Exploratory Data Analysis

CAPSTONE PROJECT

Prepared by:

Atharva Darvekar

Data Science

Abstract :

Airbnb is an American company operating an online marketplace for short- and long-term homestays and experiences. The company acts as a broker and charges a commission from each booking. The company was founded in 2008.With millions of listings on its platform, Airbnb generates a vast amount of data that can be used to gain insights into the behavior and performance of hosts and guests. Guests and hosts have used Airbnb to expand on traveling possibilities and present a more unique, personalised way of experiencing the world.This project aims to explore and analyze a dataset of approximately 49,000 Airbnb listings with 16 columns.The dataset includes both categorical and numeric variables.

Problem Statement :

1. How listing prices are distributed?
2. What are the listing counts across different neighbourhoods?
3. What are the average prices in each neighbourhood group?
4. Display price distribution using violin plot.
5. Display the top neighbourhoods by listing count.
6. Show top hosts based on listing count using bar chart.
7. How many hosts are there in each neighbourhood group?
8. Show the most reviewed room type.
9. Show average prices of different neighbourhoods in a geographical map.
10. What are the total counts of each room type?
11. What are the average prices of each room type?
12. How many reviews did each neighbourhood get?
13. How different types of rooms are distributed across NYC?
14. Display a geographical map that shows room types.
15. What are minimum stay requirements of listings?
16. What are the average prices of corresponding minimum stay requirements?

Business Objective :

Our main objective is to find out how different factors such as location, room type, etc affects the price of the listing and to recommend hosts where they can invest in properties to get better future returns.For this, we will explore and visualize the dataset from Airbnb in NYC using basic exploratory data analysis (EDA) techniques.This can help in making strategic data-driven decisions by the marketing team, finance team and technical team of Airbnb.

Imported Modules :

- Numpy:

  It supplies an enormous library of high-level mathematical functions.
- Pandas:

  It is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language.
- Matplotlib:

  It is a comprehensive library for creating static, animated, and interactive visualizations in Python.
- Seaborn:

  Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

Loading The Dataset:
Mounted using the google drive to load the dataset in the form of CSV(comma separated values)

Dataset Attributes:
The dataset contains exactly 48895 rows and 16 Columns with around 20000 missing values which should be cleaned before processing. There are 5 categorical columns namely name, host_name, neighbourhood group and room type and 11 numerical columns namely id, host_id, latitude, longitude,price, minimum_nights, number_of_reviews, last_review, reviews_per_month, calculated_host_listings_count and availability_365.

Renaming Columns:
We renamed the columns id, name, number_of_reviews, calculated_host_listings_count as listing_id, listing_name, total_reviews, host_listings_count respectively for better understanding.

Data cleaning:
Dropped columns: We have dropped the last_review column as there is no derivable information with that

Handling null values:
listing_name and host_name columns contain very few missing values. These missing values are replaced by not_known and no_name respectively using constant value imputation. reviews_per_month column contains nearly 10000 null values which are replaced by 0.

Outlier Removal:
By using box plot, we found out that price column has outliers. Hence we used an interquartile range function to remove outliers from price column.

Exploratory Data Analysis:

1. How listing prices are distributed?
→
1) The Range of the prices of the listings fall between 30 dollars to 350 dollars.
2) Majority of the listings have prices between 50 dollars to 200 dollars.

2. What are the listing counts across different neighbourhoods?
→
1) Manhattan and Brooklyn have the highest number of listings on Airbnb, with over 19,000 listings each.
2) Queens and Bronx have fewer listings compared to Manhattan and Brooklyn.
3) Staten Island has the lowest number of listings.

3. What are the average prices in each neighbourhood group?
→
1) The average price of a listing in New York City varies significantly across different neighborhoods, with Manhattan having the highest 146 dollars/day average price and the Bronx having the lowest near 77 dollars/day.
2) Most of the Listings in the Bronx have minimum prices among other neighbourhood groups.

4. Display price distribution using violin plot.
→
1) Price distribution is very high in Manhattan and Brooklyn. but Manhattan has more Diversity in the price range, as you can see in the violin plot.
2) Queens and Bronx have the same price distribution but in the Queens area more distribution in 50 to 100 but diversity in price is not like Manhattan and Brooklyn.

5. Display the top neighbourhoods by listing count.
→
1) The top neighborhoods in New York City in terms of listing counts are Williamsburg, Bedford-Stuyvesant, Harlem, Bushwick, and the Upper West Side.

6. Show top hosts based on listing count using bar chart.
→
1) Michael, David, John are the top hosts with 383, 368, 276 listings respectively.
2) There is a relatively large gap between the top two hosts and the rest of the hosts. For example, John has 276 listings, which is significantly fewer than Michael's 383 listings.

7. How many hosts are there in each neighbourhood group?
→
1) Manhattan has the largest number of hosts with 19501,Brooklyn has the second largest number of hosts with 19415.
2) Queens have 3rd most number of hosts with more than 5000 count.
3) The Bronx and Staten Island have very few hosts between 300 and 1100.

8. Show the most reviewed room type.
→
1) We can see that Private rooms received the most no of reviews/month where Manhattan had the highest reviews received for Private rooms.
2) Manhattan & Queens got the most no of reviews for Entire home/apt room type.
3) There were less reviews received from shared rooms as compared to other room types and it was from Staten Island followed by Bronx.

9. Show average prices of different neighbourhoods in geographical map.
→
1) Average price in Manhattan is high as compared to other boroughs.
2) Queens, Bronx and Staten Island tend to have a lower overall cost of living compared to Manhattan and Brooklyn.
3) These neighborhoods may be attractive to renters or buyers looking for more affordable housing options in the New York City area.

10. What are the total counts of each room type?
→

1) The majority of listings on Airbnb are for entire homes or apartments, with 22784 listings, followed by private rooms with 21996 listings, and shared rooms with 1138 listings.
2) Shared room listings are very less, comprising only 2.5% of the total share.

11. What are the average prices of each room type?
→
1) Entire home has a high average price of 160 dollars.
2) Private rooms and shared rooms have average prices 80 dollars and 60 dollars respectively.

12. How many reviews did each neighbourhood get?
→
1) Brooklyn has the largest share of total reviews on Airbnb, with 43.3%, followed by Manhattan with 38.9%.
2) Queens has the third largest share of total reviews, with 14.2%, followed by the Bronx with 2.6% and Staten Island with 1.0%.

13. How different types of rooms are distributed across NYC?
→
1) Manhattan has more listed properties with Entire home/apt around 24.6% of total listed properties followed by Brooklyn with around 19.5%.
2) Private rooms are more in Brooklyn as in 21.9% of the total listed properties followed by Manhattan with 16.9% of them. While 7.3% of private rooms are from Queens.

14. Display a geographical map that shows room types.
→
1) Geographical representation of New York city helps guests to easily check which types of rooms are there across different neighbourhoods.

15. What are minimum stay requirements of listings?
→
1) The majority of listings on Airbnb have a minimum stay requirement of 1 or 2 nights, with 12067 and 11080 listings, respectively.
2) The number of listings with a minimum stay requirement decreases as the length of stay increases, with 7375 listings requiring a minimum stay of 3 nights, and so on.

16. What are the average prices of corresponding minimum stay requirements?
→

1) Price of listing per night increases as minimum night requirement increases from 1 night to 4 nights and price decreases further as minimum night requirement increases.

Solution To Business Objective:

- Listings in Manhattan have high prices (around 150 dollars) compared to other boroughs of NYC.
- Manhattan is world-famous for its parks, museums, buildings, town, liberty, gardens, markets, island and also its substantial number of tourists throughout the year. So it makes sense that demand and price are both high in Manhattan.
- The majority of listings on Airbnb are for entire homes or apartments and also Private Rooms with relatively fewer listings for shared rooms.
- Listings with room type homes have high prices (around 160 dollars) as compared to private rooms (around 80 dollars) and shared rooms (around 60 dollars).
- Price of listing per night increases as minimum night requirement increases from 1 night to 4 nights and price decreases further as minimum night requirement increases.
- In simple terms, Guests are charged more for shorter stays and less for longer stays. (per night)
- The data indicates that there is a high level of competition among Airbnb hosts, with a small number of hosts dominating a large portion of the market. So hosts can consider investing in property in areas with relatively fewer listings in order to differentiate themselves from the competition.

Conclusion:

- In this project, we had to find insights and patterns using different types of graphs from the airbnb dataset to make data driven decisions.
- The dataset contained about 49000 records, and 16 attributes.
- We began by importing necessary libraries like pandas, numpy, matplotlib, seaborn, etc.
- Then we dealt with the dataset's missing values using constant value imputation technique.
- We dropped duplicate rows and also dropped columns which are not required for analysis.
- We removed outliers from the price column using the interquartile range method.
- Once our dataset became ready, we created various charts for data analysis.
- It was found that Manhattan and Brooklyn have high demand for airbnb rentals and the largest number of listings.

- Manhattan and Brooklyn also have the highest number of hosts, indicating a high level of competition in these boroughs.
- Some hosts have a higher number of listings which goes around 200 to 300 while other hosts have relatively fewer listings. Hence, Small number of hosts dominates the larger portion of the market.