Airbnb New York City Listing Price Suggestion

Members:

Aliaa Shafiqah Binti Jamaludin

Nursyakirah Dhaniyah Binti Abdul Munir

Yap Min Ling

Click to go to project repo

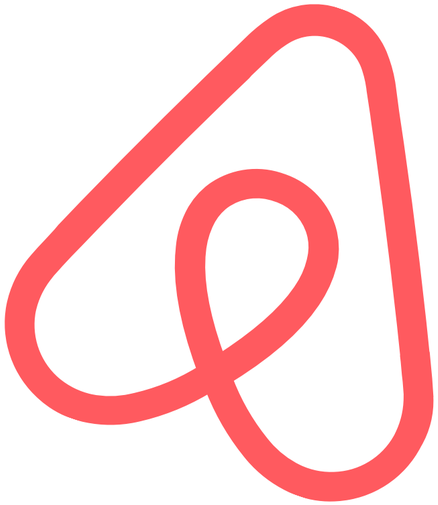
Content Overview

- Project Overview
- Business & Data Understanding
- EDA
- Reasoning To Use Current Approach
- ML Model Results
- Key Insights For Business
- Next Action Points For Business
- Key Points of Improvment/Focus

Project Overview

Airbnb, Inc. is one of the tourism industry which is an American company that operates an online marketplace for lodging, primarily homestays for vacation rentals, and tourism activities. Based in San Francisco, California, the platform is accessible via website and mobile app.

A host owns multiple properties across different neighbourhoods in New York City. She wants to list her property up on Airbnb and wants to know what is the suitable price for each respective property.



Business & Data Understanding

Business Objective

- To determine a suitable listing price for each property for the host

Data Objective

- Compare prices by neighbourhood by looking at Airbnb on a New York City map
- Analyze whether there is a difference in price by room type

Evaluation Criteria

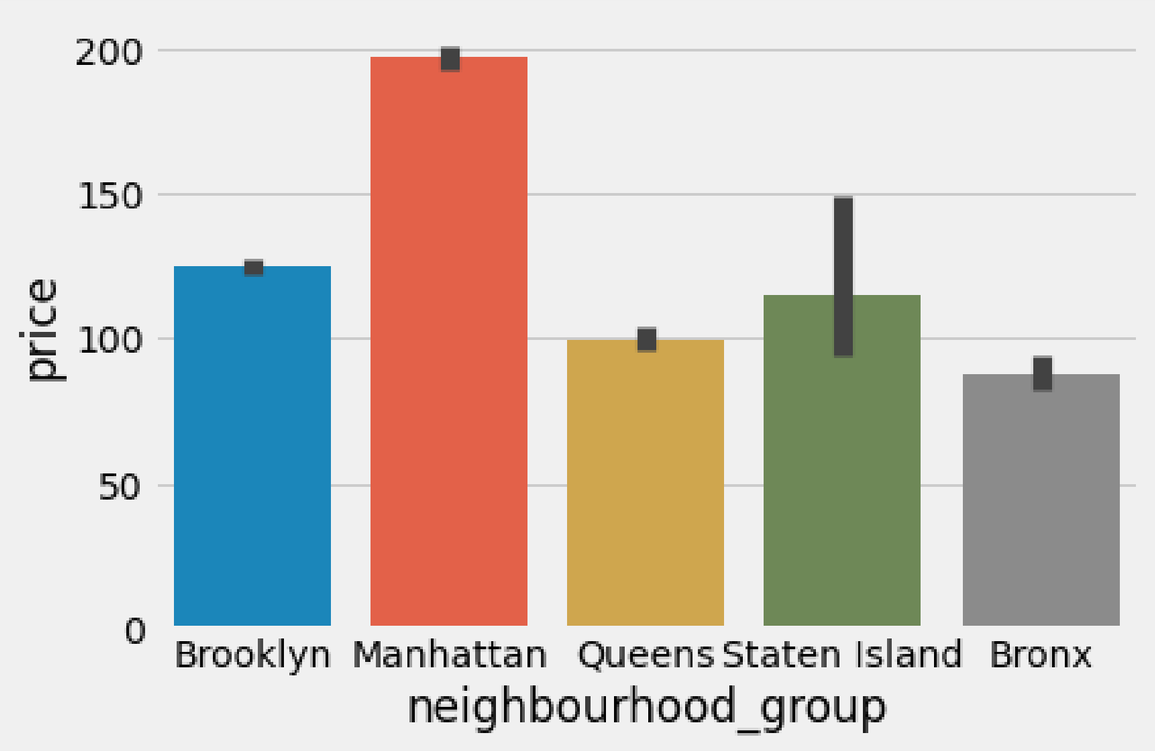
- The host uses the recommended listing price

Dataset (New York City Airbnb Open Data)

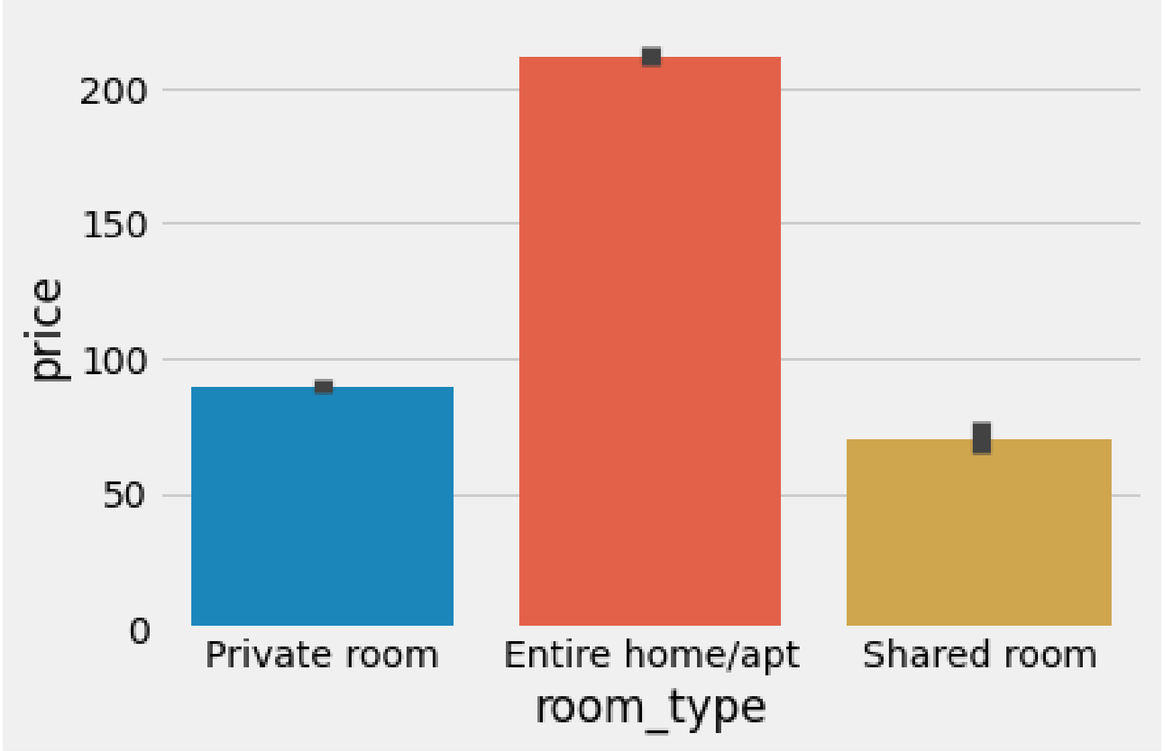
- id: listing ID
- name: name of the listing
- host_id: host ID
- host_name: name of the host
- neighbourhood_group: location
- neighbourhood: area
- latitude: latitude coordinates
- longitude: longitude coordinates
- room_type: listing space type
- price: price in dollars
- minimum_nights: amount of nights minimum
- number_of_reviews: number of reviews
- last_review: latest review
- reviews_per_month: number of reviews per month
- calculated_host_listings_count: amount of listing per host
- availability_365: number of days when listing is available for booking

	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
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	9	2018-10-19	0.21	6	365
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	45	2019-05-21	0.38	2	355
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150	3	0	NaN	NaN	1	365
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89	1	270	2019-07-05	4.64	1	194
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	10	9	2018-11-19	0.10	1	0

EDA

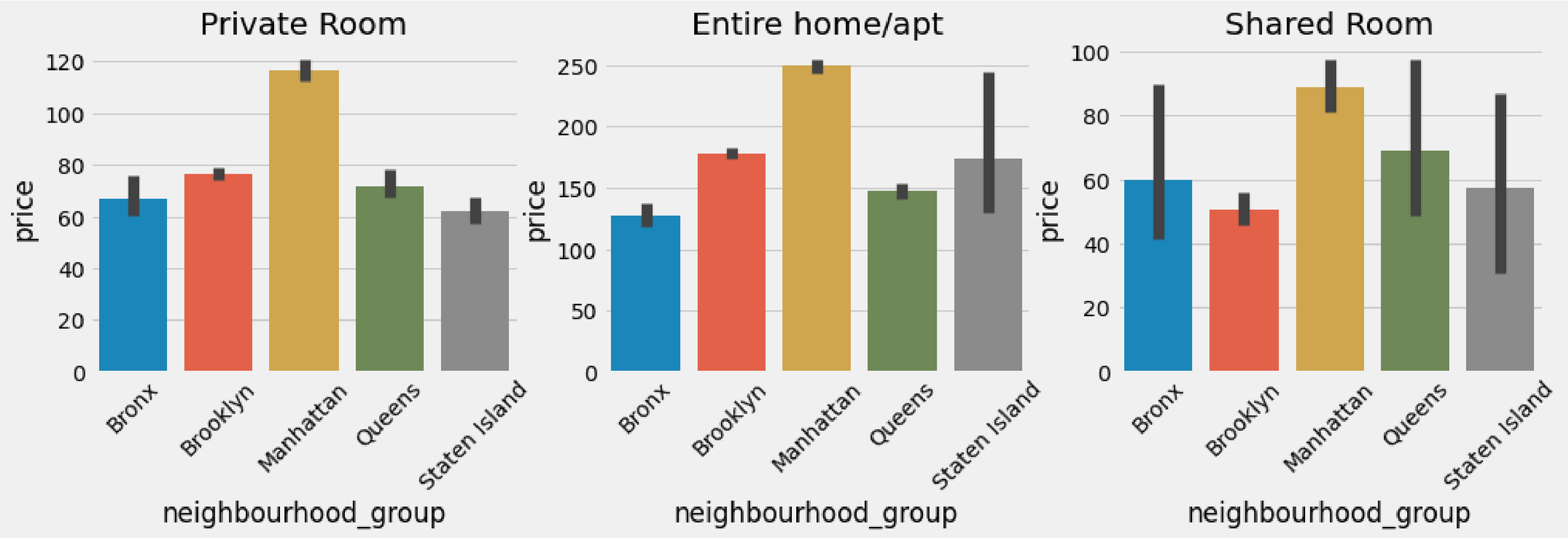


For the neighborhood group Manhattan has the highest mean price in the New York City Airbnb listing.



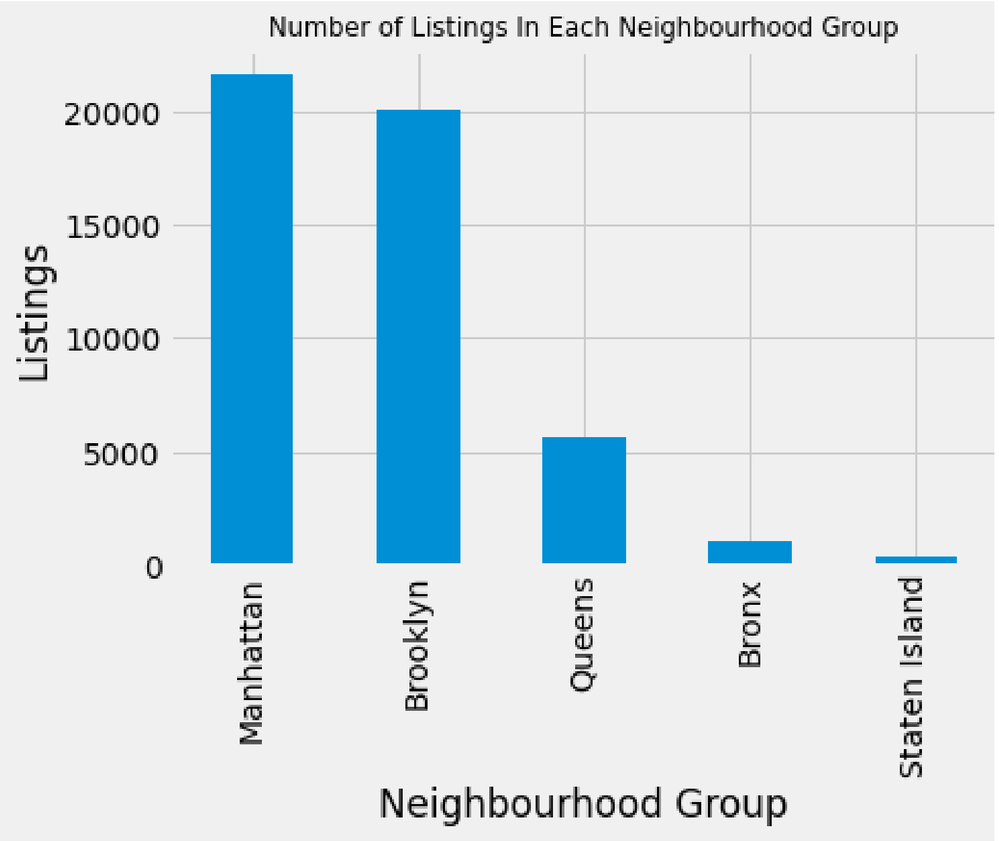
For the room type, entire home/apt has the highest mean price in the New York City Airbnb listing.

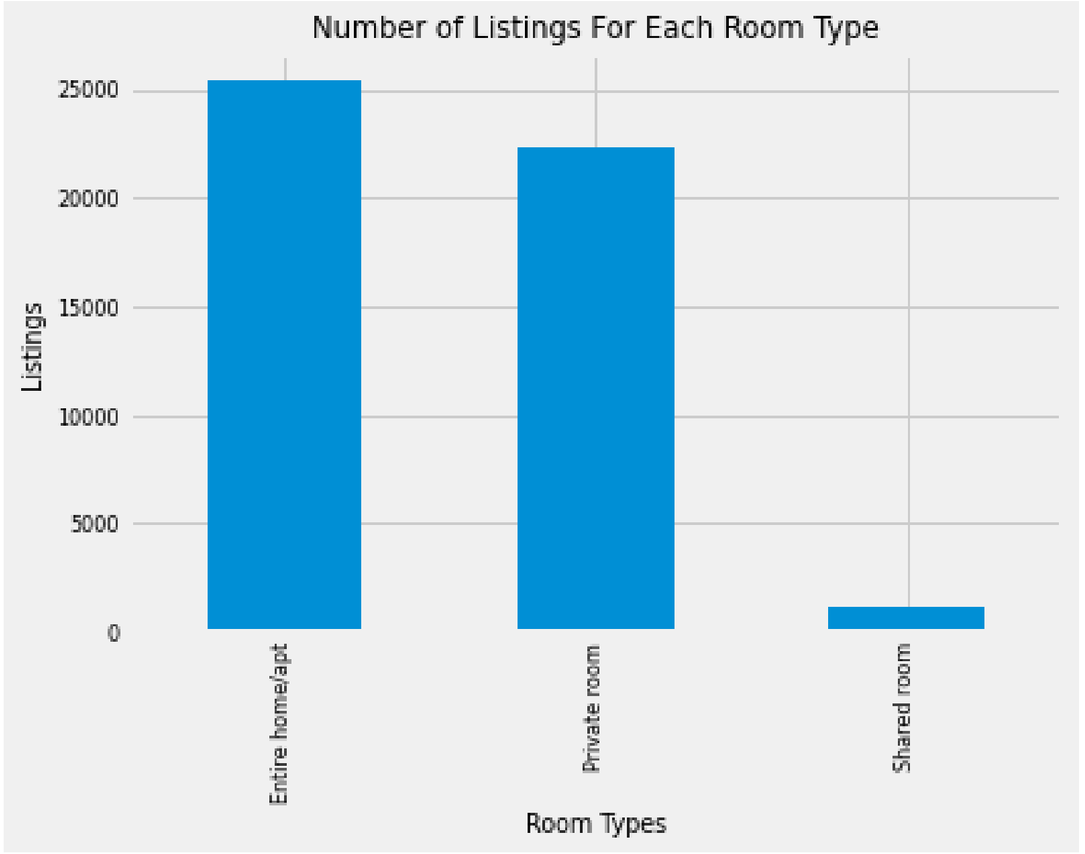
EDA



- Manhattan has the highest mean price among other neighbourhood for all room types
- Brooklyn has the least mean price for the room type - 'Shared Room'
- Staten Island has the least mean price for the room type - 'Private Room'
- Bronx has the least mean price for the room type -'Entire home/apt'

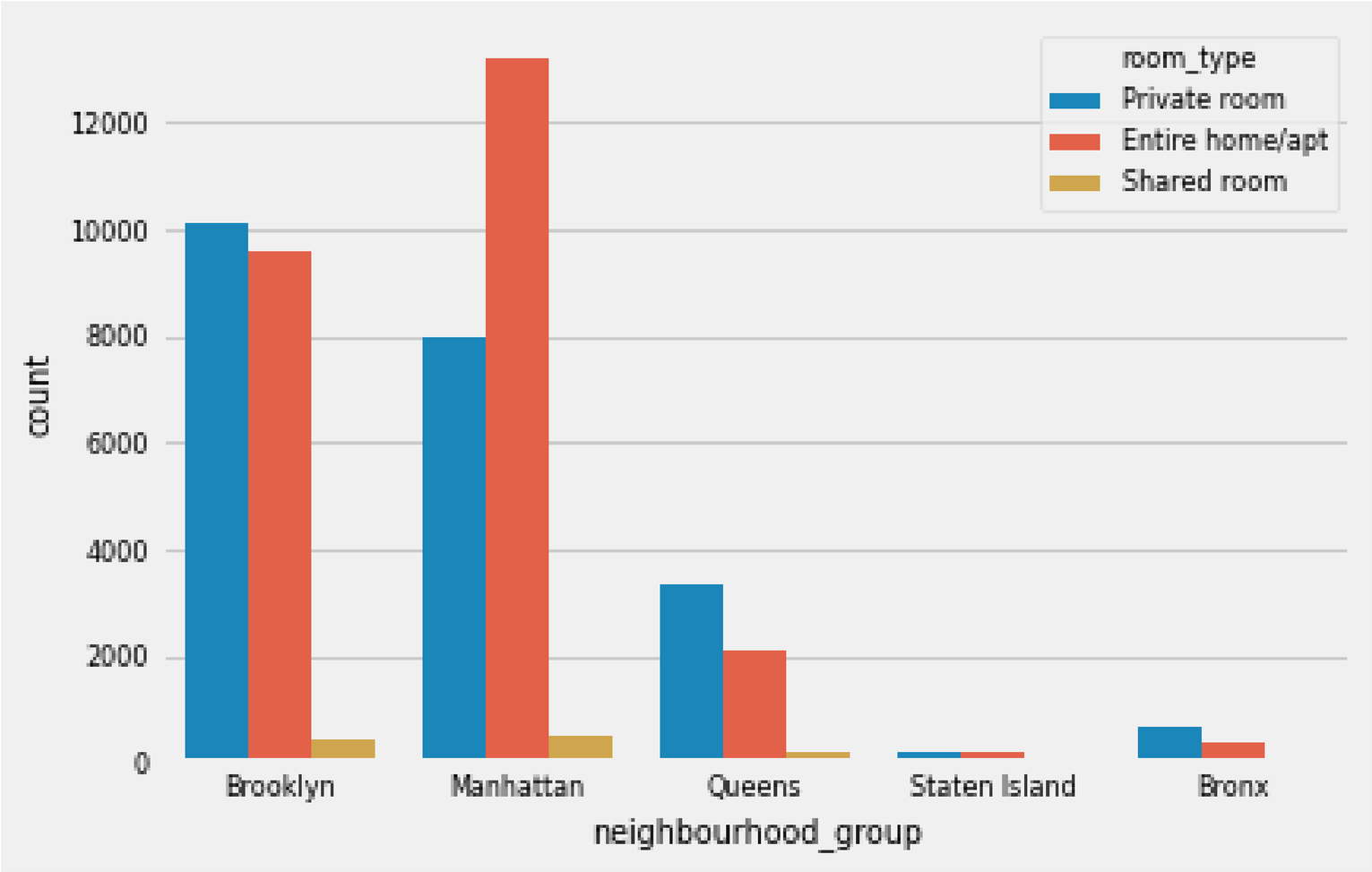
EDA





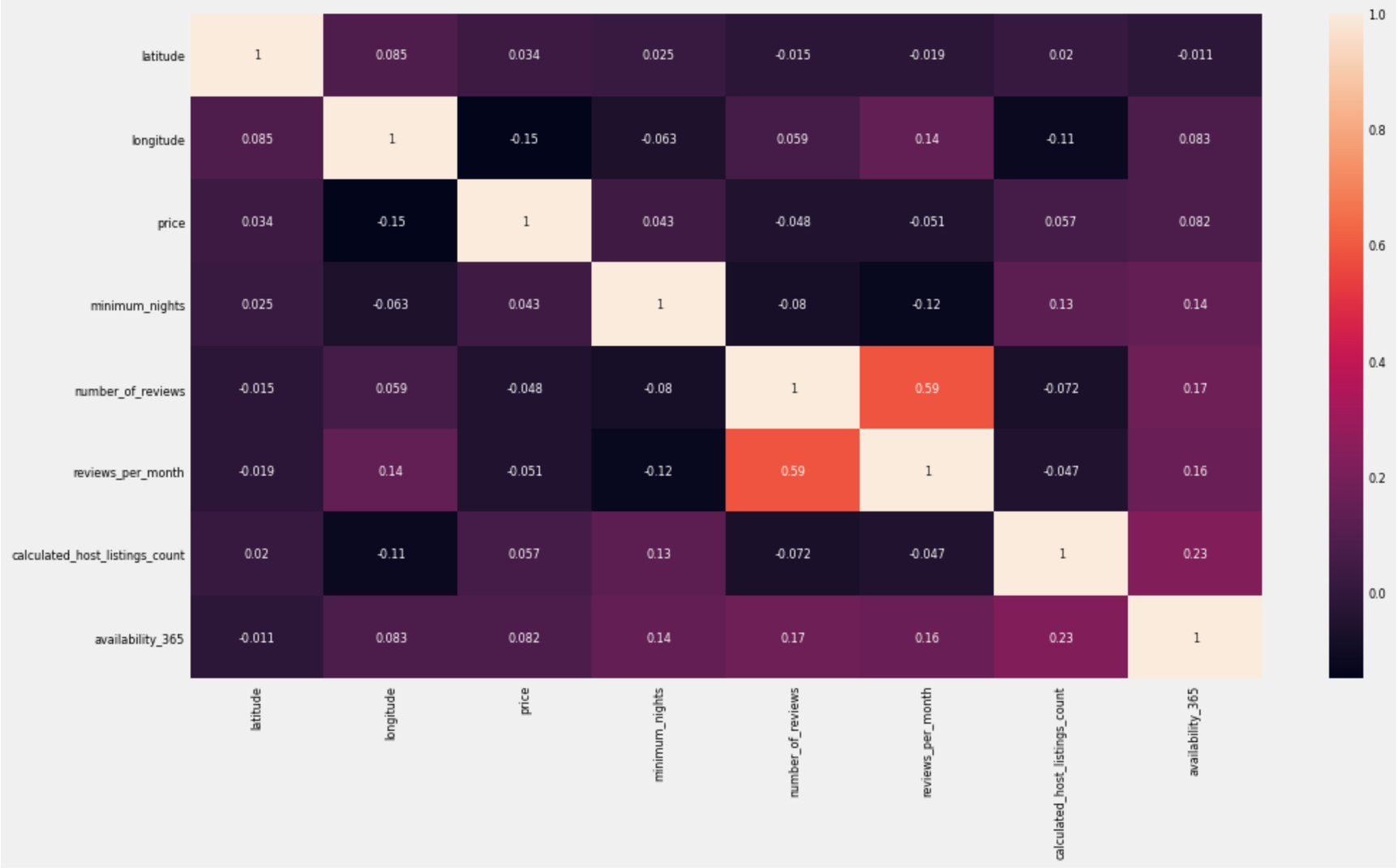
Comparing number of listings based on neighbourhood group and room type

EDA



Most of Manhattan listings constitute mostly of the room type - 'Entire home/apt'. Other neighbourhoods have a relatively equal listings between 'Private room' and 'Entire home/apt'.

EDA


latitude
longitude
price
minimum_nights
number_of_reviews
reviews_per_month
calculated_host_listings_count
availability_365
1 0.085 0.034 0.025 -0.015 -0.019 0.02 -0.011
0.085 1 -0.15 -0.063 0.059 0.14 -0.11 0.083
0.034 -0.15 1 0.043 -0.048 -0.051 0.057 0.082
0.025 -0.063 0.043 1 -0.08 -0.12 0.13 0.14
-0.015 0.059 -0.048 -0.08 1 0.59 -0.072 0.17
-0.019 0.14 -0.051 -0.12 0.59 1 -0.047 0.16
0.02 -0.11 0.057 0.13 -0.072 -0.047 1 0.23
-0.011 0.083 0.082 0.14 0.17 0.16 0.23 1
1.0
0.8
0.6
0.4
0.2
0.0

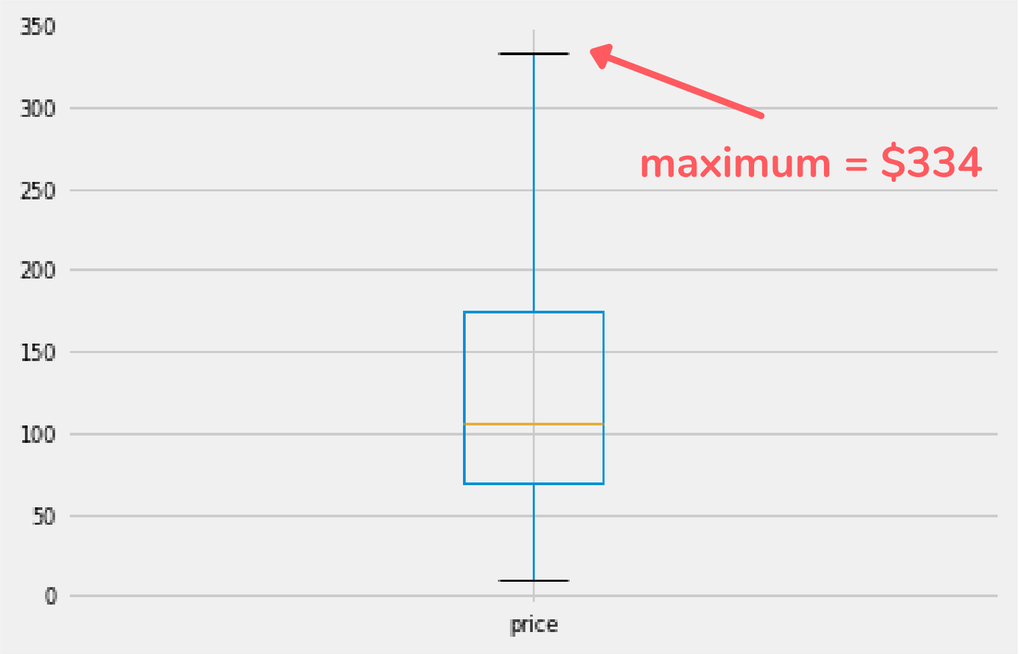
CORRELATION

	latitude	longitude	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365
latitude	1.000000	0.084830	0.033899	0.024893	-0.015357	-0.018719	0.019518	-0.010942
longitude	0.084830	1.000000	-0.149954	-0.062772	0.059015	0.138435	-0.114715	0.082669
price	0.033899	-0.149954	1.000000	0.042805	-0.047926	-0.050531	0.057462	0.081847
minimum_nights	0.024893	-0.062772	0.042805	1.000000	-0.080080	-0.124870	0.127962	0.144275
number_of_reviews	-0.015357	0.059015	-0.047926	-0.080080	1.000000	0.589311	-0.072385	0.171975
reviews_per_month	-0.018719	0.138435	-0.050531	-0.124870	0.589311	1.000000	-0.047322	0.163664
calculated_host_listings_count	0.019518	-0.114715	0.057462	0.127962	-0.072385	-0.047322	1.000000	0.225712
availability_365	-0.010942	0.082669	0.081847	0.144275	0.171975	0.163664	0.225712	1.000000

Reasoning For Using The Current Approach

The models are trained using 3 different sets of data:

- data with outliers
- data below maximum (< $334)
- data above maximum (> $334)



Models used:

- Linear regression
- Decision tree
- Random forest regression

The performance of the models are evaluated through R2, MSE and RMSE, then compared.

ML Model Results

	R2 Score			MSE			RMSE		
Datasets	**Linear Regression**	**Decision Tree**	**Random Forest Regression**	**Linear Regression**	**Decision Tree**	**Random Forest Regression**	**Linear Regression**	**Decision Tree**	**Random Forest Regression**
All	-5.08E+11	0.24	**0.59**	4.66E+10	0.07	**0.04**	215919.75	0.26	**0.19**
Below Maximum	-5.77E+15	0.29	**0.63**	3.61E+14	0.04	**0.02**	1.90E+07	0.21	**0.15**
Above Maximum	-5.44E+18	-1.10	**0.11**	2.68E+17	0.10	**0.04**	5.18E+08	0.32	**0.21**

Key Insights for Business

- Manhattan has the highest listing price
- Other neighbourhood groups listing price ranking differs across different room types
- From all the models that we used, Random Forest Regression Model provides the highest accuracy score
- Among the 3 datasets, the datasets below maximum produces the highest accuracy score (R2 score = 0.63259)

Next Action Points for Business Teams

- Analyze based on just the location (latitude, longitude, neighbourhood groups, neighbourhood)
- Try cross-validation to see if accuracy improves especially for neighbourhood groups with much lower observations (Bronx, Staten Island)

Key Points of Improvment/Focus

- Find the best fit data modeling with high accuracy score.
- Sufficient observations / stronger features with higher correlation to price to support the Machine Learning model training

Thank You