Capstone Project

AIR BNB BOOKING ANALYSIS

Team Members:

Abhishek V L

Neha R

Nisarga C

Swati G

analyticSteps

airbnb

Everything you need to know

www.analyticssteps.com

CONTENTS

- Introduction
- What is Airbnb
- Work overflow
- Data Understanding
- Analyzing the listings based on the room types
- Practical Theory
- Agenda
- Conclusion

INTRODUCTION

- Introduction Unlike hotels, which have their own pricing system, Aribnb prices are usually determined by the hosts empirically. It poses challenges for the new hosts, as well as for existing hosts with new listings, to determine the prices reasonably high yet without losing popularity. On the consumers' side, though they can compare the price across other similar listings, it is still valuable for them to know whether the current price is worthy and if it is a good time to book the rooms.
- The nightly price for Airbnb renting depends on multiple factors, and we divide the input type into 4 categories, including Continuous , categorical, text, and date features. We have extracted more than 60 features from the dataset. Here we only list a few of them that are both representative and important for the task, such as room size (accommodates, bathrooms, bedrooms, beds,etc), extra fees (security deposit, cleaning fee, extra people, etc), reviews scores {review scores rating, review scores accuracy, review scores cleanliness), location (neighbourhood, latitude, longitude,etc) facilities (transit, amenities, property type,etc) and booking related (availability, cancellation policy, host verification,etc).

What is Airbnb?

- Airbnb wants to analyze the historical data of all the listings on its platform since its initial stages and improve its recommendations to its customers.
- To do this, they need to gather the average rating, number of ratings, and prices of the Airbnb listings over the years.
- As a data engineer of the company. We took up the task of building an ETL pipeline that extracts the relevant data like listings, properties, hosts details, and load it in to a data warehouse that makes querying for the decision-makers and analysts easier.
- This dataset has around 49,000 observations in it with 16 columns and it is a mix between categorical and numeric values.
- 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 one of a kind service 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 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.

Work overflow

Work is divided into three steps:

Step 1- Data Collection and Understanding the data

Step 2- Data cleanup and Handling the missing values

Step 3 – Performing Agenda with the Visualizations.

DATA UNDERSTANDING

Field	Description
Id	Unique id
Name	Name of the listing
Host_id	Unique host_id
host_name	Name of the host
Neighbourhood_group	Location
Neighbourhood	Area
Latitude	Latitude range
Longitude	Longitude range
Room_type	Type of listing
Price	Price of listing
Minimum_nights	Minimum night to be paid for
Number_of reviews	No of reviews
Last_review	Content of the last review
Reviews_per_month	No of checks per month
Calculated_host_listing_count	Total count
Availability_365	Availability around the year

Analyzing the listings based on room types

- The number of listings for each neighborhood and the median price
- This gives us a good insight into the potential neighborhoods where there are higher number of listings which we can tap into. By analyzing the number of listings and prices for each neighborhood, we can get a clearer understanding of which neighborhood have a lot of expensive listings. Looking at the analysis done so far, we can see that certain neighborhood are indeed more 'expensive' than others.
- However, some of those neighborhood do not have as many listings as other expensive neighborhood. Since our problem was to identify factors that make a listing more expensive, we can infer that these neighborhood tend to have more expensive listings. However, a more thorough inference would be to identify neighborhood that have both a higher number of listings and higher price as lower number of listings would mean fewer available listing for a customer to choose

Common words in the summary of expensive listings

- This word cloud shows shows the most frequently used words in the summaries of the top 100 most expensive listings.
- They all have particularly 3 words in common: seattle, home, and view. Other words like : kitchen, bedroom, walk, modern.

Common words in the summary of the cheapest listings

- The word cloud, indeed there are overlapping words with the most expensive listings. Words like : Seattle, bedroom, home appeared frequently in both .So they do not tell us anything special .

Analyzing if any particular amenity results in higher prices.

- The word cloud above was taken from the top 100 listings in terms of their price. We can see that the listings with the highest prices have amenities such as washer, dryer, heating, wireless internet, smoke detector, free parking, kid friendly
- . So, an aspiring Airbnb host should ensure that his property contains these amenities so that he can charge a higher price. Similarly, if a traveller does not require any of these amenities, he can opt for a listing without them to save cost. amenities and their influence into the price will be further explored in depth in the machine learning section of the project.

Practical Theory

- Airbnb has provided many travellers a great, easy and convenient place to stay during their travels. Similarly, it has also given an opportunity for many to earn extra revenue by listing their properties for residents to stay.
- So many listings available with varying prices, how can an aspiring host know what type of property to invest in if his main aim is to list it in Airbnb and earn rental revenue? Additionally, if a traveller wants to find the cheapest listing available but with certain features he prefers like 'free parking' etc, how does he know what aspects to look into to find a suitable listing? There are many factors which influence the price of a listing. Which is why we aim to find the most important factors that affect the price and more importantly the features that is common among the most expensive listings. This will allow an aspiring Airbnb host to ensure that his listing is equipped with those important features such that he will be able to charge a higher price without losing customers. Moreover, a traveller will also know the factors to look into to get the lowest price possible while having certain features he prefers.

Agenda

We are trying to give the solution to airbnb for the following data.

- 1) Which hosts are having highest number of apartments ?
- 2) Which are the top 10 neighborhood which are having maximum number of apartments for Airbnb in the respective neighborhood ?
- 3) What are the neighborhood in each group which are having maximum prices in their respective neighborhood group ?
- 4) How neighborhood is related with reviews ?
- 5) What can we learn from predictions? (ex: locations, prices, reviews, etc)
- 6) What is the distribution of the room type and its distribution over the location ?

- 7) How does the room type is distributed over Neighborhood Group are the ratios of respective room types more or less same over each neighborhood group ?
- 8) How the price column is distributed over room type and are there any Surprising items in price column ?
- 9) Which are the top 5 hosts that have obtained highest no. of reviews ?
- 10) What is the average preferred price by customers according to the neighborhood group for each category of Room type?
- 11) What is the average price preferred for Keeping good number of reviews according to neighborhood group ?
- 12) Which host are the busiest and why?

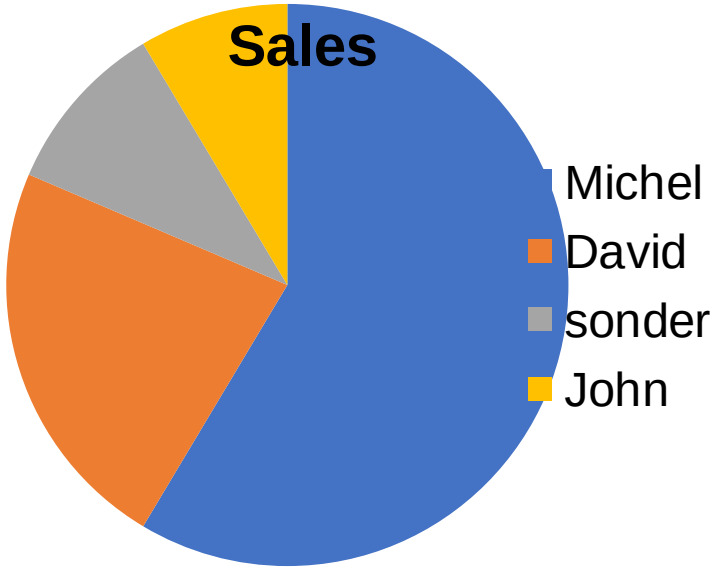
1.Which hosts are having highest number of apartments

As per the data collected from the collected data :

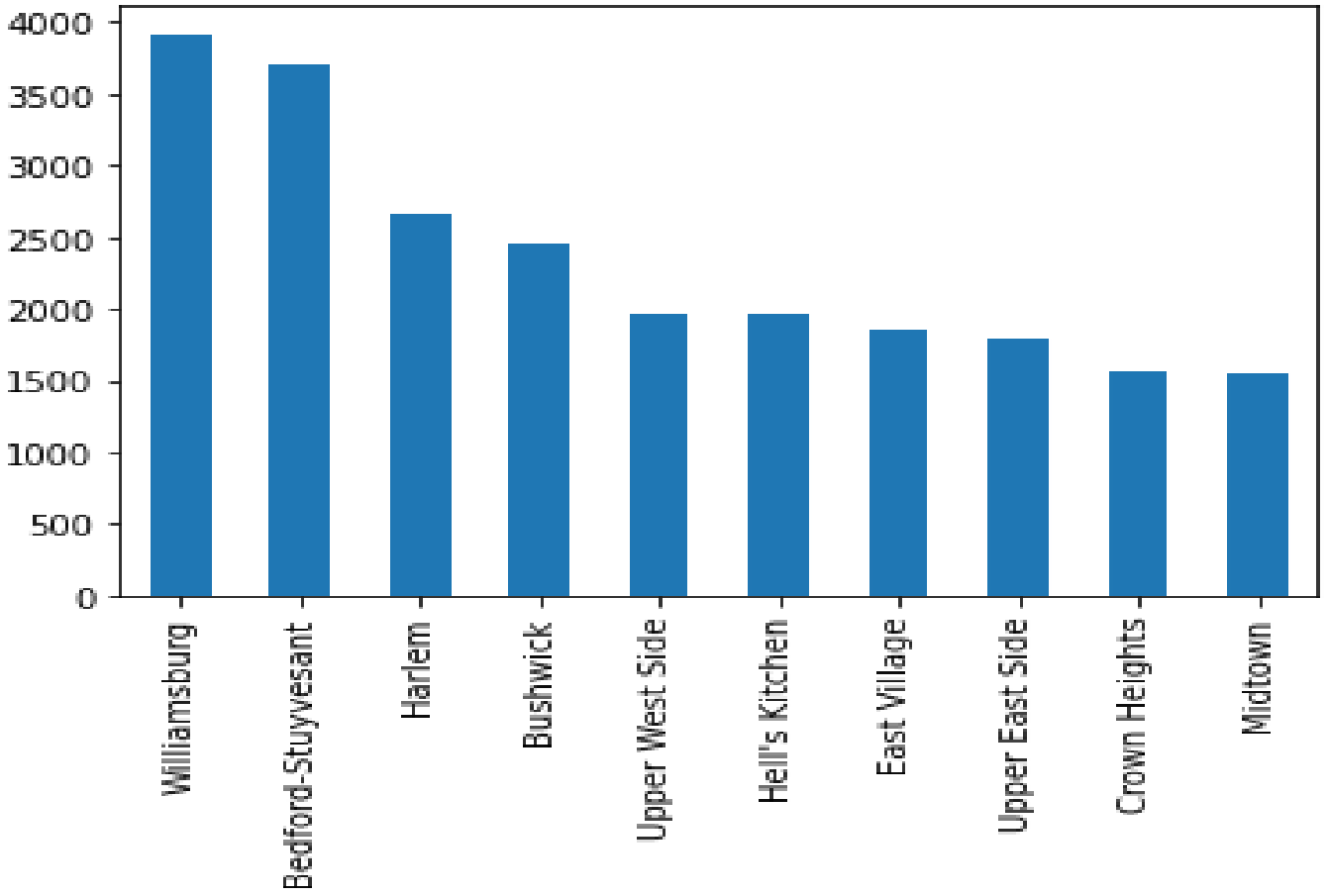
Michael has 417 apartments , David has 403 apartments, Sonder (NYC) has 327 apartments, John has - 294 apartments

From this we can see that host name Michael its appearing 417 times in the host name column , so this might imply that Michael is having highest number of rooms , but from the host id column its showing highest appearance of any hostbid is 327 .

So this clearly implies that there can be multiple person may have same name that's why we are getting different highest appearance in host name as compared to host id



2. Which are the top 10 neighborhood which are having maximum number of apartments for airbnb in the respective neighborhood ?



- Williamsburg 3920
- Bedford-Stuyvesant 3714
- Harlem 2658
- Bushwick 2465
- Upper West Side 1971
- Hell's Kitchen 1958
- East Village 1853
- Upper East Side 1798
- Crown Heights 1564
- Midtown 1545

Williamsburg has highest number of apartments and bedford-stuyvesant and so on…

3. What are the neighborhood in each group which are having maximum prices in their respective neighborhood group ?

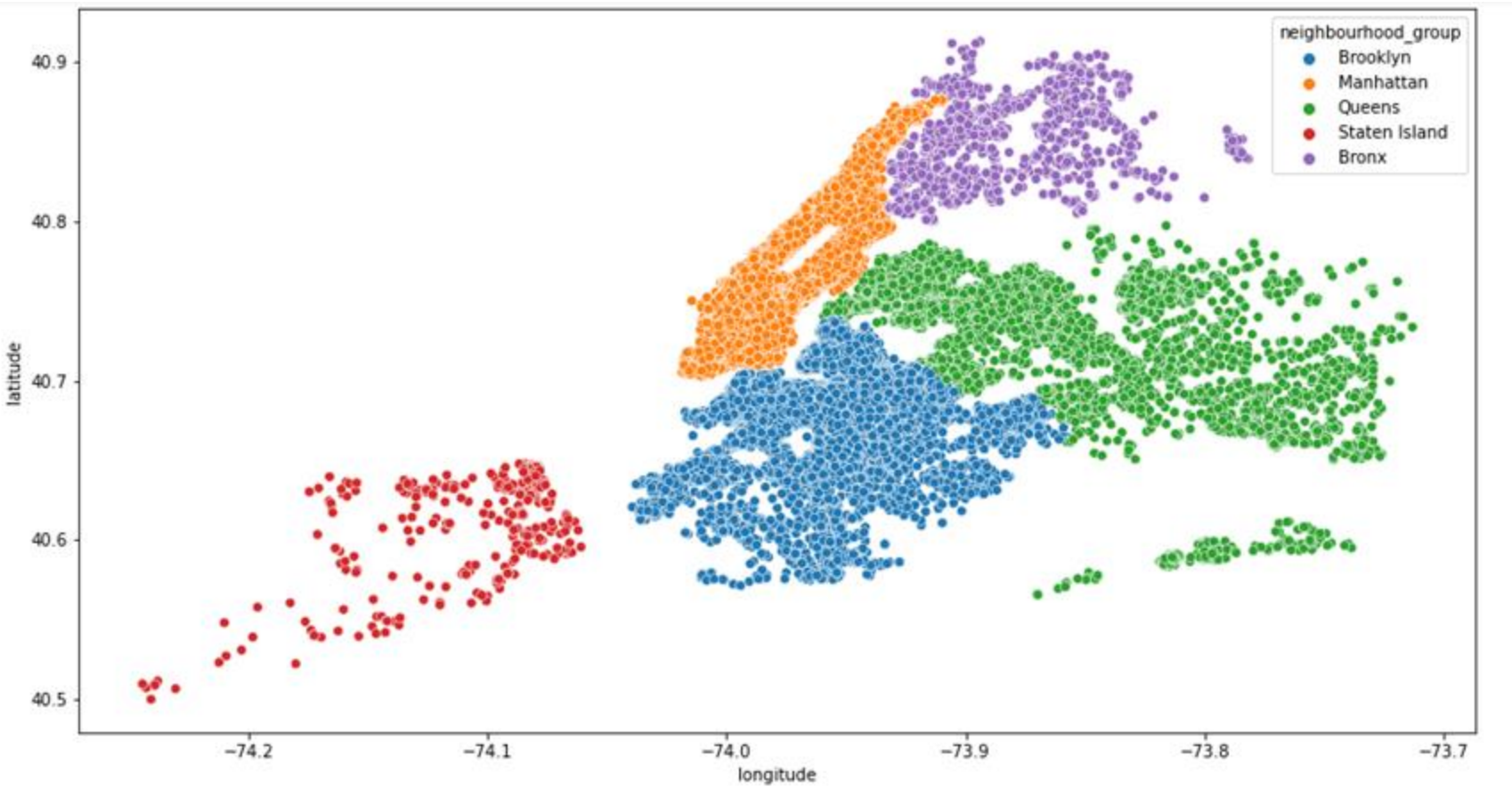
- Upper west side in Manhattan are having maximum prices that is 10000 .
- Randall Manor in Staten island are having maximum prices that is 5000.
- Riverdale in Bronx are having maximum prices that is 2500.
- Astori in Queens which are having maximum prices that is 10000.
- Green point in Brooklyn are having maximum prices that is 10000.

4.How neighborhood is related with reviews ?

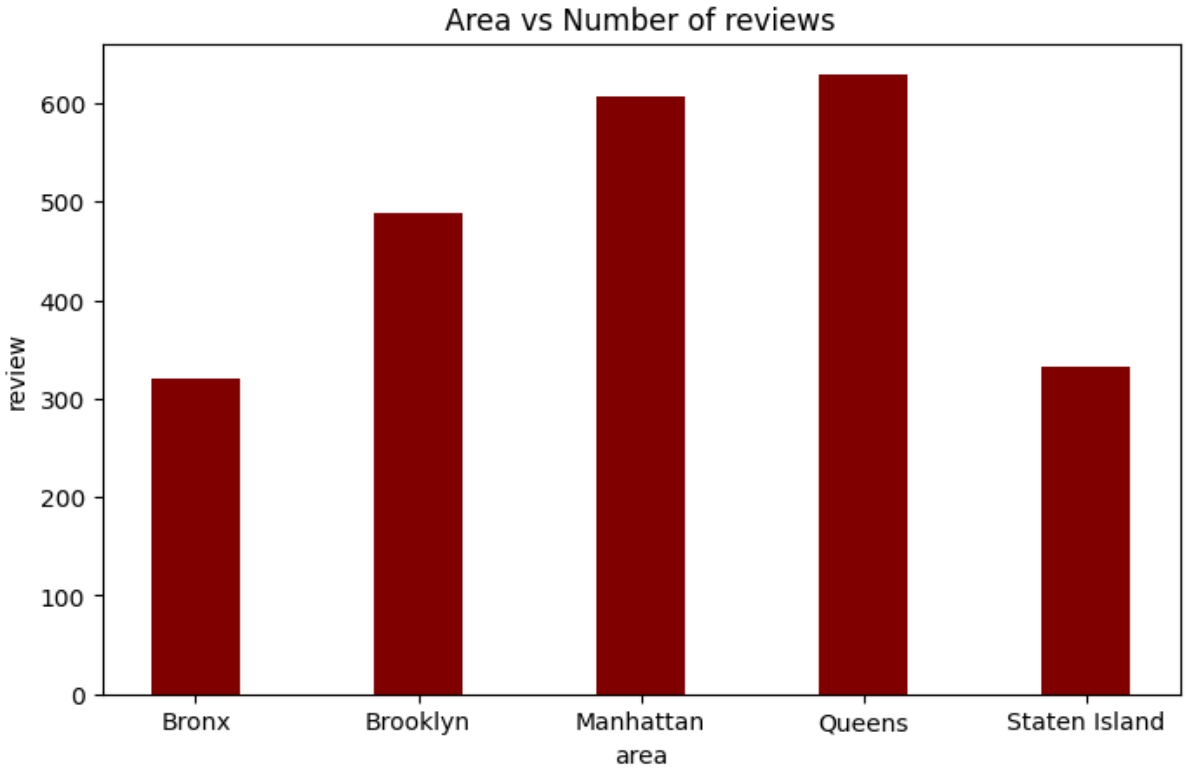
Top 5 Neighbourhood is having highest number of reviews

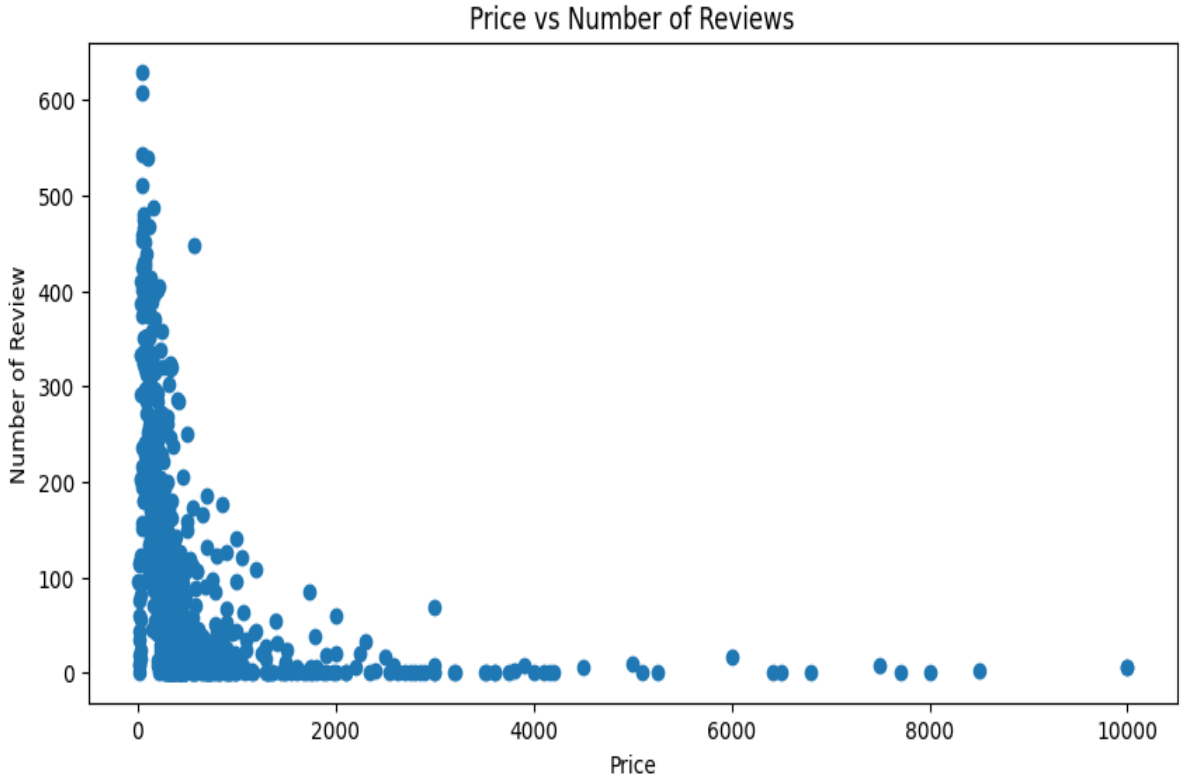
	Neighbourhood	Number of reviews
0	Bedford-Stuyvesant	110352
1	Williamsburg	85427
2	Harlem	75962
3	Bushwick	52514
4	Hell's Kitchen	50227

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



The location of each apartment using latitude and longitude values

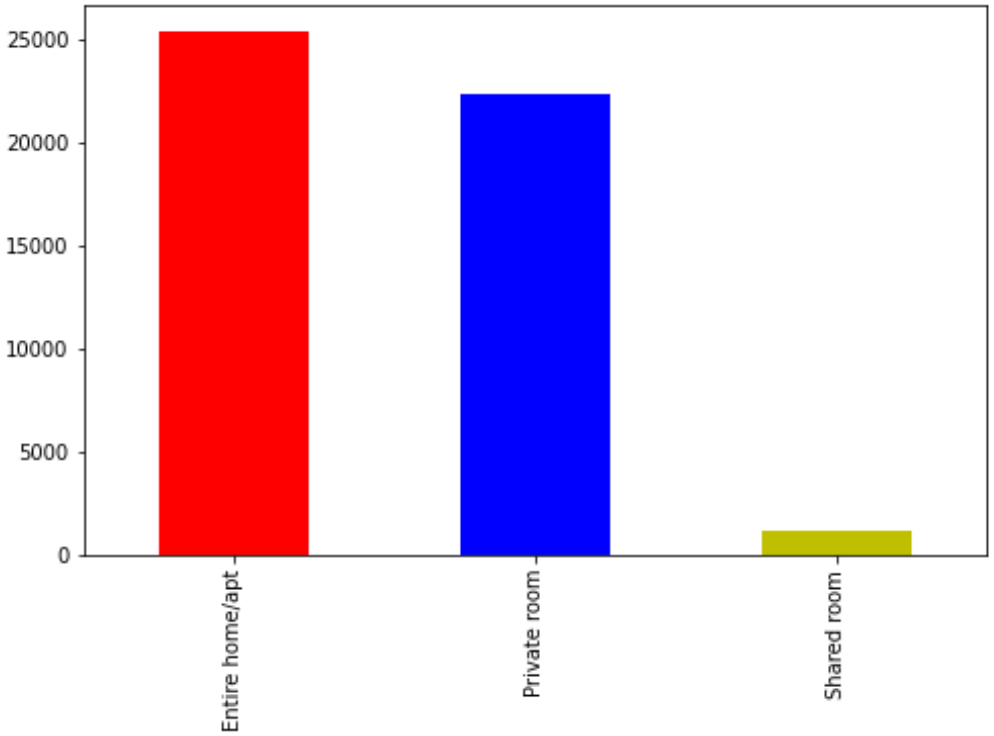


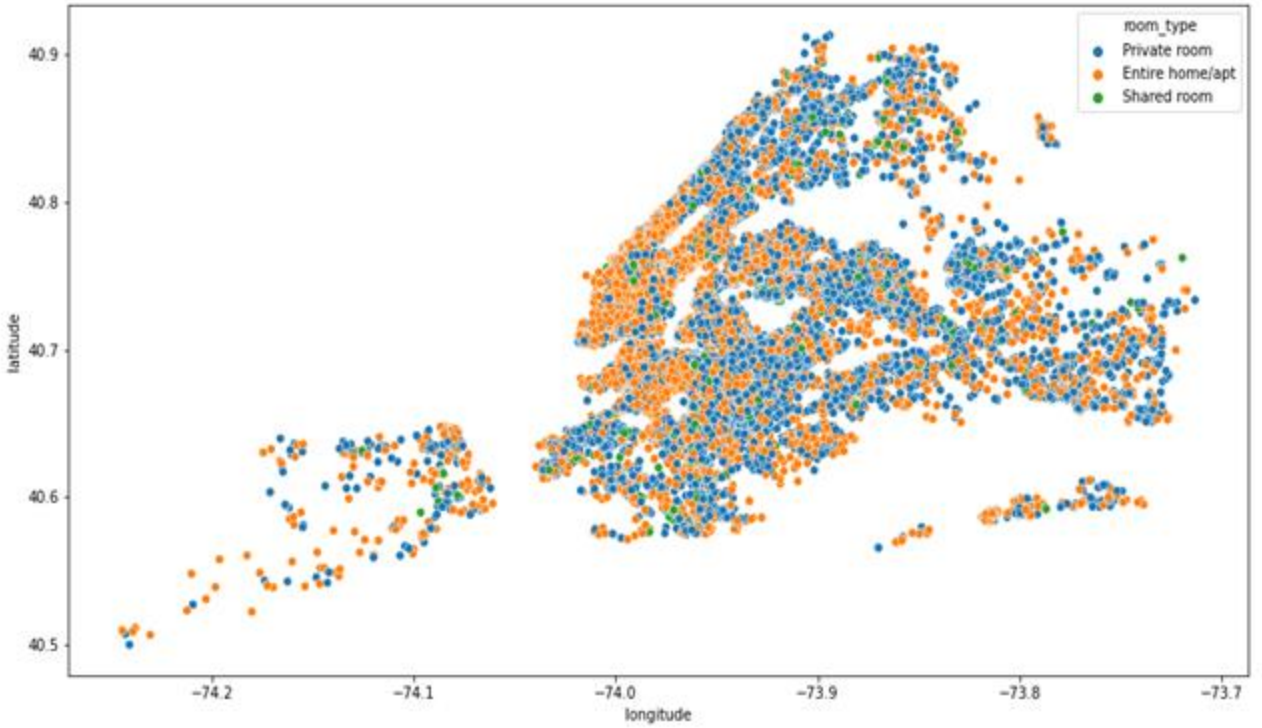


- From the above Analysis we can say that most people prefer to stay in place where price is less.

>

6.What is the distribution of the room type and its distribution over the location ?



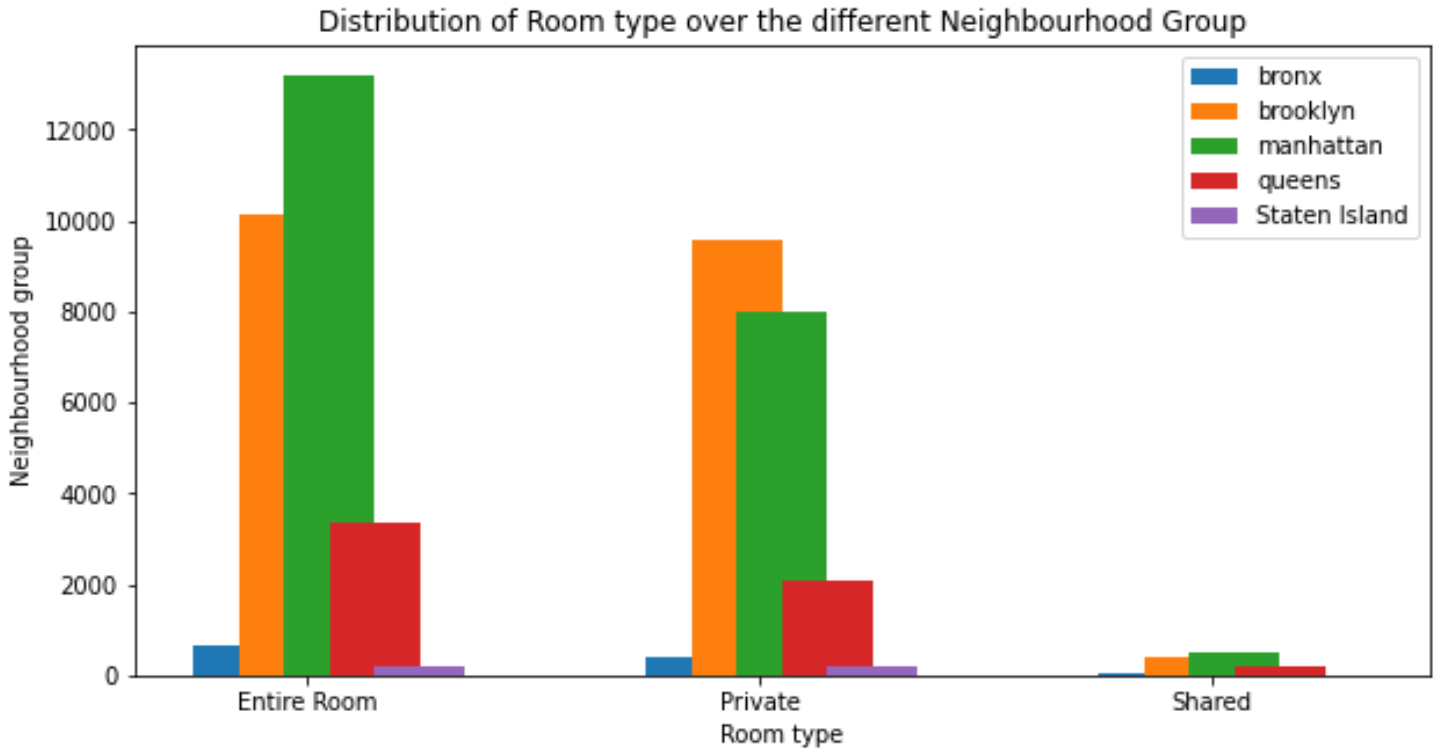


So we can notice the following

1) That maximum numbers of room are Entire home/Apartment and Private room there are only few shared rooms .

2)So mostly host prefer to give Entire home/ Apartment or Private Rooms rather than Shared rooms

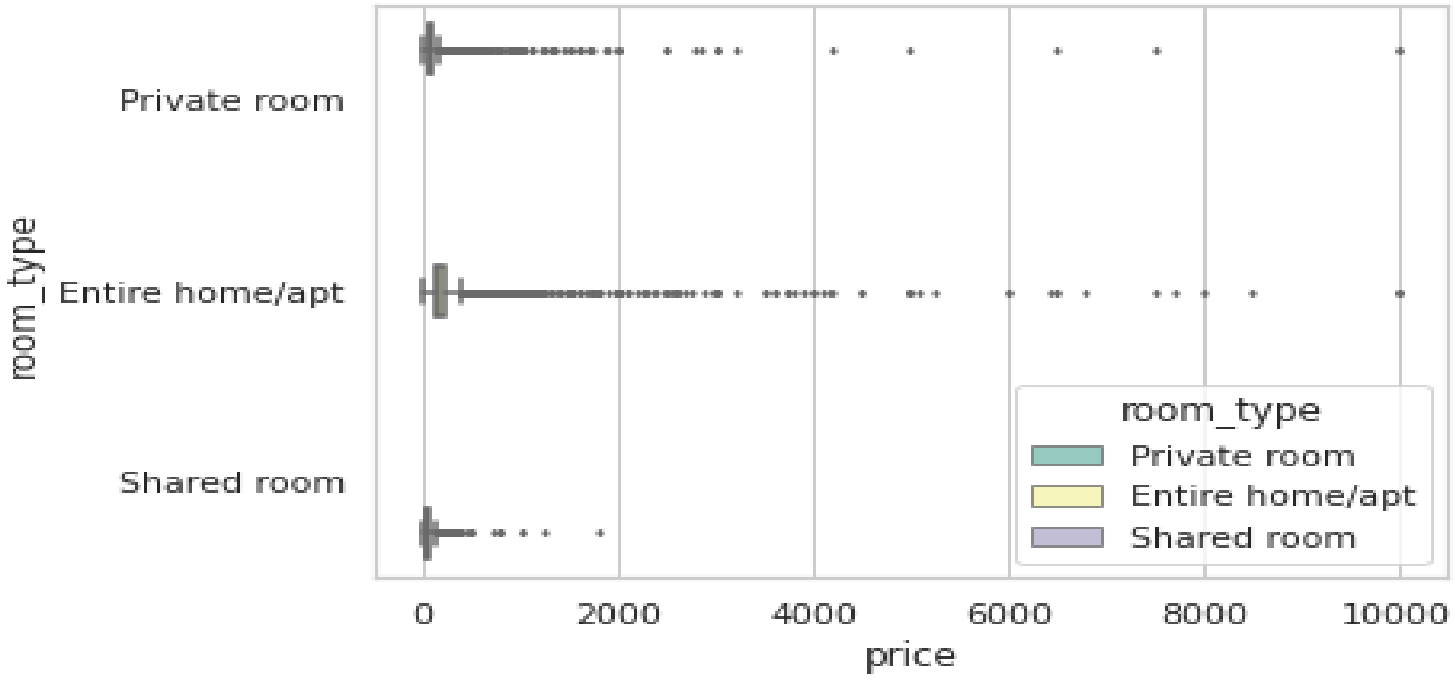
7.How does the room type is distributed over Neighborhood Group are the ratios of respective room types more or less same over each neighborhood group ?



This seems more or less same ratio in every neighbourhood

8.How the price column is distributed over room type and are there any Surprising items in price column ?

- There are many outliers for price in each of the room type category, so lets just why there is so high price or what else we can conclude for hosts having highest price for the rooms.



Host name	Reviews per month	Last review	Availability 365	price	Neighbourhood group
Kathrine	0.04	2016-02-13	0	10000	Queens
Erin	0.16	2017-07-27	0	10000	Brooklyn
Jelena	NaN	NaN	83	10000	Manhattan

Clearly if i would have working in Airbnb I would have suggested the following

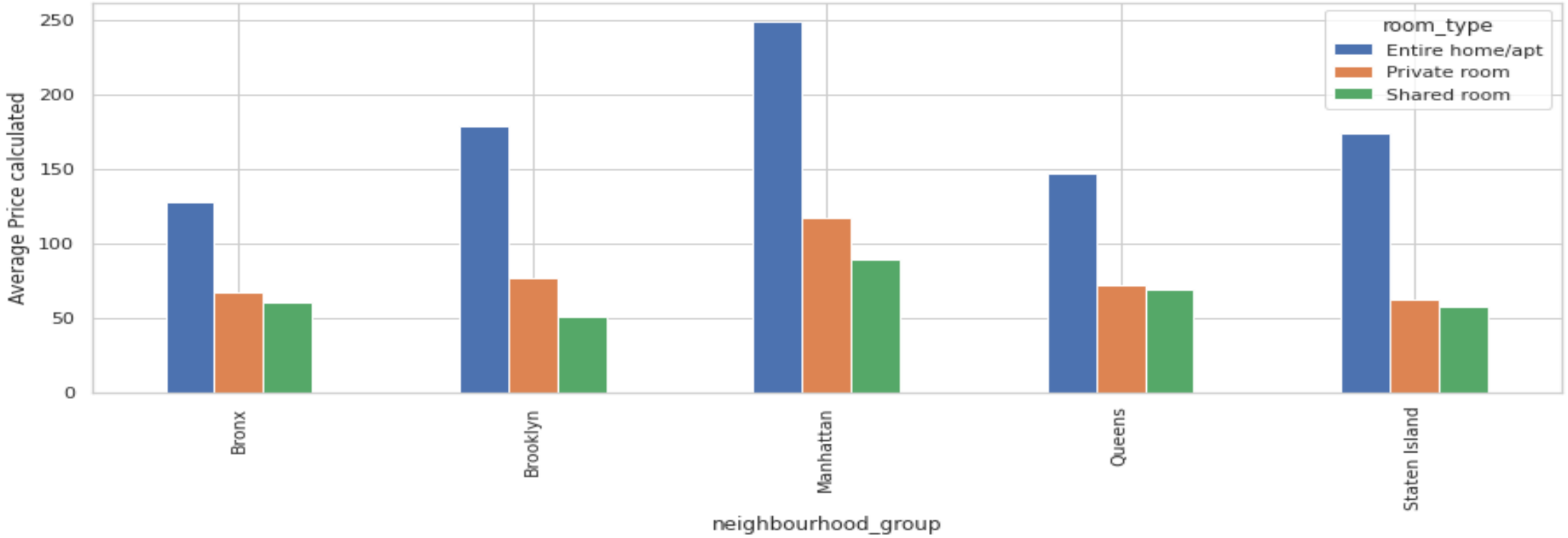
1) Katharine and Erin have price so high and having no availability then what is the benefit of keeping too high price .

2) The last review is also 2-3 years back (as the data was collected in 2019) which is also bad

3) The review may be low as there may be very few people who is staying in Katharine, Erin and jelena apartment so might have less reviews per month

4) I would have suggested to keep moderate(average) price so that more people would visit and stay in her apartment , it would also increase her reviews per month

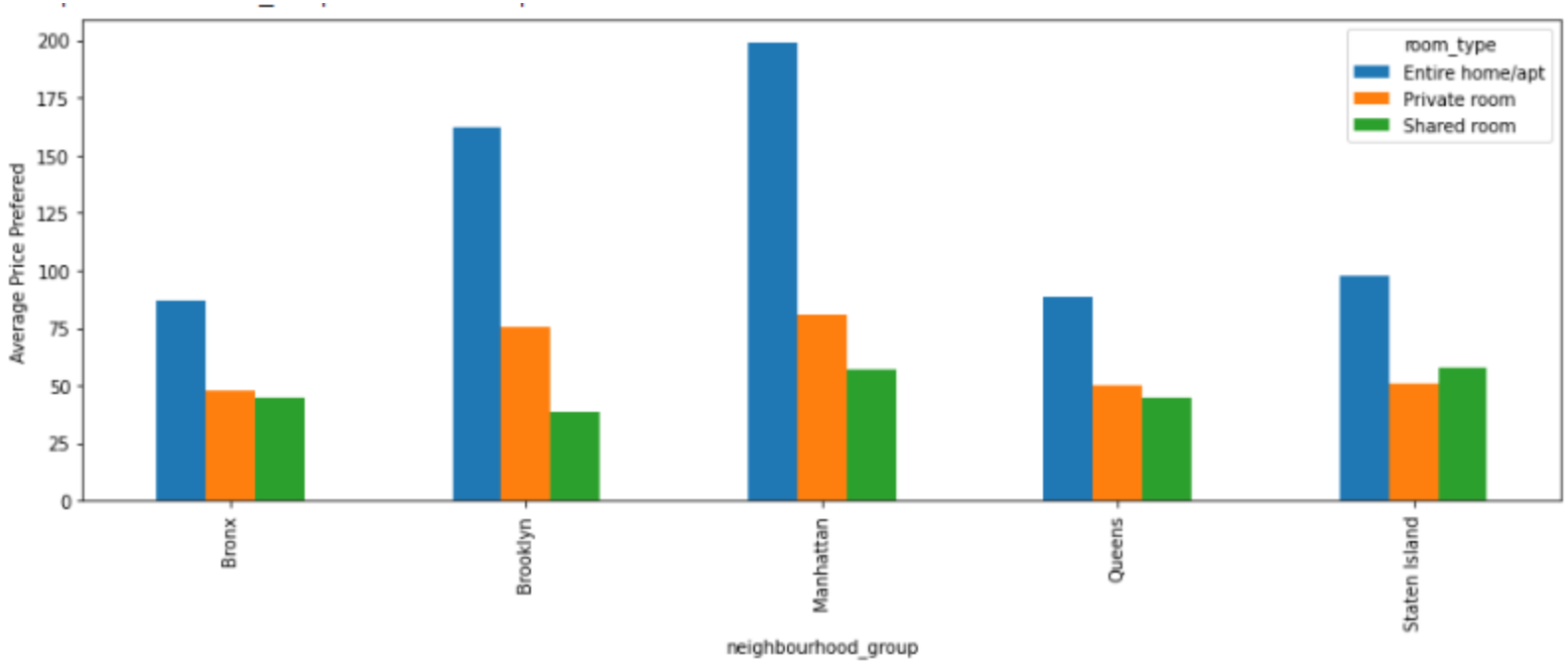
9.What is the average preferred price by customers according to the neighbourhood group for each category of Room type?



As we can see that Manhattan is most costly and Bronx is cheap for each room type

We can make it more useful for business implementation if we do some analysis on successful hosts according to the highest no of reviews so that we can suggest this price to our host for good business.

10.What is the average price preferred for Keeping good number of reviews according to neighbourhood group ?



OBSERVATIONS

- 1) clearly if we compare the results with previous result (i.e when we calculated average preferred price by people in each neighbourhood group with different room types) we can see that this result is bit different and more useful
- 2) As a analyst I would suggest to keep price in this range to get more number of reviews in specific room type and at particular place

11.Which host are the busiest and why?

- A metric is a system of measurement in this case 'busiest' which gives a relative comparison between the hosts.
- The metric mean across various properties for a host gives the average occupancy rate/percentage the host.
- The higher the percentage, the busier a host is said to be.

1) Available months = available days / (365/12)

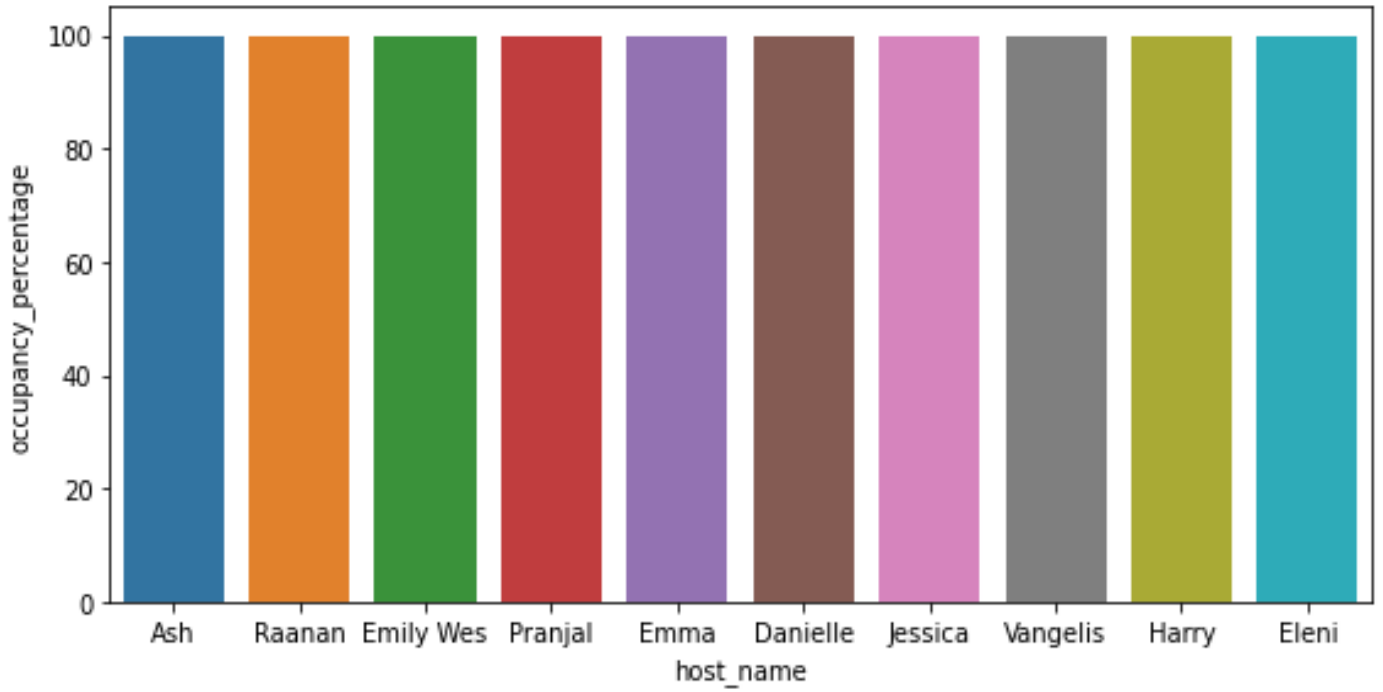
2) Total possible bookings = available days / minimum nights

3) Estimated bookings = reviews per month x available months

Using all the above calculations, the percentage of occupancy throughout the year is gives as :-

Occupancy % = estimated bookings / total possible bookings x 100

The grouped table now contains the average occupancy percentage for every host. Sorting the table to obtain the top 10 busiest hosts for Airbnb.



Conclusion

1. In all the listings registered with airbnb, more than 50% of them offer entire home/apt,45%are for private rooms,1.85 for shared rooms and 0.81 for hotel rooms.
2. The most of airbnb prices are under $1000.
3. Manhattan has the highest range of prices for the listings with an average price of 120 dollars, followed by Brooklyn with 90 dollars per
   night.
4. Queens and Staten island appear to have similar distribution, Bronx is the cheapest among all of them.
5. Minimum number of night stays has significant impact on prices.
6. The machine learning models used in this project, k –nearest neighbours model gives least accuracy and random forest regression
   predicts the sale price with best accuracy.
7. There are almost 50% positive, 37% neutral and 13%negative comments in review dataset.

Thank You