Syracuse University
School of Information Studies

IST 718: BIG DATA ANALYTICS

FINAL PROJECT REPORT

Airbnb Occupancy and Revenue Analytics Using Spark
Los Angeles County

Group 2
Advait Narvekar
Jill Karia
Punami Chowdary
Soundarya Ravi

Project Overview

Our project, *Airbnb Occupancy and Revenue Analytics Using Spark - Los Angeles County*, aims to analyze Airbnb data alongside crime statistics and public transportation availability. By examining both internal and external factors influencing occupancy rates and rental prices, our objective is to uncover key insights that help optimize profits. This analysis will provide actionable recommendations for hosts to refine pricing strategies, improve listing appeal, and make informed decisions regarding new installations or renovations of existing properties.

From our literature review (Kalehbasti et al., Dhillon and Eluri, Chapman et al.), we identified that most existing Airbnb prediction models focus primarily on historical pricing data and occupancy rates, often ignoring critical external factors that influence demand. To address this gap, our approach incorporates two external variables that are often overlooked in existing models: local crime rates and access to public transportation. This will allow our model to get a more holistic view of the airbnb market and allow hosts an opportunity to change their strategy according to the market.

In addition to predictive modeling, we are performing post-hoc analysis to generate actionable recommendations for Airbnb hosts. These insights will help hosts make informed decisions, such as: Identifying Key Amenities, recommending optimal locations and providing pricing insights. Our approach combines predictive analytics and strategic insights to help Airbnb hosts stay competitive in an ever-changing market.

Prediction, Inference, and Other Goals

Our primary goals include:

1. **Prediction**: Develop models to predict Airbnb listing prices and occupancy rates.
2. **Inference**: Identify the most influential factors driving price and occupancy rates.
3. **Actionable Recommendations**: Provide insights to help hosts optimize revenue through improved amenities, better pricing, and informed location choices.
4. **Revenue Forecasting**: Enable hosts to predict annual revenue based on occupancy and pricing trends.

Data

Our data primarily consists of 4 separate datasets that were aggregated and merged.

1. Airbnb Listings Data: The Airbnb dataset contains details about the Airbnb listing such as listing IDs, host information, neighborhood, room type, price, availability, number of reviews, and other listing details. The data was obtained from the InsideAirbnb website and has metadata provided to us in the form of a data dictionary. The data pertains to listings in Los Angeles County in California and spans from December 2023 to December 2024. There were 45,533 listings in Los Angeles during that period, so we have 45,533 rows in our dataset and 82 features.

2. Crime Data: The crime data was sourced from the LAPD website, and it includes information pertaining to various crimes committed in Los Angeles. The raw data consists of features such as date of crime, crime category, area, latitude and longitude. The raw data consists of 130,000 rows and was analyzed to identify patterns in crime categories and their geographical distribution

Aggregations: We performed aggregations on each crime category (e.g. weapons-related crimes) to transform it into the required format. We used the geopandas library in pandas to map the latitude and longitude of each crime to a zip code. Additionally, we aggregated the crime categories (e.g. weapons-related crimes) in each zip code to get a count of the total number of crimes committed in each zip code.

Merging: We then merged this dataset with the listing dataset on the zip code column. The result was a dataset that included listing details, and the corresponding crime count for the zip code that the listing is located in.

3. Public Transportation Data: The public transportation data was sourced from the Los Angeles Public Transportation website and consists of two distinct datasets: bus stops and metro stations. The bus stops dataset contains 5,332 rows and 8 columns, providing detailed information about the locations and descriptions of bus stops across Los Angeles, including latitude, longitude, and the corresponding zip codes. The metro stations dataset comprises 130 rows and 9 columns, capturing information such as station names, descriptions, latitude, longitude, and the zip codes they serve.

Data Processing

To transform the raw data into a usable format, we performed several steps. First, the latitude and longitude coordinates of each bus stop and metro station were mapped to their respective zip codes using the geopandas library. Aggregations were then applied to calculate the total number of bus stops and metro stations within each zip code.

Finally, the aggregated transportation data was merged with the listing dataset on the zip code column. This merging process produced a comprehensive dataset that included listing details along with the total count of public transportation access points for the corresponding zip codes.

Feature Engineering

To enhance the predictive power of the models, several feature engineering steps were performed:

1. Handling Missing Values

a. Missing values in numeric columns, such as bathroom_count, were treated as zero.

b. For days_since_last_review, which quantifies the number of days since the last review for an Airbnb listing, missing values were replaced with a placeholder value of -1.

c. When imputing missing numeric values, we grouped listings by zip code (292 unique zip codes) and used the mean of each group to account for location-based variations.

2. Outlier Treatment

a. The price column contained significant outliers, which were addressed by removing 4,800 extreme values that skewed the data.

b. Additionally, a log transformation was applied to the price column to compress large values and reduce the impact of outliers while retaining relative relationships.

3. Encoding Categorical Features

a. Columns like Property Type, Room Type, and Amenities had numerous unique categories. These were consolidated into super-categories to ensure practical use.

b. For Amenities, a count of grouped amenities was added as a feature.

c. The Number of Bathrooms column was adjusted, treating shared bathrooms as 0.5 and private bathrooms as 1.

d. The Host Response Time column was label-encoded using cardinality to preserve the ordinal nature of response times.

4. Temporal Features

Datetime columns were converted into numeric features, such as days_since_listing_creation.

5. Encoding Zip Codes

The 292 zip codes were encoded to allow the model to capture the regional impact of location on listing performance.

These steps ensured the dataset was robust, with cleaner and more relevant features for predictive modeling.

Methods

The project leveraged PySpark to process large datasets efficiently and employed the following supervised machine learning models to predict the price and occupancy for the airbnb listings: Linear Regression, Random Forest, Gradient Boosted Trees.

The models' performance was evaluated using standard regression metrics such as R squared, Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). These metrics provided insights into the accuracy of predictions for price and occupancy for each Airbnb listing.

Linear Regression: linear regression was chosen because it was a good baseline model that would also give us R squared values that indicate the predictive power of each of our features.

R^2 (Price): 52.97

R^2 (Occupancy): 89.99

Random Forest: Random Forest was chosen due to its ability to capture non-linear relationships, its robustness in managing diverse data types and effectiveness with incomplete data. Our Random Forest model did show improved results compared to Linear Regression.

R^2(Price): 63.66

R^2 (Occupancy): 91.61

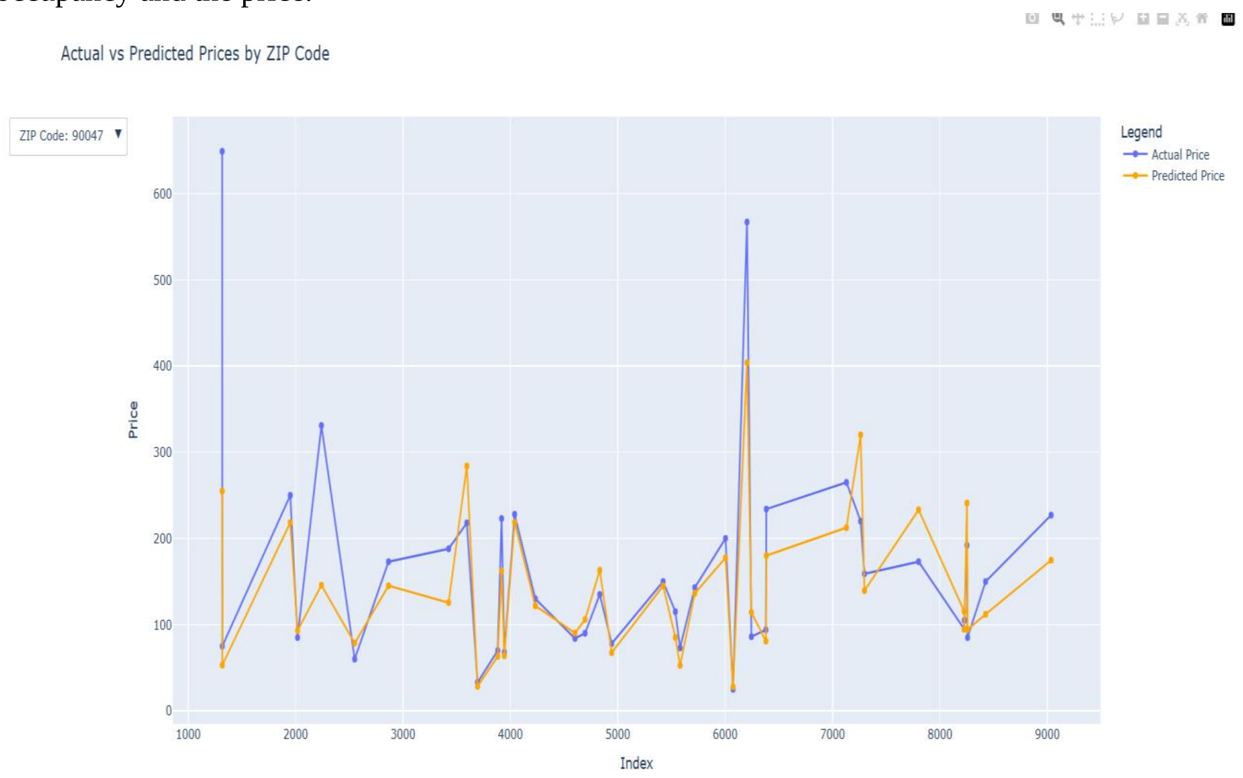
Gradient Boosted Trees: Gradient Boosted Trees provided the best results among all the models and it was chosen for the same reasons as Random Forest.

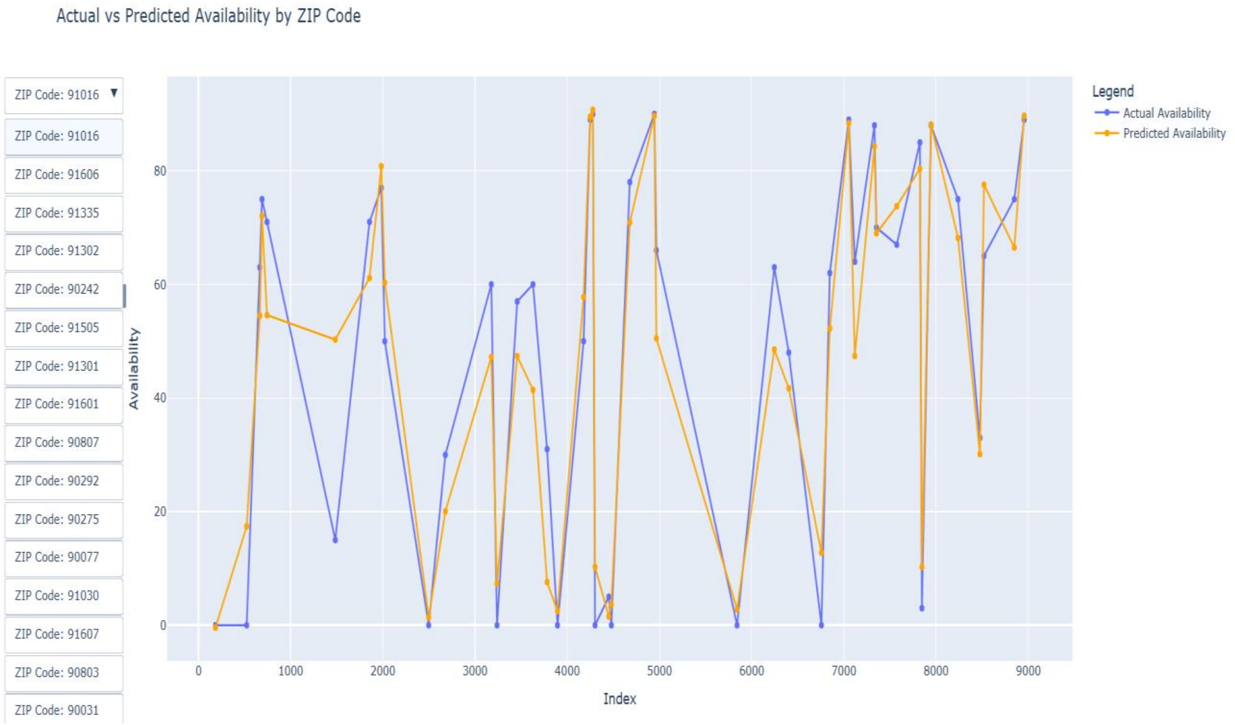
R^2 (Price): 76.80

R^2 (Occupancy): 94.6

Post- Hoc Analysis

Below is the Post-Hoc analysis conducted on each Zip code with the actual and predicted values for both occupancy and the price.





Conclusion: GBT outperformed other models for price prediction, achieving an R² of 76%.
For occupancy, the Random Forest model delivered an R² of 94%, showing robust predictive power.

Feature Importance:

In our Post Hoc analysis, we obtained feature importance scores from our Random Forest model. The intention behind this is to understand which features most influence price and occupancy. This would allow us to provide actionable insights for Airbnb hosts to optimize their listings.

The following features were found to be the most influential in determining the rental price of an Airbnb listing: Weapons-Related Crimes, Superhost Status, Free Wi-Fi and Internet, Number of Reviews, Safety Features, Swimming Amenities, Kitchen Essentials, Family-Friendly Amenities and Number of Bathrooms.

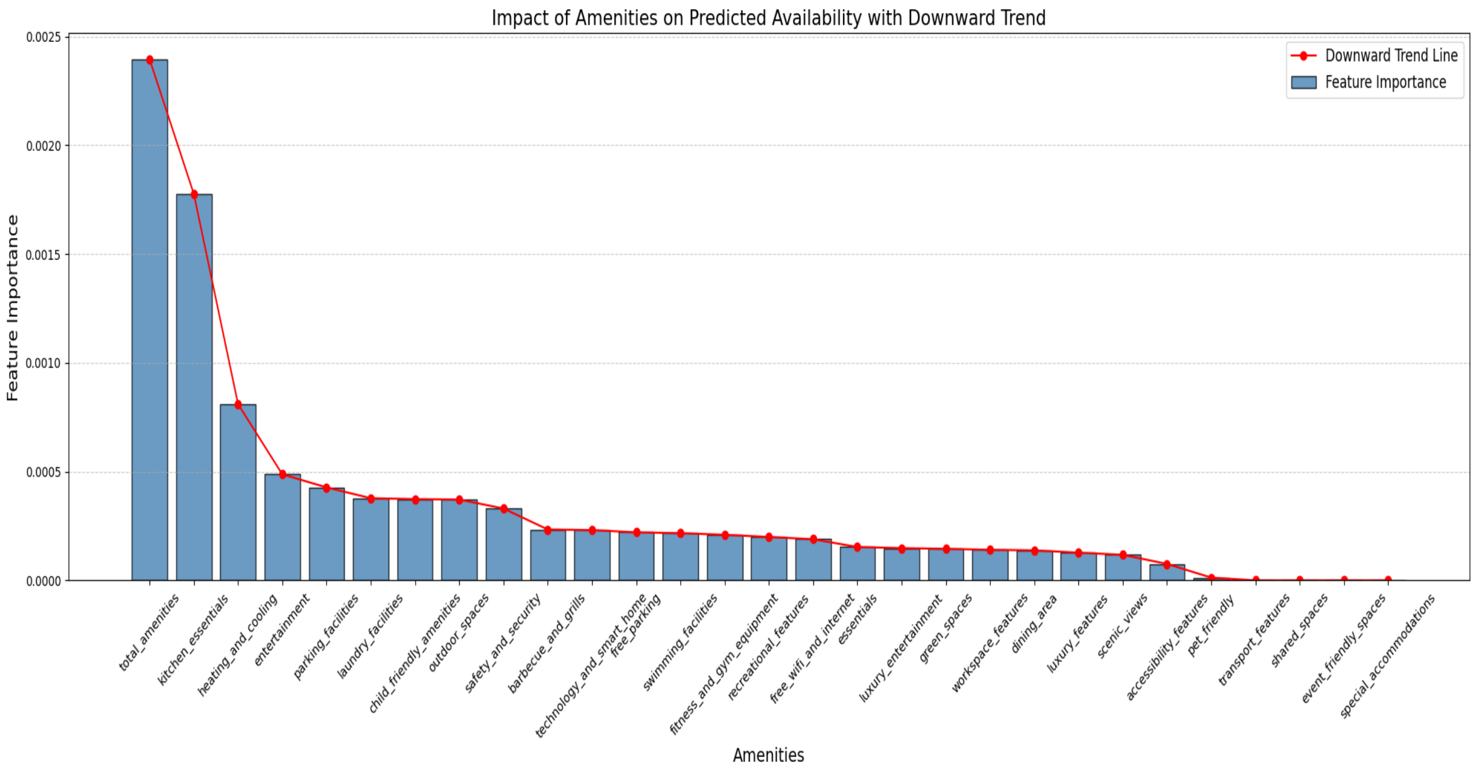
The following features were found to be the most influential in determining the occupancy rates of an Airbnb listing: Years as Host, Host Acceptance Rate, Total Listings Owned by Host, Average Minimum Nights Required, Host Response Time, Time Since Last Review, Total Amenities Offered, Private Room Listings by Host and Host Response Rate.

Recommendations for Stakeholders - Post Hoc Analysis:

In addition to predictive modeling, we are performing post-hoc analysis to generate actionable recommendations for Airbnb hosts. These insights will help hosts make informed decisions, such as: Identifying Key Amenities, recommending optimal locations and providing pricing insights. Our approach combines predictive analytics and strategic insights to help Airbnb hosts stay competitive in an ever-changing market.

Best Amenities: Amenities play a big role in attracting customers and increasing occupancy rates. By analyzing data on the performance of listings with various amenities (e.g., Wi-Fi, kitchen essentials, and parking), the model will identify the most sought-after features that drive bookings and revenue. This can help hosts use their resources effectively to increase their listings' success.

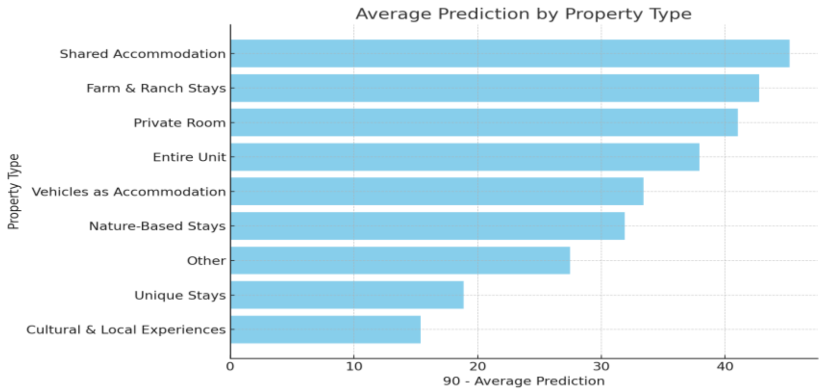
We used the results of our predictive model to highlight which amenities are most correlated with high occupancy rates. As highlighted in the graph below, the number of total amenities in an airbnb is by far the biggest predictor of occupancy rates. A close second is the presence of kitchen essentials and heating and cooling. Parking facilities, laundry facilities, and the presence of entertainment units and child friendly amenities are also significant.



Property Type:

Different property types (e.g., entire homes, apartments, shared accommodation) have different demands depending on location and guest preferences. The model will analyze occupancy rates property types to recommend the most profitable and in-demand categories. This can help hosts invest in properties that have high market demand.

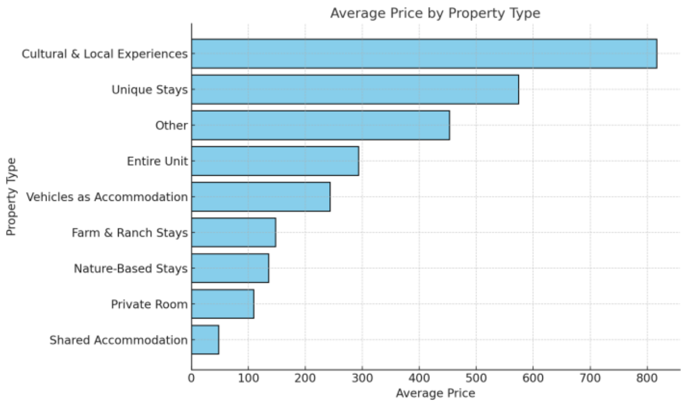
Analyzing the predictions of our model showed us that Shared Accommodations and Farm & Ranch Stays have the highest occupancy rates followed by Private rooms and Entire units. Unique stays and Cultural & Local Experiences have the lowest occupancy rates. The results are summarized in the histogram below.



To help hosts act on our recommendations and invest in the high occupancy property types identified above, we wanted to provide them with pricing insights. We analyzed the average prediction prices for these property types so that hosts know how to competitively price their listings.

We observed an interesting trend in our analysis: property types with the lowest occupancy rates, such as Unique Stays and Cultural & Local Experiences, had the highest average prices and the listing with the highest occupancy rates like Shared Accommodation and Private rooms had the lowest average price. This suggests that while Unique Stays and Cultural & Local Experiences get fewer bookings, they may cater to niche markets where guests are willing to pay a premium price for unique experience.

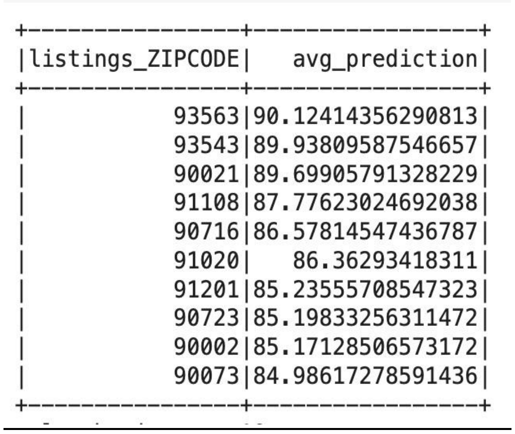
Average Price for each property type:



Location Recommendation:

Choosing the right location is important for the success of an Airbnb. We wanted to help our stakeholders reduce the risk associated with location based decision and help them purchase properties in areas that offer the best return on investment.

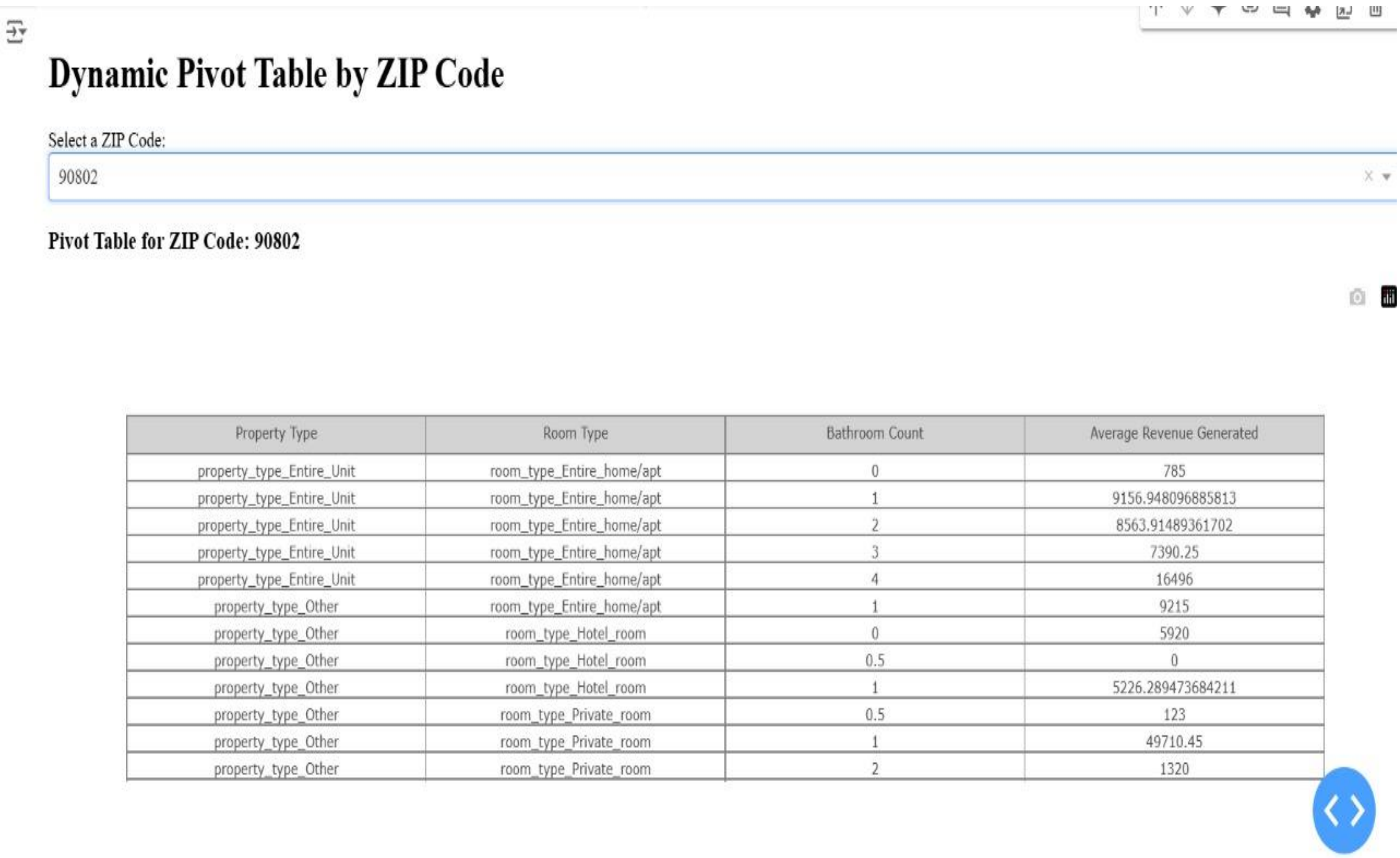
We analyzed the results from our predictive model identifying neighborhoods with the highest occupancy rates. We initially expected popular neighborhoods like 90210 or Hollywood to have the highest occupancy rates. But the zip codes listed below with the highest occupancy all belong to neighborhoods in the outskirts of Los Angeles city. It is possible that high prices and traffic issues are motivating Airbnb customers to choose rentals outside the city.


```
+---------------+-----------------+
|listings_ZIPCODE|   avg_prediction|
+---------------+-----------------+
|          93563|90.12414356290813|
|          93543|89.93809587546657|
|          90021|89.69905791328229|
|          91108|87.77623024692038|
|          90716|86.57814547436787|
|          91020|   86.36293418311|
|          91201|85.23555708547323|
|          90723|85.19833256311472|
|          90002|85.17128506573172|
|          90073|84.98617278591436|
+---------------+-----------------+
```

Revenue Predictions:

We want to give our stakeholders a forecast of their total revenue for the future so they can make informed investment decisions and set realistic expectations.

We calculated the expected revenue for the upcoming 90 days by multiplying the price and occupancy predictions from our model. This analysis was performed for each zip code and further segmented by Property Type, Room Type, and Number of Bathrooms. For example, the first row of the results indicates: In the 90802-zip code, an Entire Unit property type with a room type of "entire home/apartment" and 0 bathrooms has a projected revenue for the next 90 days.



Dynamic Pivot Table by ZIP Code

Select a ZIP Code:

90802

Pivot Table for ZIP Code: 90802

Property Type	Room Type	Bathroom Count	Average Revenue Generated
property_type_Entire_Unit	room_type_Entire_home/apt	0	785
property_type_Entire_Unit	room_type_Entire_home/apt	1	9156.948096885813
property_type_Entire_Unit	room_type_Entire_home/apt	2	8563.91489361702
property_type_Entire_Unit	room_type_Entire_home/apt	3	7390.25
property_type_Entire_Unit	room_type_Entire_home/apt	4	16496
property_type_Other	room_type_Entire_home/apt	1	9215
property_type_Other	room_type_Hotel_room	0	5920
property_type_Other	room_type_Hotel_room	0.5	0
property_type_Other	room_type_Hotel_room	1	5226.289473684211
property_type_Other	room_type_Private_room	0.5	123
property_type_Other	room_type_Private_room	1	49710.45
property_type_Other	room_type_Private_room	2	1320

Surprising Findings

Our analysis revealed several unexpected insights:

1. **Limited Seasonality**: Contrary to expectations, Los Angeles demonstrated very little seasonality in its Airbnb listings. Out of over 45,000 listings, calendar data showed that barely 3,000 exhibited any seasonal trends. As a result, we had to exclude seasonality as a significant factor in our models.
2. **Influence of Internal vs. External Factors**: Internal factors, such as listing features and amenities, played a much greater role in determining both occupancy and pricing compared to external factors like proximity to public transportation or overall crime rates.
3. **Crime and Pricing Dynamics**: Among various crime categories, only weapon-related crimes showed a notable correlation with fluctuations in price and occupancy rates. Other types of crime had minimal impact on Airbnb metrics.
4. **Occupancy and Pricing Patterns**: Shared accommodation exhibited higher occupancy rates, while cultural experience properties commanded significantly higher prices. Interestingly, more than the upscale neighborhoods, guests seemed to prefer the outskirts of Los Angeles, challenging the traditional preference for posh areas.