AIRBNB BOOKING ANALYSIS

Ishan Barway, Jasmine Sulekha Nag, Nitin M Sakhare, Roshan Alte

Data-Science Trainee

Almabetter, Bangalore

ABSTRACT

This paper uses Python and its external data processing packages to conduct an in-depth analysis of Airbnb Booking NYC 2019 review data.

Airbnb is an online marketplace for renting out homes/villas/ private rooms. The website charges a commission (3 to 20 percent) for every booking. We were provided with NYC 2019 csv file with classified labels in our data set.

Our experiment can help to understand that what could be the reason for the classification of such labels by feature selection, data analysis with Python and Matplotlib for data visualization taking into account to determine the correct documentation.

1. PROBLEM STATEMENT

Data provided by an Airbnb Booking aggregator services. Their customers can download their app on smartphones and book a property/villas from anywhere in the cities they operate in as well from the website whichever is suitable according to the customers. They, in turn, search for property/villas from various Airbnb service providers and provide the best option to their clients across available options. During this period, they have captured surge pricing types, reviews, facilities, locations from the service providers.

The main objective is to build a predictive model, which could help them in predicting the surge pricing type, room type, reviews proactively. This would in turn help them in matching the right property/villas with the right customers quickly and efficiently.

Each column in the dataframe gives us information about the property.

- name → the property is set by the host
- host_id and host_name →identification ids of the host for Airbnb
- There are five groups in neighbourhood_group, given in the datasets.
- neighbourhood →tells us which specific neighbourhood in the group the property belongs to
- latitude and longitude →give us the coordinates of the location. We can use this with folium to map all the locations
- room_type→ indicates the type of room the property is
- price will be the attribute we will try to predict
- minimum_nights →are the minimum number of nights the property has to be booked for
- number_of_reviews, last_review, and reviews_per_month→ give us information about the reviews of each property.

calculated_host_listings_count and availability_365→ tell us how many total properties the host has, and how long this property is available in a year.

2. INTRODUCTION

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. Airbnb offers people an easy, relatively stress-free way to earn some income from their property.

Data analytics is simply the analysis of various data. It is mainly used for business decision-making. Many libraries are available for doing the analysis. We have used the following libraries, for example, NumPy, Pandas, Seaborn, Matplotlib, pyplot, etc.

- **NumPy:** NumPy is a library written in Python, and used for numerical analysis in Python.
- **Pandas:** Pandas are mainly used for converting data into tabular form.
- **Matplotlib:** Matplotlib is a data visualization and graphical plotting package for Python and its numerical extension NumPy that runs on all platforms.
- **Seaborn:** Seaborn is a Python data visualization package based on matplotlib that is tightly connected with pandas' data structures.
- **Plot:** matplotlib. Pyplot is a collection of command-style functions that make Matplotlib work like MATLAB.

3. DATA EXPLORATION

a) **Data collection:** Data is the most important unit in any study that can be analysed to prefer a solution to the given problem statement.

b) **Exploratory data analysis:** Once the datasets are cleaned and free of error, they can then be analysed. A variety of techniques can be applied such as exploratory data analysis- understanding the messages contained within the obtained data and descriptive statistics finding the average, median, etc.

```
#Summary of the data
airbnb_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 48895 entries, 0 to 48894
Data columns (total 16 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   id                              48895 non-null  int64
 1   name                            48879 non-null  object
 2   host_id                         48895 non-null  int64
 3   host_name                       48874 non-null  object
 4   neighbourhood_group             48895 non-null  object
 5   neighbourhood                   48895 non-null  object
 6   latitude                        48895 non-null  float64
 7   longitude                       48895 non-null  float64
 8   room_type                       48895 non-null  object
 9   price                           48895 non-null  int64
 10  minimum_nights                  48895 non-null  int64
 11  number_of_reviews               48895 non-null  int64
 12  last_review                     38843 non-null  object
 13  reviews_per_month               38843 non-null  float64
 14  calculated_host_listings_count  48895 non-null  int64
 15  availability_365                48895 non-null  int64
dtypes: float64(3), int64(7), object(6)
memory usage: 6.0+ MB
```

c) **Data cleaning:** The method of cleaning data after it has been processed and organized is known as data cleaning. It identifies duplicates, and errors in data to remove them. The data cleaning process includes tasks such as record matching and data sorting.

```
#checking null values
airbnb_data.isnull().sum()
```

```
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
dtype: int64
```

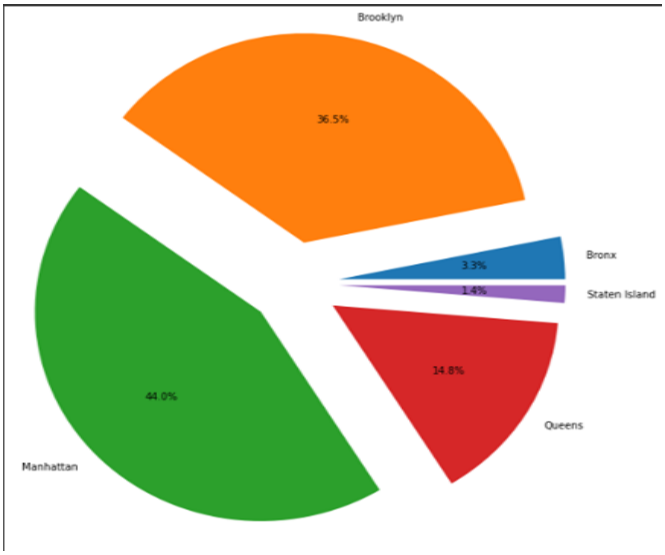
Here we are having some missing values, so we have to remove these missing values. After removing missing values, we get

```
airbnb_data.isnull().sum()
```

```
id                                    0
name                                  0
host_id                               0
host_name                             0
neighbourhood_group                   0
neighbourhood                         0
latitude                              0
longitude                             0
room_type                             0
price                                 0
minimum_nights                        0
number_of_reviews                     0
last_review                           0
reviews_per_month                     0
calculated_host_listings_count        0
availability_365                      0
dtype: int64
```

d) **Modelling and algorithms**: Mathematical formulas or models (known as algorithms), may be applied to the data to identify relationships among the variables; for example, using correlation or causation.

.



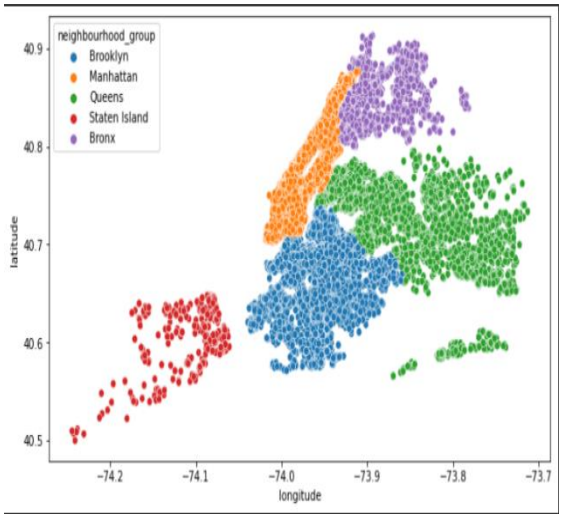
Data are obtained from the Airbnb NYC dataset shows the total number of neighbourhood and the room available in percentage, Manhattan is the highest room

available with 44% and Stalen Island is the least shared room available with only 1.4%.

4. Overall Analysis of Airbnb data and its representation separately:

Analysis of a given data is the simplest form of representing data to predict the solution. Uni means one, so in other words, the data has only one variable. Univariate data requires analyzing each variable separately. Data is represented separately

a) Bi-variate analysis on Airbnb: - Bivariate analysis is one of the simplest forms of quantitative analysis. It involves the analysis of two variables, to determine the empirical relationship between them. Bivariate analysis can help test simple hypotheses of association.

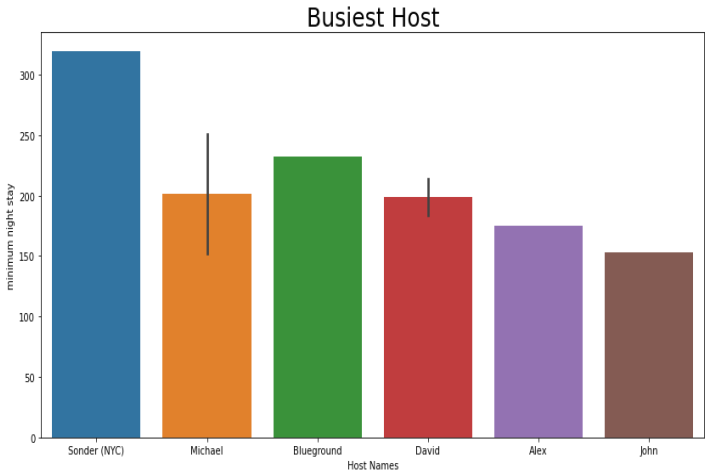


This shows us the dataset distribution in New York city with respect to latitude and longitude.

5.Understanding relationships and new insights through plots:

We can get many relations in our data by visualizing our dataset. Let's go through some techniques in order to see the insights.

a) Bar plot: A bar plot is basically used to aggregate the categorical data according to some methods and by default it's the mean. It can also be understood as a visualization of the group by action.



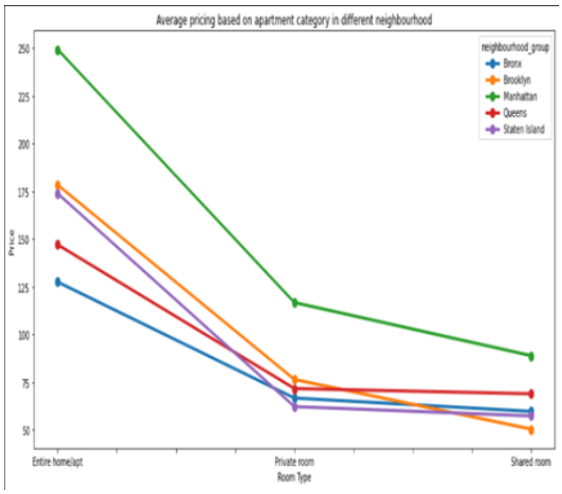
Busiest Hosts are:

1. Sonder(NYC)
2. Michael
3. Blueground

Because these hosts listed room type as Entire home and Private room which is preferred by most number of people and they are providing the rooms in the preferred locations

b) **Line Plot:** Line plots can be created in Python with Matplotlib's pyplot library.

To build a line plot, first import Matplotlib. It is a standard convention to import Matplotlib's pyplot library as plt.



This line plot shows the Average pricing based on apartment category in different neighbourhood in which Manhattan has highest avg price based on the Entire home/apt.and Brooklyn has least avg price based on shared apt.

6. Conclusion:

With the help of above performed Exploratory Data Analysis and visualization in Airbnb Dataset, suggest us about several insights. EDA help us to detect obvious errors, identify outliners in datasets. We use EDA to ensure the results produce are valid and applicable to any desired business outcomes and goals. Data cleaning concept of missing values if not handled properly then inaccurate interference occurs in the data. It can lead to wrong prediction and classification.

Pie Chart shows the total number neighbourhood and the room available in percentage in which Manhattan is the highest room available with 44% and Stalen Island is the least shared room available with 1.4% only. Bar plot shows that Entire and Private room is highly preferred by the hosts (Sonder, Michael, Blue ground). Line plot shows that Manhattan has highest average price based on Entire room type and Brooklyn has least average price based on shared room type.