METCS777

Airbnb Price Prediction & City Dynamics Analysis: LA vs NY

METCS 777 – Final Project
Yibing Wang & Yiming Chen

Introduction

What Is This Project About?

- Predict nightly Airbnb prices using PySpark ML.
- Compare key price-driving factors between **LA** and **NYC**.

Why important

- LA & NYC = largest Airbnb markets in the US.
- Hosts and platforms need pricing insights.
- Markets differ greatly in urban structure, demand, tourism.

Motivation

- West Coast vs East Coast rental patterns differ.
- Hosts need guidance on how listing attributes influence price.
- Travelers need transparency on cost drivers.
- Good case study for big data processing + ML modeling.

Dataset Overview

Data Source:

- Inside Airbnb public listings dataset
- Cities analyzed: Los Angeles (LA) and New York City (NYC)
- File format: listings.csv.gz

Key data fields:

- **Location:** latitude, longitude, neighbourhood group
- **Listing:** room_type, minimum_nights, availability_365
- **Demand:** number_of_reviews, reviews_per_month
- **Host:** host_listings_count, license (LA only)
- **Target:** price → log_price

Data Cleaning & Processing

01

Remove Invalid Rows

- Drop listings with invalid / non-numeric IDs
- Remove rows with missing price
  Exclude corrupted records & incorrect coordinates

02

Fix Data Types

- Use try_cast to safely convert numeric fields *(price, nights, reviews, availability, host counts)*
- Convert last_review → date

03

Handle Missing Values

- reviews_per_month = 0
- reviews_ltm = 0
- Keep license as-is (LA only)

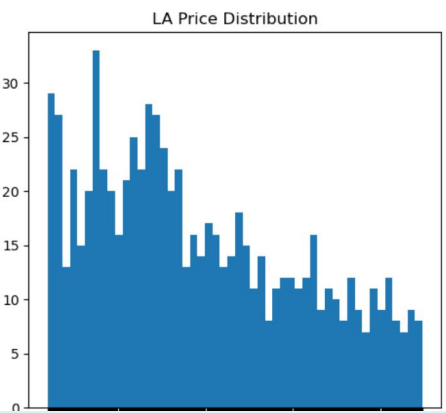
04

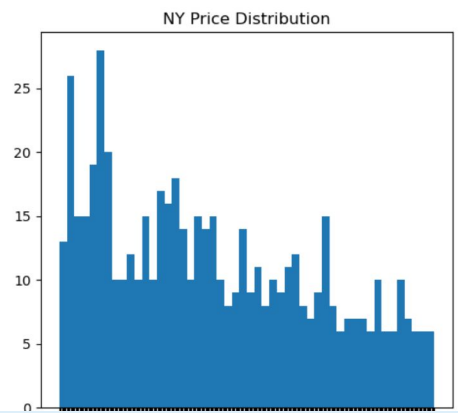
Price Outliers

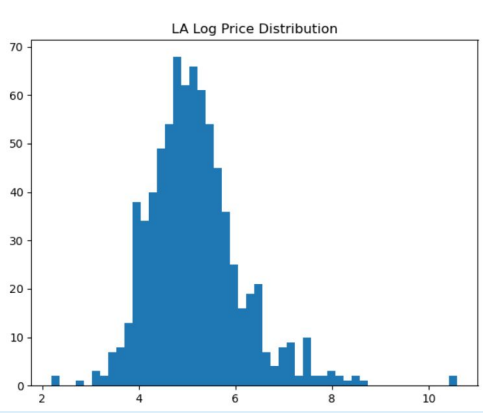
- Removed extreme prices
- Kept listings with price ≤ $2000

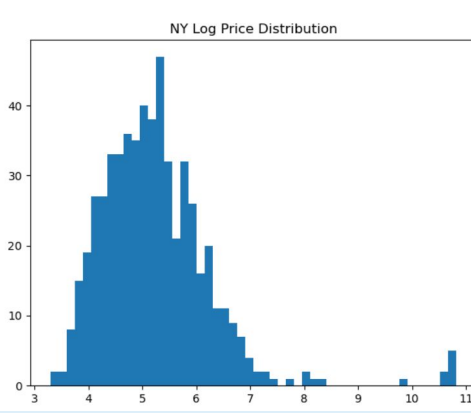
EDA Highlights

Price Distribution (LA vs NY)



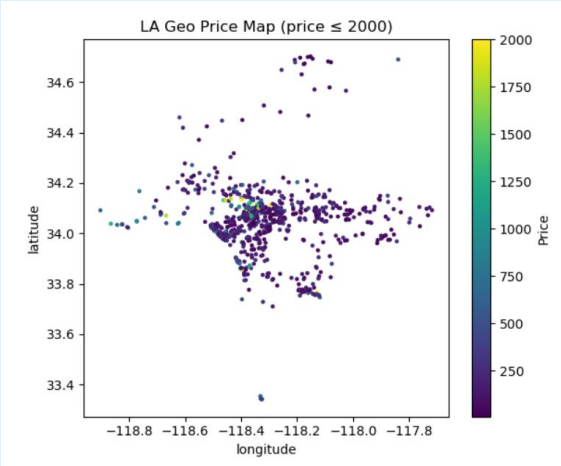


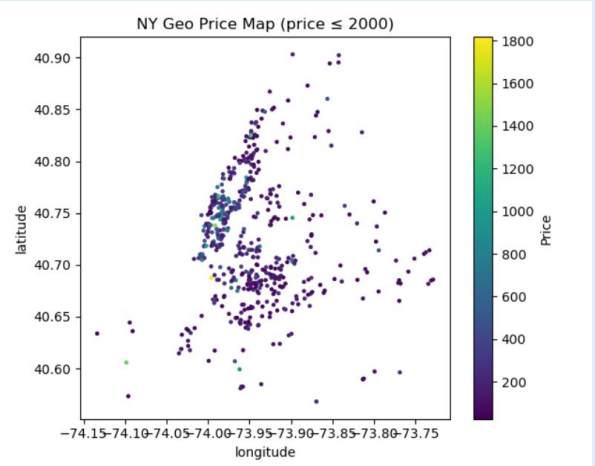




- Raw prices show a strong right-skew, with a small number of extremely high-priced listings.
- After log transformation, price distributions become much more symmetrical and close to normal.
- NY prices tend to be higher overall, with a more pronounced high-end tail than LA.
- These patterns justify applying a log transformation to stabilize variance and improve model performance.

Geographic Price Mapping





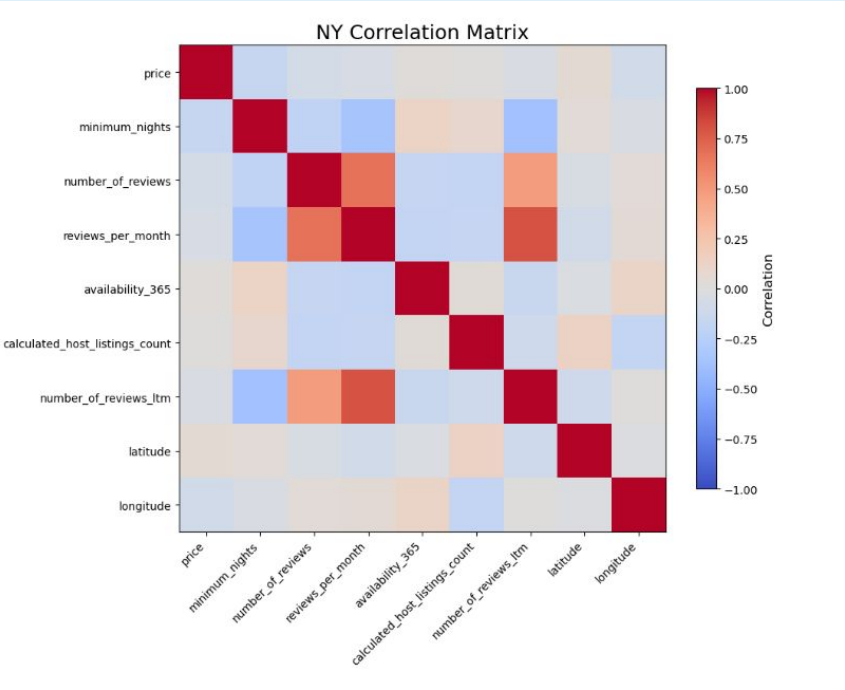
LA

- Higher-priced listings cluster around central LA and coastal areas (e.g., Santa Monica / Malibu direction).
- Prices spread more broadly due to LA's decentralized urban structure.

NY

- High-price listings are tightly concentrated around Manhattan, especially Midtown & Lower Manhattan.
- Much sharper spatial gradient than LA — prices drop quickly moving outward.

Correlation Insights (NY as example)



Strong positive correlation between review-based features

Pricing in NYC is influenced by non-linear and multi-factor interactions.

Raw features alone cannot fully capture location-driven and host-driven pricing patterns.

Feature engineering is essential — we need log-price, spatial features, and encoded categorical variables to model NYC pricing effectively.

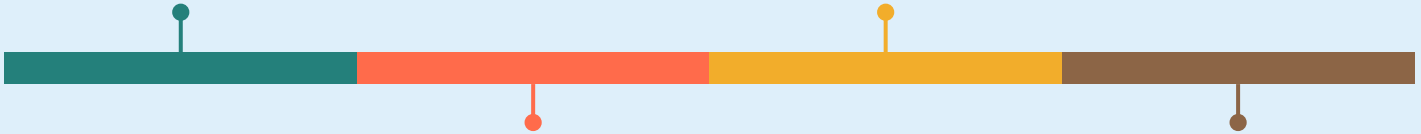
Feature Engineering

Numeric Feature Refinement

Cleaned and standardized continuous variables that influence pricing

Target Variable Transformation

- Applied log(price) to address right-skewed distribution.
- Stabilizes variance and improves model regression performance.



Categorical Feature Encoding

Converted qualitative attributes into model-friendly vectors

Final Feature Vector Assembly

All numeric + encoded categorical features were combined using VectorAssembler

Why These Features Matter

- Captures location, host behavior, demand, and property characteristics.
- Creates comparable patterns between LA vs NY for modeling and interpretation.
- Directly supports cross-city feature importance comparison.

Modeling

Modeling Approach

End-to-End PySpark ML Pipeline

- Constructed a full machine learning workflow in PySpark
- Data cleaning → feature engineering → VectorAssembler → ML models
- Ensures consistency and scalability for both LA and NY datasets

Train/Test Split

- 80% training / 20% testing
- Fixed random seed for reproducibility
- Same split for both cities → ensures fair comparison

Models Used

Linear Regression (Baseline)

- Evaluates whether a linear relationship can explain pricing

Random Forest Regressor

- Handles nonlinear interactions
- More robust to noise and outliers
- Provides interpretable feature importance

Gradient-Boosted Trees (GBT)

- Captures complex feature interactions
- Works well without extensive tuning

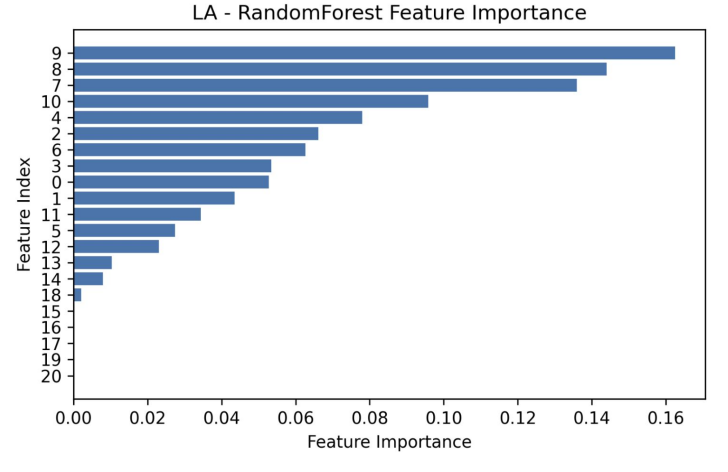
Model Performance — LA

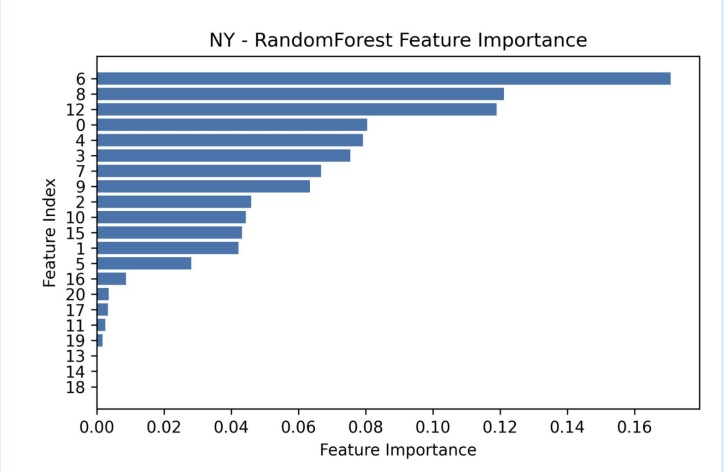
Model	RMSE	MAE	R^2	Interpretation
Linear Regression	0.723	0.545	0.378	Weak linear fit
Random Forest	0.629	0.468	0.529	Best overall performance
GBT	0.677	0.492	0.455	Moderate improvement vs LR

Model Performance — NY

Model	RMSE	MAE	R^2	Interpretation
Linear Regression	0.782	0.564	0.520	baseline
Random Forest	0.743	0.531	0.566	Best overall performance
GBT	0.854	0.551	0.429	Overfits + underperforms

Feature Importance — Random Forest





- Room type (largest effect)
- Latitude & longitude (coastal & high-demand areas priced higher)
- Host listing count (multi-unit hosts → higher price)

- Distance to city center (strongest driver)
- Longitude (Manhattan proximity)
- Room type differences
- Minimum nights requirement
- Host listing count & review activity

Cross-City Takeaways

- NY pricing is highly centralized; LA pricing is spatially dispersed.
- Review & demand signals matter more in NY.
- Room type is consistently influential across both markets.

Insights & Interpretation

New York

- Price is location-driven, with clear premiums near Manhattan.
- Distance to center, availability, and review volume are major predictors.
- Reflects a high-competition, demand-sensitive market

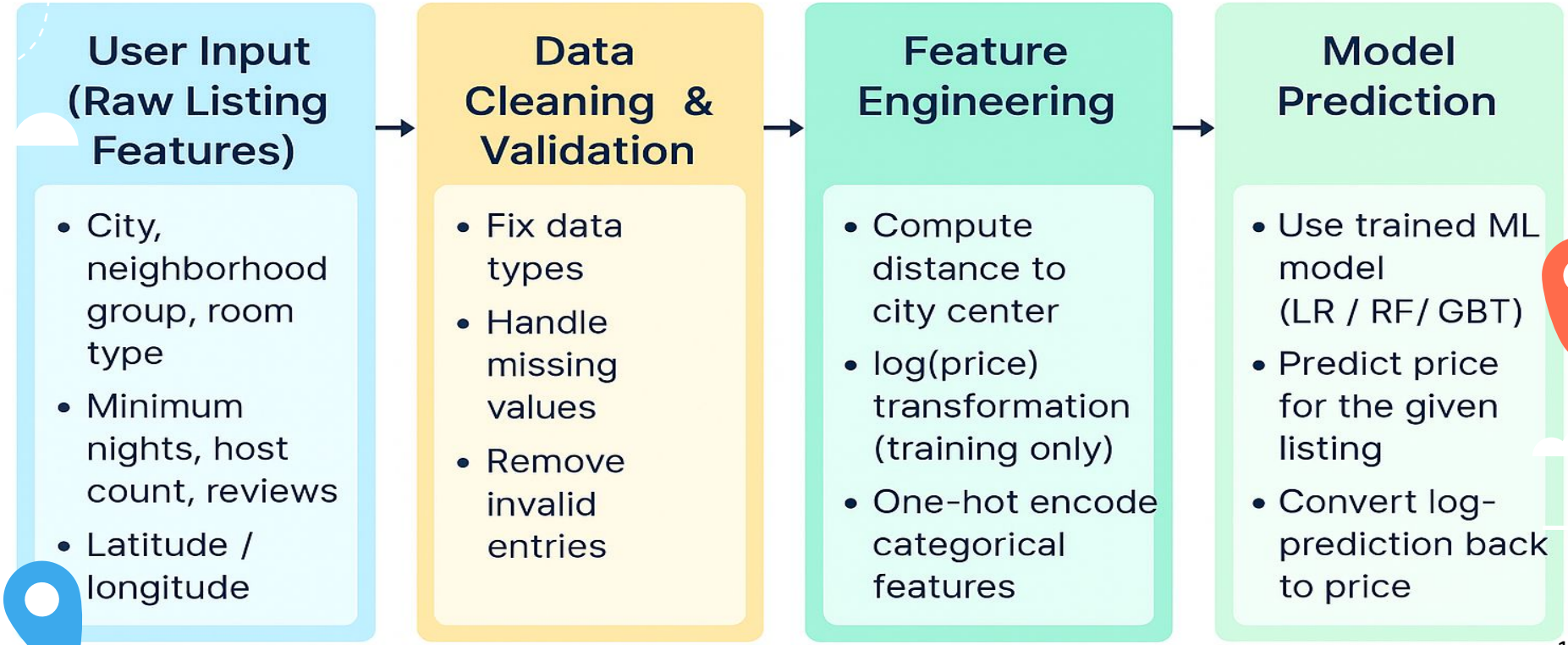
Los Angeles

- Pricing is more spread out across neighborhoods.
- Lat/long (neighborhood differences) and room type are stronger drivers.
- Market responds more to property characteristics than demand signals.

Cross-City Takeaways

- NY: location + demand signals drive price.
- LA: location + property attributes drive price.
- Feature importance aligns with EDA findings.

Demo Workflow: Predicting Airbnb Price



Conclusions

- Airbnb pricing is shaped by different forces in LA and NY
- Random Forest models performed best

NY

highly centralized
distance to Manhattan and demand indicators (reviews, availability) dominate pricing.

LA

spatially distributed
with latitude, longitude, and room type playing central roles

These results highlight how urban structure and market patterns shape Airbnb prices differently across cities, and demonstrate the value of scalable PySpark pipelines for building interpretable, city-level pricing insights.

Reference

- “How is Airbnb really being used in and affecting the neighbourhoods of your city?” *Insideairbnb.com*, insideairbnb.com/.
- “Airbnb Open Data.” *Www.kaggle.com*, www.kaggle.com/datasets/arianazmoudeh/airbnbopendata.
- Pouya Rezazadeh Kalehbasti, et al. *Airbnb Price Prediction Using Machine Learning and Sentiment Analysis*. 29 July 2019, www.researchgate.net/publication/334783073_Airbnb_Price_Prediction_Using_Machine_Learning_and_Sentiment_Analysis.
- Camatti, Nicola, et al. “Predicting Airbnb Pricing: A Comparative Analysis of Artificial Intelligence and Traditional Approaches.” *Computational Management Science*, vol. 21, no. 1, 6 May 2024, https://doi.org/10.1007/s10287-024-00511-4.