Capstone Project(EDA) Airbnb booking analysis.

by

Ansari Nazar Anwar

❖ Problem Statement:

❖ Our main aim here is to analyze the Airbnb dataset that contains information about the thousands of listings made in New York through Airbnb.

- Airbnb is an online marketplace connecting travelers with local hosts. It benefits both the hosts who want to rent out their free space and also the traveler who wants a reliable and safe place to stay for cheap.
- We will approach this problem of analyzing the dataset by creating a hypothetical situation and defining some questions and extracting the answers from the data through data analysis.
- **The hypothetical situation:**
- Let us assume that my dad is a wealthy individual who wants to invest in the New York real estate market.

- He is informed by his friend Mr. Robert Kiyosaki that for usual cash flows from his properties he can rent them out on Airbnb.
- My dad approaches me and tells me to find out some valuable information that can make his investment profitable.
- Me being a data scientist, I take the data driven approach.
- **The hypothetical situation(contd):**
- The company I work for, Alma Better, already has a well formed dataset with thousands of listings made in New York city through Airbnb.

- I use this dataset to gain all the valuable information I would need to suggest my dad to make his investment profitable.
- I have defined some questions whose answers can be extracted from the dataset and are in alignment with our investment goals.
- Let us now have a look at our questions.
- **Questions to answer:**

1. Which hosts are the busiest and why.
2. For which location do the customers pay the highest and lowest rent.

3. Top 5 highest listing areas/locations.
4. What is the average preferred price by the customer for each neighborhood group and for each room type.
5. What is the percentage of the total number of nights spent for each neighborhood group.

❖ Understanding the data:

❖ To increase the efficiency of our analysis we will first have to understand the data and also check if there are some corruptions in the data and if any found try to treat it.

- The dataset contains 48,895 observations and 16 columns.
- Each observation includes information about an individual and unique listing.
- The 16 columns represent different fields. We will understand each column now.
- **The columns involved:**

• 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 • availabilty_365

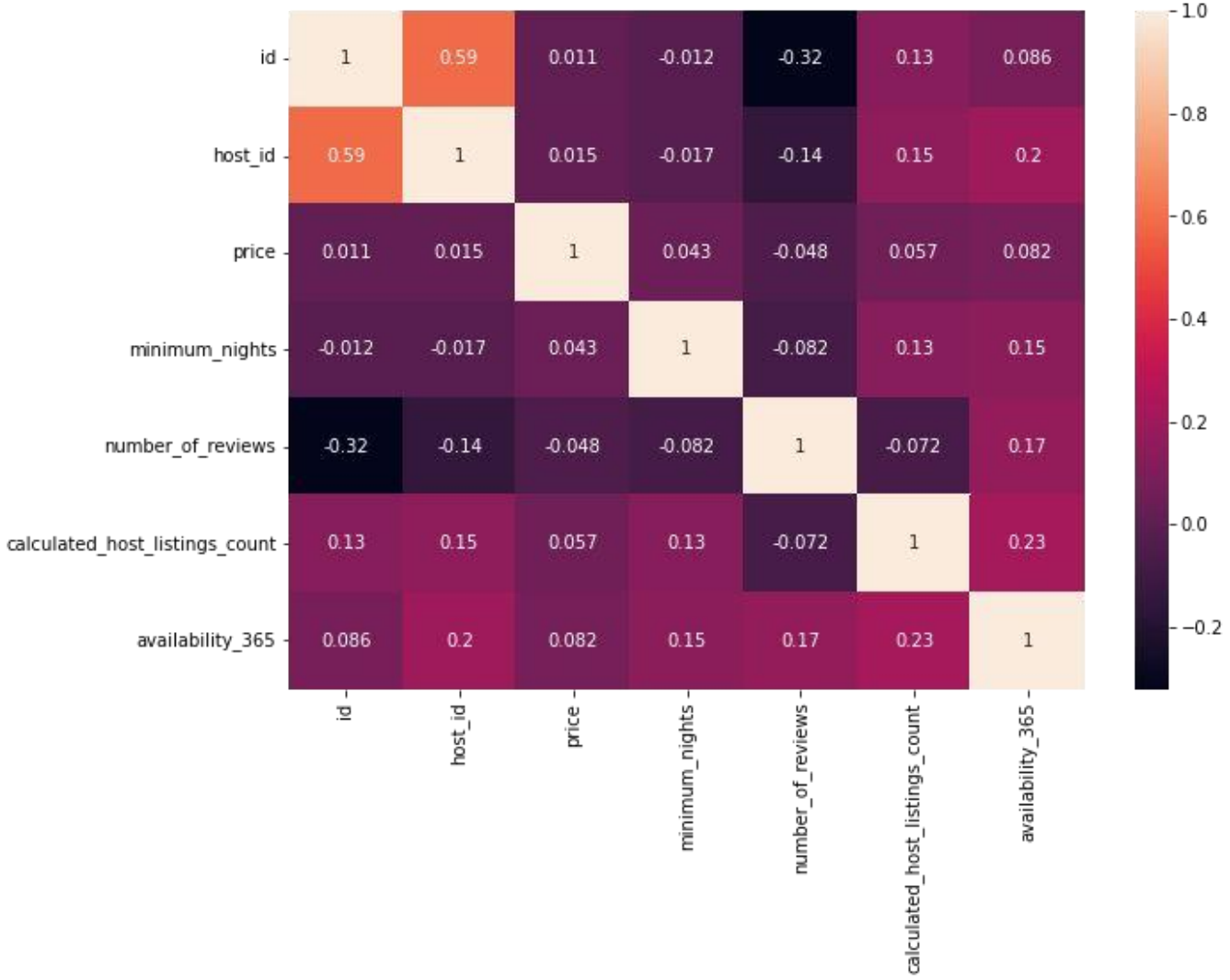
❖ Steps involved:

1. Dealing with null values:- There are two main ways of dealing with null values, One is to delete the null values and the other is to impute them with some meaning full values. In this case I have just deleted the observations with null values since they are not many and deleting them would not affect our analysis much.
2. Dealing with duplicate values:- Duplicate values can be detrimental to our analysis but luckily we don't have any duplicate values so we are good to go.
3. Feature selection:- I've deleted columns that I think won't add any value to the questions that I've aimed for the analysis.

❖ Steps Involved(Contd):

4. Figuring out the correlation: I've tried to figure out the correlation between variables by plotting the correlation matrix using a heat map, offered by the seaborn module of python. As we can see that there isn't any meaningful positive correlation so we will have to find out the latent relation of these variables with each other through further analysis.

The


id
host_id
price
minimum_nights
number_of_reviews
calculated_host_listings_count
availability_365
1 0.59 0.011 -0.012 -0.32 0.13 0.086
0.59 1 0.015 -0.017 -0.14 0.15 0.2
0.011 0.015 1 0.043 -0.048 0.057 0.082
-0.012 -0.017 0.043 1 -0.082 0.13 0.15
-0.32 -0.14 -0.048 -0.082 1 -0.072 0.17
0.13 0.15 0.057 0.13 -0.072 1 0.23
0.086 0.2 0.082 0.15 0.17 0.23 1
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

visualization of the correlation matrix.

❖ Steps Involved(Contd):

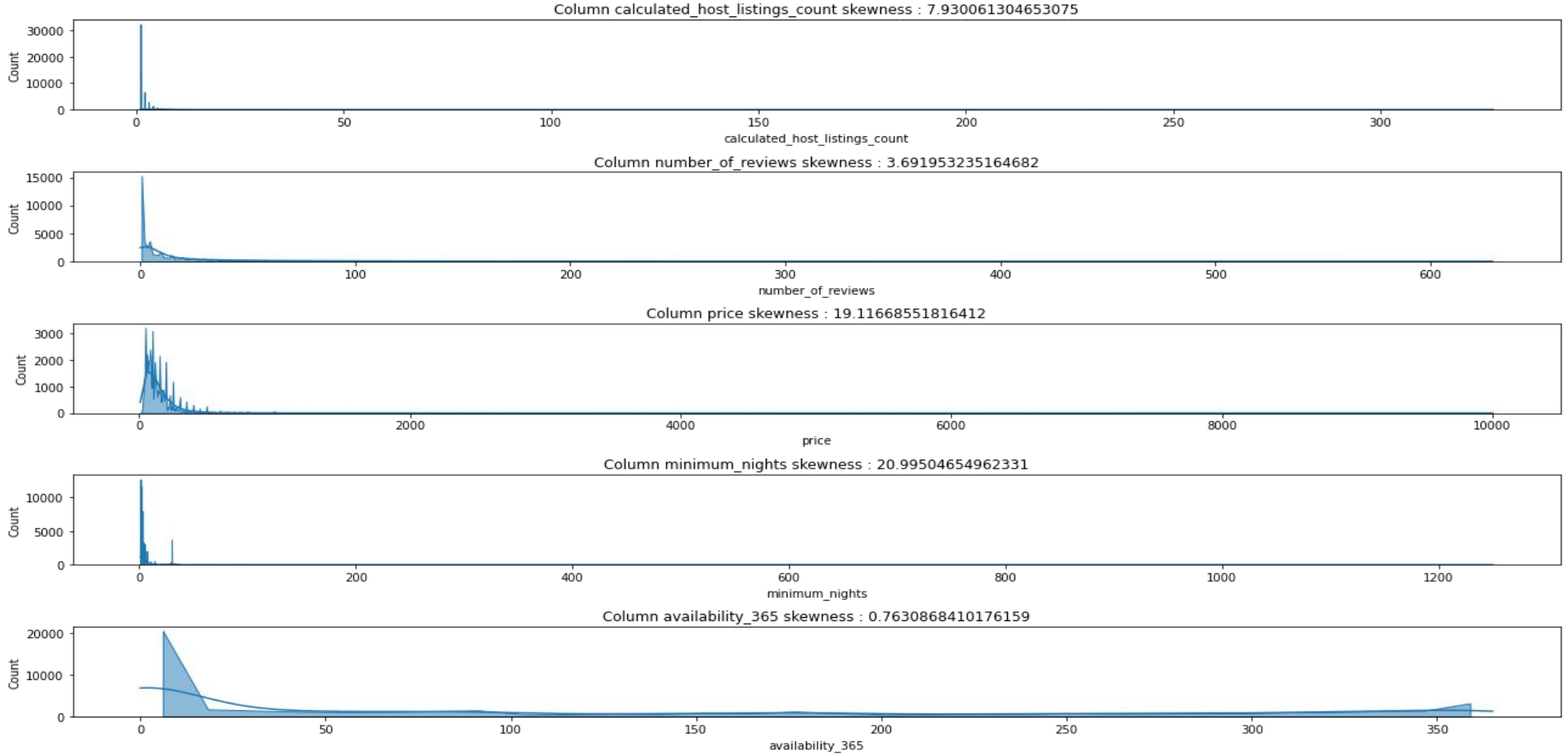
5. Checking the distribution: I've tried checking the distribution of all the important numerical variables, to check if any valuable information can be taken out of it. Here we can see that all our variables have a positive skewness except availability which is moreover a uniform

distribution, this tells us that all kinds of listings are moreover available throughout the year.(We can also see a mild to high rise in count of availability from the end of the year to the start of the year)

The visualization has been given in the next slide.

Note:- Mention the great hack....

The distribution visualisation:



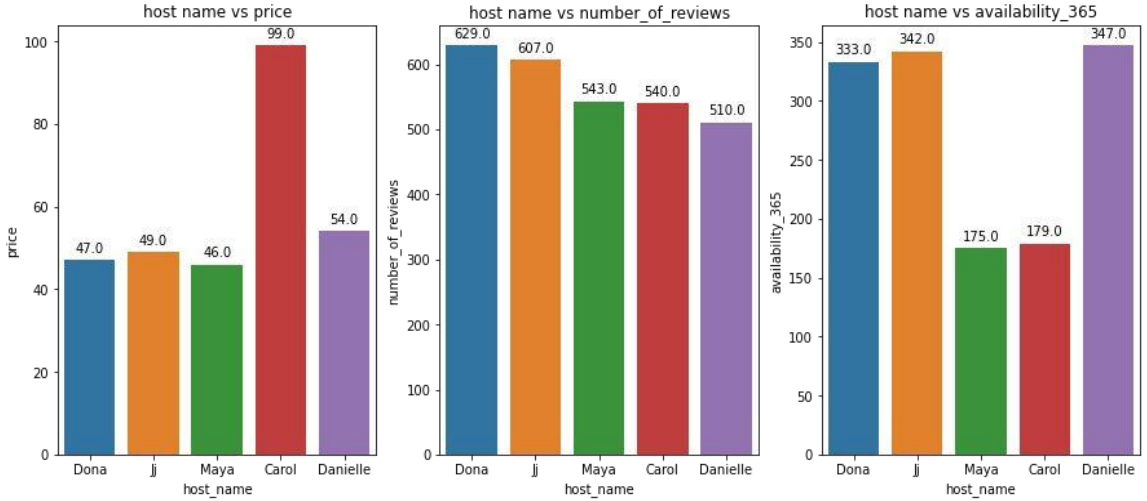
❖ Steps Involved(Contd):

6. Answers:

Q1. Which hosts are the busiest and why.

Ans: By analysis we find that the busiest hosts are Dona, JJ, Maya, Carol and Danielle(in-order).

By analysing the graphs we can see that the number of reviews directly affects the popularity of the host (in a linear fashion). It is also seen that all these guys have the right combination of price (around 70$) and availability(around half a year)



❖ Steps Involved(Contd):

6. Answers(Contd):

Q2. For which location do the customers pay the highest and lowest rent.

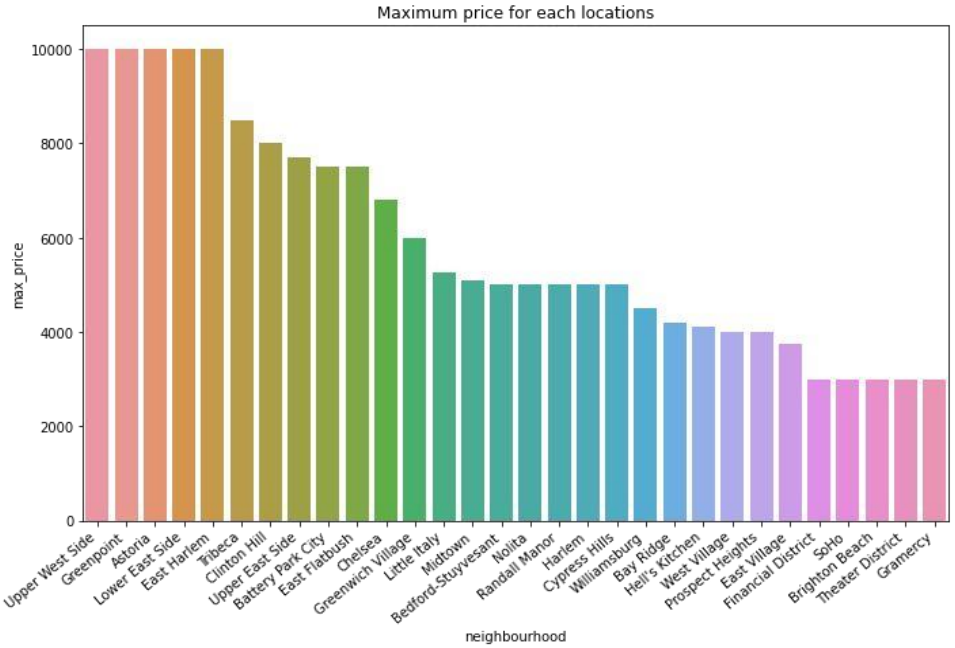
Ans: The neighbourhoods with the highest rents are:

1 . Upper West Side with a price of: 10000.0

2 . Greenpoint with a price of: 10000.0

3 . Astoria with a price of: 10000.0 4 . Lower East Side with a price of: 9999.0

5 . East Harlem with a price of: 9999.0

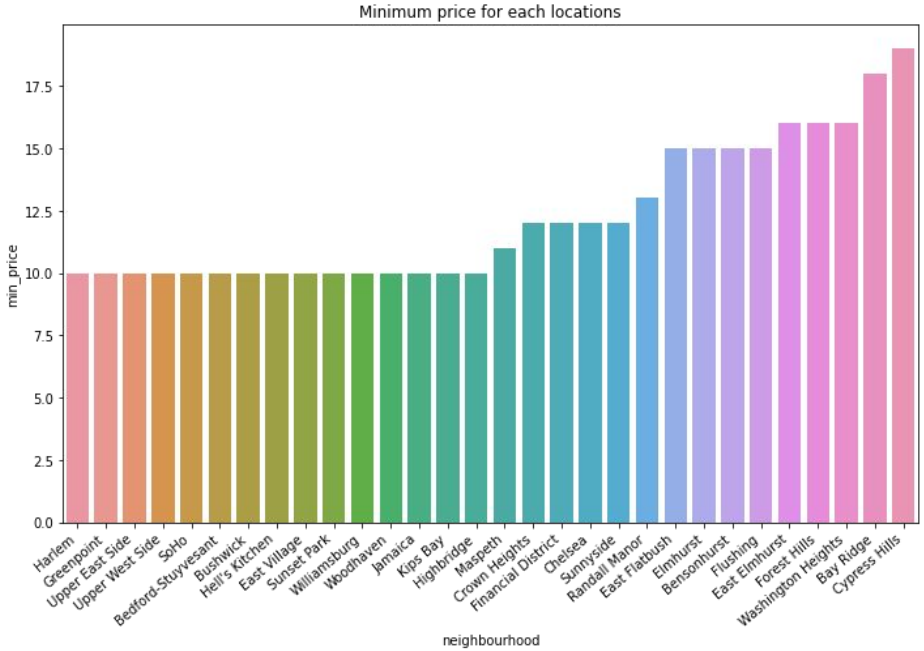


❖ Steps Involved(Contd):

6. Answers(Contd):

Q2. For which location do the customers pay the highest and lowest rent.

Ans(Contd):The neighbourhoods with the lowest rents are:
1 . Harlem with a price of: 10.0 2 . Greenpoint with a price of: 10.0 3 . Upper East Side with a price of: 10.0
4 . Upper West Side with a price of: 10.0
5 . SoHo with a price of: 10.0

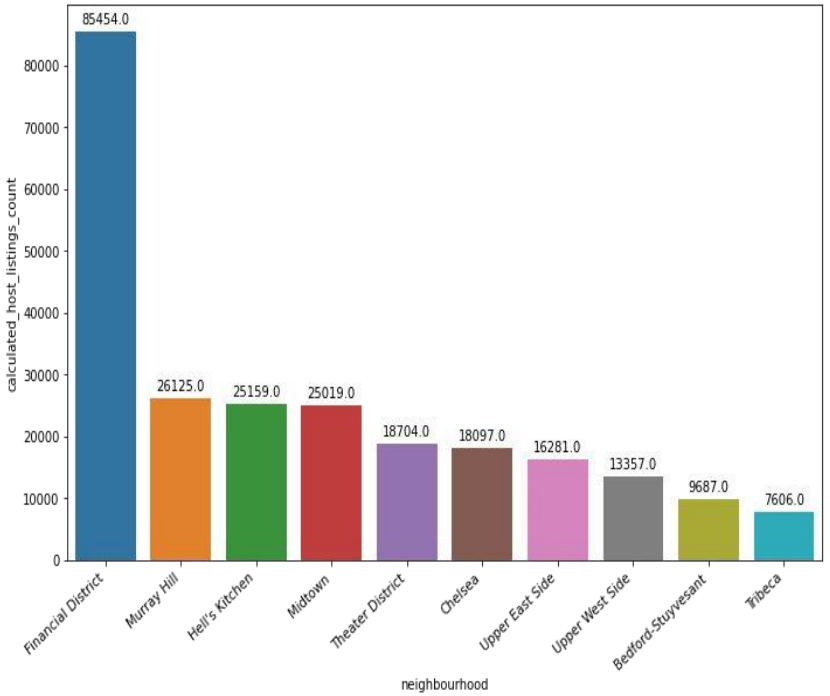


❖ Steps Involved(Contd):

6. Answers(Contd):

Q3. Top 5 highest listing areas/locations.

Ans: The top 5 highest listed areas are: 1 . Financial District with 85454 number of listings
2 . Murray Hill with 26125 number of listings
3 . Hell's Kitchen with 25159 number of listings
4 . Midtown with 25019 number of listings
5 . Theater District with 18704 number of listings

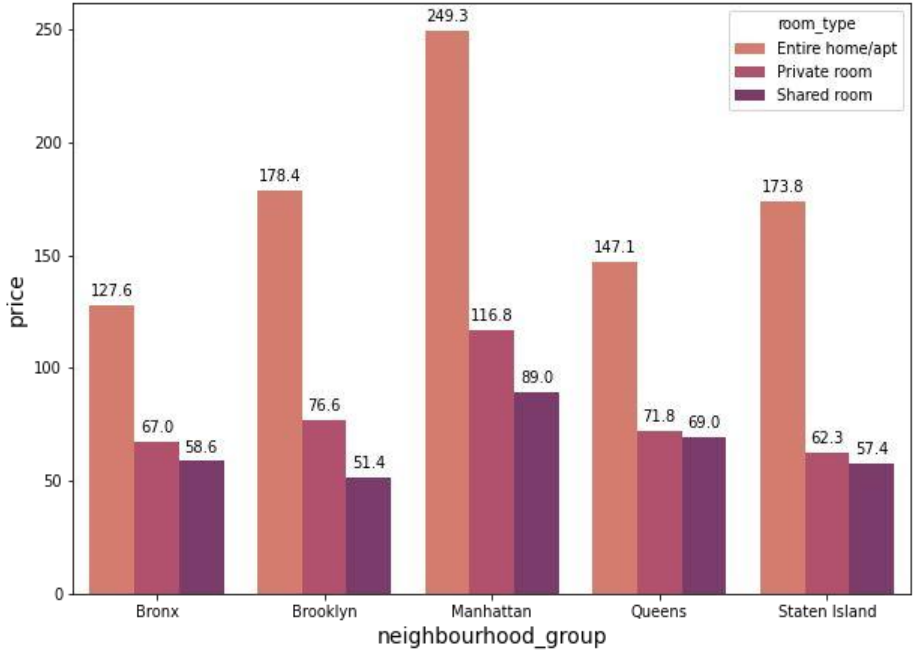


❖ Steps Involved(Contd):

6. Answers(Contd):

Q4. What is the average preferred price by the customer for each neighborhood group and for each room type.

Ans: Manhattan seems to be the neighborhood group with the highest rent and the Entire home(room_type) has the highest rent. The visualisation has all the



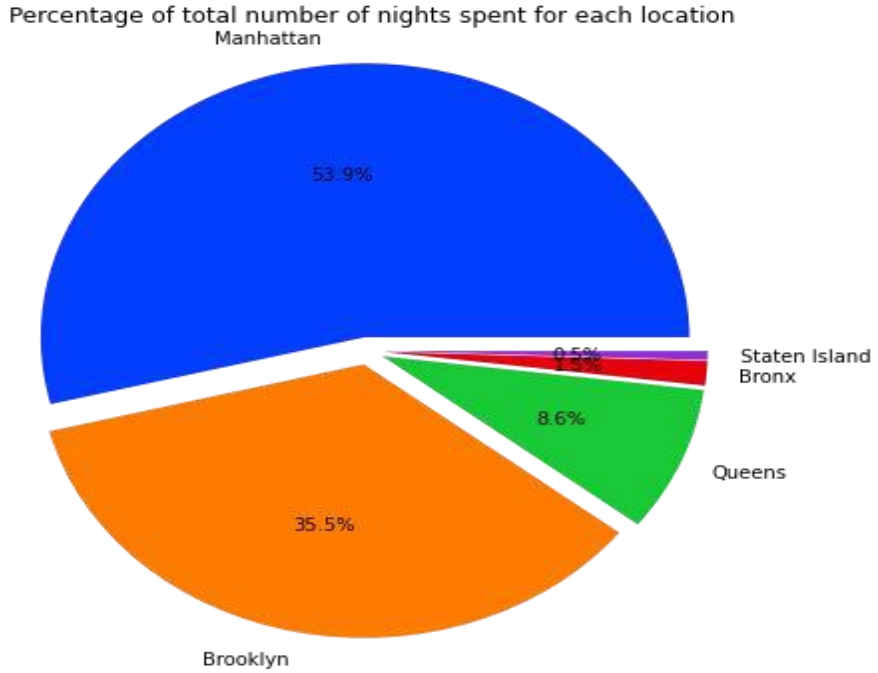
details about the average price.

❖ Steps Involved(Contd):

6. Answers(Contd):

Q5. What is the percentage of the total number of nights spent for each neighborhood group.

Ans: The percentage of the total number of nights spent for each neighborhood group is found out to be -
Manhattan-53.9%
Brooklyn-35.5%



Queens-8.6%
Bronx-1.5%
Staten Island-0.5%

❖ Summary:

Answers:

- The busiest hosts are Dona, JJ, Maya, Carol and Danielle. These seem to be the busiest hosts because they have the optimum combination of price(around 60 dollars), availability(around half a year) and number of reviews(around 550). All of these listings also have room type as private room.
- The neighbourhoods with the highest rents are Upper West side, Greenpoint, Astoria, Lower East side and East Harlem. The highest rent is 10,000 dollars.
- The neighbourhoods with the lowest rents are Harlem, Greenpoint, Upper East Side, Upper West Side and SoHo. The lowest rent is 10 dollars.

- The top 5 highest listed areas are Financial District, Murray Hill, Hell's Kitchen, Midtown and Theater District.

- The average prices for each neighbourhood group and for each room type is found.
- The percentage of total nights spent for each neighbourhood group is found.

❖ Summary(Contd):

Actions I would tell my dad to take:

- Invest in such a place which can be converted as a private room and be rented out for a price of around 60 dollars per night.
- Make sure he has proper maintenance of the place so that it is available around the year. (Especially at the beginning and end of the year)
- Encourage customers to put out reviews about the place by giving gift coupons or future discounts.

- Try buying a plot in Manhattan, because it has the highest percentage of total nights spent.
- Buying a place where the rent is the highest is also a good option because it would draw the attention of a lot of people with low prices.
- It is better to avoid buying in the neighbourhoods with highest number of listings, because the market there is already saturated with good offers.
- Try pricing the property according to the averages calculated.

THANKYOU!!

We've completed our analysis of the New York Airbnb dataset and have informed my dad about all the valuable information we've found. We've also informed him about all the actions he can take to maximize his profits.

At the end we hope that all this analysis is beneficial to my dad and he is proud of me. Which might not be possible because as the saying goes, "A happy Indian Parent is a myth". At this note we will end our Analysis.

Thank you for tagging along until the end of my journey to make my dad proud.