# Comparative Analysis of NYC Airbnb Rentals

# 5000 Data Analytics Final Project Report

# By: Thomas Nash & Daniel Mehta

# 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 price prediction modelling. 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:

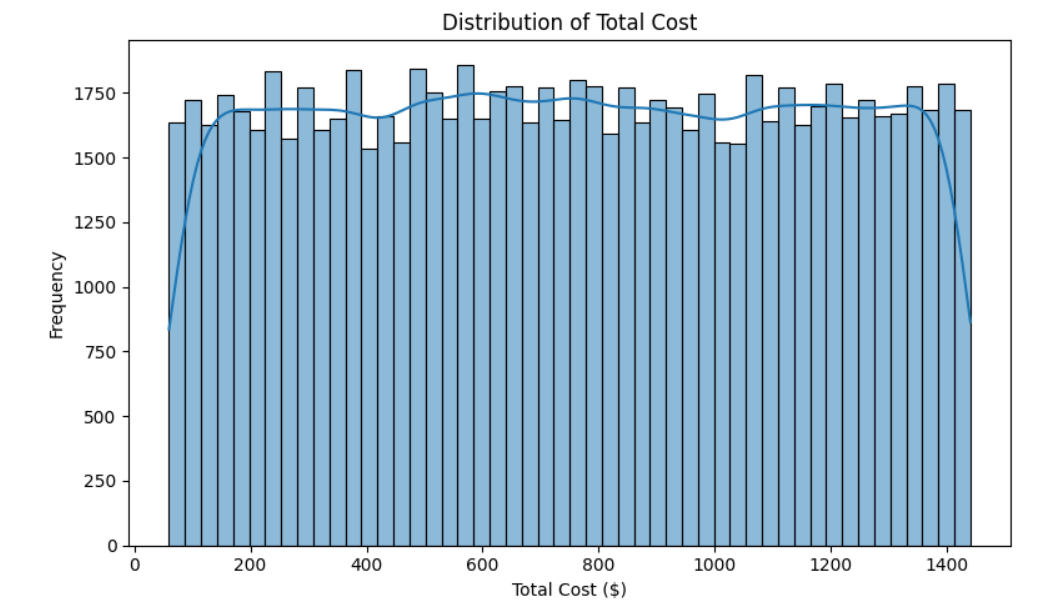
* Removal of irrelevant columns, including ID fields, geographical coordinates, and text-based descriptions
* Standardization of price and fee columns by removing currency symbols and converting to float values
* Treatment of missing values in critical columns
* Encoding of categorical variables using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

## Key Findings

### Distribution Patterns

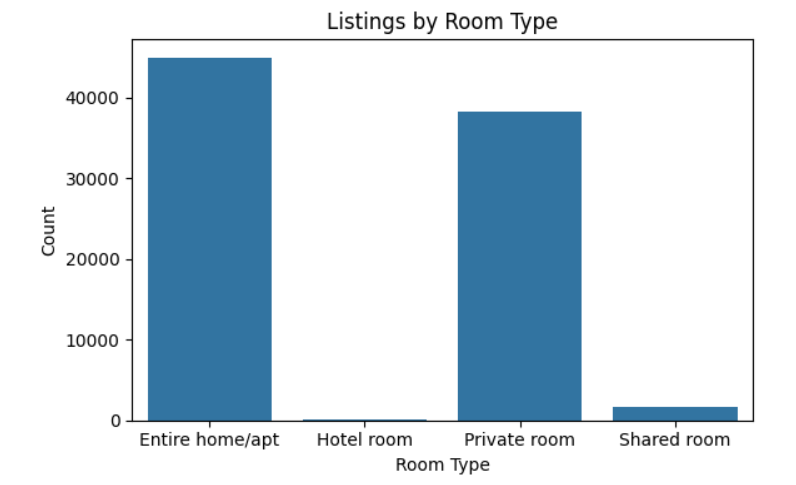
* The total cost distribution shows a right-skewed pattern, with most listings clustered in the lower price ranges
* Manhattan emerged as the borough with the highest average costs
* Entire homes/apartments consistently commanded higher prices across all boroughs

## 



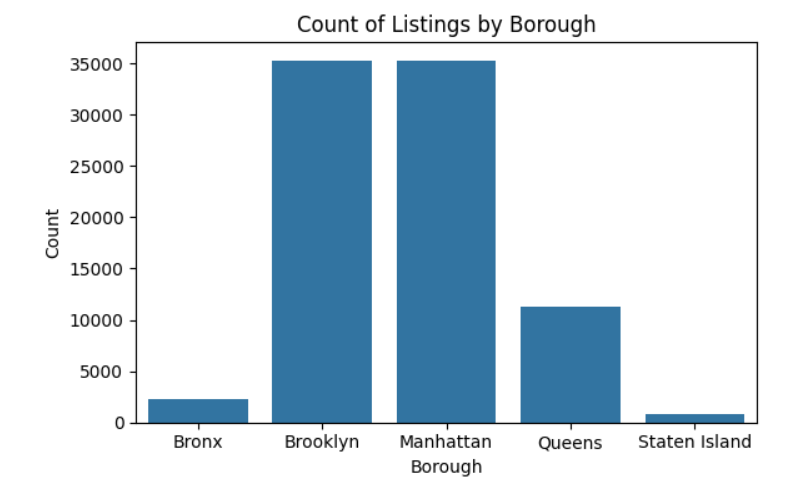
### Room Type Analysis

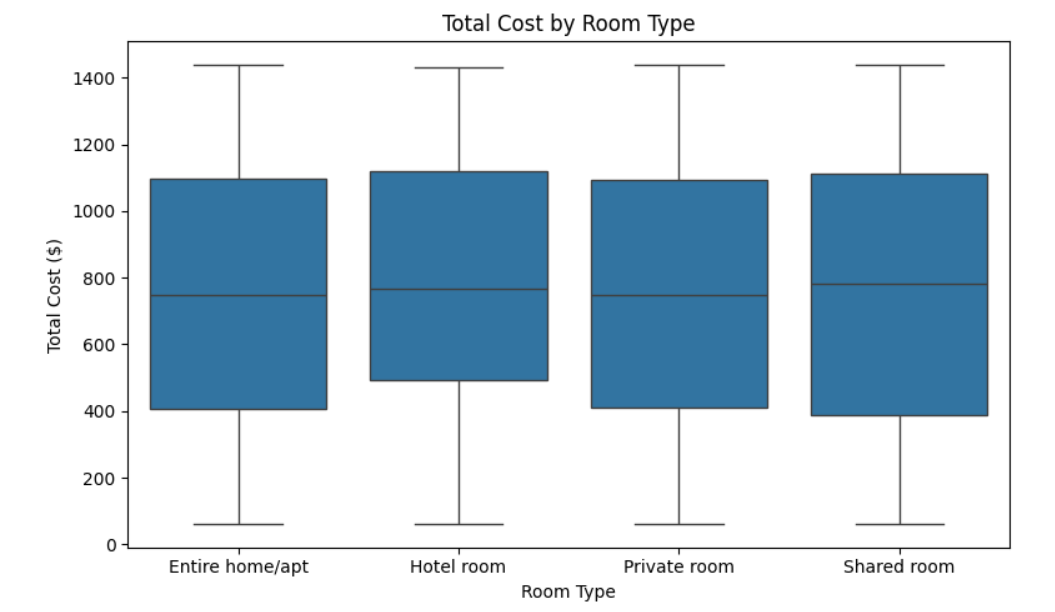
* Entire homes/apartments represent the majority of listings
* Private rooms form the second largest category
* Hotel rooms and shared rooms make up a smaller portion of the market

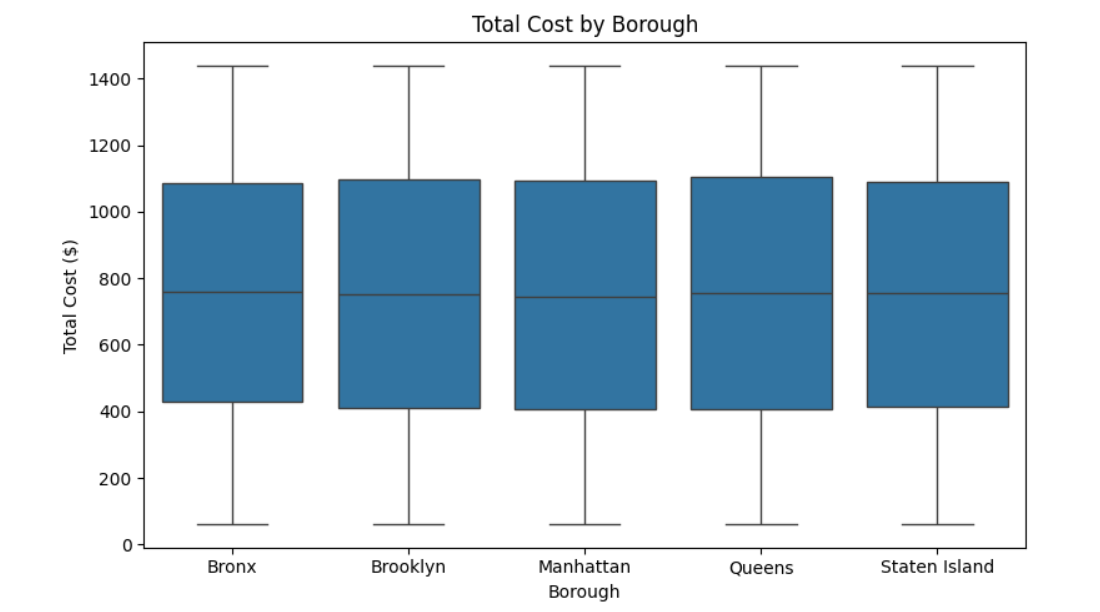


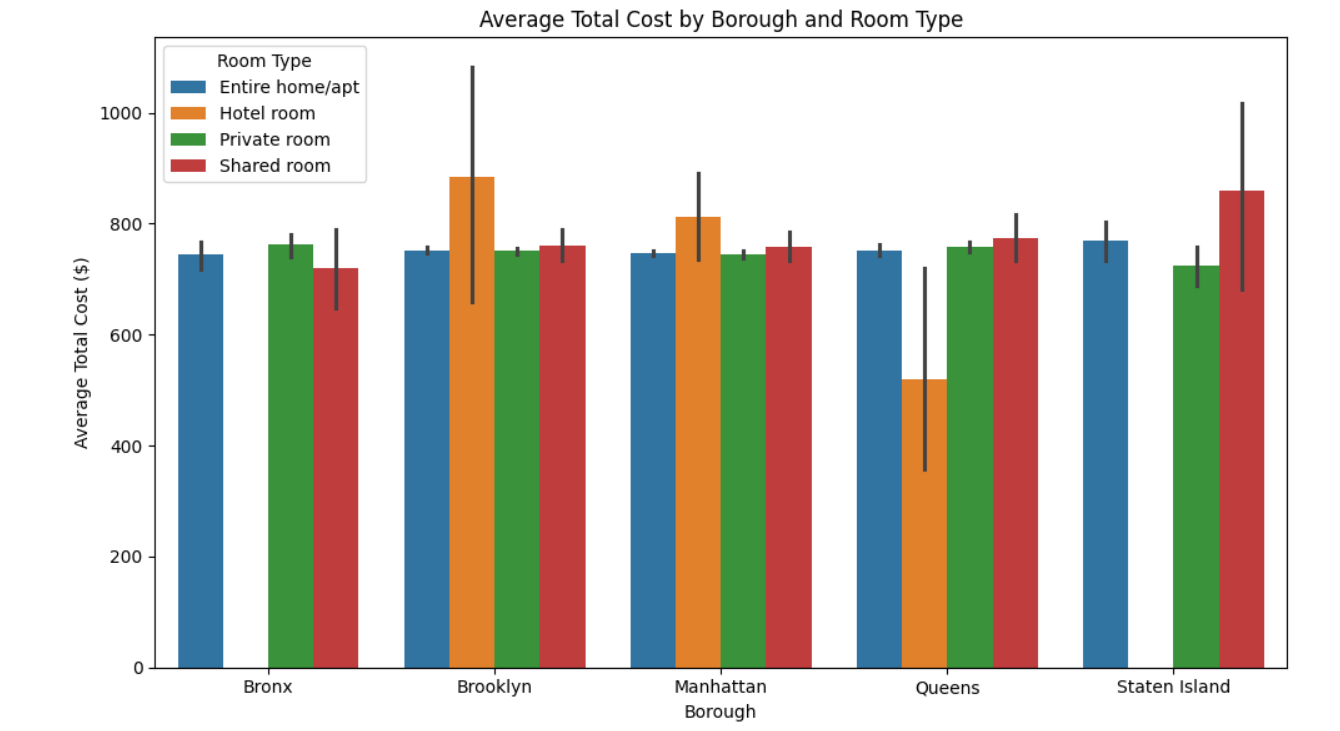
### Geographic Distribution

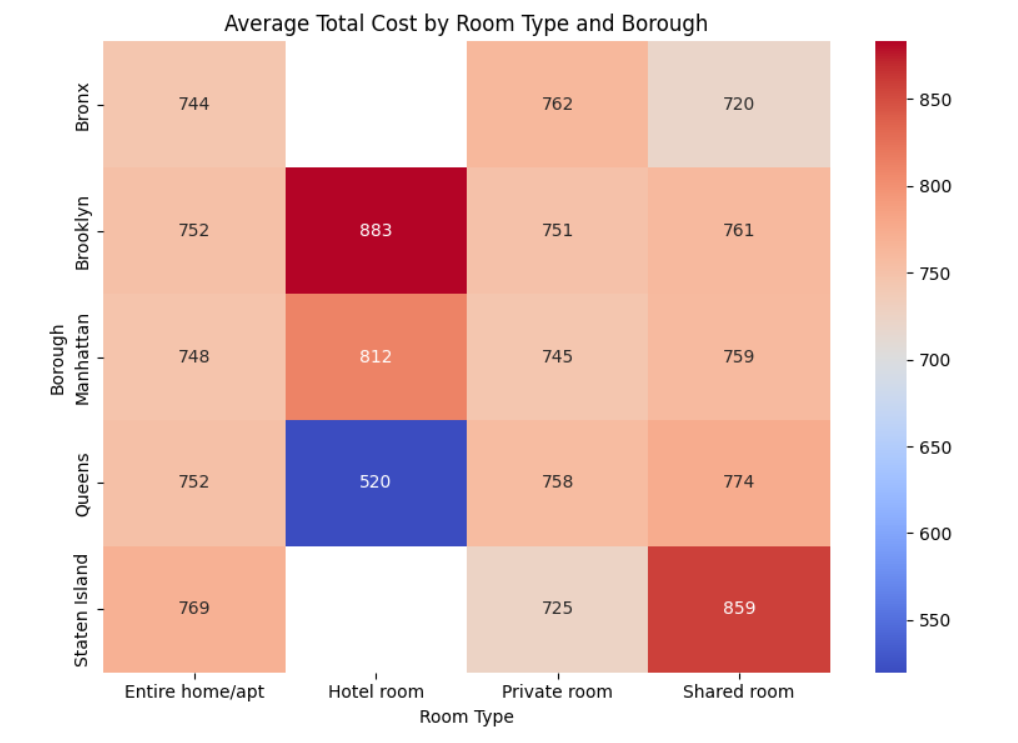
* Manhattan and Brooklyn dominate the market in terms of listing volume
* Staten Island has significantly fewer listings compared to other boroughs
* Each borough shows distinct pricing patterns based on room types





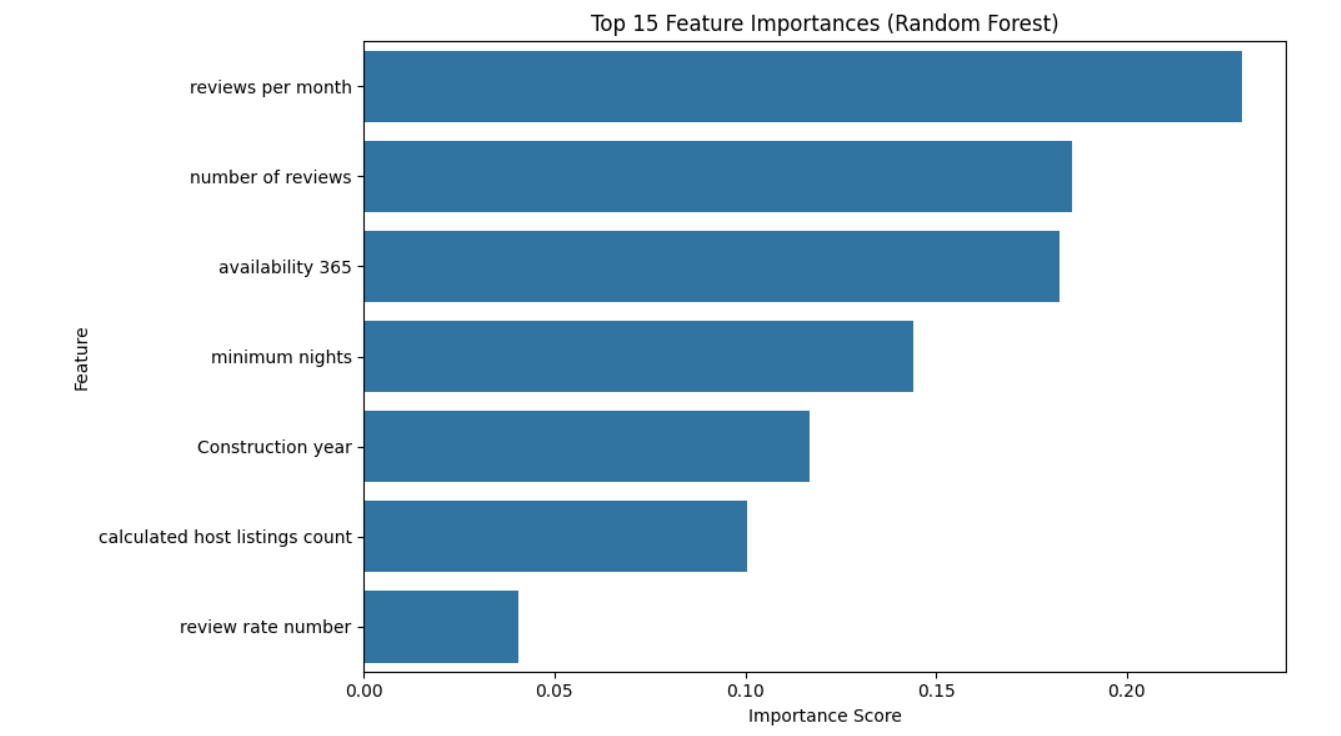


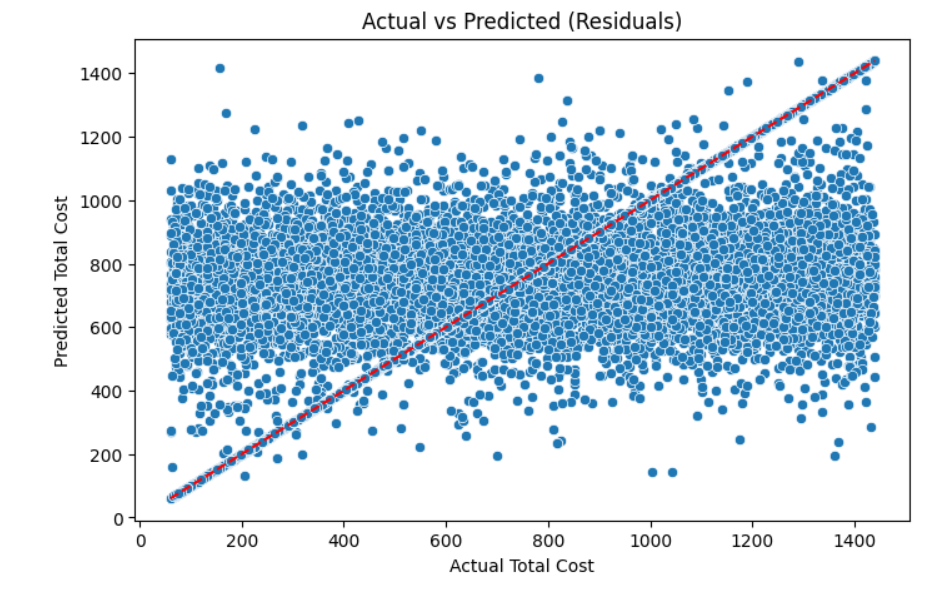


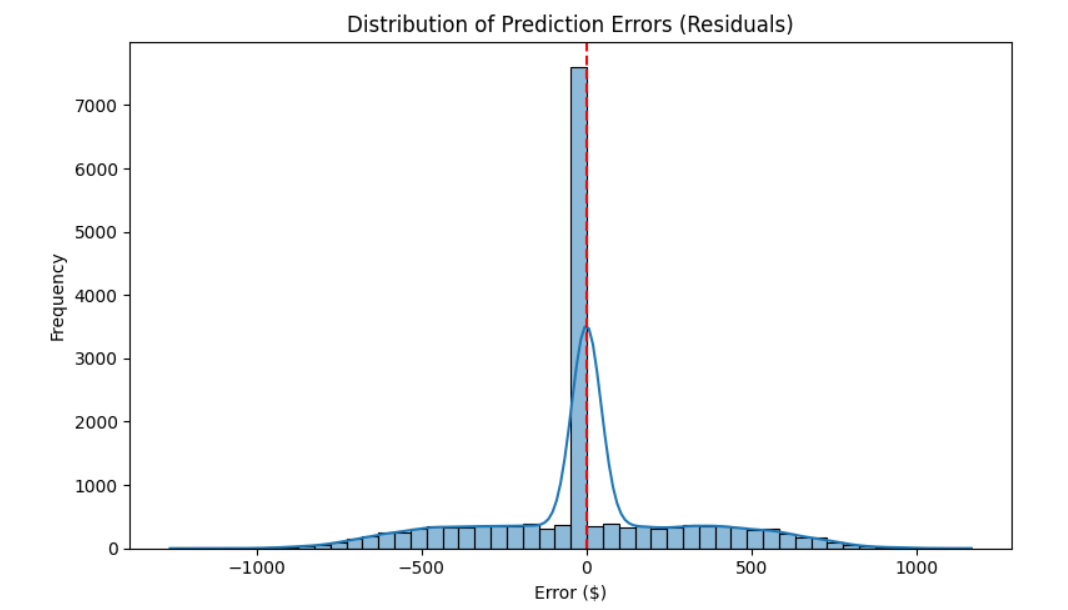


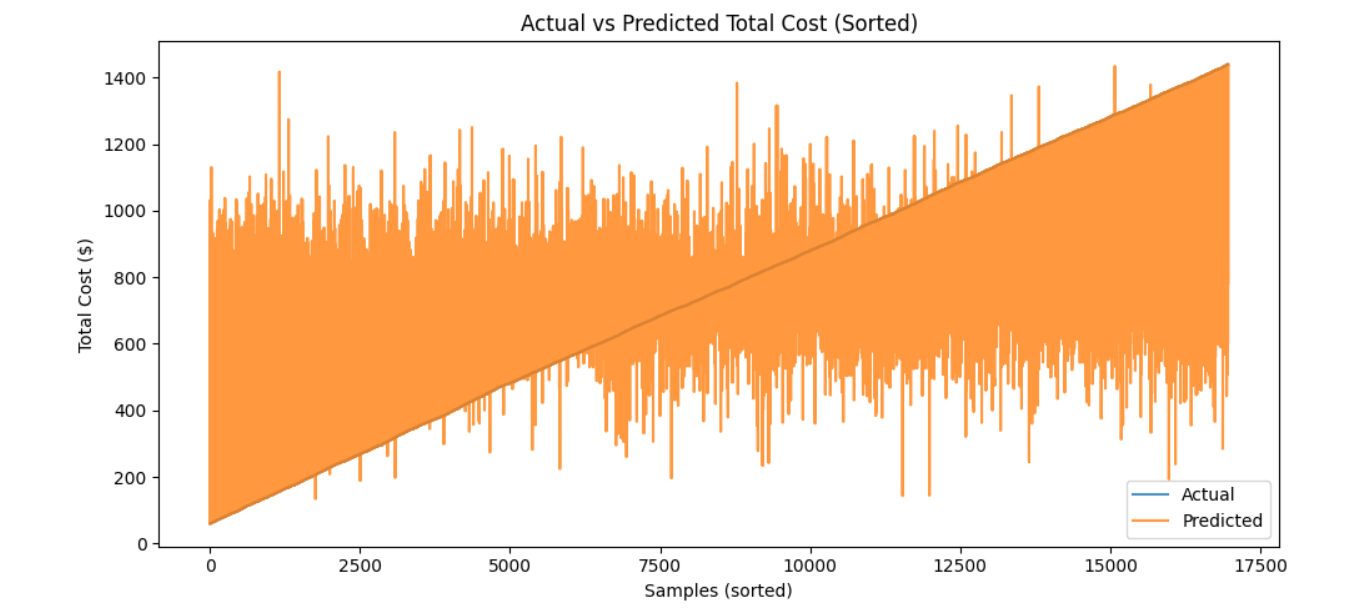
### Price Variations

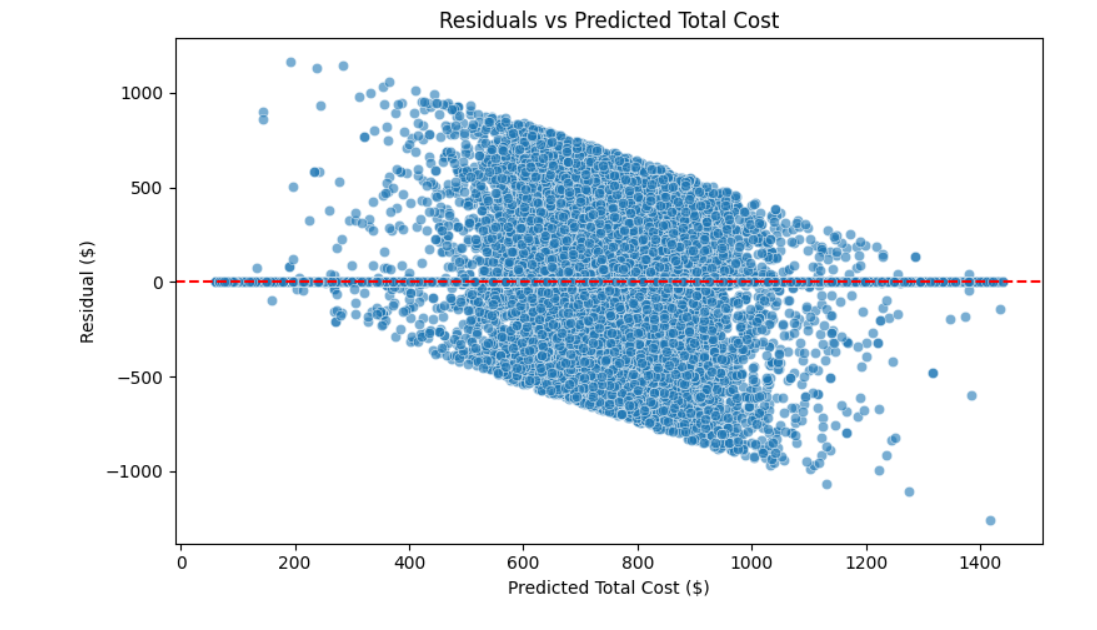
* Significant price variations exist between boroughs
* Room type is a major determinant of price
* The combination of borough and room type creates distinct price tiers











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 zero 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 Modelling Results

Using an ExtraTrees Regressor model:

* Mean Absolute Error: $200.22
* Root Mean Square Error: $311.63
* R² 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
* A moderate R² score indicates room for improvement
* Some difficulty in capturing extreme price cases

### Feature Importance

The model identified several key pricing factors:

* Construction year
* Minimum night 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
5. Consider seasonal pricing models in future analysis to capture time-based demand shifts.

## 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.