House Prices in Saratoga, New York

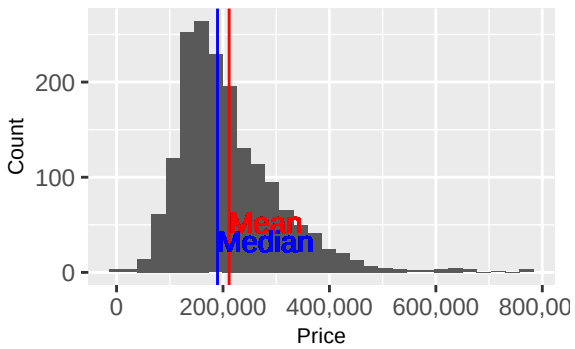
This version was compiled on November 5, 2023

This report explores how different characteristics impact house prices in Saratoga, New York. The initial model selected in the presentation was the Full Logarithmic Stepwise Forward model for its statistical robustness and performance. Upon further analysis, this model was enhanced and the Selective Logarithmic Stepwise Forward Regression model was introduced, after the variables: "Bedrooms", "Bathrooms" and "Rooms" were dropped due to multicollinearity. The variables that impacted the house prices the most in the final model were: "LivingArea", "Waterfront" and "LandValue", having a direct relationship, and "FuelType (Wood)" having an inverse relationship with Price. Future research on per square foot pricing and neighbourhood data is also suggested.

Introduction. The ongoing housing shortage and rising prices remain highly concerning for the younger generation, fueled by low supply alongside high demand stemming from challenges in construction, high immigration, and soaring interest rates (Mousina, 2023). The accurate prediction of price based on house characteristics, is a vital ability for young individuals to gauge the fair market value of a property. Knowledge of which factors influence price is a valuable financial literacy skill and assists to inform decisions about property buying or investment. Therefore, we seek to determine a model which can explain the most contributive house characteristics when estimating the price of a home.

Data Set. We investigated data collected on houses in Saratoga County, New York, USA in 2006, consisting of 1734 observations and 17 variables, including a 'test' variable that was excluded due to uncertain meaning. The key dependent variable of analysis is house price (price, Fig 1), with 15 additional explanatory house characteristic variables potentially used to predict price. These include the size of the lot in acres (lotSize), house age in years (age), land value in USD (land.Value), percentage of neighbourhood that graduated college (pct.college), type of heating system (heating.type), fuel used for heating (fuel.type), type of sewer system (sewer.type), whether the property is waterfront (waterfront), whether the property is new (new.construction), living area in square feet (Living.Area) and whether the house has central air conditioning (central.air). It also includes the number of bedrooms (bedrooms), bathrooms (bathrooms), and rooms (rooms). For preprocessing, we ensured all data entries are valid and there were no missing values.

Fig 1: Histogram of Prices

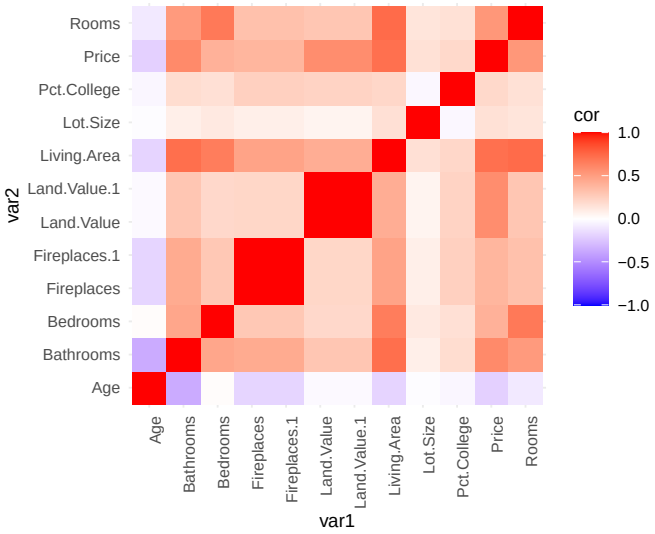


Analysis.

Model Discovery and Selection. After determining the log stepwise forward model to be our preliminary optimal model, we decided it was necessary to investigate further upon feedback around concerns of multicollinearity, and interaction effect.

Based on our findings of multicollinearity, it was necessary to select specific variables to be dropped from the model building process. Evidently, there is high correlation between the variables LivingArea, and the number of Bedrooms, Bathrooms, and Total Rooms (Fig 2). Analytically, this is a direct result of these variables being indicative of the overall size of a home, and hence living area size. Therefore, we concluded that these three variables were providing redundant information, so to stabilize the model without sacrificing predictive power, they were dropped from the model.

Fig 2: Correlation Matrix



From our interaction plot, we determined there was interaction effect between Rooms:Lot.Size, and adding in the effect did slightly improve the performance of the model. However, Rooms was related to multicollinearity, and it was crucial to address this issue first to ensure the regression model was reliable and interpretable, which meant that in our final model, no interaction effect was included.

Model Evaluation. The in-sample performance of our models was assessed using R^2 and Adjusted R^2 value. The R^2 value measures the proportion of variance in Price explained by the predictors while adjusted R^2 takes into account of the number of predictors as well. While Full Model had the highest value (Table 1), we recognised that in our dataset there were factors which should be logged as they were overly large, and hence we conducted log transformation on other models. Excluding the Full Model, all other models indicated similar in sample performance.

The out-sample performance of our models was assessed with RMSE and MAE through performing 10-fold cross validation. From Table 2, Stepwise Forward model had the smallest RMSE and MAE

Table 1. In Sample Performance for different Models

Models	R Squared	Adjusted R Squared
Full Model	0.6553	0.6509
Log Transform	0.5941	0.5889
Log Stepwise Forward	0.5919	0.5889
Log Stepwsie Backward	0.5935	0.5897
Log Full Interaction Effect	0.5947	0.5892
Log Full No Collinearity	0.5819	0.5773

Table 2. Out Sample Performance for different Models

Models	RMSE	MAE
Full Model	58530	41550
Log Transform	0.2905	0.2077
Log Stepwise Forward	0.2884	0.2068
Log Stepwsie Backward	0.2897	0.2073
Log Full Interaction Effect	0.2914	0.2081
Log Full No Collinearity	0.2952	0.2111

values at 0.2884 and 0.2068, which indicated that this was the optimal model in minimising errors. However, as mentioned in model discovery and exploration, this preliminary model had issues of multi-collinearity. We thus attempted to explore a Log Model dropping multi-collinearity variables, but found RMSE and MAE to be the highest amongst all models explored. Therefore, we concluded that the best option is to perform Stepwise Forward on a Selective Log Model which excluded variables causing multi-collinearity but could still ensure a relatively small RMSE and MAE.

Assumption Check. Our final model meets the independence assumption, with a Durbin-Watson statistic of 1.59. Testing for homoscedasticity, the model meets the assumption with only a few points appearing outside of the main cluster suggesting slight heteroscedasticity. Similarly, regarding normality, the majority of the points are very close to the line, with a few outliers near the head and tail of the data. For both assumptions, these abnormalities can be ignored due to the Central Limit Theorem as the vast majority of data points meet the assumptions. As the dataset has over 1000 samples, it is a big enough dataset to omit the effects of outlier (see Appendix).

Results. The final model we have found based on our exploration is a Selective Logarithmic Stepwise Forward Model which addresses multi-collinearity as well as bringing statistical robustness:

Log(Price) = 5.83 + 0.69Log(LivingArea) + 0.13Log(LandValue) + 0.54Waterfront + 0.06HeatType(HotAir) + 0.05HeatType(HotWater) + 0.04LotSize - 0.001Age - 0.1NewConstruct - 0.002%College + 0.06CentralAir + 0.04FuelType(Gas) = 0.86FuelType(Wood)

From Table 3, our final model does not have the smallest RMSE and MAE values comparatively. However, there is only a small difference of 0.0066 and 0.0035 respecitvely in RMSE between the most optimal model performance in Table 2 and Final Model, which is acceptable. It is evident that removing multi-collinearity variables are the cause of this slightly worse performance, but since it is essential to address this issue to derive a more reliable model, it can still be claimed that the Final Model is indeed the optimal predictor.

Table 3. Out Sample Performance for Final Model

RMSE	MAE
0.2951	0.2103

Our final model indicates LivingArea and LandValue have expected positive relationships with Price. A 1% increase in LivingArea corresponds to a 69% rise in Price on average, holding other variables constant. LandValue's smaller coefficient, of a 13% increase, suggests physical attributes may drive Price more than location. Notably, Waterfront being present predicts a 54% increase in Price versus non-Waterfront homes, all else equal. This substantial premium indicates waterfronts are highly desirable and influential on Price, underscoring buyer preferences. This serves to empower the evaluation of price between non-waterfront and waterfront properties of similar characteristics..

The negative coefficients of the model highlight that having FuelType of Wood, holding other factors constant, results in a 86% **decrease** in Price on average, suggesting this fuel source is viewed as undesirable and aligns with current cost effectiveness of alternative fuel sources. Furthermore, an interesting anomaly was observed in that a 10 percent increase in the Percent of Neighbourhood College Graduates (PercentCollege) leads to a 2% **decrease** in Price on average, holding all other variables constant. This contradicts intuitive thinking, however it may be attributable to cultural, social or other demographic information specific to the location of data collection.

Surprisingly, being new construction predicts 10% lower Price versus old homes on average, all else equal. While Age has a small 0.1% negative coefficient as expected, new builds decreasing Price is counterintuitive, raising questions about their quality or desirability. With rising prices, it would be expected for houses to appreciate over time when holding other factors constant. However, both Age and NewConstruct have conflicting interpretations, likely because renovations or restoration status are unknown. Further investigation on their precise influences is needed, since the data limitations provide insufficient insights into these relationships.

Ultimately, the aforementioned interpretation of the final model reveals some counterintuitive relationships between house characteristics and the prediction of price. This poses further research opportunities into the causes and market nuances behind these effects

Discussion and Conclusion. Under the condition of removing multi-collinearity, the Final Model determined can be considered the most optimal in performance while minimising errors. Analysing the model further, the most prominent influencers of the house price variance appears to be LivingArea, LandValue and Waterfront. A limitation of the included variables was the lack of interpretation of the quality and desirability of the location of the properties. Being pivotal components of determining property price, the addition of such variables that could quantify this information could have further supported our final model.

Future exploration should include the development of an optimised dependent variable such as Price per square foot. This metric is highly recognised as optimal for determining property desirability and quality. Additionally, our investigation was solely based on house characteristics as per the limitations of the dataset. Therefore, future research should capture and examine the influence of additional neighbourhood and demographic information, including, but not limited to, occupancy status (percentage of renters or owners), facilities and number of transport options. This may lead to a more accurate analysis and optimised model.

Appendix.

Fig 3:Homoscedasticity + Linearity

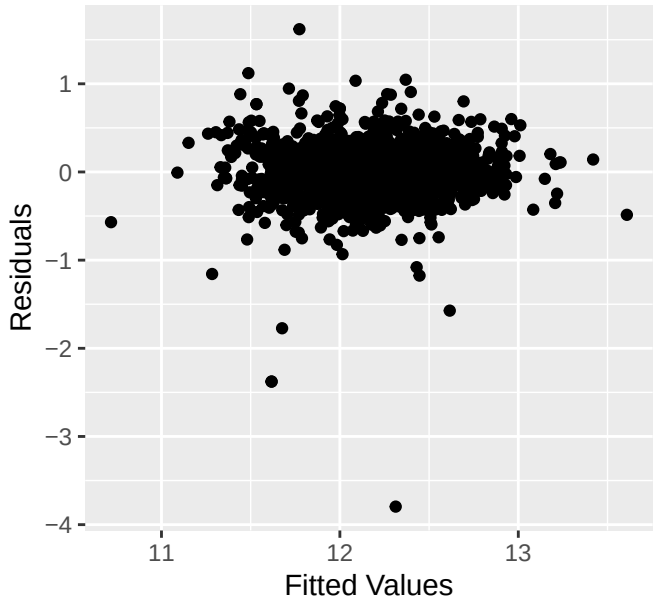
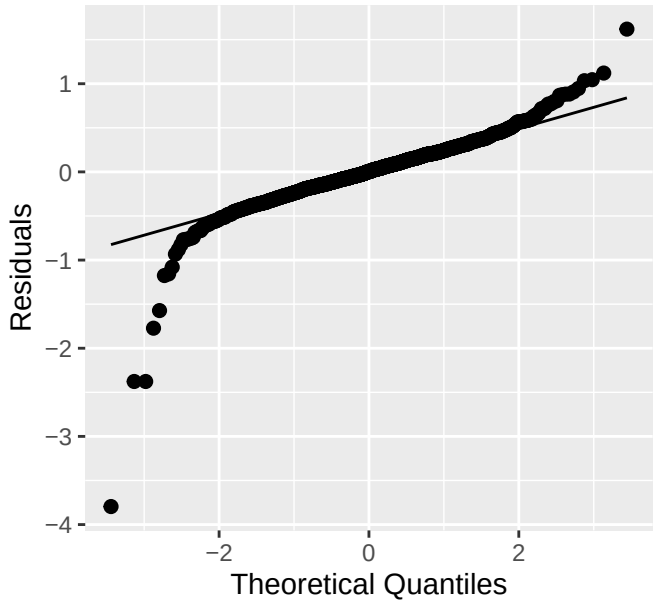


Fig 4: Normality



References. Mousina, D. (2023, April). *Econosights - Australia's housing shortage*. AMP. *https:// www.amp.com.au/ insights-hub/ blog/ investing/ econosights-australias-housing-shortage*

Packages:

Allaire, J. J., Charles Teague, Yihui Xie, and Christophe Dervieux. 2022. "Quarto." Zenodo. https://doi.org/10.5281/ZENODO.5960048.

Kuhn, Max (2008). "Building Predictive Models in R Using the caret Package." *Journal of Statistical Software*, 28(5), 1-26. doi:10.18637/jss.v028.i05, https://www.jstatsoft.org/index.php/jss/article/view/v028i05.

Team, R Core. 2022. "R: A Language and Environment for Statistical Computing." Manual. Vienna, Austria: R Foundation for Statistical Computing. https://www.r-project.org/.

Wickham, Hadley. 2016. Ggplot2: Elegant Graphics for Data Analysis. 2nd ed. 2016. Use R! Cham: Springer International Publishing : Imprint: Springer. https://cran.r-project.org/web/packages/ggplot2/index.html.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the Tidyverse." Journal of Open Source Software 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Xie, Yihui. 2022. "Knitr: A General-Purpose Package for Dynamic Report Generation in R." Manual. https://yihui.org/knitr/.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with Kable and Pipe Syntax*.

Acknowledgments. We used the package pinp (version 0.0.10) by Dirk Eddelbuettel and James Balamuta