EDA PROJECT ON AIRBNB NYC

Chandan Baraliya, Bhupendra Singh

Abstract:

This is the final technical report of our data analytic project titled "EDA on Airbnb NYC" as a part of our Data analytic course at Alma better. The goal is to analyze and predict the price and other variables in the New York Airbnb data. Also, a recommendation system will be built to recommend Airbnb listings according to the user preference.

Content:

Airbnb is an online marketplace for arranging or offering lodging, primarily home stays, or tourism experiences. The company does not own any of the real estate listings, nor does it host events; it acts as a broker, receiving commissions from each booking. The company is based in San Francisco, California, United States.

The company was conceived after its founders put an air mattress in their living room, effectively turning their apartment into a bed and breakfast, to offset the prohibitive cost of rent in San Francisco; Airbnb is a shortened version of its original name, AirBedandBreakfast.com

Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present a more unique, personalized way of experiencing the world. Today, Airbnb became one-of-a-kind service that is used and recognized by the entire world. Data analysis on millions of listings provided through Airbnb is a crucial factor for the company. These millions of listings generate a lot of data - data that can be analyzed and used for security, business decisions, understanding of customers' and providers' (hosts) behavior and performance on the platform, guiding marketing initiatives, implementation of innovative additional services and much more.

The dataset had around 49,000 observations in it with 16 columns and it is a mix between categorical and numeric values.

Problem Statement

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, etc.)
- 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?

Approach

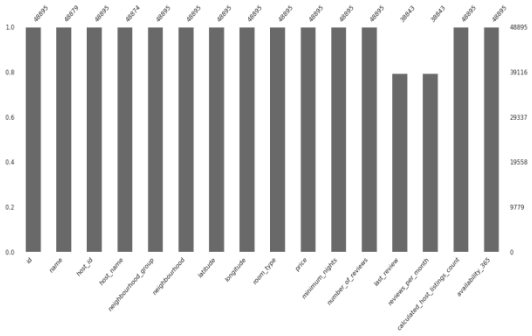
1-Let us first check our dataset's and understand it.
2-Later we will check for any missing data in the data given. Does it hamper our analysis?
3-We would check the type of data and divide it for our analysis.
4-We checked where there any outlier or unethical data in it if so, we would filer such data for specific analysis.
5-Then do Data analysis by visualization techniques.
6-And then conclude with various outcomes from it.

A. dataset's

Number of Columns:16
Number of Samples:48895

Number of quantitative variables:10
Number of qualitative variables:6
Attributes:
id, name,
host_id, host_name,
neighborhood-group,
neighborhood,
latitude, longitude,
room type, price,
minimum nights,
number_of_reviews,
last_review, reviews_per_month,
calculated_host_listings_count,
availability_365

Checking Data for any missing values.



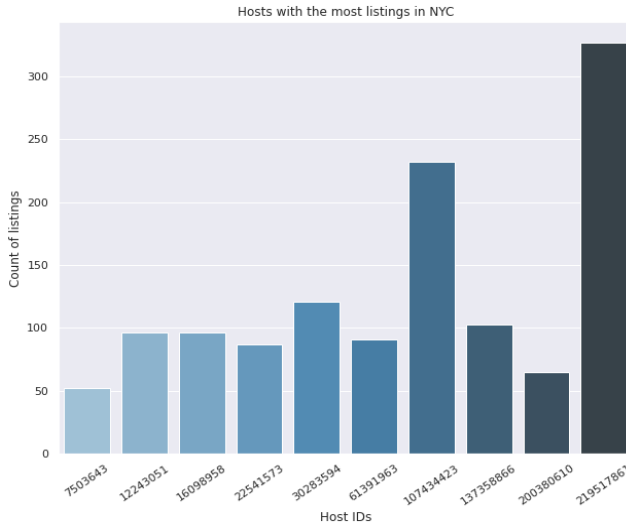
Since our dataset's contain several missing values preprocessing must be done. Missing values will either be deleted or replaced with the column mean or nan. based on how important the attribute is. Also, with respect to preprocessing the datatype of certain attributes like last_review must be changed to make processing easier. Our main goal is to analyze and find interesting patterns between the variables in our dataset's. Visualization is an important aspect of finding patterns. Hence several visualization techniques like bar graph, pie chart, Violin chart, correlation, etc. will be plotted to gain insights. We then plan to predict certain variables such as price by using predictive models. Several models will be explored and models with the best accuracy will be selected. We also got that
-Few columns like name, host name, last review had many missing values and then we replaced it with "missing". Importance for analysis, hence they were deleted.
-Reviews per month column had lot of missing rows but is important for analysis, hence missing values were replaced with the mean of that column.

Exploratory Data Analysis

Now we know that we are ready for an exploration of our data, we can make a rule that we are going to be working from left to right. The reason some may prefer to do this is due to its set approach - some datasets have a substantial number of attributes; plus, this way we will remember to explore each column individually to make sure we learn as much as we can about our dataset's.

Observation 1

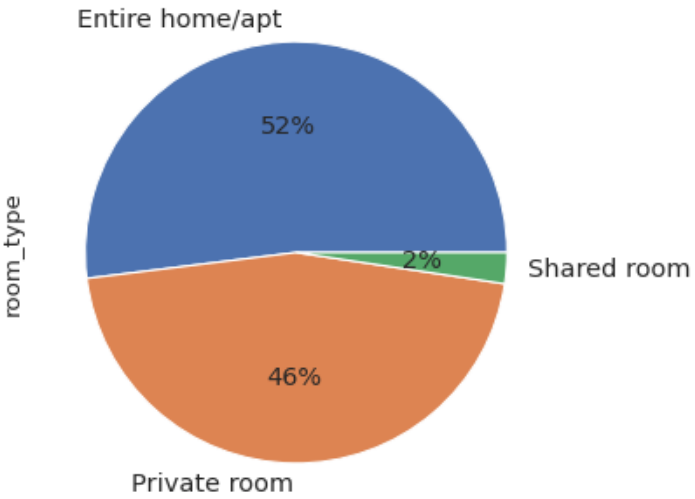
We first take the Host id and check the



maximum count list. So, we could get our top busiest hosts. We then plot it in a bar chart and find that host id 219517861 is the busiest host with more than 350 count list and is followed by 107434423.

Observation 2

We plot a pie chart and find the various distribution of home property in NYC.



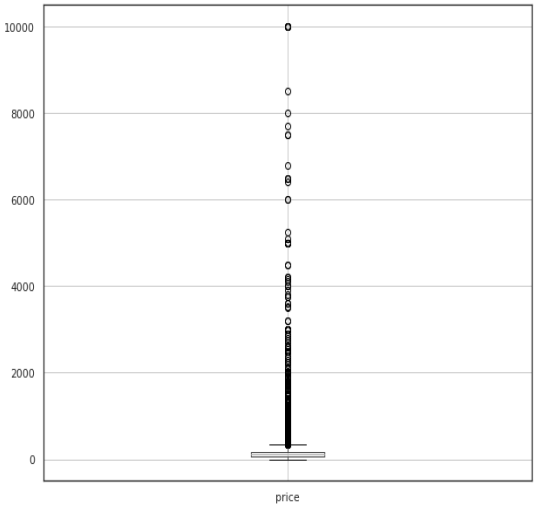
Here we find that Entire home/apt has about 52% listing in NYC and shared rooms has least of 2 %.

Let's check our Price Column which is the most important Key performer for every bus iness to run. Let's check whats the minimum and maximum price of any room types

As we found that minimum price is zero whi ch is impossible let's filter out these rows an d Airbnb is to make business.

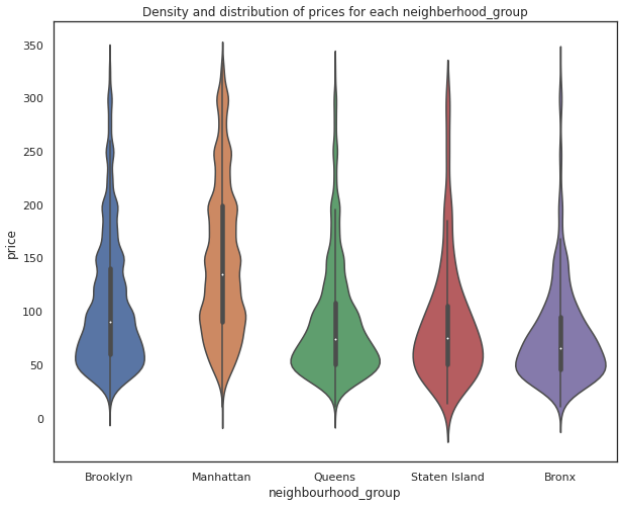
There could not be zero price. So removing price value as zero as it is not possible. So drooping rows with price as Zero.

Now, as we filter are price from zero price v alue. Let's check for any outliers in price col umn as its the most important column for ou r analysis for which we need to be cautious f or our future analysis.



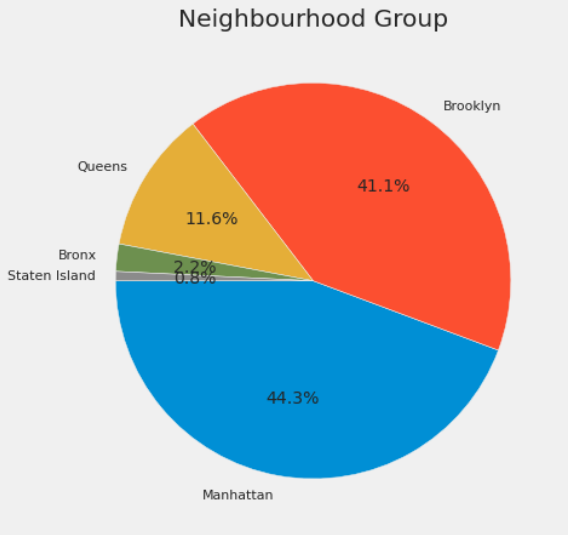
So, from above graph we could observe som e outliers present in our datasets in prices. S o, to remove the outlier lets define a functio n for it. After defining the function, we remove the outlier and then do Data Analysis on it.

Observation 3



Great with the help of violin plot with ignori ng outliers its clear that the Booking cost of room in
Manhattan is 140 dollar per night as average distribution, followed by Brooklyn with the average cost of 90 dollar per night. on the ot her hand, Queens and Staten
Island have almost same living price as appr ox 75 dollar per night, furthermore Bronx is cheapest among them all.
Even our observation somehow matches the real-
world analysis because Manhattan belongs t o expensive places to live in Newyork.

Observation 4



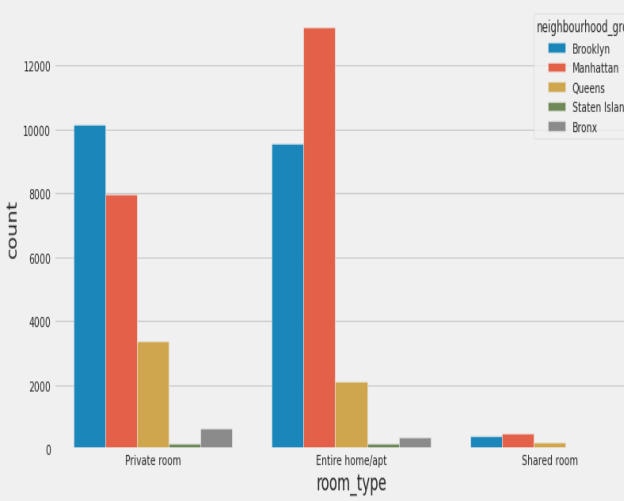
We count the list of victors count with the neighborhood and find that that Manhattan is the most preferred place for stay by people even though costly followed by Brooklyn. So, we can conclude that Manhattan is a tourist place**.**

Observation 5.

We gave a street map with various distribution of property and moving the mouse would display the price at that property. These is lowed by tourist and is a way the google maps are used much more by people to find and could give us a brief infer of spread of various properties at neighborhood and its exact location with prices. Indeed, it is the best way to get aware of various location prices and makes us an idea of higher number of properties at various

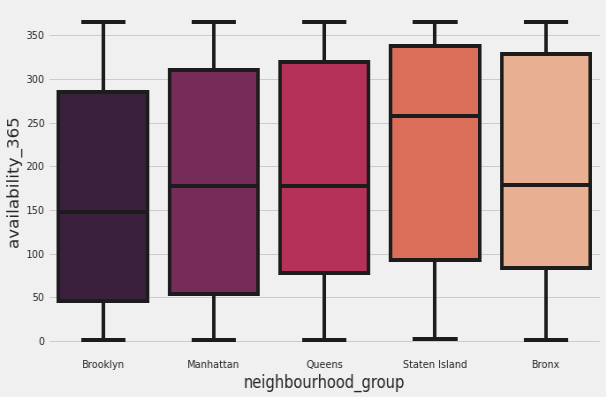
location. The Open Street map is best suited for it, and we plotted it to show it.

Observation 6



We then moved to see the various property counts at different neighborhood and find that Private property area more in Brooklyn and Entire apartment are more in Manhattan

Observation 7



To check the property availability at various location we used the availability 365 and neighborhood column and plot it to find the average availability into a box plot and we just find that Brooklyn, the second highest li sting of people in booking in NYC has a smaller number of availabilities throughout t he year and
State Island has the highest. Which states that Staten
Island is less liked by people to stay

Problem faced-

1- Huge Cluster of data with different price rate had and issue how to remove the outliers.
2- Number of reviews had a great challenge to handle at different area.
3- To plot a map in a street map was a task.
4- The regions have a great insight to say and needed to handle the latitude and longitude on a single street map it easier.
5- Few team left in between due to some personal reason which put a load to submit the task.

Conclusion

1 Host_id with 219517861 has the greatest number of listings of 300+.

2-Entire Home apartment property are more in Airbnb NYC.

3-Manhatten has highest cost of living followed by Brooklyn.

4-Manhatten is the most loved place by people in NYC.

5- We have given a map of spread of prices at various location which makes it easy for visualize the spread of room types and its rate there.

6- Manhattan has the highest Entire Apartment property list in Airbnb NYC. and Brooklyn held the first for Private room type property.

7- Staten Island has the most vacate room in Airbnb NYC.

References

1-https://www.geeksforgeeks.org/
2-https://www.almabetter.com/ (notes)
3-https://stackoverflow.com/questions/214741/what-is-a-stackoverflowerror

Remarks: - These write up is a documentation of individual work done for the project by Chandan Baraliya and is a self-documented.