Airbnb NYC Dataset Analysis Report

# 1. Dataset Overview

The dataset used for this project is the New York City Airbnb Open Data. It contains 48,631 rows and 16 columns, capturing detailed information about Airbnb listings, including listing ID, host details, neighbourhood, location, room type, pricing, number of reviews, availability, and host activity. The dataset is widely used for exploring the dynamics of Airbnb’s short-term rental market in NYC.

# 2. Purpose of the Analysis

The primary purpose of analyzing this dataset is to understand the factors that influence Airbnb listings in New York City. Specifically, we aim to:  
- Explore the distribution of listings across neighbourhood groups.  
- Analyze room types and their frequency in different neighbourhoods.  
- Examine price distributions and identify outliers.  
- Study availability patterns of listings throughout the year.  
- Identify top hosts and most-reviewed listings.  
- Build predictive models (Linear Regression and Random Forest) to predict prices and extract feature importance.

# 3. Methodology

## a. Data Cleaning

The dataset contained missing values in the 'last\_review' column. Since it does not directly impact pricing or availability analysis, these missing values were not imputed but handled carefully during visualizations and modeling.  
  
Outliers: Price distribution revealed extreme outliers (e.g., listings priced at thousands of dollars per night). Such extreme values can skew model performance and distort visualizations. For analysis, we focused on typical ranges (capping extreme outliers) while still acknowledging their presence in descriptive statistics.

## b. Exploratory Data Analysis (EDA)

We visualized and analyzed:  
- Distribution of listings by neighbourhood group: Manhattan and Brooklyn dominate the listings.  
- Room type distribution: Entire home/apt and private rooms are the most common.  
- Price distributions: Manhattan tends to have higher median prices compared to other boroughs.  
- Availability: Many listings are either rarely available (0 days) or available year-round (365 days).  
- Top hosts: A few corporate-style hosts (e.g., Sonder, Blueground) have hundreds of listings.  
- Most-reviewed listings: Some budget rooms near airports (Queens) and central Manhattan are highly popular.

## c. Predictive Modeling

Two models were tested to predict listing prices:  
- Linear Regression: Achieved RMSE ≈ 56.39 and R² ≈ 0.32.  
- Random Forest Regressor: Improved results with RMSE ≈ 54.98 and R² ≈ 0.33.  
  
Random Forest performed slightly better, capturing nonlinear relationships in the data.

## d. Feature Importance

From the Random Forest model, the most important features influencing price were:  
- Availability (number of days available in a year).  
- Room type (Private room, Entire home/apt, Shared room).  
- Reviews per month and number of reviews.  
- Minimum nights and host listing count.  
- Neighbourhood group (especially Manhattan and Brooklyn).

# 4. Findings

- Manhattan and Brooklyn dominate Airbnb activity in NYC.  
- Entire homes/apartments and private rooms form the majority of listings.  
- Prices are highest in Manhattan, moderate in Brooklyn, and lowest in Bronx/Queens.  
- A large share of listings are either never available or available the whole year.  
- Corporate hosts manage hundreds of listings, influencing availability and price trends.  
- Predictive models show that availability, room type, and reviews are the strongest predictors of price.

# 5. Conclusion

The analysis highlights the dynamics of Airbnb in New York City, showing the dominance of certain boroughs and room types, clear differences in pricing patterns, and the influence of host strategies. Outlier cleaning was necessary for meaningful visualization and modeling. While linear models capture some price variation, ensemble methods like Random Forest perform better, revealing key factors such as availability and room type as crucial in determining listing prices.