Machine Learning Module

Development Team Project

Group 1

Comprehensive Analysis of Airbnb Listing Dynamics in New York City

Table of Contents

Table of Figures

Introduction

Airbnb has five million landlords and has arranged a billion and a half stays worldwide (Airbnb Newsroom, 2024). In New York City, the variety of listings demands an analysis of how location, price, and room type influence guest choices and host earnings, and this one will uncover patterns that can inform strategic decisions and enhance guest satisfaction and host profitability.

Business Analytic Question

"How do factors such as neighbourhood, room type, and pricing affect the dynamics of Airbnb listings in New York City?" The insights gained can be used to recommend prices, propose deals, target advertising, and financially forecast for different areas.

Methodology

Data Preprocessing

The analysis utilises the AB_NYC_2019.csv dataset from Kaggle, listing Airbnb lets in New York City for 2019 (Kaggle, 2019), including attributes such as geographical coordinates, prices, room types, availability (Fig 1).

Data description

This dataset has 48,895 rows and 16 columns. The columns include:

- **id** ID for each property
- **name** Description of the property as provided by the host
- **host_id** ID for each host
- **host_name** Name of the host
- **neighbourhood_group** Name of the neighbourhood group
- **neighbourhood** Name of the neighbourhood
- **latitude** Location Latitude of the property
- **longitude** Location Longitude of the property
- **room_type** Type of Room
- **price** Price of the Room
- **minimum_nights** Minimum number of nights the room is available for booking
- **number_of_reviews** Number of reviews the property has received from past guests
- **last_review** Date of Last Review
- **reviews_per_month** Average Number of reviews received per month
- **calculated_host_listings_count** Total number of properties available for Rent from this host
- **availability_365** Number of Days out of the last 365 days that the property has been available for rent

Figure 1: Dataset Attributes

Preprocessing involved dropping irrelevant columns like *'host _name'* and *'last_review'* and handling of missing values by imputation or removal. Figure 2 shows missing values in *'reviews_per_month'*, which were set to 0, and *'name'*, which were removed.
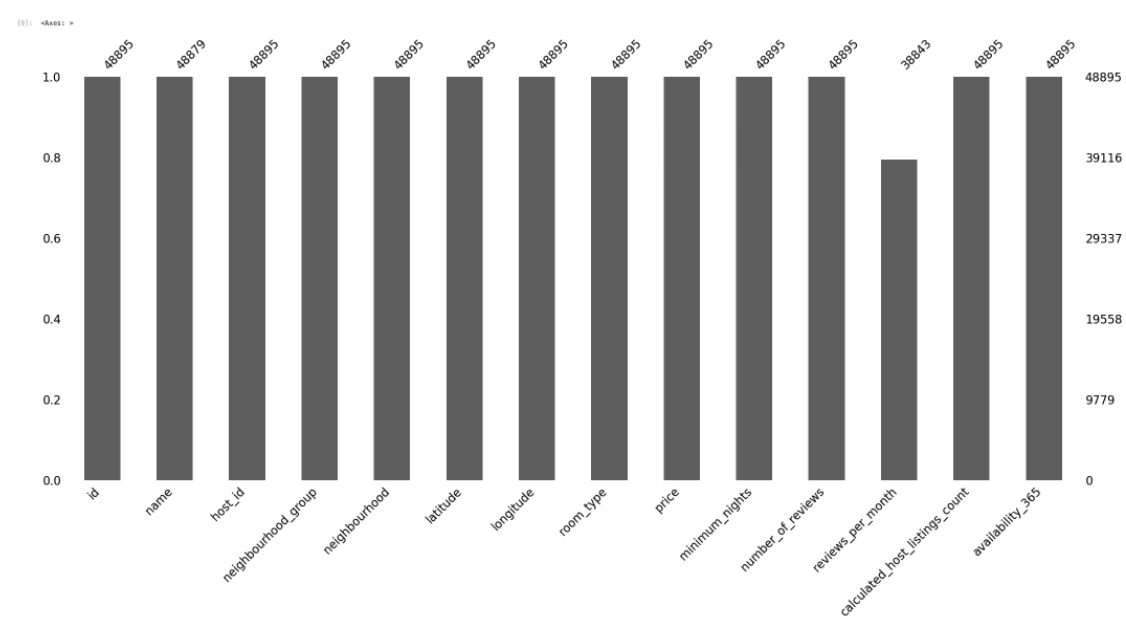

Figure 2: Plot for missing values.

Data Analysis

Distributions of selected features were investigated using density histograms and box plots (Figs 3, 4) to identify medians and outliers (Kubat 2021).
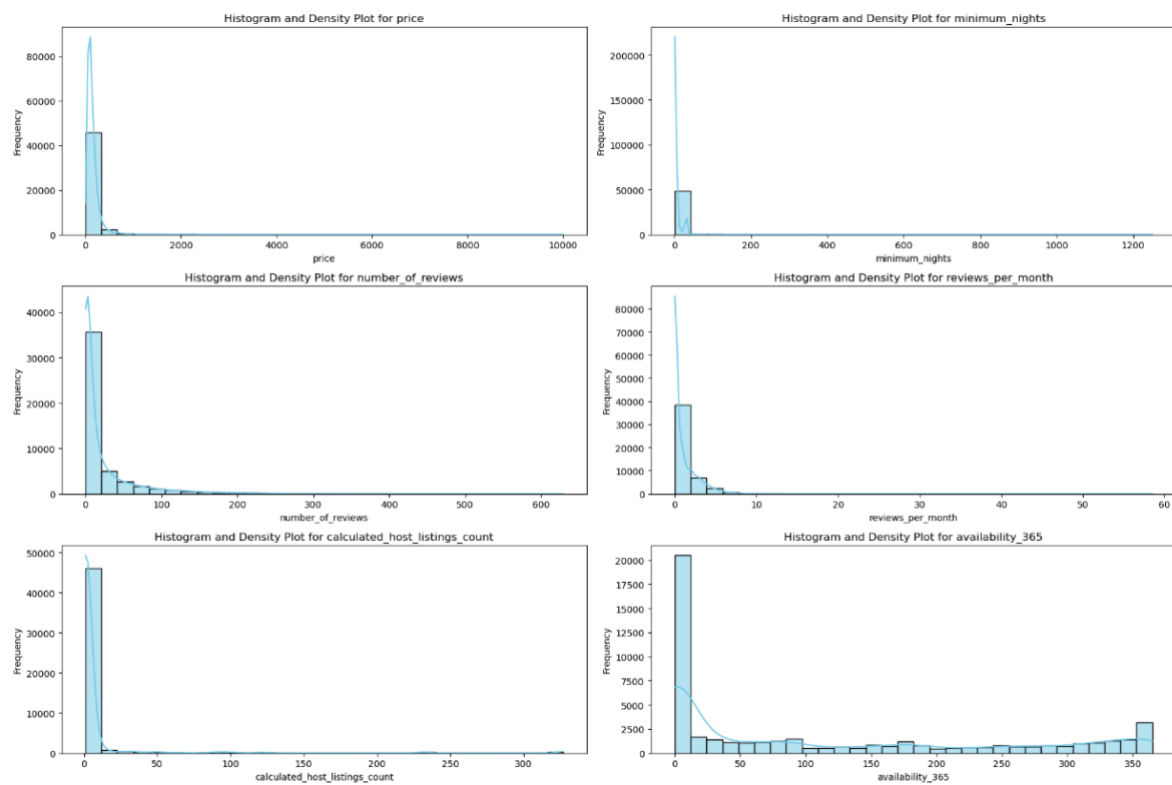

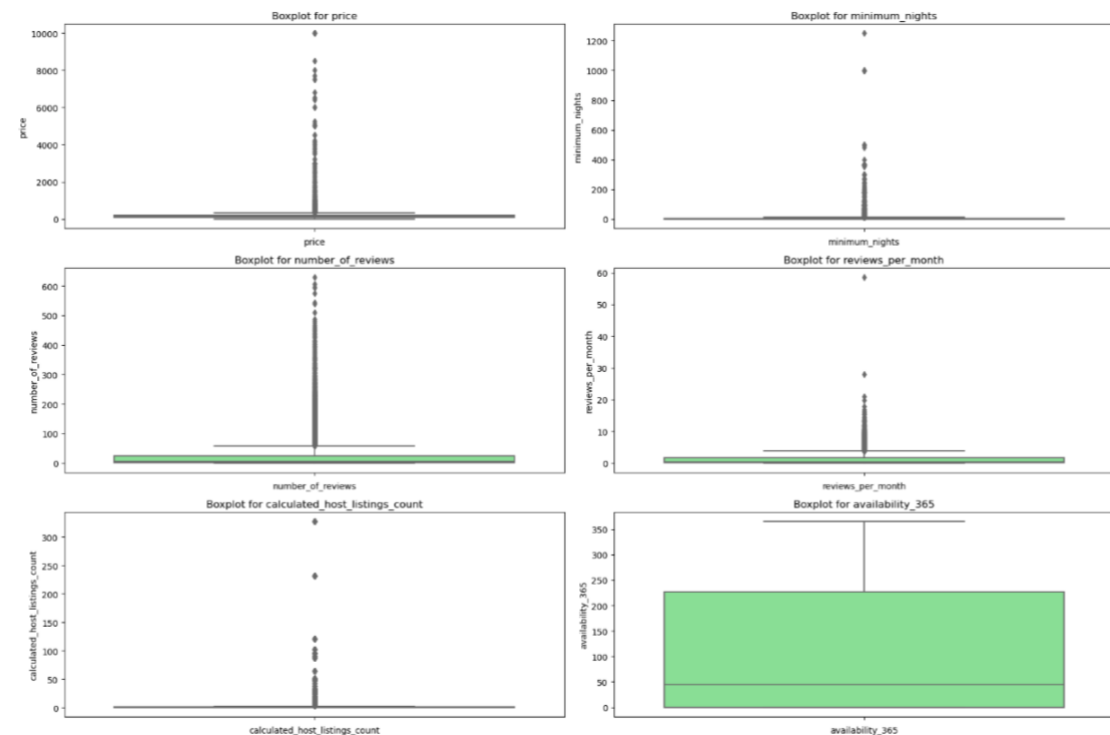Figure 3: Density Plots and Histograms for Numerical Features

Figure 4: Box Plots for Numerical Features

Density histograms showed most listings priced at the lower end of the market, indicating a concentration of affordable options. Most listings require only a few nights' stay, suggesting a market geared towards short-term accommodation. Review-related features skewed towards fewer reviews, many listings having none. The 'calculated_host_listings_count' histogram showed many hosts managing just one or a small collection of properties, illustrating the platform's foundational role in supporting small-scale hosts. The presence of some hosts with multiple listings points towards professional or semi-professional hosts leveraging Airbnb for broader property management purposes.

The 'availability_365' feature (the annual number of days a listing is available) presented a wide-ranging distribution. Low values reflect hosts' preference to rent out their properties part-time or seasonally. Listings available most days corresponding to dedicated rental properties aimed at consistent income generation.

The box plots highlighted the median values for the features, and pinpointing outliers, particularly for premium listings with higher prices and longer minimum stays, revealing a diverse Airbnb host community in New York City, from those offering private spaces for extra income to those managing several listings as a business. This diversity in hosting patterns has implications for market competitiveness and customer choice within the Airbnb ecosystem in New York City.

From Figure 5, Manhattan and Brooklyn boast the highest concentration of listings, echoing their appeal. The Bronx and Staten Island exhibit fewer listings, which could mean less tourism or emerging Airbnb markets. Figure 6's scatter plot shows a dense accumulation of listings in Manhattan, while Brooklyn enjoys a substantial spread of listings. The sparser distribution in Queens, Staten Island, and the Bronx implies a less condensed market presence.
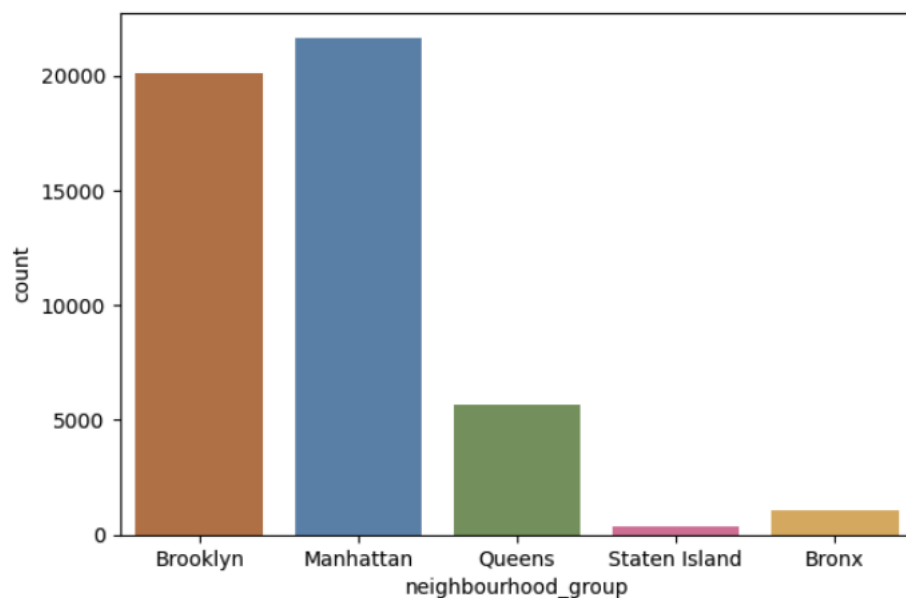

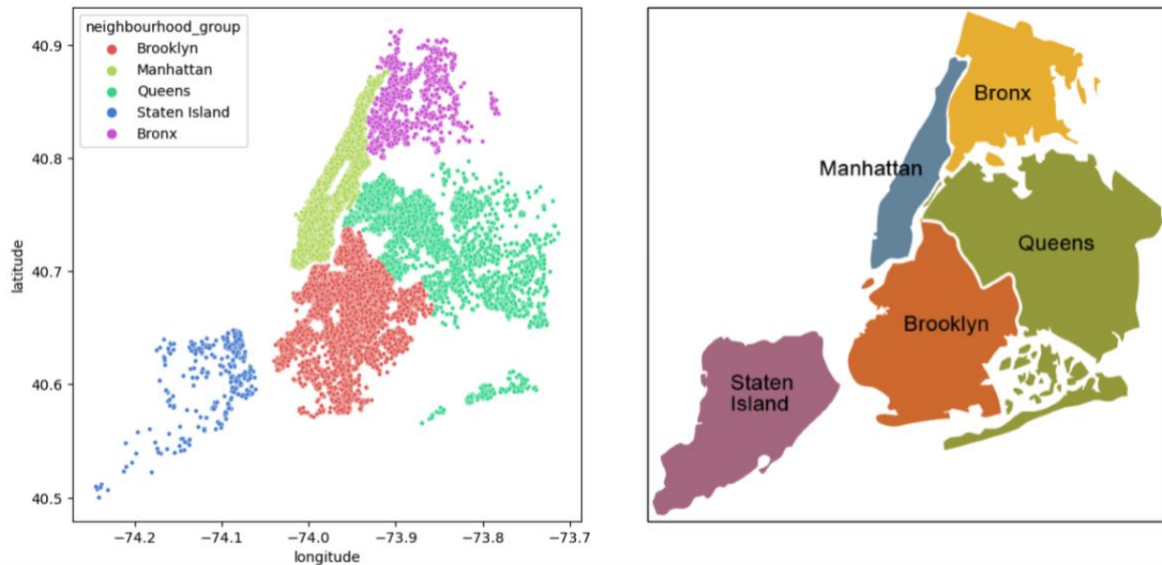Figure 5: Distribution of Listings across New York

Figure 6: Scatter Plot showing the distribution of listing across New York City

The kernel density estimation (KDE) plots (Wand & Jones, 1994) in figures 7 and 8 below reveal NYC Airbnb listings skew towards affordability, especially in Brooklyn and Queens, while Manhattan's diverse offerings span from budget to luxury, with a hint of premium properties. Short-term stays dominate, reflecting tourists and brief visits. Review distributions suggest higher turnover in Manhattan and Brooklyn, with Staten Island and the Bronx possibly featuring newer or less frequented listings. Manhattan has more multi-listing hosts, indicating professional operations. Availability trends point to a mix of strategies, from year-round rentals to occasional hosting, showcasing the market's varied dynamics.
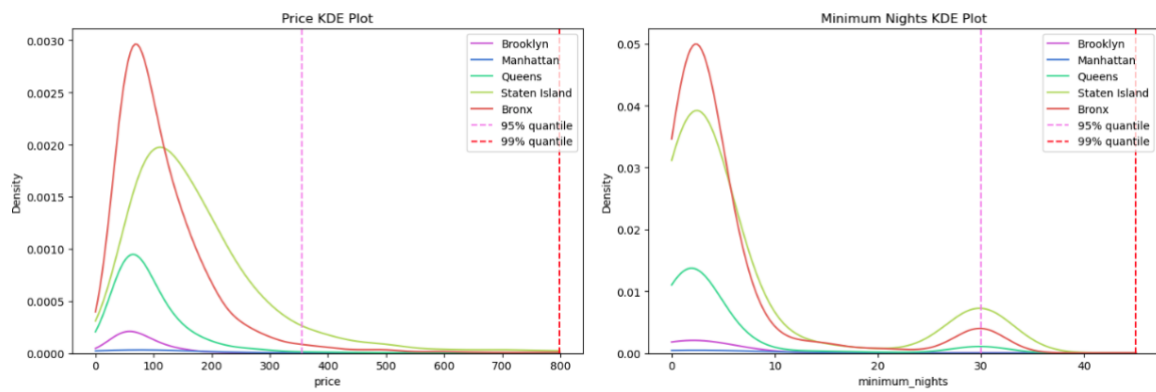
Figure 7: KDE plots from price and minimum nights across different neighbourhoods
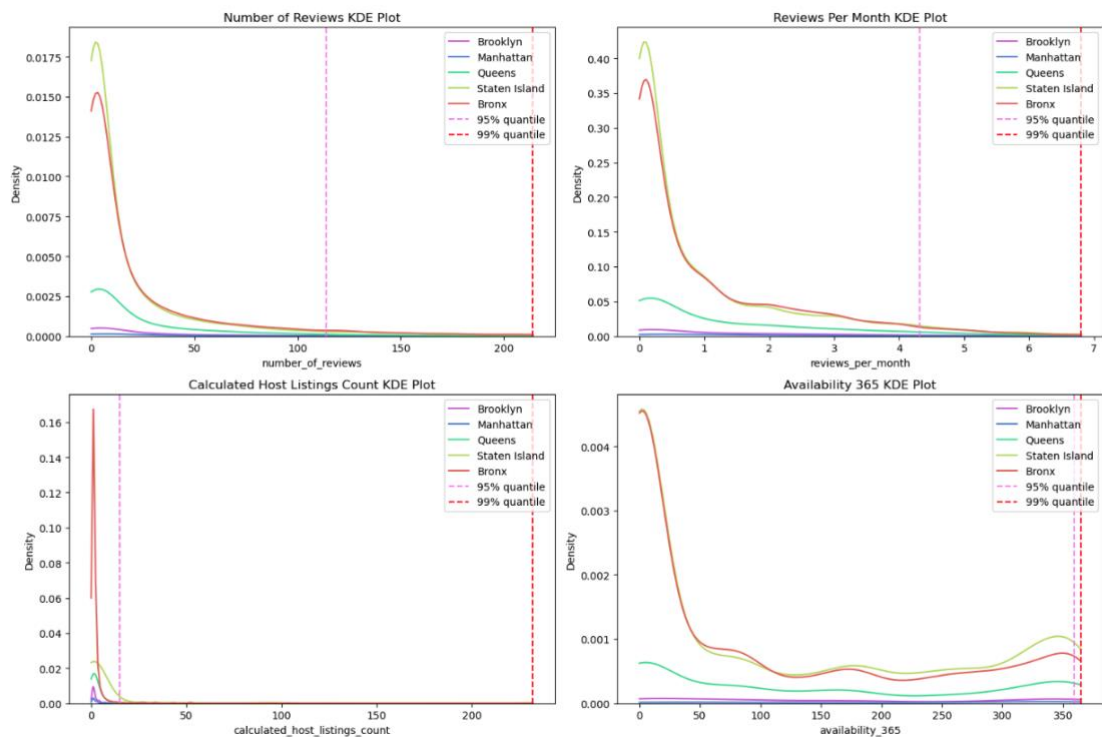

Figure 8: KDE plots across different neighbourhoods

The bar charts in Figures 9 and 10 reveal activity by neighborhood and room type. The 'Price' and 'Minimum Nights' charts show price segmentation and stay requirements, with Manhattan, a prime location, showing the highest average prices across all room types. Staten Island and the Bronx show lower prices. The reviews charts show higher engagement with listings in more frequented areas like Manhattan and Brooklyn, possibly due to higher guest turnover. In the 'Calculated Host Listings Count' chart, Manhattan has a higher

average number of listings per host, who are perhaps more experienced or professional there. The 'Availability_365' chart indicates that listings in Manhattan and Staten Island have higher availability, while other boroughs show varied availability, perhaps due to hosts' strategies or external factors.
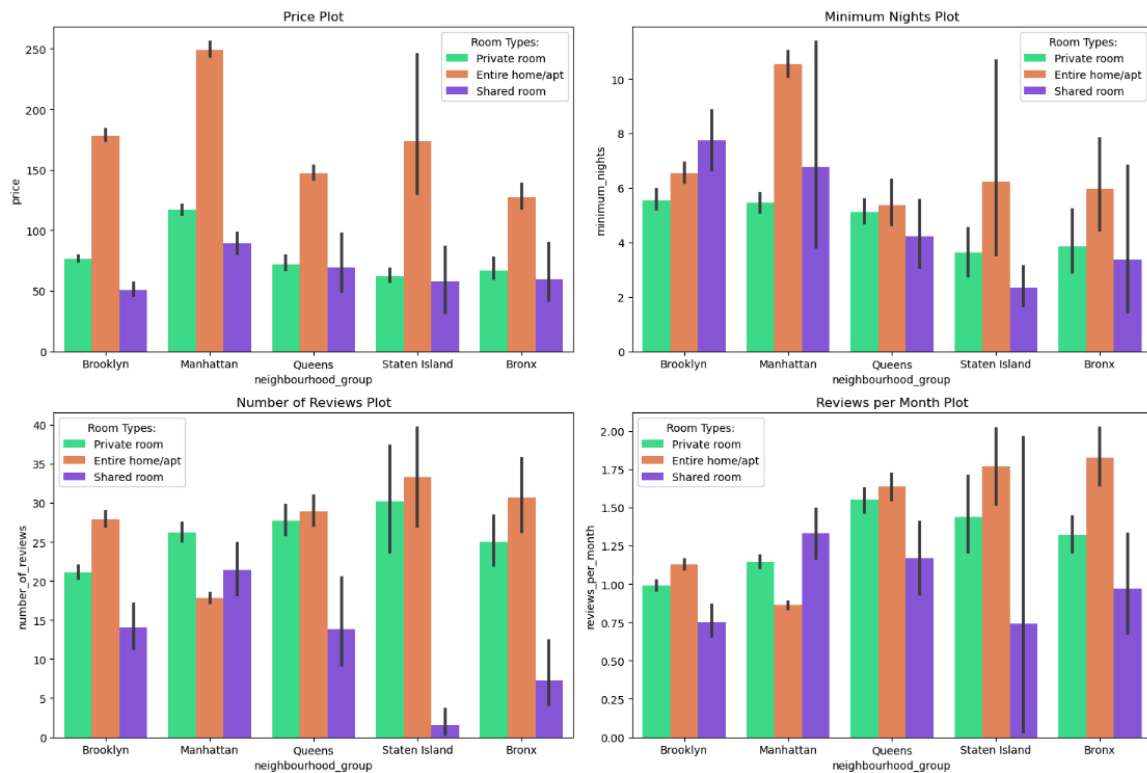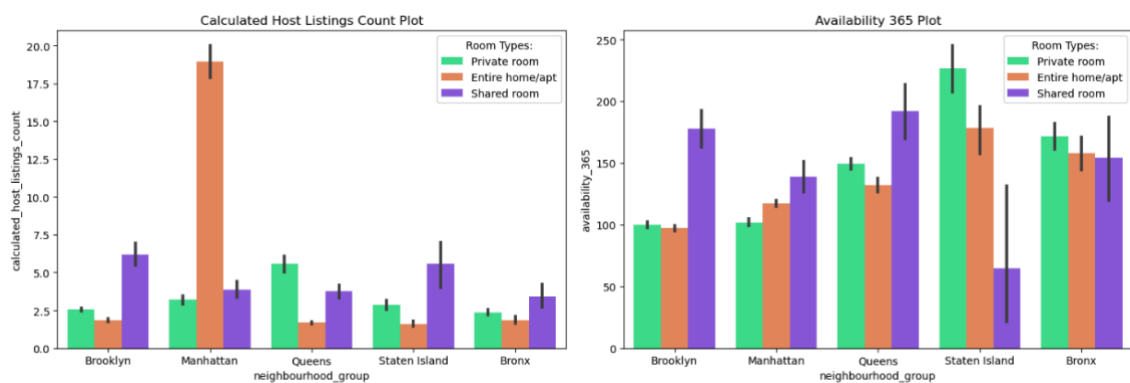

Figure 9: Bar Charts 1.0

Figure 10: Bar Charts 1.1

Discussion and Recommendation

The analysis of Airbnb listings in New York City reveals diverse offerings and host engagement. Brooklyn and Queens cater to tourists through affordable and short-term accommodations, while Manhattan offers a spectrum from budget-friendly to premium, with a tendency toward professional hosting. Growth opportunities abound in the Bronx and Staten Island, where the market is less saturated. Year-round availability in these areas also indicates a possibility for a stable income for hosts and consistent choices for guests. This approach, alongside a diversified property portfolio, could enhance bookings especially in the Bronx and Staten Island's emerging markets.

[1003 words]

References

Airbnb Newsroom. (2024). *About us - Airbnb Newsroom*. https://news.airbnb.com/about-us/

Kaggle. (2019). *New York City Airbnb Open Data*. https://www.kaggle.com/datasets/dgomonov/new-

  york-city-airbnb-open-data

Kubat, M. (2021). An Introduction to Machine Learning. 3rd ed. Cham, Switzerland: Springer.

Wand, M. P. & Jones, M. C. (1994). Kernel Smoothing. *Chapman & Hall/CRC Monographs on*

  Statistics & Applied Probability (60). Boca Raton, FL, U.S.: Chapman & Hall.

Appendix: Python Code link

https://github.com/Ngugi-Joy-Grace/airbnb-business-analysis/blob/main/analysis.ipynb