**Capstone Project Submission**

|  |
| --- |
| **Airbnb Data Analysis**:  **Business Problem :** The goal of this culmination is to do EDA on a given data file and learn from it.  We have 16 features and approximately 49,000 data instances in this dataset.  Now, the basic thing to start with is to explore the features and learn meaningful insights from them.  I started with processing NA values, univariate analysis, multivariate analysis and conclusion.  **Step 1:NA value replacement**  In this dataset we have 4 columns with "NA" values ​​namely:  1) Name column: describes information about the property.  2) Host\_name: describes the name of the host or we can say the name of individual persons.  3) Last\_review: this shows us the last review date.  4) Reviews\_per\_month: The reviews received per month are listed here.  We replaced the Name column with the corresponding value of the room\_type column.  We do nothing to the Host\_name column.  We replaced "NA" with 0 for Last\_Review.  We converted reviews\_per\_month to a categorical data type and replaced "NA" with a value of string type "Never".  **Step 2:univariate analysis**  In this step, we started with a one-dimensional analysis of individual elements.  1) We generated word\_cloud for the name column.  Then we look at the 50 most common words based on their frequency, and that's how we found out which words are useful keywords.  2) we similarly plotted the count for other features such as neighborhood  Neighborhood\_group and found out which neighborhood\_group and neighborhood and which are the most popular in terms of residence and which are the least preferred.  3) We plotted a scatter plot for the latitude and longitude properties, showing us the density of rooms/in each of the areas.  4) For the room type function, we plotted the count and looked at each type.  Since this room type has categories, we can do a similar type of segmentation in the price column and do our analysis, so we divided the price into a price range, such as from $0-80 cheap category from $80-500 available price range and over $500 Expensive. so affordable range was the most preferred category among people followed by cheap and expensive price range.  5) From the minimum columns after analysis, we concluded that people try to spend 1-4 days.  6) The number of reviews tells us that the average rating is 23 times.  7) Calculated count of host lists tells how many times a host ID is listed, shows the most famous and least known host.  8) The last review column tells us that 75% of the ratings given are around 1.5-2. on a rating scale.  **Step 3:Multivariate Analysis**  1) the relationship between neighborhood group and median price.  From this we can say that Manhattan has the highest average price and also high real estate prices are also available with this region followed by Brooklyn and Queens.  2) Relationship between neighborhood and median price.  From this we can say that ID Battery Park City has the highest median price.  11) Relationship between price and room type  From this we can say that if the customer wants to book the whole apartment, then he definitely has to pay more. A private room and a shared room follow.  3) Relationship between room type and neighborhood group.  From this we can say that Manhattan has the highest reservation of the entire apartment, followed by Brooklyn and Queens.  Similarly, from a private room, Brooklyn has the highest booking followed by Manhattan and Queens.  12) Which neighborhood generates the maximum, minimum, revenue from room types   * Entire\_home/apartment. * Williamsburg is having maximum revenue from Entire\_home/apartment. Which is around 389724$. * 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** : We performed EDA on the Airbnb dataset in this simple but powerful way. it is definitely not the end, rather this is the beginning, we can say that according to the changes in business requirements, we need to find knowledge in this direction and justify business problems. There can be n-number of questions and n-number of dimensions to explore the dataset and find insight from them, this is not a limitation unless the business constraint is solved.  Thank you  ***GITHUB link:***  https://github.com/Ganeshb979/Airbnb-Bookings-Analysis |