# 1. Environment setup

## 1.1 Cloud Environment

Platform: Google Cloud Platform (GCP)

Cluster Service: Dataproc (1 master, 2 workers recommended)

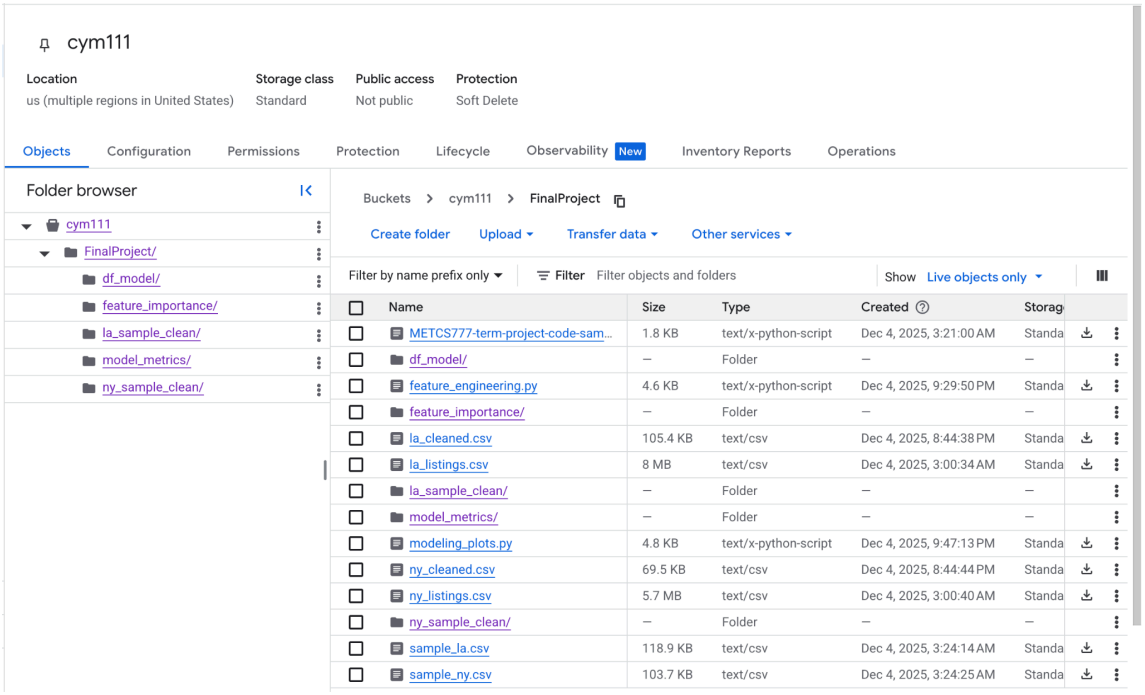
Region: us-central1

Image Version: 2.2-debian12

Python Version: 3.10+

Spark Version: 3.5+

## 1.2 Bucket Structure (GCS)




## 1.3 Dependencies

All required libraries come pre-installed in Dataproc: PySpark ML, Spark SQL, Pandas (driver only). No additional packages are needed.

# 2. How to run the code

## 2.1 Data Cleaning

The data-cleaning stage ensures that both the Los Angeles and New York datasets are consistent, complete, and suitable for modeling. We begin by removing non-predictive or identifier-level fields such as host_id, host_name, and name, keeping only variables that describe listing characteristics, location, availability, and review behavior. Rows containing missing or invalid

values in essential fields, including price, latitude, longitude, room_type, and review metrics, are dropped to maintain data integrity. This step also standardizes column names and formats so that both cities share the same schema. The cleaned datasets are stored in Google Cloud Storage as la_cleaned.csv and ny_cleaned.csv, forming the foundation for the subsequent EDA and modeling workflow.

### 2.2 Running EDA

Exploratory Data Analysis (EDA) is performed using the eda.py script on Google Cloud Dataproc, which reads the cleaned datasets directly from the storage bucket. The EDA process summarizes the statistical properties of each variable, examines distributions such as price and review counts, and highlights differences between the two cities. It also prints correlation insights to help identify which features may contribute meaningfully to pricing. The script is executed with the following command:

```
gcloud dataproc jobs submit pyspark \
  --cluster=finalproject \
  --region=us-central1 \
  gs://cym111/FinalProject/eda.py
```

All results are printed to the job output log, allowing us to verify variable behavior and assess potential modeling issues. This stage provides an essential understanding of how the LA and NY markets differ before constructing predictive models.

### 2.3 Running Feature Engineering

Feature engineering is carried out using the feature_engineering.py script, also executed on the Dataproc cluster. This stage creates modeling-ready variables by applying several transformations. Price is log-transformed to reduce skewness, and a geographic variable, dist_to_center, is computed separately for LA and NY to capture spatial effects. Categorical fields such as room_type and neighbourhood_group are encoded using Spark's StringIndexer and OneHotEncoder, while all numeric and encoded features are assembled into a single vector using VectorAssembler. The result is a streamlined dataset with three core columns, city, log_price, and features, which is fully compatible with Spark ML models. The script is run using:

```
gcloud dataproc jobs submit pyspark \
  --cluster=finalproject \
  --region=us-central1 \
  gs://cym111/FinalProject/feature_engineering.py
```

Upon completion, the script outputs the engineered dataframe (df_model), model metrics, and feature importance tables to Google Cloud Storage. These outputs serve as the basis for evaluating model performance and interpreting pricing drivers across both cities.

### 2.4 Running Modeling

The modeling stage is performed using the modeling_plots.py script, which loads the engineered dataset (df_model) from Google Cloud Storage and trains three separate regression models, Linear Regression, Random Forest, and Gradient Boosted Trees, for both Los Angeles and New York. The script begins by splitting each city's data into training and testing subsets using an 80/20 ratio. Each model is then fit on the training data and evaluated on the test set using RMSE, MAE, and $R^2$ as performance metrics. Random Forest and GBT models additionally provide feature importance values that help identify which variables contribute most strongly to price predictions. The script is executed on the Dataproc cluster with the following command:

```
gcloud dataproc jobs submit pyspark \
  --cluster=finalproject \
  --region=us-central1 \
  gs://cym111/FinalProject/modeling_plots.py
```

Upon completion, the script writes all evaluation outputs back to the GCS bucket, including FinalProject_model_metrics/ and FinalProject_feature_importance/. These results summarize model accuracy across cities and enable a direct comparison of price drivers between Los Angeles and New York. The modeling script therefore serves as the final stage of the computational pipeline and provides the quantitative basis for interpreting urban differences in Airbnb pricing.

## 3. Results of running the code with data & 4. Detailed explanation of the dataset and results

Running the modeling script generated performance metrics for all three models across both cities. Each model was evaluated using RMSE, MAE, and $R^2$ on a held-out test set. The results show that Random Forest provides the best overall performance in both Los Angeles and New York, achieving lower prediction errors and higher explanatory power compared to Linear Regression and Gradient Boosted Trees. These outputs, along with feature-importance values, were automatically saved to Google Cloud Storage and serve as the basis for analyzing how pricing factors differ between the two cities.

### 3.1 Model Performance Metrics

| City | Model | RMSE | MAE | $R^2$ |
|---|---|---|---|---|
| LA | Linear Regression | 0.723 | 0.545 | 0.378 |
| **LA** | **Random Forest** | **0.629** | **0.468** | **0.529** |
| LA | Gradient Boosted Trees | 0.677 | 0.492 | 0.455 |
| NY | Linear Regression | 0.782 | 0.564 | 0.520 |
| **NY** | **Random Forest** | **0.743** | **0.531** | **0.566** |
| NY | Gradient Boosted Trees | 0.854 | 0.551 | 0.429 |

#### 3.1.2 Interpretation of Results

The results indicate that Random Forest is the most effective modeling approach for both cities. In Los Angeles, the Random Forest model achieves the lowest RMSE and highest $R^2$, reflecting its ability to capture the heterogeneous, geographically dispersed nature of LA's housing market. New York exhibits the same pattern, with Random Forest again outperforming the other algorithms and achieving the strongest overall explanatory power. In both cases, Gradient Boosted Trees do not surpass Random Forest, likely due to the dataset's moderate size and noise level, which favor RF's robustness. These outcomes confirm that tree-based ensemble methods are well-suited for Airbnb pricing tasks and provide a reliable basis for feature-importance interpretation in the next stage of analysis.

### 3.2 Feature Importance Analysis

We analyze feature importance from both Random Forest and Gradient Boosted Trees (GBT) for Los Angeles and New York.

Below, feature_index corresponds to the position of each feature in the assembled vector.

The feature importance plots produced by the Random Forest and Gradient Boosted Tree models reference feature positions within the final assembled feature vector. Since the modeling pipeline uses a VectorAssembler, all numerical and one-hot-encoded categorical features are sequentially concatenated. Therefore, understanding model behavior requires mapping each feature index back to its semantic meaning.

**The first nine positions correspond to numerical features:**

**0: minimum_nights**

**1: number_of_reviews**

**2: reviews_per_month**

**3: availability_365**

**4: calculated_host_listings_count**

**5: number_of_reviews_ltm**

**6: dist_to_center**

**7: latitude**

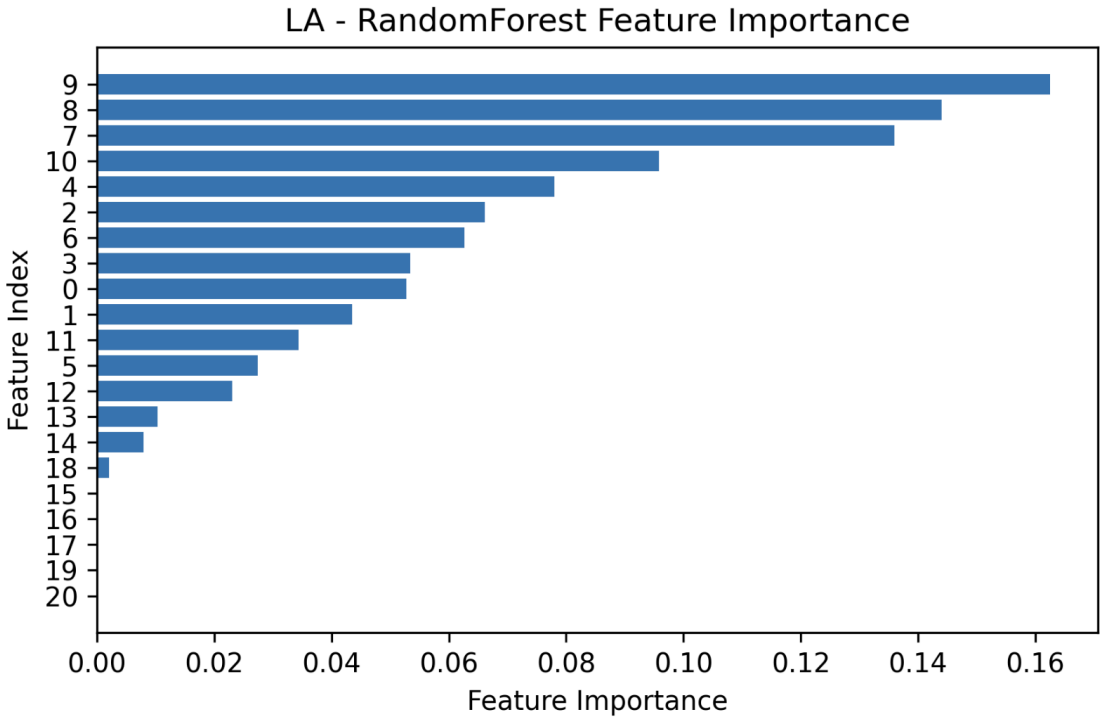
**8: longitude**

**9 - 12: the four one-hot encoded levels of room_type.**

**13 - 20: the eight one-hot encoded levels of neighbourhood_group.**

This mapping enables meaningful interpretation of the feature importance outputs. For example, if Feature Index 6 appears as a top feature, it directly indicates that distance to city center is highly predictive of the Airbnb price. Similarly, strong importance within Index 13 - 20 suggests that neighborhood groups play a substantial role in price variation.

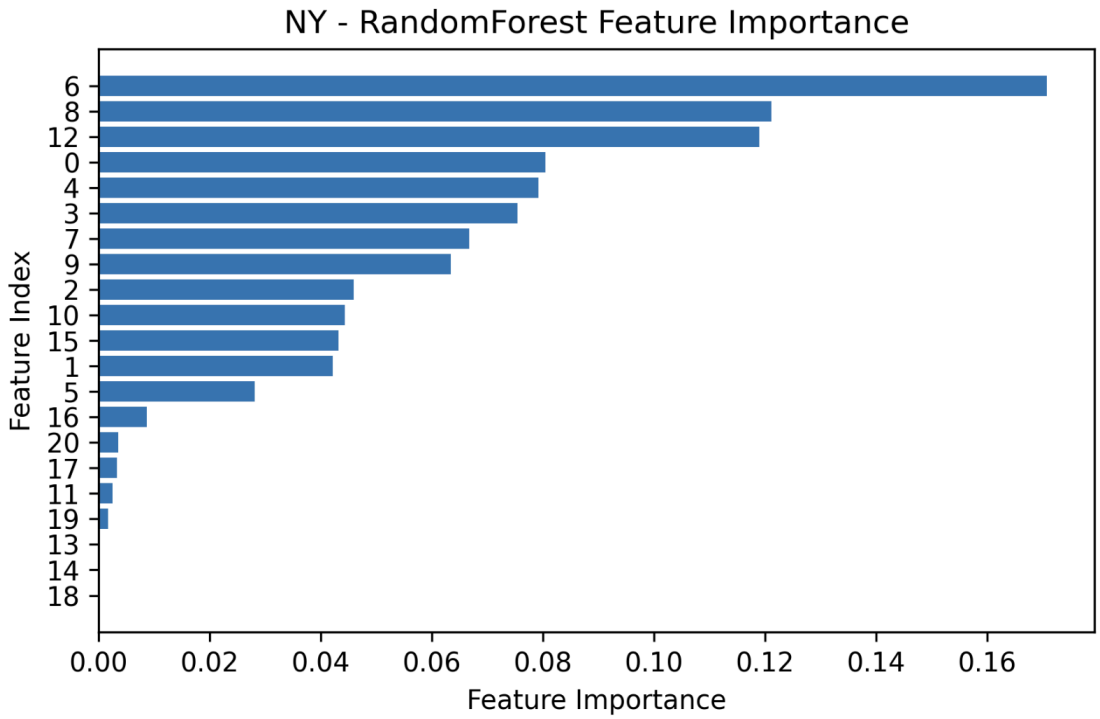
### 3.2.1 Random Forest Model




**Top 5 Most Influential Features Driving Price in LA (RandomForest Model):**

| Rank | Feature Index | Interpretation | Explanation |
|---|---|---|---|
| 1 | 9 | room_type_ohe_0 | Room type strongly affects pricing; entire homes typically charge much higher rates. |
| 2 | 8 | longitude | Prices rise toward western/coastal areas of LA, making longitude highly predictive. |
| 3 | 7 | latitude | Northern neighborhoods such as Hollywood Hills have higher prices, so latitude matters. |
| 4 | 10 | room_type_ohe_1 | Differences between room types create clear price gaps, giving this OHE feature strong weight. |

| | | | |
|---|---|---|---|
| 5 | 4 | calculated_host_listings_count | Professional hosts often manage higher-priced listings, making this a useful predictor. |

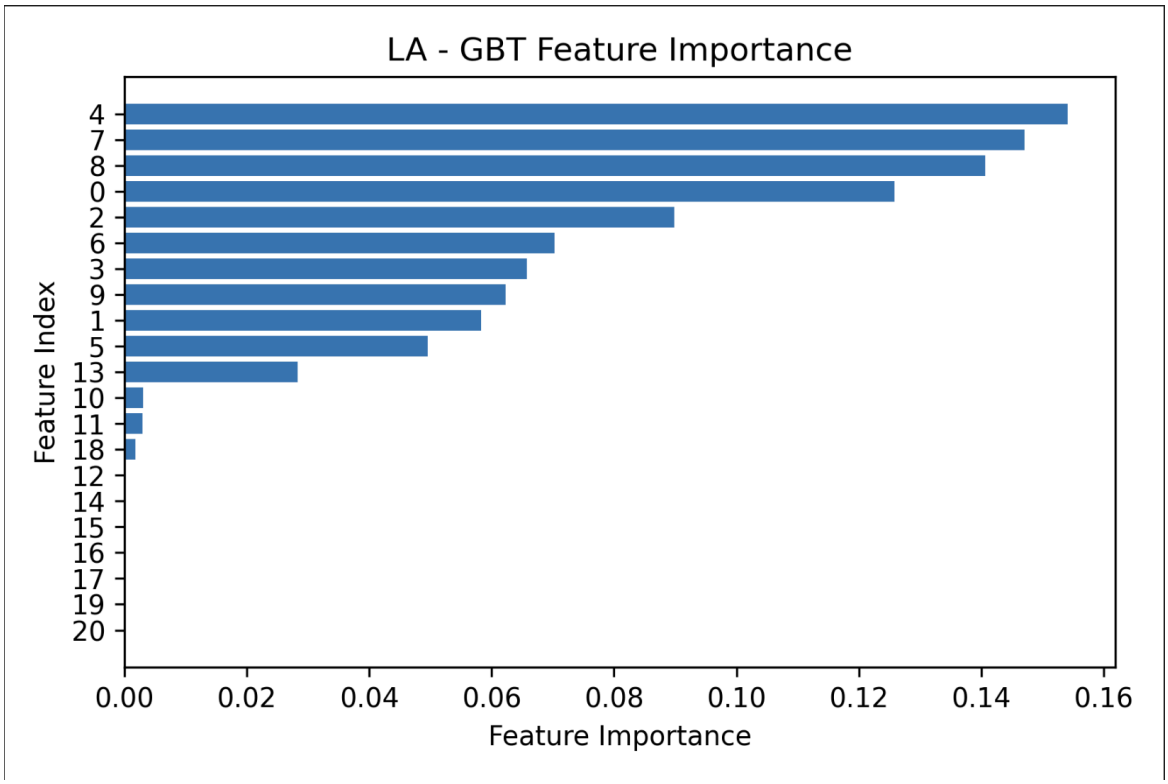



**Top 5 Most Influential Features Driving Price in NY (RandomForest Model):**

| Rank | Feature Index | Interpretation | Explanation |
|---|---|---|---|
| 1 | 6 | Distance to city center | Listings farther from Manhattan core (Midtown/Downtown) tend to have lower prices; proximity increases value. |
| 2 | 8 | Longitude | Captures east–west location within NYC; properties nearer Manhattan and the waterfront generally command higher prices. |
| 3 | 12 | room_type_ohe_3 | Represents a distinct |

| | | | |
|---|---|---|---|
| | | | room-type category that significantly affects price tiers. |
| 4 | 0 | Minimum nights requirement | Higher minimum-night stays often correspond to more premium or long-stay-oriented listings. |
| 5 | 4 | Host listing count | Hosts with many listings typically operate more professional or high-priced units. |

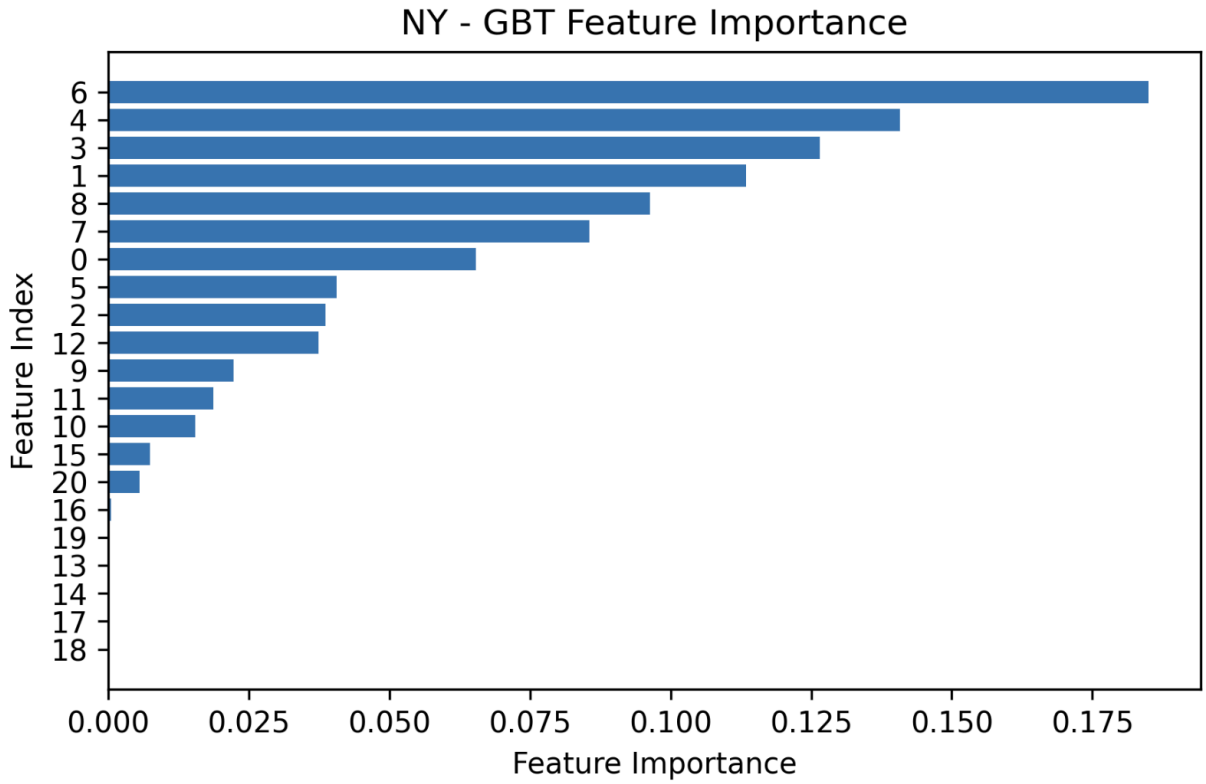
### 3.2.2 Gradient Boosted Trees Model




**Top 5 Most Influential Features Driving Price in LA (GBT Model):**

| Rank | Feature Index | Interpretation | Explanation |
|---|---|---|---|
| 1 | 4 | Host's total listing count | Professional hosts often manage higher-quality or |

| | | | |
|---|---|---|---|
| | | | more premium listings, pushing prices up. |
| 2 | 7 | Latitude | Captures north–south location in LA; areas closer to central city or desirable neighborhoods typically have higher prices. |
| 3 | 8 | Longitude | Reflects east–west variation; properties nearer the coast or high-demand districts tend to be more expensive. |
| 4 | 0 | Minimum nights requirement | Listings requiring longer minimum stays often target higher-end or long-term guests, raising overall price. |
| 5 | 2 | Reviews per month | Higher booking activity signals popularity and demand, allowing hosts to charge more. |




**Top 5 Most Influential Features Driving Price in NY (GBT Model):**

| Rank | Feature Index | Interpretation | Explanation |
|---|---|---|---|
| 1 | 6 | Distance to city center | Listings closer to Manhattan's core command significantly higher prices due to demand and convenience. |
| 2 | 4 | Host’s total listing count | Professional hosts often run high-quality or commercial-style listings, which tend to be priced higher. |
| 3 | 3 | Availability (days per year) | Lower availability often signals higher demand or occupancy, allowing hosts to charge more. |
| 4 | 1 | Number of reviews | More reviews |

| | | | |
|---|---|---|---|
| | | | indicate popularity and consistent booking activity, enabling higher pricing. |
| 5 | 8 | Longitude | Captures east–west location differences; properties closer to prime Manhattan areas or transit hubs tend to be more expensive. |

### 3.2.3 NY vs. LA: Cross-City Comparison of Price Drivers

Using both Random Forest and GBT models, we observe that NY and LA share some broad pricing patterns (location matters, review activity matters), but they also differ in what specifically drives price variation. The most influential features in each city highlight structural differences in how the Airbnb markets operate.

- **Location Sensitivity Is Strong in Both Cities, but Stronger in NY**
  Both NY and LA show clear price dependence on location, but the effect is noticeably stronger in New York. Prices in NY change sharply with even small shifts away from the city center, reflecting Manhattan's dominant role as the economic and tourism hub. In contrast, LA's pricing is influenced by location as well, but the effect is more dispersed across multiple sub-centers such as Hollywood, Santa Monica, and Downtown.
- **Review-Based Popularity Matters More in NY**
  New York listings rely more heavily on review activity to justify higher prices. Features related to number_of_reviews and reviews_per_month consistently appear among the most influential predictors in NY models, suggesting that guests value social proof in a market where building types and neighborhood quality vary widely. This effect is present but weaker in LA.
- **Host Professionalization Has More Impact in NY**
  The influence of calculated_host_listings_count is stronger in NY, indicating that hosts managing multiple units tend to operate more professionally and can command higher prices. This trend is likely amplified by NY's stricter regulations, which reduce casual hosting and highlight the role of commercial operators. LA also shows this effect, but to a smaller degree.

- **Spatial Coordinates Matter More in LA**
  Latitude and longitude emerge as highly important features in LA, reflecting the city's geographically spread-out structure. Price variation aligns with movement toward high-demand coastal and entertainment districts. In NY, the simpler vertical grid reduces the incremental value of raw coordinates, making them less informative for predicting price.
- **Availability (availability_365) Reflects Different Market Dynamics**
  Availability_365 contributes to pricing in both cities but captures different underlying patterns. In NY, low availability is often associated with consistently high demand, which supports higher prices. In LA, availability fluctuates more with seasonality and event-driven tourism, making its pricing influence less concentrated but still meaningful.

  Our project shows that Airbnb pricing in LA and NY is driven by different market dynamics, even when using the same modeling pipeline. Random Forest delivered the best predictive performance in both cities. Location is the strongest overall factor, but it influences each city differently: New York pricing is tightly centered around Manhattan, while Los Angeles shows a broader geographic pattern tied to latitude and longitude. NY relies more on demand signals such as reviews, whereas LA pricing varies more with room type and spatial distribution. Overall, the results highlight how urban structure, demand patterns, and host behavior shape price variation across cities and demonstrate the value of scalable PySpark workflows for producing interpretable insights.