OPTIMIZING AIRBNB PRICING STRATEGY IN NEW YORK CITY FOR HOSTS

Subject: DATA INTENSIVE COMPUTING (CSE 587B)

OCTOBER 2023

Naga Venkata Sahithya Alla, MS

alla16@buffalo.edu

Marziye Kouroshli, MS

marziyek@buffalo.edu

University at Buffalo
Graduate School of Education

1. Problem Statement:

The objective of this project is to analyze the NYC Airbnb dataset to explore patterns, correlations, and key factors influencing prices, occupancy rates, and customer satisfaction. By understanding these dynamics, our aim is to provide actionable insights to hosts, empowering them to optimize their listings price.

1.1 Background:

Airbnb has revolutionized the hospitality industry by allowing individuals to rent out their properties to travelers. In New York City, Airbnb is a popular choice for both hosts and guests due to its wide range of options. However, hosts often face challenges in setting the right price for their listings. Some key points to consider are:

- NYC is a top destination for travelers, attracting millions of tourists annually.
- The competition among Airbnb hosts is intense, making pricing a critical factor for attracting guests.

1.2 Significance:

Optimizing Airbnb pricing in NYC is significant for several reasons –

- NYC hosts can significantly increase their income by optimizing the pricing, contributing to their financial well-being.
- Guest Satisfaction: Accurate pricing ensures guests get the best value for their stay, leading to positive reviews and customer loyalty.
- A more efficient pricing strategy benefits both hosts and guests, creating a win-win scenario
- Accurate pricing helps hosts remain competitive in the market, attracting more bookings.

1.3 Potential Contribution:

This project has the potential to contribute significantly to the problem domain:

1. Data-driven insights: By analyzing historical Airbnb data, we can provide hosts with data-driven insights into pricing trends, demand patterns, and seasonal variations.
2. Pricing recommendations: We can develop pricing models that suggest optimal rates based on various factors, such as location, property type, and time of the year.
3. Host Success: Beyond pricing, we can offer recommendations to hosts to enhance their listings descriptions, amenities, and guest communication to improve their overall performance.
4. The project can serve as a valuable market research, helping Airbnb hosts understand their competitive landscape better.

2. Data:

2.1 Acquiring Data:

The data is sourced from Kaggle website. This data consists of 48895 rows and 16 columns. It consists of information related to geographical availability, hosts, neighbourhood groups, room types, and relevant metrics to draw meaningful conclusions and make predictions.

Link to dataset: https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data

Description of columns: (From Kaggle – Data source)

1. id: Unique ID of the Airbnb listing

2. name: Name of the listed property.

3. host_id: Unique ID of the host.

4. host_name: Name of the host.

5. neighbourhood_group: Region of the location - neighborhood (e.g., Brooklyn, Manhattan, Queens, Staten Island, and Bronx).

6. neighbourhood: Specific neighborhood where the property is located.

7. latitude: Latitude coordinate of the property.

8. longitude: Longitude coordinate of the property.

9. room_type: Type of room (e.g., Private room, Shared room, Entire home/apt).

10. price: Price per night (in USD).

11. minimum_nights: Minimum number of nights guests are required to stay.

12. number_of_reviews: Number of reviews the listing has received.

13. last_review: Date of the last review.

14. reviews_per_month: Average number of reviews per month.

15. calculated_host_listings_count: Total number of properties listed by the host.

16. availability_365: Number of days the property is available in a year.

We have introduced little noise into data by adding duplicate rows, punctuations to text data in neighbourhood column. After adding noise, the dataset is now 48901 rows and 16 columns.

2.2 Data Cleaning:

The dataset has undergone initial cleaning and processing with the below mentioned steps, and within this framework, additional steps have been implemented to further refine and cleanse the data.

2.2.1 Handling Null Values:

In the dataset, the columns – name, host_name, last_review, reviews_per_month.

- As it is difficult to handle missing values of name and host name without further knowledge on domain (especially given their free-form text nature), we are replacing them with 'Not Known' label.
- The null values in reviews_per_month are replaced with 0 rather than using the mean/ median, as reviews are typically influenced by a range of factors like location, ambiance, price, cleanliness of the rented room, etc.,
- The last_review column is dropped as there are so many null values and also it does not contribute any meaningful impact or value to our analysis.

2.2.2 Handling Duplicates:

Addressing duplicate values is a crucial step to ensure accuracy and reliability of our dataset. Duplicates refer to rows or observations that are identical in all their attribute values, which can skew our analysis and lead to inaccurate results. Hence, we dropped if there are duplicate records in the dataframe.

2.2.3 Text Data Cleaning:

- The name and neighbourhood_group column are cleaned using the following 5 steps:

    - Lowercasing words
    - Removing stop words
    - Removing Punctuations
    - Removing other characters that are not relevant
    - Removing extra spaces

  Note: neighbourhood group only benefits from lowercasing and removing punctuations from the above steps.
- The columns such as host_name, neighbourhood are transformed to lower case.
- The data in the room_type column has been converted to lowercase, and any instances of '/' have been replaced with 'or' to establish a consistent format within that column.

2.2.4 Handling inappropriate data:

- It is observed that certain rows within the dataset exhibit a price value of 0. However, it is essential to note that a price of 0 is anomalous in the context of property rentals, as it contradicts the fundamental concept of renting or listing a property with a non-zero cost.
- Upon further examination, we have identified around 11 such records. In order to uphold the quality and integrity of our dataset, we have removed these specific rows with a price value of 0. This step ensures that our dataset accurately reflects the expected pricing structure for rental properties.

2.2.5 Handling Outliers:

- The price column initially exhibits a high degree of skewness, with a skewness value of approximately 19. This level of skewness indicates a significant departure from a normal distribution. As confirmed by the distribution plot (distplot), the data is indeed skewed. To address this issue and promote a more balanced distribution for our data, we have chosen to employ the Interquartile Range (IQR) method to identify and subsequently remove outliers within the price column. This outlier removal process is essential to ensure that our model is not influenced or biased by extreme values, ultimately contributing to more robust and reliable model predictions in the future.
- The price column holds significant importance in our analysis, and for this reason, we have decided to completely remove outliers rather than replacing them with upper or lower bounds or mean/median. This approach aims to maintain the integrity of original data and prevent any distortion in the pricing information. We are striving to construct a robust model that delivers accurate and dependable results, unaffected by the extreme values, and enhance the model’s ability to generalize and make predictions effectively while preserving the critical pricing data’s fidelity.
- Similarly, removing outliers in the numerical columns 'minimum_nights', 'number_of_reviews', 'availability_365' as well.

2.2.6 Converting Categorical values into numerical:

We have applied label encoding to the ‘neighbourhood_group’ and ‘room_type’ columns to convert them into numerical data. However, we have chosen not to convert all categorical columns into numerical value at this stage. This decision is based on the interdependencies between some columns, such as ‘neighbourhood’ relying on the ‘neighbourhood_group’ column. Additionally, columns like ‘name’ and ‘host_name’ primarily consist of unique values. We plan to leverage these specific columns strategically when constructing our model.

2.2.7 Normalizing data:

- We have normalized the features of the dataset by using Min-Max Scaling. This method scales data to a range between 0 and 1.
- The formula is X_normalized = (X-X_min) / (X_max – X_min), where X is the original value, X_min is the minimum value of the feature, and X_max is the maximum value of the feature.
- The main goal is to ensure that different features have similar scales which will be beneficial in our further model training/ ML algorithm.

3. Exploratory Data Analysis (EDA):

To achieve deep understanding of data, we engage in exploratory data analysis through visualizations which will enable us to discover patterns. These insights help us to make informed decisions in our further analysis, which is crucial for effective analysis and decision-making.

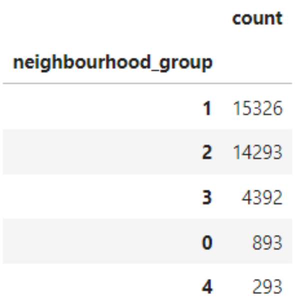
As we have transformed our categorical variables into numerical values before, here is the mapping to facilitate a better understanding of the upcoming plots and analysis:

Neighbourhood group:
Bronx: 0,
Brooklyn: 1,
Manhattan: 2,
Queens: 3,
Staten Island: 4

Room type:
Entire home or apt: 0,
Private room: 1
Shared room: 2

3.1 Analysis on neighbourhood group:

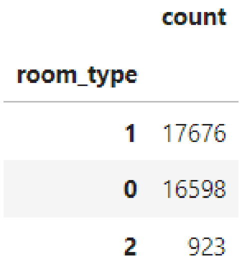
The below data reveals that 1 – Brooklyn has the highest number of property listings, and then we have 2 – Manhattan in the second position with the next highest.



	count
neighbourhood_group	
1	15326
2	14293
3	4392
0	893
4	293

3.2 Analysis on Room types:

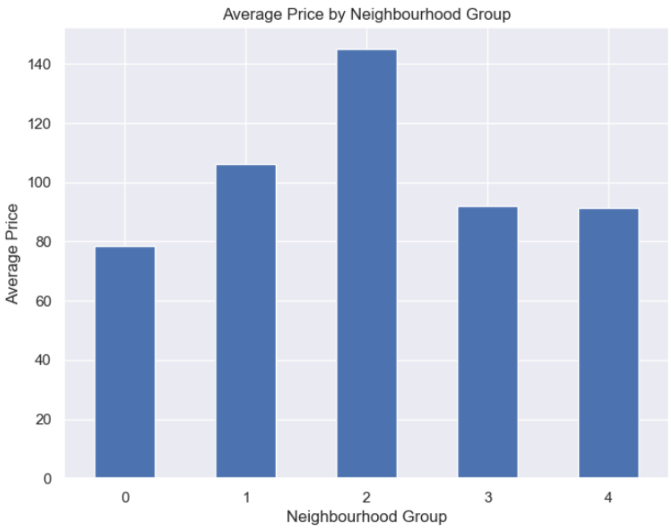
We can see that number of private rooms are highest, followed by Entire Home or apt.



	count
room_type	
1	17676
0	16598
2	923

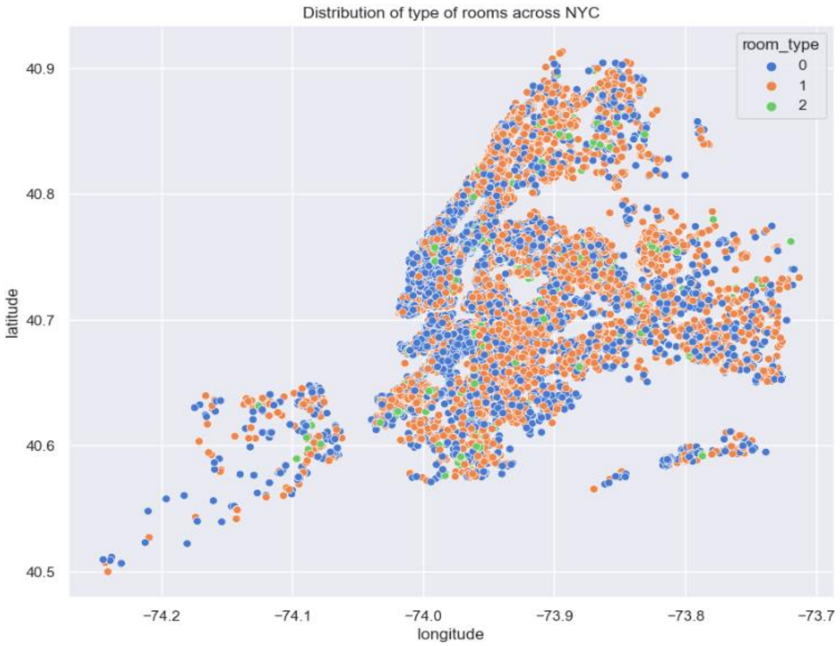
3.3 Analysis on neighbourhood group and price:

From the plot - average price by neighbourhood group, we can observe that 2 - Manhattan is the costly place among others, with average of around 140+ USD which is followed by Brooklyn with around 80 USD price on average. Bronx is the cheapest place with least average price.



3.4 Distribution of room types across NYC:

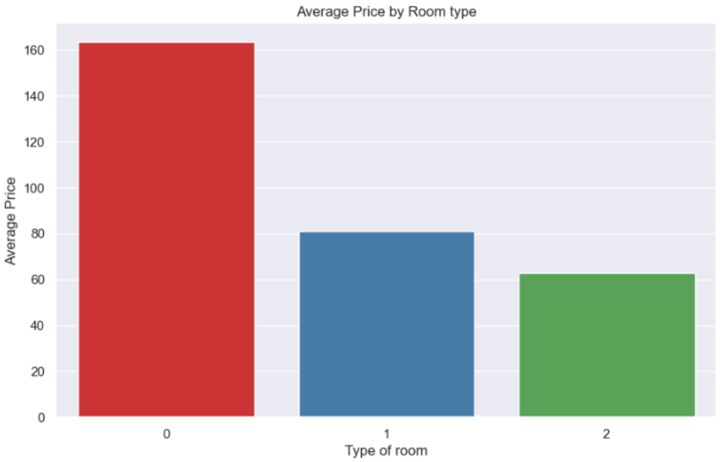
We can observe from the below plot that 2 - Shared rooms are less throughout NYC compared to 1 -Private room or 0 – Entire Home or Apt. And it makes sense as well, as many people prefer to stay in private room or rent out whole apt instead of sharing room with others.



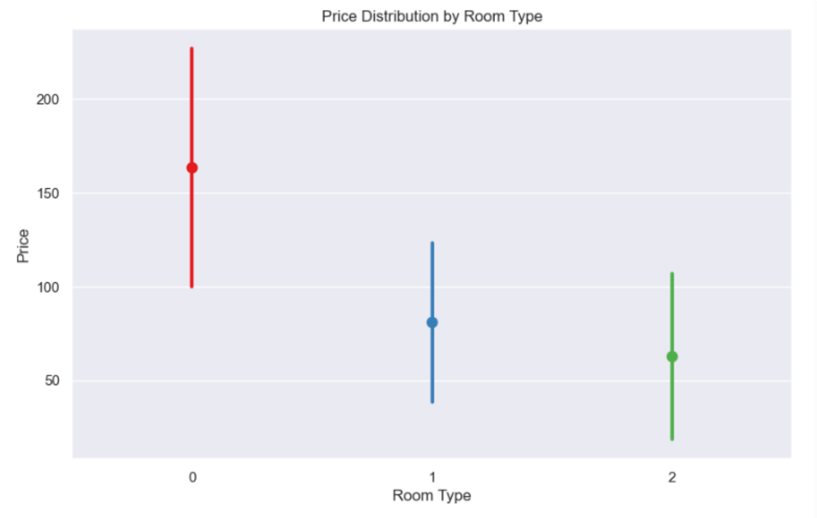
Considering both room type and neighbourhood group can be a useful approach for determining or predicting the price in our further analysis.

3.5 Analysis on room type and price:

We can observe from the above plot, that Entire home or Apt has the highest average price with around 160+ USD, followed by Private room and least being shared room (as expected).

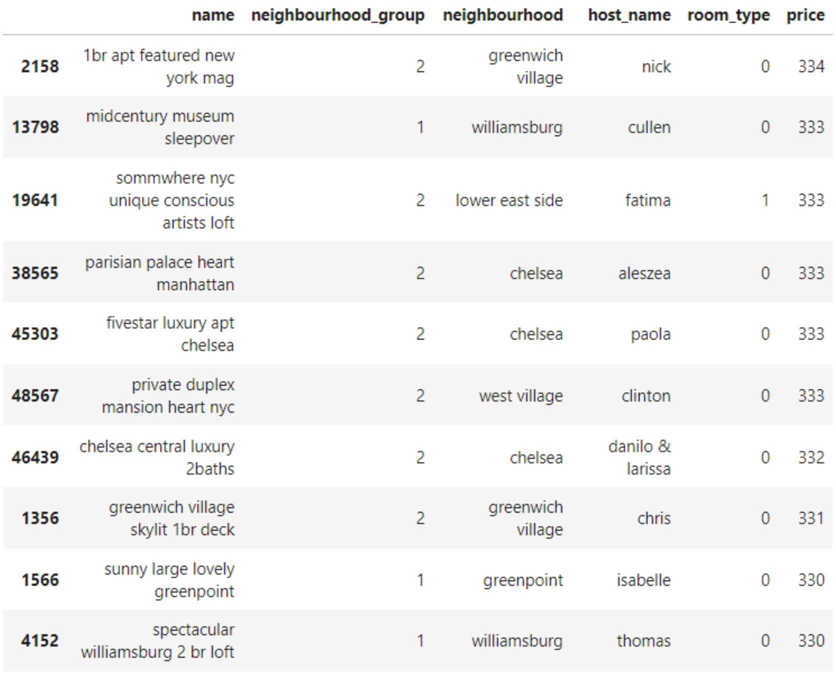


The point ploy displayed below shows the mean price for each room type along with confidence intervals allowing us to compare the tendency of the prices.



3.6 Top 10 costly and top 10 cheap listings:

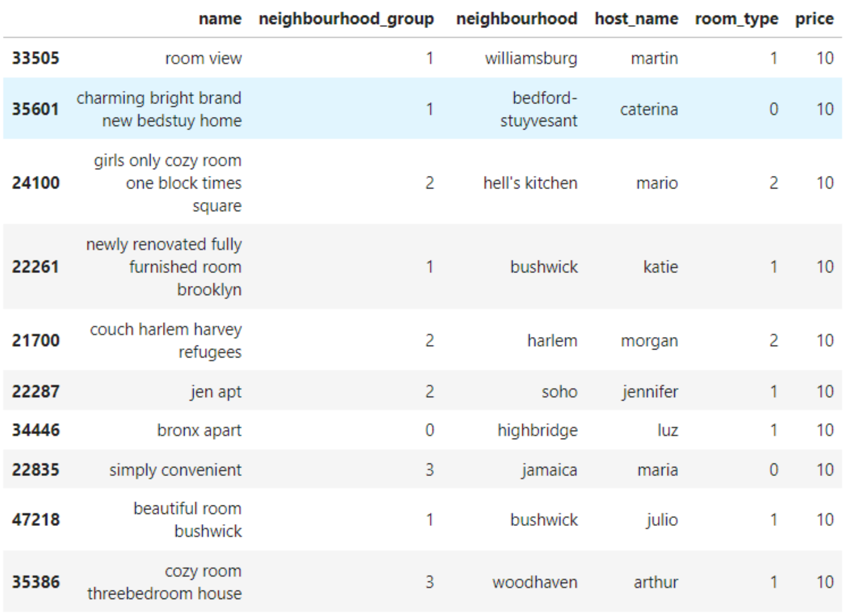
Top 10 costly listings:



	name	neighbourhood_group	neighbourhood	host_name	room_type	price
2158	1br apt featured new york mag	2	greenwich village	nick	0	334
13798	midcentury museum sleepover	1	williamsburg	cullen	0	333
19641	sommwhere nyc unique conscious artists loft	2	lower east side	fatima	1	333
38565	parisian palace heart manhattan	2	chelsea	aleszea	0	333
45303	fivestar luxury apt chelsea	2	chelsea	paola	0	333
48567	private duplex mansion heart nyc	2	west village	clinton	0	333
46439	chelsea central luxury 2baths	2	chelsea	danilo & larissa	0	332
1356	greenwich village skylit 1br deck	2	greenwich village	chris	0	331
1566	sunny large lovely greenpoint	1	greenpoint	isabelle	0	330
4152	spectacular williamsburg 2 br loft	1	williamsburg	thomas	0	330

We can observe that most of them are from 2 – Manhattan and a very few from 1 – Brooklyn. It is evident that Entire Home/ Apt room type are on top in this analysis as well.

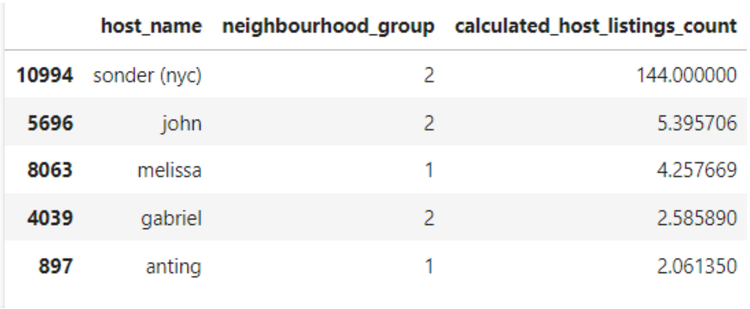
Top 10 cheap listings:



	name	neighbourhood_group	neighbourhood	host_name	room_type	price
33505	room view	1	williamsburg	martin	1	10
35601	charming bright brand new bedstuy home	1	bedford-stuyvesant	caterina	0	10
24100	girls only cozy room one block times square	2	hell's kitchen	mario	2	10
22261	newly renovated fully furnished room brooklyn	1	bushwick	katie	1	10
21700	couch harlem harvey refugees	2	harlem	morgan	2	10
22287	jen apt	2	soho	jennifer	1	10
34446	bronx apart	0	highbridge	luz	1	10
22835	simply convenient	3	jamaica	maria	0	10
47218	beautiful room bushwick	1	bushwick	julio	1	10
35386	cozy room threebedroom house	3	woodhaven	arthur	1	10

We can observe that most of the cheap listings are from 1 – Brooklyn and a few from 2 – Manhattan and Queens. It is interesting that we can find all three types of room types in the cheapest listings.

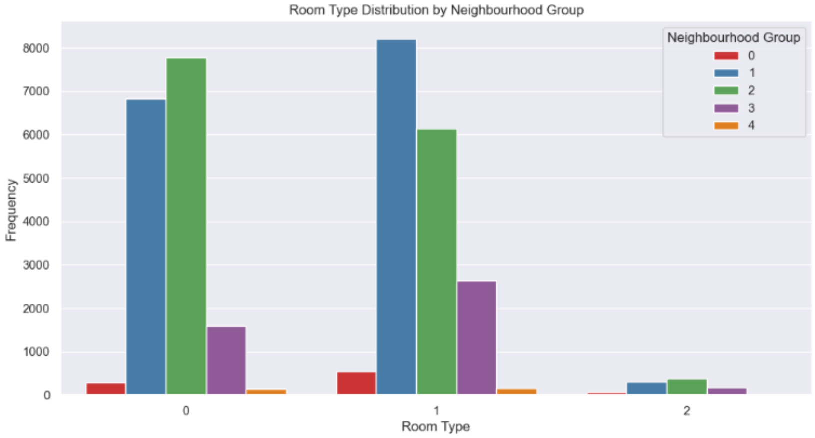
3.7 Host with highest listings:



	host_name	neighbourhood_group	calculated_host_listings_count
10994	sonder (nyc)	2	144.000000
5696	john	2	5.395706
8063	melissa	1	4.257669
4039	gabriel	2	2.585890
897	anting	1	2.061350

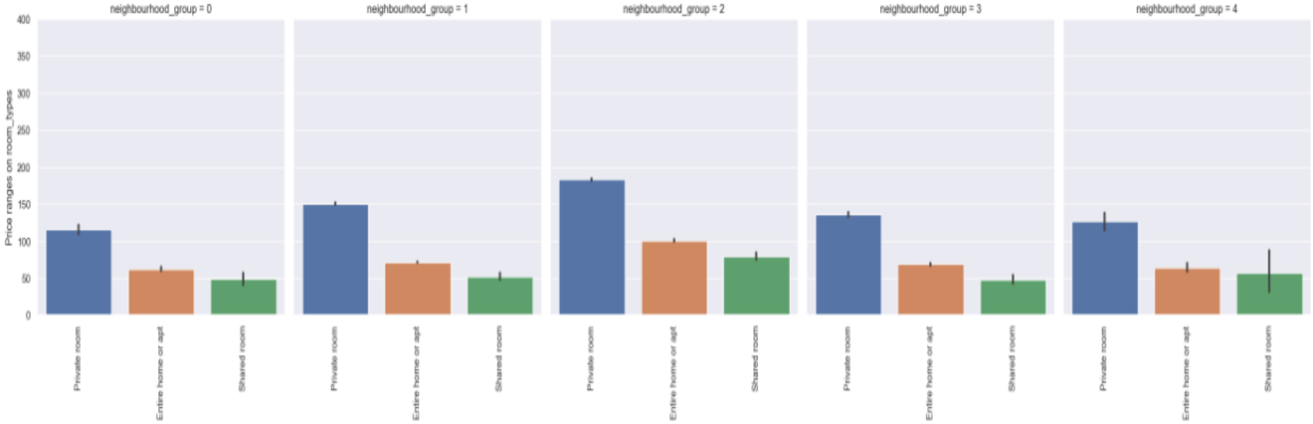
From the above results, it is evident that Sonder (NYC) owns highest number of listings and interestingly, most of them are situated in Manhattan, which is also the highest-priced location we have identified.

3.8 Analyzing Room type by Neighborhood groups:



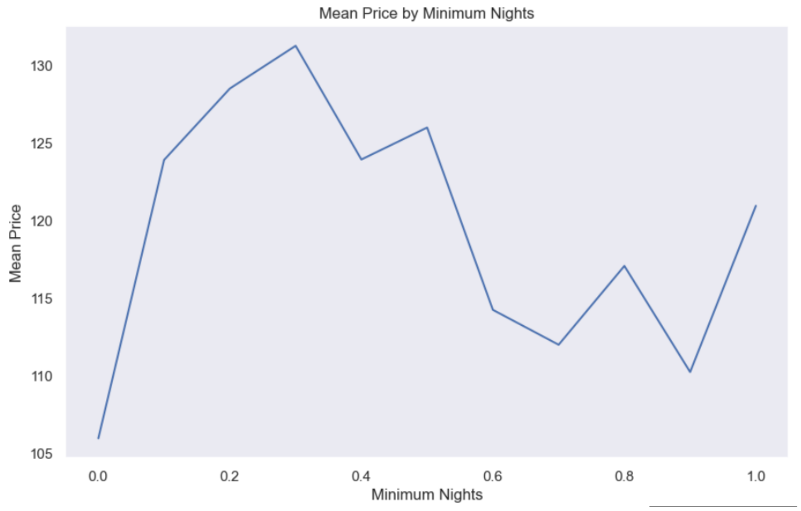
The generated count plot reveals that Entire home or apt and Private room are most common room types in NYC Airbnb listings, with Shared room being less common. Additionally, the distribution of room types differs among neighbourhood groups, highlighting variations in accommodation preferences across different areas.

3.9 Room type vs Price ranges w.r.t Neighbourhood groups:



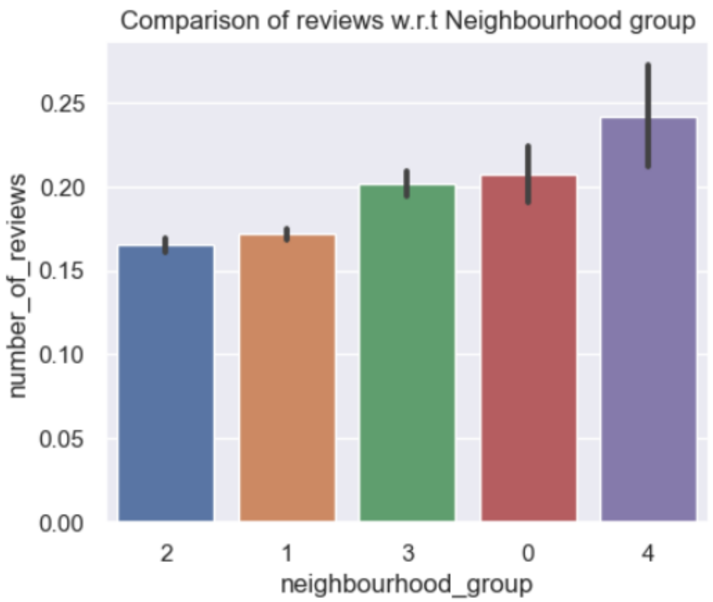
The plot creates a set of bar plots as shown, one for each neighbourhood group to visualize how the average prices vary for different room types and the generated plot suggests that Private room type is the most prevalent among NYC listings. Notably, Private room accommodations command the highest prices.

3.10 Minimum nights vs Price:



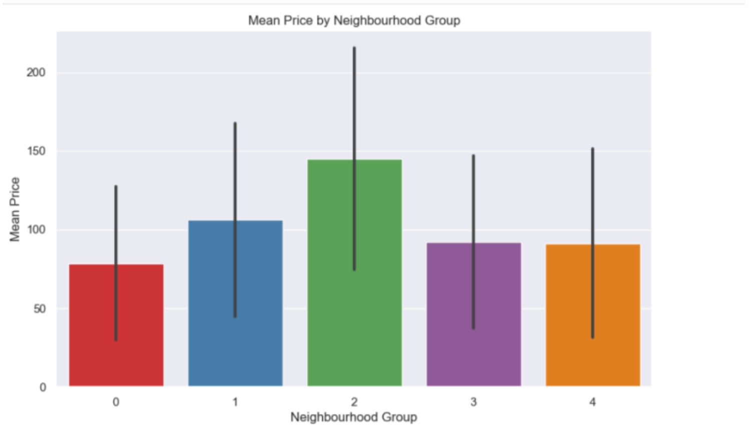
The analysis revealed a notable pattern in the dataset. As the minimum nights required for a booking increased, there was a clear trend of decreasing prices up to a certain threshold. Beyond that threshold, pricing showed occasional fluctuations. This suggests that hosts may offer discounts for longer stays, attracting guests who are willing to commit to a minimum number of nights. The fluctuations in pricing for different minimum night requirements could be influenced by factors such as seasonality, special events, or host pricing strategies. We can investigate further based on location and room type to better understand and optimize pricing and minimum night policy.

3.11 Comparison of reviews w.r.t Neighbourhood groups:



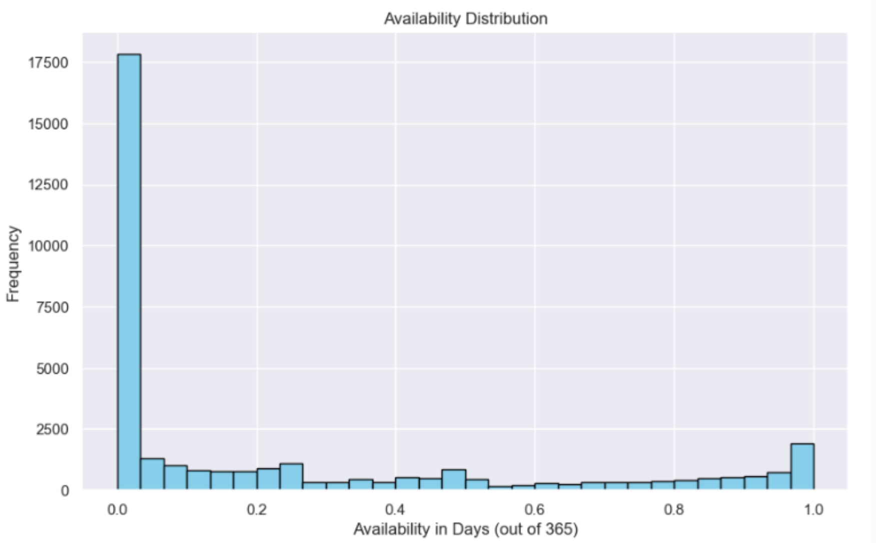
We can observe that, 4 - Staten Island neighbourhood group got the most reviews compared to others.

3.12 Mean Plot: Mean Price by Neighborhood groups:



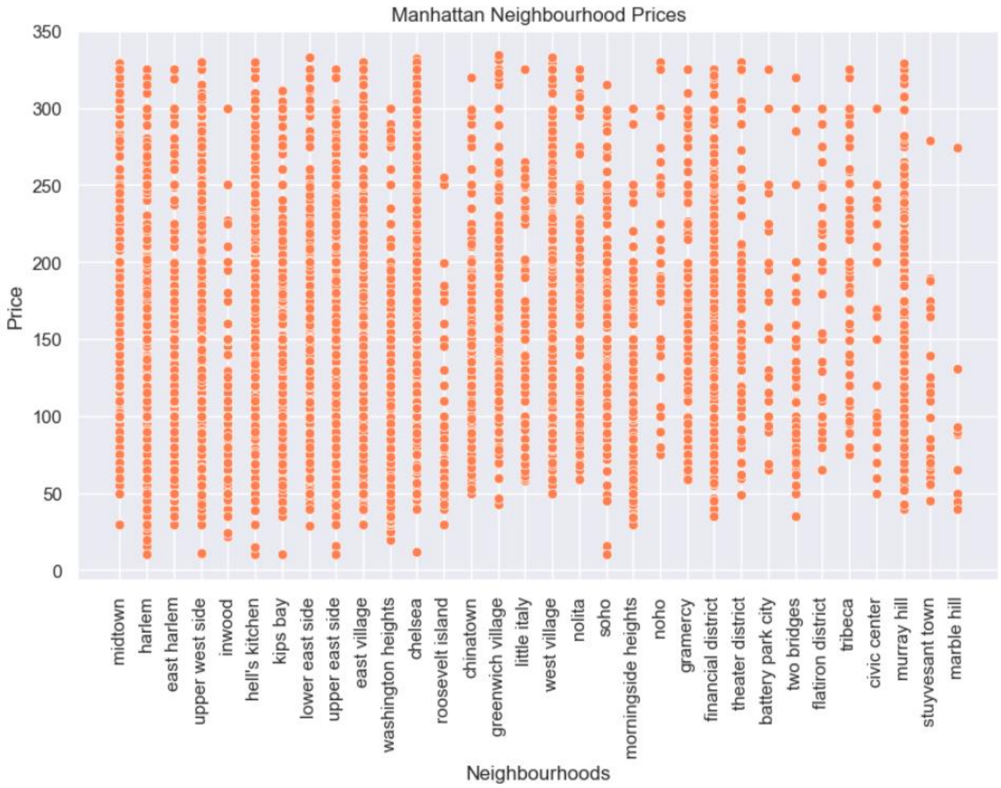
In the given mean plot, we can quickly see which neighbourhood groups tend to have higher and lower average prices, which is 2 – Manhattan with highest price and 0 – Bronx with lowest price.

3.13 Properties Availability Distribution:



From the availability distribution histogram, we can observe how the availability of listings is distributed throughout the year, identifying trends in booking patterns.

3.14 Analysis of Manhattan Neighbourhood prices:



From the plot, it appears that there is a relatively consistent distribution of prices across the neighbourhoods in Manhattan. This means that, on average, property prices do not vary significantly from one neighborhood to another within Manhattan.

References:

1. 1.1.1. What is EDA? (nist.gov)
2. Examples — Matplotlib 3.8.0 documentation
3. https://seaborn.pydata.org/
4. Find Open Datasets and Machine Learning Projects | Kaggle