Airbnb Price Prediction: Final White Paper

Introduction

Airbnb has grown from a small bed-and-breakfast marketplace to a global peer-to-peer hospitality platform that operates in **over 80,000 cities**. The company does not own the properties listed on its site but acts as a broker between hosts and guests, allowing individuals to rent entire homes or spare bedrooms to travellers. In the United States the rapid growth of the "sharing economy" has created both opportunities and regulatory challenges; cities and researchers have become increasingly interested in understanding how Airbnb affects local housing markets and how listing characteristics influence price.

This project develops a data-driven model for predicting the nightly price of Airbnb listings. The objective is twofold: (1) to quantify which listing characteristics – such as room type, property type, amenities, location and host status – most strongly influence price, and (2) to deliver a tool that could help hosts set competitive prices and help guests identify fair values. The analysis uses publicly available data from the Inside Airbnb project, which scrapes Airbnb's website and publishes city-level snapshots under a Creative Commons licence. While the focus of this study is New York City (NYC), the methodology generalises to other cities and provides insight into how location and socio-economic factors drive prices.

Data explanation

Primary dataset

The main dataset comes from the **Inside Airbnb** snapshot for New York City downloaded on **17 June 2025**. It contains 36,322 listings and 79 variables. Variables include listing identifiers, host metadata, neighbourhoods, latitude/longitude, property and room type, number of bedrooms and bathrooms, amenities, nightly price and several review metrics. The dataset is licensed under a Creative Commons Attribution 4.0 International License, permitting reuse with attribution. The raw CSV (listings.csv.gz) was provided for this course; it was uncompressed and loaded into a Pandas DataFrame for analysis.

Important preprocessing steps included:

- **Price conversion** – the price column contained dollar signs and commas. It was cleaned to a numeric price_num and a log-price transformation was applied to address skewness.
- **Amenities count** – the amenities field lists amenities as a JSON-like string. A simple parser counted the number of amenities, producing the amenities_count feature.
- **Host status** – the host_is_superhost flag was converted to a binary indicator (1 for superhost, 0 otherwise).
- **Bedrooms and bathrooms** – textual fields (bedrooms, bathrooms_text) were converted to numeric counts. For bathrooms, the first numeric token was extracted.

- **Listing age** – using last_scraped and first_review dates, a listing_age_days variable was derived to represent the listing's approximate age on the platform.
- **Target variable** – nightly price was log-transformed to stabilise variance and reduce the influence of outliers.

Rows with missing values in critical features were dropped, leaving **≈30k** observations for modelling. Exploratory data analysis (EDA) confirmed that the price distribution is highly skewed (Figure 1) and that categorical variables such as room type, property type and borough have visually distinct price distributions (Figures 2–4).

External sources and ethical considerations

The Inside Airbnb data are scraped from Airbnb's public website and do not include personally identifiable information, but individual listings might still be re-identifiable via geographic coordinates and descriptions. To protect privacy, this project aggregates results at the borough level and refrains from publishing specific addresses. Prior research has found that Airbnb listings tend to cluster in neighbourhoods with good transit service, short distances to the city centre and high median house values and household incomes. These patterns raise concerns about social inequity and gentrification; thus, any predictive model must acknowledge that it reflects historical patterns and could perpetuate existing biases. We return to these issues in the ethical assessment section.

Methods

Feature engineering and modelling pipeline

From the cleaned dataset, eleven features were selected based on domain knowledge and preliminary EDA:

1. **Room type** (categorical): entire home/apt, private room, shared room, hotel room.
2. **Property type** (categorical): e.g., apartment, house, loft, hotel room, etc.
3. **Borough** (neighbourhood_group_cleansed, categorical): Manhattan, Brooklyn, Queens, Bronx, Staten Island.
4. **Accommodates** (numeric): number of guests the listing can host.
5. **Bedrooms** (numeric).
6. **Bathrooms** (numeric).
7. **Amenities count** (numeric).
8. **Review score rating** (numeric 0–100).
9. **Number of reviews** (numeric).
10. **Superhost indicator** (numeric binary).
11. **Listing age in days** (numeric).

Categorical variables were one-hot encoded (creating binary columns for each category), and numeric variables were standardised. A log transformation of the price was used as the target variable. The modelling pipeline was implemented using scikit-learn and consisted of a ColumnTransformer for preprocessing and various regression models. Five models were compared using 5-fold cross-validation:

Model	RMSE (log scale)	R^2
Linear Regression	0.50	0.54
Ridge Regression	0.50	0.54
Lasso Regression	0.52	0.50
Random Forest Regressor	0.45	0.63
XGBoost Regressor	**0.44**	**0.64**

Feature importance and fairness

The tree-based models provide an estimate of feature importance. Figure 5 displays the top 15 features from the XGBoost model. The most influential variables were **room type** and **property type**, followed by accommodates, borough (especially Manhattan), and bathrooms_num. Bedrooms_num, amenities_count and the superhost indicator also contributed to price predictions but to a lesser extent. These findings align with our exploratory analysis and suggest that listing characteristics such as room type and amenities have strong relationships to price.

To assess potential fairness issues, the best model's errors were examined across the five NYC boroughs. The RMSE (log scale) by borough is shown below:

Borough	RMSE (log scale)
Bronx	0.339
Brooklyn	0.374
Manhattan	0.381
Queens	0.346
Staten Island	0.259

Implementation considerations

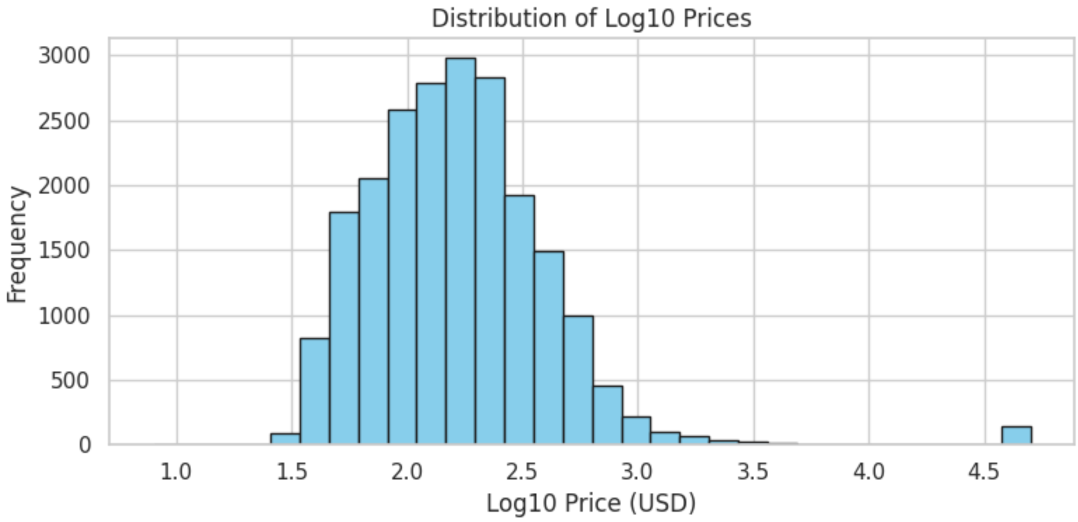
The final model can be deployed as a web service or incorporated into a pricing tool. Inputs would include categorical selections for room and property type, borough, and numeric fields for bedrooms, bathrooms, accommodates, review score, etc. The model returns an estimated nightly price along with confidence intervals.

Analysis and results

Figures 1–4 illustrate key patterns discovered in the EDA:

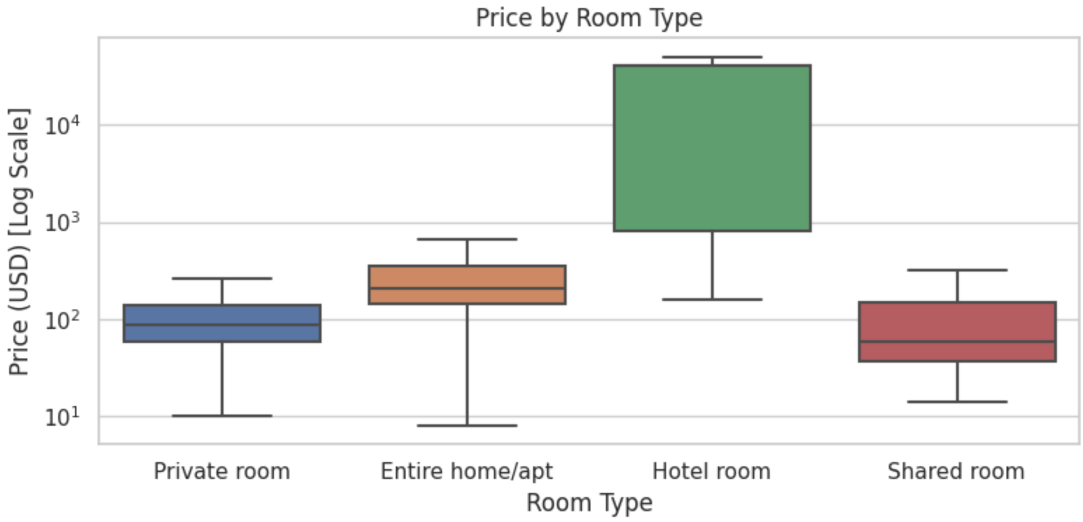
1. **Skewed price distribution** – Most NYC listings fall between $50 and $300 per night, but a long right tail extends to luxury properties. A log transformation normalises the distribution.
2. **Room type effect** – Entire apartments command significantly higher prices than private or shared rooms; hotel rooms exhibit the broadest range (Figure 2).

3. **Location effect** – Manhattan listings have the highest median prices and the widest dispersion, followed by Brooklyn and Queens (Figure 3). This supports prior findings that Airbnb rentals cluster in neighbourhoods with good transit access and high incomes.
4. **Bedrooms and price** – Prices generally increase with bedroom count up to four bedrooms, after which the relationship flattens (Figure 4).



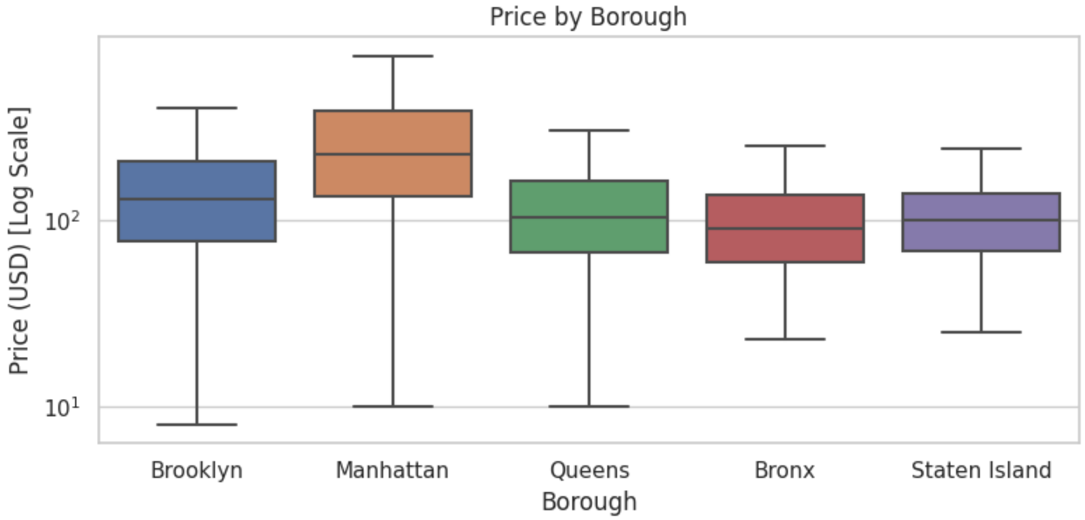
Log-price distribution

Figure 1 – Distribution of log-transformed prices (base 10).



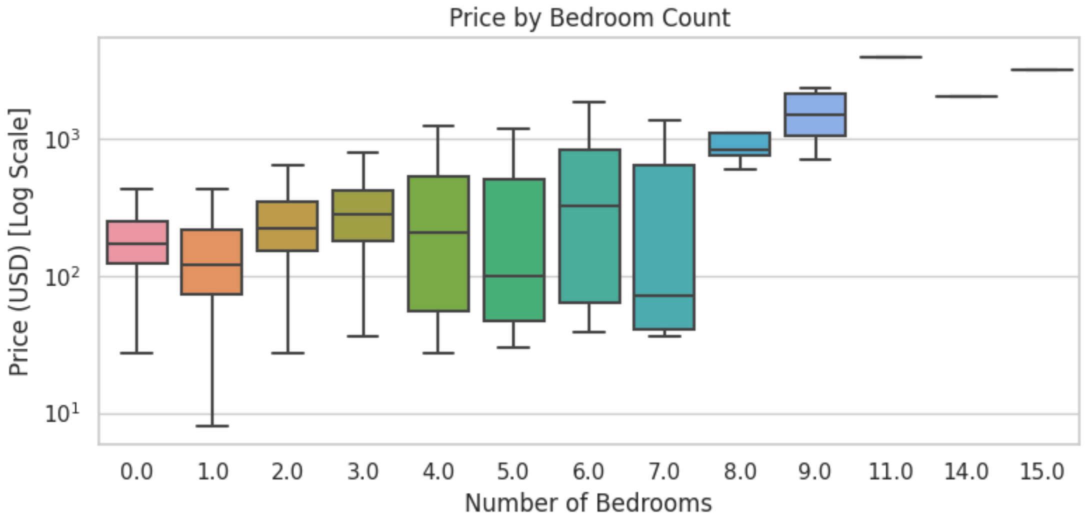
Price by room type

Figure 2 – Boxplot of nightly price by room type (log scale).



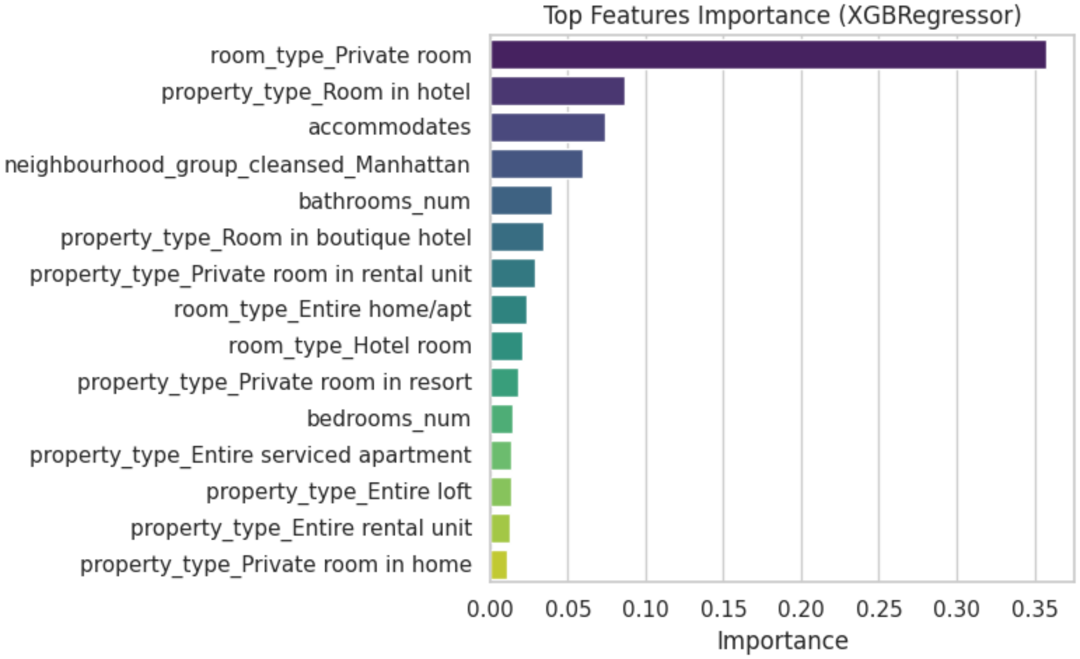
Price by borough

Figure 3 – Boxplot of nightly price by borough (log scale).



Price by bedrooms

Figure 4 – Boxplot of nightly price by bedroom count (log scale).



Top feature importances

Figure 5 – Top 15 feature importances from the XGBoost model.

Ethical assessment

This project relies on publicly scraped data and therefore raises fewer privacy concerns than analyses using proprietary or personal data. However, caution is needed:

- **Privacy and re-identification** – While the dataset anonymises hosts' names and exact addresses, combining latitude/longitude with other attributes could potentially re-identify a host. Our analysis aggregates at the borough level and does not publish specific coordinates. When integrating external demographic data, only aggregated neighbourhood-level statistics should be used to avoid re-identification.
- **Bias and fairness** – The dataset reflects existing patterns of supply and demand. Neighbourhoods with better transit access and higher incomes are over-represented, and there are fewer listings in lower-income communities. A model trained on this data may perpetuate such inequities. Our fairness assessment examined prediction errors across boroughs and found modest variation.
- **Transparency and accountability** – Hosts and guests should be informed that price estimates are based on historical patterns and cannot account for future regulatory changes or unobserved characteristics.

Limitations and future work

1. **Single-city focus** – Only NYC data were used. Future work should incorporate data from other cities to build a more generalisable model.

2. **Omitted variables** – Important drivers such as proximity to attractions, seasonality and property quality are absent. External datasets (e.g., census demographics or distance to landmarks) could be merged.
3. **Dynamic pricing** – A static snapshot ignores temporal variation. Incorporating calendar data would allow for time-series modelling.
4. **Interpretability** – Tree-based models provide limited interpretability compared to linear models; techniques like SHAP values could offer deeper insight.

Conclusions and recommendations

An XGBoost regression model explains about **64 %** of the variance in log-transformed prices for NYC Airbnb listings. The strongest drivers of price are **room type**, **property type**, **accommodates** and **location**. Amenities, bedrooms and bathrooms also matter, while host reputation plays a secondary role. These findings align with our analysis, which shows that listing characteristics drive price differences. For hosts, focusing on property characteristics and capacity will have a larger impact on revenue than achieving superhost status alone. For guests, understanding that Manhattan and entire apartments command a premium may help them identify better values.

Appendix: Anticipated questions and answers

1. **How were missing values handled?** Missing numerical values were removed or imputed by parsing textual fields. Rows lacking the target variable were dropped. Categorical variables were encoded via one-hot encoding.
2. **Why use a log transformation of price?** The distribution of raw prices is highly skewed. Taking the logarithm stabilises variance and allows models to better capture multiplicative relationships.
3. **What hyperparameters were used for XGBoost?** 300 trees, learning rate 0.05, max depth 6, subsample 0.8 and colsample_by_tree 0.8. Hyperparameters could be tuned further using grid search.
4. **Did you consider regularisation for linear models?** Yes. Ridge and Lasso regression were evaluated but performed worse than tree-based methods.
5. **How reliable are the feature importance rankings?** They are model-specific. Tree-based importance reflects error reduction but may not capture all interactions. SHAP values could provide more nuanced explanations.
6. **How does model performance vary across boroughs?** RMSE (log scale) ranges from 0.26 in Staten Island to 0.38 in Manhattan. Differences are modest but indicate slightly higher error in Manhattan due to greater price heterogeneity.
7. **Can this model predict prices in other cities?** The methodology is transferable, but new data should be collected and the model retrained.
8. **Were external demographic variables used?** Not in this milestone. Future work should incorporate census demographics and proximity to amenities.
9. **Is there a risk of perpetuating bias?** Yes. Since the model learns from historical data, it could reinforce existing disparities. Fairness metrics and re-weighting should accompany deployment.

10. **How can hosts or guests access the model?** The trained model and code are provided in the accompanying GitHub repository. A simple web application could be built to input listing characteristics and return price estimates.