Capstone Project-1

Project Title:
Airbnb Booking Analysis

Team Members

Rugada Manikanta,

Shabuddin Dhafedar,

Avanish dixit

What is AirBnb?

- Airbnb is an American company operates an online marketplace for lodging, primarily homestays for vacation rentals, and tourism activities.
- Airbnb does not own any of the listed properties; instead, it profits by receiving commission from each booking.
- Considering the Business model, it is a revolution in the hospitality industry because it can provide a stay in a very affordable cost to the guest and any host can easily register their property on the platform.


airbnb
PEOPLE
PLACES
LOVE
AIRBNB
MOMENTS

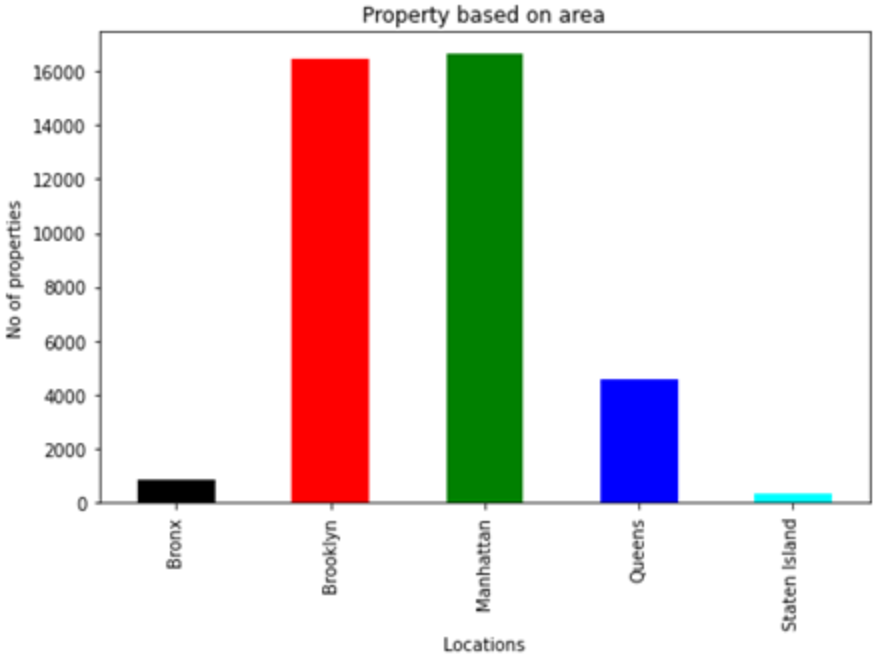
Data Pipeline

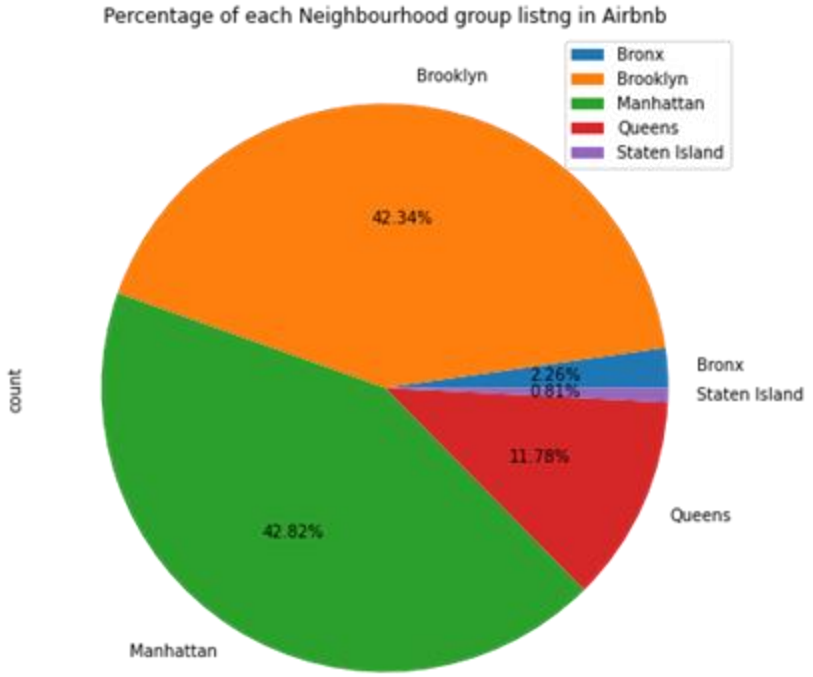
- **Collecting and Loading Data:** Hear we will load, see shape the shape data set and understand the data set features.
- **Experimenting On Data set:** Hear we make a understanding the data set of each feature, we will see data behaviour.
- **Cleaning dataset:** we will clean data set by understanding metrics of data set, we fill or remove missing values data.
- **Exploring and visualizing data:** Hear we understanding data one step more by linking one or more features, and we will develop the good visualization with data.

Brief info Regarding Data Set

- Shape: 48895 Rows, 15 columns
- Features: **Id**, *Name*, **Host_id**, *Host_name*, **Neighbourhood_group**, *Neighbourhood*, **Latitude**, *Longitude*, **Room_type**, *Price*, **Minimum_nights**, *Number_of_reviews*, **Reviews_per_month**, *Calculated_host_listings_count*, **Availability_365**
- 10- Numerical, 5-Non-numerical.
- Missing data: Name(**16**), Host name(**21**), Last review(**10052**) and Review per month(**10052**)

Neighbourhood Group

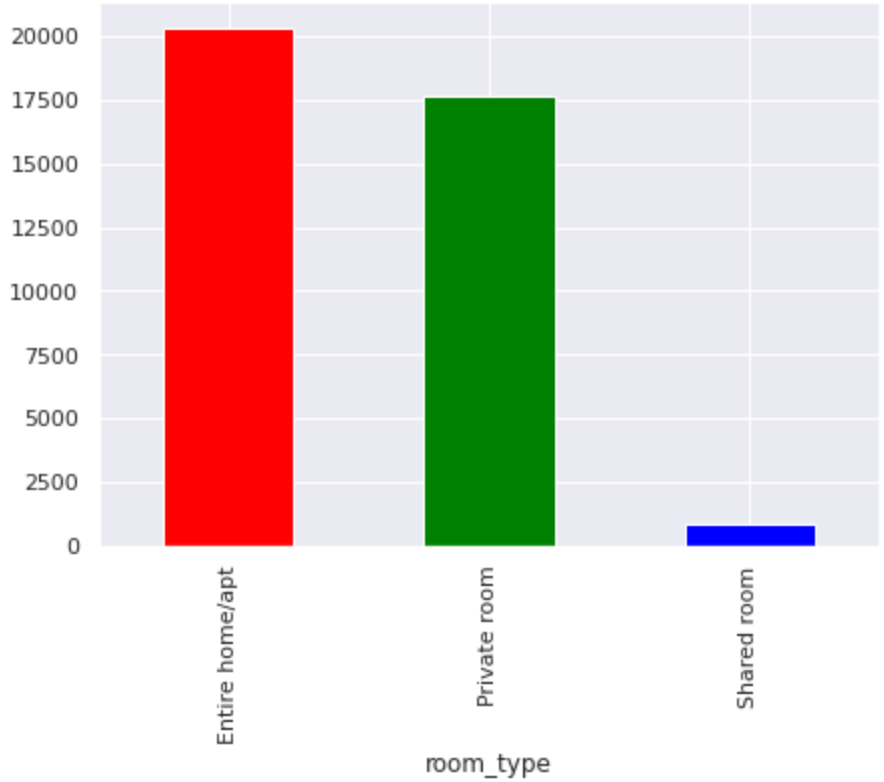




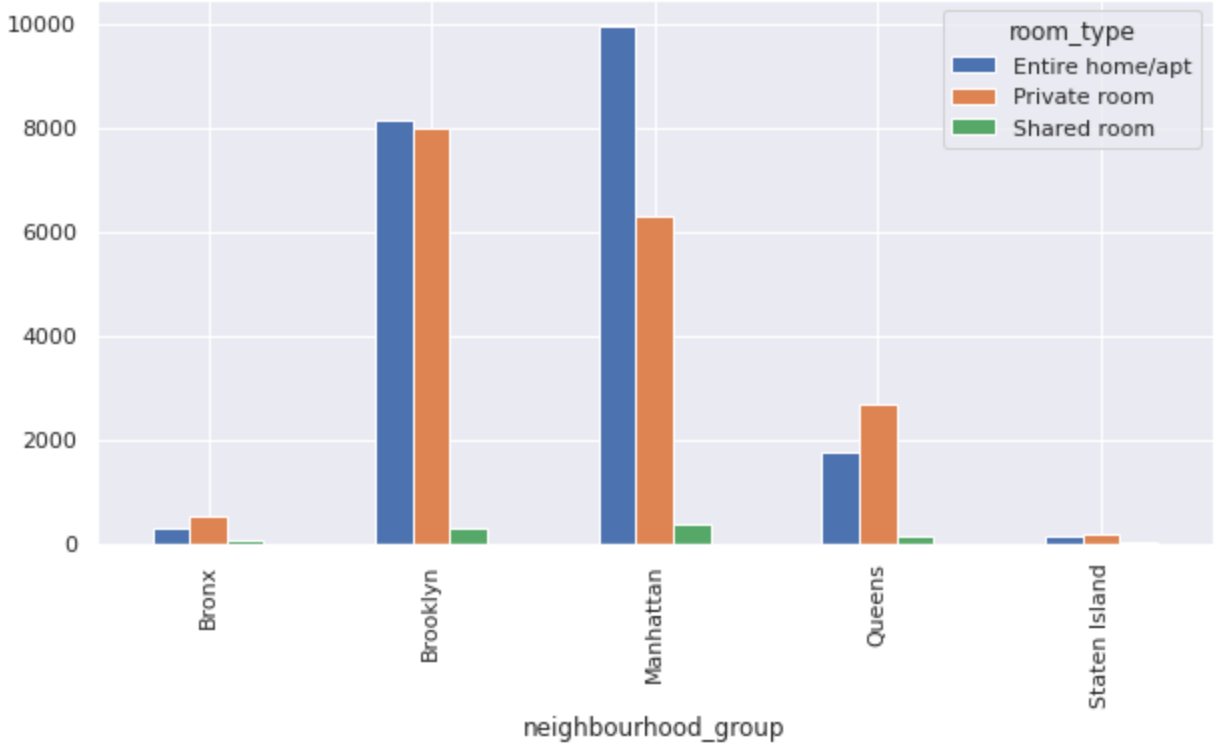
Room Type

Room are divided in three categories

- Entire Home/apartment
- Private room
- Shared room



Room Type


10000
8000
6000
4000
2000
0
room_type
Entire home/apt
Private room
Shared room
Bronx
Brooklyn
Manhattan
Queens
Staten Island
neighbourhood_group

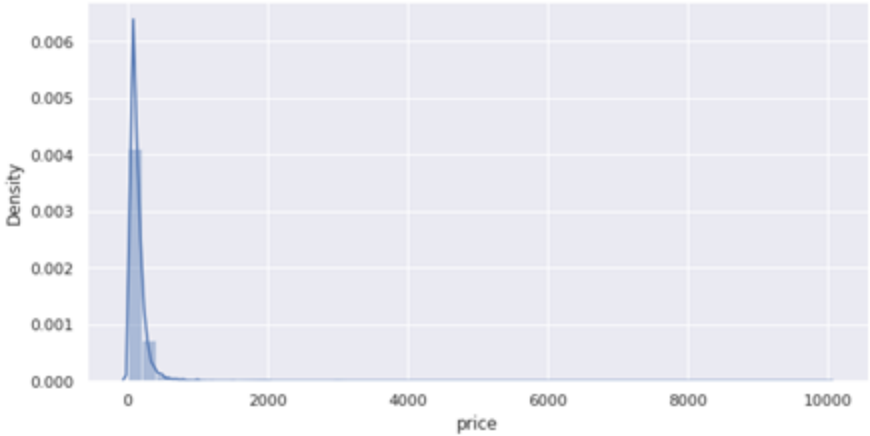
Neighbourhood

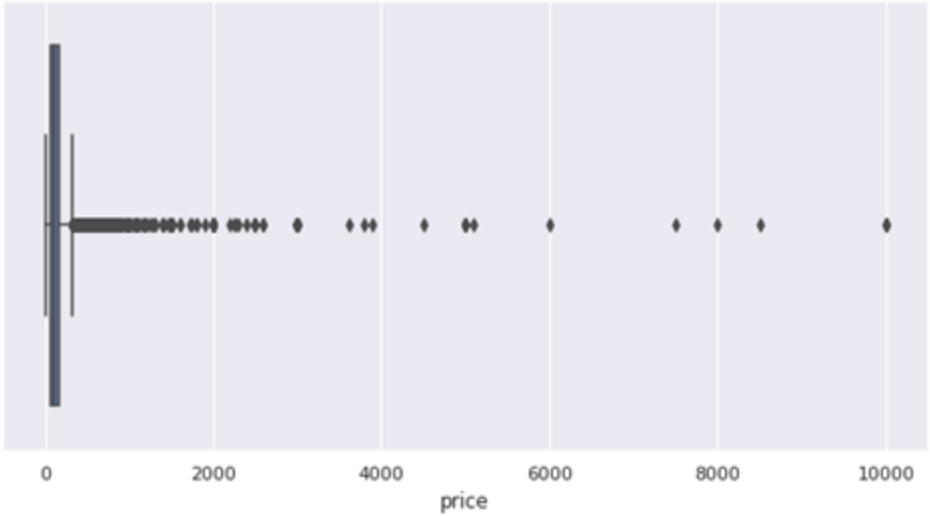
- **Williamsburg(Brookyn)** is the most listing happened across all other areas
- Properties in **silver Lake** have most number of reviews.
- Properties in **Co-op City** are mostly available throughout the year.
- **Sea Gate** neighbourhood highest average price of properties

Price

- Price of the listing varies from 0$ to 10000$ per day.
- 75% of listing were under the 175$ and mean price of listing are 154$.
- 10 properties are listed at 0$
- Most of the listing under 500$.

Price


0.006
0.005
0.004
0.003
0.002
0.001
0.000
Density
0
2000
4000
6000
8000
10000
price


0
2000
4000
6000
8000
10000
price

Price

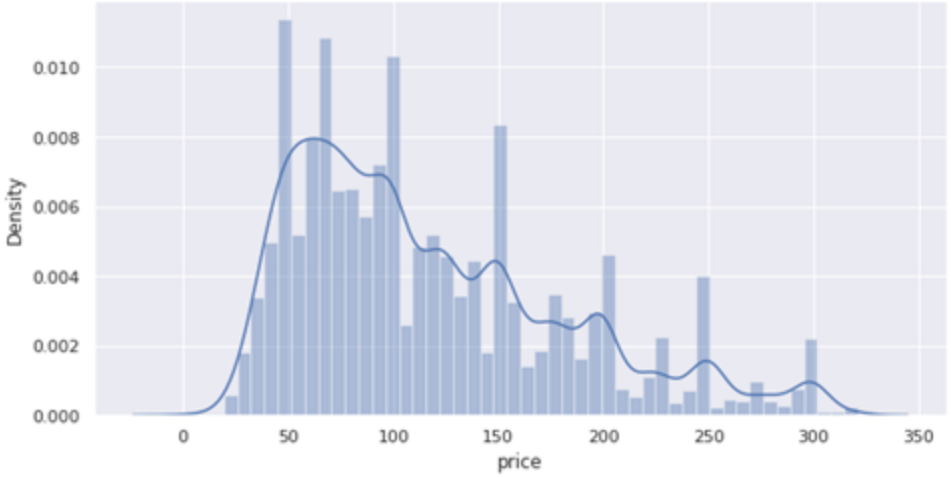
We found lot of out outliers, so we removed Interquartile range(IQR) technique

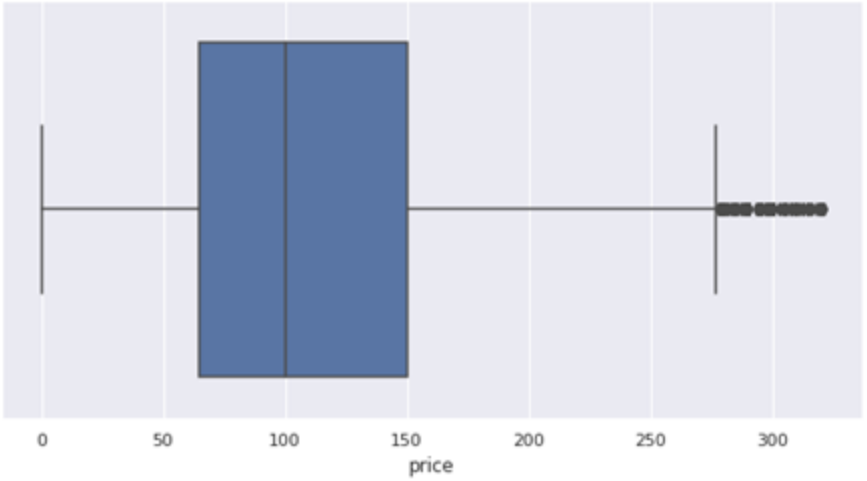
$$IQR = Q3-Q1$$

$$\text{Upper whisker} = Q3+1.5*(IQR)$$

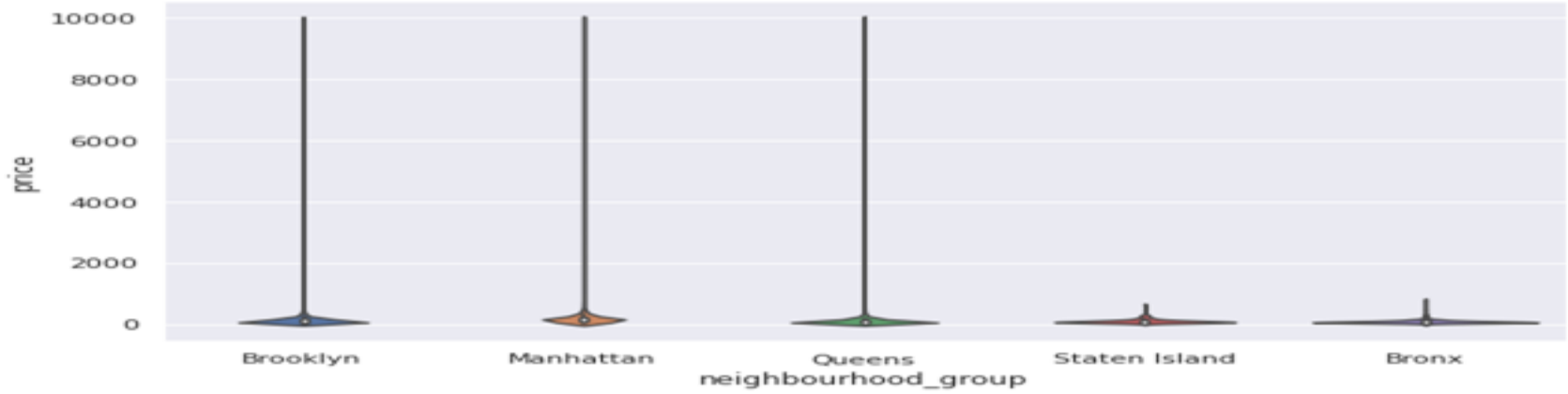
$$\text{Lower whisker} = Q1- 1.5*(IQR)$$

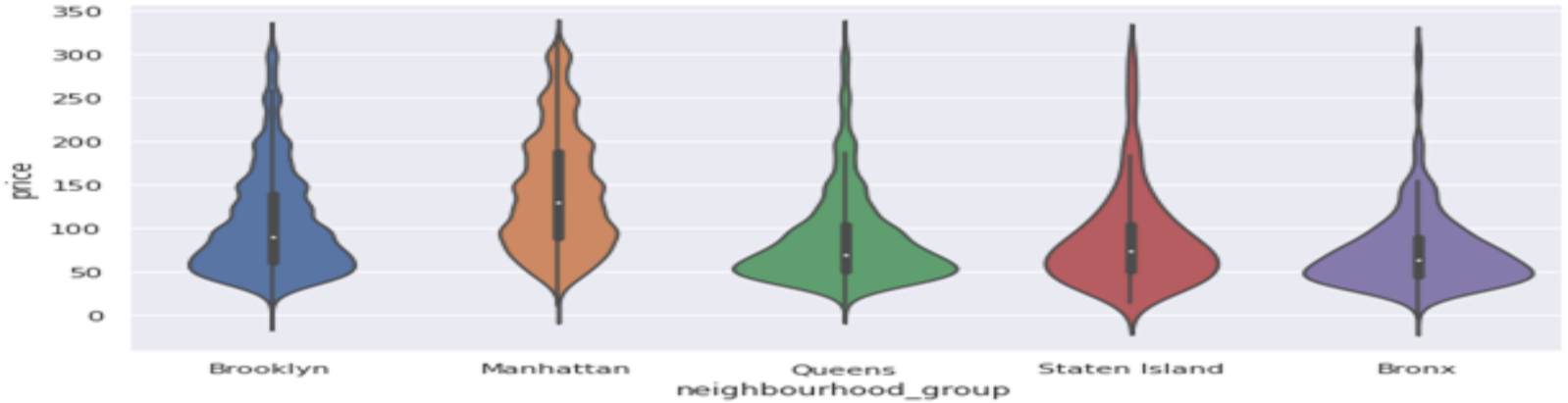
Price


0.010
0.008
0.006
0.004
0.002
0.000
Density
0
50
100
150
200
250
300
350
price


0
50
100
150
200
250
300
price

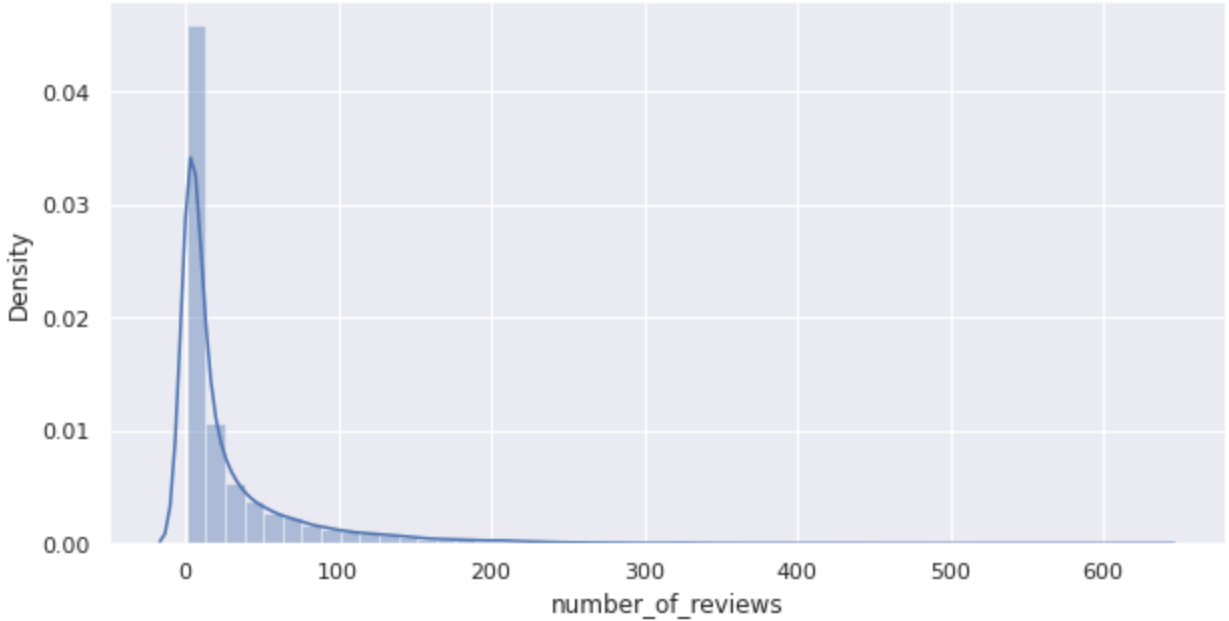
Price


10000
8000
6000
4000
2000
0
price
Brooklyn
Manhattan
Queens
Staten Island
Bronx
neighbourhood_group


350
300
250
200
150
100
50
0
price
Brooklyn
Manhattan
Queens
Staten Island
Bronx
neighbourhood_group

Number Of Reviews

There are 20271 properties which having less than 10 reviews only. This is almost 45% of total listing.



Number Of Reviews

We on something interesting that for some data Number of reviews=0

That data quantity was 10052 rows.

For same amount of data last review and review per month data was missing

As per my understanding 10052 properties are new listings, so review per month and last review was missing

```
zero_reviewd_listings=df[df['number_of_reviews']==0]
zero_reviewd_listings.info()
```

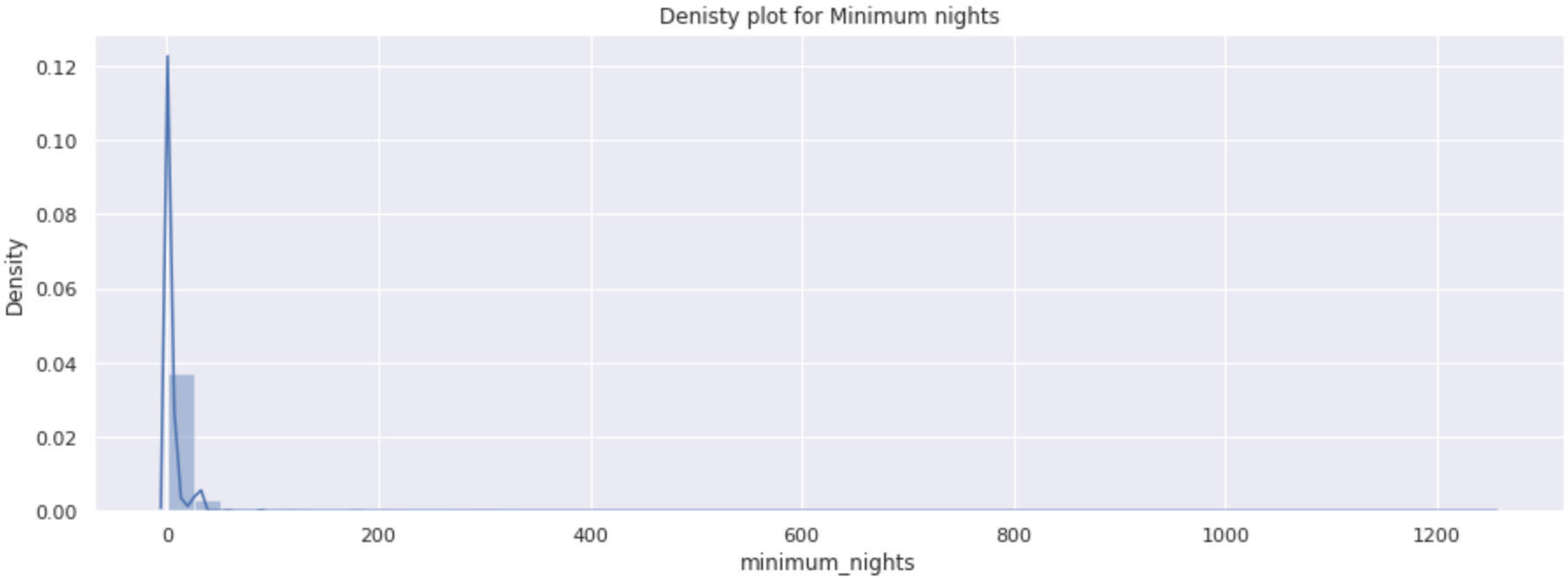
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 10052 entries, 2 to 48894
Data columns (total 16 columns):
 #   Column                          Non-Null Count  Dtype  
---  ------                          --------------  -----  
 0   id                              10052 non-null  int64  
 1   name                            10042 non-null  object 
 2   host_id                         10052 non-null  int64  
 3   host_name                       10047 non-null  object 
 4   neighbourhood_group             10052 non-null  object 
 5   neighbourhood                   10052 non-null  object 
 6   latitude                        10052 non-null  float64
 7   longitude                       10052 non-null  float64
 8   room_type                       10052 non-null  object 
 9   price                           10052 non-null  int64  
 10  minimum_nights                  10052 non-null  int64  
 11  number_of_reviews               10052 non-null  int64  
 12  last_review                     0 non-null      object 
 13  reviews_per_month               0 non-null      float64
 14  calculated_host_listings_count  10052 non-null  int64  
 15  availability_365                10052 non-null  int64  
dtypes: float64(3), int64(7), object(6)
memory usage: 1.3+ MB
```

Number Of Reviews

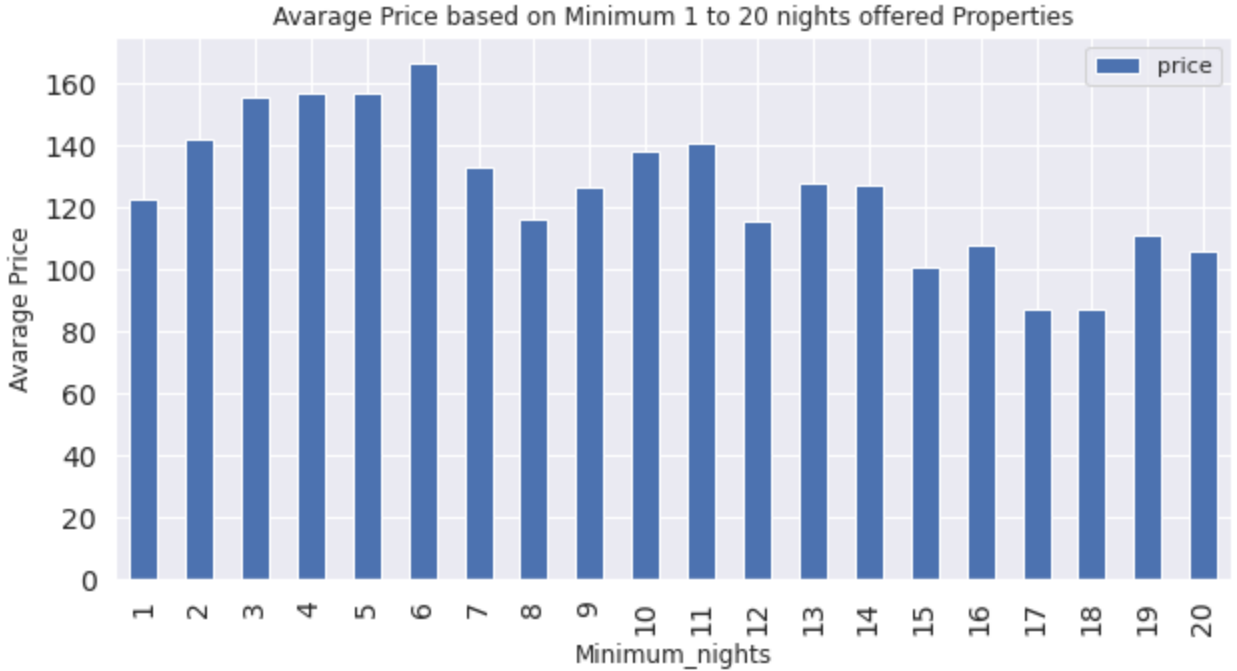
Number Of Reviews

	index	neighbourhood_group	number_of_reviews	host_name	room_type
0	11759	Queens	629	Dona	Private room
1	2031	Manhattan	607	Jj	Private room
2	2030	Manhattan	597	Jj	Private room
3	2015	Manhattan	594	Jj	Private room
4	13495	Queens	576	Dona	Private room
5	10623	Queens	543	Maya	Private room
6	1879	Manhattan	540	Carol	Private room
7	20403	Queens	510	Danielle	Private room
8	4870	Brooklyn	488	Asa	Entire home/apt
9	471	Brooklyn	480	Wanda	Private room

Minimum Nights



Minimum Nights



Hostname and Host Id’s

- No of Unique Host Id 37457
- No of unique Host Name are 11452

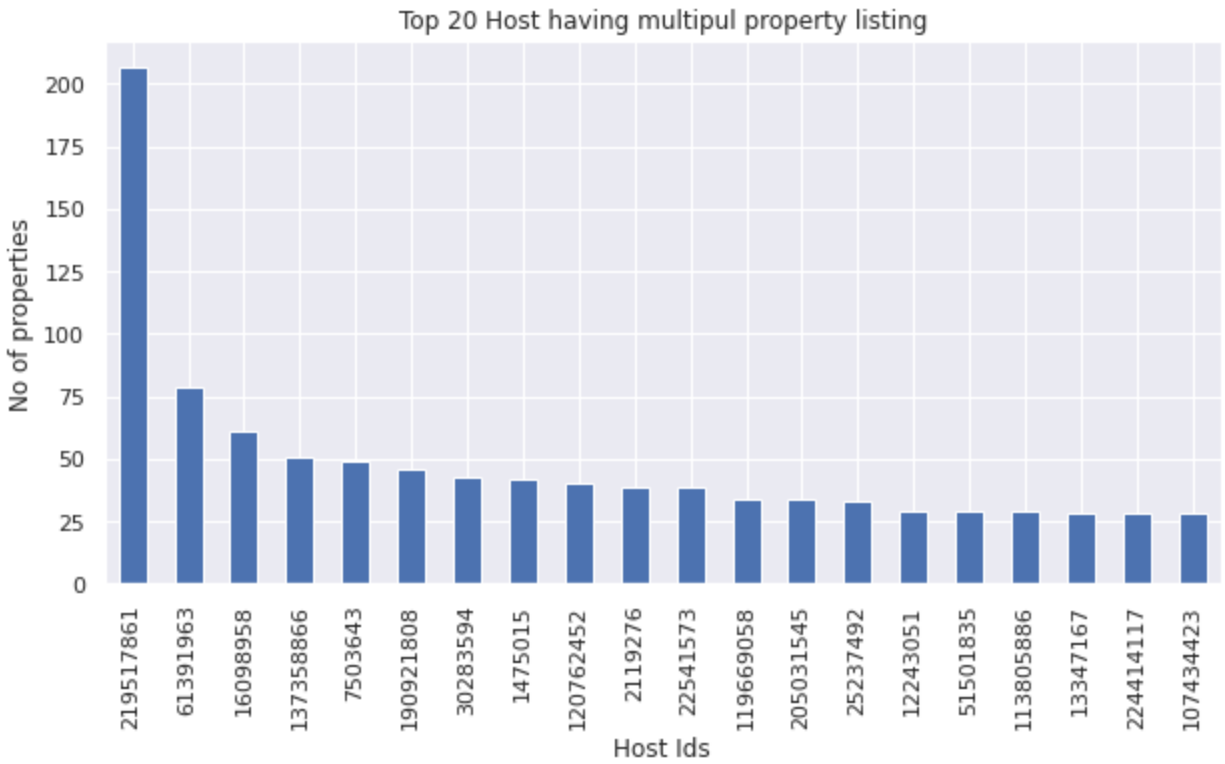
By seeing above data we can understand Host Name are very common By host id are unique.

I did experiment to prove this also, we just filtered JHON it was about 294 row in that 188 id out of that we made detail analysis on JHON of id 2787

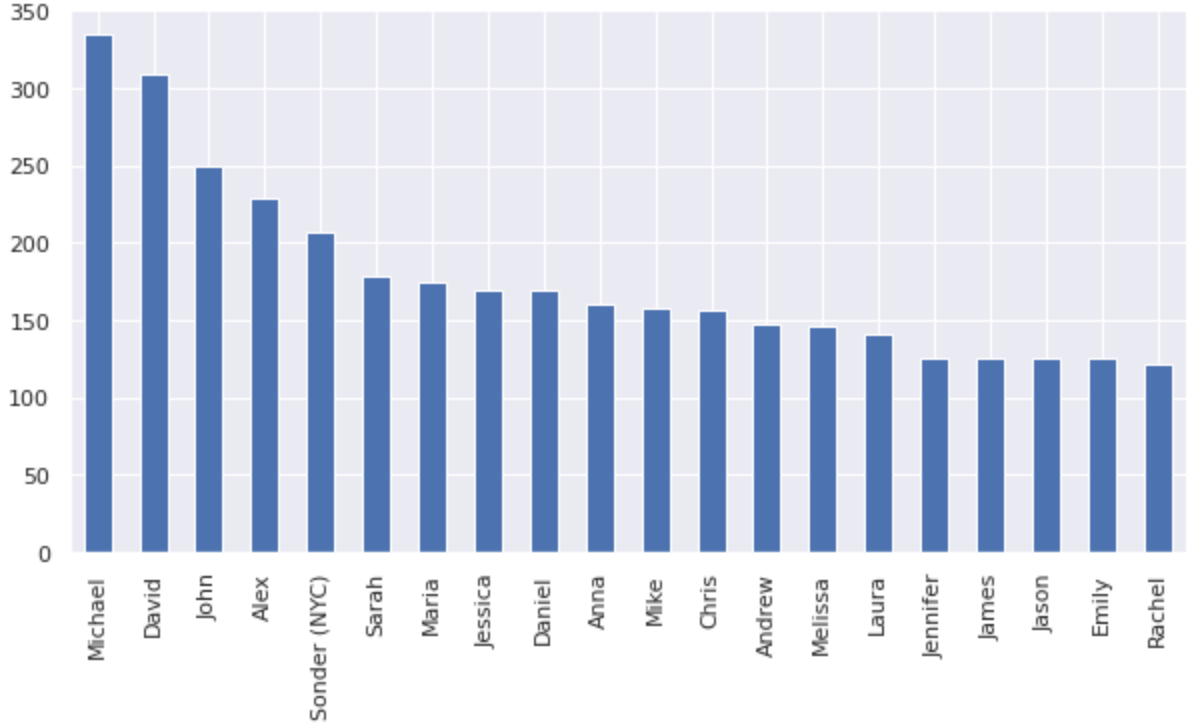
This id was repeated 6 times in calculated host list counting was shows 6.

By this we understand Id is most reliable data to use.

Hostname and Host Id’s

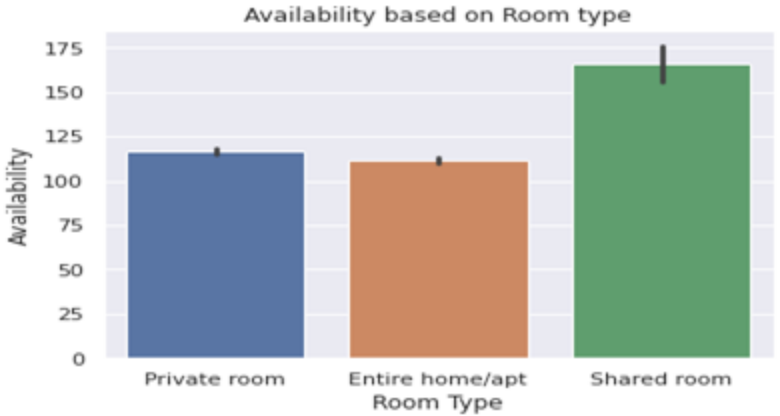


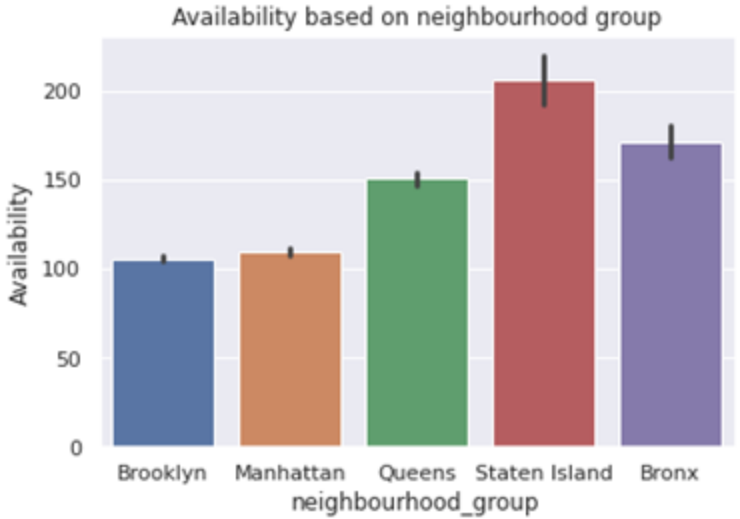
Hostname and Host Id’s



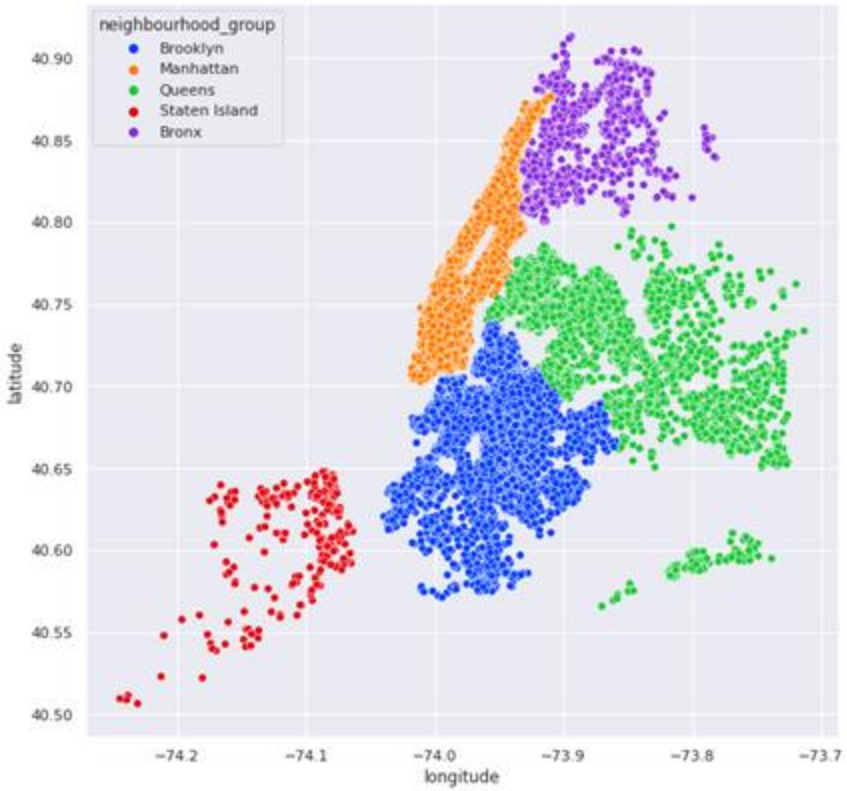
Availability 365

- There are 1295 properties are there which shows there their availability was for 365 days.
- 17533 properties are there which shows there availability was about 0 days.
- From the above scenario we can understand most of not interested to rent their property.



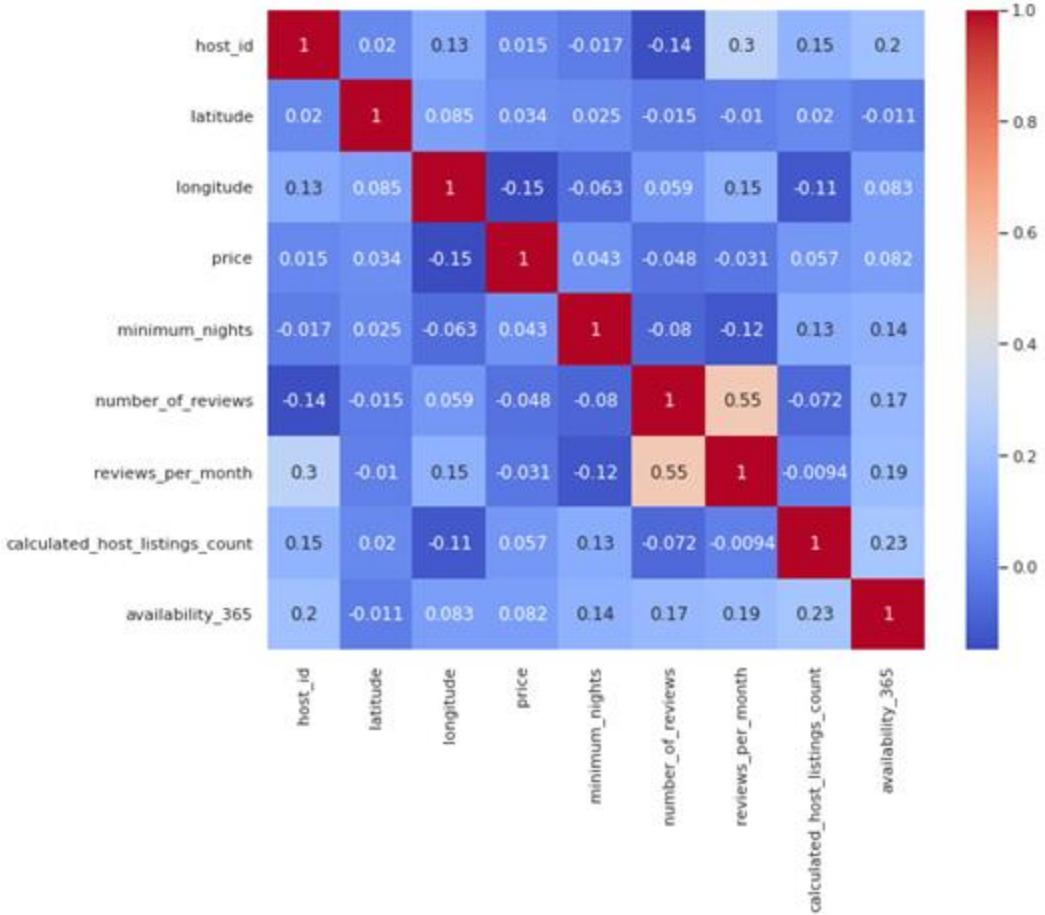


Latitude and Longitude



Correlation

There only one correlation between Review per month and number of reviews apart from that there is no strong correlation between the any numerical variables.



Conclusion

According to use we have concluded that Manhattan and Brooklyn having high number of booking, People are more interested in entire home rather than shared room, people are price conscious most of the listing are in 50-200. More people are preferred to stay 1 night

Thank You