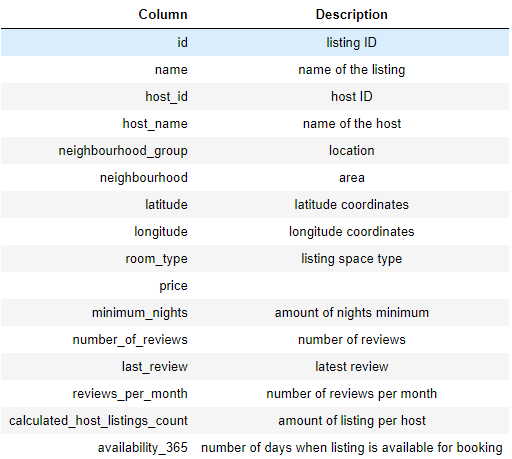
**METHODOLOGY DOCUMENT**

**Generating Insights & Recommendations for Growth of Airbnb NYC Post COVID**

**Understanding the business problem:**

* Problem Statement: Airbnb has seen a major decline in revenue due to COVID-19. Now that the restrictions have started lifting and people have started to travel more, Airbnb wants to ensure that it is fully prepared for this change.
* Read and understood the above problem statement.

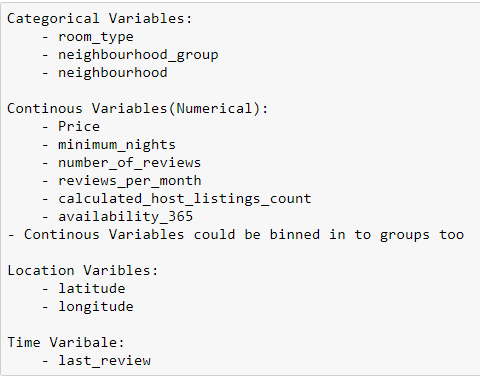
**Understanding the dataset:**

* Understood the dataset containing various Airbnb listings in New York.
* The dataset had various attributes as given below: 
* Had to use this dataset to solve the business problem.

**Data Wrangling:**

* Used Python programming language to do this step.
* First import necessary libraries such as seaborn, NumPy, Matplotlib and Panadas in Python.
* Loaded the dataset in the pandas DataFrame.
* Inspected the DataFrame by checking its shape, datatypes of each column, and information about non-null values.
* There were 48895 rows and 16 columns.
* Checked the number of null values in each column of the dataset. The columns namely name, host\_name, last\_review and review\_per\_month column contained the null values.
* Checked for duplicate rows but no duplicate rows were found.
* Dropped the column id because it was not relevant to our analysis.
* The values missing in last\_review and reviews\_per\_month were not missing at random. Customers have not provided reviews to these hosts hence NaN values. If the customers had liked they would have rated them but since they did not do it, simply means they may have not liked the host place. So, we will replace the missing values for reviews\_per\_month with 0 and keep the others as is.
* Inspected and looked at outliers in data by using boxplots. Columns namely price, minimum\_nights, reviews\_per\_month, calculated\_host\_listings\_count had many outliers.
* The outliers were not treated and kept as it is.
* For further analysis through data visualization the cleaned data was downloaded in a .csv file

**Data Analysis and Visualization:**

* The downloaded CSV file from Python was imported into Power BI.
* Only the histogram of prices and the histogram of minimum\_nights were created in Python itself. Otherwise, the majority of all visualizations were created in Power BI.
* Based on the columns, they were classified by type as given below: 
* Created the appropriate visualizations based on the type of column.

**Insights and Recommendations:**

* Based on the data analysis and visualizations, insights were generated and recommendations were presented.
* These insights and recommendations were presented in PowerPoint presentations. Some of the insights are given below:
  + Preference for Entire Room/Apt and Private Room is highest.
  + Highest properties are listed in Manhattan followed by Brooklyn.
  + Majority of Airbnb listings were priced at the lower end of the spectrum.
  + Manhattan had the highest average prices and Bronx had the lowest average prices.

**Storyboarding:**

* Before creating the presentations, storyboard my thoughts on what to include and what not to include in the presentation.
* This meant creating a visual outline of the various slides of the presentation.

**Final Presentations:**

* Finally, created the presentation using various data storytelling principles such as:
  + Converting data into business terminology.
  + Used variation (in colour, shape, size, style) only to highlight importance and relevance.
  + Only those pieces of information is presented while keeping the appropriate audience in mind.