Exploratory Data Analysis of NYC Airbnb Dataset, 2019

Rupali Dawkore
Data science trainees, AlmaBetter, Bangalore

Abstract

Airbnb, Inc based in San Francisco, California, operates an online marketplace focused on short-term homestays and experiences. Airbnb belongs to the sharing economy, which offers a unique approach to accommodation. The owner of the home can rent it out as a place to stay. It has seen explosive growth over the last decade. People prefer renting Airbnb properties because they feel more at home and are more cost- effective than staying in a hotel, and this is why Airbnb is so popular in New York.

Keywords: *graphical analysis, correlation, descriptive statistic, central tendency of measure*

1. Problem Statement

Ultimately, the project explores what factors can affect listing prices, which is a question that both Airbnb users and hosts would care most about. In this project, we will analyze the data descriptively and statistically to determine how the variables are correlated to generate hypotheses useful for future decision-making.

It's imperative to analyze the data carefully to obtain meaningful insights that can assist in making better business decisions and understanding customer and host behavior.

id: Unique identifier for each row name: Name of listings on Airbnb

host_id: Unique identifier for each host host_name: Identifies the name of the host neighbourhood_group: It has 5 unique values, namely, Brooklyn, Manhattan, Queens, Staten Island and Bronx. neighborhood: There are 221 unique neighborhoods in this column.

latitude: Contains the latitudinal information of the host's listing.

longitude: Contains the longitudinal information of the host's listing.

room_type: Specifies the different types of room.

minimum_nights: Talks about the minimum number of nights stay.

number_of_reviews: Tells us about how many visitors reviewed the listing. reviews_per_month: Rate of review per month calculated_host_listings_count: Number of listings on Airbnb by a particular host. availability_365: Availability of a particular listing throughout the year.

2. Introduction

Airbnb connects people who need a place to stay with those who have a place to rent. Different variables play a large role in determining the price of a location. It is the host's responsibility to list a fair price for their accommodations. Those seeking lodging assess listings based on a variety of characteristics, including size, location, amenities, and—most importantly—price. This dataset describes the listing activity and metrics of NYC Airbnb in 2019. It contains all the information needed to make forecasts and draw inferences about NYC hosts, costs, and geographical accessibility. Our data, which consists of 48,895 rows and 16 columns,will be explained in the next section.

This assignment uses New York City Airbnb Open Data from Kaggle.This website serves as the original source of this Airbnb open data set. To work on data, we will be using different tools that are very common for performing simple and complex analyses, like classifications of variables, histograms, textual mining, and measures of central tendency.

3. Data Sets:

All the data is sourced from Inside Airbnb, which hosts public available data from Airbnb.

The two main data sets:

• Numerical:

Some numerical attributes used in the project are id, host_id, latitude, longitude,minimum_nights, calculated_host_listings_count, availability_365, number_of_reviews among others.

• Categorical:

Some categorical attributes used in the Projects are neighbourhood_group, neighborhood and room_type.

4. Steps involved:

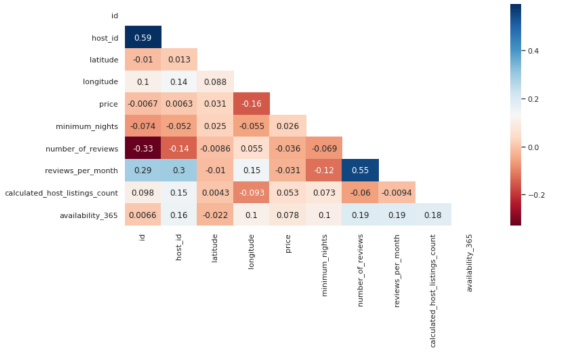
Exploratory Data Analysis *:* Once we loaded the data, we compared our target variable, 'price', with other independent variables. As a result of this process, we were able to identify various aspects and relationships between the target and independent variables. This gave us a better idea of how features behave when compared to target variables.

Data Cleaning *:* There are certain columns with a lot of null values. We can see that this dataset has around 48895 observations in it with 16 columns and it is a mix of categorical and numeric values. We can also see from the above, there are some missing NaN values that will require cleaning and handling.

EDA *:* We performed Data Analysis and Visualizations on the selected features to gain insights in the data and present answers to the objectives defined.

Feature Selection *:* In this part, we selected the columns which are most relevant for the objectives defined at the beginning of the EDA.

5. Correlation between different variables



Using the color variation, we can depict the correlation between the variables on both axes. On either side of the axis, darker colors indicate a negative correlation between both variables.

6. Data and methodology

The analysis of the NYC Airbnb dataset has been conducted mainly using the primary tool as EDA. The researched data set contains activities within this business platform for the year 2019. Primarily, it contains information about the name of the listing along with its latitude and longitude, the name of the listing's host, the room type, the price in dollars, the minimum number of nights spent here, the number of reviews, and the number of days available.

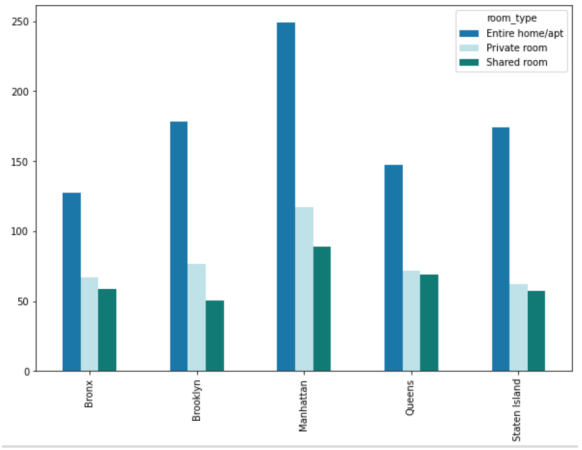
The analyzed dataset has around 48895 columns, 44% of them, located in Manhattan, Brooklyn 41%, 15% in other parts of New York City.

The percentage division of the three different room types across the region, with 'Entire home/apt' accounting for 62.78% of listings, private room with 35.03% and shared rooms representing just 2.19%.

6.1 Graphical analysis of the data set

A graphical analysis is usually the first step in EDA. In graphic data analysis, data is visualized using various types of graphs.

1. The average preferred price on the Airbnb platform by the customers on a particular location



The image above displays a bar graph of listings in Airbnb New York City broken down into the neighborhood_group where it is available. The average preferred price on the Airbnb platform by the customers on a particular location and room_type.

Bronx : Entire home/apt 378, Private room 652, Shared room 59.

Brooklyn : Entire home/apt 9553, Private room 10123, Shared room 413.

Manhattan : Entire home/apt 13190, Private room 7973, Shared room 480.

Queens : Entire home/apt 2096, Private room 3370, Shared room 198.

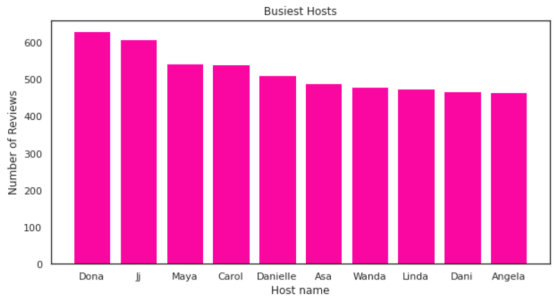
Staten Island : Entire home/apt 176, Private room 188, Shared room 9.

2. Which hosts are the busiest and why

In this airbnb dataset we identified:-
Total_no_of_host_id:48858
Total_no_of_host_name:48858
Total_unique_host_name:11450
Total_unique_host_id:37425

	host_name	room_type	number_of_reviews
3434	Dona	Private room	629
6332	Jj	Private room	607
8977	Maya	Private room	543
2164	Carol	Private room	540
2975	Danielle	Private room	510
1214	Asa	Entire home/apt	488
13839	Wanda	Private room	480
7902	Linda	Private room	474
2947	Dani	Entire home/apt	467
863	Angela	Private room	466

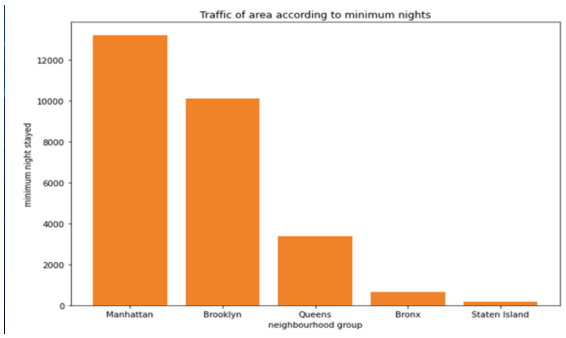
According to their reviews count. These hosts are having more reviews because they have a listing of private rooms and entire apartments and people used to stay more in private rooms and apartments.



From the above graph we can say that the busiest host are:

Dona , Jj, Maya , Carol and Danielle According to their reviews count. These hosts are having more reviews because they have a listing of private rooms and entire apartments and people used to stay more in private rooms and apartments .

3. Noticable difference between different areas

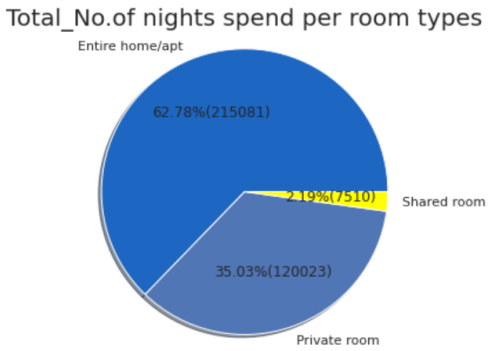


From the above analysis we can see that the minimum nights spent by the people are in Manhattan in entire home/apartment.We can also say that

1. *Manhattan*
2. *Brooklyn*
3. *Queens*

The traffic is huge because people prefer to stay in a private room and entire home/apartment.We can clearly see that these neighborhoods are providing more private rooms and the entire home/apartment.

4. The total of nights spends per room types by the customers



The three main room kinds are clearly segmented by proportion across the region, with "Entire home/apt" accounting for 62.78% of postings and "Shared rooms" accounting for just 2.19%.

6.2 Statistical characteristics of the data set

Quantifying the fundamental statistical properties of the data collection is another aspect of the EDA. Particularly among these are measurements of location (mean, minimum, maximum, median, mode, and quartiles), measures of variability (variance, standard deviation, and coefficient of variation), and measures of shape (skewness, sharpness). The researcher can spend more or less effort quantifying these values in each of the three statistical tools they have chosen.

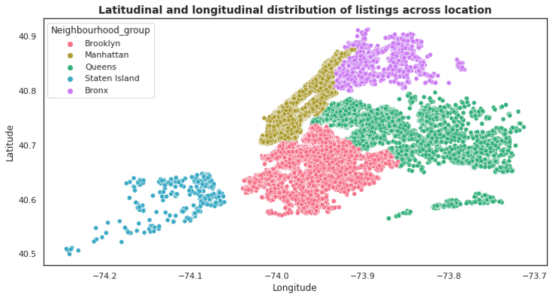
	neighbourhood_group	median_price	mean_price	min_price	max_price
0	Bronx	65.0	87.469238	0	2500
1	Brooklyn	90.0	124.410523	0	10000
2	Manhattan	150.0	196.897473	0	10000
3	Queens	75.0	99.536017	10	10000
4	Staten Island	75.0	114.812332	13	5000

With an average price of **150 dollars, Manhattan** has the **highest price range** and is the most expensive, followed by **Brooklyn** with an average of **90 dollars.** It is very obvious that Manhattan is one of the most expensive places in the world. **Bronx** has the **cheapest listings** of all with an average of **65 dollars.**

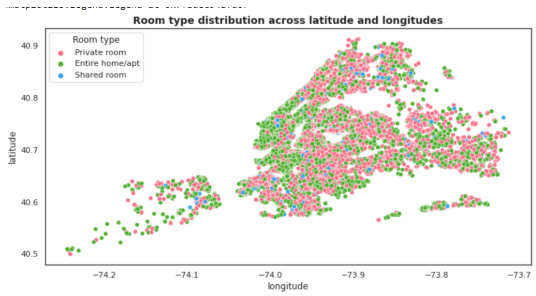
There are some **outliers** in the data, which causes the mean and median values to vary. It was discovered when looking at the price that there is a **substantial difference between the median and mean values.** The median is much smaller than the mean as a result. This is most likely a result of several outliers with much higher prices for Manhattan and Brookline that had an impact on the average.

6.3 Geographical analysis of the data set

1. ***Mapping neighborhood group***



2. Mapping neighborhood group



In terms of the distribution of room types, we can see there is a good mix of different types available across the region. When compared with shared rooms, there is dominance in private rooms and entire homes categories.

7. Conclusion:

• Our top ten hosts have a substantial number of listings, with the top host having over 300.
• The listings are dispersed among the five boroughs of New York City, with Manhattan having the most percentage and Brooklyn and Staten Island having the lowest.
• These were the three distinct room types' percentage distributions: Home or apartment as a whole: 62.78%; private room: 35.03%; shared room: 2.19%.
• According to the analysis of client needs, they strongly choose an entire home or an apartment. Offering these shared rooms carries the greatest risk of losing customers.
• Manhattan and Brooklyn are the two distinguished, expensive & posh areas of NY
• A statistical analysis shows that Manhattan has the most expensive price range with an average of 150 dollars followed by Brooklyn with an average of 90 dollars.
• There is a substantial difference between the median and mean values of the prices. The median is much smaller than the mean as a result. This is most likely a result of several outliers with much higher prices for Manhattan and Brookline that had an impact on the average.
• The Bronx provides the cheapest accommodation among all.
Though location of property has a high relation on deciding its price, but a property in popular location doesn't mean it will stay occupied in most of the time.