PROJECT

REPORT

Airbnb NYC EDA & Data Viz Python Project

Author: Saiful Islam Rupom

https://github.com/saiful-islam-rupom/airbnb-nyc-eda-and-data-viz-python-project.git

Executive Summary

In this project, I performed Exploratory Data Analysis (EDA) and Data Visualization on Airbnb listings data of New York City to extract meaningful insights, identify trends and better understand the dynamics of the short-term rental market. Here, I used different popular data science Python libraries like Numpy, Pandas, Matplotlib & Seaborn for data cleaning, visualization & analysis.

Objective

The goal of this project is to:

- Understand the structure and features of Airbnb listings in New York City.
- Detect potential outliers and clean the data properly.
- Analyze room types, prices, ratings & availability across different neighborhood groups.
- Identify key trends and patterns in host behavior, pricing, room types, availability and neighborhood statistics.
- Create visualizations to support business and user decision-making.
- Practice EDA techniques that are essential for data-driven roles in a company.

Dataset

The dataset contains 20,770 entries and 22 features, including:

- Listing IDs and names
- Host details
- Neighborhoods and locations
- Room types
- Pricing
- Availability
- Review scores

Tools & Libraries

- Python 3.10+
- Pandas – Data wrangling
- NumPy – Numerical operations
- Matplotlib & Seaborn – Static visualizations
- Jupyter Notebook – Analysis environment

Steps and Workflow

1. Data Cleaning

- Handling missing data: neighbourhood, latitude, longitude, room_type, price, minimum_nights, number_of_reviews, last_review, reviews_per_month, calculated_host_listings_count, availability_365, number_of_reviews_ltm columns had some null values.
- Deleting duplicates: 12 duplicated rows were dropped.
- Fixing data types: Transformed into proper data types of those columns whose data types were not correct.
- Removing outliers: Listings with prices > $2000 were capped to avoid skewed visualizations.

2. Feature Engineering

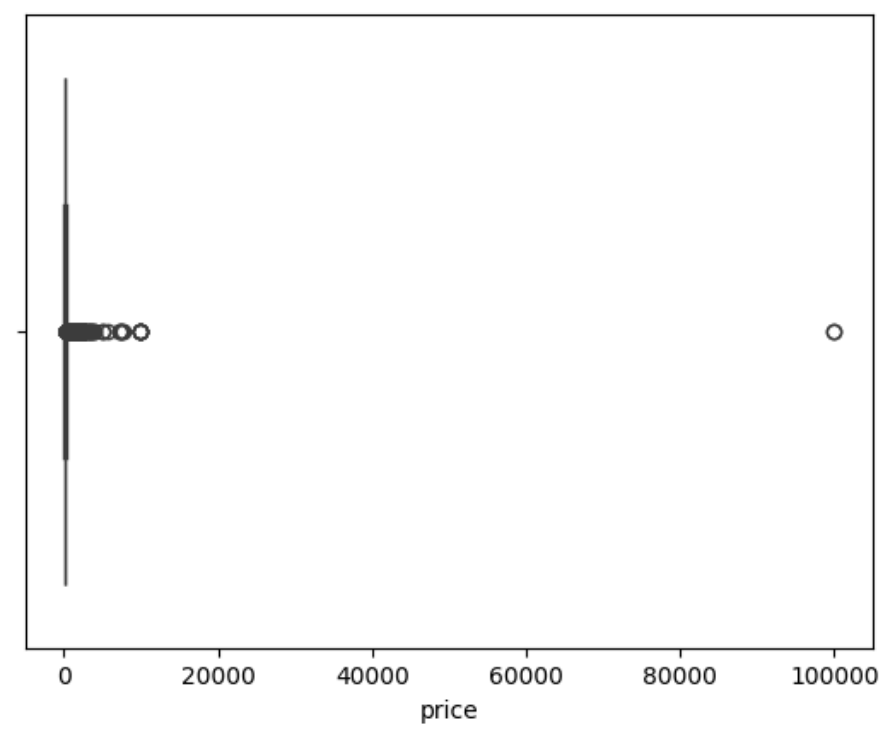
- Adding column 'price_per_bed': Create a new column using price and number of bed

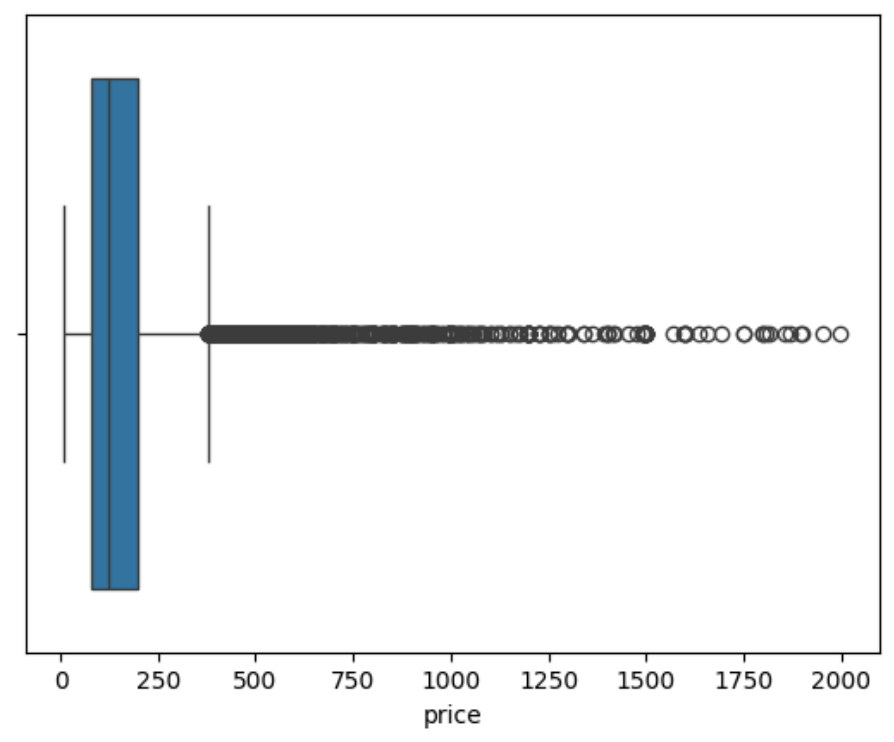
3. EDA (Exploratory Data Analysis)

1. Identifying and vizualizing outliers in Price.
2. Price Distribution
3. Distribution and percentage of different Room Types
4. Average Price by Neighbourhood Group
5. Top 10 Expensive Neighbourhoods (Avg Price)
6. Average Price per Bed by Neighbourhood Group
7. Room Type Distribution (Percentage) by Neighbourhood Group
8. Price dependency on Room Type by Neighbourhood Group (Average price)
9. Top 10 Neighbourhoods by Availability (Avg)
10. Average Availability by Room Type
11. Top 10 Hosts by Listings and their Average Ratings
12. Top 10 Hosts With The Lowest Average Ratings
13. Relationship Between Number of Reviews and Price
14. Relationships among different different columns: price, minimum_nights, number_of_reviews, availability_365

15. Geographical Distribution of Airbnb Listings in NYC and Classification by Neighbourhood Groups
16. Geographical Distribution of Different Types of Rooms Available in New York City
17. Geographical Distribution of Different Types of Rooms Available in Different Neighbourhood Groups
18. Correlation of one variable with others for each numerical column

EDA 1: Identifying and vizualizing outliers in Price.

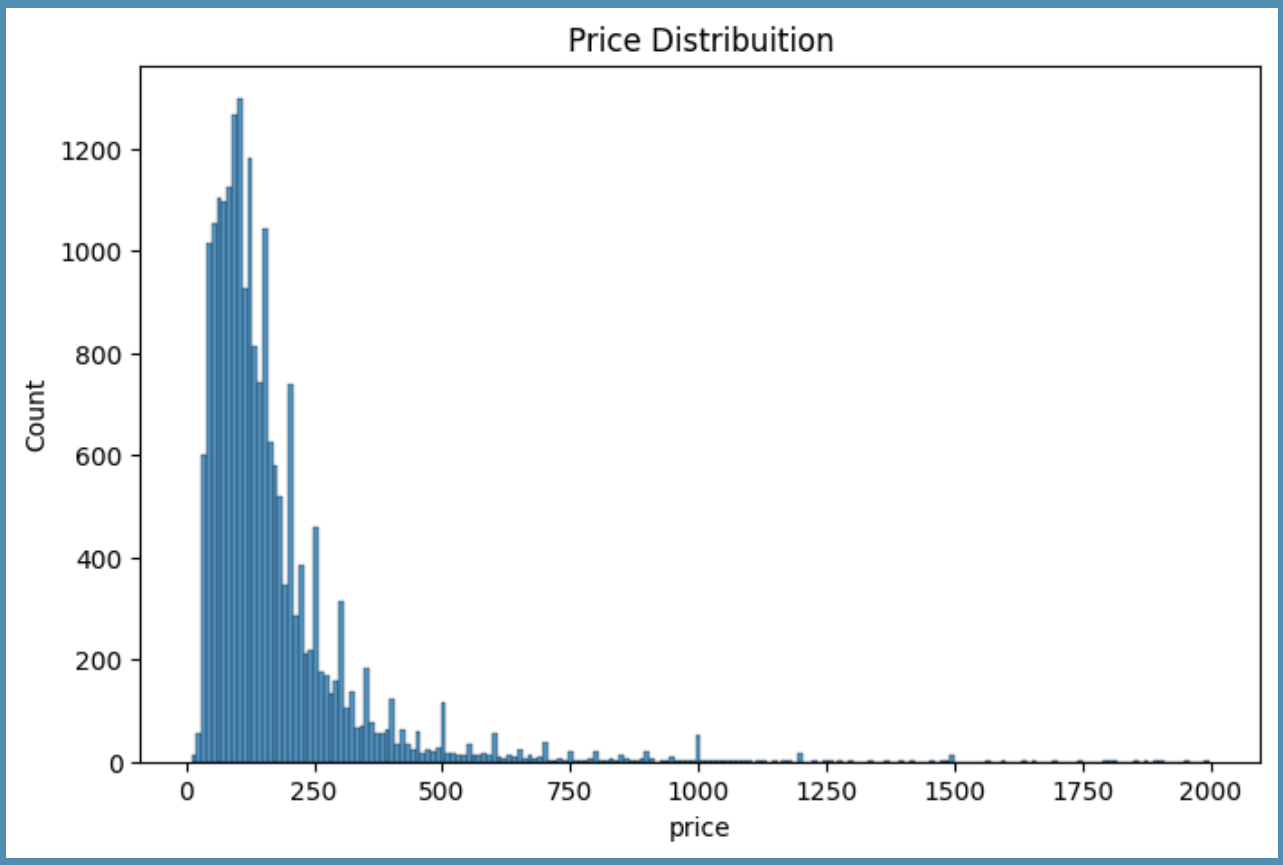




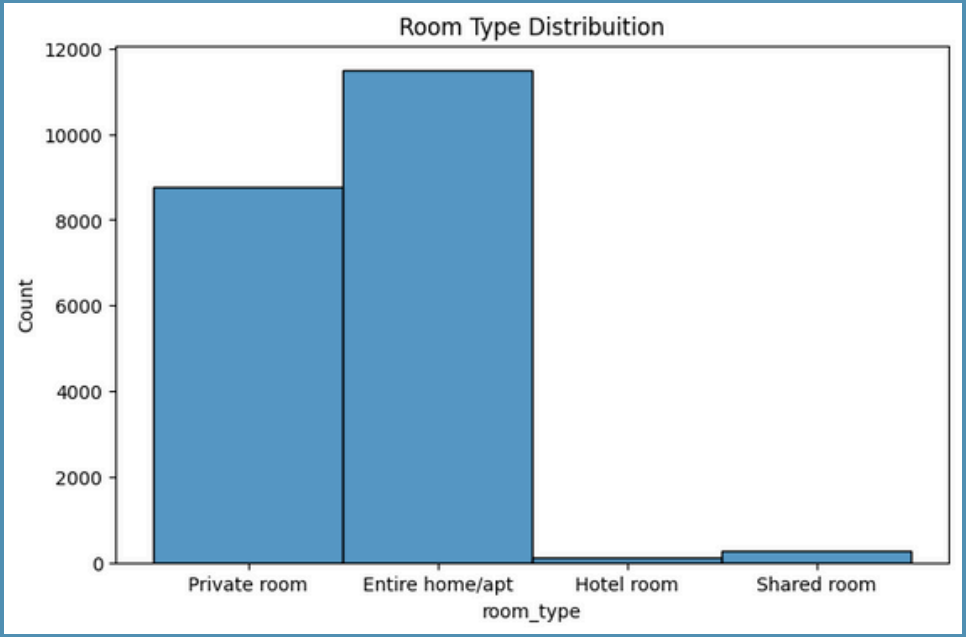
Findings & Insights:

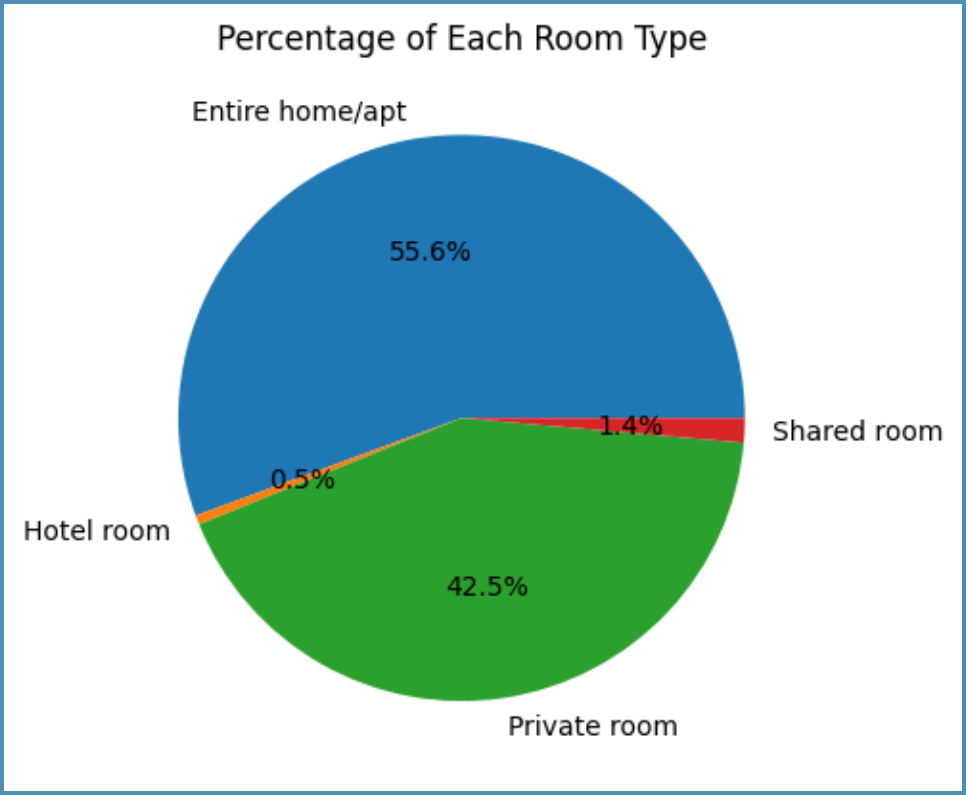
- Outliers in price that has some listings more than $10000 and also a listing of about $100000, indicating the need to filter such extreme values.

EDA 2: Price Distribution



EDA 3: Distribution and percentage of different Room Types

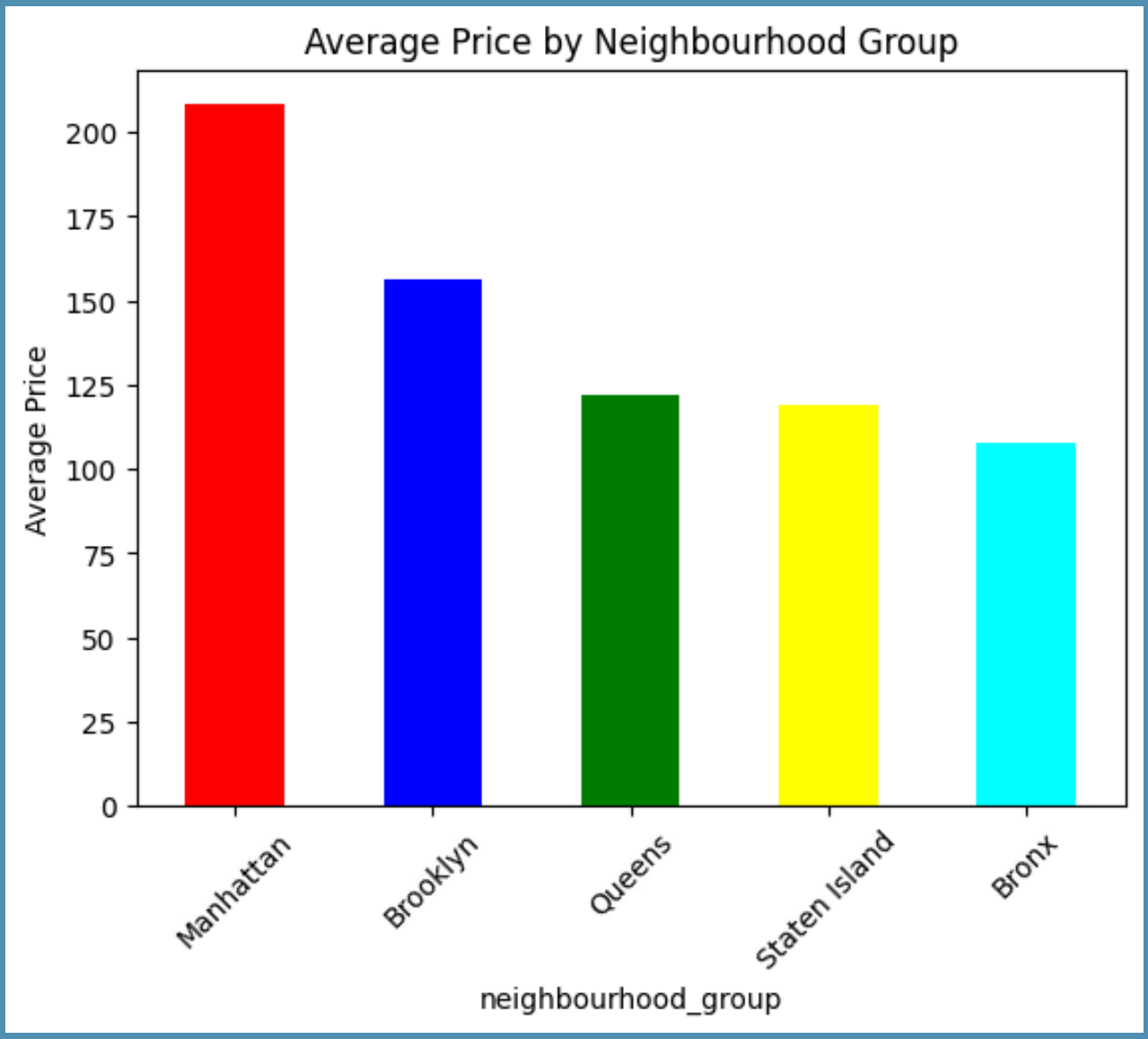




Findings & Insights:

- Entire home/apt. has got most amount of listings about 55.6%, followed Private room with 42.5% listings.
- Hotel room & Shared room got least amount of listings respectively 0.5% and 1.4%

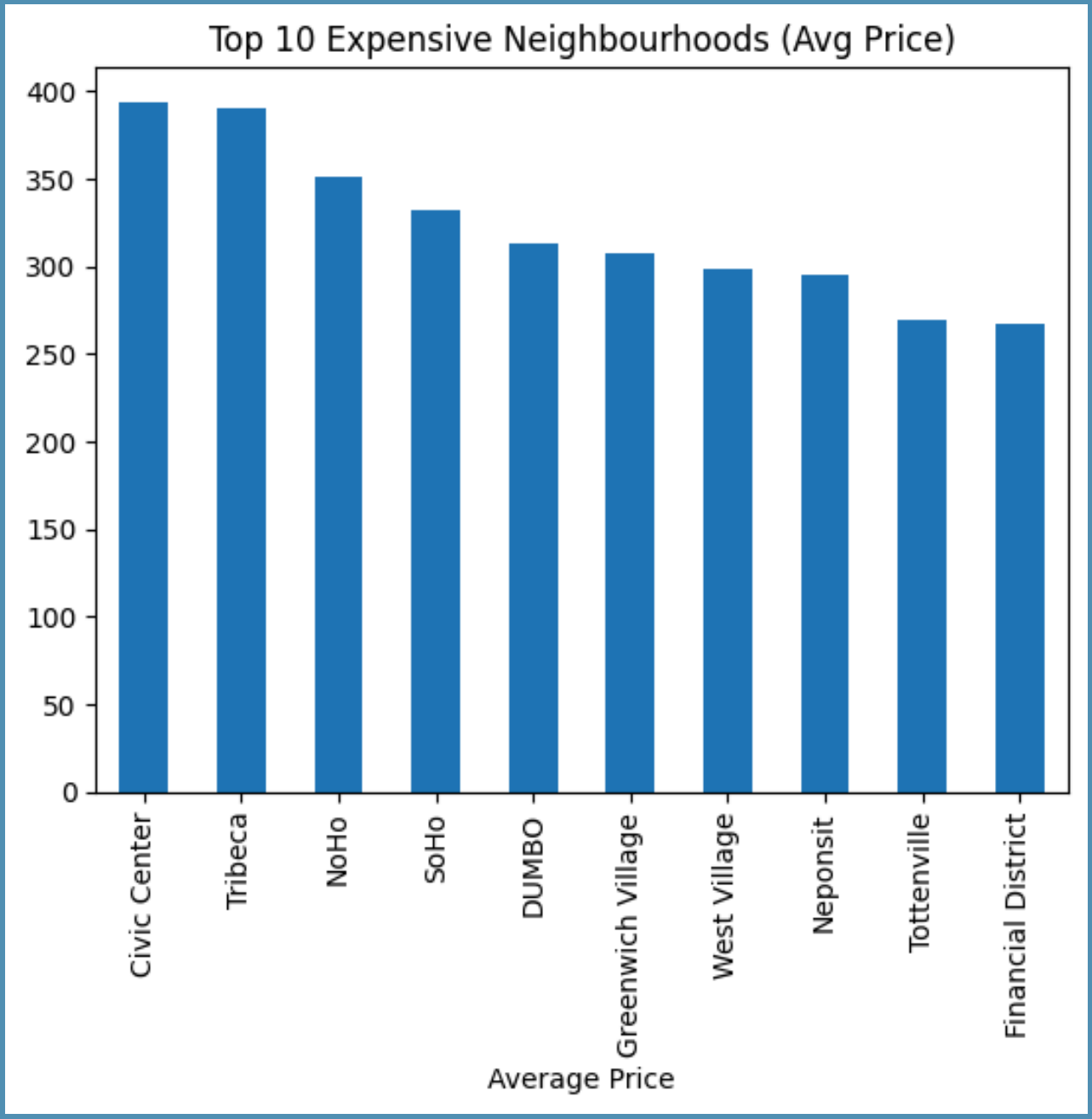
EDA 4: Average Price by Neighbourhood Group



Findings & Insights:

- Manhattan has the most expensive $208 average listing price, followed by Brooklyn.

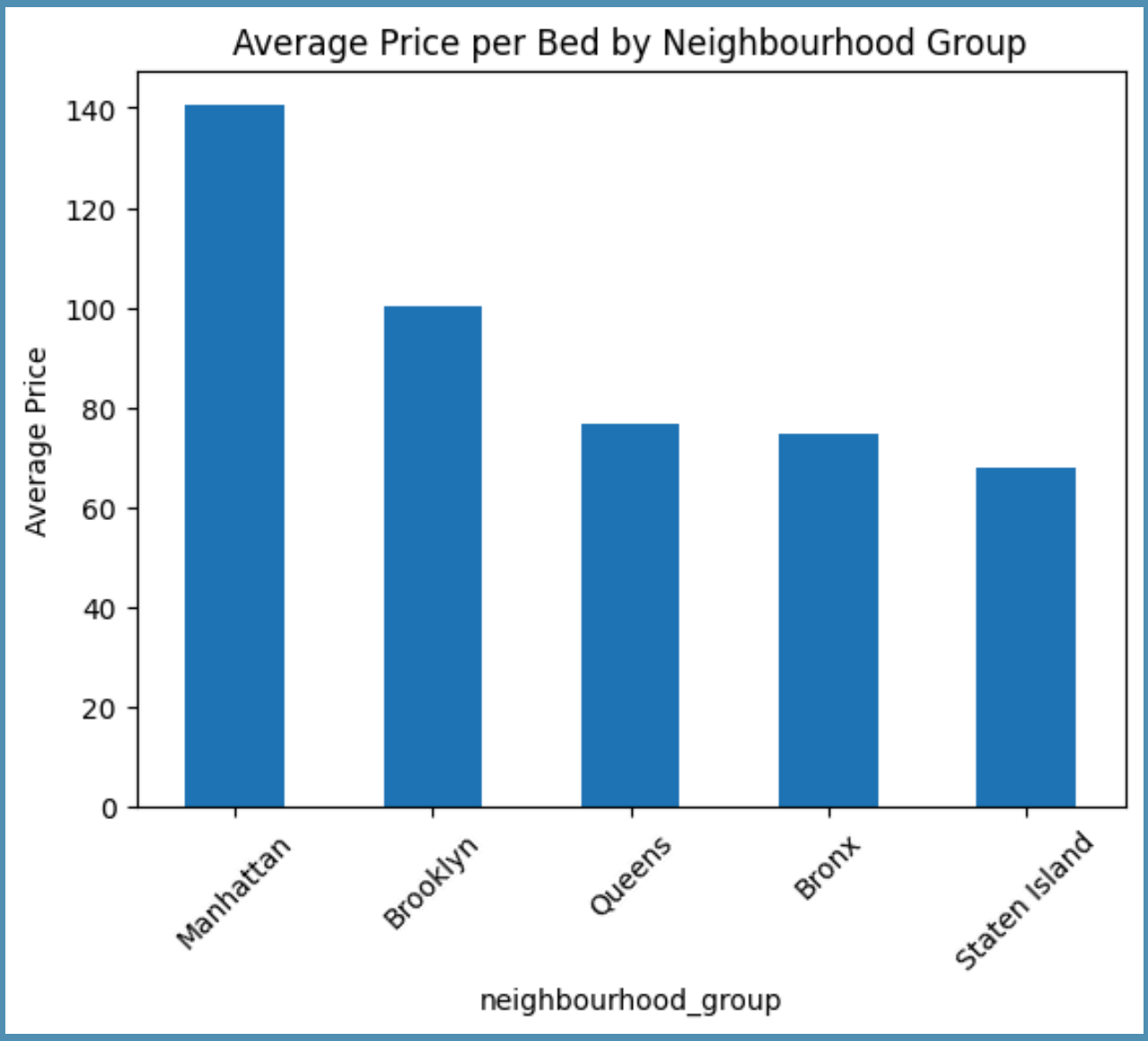
EDA 5: Top 10 Expensive Neighbourhoods (Avg Price)



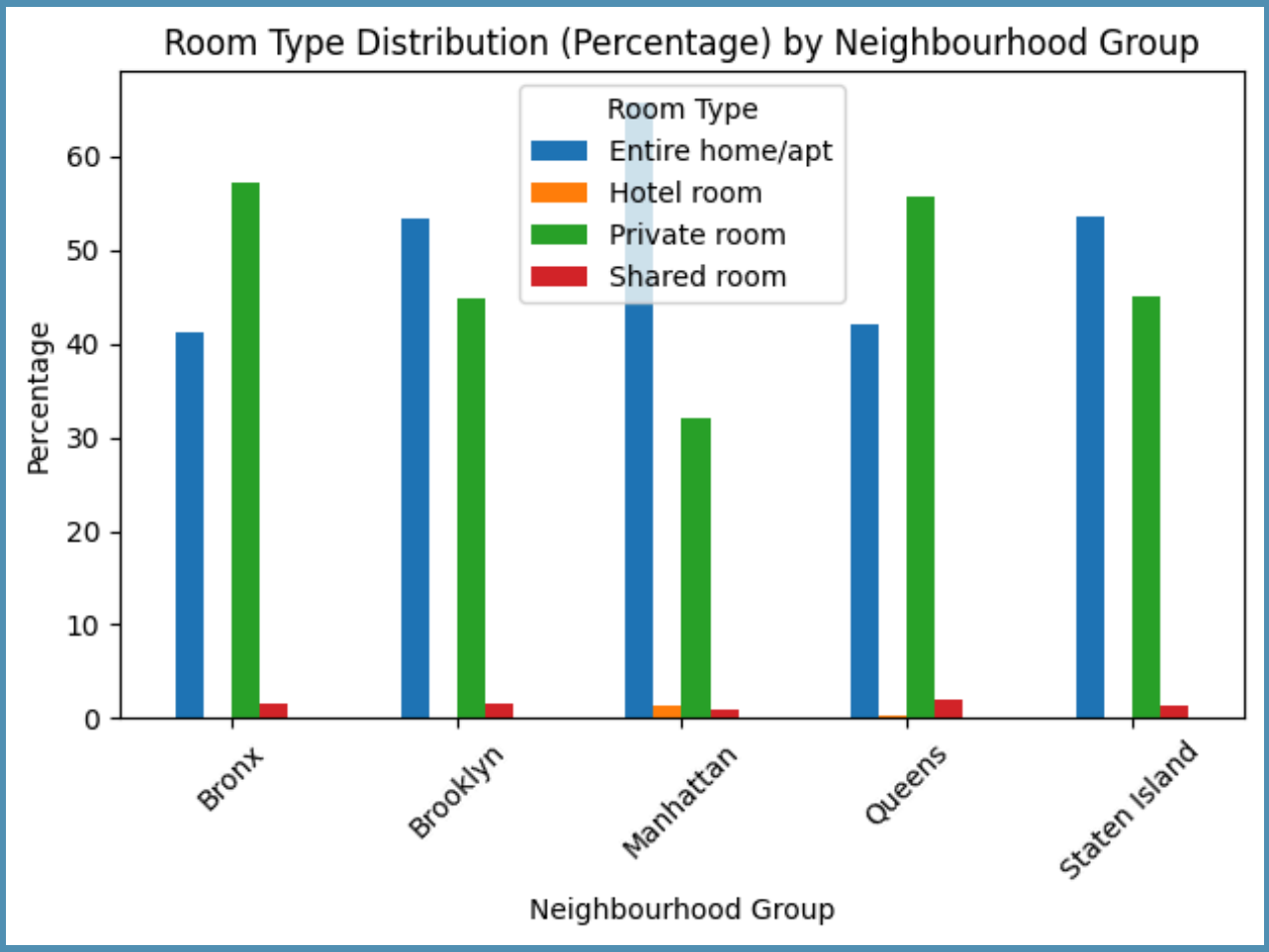
Findings & Insights:

- Top expensive locality are Civic Center & Tribeca with $390+ average listings.

EDA 6: Average Price per Bed by Neighbourhood Group



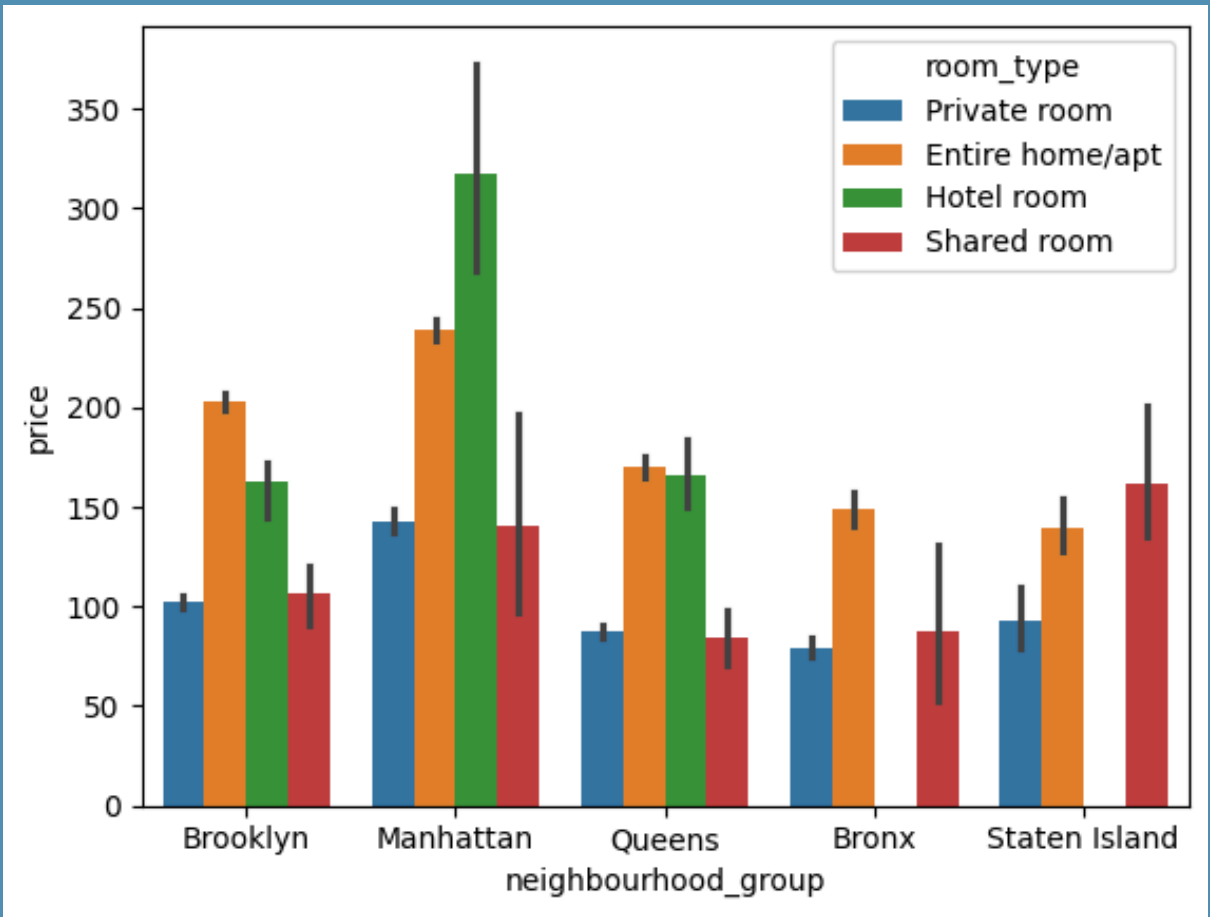
EDA 7: Room Type Distribution (Percentage) by Neighbourhood Group



Findings & Insights:

- Most number of Entire home/apt are in Manhattan about 65% of Manhattan's total listings.
- Bronx and Staten Island have no Hotel room.

EDA 8: Price dependency on Room Type by Neighbourhood Group (Average price)



Findings & Insights:

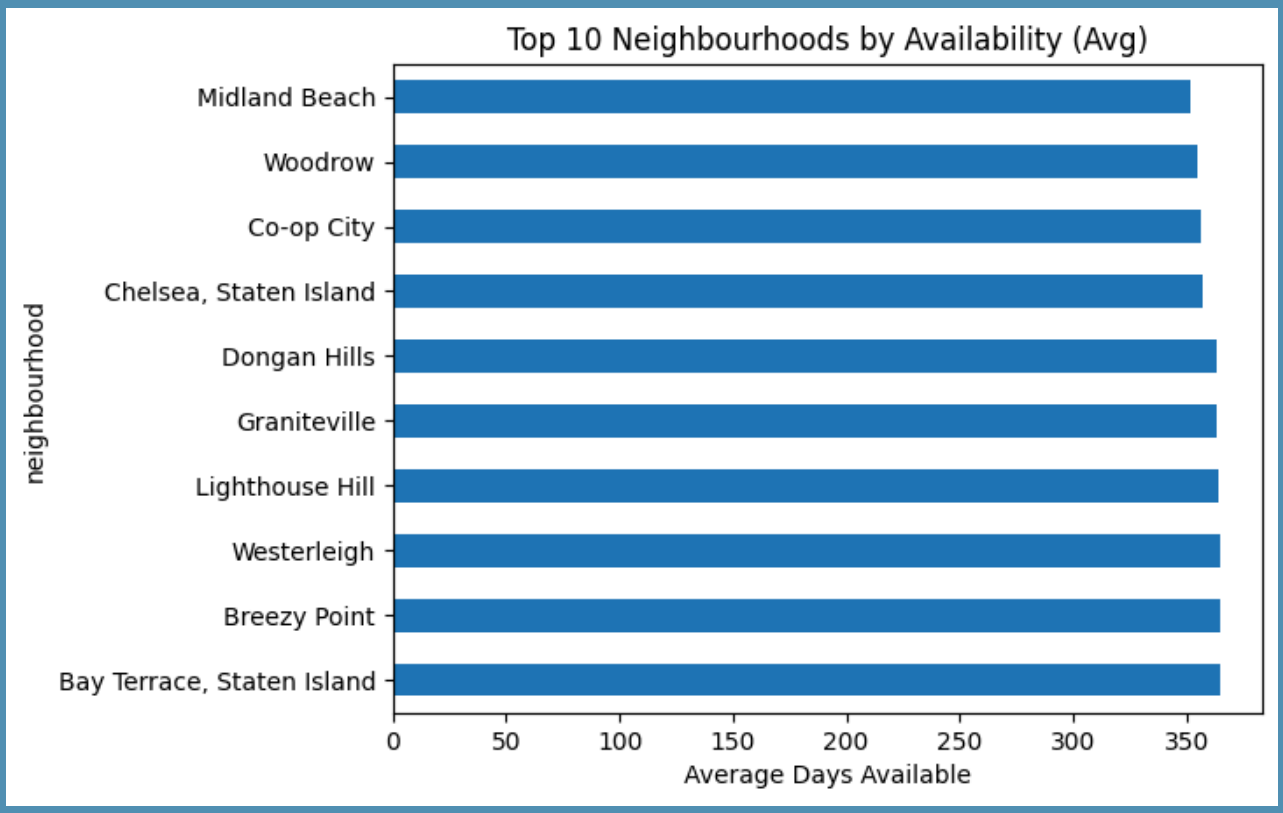
Expensive Room types by Area(Neighbourhood Group):

- Private room, Entire home/apt & Hotel rooms are expensive in Manhattan.
- Shared rooms are expensive in Staten Island about $150 average pricing.

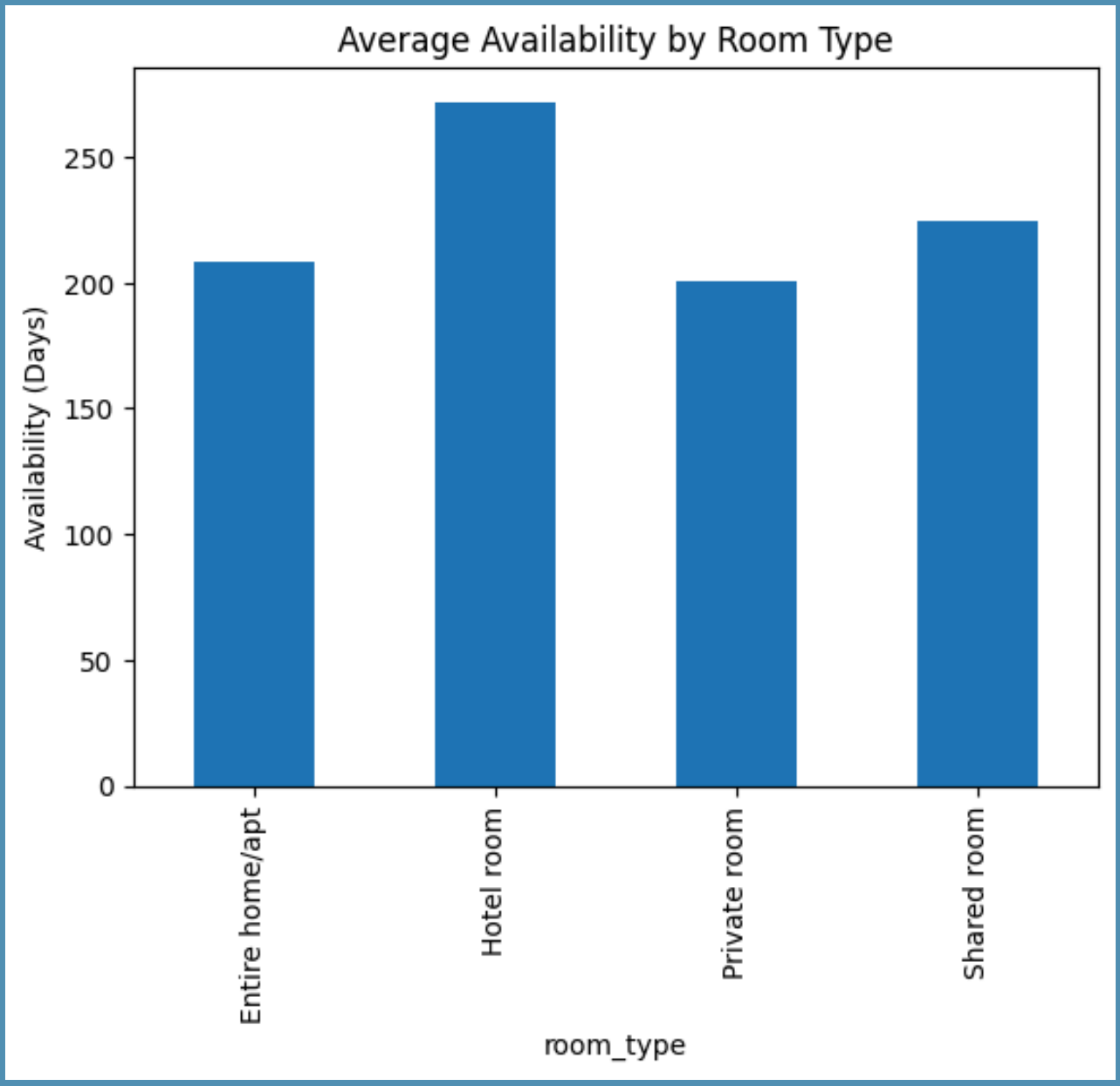
Affordable Room types by Area(Neighbourhood Group):

- Most inexpensive Private rooms are in Bronx about $75 average costing.
- Affordable Entire home/apt are available in Staten Island with $130 average pricing.
- Cheap Hotel rooms are available in both Brooklyn and Queens with $160 average pricing.
- Queens has the most affordable Shared rooms, those average price is about $80 only.

EDA 9: Top 10 Neighbourhoods by Availability (Avg)



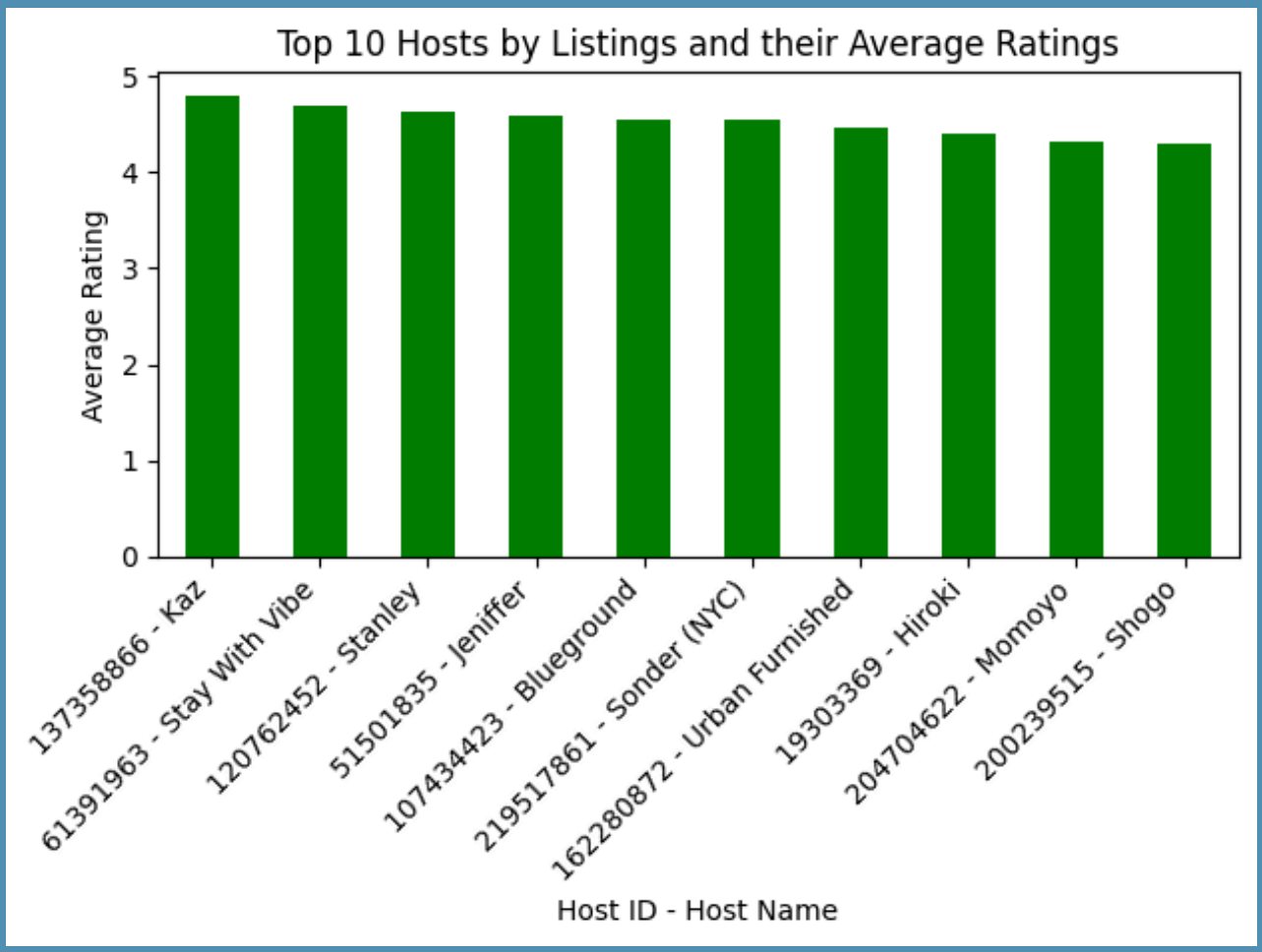
EDA 10: Average Availability by Room Type



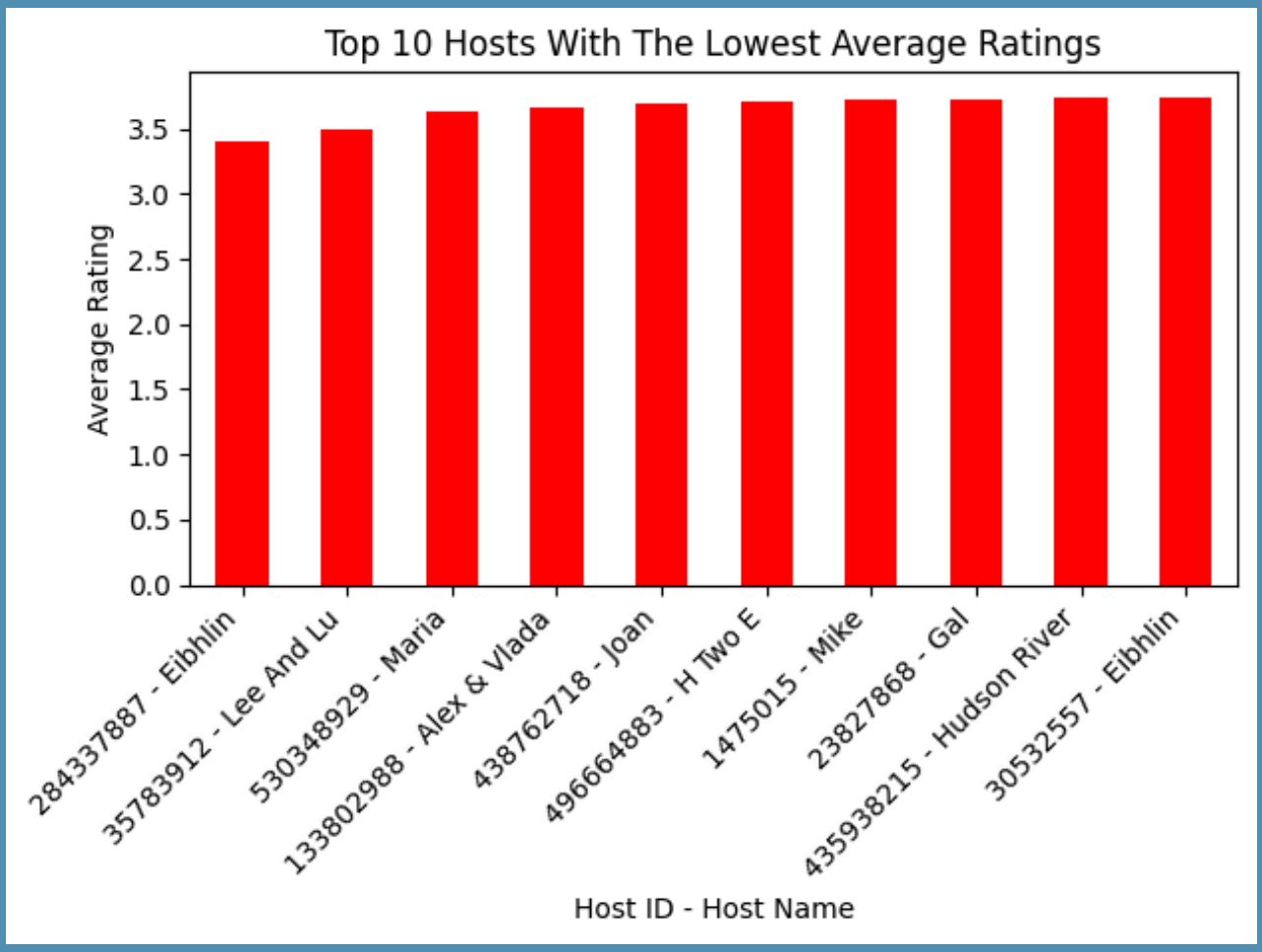
Findings & Insights:

- Most available room types are Hotel rooms with availability of more than 270 days annually.

EDA 11: Top 10 Hosts by Listings and their Average Ratings



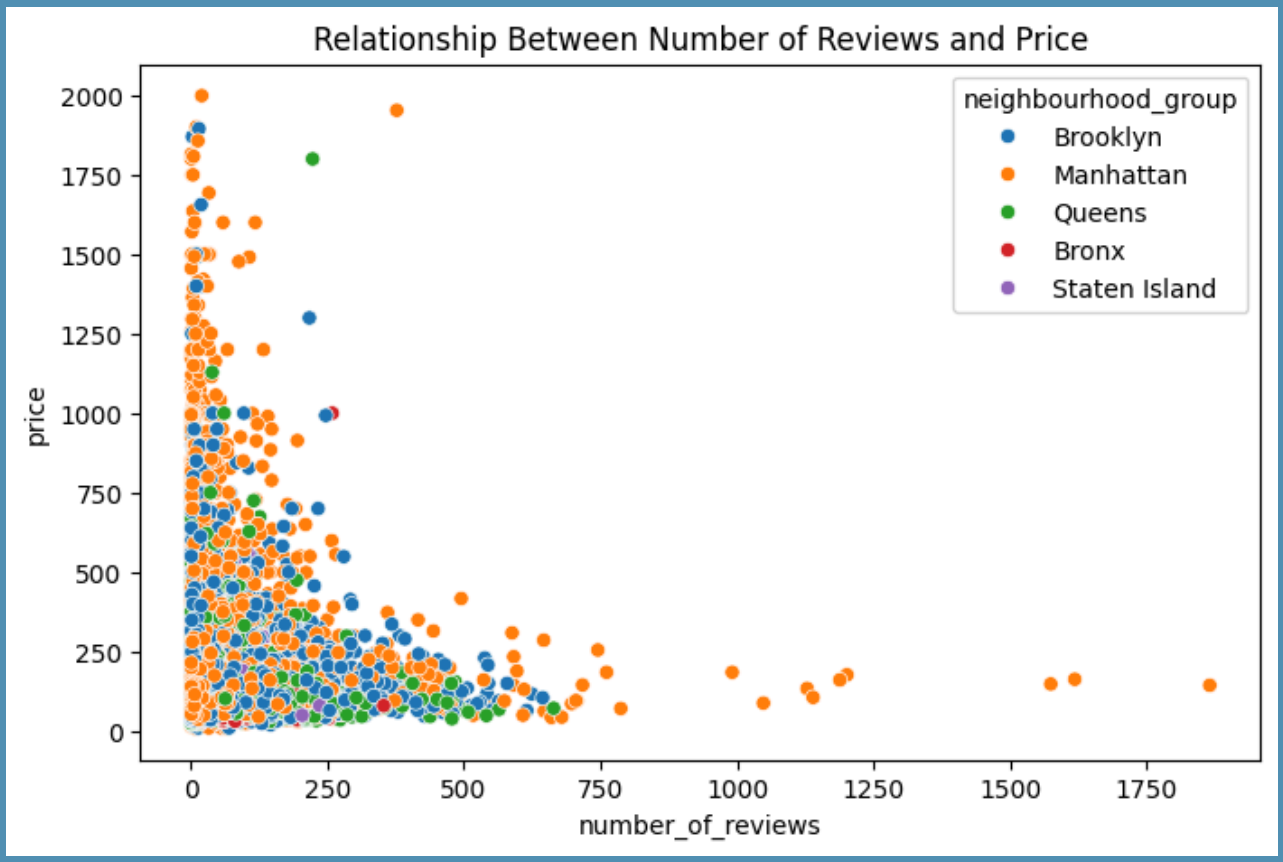
EDA 12: Top 10 Hosts With The Lowest Average Ratings



Findings & Insights:

- Clients may had bad experiances with the host (284337887 - Eibhlin) who got worst average ratings less than 3.5

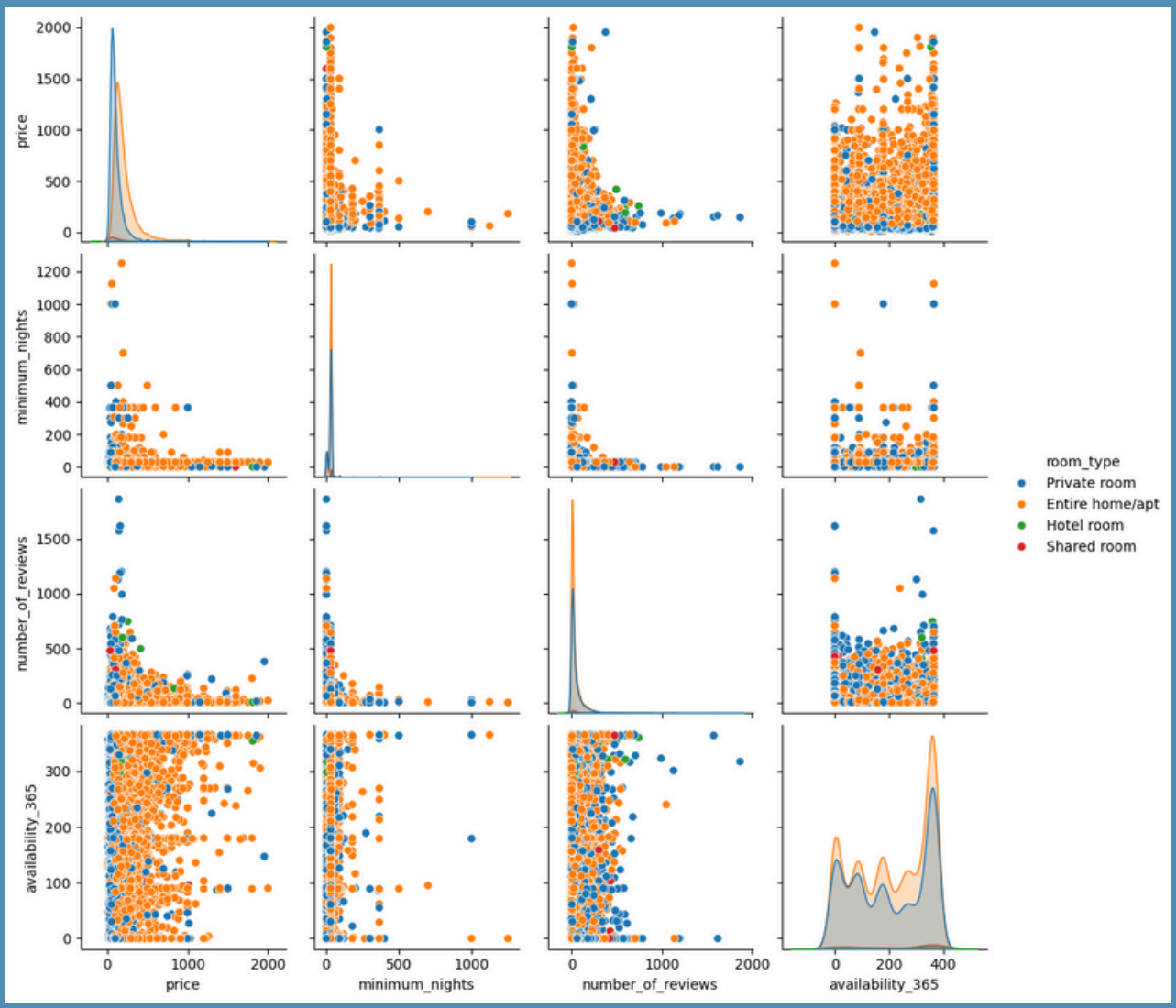
EDA 13: Relationship Between Number of Reviews and Price



Findings & Insights:

- Listings with lower prices tend to have more reviews.
- Expensive listings (>500) generally have fewer reviews, possibly due to lower bookings.

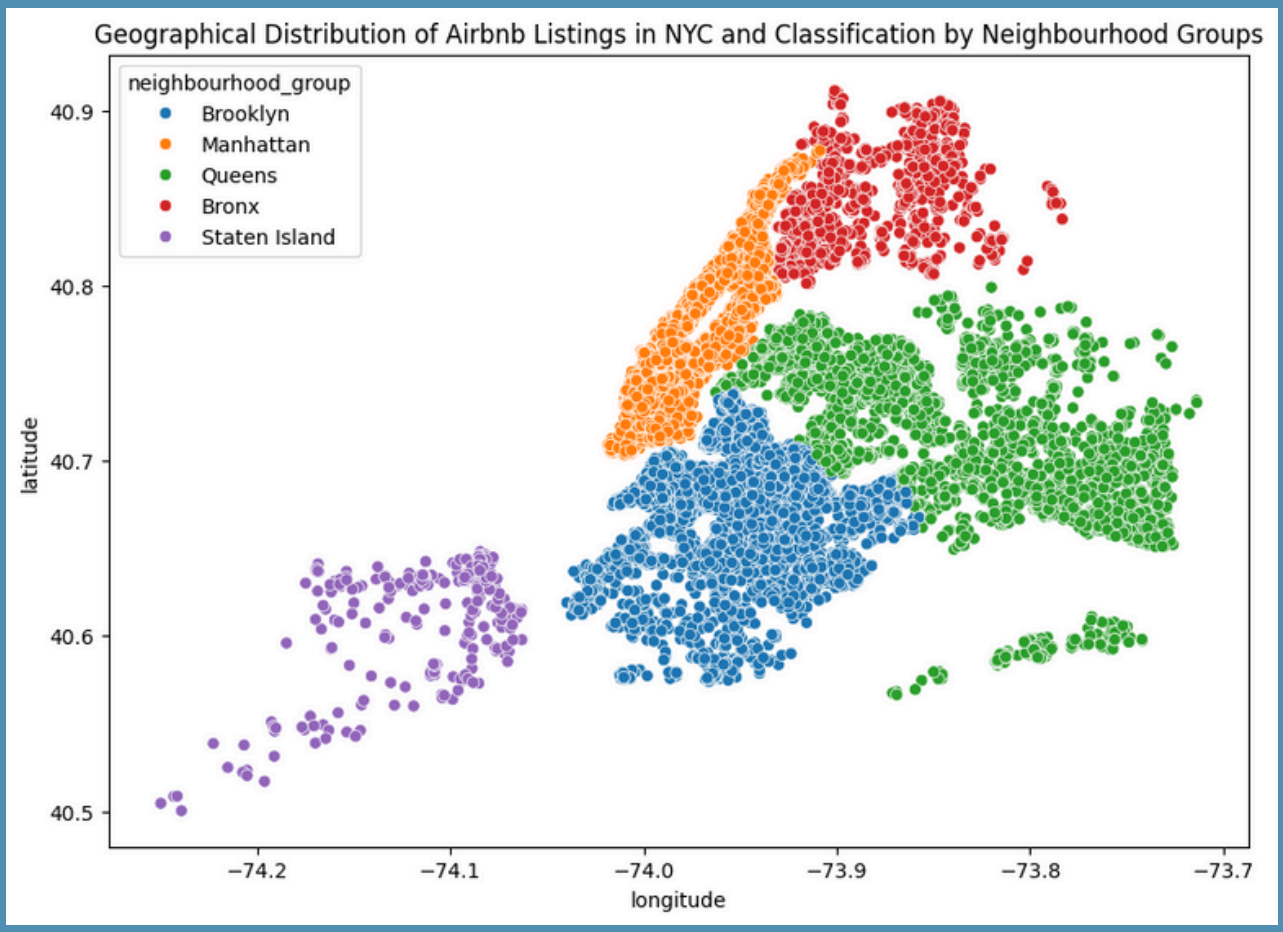
EDA 14: Relationships among different different columns: price, minimum_nights, number_of_reviews, availability_365



Findings & Insights:

- Entire home/apt dominates in both high availability and higher prices.
- Private rooms are more common in low-to-mid price ranges and have relatively high review counts.
- Hotel rooms and Shared rooms are rare and clustered toward the lower price and availability range.
- A noticeable number of listings are set to either 0 or 365 days of availability—likely reflecting management or platform defaults.

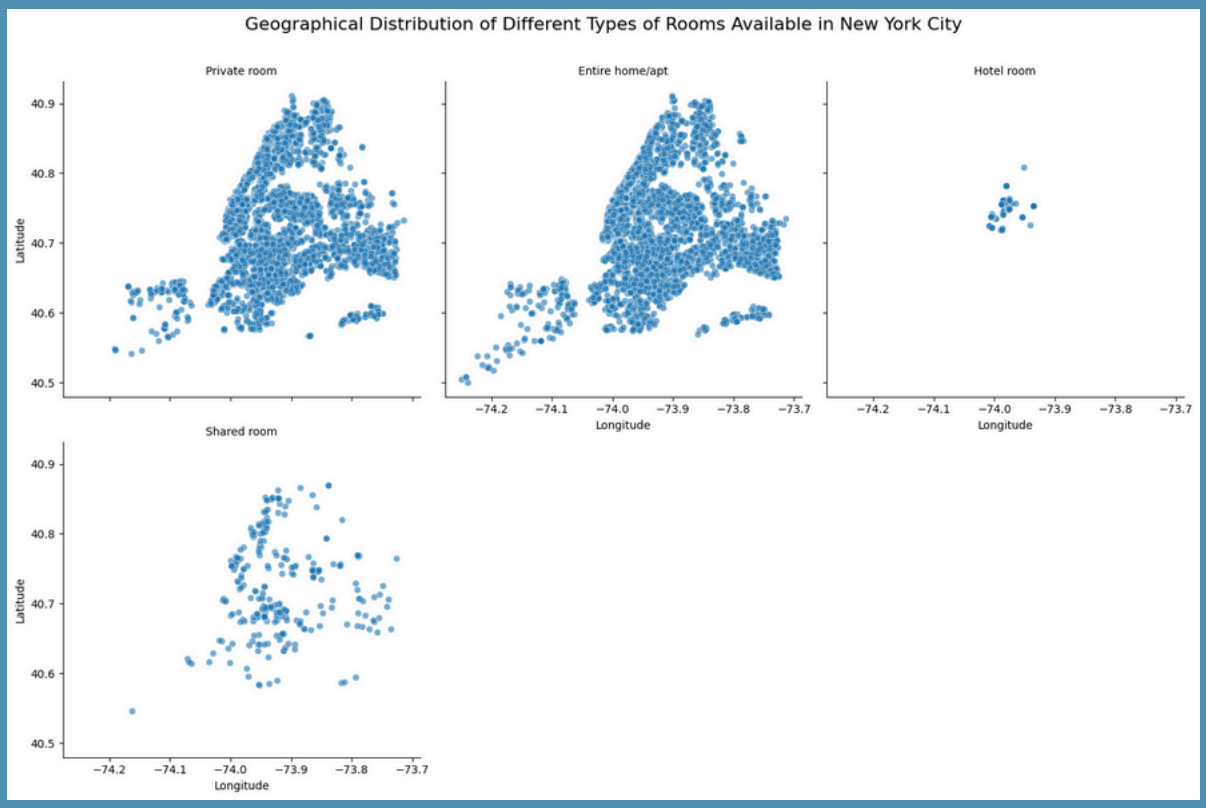
EDA 15: Geographical Distribution of Airbnb Listings in NYC and Classification by Neighbourhood Groups



Findings & Insights:

- The graph visually confirms that Airbnb activity in NYC is highly clustered in Manhattan and Brooklyn.
- Queens and the Bronx have moderate listing densities.
- Staten Island being far and isolated makes it less favorable for tourists.

EDA 16: Geographical Distribution of Different Types of Rooms Available in New York City



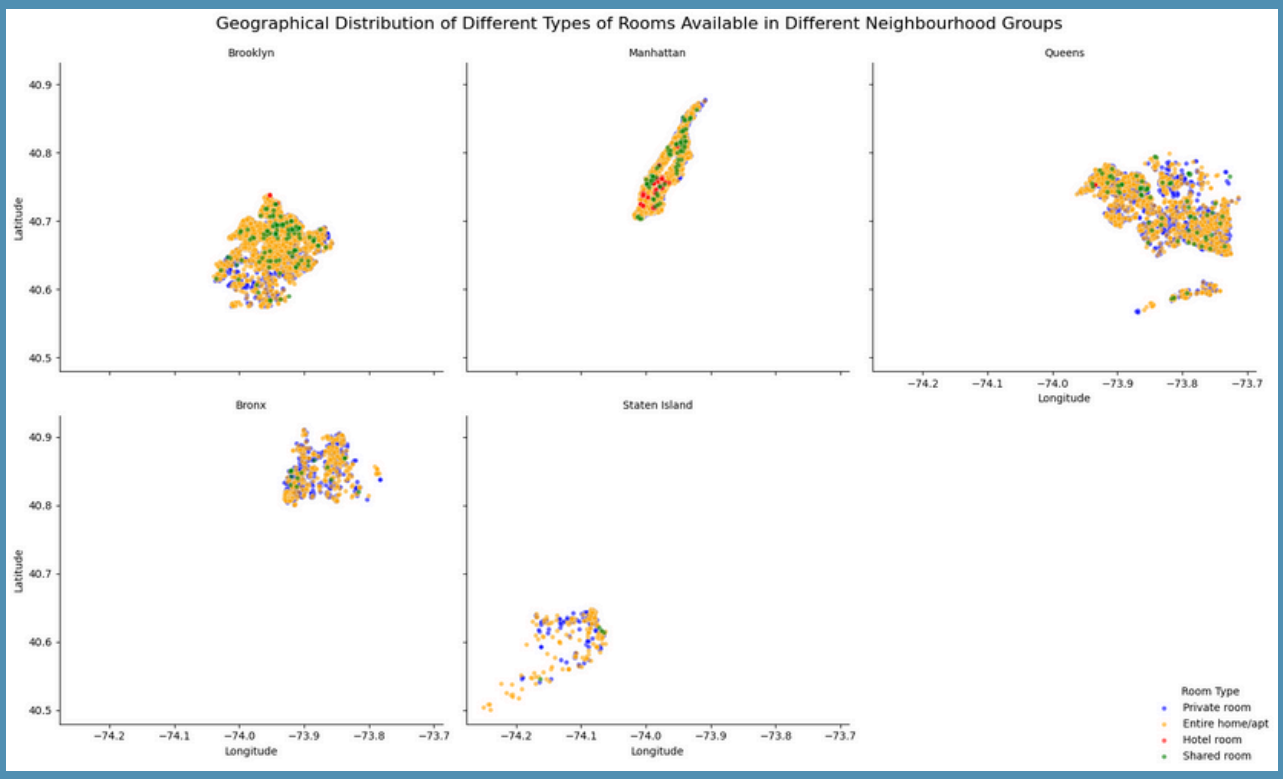
Findings & Insights:

- Private rooms and shared rooms are more common in outer neighbourhood group like Brooklyn and Bronx, likely catering to budget-conscious travelers or long-term stays.
- Shared Rooms are one of least popular room type, which aligns with lower demand for shared accommodations may be due to privacy concerns.
- These clusters can help guide pricing strategies, investment decisions, or infrastructure development for Airbnb hosts and city planners.

Comparative Summary:

Room Type	Spread Across Boroughs	Density	Popularity
Entire Home/Apt	Very widespread	Highest	Very High
Private Room	Widespread	High	High
Shared Room	Limited & scattered	Low	Low
Hotel Room	Very concentrated	Very Low	Niche

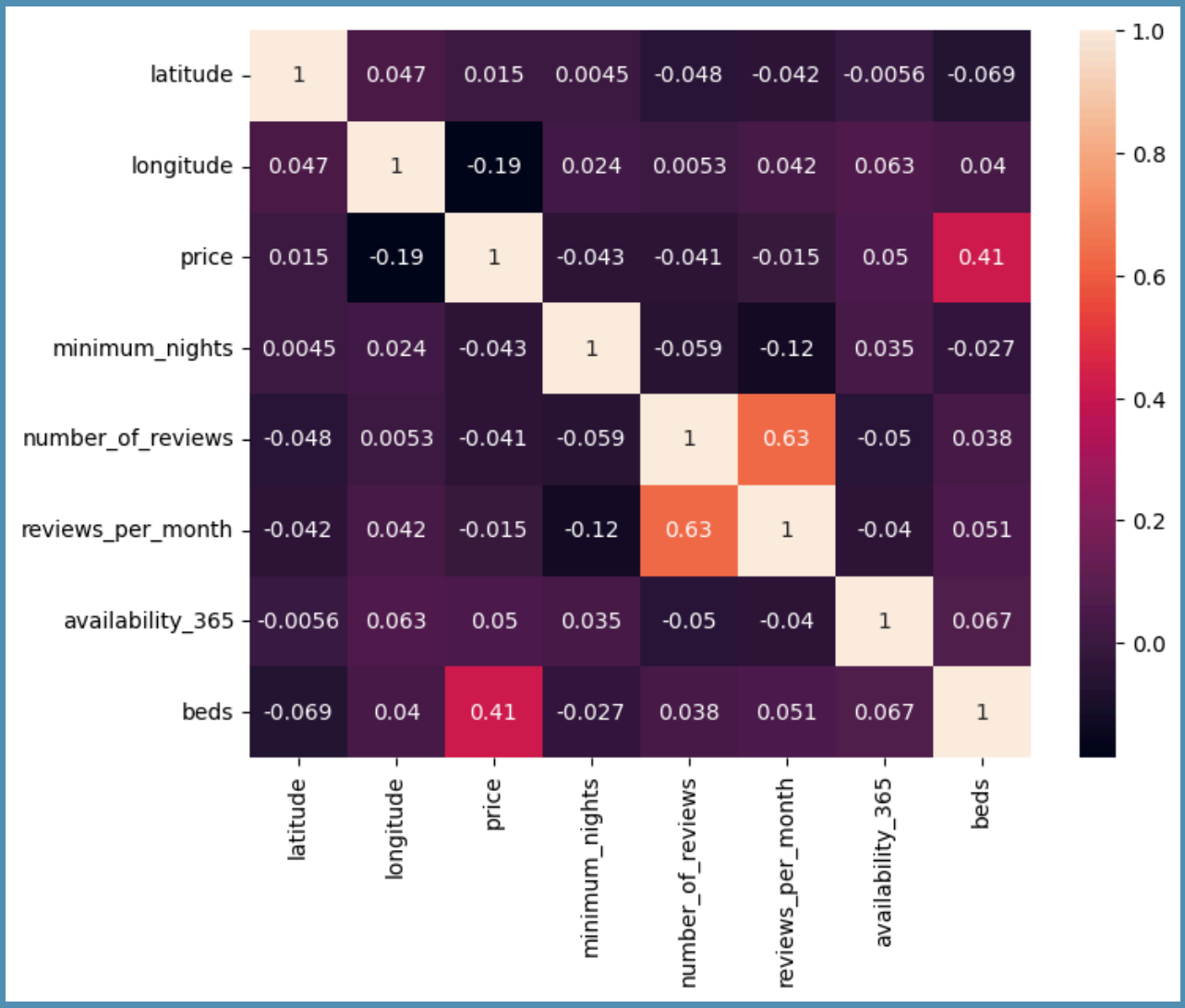
EDA 17: Geographical Distribution of Different Types of Rooms Available in Different Neighbourhood Groups



Findings & Insights:

- Private rooms (blue) and shared rooms (green) are more common in outer boroughs like Brooklyn and Bronx, likely catering to budget-conscious travelers or long-term stays.

EDA 18: Correlation of one variable with others for each numerical column



Findings & Insights:

- Price is moderately influenced by beds and slightly by location (longitude).
- Review activity is a strong indicator of listing popularity but doesn't correlate much with price.
- Geographic coordinates have minimal correlation with other listing attributes except for a slight effect on price.

Types of visualizations showed in the project

- Boxplot for detecting outliers in Price.
- Price and Room Type distribution using Histograms.
- Bar charts used to showcase- Average Price by Neighbourhood Group, Top 10 Expensive Neighbourhoods, Average Price per Bed by Neighbourhood Group, Room Type Distribution (Percentage) by Neighbourhood Group, Price dependency on Room Type by Neighbourhood Group, Top 10 Neighbourhoods by Availability, Average Availability by Room Type, Top 10 Hosts by Listings and their Average Ratings, Top 10 Hosts With The Lowest Average Ratings.
- Showing percentage distribution of different Room Types using Pie Chart.
- Scatter plots for Relationship Between Number of Reviews and Price, and showcasing different Geographical Distributions.
- Used Pairplot to see relationships among different different columns.
- Showing correlations among numerical features using Heatmap.

All Key Findings & Insights

1. Outliers in price that has some listings more than $10000 and also a listing of about $100000, indicating the need to filter such extreme values.
2. Entire home/apt. has got most amount of listings about 55.6%, followed Private room with 42.5% listings. Hotel room & Shared room got least amount of listings respectively 0.5% and 1.4%
3. Manhattan has the most expensive $208 average listing price, followed by Brooklyn.
4. Top expensive locality are Civic Center & Tribeca with $390+ average listings.
5. Most number of Entire home/apt are in Manhattan about 65% of Manhattan's total listings.
6. Bronx and Staten Island have no Hotel room.
7. Expensive Room types by Area(Neighbourhood Group):
   - Private room, Entire home/apt & Hotel rooms are expensive in Manhattan.
   - Shared rooms are expensive in Staten Island about $150 average pricing.
8. Affordable Room types by Area(Neighbourhood Group):
   - Most inexpensive Private rooms are in Bronx about $75 average costing.
   - Affordable Entire home/apt are available in Staten Island with $130 average pricing.
   - Cheap Hotel rooms are available in both Brooklyn and Queens with $160 average pricing.
   - Queens has the most affordable Shared rooms, those average price is about $80 only.

9. Most available room types are Hotel rooms with availability of more than 270 days annually.
10. Clients may had bad experiances with the host (284337887 - Eibhlin) who got worst average ratings less than 3.5
11. Relations between Price and Number of Reviews:

- Listings with lower prices tend to have more reviews.
- Expensive listings (>500) generally have fewer reviews, possibly due to lower bookings.

12. Entire home/apt dominates in both high availability and higher prices.
13. Private rooms are more common in low-to-mid price ranges and have relatively high review counts.
14. Spatial Patterns Reflect Urban and Tourist Hotspots:

- Central locations like Manhattan are likely hotspots for tourists.
- Brooklyn shows significant activity, likely due to its growing popularity and residential offerings.
- Staten Island being far and isolated makes it less favorable for tourists.

15. Private rooms and shared rooms are more common in outer neighbourhood group like Brooklyn and Bronx, likely catering to budget-conscious travelers or long-term stays.
16. Shared Rooms are one of least popular room type, which aligns with lower demand for shared accommodations may be due to privacy concerns.
17. Comparative Summary (Given above)

18. Price is moderately influenced by beds and slightly by location (longitude).
19. Review activity is a strong indicator of listing popularity but doesn't correlate much with price.
20. Geographic coordinates have minimal correlation with other listing attributes except for a slight effect on price.

Recommendations

1. Filter Out Extreme Outliers in Pricing:
   - Listings priced over $10,000 (and up to $100,000) distort analysis and user expectations. Implement a pricing cap or prompt hosts to justify ultra-high prices with premium services and visual proof.
2. Promote Affordable Listings in Outer Neighbourhood Group:
   - Encourage visibility for affordable private rooms in the Bronx and entire homes in Staten Island, which are ideal for budget travelers and longer stays.
3. Re-evaluate Hotel Room Strategies:
   - Hotel rooms are scarce in Bronx and Staten Island but show high availability (>270 days/year). Consider promoting them in underserved areas or explore partnerships to expand inventory.
4. Improve Host Standards for Low-Rated Listings:
   - Flag hosts like ID 284337887 (Eibhlin) for quality reviews or coaching if their ratings consistently fall below 3.5. This can prevent negative guest experiences.

5. Optimize Price Points for Reviews:
   - Since lower-priced listings attract more reviews, suggest optimized pricing for newer hosts aiming to boost visibility and engagement through reviews.
6. Geo-Target Marketing Campaigns:
   - Focus marketing and dynamic pricing in tourist-heavy zones like Manhattan and Brooklyn, while targeting Bronx, Staten Island and Queens for residential or extended-stay travelers.
7. Diversify Room Types in High-Demand Areas:
   - In boroughs like Brooklyn, promote underrepresented room types for example shared or hotel rooms to cater to niche or overflow markets.
8. Use Availability Data to Forecast Revenue:
   - Since entire homes dominate high-price and high-availability categories, develop tools that forecast revenue based on room type, borough, and availability trends.
9. Add Location-Based Pricing Guidance for Hosts:
   - Provide dynamic suggestions on how to price listings based on longitude/latitude and neighborhood trends to avoid under- or overpricing.
10. Tailor Listings for Target Audiences:
   - Suggest that listings in outer Neighbourhood Groups (like Bronx, Queens, Staten Island) cater to long-term stays, digital nomads, or student rentals, while Manhattan listings can be tailored for tourists and short-term visitors.

Future Work

1. Predictive Pricing Model:
   - Machine learning model can be built to predict optimal listing prices based on features like location, room type, availability, number of reviews, and host ratings. This can help new hosts price competitively.
2. Location Clustering for Investment Opportunities:
   - Clustering algorithms on latitude and longitude can be used to identify high-demand and underpriced zones. This can guide Airbnb investors or hosts on where to open new listings.
3. Time Series Analysis of Availability Trends:
   - Analysis can be performed on seasonal trends in listing availability and bookings (if date-based data is added) to help hosts optimize pricing and availability settings throughout the year.
4. Sentiment Analysis on Guest Reviews:
   - Addition of natural language processing (NLP) to extract themes and sentiment from guest reviews. This would help identify what guests like/dislike about certain room types or neighborhoods.
5. Dashboard Development for Real-Time Insights:
   - Develop an interactive dashboard using tools like Plotly Dash, Streamlit, or Tableau to explore listing attributes, price trends, and neighbourhood group-wise comparisons in real-time for both hosts and platform managers.

Conclusion

This EDA and visualization project provided valuable insights into the structure and trends of Airbnb listings in New York City. It revealed key pricing patterns, host behaviors and geographical trends that are relevant for both travelers and business analysts. The project effectively demonstrates how data science techniques can be applied to real-world datasets for exploration, business recommendations & decision-makings. Future enhancements could include leveraging advanced analytics and predictive modeling to deepen insights and improve decision-making.