NY' Airbnb Data Analysis

Lee Sunhwan

2024.12.16

Business Analytics

Motivation & Background

With the rise of the sharing economy, hospitality-sharing platforms like Airbnb have brought significant innovations to the short-term rental market. This transformation is particularly evident in New York City, where boroughs exhibit distinct profiles. For instance, Manhattan, despite not being the most populous (population by borough as of recent data: Bronx ~1,356,476, Brooklyn ~2,561,225, Manhattan ~1,597,451, Queens ~2,252,196, Staten Island ~490,687), stands out for its exceptionally high housing costs. Manhattan's home prices rank as the highest in the United States, and its cost of living surpasses the national average by approximately 122%. These factors contribute to greater price volatility within the borough's Airbnb market, and certain lodging types—especially private rooms—have an outsized influence on pricing structures.

Such a complex environment—where regional characteristics, market dynamics, and regulatory frameworks interact—requires a nuanced understanding. By examining factors such as room type, minimum nights, and neighbourhood group, we can generate valuable insights for hosts, platform operators, and tourism stakeholders. In this project, we go beyond simple averages, separating Manhattan and non-Manhattan data to uncover how local specificities, regulation, and tourism demand patterns shape price formation and volatility. This approach helps hosts develop more effective pricing strategies, policymakers and platform operators devise targeted regulatory measures, and ultimately, leads to better-informed decisions that enhance user experiences.

Through this comprehensive analysis, we aim not just to work with numerical data or predictive models, but to understand how the Airbnb market truly operates within its socio-economic and policy contexts. By integrating these findings into strategic planning and decision-making processes, we can contribute to more sustainable management of short-term rentals in economically and demographically diverse urban landscapes like New York City.

Business and Analytics Goals of Project

1. **Analysis of Regional Accommodation Prices and Variability**
   Compare accommodation prices across different regions in New York City (Manhattan, Brooklyn, Queens, Bronx, Staten Island) and analyze the factors contributing to price variability. This includes identifying why high-cost areas like Manhattan exhibit greater price volatility and uncovering the relationship between regional characteristics, regulations, and tourism demand.
2. **Significance of Room Types in Price Formation**
   Analyze the impact of room types (Entire home/apt, Private room, Shared room) on price formation, particularly focusing on how private rooms influence pricing in high-cost areas. This provides insights into optimal pricing strategies for different regions and room types.
3. **Host Strategies and Operational Patterns**
   Examine host-specific operational patterns, such as the number of listings, minimum stay requirements, and review count, to understand the differences between multi-listing hosts and single-listing hosts. This analysis identifies which host strategies generate higher revenues and attract more bookings.
4. **Feature Importance Analysis**
   Identify which independent variables (e.g., room type, minimum nights, neighborhood group) have the most significant impact on the dependent variable (price). This analysis enables data-driven decision-making by highlighting key factors influencing price formation.
5. **Social and Policy Implications**
   Based on the analysis results, understand the background and impact of policy changes such as New York City's short-term rental regulations. Discuss how the short-term rental market influences societal issues like housing shortages and provide actionable insights to policymakers and platform operators to help create balanced regulations and support measures.

This project aims to go beyond simple data analysis by uncovering the underlying economic and social factors driving the New York City Airbnb market. By doing so, it contributes to the sustainability of the hospitality industry, promotes local economic growth, and supports data-driven decision-making to create a balanced and informed ecosystem for all stakeholders.

Impact & Contribution of the Project

The sustained high accommodation prices facilitated by Airbnb in high-cost areas such as Manhattan, coupled with the concentration of supply and demand around private rooms, implies a potential encroachment on long-term housing availability. Residential spaces originally intended for permanent living are increasingly converted into short-term rental accommodations, and even relatively affordable private rooms are commercialized, exacerbating overall housing cost burdens. This feedback loop risks creating a structural imbalance in the housing market.

Such structural changes contribute to worsening housing shortages and, in the long term, threaten the stability of local communities. For example, in 2023, regulatory measures targeting short-term rentals have been introduced as a response to these challenges. The broader implications of short-term rental platforms on housing markets underscore the necessity of balancing market opportunities with social and economic sustainability.

By integrating external research and empirical evidence with the findings of this project, we can better understand how the growth of short-term rental platforms and market restructuring extend beyond individual hosts or tourists, influencing urban housing environments and policies on a larger scale. Ultimately, the insights from this project serve as a crucial foundation for evaluating the long-term impact of such platforms and for designing socially and economically sustainable housing policies.

Data Description

Since 2008, Airbnb has supported guests and hosts in expanding travel possibilities and providing more unique and personalized lodging experiences. This dataset contains information on Airbnb listings and activities in New York City for 2019.

The dataset includes detailed metrics about listings and activities across New York City's five main boroughs: Manhattan, Brooklyn, Queens, Bronx, and Staten Island. This data file contains all the necessary information to gain insights into hosts, geographical availability, and the metrics needed for making predictions and drawing conclusions.

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.647	-73.972	Private room	149
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.754	-73.984	Entire home/apt	225
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.809	-73.942	Private room	150
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.685	-73.960	Entire home/apt	89
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.799	-73.944	Entire home/apt	80

minimum_nights	number_of_reviews	last_review	reviews_per_month	calculated_host_listings_count	availability_365
1	9	2018-10-19	0.210	6	365
1	45	2019-05-21	0.380	2	355
3	0	NaN	NaN	1	365
1	270	2019-07-05	4.640	1	194
10	9	2018-11-19	0.100	1	0

The data are organized as follows. The data contains 48895 rows and 16 columns. Below is a description of 16 variables.

id: Unique ID of the Airbnb listing / name: Name of the listed property / host_id: Unique ID of the host / host_name: Name of the host / neighbourhood_group: General region of the location (neighbourhood), e.g., Brooklyn, Manhattan, Queens, Staten Island, and Bronx / neighbourhood: Specific neighbourhood where the property is located / latitude: Latitude coordinate of the property / longitude: Longitude coordinate of the property / room_type: Type of room (e.g., Private room, Shared room, Entire home/apt) / price: Price per night (in USD) / minimum_nights: Minimum number of nights guests are required to stay / number_of_reviews: Number of reviews the listing has received / last_review: Date of the last review / reviews_per_month: Average number of reviews per month / calculated_host_listings_count: Total number of properties listed by the host / availability_365: Number of days the property is available in a year.

Unnecessary variables were removed in advance for this data analysis. These variables are **'id', 'host_id', 'host_name', 'last_review', 'name', and 'neighbourhood'**.

Additionally, there were missing values in the **'reviews_per_month'** variable. The missing values in 'reviews_per_month' were filled using **KNNImputer**. First, the data was scaled using **RobustScaler**, and KNNImputer with **n_neighbours=5** was applied to fill the missing values based on the average of the 5 nearest neighbours.

```
df.isna().sum() #before Filled KNN Imputers
✓ 0.0s
```

neighbourhood_group	0
latitude	0
longitude	0
room_type	0
price	0
minimum_nights	0
number_of_reviews	0
reviews_per_month	10052
calculated_host_listings_count	0
availability_365	0

```
df.isnull().sum() #after Filled KNN Imputers
✓ 0.0s
```

neighbourhood_group	0
latitude	0
longitude	0
room_type	0
price	0
minimum_nights	0
number_of_reviews	0
reviews_per_month	0
calculated_host_listings_count	0
availability_365	0

For numerical variables, outliers were handled only for those containing outliers. Outliers were detected and processed based on the **Interquartile Range (IQR)**. Values outside **1.5 times the IQR** from the lower 5% (Q1) and upper 95% (Q3) were treated as outliers and replaced with the lower and upper limits.

```
df.head()
✓ 0.0s
```
Python

	neighbourhood_group	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365
0	Brooklyn	40.647	-73.972	Private room	149	1	9	0.210	6	365
1	Manhattan	40.754	-73.984	Entire home/apt	225	1	45	0.380	2	355
2	Manhattan	40.809	-73.942	Private room	150	3	0	1.373	1	365
3	Brooklyn	40.685	-73.960	Entire home/apt	89	1	270	4.640	1	194
4	Manhattan	40.799	-73.944	Entire home/apt	80	10	9	0.100	1	0

```
df.describe()
✓ 0.1s
```

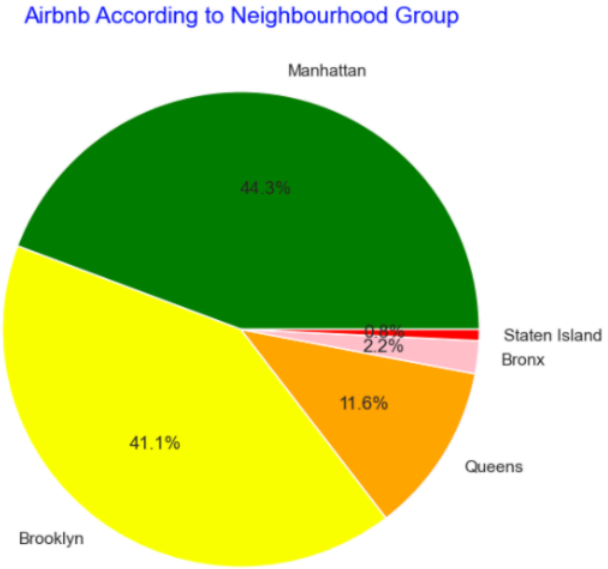
	latitude	longitude	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365
count	48895.000	48895.000	48895.000	48895.000	48895.000	48895.000	48895.000	48895.000
mean	40.729	-73.952	144.201	6.346	23.022	1.369	3.355	112.781
std	0.055	0.046	123.271	10.452	42.623	1.450	7.078	131.622
min	40.500	-74.211	0.000	1.000	0.000	0.010	1.000	0.000
25%	40.690	-73.983	69.000	1.000	1.000	0.280	1.000	0.000
50%	40.723	-73.956	106.000	3.000	5.000	1.220	1.000	45.000
75%	40.763	-73.936	175.000	5.000	24.000	1.580	2.000	227.000
max	40.913	-73.713	827.500	73.500	285.000	10.700	36.000	365.000

After all these processes, the dataset consisted of **48,895 rows and 10 columns**.

```
df.info()
✓ 0.0s
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 48895 entries, 0 to 48894
Data columns (total 10 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   neighbourhood_group             48895 non-null  object
 1   latitude                        48895 non-null  float64
 2   longitude                       48895 non-null  float64
 3   room_type                       48895 non-null  object
 4   price                           48895 non-null  float64
 5   minimum_nights                  48895 non-null  float64
 6   number_of_reviews               48895 non-null  int64
 7   reviews_per_month               48895 non-null  float64
 8   calculated_host_listings_count  48895 non-null  int64
 9   availability_365                48895 non-null  int64
dtypes: float64(5), int64(3), object(2)
memory usage: 3.7+ MB
```

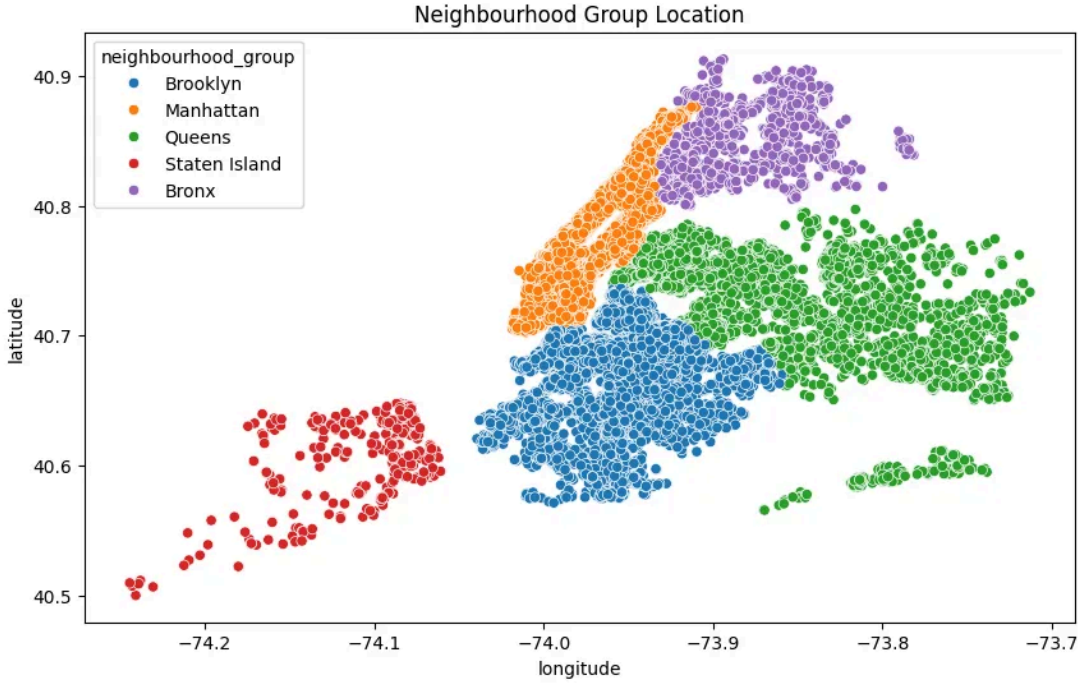
*Before modeling, categorical variables (**'neighbourhood_group'** and **'room_type'**) were converted into numerical data using **one-hot encoding**.*

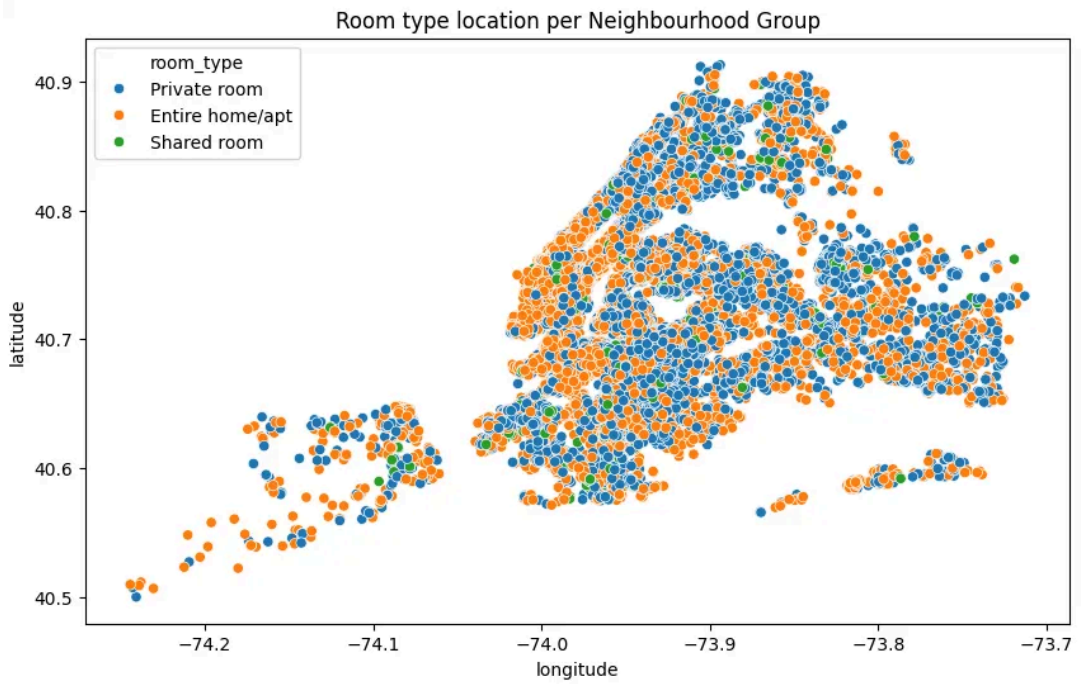
Upon examining the data, **Manhattan** accounted for the largest proportion at **44.3%**, followed by **Brooklyn** at **41.1%**. The data showed a tendency to be concentrated in Manhattan and Brooklyn.



Next part is data visualization.

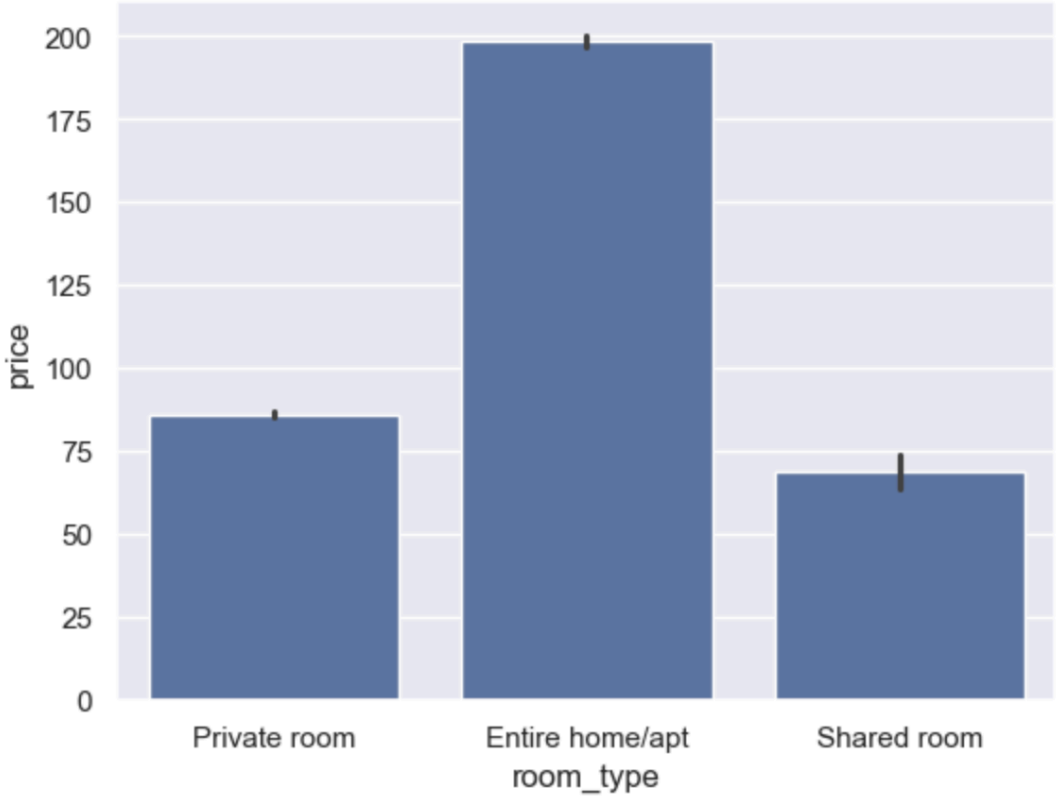
1. Visualizing Location Data Based on Neighbourhood Group and Room Type



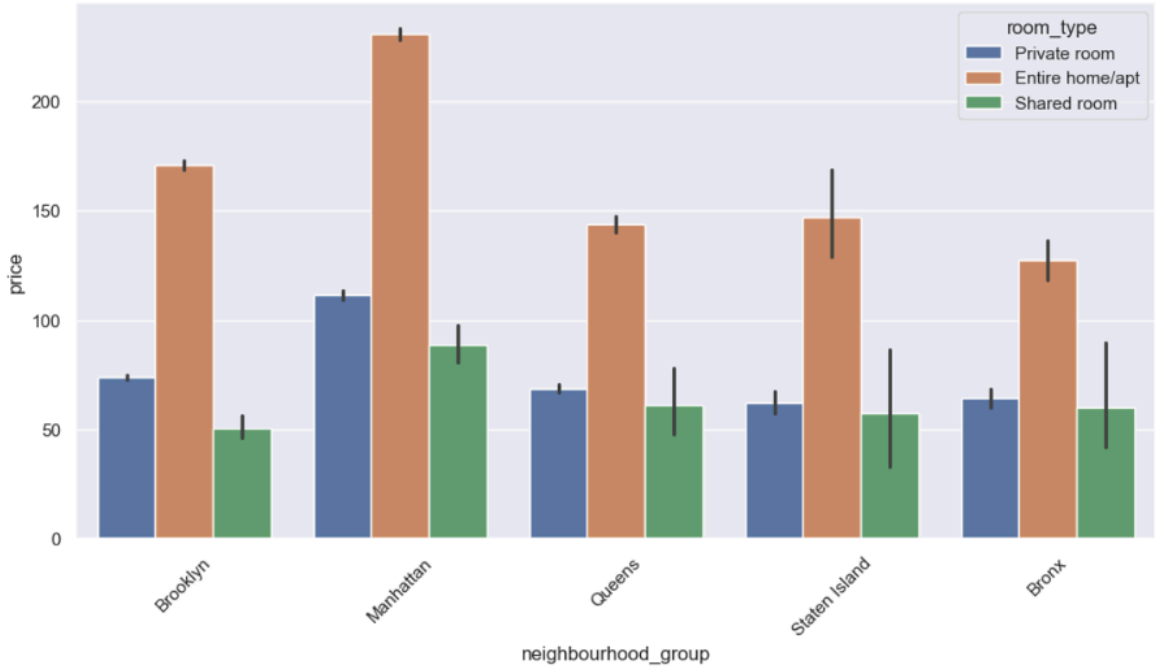


Through this, the latitude and longitude can be used to distinguish each region, and it is possible to check the distribution of room types across regions.

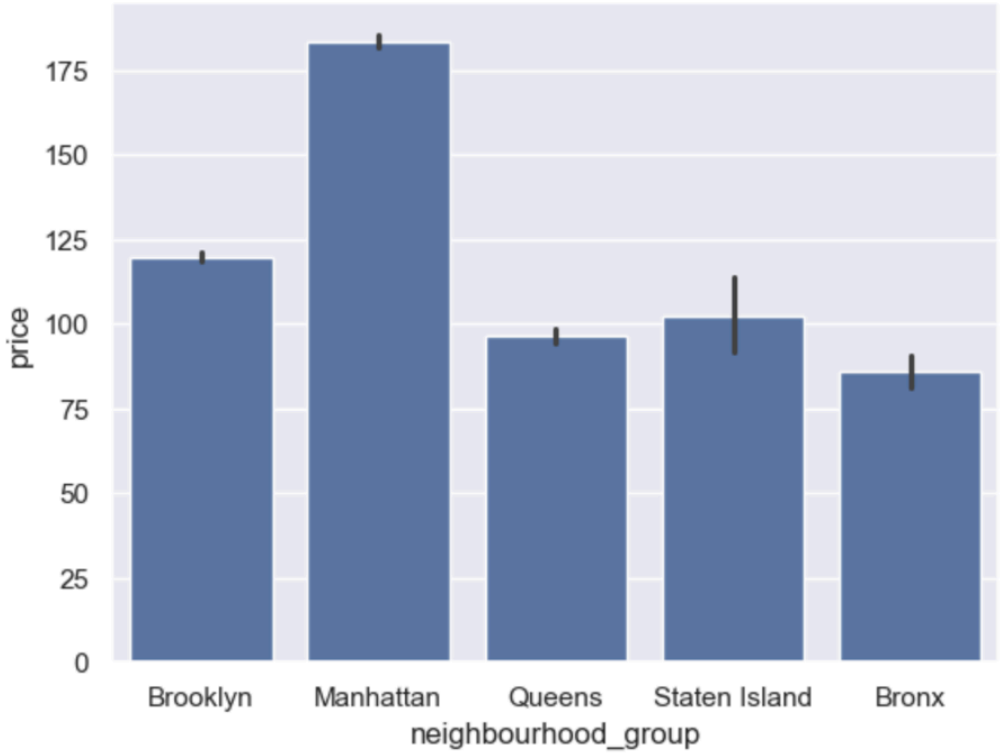
2. The relationship with room_type variable and price variable



3. Price Distribution by Room Type in Neighbourhood Groups



4. The relationship with neighbourhood_group variable and price variable



It was also confirmed that **Manhattan had a higher average price** than other regions. Based on this, I decided to **divide the data into Manhattan and non-Manhattan regions** for modeling.

```
m_true = df[df["neighbourhood_group_manhattan"] == True]
m_false = df[df["neighbourhood_group_manhattan"] == False]
```

len(m_true)	len(m_false)
21661	27234

Model Explanation

The following features were arbitrarily added before modeling.

The newly added features were created to analyze the operational status of accommodations and customer usage patterns more accurately.

NEW_estimated_listed_months estimates the number of months a listing has been active, while **NEW_availability_ratio** represents the proportion of days a house is available throughout the year. **NEW_daily_average_reviews** calculates the average number of reviews per day, and **NEW_average_stay_duration** estimates the average duration of a guest's stay. Lastly, **NEW_house_occupancy_rate** indicates the occupancy rate of a house throughout the year. These features contribute to effectively explaining the popularity, operational characteristics, and customer behavior of accommodations during the modeling process.

```
# This feature can be used to estimate for how long a house has been listed.
# This duration is calculated by dividing the total number of reviews that the house has received by the number of reviews per month.
df['NEW_estimated_listed_months'] = df['number_of_reviews'] / df['reviews_per_month']

# This feature gives the ratio of how long a house is available throughout the year.
df['NEW_availability_ratio'] = df['availability_365'] / 365

# This feature gives the daily average reviews a host receives. It divides the reviews per month by the number of days in a month.
df['NEW_daily_average_reviews'] = df['reviews_per_month'] / 30

# This feature estimates the average duration a customer stays. It divides the total number of reviews by the reviews per month.
df['NEW_average_stay_duration'] = df['number_of_reviews'] / df['reviews_per_month']

# This feature gives the occupancy  of a house throughout the year. It subtracts from 365 the number of days a house is available in a year.
df['NEW_house_occupancy_rate'] = (365 - df['availability_365']) / 365
```

The Manhattan dataframe (m_true) and the non-Manhattan dataframe (m_false) were separated, and various machine learning algorithms were applied to compare model performance. The algorithms used for evaluation include **Linear Regression**, **Ridge**, **Lasso**, **ElasticNet**, **KNN**, **Decision Tree**, **Random Forest**, and **CatBoost**.

Performance metrics such as **RMSE (Root Mean Square Error)**, **R^2 Score**, **MAE (Mean Absolute Error)**, and **MSE (Mean Squared Error)** were calculated, and the execution time for each model was recorded. The results were visualized for each performance metric to provide an intuitive comparison between models. Based on the evaluation, the **CatBoost** algorithm was determined to be the most effective model.

CatBoost (Categorical Boosting) is a Gradient Boosting algorithm known for its ability to handle complex data and interactions effectively. It was chosen for this project because, while it excels at processing categorical data automatically, it also works well with datasets where One-Hot Encoding has already been applied. Moreover, CatBoost performs strongly with default settings, eliminating the need for hyperparameter tuning, making it a practical choice for reliable analysis.

CatBoost uses decision trees and a unique learning method called Ordered Boosting, which processes data sequentially and prevents overfitting by separating training and testing data. It also handles missing values natively and provides feature importance scores, making it both efficient and interpretable.

Its strengths include robust generalization through Ordered Boosting, fast training with GPU support, and built-in handling of missing values. CatBoost also maintains strong performance with pre-encoded datasets, aligning well with this project's requirements. However, its complexity may make interpretation challenging for beginners, and memory usage can increase with large datasets.

In conclusion, CatBoost is a powerful tool for this project, offering both predictive accuracy and interpretability. While its complexity and memory usage require consideration, it remains a strong choice for handling complex data and generating reliable results.

Result of Analysis and Model Outcome

The dataset was split into training and test sets using an 80:20 ratio. Specifically, 80% of the data was allocated for training the models, and 20% was reserved for testing their performance. To ensure reproducibility of results, a random state of 17 was applied during the splitting process. And the dependent variable is **price**.

```
##from m_true dataframe
m_X_train, m_X_test, m_y_train, m_y_test = train_test_split(m_X, m_y, test_size=0.20, random_state=17)
```

```
##from m_false dataframe
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.20, random_state=17)
```

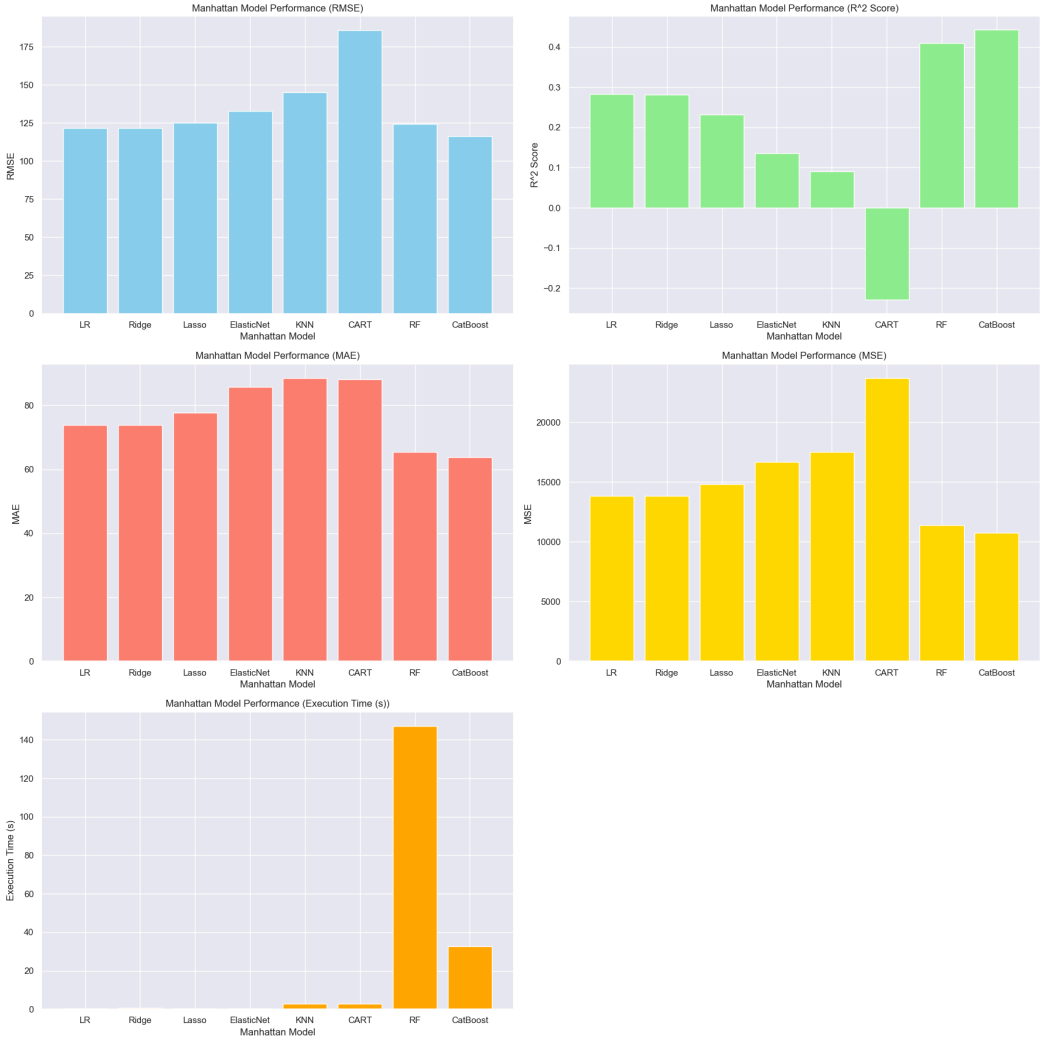
In the provided modeling process, the train-test split was performed using an 80:20 ratio, where 80% of the data was used to train the models and 20% was reserved for testing. The **train data** (**m_X_train, m_y_train or X_train, y_train**) was utilized to train the models, and predictions were made on the **test data** (**m_X_test** or **X_test**). The accuracy and performance of these predictions were then evaluated by comparing the predicted values with the actual test data.

The performance metrics, including RMSE, R^2 score, MAE, MSE, and execution time, were evaluated for various models: Linear Regression, Ridge, Lasso, ElasticNet, KNN, Decision Tree, Random Forest, and CatBoost. Across both Manhattan and Non-Manhattan datasets, Based on these results, the CatBoost algorithm was identified as the most effective model for accurately predicting the target variable, **price**.

In the Manhattan dataset, CatBoost achieved the lowest RMSE (116.2801) and MAE (63.6945) values, along with the highest R^2 score (0.4431), confirming its ability to minimize prediction errors and effectively explain the variability of the dependent variable. While Random Forest also delivered stable results with a high R^2 score (0.4094) and low MSE (11385.0619), its execution time (147.19 seconds) was significantly longer, making it less efficient. Linear models like Linear Regression and Ridge provided average performance with relatively short execution times, making them efficient alternatives when computational speed is a priority. Models such as ElasticNet, KNN, and Decision Tree performed poorly, showing lower R^2 scores and higher RMSE values.

Similarly, in the Non-Manhattan dataset, CatBoost excelled with the lowest RMSE (77.4095) and MAE (40.4597), as well as the highest R^2 score (0.3961). Random Forest ranked second in performance with a strong R^2 score (0.3713) and stable error metrics but suffered from extremely long execution times (197.13 seconds). Linear models once again provided moderate performance with shorter execution times, whereas ElasticNet, KNN, and Decision Tree displayed the weakest results, with Decision Tree even recording a negative R^2 score (-0.1665).

Model	RMSE	R^2 Score	MAE	MSE	Execution Time (s)
LR	121.5432	0.2828	73.9203	13826.193	0.34
Ridge	121.6067	0.2816	73.8771	13847.8828	0.64
Lasso	125.1809	0.2314	77.6748	14817.1002	0.48
ElasticNet	132.6337	0.135	85.683	16674.1207	0.24
KNN	145.2947	0.0903	88.5104	17536.0691	2.68
CART	185.8586	-0.2293	88.0662	23696.7674	2.94
RF	124.3179	0.4094	65.506	11385.0619	147.19
CatBoost	116.2801	0.4431	63.6945	10735.1137	32.64



<Manhattan Model Performance Metrics>

Model	RMSE	R^2 Score	MAE	MSE	Execution Time (s)
LR	81.9869	0.2729	45.606	6951.4897	0.33
Ridge	81.9875	0.2727	45.5985	6952.7141	0.19
Lasso	82.9927	0.2568	46.1089	7104.8103	0.35
ElasticNet	89.5403	0.1285	53.4293	8331.5719	0.33
KNN	100.6844	-0.0481	60.9901	10019.4652	4.47
CART	126.1227	-0.1665	55.7316	11152.1899	5.12
RF	84.3389	0.3713	41.5476	6010.6641	197.13
CatBoost	77.4095	0.3961	40.4597	5773.2519	27.94

<Non-Manhattan - Model Performance Metric>

```python
def plot_importance(model, features, num=50, save=False):
    feature_imp = pd.DataFrame({'Value': model.feature_importances_, 'Feature': features.columns})
    print(feature_imp.sort_values(by="Value", ascending=False))
```

	Value	Feature
16	17.284	room_type_private_room
1	15.864	longitude
0	14.492	latitude
2	10.653	minimum_nights
5	8.838	calculated_host_listings_count
11	5.418	new_house_occupancy_rate
6	5.255	availability_365
8	4.813	new_availability_ratio
7	4.283	new_estimated_listed_months
10	3.281	new_average_stay_duration
17	3.073	room_type_shared_room
3	2.903	number_of_reviews
4	2.265	reviews_per_month
9	1.577	new_daily_average_reviews
12	0.000	neighbourhood_group_brooklyn
13	0.000	neighbourhood_group_manhattan
14	0.000	neighbourhood_group_queens
15	0.000	neighbourhood_group_staten_island

<Manhattan>

	Value	Feature
16	31.407	room_type_private_room
1	14.177	longitude
0	11.876	latitude
11	6.406	new_house_occupancy_rate
2	6.109	minimum_nights
17	5.659	room_type_shared_room
6	4.680	availability_365
8	3.569	new_availability_ratio
7	3.180	new_estimated_listed_months
3	3.095	number_of_reviews
5	2.832	calculated_host_listings_count
4	2.683	reviews_per_month
9	1.908	new_daily_average_reviews
10	1.794	new_average_stay_duration
12	0.484	neighbourhood_group_brooklyn
14	0.123	neighbourhood_group_queens
15	0.016	neighbourhood_group_staten_island
13	0.000	neighbourhood_group_manhattan

<Non-Manhattan>

In the Manhattan data, the most important variable was **room_type_private_room**, with an importance value of **17.284**. This indicates that the 'private room' type contributes the most to the predictions of the accommodation model. Following this, **longitude** and **latitude** showed high importance values of **15.864** and **14.492**, respectively, demonstrating that the location of accommodations plays a crucial role in predicting prices or demand in the Manhattan data.

Additionally, **minimum_nights** (10.653) and **calculated_host_listings_count** (8.838) recorded moderate levels of importance, suggesting that the minimum number of nights and the number of listings managed by a host influence the model's performance.

In the Non-Manhattan data, **room_type_private_room** was again the most important variable, with an importance value of **31.407**, confirming that the 'private room' type is also a key factor in predictions outside of Manhattan. The **longitude** and **latitude** variables maintained their significance, with important values of **14.177** and **11.876**, respectively, showing that location remains an important factor.

A notable difference from the Manhattan data is that **new_house_occupancy_rate** (6.406) and **room_type_shared_room** (5.659) showed relatively high importance in the Non-Manhattan data. This suggests that the share of shared rooms and housing occupancy rates have a greater impact on predictions in Non-Manhattan areas.

Summary

The data analysis highlights the significant influence of **'room_type_private_room'** in predicting accommodation demand across both Manhattan and Non-Manhattan datasets. This underscores the importance of privacy and affordability in New York City's competitive rental market. For business and management, these findings suggest that optimizing the supply of private rooms could be a key strategy for meeting customer demand and maximizing revenue.

In Manhattan, where accommodation costs and space constraints are high, private rooms offer a balance of cost-effectiveness and comfort, making them highly attractive to both short-term and long-term visitors. Similarly, in Non-Manhattan areas, private rooms remain a popular choice, while factors like occupancy rates and shared room options also play a notable role, reflecting regional differences in customer preferences.

To capitalize on these insights, hosts and operators should focus on increasing private room availability and tailoring pricing strategies to align with location-specific demand patterns. Additionally, optimizing occupancy rates and understanding local preferences for shared accommodations in Non-Manhattan areas can further enhance market competitiveness.

References

1. https://www.spinxdigital.com/blog/what-you-can-learn-from-airbnbs-successful-startup/
2. https://www.citypopulation.de/en/usa/newyorkcity/
3. https://www.costar.com/article/230699136/manhattan-named-the-nations-most-expensive-place-to-live-by-new-report
4. https://www.wired.com/story/airbnb-ban-new-york-city
5. Koster, H. R. A., van Ommeren, J., & Volkhausen, N. (2021). Short-term rentals and the housing market: Quasi-experimental evidence from Airbnb in Los Angeles. *Journal of Urban Economics, 124,* 103356. https://doi.org/10.1016/j.jue.2021.103356