# STAT 306 Group Project Final Report

Group #: B3

**Name:**
Yoonha Jeon (42791335)
Vedant Kalyani (15907850)
Shivang Kinra (41645508)
John Chen (30288310)

# Introduction

### 1.1 Motivation and Research Question

Recently, the Canadian government unveiled a comprehensive action plan in response to the nation's highest inflation rate and ongoing concerns about housing affordability. Pamela Heaven highlighted that Canada is grappling with its highest inflation rate in 30 years, reaching 4.8% (Heaven, 2023). This significant development has sparked renewed discussions about property pricing dynamics. Building upon this idea, our project aims to uncover the factors influencing property prices while considering the confounding effects of other variables. The question we pose for our project is the following:

*"What are the key determinants of property prices in the current real estate market, focusing on property characteristics and location, and what kind of impact does each factor have?"*

For our data analysis, we opted to utilize the Housing Price & Real Estate – 2023 dataset from Kaggle, renowned for its reliability and extensive user community. This dataset was collected by scraping data from a real estate website containing publicly accessible housing listings with confirmed addresses, offering a comprehensive collection of property data including address, price, house description, location, bedrooms, bathrooms, square footage, and the website where the data of the specific house was collected.

### 1.2 Characterization of the Raw Data

```
> head(homes_data)
                        Address  Price      Description          Place Beds Bath SqFt
1               "3704 42 St SW" 979999 "CA  AB T3E 3N1"    "Glenbrook"    4  3.5 1813
2     "30 Mahogany Mews SE #415" 439900 "CA  AB T3M 3H4"     "Mahogany"    2  2.0 1029
3     "273 Auburn Shores Way SE" 950000 "CA  AB T3M 2E9"   "Auburn Bay"    4  2.5 2545
4           "235 15 Ave SW #404" 280000 "CA  AB T2R 0P6"     "Beltline"    2  2.0  898
5 "24 Hemlock Crescent SW #2308" 649000 "CA  AB T3C 2Z1" "Spruce Cliff"    2  2.0 1482
6       "591 Aboyne Crescent NE" 434900 "CA  AB T2A 5Y7"    "Abbeydale"    6  2.0 1059
                         Website
1       "Century 21 Bravo Realty"
2 "Century 21 Bamber Realty Ltd."
3                    "Exp Realty"
4   "RE/MAX Realty Professionals"
5                        "Charles"
6          "Babych Group Central"
```

The dataset consists of various columns, including:

- Address (Categorical): The physical location of the property listed for sale.
- Price (Numerical in CAD): The monetary value attached to the property in Canadian dollars, indicating the asking price set by the seller.

- Description (Categorical): The unique attributes and characteristics of the house, offering essential information to potential buyers about the property's features and amenities.
- Place (Categorical): The geographical area or neighbourhood where the property is situated, providing context about the surrounding environment and community amenities.
- Beds (Numerical): The number of bedrooms within the house, providing details about the accommodation capacity and sleeping arrangements available within the property.
- Bath (Numerical): This continuous variable represents the count of bathrooms present in the house, detailing the facilities available for personal hygiene and convenience.
- Sq.Ft (Numerical in ft^2): The total square footage or area of the property, measured in square feet, offering insights into the size and spatial layout of the house.
- Website (Categorical): The official online platform or webpage where the property listing can be accessed.

# Data Analysis

### 2.1 Data Cleaning

Our dataset contains a large volume of data that concerns a large number of house listings and variables. However, there are many issues within the data that we must address before conducting any statistical analysis. The first stage of pre-processing is conducted in Excel. Automated filtering tools are used to identify and remove duplicate entries. Manual combing identified the presence of many more duplicates which we then removed. We corrected inconsistencies in formatting to facilitate the subsequent stages of analysis. This was carried out manually to ensure that the dataset contained only relevant and correct entries, allowing for a solid basis on which to carry out analysis. Eventually, we finished with a dataset of 999 homes. To process the data into a useful format, we began by scaling the price by a factor of 10e-5.

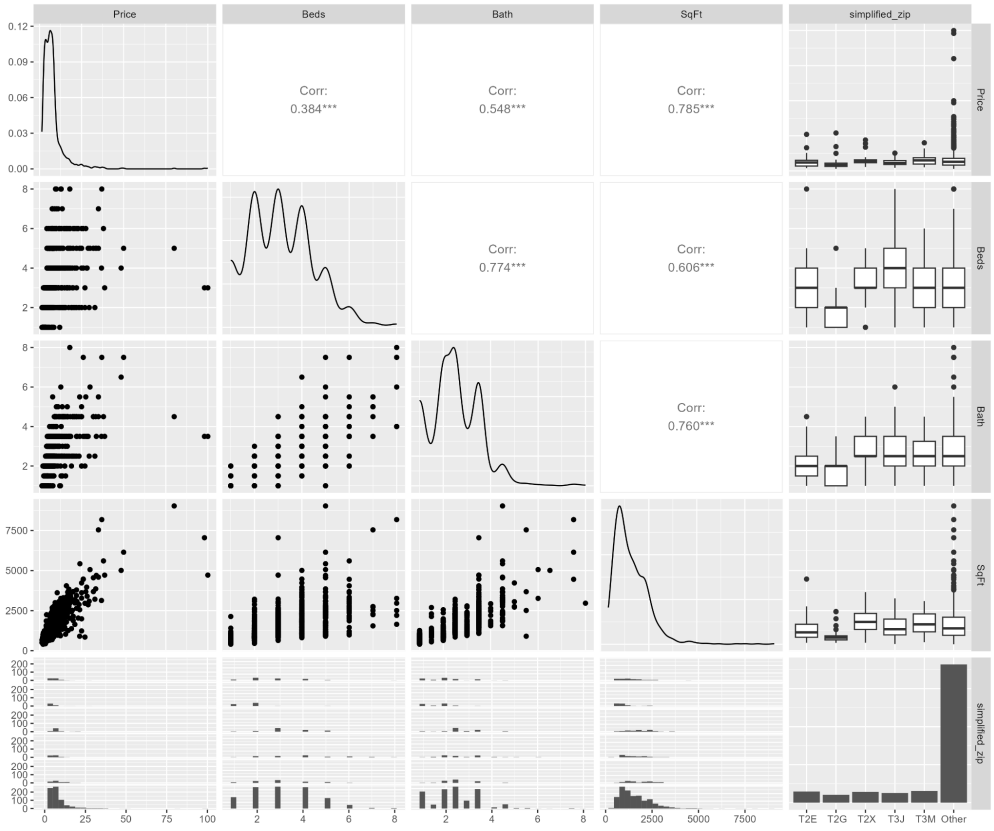
Our explanatory variable had three categorical variables with many levels within each. To simplify our data, we derived a forward sortation area (FSA) variable from the postal code found in the categorical variable Description. Rather than the full ZIP, which contains too many unique values, the FSA provides larger and more meaningful categories, avoiding the potential issue of having a large number of statistically insignificant indicator variables, while still offering a degree of granularity to the analysis. The FSA is given by the first three alphanumeric characters of a postal code (found after CA AB in Description) and points to a specific geographic area of Alberta. This allows us to use our new FSA variable, simplified_zip, as our proxy for location, rather than Description or Place. To further simplify the number of categorical variables we had, we eliminated the Website variable as well. This variable has little relevance to the aim of our study, which is to study trends within the housing market as a whole rather than comparisons between different listing sites, and including this variable would greatly increase the complexity of a model by potentially introducing a large number of categories and consequently a large volume of dummy variables. Last, we picked the 5 most recurring FSA's from our data and created dummy variables from them, leaving one last dummy variable "Other" for the remaining, less frequently seen FSA's. This made our data much easier to work with and use for modelling.

Therefore, the beta coefficient for each 'simplified_zip' category measures the difference in housing prices compared to houses in the baseline zip code category. In line with the one-hot encoding method, we can interpret the following variables as indicators of the location:

simplified_zip_T2G {0, 1} - 1 if the property is in the T2G area, 0 otherwise.
simplified_zip_T2X {0, 1} - 1 if the property is in the T2X area, 0 otherwise.
simplified_zip_T3J {0, 1} - 1 if the property is in the T3J area, 0 otherwise.
simplified_zip_T3M {0, 1} - 1 if the property is in the T3M area, 0 otherwise.
simplified_zip_Other {0, 1} - 1 if the property is in any area not covered by the top 5 zip codes, 0 otherwise.

When all the simplified_zip dummy variables are 0, it indicates that the property belongs to the reference zip code category used in our model (T2E). This categorical treatment of geographic data allows us to assess the impact of location on housing prices while controlling for other influential factors such as size, number of bedrooms, and number of bathrooms.

### 2.2 GGPAIRS/Scatterplots to Explore Relationships Between the Variables



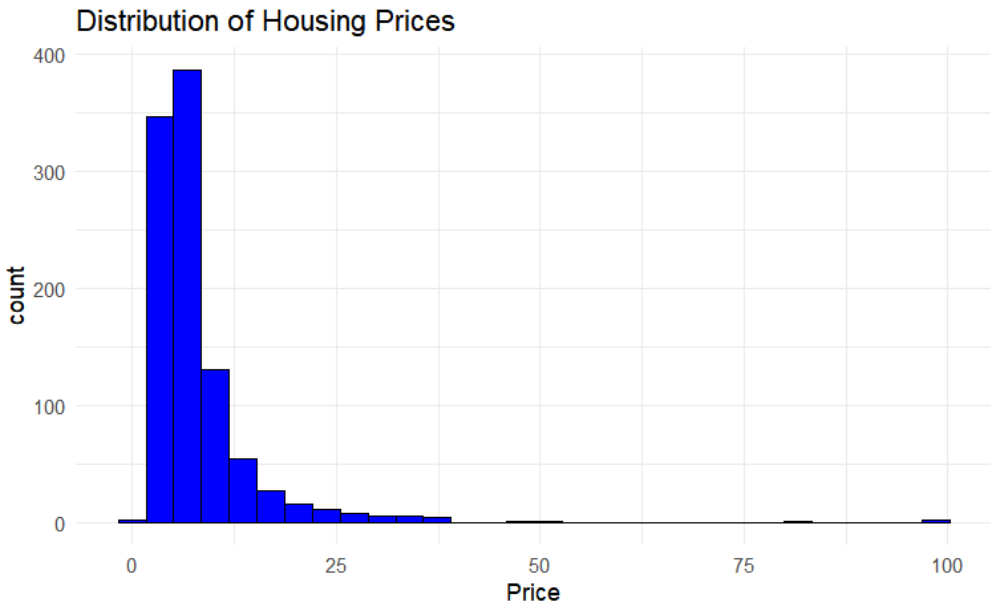

The correlation between some of the explanatory variables exhibits a notably strong correlation, particularly evident among Beds, Bath, and SqFt. This association is unsurprising given the

practical context of our analysis. In a typical home, the number of rooms, including bedrooms and bathrooms, tends to increase with the size of a home, represented by its square footage. Moreover, an increase in the number of bedrooms typically corresponds to a greater number of occupants, thereby necessitating a proportional increase in bathrooms. This correlation is indeed reflected in the data, where we observe a notably high correlation between the number of bedrooms and bathrooms. To address the issue of multicollinearity within our model, we incorporate an interaction term between Beds and Baths in the comprehensive model.

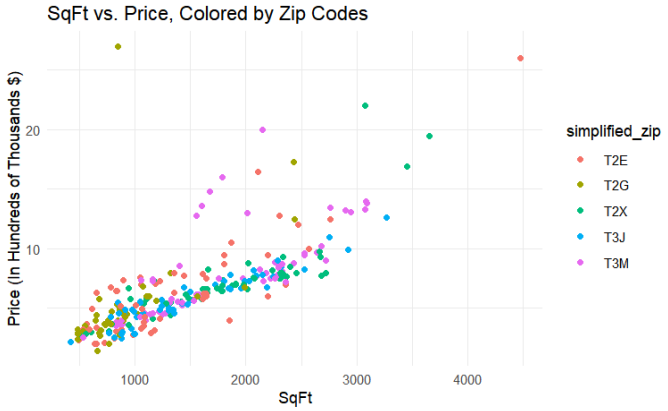
Our full model is thus defined: Price ~ Bed * Bath + SqFt + simplified_zip. These variables include the number of bedrooms (Bed), the number of bathrooms (Bath), the interaction between bedrooms and bathrooms (Bed × Bath), the square footage of the property (SqFt), and the forward sortation area (simplified_zip).

### 2.3 Data visualization in the form of scatter plot and bar graph

In this section, we examine several figures from our data.




In the figure above, we examine the full dataset of our response variable of Price. We note that house prices have a right-skew distribution with the vast majority of listings having a price of under $25 million CAD, but some reaching as much as the 100 million range.

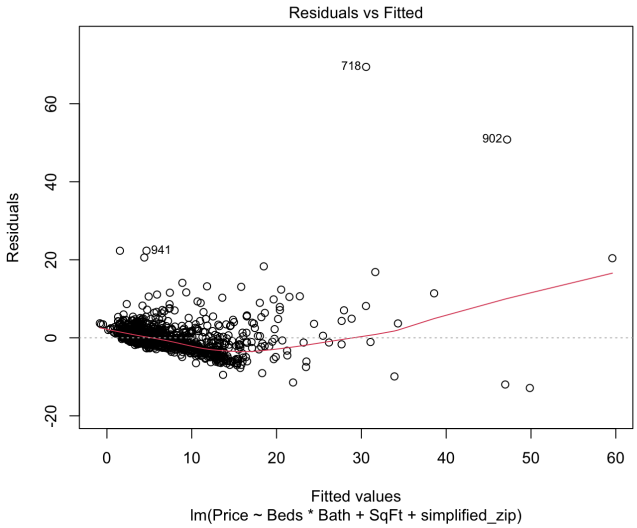



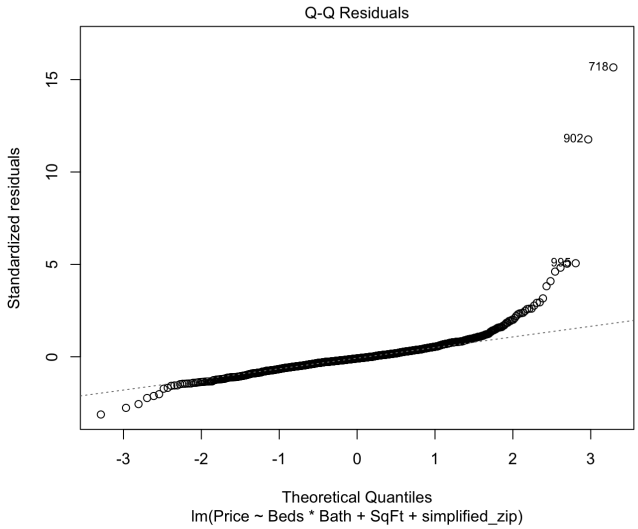

The two figures above are scatterplots illustrating the relationship between area (SqFt) and Price (in 100s of CAD), with points segmented by forward sortation area (Forward sortation area/simplified_zip). The first graph includes house listings from the five most common forward sortation areas as well as other postal codes, while the second graph includes only the former. In both plots, we observe an upward trend in price as square footage increases. However, it is difficult to determine from visual inspection alone whether there is any relationship between forward sortation area and price, whether or not we limit analysis to only the five most common, or if the “Other” category is included.

### 2.4 Model Fitting: The Linear Model

This full model:
Price = -1.8096 - 0.5242(Beds) + 0.2190(Bath) + 0.0071(SqFt) + 1.3449(simplified_zip_T2G) - 2.0281(simplified_zip_T2X) - 0.6501(simplified_zip_T3J) - 1.2368(simplified_zip_T3M) + 0.3559(simplified_zip_Other) - 0.0732(Beds:Bath)

## Residual and QQ-plot

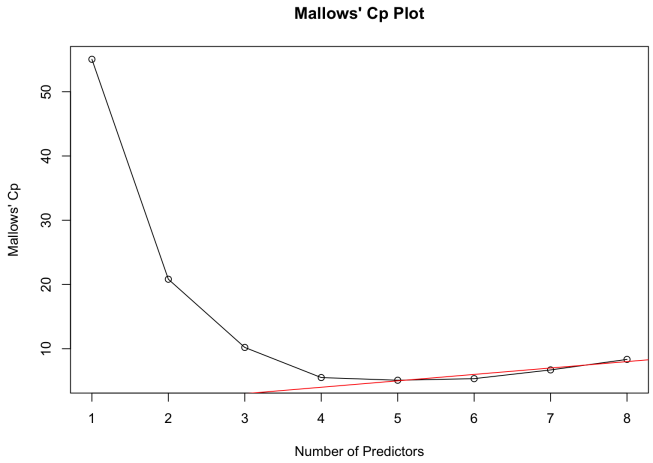



The figures provided depict key diagnostic plots for assessing the adequacy of our model. On the left, we have a plot illustrating residuals against fitted values, while on the right, a QQ-plot showcases standardized residuals against theoretical quantiles, assuming a normal distribution. The plot on the left raises concerns regarding two fundamental assumptions of our model. Firstly, the spread of residuals increases as the fitted values increase, suggesting a violation of the homoscedasticity assumption. Additionally, there seems to be a slight linear trend, which could indicate a violation of the linearity assumption. Similarly, the QQ-plot on the right reveals notable deviations from the expected theoretical quantiles, particularly evident in the right tail. These deviations align with our prior observations concerning the distribution of price data within our dataset.

These violations of the assumptions of linear regression are concerning, and we must keep them in mind when comparing the full model with subset models in the following selection process. To ensure the maximal level of predictive and explanatory validity in our model, we perform three tests to compare alternative models.

### 2.5 Cp plot of models generated using regsubsets() exhaustive model search



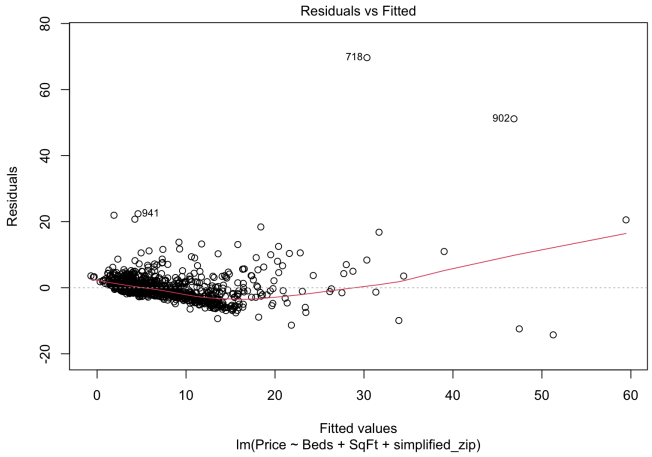

We first use the regsubsets() command from the leaps package in R to conduct an exhaustive model selection using the branch-and-bound algorithm, starting from a full model Price ~ Bed * Bath + SqFt + simplified_zip. The above figure plots the Mallows' Cp values for the ideal model for each number of predictors. According to the graph, the ideal number of predictors is five, as that is where Cp achieves its minimum value. However, while the model as a whole has predictive validity, we cannot use this if we want to examine the potential determinants of price, as only a few of the dummy variables in the model have statistically significant coefficients.

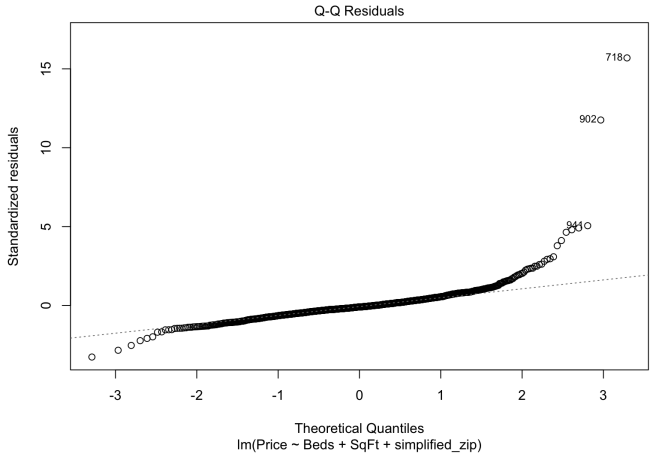
### 2.6 Stepwise Model

Next, we use a backward stepwise selection method, eliminating statistically significant parameters from the full model until the adjusted R-square value no longer increases. The final model selected using this method is:

Price = -1.2444 - 0.7177(Beds) + 0.0071(SqFt) + 1.2834(simplified_zip_T2G) - 1.9717(simplified_zip_T2X) - 0.6679(simplified_zip_T3J) - 1.2026(simplified_zip_T3M) + 0.3558(simplified_zip_Other)

<u>Residual and QQ-plot</u>



The figures above plot residuals against fitted values (left) and standardized residuals against theoretical quantiles given the assumption that the data are normally distributed.
There remains some doubt as to the validity of the model as certain model assumptions are violated. The residual plot on the left shows that variance increases for higher fitted values, violating the assumption of homoscedasticity, while there may also be a linear trend in the residuals. The QQ-plot likewise shows a large deviation from normal quantiles, which matches our initial observation of a skewed distribution of home prices and violates the normality assumption. It is crucial to note, however, that the full model exhibits all the same violations, so these alone are insufficient to eliminate the stepwise model in a comparison of relative appropriateness between the various alternatives.

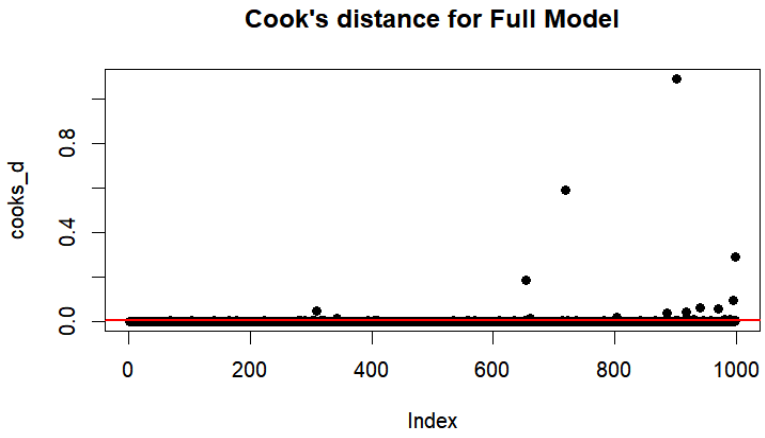
**2.7 Test-Train Splits, RMSE and Cook's Distance Comparison**

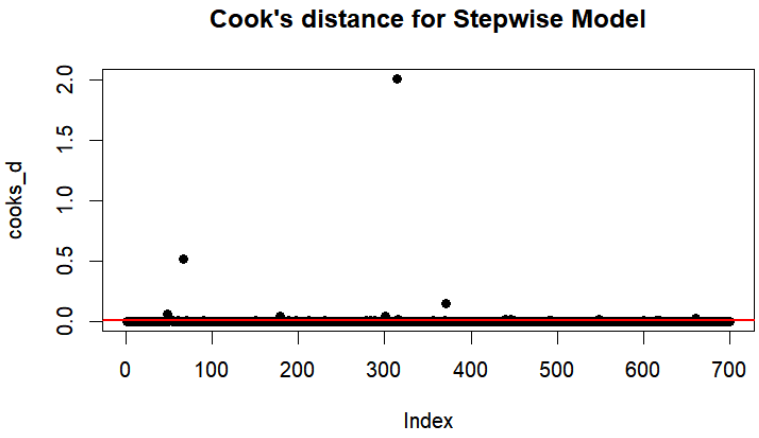
To confirm that a model of this form would have at least predictive validity, we perform a second stage of testing using a split of training and test data. We use the training set to fit a model with Price ~ Beds + SqFt + simplified_zip, the parameters selected by the stepwise selection, alongside the full model, and a third alternative that eliminates the forward sortation area from the stepwise selected model. We then compare the RMSE of the actual and fitted values produced by the three models using the training set.
The Full Model:
Price = -1.8096 - 0.5242(Beds) + 0.2190(Bath) + 0.0071(SqFt) + 1.3449(simplified_zip_T2G) - 2.0281(simplified_zip_T2X) - 0.6501(simplified_zip_T3J) - 1.2368(simplified_zip_T3M) + 0.3559(simplified_zip_Other) - 0.0732(Beds:Bath)
Gave us an RMSE value of 5.50018.

The Stepwise Mode:
Price = -1.2444 - 0.7177(Beds) + 0.0071(SqFt) + 1.2834(simplified_zip_T2G) - 1.9717(simplified_zip_T2X) - 0.6679(simplified_zip_T3J) - 1.2026(simplified_zip_T3M) + 0.3558(simplified_zip_Other)
Gave us an RMSE value of 5.501711.

The No Simplified Zip Model:
Price = -0.9568 - 0.7628(Beds) + 0.0070(SqFt)
Gave us an RMSE value of 5.563402

From this, we can see that the Full Model has a slight advantage in terms of RMSE value to the Stepwise Model that we calculated.




We used Cook's distance to identify influential data points that could affect our model's predictions. These points stand out due to their leverage or significant deviation from the main trend. Most data points have low Cook's distance, indicating minimal influence. However, a few have higher values, suggesting an impact worth examining. These may be outliers like luxury estates or historical buildings. Depending on our goal (modelling broad market trends or capturing market diversity) we might exclude or include these outliers. We chose to include them, recognizing the diversity and complexity of real estate markets.

### 2.8 Final Model Selection

From the previous sections, we can see that the Full Model and the Stepwise Model are both comparable to each other with the Adjusted R-squared values (0.6354 and 0.6358 respectively), the RMSE values (5.50018 and 5.501711 respectively) as well as similar Residual, QQ and Cook's Distance Plots. The Full Model gave us the lowest RMSE value, indicating the highest predictive accuracy among the models evaluated. Yet, the Stepwise Model is also a viable option when trading off a small bit of predictive accuracy for simplicity and interpretability

However, due to the slightly higher predictive power of the Full Model with its lower RMSE value, we choose the Full Model as our Final Model.

Final Model:
Price = -1.8096 - 0.5242(Beds) + 0.2190(Bath) + 0.0071(SqFt) + 1.3449(simplified_zip_T2G) - 2.0281(simplified_zip_T2X) - 0.6501(simplified_zip_T3J) - 1.2368(simplified_zip_T3M) + 0.3559(simplified_zip_Other) - 0.0732(Beds:Bath)

## Conclusion

The model we finalized incorporates explanatory variables that capture the interaction between the number of bedrooms and bathrooms, square footage, and the influence of location through forward sortation areas (FSAs). This model highlights how property features interact with their geographic surroundings. The use of Cook's distance was pivotal in identifying outliers with substantial influence on our model, emphasizing the importance of rigorous data cleaning and the careful consideration of outliers. Despite encountering challenges due to violations of assumptions in linear regression, our careful selection process, which included exhaustive and stepwise model selection methods, revealed that the full model exhibited slightly superior predictive power. This was evidenced by its lower root mean square error (RMSE) value compared to alternative models.

This exploration reaffirms the significant influence of location, size, and property features on housing prices (with SqFt being the most significant parameter based on our summaries), offering a refined lens through which the real estate market's complexity can be understood. By recognizing the diversity and complexity of the market, our study contributes to the broader discourse on housing affordability and market dynamics, providing stakeholders with insights that are both detailed and actionable. This research, rooted in a robust analytical framework, emphasizes the need for ongoing analysis in the face of evolving market conditions and underlines the value of statistical difficulty in dissecting the multifaceted nature of real estate pricing.

## References


Heaven, P. (2023, November 23). *Posthaste: Canada's housing market faces biggest test since 1990s recession*. MSN. https://www.msn.com/en-ca/money/topstories/posthaste-canadas-housing-market-faces-biggest-test-since-1990s-recession/ar-AA1kpFso

Reena, R. (2023, October 8). *Housing Price & Real Estate - 2023*. Kaggle. https://www.kaggle.com/datasets/reenapinto/housing-price-and-real-estate-2023?resource=download