A Bayesian Approach to Property Price Prediction.

Swathi Vangala * ,

STAT 52900 Applied Decision Theory and Bayesian Analysis
Computational Data Science

Abstract

In the realm of real estate valuation, accurately predicting property prices is a complex challenge, influenced by myriad factors such as location, area, and demographic trends.This project introduces a Bayesian approach to property price prediction in King County, enhancing traditional methods by integrating various influencing factors like location and property characteristics. Employing Bayesian linear regression with hierarchical modeling and Markov Chain Monte Carlo (MCMC) methods, our analysis robustly estimates posterior distributions using a dataset of over 21,000 property listings. Conducted using R, this method demonstrates superior predictive capabilities compared to standard linear regression, offering valuable insights for real estate valuation.

1 Introduction

In the realm of real estate, predicting property prices with accuracy is a vital yet challenging task, especially in fluctuating economic climates like a global recession. Our research, "A Bayesian Approach to Property Price Prediction," addresses this challenge by employing Bayesian hierarchical linear regression. This method stands apart for its ability to incorporate prior knowledge and hierarchical data structures, offering a robust alternative to traditional models. Focusing on King County, our study not only aims to predict property prices accurately but also identifies regions within the county with potentially higher property values.

Our approach is detailed through a structured report. The initial sections set the stage by outlining the project's objectives and reviewing relevant literature. Subsequent sections delve into our methodology, particularly the data collection process and the Bayesian model's implementation. We carefully selected prior distributions and analyzed posterior distributions, thus ensuring that our model captures the complex dynamics of the property market. The core of our analysis lies in comparing the hierarchical Bayesian model's performance with standard linear regression, demonstrating the former's superior predictive capabilities.

The final section synthesizes our findings, showcasing the Bayesian model's effectiveness in King County's property price prediction. The comparison with linear regression highlights the Bayesian model's nuanced understanding of market trends and its potential in guiding both buyers and sellers in making informed decisions.

2 Literature review

Recent studies have leveraged machine learning to enhance predictive analysis in the housing market [1,2,3]. Phan [1] highlighted price variations in Melbourne and found Stepwise regression with Support Vector Machines to be highly predictive. Similarly, Yu et al. [2] used deep learning to forecast China's housing prices, while Vineeth et al. [3] tested various algorithms, including regression models and Neural Networks, for their effectiveness in price prediction.

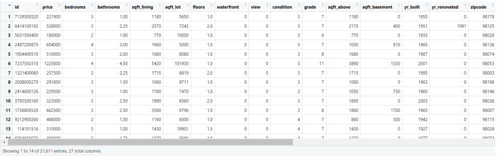
While these methods have proven useful, Bayesian hierarchical linear regression techniques stand out for their comprehensive analytical power [4,5,6]. They offer not just single model estimates but a distribution of models weighted by the data-informed likelihood of their accuracy. This Bayesian method excels by delivering an uncertainty quantification in predictions, an aspect of machine learning that traditional point estimators typically overlook [7,8,9].

Our approach, which employs Bayesian hierarchical linear regression, advances these concepts further. It transcends the capabilities of standard machine learning techniques by effectively integrating prior domain knowledge about model parameters . The hierarchical aspect of our model allows for a nuanced consideration of nested data structures, offering an intricate portrayal of market dynamics. Consequently, Bayesian hierarchical linear regression not only yields a full posterior distribution over model parameters for enhanced uncertainty quantification but also proves to be a more efficient and powerful tool when faced with complex and hierarchical data typical in real estate pricing scenarios [7,8,9].

3 Méthodology

3.1 Data Collection:

The dataset for this study was acquired from a publicly available Kaggle dataset[11], consisting of residential property sales in King County, which includes Seattle. It encompasses a comprehensive collection of 21,611 observations and 27 variables, each representing a property sale. The data captures a variety of attributes such as unique identifiers, sale date, price, number of bedrooms and bathrooms, living area square footage, lot size, the number of floors, and whether the property is waterfront, among others.



	id	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	sqft_basement	yr_built	yr_renovated	zipcode
1	7129300520	221900	3	1.00	1180	5650	1.0	0	0	3	7	1180	0	1955	0	98178
2	6414100192	538000	3	2.25	2570	7242	2.0	0	0	3	7	2170	400	1951	1991	98125
3	5631500400	180000	2	1.00	770	10000	1.0	0	0	3	6	770	0	1933	0	98028
4	2487200875	604000	4	3.00	1960	5000	1.0	0	0	5	7	1050	910	1965	0	98136
5	1954400510	510000	3	2.00	1680	8080	1.0	0	0	3	8	1680	0	1987	0	98074
6	7237550310	1225000	4	4.50	5420	101930	1.0	0	0	3	11	3890	1530	2001	0	98053
7	1321400060	257500	3	2.25	1715	6819	2.0	0	0	3	7	1715	0	1995	0	98003
8	2008000270	291850	3	1.50	1060	9711	1.0	0	0	3	7	1060	0	1963	0	98198
9	2414600126	229500	3	1.00	1780	7470	1.0	0	0	3	7	1050	730	1960	0	98146
10	3793500160	323000	3	2.50	1890	6560	2.0	0	0	3	7	1890	0	2003	0	98038
11	1736800520	662500	3	2.50	3560	9796	1.0	0	0	3	8	1860	1700	1965	0	98007
12	9212900260	468000	2	1.00	1160	6000	1.0	0	0	4	7	860	300	1942	0	98115
13	114101516	310000	3	1.00	1430	19901	1.5	0	0	4	7	1430	0	1927	0	98028
14	6054650070	400000	3	1.75	1370	9680	1.0	0	0	4	7	1370	0	1977	0	98074

Showing 1 to 14 of 21,611 entries, 27 total columns

Figure 1: Original Dataset

3.2 Data Preprocessing:

Our dataset underwent a comprehensive preprocessing regimen to ensure the quality and robustness of the subsequent analysis. The following steps were meticulously carried out:

3.2.1 Log Transformation:

To address right-skewness and bring the distribution of continuous variables closer to normal, we performed log transformations on the price and various size metrics, including square footage of living space, lot size, above-ground living area, and basement area. This transformation stabilizes variance and makes the data more suitable for the assumptions of our linear modeling approach.

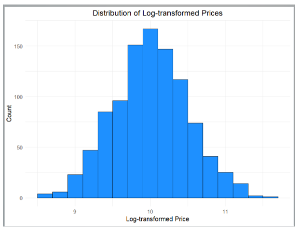


Figure 2: Histogram of Log- Transformed House Prices

3.2.2 Creation of Binary Features:

Binary variables were engineered to capture categorical property attributes. Specifically, we identified whether a property had a basement (basement) and whether it had undergone renovation (renovated). This step was crucial for capturing the impact of these features on property prices.

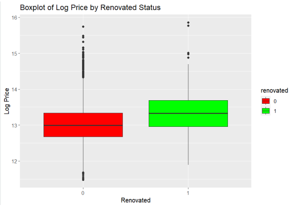


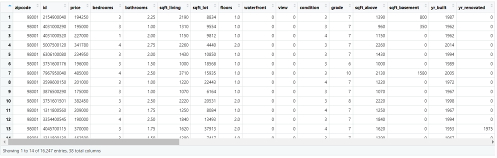
Figure 3: Log price vs Renovated

3.2.3 Demeaning of Variables:

To enhance comparability and interpretability, we centered several variables around their means (demeaning). This process involved subtracting the mean value of each variable from individual observations, thereby creating new variables that reflect deviations from the average. Demeaned variables included the number of bedrooms, bathrooms, floors, and other relevant metrics.

3.2.4 Data Cleaning:

The cleaning process resulted in a refined dataset, reducing the original 21,613 observations to 16,247 due to the removal of missing values and the exclusion of outliers and influential points. This was an important step to improve the model's accuracy and generalizability.



	zipcode	id	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	sqft_basement	yr_built	yr_renovated
1	98001	2154900040	194250	3	2.25	2190	8834	1.0	0	0	3	7	1390	800	1987	0
2	98001	4031000290	195000	3	1.00	1310	9554	1.0	0	0	3	7	960	350	1962	0
3	98001	4031000520	227000	1	2.00	1150	9812	1.0	0	0	4	7	1150	0	1962	0
4	98001	5007500120	341780	4	2.75	2260	4440	2.0	0	0	3	7	2260	0	2014	0
5	98001	6306100080	234950	3	2.00	1430	10850	1.0	0	0	3	7	1430	0	1994	0
6	98001	3751600176	196000	3	1.50	1000	18568	1.0	0	0	3	6	1000	0	1989	0
7	98001	7967950040	485000	4	2.50	3710	15935	1.0	0	0	3	10	2130	1580	2005	0
8	98001	3599600150	201000	3	1.00	1220	22443	1.0	0	0	4	7	1220	0	1972	0
9	98001	3876500290	175000	3	1.00	1070	6164	1.0	0	0	3	7	1070	0	1967	0
10	98001	3751601501	382450	3	2.50	2220	20531	2.0	0	0	3	8	2220	0	1998	0
11	98001	1311800560	209000	3	1.75	1250	8084	1.0	0	0	4	7	1250	0	1967	0
12	98001	3354400545	190000	4	2.50	1840	13493	2.0	0	0	3	7	1840	0	1994	0
13	98001	4045700115	370000	3	1.75	1620	37913	2.0	0	0	4	7	1620	0	1953	1975
14	98001	1311900130	162500	3	1.50	1390	7417	1.0	0	0	3	7	1390	0	1967	0

Showing 1 to 14 of 16,247 entries, 38 total columns

Figure 4: Cleaned Dataset

3.3 Linear Regression Modeling:

The linear regression modeling phase involved constructing a predictive model to establish the relationship between the log-transformed price of properties and a set of explanatory variables. The general form of the linear regression model we employed is:

$\log(Price) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \epsilon$

- $\log(Price)$ is the natural logarithm of the property price, ensuring a normal distribution of residuals, which is a key assumption in linear regression.
- β_0 is the intercept term, representing the expected log price when all other explanatory variables are equal to zero.
- $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients associated with each explanatory variable $X_1, X_2, \ldots, X_n$, quantifying their individual impact on the log-transformed price.
- ϵ is the error term, accounting for the variation in log price not explained by the model.

The explanatory variables included both continuous features, such as the square footage of the living area, and categorical features, like the presence of a basement or renovation status. We also included locational variables, such as zip codes, to capture the geographical influences on property prices.

To fit the linear model, we utilized the Ordinary Least Squares (OLS) method, which aims to minimize the sum of squared differences between the observed and predicted log prices.

3.4 Bayesian Hierarchical Linear Regression:

At the heart of our statistical analysis lies the Bayesian hierarchical linear regression model, which is adept at managing the complex, multilevel structure inherent in geographically-indexed data. Our model is explicitly designed to handle the natural hierarchy within the data—specifically, the properties nested within different zip codes. This sophisticated approach allows for variability among different locations, acknowledging that each zipcode carries its unique baseline property price level. y$_i = \mu + \alpha_{j[i]} + \sum_{k=1}^{K} \beta_k \cdot x_{k,i} + \epsilon_i$

- y_i is the log-transformed price of the i-th property.
- μ is the overall mean log price across all properties.
- $\alpha_{j[i]}$ is the varying intercept for each zip code j, allowing for zip code-specific deviations from the overall mean.
- β_k represents the effect size of the k-th covariate.
- $x_{k,i}$ is the value of the k-th covariate for the i-th property.
- ϵ_i is the error term associated with the i-th property, assumed to follow a normal distribution with mean 0 and variance σ^2.

Priors: We define priors for the intercepts and the regression coefficients. The priors reflect our initial beliefs about the parameters before observing the data:

Our likelihood function, which is the probability of the observed data given the parameters, is defined by the normal distribution of the residuals: $\epsilon_i \sim Normal(0, \sigma^2)$

The intercepts j for each zip code are assumed to be normally distributed,

$\alpha_j \sim Normal(0, \sigma_\alpha^2)$ expressing the belief that while different zip codes may have different average prices, they are variations around a general average.

- The regression coefficients k are also assigned normal priors, which provides regularization and helps to avoid overfitting in the presence of collinearity among the covariates.
- The priors for the standard deviations, and , are typically set to be broad and non-informative, reflecting a lack of strong prior beliefs about the scale of these parameters.

The combination of our priors and the likelihood through Bayes' theorem provides the posterior distributions for our model parameters. These posteriors are estimated using Markov Chain Monte Carlo (MCMC) methods, facilitated by the `rstan` package in R. We implemented this model in R using the `rstan` package, which interfaces seamlessly with Stan, a powerful platform for Bayesian statistical modeling and computation. For the model setup, we constructed a matrix of predictor variables XX, the response variable vector yy, and an index array to map each observation to its corresponding zipcode. The `rstan` package allows us to perform Markov Chain Monte Carlo (MCMC) sampling, facilitating the estimation of the posterior distributions of our model parameters.

The application of a Bayesian hierarchical framework not only provides us with estimates of the fixed effects but also quantifies the uncertainty associated with these estimates, yielding a more nuanced interpretation of the results. This is particularly advantageous in real estate market analysis, where understanding the variability and confidence in our predictions is as critical as the predictions themselves.

3.5 Model Estimation and Diagnostics:

Our Bayesian hierarchical linear regression model's estimation was meticulously performed using Markov Chain Monte Carlo (MCMC) sampling, with visual diagnostics playing a crucial role in evaluating the model's performance. The estimation process involved analyzing several chains to ensure a comprehensive exploration of the parameter space.

3.5.1 Estimation Methodology:

Our Bayesian hierarchical model was estimated using advanced MCMC algorithms facilitated by the rstan interface for the Stan programming language. Stan utilizes the No-U-Turn Sampler (NUTS), an adaptive form of Hamiltonian Monte Carlo (HMC), which efficiently navigates the complex parameter space of hierarchical models. This choice is motivated by HMC's superior performance over traditional MCMC methods, such as Gibbs sampling or Metropolis-Hastings, particularly in dealing with the high-dimensional parameter spaces typical of hierarchical models.

3.5.2 Prior Distributions:

We assigned non-informative, broad priors to most parameters to minimize their influence on the posterior. Specifically, normal priors with a mean of zero and a large variance were used for the regression coefficients, reflecting a lack of strong prior belief regarding their values.

3.5.3 Convergence Criteria:

We assessed convergence of the MCMC chains using several diagnostics. The Gelman-Rubin statistic was computed for each parameter, with values approaching 1 indicating convergence. Additionally, we inspected trace plots to visually confirm that the chains had mixed well and were sampling from the entire posterior distribution. A burn-in period was determined empirically, and we used thinning to reduce autocorrelation in the samples, ensuring independent draws from the posterior.

3.5.4 Trace Plots:

In our project, The trace plots for the regression coefficients depicted in the figure[5] provided substantial insight. Each trace plot represents the sampling paths of individual beta coefficients across iterations for two separate chains. The overlapping and dense nature of the plots suggests good mixing and convergence, with the chains appearing to stabilize around certain values, indicating reliable estimates.

In addition to trace plots, the figure[6] for adjusted coefficients similarly indicate convergence, as demonstrated by the overlapping patterns for different chains. This convergence is critical because it implies that our posterior samples are representative of the true posterior distribution.

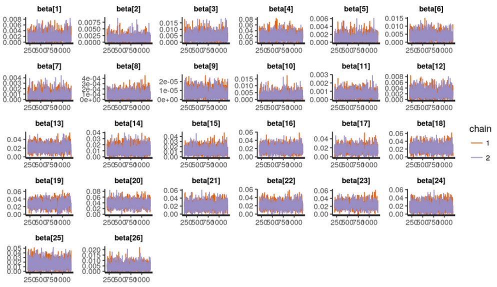


Figure 5: MCMC Trace plots for Model Coefficients

3.5.5 Residual Analysis:

Residual analysis involves examining the differences between observed values and those predicted by the model. In a well-specified model, we expect residuals to be randomly distributed with no discernible pattern when plotted against fitted values or any other variable. Systematic patterns in these plots can indicate potential issues with the model such as non-linearity, heteroscedasticity, or omitted variables.

Figure 6: Trace plots of Random Effects for Hierarchical Model

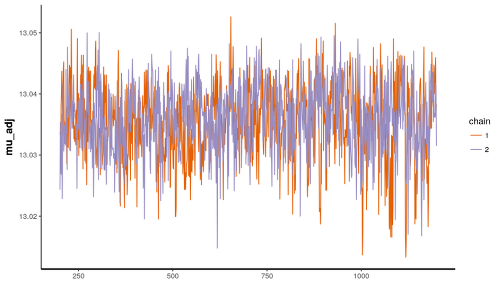


Figure 7: MCMC Sampling Trace plot for Overall Intercept Adjustment

For our Bayesian hierarchical linear regression model, we plotted residuals against fitted valuesin Figure [8] to assess model adequacy. We looked for random scatter in these plots as evidence that the model was capturing the data's underlying structure effectively. Residual plots did not reveal any obvious patterns, such as clear curves or fan shapes, suggesting that the model did not suffer from obvious misspecification. This random distribution of residuals supports the assumption that our model adequately captures the relationships between the predictors and the response variable.

.

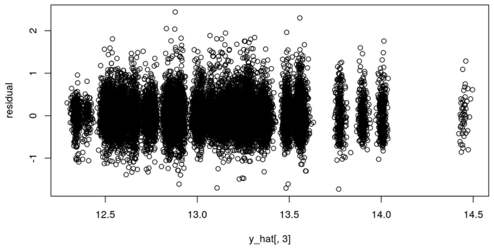


Figure 8: Residuals vs. Predicted Values

4 Results

4.0.1 Random Effect

In the random effects plot as shown in Figure[9] for our model, the x-axis represents different zip codes as categorical variables, while the y-axis, labeled 'Chain[1]', displays the estimated random effect for each zip code. Positive values on the y-axis suggest that a zip code's average log price is above the overall mean, and negative values indicate the opposite. The height and direction of each bar in the plot signify the magnitude and direction of each zip code's deviation from the overall mean, with taller bars indicating greater deviations. This variation in bar heights highlights the diversity in property price levels across different zip codes and underscores the importance of including random effects in our model. These effects capture local-level variations in property prices, offering insights that fixed effects might not fully reveal.

The presence of both positive and negative random effects illustrates that some areas have higher or lower property values than might be expected based on the overall average. This insight can inform targeted strategies for real estate investment, development, and policy-making. The plot also serves as a diagnostic tool, ensuring that the model accounts for the nested data structure and providing a check against over-simplified models that assume homogeneity across all geographic areas.

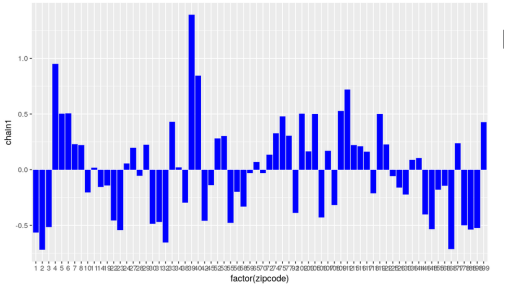


Figure 9: Bar Plot of Random Effects for Zip Codes

4.0.2 Linear Regression Residuals:

The diagnostic plots from the linear regression model reveal patterns in the Residuals vs Fitted plot figure [10], suggesting potential heteroscedasticity or non-linearity, and deviations from normality in the Q-Q plot. These signs indicate that the assumptions of linear regression are not adequately met. Consequently, we opted for a hierarchical

Bayesian model, which better accommodates the complexity of the data through random effects and prior information, thus providing a more suitable and nuanced analytical approach.

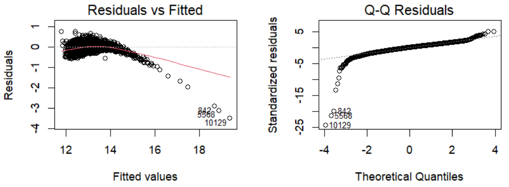


Figure 10: Residuals vs Fitted

4.0.3 Model Performance and R-square Value

In assessing the performance of our Bayesian hierarchical linear regression model, a key metric of interest is the R2 value, also known as the coefficient of determination. For our model, the R2R2 value was calculated to be approximately 0.582. This suggests that about 58.2% of the variance in property prices within King County is accounted for by the model.

This level of explanatory power is significant, particularly in the complex and multifaceted domain of real estate valuation. It indicates that a majority of the fluctuations in property prices can be attributed to the variables and hierarchical structure incorporated in our model. This finding underscores the model's capacity to effectively capture and quantify the key drivers of real estate prices in the region.

4.1 Conclusion:

In conclusion, the transition from traditional linear regression to a hierarchical Bayesian model was necessitated by significant deviations from key linear regression assumptions, including heteroscedasticity and non-normality of residuals. The hierarchical Bayesian model proficiently rectified these issues, utilizing the data's hierarchical nature with zip codes as crucial random effects to capture distinct local market variations. This model not only improved adaptability and precision in prediction but also deepened the understanding of various factors influencing property prices. A notable aspect of this model's efficacy is its R2 value of approximately 0.582, indicating that it explains around 58.2% of the variance in property prices, a significant achievement given the complexities of real estate valuation. By integrating each location's specific characteristics, the model provides more accurate and locally tailored price predictions, demonstrating the effectiveness and versatility of Bayesian methods in analyzing intricate data landscapes.

References

[1] L. Yu, C. Jiao, H. Xin, Y. Wang, and K. Wang, " Prediction on Housing Price Based on Deep Learning," World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering, vol. 12, pp. 90–99, 2018

[2] N. Vineeth, M. Ayyappa, and B. Varadharajulu, "House Price Prediction Using Machine Learning Algorithms," in International Conference on Soft Computing Systems, Apr. 2018, pp. 425–433

[3] M. West, "Outlier Models and Prior Distributions in Bayesian Linear Regression," Journal of the Royal Statistical Society. Series B (Methodological), vol. 46, no. 3, pp. 431–439, 1984.

[4] . J. Mitchell and J. J. Beauchamp, "Bayesian variable selection in linear regression. Journal of the american statistical association," J Am Stat Assoc, vol. 83, no. 404, pp. 1023–1032, 1988.

[5] D. J. C. MacKay, "Bayesian nonlinear modeling for the prediction competition," ASHRAE Trans, vol. 100, no. 2, pp. 1053–1062, 1994

[6] A. E. Raftery, D. Madigan, and J. A. Hoeting, "Bayesian model averaging for linear regression models," J Am Stat Assoc, vol. 92, no. 437, pp. 179–191, Mar. 1997

[7] T. P. Minka, "Bayesian linear regression," Technical Report,MIT. 2000.

[8] S. M. Lynch, Introduction to Applied Bayesian Statistics and Estimation for Social Scientists, vol. 1. New York: Springer, 2001.

[9] P. Pérez, G. de Los Campos, J. Crossa, and D. Gianola, "Genomic-Enabled Prediction Based on Molecular Markers and Pedigree Using the Bayesian Linear Regression Package in R," Plant Genome, vol. 3, no. 2, 2010.

[10] https://www.kaggle.com/datasets