Exploratory Data Analysis-

AIRBNB BOOKING ANALYSIS



* **By Kaiwalya Zankar**

**DataScience Trainee At AlmaBetter**

* **INTRODUCTION-**

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

Airbnb connects people who want to rent out their homes with people looking for accommodations in that locality. New York is the most populous city in the United States, and one of the most popular tourism and business places globally.

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

* **OBJECTIVE-**
* Data Prepping/Wrangling.
* Checking the Null values for cleaning the Dataset for further analysis.
* Checking the unique values for Analyzing the Dataset for further analysis
* Exploration of Neighborhood group, Neighborhood, Room type, Price, Reviews.
* Correlation between variables.
* Analysis of V hosts and key findings.
* **PROBLEM STATEMENT-**

Since 2008, guests and hosts have used Airbnb to expand on travelling possibilities and present a more unique, personalized way of experiencing the world, today, Airbnb became one-of-a-kind services that is used and recognized by the whole 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 data set has around 49000 observations in it with 16 columns and it is a mix between categorical and numeric values.

* **METHODOLOGY- Exploratory Data Analysis**

In this EDA, an analysis is done on a dataset which helps us to visualize the different factors in data which help us to conclude the different insights, predictions of the data,

In this project I have 1st clean the data, make the data as much we can utilize each factor to its maximum limit, then I have done the EDA to obtain the perception from the given data.

* **DATA PREPPING-**

The dataset contains Airbnb Booking with respect to the factors associated with it from 2008 to 2022. In which there are total 16 variables having 48895 observations in which there are 3 categorical variables and rest of numerical variables. In this dataset the observations represent the features of hotel booking, with respect to availability of resources such as price, neighborhood, neighborhood type, type of room available, reviews of that particular room.

As the dataset is real business data, it can be used for scientific and educational purposes and many other fields.

Start from here->

**Columns used in the analysis-**

* **id**
  + The id column represents the unique id allocated for Airbnb room by the firm itself
* **name**
  + The name column represents the name or description of the particular property
* **host\_id**
  + This column represents the unique id given by the Airbnb to the particular host
* **host\_name**
  + This column represents the name of the host or the owners of that particular property
* **neighborhood\_group**
  + New York City is split up into five boroughs and this boroughs is called as neighborhood groups which are  [the Bronx](https://en.wikipedia.org/wiki/The_Bronx), [Brooklyn](https://en.wikipedia.org/wiki/Brooklyn), [Manhattan](https://en.wikipedia.org/wiki/Manhattan), [Queens](https://en.wikipedia.org/wiki/Queens), and [Staten Island](https://en.wikipedia.org/wiki/Staten_Island). Each [borough](https://en.wikipedia.org/wiki/Borough_(New_York_City)) (neighborhood\_group) has the same boundaries as a [county](https://en.wikipedia.org/wiki/County_(United_States)) of the state. The county governments were dissolved when the city consolidated in 1898, along with all city, town, and village governments within each county. The term *borough* was adopted to describe a unique form of governmental administration for each of the five fundamental constituent parts of the newly [consolidated city](https://en.wikipedia.org/wiki/History_of_New_York_City_(1898-1945))
* **Neighborhood** 
  + Basically, the neighborhood is the cities located in the neighborhood group.
  + And this column represents the neighborhood of the property
  + In the given data set we have around 221 unique neighborhoods (cities).
  + We can also consider the neighborhood as the tehsil.
* **Latitude**
  + This column represents the measurement of the property distance from north or south of the equator
* **longitude**
  + This column represents the measurement of the property distance fromeast or west of the prime meridian.
* **room\_type** 
  + This column represents the type of room that property consist of.
  + In our data set there are basically three types of room type: Entire home/apartment, Private room and Shared room.
* **Price** 
  + This column represents the price of that property per night
* **minimum\_nights**
  + This column represents the least numbers of nights that is required for the customers to stay in that property
* **number\_of\_reviews**
  + This column represents that how many reviews got by that property.
* **Last\_review**
  + This column represents the last date of the review given to that property
* **Reviews\_per\_month**
  + This column represents the average number of reviews given to the particular property in throughout month.
* **Calculated\_host\_listing\_count**
  + This column represents the number of hosts for the property
* **Availability\_365**
  + This column represents that for how many days the current property is available in 365 days.
  + That is number of days the property available in year.
* **CHALLENGES FACED-**
* **Size**: The size of the dataset right away is the strength as well as the most challenging aspect of this dataset. Although there are 16 fields that describe every factor, it’s only natural that such a wide range of information may not be available for every property This results in a large number of missing or 0 values.
* **NaN values**: NaN values in the dataset had to be cleaned up accordingly. For instance, NaN values in Numerical and categorical fields like last\_review, reviews\_per\_month and name, host\_name were replaced with **unknown** and **not\_known** respectively.
* **APPROACH USED-**

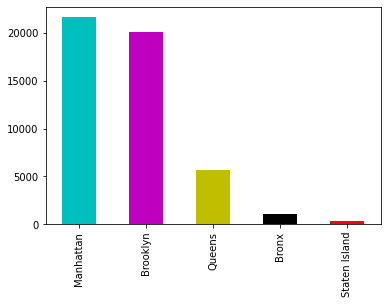
As the data includes 16 columns which are both numerical and categorical. To perform the analysis and different visualizations of factors associated with the Airbnb booking, we performed the pre-processing of data which could be removing NAN values or creating a new dataframe itself.

And, decided the factors which are important for analysis from the view of EDA.

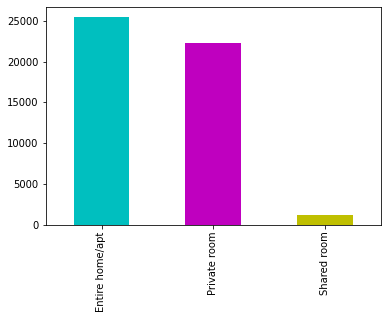
* **TOOLS USED-**
* **NumPy: -** NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.
* **Pandas: -** pandas is a [software library](https://en.wikipedia.org/wiki/Software_library) written for the [Python programming language](https://en.wikipedia.org/wiki/Python_(programming_language)) for data manipulation and [analysis](https://en.wikipedia.org/wiki/Data_analysis). In particular, it offers [data structures](https://en.wikipedia.org/wiki/Data_structure) and operations for manipulating numerical tables and [time series](https://en.wikipedia.org/wiki/Time_series). The name is derived from the term "[panel data](https://en.wikipedia.org/wiki/Panel_data)", an [econometrics](https://en.wikipedia.org/wiki/Econometrics) term for [data sets](https://en.wikipedia.org/wiki/Data_set) that include observations over multiple time periods for the same individuals. Its name is a play on the phrase "Python data analysis" itself.
* **Matplotlib: -**Matplotlib is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack. One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram etc.
* **Seaborn: -** Seaborn is a library for making statistical graphics in Python. It builds on top of [matplotlib](https://matplotlib.org/) and integrates closely with [pandas](https://pandas.pydata.org/) data structures. Seaborn helps you explore and understand your data. Its plotting functions operate on dataframes and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots. Its dataset-oriented, declarative API lets you focus on what the different elements of your plots mean, rather than on the details of how to draw them.
* **ANALYSIS-**

**Exploratory Data Analysis-**

1] Exploration of Neighborhood Group-

****

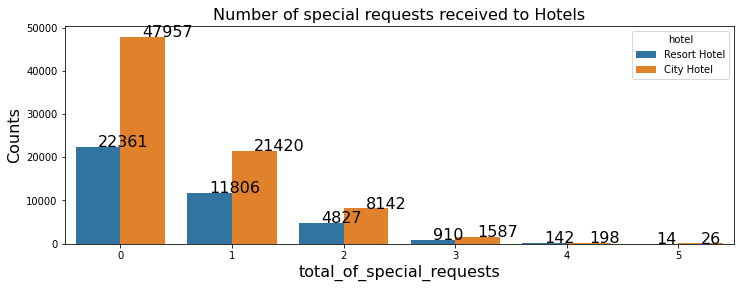
2] Exploration of Roomtype



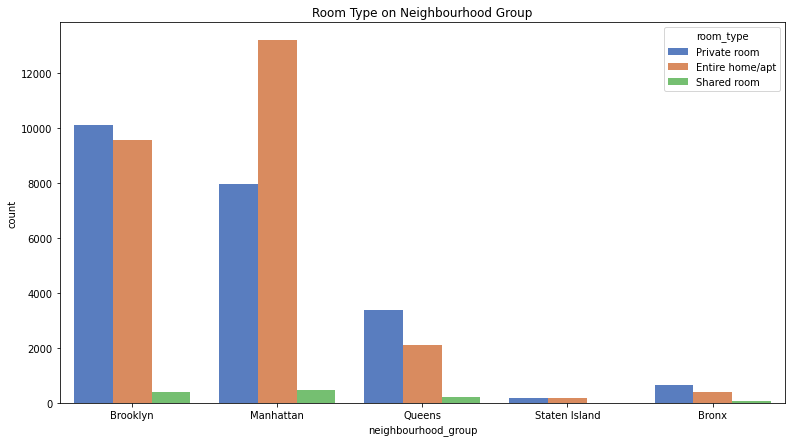
3] Availability of Room-



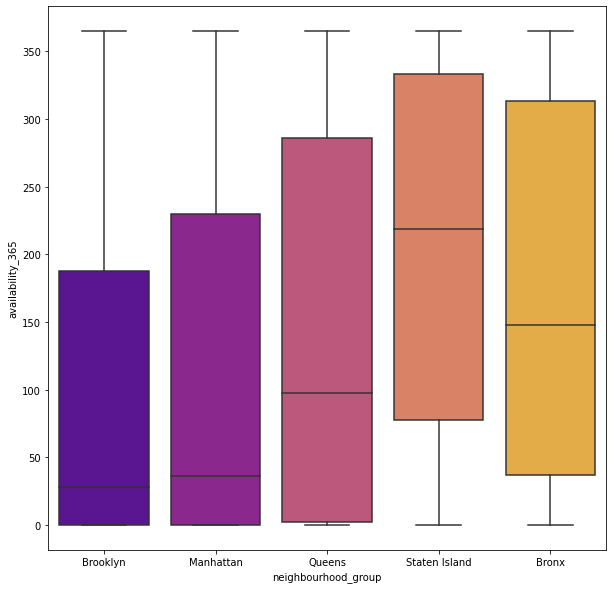
4] Number of special requests received from customers-



5] Room types and Neighborhood Group**-**



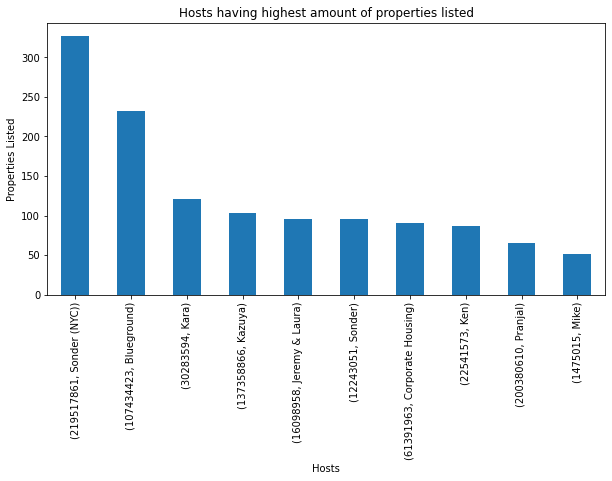
6] relation between neighborhood group and availability of room-



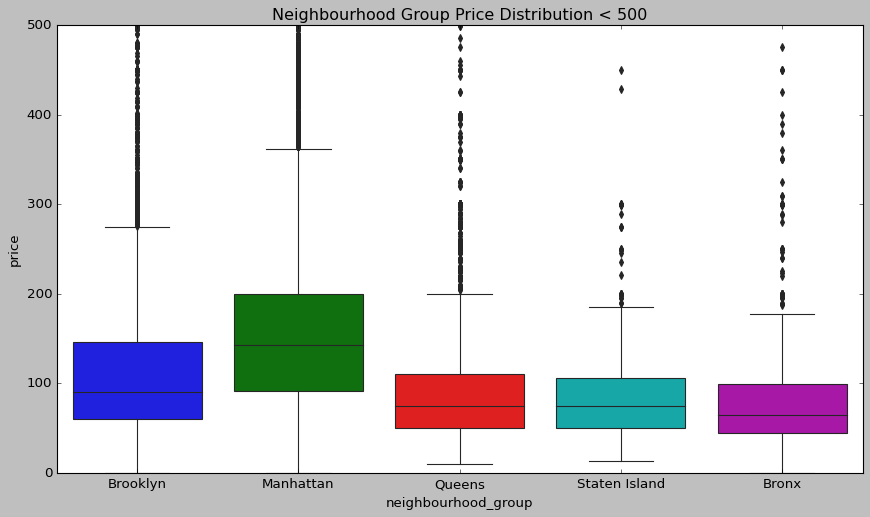
7] What can we learn about different hosts and areas

|  |  |  |  |
| --- | --- | --- | --- |
|  | **host\_name** | **neighbourhood\_group** | **calculated\_host\_listings\_count** |
| **13217** | Sonder (NYC) | Manhattan | 327 |
| **1834** | Blueground | Manhattan | 230 |
| **9742** | Michael | Manhattan | 212 |
| **3250** | David | Manhattan | 202 |
| **9741** | Michael | Brooklyn | 159 |
| **6808** | John | Manhattan | 151 |
| **3249** | David | Brooklyn | 142 |
| **7275** | Kara | Manhattan | 135 |
| **432** | Alex | Manhattan | 134 |
| **9856** | Mike | Manhattan | 134 |

8] Hosts having highest amount of properties listed-



9] Neighborhood group with price-



**Descriptive Analysis-**

Descriptive Statistics is summarizing the data at hand through certain numbers like mean, median etc. so as to make the understanding of the data easier. This means that the descriptive statistics are just the representation of the data (sample) available and not based on any theory of probability.

### Top 10 Entire home/apt listing with location

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Host\_id | Host\_name | Neighborhood | Neighborhood group | apartment |
| **20219** | 219517861 | Sonder (NYC) | Financial District | Manhattan | 210 |
| **6452** | 7503643 | Vida | Greenpoint | Brooklyn | 52 |
| **8434** | 12243051 | Sonder | Financial District | Manhattan | 51 |
| **20222** | 219517861 | Sonder (NYC) | Murray Hill | Manhattan | 50 |
| **17840** | 120762452 | Stanley | Murray Hill | Manhattan | 49 |
| **19908** | 205031545 | Red Awning | Midtown | Manhattan | 49 |
| **12278** | 30283594 | Kara | Hell's Kitchen | Manhattan | 41 |
| **17458** | 107434423 | Blueground | Chelsea | Manhattan | 37 |
| **17635** | 113805886 | Yaacov | Upper East Side | Manhattan | 33 |
| **12280** | 30283594 | Kara | Midtown | Manhattan | 33 |

## What can we learn about different hosts and areas

|  |  |  |  |
| --- | --- | --- | --- |
|  | Host\_name | Neughborhood  Group | Calculated host  Listing count |
| **13217** | Sonder (NYC) | Manhattan | 327 |
| **1834** | Blueground | Manhattan | 230 |
| **9742** | Michael | Manhattan | 212 |
| **3250** | David | Manhattan | 202 |
| **9741** | Michael | Brooklyn | 159 |
| **6808** | John | Manhattan | 151 |
| **3249** | David | Brooklyn | 142 |
| **7275** | Kara | Manhattan | 135 |
| **432** | Alex | Manhattan | 134 |
| **9856** | Mike | Manhattan | 134 |

* Is there any noticeable difference of traffic among different areas and what could be the reason for it.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Neighborhood**  **Group** | **Room type** | **Minimum**  **nights** |
| **6** | Manhattan | Entire home/apt | 13199 |
| **4** | Brooklyn | Private room | 10132 |
| **3** | Brooklyn | Entire home/apt | 9559 |
| **7** | Manhattan | Private room | 7982 |
| **10** | Queens | Private room | 3372 |
| **9** | Queens | Entire home/apt | 2096 |
| **1** | Bronx | Private room | 652 |
| **8** | Manhattan | Shared room | 480 |
| **5** | Brooklyn | Shared room | 413 |
| **0** | Bronx | Entire home/apt | 379 |
| **11** | Queens | Shared room | 198 |
| **13** | Staten Island | Private room | 188 |

* **CONCLUSION-**
* From the given Dataset, there are 2 unique room type and 5 neighborhood group ranking followed by Entire Apartment> Private Room > Shared Room and for neighborhood group ranking followed Manhattan at the top and Staten Island at bottom Neighborhood
* Sonder (NYC) has the highest number or properties in given data
* Williamsburg has the highest properties of Brooklyn neighborhood group and also Brooklyn has a greater number of Private Rooms unlike Manhattan has entire apartment
* Also, the price of Manhattan is most expensive unlike shared room is least expensive one.
* From the analysis we can say that people are preferring entire home/apt over private which are present in Manhattan, Brooklyn, Queens and people preferring listing which are less in price

**THANK YOU !!!!...**