# Final Project Report

Aviv Rabi (207667635), Ron Bartal (314724600), Naomi Chauvart (337917843)

## Introduction

Pricing is a central business problem in short-term rental platforms. Hosts aim to maximize revenue while maintaining competitive prices that attract bookings, and platforms benefit directly from improved pricing accuracy through higher conversion rates and commission-based revenue. At scale, however, developing and maintaining a dedicated pricing model for every city is costly, raising a strategic question: when does investing in a city-specific pricing model meaningfully outperform a single global solution? This project investigates that trade-off by comparing global and city-specific pricing models and identifying the conditions under which local specialization becomes worthwhile, while prioritizing qualitative comparison over quantitative optimization. We focus on Paris as a representative large and heterogeneous market, evaluating whether a locally trained model can outperform a global baseline and how this advantage depends on the amount of available city-level data.

To answer this question, we compare three modeling approaches: a global model, a local model and hybrid variants. This comparison supports a tiered deployment strategy in which a global model serves as a default baseline for broad coverage, while specialized local models are deployed selectively in high-volume markets where the accuracy gains justify the additional development and maintenance effort.

## Data Collection and Integration

We integrate two complementary data sources: a given large-scale global dataset from bright data covering listings from multiple cities worldwide, used to train a global baseline model and a city-specific Paris dataset collected via Inside Airbnb, which serves as the authoritative source for local training and evaluation, and a Although both datasets describe listings at the property level, they originate from different pipelines, schemas, and scrape dates, requiring explicit Standardization.

### Data Integration and Consistency Validation

A key assumption is that the listing identifier refers to the same real-world property across sources. We validated this by matching all Paris listings extracted from the global dataset against the Paris dataset using the identifier and spatial checks. Out of 14,387 Paris listings in the global data, 13,371 (92.9%) matched exactly by identifier. Among these, 99.7% were also spatially consistent within ~10 meters sensitivity factor, indicating high cross-source consistency and rare mismatches.

**Price Standardization and Quality Assessment**

Inside airbnb data reported prices in local currencies (EUR) and were converted to USD using the median exchange rate in the local data scraping month (June 2025). Joint analysis of matched Paris listings showed that 74.9% of Paris-source prices were valid, compared to 55.7% in the global source. For listings with valid prices in both datasets, Alongside high correlation (0.84), global prices were systematically higher by approximately 35 USD on average. To reduce sensitivity to such shifts and to heavy tails in the price distribution, all models operate on log-transformed prices.

**Leakage-Aware Dataset Construction**

Ground-truth prices are assumed to be most reliable in the Paris dataset. We therefore constructed a fixed hold-out test set from Paris only, consisting of 20% of valid listings properties). The remaining Paris listings form a local training pool used for progressive subsampling. To prevent leakage, all Paris test identifiers were removed from the global dataset via an anti-join before preprocessing, and zero overlap was verified. This leakage check is enforced at the start of every pipeline stage.

**Calibration Proof of Concept**

The observed price gap motivated a lightweight calibration proof of concept. We trained a simple Huber regressor to model residual corrections between global predictions and local ground-truth prices. This experiment served solely as a feasibility demonstration and was not used in the final evaluation pipeline. In the main experiments, calibration models are trained only on designated training splits, as detailed in the Methodology section.

**Final Artifacts and Reproducibility**

The data pipeline produces three Parquet artifacts: a listings fixed Paris test set, a listings Paris local training set, and a global training dataset containing approximately 1.33 million, 11 thousands and 40 thousands listings respectively. Persisting these artifacts ensures reproducibility and reduces computational overhead in downstream modeling. Additionally, two Paris-specific enrichment datasets,locations of major monuments and metro stations,were collected to derive distance-based geographic features for the local model.

* Plots examples are attached in the appendix.

**<u>Data Analysis</u>**

We began by cleaning and preparing the data for analysis. The preprocessing pipeline included outlier filtering from the train sets, such as quantile-based filtering for log-price while ensuring consistent ranges between global and local train datasets. Missing value imputation employed domain-aware strategies, for example, imputing missing categorical ratings (accuracy,

cleanliness) with overall ratings rather than means to preserve listing quality indicators. To limit feature expansion while retaining informative attributes, we applied Zipf's Power Law for dimensionality reduction. For instance, we selected the top 30 countries plus an "other" category to capture the most relevant geographic diversity without excessive feature growth.
For the test dataset, we applied only transformations that preserved listing characteristics, such as dimensionality reduction, while avoiding filters that would remove natural variation, such as filtering by number of guests. This ensures the model is evaluated on realistic data that reflects the full range of listing characteristics it would encounter in deployment.

Feature selection focused on correlation and contribution to the target variable (log-price). We employed distribution plots, trend lines, and correlation heatmaps to identify relevant features. The local dataset contains all global attributes plus additional Paris-specific features engineered to capture city-specific pricing factors. These include neighborhood identifiers, distance from major tourist monuments (such as the Eiffel Tower), These geographic features were designed to capture location-based pricing dynamics unique to Paris, where proximity to landmarks can significantly influence rental rates. An example for an examined Paris specific feature which we decided to not include in the final models is the distance to the nearest metro station. This feature showed neglectable correlation, we assumed it's because Paris is so well networked in metro stations that the differences are minor.
* Plots examples are attached in the appendix.

**Methodology**

This project investigates the conditions under which a city-specific model outperforms a globally trained model. We hypothesized that the primary condition would be training data size, meaning that training a local model would only be worthwhile for sufficiently large markets. This hypothesis was evaluated using subsets of the local dataset with increasing size, as explained below. Evaluation was performed using Mean Absolute Error (MAE) of three distinct models on a local, held-out test set:

1. **Global Model (Baseline)**: A Gradient Boosted Trees (GBT) model trained on the full global dataset.
2. **Local Model**: A Random Forest model trained from scratch in each experiment iteration on increasing subsets of the local training data.
3. **Hybrid Model**: Uses the Global Model's predictions as a baseline and refines them using a Huber Regressor for residual learning. This projects global predictions onto the local pricing distribution (addressing differences noted in the Data Integration section). We tested two regressor versions: one utilizing only the global and local predictions, and

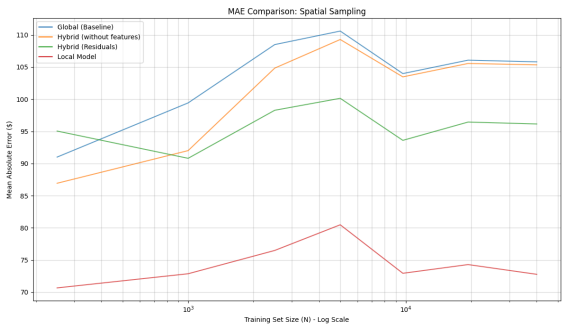
another incorporating local features as well. The motivation was to mimic a fine-tuning approach, aiming to leverage massive global data alongside specific local signals, which is theoretically advantageous in small-data scenarios.

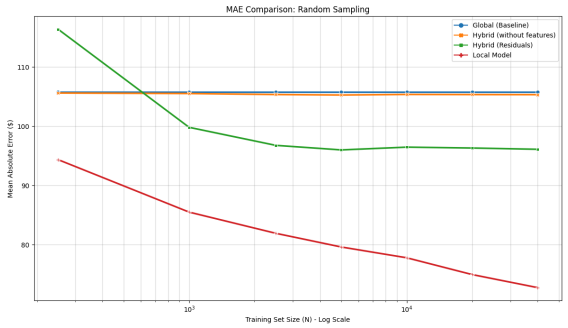
To simulate small markets, we sampled the local training subsets using two parallel methods:

- **Random Sampling**: Incrementally increasing the number of randomly selected listings, creating a geographically sparse dataset (fewer listings per geographic polygon).
- **Spatial Sampling**: Paris has 20 neighborhoods (arrondissements). We sampled an increasing number of neighborhoods to form subsets of increasing size. These subsets were evaluated on a matching neighbourhood subset from the test dataset.

## Evaluation and Results

As mentioned in the previous section, the final evaluation comparing the three model architectures used MAE, which is more robust to a wide range of values. Our expectation of a clear performance hierarchy was limited to the large-data regime. For small training sets, we hypothesized that the hybrid model would outperform the local model, as the latter is more prone to overfitting when trained on limited data. Only once sufficient local data becomes available did we expect the local model to consistently achieve the lowest error.




In practice, the local model consistently achieved the lowest error across all data sizes. The global model remained stable and insensitive to the local training size, as expected. The hybrid models provided only marginal improvements over the global baseline and did not outperform the local model, even in the low-data regime. This suggests that, in our setting, the local signal was strong enough to outweigh the overfitting risk associated with smaller training sets. For spatial sampling, the performance hierarchy was preserved. The error on the smallest dataset was very low, likely due to minimal noise and better alignment with the crafted test subset. Error peaked at a train set size of 5,000 (collected from 4 different neighborhoods) , then began to decrease as the model learned and generalized better.

## Limitation and Reflection

One limitation lies in the temporal aspect of the datasets. In this project, we focused on predicting pricing without accounting for date. In reality, timing has a significant influence because different periods of the year can have substantial pricing differences, for example, Christmas and New Year's can be very expensive in Europe compared to periods without holidays. There was a three-month gap between data collection: the global data was scraped in September while the local data was scraped in June. This temporal difference could lead to price deviations that we would be unable to explain, as was explored in the data integration section. We attempted to address this using the hybrid model approach, which worked partially, but we believe that having the data scraped on the same day would have yielded better results.

A difficulty we encountered concerns the requirement to use the data sources specified in the project instructions. This constraint prevented us from collecting data at a synchronized time point. We would have preferred to create a global dataset comprising several large cities collected in the same month as the Paris dataset. This approach would have simplified the integration process and reduced error rates stemming from temporal inconsistencies between datasets.

Another limitation concerns the models we chose to use. Since our data is tabular, we used appropriate models for this format (as mentioned in the methodology section). However, best practice today involves using neural networks. With more time and compatible computing resources (GPU on the workers), we could have training neural networks expecting lower error rates.

## Conclusion

Based on our experimental results, the local model consistently achieved the best performance across all subsets in both sampling strategies. This indicates that for large, western, structured, data-rich markets, such as Paris, investing in a dedicated local model is preferred.

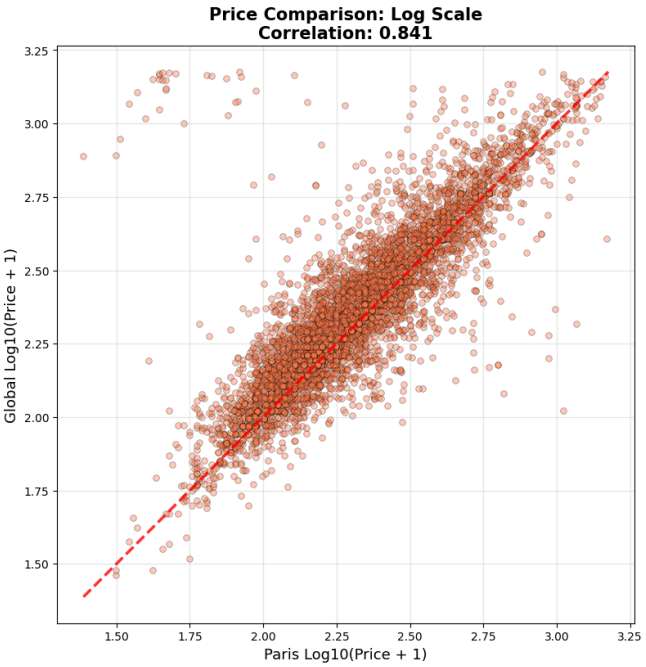
Our suggestion for smaller cities is based on the spatial experiment results. We assume that spatial sampling better represents smaller but Western-style cities, such as Tel Aviv. Under this sampling strategy, we observe that the simple hybrid model (the version without the features) yields a substantial improvement on the small datasets over both the global baseline and the more complex hybrid variant. This leads to a practical conclusion for such small-western-cities: when sufficient time and resources are available, training a dedicated city-specific model is preferable, however, when rapid results are required, collecting price data alone and applying calibration constitutes a strong and viable alternative.

## Appendix

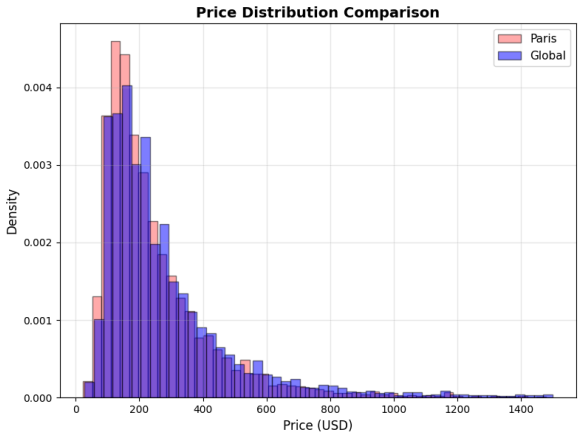
**Git Repo:** No video is required, there is a single notebook in the repo to run using “Run All” as explained in the repo’s ReadMe file.

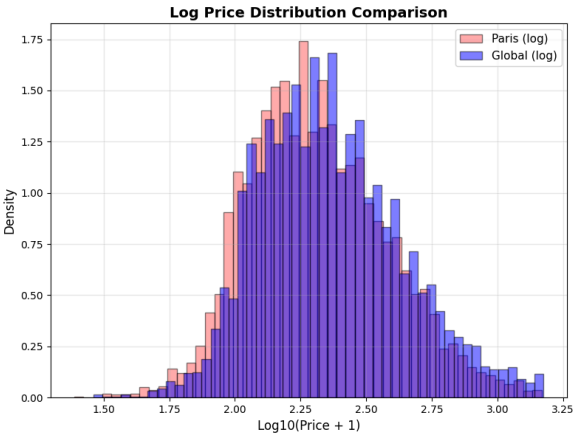
### Data Collection and Integration:

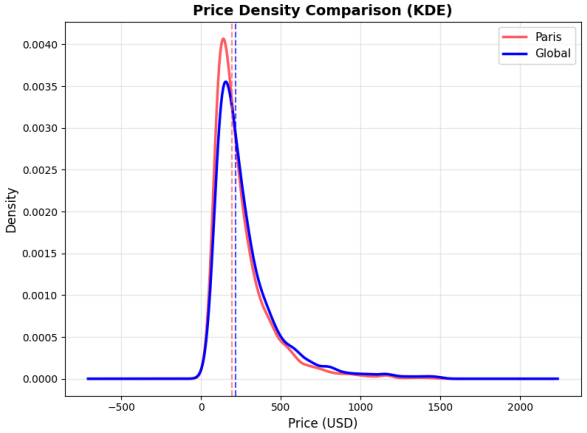
Price and Log-Price Distribution Comparisons

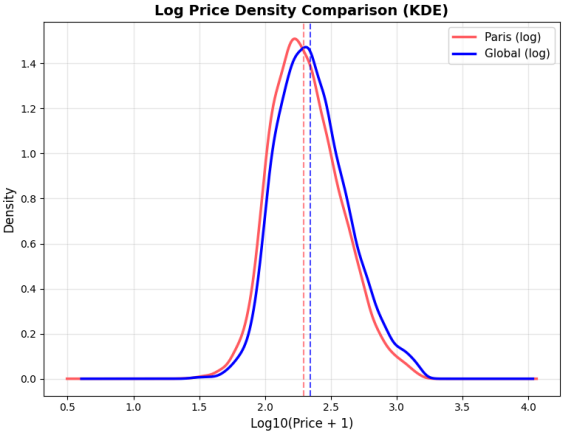




Log scale shows **stronger** correlation (difference: 0.039)

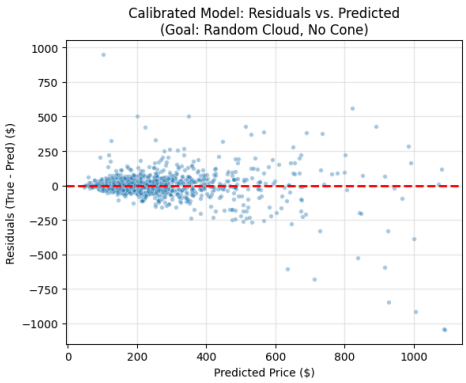


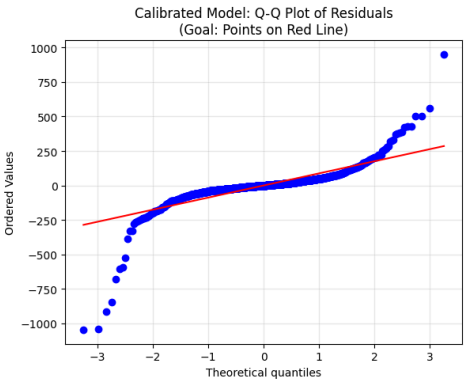



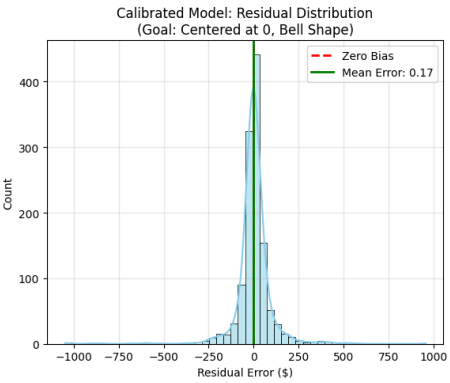



Residual Analysis and POC for the Hybrid Model



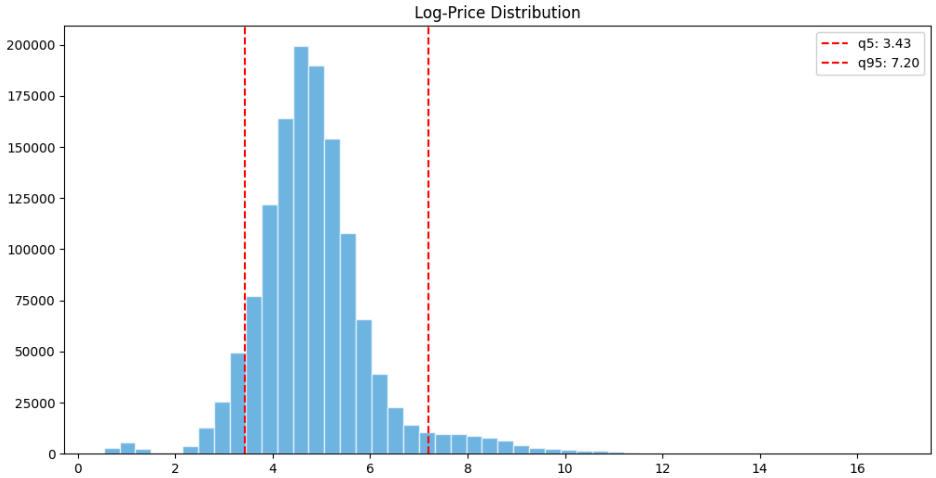

As explained in the report, we conduct a Huber regression on the valid set of prices in both global and local data and receive a corrected (or projected on the local space) distribution for the global model.

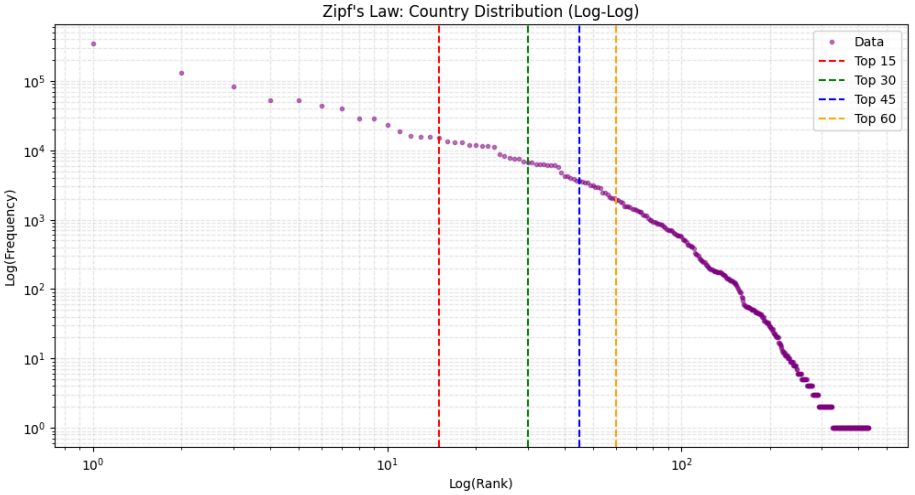
| | **RMSE** | **MAE** | **Bias** |
|---|---|---|---|
| **Baseline ( Global Data)** | $130.52 | $64.22 | $35.74 |
| **After Log-Calibration** | $107.02 | $53.61 | $-0.17 |
| **Improvement** | 18.0% | 16.5% | 99.5% |

**Additional Data Analysis Examples:**

All the plots are in the data analysis and feature extraction notebook in the repo.

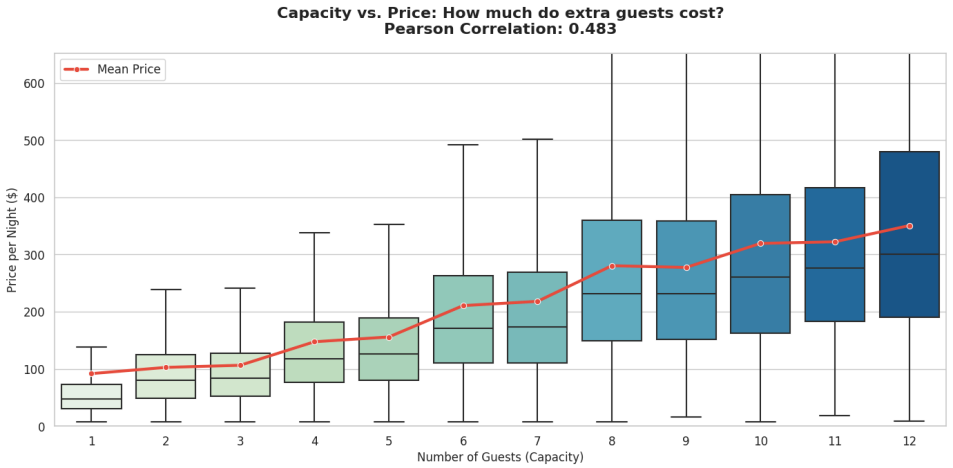
Data cleaning:

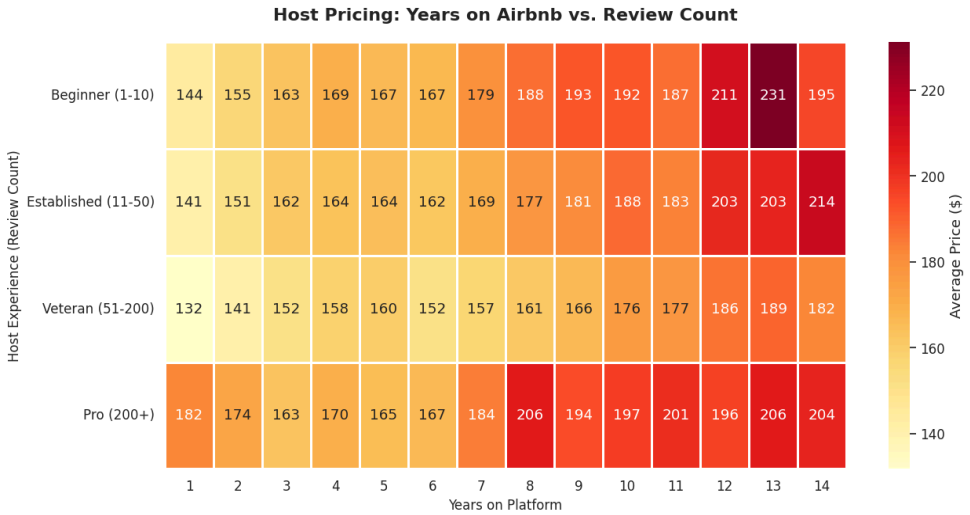




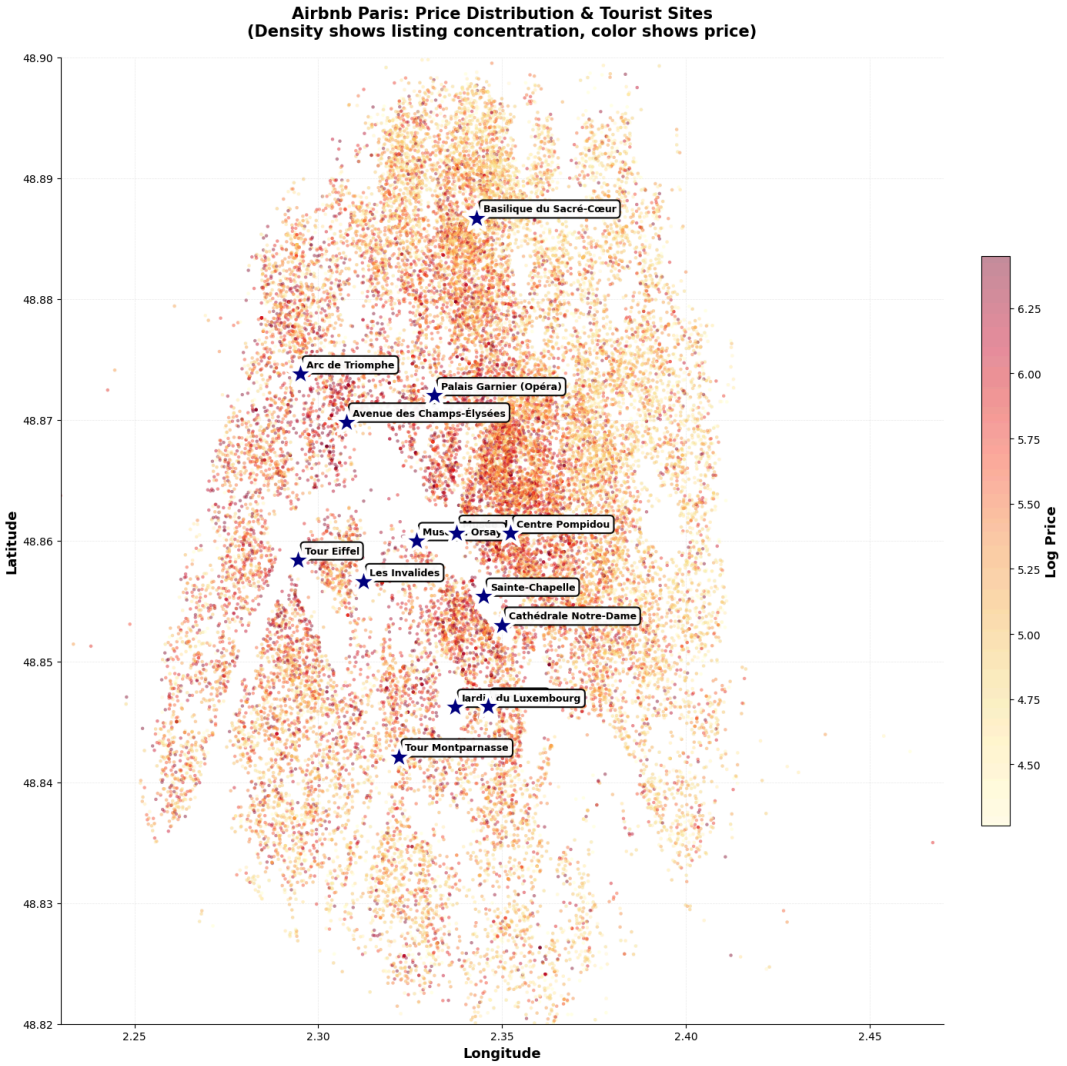

```
--- Coverage Analysis for Country ---
Top k      | Records Covered | Percentage
-------------------------------------------
15         | 909881          | 73.51%
30         | 1060689         | 85.70%
45         | 1138734         | 92.00%
60         | 1179600         | 95.31%
```

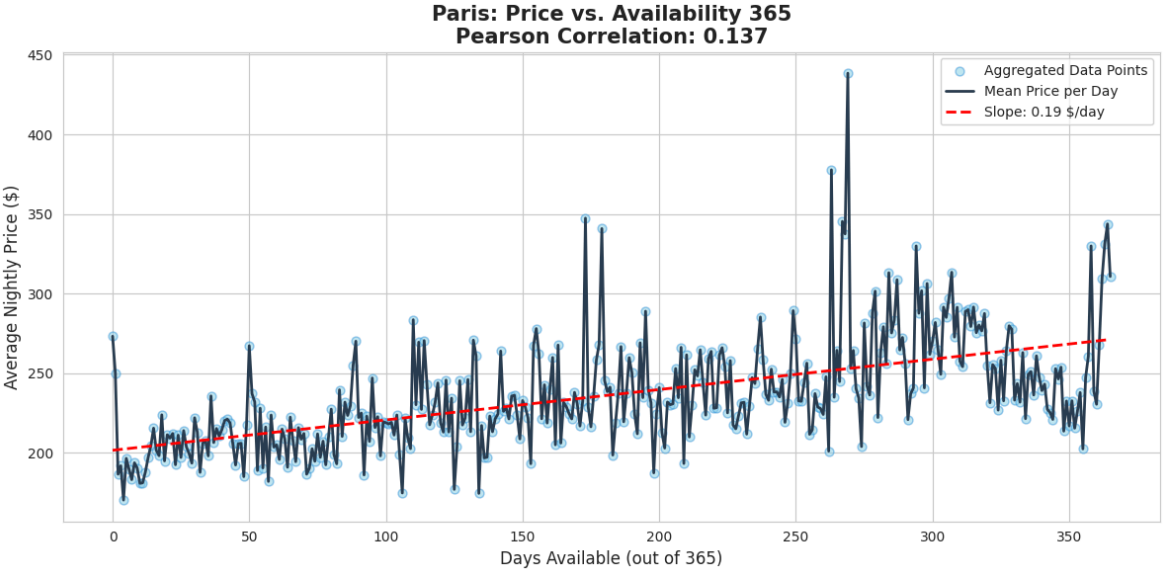
## Features selection:

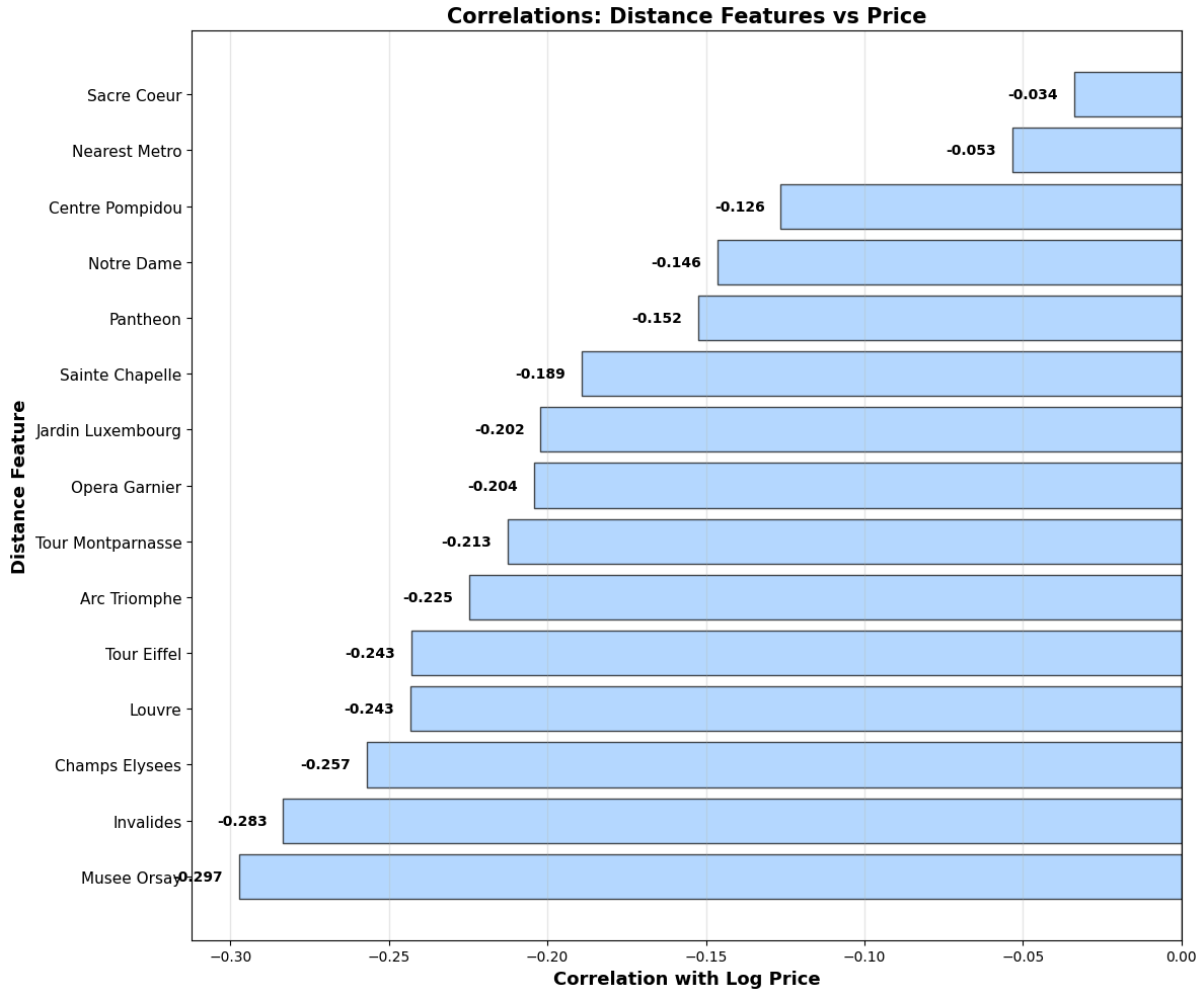


Paris-specific features:




Correlations: Distance Features vs Price
Distance Feature
Sacre Coeur
Nearest Metro
Centre Pompidou
Notre Dame
Pantheon
Sainte Chapelle
Jardin Luxembourg
Opera Garnier
Tour Montparnasse
Arc Triomphe
Tour Eiffel
Louvre
Champs Elysees
Invalides
Musee Orsay
-0.034
-0.053
-0.126
-0.146
-0.152
-0.189
-0.202
-0.204
-0.213
-0.225
-0.243
-0.243
-0.257
-0.283
-0.297
−0.30
−0.25
−0.20
−0.15
−0.10
−0.05
0.00
Correlation with Log Price