**Airbnb Dataset Exploratory Data Analysis**

**Bishakha Kumari, Aashish, Mrityunjay Dubey**

**Data Science Trainees @Almabetter**

| 1. **About Airbnb:**   Airbnb is an online marketplace connecting travellers 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, Airbnb enables travellers 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.   1. **How it works?**  * Hosts list out their property details on Airbnb along with other factors like pricing, amenities provided etc. * Airbnb sends a professional photographer (if available) to the property location in order to take high quality photographs. * Travellers search for a property in the city where they wish to stay and browse available options according to price, amenities etc. * Booking is made through Airbnb where traveller pays the amount mentioned by host and some additional money as transaction charges. * Host approves the booking. Traveller stays there and finally Airbnb pays the amount to the host after deducting their commission | 1. **Overview of Our Airbnb Dataset**   Our Airbnb dataset consists mainly the NYC’s five boroughs data which are Manhattan, Brooklyn, Queens, Bronx and Staten Island. The shape of our Airbnb dataset has initially 48895 records and 16 features in it.  The dataset is a mixture of Categorical and Numerical features. Each record in a dataset signifies a particular Airbnb listing and the features associated to it like its id, host, price, reviews etc. By looking at the head of a dataset we can get an overall idea of its columns and their description.   1. **All the columns that are present initially in our dataset with their short descriptions are listed below:**  * id – The id of the listing * name – The name of the listing * host\_id – This is to signify the id of host who are associated to the respective listing * host\_name – The name of the host * neighbourhood\_group – This column is signifying the five boroughs of NYC * neighbourhood - the neighbourhood of the respective neighbourhood groups. * latitude and longitude - the geographical co-ordinates of respective listings. * room\_types - listing space type * price - price of listing in dollars * minimum\_nights - This is the amount of nights minimum |
| --- | --- |
| **Description of columns continued**   * availability\_365 – The number of days when listing is available for booking within a year. * calculated\_host\_listings\_count – The amount of listing per host * number\_of\_reviews - Total number of reviews. * reviews\_per\_month - Average of reviews per month. * last\_review - Date of the last review recorded   **5. Steps Involved in our exploratory data analysis of Airbnb NYC dataset**  **5.1 Getting the overview of our dataset**  The first and foremost thing for the analysis of any data is to get the idea and overview of how our data looks like, what are its features, what is its shape, determining the data types of the columns and changing the data types as and when required. Initially we have set the name of our dataset to ‘airbnb\_df’.  Shape - the shape of our dataset was - (48895, 16)  Structure - To get the basic structure of our dataset we have simply called head function to print the first five rows of our dataset like ‘airbnb\_df.head()’. By doing this we will be able to visualize the structure of the dataset and many a times it gives a lot information about the presence of null values, categorical columns, numerical columns etc.  Information of columns, their data types and presence of null values – We will get the information about all the features running the code ‘airbnb\_df.info()’. It will give summarize view of the columns and their data types. By doing this we can verify the | data types and presence of any null count in features.    **5.2 Feature Engineering**  Next, we have selected the columns that we are going to use in our analysis and have removed the rest of the unnecessary columns. We have removed ‘id’ and ‘last\_review’ columns.  After eliminating the columns, we have checked the shape of the dataset to ensure the columns are eliminated and dataset is updated.  Updated shape of the airbnb\_df is (48895, 14)  **5.3 Handling the null values**  From the info() method we have noted there are some null values present in different features of dataset, so first we have got the total number of null counts in the respective columns using: **‘airbnb\_df.isnull().sum()’,** below is the output that we have got: |

| As we can see from the output that in some of the columns there are very low number of null count and some have very high number of null counts, so we have followed different procedures to handle these two scenarios to protect significant data loss.  Firstly, we have handled the columns that have very high null count by replacing the null values with the most appropriate option. Generally, the most appropriate option is mean, median or mode depending on the respective type of columns i.e., for categorical columns the most appropriate value to replace null values is mode and for Numerical ones it is median or mean.  In our case, the numerical column ‘reviews\_per\_month’ has very high null count so we have to replace it with median of the column as the data points are not normally distributed so median is the right choice instead of mean. | Then the all the records that still have null count but in a very less number as compared to the total records in dataset, are removed using:  ‘**airbnb\_df.dropna(subset = ['name', 'host\_name'], inplace = True)**’  **5.4 Analysis of the dataset**  Once the dataset set is prepared according to our need, we will start to dug out the information from it. We have used following techniques to draw information about our dataset:   * Correlation Heatmap * Univariate Analysis * Bivariate Analysis   **Correlation Heatmap -** A correlation heatmap is a heatmap that shows a 2D correlation matrix between two discrete dimensions, using coloured cells to represent data from usually a monochromatic scale The colour of the cell is proportional to the number of measurements that match the dimensional value. This makes correlation heatmaps ideal for data analysis since it makes patterns easily readable and highlights the differences and variation in the same data. |
| --- | --- |

| **Univariate Analysis -** Univariate analysis is a basic kind of analysis technique for statistical data. Here the data contains just one variable and does not have to deal with the relationship of a cause and effect. It is basically the simplest form to analyse data. Uni means one and this means that the data has only one kind of variable. The major reason for univariate analysis is to use the data to describe. The analysis will take data, summarise it, and then find some pattern in the data.  Following are the ways though which we have conducted this analysis:   * Frequency distribution table * Bar Chart * Pie Chart.   Inferences that we got from the univariate analysis are as follows:  >The host that has maximum number of Airbnb listings.  >The location which has maximum number of listings  >The room type that is most often offered by listings.  **Bivariate Analysis -** Bivariate analysis is stated to be an analysis of any concurrent relation between two variables or attributes. This study explores the relationship of two variables as well as the depth of this relationship to figure out if there are any discrepancies between two variables and any causes of this difference. Some of the examples are percentage table, scatter plot, etc.  Following are the ways though which we have conducted this analysis:   * Bar Chart * Scatter Plot. * Box Plot | Inference drawn from the bivariate analysis are as follows:  >the variation of price with location, room type and number of reviews  > the variation of minimum nights with room type and number of reviews  >the variation of number of reviews with location.  **6. Inferences drawn from analysis**  > Manhattan is the costliest neighbourhood group and most of the hosts in here has maximum number of listings associated to them as compared to that of any other neighbourhood group.  >The Entire home/apt listing type is costliest and maximum among all other listing types available.  > Guests are preferring private rooms over Entire home/apt for shorter duration.  >Private rooms are getting a greater number of reviews which makes sense as it is less costly so a greater number of people can afford it. Also, the minimum nights dependency is less.  >The location with a smaller number of listings is getting more reviews compared to that of the locations with higher listings.  **7. Graphs and Plots used for visualization**  \*Scatter plot – used to show the relation between features like location and price, location and reviews, rooms and price etc.  \*Bar chart – used to show relation and counts of a feature as per the other feature.  \*Count plot – It is similar as bar chart. We used it to count number of appearances of a particular value of a feature.  \*Pie chart – used to show the distribution percentage of data points of a categorical column.  \*Box Plot – used to identify outliers in a particular feature and get the 5-point summary of the selected feature which is min, max, median, 1st and 3rd quartile. |
| --- | --- |

| **8. Conclusion**  We have finally reached the end of our project. Starting from loading the raw data to get it cleaned and free from any non-uniformity and null values and getting the desired data for analysis was a challenge. In this process we have eliminated the features that are not required, removed outliers to get an unbiased result, added additional features to the dataset as and when required.  We have also used different visualization techniques, graphs and plots to get a clearer picture of what is preferred most by the guest and what they are looking for in Airbnb.  We got to know that people are preferring the locations that are the most popular and have numerous attractions the most.  The Entire home/apt are the costliest and maximum in Manhattan and preferred for longer duration bookings.  **References –**  Almabetter course  Geeks for geeks  Wikipedia  Tutorials point |  |
| --- | --- |