AirBnb Analysis NYC

Section I: Introduction

Airbnb, born with a creative spirit, is an online marketplace that connects people who want to rent out their homes with people who are looking for accommodations in that locale, ultimately providing a more unique and personalized way of experiencing the world. The source of motivation of our problem statement as a social impact is to help people in making informed and meaningful decisions when choosing for their accommodation whereas, for the organization it can be used by decision-makers to take appropriate action to enhance productivity and business gain. The basic idea is to understand the key metrics to explore data in meaningful ways as data is merely facts and figures.

Section II: Data Informatics

The dataset consists of almost **50 thousand** listing activity and **16 metrics** in NYC, NY for the year **2019**. The data file includes all the information to find out more about *hosts, geographical availability, review ratings* etc. necessary to make predictions and draw suitable conclusions. Here for our analysis, we will be organizing, interpreting, re-structuring and present the data into useful information that provides comprehensive results with visualization representations in an easily digestible format.

Section III: Data Pre-Processing and Wrangling

The raw data had various discrepancies and inconsistent across multiple features. Hence, with data cleaning and wrangling, it provided us much feasibility to gain insights for our analysis. Below are some of the operations and analysis that we used for this:

(A) Data Pre-Processing

After looking at the head of the dataset we already were able to identify some NaN values. Then, using sum function, we found the count of null for each column. The below pic depicts null value count for each of the column:

```
id                                  0
name                               16
host_id                             0
host_name                          21
neighbourhood_group                 0
neighbourhood                       0
latitude                            0
longitude                           0
room_type                           0
price                               0
minimum_nights                      0
number_of_reviews                   0
last_review                     10052
reviews_per_month               10052
calculated_host_listings_count      0
availability_365                    0
```

Fig 1: Total Count of null values across features

(B) Data Wrangling

In our case, missing data that is observed does not need too much special treatment. Looking into the nature of our dataset we found out: columns "*name*" and "*host name"* were irrelevant and insignificant for our data analysis, whereas for columns "*last_review"* and "*review_per_month*" need very simple handling. To elaborate, *"last review"* is date; if there were no reviews for the listing - date simply will not exist. Therefore, we removed all those columns that were insignificant and imputed other columns for missing values of data.

Section IV: DB Normalization

Here we identified some of the core table and inter-dependency between them, in order to create our primary table i.e., ***Airbnb.*** After normalizing the data, the below DB Schema is generated:

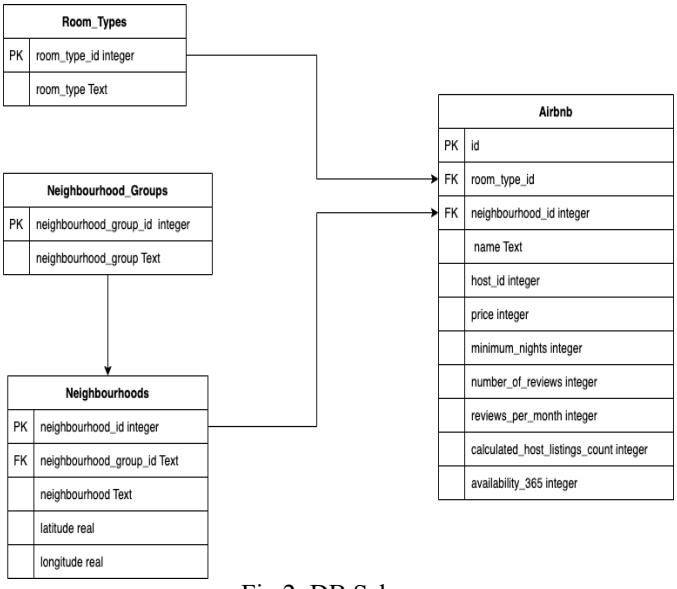

Fig 2: DB Schema

Section IV: Visual Representations and Inferences

Using various libraries for our graphical representation, the below details and plots represents the key insights for various of our features.

(A) Host Data

Here the below bar plot represents the data of the host id with most number of listings:

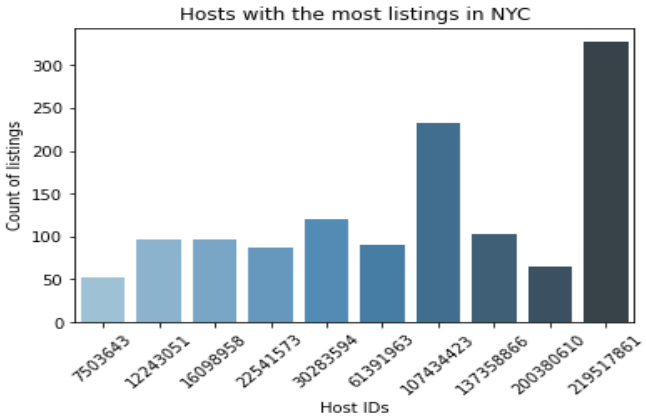

Fig 3: Top Hosts on Airbnb in NYC

(B) Reviews Data

Here, the below boxplot depicts the relation between *number_of_reviews* and the *neighbourhood group*

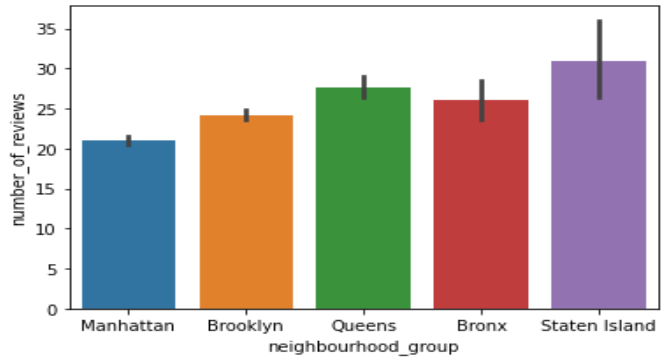

Fig 4: Comparison of number of reviews and neighbourhood group

(C) Neighbourhood Data

Here the below pie-chart show the no. of listings according to the neighborhood in NYC.

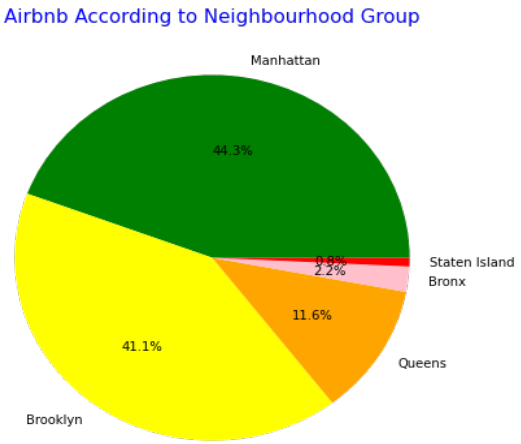

Fig 5: Airbnb distribution as per neighbourhood

(D) Geographic Data

Availability in number of days per year of Airbnb according to geographic location

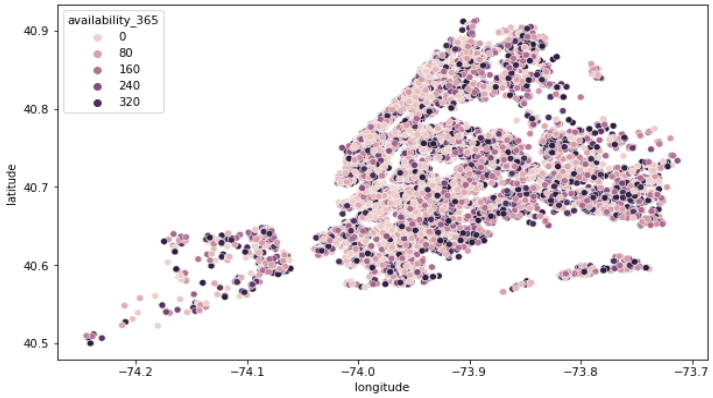

Fig 6: This graph shows availability in number of days depending on geographic location

(E) Room Type Data

Here, the pie-chart represents the listings distribution with their room type

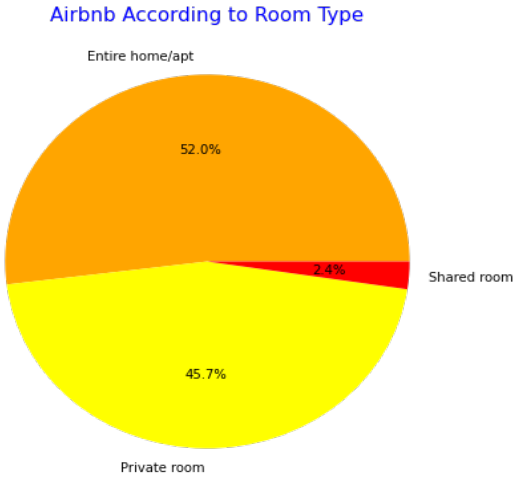

Fig 7: Distribution of listed Airbnb according to room type

Here the below histogram represents the listings count of Airbnb in neighbourhood with their room category

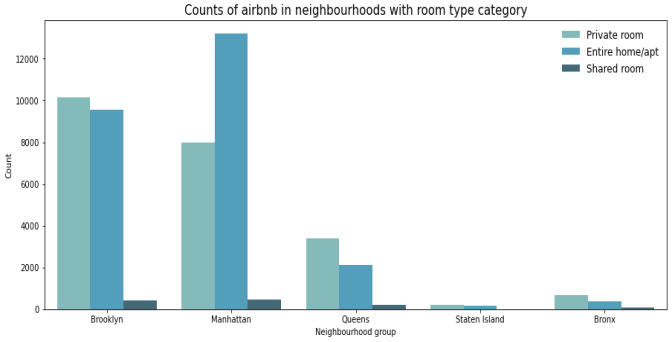

Fig 8: Count of Airbnb in neighbourhoods with room type category

The below plots depict the listings count of Room types in localities of most popular nieghbourhood in NYC i.e., Manhattan and Brooklyn

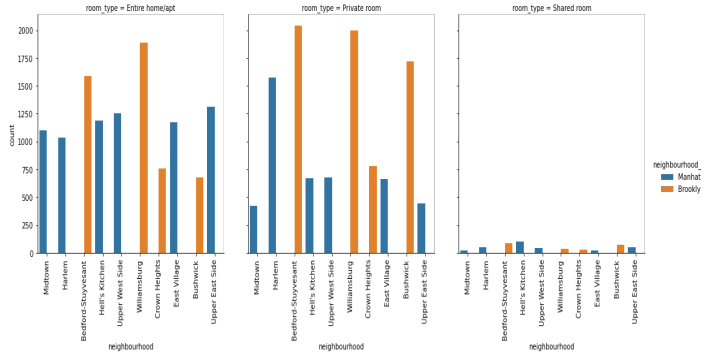

Fig 9: Count of various room types

Here based on the above plot, we can draw some inferences:

- We can see that for these 10 neighborhoods only 2 boroughs are represented: Manhattan and Brooklyn; that was somewhat expected as Manhattan and Brooklyn are one of the most traveled destinations, therefore would have the most listing availability.
- We can also observe that Bedford-Stuyvesant and Williamsburg are the most popular for Manhattan borough, and Harlem for Brooklyn.

(F) Price Data

Here, using Violin plot we determine the relation between locale density and distribution of prices for each neighbourhood group

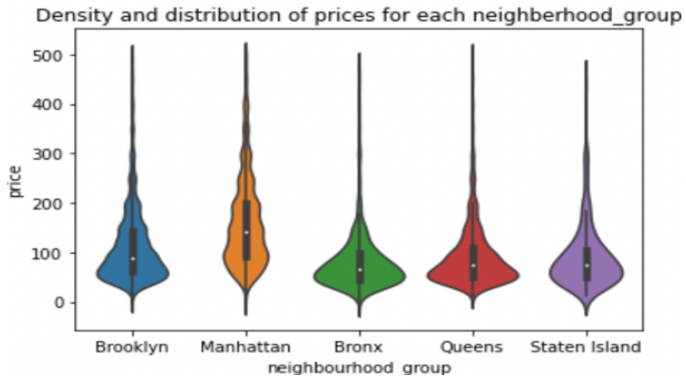

Fig 10: Density and distribution of prices for each neighbourhood group

Here based on the above plot we draw some inferences:

- 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.

Here, the below Heat map, in Figure 11 gives information about the prices in various regions of New York City. Here *blue* means the lowest price and *red* means most expensive area.

Fig 11: Heat Map of Airbnb prices in NYC

Section V: Summary:

The Airbnb *('AB_NYC_2019')* dataset for the 2019 year is a very rich dataset with a variety of features, that allowed us to perform deep data exploration on each significant column. Firstly, we found the hosts that take good advantage of the Airbnb platform and provided the most listings i.e., we found that our top host has 327 listings. Further, we graphically represented the reviews data w.r.t the neighborhood data and analyzed existing listings for their popularity and market demand. Then the data is classified based on their spatial geographical distributions and room category types. Lastly, one of the most important factor "*Pricing*" is analyzed w.r.t multiple features and heat map was generated for lucid understanding and taking meaningful decisions for choosing a listing. For our future data exploration, it would have been nice to have couple additional features, such as positive and negative numeric (0-5 stars) reviews or 0–5-star average review for each listing, which would additionally help us to determine the best-reviewed hosts for NYC. Overall, we discovered a very good number of interesting relationships between features and tried to explain each step of the process by visual representation for the same. This data analytics can very much be mimicked on a higher level on Airbnb Data/Machine Learning team for better business decisions, control over the platform, marketing initiatives, and much more. Therefore, we hope this project will along with business will also serve for the community as a whole.

Section VI: References

[1] Julia M. Núñez-Tabales, Miguel Ángel Solano-Sánchez and Lorena Caridad-y-López-del-Río "Ten Years of Airbnb Phenomenon Research:" paper published on 1 August 2020.

[2] Najmeh Hassanli, Jenni Small, Simon Darcy, "The representation of Airbnb in nespapers" published online on24 Sep 2019.

[3] Lusia Andreu, Enrique Bigne, Suzanne Amaro, Jesus Palomo"Airbnb Research" Published on 11 February 2020

Important Links:

- http://insideairbnb.com/
- https://www.kaggle.com/
- https://towardsdatascience.com/