STAYZE – Rent Prediction

TEAM – DATA BLITZ

Dashang Makwana
Nevina Dalal
Simran Mulani
Vishakha Singh

Problem Statement

We are provided with all the demographics for Stayze; about which all properties/houses are associated with it, their locations and other such details. We are asked to study the data and Predict the Rent.

Our primary stakeholder is the sales manager of Stayze.

Datasets

We were provided with the Train and Test datasets with the following information and details:

Name

Details of the listing

Price

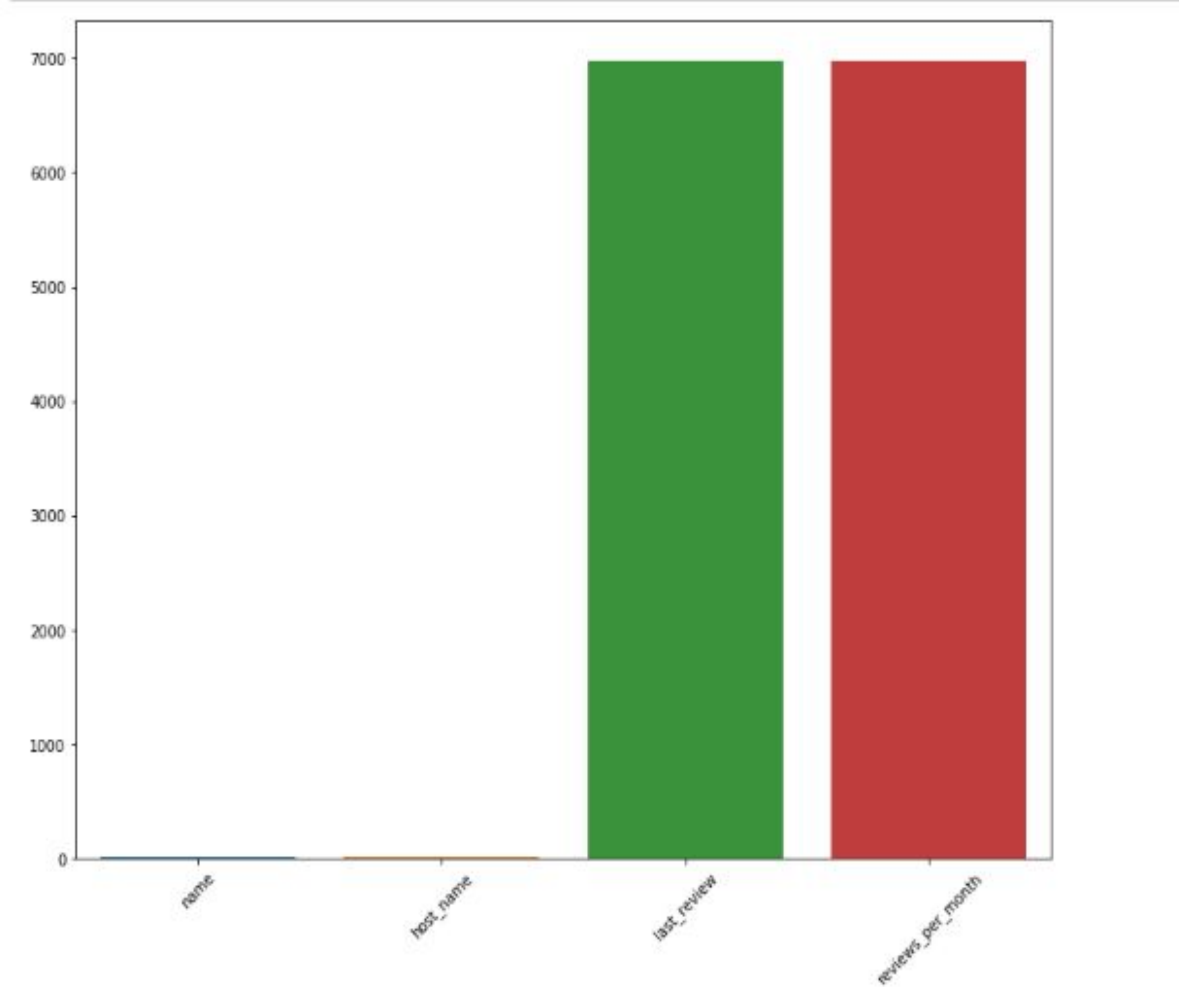
Reviews, availability etc.

Pre-processing

Removed unwanted variables : name and host name

Replaced missing values in reviews with 0

Label encoded the Category columns :neighbourhood, neighbourhood group, room type



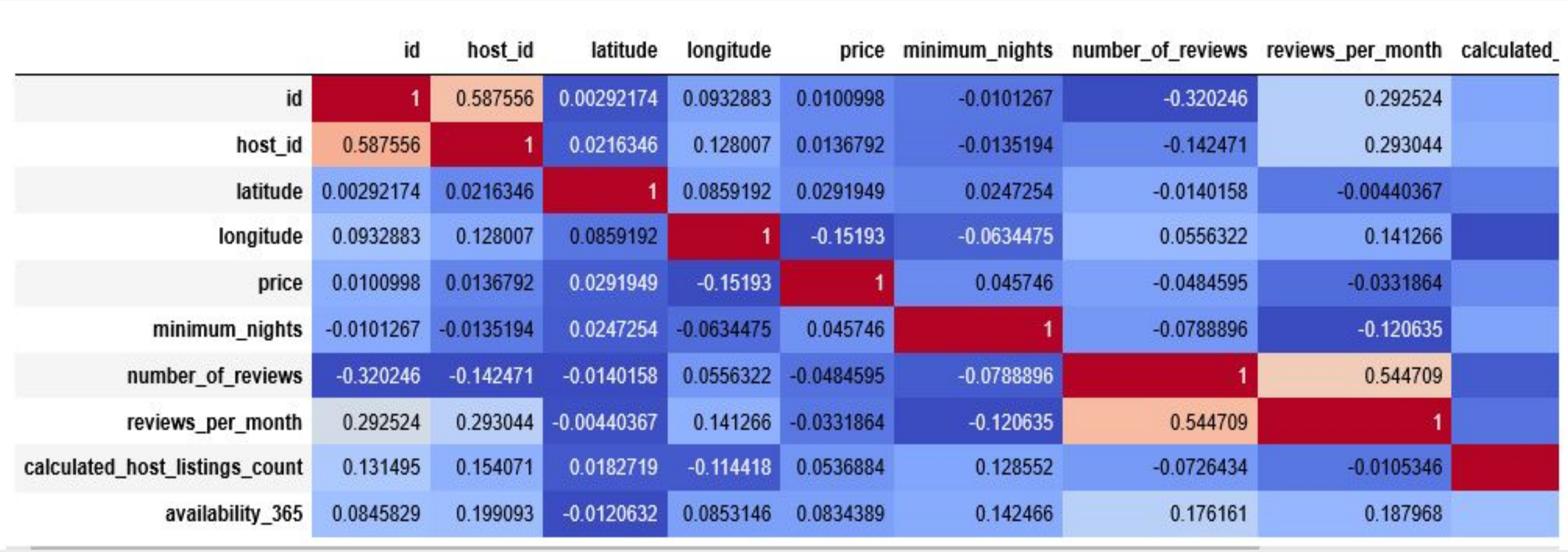
Descriptive Statistics

We are considering 34226 rows in the train data

The price of the listing ranges from 0 USD to 10000 USD, with a mean price of 153 USD and a standard deviation of 243 USD suggesting a very broad price range distribution.

Availability of the property ranges from a day to all year round, with a median of 44 days.

Correlation

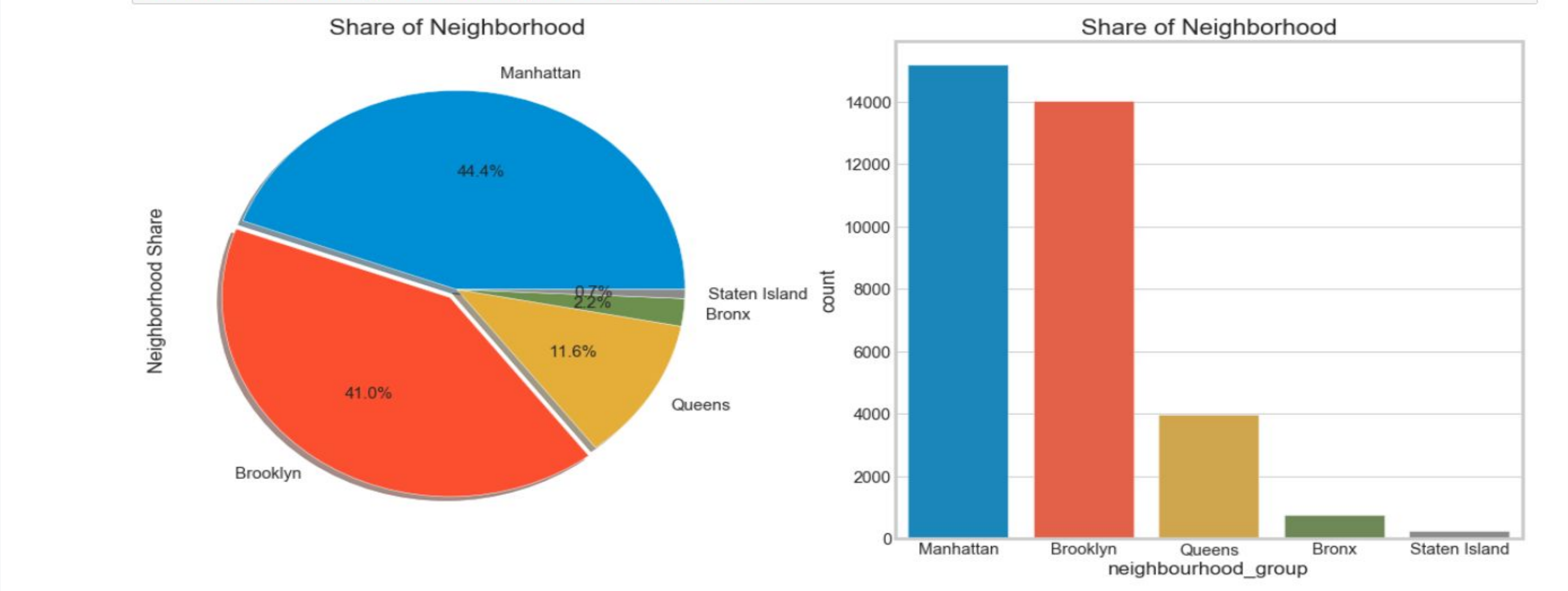


	id	host_id	latitude	longitude	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_
id	1	0.587556	0.00292174	0.0932883	0.0100998	-0.0101267	-0.320246	0.292524	
host_id	0.587556	1	0.0216346	0.128007	0.0136792	-0.0135194	-0.142471	0.293044	
latitude	0.00292174	0.0216346	1	0.0859192	0.0291949	0.0247254	-0.0140158	-0.00440367	
longitude	0.0932883	0.128007	0.0859192	1	-0.15193	-0.0634475	0.0556322	0.141266	
price	0.0100998	0.0136792	0.0291949	-0.15193	1	0.045746	-0.0484595	-0.0331864	
minimum_nights	-0.0101267	-0.0135194	0.0247254	-0.0634475	0.045746	1	-0.0788896	-0.120635	
number_of_reviews	-0.320246	-0.142471	-0.0140158	0.0556322	-0.0484595	-0.0788896	1	0.544709	
reviews_per_month	0.292524	0.293044	-0.00440367	0.141266	-0.0331864	-0.120635	0.544709	1	
calculated_host_listings_count	0.131495	0.154071	0.0182719	-0.114418	0.0536884	0.128552	-0.0726434	-0.0105346	
availability_365	0.0845829	0.199093	-0.0120632	0.0853146	0.0834389	0.142466	0.176161	0.187968	

No strong correlation except number_of_reviews vs reviews_per_month

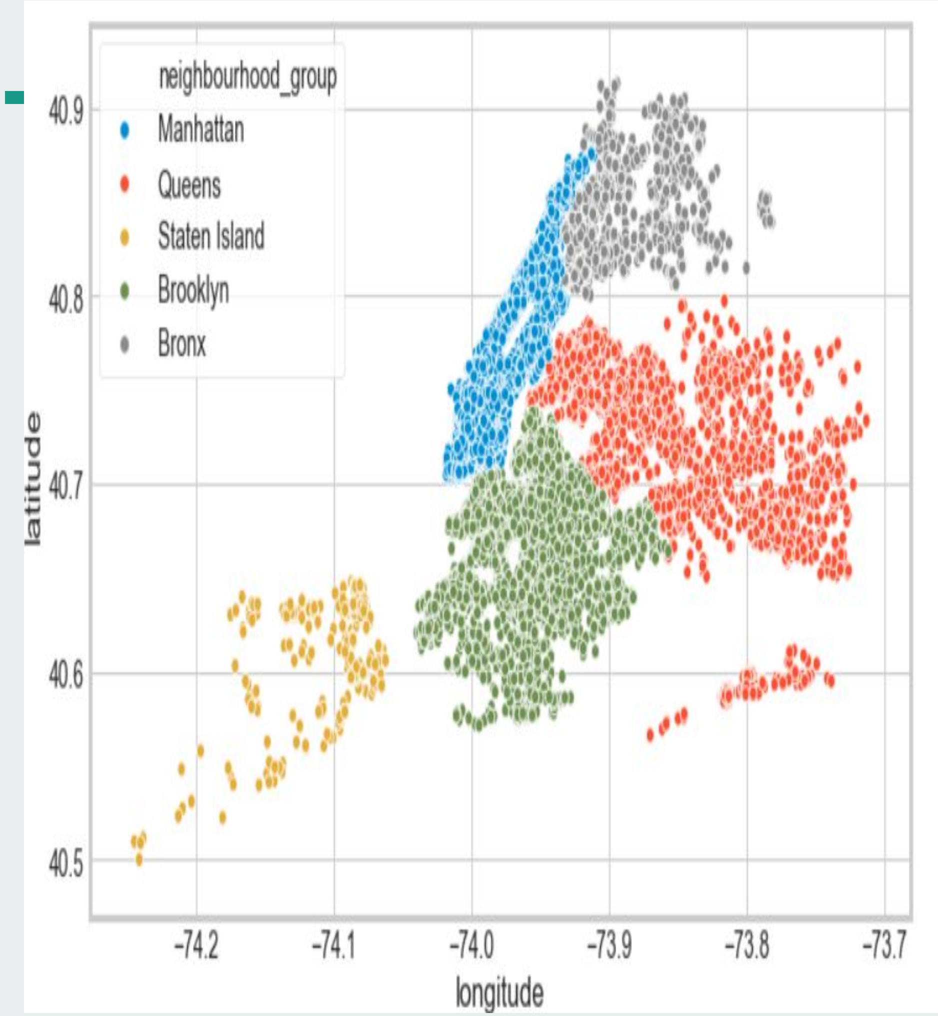
Longitude is slightly negatively correlated with the price which could suggest that as we move west the prices increase

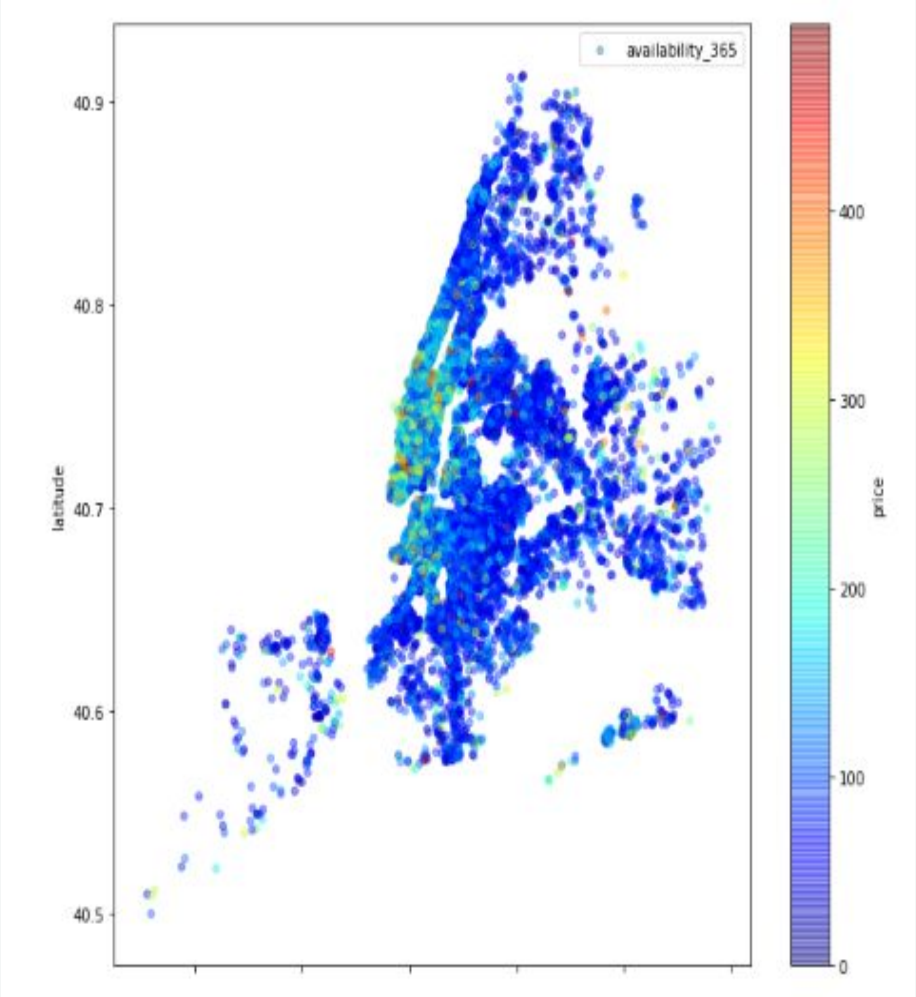
Neighbourhoods



Clearly through this graph we understand that predominantly Stayze has more property in Manhattan and Brooklyn

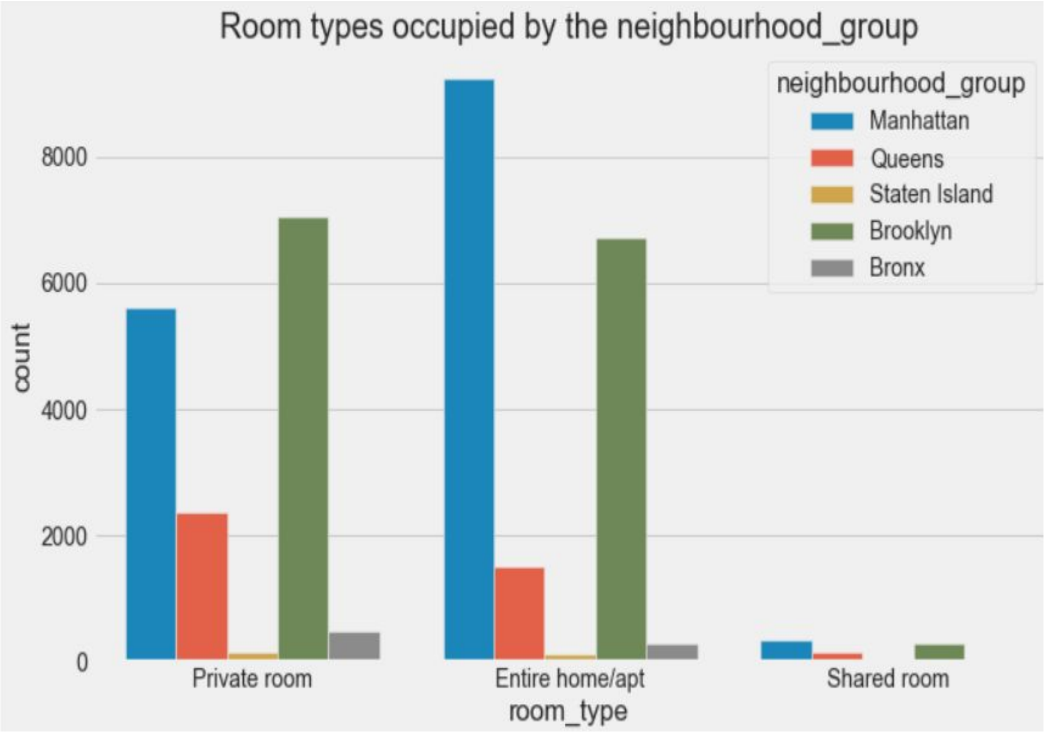
Area Wise Distribution of Price

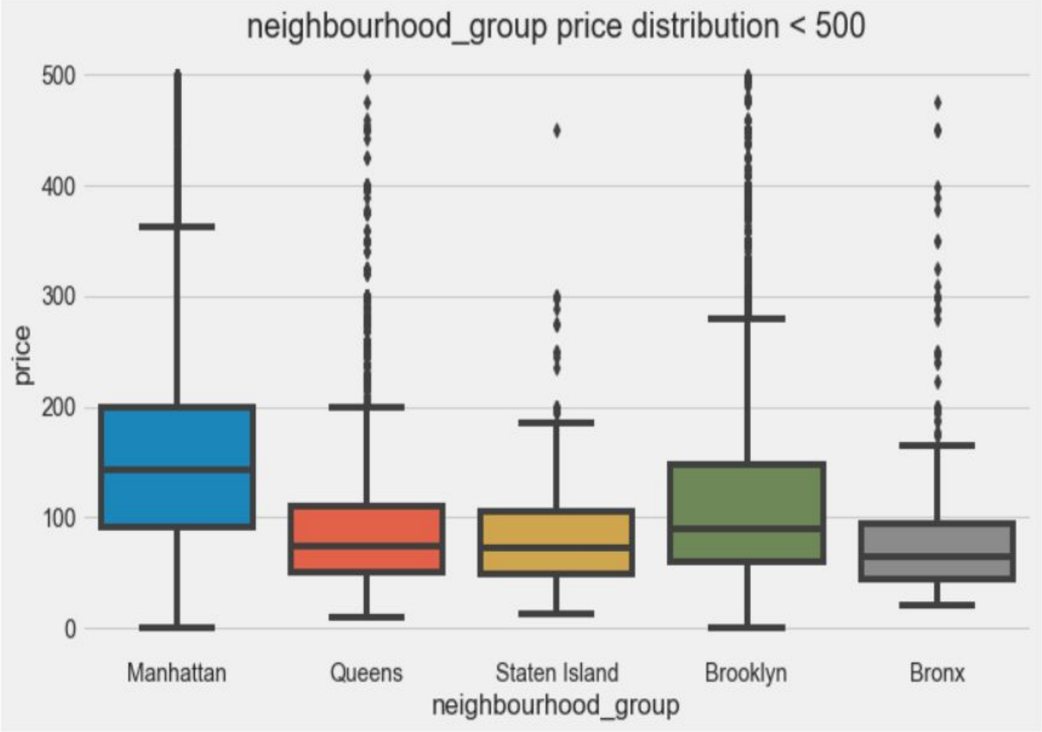




Prices in more popular areas like Manhattan and Brooklyn are more. The properties in these areas are also available for lesser number of days.

Room Type in Each Neighbourhood

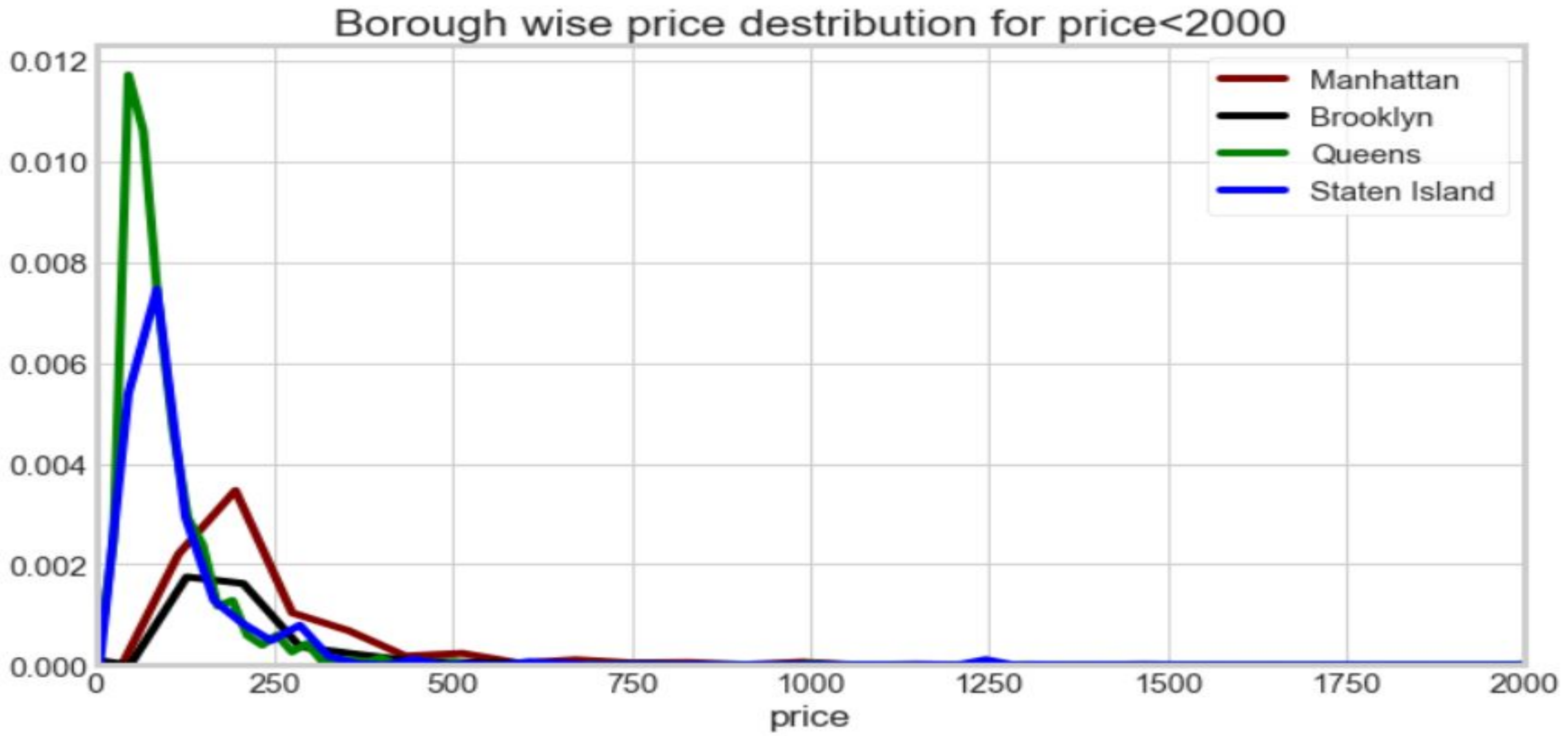




From the above graph we can deduce that people prefer Private rooms and Entire-house over Shared rooms

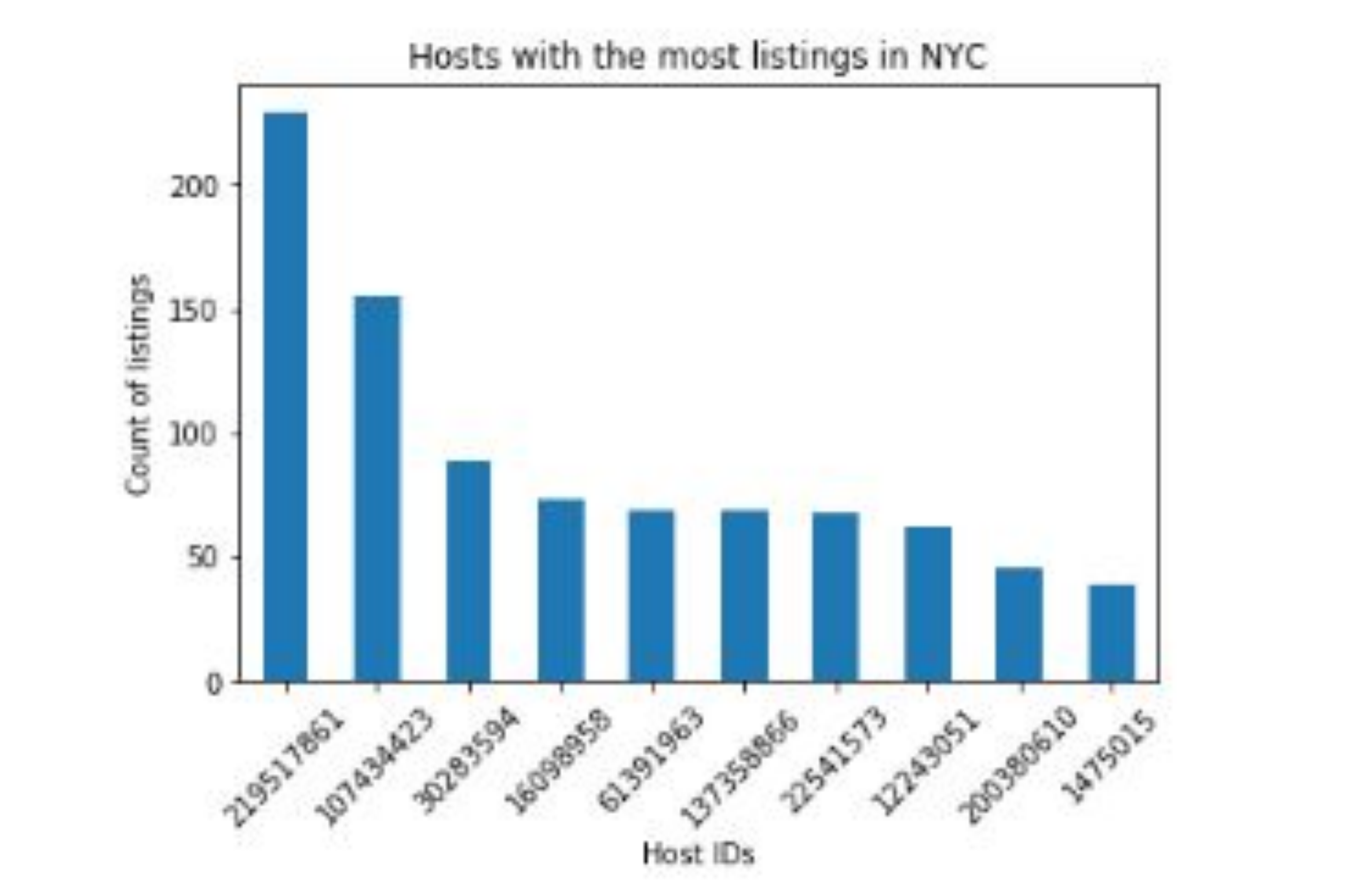
Manhattan has the most expensive accomodation followed by Brooklyn

Distribution of Prices



Area wise distribution of price

Top 10 Hosts

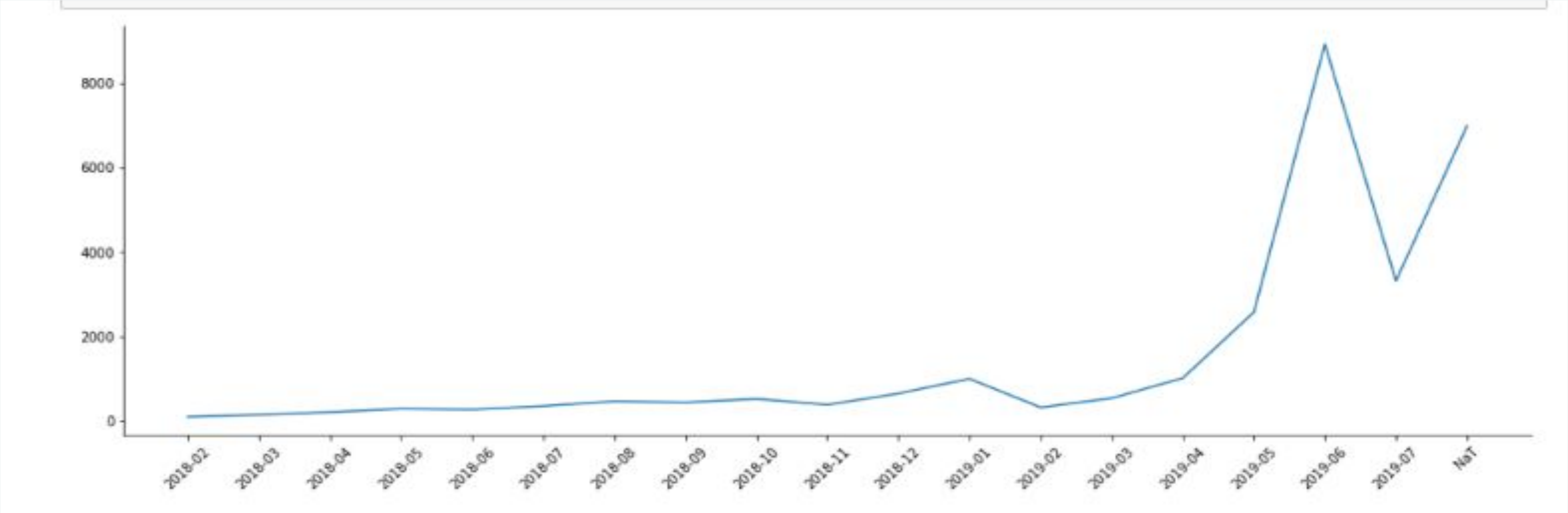


9 / 10 Host ID’s are situated in Manhattan

9 / 10 Host ID’s belong to the ‘Entire Home’ Room type

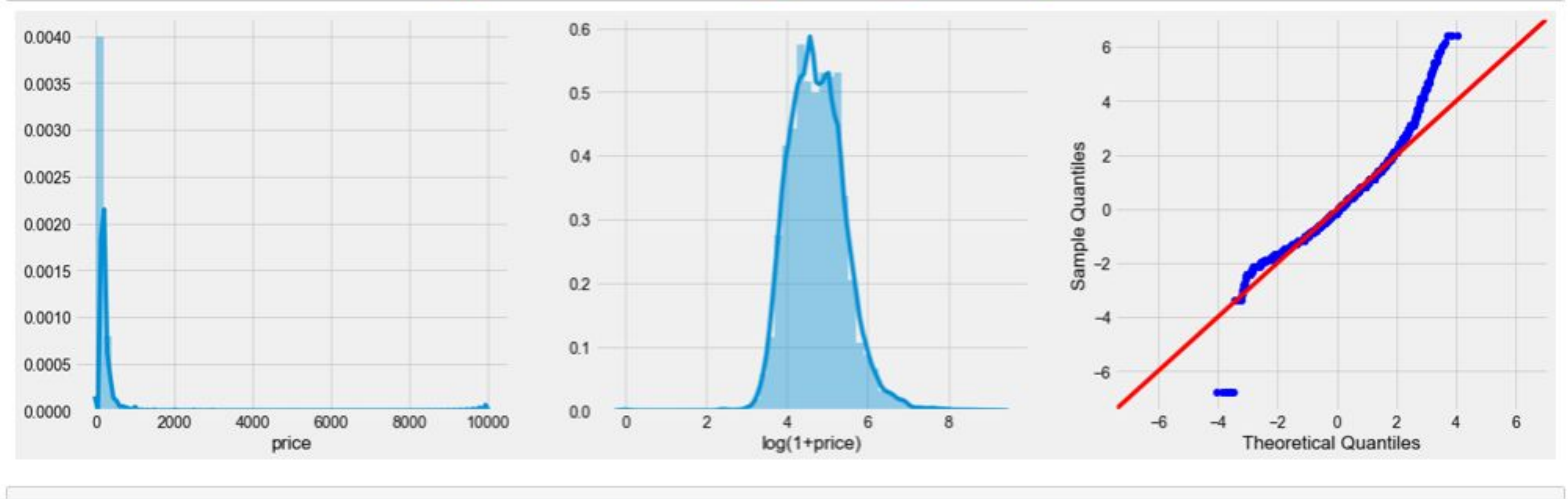
5 / 10 Host ID’s have availability of 300+ days out of 365days

Last Reviews of listings



Plotted the last review in months and saw that most of the properties were given their last review in May 2019.

Skewness in Price



Log transformed price for some of models to remove skewness

Models Used

Models	Test RMSE	Platform Result
Linear Regression	234.97	222.36
Gradient Boosting Regressor	143.058	217.11
Random Forest Regressor	139.897	220.12
Random Forest with RandomizedSearch CV	223.14	211.1
Random Forest with Feature Selection	232.51	233.86
Lasso	109.370	234.82
Ridge	135.648	222.27
Bagging	138.513	220.69
Adaboost	209.96	233.24
XG Boost	174.837	212.37
Random Forest with Feature Selection and Randomized Search CV	223.56	213.07

Business Insights

1. Manhattan is the most expensive among all areas whereas Bronx is the least priced one.

2. People generally go for Entire houses or private rooms over shared rooms hence their prices could accordingly be changed depending upon the demand.

Additional Data for better Insights..

If the following data was also provided we would’ve been able to explore the data in more depth:

1. **Property Specifications** : area, room amenities, parking space, society amenities, smoking area, no. of beds, ppl capacity)
2. **Review category** : good/bad
3. **Host data** : personal details , profession
4. **Dates**: Dates on which weekly prices are recorded and availability dates as on availability recorded.

More Business problems that could be solved..

1. Customer(home owner) attrition
2. Leverage reviews
3. Dynamic pricing
4. How to on-board new home owners for the company

THANK YOU