EDA

Airbnb Booking Analysis

Data Science Pro

Almabetter

Bengaluru, Karnataka

Submitted By-

Ana khan

Group Members

Ana khan

Abhinav Singh

Tanay Tupe

Abstract

Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present a more unique, personalized way of experiencing the world. The Dataset we have used for exploration is NYC Airbnb dataset. Data analysis on listings provided through Airbnb is a crucial factor for the company. The purpose of this project is to explore Airbnb dataset, map the result clearly through visualization tools, and give new insight to the public and other relevant parties.The sole purpose is to analyze it which can further be used for security, business decisions, understanding of customers' and providers' (hosts) behavior, guiding marketing initiatives, implementation of innovative additional services and much more.

This dataset has around 49,000 observations in it with 16 columns and it is a mix between categorical and numeric values.

The dataset contains some very regular information about hosts like their ID and name. Along with this we have relevant information about the Airbnbs i.e. Property name , Room type, Neighbourhood, Availability, reviews, location and minimum stay.

Problem Statement

Exploratory Data Analysis helps to deep dive into the dataset as well as it turns the unusable data into usable dataset. Now let's get familiar with the dataset by diving deep into each column.

id: Unique ID for each Airbnb (Numeric)

name : Distinct name for each Airbnb, 16 null values.

host_id : Unique ID assigned to each Airbnb owner or host of the property(Univariate)
host_name : Name of the Airbnb Property owner, 21 null values(Univariate)
neighbourhood_group : Depicts the presence of different Airbnbs in 5 distinct neighbourhood groups. (Categorical)
neighbourhood : Depicts the presence of different Airbnbs in 221 distinct neighbourhood cities. (Categorical)
number_of_reviews Total number of reviews currently listing have. There are 1 million + reviews till date(Univariate)
latitude:"latitude" It represents the measurement of distance north or south of the Equator.(Continuous)
longitude:"longitude" It illustrates the measurement of east or west of the prime meridian.(Continuous)
Room_type: It has three different types of stay options.i. e. 'Entire home/Apt' , 'Private room' ,'Shared room'. (Categorical)
Minimum_nights: It illustrates minimum number of nights for booking a particular listing/Airbnb(Continuous)
price: Price details of all listings in US Dollars.(Univariate)
last_review : Date on which last review was received by property(Univariate)
reviews_per_month: Average number of reviews listing have per month(Univariate)
calculated_host_listings_count: Total counts of listings per host. (Continuous)
availability_365 : Number of days when listing is available for booking in the year.(Numeric)

Introduction

We have analysed the key features of the airbnb properties in order to get insights about the visitors and hosts of the properties. For better visualization we have got rid of all null values present in the dataset. We have analysed on the basis of two main EDA techniques i.e. Spatial Visualisation and User Review Mining. Spatial visualisation of the properties mainly depicts the relation between price and location of the airbnb. User review mining provides valuable information about the experience of the stay at airbnb, preference, and overall satisfaction of the visitor.

STEPS INVOLVED

1. Data Preparation And Cleaning:

Null Value Treatment: There are 4 columns (name, host_name, last_review, number_of_review) with null values.

There are certain columns which are not relevant for EDA, therefore it is best to delete them.

We have removed the “host_name” column which is the name of the owner of the airbnb property and we have also deleted the “last_review” column as it is in the date format and it provides the date when the airbnb was last reviewed. So it is also not relevant for our EDA.

Replace Null Value: We have replaced null values in the “name” column with “Not_available” to make it presentable. Along with this we have replaced null values in the “reviews_per_month” with “0”(numeric zero) to avoid any unprecedented error.

Now our data is cleaned of any Null values.

2. EDA and Data Visualization

We have divided our analysis into two parts. -

1. Spatial Analysis
2. User Review Mining

Spatial Analysis:

Analysing different neighbourhoods and their price analysis:

To begin with we have sorted out the data and found out the top 10 most popular neighbourhood cities. The list goes like this:

1. Williamsburg
2. Bedford-Stuyvesant
3. Harlem
4. Bushwick
5. Upper West side
6. Hell’s Kitchen

7. East Village
8. Upper East Side
9. Crown Heights
10. Midtown

Highest number of airbnbs are located in and around Williamsburg, Bedford-Stuyvesant and Harlem. Now if we compare Neighbourhood cities on the basis of average price. We have observed that the top 3 neighbourhood cities offer lowest average prices that could be a possible reason for their wide popularity.

Analysing the type of listings:

There are three types of rooms available throughout all the airbnbs in New York.

1. Entire home/Apartment
2. Private Room
3. Shared Room

After analysing all the room types available throughout all the neighbourhood groups.Three major observations are ::

- Entire Home/ Apartment type is a highly demanded room type among all the available options.
- Private room types are preferred over shared room types. Shared room types are least preferred.
- It highlights that most visitors who opt for airbnbs are mainly families or couples.
- Among all the five neighbourhood groups “Manhattan” has the maximum number of airbnbs of almost all the three types. Brooklyn stood at 2nd position when it comes to the presence of airbnbs.

Price Analysis for types of listings:

We plotted average prices of the three types of listings available for all the 5 neighbourhood groups present. The major observations are :

- “Manhattan” has the highest average price for all the three room types available. Entire home/Apt type has the average price approx 250 bucks, around 117 bucks for private rooms, and around 89 bucks for shared room.
- “Queens” and “Brooklyn” have almost the same average price for private rooms and “Staten Island” and “Bronx” have a minute difference between their private room average price.
- “Staten Island” and “Brooklyn” have a very minute difference in prices for the entire home/ apt room type.
- “Bronx” has the lowest average price for the entire home/ apt room type, “Staten Island” has the lowest average price for the private room type, and “Brooklyn” has the lowest average price for the shared room type.

Analysing Relation Between Minimum Nights and Number of Reviews:

We have plotted the airbnbs over the New York map in order to get a clearer picture of the presence of airbnbs.
Here, “Manhattan” has the highest number of airbnbs and if we compare area wise then Manhattan is the second largest. Therefore, we can say that the ratio of number of airbnbs to area for the Manhattan is the highest whereas the same is the lowest for “Staten Island”

Next to this we have plotted a scatter plot of the number of reviews less than 10 over the NYC map. Here we have formed a relation between minimum_nights and number_of_reviews using a scatter plot. The main

purpose was to find how the limitation on minimum stay affects the reviews of the visitors.

We have chosen to plot it below 10 because 75% of the number of reviews are less than 5, which clearly depicts that most of the visitors have reviewed in this range. After observing the graph we can say that the number of reviews for the airbnbs are more in the range of (0-2) majorly located in Manhattan and Brooklyn as the area is dark purple for these neighbourhood groups.

So, we come to know that airbnbs with no minimum stay or zero minimum stay have received the maximum number of reviews. We can clearly say that as the number of minimum nights increases the number of reviews for the property falls proportionately.

Insights about the host of the property

Now we will get some insights on the host of the airbnb. In this dataset we only have the host_id of the owner(host_name removed). So we have compared and found the top ten hosts of the airbnb. Out of all the hosts, host id 219517861 owned the maximum number of airbnbs in NYC and it owned 26% listings of the top ten listings.

Availability of Airbnbs throughout the year for different Neighbourhood Groups:

We have taken the average availability in terms of the number of days for different 5 neighbourhood groups. If we compare each group on the basis of mean room availability, then we can say that in Staten Island rooms are available for more than 200 days in a year, which is the highest of all. On second position we have the Bronx, wherein average room availability is more than 150 days. Next below that we have Queens, which have average room

availability of 100 days. Also, Manhattan and Brooklyn have low availability of rooms, which is below 50.

Analysing the relation between “INT” and “FLOAT” values:

Notes:

Sr. No	Correlation Range	Relation
1	-0.1 to +0.1	Very Weak linear relation
2	-0.2 to +0.2	Weak linear relation
3	0	No relation

By observing the heat map we can say that:

- The variable “price” has a very weak but positive correlation with “minimum_nights”, “calculated_host_listings_counts” and “availability_365”. It has negative and weak correlation with ” number_of_reviews” and ” reviews_per_month”
- We can see that there is a positive but low correlation between reviews_per_month and number_of_reviews of around 0.5 which is obvious as number of reviews are much related and dependent upon number of reviews in a month.

CONCLUSION:

- Among all the neighbourhood groups, Manhattan has the highest number of airbnbs and bookings. It is the most demanding destination, which provides the reason for it’s high average price. . Following it, at

second position we have Brooklyn. It has the second highest number of airbnbs.

- As “WIlliamsburg” and “Bedford-Stuyvesant” offer airbnbs at reasonably low prices, they are the most popular neighbourhoods.
- 98% bookings are done in entire home/apt type airbnbs, which possibly means that most of the visitors book airbnbs for family outings and they value personal space and time. Very few people book shared rooms, which shows that people who are looking for budget friendly rooms do not go for airbnb, the reason for this could be anything like the condition of the room or sense of security.
- Out of 49k + bookings , 13k bookings are done in airbnbs where minimum nights is 1 or zero. This shows that visitors prefer to stay for a shorter period of time and top of that properties with less number of minimum nights like 0 or 1 have received the maximum number of reviews.

Future Prospects:

- In Manhattan the number of airbnbs having private rooms can be increased as the presence of the entire home/apt type is way more than private rooms. Customers will have a choice to make between the two types and the price for an entire home type airbnb is very high if we compare it with a private room. So, it is likely that visitors will prefer private rooms because customers who don’t want to pay high but still want space then they can go with private rooms in Manhattan.
- Number of minimum nights can be increased to some extent in demanding locations and also in most popular neighbourhood cities i.e. “WIlliamsburg” and “Bedford-Stuyvesant” in order to increase revenue.