**Exploratory Data Analysis on Airbnb Bookings**

**Sumit Ghanghas**

**Data Science Student**

**Alma Better**

**CONTENTS**

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

**Introduction:**

Airbnb is an American Company since 2008, it is an online marketplace that connects people who want to rent out their homes with people who are looking for accommodations in specific locales.

The dataset from Airbnb is based on NY. NY is amongst the most expensive places to live in the USA. We would like to perform an in-depth analysis on one of the most densely populated cities of the world. Our dataset is feature rich containing, location with coordinates, prices, host name, room types, availability throughout season.

With these features we’ve done exploratory data analysis and tried to extract information like most expensive places to live in NY, is location really varies with occupancy rate, what type of room people tends to choose most, is there any particular season for tourists or locale when we can follow a surge in prices or occupancy rate of properties etc.

**Problem Statement:**

Data analysis on millions of listings provided by Airbnb is a main factor for the company. These millions of listings have 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 etc.

* With help of python data visualisation, 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 neighbourhood, neighbourhood\_group
* Why type of reviews is made by the most of costumer’s
* Limitation in Airbnb’s data
* Scope of Improvement (How we help to resolve the problem)

**Dataset Analysis ( variable name ):**

The dataset contains 48895 observations with 16 features. This data file includes all needed information to find out more about hosts, geographical availability, necessary metrics to make predictions and draw conclusions. Let us look through our features,

* **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 neighbourhood group present in the particular city ( ex :- NYK , San Francisco etc. )
* **Neighbourhood :-** It tells the neighbourhood present in neighbourhood\_group in the city.
* **Latitude :-** it gives the coordinate of listing
* **Longitude :-** it gives the coordinate of listing
* **Room\_type :-** it tells the type to categorise 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
* **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

latitude and longitude have represented a coordinate, neighbourhood\_group, neighbourhood and room\_type are columns of categorical type. last\_review is a column of date type; we will convert it as required.

The distribution of numerical columns are as follows,

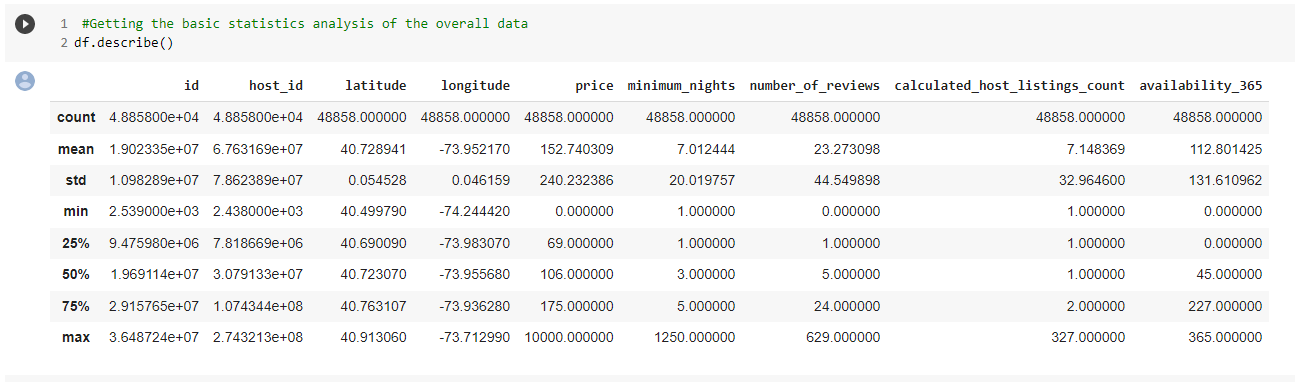
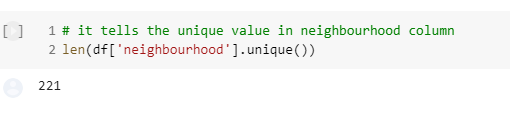


Fig 1. Statistical Distribution of Numerical Features

Other 3 important columns are,

* **neighbourhood\_group:** It contains 5 unique hoods which are Manhattan, Brooklyn, Queens, Bronx & Staten Island.
* **neighbourhood:** It contains 211 unique neighbourhoods.

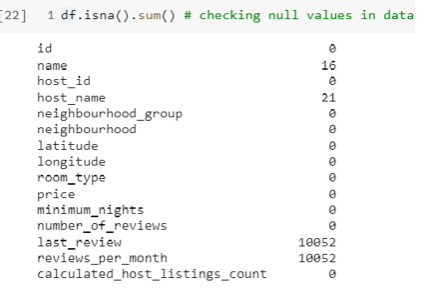
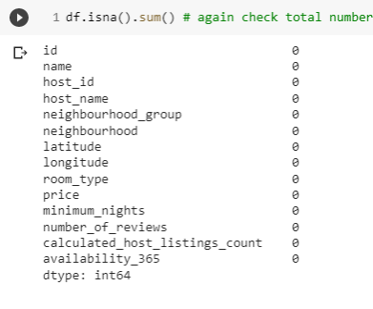


* **room\_type:** It contains 3 unique room types which are Entire home/apt, Private room, Shared room

**Data Cleaning :**

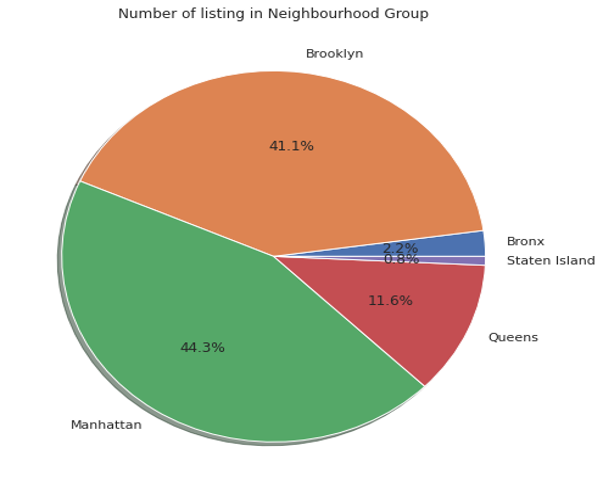
**Fixing the null values**

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 )



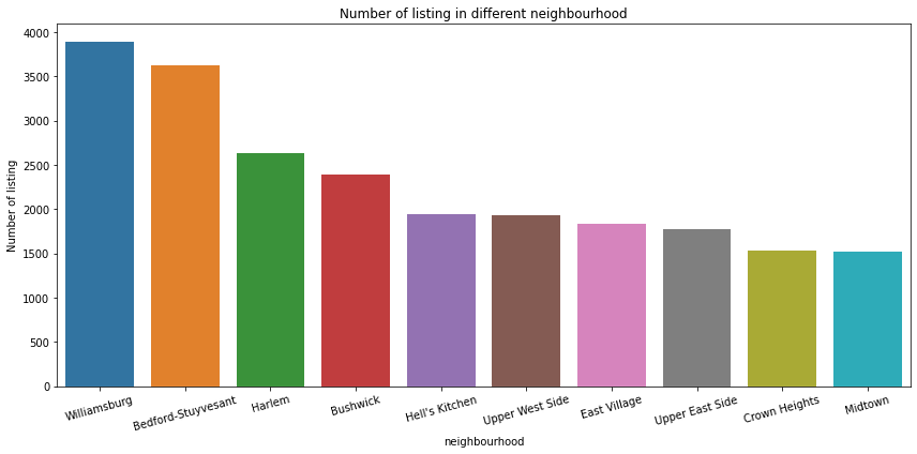
* The highest missing value is in reviews\_per\_month and last\_reviews column, I think this is because most people don't write the reviews . host\_name and name also have the missing value this may be because of some technical issue.
* We can handle the missing value in review\_per\_month by default value.
* We can handle the missing value in name and host\_name only by dropping them.

**No of list made by host across Neighborhood Group**



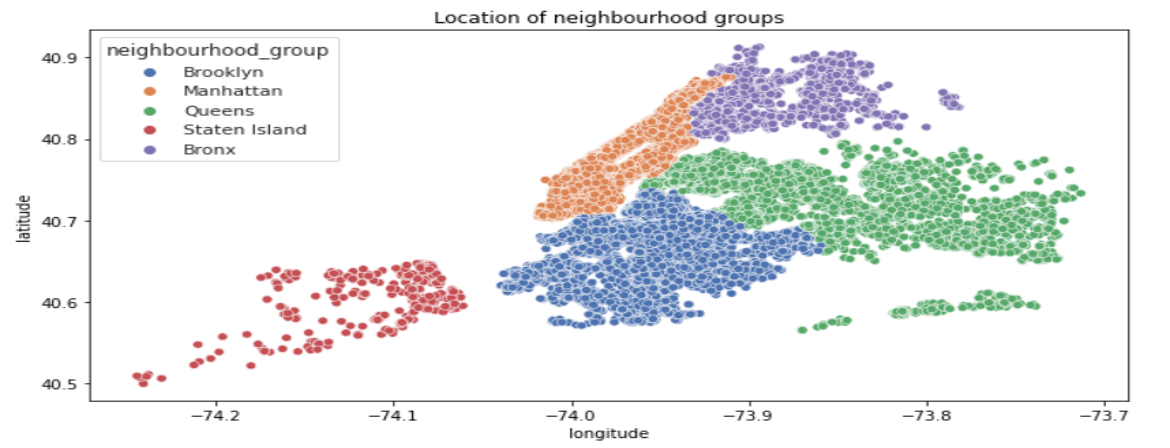
* It is observed that Manhattan has the highest number of listings of 21661 which is 44.3% of total listing.
* Brooklyn has the second highest number of listings in 20104 which is 41.1% of total listings.
* Queens are at third place with 5666 listing and Bronx and Staten have least number of listing
* 85% of Airbnb are distributed in Manhattan and Brooklyn whereas other very less Airbnb locations.
* Most of the people stay in Manhattan and Brooklyn.

**No of list made by host across Neighbourhood**

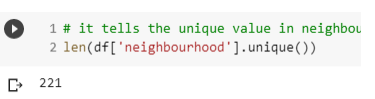


* Most of the Airbnb locations are available in this top 15 neighbourhood and people prefer to stay here.
* The top 2 neighbourhood which is most prefer by people is Williamsburg and Bedford-Stuyvesant
* Most of peoples wants to live in Williamsburg and Bedford-Stuyvesant

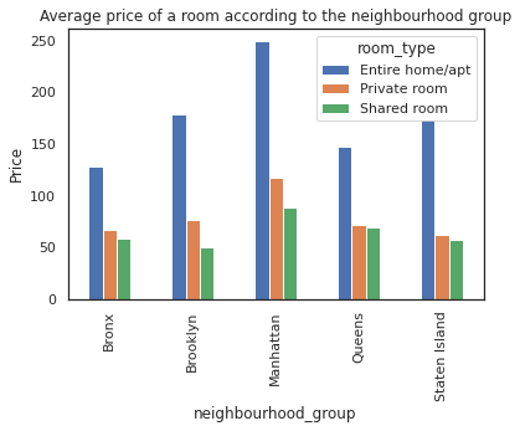
**Location of Neighbourhood Groups**



**There are 221 unique neighbourhoods falls under 5 groups**

**Price distribution across neighbourhood\_group**

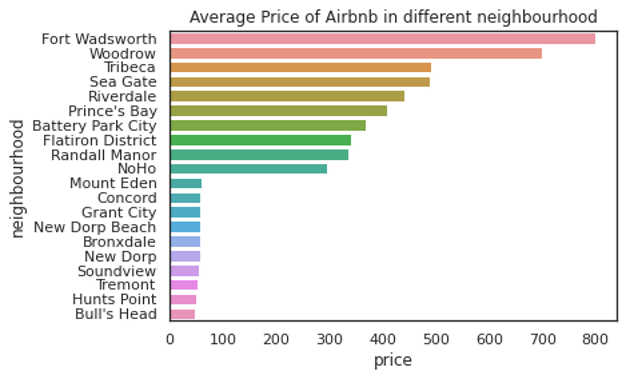
**( which room type are most expensive and where it is located )**



* 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 obviously the most expensive place to live in!
* In Manhattan you opt for an entire home 40% more than you opt for an entire home in Brooklyn.
* location wise distribution of price shows that Manhattan has expensive and Bronx has low priced rooms.
* If you opt for Bronx you have to pay 50% less than Manhattan

. **Price distribution across neighbourhood**

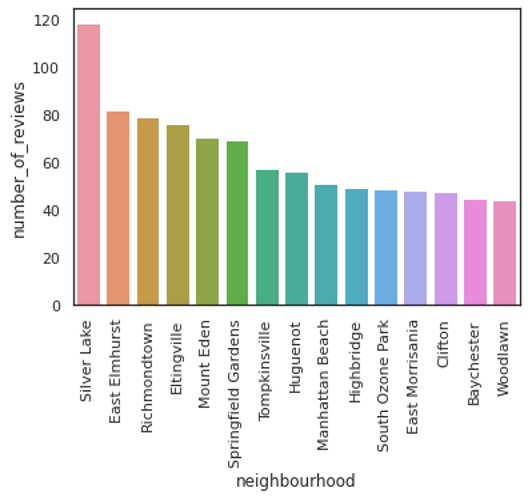
**( which room type are most expensive and where it is located )**



* Most expensive neighbourhood is fort Wadsworth followed by Woodrow then Tribeca.
* The cheapest neighbourhood is hunt point followed by bull's head then Soundview.

**Popular neighbourhood by review**

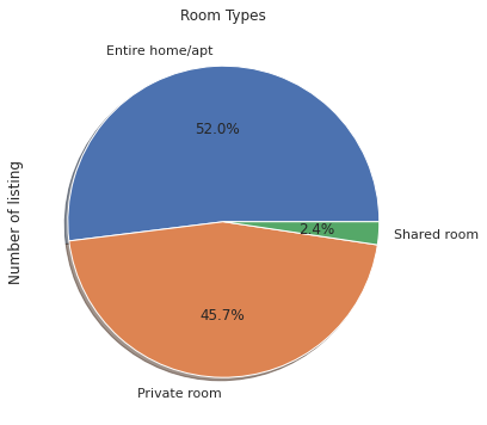
**( which room type are most and where it is located )**



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

**Preferred room types**

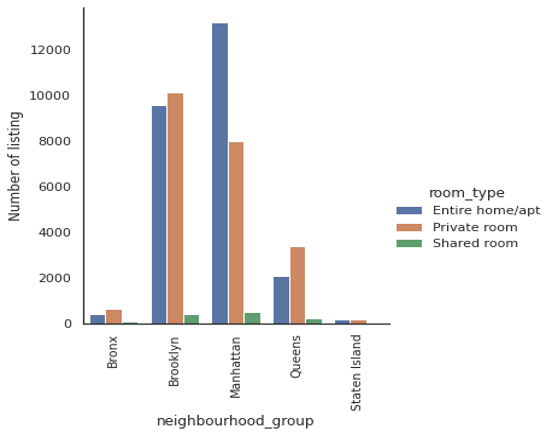
**( which room type people are preferred to stay longer )**



* The demand for an entire home and private room is more high and people also choose an entire home and private room.
* As per the dataset 97.7% of them are entire home or private rooms and only 2.4% of them are shared rooms.

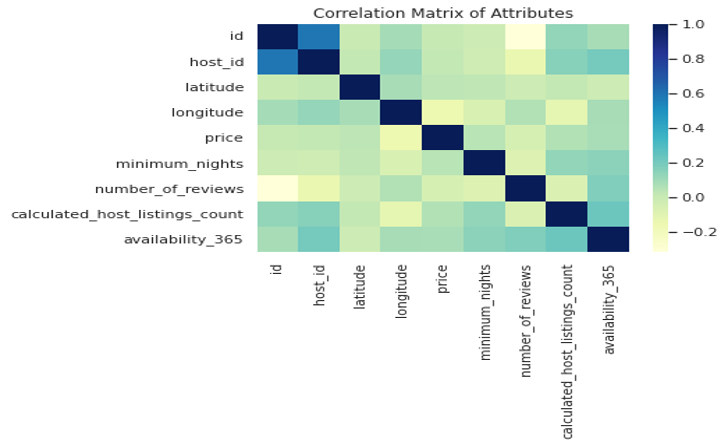
**Preferred room types**

**( 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**

**Limitation:**

* 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.

THE END