Airbnb Booking Analysis

Parth Sharma-Cohort Amsot

Abstract:

The aim is to establish an approach to understand and draw the insights from Airbnb dataset through exploratory data analysis and provide insights potentially useful for its regulation. The document reveals Airbnb is spread across the neighbourhood of New York City, its popularity among the customer, their business and the most appropriate airbnb type across the NYC.Here we will understand and learn the relationship between Airbnb room type in a particular neighbourhood and housing prices. Finally the document provides the conclusive insights that are drawn from the data.

1. Problem Statement

Airbnb has become a very popular choice among travelers around the world for the kind of unique experiences that they provide and also for presenting an alternative to costly hotels. The data provided contains list of the hosts in the different neighbourhood groups of NYC.

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) behaviour and performance on the platform, guiding marketing initiatives, implementation of innovative additional services and much more.

The objective behind the Exploratory Data Analysis is to get the insight of the data and learn about the different factors that are relates with the availability, neighbourhood, price, neighbourhood groups, room type and the spread of the airbnb across the NYC. Here, we have a short summary of the input features.

- id: ID for entry
- name: This represents the somewhere the name of the hotels or apartment and somewhere a short description.
- host_id: Numerical data type has the id of the host.
- host_name: contains the name of the host.
- neighbourhood_group: Categorical data, NYC is divided into some neighbourhood groups.
- neighbourhood: Category showing different neighbourhood in these neighbourhood groups.
- latitude: Numerical data type , geographic position is defined in terms of latitude and longitude. This column gives the latitude.

- longitude: Numerical data type , geographic position is defined in terms of latitude and longitude. This column gives the longitude.
- room_type: It contains 3 categories private rooms, entire home/apt, shared room.
- price: It contains the total price per day for each host.
- minimum_nights:It represents the minimum number of nights people stayed in airbnb.
- reviews_per_month: This input variable represents the number of reviews a host got per month .
- last review: This column represents the last review date by any customer at a particular host.
- calculated_host_listings_count: this column shows us the number of times a particular host showed up in the dataset i.e. depending upon the room type, neighbourhood group how many times the host is present.
- availability_365:It showed the availability of the rooms across different host associated to airbnb

2. Introduction

Airbnb has become a very popular choice among travelers around the world for the kind of unique experiences that they provide and also for presenting an alternative to costly hotels. The data provided contains list of the hosts in the different neighbourhood groups of NYC.

The thing that this study focuses on is a clear understanding and getting the insights of the Airbnb dataset, to get clear thoughts on various input features that are available with us.

3. Steps involved:

Earlier, in the previous section we have just understood and got familiar with the data that is present with our Airbnb dataset. Now we will get to the further steps of data exploration and visualization.

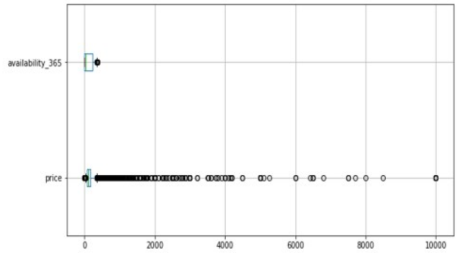
- **Exploratory Data Analysis:**
  EDA or Exploratory Data Analysis is the critical process of performing the initial investigation on the data to find the anamolies in our data and shape it such that it is useful for taking some insights to sole our purpose. There are certain step that we follow initially we will clean our data and make it free from anamolies such as Nan values,missing values and such values that could hinder the accuacy of our analysis.
  After loading the dataset first we will study the input variables and depending upon the questions that we discover the answer we will drop those columns that are of least relevance to us.

- **Libraries used:**
  - NUMPY.
  - PANDAS.
  - MATPLOTLIB.
  - SEABORN.

- **Understanding the data:**
  After loading the dataset first we will study the input variables and then learned about the data type of each column. Studied the descriptive summary of the dataset provided.

- **Data Cleaning and Null values Treatment**
  Firstly, after having the understanding of the data, went to look up for the column containing the missing values.
  Depending upon the questions that we discover the answer we will drop those columns that are of least relevance to us for e.g. id.
  Our dataset contains a large number of null values especially in the last review, name, host name as well as the review per month which might tend to disturb our accuracy hence we dropped them at the beginning of our analysis in order to get a better outputs. So, as said earlier we will drop the columns with least relevance to the scope of our analysis. Impute mean, median or zero where required.

- **Dealing with the Outliers :**
  We went to check for the outliers in our numerical d-type input variables and found out that the price column is extremely skewed towards the right and hence decided to do the further analysis by the taking the median of the price rather then mean.
  Also shown below the plot that shows that by how much degree the price column is skewed towards the right and availability has very less skewness.



- **Visualisation**
  Data visualization is the most important step while doing the analysis. It is more impressive, interesting and understanding when we represent our study or analysis with the help of colours and graphics. Using visualization elements like graphs, charts, maps, etc., it becomes easier to understand the underlying structure, trends, patterns and relationships among variables within the dataset.
  The libraries use for the plotting and visualisation are **Matplotlib** and **Seabon**.
  Matplotlib is a cross-platform, data visualization and graphical plotting library for Python and its numerical extension NumPy. As such, it offers a viable open source alternative to MATLAB.
  A Python matplotlib script is structured so that a few lines of code are all that is required in most instances to generate a visual data

plot. The matplotlib scripting layer overlays two APIs:

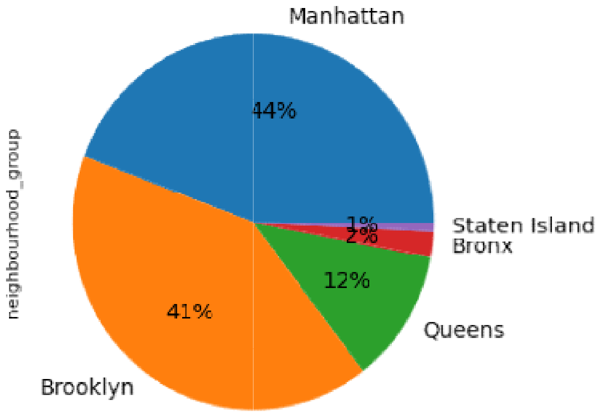
- The pyplot API is a hierarchy of Python code objects topped by matplotlib.pyplot
- An OO (Object-Oriented) API collection of objects that can be assembled with greater flexibility than pyplot. This API provides direct access to Matplotlib's backend layers.

Note: During the course of our visualisation both the above described scripts are used.

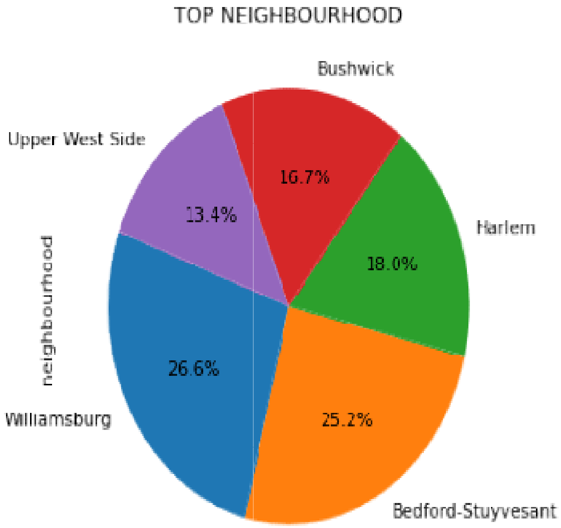
Seaborn is a library in Python predominantly used for making statistical graphics. Seaborn is a data visualization library built on top of matplotlib and closely integrated with pandas data structures in Python. Visualization is the central part of Seaborn which helps in exploration and understanding of data.

Some of the plots are show below:

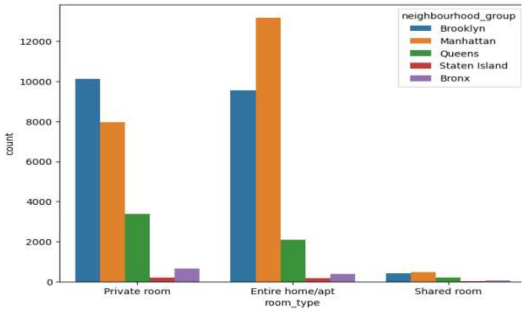
- **<u>Neighbourhood group with most Airbnb:</u>**
  - ➢ Here we have an observation that 44% of the overall airbnb are in Manhattan alone, followed by Brooklyn.
  - ➢ Least number has been shown by the Staten Island.



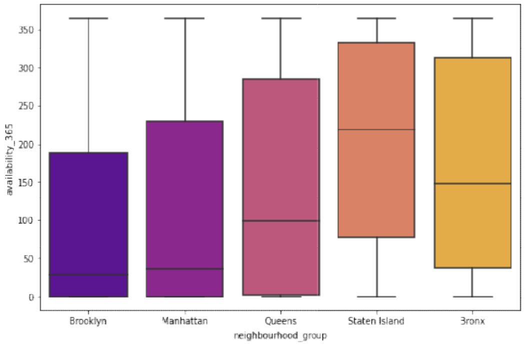
- **<u>Top 5 neighbourhoods with most Airbnb:</u>**
  Here we can observe that Williamsburg as a neighbourhood posses the maximum number of listings followed by Bedford-Stuyvesant.



- **<u>Preferred room type for each neighbourhood:</u>**
    - ➢ Interesting to understand from the plot that Entire home in Manhattan are the most while Brooklyn is tops in the Private rooms type



- **<u>Availability for each room type in each neighbourhood group:</u>**
    - ➢ It is quite interesting to asses the availability of different room type across different neighbourhood group across the year



8. Conclusion:

- Manhattan has the major proportion of the overall Airbnb present in the NYC, almost 44% i.e. 21k,followed by Brooklyn.
- Staten Island has the minimum number of the airbnb across NYC, its share is approximately 1% i.e. 373 airbnb.
- The most present room type across NYC is Entire home/apt i.e. 52%.Private rooms are also present in a greater chunk.
- Shared rooms are the least present in the overall NYC neighbourhood.
- Fort Wadsworth as a neighbourhood with maximum average price per day.
- Fort Wadsworth belongs to Staten Island neighbourhood group
- Williamsburg as a neighbourhood posses the maximum number of listings approximately 3917 followed by Bedford-Stuyvesant 3713.
- Upper West Side has the minimum number of 1969 airbnb across all the neighbourhood of NYC.
- Manhattan as a neighbourhood group is the most expensive across all the room type whether it is Entire home/apt, private room or shared room.
- Brooklyn is the second most expensive neighbourhood across Entire home/apt and Private rooms.
- Entire home/apt in Manhattan are the most while Brooklyn has the most number of Private rooms type.

- Though very less but Manhattan has the most number of Shared rooms.
- People choose Entire home/apt are a preferred room type to stay.
- Almost 12500 people accommodated in airbnb for 1 day only.
- Almost 3500 people used airbnb for almost a month and we can assume that this group might the one that used the private home/apt as room type.
- Brooklyn has the maximum number of reviews out the entire neighbourhood group followed by Manhattan.
- Staten Island has the minimum numbers of reviews out of all the neighbourhood groups

References-

1. Towards Data Science.
2. Tutorials point.
3. Analytics Vidhya.