Kwasi Mensah

IST 718

Big Data Analytics

Final Project

Rising Cost of Food in the United States

**Objective:**

1) Obtain data and understand data structures and data elements.

2) Scrub data using scripting methods, to include debugging, for data manipulation in R and other tools.

3) Explore data using essential qualitative analysis techniques including descriptive statistics.

4) Model relationships between data using the appropriate analytical methodologies matched to

the information and the needs of clients and users.

5) Interpret the data, model, analysis, and findings. Communicate the results in a meaningful way.

**Introduction**:

Our stakeholders, ABC, a lobbyist group based in NYC, have expressed their interest in understanding the reasons behind the rising cost of food in the United States. They also want us to investigate if there are similar cost increases in other sectors of production. This request stems from their role as lobbyists for the Agriculture Improvement Act, which is scheduled to expire in September of this year.

Additionally, they would like us to determine the percentage of US imports and exports concerning food items, examine the historical increase in the price of grocery items in the US by group, identify which food products directly influence changes in the US CPI for Food in Urban Cities, explore external factors potentially linked to the change in the US CPI for Food in Urban Cities, assess the feasibility of predicting increases in the inflation rate and the US CPI, and identify the primary factors, other than inflation, contributing to the recent rise in food prices in the US.

**Question:**

1. **How has the average price of groceries changed over time considering impacts such as inflation?**
2. **What are the primary factors, other than inflation, causing the recent increase in food prices in the US?**
3. **Should the government provide additional resources like SNAP (Supplemental Nutrition Assistance Program) to help mitigate the impact of increasing grocery prices or would investing in other sectors address this issue effectively?**

**Data Obtained & Cleaned for Analysis:**

In order to perform our analysis for our stakeholders we acquired multiple data sets to try to give us some insight to help answer these questions. We decided on using the Bureau of Labor Statistics to find data sets that include the inflation rate in the US, energy, the US unemployment rate, and the Consumer & Producer Price Index. To get these multiple data sets we made use of the multiscreen data tool provided on the BLS website to find specific data sets to help build a dataset for analysis. We also leaned towards the most popular surveys when it came to the CPI for specific food items since our previous topic modeling was inconclusive in finding what food products rose in price from the public.

When looking into data that provided insight of the US importation and exportation practices, we decided to investigate the datasets and documentation provided by the United States Department of Agriculture, as well as the BLS data base for the find data on the percentage of change. To find our transportation data set for insight we decided to use the data available on the Bureau of Transportation Statistics website. Lastly, for us to help give recommendations to our stakeholders we decided to go to the Census Bureau website to find data on SNAP statistics, its demographic, and characteristics. When obtaining the data, we decided to use the most recent available population which was 2021. With the data tool available on the website, we were able to create custom tables in order to filter the Census data for our analysis. We were also able to find data on SNAP on the USDA website to help add to backing behind our recommendations. In addition, some of the SNAP data were used in pivot tables to simplify the data that was collected.

Luckily most of the data sets that we obtained were already cleaned by the government website database sites. For the BLS data sets we had to remove the first 9-10 lines in each dataset to clean our data frames for our analysis, and convert the period column to year , and turn the column to the index so we could create line graphs and possible time series for forecasting. In addition, new data frames were created in order to calculate the average value being analyzed per dataset, this was due to the fact our data was set up by months as the column headers and years as the index.

Our USDA data sets came already cleaned as well as our transportation statistics data as well. For our SNAP data collected from the Census Bureau database the data was preselected through their table creator on their website. Once downloaded pivot tables were created in excel to analyze the available data collected on SNAP. To fill any missing values in our datasets we used the Linear Interpolation method in python to fill any missing N/A’s, if there were not a large number observed, if so N/A’s were dropped instead. This method estimates missing values based on a linear relationship between the neighboring observations.

Lastly, to perform our regression model to find out what variable that we collected have a direct relationship to the increasing food prices in the US, we combined all food item datasets that had the percent change in average price from 1992-2023, in one dataset by left joining them by the like column year. Another data frame for modeling was created that had the factors from the Producer Price Index, as well as the US inflation and unemployment rates. The US CPI for food in urban cities averaged percent change was added to each dataset created to represent our dependent variable. These data frames where named df\_merged and df\_merged1 for our model analysis.

**Exploration:**

**A picture containing line, plot, text, diagram

Description automatically generated**A picture containing line, plot, text, diagram

Description automatically generated

* We see that there has been a larger increase in imported goods (to the left) when compared to exported goods (to the right) for the United States.

**A picture containing line, plot, diagram, text

Description automatically generated**

* We see here that there has been a steady increase in the percentage of change in the consumer price index for food in the United States urban cities.

**A picture containing line, plot, diagram, text

Description automatically generated**

* We see how the inflation rate has varied in the US over time, and we notice that it is beginning to drop since its sudden spike after the COVID-19 Pandemic.

**A picture containing diagram, line, plot, text

Description automatically generated**

* The measurement of the Producer Price Index (PPI) for transportation services gauges the average fluctuations in the earnings received by producers when selling their transportation services. It serves as an indicator of the expenses businesses encounter when acquiring transportation services, such as airfare. However, it's important to note that the PPIs do not encompass the entire cost associated with purchasing transportation services, as they exclude sales and excise taxes that businesses and households are obligated to pay for these services.

**A picture containing text, screenshot, diagram, line

Description automatically generated**

* We see that there is a large number of the population who are either in poverty and/or are receiving assistance from the government.

**SNAP Facts for New York**

****

**Interpretation (Model):**

**Model 1**

The adjusted R-squared of M2 indicates that 98% of the variation in the dataset can be explained by the dependent variable US\_CPI\_Food. The F-statistic, which measures the overall significance of our model, has a score of 79.38 with a p-value of 0.000348, indicating that our model is statistically significant. Among the variables, Bananas, Potatoes, Tomatoes, Rice, Sugar, and Chicken are found to be statistically significant.

A screenshot of a computer

Description automatically generated with low confidence

However, we need to be cautious about the high R-squared value of 98%, as it suggests a potential issue of overfitting between the model and the data. To address this concern and reduce noise, we will employ two techniques. These techniques will help us identify any new statistically significant variables and mitigate the effects of overfitting.

**Reduce Noise**

**A picture containing text, screenshot, menu, number

Description automatically generated**

The importance values obtained from both the Random Forest Regressor and Gradient Boosting Regressor are normalized and add up to 1. The Gradient Boosting Regressor is used to identify significant variables when predicting the dependent variable. It focuses on reducing noise, refining predictions, and improving accuracy by minimizing residual error and the loss function, typically measured by Mean Squared Error (MSE).

The Random Forest Regressor, on the other hand, is employed to determine the significant variables when predicting the US CPI change in food prices over time. It helps reduce noise by addressing overfitting and capturing general patterns in the data. It achieves this by averaging the predictions of multiple decision trees, which helps reduce biases and errors.

The top 5 variables of importance identified by the Random Forest Regressor for predicting the US CPI for food in urban cities are Ground beef, Pasta, Ice Cream, Coffee, and Bacon. Meanwhile, the top 5 variables of importance identified by the Gradient Boosting Regressor are Coffee, Steak, Pasta, Ice Cream, and Bacon. Notably, Bacon, Ice Cream, Coffee, and Pasta are common variables among both models. These findings suggest that Bacon, Ice Cream, Coffee, and Pasta play a significant role in predicting the US CPI for food in urban cities according to both the Random Forest Regressor and Gradient Boosting Regressor.

**Model 2**

The Adjusted R-squared of M2 is 100%, indicating that the variation in the data set created can be fully explained by the US\_CPI\_Food dependent variable. This implies that the independent variables included in the model (AvgInflation, Truck, Percent Change in US import consumption, and Year) collectively account for all the variability in the dependent variable. The F-Statistic for the model has a score of 3455 with a very low p-value of 1.86e-13, which is less than the significance level of 0.05. This indicates that the model is statistically significant, suggesting that at least one of the independent variables has a significant impact on the US\_CPI\_Food. The statistically significant variables in the model are AvgInflation, Truck, Percent Change in US import consumption, and Year. These variables have been found to have a statistically significant relationship with the US\_CPI\_Food, meaning they have a measurable impact on the dependent variable.

A screenshot of a computer

Description automatically generated with medium confidence

The R-squared value of 100% suggests that the model perfectly fits the data, which may indicate potential overfitting. Overfitting occurs when a model becomes too complex and starts capturing noise or idiosyncrasies in the data, leading to poor performance on new data. It is important to address overfitting to ensure the model's generalizability. To reduce noise and identify any new statistically significant variables, additional techniques can be applied. These techniques could include regularization methods like Lasso or Ridge regression, feature selection methods, or cross-validation to evaluate the model's performance on unseen data.

**Reduce Noise**

**A picture containing text, screenshot, number, font

Description automatically generated**

The Random Forest Regressor has identified Truck, Year, Percent Change in US import consumption, Rail, and water transportation as the top five important variables when predicting the US CPI for food in urban cities. On the other hand, the Gradient Boosting Regressor, known for its expressive nature, has also highlighted Year, Truck, Rail, Percent Change in US export consumption, and Percent Change in US import consumption as the top five important variables. These findings suggest that both models consider Year and Truck as important factors. Additionally, Percent Change in US import and export consumption, as well as Rail transportation, are significant variables for predicting the US CPI for food in urban cities.

**Model 3**

A picture containing text, line, plot, diagram

Description automatically generated

**A picture containing text, line, plot, diagram

Description automatically generated**

We performed a one step time series analysis above are the graphed results. Unfortunately, we were not able to gain many insights from these times series after about a 9-year period. When trying to gain insights from these graphs, we can see we the inflation rate is going to decrease over time and the US CPI for food may drop very slightly

**Conclusion**

When examining the demographic of individuals benefiting from the Agriculture Improvement Act (SNAP), it becomes clear that this program reaches a diverse pool, primarily consisting of individuals living in poverty or a large portion of retired citizens. Given this observation, I highly recommend that ABC continues its investment in lobbying for the renewal of this act. In the United States, the impact of inflation is felt across nearly every sector, resulting in increased costs and prices of production. While food costs are beginning to return to levels like the previous year, a closer examination of price changes over time reveals a consistent upward trend in the cost of living, particularly concerning when trying to budget for rising food expenses.

To address the effects of rising inflation rates and escalating costs of goods and services, it becomes important for the United States to focus on boosting exports rather than relying heavily on imports. This strategic shift would help alleviate the direct impact on consumers and the overall economy. Due to the escalating inflation, transportation expenses, energy prices, and grocery costs, I strongly recommend that ABC stakeholders advocate for an increase in SNAP funding, rewards, and lower the income requirements associate with receiving the benefits. This step recognizes the need for additional support to cope with the challenges posed by rising costs.

Lastly, it would be beneficial for ABC stakeholders to specifically lobby for the creation of a new assistance program aimed at helping residents in New York cope with the rising costs of living, particularly in relation to food. By addressing the specific needs of New York residents, ABC can play a pivotal role in alleviating the financial burdens associated with the current cost trends. ABC Stakeholders' continued investment in lobbying for the renewal of the Agriculture Improvement Act (SNAP) is crucial, given its positive impact on a diverse group of individuals who are either in poverty, have families or are elderly. We believe advocating for increased SNAP funding and rewards, along with the creation of a targeted assistance program for New York residents, will address the challenges posed by inflation, transportation, energy, and grocery costs.

**Resources**[Transportation Economic Trends: Transportation Costs - Businesses (bts.gov)](https://data.bts.gov/stories/s/Transportation-Economic-Trends-Transportation-Cost/2yqq-baqd/)

[PPI Databases : U.S. Bureau of Labor Statistics (bls.gov)](https://www.bls.gov/ppi/databases/)

[USDA ERS - U.S. Agricultural Trade](https://www.ers.usda.gov/topics/international-markets-u-s-trade/u-s-agricultural-trade/)

[United States - Census Bureau Tables](https://data.census.gov/table?q=United+States&t=Earnings+(Individuals):Income+(Households,+Families,+Individuals):Income+and+Earnings:SNAP/Food+Stamps&g=010XX00US_010XX01US&y=2016)

[SNAP Benefits Recipients (census.gov)](https://www.census.gov/programs-surveys/saipe/guidance/model-input-data/snap.html)

[USDA ERS - U.S. Food Imports](https://www.ers.usda.gov/data-products/u-s-food-imports/)

[BLS Data Finder](https://beta.bls.gov/dataQuery/find?fq=survey:%5bcu%5d&s=popularity:D)

[Databases, Tables & Calculators by Subject (bls.gov)](https://www.bls.gov/data/)

[Yearly trends in SNAP participants, unemployment, and poverty | Food and Nutrition Service (usda.gov)](https://www.fns.usda.gov/yearly-trends)

[Consumer Price Index (CPI) Databases : U.S. Bureau of Labor Statistics (bls.gov)](https://www.bls.gov/cpi/data.htm)

[B19058: PUBLIC ASSISTANCE INCOME OR ... - Census Bureau Table](https://data.census.gov/table?q=United+States&t=Income+and+Poverty:SNAP/Food+Stamps&g=040XX00US36&tid=ACSDT1Y2021.B19058)

[Top Picks (Most Requested Statistics) : U.S. Bureau of Labor Statistics (bls.gov)](https://data.bls.gov/cgi-bin/surveymost?ap)

[Historical Inflation Rates: 1914-2023 (usinflationcalculator.com)](https://www.usinflationcalculator.com/inflation/historical-inflation-rates/)

[A Guide to Time Series Forecasting with ARIMA in Python 3 | Digital Ocean](https://www.digitalocean.com/community/tutorials/a-guide-to-time-series-forecasting-with-arima-in-python-3)

[A Guide to Time Series Visualization with Python 3 | DigitalOcean](https://www.digitalocean.com/community/tutorials/a-guide-to-time-series-visualization-with-python-3)

**Appendix:**

**Attached Separately**