Statistics and Probability Final Project
Group 26

Università Bocconi | January 11, 2025
Members: Caretti Giorgio Filippo (3160916), Lorenzana Luca Garcia (3161033), Milani Luca (3160856), Muraro Margherita (3164644).

1 Introduction

Understanding the factors driving housing prices is essential for effective economic planning and fostering social well-being. Housing prices reflect not only the physical attributes of individual properties but also the broader regional and administrative dynamics that shape urban markets. This study utilizes a comprehensive dataset of housing prices in Melbourne, spanning from January 2016 to September 2017. The dataset includes key property characteristics such as size, type, and distance from the central business district (CBD), as well as administrative divisions across 31 councils and 8 regions in the Greater Melbourne area. The objective of our study is to investigate the variability in housing prices across regions and council areas, accounting for property characteristics, and to evaluate the prediction power of the model developed.

2 EDA

We begin by exploring the distribution of housing prices (`Price`), which exhibits a highly skewed distribution with the majority of properties priced below $2 million. To address this skewness and variance instability, a log transformation is applied to `Price`, resulting in a more normalized distribution. Examining the relationships between `LogPrice` and key predictors reveals several important patterns that inform model design. Scatterplots indicate a positive relationship between `LogPrice` and the number of `Rooms`, though the marginal relationship diminishes beyond four rooms. This non-linear trend likely reflects the fact that larger houses are often located farther from the Central Business District (CBD), where housing prices tend to be lower. Boxplots of `LogPrice` by `Type` reveal that houses and villas (`Type = h`) command higher prices compared to townhouses and units 1. This variability is likely driven by structural features like number of rooms, which also exhibit greater heterogeneity in these property types. Consequently, an interaction term between `Rooms` and `Type` was introduced in the model to capture this relationship 7.

The number of car spaces (`Car`) is another important predictor. Scatterplots show a positive relationship between `LogPrice` and car spaces, with diminishing marginal effects beyond four spaces 2. Interestingly, properties with no car spaces still show high prices, whose scatterplots suggest that they are concentrated near the CBD. These observations informed the inclusion of an interaction term between `Car` and `Distance` in the model, capturing the differing impacts of car spaces based on proximity to the city center. Structural attributes such as the number of rooms and car spaces show significant relationships with housing prices, consistent with the findings of Sirmans et al. [2005], who emphasize the role of such features in hedonic pricing models.

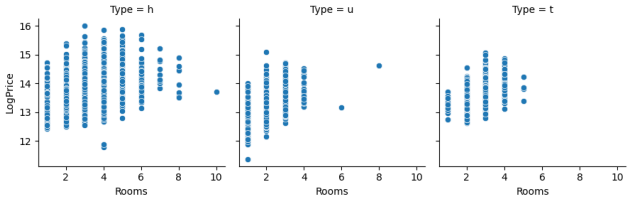

Figure 1: Scatterplots `LogPrice` and `Rooms` per `Type`.

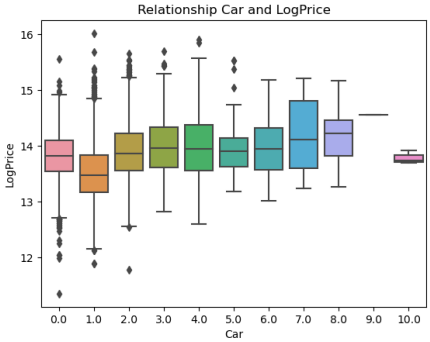

Figure 2: Boxplots `LogPrice` and `Car`.

Spatial analysis reveals significant variability across council areas and regions. Some council areas, such as *Boroondara*, exhibit a high volume of observations, high median prices, and greater price variability, highlighting their status as hot real estate markets 3. Similarly, *Southern Metropolitan*, as the region with the most transactions and diverse property types, shows the highest variability. These patterns justify the inclusion of random effects for council areas and regions in the model to account for unobserved spatial influences on housing prices 4. This aligns with Cheshire and Sheppard [1995], who discuss the influence of location-specific amenities and regional differences on property values.

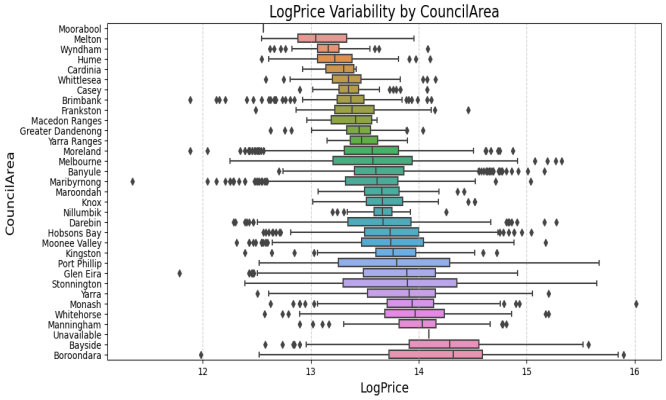

Figure 3: Boxplots `LogPrice` by `CouncilArea`.

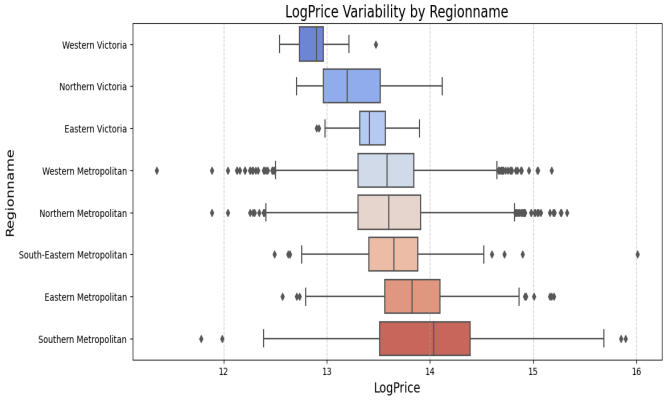

Figure 4: Boxplots `LogPrice` by `Regionname`.

Temporal trends in the dataset highlight seasonal variations in transaction prices 5 and volumes 6. Time-series plots reveal peaks in transaction volumes during *November 2016*, *May 2017*, and *July 2017*, while boxplots indicate that *December* exhibits the highest median prices during the observed period. These findings support the use of random effects for months and the inclusion of a seasonality variable in the model, allowing it to capture cyclical trends effectively.

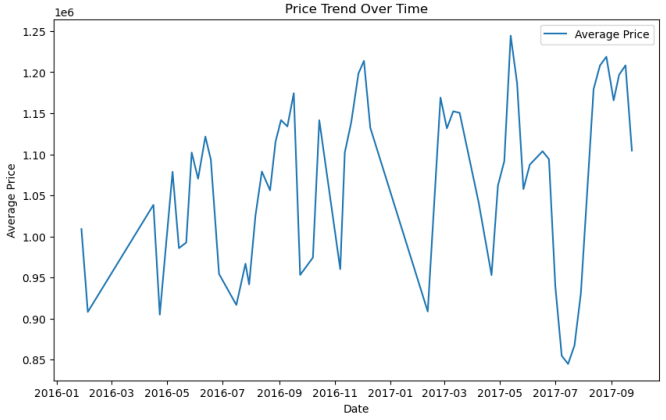

Figure 5: Average `Price` over `Month`.

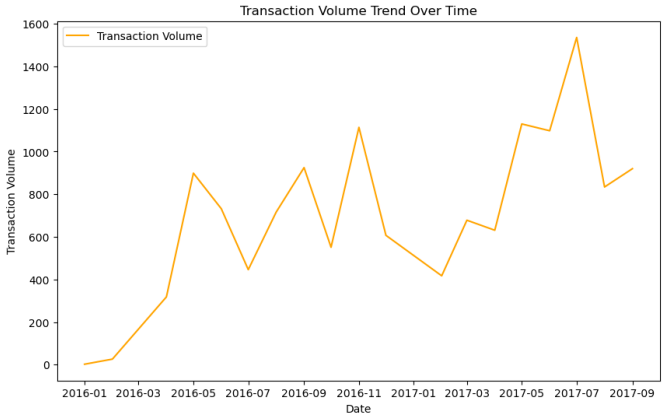

Figure 6: Number of deals over `Month`.

Finally, the combined findings from spatial, temporal, and structural analyses underscore the importance of incorporating interaction terms, random effects, and log transformation into the hierarchical Bayesian model.

3 Model

This section uses a Bayesian modeling approach with Markov Chain Monte Carlo (MCMC) methods to predict housing prices in Melbourne, accounting for various fixed and random effects. In the context of predicting housing prices, MCMC is particularly advantageous for modeling uncertainty, incorporating prior knowledge, and estimating parameters within a hierarchical framework.

The analysis began with comprehensive data cleaning and feature engineering to prepare the dataset for modeling. Missing values in the `CouncilArea` column were imputed based on mappings from the `Postcode`,

ensuring accurate alignment of council areas with geographic locations. Other numerical variables were filled with their respective medians. Outliers in the `Price` variable were identified and removed using the Interquartile Range (IQR) method to ensure that extreme values did not distort the analysis or bias the results, thereby improving the reliability and accuracy of the model. Finally, `LandSize` and `BuildingArea` were removed from the analysis because, despite their theoretical importance in predicting housing prices, they had inconsistent and incorrect values in this dataset. Including them would have added noise without increasing the model's statistical power.

The `Date` column was converted into a proper date type, enabling the creation of `Year` and `Month` columns for temporal analysis. Dummy variables were created for `Type`, reflecting property categories such as houses, townhouses, and units. Categorical variables such as `CouncilArea`, `Regionname`, `Month`, and `Method` were converted into factors. A `Seasonality` variable was constructed using sine and cosine transformations of the `Month` to capture cyclical trends, while `Price` was log-transformed to create `LogPrice`, addressing variance instability and skewness. The dataset was then split into training and test sets using an 80/20 ratio to ensure that the model was trained on the majority of the data while reserving a portion for evaluation.

After preparing the dataset, a Bayesian hierarchical model was specified. The likelihood of the observed log-transformed price, LogPrice_i, for observation i, was modeled as:

$$\text{LogPrice}_i | \mu_i, \sigma^2 \sim N(\mu_i, \sigma^2),$$

where μ_i is the mean log-price and σ^2 is the residual variance.

The linear predictor for μ_i included both fixed and random effects:

$$\text{LogPrice}_{ij} = \beta_0 + \sum_k \beta_k X_{ijk} + u_{\text{CouncilArea}} + u_{\text{Regionname}} + u_{\text{Month}} + \epsilon_{ij},$$

where:

- β_0 is the intercept.
- $\sum_k \beta_k X_{ijk}$ represents the fixed effects, with β_k as coefficients for predictors X_k. These predictors included variables such as `Rooms`, `Car`, `Distance` from the Central Business District (CBD), `Seasonality`, `YearBuilt`, and `Method`, as well as interaction terms like `Rooms` × `Type` and `Car` × `Distance`.
- $u_{\text{CouncilArea}}$, $u_{\text{Regionname}}$, u_{Month} are random effects, capturing unobserved spatial and temporal variability. These are modeled as $N(0, \sigma_k^2)$, with mean 0 and variance σ_k^2.
- ϵ_{ij} is the residual error term, accounting for unexplained variability.

The priors were weakly informative:

$$\beta_j \sim N(0, \tau^2), \quad \text{with large } \tau^2 \text{ (e.g., 10,000).}$$

Variance parameters followed an Inverse-Gamma prior:

$$\sigma^2 \sim \text{Inverse-Gamma}(\nu, \lambda), \quad \text{with } \nu = 0.002 \text{ and } \lambda = 0.002.$$

The MCMC sampler was implemented, running for 50,000 iterations with a burn-in of 10,000 and thinning applied (retaining every 20th sample). Model updates included:

- **Metropolis Update for Fixed Effects (β):** Proposals were generated from a Normal distribution centered at the current value, and acceptance ratios were computed.
- **Gibbs Update for Random Effects:** Random effects were updated based on their conditional posterior distributions.
- **Gibbs Update for Variances:** Variances ($\sigma^2, \sigma_s^2, \sigma_r^2, \sigma_m^2$) were updated using conjugate Inverse-Gamma posterior distributions.

4 Results

The model revealed key insights. Among the fixed effects, structural features such as `Rooms` ($\beta = 0.148$, this indicates that a one-unit increase in `Rooms` is associated, on average, with a relative increase of $e^{0.148} - 1 \approx 15.96\%$ in the predicted price) and `Car` ($\beta = 0.058$) positively influenced housing prices. Distance from the CBD had a negative effect ($\beta = -0.030$). Property type was significant, with townhouses ($\beta = -0.410$) and units ($\beta = -1.082$) reducing log-prices. Interaction terms, such as `Rooms` $\times$ `Type` ($\beta = 0.254$), suggested that additional rooms in units have a relatively higher impact on price compared to houses 7. Random effects captured spatial variability through $u_{\text{CouncilArea}}$ ($\sigma_s^2 = 0.075$) 8 and u_{region} ($\sigma_r^2 = 0.0053$). Temporal variability was relatively minor ($\sigma_m^2 = 0.0005$).

The model's performance was assessed using predictions on the test set. On the log-scale, the Root Mean Squared Error (RMSE) was 0.234, indicating accurate predictions of the log-transformed housing prices. Residual standard deviation (σ) was 0.241, demonstrating that the model captures most of the variability in housing prices. The similarity between these two values indicates that the model's predictions are well-calibrated and exhibit minimal systematic bias.

On the original price scale, RMSE was \$252,217, reflecting the average squared deviation of predictions from actual prices. Mean Absolute Error (MAE) was \$177,196, which represents the average absolute difference between observed and predicted prices. Unlike RMSE, which is more sensitive to large outliers, MAE provides a complementary perspective by focusing on the magnitude of prediction errors regardless of direction. Together, these metrics underscore the model's strong predictive power, effectively capturing the variability in housing prices.

Parameter	Mean	Lower95	Upper95
(Intercept)	15.97	15.59	16.34
Rooms	0.15	0.14	0.16
Typet	-0.41	-0.48	-0.34
Typeu	-1.08	-1.12	-1.04
Car	0.06	0.05	0.07
Distance	-0.03	-0.03	-0.03
Seasonality	0.00	-0.01	0.02
YearBuilt	-0.00	-0.00	-0.00
MethodS	0.07	0.06	0.09
MethodSA	0.04	-0.02	0.10
MethodSP	0.04	0.02	0.06
MethodVB	-0.01	-0.03	0.02
Rooms:Typet	0.08	0.05	0.10
Rooms:Typeu	0.25	0.24	0.27
Car:Distance	-0.00	-0.00	-0.00

Figure 7: Posterior Summary of Fixed Effects

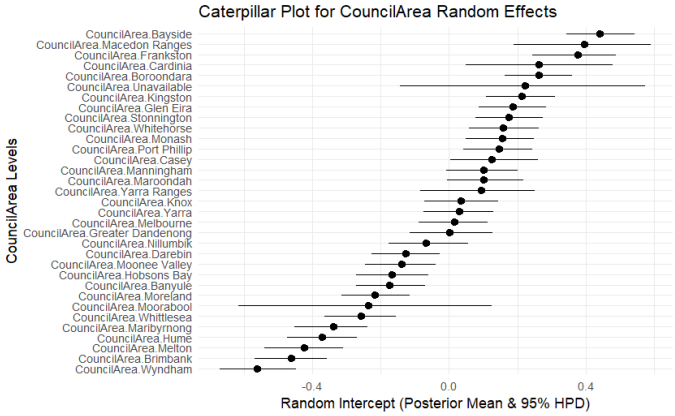

Figure 8: Caterpillar plot for CouncilArea random effect

5 Conclusion and Discussion

This study underscores the multifaceted nature of housing price dynamics in Melbourne, effectively capturing the interplay between spatial, structural, and temporal factors using a Bayesian hierarchical framework with MCMC. The model highlights key predictors, such as the positive influence of structural features like the number of rooms ($\beta = 0.15$) and car spaces ($\beta = 0.06$) and the negative effect of distance from the CBD ($\beta = -0.03$), while interaction terms reveal nuanced relationships, including the more significant impact on the price of rooms in units compared to houses ($\beta = 0.254$). Random effects for council areas, regions, and months further underline the critical role of spatial and temporal variability in heterogeneous markets.

The model demonstrated robust predictive accuracy with a log-scale RMSE of 0.234 and a price-scale MAE of \$177,196, indicating its effectiveness in estimating housing prices. These insights align with findings from Cheshire and Sheppard [1995] on location-specific amenities and Sirmans et al. [2005] on structural features' role in pricing. Future research could enhance forecasting by integrating additional covariates, refining priors, or applying this framework to other regions.

References

Paul Cheshire and Stephen Sheppard. On the price of land and the value of amenities. *Economica*, 62(246): 247–267, 1995. URL https://doi.org/10.2307/2554906. Accessed 2 Jan. 2025.

G. Stacy Sirmans, David A. Macpherson, and Emily N. Zietz. The composition of hedonic pricing models. *Journal of Real Estate Literature*, 13(1):3–43, 2005. URL http://www.jstor.org/stable/44103506. Accessed 2 Jan. 2025.