DATA 501 Final Report

Predictive Power: Housing Price Model for Calgary's Market

Jason Osmond & David Zhang

April 12, 2024

Table of Contents

1. Introduction

1.1. Motivation

1.1.1. Context
Predicting housing prices in Calgary using variable analysis, in a changing city like Calgary home prices vary depending on which side of the city you reside. This project will attempt to uncover the differences in price caused by characteristics around the city and use that in its predictions.

1.1.2. Problem
When purchasing a property, customers need to consider the most important factors for the region their property is in to gain a holistic idea of the value of the property.

1.1.3. Challenges
The challenge of this problem is figuring out the most optimal weight for each feature in consideration of the quadrant of the city so that we can better understand the dynamic nature between sellers and buyers. As often, listing home prices are not totally based on the characteristics of the house and its surrounding but what the seller believes it should be worth.

1.2. Objectives

1.2.1. Overview
We want to give more power to the consumers by providing a model of regional based housing prices in Calgary. Providing them with a model and our predictions, home buyers will be able to make better purchase and investment decisions for their homes. For scope, our analysis will focus exclusively on single-family homes.

1.2.2. Goals & Research Questions
1. What distinguishing property characteristics hold greater value in different quadrants of Calgary?
2. How do property prices vary across different city quadrants (NW, NE, SW, SE), and are there specific areas showing consistent pricing trends?
3. Considering the absence of historical data, can we build an accurate predictive model based solely on the current listings?

1.3. Methodology

1.3.1. Data
The dataset for this project originates from Re/Max Central, a real estate brokerage, Re/Max uses a Multiple Listing Service (MLS) database, which ensures consistent property data and prices across all brokerages. Data was acquired using a web scraper and includes single-family detached home listings from January 2024, with about 400 current listings. Due to privacy constraints, no historical data is available. The dataset

size complies with the Central Limit Theorem (CLT), allowing us to assume a normal distribution after data cleaning. The dataset includes 9 attributes:

1. **Price (CAD)**: The sale price of the property.
2. **Address**: The location of the property.
3. **Square Footage**: The total square footage of the unfurnished property, including all floors and basements.
4. **Year Built**: The year when the construction of the property was completed, excluding renovations.
5. **Bedrooms**: The total number of bedrooms, without distinction between types.
6. **Style**: The number of stories the property has, simplified to storey counts (except for 4-level split). Bungalow style houses have been considered as 1 storey.
7. **Basement**: Indicates the presence of a basement and its completion status.
8. **Distance to Core KM**: Distance to the Core Shopping Center in downtown Calgary, manually added through Google Maps.
9. **City Quadrant**: Section of the city. i.e. NW, NE, SW, and SE.

1.3.2. Approach

We're using a mixed-method approach for real estate price prediction, segmenting the city into four areas (NW, NE, SW, SE) and applying separate regression analyses. We'll check for linearity, avoid multicollinearity, ensure constant error term variance (homoscedasticity), and verify normal distribution of residuals.

The linear regression insights will guide the inputs for our Multilayer Perceptron (MLP) model, helping us select significant features for accurate predictions. MLPs don't have pre-existing conditions but their performance can be affected by data scale and overfitting. We'll normalize the data, avoid overfitting, and use a slow and small start for initialization and hyperparameter tuning. The activation function will be linear with mean square error loss for this regression task.

1.3.3. Workflow

1. Approach for Data/Visual Analytics Solution (Multiple Linear Regression)
2. Challenging Step - Multilayer Perceptron (MLP) Model
3. Random Forest and Gradient Boost Comparison Models

The implementation of the Multilayer Perceptron (MLP) model is identified as the particularly challenging step in the workflow. If challenges arise during MLP implementation, we'll start with a gradual approach, adjust parameters for effective prediction in each Calgary quadrant, or pivot to an alternative approach if needed.

1.3.4 Workload Distribution

Our team takes a balanced approach to the project, with Jason specializing in data visualization and technical aspects such as coding, and David focusing on the statistical and analytical side. Both team members actively contribute to the data analytics approach, addressing model conditions and diagnostics, and collaboratively analyzing and presenting the findings.

1.4. Contributions

Our project introduces an original solution designed to empower consumers in Calgary with valuable insights for making informed real estate investment decisions. The uniqueness lies in the combination of a multilinear regression model and a tailored Multilayer Perceptron (MLP) model, providing nuanced perspectives on optimal property feature weights across different city quadrants. Our approach can be utilized independently or in conjunction with other sources, offering consumers a comprehensive tool for enhancing decision-making when investing in real estate.

2. Related Works

2.1. Technology Scan

2.1.1. Methodology 1

The Mortgage Sandbox website's analysis of Calgary's real estate market uses a multifactor approach. It ranks the city, presents historical sales data, and examines median and benchmark sale prices using short and long-term data. It categorizes Calgary's districts by directions, covers homeowner anchoring effects, rental market dynamics, and high mortgage rates' impact. It includes visual data like sales versus listings and new house price trends, concluding with insights on new home construction record highs. The analysis, based on historical data and market intelligence, doesn't include individual home features. It infers from broader market conditions, which we believe only explains a portion of Calgary homes' price differences.

2.1.2. Methodology 2

In our second technical scan, we explored existing methodologies for predicting Calgary's housing prices. Paper [2] uses PCA and algorithms to reduce the feature set from 36 to 20, crucial for mitigating overfitting and enhancing model efficiency. Machine learning models, including Decision Tree, Random Forest, and XGboost, were used, with XGBoost achieving the highest accuracy of 69.7%. Predicting Calgary's house prices is challenging due to factors like politics, economy, and social aspects.

2.2. Academic Papers

2.2.1. Housing Price Prediction Based on Multiple Linear Regression

The study employs statistical methodologies, specifically multiple linear regression and the Spearman correlation coefficient, to analyze and predict housing prices in Boston. The Spearman correlation coefficient identifies significant factors influencing housing prices, including the proportion of lower-income groups, property land area, average number of rooms, and proximity to employment zones. A predictive model for housing prices is constructed using multiple linear regression, trained and tested on the Boston housing price dataset. The model effectively predicts housing prices, with consistency observed between predicted and actual prices. However, contradictory findings were noted in the correlation between property price and distance to the city core when compared to another study. The study underscores the potential of Machine Learning algorithms for more accurate predictions by leveraging complex data patterns, with plans to enhance the prediction model using ML.

2.2.2. An Efficient System for the Prediction of House Prices using a Neural Network Algorithm

This study aims to predict house prices using ML algorithms, focusing on deep learning, with data from King County listings. Four models - deep learning neural network, linear regression, KNN, and random forest - are developed and evaluated. Performance analysis shows that deep learning, explaining nearly 89% of the variance, outperforms traditional algorithms. However, predictions for higher-priced houses suggest that different factors may influence their prices.

2.2.3. Reasonable Price Recommendation on Airbnb Using Multi-Scale Clustering

The study proposes a method for recommending Airbnb house prices using multi-scale clustering. The Multi-Scale Affinity Propagation (MSAP) method aggregates houses based on landmarks and facilities, and predicts prices using a Linear Regression model with Normal Noise (LRNN). The methods include Multi-Scale Affinity Propagation Clustering, which aggregates houses considering landmarks and facilities, and Linear Regression with Normal Noise, which predicts house prices considering factors like facility status, distance to landmarks, and landmark popularity. The approach had decent results, with prediction results ranging between 20% to 50% depending on the city. This approach could enhance our model by factoring in nearby parks and amenities, better accounting for location in our prediction model.

3. Workflow

3.2. Approach for Data/Visual Analytics Solution

Possible Implementations:

- **Simple Linear Regression (SLR)**: Considering one variable at a time.
- **Multiple Linear Regression (MLR)**: Incorporating multiple variables.
- **Principal Component Analysis (PCA)**: For dimensionality reduction in case of multicollinearity.

Implementation Decision: MLR was adopted as the primary model structure, with attention to diagnostics like linearity, multicollinearity, homoscedasticity, and normality of residuals using standard plots summary and variance inflation factor function from R. MLR was the best choice for its balance between complexity and interpretability.

3.3. Multilayer Perceptron (MLP) Model with One-Hot Encoding for City Quadrants

Implementation Details with One-Hot Encoding: To accurately model the impact of a property's characteristics with respect to its location in the city on its price, city quadrants (NE, NW, SE, SW) were included as categorical features. Since MLP models require numerical input, these categorical variables were transformed into a format the model could process using one-hot encoding.

Integration into the MLP Model: After applying one-hot encoding to the quadrant data, the resulting binary columns were included alongside other numerical features (like square footage, year built, etc.) as inputs to the MLP model. This approach allowed the MLP to consider both the physical attributes of the properties and their locations within Calgary when predicting prices.

3.4. Random Forest and Gradient Boost Comparison Models

Implementation: We use Python libraries like numpy, pandas, sklearn, matplotlib, and seaborn to predict house prices. The data is preprocessed, standardized, and reduced in dimension using PCA. Both ML models are trained and evaluated using mean squared error and R-squared. This process was the general approach done by some of the research papers mentioned in 2.2..

Optimization: Implement Random Forest model for regression, optimizing hyperparameters such as number of trees, maximum tree depth, and feature selection to enhance predictive performance.

Evaluate: Assess the effectiveness of the model by employing measures such as mean squared error and R-squared. Compare the model's prediction performance with the Multilayer Perceptron (MLP) method.

3.5. Analyze and Present Findings

Variance Inflation Factor (VIF) Analysis: We used VIF to quantify how much the feature variances are inflated due to multicollinearity. A VIF value above 5 as per different threshold, indicates a high multicollinearity that could distort our regression coefficients and their significance.

MLP Model Loss Over Epochs: Plotting the loss value (Mean Squared Error, MSE) for both the training and validation sets over the number of epochs. This visualization should highlight the point at which the model starts to overfit, indicated by the validation loss increasing or plateauing while the training loss continues to decrease.

MLP Best Loss Value Reporting: Highlight the epoch at which the model achieved the best loss value on the validation set. This is critical for understanding the model's performance and the effectiveness of early stopping in preventing overfitting.

4. Results

4.1. Measurement of Success

The primary goal of our analysis was to accurately predict housing prices in Calgary based on a comprehensive set of features and city quadrants, including both quantitative characteristics like square footage and qualitative attributes like the quadrant location. Success was measured by our models' ability to identify significant predictors of housing prices and their precision in price prediction, as reflected by statistical measures such as p-values and the Mean Squared Error (MSE) for predictions.

4.2. Applied Models and Experiments

Multiple Regression Model: This model was utilized to evaluate the influence of various housing attributes on the price. The experiment conducted using this model enabled us to gain a deeper understanding of the key factors that drive housing prices. By analyzing the coefficients of the regression, we were able to quantify the impact of each characteristic on the price, providing valuable insights for both buyers and sellers in the real estate market.

Multilayer Perceptron (MLP) Model: We implemented this type of artificial neural network to predict housing prices using a comprehensive set of features. The MLP model, with its ability to model non-linear relationships and interactions between features, allowed us to further delve into the potential of ML for real estate valuation. Initial implementations struggled to accurately predict houses with a price exceeding $1 million (see Figure E2). We believe these 'luxury houses' are difficult to predict as price is determined by factors beyond the basic characteristics of the property. Thus, houses exceeding $1 million were excluded from the training and testing dataset for all models.

Random Forest Regressor Model: This model was used to predict housing prices by creating a multitude of decision trees at training time and outputting the mean prediction of the individual trees. The Random Forest model helps in handling the bias-variance trade-off and provides a more robust prediction by reducing overfitting.

Gradient Boost Prediction Model: This model was implemented to compare the prediction accuracy of our housing price model. Gradient Boosting operates by building new models that predict the residuals or errors of prior models and then adds them together to make the final prediction. It allows for the optimization of arbitrary differentiable loss functions, making it adaptable and flexible for various datasets.

4.3. Quantitative Results

MLR Results

In the **SW** quadrant, house prices are positively influenced by **square footage** and y**ear built**, but **negatively by distance from the city** (see Figure A2). Each km from the city reduces price by $14,420[1.53e-12] (where [1.53e-12] is the p-value), each additional sqft increases price by $451.7[<2e-16], and each newer year adds $2,870[0.0143] to the price.

In the **SE** quadrant, only **square footage** significantly predicts house prices (see figure B2). Each additional sqft increases the price by $376.33[4.98e-15]. Other factors like year built, bedrooms, style, and distance to core don't significantly influence pricing.

In the **NW** quadrant, **square footage** and **number of bedrooms** significantly predict house prices (see Figure C1 and C2). Each additional sqft increases the price by $526.89[7.95e-12]. However, having 3, 4, or 5 bedrooms negatively impacts the price, with coefficients ranging from -$209,891.25 to -$382,678.96 [<0.05]. Other factors like year built, style, and distance to core don't significantly influence pricing.

In the **NE** quadrant, **square footage** and **number of bedrooms** significantly predict house prices (see Figure D1 and D2). Each additional sqft increases the price by $229.15[<2e-16], and more bedrooms generally lead to higher prices, with coefficients ranging from $44,419.28 for 3 bedrooms to $140,583.60 for 7 bedrooms[<0.05]. **House styles** have mixed effects, with Style3 significantly reducing prices by -$68,478.84[0.011882]. Other factors like year built and distance to core don't significantly influence pricing.

Using VIF to test multicollinearity, we found all data was within an acceptable range (less than 5). Thus, there are no significant signs of collinearity in the data (see Tables A1, B1, C1 and D1).

MLP Results

The MLP training process spanned around **30,000 epochs**. Notably, as we approached this number of epochs, we observed that the MLE test loss began to stabilize around **55 billion** (See Appendix E1). This stabilization is a critical indicator that our model reached a point where further training would likely yield minimal improvements in predictive accuracy on our test data. This finding is significant as it suggests that the chosen architecture and hyperparameters are appropriate for our dataset, as evidenced by the convergence of loss.The 30,000-epoch mark was selected as a stopping point for training, helping us avoid overfitting while maximizing computational efficiency.

After removing listings valued over $1,000,000, we achieved the following results for each of our prediction models.

Prediction Model	MSE (billion)	RMSE	Accuracy within $20,000	Accuracy within $50,000
Multi-Layer Perceptron	44	$209,761	15%	48%
Random Forest	7.8	$88,822	18%	49%
Gradient Boost	8.3	$91,118	16%	52%

Table 1: Results from our three prediction models. Does not include listings >$1 million.

The prediction results, shown in Table 1, indicate our MLP model is underperforming when compared to more traditional ML approaches. Random Forest Regressor has the best overall performance with the lowest MSE and highest accuracy within $20,000. The Gradient Boost model achieved the highest accuracy within $50,000.

4.4. Analysis of Results

4.4.1. What Our Multiple Linear Regression Told Us

From our MLR models, we learned which features of a house tend to increase or decrease its price. For instance, having more square footage generally raises a house's price—a bigger house costs more. However, not every feature we thought might matter turned out to have a clear impact. Sometimes, factors like the year a house was built and even the size of house didn't show a strong enough connection to price.This is evidence that the difference in pricing based on market demand is drastically different for some quadrants.

4.4.2. Our MLP Model's Performance

Moving to the MLP model, we ventured into predicting house prices using a neural network, which is a bit like a very complex math puzzle that learns from lots of data to make predictions. The test loss, a measure of how far off the model's predictions were from the actual prices, stabilized around 55 billion MSE. This number might sound huge, but in the context of ML and considering the scale of house prices, it shows us the model is not doing too badly. Stabilization means our model reached its learning potential with the given data and setup. This is also the overall MSE over the entire price range, from the graphical display of Actual vs predicted housing price. The model is much better at predicting prices in the price range of 500,000 CAD to 900,000 CAD. After removing the listing exceeding $1 million, MSE was reduced to $44 billion, a significant improvement. When comparing the predictions with the random forest regressor and the gradient boost models, the MLP slightly underperformed in all categories (see Table 1) . From these comparisons, we can conclude the MLP model, while not significantly behind other models, is inferior to the random forest and gradient boost models.

4.4.3. Summarizing Our Journey

Throughout this analysis, we've combined traditional statistical methods and modern ML to understand and predict housing prices. Statistical significance helped us identify which features of houses are truly important for their prices. The MLP model showed us that with enough data and careful training, we can predict housing prices to a somewhat accurate degree. Each step of the way, from understanding what makes a house more or less expensive to training a computer model to predict those prices, has brought us insights into the complex world of real estate values using housing characteristics.

5. Discussion

5.1. Overall, is this approach promising?

Our approach shows promise in effectively combining both traditional statistical methods and modern ML techniques to predict housing prices in Calgary. By employing multiple regression analysis we were able to identify the primary influencers in our data and apply those features in our MLP model. Our objective was to develop an effective model to predict housing prices based on property characteristics. Currently, our prediction's RMSE is ~$230,000, which is insufficient and will need improvement.

5.2. What different approaches are better?

Our robust approach could be enriched by exploring alternative methodologies. Studies like Patankar's [5] suggest more advanced deep learning techniques [5] could enhance predictive accuracy. Despite potential improvements, achieving drastically better results may still be challenging, as any model will have difficulty making accurate predictions with the limited information.

5.3. What follow-up work should be done next?

Based on the inaccuracies of the model we can determine there is more work needed to improve the predictive power of our models. We believe this is because we are missing vital information needed to make a strong prediction. Incorporating additional property features in a meaningful way should help us achieve better results. One such property feature we would like to explore would be location. Location is a major influencer of property value and including more data about the location of each property we expect will improve the model's accuracy. Continually adding more data over time should also improve accuracy and counteract seasonality of the data.

5.4. What did we learn by doing this project?

Creating a house price prediction model was challenging. We gained insights from implementing a Multi-Layer Perceptron (MLP) algorithm, refining our understanding of Multiple Linear Regression (MLR) for feature selection, and applying theoretical concepts to practical challenges.

Conclusions

This project serves as a platform for both academic and professional growth, offering the opportunity to deepen our understanding of Real Estate and its management, particularly in the context of Calgary. Through the practical application of analytical models, such as multilinear regression and Multilayer Perceptron (MLP), we aim to gain insights into regional dynamics and factors influencing property prices. As individuals and potential homebuyers, this knowledge will directly benefit us by providing valuable insights into the local housing market, enhancing our ability to make informed investment decisions. Additionally, the project enhances our collaborative project management and communication skills, bridging the gap between academic insights and real-world applications in the dynamic field of real estate.

References

Mortgage Sandbox. (n.d.). Real Estate Trends and Price Forecast - Metro Calgary. Mortgage Sandbox. Retrieved February 5, 2024, from https://www.mortgagesandbox.com/calgary-real-estate-forecast

Di Benedetto, D., Gao, L., Huang, Y., Sharma, N., & Wang, D. (2018). HOUSING PRICE PROJECT REPORT. m2pi. Retrieved from https://m2pi.ca/project/2020/bc-financial-services-authority/BCFSA-final.pdf

Re/Max Central. (n.d.). Official Website. Re/Max Central. Retrieved January 31, 2024, from https://www.remaxcentral.ab.ca/

Zhang, Q. (2021). Housing Price Prediction Based on Multiple Linear Regression. *Scientific Programming*, 2021, Article 7678931. https://doi.org/10.1155/2021/7678931

Patankar, N. S., Vengatesan, K., Deshmukh, Y. S., Chintamani, R. D., & Shelake, N. L. (2022). An Efficient System for the Prediction of House Prices using a Neural Network Algorithm. In *2022 International Conference on Edge Computing and Applications (ICECAA)* (pp. 1102-1106). IEEE. https://doi.org/10.1109/ICECAA55415.2022.9936419

Li, Y., Pan, Q., Yang, T., & Guo, L. (2016). Reasonable price recommendation on Airbnb using Multi-Scale clustering. In *2016 35th Chinese Control Conference (CCC)* (pp. 7038-7041). IEEE. https://doi.org/10.1109/CHICC.2016.7554467

Osmond J., Zhang, D., (2024). Predictive Power: Housing Price Model for Calgary's Market [Data set]. https://github.com/TheJasonOsmond/Calgary-Property-Price-Prediction-Model

Appendix

A. Results of Linear Regression for the South West (SW) Quadrant

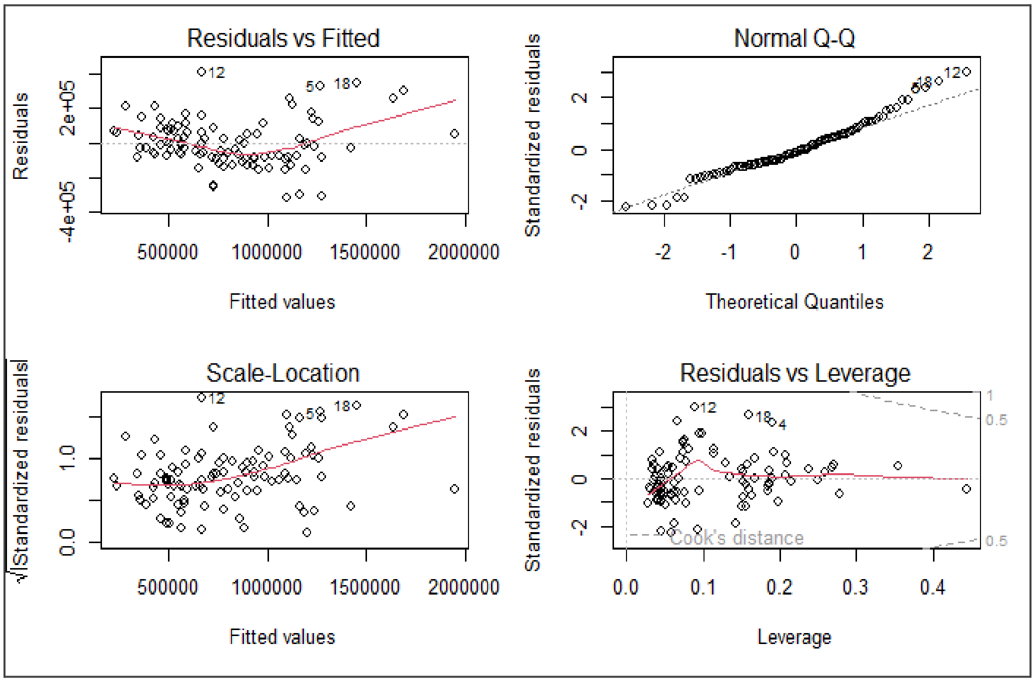

Figure A1: Standard assumption plots (SW).

	GVIF	DF	GVIF^(1/(2*Df))
Square.Footage	1.997236	1	1.413236
Year.Built	2.020444	1	1.421423
Bed.Rooms	2.522716	4	1.122622
Style	3.186715	3	1.213083
Distance.to.City.Core.KM	1.409247	1	1.187117

Table A1: VIF table for multicollinearity (SW).

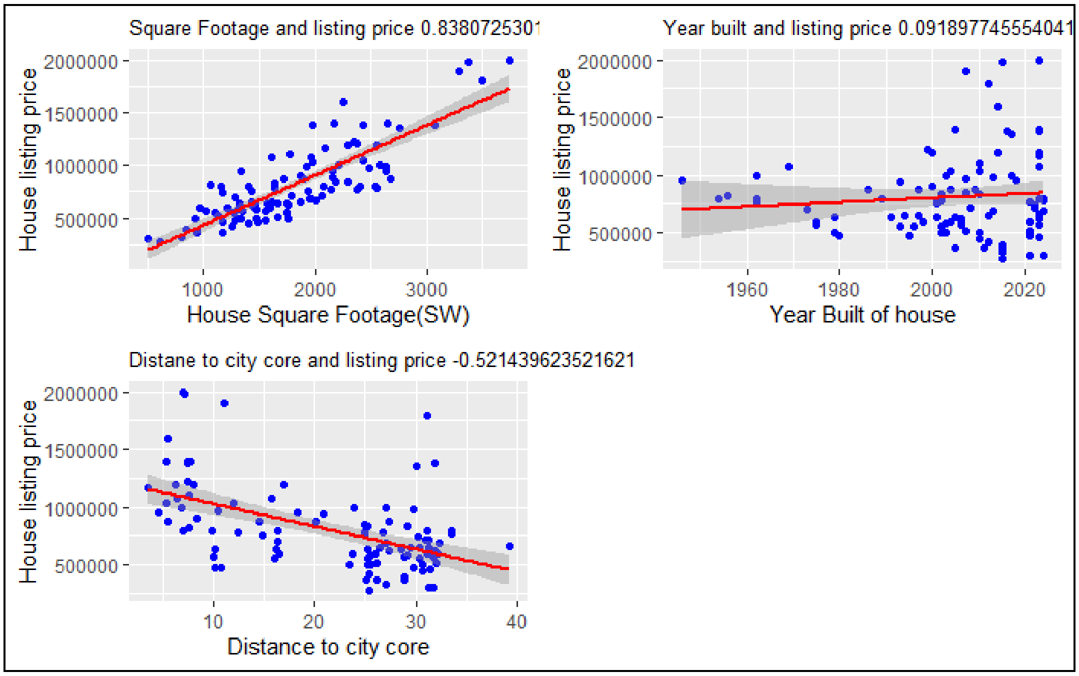

Figure A2: Numerical vs response variables scatter-plot (SW).

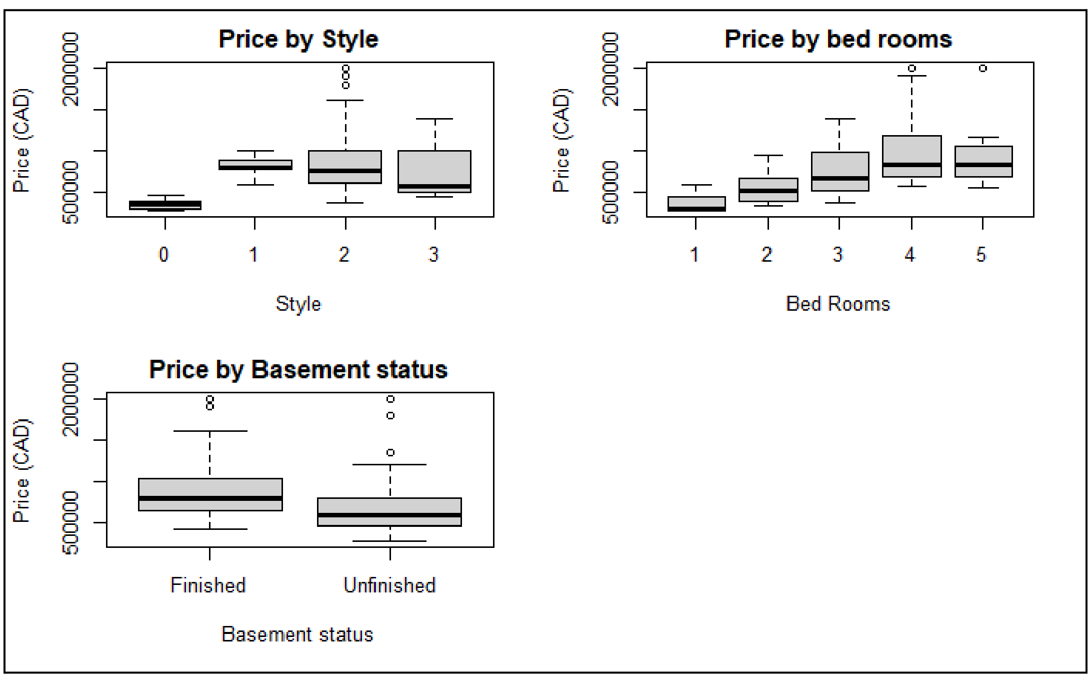

Figure A3: Categorical vs response variables (SW).

B. Testing Assumptions for Linear Regression for the South East (SE) Quadrant

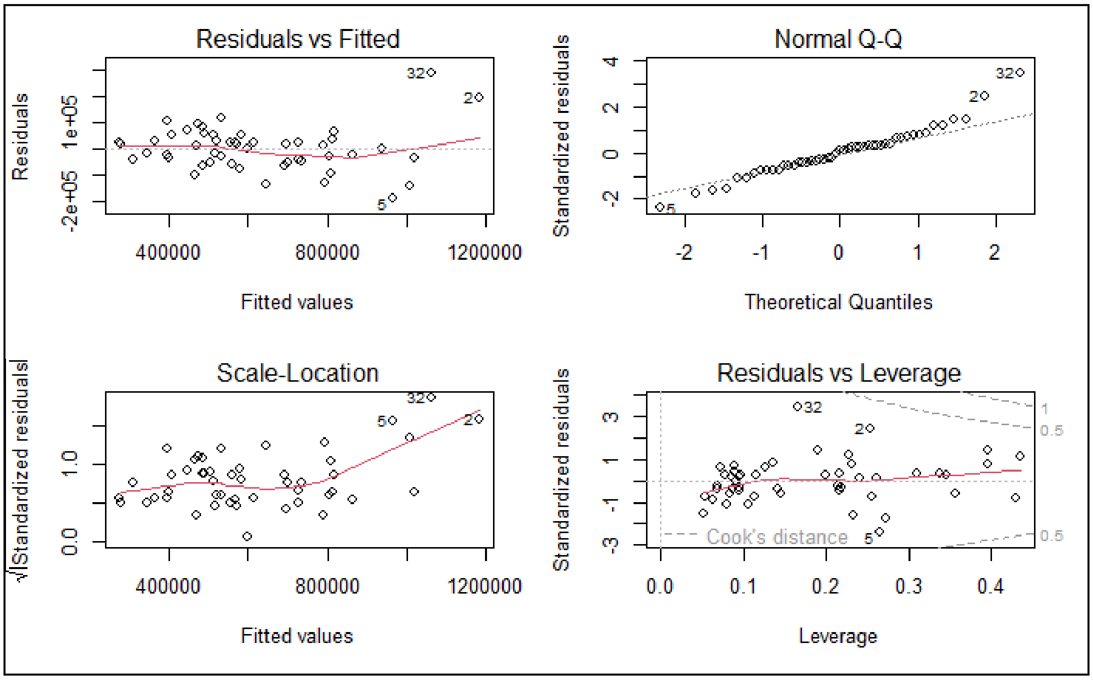

Figure B1: Standard assumption plots (SE).

	GVIF	DF	GVIF^(1/(2*Df))
Square.Footage	2.112629	1	1.453489
Year.Built	4.277786	1	2.068281
Bed.Rooms	4.428224	4	1.204422
Style	3.818366	2	1.397878
Distance.to.City.Core.KM	3.705221	1	1.924895

Table B1: VIF table for multicollinearity (SE).

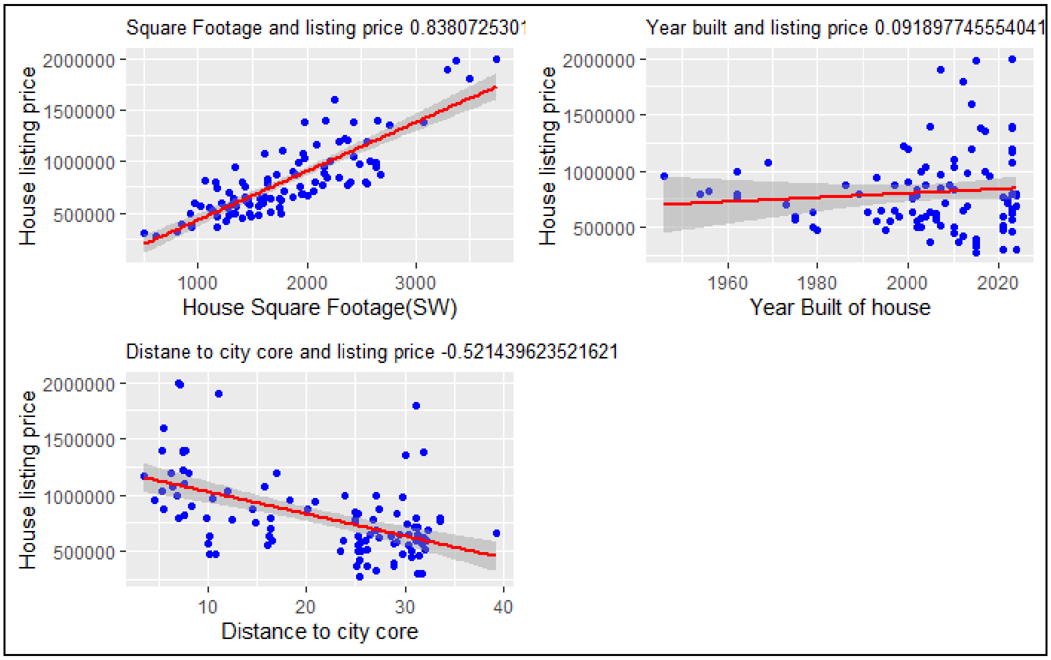

Figure B2: Numerical vs response variables scatter-plot (SE).

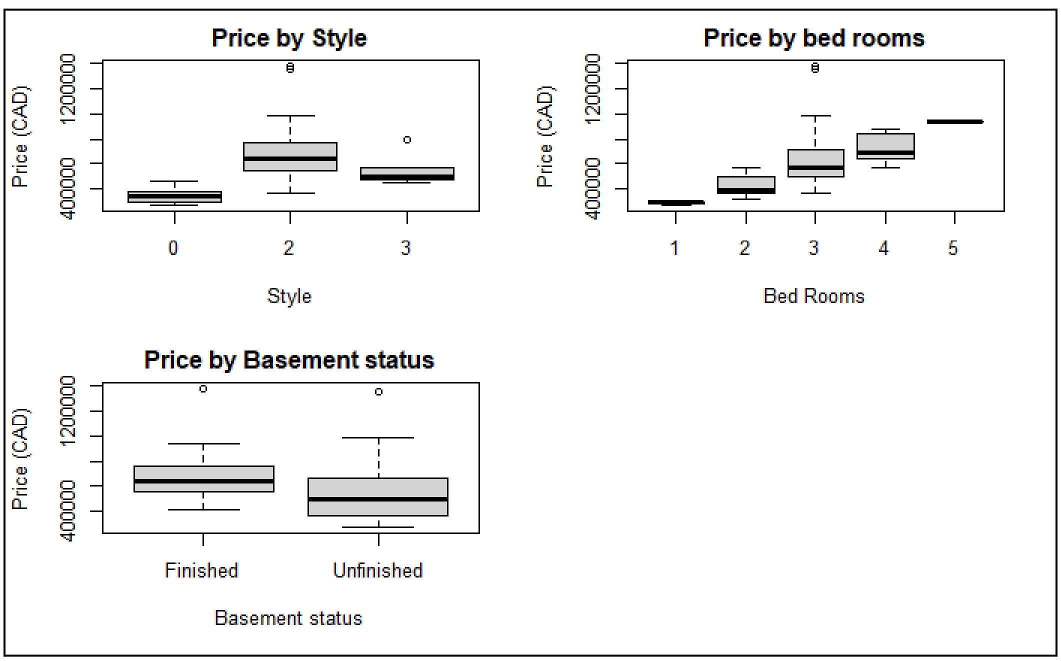

Figure B3: Categorical vs response variables (SE).

C. Testing Assumptions for Linear Regression for the North West (NW) Quadrant

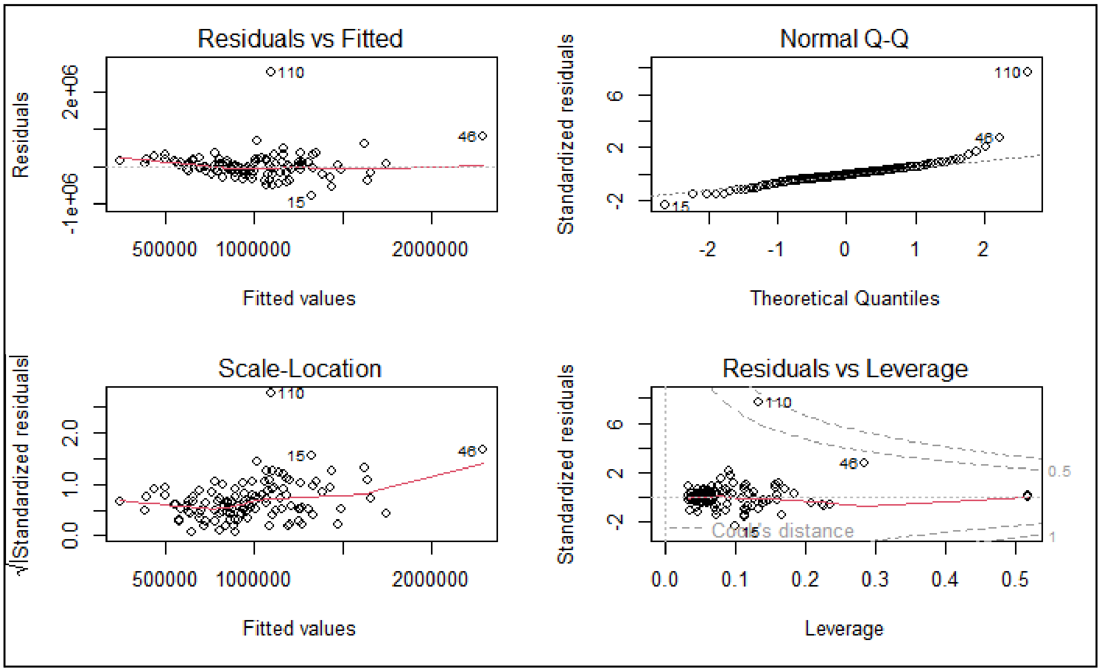

Figure C1: Standard assumption plots (NW).

	GVIF	DF	GVIF^(1/(2*Df))
Square.Footage	2.129481	1	1.459274
Year.Built	1.990494	1	1.410849
Bed.Rooms	1.816594	4	1.077475
Style	2.002222	3	1.122670
Distance.to.City.Core.KM	1.328154	1	1.152456

Table C1: VIF table for multicollinearity (NW).

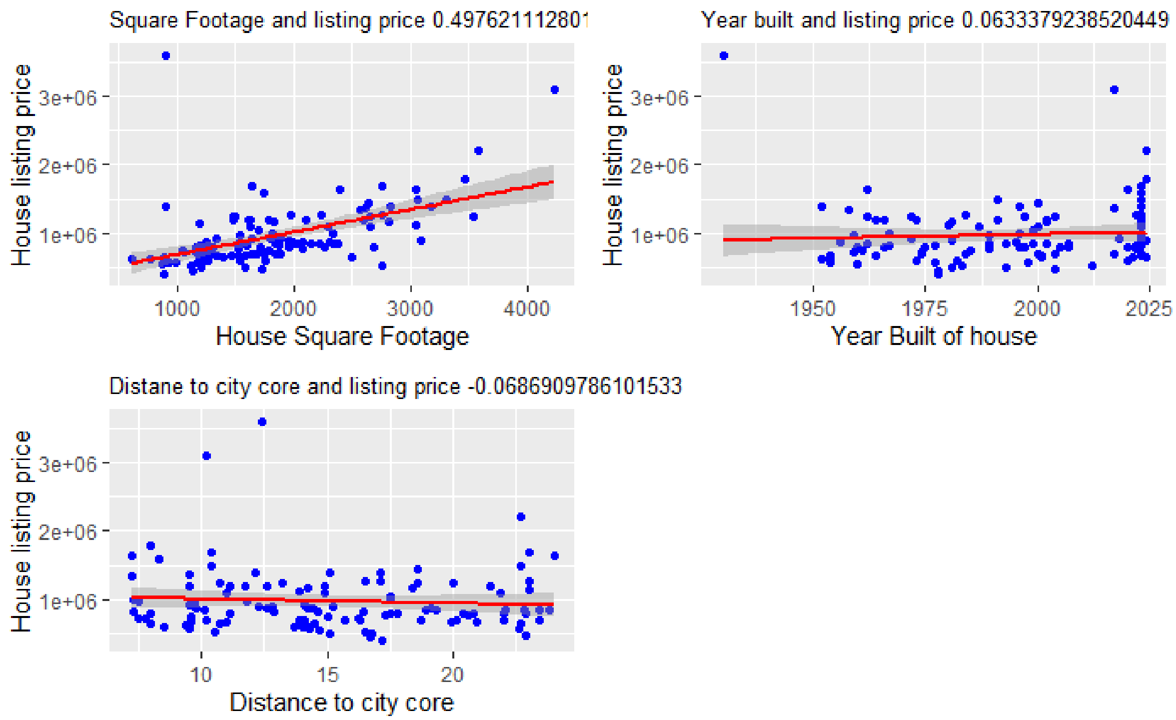

Figure C2: Numerical vs response variables scatter-plot (NW).

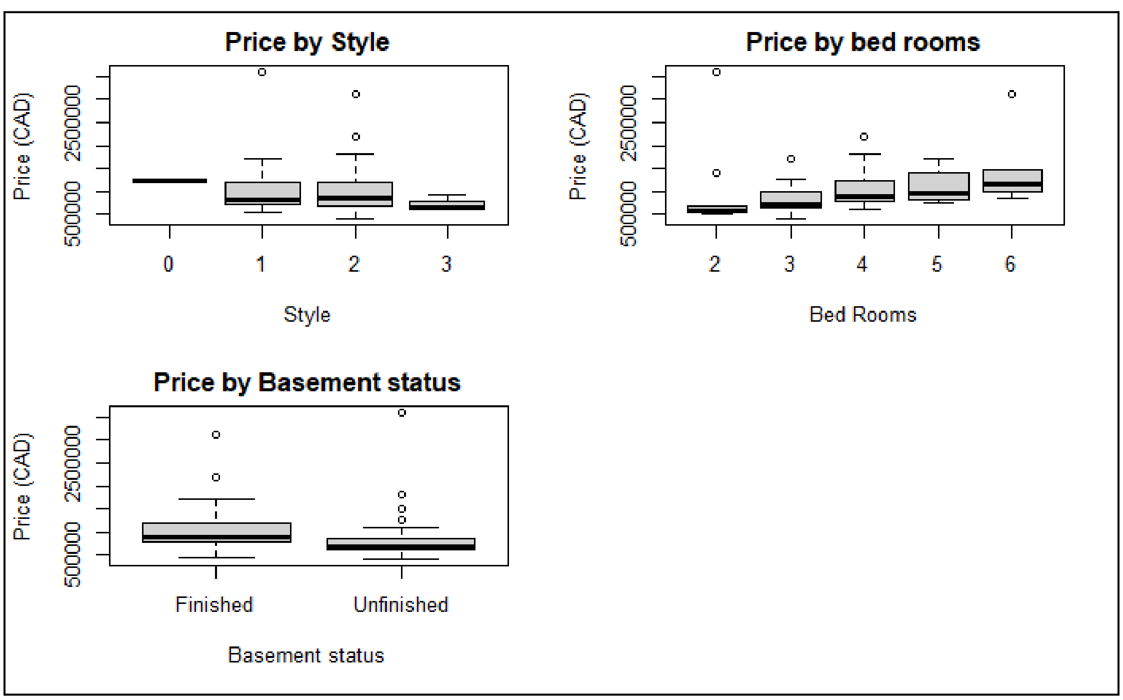

Figure C3: Categorical vs response variables (NW).

D. Testing Assumptions for Linear Regression for the North East (NE) Quadrant

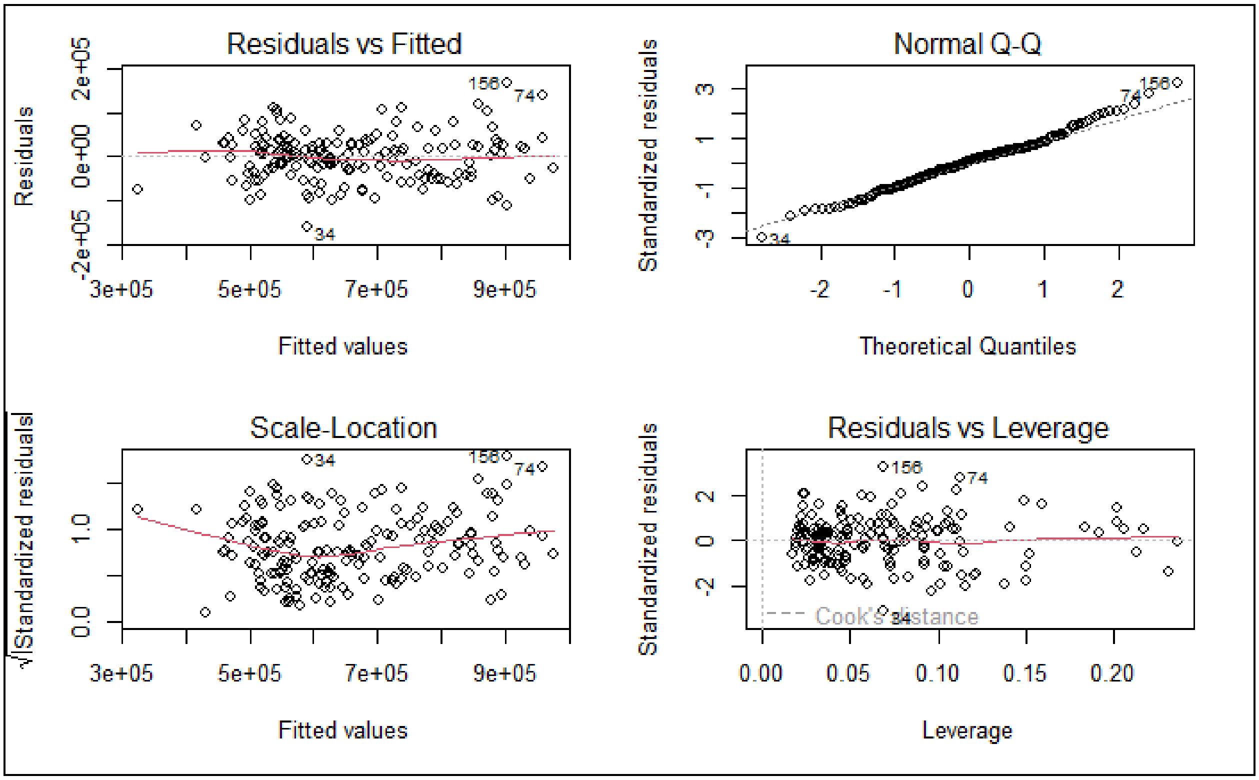

Figure D1: Standard assumption plots (NE).

	GVIF	DF	GVIF^(1/(2*Df))
Square.Footage	2.476802	1	1.573786
Year.Built	5.330436	1	2.308774
Bed.Rooms	1.627686	5	1.049922
Style	2.035701	3	1.125777
Distance.to.City.Core.KM	4.240820	1	2.059325

Table D1: VIF table for multicollinearity (NE).

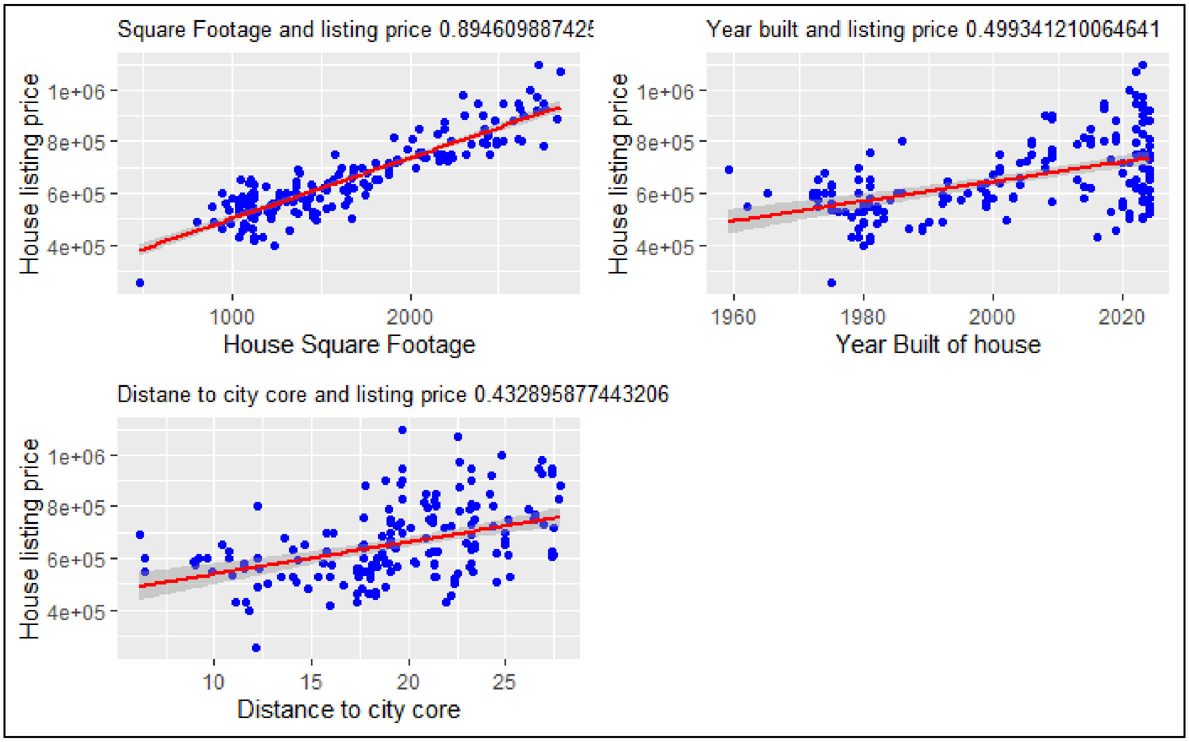

Figure D2: Numerical vs response variables scatter-plot (NW).

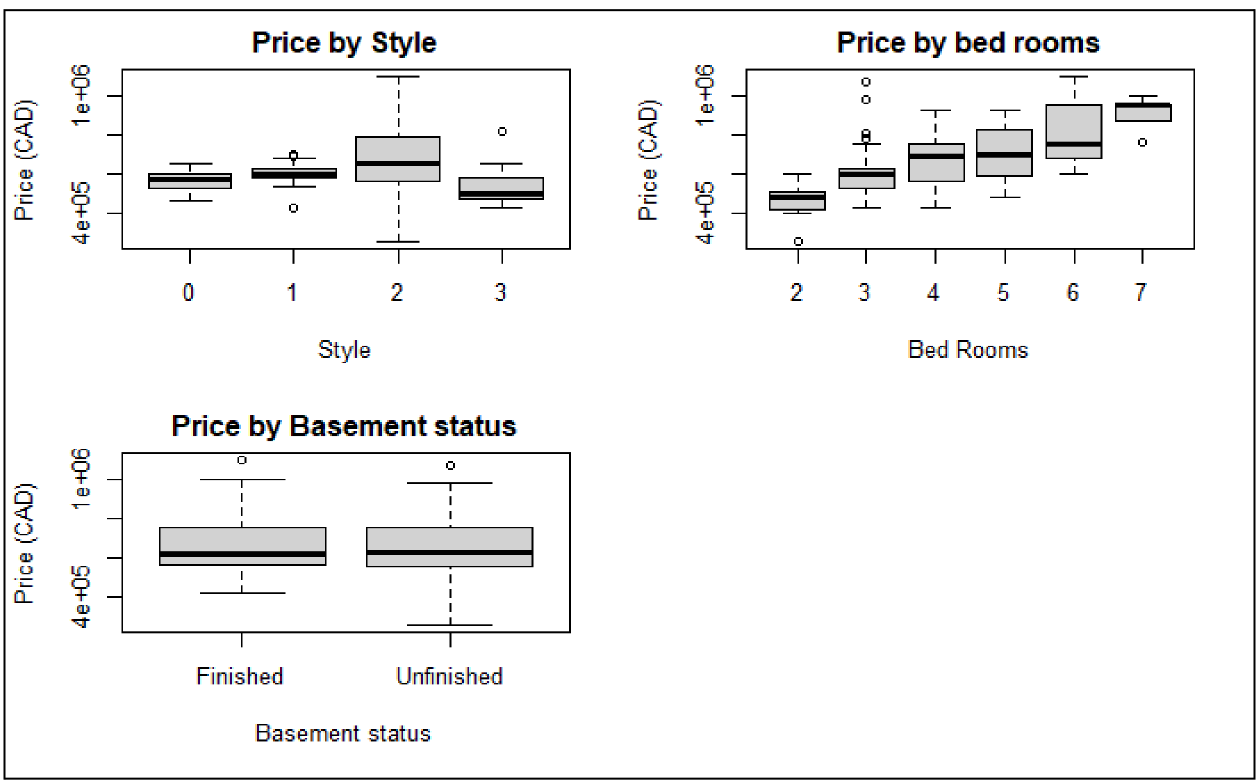

Figure D3: Categorical vs response variables (NW).

E. MLP Results

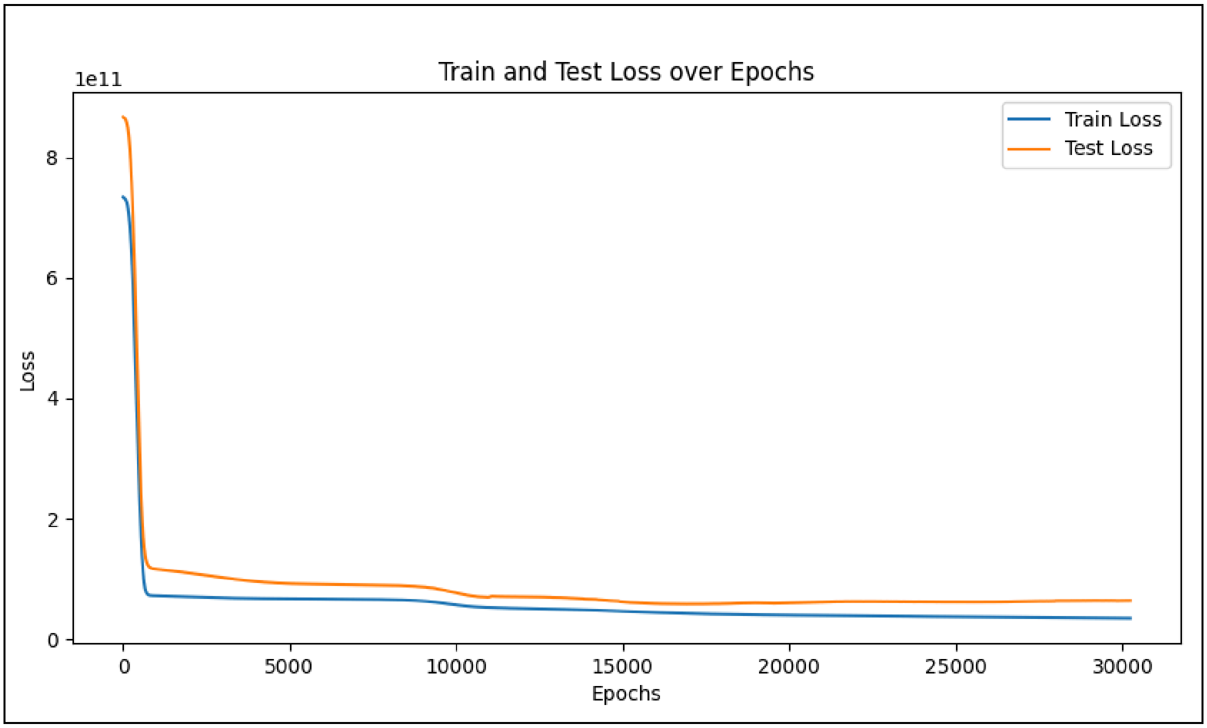

Figure E1: Plot of loss against the numbers of training epoch

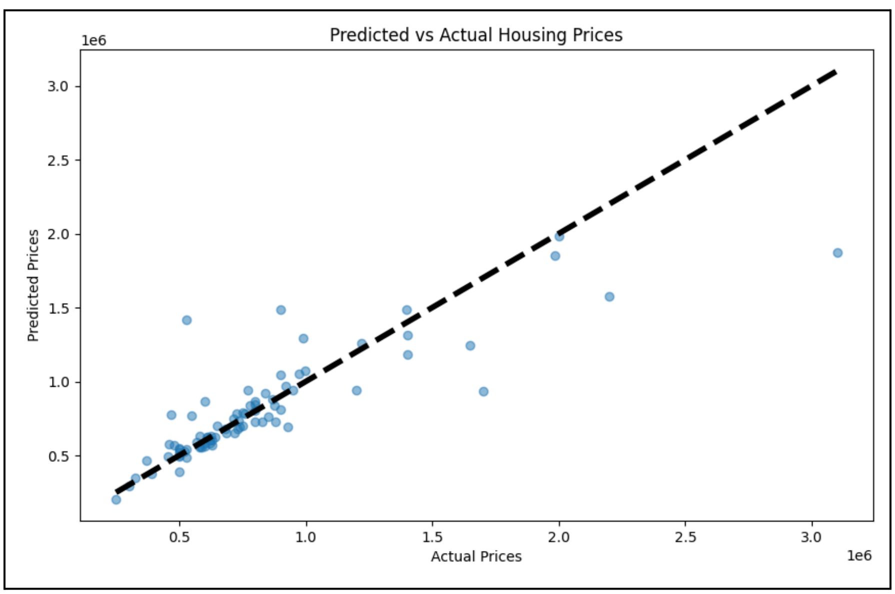

Figure E2: MLP Predictions (including listings over $1 million)

This scatter plot illustrates the relationship between actual and predicted housing prices, where points clustered along the dashed line indicate accurate predictions, and deviations from the line reflect prediction errors.

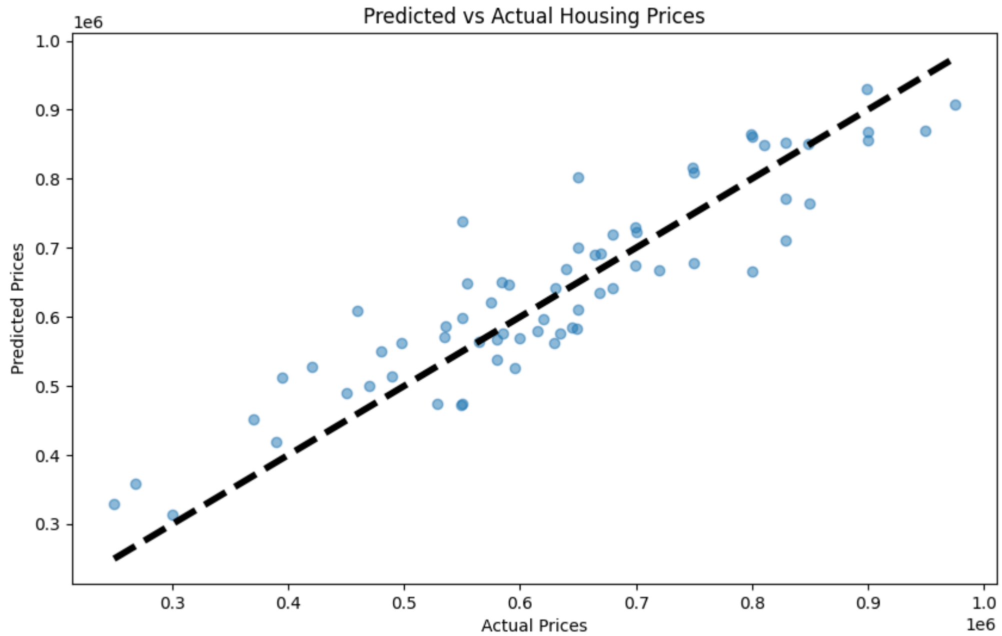

Figure E3: MLP Predictions (excluding listings over $1 million)

F. MLR Results

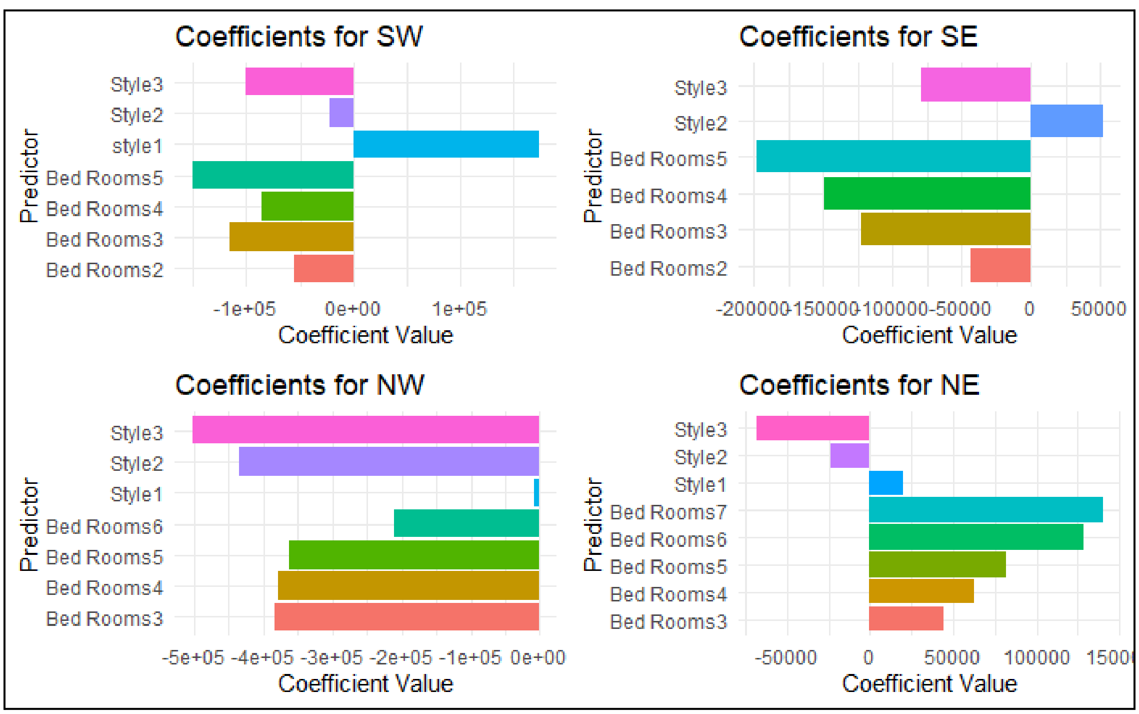

Figure F1: Coefficients for all quadrants for categorical predictors

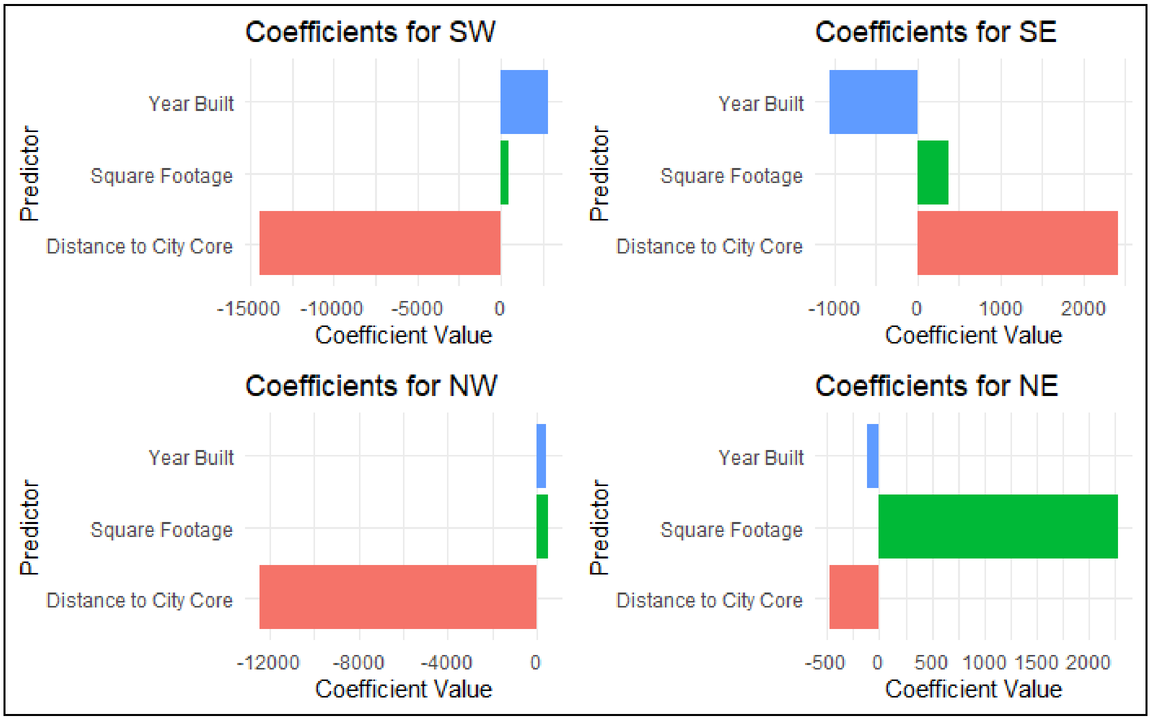

Figure F2: Coefficients for all quadrants for numerical predictors

G. House Price Graph

Figure G1: A box plot of property prices in each quadrant

H. Comparison Prediction Models

Figure H1: Random Forest Regressor Predictions (excluding listings over $1 million)

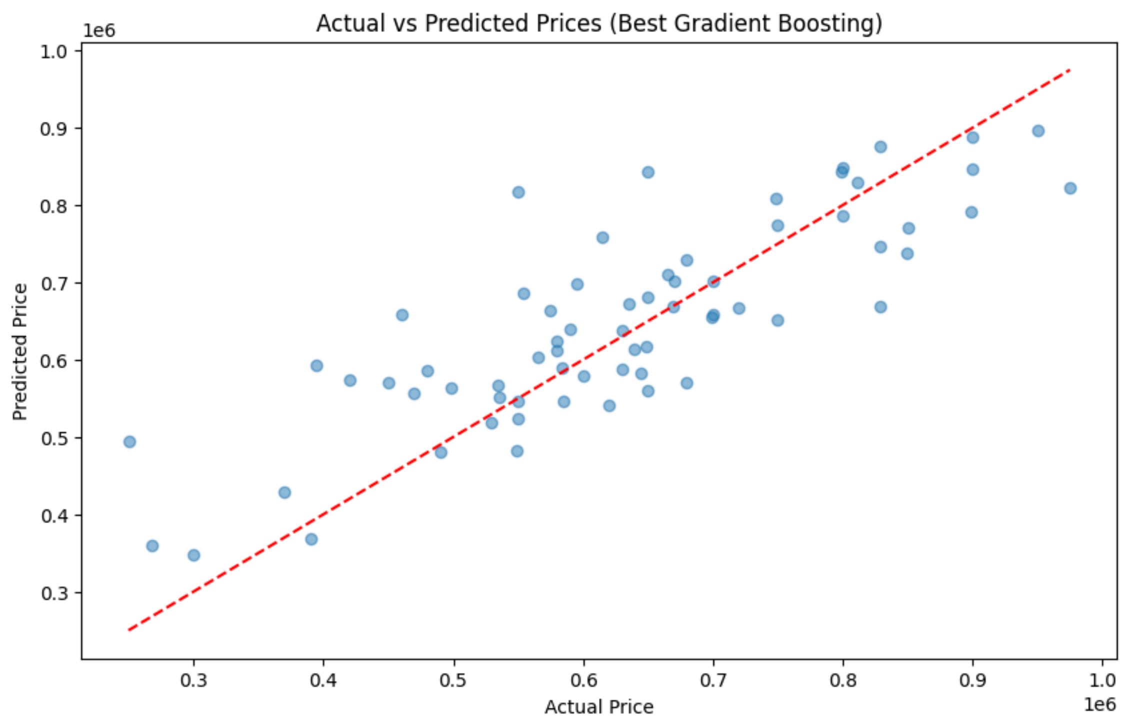

Figure H2: Gradient Boost Predictions (excluding listings over $1 million)