📊 Airbnb NYC 2019 EDA – Project Documentation

# 🔍 Project Overview

This project involves an in-depth Exploratory Data Analysis (EDA) of Airbnb listings in New York City for the year 2019. Using data science techniques, we cleaned, analyzed, and visualized the dataset to uncover meaningful insights regarding pricing, popularity of room types, location trends, and host behaviors. The analysis will help both hosts and customers make informed decisions and can serve as a base for future predictive modeling.

# 📁 Dataset Information

• Source: Inside Airbnb (http://insideairbnb.com)

• Format: CSV

• Year: 2019

• Total Records: Approx. 49,000

• Columns and Descriptions:

- id: Unique listing ID  
- name: Title of the listing  
- host\_id: Unique host ID  
- neighbourhood\_group: Borough (Manhattan, Brooklyn, etc.)  
- neighbourhood: Specific area within borough  
- latitude & longitude: Location coordinates  
- room\_type: Type of place (Entire home, Private room, etc.)  
- price: Per night cost in USD  
- minimum\_nights: Minimum stay length  
- number\_of\_reviews: Number of reviews received  
- last\_review: Date of last review  
- reviews\_per\_month: Avg. number of reviews per month  
- calculated\_host\_listings\_count: Total listings by host  
- availability\_365: Number of available days in a year

# 🧹 Data Cleaning & Preprocessing

• Removed rows with null 'name' values.  
• Filled missing 'reviews\_per\_month' values with 0.  
• Replaced nulls in 'calculated\_host\_listings\_count' with median value.  
• Converted 'last\_review' column to datetime format.  
• Checked and corrected data types for numerical and categorical fields.  
• Removed extreme outliers in 'price' and 'minimum\_nights' to improve plot clarity.

# 📊 Analysis Performed

• Missing value analysis and statistical summary  
• Count of room types and neighborhood distribution  
• Price distribution per borough and room type  
• Review trends over time  
• Host listing behavior patterns  
• Availability across NYC regions

# 📈 Visualizations Created

• Room Type Distribution – Bar Chart & Pie Chart  
• Price Distribution – Histogram and Boxplot  
• Availability vs Price – Scatter Plot  
• Review Activity – Line Plot  
• Neighborhood Popularity – Count Plot & Heatmap  
• Host Listing Count – Histogram  
• Correlation between numerical variables – Heatmap

# 🛠️ Tools & Technologies Used

• Python (Pandas, NumPy, Matplotlib, Seaborn)  
• Jupyter Notebook / Google Colab for execution  
• GitHub for version control

# ✅ Key Insights

• Manhattan is the most expensive borough for Airbnb stays.  
• Private rooms tend to get more reviews, indicating higher booking frequency.  
• Some hosts own a very high number of listings.  
• Listings with 0 reviews are mostly new or inactive.  
• Listings under $100/night are most common.  
• High review rates are seen in Brooklyn and Manhattan neighborhoods.

# 🚀 Future Scope

• Incorporate predictive models for price or booking success.  
• Perform NLP-based sentiment analysis on customer reviews.  
• Add interactive dashboards using Plotly or Tableau.  
• Merge this dataset with COVID/post-2020 data to assess long-term impact.

# 📌 Conclusion

This project provides a comprehensive look at Airbnb’s presence in NYC in 2019. From neighborhood analysis to pricing patterns,   
the insights gained are valuable for stakeholders including hosts, travelers, and Airbnb itself. The structured approach to cleaning,   
visualization, and analysis gives a strong foundation for deeper studies and practical applications like price prediction or listing optimization.