# Airbnb Data Analysis Project Report

**Kartik Agarwal**

November 6, 2025

—

Data Analysis Project

—

Airbnb

# 1. Introduction

The purpose of this project is to analyze Airbnb open data to uncover key insights into pricing patterns, room types, and neighborhood trends. The analysis focuses on understanding how various factors—such as location, room type, and review metrics—affect listing prices and availability. By leveraging data analysis and visualization techniques, this project aims to extract meaningful insights for hosts, travelers, and market researchers.

---

# 2. Objectives

The main objectives of this analysis are:

- To clean and preprocess the Airbnb dataset for accurate analysis.
- To explore patterns in **listing prices** and **room types**.
- To analyze **geographic distribution** of listings across neighborhoods.
- To visualize relationships between **price**, **room type**, and **location**.
- To identify **key insights** that can guide both hosts and guests.

---

# 3. Dataset Description

The dataset used in this project is the **Airbnb Open Data** file (Airbnb_Open_Data.csv).

### Dataset Details:

- Total Listings (Before Cleaning): **102,599**
- Total Columns: **26**

### Main Columns:

| Category | Features |
|---|---|
| **Listing Info** | ID, Name, Host Name |
| **Location Data** | Neighborhood, Latitude, Longitude, Country |
| **Pricing Info** | Price, Service Fee |
| **Room Details** | Room Type, Minimum Nights, Construction Year |
| **Review Metrics** | Number of Reviews, Review Rate, Last Review Date |
| **Availability** | Availability 365, Number of Bookings |

---

# 4. Tools and Technologies

The project was developed in **Python** using the following libraries:

- **Pandas** → Data cleaning and manipulation
- **NumPy** → Numerical computation
- **Matplotlib** → Data visualization
- **Seaborn** → Statistical visualization
- **Jupyter Notebook** → Interactive analysis environment

---

# 5. Methodology

## Step 1: Data Loading

- Loaded the dataset into a Pandas DataFrame.
- Handled data type warnings for mixed-type columns.

## Step 2: Data Cleaning

- **Handled Missing Values:** Replaced or removed NaN values in columns.
- **Converted Dates:** Transformed 'last review' to datetime format.
- **Cleaned Price Columns:** Removed symbols (like $) and converted to numeric format.
- **Dropped Irrelevant Columns:** Such as license and house_rules due to insufficient data.
- **Removed Duplicates:** Ensured unique entries.
- **Final Dataset Size: ~101,410 listings** after cleaning.

## Step 3: Exploratory Data Analysis (EDA)

The EDA process involved understanding and visualizing the key trends:

**A. Price Distribution**

- Histogram and KDE plots were used to observe the spread of listing prices.
- Found a wide range of accommodation prices, indicating market diversity.

**B. Room Type Analysis**

- Count plots revealed four major room types:
  - Entire home/apt
  - Private room
  - Shared room
  - Hotel room
- Entire homes and private rooms dominate listings.

**C. Neighborhood Distribution**

- Listings were analyzed across neighborhood groups:
  - Manhattan
  - Brooklyn
  - Queens
  - Bronx
  - Staten Island
- Manhattan and Brooklyn have the highest number of listings.

**D. Price vs. Room Type**

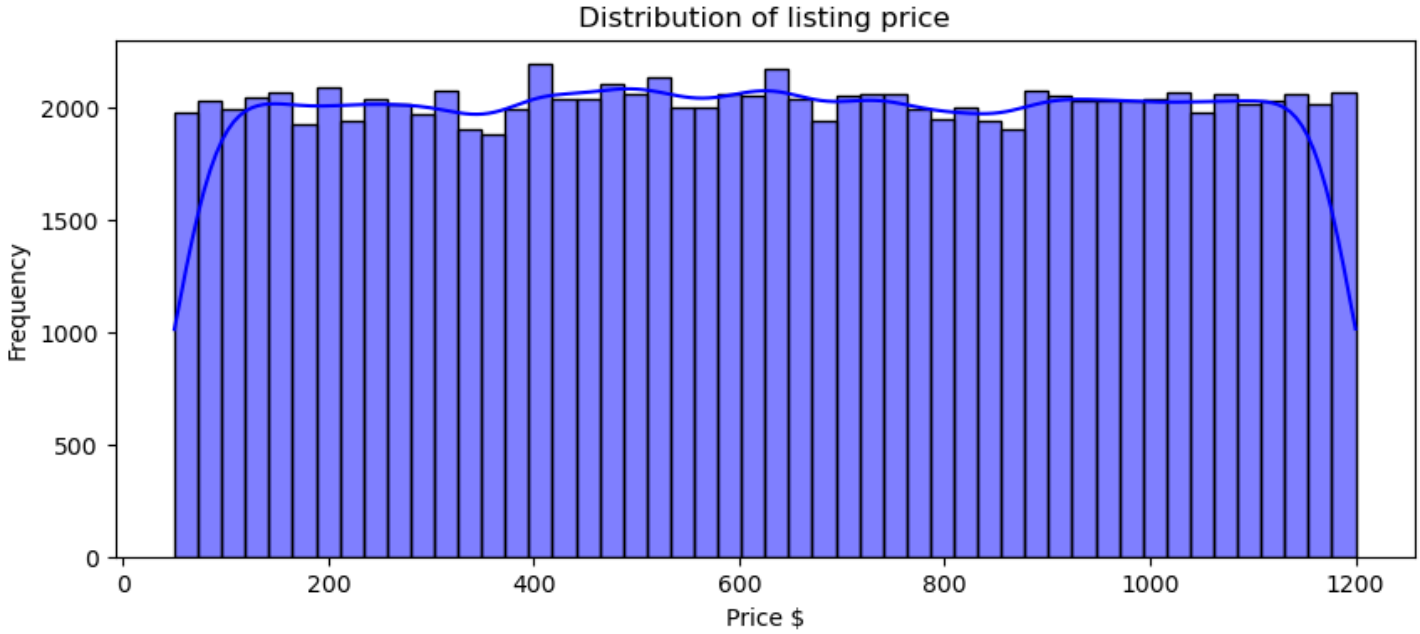
- Box plots were used to compare prices across room types.
- Shared rooms have lower prices, while private and entire homes show higher variability.

---

# 6. Key Findings

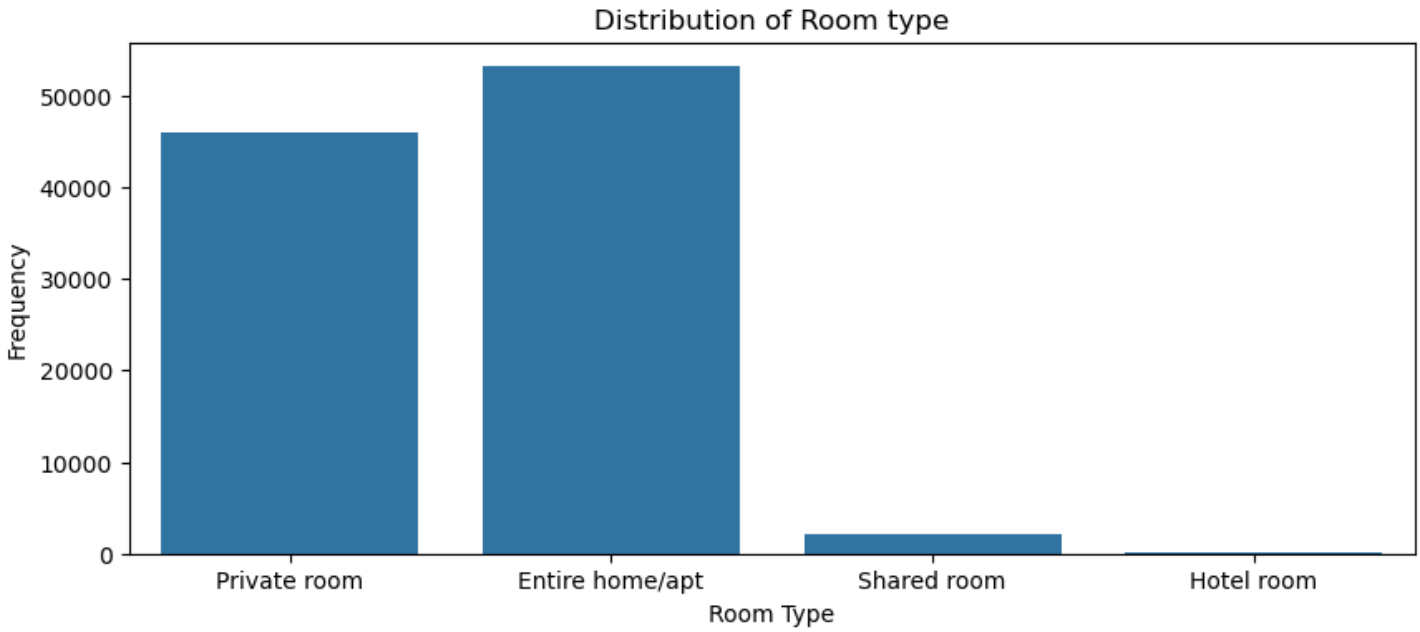
## Price Distribution:

- Prices vary widely, with no extreme skewness, reflecting a balanced market.



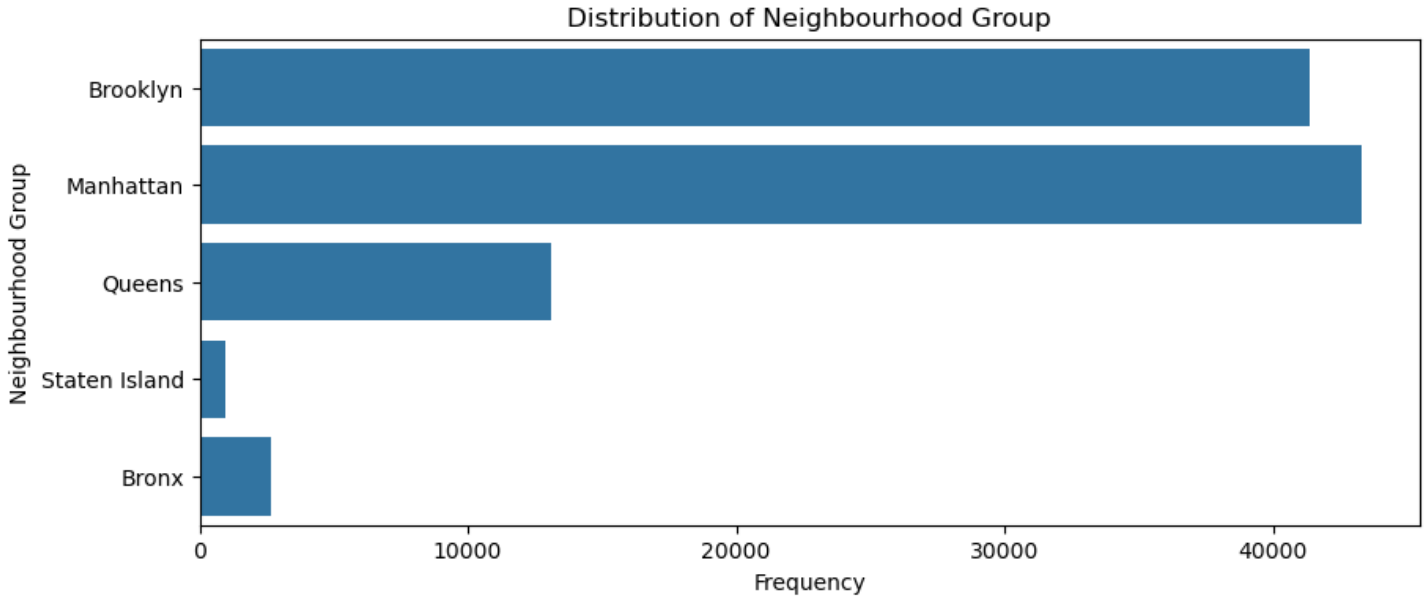

## Room Type Popularity:

- **Entire home/apt** and **Private room** dominate, together forming the majority of listings.
- **Shared** and **Hotel rooms** form a smaller segment.



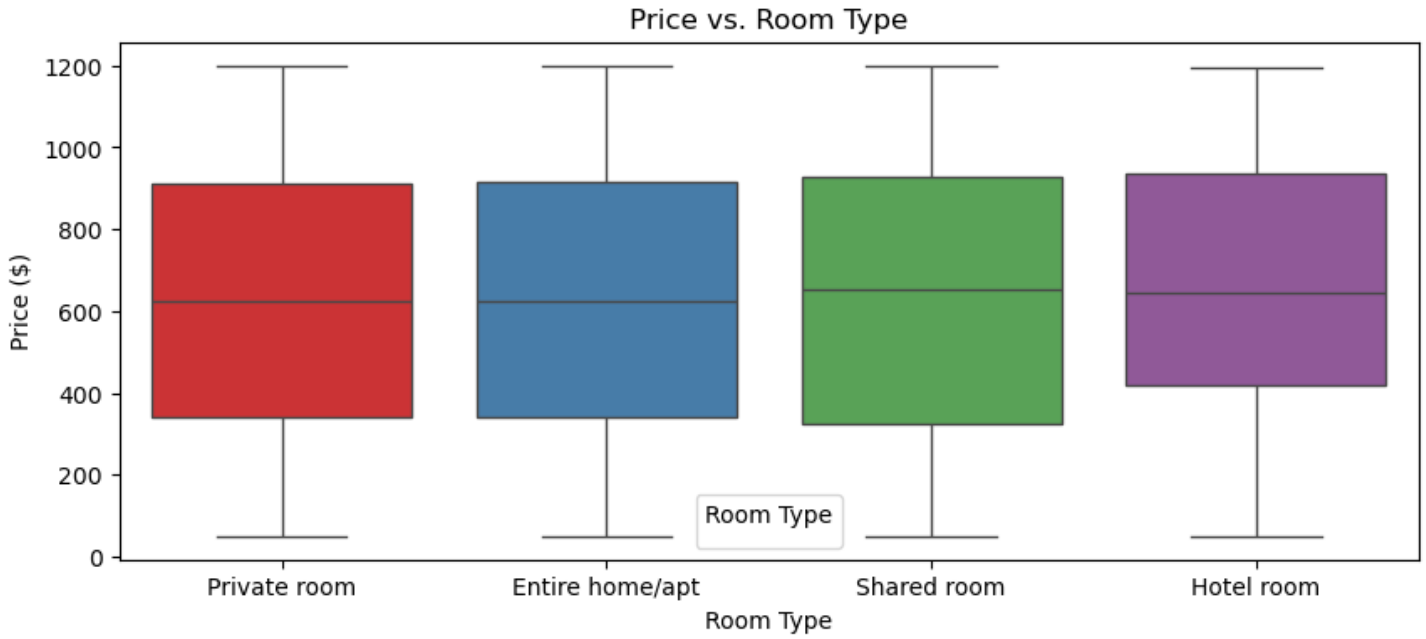

## Neighborhood Trends:

- **Manhattan** and **Brooklyn** host the largest number of listings.
- **Queens**, **Bronx**, and **Staten Island** have relatively fewer.




### Pricing Patterns:

- **Shared rooms** are the most affordable.
- **Private and Entire homes** show greater price variation, often linked to location and amenities.




---

# 7. Visualizations

The following visualizations were created in Data_analysis.ipynb:

- Price distribution histogram and KDE plot.
- Count plots for room type distribution.
- Bar chart of listings per neighborhood group.
- Box plot of prices by room type.

Each visualization contributed to identifying clear market trends and price clusters.

---

# 8. Results and Insights

- The Airbnb market is **diverse and location-sensitive**.
- **Room type** significantly influences pricing.
- **Neighborhood** plays a major role in determining the number of listings and pricing tiers.
- Data cleaning was essential to ensure accuracy, as the raw dataset contained duplicates and inconsistent formats.

---

# 9. Limitations

- Geographic coordinates were available but not visualized on maps.
- Some data columns had missing or inconsistent values.
- Predictive modeling and correlation analysis were not implemented in this version.

---

# 10. Future Enhancements

- Incorporate **geospatial visualization** using folium or plotly.
- Perform **time series analysis** on reviews and availability.
- Build a **price prediction model** using machine learning.
- Conduct **correlation and regression analysis** for deeper insights.
- Explore **host performance metrics** and their impact on reviews.

---

# 11. Conclusion

This project successfully analyzed Airbnb data to uncover insights into listing trends, room type preferences, and neighborhood-level pricing. Through systematic data cleaning, visualization, and exploratory analysis, it highlights the diversity and complexity of Airbnb's marketplace. The findings can serve as a foundation for future work in predictive modeling and market optimization.

---

# 12. References


- Airbnb Open Data (Public Dataset)
- Python Documentation
- Pandas, NumPy, Matplotlib, Seaborn Official Docs