**Airbnb Bookings Analysis- EDA**

**Upasana kumari,**

**Lonka ArunTeja,**

**Abriti nanda**

**Data Science Trainees**

**Alma Better, Bangalore**

**Abstract:**



Airbnb, Inc. is an American company that works an online marketplace for lodging, primarily homestays for vacation rentals, and tourism activities.

They collected various data which was helpful in finding solutions for different business problems and help to implements additional services and customer relationship.

Our Exploratory Data Analysis (EDA) can help to understand business relations such as relation between hosts and hosting listings count, Max prices in different neighborhood groups, finding busiest Hosts, surge in minimum nights at different neighborhood groups considering type of room.

***Keywords: Exploratory Data analysis, Business performance, business decisions, understanding customer behavior, implementation of additional services.***

**1.Problem Statement**

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.

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

* **Various Data collected:**
* **id**: unique reference number for each different hotel.
* **name**: name of different hotels of various neighborhood groups.
* **host\_id**: unique reference id of each individual host.
* **host\_name**: name of host hosting different hotels.
* **neighbourhood\_group:** aggregate group of neighborhood cities of some particular regions.
* **neighbourhood:** cities present in NYC.
* **latitude:** latitude is a geographic coordinate that specifies the north–south position of a point on the Earth's surface. Latitude is an angle which ranges from 0° at the Equator to 90° at the poles.
* **longitude**: Longitude is a geographic coordinate that specifies the east–west position of a point on the Earth's surface, or the surface of a celestial body.
* **room\_type**: Different room types available for booking, which contains Private room, Entire home/apt, Shared room.
* **price:** price per each night stay of different room types at various hotels.
* **minimum\_nights**: minimum nights booked in particular hotel.
* **number\_of\_reviews**: count of reviews got for each hotel.
* **last\_review:** date of last review got by a customer to a particular hotel.
* **reviews\_per\_month:**  count of reviews getting per month of a particular hotel.
* **calculated\_host\_listings\_count**: It represents total number of listings made by a specific host. In some cases, the properties are same but some of the other features differ like(room\_type).
* **availability\_365:** number of available days for booking in a year.

**2. Introduction**

### Airbnb collects various useful data and uses for the analysis for business development and customer relationship. It is a company for lodging, homestays, vacations, and tourism activities. So indirectly the company is dealing with Real Estate, although their business is running on the internet through their website. They are collecting large amount of data by their visitors from the website of Airbnb. After checking the data, we can analyze and understand the different features those are affecting Airbnb business and client-side requirements.

### Data is from the New York City, USA region. So, this project will help the Airbnb for their business improvement in the New York City. Also, from this data Airbnb could change modify their future plans for giving better experience to their clients.

### **Introduction to Exploratory Data Analysis**



* Exploratory Data Analysis (EDA) is the process of visualizing and analysing data to extract insights from it. In other words, EDA is the process of summarizing important characteristics of data in order to gain better understanding of the dataset.
* EDA is very essential because it is a good practice to first understand the problem statement and the various relationships between the data features before getting your hands dirty.

**Understanding EDA**

To understand the steps involved in EDA, we will use Python as the programming language and Jupyter Notebooks because it’s open-source, and not only it’s an excellent IDE but also very good for visualization and presentation.

***Technically, The primary motive of EDA is to***

* Examine the data distribution
* Handling missing values of the dataset(a most common issue with every dataset)
* Handling the outliers
* Removing duplicate data
* Encoding the categorical variables
* Normalizing and Scaling

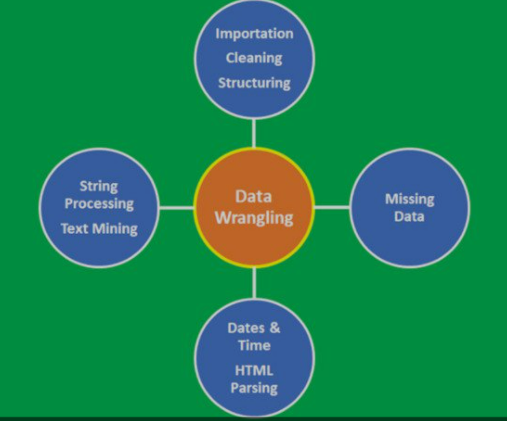
## **3. Questions we have focused in the entire EDA.**

* 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?

# **4.Importing Python libraries & Stage setup**

* A Python library is a collection of related modules. It contains bundles of code that can be used repeatedly in different programs. It makes Python Programming simpler and convenient for the programmer.
* Python have more than 137,000 libraries, Python can be used to create applications and models in a variety of fields, for instance, machine learning, data science, data visualization, image and data manipulation, and many more…
* In this Project we are mainly Using some libraries.(i.e., Numpy,Pandas,matplotlib,seaborn,etc..)
* After importing libraries, we will come up with an imaginary story to explain the data
* Data visualization also helps to tell stories by curating data into a form easier to understand, highlighting the trends and outliers.
* A good visualization tells a story, removing the noise from data and highlighting the useful information.
* Good visualization combined with daily life stories/any well-known frictional characters, they can impact more and we will understand simply and remembering easily ..

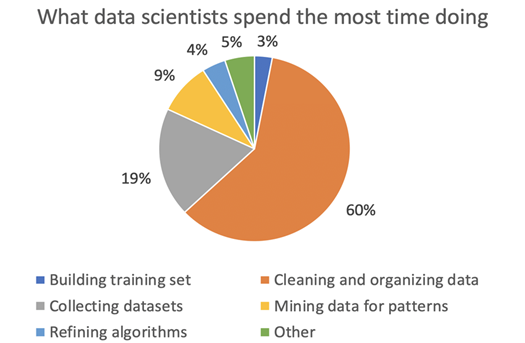
# **5. Data Wrangling:**



Data pre-processing is an important task. It is a data mining technique that transforms raw data into a more understandable, useful and efficient format.

Data has a better idea. This idea will be clearer and understandable after performing data pre-processing.

The data cleaning process detects and removes the errors and inconsistencies present in the data and improves its quality. Data quality problems occur due to misspellings during data entry, missing values or any other invalid data.



## **Dropping unnecessary data:**

## 

## As “id”, “last\_review” and “reviews\_per\_month” have more than 10,000 null values, it affects the outcomes of Data analysis; So, we are removing these columns and also as we are not doing any analysis specifically on latitude and longitude, we're also removing these variables as well.

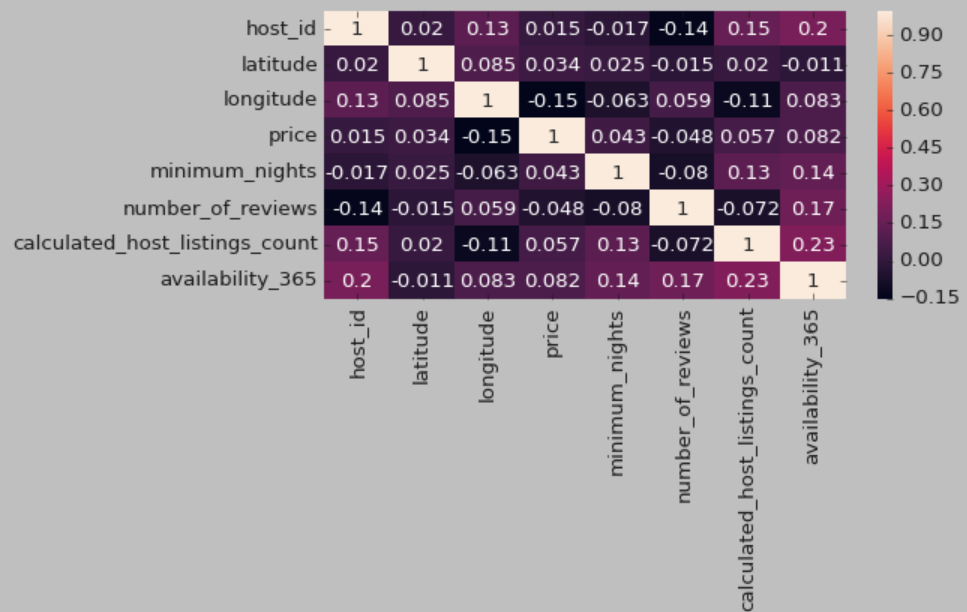
## **Verifying Data quality:**

We have gone through whole data and checked null values and reviewed any missing data or wrong data. And prepared whole data ready for exploratory data analysis

## **Basic data exploration:**

Using describe () and info () and size functions of pandas. Gone through a basic exploration of data before entering into EDA.

**6. Exploratory Data analysis:**

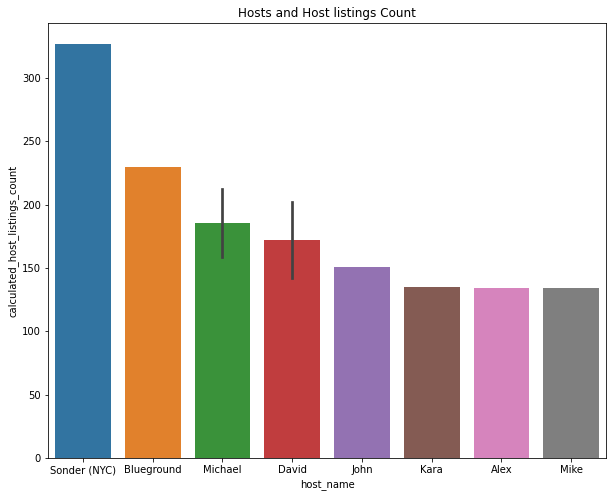
* **Tools used**
* **Programming language**: Python
* **Libraries**: Pandas, Seaborn, Matplotlib
* **MS Excel**
* **Tableau**
* **Correlation between data**
* ****

**Map of New York city (NYC)**

* **Top Hosts and their listings count:**

For analyzing this we used group by function and took 'host\_name', 'neighbourhood\_group', and

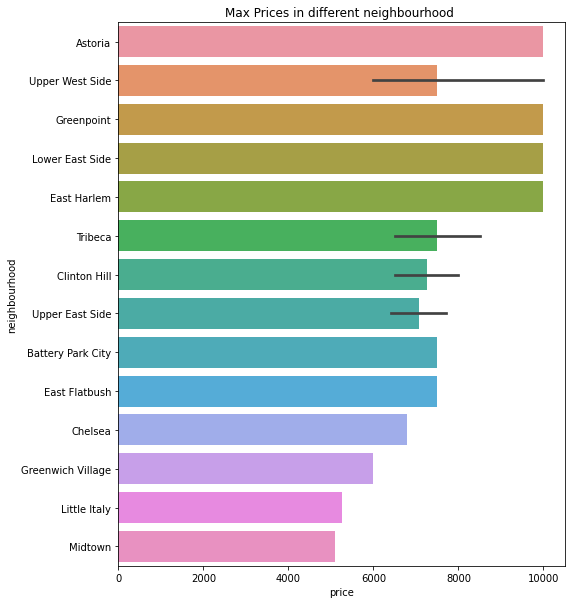
"calculated\_host\_listings\_count" and calculated the top 10 hosts.



**Here are some interesting findings, we found fascinating:**

* 8 out of top 10 hosts are from the ''Manhattan'' neighbourhood group.
* 2 out of top 10 hosts are from the ''Brooklyn" neighbourhood group.
* [ Sonder (NYC), Blue ground, Michael, David, John, Kara, Alex, Mike] are the top hosts of Manhattan neighbourhood group.
* [ Michael, David] are the top hosts in Brooklyn neighbourhood group.
* Manhattan neighbourhood group hosts are out-performing in listings.
* **Max prices in different neighborhood:**

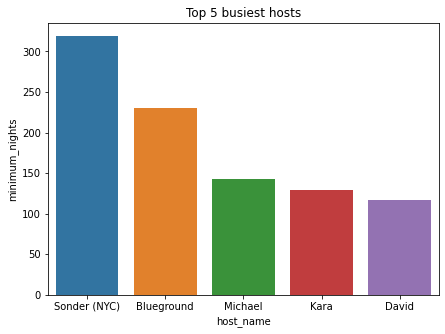
For this question, we approached with, 'name','neighbourhood\_group','neighbourhood','price','minimum\_nights', 'number\_of\_reviews'



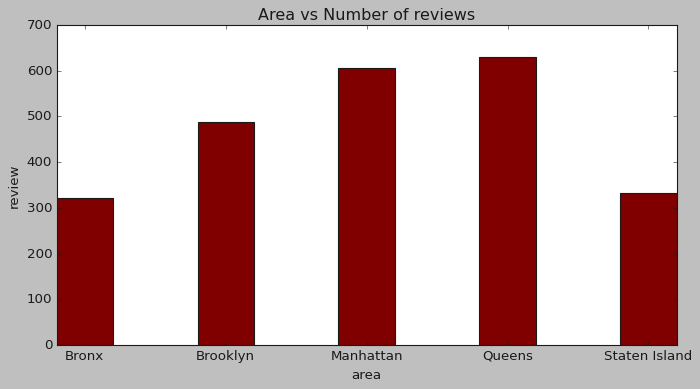
**Price Vs Location analysis shown some interesting facts:**

* The highest price is 10,000 $ and can be seen in Astoria, Upper west side, lower east side, Greenpoint and East Harlem.
* Most minimum nights spent in top 20 price list are,
  + Luxury TriBeCa Apartment at an amazing price - 180 nights
  + Furnished room in Astoria apartment - 100 nights
  + Quiet, Clean, Lit @ LES & Chinatown - 99 nights
* Highest priced rooms (i.e., 10,000usd) are all present in Manhattan, Brooklyn and Queens neighborhood group.
* **Finding busiest hosts:**

Next, we were interested in finding the busiest hosts by considering “minimum nights” bookings in their hotels. We took, ‘host\_name','neighbourhood\_group','room\_type', "minimum\_nights" and got the following result.



* 5 out of top 5 are all from " Manhattan" neighbourhood group.
* Sonder (NYC), Blue ground, Michael, Kara, David are the top 5 most busiest hosts.



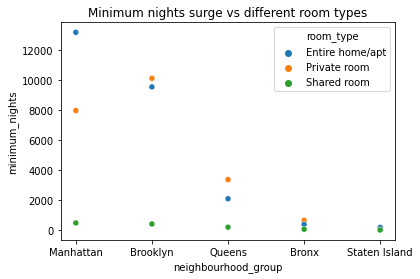
Next, we were interested in finding the busiest hosts by considering “Reviews” bookings in their hotels.

* **Traffic among different room types and different neighborhood**

Our main motive through this step was to find traffic among different types of rooms at different neighborhood groups.

For this we took, 'neighbourhood\_group', 'room\_type', 'minimum\_nights' to analyze this question.

Result we got from above analysis



**Key findings:**

* In Manhattan, people are preferring "Entire Home/apt".
* But, in Brooklyn, Queens and Bronx people are preferring private rooms.
* In Staten Island, people are having equal preference over all three types of rooms.

**7.Tableau dashboarding:**

We used tableau for analyzing data

Visually and to make various relations on tableau dashboard.

Below are some of the visualizations of relations we have worked on,

* Neighbourhood Group and Host Listings Count.
* Max Number of Reviews.
* Max Prices in Different Neighbourhood Groups.
* Maximum Number of Nights in Different Room Types.
* Costly locations based on price vs location relationship on maps.

Dashboard

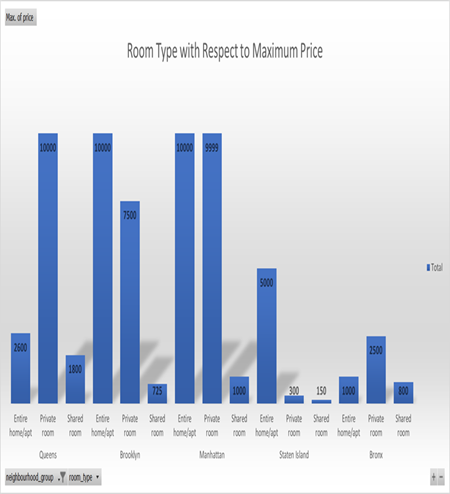


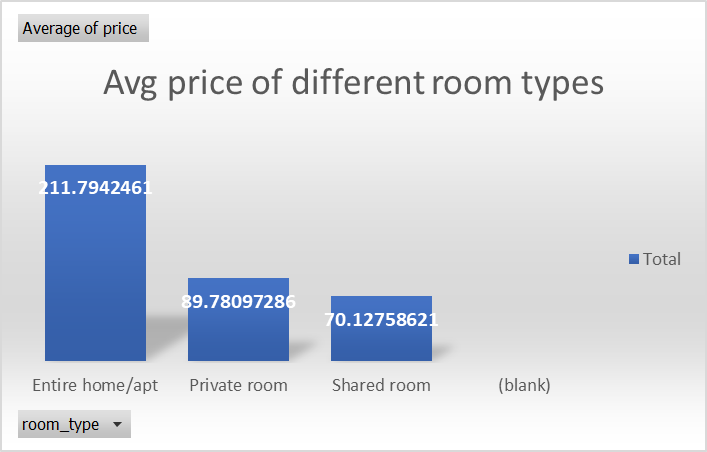
**8. EDA using MS Excel:**

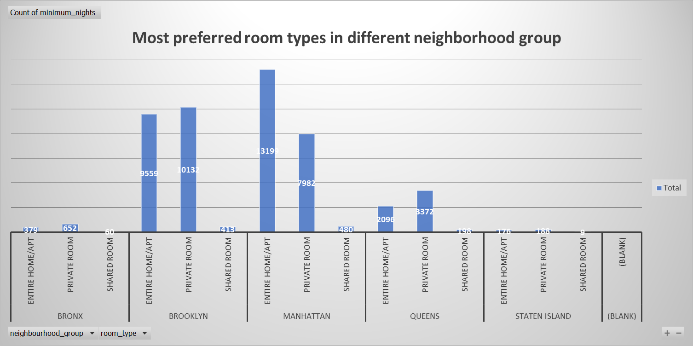
Analyzed data and made graphical representation of below relations.

Avg price in different neighborhood

Avg prices of different room types



****

****

**9. Challenges faced:**

* Verifying quality of such huge data and looking for error values.
* Dropping down irrelevant data and making the whole data getting ready for full pledged data analysis.
* Understanding and visualizing complex numerical data, and communicating business solutions.
* Analyzing and solving various queries and presenting clear cut outputs.

**10. Conclusion:**

1. The people who prefer to stay in Private room they won't stay longer
2. The people who prefer to stay in Entire home or Apartment they are going to stay bit longer in that particular Neighborhood only.
3. Most people prefer to pay less price accommodation.
4. More number of Reviews for particular Neighborhood group that means that place is a tourist place.
5. If people are not staying more then one night means they are travelers.
6. We proceeded with analyzing boroughs and neighborhood listing densities and what areas were more popular than another.

**11. Summary:**

From the entire analysis on Airbnb bookings analysis, Our assumptions before analysis went totally different after getting results from the analysis.

Airbnb dataset-2019 appeared to be a very rich dataset with a variety of columns that allowed us to do deep data exploration on each significant column presented.

First, we have found hosts that take good advantage of the Airbnb platform and provide the most listings; we found that our top host has 327 listings. After that, we proceeded with analyzing boroughs and neighborhood listing densities and what areas were more popular than another.

From the entire analysis on Airbnb bookings analysis, our assumptions before analysis went totally different after getting results from the analysis. The whole EDA process gave very fascinating results and insights that will be helpful for business development and expansion, budget allocations and focusing on things people prefer.

**References-**

1. AlmaBetter
2. Kaggle
3. Coursera
4. W3schools
5. Greeck for greeck
6. Analytical vidhya

***Thank you***