An Exploratory Analysis for Beginners

Housing Price Index, Consumer Price Index and Energy Prices in Canada

Umaid Muhammad Zaffar
20963456
DATASCI 702
Supervised by **Michael Jack Davis**

Table of Content

Introduction

The Canadian economy is characterized by its resilience and diversity, and key economic indicators such as the Housing Price Index (HPI), Consumer Price Index (CPI), and energy prices play pivotal roles in shaping its trajectory. This survey paper embarks on a comprehensive exploration of time series analysis specific to Canada, focusing on HPI, CPI, and energy prices. It serves as a valuable resource for those seeking to comprehend the dynamics of housing prices in the Canadian context.

The motivation behind this paper stems from the growing importance of HPI, CPI, and energy prices in the Canadian economic landscape. These indicators wield significant influence over the financial well-being of Canadians, impacting decisions made by individuals, businesses, and policymakers alike. The need for a comprehensive understanding of the time series data associated with these factors becomes increasingly evident.

This paper follows a structured approach, offering a step-by-step guide to unravel the intricacies of time series analysis. Beginning with fundamental concepts, it navigates through the core principles of time series, elucidating vital components such as stationarity, seasonality, and trend analysis.

Furthermore, the paper conducts a deep-dive into the unique characteristics of Canadian HPI, CPI, and energy price time series data. By analyzing historical trends, patterns, and correlations within these datasets, it provides valuable insights into their economic significance within the Canadian context.

In summary, this survey paper acts as an educational tool and reference guide tailored to those interested in analyzing HPI, CPI, and energy price time series data specific to Canada. By combining foundational time series analysis principles with real-world insights into the Canadian economy, we endeavor to equip readers with the knowledge and tools needed to navigate the dynamic realm of Canadian housing price analysis.

Basic Terminologies

What is the House Price Index?

The Housing Price Index (HPI) is a statistical measure that tracks changes in the prices of residential properties over time. It provides insights into the trends and fluctuations in the real estate market, specifically in the prices of homes or properties. HPI is an essential economic indicator used by individuals, businesses, and policymakers to gauge the health and stability of the housing sector.

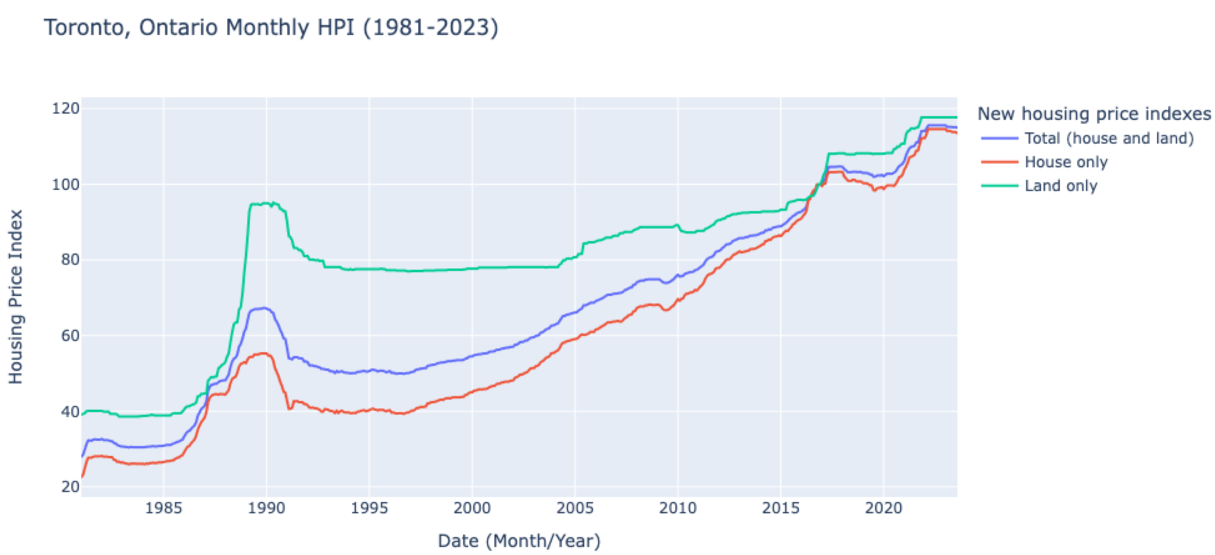


Figure 1: Toronto House Price Index Monthly

Calculations:

1. Selection of a Reference Period: A base period is chosen as a reference point for comparison. This is usually a specific month or quarter, and its HPI value is set to 100.
2. Selection of a Sample of Properties: A representative sample of residential properties is selected. These properties should reflect the diversity of the housing market in terms of location, size, type (e.g., single-family homes, apartments), and other relevant characteristics.
3. Data Collection: Data on the selling prices of these selected properties are collected at regular intervals (e.g., monthly or quarterly) over time.

4. Calculation of Price Changes: The price changes for each property are calculated by comparing their selling prices in the current period to their prices in the base period. This calculation is typically done using a formula like the Laspeyres index:

$$HPI = \frac{P_t}{P_0} \times 100\%$$

Where:

- HPI is the Housing Price Index for the current period.
- P_t is the aggregate price of properties in the current period.
- P_0 is the aggregate price of properties in the base period.

What is the Consumer Price Index?

The Consumer Price Index (CPI) is a widely used economic indicator that measures the average change in prices paid by consumers for a basket of goods and services over time. It reflects the rate of inflation and is a crucial tool for assessing changes in the cost of living for households.

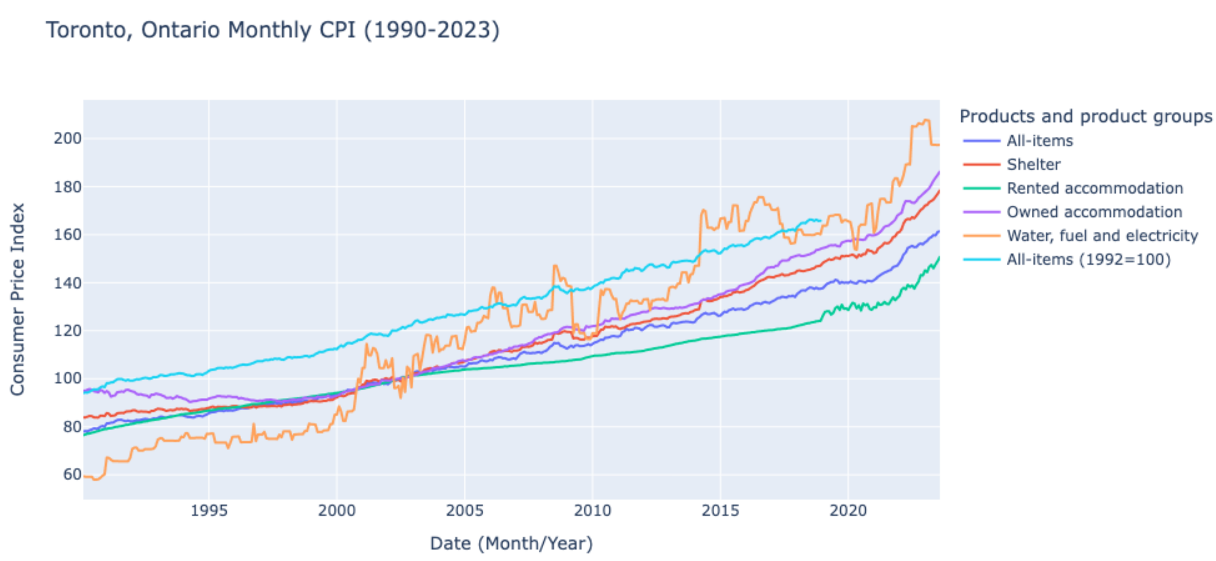


Figure 2: Toronto Consumer Price Index Monthly

1. Selection of a Basket of Goods and Services: A representative basket of goods and services is selected to reflect the typical consumption patterns of an average

consumer or household. This basket includes items such as food, clothing, housing, transportation, healthcare, and more.

2. Data Collection: The prices of the items in the basket are collected regularly (e.g., monthly) from various outlets, stores, or providers.
3. Calculation of Price Changes: The percentage change in prices for each item in the basket is calculated by comparing its current price to its price in the base period. This is typically done using the Laspeyres formula:

$$CPI = \frac{\Sigma P_t \times W_0}{\Sigma P_0 \times W_0} \times 100\%$$

Where:

- CPI is the Consumer Price Index for the current period.
- P_t is the price of each item in the current period.
- P_0 is the price of each item in the base period.
- W_0 is the weight of each item in the base period (reflecting its importance in the consumer's budget).

What is Energy Price?

Energy prices encompass the costs of various energy forms such as oil, natural gas, electricity, coal, and renewables, and their fluctuations are pivotal in shaping a nation's economic landscape. A stable and reasonable energy price propels industries, fuels transportation, and supports household energy needs, essentially acting as a catalyst for economic growth. On the flip side, unforeseen escalations in energy prices can dampen economic growth, as industries and consumers adjust their spending due to increased operational and living costs. Moreover, the trade dynamics of a nation are also intricately tied to energy prices. For instance, countries abundant in energy resources, like Canada, see their trade revenues soar with a rise in global energy prices, while those reliant on imports can experience a strain on trade balances. Furthermore, the energy sector's health directly impacts job markets, investments, and fiscal policies. In essence, energy prices serve as a barometer for a nation's economic health, reflecting the interplay of global demand-supply chains, geopolitical events, and internal policy decisions.

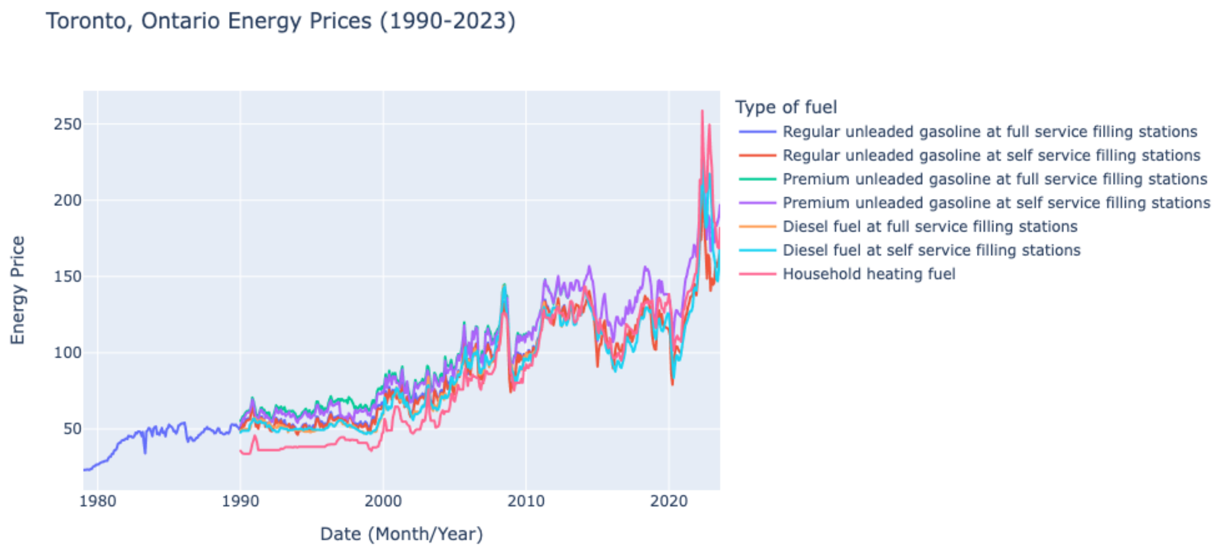


Figure: Energy Prices Monthly

Time Series Preparation

We can see that HPI and CPI are time series and the following paper goes through a brief method on processing this data for time series analysis. More specifically, we will be focusing on **Granger Causation**, **Seasonal Decomposition**, **Modeling** and **Forecasting**.

Dataset

The dataset was sourced from the Statistics Canada website https://www150.statcan.gc.ca/n1/en/type/data?MM=1, combining data from three distinct categories into a consolidated dataset. Specifically, the Consumer Price Index (CPI) data was extracted from the following repository:

https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1810000413, which encompasses CPI records spanning from 1918 to 2023. Users have the option to select a reference period for establishing the base. In the example provided, the period was truncated to cover the years from 1990 to 2023.

Similarly, the Housing Price Index (HPI) data was obtained from https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1810020501 and includes data from 1981 to 2023. Here too, users can specify a reference period for the HPI dataset.

Additionally, energy prices data were sourced from https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1810000101, spanning from 1979 to 2023. For the purpose of the illustration, the time frame was reduced to encompass the years from 1990 to 2023.

Preprocessing

1. Dropping Irrelevant Columns:
   Irrelevant columns are removed from each DataFrame containing HPI, CPI, and energy price data. These columns include information such as unique identifiers (`DGUID`), units of measurement (`UOM`), and other metadata that are not needed for the analysis.

2. Formatting Date:
   The "`REF_DATE`" column in each DataFrame is converted to a datetime format using the pd.to_datetime() function. This ensures that the date information is in a standardized format for further analysis.

3. Filtering Data:
   Additional filtering is applied to each DataFrame:

a. In the HPI DataFrame, only rows where the "`New housing price indexes`" column is "`Total (house and land)`" are retained. This filters the data to focus on total housing price indexes.
b. In the energy DataFrame, only rows where the "`Type of fuel`" column is "`Regular unleaded gasoline at self service filling stations`" are retained. This filters the data to focus on regular unleaded gasoline prices.
c. In the CPI DataFrame, only rows where the "`Products and product groups`" column is "`All-items`" are retained. This filters the data to focus on the consumer price index for all items.

4. Date Range Selection:
   Further filtering is applied to each DataFrame to select data only from February 1990 onwards. This narrows down the dataset to a specific time period of interest.

5. Merging DataFrames:
   Finally, the three DataFrames containing HPI, CPI, and energy price data are merged into a single DataFrame named merged_data using the "`REF_DATE`" column as the common key. Suffixes are added to columns with the same name in the merged DataFrame to distinguish between the sources of data (HPI, CPI, and energy prices).

Granger Causation

Granger causation is a statistical concept and test used to assess whether one time series can predict another time series. It is named after Clive Granger, a Nobel laureate in economics, who developed the concept in the context of econometrics. Granger causality helps us understand the potential causal relationship between two time series variables.

Here are the key points to understand about Granger causation:

1. Causation vs. Correlation:
   Granger causation does not imply a cause-and-effect relationship in the traditional sense of causation. Instead, it explores whether one time series contains information that can help predict another time series. It deals with statistical causality, not necessarily a true underlying cause.

2. Lagged Variables:
   Granger causation relies on the concept of lagged variables. In a time series, you can examine whether the past values of one variable (the potential cause) can help predict the future values of another variable (the potential effect).

3. Null Hypothesis:
   The Granger causality test involves formulating a null hypothesis (H0) that there is no causal relationship between the two time series. In other words, the past values of one series do not contain information that helps predict the future values of the other series.

4. Statistical Test:
   To test for Granger causation, statistical tests are performed. These tests involve estimating autoregressive models (AR models) for each of the two time series, one with and one without lagged values of the potential cause variable. The improvement in prediction accuracy with the inclusion of lagged values is used to determine causality.

5. Interpretation:
   If the inclusion of lagged values of the potential cause variable significantly improves the prediction of the potential effect variable, the null hypothesis is rejected. This suggests that there is evidence of Granger causality, indicating that the potential cause variable has predictive power for the potential effect variable.

Granger causality is a valuable tool for exploring relationships between time series variables and has applications in fields such as economic forecasting, finance (e.g., stock price prediction), and climate science (e.g., climate variables influencing each other). It is important to note that while Granger causality can detect statistical relationships, it does not establish a true causal mechanism or imply causation in a broader sense. To ensure that Granger causality tests provide meaningful and accurate results, it is essential to verify that the time series data involved are stationary. This typically involves checking for stationarity through statistical tests and, if needed, applying transformations like differencing to make the data stationary. Only then can you confidently proceed with Granger causality analysis to investigate potential causal relationships between the time series variables.

Stationarity

The Augmented Dickey-Fuller (ADF) test is used to assess whether a time series is stationary by examining the presence of unit roots, which indicate non-stationarity. The test involves estimating an autoregressive model of the time series and then evaluating whether the coefficient of the lagged first difference (representing the impact of previous periods on the current value) is significantly different from zero. If the coefficient is significantly different from zero, indicating that differencing the series removes the unit root, the series is considered stationary. Conversely, if the coefficient is not significantly different from zero, suggesting the presence of a unit root, the series is non-stationary. The ADF test is valuable in time series analysis for ensuring that the data's statistical properties, such as mean and variance, do not change over time, which is a fundamental assumption in many time series models and analyses. We use the function `adfuller` from `statsmodel.tsa.statstools`. First, we will conduct the stationarity check on one of the time series and explain each element in detail.

```
sts.adfuller(merged_df.VALUE_HPI)
```

```
✓ 0.1s
```

```
(0.14562956601923868,
 0.9690418522709915,
 9,
 266,
 {'1%': -3.455175292841607,
  '5%': -2.8724677563219485,
  '10%': -2.57259315846006},
 -229.99219849316398)
```

Figure 3: Augmented Dickey-Fuller Non-Stationary Demonstration

The null hypothesis of the ADF test is that the time series has a unit root, which means it is non-stationary. In simpler terms, it assumes that the time series exhibits a stochastic trend and is not stationary. The alternative hypothesis (H1) is that the time series is stationary, meaning it does not have a unit root.

The first value in the figure above indicates the **T statistic**. Conveniently enough, Python also provides us with the 1%, 5% and 10% critical values from the Dickey-fuller test. Having all three comes in handy because we might use any of them as levels of significance in our analysis. Furthermore, you can notice that the T statistic is greater than each of the critical values. So in this case, we do not find enough evidence for stationarity in this dataset.

The second value is associated with the **P Value.** P Value indicates the percentage of not rejecting the null hypothesis. In this case there is a 96% chance of not rejecting the null hypothesis. This indicates that it is extremely likely that the time series is non-stationary.

The third line indicates the **number of lags** used in the regression when determining the t statistic. In this case it is 9 that means there is some autocorrelation going back 9 periods.

The fourth line indicates the **number of observations** used in the analysis which depends on the number of lags. The two of these sum up to the total number of observations in the time series.

The last line is the **maximized information criteria** provided that there is some apparent autocorrelation. The lower the value, the easier it is to predict the future data.

Non Stationary to Stationary

There are a lot of methods to convert non-stationary data to stationary data: Differencing, Log Transformation, Detrending, Seasonal Decomposition, Box-Cox transformation, Deseasonalization. For the purpose of this paper, we will be going through the most basic version which is differencing. To perform this transformation we subtract the current value in the period with the previous period. This method calculates the difference between consecutive observations. It is useful when you have a time series with a clear trend, and differencing helps remove the trend component. Performing ADFuller test on this transformed series gives us this result:

```
1  sts.adfuller(hpi_series_diff)
✓ 0.0s

(-3.389854412688381,
 0.011309634739783988,
 8,
 266,
 {'1%': -3.455175292841607,
  '5%': -2.8724677563219485,
  '10%': -2.57259315846006},
 -230.17014789998717)
```

Figure 4: Augmented Dickey-Fuller Stationary Demonstration

We can clearly see here that the T statistic is less than 1% critical value and the P value is 1.1% which means it is now almost certain that the series is non-stationary.

In order to interpret the results of Granger Causality, we will list down the p-value along with the null hypothesis and explain them in relation with the time series.

The SSR (Sum of Squared Residuals) based F-test, also known as the Error Reduction Ratio F-test, is a statistical test used in the context of Granger causality testing and time series analysis. It assesses whether adding lagged values of a potential cause variable (or variables) to a regression model significantly reduces the sum of squared residuals compared to a model without those lagged values. In other words, it evaluates whether including past values of a variable improves the prediction of the dependent variable.

The p-value associated with the F-statistic indicates the probability of obtaining the observed F-statistic value under the null hypothesis (H0). A low p-value (typically below a significance level like 0.05) suggests that there is strong evidence against the null hypothesis, indicating Granger causality.

The following table summarizes the result of granger causality with the associated p value for SSR based F-test, SSR based chi2 test, Likelihood ratio test, Parameter based F-test.

Cause	Effect	SSR F test	SSR chi2 test	LLR test	Param F test
HPI	CPI	0.0026	0.0015	0.0020	0.0026
HPI	Energy	0.0048	0.0029	0.0037	0.0048
CPI	HPI	0.1569	0.1352	0.1415	0.1569
CPI	Energy	0.1825	0.1592	0.1658	0.1825
Energy	HPI	0.9202	0.9138	0.9143	0.9202
Energy	CPI	0.0018	0.0009	0.0013	0.0018

Table 1: Granger Causality P Value for 5 Lags

From these results we can deduce some results on potential causes and their effect. For Example HPI being a cause for CPI and Energy is quite likely because the P Value is lowe than 1% (Significant Value). If our P Value is low then we are able to reject the null hypothesis and HPI does indeed granger cause CPI and Energy Prices. However, we can see that CPI does not cause any change to HPI or Energy which is quite intuitive since CPI is the one which is actually calculated from Energy Prices. Lastly we can see that Energy Prices do not cause any change to HPI since P Value is quite large while it does affect CPI since in principle, Energy price constitutes the formula for calculating CPI.

Seasonality

The `seasonal_decompose` function from the `statsmodels.tsa` module in Python typically generates four figures when decomposing a time series:

Original Time Series:
The first figure displays the original time series data as it is without any decomposition. This plot provides a visual representation of the raw data, including any underlying trends, seasonality, and noise.

Trend Component:
The second figure shows the trend component extracted from the original time series. The trend represents the long-term movement or underlying pattern in the data, abstracting away shorter-term fluctuations and seasonal effects.

Seasonal Component:
The third figure illustrates the seasonal component of the time series. This component captures periodic patterns or seasonal variations that repeat at regular intervals within the data, such as daily, weekly, or yearly cycles.

Residuals (or Remainder):
The fourth figure displays the residuals, also known as the remainder or error component. These residuals represent the unexplained variation in the data after removing the trend and seasonal components. Residuals are essentially the noise or random fluctuations in the time series.

Each of these four figures provides valuable insights into the decomposition of the time series, allowing you to visually assess the presence of trends, seasonality, and any remaining patterns or irregularities in the data. This decomposition is useful for further analysis, forecasting, and modeling as it helps isolate and understand the underlying components of the time series.

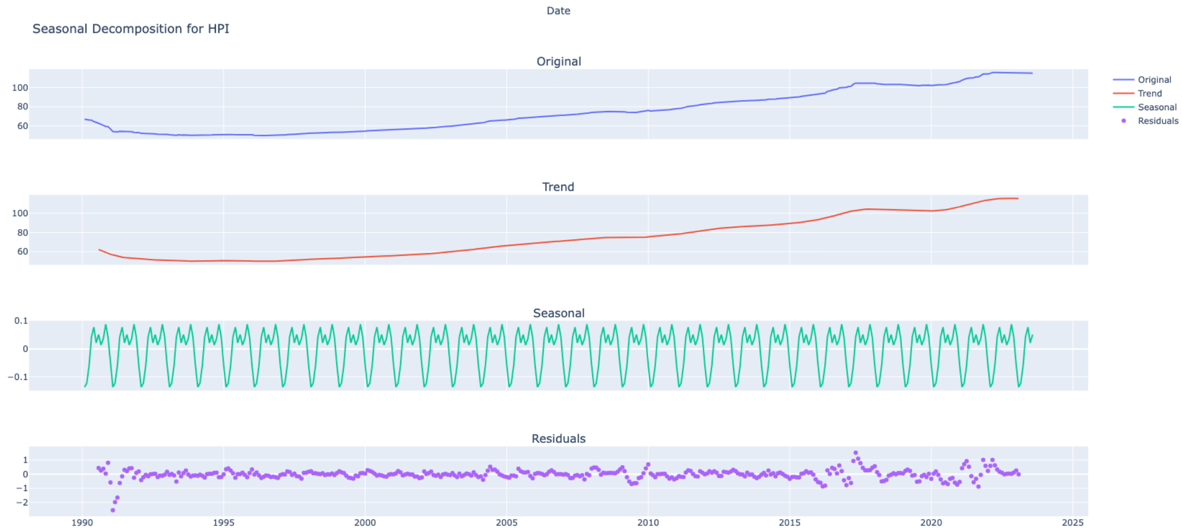


Figure 5: Seasonal Decomposition for HPI

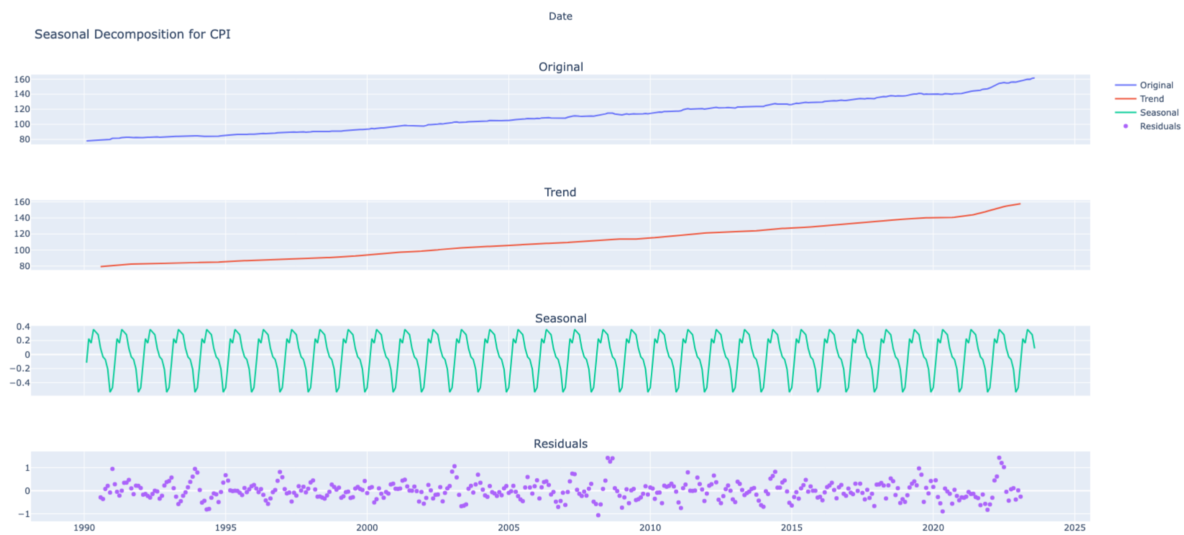


Figure 6: Seasonal Decomposition for CPI

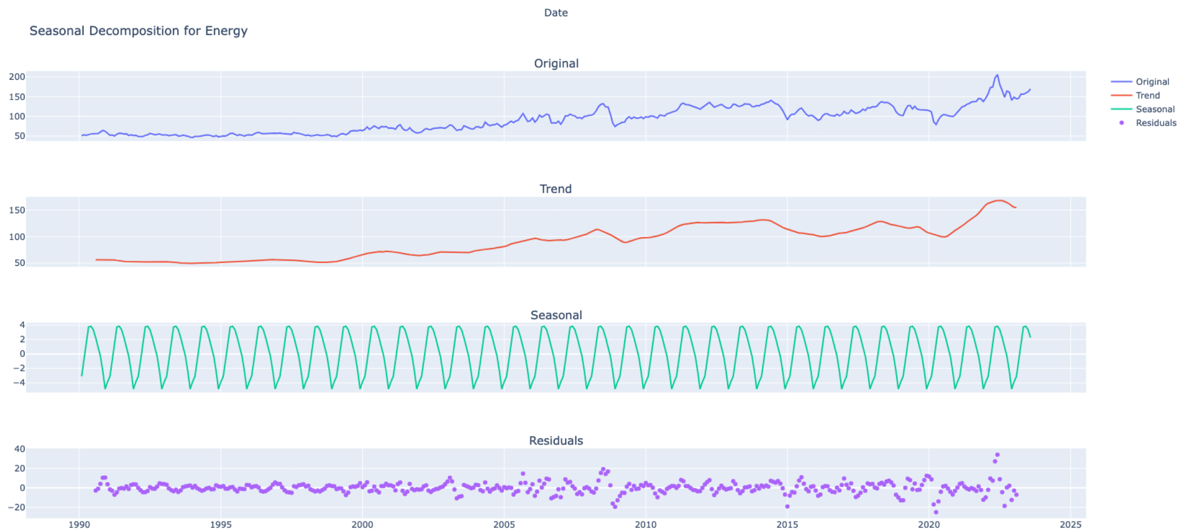


Figure 7: Seasonal Decomposition for Energy Prices

Analysis

The House Price Index in Toronto has experienced a general increase from 1990 to 2023, with a notably rapid growth between 2010 and 2020. There's a clear seasonal pattern in house prices, indicating certain times in the year when prices typically rise or fall. The residuals, although centered around zero, have occasional spikes indicating unaccounted events or anomalies in the data.

The Consumer Price Index in this dataset has shown a steady increase over the years from 1990 to 2023. There's a clear seasonal pattern that repeats yearly, suggesting some regular annual economic influences. The residuals suggest some early variance, stabilization, and then an increase in anomalies or unexpected events in recent years.

The energy prices remained relatively stable for a long period but started rising around 2015. This upward trend, combined with the clear seasonal patterns and increased residuals' volatility, indicates a more complex and dynamic energy market in recent years.

Autocorrelation

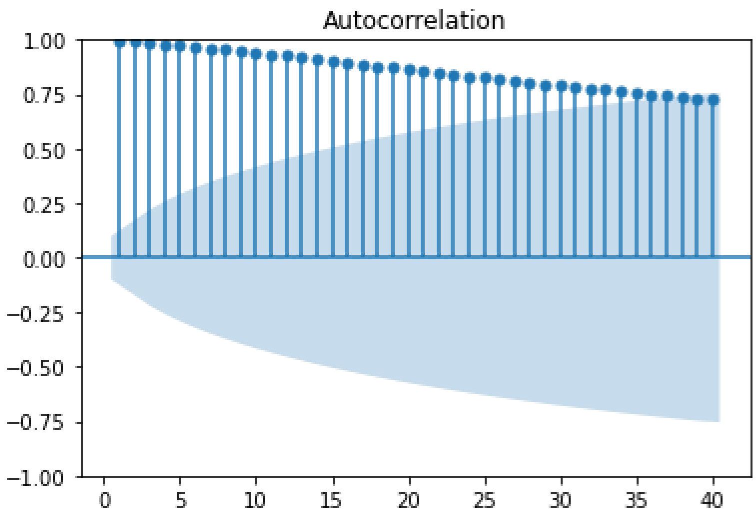


Figure 8: HPI Autocorrelation

The autocorrelation starts near 1 at lag 0, which is expected since a series is perfectly correlated with itself. It gradually decreases but remains positive and significant for all 40 lags. This indicates a strong positive autocorrelation, suggesting a slow decay or trend in the HPI time series. The persistently high autocorrelation indicates that HPI values from several previous periods (lags) could be useful predictors for future values.

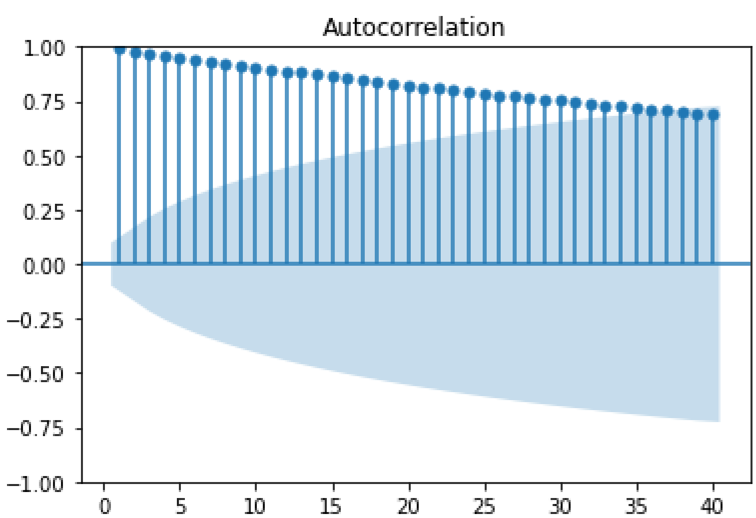


Figure 9: CPI Autocorrelation

Just like HPI, the CPI autocorrelation starts near 1 at lag 0. It also gradually decreases, but the decay seems to be a bit faster compared to HPI. Still, it remains positive and significant for a majority of the 40 lags. This indicates a strong positive autocorrelation, though slightly weaker than that of HPI. This suggests that while previous values of CPI are useful predictors, they might become less influential as the lag increases, especially when compared to HPI.

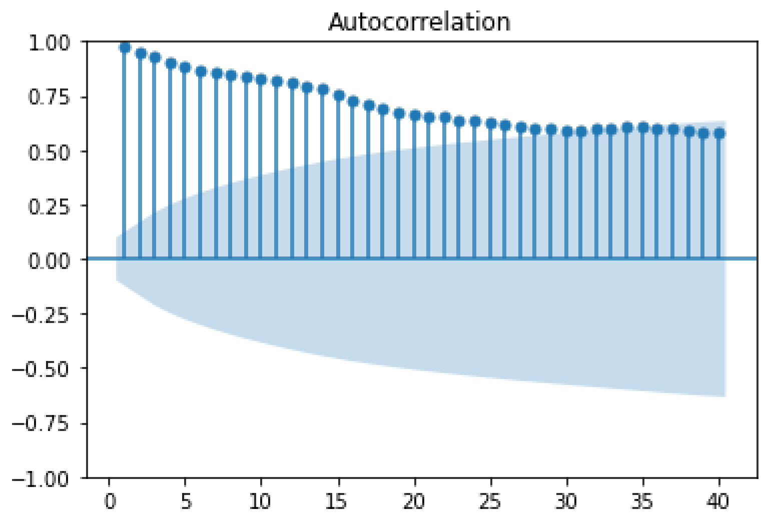


Figure 10: Energy Prices Autocorrelation

The autocorrelation pattern for Energy Prices appears similar to that of CPI, starting near 1 at lag 0 and gradually decreasing. The decay is somewhat steeper than HPI but is roughly in line with the CPI's decay pattern. This indicates that the autocorrelation is positive and strong, but not as persistent as HPI's. It suggests that while past values of Energy Prices can be predictors, their influence wanes a bit more rapidly than HPI.

Partial Autocorrelation

Partial autocorrelation plots (PACF) are similar to autocorrelation plots (ACF), but they show the correlation of a time series with its own lagged values, controlling for the values of the time series at all shorter lags. It helps in determining the order of an autoregressive (AR) term in a time series model.

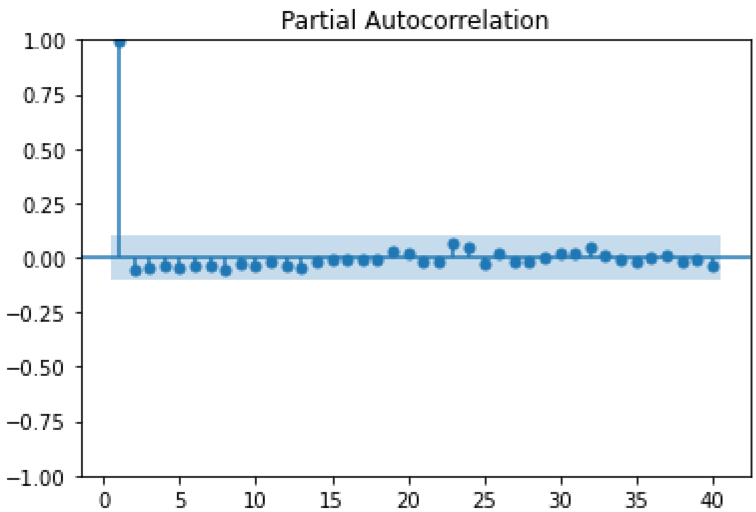


Figure 11: HPI Partial Autocorrelation

At lag 0, we see a value of 1, which is expected since any time series is perfectly correlated with itself. The significant spike at lag 1 indicates that the value of HPI has a strong correlation with its immediate previous value, after adjusting for other lags. For lags beyond 1, the partial autocorrelation values hover around zero and remain within the confidence bounds (the shaded region). This suggests that, once the immediate previous value (lag 1) is accounted for, the values of HPI at other lags don't provide much additional explanatory power.

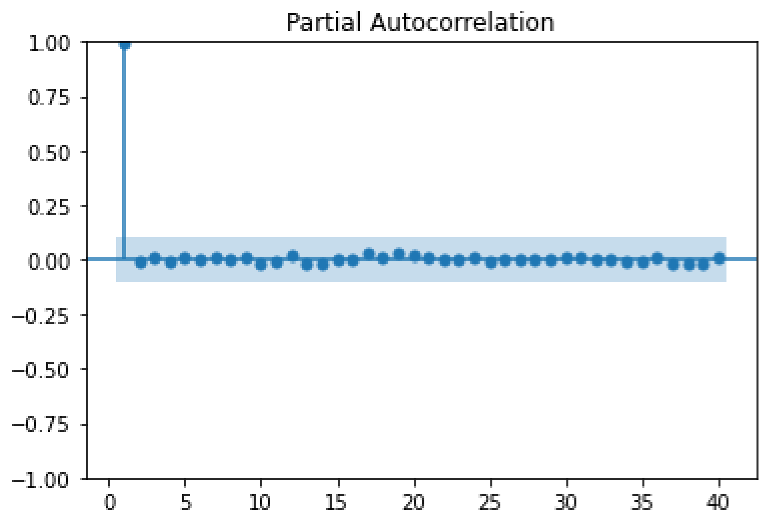


Figure 12: CPI Partial Autocorrelation

As with HPI, the PACF for CPI starts with a value of 1 at lag 0. There's a notable spike at lag 1, signifying a strong correlation of the CPI value with its previous value, after adjusting for the influence of other lags. Beyond lag 1, the partial autocorrelation values for CPI also fluctuate around zero and stay within the confidence bounds. This indicates that, once we account for the immediate previous value (lag 1), the other lagged values don't offer much more information for explaining the variation in CPI.

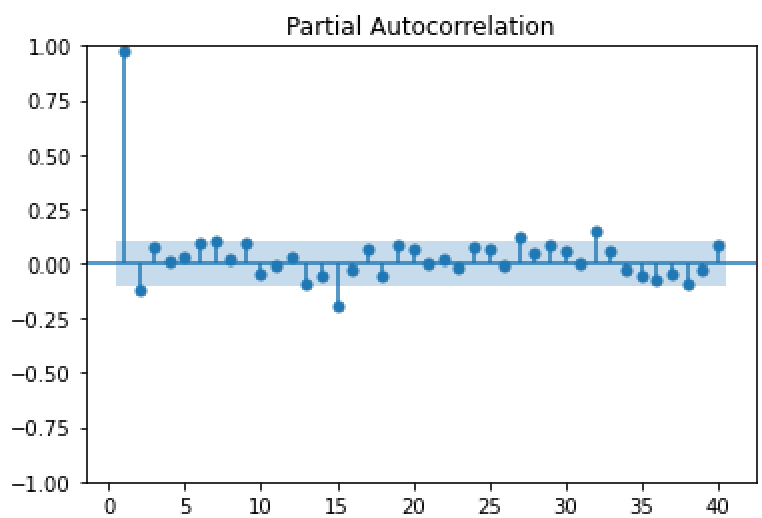


Figure 13: Energy Prices Partial Autocorrelation

At lag 0, the PACF value is 1, as expected since any series is perfectly correlated with itself. There's a significant spike at lag 1, indicating a strong correlation between the current value of Energy Prices and its immediate previous value, after controlling for other lags. Beyond lag 1, the PACF values fluctuate around zero. Although some values lie slightly outside the confidence bounds (shaded region), they aren't particularly prominent or consistent enough to suggest a strong partial autocorrelation at higher lags.

Modeling and Forecasting

Modeling

We know that HPI and Energy Prices contain information to predict CPI prices. That is, past values of HPI and Energy prices influence the value of CPI along with its own lagged values. So we will use this information to postulate a model to forecast values of CPI. In order to do this we use the `auto_arima` function. `auto_arima` is a function from the `pmdarima` library, designed to automatically discover the optimal order for an ARIMA model. ARIMA stands for AutoRegressive Integrated Moving Average. The function performs a grid search over multiple combinations of `p, d, q` (for ARIMA) and `P, D, Q` (for seasonal components) parameters, and returns the best ARIMA model according to the provided criteria (like AIC, BIC, etc.). In this process we use the data from 1990 to 2020 for training the model and from 2020 onwards to test model accuracy.

Component of ARIMA

AR (Autoregression): Denoted by the parameter `p`.
It refers to the use of past values in the regression equation for the time series. The premise being that past values have an effect on current values.

I (Integrated): Denoted by the parameter `d`.
Represents the number of differences needed to make the time series stationary (i.e., data values are not a function of time). A stationary time series' properties do not depend on the time at which the series is observed.

MA (Moving Average): Denoted by the parameter `q`.
Uses past forecast errors in a regression-like model. It means that the regression error is actually a linear combination of error terms whose values occurred both in the current and at various times in the past.

Exogenous Variables

Exogenous variables are external variables that aren't a part of the time series itself but might have an influence on its behavior. In the code you've provided, `VALUE_HPI` (House Price Index) and `VALUE_ENERGY` (Energy Prices) are treated as exogenous variables to the `VALUE_CPI` series.

SARIMAX Results

Dep. Variable:	y	No. Observations:	359
Model:	SARIMAX(0, 1, 0)x(1, 0, [1], 12)	Log Likelihood	-152.801
Date:	Tue, 31 Oct 2023	AIC	313.602
Time:	17:12:47	BIC	329.125
Sample:	02-01-1990	HQIC	319.776
	- 12-01-2019		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	0.0096	0.008	1.250	0.211	-0.005	0.025
ar.S.L12	0.9460	0.038	24.847	0.000	0.871	1.021
ma.S.L12	-0.7916	0.067	-11.814	0.000	-0.923	-0.660
sigma2	0.1353	0.009	14.745	0.000	0.117	0.153

Ljung-Box (L1) (Q):	0.15	Jarque-Bera (JB):	5.75
Prob(Q):	0.70	Prob(JB):	0.06
Heteroskedasticity (H):	2.28	Skew:	0.05
Prob(H) (two-sided):	0.00	Kurtosis:	3.61

Figure 14: Best Model Summary Sample

We can observe in the above model that the order is (0,1,0) meaning that AR order (`p`): 0, Differencing order (`d`): 1, MA order (`q`): 0. Seasonal AR order (`P`): 0, Seasonal Differencing order (`D`): 1, Seasonal MA order (`Q`): 1, Seasonal frequency (m): 12. We set the seasonal frequency to 12 because we observed in the seasonal decomposition that there was a cyclical pattern every year.

Forecasting

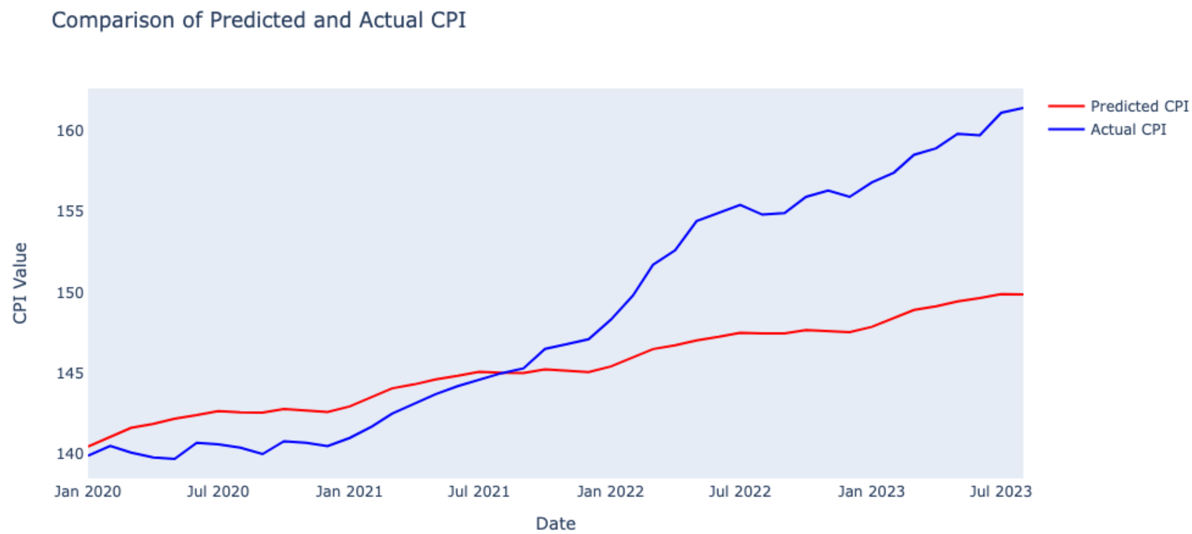


Figure 15: Comparison of Predicted CPI incorporating a Seasonal Trend of 12 Month VS Actual CPI

After plotting the forecasting CPI predictions with Actual CPI from January 2020 to August 2023 we can see that the trend for Predicted CPI greatly resembles the Actual CPI. The predictions do not precisely match the Actual CPi because of the current unprecedented economical situation of the country. Now lets try the same model but while incorporating seasonality from the last 6 periods inst

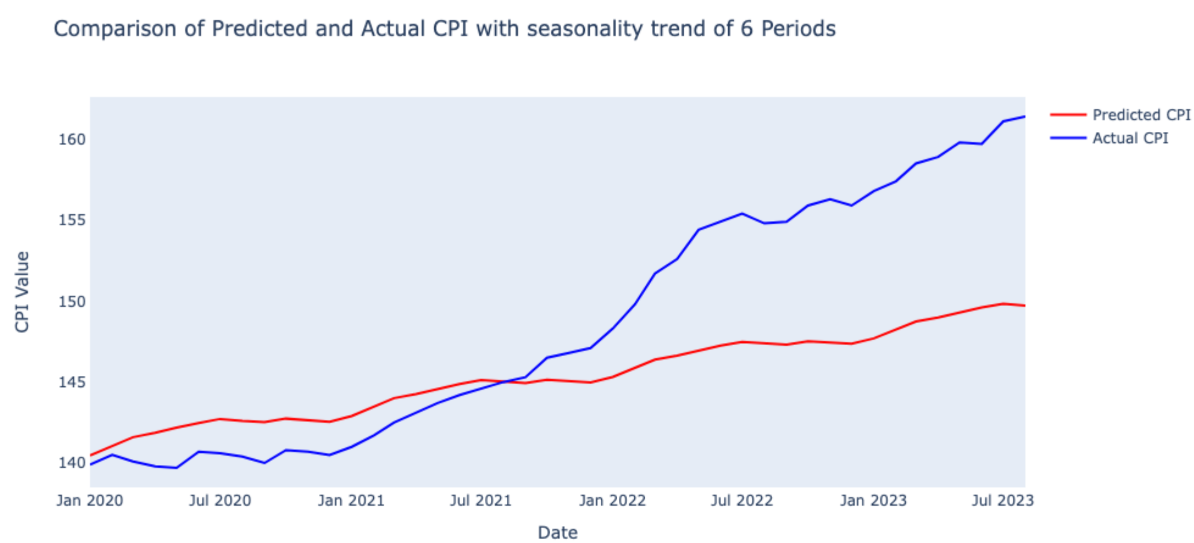


Figure 16: Comparison of Predicted CPI incorporating a Seasonal Trend of 6 Month VS Actual CPI

Here we can see that the predicted line is almost exactly the same as the one with seasonality trends recurring in 12 months. However, if we take a closer look by carefully analyzing the Mean Squared Errors. The MSE of 12 Period Season Prediction is **33.2416** while MSE of 6 Period Season Prediction is **34.0867** which is quite similar but confirms that MSE of 12 Period Season Prediction is better.

We can also inspect some more Auto ARIMA Models to make sure that our 12 Period Seasonal Model is the best model by inspecting the MSE of each of them. We keep the same exogenous variables for each test and record results.

Seasonal Trend	Order (p, d, q, P, D, Q)	Mean Squared Error
16	(0 ,1 ,0, 2, 0, 1)	42.20
12	(0, 1, 0, 1, 0, 1)	33.24
6	(0, 1, 0, 2, 0, 2)	34.08
4	(3, 1, 1, 1, 0, 1)	38.14
3	(0, 1, 0, 2, 0, 2)	46.38

Table 2: MSE for Forecasting Models

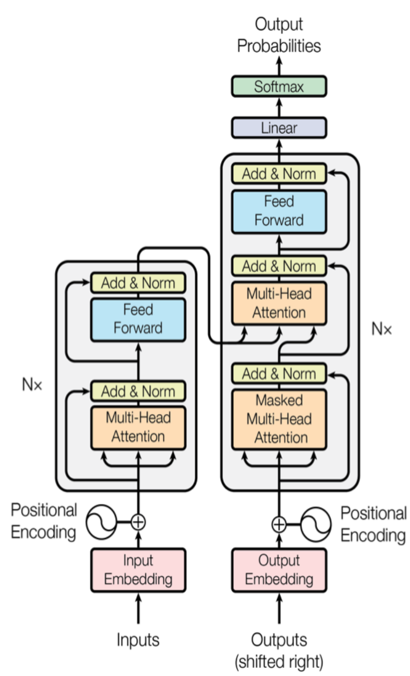
From the table we can confirm that using the fact that there is a cyclical pattern during the last 12 months we can postulate a model that can help us forecast the CPI values. Since the P and Q in this model is 1 means that the model takes into account the lag from exactly 12 periods (1 year) while considering the Autoregressive order and the Moving Average.

Neural Networks

Neural networks form the bedrock of modern artificial intelligence (AI) and machine learning. Inspired by the biological neural networks in human brains, these computational models are designed to recognize patterns and solve complex problems. A typical neural network comprises layers of interconnected nodes or neurons, each processing input data and forwarding the output to subsequent layers. Through training, neural networks adjust their internal parameters, enabling them to interpret and respond to diverse data inputs effectively. These models have been pivotal in advancements across various fields, including image and speech recognition, natural language processing (NLP), and complex decision-making.

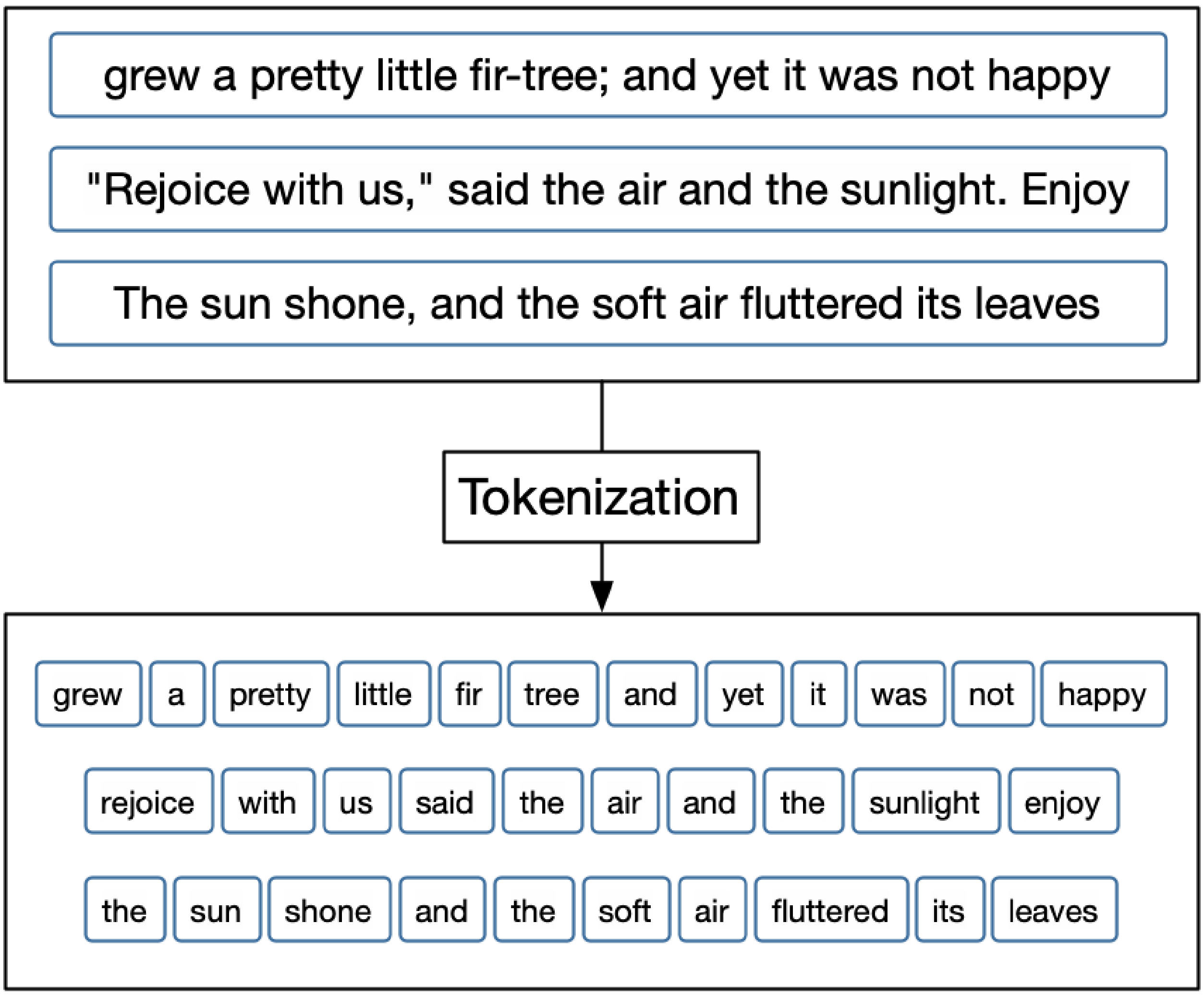
Evolution to Transformers

Transformers, introduced in the groundbreaking paper "Attention Is All You Need" in 2017, mark a significant evolution in neural network architecture. Traditional neural network designs, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have limitations—CNNs are less suited for sequential data, and RNNs struggle with long-range dependencies. Transformers overcome these challenges with a unique structure based on 'self-attention'. This mechanism allows the network to process entire sequences of data in parallel, significantly improving efficiency and effectiveness in handling sequential data. Transformers can weigh the significance of different parts of the input, focusing on relevant segments for tasks like translation or summarization.



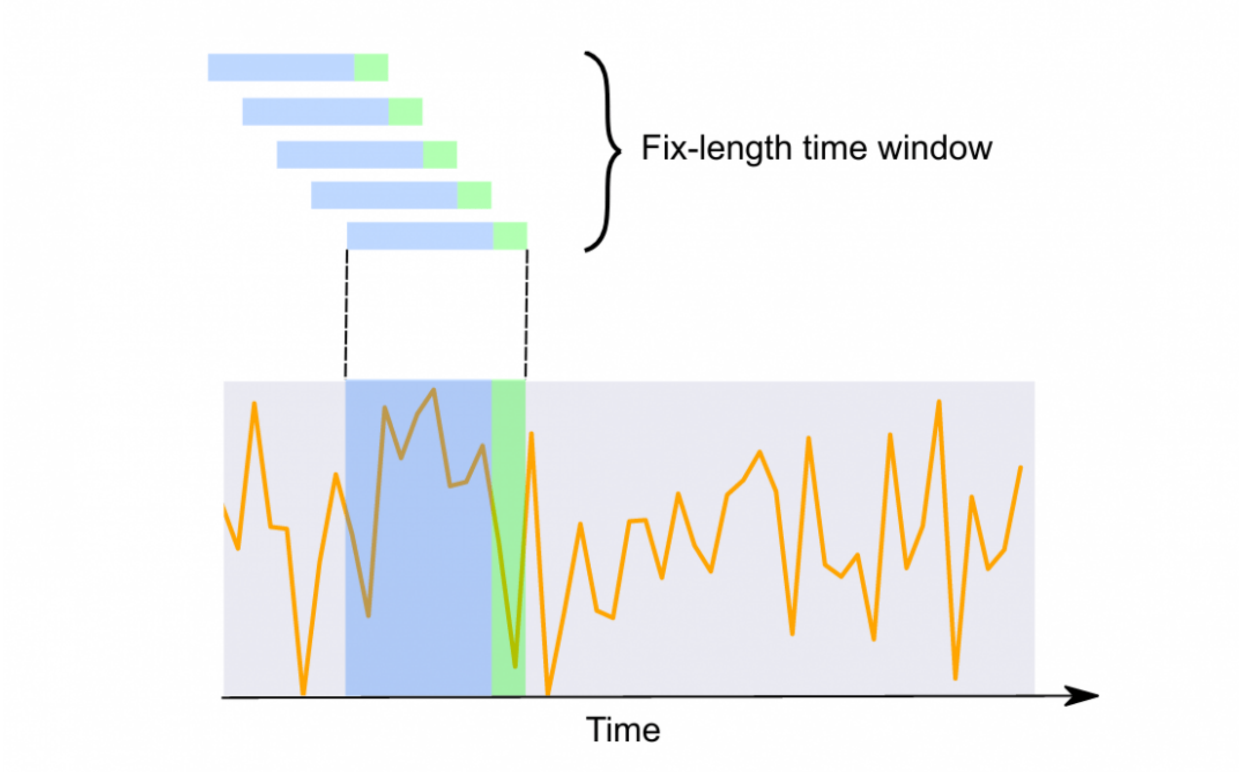
Transformers in Natural Language Processing

In NLP, transformers have revolutionized the field. Unlike RNNs that process text sequentially, transformers view an entire sequence at once, capturing intricate dependencies irrespective of distance within the sequence. This capability has led to state-of-the-art performance in language translation, text generation, and sentiment analysis. Models like OpenAI's GPT series and Google's BERT are based on transformer architecture, demonstrating exceptional proficiency in understanding and generating human language.



Transformers for Forecasting

Beyond NLP, transformers have shown promise in forecasting, a domain traditionally dominated by time-series models. Forecasting involves predicting future values based on past data, which is inherently sequential. Transformers can be adapted for forecasting by treating time-stamped data as a sequence, similar to words in a sentence. This allows the model to capture temporal dependencies, a crucial factor in accurate forecasting. For example, in stock market prediction, a transformer can learn complex patterns in historical price data, considering both short-term fluctuations and long-term trends. Similarly, in weather forecasting, transformers can process vast arrays of historical weather data, identifying patterns that traditional models might miss.



Analogy: Transformers in Language and Forecasting

To understand how transformers apply to forecasting, consider an analogy with language processing. In language, especially in tasks like text generation, a transformer model predicts the next word in a sentence based on the context provided by previous words. This is similar to forecasting, where future data points are predicted based on past data.

Language Prediction: Setting the Scene

Imagine a transformer working on a sentence completion task. Given the beginning of a sentence, the model predicts the next word by considering the context—each word that has come before. It understands not just the immediate predecessor but the entire sequence, capturing nuances and dependencies that influence what word comes next. For instance, in the sentence "The weather today is...", a transformer, analyzing past words, might predict "sunny" or "rainy" as the next word.

Forecasting: Parallel in Time

Now, apply this concept to forecasting. Instead of words, the sequence consists of historical data points. Let's say we're predicting stock prices. The transformer looks at past prices in a manner akin to how it examines previous words in a sentence. Just as it understands that certain words are more likely to follow others, it recognizes patterns in how stock prices evolve over time. It acknowledges that today's price isn't just a function of yesterday's but a culmination of a trend over many days, weeks, or even months.

The Role of Self-Attention

The key to this capability is the transformer's self-attention mechanism. In language, it allows the model to weigh the importance of each preceding word in predicting the next. In forecasting, this translates to understanding which past data points (e.g., stock prices from specific days) are most influential in predicting the future. The transformer can thus make informed predictions by considering both recent events and long-term trends.

TimeGPT is a generative pre-trained model specifically designed for forecasting time series data. This model functions by using historical values of the time series and exogenous variables (external factors that might influence the time series) as inputs to make predictions. TimeGPT is versatile and can be applied to a wide range of tasks, including but not limited to demand forecasting, anomaly detection, and financial forecasting.

TimeGPT

The working principle of TimeGPT is analogous to how humans read a sentence, processing the data from left to right. It analyzes a window of past observations from your time series, treating each data point as a "token." Just as in natural language processing, where each word (or token) contributes to the understanding of a sentence, each data point in TimeGPT contributes to the understanding of the time series pattern. TimeGPT leverages these temporal patterns, which it has learned during training on vast datasets, to make accurate forecasts.

The TimeGPT API provides a user-friendly interface to this model, allowing you to harness its forecasting capabilities. Through this API, you can forecast future events and engage in various time series-related tasks. This includes exploring what-if scenarios, detecting anomalies in your data, and other applications where understanding future trends based on historical data is crucial.

To use TimeGPT for forecasting, you typically would:

- Prepare Your Data: Format your time series data and any relevant exogenous variables.
- Access T- API: Utilize the TimeGPT API to input your data into the model.
- Model Training/Loading: Depending on your setup, either train the model with your data or use the pre-trained model.
- Forecasting: Use the model to predict future data points based on the historical data you provided.
- Interpret Results: Analyze the forecasts generated by TimeGPT to make informed decisions or gain insights into future trends.

TimeGPT's advanced approach to time series analysis makes it a powerful tool for scenarios where understanding and predicting temporal data patterns are essential

Conclusion

The trajectory of the CPI from January 2020 to July 2023 suggests a significant relationship between it and the HPI, with the latter seemingly offering valuable insights for forecasting CPI values. One reason for this one-sided predictive power could be the nature and structure of the housing market in relation to the broader economy. Housing prices, encapsulated within the HPI, are a reflection of a myriad of macroeconomic variables such as interest rates, employment rates, and broader economic sentiment. A rise in the HPI could suggest increased consumer confidence, higher employment rates, and possibly even easy lending conditions set by the central bank, all of which have direct implications for the CPI. This makes the HPI a comprehensive and leading indicator in forecasting the CPI.

On the other hand, the CPI, being a broader measure, includes a vast range of goods and services and may not necessarily provide specific insights into the housing market. Simply put, while the price of a loaf of bread might increase due to inflation, it doesn't directly correlate to a surge in property values.

Similarly, the choice of using energy prices as exogenous variables to predict the CPI is grounded in logic. Canada, as a significant exporter of energy resources, sees its domestic economy intricately linked to global energy market fluctuations. Rising global energy prices could mean increased national revenues from energy exports, but it also implies higher domestic fuel and heating costs, directly impacting the CPI. The rationale behind not using CPI to forecast energy prices is clear: while the cost of various goods and services might change, they don't necessarily have a direct bearing on global energy demand and supply dynamics.

In essence, while the CPI captures the average price changes in goods and services, factors such as the HPI and energy prices play a foundational role in its movement. Their prominence in the Canadian economic structure makes them more explanatory in nature for CPI fluctuations, rather than the other way around

References and Acknowledgements

The case study conducted in this paper for HPI, CPI and Energy Prices was conducted solely based on our own experiments and were not copied from any external links nor replicated from other researchers. However, information and domain knowledge has been referenced below.

1. https://www.machinelearningplus.com/time-series/granger-causality-test-in-python/
2. https://www.analyticsvidhya.com/blog/2021/08/granger-causality-in-time-series-explained-using-chicken-and-egg-problem/
3. https://www.statsmodels.org/stable/generated/statsmodels.tsa.stattools.grangercausalitytests.html
4. https://www150.statcan.gc.ca/n1/en/type/data?MM=1
5. https://www.kaggle.com/code/anki112279/heatmap-by-canada-s-provinces/input
6. https://www.udemy.com/course/time-series-analysis-in-python/