

Capstone project -1 Exploratory Data Analysis on Airbnb Bookings



BY - ANKIT RAI

**CONTENTS**

* **Introduction**
* **Problem Statement**
* **Dataset Analysis**
* **Plot Analysis**
* **Limitation**
* **Scope of Improvement**
* **Conclusion**

**INTRODUCTION:-**

[Airbnb](https://www.airbnb.com/), 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 that recently moved to San Francisco. Unable to afford the rent for their loft at the time, the pair decided to make up the money they needed by renting out their apartment to people who couldn’t find hotels to stay at while attending nearby trade shows. They set up air mattresses in the apartment’s living room for their guests to sleep on and cooked homemade breakfast in the morning. Since then,Airbnb has become one of the trailblazers of peer-to-peer property rental.

**Problem Statement:**

Airbnb Problem Statement Airbnb is very successful, but many problems surround them. Airbnb has been very well known as an alternate solution to hotels. Airbnb is sometimes cheaper and can be a better solution. Airbnb has to decide whether to define their business model in a court of law and in the court of public opinion, or they will have to change it. Many other companies started to come up with their own concepts similar to Airbnb, so it is crucial to show how they differentiate themselves from their competition. Not only does Airbnb have competition with other concepts like them, but they are also in a competitive pool against rentals, one stop vacations, and others

We need to explore and analyze the data to discover key understandings (not limited to these) such as:

* What can we learn about different hosts and areas?
* What can we learn from predictions? (ex: locations, prices, reviews)
* Which hosts are the busiest and why?
* Is there any noticeable difference of traffic among different areas and what could be the reason for it?

Dataset Analysis

The dataset contains 48895 observations with 16 features. This data file includes all needed information to find out more about hosts, geographical availability, necessary metrics to make predictions and draw conclusions. Let us look through our features,

* Id: a unique id identifying an Airbnb listing or property
* name:  name representing the accommodation
* host\_id: a unique id identifying an Airbnb host
* neighbourhood\_group: a group of area
* neighborhood: area falls under neighbourhood\_group
* latitude: coordinate of listing
* longitude: coordinate of listing
* room\_type: type to categorize listing rooms
* price: price of listing
* minimum\_nights: the minimum nights required to stay in a single visit
* number\_of\_reviews: total count of reviews given by visitors
* last\_review: date of last review given
* reviews\_per\_month: rate of reviews given per month
* calculated\_host\_listings\_count: total no of listing registered under the host
* availability\_365: the number of days for which a host is available in a year.

latitude and longitude have represented a co-ordinate, neighbourhood\_group, neighborhood and room\_type are columns of categorical type. last\_review is a column of date type; we will convert it as required.

The distribution of numerical columns are as follows,

Other 3 important columns are,

* neighbourhood\_group: It contains 5 unique hoods which are Manhattan, Brooklyn, Queens, Bronx & Staten Island.
* neighbourhood: It contains 211 unique neighborhoods.
* room\_type: It contains 3 unique room types which are Entire home/apt, Private room, Shared room

The distribution of our numerical columns has positive skewness.

Out of all columns, 4 columns containing null values which are name, host\_name (looks like listing name and host\_name doesn't really matter to us for now) and last\_reviews, reviews\_per\_month (obviously, if a listing has never received a review, it’s possible and valid). So, those null values have been replaced with 0 during our analysis.

Data structure for given data sets

calculated\_host\_listings\_count

**Location**

Latitude

**Date**

Minimum nights

**Unique**

Id

Host Id

**Category**

Neighbourhood group

Neighbourhood

Room type

availability\_365

**Numeric**

**Object**

Reviews per month

Price

Host name

Name

**Airbnb booking data**

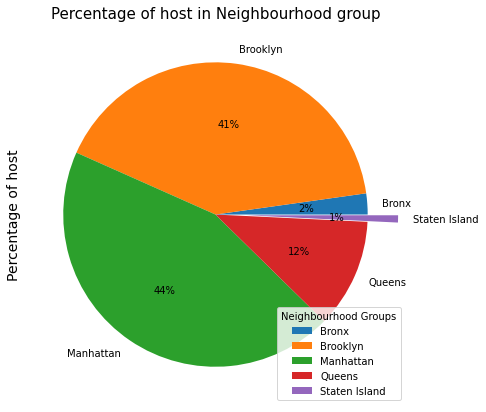
Number of reviews

Longitude

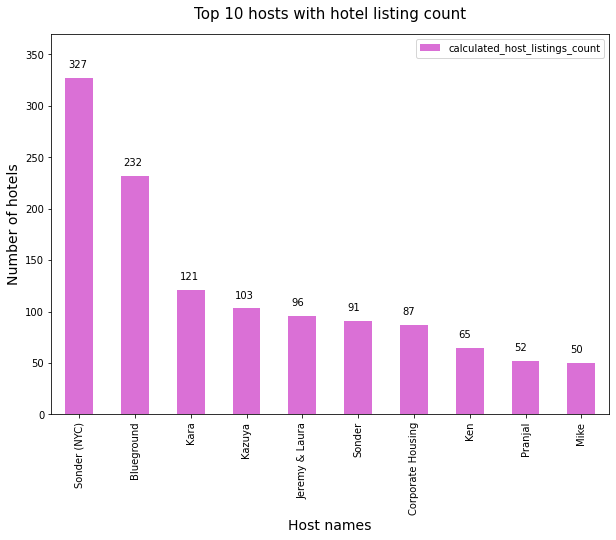
last\_review

Plot Analysis

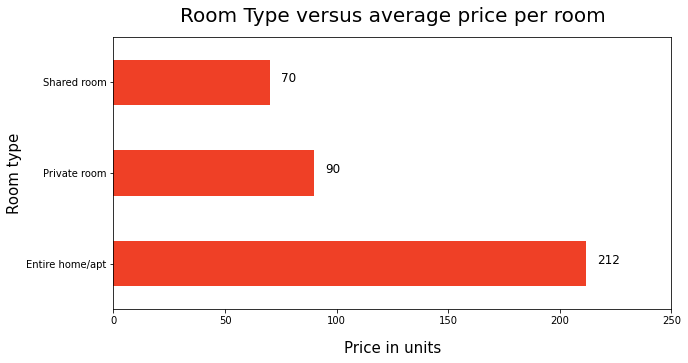
This is a heatmap-based correlation matrix. Except for reviews per month and number of reviews, which are obviously in the same category, almost all pairs of features have very little correlation. If all features are assumed to be independent, and price is used as the target variable, then MSE will be higher during regression because the target variable has a very low correlation with the features.



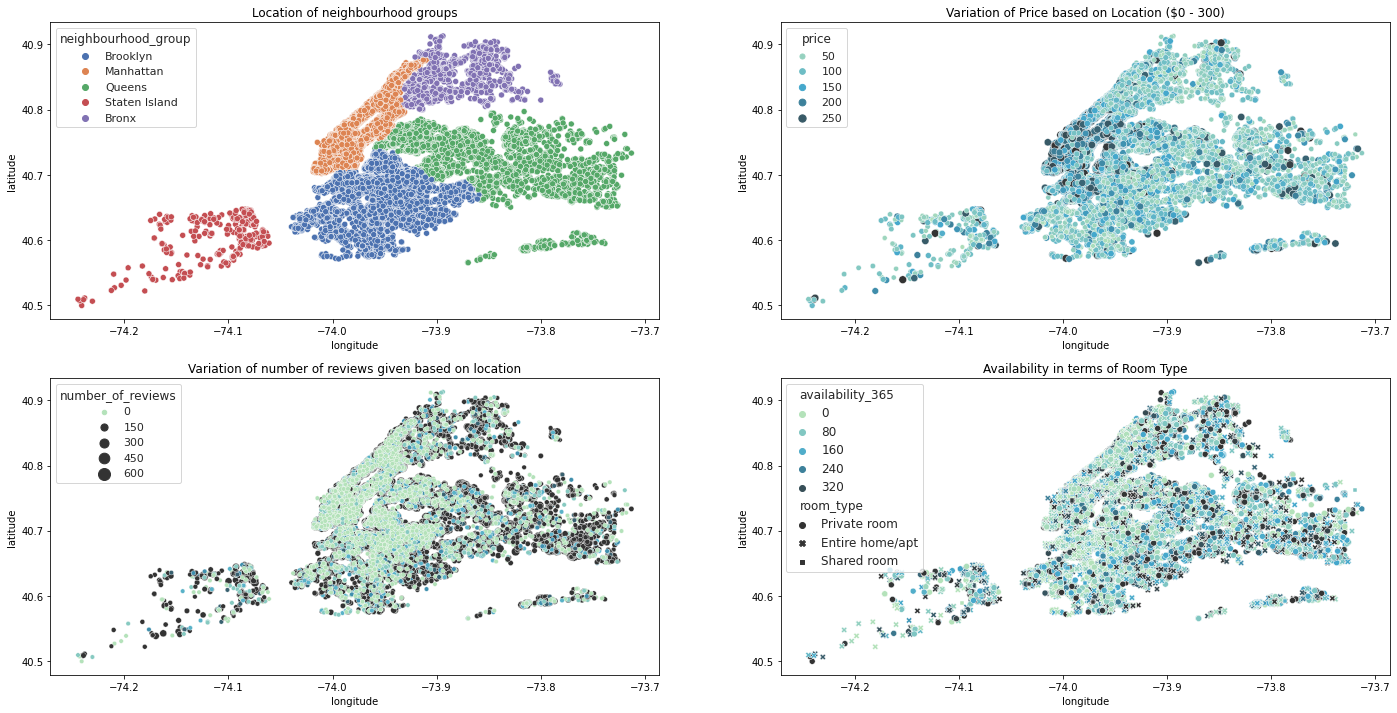
* In the first plot, we can see where our dataset is currently located in New York City.



* In the second plot we can check the number of hotels wrt the host name.



* We can check the number of hotels based on the host name in the third plot.
* Hotels provide three different types of rooms.
* When compared to the other two room types, the entire house/apt is more expensive.
* When compared to the other two room types, shared rooms are less expensive.
* THE ABOVE GRAPH CLEARLY SHOW THAT THE ENTIRE HOME/APARTMENT IS MOSTLY PREFERRED



**Let's talk about the plots mentioned above.**

**The first plot shows the neighborhood of New York City to which our dataset currently belongs. Because our 75th percentile data is in the range of USD 175, we only considered listings with a price range of up to USD 300 in the second plot. We can examine how price variations are distributed throughout the city. The areas south of Manhattan and north of Brooklyn are among the most expensive in New York. (From timeout.com.) The third plot shows a rising trend in the number of reviews on the outskirts of the city. We attempted to visualize availability in terms of room type in the previous plot. Availability varies by room type. Though availability by room type is well distributed, we can still observe a pattern in which the heart of New York remains the busiest or most booked for the majority of the time.**

**Limitation:**

Despite the fact that the dataset is very feature rich, has low correlation, and contains enough data to perform regression on price prediction, the correlation with the target price is also low. As a result, the MSE will be high. Furthermore, the features dataset provides in terms of the modern world are of very poor quality in determining property valuation. Furthermore, because the features are positively skewed, we must treat them prior to prediction.

* It would be interesting to conduct a sentiment analysis with property valuations in order to have a better analysis of the quality of the properties.
* Because host user ratings are not available, it would have been preferable to rank our hosts based on user satisfaction and ratings. Furthermore, in those instances, Further research can be conducted to determine how guests rate in terms of price or room type, or whether the rating determines the property's valuation. A low-rated property usually has a lower price.
* The exact number of guests is also missing; it is assumed that the guests are counted by col: last review. Even if a new host has never been rated, that does not mean no guests have ever stayed there.

**Scope of Improvement:**

Because the dataset only contains a few qualifying attributes for valuing a property, more features such as bedroom, bathroom, property age (which may be one of the most important), tax rate applicable, and distance to nearest airport, hospital, or schools can be added.

With ratings, hosts can be classified and ranked, and special discounts or offers can be given to the highest rated hosts in accordance with marketing strategy.

Time series analysis can be used to forecast occupancy rates based on the tourist season.

**Conclusion:**

* Based on the findings, it is possible to conclude that: • Most visitors do not prefer shared rooms; instead, they prefer to visit a private room or an entire home.
* Manhattan and Brooklyn are New York's two most distinguished, expensive, and affluent neighborhoods.
* Some properties have minimum nightly stays of more than 365 days, which may be appealing to students, low-income employees, and immigrants.
* While the location of a property has a strong influence on its price, a property in a popular location does not guarantee that it will be occupied the majority of the time.
* Because the features in this dataset are of very poor quality in determining property valuation, performing a regression on it may result in a high error rate. The correlation heatmap demonstrates this. We. We can see this by looking at correlation heatmap. • We would require additional features such as bedrooms, bathrooms, property age (which I believe is very important), tax rate applicable on land, room extra amenities, and distance to nearest hospital, stores, or schools. These characteristics may have a strong relationship with price.
* Using time series analysis, it is possible to make predictions about occupancy rates at specific times of the month or season.
* It would be preferable if we had average guest ratings for a property; this would help us understand the property better and could also be used to determine price. A low-rated property usually has a lower price.