**Capstone Project Submission**

**Instructions:**

i) Please fill 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)** |
| **Airbnb Data Analysis**:  **Business Problem :** Objective of this capstone is to do EDA on the given dataset and find out the insights from it.  In this dataset we are provided with 16 features and around 49k data instances.  Now the basic thing to start with is the exploring the features and finding out the meaningful insights from it.  I started with doing NA value handling, univariate analysis, multivariate analysis and conclusion.  **Step\_1:NA value replacement**  In this dataset we are having 4 columns with ‘NA’ values namely:   1. Name column: this describes the information regarding the property. 2. Host\_name: this describes the name of host or we can say individual persons name. 3. Last\_review: this shows us the last date of review. 4. Reviews\_per\_month: this shows the reviews got per month.   Name column we replaced it with corresponding value of room\_type column.  Host\_name column we dint do any thing.  Last\_Review we replaced ‘NA’ with 0 value.  Reviews\_per\_month we converted it into categorical data type and replaced ‘NA’ with ‘Never’ string type value.  **Step\_2:univariate analysis**  In this step we started with univariate analysis of individual feature.   1. For name column we generated word\_cloud.   Then we take a look into top 50 common words based on frequency of them this way we came to know about which words are useful of key words.   1. similarly we plotted the count plot for other features like neighbourhood   neighbourhood\_group and found which neighbourhood\_group and neighbourhood and most popular in terms of staying and which are least preferred.   1. We plotted scatter plot for lattitude and longitude feature this shows us the density of rooms/in each of the regions. 2. For room type feature we plotted the count plot and looked at each type.   As this room\_type is having categories we can do similar type of segmentation in price column and do our analysis, so we divided the price into range of price like from 0-80$ cheep category from 80-500$ affordable price range and above 500$ expensive.so affordable range was the most preferable category among the people followed by cheep and expensive price range.   1. From minimum\_neighs columns after analysis it we concluded that people try to spend 1-4 days. 2. Number\_of\_reviews tells us that average ratting is 23 times. 3. Calculated\_host\_listing\_counting tells how many time the host\_id is listed this shows the most famous host and least famous host. 4. Last review column tells us that 75% of times rating given is around 1.5-2. on ratting scale.   **Step\_3:Multivariate Analysis**   1. relation between neighbourhood\_group and median price.   From this we can say that Manhattan is having highest mean price and also high price property are also available with this region followed by Brooklyn and Queens.   1. Relationship between neighbourhood and median price.   From this we can say that Battery Park City id having highest median price.  11)Relationship between price and room\_type  From this we can say that if customer wants to book entire apartment then definitely they have to pay more.followed by private room and shared room.   1. Relationship between room\_type and neighbourhood\_group.   From this we can say that Manhattan is having highest booking of entire apartment followed by Brooklyn and Queens.  Similarly fro private room Brooklyn is having highest booking followed by Manhattan and Queens.  12Which neighbourhood are generating maximum,minimum,revenues from room types   * Entire\_home/apartment. * Williamsburg is having maximum revenue from Entire\_home/apartment. Which is around 389724$. * New\_Drop is the least or having minimum share of income form Entire\_home/apartment. * Maximum revenue from Private Room. * Williamsburg is having maximum revenue generation from Private Rooms around 171265$ * Graniteville is having minimum revenue generation from Private Rooms around 20$. * Maximum revenue from shared Room. * Hell’s Kitchen is having maximum revenue generation from Private Rooms around 9488$. * Randoll Manor is having minimum revenue generation from Private Rooms around 13$.   **Step\_4:**  **Conclusion** : In this simple yet power full way we had done the EDA on Airbnb dataset. certainly this is not the end rather this this the start we can say as per business requirement changes we need to find the insights in that direction and justify the business problems. There can be n-number of questions and n-number of dimension to explore the dataset and find the insight from them, this there is no limit unless the business constrain is solved.  Thank you:- |
| **https://drive.google.com/drive/folders/1TLOlqf34K6MHWqjXLJaI1cYIrrKIJdrH?usp=sharing** |
|  |