Airbnb listings price prediction

Big Data 2025 project presentation

OUR TEAM

Erokhin Evgenii

Davydov Danil

Poliakov Egor

Mitrokhin Ilia

CONTENTS

Introduction and goals of project

01

Hosts on Airbnb often struggle to determine the optimal nightly rental price for their properties. Setting a price too high may lead to low occupancy, while pricing too low could result in lost revenue. Current pricing decisions are often based on intuition or manual comparisons with similar listings, which can be time-consuming and inaccurate.

To address this, we aim to develop a regression model that analyzes the provided data by the host—including location, property features, amenities, reviews—to recommend competitive and profitable rental prices. This solution will help hosts optimize their pricing strategy, improve occupancy rates, maximize earnings, and simplify the process of price setting.

Dataset characteristic

02

```
dtypes: float64(33), int64(1), object(55)
memory usage: 336.1+ MB
```

~916,000 the total dataset our part is ~458, 000 entries (our part are even rows according to arrangement)

Missing Values Analysis:

	MissingCount	MissingPercentage
Has Availability	485647	98.119623
Square Feet	482745	97.533306
License	480358	97.051039
Host Acceptance Rate	452696	91.462237
Monthly Price	398863	80.585873
...	...	...
Host URL	0	0.000000
Host ID	0	0.000000
Experiences Offered	0	0.000000
Scrape ID	0	0.000000
ID	0	0.000000

Is our dataset appropriate for achieve our business goal?

After analyzing the characteristics of our dataset, we understand that it is comprehensive and suitable for obtaining the desired results in our task.

Price was converted from local currency to $

Link to detailed description of all features

Data analyze

03

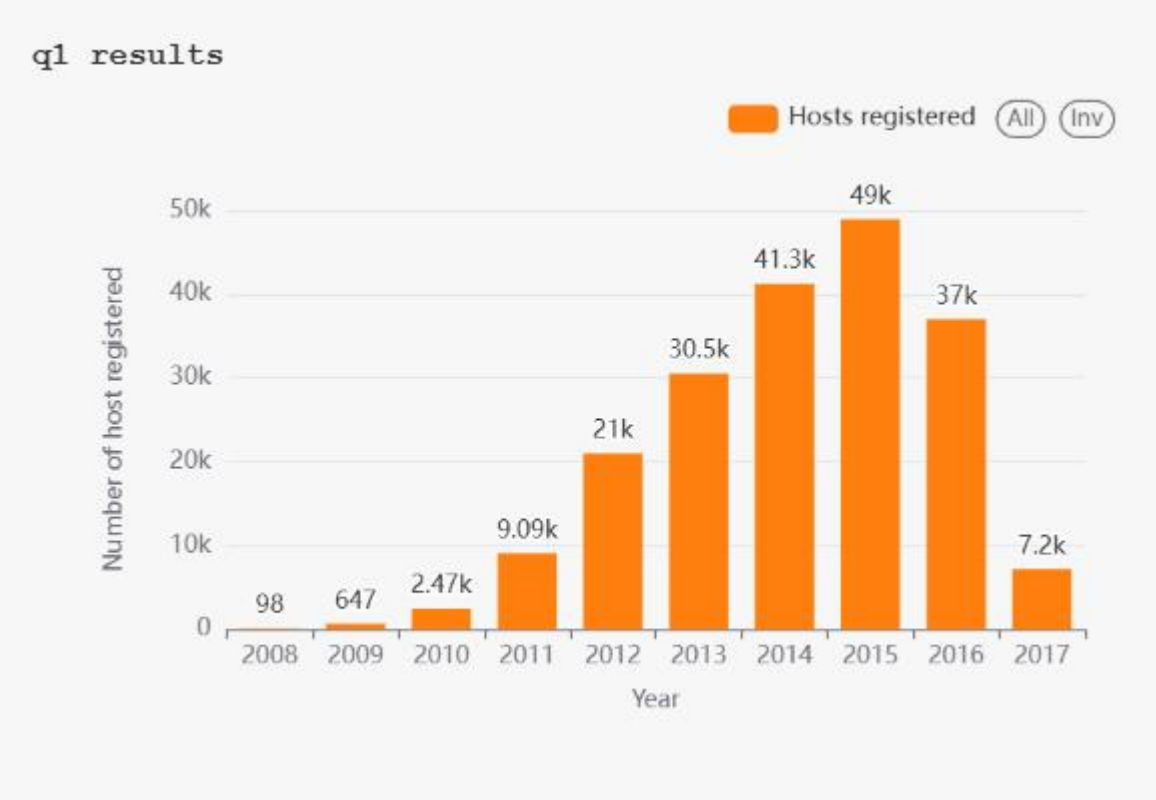
We performed explored data analyze to find key insights and trends from our data:

- Distribution of hosts registered on platform by year
- Do hosts have provided a description in their profiles
- Distribution of average price of listings across different countries
- Distribution of listings across different cities
- Relationship between listing ratings and average price
- Relationship between property types and their average prices

Data analyze results

04

Distribution of hosts registered on platform by year

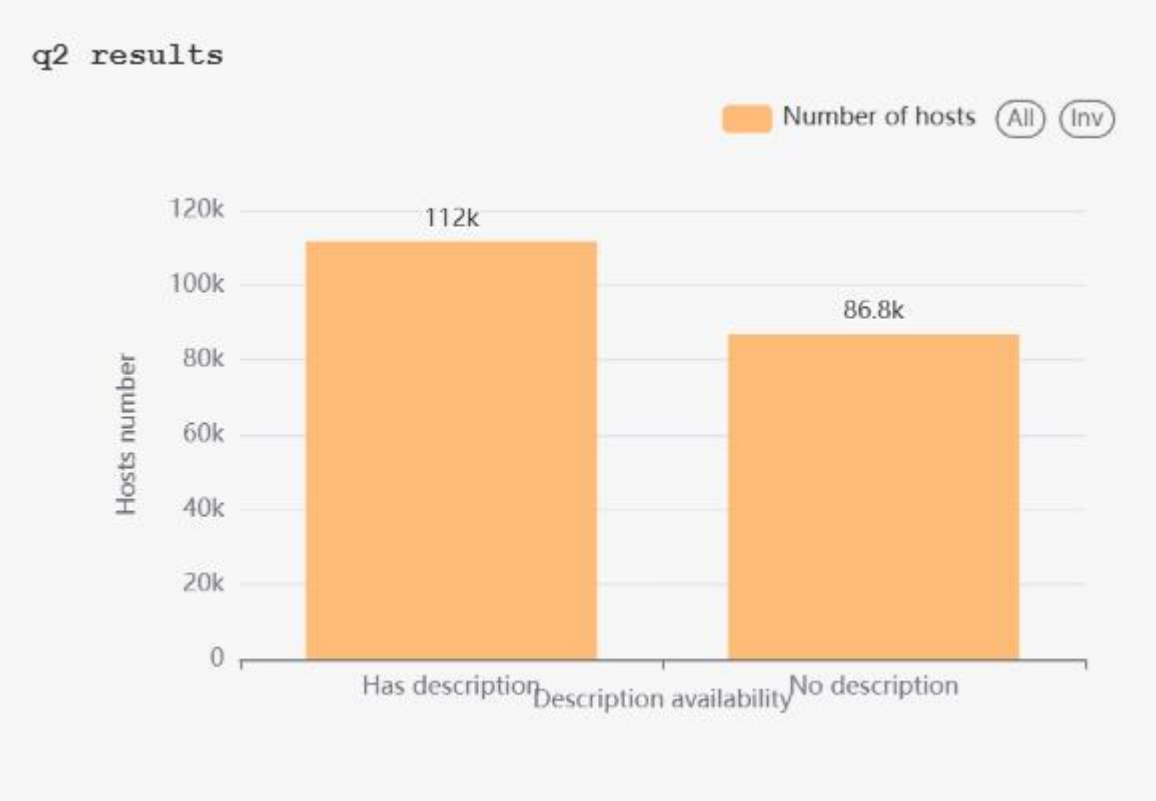


Correlation with our business goal: Understanding host growth trends helps identify market saturation levels. If many new hosts joined recently, pricing may need to be more competitive. This informs how aggressively to price relative to market dynamics.

Data analyze results

04

Do hosts have provided a description in their profiles

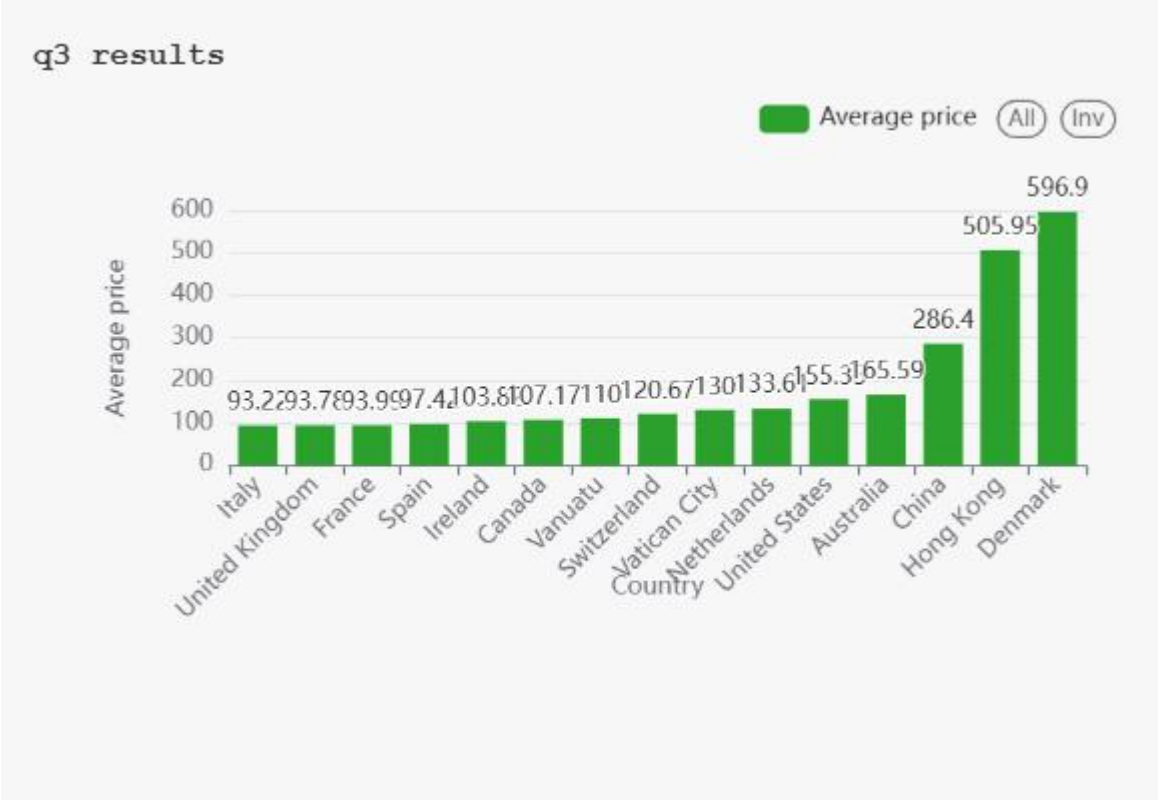


Correlation with our business goal: Profile completeness (like descriptions) often correlates with listing quality and professionalism, which can justify higher prices. This could be a feature in your pricing model to account for perceived host quality.

Data analyze results

04

Distribution of average price of listings across different countries

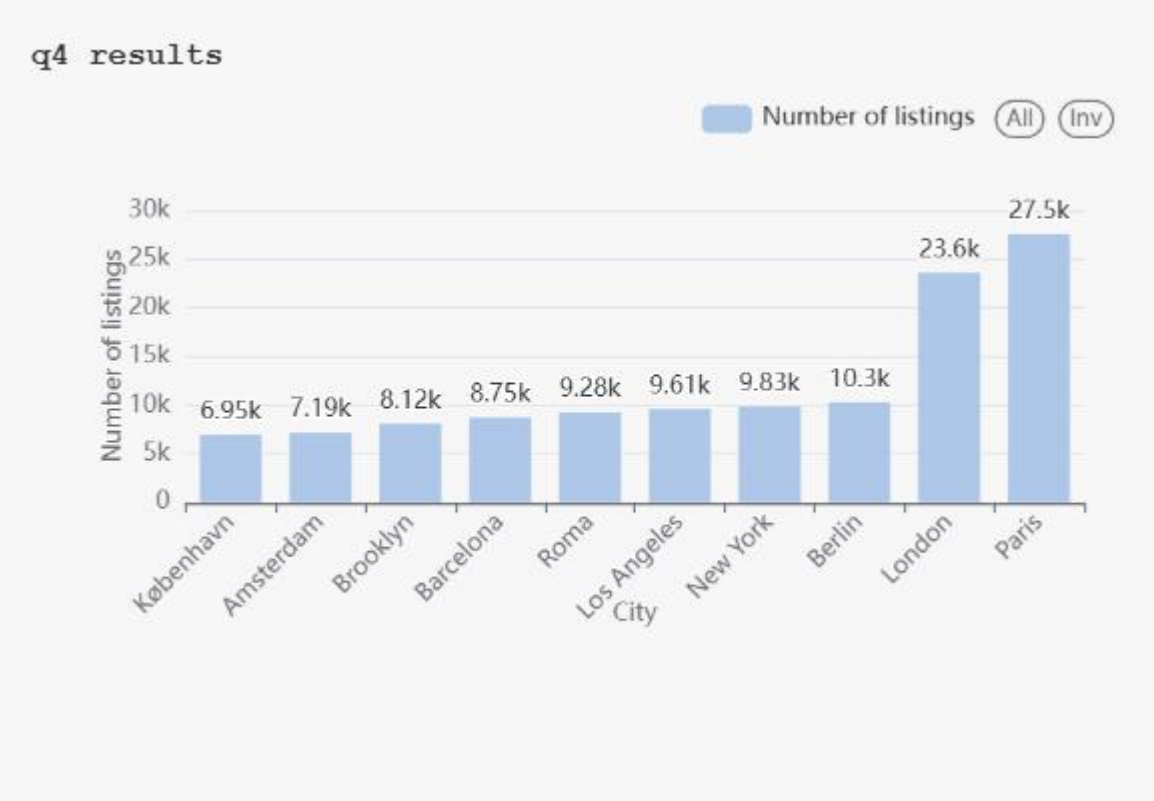


Correlation with our business goal: Reveals baseline price expectations by location, which is fundamental for location-based pricing. This helps hosts understand how their country's market context affects pricing strategies.

Data analyze results

04

Distribution of listings across different cities



Correlation with our business goal: Identifies supply concentration - cities with more listings may require more competitive pricing. Helps model account for local market density when recommending prices

Data analyze results

04

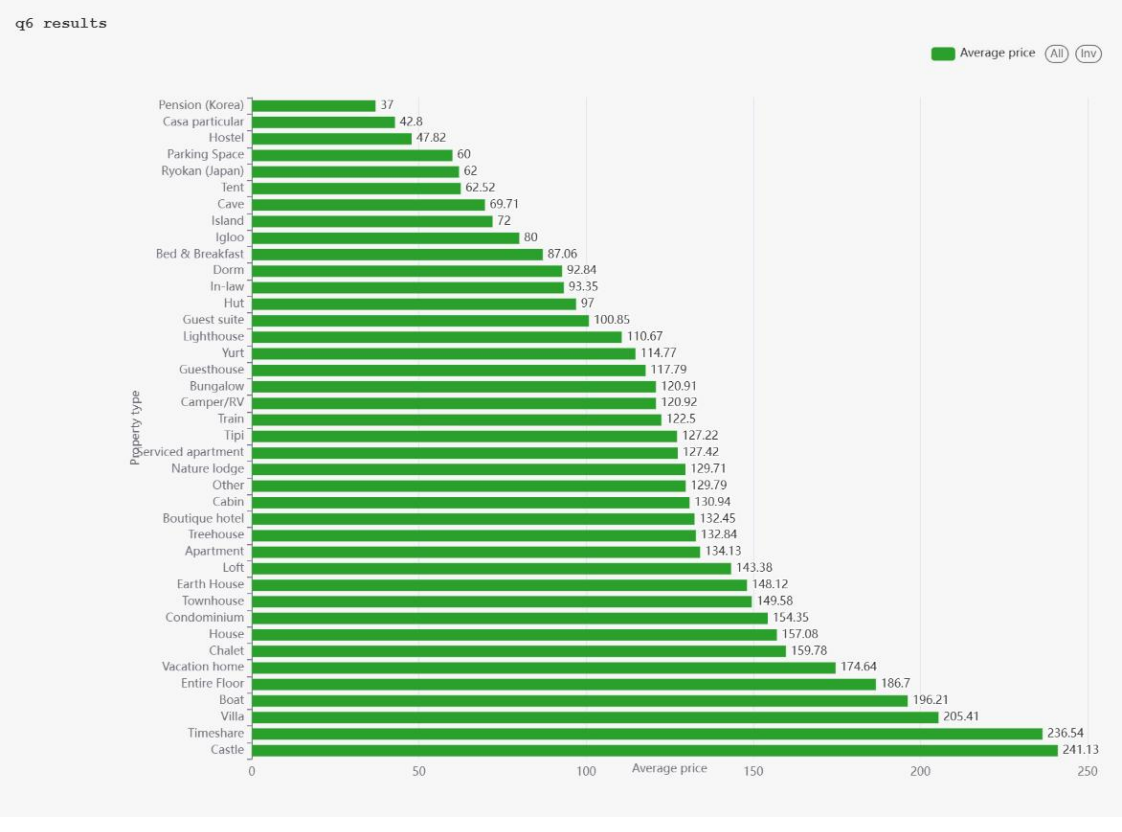
Relationship between listing ratings and average price

Correlation with our business goal: Directly relevant as ratings are proven price drivers. This analysis validates that higher-rated listings can command premium prices, informing how strongly to weight review scores in your model.

Data analyze results

04

Relationship between property types and their average prices

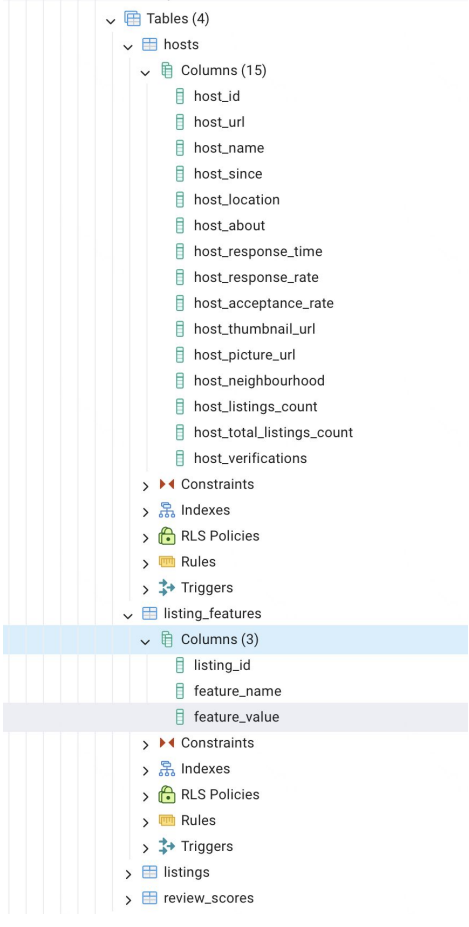


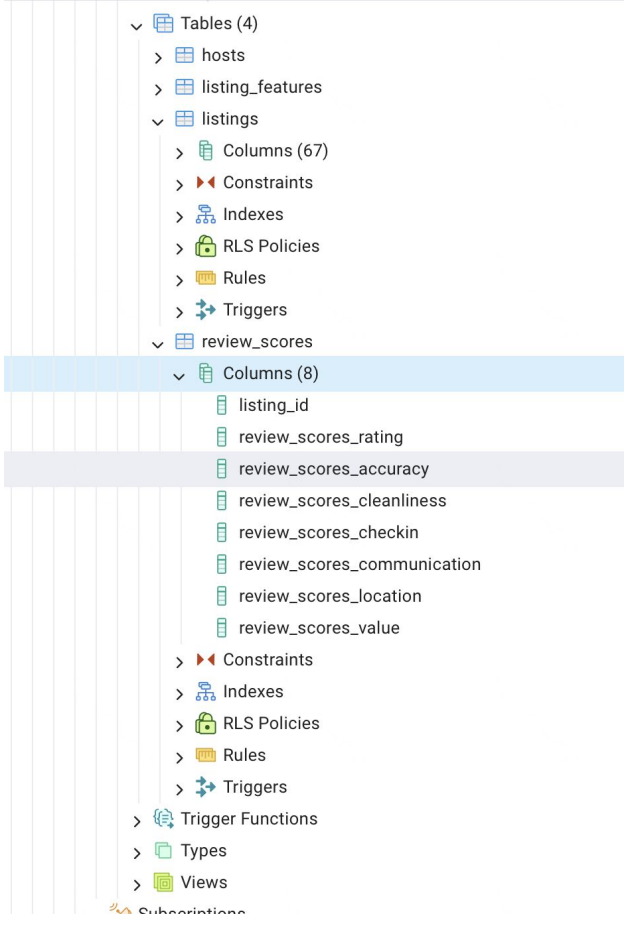
Correlation with our business goal: Establishes baseline price ranges by property category (apartment, house, etc.). This ensures your model provides type-appropriate recommendations and identifies which property features have premium value.

Results of each stage

05

Stage 1





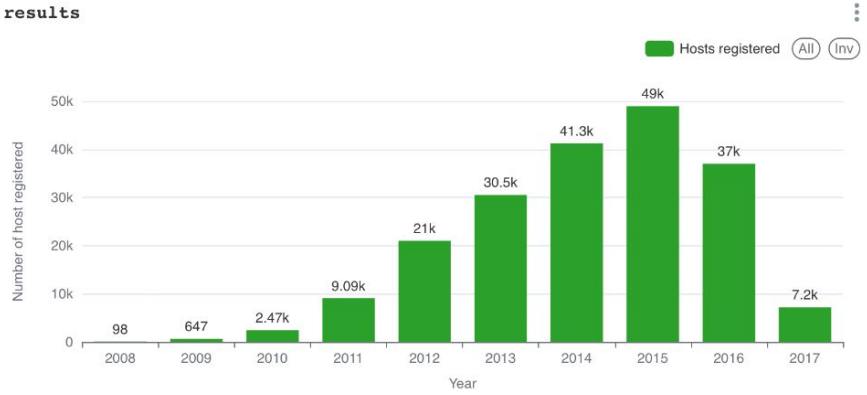
Results of each stage

05

Stage 2

- As the result of a stage 2 we obtained EDA tab on Superset Dashboard
- We analyzed key results and provide explanation

q1 results



What and why to investigate

This query examines when hosts registered on the platform by year. The insights worth investigating include:

Annual registration trends to track platform growth over time Peak registration years to identify successful growth periods Slowing registration periods to spot potential market challenges

Key insights

Based on the host registration data from 2008-2018:

- **Growth Pattern:** The platform experienced minimal registrations in 2008-2010, followed by rapid growth peaking in 2015 with approximately 49,000 new hosts.
- **Peak and Decline:** After reaching its peak in 2015, registrations dropped significantly through 2018, returning to much lower levels.

Results of each stage

05

Stage 3

- We created pipeline to fit the data to the model
- We obtain champion model: GBT regressor
- We perform grid search of the hyper-parameters

results of hyper-parameter optimization linear regression

reg_param	elastic_net_param	rmse
0.01	0.2	74.48334703899742
0.01	0.6	74.38998113751184
0.1	0.2	74.19036222418374
0.1	0.6	73.82293867129293

Linear model with reg_param = 0.1 and elastic_net_peram = 0.6 having rmse~73.82 is the best linear model

results of hyper-parameter optimization GBT regression

max_depth	min_instances_per_node	rmse
3	1	77.77290373933313
3	2	77.77290373933313
5	1	71.05597877690413
5	2	71.03169946311532

GBT model with max_depth = 5 and min_instances_per_node = 2 having rmse~71.03 is the best GBT model

Comparison table

model	rmse	r2
LinearRegressionModel: uid=LinearRegression_4d786dddc613, numFeatures=9436	71.94502471391714	0.725904604749003
GBTRegressionModel: uid=GBTRegressor_ccf92d010b38, numTrees=10, numFeatures=9436	71.00371143654189	0.7459944557845417

Challenges encountered

06

- Huge dataset with a lot of text features which affect data loading and data preparation
- Problem with view in superset
- Problem with passing python environment via spark submit

Demo

07

THANKS

DO YOU HAVE ANY QUESTIONS