Group B7: Group Project

Kevin Shiao, Allan Wang, Tiffany Zhang

Introduction

Source of Data

Our analysis is based on the “B.C Housing Affordability” release published on June 20th 2025 by BC Stats under the Open Government License - British Columbia. The files we will use are

- 2025Q1 - B.C. Housing Affordability Payment as Percent of Income Analysis.pdf
- CMA_quarterly_results_filtered.csv

House price data is provided by the BC assessment Authority (BCA). Income Data, Prime Rates, and Consumer Price index data is provided by Statistics Canada (StatsCan). Below is the website where you can find all the related files.
https://catalogue.data.gov.bc.ca/dataset/b-c-housing-affordability

Description of Variables

Variable	Description
CMA ID	The unique identifier of the Canadian Metropolitan Area. Defined in Statistics Canada’s Standard Geographical Classification (SGC) system.
CMA Name	The Name of the Canadian Metropolitan Area. Defined in Statistics Canada’s SGC system.
Year	Calendar year of the observation
Quarter	Fiscal quarter of the year of the observation
Housing Type	Category of dwelling
Median Housing Price	Median sale price of homes sold (CAD) Collected by BCA.

Median After Tax Family Income	Median After-Tax family income (CAD). Collected by Statistics Canada via T1 file combined with T4 tax file. Used ARIMA Modeling for the more recent years.
Prime Interest Rate	The Bank prime lending rate in effect during that quarter (%) Fetched from Bank of Canada Prime Interest Rate. Data publicly available.
Quarterly Payment	Estimated quarterly payment of mortgages (CAD)
Payment to Income Percent	Housing affordability metric which is a combination of house prices, income, and mortgage rates. (%)
Historical Average Payment to Income Percent	Long-term average of the payment-to-income percentage for that CMA, against which the current value can be compared. (%)
Number of Resales	Total count of home resale transactions recorded in that CMA during the quarter. (Count) Collected by BCA.

Relevant Information regarding the data

- “Median after-tax family income was gathered from StatsCan and mapped to each CMA region for all years where CMA level income data was available“ (BC Stats, 2025, p. 12).
- “The median after-tax family income for the more recent (and not yet available) years was estimated using ARIMA modeling for each CMA region“ (BC Stats, 2025, p. 12).
- “To estimate historic BC housing affordability back to 1979, multiple historic data sets were combined to provide median family income estimates that align with the current income data series“ (BC Stats, 2025, p. 12).
- “Income data is collected/reported annually but was adjusted (smoothed) in this model to change on a quarterly basis to better reflect actual income trends over time“ (BC Stats, 2025, p. 12).
- “The Average Median After-Tax Family Income and Average Median House Prices was combined with the average Prime Rate for the same period to calculate a quarterly Payment as Percent of Income” (BC Stats, 2025, p. 12).
- “These three data series are used to estimate a monthly payment for housing as a percentage of the total income across the 4 housing categories” (BC Stats, 2025, p. 12).

Research Question and Motivation

Our main research question is, **"How do local market activity (number of quarterly resales) and borrowing costs (Prime Interest Rate) jointly explain variations in housing affordability (Payment to Income Percent) across Canadian Metropolitan areas from 2000 through Q1 2025, given the confounding effects of region (CMA)?"**

Housing affordability is a pressing socioeconomic issue. It affects household well-being, local economies, and financial stability. By quantifying the roles of

1. Region
2. Market Activity
3. Borrowing Costs

We can answer:

- Which CMAs are inherently more or less affordable, after accounting for market churn and interest rates
- How sensitive is affordability to shifts in resale volumes versus interest rate moves
- Where should policymakers, lenders, and first time buyers focus their attention when the prime interest rate changes or market activity spikes.

This analysis will deliver clear, region-specific insights on the relative importance of supply-side factors, demand-side factors, and financing costs. It will inform targeted interventions (for example, rate buffers or resale-tax incentives) and enable early warning signals for emerging affordability crises. If we understand these drivers, policymakers could craft new laws and policies to improve housing affordability for lower and middle income families across Canada.

Analysis

Data Visualizations

- Payment to Income Percentage reached its lowest point around 2002 at roughly 17%
- From 2020 to 2023, Payment to Income Percentage climbed sharply from around 25% to nearly 49%.
- The Prime Interest Rate was highest in 2000 at around 7% and then fell during the early 2000s
- Interest rates were stagnant from 2009-2019, then leading into an increase from 2020 onwards

- Apartments have the lowest prices overall compared to Row Housing and Single Detached houses. They also have the tightest spread showing they

are the most affordable and consistent housing type across all BC Census Metropolitan Areas
- Row housing sits in the middle with fewer extreme outliers than single detached.
- Single Detached homes have the highest median and the largest interquartile range. They also have the most extreme high price outliers as well
- Vancouver has a more active housing market compared to the other metropolitan areas. Its median quarterly resale count is around 4500 while the others sit below 400
- Vancouver also shows high volatility with its interquartile range spanning roughly around 3000 to 8000 resales per quarter and extreme outliers with around 20000 resales in a quarter.
- Chilliwack, Kamloops, and Nanaimo all have very similar and tighter distributions. Medians of around 200 and IQRs around 500
- This shows that Vancouver dominates provincial resale activity while smaller markets stay relatively stable and low volume

Visualizations Between Response and Explanatory

Distribution of Payment to Income Percent
Observing the plot, we immediately notice the distribution of payment to income percent is unimodal and right skewed. With the help of R, we find the median of payment to income percent is 31.2%, meaning the median family spends 31.2% of their income on housing payment. This is also quite revealing about the economy, such that there have been many more times historically that families had to partition a huge portion of their income (and potentially even more than what they earned) to pay the mortgage.

Payment Income Percent vs Prime Interest Rate By Region
From the plot above, we can see that generally, during times with higher prime interest rates, the payment to income percent is higher. Payment to income percent and prime interest rate are positively correlated. We can see that Vancouver generally has the highest payment to income percent, while prime interest rate does not differ by region.

Payment Income Percent vs Number of Resales By Region
From the plot above, we can see that Vancouver has a dominating number of resales when compared to other regions. Most of the scatter points are closer to the median of 31.2% that we mentioned above, but there are a handful of points that fall close to the top of the plot, which is expected, based on the right skewed distribution of payment to income percent. From this graph, it is a little difficult to tell the correlation

between payment to income percent and the number of resales, due to the points close to the top of the plot and how far to the right of the plot the resale for Vancouver extends to. It seems as though the correlation coefficient could be positive but it would be close to 0.

A plot of how the prime interest rate has changed depending on the year is also important in helping us with interpretation of results. The trend seems to follow an upward opening parabola.

Model Specification

Before continuing, we have cleaned the dataset, and we have filtered for the 4 most populated regions [Reference 2] to be our main regions of analysis, they are Abbotsford - Mission, Kelowna, Vancouver, and Victoria.

We begin with an additive linear regression model of housing-payment burden on region, borrowing costs, and market activity:

$$Yi = 0 + 1z1 + 2z2 + 3z3 + 4x1 + 5x2 + .$$

Where:
Yi = Payment to Income Percent of the i'th observation.
z1i = 1 If the i'th observation is from Kelowna. 0 otherwise.
z2i = 1 If the i'th observation is from Vancouver. 0 otherwise.
z3i = 1 If the i'th observation is from Victoria. 0 otherwise.
x1i = Prime Interest Rate of the i'th observation.
x2i = Number of Resales of the i'th observation.
For i = 1, … , 1616.

The model fitted to our data is:

$$y = 14.171 - 0.5963z1 + 7.7795z2 + 0.9351z3 + 4.0926x1 + 0.0008x2.$$

The adjusted R-squared is approximately 0.2541, and the corresponding coefficients of **cma_nameVancouver**, ***prime_interesterate*** and ***number_of_resales*** are highly significant ($p < 0.01$). This also means if we created a 99% confidence interval for them, the intervals will not contain 0.

To check model assumptions we plot residuals vs. fitted values and we get:
Observation: We can see there is a "funnel" shape in the plot, which indicates there could be non-constant variance in the error terms (in this case, variance is larger for larger fitted values of response).

We also check the QQ plot:

Observation: we see that the distribution of residuals are quite obviously right skewed. However, since the main goal of our study is inference, we may rely on the Central Limit Theorem for an approximately normal sampling distribution of the estimators for the parameters as we have a sample size of 1616. The standard error for the estimators are also valid, since we have Var(B)=2*(XTX)-1, it does not rely on the distribution of errors to be normal, thus the inferences should still be fine.

Check Collinearity

Having fitted the basic additive model, it would be simple now to check for collinearity within our predictors. After using VIF from the car package, we found adjusted GVIFs of 1.136, 1.009, 1.474 for CMA, prime interest rate, and number of resales, respectively. These values are well below 5, which is a standard threshold of concern. Thus, we can be comfortable with our predictor variable selections.

Back to stabilizing variance, we try to take the natural logarithm of our response variable and fit a new model:

$$\log(y)=2.9+0.0284z1+0.1852z2+0.0773z3 +0.1017x1+0.0000335x2.$$

The adjusted R-squared increases to 0.2635 and coefficients for **Vancouver CMA, Prime Interest Rate** and **Number of Resales** remain significant at 1% significance level. After transforming the response, we also note that the coefficient for **Victoria CMA** has become significant at 1% significance level as well.
Observing the residual plot again, we can see that the “funnel” disappears and there is no obvious pattern in the residual plot, points seem randomly scattered.
Observation: After the log-transformation, most residuals lie on the QQ line, so the central part of the error is nearly normal. The extreme tails show a pattern of heavy tails.

Interaction Model:

We want to explore whether there are interactions between explanatory variables. Therefore, we next explored whether the marginal effects of our three predictors vary in combination. We can be confident that there is no interaction between CMA and Prime Interest Rate, as the interest rate is the same for every CMA in a given year. We will then test the interaction models with CMA * Resales.

1. CMA x Prime Interest Rate
To verify this, we fitted a model with **CMA * Prime Interest Rate** interaction term, with reference to the notations defined above:

$$\log(y)=2.876+0.0616z1+0.2068z2+0.1179z3 +0.1075x1+0.0000335x2-0.008z1x1-0.005z2x1-0.010z3x1.$$

From the summary we did indeed find the interaction coefficients to be **non-significant**, which should also be true from the physical perspective of the world, as interest rates do not differ by regions within BC. Thus, there is no need to test for the full model either, as we have a confirmation from the physical perspective.

2. CMA x Resales

To test this, we fitted a model with **CMA * Number of Resales** interaction terms.

$$\log(y)=2.556+0.1178z1+0.5418z2+0.178z3 +0.1153x1+0.0008x2-0.00038z1x2-0.00078z2x2-0.00052z3x2.$$

From summary, we found that the adjusted R-squared has increased to 0.4093, and all of the coefficients are significant at 1% significance level. This is a very good sign, and indicates that we should include this interaction term in our model.

Then we check the residual plot and QQ plot again:

From the plots we can see points are randomly scattered around the center line in the residual plot, and the pattern in the QQ Plot did not change much. Therefore, we can be confident that the interaction between CMA and Resales should be kept in our model.

3. Prime Interest Rate x Resales

To test this, we fitted a model by adding on ***Number of Resales * Prime Interest Rate*** interaction on top of the interaction model that we fitted in the last step.

$$\log(y)=2.558+0.1178z1+0.5411z2+0.178z3 +0.1148x1+0.0008x2-0.00038z1x2-0.00078z2x2-0.00052z3x2+0.00000025x1x2.$$

From summary, we found that the adjusted R-squared actually decreased from the previous model (0.4093 to 0.4089), and the p-value of the new interaction term is 0.89493, which is not significant at significance level of 1%. Therefore, the interaction between Prime Interest Rate and Resales should not be retained**.**

Conclusion

Considering the models that we have fitted, we believe the best one was the second interactive model that we fitted, as shown below. We have also found 99% confidence intervals for the parameters, given this sample.

Model Interpretation

$$\log(y)=b0+b1z1+b2z2+b3z3 +b4x1+b5x2-b6z1x2-b7z2x2-b8z3x2,$$

$$\log(y)=2.556+0.1178z1+0.5418z2+0.178z3 +0.1153x1+0.0008x2-0.00038z1x2-0.00078z2x2-0.00052z3x2.$$

Interpretations for the final model:

Since we transformed the model by logging the response, it is best to exponentiate our model for the best interpretation (so we can talk about the response as it should be, the payment to income percent).

$$y=e(2.556+0.1178z1+0.5418z2+0.178z3 +0.1153x1+0.0008x2-0.00038z1x2-0.00078z2x2-0.00052z3x2).$$

In the interaction model, the coefficients related to the intercepts for different cities are not particularly interesting in terms of their implication. This is the expected payment to income percent, when both x1 (prime interest rate) and x2 (number of housing sales) are equal to 0. From the physical perspective, this has not occurred throughout the history of our dataset (prime interest rate has always been higher than 0).

However, they are still worth analyzing:
b0 =2.556, e2.556=12.89, 99% CI= [2.467, 2.645], exp 99% CI = [11.79, 14.09]

- The expected payment to income percent in Abbotsford - Mission is 12.89% when the prime interest rate is 0% and there are 0 resales.
- We are 99% confident that the true expected payment to income percent in Abbotsford - Mission when the prime interest rate is 0% and there are 0 resales is between 11.74% and 14.09%.

b1 =0.1178, e0.1178=1.125, 99% CI= [0.026, 0.210], exp 99% CI = [1.026, 1.234]

- The expected payment to income percent in Kelowna is 1.125 times higher than that of Abbotsford - Mission, when the prime interest rate is 0% and there are 0 resales. Thus, our estimate for the payment to income percent for Kelowna under these conditions is 14.501%.

- We are 99% confident that this true expected factor for the difference is between 1.026 and 1.234 when the prime interest rate is 0% and there are 0 resales.

b2 =0.5418, e0.5418=1.720, 99% CI= [0.446, 0.638], exp 99% CI = [1.562, 1.893]

- The expected payment to income percent in Vancouver is 1.72 times higher than that of Abbotsford - Mission, when the prime interest rate is 0% and there are 0 resales. Thus, our estimate for the payment to income percent for Vancouver under these conditions is 22.171%.
- We are 99% confident that this true expected factor for the difference is between 1.562 and 1.893 when the prime interest rate is 0% and there are 0 resales.

b3 =0.178, e0.178=1.195, 99% CI= [0.084, 0.273], exp 99% CI = [1.087, 1.313]

- The expected payment to income percent in Victoria is 1.195 times higher than that of Abbotsford - Mission, when the prime interest rate is 0% and there are 0 resales. Thus, our estimate for the payment to income percent for Victoria under these conditions is 15.404%.
- We are 99% confident that this true expected factor for the difference is between 1.087 and 1.313 when the prime interest rate is 0% and there are 0 resales.

b4 =0.1153, e0.1153=1.122, 99% CI= [0.102, 0.129], exp 99% CI = [1.107, 1.137]

- Each 1% increase in prime interest rate is associated with an expected increase in payment to income percent by a factor of 1.122, holding other variables constant.
- We are 99% confident that the true multiplicative factor on the expected payment to income percent for 1% increase in prime interest rate is between 1.107 and 1.137.

b5 =0.000801, e0.000801=1.0008013, 99% CI= [0.0007, 0.0009], exp 99% CI = [1.0007, 1.0009]

- In Abbotsford - Mission, each 1 more house sold is associated with an expected increase in payment to income percent by a factor of 1.001, holding other variables constant.
- We are 99% confident that the true multiplicative factor on the expected payment to income percent for 1 more house sold in Abbotsford - Mission is between 1.0007 and 1.0009.

b6 =-0.000378, e-0.000378=0.9996, 99% CI= [-0.000545, -0.000211], exp 99% CI = [0.9995, 0.9998]

- In Kelowna, each 1 more house sold is associated with an expected increase in payment to income percent that is 0.9996 times that of the Abbotsford - Mission region, holding other variables constant.

- Thus, for Kelowna, our estimate for the associated expected increase in payment to income percent for 1 more house sold is a factor of 1.0004010, holding other variables constant.
- We are 99% confident that the true multiplicative factor for the difference is between 0.9995 and 0.9998, holding other variables constant.

b7 =-0.00078, e-0.00078=0.9992, 99% CI= [-0.000916, -0.000644], exp 99% CI = [0.9991, 0.9994]

- In Vancouver, each 1 more house sold is associated with an expected increase in payment to income percent that is 0.9992 times that of the Abbotsford - Mission region, holding other variables constant.
- Thus, for Vancouver, our estimate for the associated expected increase in payment to income percent for 1 more house sold is a factor of 1.0000007, holding other variables constant.
- We are 99% confident that the true multiplicative factor for the difference is between 0.9991 and 0.9994, holding other variables constant.

b8 =-0.000522, e-0.000522=0.9995, 99% CI= [-0.000675, -0.000369], exp 99% CI = [0.9993, 0.9996]

- In Victoria, each 1 more house sold is associated with an expected increase in payment to income percent that is 0.9995 times that of the Abbotsford - Mission region, holding other variables constant.
- Thus, for Victoria, our estimate for the associated expected increase in payment to income percent for 1 more house sold is a factor of 1.0003009, holding other variables constant.
- We are 99% confident that the true multiplicative factor for the difference is between 0.9993 and 0.9996, holding other variables constant.

Synopsis

Throughout this study, we explored different models to select the best one to answer the question: “How do local market activity (number of quarterly resales) and borrowing costs (Prime Interest Rate) jointly explain variations in housing affordability (Payment to Income Percent) across Canadian Metropolitan areas from 2000 through Q1 2025, given the confounding effects of region (CMA)?”.

We explored whether there was collinearity, assessed the performance of different models, and investigated interactions between our predictors. Based on our results, it does seem plausible that payment to income percent is explained by the number of resales and prime interest rate, given the confounding effects of regions. With our final model, we achieved an adjusted R-squared of 0.4093, which is decent in this context.

Before interpreting the results in detail, it is important that we establish, a 100% payment to income percentage indicates that 100% of a family's income goes towards paying a mortgage, which means houses are very unaffordable. It is also important to note that this percentage could go higher than 100% as well.

Referring to the results we obtained above, it appears that there is a positive association between the prime interest rate and payment to income percent, such that for every 1% increase in prime interest rate, there is an associated difference by a factor of 1.122 for the payment to income percent, holding other variables constant. This aligns with what we know about the housing market as common knowledge; if the borrowing cost is higher, housing affordability is lower (higher payment to income percent).

We also note that there is a positive association between the number of resales and payment to income percent. However, the effect on the associated payment to income percent is not the greatest, though it is certainly not zero. This also differs for different regions, as more populated regions will have more housing resales; for every house sold, there is an associated difference by a factor of (1.008013, 1.0004010, 1.0000007, 1.0003009) for (Abbotsford - Mission, Kelowna, Vancouver, and Victoria), respectively. We can also tell that attempting to explain payment to income percent with interaction between number of resales and region has increased the adjusted R-squared substantially.

Addressing the limitations of the study, the non-normal distribution of our residuals is something that we discovered during analysis. Though we attempted different transformation techniques, the normality of error terms is not fixed. Thus, we propose to rely on the CLT; based on our final residual plot, it seems the variance of the error term has been stabilized; we can see it is finite, and we have a very large sample. Some may argue that there may not be independence of error terms in our data, which does not guarantee our estimators to be unbiased. However, there have been similar studies (Activity 14) that performed linear regression on data sampling regions repeatedly over time.

In conclusion, economic growth, market activity, policies of banks and government, and the region you live in can really affect the pressure to purchase housing units. Though the independence of errors may not be fully guaranteed due to the repeated sampling structure, the CLT supported with our large sample size and stabilized variance supports the validity of our estimators for practical inference, consistent with established practices in similar longitudinal studies.

References

1. https://catalogue.data.gov.bc.ca/dataset/b-c-housing-affordability

2. https://www12.statcan.gc.ca/census-recensement/2021/as-sa/fogs-spg/page.cfm?lang=E&topic=1&dguid=2021A000259