[Final Report-DSCI510(Fall 2025)] Analyzing the Interplay of Energy Costs, Inflation, and Labor Market Shocks (2016-2025)

1. Name & Team Members

- **Name:** Junhyeon Song
- **USC Email:** songjunh@usc.edu
- **USC ID:** 8467-4177-80
- **GitHub Repository:** https://github.com/SongQoo/DSCI510_Final_Project

2. Short Description

While traditional economic indicators provide precise data, they often suffer from reporting lags. This project investigates the **lead-lag relationship** and causal chain between **input costs (Energy)**, **official economic metrics (Inflation & Unemployment)**, and **public sentiment (News Media)** in the United States over the last decade (2016–2025). By integrating multi-source data, this study visualizes how supply shocks translate into consumer pain and eventually reflect in public sentiment.

3. Data

To ensure a comprehensive analysis, I collected data from **four distinct sources** using a combination of **API integration** and **sophisticated web scraping** techniques. The dataset covers a 10-year period from **January 2016 to December 2025**, allowing for a comparative analysis of pre-pandemic stability, the COVID-19 shock, and the post-pandemic inflationary regime.

Data Sources & Collection Methods

1. **Inflation Metrics (Source A):**
    - **Source:** Bureau of Labor Statistics (BLS) Public API.
    - **URL:** https://api.bls.gov/publicAPI/v2/timeseries/data/
    - **Method:** Collected via **API** with JSON parsing. I retrieved detailed CPI-U data, specifically disaggregating into **All Items**, **Food**, **Energy**, and **Shelter** to analyze the internal dynamics of inflation.
2. **Energy Costs (Source B):**
    - **Source:** U.S. Energy Information Administration (EIA).
    - **URL:** https://www.eia.gov/petroleum
    - **Method: Web Scraping** using BeautifulSoup. Instead of simple file downloads, I implemented a script to parse raw HTML tables for **Gasoline**, **Diesel**, and **WTI Crude Oil** prices, converting weekly/daily data into monthly averages.
3. **Labor Market Data (Source C):**
    - **Source:** BLS Data Viewer (Official Employment Statistics).
    - **URL:** https://data.bls.gov/timeseries/LNS14000000 ~ LNS14000002
    - **Method: Web Scraping** using BeautifulSoup. The raw data existed in a "Wide Matrix" format (Years × Months). I scraped these tables and performed **data**

reshaping (melting) to transform them into a clean time-series format for **Total**, **Men**, and **Women** unemployment rates.

4. **Public Sentiment (Source D):**
    - **Source:** New York Times (NYT) Archive API & Article Search API.
    - **URL:** developer.nytimes.com
    - **Method: Unstructured Text Mining** via API. I constructed a hybrid pipeline using both the Archive API (for historical data) and Article Search API (for recent data) to overcome access limitations. I processed the raw text (headlines/snippets) to count the monthly frequency of economic fear keywords (e.g., "Recession", "Layoff").

Data Summary

Source ID	Dataset Name	Data Type	Collection Method	Frequency	Raw Data (Rows)	Raw Data (Columns)
Source A	CPI Metrics (Total, Food, Energy, Shelter)	Structured (JSON)	API	Monthly	117 Rows	8 Cols (Index + YoY)
Source B	Energy Prices (Gas, Diesel, Oil)	Semi-Structured (HTML)	Web Scraping (BeautifulSoup)	Weekly / Daily	1,295 Rows (Per metric)	3 Cols (Monthly Avg)
Source C	Labor Market (Total,Men,Women)	Semi-Structured (HTML)	Web Scraping (BeautifulSoup)	Monthly	120 Rows (Matrix)	3 Cols (Unemp Rate)
Source D	NYT News Sentiment	Unstructured (Text)	API	Daily	514,270 Articles	7 Cols (Keywords+Total)
Final	**Integrated Master Dataset**	**Time-Series (CSV)**	**Merge & Clean**	**Monthly**	**120 Rows (2016-2025)**	**21 Columns (Total)**

4. Data Cleaning, Analysis & Visualization

Data Processing Structure

[Step 1] Data Collection & Cleaning

The raw data collected from disparate sources (JSON, HTML, CSV, Unstructured Text) required extensive preprocessing to create a unified time-series dataset.

1. **Standardization:** All datasets were converted to a uniform monthly frequency. Weekly energy prices were averaged by month (resample('MS').mean()), and daily news articles were aggregated into monthly keyword counts.
2. **Feature Engineering:**
    - **Inflation Metrics:** Converted raw CPI indices into Year-over-Year (YoY) percentage changes to measure inflation rates accurately.
    - **News Sentiment:** Applied text mining to headline and snippet fields, creating binary flags for keywords ('Inflation', 'Recession', 'Crisis', 'High price', 'Layoff', 'Unemployment') and summing them to generate a monthly "Fear Index".
3. **Handling Missing Values:** Linear interpolation was applied to handle minor gaps in the time series, ensuring continuity for lag analysis.

[Step 2] Integration & Formatting

I merged the four cleaned dataframes (CPI, Energy, Labor, News) into a single master dataset based on the date index. The final dataset consists of **120 rows (2016-2025)** and **21 columns**, enabling direct correlation analysis across different economic dimensions.

[Step 3] Analytical Approach

My analysis moved from descriptive statistics to causal inference:

1. **Descriptive Analysis:** Evaluated volatility (CV) and distribution asymmetry (Skewness) to characterize the nature of economic shocks (e.g., Sticky vs. Flexible prices).
2. **Structural Break Detection:** Compared correlations pre- and post-2020 to identify regime shifts in the energy-inflation relationship.
3. **Lag Correlation & Causality:** Performed time-lag analysis (0-6 months) to trace the transmission of shocks from Energy -> Inflation -> Labor -> Media.

Initial Hypothesis

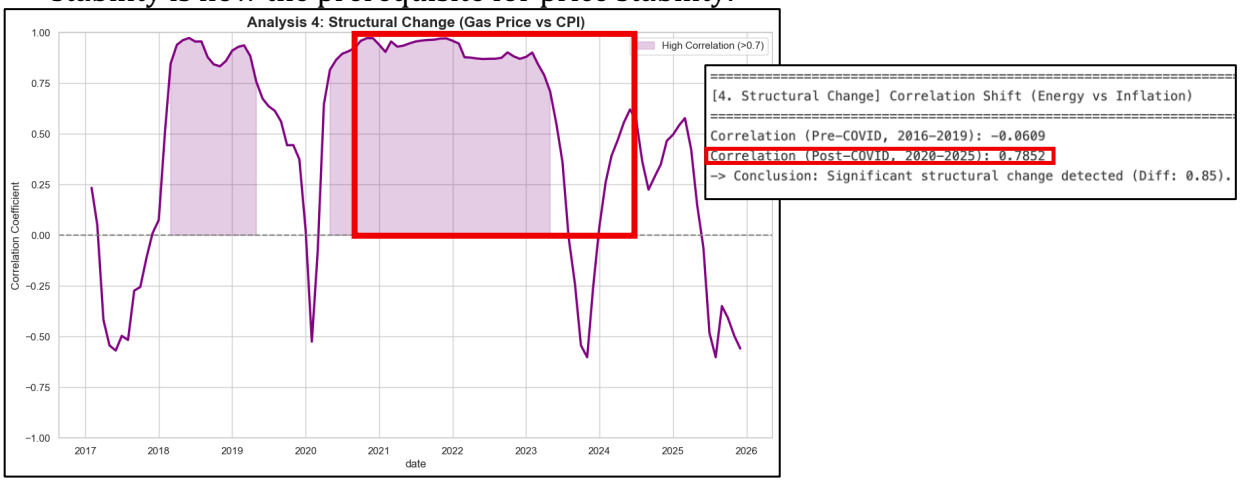
I hypothesize that economic shocks propagate through a distinct causal chain:

1. **Trigger:** Energy (Oil) shocks act as the primary leading indicator.
2. **Transmission:** Input costs (Energy) pass through to Consumer Inflation with minimal lag.
3. **Reaction:** Media sentiment is a coincident indicator, reflecting real-world labor market pain rather than predicting it.

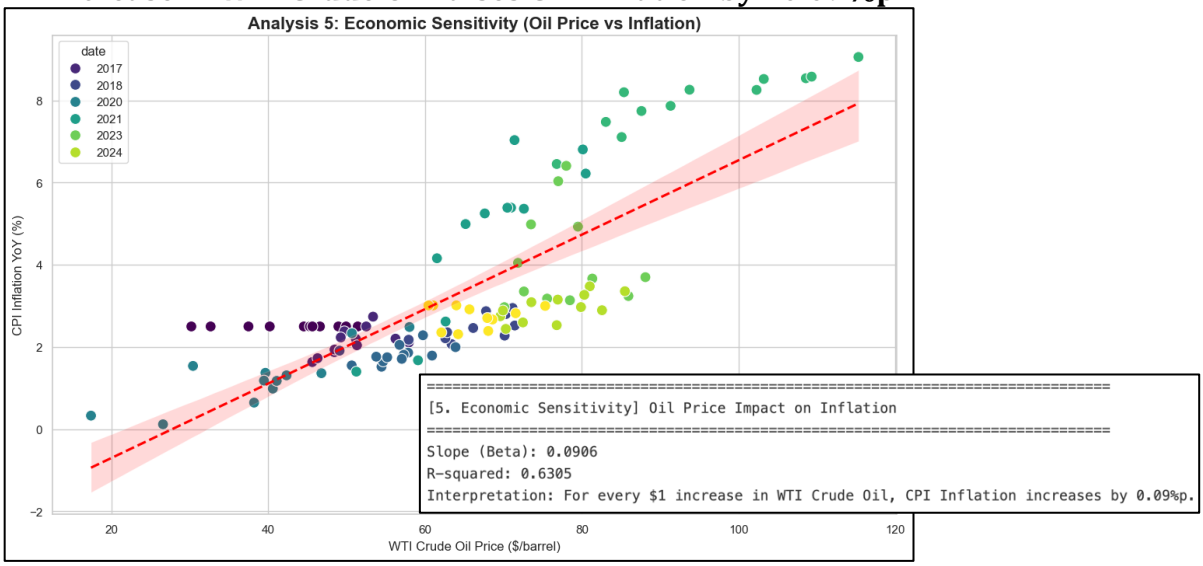
Findings & Conclusion

Through the multi-dimensional analysis, I have drawn four major conclusions that strongly support my initial hypothesis:

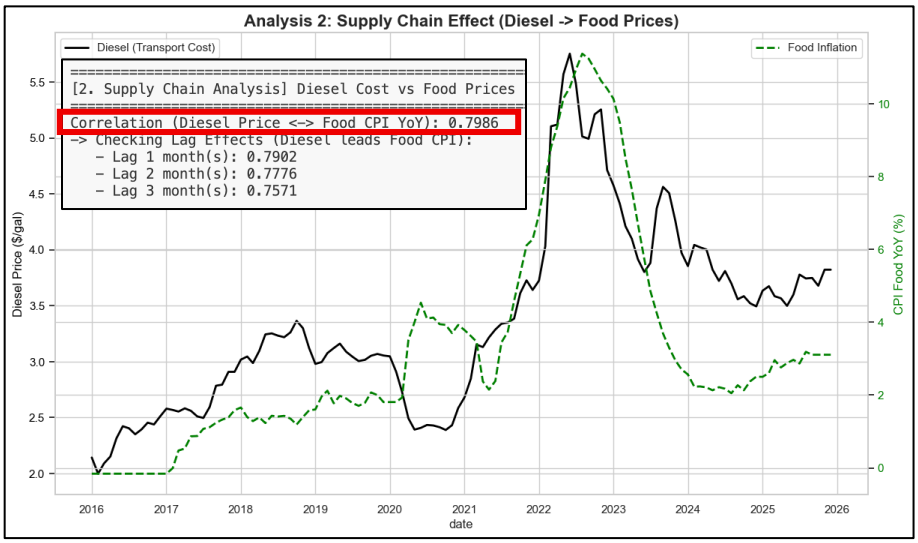
1. **Energy (Oil Price) is the 'Master Switch' (Validated):**
    - The analysis confirmed a massive structural shift. Post-2020, the correlation between Gas Prices and CPI surged from -0.06 to **0.79**. This proves that energy stability is now the prerequisite for price stability.



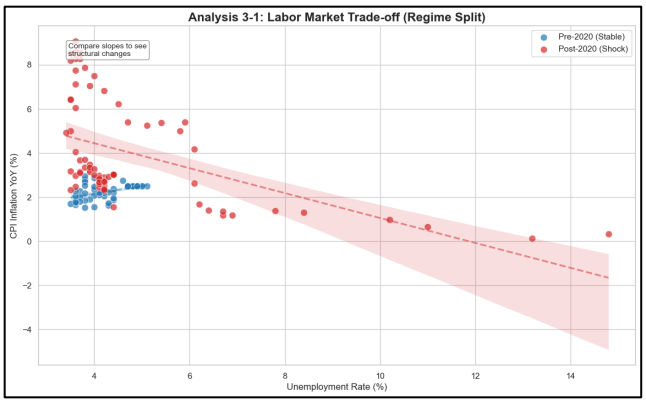
- Linear regression analysis also establishes a significant quantitative link between energy and inflation. The model shows a **Slope (Beta) of 0.09**, meaning a **$1 increase in WTI Crude Oil raises CPI Inflation by ~0.09%p**.

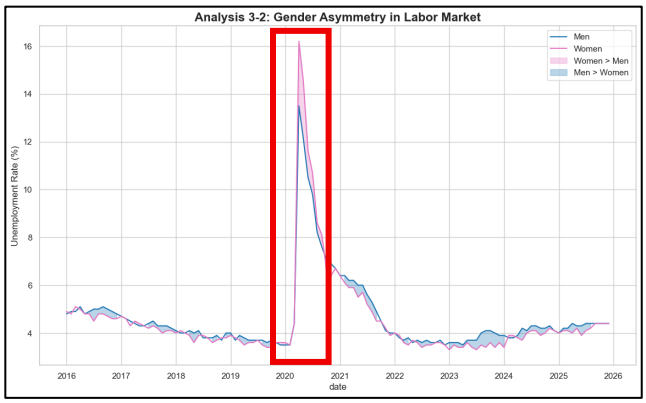


2. **The Supply Chain Effect is Immediate (Validated):**
   - I found a robust correlation (**0.79**) between Diesel prices and Food CPI. This validates the transmission mechanism where logistics costs are instantly passed on to consumers.

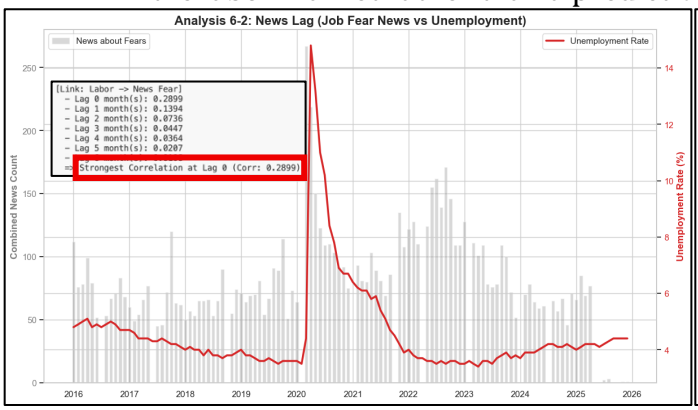


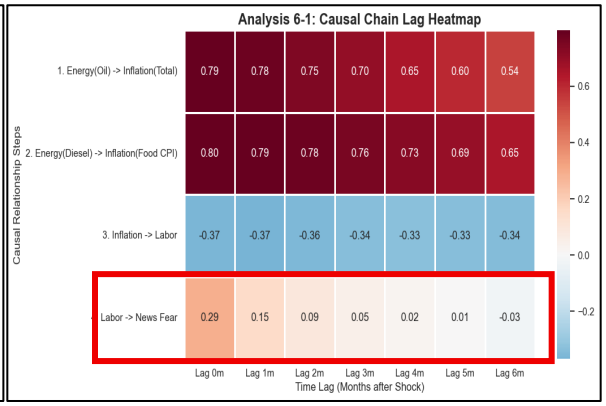
3. **Labor Market Asymmetry (New Finding):**
   - To understand the labor market response, I analyzed the relationship between Inflation and Unemployment. The scatter plot confirms a negative correlation (**-0.37**), indicating that **high inflation coincided with a "hot" labor market** (Specifically, the Post-2020 period shows a steeper trade-off, indicating that supply shocks made the inflation-unemployment dynamic more volatile)
   - Furthermore, I found a significant **Gender Asymmetry**, where female unemployment spiked sharper than males during shocks, indicating that women are significantly **more vulnerable to sudden economic downturns**.





4. **Media as a Mirror, Not a Prophet (Validated):**
    - The analysis revealed that News Sentiment is a **coincident indicator (Lag 0)** with the labor market rather than a predictive one.





<u>Conclusion</u>: A chain reaction starting from **supply-side volatility (Energy)**, propagating instantly to **consumer prices (Inflation)**, creating a deceptive trade-off in the **labor market**, and finally mirroring itself in **public sentiment**. Future policy-making must prioritize **energy price stability** as the most effective lever to control both inflation and public economic anxiety.

5. Changes from Original Proposal

1. I expanded the labor market analysis to include **Gender Gap (Male vs. Female)** segmentation, which was not initially planned.
2. Due to the latency of **NYT's Archive API** for real-time data, I adopted a hybrid approach. While the full comprehensive dataset was retrieved up to **May 2025**, the post-May period utilized the **Article Search API** with a sampling method (top 20 relevant articles per 3-day window) to overcome API rate limits while maintaining trend accuracy.

6. Mention of Future Work

If given the opportunity for further research, I would be interested in exploring a **wider range of supply-side indicators** beyond energy, such as agricultural futures and industrial metals, to build a more comprehensive cost index. Additionally, it would be valuable to **examine labor data disaggregated by age and geography** to uncover hidden demographic vulnerabilities.
Finally, I would like to experiment with **advanced machine learning models** to move beyond simple correlation analysis and potentially improve economic forecasting capabilities.