# 20592 - Statistics and Probability: Final Project

**Group 10**
Bin Arshad Muhammad, Guler Zeynep, Kalak Utku, Karahan Arzum

## 1 Introduction

This report presents a comprehensive Bayesian regression analysis to explore the relationship between features in the given housing dataset, particularly focusing on the calculated dependent variable *Price per square meter*. The analysis involves exploratory data visualization, model diagnostics, and an evolution of goodness-of-fit measures to assess model performance and reliability.

## 2 Methodology

### 2.1 Data Filtering and Cleaning

To ensure a robust and clean dataset, we removed outliers and irrelevant data points to enhance the accuracy of the subsequent machine learning models. For the *BuildingArea* column, we identified and removed the extreme outliers using the Interquartile Range (IQR) method within each property type group, accounting for variability across property categories. We applied the same procedure to the *Landsize* column as well. Additionally, for *YearBuilt*, we excluded the entries before 1900 to focus on modern housing data.

Regarding missing values, we used a two-step approach. First, we filled *YearBuilt* and *CouncilArea* based on the mode within the same latitude and longitude coordinates. If no direct match was found, we imputed values using data from nearby properties within a small proximity threshold, ensuring consistency and accuracy. Furthermore, we retained the rows with missing values to prevent unintended data loss, addressing these values later using the ML algorithm for imputation. For the remaining missing data in the columns *BuildingArea*, *YearBuilt*, and *Car*, we applied the K-Nearest Neighbors (KNN) algorithm, grouping the data by *CouncilArea*, *Suburb*, and *Type* to ensure the imputation process was context-aware and aligned with the specific characteristics of each group.

To evaluate the quality of the imputation, we performed the DKW test, which compares the empirical cumulative distribution functions (ECDFs) of the original and imputed datasets. The imputation of *BuildingArea* performed well across all property types.

We dropped the following regions as there was limited data in those areas: *Eastern Victoria*, *Northern Victoria*, *Western Victoria*. Furthermore, we created some variables, such as the logarithmic value of *Distance* to normalize the data and mitigate the risk of extreme values' impact. We also introduced a squared value of *Bathroom* to capture potential non-linear relationships, particularly for properties with a higher number of bathrooms. Lastly, we created the dependent variable by dividing the *Price* by *BuildingArea* column, with an exception for types *h* and *t*, where we divided *Price* by the sum of *BuildingArea* and *Landsize* to account for their joint contribution. We made this adjustment to neutralize the dominant effect of building area while incorporating the significance of land size for types *h* and *t*. We employed oversampling techniques to address underrepresented cases, particularly for interactions like *Relative_Landsize:Type_h* and *Relative_Landsize:Type_t*, ensuring balanced data for model training. As the next step, we identified the outliers in this column using the IQR method and filtered the dataset to exclude extreme values.

Lastly, we computed the *Vicinity Density* column using KDE. By applying KDE with a Gaussian kernel and a bandwidth of 0.01, we calculated the density of properties based on their latitude and longitude. It provided a nuanced measure of population density by filling gaps where traditional metrics might fail.

### 2.2 Variable Selection

We used the calculated column, *Price per square meter*, as the dependent variable. We selected the following independent variables to conduct a reliable and parsimonious analysis: *Log Distance*, accounting for the logarithm of a property's distance to the city center, *Sq Bathroom*, capturing the non-linear relationships through the square of the number of bathrooms, *Property Count*, indicating the number of properties within a suburb, *Vicinity Density*, providing insights into the influence of spatial property density within a given neighborhood, *Relative Landsize*, representing the difference between a property's land and the average for its type. Additionally, we included some interaction terms involving *Relative Landsize* and property types to capture type-specific effects of land size. We chose aforementioned variables based on their relevance and potential influence on property pricing. They collectively contribute to a comprehensive understanding of property pricing dynamics.

### 2.3 Bayesian Regression

Our Bayesian regression model is built on probabilistic reasoning to estimate the relationships between predictors and the target variable. The hypothesis was that the dependent variable, *price per sqm*, is influenced by a combination of the predictors, region-specific effects, and random variability.

Firstly, we constructed the design matrix by including the previously selected predictors. To ensure consistency across predictors with varying scales, we standardized them using *StandardScaler*. Standardization transformed each feature to have a mean of 0 and a standard deviation of 1. This step ensured that all predictors contributed equally to the regression model. As a result, the estimated effects in the model correspond to a one standard deviation change in the predictors rather than a unit change.

The Bayesian model had three main components: intercept prior, horseshoe prior for predictors, and hierarchical priors for region-type effects. We modeled the global intercept as a normal distribution with a mean of 0 and a standard deviation of 10. To manage sparsity and promote shrinkage for less significant predictors, we used a hierarchical horseshoe prior. The global and local shrinkage parameters presented a Half-Cauchy distribution, whereas the regression coefficients were modeled as normal distributions.

We incorporated hierarchical priors for region-type effects to account for regional heterogeneity, reflecting how prices vary across different regions and property types. Based on the Domain Price by Square Meter Report 2024, which aggregates square meter pricing data across regions in Australia, we set the prior mean for these effects at 7300. To accommodate the variability in prices across regions, we assigned a high prior variance of 1500. Each region-type combination has its own effect, following a normal distribution centered on the observed mean for that group, with variability informed by its observed standard deviation.

We modeled the likelihood function using a student-t distribution, which is robust to outliers and heavy-tailed data. The model predicts the mean property price as a linear combination of the predictors, region-type effects, and an error term. We placed a Half-Normal prior for the error term's variance, as we expect to have positive and relatively small deviations.

We used a Bayesian sampling framework, where posterior distributions for the model parameters are estimated using Markov Chain Monte Carlo (MCMC), with 2000 iterations and a target acceptance rate of 0.95.

## 3 Results

The global intercept has a mean estimate of -1.575 with a standard deviation of 9.932. The 97% Highest Density Interval (HDI) ranges from -20.321 to 16.765, reflecting high uncertainty. Therefore, the intercept contributes little explanatory power to the model, with the predictors playing a substantial role in capturing property price variability.

Among the main predictors, *Log Distance* shows a strong negative impact with a mean estimate of -779.089 and an HDI of [-797.775, -759.333]. This result suggests that properties further from the city center tend to have significantly lower prices per square meter, consistent with the expectations. In contrast, *Sq Bathroom* exhibits a positive impact with a mean of 137.391 and an HDI of [126.442, 147.718], indicating that properties with more bathrooms generally have higher prices. *Property Count* demonstrates a modest positive effect with a mean estimate of 26.131 and an HDI of [15.031, 36.685], showing that a higher number of properties in the suburb increases property prices due to popularity. On the other hand, *Vicinity Density* has a negative mean effect of -76.683 and an HDI of [-91.261, -61.478], showing that higher population density in the vicinity reduces property prices. This outcome suggests that areas with a higher density of properties experience higher prices, reflecting the Demand-Supply theory.

The interaction terms involving relative land size and property types show varying impacts on pricing, depending on the property type. For instance, *Relative_Landsize:Type_h* and *Relative_Landsize:Type_t* have negative effects, with mean effects of -405.535 (HDI: [-418.024, -393.016]) and -285.356 (HDI: [-393.492, -268.901]), respectively. In contrast, *Relative_Landsize:Type_u* exhibits a positive impact, with a mean effect of 68.504 (HDI: [52.389, 84.825]). These results indicate that for properties of types *h* and *t*, increased relative land size reduces price per square meter, whereas for type *u*, it corresponds to higher prices per square meter.

Region-type effects capture variability across geographic and categorical combinations, showing how prices are influenced by location and type. The mean effects consist of a wide range, from 1748.752 (Western Metropolitan_h) to 7611.352 (Southern Metropolitan_u). Units (*u*) generally emerge as the most expensive property type across most regions, with the highest effect observed in Southern Metropolitan (7611.352), followed by Eastern Metropolitan (7403.415). Conversely, houses (*h*) tend to be the least expensive property type, with the lowest effect in Western Metropolitan (1748.752) and Northern Metropolitan (1809.227). For example, units in the Southern Metropolitan and Northern Metropolitan, with mean effects of 7611.352 and 6248.952, respectively, have higher prices per square meter, likely because of higher demand and premium urban locations. On the other hand, houses in Western Metropolitan and Northern Metropolitan, with mean effects of 1748.752 and 1809.227, respectively, exhibit lower prices, potentially due to larger land availability and lower demand in the regions. These differences underscore the importance of location and property type in determining property values.

Finally, the small Monte Carlo Standard Error (MCSE) values confirm the robustness of the results and indicate stable sampling for mean estimates. Effective Sample Size (ESS) values are high, ensuring reliable posterior estimates and well-mixed MCMC chains. Additionally, *R_hat* values are all 1.0, showing that the chains converged appropriately and that the posterior distributions are reliable.

## 4 Performance Measures

### 4.1 Multicollinearity Check

The Variance Inflation Factor (VIF) analysis was used to check for multicollinearity. It indicated minimal multicollinearity among the predictors in the model, with all VIF values well below the critical threshold of 10. *Log_Distance* exhibits a VIF of 1.97, showing a low level of collinearity. Similarly, *Sq_Bathroom*, *Propertycount*, and *vicinity_density* display VIF values of 1.04, 1.01, and 1.77, respectively, reflecting their independent contributions to the model.

The categorical variables *Relative_Landsize* also show very low VIF values ranging from 1.02 to 1.29, this confirms negligible multicollinearity. The results suggest that the predictors are well-suited for regression modeling without any concern about multicollinearity distorting the estimates of the parameters.

### 4.2 Residual Plots

We plotted the predicted price per square meter against their corresponding actual values. It demonstrated a generally well-performing model, with most predictions closely aligned with the perfect-fit line. However, we identified variability at higher price ranges, particularly above 9,000, where the data is less densely represented.

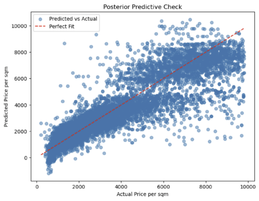

Figure 1: Posterior Predictive Check: Predicted vs Actual Price per sqm.

The residual distribution further highlights the model's strong performance, with residuals following a near-normal distribution. The symmetry indicates the model's predictions are unbiased and errors are evenly distributed. The narrow peak around zero reinforces the model's reliability, while minor deviations in the tails suggest an opportunity for further fine-tuning.

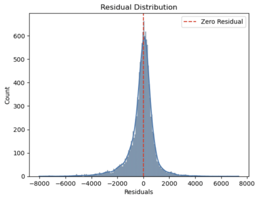

Figure 2: Residual Distribution: Residuals follow a near-normal distribution.

### 4.3 Goodness-of-fit

The resulting $R^2$ value of 0.8103 is highly promising, suggesting that the model explains 81.03% of the variability in the data, effectively capturing the underlying patterns. The Normalized Mean Square Error (NMSE) of 0.19 indicates acceptable precision, although a lower value closer to 0 would be ideal. However, an NMSE of 10-20% is generally regarded as acceptable in housing price predictions (Vexpower). Given the high $R^2$ value, the regression model can be deemed reliable.

## 5 Conclusion

The project successfully demonstrates the application of Bayesian regression modeling to analyze housing price dynamics. By cleaning and imputing data, filtering outliers, and engineering meaningful features, we ensured an effective data analysis. The model effectively captured significant relationships between prices and predictors such as distance to the city center, number of bathrooms, and regional effects. Through the inclusion of hierarchical priors, we addressed regional heterogeneity, while by performing the DKW test, we ensured imputation validity and model reliability.

Performance metrics, including an $R^2$ value of 0.8103 and low residual errors, validate the model's accuracy and predictive power. The model provides a solid foundation for understanding pricing dynamics. As a result, this analysis underscores the strength of Bayesian approaches in delivering predictable and reliable insights into complex datasets.

## 6 References

OpenAI. ChatGPT (Version January 2025). OpenAI, 2025, https://openai.com/chatgpt

Domain Price by Square Meter Report 2024. CRTV-3256_DI-PriceBySquareMetre-Report, Domain, 2024, https://s3.ap-southeast-2.amazonaws.com/ffx.adcentre.com.au/domain/2024/CRTV-3256_DI-Price+by+Square+Metre+Report/CRTV-3256_DI-PriceBySquareMetre-Report-A4-Digital-FA.pdf

Vexpower. "Mean Absolute Percentage Error (MAPE)." Vexpower, n.d., https://www.vexpower.com/brief/mean-absolute-percentage-error