



Airbnb EDA

- *Creating Visualizations and storytelling*

Dataset : Airbnb NYC

Project Summary

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.

This dataset has around **49,000** observations in it with **16** columns and it is a mix between *categorical and numeric values*.he platform, guiding marketing initiatives, implementation of innovative additional services and much more.

This dataset has around 49,000 observations in it with 16 columns and it is a mix between categorical and numeric values.



Problem Statement:

Let's Explore and analyze the Data Set and find some insights(Few Questions Listed Below)

- 1.What can we learn about different hosts and areas?*
- 2.What we learn from room type and their prices according to area?*
- 3.What can we learn from Data? (ex: locations, prices, reviews, etc)*
- 4.Which hosts are the busiest and why is the reason?*
- 5.Which Hosts are charging higher price?*
- 6.Is there any traffic difference among different areas and what could be the reason for it?*
- 7.What is the correlation between different variables ?*
- 8.What is the room count in overall NYC according to the listing of room types?ta? (ex: locations, prices, reviews, etc)*

Describing the dataset :

Data Sample :

	id	host_id	latitude	longitude	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365
count	4.889500e+04	4.889500e+04	48895.000000	48895.000000	48895.000000	48895.000000	48895.000000	38843.000000	48895.000000	48895.000000
mean	1.901714e+07	6.762001e+07	40.728949	-73.952170	152.720687	7.029962	23.274466	1.373221	7.143982	112.781327
std	1.098311e+07	7.861097e+07	0.054530	0.046157	240.154170	20.510550	44.550582	1.680442	32.952519	131.622289
min	2.539000e+03	2.438000e+03	40.499790	-74.244420	0.000000	1.000000	0.000000	0.010000	1.000000	0.000000
25%	9.471945e+06	7.822033e+06	40.690100	-73.983070	69.000000	1.000000	1.000000	0.190000	1.000000	0.000000
50%	1.967728e+07	3.079382e+07	40.723070	-73.955680	106.000000	3.000000	5.000000	0.720000	1.000000	45.000000
75%	2.915218e+07	1.074344e+08	40.763115	-73.936275	175.000000	5.000000	24.000000	2.020000	2.000000	227.000000
max	3.648724e+07	2.743213e+08	40.913060	-73.712990	10000.000000	1250.000000	629.000000	58.500000	327.000000	365.000000

Getting the info of the data :

```
<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
```

Checking for any missing values in the data:

4 columns have null values-

- *name*
- *host_name*
- *last_review*
- *reviews_per_month*

```
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
```

Getting all the attributes in the dataset:

The given are the attributes present in the dataset

```
Index(['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_listings_count',  
      'availability_365'],  
      dtype='object')
```

So, we drop the irrelevant columns from our dataset which are-

- *latitude*
- *longitude*
- *last_review*
- *reviews_per_month*

Displaying the first 10 values in the data:

Now we have the data looking like this-

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	room_type	price	minimum_nights	number_of_reviews	calculated_host_listings_count	availability_365
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	Private room	149	1	9	6	365
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	Entire home/apt	225	1	45	2	355
2	3647	THE VILLAGE OF HARLEM... NEW YORK I	4632	Elisabeth	Manhattan	Harlem	Private room	150	3	0	1	365
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	Entire home/apt	89	1	270	1	194
4	5022	Entire Apt. Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	Entire home/apt	80	10	9	1	0
5	5099	Large Cozy 1 BR Apartment In Midtown East	7322	Chris	Manhattan	Murray Hill	Entire home/apt	200	3	74	1	129
6	5121	BlissArtsSpace!	7356	Garon	Brooklyn	Bedford-Stuyvesant	Private room	60	45	49	1	0
7	5178	Large Furnished Room Near B'way	8967	Shunichi	Manhattan	Hell's Kitchen	Private room	79	2	430	1	220
8	5203	Cozy Clean Guest Room - Family Apt	7490	MaryEllen	Manhattan	Upper West Side	Private room	79	2	118	1	0
9	5238	Cute & Cozy Lower East Side 1 bdrm	7549	Ben	Manhattan	Chinatown	Entire home/apt	150	1	160	4	188

Solutions as per our problem statements::

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

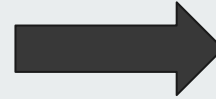
	host_name	neighbourhood_group	calculated_host_listings_count
13217	Sonder (NYC)	Manhattan	327
1834	Blueground	Manhattan	232
1833	Blueground	Brooklyn	232
7275	Kara	Manhattan	121
7480	Kazuya	Queens	103

- We find that Host name **Sonder(NYC)** has listed highest number of listings in **Manhattan** followed by Blueground

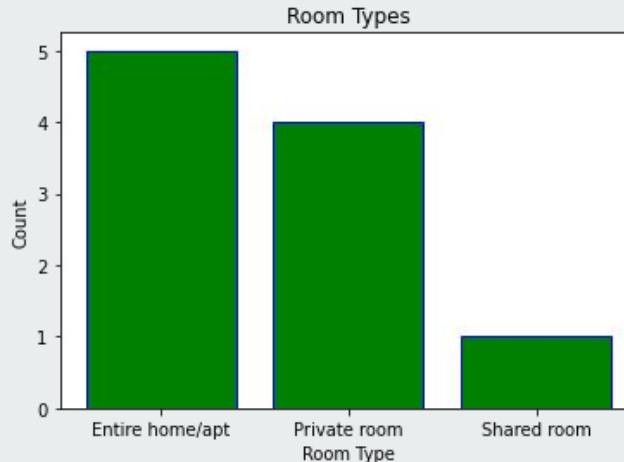
2.What we learn from room type and their prices according to area?

- The prices for Entire home/apt are the highest in Brooklyn and Manhattan.
- For the Room type and their respective demands:

We found that **Entire home/apt** is the highest number of room types overall and prices are high in the **brooklyn** and **Manhattan** for entire home/apt.



	neighbourhood_group	room_type	price
3	Brooklyn	Entire home/apt	10000
6	Manhattan	Entire home/apt	10000
10	Queens	Private room	10000
7	Manhattan	Private room	9999
4	Brooklyn	Private room	7500
12	Staten Island	Entire home/apt	5000
9	Queens	Entire home/apt	2600
1	Bronx	Private room	2500
11	Queens	Shared room	1800
0	Bronx	Entire home/apt	1000

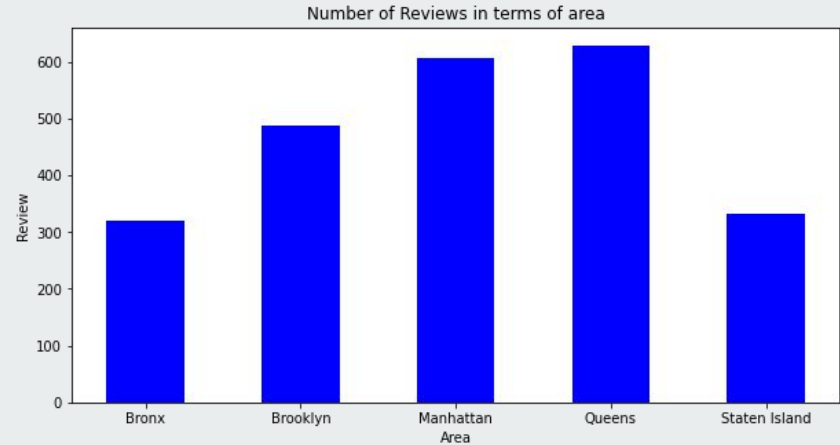


3.What can we learn from Data? (ex: locations, prices, reviews, etc)

- For No. of reviews depending upon the area:

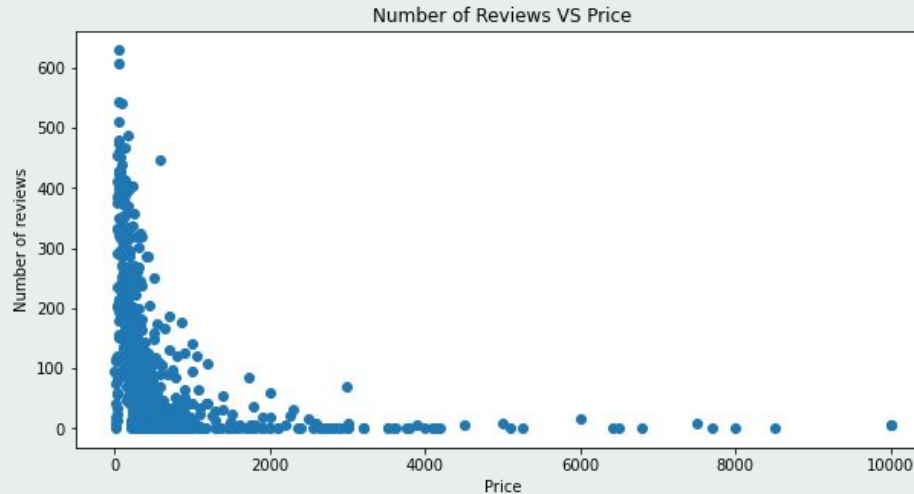
Queens has the highest no. of reviews followed by **Manhattan** and **Brooklyn**

	neighbourhood_group	number_of_reviews
0	Bronx	321
1	Brooklyn	488
2	Manhattan	607
3	Queens	629
4	Staten Island	333



- For Number of reviews vs the price :

From the given visualization we can say that most number of people like to **stay in less price and their reviews are higher in those areas**



	price	number_of_reviews
0	0	95
1	10	93
2	11	113
3	12	8
4	13	9
5	15	19
6	16	43
7	18	1
8	19	76
9	20	116

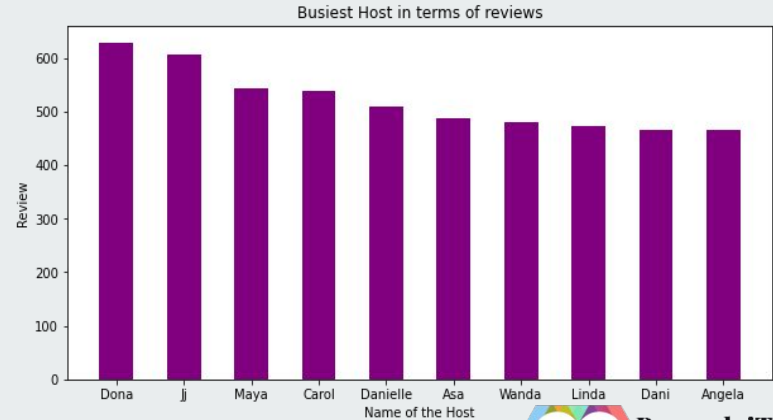
4. Which hosts are the busiest and why is the reason?

So, from the visualizations, we have found the busiest hosts as:

1. ***Dona***
2. ***Jj***
3. ***Maya***
4. ***Carol***
5. ***Danielle***

Because these hosts listed their room type as Entire home and Private room which is preferred by most number of people and also their reviews are higher.

	host_id	host_name	room_type	number_of_reviews
24484	47621202	Dona	Private room	629
7707	4734398	Jj	Private room	607
22213	37312959	Maya	Private room	543
4590	2369681	Carol	Private room	540
19443	26432133	Danielle	Private room	510
13707	12949460	Asa	Entire home/apt	488
1974	792159	Wanda	Private room	480
5056	2680820	Linda	Private room	474
161	42273	Dani	Entire home/apt	467
18289	23591164	Angela	Private room	466



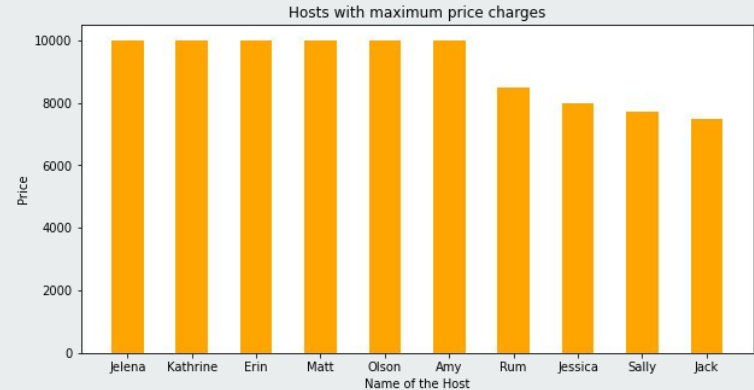
5. Which Hosts are charging higher price?

- From the graphs, we have seen that the 10 Hosts who are charging maximum price are :

Jelena, Kathrine, Erin, Matt, Olson, Amy, Rum, Jessica, Sally, Jack

& the *Max Price* is **10000 USD**

	host_id	host_name	room_type	neighbourhood_group	price
27920	72390391	Jelena	Entire home/apt	Manhattan	10000
17060	20582832	Kathrine	Private room	Queens	10000
8055	5143901	Erin	Entire home/apt	Brooklyn	10000
7325	4382127	Matt	Entire home/apt	Manhattan	9999
2659	1235070	Olson	Entire home/apt	Manhattan	9999
6628	3906464	Amy	Private room	Manhattan	9999
16096	18128455	Rum	Entire home/apt	Manhattan	8500
2561	1177497	Jessica	Entire home/apt	Brooklyn	8000
33424	156158778	Sally	Entire home/apt	Manhattan	7703
10113	7407743	Jack	Entire home/apt	Manhattan	7500



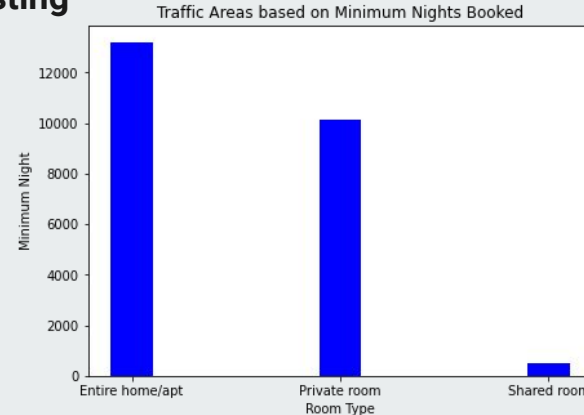
6. Is there any traffic difference among different areas and what could be the reason for it?

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	room_type	price	minimum_nights	number_of_reviews	calculated_host_listings_count	availability_365
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	Private room	149	1	9	6	365
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	Entire home/apt	225	1	45	2	355
2	3647	THE VILLAGE OF HARLEM.....NEW YORK I	4632	Elisabeth	Manhattan	Harlem	Private room	150	3	0	1	365
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	Entire home/apt	89	1	270	1	194
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	Entire home/apt	80	10	9	1	0
5	5099	Large Cozy 1 BR Apartment In Midtown East	7322	Chris	Manhattan	Murray Hill	Entire home/apt	200	3	74	1	129
6	5121	BlissArtsSpace!	7356	Garon	Brooklyn	Bedford-Stuyvesant	Private room	60	45	49	1	0
7	5178	Large Furnished Room Near B'way	8967	Shunichi	Manhattan	Hell's Kitchen	Private room	79	2	430	1	220
8	5203	Cozy Clean Guest Room - Family Apt	7490	MaryEllen	Manhattan	Upper West Side	Private room	79	2	118	1	0
9	5238	Cute & Cozy Lower East Side 1 bdrm	7549	Ben	Manhattan	Chinatown	Entire home/apt	150	1	160	4	188

- For Traffic areas according to the minimum no. of night stays
- From this visualization, We find that, **most of the people like to stay at Entire home/apt and Private room** which are present in *Manhattan, Brooklyn & Queens*.
- Visitors **prefer to stay in rooms whose listing price is less**.

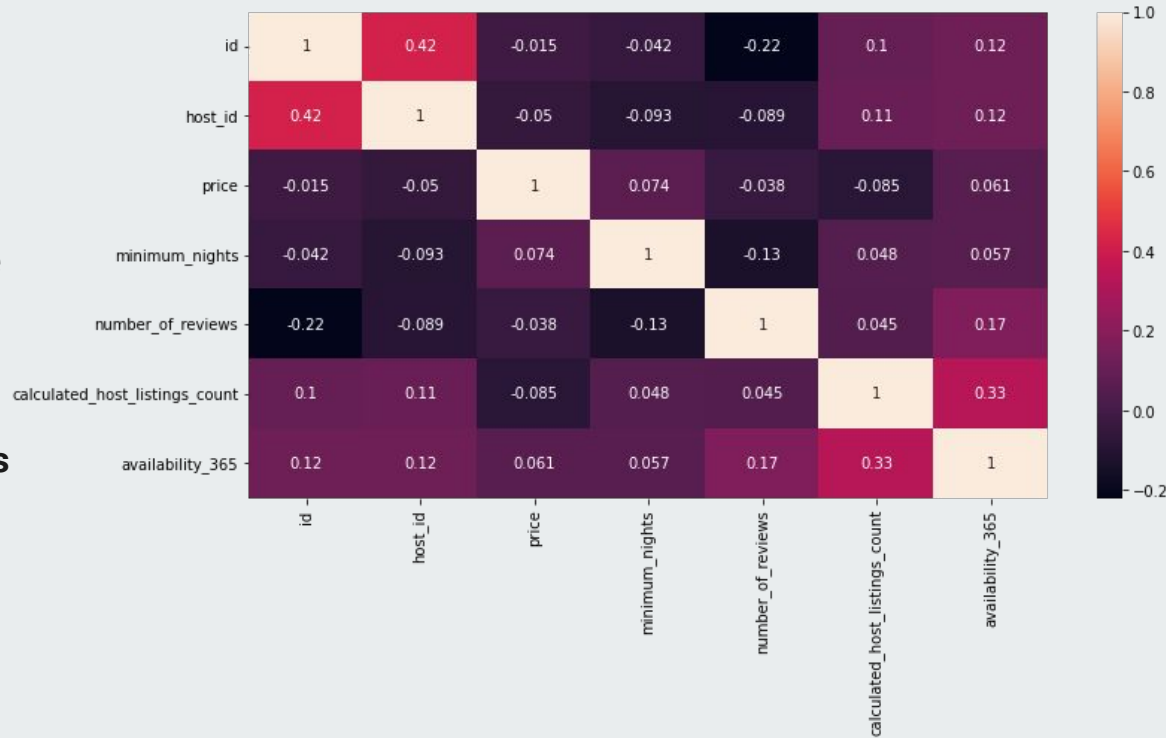


	neighbourhood_group	room_type	minimum_nights
6	Manhattan	Entire home/apt	13199
4	Brooklyn	Private room	10132
3	Brooklyn	Entire home/apt	9559
7	Manhattan	Private room	7982
10	Queens	Private room	3372
9	Queens	Entire home/apt	2096
1	Bronx	Private room	652
8	Manhattan	Shared room	480
5	Brooklyn	Shared room	413
0	Bronx	Entire home/apt	379



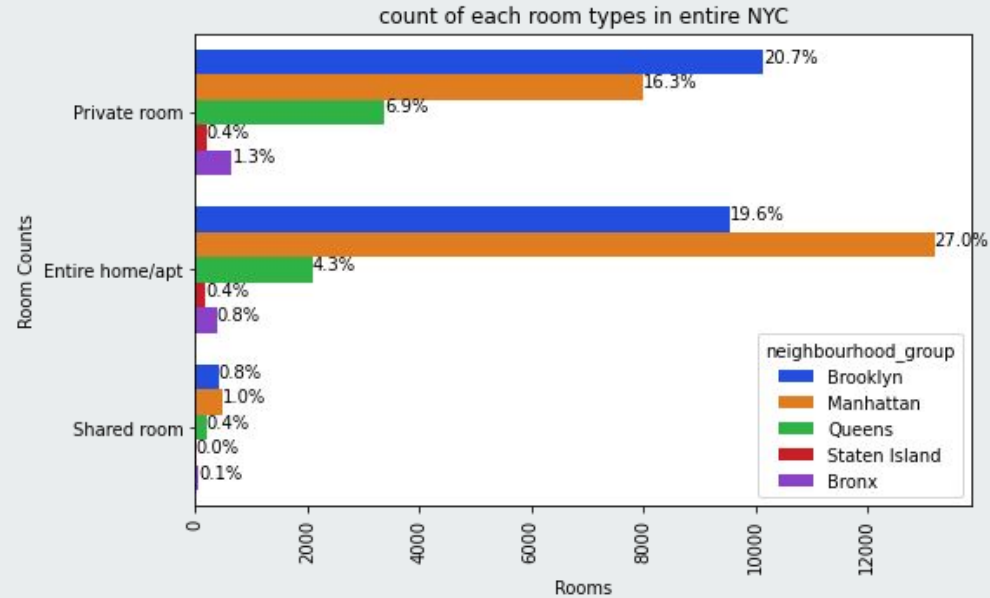
7. What is the correlation between different variables ?

- This is a heatmap that shows the **correlation coefficients between the different columns in the airbnb DataFrame.**
- The **darker the color, the lesser is the correlation** between Attributes.



8. What is the room count in overall NYC according to the listing of room types?

- **Manhattan** has the most listed properties with Entire home/apt around 27% of the total listed properties followed by **Brooklyn** → 19.6%.
- Private rooms are more in **Brooklyn** as in 20.7% of the total listed properties followed by **Manhattan** with 16.3% of them. While 6.9% of private rooms are from **Queens**.
- We can infer that **Brooklyn, Queens, Bronx** has more private room types while **Manhattan** which has the highest no of listings in the entire NYC.






Conclusion:

1. We find that Host name **Sonder(NYC)** has listed highest number of listings in **Manhattan** followed by Blueground.
2. We found that Entire home/apt is the highest number of room types overall and prices are high in the **Brooklyn** and **Manhattan** for entire home/apt.
3. Most number of people like to stay in less price and their reviews are higher in those areas.
4. We have found Busiest hosts : **Dona, Jj, Maya, Carol, Danielle**

Because these hosts listed their room type as Entire home and Private room which is preferred by most number of people and also their reviews are higher.

5. 10 Hosts who are charging maximum price are: **Jelena, Kathrine, Erin, Matt, Olson, Amy, Rum, Jessica, Sally & Jack.**

Max Price is **10000 USD**

- 
6. Most of the people like to stay at Entire home and Private room which are present in **Manhattan, Brooklyn & Queens**

&

also visitors prefer to stay in room whose listing price is less.

7. The correlation between different variables are shown in the heatmap
8. Manhattan has more listed properties with Entire home/apt around 27% of total listed properties followed by Brooklyn with around 19.6%. Private rooms are more in Brooklyn as in 20.7% of the total listed properties followed by Manhattan with 16.3% of them. While 6.9% of private rooms are from Queens. We can infer that Brooklyn,Queens,Bronx has more private room types while Manhattan which has the highest no of listings in entire NYC has more Entire home/apt room types.



Thank You !