CSC110 Project : Final Submission

Aidan Li, Youssef Soliman, Min Gi Kwon, Tej Jaspal Capildeo

December 2021

1 Project Title

A study of the changes of Housing Price Index during the COVID-19 pandemic in different areas of Canada.

2 Problem Description

Research Question: "How has the COVID-19 pandemic impacted prices in the Canadian housing market?"

Housing is important. People need a place to rest safely and take shelter, so housing satisfies a basic necessity needed to sustain life. Also, housing usually comprises the largest component of a person's wealth, and can be an invaluable investment asset. This means a country's housing market, which determines the price of all housing using supply and demand, is extremely important to consumers, as well as for the country's economy as a whole.

The Canadian housing market in particular has seen prices rise significantly almost every year for the last two decades. Certain cities, like Toronto and Vancouver, have seen higher price increases than others (Teranet-NBC, 2021). This is bad for young Canadians, who are increasingly less likely to ever be able to afford a house to call their own. It also signals vulnerability in the Canadian economy, as the Canadian housing market comprises the largest portion of Canada's GDP (Statista, 2021). Analysts claim the Canadian housing market is

one of the most overvalued in the world and has a high risk of a sharp correction (CHMC, 2021).

As COVID-19 rocked the world, many economies suffered. People got sick, businesses had to close down, and healthcare systems were strained. Many governments attempted to stimulate spending in their economies to resuscitate their faltering economies. Since the Canadian economy is so closely tied to its housing market, it is highly likely that the Canadian housing market was affected by the pandemic.

We want to investigate how the pandemic has affected prices in the housing market for the 11 largest cities in Canada, as well as in some of these cities individually where prices have typically risen faster. Prices may have dropped, bucking the trend of increasing house prices. Alternatively, the trend of increasing house prices may have gotten even worse; prices may have increased at a rate beyond what would be expected without the pandemic. It could also be possible that prices continued to increase at the same rate, and that COVID-19 was no match for expedient investors and home buyers.

We want to find out if prices in the housing market for the 11 largest cities in Canada deviate from the trend, as well as examining changes in individual cities in order to see if certain regions were affected more than others.

3 Dataset

The dataset comes from a collaboration between Teranet and the National Bank of Canada. Teranet is a private company that facilitates Canada's electronic land registration systems and commerce, and is owned by Borealis, the infrastructure arm of the Ontario Municipal Employees' Retirement System (OMERS). The dataset is a CSV file that has a total of 25 columns. Our dataset is structured as follows:

- The *Transaction Date* variable contains the month and year of the observation that the following columns/variables were calculated for.
  - The base *Transaction Date* (where the price index is 100) is June 2005.
- There are 12 regions and each region has two data variables/columns:

  - The *index* variable is the percent multiplier of the average house value compared to the base date in June 2005.
  - The *Sales Pair Count* variable is the number of houses included in the calculation of *index*.

- The *c11* group represents the composite of the 11 largest cities in Canada with the biggest housing markets. The other 11 groups are cities, denoted by *province_city*.

More information regarding how index is calculated can be found in section 1.3 of this methodology report.

4 Computational Overview

4.1 Data Transformation

First, we parsed our data from the CSV file using `parse.py`. Since our data set contained twelve different geographical regions that can be plotted and analyzed, we programmatically isolated each region with its own index and sales pair count in order for their individual regression lines to be calculated and toggled on/off. We organized our data into a dictionary mapping the regions (in string format) to their data entries as IndexClass entries (from `housing_entry.py`) with date, house index, and sales pair count.

From here, we converted the entries to Pandas dataframes, and split up the dataframe containing dates and house indexes for the key location into two dataframes: a training dataframe and a testing dataframe. The testing dataframe

contained all entries of data from 2020 onwards, when COVID-19 emerged. The training dataset included all entries prior to 2020.

We used the training dataframe to calculate three different prediction models: linear regression, exponential regression, and SVM regression. Then, we compared the test data to the extrapolated estimates we calculated from our models.

For all dataframes, we converted the values from the `Transaction Date` column, which contained dates, into the number of days that passed since the baseline date of 1 Jul 1990, and stored the new values into a separate column inside the dataframe. This conversion allowed the regression to be performed with a finer numerical value as the mathematical functions used in regression do not support date objects.

4.2 Regression Methods

Firstly, we implemented a Support Vector Regression model in `plot.py`. This model uses a supervised learning algorithm which can be used to predict real discrete numbers. This model is often used as a simple prediction model at it is really good at fitting to predictable data without over fitting (if the configured parameters are within reason). Internally, this algorithm takes in the data, finds two boundary lines and a hyper plane (best fit) line, then attempts to fit the hyper plane within these boundary lines using a set threshold value that can be configured. We resorted to sklearn's implementation of this algorithm as its effecient implementation was beyond the scope of this course.

We implemented our very own algorithm in `regression.py` to find the slope and intercept of a least squares regression line. Our function takes an input of entries with a date and corresponding house price index. The date is converted to a numerical value representing the days that passed from the baseline date of Jul 1, 1990. Then, our function manually calculates the regression line to predict index prices based on the date using the following mathematical formulae:

$$slope = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$intercept = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

where y is an entry of index prices, x is an entry of date, and n is the total number of entries. The numerical values of the date is the predictor/independent variable, and the house pricing index is the reactor/dependent variable.

Knowing the slope and intercept allowed us to model linear regression lines and exponential regression lines predicting house index prices by the date for the regions of Canada in our dataset.

Our linear models were in the form $y = mx + c$, where y is the house price index, x is the date, m is the line's slope/gradient, and c is the intercept.

Our exponential models applied the natural logarithm to the index values before calculating the slope and intercept. Thus, our models were in the form $\ln y = x \ln m + \ln c$, which we manipulated to $y = cm^x$.

We extrapolated our models to predict values for index prices in 2020 and 2021, when the COVID-19 pandemic was most prolific. We then compared the actual index price values in our data to our predicted values from the models. We expect the indexes for 2020 and 2021 to all have positive residual errors, which shows how COVID-19 may have affected housing prices.

Finally, we calculated the Root Mean Square Error values of our models for the various locations to evaluate the accuracy of our regression models.

4.3 Visual Presentation

We used the Plotly library to plot our linear and exponential regression models using `plot.py`, as well as the points in the years 2020 to 2021 to give a visual representation. We instantiated a Figure object from the Plotly library, which is shown by our program opening a HTML page. The figure is interactive, as the regions on the legend of the figure can be clicked to toggle on and off the line graph. The figure also includes a vertical line on 2020 to delineate the time before COVID-19 and after COVID-19.

The rationale is as follows: if COVID-19 did not have an effect on housing prices, the data from 2020 and 2021 would have followed our models closely. If COVID-19 did have an effect on housing prices, the data from 2020 and 2021 would have veered off from our model prediction line.

We also designed an interactive map of Canada using a Plotly geographic scatter plot to complement our presentation.

- All 11 regions were assigned a co-ordinate point on the map matching their geographical location.
- There is a time slider at the bottom, so the user can see how the sales pair count (number of houses used to calculate index) and index changed over time.
- Heatmap: The radius of the coloured circle around the co-ordinate points is proportional to the number of sales pair counts, and the colour changes gradients depending on the value of its index.

Since a high housing index could be caused by a few very expensive houses sold in a low sales pair count, this map also helps the user interpret our results. If the sales pair count was high along with the index, it would show that the average pricing of housing was truly high that year.

5 Instructions

5.1 Obtaining Datasets

Please click on this link. https://housepriceindex.ca/index-history/

Fill in your details, read over the terms and conditions, accept them, and then proceed to submit the details.

A download button should appear. Press download. The name of the downloaded csv should be House_Price_Index.csv. Place the file in the main project folder.

5.2 Running the program

1. Please install all Python libraries listed under the requirements.txt file that has been submitted.

2. Run the main.py file.

3. HTML files offering the user interactive scatter plots, the RMSE tables, and a map will open in three separate windows.

4. Double click on a location in the legend to isolate it in the graph preview

6 Changes

One of of our main goals was to augment the complexity of our programs and enhance clarity of the data being displayed. To do this, we implemented several changes.

Initially, we planned to use linear regression alone to analyze the housing price data from the Canadian market. We included multiple regression methods to our analysis to be more thorough in our modelling of the data.

We also incorporated a geographical scatter plot of Canada with our data, as previously explained. This aids visualization of the changing Canadian housing market by location.

Lastly, we included interactivity in our program to make our models intuitive for user exploration. This was accomplished in Plotly by allowing the user to toggle between the plots of the true prices for c11 and all cities, the linear regression models, the exponential regression models, and the support vector regression models. This allows the user to easily isolate and compare the actual data with the various regression models for the locations the user wants.

7 Discussion

7.1 Results

The results of our computational exploration greatly supplemented our analysis of the impacts of COVID-19 on the Canadian housing market. Utilising the various regression models, our program made predictions of the house index price values that would have been expected without COVID-19's intervention.

By comparing the actual data with our program's predictions, we found a general trend: there was a large and disproportionate increase in the index price for most of the cities over the years 2020 and 2021. Only Calgary, Edmonton, and Quebec bucked the trend. This trend is useful in answering our research question, as it demonstrates a correlation between COVID-19 and a large increase in prices in the Canadian housing market. We cannot say that correlation implies causation, but given the effects of COVID-19 on the economy, the link between the two events is obvious.

Furthermore, the gradient of the plot of the actual values over the years 2020 and 2021 is much steeper for the majority of that portion of the plot than either of the regression models. This suggests that it is very likely that in the near future, prices will continue to increase at a disproportionately higher rate than expected.

To evaluate the accuracy of our regression models, we calculated the Root Mean Square Error values of our models for the various locations. Generally, the SVM model was the most accurate as it had the smallest training RMSE values, which are displayed in the table below.

RMSE Values for SVM Model			
Area Name	Training RMSE (4d.p.)	Test RMSE (4d.p.)	RMSE Ratio (Test/Training)
Composite 11	6.2072	10.6418	1.7144
BC, Victoria	12.8794	12.0696	2.2305
BC, Vancouver	13.3161	22.8899	3.6714
AB, Calgary	19.0731	43.7966	1.7665
AB, Edmonton	23.9309	53.3782	0.9371
MB, Winnipeg	13.3042	30.5480	2.2961
ON, Hamilton	12.4000	63.6866	5.1360
ON, Toronto	14.4739	48.3460	3.3402
ON, Ottawa	8.7079	31.9704	3.671
QC, Montreal	8.3263	14.7083	1.7664
QC, Quebec	12.3158	37.5930	3.0524
NS, Halifax	9.8872	17.3356	1.7533

On average, our Test RMSE was much higher than our Training RMSE. A higher RMSE shows that our model was unfit to extrapolate data for years 2020

and 2021. This was in-line of our expectations since we had already suspected COVID-19 may have impacted house indexes from our preliminary research. We can use the RMSE ratio as an indicator to compare the weight that COVID presumably had on each city individually. From the data gathered, we can see that the highest RMSE ratio was found in Hamilton, Ontario which is very evident by the graph in our analysis. The lowest RMSE ratio was actually in Edmonton, Alberta where as seen by the graph, was actually under predicted by our models. Judging by the huge spike in housing prices around the 2008 housing crisis, we can assume that our model was negatively impacted by the irrationality in our data around that time period.

7.2 Limitations

Our analysis does have limitations. For one, it is impossible to isolate COVID's effects on housing prices without any confounding external factors. Other factors may have contributed to the increase in housing prices, making it easy to mistake correlation for causation. These factors themselves may have also been caused by COVID-19 to some extent; it is impossible to discern and quantify COVID-19's impacts.

Also, our models are imperfect. Though we may have evaluated their accuracy through their Root Mean Squared Error values, our models still have room for error so we cannot be completely certain of our conclusion (especially since we are extrapolating the models).

For example: in 2008, there was a severe global financial crisis which had wide-ranging effects, and resulted in volatile housing prices in the Canadian housing market. Although our trained models did include this data, our models cannot accurately reflect the sudden spike/dip in prices in the various housing markets due to their modelling equations. Therefore, our predictions are further from the actual prices around that time period.

7.3 Further Exploration

While this program proved very useful at making predictions for the index price values for the Canadian housing market, there is still room for further analysis and exploration.

This program could be used to analyze other housing markets around the world, and figure out how different countries' housing markets have been affected by the pandemic. It could be interesting to see the difference between developed and developing countries' housing prices. Then again, many developing countries may not have the data necessary for this investigation.

Alternatively, we could look inwards by focusing on different types of housing within a city, such as apartments versus homes, or differences between districts/neighbourhoods in a city. We could see how different neighbourhoods have been perceived historically and how that perception has changed over time. We could also look at commercial properties as well, or look at rent prices.

We could also explore the question "**Why** has the COVID-19 pandemic impacted prices in the Canadian housing market?" Since our program has helped us draw a conclusion about how the pandemic impacts the housing market, finding out why this is the case can be very useful, and can possibly mitigate the harsh effects on house prices.

7.4 Conclusion

In this project, we utilised Python and its libraries for statistical modelling to develop a program modelling predicted housing prices in Canadian cities. The plotted graph allows for the easy and accurate comparison of our models' predicted prices with actual housing prices in Canada. We were then able to compare these sets of prices to conclude that the COVID-19 pandemic has caused a disproportionate increase in housing prices, while displaying the results in a comprehensive and lucid manner.

8 References

CHMC (2021). Housing market assessment. Retrieved 5 Nov 2021 from https://www.cmhc-schl.gc.ca/en/professionals/housing-markets-data-and-research/market-reports/housing-market/housing-market-assessment.

Statista (2021). Canada: GDP by industry 2021. Technical report, StatistaResearch Department. Retrieved 5 Nov 2021 from https://www.statista.

com/statistics/594293/gross-domestic-product-of-canada-by-industry-monthly/.

Teranet-NBC (2021). Teranet-National Bank House Price Index. Retrieved 5 Nov 2021 from https://housepriceindex.ca.

Libraries:

(pandas) McKinney, W., & others. (2010). Data structures for statistical computing in python. In Proceedings of the 9th Python in Science Conference (Vol. 445, pp. 51–56). https://pandas.pydata.org/docs/.

(plotly) Plotly Technologies Inc (2015). Collaborative data science. https://plot.ly

(scikit-learn) Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... others. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825–2830. https://scikit-learn.org/0.21/documentation.html