A Bayesian Approach to Property Price Prediction.

Swathi Vangala

Overview

- Introduction
- Objective
- Data Synopsis
- Methodology
- Model Design
- Results
- Conclusion
- Future Work

Introduction

- House price evaluation is crucial in real estate as it informs pricing negotiations and strategic decisions for companies, financial institutions, and investors.
- Indeed, driven by the strong business needs, many statistical models have been proposed for house price evaluation in the past few years.
- The Bayesian approach enhances house price predictions by incorporating prior knowledge and uncertainty into the model, allowing for more nuanced and potentially more accurate estimations than traditional regression methods.

Objectives

- Deploy advanced Bayesian methodologies to enhance property price prediction precision.
- Extract deeper probabilistic insights from housing data for nuanced property price inferences.
- Effectively accommodate the inherent uncertainties and varied price dynamics across different neighborhoods.

Data Synopsis

- Source: Kaggle - https://www.kaggle.com/datasets
- Original Dataset Size: 21614 properties with 27 features each.
- Primary Features: Bedrooms, bathrooms, square footage, lot size, location (zipcode, latitude, longitude), sale date, construction and renovation years.

id	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	sqft_basement	yr_built	yr_renovated	zipcode	lat	long
7129300520	221900	3	1	1180	5650	1	0	0	3	7	1180	0	1955	0	98178	47.5112	-122.257
6414100192	538000	3	2.25	2570	7242	2	0	0	3	7	2170	400	1951	1991	98125	47.721	-122.319
5631500400	180000	2	1	770	10000	1	0	0	3	6	770	0	1933	0	98028	47.7379	-122.233
2487200875	604000	4	3	1960	5000	1	0	0	5	7	1050	910	1965	0	98136	47.5208	-122.393
1954400510	510000	3	2	1680	8080	1	0	0	3	8	1680	0	1987	0	98074	47.6168	-122.045

Data Synopsis

- Data Preprocessing:
  - Log Transformed: Price and size metrics for normal distribution.
  - Binary Features: Basement presence and renovation status.
  - Demeaned Variables: Centered around mean for comparability.

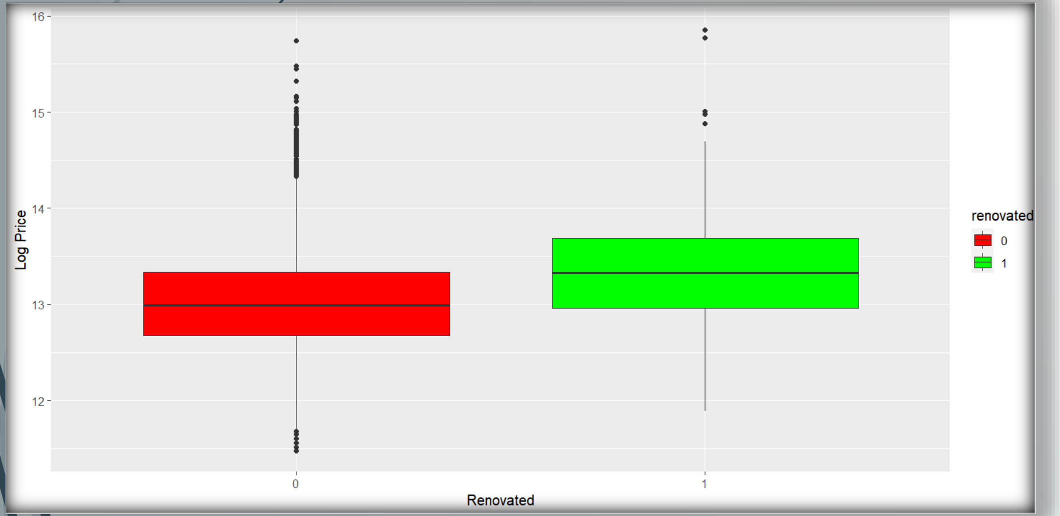

Fig 1: Log price vs Renovated

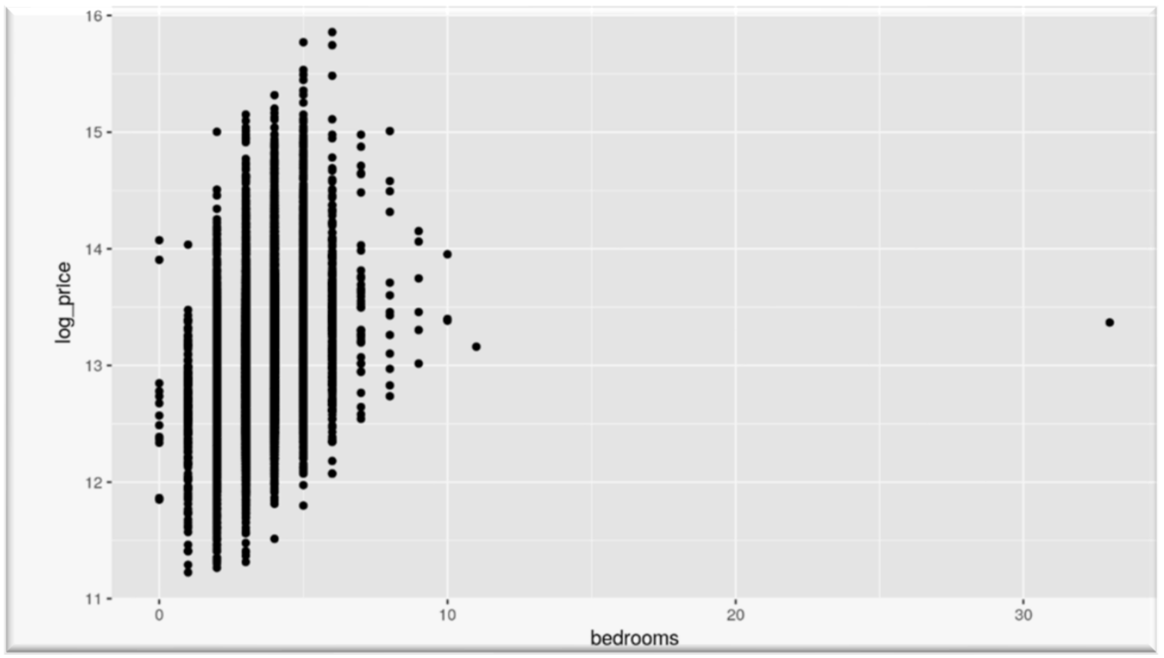

Fig 2: Log price vs Bedrooms

Data Synopsis

- Cleaned Dataset size : 16247 observations with 38 variables each

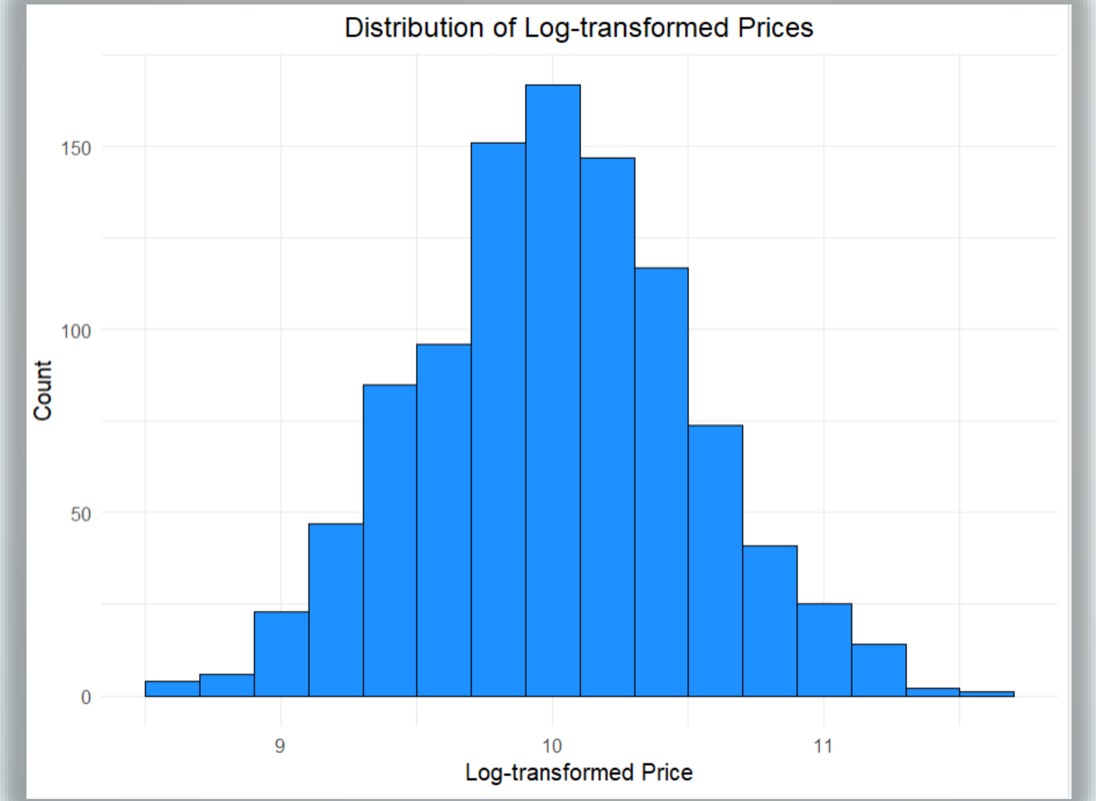

Fig 3: Histogram of Log-Transformed House Prices"

Methodology

- Employ Bayesian Hierarchical linear regression model, adjusting for the influence of key features on property prices.

- Use MCMC (Markov Chain Monte Carlo) methods for posterior distribution estimation.

Model Design: Hierarchical Linear Regression Model

- Priors:

$$\epsilon_i \sim \text{Normal}(0, \sigma^2)$$
$$\alpha_j \sim \text{Normal}(0, \sigma_\alpha^2)$$

- Model Equation:

$$y_i = \mu + \alpha_{j[i]} + \sum_{k=1}^{K} \beta_k * x_{k,i} + \epsilon_i$$

Where j represents different zip codes, i is the row index, k is the index for each covariate.

Results : Trace plots

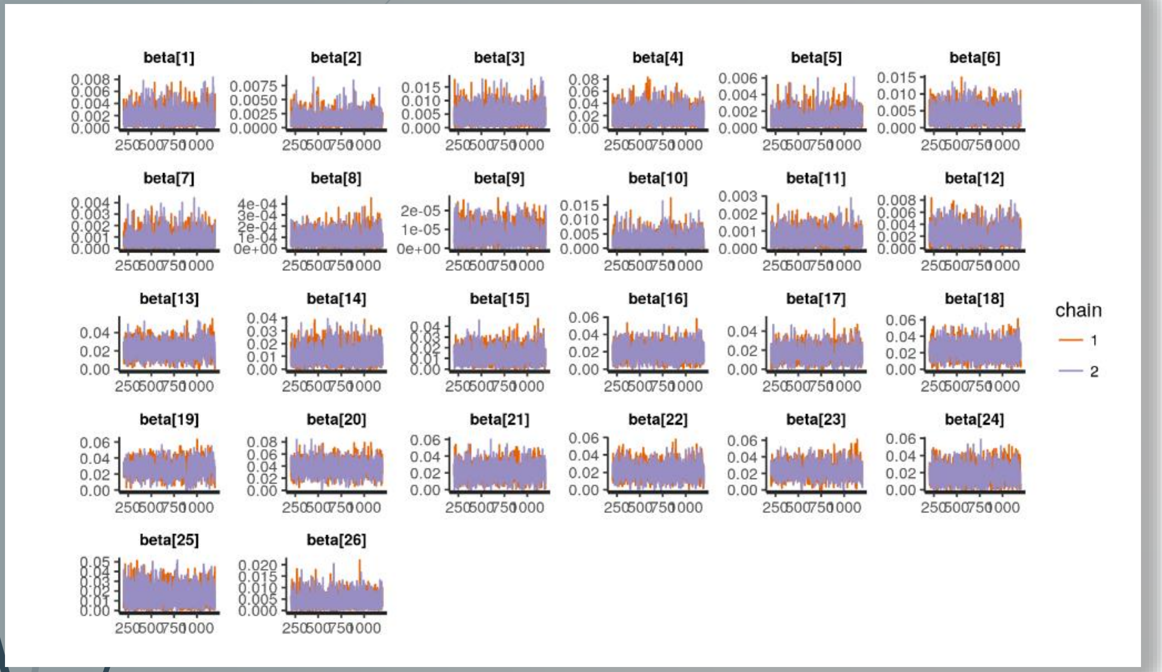

Fig 5: MCMC Trace plots for Model Coefficients

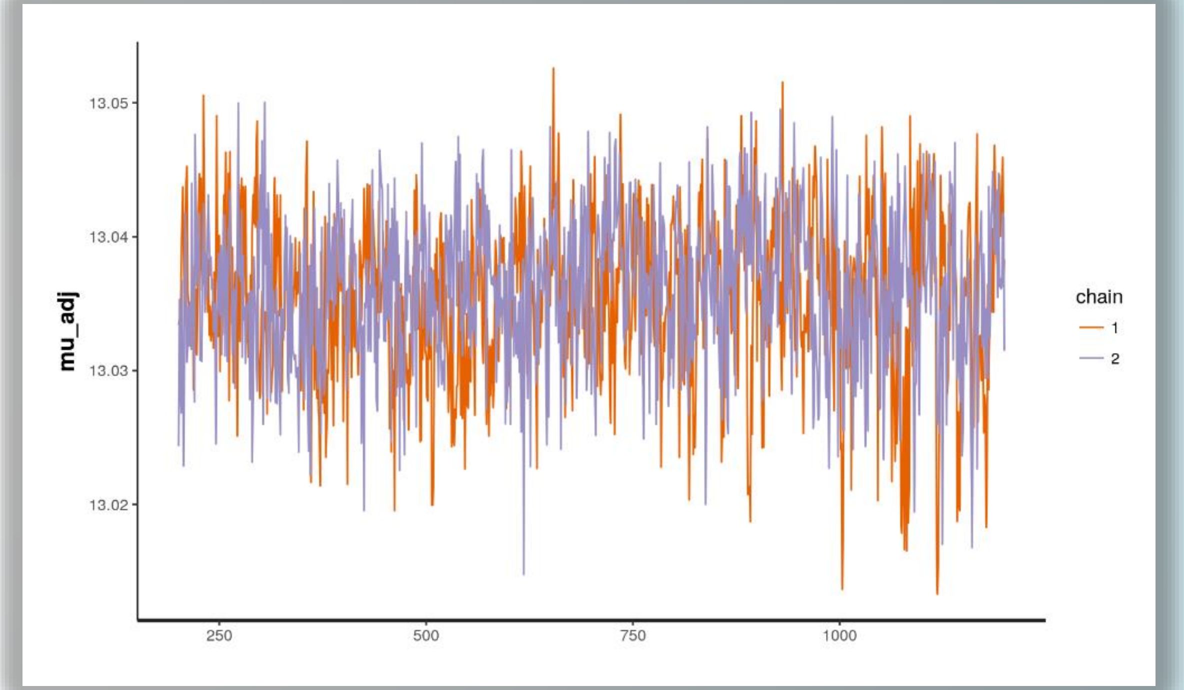

Fig 6: MCMC Sampling Trace plot for Overall Intercept Adjustment

Results : Trace plots

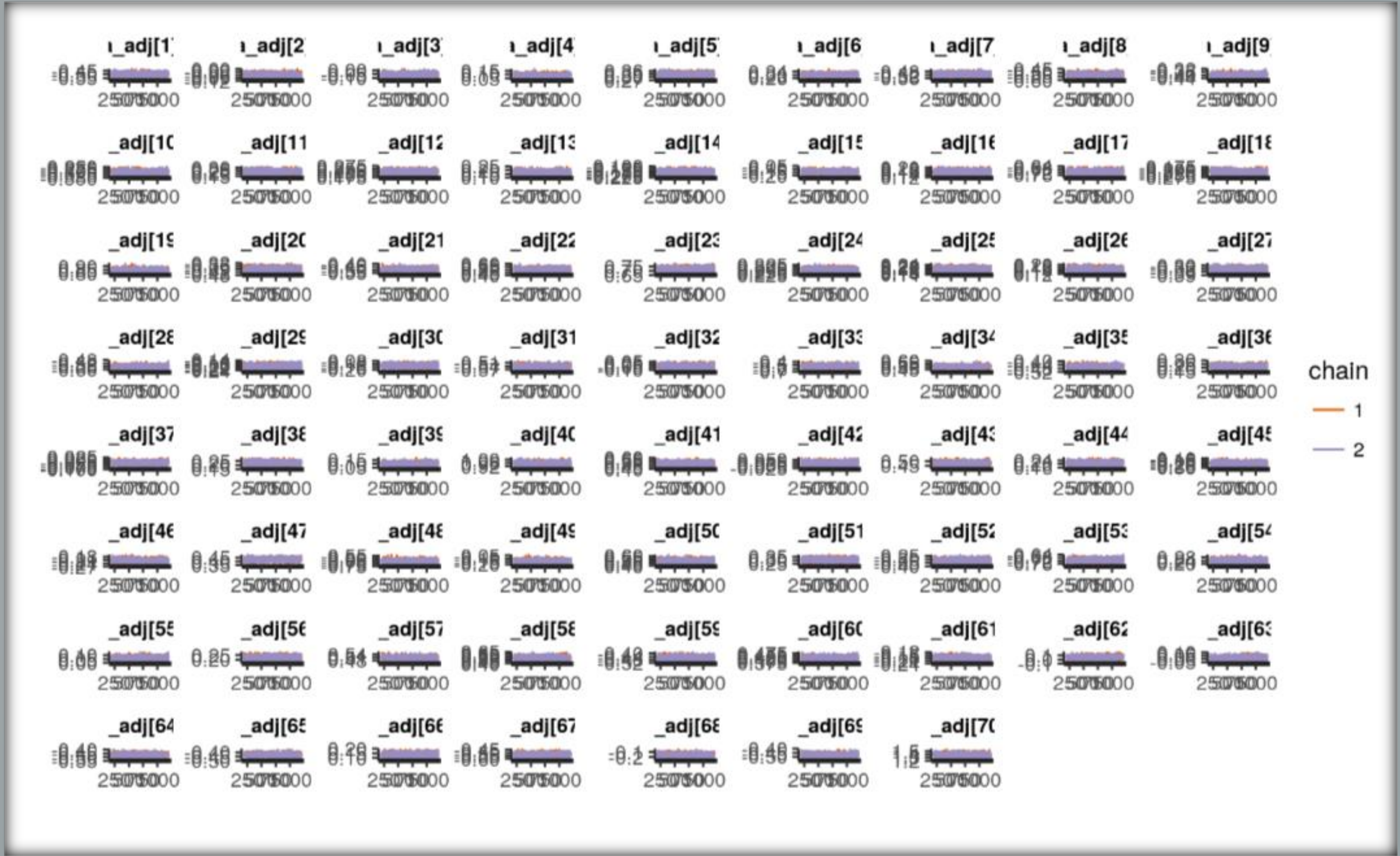

Fig 7: Trace plots of Random Effects for Hierarchical Model

Results: Residual plots

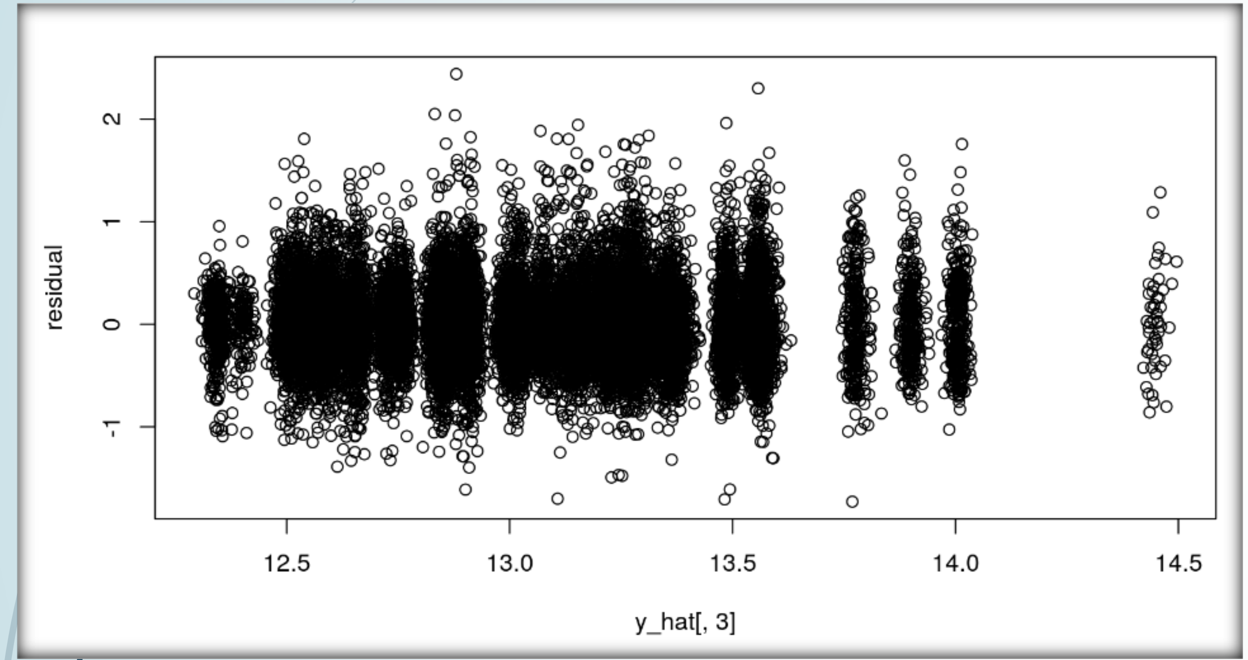

Fig 8: Residuals vs. Predicted Values

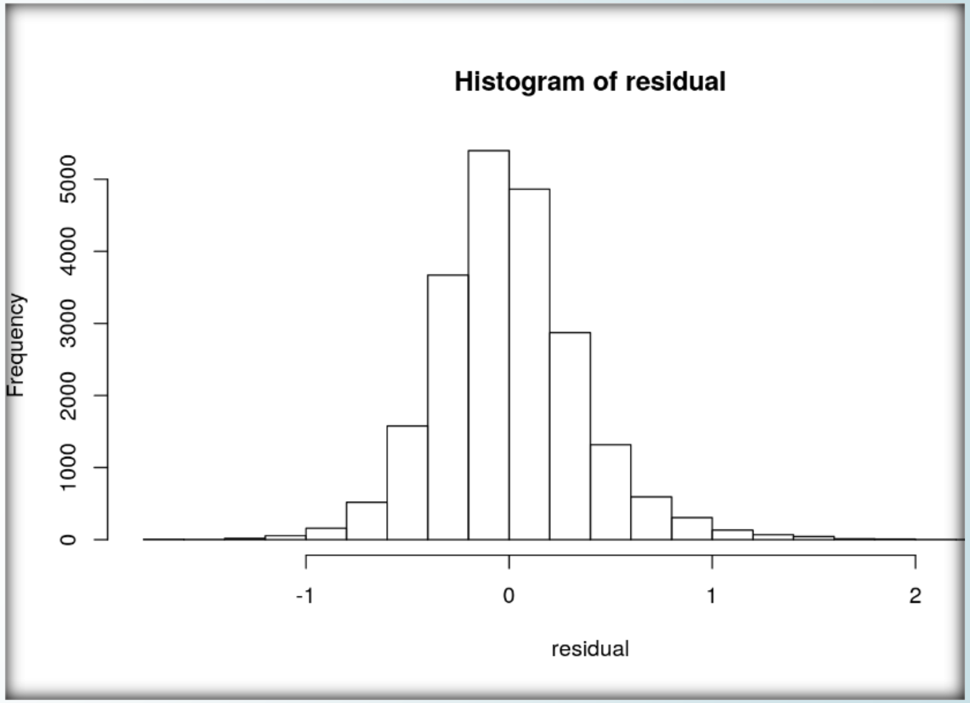

Fig 9: Histogram of Model Residuals

Results: Random effect

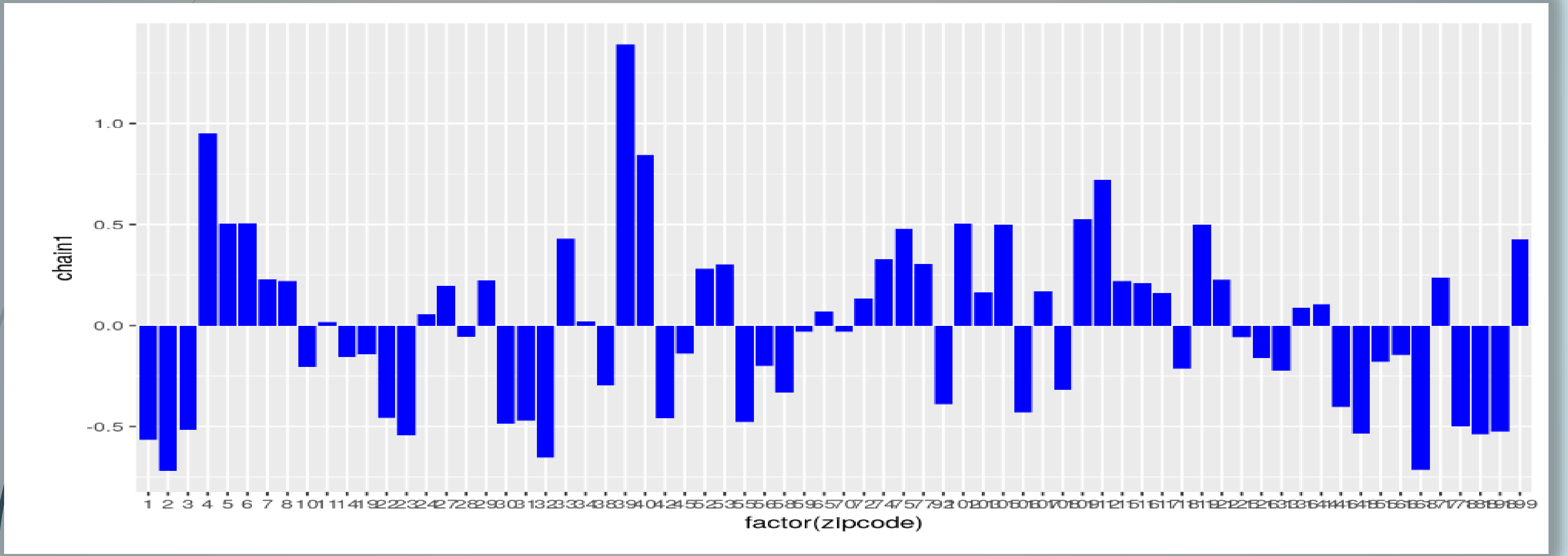

Fig 10: Bar Plot of Random Effects for Zip Codes

Linear regression Model:

Residuals vs Fitted shows a slight pattern with the residuals curving, which could indicate that the model is not capturing some non-linear effects.

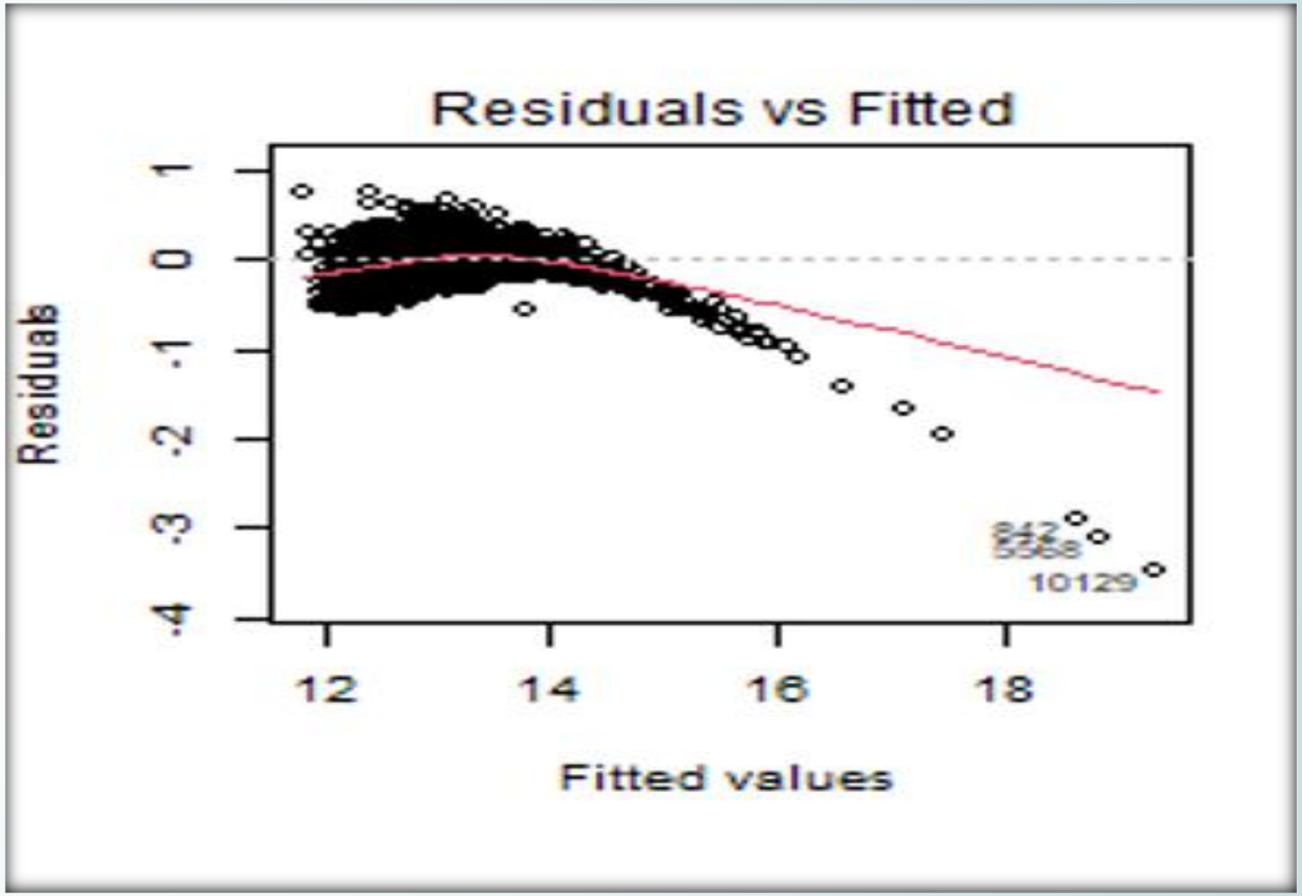

Fig 11: Residuals vs Fitted

Limitations

- Hierarchical models, especially Bayesian ones, are computationally intensive and require more processing power and time.
- Hierarchical models can overfit the data, especially when there are many parameters relative to the amount of data.
- While the hierarchical structure is designed to model nested data, it might not capture all levels of interaction or might be too complex for the available data.
- The residual plots indicated room for improvement, which could mean the current model does not capture all the underlying patterns in the data.

Conclusion

- The Bayesian hierarchical model identified square footage, number of bedrooms and bathrooms, and renovation status as key drivers of house prices, with significant price variations across zip codes.
- With an R^2 value of 0.58 , the model demonstrates an ability to explain the variance in house prices.
- The diagnostic plots indicate the model captures group-level variations well, though some outliers and patterns in residuals suggest room for improvement.
- Although the primary model did not show particularly strong predictive capabilities, it still provides valuable insights into a different research query: it examines the variation in house prices across various zip codes when accounting for other variables.

Future work

- Incorporate additional data points, potentially from different regions or time periods, to increase the robustness and generalizability of the model.
- Include new variables that may influence house prices, such as economic indicators, crime rates, school district quality, or public infrastructure.
- Experiment with different hierarchical structures to improve model fit and predictive accuracy.

Thank You