Management Summary
Airbnb Price Analysis

Viktor Hoffmann Rahul Maddineni Stefan Merdian Teresa Schuch

1 Overview

This project aimed to uncover key factors influencing Airbnb prices in New York City, focusing on the impact of Points of Interest (POIs), accessibility to public transportation, and neighborhood crime rates. Using clustering techniques, areas were segmented by Airbnb locations, and patterns were analyzed in relation to external predictors. Machine learning models quantified these relationships, providing insights into how location-based factors shape Airbnb pricing trends.

2 Data Sets

To address our research questions, we utilized various datasets, including Airbnb listings, Points of Interest, parks, and crime data. Although a transportation dataset was initially included, it was later excluded as its key features were already captured in other datasets. Using clustering techniques, we segmented New York City areas based on Airbnb prices and popularity. These clusters were analyzed alongside external factors such as amenities, landmarks, and accessibility to identify patterns and correlations.

3 Challenges

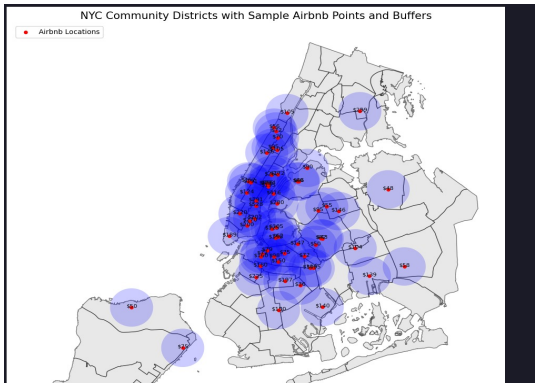
To begin with, the task itself was challenging to begin with as none of us had adequate experience in performing location analysis and incorporating it as a key step in the assignment pipeline. The project posed several other challenges that required adaptive problem-solving strategies. Firstly, the lack of unique identifiers across datasets created difficulties in aligning Airbnb data with points of interest, crime statistics, and transportation access. This was mitigated by developing custom scripts to process and clean data iteratively. Another significant hurdle was the high volume of missing data in key columns like "price," "beds," and "review scores," which restricted the dataset's completeness. Strategies such as imputing values for specific columns or filtering incomplete records helped manage this issue, albeit at the cost of reducing sample size.

Additionally, processing geospatial data proved computationally demanding, particularly during clustering and buffer creation. Optimization techniques like spatial indexing were employed to enhance efficiency. Lastly, challenges emerged from unbalanced review scores, where most listings had high ratings, complicating the analysis of external factors. This imbalance was addressed by focusing on normalized and mean-based scores to better capture the variations in listing quality and customer satisfaction.

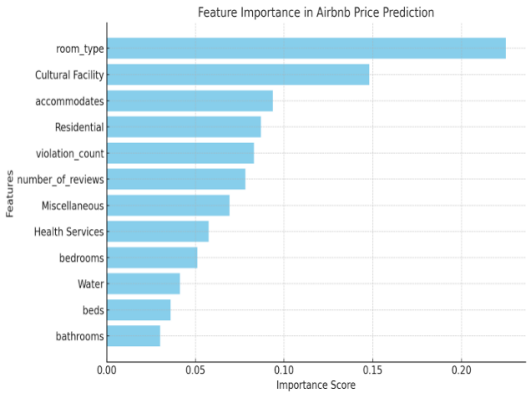
The merged dataset also provided challenges of its own which were eventually addressed by scaling and eliminating highly collinear features.

4 Insights

Clustering techniques and geospatial data handling was instrumental in understanding of pricing dynamics. By segmenting listings based on their proximity to points of interest (POIs), crime statistics, and amenities, the analysis uncovered clear patterns. Areas with a high density of cultural facilities or recreational zones were consistently associated with higher prices. Buffer zones and spatial joins with external datasets as shown in Figure 1, enabled the aggregation of critical statistics within defined radii, further strengthening the predictive capabilities of the model.



(a) Districts with Airbnb Points and Buffers



(b) Feature Importances

The analysis revealed that certain features (Figure 2) had a significant impact on Airbnb prices in New York City, with room_type, cultural facilities, and accommodates emerging as the most influential factors. Listings offering private or entire spaces, proximity to cultural landmarks, and the capacity to accommodate more guests consistently correlated with higher pricing. In contrast, features such as bathrooms, beds, and water had less predictive power, indicating that while relevant, they played a less critical role in the model's ability to predict prices.

5 Conclusion

The project highlighted the importance of structured, computationally feasible location analysis. While limited by the availability of free data, the analysis identified room type, cultural facilities, and accommodates as key factors influencing Airbnb prices, with room type having the highest importance (0.2249). Crime statistics also impacted prices, while public transportation's role was inconclusive due to collinearity. Geospatial clustering showed higher prices in amenity-dense areas. The model achieved reasonable accuracy with an RMSE of 0.5 and a 9% prediction error, effectively capturing key pricing drivers.