Predicting House Prices Using Key Property Features Through Multiple Regression

Hamza Ahmed, Cifti Saggu, Daniella Jaqin, Mike Xu, Zichen Liu

Abstract

This report examines key variables influencing house prices in New York, focusing on the relationships between property characteristics and pricing to provide insights for stakeholders in the real estate market. Using multiple regression analysis, we found that the number of bathrooms and rooms have the most significant positive impact on housing prices, reflecting the value buyers place on space and amenities in a densely populated city like New York. Additionally, the age of a house negatively impacts its price, indicating a preference among buyers for newer properties.The model's performance indicators, including MAE and RMSE, suggest reliable accuracy in estimating price fluctuations. Additionally, consistent AIC and BIC values indicate a good balance between accuracy and simplicity for our selected model. This report serves as a useful reference for property valuation and supports stakeholders in making informed decisions.

Introduction

The current housing market is a complex and dynamic environment influenced by various factors. Being able to understand the key drivers of house prices is crucial for stakeholders, including potential buyers, sellers, and investors. This analysis aims to answer several important questions. These questions are: What property features most significantly impact house prices? How do these features interact with one another to influence overall value? By utilising multiple regression analysis, we seek to quantify the relationships between key numeric variables and house prices, providing insights that can guide decision-making in the real estate market.

Data set

The dataset used for this analysis was collected from The Data And Story Library (DASL), capturing a range of properties across New York. It includes numerous variables related to property characteristics, focusing on numeric predictors that correlate well with regression analysis. The key variables in this dataset include Lot Size, Age, Land Value, Living Area, Pct College, Bedrooms, Bathrooms, and Rooms. The dataset excludes categorical variables, focusing on continuous numeric features that provide measurable attributes relevant to property valuation. This approach enables a clearer analysis of how these features affect house prices, allowing for a more systematic understanding of their impacts.

Analysis

To analyse the dataset using multiple regression, we focused on the numeric predictor variables, specifically Lot Size, Age, Land Value, Living Area, Pct College, Bedrooms, Bathrooms, and Rooms. We chose to exclude categorical variables, as many such as waterfront or fireplaces are based on personal house preferences rather than common, quantifiable features that may have a consistent impact on price. Continuous numeric features further provide measurable attributes that are easier to evaluate a prediction of price.

To ensure the model's validity, the numeric predictors were systematically checked against the key multiple regression assumptions. For linearity, we visualized each predictor's relationship with log-transformed price, leading to the log-transformation of age due to slight curvature. **(Appendix 1)** Independence was verified by assessing the Variance Inflation Factor (VIF) values, where all values were below 5, indicating acceptable levels of multicollinearity **(Appendix 2)**. Homoskedasticity was examined through a residual plot **(Appendix 3)**, suggesting constant variance of residuals. Normality was checked with a Q-Q plot **(Appendix 4)**, which showed residuals following the quantile line closely, with minor deviations at the tails. Based on these checks, we confirmed that the final model meets the assumptions required for multiple regression without further adjustments.

Results

For our results, we evaluated three model selection methods, the forward selection, backward selection, and exhaustive search. Interestingly, all three models produced identical performance metrics **(RMSE: 0.2686, MAE: 0.2042, R^2: 0.5613)** and shared the same **AIC (451.24)** and **BIC (483.64)** values. Since the statistical criteria did not differentiate them, we selected the forward model for its practical advantages.Forward selection is computationally efficient compared to exhaustive

search, which becomes costly with additions to the dataset. While less expensive than exhaustive search, backward selection can still retain interdependent or non-informative variables, raising the risk of overfitting. The forward model avoids these issues by adding only essential variables, resulting in a more robust and interpretable model that mirrors real-world property evaluation. Its gradual selection approach also builds stakeholder confidence by focusing on factors typically prioritised in property assessments making forward selection a practical, reliable choice.

Our final model for predicting house prices is given by the equation:

$$\log(\text{Price}) = 11.395433 + 0.000377 \times \text{Living Area} + 0.103651 \times \text{Bathrooms} - 0.047154 \times \log(\text{Age}) + 0.007997 \times \text{Rooms}$$

This equation shows how each predictor contributes to house price, both statistically and practically. Adding a bathroom or a room results in about a 10.37% and 0.8% price increase, respectively, due to the significance of each "unit increase." In real terms, an additional bathroom or room is a substantial change, aligning with how buyers perceive these features in property value. Living area, with a smaller coefficient (0.0004), can still impact price substantially but with a larger 'unit size' (e.g., 1,000 square feet). The negative coefficient for age (-0.0472) reflects a preference for newer homes among stakeholders. This interpretation balances statistical results with real-world relevance, ensuring our insights are meaningful for stakeholders.

Discussion

Our analysis uses a multiple regression model to examine key property characteristics that influence New York home prices, focusing on numeric predictor variables.

-Appendix 1 shows the linear relationship between the log-converted price and the predictors (log age, living area, rooms and bathrooms), confirming the correlation of the predictors.

-In Appendix 2, VIF values verify independence, and all values are below 5.

- The residuals plot in Appendix 3 shows that the variance of the residuals is constant.

- The Q-Q chart in Appendix 4 shows that residuals follow a normal distribution. Our final model provides a look at how certain variables (living area, number of bathrooms, rooms, and age) affect house prices:

- Although the coefficient for living area is small (0.0004), it doesn't fully capture the value buyers place on space as large living areas (e.g.1,000 square feet) reflect a substantial relative impact on price. Each additional bathroom raises the price by approximately 10.37%, highlighting the high value buyers place on space and amenities

- Age has a negative coefficient, indicating that as homes age, their value generally declines.

- The increase in the number of rooms resulted in a small increase in prices (0.8%). These findings have benefits for stakeholders, including potential future homeowners, investors and policymakers, who rely on this knowledge to make decisions.

Limitations and Future Improvements

Our dataset presented limitations due to its predominantly categorical nature, which restricted the continuous predictors available for modelling.Of the 17 initial variables, 8 continuous ones were considered for the models, but only those that met the linearity, normality, independence, and homoskedasticity assumptions were used. This reduced the model's depth and limited its ability to capture finer details. To enhance our model, future work could benefit from a more comprehensive dataset that includes location-specific and economic indicators, such as inflation and interest rates, as well as additional continuous predictors for greater granularity.

Real estate prices often exhibit non-linear trends that a linear model like multiple regression cannot fully capture, limiting the scope of housing features that our model can analyse. Variables with non-linear relationships were either excluded or transformed, which can lead to bias by either omitting their impact on price if excluded, or perhaps underrepresent their effects if transformed. Exploring non-linear models, such as decision trees or random forests, could better accommodate categorical and non-linear variables, capturing complex relationships that a linear model may overlook. Finally, since the model is specific to New York, its application to other regions or markets is limited, suggesting that further region-specific models may be required for accurate predictions elsewhere.

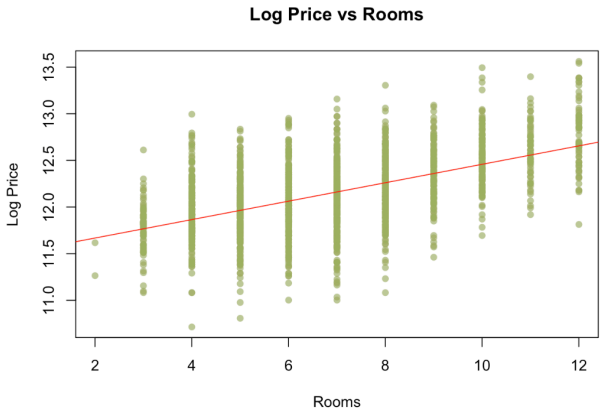
Conclusion

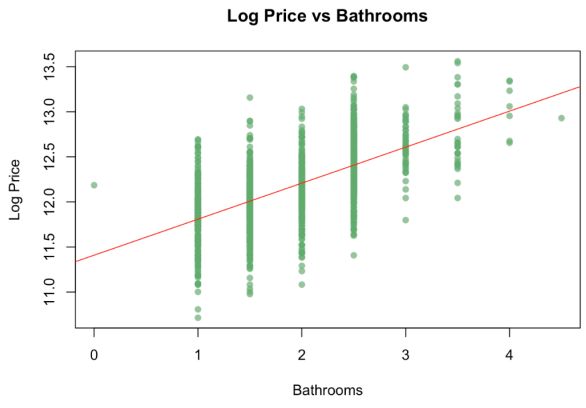
In summary, the regression model we used in this report identifies the key variables that influence New York house prices, providing a useful tool for property valuation. Hopefully, our findings can provide meaningful insights to those involved in the housing market that can help them make informed decisions.

Git Repository

https://github.sydney.edu.au/djaq0514/L12G03

Appendix 1

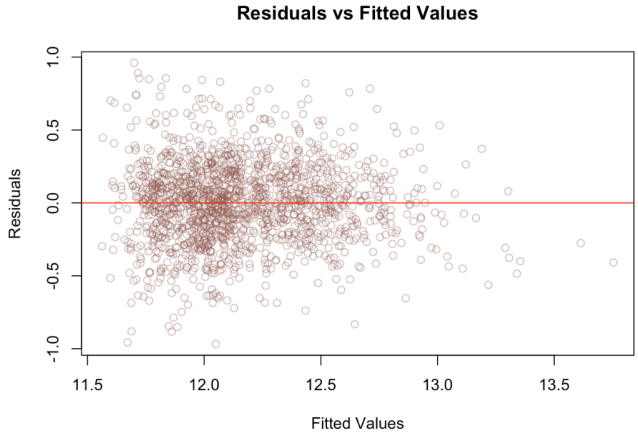




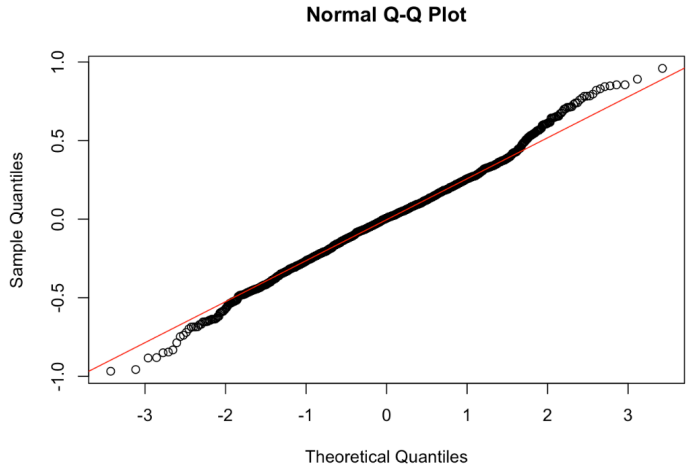
Appendix 2

Predictor	VIF Value
log_age	1.249816
living_area	3.181138
bathrooms	2.287007
rooms	2.109127

Appendix 3



Appendix 4



References

1. **Quarto Documentation.** (2024). *PDF Basics.* Retrieved from https://quarto.org/docs/output-formats/pdf-basics.html#latex-output.

2. **Quarto Documentation.** (2024). *Article Layout.* Retrieved from https://quarto.org/docs/authoring/article-layout.html#pdflatex-layout.