**Manan Aggarwal(23112053)**

**EDA on AIRBNB Data from 2019**

**Introduction**

This report details the exploratory data analysis (EDA) conducted on the Airbnb datase. The purpose of this analysis is to gain insights into the data, identify missing values, visualize distributions, and clean the dataset for further analysis.

**1. Importing Libraries**

The following libraries were used:

* pandas and numpy for data manipulation.
* matplotlib.pyplot, seaborn, and plotly.express for visualizations.

**2. Loading and Exploring Data**

* The dataset was loaded using pd.read\_csv().
* The first few rows were displayed using head() to understand its structure.
* The shape of the dataset was checked using shape, revealing the number of rows and columns.
* Categorical and numerical features were identified using select\_dtypes().

**3. Handling Missing Data**

* Missing values were counted using isnull().sum().
* A bar plot was created to visualize the percentage of missing values for each feature.
* A heatmap was plotted using sns.heatmap(train\_df.isnull()) to show the distribution of missing data.
* Strategies to handle missing values:
  + Missing name and host\_name were filled with "not known".
  + Missing reviews\_per\_month was replaced with 0.
  + Columns id and last\_review were dropped due to irrelevance.

**4. Statistical Summary**

* The describe() function was used to display summary statistics, such as mean, median, and standard deviation of numerical features.

**5. Univariate and Bivariate Analysis**

**Univariate Analysis:**

* A histogram with KDE was plotted using sns.histplot() to analyze the distribution of Airbnb prices.
* Other individual feature distributions were also explored.
* This helped in identifying common value ranges, outliers, and skewness in the data.

**Bivariate Analysis:**

* Correlation between numerical features was analyzed using sns.heatmap(train\_df.corr()).
* Relationships between categorical and numerical features, such as room type vs. price, were visualized using boxplots and scatterplots.
* This helped in understanding how different factors impact pricing and availability.

**6. Data Insights**

* The price distribution showed a right-skewed pattern, indicating that most Airbnb listings are relatively affordable, but there are a few high-priced outliers.
* Certain room types (e.g., entire homes/apartments) tend to have significantly higher prices than shared rooms.
* The number of reviews per listing is generally low, with many properties having zero reviews, indicating that many listings might be new or inactive.
* Missing values in reviews\_per\_month were mostly in listings with zero reviews, justifying filling them with 0.

**Conclusion**

The EDA helped in understanding the structure of the dataset, handling missing values, and visualizing important trends. This process is essential for preparing the data for further predictive modeling or business insights.

**Reasons for Analysis:**

1. **Understanding Data Quality** – Identifying and handling missing values ensures accurate analysis.
2. **Feature Engineering** – Understanding data distributions helps in deciding transformations or feature selection.
3. **Market Insights** – The price distribution analysis provides useful information for Airbnb hosts and market trends.
4. **Identifying Relationships** – Bivariate analysis helps uncover trends between features, aiding in business decision-making.
5. **Detecting Outliers** – Univariate analysis helps in spotting anomalies that could impact modeling results.