The relationship between Google search Trends and Inflation

Logan Kyle Lomonaco

The University of texas at Dallas

EPPS6323 Knowledge Mining

Dr Karl Ho

May 12th, 2025

Abstract: This paper aims to see what the relationship between Google search Trends and Inflation is. This paper uses various AI models to construct a corpus of search terms related to everyday essential goods that is used to then gather data from Google trends. Combined with established economic indicators From FRED, this paper looks at google trends data, CPI, and The Michigan Consumer sentiment Index to investigate the relationship between them. This study finds that while traditional economic indicators nearly perfectly predict CPI, integrating Google search trends and lagged measures for key goods provides early signals of inflationary pressures. In contrast, consumer sentiment is shaped by a broader set of factors, indicating that digital trends serve as a valuable complement rather than a substitute for conventional economic metrics.

Introduction

In economics the study of overall economic health is paramount to making policy decisions, which has led to a significant amount of resources being dedicated to measuring key economic indicators such as GDP, Inflation, employment, and consumer confidence. However, it may be possible to utilize knowledge mining in order to evaluate such measures in real time and improve our understanding of these economic indicators better than existing methods. There are a number of possible sources of data that have the potential to become economic tools in predicting economic health, one of which is web search data. Google Trends provides a way to access such a source, and allows for one to evaluate the relationship between economic indicators and search trends. This paper seeks to use this to answer the question; What is the relationship between Google search Trends and Inflation? This paper aims to contribute to the literature by expanding the understanding of the relationship between Google searches and Inflation, examining the viability of utilizing knowledge mining techniques to better understand such relationships, and determining if utilizing these tools can supplement or perhaps even surpass traditional methods of studying economic health.

Overview placeholder

Theoretical framework

Background

The typical measures of inflation are often time consuming to collect and require significant resources to maintain, but provide an essential economic indicator for measuring the health of the economy. The main measure used for measuring inflation in the United States is the Consumer Price Index (CPI). Information for the CPI is normally collected by the Bureau of Labor Statistics (BLS) by sending data collectors who visit or call a sample of retailstors, rental units, and service establishments in order to record prices of a basket of goods (Bureau of Labor Statistics 2025). The CPI measures the changes in prices of all goods and services purchased for consumption by urban households, Including user fees like water and sewer service as well as sales and excise taxes paid by the consumer (Bureau of Labor Statistics 2025). However, income taxes and investment items such as stocks, bonds, and life insurance are excluded from CPI (Bureau of Labor Statistics 2025). The process of collecting this information can be time consuming, resource intensive, and only covers a sample representing the basket of goods, but has been one of the most robust measures of inflation available and has become the standard as a result.

Another important indicator of economic health is the Surveys of Consumers from the University of Michigan. This survey measures consumer confidence in the United States economy each month by surveying households randomly selected from the contiguous United States (48 states plus the District of Columbia) to participate(University of Michigan 2025). The questions asked in the Survey cover three broad areas of consumer confidence: personal finances, business conditions, and future buying plans (University of Michigan 2025). This survey can provide insight into how consumers feel about their current financial situation, their expectations for the future,

and their plans for spending and saving. This can be a valuable tool as an indicator for economic activity, as changes in consumer sentiment can foreshadow shifts in consumer spending and investment.

While very different, both of these economic indicators are important when understanding inflation. While the CPI is a direct measure of the changes in prices for a select basket of goods and services, The Michigan Consumer Sentiment Index reflects consumers' expectations and perceptions about the economy and inflation. These contrasting measures are both important when trying to predict and measure economic health and inflation as a whole.

Theory

The idea is that both the CPI and The Michigan Consumer Sentiment Index can be used as important measures related to inflation, but what about search trends? While search trends are not formal measures for inflation, it may be possible to use them to predict economic trends in a similar manner if one were to consolidate the insights from the mass amounts of search data available. Thus, in order to test this, this paper seeks to try and use search data to predict both CPI and The Michigan Consumer Sentiment Index as a way to better understand the relationship between them and Google searches and to see if It might be viable to use search terms as measures of economic health.

Data

The data utilized in this paper has been organized into a dataset containing monthly data ranging from 2004-01-01 to 2025-20-01 from the United States in order to capture the most data while still accommodating the availability of data as Google

trends data was limited to a range from 2004-01-01 to the present. This data contains information regarding a range of google trends data related to important essential goods for everyday people, essential economic indicators to serve as control variables that help prevent the over estimation of the impact of search terms, and CPI and The Michigan Consumer Sentiment Index as well. The following sections will go into detail about how each of these is defined in the data, the source of the data, and defines the key variables.

Sources

The data collection stars with the process of selecting which search terms to use to build the corpus of search trends data. To this end, it was decided to utilize a number of AI models and Knowledge Mining techniques to create a list of search terms based on relevance to important everyday essentials in the average person's lives. This resulted in creating a list of goods that were both relevant to the research, and were compatible with Google trends as there are a number of restrictions that have to be accounted for. Such restrictions include the limit on search terms in a single query being limited to five, resulting in each good on the list being defined by only five relevant search terms. Another limitation what that directly importing data from Google trends into Rstudio utilizing Libraries such as *gtrendsR* and *gtrendR* proved to be difficult to the point that manual download of the data was the better option, resulting in the gathering of data being time and resource intensive and limiting how much data could be reasonably gathered. However, despite the severe limitations of gathering Google trends data, a notable corpus of goods and their related search terms was gathered for this analysis.

The process of generating the corpus involved giving the same prompt to four different Ai models: Grok3, Copilot, Propensity AI, and ChatGPT. Each of the Models then generated output that built an individual corpus of essential everyday goods and relevant search terms. The output of each model was then fed back into Grok3 in order to consolidate the information and create the final corpus(s). This combination of AI models output was used to generate a primary corpus of essential everyday goods as well as a secondary corpus of more expensive or luxurious versions of goods. This distinction was made for two reasons, first being that normally inflation measures consider the basic version of a good or service rather than premium versions.The second reason for the two lists was that separating the search terms allowed for a more robust coverage of of the goods as in Google trends the number of terms are limited to five per query. While it was later determined that adding too many variables to the equations would negatively impact the analysis, thus necessitating a focus on the first set of terms rather than utilizing both, This separation still helped better define the corpus depicted in Table 1 below. Table 2 depicts the secondary corpus of More expensive versions of goods.

Table 1: Goods and Search terms for standard goods

Good	Search Terms
Milk	milk + whole milk + milk price + milk cost + milk prices
Bread	bread + white bread + whole wheat bread + bread price + bread cost
Beef	beef + ground beef + hamburger meat + beef price + ground beef price
Chicken	chicken + chicken breast + chicken price + chicken cost + price of chicken
Eggs	eggs + large eggs + eggs price + eggs cost + price of eggs
Coffee	coffee + ground coffee + coffee price + coffee cost + price of coffee

Table 1: Goods and Search terms for standard goods

Rice	rice + white rice + brown rice + rice price + price of rice
Apples	apples + apple price + price of apples + cost of apples + apples cost
Gasoline	gasoline + regular gasoline + gas price + gas cost + price of gas
Diesel Fuel	diesel + diesel fuel + diesel price + diesel cost + price of diesel
Electricity	electricity + electric bill + electricity price + electricity cost + power bill
Natural Gas	natural gas + gas bill + natural gas price + natural gas cost + heating gas price
Heating Oil	heating oil + fuel oil + heating oil price + heating oil cost + price of heating oil

This table shows the good and the combination of search terms related to said standard goods that were utilized to gather data from google trends. The search terms were used to find search trends data from the United States on the monthly level under the 2004 to present category.

Table 2: Goods and Search terms for goods

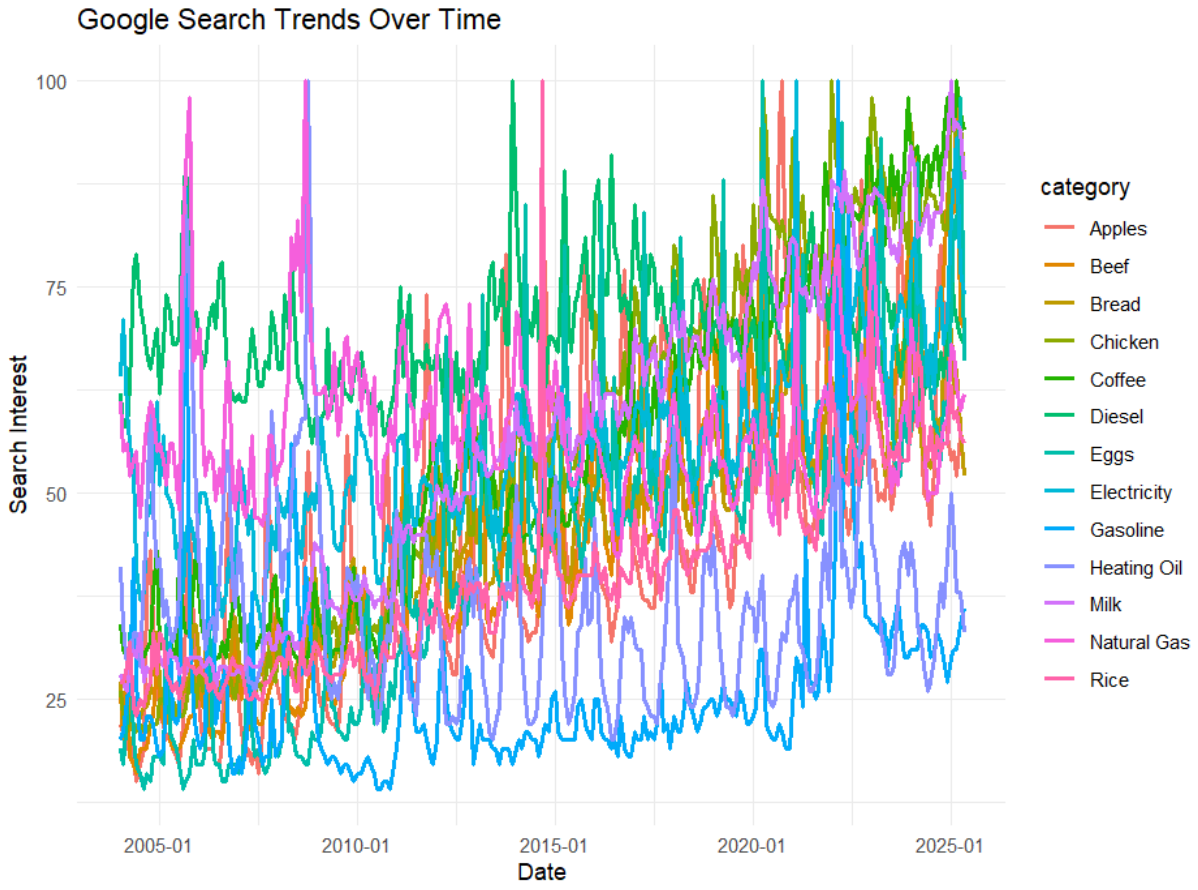
Good	Search Terms
Organic Milk	organic milk + organic whole milk + organic milk price + organic milk cost + price of organic milk
Artisan Bread	artisan bread + sourdough bread + artisan bread price + sourdough price + cost of artisan bread
Steak	steak + ribeye steak + steak price + ribeye price + cost of steak
Organic Chicken	organic chicken + free-range chicken + organic chicken price + free-range chicken cost + price of organic chicken
Organic Eggs	organic eggs + free-range eggs + organic eggs price + free-range eggs cost + price of organic eggs
Specialty Coffee	specialty coffee + Starbucks coffee + specialty coffee price + Starbucks price + cost of specialty coffee
Basmati Rice	basmati rice + organic rice + basmati rice price + organic rice cost + price of basmati rice
Organic Apples	organic apples + honeycrisp apples + organic apple price + honeycrisp price + cost of organic apples
Premium Gasoline	premium gasoline + high octane gas + premium gas price + premium gas cost + cost of premium gas
Renewable Electricity	renewable electricity + green energy + renewable energy price + green energy cost + cost of renewable energy

This table shows the good and the combination of search terms related to said *expensive versions of* goods that were utilized to gather data from google trends. The search terms were used to find search trends data from the United States on the monthly level under the 2004 to present category.

While the inclusion of the expensive versions of goods was never realized due to concerns of overfitting the model (degrees of freedom would be an issue), and time constraints as separate regressions would have doubled the number of models in this analysis (from eight to sixteen), This would be an excellent opportunity for future research that will be touched on in the conclusions section of this paper.

The data for the goods was collected, compiled into a dataframe on Rstudio, and then organized by date and value as assigned by Google with values representing search interest. This general overview of the data can be seen below in Figure 1.

Figure 1: visual overview of the Google trends data



A value of 100 is the peak popularity for the term, 50 means that the term is half as popular, 0 means there was not enough data

This data shows that some search trends are more relevant than others, with some of the standouts being Gasoline and Heating oil being the least searched and terms such as Coffee, Chicken, Milk, and Electricity being the most searched.

The other portion of the data is comprised of data sets collected from The St. Louis Fed data base: Federal Reserve Economic Data (FRED), where The following Data sets were downloaded from: Consumer Price Index for All Urban Consumers: All Items in U.S. City Average (CPIAUCSL), University of Michigan: Consumer Sentiment (UMCSENT), Unemployment Rate (UNRATE), Real Disposable Personal Income (DSPIC96), Producer Price Index by Commodity: All Commodities (PPIACO), Federal Funds Effective Rate (FEDFUNDS), Advance Retail Sales: Retail Trade (RSXFS), S&P CoreLogic Case-Shiller U.S. National Home Price Index (CSUSHPINSA), and Total Consumer Credit Owned and Securitized (TOTALSL) (also note that while Average Hourly Earnings of All Employees, Total Private (CES0500000003) was originally added, it was later dropped due to it missing data from 2004 to 2006 which limited its use in the analysis unless two years of data was dropped to accommodate).

Dependent and Explanatory Variables

Utilizing the downloaded data, The dependent variables are Consumer Price Index for All Urban Consumers: All Items in U.S. City Average (CPIAUCSL) or CPI, and University of Michigan: Consumer Sentiment (UMCSENT) or Consumer sentiment. These variables are being evaluated to see the relationship between them and THe Google trends data, with the other data sets from Fred being used as control variables to avoid over predicting the importance of the Google search terms. CPI and Consumer

sentiment are being estimated separately with each one being the sole dependent variable in their respective models as the other serves as an additional control variable.

Methods.

The analysis utilizes a total of eight models, four for measuring CPI, and four measuring Consumer Sentiment. These models are the same for each dependent variable, just duplicated with CPi and Consumer sentiment being swapped in each case. There are four types for equations utilized in this paper for this analysis: Linear model, Linear model with lagged Google trends data, a LASSO model, and a Random Forest model. The following section looks into each of the equations and explains how the variables are defined.

Equations and Random Forest models.

$$\text{cpi} \sim \text{apples} + \text{beef} + \text{bread} + \text{chicken} + \text{coffee} + \text{diesel} + \text{eggs} + \text{electricity} + \text{gasoline} + \text{heating\_oil} + \text{milk} + \text{natural\_gas} + \text{rice} + \text{consumer\_sentiment} + \text{unemployment\_rate} + \text{real\_income} + \text{producer\_price} + \text{fed\_rate} + \text{retail\_sales} + \text{home\_price} + \text{consumer\_credit} \quad (1)$$

$$\text{cpi} \sim \text{apples\_lag} + \text{beef\_lag} + \text{bread\_lag} + \text{chicken\_lag} + \text{coffee\_lag} + \text{diesel\_lag} + \text{eggs\_lag} + \text{electricity\_lag} + \text{gasoline\_lag} + \text{heating\_oil\_lag} + \text{milk\_lag} + \text{natural\_gas\_lag} + \text{rice\_lag} + \text{consumer\_sentiment} + \text{unemployment\_rate} + \text{real\_income} + \text{producer\_price} + \text{fed\_rate} + \text{retail\_sales} + \text{home\_price} + \text{consumer\_credit} \quad (2)$$

$$\begin{aligned}\text{cpi} \sim\ & 98.45 + (0.0275 \cdot \text{apples}) + (0.0289 \cdot \text{beef}) + (-0.0004 \cdot \text{bread}) + \\ & (0.0345 \cdot \text{chicken}) + (0.0043 \cdot \text{diesel}) + (-0.0008 \cdot \text{eggs}) + (-0.0421 \cdot \text{electricity}) + \\ & (-0.0437 \cdot \text{gasoline}) + (-0.0197 \cdot \text{heating\_oil}) + (0.0263 \cdot \text{natural\_gas}) + \\ & (-0.0430 \cdot \text{rice}) + (-0.0166 \cdot \text{consumer\_sentiment}) + \\ & (0.6714 \cdot \text{unemployment\_rate}) + (0.2769 \cdot \text{producer\_price}) + \\ & (0.00002 \cdot \text{retail\_sales}) + (-0.0559 \cdot \text{home\_price}) + (0.00002 \cdot \text{consumer\_credit})\end{aligned} \quad (3)$$

$$\text{consumer\_sentiment} \sim \text{apples} + \text{beef} + \text{bread} + \text{chicken} + \text{coffee} + \text{diesel} + \text{eggs} + \text{electricity} + \text{gasoline} + \text{heating\_oil} + \text{milk} + \text{natural\_gas} + \text{rice} + \text{cpi} + \text{unemployment\_rate} + \text{real\_income} + \text{producer\_price} + \text{fed\_rate} + \text{retail\_sales} + \text{home\_price} + \text{consumer\_credit} \quad (4)$$

consumer_sentiment ~ apples_lag + beef_lag + bread_lag + chicken_lag + coffee_lag + diesel_lag + eggs_lag + electricity_lag + gasoline_lag + heating_oil_lag + milk_lag + natural_gas_lag + rice_lag + cpi + unemployment_rate + real_income + producer_price + fed_rate + retail_sales + home_price + consumer_credit (5)

consumer_sentiment ~ 172.84 + (−0.1416 · CPI) + (0.0354 · apples) + (0.1926 · coffee) + (0.0914 · diesel) + (0.0785 · eggs) + (0.0289 · electricity) + (−0.1712 · gasoline) + (−0.2532 · heating_oil) + (0.2556 · milk) + (0.1313 · natural_gas) + (0.0829 · rice) + (−3.6375 · unemployment_rate) + (−0.3441 · producer_price) + (1.0072 · fed_rate) + (0.00005 · retail_sales) + (−0.1318 · home_price) (6)

Equations 1 and 4 represent the linear models that use the the predictor variables representing the Google trends data related to the good the variables are named after (apples, beef, bread, chicken, coffee, diesel, eggs, electricity, gasoline, heating_oil, milk, natural_gas, rice) with the search terms that are defined in Table 1, and the economic indicators to be used as controls pertaining to the data from FRED with each variable defined as: cpi represents (CPIAUCSL), consumer_sentiment represents (UMCSENT), unemployment_rate represents (UNRATE), real_income represents (DSPIC96), producer_price represents (PPIACO), fed_rate represents (FEDFUNDS), retail_sales represents (RSXFS), home_price represents (CSUSHPINSA), consumer_credit represents (TOTALSL). The main difference between Equations 1 and 4 is that equation 1 is predicting cpi, while equation 4 is predicting consumer_sentiment. The goal of these equations is to gather a general understanding of the relationship between the CPI or Consumer sentiment and search trends data with the control variables in place.

Equations 2 and 5 are similar to equations, but differ in that they used lagged versions of the google trend data (lagged by 1 month) rather than the standard, with the

Trends data marked with “_lag” to indicate such. The goal is similar, with these equations seeking to gather a general understanding of the relationship between the CPI or Consumer sentiment and lagged search trends data with the control variables in place. This set of models was chosen to examine if search activity from a previous period has a relationship with present values.

Equations 3 and 6 focuses on using a LASSO model in order to add a penalty to shrink coefficients in order to aid in feature selection and reducing overfitting. Thus each of the predictor variables is given a weight that it is multiplied by that is detailed more in the results section of this paper. However, the equations are shown with the weights in order to better explain the basic principle of these equations. Similarly to other equations, equation 3 is predicting cpi, while equation 6 is predicting consumer_sentiment.

There are also two Random forest models that are based on the estimation variables present in equations 1 and 4 with each one looking into ranking the variables in order of importance when predicting cpi or consumer_sentiment respectively. This is done to help determine which variables are the most important by seeing how a model's accuracy is negatively impacted by the removal of said variables. These models are implemented to expand the understanding of the relationship between the estimator variables and cpi or consumer_sentiment. These random forest models utilize a 80/20 training and testing split as there was enough data to justify the allocation of more training data for more robust results.

Packages and Libraries.

The *ggplot2* library is used for creating advanced and customizable graphs and plots. This library was used for plotting Google Trends data and visualizing the relationship between actual and predicted values through line plots and scatter plots. The *lubridate* package helps simplify working with dates and times. This library helps convert date strings from CSV files into proper Date objects (for example, using the ym() function), which is essential for time-series analysis and ensuring correct ordering or calculations based on date. This was used to format the various data sources as they all were merged on date, making this library one of the keystones of this research. The *dplyr* library is a tool for data manipulation and transformation. It was used to merge data sets (with functions like bind_rows() and full_join()), filter rows, and pipe operations together for a streamlined workflow. It was very useful for handling and tidying the Google Trends and FRED data in a more intuitive manner than what likely would have been the alternative. The *randomForest* package helps streamline the implementation of a Random Forest algorithm, which is an ensemble method that creates multiple decision trees and averages their results to improve prediction accuracy while handling non-linear relationships. In this case it was used to produce a model for predicting CPI and another for Consumer sentiment in order to determine the importance of different variables using feature importance scores. Finally, the *glmnet* library helps fit linear models with regularization options such as LASSO and Ridge. In the case of this paper, it was used to perform LASSO regressions on the CPI and Consumer sentiment variables.

It is also worth mentioning the gtrendsR and gtrendR libraries, despite the fact that they were not used in the research due to issues of importing the google trends

data. Both libraries serve a similar purpose of interfacing with Google Trends So that one could query the data rather than of manually downloading CSV files, with the plan being to use one or both to obtain real-time or historical search interest data for the keywords (such as "apples," "beef," "bread," etc.). However it would appear that Google, like many other platforms, is taking greater steps to prevent web scraping or other forms of unapproved access, resulting in a situation where it was easier to download the data manually. However, the idea of using google trends data was due to the existence of these packages, so it is worth noting them as a possible library for similar research in the future when there are tweeks or updates that make the packages viable.

Results

The following sections detail the results of the analysis, and the key insights derived from the models specified in the methods section.

CPI linear regression

The results for the first linear regression show an Extremely high R-squared value of approximately 99.51%, indicating that the predictors in the model in theory explain nearly all of the variation in CPI, however looks like a case of overfitting or multicollinearity among predictors when considering the Actual vs Predicted CPI graph in Figure 2. It is also observed that goods such as beef and milk have significant positive associations with CPI,while bread, chicken, and gasoline show significant negative relationships; this would suggest that consumer web search behavior for standard goods may reflect substitution or market adjustment effects. The unemployment rate, real income, producer price, fed rate, retail sales, home price, and

consumer credit emerge as the most statistically significant. Particularly, the unemployment rate shows a robust positive effect on CPI. It is also notable that consumer_sentiment is not statistically significant, implying that it does not add much in the way of predictive power regarding CPI. The main findings are recorded in table 3.

Table 3: residuals and CPI linear regression results

```
Residuals:
       Min     1Q  Median     3Q     Max
-8.2459 -1.3291 -0.2229  1.2498  9.1518

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         8.145e+01  7.416e+00  10.984  < 2e-16 ***
apples              3.965e-02  1.751e-02   2.265 0.024441 *
beef                1.012e-01  2.967e-02   3.411 0.000762 ***
bread               -2.474e-01  6.071e-02  -4.075 6.33e-05 ***
chicken             -3.170e-01  4.635e-02  -6.838 7.04e-11 ***
coffee              1.916e-02  4.248e-02   0.451 0.652405
diesel              -1.123e-02  3.055e-02  -0.367 0.713637
eggs                3.996e-03  1.985e-02   0.201 0.840638
electricity         5.864e-02  3.067e-02   1.912 0.057122 .
gasoline            -7.456e-02  2.364e-02  -3.154 0.001822 **
heating_oil         1.641e-02  2.696e-02   0.609 0.543299
milk                3.897e-01  6.698e-02   5.818 1.96e-08 ***
natural_gas         1.888e-02  3.162e-02   0.597 0.551024
rice                2.794e-02  3.466e-02   0.806 0.421062
consumer_sentiment -1.645e-02  2.936e-02  -0.560 0.575865
unemployment_rate   2.246e+00  2.055e-01  10.929  < 2e-16 ***
real_income         -1.896e-03  5.303e-04  -3.576 0.000424 ***
producer_price      2.340e-01  2.815e-02   8.312 7.93e-15 ***
fed_rate            1.857e+00  1.764e-01  10.527  < 2e-16 ***
retail_sales        8.557e-05  2.286e-05   3.744 0.000229 ***
home_price          -8.353e-02  2.221e-02  -3.761 0.000215 ***
consumer_credit     3.010e-05  1.783e-06  16.880  < 2e-16 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 2.494 on 232 degrees of freedom
Multiple R-squared:  0.9951,    Adjusted R-squared:  0.9947
F-statistic:  2246 on 21 and 232 DF,  p-value: < 2.2e-16
```

CPI linear regression with lags

This paper finds a similar fit of Multiple R-squared ~99.54% with a slightly lower residual error compared to the linear mode without lags, suggesting that lagged (by one month) Google trends variables do not lose significant explanatory power. This implies that this model is marginally better at minimizing prediction errors compared to the linear mode without lags. Most digital predictors remain significant when lagged, though some coefficients such as natural gas and electricity adjust in magnitude and become

significant. Thus, it would seem that last month's Google search behavior may be able to serve as an early indicator of current inflationary pressures. This indicates that further study of lags might be worthwhile in future research as there might be an effect multiple months after that fact that could have a profound impact. The main findings are recorded in table 4.

Table 4: residuals and CPI linear regression with lags results

```
Residuals:
       Min     1Q  Median     3Q     Max
-6.7994 -1.4830 -0.1423  1.2377  9.4419

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          8.451e+01  6.727e+00  12.563  < 2e-16 ***
apples_lag           4.206e-02  1.659e-02   2.535 0.011901 *
beef_lag             5.988e-02  2.741e-02   2.184 0.029939 *
bread_lag            -2.915e-01  5.604e-02  -5.203 4.33e-07 ***
chicken_lag          -3.206e-01  4.425e-02  -7.245 6.40e-12 ***
coffee_lag           6.953e-02  4.129e-02   1.684 0.093552 .
diesel_lag           9.409e-04  2.888e-02   0.033 0.974039
eggs_lag             -1.246e-02  1.946e-02  -0.640 0.522632
electricity_lag      6.005e-02  2.924e-02   2.054 0.041089 *
gasoline_lag         -8.775e-02  2.252e-02  -3.897 0.000128 ***
heating_oil_lag      1.766e-03  2.631e-02   0.067 0.946538
milk_lag             3.574e-01  6.424e-02   5.563 7.31e-08 ***
natural_gas_lag      6.344e-02  3.007e-02   2.110 0.035958 *
rice_lag             4.512e-02  3.338e-02   1.352 0.177795
consumer_sentiment -3.011e-02  2.830e-02  -1.064 0.288475
unemployment_rate   2.426e+00  2.030e-01  11.950  < 2e-16 ***
real_income          -2.406e-03  4.863e-04  -4.948 1.45e-06 ***
producer_price       1.884e-01  2.710e-02   6.950 3.69e-11 ***
fed_rate             1.949e+00  1.695e-01  11.500  < 2e-16 ***
retail_sales         1.176e-04  2.117e-05   5.556 7.59e-08 ***
home_price           -1.177e-01  2.186e-02  -5.385 1.78e-07 ***
consumer_credit      3.224e-05  1.726e-06  18.683  < 2e-16 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 2.409 on 231 degrees of freedom
Multiple R-squared:  0.9954,    Adjusted R-squared:  0.995
F-statistic:  2385 on 21 and 231 DF,  p-value: < 2.2e-16
```

CPI LASSO model

The LASSO model works by shrinking some coefficients, in some cases to zero, in order to prevent overfitting and aids in variable selection. We see that apples, beef, chicken, diesel, eggs, electricity, gasoline, heating_oil, natural_gas, rice have survived the process with non-zero coefficients along with many of the key control variables and the unemployment rate being especially notable with its high score. This leads to the

understanding that only some of the Google trends are significant after accounting for multicollinearity and overfitting. Full results can be found in table 5 below.

Table 5: CPI LASSO model results

```
22 x 1 sparse Matrix of class "dgCMatrix"
                                s0
(Intercept)           9.845228e+01
apples                2.751446e-02
beef                  2.886313e-02
bread                 -4.149391e-04
chicken               3.452268e-02
coffee                .
diesel                4.309143e-03
eggs                  -8.472224e-04
electricity           -4.210860e-02
gasoline              -4.372158e-02
heating_oil           -1.974552e-02
milk                  .
natural_gas           2.629624e-02
rice                  -4.304295e-02
consumer_sentiment -1.660535e-02
unemployment_rate   6.714443e-01
real_income           .
producer_price        2.769529e-01
fed_rate              .
retail_sales          2.251072e-05
home_price            -5.591857e-02
consumer_credit       2.468206e-05
```

CPI Random Forest Model

The random forest model has two important measures: %IncMSE which is how much the error increases when a predictor is removed, and IncNodePurity which is the reduction in variance (or impurity) attributable to each predictor. With this in mind, the results show that Traditional economic indicators such as unemployment rate, real income, producer price, fed rate, retail sales, home price, and consumer credit are among the highest in importance along with Google Trends variables, items like coffee, chicken, and milk. This shows that some non-linear relationships between some search trend variables and CPI are especially significant and that they deserve further study in regard to inflation dynamics and consumer behaviour.

Table 6: CPI Random Forest model importance scores

	%IncMSE	IncNodePurity
apples	3.957902	21.97344
beef	1.564951	174.51227
bread	2.265629	455.96429
chicken	7.418393	5525.60582
coffee	8.905573	9262.74884
diesel	3.821961	20.08769
eggs	2.127939	620.31582
electricity	2.115320	20.19945
gasoline	3.994951	15.76291
heating_oil	3.345135	18.20560
milk	7.338589	6136.34822
natural_gas	2.529521	25.44759
rice	3.330227	1169.28435
consumer_sentiment	7.542949	720.85541
unemployment_rate	10.072189	776.94726
real_income	13.089366	14242.27784
producer_price	13.429106	1951.99630
fed_rate	10.927997	513.06562
retail_sales	13.279699	15420.99874
home_price	16.415460	3216.71397
consumer_credit	16.874662	21460.96068

Consumer sentiment linear regression

This model sees a lower R-squared value of ~82.93%, which is significantly less than the models predicting CPI. This may be explained as consumer sentiment being a more subjective measure of economic health that is influenced by factors outside of the scope of the chosen predictors. While most Google trends variables are not statistically significant, eggs (positive) and heating oil (negative) are notable exceptions. It is also notable that Unemployment and producer price have a negative relationship while fed rate and retail sales have a positive one. It is also notable that CPI is not a significant Predictor of Consumer sentiment. Results can be found in table 7.

Table 7: residuals and Consumer sentiment linear regression results

```
Residuals:
       Min     1Q   Median    3Q     Max
-17.3724  -3.3400   0.8021   3.6121  13.6582

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          1.777e+02  1.677e+01  10.597  < 2e-16 ***
cpi            -8.214e-02  1.466e-01  -0.560   0.5759
apples               5.743e-02  3.938e-02   1.458   0.1461
beef           -3.793e-02  6.790e-02  -0.559   0.5770
bread          -1.258e-02  1.404e-01  -0.090   0.9287
chicken              5.932e-03  1.135e-01   0.052   0.9584
coffee               1.795e-01  9.424e-02   1.904   0.0581 .
diesel               6.900e-02  6.815e-02   1.013   0.3123
eggs                 9.427e-02  4.393e-02   2.146   0.0329 *
```

```
electricity           3.306e-03  6.908e-02   0.048   0.9619
gasoline       -9.660e-02  5.357e-02  -1.803   0.0727 .
heating_oil    -2.490e-01  5.805e-02  -4.290 2.62e-05 ***
milk                  2.431e-01  1.594e-01   1.525   0.1286
natural_gas           7.928e-02  7.052e-02   1.124   0.2621
rice                  6.381e-02  7.746e-02   0.824   0.4109
unemployment_rate -3.740e+00  5.091e-01  -7.347 3.44e-12 ***
real_income           1.171e-03  1.215e-03   0.964   0.3361
producer_price -4.520e-01  6.524e-02  -6.929 4.15e-11 ***
fed_rate              1.033e+00  4.744e-01   2.177   0.0305 *
retail_sales          1.174e-04  5.203e-05   2.257   0.0250 *
home_price     -2.197e-01  4.905e-02  -4.479 1.18e-05 ***
consumer_credit   -4.818e-06  5.940e-06  -0.811   0.4182

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 5.574 on 232 degrees of freedom
Multiple R-squared:  0.8293,    Adjusted R-squared:  0.8138
F-statistic: 53.66 on 21 and 232 DF,  p-value: < 2.2e-16
```

Consumer sentiment linear regression with lags

The results show similar performance R-squared ~82.71% compared to the non-lagged regression, and found significant lagged effects for apples_lag (positive), heating_oil_lag (negative), and milk_lag (positive). Macroeconomic controls remain consistent, reiterating that economic conditions such as unemployment and producer prices critically shaping sentiment. Yet again, CPI is not significant when predicting consumer sentiment. This analysis shows that searches for goods such as apples and milk can capture early shifts in consumer mood. The whole results can be found in table 8 below.

Table 8: residuals and Consumer sentiment linear regression results

```
Residuals:
       Min     1Q   Median    3Q     Max
-16.5746  -3.3619   0.5281   3.4672  15.4934

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          1.832e+02  1.626e+01  11.269  < 2e-16 ***
cpi           -1.619e-01  1.522e-01  -1.064  0.28847
apples_lag            7.818e-02  3.866e-02   2.022  0.04431 *
beef_lag              7.975e-02  6.401e-02   1.246  0.21408
bread_lag      -3.097e-02  1.373e-01  -0.226  0.82176
chicken_lag    -1.233e-01  1.134e-01  -1.087  0.27805
coffee_lag            9.277e-02  9.616e-02   0.965  0.33565
diesel_lag            7.488e-02  6.680e-02   1.121  0.26348
eggs_lag              5.274e-02  4.503e-02   1.171  0.24274
electricity_lag       6.490e-03  6.842e-02   0.095  0.92451
gasoline_lag   -7.640e-02  5.368e-02  -1.423  0.15604
heating_oil_lag   -2.533e-01  5.870e-02  -4.316 2.36e-05 ***
milk_lag              3.989e-01  1.565e-01   2.550  0.01143 *
natural_gas_lag       2.591e-02  7.038e-02   0.368  0.71316
rice_lag              4.742e-02  7.765e-02   0.611  0.54202
```

```
unemployment_rate -3.457e+00  5.539e-01  -6.241 2.05e-09 ***
real_income           1.054e-03  1.184e-03   0.890  0.37429
producer_price -4.172e-01  6.342e-02  -6.578 3.15e-10 ***
fed_rate              1.305e+00  4.852e-01   2.689  0.00769 **
retail_sales          1.124e-04  5.176e-05   2.172  0.03087 *
home_price     -2.262e-01  5.168e-02  -4.377 1.82e-05 ***
consumer_credit   -1.264e-06  6.342e-06  -0.199  0.84217

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 5.586 on 231 degrees of freedom
Multiple R-squared:  0.8271,   Adjusted R-squared:  0.8114
F-statistic: 52.62 on 21 and 231 DF,  p-value: < 2.2e-16
```

Consumer Sentiment LASSO model

The LASSO model saw beef, bread, chicken, real_income, and consumer_credit effectively dropped from the model, but apples, coffee, diesel, eggs, electricity, gasoline (negative), heating_oil (negative), milk, natural_gas, and rice remain, with diesel, eggs, and rice exhibiting significant effects that surpass very important economic indicators. Additionally, CPI exhibits a small negative association with consumer sentiment. The whole results are reflected in table 9 below.

Table 9: Consumer Sentiment LASSO model results

```
22 x 1 sparse Matrix of class "dgCMatrix"
                                s0
(Intercept)           1.728413e+02
cpi            -1.416344e-01
apples                3.542740e-02
beef                  .
bread                 .
chicken               .
coffee                1.926121e-01
diesel                9.143859e-02
eggs                  7.853803e-02
electricity           2.897851e-02
gasoline       -1.712273e-01
heating_oil    -2.532458e-01
milk                  2.556421e-01
natural_gas           1.312569e-01
rice                  8.292345e-02
unemployment_rate -3.637512e+00
real_income           .
producer_price -3.440713e-01
fed_rate              1.007246e+00
retail_sales          5.412120e-05
home_price     -1.318038e-01
consumer_credit       .
```

Consumer sentiment Random Forest Model

The Results of the random forest model have interesting implications as not only are economic indicators the most significant predictors, CPI is the second most potent just

under unemployment. It is also notable that Milk is the most significant of the Google trends data being somewhat comparable to producer price, suggesting that searches may signal positive or reassuring trends in consumer sentiment. While market fundamentals are still the predominant factor, this shows that search data can be important to understand consumer sentiment. Complete results are in Table 10 below.

Table 10: Consumer sentiment Random Forest Model importance scores

	%IncMSE	IncNodePurity
cpi	16.836605	2982.5905
apples	6.345709	240.5698
beef	5.868105	203.3543
bread	4.966973	304.3743
chicken	8.332912	783.7860
coffee	6.456002	441.5543
diesel	5.245344	328.6686
eggs	5.857014	383.3080
electricity	3.776974	214.9263
gasoline	7.727740	597.5830
heating_oil	5.028160	408.2169
milk	10.446502	1069.8609
natural_gas	5.433679	479.3535
rice	4.461715	242.8860
unemployment_rate	20.302104	6304.4848
real_income	13.673201	2891.3980
producer_price	10.973173	982.2575
fed_rate	11.279843	1530.2848
retail_sales	12.493734	2291.4270
home_price	14.080963	2633.4453
consumer_credit	15.142770	3091.4372

Actual vs Predicted: evaluating model fit to the data

With a significant number of models in this analysis with differing outcomes, determining which ones fit the data better can be valuable as it allows focus on the model that is more accurate to the data. To this end, each model with the exception of the Random Forest models was subject to an actual vs predicted analysis to see how well each model fits the data. All actual vs predicted value tests utilize a 80/20 training/testing split as there was enough data to support the more robust training set. The results are shown in figure 2 below, with CPI models on the left and Consumer sentiment models on the right. Overall. It would appear that the LASSO model for CPI had the best fit out of the three. However, this is a far from perfect fit as all models evaluating CPI in this

test didn't cover the higher value data points in the top right of the graphs. In contrast, the Linear regression did the best for fitting the data in the case of Consumer Sentiment. This appears to be because the other codels tried to correct in a way that made them have an inferior fit for the data by comparison as they moved away from the central cluster of data points.

Figure 2

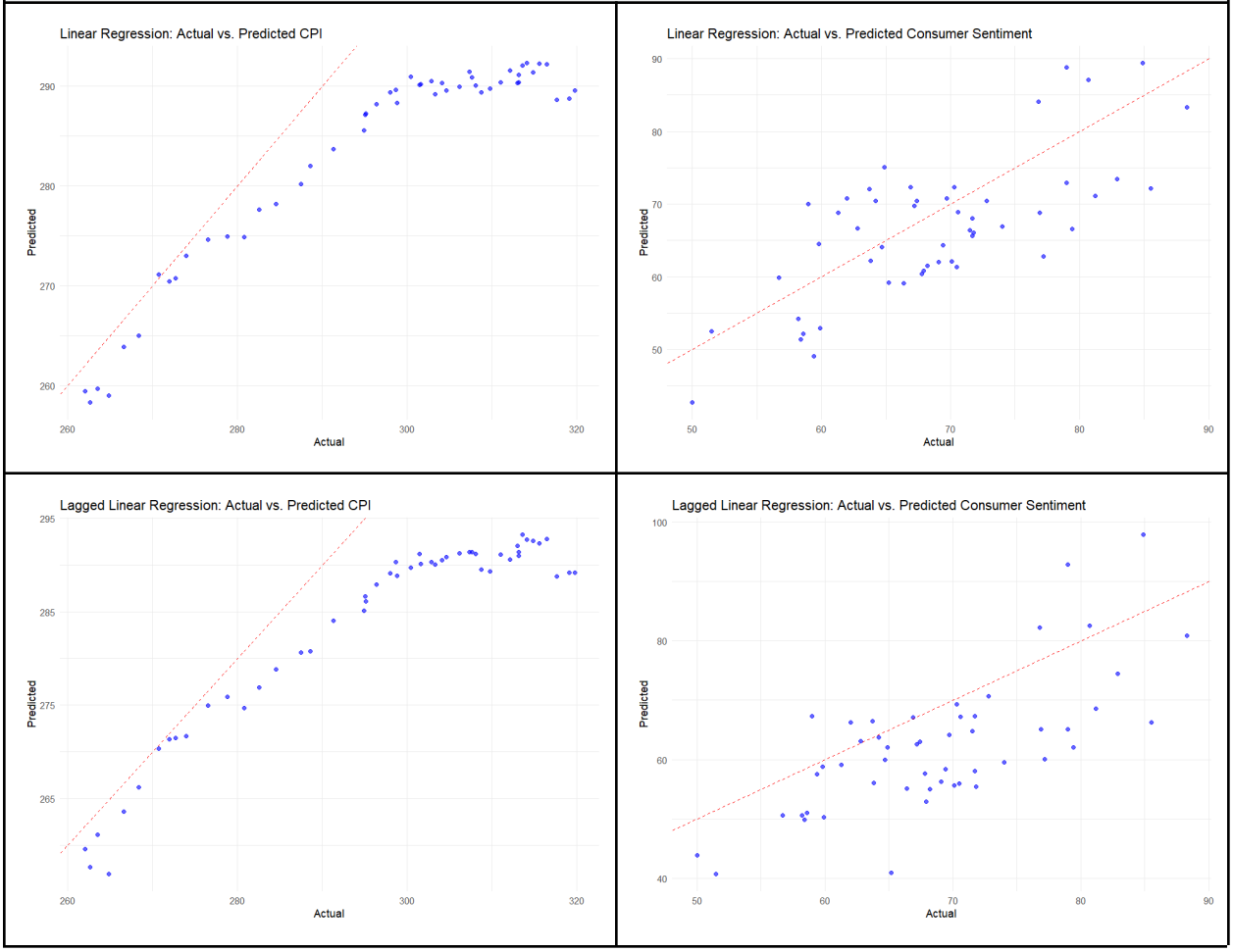
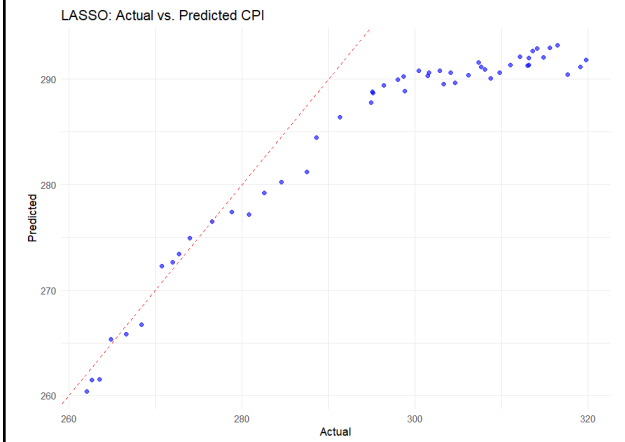
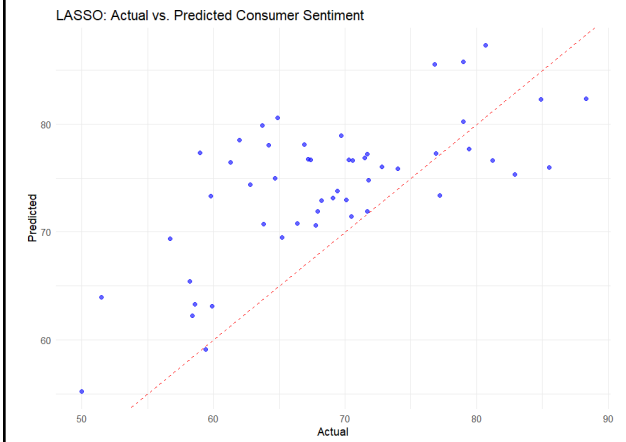


Figure 2





Conclusions

Overall, this paper has found that both traditional economic indicators and Google search trends can play a significant role in understanding fluctuations in bothe CPI and consumer sentiment, albeit in differing ways.

It was found that in the case of CPI models, the standard linear regression with an incredibly high R-squared of ≈99.51% demonstrates that, in theory, almost all variability in CPI can be explained by the included predictors. However, the exceptional fit raises concerns about overfitting and multicollinearity, as suggested by the Actual vs. Predicted plots. In this model, key goods such as beef and milk exert significant positive effects, while bread, chicken, and gasoline have significant negative effects. This points toward potential market substitution or adjustment behaviors. The macroeconomic controls such as unemployment rate, real income, producer price, fed rate, retail sales, home price, and consumer credit are notably influential. When the model is changed to incorporate one-month lagged Google trends variables, a similarly high fit of ≈99.54% R^2 and a slightly lower residual error are found. The lagged specifications not only preserved explanatory power but also hints at the potential to use prior month search

data to serve as early indicators of inflationary pressure. Additionally, the LASSO and Random Forest models complement these findings by refining the predictor set and capturing nonlinear effects, respectively. This reinforces the findings that while Google trends matter, traditional fundamentals remain a strong method of examining economic health.

In contrast, the consumer sentiment models have R-squared values of ≈82.9% and ≈82.71% in the case of the lagged version, which indicate that consumer sentiment is inherently more subjective and/or is influenced by a broader range of factors not fully captured by chosen the predictors. Although most Google search trends in these models are not statistically significant, some such as eggs and heating oil (consistently negative) and, in the lagged model, apples and milk emerge as notable predictors. Notably, macroeconomic variables still maintain their relevance with the exception of CPI as it was found to not be a significant driver of sentiment in these linear models. The LASSO model trims down the predictor set to the most relevant, and the Random Forest analysis ranking variables in terms of how much the error increases when they are is removed, and the attributable reduction in variance. It was found that economic indicators were the most important, with unemployment at the top and, surprisingly, CPI coming in as a strong second most important. It was also found that Google search trends related to milk can be as influential as core economic variables.

So, what is the relationship between Google Search Trends and Inflation? The findings indicate a robust relationship between Google search trends and inflation. Google search behavior variables in the models explored in the work such as beef, milk, and other standard goods have a statistically significant association with CPI. These

searches likely reflect underlying consumer behaviors such as substitution or market adjustments, which in turn impact price levels. Additionally, the predictive power of lagged search variables would suggest that previous month's search behavior can serve as a leading indicator of inflationary pressures. Overall, while traditional economic indicators are still essential for understanding inflation, integrating Google Trends data offers a valuable, real-time supplement that can enhance the accuracy of models and provide early warning signals for shifts in inflation.

In summary, although CPI is captured with near-perfect fidelity by a set of predictors, caution is warranted regarding overfitting, and the incorporation of lagged search behavior appears promising in forecasting inflation. Conversely, the lower explanatory power for consumer sentiment underscores its multifaceted nature, where traditional economic conditions dominate but subtle signals from digital behavior nonetheless contribute to early detection of shifts in public mood. Together, these findings advocate for a mixed approach—where both established economic metrics and innovative digital data are utilized—to unravel complex modern economic dynamics and enhance forecasting accuracy.

Future research

There are a number of different avenues to expand the work here in future work, as the topic not only allows for different data to be used, but also for different models to be tested. Thus, there are a great deal of options when considering how to expand into future research.

One such option would be to use social media data and collect sentiments related to posts referencing economic issues to gauge overall economic sentiment. This

Idea was originally considered in the proposal for what would become this paper. However, the difficulty of acquiring data made this idea nonviable as many prominent social media platforms have moved to severely restrict API access which makes web scraping social media posts difficult if not impossible without permission being granted by the companies hosting the platforms. Despite this, if policy changes or new platforms become prominent that have more open APIs, then this type of analysis may become viable.

Another option would be to look into analyzing whether there are different results when using premium versions of goods as outlined in table 2. While the current analysis was limited by the degrees of freedom and doing a second set of models would require eight more models, this would be a viable addition as it would only require changing the corpus for different search terms. Additionally, researching sentiments related to news stories both through options in google trends or through web scraping would also be a notable addition to the overall topic. Often the news can influence sentiments through the stories they run, so examining trends in the stories and wording could be a viable research path in itself. Furthermore, changing the models and data sources would also allow for further research into the topic. Changing Google trends data for another search engine's data (perhaps Bing), or comparing multiple, could be an option if that data could be accessed. There are also other viable economic indicators and variations of said indicators that would also be worth experimenting with to see if the models can be refined with their implementation. There are also a huge range of possible search terms relevant to other aspects of inflation and beyond that could be tested.Thus, there are a great number of options as to how one could expand the research.

References:

AI disclosure: the following Ai models were used to generate the corpus of search terms in this paper: Grok3, Copilot, Propensity AI, and ChatGPT. Grok3 and Copilot were also used to refine and bug fix the Rstudio code used in the analysis that this paper covers.

Bureau of Labor Statistics. n.d. “Overview of the Consumer Price Index.” Bureau of Labor Statistics. Retrieved May 12, 2025, from https://www.bls.gov/cpi/overview.htm.

Federal Reserve Bank of St. Louis. n.d. “Advance Retail Sales: Retail Trade (RSXFS).” FRED. Retrieved May 12, 2025, from https://fred.stlouisfed.org/series/RSXFS.

Federal Reserve Bank of St. Louis. n.d. “Consumer Price Index for All Urban Consumers: All Items in U.S. City Average (CPIAUCSL).” FRED. Retrieved May 12, 2025, from https://fred.stlouisfed.org/series/CPIAUCSL.

Federal Reserve Bank of St. Louis. n.d. “Federal Funds Effective Rate (FEDFUNDS).” FRED. Retrieved May 12, 2025, from https://fred.stlouisfed.org/series/FEDFUNDS.

Federal Reserve Bank of St. Louis. n.d. “Producer Price Index by Commodity: All Commodities (PPIACO).” FRED. Retrieved May 12, 2025, from https://fred.stlouisfed.org/series/PPIACO.

Federal Reserve Bank of St. Louis. n.d. “Real Disposable Personal Income (DSPIC96).” FRED. Retrieved May 12, 2025, from https://fred.stlouisfed.org/series/DSPIC96.

Federal Reserve Bank of St. Louis. n.d. “S&P CoreLogic Case-Shiller U.S. National Home Price Index (CSUSHPINSA).” FRED. Retrieved May 12, 2025, from https://fred.stlouisfed.org/series/CSUSHPINSA.

Federal Reserve Bank of St. Louis. n.d. “Total Consumer Credit Owned and Securitized (TOTALSL).” FRED. Retrieved May 12, 2025, from https://fred.stlouisfed.org/series/TOTALSL.

Federal Reserve Bank of St. Louis. n.d. “Unemployment Rate (UNRATE).” FRED. Retrieved May 12, 2025, from https://fred.stlouisfed.org/series/UNRATE.

Google Trends. n.d. “Google Trends.” Google. Retrieved May 12, 2025, from https://trends.google.com/trends/.

University of Michigan, Institute for Social Research. n.d. “Consumer Sentiment (UMCSENT).” Retrieved May 12, 2025, from https://soc.isr.umich.edu/about-the-survey.html.