**EDA ON AIRBNB**

**COHORT: MONTREAL**

**Abstract:**

Airbnb, as in “Air Bed and Breakfast,” is **a service that lets property owners rent out their spaces to travelers looking for a place to stay**. Travelers can rent a space for multiple people to share, a shared space with private rooms, or the entire property for themselves

We are given a data set from Airbnb which holds values related to the tourists looking for renting a space in New York City.

**1.Problem Statement**

**The given dataset has around 49,000 observations in it with 16 columns and it is a mix between categorical and numeric values , related to listings of Private, Shared or other type of rooms available in NYC !**

**id** :indicates the unique id of an entry in data set

**name** : indicates names of listings area in New York City

**host\_id** :indicates unique id for a host in New York City

**host\_name** : shows the name of the host

**neighbourhood\_group**: indicates names of the cities where these listings present

**neighbourhood**: displays the names of different neighbouring cities for each

neighbourhood\_group

**latitude**: displays latitude values of a listing.

**longitude**: displays longitude values of a listing.

**room\_type**: indicates the type of room available.

**price**: indicates the price.

**minimum\_nights**: indicates minimum number of nights spent.

**number\_of\_reviews**: indicates total number of reviews for a particular listing.

**reviews\_per\_month** :indicates reviews per month for a particular.

**calculated\_host\_listings\_count** : shows the count of each host for a certain listing

**availability\_365** : indicates the number of days a particular listing is available in a year.

**minimum\_price**: It’s a new column added which will give the product of **price** and **minimum\_nights.**

**2.Objective**

Airbnb provides lodging to travelers but not only limits to hotels but also provides with private rooms, shared rooms or entire apartments.

Following this trend, we have a data set with values related to Airbnb listings in New York City.

Our goal here is to analyze the data set and gather useful key points related to the listings, areas with their room types, prices, reviews, availability, hosts in every aspect so that in the end it benefits the Airbnb community including their customers.

**3. Approach Used**

* **Data Analyzing**

To go through the data set try to

comprehend the raw form to get as

bird eye overview of data set.

* **Data Cleaning**

To check through the values in

columns, fix all the blank, NULL

values, dealing with the duplicated

values, finding and then fixing outlier.

* **Correlation**

Before starting our EDA, a better step

is to know the relation of columns

with each other to know how change

in the value of a specific column is

affecting the rest

* **Exploratory Data Analysis**

EDA was done in 3 basic steps:

* **Univariate Analysis**
* **Bivariate Analysis**
* **Multivariate Analysis**

The difference between these is just

the number of variables being

compared at a time, in Univariate

Analysis we only compared a single

variable’s values, in Bivariate the

number of variables increases to 2 and

in Multivariate

Analysis 2 or more variables are

compared with each other to produce

result.

* **Visualizing Using Various Plots**

Finally, after comparing all the

variable values with each other, the

most important step is to visualize

the information so that the results

could loudly and clearly

communicate with the reader.

Bar graphs, Pie chart, Point plot,

Histograms, Correlation Heatmap

etc. were used in our case to

visualize the result.

**4. Challenges Faced**

Fixing missing values and null values in a data set is required before executing any operation in a data set via **Exploratory Data Analysis.**

Null and missing values from ‘name’ and ‘host\_name’ columns were fixed by replacing them by with ‘Unknown’ and ‘Host\_Unknown ’

The number of null values in ‘reviews\_per\_month’ column was quite high. Hence, replaced all the null values with 0 to allow any further numerical operation to execute in our analysis.

‘last\_review’ column is used for timestamp to evaluate month and year.

**Fixing Outliers:**

Using **Quantile** approach: By assigning a minimum threshold value (0.1percentile) and a maximum threshold (98percentile) and only taking the values in between to get a clean data set.

Using **Quartile** approach: By using Interquartile Range(IQR) Formula which is:

**Interquartile range = Upper Quartile – Lower Quartile where,** Lower Quartile(Q1)=((n+1)/4)th Term Upper Quartile(Q3)=(3(n+1)/4)th Term N = Number of elements in data set

Using Quantile approach was better as it was just enough to make our data set analysis ready.

**5. Conclusion**

* First of all we saw the correlation between various columns using a heatmap to get a upper hand in our later analysis and found out that there's correlation among **host\_id to reveiws\_per\_month** **&** **availability\_365**. And there's noticiable correlation between **min\_nights to no\_of\_listings\_count & availability\_365**. Price also shows some correlation with availability\_365 & host\_listings\_count.
* According to the availability and price mean graph **Bronx,Queens and Staten Island** best suited for customers whereas **Manhattan** is the least suitable because of its less availability and high price.
* After checking the pie chart is clearly understandable that
* "Manhattan" and "Brooklyn" are the most visited places by the hosts.
* After this we have "Queens" and then rest of the places.
* ok, We could also see that majority of the hosts prefer to book
* Entire home/apartment then private rooms and then finally shared rooms.
* **Queens** is the most reviewed **Neighbourhood group** in entire NYC with total reviews of **≈629** whereas **Bronx** is the least reviewed Neighbourhood group with **≈321** total reviews.
* Reviews per Month is highest in **Manhattan** (might be because it has the highest listing) followed by **Bronx,Queens,Broonkyl** and at last **Staten Island** (has the least number of listing).
* About the hosts **Michael** holds the most number of listing properties in NYC i.e ≈407 followed by **David** with ≈390 listings.
* Even after holding the most number of listings **Dona** proves to be The People's Champ as he is **Most reviewed host** with **≈630** reviews followed by **Jj** with **≈607** reviews.
* Density of Month graph showed that **JUNE** was the busiest month followed by **JULY** and **FEBRUARY** was the least. Density of Year graph showed that **2019** was the busiest of all years whereas **2011** and **2012** were the least.
* Concluded best suited Places for visitors on the basis of Price & Availability
* **Private Bedroom in Manhattan** is the most famous listing property as per total reviews **≈667** followed by **Room near JFK Queen Bed** with **≈630** total reviews.
* On the basis of monthly reviews **Enjoy great views of the City in our Deluxe Room!** is at the top with a total of **≈59** reviews per month.
* Top neighbourhood is **Bedford-Stuyvesant** with total reviews of ≈109482
* Using **latitude** & **longitude** values we actually marked the position of each listing in **each neighbourhood\_group** on the map  
  of **New York City** !!

Colab link:-

<https://colab.research.google.com/drive/1-fSbqSRAD8CAEqR-K5tMp7UIlYOBaIyj?usp=sharing>