Airbnb Booking Analysis

(Abhishek V L, Neha R, Nisarga C, Swati R G)
Data science trainees,
Alma Better, Bangalore

ABSTRACT

Airbnb is an online marketplace that connects people who want to rent out their homes with people who are looking for accommodations in specific locales and hospitality service for people to lease or rent short-term lodging including holiday cottages, apartments, home stays, hostel beds, or hotel rooms. Basically they provide the platform that is shared by hosts and visitors worldwide

Airbnb is one kind of service that connects the guest and host to share their property. Basically, property owner wants to utilize their property in the right direction so that for host it can be an income source and for guest, it can be a destination for a stay. The problem for Airbnb is there are tons of data generated through the hosts and guests as well. So, to find the right direction according to marketing and find the correct business-driven solution we did the data analysis on 49,000 observations in it with 16 columns and it is a mix of categorical and numerical values.

1. PROBLEM STATEMENT

The datasets used for analysis provide detailed information about a listing's hosts, rooms, customer reviews, prices, Neighborhood and ratings among other important attributes.

By exploring prior work correlating and visualizing various attributes of Airbnb's listings. This helped to understand the different approaches people have taken to analyze, visualize, and co-relate Airbnb data.

Airbnb Super hosts i.e. hosts with consistently high ratings and sales. This helped to shortlist the attributes.

Due to the scarcity of real business data for scientific and educational purposes, these datasets can have an important role for research and education in revenue management, machine learning, or data mining, as well as in other fields.

After the EDA, we saw how each analysis is contributed positively for the business objective to be achieved and the suggestions were made in Airbnb booking and how factors interacted with each other's. They are:

1. How neighborhood is related with reviews?
2. Which are the top 5 hosts that have obtained highest no. of reviews?
3. Which hosts are having highest number of apartments?
4. Which are the top 10 neighbourhood? Which are having

maximum number of apartments for Airbnb?

5. What is the neighbourhood in each group which is having maximum prices in their respective neighbourhood_group?
6. What can we learn from predictions? (Ex: locations, prices, reviews, etc.)
7. What are the distribution of the room type and its distribution over the location?
8. How does the room type is distributed over Neighbourhood_Group are the ratios of respective room types more or less same over each neighbourhood_group?
9. How the price column is distributed over room_type and are there any Surprising items in price column?
10. What is the average preferred price by customers according to the neighbourhood_group for each category of Room_type?
11. What is the average price preferred for Keeping good number_of_reviews according to neighbourhood_group ?
12. Which host are busiest and why?

1.1 Data set

So, we can see our dataset has 48895 data and 16 columns. Let's try to understand about the columns we've got here.

- id : a unique id identifying an Airbnb listing
- name : name representating the accommodation
- host_id : a unique id identifying an Airbnb host
- host_name : name under whom host is registered
- 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.

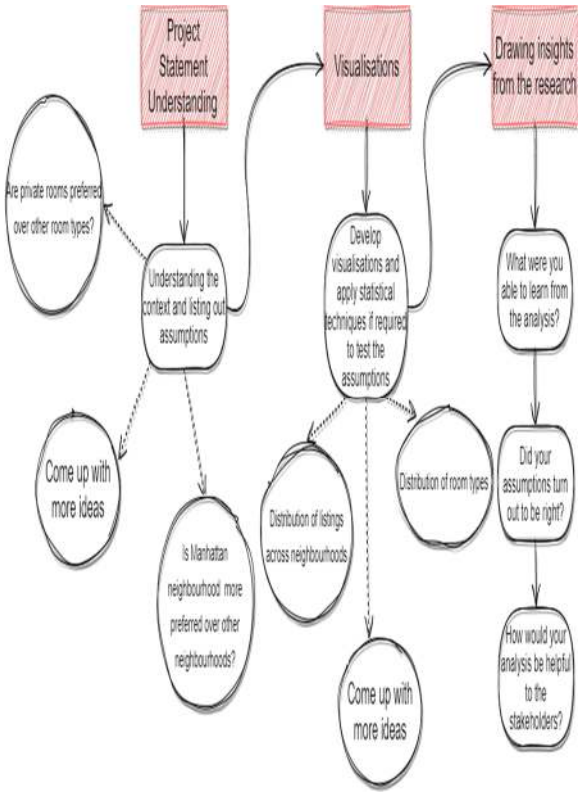
Latitude and longitude has represented a co-ordinate, neighbourhood_group, neighbourhood and room_type are columns of categorical type.

Last_review is a column of date type, we will convert it as required.

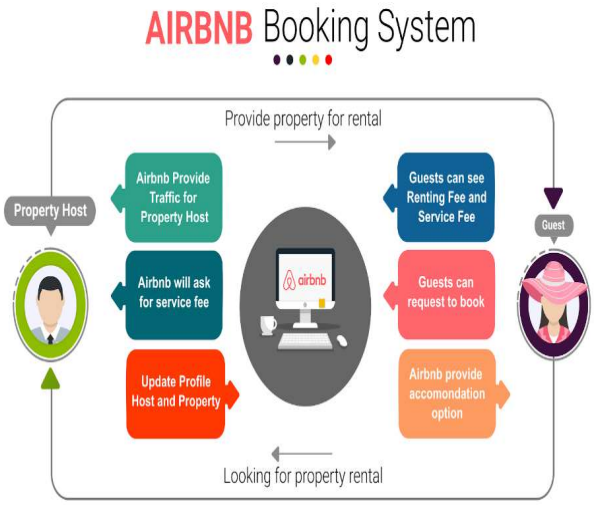
1.2 Main Libraries Used:

- Pandas for data manipulation, aggregation.
- Matplotlib and seaborn for visualization and behavior with respect to the target variable.
- NumPy for Computationally efficient operations.

1.3 Project Architecture:



2. INTRODUCTION



Airbnb provides various rental options for different customer segments. Based on customer budget, they can either opt for an entire house or just a room or even better share a room. With a range of prices as low as 700 to as high as 50,000, comes a range of amenities, such as selection on a number of beds, bedrooms, kitchen, air conditioning, heating washing machine, breakfast, beachfront, gym, pool etc to name a few.

There are four major types of places:

1. Entire place
2. Private room
3. Hotel room
4. Shared room

For the hosts, Airbnb has a super host program providing exclusive benefits and higher visibility.

3. STEPS INVOLVED

3.1 Null Values Treatment

- We created a copy of the given dataset, so that our original dataset remains unchanged.
- We can check there are 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 we will just isna(0) to those null values.
- We can conclude from that the Airbnb dataset contains only a few NA values which further help us argue that the analysis done on this data is performed without much loss of information.

3.2 EDA Performing

3.2.1 EDA based on highest number of reviews:

Analyzed the top 5 neighborhoods having the highest number of reviews per month, and in that we come to know that the neighborhood of Theater district-58.90, Rosedale-20.94, Springfield gardens-19.75, East elmhrust-16.22, Jamaica-15.32 respectively.

Analyzed the top 5 neighborhoods having the highest number of reviews, and in this we come through that Bedford-Stuyvesant has the highest number of reviews with 110352 and so on.

3.2.2 EDA based on top 5 hosts that have obtained highest no. of reviews :

Analyzed the top 5 hosts that have obtained highest no. of reviews and we got to know that Maya, Brooklyn & Breakfast–len-, Danielle, Yasu & akikkio, Brady are the ones who got highest reviews.

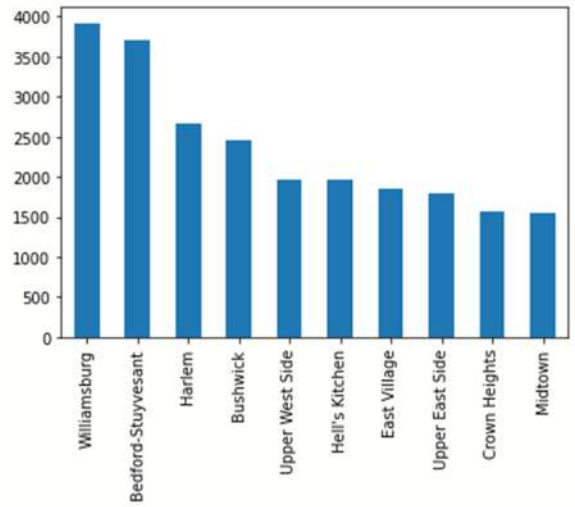
3.2.3 EDA based on hosts that are having highest no. of apartments:

From the analysis we can see that host name Michael its appearing 417 times in the host_name column, so this might imply that Michael is having highest number of rooms , but from the host_id column its showing highest appearance of any host_id is 327 , so this clearly implies that there can be multiple person may have same name that's why we are getting different highest appearance in host_name as compared to host_id.

And also after performing the analysis we come to know that Sonder (NYC) is having maximum numbers of rooms for the guest, For Airbnb he might be very important person then.

3.2.4 EDA based on Which are the top 10 neighborhood & which are having maximum number of apartments for airbnb:

From the below bar graph, we can see the top
10 neighbourhood which are having maximum number of apartments for Airbnb in the respective neighbourhood:



3.2.5 EDA based on What is the neighborhood in each group which is having maximum prices in their respective neighbourhood group:

- Top 3 neighborhood in Manhattan which are having maximum prices

	Neighborhood	price
0	Upper West Side	10000
1	East Harlem	9999
2	Lower East Side	9999

- Top 3 neighborhood in Staten Island which are having maximum prices

	Neighborhood	price
0	Randall Manor	5000
1	Prince's Bay	1250
2	St. George	1000

- Top 3 neighborhood in Bronx which are having maximum prices

	Neighborhood	price
0	Riverdale	2500
1	City Island	1000
2	Longwood	680

- Top 3 neighborhood in Queens which are having maximum prices

	Neighborhood	price
0	Astoria	10000
1	Bayside	2600
2	Foresthills	2350

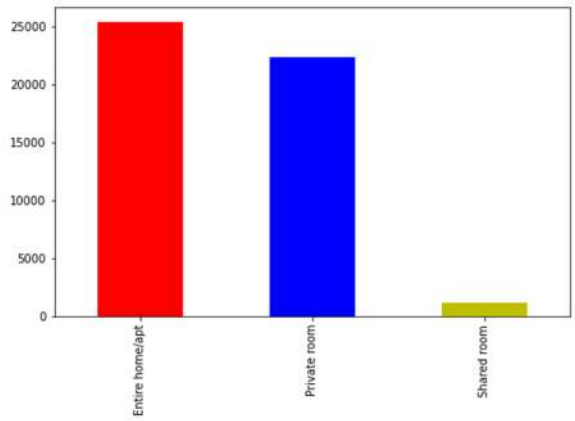
- Top 3 neighborhood in Brooklyn which are having maximum prices

	Neighborhood	price
0	Greenpoint	10000
1	Clinton Hill	8000
2	East Flatbush	7500

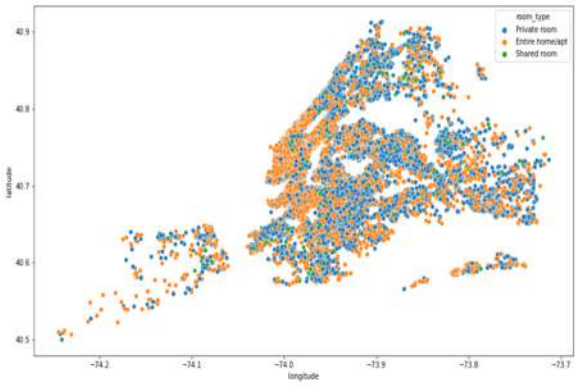
3.2.6 EDA based on What can we learn from predictions? (ex: locations, prices, reviews, etc):

We can definitely observe a couple of things about distribution of prices for Airbnb in NYC boroughs. First, we can state that Manhattan has the highest range of prices for the listings with $150 price as average observation, followed by Brooklyn with $90 per night. Queens and Staten Island appear to have very similar distributions; Bronx is the cheapest of them all. This distribution and density of prices were completely expected; for example, as it is no secret that Manhattan is one of the most expensive places in the world to live in, where Bronx on other hand appears to have lower standards of living.

3.2.7 EDA based on What is the distribution of the room type and its distribution over the location:



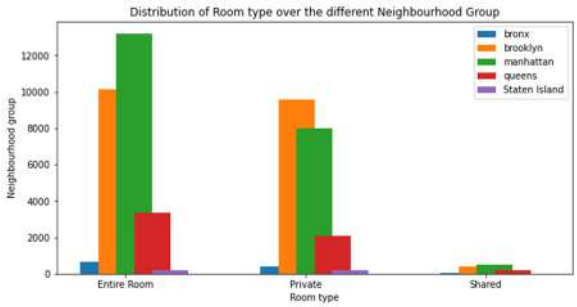
From the analysis we can understand Brooklyn and Manhattan stands within the most urban and active area, in terms of listing areas and pricing. Manhattan & Brooklyn has highest average room price, though Staten Island is not far behind. Shared room has relatively low price and also low in count in the entire neighborhood, whilst Manhattan has most number of Entire home/apt category, but Brooklyn has most number of Private room category.



Clearly, room type Entire home/apt has maintained higher price range in almost all neighborhoods.

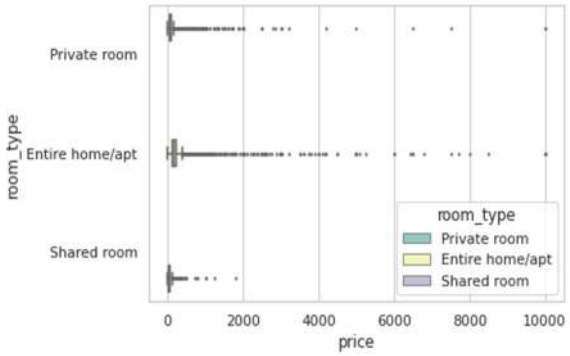
Entire home/apt has more than 50% proportion in New York City and it too has highest average price also. Shared room is the cheapest, but only has 2.4% proportion. No wonder New York life is of high standard.

3.2.8 EDA based on How does the Room type is distributed over Neighbourhood Group are the ratios of respective room types more or less same over each neighbourhood group:



From this analysis we can see that more or less same ratio in every neighbourhood_group with respect to the room_type.

3.2.9 EDA based on How the price column is distributed over room type and are there any Surprising items in price column:



From this analysis we can notice that there are many outliers for price in each of the room_type category, so let's just why there is so high price or what else we can conclude for hosts having highest price for the rooms**.**

3.2.10 EDA based on What is the average preferred price by customers according to the neighbourhood group for each category of Room type:

From the analysis we can see that Manhattan is most costly and Bronx is cheap for each room_type,

But I think we can make it more useful for business implementation if we do some analysis on successful hosts according to the highest no of reviews so that we can suggest this price to our host for good business.

3.2.11 EDA based on What is the average price preferred for Keeping good number of reviews according to neighbourhood group:

1) Clearly if we compare the results with previous result (i.e. when we calculated average preferred price by people in each neighbourhood_group with different room_types) we can see that this result is bit different and more useful

2) As a analyst I would suggest to keep price in this range to get more number of reviews in specific room type and at particular place

Brooklyn and Manhattan stands pretty fall in terms of review rate per month. Also, we can notice a negative relation between price & no. of reviews. Where, costlier properties have significantly less no of reviews, but cheaper properties had large number of reviews. Usually, cheaper rooms have more number of guest visits than costlier one, we know no of reviews is directly proportional to no. of guests.

3.2.12 EDA based on Which host are busiest and why:

According to the assumptions and calculations done above to calculate the metric, a property with 1 customer over the entire period of business as the property's total possible booking records a 100% when the estimated bookings is also 1. In simpler terms, if the expected booking count is calculated to be 1 and the property hosts 1 customer, then the property is said to be 100% busy.

4. CONCLUSION:

After analyzing almost 49,000 properties on Airbnb, it seems the platform has given an additional source of income of the owners and choice of the customer (Private room, Shared rooms, Dormitory rooms, tent house, and their comfortability) over the traditional hotels.

- Manhattan and Brooklyn are the two distinguished, expensive & posh areas of NY
- Customers pay highest amount in Brooklyn, Queens and Manhattan that is 10,000 and lowest amount is 10$.

- For the three types of room type (i.e. Entire home, Shared room, & Private room) average price of entire home is around $157, for shared room is around 60, and for private room is around 75.

- 'Entire home/apt' room type has the highest number of listing of 52% and ‘Shared Room’ is the least listed room type at only 2.4% in total.

- People stay for longer duration of time in Private rooms in Brooklyn and Manhattan.

- The main takeaways of the Seattle data analysis include:
  Basic characteristics of the place (number of bedrooms, bathrooms, beds and accommodates) affect the reservation price.
  The reservation price varies depending on the time of the year. For example, the busiest time to visit Seattle is summer.

Overall, we discovered a very good number of interesting relationships between features and explained each step of the process. This data analytics is very much mimicked on a higher level on Airbnb Data for better business decisions, control over the platform, marketing initiatives, implementation of new features and much more. Therefore, I hope this kernel helps everyone!