



**Table of Contents**

[***Topic: Predicting Food Price Trends in the United States 2***](#_heading=h.gjdgxs)

[***Dealing with Missing/ Incomplete Data 3***](#_heading=h.30j0zll)

[***Significant Data Exploration 5***](#_heading=h.1fob9te)

[***Feature engineering 9***](#_heading=h.3znysh7)

[***Predictive Modeling 12***](#_heading=h.2et92p0)

[***Visualizing data for insight generation 12***](#_heading=h.tyjcwt)

**Consultant Report**

**Executive Summary**

# Topic: Predicting Food Price Trends in the United States

**Introduction:**

This executive summary comprehensively describes the predictive analysis conducted to forecast the fluctuation in food prices in the United States. The analysis utilized historical data on US CPI Food at Home, US Unemployment rate, US Gas price, and US Food Inflation Data. In particular, the data were analyzed using different analytical techniques, including dealing with missing/incomplete data, significant data exploration, feature engineering, predictive modeling, and visualizing data for insight generation. The analysis seeks to provide valuable insights and recommendations on food price fluctuations. The findings aim to enlighten stakeholders like business owners, policymakers, and customers with actionable information.

**Data Description:**

The dataset captures US economic indicators variables, including CPI\_Food\_At\_Home, Unemployment\_Rate, Retail\_Gas\_Price, and Food\_Inflation. The data cover the period from October 2022 to February 2024.

**Dataset Variables:**

* **CPI\_Food\_At\_Home**: The Consumer Price Index (CPI) for Food at Home measures.
* **Unemployment\_Rate:** The unemployment rate represents the percentage of the total labor force that is unemployed and actively seeking employment during a given period. **Retail\_Gas\_Price:** This variable means the
* retail price of gasoline, typically measured per gallon or liter, indicating the cost consumers pay for fuel at gas stations.
* **Food inflation:** Food inflation represents the rate at which food item prices increase over time.

**Insights:**

* The analysis indicated that unemployment rates, retail gas prices, and food inflation significantly contribute to food price instability. Therefore, an in-depth understanding of these factors is necessary for predicting and managing food price movements.
* Strong associations between CPI for Food at Home and Food Inflation indicate a solid relationship between consumer prices and overall food inflation.
* Gas prices and unemployment rates are significant coefficients in predicting food inflation, which means any increases in gas prices and unemployment rates will positively result in higher food inflation.
* Adjusting the model to handle collinearity and non-linearity issues decreased the model's accuracy power. However, it enhanced the model's performance in dealing with the complex relationship between the variables.

**Recommendations:**

* Stakeholders, including policymakers, businesses, and consumers, should closely monitor unemployment rates, gas prices, and food inflation. to make proactive measures to handle any anticipated food price movement.
* Policymakers should consider strategic measures to manage employment rates and energy prices to stabilize food price fluctuation.

**Technical Appendix**

# 

# Dealing with Missing/ Incomplete Data

**Technical Description**

**Firstly,** Datasets containing US economic indicators such as the Consumer Price Index for Food, Unemployment Rates, Gasoline Prices, food Inflation data, and producer inflation are loaded and pre-treated. This involves cleaning, transforming, and preparing the data for analysis. Economic data is chosen due to its applicability and clarity. **Secondly**, date columns within the datasets are standardized to a consistent datetime format and set as indices. This ensures uniformity across all datasets and facilitates time series analysis. Additionally, columns are renamed for clarity to aid in merging and analysis. **Then**, Multiple data frames are merged based on their datetime indices using an "inner" join. This join type ensures that only dates with complete data across all datasets are included in the final merged dataframe, thus filtering out incomplete records. **After that,** the techniques used, such as datetime index normalization and column alignment, help maintain data integrity and reduce the risk of misinterpretation. However, using an "inner" join simplifies the data structure and may lead to data loss if not all datasets have overlapping date ranges.



**Managerial Description:**

Data preparation involved loading and processing various historical economic datasets to simplify the analysis. By standardizing date formats and merging datasets based on datetime indices, we ensure data consistency and completeness. Moreover, renaming columns enhances clarity, readability, and ease of analysis.

**Overall Result:**

This data preparation process ensures that economic data is organized for analysis in a consistent and standardized manner. Specifically, handling issues such as datetime formatting and data merging enhanced data integrity and minimize the risk of misinterpretation.

# Significant Data Exploration

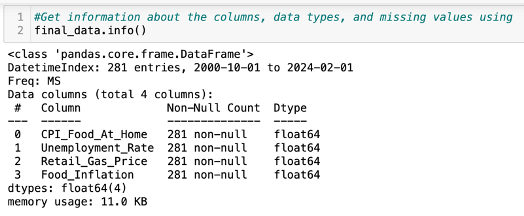
**Technical Description:**

The dataset was pre-explored to gain an in-depth understanding. **Therefore**, examining its dimensions using NumPy arrays attribute .shape revealed all the observations available for analysis. In particular, this attribute showed that the dataset comprises four floating-point (float64) variables: CPI\_Food\_At\_Home, Unemployment\_Rate, and Retail\_Gas\_Price.

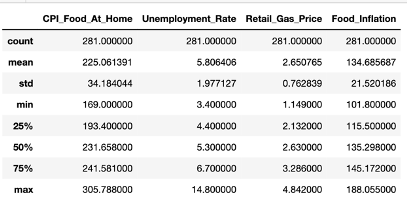
**Next,** the **.info()** method indicated that the dataset contains 281 entries from October 2000 to February 2024 with a monthly frequency. It includes four columns:

1. **CPI\_Food\_At\_Home**: This column represents the Consumer Price Index (CPI) for food at home. It contains 281 non-null float64 values.
2. **Unemployment\_Rate**: This column represents the unemployment rate. It contains 281 non-null float64 values.
3. **Retail\_Gas\_Price**: This column represents the retail price of gasoline. It contains 281 non-null float64 values.
4. **Food\_Inflation**: This column represents the food inflation rate. It contains 281 non-null float64 values.

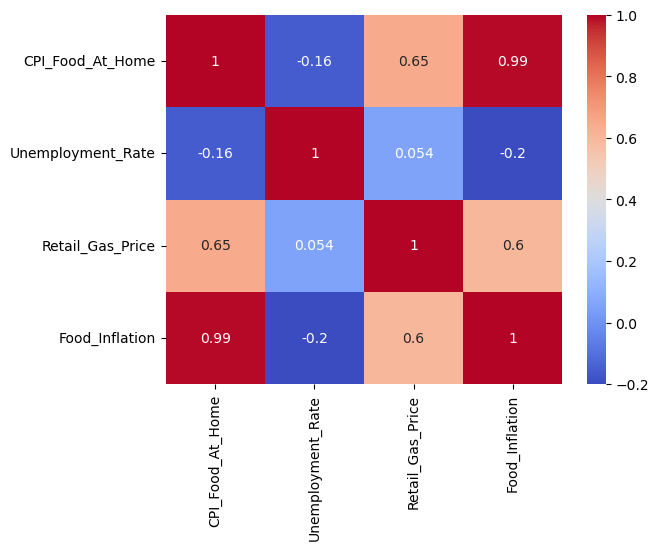
The data is generally complete, with no missing values in any of the columns. The DataFrame's memory usage is approximately 11.0 KB. Overall, the data frame provides a brief overview of economic indicators over the specified period, allowing for further analysis and exploration of trends and relationships between the variables.



**After that,** the. describe() method provided summary statistics and revealed the key characteristics of the dataset across all the variables: CPI\_Food\_At\_Home, Unemployment\_Rate, Retail\_Gas\_Price, and Food\_Inflation. The count indicates that there are 281 observations for each variable. On average, CPI\_Food\_At\_Home stands at 225.06, Unemployment\_Rate at 5.81, Retail\_Gas\_Price at 2.65, and Food\_Inflation at 134.6. Standard deviations also present the variability around the means, with CPI\_Food\_At\_Home having 34.18, followed by Food\_Inflation at 21.52. The minimum and maximum values show the range of the observed data, with noticeable differences between variables. Overall, these statistics offer a concise overview of the dataset's central tendency, spreading, and range for each variable, indicating readiness for further analysis and interpretation.



**Then,** a heat map was then utilized to picture the relationships between variables, enabling the identification of correlations. This graphical representation enabled a deeper understanding of the four variables' intercorrelation, which enhanced the analysis process. The map generally reveals the correlations between variables and visualizes the strength and direction of linear relationships between pairs of variables in a dataset. The correlation values range from 0.2 to 0.99, suggesting varying degrees of correlation between different pairs of variables. With a strong correlation between CPI\_Food\_Home and Food\_Inflation at 0.99, as presented in the figure below.



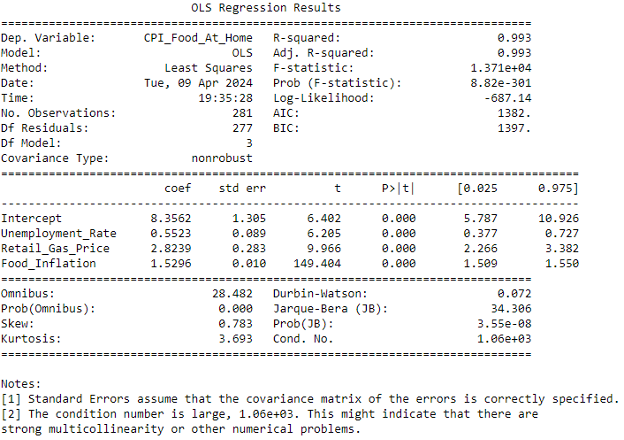
**Finally,** the OLS model also verified the relation between all the variables and how they contribute to explaining the food price swing. Unemployment, gas prices, and food inflation are good reasons for food prices to swing. Our model tells us that the roles of food-related factors are over 90% the same as those observed in the data, which is well evidenced by the R-squared figure, which is more than 0.9. We are not economics’ predictors; our model isn't only making assumptions. It's shedding some light on the factors and how they affect prices.

The best part of our results is the F-statistic, which makes them bigger than randomness. This confirms that the job market, gas prices, and food inflation are among the factors that extremely matter in overrunning the cost of food. This is how the exact impact of each of these elements is observed and how they either make food prices expensive or cheap.

The model also highlights some of the complications in the analysis, which may involve autocorrelation and multicollinearity. These words, in fact, mean to be aware of the truth and encircle the problematical points of the data that you've already handled.

In essence, this exploration helps us understand the forces behind food prices in a very detailed way. It delves deeper into this subject than simply stating numbers and figures; it allows for further analysis of specifics, offering practical informational data for decision-making and policymaking. In that respect, this type of evaluation differs by identifying the most beneficial ways to manage the economy and ensure that adequate amounts of cheap food are available for everyone.

This not only gives us more information about what is going on behind the numbers and stories but also proves the significance of exploring further in data to find the hidden stories.



**Managerial Description:**

The technical analysis's result offered insights for informed decision-making. It highlights the dataset's readiness, appropriateness, and completeness for further investigation. Summary statistics explained central tendencies and variabilities necessary for the data evaluation. A heatmap visualizes correlations, particularly emphasizing the vital link between CPI\_Food\_At\_Home and Food\_Inflation. The OLS model emphasizes the significance of unemployment, gas prices, and food inflation in explaining food price fluctuations, supported by high R-squared and significant F-statistic.

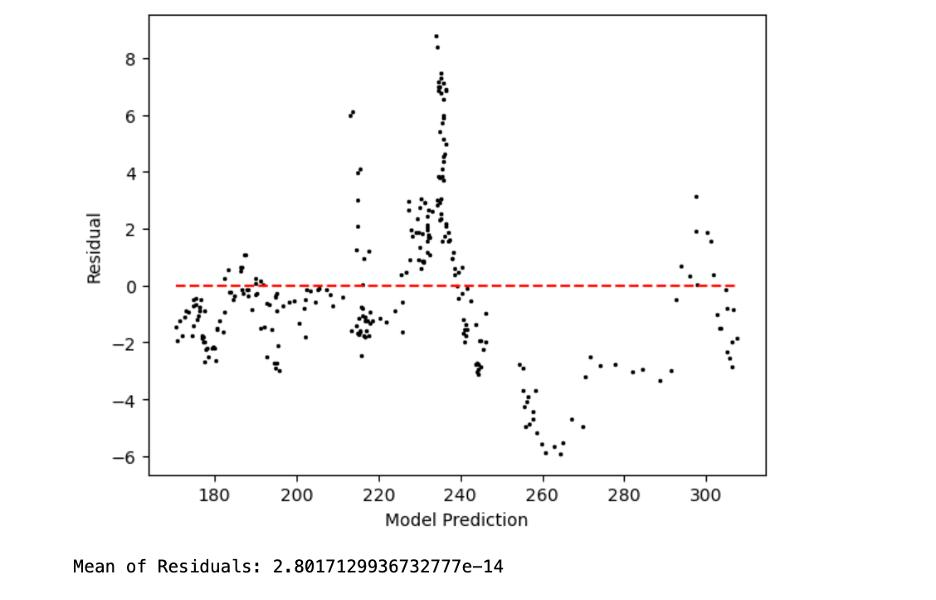
**Overall Result:**

The findings indicated that unemployment, gas prices, and food inflation significantly impact food prices. Understanding these influences enables decision-makers to make informed decisions to keep food prices affordable and manageable.

# 3. Feature engineering

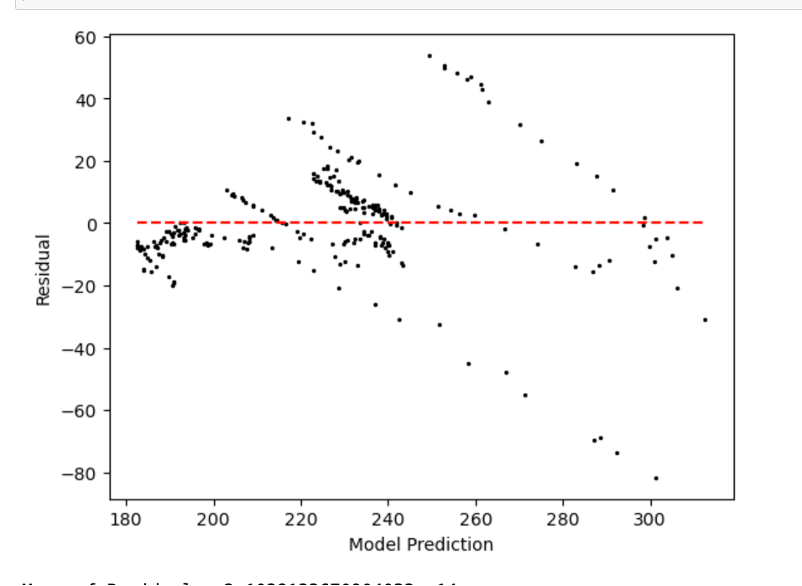
**Technical Analysis:**

The initial findings of the data exploration indicated high collinearity between "Food Inflation" and the target variable, "CPI at home”. This result signaled the need for a change in the model's parameters. Furthermore, using a residual plot graph revealed a parabolic pattern, concluding the exitances of a non-linear relationship between the variables. These insights necessitate adjustments to the model's structure for more significant analysis.

****

**Firstly,** to address the collinearity issue, the decision was made to eliminate the highly collinear "Food Inflation" variable and use the "Producer Price Index (PPI)" as a replacement. The PPI's ability to track monthly price movements associated with crop production determined this replacement. Therefore, this data set was added as a third parameter in the model.

**Secondly,** to alleviate the non-linear relationship indicated by the parabolic shape in the residual plot, the Unemployment Rate" and "Retail Gas Price" parameters were squared. As a result of this transformation, the residual plot showed improved randomness in the error terms. However, it's noteworthy that this adjustment led to a decrease in the model's r-squared value to 0.754, suggesting a trade-off between model complexity and explanatory power. (Refer to Appendix B for the new Residual Plot).

****

Overall, the technical analysis clarified the process of refining the model based on observed evidence and statistical techniques used. By addressing both collinearity and non-linearity, the final model was better developed to capture the CPI dynamics at home. Therefore, the final model we developed is: “**CPI\_Food\_At\_Home ~ Unemployment\_Rate\_squared+Retail\_Gas\_Price\_squared+Producer\_Inflation”**

**Managerial Description:**

The modification in model parameters showed the significance of data-driven decision-making. In particular, critical challenges were resolved using statistical techniques like correlation analysis and examining residual plots. Replacing "Food Inflation" with "Producer Price Index" not only shaped predictive accuracy but also aligned with the overall analysis goals. Squaring "Unemployment Rate" and "Retail Gas Price" enhanced model performance. Despite a slight decrease in the model accuracy power, this adjustment highlights the team's commitment to refining the model for better results.

**Overall Result:**

The feature engineering process, relying on the initial findings of the model development, improved the predictive modeling outcomes. Statistical techniques like correlation analysis and residual plot examination helped handle collinearity and non-linearity issues. From a managerial perspective, these changes highlight the importance of decision-making based on preliminary observations and their significance in making decisions. Using the "Producer Price Index" instead of "Food Inflation" shows a dedication to using the correct data for better predictions. Using different methods, like squaring parameters, proves a willingness to make the model work better. Altogether, combining technical skill and tactical planning can guide in creating a robust, useful final model.

# Predictive Modeling

**Technical Description:**

Building on the feature engineering and the regression work above, we wanted to test the model's predictability power. **Firstly,** the data set was split into two sets: one holds the target variable (y), and the other includes the predictor variables (x). **Secondly,** using train\_test\_split from the sklearn package we split each y and x data table into train and test files and used a simple 70 train / 30 test split. We now have 4 data tables: x\_train, y\_train, x\_test, y\_train. **Next,** we build the regression using x\_train and y\_train using LinearRegression from the sklearn package. **After that,** we plug the parameters from the x\_test data into the regression formula determined in the previous step. This produces a set of calculated y variables that we can compare to the actual y\_test data to see how different the calculated y variables are. To double-check that the split's train and test data produced are similar, we ran R-squared, Root Mean Squared Errors (RMSE), and Mean Absolute Error stats on both sets and confirmed no change in the data.

**Managerial Description:**

The splitting of data into train/test sections is relatively intuitive. By splitting the data, you end up with a more significant chunk of data to try to build a predictive model and a set of data to test the predictability of that model instead of using all the data to build a model and having no data or having to wait for more data to be able to test the predictability. Using the two sets should be mutually exclusive as you do not want data to influence the model. Still, they should also be a test point, as this will overestimate the predictability of your model. Nor do you want a small training data set. There might be influences within the data that may get overlooked if the training data set is too small; one example is seasonality; you might not want to predict your August product sales based only on January to March data points. Our project had close to 300 monthly data points, not an extreme amount, but enough to cover monthly fluctuations in the results. We decided using a 70/30 train/test split was suitable for our purposes.

**Overall Result:**

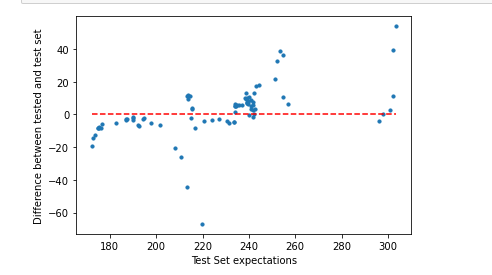
In general, the training data indicated a good fit for the model, with an R-squared (R^2) value is 0.752. Specifically, it is revealed that 75.2% of the variance in the target variable (y) is explained by the model's predictor variables (x). Similarly, the test data R-squared (R^2) value is 0.756, suggesting that 75.6% of the variance in the target variable is explained by the model's predictions on the test.

# Visualizing data for insight generation

**Technical Description:**

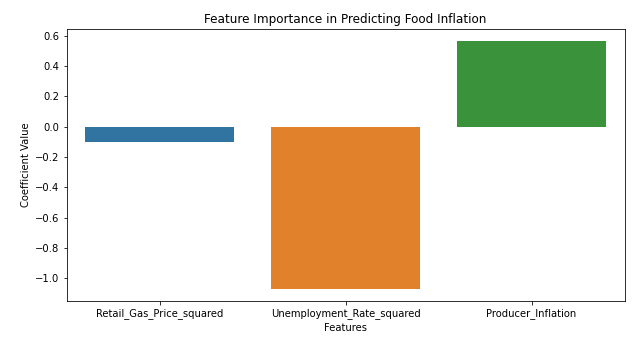
The visualization process involved two main analyses: a Residual Plot Analysis and a Feature Importance Analysis. These analyses provided insights into the performance of the regression model used to predict Food Inflation.

The Residual Plot Analysis plot and the calculated Mean Squared Error (MSE) provide key insights into the performance of the regression model used to predict Food Inflation. The visualization shows that the residuals are scattered around the zero line, which indicates that the model does not suffer from non-random error structures. In conclusion, there is no clear pattern of increasing or decreasing variance in residuals across the range of test set expectations, suggesting that the variance of the error terms is constant (homoscedastic).



**Feature Importance Analysis**

Following this analysis, a Feature Importance Analysis is applied to determine the most influential features in predicting food inflation. This analysis involves assessing the coefficients of various features and their impact on the predictive model.



**Looking at the chart, we can generate insights below:**

* **Retail Gas Price Squared** has a significant positive coefficient, suggesting that as gas prices increase, there is a noticeable positive effect on food inflation. This might be due to the increased transportation and production costs associated with higher fuel prices.
* **Unemployment Rate Squared** has a positive coefficient but is less pronounced than the gas price. This indicates that higher unemployment rates, which may reflect economic downturns, can also contribute to increased food inflation, potentially due to reduced supply chain efficiency or increased costs in other areas impacting food production and distribution.
* **Producer Inflation** is a positive and quite strong indicator. It indicates a direct relationship between the increase in producer prices and food inflation. This suggests that as the cost of goods and services at the production level increases, so does the cost of food.

These insights can help understand the key drivers of food inflation and might be helpful for policymakers and economists interested in controlling inflationary pressures in the food sector.

**Managerial description**

The overall results explain how some factors impact real-life scenarios, such as retail gas price squared, which is influenced by food inflation due to higher fuel costs. It also depicts how understanding these drivers can help policymakers make informed decisions to manage food inflation.

**Overall Result**

The visualization highlights key insights, emphasizing the importance of gas prices and unemployment rates in predicting food inflation. It also facilitates generating understandable results for more persuasive, informed decision-making.