Statistics Final Project

Michael Ladaa - 3173924, Yash Patel - 3318837,
Matteo Roda - 3160426, Omar Mukhtar - 3182217

January 2025

1 Introduction

The housing market plays an important role in our day to day lives economically and socially. Given this importance, much research goes into understanding the factors that influence house prices. In this paper, a housing dataset from Melbourne outlining house sales between 2016 and 2017 was analysed to provide such insights. The focus of this paper is to determine the relationship between the year the property was constructed and the price of sale. The motivation for this stems from an expected non-linear relationship between the year built and the price. Old builds may be viewed as 'vintage' or 'antique', driving up the price. Similarly, new builds are priced higher due to developers recuperating costs as well as heightened demand. Meanwhile, in the range of this dataset, middle-aged houses come in post world war eras maybe indicating poorer build quality and lower price. To achieve this, we follow an approach utilizing MCMC and Newton's Method.

2 EDA

The initial exploratory analysis of the dataset highlighted many results, including those that were of interest to our preliminary idea. The dataset is composed of 21 features ranging from geospatial data to specific house characteristics to sale characteristics. While mostly clean, the exploration revealed some missing-ness in data which is handled in the next section Data Cleaning. Clearly but unsurprisingly we can see a clear trend between the location of the houses and their price, with central locations being more sought after in Figure 1. Moreover, we observe a positive impact of the number of rooms on price, albeit a diminishing effect around 5 rooms or greater. This could be attributed to average family sizes dictating the demand for the number of rooms.

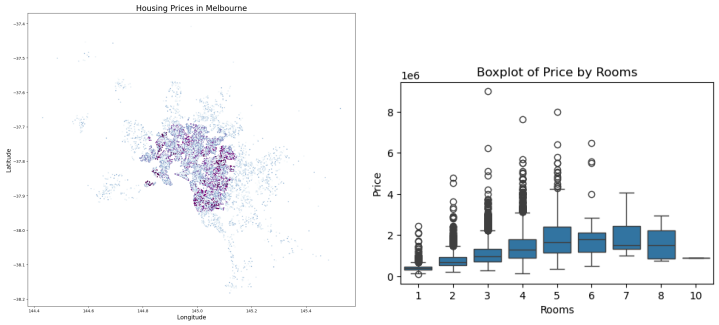

Figure 1: Map of Houses Price Distribution (left) and Box Plot of Rooms Against Price(right)

Furthermore, the violin plots in Figure 2 highlighted the impact of year built on the house prices. By dividing the observations into buckets of year built that contain an equal number of observations in each bucket, the plot on the left exhibits a slight upward curve. This effect is even more pronounced when removing extremely high-valued houses sold at over 4 million dollars, of which there were few observations. It is worth noting that one outlier was dropped as the year built predates the oldest registered house in Australia.

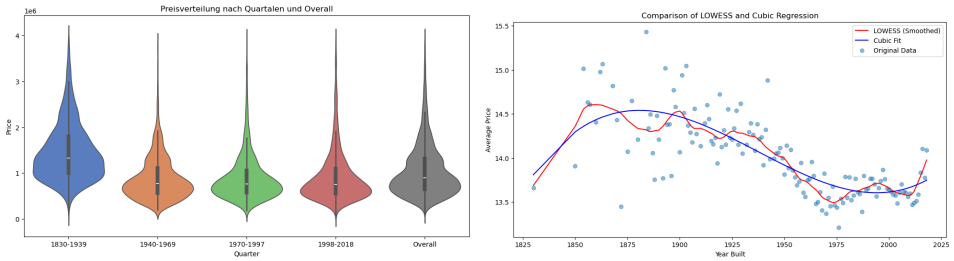

Figure 2: Violin Plot (left) and Average Log Price per Year (right)

We then proceeded to examine this relationship further. In Figure 2, using a smoothed LOWESS curve and a cubic fit line of year built on the logarithm of price highlights the non-linear relationship we expected. A high R-squared score of 0.90 suggests the existence of a potential cubic relationship.

3 Data Cleaning

Before proceeding with deeper analysis, we needed to address the missing values in the dataset. The dataset contained four columns with missing values: Car (62), BuildingArea (6450), YearBuilt (5375), and CouncilArea (1369). Upon closer inspection, we also identified a significant number of zero values in the Land size column, totaling 1939 data entries. Since a land size of zero is unrealistic for real estate data, we treated these values as missing. A similar approach was applied to a missing values in the columns Building Area, Bedroom, and Bathroom. In order to address these missing values, we chose to apply different imputation methods, tailored to each column.

3.1 Mode Imputation

For the Council Area column, we implemented a mode-based imputation method. Here, we imputed missing values based on the most frequent Council Area for each given Postcode. This approach decreased our number of missing values in this column to 3. For the remaining missing values, we imputed the overall most frequent Council Area across the entire dataset.

3.2 Iterative Imputer

For the columns Bathroom, Bedroom, Land size and Building Area we decided to implement an Iterative Imputer. This method allows us to leverage the relationships between the different columns in the dataset to estimate the missing values efficiently. This was done by iteratively predicting and refining the missing values for each column, based on available data.
As an estimator of the Imputer, we employed a Random Forest Regressor, specifying 20 trees and allowing up to 100 iterations for convergence. This approach allowed us to preserve the data distribution while filling the missing values.

3.3 K − Means Clustering + Bayesian Imputation

In order to handle missing values in the Year Built column, we used both clustering and Bayesian imputation.

3.3.1 K-Means Clustering

We applied K-Means Clustering to group similar real estate based on selected features: Rooms, Price, Land Size, Building Area, and Distance. Given the size, price and location of the housing units, we believed to be able to capture houses from similar time eras in the cluster.
To prepare the data for clustering, we standardized the features using a Standard Scaler to ensure that each feature was normally gaussian distributed, as K-Means is sensitive to the scale of the data.
The number of clusters was set to 10, and the K-Means algorithm was applied with a random seed. The resulting cluster labels were added as a new column Cluster to the dataset, allowing us to analyze the data in terms of these groups.

3.3.2 Bayesian Imputation

For imputing missing values in the Year Built column, we implemented an imputation technique based on Bayesian Inference. The method works by first estimating a Kernel Density Estimation based on the known values of Year Built within each cluster. This serves as a prior distribution for the missing values.
We then calculated a likelihood for each missing value based on the similarities of the features defined

before. The similarity was computed using a weight derived from the differences between the current missing entry and the known values in the cluster. These weights were normalized to form a probability distribution. Using the posterior distribution, we sampled from the existing values of Year Built within the cluster, weighted by their similarity to the missing entry, to impute the missing values.

3.4 Zero Replacement

Lastly, we were left with 62 missing values for the car column. As we interpret a missing data entry as having no car spot, we simply replaced these missing values by zeros.

4 Model Specification & Initial Regression

4.1 Introduction

In exploring factors influencing housing prices in Melbourne, we identified a potential non-linear relationship between the year a property was built and its price. Visualising the data in Figure 2 revealed the potential for a cubic trend, which became more pronounced when transforming the response variable to $\log(Price)$. This transformation was correspondingly employed to enhance the interpretability of the regression coefficients.

4.2 Methodology

A cubic regression model was employed to capture this non-linear relationship. The regression included 'YearBuilt', its second- and third-order terms 'YearBuilt^2' and 'YearBuilt^3', and additional covariates consisting of 'Rooms', 'Distance', 'Landsize' and an interaction term 'Rooms:Distance'.
These covariates were found to be most significant when trialling alternative regressions. The interaction term was included to account for potential heterogeneity in the impact of the number of rooms based on proximity to the Central Business District (CBD). All variables were standardised before the regression.
The model equation was specified as:

$$\begin{aligned}\log(Price) = \beta_0 &+ \beta_1 YearBuilt + \beta_2 YearBuilt^2 + \beta_3 YearBuilt^3 \\ &+ \beta_4 Rooms + \beta_5 Distance + \beta_6 Landsize \\ &+ \beta_7 (Rooms \times Distance) + \epsilon\end{aligned}$$

4.3 Result

The regression results are summarised as follows:

- **Significance:** All predictors were highly significant ($p < 0.05$), with *YearBuilt* and its polynomial terms $YearBuilt^2$ and $YearBuilt^3$ showing strong evidence of a cubic relationship.
- **R-squared:** The model explains 50.53% of the variation in $\log(Price)$, with $R^2 = 0.5053$, indicating that the chosen predictors capture a substantial portion of the variability in housing prices.
- **Residual Analysis:** The residual standard error of 0.3706 suggests a reasonable fit for the data, with residuals centered around zero and no extreme outliers.

4.4 Interpretation

The quadratic and cubic terms for YearBuilt ($\beta_2 = 0.06033$, $\beta_3 = 0.00316$) highlight a non-linear relationship between housing prices and the year built. Our explanation for this phenomenon is that newer houses are often built for demand with modern amenities and lower maintenance costs and so attract a higher price, whilst older properties can also attract premium prices, especially if they possess unique architectural features or are situated in historically significant neighbourhoods. This leaves a period in-between these two, in which houses built do not fit into either of these categories and hence lead to a relative fall in price.
The interaction term ($\beta_7 = -0.0665$) reflects how the number of rooms becomes less impactful as the distance from the CBD increases, likely due to reduced demand in suburban areas.

4.5 Conclusion

We determine that our results validate the hypothesis of a cubic relationship between housing prices and the year in which the property was built. By incorporating an interaction effect, the model also captures the spatial heterogeneity of Melbourne's housing market, namely with regard to number of rooms and pricing. This analysis lays a robust foundation for further exploration using advanced methods with MCMC and Newton's optimisation, as performed later.

5 MCMC & Newton's Method

5.1 Markov Chain - Monte Carlo

Building on these promising results, we utilise MCMC to sample the parameters from the specified functional form. Given the sparsity of data in the early years of the dataset, the motivation behind this approach is to get a more accurate estimate of these parameters. The priors used in the model were chosen based on the regression results and under the assumption the log of price assumes a normal distribution. A burn-in period of 1000 iterations was chosen with another 2000 samples generated afterwards with 4 chains processing simultaneously.

The convergence diagnostic suggest convergence of the chains with ESS sizes between 4700 and 8500 indicating lower autocorrelation between samples. Moreover, the trace plots below provide further evidence of convergence as evidenced by the trace plot in Figure 3 with a constant mean and consistent variance indicating low autocorrelation between samples. For complete reference of all trace plots, consult the associated jupyter notebook.

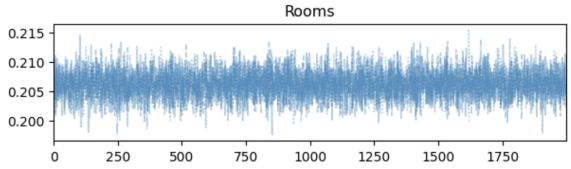

Figure 3: Trace Plot of 'Rooms' Parameter

Given the convergence of the chain, in the next section, Newton's Method, we use the mean of these samples as an initialisaion for Newton's Method to find the optimal parameters describing the desired relationship.

5.2 Newton's Method

To estimate the optimal coefficients for the parameters of our regression model, we employed the Newton's method. This aims to refine the initial coefficients, which were derived from the MCMC sampling, by minimizing the negative log-likelihood of the observed data.
In each iteration, Newton's method uses the gradient of the negative log-likelihood function, which indicates the direction in which the coefficients should be updated. Moreover, it takes the Hessian matrix, which provides information on the curve of the likelihood function. By combining these, we were able to scale the updates and converge more efficiently to the optimal coefficients.
The formula applied is the following:

$$\beta_{\text{new}} = \beta_{\text{old}} - \left(H^{-1} \cdot \nabla L(\beta)\right) \tag{1}$$

where $\nabla L(\boldsymbol{\beta}_i)$ is the gradient and H is the Hessian matrix.
The process continues until the change in the coefficients between iterations is below a specified tolerance, which we set at 1e-6, indicating that the model has converged to an optimal solution. The final coefficients obtained were $[\beta_0 = 13.75, \beta_1 = -0.12, \beta_2 = 0.34, \beta_3 = 0.35, \beta_4 = 0.43, \beta_5 = 0.85, \beta_6 = 0.0086, \beta_7 = -0.27]$ thus confirming the existence the non-linear relationship we expected. The β_1 coefficient tells us that for every standard deviation increase in year built, how prices decrease by about 10.9%. Similarly, the positive coefficient of β_2 captures the upward curve we imagine, indication that old homes rise again in price with 1 standard deviation increasing price by about 39%. Likewise, a standard deviation change in the cubic term, with coefficient, β_3 results in a 41.5% increase in price capturing this cubic component. In the interest of space, the other coefficients will not be interpreted given they are not of primary interest to the paper.

6 Conclusion

Overall, our results indicate a non-linear relationship between price of sale and year built of properties. From our results, it is clear that the relationship can be modeled by both a quadratic and a cubic component. These results provide useful insights into potential long-term holding strategies of properties and potential post-war constructions having lower values and demand.