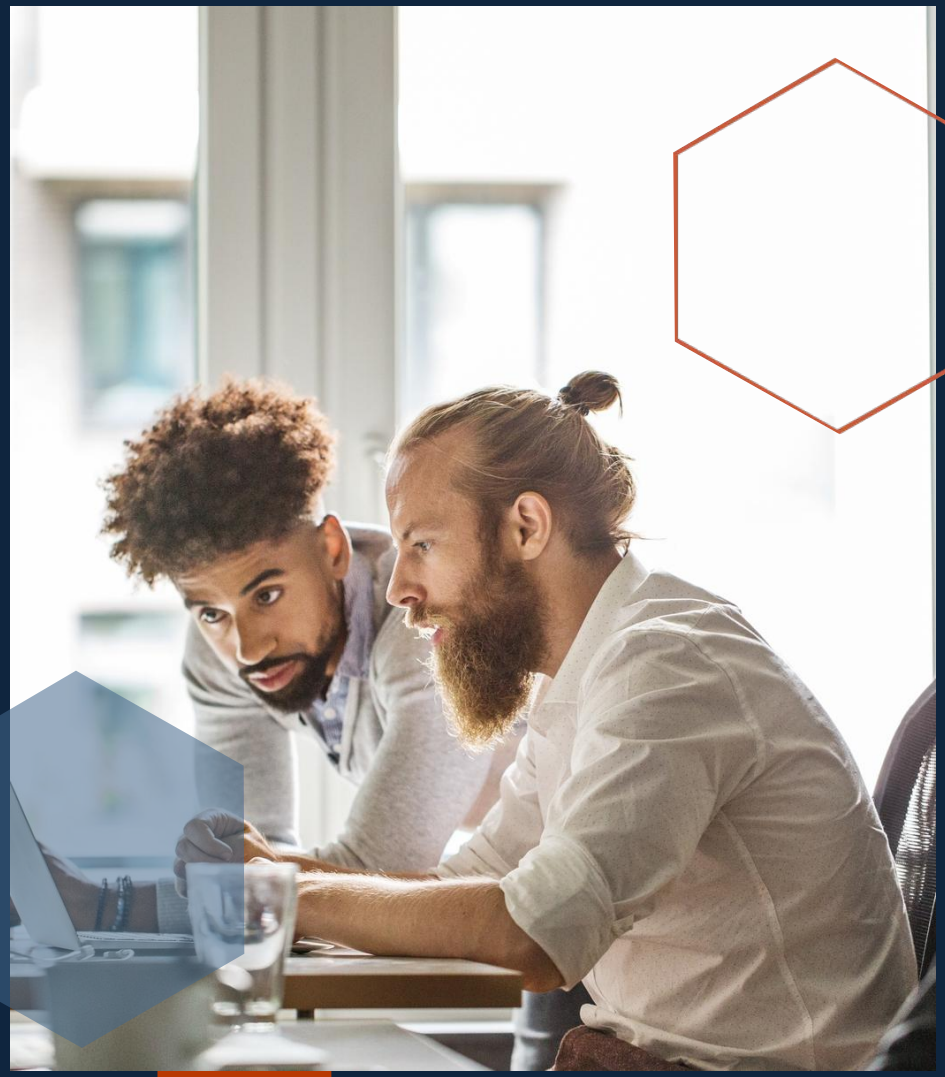
Capstone Project

Airbnb Bookings Analysis (EDA)

By- Navneet Keshri



Agenda

Introduction

Primary Goals

About Dataset

Data Analysis

Conclusion

Introduction

1. Airbnb is an online marketplace that connects people who want to rent out their homes with people who are looking for accommodations in specific locales.
2. Airbnb offers people an easy, relatively stress-free way to earn some income from their property. Guests often find Airbnb is cheaper, has more character, and is homier than hotels.
3. Airbnb makes the bulk of its revenue by charging a service fee for each booking.
4. Cons of using this service include not getting what you expected, and, for hosts renting your place to someone you haven't had the chance to meet first.

Problem Statements

AI

DATA ANALYSIS

Explore the data, investigate the variables, and clean the data for EDA using Numpy and Pandas.

BUSINESS ANALYSIS

Find out the most important variables to focus on from a business point of view (Busiest days for hosts, popular neighborhoods, etc.)

FINANCIAL ANALYSIS

Do the cost and revenue analysis and find out groups where Airbnb can focus to increase its profit, average price/day, the effect of price on room type, etc.

SENTIMENT ANALYSIS

Analyze the sentiment from ratings of the customers. Find out the relationship between price, availability, and customer ratings if there is any.

DATA VISUALIZATION

Simultaneously visualize different variables and see bar charts, graphs, tables, etc. using Matplotlib and Seaborn.

ABOUT DATASET

```
airbnb.shape
```

```
(48895, 16)
```

48895 Rows and 16 Columns with some null values in name, host_name, last_review, and reviews_per_month columns

```
airbnb.info()
```

```
<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
```

Feature Description

1. id – This is a unique identification number linked to each row.
2. name – Name of Hotel/Apartment listed by local hosts.
3. host_id – Identification number assigned to the local host.
4. host_name – Name of the host.
5. neighbourhood_group – These are the names of the areas where hotels are located (5 groups - Manhattan, Brooklyn, Queens, Bronx, Staten Island)
6. neighborhood – Names of the sub-areas under 5 groups where hotels are situated.

```
airbnb['name'].value_counts()

Hillside Hotel
Home away from home
New york Multi-unit building
Brooklyn Apartment
Loft Suite @ The Box House Hotel
```

```
airbnb['neighbourhood_group'].value_counts()

Manhattan        21661
Brooklyn         20104
Queens            5666
Bronx             1091
Staten Island      373
Name: neighbourhood_group, dtype: int64
```

Feature Description

7. latitude – Latitude coordinates for the location of the hotel/apartment.
8. longitude – Longitude coordinates for the location of the hotel/apartment.
9. room_type – Type of rooms (3 types - entire home/apt, private room, shared room)
10. price – Rent of the property listed in USD (Range – $0 to $10000)
11. minimum_nights – Number of nights the customer rented a hotel

```
airbnb['room_type'].value_counts()

Entire home/apt    25409
Private room       22326
Shared room         1160
Name: room_type, dtype: int64
```

Feature Description

12. number_of_reviews – Number of reviews for the hotel by users.
13. last_review – Last date when a hotel is reviewed by the user.
14. reviews_per_month – Count of how many reviews the hotel/apartment gets per month.
15. calculated_host_listings_count –Number of times the listing is done by a particular host.
16. availability_365 – Number of days the hotel/apartment is available for rent (Ranging from 0 to 365 as per the availability)

Data Cleaning

```
# checking for the null values in the dataset
airbnb.isna().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
```

Before Cleaning

1. Drop the rows of a column which have less number of null values (name, host_name)
2. We can drop the columns last_review and reviews_per_month but in the analysis, it is not so useful so we are just neglecting them.

Data Cleaning

```
airbnb['availability_365'].min(), airbnb['availability_365'].max()
```

```
(0, 365)
```

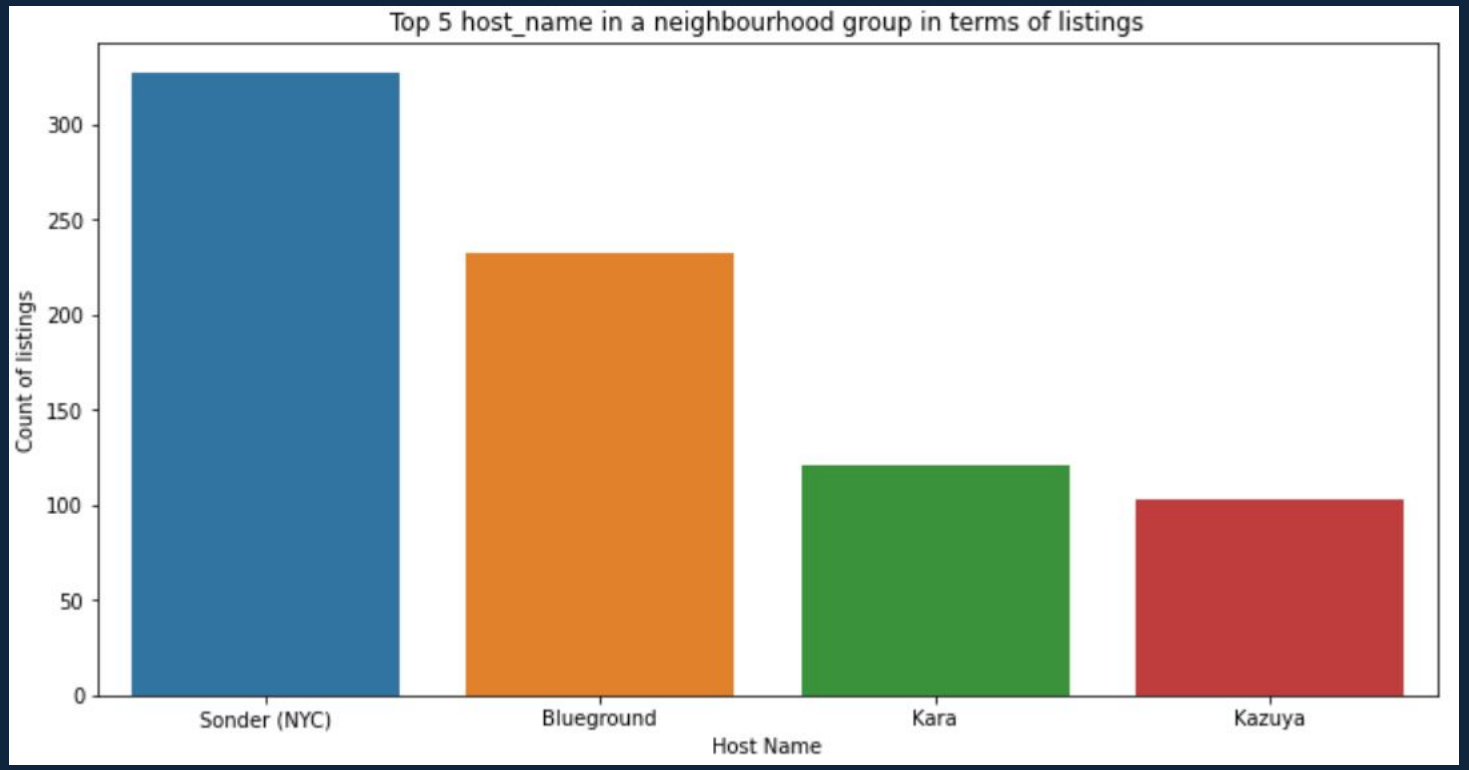
availability in the number of days which cannot be 0 so replacing it with the mean integer value of availability_365

Data After Clean up

```
airbnb.shape
```

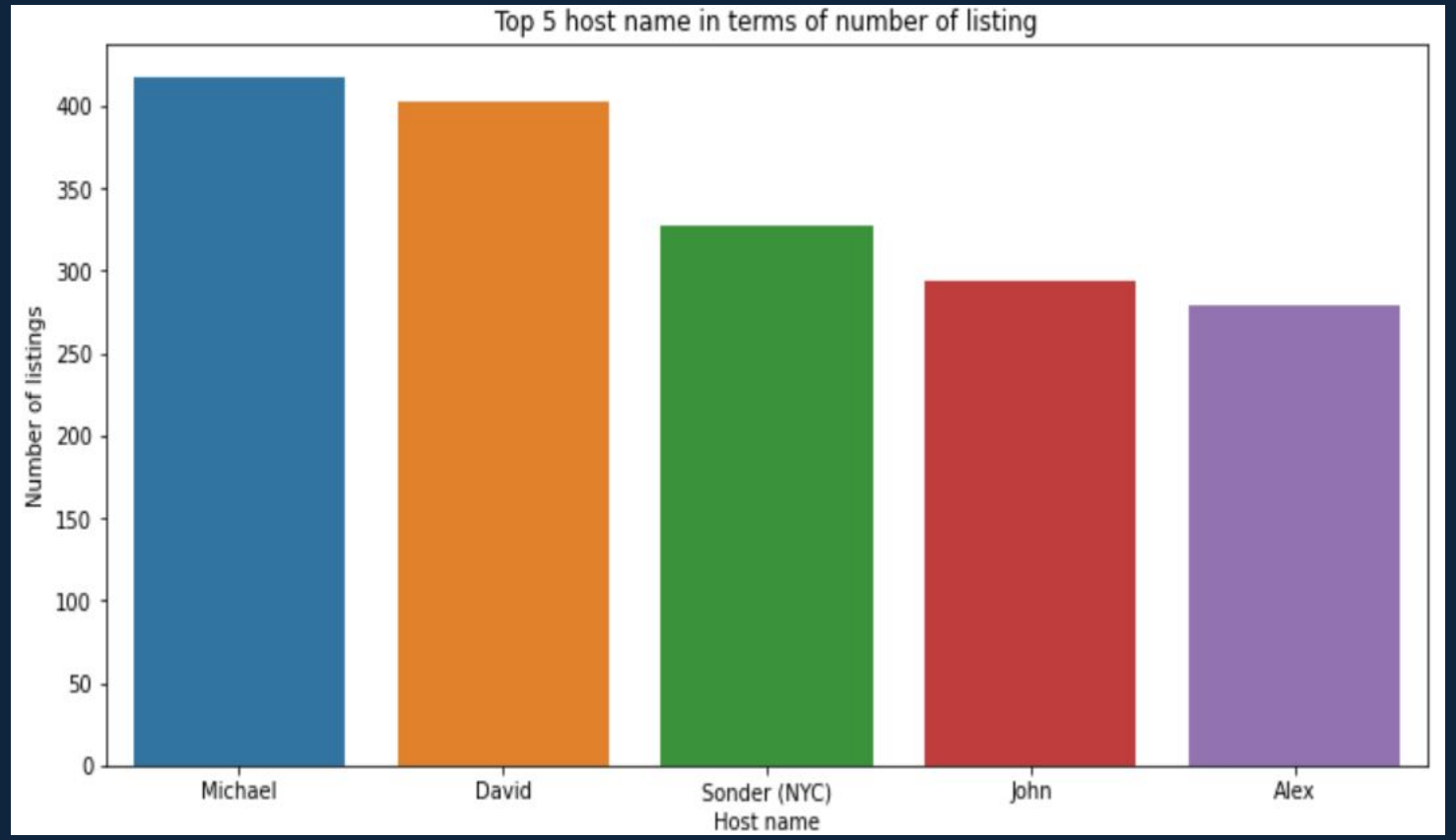
```
(48858, 16)
```

Data Analysis and Visualization



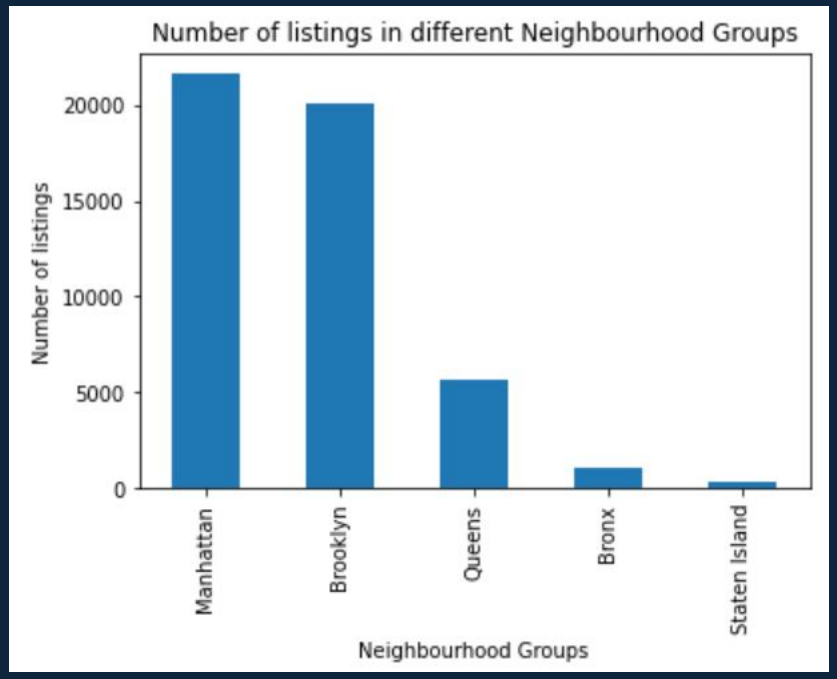
Host name Sonder has a maximum listing which is 327, All located in the Manhattan neighbourhood group

Data Analysis and Visualization

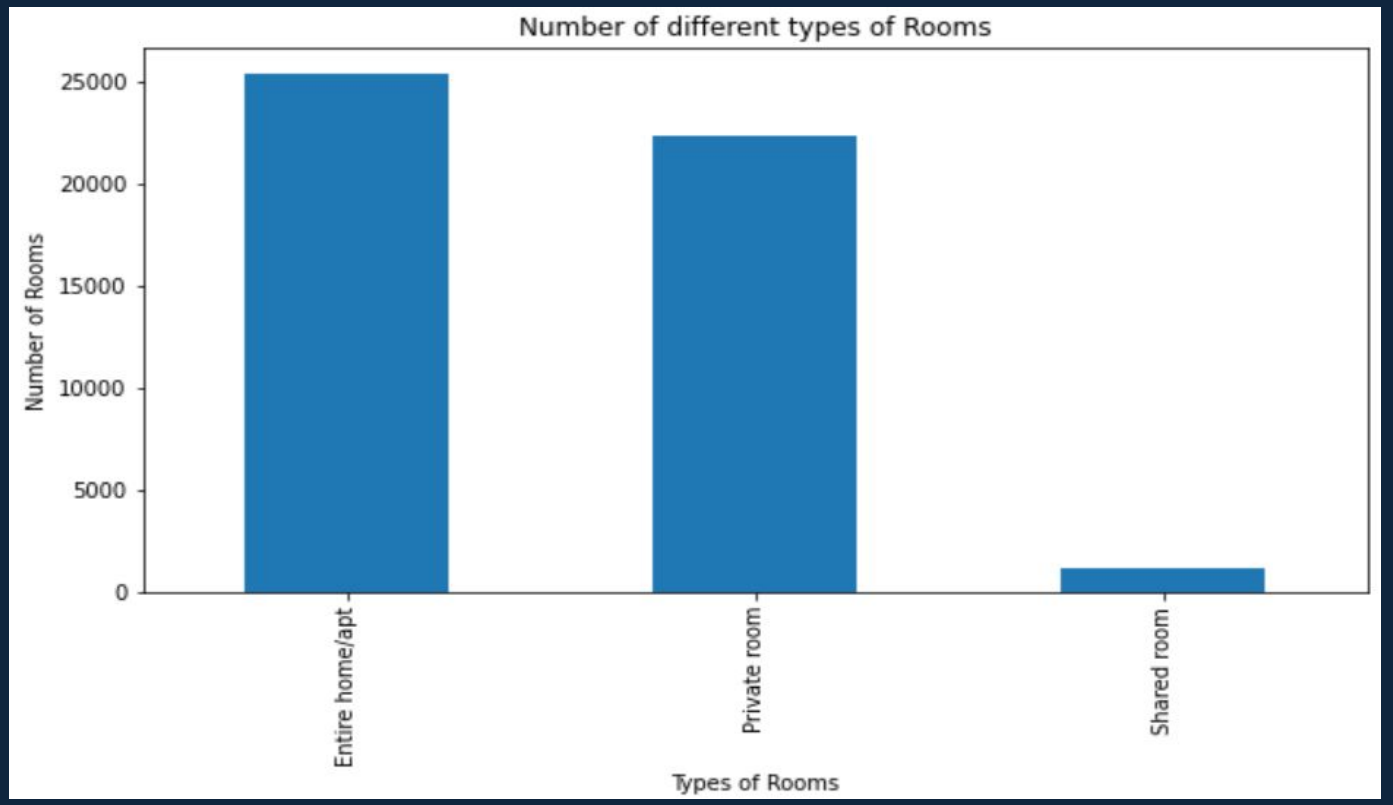


Michael has a maximum count in host_name which means Sonder has listings only in Manhattan but Michael has listings in different neighbourhood_groups or maybe there are multiple Michael as host_name

Data Analysis and Visualization

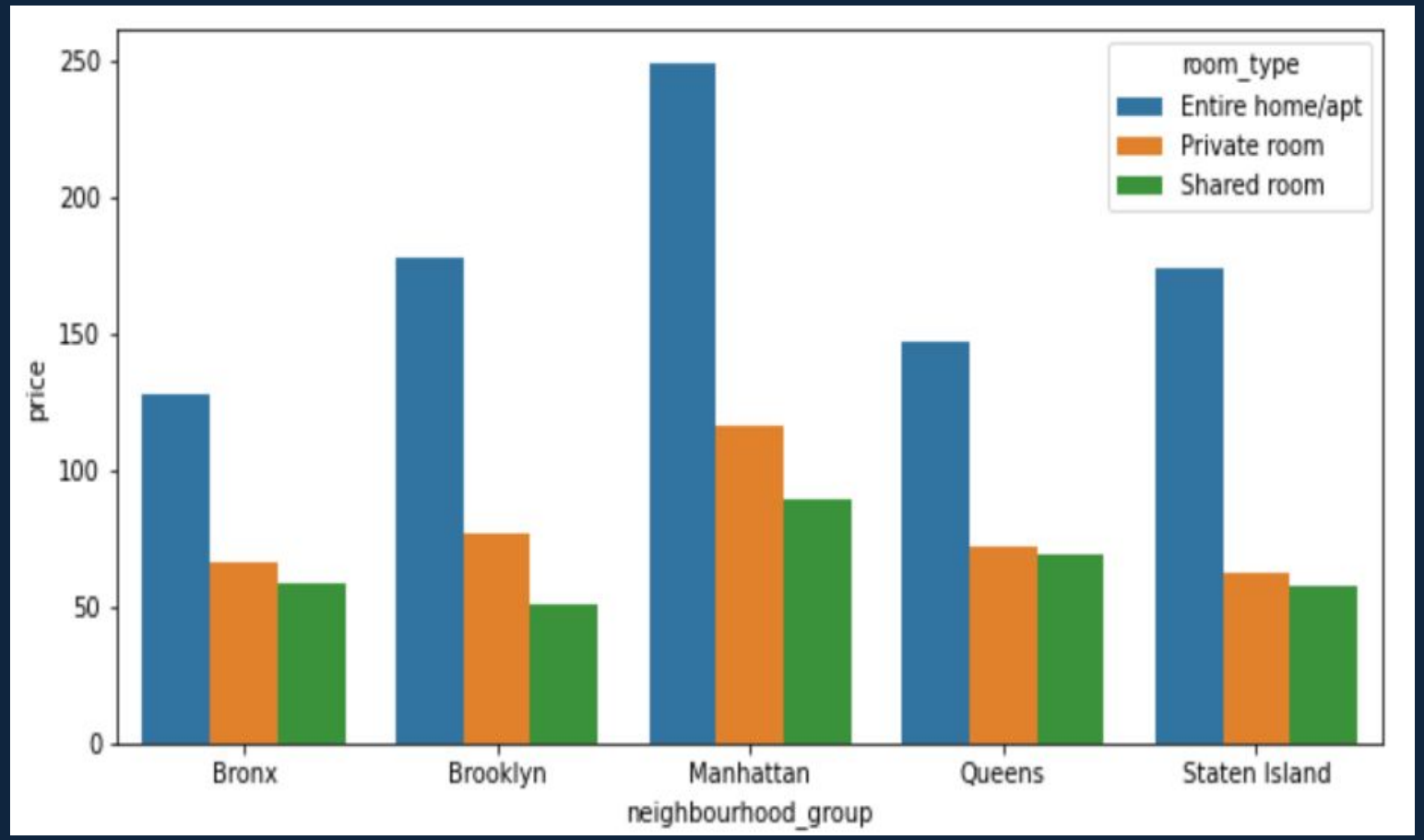


Manhattan has the most number of listings followed by Brooklyn



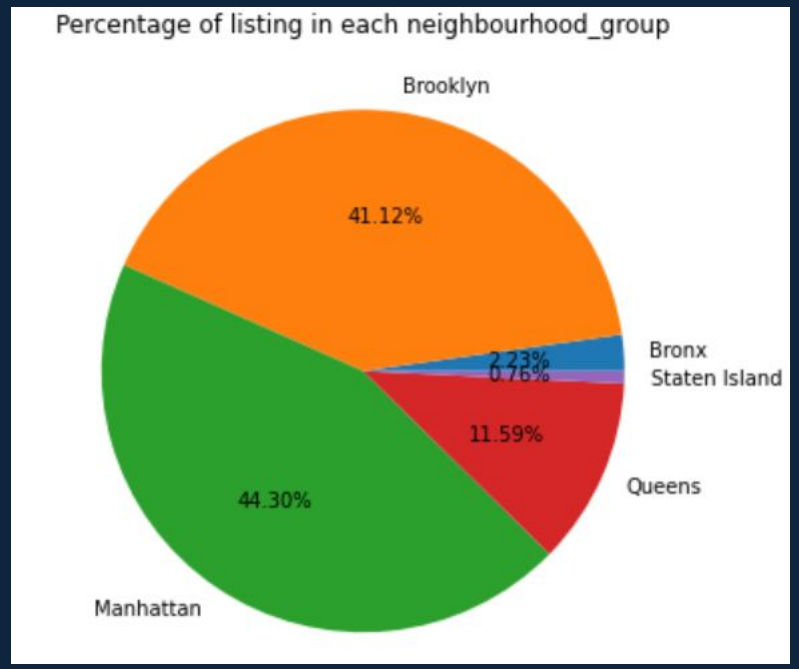
The entire home/apartment is the most preferred room type followed by a private room. Shared room is preferred rarely.

Data Analysis and Visualization

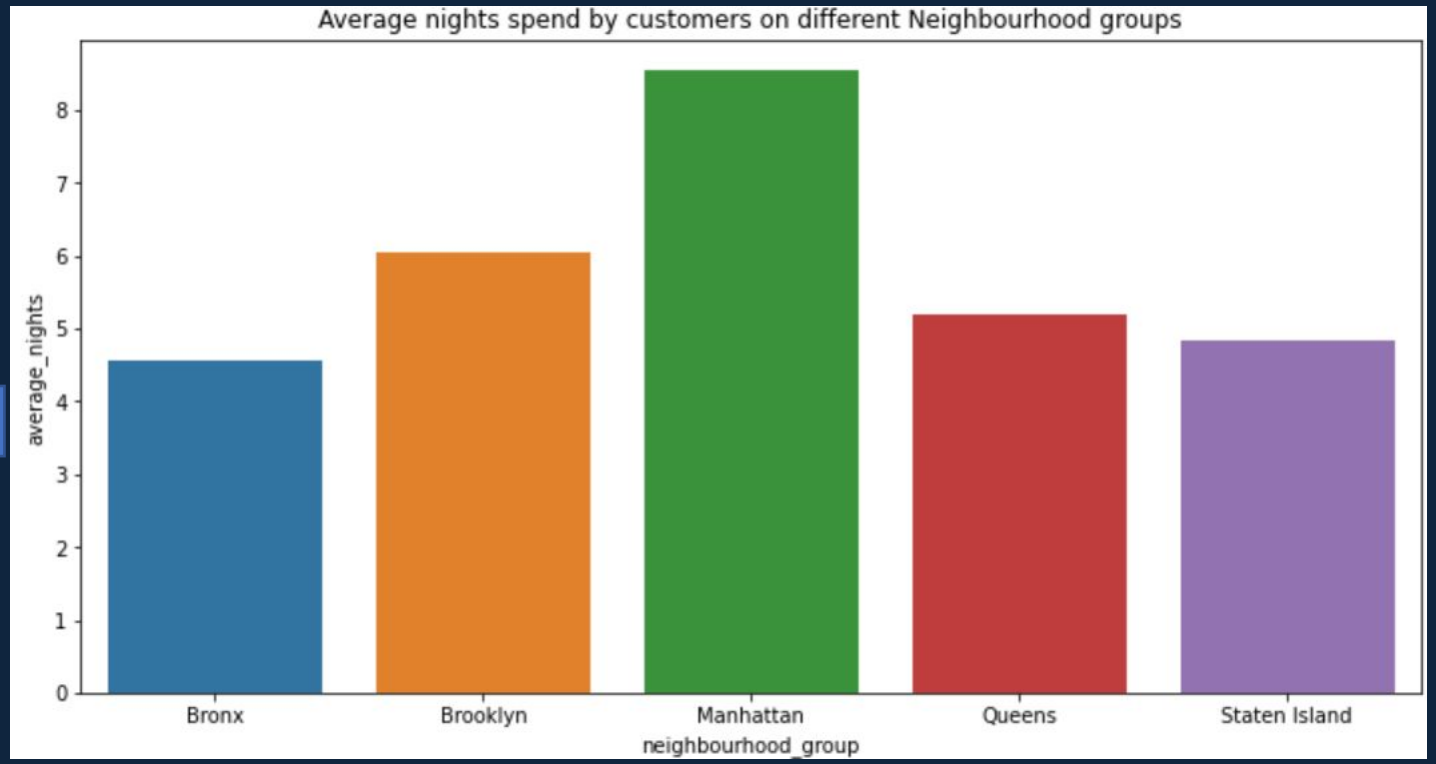


Entire home/apartment is the most preferred room type even though it is the most costly in all the neighbourhood groups

Business Analysis – Analyze data from a business perspective

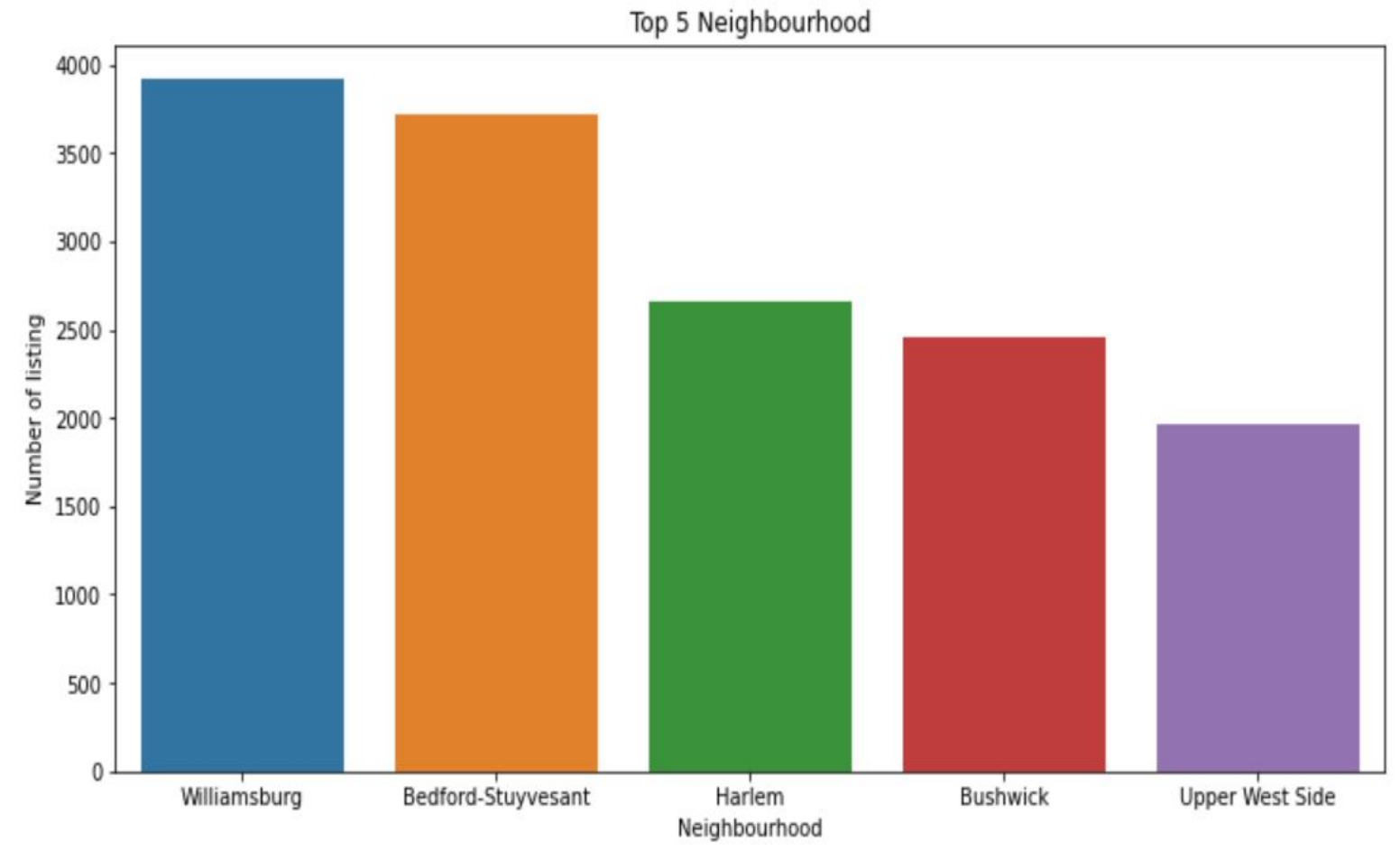


85% of listing on Airbnb come from Manhattan and Brooklyn. If we consider Queens also it will be around 97%. It is important from a business perspective



The average length of stay (ALOS) which is an Important Business KPI again shows the importance of Manhattan and Brooklyn. Users prefer to stay more in Manhattan and Brooklyn but this may be due to most number of hotels/apartments located in Manhattan and Brooklyn

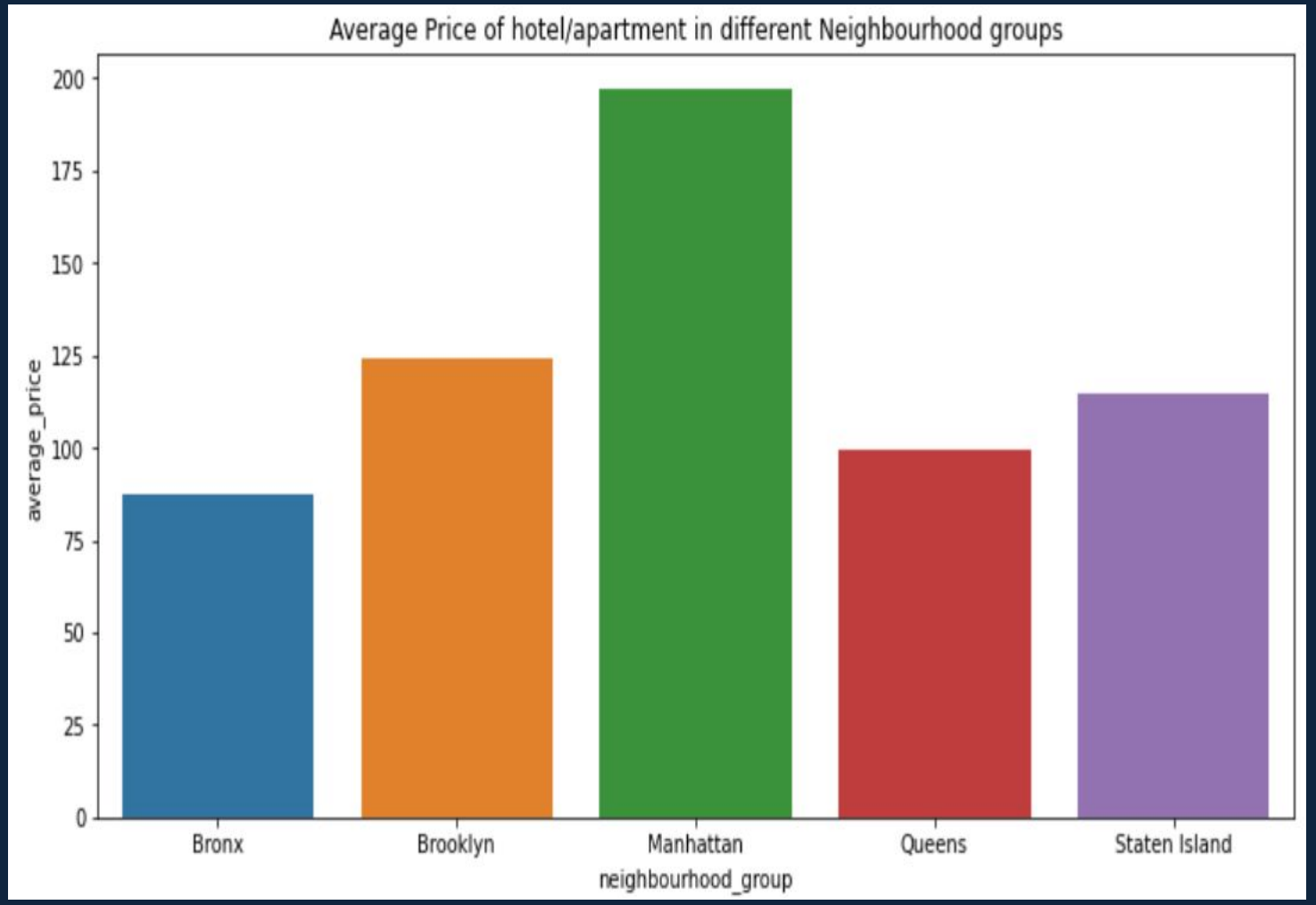
Business Analysis – Analyze data from a business perspective



The top 5 neighbourhoods are Williamsburg, Bedford-Stuyvesant, Harlem, Bushwick, and Upper West Side.

Airbnb needs to keep an eye on these neighbourhoods and try to list more hotels/apartments from these beighbourhoods

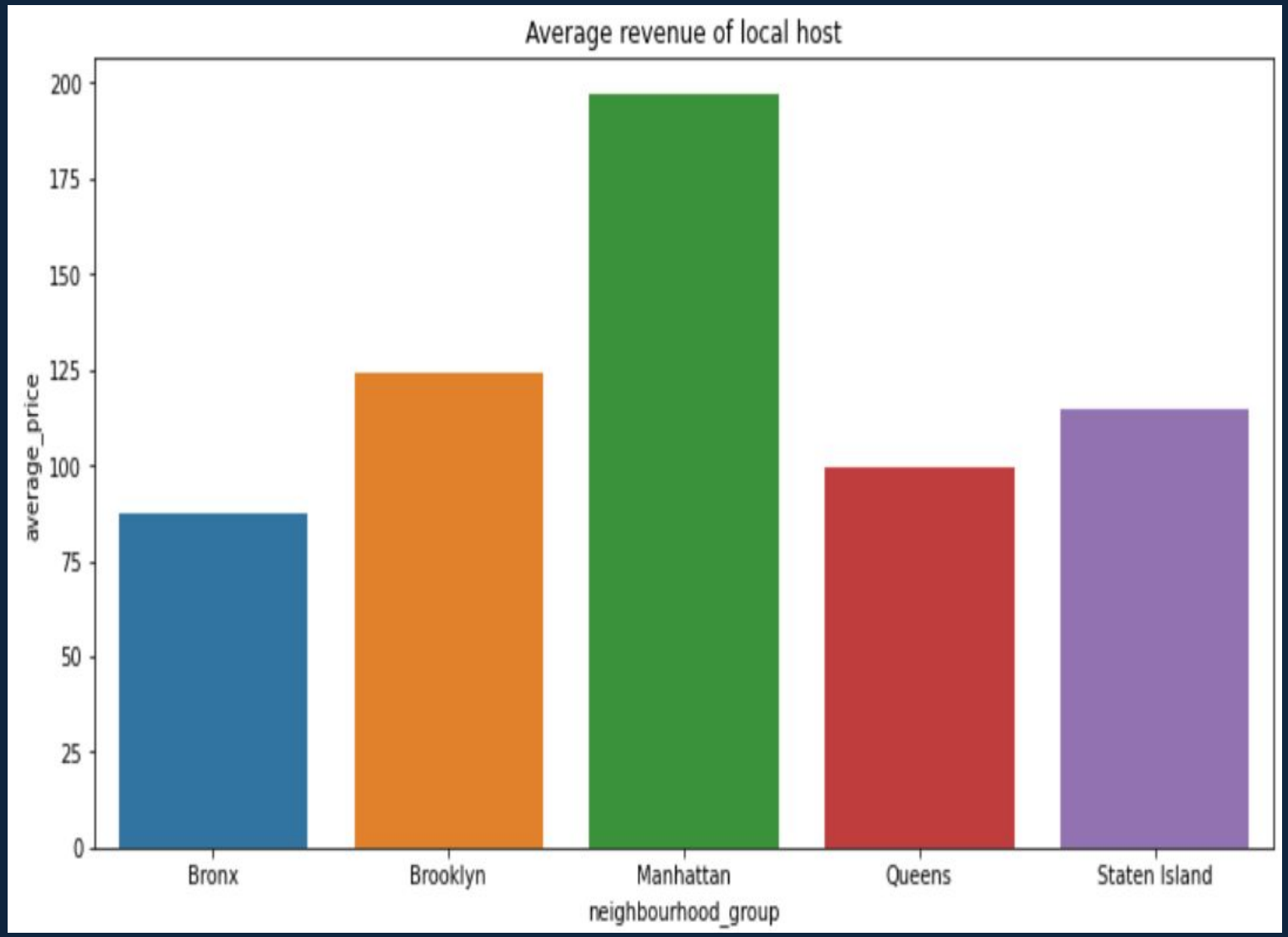
Financial Analysis – Analyze data from a cost and revenue perspective



Although Manhattan and Brooklyn have the highest number of hotels yet they are costly, which means they are important from the finance point of Airbnb.

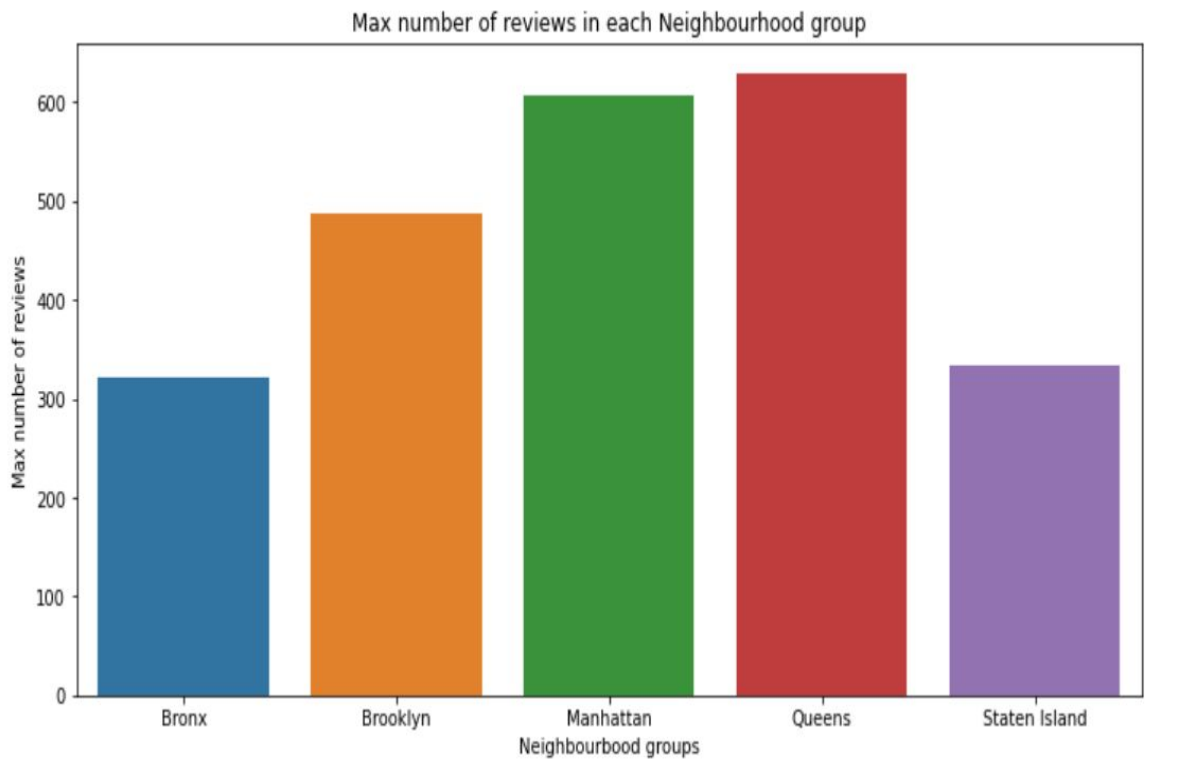
Important to note that Staten Island was not in the top 3 in the number of listings yet it is in the top 3 in the average price of the hotel which may be because not enough hotels in Staten Island. (Airbnb should focus on listing more hotels/apartments from Staten Island to match demand and supply)

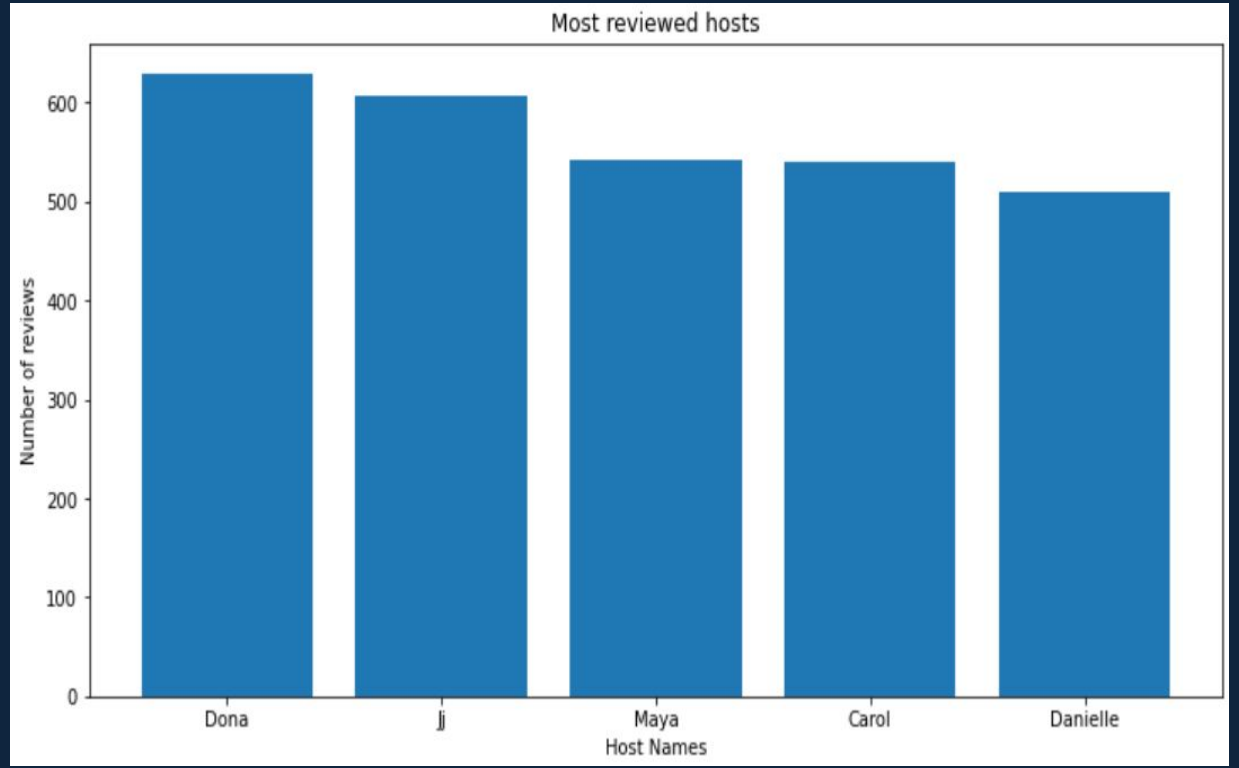
Financial Analysis – Analyze data from a cost and revenue perspective



From a local host point of view –The average revenue of a local host per neighbourhood group can be a good criterion to analyze revenue from different neighbourhood groups.
Manhattan and Brooklyn provide more average revenue to local hosts
This data is similar to the price per neighbourhood group

Sentiment Analysis – Analyze user's ratings

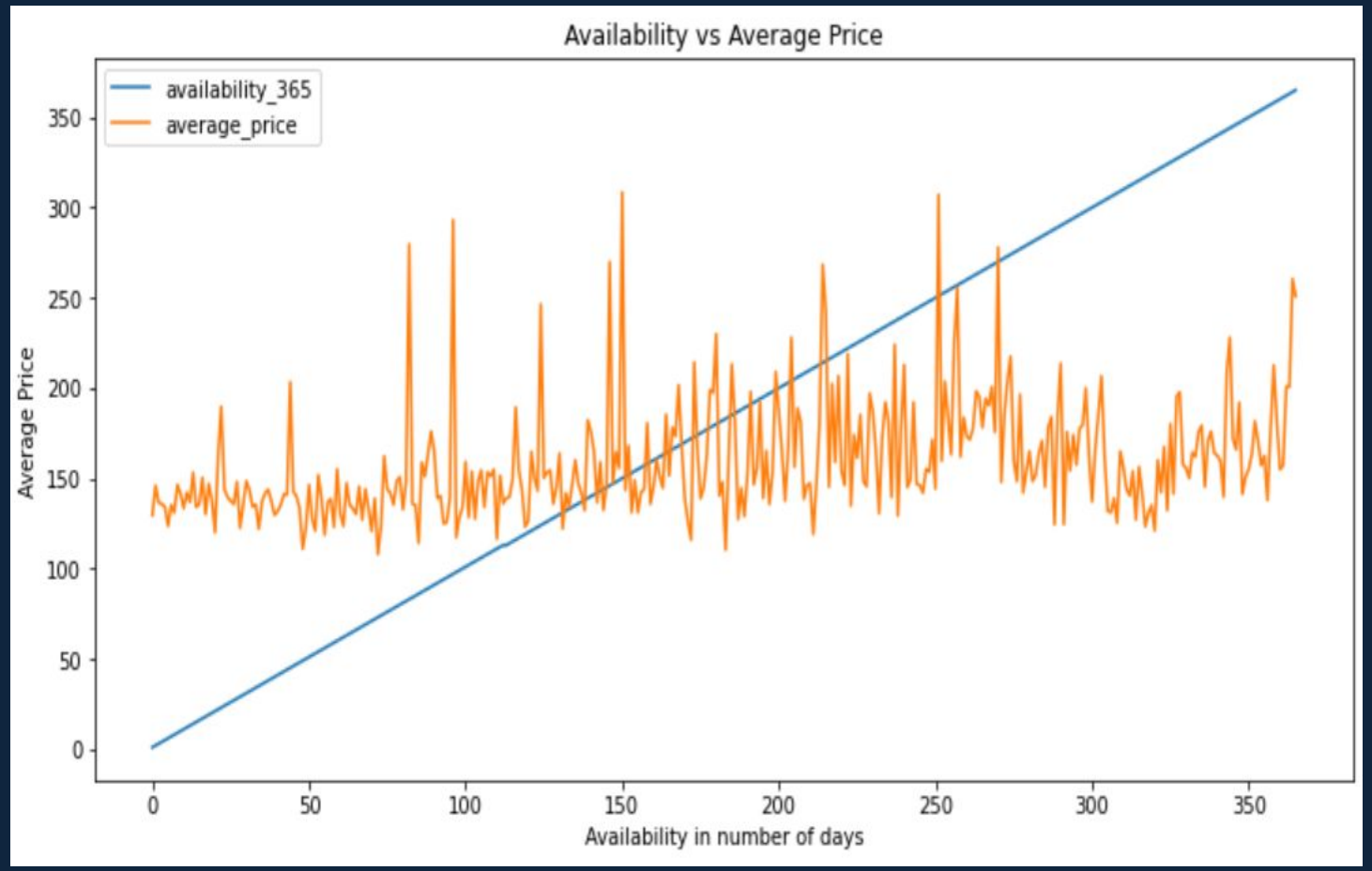




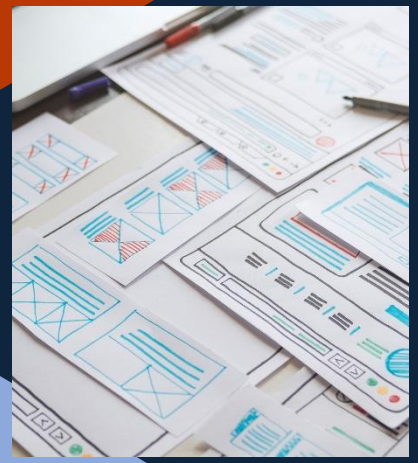
Reviews show that even though Manhattan and Brooklyn have a max number of listings yet the max number of reviews belongs to Queens which means at least one hotel in Queens is very popular among the users. Most rated hotel/apartment name - Room near JFK Queen Bed

The top 5 most-rated local hosts are Dona, Jj, Maya, Carol, and Danielle. Among these top 5, three hotels/apartments belong to Queens, and the other two belong to Manhattan

Sentiment Analysis – Analyze data from ratings perspective



There is no significant relationship between availability and average price but the price shoots up from an average of USD 150 to USD 250-300 in some cases which may be hotel/apartment specific.



Areas of focus

Average price/listing

Staten Island was not in the top 3 in the number of listings yet it is in the top 3 in the average price of the hotels. Need to focus on more listings

Manhattan and Brooklyn are also important from a business point of view as the average price of a hotel/apartment is the highest

Ratings-based opportunities

Some hotels/apartments in Queens should be given more importance as the maximum number of reviews belongs to Queens which means some hotels/apartments in Queens are very popular among the users. Most rated hotel/apartment name - Room near JFK Queen Bed

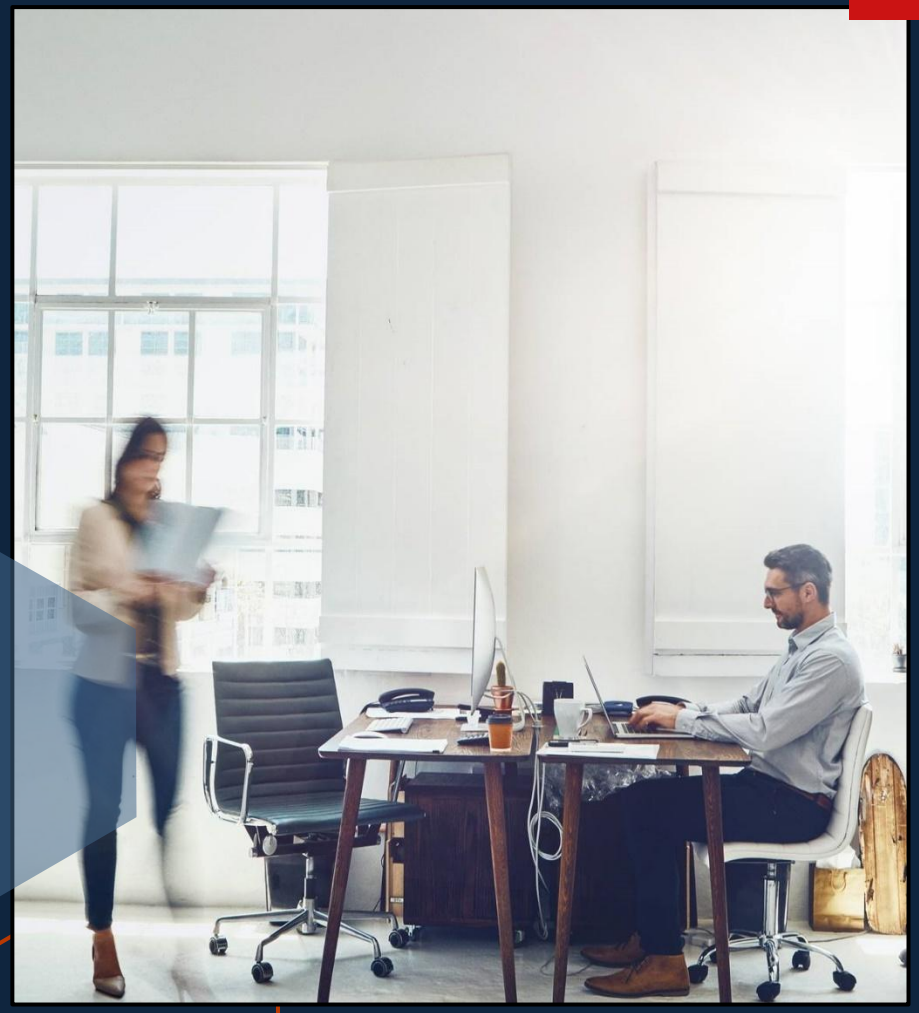
Major Findings

Manhattan has the most number of listings (85% of listing on Airbnb come from Manhattan and Brooklyn) with the Host name Sonder having a maximum listing which is 327 in Manhattan.

People prefer to stay in Entire homes/apartments and Private rooms which is why the average price of private rooms is higher than Shared rooms.

The top 5 neighbourhoods are Williamsburg, Bedford-Stuyvesant, Harlem, Bushwick, and Upper West Side. The entire home is the most preferred room type.

There is no significant relationship between availability and average price but the price shoots up from an average of USD 150 to USD 250-300 in some cases which may be hotel/apartment specific.



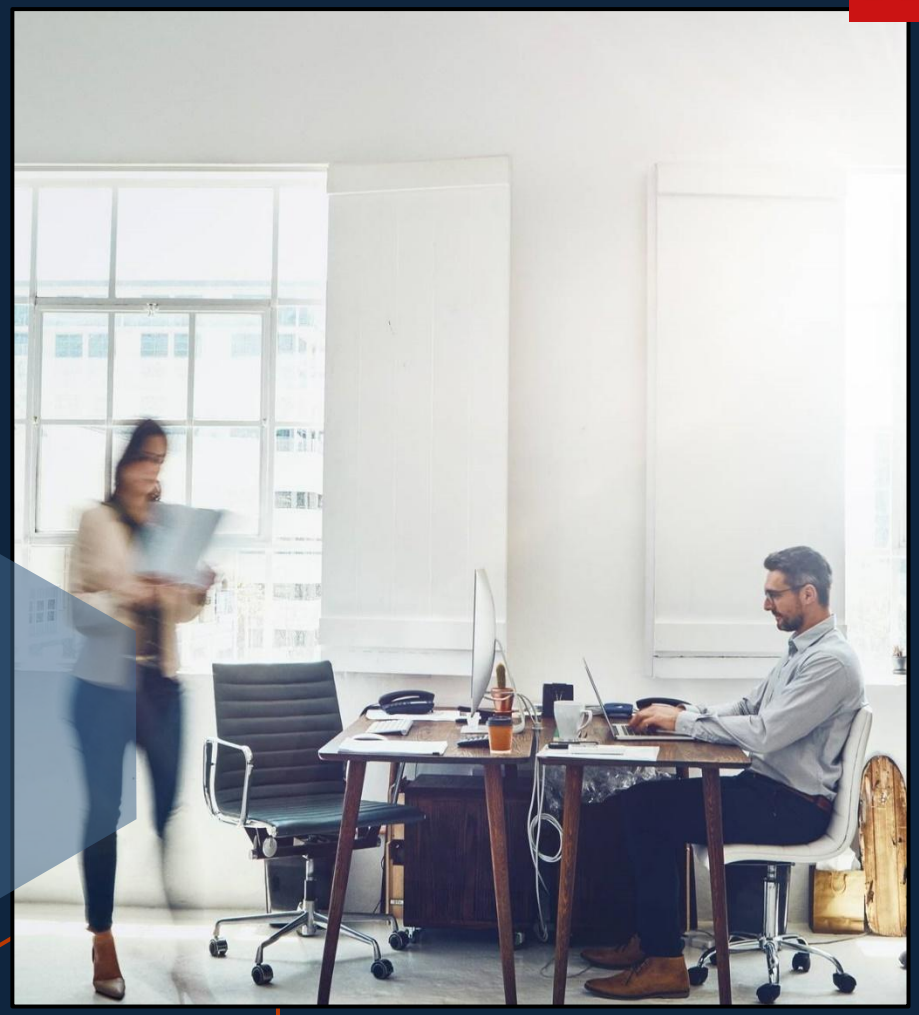
Conclusion

* Through this Exploratory Data Analysis (EDA) we analyzed Airbnb Bookings Analysis using Pandas and used Seaborn and Matplotlib for the visualization purpose.

* Started with understanding the variables, we prepared the data after cleaning it and found the relationship between variables.

* We also did analyze the data from a business perspective, financial perspective, and sentiment perspective. We used Data Visualization for a better understanding and presentation purpose.

* Area of focus and assumptions has been clearly stated whenever required. Now is up to Airbnb to utilize this analysis efficiently:)



“Most decisions are not binary, and there are usually better answers waiting to be found if you do the analysis and involve the right people.”

Jamie Dimon

Thank you

Navneet Keshri

navneet2409jnv@gmail.com

linkedin.com/in/navneet-keshri-28650918b