Databnb: Modeling Airbnb Prices in Los Angeles: Exploring Key Predictors Beyond Room Type

GitHub Repository: Lab 2 Repository

Carlos Santander Kirthi Shanbhag Man Vilailuck

April 17, 2025

Introduction & Motivation

Airbnb has transformed the short-term rental market, offering a wide range of lodging options in cities like Los Angeles. While it's well known that factors like property type and seasonality affect prices, there is less clarity around how less obvious features, such as number of amenities, occupancy history, and reviews, influence nightly rates.

In this study, we ask: *What is the relationship between Airbnb listing price and listing-specific characteristics like number of amenities, number of bedrooms, and estimated occupancy in Los Angeles?* Our goal is to inform both hosts and platforms like Airbnb about the most influential drivers of pricing by building a regression model that captures subtle, yet impactful listing characteristics.

Data and Methodology

We focus on recent listings, from the year 2024 onwards due to factors like Market Relevance, where Older listings may include properties that are no longer active or priced based on outdated market conditions. Using recent listings ensures our analysis reflects the current pricing landscape. Another factor we considered was User Behavior: Airbnb guests tend to rely more on recent reviews when making booking decisions. By focusing on listings with recent activity, we're modeling the pricing strategies of listings that are actually competitive and in-demand today.

We focused on guest-relevant features that are often overlooked in pricing models, such as amenities and occupancy history. These variables are price, which are continuous, interpretable, and central to airbnb decisions, amenities, which are understudied but potentially influential, and number of accommodations/beds. These variables reflect value-added features that go beyond typical filters like location or season. At the beginning, our data, consisted of *45031 observations and 82 variables*.

The data was cleaned to remove NA values for price and NA values for all the other selected columns. Only relevant columns important for this study and columns with valid values were selected. The rest were all dropped from the final dataset. Some of the features consider for this study are as follows.

Amenities : A JSON Object of amineties listed for the property. - The column consists of JSON text and each amenities listed is separated by ','. Minimum amenities found were ranging from 1 to 120. - For our analysis and to avoid coercion warnings all the columns such as 'neighbour', 'Amenities' which were earlier non-numeric were converted to numeric values. - Few specific important amenities such as hot tub,wifi,kitchen,pool etc were converted to separate column with binary values. - Most of the individual amenities did not add any significant value to the model hence we introduced a new column count of amenities 'num_amenities'

Other features used include: - `neighbourhood_cleansed`: the neighborhood inferred from geolocation - `price`: nightly price (originally stored as text with a dollar sign, cleaned to numeric) - `number_of_reviews`: total number of reviews received

The final dataset consists of 8577 rows and 43 columns i.e. 14% were dropped from original dataset for last review year 2024. This is a good count of dataset to proceed for building regression model.

Exploratory Data Analysis

We explored key variables (accommodates, bedrooms, amenities, etc.) to understand distributions and correlations. All variables showed expected right skew. Bedrooms and accommodates had strong linear trends with price. Histogram plots, pairwise correlations, and summary statistics are included in the Appendix.

No of Reviews has the least correlation with response. Hence we will drop this variable and focus on the rest of the above independent variables. *Accessing the relationship between individual X features to access normality.*

The predictors and the response plots show a lot of skewness in the data.Linear regression is relatively robust to violations of normality, especially with large sample sizes. The Central Limit Theorem suggests that the sampling distribution of the estimates approaches normality as the sample size increases. Despite the observed skewness in both the predictor and response variables, we will proceed with constructing and evaluating regression models using various predictors. This approach allows us to understand the relationships between variables and assess the predictive power of different combinations of predictors.

Data Partitioning and Validation Partitioning the data into training and testing subsets to evaluate the model's performance, and employ cross-validation techniques to assess its robustness

ANOVA Test for Model Comparison

To statistically compare nested models, we conducted an *ANOVA (Analysis of Variance)* test between the following models:

- *Model 1 (`model_X1X2X3`)*: A log-linear model with predictors `accommodates`, `bedrooms`, `room_type`, and `estimated_occupancy_l365d`
- *Model 2 (`model_log_final`)*: The same model as above, but with the addition of the `has_hot_tub` variable

This comparison helps assess whether including `has_hot_tub` leads to a significantly better fit. The results are shown below.

```
anova(model_X1X2X3, model_log_final)
```

```
Analysis of Variance Table

Model 1: log(price) ~ accommodates + bedrooms + I(room_type == "Entire home/apt") +
    estimated_occupancy_l365d
Model 2: log(price) ~ accommodates + bedrooms + I(room_type == "Entire home/apt") +
    estimated_occupancy_l365d + has_hot_tub
  Res.Df    RSS Df Sum of Sq     F    Pr(>F)
1   2575 675.34
2   2574 664.90  1    10.444 40.43 2.403e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A low p-value (typically < 0.05) in the last column indicates that the added variable (has_hot_tub) significantly improves the model fit.

Validating the Model

We evaluated our final model using the confirmation set (test data). The output below summarizes model performance and significance of each predictor. *Key Takeaways from the Confirmation Model: All coefficients are statistically significant at $p < 0.001$, `accommodates`, `bedrooms`, and `has_hot_tub` positively influence log(price), Adjusted $R^2 = 0.61$ indicates solid model performance on holdout data*

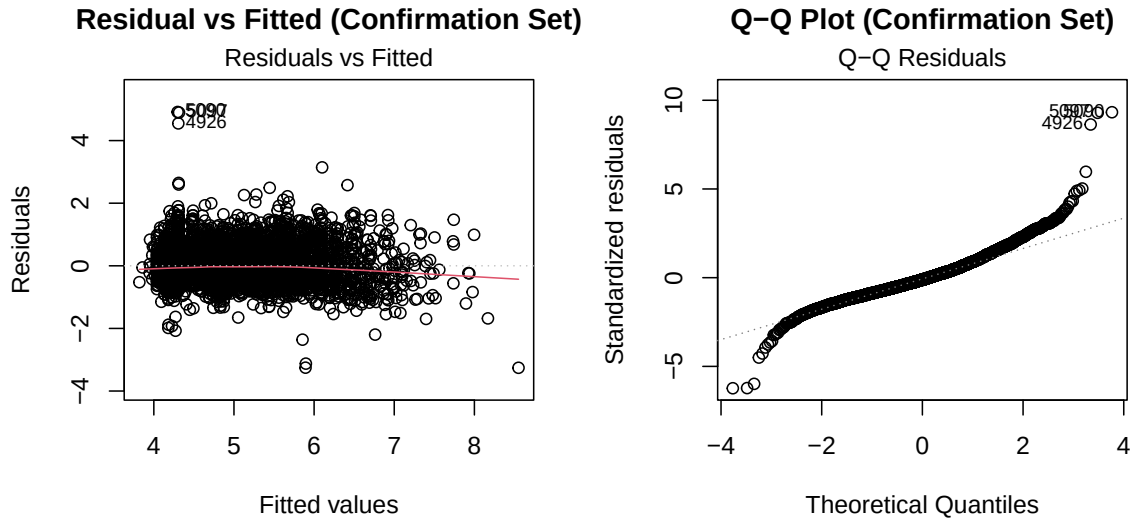
```
Call:
lm(formula = log(price) ~ accommodates + bedrooms + I(room_type ==
    "Entire home/apt") + estimated_occupancy_l365d + has_hot_tub,
    data = confirmation_set)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2573 -0.3389 -0.0556  0.2691  4.9134

Coefficients:
                                     Estimate Std. Error t value Pr(>|t|)
(Intercept)                         3.9037292  0.0164610 237.151  < 2e-16
accommodates                        0.0899234  0.0048056  18.712  < 2e-16
bedrooms                            0.2306863  0.0101631  22.698  < 2e-16
I(room_type == "Entire home/apt")TRUE  0.5588571  0.0173790  32.157  < 2e-16
estimated_occupancy_l365d          -0.0007252  0.0000917  -7.908 3.08e-15
has_hot_tub                         0.1987092  0.0263810   7.532 5.72e-14

(Intercept)                         ***
accommodates                        ***
bedrooms                            ***
I(room_type == "Entire home/apt")TRUE ***
estimated_occupancy_l365d           ***
has_hot_tub                         ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5271 on 6015 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.6144,    Adjusted R-squared:  0.614
F-statistic:  1917 on 5 and 6015 DF,  p-value: < 2.2e-16
```



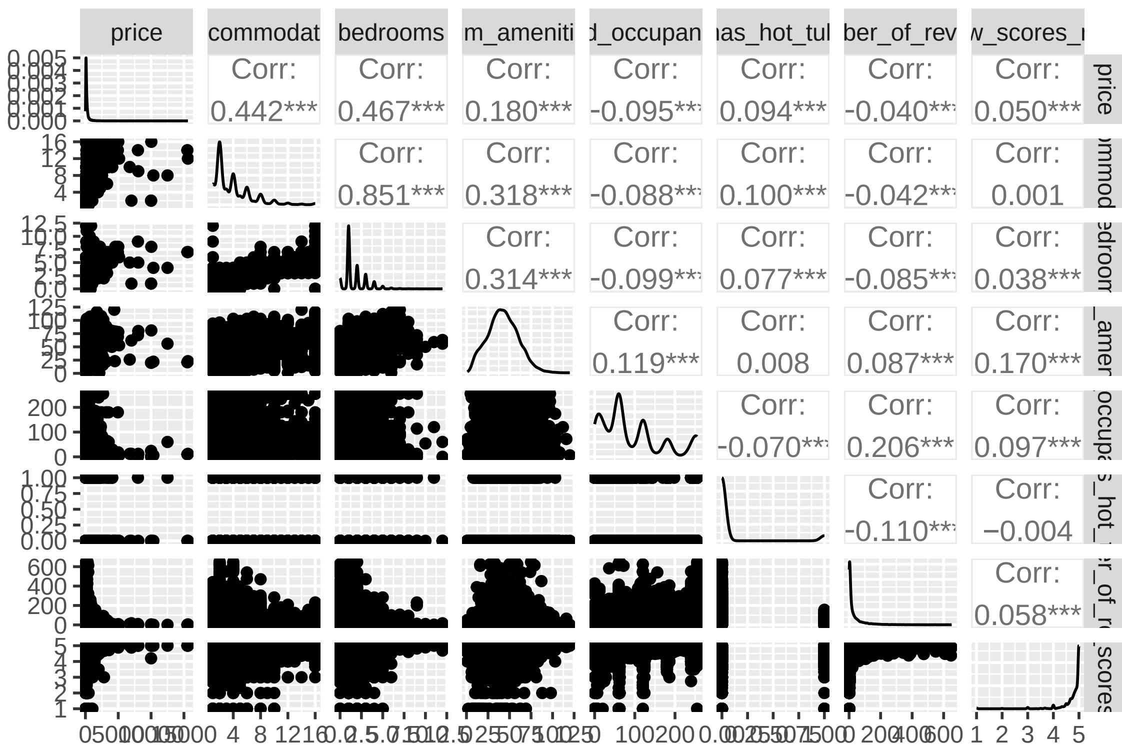
Key Takeaways & Future Work

Our analysis looked at which listing features, beyond just room type, help explain Airbnb pricing in Los Angeles. Using a log-transformed linear regression model, we identified five key predictors: accommodates, bedrooms, room type, estimated occupancy, and the presence of a hot tub. Together, these explain about 36% of the variation in nightly prices. ANOVA tests confirmed that amenities like hot tubs add meaningful value, and diagnostic plots showed our model performed reasonably well, despite some imperfections. We also compared actual and predicted prices from a holdout (confirmation) set to see how well our model generalized. While results aligned overall, some high-end listings showed wider dispersion, suggesting those prices are harder to capture with structured data alone.

For hosts, the results point to a few clear takeaways: listings that accommodate more guests and offer popular amenities like hot tubs tend to command higher prices. For Airbnb as a platform, our findings highlight the potential of enhancing pricing recommendations by incorporating features like amenity count and recent activity, details often overlooked in traditional pricing algorithms. Of course, pricing decisions are rarely driven by data alone. Visual appeal, guest reviews, and short-term demand surges (like holidays or local events) also play a big role. Our current model doesn't account for those. In future work, we'd love to explore how adding unstructured data, such as review text, listing photos, or neighborhood-level context, might improve prediction accuracy and give a fuller picture of what drives Airbnb pricing. Ultimately, even a fairly simple model, when thoughtfully designed and interpreted, can shed light on how travelers value different listing attributes, and how hosts and platforms can use that information to make smarter pricing decisions.

Appendix

A1. Pairwise Correlations



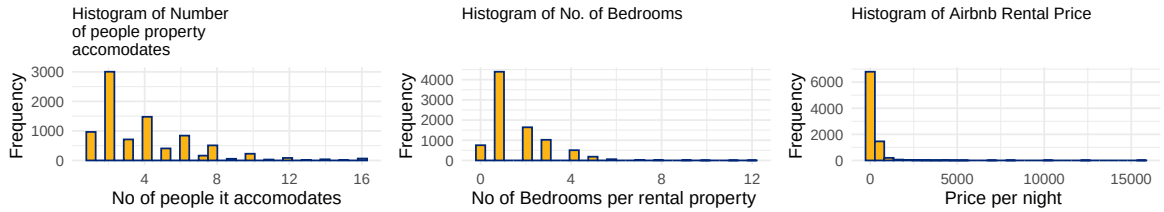
A2. Descriptive Statistics

	vars	n	mean	sd	median	trimmed	mad	min
price	1	8609	240.88	503.40	135.00	161.67	94.89	9
accommodates	2	8609	3.88	2.79	3.00	3.46	1.48	1
bedrooms	3	8601	1.67	1.26	1.00	1.52	0.00	0
num_amenities	4	8609	42.30	17.56	42.00	41.96	17.79	1
estimated_occupancy_l365d	5	8609	91.26	75.71	60.00	82.31	71.16	0
has_hot_tub	6	8609	0.07	0.26	0.00	0.00	0.00	0
number_of_reviews	7	8609	35.35	63.92	10.00	20.51	13.34	1
review_scores_rating	8	8609	4.78	0.43	4.91	4.87	0.13	1

	max	range	skew	kurtosis	se
price	15607	15598	14.87	337.17	5.43
accommodates	16	15	1.62	3.06	0.03
bedrooms	12	12	1.57	3.98	0.01
num_amenities	120	119	0.24	-0.08	0.19
estimated_occupancy_l365d	255	255	0.86	-0.29	0.82
has_hot_tub	1	1	3.23	8.45	0.00

```
number_of_reviews           651    650  3.99    22.22 0.69
review_scores_rating          5      4 -4.95    33.20 0.00
```

A3. Histograms of Key Predictors



A4. Actual vs Predicted Prices