Statistics and Probability Final Project

Marie Cieslar, Gaia Iori, Marta Laskowska, Javiera Rubio
Group 11

January 11th, 2025

1 Introduction

Accurately predicting housing prices is a critical task in real estate analytics, providing valuable insights for buyers, sellers, and policymakers. Our project focuses on analyzing factors influencing housing prices in Melbourne, Australia and developing a predictive model to estimate them, based on similar characteristics, using data collected in 2016 and 2017. The dataset comprises a variety of features, including structural attributes of properties, geographic information, and transaction details. By leveraging these features, our models aim to uncover key drivers of housing prices and provide reliable predictions.

2 Data Preprocessing

Our first step was to optimize the dataset for predictive modeling while maintaining its interpretability and analytical robustness. We systematically addressed outliers and missing values in the data. After initial detection, we filtered out outliers from two variables, `Landsize` and `YearBuilt`, using upper and lower bound restrictions based on interquartile ranges.

Additionally, we detected four variables with missing values: `BuildingArea`, `YearBuilt`, `CouncilArea`, and `Car`. Notably, there were significant differences between properties with missing and non-missing values for `YearBuilt`, suggesting that more expensive, potentially luxurious properties may not provide full transparency regarding their characteristics. After identifying the percentage and patterns of missingness for each variable, we chose a fitting approach for handling the missing values. Due to the low proportion of missing values in the variable `Car` we chose to drop all rows containing missing values. Following preliminary analysis, we imputed `YearBuilt` by sampling from region-specific distributions of observed values, as we hypothesized differences in property age within different regions. `BuildingArea` was imputed using K-Nearest Neighbors (KNN) with an RMSE optimized parameter of K neighbors. Lastly, for `Landsize`, zero entries, indicative of void data collection, were replaced with corresponding `BuildingArea` values to improve data integrity.

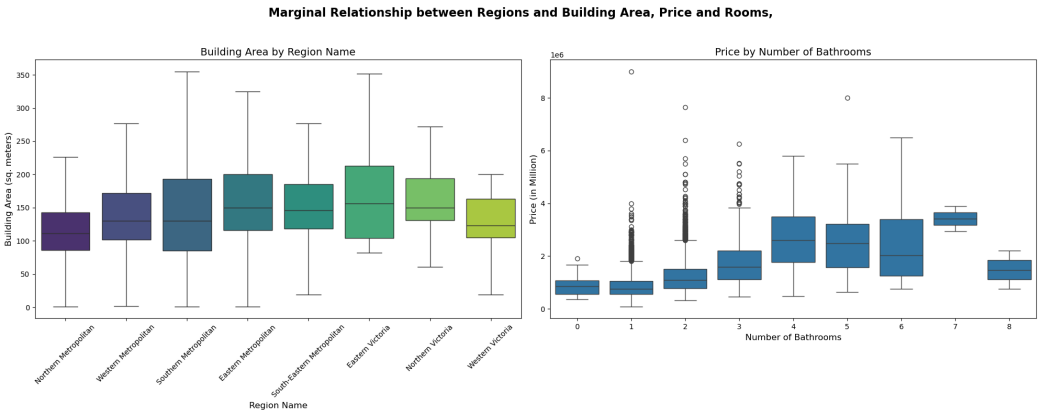
Furthermore, we created several new features to enhance predictive power, such as `HouseAge`[1]. Continuous variables like `Rooms`, `Bathroom`, `Car`, and `BuildingArea` were log-transformed to address skewness, stabilize variance and account for decreasing returns on price.

To refine the dataset, we conducted correlation analyses using heatmaps and matrices, accounting for various variable types, identifying those that are strongly correlated and avoiding overfitting, setting the threshold at 0.7. To capture linear and non-linear temporal patterns and trends, we used encoded time variables indicating `Date`.

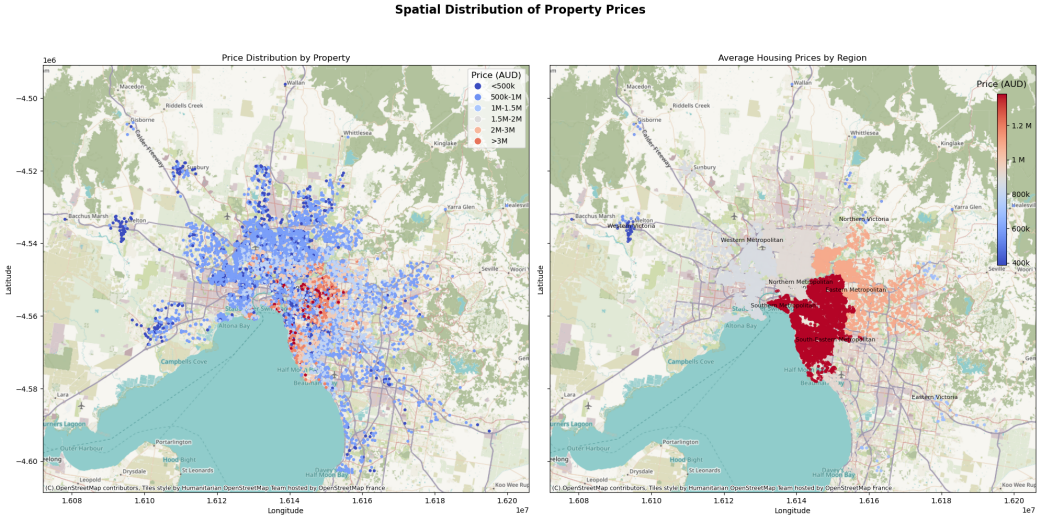
3 EDA

To gain a deeper understanding of housing prices in relation to other factors, we conducted a Bivariate Analysis. The data visualizations revealed notable patterns, such as properties with a greater number of `Bathrooms` displaying an upward trend in `Price`, while the distribution of building areas varied significantly across regions.

[1] Smith, J., & Doe, A. (2023). *House price prediction.* AgEcon Search.



In addition, we investigated spatial trends by analyzing average property prices grouped by region. We identified four main clusters and, for subsequent modeling purposes, grouped the regions into these four categories.



4 Linear Regression

Based on the concept of hedonic house pricing models [2], we developed a Linear Regression to investigate what factors influence housing prices in Melbourne. We evaluated its predictive performance by randomly dividing the dataset, while stratifying by region to account for location differences.

To avoid multicollinearity, we excluded the variable `Rooms`, due to its strong correlation with the number of bedrooms, and omitted one dummy variable from each set to serve as a benchmark for our model. To address the variability of building areas in various regions, as observed in the EDA, we introduced interaction terms. We decided on the following specification for the linear regression model: [3]

$$\begin{aligned}\log(\text{price}) = {} & \log(\text{Landsize}) + \log(\text{Distance}) + \log(\text{PropertyCount}) + \text{time}_{\text{cont}} + \text{time}_{\text{cos}} + \text{time}_{\text{sq}} \\ & + \log(\text{Bathroom}) + \log(\text{Car}) + \log(\text{Bedroom2}) + \text{HouseAge} + \log(\text{BA}) \\ & + \log(\text{BA}) \cdot \text{RG}_{\text{EMetropolitan}} + \log(\text{BA}) \cdot \text{RG}_{\text{NSSEMetropolitan}} \\ & + \log(\text{BA}) \cdot \text{RG}_{\text{WNEVictoria}} + \text{RG}_{\text{EMetropolitan}} + \text{RG}_{\text{NSSEMetropolitan}} + \text{RG}_{\text{WNEVictoria}} \\ & + \text{Method}_{\text{PI}} + \text{Method}_{\text{SA}} + \text{Method}_{\text{SP}} + \text{Method}_{\text{VB}} + \text{Type}_{\text{t}} + \text{Type}_{\text{u}}.\end{aligned}$$

The model demonstrates moderately high explanatory power, with an adjusted R^2 of 0.64, a metric accounting for potential overfitting, indicating that approximately 64% of the variance in log-transformed property prices is

[2] Smith, J., & Doe, A. (2023). *A review of hedonic pricing models in housing research.* ResearchGate.
[3] In the equation, RG stands for RegionGroup, and BA stands for BuildingArea.

explained by the predictors.

Considering the statistical significance of the coefficients at the 0.05 significance level, the following conclusions can be drawn. Property type significantly influences prices. Townhouses are, on average, 7.4% less expensive than houses, while units and duplexes exhibit an even larger price reduction of, on average, 29.4%.

The dynamics of property prices are shaped by a combination of locational and structural factors. Specifically, a 1% increase in land size raises prices by around 11.6%. In contrast, a 1% increase in distance from the Central Business District (CBD) and property density reduce prices by around 42% and 1.7% accordingly. An increase in the number of bedrooms, bathrooms and car slots, is associated with a significant increase in the property prices, while the first two factors influence the price more significantly.

As initially observed, the model provides evidence that there are significant differences between the impact of certain regions on the property price. For a 1% increase in building area, property prices in the Northern, Southern and South-Eastern (NSSE) Metropolitan region increase by 13.38 p.p. more compared to the Western Metropolitan region. There is no strong evidence of a heterogenous impact of building area on property prices in the other two regions. Contrary to initial expectations, properties in the NSSE Metropolitan region are, on average, priced 31.59% lower than houses with similar characteristics in the Western Metropolitan region.

To ensure that the developed model meets the assumptions of linear regression, we verified the absence of autocorrelation in the residuals and tested whether the error terms follow a normal distribution. The Durbin-Watson statistic (2.01) suggests the absence of significant autocorrelation in the residuals, however, the Jarque-Bera test yields p-value less than 0.001, indicating that the residuals deviate from normality. Although considered as a potential limitation, this does not affect the unbiasedness or consistency of the regression coefficients.

5 Bayesian Linear Regression and Markov Chains Monte Carlo

To predict housing prices, we implemented a Bayesian Linear Regression model following the same specification as described above and leveraging Gibbs Sampling to approximate the posterior distributions of the parameters.

Let:

- y = vector of observed prices
- X = matrix of predictors
- β = vector of coefficients
- σ^2 = variance of the errors
- $\mu = X\beta$ = predicted mean

We assume a normal likelihood function for the residuals, therefore the log-likelihood function is described as:

$$\ell(y \mid X, \beta, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - X_i\beta)^2$$

Where n is the number of observations ($n = \text{len}(\mathbf{y})$), $\mu_i = X_i\beta$ is the predicted mean for observation i, and $(y_i - \mu_i)$ is the residual for observation i.

Because of our limited domain knowledge regarding the real estate market, we define non-informative priors for our coefficients as:

$$\beta_j \sim \text{Normal}(0, 20^2)$$

While for the variance, we followed the recommendations found in the literature regarding the choice of the Inverse-Gamma prior as it ensures that the values are always positive,

$$\sigma^2 \sim \text{InverseGamma}(\alpha = 2, \beta = 1)$$

The posterior distribution combines the log-likelihood function and the prior distributions of the parameters:

$$\log \mathit{posterior}(\beta, \sigma^2 \mid y, X) = \ell(y \mid X, \beta, \sigma^2) + \log\left(\mathit{prior}(\beta)\right) + \log\left(\mathit{prior}(\sigma^2)\right)$$

After splitting the dataset into training and test sets, we standardized continuous predictors using z-score normalization to ensure all variables were on a comparable scale.

Finally, we employed Gibbs Sampling to iteratively draw samples from the conditional posterior distributions. After 10,000 iterations, with a burn-in period equal to 2000, the retained samples approximate the joint posterior distribution of β and σ^2.

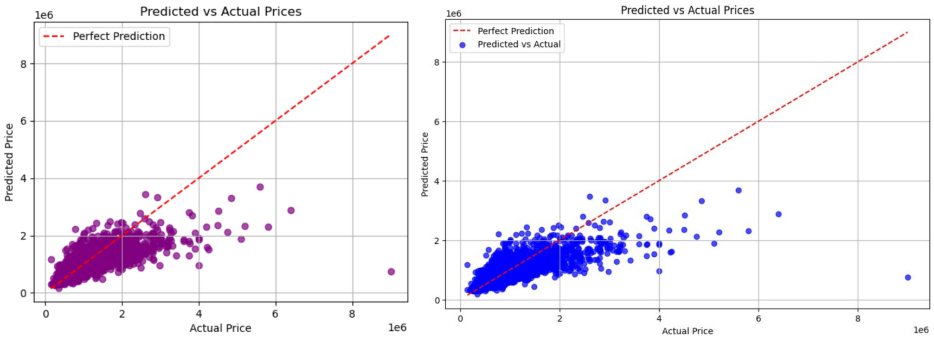
6 Results and Models Prediction Comparison

First, the benchmark property of our model is in the Western Metropolitan region, which in our EDA showed a middle-range average price compared to other region groups. Both models show a positive (mean) [4] coefficient regarding the Eastern Metropolitan region, which confirms the initial findings, yet for the NSSE Metropolitan Regions the (mean) coefficient is negative, even if the initial analysis shows a higher average price, which is most likely due to the grouping criteria we used for the feature engineering aspect of these models.

Second, both models estimate positive (mean) coefficients for the internal characteristics of the property, i.e. number of bathrooms, car spaces and bedrooms, which confirms the pattern discovered in the EDA. It's interesting to note that the linear regression estimates coefficients which are higher than their estimated mean in the Bayesian model.

Third, the method of sale for our benchmark property is "S - property sold", and both models estimate negative (mean) coefficients for the four other methods. This result coincides with our initial intuition regarding a lower price for properties sold via auctions or bids.

Below is a graphic comparison of the price predictions for each model. The Bayesian linear regression model is shown on the left side, while the linear regression is shown on the right side.



7 Conclusion and Discussion

In our study, we analyzed diverse patterns in housing prices in Melbourne using both Linear Regression and Bayesian Linear Regression models. The findings indicate key drivers of building prices such as the number of bathrooms and bedrooms, the property type and the proximity from the CBD and regional differences. Western Metropolitan properties indicate on average higher prices than NSSE Metropolitan properties, however, increases in building area have a greater impact on prices in the NSSE region. The observed price variations may be attributed to factors such as regional development, infrastructure, and urban desirability. Properties closer to urban centers and those with historical or architectural value are hypothesized to command price premiums. Conversely, lower prices in densely developed areas and for properties sold through certain methods, such as vendor bids, could reflect oversupply, reduced demand, or buyer perceptions. While our models provide meaningful insights, additional variables such as real estate agent information, suburb characteristics, street-specific data, and governing council to account for potential subsidies or tax-reductions, could further improve its explanatory power. We therefore recommend for future research purposes to explore those factors to help better capture the complex dynamics of Melbourne's housing market.

[4] The Bayesian Linear Regression estimates the mean of the coefficients