Capstone Project - 1

Exploratory Data Analysis on Airbnb Bookings

Anas Malik
Data Science Enthusiast, Almabetter

Table of Content

- Problem Statement
- Data Analysis
- 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.

I have the dataset from Airbnb in New York City. New York is one of the most expensive cities to live in the USA. I loved this project because it gave me an opportunity to analyze in-depth on one the most densely populated cities in the world. The dataset contains id, name, host_id, host_name, neighbourhood_group, neighborhood, latitude, longitude, room_type, price , minimum_nights, number_of_reviews, last_review, reviews_per_month, calculated_host_listings_count and availaibility_365. Among these, I tried to get information like which place is most expensive to live in, what type of rooms people prefer to stay in, and which hosts have the highest listings.

While taking care of the null values, I performed various tasks to achieve answers for the required questions. I created a copy of the dataframe so that changes should not reflect inside the main dataframe. I analyzed the data and can conclude that Manhattan, Brooklyn and Queens are the most expensive areas in New York. People tend to prefer Entire home/apt and private rooms. Most of the people prefer to go with the cheaper prices but still they do not prefer shared rooms.

Introduction to Dataset

The dataset contains 48895 observations with 16 columns. This dataset is all about the booking information on Airbnb. Let us look at the columns that are present in the dataset.

- **id** - a unique id for each property listing.
- **name** - name of the property listed.
- **host_id** - a unique id for registered hosts(providers).
- **host_name** - the name for the registered hosts.
- **neighbourhood_group** - group of areas present in NYC.
- **neighborhood** - area under neighborhood_group.
- **latitude** - coordinates of latitude for listing.
- **longitude** - coordinates of longitude for listing.
- **room_type** - type of the room listed.
- **price** - price of listings.
- **minimum_nights** - the minimum number of nights people stayed.
- **number_of_reviews** - the number of reviews from people.
- **last_review** - It is the date when property was last reviewed.
- **reviews_per_month** - number of reviews per month.

- **calculated_host_listings_count** - number of listings from each host.
- **availibility_365** - It is the number of days property available in a year.

Problem Statement

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 analyzed and used for security, business decisions, understanding of customers' and providers' (hosts) behavior and performance on the platform, guiding marketing initiatives, implementation of innovative additional services and much more.
We need to explore and analyze the data to discover key understandings (not limited to these) such as:

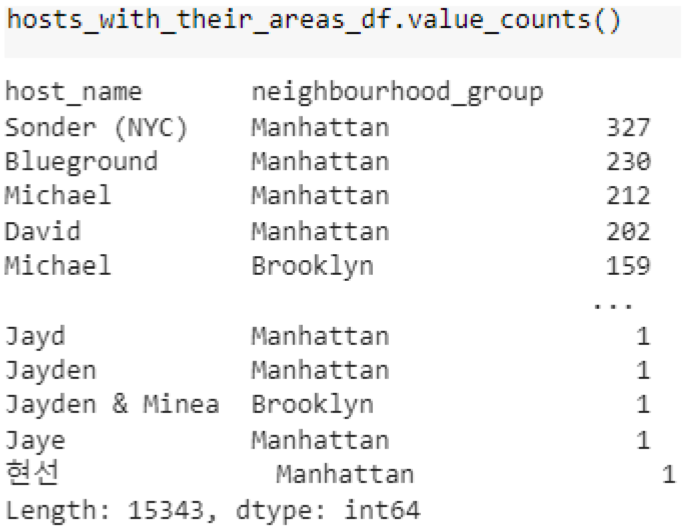
• What can we learn about different hosts and areas?
• What can we learn from predictions? (ex: locations, prices, reviews)
• Which hosts are the busiest and why?
• Is there any noticeable difference of traffic among different areas and what could be the reason for it?

Data Analysis

Q1. What can we learn about different hosts and areas?

To analyze the hosts who have listed the maximum number of listings in which areas, I first take care of the null values from the host_name column.

I created another dataframe with two columns, i.e, host_name and neigbourhood_group, where all the values are not null. Then I counted the values from both columns where the host name and area are same.



```
hosts_with_their_areas_df.value_counts()

host_name        neighbourhood_group
Sonder (NYC)     Manhattan              327
Blueground       Manhattan              230
Michael          Manhattan              212
David            Manhattan              202
Michael          Brooklyn               159
                                       ...
Jayd             Manhattan                1
Jayden           Manhattan                1
Jayden & Minea   Brooklyn                 1
Jaye             Manhattan                1
현선               Manhattan                1
Length: 15343, dtype: int64
```

<u>Observation:</u>

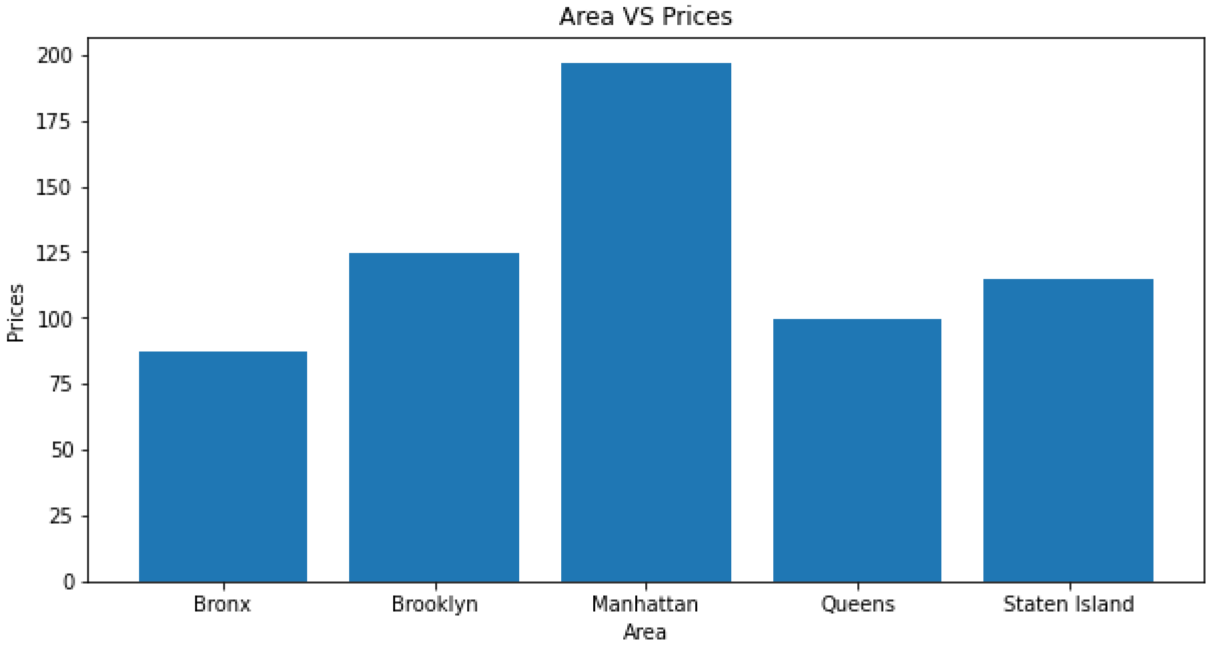
1. About the Areas:
   - Manhattan has the highest number of listings.
   - Brooklyn has the second highest number of listings.
2. About the Hosts:
   - The hosts who have done most number of listings are:
     1. Sonder (NYC).
     2. Blueground.
     3. Michael
     4. David

Q2. What can we learn from predictions? (ex: locations, prices, reviews)

To deal with this problem, I divided this part into three separate parts which are as follows:

➢ **First analyze the area with their mean prices:**

At first, I created a dataframe by doing a groupby operation on the neighbourhood_group column which does a grouping on different areas in NYC. Then, I took the mean of all the rent prices in those areas.

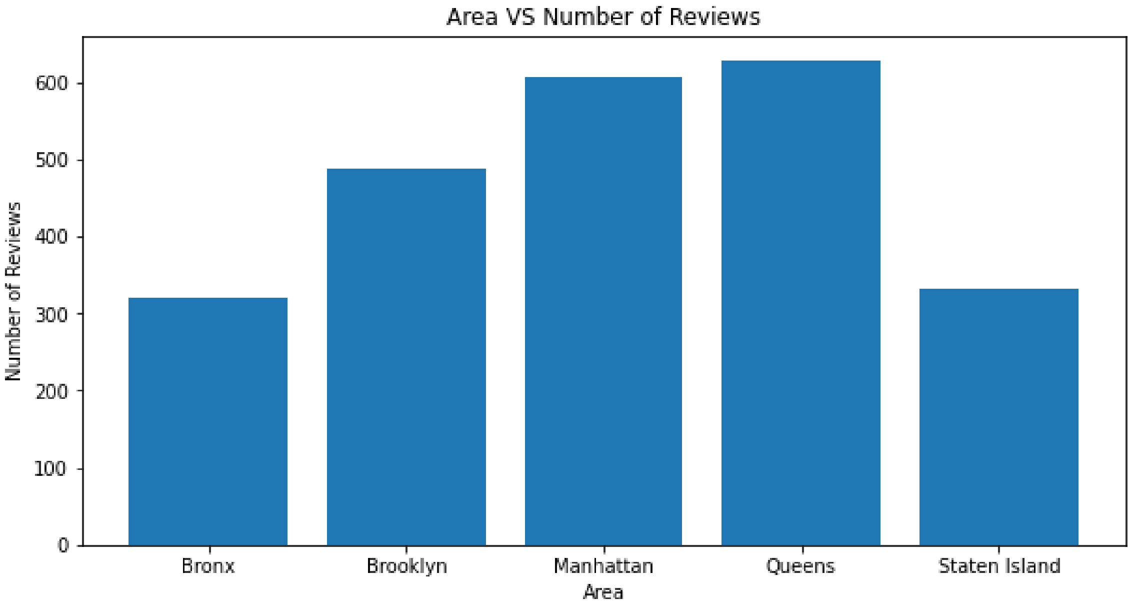


Observation:

- The average rent price in Manhattan is the highest followed by Brooklyn and Queens.
- The Bronx and Staten Island have the cheapest price for rent.

➢ Now Analyzing the areas with the people reviews:

To analyze this, I firstly look over the columns which can be very helpful to get the answer. I found that the two columns ,i.e., neigbourhood_group and number_of_reviews could be very helpful for the same. So, I again done a groupby operation on the same column and then from the number_of_reviews column took the maximum number of reviews for a particular area.

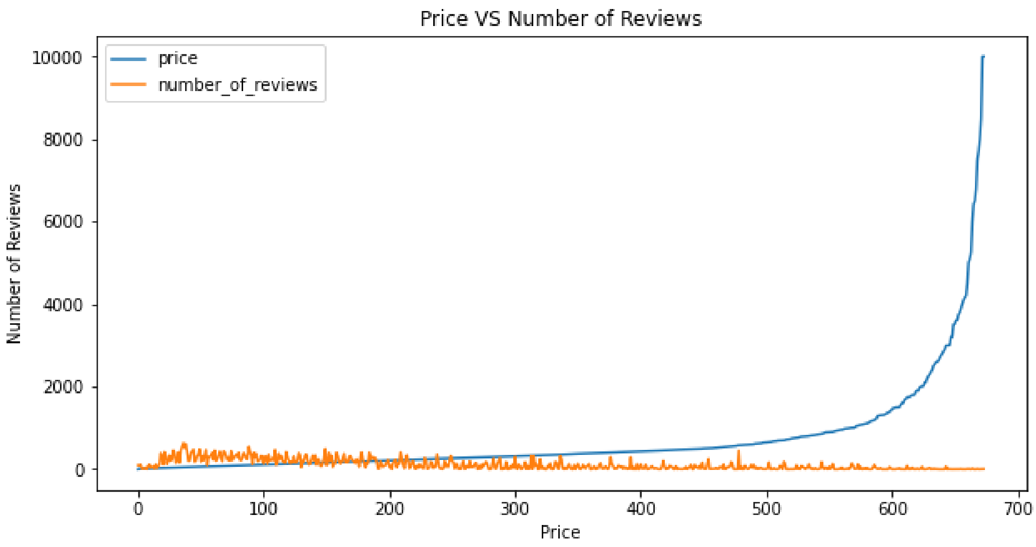


Observation:

- Queens have the most number of reviews from people followed by Manhattan and Brooklyn.
- Staten Island and the Bronx have received the lowest number of reviews.

➢ **Now Analyzing the property prices with number of reviews:**

While analyzing this problem, I personally enjoyed it very much because the outcome is something I expected desperately.
For this, I did a groupby operation on price and took the maximum for the number-of_reviews from the people.



Observation:

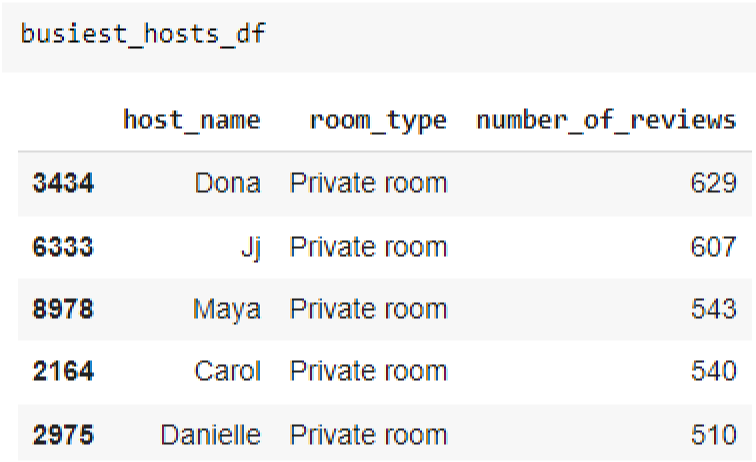
- More people have given their reviews where the rent prices are low.
- The number of reviews decreases as the rent price increases.
- Therefore, It can be concluded that people prefer cheaper rent prices for their stay.

Q3. Which hosts are the busiest and why?

To analyze this, I looked at the columns which could be very helpful and found some columns ,i.e., host_name, room_type and number_of_reviews.

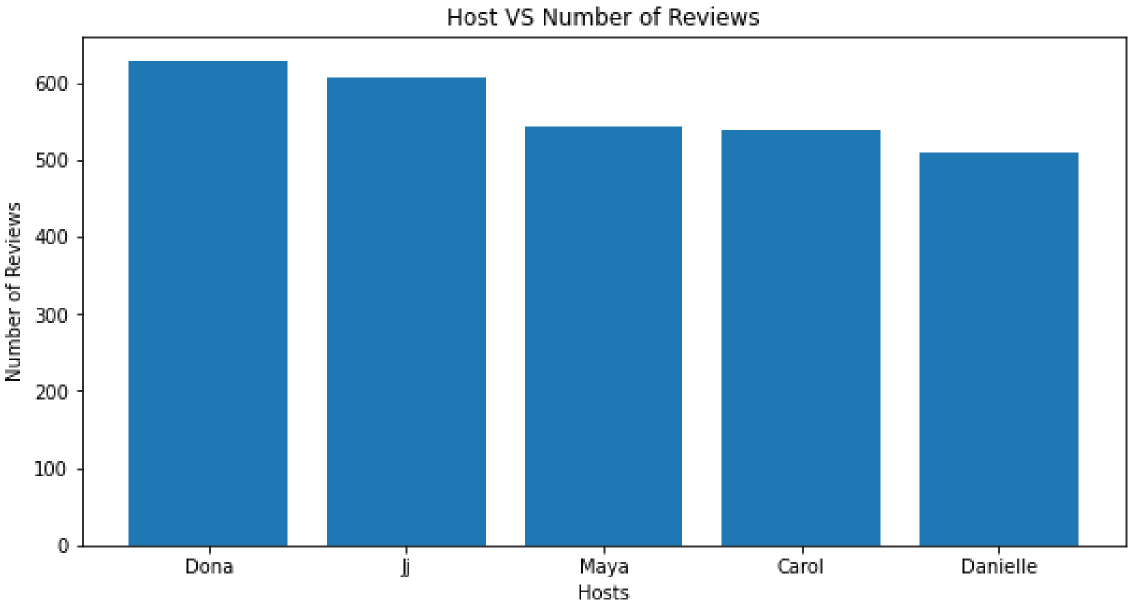
➢Lets first analyze the host with more number of views:

After that, I created a dataframe by grouping on host_name and room_type and then took the maximum number of reviews.



```
busiest_hosts_df
```

	host_name	room_type	number_of_reviews
3434	Dona	Private room	629
6333	Jj	Private room	607
8978	Maya	Private room	543
2164	Carol	Private room	540
2975	Danielle	Private room	510



Observation:

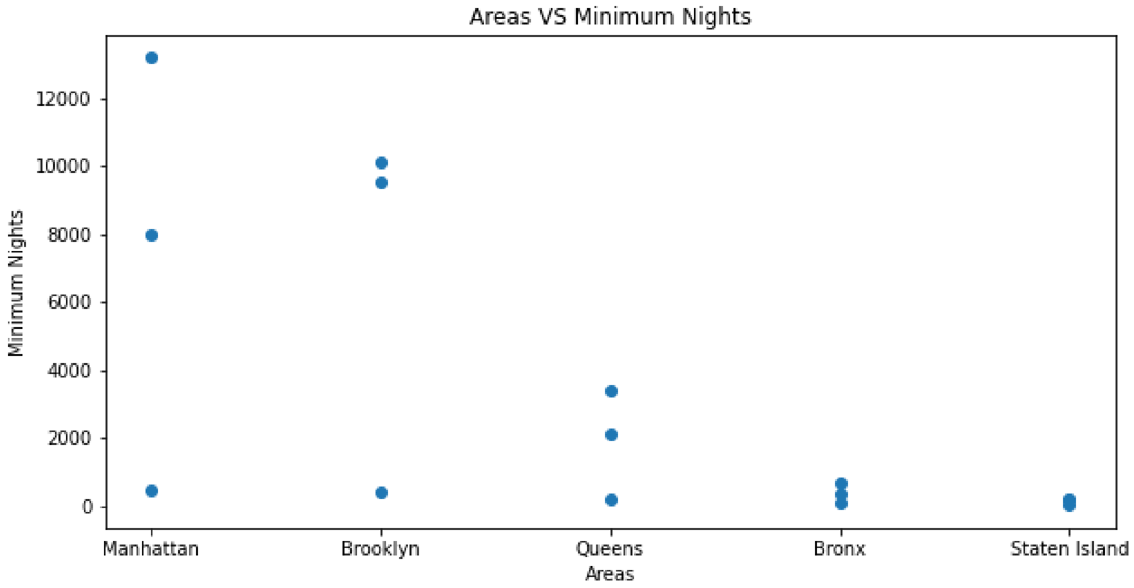
- The more reviews for a host, more will be its people's favorite or demanded.
- Dona has the most number of reviews and then followed by others.
- The common thing among these hosts is that they have listed more private rooms.

Q4. Is there any noticeable difference of traffic among different areas and what could be the reason for it?

This part of the problem made me think more to grab the answer for the same and honestly this was very time consuming for me. So I divided this part also into two different parts which are as follows:

- **Lets first analyze the minimum nights in particular area:**

  At first, I created a separated dataframe by grouping on neigbouhood_group, room_type and then counted the maximum number of nights people stay in particular area.

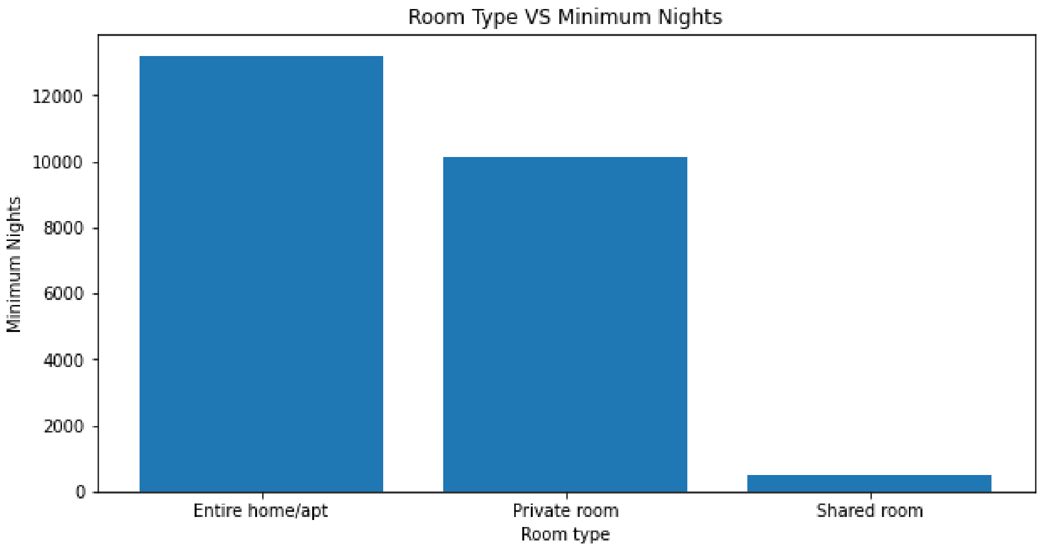


<u>Observation:</u>

- Although Manhattan has the highest rent price, people stay longer in Manhattan followed by Brooklyn and Queens.
- There can be the reason that these areas have more public attractions or tourist places or that they are well developed cities people travel for their work purpose.
- People stay in the Bronx and Staten Island for a short period as compared to other areas.

- **Now let's analyze the minimum nights in particular room type:**

  I got the answer for this question from the same dataframe where counting for each type of property clears that which room_type is most preferable by people.



Observation:

- There are a total three categories for room type.
- People prefer to rent a Entire home/apt for longer duration of stays.
- For a short duration, people go with the private rooms.
- There are very few people who prefer shared rooms.

Conclusion

- Most of the people don’t want to spend more on the rent price. They prefer listings with lower prices.
- For Longer Duration of stays, people prefer to opt for an Entire home/apt room type while for shorter stays, they prefer private rooms.
- In NYC, Manhattan has a large number of listings and it is very expensive to live in the USA.
- Staten Island and Bronx have the lower number of listings and also lower prices. Hence, living in these areas is affordable.
- Location of property has high dependency on deciding its price.

Thank You