Newsflash: Current Events Affect Commodities

Sheza Munir (shezamnr)
Era Parihar (erap)
Project Code Notebook

Motivation:

The project proposes an analytical framework to understand the relationship between emerging trends from large-scale text data, such as news articles, and commodities/macro-economics. This framework aims to correlate news trends with the performance of specific market sectors or assets.
This project stems from the realization that understanding the intricate relationship between real-world events, as captured in news articles, and economic indicators is crucial for informed decision-making across various sectors. By analyzing large-scale text data alongside commodity prices and macroeconomic data, our goal is to unearth patterns, correlations, and trends that can aid economic forecasting, strategic business planning, and trade policy analysis. The project not only contributes to the fields of data science and economics but also holds the potential to inform policy decisions, empower businesses, and shed light on the dynamic interplay between global events and economic landscapes. Through this interdisciplinary exploration, we aspire to provide valuable insights that go beyond numerical correlations, encompassing the nuanced narratives that shape our economic world.
Similar studies and reports have explored the correlation between news sentiment, economic indicators, and financial markets. Research by scholars like Tetlock et al. (2007) and Bollen et al. (2011) delves into sentiment analysis of news and its impact on stock markets, providing insights into the interconnectedness of news media and financial outcomes. Additionally, reports from financial institutions such as JPMorgan and Goldman Sachs often analyze the influence of news events on commodity prices and economic trends, offering valuable perspectives on the intersection of news and economic variables.

Data Sources:

Primary Dataset Description:

Key Features:
The dataset contains articles from HuffPost with the following features:
Link: The URL directing to the full article.
Headline: The title or main descriptor of the article's content.
Category: The main classification or genre of the news (e.g., "U.S. NEWS", "WORLD NEWS").
Short Description: A brief summary or snippet from the article.
Authors: The journalist(s) or writer(s) who authored the piece.
Date: The publication date of the article.

Estimated Size: The dataset set contains 209528 rows of data with above-mentioned columns and the total size would be around 87.3 MB.

Location: The dataset is taken from Kaggle, here is the link to the dataset: https://www.kaggle.com/datasets/rmisra/news-category-dataset/data. The dataset encompasses news headlines from HuffPost spanning from 2012 to 2022. It is notable that HuffPost ceased the maintenance of an extensive archive post-2018, thus making such a comprehensive dataset unavailable for fresh collection in the current times.

Format: JSON

Access Method:
The dataset is publicly available at the Kaggle URL given above. We directly downloaded it from this URL.

Relevant Features:
Depending on the question listed above, different features might be relevant. If analyzing trends in news categories over time, then "Category" and "Date" would be crucial. For studying the frequency of authors' contributions, "Authors" would be of primary importance.

Feature Extraction:
We will extract relevant features from the news articles. Possible features include:

- Sentiment analysis scores (positive, negative, neutral).
- Keywords or topics mentioned.
- Event or news category tags.
- Length of articles.
- Named entities (e.g., countries, companies, or individuals mentioned).
- Aggregate features over different time periods (e.g., daily, weekly, or monthly) to capture trends and patterns.

Secondary Dataset Descriptions:

Dataset: Daily US Wheat
Short description: The dataset captures daily stock market performance data, including the opening, highest, lowest, and closing prices of stocks for specific dates.
Estimated size:
Records: 2,273 rows
Bytes: Approximate estimation can be given as: Each row has its date and four float values, the data is around 90 KB.
Location: https://www.kaggle.com/datasets/nickwong64/daily-wheat-price
Access method: Public access, downloaded from Kaggle

Dataset: **Monthly Unemployment Rate**
Short Description: This dataset showcases the unemployment rate for specific months between January 2000 to November 1949.
Estimated Size: Contains 285 records. The byte size is 16 kb on disk
Location: Data provided in the image format; no direct URL or other access method provided.
Access Method: The dataset is taken from https://fred.stlouisfed.org/series/UNRATE, it's open to the public and is easily accessible.

Dataset: **Daily Crude Oil Price**
Short Description: The dataset displays the daily Crude Oil Prices for West Texas Intermediate (WTI) in January 2000 and December 2023. Two columns represent price data as per different codes: DCOILWTCO_20231018 and DCOILWTCO_20231025.
Note: West Texas Intermediate (WTI) Crude Oil Prices from Cushing, Oklahoma is widely considered a benchmark for U.S. oil prices. WTI is a particular grade of crude oil often used for oil price references in the

U.S. Energy Information Administration reports. The prices at Cushing, Oklahoma specifically are important because Cushing is a key trading hub for crude oil in North America.
Estimated Size: Contains 6210 records. The byte size is 143 kb.
Location: https://fred.stlouisfed.org/series/UNRATE
Format: Image representation of a table. The original data might be in formats such as CSV or Excel.
Access Method: The Dataset is taken from https://fred.stlouisfed.org/series/DCOILWTICO and it is open to the public and is easily accessible.

Data Manipulation Methods:

Data Conversion: Converting Date Formats and Strings to Floats
We have datasets with dates in a format that is not datetime. We converted these into datetime objects to work effectively with time series data. The numbers in the secondary datasets were strings in the dataset so we converted them to floats. We used Pandas for these conversions.

```
crude_oil_dataset['observation_date'] = pd.to_datetime(crude_oil_dataset['observation_date'])
unemployment_rate_dataset['DATE'] = pd.to_datetime(unemployment_rate_dataset['DATE'])
wheat_dataset['date'] = pd.to_datetime(wheat_dataset['date'])
```

```
crude_oil_dataset['DCOILBRENTEU_20231025'] = pd.to_numeric(crude_oil_dataset['DCOILBRENTEU_20231025'], errors='coerce')
unemployment_rate_dataset['UNRATE'] = pd.to_numeric(unemployment_rate_dataset['UNRATE'], errors='coerce')
wheat_dataset['close'] = pd.to_numeric(wheat_dataset['close'], errors='coerce')
```

Handling Missing Data: Imputation and Removal
Dealing with missing or anomalous data is another challenge we encountered. Our approach was to either fill in missing data with average values for the respective feature if we can make reasonable estimations or to remove rows if sufficient data is available. This decision depended on the specific dataset and context.

```
[ ] crude_oil_dataset = crude_oil_dataset.dropna()
    unemployment_rate_dataset = unemployment_rate_dataset.dropna()
    wheat_dataset = wheat_dataset.dropna()
```

Joining Datasets: Daily Data Integration
Joining datasets was crucial for our analysis. We have three daily datasets (textual analysis, crude oil prices, wheat prices), and one monthly dataset (unemployment rate). To prevent loss of data, we repeat the monthly unemployment rate to convert it into daily and combine.

```
unemployment_rate_resampled = unemployment_rate_dataset.resample('D').ffill()
combined_dataset = pd.concat([crude_oil_dataset, unemployment_rate_resampled, wheat_dataset], axis=1, join='inner')
```

Handling Different Timelines: Filtering Data
The timelines for our datasets are different. For instance, crude oil data covers a range from 1986 to 2023, while the textual data spans from 2012 to 2018. To address this, we specified the time range that was common to all (2012-2018), and filtered the data accordingly. One example:

```
crude_oil_dataset = crude_oil_dataset[(crude_oil_dataset['observation_date']
                                       >= '2012-01-01') & (crude_oil_dataset['observation_date'] <= '2018-12-31')]
```

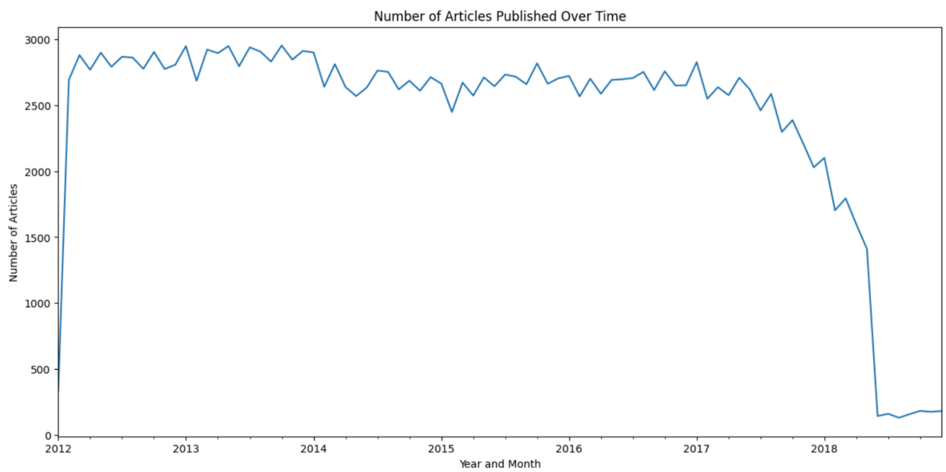
Data Conversion: Converting Date Formats
We are converting the 'date' column in the news_data DataFrame to a datetime object. This is typically done to enable time-series analysis or manipulation of the dates more easily within the DataFrame.

```
news_data['date'] = pd.to_datetime(news_data['date'])
news_data = news_data[(news_data['date'].dt.year >= 2012) & (news_data['date'].dt.year <= 2018)]
```

Remove stopwords, punctuations, and tokenize the columns

We have created a set of English stopwords for later use in filtering out these common words that often carry less meaningful content. We have also defined a function preprocess_text that takes a string input, removes punctuation, tokenizes the string into words, filters out stopwords and words that are not purely alphabetical, and then joins the remaining words back into a processed string.After that we have applied the preprocess_text function to two columns ('headline' and 'short_description') in the news_data DataFrame and stores the preprocessed text in new columns ('processed_headline' and 'processed_description').

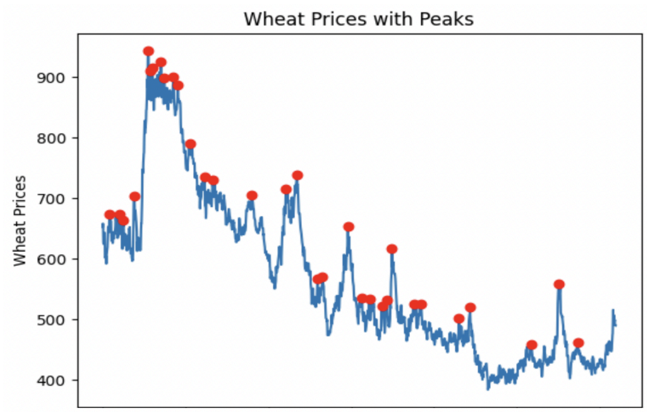
Following shows the trend of the number of articles accumulated. The plot shows the number of articles published each month over a period from 2012 to 2018. The data seems relatively stable until late 2017, where there's a noticeable decline, culminating in a sharp drop towards the end of 2018. This could indicate a data collection issue, a change in the publication's output, or an actual decrease in the number of articles published.



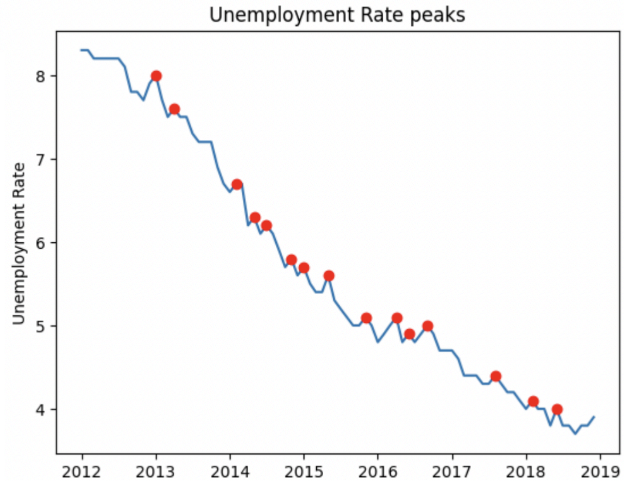
Here is the word cloud for the top 50 words with highest frequency from the dataset:

Analysis:

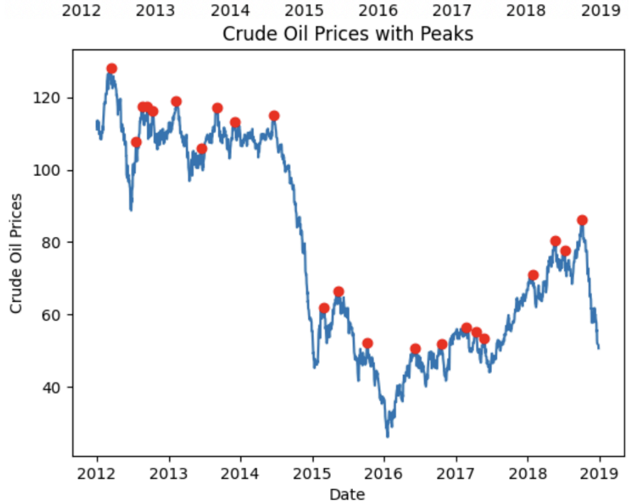
EDA on wheat, crude oil and unemployment data:



The wheat prices exhibited a consistent decline from 2012 to 2018, with a noticeable annual pattern. Notably, there is a recurring peak around the middle of each year. This annual surge can be linked to the harvesting season for wheat in the United States, which typically occurs in July.

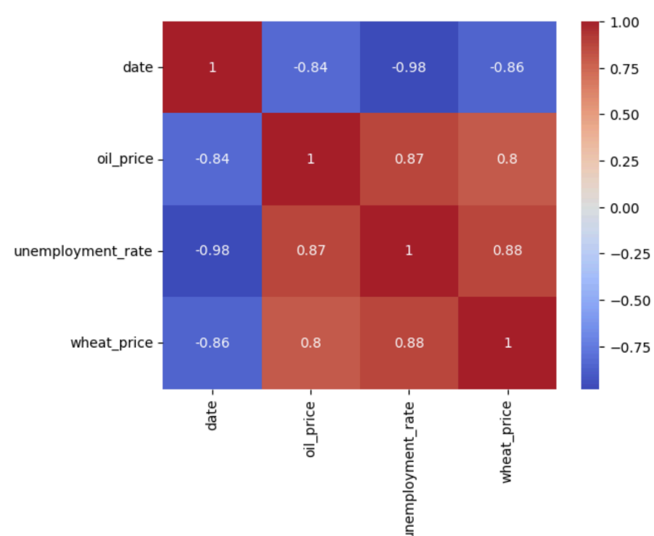


The unemployment rate exhibited a consistent decrease from 2012 to 2018, suggesting that initiatives to reduce unemployment might be effective. However, it is essential to refrain from definitively concluding their success, as other factors, like challenges in filing for unemployment, may influence these trends. Furthermore, there isn't a distinct month or specific time of the year with consistently high unemployment rates annually; rather, it appears to vary.



Crude oil prices exhibit a general steady rate, notably marked by a sharp downturn in 2014, and a steady increase after 2016. The substantial drop in 2014 was accelerated by a surplus of petroleum in the global market. Peak prices were observed in mid-June 2014, contrasting with the lowest recorded price in late January 2015 (source).

Correlation between these 3 economic indicators:
1. Decreasing Trends Over Time: The 'date' column strongly correlates negatively with unemployment_rate (-0.98) and wheat_price (-0.86), indicating a consistent decline, suggesting economic improvement or stability.
2. Oil Price Decline: Oil_price moderately correlates negatively with 'date' (-0.84), suggesting a general decrease in oil prices over time, influenced by factors like technology or global demand changes.



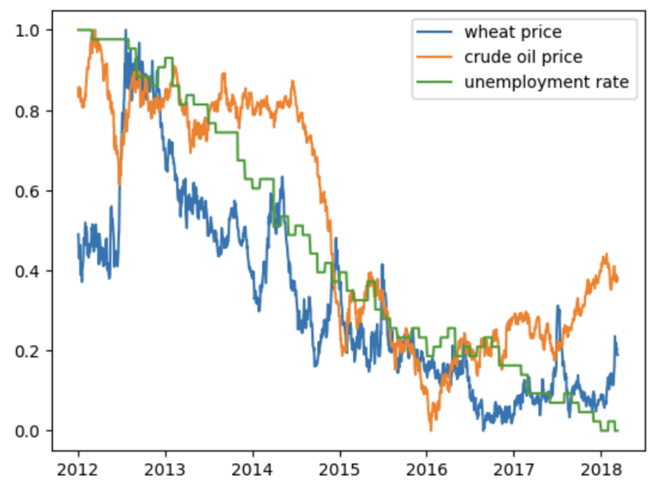
3. Oil Price and Employment: A positive correlation (0.87) between oil_price and unemployment_rate suggests potential impacts on employment due to higher oil prices, affecting production costs.
4. Oil Price and Wheat Cost Connection: Positive correlation (0.80) between oil_price and wheat_price indicates a positive relationship between energy prices and agricultural expenses.
5. Unemployment and Wheat Costs: The 0.88 correlation between unemployment_rate and wheat_price suggests a link, where economic downturns and higher unemployment may impact demand for commodities like wheat, affecting its price.

Spearman Rank-Order Correlations:
The plot on the right shows normalized values of the three datasets overlaid on the same time axis.

```
# do a spearman analysis on the data
from scipy.stats import spearmanr
print(spearmanr(combined_dataset['wheat_price'], combined_dataset['oil_price']))
print(spearmanr(combined_dataset['wheat_price'], combined_dataset['unemployment_rate']))
print(spearmanr(combined_dataset['oil_price'], combined_dataset['unemployment_rate']))
```

```
SignificanceResult(statistic=0.7884709898932397, pvalue=4.702649688605433e-11)
SignificanceResult(statistic=0.8460297403496967, pvalue=7.073540697392186e-14)
SignificanceResult(statistic=0.7370847797150382, pvalue=3.4655265924540187e-09)
```

1. Spearman Rank-Order Correlation between Wheat Price and Oil Price:
- Correlation Coefficient (statistic): 0.79, P-value: 4.702649688605433e-11
The correlation coefficient of 0.79 suggests a strong positive monotonic relationship between wheat price and oil price. The p-value is very small (close to zero), indicating that this correlation is statistically significant. In other words, there is evidence to suggest that as one variable increases, the other tends to increase as well, and vice versa.



2. Spearman Rank-Order Correlation between Wheat Price and Unemployment Rate:
- Correlation Coefficient (statistic): 0.85, P-value: 7.073540697392186e-14
The correlation coefficient of 0.85 indicates a strong positive monotonic relationship between wheat price and the unemployment rate. The p-value is very small, suggesting that this correlation is statistically significant. It implies that as wheat prices increase or decrease, there is a tendency for the unemployment rate to follow suit.

3. Spearman Rank-Order Correlation between Oil Price and Unemployment Rate:
- Correlation Coefficient (statistic): 0.74, P-value: 3.4655265924540187e-09
The correlation coefficient of 0.74 shows a strong positive monotonic relationship between oil price and the unemployment rate. The p-value is very small, indicating statistical significance. This suggests that as oil prices change, there tends to be a corresponding change in the unemployment rate.

In summary, these Spearman correlation results provide evidence of strong positive monotonic relationships between pairs of variables in our datasets. The p-values are very small, indicating that these correlations are unlikely to be due to random chance.

Analyzing the Influence of Macro-economic Data on Commodities:

- **Technique: Regression analysis.**
- **Explanation**: We investigated whether macro-economic data, such as unemployment rates, influences commodities like crude oil and wheat and vice versa. We performed regression analyses to determine the extent of this influence.

```
model0 = smf.ols(
    "Q('oil_price') ~ Q('unemployment_rate')",
    data=combined_dataset).fit()
model0.summary()
```

OLS Regression Results

Dep. Variable:	Q('oil_price')	**R-squared:**	0.763
Model:	OLS	**Adj. R-squared:**	0.758
Method:	Least Squares	**F-statistic:**	144.9
Date:	Tue, 05 Dec 2023	**Prob (F-statistic):**	1.14e-15
Time:	06:15:33	**Log-Likelihood:**	-191.18
No. Observations:	47	**AIC:**	386.4
Df Residuals:	45	**BIC:**	390.1
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-30.5606	9.112	-3.354	0.002	-48.913	-12.208
Q('unemployment_rate')	18.2538	1.516	12.039	0.000	15.200	21.308

Omnibus:	1.009	**Durbin-Watson:**	0.317
Prob(Omnibus):	0.604	**Jarque-Bera (JB):**	0.957
Skew:	0.162	**Prob(JB):**	0.620
Kurtosis:	2.380	**Cond. No.**	26.7

Regression model: oil price (dependent variable) and unemployment rate(independent variable):

The regression analysis indicates a significant relationship between crude oil prices and the unemployment rate (F-statistic: 144.9, p-value: 1.14e-15). The model, with an R-squared of 0.763, explains 76.3% of oil price variability based on unemployment rate variations. A one-unit increase in unemployment is associated with an estimated 18.2538-unit rise in oil prices. Both intercept and coefficient are statistically significant (p-values near zero). These findings suggest a strong positive link between oil prices and unemployment, offering valuable insights into economic and energy market dynamics. However, caution is advised in inferring causation solely from correlation and regression.

```
model1 = smf.ols(
    "Q('oil_price') ~ Q('wheat_price')",
    data=combined_dataset).fit()
model1.summary()
```

OLS Regression Results

Dep. Variable:	Q('oil_price')	**R-squared:**	0.691
Model:	OLS	**Adj. R-squared:**	0.684
Method:	Least Squares	**F-statistic:**	100.6
Date:	Tue, 05 Dec 2023	**Prob (F-statistic):**	4.75e-13
Time:	06:15:33	**Log-Likelihood:**	-197.43
No. Observations:	47	**AIC:**	398.9
Df Residuals:	45	**BIC:**	402.6
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-25.1038	10.381	-2.418	0.020	-46.011	-4.196
Q('wheat_price')	0.1809	0.018	10.029	0.000	0.145	0.217

Omnibus:	1.008	**Durbin-Watson:**	0.691
Prob(Omnibus):	0.604	**Jarque-Bera (JB):**	1.060
Skew:	0.273	**Prob(JB):**	0.588
Kurtosis:	2.506	**Cond. No.**	2.48e+03

Regression model: crude oil price (dependent variable) and wheat price (independent variable):

The regression analysis on crude oil and wheat prices reveals a significant relationship (Prob (F-statistic): 4.75e-13). An R-squared of 0.691 shows 69.1% of oil price variability linked to wheat prices. The highly significant wheat price coefficient (0.1809) suggests a corresponding 0.1809-unit oil price increase with a one-unit rise in wheat prices. While insightful, the analysis calls for attention to limitations and potential model refinement for enhanced explanatory power and robustness. Compared to the correlation values, this relationship holds true.

```
model2 = smf.ols(
    "Q('unemployment_rate') ~ Q('wheat_price') + Q('oil_price')",
    data=combined_dataset).fit()
model2.summary()
```

OLS Regression Results

Dep. Variable:	Q('unemployment_rate')	**R-squared:**	0.826
Model:	OLS	**Adj. R-squared:**	0.818
Method:	Least Squares	**F-statistic:**	104.4
Date:	Tue, 05 Dec 2023	**Prob (F-statistic):**	1.97e-17
Time:	06:15:34	**Log-Likelihood:**	-41.074
No. Observations:	47	**AIC:**	88.15
Df Residuals:	44	**BIC:**	93.70
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.3995	0.401	3.492	0.001	0.592	2.207
Q('wheat_price')	0.0047	0.001	3.987	0.000	0.002	0.007
Q('oil_price')	0.0239	0.005	4.408	0.000	0.013	0.035

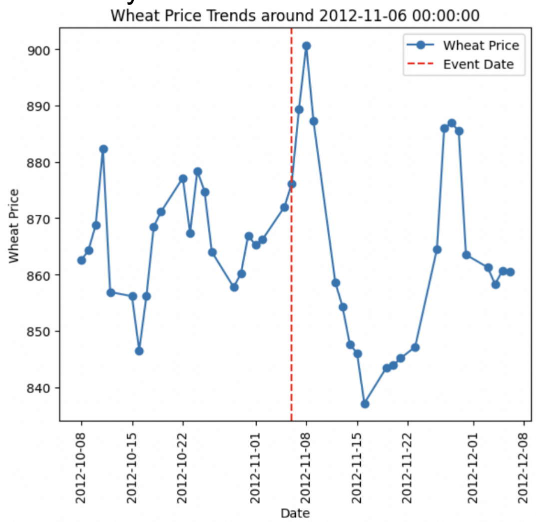
Omnibus:	0.483	**Durbin-Watson:**	0.433
Prob(Omnibus):	0.785	**Jarque-Bera (JB):**	0.071
Skew:	0.052	**Prob(JB):**	0.965
Kurtosis:	3.160	**Cond. No.**	2.66e+03

Regression: unemployment (dependent variable) and wheat and oil (independent variable):

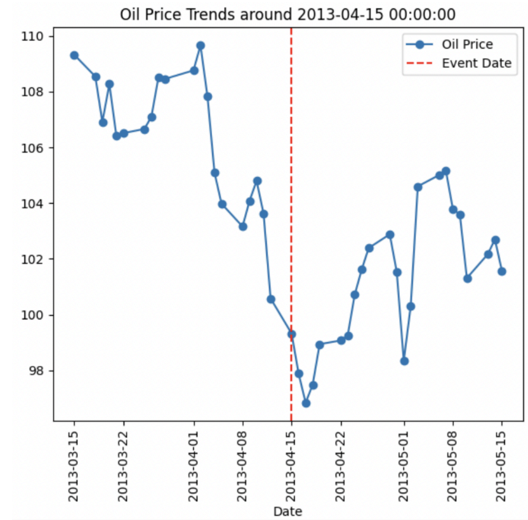
The analysis reveals compelling insights into the factors influencing the unemployment rate. The model demonstrates high statistical significance (F-statistic: 104.4, p-value: 1.97e-17), indicating that at least one of the independent variables significantly contributes to explaining the variation in the unemployment rate. The R-squared value of 0.826 highlights the model's robust explanatory power, that approximately 82.6% of the variability in the unemployment rate can be attributed to variations in wheat prices and crude oil prices. Both wheat price (0.0047) and crude oil price (0.0239) exhibit statistically significant positive coefficients, implying that an increase in either wheat prices or crude oil prices is associated with a corresponding increase in the unemployment rate.

Key News Events and Their Impact on Crude Oil and Wheat Prices:

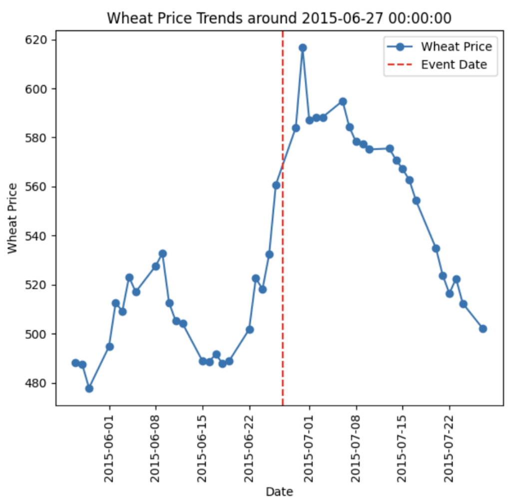
- **Technique**: Manual text mining, visualizations.
- **Explanation**: We took major events in the US from 2012 to 2018 (source) and aimed to understand which kind of key news events influenced the crude oil, and wheat prices and to what degree.



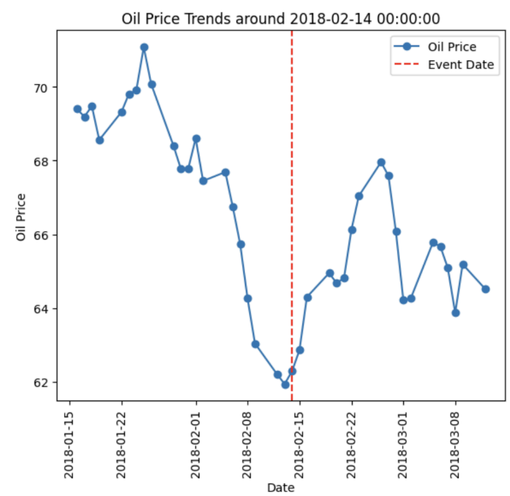
US elections - increased economic stability - indirect impact



Boston bombing-destabilize economy, oil prices increase



Wheat harvest in July, cost goes down

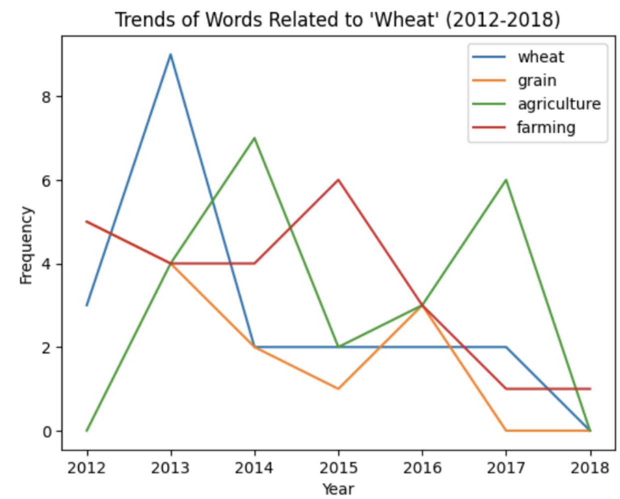


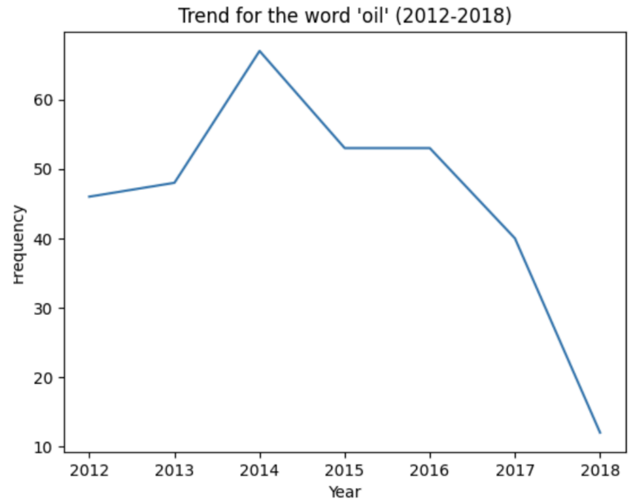
Marjorie Douglas Shooting-destabilize economy, oil prices increase

EDA : Analyze the frequency of keywords over time.

Evolving Trends in Large-Scale Text Data:

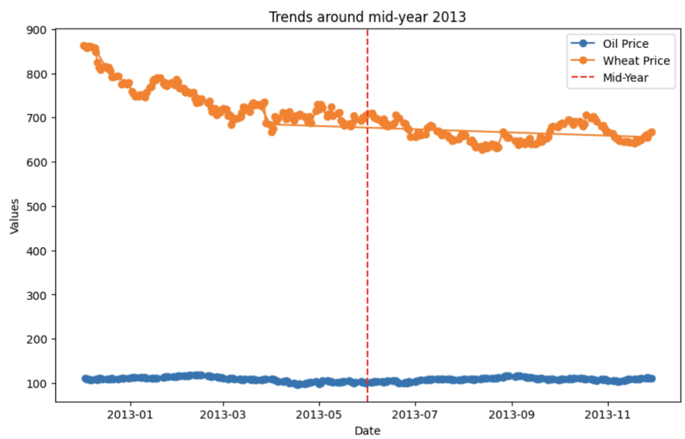
- **Technique**: Text mining, time series analysis.
- **Explanation**: We will analyze the evolution of trends, keywords, or phrases from the large-scale text data (news articles) over time. By using text mining and time series analysis, we expect to identify how certain topics or themes gain or lose relevance over the years.

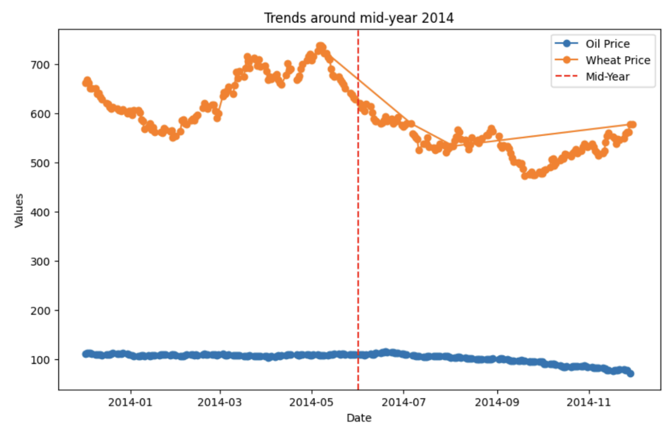


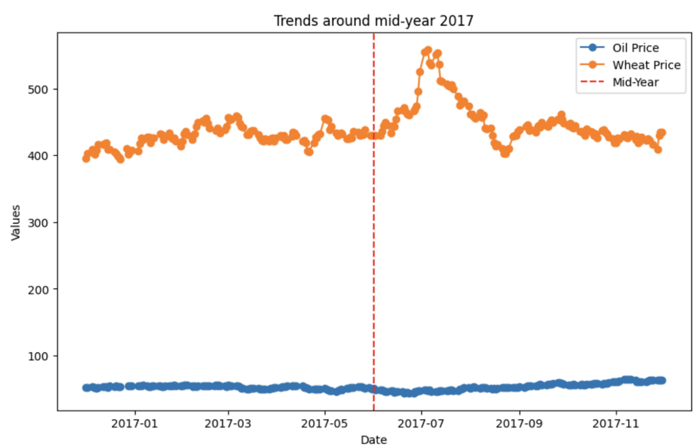


As we can see from the plots, wheat had major occurrences of wheat related words and oil in the year 2013, 2014, 2017.

Now comparing the results for those years. And further analyze the correlation between wheat-related keyword frequencies with news trends and economic indicators.







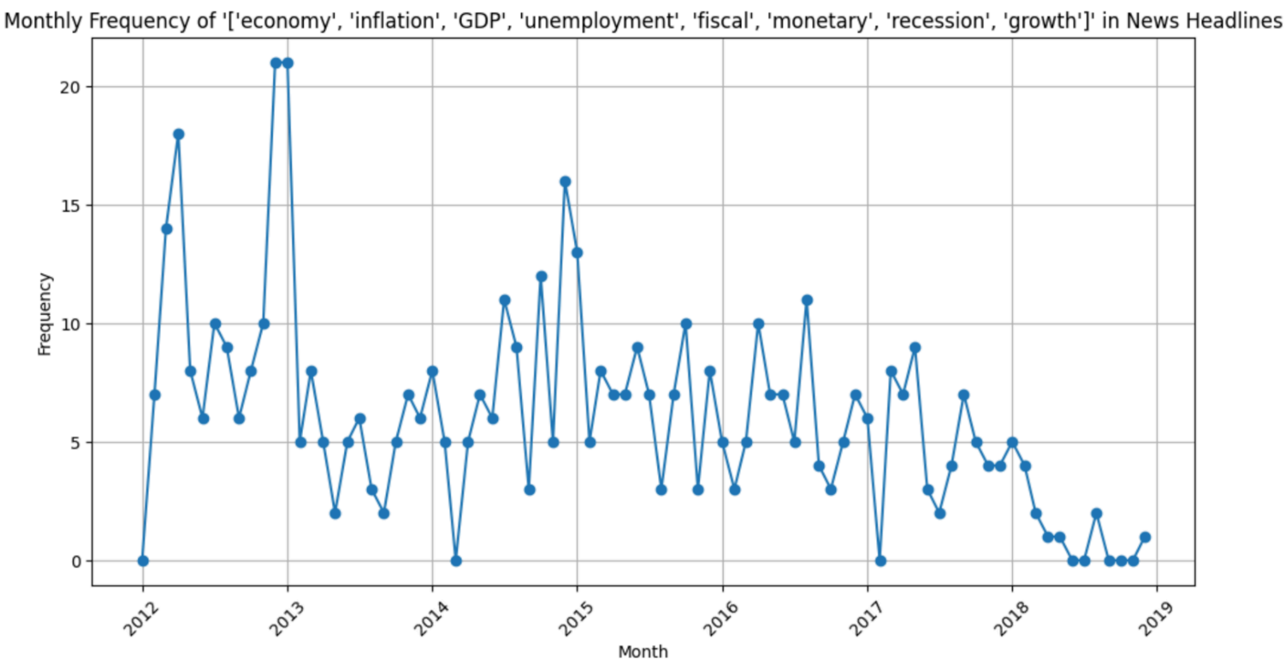
We observe that the influence of economic factors in news trends doesn't have much influence on the actual prices of wheat and oil as seen in the above plots.

Showcasing the Relationship Between News Trends and Macro-economic Indicators:

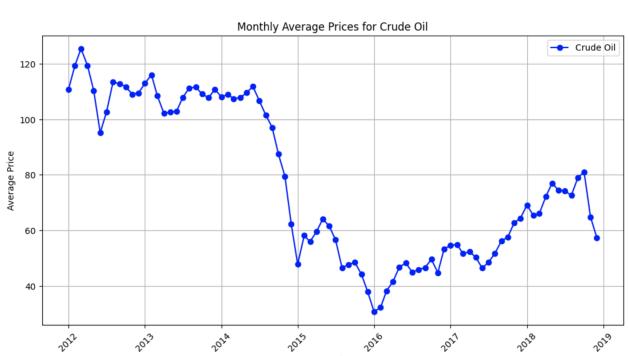
- **Technique**: Visualization, time series analysis.
- **Explanation**: To present the relationship between news trends and macro-economic indicators, we will create visualizations. Time series analysis will allow us to highlight how these trends correlate over time. This will provide a visual narrative of the data's dynamics.

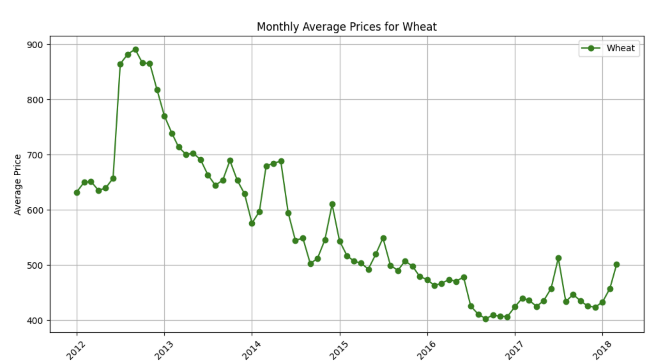
Code steps:
Firstly, code for filtering datasets based on a date range. It defines a start and an end date and then filters two datasets, crude_oil_dataset and wheat_dataset, to include only the data within this date range.
Then, DataFrame called news_data with 'date' and 'headline' columns is manipulated. Dates are converted to datetime objects, a list of economy-related keywords is defined, and the occurrences of these keywords in headlines are counted and grouped by month.
Then, code that converts a 'observation_date' column of the crude_oil_data_filtered DataFrame to datetime format, then groups the data by month to calculate the monthly average price, and finally converts the time series data into a DataFrame for further analysis with renamed columns.
Next for a wheat_data_filtered DataFrame. It converts the 'date' column to datetime, calculates monthly average prices, and again converts the time series into a DataFrame with renamed columns for further analysis.

Analysis:



The first graph shows fluctuations in the frequency of keywords related to the economy, such as 'economy', 'inflation', 'GDP', 'unemployment', 'fiscal', 'monetary', 'recession', and 'growth' in news headlines from 2012 to the beginning of 2019. There are peaks and troughs indicating varying levels of news coverage on these topics over time.





The second graph displays the monthly average prices for wheat over the same time period. There's a sharp peak around 2012, followed by a general decline and then some volatility but in a lower price range. This could indicate market reactions to economic conditions, possibly influenced by the events discussed in the news.
The third graph tracks the monthly average prices for crude oil. It also shows a general decline post-2012, with significant volatility. The price of crude oil is often subject to many factors, including economic news, geopolitical events, and supply and demand dynamics.
To determine the influence of news on wheat and oil prices, one would typically conduct a more in-depth statistical analysis,

Relationship between oil-related words and crude oil prices:

```
The Pearson correlation coefficient between the frequency of oil-related words and crude oil prices is: 0.01651940603731362
```

The Pearson correlation coefficient of 0.0165 suggests an extremely weak or negligible linear relationship between the frequency of oil-related words in news headlines and crude oil prices. In simpler terms, changes in the frequency of these words do not consistently or significantly affect crude oil prices. The positive sign of the coefficient indicates a weak positive relationship, meaning that when the frequency of oil-related words increases, crude oil prices tend to increase slightly. However, this relationship is so weak that it's practically insignificant. In summary, this coefficient indicates that the frequency of these words in news headlines is not a reliable predictor of crude oil price movements; other factors have a more significant impact on oil prices.

Ethical Considerations:

The use of news features and commodity pricing data, particularly when combined, does raise some ethical concerns:

1. Potential for Misinterpretation: The correlation between news features and commodity price fluctuations does not imply causation. Misinterpretation can lead to wrong decisions, and stakeholders might be misled if they assume that certain news features directly cause price changes.
Mitigation: we have clearly emphasized in all visualizations, and analyses that correlations do not imply causation. Ensured thorough statistical testing and provided contextual explanations for observed trends.
2. Data Privacy and Source Anonymity: Using news features might touch on sensitive topics or indirectly reference specific events, entities, or individuals. It's essential to respect the anonymity and privacy of sources.
Mitigation: We ensured that data is anonymized and that any sensitive or personal information is removed or obfuscated. Always credited data sources appropriately and avoided using unauthorized or copyrighted content without permission.
3. Bias in News Reporting: News articles and reports may contain inherent biases based on the outlet, region, or reporter. Relying on biased news can skew the analysis.
Mitigation: We are utilizing a diverse range of news articles to capture a more comprehensive and unbiased view of events.

Conclusion:

In conclusion, our project serves as a comprehensive exploration into the intricate relationships between real-world events, as captured in news articles, and economic indicators. The motivation behind this endeavor lies in the recognition that understanding these relationships is essential for making informed decisions across diverse sectors, ranging from economic forecasting to strategic business planning and trade policy analysis. By leveraging large-scale text data alongside commodity prices and macroeconomic data, our interdisciplinary approach aims to uncover nuanced patterns, correlations, and trends that go beyond mere numerical associations. Through extensive data manipulation, correlation analyses, regression modeling, and ethical considerations, we have embarked on a journey to provide valuable insights into the dynamic interplay between global events and economic landscapes. As we progressed through our analyses and visualizations, we gained a deeper understanding of how news influences economic factors, and came to the conclusion that the indicators have a strong correlation amongst themselves but the news reported often does not reflect the true situation of the economy.

Statement of Work:

Motivation	Era and Sheza
Data Sources	Sheza
Data manipulation methods	Era (text data) and Sheza (commodity datasets)
Analysis	Sheza (Q1 and Q2) and Era (Q3 and Q4)
EDA	Sheza (commodity datasets) and Era (text data)
Ethics	Sheza
Conclusion	Sheza
Statement of work	Era and Sheza