COMPARATIVE ANALYSIS OF NYC AIRBNB RENTALS

5000 Data Analytics Final Project Presentation

By: Daniel Mehta & Thomas Nash

Problem Statement

A comparative analysis of value and performance between two Airbnb rental types across the five boroughs of New York City.

Executive Summary

This analysis examined Airbnb rental data across New York City's five boroughs, focusing on pricing patterns and predictive modeling. The study utilized a dataset of 102,599 listings, which was cleaned and processed to 85,029 valid entries for analysis.

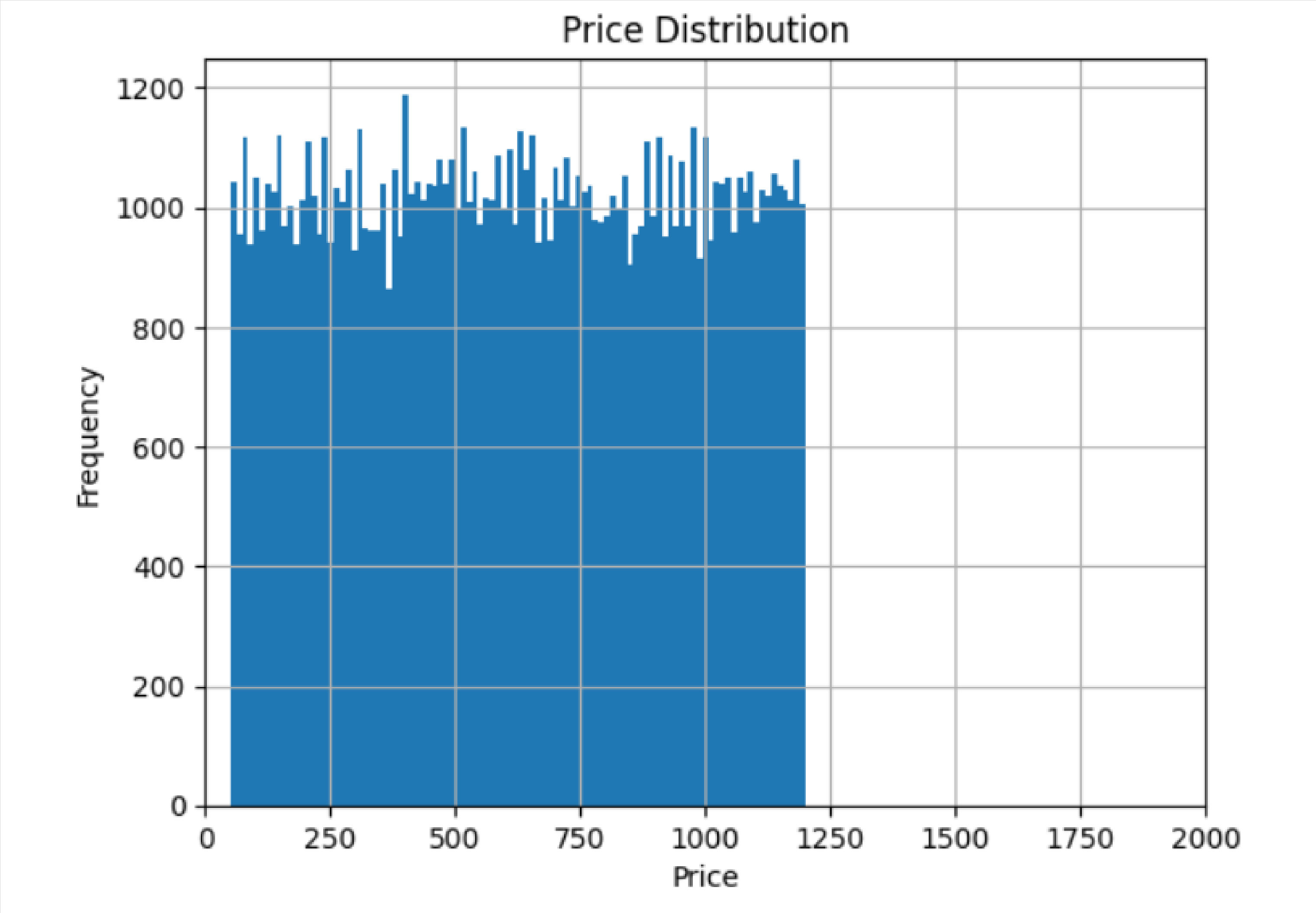
Data Preprocessing & Cleaning

The initial dataset underwent several cleaning steps:

- **Removal of irrelevant columns including ID fields, geographical coordinates, and text-based descriptions**
- **Standardization of price and fee columns by removing currency symbols and converting to float values**
- **Treatment of missing values in critical columns**
- **Encoding of categorical variables using one-hot encoding**
- **Creation of derived features including total cost and price per night**

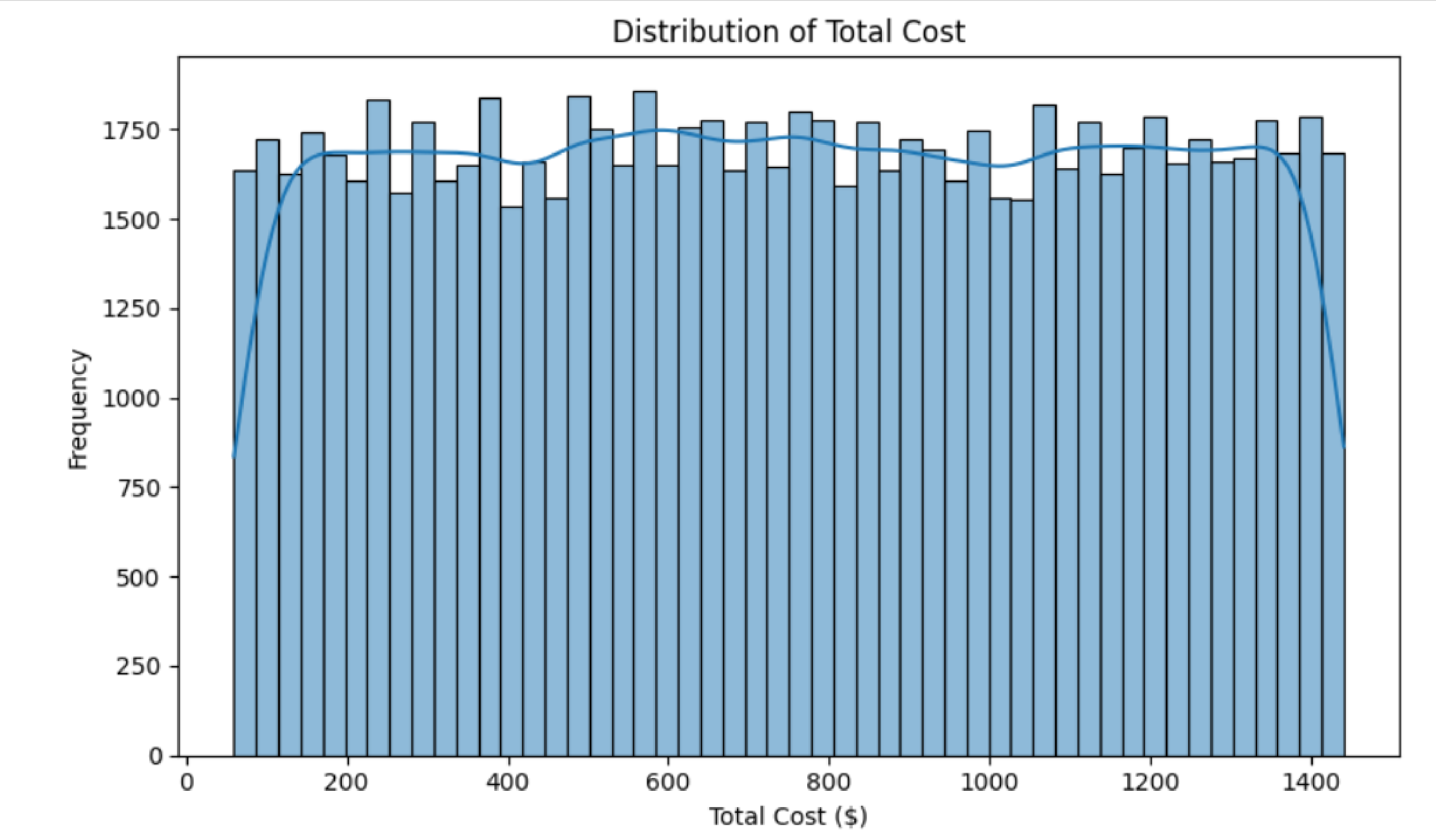
Key Findings:
Distribution Patterns

- The total cost distribution shows a right-skewed pattern, with most listings clustered in the lower price ranges
- Manhattan emerged as the borough with the highest average costs
- Entire homes/apartments consistently commanded higher prices across all boroughs



Key Findings:
Distribution Patterns

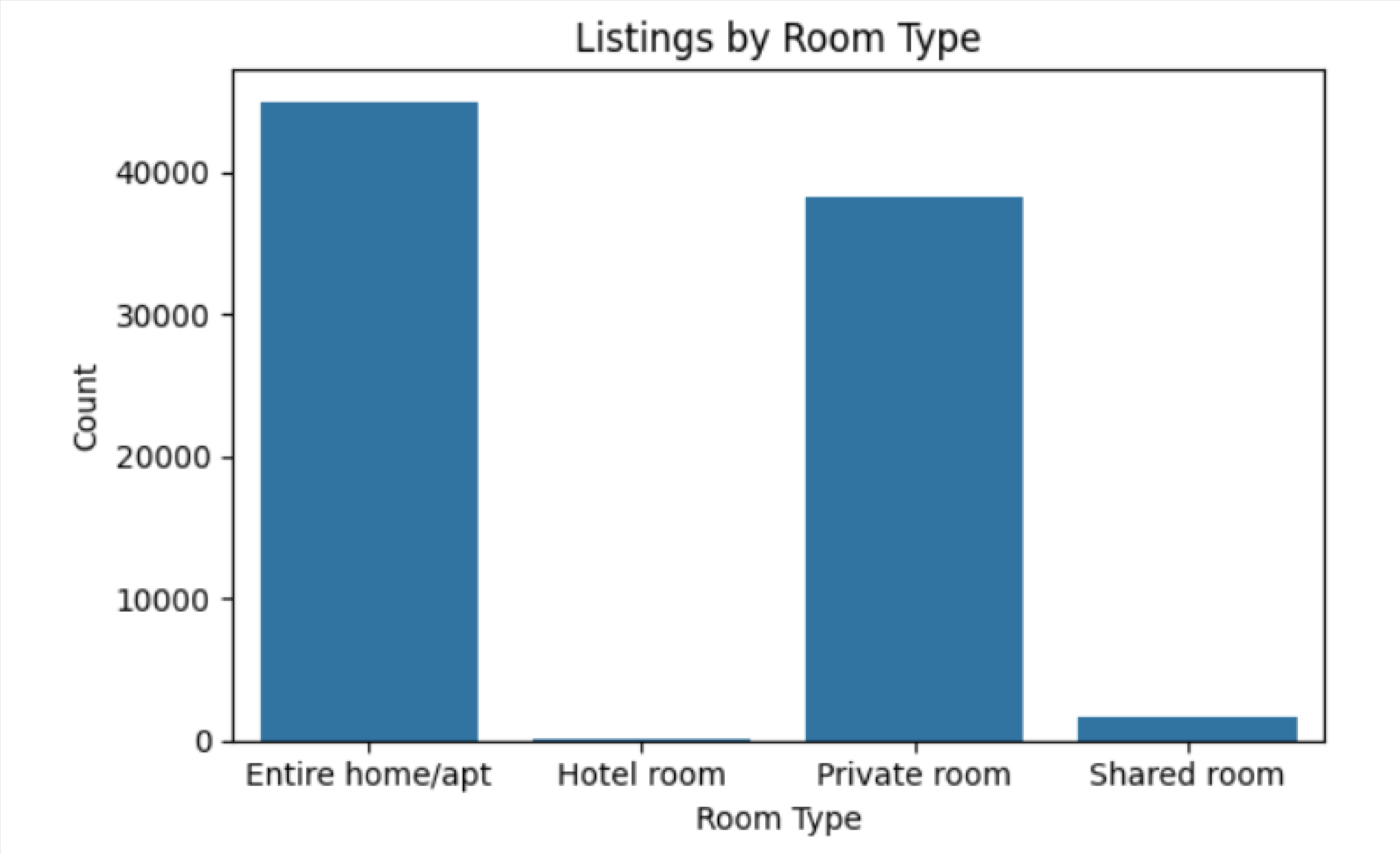
- The total cost distribution shows a right-skewed pattern, with most listings clustered in the lower price ranges
- Manhattan emerged as the borough with the highest average costs
- Entire homes/apartments consistently commanded higher prices across all boroughs



Key Findings:

Room Type Analysis

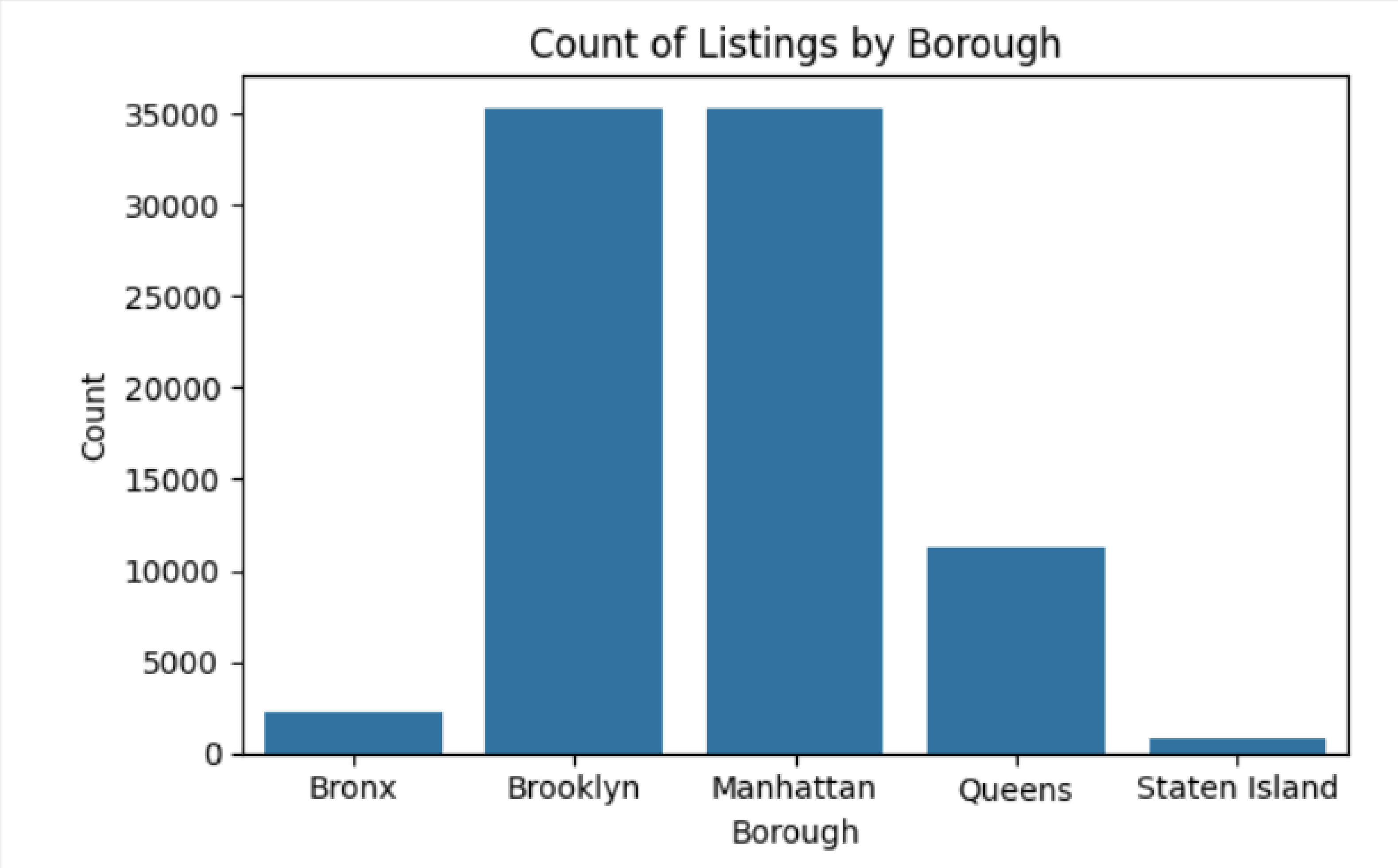
- Entire homes/apartments represent the majority of listings
- Private rooms form the second largest category
- Hotel rooms and shared rooms make up a smaller portion of the market

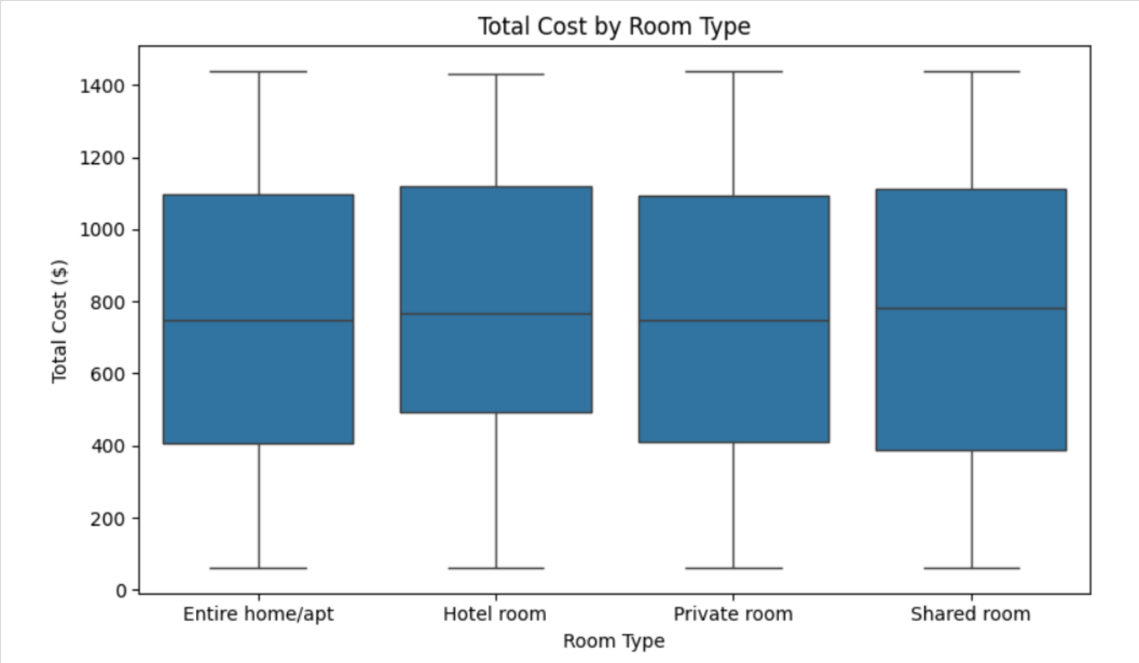


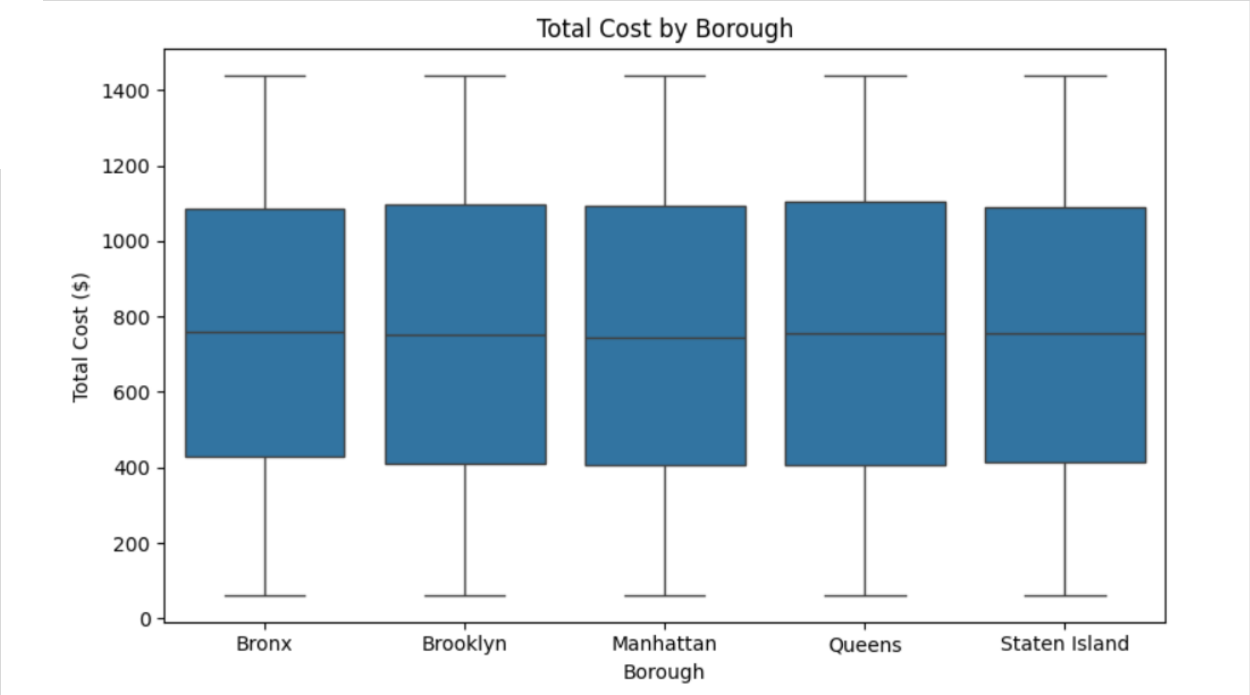
Key Findings:

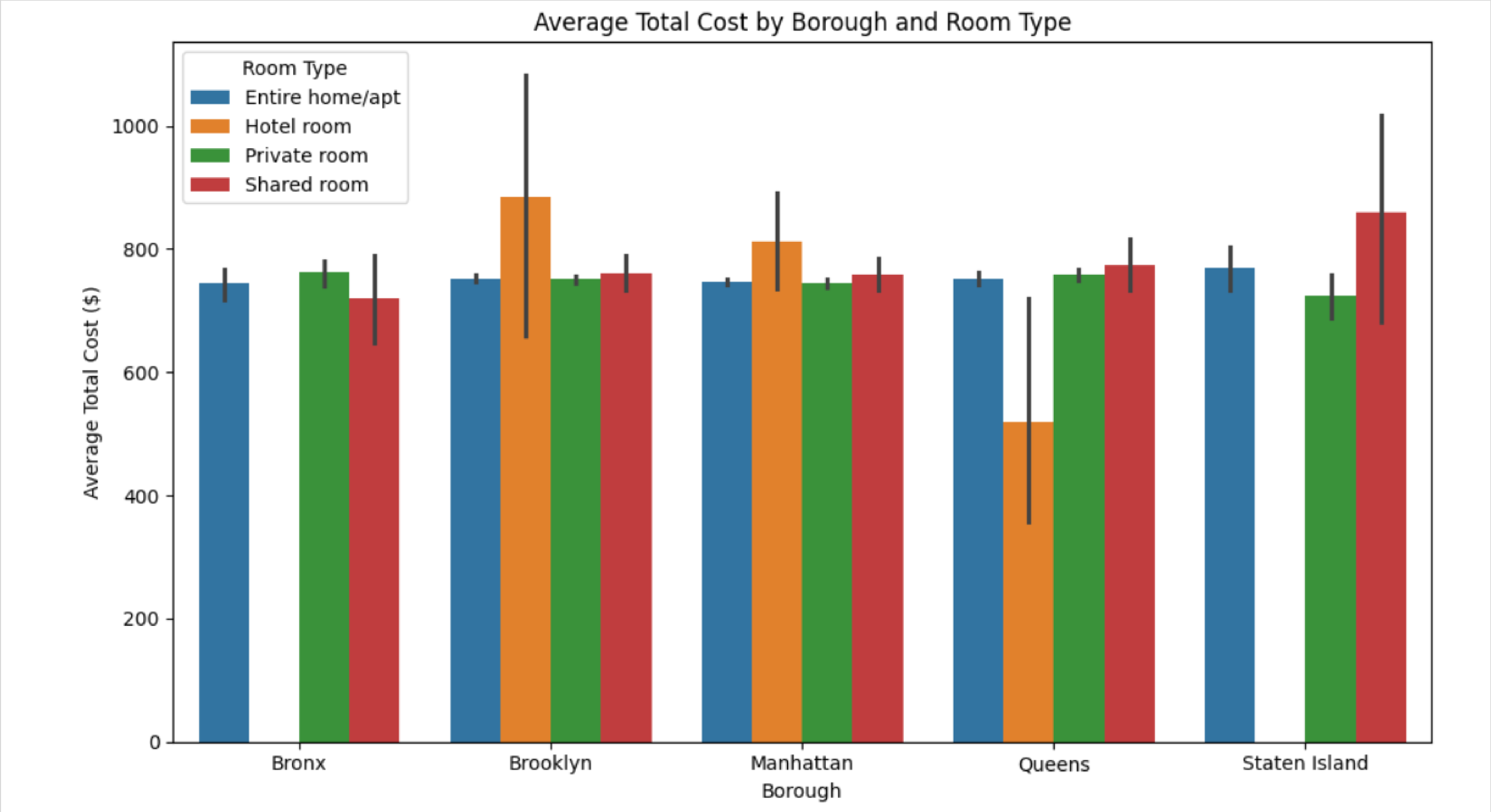
Geographic Distribution

- Manhattan and Brooklyn dominate the market in terms of listing volume
- Staten Island has significantly fewer listings compared to other boroughs
- Each borough shows distinct pricing patterns based on room types

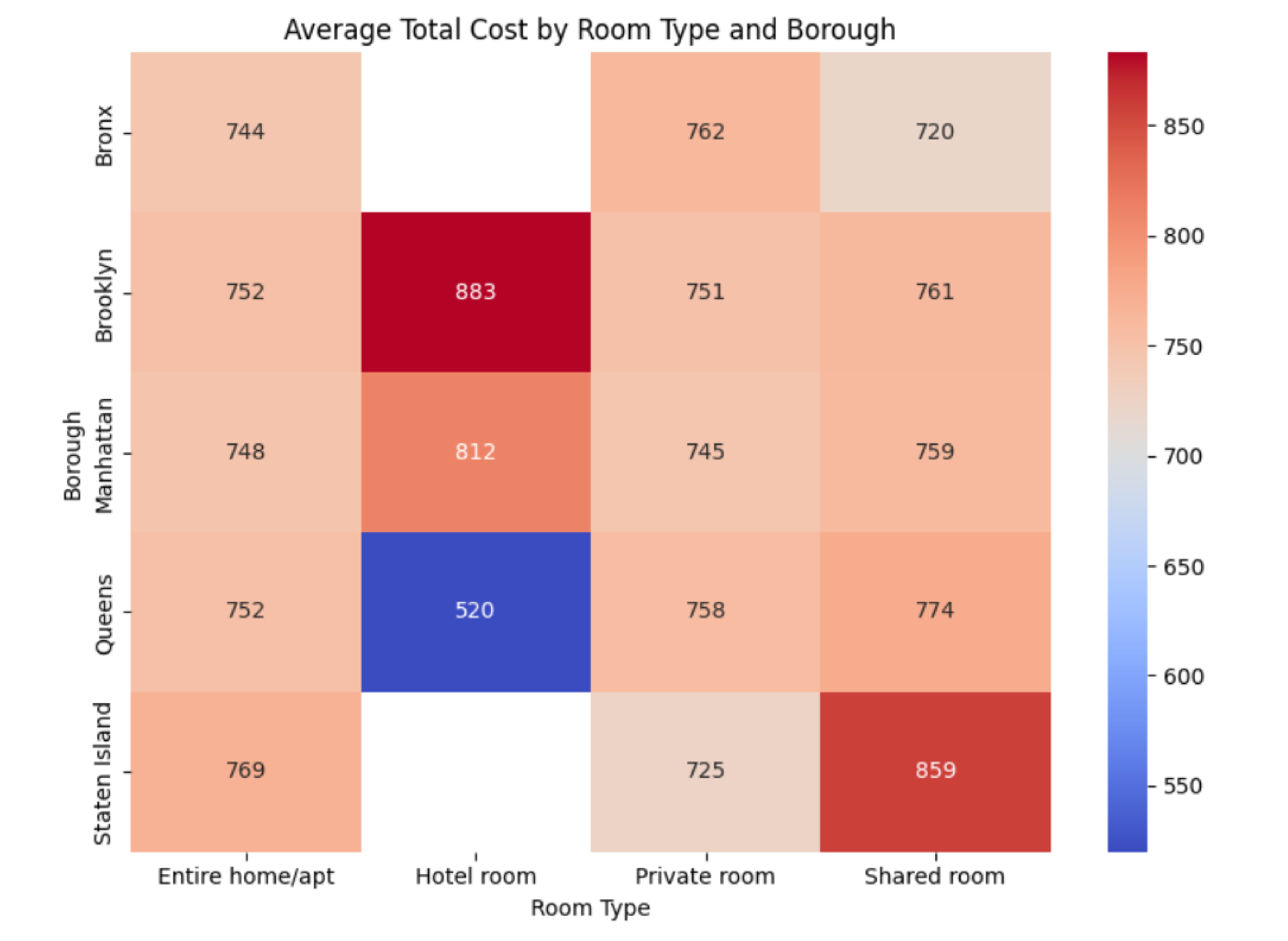



Total Cost by Room Type
Total Cost ($)
0
200
400
600
800
1000
1200
1400
Entire home/apt
Hotel room
Private room
Shared room
Room Type


Total Cost by Borough
Total Cost ($)
0
200
400
600
800
1000
1200
1400
Bronx
Brooklyn
Manhattan
Queens
Staten Island
Borough



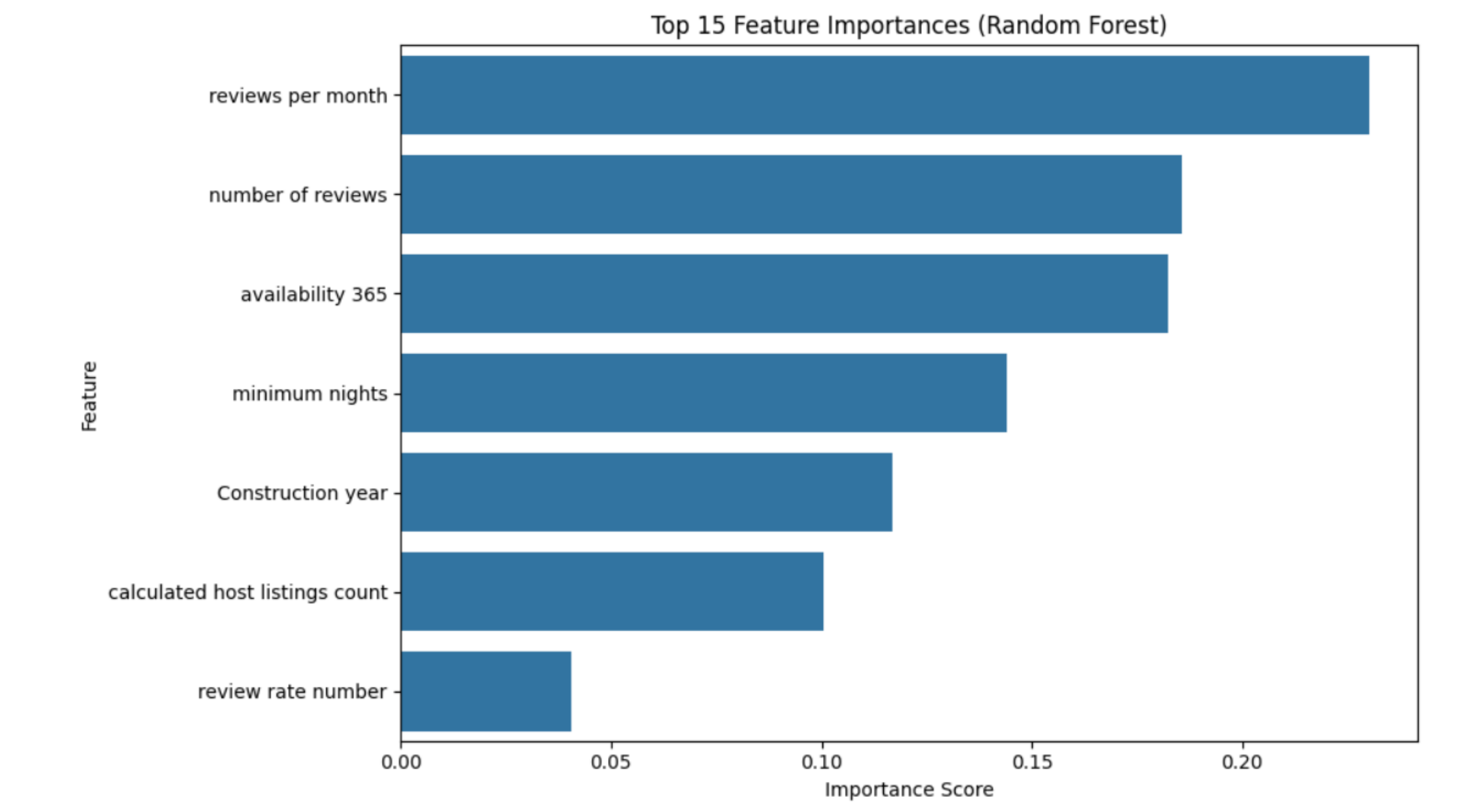
Average Total Cost by Borough and Room Type
Room Type
Entire home/apt
Hotel room
Private room
Shared room
Average Total Cost ($)
0
200
400
600
800
1000
Bronx
Brooklyn
Manhattan
Queens
Staten Island
Borough

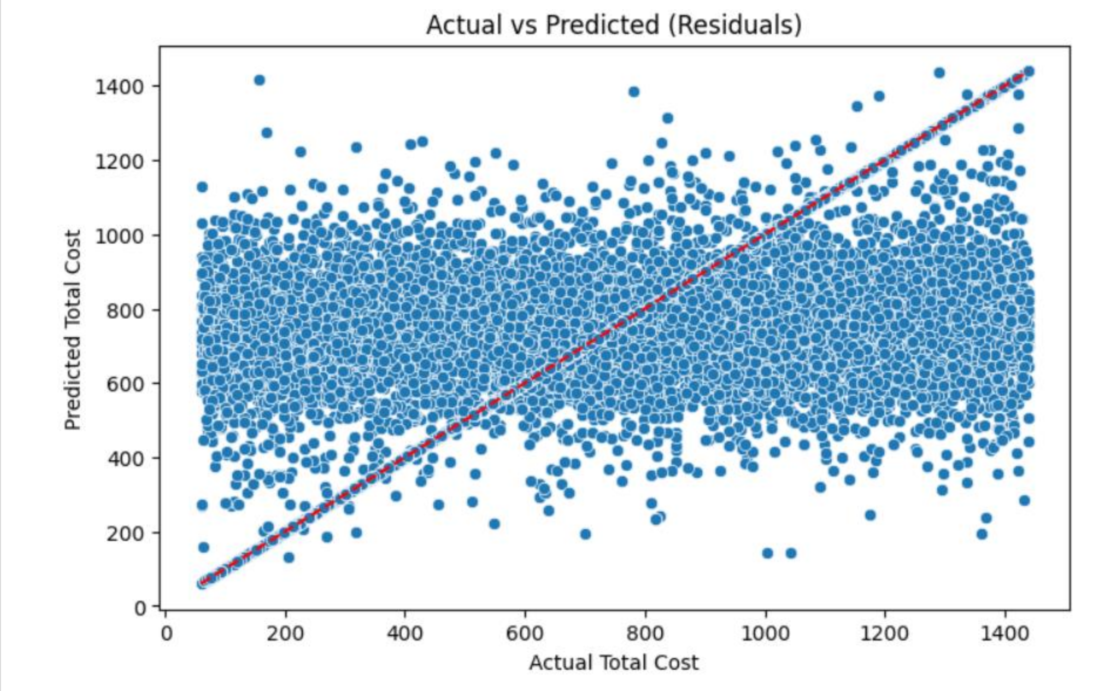

Average Total Cost by Room Type and Borough
Bronx
Brooklyn
Manhattan
Queens
Staten Island
Borough
744
762
720
752
883
751
761
748
812
745
759
752
520
758
774
769
725
859
Entire home/apt
Hotel room
Private room
Shared room
Room Type
850
800
750
700
650
600
550

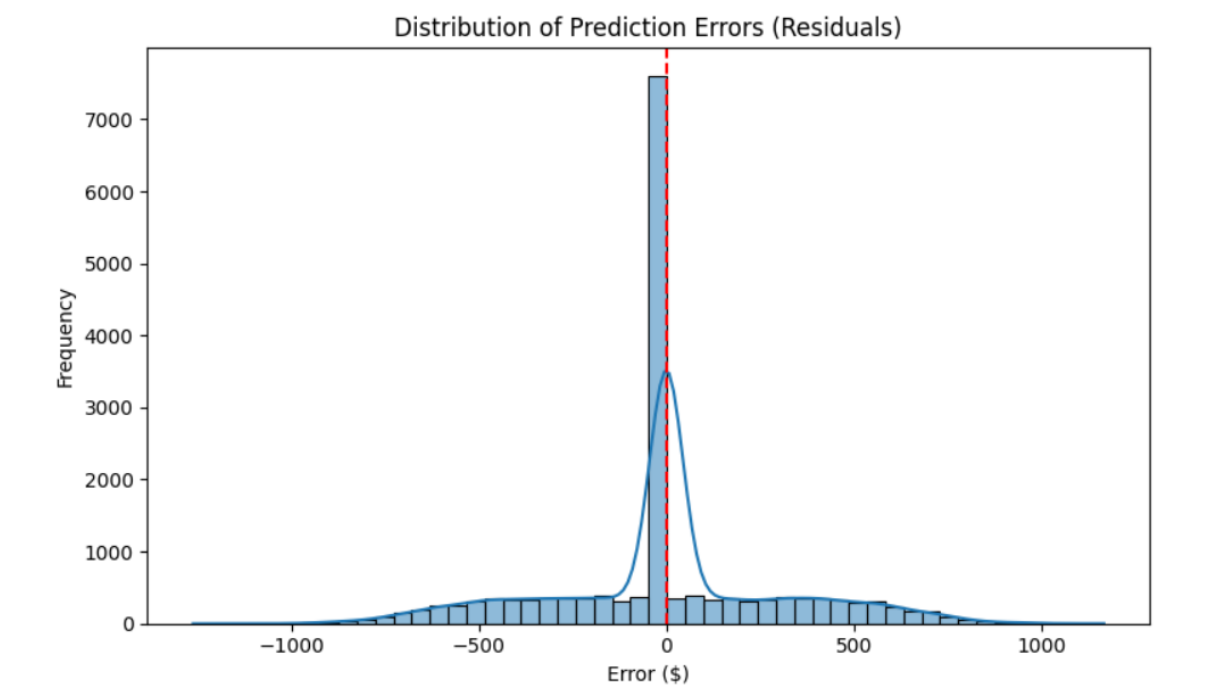
Key Findings:

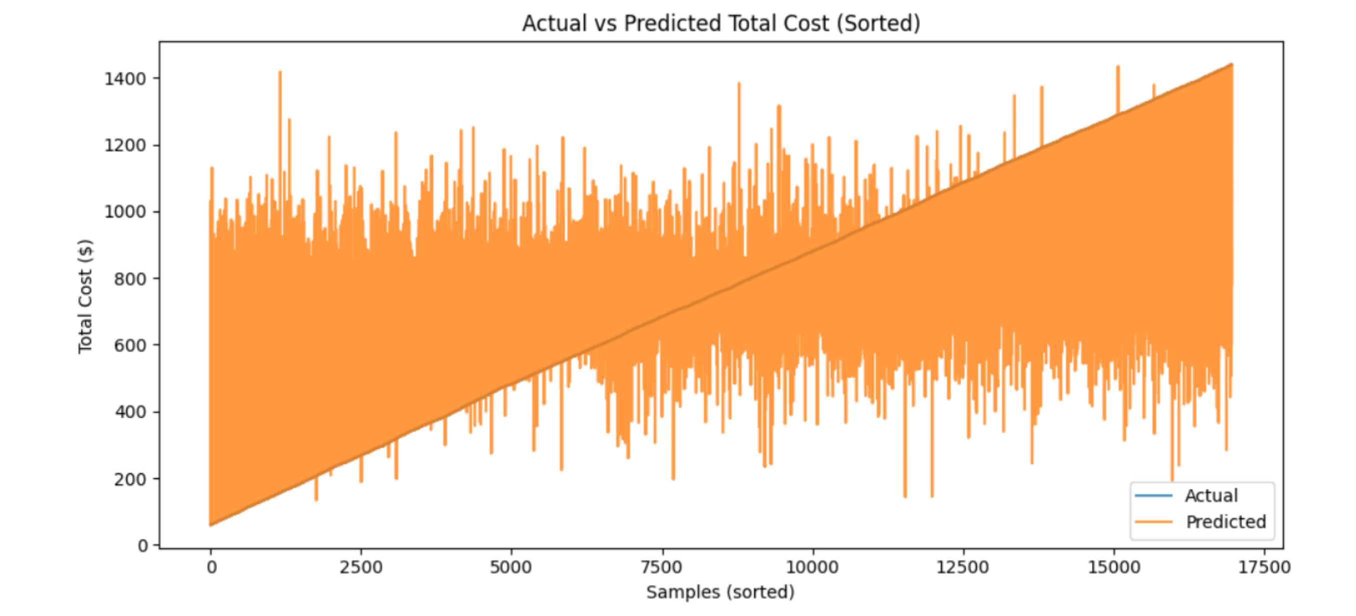
Price Variations

- Significant price variations exist between boroughs
- Room type is a major determinant of price
- The combination of borough and room type creates distinct price tiers

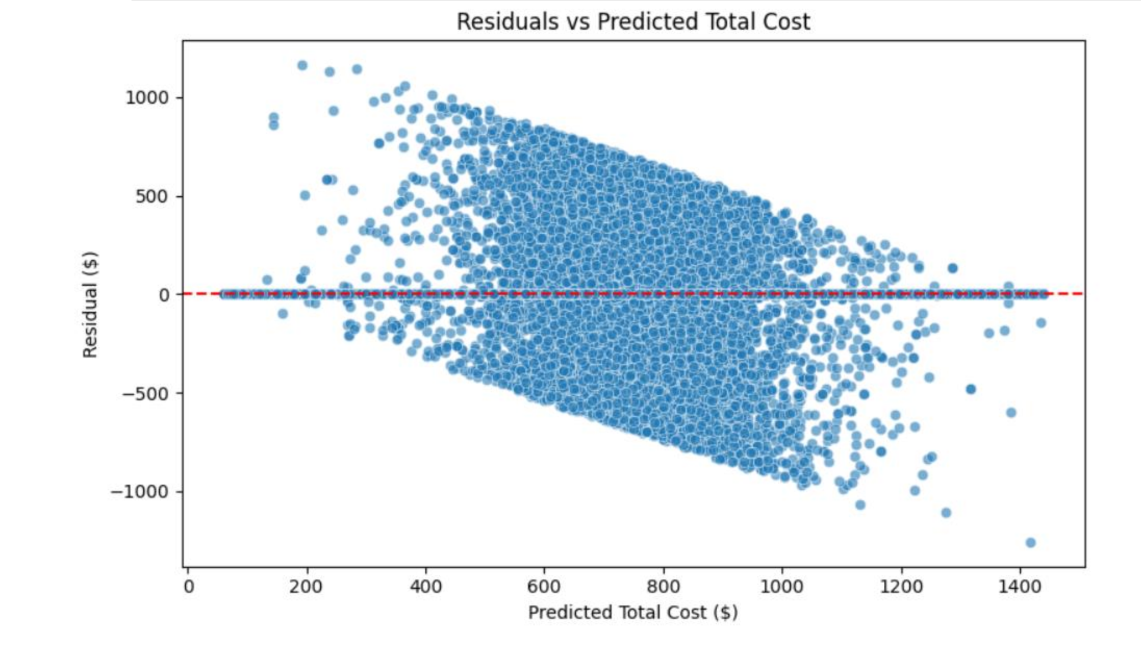



Actual vs Predicted (Residuals)
Predicted Total Cost
Actual Total Cost
0
200
400
600
800
1000
1200
1400


Distribution of Prediction Errors (Residuals)
Frequency
Error ($)
0
1000
2000
3000
4000
5000
6000
7000
−1000
−500
500



Actual vs Predicted Total Cost (Sorted)
Total Cost ($)
1400
1200
1000
800
600
400
200
0
0
2500
5000
7500
10000
12500
15000
17500
Samples (sorted)
Actual
Predicted



Residuals vs Predicted Total Cost
Residual ($)
1000
500
0
−500
−1000
0
200
400
600
800
1000
1200
1400
Predicted Total Cost ($)

Overall Results

We compared multiple models including Linear Regression, XGBoost, Random Forest, and ExtraTrees. ExtraTrees performed the best.

Strengths

- Model performs well overall – it captures general patterns in total cost across listings
- Prediction errors are centered around 0 with no major bias in over or under-prediction
- Most predictions are accurate within $200

Weaknesses

- Model struggles with very expensive listings – tends to underpredict the high end
- Guest activity features (reviews, availability) are more important than location
- Some predictions are too "average" – doesn’t always capture extreme cases

Predictive Modeling Results

Using an ExtraTrees Regressor model:

- Mean Absolute Error: $200.22
- Root Mean Square Error: $311.63
- R^2 Score: 0.3839

Model Performance Analysis

Strengths:

- Balanced predictions around the true values
- Reasonable accuracy for typical price ranges
- No systematic bias in predictions

Limitations:

- Lower accuracy for high-priced listings
- Moderate R^2 score indicating room for improvement
- Some difficulty capturing extreme price cases

Feature Importance

The model identified several key pricing factors:

- Construction year
- Minimum nights requirement
- Review-related metrics
- Borough location
- Room type

Recommendations

1. Price optimization strategies should consider both borough and room type
2. Different pricing models may be needed for luxury/high-end properties
3. Focus on accurate pricing in the most common price ranges
4. Consider seasonal variations in future analyses

Methodology Notes

The analysis employed various visualization techniques including histograms, box plots, heat maps, and scatter plots to understand the data distribution and relationships. The predictive modeling phase used the ExtraTrees algorithm with an 80/20 train-test split. We selected ExtraTrees due to its superior performance across all metrics during model comparison.