**EDA PROJECT ON AIRBNB NYC DATASET**

**by:-Prashant singh**

**Sachin**

**Jaanhvi khandelwal**

**Pankaj kumar pandit**

**Submitted to:-Almabetter**

# **Abstract:**

This is the final technical report of our data analytic project titled “EDA on Airbnb NYC” as a part of our Data analytic course at Alma better. The goal is to analyze and predict the price and other variables in the New York Airbnb data. Also, a recommendation system will be built to recommend Airbnb listings according to the user preference.

# **Content:**

Airbnb is an online marketplace for arranging or offering lodging, primarily home stays, or tourism experiences. The company does not own any of the real estate listings, nor does it host events; it acts as a broker, receiving commissions from each booking. The company is based in San Francisco, California, United States.The company was conceived after its founders put an air mattress in their living room, effectively turning their apartment into a bed and breakfast, to offset the prohibitive cost of rent in San Francisco; Airbnb is a shortened version of its original name, AirBedandBreakfast.com Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present a more unique, personalized way of experiencing the world. Today, Airbnb became one-of-a-kind service that is used and recognized by the entire 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 dataset had around 49,000 observations in it with 16 columns and it is a mix between categorical and numeric values.

# **Problem Statement**

Explore and analyze the data to discover key understandings (not limited to these) such as:-

* 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?

# Neighborhood Group vs. Availability Room.

* Properties in every neighborhood group.
* Map of neighborhood group.
* Price distribution in every types of room.

# **Approach:-**

# **1.**Let us first check our dataset's and understand it.

# **2.**Later we will check for any missing data in the data given. Does it hamper our analysis?

**3.**We would check the type of data and divide it for our analysis.

**4.**We checked where there any outlier or unethical data in it if so, we would filer such data for specific analysis.

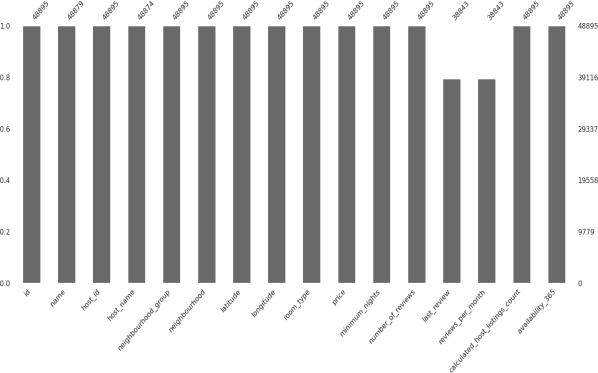
Number of quantitative variables:10 Number of qualitative variables:6 Attributes:

id, name,

host\_id, host\_name, neighborhood-group, neighborhood, latitude, longitude, room type, price, minimum nights, number\_of\_reviews,

last\_review, reviews\_per\_month, calculated\_host\_listings\_count, availability\_365

# **Checking Data for any missing values**



# .

Since our dataset's contain several missing values preprocessing must be done. Missing values will either be deleted or replaced with the column mean or nan. based on how important the attribute is. Also, with respect to preprocessing the datatype of certain attributes like last\_review must be changed to make processing easier. Our main goal is to analyze and find interesting patterns between the variables in our dataset’s. Visualization is an important aspect of finding patterns. Hence several visualization techniques like bar graph, pie chart, Violin chart, correlation, etc. will be plotted to gain insights. We then plan to predict certain variables such as price by using predictive models. Several models will be explored and models with the best accuracy will be selected. We also got that

Few columns like name, host name, last review had many missing values and then we replaced it with “missing”. Importance for analysis, hence they were deleted.

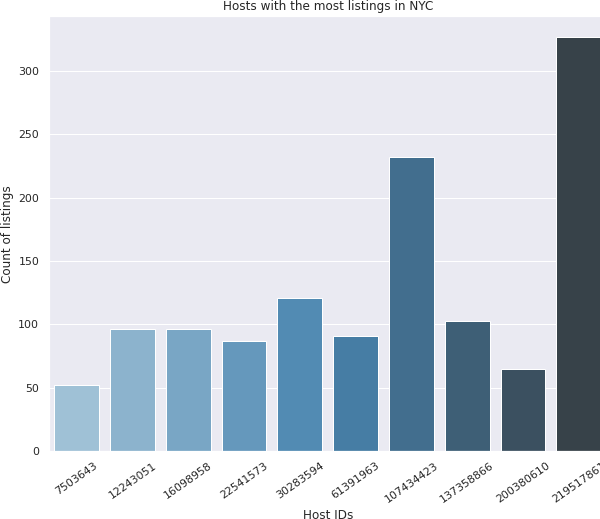
Reviews per month column had lot of missing rows but is important for analysis, hence missing values were replaced with the mean of that column.

**EXPLORATORY DATA ANALYSIS**

Now we know that we are ready for an exploration of our data, we can make a rule that we are going to be working from left to right. The reason some may prefer to do this is due to its set approach - some datasets have a substantial number of attributes; plus, this way we will remember to explore each column individually to make sure we learn as much as we can about our dataset's.

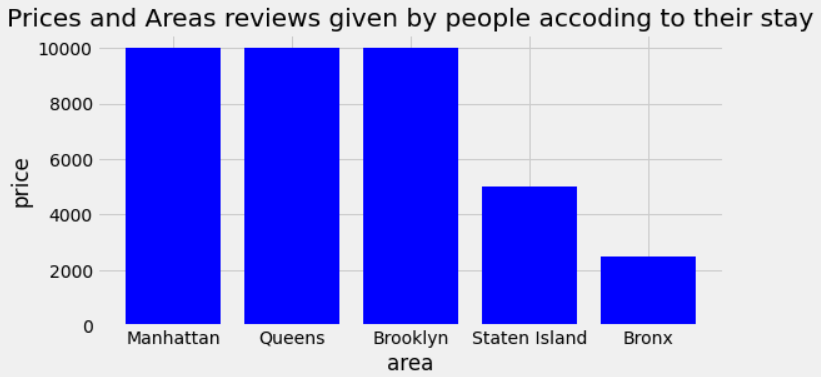
# 

# **Observation 1:-**What can we learn about different hosts and areas?

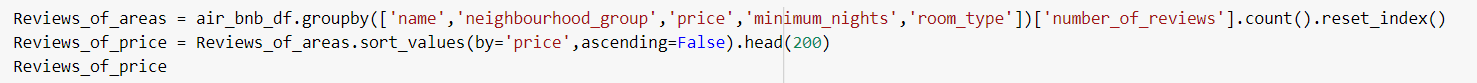


We first take the Host id and check the maximum count list. So, we could get our top busiest hosts. We then plot it in a bar chart and find that host id 219517861 is the busiest host with more than 350 count list and is followed by 107434423.

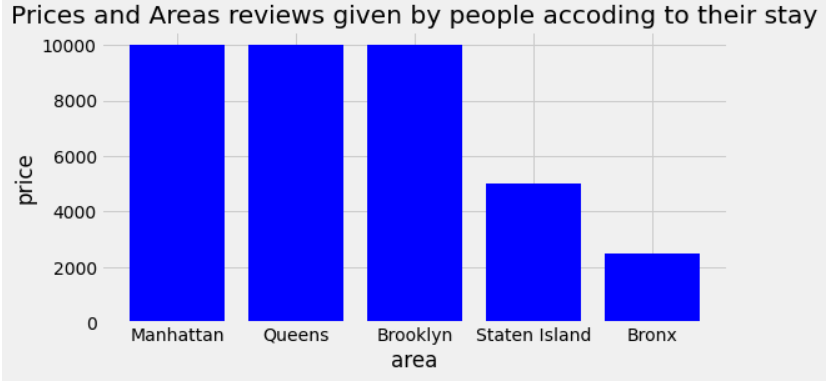
**Observation 2:-**What can we learn from predictions?



First we make a separate Dataframe using groupby function as shown in the fig.



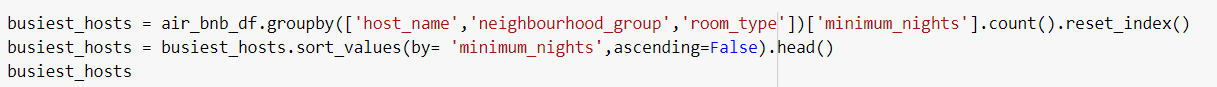
Then we plot a graph between price and area distribution

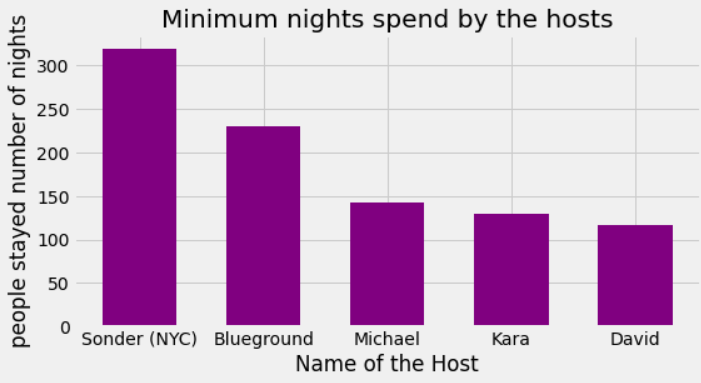


As we can see in the graph above as shown as result in DataFrame that most of the people prefer to stay at the places where prices are low and most of them have taken Entire home/apt. But in "Queens" minimum number of nights were stayed by people it means here people stays a bit longer most of them have taken private room.

**Observation 3:-**Which hosts are the busiest and why?

First we make a saparate Dataframe using groupby function



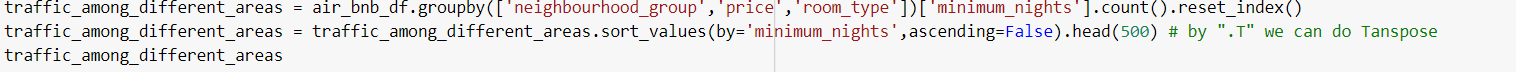
Then we plot a graph using plt.bar function having x label as name of hosts and y label as minimum night spent by the host.

As we can see from the graph above Most Busiest host among all is Sonder (NYC) after him Blueground, Michael, Kara, David and soo on...

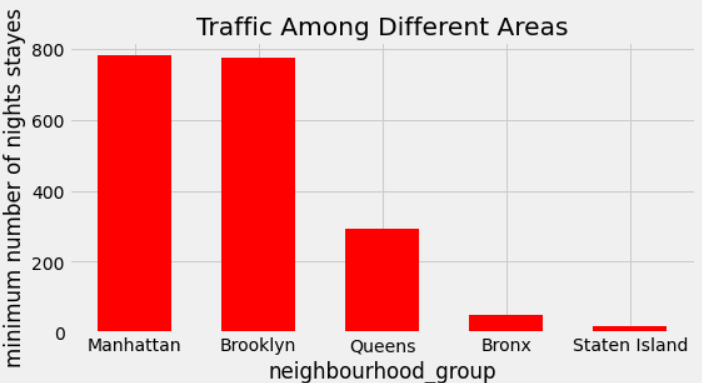
Host is busy because maintaining there place as you can see Manhattan is the most popular place and there Entire home/apt is prefered by most of the people.

**Observation 4:-**Is there any noticeable difference of traffic among different areas and what could be the reason for it?

First me make a saparate dataframe using groupby function having following columns



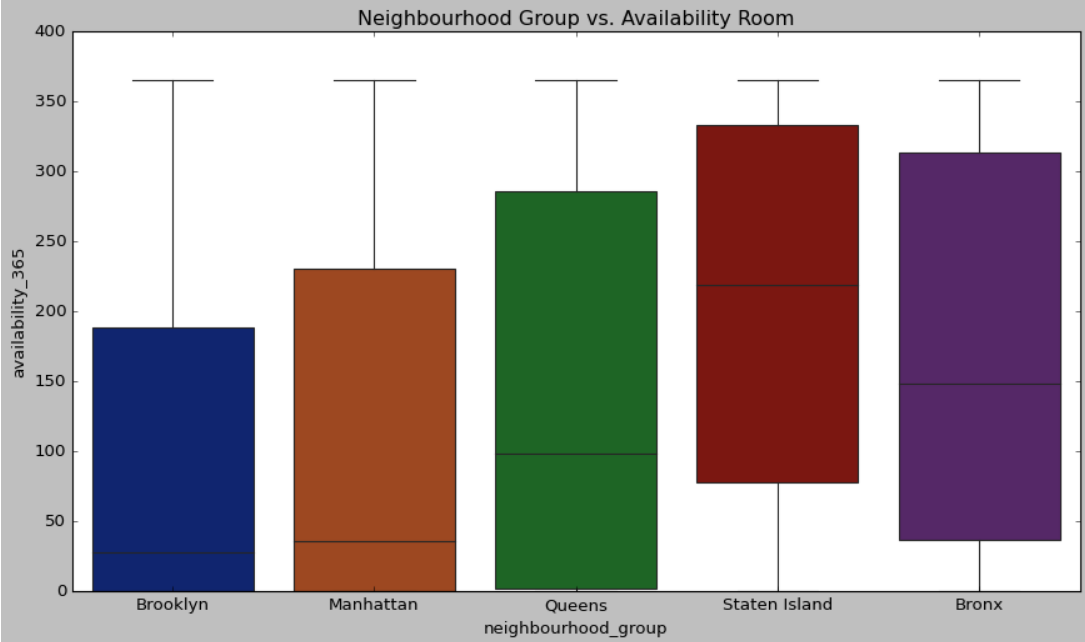
Then we drawn a graph using plt.bar function having x labeled neighbourhood\_groups and y labeled as minimum number of nights stayed by the guest.



As we can see from the graph above most of the people are preferring to rent homes for more than 700 nights these people mostly take whole apartments on rents for long period of time in manhattan and Brooklyn but on the polar opposites people in Bronx and Staten island does not prefer to stay long.

**Observation 5 :-**Neighborhood Group vs. Availability Room

We have used a box plot to show the availability of the rooms in various neighborhood groups having y labeled as availibility\_365 and x labeled as neighbourhood groups.



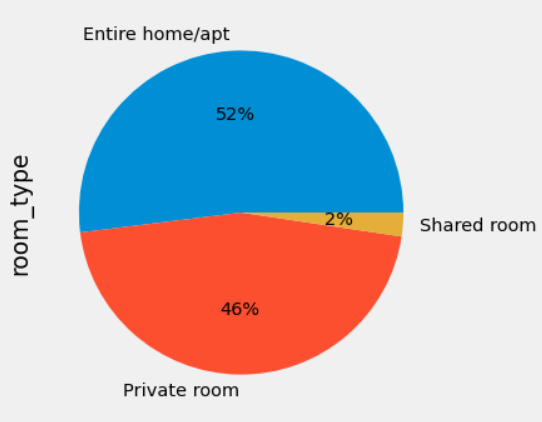
As shown in boxplot graph that in Brooklyn most of the properties are available between 0 to 200 little bit of number increased by manhattan and most diverse room availaibilitiy is of Queens.

**Observation 6 :-**Types of properties listed

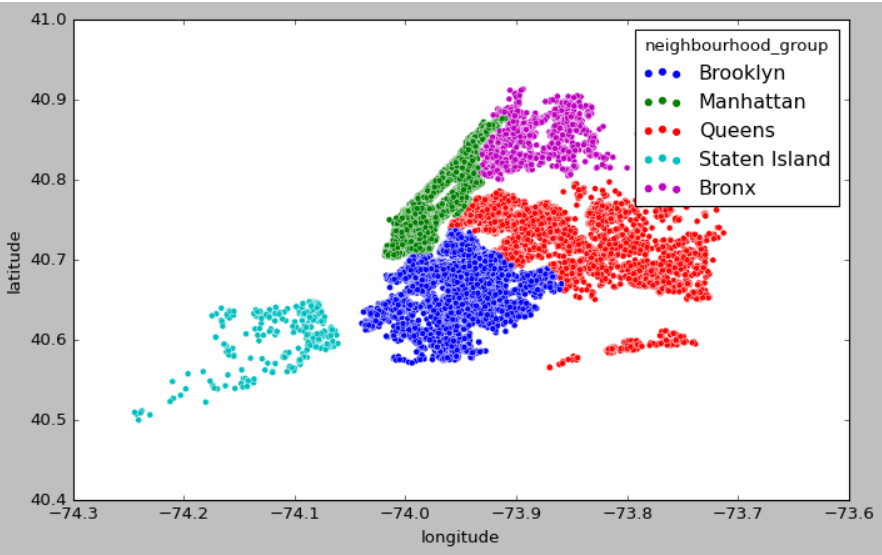
We plot a pie chart and find the various distribution of home property in NYC.Here we find that Entire home/apt has about 52% listing in NYC and shared rooms has least of 2%.Let’s check our Price Column which is the most important Key performer for every business to run. Let’s check what’s the minimum and maximum price of any room types

As we found that the minimum price is zero which is impossible let’s filter out these rows and Airbnb is to make business.There could not be zero price. So

removing the price value as zero as it is not possible. So drooping rows with price as Zero.Now, as we filter are price from zero price Value. Let’s check for any outliers in price column as its the most important column for our analysis for which we need to be cautious for our future analysis.



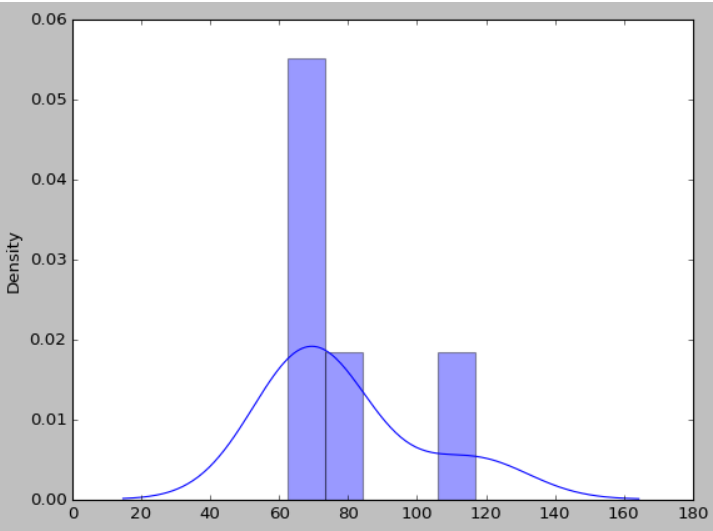
**Observation 7 :-**Map of Neighborhood group



**Observation 8 :-**Price distribution in different types of rooms

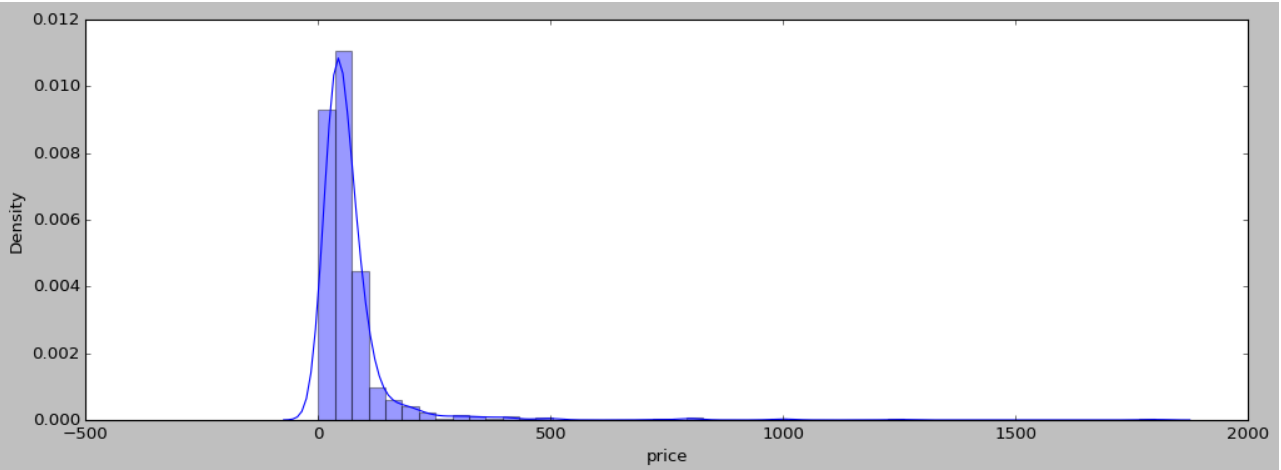
1.Private rooms:-

Private rooms on average are priced from 60-120 dollars per night on an average depending upon the neighborhood group it is located.



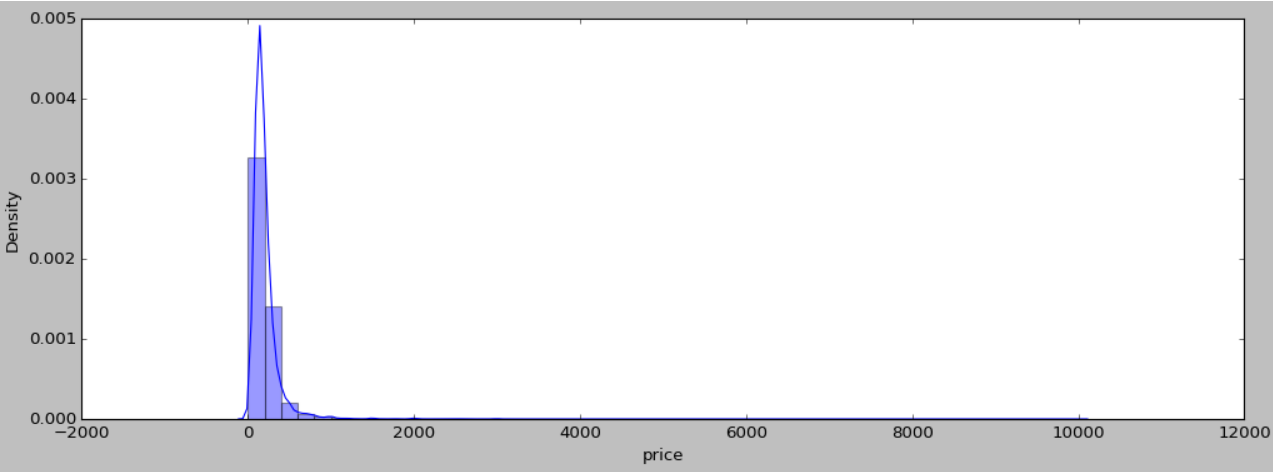
As shown in the graph most of the private rooms are starting from 20 dollars and goes upto 120 dollar but most the private rooms are popular in the price bracket of 60 to 120 dollars.

2.Shared rooms:-



Most of the shared rooms have the price range between 50-70 dollars per night depending upon the neighborhood groups.

Entire home/apt:-



Most of the Entire home/apt are priced between 0 to 1000.

# **Problem faced:-**

1- Huge Cluster of data with different price rates had an issue on how to remove the outliers.

2- Number of reviews was a great challenge to handle in different areas.

3- To plot a map in a street map was a task. 4- The regions have a great insight to say and needed to handle the latitude and longitude on a single street map easier.

5- Few team left in between due to some personal reason which put a load to submit the task.

# **Conclusion:-**

1 Host\_id with 219517861 has the greatest number of listings of 300+.

2. Entire Home apartment properties are more in Airbnb NYC.

3. Manhattan has the highest cost of living followed by Brooklyn.

4. Manhattan is the most loved place by people in NYC.

5. We have given a map of spread of prices at various location which makes it easy for visualize the spread of room types and its rate there.

6. Manhattan has the highest Entire Apartment property list in Airbnb NYC. and Brooklyn held the first for Private room type property.

# 

# 

# **References:-**

1.<https://www.almabetter.com/>(notes)

2. https://stackoverflow.com/questions/214741/wha t-is-a-stackoverflowerror

3. <https://startuptalky.com/airbnb-success-story/>

**Remarks:-These write up is a documentation of group work done for the project by Prashant singh,Sachin,Pankaj kumar pandit,jaanhvi khandelwal and is self- documented.**