Capstone Project – 1 (EDA)

Airbnb Booking Analysis

By

Team Leader

Ade suchit shrimant

Index

1. About Airbnb
2. Problem Statements
3. Data Exploration
4. Data Featuring
5. Exploratory Data Analysis
6. Conclusion
7. Challenges

About Airbnb

- Airbnb, was founded in 2008 in San Francisco, California.
- It’s original name, Air_Bed and Breakfast.com
- It mainly based on online marketplace.
- It operates as a broker for hotel booking, home and apartment for rentals, based on requirements.
- Company charges a commission from each booking.

Problem Statements

Questions:

with the help of exploratory data analysis techniques, try to answer the following problem statements.

- What can we learn about different hosts and areas?
- On the basis of availability of listings around the year give an availability status to each of the listings.
- Plotting neighbourhood group map with their respective coordinates.
- Show the trend of ‘room type’ for different ‘neighbourhood group’.
- Which apartment have been more used by travellers?
- Which hosts are busiest and why?
- Is there any noticeable difference of traffic among different areas and what could be the reason for it?
- How many properties are available for more than 100 days?
- Check how the average price varies for different room types?
- Pair plot
- Show price update column?

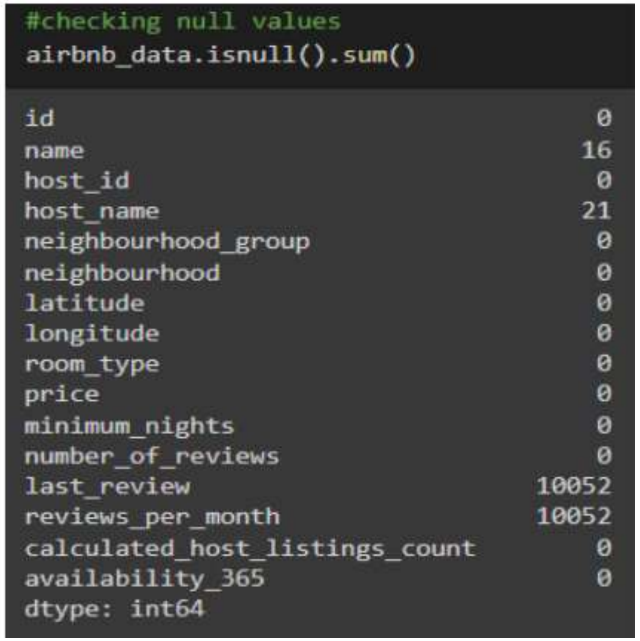
Data Exploration

Data set-Airbnb NYC 2019 contains observation of bookings done in 5 neighbourhood groups in New York City.

- Rows-48895
- Columns- 16
- Room types- Entire homes/apt, private rooms, shared rooms.
- Neighbourhood groups- Manhattan, Queens, Staten Island, Bronx, Brooklyn.

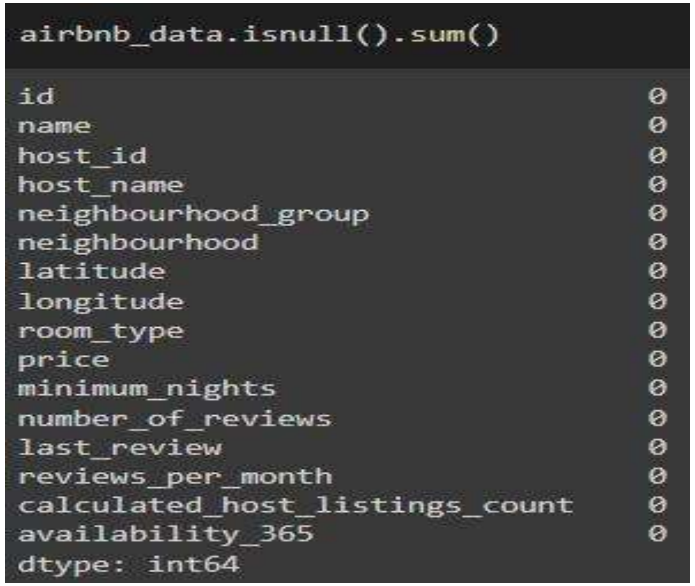
Data exploration

Data set contains null values.



```
#checking null values
airbnb_data.isnull().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
```



```
airbnb_data.isnull().sum()

id                                0
name                              0
host_id                           0
host_name                         0
neighbourhood_group               0
neighbourhood                     0
latitude                          0
longitude                         0
room_type                         0
price                             0
minimum_nights                    0
number_of_reviews                 0
last_review                       0
reviews_per_month                 0
calculated_host_listings_count    0
availability_365                  0
dtype: int64
```

Replacing all the null values.

Data Featuring

Given data set consists the booking information of Airbnb from 2008 till 2019 in a form of csv file.

The features in the dataset can be described as follows:

List of fields	
▪ Id	▪ Latitude
▪ Name	▪ Longitude
▪ Host_id	▪ Price
▪ Host_name	▪ Minimum_nights
▪ Neighbourhood_group	▪ Number_of_reviews
▪ Neighbourhood	▪ Last_review
▪ Room type	▪ Reviews_per_month
▪ Calculated_host_listings_count	▪ Avaliability_365

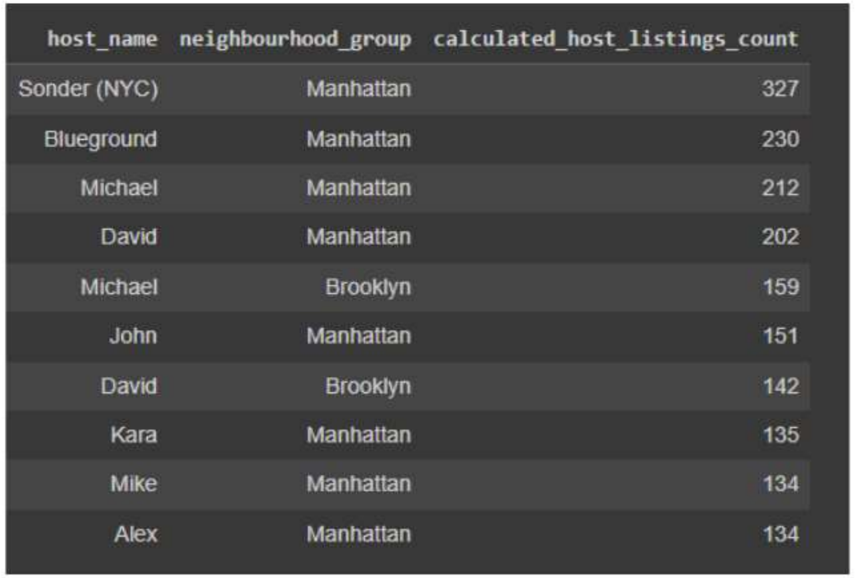
Data Featuring

- Id= unique id assigned to the entry.
- Name= this is a column containing the name provided by each host for customer reference.
- Host id and host name= many hosts serve many objects. This host id and host name contains this records.
- Neighbourhood and neighbourhood group = these columns contain information about the city and area of properties offered by Airbnb New York.
- Longitude and latitude= contains the longitude and latitude of the property location.
- Room type = private room/ entire room and shared room.
- Price= An important column that contains price values for all these properties.
- Minimum nights= this gives you information about the minimum number of nights a host offers for a particular accommodation.
- number of reviews and reviews month= it includes the number of reviews and ratings per month for these accommodations, as well as information about the host’s hospitality.
- Availability 365= provides information about offer availability.

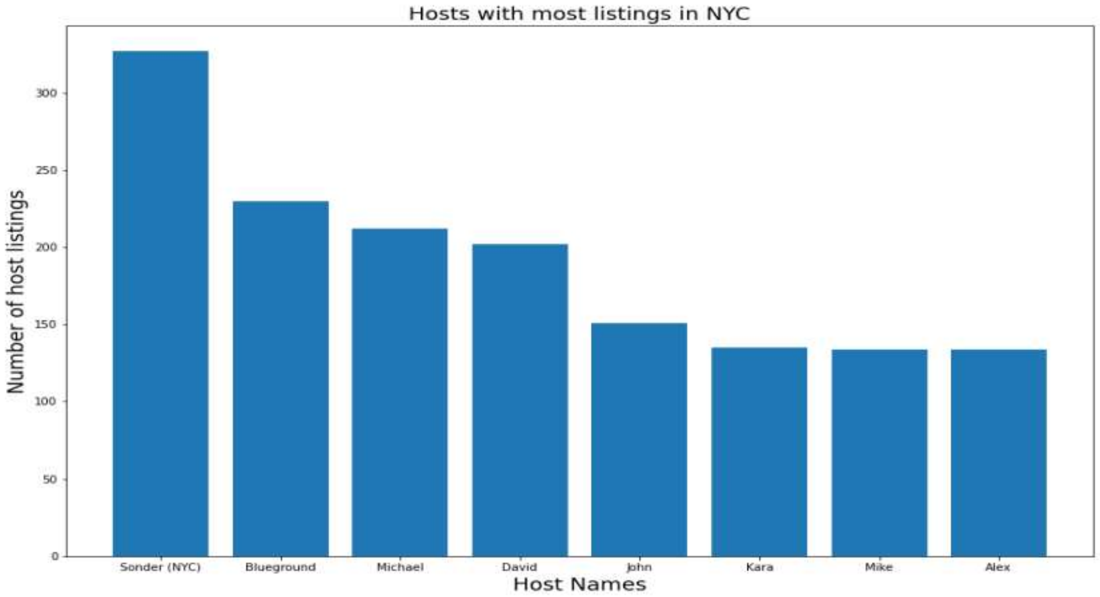
Data Analysis

Host with most listed count in the NYC

- Sonder has the highest count in listing.
- Then blueground and michael. are the highest host count area.



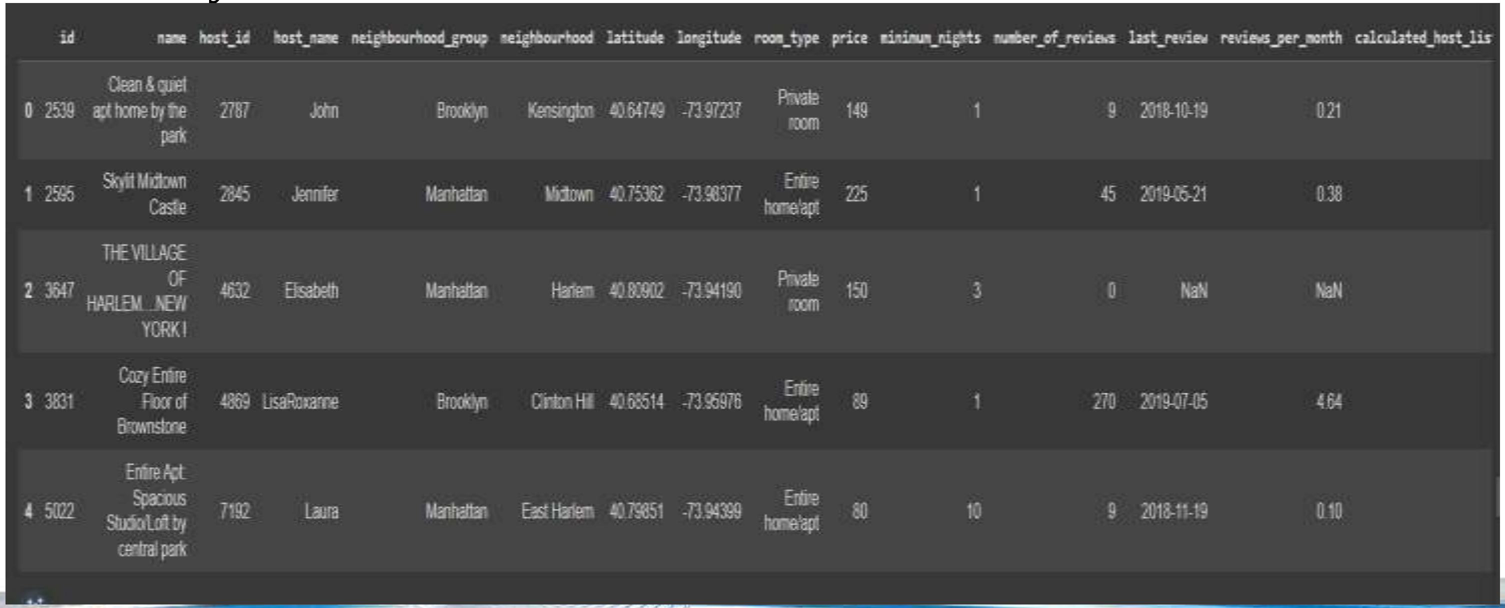
host_name	neighbourhood_group	calculated_host_listings_count
Sonder (NYC)	Manhattan	327
Blueground	Manhattan	230
Michael	Manhattan	212
David	Manhattan	202
Michael	Brooklyn	159
John	Manhattan	151
David	Brooklyn	142
Kara	Manhattan	135
Mike	Manhattan	134
Alex	Manhattan	134



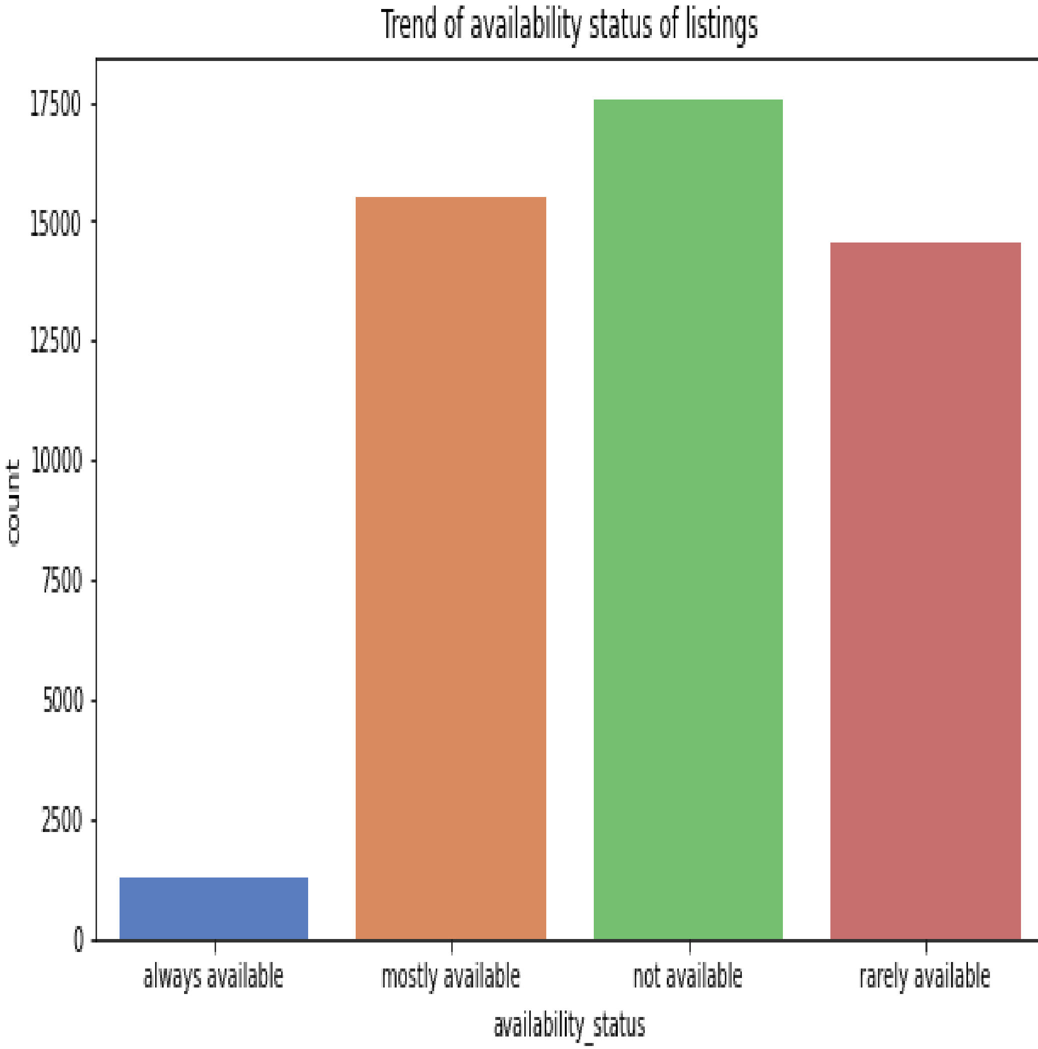
Data Analysis

Availability status

- around the year 17500 rooms are not available
- 1250 rooms are always available in the year.
- 15000 rooms are mostly available in the year.



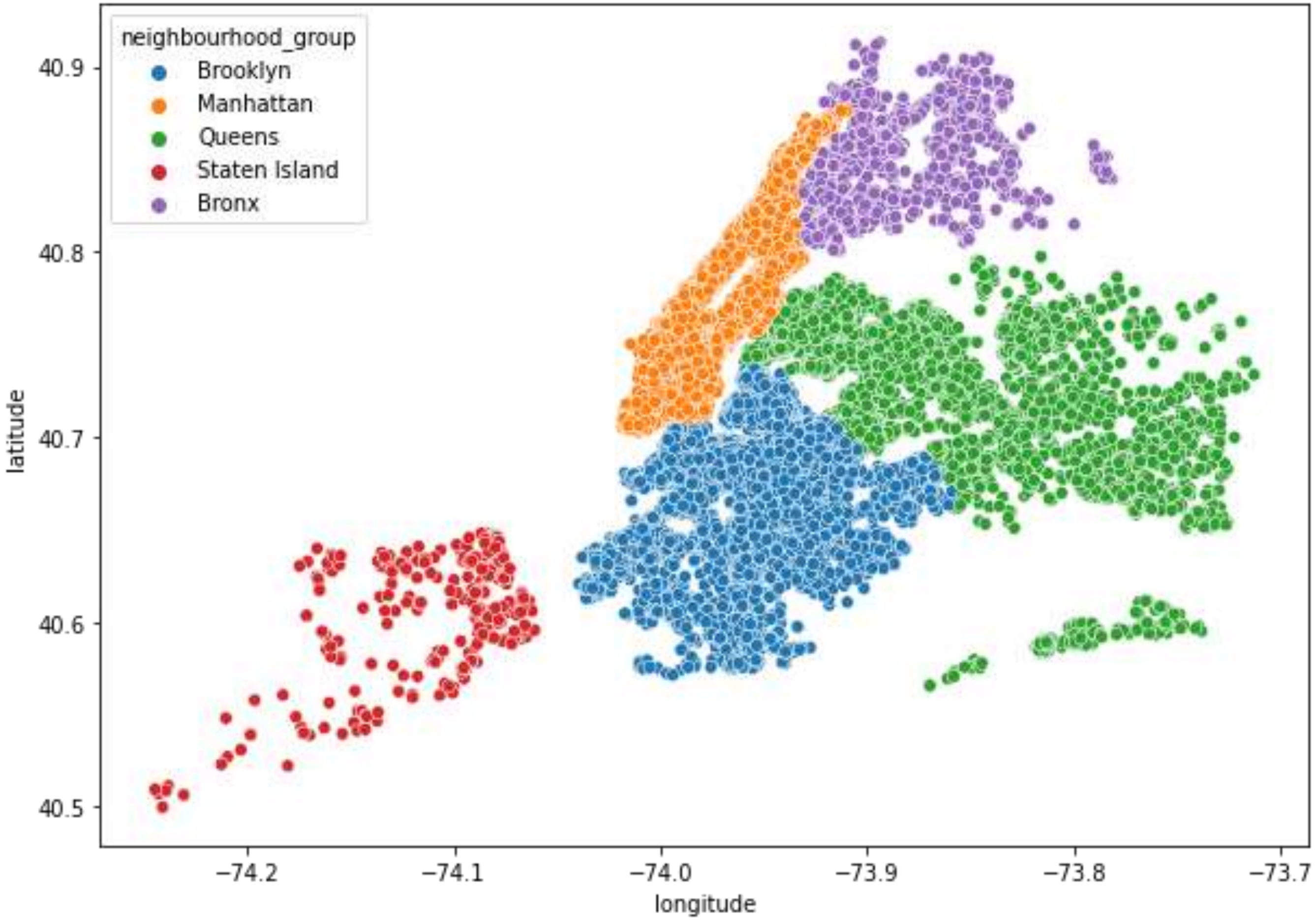
	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review	reviews_per_month	calculated_host_lis
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	9	2018-10-19	0.21	
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	45	2019-05-21	0.38	
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150	3	0	NaN	NaN	
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89	1	270	2019-07-05	4.64	
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	10	9	2018-11-19	0.10	



Data analysis

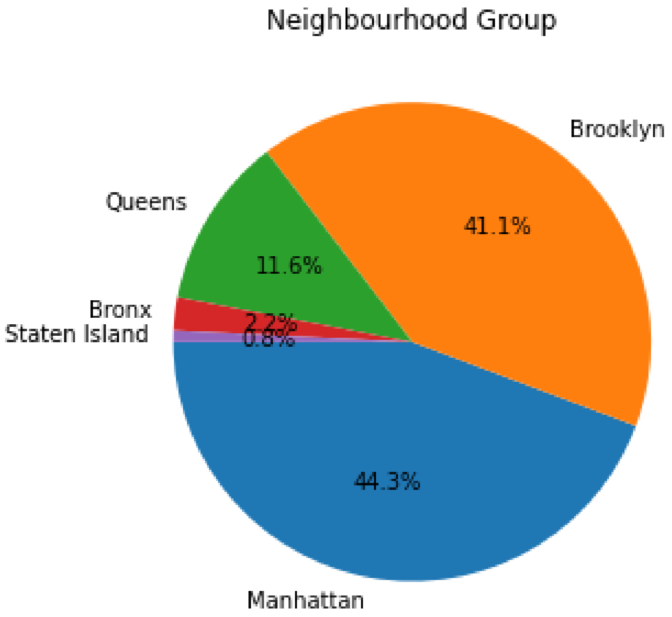
Neighbourhood group with their respective coordinates.

- Queens has a maximum coordinates
- Bronx has a minimum coordinates

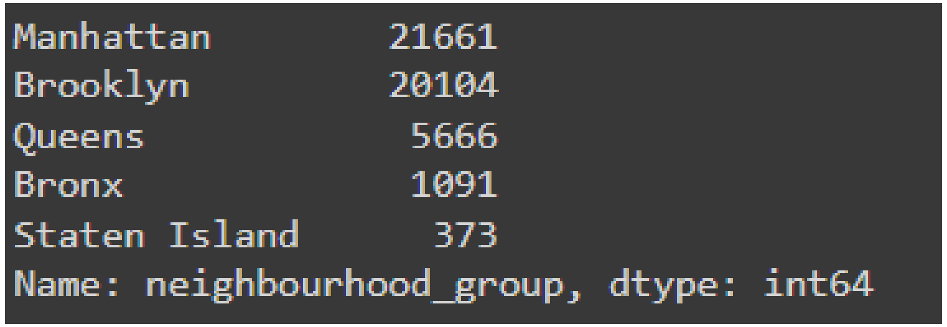


Data analysis

Room type vs different neighbourhood group



- Manhattan has a highest number of neighbourhood group.
- Staten island has a lowest number of neighbourhood group.

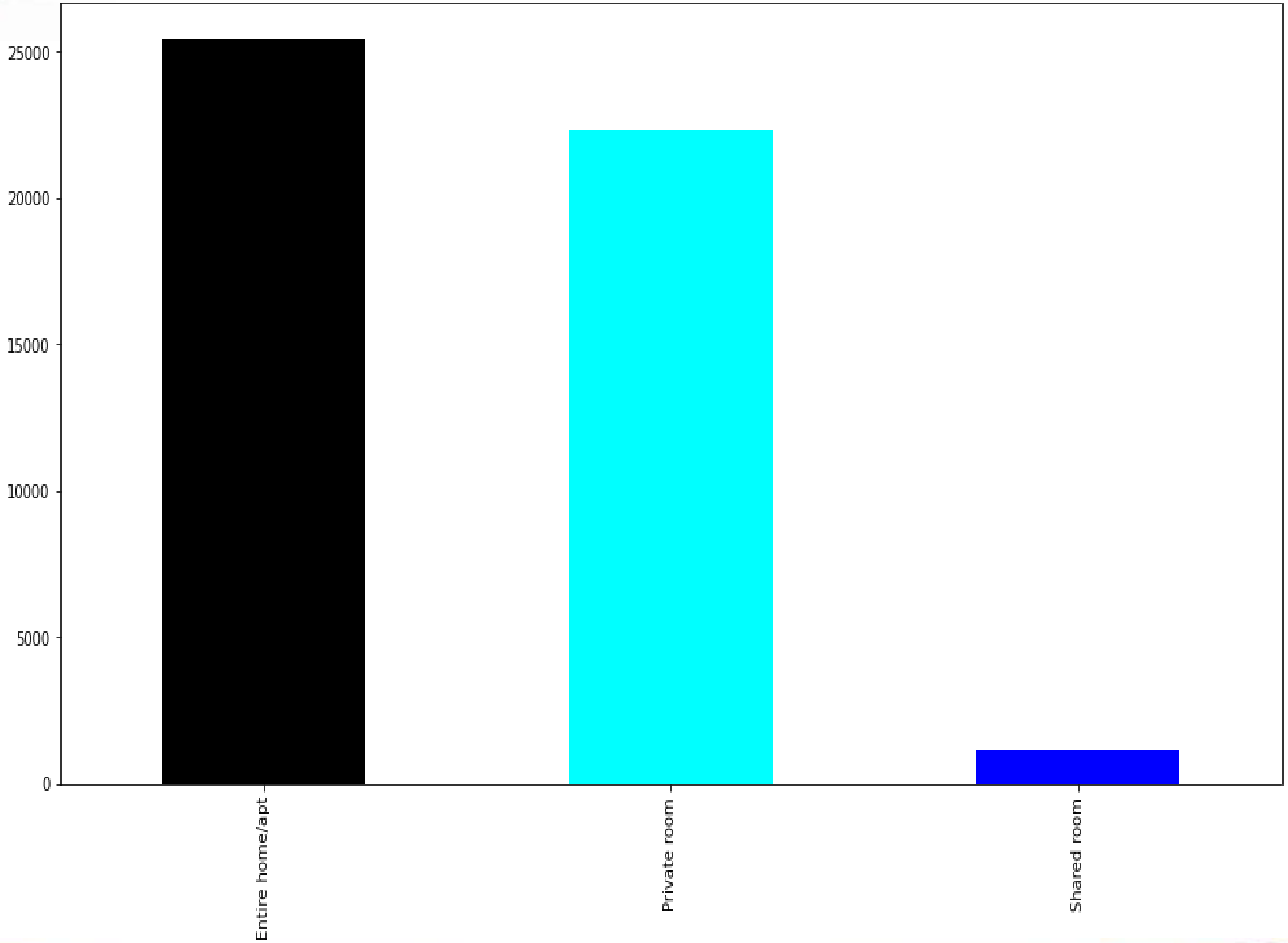


```
Manhattan        21661
Brooklyn         20104
Queens            5666
Bronx             1091
Staten Island      373
Name: neighbourhood_group, dtype: int64
```

Data Analysis

Appartment have been more used by traveller

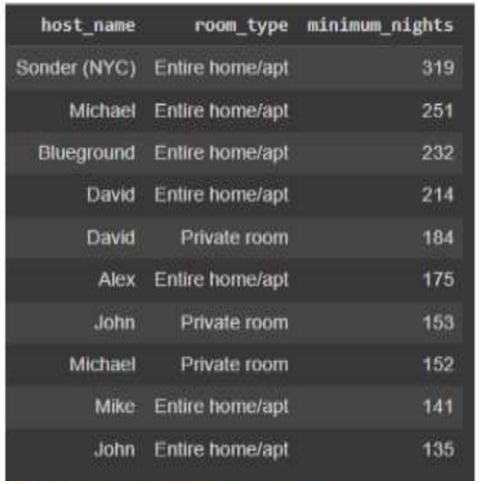
- Entire home/apt is highest used by traveller.
- Shared room is lowest used by traveller.



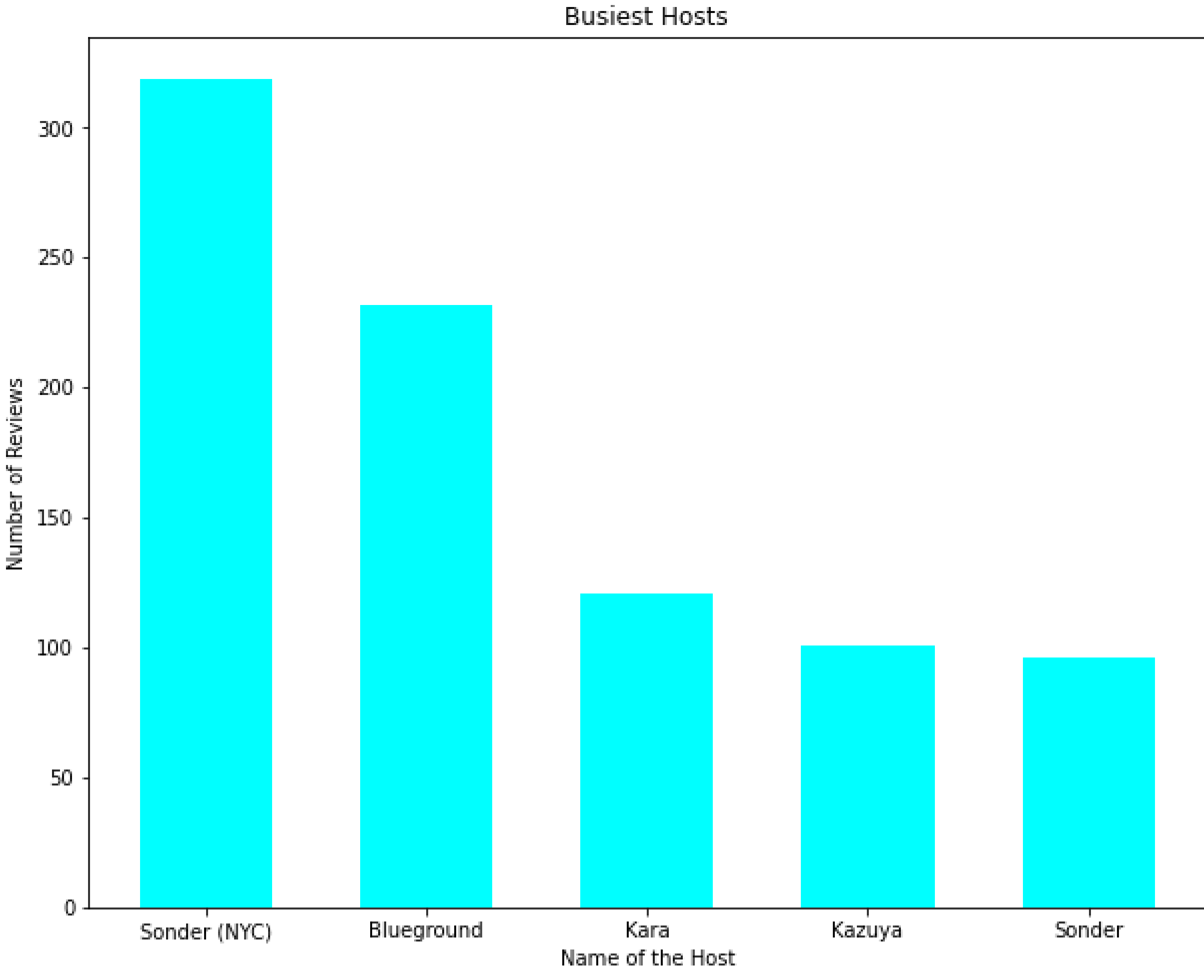
Data Analysis

Busiest host

- Sounder prefer entire room/apt and he is the most busiest host followed by Michael and Blueground.
- Michael, David and John are all on the top 10 but they preference is shared and entire room/apt both.



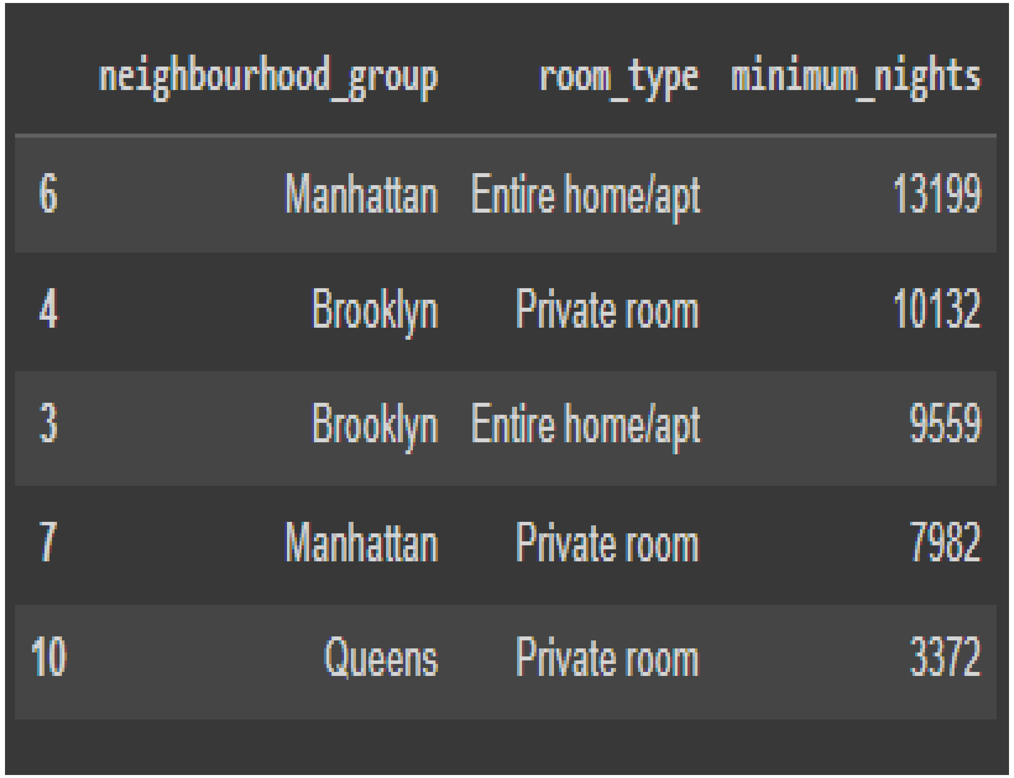
host_name	room_type	minimum_nights
Sonder (NYC)	Entire home/apt	319
Michael	Entire home/apt	251
Blueground	Entire home/apt	232
David	Entire home/apt	214
David	Private room	184
Alex	Entire home/apt	175
John	Private room	153
Michael	Private room	152
Mike	Entire home/apt	141
John	Entire home/apt	135



Data Analysis

Traffic areas

- Among the heavy traffic areas are manhattan and Brooklyn neighbourhood_group.

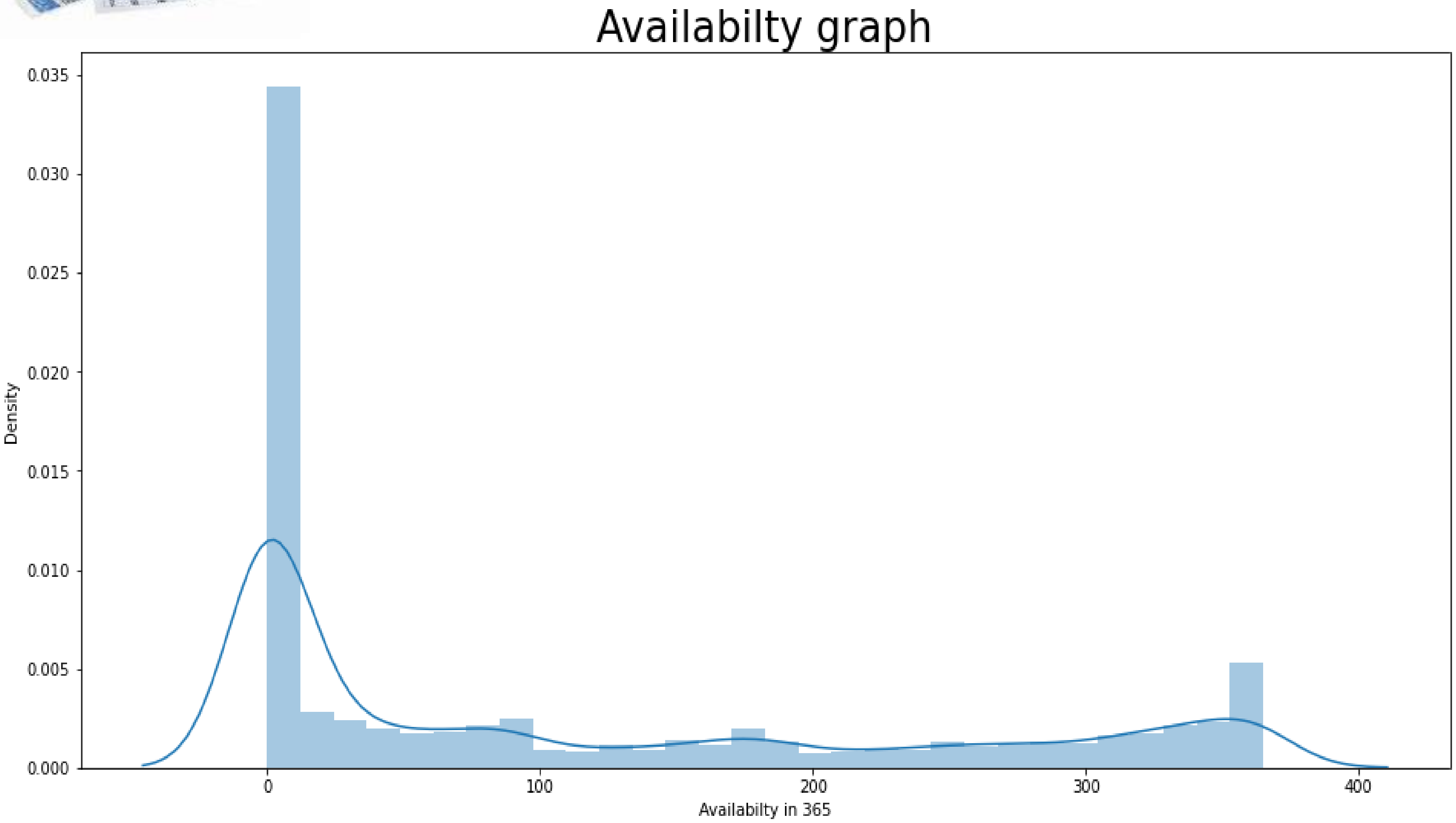


	neighbourhood_group	room_type	minimum_nights
6	Manhattan	Entire home/apt	13199
4	Brooklyn	Private room	10132
3	Brooklyn	Entire home/apt	9559
7	Manhattan	Private room	7982
10	Queens	Private room	3372

Data Analysis

Properties available more than 100 days

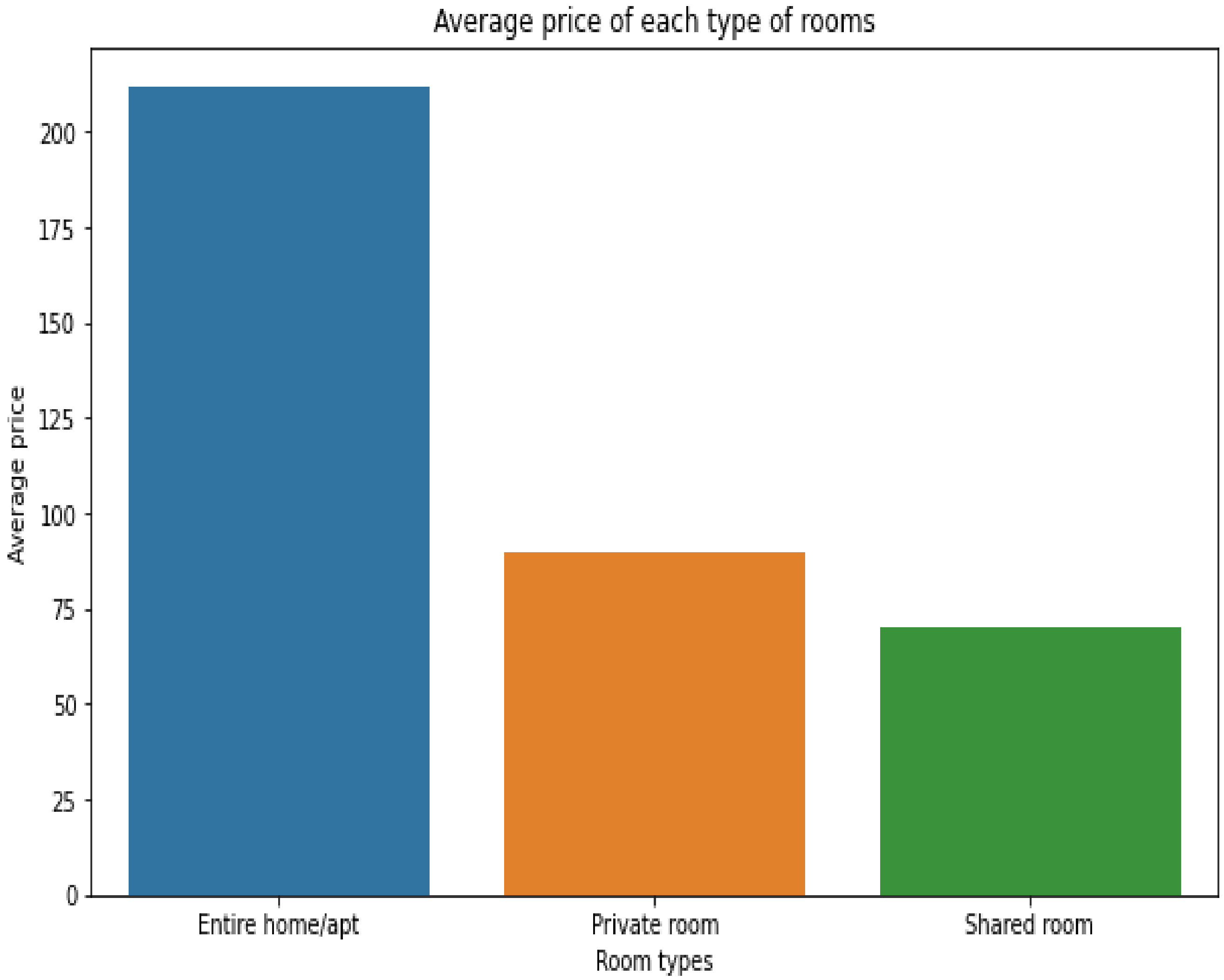
- 39.19% properties are availabel for more than 100 days. Most of the properties doesn’t work for full year.



Data Analysis

Avarage price vs different room type

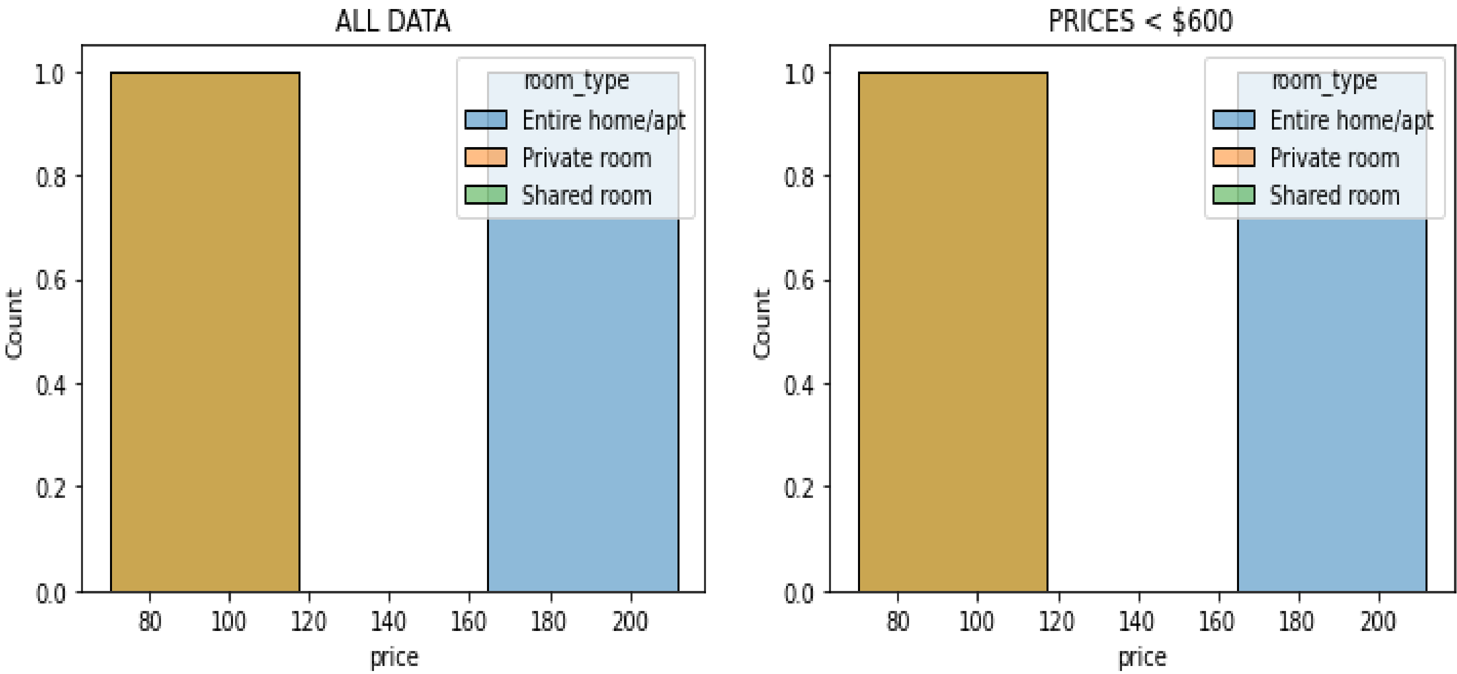
- Entire home/apt has the highest price.
- Shared room has the lowest price.
- Private room has the average price .



Data analysis

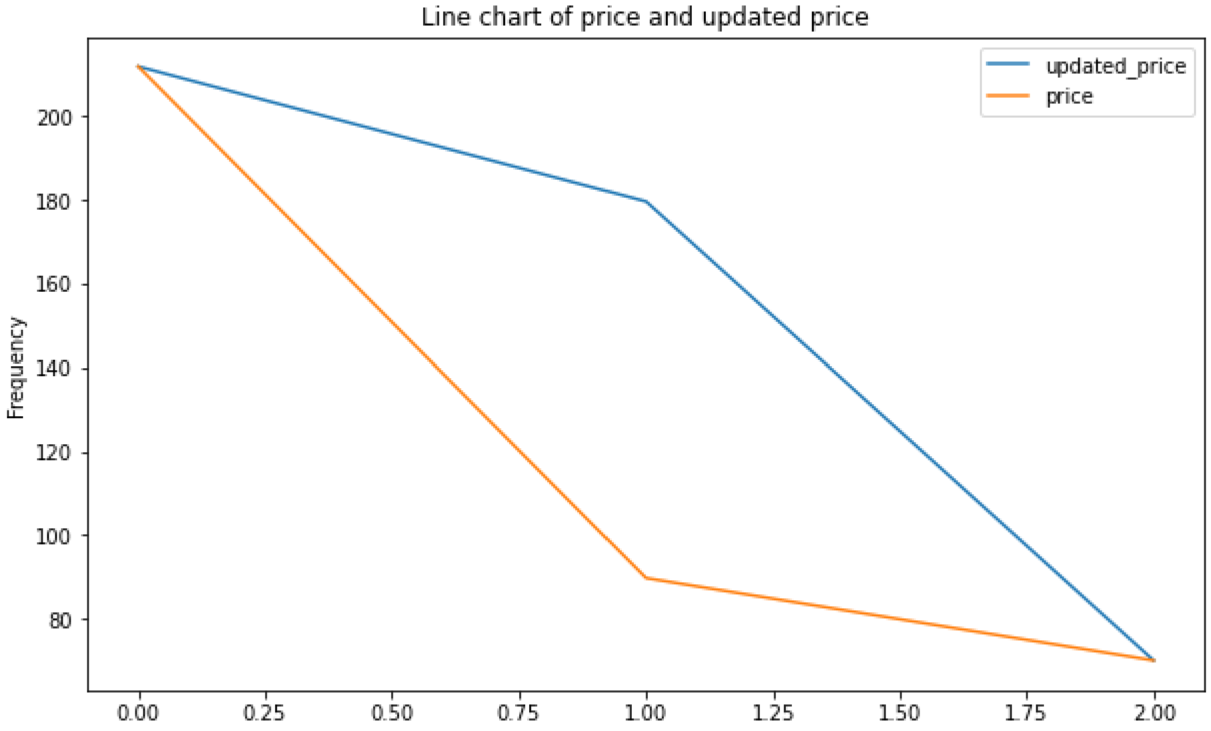
Pair plot

- From the above histplot shows the relation between the price of each room type thats is private room vs Entire room/apt vs shared room .From the first figure we are showing that Price less than 600 & From the 2nd data we are showing the prices of Private room,Entire room/apt and shared room less than $600.



Data analysis

Relationship between the price and updated price.



- From the above line chart we are showing the relationship between the price and updated price . Hence, we are observing that there is bit vartion between price and updated price sometimes it is high and sometimes its low. It will vary.

conclusion

1. Most number of listing from Manhattan by host name Sonder(NYC) and then blureground and Michael.
2. 1250 rooms are always available in the year and 17500 rooms are not available in the year.
3. In neighbourhood group maximum coordinates is Queens and Brooklyn and minimum coordinates is staten island and Bronx
4. Manhattan has highest number in Airbnb spread across the area of neighbourhood group. Then Brooklyn then queens.
5. The entire home and private rooms has the maximum used by traveller. and shared rooms minimum used by traveller.
6. Sonder(NYC) is the most bussiest host and then Michael. Because these host listed room type as entire room and private roo m which is preferred by most number of people.
7. among the heavy traffic areas are Manhattan and Brooklyn neighbourhood_group.
8. There are only 39.19% of properties are available for more than 100 days.
9. The most expensive room types are the Entire home and then private rooms.
10. From the above histplot shows the relation between the price of each room type that's is private room vs Entire room/apt vs shared room .From the first figure we are showing that Price less than 600 & From the 2nd data we are showing the prices o f Private room , Entire room/apt and shared room less than $600.
11. From the above line chart we are showing the relationship between the price and updated price . Hence, we are observing th at there is bit vartion between price and updated price sometimes it is high and sometimes its low. It will vary.

challenges

- Huge chunk of data was to be handled keeping in mind not to miss anything which is even of little relevance.
- Computation time.

Thank you.