Exploratory Data Analysis and Price Prediction of Airbnb in New York City

Jiani Lyu, Chelsea Liu, Yuewei Shi, Yanchi Jin, Yajie Zeng

Abstract

This study delves into the intricate dynamics of Airbnb pricing in New York City, aiming to unearth key determinants that influence listing prices and develop robust predictive models. Leveraging a dataset featuring a diverse array of Airbnb listings across the city, this research integrates variables such as geographical coordinates, room types, pricing, and amenities with additional data on rental income and transportation accessibility. Advanced machine learning techniques, including decision trees, random forests, gradient boosting machines, and support vector machines, were employed to model and predict Airbnb prices. Significant predictors identified include proximity to transit, housing type, and neighborhood characteristics. The findings emphasize the importance of utilizing sophisticated data sources and analytical methods to enhance prediction accuracy, providing valuable insights for hosts and policymakers engaged in strategic decision-making.

Keywords: Airbnb, predictive modeling, urban housing, machine learning, New York City

1 Introduction

Airbnb has become a pivotal element of the sharing economy, especially in urban settings like New York City, where it plays a substantial role in shaping the accommodation market. This platform not only provides alternative lodging options for tourists but also presents economic opportunities and challenges for local residents and policymakers. As such, understanding the determinants of Airbnb pricing is crucial for stakeholders to make informed decisions that balance benefits and disruptions within the urban housing market.

This research focuses on analyzing Airbnb pricing dynamics in New York City. The study is driven by three primary objectives: firstly, to identify key factors that significantly influence the prices of Airbnb listings; secondly, to evaluate the impact of transportation accessibility on Airbnb pricing, recognizing its importance in urban travel and accommodation choices, and incorporating it into our predictive models as a significant variable; and thirdly, to use these

insights to develop models capable of accurately forecasting price fluctuations. The analysis leverages a comprehensive dataset of Airbnb listings, emphasizing intrinsic attributes such as location, room type, and amenities, as well as extrinsic factors like accessibility to public transportation.

This research is crucial for enabling Airbnb hosts to optimize pricing strategies and for policymakers to understand the platform's impact on urban housing markets.

2 Literature Review

The literature exploring factors affecting Airbnb pricing in urban environments like New York City underscores the influence of multiple interrelated elements. Research incorporating geospatial analysis and machine learning models has consistently shown that proximity to key urban amenities, such as transportation hubs, can significantly impact rental prices, emphasizing the value that guests place on accessibility (Zhang & Gao, 2019).

Property attributes, including size and quality, also play a critical role. Studies from diverse urban contexts demonstrate that larger and higher-quality accommodations command higher prices, a trend that is observable across various international cities and likely pertinent to New York City as well (Miller & Brown, 2019; Harper & Lim, 2021).

Host-related factors introduce another layer of complexity. While characteristics like "Superhost" status might not significantly impact pricing in rural settings, their influence in a competitive urban market like New York City can be more pronounced, reflecting the nuances of consumer preferences and market saturation (Kisieliauskas, 2023).

Additionally, the broader economic and social environment can influence Airbnb pricing. Studies have linked economic indicators and social factors such as local events or seasonal variations to fluctuations in Airbnb pricing, suggesting a dynamic interplay between local economic conditions and short-term rental prices (Jensen et al., 2020; Lee & Nguyen, 2019).

Moreover, the technological amenities offered within Airbnb listings have been identified as key price determinants, with modern and technologically equipped homes attracting higher rental premiums (Kim & Park, 2021).

These studies collectively illustrate the complex and multifaceted nature of Airbnb pricing in urban areas, highlighting the importance of a comprehensive approach that considers geographical, property-specific, host-related, and socio-economic factors.

3 Dataset

3.1 Dataset Overview

- Airbnb listings data of NYC: The study employs a comprehensive dataset of Airbnb listings in New York City, featuring various attributes including location, room type, pricing, and reviews. This dataset is enriched with the DOF Condominium Comparable Rental Income Dataset, which provides additional insights into the rental incomes and market values across NYC neighborhoods, thereby offering a more holistic understanding of the city's rental market dynamics.
- Transformation Data of NYC: In theory, transportation is an important factor affecting the price of accommodation. Therefore, this paper introduces some traffic data from the U.S. Environmental Protection Agency, including network density, population-weighted centroid distance to the nearest bus stop, and so on. At the same time, there are some useless variables, such as state FIPS codes and county FIPS codes. This article removes them before data processing. In addition, the paper added the latitude and longitude of each listing, allowing the study to perform map visualization analysis and match each listing to traffic conditions based on geographic coordinates.

3.2 Data Cleaning and Preprocessing

Prior to feature engineering, data cleaning and preparation are very important processes. In order to increase the robustness of the model, the effects of off-season and off-season should be excluded. Therefore, this paper selects data from March 2023 to February 2024 and takes the average price during that time period as the target object.

There are several important steps in the data cleansing process. First, columns with more than 70% missing values are excluded because too many missing values can affect the results of the model. Then, the outliers are eliminated using the isolated forest method. Because it does not always exclude all outliers, this article uses a box chart to check the results. As you can see from the figure (Figure 1), only a few values are greater than 2000. Therefore, these outliers are excluded in this paper, and good results are obtained.

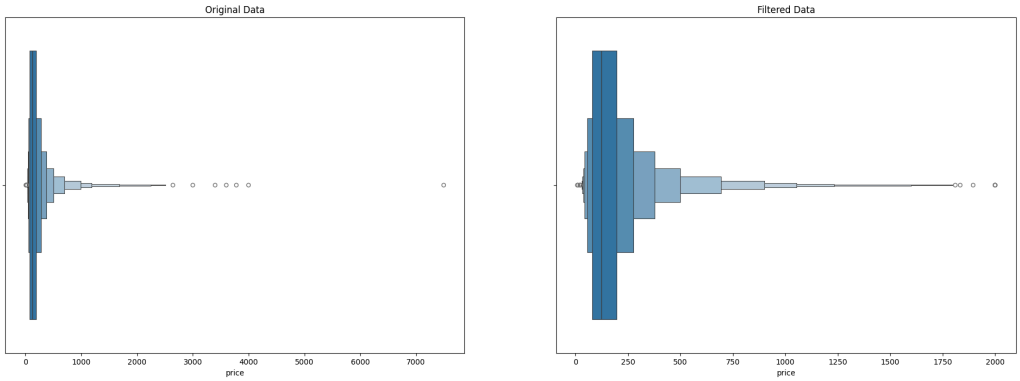

Figure 1 Boxen Plot of Price

Third, this paper needs to understand the situation of these variances through the correlation heat map because it needs to pay attention to the independent variables and collinearity problems. On the one hand, the paper draws a heat map of the correlation between price and each variable, and if the correlation between the two is small, it means that the variable is probably an independent variable, which they can remove in the future. As you can see from Figure 2, there is always some correlation between these variables and prices. Therefore, this paper keeps all variables for the time being and plans to remove the independent variables through feature engineering. On the other hand, this paper draws a heat map of the correlation between all the variables to see if there is a collinearity problem. As can be seen from the figure (Figure 3), there is a strong correlation between some variables. Therefore, this article uses VIF to check and remove these variables with serious collinearity problems.

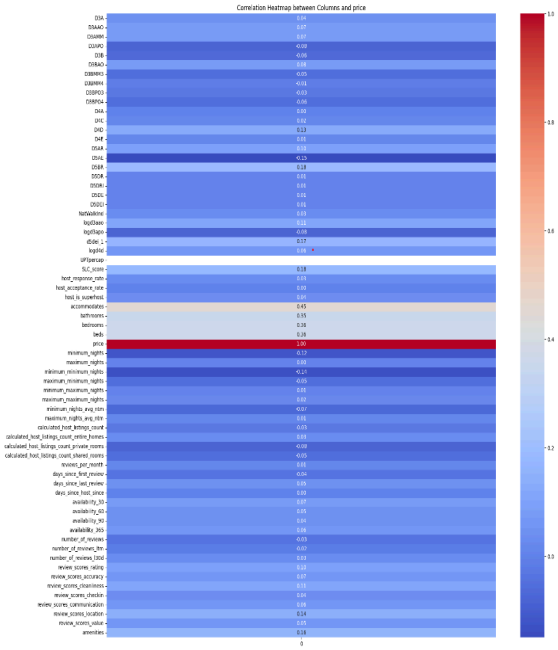

Figure 2 Correlation Heatmap between Columns and Price

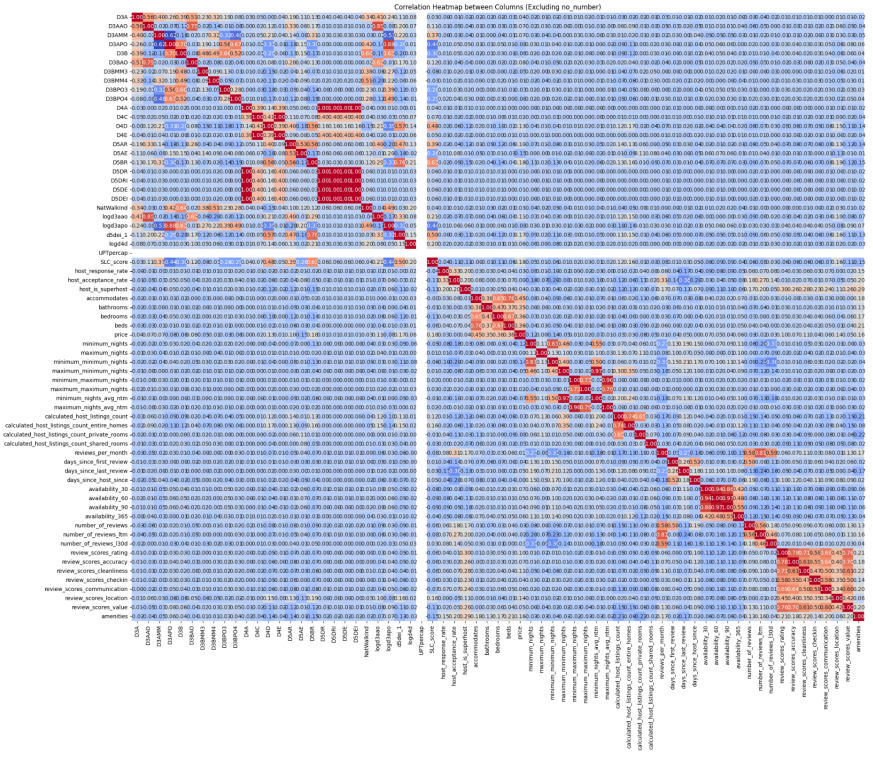

Figure 3 Correlation Heatmap between Columns

Fourth, this paper encodes the variables whose contents are non-numeric. Before encoding, this article needs to know each variable (Figure 4) to determine the encoding method. As can be seen from the figure, some variables contain only "True" and "False", while others contain different text. Therefore, for variables containing only "True" and "False", this paper uses label encoding; For variables containing different text, this article uses one-hot encoding.

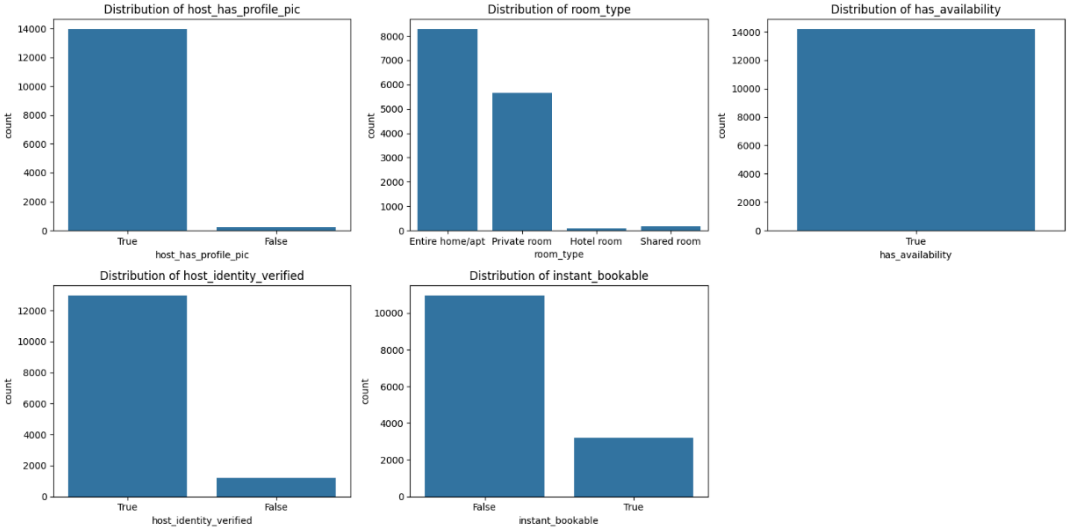

Figure 4 Distribution of the Variables Factors that Need to be Coded

3.3 Exploratory Analysis and Visualization

Then EDA is carried out to facilitate feature selection and model processing. First, this paper observes the distribution of prices. As can be seen from the figure (Figure 5), the distribution of price variables is uneven and presents a left-leaning shape, so this paper recorded it (Figure 6), and the result is roughly normal distribution. Therefore, subsequent dependent variables are logged.

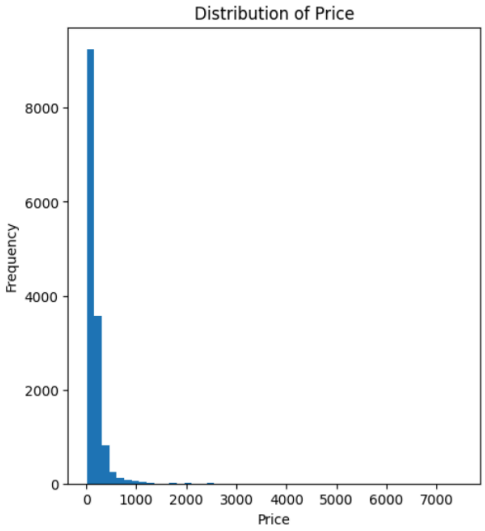

Figure 5 Distribution of the Original Price

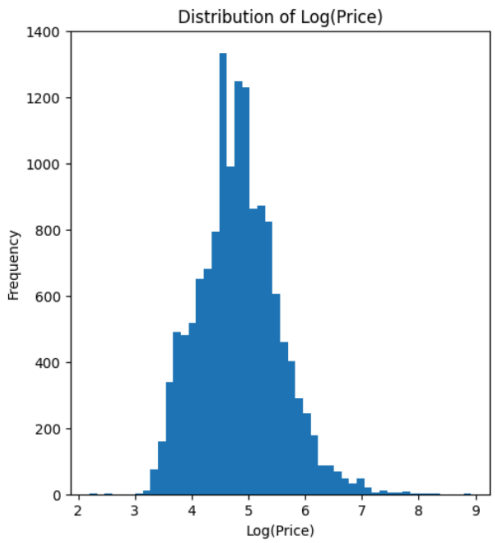

Figure 6 Distribution of the Log (Price)

The following is a general discussion on price distribution. The figure (Figure 7) shows that there are many houses in the 500-1000 price range, and the distribution of each price is relatively dispersed, with no obvious agglomeration.

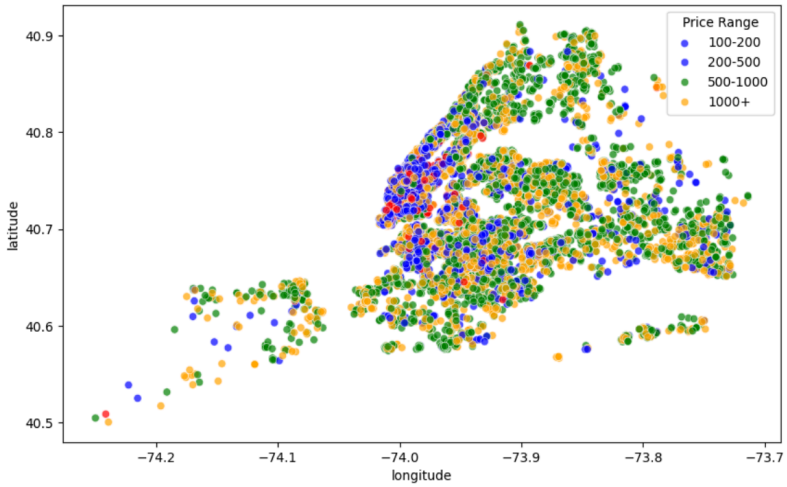

Figure 7 The Price of Each Listing on Map

Finally, this paper selects the variables that have a greater impact on the housing price in the theoretical sense from many variables to provide reference for later feature projects. This paper selected host response rate, host acceptance rate, walking index, street line convenience, number of accommodations, number of bathrooms, bedrooms, and beds, and review scores. Among them (Figure 8), high response rate, street line convenience score, and review score are positively correlated with price.

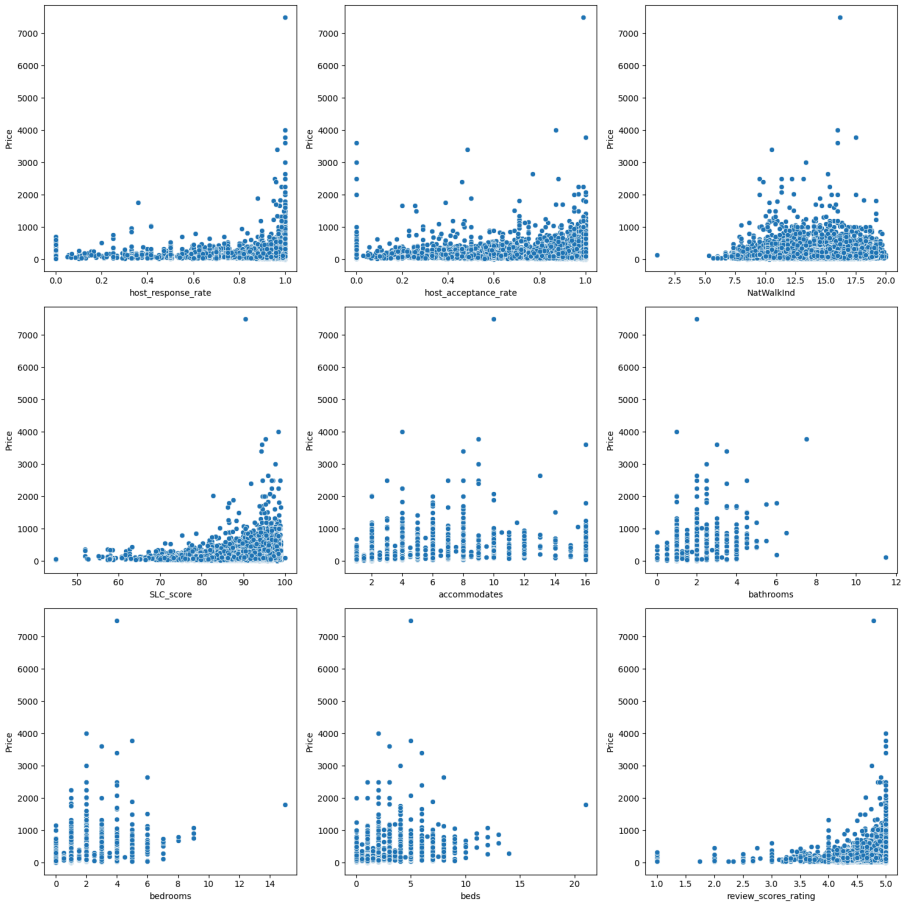

Figure 8 Correlation with Price and Important Variables

4 Methodology

4.1 Impact of Traffic Accessibility on Property Prices

To explore the relationship between traffic accessibility and house prices, and to understand how traffic variables influence housing prices, we employed cluster analysis as our primary method. This technique allows us to identify distinct pricing patterns under various traffic conditions within the housing market.

We utilized the K-means clustering algorithm for our analysis. This method was chosen due to its simplicity, efficiency, and applicability to various datasets. We standardized the housing prices and traffic variables to ensure each feature contributed equally to the analysis.

To determine the optimal number of clusters, we used both the Elbow Method and the Silhouette Score. The Elbow Method, illustrated in Figure 9, plots the number of clusters against the sum of squared errors (SSE) to find the inflection point. The Silhouette Score, shown in Figure 10, measures the tightness and separation of the clusters, aiding in selecting the most suitable number of clusters. Combining these methods, we opted for three clusters (k = 3).

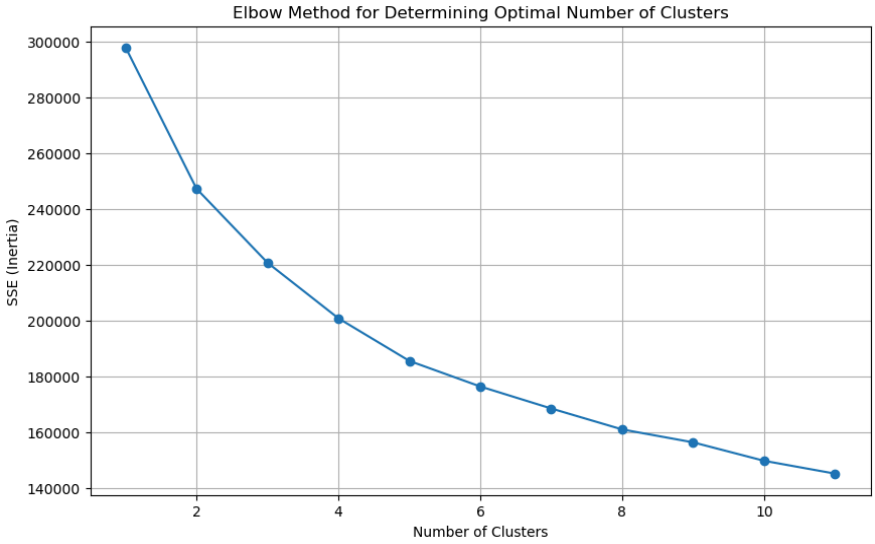

Figure 9 Elbow Method for Determining Optimal Number of Clusters

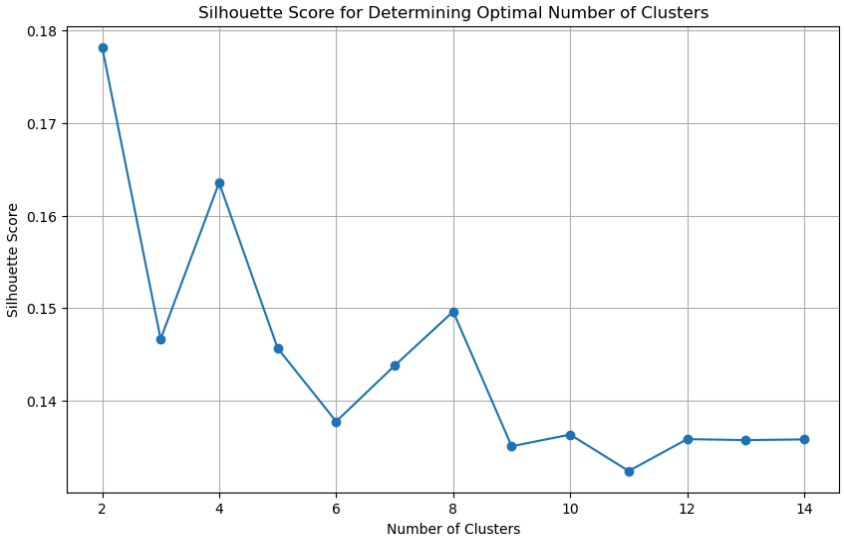

Figure 10 Silhouette Score for Determining Optimal Number of Clusters

4.2 Feature Engineering

The project began with a linear regression model, applying a logarithmic transformation to the price and standardizing the features to minimize scale discrepancies and normalize the target variable. This increased the model's training efficiency and prediction accuracy. This project uses Ordinary Least Squares (OLS) for feature selection based on p-values to simplify the model and reduce overfitting while retaining features with significant predictive power.

A decision tree model was trained using all features and then selected features based on their importance scores, setting a threshold of 0.01. This method aimed to reduce data dimensionality and noise, focusing on the most impactful features. It utilized the subset of features selected by the decision tree to train a random forest model, enhancing accuracy and generalization capability. The ensemble nature of random forests, which comprises multiple decision trees, allows for better handling of complex interactions between features, likely outperforming a single decision tree model.

The random forest model was used to reassess feature importance, further refining the feature subset to focus on those with the most significant impact on predictions. An optimized random forest model was trained with the newly selected features and an adjusted max_features ratio. This model aimed to achieve the best balance between performance and model complexity.

4.3 Airbnb Price Prediction Models

Gradient Boosting is selected for its strength in managing non-linear relationships and its proficiency in handling complex, heterogeneous datasets. Gradient Boosting builds on an ensemble of decision trees and optimizes prediction accuracy by iteratively minimizing errors. Its ability to model intricate patterns in data through feature interactions makes it particularly suitable for capturing the dynamic pricing factors of Airbnb listings.

Lasso and Ridge Regression are included for their regularization properties, which are crucial for preventing overfitting—a common challenge in predictive modeling. Lasso regression (L1 regularization) assists in feature selection by reducing the coefficients of less important features to zero, thus simplifying the model and improving interpretability. Ridge regression (L2 regularization), conversely, reduces all coefficients toward zero but does not eliminate them, which is beneficial when many variables are correlated and each carries some information about the dependent variable.

Support Vector Regression is chosen for its effectiveness in high-dimensional spaces and its robustness against outliers. It is capable of modeling non-linear relationships using the kernel trick, which is essential for addressing the complex pricing dynamics of Airbnb listings. SVR works by finding the best-fitting hyperplane in a multidimensional space, an approach that is crucial for achieving accurate predictions in a variable-rich environment like New York City's Airbnb market.

5 Results

5.1 Effects of Traffic Accessibility on Housing Prices

After conducting the cluster analysis, we utilized various visualization techniques to interpret the results. Figure 11 depicts the geographic distribution of our clusters, color-coded to indicate different clusters. This includes histograms and parallel coordinate plots to show the distribution of housing prices under different traffic conditions and to highlight the differences in characteristics between clusters.

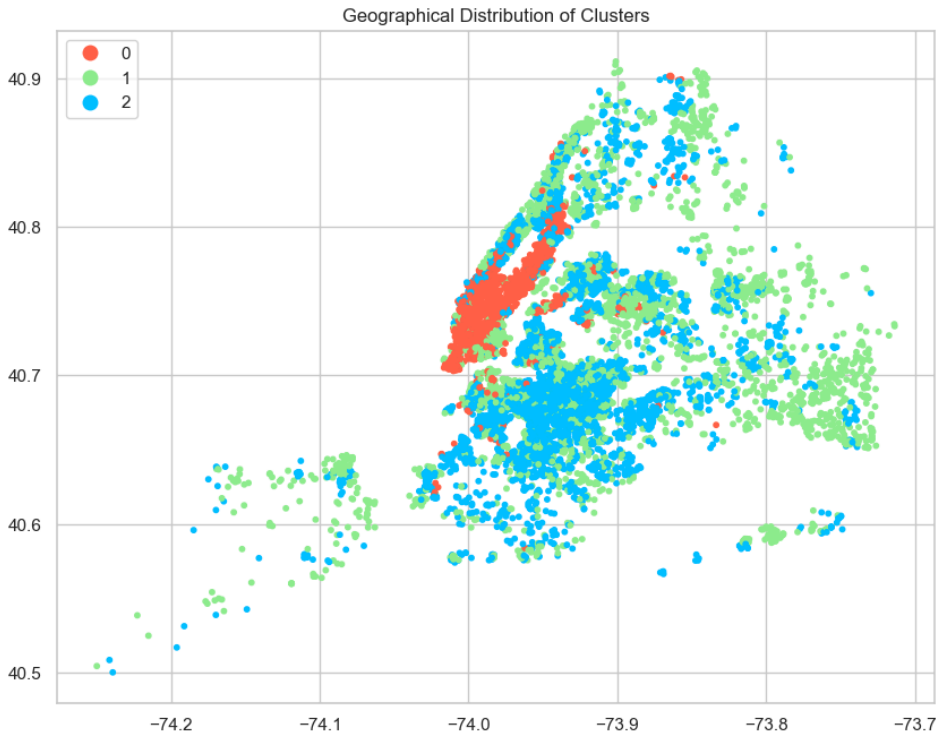

Figure 11 Geographical Distribution of Clusters

These histograms (Figure 12) display the distribution of different traffic variables within each cluster. Distinct colors represent different clusters, indicating variations in traffic characteristics across clusters. Peaks suggest a higher prevalence of specific traffic features within a cluster. From the distribution charts, we conducted an in-depth analysis of how different traffic variables are distributed across clusters and their potential impacts on housing prices. Particular attention was given to variables such as D3BMM3 and D3BMM4, which showed significant distribution variations across clusters, indicating that the density of multimodal intersections could significantly impact housing prices.

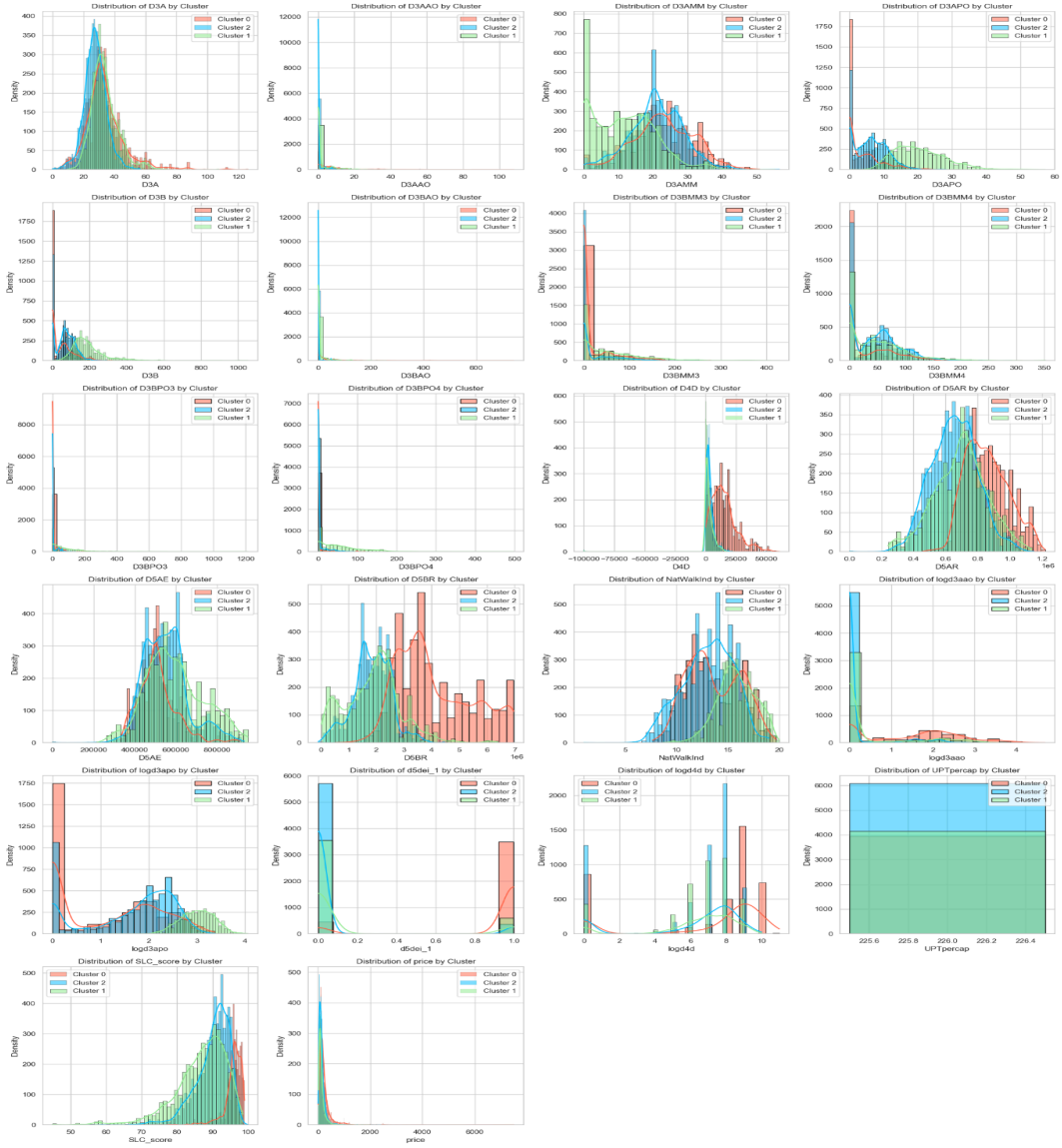

Figure 12 Distribution of Variables by Cluster

Detailed variable analysis includes:

- D3A (Total road network density): High peaks in Cluster 0 suggest areas with dense road networks, indicative of high urbanization and commercial development. This density likely correlates with higher housing prices due to better accessibility to urban services and amenities.
- D3AAO (Network density of auto-oriented links per square mile): High values in Cluster 0 indicate an area favorable to vehicular traffic, which may attract residents who depend on vehicles for commuting, potentially affecting housing prices and residential preferences.
- D3AMM (Network density of multimodal links per square mile): Prominent in Cluster 1, suggesting regions favorable to both pedestrian and public transit, likely correlating with higher residential quality and potential increases in housing prices.
- D3APO (Network density of pedestrian-oriented links per square mile): High values in Cluster 1 suggest areas with excellent pedestrian facilities, likely enhancing residential appeal and boosting housing prices.

Parallel coordinate plot (Figure 13) illustrates patterns and relationships across multiple dimensions. Each line represents a case (a house in this context), and its trajectory across the axes shows its standardized scores on traffic variables. Different colors help identify which cluster a specific case belongs to, allowing observations of variable distributions and interrelations across clusters. Detailed analysis of each cluster reveals:

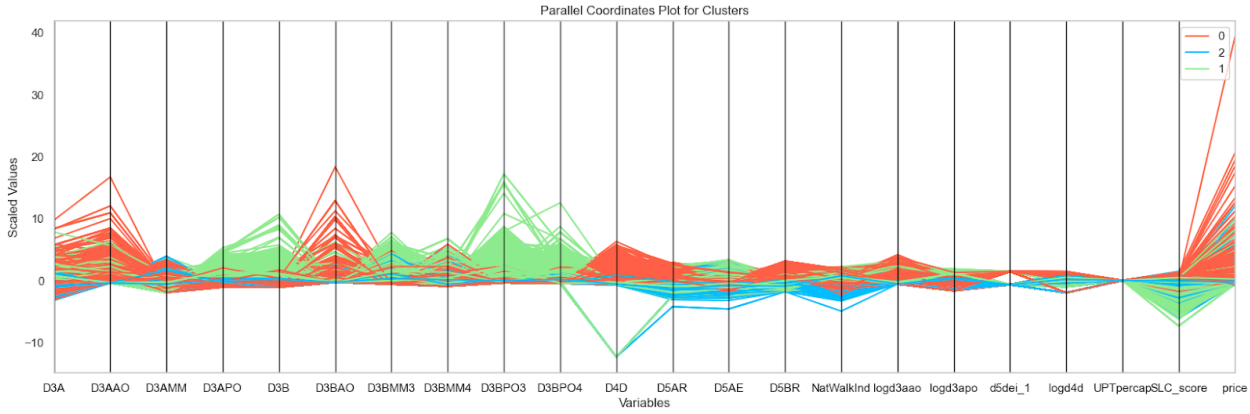

Figure 13 Parallel Coordinates Plot for Clusters

- Cluster 0 (Red): High peaks for D3AAO, D3BMM4, and price, indicating high auto and multimodal traffic facilities, typical of urban centers or highly developed commercial areas. Low values for D3APO suggest these areas, while well-serviced by traffic infrastructure, are less pedestrian-friendly.
- Cluster 1 (Green): High values for D3APO, D3BPO4, and NatWalkInd indicate good pedestrian infrastructure and high walkability, typical of residential areas or pedestrian-friendly urban zones. Lower values for D3AAO and price suggest less vehicular road density and lower housing prices, likely indicative of peripheral or less developed areas.
- Cluster 2 (Blue): High SLC_score signifies a region with superior life quality, suitable for those seeking high-standard living conditions. Lower values for D3BMM3 and D4D suggest less developed public transportation services compared to other areas.

For the conclusion, what we get is:

- Cluster 0 likely represents commercial centers or highly developed areas, characterized by high housing prices and extensive automated traffic facilities.
- Cluster 1 appears to be quieter residential areas, pedestrian-friendly, and suitable for living but with lower housing prices.
- Cluster 2 may indicate areas with a high overall quality of life, appealing to residents desiring high-quality living conditions.

5.2 Feature Selection

Use a linear regression model for basic analysis. Through feature selection using OLS, significant features are screened based on P values, which helps simplify the model and reduce overfitting while retaining features with significant predictive power for the target variable. According to Figure 14, although linear regression provides initial insights, its proportion of variance explained (R^2 = 0.573) and prediction error (RMSE = 0.467) indicate room for improvement.

OLS Regression Results

Dep. Variable:	log_price	R-squared:	0.562
Model:	OLS	Adj. R-squared:	0.560
Method:	Least Squares	F-statistic:	249.8
Date:	Fri, 03 May 2024	Prob (F-statistic):	0.00
Time:	20:33:34	Log-Likelihood:	-7369.1
No. Observations:	11347	AIC:	1.486e+04
Df Residuals:	11288	BIC:	1.529e+04
Df Model:	58		
Covariance Type:	nonrobust		

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 1.8e-28. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
Significant features: ['id', 'D3AMM', 'D3BAO', 'D5AR', 'D5AE', 'D5BR', 'NatWalkInd', 'logd3apo',
Final Model RMSE on Test Set: 0.4671419052638254
Final Model R² on Test Set: 0.5725421601491936

Figure 14 OLS Regression Results

From Figure 15, the decision tree improved upon this, providing a more robust model with an R^2 of 0.618 and an RMSE of 0.44, explaining approximately 61.8% of the variability. The purpose of decision tree feature selection is to reduce the dimensionality and noise of the data and find the most influential features.

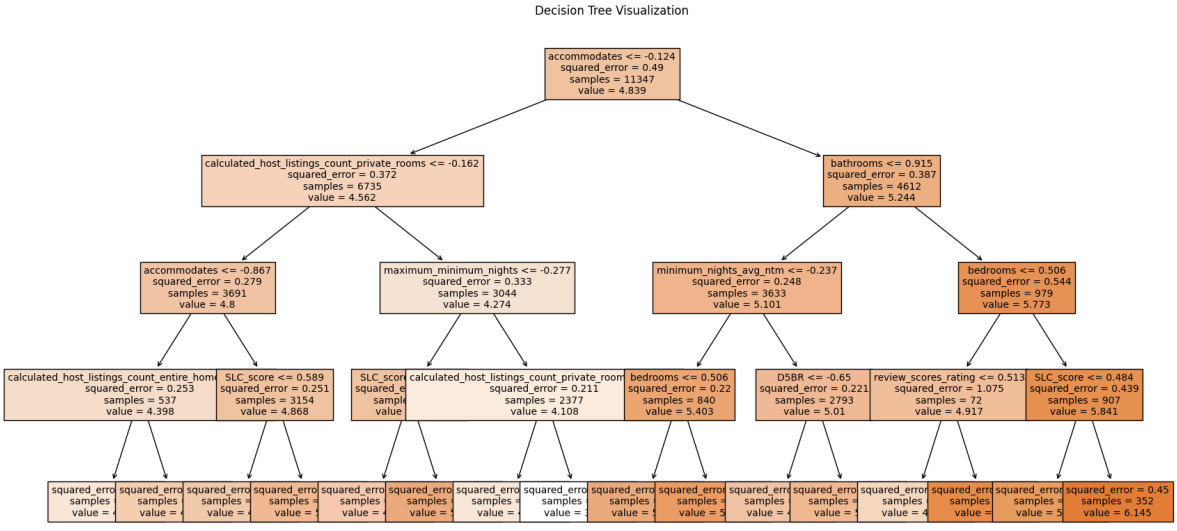

Figure 15 Decision Tree Visualization

In the data and training the random forest with decision tree-selected features significantly improved model performance, achieving an R^2 of 0.750 and an RMSE of 0.36 on the test set,

indicating superior predictive accuracy and lower error, indicating that the random forest model fits the data better than the decision tree.

The optimized random forest model further refined these results, achieving an R^2 of 0.726 and an RMSE of 0.37 on the test set. In Figure 16, It can be seen from the residual plot that the residuals of the optimized random forest model are relatively concentrated and evenly distributed, indicating that the consistency and accuracy of model predictions are relatively good. Most data points are tightly around 0, indicating that the error is small.

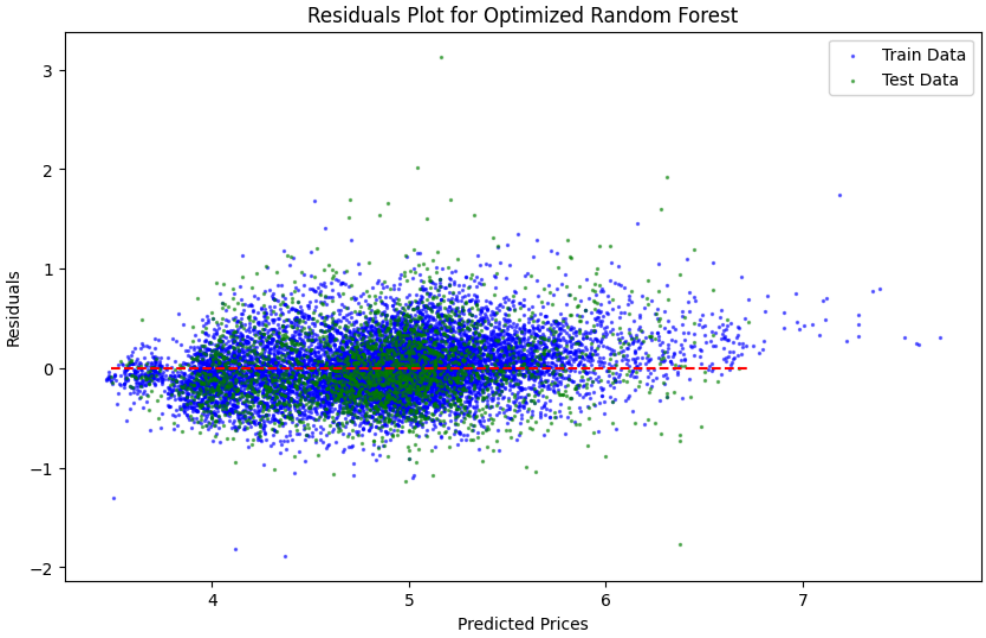

Figure 16 Residual Plot for Optimized Random Forest

This sequential analysis, from a simple linear regression to an optimized random forest, illustrates a deepening complexity and effectiveness in the models used. Each step enhanced the model's predictive precision and explanatory power. While linear regression provided a good starting point, decision trees and random forests demonstrated more robust capabilities in managing complex data structures, particularly with feature selection and model optimization.

The progression enhanced our understanding of the data and laid a solid foundation for further research and practical application. As can be seen in Figure 17, this step-by-step optimization established the random forest model as the final predictive model, outperforming both the singular decision tree and the baseline linear regression regarding robustness and accuracy, particularly in handling large-scale data and complex feature interactions. This comprehensive

approach ensures that the random forest model, with its refined feature selection and parameter tuning, provides a reliable tool for predictive analytics.

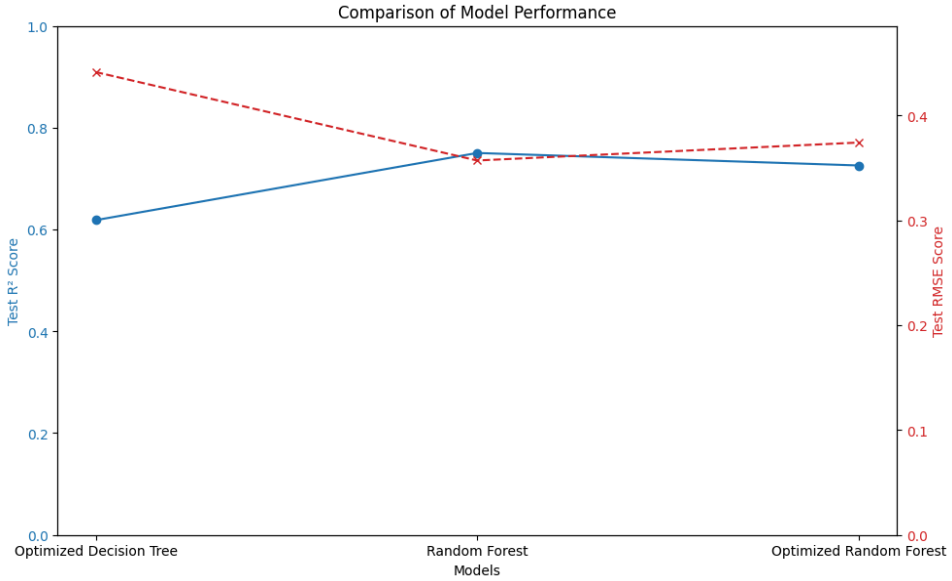

Figure 17 Comparison of R^2 & RMSE Scores Across Models

5.3 Price prediction and model selection

The study utilized Gradient Boosting, a sophisticated machine learning approach, to predict Airbnb prices. Configured with 100 estimators and a learning rate of 0.1, this model was tailored to manage the complex interactions within the dataset effectively. It was trained on features selected from a preliminary Random Forest analysis, ensuring that only the most impactful attributes influenced the predictions. The Gradient Boosting model demonstrated robust performance, capturing a significant portion of the variance in the data, evidenced by an R^2 of 0.824 for the training set and 0.751 for the testing set. The RMSE values were equally impressive, recorded at 0.29 for training and 0.36 for testing, underscoring the model's accuracy in price forecasting. Furthermore, an analysis of feature importance revealed key drivers in pricing, providing actionable insights for further refining the model.

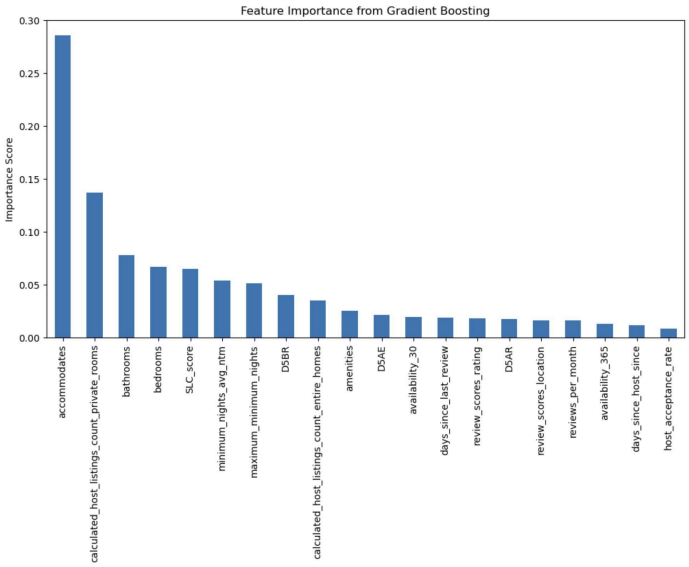

Figure 18 Feature Importance from Gradient Boosting

To enhance our model's predictive stability and prevent overfitting, the research employed Lasso and Ridge regression techniques, known for their regularization capabilities. Lasso regression, optimized through cross-validation, favored a minimal alpha of 0.001, indicating a preference for a model that retains most features. However, its moderate R^2 of 0.542 on the test set, accompanied by an RMSE of 0.48 for both training and testing, suggested a balance between complexity and predictive accuracy, potentially pointing to underfitting. This regression method also facilitated an in-depth analysis of how feature contributions varied with different levels of regularization, aiding in precise feature selection.

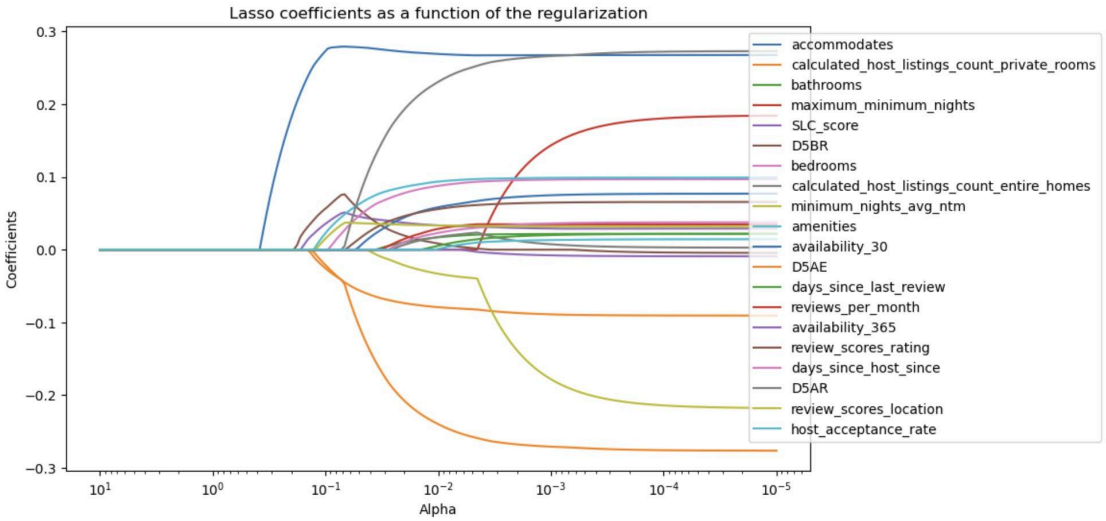

Figure 19 Lasso Coefficients as a Function of the Regularization

Lastly, Support Vector Regression with an RBF kernel was included in our suite of models to capture non-linear relationships. SVR achieved a training R^2 of 0.762 and a testing R^2 of 0.695, with RMSE scores slightly higher than Gradient Boosting, indicating its effectiveness yet slightly lesser precision. The residual plots from SVR demonstrated a generally good fit as shown in Figure 21, with most errors clustering around zero, although some discrepancies pointed towards potential model improvements.

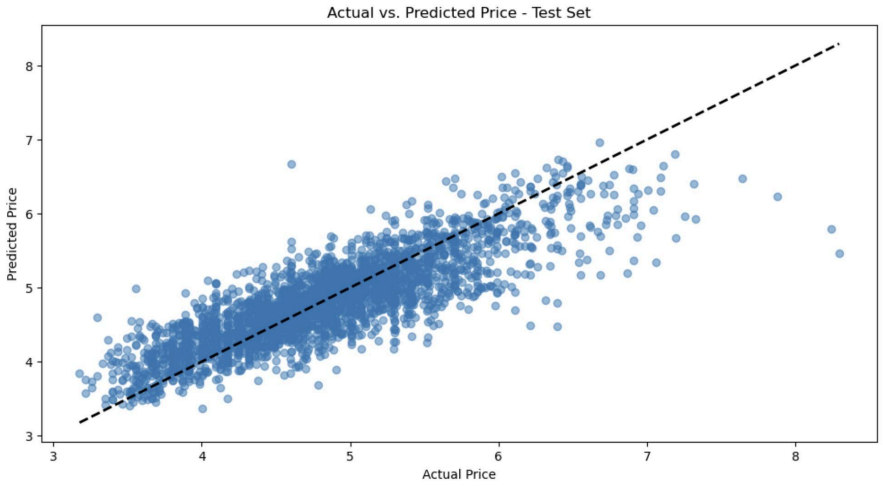

Figure 20 Support Vector Regression Actual vs Predicted Price -Test Set

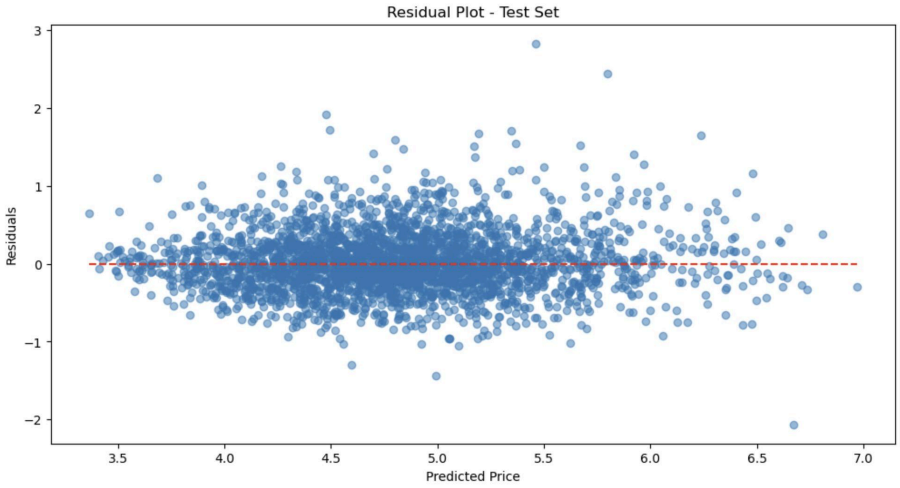

Figure 21 Support Vector Regression Residual Plot - Test Set

The selection of the final model for price prediction was based on a comprehensive evaluation of each model's accuracy, robustness, and ability to generalize unseen data. Gradient Boosting

emerged as the preferred model due to its superior R^2 and RMSE scores, indicating high predictive accuracy and the ability to capture complex nonlinear relationships within the data.

The optimized Gradient Boosting model, with its refined feature set and tuned parameters, stands as the recommended model for stakeholders looking to predict Airbnb prices in New York City effectively. This model offers a robust tool for Airbnb hosts to optimize pricing strategies and for policymakers to understand the implications of Airbnb on the urban housing market.

Further research could explore the integration of additional data sources, such as real-time economic indicators or more granular local events, which could enhance the model's predictive power. Additionally, deploying the model in a real-time pricing tool could provide actionable insights for immediate application in the market.

6 Discussion

The findings of this project underscore the significant role that both geographic proximity to transportation and property-specific attributes play in determining Airbnb pricing in New York City. This correlates well with prior research suggesting that proximity to urban amenities, such as transit hubs, significantly impacts rental prices. Our analysis extends these findings by quantitatively demonstrating how variations in access to public transportation can affect accommodation prices in an urban environment.

The practical implications of these findings are significant for Airbnb hosts and urban policymakers. For hosts, understanding that proximity to transit can command higher rental prices may influence decisions about listings and marketing strategies. For policymakers, the study highlights the importance of considering how Airbnb impacts urban housing markets and transit systems, suggesting that strategic planning and regulation could mitigate potential negative effects on local communities.

However, the study's limitations must be acknowledged. The data used, while comprehensive, does not account for the full range of seasonal and event-driven fluctuations that can affect pricing, such as major holidays or large-scale public events. Additionally, the models employed, while sophisticated, assume static relationships and may not fully capture the dynamic nature of the real estate market.

Future research could build on this study's findings by incorporating dynamic pricing models that account for real-time changes in market conditions. Further exploration of the causal

relationships between Airbnb operations and long-term housing market trends could also provide valuable insights for urban development strategies.

7 Conclusions

This study has detailed the significant factors that influence Airbnb pricing in New York City, with a particular emphasis on traffic accessibility and the attributes of the property. Specifically, the analysis revealed that properties closer to major transit routes and hubs command higher prices. This correlation suggests that guests place a premium on convenience and accessibility, which aligns with trends observed in urban accommodation preferences.

Additionally, property attributes such as modern amenities, size, and unique characteristics (e.g., balconies or views) were found to significantly affect pricing. These insights are crucial for Airbnb hosts who are seeking to optimize their rental income through strategic pricing and property improvements.

Policymakers should consider these insights when evaluating urban planning and transit development, ensuring that infrastructure improvements align with high-demand Airbnb areas to maximize economic benefits while mitigating potential impacts on housing markets. Moreover, understanding these pricing determinants can aid in the formulation of guidelines that help maintain fair pricing practices within the short-term rental market.

In conclusion, this research sheds light on the pivotal factors that drive Airbnb pricing dynamics and provides actionable data that can assist hosts in optimizing their listings for better profitability. It also offers policymakers a grounded perspective for developing informed regulations that support sustainable tourism and housing strategies. Future studies could further explore the direct correlations between short-term rental trends and long-term housing availability to refine and update urban housing policies.

References

- Harper, D., & Lim, C. (2021). The impact of environmental quality on Airbnb pricing. Journal of Environmental Economics and Management, 106, 102383. https://doi.org/10.1016/j.jeem.2021.102383
- Jensen, K., Liu, S., & Yu, J. (2020). Hotel pricing dynamics and Airbnb. Tourism Management, 81, 104145. https://doi.org/10.1016/j.tourman.2020.104145
- Kim, J., & Park, H. (2021). The impact of technological amenities on Airbnb prices. Hospitality Management, 92, 102719. https://doi.org/10.1016/j.ijhm.2021.102719
- Kisieliauskas, J. (2023). Host-related factors influencing Airbnb prices in rural areas. EconPapers. https://doi.org/10.15544/mts.2023.37
- Lee, S., & Nguyen, H. (2019). Safety perceptions and their effects on Airbnb pricing. Crime Science, 8(1), 21. https://doi.org/10.1186/s40163-019-0106-5
- Miller, T., & Brown, A. (2019). Determinants of Airbnb prices in European cities: A spatial analysis. Journal of Urban Economics, 87, 85-99. https://doi.org/10.1016/j.jue.2015.04.003
- Zhang, X., & Gao, J. (2019). Transportation accessibility and Airbnb: A geospatial analysis. Transportation Research Part D: Transport and Environment, 71, 23-36. https://doi.org/10.1016/j.trd.2019.02.012

Appendices

Team Members	Roles and Works	Approximate Percentages
Jiani Lyu	Price Prediction and models	20%
Yanchi Jin	Data collection, poster and paper organization	20%
Chelsea Liu	K-means Clustering and visualization	20%
Yajie Zeng	Feature Selection: Linear regression, Decision Tree, Random forest	20%
Yuewei Shi	Data Cleaning and EDA	20%