Project Proposal *(Revised)*

DSCI510-Principles of Programming for Data Science

Junhyeon Song(Email: songjunh@usc.edu, Github: @SongQoo, USC ID: 8467-4177-80)

1. Problem Statement

Inflation has recently become one of the most important economic issues affecting consumers in the United States. While traditional economic indicators like CPI (Consumer Price Index) provide precise data, they often suffer from a reporting lag. This project aims to investigate the **lead-lag relationship** between **input costs** (Energy), **official economic metrics** (Inflation/Unemployment), and **public sentiment** (News Media). Specifically, I want to answer like this: *"Do spikes in energy prices predict inflation? And does media panic (keyword frequency) precede actual labor market contractions?"*

2. Data Collection Plan (Multi-Source & Complex Integration)

To address the requirements for data collection, I have architected a pipeline that integrates **four distinct data sources** using diverse technical approaches:

- **Source A: Inflation Metrics (API) –** Collected via **BLS Public API** with JSON handling.
- **Source B**: **Energy Costs (Web Scraping) –** Collected from **EIA.gov** using **Regular Expressions (Regex)** to parse unstructured HTML tables dynamically.
- **Source C**: **Labor Market Data (Web Scraping)** – Collected from **BLS Data Viewer** by parsing HTML tables and reshaping them into time-series format using Pandas.
- **Source D**: **Public Sentiment (Text Mining) –** Collected via **NYT Archive API**, implementing an NLP script to count the monthly frequency of fear keywords (e.g., "Recession", "Layoff").

3. Planned Analysis & Visualizations

The project will leverage Python libraries such as Pandas, NumPy, and Seaborn to integrate disparate datasets, employing **time-series resampling** to align weekly energy prices and daily news sentiment into a standardized monthly frequency. Analysis will focus on **lead-lag correlations**, investigating whether energy cost spikes predict CPI inflation and if rising media "fear" keywords serve as a leading indicator for unemployment. Visualizations will feature **dual-axis plots** overlaying public sentiment against official economic metrics, along with correlation heatmaps to quantify how these interdependencies have shifted across the pre-pandemic, shock, and recovery phases.