Capstone Project - 1

Airbnb Bookings Analysis

Team Member

Sanjay Yadav

Introduction

- 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.

Objectives of the project:

- Explore and clean the data for any missing values and duplicate data.

- Perform analysis on the price distribution in various major cities of New York.

- Check if the price has a relation with the minimum stays as well as property type along with location.

- Perform geographical plotting of the price and location, and to see of if I can find any interesting aspects in New York.

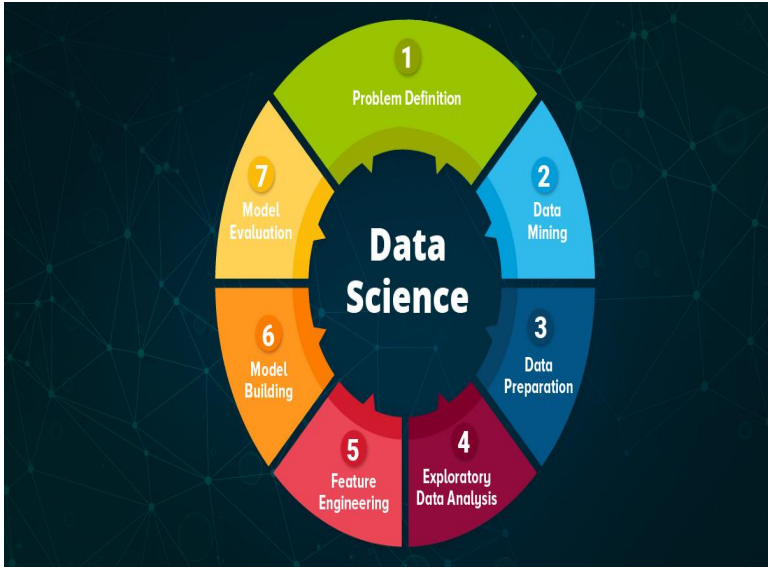
- Analyze the behavior of hosts with multiple listings.

- Understand the relation between reviews count and neighborhood.

- Create appropriate visualizations and build a summary of findings from the data.

About :

Exploratory data analysis is an approach of analyzing data sets to summarize their main characteristics, often using statistics and other data visualization methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task and handling missing values and making transformations of variables as needed.



Data Summary

Data consists of the Airbnb listings which includes of 16 columns :

- id – represents the id of the listing
- name – It is the name of the listing
- host_id – Host id is a unique Id associated to each host in Airbnb
- host_name – The host's name
- neighbourhood_group – The city in which the listing is present
- neighbourhood – Location of the listing
- latitude – Latitude value of the listing
- longitude – Longitude value of the listing

- room_type – Category of the type of listing like entire house, private room etc.
- price – Cost per night
- minimum_nights – Minimum number of nights that the listing should be booked for
- number_of_reviews – Total reviews received by the listing
- last_review – Date when the property received its last review
- reviews_per_month – The ratio of reviews received, and the time property is listed
- calculated_host_listings_count – The count of listings a host has on Airbnb
- availability_365 – The number of days a listing is available in a year
- Records count or the row count of the data is **49,000**. Size of the data: 6.8 MB

Importing Libraries

We will start by importing the libraries we will require for performing EDA. These include NumPy, Pandas, Matplotlib, Seaborn and Plotly.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
```

Reading Data

Reading the data from a CSV file into a Pandas DataFrame.

```
<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
```

Removing Missing Values

We can conclude the no. of missing values in following columns:
name - 16
host_name - 21
last_review - 10052
reviews_per_month - 10052

Steps:
Removing records with missing values and dropping column last_review.

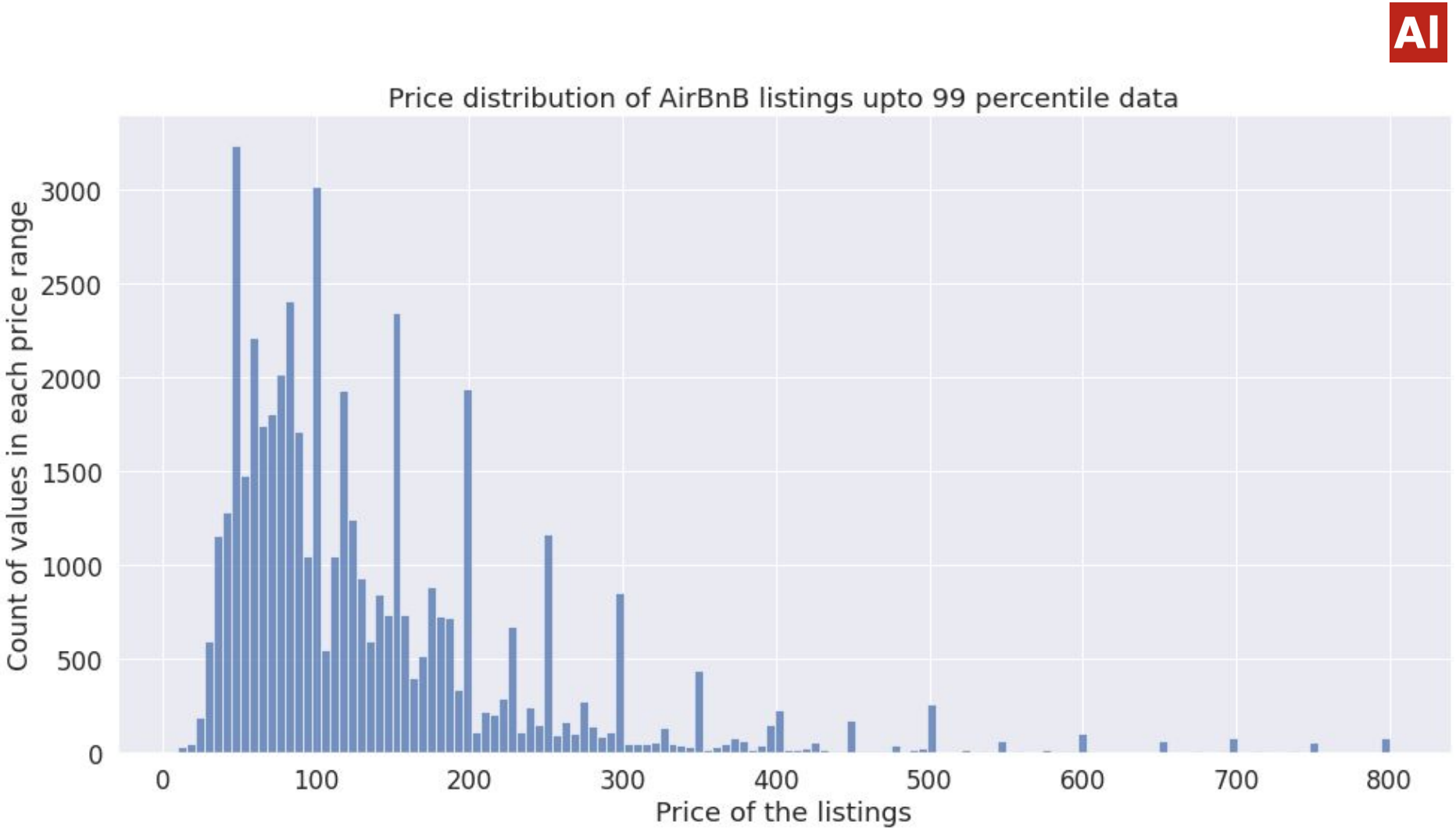
```
df.isnull().sum()
```

```
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
```

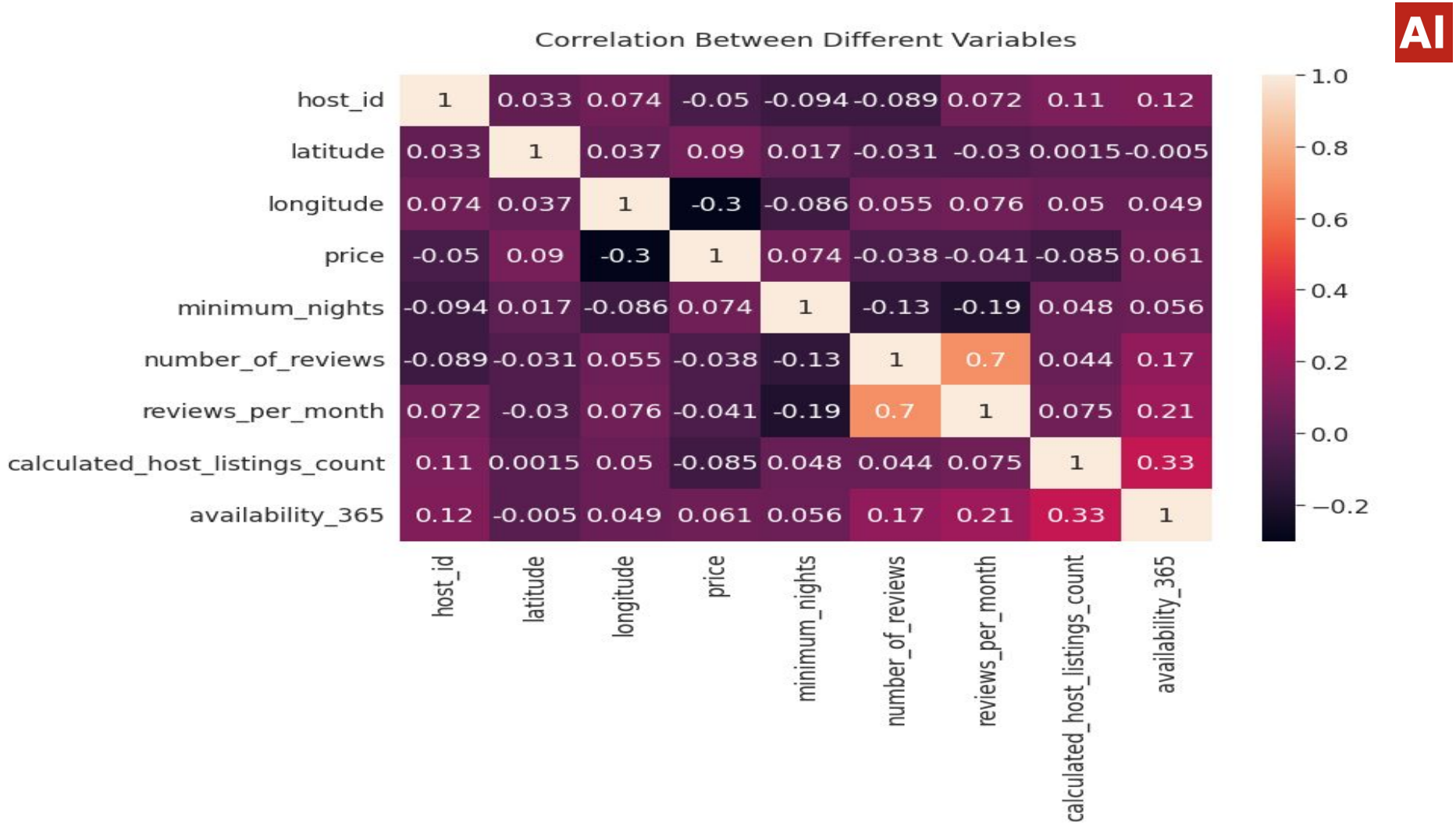
Outliers in Price

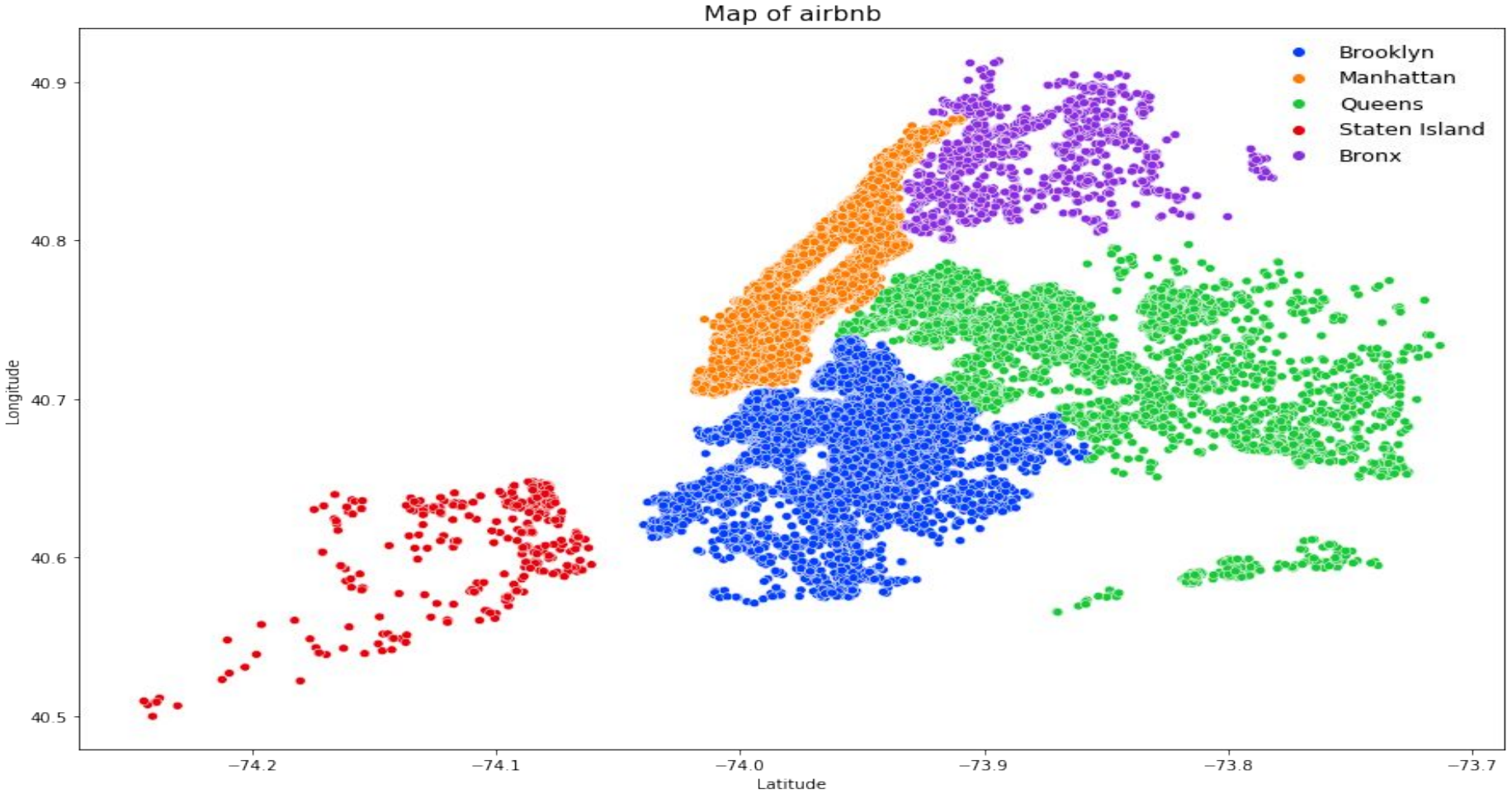
Steps :

- Removing records where listing price is 0.
- Since price column contain values as listed by hosts so we can see high variability depending upon the location and other factors.
- But for our analysis we are going to be working with 99th percentile price since maximum listings fall in that region.
- 99th percentile prices lie under $800.
- We have around 474 records with higher price than $800.
- We are only taking listings with price less than $800 in our dataset.

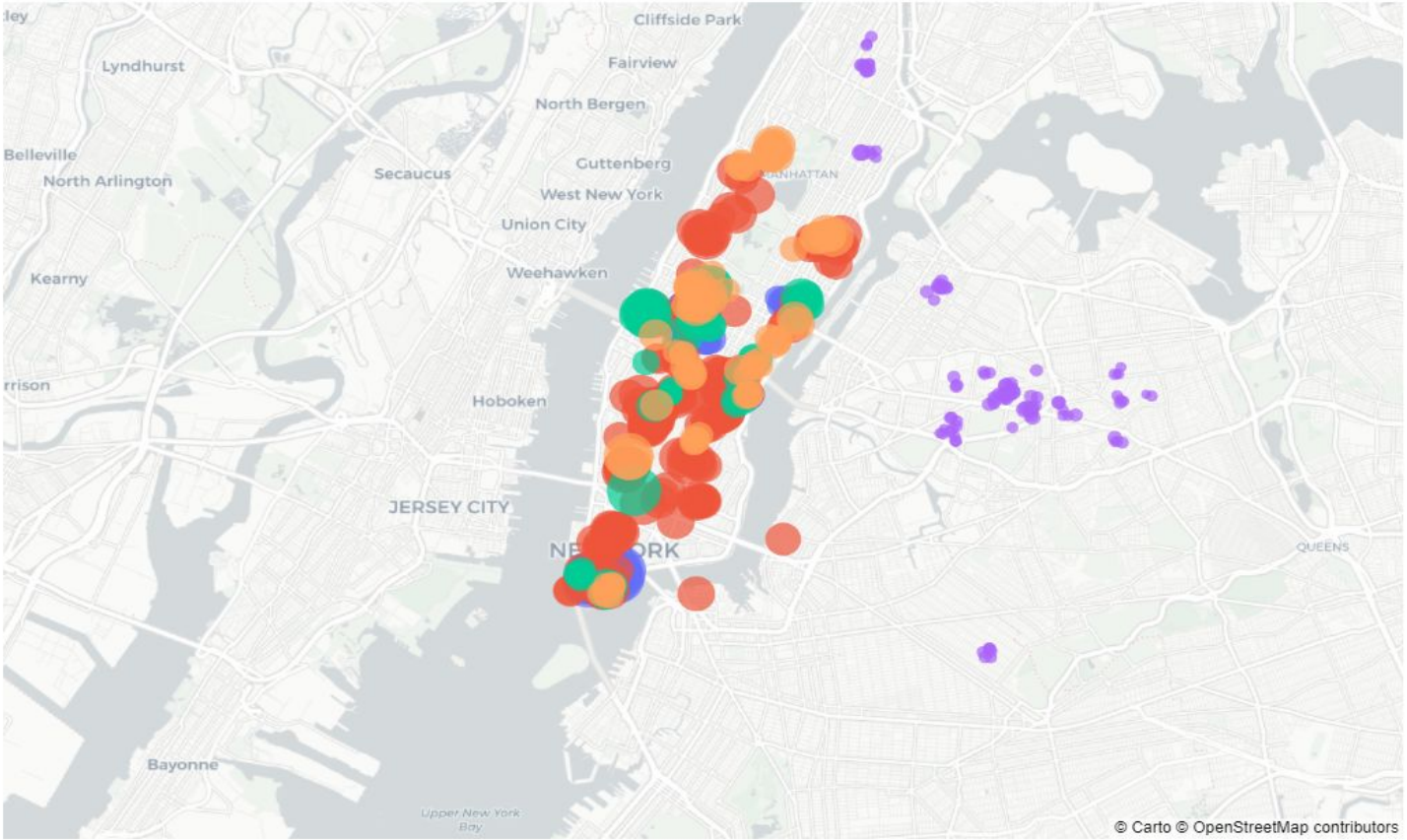

Price distribution of AirBnB listings upto 99 percentile data
Count of values in each price range
0
500
1000
1500
2000
2500
3000
0
100
200
300
400
500
600
700
800
Price of the listings

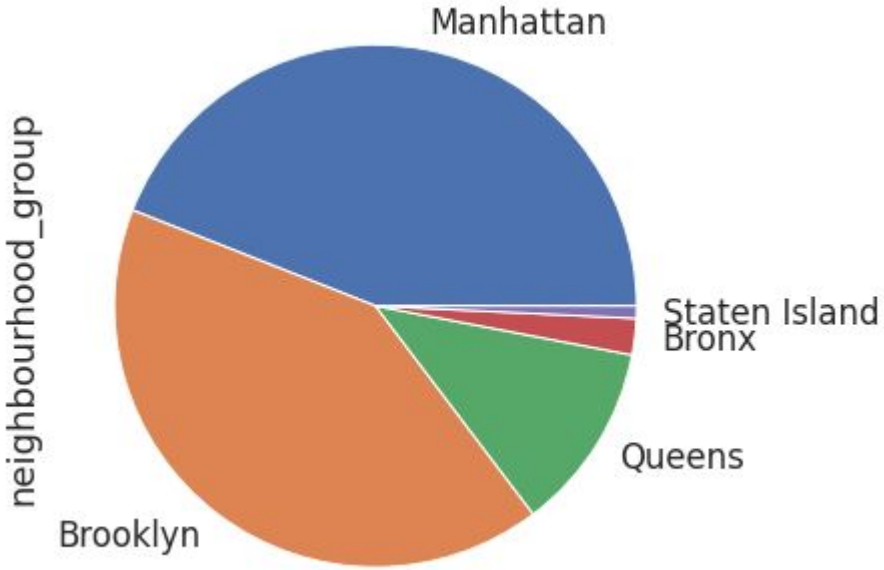

Price distribution of AirBnB listings above 99 percentile data
Count of values in each price range
0
25
50
75
100
125
150
175
2000
4000
6000
8000
10000
Price of the listings per night


Correlation Between Different Variables
host_id 1 0.033 0.074 -0.05 -0.094 -0.089 0.072 0.11 0.12
latitude 0.033 1 0.037 0.09 0.017 -0.031 -0.03 0.0015 -0.005
longitude 0.074 0.037 1 -0.3 -0.086 0.055 0.076 0.05 0.049
price -0.05 0.09 -0.3 1 0.074 -0.038 -0.041 -0.085 0.061
minimum_nights -0.094 0.017 -0.086 0.074 1 -0.13 -0.19 0.048 0.056
number_of_reviews -0.089 -0.031 0.055 -0.038 -0.13 1 0.7 0.044 0.17
reviews_per_month 0.072 -0.03 0.076 -0.041 -0.19 0.7 1 0.075 0.21
calculated_host_listings_count 0.11 0.0015 0.05 -0.085 0.048 0.044 0.075 1 0.33
availability_365 0.12 -0.005 0.049 0.061 0.056 0.17 0.21 0.33 1
host_id
latitude
longitude
price
minimum_nights
number_of_reviews
reviews_per_month
calculated_host_listings_count
availability_365
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

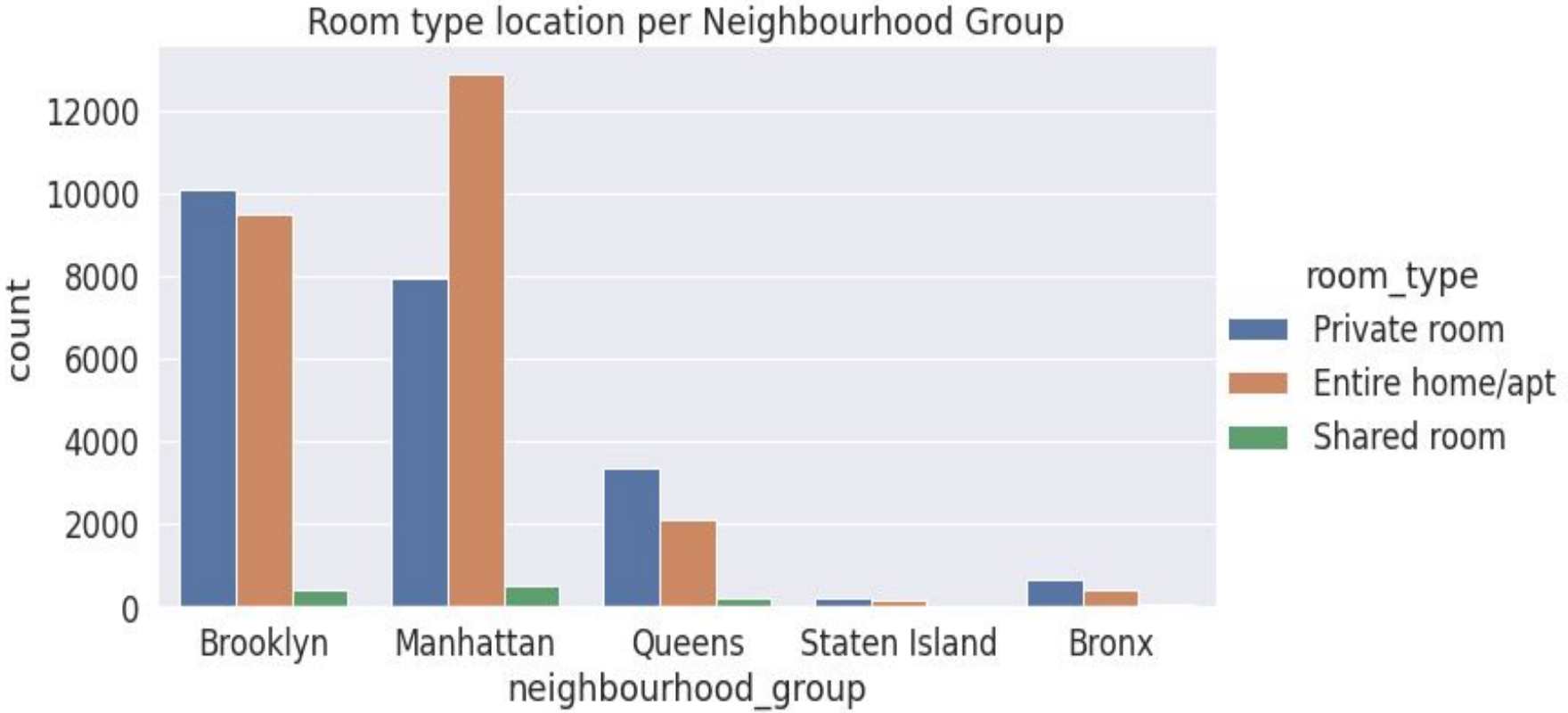

Map of airbnb
Brooklyn
Manhattan
Queens
Staten Island
Bronx
40.9
40.8
40.7
40.6
40.5
Longitude
−74.2
−74.1
−74.0
−73.9
−73.8
−73.7
Latitude

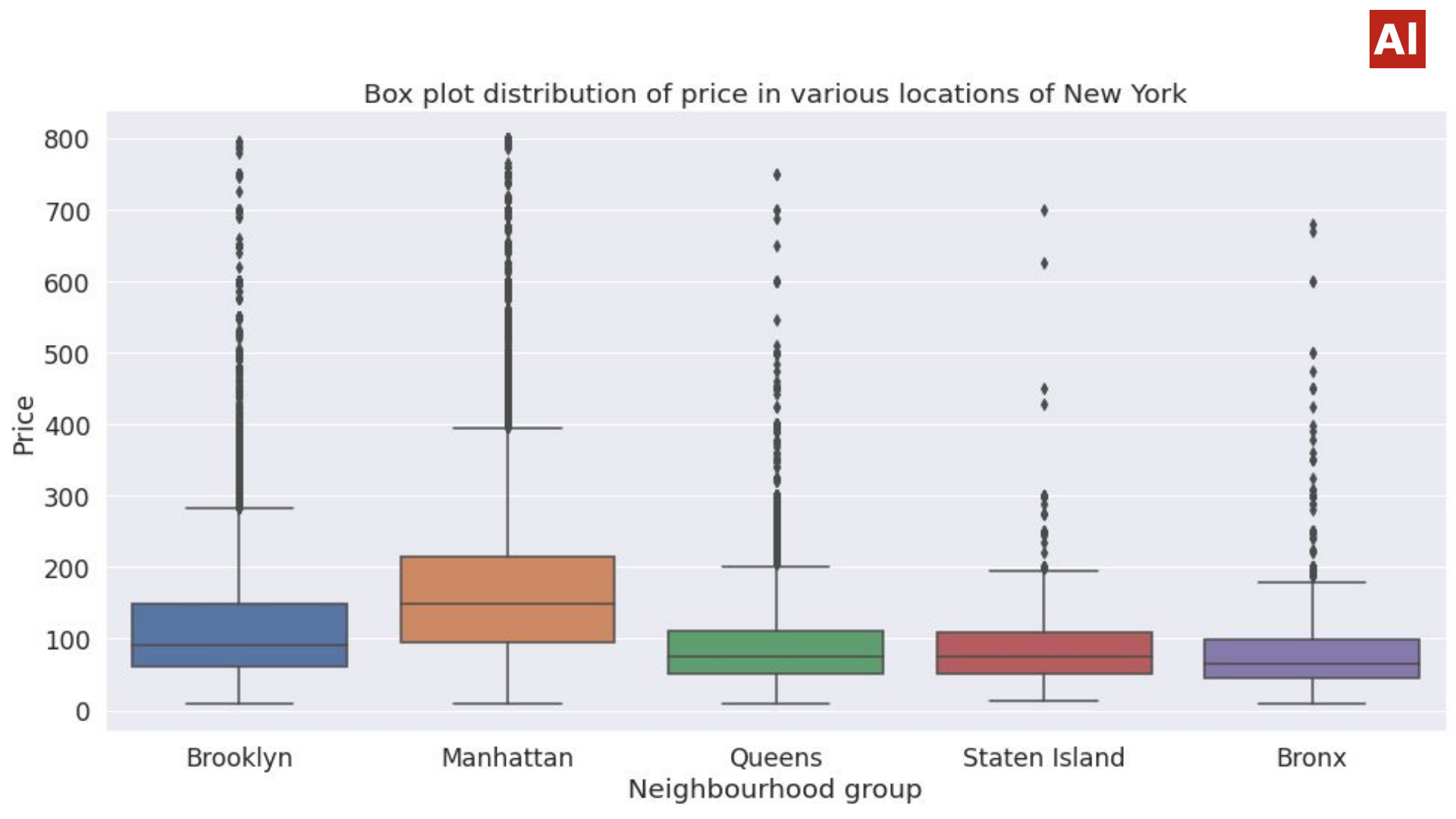
Top 5 hosts and their hosted Locations

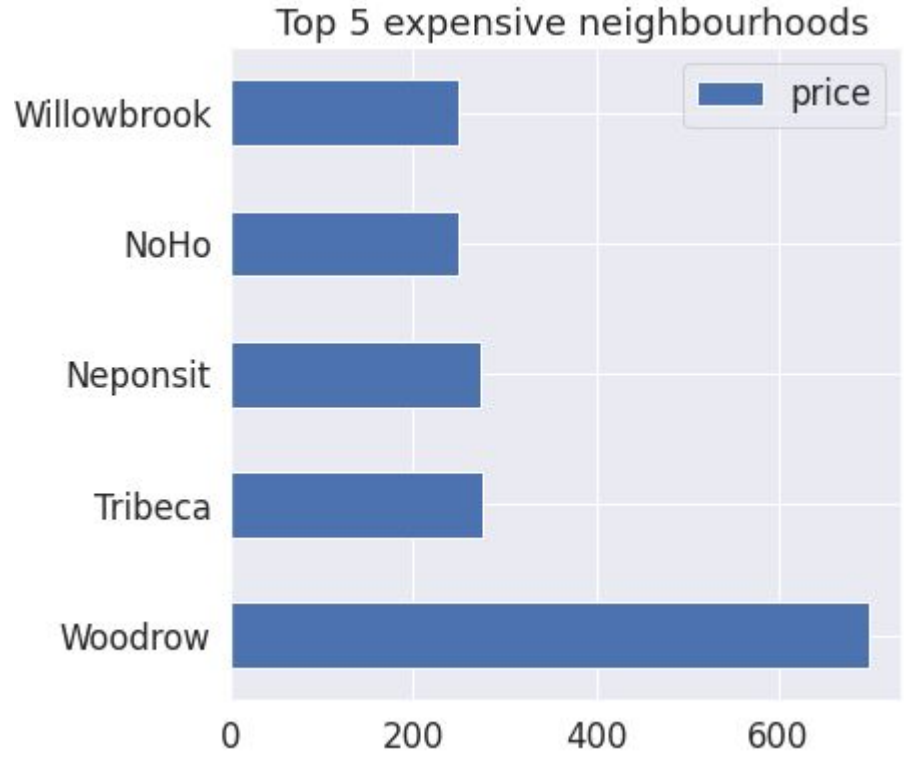


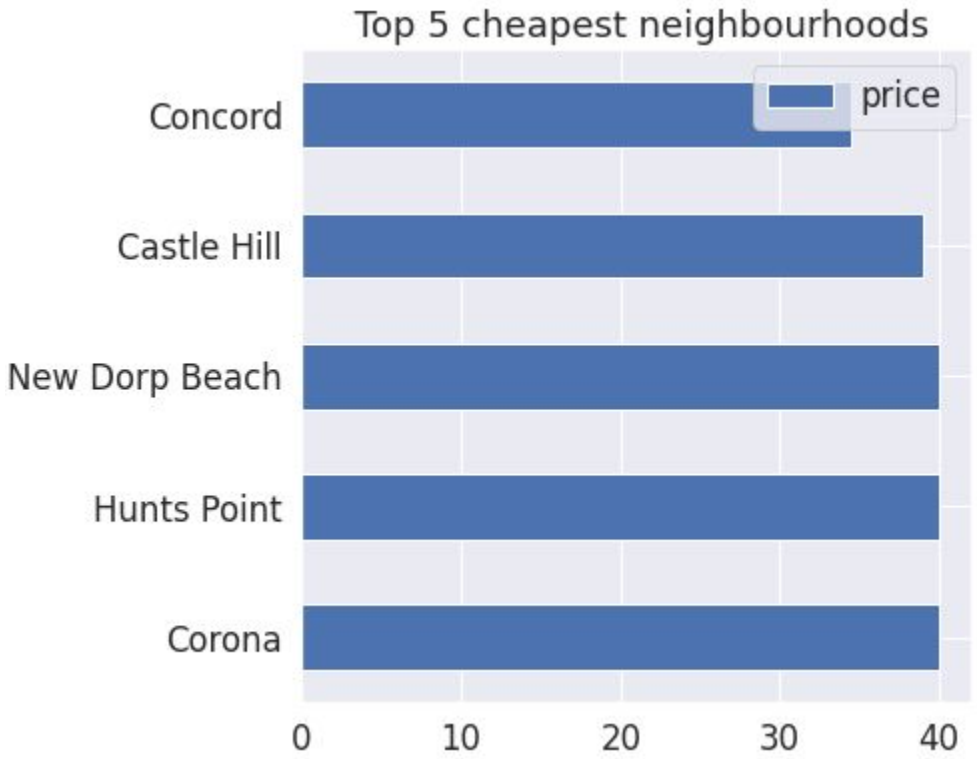


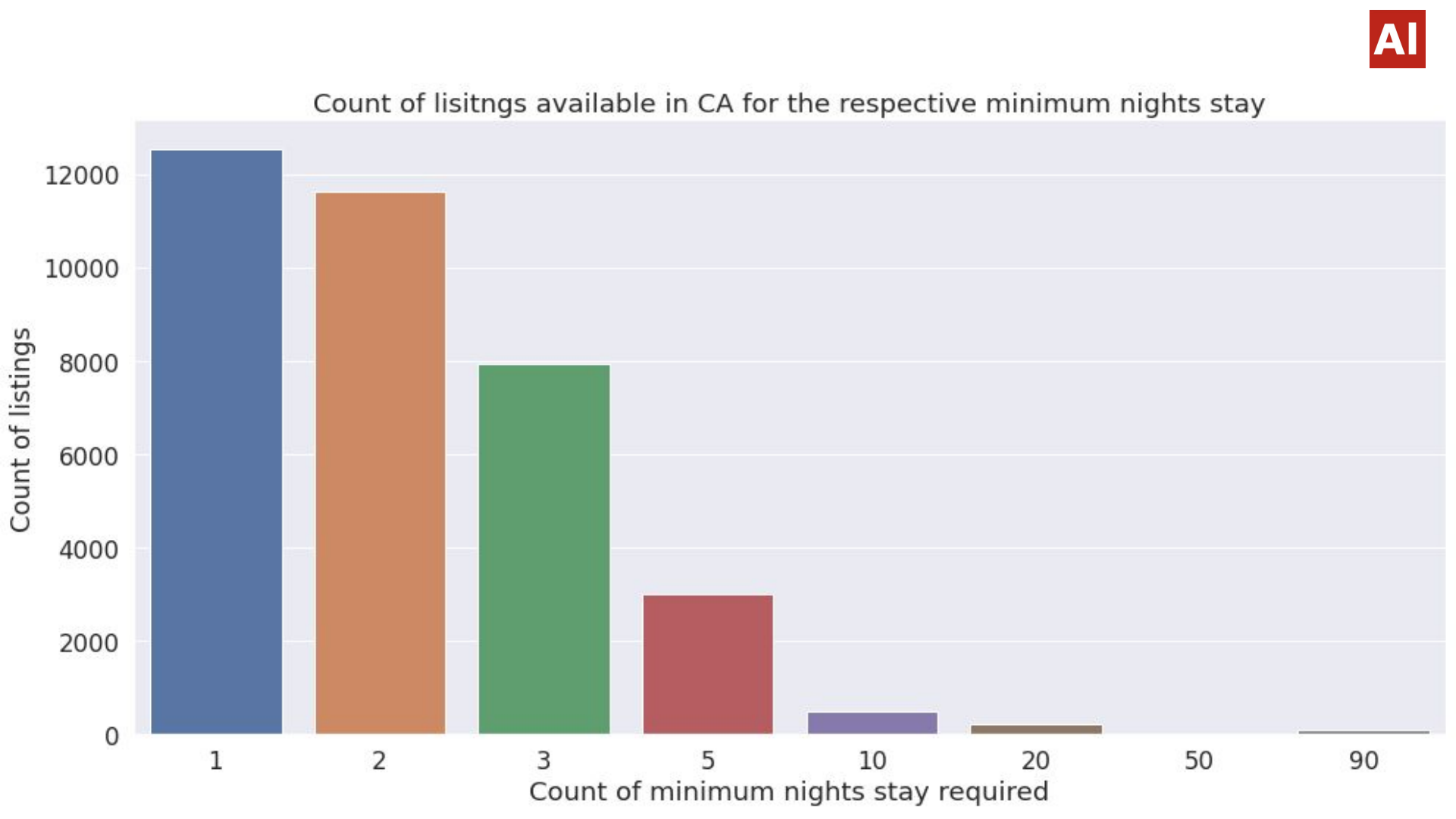
	neighbourhood_group	Percentage
Manhattan	21303	44.02
Brooklyn	19983	41.30
Queens	5648	11.67
Bronx	1086	2.24
Staten Island	369	0.76

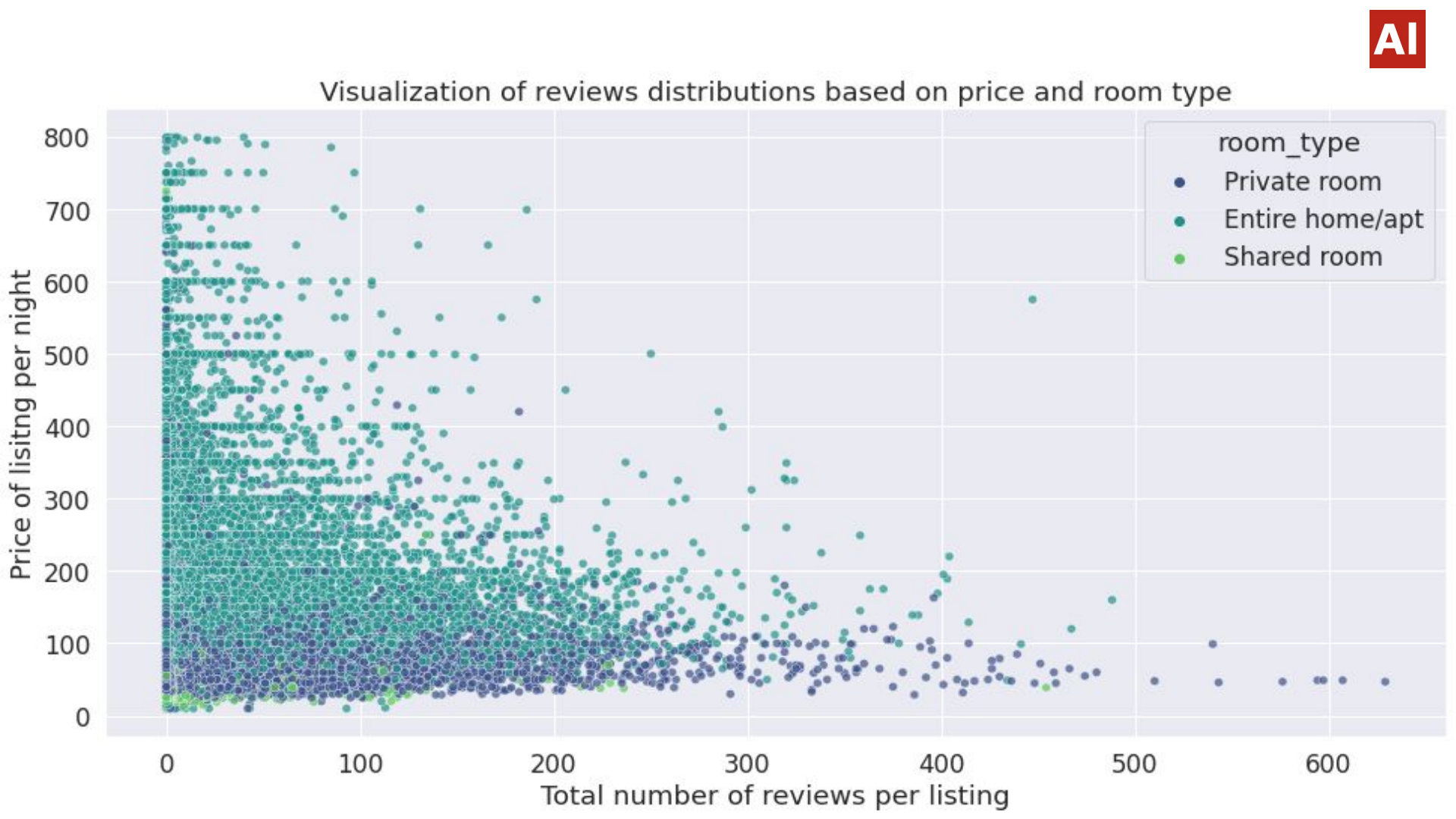

Room type location per Neighbourhood Group
count
12000
10000
8000
6000
4000
2000
0
Brooklyn
Manhattan
Queens
Staten Island
Bronx
neighbourhood_group
room_type
Private room
Entire home/apt
Shared room


Box plot distribution of price in various locations of New York
Price
800
700
600
500
400
300
200
100
0
Brooklyn
Manhattan
Queens
Staten Island
Bronx
Neighbourhood group


Top 5 expensive neighbourhoods
price
Willowbrook
NoHo
Neponsit
Tribeca
Woodrow
0
200
400
600


Top 5 cheapest neighbourhoods
price
Concord
Castle Hill
New Dorp Beach
Hunts Point
Corona
0
10
20
30
40


Count of lisitngs available in CA for the respective minimum nights stay
Count of listings
12000
10000
8000
6000
4000
2000
0
1
2
3
5
10
20
50
90
Count of minimum nights stay required

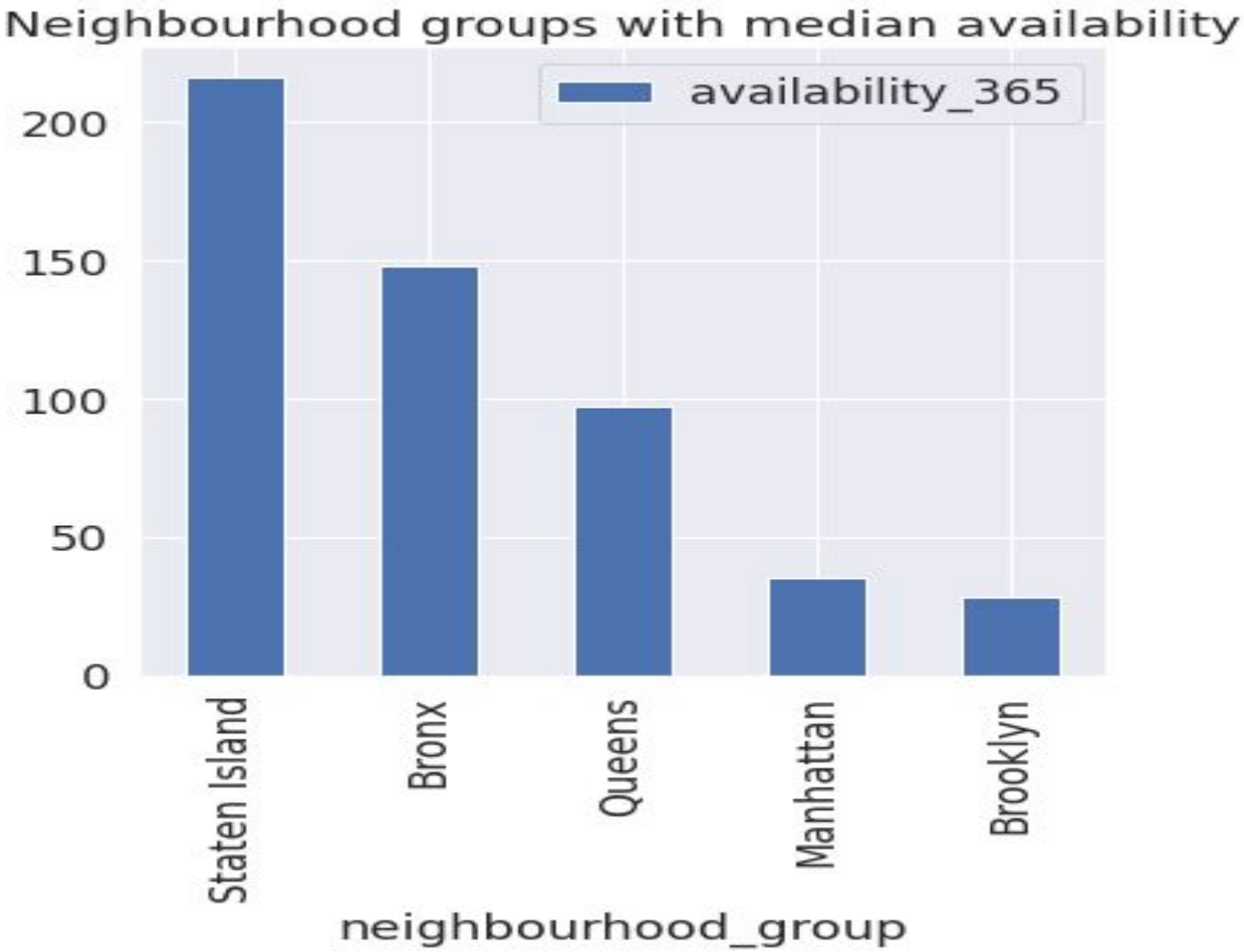

Visualization of reviews distributions based on price and room type
room_type
Private room
Entire home/apt
Shared room
Price of lisitng per night
0
100
200
300
400
500
600
700
800
Total number of reviews per listing
0
100
200
300
400
500
600

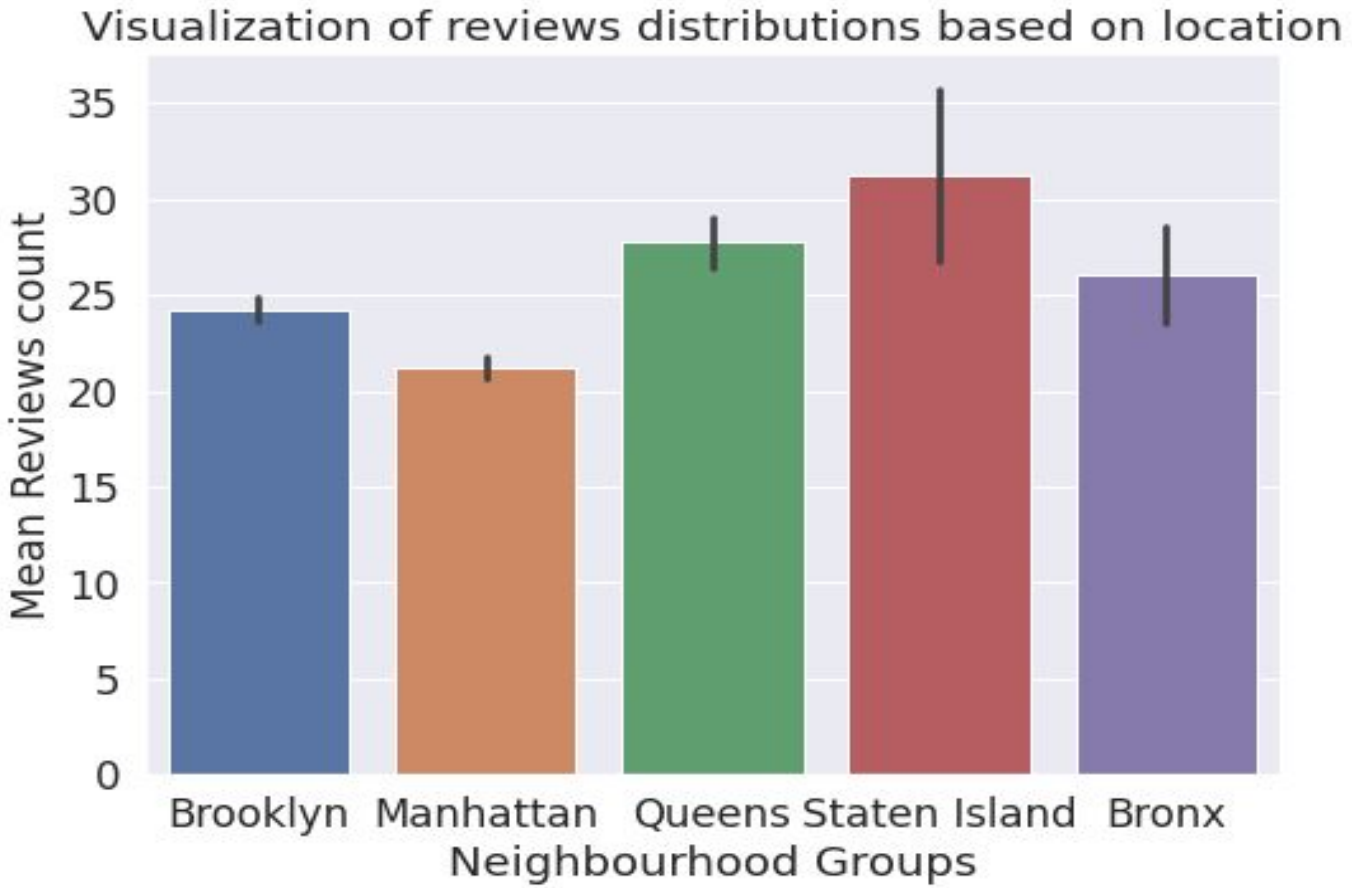
Top Busiest Hosts

	host_name	host_id	neighbourhood	reviews_per_month
22442	Louann	228415932	Rosedale	20.940000
26943	Nalicia	156684502	Springfield Gardens	18.126667
5206	Brent	217379941	Springfield Gardens	15.780000
10020	Dona	47621202	Jamaica	13.990000
8713	Danielle	26432133	East Elmhurst	13.604000
617	Aisling	256290334	Richmond Hill	13.420000
34118	Stephanie	257832461	Bushwick	13.330000
23170	Malini	111841534	Jamaica	13.150000
4543	Ben	27287203	Upper West Side	13.130000
25270	Melissa	222098649	Jamaica	13.110000

Staten Island and Queens have highest traffic

	neighbourhood_group	reviews_per_month
4	Staten Island	1.593469
3	Queens	1.570621
0	Bronx	1.478941
1	Brooklyn	1.053081
2	Manhattan	0.983633


Neighbourhood groups with median availability
availability_365
200
150
100
50
0
Staten Island
Bronx
Queens
Manhattan
Brooklyn
neighbourhood_group


Visualization of reviews distributions based on location
Mean Reviews count
0
5
10
15
20
25
30
35
Brooklyn
Manhattan
Queens
Staten Island
Bronx
Neighbourhood Groups

Conclusions

- Noticed that Manhattan has nearly 4 times the listings when compared to Queens.

- Noticed that the price distribution is right skewed with very few listings more costly than $1000 per night.

- The median price of listings in Manhattan is around $150 which is double the median price in Queens .

- There are a lot of listings available for a minimum stay of 1, 2 as well as 30 days.

- Also one can notice that, majority of the listings are of Entire home and Private room type and minimum for Shared room.

- Most of the highly reviewed listings are present in Staten Island ,Bronx and Queens.

- Noticed that the most common words used in listing names are : Bedford-Stuyvesant,East Village,East Side, Upper West, Hell Kitchen , Crown Heights.

Challenges

- Large dataset containing around 48900 rows and 16 columns.
- Managing null values and outliers.
- Reaching a proper conclusion.

THANK YOU!!