**Data Analysis Report on NYC Airbnb Dataset**

**1. Introduction**

This report aims to conduct an exploratory data analysis (EDA) on the NYC Airbnb dataset. The objective is to clean the data, understand pricing patterns, booking behavior, and key influencing factors.

**2. Data Cleaning**

**Steps Taken:**

1. **Handling Missing Values:**
   * name and host\_name filled with "Unknown".
   * reviews\_per\_month set to 0 for missing values.
   * last\_review converted to datetime, with missing values as "No Review".
2. **Outlier Removal:**
   * Used Interquartile Range (IQR) method to filter extreme values in price and minimum\_nights.
3. **Duplicate Removal:** Removed duplicate rows for data integrity.
4. **Standardization of Categorical Data:** Corrected inconsistencies in neighbourhood\_group and room\_type.

**3. Exploratory Data Analysis (EDA)**

**3.1 Univariate Analysis**

* **Price Distribution:** Right skewed with most listings below $500, but extreme outliers above $9000.
* **Minimum Nights:** Majority between 1-10 nights; outliers exceed 1000 nights.
* **Neighbourhood Group Frequency:** Manhattan and Brooklyn dominate the market.
* **Room Type:** Entire home/apartment is the most frequent listing.

**Visualizations:**

* Histograms and bar charts for numerical and categorical features.

**3.2 Bivariate Analysis**

* **Price vs Minimum Nights:** High prices associated with shorter stays, but some extreme cases with long minimum nights exist.
* **Price vs Reviews:** Lower-priced listings attract more reviews, indicating higher customer engagement.
* **Price by Room Type:** Entire homes have the highest median price, while shared rooms are the cheapest.
* **Price vs Availability (365 days):** Listings with higher availability tend to have moderate prices, while year-round high-priced listings are rare.

**Visualizations:**

* Scatter plots, box plots, and violin plots.

**3.3 Multivariate Analysis**

* **Correlation Heatmap:**
  + Strong Positive Correlation: number\_of\_reviews and reviews\_per\_month (0.58), indicating that frequently reviewed listings get consistent reviews.
  + Moderate Negative Correlation: price and longitude (-0.29), suggesting higher prices in certain areas.
  + Weak Correlations: Minimal linear relationships between price and availability.
* **Pair Plot:** Visualizes relationships between numerical features and highlights potential clusters and outliers.

**4. Conclusion**

1. Most listings are affordable, but luxury properties skew the average price.
2. Lower-priced listings get more reviews, indicating higher demand.
3. Manhattan dominates high pricing, while Bronx is the cheapest.
4. Minimal correlation between price and numerical features suggests location and room type drive pricing.