FINAL PROJECT

1. Problem Formulation and Significance

The short-term rental market has grown rapidly over the past decade, with platforms such as Airbnb enabling individual hosts to compete in what is effectively a global hospitality market. For hosts, pricing decisions directly affect occupancy, revenue, and long-term competitiveness. However, pricing is difficult because demand is influenced by multiple interacting factors such as location, listing characteristics, and host behavior. This project focuses on understanding and predicting listing-level revenue performance rather than nightly price alone.

The problem addressed in this project is to predict revenue per available room (RevPAR) for Airbnb listings using observable listing and host characteristics. RevPAR is chosen as the target variable because it combines both price and occupancy, making it a more informative performance metric than either variable in isolation. The core research question is: to what extent can Airbnb listing attributes and host characteristics explain variation in RevPAR, and how well can a regression model predict RevPAR for unseen listings?

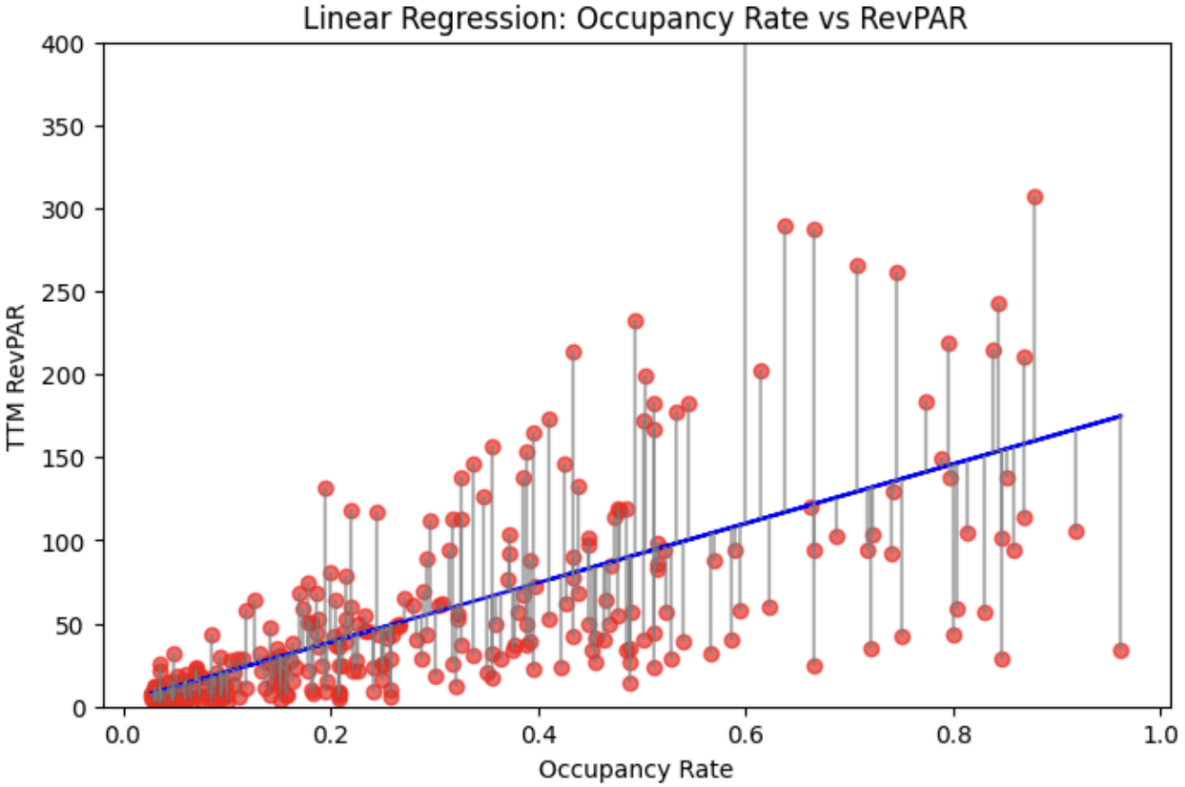
Multiple linear regression is used as the primary modeling approach, with model performance evaluated using out-of-sample metrics. The task is important because accurate revenue prediction can help hosts make better pricing and investment decisions, while also illustrating how interpretable machine learning models can be applied to real marketplace data. The dataset used in this project consists of approximately 300 Airbnb listings, each representing a unique property.

The data include both quantitative and qualitative features. Quantitative variables include number of *beds, average rate, average occupancy, reserved days* and *available days* (all for trailing twelve months). Qualitative variables include *room type* and *superhost* status, which are encoded for use in the regression model. The target variable is *trailing twelve-month RevPAR*. The dataset is well suited for this project because it contains sufficient observations for regression analysis, includes economically meaningful predictors, and reflects real decision variables faced by hosts.

2. Exploratory Data Analysis

Initial data exploration focused on understanding distributions, identifying potential relationships, and checking data quality. Summary statistics showed substantial variation in RevPAR across listings, indicating that some listings perform significantly better than others. Average rate and occupancy rate both exhibited wide ranges, suggesting that pricing strategy and demand conditions differ substantially across properties.

Visual inspection revealed a strong positive relationship between occupancy rate and RevPAR, which is expected since higher occupancy directly increases realized revenue.

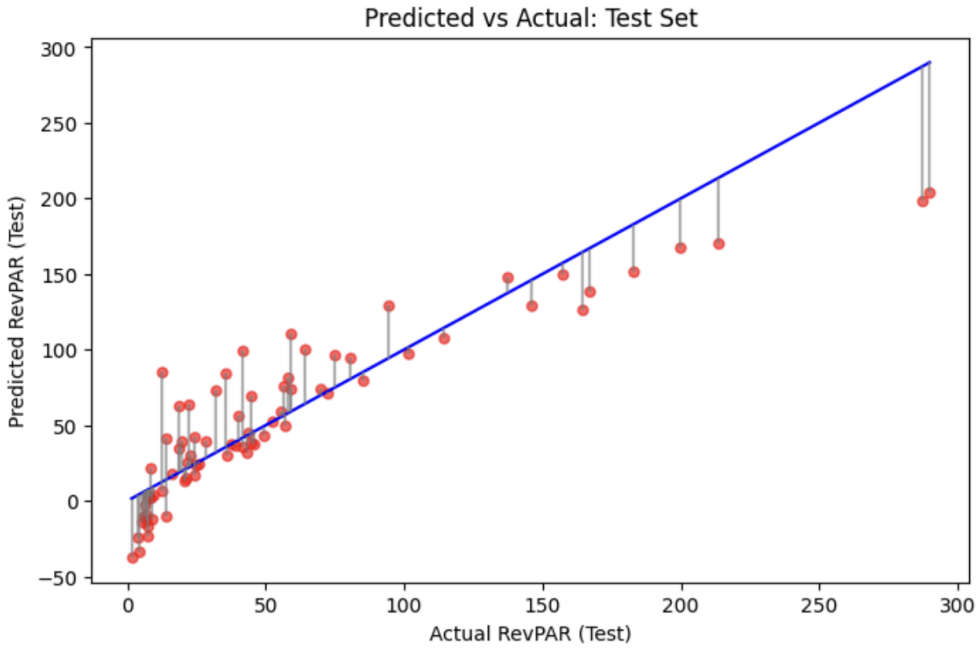


Missing data were minimal. Only three observations contained missing values in the beds variable and were removed from the analysis. Given the small number of missing entries relative to the dataset size, dropping these observations was unlikely to bias results. No strong class imbalance issues were present since the task is continuous regression rather than classification. Quantitative variables were used in their natural scales, as linear regression does not require normalization for correctness, and interpretability was preserved by keeping coefficients in meaningful units.

Data visualizations in the forms of histograms and scatterplots, were used to assess distributions and relationships. These plots indicated mild nonlinearity and heteroscedasticity, particularly at higher fitted values of RevPAR, which informed later diagnostic analysis.

3. Model Selection, Application, and Evaluation

Multiple linear regression was selected as the primary model because the goal of the project is not only prediction but also interpretability. Linear models allow direct interpretation of how each feature contributes to expected revenue, which is valuable in a business and policy context. The dataset size and feature structure also make linear regression an appropriate baseline model. The model was implemented from scratch using matrix algebra and the normal equation. This ensured a clear understanding of the underlying mechanics and minimized reliance on black-box implementations. The dataset was split into training and test sets, with approximately 75% of observations used for training and the remaining 25% reserved for test.



Model performance was evaluated using root mean squared error (RMSE) and R squared. These metrics were computed manually using predictions from the custom implementation. The base model achieved strong performance, with an out-of-sample R^2 of approximately 0.81 and an RMSE of around 27, indicating that a large proportion of variation in revenue can be explained by the selected features.

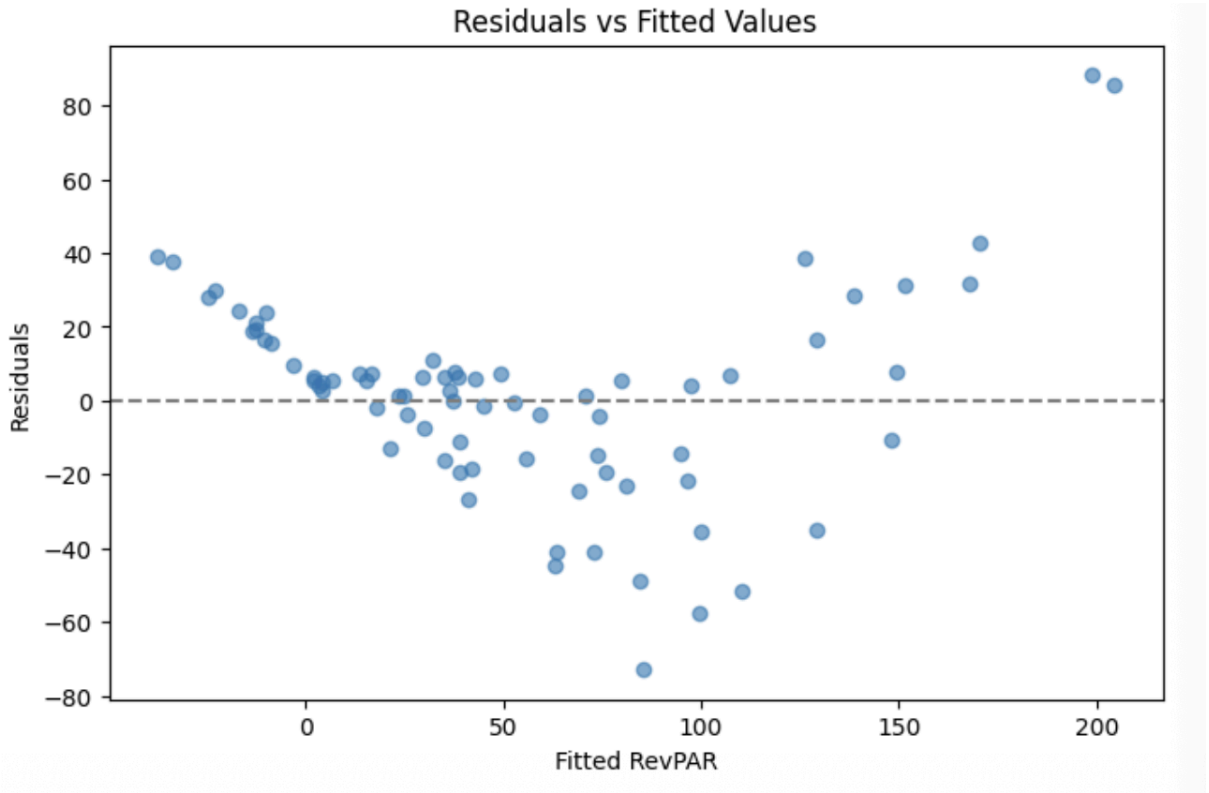
An alternative model including an interaction term was also tested to capture potential nonlinear effects. While this model improved in-sample fit, it performed worse on the test set, indicating overfitting. Based on this comparison, the simpler model was retained as the final specification. This trade-off highlights the importance of generalization rather than purely maximizing in-sample performance.

The table below outlines the different values obtained from the two different models:

	R^2	**RMSE**
Base model	*0.810*	*27.41*
Alternative model	*0.762*	*30.69*

A higher R squared and a lower RMSE is desired for the model, and we can observe that the base model satisfies both these criteria. For the base model, the one I proceeded to keep, we can interpret the $R^2 = 0.810$ as that the model reduces unexplainable variance by 81%, meaning that the remaining 19% is dependent on factors I do not observe or model. The RMSE of 27.41 is in the unit of a Swedish Crown (SEK), the currency used for the target variable. This means that the typical prediction error for the model is around 27 SEK ($2,91 as of today), which is arguably relatively low for listing prices of up to several hundreds of dollars.

Residual diagnostics were conducted to evaluate model assumptions. Residuals versus fitted values showed mild heteroscedasticity and a weak U-shaped pattern, suggesting that some nonlinear structure remains unmodeled. A histogram of residuals showed heavier tails than a normal distribution, indicating the presence of outliers or high variance listings. These issues were acknowledged rather than ignored, reinforcing that the model is a useful but imperfect approximation.



Finally, the results were verified using scikit-learn's LinearRegression implementation. The sklearn results fully matched those of the custom model, confirming the correctness of the implementation and evaluation pipeline.

4. Results, Conclusions, and Real-World Implications

The results demonstrate that RevPAR for Airbnb listings can be predicted with reasonably high accuracy using a small set of intuitive features. From a practical perspective, these findings suggest that hosts should consider all of the predictors used in this model to set a competitive price for the market. The model highlights that revenue optimization is a balancing act rather than a simple price maximization problem. Several limitations should however be noted. The linear model does not fully capture nonlinear effects or extreme outliers, and the dataset is limited to a specific market and time period. External factors such as seasonality, local events, and regulatory constraints are not included. More complex, qualitative factors such as neighborhoods, crime rates and any other hard-to-measure parameter are not accounted for at all, which intuitively contribute a good amount. Future work could explore nonlinear models, log transformations, or tree-based methods to address some of these limitations.

This project strengthened several skills directly relevant to future academic and professional work. These include implementing machine learning models from first principles, designing reproducible analysis pipelines, interpreting statistical results in a business context, and clearly communicating findings. The ability to move from raw data to defensible conclusions mirrors real-world analytical work in consulting, finance, and data science.