Final Project Part 3 Report

Gadiel David Flores Wendy Huang

2024-12-06

Contents

Contributions **1**

Introduction **2**

Methods **3**

1. Data Preparation . . . 4
2. Assumptions and Multicollinearity . . . 4
3. Model Building . . . 4
4. Model Diagnostics . . . 4
5. Model Performance . . . 5
6. Model Validation . . . 5

Results **5**

Conclusion and Limitations **8**

Ethics Discussion **8**

References **10**

Contributions

- **Gadiel David Flores:** Data Processing, Methods, flowchart and Results
- **Wendy Huang:** Introduction ,Conclusion, Limitation and Ethnic Discussion

Introduction

Housing affordability has become a pressing issue in many urban cities, including Toronto, where rising house prices have placed both homeowners and tenants out of reach for many years. This paper is conducted to investigating several key house indicators to the influence of house price. This is essential for policymakers, real estate professionals and potential homeowners seeking to navigate this challenging market.

Based on prior experience, we hypothesize that housing demand, supply, and inflation are key determinants of house prices. To test this hypothesis, we identified critical indicators representing these factors: house absorption, construction activity, and the Consumer Price Index (CPI), which serves as a proxy for investment conditions. We sourced dataset with relevant predictors from Statistics Canada (Government of Canada, Statistics Canada, 2017) and conducted an in-depth analysis of the variables. After cleaning the data and assessing multicollinearity using the Variance Inflation Factor (VIF), we refined our predictors to include CPI, unabsorbed homes, and completed construction homes as the most relevant variables.This study employs a multiple linear regression model to examine the relationship between house prices and these predictors, providing insights into the factors that drive housing market dynamics.

We conducted literature reviews to validate the selection of our predictors based on prior studies. Housing supply is consistently shown to be a critical factor influencing house prices. Research indicates that higher sales activity (houses sold) is often associated with increased demand, which drives prices upward. Conversely, an oversupply of unsold houses tends to push prices downward. These relationships are commonly examined using hedonic and linear regression models, where transactional data serve as key predictors for estimating price changes effectively (Geerts et al., 2023). The number of newly constructed and completed houses influences house prices by affecting market supply. An increase in construction tends to stabilize prices by meeting demand, while lower construction rates can contribute to supply constraints and higher prices. "Global House Prices: Trends and Cycles" highlights the influence of new constructions and completions on house prices (Gnan, 2021). Regression analyses reveal that increases in supply through construction stabilize prices, while supply shortages contribute to price hikes.

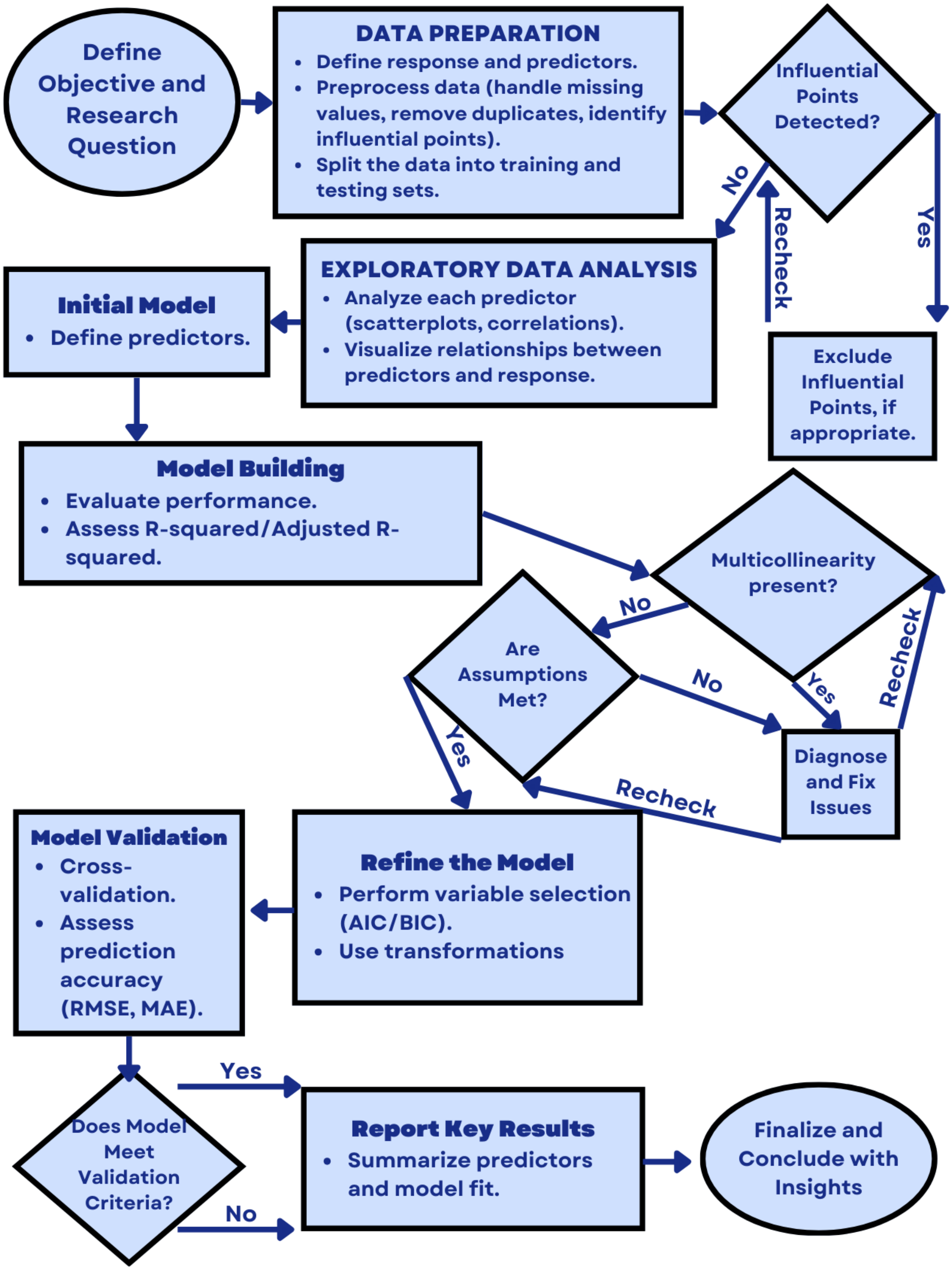
In the context of CPI, the CEPR report discusses their inclusion in econometric models to study the effects of borrowing costs on housing affordability and prices (Muellbauer & Murphy, 2021). It finds that higher CPI indicates higher inflation, reflected in higher mortgage rates and reducing both housing demand and price growth, a relationship frequently captured through regression techniques.

Our literature review confirms that the selected predictors—CPI, unabsorbed homes, and completed construction homes—have been previously studied and exhibit a linear relationship with house prices. Linear regression is recognized as a valid method for this analysis, offering a clear way to quantify the relationships between house price (the response) and these predictors. The validity of the linear relationship will be tested and thoroughly discussed later, as this is a key objective of this paper.

For this analysis, the focus will be on identifying and validating the linear relationship between the predictors and the response variable. We will employ exploratory data analysis (EDA) and Variance Inflation Factor (VIF), evaluate model performance using metrics such as AIC, BIC, R^2, adjusted R^2 and RMSE. Through these statistical methods, we aim to establish that a linear regression relationship exists between house prices and the predictors: CPI, unabsorbed homes, and completed construction homes.

The structure of this paper is as follows: Methods section outlines the primary data preparation process, detailing how we refined the dataset predictors to identify the most relevant variables using exploratory data analysis (EDA) and Variance Inflation Factor (VIF) testing. Additionally, we describe the steps involved in model development, including building, diagnosing, evaluating performance, and validating the multiple linear regression model. In result section, we present the performance of the multiple linear regression model based on the selected predictors, highlighting key findings and insights. Limitation section discusses the limitations of the study, including constraints in the dataset selection, processing, and the model's scope. Ethical considerations related to the study, including data usage and implications, are addressed in Ethics section.

Methods


Define Objective and Research Question
DATA PREPARATION
Define response and predictors.
Preprocess data (handle missing values, remove duplicates, identify influential points).
Split the data into training and testing sets.
Influential Points Detected?
No
Recheck
Yes
Exclude Influential Points, if appropriate.
EXPLORATORY DATA ANALYSIS
Analyze each predictor (scatterplots, correlations).
Visualize relationships between predictors and response.
Initial Model
Define predictors.
Model Building
Evaluate performance.
Assess R-squared/Adjusted R-squared.
Multicollinearity present?
No
Yes
Recheck
Are Assumptions Met?
No
Yes
Recheck
Diagnose and Fix Issues
Model Validation
Cross-validation.
Assess prediction accuracy (RMSE, MAE).
Refine the Model
Perform variable selection (AIC/BIC).
Use transformations
Does Model Meet Validation Criteria?
Yes
No
Report Key Results
Summarize predictors and model fit.
Finalize and Conclude with Insights

1. Data Preparation

According to the topic of house prices and its influencers, we download the dataset with relevant predictors from Statistics Canada (Statistics Canada, 2024). The New Housing Price Index (NHPI) was defined as the response variable, with 13 predictors identified, including absorption (detached-absorption, semi-absorption, detached-unabsorbed, semi-unabsorbed), GDP, CPI, and construction (starting-detached-construction, starting-semi-construction, under construction detached, under construction-semi, completed-construction-detached, completed-construction-semi).

To prepare the data for analysis, we used `tidyverse` for data manipulation and `car` for diagnostics (The R project for statistical computing, 2023). The dataset was cleaned by removing missing values, duplicate observations, and influential outliers identified using Cook's Distance. Finally, the data was randomly split into training (50%) and testing (50%) sets to ensure unbiased and robust analysis.

2. Assumptions and Multicollinearity

The model's assumptions were validated using residual plots, Q-Q plots, and statistical tests like Cook's Distance. Linearity and homoscedasticity were assessed through residuals vs. fitted values and residuals vs. predictors plots(Consumer Price Index (CPI), Unabsorbed Homes and Completed Construction Homes), while normality of residuals was evaluated with Q-Q plots and histograms, which showed minor deviations in the tails. These diagnostic checks ensured the model's validity and strengthened confidence in its results and inferences. A further BOX-COX transformation was applied to the response variable (NHPI) to address issues of skewness and heteroscedasticity.

We used `ggplot2` for visualizations and `car` for diagnostics to explore the relationships between the 13 relevant predictors and the response variable (NHPI). Scatterplots and Variance Inflation Factors (VIF) were used to verify linearity and identify multicollinearity issues among all the predictors. Exploratory data analysis (EDA) was conducted on both training and test datasets to ensure consistency in variable distributions and to detect any anomalies. Multicollinearity was addressed through transformations, which improved model accuracy by refining variable relationships.

3. Model Building

Using the `lm` function in R, an initial regression model was constructed by selecting predictors based on their theoretical relevance and statistical significance, naming the predictors: CPI, unabsorbed homes, completed construction homes, and absorption. To improve model performance, log transformations were applied to skewed predictors absorption, completed construction homes, and unabsorbed homes, addressing non-linearity and improving their relationships with the response variable (NHPI). A Box-Cox transformation was applied to the New Housing Price Index (NHPI) to correct residual skewness and stabilize variance, ensuring better adherence to regression assumptions. The final model underwent rigorous diagnostic checks, including residual plots, Q-Q plots, and statistical tests like Cook's Distance, to validate linearity, homoscedasticity, and normality of residuals. These refinements ensured the model not only satisfied regression assumptions but also achieved optimal accuracy and robustness in explaining the variability in NHPI.

4. Model Diagnostics

After implementing transformations and adjustments, diagnostics were rechecked using residual plots to confirm improvements in linearity, homoscedasticity, and normality. Residuals vs. fitted values and residuals vs. predictors plots were examined to ensure that the transformations effectively addressed patterns of non-linearity and heteroscedasticity. Q-Q plots and histograms of residuals were revisited to evaluate the distribution of residuals, confirming that deviations from normality, particularly in the tails, were significantly reduced. Cook's Distance was revisited to ensure no new influential points emerged that could unduly affect the regression coefficients, and Variance Inflation Factor (VIF) values were reassessed to verify that multicollinearity levels among predictors, CPI and absorption, were within acceptable thresholds. These iterative steps validated the effectiveness of the transformations and adjustments. If any persistent issues were detected, the process was revisited at Step 3, refining predictor transformations or re-evaluating model specifications to further enhance accuracy and adherence to regression assumptions.

5. Model Performance

Model performance was assessed using metrics such as R^2, Adjusted R^2, AIC, BIC, and RMSE. The model demonstrated a strong fit,indicating that the model effectively explained the majority of the variability in the response variable. A high R^2 and Adjusted R^2 value suggest that a large proportion of the variability in NHPI was captured by the model, with the Adjusted R^2 taking into account the number of predictors used, thus avoiding overfitting. The AIC and BIC, which penalize for model complexity, were both relatively low, indicating a good balance between model fit and complexity. The RMSE, which reflects the average prediction error, was also low, suggesting that the model made accurate predictions with minimal deviation from the observed values. Most of the predictors including CPI, completed construction homes, and absorption, were statistically significant, with low p-values confirming strong relationships with the response variable (NHPI), further validating the model's reliability and its effectiveness to capture the key factors influencing NHPI.

6. Model Validation

Model validation was conducted using performance metrics on test data to assess its generalizability and predictive accuracy. Key metrics, such as Mean Squared Error (MSE) and R^2, were calculated to evaluate how well the model performed on unseen data. A low MSE indicates that the model's predictions are close to the actual values, suggesting high predictive accuracy. Conversely, a high MSE signals large deviations between predicted and observed values, implying that the model struggles to generalize to new data. Similarly, a high R^2 indicates that the model explains a large portion of the variability in the response variable, showing strong predictive performance. A low or negative R^2, on the other hand, suggests that the model performs poorly, failing to capture the underlying patterns in the test data. These results are critical for understanding how well the model can be expected to perform on future datasets.

Results

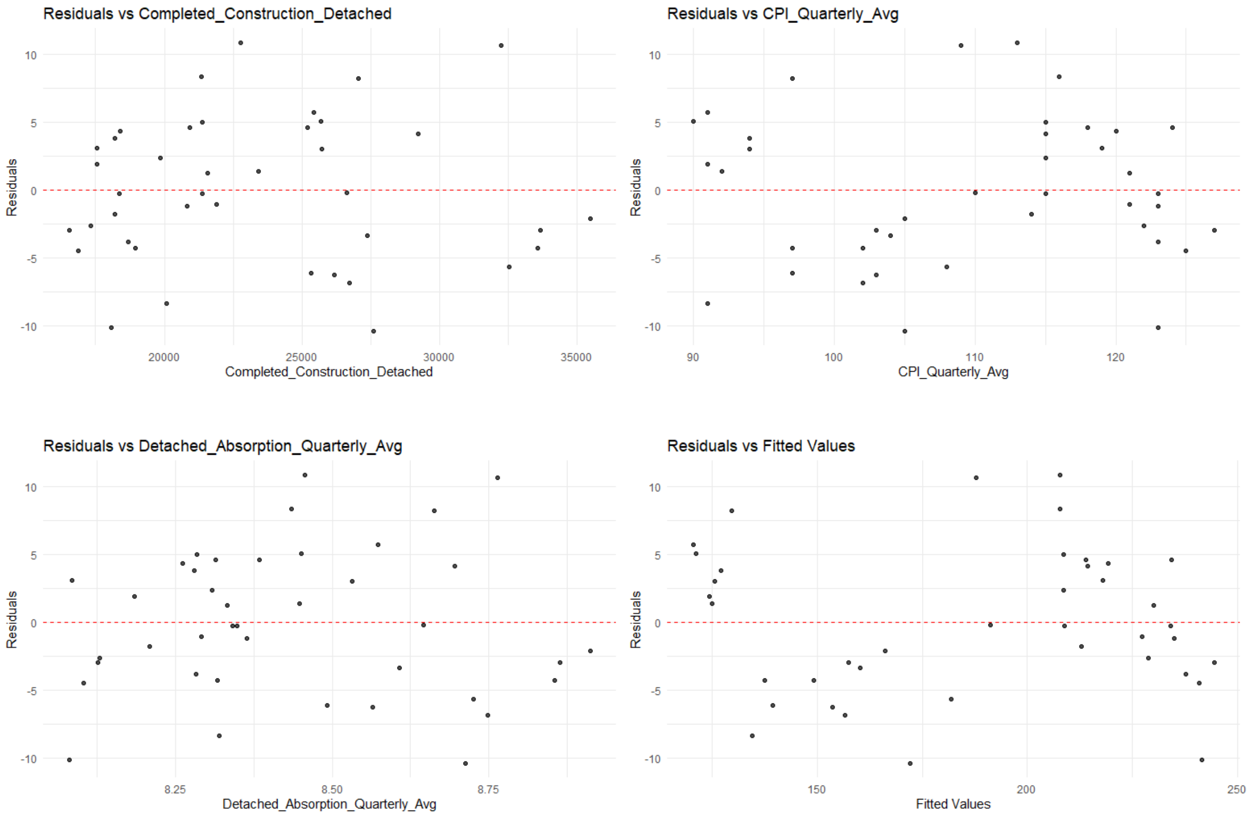
This section presents the results of the multiple linear regression (MLR) analysis. The response variable for this analysis was the New Housing Price Index (NHPI), and thirteen predictors were selected to represent factors such as absorption (detached- absorption, semi-absorption, detached-unabsorbed, semi-unabsorbed), interest rates, GDP, CPI, and various housing construction metrics(starting-detached-construction, starting-semi-construction, under construction detached, under construction-semi, completed-construction-detached, completed-construction-semi). These predictors were chosen based on their theoretical relevance to housing price trends and their potential statistical significance. The focus of this model is explanatory, as the aim is to understand the relationships between variables.

We begin by identifying influential observations. Although removing influential observations (rows 1 and 4) improved model fit and precision, we decided against excluding them to avoid overfitting and to maintain the integrity of the dataset.

The regression analysis conducted for the NHPI response variable exhibits a strong fit on the training data, with an R-squared value of 0.9832 and an Adjusted R-squared of 0.9807, indicating that the predictors explain over 98% of the variability in the response. These metrics demonstrate excellent explanatory power while balancing robustness, as the Adjusted R-squared accounts for the number of predictors in the model. Additionally, the Root Mean Squared Error (RMSE) of 5.4798 reflects low average deviations between predicted and observed values, while the Akaike Information Criterion (AIC) of 263.6004 and Bayesian Information Criterion (BIC) of 275.4225 suggest a reasonable trade-off between model complexity and fit.

Key predictors such as CPI, Unabsorbed Homes, and Completed Construction Homes emerged as significant contributors to the model, with consistently low p-values and strong relationships with the response variable. However, issues of multicollinearity remain evident, as demonstrated by high Variance Inflation Factor (VIF) values for predictors like CPI, GDP and Absorption. Multicollinearity inflates standard errors, potentially reducing the reliability of coefficient estimates.

Transformations applied to the predictors, such as logarithmic adjustments to Absorption, Completed Construction, and Unabsorbed homes, addressed non-linear relationships observed in the scatterplots, improving the linearity assumption of regression. The removal of predictors like GDP, and Starting Construction was motivated by their non-significant p-values and high multicollinearity, streamlining the model while retaining its predictive power. Additionally, the application of a Box-Cox transformation to the response variable mitigated issues of skewness and heteroscedasticity, ensuring constant variance across the fitted values.



Residual diagnostic plots indicate that the model reasonably satisfies the key assumptions of linear regression. Residuals are well-centered around zero, with no clear patterns in residuals versus fitted values, suggesting that the linearity and homoscedasticity assumptions are mostly satisfied.

After applying all the necessary transformations to address non-linearity and improve model fit, the updated regression model demonstrated strong performance metrics. The R^2 value of 0.9742 indicates that 97.42% of the variance in the response variable is explained by the predictors, with an adjusted R^2 of 0.9712 accounting for the number of predictors used. The model's Akaike Information Criterion (AIC) of 237.3722 and Bayesian Information Criterion (BIC) of 247.5055 suggest a well-optimized model with minimal overfitting. Additionally, the Root Mean Squared Error (RMSE) of 4.048 confirms a low average prediction error, indicating that the transformations significantly improved the model's predictive accuracy.

Despite strong performance metrics on the training data, the model's performance on the test set highlights significant shortcomings. The Mean Squared Error (MSE) on the test data is alarmingly high (1.303812×10^{15}), and the R-squared value is highly negative (-4.163317×10^{12}), indicating that the model fails to generalize to unseen data. These results suggest severe overfitting, where the model captures noise in the training set rather than generalizable patterns, rendering it ineffective for prediction.

Overall, while the regression model demonstrates robust explanatory power on the training data, its limited performance on the test data is less concerning as the main objective of this model is explanatory rather than predictive. The model effectively identifies key relationships and provides valuable insights into the factors influencing the response variable, which aligns with its intended purpose. As an explanatory tool, the model serves its purpose well by shedding light on underlying patterns and relationships.

Table 1: Model Coefficients and Summary

Term	Estimate	Std. Error	t value	p value
(Intercept)	-1617.6578	101.6560	-15.9131	**0.0000**
Detached_Absorption_Quarterly_Avg	-31.1436	18.2679	-1.7048	**0.0971**
Detached_Unabsorbed_Quarterly_Avg	0.0056	0.0020	2.8262	**0.0077**
Completed_Construction_Detached	28.8190	19.6672	1.4653	**0.1518**
CPI_Quarterly_Avg	373.0912	11.9951	31.1036	**0.0000**

Conclusion and Limitations

The final regression model effectively addresses the research question by identifying key relationships between the New Housing Price Index (NHPI) and its significant predictors, namely, CPI, Unabsorbed Homes, and Completed Construction Homes. The New Housing Price Index (NHPI) exhibits a positive relationship with the Consumer Price Index (CPI), completed construction homes, and absorption, while showing a negative relationship with unabsorbed homes. Additionally, although GDP and starting construction were removed from the final model due to a lack of statistical significance, both variables demonstrate a positive influence on the NHPI. These relationships suggest that housing prices are shaped by a delicate balance of demand-side pressures (CPI, Absorption) and supply-side constraints (Unabsorbed Homes, Completed Construction). The positive and negative effects of these predictors align with economic theory, underscoring the utility of the model for understanding housing market dynamics.

The final model explains a substantial portion of the variability in NHPI with an adjusted R^2 of 0.9712,indicating that 97.12% of the variation in NHPI is explained by the selected predictors. Transformations, such as logarithmic adjustments to variables like Absorption, Completed Construction Homes, and Unabsorbed Homes, successfully addressed non-linearity, making them suitable for regression analysis. The Box-Cox transformation normalizing its distribution and ensuring that the model assumptions were better satisfied. All the transformations improves the model's explanatory power and ensuring compliance with key linear regression assumptions.

The findings emphasize the need for careful consideration of both theoretical and statistical relevance when selecting predictors. This model serves as a valuable tool for understanding market dynamics rather than precise forecasting, i.e. this is a descriptive model more than a predictive model. Future studies could address these limitations by employing alternative modeling approaches. For example, the high varaiance inflaction factor values (VIF) for predictor such as CPI and Absorption suggest intercorrelations that inflate standard errors, potentially affecting the reliability of individual coefficient estimates. Since this model focus on the explanatory instead of prediction, the model performs poor on test data with a negative R^2 and sightly high Mean Squared Error (MSE). This also shows that while the model explains patterns within the training set, it struggles to generalize to new data, limiting its predictive reliability. What's more ,the exclusion of theoretically relevant predictors, such as GDP and Starting Construction, due to multicollinearity or insignificance, may limit the comprehensiveness of the model. Additionally, the decision not to remove influential observations (rows 1 and 4) to preserve dataset integrity could introduce noise or bias into the analysis. By building on this work, future research can deepen our understanding of housing price determinants and improve strategies for managing market volatility.

Ethics Discussion

The dataset used in this analysis originates from Statistics Canada, a reputable and reliable source, which ensures the data's credibility (Government of Canada, 2024). However, the sensitivity of the response variable—house prices—requires careful ethical consideration. Housing affordability impacts individuals and communities significantly, and any conclusions drawn from this analysis may influence public perception or policy decisions.

First, transparency is essential. We document every step of data cleaning, analysis, and model selection to ensure reproducibility and avoid misrepresentation of findings. This transparency fosters trust and helps prevent the ethical pitfalls of obscuring methodology. We are fully aware that the results are closely tied to regional economic factors like GDP and CPI. We took extra care to avoid errors that could lead to incorrect conclusions.

Second, respect for sensitivity is paramount. House prices are not just economic indicators; they directly affect people's livelihoods and well-being. It is crucial to present results in a manner that avoids perpetuating inequalities or stigmatizing specific groups or regions. On the other hand, the aim of this paper is to provide insights for policymakers and the estimators. While it is challenging to avoid perpetuate inequalities or stigmatize particular groups or region, we strive to present the results in a way that reflects the data accurately while minimizing harm to any particular group.

Third, fairness in data representation is critical. The predictors—such as sold and unsold houses, construction activity, and CPI were analyzed impartially to avoid introducing biases. Any biases identified in the dataset or methodology must be explicitly addressed to maintain the analysis's integrity.

Finally, as stewards of statistical work, we are responsible for cultivating virtues such as diligence, accountability, and equity, while avoiding vices like negligence or bias. By adhering to these principles, we ensure that the research contributes positively to understanding the impacts of supply-side factors on housing prices in Toronto without causing harm or perpetuating inequities.

References

Geerts, M., vanden Broucke, S., & De Weerdt, J. (2023, May 14). A survey of methods and input data types for house price prediction. MDPI. https://doi.org/10.3390/ijgi12050200

Gnan, E. (2021, June 25). Monetary policy and housing markets: Interactions and side effects. ECOSCOPE. https://oecdecoscope.blog/2021/06/25/monetary-policy-and-housing-markets-interactions-and-side-effects/

Government of Canada, S. C. (2024, December 5). Statistics Canada: Canada's National Statistical Agency. Statistics Canada: Canada's national statistical agency. https://www.statcan.gc.ca/en/start

Government of Canada, Statistics Canada, S. C. (2017, March 9). New housing price index (2007=100). https://doi.org/10.25318/1810005201-eng

Government of Canada, Statistics Canada, S. C. (2024, November 19). Consumer price index, monthly, not seasonally adjusted. Consumer Price Index, monthly, not seasonally adjusted. https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1810000401

Muellbauer, J., & Murphy, A. (2021). What drives house prices: Lessons from the literature | CEPR. https://cepr.org/voxeu/columns/what-drives-house-prices-lessons-literature

R Foundation for Statistical Computing (Ed.). (2023). The R project for statistical computing. R. https://www.r-project.org/