**Exploratory Data Analysis on NYC Airbnb 2019 dataset**

**Ajinkya Morade, Nitesh Gajakosh**

**Problem Statement**

Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present a more unique, personalized way of experiencing the world. Today, Airbnb became unique kind services that is used and recognized by the whole world. Data analysis on millions of listings provided through Airbnb is a crucial factor for the company. These millions of listings generate a lot of data - data that can be analysed and used for security, business decisions, understanding of customers' and providers' (hosts) behaviour and performance on the platform,

guiding marketing initiatives, implementation of innovative additional services and much

more.

**This dataset has around 49,000 observations in it with 16 columns and it is a mix between categorical and numeric values.**

Explore and analyze the data to discover key understandings (not limited to these) such as:

* What can we learn about different hosts and areas?
* What can we learn from predictions? (Ex: locations, prices, reviews, etc)
* Which hosts are the busiest and why?
* Is there any noticeable difference of traffic among different areas and what could be  the reason for it?

**Introduction**

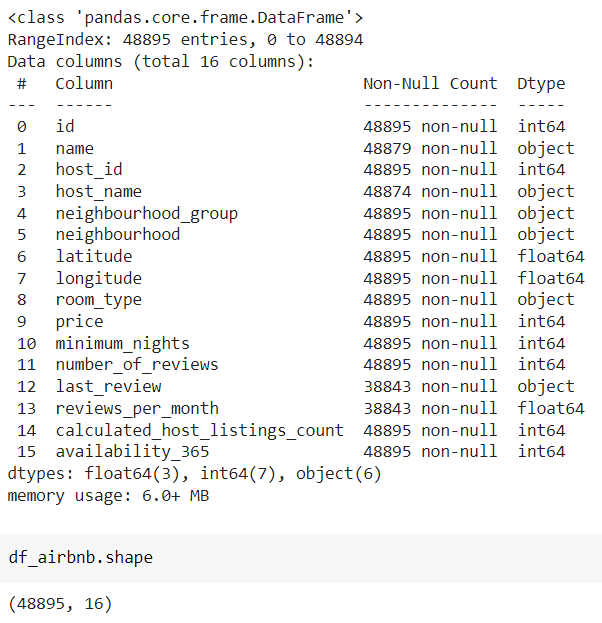
The **Airbnb** story started back in 2008, with the two founders, Joe Gebbia and Brian Chesky (Nathan Biecharczyk was invited to join later), and 3 air mattresses in San Francisco, California. Their problem: they couldn’t pay rent. They set up a simple website with a map and rented out three mattresses and promised home-made breakfast. Three people came and they decided to build on it. They refined the website with money from selling cereal boxes during the presidential campaign in 2008, and took high-quality pictures of the rented space to boost click-rates. In the winter of 2012, Airbnb overtakes Hilton Hotels in nights booked. Now, it has over 1,500,000 listings in 34,000 cities and 190 countries. Bookings can be made online and on mobile devices. In 2015, the value of the company was $25.5 billion.

****

**Methodology**

1. **Data**
   1. **Acquiring a loading Data**

* At first, we imported required libraries in collab environment such as NumPy, pandas, matplotlib, seaborn.
* Then we mounted drive and imported CSV file provided using read\_csv function in pandas.
  1. **Data Exploration and variable Identification**
* In this step we checked information related to data.
* Also, we checked shape of dataset which we found out that this dataset has around 48895 observations in it, with 16 columns (features) and it is a mix between categorical and numeric values.

****

* 1. **Understanding Data**

**Following are the various columns(features) and what they represent**

**'id’: -** This column represents property id.

**'name’: -** This column represents property Name and Description.

**'host\_id’: -** Particular properties were hosted by particular hosts who are represented by host id column.

**‘host\_name’: -** Particular properties were hosted by particular hosts who are represented by host name column.

**'neighbourhood\_group’: -** It represent cities in New York i.e., 'Brooklyn', 'Manhattan', 'Queens', 'Staten Island’, 'Bronx’.

**'neighbourhood’: -** It’s represented particular area in particular city for example, Kensington is one of area in Brooklyn city.

**‘latitude’, ‘longitude’: -** These columns represent location of particular Airbnb listing.

**‘room\_type’****: -** It contain three categorical values i.e., 'Private room', 'Entire home/apt', 'Shared room’ which represent the what is room type.

**‘price’: -** It represent the price of particular room per night.

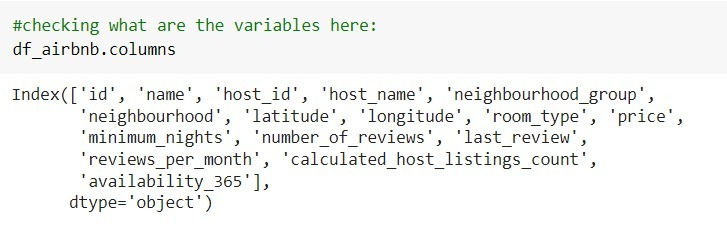
**‘minimum\_nights’: -** Represent minimum night spend by guests in particular host’s listing.

**‘number\_of\_reviews’, ‘last\_review’, ‘reviews\_per\_month’: -** Represents number of reviews, last review, reviews per month of specific listing.

**‘calculated\_host\_listings\_count’:-** Represent total number of times host listed property.

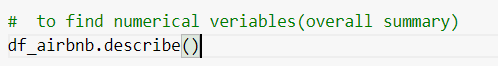
**‘availability\_365’: -** Represent number of days available in year for specific listing.

In this next step we find out various **Categorical** and **Numerical** variables.



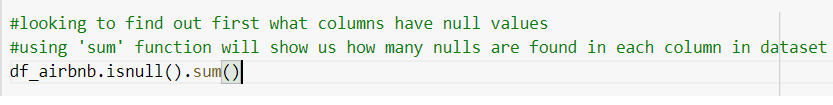
By observations we get to that **'name'** column represents **property name** and particular properties were hosted by particular hosts who are represented by **‘host\_name’** column.

But a particular host\_name can have multiple properties in an area. So, host\_name is like one of the **categorical variables** like **neighbourhood (areas), neighbourhood\_groups, and room\_type**.



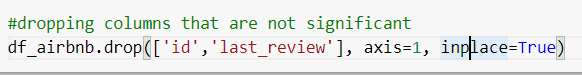
The function used above told us about various **Numerical Variables** in our dataset such as **id, latitude, longitude, price, minimum\_nights, number\_of\_reviews, last\_review, reviews\_per\_month, calculated\_host\_listings\_count, availability\_365**.

* 1. **Handling NaN Values**

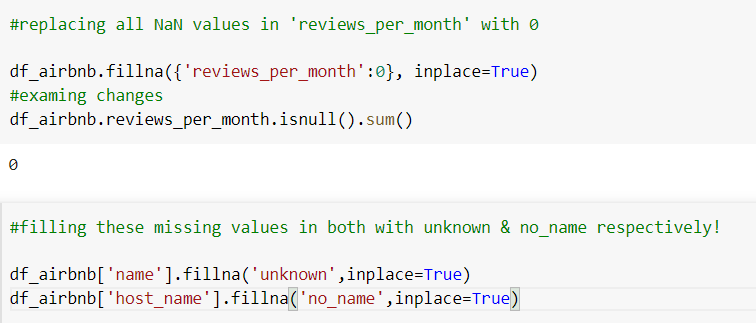
****

This function told us about number of Null values in each column to remove these values we either dropped the column (which are not that much important in our analysis) or replaced NaN values with some relevant substitutes.

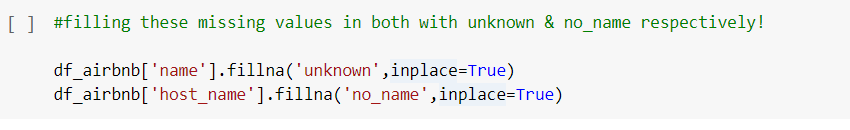
* We got to know that columns such as **'id’, 'last\_review' are of no use for this particular analysis**. To elaborate, "last\_review" is date; if there were no reviews for the listing - date simply will not exist. In our case, this column is irrelevant and insignificant therefore appending those values is not needed.



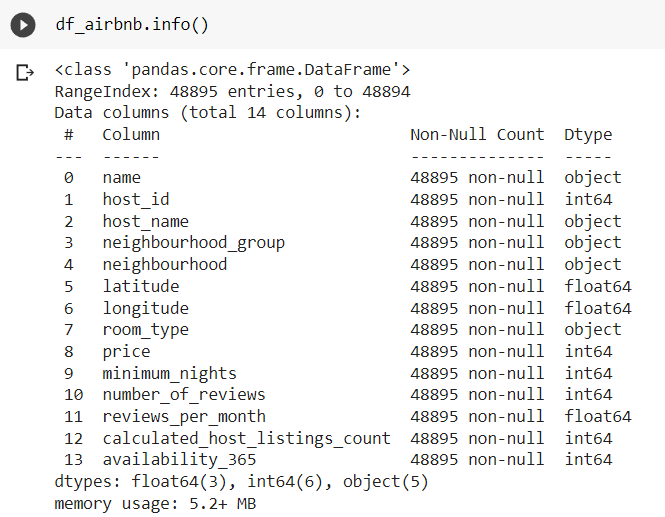
* For "review\_per\_month" column we can simply append it with 0.0 for missing values; we can see that in "number\_of\_review" that column will have a 0, therefore following this logic with 0 total reviews there will be 0.0 rate of reviews per month.



* Also, host\_name and name are not that much important in our analysis so we filled those columns with substitute ‘no\_name’ and ‘unknown’ respectively.



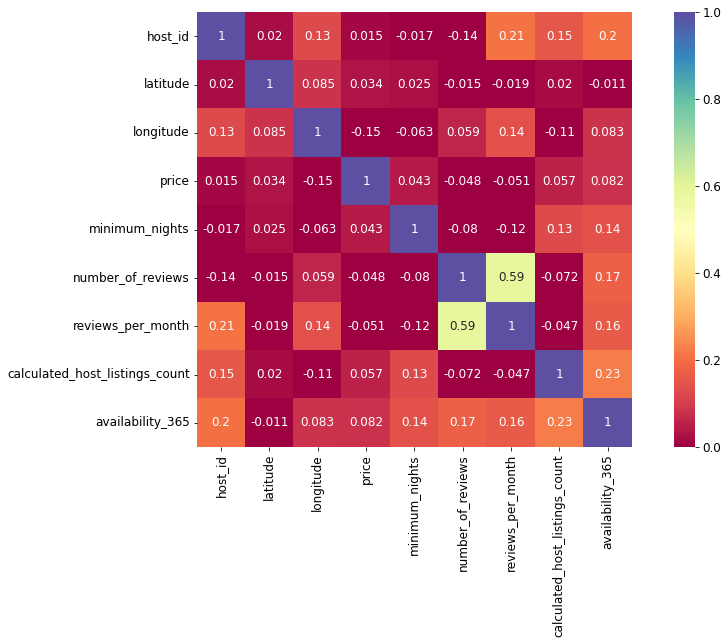
Updated data frame



1. **Exploring and Visualizing Data**

At first, we have to find out about correlation between various features in our dataset.

To find out correlation we used heat map in matplotlib library.



From the heatmap we can see: -

There's correlation among host\_id to reveiws\_per\_month & availability\_365. Also, there is correlation between minimum\_nights to calculated\_host\_listings\_count and availability\_365.

Price also show correlation with availability\_365 and calculated\_host\_listings\_count. no\_of\_reviews and reviews\_per\_month gives almost the same information. so, we can carry out analysis with any of the two variables. Also, no\_of\_reviews column is correlated to availability\_365.

As we got to know about correlation between various features in our dataset and also, we handled NaN values in our dataset, we can dive into analysis of our dataset to find out some important conclusions.

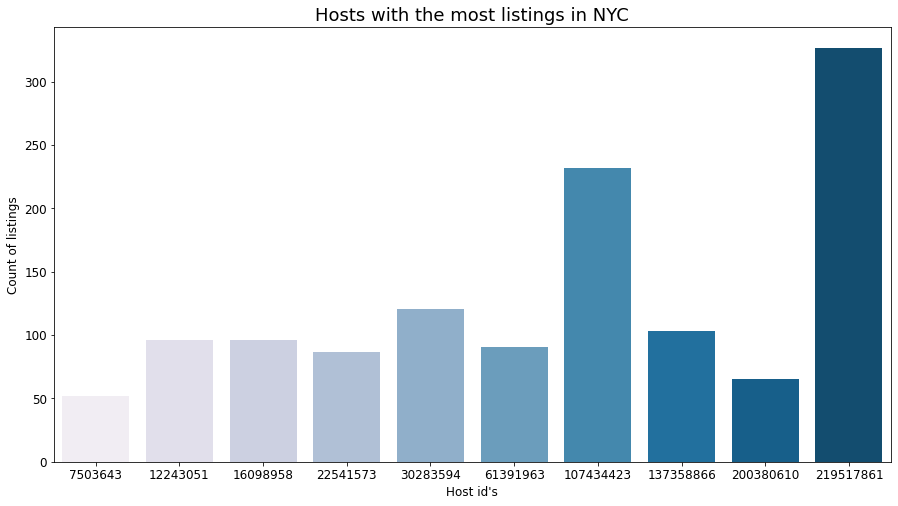
To analyse data, we used two methods Single variable analysis, Bi-variable analysis.

We will elaborate about it in next section

* 1. **Single Variable Analysis**

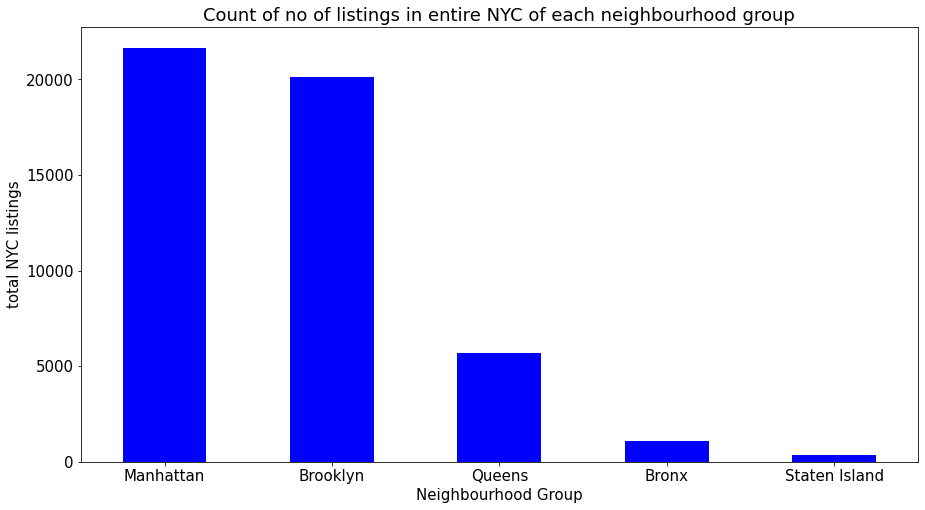
In this analysis we consider single variable against single parameter

1. ***top 10 hosts (IDs) have the most listings***

****

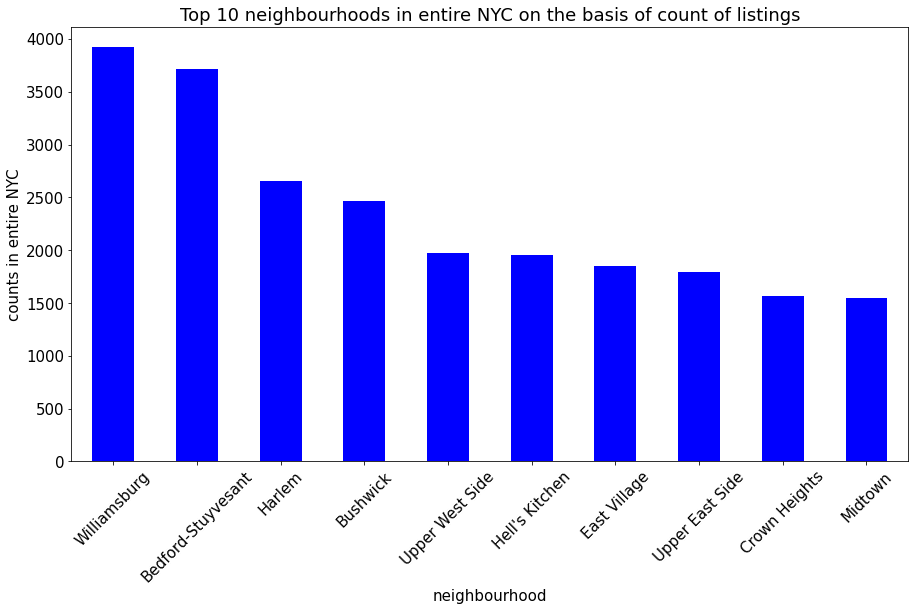
This graph shows which host ID have highest no of listings in this dataset.

1. ***the neighbourhood group vs no of listings in entire NYC***

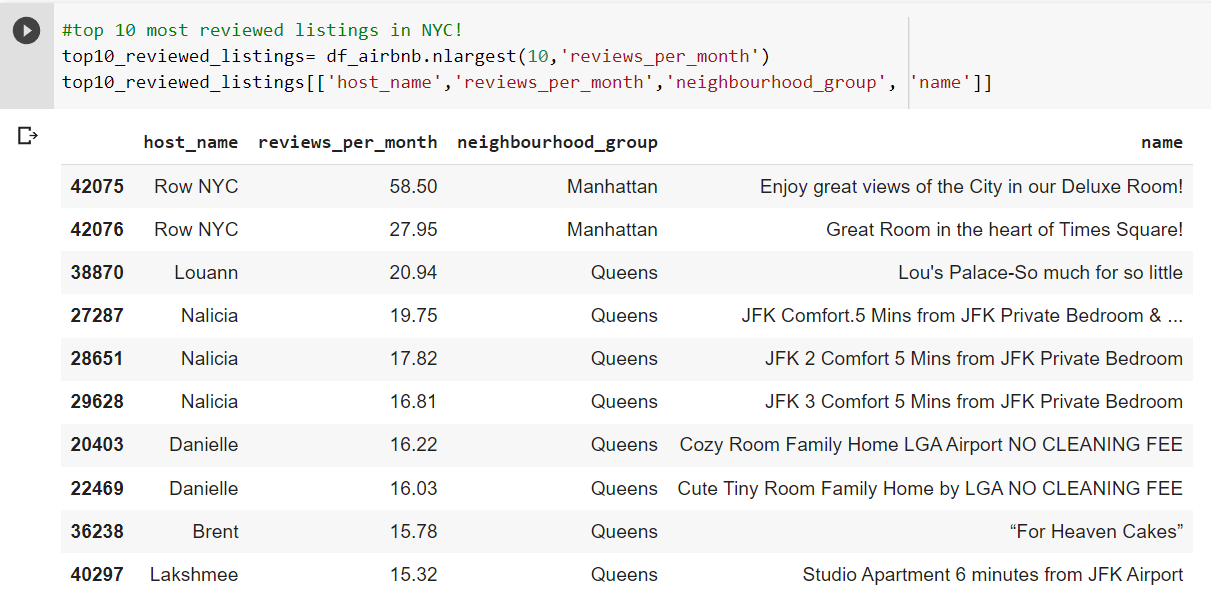


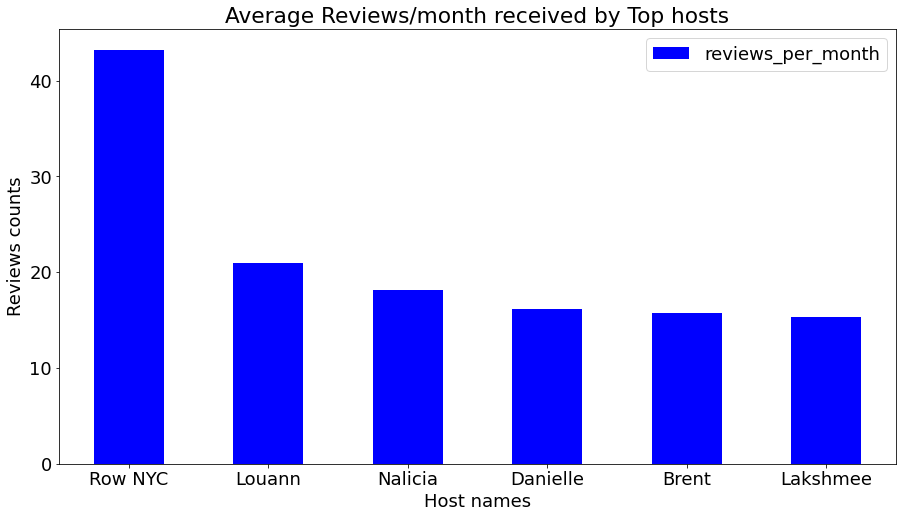
The neighbourhood group 'Manhattan' has the highest no of listings in entire NYC. This is reasonable as Manhattan is one of the main neighbourhood groups in NYC.

1. ***Top 10 neighbourhoods in entire NYC on the basis of count of listings***



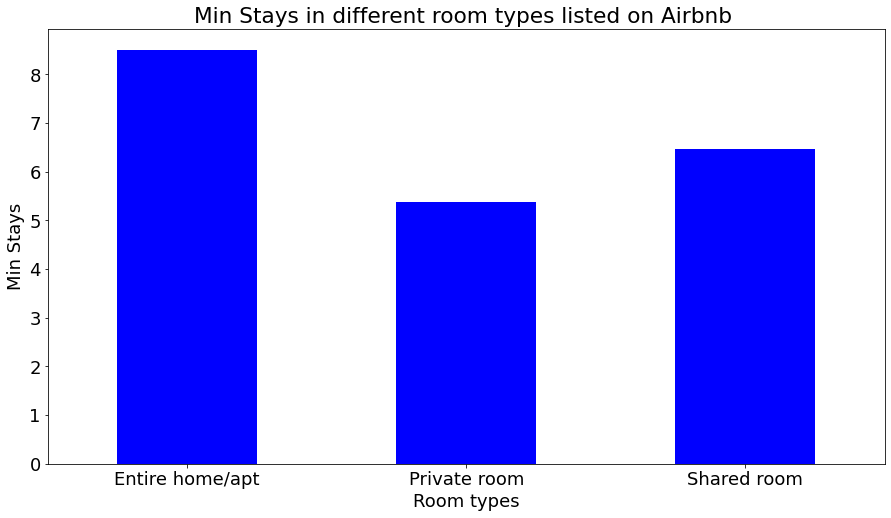
1. ***Average Review/Month received by top 10 hosts***





So, this are the hosts with highest number of reviews, the highest being ROW NYC (Enjoy great views of the city in our Deluxe Room) with 58 reviews/month followed by Louann (Great Room in the heart of Times Square) in the heart of Times Square with a total of 27 reviews/month. Top 2 are from Manhattan so we can decide the best suited place for a comfortable stay but the prices are high to afford.

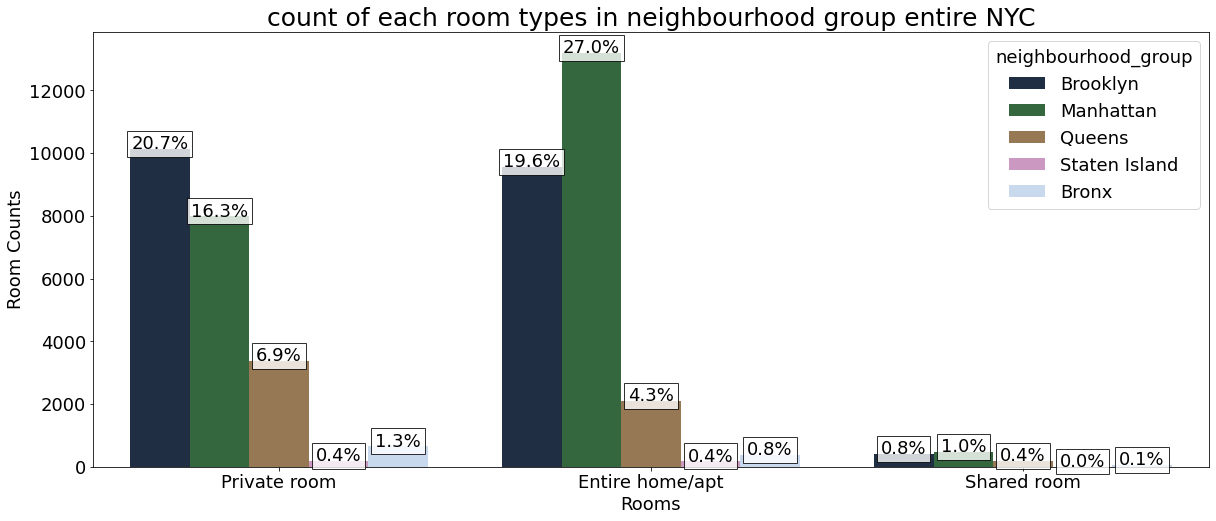
1. ***On an average for how many nights people stayed in each room types***



Looking at these it clearly indicates that people mostly prefer living in an entire home/apt on an average of more than 8 nights followed by guests who stayed in shared room where average stay is 6-7 nights.

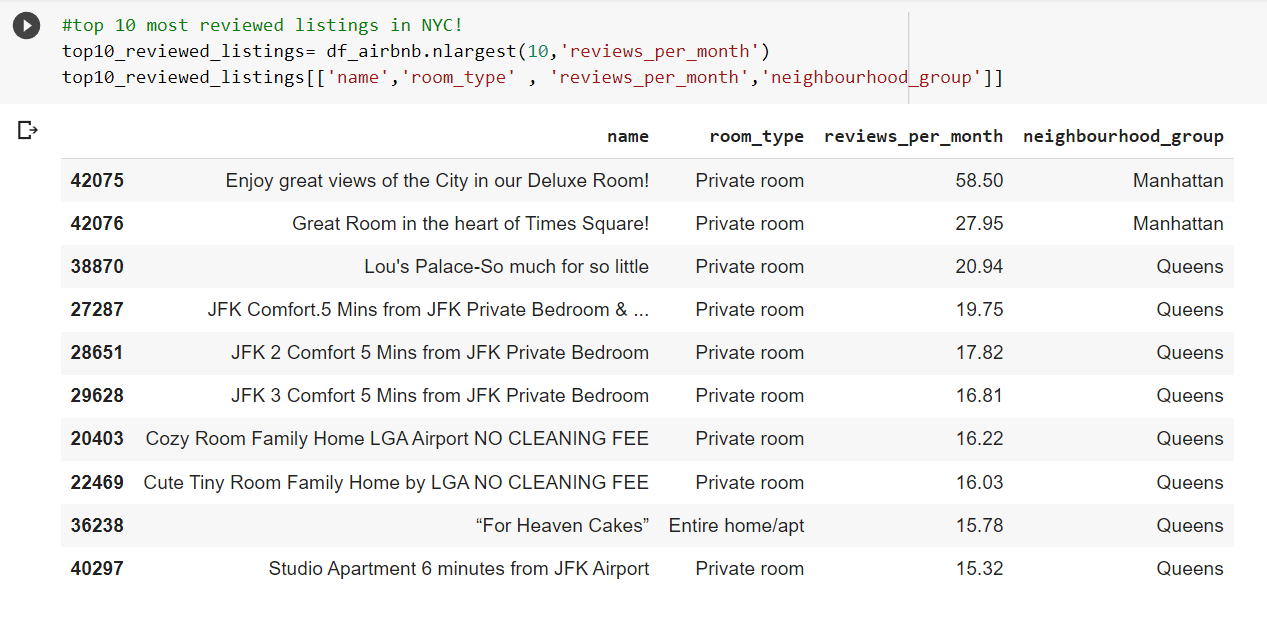
* 1. **Bi-variable Analysis**

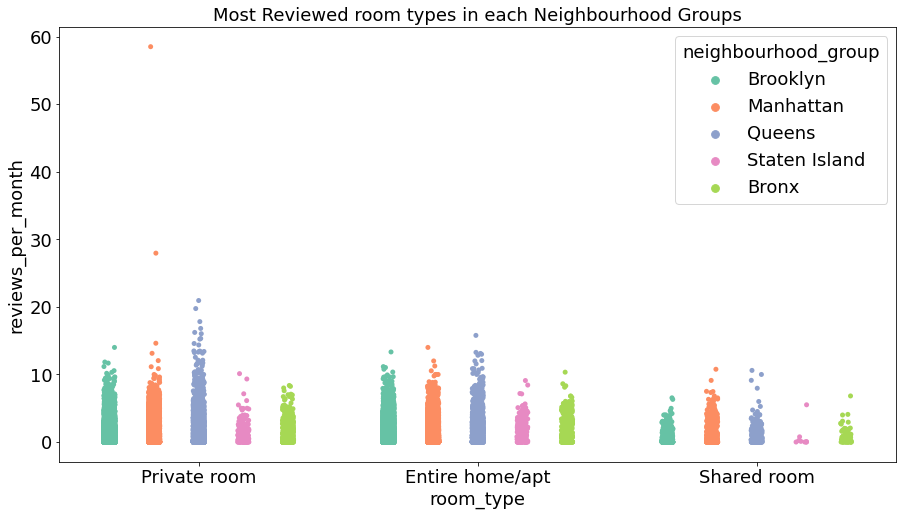
1. ***count of each room types in neighbourhood group entire NYC***

******

Manhattan has more listed properties with Entire home/apt around 27% of total listed propert-ies followed by Brooklyn with around 19.6%. Private rooms are more in Brooklyn as in 20.7% of the total listed properties followed by Manhattan with 16.3% of them. While 6.9% of private rooms are from Queens. Very few of the total listed have shared rooms listed on Airbnb where there's negligible or almost very rare shared rooms in Staten Island and Bronx. We can infer that Brooklyn, Queens, Bronx has more private room types while Manhattan which has the highest no of listings in entire NYC has more Entire home/apt room types.

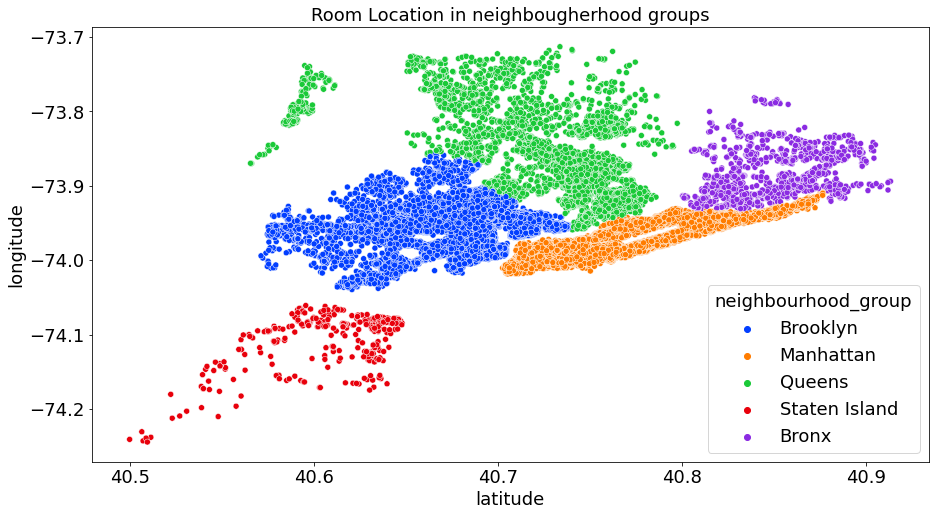
1. ***monthly reviews variations with room types in each neighbourhood groups***

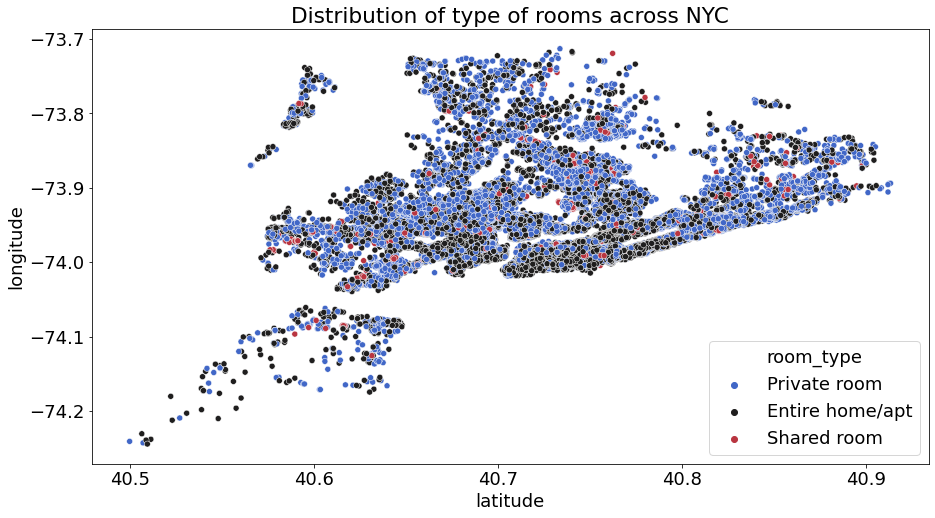
******

******

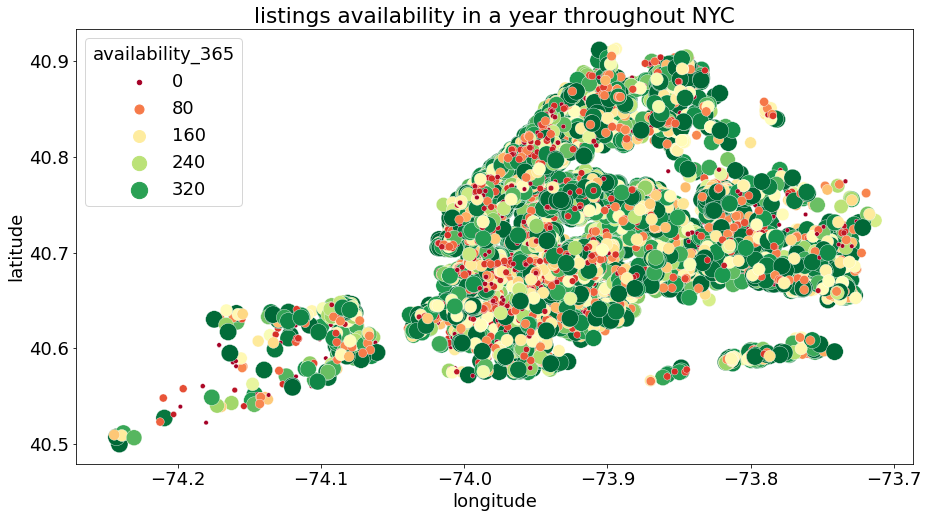
We can see that Private room received the most no of reviews/month where Manhattan had the highest reviews received for Private rooms with more than 50 reviews/month. Manhattan & Queens got the most no of reviews for Entire home/apt room type. There were less reviews received from shared rooms as compared to other room types and it was from Staten Island followed by Bronx***.***

1. ***Room types and their relation with availability and also with different neighbourhood groups***

****

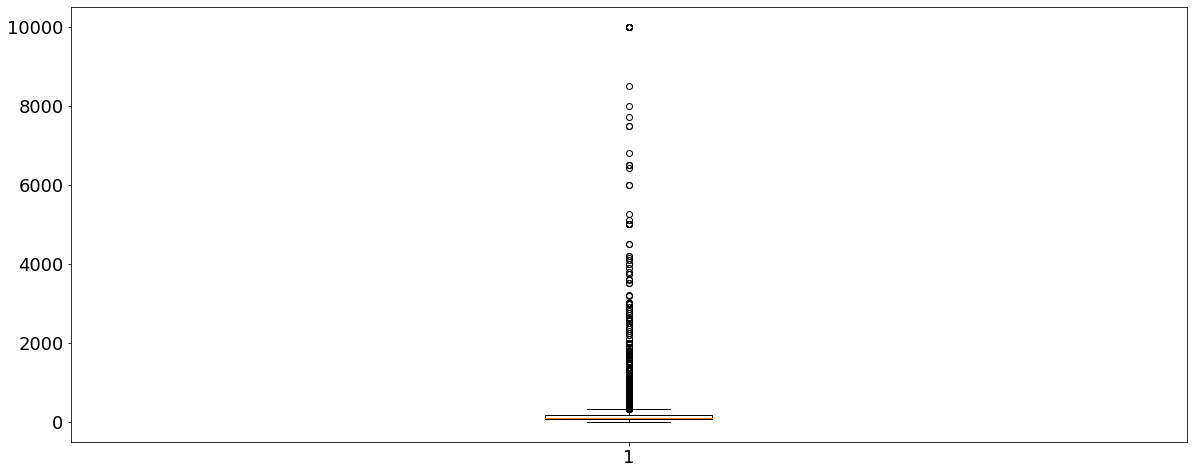
****

By the two scatterplots of latitude vs longitude, we can infer there's is very less shared room throughout NYC as compared to private and Entire home/apt. 95% of the listings on Airbnb are either Private room or Entire home/ apt. Very few guests had opted for shared rooms on Airbnb. Also, guests mostly prefer this room types when they are looking for a rent on Airbnb as we found out previously in our analysis.

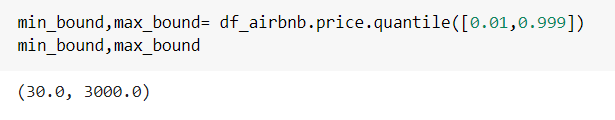
****

Bronx & Staten Island has listings which are mostly available throughout the year, this might be the case as they are not much costlier as compared to other neighbourhood groups such as in Manhattan, Brooklyn & Queens.

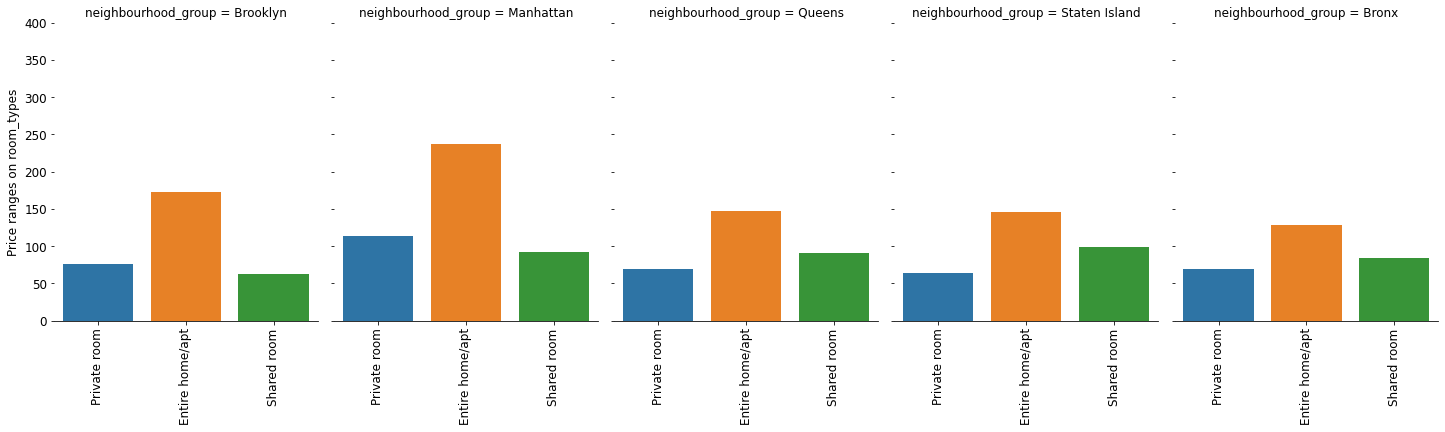
* 1. **Price Feature**
* We used visualization method i.e., **boxplot** to find outliers in price column.



* As, boxplot shows that feature column (i.e., Price) has many outliers. So, we have to remove those outliers for better result.
* So, we use **quantile approach** to remove these outliers. In this technique, the outlier is capped at a certain value above the 90th percentile value or floored at a factor below the 10th percentile value.

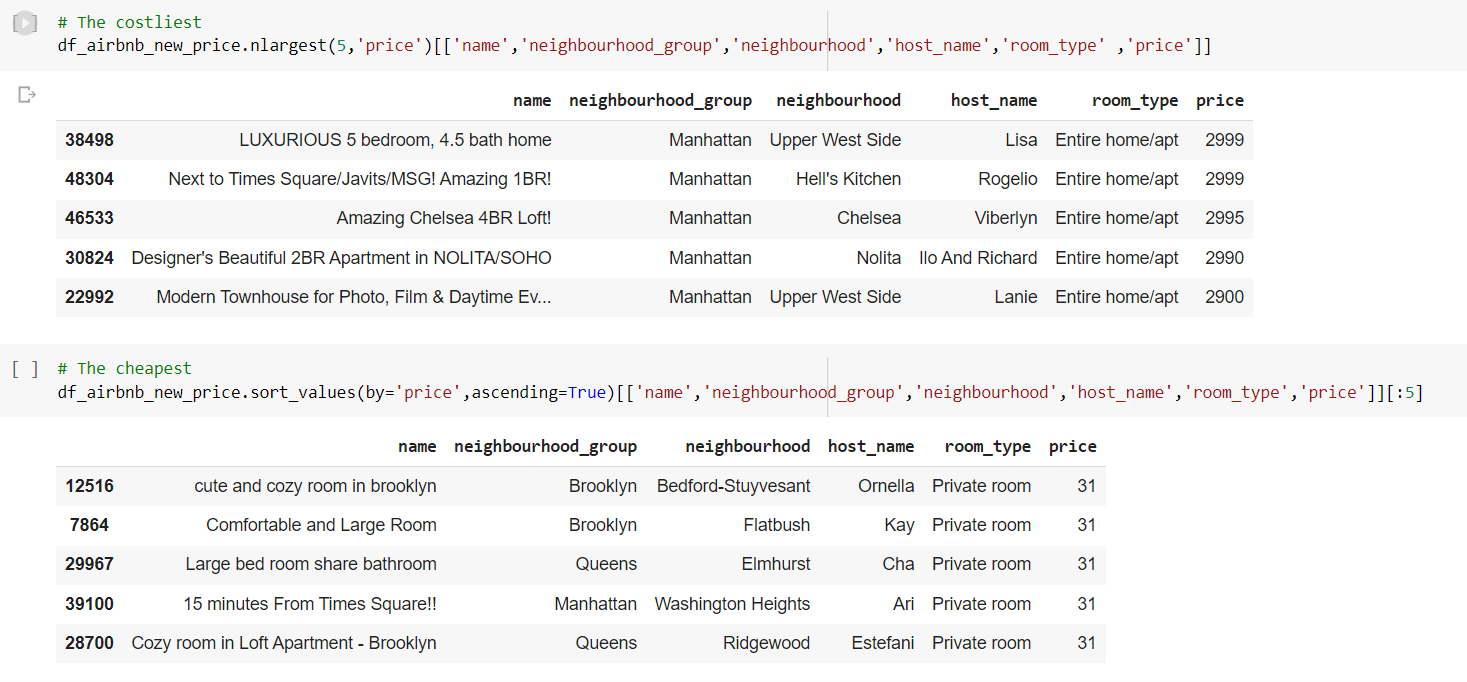


1. ***Room types Vs price in different neighbourhood groups***



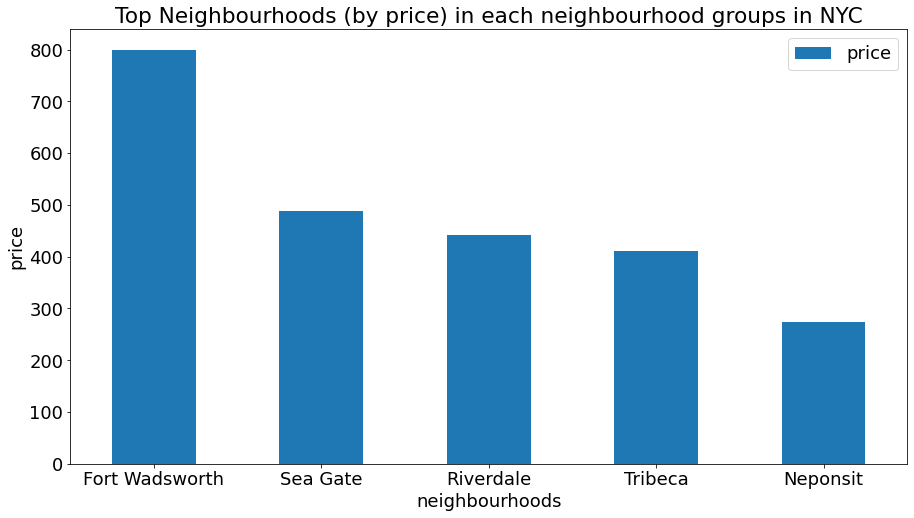
Looks like a property/listing with Entire home/apt as room type wins the show at NYC followed by private rooms. Manhattan has the highest price for room types with Entire home/apt ranging to nearly 240 USD/night, followed by Private room with 110 USD/night. And it's obvious being the most expensive place to live in.

1. ***The costliest and cheapest listings & their respective hosts in entire NYC***



All top 5 listing (by price) are from Manhattan. Cheapest listings mostly come from Brooklyn, Queens and all are private rooms which makes sense.

1. ***Top neighbourhoods’ groups in NYC with respect to average price/day of Airbnb listings***



Among the top neighbourhoods in each neighbourhood groups in NYC with respect to avera-ge price/day of Airbnb listings, top 2 of them namely: Fort Wadsworth & Sea Gate**,** origins  from Staten Island & Brooklyn.

**Conclusion**

* If a person trying to book a listing for stay/rent he/she will look into these following factors while booking: neighbourhood group, neighbourhood, room type, price, number of reviews and availability.
* The neighbourhood group ‘Manhattan’ has highest number of listings in entire NYC. Also top 5 costliest listings are present in it.
* Hillside Hotel is found to have listed more listings in entire NYC, followed by Home away from Home.
* We can infer that Brooklyn, Queens, Bronx has more private room types while Manhattan which has the highest no of listings in entire NYC has more Entire home/apt room types.
* Host named as Row NYC holds the title as the most reviewed host with more than 40 reviews/month on average.
* We can see that Private room received the most no of reviews/month where Manhattan had the highest reviews received for Private rooms. Manhattan & Queens got the most no of reviews for Entire home/apt room type. There were less reviews received from shared rooms as compared to other room types and it was from Staten Island followed by Bronx.
* Also, guests mostly prefer this room types when they are looking for a rent on Airbnb as we found out previously in our analysis.
* Manhattan has the highest price for room types with Entire home/apt ranging to nearly 240 USD/night, followed by Private room with 110 USD/night. And it's obvious being the most expensive place to live in.