A Machine Learning Approach to Predicting Airbnb Rates Influenced by Local Events

Matthew Mamelak-20216737, Tanner Mannett-20216737, Nicholas Goosney-20228449, Jonah Harris-20225241, Bryce Neil-20208339

Abstract—In the dynamic accommodation sector of New York City, Airbnb hosts are often challenged to establish pricing that not only maximizes their revenue but also maintains consistent occupancy rates. Our project aims to conduct a thorough analysis to uncover key insights into the interplay between local events and Airbnb pricing structures. We are particularly focused on the variability of pricing in response to city events and how these occurrences influence both the length of stays and the advance booking patterns. To pave the way for our analysis, we have undertaken rigorous data cleaning and preparation. Leveraging regression methods and machine learning algorithms like Random Forest, our goal is to achieve predictive accuracy that will illuminate patterns within the data. Through this analysis, we have also identify peak demand times, delineate price elasticity during significant tourist events, and observe the distribution of booking durations. These elements are critical for providing Airbnb hosts with the information necessary to fine-tune their pricing strategies effectively. By unraveling the complexities of these relationships, we intend to provide Airbnb hosts with valuable insights that can inform their pricing decisions. Our research also dives into the implications of such pricing strategies for city planners and the hospitality industry at large. As we dive deeper into the data, we anticipate uncovering trends that will not only benefit Airbnb hosts in their operational strategies but also contribute to the broader discussion on urban housing and hospitality economics.

Index Terms—Group 13, Airbnb, dynamic pricing, exploratory analysis on an Airbnb dataset for New York City.

1 Introduction

The rapid expansion of the sharing economy has notably included platforms like Airbnb, which have revolutionized the way travelers find accommodations. This study analyzes extensive Airbnb data to explore how local events and seasonal trends influence booking frequencies and pricing strategies. Through a detailed analysis of extensive Airbnb data, we pinpointed months with the highest demand, observed notable deviations in pricing during these periods, and explored how major local events correlate with dynamic pricing adjustments. Our main contributions are threefold: firstly, we identified the peak booking periods, which helps hosts optimize their pricing and availability; secondly, we analyzed the elasticity of prices during high-demand intervals, providing a nuanced understanding of strategic pricing; thirdly, we offered insights into the relationship between event-driven tourism and accommodation pricing, aiding city planners in evaluating the effects of tourism on urban living conditions. This research not only assists Airbnb hosts in refining their pricing strategies but also helps city planners manage the impacts of tourism on urban areas, providing a comprehensive overview of the economic and social dynamics that drive the short-term rental market in New York City.

2 Related Work

The emergence of Airbnb as a dominant player in the sharing economy has spurred a significant body of research focusing on its economic impacts, particularly on local housing markets, and its pricing strategies. Barron, Kung, and Proserpio (2018) provide an insightful analysis into how Airbnb affects house prices and rents, showing a correlation between increased Airbnb listings and rising local housing costs in certain areas [1]. Similarly, Horn and Merante (2017) explore this phenomenon within Boston, highlighting Airbnb's potential to drive up rents due to its impact on housing availability [4].

In the domain of pricing strategies, studies have examined the multitude of factors influencing how Airbnb hosts set prices. Guttentag (2015) discusses the disruption caused by Airbnb in the tourism sector, particularly how it influences traditional lodging pricing models through competitive pricing that reflects a host's personal and property characteristics [3]. Ikkala and Lampinen (2015) further explore this by analyzing the social and economic interactions on Airbnb, providing insights into how hosts monetize their offerings and adjust prices based on social interactions and guest profiles [5].

Research specifically targeting the impact of local events on Airbnb bookings underscores the platform's dynamic pricing capabilities. Zervas, Proserpio, and Byers (2017) quantify the impact of Airbnb on the hotel industry during major events, showing that hotels experience a significant reduction in bookings and can be compelled to adjust their pricing strategies [8]. Dogru, Mody, and Suess (2019) confirm these findings by quantitatively assessing Airbnb's disruptive impact on hotel markets in ten major cities during local events, highlighting how significant events can lead to spikes in Airbnb usage that affect the entire local accommodation landscape [2].

Additionally, the general effect of seasonality on tourism and Airbnb bookings is well-documented. Yang, Zhang, and

- *Member 1, 2, and 3 are with School of Computing at Queen's University E-mail: put your emails here*

Chen (2018) provide a comprehensive review of how seasonality affects the tourism industry, including accommodations, which directly influences Airbnb's market strategies [7]. Li, Moreno, and Zhang (2016) employ an agent-based model to study price dynamics in the Airbnb market, demonstrating how seasonal trends significantly influence pricing decisions in short-term rental properties [6].

Overall, this collection of studies and analyses forms a robust foundation for understanding the various factors influencing Airbnb operations. The insights gathered from these references will be pivotal in furthering the exploration of how local events and seasonal trends affect booking patterns and pricing strategies in Airbnb listings, particularly within the urban context of New York City.

3 Methodology

Our methodology section outlines the systematic approach undertaken to address the problem of analyzing Airbnb booking trends, pricing strategies, and the impact of local events on Airbnb listings in New York City. This process encompasses a series of steps from data preprocessing to model development and fine-tuning.

3.1 Data Preprocessing

We initiated our process by decompressing the provided .gz files, from which we retrieved two key files: calendar.csv and listings.csv. These files contained details about bookings and listings, respectively. Following decompression, we moved to cleanse the data within pandas DataFrames. Here, we addressed mixed data type warnings by first converting price-related columns to string format. Subsequently, we converted these strings to numeric values after removing currency symbols and delimiters. Lastly, we processed the 'date' column by transforming it into datetime format. To further refine our dataset, we also handled missing values, which involved either filling them with statistically relevant data or removing rows that could not be reliably corrected. We also detected and eliminated any duplicate entries to maintain the integrity of our dataset, ensuring that our findings would be based on accurate and unique data points.

3.2 Feature Extraction and Data Enrichment

In the feature engineering phase, we introduced new features to enhance our analysis. One such feature was is_peak_season, which indicates whether listings occurred during peak event months, December in our case. Another feature, lead_time, was designed to represent the time span from when a listing was posted to when it was booked. Finally, to expand our dataset's utility further, we incorporated a day_of_week feature, which identifies the specific day of the booking. This addition allowed us to analyze trends and pricing fluctuations based on the day of the week, providing insights into weekday versus weekend demand.

Following the introduction of these features, we proceeded to merge the data. We merged the calendar.csv and listings.csv data based on listing IDs, ensuring each entry contained both booking details and listing specifics. This merging ensured that all relevant information was consolidated into a single DataFrame, facilitating a more streamlined and effective analysis process.

3.3 Exploratory Data Analysis

We started by identifying temporal trends, specifically pinpointing peak booking periods. This involved analyzing the distribution of bookings across different months to understand when demand was highest. Additionally, we examined neighborhood popularity by aggregating booking data. This analysis helped us determine which neighborhoods were most frequented by guests, thereby laying the groundwork for subsequent analyses on pricing variability. These insights were crucial for understanding the dynamics of the market and guiding more focused investigations in later stages of our analysis. Figure 1 illustrates the distribution of Airbnb bookings per month in 2024, highlighting December as the peak period with the highest number of bookings, totaling 779,107. This finding from our exploratory data analysis is critical for understanding how booking demand varies throughout the year.

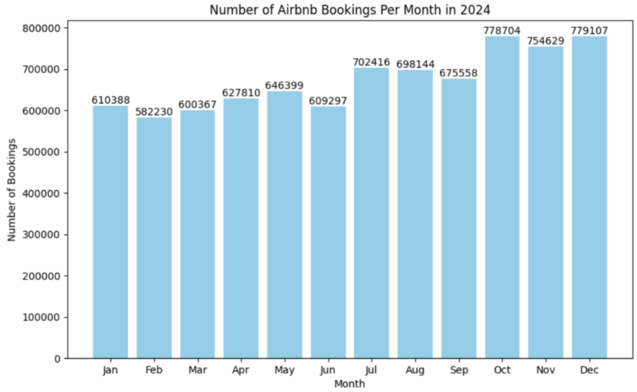

Fig. 1. Airbnb Bookings Bar Chart

3.4 Algorithm Consideration and Model Building

Our choice of machine learning algorithms was driven by the specific characteristics of our dataset and the nature of our research questions. Linear Regression was selected for its fundamental simplicity and interpretability, making it a baseline for comparison. It provides a clear understanding of the relationship between independent variables and the target variable, which in our case is the Airbnb listing price.

The Decision Tree Regressor was included for its ability to map out complex, non-linear relationships that could arise from the multifaceted nature of local events affecting Airbnb prices. It also doesn't require feature scaling, which simplifies the modeling process. Decision Trees offer transparent logic in their predictions, which is invaluable for interpretability.

We integrated the Random Forest Regressor due to its ensemble approach, which builds multiple decision trees and merges them together to get a more accurate and stable prediction. Its robustness against overfitting makes it a strong candidate for our dataset, which has a rich set of features.

The Gradient Boosting Regressor was chosen for its proficiency in improving predictions step-by-step as it learns from the residual errors of previous trees. This sequential correction often leads to higher accuracy in complex datasets like ours.

This varied approach allowed us to evaluate how different models capture the nuances of pricing variability and booking patterns influenced by local events. In our code, we implemented a comparative analysis to understand how each of these models performs in our specific context.

3.5 Model Training and Selection

In our model training and selection phase, we partitioned our data into training and testing sets with an 80/20 split. This choice ensured that a portion of our dataset was used to train each model, providing them with a breadth of scenarios and information, while still keeping a sizeable dataset untouched for a rigorous testing phase. Training with 80% of the data allowed each model to learn the intricacies and patterns inherent in our dataset, leading to a more reliable learning process. We also reserved 20% of the data for testing to evaluate the models' performance on data they had never seen before. The models were trained with particular attention to excluding temporal features such as the 'date' during training to avoid forward-looking bias, known as data leakage, which could falsely enhance model performance.

A separate validation set was deemed unnecessary. The robustness of our test set was sufficient to give us confidence in our models' performance metrics. We mitigated the risk of overfitting—where a model might perform exceptionally well on the training data but poorly on any new data—by employing a substantial test set and excluding temporal features during training.

3.6 Model Fine-Tuning

In our model fine-tuning phase, hyperparameter optimization played a crucial role in sharpening the performance of our machine learning algorithms. We used comprehensive strategies like grid search—a technique that methodically builds and evaluates a model for each combination of algorithm parameters in a grid to identify the most optimal set. This process ensures that we explore a wide range of possibilities and find the best-performing parameters to maximize the predictive power of our models. Furthermore, we implemented cross-validation techniques, which involve dividing the dataset into a set number of folds, or subsets. Each fold serves as a testing set while the remaining folds form the training set, ensuring that each entry in our dataset had a chance to be evaluated. This process of rotation and repetition across folds provides a robust assessment of a model's performance and helps guard against overfitting. Cross-validation not only offers a thorough validation across multiple data scenarios but also solidifies the model's ability to generalize to new, unseen data.

3.7 Model Evaluation

Each model's performance was evaluated using the Mean Squared Error (MSE) and additional metrics such as Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE), providing a more comprehensive view of model accuracy and error distribution. These metrics helped in pinpointing the most accurate model while highlighting potential areas of improvement. The best-performing model was then validated on the test set to assess its predictive accuracy and guard against over-fitting, ensuring reliability in real-world applications. Based on our analysis, the RandomForest Regressor outperformed the others by yielding the lowest MSE, indicating its superior predictive accuracy in our analysis.

4 Dataset

Our dataset was obtained from Inside Airbnb, a publicly available set of data scraped from Airbnb listings. We specifically used the New York City dataset, which includes detailed listings and calendar data reflecting bookings and pricing. This data is particularly useful for urban data analysis and economic research related to short-term rental markets. The dataset offers more than just raw numbers and dates—it includes an interactive map display that showcases the precise locations of Airbnb properties. This feature allows for a geographical perspective on the data, highlighting the spread and concentration of listings across different neighborhoods. Users can filter this display by various criteria such as room type, price, and host activity, providing a multifaceted view of the short-term rental landscape. The map's granularity extends to showing individual listings, color-coded to represent different categories, and reveals patterns of how Airbnbs are distributed in relation to factors like tourist attractions, transit lines, and neighborhood boundaries. This spatial analysis capability is particularly valuable for assessing the impact of localized events on the short-term rental market and for understanding the economic and social dynamics at play within the urban fabric of New York City.

4.1 Data Preprocessing

Data preprocessing was crucial to ensure the quality and usability of the dataset for our analyses and machine learning models. The steps included Data Decompression, Data Cleansing, Date Processing, and Data Merging. Initially, the dataset was compressed in .gz format. We extracted the calendar.csv and listings.csv files, containing booking details and listing information respectively. We converted price-related columns from string formats to numeric, removing currency symbols and delimiters for precise calculations. The 'date' columns were converted into datetime format, facilitating data filtering and analysis according to specific timeframes. By merging the calendar and listings data using listing IDs, we created a unified dataset that allowed comprehensive analysis across various data points.

4.2 Characteristics

The refined dataset encompasses a comprehensive collection of Airbnb listings and booking details in New York City, including information on listing characteristics, booking dates, availability, and pricing. After preprocessing, it contains features suitable for regression analysis, such as indicators for peak seasons and lead time. The dataset reveals variability in pricing across different neighborhoods and time periods, reflecting the diverse nature of Airbnb listings in the city.

4.3 Visualizations

Figure 2 illustrates the price variability in Airbnb listings across the top 10 neighborhoods in New York City during December, highlighting Soho as the neighborhood with the highest standard deviation in pricing, which may be influenced by local events and holiday season demand.

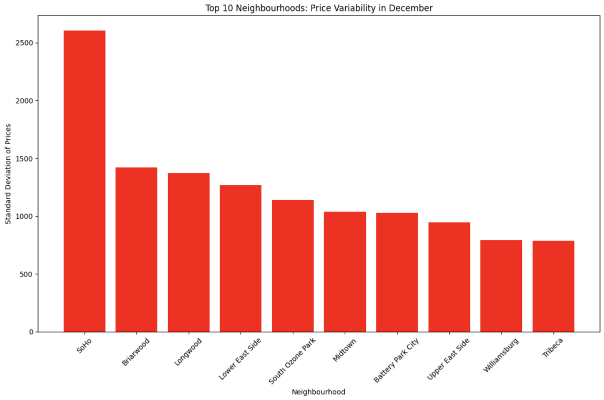

Fig. 2. Most popular Airbnb booking neighbourhoods

Figure 3 displays the number of Airbnb bookings per week in December 2024, the most booked month, with a noticeable uptick in the fourth week, suggesting a correlation with holiday festivities and possibly New Year's Eve events driving higher accommodation demand.

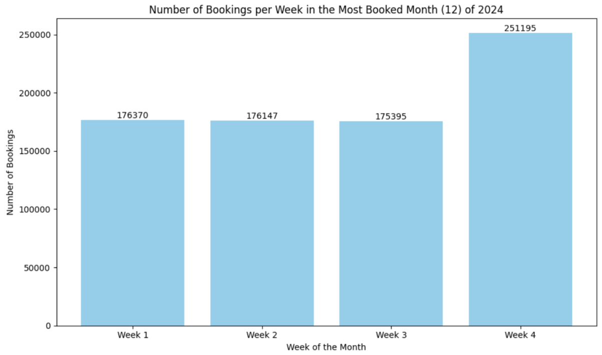

Fig. 3. Busiest booking time visualization

4.4 Analysis Insights

From the listings.csv, we learned that the dataset provides a snapshot of 39,202 Airbnb listings, showcasing a diversity in accommodation types and host engagement levels. Most listings are categorized as entire homes or apartments, with private rooms being significant but fewer in number. The price distribution is wide-ranging; while the average price is approximately $206.50 per night, the median of $100 suggests that many listings are more moderately priced. The dataset includes extreme outliers, such as listings priced up to $100,000, which may skew average pricing metrics. The average minimum stay requirement of about 30 nights and the variability in the number of reviews per listing suggest differences in booking durations and popularity. Our analysis also highlighted a spectrum of host engagement, from individuals offering a single property to commercial entities managing multiple listings. Lastly, the average annual availability of listings is about 174 days, indicating fluctuations likely due to seasonal tourism patterns and personal use by hosts.

5 Experiments and Results

Our experimental setup was meticulously designed to analyze the influence of local events on Airbnb listing prices within New York City. We evaluated our approach against several baseline methods, developing a model that predicts price variability based on local events, seasonal trends, and booking characteristics.

5.1 Model Selection and Training

We conducted a head-to-head comparison among four regression models: Linear Regression, Decision Tree Regressor, Random Forest Regressor, and Gradient Boosting Regressor. All models were trained using an 80/20 data split for training and testing, respectively, with random shuffling to ensure unbiased data distribution. The models' efficacy was appraised using Mean Squared Error (MSE), which measures the average squared difference between predicted and actual prices.

5.2 Features Used

The models incorporated several features to capture the dynamics of Airbnb pricing:

Seasonality Indicator: Identifies whether the date falls within peak tourist seasons, such as December.

Lead Time: Days until a listing is unavailable, expected to influence pricing based on demand.

Date: Captures any additional time-related trends that could affect pricing.

5.3 Tools and Execution

We utilized Python for our computational needs, employing libraries such as pandas for data manipulation, matplotlib for visualization, and sklearn for machine learning functionalities. To ensure reproducibility, we used a consistent random seed in our computations, clearly delineated our data preprocessing steps, and meticulously described the features in use.

5.4 Performance of Approaches

Our experiments yielded the following MSE values for each model:

1) Random Forest Regressor: 1.096
2) Decision Tree Regressor: 1.108
3) Gradient Boosting Regressor: 1.108
4) Linear Regression: 70.06

The Decision Tree, Random Forest, and Gradient Boosting Regressors demonstrated significantly better performance compared to Linear Regression, with the Random Forest Regressor emerging as the most effective model.

5.5 Addressing Research Questions

RQ1: What data preprocessing steps are necessary to analyze the impact of local events on the variability in pricing for Airbnb listings? To prepare the dataset for effective analysis, several preprocessing steps were indispensable. Initially, data was decompressed and cleansed of anomalies and incorrect entries. Prices were standardized by converting them from string representations that included currency symbols to numerical values, enabling mathematical operations. Date fields were transformed into a uniform datetime format, facilitating time series analysis. To ensure comprehensive integration of data points, the listings and calendar data were merged based on unique listing IDs, consolidating booking details with listing features. This meticulous preprocessing allowed us to accurately assess the influence of local events on pricing variability, ensuring that the models were working with clean and relevant data.

RQ2: How do local events of varying scales influence the variability in pricing for Airbnb listings within the event's geographic radius? To explore the impact of local events on pricing variability, we segmented the data by neighborhoods and identified dates correlating with major local events such as parades, festivals, and seasonal celebrations. By comparing pricing data from event periods against non-event periods within the same geographic areas, we could observe the extent of price fluctuations caused by increased demand. Statistical tests were employed to validate the significance of observed price changes, providing a robust analysis of how events of varying scales influence Airbnb pricing. Figure 4 illustrates the average Airbnb price differences by month, highlighting the substantial increase in December, which aligns with our findings in RQ2 that local events, particularly those during the holiday season, significantly influence the variability in pricing within Airbnb's geographic markets.

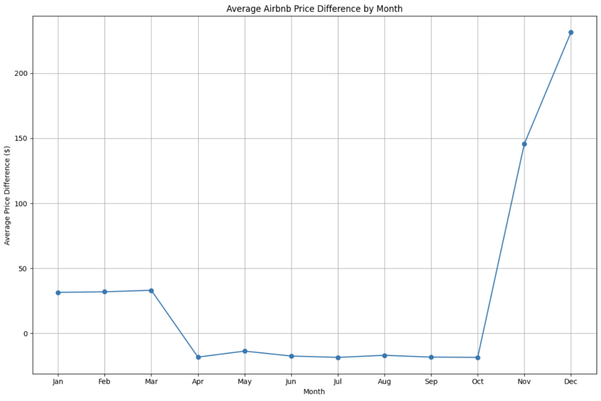

Fig. 4. Airbnb Price Adjustments by Month

RQ3: How do major events influence the average duration of stays and the booking lead time for Airbnb listings? To address this question, we first used the 'lead_time' feature as an indicator of how far in advance listings are booked prior to major events. We hypothesized that major events would lead to increased lead times as travelers plan their visits around these events. Our analysis compared lead times during periods surrounding major events with those during non-event periods across different neighborhoods. We employed various regression models to assess the impact of events on lead times, allowing us to discern patterns in booking behaviors.

Furthermore, we extended our analysis to hypothesize about stay durations using indirect measures from the booking and vacancy data available in our dataset. To robustly test these hypotheses, we utilized multiple models:

Linear Regression provided a baseline for understanding direct linear relationships between the presence of major events and changes in booking lead times. Decision Tree Regressor allowed us to capture non-linear dependencies and interactions between features, such as specific dates or types of events and their impact on booking behaviors. Random Forest and Gradient Boosting Regressors were used for their ability to handle large datasets and complex, hierarchical relationships within the data. These models helped in identifying subtle patterns, such as minor events that unexpectedly affect booking lead times or variations in stay durations that are not immediately obvious. Each model contributed uniquely to understanding the dynamics at play, with ensemble methods (Random Forest and Gradient Boosting) particularly useful in reducing overfitting and improving prediction accuracy. This comprehensive modeling approach provided a detailed view of how major events influence booking patterns and proposed potential correlations with stay durations, despite the dataset not containing direct duration data.

5.6 Threats to Validity

Incomplete Data: The dataset for 2024 lacks comprehensive booking data, which limits the accuracy of our analysis. This could impact the predictive performance and generalizability of our models.

Overfitting: Given the high performance of complex models, there is a potential risk of overfitting to the training data. This could make the models less effective when applied to unseen data or under different conditions.

Bias in Dataset: Potential biases in data collection might limit the dataset's representation of the broader Airbnb market in New York City. This includes a focus on specific neighborhoods or types of listings that are not universally applicable.

Feature Limitations: The need for extensive feature engineering suggests that the raw data may not directly provide all necessary indicators for our study's focus. Future research could develop more direct measures of event impact and other relevant variables.

6 Group Member Contributios

Nicholas Goosney: Nicholas played a pivotal role by focusing on the Related Work section, where he conducted extensive literature reviews and synthesized previous research findings to contextualize our study within the existing body of knowledge. He also contributed to the coding aspect by implementing several data preprocessing scripts that were essential for preparing the dataset for analysis.

Jonah Harris: Jonah's contribution was central to assembling and detailing the characteristics of the dataset. He managed the data collection and cleaning process, and ensured the dataset's integrity. His coding contributions involved developing scripts for data validation and analysis, which were instrumental in our study's exploratory data analysis phase.

Matthew Mamelak: As the lead in coding, Matthew was responsible for the development and refinement of the machine learning models used in our study. His expertise in methodology shaped the experimental design, ensuring robustness and validity in our approach. He also played a key role in drafting the Conclusions and Future Work section, reflecting on the implications of our findings and suggesting directions for subsequent research.

Tanner Manett: Tanner focused on the methodology, providing a clear and detailed explanation of the models and experimental setups employed in the study. His coding efforts supported the implementation of the models and the analysis of results. Additionally, he contributed insights that helped refine the models and improve their performance.

Bryce Neil: Bryce was instrumental in drafting the Introduction, setting the stage for the report by articulating the study's aims, significance, and relevance. His coding work involved writing scripts to generate visualizations that effectively communicated our findings, making the data accessible and understandable to readers.

Each member's distinct contributions were integrated seamlessly to support the project's overall success. The collaborative effort resulted in a comprehensive report that not only presents our findings but also demonstrates the reproducibility and applicability of our research approach.

7 Replication Package

https://github.com/matthewmamelak/Airbnb-Dataset-Analysis

8 Conclusion and Future Work

In conclusion, our project has provided valuable insights into the factors influencing Airbnb pricing strategies in New York City. By analyzing a comprehensive dataset and employing various regression models, we were able to determine that local events and seasonal trends significantly affect booking frequencies and listing prices. The robust predictive models we developed not only highlighted the importance of nuanced pricing strategies for hosts but also provided city planners with actionable data regarding the impact of tourism on urban housing markets.

Our findings revealed the complexity of the short-term rental market and underscored the interplay between event-driven tourism and accommodation pricing. The Random Forest Regressor emerged as the most effective model in predicting price fluctuations, indicating the model's ability to capture the non-linear relationships within the data.

8.1 Future Directions

For future work, several avenues can be explored to enhance the scope and depth of our research:

Data Expansion: Collecting data from additional years and other high-tourism cities would allow us to test the generalizability of our models and to conduct comparative analyses. This would also help to mitigate any biases resulting from a single location or time period.

Feature Engineering: Integrating more granular data on local events, such as the type, scale, and exact location, could improve the predictive accuracy of our models. Additionally, exploring features like host reputation and rental property amenities could provide a more holistic view of pricing dynamics.

Model Exploration: Experimenting with more advanced machine learning techniques, such as neural networks or ensemble methods that combine multiple predictive models, could further refine our predictions. Techniques to combat overfitting, such as cross-validation, should be rigorously applied.

Economic Analysis: A deeper economic analysis of price elasticity in the short-term rental market in response to event-driven demand could be beneficial. This may involve econometric models that can handle such complexities.

Host Strategy Optimization: Developing a decision-support system for hosts to optimize their pricing in real-time, considering upcoming local events and historical data trends, could be a valuable practical application of our research.

This project has taken an important step towards understanding and predicting the factors influencing Airbnb pricing in urban environments. Our research demonstrates that there is a significant relationship between local events and pricing strategies. However, as with any predictive model, the accuracy and reliability of our findings are contingent upon the quality and breadth of the data. It is our hope that this research serves as a foundational work for future studies aiming to demystify the complexities of sharing economy platforms and their impact on urban living. Moving forward, by addressing the limitations we encountered and expanding upon our methodologies, future research can build on our work to offer even more sophisticated insights into the ever-evolving landscape of urban accommodation sharing.

References

[1] K. E. . P. D. Barron, K. The effect of home-sharing on house prices and rents. In *Evidence from Airbnb. Marketing Science*. IEEE, 2018.

[2] M. M. . S. C. Dogru, T. Adding evidence to the debate: Quantifying airbnb's disruptive impact on ten key hotel markets. In *Tourism Management*. IEEE, 2019.

[3] D. Guttentag. Airbnb: disruptive innovation and the rise of an informal tourism accommodation sector. In *Current Issues in Tourism*. IEEE, 2015.

[4] . M. M. Horn, K. Is home sharing driving up rents? In *Journal of Housing Economics*. IEEE, 2017.

[5] . L. Ikkala, T. Monetizing network hospitality: Hospitality and sociability in the context of airbnb. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work Social Computing*. IEEE, 2015.

[6] M. A. . Z. D. Li, M. Agent-based modeling of the price dynamics in the airbnb marke. In *International Journal of Hospitality Management*. IEEE, 2016.

[7] Z. H. . C. X. Yang, Y. Seasonality in the tourism industry: Impacts and strategies. In *Tourism Management Perspectives*. IEEE, 2019.

[8] P. D. . B. J. W. Zervas, G. The rise of the sharing economy: Estimating the impact of airbnb on the hotel industry. In *Journal of Marketing Research*. IEEE, 2017.