Capstone Project

Airbnb Bookings Analysis

Contents

- Introduction
- Data Summary
- Variable Identification
- Handling NaN values
- Finding Correlation
- Exploring and Visualizing Data
- ‘Price Feature’
- Conclusion

Introduction

Airbnb was conceived years ago by two roommates who rented out an air mattress in their living room. This turned their whole apartment into a bed and breakfast. This was done to sustain the high-priced living in San Francisco. This San-Francisco based start-up offers you someone’s home as a place to stay instead of a hotel. Airbnb was started in 2008. 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 services 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.

Data Summary

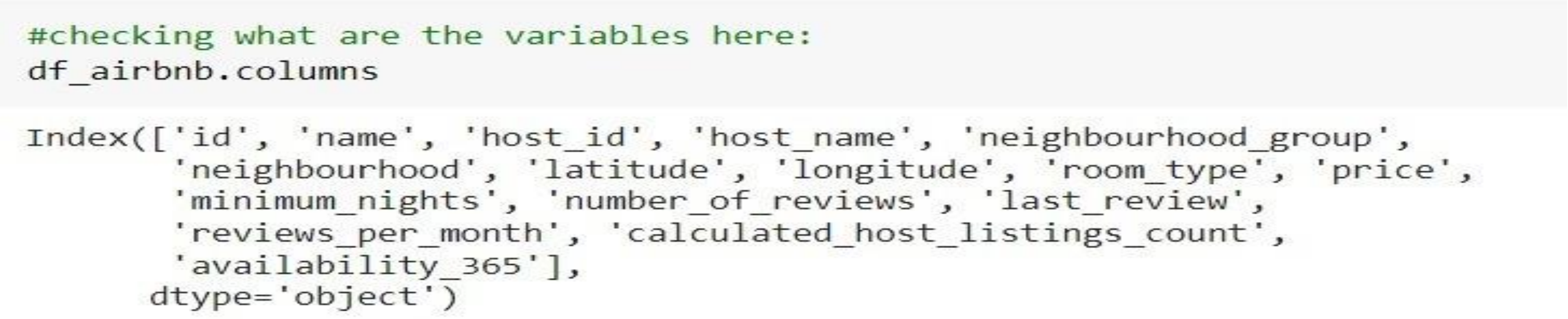
- **'id' :-** This column represents property id
- **'name' :-** This column represents property Name and Description
- **'host_id' :-** Particular properties were hosted by particular hosts who are represented by host id column
- **'host_name' :-** Particular properties were hosted by particular hosts who are represented by host name column
- **'neighbourhood _group' :-** It represent cities in New York i.e. 'Brooklyn', 'Manhattan', 'Queens', 'Staten Island', 'Bronx'
- **'neighbourhood ' :-** It's represented particular area in particular city for example, midtown is one of area in Manhattan city
- **'latitude', 'longitude' :-** These columns represent location of particular Airbnb listing

Data Summary

- **‘room_type’ :-** It contain three categorical values i.e. 'Private room', 'Entire home/apt', 'Shared room’ which represent the what is room type
- **‘price’ :-** It represent the price of particular room per night
- **‘minimum_nights’ :-** Represent minimum night spend by guests in particular host’s listing
- **‘number_of_reviews’, ‘last_review’, ‘reviews_per_month’ :-** Represents number of reviews, last review, reviews per month of specific listing
- **‘calculated_host_listings_count’ :-** Represent total number of times host listed property
- **‘availability_365’ :-** Represent number of days available in year for specific listing

Variable Identification

In this step we find out various **Categorical** and **Numerical** variables



```
#checking what are the variables here:
df_airbnb.columns
```

```
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')
```

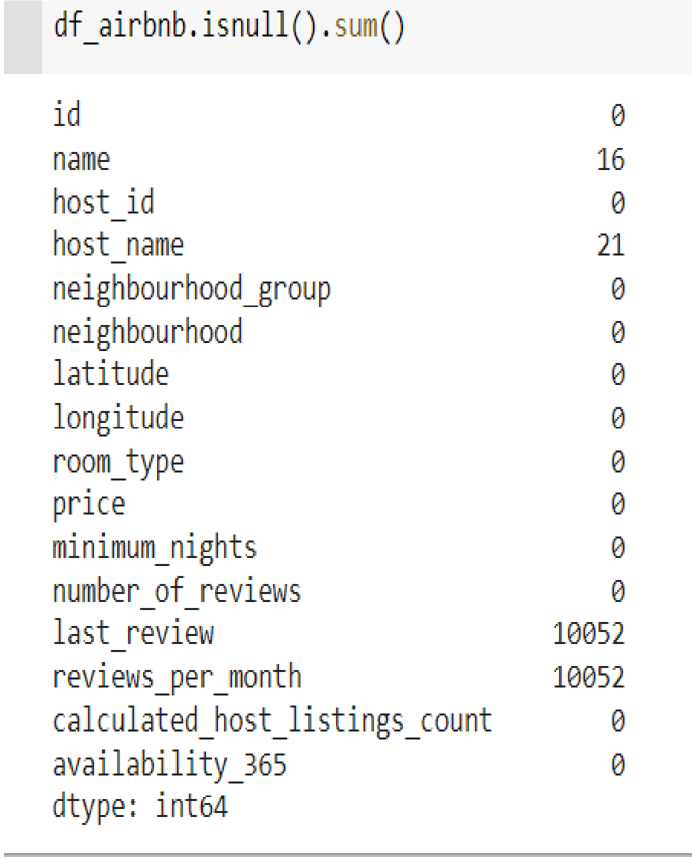
- By observations we get to that '**name**' column represents property name and particular properties were hosted by particular hosts who are represented in 'host_name' column. But a particular host_name can have multiple properties in an area. So, the ‘host_name’ is one of **categorical variables like neighbourhood (areas), neighbourhood _group, and room_type.**

- **id, latitude, longitude, price, minimum_nights, number_of_reviews, last_review, reviews_per_month, calculated_host_listings_count, availability_365 are numerical variables.**

Handling NaN value

we either dropped the column (which are not that much important in our analysis) or replaced NaN values with some relevant substitutes.

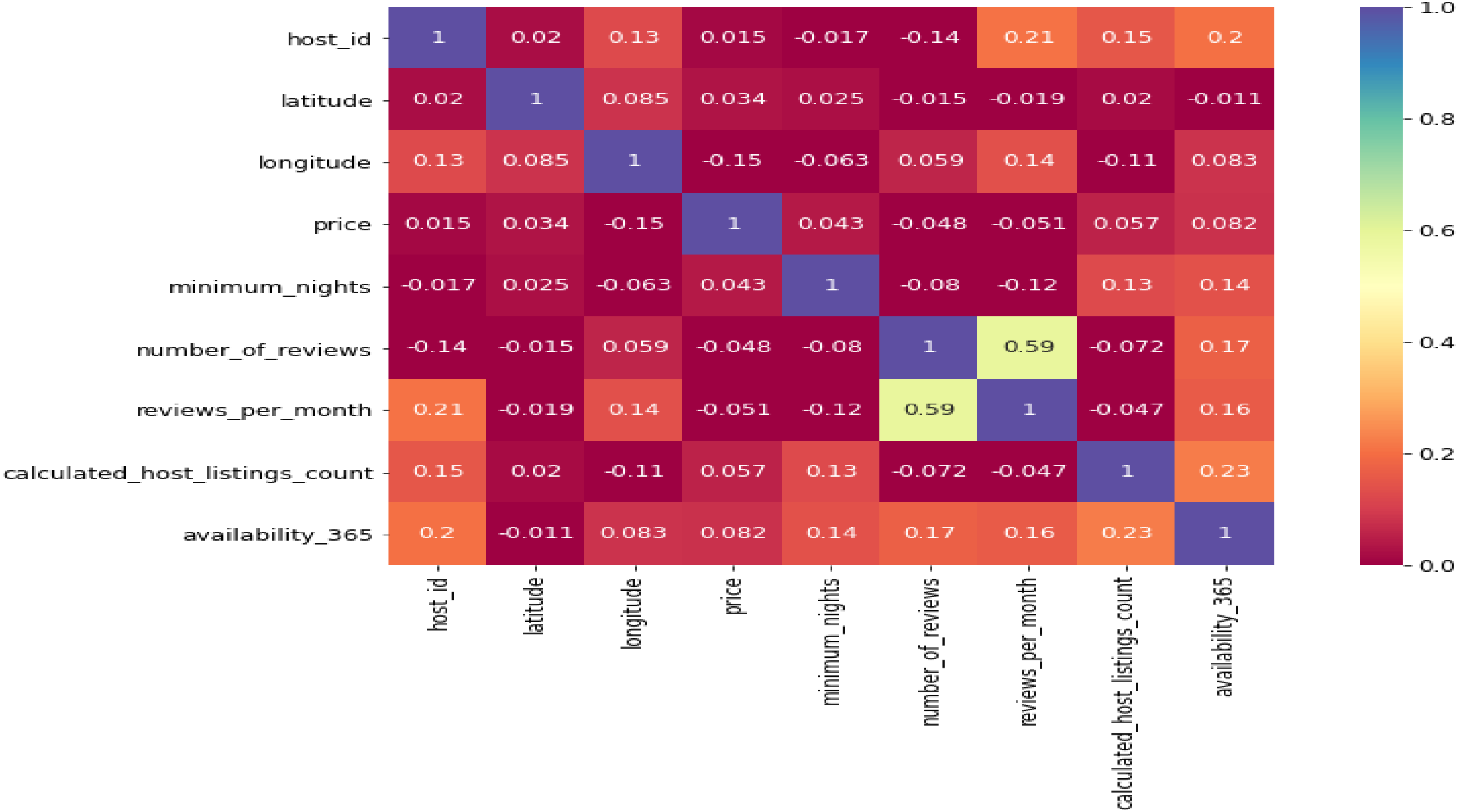
- We got to know that columns such as 'id’, 'last_review' are of no use for this particular analysis. To elaborate, "last_review" is date; if there were no reviews for the listing - date simply will not exist. In our case, this column is irrelevant and insignificant therefore appending those values is not needed.
- For "review_per_month" column we can simply append it with 0.0 for missing values; we can see that in "number_of_review" that column will have a 0, therefore following this logic with 0 total reviews there will be 0.0 rate of reviews per month.
- Also, host_name and name are not that much important in our analysis so we filled those columns with substitute ‘no_name’ and ‘unknown’ respectively.



```
df_airbnb.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
```

Correlation Matrix



	host_id	latitude	longitude	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365
host_id	1	0.02	0.13	0.015	-0.017	-0.14	0.21	0.15	0.2
latitude	0.02	1	0.085	0.034	0.025	-0.015	-0.019	0.02	-0.011
longitude	0.13	0.085	1	-0.15	-0.063	0.059	0.14	-0.11	0.083
price	0.015	0.034	-0.15	1	0.043	-0.048	-0.051	0.057	0.082
minimum_nights	-0.017	0.025	-0.063	0.043	1	-0.08	-0.12	0.13	0.14
number_of_reviews	-0.14	-0.015	0.059	-0.048	-0.08	1	0.59	-0.072	0.17
reviews_per_month	0.21	-0.019	0.14	-0.051	-0.12	0.59	1	-0.047	0.16
calculated_host_listings_count	0.15	0.02	-0.11	0.057	0.13	-0.072	-0.047	1	0.23
availability_365	0.2	-0.011	0.083	0.082	0.14	0.17	0.16	0.23	1

Exploring and Visualizing Data

Single Variable Analysis

In this analysis we consider single variable against single parameter.
We considered following important features to analyze against each other

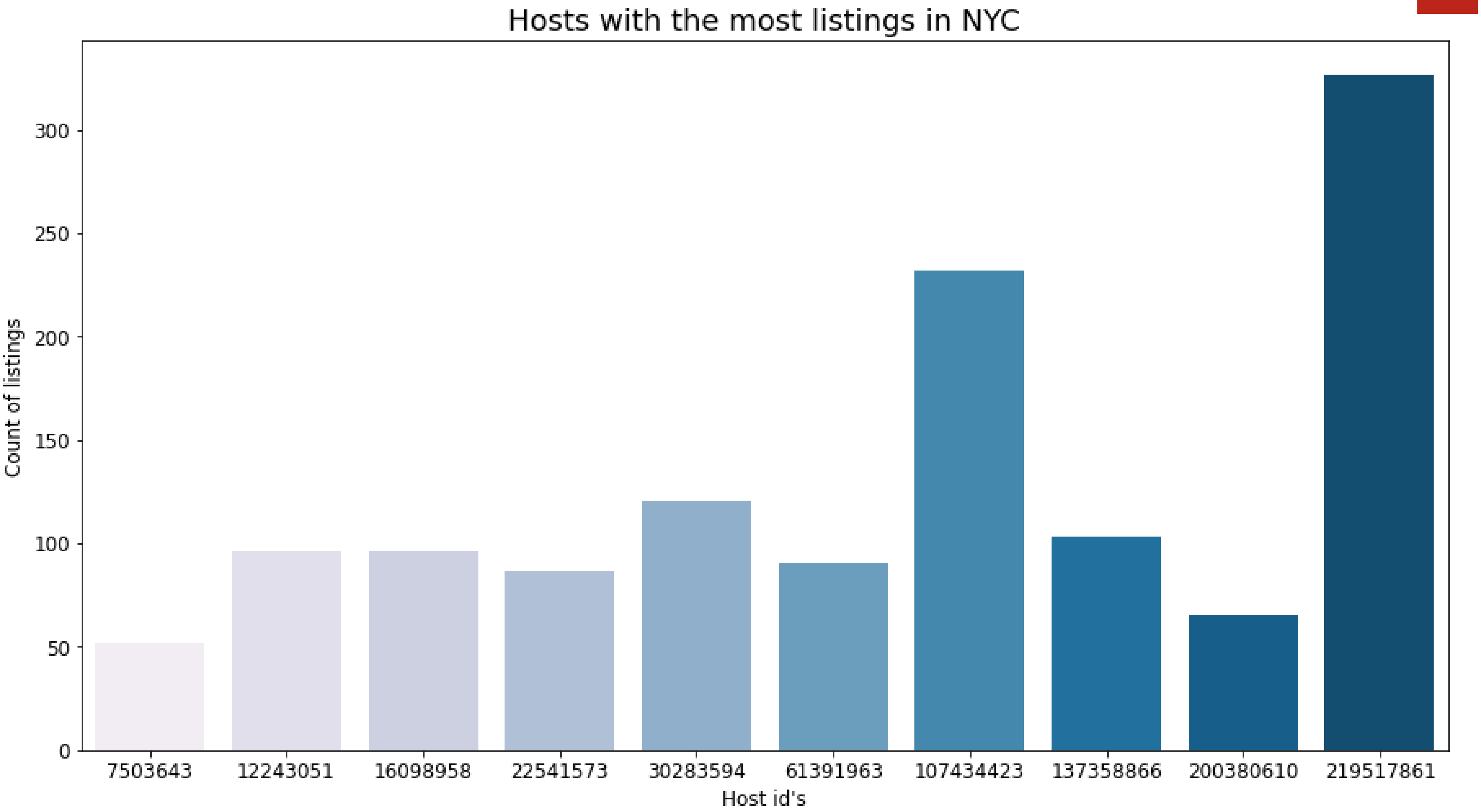
- Top 10 hosts (IDs) have the most listings
- The neighbourhood group vs no of listings in entire NYC
- Top 10 neighbourhood s in entire NYC on the basis of count of listings
- Top 10 reviewed hosts on the basis of reviews/month
- On an average for how many nights people stayed in each room types

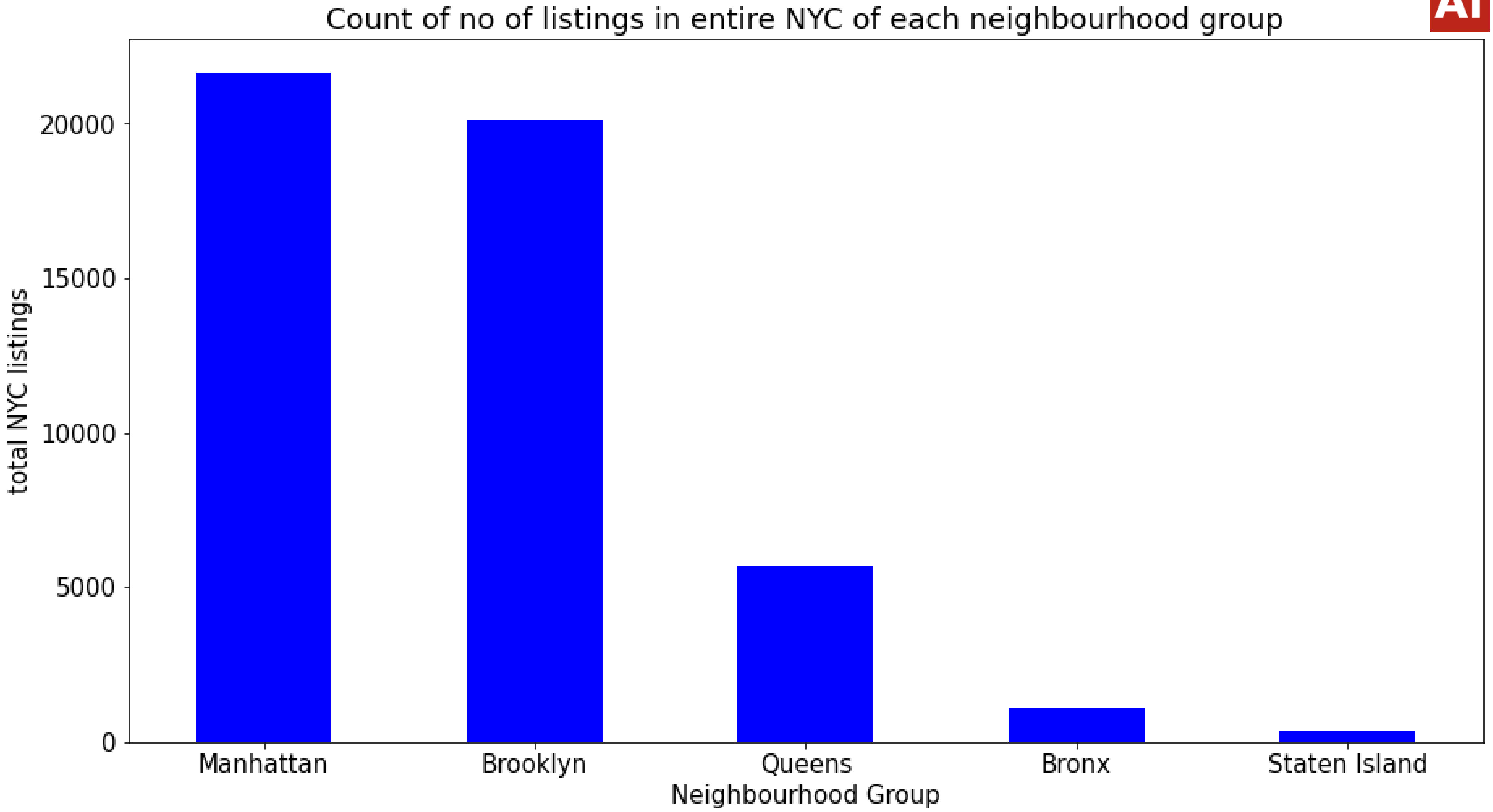
Bi-variable Analysis

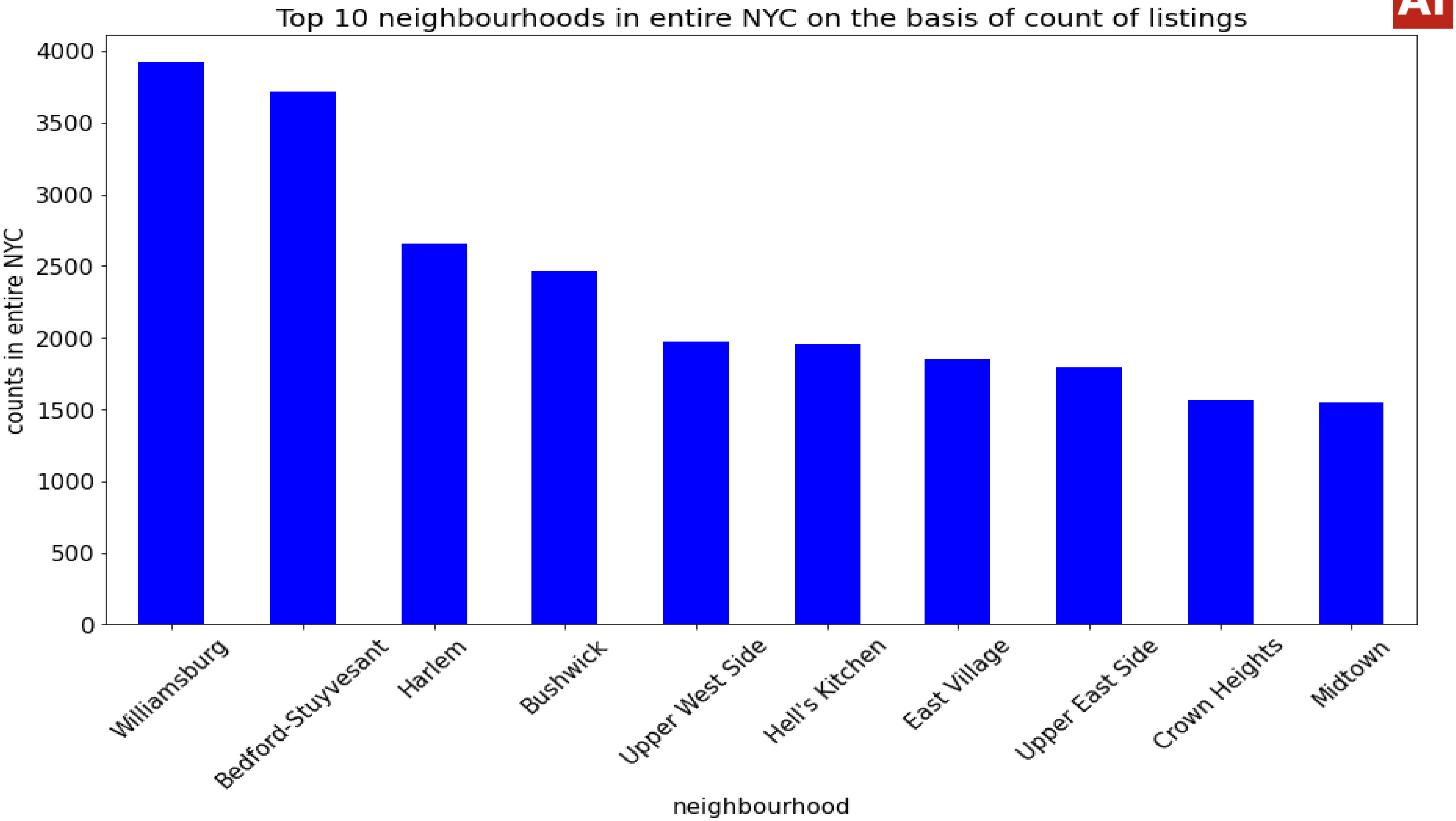
In this analysis we consider two variables against single parameter.
We considered following important features to analyze against each other

- Count of each room types in neighbourhood group entire NYC
- Most Reviewed room types in each neighbourhood Groups
- Room types and their relation with availability and also with different neighbourhood groups

Single Variable Analysis…

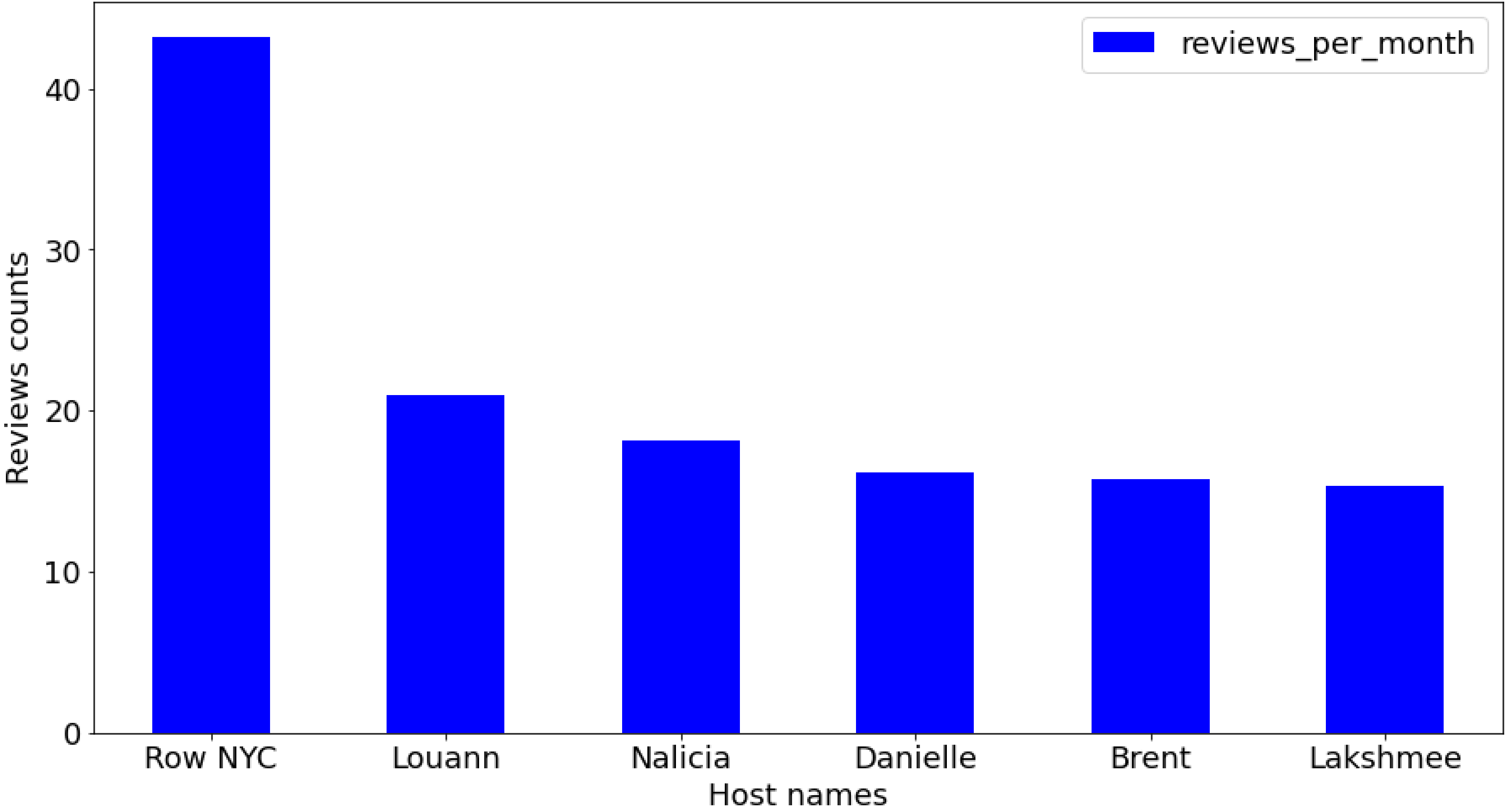


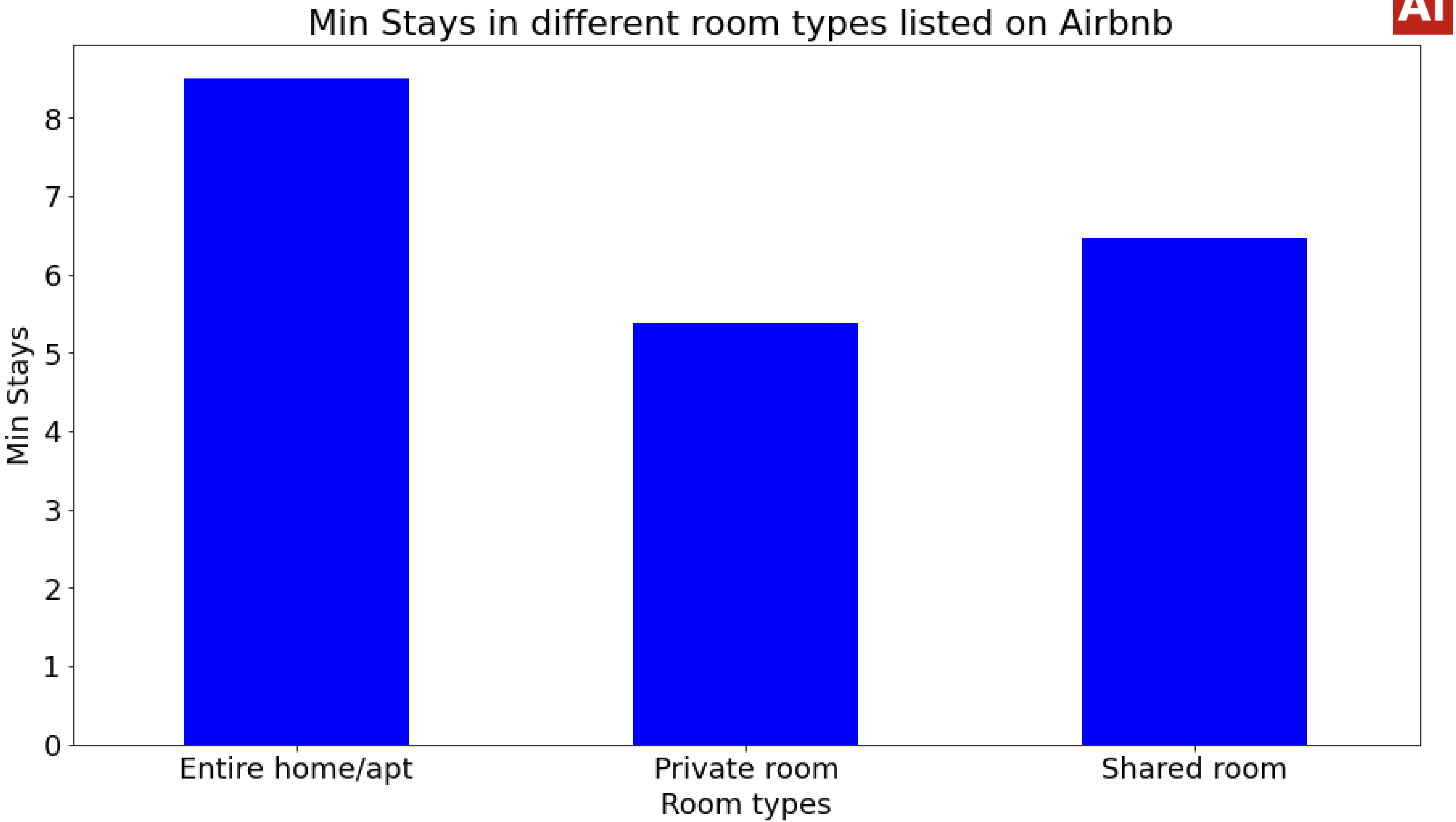

Count of no of listings in entire NYC of each neighbourhood group
total NYC listings
0
5000
10000
15000
20000
Manhattan
Brooklyn
Queens
Bronx
Staten Island
Neighbourhood Group



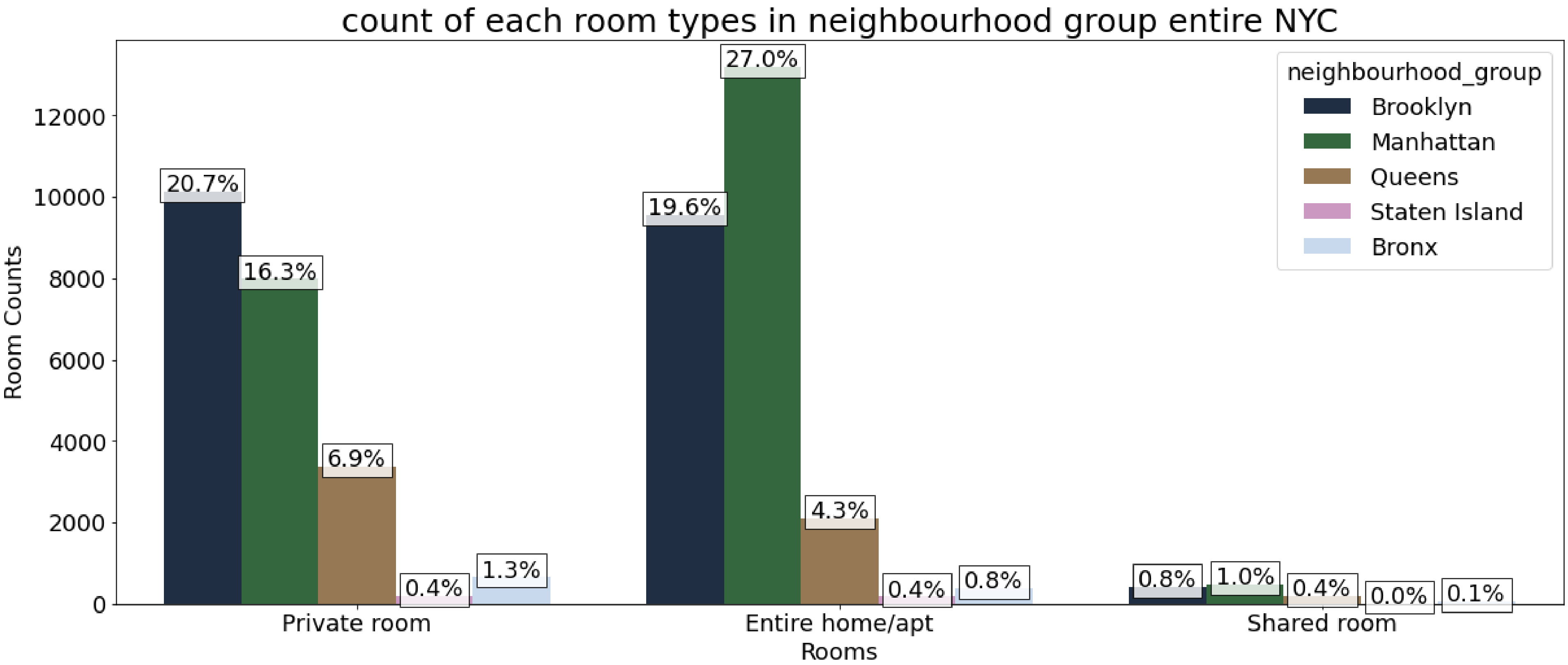
Top 10 neighbourhoods in entire NYC on the basis of count of listings
counts in entire NYC
4000
3500
3000
2500
2000
1500
1000
500
0
Williamsburg
Bedford-Stuyvesant
Harlem
Bushwick
Upper West Side
Hell's Kitchen
East Village
Upper East Side
Crown Heights
Midtown
neighbourhood

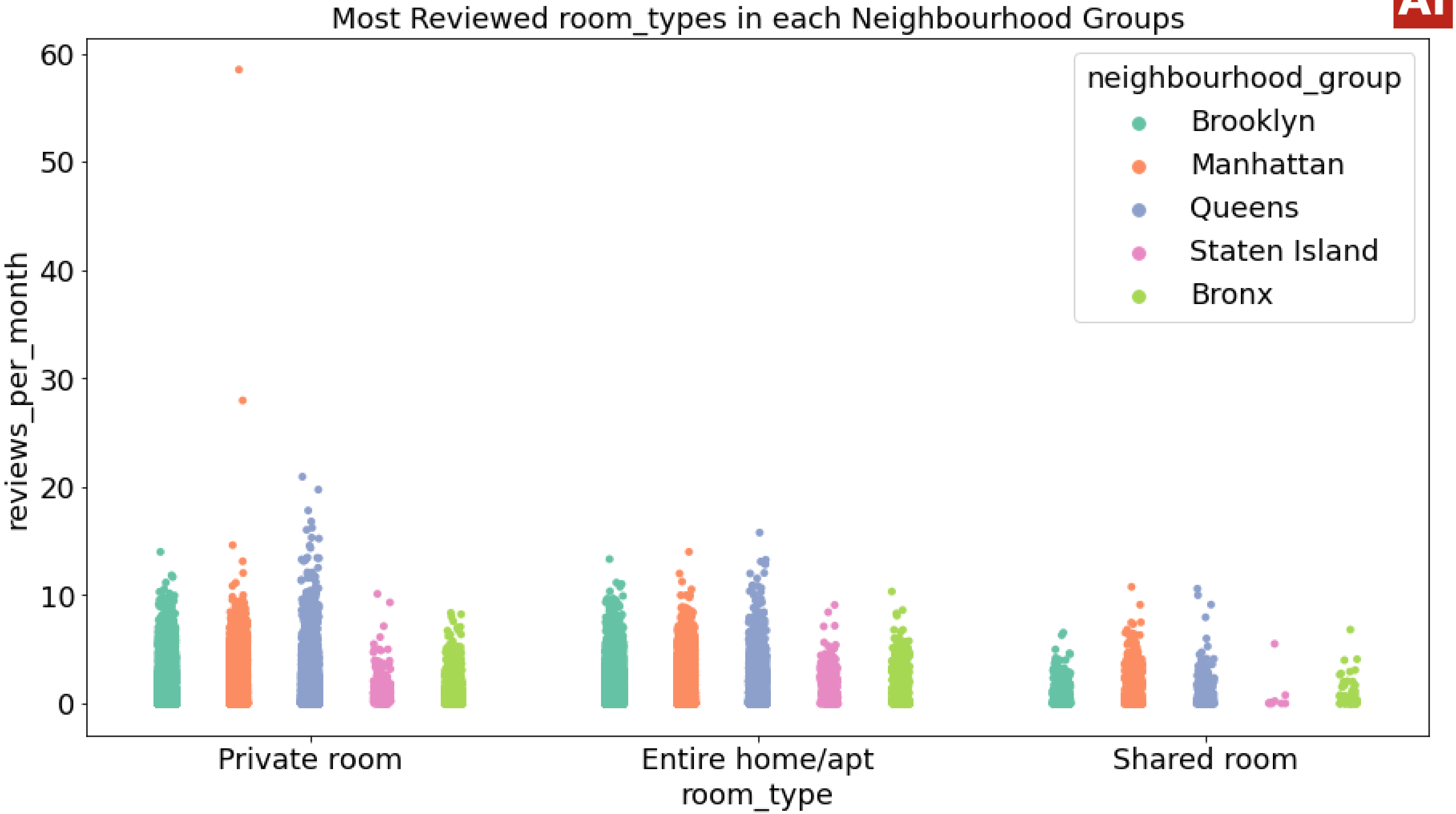
Average Reviews/month received by Top hosts

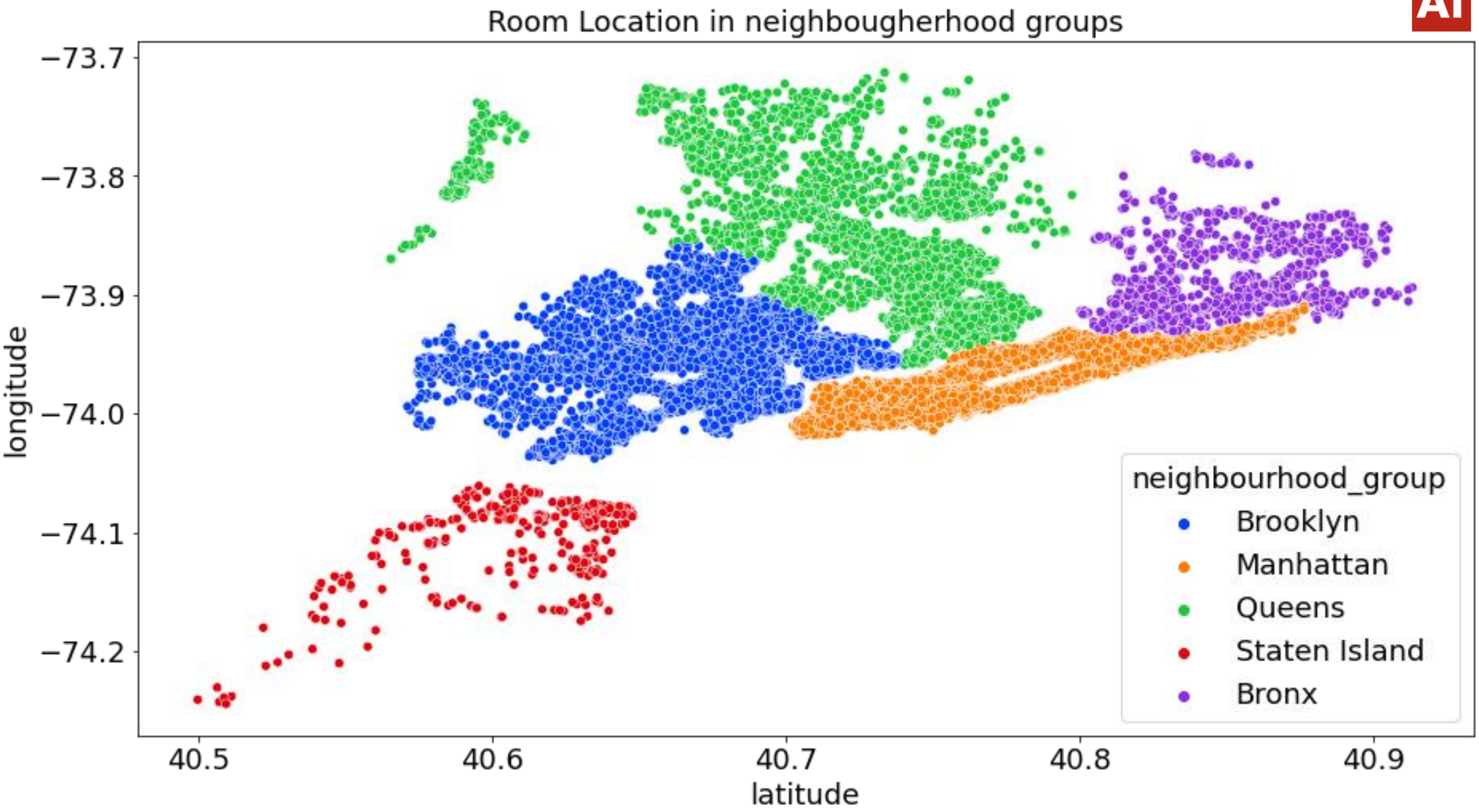


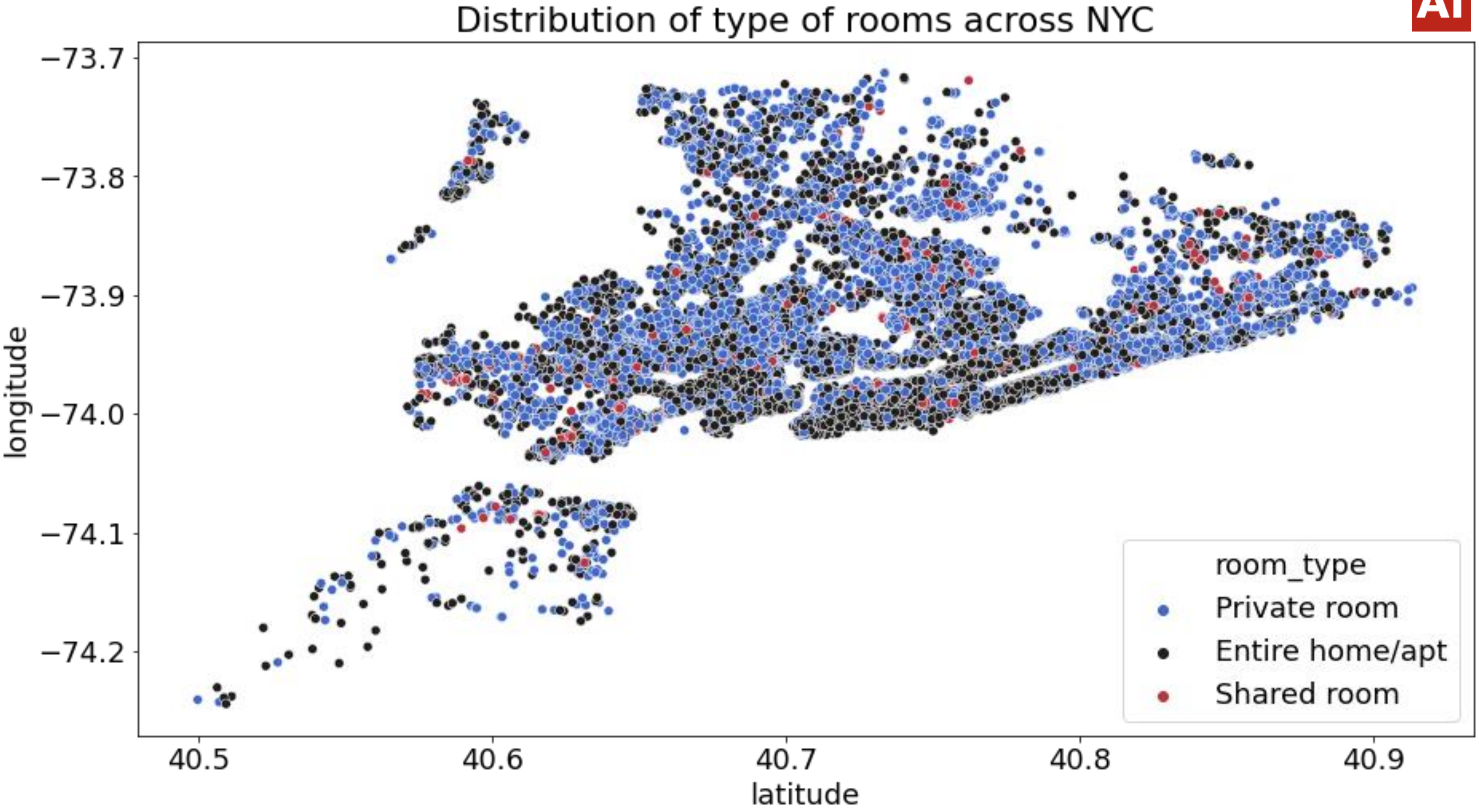

Min Stays in different room types listed on Airbnb
Min Stays
0
1
2
3
4
5
6
7
8
Entire home/apt
Private room
Shared room
Room types

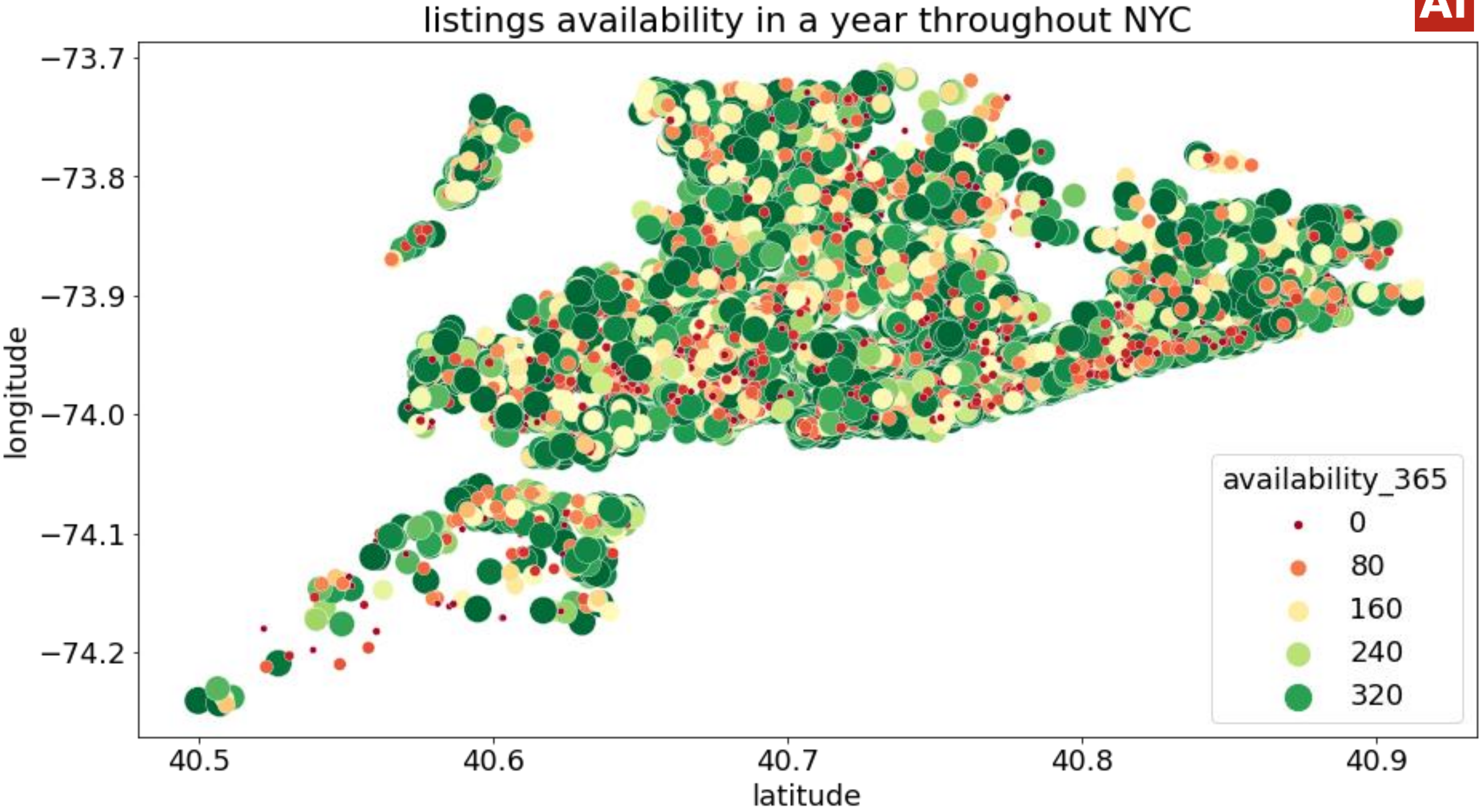
Bi-Variable Analysis...




Most Reviewed room_types in each Neighbourhood Groups
reviews_per_month
0
10
20
30
40
50
60
Private room
Entire home/apt
Shared room
room_type
neighbourhood_group
● Brooklyn
● Manhattan
● Queens
● Staten Island
● Bronx


Room Location in neighbougherhood groups
−73.7
−73.8
−73.9
−74.0
−74.1
−74.2
longitude
40.5
40.6
40.7
40.8
40.9
latitude
neighbourhood_group
● Brooklyn
● Manhattan
● Queens
● Staten Island
● Bronx


Distribution of type of rooms across NYC
−73.7
−73.8
−73.9
−74.0
−74.1
−74.2
longitude
40.5
40.6
40.7
40.8
40.9
latitude
room_type
● Private room
● Entire home/apt
● Shared room

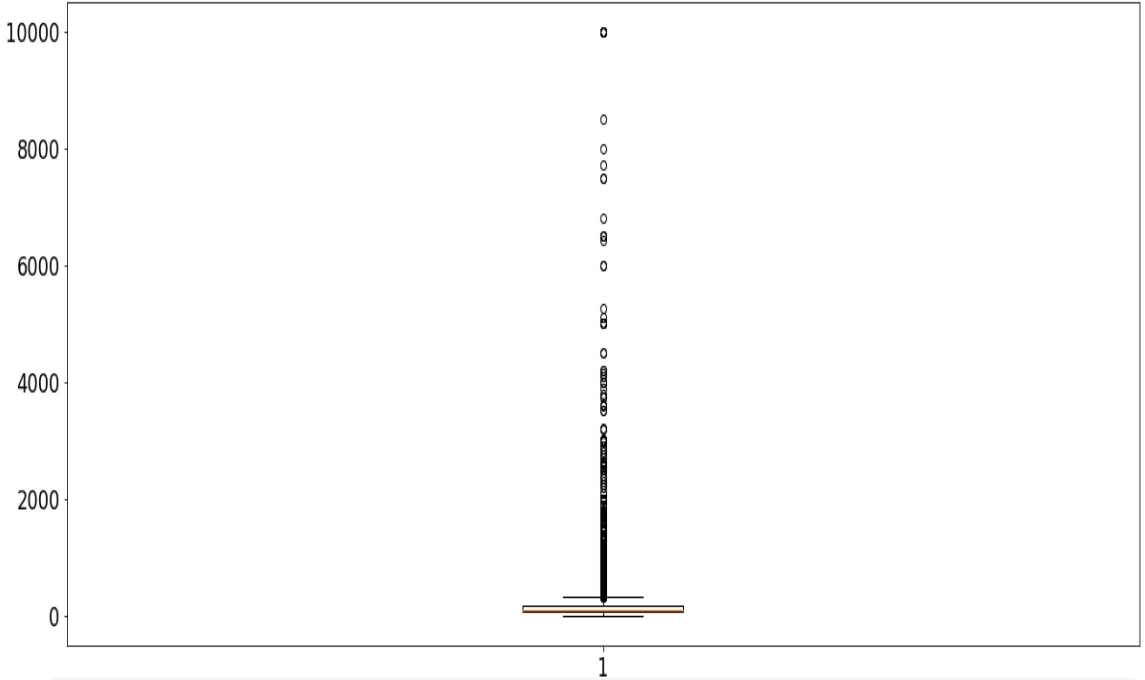

listings availability in a year throughout NYC
−73.7
−73.8
−73.9
−74.0
−74.1
−74.2
longitude
40.5
40.6
40.7
40.8
40.9
latitude
availability_365
0
80
160
240
320

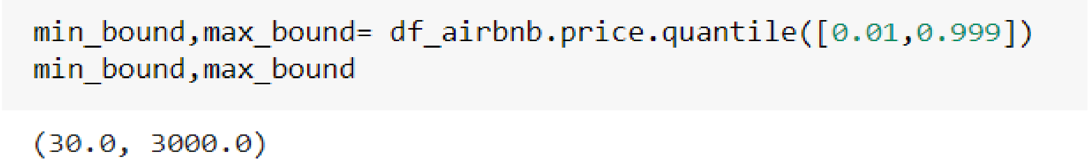
‘Price Feature’

- We used visualization method i.e. **boxplot** to find outliers in price column.

- As, boxplot shows that feature column (i.e. Price) has many outliers. So, we have to remove those outliers for better result.

- So, we use **quantile approach** to remove these outliers. In this technique, the outlier is capped at a certain value above the 90th percentile value or floored at a factor below the 10th percentile value.

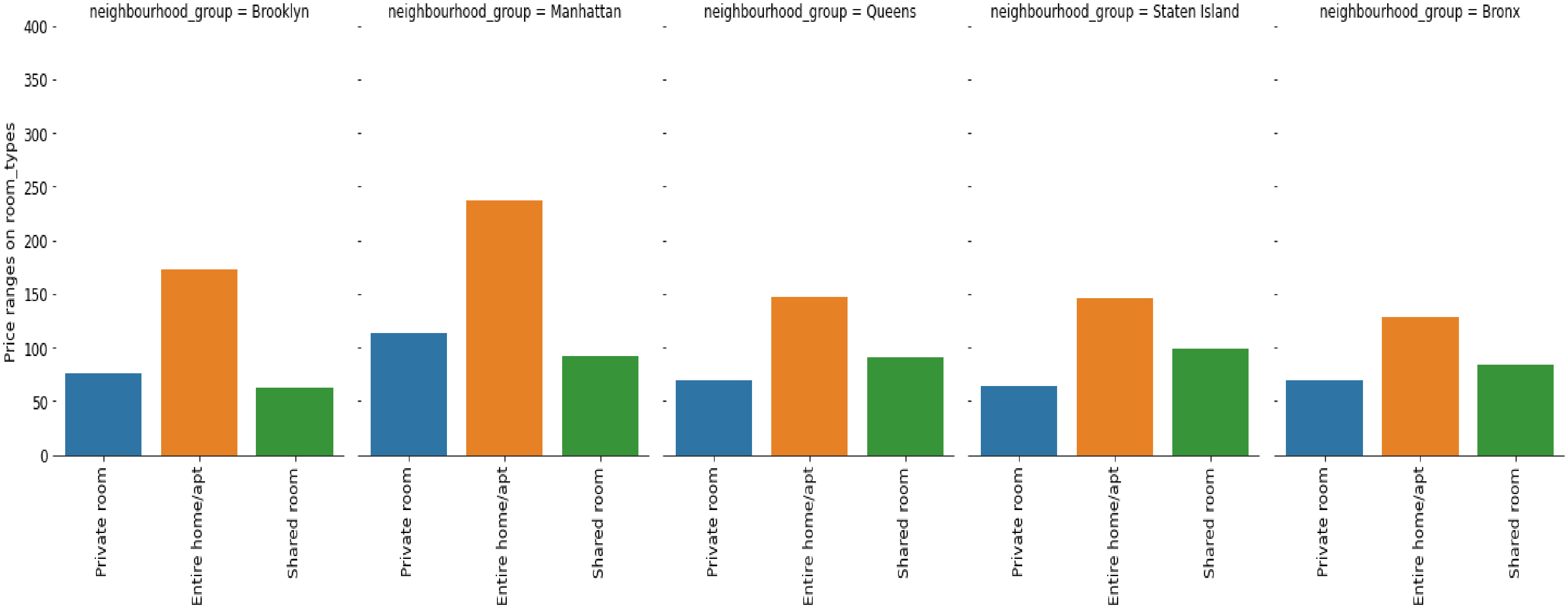




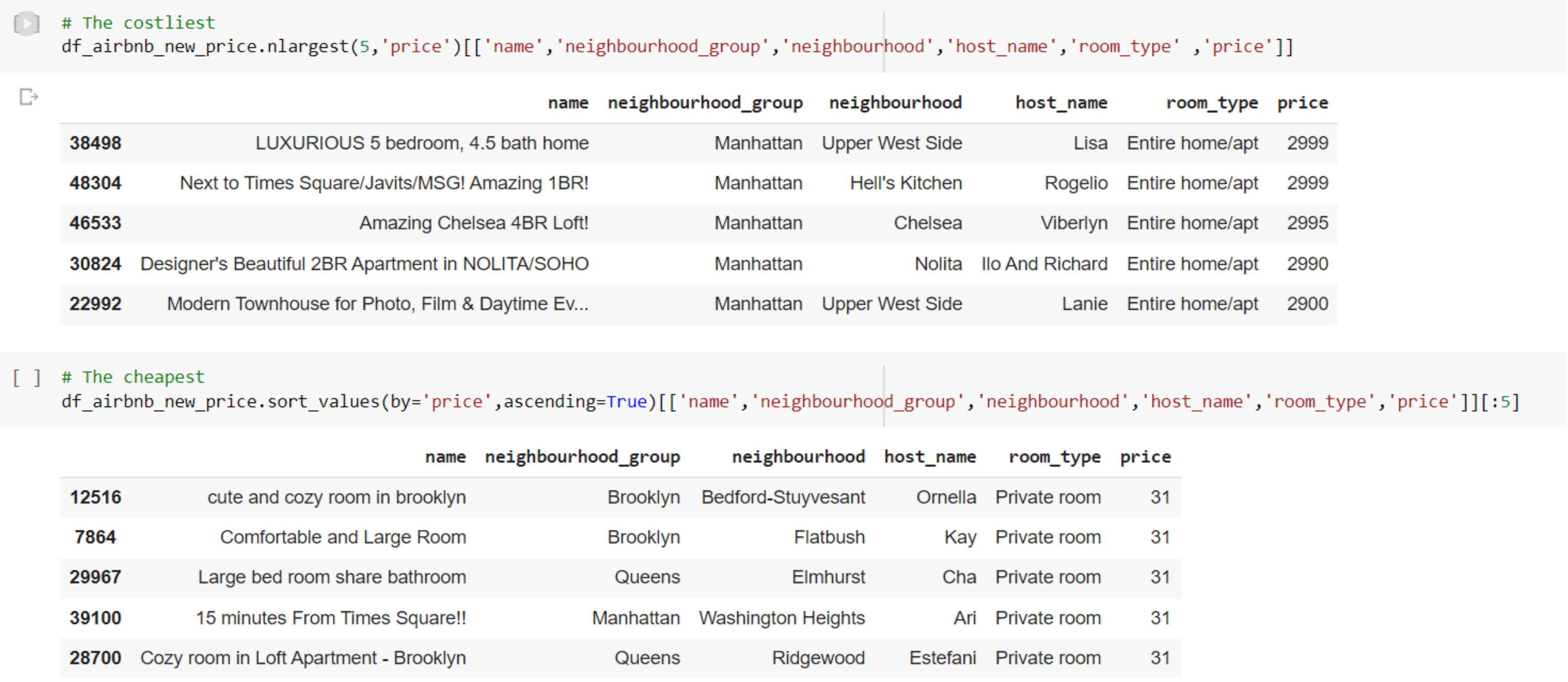
```
min_bound,max_bound= df_airbnb.price.quantile([0.01,0.999])
min_bound,max_bound
```

```
(30.0, 3000.0)
```

Room types Vs price in different neighbourhood groups



The costliest and cheapest listings & their respective hosts in entire NYC



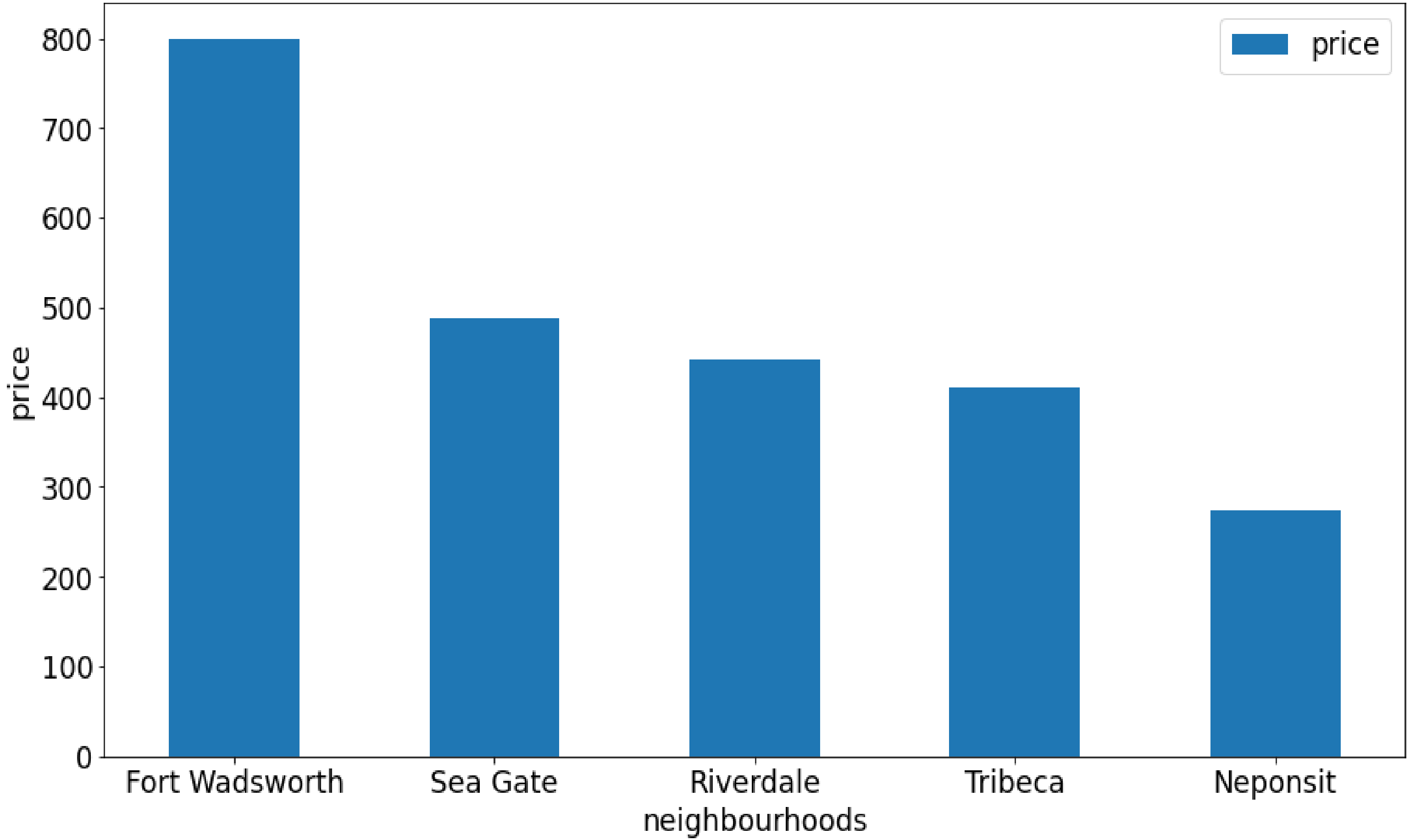
```
# The costliest
df_airbnb_new_price.nlargest(5,'price')[['name','neighbourhood_group','neighbourhood','host_name','room_type' ,'price']]
```

	name	neighbourhood_group	neighbourhood	host_name	room_type	price
38498	LUXURIOUS 5 bedroom, 4.5 bath home	Manhattan	Upper West Side	Lisa	Entire home/apt	2999
48304	Next to Times Square/Javits/MSG! Amazing 1BR!	Manhattan	Hell's Kitchen	Rogelio	Entire home/apt	2999
46533	Amazing Chelsea 4BR Loft!	Manhattan	Chelsea	Viberlyn	Entire home/apt	2995
30824	Designer's Beautiful 2BR Apartment in NOLITA/SOHO	Manhattan	Nolita	Ilo And Richard	Entire home/apt	2990
22992	Modern Townhouse for Photo, Film & Daytime Ev...	Manhattan	Upper West Side	Lanie	Entire home/apt	2900

```
[ ] # The cheapest
    df_airbnb_new_price.sort_values(by='price',ascending=True)[['name','neighbourhood_group','neighbourhood','host_name','room_type','price']][:5]
```

	name	neighbourhood_group	neighbourhood	host_name	room_type	price
12516	cute and cozy room in brooklyn	Brooklyn	Bedford-Stuyvesant	Ornella	Private room	31
7864	Comfortable and Large Room	Brooklyn	Flatbush	Kay	Private room	31
29967	Large bed room share bathroom	Queens	Elmhurst	Cha	Private room	31
39100	15 minutes From Times Square!!	Manhattan	Washington Heights	Ari	Private room	31
28700	Cozy room in Loft Apartment - Brooklyn	Queens	Ridgewood	Estefani	Private room	31

Top neighbourhoods in each neighbourhood groups with respect to average price/day of Airbnb listings



Conclusion

- If a person trying to book a listing for stay/rent he/she will look into these following factors while booking: neighbourhood group, neighbourhood, room type, price, number of reviews and availability throughout the year.

- The neighbourhood group 'Manhattan' has highest number of listings in entire NYC. Also top 5 costliest listings are present in it.

- People mostly prefer living in an entire home/apt on an average of more than 8 nights followed by guests who stayed in shared room where average stay is 6-7 nights.

- 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.

- 95% of the listings on Airbnb are either Private room or Entire home/apt. Very few guests had opted for shared rooms on Airbnb. Also, guests mostly prefer private or entire home/apt room types when they are looking for a rent on Airbnb.

- Bronx & Staten Island has listings which are mostly available throughout the year, this might be the case as they are not much costlier as compared to other neighbourhood groups such as in Manhattan, Brooklyn & Queens.