Vacation Rental Market Analysis and Revenue Forecasting

Sushma Ramesh , Anish Shivram Gawde , Roshan Chouhan

Northeastern University, Khoury College of Computer Sciences, Boston, MA 02115, USA
ramesh.sus@northeastern.edu,
gawde.ani@northeastern.edu,
chouhan.ro@northeastern.edu

Abstract

This project analyzes the vacation rental market, combining data analysis, machine learning, and intuitive visualizations to provide actionable insights for property owners and vacationers. It enables seamless management of rental properties, bookings, reviews, and listings while offering tools for optimizing revenues and staying competitive. Vacationers benefit from transparent reviews and personalized recommendations to find the perfect stay, while machine learning models predict revenues and reveal market trends to support informed decisions. Developed by a team specializing in data analysis, application development, and database management, the system streamlines vacation rental operations and enhances decision-making for all stakeholders.

Introduction

The *Vacation Rental Market Analysis and Revenue Forecasting* project leverages a robust system designed to analyze and predict revenue trends within the vacation rental market. The data is sourced from vacation rental platforms, such as Airbnb. By incorporating machine learning algorithms, the system provides accurate predictions of revenue based on historical data, occupancy rate, and the amenities provided by the property. This functionality helps property owners optimize pricing strategies, improve occupancy rates, and make informed decisions, ultimately enhancing the overall vacation rental management experience.

Database Design

The system employs a relational database comprising 12 well-structured tables to handle various facets of the vacation rental market, including properties, user feedback, booking transactions, and owner details. Primary keys ensure unique records, while foreign keys establish relationships across tables, such as linking the Bookings table to Users and Properties.
To ensure fast and efficient data retrieval, indexes are created on frequently queried columns like user_id, property_id, and booking_id. This design supports both operational tasks, like property and booking management, and analytical functions, such as revenue forecasting. By organizing data in an interconnected and optimized manner, the database delivers quick response times and robust performance, even with large datasets.

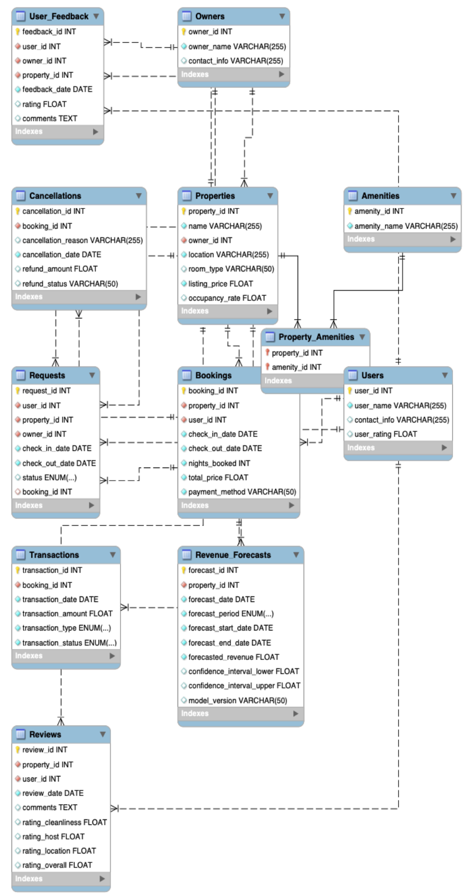


Figure: Database Schema

Key Design Decisions

Normalization:

The database schema for our project was normalized to ensure data consistency, reduce redundancy, and improve efficiency. We adhered to the principles of the Third Normal Form (3NF) to design the database, which involves three stages of normalization:

- **1NF (First Normal Form):** Each column stores atomic values, meaning no repeating groups or multi-valued attributes. Additionally, each table has a unique primary key to uniquely identify records. For example, in the Properties table, the property_id serves as the primary key, ensuring that each property is uniquely identifiable. Also, the Amenities table was extracted from the Properties table to make sure that the atomicity principle is followed and normalization criteria are met.

- **2NF (Second Normal Form):** We eliminated partial dependencies by ensuring that non-key attributes depend on the whole primary key. This was achieved by creating relationships between tables using foreign keys. For instance, the Bookings table references the Users and Properties tables through foreign keys (user_id and property_id), ensuring that the data in Bookings is dependent on both the booking and the property being booked.

- **3NF (Third Normal Form):** We removed transitive dependencies, meaning that non-key attributes do not depend on other non-key attributes. For example, information about the property owner is stored in the Owners table, and the Properties table only contains the owner_id as a foreign key, ensuring that the properties and their owners are connected without storing redundant data in multiple places.

Relationships:

The database schema includes various types of relationships between tables, ensuring that the data is logically connected and easily accessible for analysis:

- **One-to-Many Relationships:**
    - **Property to Reviews:** A property can have multiple reviews submitted by different users, which are stored in the Reviews table.
    - **Property to Bookings:** A property can be booked multiple times, and each booking is stored in the Bookings table, linking back to the relevant property.
    - **User to Feedback:** A user can provide feedback for multiple properties, stored in the User_Feedback table, with the user_id linking each feedback to a specific user.

- **Many-to-One Relationships:**
    - **Booking to Property:** Each booking corresponds to one property, ensuring that each booking record in the Bookings table is tied to a specific property.
    - **Review to User:** Each review is linked to a specific user who submitted it, ensuring that the Reviews table contains a reference to the corresponding user.

- **Many-to-Many Relationships:**
    - **Properties to Amenities:** The many-to-many relationship between Properties and Amenities represents a scenario where a property can have multiple amenities, and an amenity can be associated with multiple properties.

Schema Structure Summary

Core Tables:

- **Users:** Stores information about users, including contact details and ratings.
- **Owners:** Contains details about property owners, including contact information.
- **Properties:** Stores information about rental properties, including location, room type, and pricing.
- **Bookings:** Captures details of bookings, such as check-in and check-out dates, total price, and payment methods.
- **Reviews:** Stores user reviews and ratings for properties, covering multiple dimensions like cleanliness and location.
- **User_Feedback:** Captures general feedback provided by users for properties, including comments and ratings.

Auxiliary Tables:

- **Amenities:** Lists available amenities, such as Wi-Fi, parking, and air conditioning.
- **Property_Amenities:** Establishes a many-to-many relationship between properties and amenities.
- **Requests:** Tracks reservation requests made by users, including their status (Pending, Approved, Rejected).
- **Transactions:** Logs payment details for bookings, including transaction type and status.

- **Cancellations:** Tracks details of canceled bookings, including reasons and refund amounts.
-

Machine Learning Tables:

- **Revenue_Forecasts:** Stores forecasted revenue data for properties, including confidence intervals and model version metadata.
- **Indexing**: Each table includes indexes to enhance search performance based on frequently queried fields (e.g., user_id, property_id, booking_id).
- **Normalization**: The schema is normalized to avoid data redundancy. Foreign keys link related tables to maintain referential integrity.

This database design will allow efficient data management for the vacation rental market system, supporting both operational tasks (like property listings and bookings) and analytical tasks (like revenue forecasting and market trends analysis).

Application Description

The **Vacation Rental Analytics and Revenue Forecasting** application is designed to interact with a robust database, providing a user-friendly interface for querying, retrieving, and analyzing data related to vacation rental properties, bookings, revenue forecasts, and market trends. Its primary purpose is to simplify property management and decision-making for both property owners and renters. The application enables property owners to manage listings, set dynamic pricing, and track performance metrics, while renters can browse available properties, submit booking requests, and provide feedback. Additionally, the system integrates machine learning models for pricing predictions and revenue forecasting, enhancing its ability to offer data-driven insights and improve rental market strategies.

Main Features:

- **SQL Query Execution:** Users can interact with the system by executing SQL queries through a query interface. The application provides real-time query results, displayed in an easy-to-read table format, to assist users in making data-driven decisions.

- **Predefined Queries:** To simplify the process, the system provides a set of predefined queries for common tasks, such as retrieving detailed information on properties, user reviews, and revenue predictions. This feature reduces the need for manual query writing, making it more user-friendly.

- **Error Handling and Reusability:** The application is designed to handle errors gracefully. In case of a failed query or connection issue, the system will display clear error messages, allow users to restart the process, and roll back any incomplete transactions to maintain data integrity.

Data Visualization:

Data visualization plays a pivotal role in uncovering trends, patterns, and insights from the vacation rental data. Key visualizations and insights for this project include:

1. **Price Distribution**: Visualizing the distribution of property prices by location, property type, and seasonality helps identify pricing trends. Boxplots and histograms were used to detect price ranges and the effect of factors such as property size or amenities on pricing.
2. **Occupancy Rate Trends**: Time series plots were used to show occupancy rate trends over time, helping to identify seasonal variations, peak rental periods, and occupancy fluctuations across different locations and property types.
3. **Revenue Forecasting**: Visualizations like line charts and scatter plots are employed to present the predicted revenue for properties, helping property owners adjust pricing strategies according to demand fluctuations and competitor performance.
4. **Market Segmentation**: Clustering analysis results were visualized using scatter plots to identify property segments with similar characteristics. This provides insights into targeted pricing strategies and helps market properties to the most relevant audience.
5. **Location Demand**: Heatmaps were used to show the geographic distribution of rental demand across different regions, helping property owners understand high-demand locations and adjust their pricing strategies accordingly.

Extensible Architecture:

The architecture of this system is designed to be highly extensible to accommodate future growth and additional features. Key features of this extensible architecture include:

1. **Modular Data Storage**: The database schema is designed with multiple interconnected tables (e.g., Properties,

Owners, Bookings, Reviews) that can easily be expanded to include additional data points such as property images, amenities, or historical pricing trends. This modular approach ensures that new features or data sources can be incorporated seamlessly without disrupting existing functionality.
2. **Scalability**: The system is built to scale horizontally, meaning that as the volume of data grows, additional resources can be added to handle increased demand. This scalability ensures that the system remains responsive even as more properties, bookings, and reviews are added.
3. **APIs for Integration**: The architecture allows for easy integration with external APIs, such as weather data providers, local event schedules, and new vacation rental platforms. This enables the system to adapt to new data sources without requiring major redesigns.
4. **Data Pipelines**: Data pipelines are established to automate the process of data collection, cleaning, and preprocessing. This ensures that the dataset is always up-to-date and ready for analysis without manual intervention. The system also supports continuous data flows, which allows for real-time market analysis and dynamic pricing strategies.

Data Collection

The data for this project was collected from Airbnb's official data source called Inside Airbnb to ensure a comprehensive and reliable dataset for vacation rental analysis. The primary data sources include vacation rental platforms, user-generated inputs, and external data sources. Data was gathered through:

- **Vacation Rental Platforms**: Data scraping techniques were employed to extract property details, pricing information, availability, and reviews from leading vacation rental platforms like Airbnb. This data provided valuable insights into property performance, location demand, and pricing trends.
- **User Input**: Property owners and renters contributed data by submitting new rental listings, updating existing listings, and providing reviews and ratings. This ensures the database remains up-to-date and includes firsthand information on properties and their market performance.
- **Synthetic Data Generation:** Some important data like transaction history, cancellation history and user contact details were unavailable for the specified properties and users. Hence, we incorporated techniques to produce synthetic data for some tables and columns.

Data Cleaning:

Data cleaning is a crucial step to ensure the quality and integrity of the dataset. For this project, the following cleaning tasks were carried out:

- **Duplicate Removal**: Duplicate entries, which can skew analysis results, were identified and removed to maintain the uniqueness of each rental listing.
- **Handling Missing Data**: Missing values in critical fields, such as rental prices, occupancy rates, and reviews, were handled using imputation techniques:
    - **Numerical Values**: Median or mean imputation was used for numerical fields.
    - **Categorical Data**: Mode imputation was used for categorical fields. In cases where imputation wasn't suitable, rows with excessive missing data were removed to ensure dataset accuracy.
- **Standardizing Data Formats**: Various fields such as location, pricing, and review ratings were standardized to maintain uniformity across the dataset. For example, location names were standardized to a consistent format (e.g., city, state, country), and price data was converted to the same currency unit.
- **Removing Outliers**: Outliers in key metrics like rental price or occupancy rate were detected using statistical methods (e.g., Z-scores) and removed to prevent them from distorting analysis results.
- **Data Normalization**: Data normalization was applied to ensure consistency across different property types, locations, and categories. This included scaling numeric variables to a common range to facilitate better comparison and model performance.

Metrics:

To evaluate the performance of the regression model, various statistical metrics are used. These metrics provide insights into the accuracy of predictions, the magnitude of errors, and the model's ability to explain the variance in the target variable.

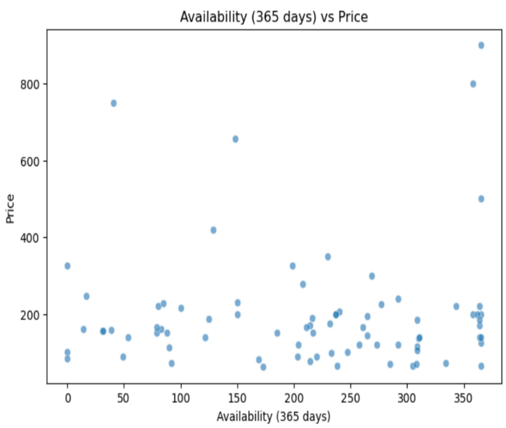
- **Mean Absolute Error(MAE):** Measures prediction error magnitude.

- **Mean Squared Error (MSE)**: Measures squared prediction error magnitude.
- **R2:**Indicates the proportion of variance in the target variable explained by the model.

Exploratory Data Analysis

EDA was performed to understand the underlying structure of the data, identify patterns, detect outliers, and assess relationships between variables. This step ensures that the data is clean, relevant, and ready for building an effective predictive model.

Scatter Plot: Availability (365 days) vs. Price



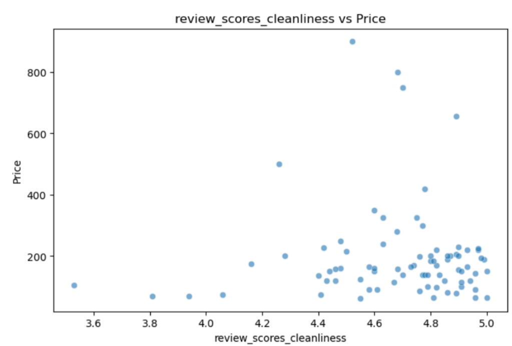
Key Observations:

1. **Price Distribution Across Availability**:
   - The majority of properties fall within a price range of **$0-$200**, regardless of availability.
   - A few high-priced properties (above **$500**) are scattered throughout, indicating luxury listings that maintain availability year-round.
2. **Availability Trends**:
   - Properties with **lower availability (0-100 days)** are just as likely to be high-priced as properties with higher availability, suggesting variability in pricing strategies based on target markets rather than availability alone.
   - Clusters of properties are observed near **full-year availability (300+ days)**, indicating consistent demand for year-round bookings.
3. **Outliers**:
   - There are extreme outliers with **very high prices (>$800)** even at lower availability, which might represent premium or unique listings like luxury villas or niche properties.

Applications:

1. **Property Owners:**
   - Properties with consistent availability throughout the year may command premium prices, especially in high-demand areas.
   - Owners can analyze competitor pricing strategies to optimize their rates based on availability.
2. **Dynamic Pricing Strategies:**
   - For properties with low availability, owners can consider higher rates during peak seasons to maximize profitability.
3. **Demand Prediction:**
   - Helps predict demand for properties based on availability trends and adjust listing duration accordingly.
4. **Market Segmentation:**
   - Identifies specific market segments (budget, mid-range, luxury) and how they correlate with availability duration.

Scatter Plot: Review Scores for Cleanliness vs. Price



Key Observations:

1. **General Trend**:
   - Properties with higher **cleanliness review scores (4.6 to 5.0)** are more densely clustered within the **$0-$200 price range**, indicating affordability does not compromise cleanliness for most properties.
   - There is no strong correlation between **cleanliness scores** and

price, as high-priced properties are observed across the spectrum of review scores.

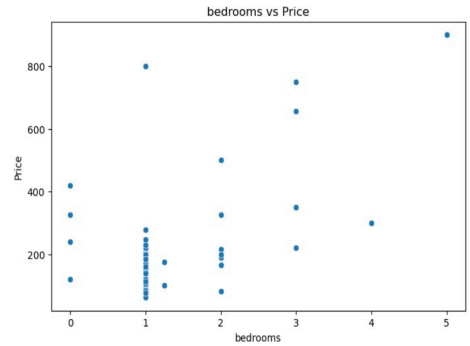
2. **High-Priced Outliers**:
   - A few outliers with **high prices (>$600)** have cleanliness scores between **4.4 to 4.8**, suggesting that cleanliness alone may not drive higher pricing but could be paired with other factors like amenities, location, or property type.

3. **Properties with Lower Scores:**
   - Listings with cleanliness scores below **4.0** are limited in frequency and tend to have lower prices, potentially indicating a lower demand for these properties due to perceived poor quality.

Applications:

1. **Market Positioning:**
   - High cleanliness scores can enhance brand reputation and justify premium pricing.
   - Budget properties can use cleanliness as a differentiator to stand out in competitive low-cost markets.
2. **Customer Insights:**
   - Cleanliness scores are a major factor in customer decision-making and willingness to pay higher prices.
3. **Operational Strategy:**
   - Encourages property managers to invest in housekeeping and cleanliness to boost review scores and revenue potential.
4. **Marketing:**
   - Highlight cleanliness in marketing for properties with strong cleanliness scores to attract cleanliness-conscious guests.

Scatter Plot: Bedrooms vs. Price



Key Observations:

1. **Outliers:**
   - Some properties with fewer bedrooms have significantly higher prices, possibly due to premium locations or additional amenities.
2. **Clustered Trends:**
   - The majority of properties are clustered around 1–2 bedrooms in the $50–$200 range, representing budget and mid-range family accommodations.

Applications:

1. **Upgrade Decision-Making:**
   - Encourages property owners to consider adding bedrooms for higher price potential, particularly in high-demand areas.
2. **Market Segmentation:**
   - Insights help target customers based on group size (e.g., solo travelers, families, or large groups).

Machine Learning Models

Model 1: Linear Regression

Why Linear Regression Was Chosen:

- **Simplicity and Interpretability:** Linear regression provides an easy-to-understand baseline model to establish the relationship between features (like location, room type, etc.) and revenue.
- **Baseline Comparison:** Acts as a starting point to evaluate the effectiveness of more advanced models.
- **Efficiency:** Requires minimal computational resources, making it a quick method for initial insights.

Data Preparation and Sampling:

- **Feature Scaling:** Applied Min-Max scaling to ensure numerical features were on a consistent scale.
- **Handling Multicollinearity:** Checked for multicollinearity using Variance Inflation Factor (VIF) and removed highly correlated features.
- **Data Partitioning:** Split into training (80%) and testing (20%) datasets.

Performance metrics:

Mean Squared Error (MSE): 6656.74
Root Mean Squared Error (RMSE): 81.59
Mean Absolute Error (MAE): 61.83
R^2 Score: 0.39

Insights:

Linear regression struggled to capture non-linear interactions between features, resulting in moderate prediction accuracy. However, it provided a good starting point for feature evaluation.

Model 2: Random Forest Regressor
Why Random Forest Was Chosen:

- **Ensemble Learning:** Combines predictions from multiple decision trees to improve accuracy.
- **Robustness:** Reduces overfitting by averaging predictions across trees.
- **Feature Importance:** Provides insights into the most influential features.

Data Preparation and Sampling:

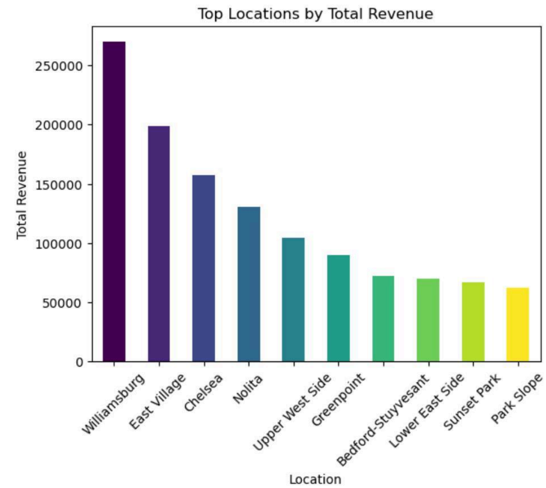
- **Feature Engineering:** Included interaction terms (e.g., occupancy rate × season type) to enhance model expressiveness.
- **Hyperparameter Tuning:** Used RandomizedSearchCV to optimize parameters like n_estimators, max_depth, and min_samples_split.
- **Data Partitioning:** Ensured balanced representation of peak and off-peak seasons in training data.
- R2 Score: 0.742947890936718
- Mean Absolute Error (MAE): 110.82904249142419
- Root Mean Squared Error (RMSE): 344.7090211019369

Insights:

Random Forest delivered the best results, effectively capturing complex relationships and reducing errors. It was robust against overfitting and provided a comprehensive understanding of feature importance, making it the top choice for deployment.

Reports

Bar Graph: Top Locations by Total Revenue



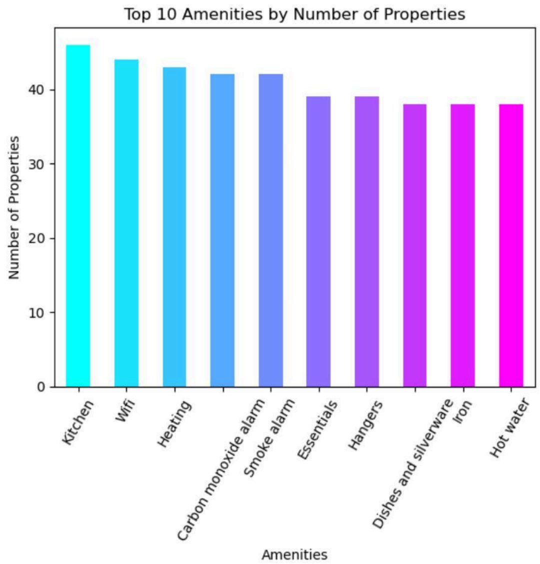
Key Observations:

1. **Williamsburg Leads the Pack**:
   - Williamsburg generates the highest total revenue, crossing **$250,000**. This indicates its popularity among guests, likely due to a combination of trendy appeal, proximity to major attractions, and a vibrant neighborhood vibe.
2. **Revenue Clusters**:
   - Locations such as **East Village**, **Chelsea**, and **Nolita** also perform exceptionally well, generating revenue between **$150,000 and $200,000**. These areas are known for their rich cultural, culinary, and lifestyle offerings, appealing to diverse visitor demographics.

Applications:

1. **Revenue Maximization**:
   - Hosts in Williamsburg and East Village should leverage their high-revenue potential by focusing on premium pricing strategies, especially during peak seasons or events.
   - For emerging areas like Sunset Park, consider targeted promotions or investments in unique amenities to boost visibility and bookings.

Bar Graph: Top 10 Amenities by Number of Properties

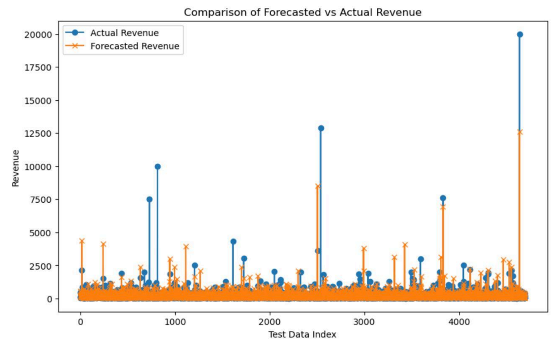


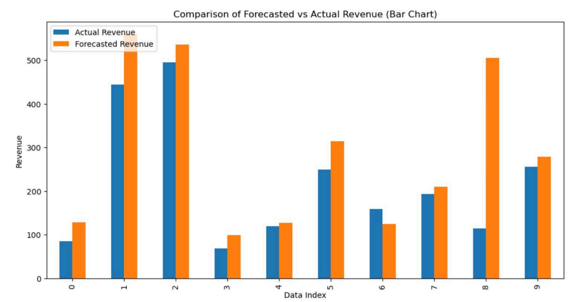
Key Observations:

1. **Popular Amenities**:
   - Kitchen, Wi-Fi, and Heating are the most commonly offered amenities, each available in over 40 properties. These amenities cater to fundamental guest needs, including cooking, connectivity, and comfort during different seasons.
   - Safety features like Carbon Monoxide Alarms and Smoke Alarms are also widely provided, indicating compliance with safety standards.
2. **Essential Items**:
   - Amenities such as Essentials (basic toiletries, linens), Hangers, Dishes and Silverware, and Hot Water highlight a focus on enhancing guest convenience.

Applications:

1. **Improving Competitive Edge**:
   - For properties lacking one or more of these top amenities, adding them could significantly increase bookings and guest satisfaction.
   - Consider enhancing the presence of underutilized but impactful amenities like irons or hot water in budget properties.

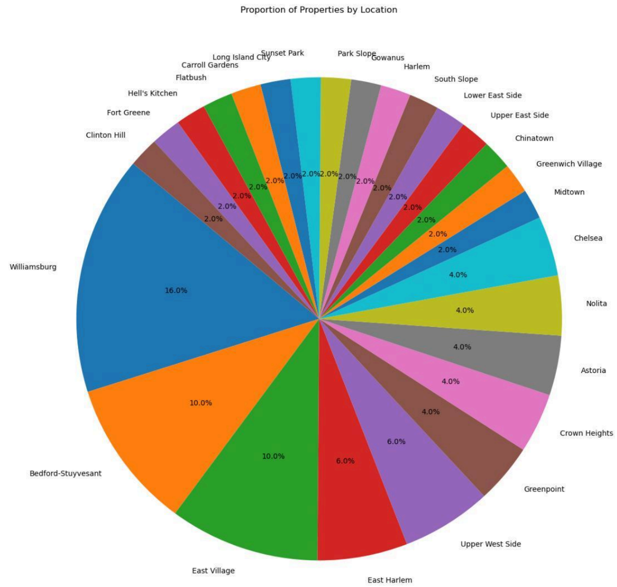




Key Insights:

- General Alignment Between Forecasted and Actual Revenue:
- For most data points, the forecasted revenue (orange crosses) closely follows the actual revenue (blue dots), indicating that the model performs reasonably well in predicting revenue trends.
- Outliers with High Actual Revenue:
- There are some significant outliers where actual revenue is much higher than forecasted revenue. These could be due to unusual factors such as special events, high-demand periods, or exceptional property features not captured in the model.
- Underestimation of Revenue Peaks:
- The model struggles to predict extreme spikes in revenue accurately. This suggests that the model may not fully capture the factors influencing high-revenue properties or scenarios (e.g., premium locations, seasonal trends).

Pie Chart - Top 15 Proportion of Properties by Location:



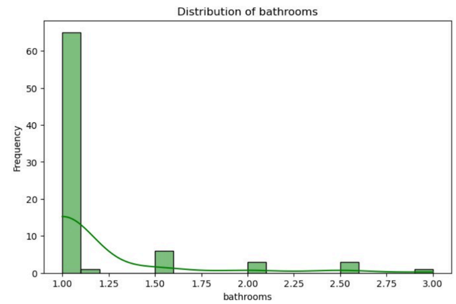
Key Observations:

1. **Most Popular Locations:**
    - **Williamsburg** dominates the location charts, with **16%** of properties belonging to it.
    - Following it are **Bedford and East-Village** with **10%** of the properties.
    - Popular areas like Manhattan and Midtown have less properties, suggesting high property values.

Applications:

1. **Investment Strategies**:
    - Investors and property managers can prioritize Williamsburg, Bedford and East Village for property investments due to their large share and high demand.
2. **Pricing Strategies**:
    - Williamsburg might sustain higher prices due to demand, whereas smaller areas might need competitive pricing.
    - Use competitive pricing in less popular areas to appeal to budget-conscious travelers.

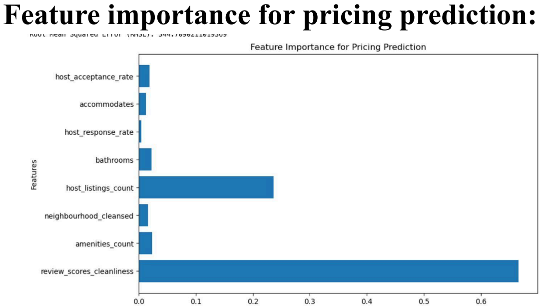
Bar Graph : Bathrooms and their frequency:



Key Observations:

1. **Single-Bathroom Listings Dominate**:
    - Listings with **1 bathroom** are the most frequent, with a count nearing **60**. This suggests that single-bathroom properties are the most popular or abundant among the dataset.
2. **Gradual Decrease with Higher Bathroom Counts**:
   As the number of bathrooms increases, the frequency of listings declines. For instance:
    - **1.25–1.75 bathrooms** are moderately common, with frequencies between **30 and 40**.
    - Properties with **2 or more bathrooms** are less frequent, with counts below **20**, indicating they cater to a niche segment.

Feature importance for pricing prediction:

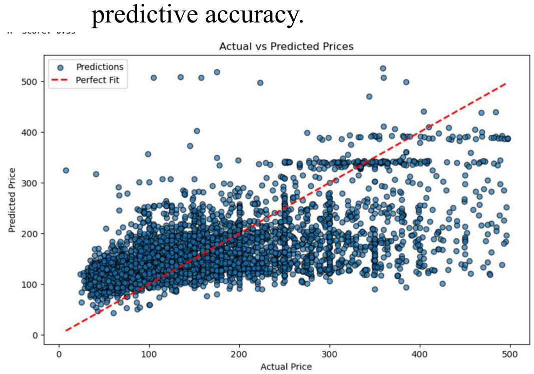


Key Features:

- review_scores_cleanliness has the highest importance, indicating it strongly influences the price of properties. Customers likely value cleanliness, making it a significant factor in pricing.
- host_listings_count is the second most important feature. This might indicate that experienced hosts with more listings can charge competitive or higher prices due to

trustworthiness and optimized property management.

- Other features like bathrooms, amenities_count, and accommodates have minimal contributions, showing they have limited influence on price predictions in this model.
- Negligible Features:
- host_response_rate, host_acceptance_rate, and neighbourhood_cleansed contribute minimally, implying they may not directly affect price or are overshadowed by other dominant factors.
- Model Performance:
- The reported Root Mean Squared Error (RMSE) value of 344.79 indicates the average deviation between predicted and actual prices. A lower RMSE implies better predictive accuracy.



Key Insights

Good Fit for Lower and Mid-Range Prices:

The model performs reasonably well for lower and mid-range prices, as seen by the clustering of points near the red "perfect fit" line for actual prices under 200.

This indicates the model captures trends accurately in these ranges.

Overestimation at Higher Prices:

For higher actual prices (above ~300), the model tends to overpredict, with many predicted prices exceeding the actual values. This suggests the model struggles to generalize well for high-priced properties.

Conclusion

The Vacation Rental Analytics and Revenue Forecasting system seamlessly integrates data management and machine learning to empower both property owners and renters. By offering dynamic pricing forecasts, occupancy rate analysis, and market trend insights, the system allows owners to optimize their listings and revenue strategies while providing renters with an intuitive platform to discover and book properties. Personalized recommendations, user reviews, and real-time price adjustments enhance decision-making and the overall user experience. Continued development in machine learning models, scalability, and personalization will further refine its capabilities, ensuring sustained relevance and impact within the vacation rental industry.

Insights Gained

Data-Driven Decisions

- The system enables property owners and renters to leverage a robust data structure capturing location-based demand, seasonal trends, and user feedback.
- Real-time dynamic property management features foster better market alignment.

User-Centric Design

- A seamless interface allows renters to easily search and book properties while providing owners with tools to manage listings and analyze revenue trends.
- Comprehensive dashboards deliver actionable insights into property performance, occupancy rates, and review analytics.

Machine Learning Integration

- Advanced forecasting models analyze historical data, competition, and seasonal fluctuations to predict rental revenues.
- Market trend insights help owners identify peak demand periods and optimize pricing strategies.

Challenges and Solutions

Seasonal Data Variability

Challenge: Limited historical data for specific locations reduced forecasting precision. Solution: Incorporated competitor analysis and external datasets to improve predictions for underserved areas.

User Engagement with Reviews

Challenge: Low participation in submitting reviews impacted property rankings and market insights. Solution: Enhanced user incentives and simplified review submission processes to increase engagement.

Future Directions

Enhanced Features

- Dynamic Pricing Models: Introduce AI-driven pricing adjustments that respond to real-time market changes and demand spikes.
- Advanced Search Capabilities: Implement natural language search features to improve property discovery for renters.

User Personalization

- Customized Recommendations: Leverage historical user data to suggest properties and price ranges tailored to individual preferences.

Advice for Future Projects

Plan for Scalability

- Design database and application layers with future growth in mind to handle increased user demands and larger datasets.

Prioritize User Experience

- Balance advanced features with ease of use to ensure adoption across diverse user groups.

Leverage Automation

- Automate repetitive tasks, such as data cleaning and model evaluation, to focus on innovation and feature development.

Final Thoughts

This system showcases how data-driven tools and machine learning can revolutionize the vacation rental market. By merging a user-friendly interface with advanced analytics, it sets the foundation for smarter decision-making and enhanced property management. The inclusion of personalized features, dynamic forecasting, and scalable infrastructure ensures the platform remains impactful and adaptable as the market continues to evolve.

References

[1] Airbnb. Airbnb data and listings. https://insideairbnb.com/get-the-data/, 2024.

[2] Hyndman, R.J., & Athanasopoulos, G. *Forecasting: Principles and Practice*, 2nd edition, OTexts, 2018. https://otexts.com/fpp3/

[3] Hochreiter, S., & Schmidhuber, J. *Long Short-Term Memory*. Neural Computation, 9(8):1735–1780, 1997. https://www.jmlr.org/papers/volume3/hochreiter97a/hochreiter97a.pdf

[4] Bird, S., Klein, E., & Loper, E. *Natural Language Processing with Python*. O'Reilly Media, 2009. https://www.nltk.org/book/

[5] MacQueen, J. *Some methods for classification and analysis of multivariate observations*. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1, 281-297, 1967.

[6] Pedregosa, F., et al. *Scikit-learn: Machine learning in Python*. Journal of Machine Learning Research, 12:2825-2830, 2011. https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf