Technical Documentation

Airbnb Bookings Analysis

Capstone Project 1

Submitted by:
Chetan Prakash

Index :

Abstract :

Airbnb is a popular online marketplace that allows people to find and rent lodging from local hosts in over 220 countries and regions. Founded in 2008, Airbnb has revolutionized the travel industry by offering unique and affordable accommodations that go beyond the traditional hotel experience. Through its platform, guests can book a wide range of properties, including private rooms, apartments, villas, treehouses, and even castles.

Airbnb has disrupted the hospitality industry and has become a preferred choice for travelers seeking a more authentic and personalized experience. Hosts on the platform can earn extra income by renting out their spare rooms or entire homes, while guests can enjoy local experiences and connect with their hosts and communities. The platform's success is largely due to its user-friendly interface, robust community, and commitment to safety and security.

Airbnb has faced some challenges over the years, including regulatory hurdles and concerns over safety and privacy. However, the company has continued to innovate and adapt to the changing needs of its users. With a growing user base and a commitment to sustainability and responsible tourism, Airbnb looks set to continue its success in the travel industry.

Introduction :

Airbnb is a game-changing platform that has transformed the way people travel and experience new destinations. Founded in 2008, it has grown into a global phenomenon, offering a unique and personalized way to book lodging accommodations in over 220 countries and regions. With its user-friendly interface, diverse range of properties, and commitment to creating authentic local experiences, Airbnb has disrupted the traditional hospitality industry.

The concept behind Airbnb is simple: to connect travelers with local hosts who are willing to rent out their homes or spare rooms. This not only provides guests with a more personalized and affordable travel experience but also offers hosts the opportunity to earn extra income from their properties. Airbnb has built a strong community of hosts and guests, fostering connections and cultural exchanges between people from all over the world.

Despite its success, Airbnb has faced its fair share of challenges, including regulatory hurdles and concerns over safety and privacy. However, the company has worked tirelessly to address these issues and to create a safe and secure platform for its users. As a result, Airbnb has continued to grow and evolve, expanding its offerings and developing new features to enhance the user experience.

Overall, Airbnb has revolutionized the way people travel, offering a unique and personalized alternative to traditional hotel stays. With its commitment to sustainability and responsible tourism, it looks set to continue its success and inspire further innovation in the travel industry.

Problem Statement :

Airbnb has faced a number of challenges over the years, including regulatory hurdles and concerns over safety and privacy. In some cities, the platform has been met with opposition from residents who are concerned about the impact of short-term rentals on their neighborhoods. Additionally, there have been cases of hosts and guests experiencing safety and security issues, such as theft, vandalism, or even assault. These incidents have raised questions about the responsibility of Airbnb and its hosts to ensure the safety of guests, and the need for better regulation and enforcement to protect both hosts and guests.

Moreover, the COVID-19 pandemic has severely impacted the travel industry, including Airbnb. The company has faced cancellations and a decline in bookings due to travel restrictions and health concerns. The pandemic has also highlighted the importance of health and safety measures, such as enhanced cleaning protocols and contactless check-in, which have become crucial for the recovery of the travel industry.

Therefore, the problem statement for Airbnb is how to address these challenges and continue to provide a safe, secure, and responsible platform for both hosts and guests in the face of regulatory, safety, and health concerns.

What is Exploratory Data Analysis (EDA)?

Exploratory data analysis (EDA) is a method used to analyze and summarize a dataset in order to understand its characteristics and patterns. EDA can be used to clean and preprocess the data, as well as identify any outliers or anomalies that may be present. Some common techniques used in EDA include visualizing the data using graphs and plots, calculating summary statistics, and identifying correlations and relationships between variables.

The following are the various steps involved in the EDA process:

Data collection- Collecting the relevant data that is necessary for the analysis. This can include gathering data from various sources such as databases, surveys, and experiments. We took the datasets from Kaggle.

Problem Statement - This is the initial step on understanding the attributes of the datasets and based on that the aim of EDA is defined.

Data Cleaning - Cleaning the data by removing any missing or corrupted values, and correcting any errors or inconsistencies.

Data Exploration - Exploring the data by creating visualizations, performing summary statistics, and identifying patterns and trends. This step helps to gain a better understanding of the data and identify any potential issues or outliers.

Data Transformation - Transforming the data to make it more suitable for analysis. This can include normalizing data, creating new variables, or removing outliers.

Data Visualization - Creating visual representations of the data to better understand and communicate the findings. This step can include creating charts, graphs, maps, and other visualizations to help convey the key insights of the analysis.

Data Interpretation - Interpreting the results of the analysis and drawing conclusions from the data. This step includes communicating the findings to stakeholders and making recommendations for further action.

Conclusion - Based on the analysis, a final set of observations are made and recommendations are derived from that.

Dataset - Airbnb NYC 2019

The data set contain these columns :

- **Id:** This column contains the unique ids of clients who made bookings.
- **Name:** This column contains the host name at different locations.
- **Host id:** This column contains the unique host id so the client can find the required host easily.
- **Host name:** This column contains the host name.
- **Neighbourhood Group:** This column contains the name of the neighborhood group. It means the larger geographical area or region within a city where a particular location is located.
- **Neighborhood:** This column contains the name of the neighborhood. It means particular locality in a large area like in New York City, the neighborhood column would indicate the specific neighborhood, such as the lower east side, Harlems, or

Williamsburg.

- **Latitude:** This location contains the latitude of a particular location in a particular area.

- **Longitude:** This location contains the longitude of a particular location in a particular area.

- **Room Type:** This column contains the different room types in different locations like Private room, Entire Apartment, Sharing Room, Entire Home etc.

- **Price:** This column contains the pricing of different room types according to its location.

- **Minimum Neights:** This column contains minimum nights that clients should stay or pay rent for.

- **Number of Reviews:** This column contains the total number of reviews that a particular listing has received from previous guests. This helps to predict the popularity of a particular host.

- **Last Review:** This column contains the most recent review left by the guest for a particular listing.

- **Reviews per Month:** This column contains the average number of reviews that particular listing receives per month.

- **Calculated host listing count:** This column contains the total number of listings that a particular host has on Airbnb. Host with a highly calculated host listing count indicates the experience and expertise of the host.

- **Availability 365 days:** This column contains information about the host who opens for how many days in a year. This can affect profitability.

Data Cleaning and Preparation:

Data cleaning is the process of identifying and correcting or removing errors, inconsistencies, and missing values in a dataset. It is an important step in the data preprocessing phase, as it can improve the quality and reliability of the data and make it more suitable for analysis.

Some common data cleaning techniques include:

Removing duplicate records: Identifying and removing duplicate records from the dataset.

Handling missing values: Identifying and handling missing values, which can be done by either removing the entire record or replacing the missing value with a suitable estimate.

Formatting and type conversion: Formatting data values to ensure consistency and converting them to the appropriate data type.

Outlier detection: Identifying and handling outliers, which are extreme values that can skew the results of the analysis.

Normalization and scaling: Normalizing and scaling numerical variables to ensure that they are on the same scale and can be compared more easily.

Text cleaning: Cleaning the text attributes like removing stop words, stemming, lemmatization.

It's important to note that data cleaning can be an iterative process, as errors and inconsistencies may be discovered during the cleaning process, and multiple rounds of cleaning may be necessary to achieve high-quality data. Additionally, data cleaning should be done with care as it could lead to loss of important information.

Steps that would taken:

➔ Step 1 - First of all I try to find information about the dataset by using the df.info() function. By this function I find that there are 10 numerical and 5 categorical columns.

➔ Step 2 - Then I try to find the duplicate values but there is no duplicate value so we can move further.

➔ Step 3 - Then I use the df.describe() function that describes all the things regarding columns like count, mean, std, min, 25%, 50%, 75%, max. This gives a general overview for our dataset.

➔ Step 4 - Then I use df.describe().columns to find the numerical columns name. It shows 5 names: id, host_id, latitude, longitude, price, minimum_nights,number_of_reviews, reviews_per_month, calculated_host_listings_count, availability_365.

➔ Step 5 - Then I try to find the null values so we use df.isnull().sum() this function gives all the number of null values in different columns like name : 16, host_name : 21, last_review : 10052, and review_per_month : 10052.

➔ Step 6 - Then I make a copy for our original dataset by using df.copy() function so that it is not affected.

➔ Step 7 - Then I rectify all the null values with the help of df1.fillna() functions like name : no name, host_name : No name, and last_review : no review and rest all the numerical columns as 0.

➔ Step 8 - Then we find the values with the help of values count() function.

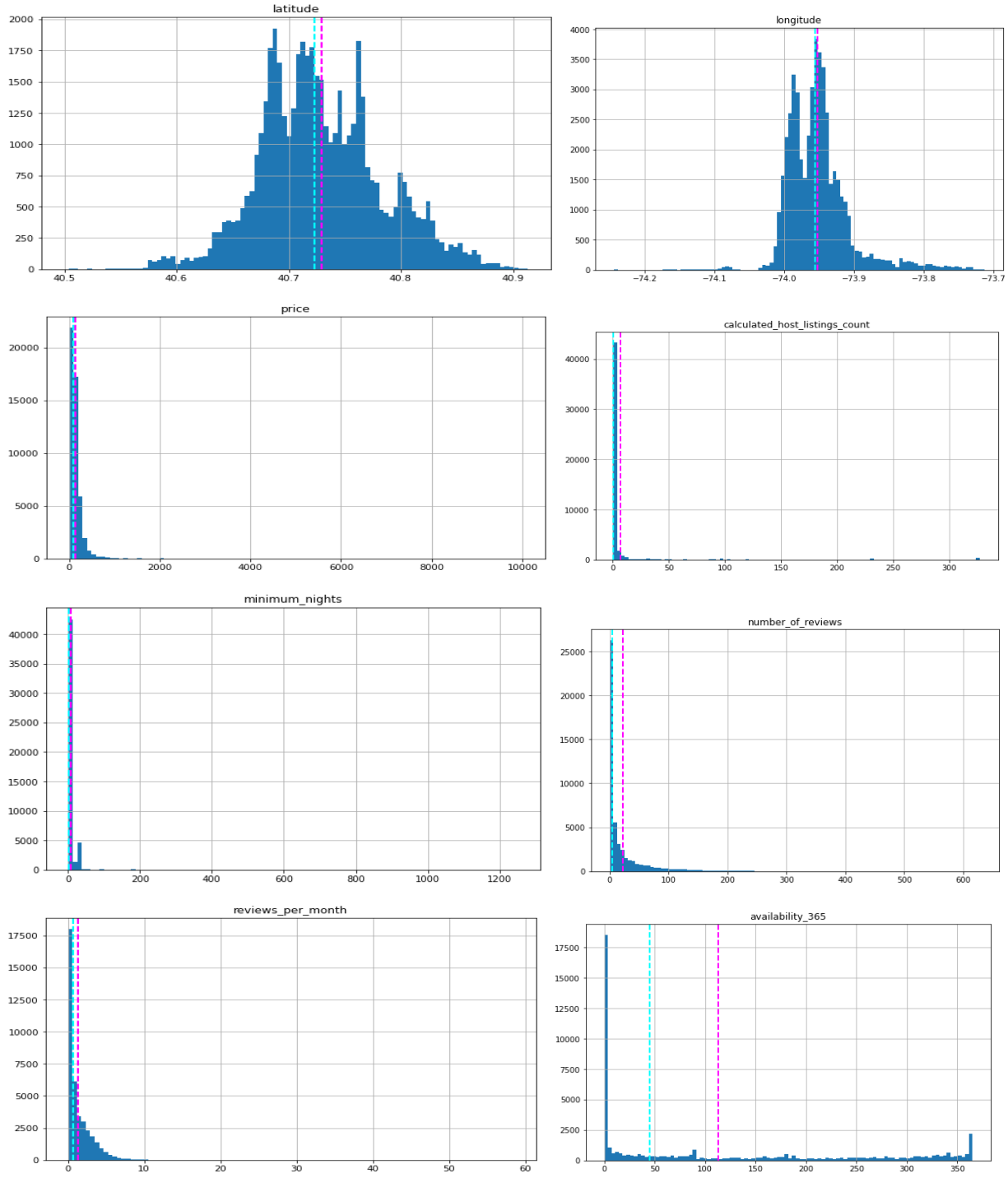
EDA Findings :

Exploratory Data Analysis, or EDA, is an important step in any Data Analysis or Data Science project. EDA is the process of investigating the dataset to discover patterns, and anomalies (outliers), and form hypotheses based on our understanding of the dataset. EDA involves generating summary statistics for numerical data in the dataset and creating various graphical representations to understand the data better. In this article, we will understand EDA with the help of an example dataset. We will use Python language (Pandas library) for this purpose.

★ Map of New York:

This is the map of New York City from where this dataset belongs. So we can take ideas for different locations.

★ Mean and median value of all the numerical value:

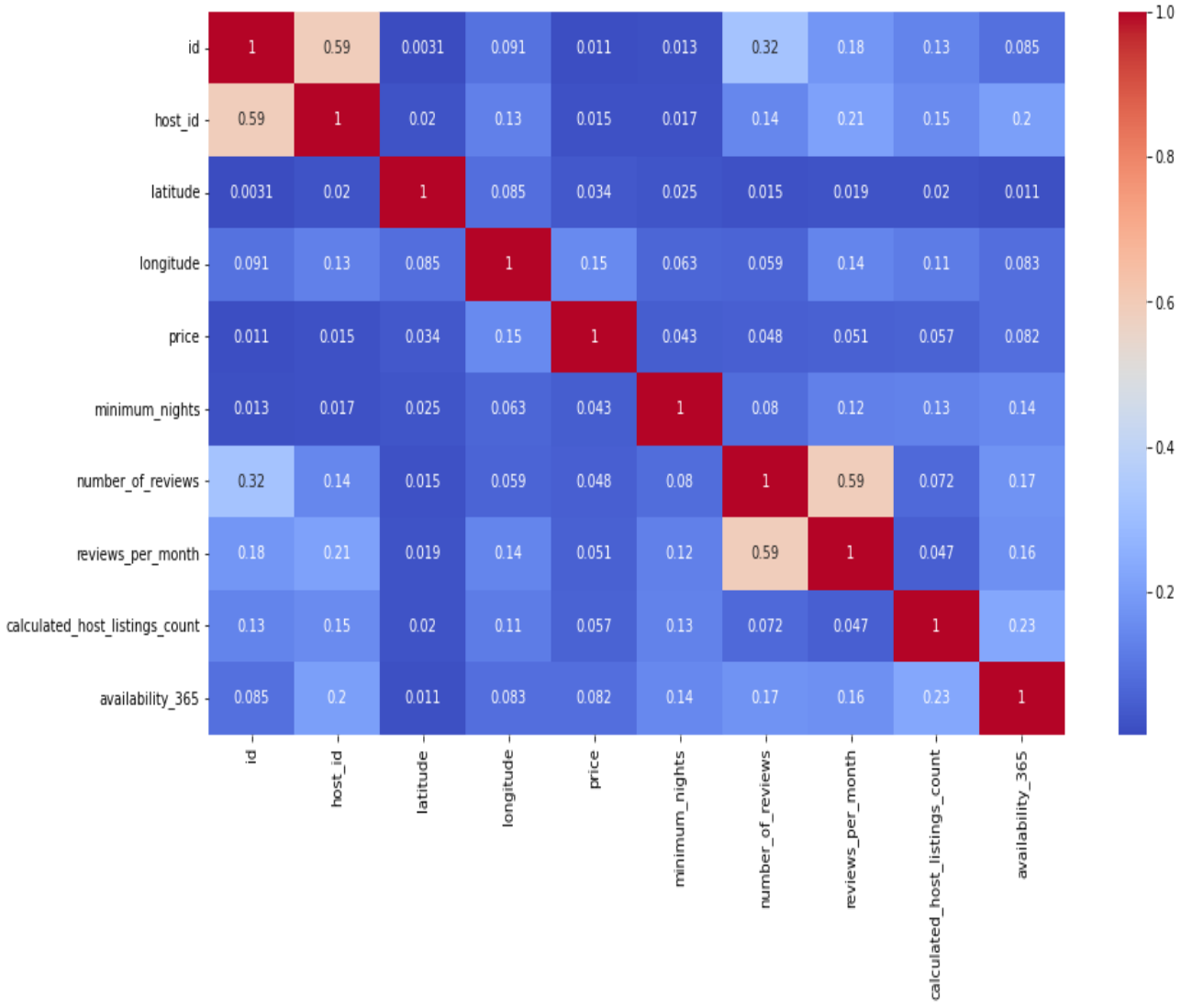


These graphs represent the mean and median value of all numeric

columns. Mean is also known as the average, is calculated by adding up all the values of all the dataset and divided by total number of values. Median shows the middle value in the database, that is half values are up and half down.

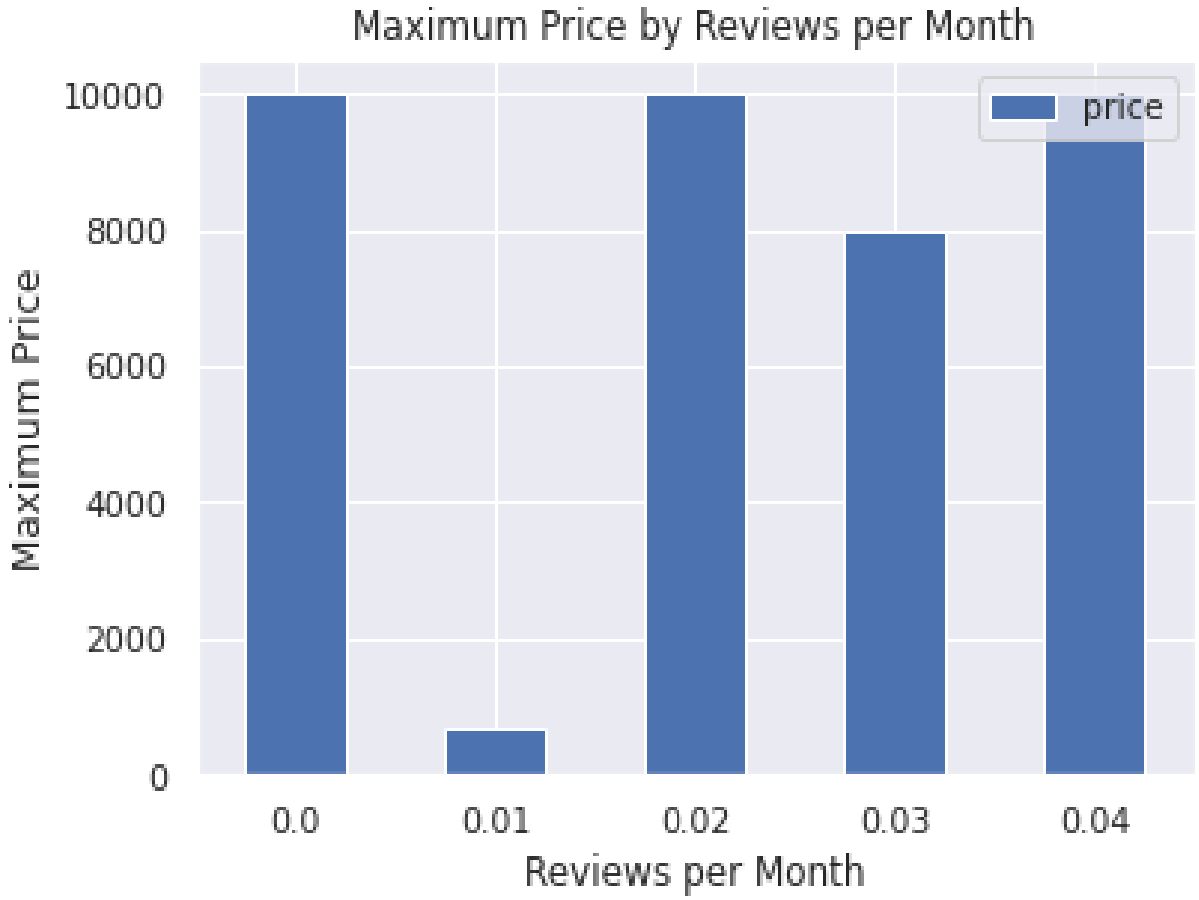
★ Correlation between all the numerical data:

This is a heat map that shows the relation between all the numeric values of the column.



As we can see from this chart there is not that much correlation between the given category because the maximum value here we found is 0.32 between Host_id and number_of_reviews and some more correlation between host_listing_count to availability_365.

★ Best rating hotel according to their price :



As we can see from this analysis, generally people review hotels who have less price.

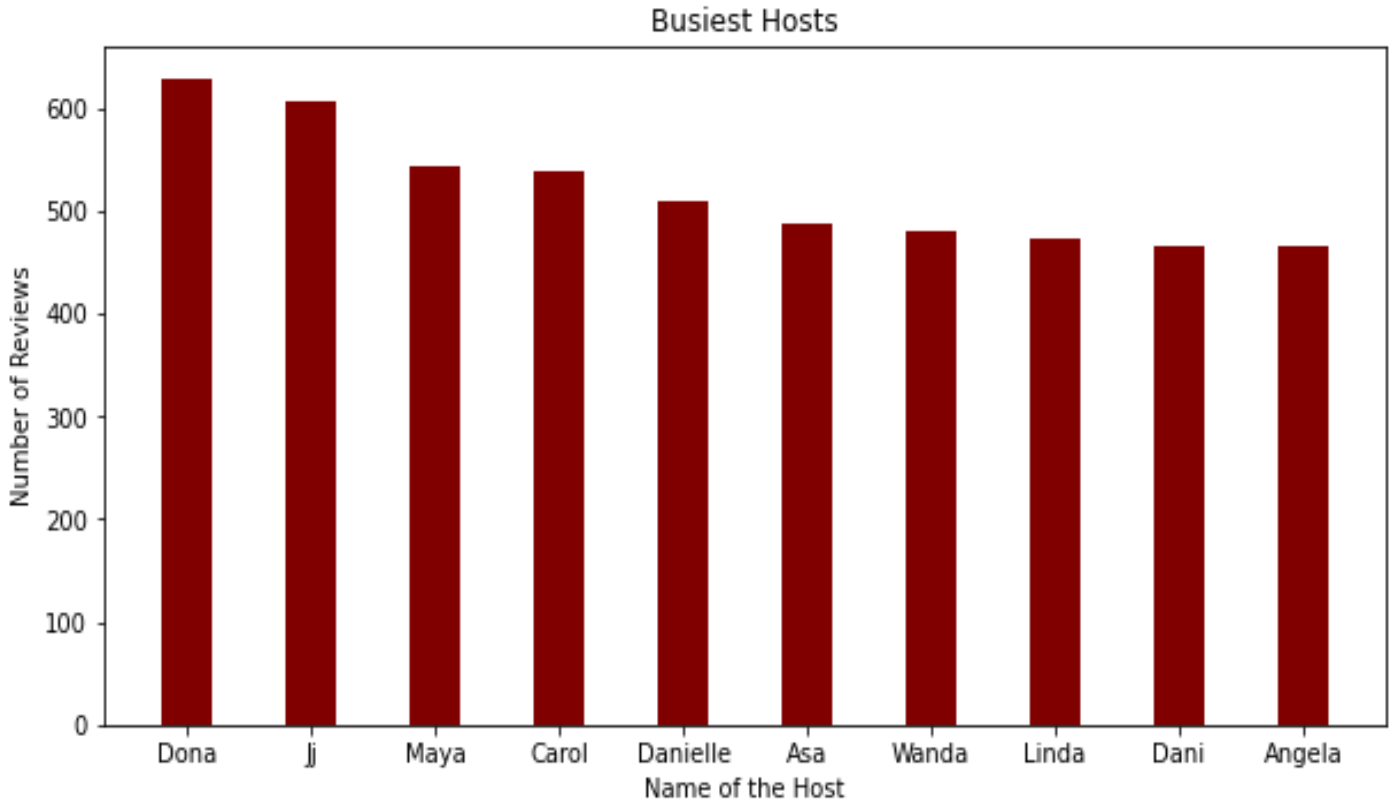
From the 1st bar that price is around 1000 whose reviews is 0 and at the 3rd bar price is again around 1000 rupees whose reviews is 0.02.

★ For each neighborhood count how many of them prefer the same location :

Index	host_name	neighbourhood_group	calculated_host_listing_count
13217	Sonder(NYC)	Manhattan	327
1834	Blueground	Manhattan	232
1833	Blueground	Brooklyn	232
7275	Kara	Manhattan	121
7480	Kazuya	Queens	103

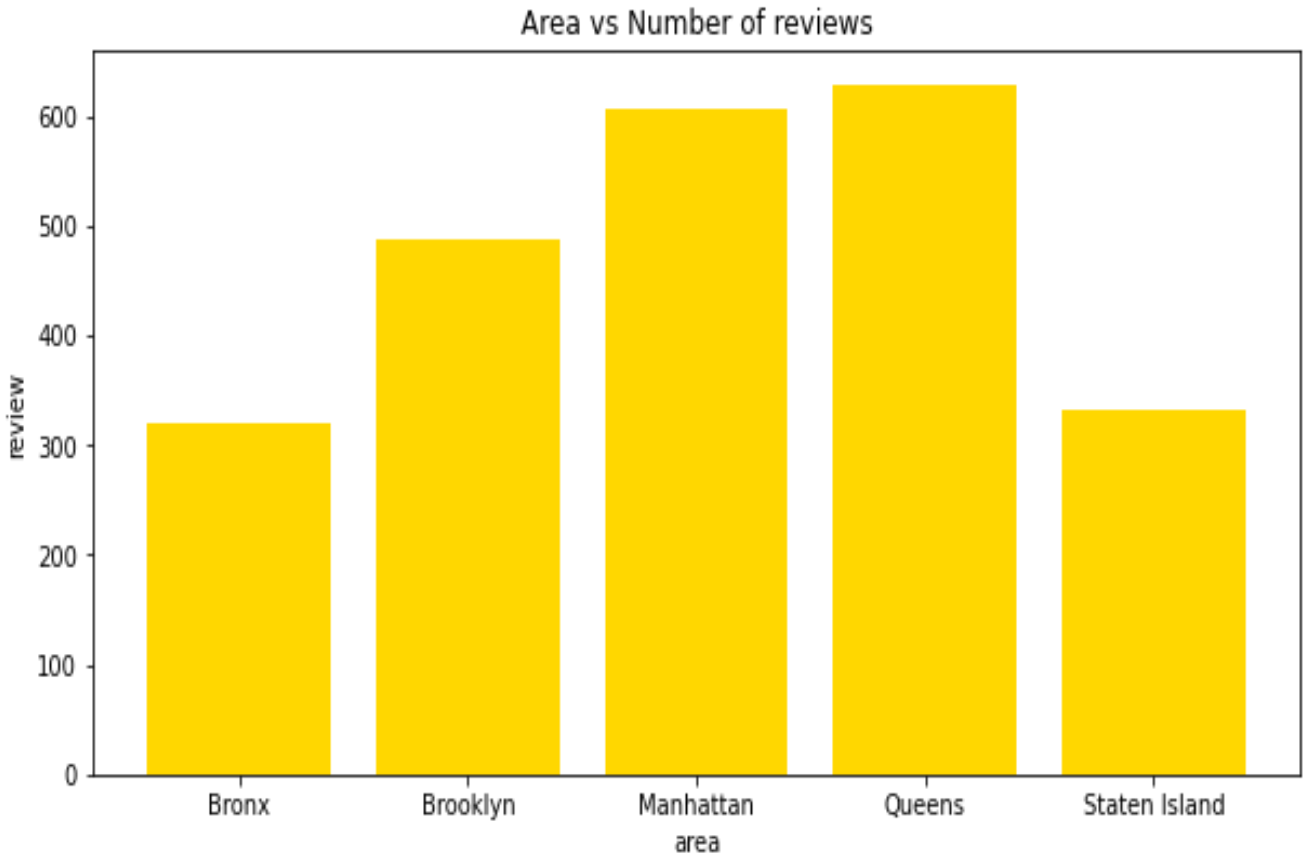
This table shows which host locality is most of the time chosen by guests. So Manhattan is the location that is chosen mostly means it's a popular place between guests. So we have to find out the main reason why these location hosts are most popular so we can make changes that are possible and provide guests more comfortable and attractive.

★ Most busiest location and why :



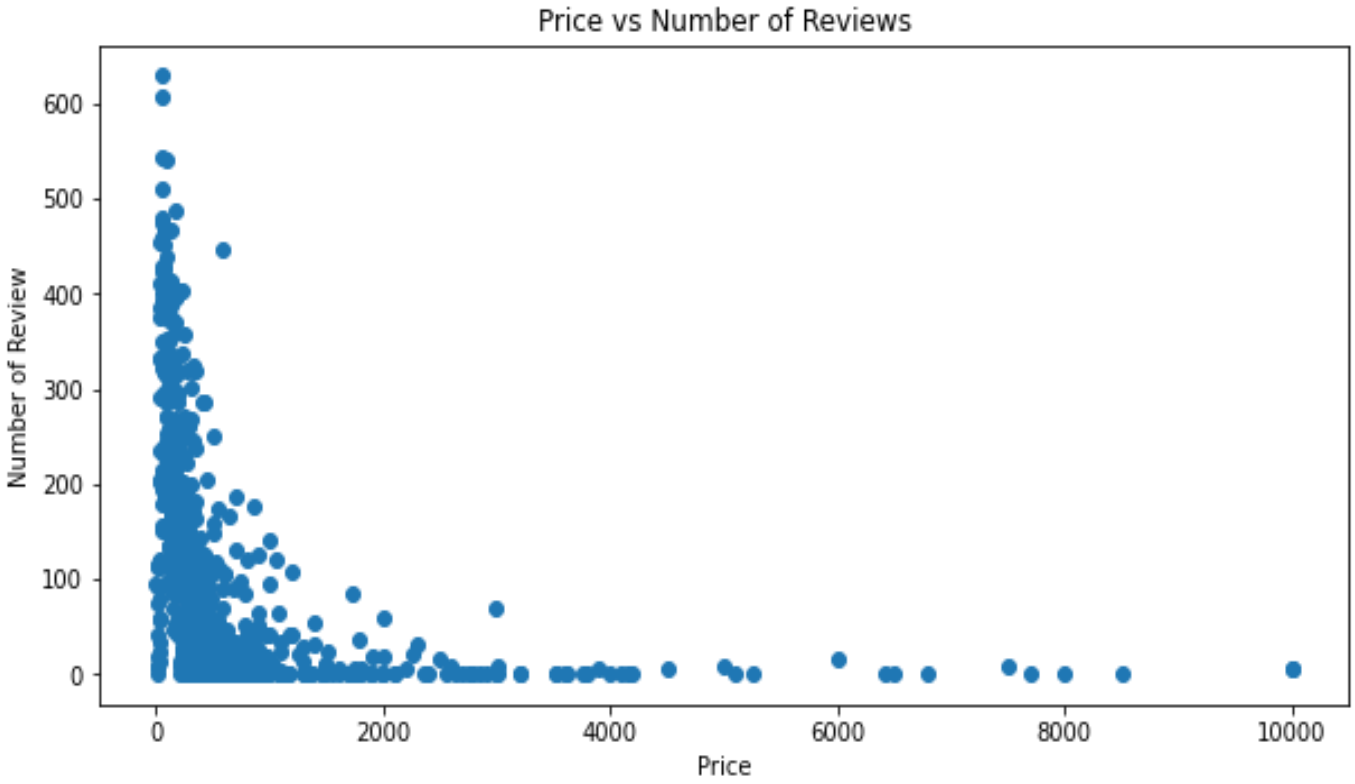
- As we can see, Dona, Jj, Maya, Carol, Danielle are the top 5 busiest hosts.
- According to the chart with the highest reviews, people choose them more.
- We have to make guests comfortable and make them secure so they give good reviews.
- Good reviews show that hosts have high expertise so they choose them more.
- So we try to make each and every review good so that’s helps in our business.

★ Area have highest number of reviews :



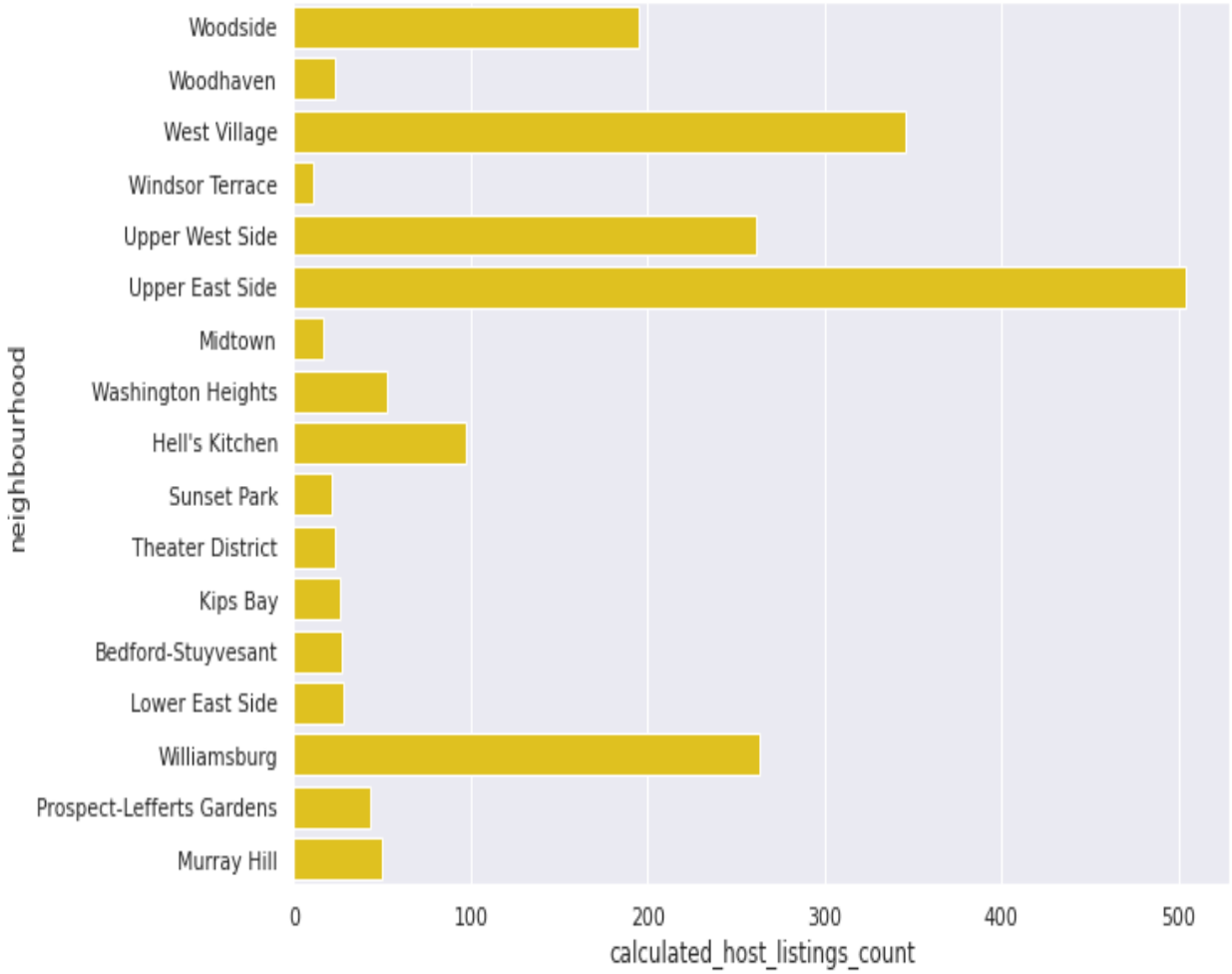
- As we see, the highest number of reviews comes from the Queens area which is 629.
- So if the number of reviews is highest people visit them most.
- In these top 10 Angela is the lowest one so we have to improve it.
- With this graph we can check which locality hosts have minimum reviews so we can correct them.

★ Price of host according to their number of reviews :



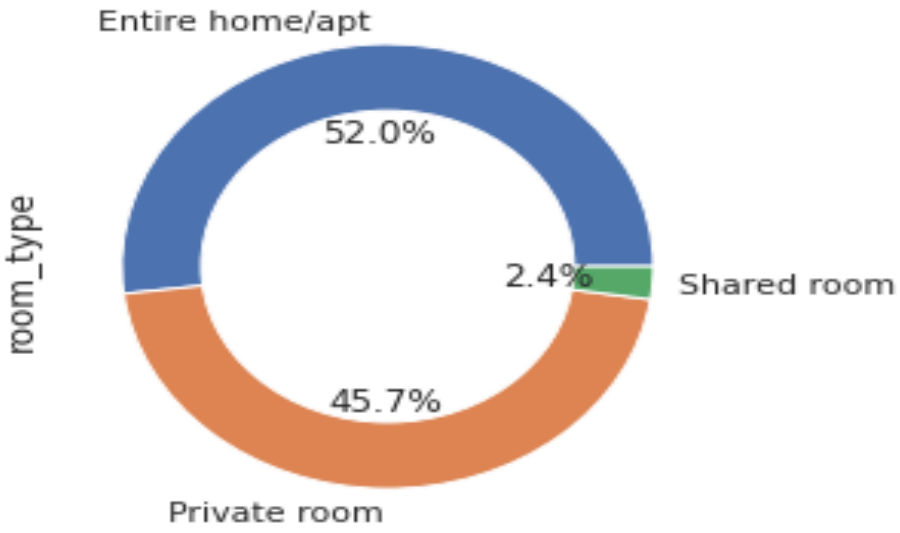
- Thai graphs show the pricing according to the number of reviews.
- According to this graph we can see that generally the price is less and where the price is less people reviewed most.
- So high reviews means expertise and guests list them more.
- So we try to make the price less for high pricing hosts so guests choose them also.

Distribution of listings across the neighborhood:



- With this chart we can say that most of the listings are located in the Upper East Side.
- So we have to check why Upper East Side is most listed and why others are not and make changes according to it.
- And in these top 10 Windsor Terrace have a minimum calculated host listing count so we can take steps to cover it.

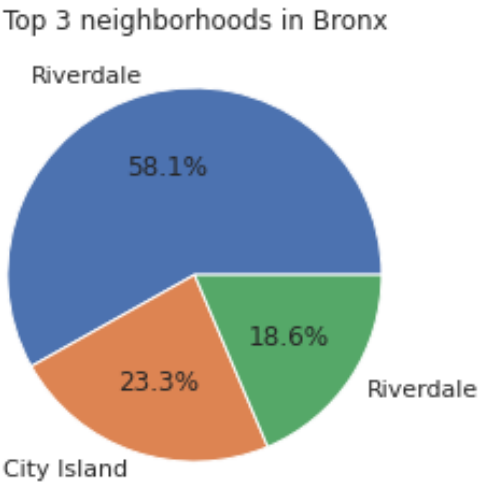
★ Room type distribution :

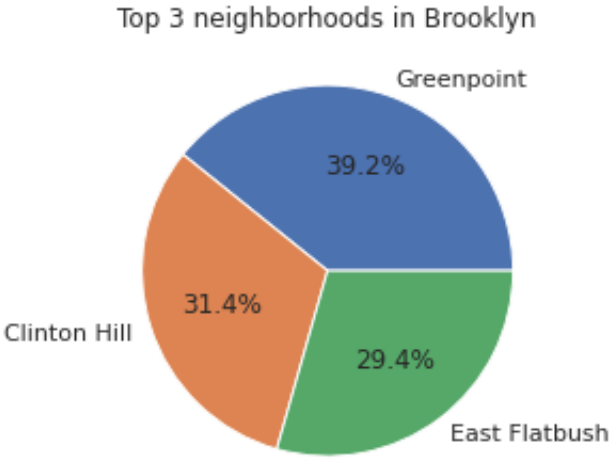


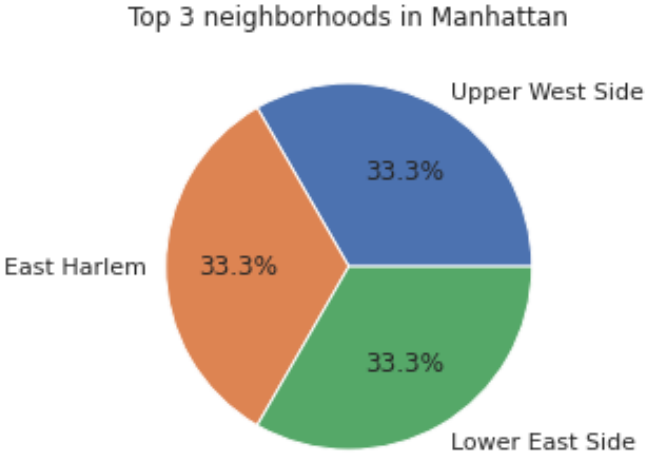
The "room types" column in Airbnb datasets provides information about the type of accommodation that is being listed, such as an entire apartment or a private room within a shared apartment. Analyzing the distribution of room types can provide insights into the preferences of hosts and guests in a particular market, as well as the availability and competition for different types of accommodations.

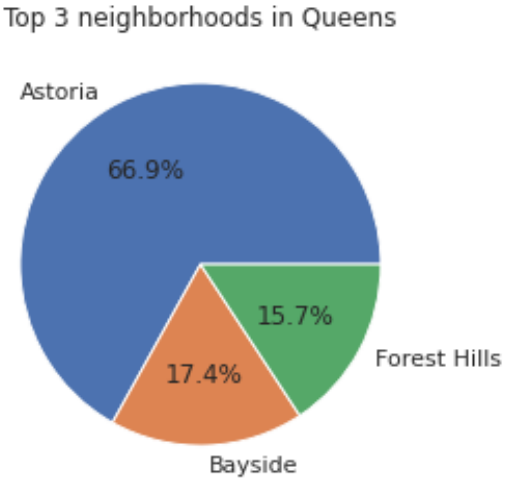
According to this graph we have 52% Entire home/apt, 45.7% Private room, and 2.4% Shared room. So this type of pattern shows that most of the guests prefer the entire home/apt more than any type so we have to increase the number of this room type at all hosts so bookings will increase.

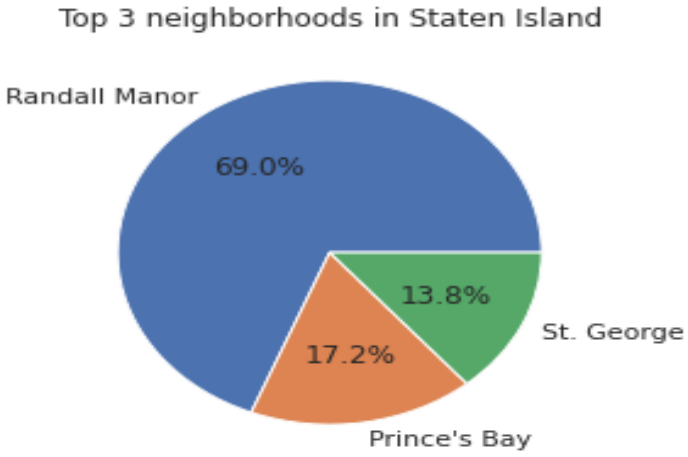
★ Grouping neighbor according to location :







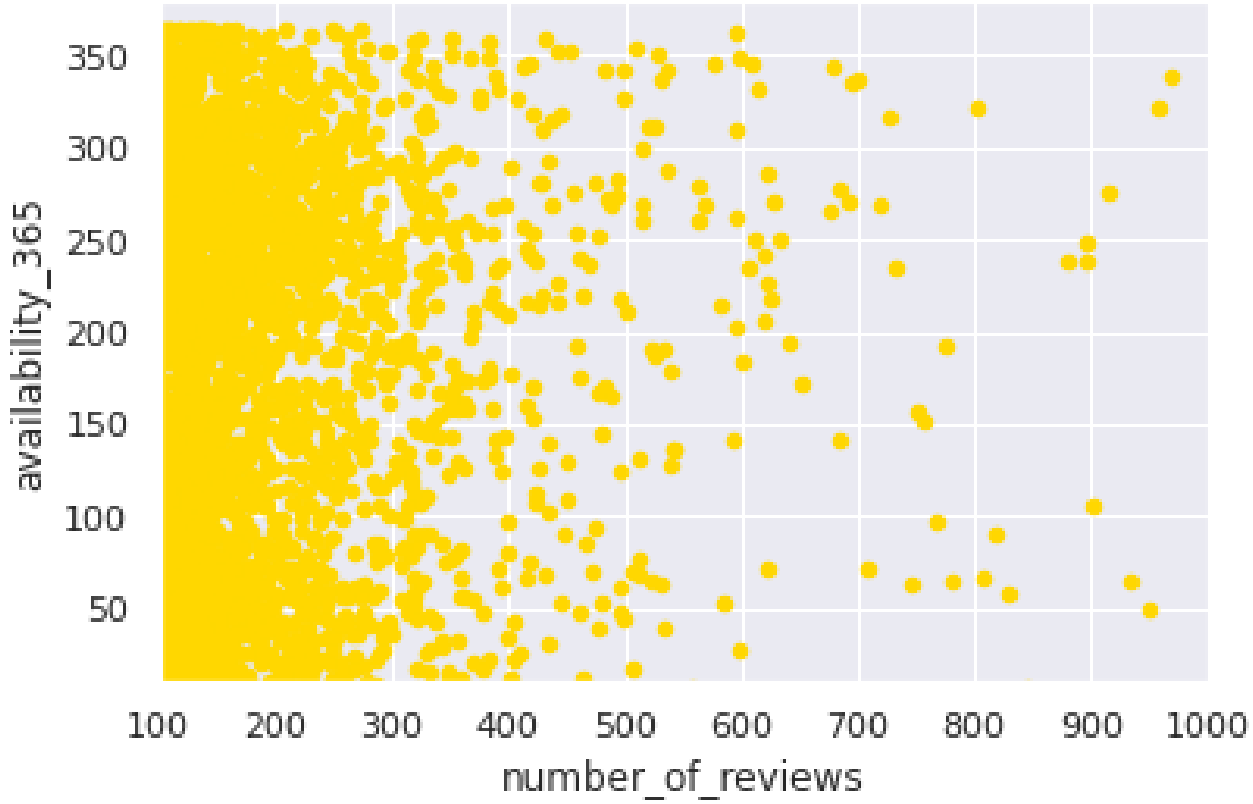




These graphs can provide insights into the geographical distribution of Airbnb listings in a particular area, as well as the availability and demand for accommodations in different parts of the city or region.
It can also provide information about the popularity of different neighborhoods among tourists or business travelers.
Overall, grouping neighborhoods by location can provide valuable insights for hosts and analysts looking to better understand the Airbnb market in a particular area, and can inform decisions about pricing, marketing, and location targeting.
As we can see, the top 3 neighborhoods in the Bronx (Riverdale, Riverdale, and City Island), top 3 neighborhoods in Brooklyn (Green point, Clinton Hill, and East Flatbush), and so on. So these basically indicate the best host in a particular locality and due to which some causes guests prefer them.

★ Most demanding host for Airbnb :

In this scatter plot graph we can analyze the most demanding host on the basis of it’s availability means that we clearly see from this graph that mostly guests reviewed those hosts whose availability is the most throughout the year and guests reviewed on many basis as we already seen in our last few graphs.
It's important to note that while high demand can be an indicator of success on Airbnb, it's not the only measure of success. Other factors, such as guest satisfaction and profitability, should also be considered when evaluating the performance of Airbnb hosts.



Endnotes :

- Manhattan is the most focused place in NYC for hosts to do their business.
- Customers pay the highest amount in Brooklyn, Queens and Manhattan (i.e. $10000 -$10).
- For the three types of room(Entire home/apt, private room and shared room) avg price is highest for Entire home/apt is 211.79, Private room is 89.78 and sharing room is 70.12.
- There are a total 1294 locations which are available for 365 days.

This information helps to find people's location where they can go anytime.

- Then we see how to best rating hotels according to their price. This will help in business from an improvement point of view that how we can improve so our rating, reviews, listing count will increase.
- And we see mostly people review those hosts who have less price so those hosts who don’t get enough reviews because may their price is too high make them less.
- Upper East Side part having maximum number of list count
- People choose Entire home/apt generally like so we have 52% of it, Private room 45.7% and Shared room 2.4 % so we have given more focus on the Entire home/apt option so our profit will increase.
- And we have to appreciate our premium customers like Gurpreet Singh
- And we also check from the group pie chart which host is doing well in a particular location.