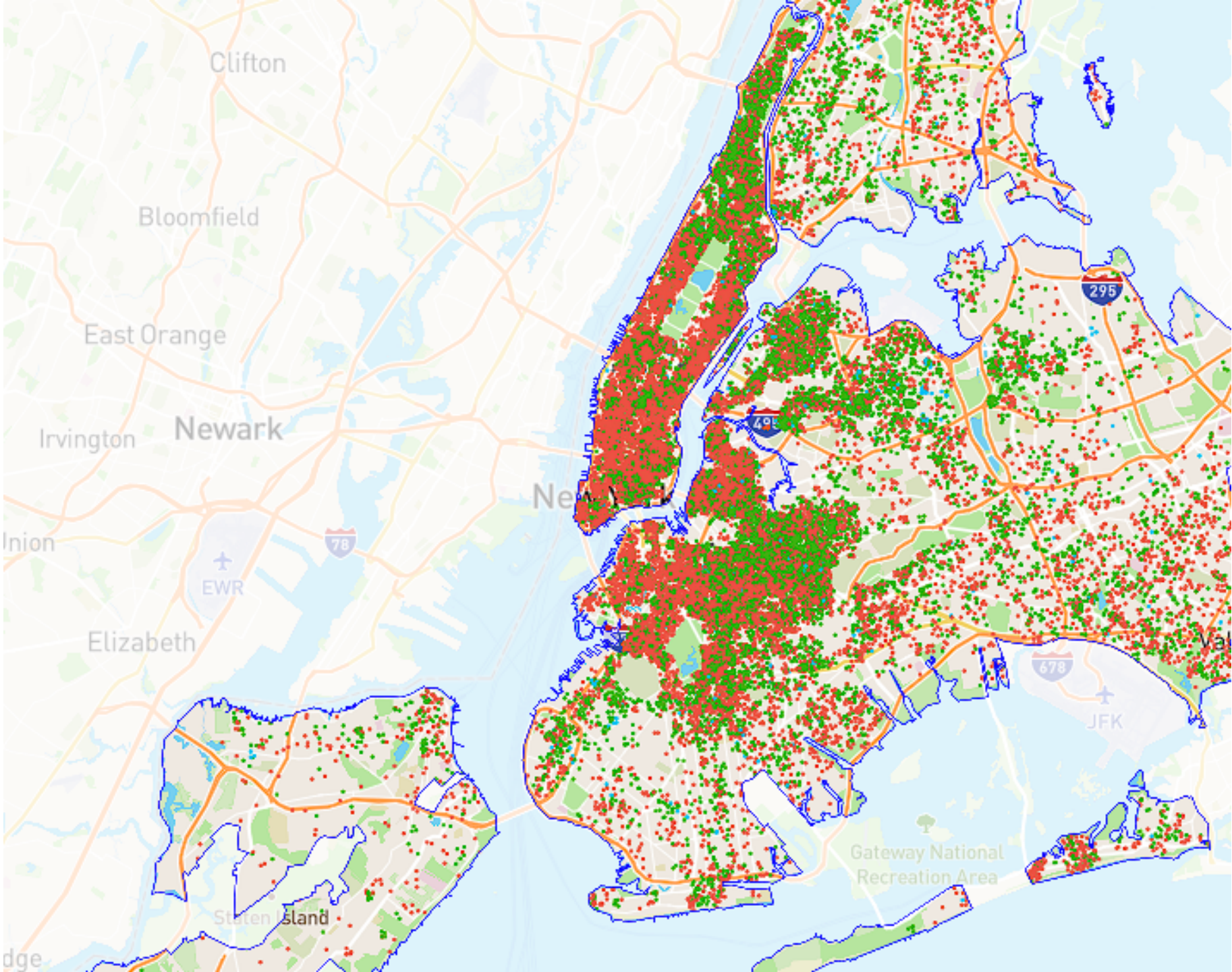
# Predicting Short-term Rental Prices




Springboard – Capstone 2

Nizar Altawam

2023

# Defining Short-term Rentals (STRs)

STRs are a type of accommodation that allows travelers to rent a property for a short period of time, usually less than a month.

They have become increasingly popular in recent years, especially in urban areas like New York City (NYC), where they offer an alternative to hotels and other traditional lodging options.

However, STRs also pose various challenges for the hosts, guests, and regulators, such as pricing, quality, safety, and legality.

# Project Context and Motivation

The motivation for this project is to provide consulting services for a start-up company that operates in the NYC market.

It is seeking help optimizing its pricing strategy on its platform and increasing its revenue and market share

The aim is to build a machine learning prediction model for STR prices

# Objectives and Scope

| Using historical data on short-term listings scraped from Airbnb's website to: | | |
|---|---|---|
| Identify key insights and highlight significant trends | Train a predictive model using important features | Conduct a profitability analysis to draw actionable insights |

# Data Collection and Wrangling

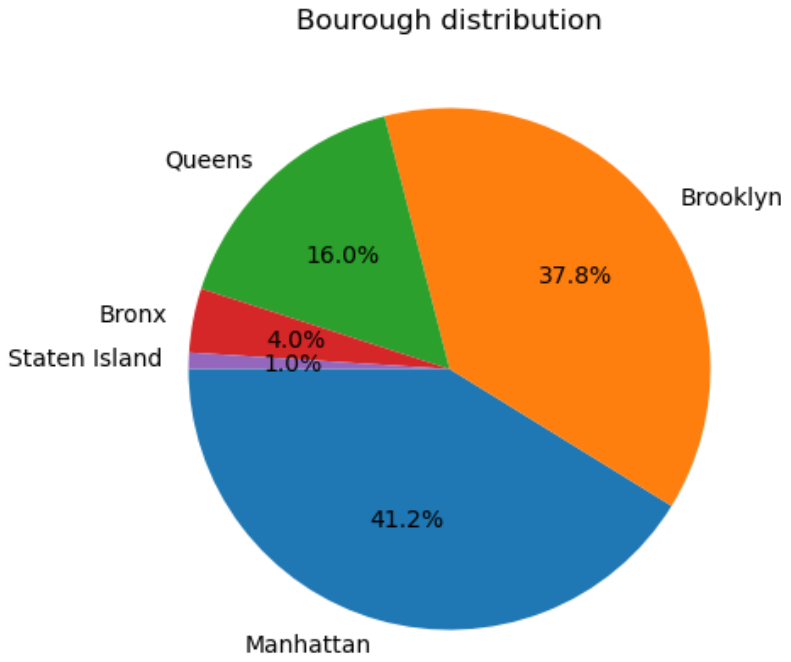
- **Source**: InsideAirbnb.com - A website that collects and publishes historical Airbnb listing data.
- **License**: CC BY 4.0
- The data covers a 12-month period (ending December 4th, 2022).
- Three separate datasets were obtained:
  - “listings.csv.gz” - Detailed Listings data
  - “calendar.csv.gz” - Detailed Calendar Data
  - “listings.csv” - Summary information and metrics for listings in New York City.
- **Wrangling**: dropped missing/incorrect values, outliers, and irrelevant features

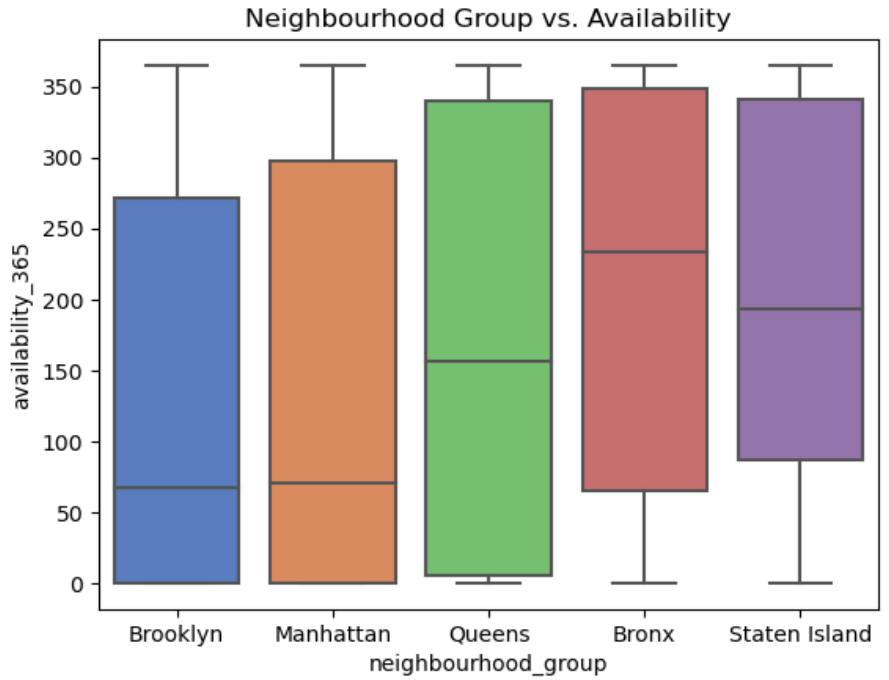
# Exploratory Data Analysis

- **Data distribution analysis**: examined the distribution of key features and their relationship with other features

- **Correlation analysis**: performed correlation analysis to examine the price distribution across the boroughs and room types

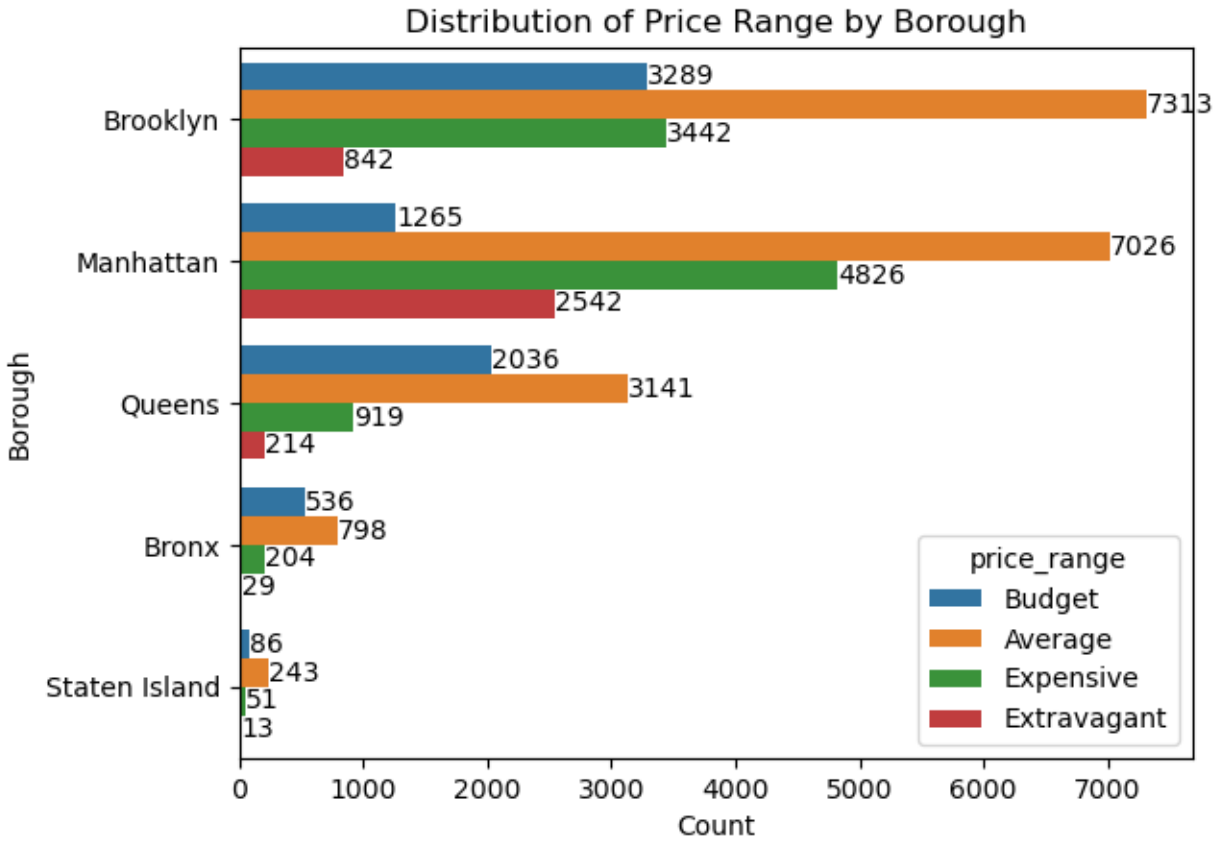
- **Key findings:** as follows…

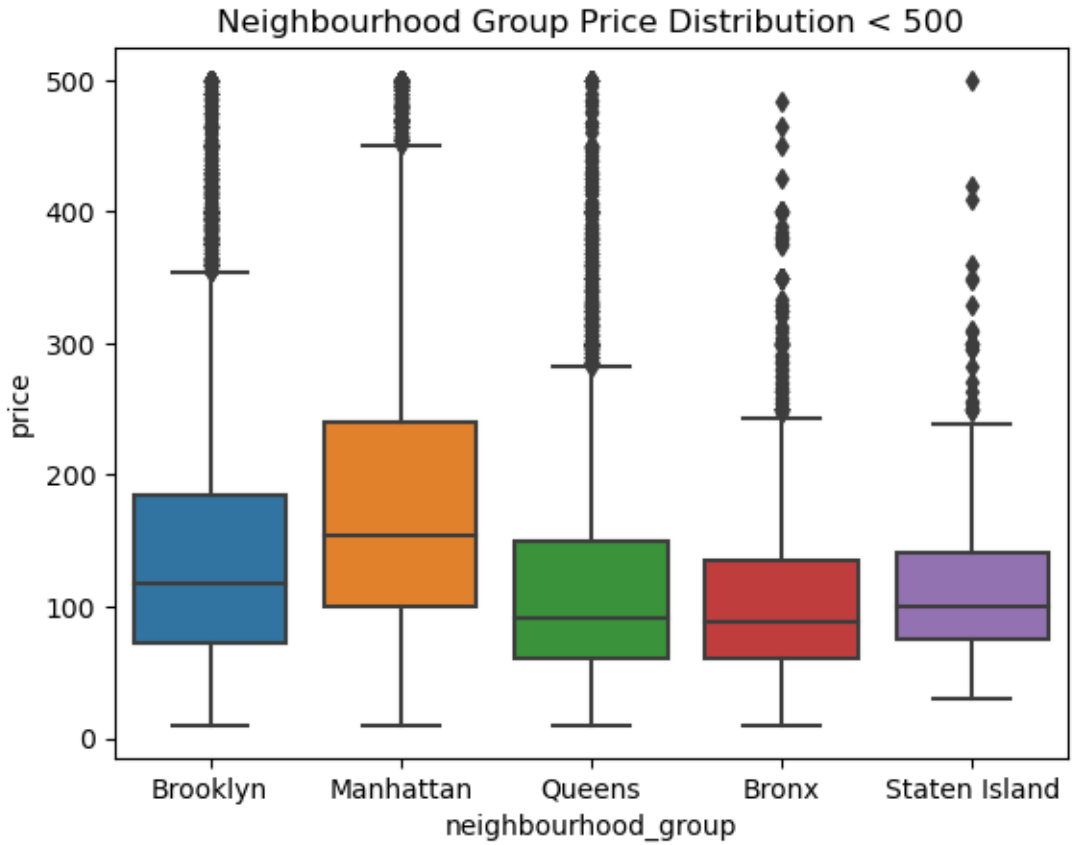


# Key findings

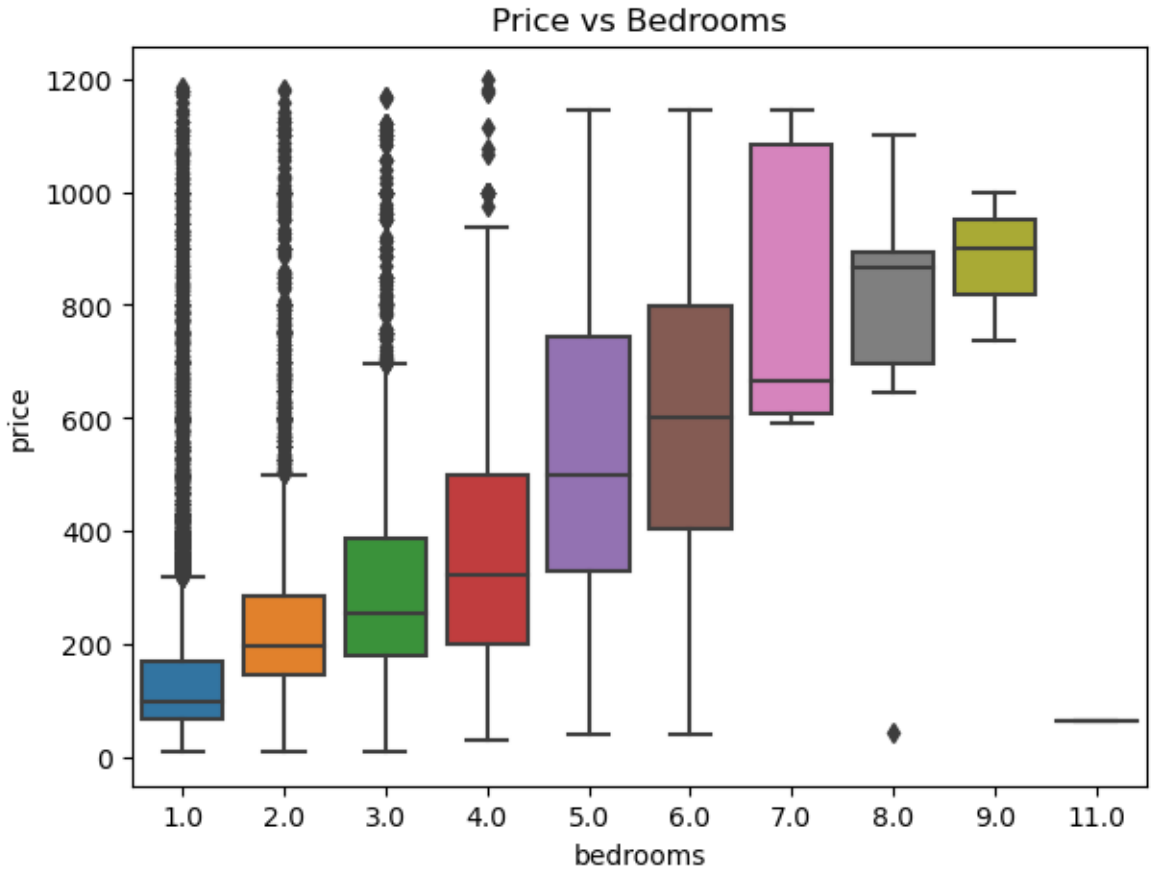
1. Brooklyn and Manhattan had the largest share of listings (37.8% and 41.2%, respectively) and the highest median prices ($118 and $154, respectively)

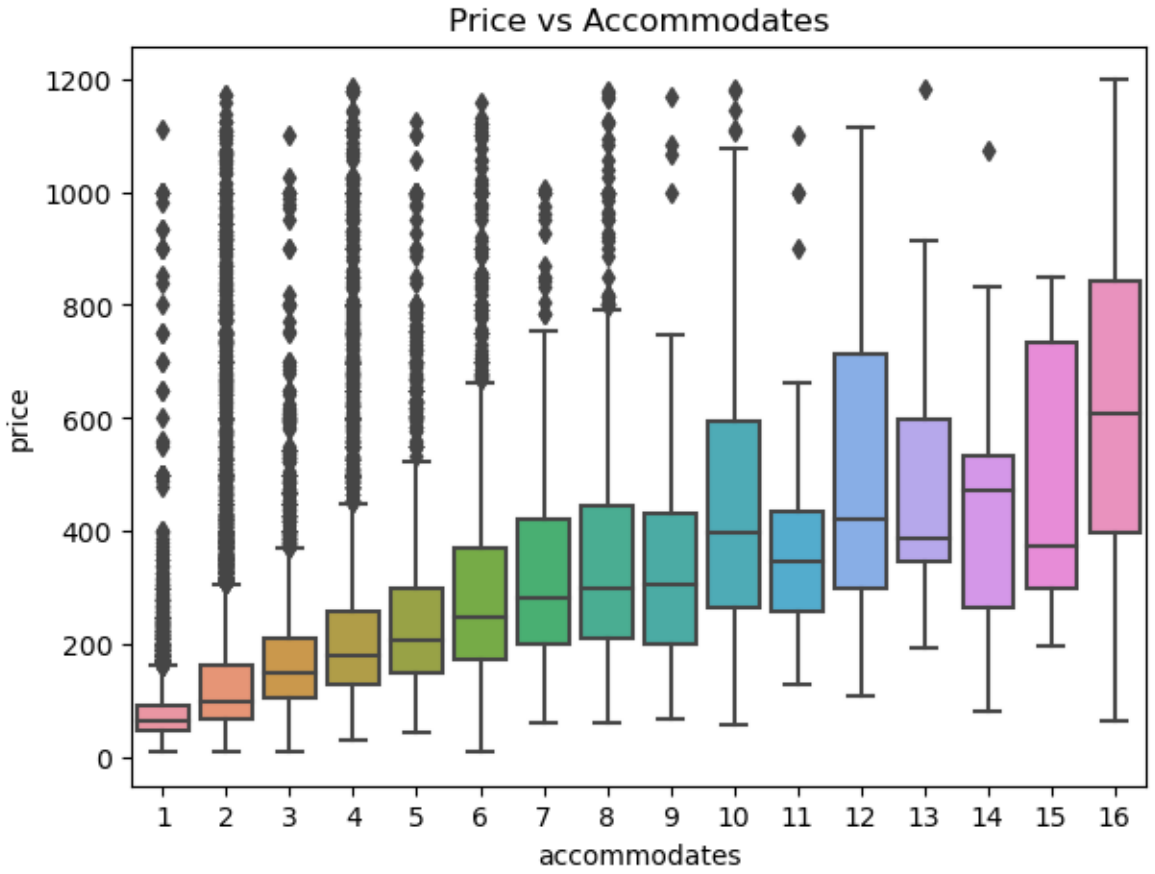


# Key findings (cont.)

2. More bedrooms and guests accommodated were typically associated with higher prices

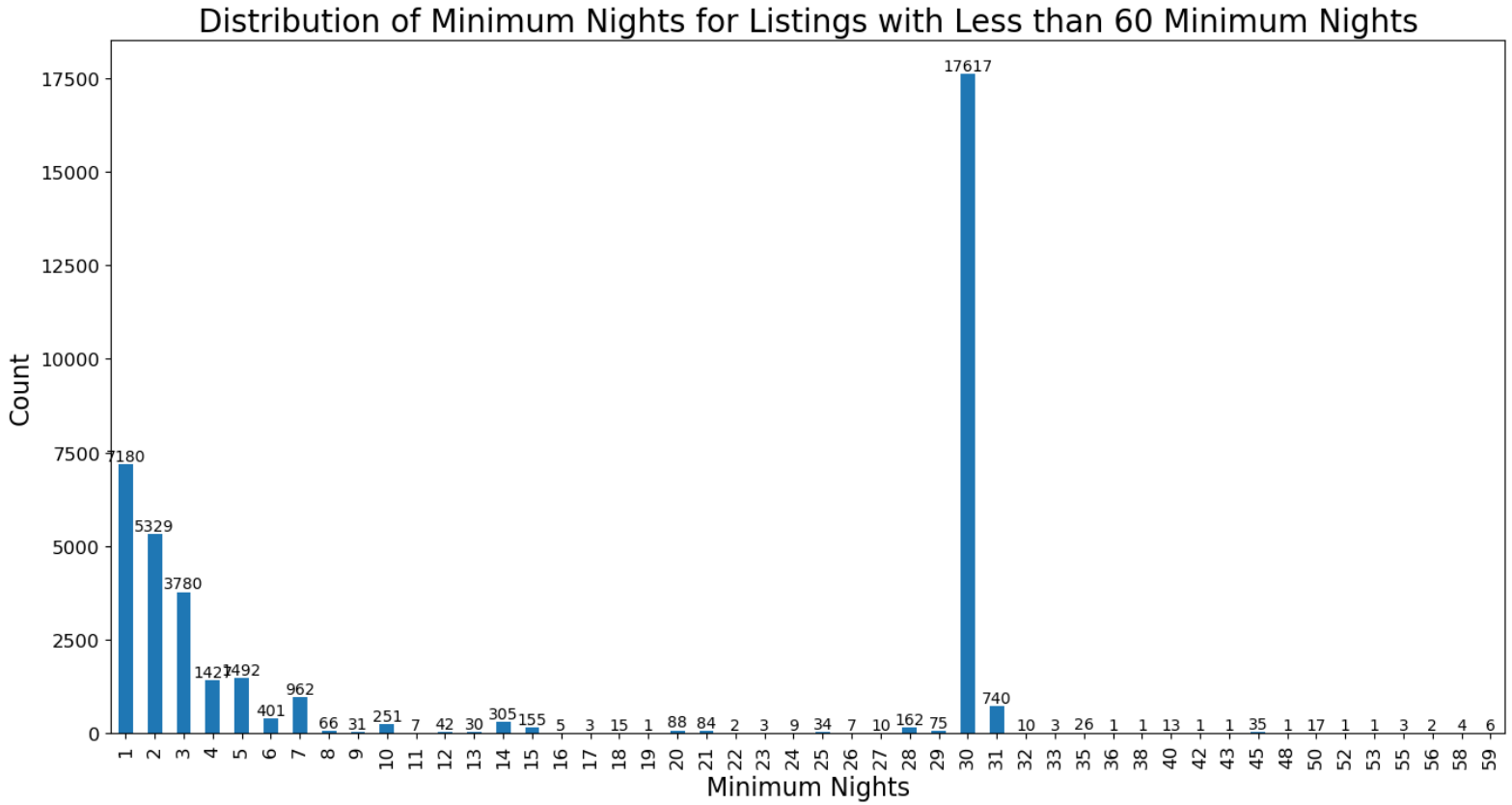


# Key findings (cont.)

3. There was a shift to longer-term stays (> 30 nights) possibly due to regulations or the recent pandemic

*note: Due to this imbalance in distribution, I will amend the definition of STR in the context of this project to include listings with less than **31 days**



# Data Pre=processing

- **Categorical feature encoding**: encoded borough, neighborhood, and room type features using one-hot and ordinal encoding
- **Data splitting**: split the data into training and testing datasets with a 70/30 ratio
- **Baseline model**: created a dummy model for comparison using the mean value of price as a constant prediction
- **Models**: applied two machine learning algorithms:
  - Random Forest Regressor (RFR) and an XGBoost for Regression(XGBR).
  - Both models were tuned using grid search and cross validation to find the optimal hyperparameters.

# Modeling

- **Evaluation metrics:** used root mean squared error (RMSE) and mean absolute error (MAE) metrics to evaluate the performance of the models
  - The initial RFR model had an RMSE 111.6 with predictions being off by an average of $66.14.
- **Feature engineering**:
  - In attempt to improve model performance, a new feature was created using the distance from a popular landmark within the same borough of each listing.

# Performance Evaluation

## Results:

> After adding the landmark proximity feature, the RFR performance improved to an RMSE of 108.4 and a prediction MAE of $64.75

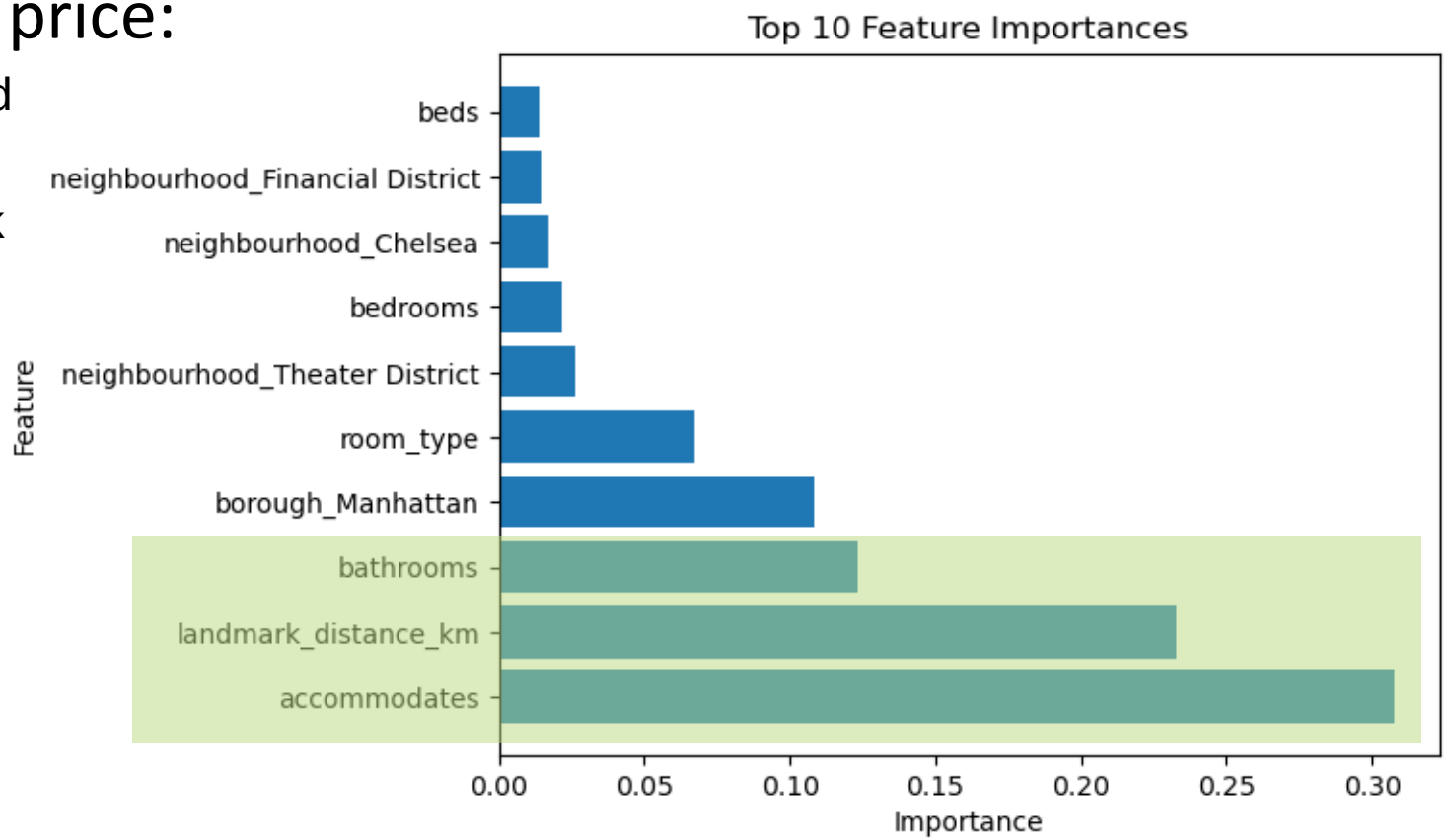
> The XGB model was also retrained with the new feature and scored an RMSE of 105.7 and MAE of $63.45.

This slight improvement in performance compared to the RFR was noted, however running this model utilized more computational resources and took twice as long as the RFR needed to be trained.

# Feature Importance Analysis

- Using RFR model for feature importance* to identify which features had the most influence on price:
  - Number of guests accommodated
  - number of bathrooms
  - distance from a popular landmark

**It's important to note that the importance score is a relative measure within the context of the specific model and dataset. It does not necessarily reflect the absolute importance of a feature or imply causality*



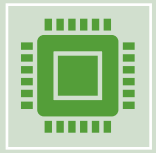
# Profitability Analysis

- To understand the model's performance and estimate potential revenue we conducted a hypothetical profitability analysis.
- The analysis compared our model to a hypothetical model that takes the average price of listings aggregated by neighborhood and property type
- Occupancy rate was estimated for each listing using availability information in the detailed calendar data. This was then factored into the revenue estimation.
- The results indicate, for the listings assessed and over the 12-month period, our model would generate approximately a **1.4% increase in revenue**.

# Future Steps

Additional feature engineering: explore other features that may improve the prediction accuracy such as amenities, reviews, ratings, or sentiment analysis

Applying models on other cities: test the model on other cities with similar or different market conditions to evaluate its robustness and transferability



Using different models: try other machine learning algorithms or techniques such as neural networks, ensemble methods, or deep learning

# Conclusion

This project presented a machine learning prediction model for short-term listing prices in New York City using historical Airbnb data

The data was collected, wrangled, and analyzed to explore the distribution and correlation of price with various features such as location, room type, capacity, and availability

Two regression models were applied and tuned: Random Forest and XGboost. A new feature was also created using the distance from a popular landmark

The RFR model performed slightly better than the XGboost model in terms of computational efficiency. The most important features for predicting price were guests accommodated, number of bathrooms, and distance from a landmark

We discussed any assumptions & limitations that could affect the validity and reliability of the predictions in the project report.