Exploratory Data Analysis of NYC Airbnb Dataset, 2019

Naghma Firdous
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 analyse the data descriptively and statistically to determine how the variables are correlated to generate hypotheses useful for future decision-making. It's imperative to analyse the data carefully to obtain meaningful insights that can assist in making better business decisions and understanding customer and host behaviour.

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.
neighbourhood: There are 221 unique neighbourhoods in this column.
latitude: Contains the latitudinal information of host's listing.
longitude: Contains the longitudinal information of host's listing.
room_type: Specifies the different types of room.
minimum_nights: Talks about the minimum number of night stay.
number_of_reviews: Tells us about how many visitors reviewed the listing.
last_review: Contains the date of the latest review for that 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 project are neighbourhood_group, neighbourhood and room_type.

4. Steps involved:

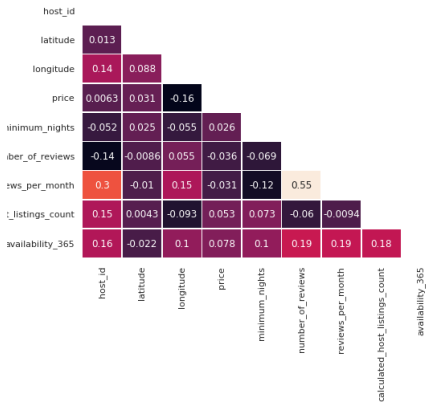
Exploratory Data Analysis : Once we loaded the data, we compared our target variable, 'price', with other independent variables. As a result of this process, we were able to identify various aspects and relationships between the target and independent variables. This gave us a better idea of how features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

Data Cleaning : There are certain columns with 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 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 analysed 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 52.3% of listings, private room with 45.5% and shared rooms representing just 2.2%.

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.

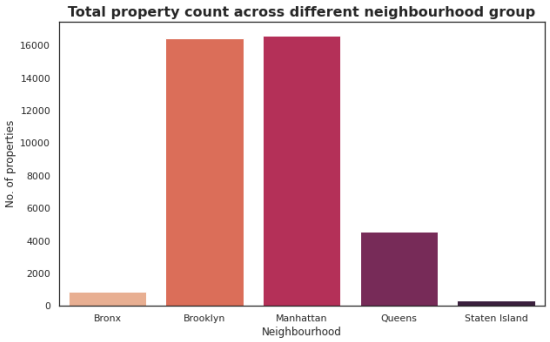


Fig: Top property count across different neighbourhoods

The image above displays a bar graph of listings in Airbnb New York City broken down into the neighbourhood where it is available. This graph makes it quite evident that the **most sought-after locations** on Airbnb in New York City are **Manhattan** and **Brooklyn**. The **Bronx** and **Staten Island's** lodging options are quite limited.

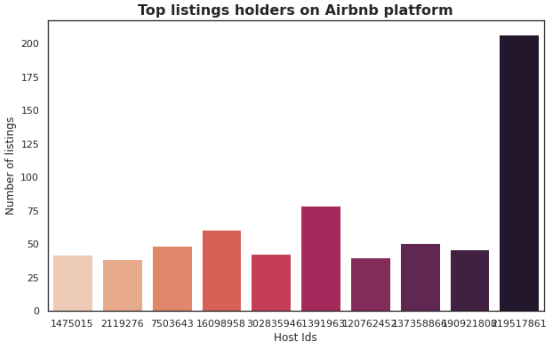


Fig: Top Listing holders on Airbnb

We can observe that the *top ten hosts have a sizable number of listings* on the Airbnb website. More than **300 properties** are listed on the platform under the **first host**.

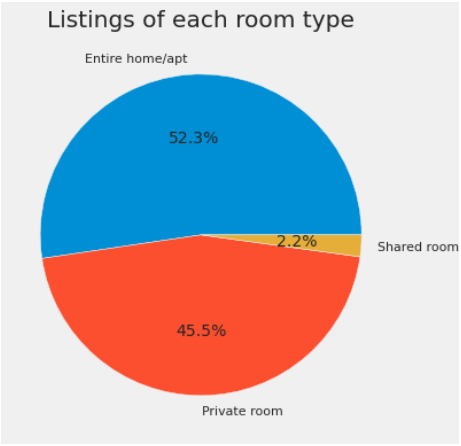


Fig: Percentage of room distribution

The three main room kinds are clearly segmented by proportion across the region, with "**Entire home/apt**" accounting for **52.3%** of postings and "**Shared rooms**" accounting for just **2.2%.**

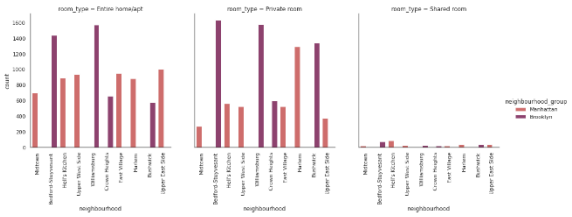


Fig: Top 10 neighbourhoods with maximum listings

One thing to note is that there are just Manhattan and Brooklyn included among the top ten boroughs. **Williamsburg** and **Bedford-Stuyvesant** are more *well-known neighbourhoods in Brooklyn*. The fact that "Shared rooms" barely cracks the top 10 neighbourhoods is astonishing.

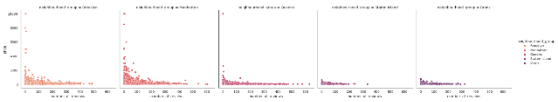


Fig: Analyzing the number of reviews and room availability with respect to price

We can see a negative relationship between price and the number of reviews. There are more reviews for properties with lower prices since they are booked more frequently

A stacked bar graph is one particular type of bar graph that can give us a great deal of

information. In this graph, for instance, the share of each form of lodging is displayed along with the type of accommodations offered by Airbnb in various areas of New York City. For instance, room rentals are more common in the Bronx than complete houses or flats in Manhattan. Manhattan has the most shared rooms, however compared to other types of lodging, this type is the least prevalent there. Renting entire apartments or rooms, which is not the case in other sections of New York, is much less popular than renting private rooms in the Bronx.

6.2 Geographical analysis of the data set

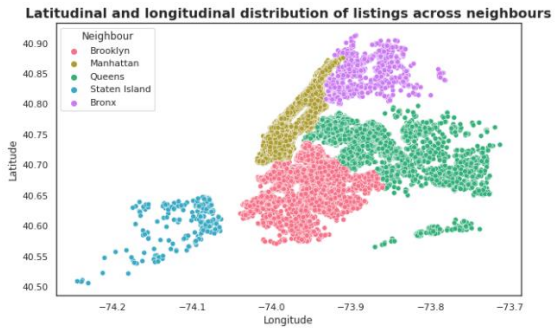


Fig: Geographical dist. of listings across neighbourhood

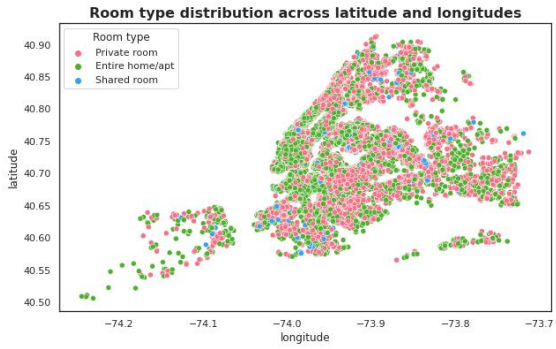


Fig: Geographical dist. of room type

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 dominancy in private rooms and entire homes categories*.

6.3 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.

Statistical characteristics		
Measures of position	mean	MEAN
	minimum	MIN
	maximum	MAX
	median	MEDIAN

	neighbourhood_group	median_price	mean_price	min_price	max_price
0	Bronx	65.0	79.558857	0	800
1	Brooklyn	94.0	121.463289	0	10000
2	Manhattan	140.0	180.071596	10	9999
3	Queens	72.0	95.783683	10	10000
4	Staten Island	75.0	89.964968	13	625

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**.

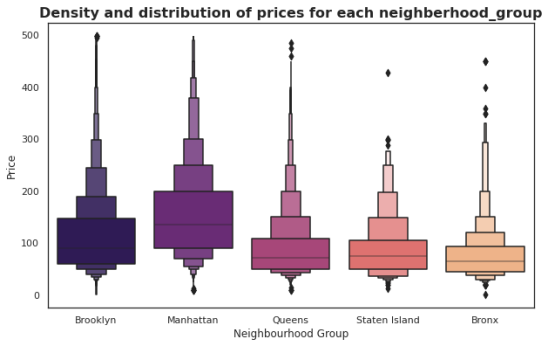


Fig: Density and distribution of price

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.

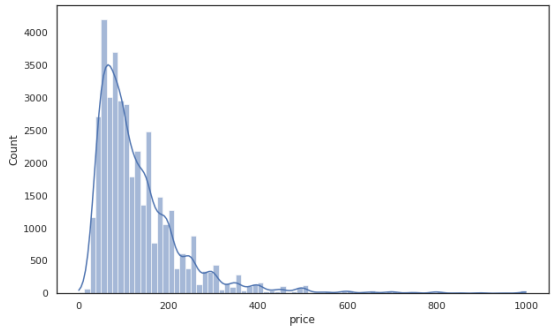


Fig: Frequency of listings with respect to the price using Histogram

By using the histogram, we can now see how prices are distributed. We have *a large number of values concentrated below 200 dollars*.

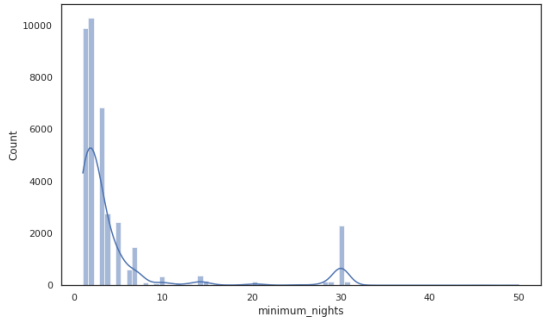


Fig: Frequency of listings with respect to the minimum number of nights using Histogram

In this case, most of the listings have listed out their minimum night record below 10. One unusual thing to note *here is the peak in the listing frequency for the **minimum night of 30***. It is possible that some owners have listed their properties on a monthly rental basis, which may explain this.

6.4 Univariate Analysis (Textual Data Mining):

For textual data mining on the name column, we will be using the **Wordcloud** library. Word clouds use frequency counts of the words as input and return a beautiful graphic display of the *most frequently occurring words* with their size proportional to their relative frequency. We can observe that our hosts utilise a variety of name conventions for their listings. We can find some intriguing tendencies by using the word cloud as an analytical tool to find out more about our hosts' behaviour and mentality.

Fig: Most frequent used words in listing's name

Hosts are using **simple** and **geographically focused** keywords. Given that the terms "**Manhattan**" and "**Brooklyn**" are prominently printed, the location is the main clue in this example. Adjectives like "**lovely**," "**peaceful**," "**cosy**," and "**gorgeous**" are used to describe the bedrooms and apartments, indicating that the comfort of the guests comes first. Private rooms are frequently mentioned in the city, which is a sign of their popularity.

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: 52.3%; private room: 45.5%; shared room: 2.2%.
- 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.

- Williamsburg and Bedford-Stuyvesant were found to be more popular neighbourhoods in the Brooklyn borough.
- 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.
- Bronx provides the cheapest accommodation among all.
- The majority of the listings have a minimum night record below 10 but there is a considerate frequency of listings for the minimum night of 30.
- It is evident that hosts are using simple and location-oriented keywords to differentiate their listings.
- Several mentions of “private rooms” indicate the popularity of this room type in the city.

8. References

1. Komorowski, Matthieu & Marshall, Dominic & Salciccioli, Justin & Crutain, Yves. (2016). Exploratory Data Analysis. 10.1007/978-3-319- 43742-2_15.

2. https://towardsdatascience.com/using -eda-to-generate-businessunderstanding-7f07f81e5af6

3.https://www.researchgate.net/publication/350400604_Exploratory_data_analysis_as_a_tool_for_risk_management_of_accommodation_services_Airbnb_New_York_City