Research Proposal and Data Introduction

A Preliminary Multiple Linear Regression Model

Alyna Qi Heidi Wang John Zhang

November 21, 2024

Contents

1 Introduction

In the real estate industry, understanding the factors that influence home prices is essential for homeowners, investors, city planners, and policymakers. Accurately predicting these factors is crucial for informed real estate investment decisions. This study delves into the dynamics of Melbourne's housing market, examining how various key factors collectively impact market trends.

This study utilizes Multiple Linear Regression (MLR) to model the relationship between the dependent variable (house price) and multiple independent variables, such as the number of bedrooms, bathrooms, distance to the city center, land size, and house type. MLR is particularly effective because it is assumed that these predictors can be linearly related to house prices, allowing for the quantification of the separate impact of each predictor while controlling for the influence of others. This provides a clear and structured method to understand how various factors contribute to property values.

Several peer-reviewed studies are referenced to explore factors influencing house prices. He and He (2021) utilized linear and logistic regression to assess the impact of Melbourne's location and property type on house prices, concluding that these factors significantly affect market value. This aligns with the current study's objectives, which similarly considers variables such as the number of rooms and building age. Yusof and Ismail (2012) employed multiple regression to explore how intrinsic house characteristics, such as the number of bedrooms and bathrooms, influence price variations, highlighting their significant impact on house prices. This provides a method to quantify real estate market behaviors effectively. Chen and Hao (2008) examined the effect of distance from the CBD on house prices in Shanghai through hedonic analysis, underscoring the critical role of location in property valuation. While these studies vary in geographic focus, their methodologies and conclusions offer important insights for understanding Melbourne's housing market. Collectively, they illustrate the utility of multiple linear regression in isolating the effects of various property characteristics on market values, thus providing a robust framework for academic and practical applications in real estate analysis.

Ultimately, with the model developed, we are able to not only explain the factors that shape current house prices, but also predict future changes in house prices. This predictive ability is a valuable tool for real estate investors and policy makers, enabling them to make more informed decisions in a market full of uncertainty.

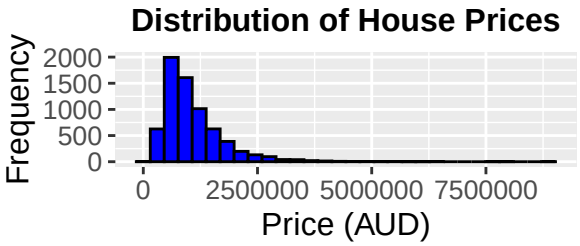
2 Data Description

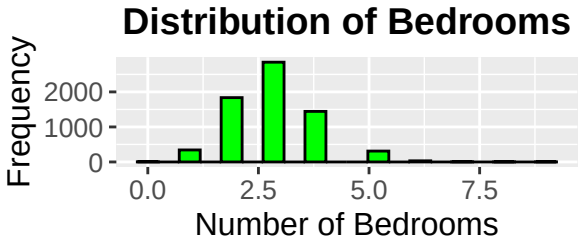
Table 1: Summary Statistics for Numeric Variables

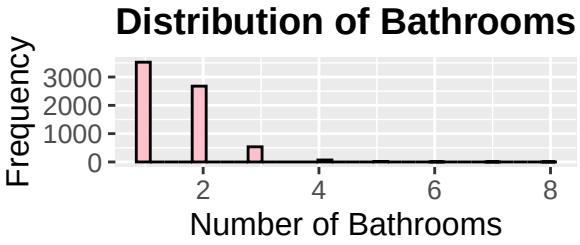
	Value
Price_Min	131000.00
Price_Mean	1077603.78
Price_Median	890000.00
Price_Max	9000000.00
Bedroom2_Min	0.00
Bedroom2_Mean	2.95
Bedroom2_Median	3.00
Bedroom2_Max	9.00
Bathroom_Min	1.00
Bathroom_Mean	1.59
Bathroom_Median	1.00
Bathroom_Max	8.00
Distance_Min	0.00
Distance_Mean	10.15
Distance_Median	9.20
Distance_Max	47.40
Landsize_Min	0.00
Landsize_Mean	487.50
Landsize_Median	404.00
Landsize_Max	37000.00

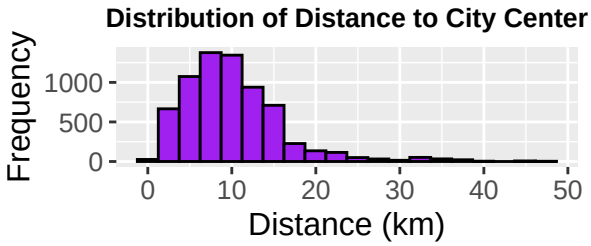
Table 2: Frequency Table for House Type

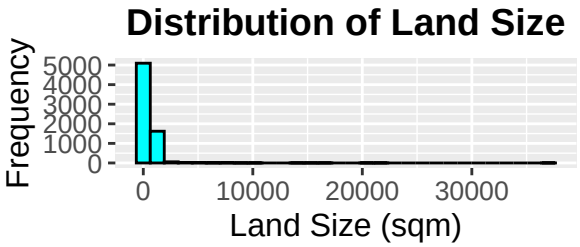
House Type	Frequency
h	4660
t	642
u	1528

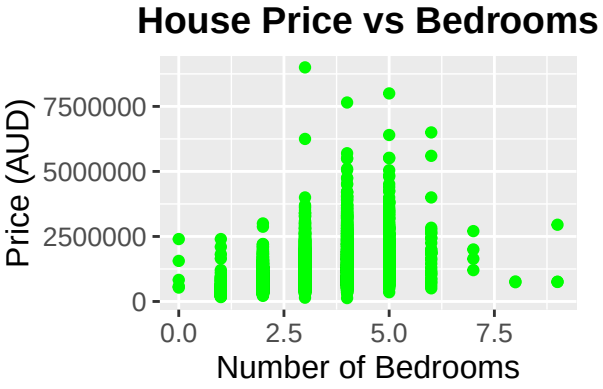


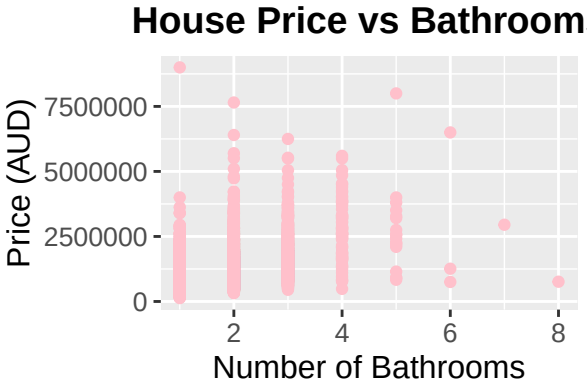


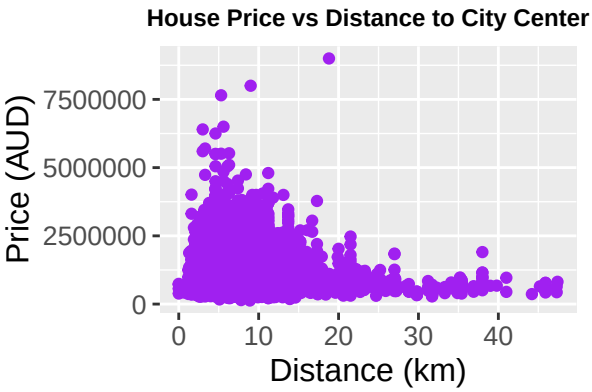


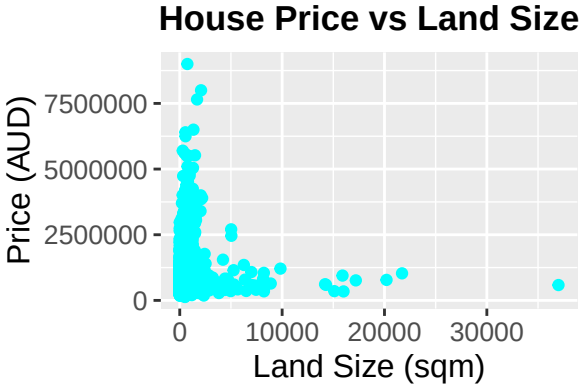












The dataset used in this analysis is derived from Melbourne Housing Snapshot on Kaggle (Becker, 2018). It was originally collected by Tony Pino from publicly available data on the real estate website Domain.com.au

(Pino, 2018) and collated to contain 13,580 observations and 21 variables related to Melbourne housing characteristics.

2.1 Why Linear Regression?

Linear regression is appropriate for this analysis, since the response variable is continuous, with predictors comprising both numeric and categorical variables. Each observation is assumed to be independent, and preliminary visual checks indicate that the response variable approximates a normal distribution, meeting the assumptions for linear regression.

2.2 Histogram and Scatter Plots Description

The histograms illustrate the distribution of each predictor and the response variable. House prices exhibit a right-skewed distribution, indicating that most properties are priced below AUD 2.5 million, with a few high-end outliers. The predictors such as the number of bedrooms and bathrooms follow relatively normal distributions, though some extreme values are present. Distance from the city center is skewed towards properties located within 10-20 km from the Central Business District (CBD). The scatter plots demonstrate a generally positive relationship between house price and predictors like the number of bedrooms, bathrooms, and land size, while distance shows a negative correlation.

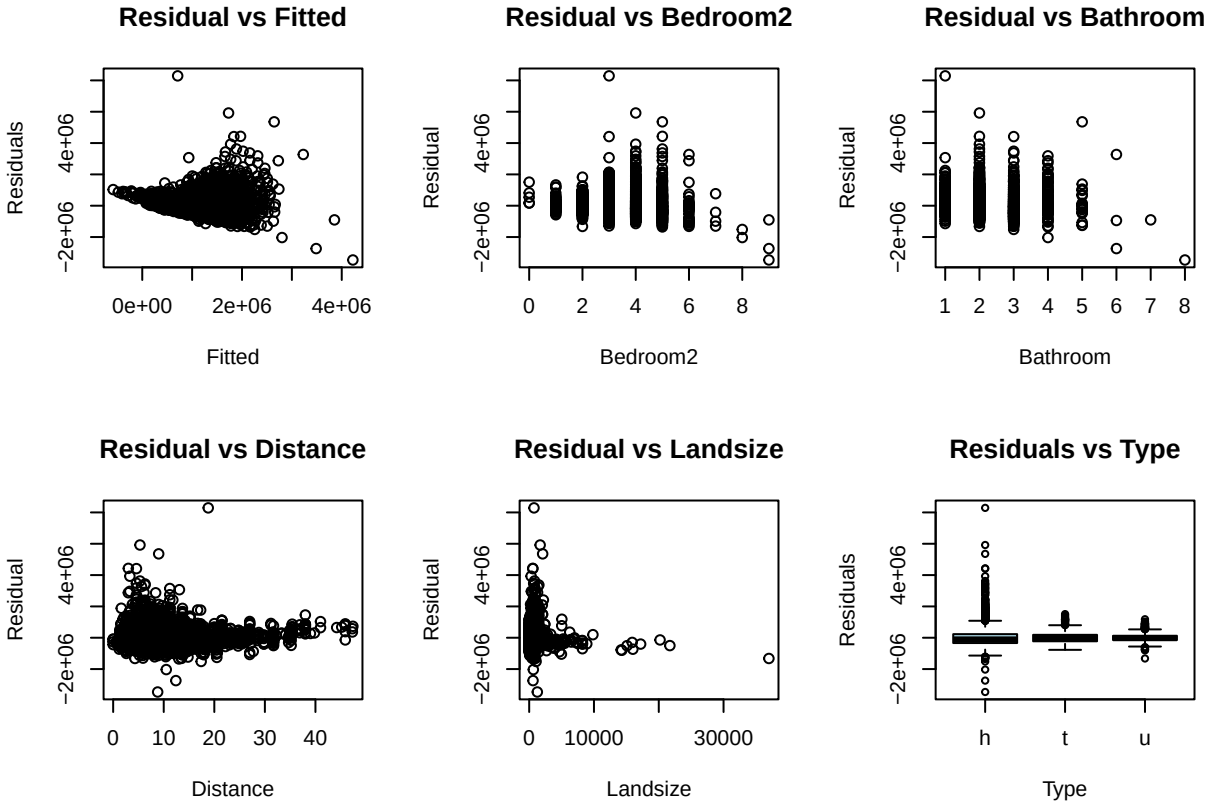
3 Ethics discussion

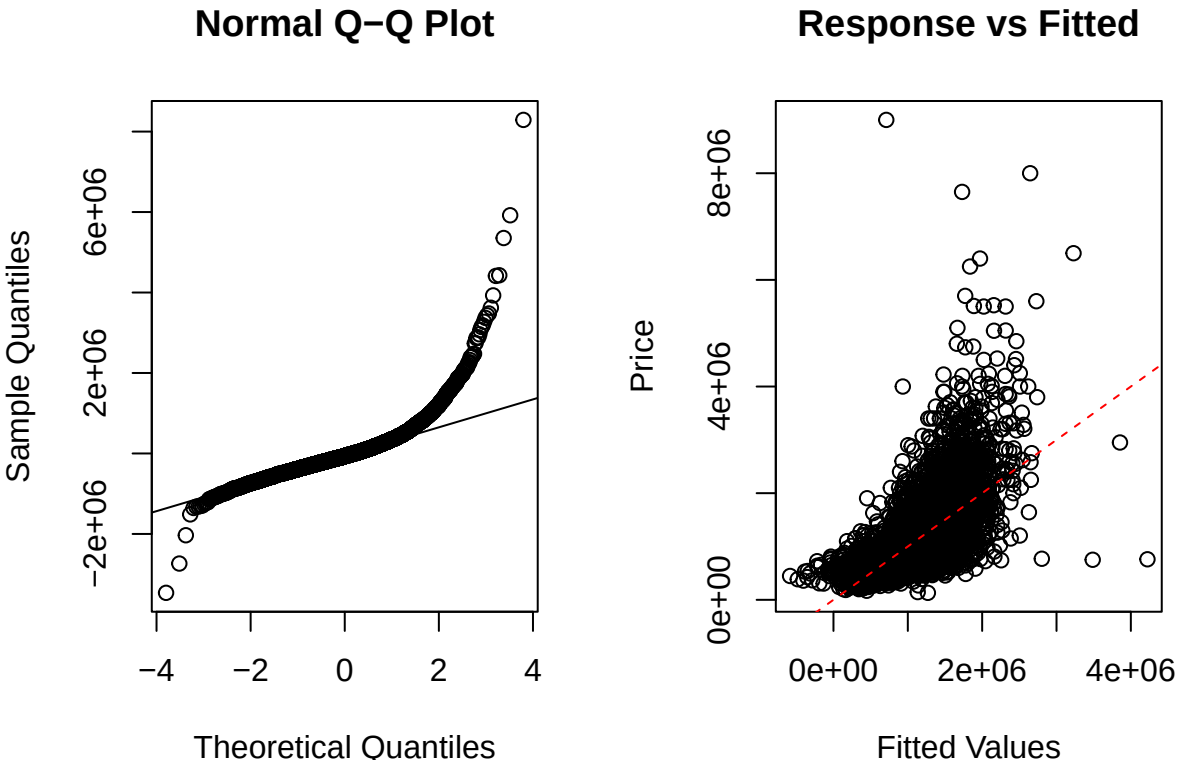
Based on the ethical considerations outlined in the first ethics module, Tony Pino's Melbourne housing dataset on Kaggle appears to be trustworthy. As detailed in the description section of the dataset, the data is crawled from public results published weekly by Domain.com.au, which indicates that its source is transparent. The dataset was collected from a reliable source with detailed descriptions, demonstrating that it has been rigorously vetted to meet ethical guidelines for data accuracy and source transparency. In addition, the anonymization of the data addresses privacy concerns, making it suitable for use in housing market analysis without compromising individual privacy. Additionally, the dataset is used by a broad community on Kaggle, which helps to ensure that the dataset has been informally peer-reviewed through user feedback and applications, thus strengthening its reliability. The process of updating the dataset on a regular basis by reputable real estate websites also enhances its relevance and timeliness, which is a key factor in its applicability to dynamic market analyses such as real estate trends and pricing assessments.

4 Preliminary Result

A preliminary linear regression model was fitted using `Price` as the response variable, with five predictors: `Bedroom2`, `Bathroom`, `Distance`, `Landsize`, and `Type`. The model aimed to determine significant factors influencing housing prices in Melbourne.

4.1 Residual Analysis and Diagnostic Plots





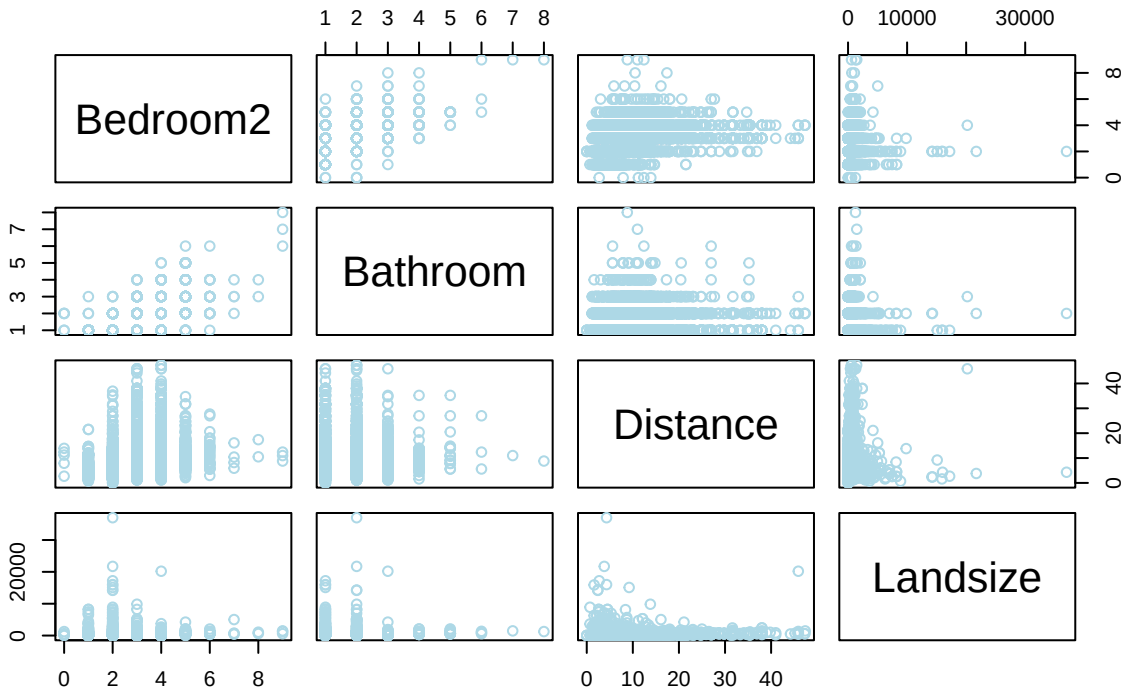

Pairwise Plots of Predictors
Bedroom2
Bathroom
Distance
Landsize
1 2 3 4 5 6 7 8
0 10000 30000
0 4 8
1 3 5 7
0 20 40
0 20000
0 2 4 6 8
0 10 20 30 40

Table 3: Summary of Assumptions and Conditions for the Preliminary Linear Model

Aspect	**Description**
Linearity	The Residual vs Fitted plot shows a noticeable pattern rather than random scatter. Residuals fan out at higher fitted values, which indicates non-linearity. Therefore, the linearity assumption is violated, implying that the model might not be properly capturing all the complexity in the relationship between predictors and response.
Uncorrelated Errors	In the Residuals vs Predictors plots: There are noticeable patterns in the residuals, especially in 'Distance' and 'Landsize', which show a widening spread and clustering in residuals. This suggests some correlation or dependence among residuals, implying that the errors are not independent. Additionally, the Residuals vs Type plot (boxplot) shows variations in the residual distributions for different categories (h, t, u). While there is some similarity, differences in variance imply potential dependence
Constant Error Variance	The funnel shape (wider spread as fitted values increase) indicates where the variance of residuals increases with the fitted values. This means the error variance is not constant. Such non-constant variance often suggests that the model does not perform well for different ranges of fitted values, and it violates the constant error variance assumption.
Normal Errors	The Q-Q plot deviates significantly from the diagonal line, especially in the tails. This implies non-normal distribution of residuals, particularly for extreme values, which affects the validity of statistical inferences.
Condition 1: Conditional Mean Response	The Response vs Fitted plot should ideally show points along the diagonal line, indicating a good fit. However, the provided plot shows significant deviations, especially for higher fitted values, meaning the model does not adequately capture the linear relationship between the predictors and the response. Thus, the conditional mean response condition is violated, suggesting underfitting, particularly for high-priced properties.
Condition 2: Conditional Mean Predictors	The Pairwise Scatter Plots of predictors should show no clear non-linear patterns if the predictors are adequately modeled. However, clustering is evident, especially between Bedroom2 and Bathroom, and there are potential non-linear relationships among some predictors. This indicates possible multicollinearity and violates the conditional mean predictor condition.

Table 4: Summary of Preliminary Model Coefficients

	Coefficients Summary			
Predictor	**Estimate**	**Std. Error**	**t-value**	**p-value**
(Intercept)	591032.538	26852.665	22.010	<0.001
Bedroom2	180672.269	9488.996	19.040	<0.001
Bathroom	289515.819	10927.399	26.494	<0.001
Distance	-38917.887	1061.110	-36.677	<0.001
Landsize	25.975	6.678	3.890	<0.001
Typet	-348425.626	21653.640	-16.091	<0.001
Typeu	-416276.230	18012.374	-23.111	<0.001

Note: Significance codes: *** p<0.001, ** p<0.01, * p<0.05

4.2 Discussion of Model Estimates

The preliminary linear regression model indicates that Distance has a significant negative impact on housing prices. This suggests that properties closer to the city center are more valuable, aligning with traditional real estate principles that emphasize central location advantages. Additionally, Type plays a significant role, with houses being valued higher compared to units or townhouses. This is likely because houses offer more land and space, making them more desirable. These results suggest that proximity to urban amenities and property type are key factors in determining property values in Melbourne.

4.3 Comparison with Literature

The findings from the preliminary model are consistent with several points in the literature but also show notable differences. Chen and Hao (2008) identified that in Shanghai, property prices decrease approximately 5% per kilometer away from the CBD, highlighting the importance of location in property valuation, which parallels the negative effect of distance observed in this model . He and He (2021) also found that location and property type are influential predictors of housing prices in Melbourne, similar to the effects seen in this study, confirming the robustness of these factors in different urban settings. However, unlike the typically strong linear fits found in Yusof and Ismail (2012), the residual plots of the current model reveal non-linearity and non-constant error variance. This suggests that a simple linear model might not sufficiently capture the complexities of Melbourne's housing market, whereas the literature often reports better-fitting models using additional adjustments or non-linear relationships . These discrepancies highlight the need for more sophisticated modeling techniques, such as interaction terms or non-linear approaches, to improve the current model's accuracy.

A Reference