Question 3 - Describe in words how you would automate the process of extracting the data.

As wanted goal of capturing, transforming, and displaying U.S. inflation data, I would start by identifying how to retrieve this information. Given the goal of accessing U.S. inflation series (CPI) from the BLS (Bureau of Labor Statistics), the next step is to explore the BLS website, https://www.bls.gov/, to find the endpoint URL for the data. At this URL, we can find the latest API Signatures, specifically BLS Public Data API Signatures (Version 2.0). This link provides the necessary information for data extraction.

Next, we need to find the series IDs for:

- CPI All items, seasonally adjusted
- CPI All items, less food and energy, seasonally adjusted
- CPI Gasoline (all types), seasonally adjusted

The following links help identify and verify these series IDs:

- https://www.bls.gov/cpi/factsheets/cpi-series-ids.htm
- https://www.bls.gov/cpi/additional-resources/cpi-item-aggregation.htm
- https://data.bls.gov/cgi-bin/surveymost?bls

The series IDs are:

- CPI All items, seasonally adjusted: CUSR0000SA0
- CPI All items, less food and energy, seasonally adjusted: CUSR0000SA0L1E
- CPI Gasoline (all types), seasonally adjusted: CUSR0000SEGA

With the series IDs and endpoint, we'll set a start year and end year, covering 2014 to 2024.

After extracting the data, we will format it into a dataframe using pandas, pivot it for the desired structure, and save this pivot table in CSV format for analysis. Later, we can extract this data using FastAPI for further usage.

To fully automate, we need to change the trigger from manual to automatic. We can adjust the script to retrieve only the previous month's data by filtering just the previous month from the dataframe. We append the previous month's data to the CSV file in the project directory and run the process monthly at the beginning of each month, giving the BLS enough time to update their database. The CSV will be updated with data from month -1 (M-1). This can be automated using the Windows Task Scheduler to run the Python scripts.

As a future improvement, we could move the solution to a cloud-based platform like AWS, storing the CSV file in a bucket.

Question 4 - Explain how you would relate the prices series (All items) with the Gasoline (Gasoline) prices series.

To relate the price series (All items) with the Gasoline prices series, I would first retrieve the CSV file that has these data. After that, I would load the CSV file into a pandas DataFrame, ensuring the data is clean by handling any missing values and converting the date column to a datetime format.

Next, I would perform a linear regression analysis to understand the relationship between the All items price series and the Gasoline price series. Linear regression helps determine how changes in Gasoline prices can predict changes in All items prices.

To fully understand if the model works, I would do a few tests, such as the Durbin-Watson test to check for the presence of autocorrelation in the residuals from the regression analysis and the Student's t-test to check whether the coefficients of the regression are significantly different from zero. The Durbin-Watson test is important because autocorrelation can invalidate some of the assumptions of linear regression, and the t-test helps determine if Gasoline prices have a statistically significant effect on All items prices.

After performing these tests, I would analyze the coefficients, p-values, and R-squared values from the regression output. The coefficients indicate the magnitude and direction of the relationship between Gasoline prices and All items prices. The p-values help determine the significance of these relationships, with low p-values (typically < 0.05) indicating that the relationships are statistically significant. The R-squared value indicates how well the independent variable (Gasoline prices) explains the variation in the dependent variable (All items prices). A higher R-squared value means a better fit of the model to the data.

To ensure robustness, I would also check the F-statistic and Prob (F-statistic) to confirm the overall significance of the model. Finally, I would review any potential issues like autocorrelation or non-normality in the residuals, using tests like the Ljung-Box and Shapiro-Wilk tests, and make any necessary adjustments to the model to ensure accurate and reliable results.

Analysis of the Relationship Between Gasoline Prices and Overall Price Levels Using Linear Regression

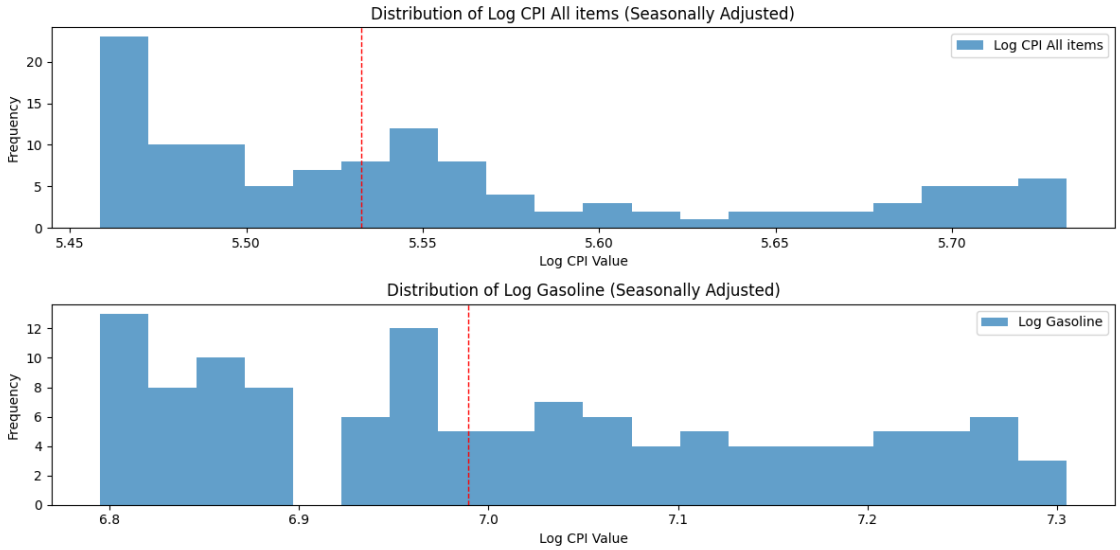
Table of Contents

Introduction

The data for this analysis, sourced from the U.S. Bureau of Labor Statistics, covers the period from 2014 to 2024 and includes seasonally adjusted CPI for all items and gasoline prices. Log transformation was applied to stabilize variance, linearize relationships, and reduce the impact of outliers. The analysis uses linear regression to explore the relationship between log-transformed CPI and gasoline prices. Various statistical tests, including the Durbin-Watson statistic, Student's t-test, ANOVA, Levene's test, and Shapiro-Wilk test, were conducted to ensure the validity and reliability of the model.

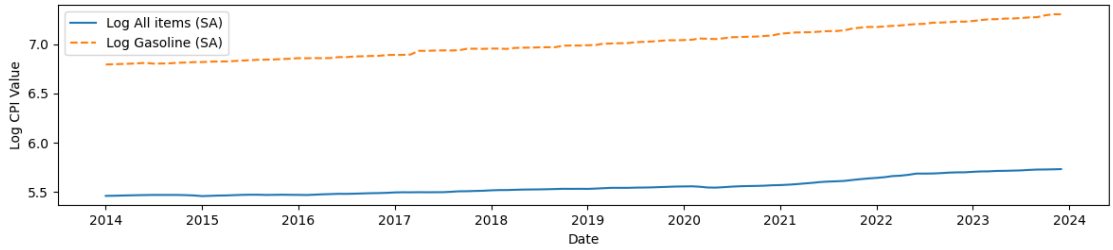
Distribution Analysis of Log-Transformed CPI All Items and Gasoline Prices

Analyzing the distribution of log-transformed CPI All Items and Gasoline prices is crucial to ensure the data is suitable for regression analysis. Log transformation stabilizes variance and reduces the impact of outliers, improving the accuracy of our statistical tests and models.



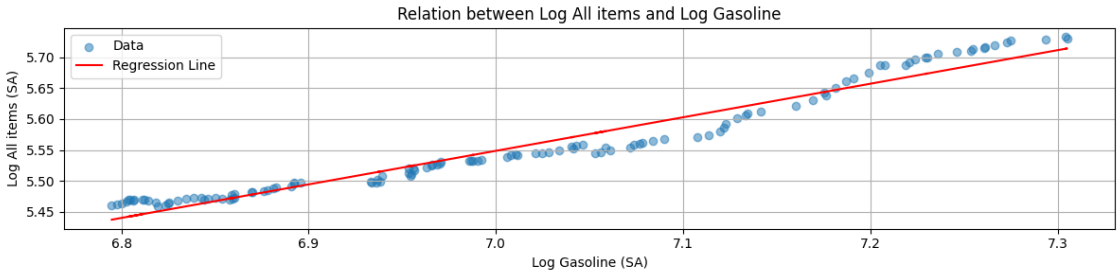
The histograms show that the log-transformed data for CPI All Items and Gasoline prices are reasonably well-behaved, reducing skewness and stabilizing variance. This supports using these log-transformed values in our linear regression analysis to explore the relationship between gasoline prices and overall inflation. The nearly normal distributions, indicated by red dashed median lines, confirm that the data is well-prepared for regression analysis. This distribution analysis ensures that our approach will provide reliable insights into the relationship between these variables.

Log-transformed CPI All Items vs. Gasoline Prices, 2014-2024



The first image shows a time series plot of the log-transformed CPI for all items and log-transformed gasoline prices from 2014 to 2024. The trends for both variables are steadily upward, indicating a stable relationship over time

Log-transformed CPI All Items vs. Gasoline Prices with Data Points and Regression Line



The second image illustrates a strong linear relationship between log-transformed gasoline prices and log-transformed CPI values, as shown by the tight clustering around the regression line. This visual evidence supports the conclusion that changes in gasoline prices are closely associated with changes in the overall CPI.

The analysis of the relationship between the price series for All items and Gasoline prices shows significant findings. Using linear regression, the results indicate a strong relationship between these two variables.

Metrics & Tests

R-squared and Adjusted R-squared

These metrics indicate a strong fit of the model to the data.

Metric	Value	Explanation
R-squared	0.951	95.1% of the variability in CPI All Items can be explained by gasoline prices.
Adjusted R-squared	0.950	Adjusts R-squared for the number of predictors.

These metrics indicate a strong relationship between gasoline prices and CPI All Items, confirming that changes in gasoline prices can largely explain changes in overall CPI.

Coefficient for log Gasoline

This coefficient quantifies the relationship between gasoline prices and CPI All Items.

Metric	Value	Explanation
Coefficient for log_Gasoline	0.5429	Suggests that a 1% increase in gasoline prices is associated with a 0.5429% increase in CPI All Items.
Standard Error for log_Gasoline	0.011	Indicates a precise estimate of the coefficient.

This coefficient quantifies the positive impact of gasoline prices on CPI All Items, showing that gasoline price changes significantly affect overall price levels.

F-statistic and Prob (F-statistic)

These metrics check the overall significance of the regression model.

Metric	Value	Explanation
F-statistic	2277	Indicates that the overall model is statistically significant.
Prob (F-statistic)	5.46e-79	Confirms the overall significance of the regression model.

These metrics confirm that the overall regression model is statistically significant, validating the model's ability to explain the relationship between gasoline prices and CPI All Items.

p-value for log_Gasoline

The small p-value confirms the statistical significance of the relationship.

Metric	Value	Explanation
t-value for log_Gasoline	47.719	Indicates the coefficient for log_Gasoline is significantly different from zero.
p-value for log_Gasoline	5.46e-79	A very small p-value suggests a highly significant relationship between gasoline prices and CPI All Items.

The small p-value confirms that the relationship between gasoline prices and CPI All Items is statistically significant, meaning the observed relationship is not due to random chance.

Durbin-Watson Statistic

The Durbin-Watson statistic was calculated to detect the presence of autocorrelation in the residuals of the regression model.

Metric	Value	Explanation
Durbin-Watson	0.04	Indicates the presence of strong positive autocorrelation in the residuals, suggesting that the residuals are not entirely independent and that there may be underlying patterns not captured by the simple linear model.

The Durbin-Watson statistic reveals issues with autocorrelation in the residuals, which may imply that the model's assumptions are violated and could affect the reliability of the results.

Ljung-Box Test for Autocorrelation in Residuals

The Ljung-Box test checks for autocorrelation in the residuals of the regression model.

Metric	Value	Explanation
Ljung-Box statistic	819.47	Indicates substantial autocorrelation in the residuals, suggesting that the residuals are not purely random.
p-value for Ljung-Box	1.104518e-167	A very small p-value confirming significant autocorrelation in the residuals of the regression model.

The Ljung-Box test confirms significant autocorrelation in the residuals, suggesting that the model may need refinement to address these patterns for more accurate predictions.

ANOVA (Analysis of Variance)

ANOVA was performed to test the overall significance of the regression model.

Metric	Value	Explanation
F-statistic	8502.52	Indicates that the overall model is statistically significant.
p-value	3.08e-188	Confirms the overall significance of the regression model.

ANOVA results reinforce the statistical significance of the regression model, supporting the robustness of the relationship between gasoline prices and CPI All Items.

Median Analysis

Median values provide a measure of central tendency for the log-transformed data.

Metric	Value	Explanation
Median Log CPI All items	5.53	Median value of log-transformed CPI All Items.
Median Log Gasoline	6.99	Median value of log-transformed gasoline prices.

Median values provide additional context on the central tendency of the log-transformed data, confirming the typical levels of the variables in the study.

Levene's Test for Homogeneity of Variances

Levene’s test checks for equal variances across groups.

Metric	Value	Explanation
Levene’s test statistic	47.24	Indicates significant differences in variances.
p-value	5.46e-11	Confirms significant differences in variances.

Levene's test indicates variability in the data, which may affect the assumptions of homoscedasticity in the regression analysis.

Shapiro-Wilk Test for Normality

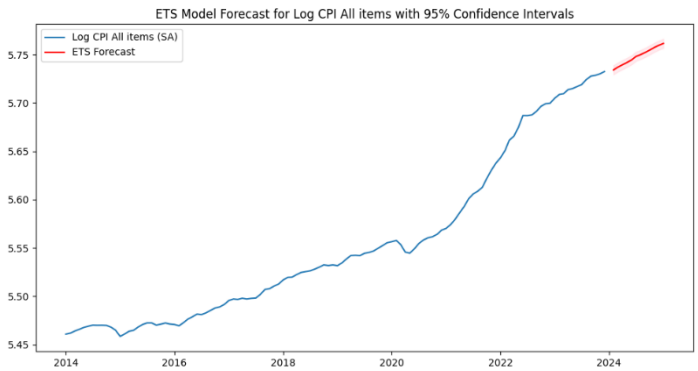
The Shapiro-Wilk test assesses the normality of the log-transformed data.

Metric	Value	Explanation
Shapiro-Wilk statistic (Log CPI All items)	0.87	Indicates non-normal distribution of log CPI All Items data.
p-value (Log CPI All items)	7.20e-09	Confirms non-normality of log CPI All Items data.
Shapiro-Wilk statistic (Log Gasoline)	0.94	Indicates non-normal distribution of log gasoline prices data.
p-value (Log Gasoline)	4.06e-05	Confirms non-normality of log gasoline prices data.

The Shapiro-Wilk test results indicate non-normality of the log-transformed data, which suggests that the data distributions are not perfectly normal and could affect certain statistical assumptions.

ETS Model Forecast

The Exponential Smoothing (ETS) model is used to forecast the log-transformed CPI All Items values from 2014 to 2024. This model accounts for seasonality and trends in the data, providing a comprehensive prediction framework. The forecast includes a 95% confidence interval, visually represented in the accompanying graph.



The consistent upward trend observed in the ETS model forecast for CPI All Items supports the relationship between gasoline prices and overall inflation. This indicates that changes in gasoline prices have a predictable and sustained impact on the general inflation trend, reinforcing the importance of gasoline prices as a key factor in understanding and predicting overall inflation.

Conclusion

The analysis demonstrates a strong and statistically significant relationship between gasoline prices and overall price levels (CPI All Items). The linear regression model shows that 95.1% of the variability in CPI All Items can be explained by changes in gasoline prices. However, it is important to note that this is a simplified model and does not account for external factors that can influence inflation. These findings indicate that while gasoline prices play a substantial role in overall inflation trends, they are not the sole factor. The robustness of the results is confirmed by various statistical tests, highlighting the importance of considering gasoline prices in understanding and predicting overall inflation.