WHAT FACTORS PREDICT INFLATION?

With Linear Regression Modeling

TRISH GIRMUS

5/2/2022

TABLE OF CONTENTS

"Inflation is not only unnecessary for economic growth. As long as it exists it is the enemy of economic growth."

- *Henry Hazlitt*

BUSINESS PROBLEM

In February 2022, the Consumer Price Index (CPI) reached its highest percentage the United States has witnessed in 40 years. At 7.9% (279) from a year earlier, the concern of inflation with the rising costs of goods and services is affecting consumers, one way or another. The focus of this project was to predict the feature(s) which attributed to the latest CPI measure. Linear regression modeling was used to predict the features.

BACKGROUND/HISTORY

What is inflation? According to Investopedia, "inflation is a measure of rising prices of goods and services in an economy" (Investopedia Team, 2021). Examples of these goods and services include housing, food, electricity, gasoline, transportation, and industrial goods. Typically, this measure occurs over a year and looks at the average rise in prices. Over time, the purchasing power of a dollar decreases and therefore costs more to buy goods and services. When a CPI measure increases, so does the cost of everything else. A lower CPI measure means a lower cost in goods and services.

To show an example of the power of a dollar, in 1978 if the cost of an item was $5, in 2022 to purchase that same item it would cost $22.05. This is a 341.0% cumulative rate of inflation (https://www.usinflationcalculator.com/).

There are two ways to measure inflation: the Consumer Price Index (CPI) and Producer Price Index (PPI). For this project, the measure used to identify inflation was CPI. CPI is defined by the U.S. Bureau of Labor Statistics as "a measure of the average change over time in the prices paid by urban consumers for a market basket of consumer goods and services" (2022). PPI in contrast, reports a change in price which influences domestic producers (Investopedia Team, 2021).

Consumers receive household surveys from government agencies to determine the basket cost of frequently purchased items over a period of time. This is calculated by using a weighted average of the prices for those items (One Minute Economics, 2020). This basket price is then reported to the U.S. Bureau of Labor Statistics (BLS) each month and identifies the current CPI measure. There are eight fundamental categories according to the BLS which measure CPI.

These include:

- Food & Beverage
- Housing
- Apparel
- Education & Communication
- Medical Care
- Recreation
- Transportation
- Other Goods & Services
  (Source: https://advisor.visualcapitalist.com/inflation-over-last-100-years/)

There are a few factors which can attribute to increases in price/inflation: production costs, demand, and fiscal policy (Investopedia Team, 2021). However, over the past two years, current and world events are also contributing to inflation which include COVID-19/pandemic and the war in Ukraine. Consumers have witnessed higher wages, low interest rates, high housing costs, and high demand with low supply, as a few examples. The analysis and modeling conducted provides more insight into this topic.

DATA EXPLANATION

Due to the currency, timelines in history with dates, events, etc., laws, economic outlooks and other various factors, the United States was the only country selected for this project. Secondary data was used for this project from several sources, with the majority collected from the U.S. Bureau of Labor Statistics website. It contained 11 categories/variables. The time range selected was November 1978 to February 2022. This was due to missing data in earlier years of some categories. Additional data was collected, and four variables were added to the dataset. All data used for the project was both quantitative and qualitative.

In total, 15 variables were initially selected for analysis and modeling. Below is a list of the variable names and the definition of each.

Name [variable_name]:

- Year [year] – years include 1978 to 2022
- Month [month] – includes every month of the year, starting in November 1978 through February 2022
- CPI [cpi] – U.S. city average, all items, urban wage earners and clerical workers, not seasonally adjusted

- Gasoline [gas] – CPI U.S. city average price, not seasonally adjusted (unleaded regular, per gallon/3.785 liters)
- Electricity [electricity] – CPI U. S. city average price per KWH, not seasonally adjusted
- Food [food] – CPI U.S. city average, seasonally adjusted (all urban consumers)
- Shelter [shelter] – CPI U.S. city average, seasonally adjusted (all urban consumers)
- Transportation [transportation] – CPI U.S. city average, seasonally adjusted (all urban consumers)
- Medical Care [medical_care] – CPI U.S. city average, not seasonally adjusted (all urban consumers)
- Unemployment Rate [unemployment_rate] – seasonally adjusted, ages 16 years and over
- Employment Population Ratio [employment_pop_ratio] – seasonally adjusted, ages 16 years and over, all industries & occupations
- Presidential Political Party [pres_political_party] – shows political affiliation of U.S. president by month
- Consumer Confidence Index [cci] – amplitude adjusted, long-term average
- Crude Oil [crude_oil] – in U.S. dollars per barrel, located in WTI – Cushing, Oklahoma
- Cause [cause] – cause of oil prices by year

The variable [crude_oil] had missing values from November 1978 through December 1985. Zeros were filled in those missing cells in the .csv file when compiling the data. The remaining data did not require any preparation. All data selected was compiled from its respective resources and arranged into one .csv file.

METHODS

The dataset file was read and then saved into a Pandas DataFrame in Python, which was the programming language used for analysis and modeling. The data initially contained 15 variables and 520 rows of data. Most data were measured as an Index. To make all data the same measure, data not containing values as Indexes were standardized during the data cleaning process (see below).

Cleaning the data:

- A time variable was created for the inflation model. The variables [year] and [month] were not sufficient and therefore were converted from integers to objects. An index column with these two variables was then created as [ym]. The variables [year] and [month] were dropped and replaced with [ym]. By creating a composite variable using the year and month appended (i.e., 1978_11), it was a better match.
- The variable [cause] did not include specific data for events by month. A categorical event was added only to each January of the corresponding year and had missing values for February through December. Due to the complexity and research required to dummy encode and standardize this variable, it was dropped and not used for analysis or modeling.
- The units of measurement across the variables were not universal, and therefore required the data to be standardized. Relative index variables were created for data containing the variables [crude_oil], [cci], [unemployment_rate], [gas], and [employment_pop_ratio]. Those variables

were then dropped, and the new formatted variables were created [previous variablename_idx].

- The formatting for this data was calculated by dividing each value by the first data point (November 1978) and multiplying by 100. This is the methodology used for calculating the CPI (Indeed Editorial Team, 2022).

The next step of the process was Exploratory Data Analysis (EDA). Line charts were created using the matplotlib library in Python to visualize the relationships between the variables. The first line chart created contained the variables [cpi], [food], [shelter], [transportation], [medical_ care], and [crude_oil_idx]. All variables minus [crude_oil_idx] showed a steady increase over 43 years with [medical_care] increasing the most. (Figure 1)

Subsequent line charts included relationships between the following variables over the same time span:

- [gas_idx] and [unemployment_rate_idx] (Figure 2)
- [cci_idx] and [employment_pop_ratio_idx] (Figure 3)
- [gas_idx] and [crude_oil_idx] (Figure 4)

The relationship patterns with the two pairs of variables were closely aligned. Sharp increases or decreases signaled events in 2008, 2009, (the financial crisis leading into the great recession) and 2020 (COVID-19 pandemic).

A Seaborn pairplot was then added to visualize other relationships between the variables. A positive correlation showed with the variables [cpi] & [electricity], [cpi] & [food], [cpi] & [shelter], [cpi] & [transportation], and [cpi] & [medical_care]. Then, a correlation heatmap was used to determine how closely the relationships were between each pair (Figure 5). It showed a highly positive relationship between those pairs of data. All variables had an r value of 1, except for electricity (0.96).

After reviewing the correlation heatmap, a new subset dataframe was created using variables [electricity], [transportation], [medical_care], [unemployment_rate_idx], [cci_idx], [crude_oil_idx], and [pres_political_party]. The variables were defined as predictor variables and then the values for those were standardized (Zach-Statology, 2021). This was the second standardization for the subset. Additional pairplot, correlation, and heatmap visualizations were also created to review relationships with the variables.

Additional visualizations were created to review other relationships with variables. When reviewing a line chart of [medical_care] & [unemployment_rate_idx], it showed regardless of an increase or decrease of the unemployment rate index over time, medical care continues to increase (Figure 6). To visualize the categorical variable [pres_political_party], pairplots and boxplots were created (Pathak, 2020). The variables [food] and [cci_idx] used for the pairplot in Figure 7 showed that food values can be expected to be higher for a lower cci index. There is a wider range for Republicans. When Democrats are in office, the food index doesn't change much during the presidential term, but the range changes a lot when Republicans are in office. The boxplot in Figure 8 visualized the variables [employment_pop_ratio_idx] & [pres_political_party]. It showed that employment is greater when Democrats are in office. Outliers were present for the Republicans, which may be due to the COVID-19 pandemic.

ANALYSIS

Modeling was performed with the linear regression function from the sklearn library (Pathak, 2020). Several models were created using different sets of variables to predict. Predictor variables were determined from EDA. The first model used the independent variable [shelter] and the dependent variable [cpi]. A train and test dataset using the train_test_split package from sklearn was created to prevent overfitting from occurring. A test size of 0.40 was used. The coefficient parameter was 0.728. The prediction graph in Figure 9 showed good results for the prediction model. An ordinary least square (OLS) regression results summary was then created. The function comes from the statsmodel library. The results showed the coefficient of determination, or R-squared value of 0.992 which indicated the variables were perfectly correlated.

A second model was created using the independent variable [food] and dependent variable [cpi]. Test size remained 0.40. The coefficient parameter was 0.940. Figure 10 displays good results again for the prediction model. The OLS regression results showed an R-squared value of 0.995.

A third model was created using two independent variables: [medical_care] and [food]. The dependent variable remained [cpi]. The test size 0.40 was used. The coefficient parameter for [medical_care] was -2.656 and [food] was 0.986. The prediction model again showed good results (Figure 11). The OLS regression results show an R-squared value of 0.995. Because these variables were both highly correlated, multicollinearity was present in these results.

A fourth model was created using the independent variable [gas_idx] and dependent variable [cpi]. Test size again remained 0.40. The coefficient parameter was 0.347, which was the lowest thus far when creating models. The prediction model in Figure 12 does not indicate these results as good predictors. The R-Squared value was 0.679, the lowest of any models. It is not a good predictor of CPI.

A final model was made using 10 independent variables: [electricity], [food], [shelter], [transportation], [medical_care], [crude_oil_idx], [cci_idx], [umemployment_rate_idx], [gas_idx], and [employment_pop_ratio_idx]. The dependent variable was again [cpi]. The test size for this model was 0.30. The training and test scores confirmed the warning in the multivariate model of multicollinearity. The R-squared value was 1.00 which indicated that the variables are too highly correlated with one another and not good for predicting CPI.

CONCLUSION

As most of the variables were already factors to measure CPI, the correlation results during EDA indicated that these variables may not be good predictors for modeling. Linear regression modeling confirmed these results when trying to use the variables to predict CPI. The current predictors should be removed. Additional data would be required to predict what variables attributed to the current CPI measure.

ASSUMPTIONS

An assumption was made with the origination of selected data used for the project (based on the credibility of the source) that it would clearly solve the business problem. Unfortunately, it did not. More resources, information, and research are required. It was also assumed using Linear Regression modeling would be sufficient to solve the business problem.

LIMITATIONS

The biggest limitation for this project was time. Inflation itself is a topic that encompasses many facets. Many hours were spent on conducting research to understand and comprehend what elements were important to include for the project.

CHALLENGES

There were several challenges imposed with this project. Besides the limitation of time, understanding inflation without having a strong background and overall understanding of it was very challenging. For someone at a beginner level for learning vs. a Data Scientist, this topic was too complex for a four-week project. Another challenge was not having a solid knowledge of statistics. It was also difficult to interpret the results from the regression summary. The data selected was also a challenge for a few reasons. The units for each variable were not uniform and required standardizing, which required research and more time to determine how to achieve. There was also not enough data to predict CPI with, which could be seen after EDA, but unfortunately there was not enough time to go back and find additional data to rework. Trying to use categorical data was also difficult, especially with adding the [cause] variable. It was beyond the knowledge and experience level trying to standardize and quantify events. Categorical data is like a soft science. Psychology is a good example of this. There was also concern since there were not specific monthly events for the [cause] variable, the data might become biased.

FUTURE USES/ADDITIONAL APPLICATIONS

As the CPI measure continues to increase in 2022, a future use of this study could be to continue to evaluate the data and add more variables to see what other predictors impact CPI. Adding more variables would require finding additional data to work with. A time series study might also be another option for future use to see trends with inflation and CPI. An additional application might also be to look at expanding the project as global, which could give additional insights with more data as well.

RECOMMENDATIONS

It is recommended to find and include additional data for this project. The variables in the current project were determined to not be good predictors and therefore new variables need to be added to the study for further analysis. Another recommendation is to try a different model and determine if better results will occur. Polynomial regression might be an option. Unfortunately, time did not permit to explore other options with modeling.

IMPLEMENTATION PLAN

The first step in the implementation plan is to collaborate with others and ask for feedback on suggestions for finding additional variables that make sense to use in the study. While gathering this information, continue to research about inflation to get a better understanding of the topic. Maybe there is a different approach or identify a different business problem. Also, research other types of models to use that might be better suited than linear regression. Once more information is acquired, it will be added to the dataset and then data cleaning, EDA, and analysis/modeling will be reworked.

ETHICAL ASSESSMENT

There should not be a compromise with ethics for this project for a couple of reasons. One, the data used initially was collected from a secondary source. There is no bias with the data and it was not used for reasons that could impose any potential harm. Secondly, this information is being shared and told through a Data Analyst perspective vs. a professional/government official. The words used to communicate and describe this information should not cause any widespread panic as there is no intention to create action or alarm.

APA REFERENCES

How to calculate consumer price index. Indeed Career Guide. (n.d.). Retrieved from https://www.indeed.com/career-advice/career-development/how-to-calculate-cpi

One Minute Economics (July 19, 2020) *Consumer Price Inflation (Consumer Price Index/CPI) and Asset Price Inflation Compared in One Minute*. Retrieved from https://www.youtube.com/watch?v=TqXvpyNFW1U

Neufeld, D. (2021, June 11). *Visualizing the history of U.S. inflation over 100 years*. Advisor Channel. Retrieved from https://advisor.visualcapitalist.com/inflation-over-last-100-years/

Pathak, P. (2020, July 16). *Implementing linear regression using Sklearn*. Medium. Retrieved from https://medium.com/analytics-vidhya/implementing-linear-regression-using-sklearn-76264a3c073c

Staff, U. S. I. C. (2022, April 12). *Inflation calculator: Find US Dollar's value from 1913-2022*. US Inflation Calculator |. Retrieved from https://www.usinflationcalculator.com/

Team, T. I. (2022, February 24). *What causes inflation and who profits from it?* Investopedia. Retrieved from https://www.investopedia.com/ask/answers/111314/what-causes-inflation-and-does-anyone-gain-it.asp

U.S. Bureau of Labor Statistics. (2022, April 29). U.S. Bureau of Labor Statistics. Retrieved from https://www.bls.gov/

U.S. Bureau of Labor Statistics. (n.d.). *CPI Home*. U.S. Bureau of Labor Statistics. Retrieved from https://www.bls.gov/cpi/

DATA REFERENCES

CPI: https://beta.bls.gov/dataViewer/view/timeseries/CWUR0000SA0

Gasoline: https://beta.bls.gov/dataViewer/view/timeseries/APU000074714

Electricity: https://beta.bls.gov/dataViewer/view/timeseries/APU000072610

Food: https://beta.bls.gov/dataViewer/view/timeseries/CUSR0000SAF1

Shelter: https://beta.bls.gov/dataViewer/view/timeseries/CUSR0000SAH1

Transportation: https://beta.bls.gov/dataViewer/view/timeseries/CUSR0000SAS4

Medical Care: https://beta.bls.gov/dataViewer/view/timeseries/CUUR0000SAM

Unemployment Rate: https://data.bls.gov/pdq/SurveyOutputServlet

Employment Population Ratio: https://beta.bls.gov/dataViewer/view/timeseries/LNS12300000

President Political Party: https://en.wikipedia.org/wiki/List_of_presidents_of_the_United_States

Consumer Confidence Index: https://data.oecd.org/leadind/consumer-confidence-index-cci.htm

Crude Oil: https://www.eia.gov/dnav/pet/pet_pri_spt_s1_m.htm

Cause: https://www.thebalance.com/oil-price-history-3306200

APPENDIX

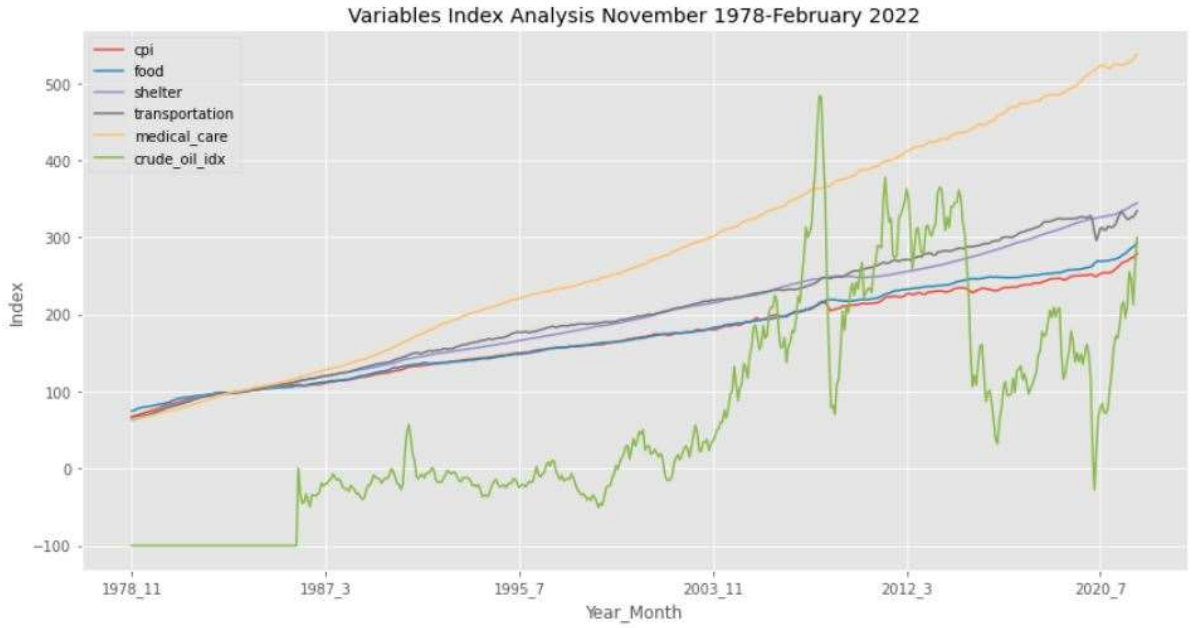

Figure 1- Line Chart of Variables Over 43 Years

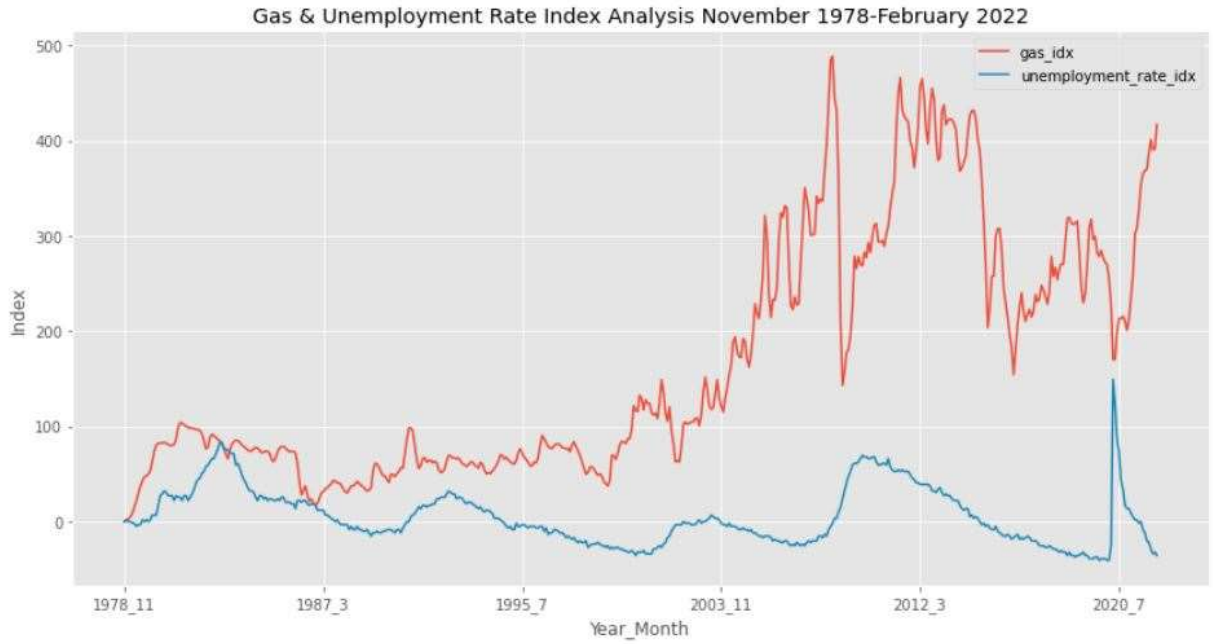

Figure 2- Line Chart of Gas & Unemployment Rate Variables Over 43 Years

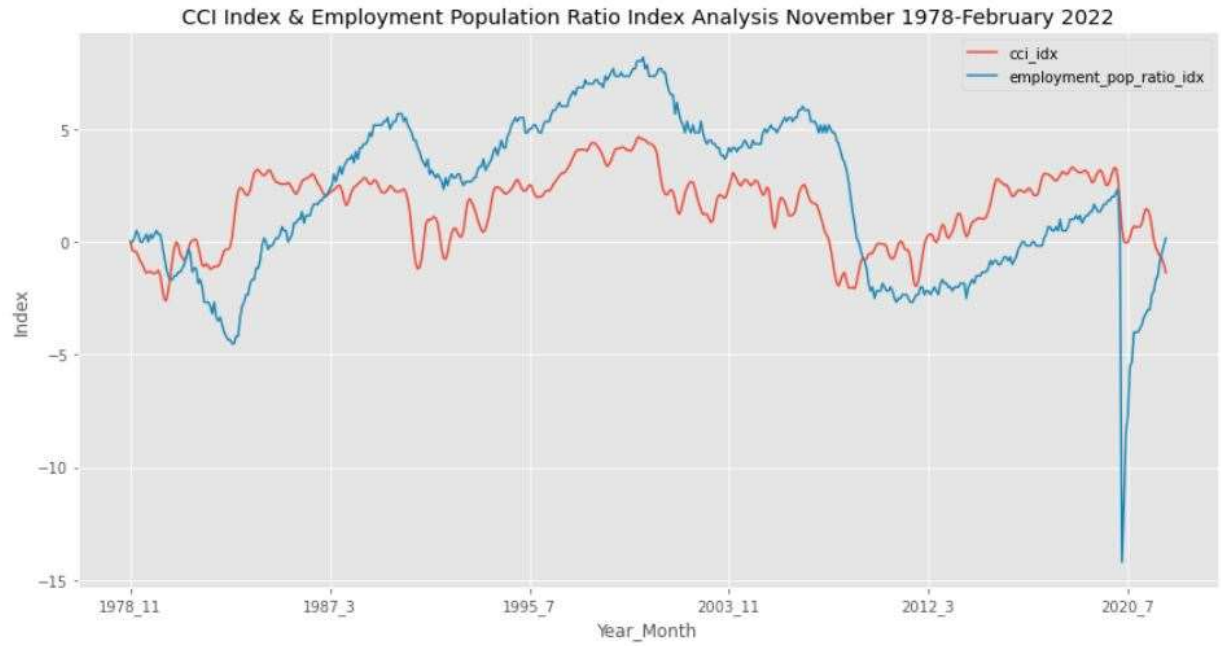

Figure 3- Line Chart of CCI Index & Employment Population Ratio Index Over 43 Years

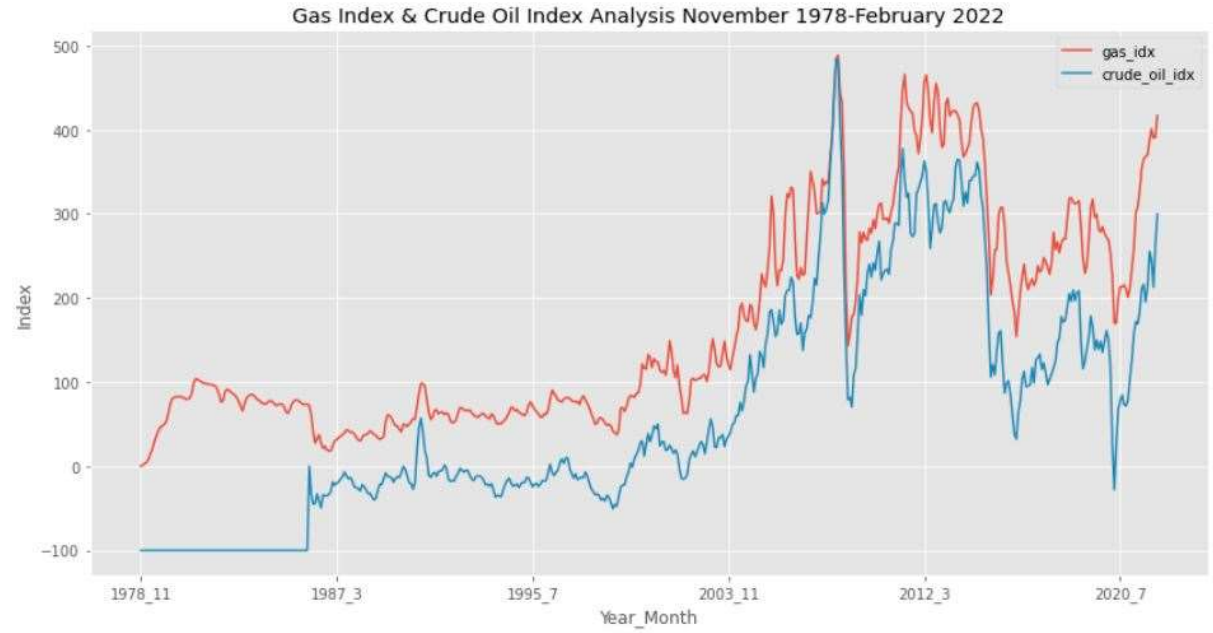

Figure 4 – Line Chart of Gas Index & Crude Oil Index Over 43 Years

Figure 5- Correlation Heatmap Using Seaborn

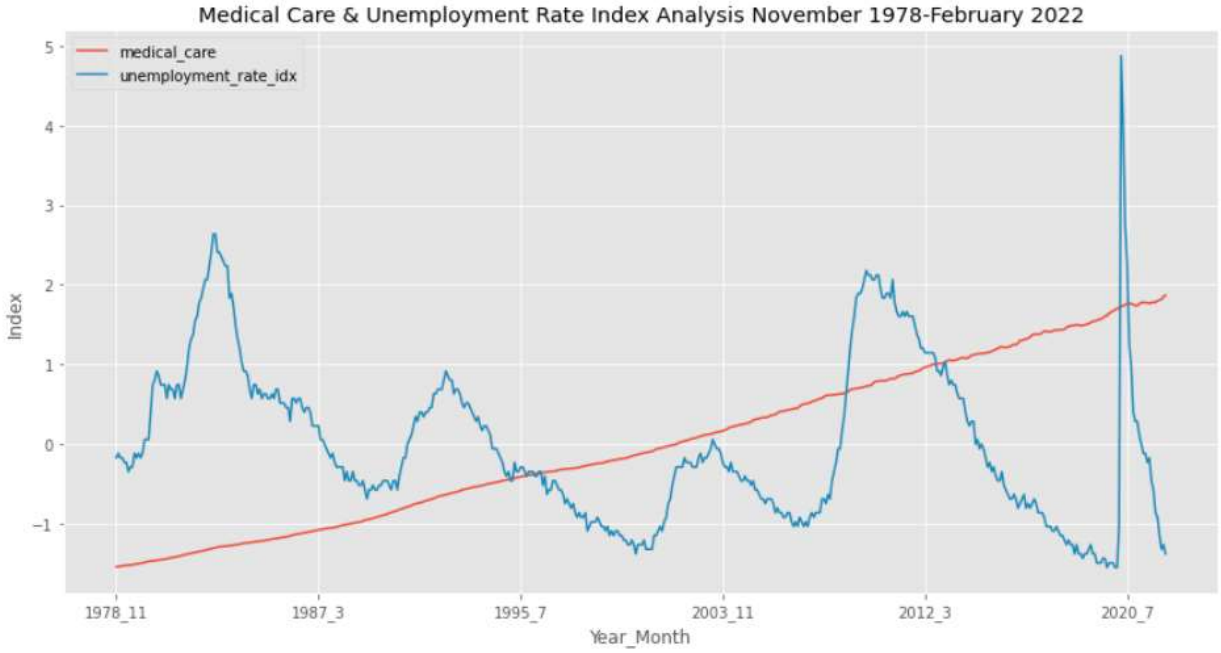

Figure 6 - Line Chart of Medical Care & Unemployment Rate Index Over 43 Years

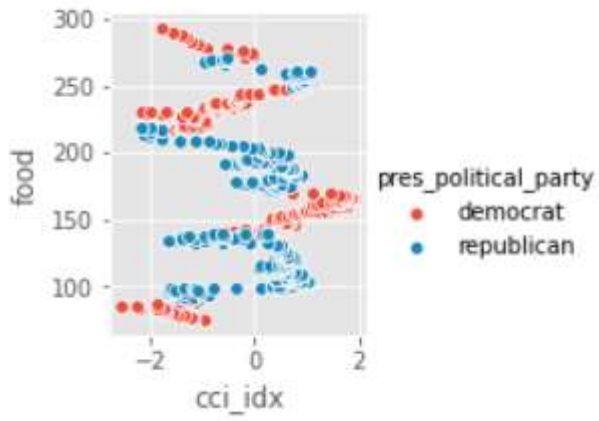

Figure 7- Pairplot of Food & CCI Index When Democrats or Republicans are in Office

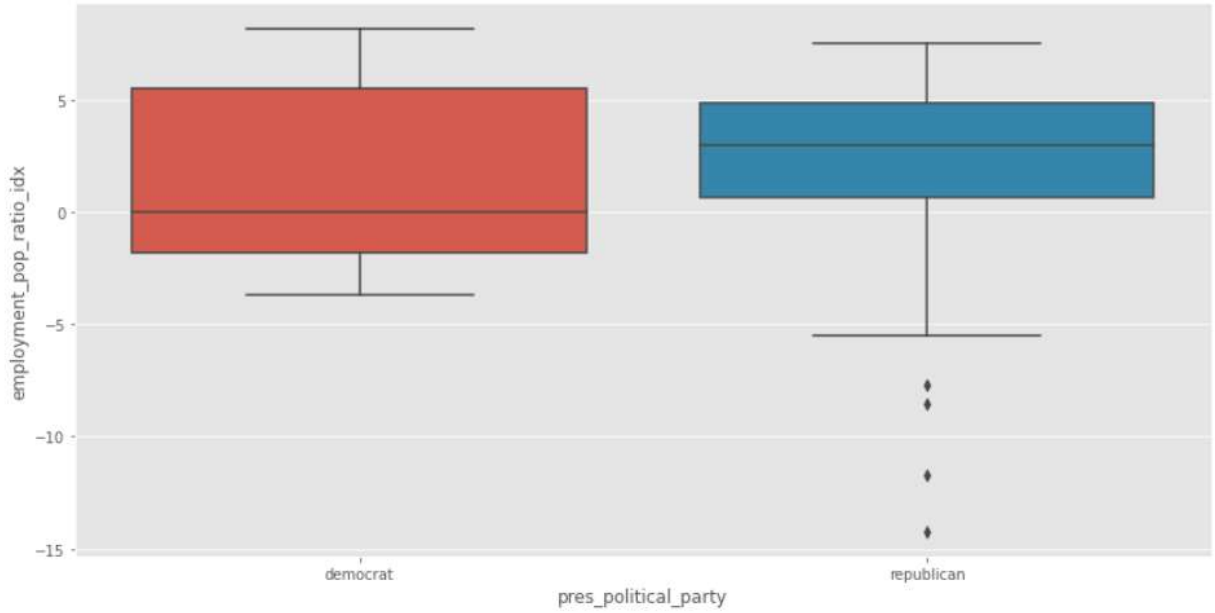

Figure 8 - Boxplot of Employment Population Ratio & Presidential Political Party

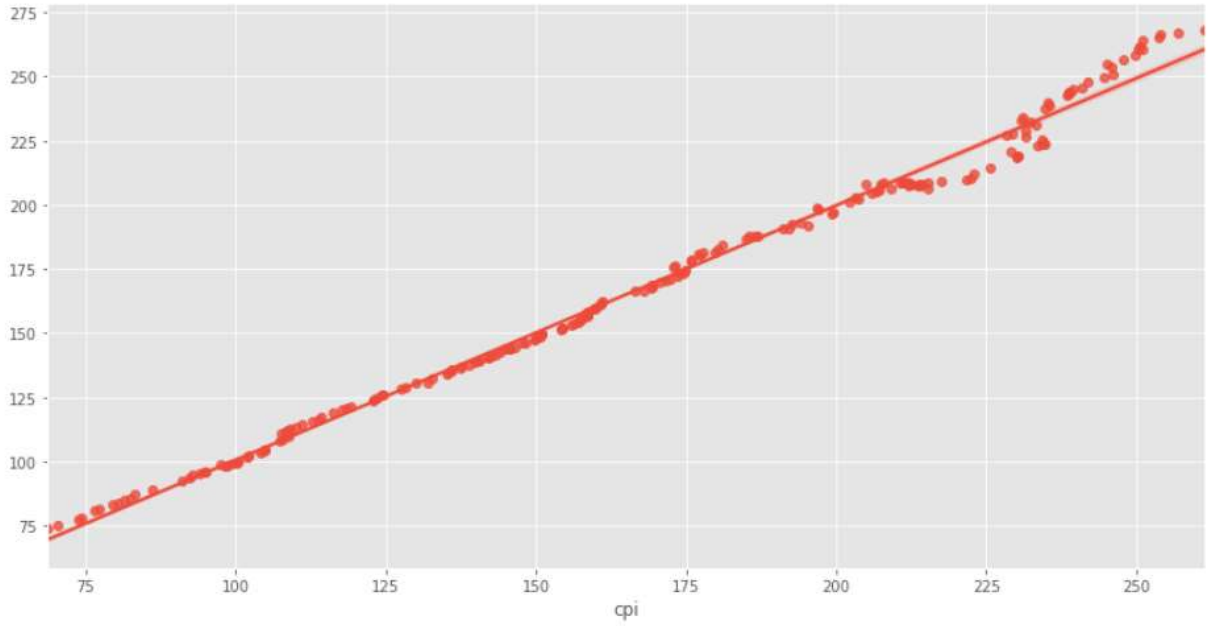

Figure 9- Prediction Model - showing good results of predictors. Independent variable is Shelter, dependent variable is CPI.

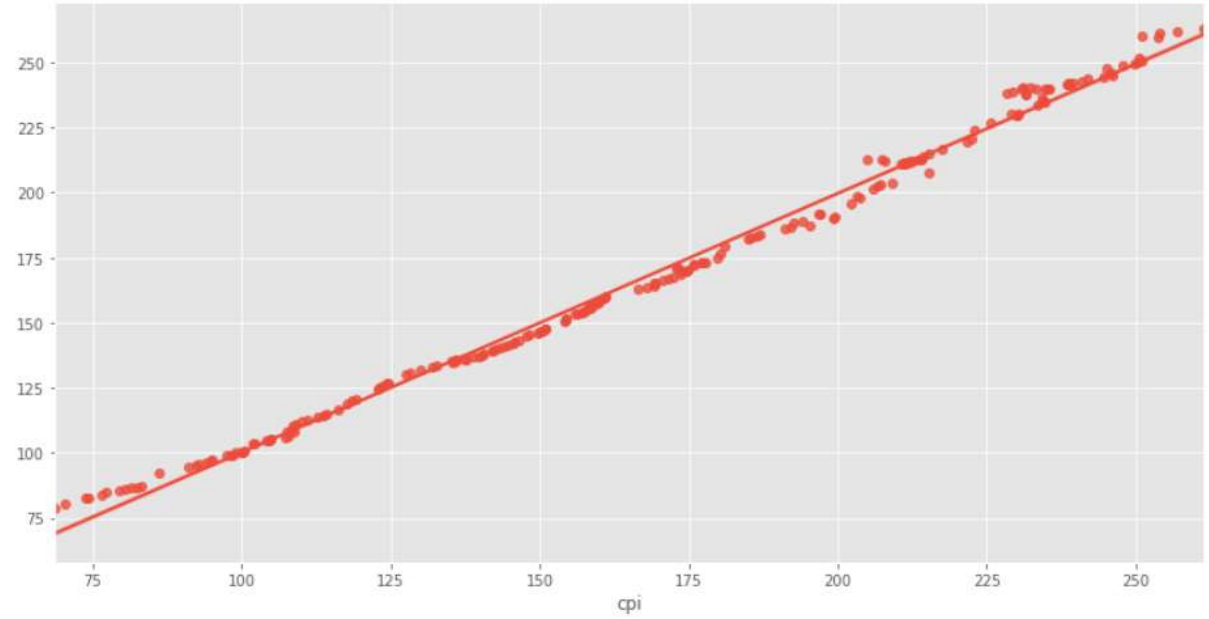

Figure 10- Prediction Model - showing good results. Independent variable is Food, dependent variable is CPI.

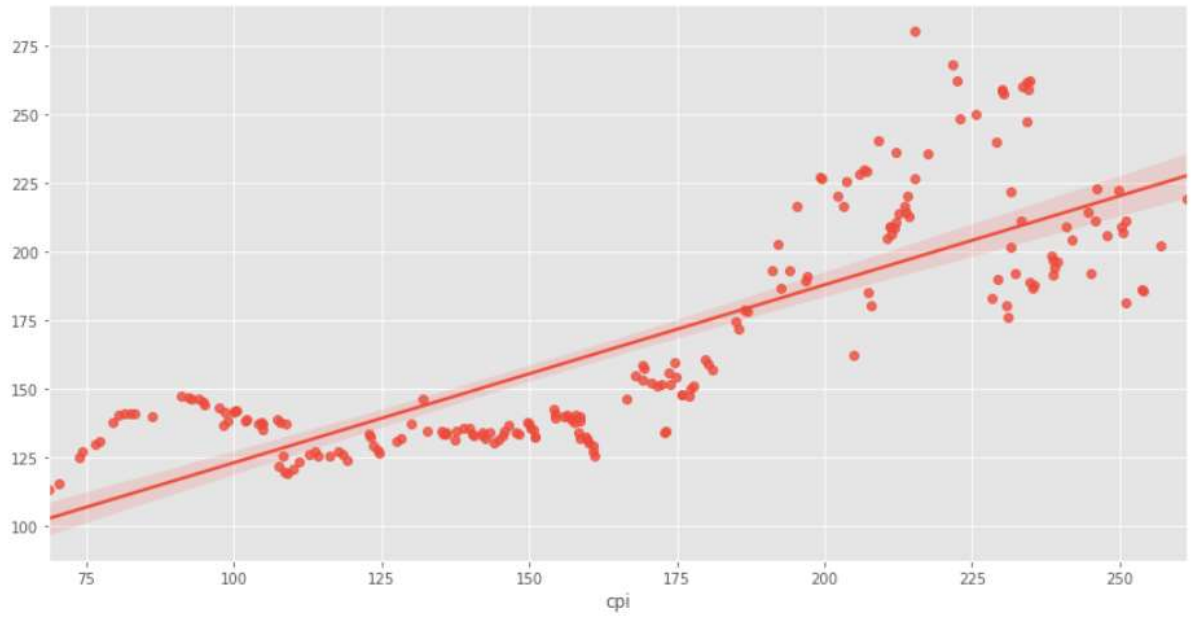

Figure 11- Prediction model with independent variable Gas Index and dependent variable CPI. The prediction model does not indicate good predictors.

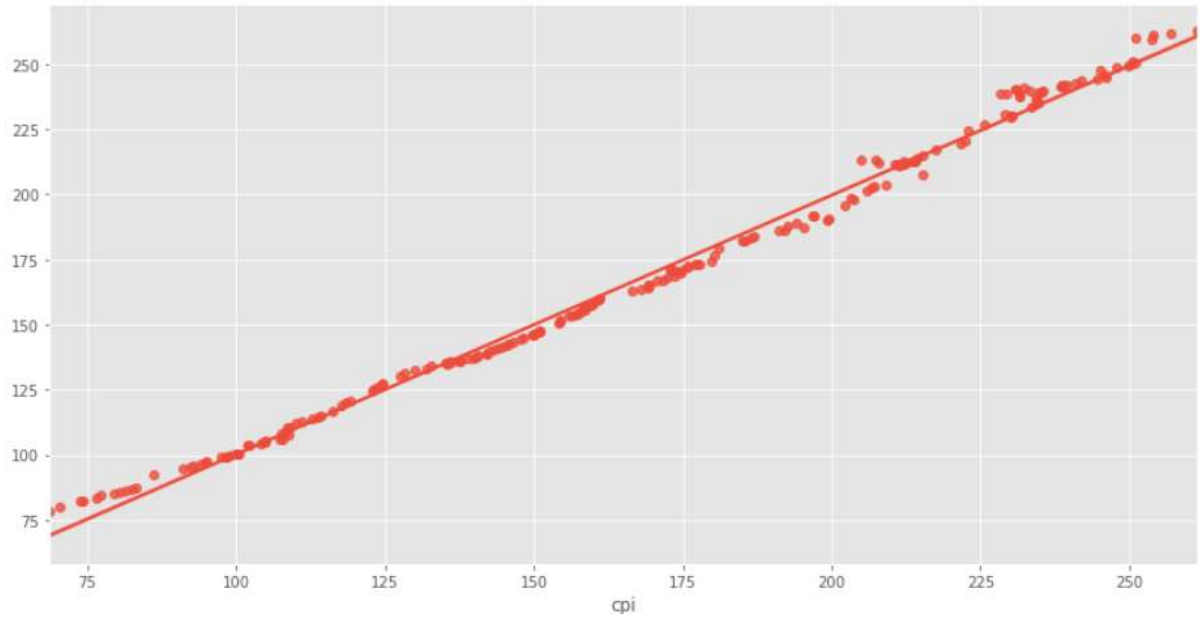

Figure 12 - Prediction model using independent variables Medical Care & Food. Dependent variable is CPI. Graph displays good predictors.

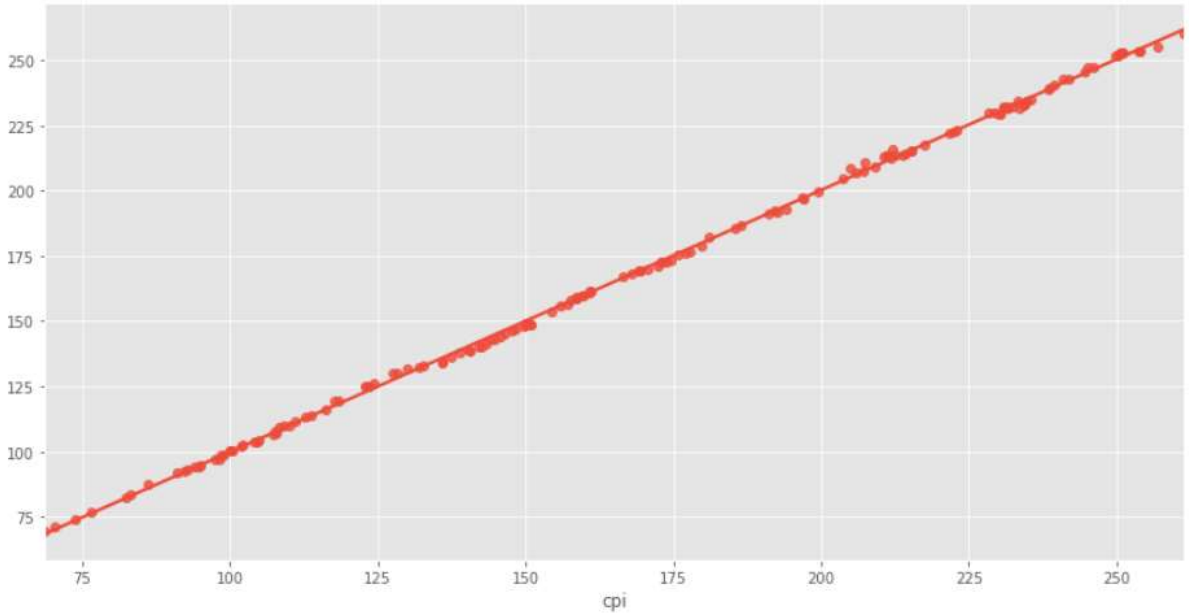

Figure 13- Prediction model using independent variables: Electricity, Food, Shelter, Transportation, Medical Care, Crude Oil Index, CCI Index, Unemployment Index, Gas Index, & Employment Population Ration Index. Dependent variable is CPI.

10 QUESTIONS AN AUDIENCE WOULD ASK YOU -INCLUDES ANSWERS

1) Does a high CPI measure represent a positive or negative signal in the context of inflation?

-A high CPI measure represents a negative signal in the context of inflation. When a CPI measure increases, so does the costs of everything else.

2) How does a market basket determine the Consumer Price Index (CPI)?

-Consumers receive household surveys from government agencies to determine the basket cost of frequently purchased items over a period of time. This is calculated by using a weighted average of the prices for those items (One Minute Economics, 2020). This basket price is then reported to the U.S. Bureau of Labor Statistics (BLS) each month and identifies the current CPI measure. There are eight fundamental categories according to the BLS which measure CPI: food & beverage, housing, apparel, education & communication, medical care, recreation, transportation, and other goods & services.

3) Why did you use November 1978-February 2022 for your data?

-The majority of data selected for this project was selected from the U.S. Bureau of Labor Statistics. It had missing data in earlier years on several categories. November 1978-February 2022 contained all data and therefore was selected.

4) Why are you focusing on the United States only for this project versus globally?

-Due to the currency, timelines in history with dates, events, etc., laws, economic outlooks and other various factors, the United States was the only country selected for this project.

5) Why did you select this particular data for your project?

-The data selected for this project was in alignment with the topic of predicting which factors attributed to the latest CPI measure. While the majority was chosen from the U.S. Bureau of Labor Statistics website, additional data was collected to determine if other factors could predict the CPI measure increase.

6) Which considerations made you select the predictor variables?

-When reviewing the variables during exploratory data analysis, it was determined which variables would be selected as predictor variables. A Seaborn pairplot was used to visualize those relationships as well as a correlation heatmap.

7) What measurement were the variables in?

-The data was measured as an Index. Most data before data cleaning were already measured as an Index. Therefore, it was necessary to standardize all data, so it was in the same measure.

8) How did you make all variables the same units?

-The formatting for this data was calculated by dividing each value by the first data point (November 1978) and multiplying by 100. This is the same methodology used for calculating the CPI (Indeed Editorial Team, 2022).

9) Why didn't you use the feature [cause] for any analysis/modeling?

-The feature [cause] did not contain data per month, it was only listed by year. It did not specify which month the event started, or when it ended. Assumptions could have been made incorrectly and did not want to create bias or inaccurate data. Due to the complexity and time required to determine how to use this feature accurately, it was dropped and not used for analysis or modeling.

10) Why did you choose linear regression for modeling?

-Linear Regression modeling was assumed the best choice for solving the business problem. The reason is it was selected was to be able to better understand and predict the relationships of independent and dependent features in a statistical sense. It was believed that the relationships of the features were linear, and that this type of model would best predict the business problem.