**Airbnb 2019 Data Analysis Project Report**

**A comprehensive exploratory data analysis project deployed using AWS**

Nepal, Peshal., Andres, Duarte, Arvindhai, Milan, Patrick Odita & Tolulope Victoria Hassan.

Big Data Analytics, Georgian College

24F Data Analytics Programming - 02

Professor Roshan Sahu

December 4, 2024

### **Summary of the Analysis**

### This project highlighted the effective use of data analysis techniques to acquire, process, and analyze a meaningful dataset, generating actionable insights aligned with client interests. Using the Airbnb NYC 2019 dataset, which provides detailed information on room listings such as pricing, location, and reviews, we explored key trends and patterns to guide strategic decisions.

### **Data Acquisition and Storage**

### The dataset was fetched from a remote source using Python and stored in an Amazon RDS MySQL database to ensure accessibility and scalability. Custom data processing scripts were used to clean the data by addressing Nan values, enforcing strict data type constraints, and standardizing formats, significantly enhancing data reliability for subsequent analyses. Dataset Link: <https://www.kaggle.com/api/v1/datasets/download/dgomonov/new-york-city-airbnb-open-data>

### **Dataset**

### The dataset holds 48,895 Airbnb listings in New York City with 16 columns, including listing ID, host information, neighborhood, room type, price, minimum nights, reviews, and availability. While most fields are complete, some columns, like reviews per month and last review, have missing values. It offers insights into price trends, host activity, and neighborhood preferences in NYC.

### **Methodology**

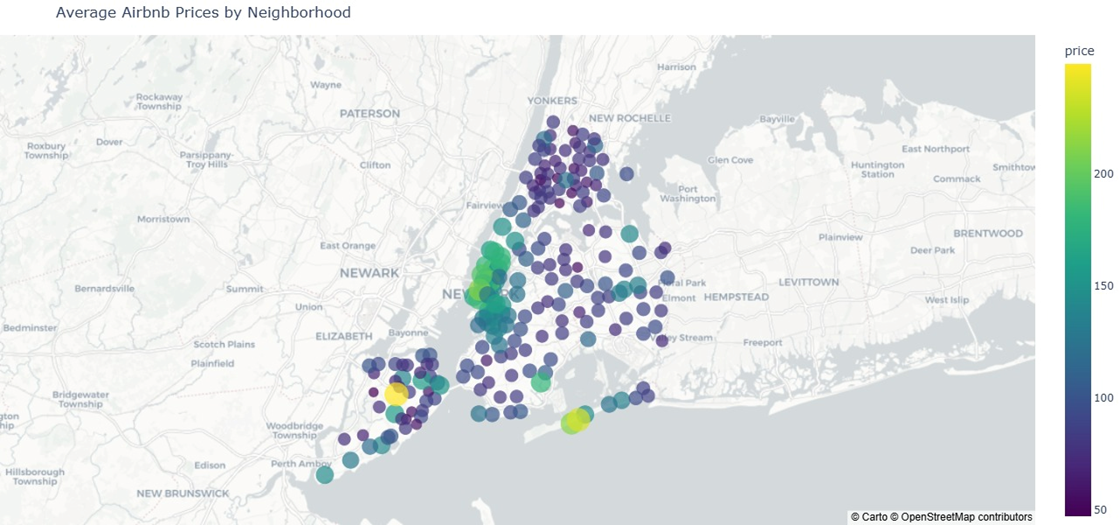
### An interactive web application was developed to visualize Airbnb data and make it accessible globally.

### **Process:**

* **Application Development:** The web application was built using Python with Flask for the backend and HTML, CSS, and JavaScript for the frontend, ensuring a responsive and interactive user interface.
* **Data Processing and Visualization:** The Airbnb dataset was cleaned by removing duplicates, filtering outliers in the 'price' column using the IQR method, and validating ranges for fields like 'availability\_365' (0–365) and 'minimum\_nights' (1–364). Geographic coordinates and text fields were standardized, ensuring data accuracy. The cleaned data was visualized through interactive box plots, scatter plots, and maps, created using Matplotlib and Plotly for seamless browser integration.
* **EC2 Instance Setup:** An AWS EC2 instance was launched and configured with Python, Flask, and project dependencies. The application code was deployed in a virtual environment for isolated and efficient deployment.
* **Database Integration and Deployment:** The application was connected to an Amazon RDS database for secure and efficient data management. The Flask app was hosted on the EC2 instance, with its public IPv4 address enabling global accessibility.

#### **Results of the analysis**

1. Average Airbnb Prices by Neighborhood:



**Overview:**

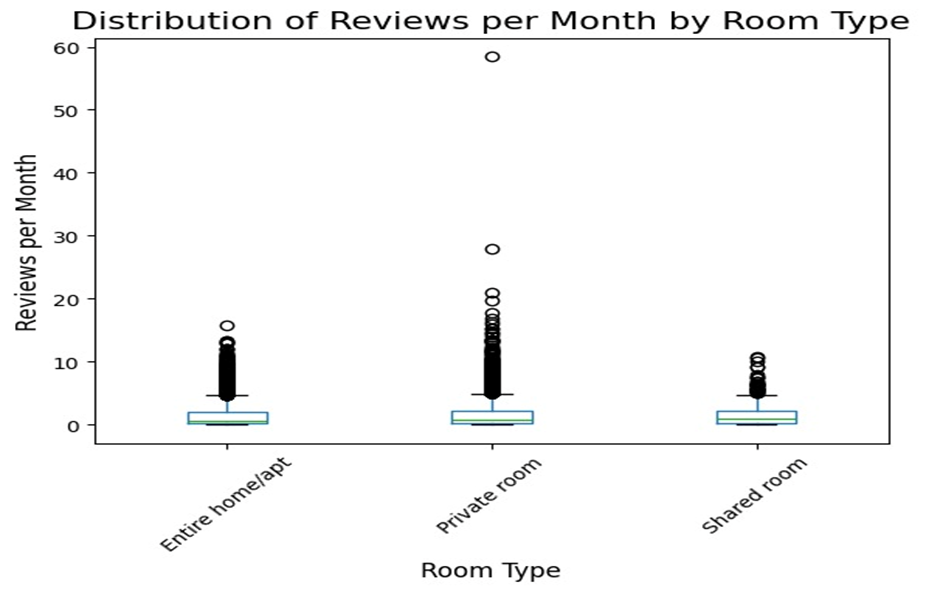
Manhattan was found to have the highest average Airbnb prices, which are mostly concentrated in central neighborhoods, as prices tend to fall in places farther from the city center, such as Brooklyn and Queen.

**EDA Analisys:**

Prices change visibly from neighborhood to neighborhood, with Manhattan having the highest average prices in the city. More central neighborhoods such as Midtown Manhattan and areas near other tourist attractions have higher average prices. *Neighborhoods farther away from the urban centers such as Brooklyn and Queens have lower average prices as they move away from the city center, creating a good deal for visitors on a budget.*

**Insights/Recommendations:**

Proximity to central business districts and attractions directly increases prices. Hosts in high-demand areas can receive help from seasonal fluctuations in demand. Hosts in peripheral neighborhoods should focus on alternative offerings and their lower prices.

1. Distribution of Reviews per Month by Room Type

**Overview:**

Private rooms clearly have more reviews than other room types, which may show consistent popularity. Except for a few outliers, entire homes/apartments and shared rooms do not have high activity.

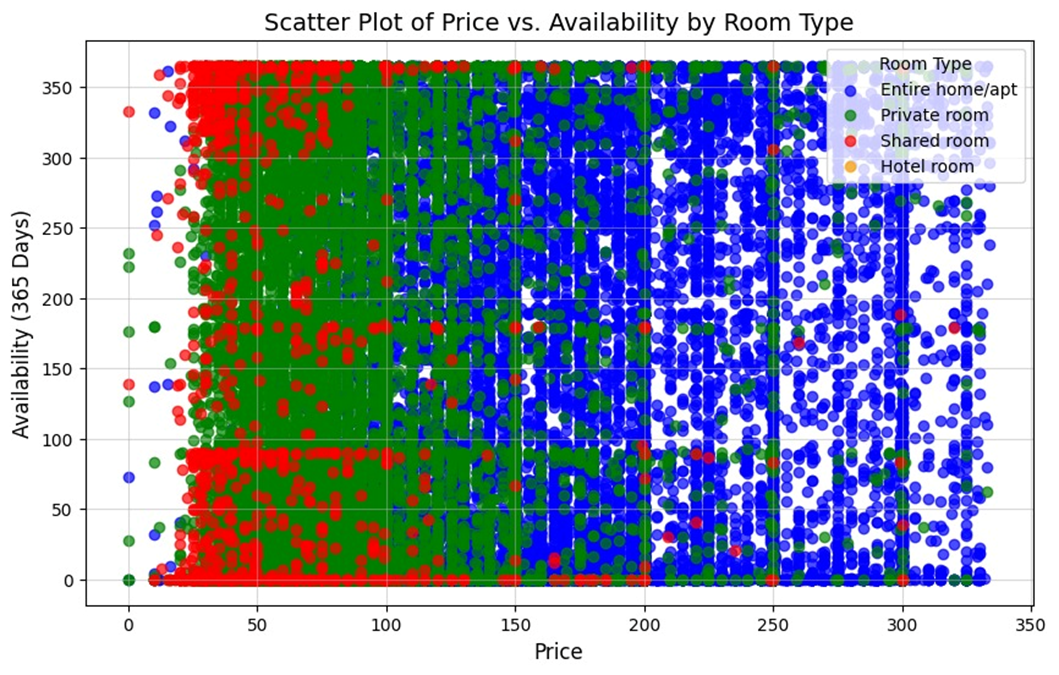
**EDA Analisys:**

Private rooms have the highest number of reviews by far even discounting peak demand, the entire houses are somewhere in between, while the least interest shown through reviews is for shared rooms, private rooms are the most popular.

**Insights/Recommendations:**

Private rooms obtain more than half of the total number of reviews. The shared rooms hardly arouse any interest. These facts should be a good sign for future investors in this market given the visible differences in interest from potential clients.

1. Scatter Plot of Price vs. Availability by Room Type



**Overview:**

All houses clearly have higher prices and less constant availability, as opposed to shared rooms which have the lowest prices followed by private rooms both with year-round availability.

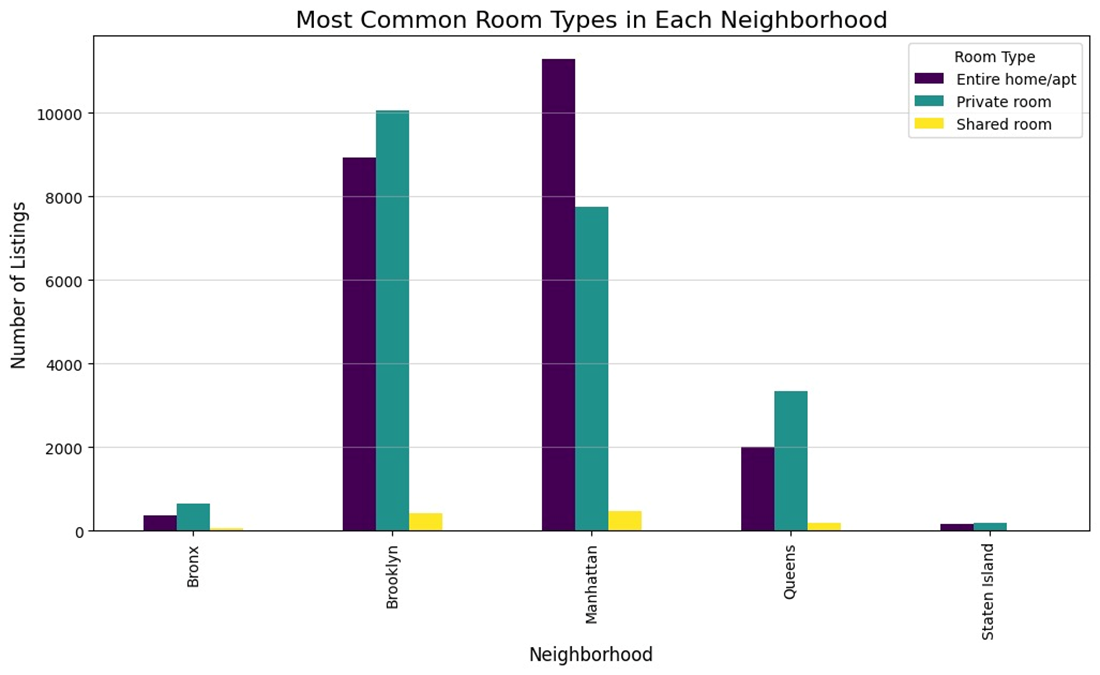
**EDA Analisys:**

Entire house/apartment (blue): Some available all year round, have the highest prices among the different options. Private room (green): Very affordable prices with high availability all year round. Shared room (red): Has the lowest prices and year-round availability. Hotel room (orange): Few listings, with moderate prices and variable availability throughout the year.

**Insights/Recommendations:**

Private rooms are the most common as they mix availability and low prices and can increase reservations by optimizing listings. The entire houses are more suitable for higher budgets and longer stays due to the annual availability.

1. Most Common Room Types in Each Neighborhood



**Overview:**

Most of the demand is in Manhattan and Brooklyn with a clear preference for private rooms and entire homes, with some movement in Queens, although in all cases the presence of activity for shared rooms is exceptionally low.

**EDA Analysis:**

Manhattan: Entire homes/apartments are the first choice, followed by private rooms. There is almost no demand for shared rooms. Brooklyn: Private rooms and entire homes have similar demands, again there is extraordinarily little demand for shared rooms. Queens: All demand is for private rooms, with fewer entire homes and barely any demand for shared rooms. Bronx and Staten Island: Very few listings, though again most are for private rooms.

**Insights/Recommendations:**

Manhattan has strong potential to attract more high-budget travelers, Brooklyn is for more balanced audiences. Queens hosts are recommended to focus on offering private rooms to take advantage of its popularity. Bronx and Staten Island should diversify their supply.

#### **Conclusions**

This project used AWS and Python to conduct an in-depth analysis of Airbnb data in New York City, revealing insights into pricing, availability, and room type patterns. Using Flask, an interactive web application was developed to visualize findings, supported by rigorous data cleaning to remove duplicates, filter outliers, and standardize key fields. The analysis showed that Manhattan has the highest prices, driven by proximity to business hubs and tourist attractions, while Queens and Brooklyn offer more affordable options. Private rooms appeared as the most popular choice, offering affordability and consistent availability, while full houses cater to higher budgets with less year-round availability. Shared rooms saw minimal demand, highlighting privacy as a key preference for guests. These findings offer valuable guidance for hosts and investors to tailor offerings based on location, room type, and guest preferences.