**Capstone Project Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

|  |
| --- |
| **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:** This dataset has around 49,000 observations in it with 16 columns and it is a mix between categorical and numeric values.  Airbnb is an online marketplace connecting travelers with local hosts. On one side, the platform enables people to list their available space and earn extra income in the form of rent. On the other hand, Airbnb enables travelers to book unique homestays from local hosts, saving them money and giving them a chance to interact with locals. Catering to the on-demand travel industry, Airbnb is present in over 190 countries across the world.  The data we are going to analyze is the data of Airbnb NYC (2019). Our main objectives of analysis will be above four statements which can be briefed as learnings from hosts, areas, price, reviews, locations etc. but not limited to.we will also try to explore some more insights.  **Problem Statement:**   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 column 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 listing ID * **name**: Name of the listing * **host\_id**: Unique host ID * **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**: Amount of nights minimum * **number\_of\_reviews**: Number of reviews * **last\_review**: Latest review * **reviews\_per\_month**: Number of reviews per month   **Approach used:** the approach we have used in this project is can be defined in the given format-   1. **Loading our data :** In this section we just loaded our dataset in coloab 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. **Summarizing and conclusion:** In the end we ended up summarizing our analysis with the proper conclusion about the problem statement   **Challenges faced:**   * I faced challenges in selecting the features to interpret ,like price column may represent price for minimum\_nights or may represent per day price ,but after the deep literature survey I got to know that this is price per day * I also faced challenges in formatting metrics for busiest host   **Conclusion:**  This Airbnb-NYC(2019) dataset is a very informative dataset having 48895 rows and 16 columns. I found that SONDER(NYC) has the highest number of listings i.e 327 listings .I found that the highest number of listings in any neighbourhood group is Manhattan. The Williamsburg neighbourhood has the most number of listings among all neighborhoods .Upper West Side, Astoria and Greenpoint neighbourhoods have the costliest listing in NYC. Bedford-Stuyvesant neighbourhood has the highest number of total reviews and highest number of reviews\_per\_month. And Maximum listings are listed on Manhattan and Brooklyn neighbourhood\_groups. Staten Island and Bronx neighbourhood\_group have very less numbers of listings. Most of the listings on Airbnb in NYC are either Entire Home/Apartment or Private Room. The people who prefer to stay in the entire home/apartment are likely going to stay longer, whereas people who prefer to stay in private\_room are likely to stay for a shorter period of time than the people who prefer to stay in entire home/apartment. Many rows are having values as 0 in the price column, so this seems like an error which must be rectified by Airbnb. Keeping the high price of the listing and having 0 availability isn't benefiting the host as the consumer is ready to pay the price but even after that there are no available rooms then what's the benefit of paying such a premium. Maya (host) has the 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 neighbourhood\_group. Average prices of all the room\_type in Bronx neighbourhood\_group is less than all the other neighbourhood\_groups. |