Capstone Project

Project Title – Airbnb Booking Analysis

Team Members

Ankur Bhattacharjee
Bhabakriskna Talukdar
Mayank Tiwari
Md Suhel Ansari
Pratheek T M

AirBnb Booking analysis

1. Understanding the Airbnb dataset.
2. Removing impurities and handing null values.
3. Exploratory Data Analysis:
   - Univariate analysis
   - Bivariate & multivariate analysis.
4. Hypothesis evaluation using statistical test.
5. Conclusion.

Understanding AirBnb

Airbnb, Inc. is an American company that operates an online marketplace for lodging, primarily homestays for vacation rentals, and tourism activities. Airbnb does not own any of the listed properties; instead, it profits by receiving commission from each booking.

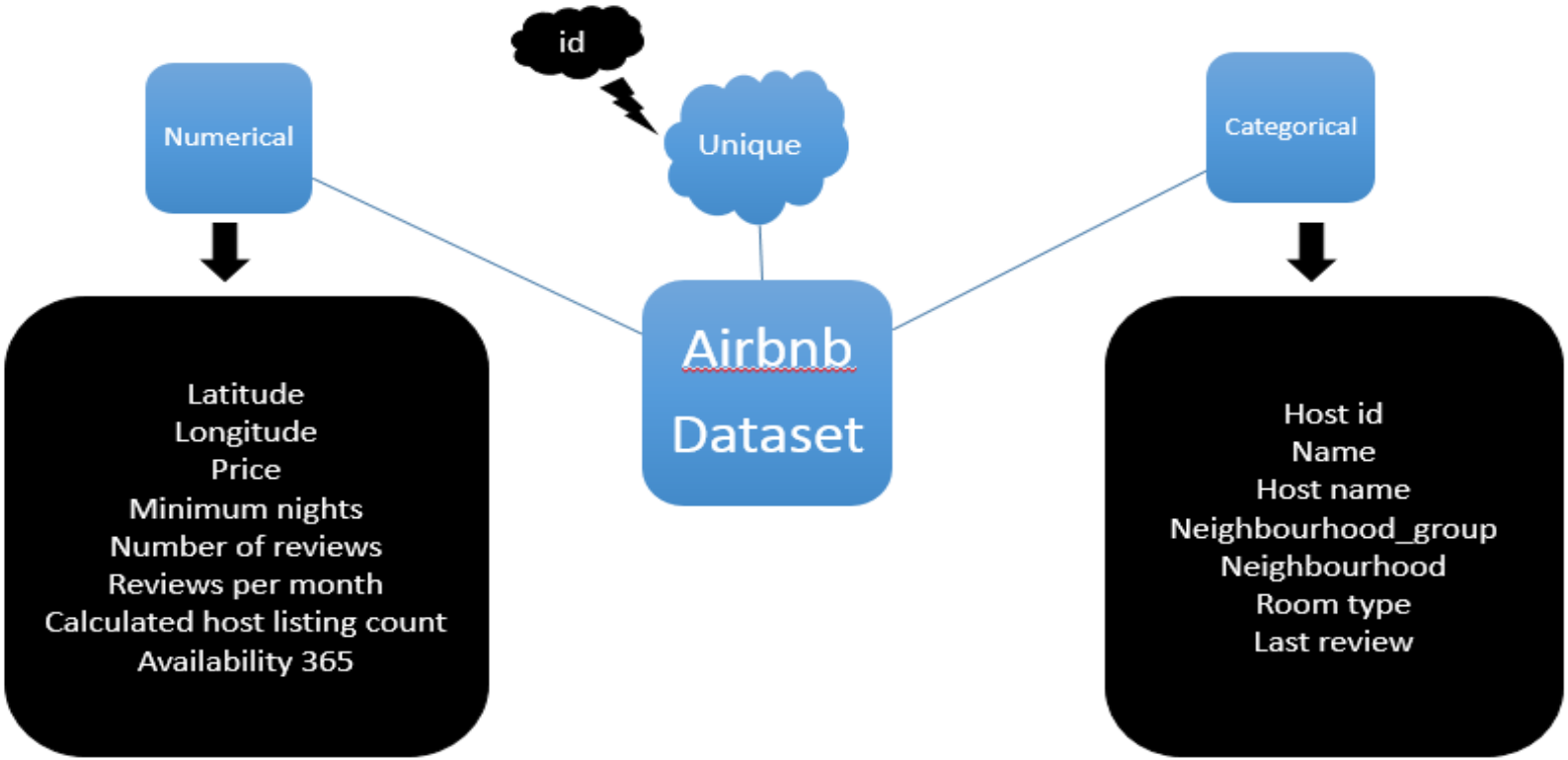
The idea of this project is to analyze the Airbnb data and to get insights of the data to discover the key understanding of their model.

Using this data we can analyze 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.

Problem Statements:

1. What can we learn about different hosts and areas?
2. What can we learn from the predictions?(ex: locations, prices, reviews, etc.)
3. Which hosts are the busiest and why?
4. Is there any noticeable difference of traffic among different areas and what could be the reason for it.

Data structure:



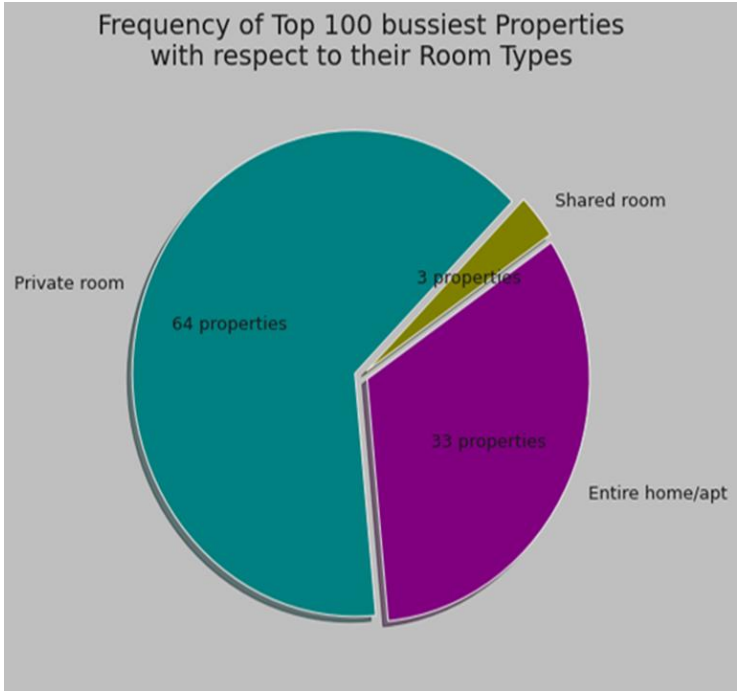
Data description:

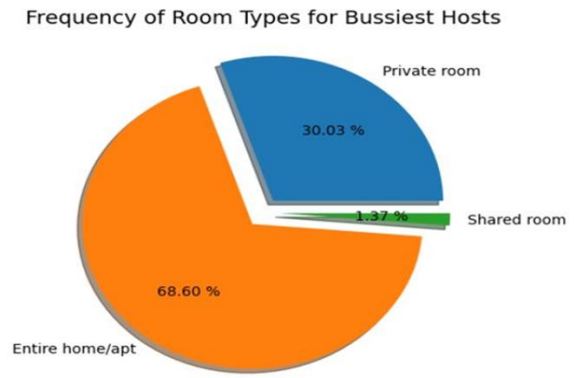
- We have 48895 rows and 16 columns.
- Here, we have 3 floating data types, 7 integer types and 6 object data types.
- We found that there are some null values present in some columns.
- ‘reviews_per_month’ and ‘last_review’ columns were having 20.56 % of the data null.
- ‘name’ and ‘host_name’ had very minimal amount of null values.
- Rest of the columns had no null values.

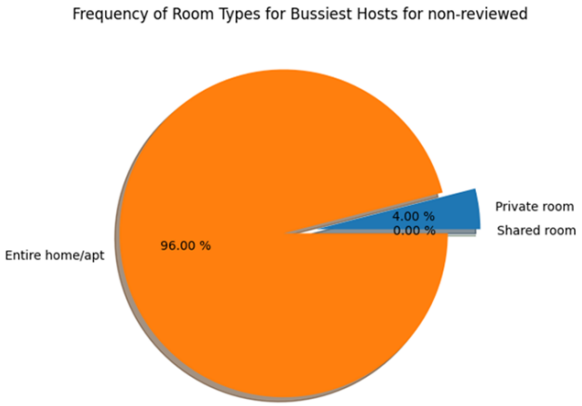
Univariate Analysis

	latitude	longitude	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365
count	48895.000000	48895.000000	48895.000000	48895.000000	48895.000000	38843.000000	48895.000000	48895.000000
mean	40.728949	-73.952170	152.720687	7.029962	23.274466	1.373221	7.143982	112.781327
std	0.054530	0.046157	240.154170	20.510550	44.550582	1.680442	32.952519	131.622289
min	40.499790	-74.244420	0.000000	1.000000	0.000000	0.010000	1.000000	0.000000
25%	40.690100	-73.983070	69.000000	1.000000	1.000000	0.190000	1.000000	0.000000
50%	40.723070	-73.955680	106.000000	3.000000	5.000000	0.720000	1.000000	45.000000
75%	40.763115	-73.936275	175.000000	5.000000	24.000000	2.020000	2.000000	227.000000
max	40.913060	-73.712990	10000.000000	1250.000000	629.000000	58.500000	327.000000	365.000000

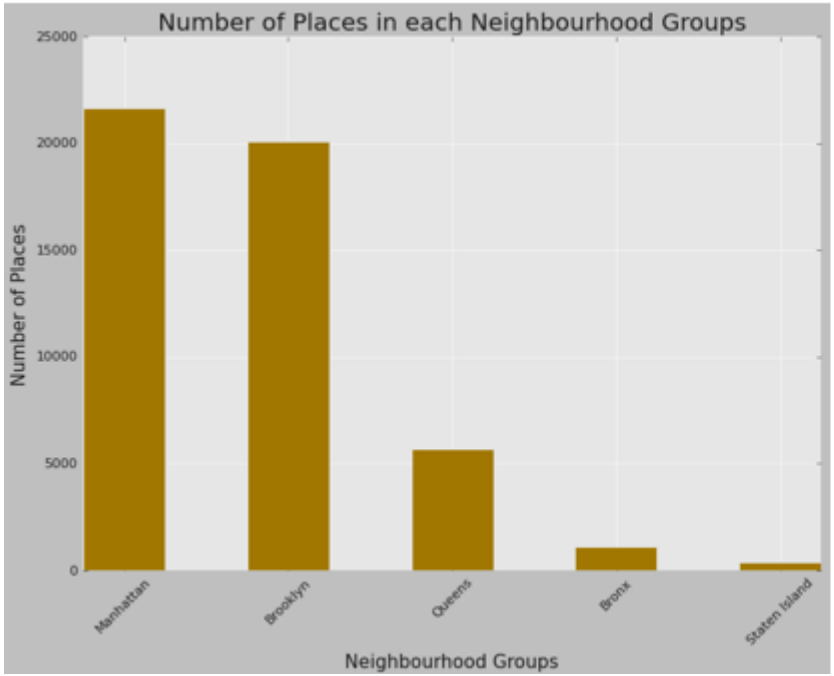
Univariate Analysis

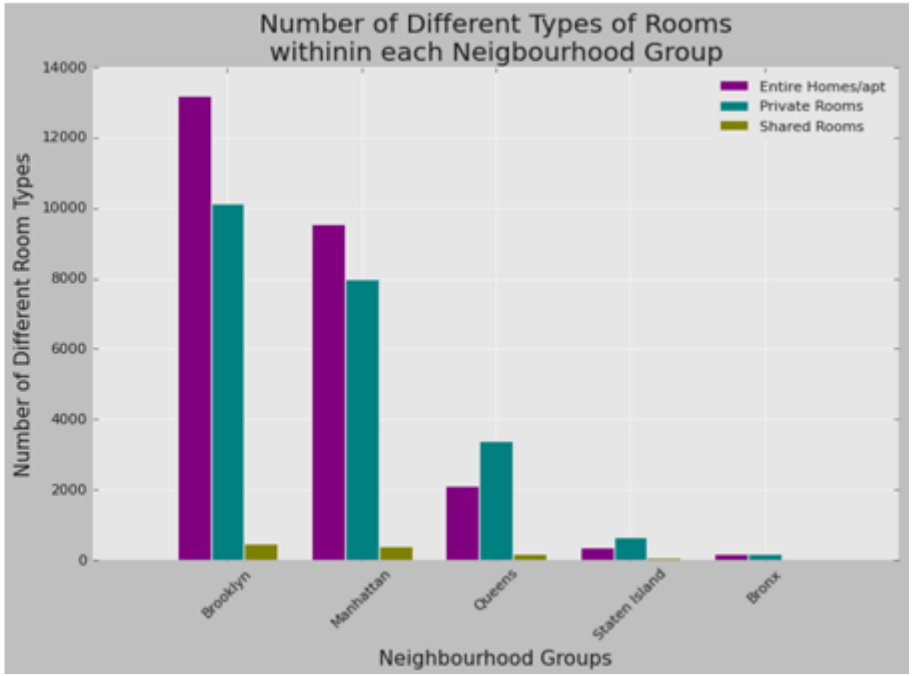




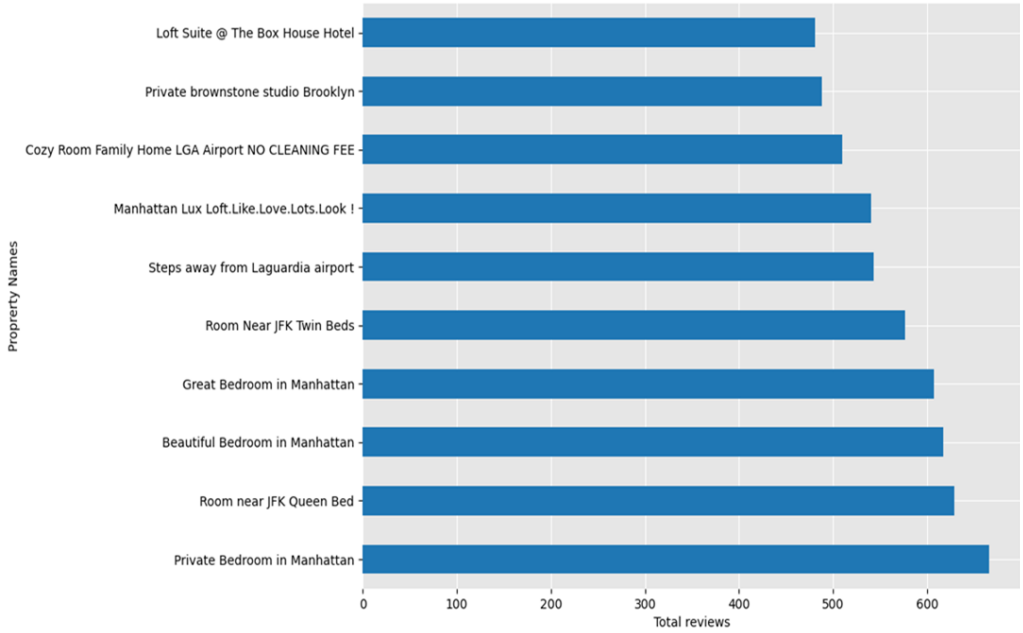


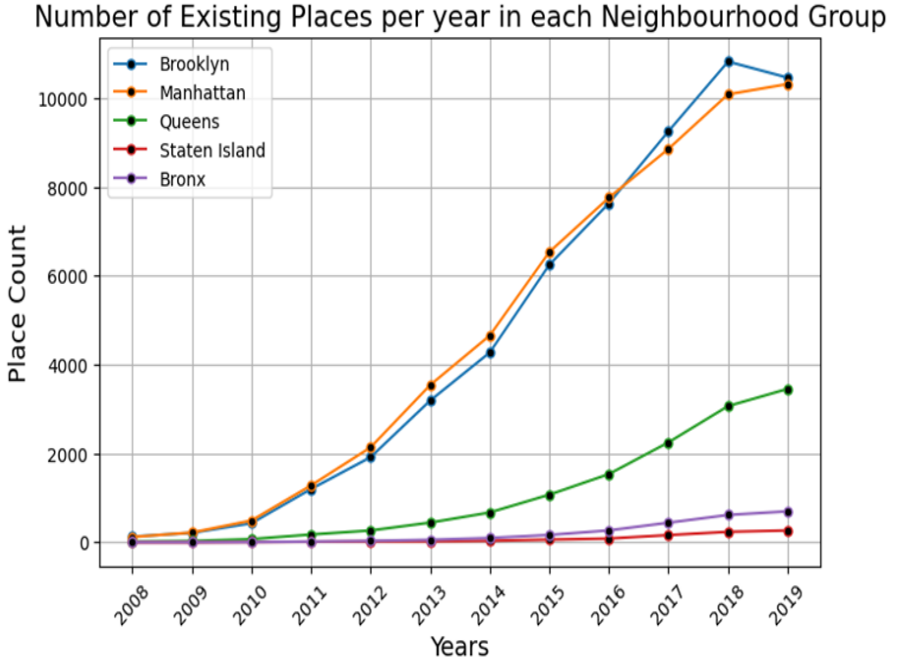
Bivariate Analysis



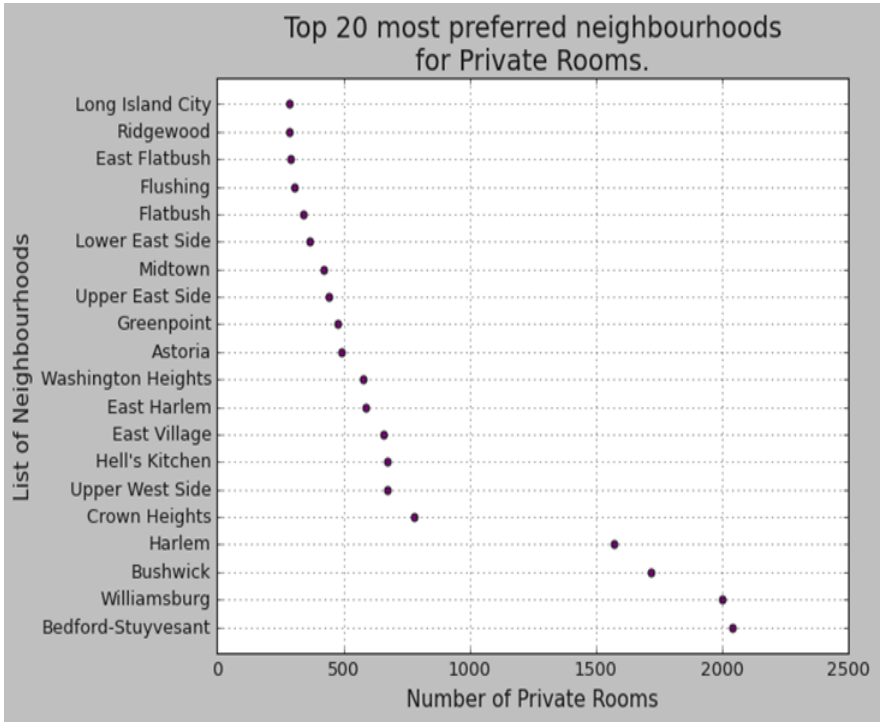


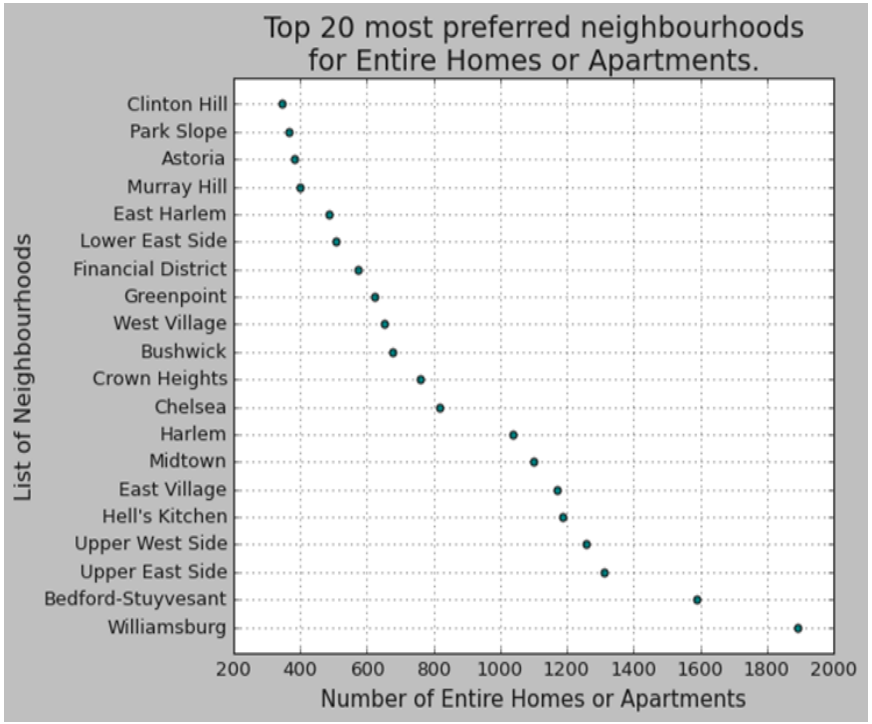
Bivariate Analysis



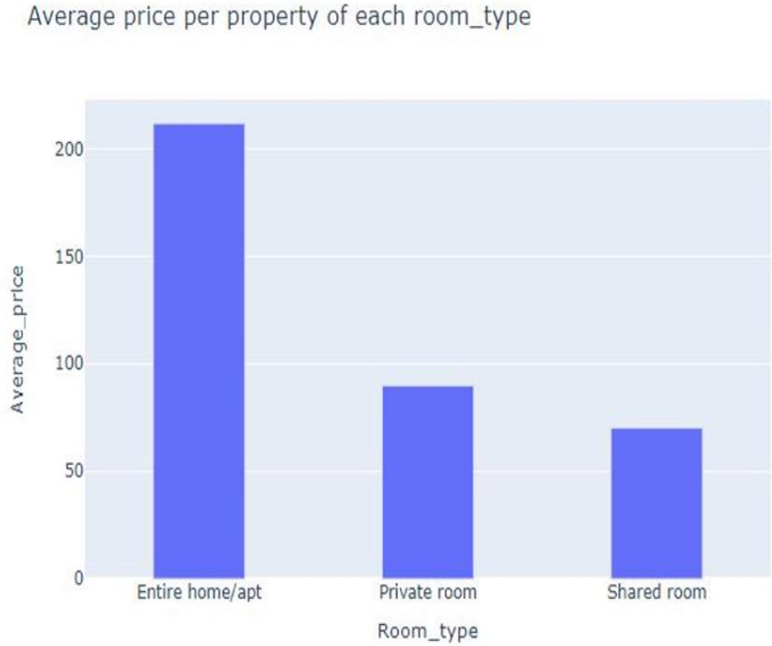


Bivariate Analysis

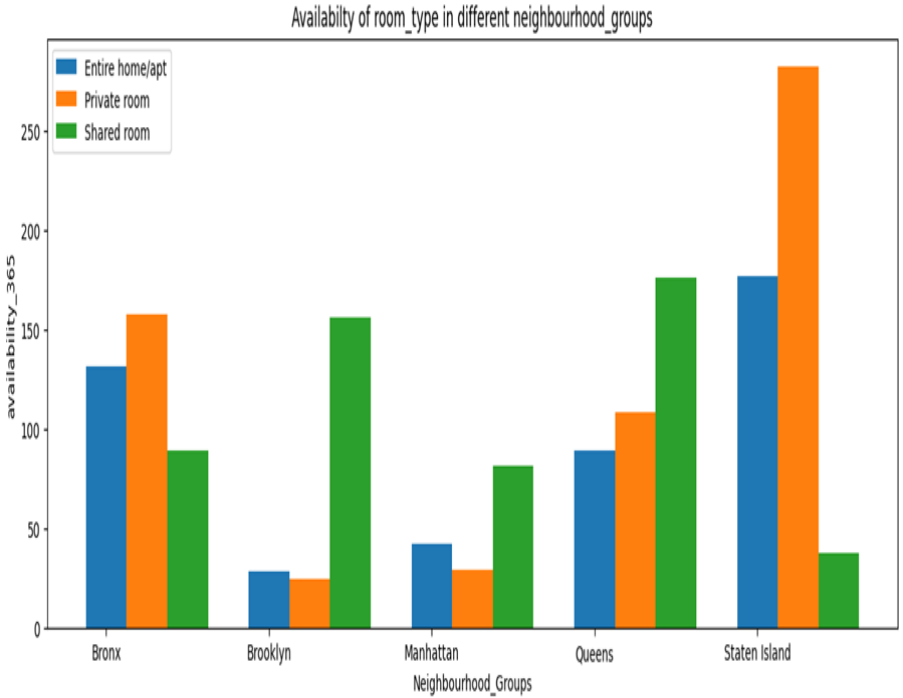


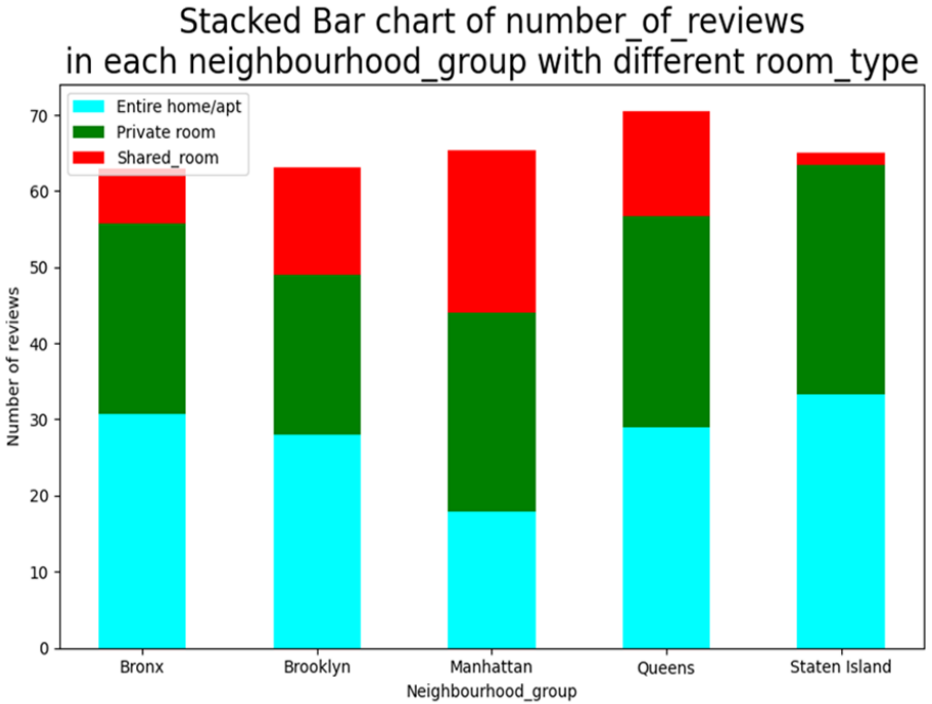


Bivariate Analysis

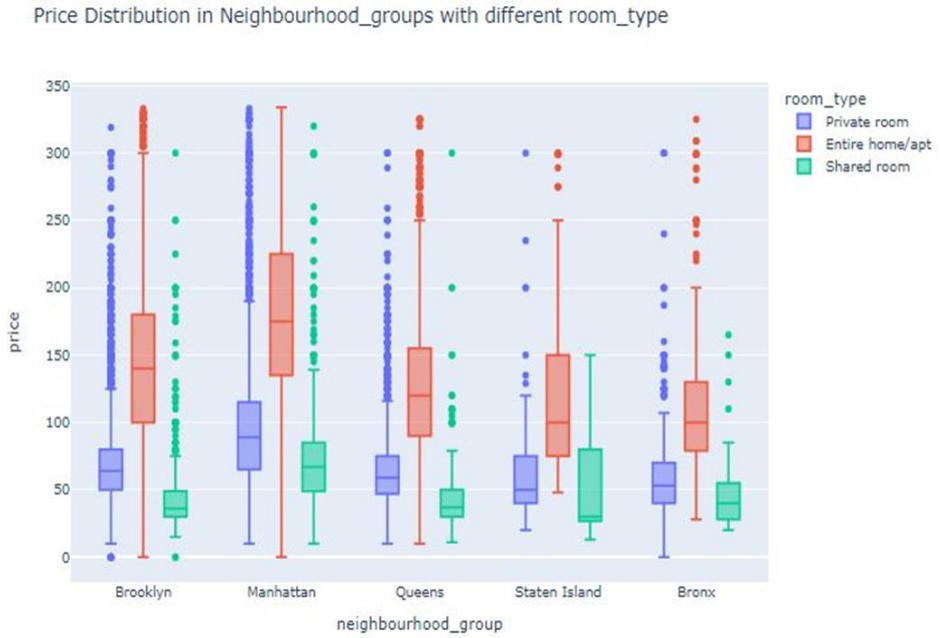


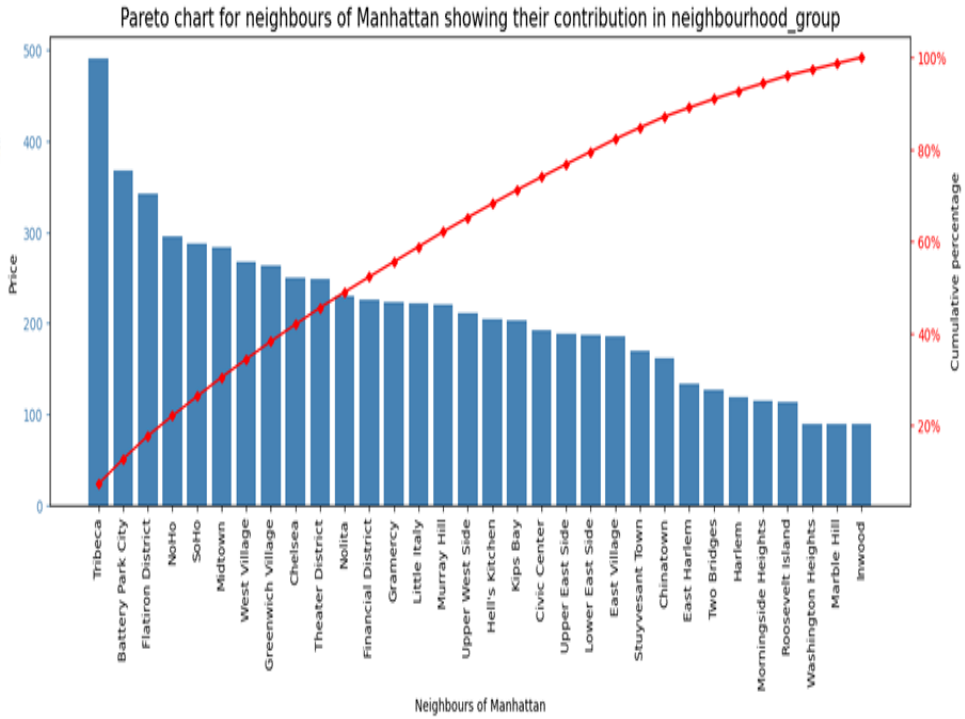
Bivariate Analysis



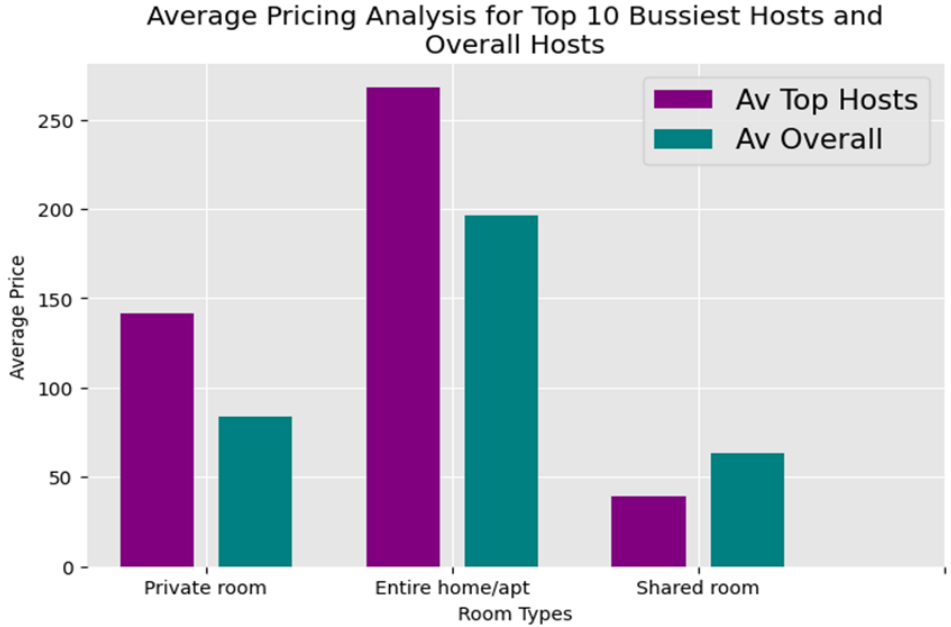


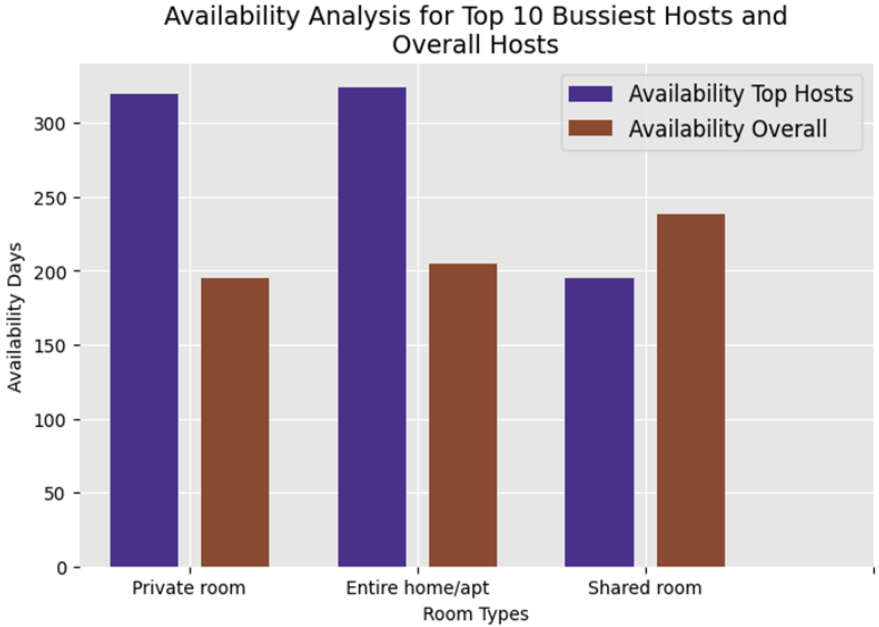
Bivariate Analysis



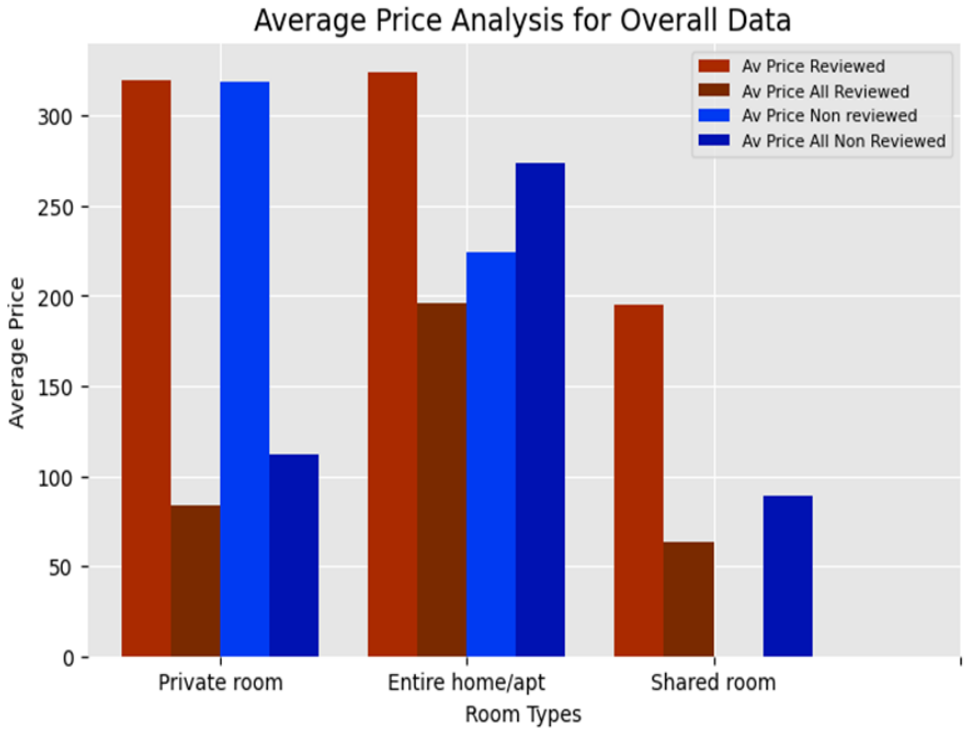


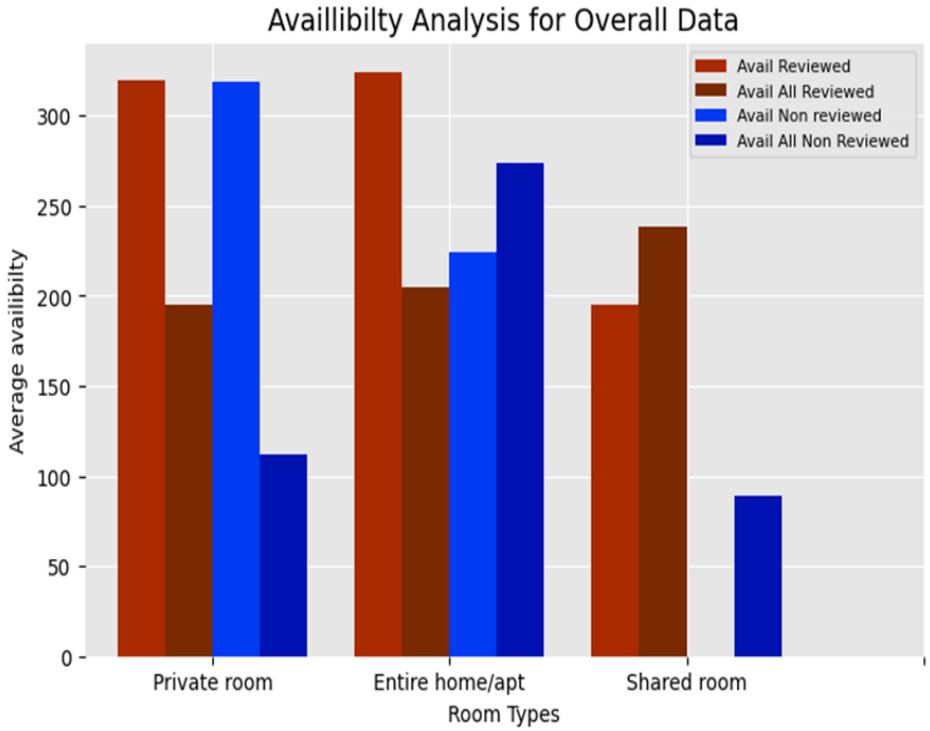
Bivariate Analysis



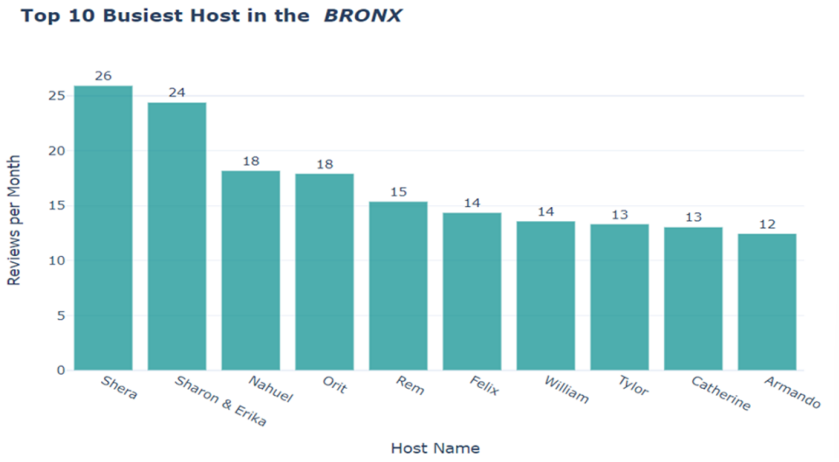


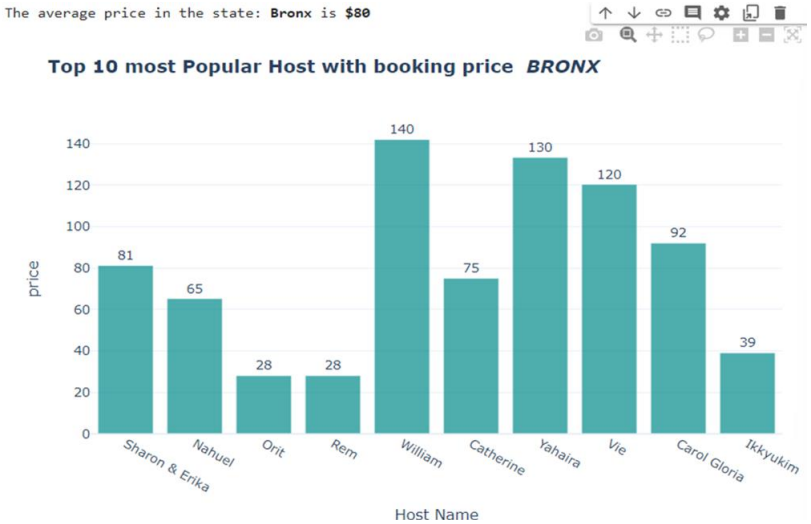
Bivariate Analysis

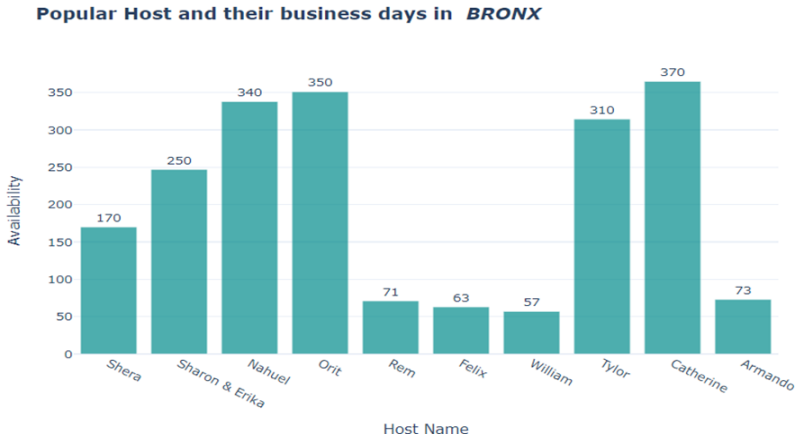


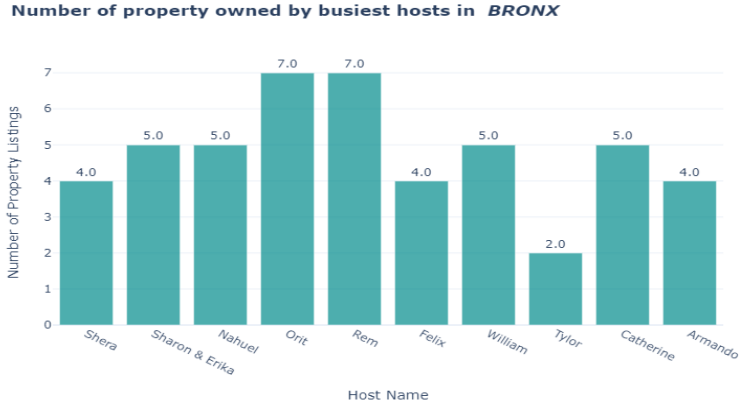


Bivariate Analysis- Bronx





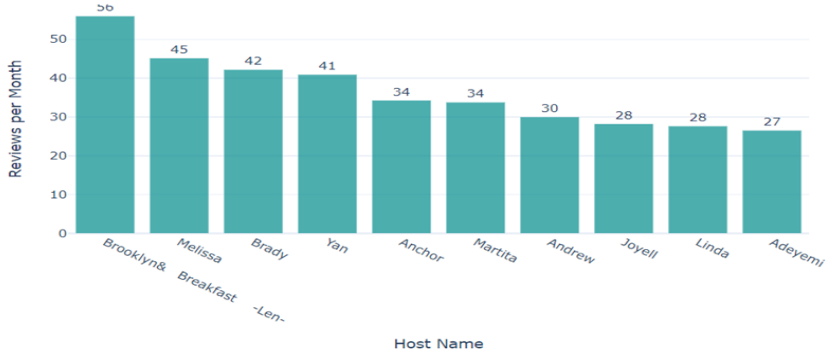




Bivariate Analysis- Brooklyn

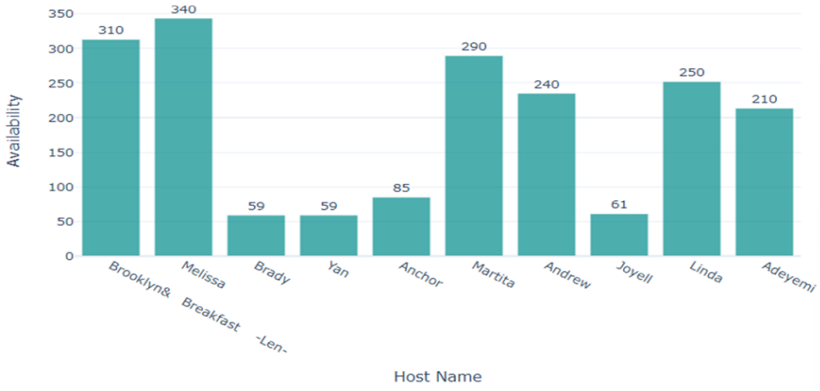
The average price in the state: **Brooklyn** is **$121**

Top 10 Busiest Host in the *BROOKLYN*



Top 10 most Popular Host with booking price *BROOKLYN*

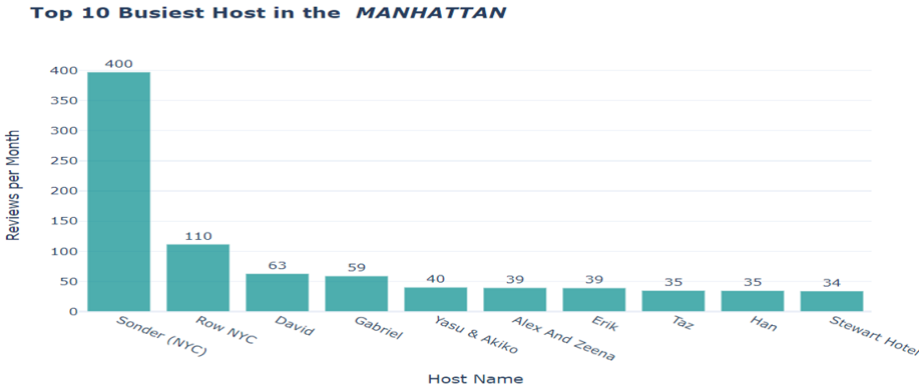
Popular Host and their business days in *BROOKLYN*

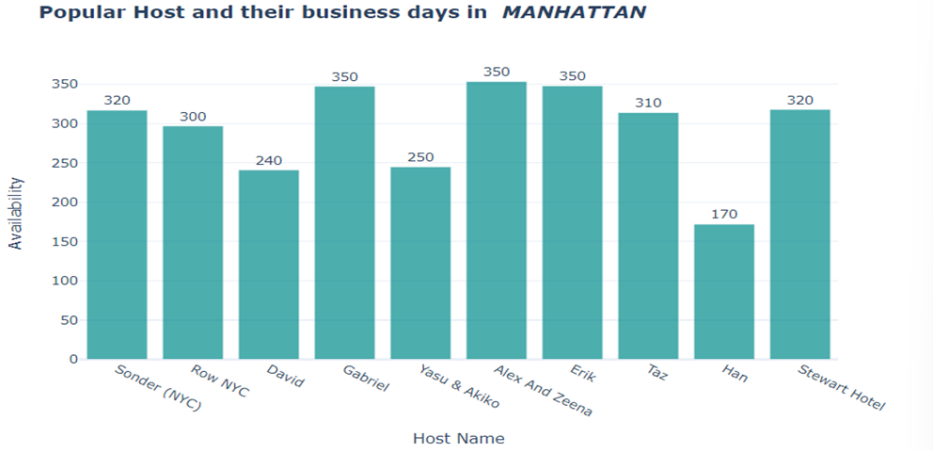


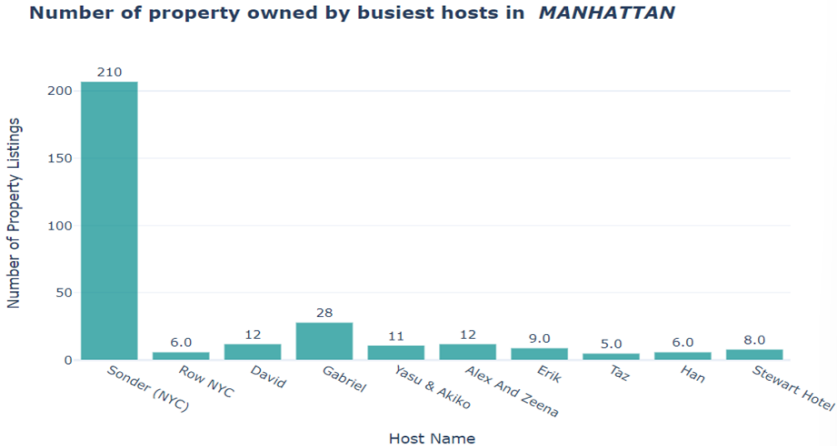
Number of property owned by busiest hosts in *BROOKLYN*

Bivariate Analysis- Manhattan

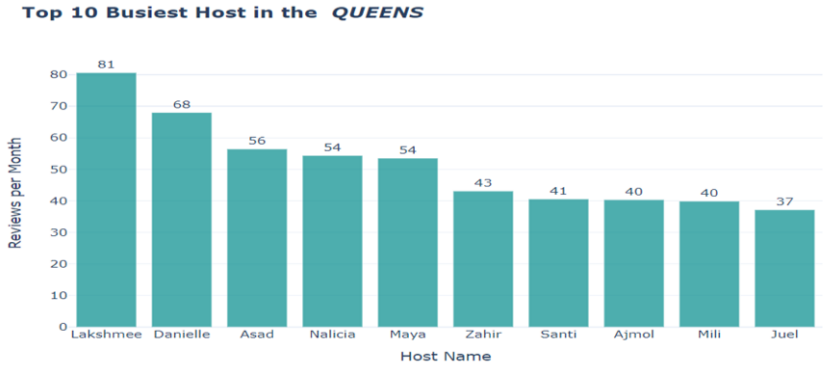
The average price in the state: **Manhattan** is **$180**



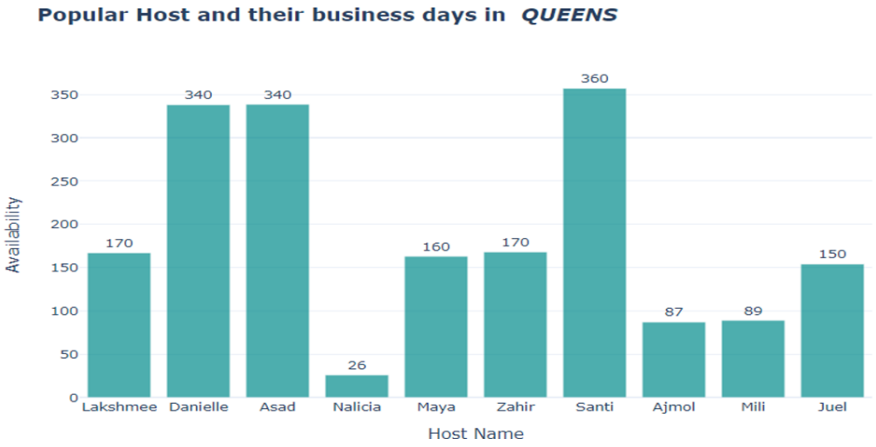


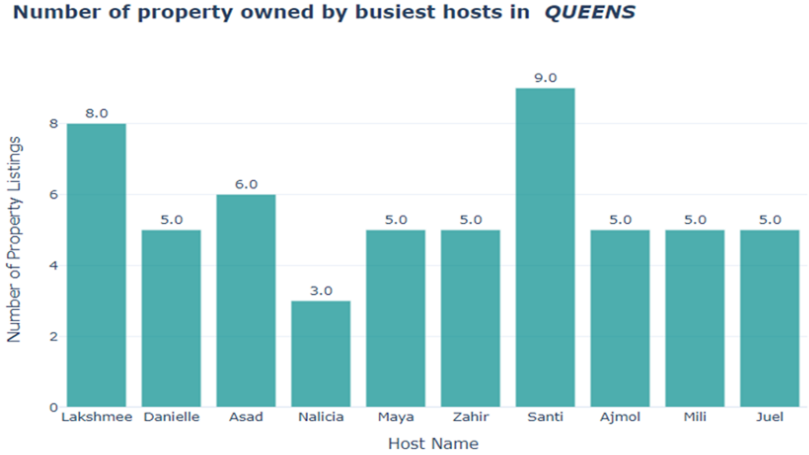


Bivariate Analysis- Queens



The average price in the state: **Queens** is **$96**

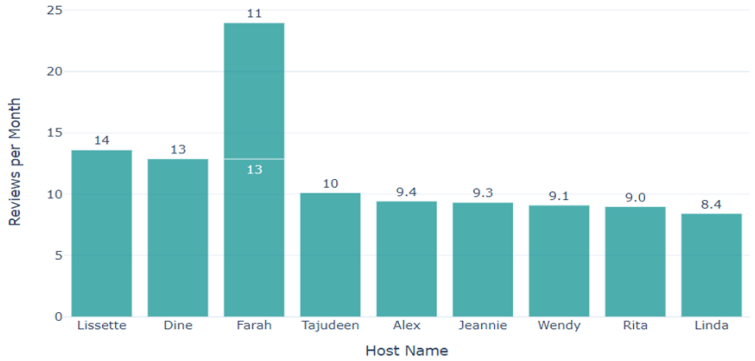




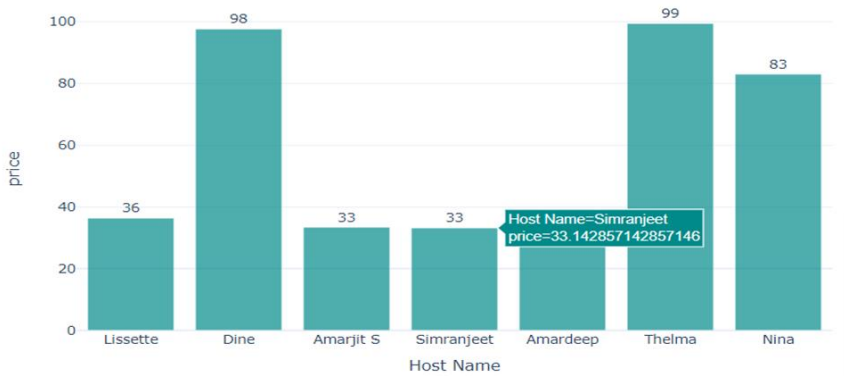
Bivariate Analysis- Staten Island

The average price in the state: **Staten Island** is **$90**

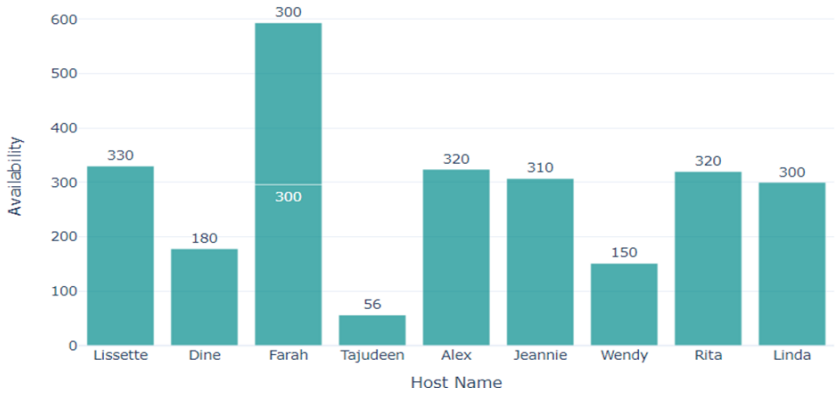
Top 10 Busiest Host in the *STATEN ISLAND*



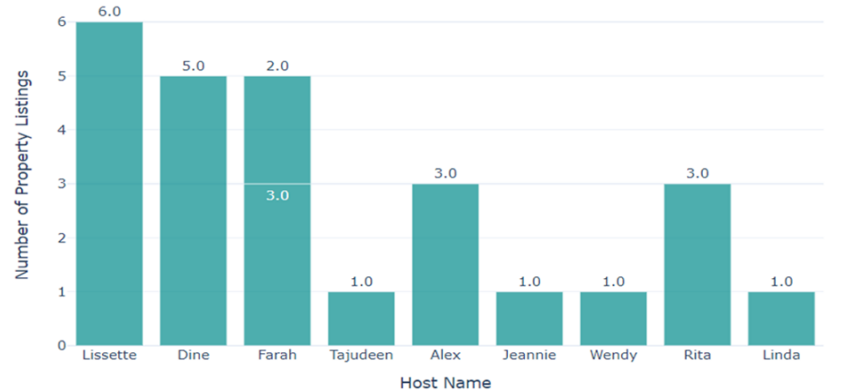
Top 10 most Popular Host with booking price *STATEN ISLAND*



Popular Host and their business days in *STATEN ISLAND*



Number of property owned by busiest hosts in *STATEN ISLAND*



Hypothesis evaluation:

Hypothesis : Average Growth of both Brooklyn and Manhattan are same.

Test statistic : Mann Whitney U Test

$$U = \min(U_A, U_B)$$

$U_A = ranksum_A - \frac{n_1(n_1+1)}{2}$; n1, n2 is the sample size of A & B.

$U_B = ranksum_B - \frac{n_2(n_2+1)}{2}$

- $U_{critical}=37 U_{stat}=71 U_{stat}>U_{critical}$

Conclusion: Average Growth of Brooklyn and Manhattan are same. The difference between them is not significant.

Therefore, we fail to reject **Hypothesis**, that is, **Average Growth of both Brooklyn and Manhattan are same.**

Hypothesis evaluation:

Hypothesis : The host which has more number of properties. Are they charges less for room price?

Test statistic : spearman rank correlation test

Spearman Correlation coefficient $\rho = 1 - \frac{6 * \sum di^2}{n(n^2-1)}$;

di= difference of ranks of two categorical variables

n= sample size

Spearman correlation coefficient is -0.10608308461077175 & p-value of test is 2.3661834310433083e-122

Conclusion: As the p-value is less than 0.05 significance level, so we will reject the null hypothesis. The sign of correlation is negative, which indicates that as the number of host listings count increases the charge for price decreases.

Hypothesis evaluation:

Hypothesis : Hypothesis: whether more number of properties a host has higher the number of reviews per month.

Test statistic : spearman rank correlation test

Spearman Correlation coefficient $\rho = 1 - \frac{6*\sum di^2}{n(n^2-1)}$;

di= difference of ranks of two categorical variables

n= sample size

Spearman correlation coefficient is .38 and P-value is 0.0.

Conclusion: As the p-value is less than 0.05 significance level, so we will reject the null hypothesis. The sign of correlation is positive, which indicates that as the number of host listings count increases the reviews per month also increases.

Hypothesis evaluation:

- Hypothesis : Mean price of all neighborhood groups are equal or not

Test statistic : Kruskal Wallis *Test*

$$H = \left[\frac{12}{n(n+1)} \sum_{j=1}^{c} \frac{T_j^2}{n_j} \right] - 3(n-1)$$

Where n = sum of sample sizes for all samples

c = number of samples

Tj = sum of ranks in the jth sample

nj = size of the jth sample

Kruskal Result(statistic=120.54185789749667, p-value=4.091877308111463e-25)

Conclusion: As the p-value is less than 0.05 significance level, so we will reject the null hypothesis. So we conclude that the mean price of property are not equal.

Custom Question

Q: Hi, Myself Riyan Mascarenhas from India, I've been appointed as an Assistant Professor in Columbia University in New York City. Since, I need time to settle down in the neighbourhood, I'm planning to go for renting a home/apartment for an year or two. So can you please suggest me some place names in or near the neighbourhood?

A:We can answer this query with the **Geographical location** of Required Landmark and assuming the maximum walkable distance is less than **2 Kilometres** from the Required Landmark. We need to consider Price, Type of the Place and Availability while suggesting a list of Preferable Places. Here we're considering all Properties with,

(i) Price lesser than the mean of all prices.

(ii) Availability greater than 120 days a year.

(iii) Room Type is Entire Home/Apartments

Geo-location of Required Columbia University is (40.80775585, -73.96164946987652). There are totally 3252 Places within 2 Kilometre radius of Columbia University

Custom Question (contd.)

Top 20 most Preferable Places are :

	name	price	distance	availability	latitude	longitude
0	Big one bedroom apt, new and beautiful.	11	0.98	273	40.79900	-73.96315
1	Cozy+Sunny 2 bedroom apt. close to Central Park	49	0.73	279	40.80719	-73.95303
2	Lovely 1B1B Apt in UWS Manhattan	60	1.41	157	40.81847	-73.95259
3	Wonderful Large 1 bedroom	75	1.60	362	40.82135	-73.95521
4	Full-Service Studio Apt in Brownstone/Townhouse	80	1.70	193	40.81726	-73.94583
5	The best studio in town	90	1.94	311	40.79596	-73.94463
6	Beautiful Large 1 BED in Upper Manhattan	90	1.89	269	40.82339	-73.95271
7	Studio Sanctuary in Landmark Brownstone	91	1.08	216	40.80480	-73.94951
8	Calming & Bright 1 BR Apt in Upper West Side	93	1.95	325	40.81982	-73.94481
9	Harlem cosy renovated unit for two	94	1.68	316	40.81618	-73.94517
10	Direct Central Park View from 6th floor Studio	98	1.25	359	40.79655	-73.96285
11	GREAT COZY APT	98	1.67	170	40.81143	-73.94248
12	Beautiful 2 bedroom private suite	99	1.08	296	40.80297	-73.95050
13	Harlem Monthly Rental True Two Bedroom 1 Bath	99	1.90	220	40.80060	-73.94115
14	STYLISH 2 BEDROOM APT RIGHT ON CENTRAL PARK WEST	99	1.26	201	40.79641	-73.96191
15	New Spacious 3BR/1BA Apt mins to Columbia Univ...	99	1.23	189	40.80927	-73.94720
16	Entire Luxury Newly Renovated Studio	99	1.99	135	40.79203	-73.97299
17	Spacious Harlem Condo	100	1.91	365	40.81691	-73.94251
18	Artist's Harlem Apartment	100	1.40	332	40.81003	-73.94535
19	Cozy PRIVATE Studio Apartment UWS and Jazz Tour.	100	1.08	281	40.79820	-73.96394

Custom Question

Q: Going through some negative reviews and comments, the manager wants to make some changes in the rules and regulation. But before making any changes, he wants to check whether there is some relationship in choosing the room type with the minimum nights to spend.

A:To check the dependency of ‘room_types’ with the ‘minimum_nights’.

- **Null hypothesis H0**: ‘room_type’ and ‘minimum_nights’ are independent, i.e. there is no relationship between them.

- **Alternative hypothesis H1**: ‘room_type’ and ‘minimum_nights’ are dependent, i.e. there is some relationship between them.

Chi-square test Formula:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

- Oi - Observed value
- Ei - Expected value

Custom Question (contd.)

Contingency Table :

time_interval room_type	1_or_2_days	Month	Quarter	Week	Year & +	Total
Entire home/apt	10463	4175	377	10267	127	25409
Private room	13154	2260	164	6686	62	22326
Shared room	799	151	9	193	8	1160
Total	24416	6586	550	17146	197	48895

The calculated value of test statistic is 1749.0172864434269

The critical value of test statistic is 15.50731305586545.

The p-value for the test is 0.0

- As the p-value is **less than the significance level '0.05'**, so we will **reject the null hypothesis**. Therefore, we will conclude that the ***'room_types' and 'minimum_nights' are not independent***, i.e. there is some significant relationship between them.
- While choosing the room type, there is a dependency of minimum nights to spend.

Custom Question

Q:I want to open a furniture store in NY. I want to know the trend for room type in NY across neighborhoods.

A:Since the correlation between columns 'number_of_reviews' and 'reviews_per month' is good, we're assuming that the data in column 'reviews_per_month' as 'number of reviews' a property received over it's lifetime divided by total number of months that property is existing from.

With this assumption we're creating two columns: 'duration' and 'possible_year_of_start'.

Next, we're counting all the properties with

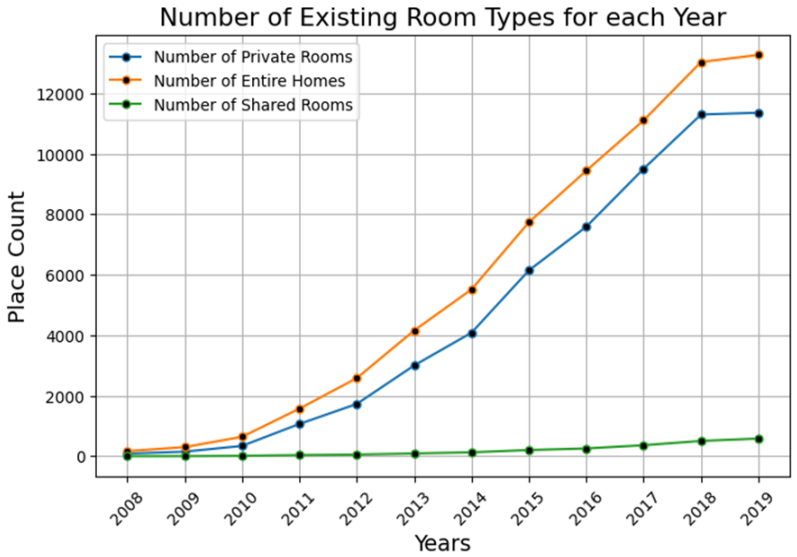
(i) **'last review year' greater than the desired year**

(ii) **'possible year of start' lesser than the desired year** as Properties Existing in that particular year ('desired year' is from 2008 to 2019).

Custom Question (contd.)

Number of Places per Year for each room type

	Year	Number of Private Rooms	Number of Entire Homes	Number of Shared Rooms
0	2008	91	173	3
1	2009	154	304	7
2	2010	344	645	16
3	2011	1070	1576	41
4	2012	1731	2586	58
5	2013	3003	4159	95
6	2014	4086	5518	133
7	2015	6145	7737	208
8	2016	7589	9440	260
9	2017	9505	11106	367
10	2018	11297	13033	509
11	2019	11356	13266	587

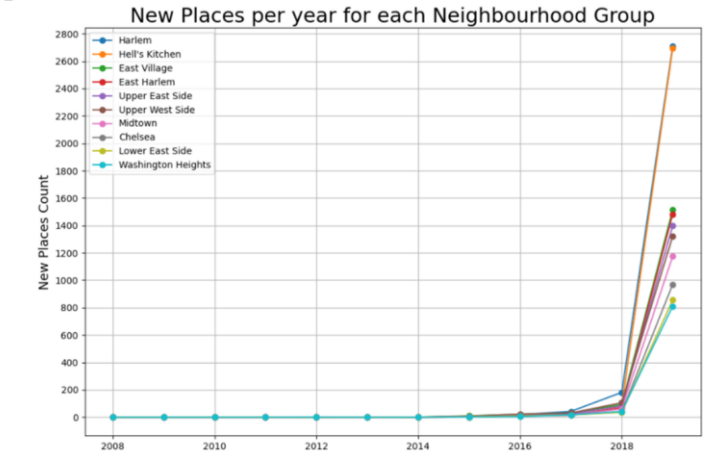


The above graph shows that the trend for private rooms and entire rooms is increasing significantly over the years, while there is increase for shared rooms, but is marginal.

Custom Question

Q:Which host has the potential to open a franchise of their own under Airbnb.

A:From the Primary Objective (1), We observe that Manhattan and Brooklyn have number of property listings consistently increased. Hence, we chose to do our following analysis on neighborhoods Manhattan and Brooklyn.

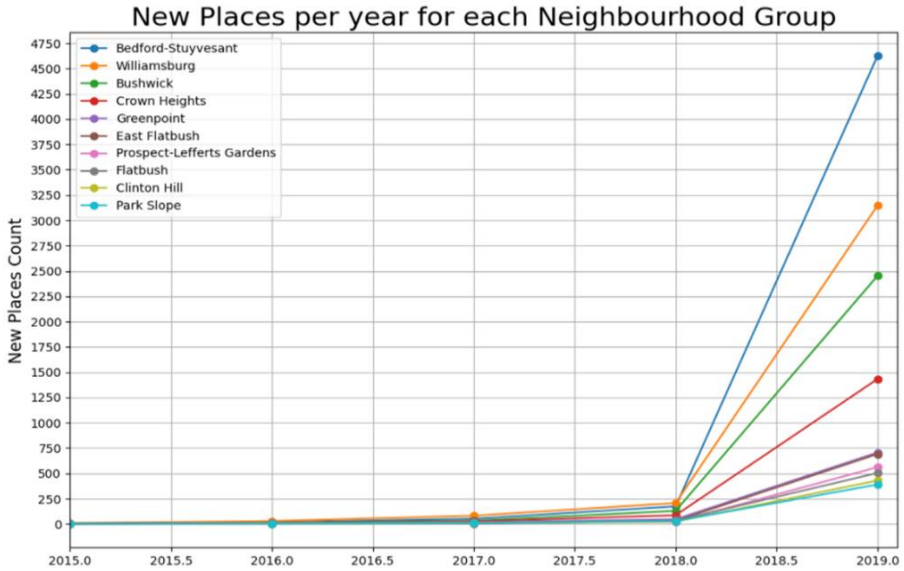


We observe that **top** reviewed **neighbourhoods** are **Harlem, Hell's Kitchen, East Village, East Harlem, Upper East Side, Upper West Side, Midtown,Chelsea, Lower East Side, Washington Heights.** So, Hosts with property listing over these neighbourhood is more likely to have a franchise of their own.

OBSERVATION:

We observe that Sonder (NYC) has made maximum investments on property in 2019. However, Gabriel and John have consistently invested from the possible year of start of their business at those neighbourhood which has garnered significant ratings over the years. Hence, Sonder (NYC), Gabriel and John have the potential to create their own franchise under Airbnb at Manhattan.

Custom Question (contd.)



We observe that **top** reviewed **neighbourhoods** are **Bedford-Stuyvesant,Williamsburg, Bushwick, Crown Heights, Greenpoint, East Flatbush, Prospect-Lefferts Gardens, Flatbush, Clinton Hill, Park Slope**. So, Hosts with property listing over these neighbourhood is more likely to have a franchise of their own.

OBSERVATION:

We observe that Melissa has made maximum investments on property in 2018. However, Anchor and Randy has consistently invested from the possible year of start of their business at those neighborhood which has garnered significant ratings over the years. Hence, Melissa, Anchor and Randy has the potential to create their own franchise under Airbnb at Brooklyn.

Conclusions:

1. Manhattan and Brooklyn are the neighbourhoods with the most properties available and with the busiest hosts.
2. Manhattan has the most expensive properties on avg. and entire home/apartments being most in demand. Rest of the neighbourhoods share almost same avg. price.
3. Hosts who have more number of properties and that of being either private rooms or entire home/apartments are in the busiest hosts list. Also they offer room in less price than others.
4. Traffic in a neighbourhood is dependent on the number of property it has and its reviews/month.

Thank you