Exploratory Data Analysis on AirBnB NYC 2019

Smriti Bhattrai, Aman Gupta, Kiran Mahara
Data science trainees,
AlmaBetter, Bangalore

Abstract:

AirBnB is an online platform that allows property owners to list their place as holiday Accommodation and allows travelers to find a place to stay while they are away from home.

On the website, homeowners can create a listing for their property and that listing will include a written description, photos and a list of amenities, as well as information about the local area.Travellers can use filters to search for holiday accommodation that's right for them – such as the number of bedrooms, location and price.

1.Problem Statement

This data set contains different variables defining the booking and stay at a particular host along with its pricing, room types . We are trying to analyze the data to find insights into what goes into a booking and also identify the factors that could influence this process the most**.**

Attributes of data

- **Id** -Unique identifier for each row.
- **Name** - Listing name which user sees while booking (hotel name).
- **Host_Id** - Every host in Airbnb gets an unique id, host id represents that id only.
- **Host_name** - Name of the host.
- **Neighbourhood_group -** Whole New York is divided into 5 boroughs(regionally) This feature shows in which borough a particular listing is located.
- **Neighborhood** - Each borough is further subdivided into neighborhoods. This feature shows in which neighborhood a particular listing is located.
- **Latitude and longitude** -These shows the geographical location of a listing.
- **Room_type** - The type of room rental is offered (private, shared , entire home).
- **Price** - Price per night that a rental is charging.
- **Minimum_nights** - Minimum number of nights for which rental can be booked.
- **Number_of_reviews** - Number of reviews rental has got till now.
- **Reviews_per_month** - An average number of reviews a rental has got till now.
- **Last_review** - Date on which a rental got its last review.

- **Calculated_host_listings_count** - Number of rentals a host is hosting in the dataset.

- **Availability_365 -** Number of days for which a rental is available to book.

2. Introduction

Airbnb, as in "Air Bed and Breakfast," is a service that lets property owners rent out their spaces to travelers looking for a place to stay. Travelers can rent a space for multiple people to share, a shared space with private rooms, or the entire property for themselves.Airbnb was started in 2008 by Brian Chesky and Joe Gebbia, two industrial designers.

3. Reasons for Analysis

1. The purpose of this analysis was to gather the key features which are relevant for the airbnb business growth.
2. Distribution of hosts in different neighborhoods in different Neighborhood groups.
3. To study and analyze different types of room type, availability and top host i.e., hosting most in that neighborhood.
4. Understanding about the price distribution in different boroughs.

4. Steps involved:

- **Data Wrangling.**
  We have loaded the dataset from our drive and ran a basic info commands like head, tail info, describe, etc to get the basic information of data. We used the shape command to get the number of rows and columns in our dataset

- **Null values Treatment**
  Our dataset contains a large number of null values in two columns last_review and reviews_per_month and that is related to the number of reviews so we replaced all those null values with zero in spite of deleting them as both the columns are not contributing enough in analysis and deleting the rows only lead to loss of data.
  .

- **Outlier Treatment**
  We haven't removed or treated the outliers here as we are not building a model. Instead while analyzing the data we kept the thresholds to keep outliers away and visualize the data properly and maintain the originality of data at same time.

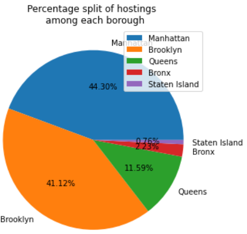
- **Exploratory Data Analysis**
  In EDA we have used different kinds of plots such as bar plot, distribution plots, etc and various other exploration techniques to bring out the best insight out of the data.

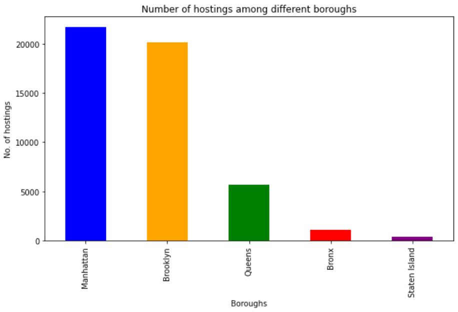
5. EDA

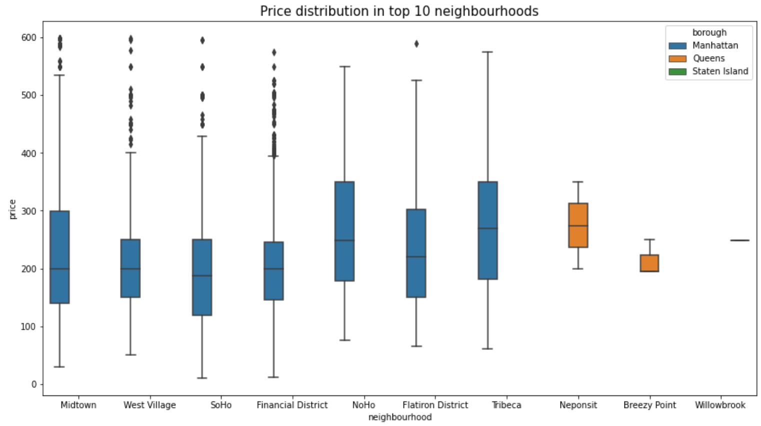
In order to go through the whole data and dig out as much information as possible we used a approach of dividing whole analysis in a kind of different modules and then going through each thoroughly. We played around different visualization techniques including different kind of charts including bar plots, pie charts, distribution plots, geo plots, etc. Our analysis is divided into following modules:

1. Exploring about different areas (neighborhood and boroughs.

In this section we completely analyzed different boroughs (neighborhood groups) and neighborhoods. We used different kinds of plots like distribution plots to check distribution of prices among different boroughs, bar plots to check for top neighborhoods, number of rentals in each borough and Some of them are shown below.

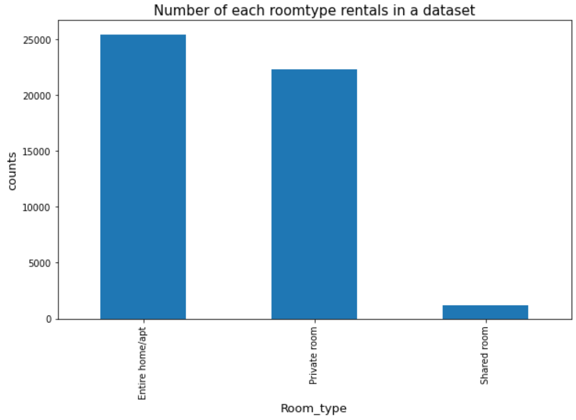


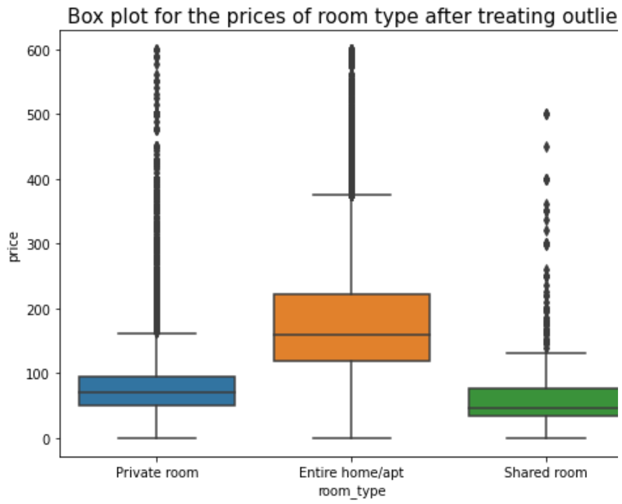




2. Visualizing abou different room types

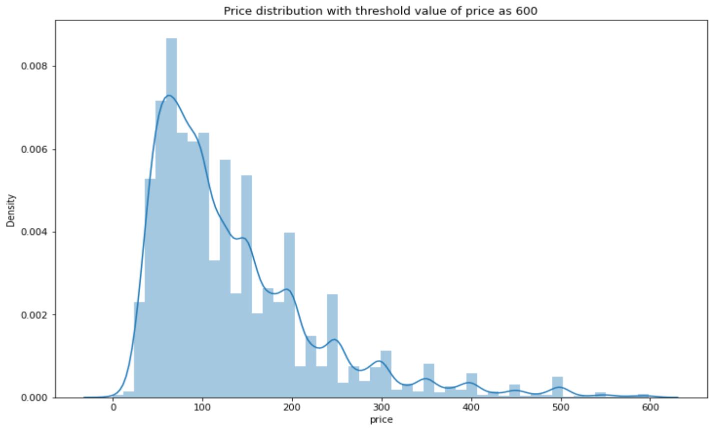
We again analyzed room type with respect to different features like borough, prices ,etc using different kind of charts and able to understand that which room type is most costlier, how they are arranged among different boroughs, etc. and able to grab some of key points like entire homes and private rooms are most in demand and also available in abundance.

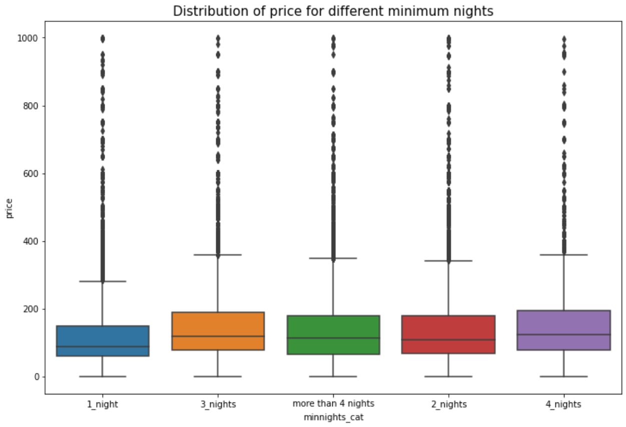




3. Analyzing about prices.

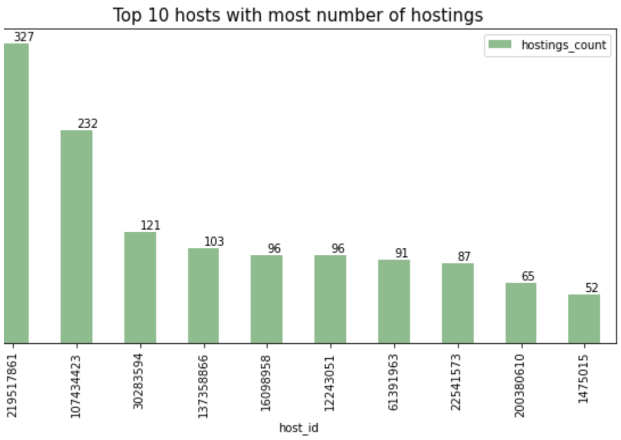
In this sector we checked how price is distributed among whole dataset irrespective of boroughs, we used scatter plot to check is there any visible relation between number of reviews and hotels. We did categorical conversion of availability_365 and minimum night to get any possible relation between price and them using box plot.

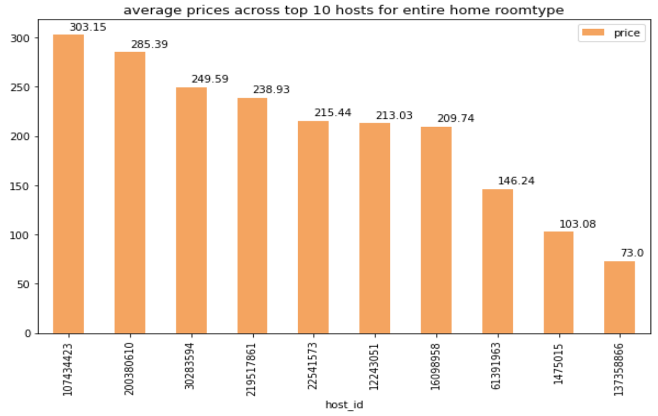




4. Analyzing about different hosts.

We selected top 10 host, they have been selected in terms of number of hosting they have and then further analyzed other features of these top hosts to see how they differ from the rest of the other like on an average what is the price they are offering as compared to average prices we have, their Boroughs, etc.



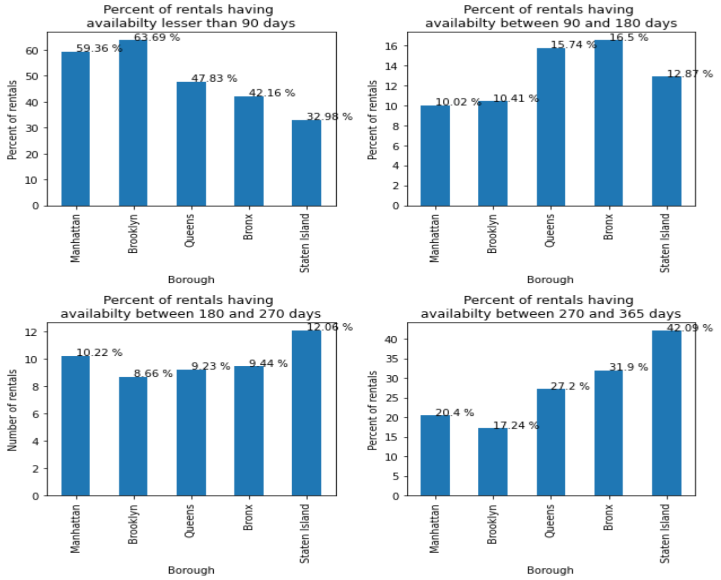


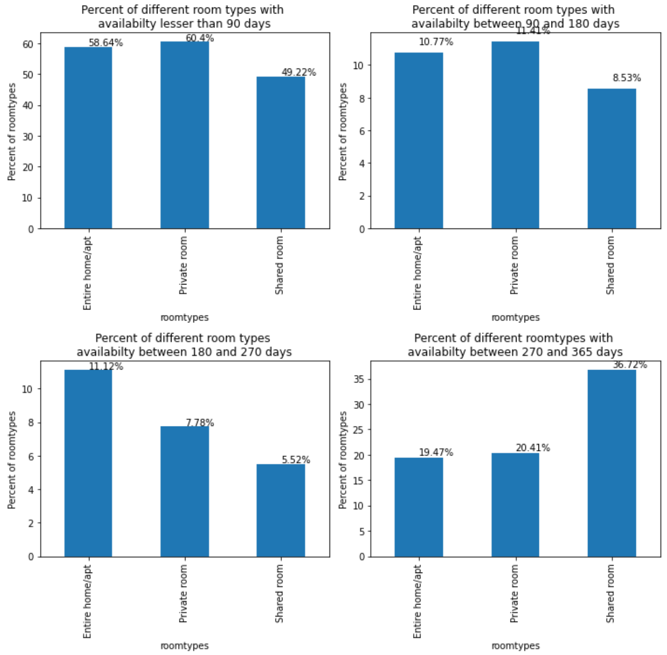
5. Analyzing about availability_365

Availability_365 depicts the number of days in a year property is available to rent out.For better visualization we categorized the feature into 4 categories

1. less than 90 days
2. between 90 and 180 days
3. between 180 and 270 days
4. between 270 and 365 days

We then used bar plots to visualize the percent of total rentals in each borough available in these categories.





6. Conclusion.

1. Manhattan and Brooklyn attract the most tourists and that’s why they have the most rentals. So the hosts looking for the best borough to have rental should go for Manhattan and Brooklyn.
2. Customers respect privacy and mostly prefer private rooms and apartments. So there is very minute availability of shared rooms as compared to private and entire homes.
3. More than half of the rentals of Manhattan and

Brooklyn have availability less than 90 days. This suggests that Manhattan and Brooklyn rentals are heavily booked.

4. Generally rentals with a minimum one night's offering have lesser prices .
5. Rentals whose availability is less than 90 days are generally the ones with lower price.
6. Manhattan is the most expensive borough in each of the room types.
7. Entire homes are one of the most available rentals in New York City whereas shared rooms are very less preferable.
8. For the three room types, the average price of an entire home is around $150, for shared room is around 50 Dollar, and for private room is around 75 Dollar.
9. Most hosts only have one rental but at the same time there is a host with host ID 219517861 manages 327 rentals and is one of the busiest hosts. He has all his rentals located in Manhattan.
10. Price seems to be very lessly relatable with the number of reviews although it seems out that one with more number of reviews are having lesser prices.

References-

Google
GeeksforGeeks
Analytics Vidhya