Capstone Project

Airbnb Bookings Analysis

Team members

Kajal Dhun

Navinkumar Sambari

Shivam Singh

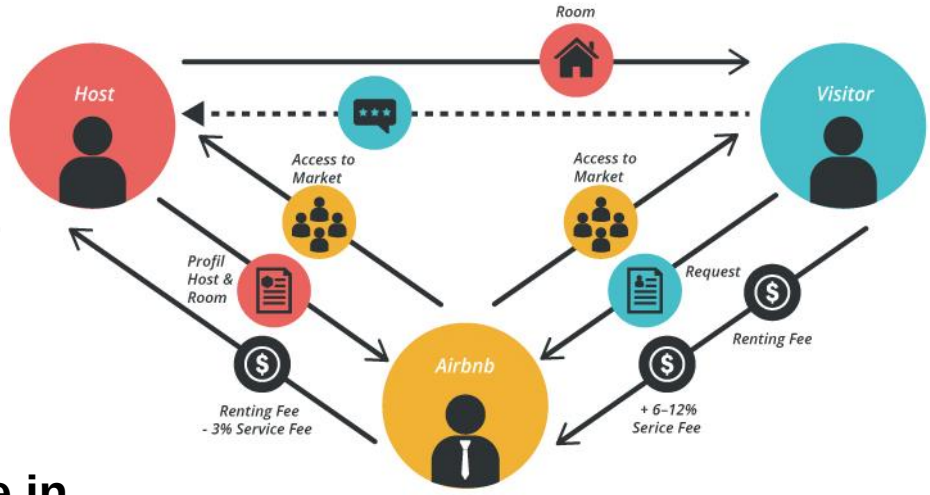
Tanu Rajput

Content

- Introduction
- Description of Data
- Data Cleaning
- Exploratory Data Analysis
- Result & Conclusion

Introduction

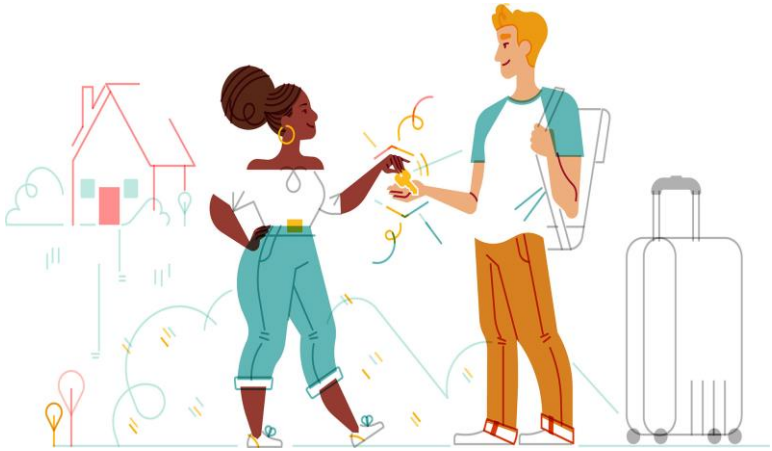
- **Airbnb started back in 2008, with the two founders, Joe Gebbia & Brian Chesky (Nathan Blecharczyk was invited to join later), and 3 air mattresses in San Francisco, California.**



- **Airbnb is an online platform that connects hosts renting out space in their homes with guests seeking lodging for generally cheaper prices than a hotel. Airbnb takes a 3% commission from bookings as well as a 6%-12% servicing fee from guests.**

  - **It currently covers more than 100,000 cities and 220 countries worldwide.**

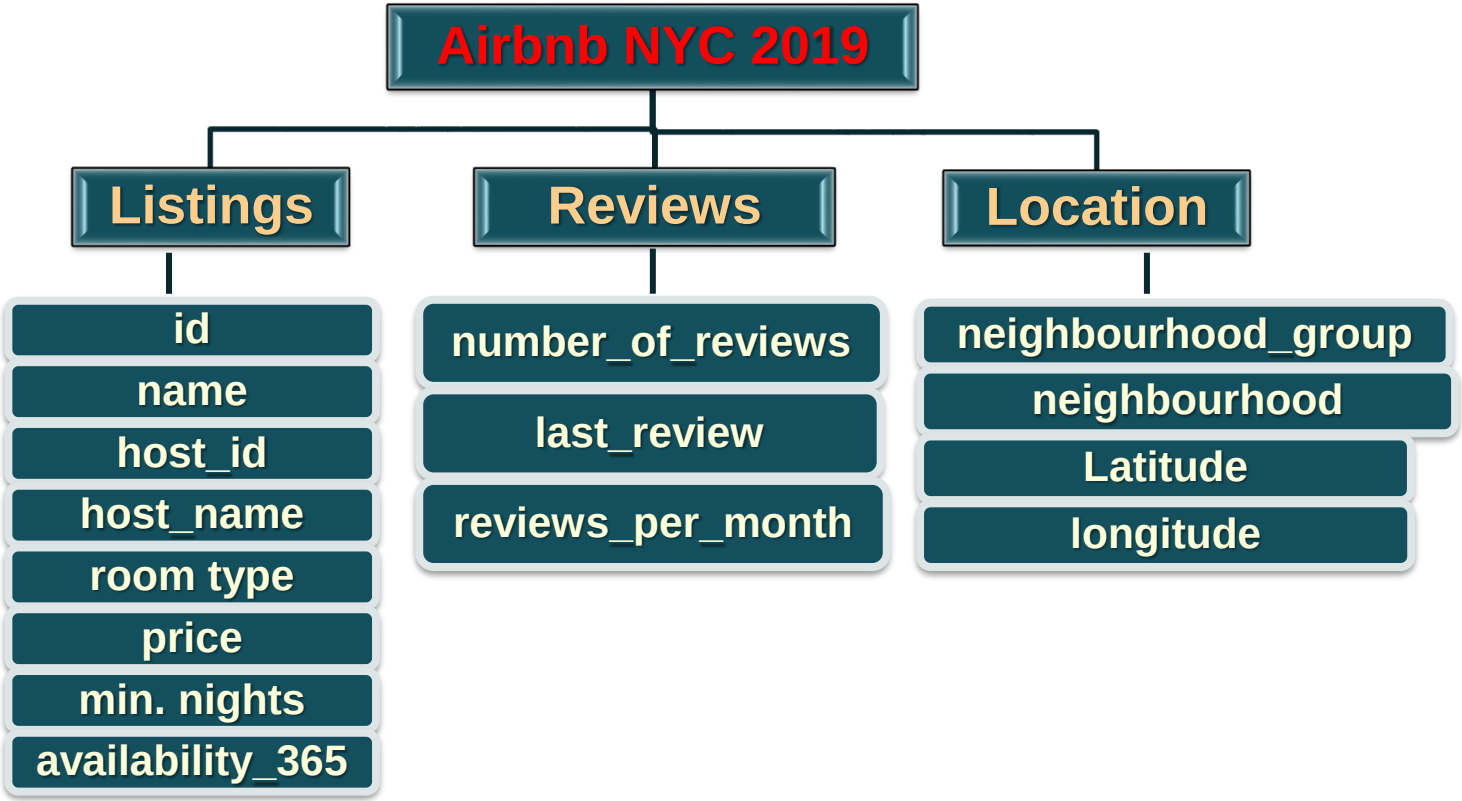
How Airbnb works?



For hosts

For guests

Data Description



A quick peek at the data:

- There are total **48895 rows** and **16 columns.**
- 16 columns are divided into ‘categorical data’ & ‘numeric data’ .
- Datatypes = int, float, object.
- Columns containing “**Null values**” = name, host_name, last_review, reviews_per_month.

```
<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                            48879 non-null  object
 2   host_id                         48895 non-null  int64
 3   host_name                       48874 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                     38843 non-null  object
 13  reviews_per_month               38843 non-null  float64
 14  calculated_host_listings_count  48895 non-null  int64
 15  availability_365                48895 non-null  int64
dtypes: float64(3), int64(7), object(6)
memory usage: 6.0+ MB
```

- **There is a total of 48895 listings in NYC.**
- **In every column below, there is a huge gap between third quartile value(75%) and maximum value.**
- **The minimum value for the price is 0. So, by all this, we come to know that there may be a possibility of outliers present in the dataset.**

```
df.describe()
```

	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365
count	48895.000000	48895.000000	48895.000000	38843.000000	48895.000000	48895.000000
mean	152.720687	7.029962	23.274466	1.373221	7.143982	112.781327
std	240.154170	20.510550	44.550582	1.680442	32.952519	131.622289
min	0.000000	1.000000	0.000000	0.010000	1.000000	0.000000
25%	69.000000	1.000000	1.000000	0.190000	1.000000	0.000000
50%	106.000000	3.000000	5.000000	0.720000	1.000000	45.000000
75%	175.000000	5.000000	24.000000	2.020000	2.000000	227.000000
max	10000.000000	1250.000000	629.000000	58.500000	327.000000	365.000000

Data Cleaning

- **Two columns** contain more than **10,000 null values**.

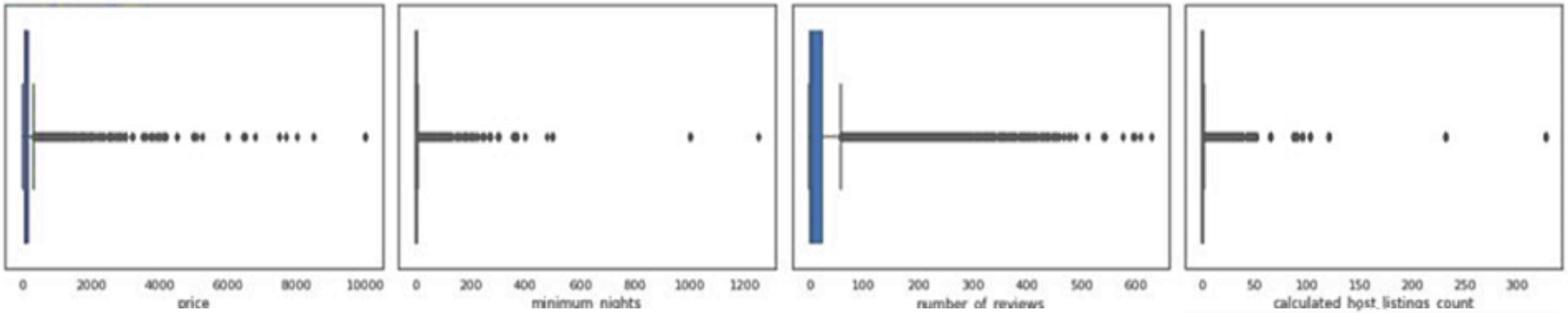
- We observe that null values are present in irrelevant columns so we directly clean the data by dropping certain columns that are not needed for analysis i.e., 'name' ,'host_name', 'last_reviews', 'latitude', 'longitude'

```
df.isnull().sum()
```

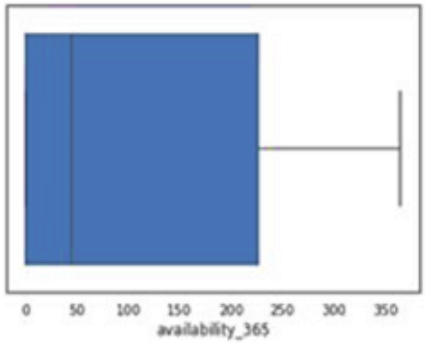
```
id                                  0
name                               16
host_id                             0
host_name                          21
neighbourhood_group                 0
neighbourhood                       0
latitude                            0
longitude                           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
```

Outliers and Handling Data:

- In the below results, columns, especially, ‘price’, ‘minimum_nights’, ‘calculated_host_listings’ have Outliers. We decided to remove outliers of ‘price’ and ‘minimum_nights’ by using the Quantile method.
- We will not be taking any action on ‘calculated_host_listings_count’ even though it contains outliers.
- There is no single outlier we found in the ‘availability_365’ column.



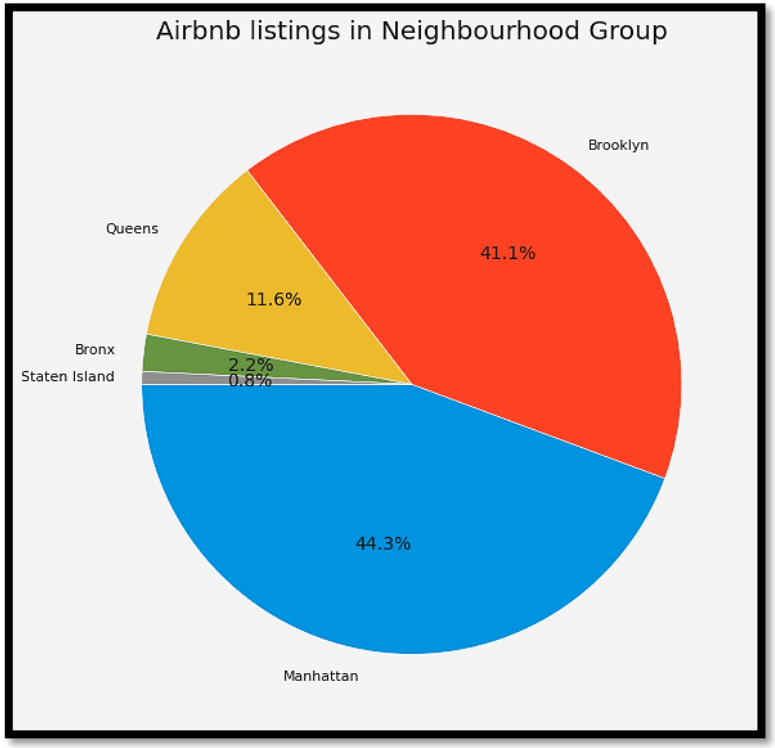
✓ After cleaning Dataset, we have 48796 rows and 11 columns.



EDA

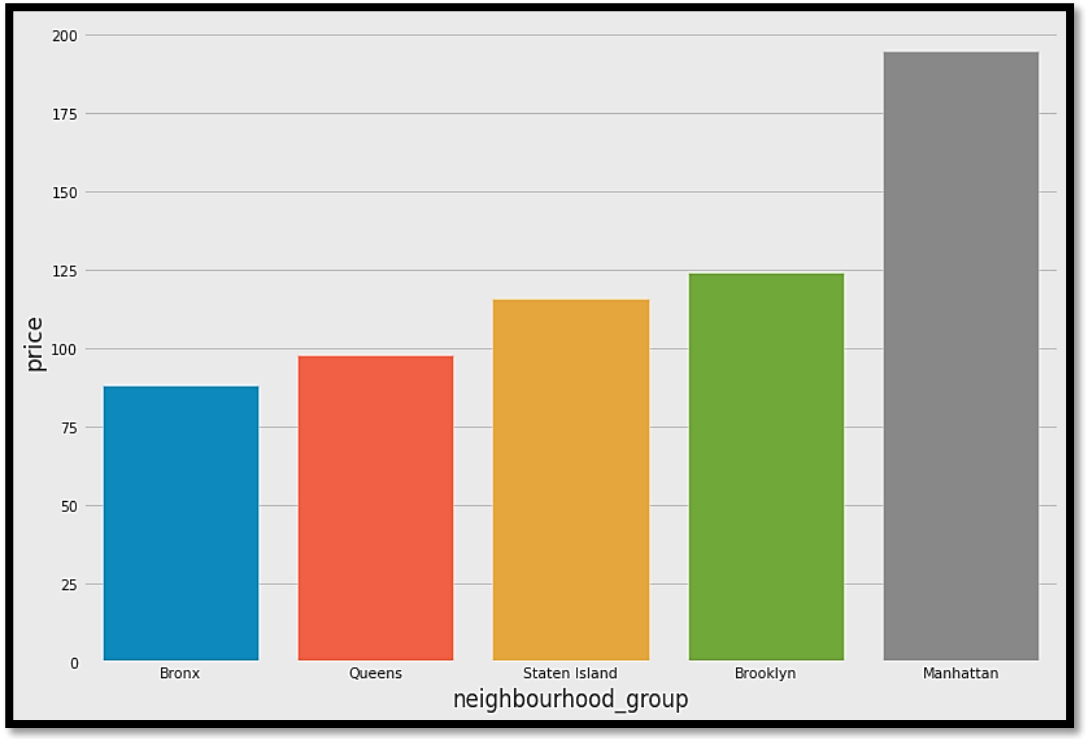
Number of Airbnb listings in neighbourhood group.

- The chart shows that 44.3% of Airbnb housings are located in Manhattan, and 41.12% of Airbnb housings are located in Brooklyn.
- "Staten island" has the least number of Airbnb listings.



Avg. price of Airbnb listings in neighbourhood group.

- Manhattan receives the highest average price of $194.8 because of its highly demand.
- Manhattan has the most expensive rentals compared to the other neighbourhood group.
- Bronx receives the lowest the average price of $88.0

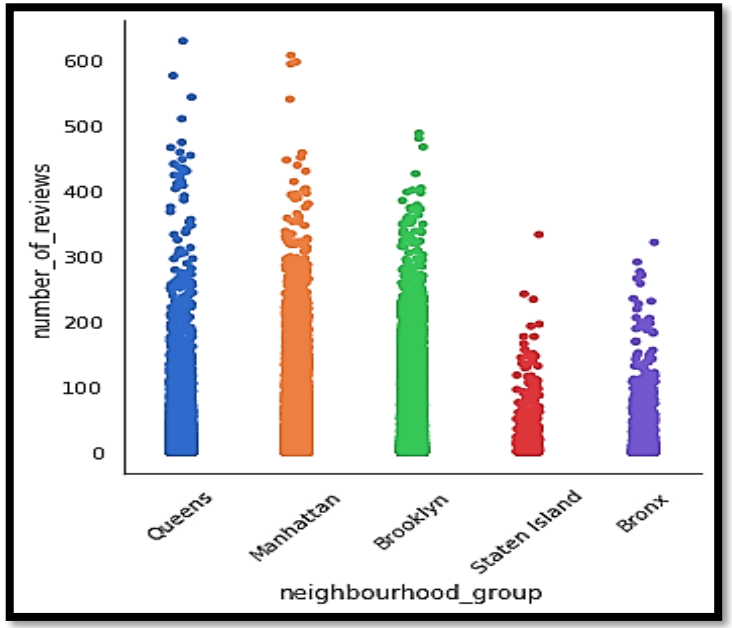


Neighbourhood group contains the listings with most reviewed.

- Queens and Manhattan have the listings with most reviewed.
- Bronx has the listings with less reviewed.

	id	neighbourhood_group	number_of_reviews
0	9145202	Queens	629
1	903972	Manhattan	607
2	903947	Manhattan	597
3	891117	Manhattan	594
4	10101135	Queens	576
...	...	...	...
48791	31797655	Manhattan	0
48792	2224896	Manhattan	0
48793	9794251	Manhattan	0
48794	2222428	Manhattan	0
48795	36487245	Manhattan	0

48796 rows × 3 columns



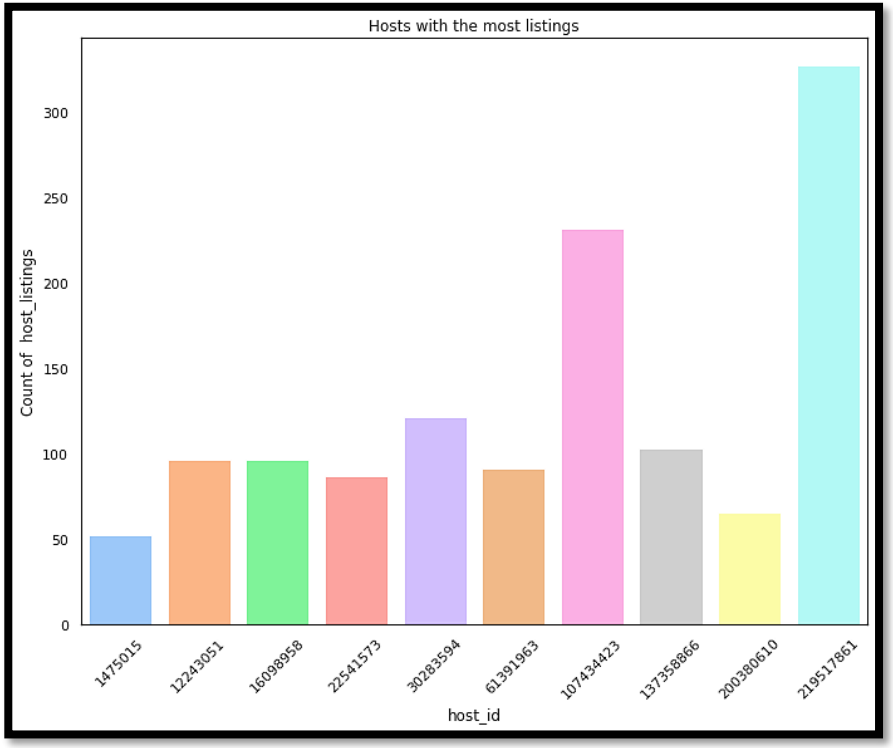
Top host with the highest number of Airbnb listings.

- the host with host id number = 219517861. (His name is Sonder NYC) who has 327 listings is the top host in the whole of NYC.

- **JUST OUT OF CURIOSITY!!!**

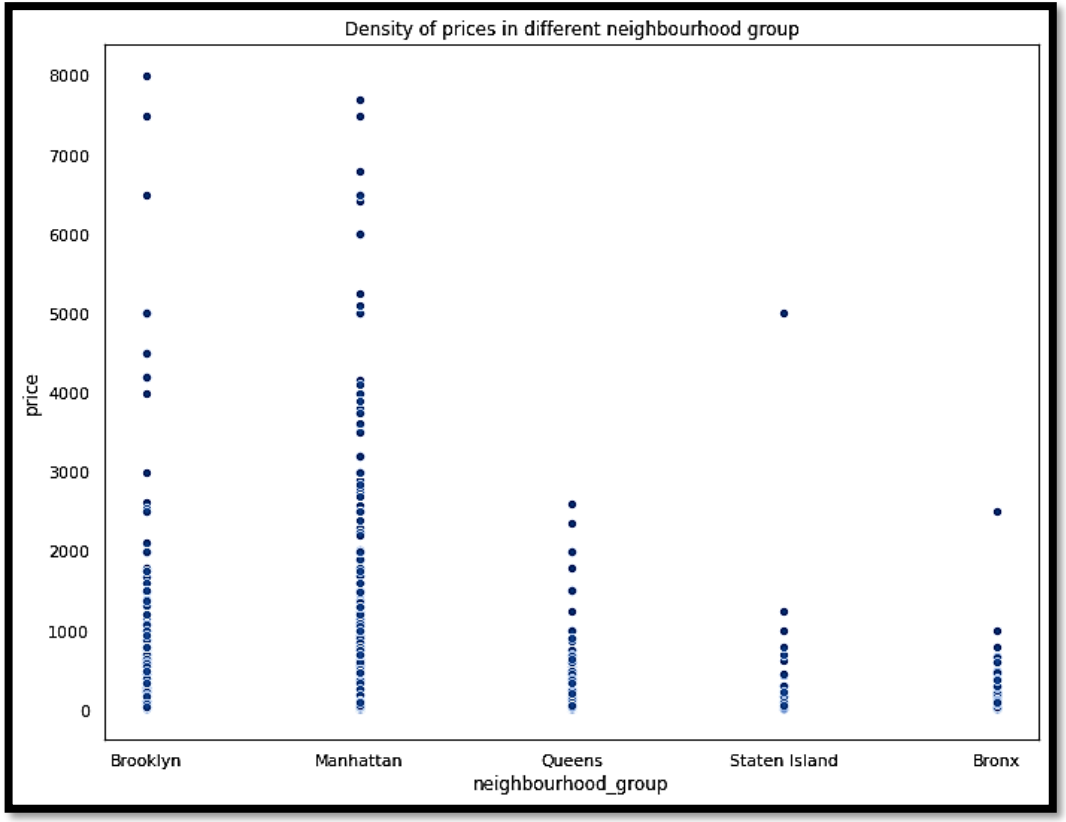
Host Id	Host name	No. of listings
219517861	Sonder NYC	327

- If all his listings were booked! for one night, he earns total **$82795**
- The average money he earns with all his listings is **$253.19.**
- After deducting Airbnb's commission, he gets the actual amount which is **$80311.15**



Price distribution of listings in neighbourhood group.

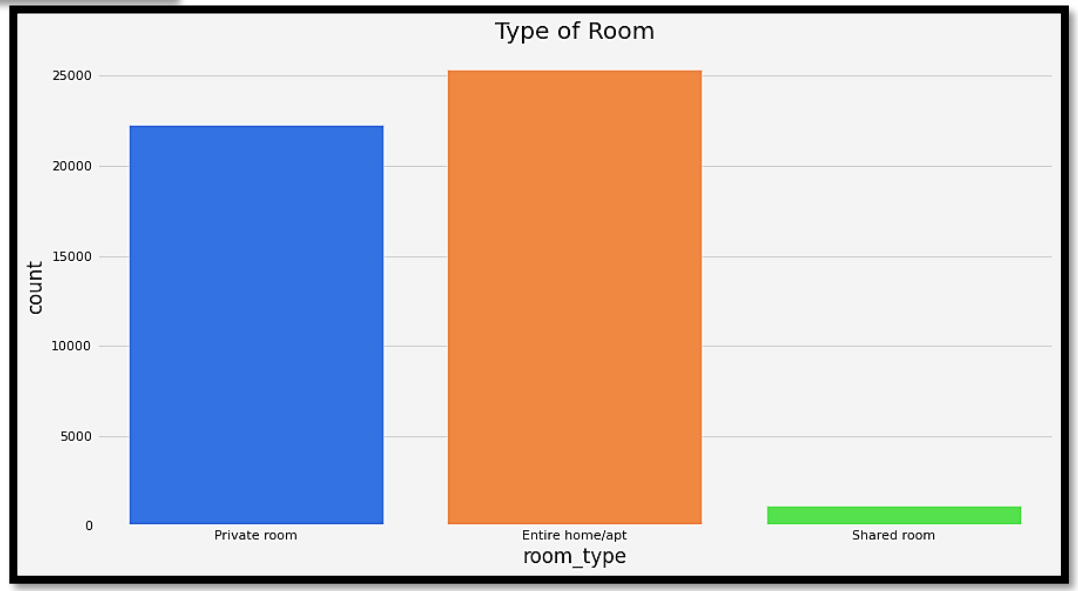
- **Brooklyn has the highest the density of price distribution.**
- **Bronx has the lowest.**



Counting the different room type!

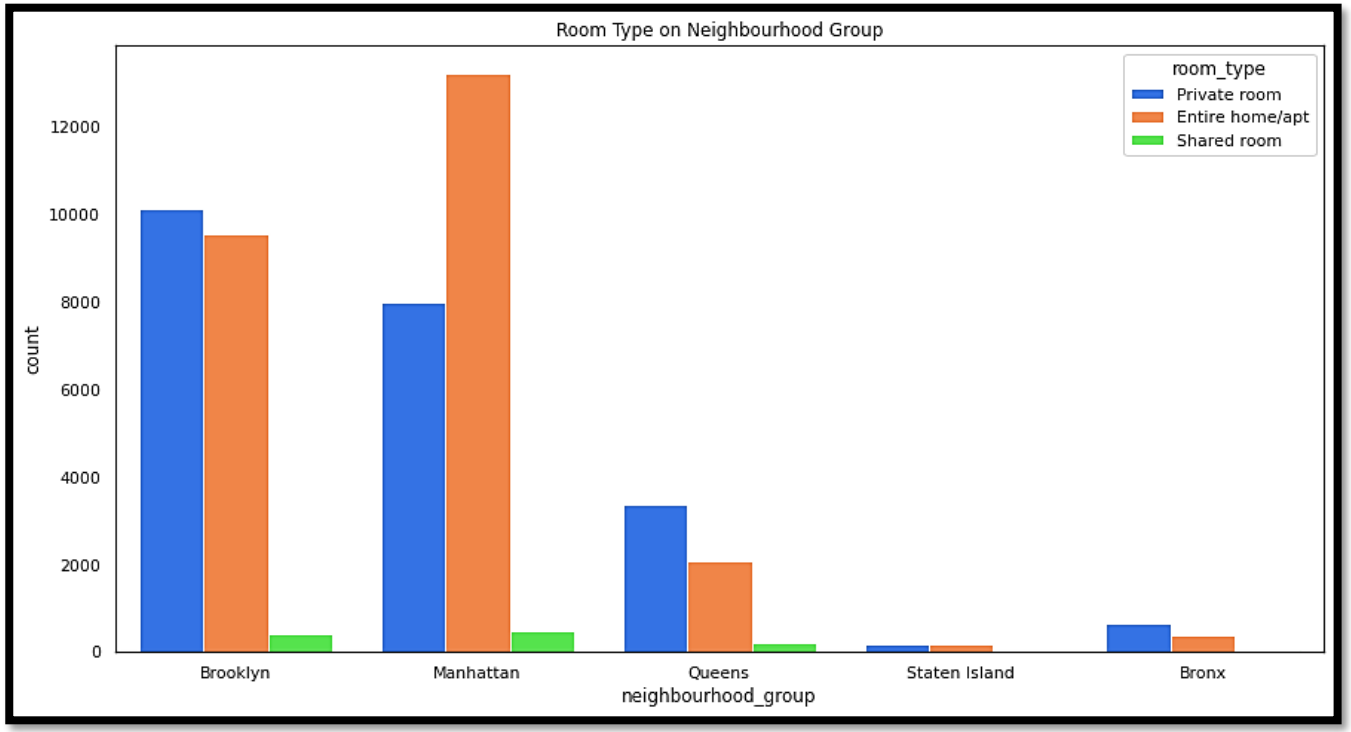
```
[253] counting_room_type = new_df1.room_type.value_counts()
      print(counting_room_type)

      Entire home/apt    25381
      Private room       22292
      Shared room         1123
      Name: room_type, dtype: int64
```

- In 2019, there are a total of 25381 ‘Entire home/apt’ room type in NYC, while shared room count is less.



Total of different room type in each neighbourhood group.

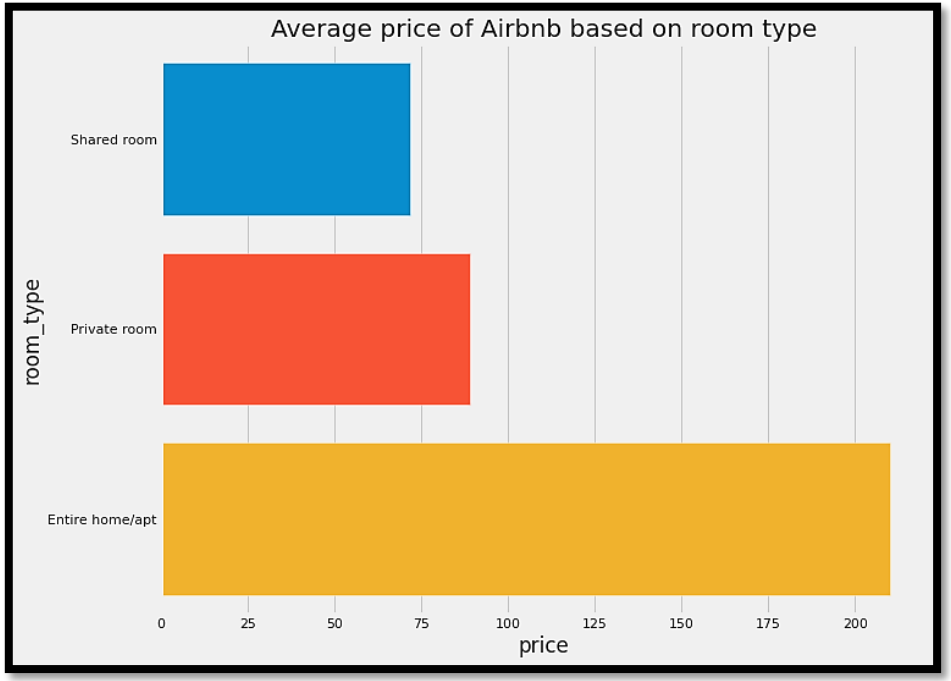
1. Brooklyn has a high number of private room spaces
2. Manhattan has a high number of entire homes/apt. room type
3. Queens have the highest private room spaces which is much lesser than Brooklyn and Manhattan.
4. Staten island and the Bronx has negligibly shared rooms.!



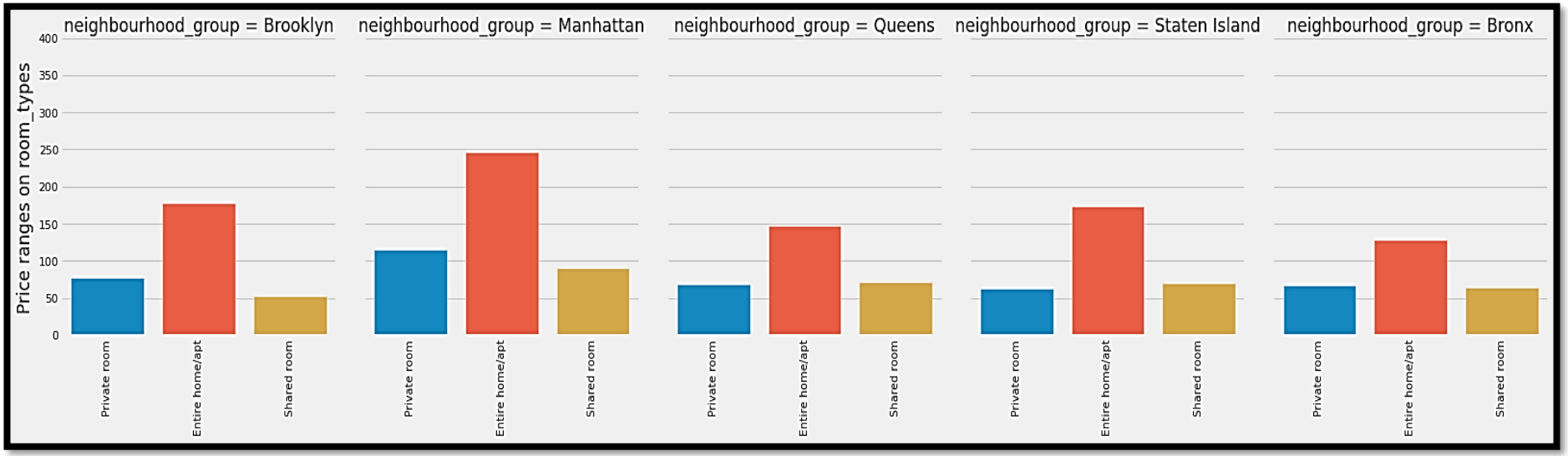
Average price of each room type.

- Most expensive room type is Entire home/Apt followed by a private and share room.
- Average Price of a Private room is like about 50% cheaper than the Entire room.

	room_type	price
2	Shared room	71.796082
1	Private room	89.000314
0	Entire home/apt	210.081124



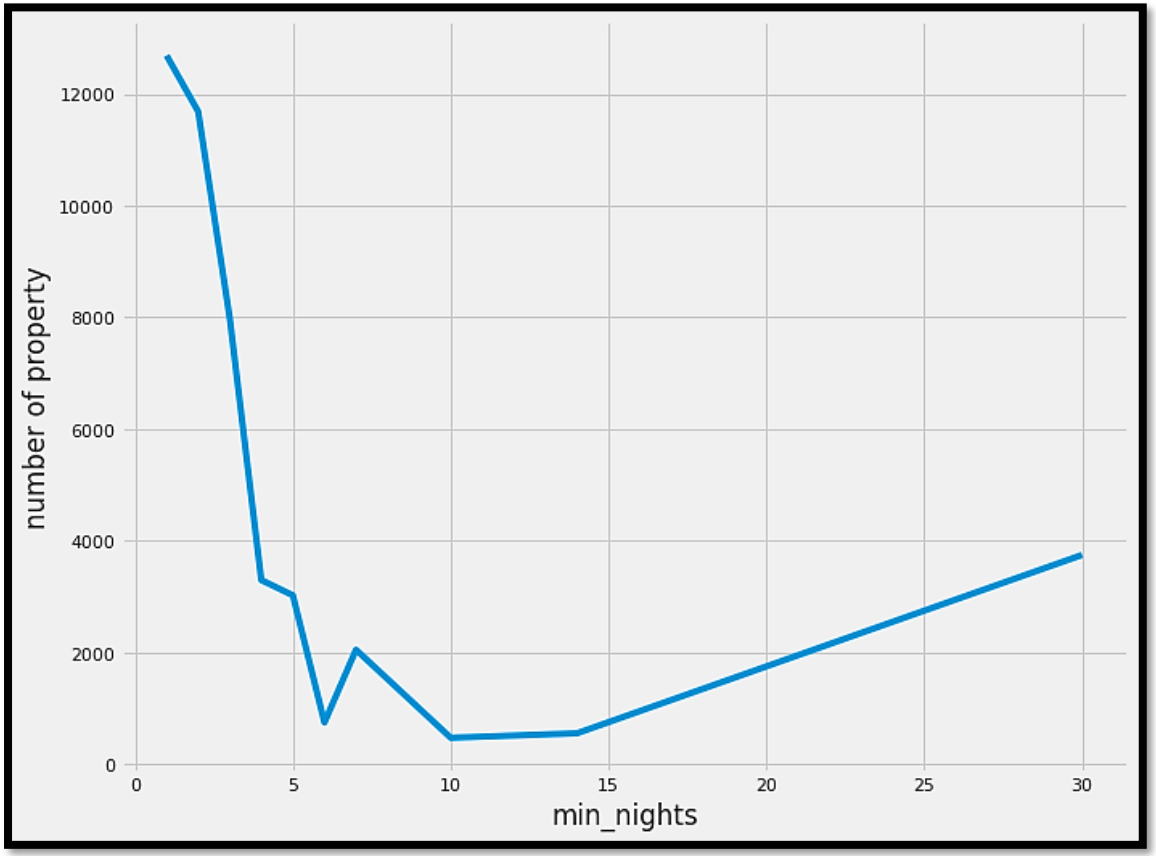
Price range of each room types in different neighbourhood group



- So, the average price of an Entire home/apt in Manhattan is high, which is around $ 246.5 while the average price of an Entire home/apt in the Bronx is low, i.e., $127.5.
- Talking about shared room prices, again Manhattan has high compare to the other boroughs

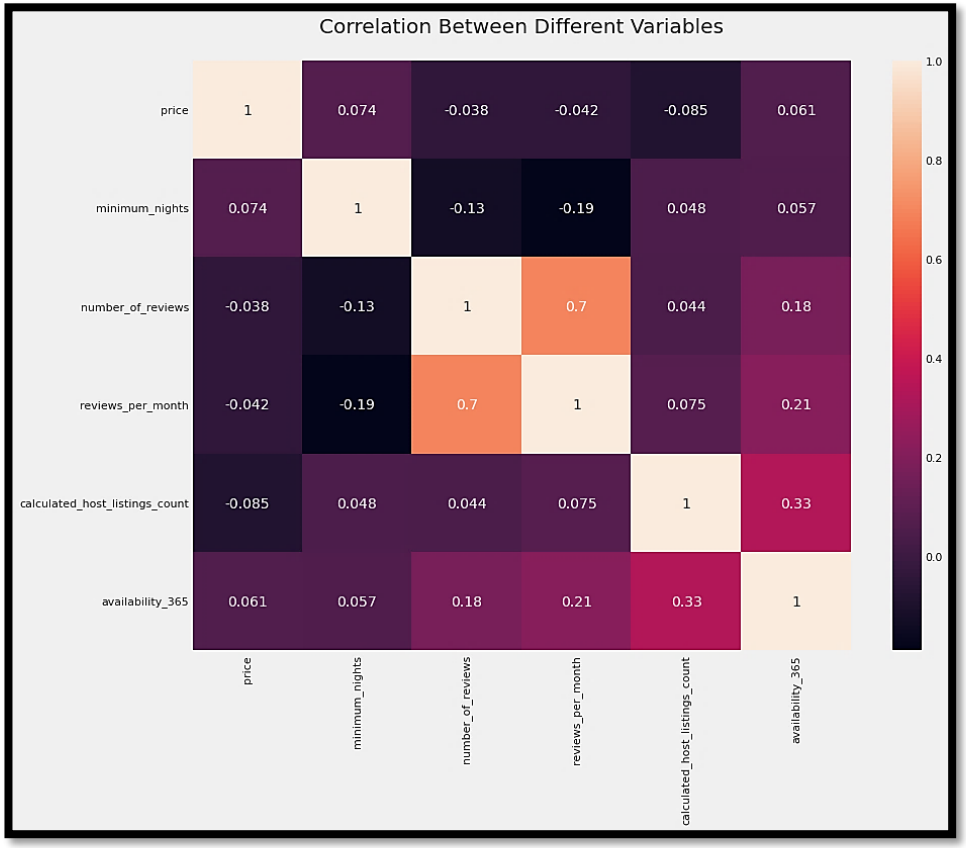
Total no. of listings vs minimum nights!

- So, the graph tells us about the total number of listings providing the total number of minimum nights.
- More than 12000 listings provides a minimum night of 1.



Correlation b/w columns

- We are getting a good correlation between the number of reviews with subsequent variable reviews per month, which is 0.7, but the physical interpretation will give us nothing and hence we have nothing much on this.



Result and Conclusion

After exploring and analyzing through data and performing visualization, we obtained some interesting insights into the Airbnb domain….

1. **"Manhattan" has the most expensive bookings compared to the other neighbourhood group.**
2. **Manhattan is also considered as the best location based on the graph of neighbourhood group vs a number of reviews.**
3. **Busiest Host = the host(host id = 219517861 and his name = Sonder NYC) who has 327 listings' is considered as the busiest host in NYC and he belongs to the Manhattan.**
4. **According to our analysis, we noticed some differences in traffic among different areas. Manhattan, Brooklyn & some parts of Queens have a high traffic of Airbnb bookings.**