Kayla Lee 301393747
Jacob He 301391374

CMPT 353 Final Project: Is there a Relationship Between the Price of Gas and the Stock Market?

1. Introduction

Our project explores the cost of retail gas in the US and its relationship with the overall stock market, as well as different stock market indexes. We also investigate the different features that affect the price of gas, such as stock indexes, the cost of crude oil (2023), and the demand for gas in the US (Wagner, 2022), and create a model to predict the price of gas depending on those features and determine how influential each feature is in the price of gas. These steps help us answer our question: **Is there a relationship between the cost of gas and the stock market?**

We hope our conclusions from this analysis can be applied in the real-world for further analysis of the trends in gas prices and further research into the relationship between the stock market and the price of gas.

2. Acquiring and Cleaning the Data

The data used in this project was retrieved from Kaggle, websites that track stock indexes (such as the S&P Dow Jones), and government websites. We began our analysis with the general stock market, so we chose to compare the price of gas to the S&P 500 stock index since it can be used as a gauge for the overall stock market's performance (Brock, 2023). Hence, in this report we will use the S&P 500 stock index as an indicator of the general stock market. When deciding which currency to use for our data, we chose data reported in USD since the stock indexes are reported in USD so the values would be a direct comparison without needing to convert currency. The units of measurement for the cost of gas is dollars per gallon (USD), the units of measurement for the cost of crude oil is dollars per barrel (USD), and the units of measurement for the demand of gas in the US is barrels of retail gas supplied.

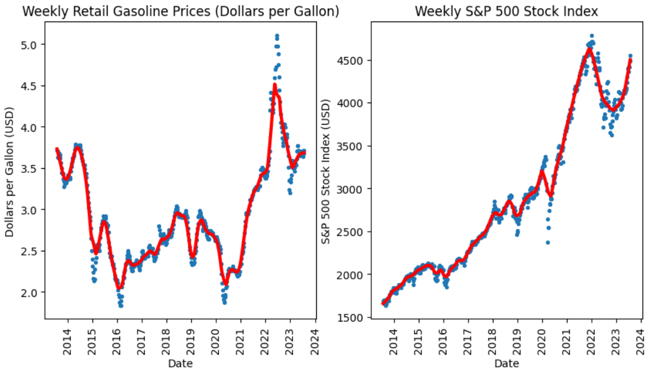
Since some of our data was collected weekly (such as the cost of gas) and some data was collected daily (such as the S&P stock index), we performed resampling of the daily data to group the daily data into weekly averages. Since the cost of gas data was sampled every Monday, we resampled our daily data on Mondays as well to yield a fair comparison between the different indexes and features.

The different data collected did not share the same collection start and end dates, as gas prices were sampled from the 1990s-present, while the S&P stock index was only collected 2013-present. Thus we converted all the dates to a datetime object, and excluded the rows that dated prior to July 2013, to make all the dates align. We followed the same procedure for other dataframes we created, including the other stock indexes and features that influence the cost of gas. Finally we merged all the index data and cost of gas data columns into one dataframe to start analysis. Lowess Smoothing was used to visualize the outliers and apply smoothing to the individual cost of gas data and S&P 500 stock index data.

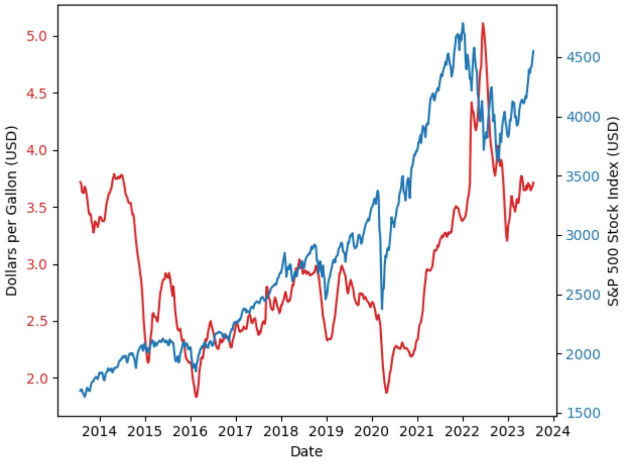
To do ETL (Extract, Transform, Load), we wrote a python program, etl.py, to handle data from a CSV file or Excel file. We take 3 parameters: an input file path to our input file, and two numbers representing the row numbers in the CSV file that need to be excluded from the processing, such as descriptions in the beginning of the data files. Rows above and below these numbers will be skipped. The code can also be modified to drop specific columns in the data. The output CSV file will have the same format as the input, but with certain rows excluded and optionally, columns removed.

3. Data Analysis

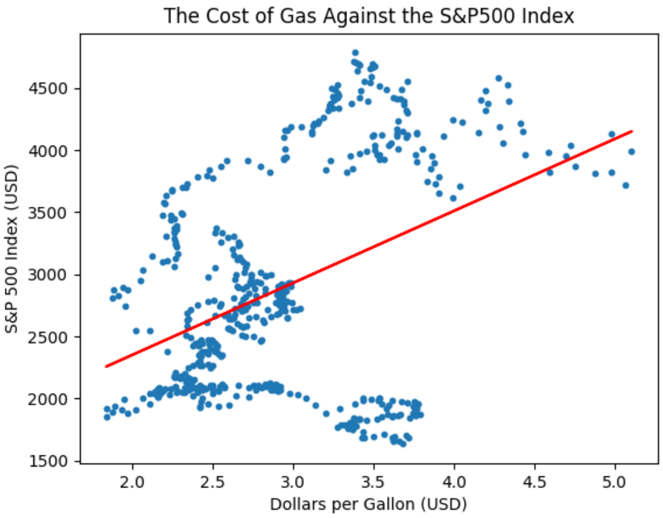
a. The relationship between the cost of gas and the S&P 500



Since we decided to use the S&P 500 stock index as an indicator for the overall stock market, we graphed both the gas and S&P 500 data individually to visualize the overall shape and spread of each column.



We then plotted both the gas and S&P data on the same graph to better visualize how their data points compare throughout the years.

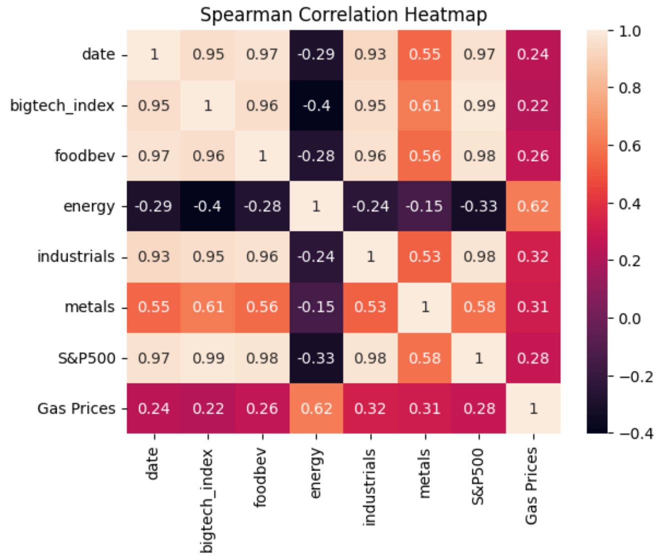


To determine whether the two variables had a linear relationship, we graphed the S&P 500 index (y values) against the cost of gas (x values).

In The Cost of Gas Against the S&P 500 Index Graph, a linear best fit line was added to identify trends and visualize the residuals, how vertically far each data point is from the best fit line. The data points did not seem to follow any pattern or increase linearly, so we concluded that the data is **non-linear**. We then calculated the Spearman Correlation Coefficient, which can capture monotonic relationships between non-linear variables, to be 0.27, indicating a low correlation between the price of gas and the S&P 500 index. Thinking there may be a delayed correlation between the variables, we shifted the S&P data by multiple weeks against the original gas data and calculated each new correlation value. We calculated this for up to 4 weeks, however all calculations stayed within 0.2-0.3. We shifted the gas data and performed the calculations accordingly again, however we still only calculated correlations of 0.2-0.3. Thus, we can conclude that there is a very low correlation between the cost of gas and the S&P 500 index. Since we used the S&P 500 index as an indicator for the overall stock market, we can also conclude that there is a **very low correlation between the cost of gas and the overall stock market.**

b. The relationship between the cost of gas and individual stock indexes

The Spearman correlation heatmap between Gas and the S&P index indicated weak correlation (correlation coefficient of 0.27). As the initial analysis showed no significant correlation between Gas and the overall stock market, we decided to study the correlations with other individual stock market indexes, including Hi-Tech, Food, Industry, Metals, and Energy.



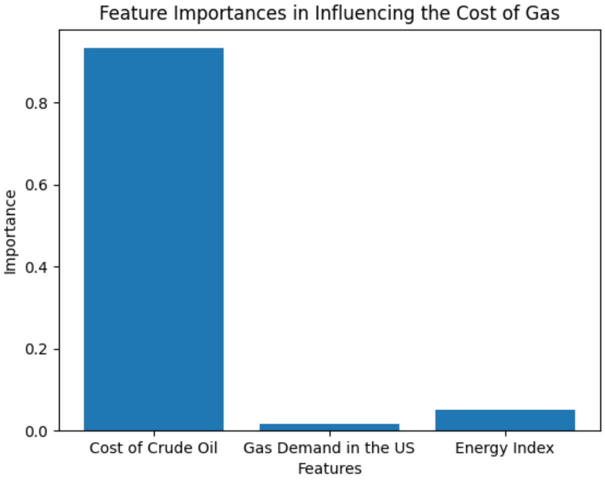
The subsequent Spearman correlation analysis revealed the following correlations between the cost of gas and various indexes:

The index that showed the highest correlation with the cost of gas was the energy index, with a correlation coefficient of 0.62. No other index exhibited a correlation remotely close to this value. The strong correlation coefficient of 0.62 suggests that Gas and Energy are positively related. This finding is reasonable as Gasoline is an energy-related product, and fluctuations in energy prices can significantly influence Gas prices. However, energy is only a small fraction of the overall stock market, so it hardly affects the overall trend.

It's worth noting that some of the indexes showed high correlations with each other, but for the purpose of this project we focused solely on the correlation between the price of gas and the various indexes.

c. How influential is the energy stock index in determining the cost of gas?

After determining that the energy index had a correlation to the cost of gas, we created a model to see how influential the energy index was compared to other features in influencing the price of gas. After splitting the data into training and validation data, we chose to use the Random Forest Regressor model due to its ability to make decisions based on features, and provide information of feature importances. The Random Forest Regressor, which creates multiple decision trees that all vote on what they predict the price of gas to be. Our model had a training score of 0.96 and a validation score of 0.89. After also calculating the feature importances, we saw that the cost of crude oil had the greatest influence on the price of gas, while the energy index did not have as much importance. From this, we can conclude that while the energy index is correlated to the price of gas, it is not highly influential that the model used in determining the price of gas.



4. **Limitations**

    a. **Limitation of Historical Data**

It's understandable that having only 10 years of available stock index and S&P data can be a limitation. The stock market is influenced by various factors, and relying solely on a decade of data might not provide a comprehensive understanding of long-term trends and patterns, thus limiting our conclusions, especially when the past decade has been volatile and may not represent a reliable indicator for future stock trends.

    b. **Feature Dependence on Future Data**

For the model's gas price prediction, it requires features that are dependent on future data, such as the cost of crude oil and stock index data. Since predicting future values of features is not currently feasible due to data unavailability, our model is unable to make accurate predictions for future gas prices. With more time and research, other models could be created to predict the price of crude oil and other features, which could then be used as data for our model.

c. Sparse Data points

Due to publicly available data having sparse data points, only once a week for gas, we were unable to zoom in and identify specific trends that could take place within a single week. Because stock markets are often volatile, there are potentially a lot of economic trends that only last for a short duration that we are missing, reducing the accuracy of our analysis.

5. **Conclusion**

From our analysis of the relationships between the cost of gas and the overall stock market (through the S&P 500 stock) and individual stock market indexes, we have concluded that the overall stock market has a low correlation to the cost of gas, however there is a strong correlation between the energy stock index and the cost of gas. However, the feature importance values that were generated from our Random Forest Regressor model proved that the energy index is not influential in determining the price of gas compared to other features that influence the price of gas.

6. Project Experience Summary

Jacob:

- Found appropriate sources of data for various stock indexes (metal, food, etc)
- Implemented Extract Transform Load to refine raw data from 7 different sheets of varying formats and row configurations into an easily workable format.
- Computed the Spearman correlation analysis to create a correlation heatmap of all stock indexes and the price of gas.

Kayla:

- Wrote a script to extract and clean data from the gas and S&P 500 data files.
- Performed statistical analysis and created graphs to determine trends of the gas and S&P data and calculate the correlation accordingly for different time shifts.
- Created a Random Tree Regressor model to yield the feature importance traits to determine the most influential features that determine the price of gas.

7. Appendix

Citations:

Brock, T. (Ed.). (2023, July 29). *What does the S&P 500 index measure and how is it calculated?*. Investopedia. https://www.investopedia.com/ask/answers/040215/what-does-sp-500-index-measure-and-how-it-calculated.asp

US Energy Information Administration. (2023, February 22). *U.S. Energy Information Administration - EIA - independent statistics and analysis*. Factors affecting gasoline prices - U.S. Energy Information Administration (EIA). https://www.eia.gov/energyexplained/gasoline/factors-affecting-gasoline-prices.php

Wagner, H. (2022, November 10). *What determines gas prices?*. Investopedia. https://www.investopedia.com/articles/economics/08/gas-prices.asp

Data downloaded from:

Gas:
https://www.eia.gov/dnav/pet/hist/LeafHandler.ashx?n=pet&s=emm_epm0_pte_nus_dpg&f=m
S&P 500:
https://www.kaggle.com/datasets/andrewmvd/sp-500-stocks?resource=download
Energy:
https://www.spglobal.com/spdji/en/indices/equity/sp-500-energy-sector/

Industrials:
https://www.spglobal.com/spdji/en/indices/equity/sp-500-industrials-sector/#overview
Food/Bev:
https://www.spglobal.com/spdji/en/indices/equity/sp-food-beverage-select-industry-index/
Big Tech:
https://www.nasdaq.com/market-activity/index/ndxt/historical
Precious Metals:
https://www.spglobal.com/spdji/en/indices/equity/dow-jones-precious-metals-index/#overview
Crude Oil:
https://www.macrotrends.net/2516/wti-crude-oil-prices-10-year-daily-chart#google_vignette
Gas Demand:
https://www.eia.gov/dnav/pet/hist/LeafHandler.ashx?n=PET&s=wgfupus2&f=W