GREYATOM

Stayze

Sign up Log in

Homestay Price Prediction

BOMBAY HACKERS

Welcome to New York

Welcome Home

This project is a product of Learnings and guidance under the team from Grey Atoms -

Manish Kukreja

Ayushmaan

Sagar Dawda

Amitansh Gupta

Problem Statement

Given the available data, Stakeholders would like to know potential price and location of properties in New York for investment for which high returns can be expected.

Potential Business Opportunities

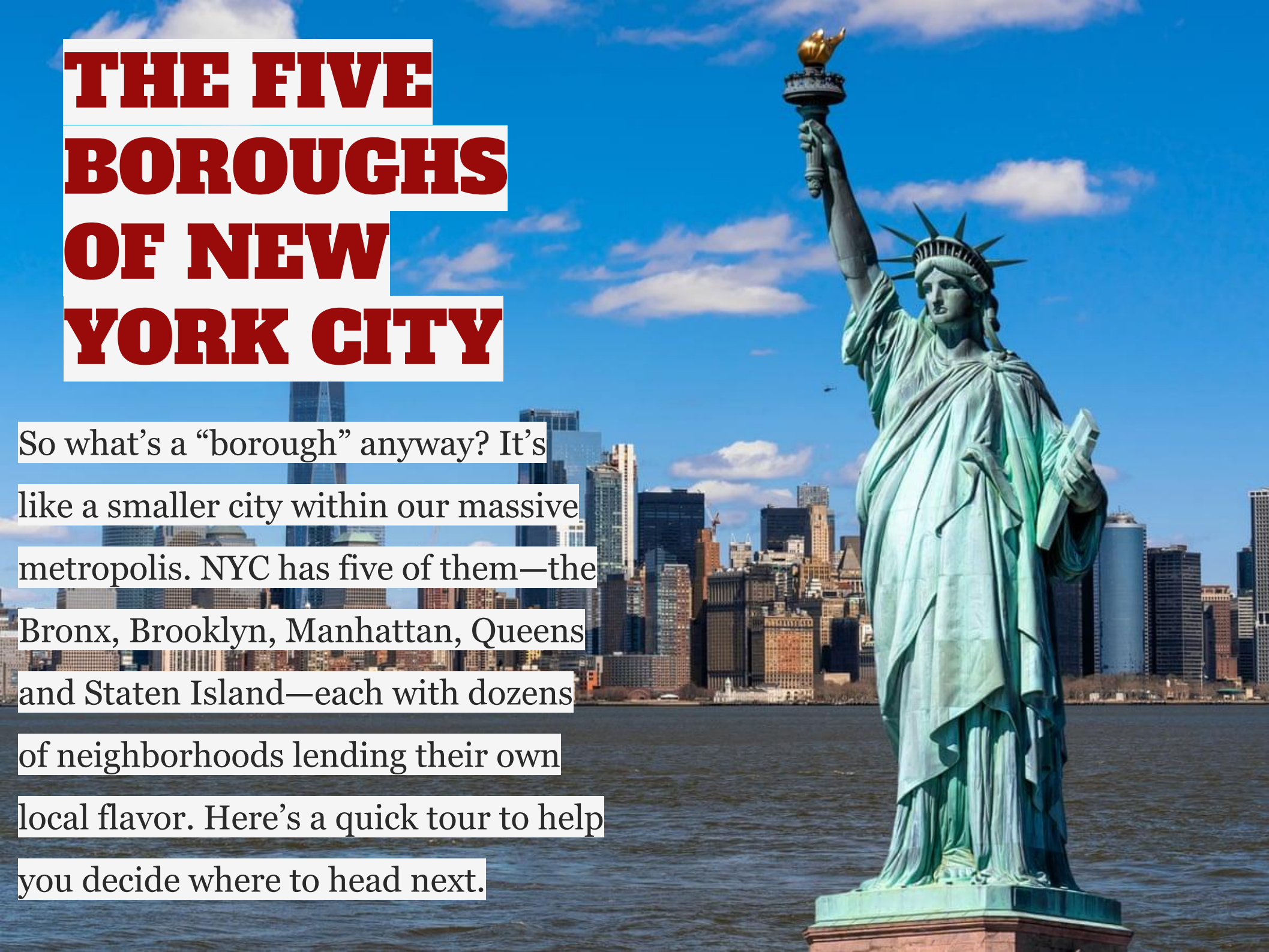
Analysis and information related to given data can help us get aware of-

1. Most in-demand Neighbourhoods for Homestay.
2. Types of rooms most preferred in different neighbourhoods.
3. Peak Time for Tourism
4. Most popular, Most costly and Cheapest Homestays and Hosts.
5. Exploring trends and patterns for revenue generation.
6. Devising a targeted campaign based on diversity of neighbourhood and potential customers.
7. Moreover, the development of the Automation Model can help us understand the most important parameters that provide the best predicted results, and can also provide us with tailored model, making cost analysis much simpler in future.

Various Stakeholders Considered

1. **Chief Executive Officer** (CEO) for the final Decision
2. The **Host** - for the tie-up
3. **FPA Team**- for analysing reasonable Commissions
4. **Operations Team** - for keeping positive flow of communication with host.
5. **Sales and Marketing Team** - For targeting the potential customers for the diverse neighbourhoods.

THE FIVE BOROUGHS OF NEW YORK CITY



So what's a "borough" anyway? It's like a smaller city within our massive metropolis. NYC has five of them—the Bronx, Brooklyn, Manhattan, Queens and Staten Island—each with dozens of neighborhoods lending their own local flavor. Here's a quick tour to help you decide where to head next.

BROOKLYN

Hipsters and history buffs agree: Brooklyn is everyone's style. The most populous borough has no one defining draw.

1. Williamberg- Live music and incubation for bands.
2. Bedford- Suburban feel. Owned apartments.Considered best to live.
3. Bushwick- Latino American Majority. immigrants from Mexico and El Salvador.
4. Crown Heights - West Indian CarnivaIndian and Afro-Americans. West l(labor day) attract ~3M people.

*The diameter of circle is proportional to number of properties in the region.

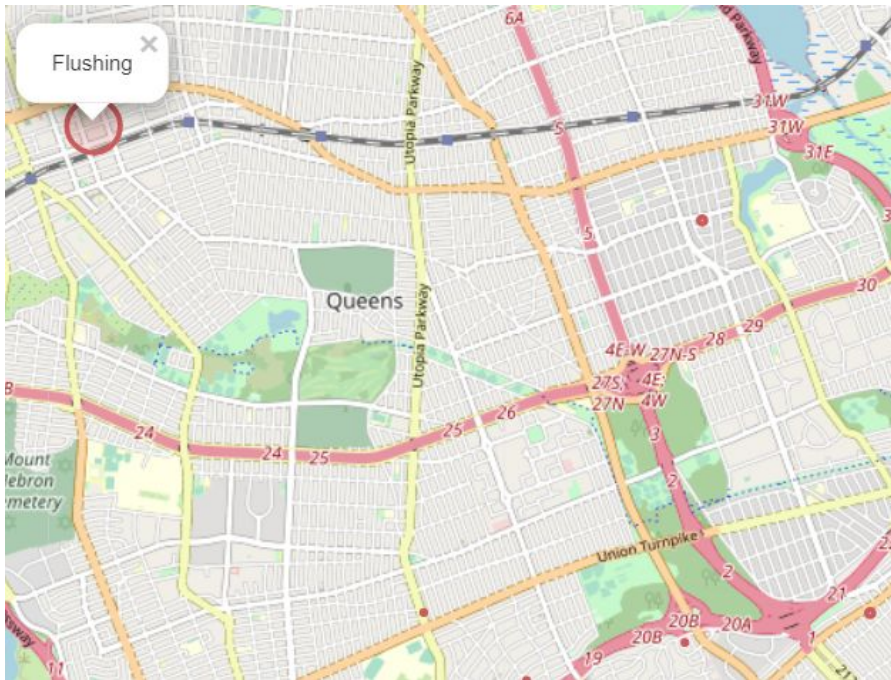
MANHATTAN

1. Harlem - 77 % Afro-American population and Asian immigrant influx.
2. Hell's Kitchen and Chelsea - Emerged from cheap rents for aspiring entertainers like Madonna and S.Stallone. Home to broadcast ans music industry.
3. Midtown - Corporate HQ area. Eg. E&Y,NY institute of Finance, Morgan Stanley, Warner Bros, MARVAL etc.
4. Upper East Side and west side - Asian Population influx.
5. East Village - ISKCON founded here.
6. Financial District - World Trade centre.

QUEENS

Queens is one of the most diverse places in the world. Many get their fill from the food scene alone, which ranges from exquisite Greek souvlaki to the best hot pot outside of Sichuan. The borough is also a major destination for sports fans, nature lovers and modern-art aficionados.Take your pick from the array of amazing ethnic restaurants in Queens.

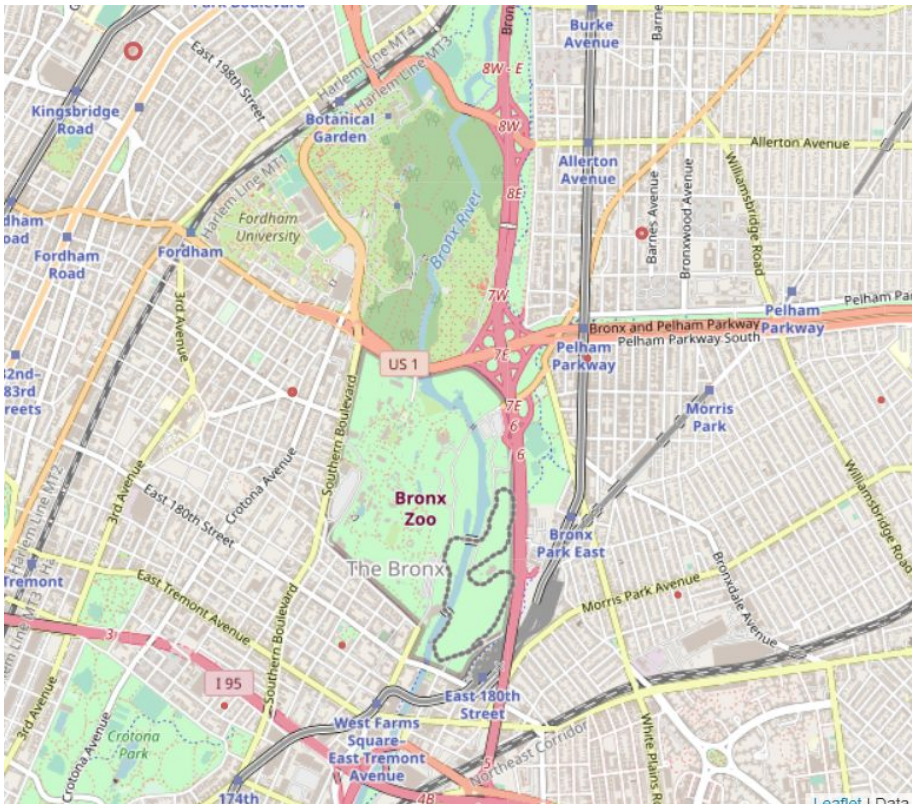
Flushing - 70 % Asian population



THE BRONX

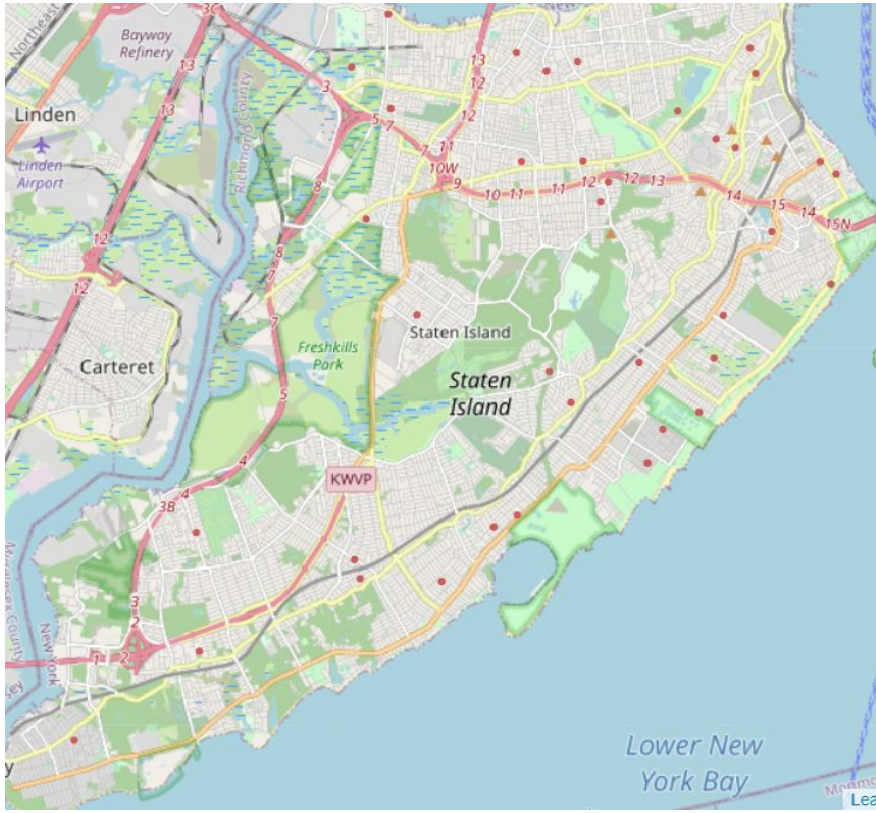
What do home-run heroes, hip-hop history and hand-pulled mozzarella have in common? You can find them all in the Bronx—that is, when you're not wandering hundreds of acres of parkland and **Bronx Zoo**, touring historic homes or strolling along New York City's answer to the Champs-Élysées.Also, New York Botanical Garden

- Yankee Stadium.



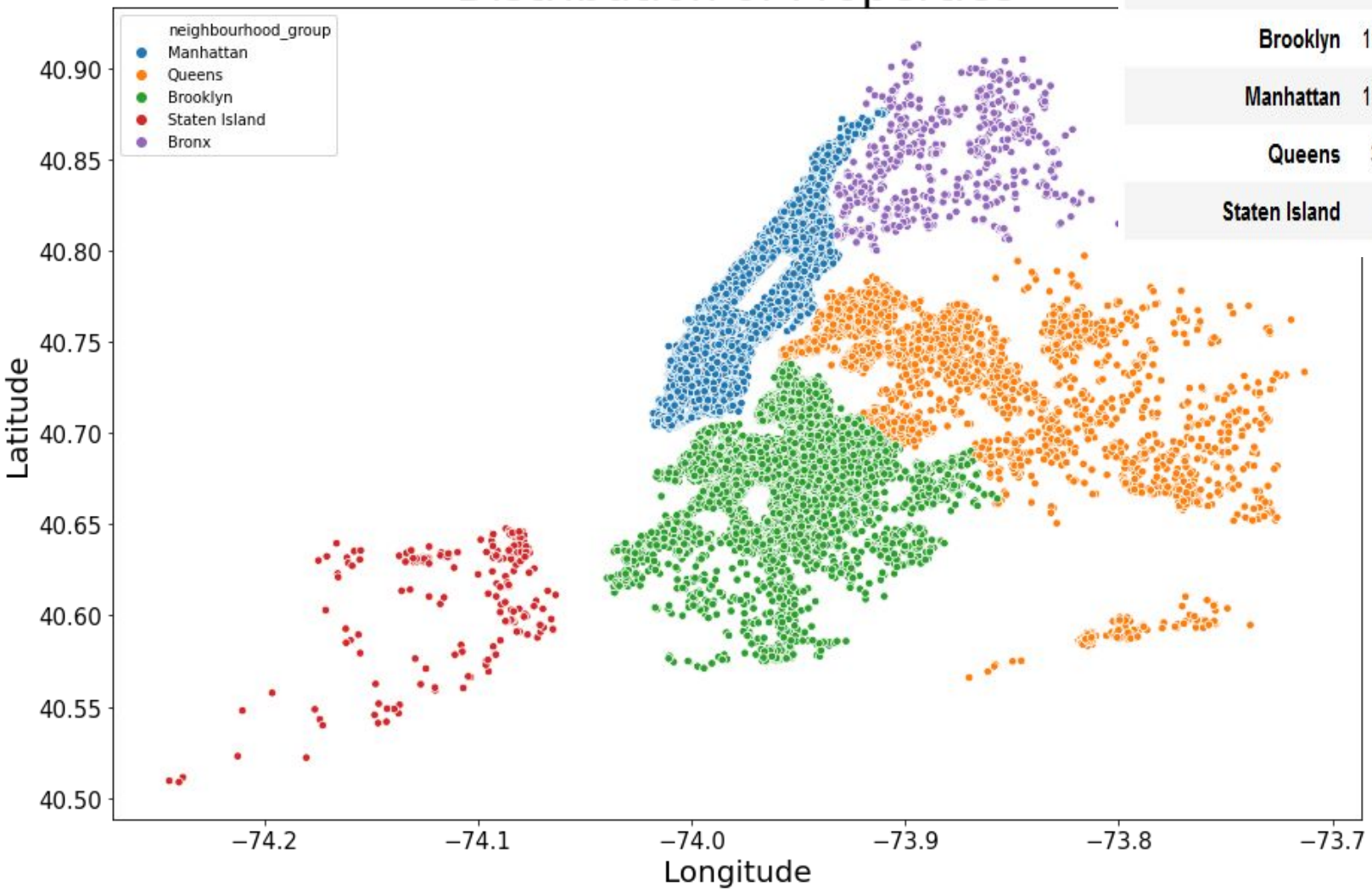
STATEN ISLAND

Easily accessible by a scenic ride on the **Staten Island Ferry**, the **greenest borough** feels like a getaway within the City. Beyond a charming North Shore rich with maritime history, Staten Island is best known for its beaches, vast parkland and even a fully preserved colonial village.you can also take the opportunity to visit the Staten Island Museum in Snug Harbor.It also has nearly dozens of Golf Courses.



Distribution of property along Latitude and Longitude coordinates

neighbourhood_group	count	unique	top	freq
Bronx	618	3	Private room	366
Brooklyn	11512	3	Entire home/apt	5751
Manhattan	11671	3	Entire home/apt	6990
Queens	3213	3	Private room	1861
Staten Island	216	3	Private room	115



More than 50 % of the properties are located in the Manhattan and Brooklyn and outliers in the prices is also more in thes areas.

Most popular neighborhoods

neighbourhood_group	neighbourhood	no. of properties
Brooklyn	Williamsburg	2756
	Bedford-Stuyvesant	2577
Manhattan	Harlem	1877
Brooklyn	Bushwick	1719
Manhattan	Upper West Side	1403
	Hell's Kitchen	1383
	East Village	1309
	Upper East Side	1245

Inference:

About 85% properties in the data belong to Manhattan and Brooklyn.

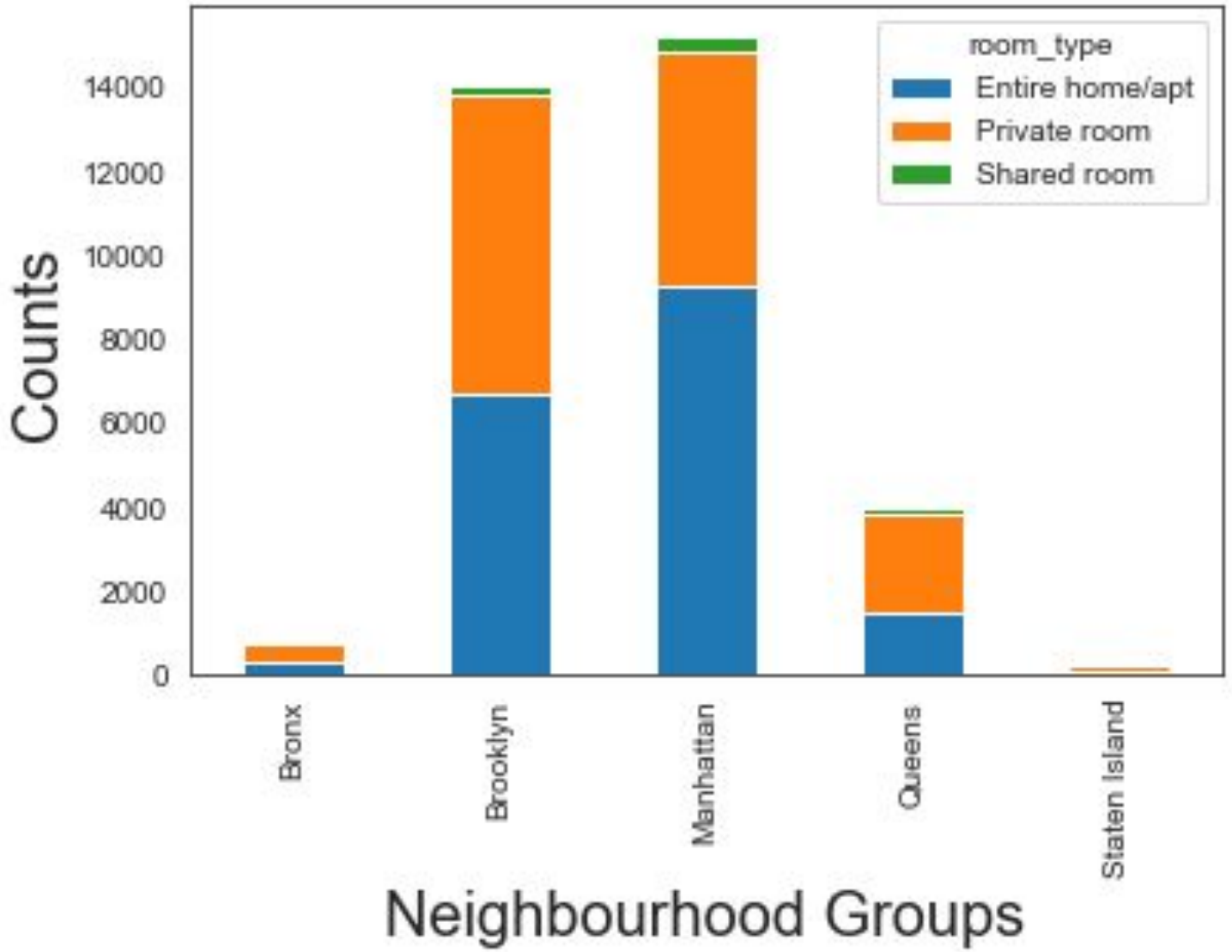
Staten Island has only 248 properties listed

(Queens: 3985, Bronx: 767)

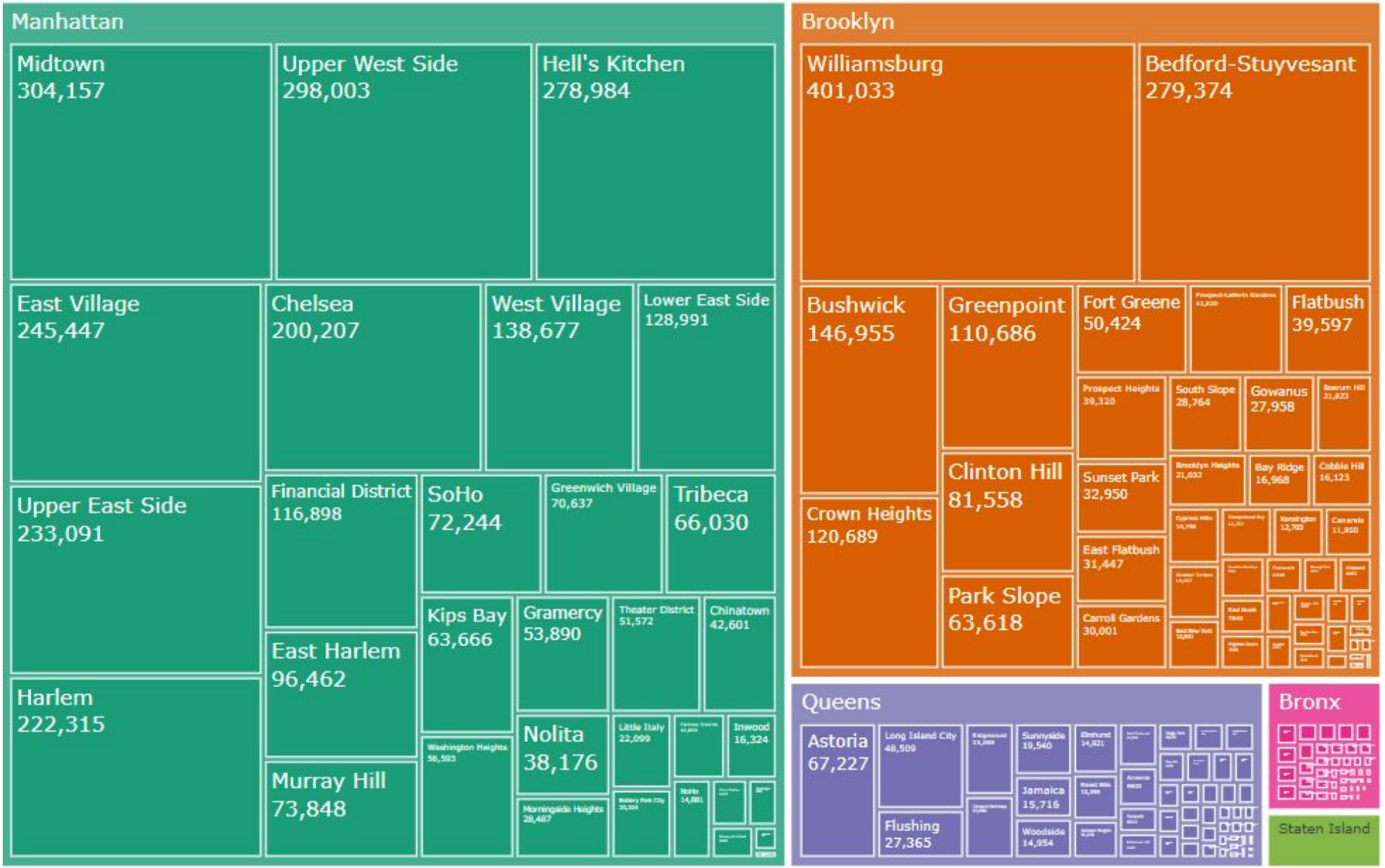
Location information is vital for statistical analysis

Types of room in Neighbourhood Groups

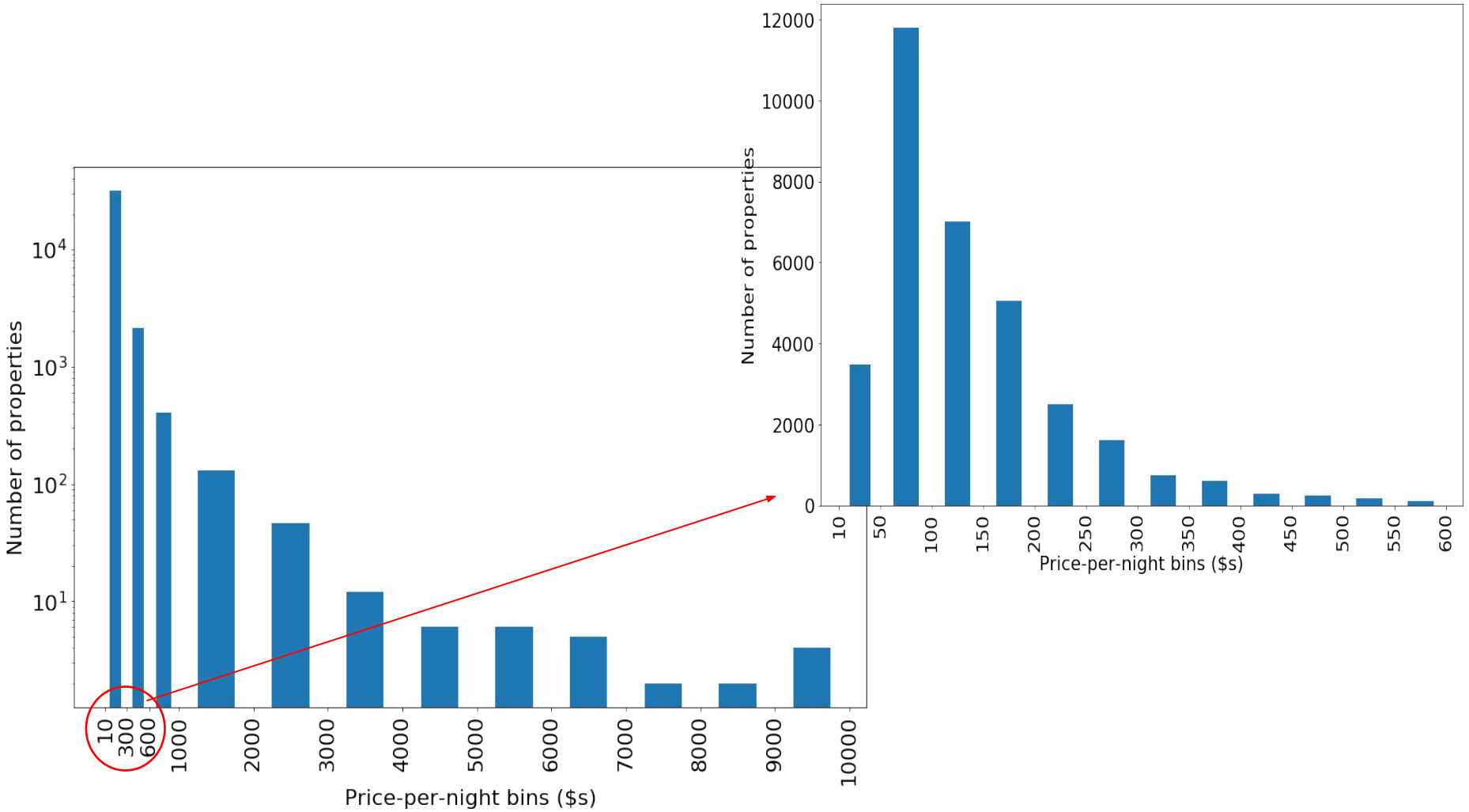
Manhattan and Brooklyn are the most in demand destinations. Also, Preference is higher for Whole apartment and Private rooms. Hence, it is the most important customer base.



Potential Revenue Generation wrt Neighbourhood



Price Distribution of properties

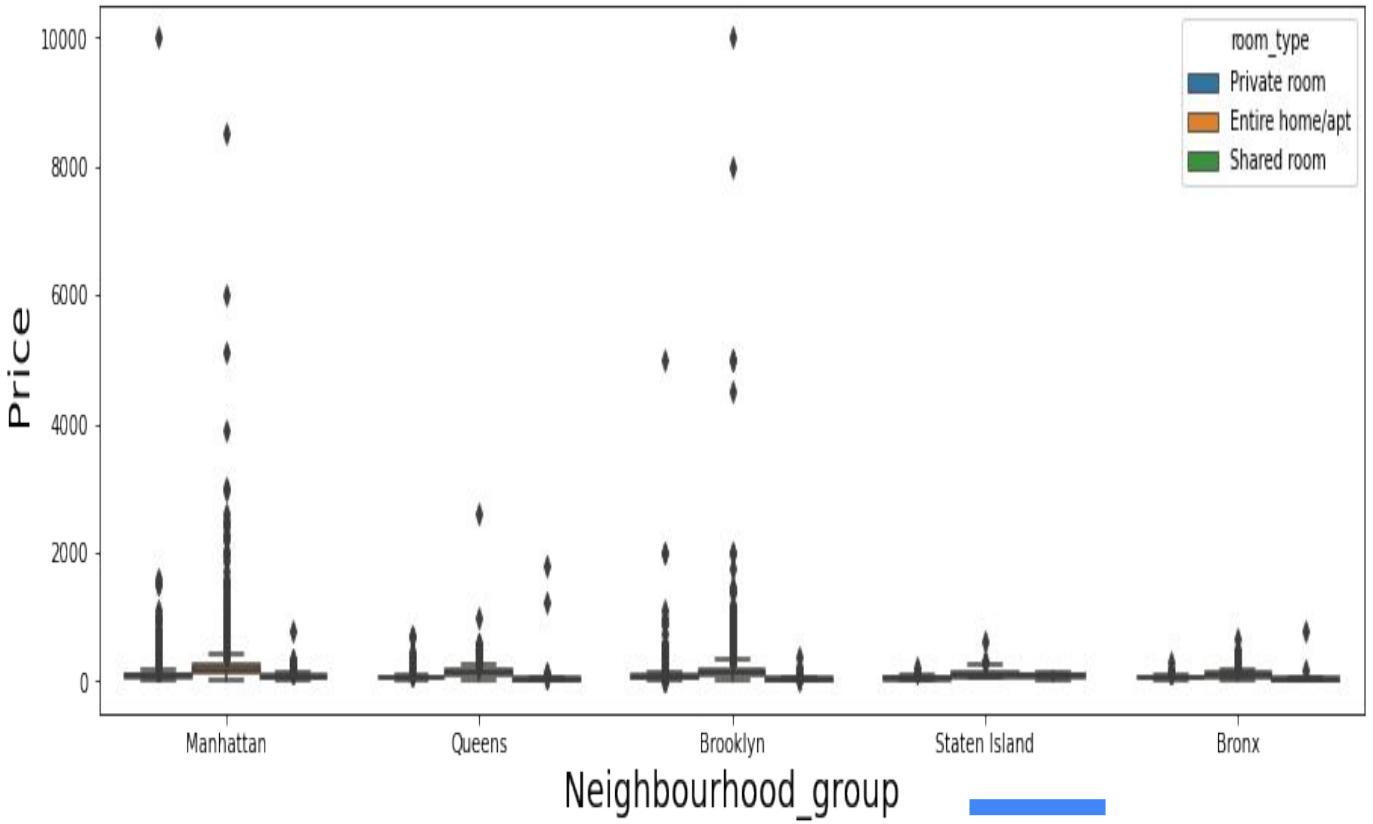


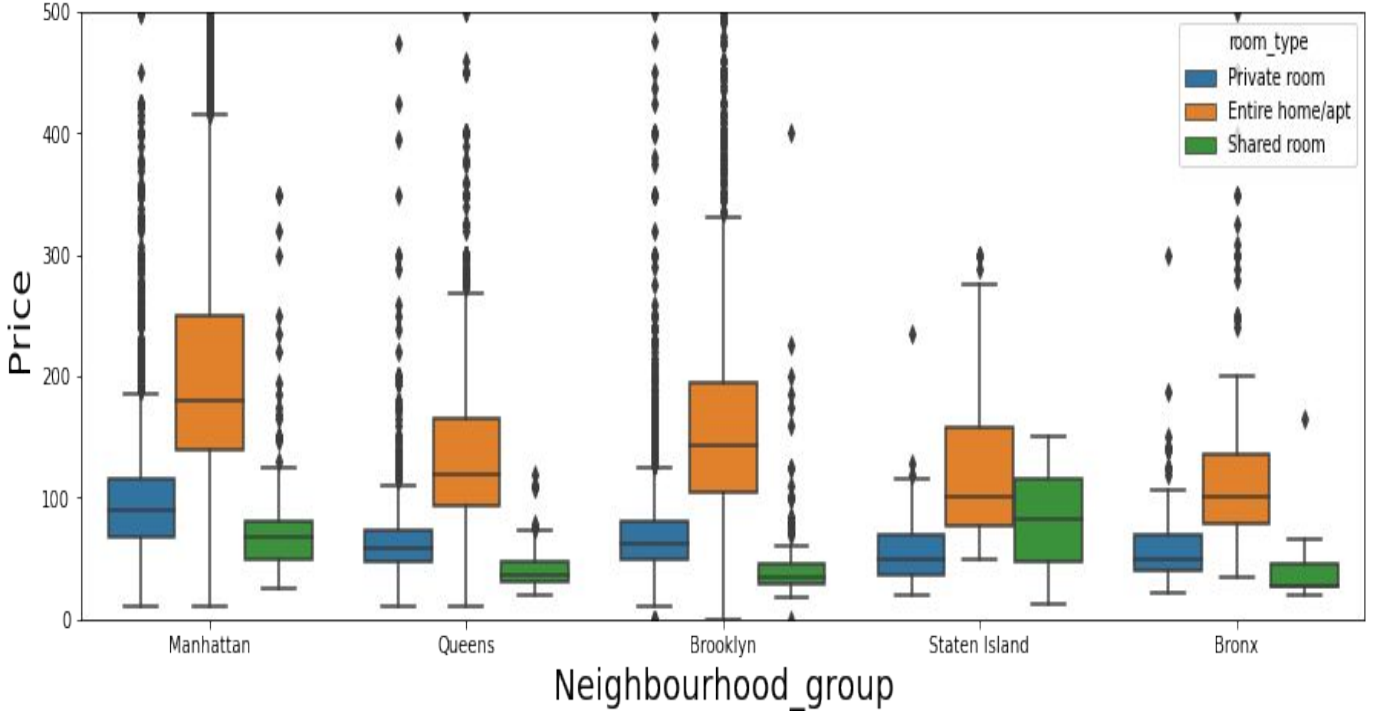
Inference:

Distribution is positively skewed

Majority of the properties (about 1/3 rd) are priced less than 100 $s per night.

Only 892 out of 34226 properties charges more than 500 $s per night





1. Lot of outliers are present neighbourhood_groups (Manhattan and Brooklyn) when compared to other groups for different room_types
2. The mean is larger than median which clearly shows that the distribution is skewed
3. In order to predict price nicely we have to do some transformations to convert this skewed distribution to normal distribution
4. We have used log transformation to convert this distributions.

Costliest and Cheapest Properties (Top 10)

Costliest Properties

host_name	neighbourhood_group	neighbourhood	room_type	price
Erin	Brooklyn	Greenpoint	Entire home/apt	10000
Jelena	Manhattan	Upper West Side	Entire home/apt	10000
Matt	Manhattan	Lower East Side	Entire home/apt	9999
Amy	Manhattan	Lower East Side	Private room	9999
Rum	Manhattan	Tribeca	Entire home/apt	8500
Jessica	Brooklyn	Clinton Hill	Entire home/apt	8000
Sally	Manhattan	Upper East Side	Entire home/apt	7703
Jack	Manhattan	Battery Park City	Entire home/apt	7500
Kevin	Manhattan	Chelsea	Entire home/apt	6800
Jonathan	Brooklyn	Clinton Hill	Entire home/apt	6500

Cheapest Properties

host_name	neighbourhood_group	neighbourhood	room_type	price
Katie	Brooklyn	Bushwick	Private room	10
Julio	Brooklyn	Bushwick	Private room	10
Maria	Queens	Jamaica	Entire home/apt	10
Jennifer	Manhattan	SoHo	Private room	10
Amy	Manhattan	Upper East Side	Entire home/apt	10
Maria	Queens	Jamaica	Entire home/apt	10
Martin	Brooklyn	Williamsburg	Private room	10
Salim	Manhattan	Upper West Side	Private room	10
Caterina	Brooklyn	Bedford-Stuyvesant	Entire home/apt	10
Sally	Manhattan	East Village	Entire home/apt	10
Rachel	Brooklyn	Sunset Park	Entire home/apt	10

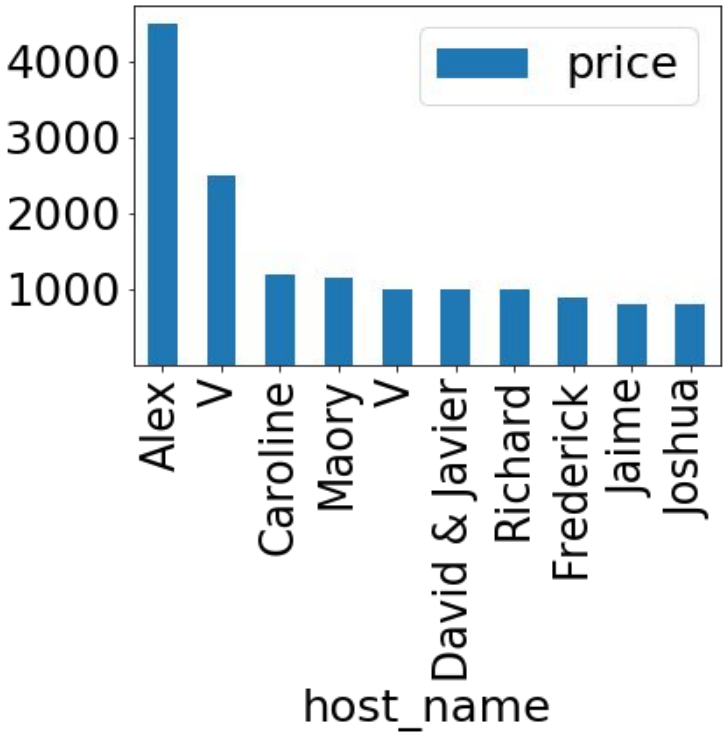
Inference:

room_type ‘Entire home/apt and Private room ’ are listed in both costliest and cheapest properties

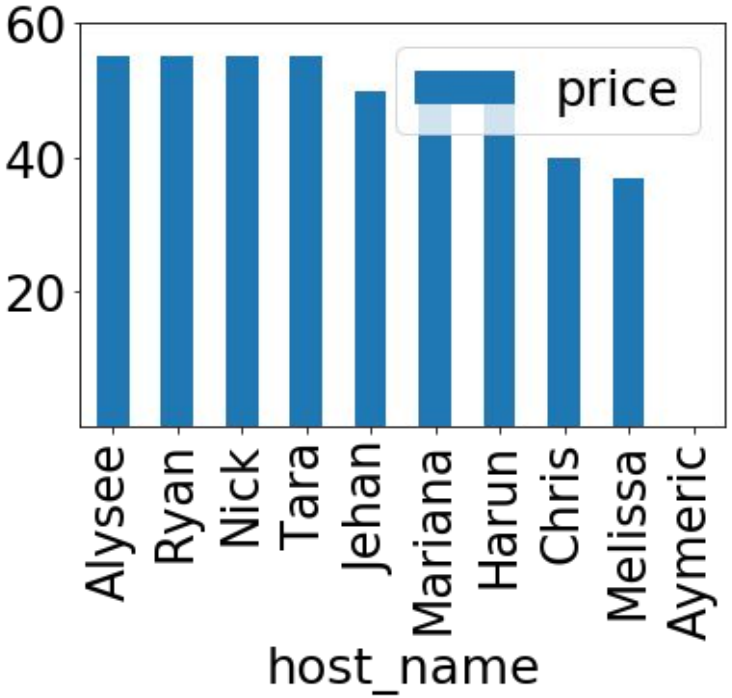
 room_type is not a good indicator of price per night

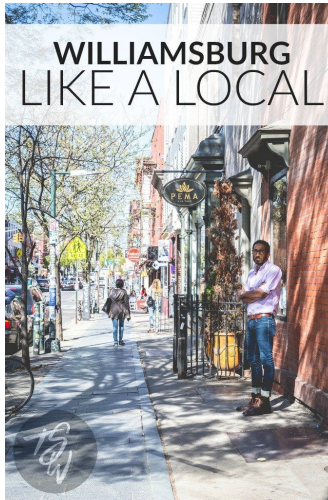
Price variation for **'Entire home/apt'** in the most popular neighbourhood: Williamsburg, Brooklyn

Costliest Hosts



Cheapest Hosts



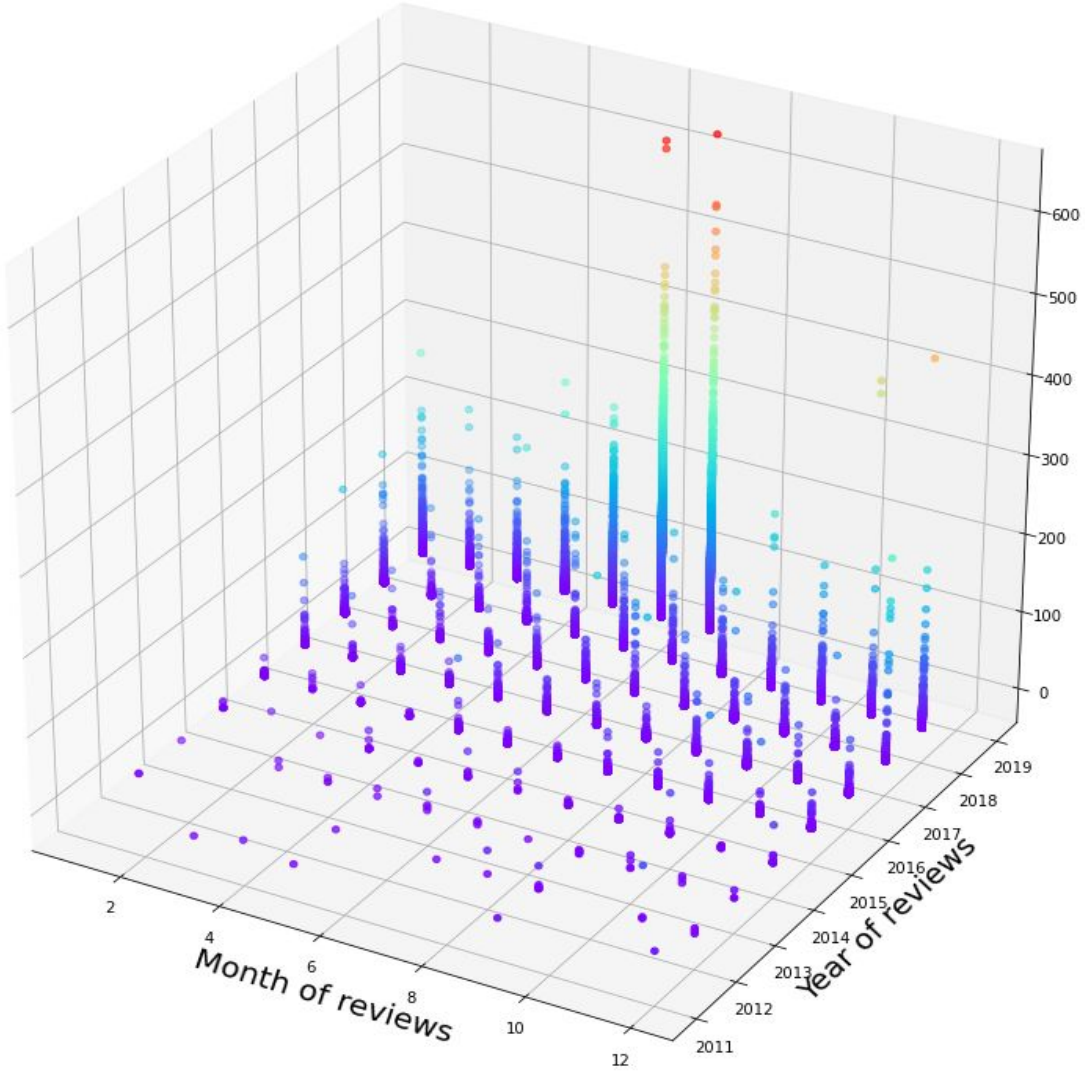


Inference:

"host_id" is a good indicator of price per night

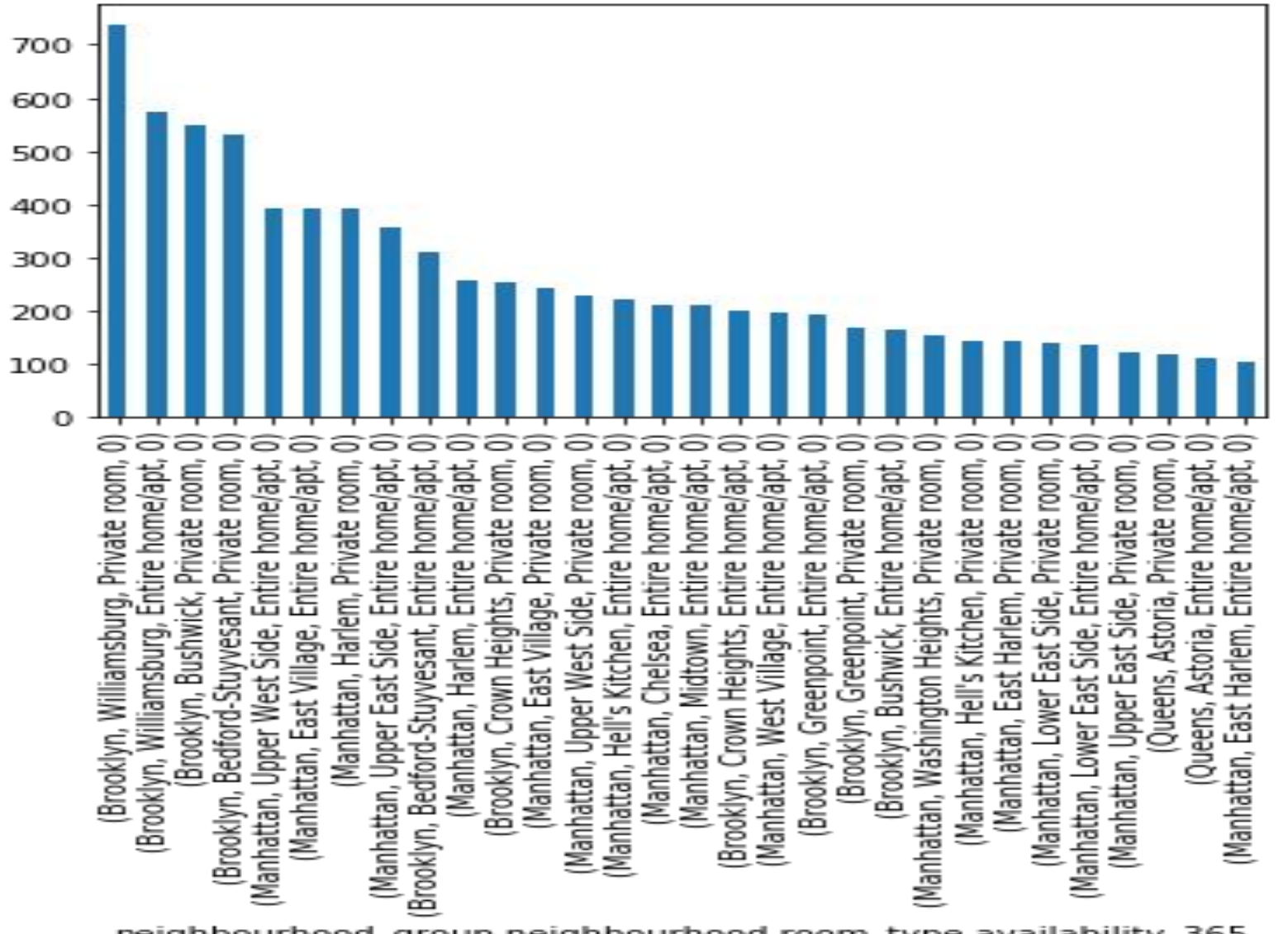
Best Time for Tourism

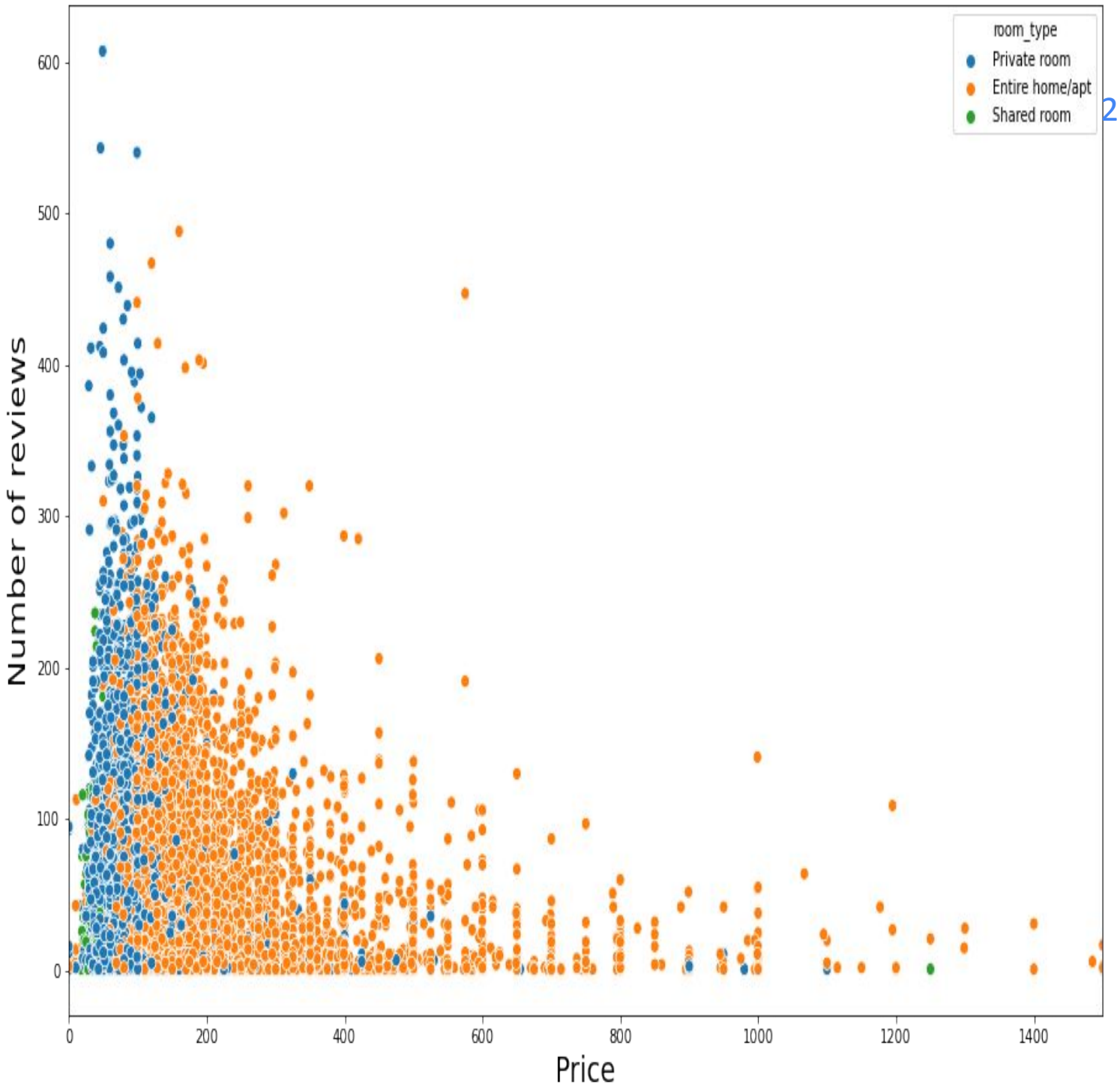
- June and July



	rainfall	high	low
jan	3.9"	39°F (4°C)	26°F (-3°C)
feb	3.0"	42°F (6°C)	29°F (-2°C)
mar	4.0"	50°F (10°C)	35°F (2°C)
apr	3.9"	60°F (16°C)	44°F (7°C)
may	4.4"	72°F (22°C)	55°F (13°C)
june	3.5"	75°F (24°C)	60°F (16°C)
july	4.0"	85°F (29°C)	70°F (21°C)
aug	4.1"	88°F (31°C)	75°F (24°C)
sep	4.0"	77°F (25°C)	60°F (16°C)
oct	3.4"	67°F (19°C)	51°F (11°C)
nov	4.5"	54°F (12°C)	41°F (5°C)
dec	3.6"	44°F (7°C)	32°F (0°C)

Most in Demand (wrt neighbourhood group,neighbourhood, room type & availability)



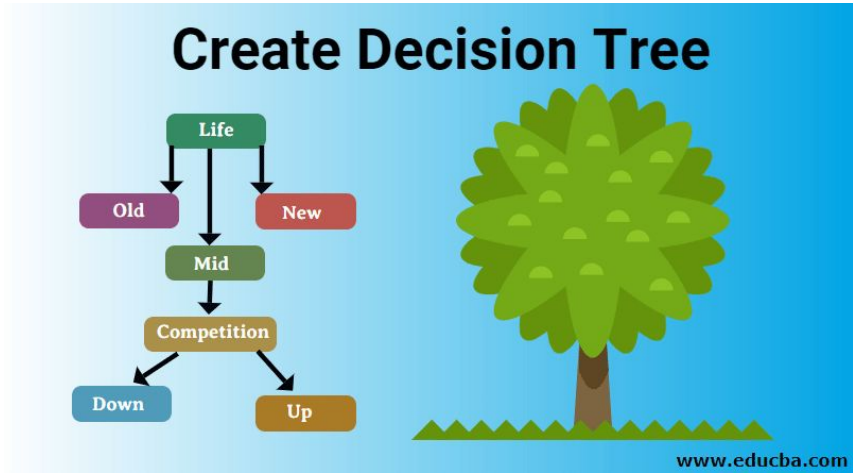


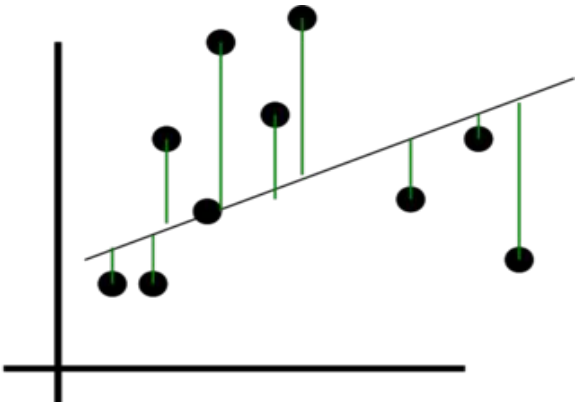
1. Most number of reviews are there for low price properties
2. We can infer more people prefer low price properties rather than high price room-type

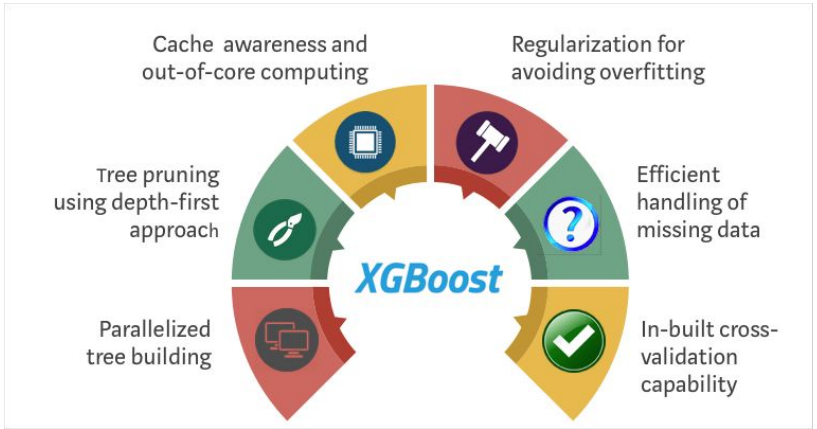
Modelling

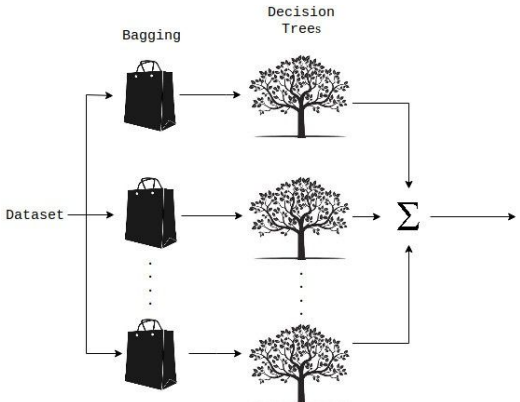
We experimented with various ML Regression Models to understand the parameter behaviour and provide you with most accurate and stable predictions. We used -

1. Linear Regression
2. XG Boost
3. Lasso Regression
4. Decision Tree Regression
5. Random Forest using Seach CV

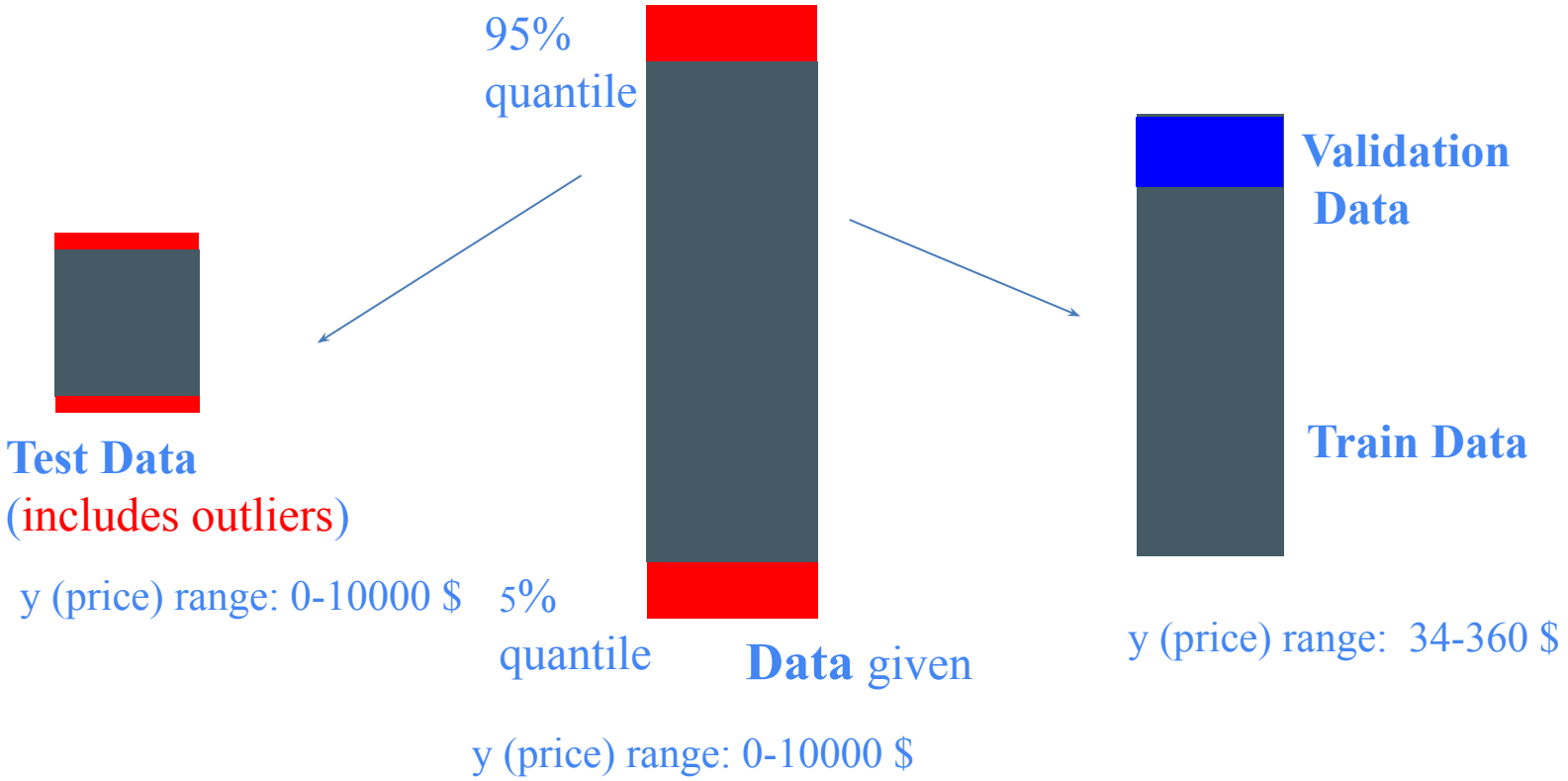






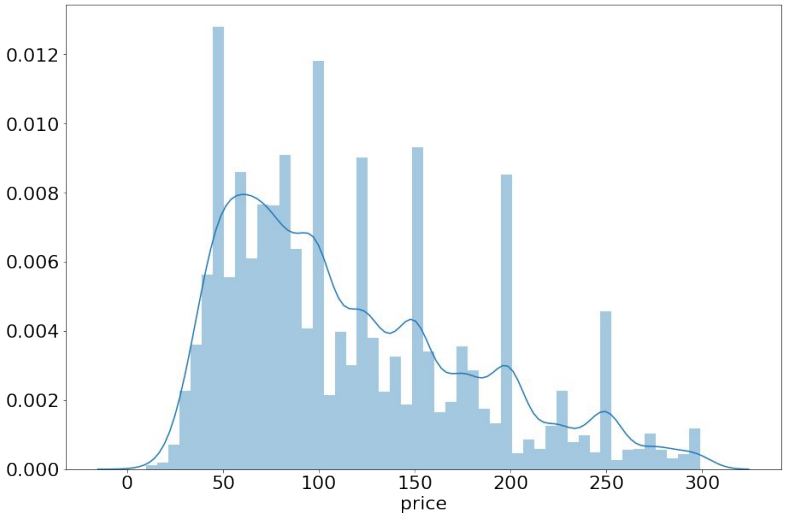


Outlier Removal



- Data provided (in Train.csv file) is divided into train, validation and test data.
- Data outside 5-95% quantile (price value) considered as outliers.
- Outliers are not included in train and validation data as we are developing generic model.
- Outliers are included in test data for rigorous testing of model.

Effect of Logarithmic transformation on accuracy of the model



*Price distribution **y***

*Price distribution **log(y)***

Transformation from positively skewed to near normal distribution using ***log***

	Linear Regression .score	RMSE
Without log10 transformation	0.41	55.33
With log10 transformation	0.50	55.94

Price range considered for modelling10-300 $

LinearRegression.score = $1-u/v$

where $u = sum(y_{true} - y_{predit})^2$

$v = sum(y_{true} - y_{true\_mean})^2$

Logarithm transformation of the dependant variable (price) is considered for modeling

Linear Regression Model Experiments :

These are Predictions based on various set of independent variables

Expt	Lat	Long	Distance	Nb-G	Nb	room-typ	id	host_id	# of reviews	Log(price)	Outliers	Score	Rmse-test	Rmse-train	hotEncoding
1	✓		✓							✓	✓	0.116	69.039	no	no
2		✓	✓							✓	✓	0.145	67.823	no	no
3		✓	✓							x	✓	0.117	66.619	no	no
4		✓	✓							x	x	0.035	197.26	no	no
5		✓	✓			✓				x	x	0.086	191.94	no	no
6	✓		✓			✓				✓	✓	0.376	61.42	no	no
7	✓		✓	✓		✓				✓	✓	0.396	60.42	no	no
8	✓	✓	✓	✓		✓				✓	✓	0.4	60.95	no	no
9	✓	✓	✓			✓				✓	✓	0.549	43.58	44.26	Yes
10	✓	✓	✓	✓		✓				✓	✓	0.565	43.26	43.77	Yes
11	✓	✓	✓	✓	✓	✓				✓	✓	-5.42E+13	43.26	43.77	Yes
12	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	0.603	41.392	41.81	Yes

Here, we observed that host_id, latitude, longitude, neighbourhood_group and room_type are the most important variables for Price prediction.

In Expt 12: we have also used other variables likemonth_of_reviewandYear_of_review)

**Hot encoding has been done for only 3 variables (room_type, Neighbourhood_group and Neighbour)

Performance of Models

Model	RMSE on Train Data	RMSE on Validation Data	RMSE on Test Data (with outliers)
Linear Regression	54.45	55.94	215.18
XG Boost	117.06	117.56	250.91
Lasso	72.53	73.76	226.20
Decision Tree Regressor	0.00	70.31	208.0
Random Forest (with search CV)	51.67	53.38	213.93

- **Random Forest model is giving the best performance among the models considered and is used for the final price prediction**

**Decision Tree Regressor is overfitting the data

Summary of Observations

1. Out of 34266 locations, around 29266 homestay locations are in Manhattan and Brooklyn (85% of total data).
2. Moreover, Williamsburg and Bedford alone constitute more than 5000 properties.
3. Preference is higher for Whole apartment and Private rooms.
4. Majority of the properties (about 1/3 rd) are priced less than 100 $ per night.
5. Only 892 out of 34226 properties charges more than 500 $ / night
6. “host_id” is a good indicator of price per night
7. Private rooms in Williamsburg are most in demand, with price ranging between $50 to $100.
8. No. of reviews are increasing with the passing years.Maximum reviews have been noticed during June and July. (in Summer season)
9. Most number of reviews are there for low priced properties

Recommendations

1. **For CEO** - 85 % properties belong to Manhattan and Brooklyn. Hence, We should consider these areas maximum business generators.
2. **FPA Team**- Among the Neighbourhoods except Manhattan and Brooklyn, Flushing in Queens seem the only potential business generator, as properties in other neighbourhoods are sparsely located and populated.
3. **Operations Team** - Michael is the most popular host. We can look for tie-ups with hosts like him.
4. **Sales and Marketing Team** - June and July are best time for tourism. Also, West Carnival Festival attracts around 3 Million visitors. We can plan our campaigns accordingly.


STAYZE
Welcome Home
STAYZE
Welcome Home

TOUGH TIMES DON’T LAST BUT TOUGH TEAMS DO

SVLS Rao (Mumbai)

Rohit Gupta (Jammu)

Ritesh Patel (Mumbai)

Kondayya Naidu (Mumbai)

References

1. Code Along Saturday by Grey Atoms
2. Wikipedia
3. keggle.com
4. github.com
5. NYC Website
6. google.com (for photographs)

Python Libraries Used-
1. Pandas
2. Numpy
3.Matplotlib
4.Folium
5.Plotly
6.Seaborn
7. Sci-kit learn

For Code and Datasets, Please visit-

- https://github.com/rohitgupta29/Hackathron_1/blob/master/EDA_Bombay%20Hackers_oct10.ipynb (EDA)
- https://github.com/gkondayya/Hackathon/blob/master/final_gk.ipynb (ML Automation and Prediction)


Thank You

Appendix - Potential Revenue wrt Room Type