Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:

1. **Vaibhavkumar Gupta** (vaibhavguptavkg@gmail.com)
   - Analyzed availability_365 column
   - Filled price column price=0 with respective median price
   - Plotted Categorical Plot using Klib library
   - Analyzed host_name and host_id
   - Analyzed the relation of reviews_per_month with each neighbourhood
   - Analyzed listings according to neighbourhood_group
   - Analyzed distribution of price column across each room_type
   - Found out average price in each neighbourhood_group for each room_type for getting good number of reviews
   - Figured out a metric for knowing busiest hosts

2. **Bhavik Ashokkumar Verma** (vermabhavik585@gmail.com)
   - Analyzed price column
   - Replaced NA values in reviews_per_month by 0
   - Plotted Pearson's correlation matrix using Klib library
   - Analyzed host_name and host_id
   - Analyzed the relation of total_number_of_reviews with each neighbourhood
   - Plotted Map using Folium and added the markers by using list of latitudes and longitudes.
   - Analyzed room_type distribution across each neighbourhood_group
   - Found out average price in each neighbourhood_group for each room_type
   - Helped to understand correlation with various columns to figure out good metric for knowing busiest hosts

3. **Dilkhush Sharma** (kumardilkhush.rds@gmail.com)
   - Analyzed neighbourhoods
   - Analyzed room_type distribution across each neighbourhood
   - Analyzed price trend across neighbourhood
   - Analyzed room_type distribution on Airbnb
   - Plotted multiple scatter plots and bar plots
   - Analyzed hosts having highest number_of_reviews
   - Helped in knowing busiest hosts

4. **Priyanka Pal** (palpriyanka00029@gmail.com)
   - Analyzed neighbourhoods
   - Analyzed room_type distribution across each neighbourhood_group
   - Analyzed hosts having highest reviews_per_month
   - Analyzed price trend across neighbourhood
   - Analyzed room_type distribution on Airbnb
   - Helped in knowing busiest hosts

5. Shayan Somanna (shayan.somzz@gmail.com)
   - Analyzed neighbourhood_group
   - Analyzed room_type distribution across each neighbourhood_group
   - Analyzed price trend across neighbourhood_group
   - Analyzed room_type distribution on Airbnb
   - Analyzed hosts having highest reviews_per_month
   - Helped in knowing busiest hosts

Please paste the GitHub Repo and Google Drive Folder link.

Github Link: - https://github.com/dil150898?tab=repositories
Google Drive Folder:
https://drive.google.com/drive/folders/14icjBJzuGFtSr8kBewta0Gli7iLK7Xl7?usp=sharing

Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)

About the dataset:

The Airbnb dataset has around 49000 observation and 16 columns.

The dataset has mix data type of numerical, unique, date and categorical type.

Airbnb, as in "Air Bed and Breakfast," is a service that lets property owners rent out their spaces to travellers 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.

Airbnb was started in 2008 by Brian Chesky and Joe Gebbia, Airbnb is based on a peer-to-peer business model. This makes it simple, easy to use, and tends to be more profitable for both parties. The model also gives you the opportunity to customize and personalize your guests' experience the way you want.

The data we are going to analyse is the data of Airbnb NYC (2011-19). Our main objectives of analysis will be above four statements which can be briefed as learnings from hosts, areas, price, reviews, locations etc. but we will not be limited to here. We will also try to explore some more insights.

Problems we have answered:

1. Which hosts have the highest number of apartments?
2. Which are the top 10 neighbourhoods which are having the maximum number of apartments on Airbnb in the respective neighbourhood?
3. Which neighbourhood are having maximum prices in their respective neighbourhood_group?
4. How is the neighbourhood related to reviews?
5. What can we learn from predictions? (ex: locations, prices, reviews, etc.)
6. What is the distribution of the room type and its distribution over the location?
7. How is the room_type distributed over neighbourhood_group are the ratios of respective room_types more or less the same over each neighbourhood_group?
8. How is the price columns distributed over room_type and are there any Surprising items in the price column?
9. Which are the top 5 hosts that have obtained the highest no. of reviews?
10. What is the average preferred price by customers according to the neighbourhood_group for each category of room_type?
11. What is the average price preferred for getting a good number_of_reviews according to neighbourhood_group ?
12. Which hosts are busiest? (Most important)

Understanding the variables:

- **id**: Unique house/apartment ID
- **name**: Name of the house/apartment's owner
- **host_id**: Unique host ID provided by government
- **host_name**: Name of the host
- **neighbourhood_group**: Location
- **neighbourhood**: Area
- **latitude**: Latitude coordinates
- **longitude**: Longitude coordinates
- **room_type**: Listing space type
- **price**: price in dollars
- **minimum_nights**: Minimum amount of night for booking
- **number_of_reviews**: Number of reviews
- **last_review**: Latest review
- **reviews_per_month**: Number of reviews per month

The approach we have used in this project is defined in the given format:

1) Loading our data: In this section we just loaded our dataset in colab notebook and read the csv file.

2) Data Cleaning and Processing: In this section we have tried to remove the null values and for some of the columns we have replaced the null values with the appropriate values with reasonable assumptions.

3) Analysis and Visualization: In this section we have tried to explore all variables which can play an important role for the analysis. In the next parts we have tried to explore the effect of one over the other. In the next part we tried to answers our hypothetical questions.

4) Future scope of Further Analysis: There are many apartments having availability as 0 and date of last_review is very old, which can mean that they must have stopped their business, we can find the relation with neighbourhood with these apartments if we could dig much, various micro trends could be unearthed, which we are not able to cover during this short duration efficiently. There are various columns which can play an important role in further analysis such as number of reviews and reviews per month finding its relation with other factors or other grouped factors can play an important role.

Challenges faced:

● I faced challenges in in filling the price column with the median value of the same type of room_type and price trend across the neighbourhood.

●also faced problem while analysing the neighbourhood.

● And formatting formula for busiest host.

Conclusion:

Airbnb-NYC(2019) is a dataset of house booking having 48895 rows and 16 columns.

In the data Sonder(NYC) is having highest number of houses holding 327 listings.

In neighbourhood_group "Manhattan" having highest number of business or listing.

Williamsburg neighbourhood has the most number of listings among all neighborhoods .

While looking for costliest listing in NYC we got that 'Upper West Side, Astoria and Greenpoint' neighbourhood having the costliest listing.

Bedford-Stuyvesant neighbourhood is having highest number of total reviews and highest number of reviews_per_month also.

Manhattan and Brooklyn neighbourhood_groups are similar to each other in the count of listing and having highest number of listing. Staten Island and Bronx neighbourhood_group have very less numbers of listings according to Manhattan and Brooklyn.

Entire Home/Apartment or Private Rooms are most listed on Airbnb-NYC.

Guests prefer to stay in private rooms are stay for a shorter period of time according to the guests prefer to stay in the entire home/apartment.

Many price is having 0 in the price column, it seems like a glitch which must be rectified by Airbnb.

Maya (host) having heighest total number_of_reviews.

Average prices of all the room_types in Manhattan are more than the average price of each room_type in other all neighbourhood_group.

Average prices of all room_type in Bronx neighbourhood_group is less than all the other neighbourhood_groups.