Introduction

In the dynamic landscape of the hospitality industry, the short-term rental market has undergone remarkable transformation, with the advent of platforms like Airbnb in 2008. This innovative approach to accommodation has not only redefined the way people travel but has also given rise to an entire ecosystem of opportunities for hosts and travelers alike. The swift evolution of this segment has led to the emergence of a global industry, which, as of 2022, boasts a staggering revenue of USD 82.78 billion, engaging approximately 701.9 million users worldwide. This impressive growth trajectory highlights the profound impact of the sharing economy on the way people access and experience lodging.

At the heart of this industry's success is the unique concept of home-sharing, where private individuals open up their residences to travelers for short-term stays. The pricing dynamics in this space are inherently influenced by a multitude of factors, ranging from property location and amenities to local events and seasonal trends. As a result, the pricing of short-term rentals often hinges on the comparative valuation of similar listings, with hosts striving to set rates that attract guests while ensuring their offerings remain competitive in the market.

It is within this context that we develop a project focused on leveraging the power of machine learning to address the intricacies of short-term rental pricing. This project seeks to enhance pricing strategies for property owners by estimating a listing price based on several attributes. The primary objective is to develop, select, and optimize a data-driven pricing model that surpasses the traditional benchmark of average prices for comparable listings.

Project Vision and Objectives

This project envisions the creation of a novel machine learning-based pricing model that empowers hosts in the short-term rental market to make well-informed pricing decisions. Rather than relying solely on the conventional approach of setting prices based on local averages, the proposed model will leverage a comprehensive dataset drawn from Airbnb listings, calendar data, and guest reviews for three distinct cities: New York City (NYC), Jersey City (JC), and Rio de Janeiro (Rio). The inclusion of Jersey City stems from its geographical proximity to NYC, offering a unique vantage point to compare and contrast relevant variables in the pricing models of both cities. Additionally, the incorporation of Rio's data provides an opportunity for cross-market analysis, facilitating insights into the variances between domestic and international rental markets.

As with any ambitious endeavor, certain constraints shape the scope of this project. The dataset for analysis comprises Airbnb listings from the past 12 months, ensuring that the model's training and evaluation remain rooted in recent trends and dynamics. This pragmatic timeframe aims to enhance the model's relevance and accuracy by reflecting the most up-to-date market conditions.

Stakeholders and Data Source

This undertaking is not only of intrinsic value to the short-term rental industry but also holds implications for property owners, travelers, hotels, real estate investors, community leaders, and local governments. By equipping hosts with a sophisticated pricing tool, this project contributes to the industry's overall health while enabling travelers to access diverse and attractively priced accommodations.

The data necessary for this project was sourced from "http://insideairbnb.com," an advocacy group dedicated to understanding Airbnb's impact on residential communities. The website's regular scraping of Airbnb's data provides a robust and comprehensive dataset that serves as the foundation for the proposed machine learning model.

In the subsequent sections of this white paper, we will delve into the methodology employed for data preprocessing, model selection, training, and optimization. We will also discuss the implications of our findings for various stakeholders in the short-term rental ecosystem. Ultimately, this project seeks to chart a course toward more informed, data-driven pricing strategies, enabling hosts to optimize their offerings and enhancing the overall experience for both hosts and travelers.

Data Used in the Study

This study draws upon a rich and extensive dataset collected from three diverse cities: Jersey City, New York City, and Rio de Janeiro. Each dataset comprises four primary components: Calendar, Listings, Reviews, and Locale, collectively providing a comprehensive view of the short-term rental market in these urban centers.

Jersey City Data:
- Calendar Dataset: This dataset encompasses 935,900 entries with 7 columns, serving as a chronological record of property bookings.
- Listings Dataset: Comprising 2,566 entries and a sprawling 75 columns, this dataset contains detailed information about individual listings, such as property attributes, host details, and pricing.
- Reviews Dataset: With 1,064,458 rows and 2 columns, this dataset holds invaluable feedback and reviews from guests, shedding light on host performance and property quality.
- Locale Dataset: This dataset includes 230 entries across 2 columns, providing location-related information associated with each listing.

New York City Data:
- Calendar Dataset: A robust dataset of 14,551,462 entries and 7 columns offers a comprehensive overview of property bookings over time.
- Listings Dataset: Comprising 39,881 entries and 75 columns, this dataset presents extensive details about listings, including host characteristics and property specifications. The Neighborhood column alone encompasses 23,466 unique entries, reflecting the diverse localities within the city.

- Reviews Dataset: This dataset contains a remarkable 1,064,458 rows and 6 columns, encapsulating guest reviews and insights into the guest-host interaction.
- Locale Dataset: This dataset is relatively succinct, consisting of 6 entries across 2 columns, focusing on locale-specific information associated with each listing.

Rio de Janeiro Data:
- Calendar Dataset: With an impressive 9,623,164 rows and 7 columns, this dataset offers an expansive view of property bookings in Rio de Janeiro.
- Listings Dataset: Comprising 26,366 entries and 75 columns, this dataset provides comprehensive details about listed properties, encompassing host information, property features, and pricing.
- Reviews Dataset: This dataset encompasses 458,439 rows and 6 columns, housing valuable guest reviews and insights into host performance.
- Locale Dataset: This dataset, spanning 160 entries and 2 columns, offers specific location-related information pertaining to each listing.

The sheer volume and diversity of data from these three cities empower our analysis, enabling us to explore the dynamic short-term rental market thoroughly. The datasets provide a robust foundation for developing and optimizing our data-driven pricing model, ensuring that our findings are grounded in a wealth of real-world information.

Data Wrangling

Data wrangling plays a critical role in the preprocessing phase of this study, enabling us to transform raw data into a clean, structured format suitable for analysis. In this section, we address key data wrangling steps undertaken to ensure the quality and integrity of the datasets from Jersey City, New York City, and Rio de Janeiro.

Handling Missing Values:
One common challenge observed across all three datasets was the presence of missing values. Several columns exhibited a significant proportion of null entries, which required careful handling. For instance, the "Bathroom" column contained no entries, while "Bathroom_text" suffered from substantial missing values. In addition, a noteworthy portion of listings lacked neighborhood information. Moreover, there were instances of missing reviews for listings, which could be attributed to new properties without any prior guest feedback.

To address these issues, we adopted a systematic approach:
- Column Deletion: Columns with a large proportion of null values and limited relevance to our analysis were removed. These included "Bathroom," "Calendar_updated," "License," "Host_neighbourhood," "Host_about," "Neighborhood_overview," "Neighbourhood," and "Host_location."

- Row Deletion: Listings with null entries in critical columns, such as "Host_name" and "Bathroom_text," were removed. This process helped to ensure that only listings with complete and essential information were retained for analysis.

By implementing these data wrangling techniques, we streamlined the datasets, reducing noise and facilitating more focused analysis. This step ensures that our subsequent modeling and pricing predictions are based on a clean, comprehensive dataset that accurately represents the short-term rental market in the selected cities. The resulting dataset is well-suited for training and optimizing our data-driven pricing model, leading to more accurate and actionable insights for property hosts and stakeholders.

Inspection of Cleaned Listings Dataset

Following the data wrangling process, we now have three distinct and cleaned listings datasets, one for each of the selected cities: Jersey City, New York City, and Rio de Janeiro. Each of these datasets has been meticulously prepared for further analysis. Below is an overview of the dimensions and structure of the cleaned listings datasets:

Jersey City Listings Dataset:
- Number of Rows: 1,561
- Number of Columns: 67

New York City Listings Dataset:
- Number of Rows: 38,747
- Number of Columns: 67

Rio de Janeiro Listings Dataset:
- Number of Rows: 25,311
- Number of Columns: 67

These cleaned and structured datasets now serve as the foundation for our subsequent data analysis and the development of our data-driven pricing model. With consistent columns and reduced missing values, these datasets are primed for exploratory data analysis, feature engineering, and model training. The uniformity in the number of columns across the datasets allows for meaningful comparisons and insights into the unique characteristics of the short-term rental market in each of these three vibrant cities.

DataSet Features - Neighbourhood Listing Distributions

Jersey City (JC)

In our exploration of the cleaned listings datasets, we focus on the distribution of listings across neighborhoods, which can provide valuable insights into the short-term rental market dynamics in each city.

Here, we present the distribution of listings in Jersey City (JC) based on the "neighbourhood_cleansed" feature, shedding light on the preferences and characteristics of these neighborhoods.

```
Ward E (councilmember James Solomon)          1076
Ward D (councilmember Michael Yun)             615
Ward F (councilmember Jermaine D. Robinson)    416
Ward C (councilmember Richard Boggiano)        215
Ward A (councilmember Denise Ridley)           134
Ward B (councilmember Mira Prinz-Arey)         105
Name: neighbourhood_cleansed, dtype: int64
```

The concentration of listings in certain neighborhoods, such as Ward E, suggests a demand for affordable yet conveniently located accommodations, likely driven by price-sensitive renters seeking easy access to Manhattan. This information not only aids in understanding the distribution of short-term rentals but also informs pricing strategies for hosts and offers valuable insights for travelers seeking specific neighborhood experiences in Jersey City.

New York City (NYC)

In the context of New York City's dynamic short-term rental market, the distribution of listings across different neighborhood groups provides valuable insights into the preferences and demand patterns of travelers and hosts. Here, we present the distribution of listings in NYC based on the "neighbourhood_group_cleansed" feature, shedding light on the prominence of certain boroughs and neighborhoods within the city.

```
Manhattan        16748
Brooklyn         14817
Queens            6170
Bronx             1566
Staten Island      446
```

The dominance of Manhattan and Brooklyn in terms of the sheer number of listings underscores their appeal to travelers. Manhattan, as the city's central hub, attracts guests seeking easy access to iconic landmarks, while Brooklyn's cultural diversity and neighborhood charm make it a sought-after destination.

The notable presence of Queens emphasizes the allure of neighborhoods close to Manhattan, where travelers can enjoy a convenient commute to the city's core while potentially benefiting from more budget-friendly options.

Understanding the distribution of listings across NYC's boroughs and neighborhoods is instrumental in tailoring pricing strategies, targeting specific traveler demographics, and optimizing the short-term rental experience across the city's diverse landscape.

Rio de Janeiro

Exploring the distribution of listings across different neighborhoods in Rio de Janeiro provides valuable insights into the diverse offerings of this vibrant city. Here, we present the distribution of listings in Rio de Janeiro based on the "neighbourhood_cleansed" feature, highlighting the prominent neighborhoods within the city.

Copacabana	7525
Barra da Tijuca	2723
Ipanema	2603
Jacarepaguá	1457
Recreio dos Bandeirantes	1327
Leblon	1262
Botafogo	1115
Santa Teresa	918
Centro	689
Flamengo	568
Leme	481
Laranjeiras	396
Tijuca	392
Camorim	303
Lagoa	218
Glória	208
São Conrado	195
Gávea	189
Catete	180
Jardim Botânico	172
Vidigal	153
Humaitá	148
Taquara	138
Itanhangá	132
Freguesia (Jacarepaguá)	130
Vila Isabel	128

These distributions reflect the diverse landscape of Rio de Janeiro's short-term rental market, with listings concentrated in both the iconic Zona Sul (South Rio) neighborhoods like Copacabana, Ipanema, Leblon, and Leme, as well as the expanding Zona Oeste (West Rio) neighborhoods like Barra da Tijuca, Jacarepaguá, and Recreio dos Bandeirantes. Travelers can choose from a wide array of experiences, from the bustling beaches of Copacabana to the more tranquil settings of Recreio dos Bandeirantes.

Understanding these neighborhood distributions is pivotal for hosts in setting competitive prices and for travelers in finding accommodations that align with their preferences, whether they seek the vibrant energy of the city center or the serene beauty of its beaches and hills.

Types of Properties Listed

The types of properties listed in Jersey City, New York City, and Rio de Janeiro offer valuable insights into the diverse accommodation options available to travelers in each city. The table 1 below provides a breakdown of the types of properties listed in each city, expressed as a percentage of the total available cases:

	Jersey City	%	NYC	%	Rio	%
Entire home/apt	1917	0.748536	22724	0.571716	20210	0.768120
Private room	609	0.237798	16301	0.410119	5557	0.211204
Hotel room	33	0.012886	170	0.004277	51	0.001938
Shared room	2	0.000781	552	0.013888	493	0.018737

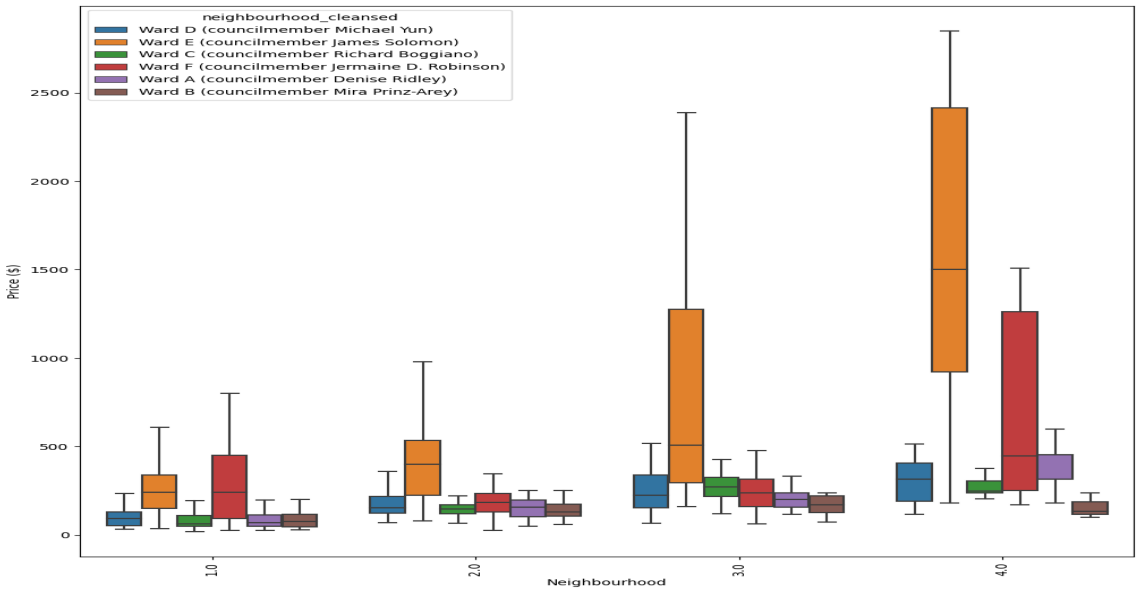
Table 1

These findings emphasize the varied preferences of travelers across these three cities, with NYC's higher percentage of private room listings reflecting its role as a global travel hub. In contrast, Jersey City and Rio de Janeiro predominantly offer entire homes/apartments, aligning with the desire for privacy and space in these locales. The diversity in property types allows travelers to tailor their accommodation choices to their unique needs and budgets.

Price Statistics per Number of Rooms by City

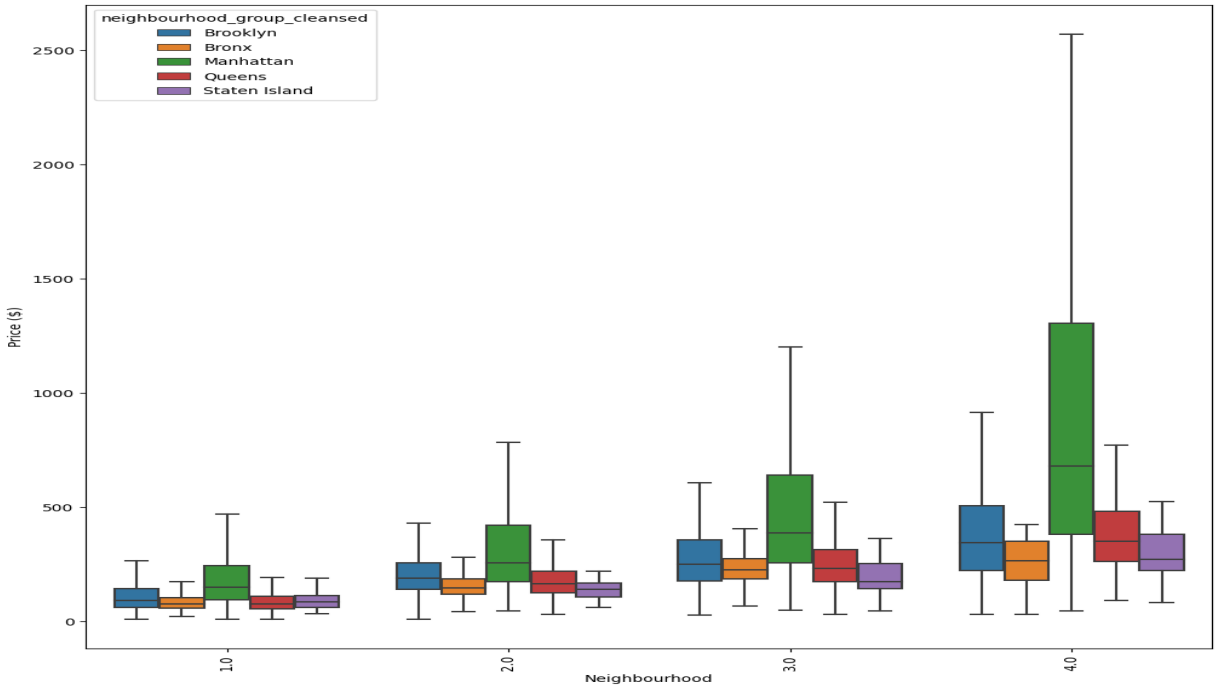
Analyzing the price statistics per number of rooms in Jersey City, New York City (NYC), and Rio de Janeiro offers valuable insights into the variations in pricing based on different factors, including neighborhood. Here, we summarize the observed trends:

Jersey City:



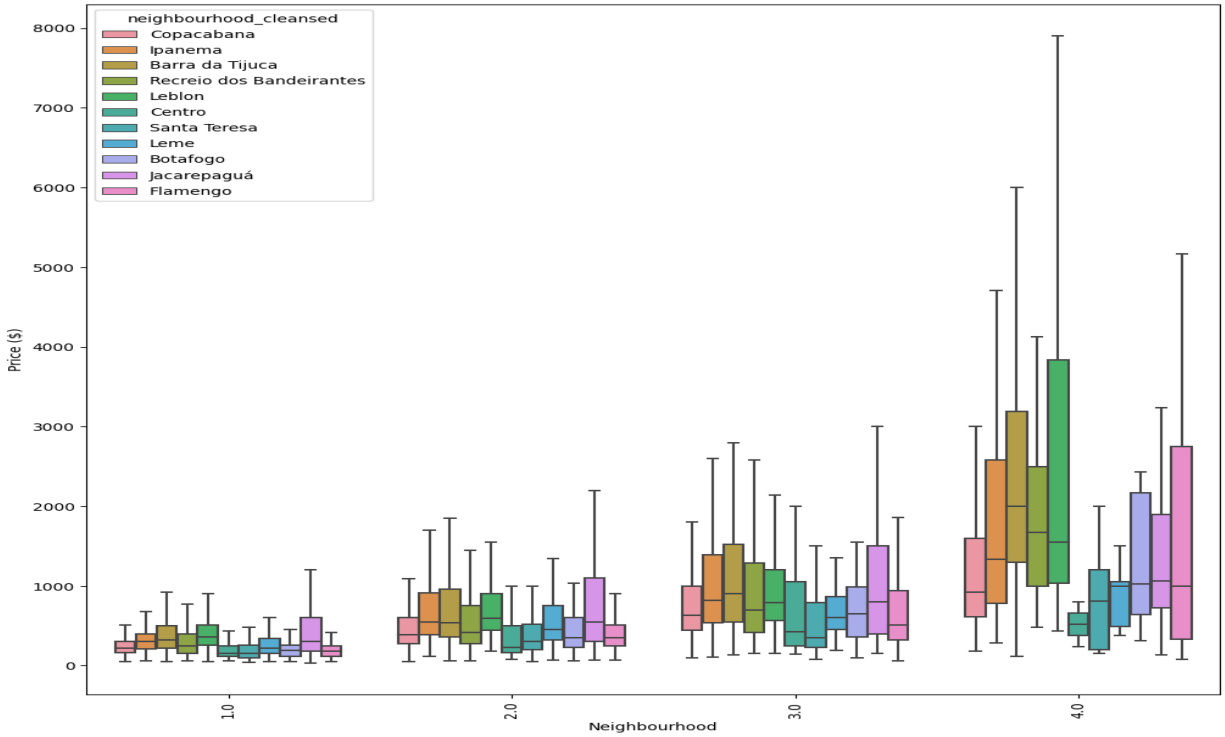
In Jersey City, there are significant price differences based on neighborhoods. This is particularly noticeable for listings with single rooms or private rooms, where the neighborhood seems to play a crucial role in pricing. Certain neighborhoods, likely those with convenient proximity to Manhattan, may command higher prices, while others offer more budget-friendly options. For listings with multiple beds, the price difference between neighborhoods appears to be less pronounced, indicating a relatively consistent pricing pattern for larger accommodations.

New York City (NYC):



In NYC, the price variation between neighborhoods is generally less pronounced, except for high-price neighborhoods like Manhattan. NYC's diverse and competitive market may contribute to this relative price consistency, with travelers having a wide array of options across neighborhoods. Manhattan price premium increases with the number of beds available.

Rio de Janeiro:

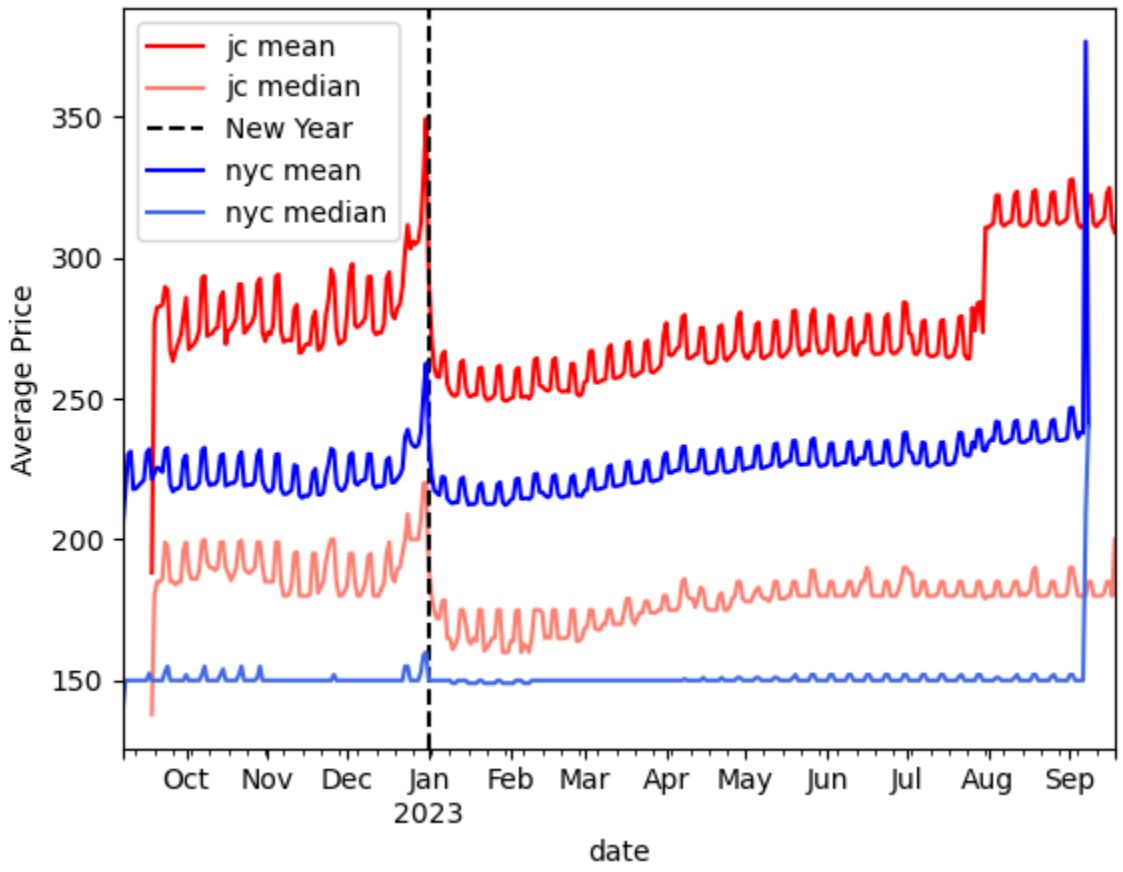


Similar to NYC, Rio de Janeiro exhibits a lower price variation between neighborhoods, except for high-priced neighborhoods such as Ipanema and Leblon. Rio's pricing patterns may be influenced by its diverse attractions and scenic locations, contributing to a more balanced pricing landscape for most listings. The significant price differences in affluent neighborhoods suggest a premium associated with proximity to iconic beaches and landmarks.

These observations underscore the importance of neighborhood considerations in pricing for short-term rentals, especially in cities where neighborhood characteristics strongly influence traveler preferences. Understanding these nuances is essential for both hosts and travelers, enabling informed decisions regarding pricing strategies and accommodation choices.

Visual Analysis of Aggregated Time Series of Listing Prices

Analyzing the aggregated time series of listing prices for Jersey City, New York City (NYC), and Rio de Janeiro provides valuable insights into the seasonality and price dynamics within these short-term rental markets. Here are the key findings based on mean and median price series:

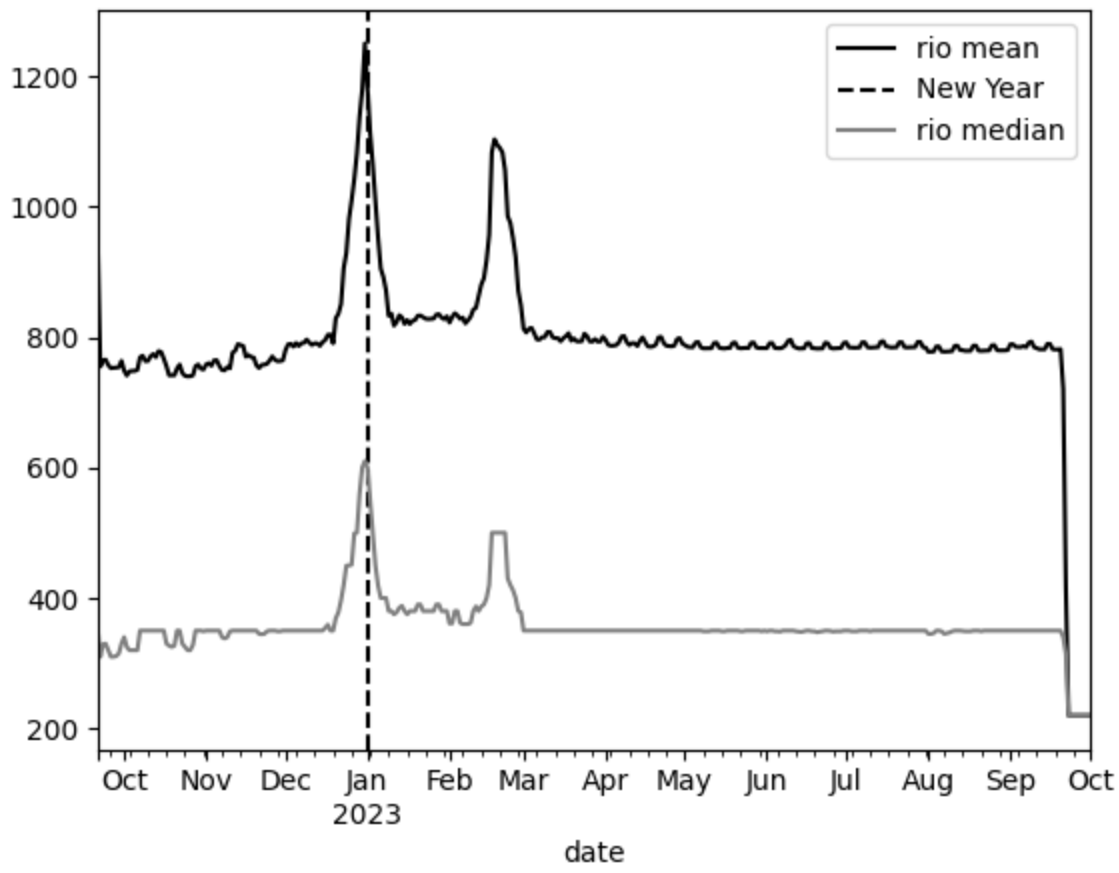


All three datasets exhibit strong seasonality, characterized by recurring patterns in listing prices over time. The peaks in prices are particularly prominent during year-end holidays, indicating higher demand during these periods.

In Jersey City (NJ), there is a distinct division between high-demand and low-demand seasons. End of year holidays and New Year's Eve see significant increases in listing prices, reflecting the holiday demand surge.

Both Jersey City and NYC experience notable price spikes during end of year holidays and New Year's, suggesting that these holidays drive substantial price increases in the region.

In Rio de Janeiro (Rio), the end-of-year price effect lasts longer, and there is also a prolonged Carnival effect, highlighting the impact of these major events on listing prices.

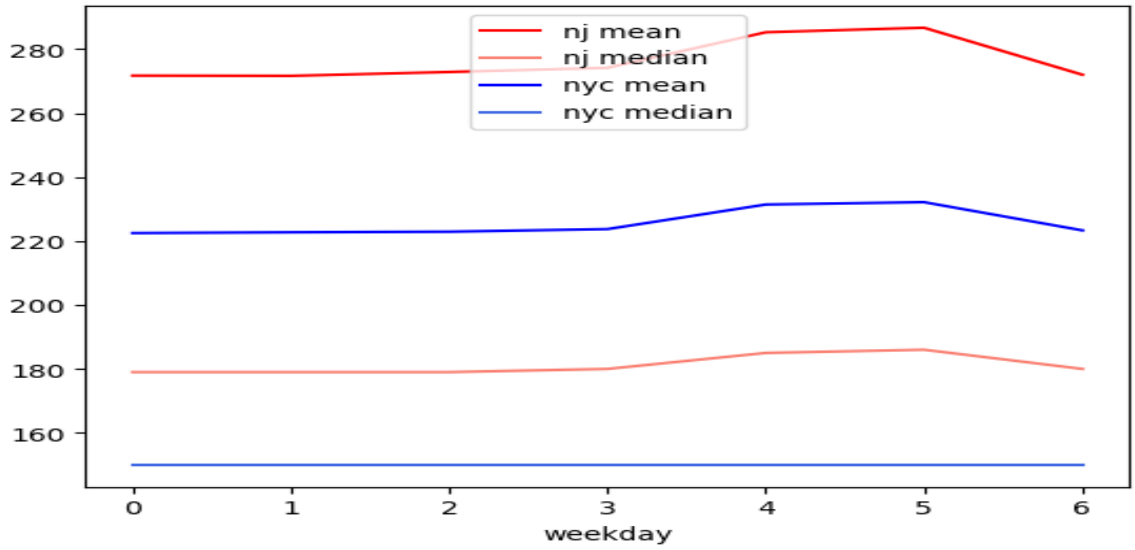


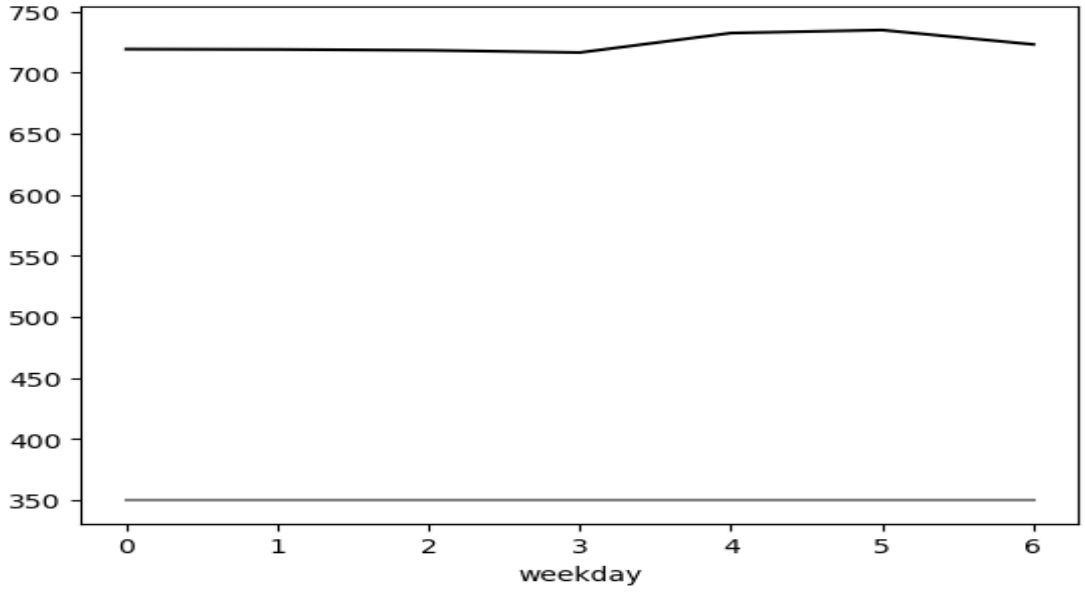
These observations underscore the significance of seasonality in short-term rental pricing across the three cities. The strong influence of year-end holidays, particularly Christmas and New Year's, on listing prices is evident. Additionally, Rio's extended seasonality, driven by both end-of-year celebrations and Carnival, highlights the unique dynamics of the city's short-term rental market.

Understanding these seasonal trends is crucial for hosts and travelers alike, as it allows for informed decisions regarding booking times, pricing strategies, and accommodation choices based on individual preferences and budget considerations.

Effect of Weekday on Prices

In our analysis, we examined the impact of weekdays on listing prices in Jersey City, New York City (NYC), and Rio de Janeiro. Here are the key findings:





Weekday Premiums:
opportunities.

Table 2 and 3 show the expected price increase due to seasonality

	Jersey City	NYC	Rio
price	0.054186	0.034464	0.028127
price	0.046688	0.040488	0.016723
price	0.086090	0.062983	0.028815

Table 2 - average week day price premium for Thursday, Friday and Saturday for each city.

Across all cities, Friday and Saturday emerged as the most expensive days for short-term rental listings. However, the premium for these days was relatively modest, with an approximate increase of around 5% for listings with 1 and 2 bedrooms and a slightly higher increase of approximately 9% for 3-bedroom listings.

Jersey City exhibited clear price differentiation based on the day of the week, with noticeable variations in listing prices from one day to another.
In contrast, both NYC and Rio showed variations in average prices based on the day of the week, but these differences were less pronounced for median prices.

bedrooms	Jersey City	NYC	Rio	Rio_carnaval
1	0.222046	0.163604	0.762932	0.376349
2	0.250668	0.160109	0.632019	0.360687
3	0.294405	0.244866	0.777436	0.498302

Table 3 - average season price premium for end of year holidays and for Rio Carnaval for each city.

The analysis suggests that seasonality is correlated with price levels. Listings in all three cities experience seasonality driven by factors such as year-end holidays and, in the case of Rio, Carnival.
In NYC, similar to Jersey City, seasonality appears to be independent of the number of bedrooms, and listings show sensitivity to both the day of the week and year-end holidays.
In Rio, seasonality is characterized by a bimodal distribution, with significant pricing effects observed during end-of-year holidays and Carnival. Additionally, weekday seasonality is evident in Rio's listings. The price premiums for end-of-year and Carnival periods in Rio are notably higher than those in Jersey City and New York, with an increase of over 60% for end-of-year and approximately 30% for Carnival.

These findings underscore the multifaceted nature of seasonality in short-term rental markets, influenced by various factors including day of the week, major events, and price levels. Hosts and travelers can use this information to optimize their strategies for booking and pricing,

aligning with their preferences and budget constraints, and capitalizing on seasonal opportunities.

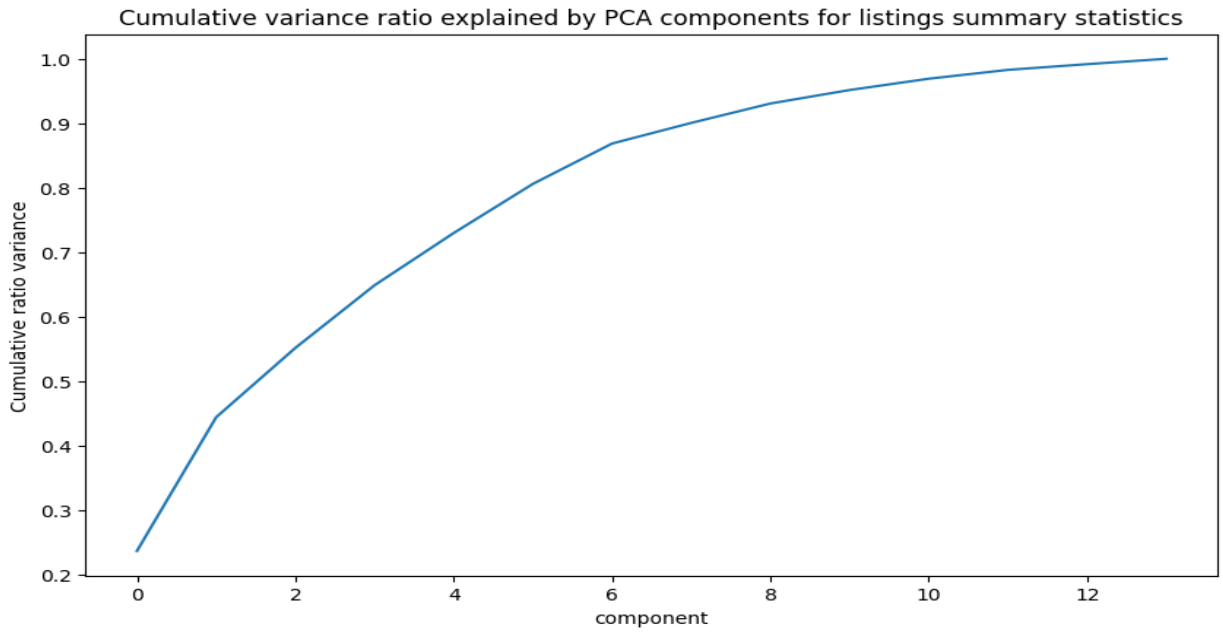
Feature Engineering

In the process of building a pricing model for short-term rentals, feature engineering and feature selection are critical steps to identify the most relevant variables that will contribute to accurate pricing predictions.

We utilize two methods for feature engineering

1. Principal Component Analysis (PCA) Transformation:

The graph below illustrates the cumulative variance explained by Principal Component Analysis (PCA) components for Jersey City. This graphical representation provides insights into how much of the total variance in the dataset is accounted for as we consider each successive PCA component.

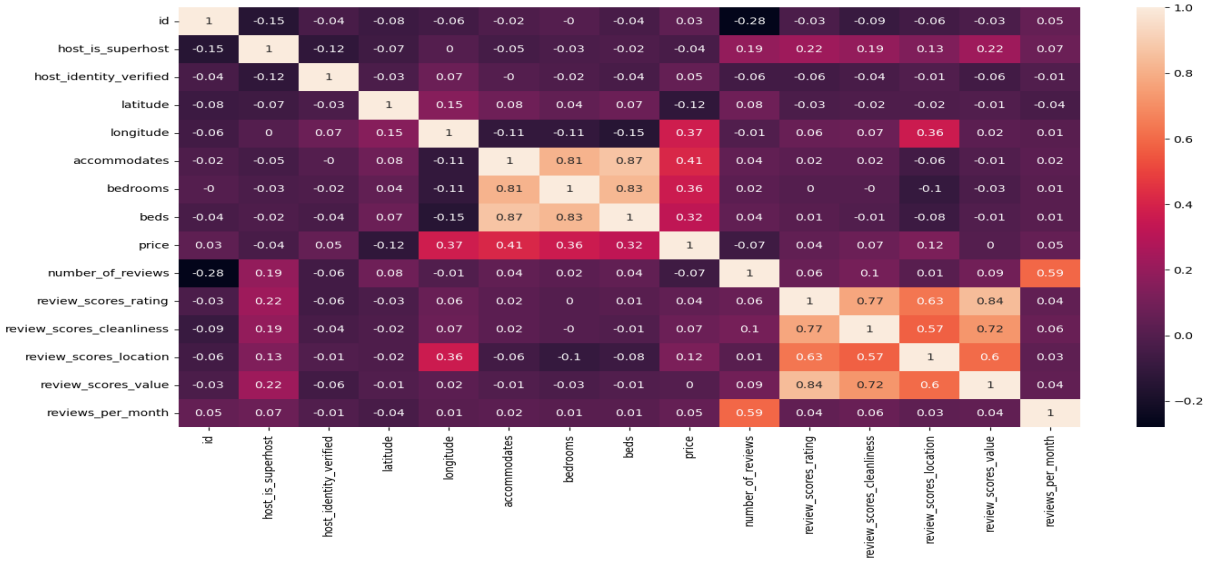


The graphical analysis of Principal Component Analysis (PCA) reveals that each principal component explains only a marginal fraction of the overall variance in the data. Consequently, PCA does not appear to be an effective method for feature selection in this case. Instead, we will place our emphasis on correlation analysis as the primary approach for feature engineering.

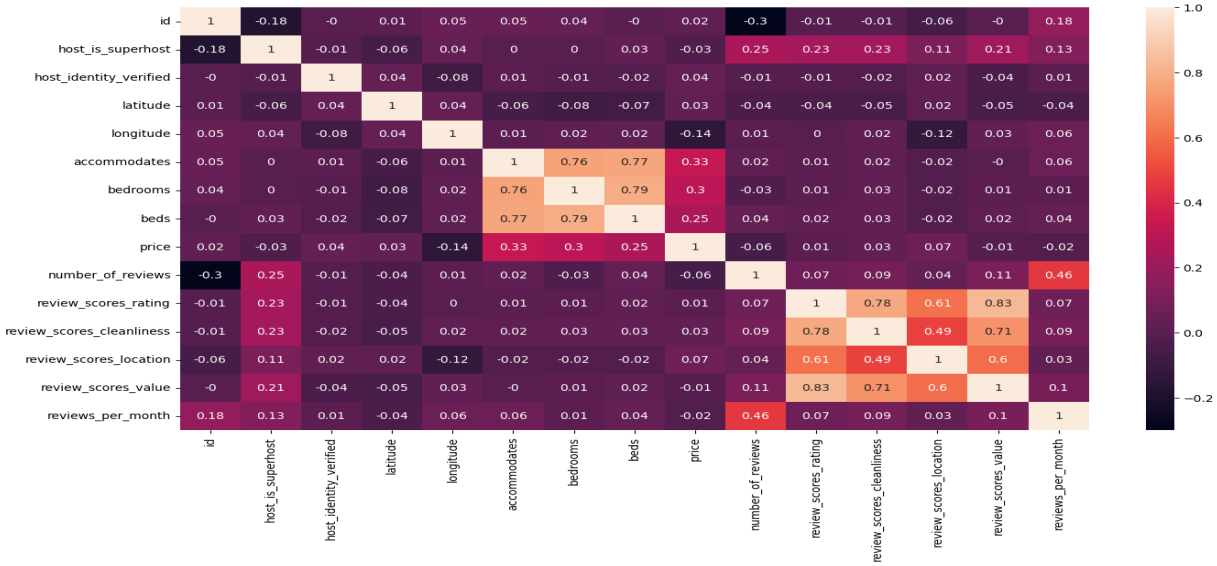
By assessing the relationships between individual features and the target variable through correlation analysis, we aim to identify the most relevant predictors for our pricing model.

2. Correlation Heatmap:

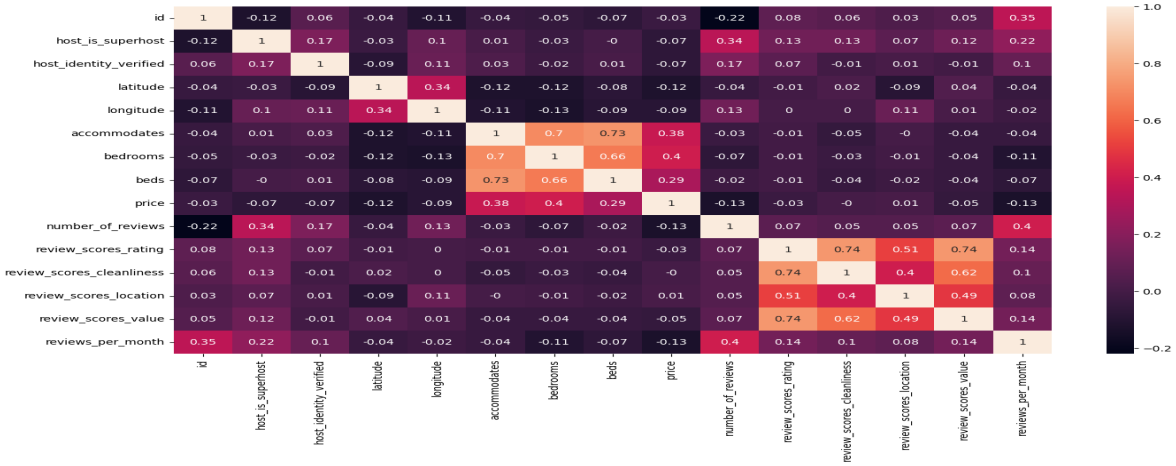
Jersey City:



New York City:



Rio de Janeiro:



The PCA analysis for dimension reduction didn't yield conclusive results due to the relatively small explanatory power of each principal component. It required the consideration of six principal components to explain approximately 90% of the variation.

On the other hand, the correlation heatmap proved to be a valuable tool in our feature engineering efforts. It revealed several key insights:

Notably, we found positive correlations between listing price and features such as 'accommodates' (the number of people the listing can accommodate), 'bedrooms,' and 'beds.' Additionally, positive correlations were observed with ratings for cleanliness, location, and value across most cities. In Jersey City and New York, host identity verification also displayed a positive correlation with price, although this might be attributed to the number of rooms rented rather than the entire apartment.

Surprisingly, we did not observe a significant positive correlation between the number of reviews or reviews per month and pricing. This suggests that other factors play a more prominent role in determining listing prices.

While 'accommodates,' 'bedrooms,' and 'beds' all showed positive correlations with price, they exhibited high levels of correlation with each other. To mitigate multicollinearity among explanatory variables, we have opted to include only the 'accommodates' feature in our modeling to avoid issues related to cross-correlation.

The correlation between latitude and longitude displayed mixed results. However, we intend to utilize neighborhood information in our modeling, as it offers a more granular understanding of location-based factors that influence pricing.

In summary, correlation analysis played a crucial role in feature selection and engineering, allowing us to identify key predictors that are positively associated with listing prices. By considering these insights, we can develop a robust pricing model that incorporates the most relevant features for predicting short-term rental prices in the selected cities.

We chose the following features for our modeling:

```
features = ['neighbourhood_cleansed', 'room_type','latitude','longitude', 'bedrooms', 'beds',
            'accommodates', 'review_scores_rating', 'review_scores_cleanliness',
            'review_scores_location', 'review_scores_value']
```

To generate a baseline model, we divided our clean dataset, containing only the relevant features, into a training set (70%) and a testing set (30%). We then applied four regression models: Linear Regression, Lasso Regression, Random Forest Regression, and XGBoost Regression.

Here are the results in terms of R-squared (R^2) and Root Mean Squared Error (RMSE) for each of these models for Jersey City:

Linear Regression:
- R^2: 61.78
- RMSE: 468.86

Lasso Regression:
- R^2: 53.10
- RMSE: 490.28

Random Forest Regression:
- R^2: 74.05
- RMSE: 453.39

XGBoost Regression:
- R^2: 71.50
- RMSE: 444.25

These metrics serve as an initial assessment of the model's performance. A higher R^2 value indicates a better fit of the model to the data, while a lower RMSE suggests that the model's predictions are closer to the actual prices.

The Random Forest Regressor and XGBoost (XGB) models have emerged as the top-performing models in our analysis. Given their strong performance, we have decided to focus our efforts on optimizing these two models further.

Preprocessing and GridsearchCV

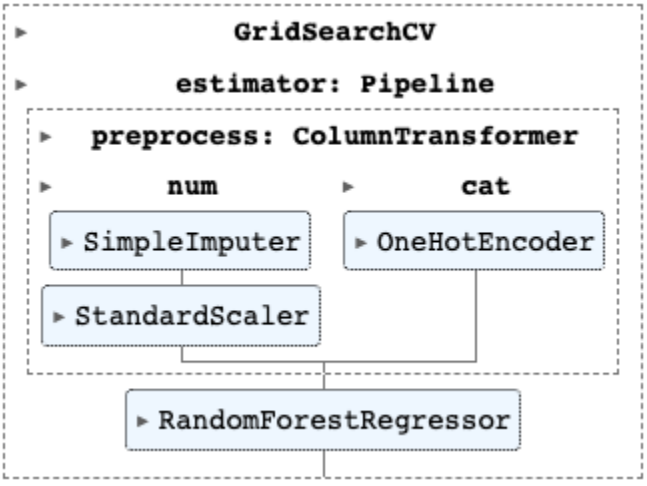
We conducted extensive preprocessing of the dataset and trained models to optimize both the Random Forest and XGBoost models. Our data preprocessing pipeline followed this structured approach:

To address any missing values in the numerical features, we employed a straightforward mean imputer, ensuring that missing data did not impact model training. Subsequently, we applied standard scaling to normalize the numerical data, which aids in bringing all features to a common scale for improved model performance.

For the categorical data, we performed one-hot encoding. This transformation converted categorical variables into a numerical format that can be effectively used by machine learning models. One-hot encoding allows us to represent categorical attributes as binary vectors, enabling the models to understand and learn from these features.

To fine-tune and optimize both the Random Forest and XGBoost models, we implemented a grid search cross-validation approach. This technique systematically explores various combinations of hyperparameters for each model. By assessing model performance through cross-validation, we selected the hyperparameter configurations that produced the best results in terms of accuracy and predictive power.

Pipeline Schema:



By implementing this structured pipeline, we aimed to ensure that our models were trained on high-quality, well-preprocessed data and optimized for their respective algorithms. This approach helps us achieve the best possible predictive performance in our pricing model for short-term rentals.

Optimized Models Results

The Grid Search hyperparameter optimization process had a notable impact in the root mean squared error (RMSE) for all XGBoost models, demonstrating improved model performance. While the Jersey City model showed only marginal improvement on the coefficient of determination (R-squared), there were slight fluctuations in RMSE for the Rio and New York City models—though these changes were also marginal. These results collectively bolster our confidence in the models, suggesting that they are not overfitting the current data.

The table 4 below show the optimized XGBoost results for the cities.

	R^2	rmse
JC	75.37	220.627080
Rio	55.68	912.965461
NYC	69.16	337.548466

Table 4 - Optimized XGBoost R-squared (R^2) and Root Mean Squared Error (RMSE).

It's important to note that the RMSE for the Rio model appears larger than the others. This discrepancy can be attributed to the currency used for pricing, as Rio's RMSE is quoted in Brazilian Reais (BRL) rather than the currency used for Jersey City and New York City. The currency factor contributes to the numerical differences in RMSE and should not be interpreted as a direct reflection of model performance.

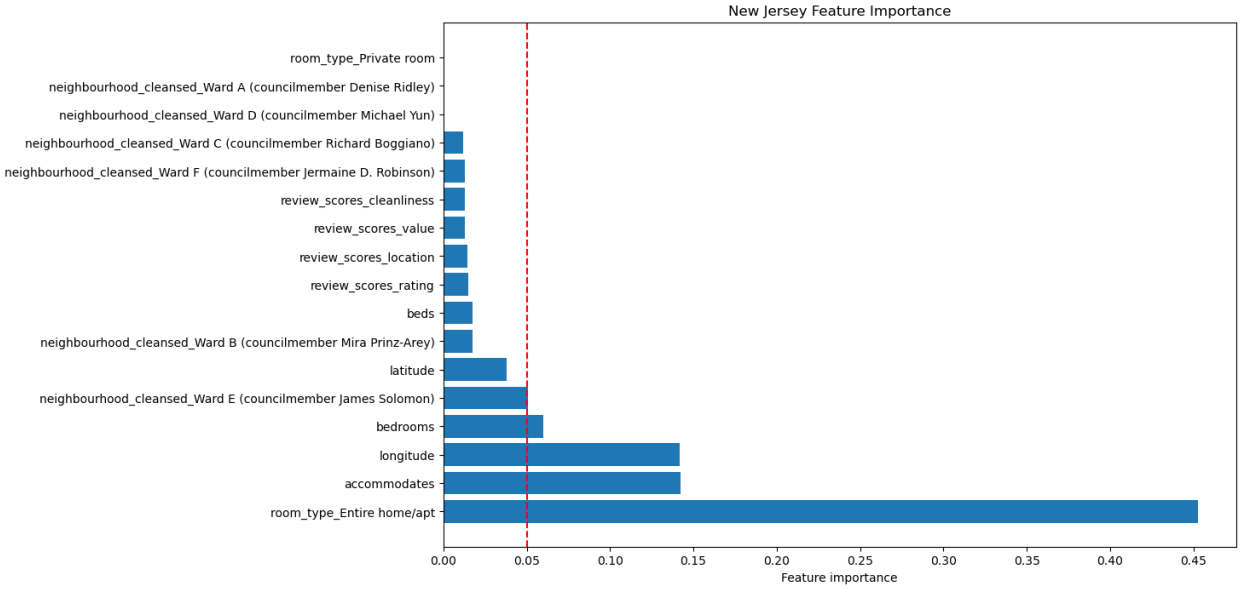
Overall, the improved RMSE values and the stable or slightly fluctuating R-squared values indicate that the models are performing well and are robust enough to provide reliable pricing predictions for short-term rentals in these cities. Further model evaluation and validation may be conducted to ensure their effectiveness in real-world scenarios.

Interpreting Model Results and Insights

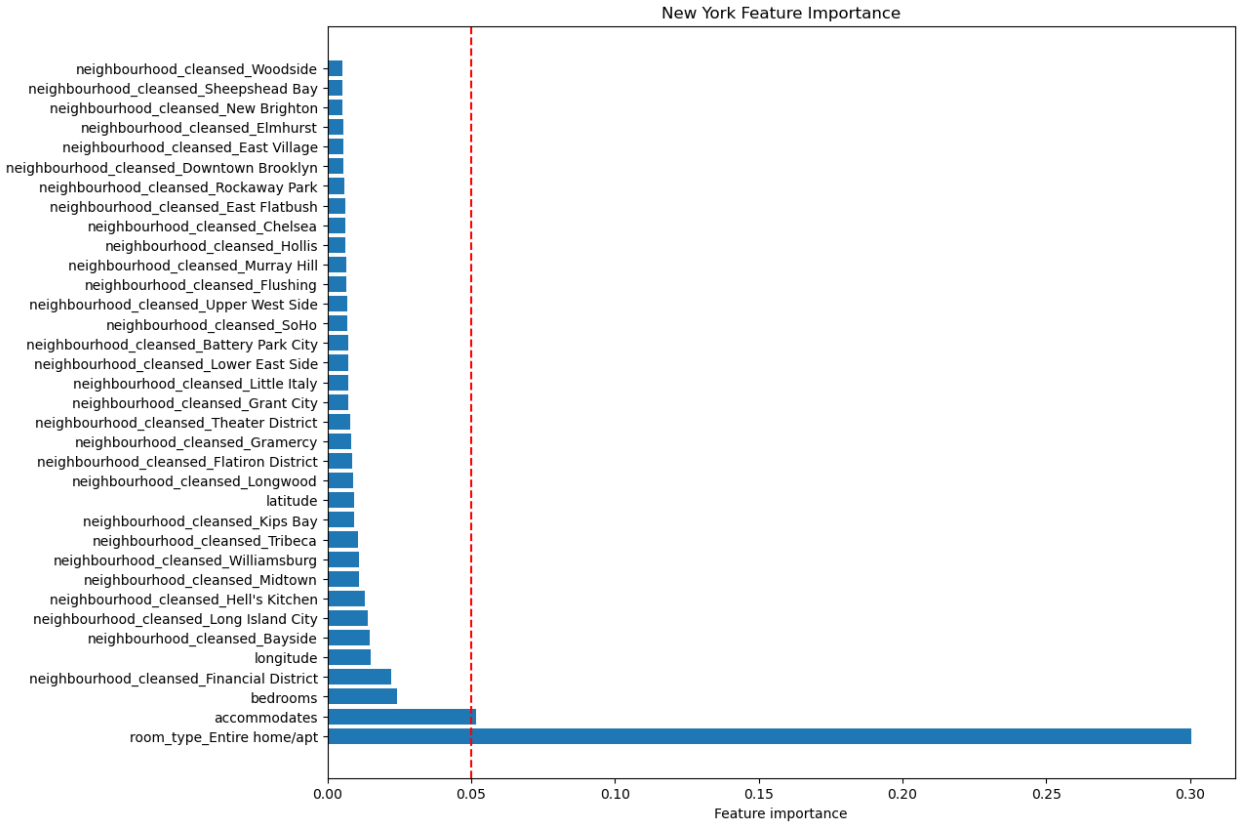
Feature Importance

We conducted a comprehensive analysis of feature importance for all the cities in our dataset. This analysis was undertaken to extract valuable insights that would contribute to the modeling of listing prices across these diverse urban areas. Feature importance analysis serves as a crucial step in understanding the factors that significantly influence pricing decisions in the short-term rental market.

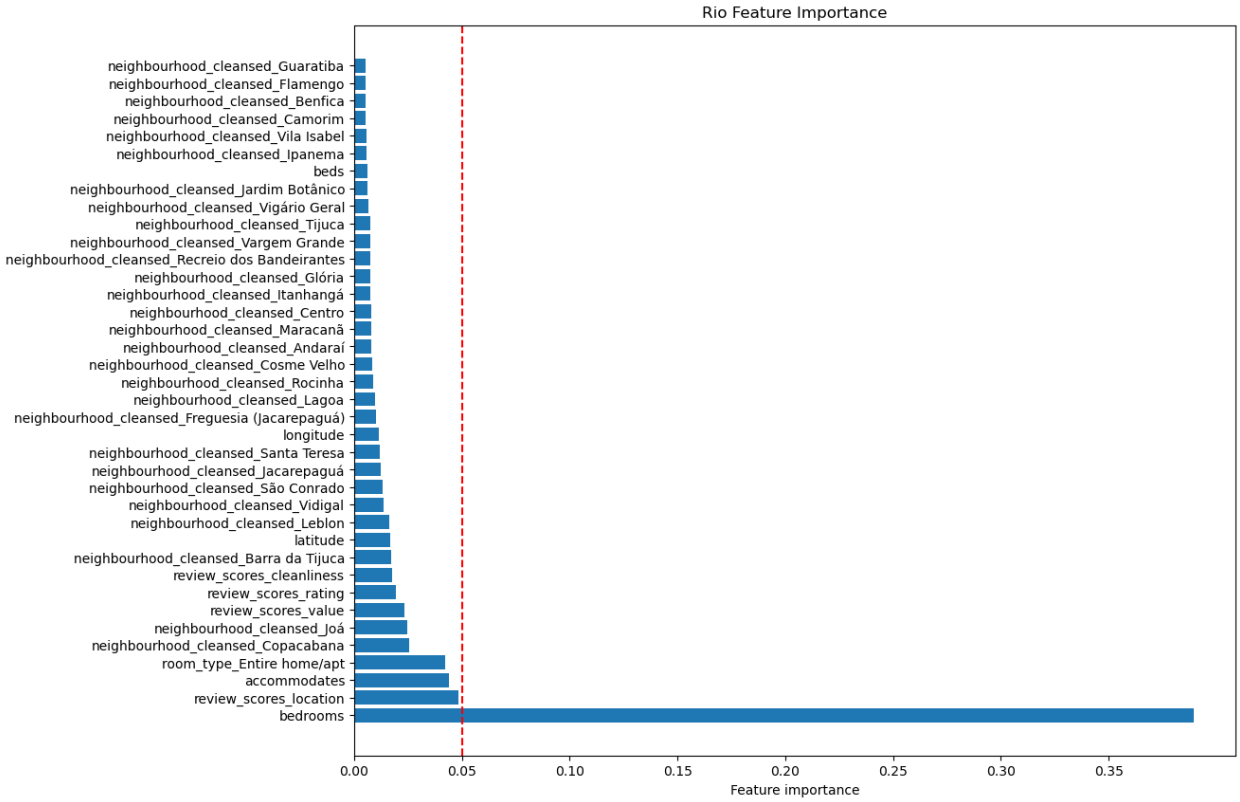
Jersey City:



New York City:



Rio de Janeiro:



The importance of various features in Jersey City, New York City, and Rio de Janeiro reveals insightful patterns specific to each city's short-term rental market:

The importance of features is relatively similar in Jersey City and New York City. The most influential features are 'room type' and 'the number of people a listing accommodates.' Additionally, the number of bedrooms and longitude also hold notable importance.

Among Jersey City's neighborhoods, the 'Ward E' district stands out as the most significant feature. This district's proximity to Manhattan and convenient access to the PATH transportation system make it a key determinant of listing prices. In New York City, the 'Financial District' emerges as the highest-ranking neighborhood in terms of feature importance, reflecting its distinct impact on pricing.

In contrast, Rio de Janeiro exhibits a unique feature importance pattern. Here, the 'number of bedrooms' and the 'review score of listing locations' are the top two features influencing pricing decisions. The 'number of people a listing accommodates' and 'room type' follow closely in terms of importance. Notably, review score metrics, including location, value, overall rating, and cleanliness, occupy prominent positions among the top 10 feature importances. This emphasis on review scores suggests that in a city where safety concerns exist, reviews provide valuable

insights and may serve as proxies for assessing the safety of both listings and their neighborhoods.

Overall, our tuned models demonstrated superior performance for Jersey City and New York City, possibly due to the narrower price range within neighborhoods in these cities. In contrast, Rio de Janeiro exhibits a wider price range even within sought-after neighborhoods, reflecting unique safety challenges.

The inclusion of 'longitude' and 'latitude' as features helped capture the location aspect of listings. However, for Rio de Janeiro, where distinct challenges such as safety risks are present, considering spatial regression and spatial indexing techniques may offer an even better model for understanding the complex dynamics of listing prices in this unique urban environment.

By recognizing these distinctive feature importance patterns, we are better equipped to tailor our pricing models to the specific needs and characteristics of each city's short-term rental market, ultimately providing more accurate and informative pricing predictions for users and stakeholders.

In conclusion, our pricing models for Jersey City and New York City have demonstrated strong performance and stability. These models, which have undergone rigorous fine-tuning and optimization, are poised for practical implementation and utilization by the public. Such implementation can take the form of an accessible API (Application Programming Interface) or a user-friendly web application with an intuitive interface.