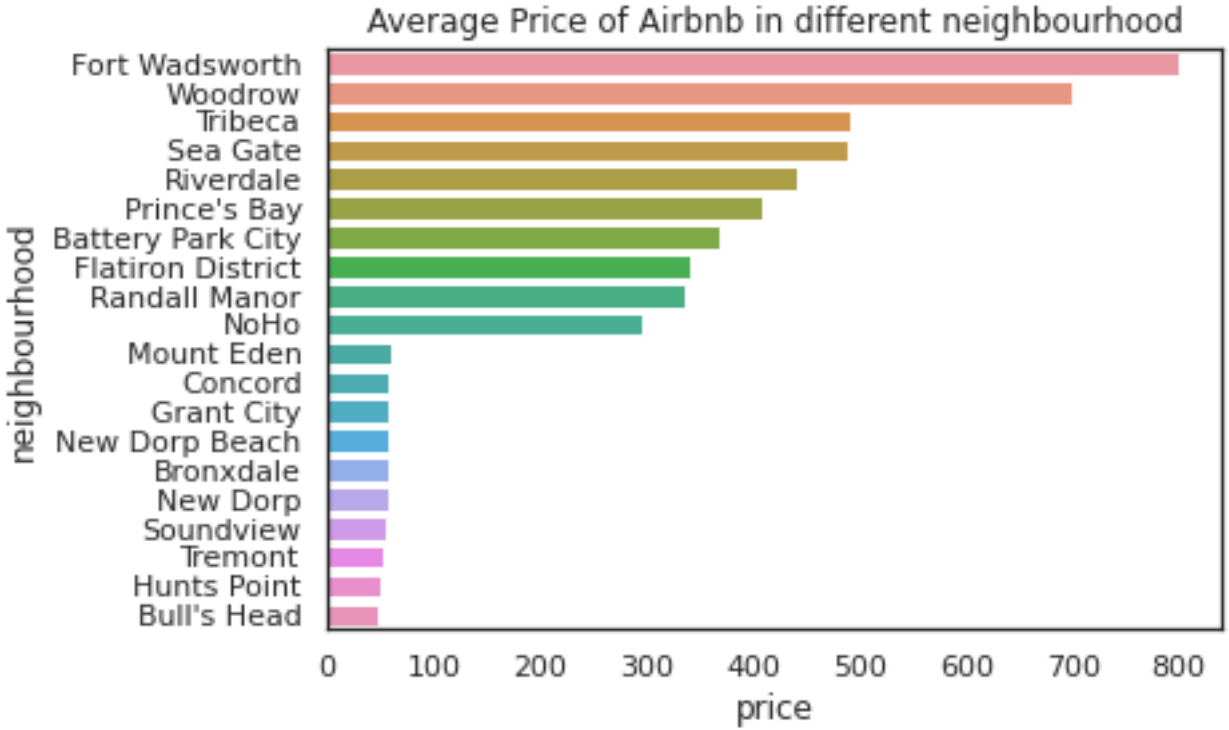
- In Manhattan you opt for entire home 40% more amount then opt entire home in Brooklyn.



Price distribution across neighborhood

(which room type are most expensive and where it is located)

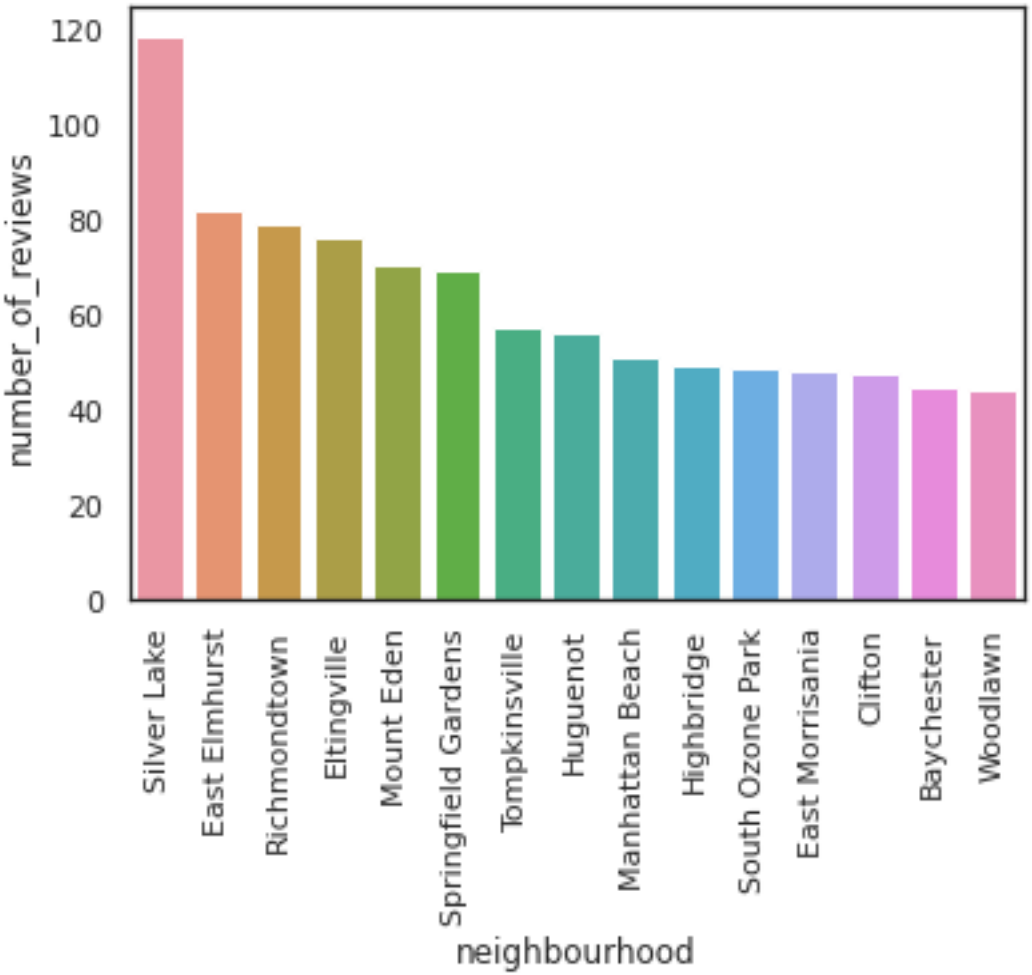
- Most expensive neighborhood is fort Wadsworth followed by Woodrow then Tribeca.
- Most cheapest neighborhood is hunts point followed by bull's head then Soundview.



Popular neighborhood by review

(which room type are most and where it is located)

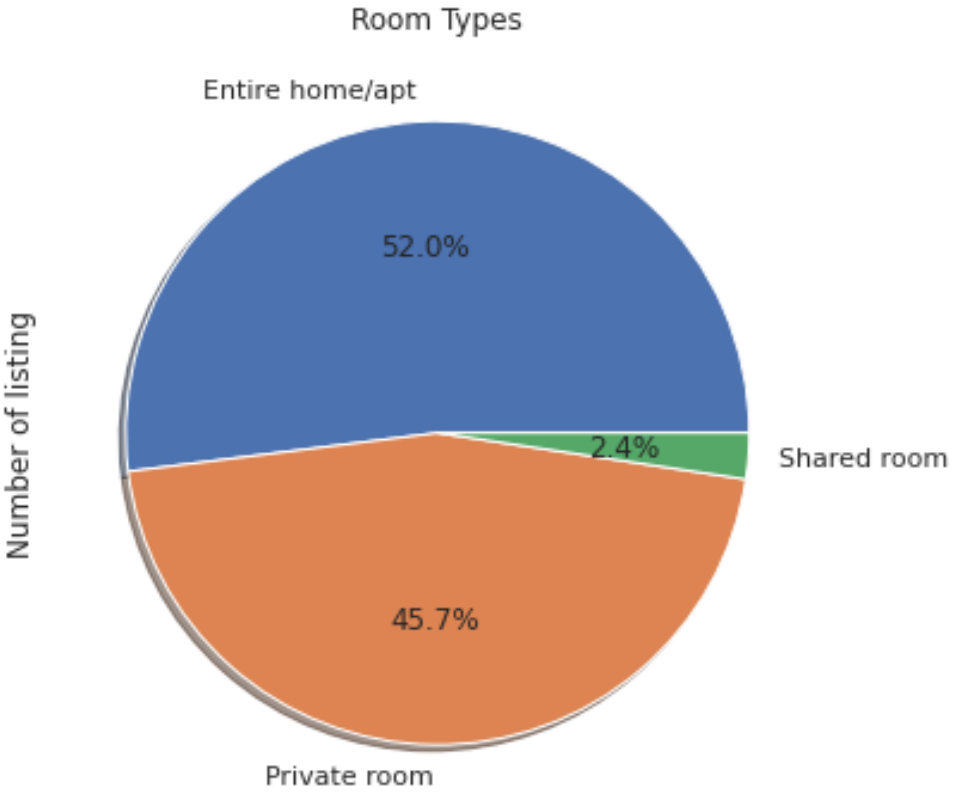
- The most review neighborhood is silver lake with average reviews of 118 per month, followed by East Elmhurst with average review of 83



Preferred room types

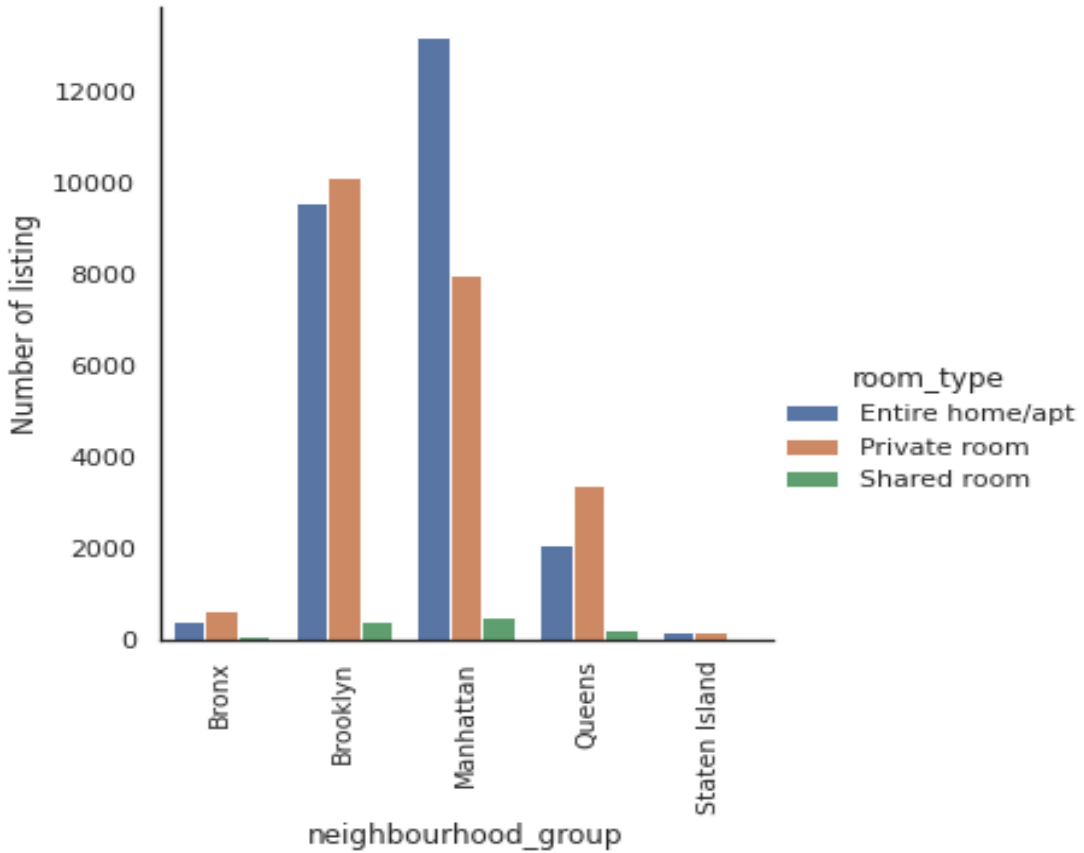
(which room type people are preferred to stay longer)

- The demand of entire home and private room is more high and people also choose entire home and private room.
- As per the dataset 97.7% of them are entire home or private room and only 2.4% of them are share room.



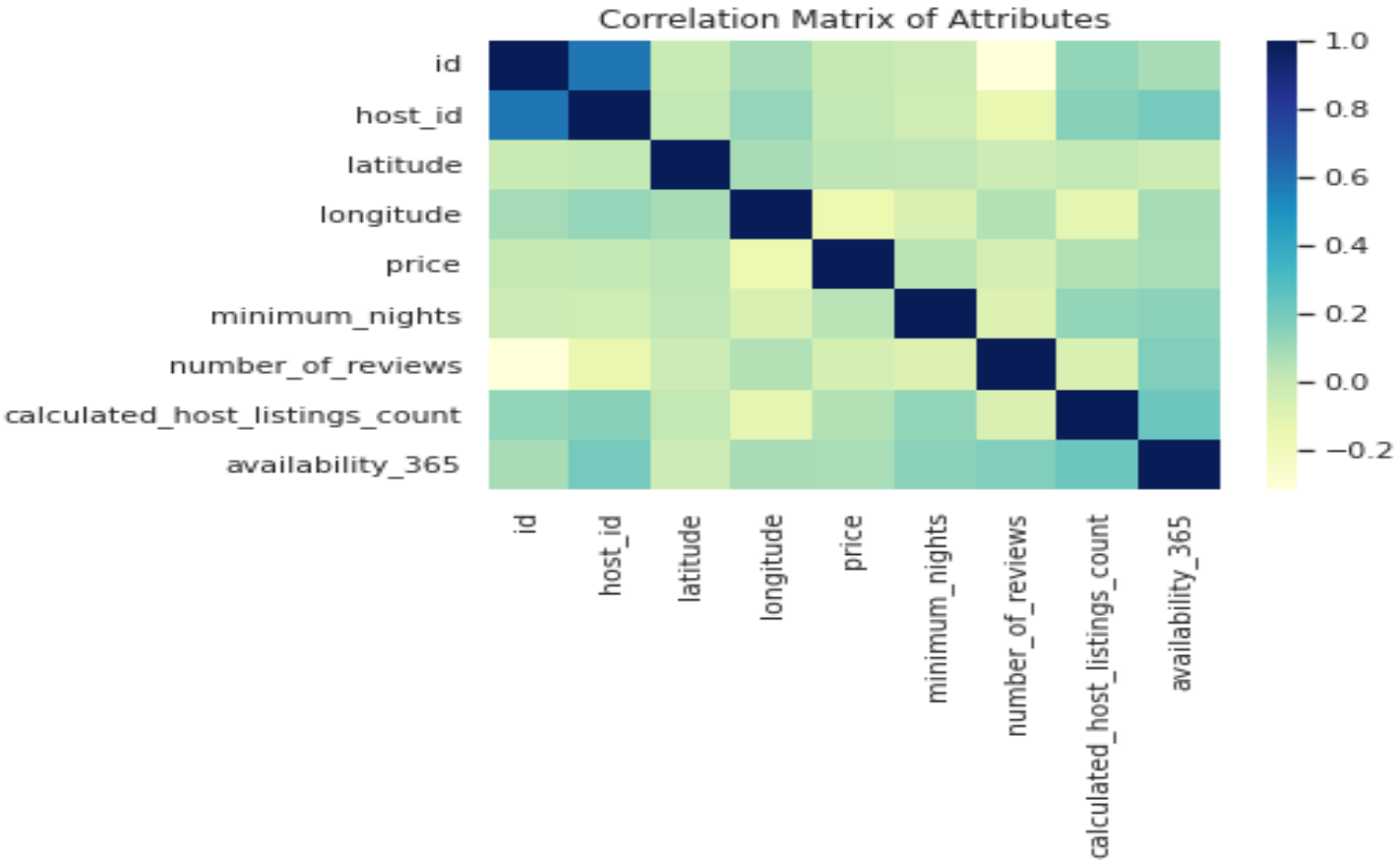
Preferred room types
(Location)

- The demand of entire home is more in Manhattan followed by Brooklyn and the price is also high.
- Demand of private room is more in Brooklyn followed by Manhattan.



Correlation Matrix :-

All the features are less correlated with price, regression taking price as target, will be less accurate



Limitations

- Dataset features in terms of modern world, are of very poor quality in deciding the valuation of a property
- User ratings of hosts aren't available, it would've been better to rank our hosts based on user satisfaction and ratings. Normally a low rated property tends to lower their price.
- In order to have a better analysis regarding the quality of the properties, it would be interesting if we had an analysis of sentiments with property valuations.
- The exact number of guests count also missing

Scope of Improvement

- ★ As dataset has less number of features to decide a property, more features can be added like bedroom, bathroom, property age (it might be one of the most important one), tax rate, distance to nearest airport, hospital, metro station or schools etc.
- ★ In presence of ratings, hosts can be classified and ranked, gives a special discount or offer to highest rated hosts following marketing strategy.
- ★ Time series analysis can be done to make decision on the rate in the tourist season.

Conclusion

From the entire analysis, it can be concluded that

- Most visitors don't prefer shared rooms, they tend to visit private room or entire home.
- We can conclude that throughout New York city there is a larger preference towards Entire apt as compared to private or shared rooms irrespective to the prices.
- Most of the people prefer to stay in Manhattan and Brooklyn where as least like place is staten island and Bronx.
- Though location of property has high relation on deciding its price, but a property in popular location doesn't mean it will stay occupied in most of the time.
- Performing a regression on this dataset may result in high error rate, as the features given in this dataset, are of very poor quality in deciding the property valuation. We can see this by looking at correlation heatmap. We would need more features like bedrooms, bathroom, property age (guessed it'd be a very important one), tax rate applicable on land, room extra amenities, distance to nearest hospital, stores or schools. These features might have a high relation with price.

Thank You

Team :- Team Denver

Team Member :- Sumit , Bhoomika

Batch Name - Cohort Cairo

As part of EDA Capstone Project by

AlmaBetter