From Bump to Pump: Crude Oil, Gasoline Pricing, and the Financial Ripple Effects Across Energy and Automaker Sectors

Filya Geikyan, Rhea Bose, TsaiChen Lo, Zejia Yang

May 18, 2025

1 Project Statement

Gasoline is the most widely used fuel in the United States and represents the largest output of domestic oil refineries. As the world transitions toward cleaner energy, understanding the dynamics of gasoline pricing remains essential, not only for energy economics, but also for understanding consumer behavior and future developments in transportation.

2 Non-Technical Executive Summary

Research Questions

In this report, we set out to understand the pricing dynamics of gasoline and its effects on various stocks in the energy markets.

Gasoline prices are shaped by a complex supply chain—from crude extraction (bump) to consumer access (pump)—and understanding this pipeline helps identify the key drivers behind price fluctuations. With available data, we can examine whether gasoline prices are more sensitive to upstream factors like crude oil prices, midstream influences such as refining technologies and policy (e.g., taxes), or downstream elements like marketing and distribution.

Gasoline prices also serve as a strong indicator of economic conditions and directly impact companies across the energy chain, as reflected in their stock fluctuations. By better predicting these dynamics and capturing their nuances, investors can more effectively manage portfolios, and policymakers can design targeted interventions. This research helps bridge market behavior, production economics, and financial performance in an industry that shapes both economic trends and sustainability efforts.

The overarching question is:

> **How Do Crude Oil Prices and Gasoline Price Makeup Percentage Shape Fuel Prices and Influence Stock Performance Across Energy Streams and Automakers?**

More specifically, this can be divided into several sub-questions.

- **Crude-to-Gasoline Pricing Dynamics**

1. How do fluctuations in crude oil prices and the makeup percentage of gasoline costs (refining, distribution, taxes, etc.) impact US regular gasoline prices?
2. Can we build models to predict future gasoline prices based on its historical data and all the data above?
3. How do factors influencing gasoline prices vary geographically across different regions of the US?

- **Ripple Effects Across the Energy Supply Chain**

  1. How do changes in gasoline and crude oil prices affect the stock performance of energy companies?
  2. Do companies in different parts of the supply chain—upstream (extraction), midstream (transportation), and downstream (refining and retail)—respond differently to changes in fuel prices?
  3. Can we group or cluster energy companies based on how closely their stock prices move with fuel prices, and do these groups reflect their actual business roles?

- **Impact on Automakers and the Shift Toward Electric Vehicles**

  1. How do gasoline price changes affect the stock prices of traditional automakers versus electric vehicle manufacturers?

Findings

- **Crude-to-Gasoline Pricing Dynamics**

  1. Crude oil prices and tax rates are the most significant drivers of US regular gasoline price fluctuations. Changes in distribution and marketing costs also play a notable role, while refining and crude oil makeup percentages have relatively lower predictive power.
  2. Feature importance clustering reveals regional differences in sensitivity to gasoline prices, shaped by variations in regulations, tax regimes, and infrastructure.

- **Ripple Effects Across the Energy Supply Chain**

  1. Upstream firms exhibit stronger co-movement with crude oil prices than gasoline prices, reflecting their direct exposure to commodity markets.
  2. Causal sensitivity is not confined to upstream firms, some downstream and integrated firms also exhibit short term predictive relationships with crude oil prices.

- **Impact on Automakers and the Shift Toward Electric Vehicles** EV-sector returns derive more signal from gasoline dynamics than crude, while traditional auto returns are more sensitive to crude-oil shocks

3 Technical Exposition

3.1 Crude-to-Gasoline Pricing Dynamics

Gasoline pricing is influenced by complex temporal and spatial dynamics. We conducted an in-depth analysis to identify key drivers of gasoline prices, examining the roles of crude oil prices and component costs at different streams. Building on these findings, we aim to provide a basis for analyzing the performance of companies involved in different stages of gasoline production and marketing.

Data Processing and Feature Selection

To examine how fluctuations in crude oil prices and the components of gasoline pricing influence US retail gasoline prices, we began by integrating multiple data sources into a unified dataset through data wrangling and feature engineering. Table 1 outlines the key variables extracted from the original datasets, including their frequency units and descriptions.

Table Name	**Field**	**Frequency**	**Description**
Weekly Gasoline Prices	`Price`	Weekly	National average US retail gasoline prices; includes regular conventional, reformulated, and all formulations.
Monthly Gasoline Makeup Percentages	`Retail_Price`	Monthly	Total monthly average US retail gasoline price from the cost breakdown data set.
Commodities	`Value`	Daily	Daily prices of crude oil (WTI and Brent); aggregated to weekly averages.
Monthly Gasoline Makeup Percentages	`Crude Oil`	Monthly	Percentage of retail gasoline price attributed to crude oil cost.
Monthly Gasoline Makeup Percentages	`Refining`	Monthly	Percentage of retail gasoline price attributed to refining cost.
Monthly Gasoline Makeup Percentages	`Distribution_and_Marketing`	Monthly	Percentage of retail gasoline price attributed to distribution and marketing.
Monthly Gasoline Makeup Percentages	`Taxes`	Monthly	Percentage of retail gasoline price attributed to taxes.

Table 1: Summary of Data Fields Used from Original data sets

The first two rows, `Price` and `Retail_Price`, serve as the dependent variables in our following analysis. Given the time series property of the dataset, where all variables can be treated as continuous, it is reasonable to assume that current values are strongly influenced by their preceding values. Therefore, instead of modeling absolute price levels, we focus on the rate of change.

To capture this, we compute the log return for each variable, defined by the equation below. Compared to simple percentage change, log return offers several advantages: it is time-additive, better captures compounding effects, and provides a more accurate reflection of long-term trends.

$$\text{Log Return:} \quad r_t = \ln\left(\frac{P_t}{P_{t-1}}\right) = \ln(P_t) - \ln(P_{t-1})$$

Given the different frequencies of the original datasets (i.e., daily, weekly, and monthly), it was necessary to apply appropriate aggregation and interpolation techniques to align all variables onto a

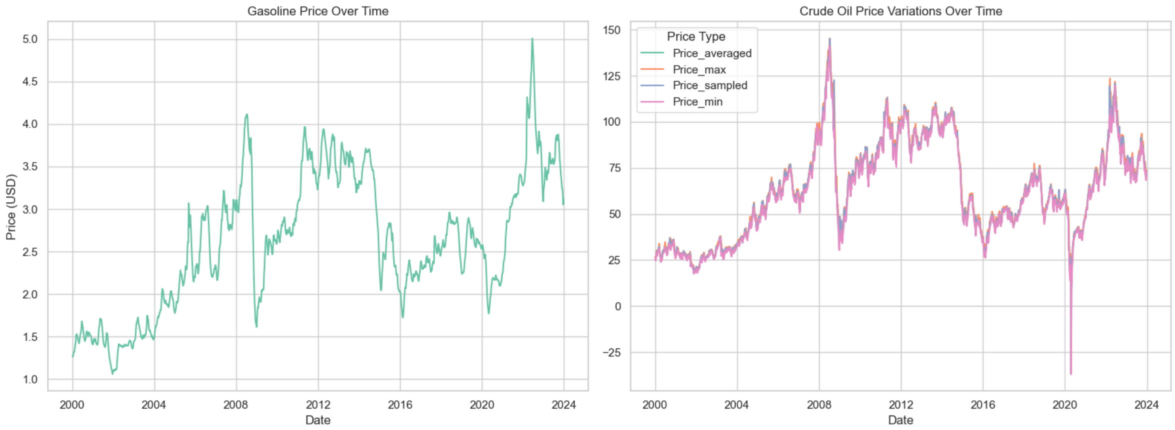
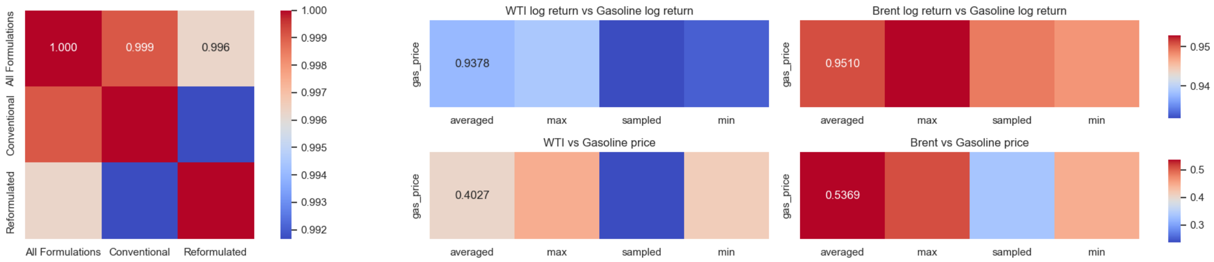


Figure 1: Gasoline and Crude Oil (Brent) Price over Time

common time index. We separately examined both weekly and monthly effects to capture short-term and long-term dynamics, which may reflect the immediate or lagged impact of certain factors.

Weekly Analysis. For the week-wise analysis, we selected `Price` from the Weekly Gasoline Prices dataset as the dependent variable, as it is inherently reported on a weekly basis. As shown in Figure 2a, among the three regular gasoline price categories—Conventional, Reformulated, and All Formulated—All Formulated exhibits the strongest correlation with the other two. This suggests that it is the most representative measure of overall regular gasoline price trends, which naturally holds as All Formulated typically represents the aggregate gasoline market, including all types sold. Therefore, it was chosen as the dependent variable for our weekly model.

For the independent variable `Value` from the Commodities dataset representing crude oil prices, we derived weekly features by aggregating the daily data. Specifically, we computed the weekly average, minimum, maximum, and a randomly sampled daily price within each week. These derived statistics are shown in Figure 1. Their prices exhibit highly similar fluctuations, giving us confidence that there is a correlation—or even causality—between the two prices. We conducted a correlation test to select the best aggregation method. As illustrated in Figure 2b, Brent crude oil prices show a stronger correlation than WTI, which contrasts with common belief, as WTI reflects more US-specific characteristics while Brent is more globally representative.



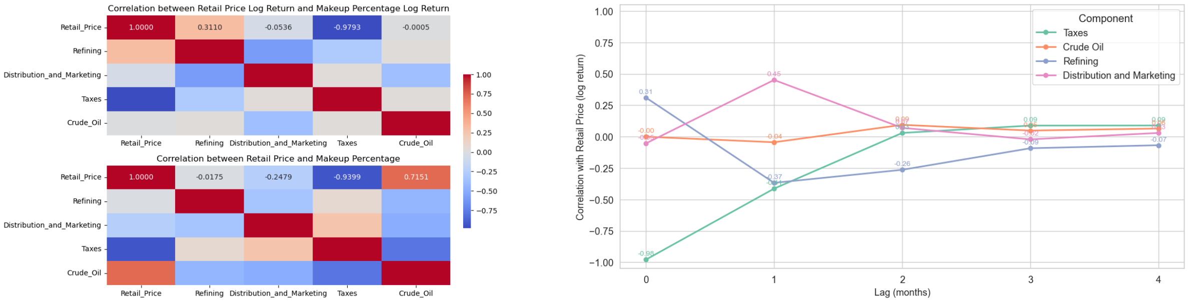
(a) Correlation among Gasoline Price Types

(b) Correlation between Gasoline and Crude Oil Prices

Figure 2: Gasoline Price Correlations with Itself and Crude Oil

Monthly Analysis. For the monthly analysis, we selected `Retail_Price` from the Monthly Gasoline Makeup Percentages dataset as the dependent variable. We examined its relationship with various components of the gasoline price makeup (e.g., `Refining`, `Taxes`, etc.) using both their absolute values and log returns. Monthly approach emphasizes long-term trends and enables us to investigate structural factors such as tax policy, market conditions, and improvements in refining efficiency.

From the correlation heatmap fig 3a, we observe a strong negative correlation between `Taxes` and `Retail_Price`, both in terms of absolute values and log returns. Interestingly, `Crude_Oil` shows little to no significant correlation in its log return form. This can be partially attributed to the monthly aggregation of crude oil data. From their daily change in Commodities, crude oil prices exhibit substantial short-term volatility, and their impact on gasoline prices is typically immediate. However, monthly resampling smooths out this volatility, making crude oil less predictive in a monthly context.



(a) Correlation between Retail Price (log return) and Makeup Percentage (log return)

(b) Correlation between Lagged Retail Price (log return) and Makeup Percentage (log return)

Figure 3: Gasoline Price Correlations with Makeup Percentages

To further investigate the **lag effects** of various components on `Retail_Price_logreturn`, we computed correlations between the log return of each component and the future log return of `Retail_Price` at different time lags. As illustrated in Figure 3b, we find that: `Taxes` exhibit a lag effect of 0 to 1 month, `Refining` shows influence over a 0 to 2-month window.

Given these insights, we prioritize the **weekly modeling approach** for its ability to capture more dynamic and responsive short-term fluctuations in gasoline pricing.

Modelling and Prediction

In this section, we aim to predict weekly fluctuations in gasoline prices using features derived from both crude oil pricing (specifically Brent crude) and gasoline price makeup components. All data are aligned to a weekly frequency. For the monthly gasoline makeup percentages, we applied linear interpolation to generate weekly values. We organize the input features into three categories based on their type, shown 2.

Feature Type	**Features**
Value	Gasoline_Price, Brent_Price
Percentage	Crude_Oil, Refining, Taxes, Distribution_and_Marketing
Log Returns	Gasoline_Price_logreturn, Brent_Price_logreturn, Crude_Oil_logreturn, Refining_logreturn, Taxes_logreturn, Distribution_and_Marketing_logreturn

Table 2: Feature Classification for Weekly Gasoline Price Modeling

To model gasoline price changes, we use **XGBoost** (Extreme Gradient Boosting), a highly efficient and scalable implementation of gradient-boosted decision trees. XGBoost is particularly suitable for this task due to the following reasons:

- It effectively models complex, non-linear interactions between variables.
- It provides built-in mechanisms for regularization, preventing overfitting.
- It outputs interpretable **feature importance**, allowing us to evaluate which predictors are most influential.

We trained an XGBoost regression model to predict the log return of weekly gasoline prices using the log returns of several independent variables as features. Prior to training, we standardized all feature variables to have a mean of 0 and a standard deviation of 1.

This **z-score normalization** was necessary to ensure that all features were on a comparable scale, especially since log returns tend to be small in magnitude. While normalization is not required for correlation analysis we did before—since the test itself accounts for scaling—it can improve interpretability of the metric Mean Squared Error (MSE), since we don't have a baseline.

After training the model, we assessed its performance on a hold-out test set, achieving an MSE of 0.4668. Given that the target variable was standardized, this can be interpreted as an average prediction error of approximately 0.68 standard deviations from the true value. Finally, we extracted feature importances from the trained model to identify which factors most significantly drive short-term fluctuations in gasoline prices.

Rank	**Feature (log return)**	**Importance**
1	Taxes	0.54428
2	Brent_Price	0.16922
3	Distribution_and_Marketing	0.15251
4	Refining	0.07422
5	Crude_Oil	0.05977

Table 3: Gasoline Prices Feature Importance (log return, normalized) from XGBoost

It reveals that weekly fluctuations in gasoline prices are primarily driven by changes in **tax rates**, as indicated by the highest importance of the Taxes_logreturn feature (54.4%). The Brent_Price_logreturn also plays a significant role (16.9%), reflecting the strong influence of **global crude oil price volatility** on gasoline pricing. Additionally, short-term variations in distribution and marketing costs contribute notably to price changes (15.2%). In contrast, components such as refining and crude oil exhibit lower predictive power, suggesting that their changes are either more random or indirectly driven by other factors, and might not directly influence gasoline prices.

Spatial Clustering based on Feature Importances

Following the same methodology, in Weekly Gasoline Prices we have the weekly `Prices` for 30 different regions across the US. We looked at the feature importance distribution for each region individually, and compare the pattern between them to check whether there is any interesting spatial pattern.

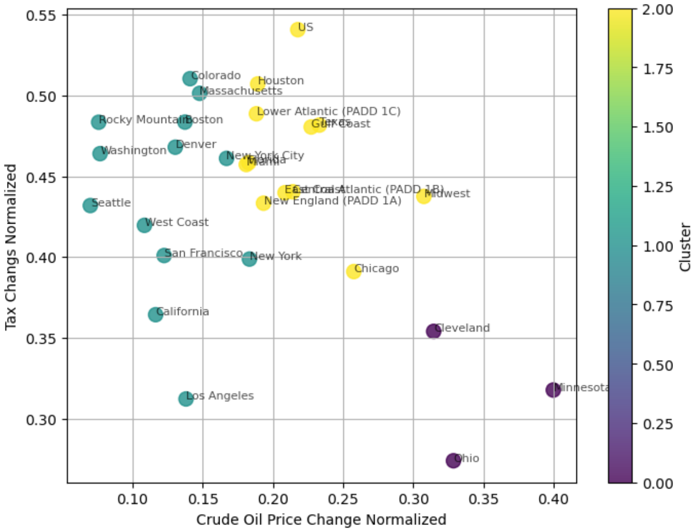


Figure 4: Gasoline Prices Feature Importance Clustering against Two Features

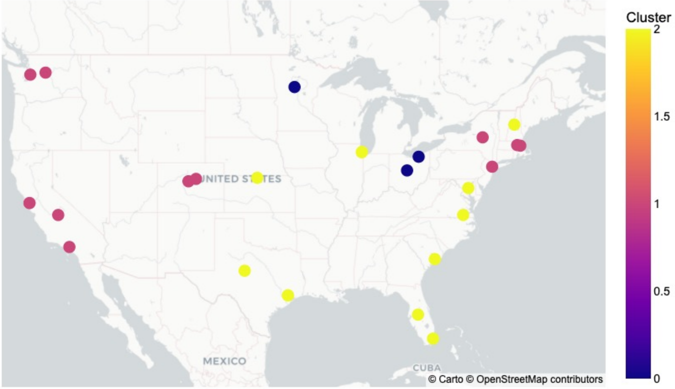


Figure 5: Gasoline Prices Feature Importance Spatial Clustering

We plotted scatter plots of the regions against the two most significant contributors (`Tax` and `Brent Price`) and observed distinct boundaries among the three clusters, as shown in fig 4.

- **Left (teal)**: Exhibits low crude oil price change influence and mixed levels of tax change. A possible explanation is that these states tend to have **highly regulated gasoline markets** and often enforce environmental standards and tax adjustments that reduce the direct impact of crude oil fluctuations on retail prices. Additionally, **urban density and established public transport systems** may buffer demand-driven volatility.
- **Middle (yellow)**: Shows moderate to high influence from both crude oil and tax changes. These are diverse and economically mixed regions, where gasoline prices are shaped both by **global oil market exposure** and **state-level tax policies**. For instance, Houston is a key oil industry hub where infrastructure may moderate some crude volatility.
- **Right (purple)**: Presents high crude oil price sensitivity and low tax change variation. These are predominantly Midwestern, industrial, and inland regions where state fuel taxes are relatively stable and infrequently adjusted. Additionally, with **less regulatory intervention and simpler distribution chains**, the crude oil component plays a more dominant role in retail pricing.

Across all analyses, taxes and crude oil prices consistently emerge as the most influential factors driving gasoline prices. Additionally, the spatial distribution of feature importance highlights distinct regional patterns, where different areas are influenced to varying degrees by these key drivers.

Causal Claims: We would like to clarify the rationale behind using causal terms such as *key drivers, influencing, and causing.* These are not directly concluded from correlation analysis or predictive modeling alone, but are instead inferred based on natural reasoning and simplified assumptions. For example, given the fact that gasoline is refined from crude oil—and that crude oil accounts for the largest component of gasoline costs—it is reasonable to infer that crude prices causally influence gasoline prices.

This inference is supported by the strong correlation observed between Brent crude oil and gasoline prices (fig 2b), as well as by visual evidence from the graphs showing that changes in crude oil prices typically precede changes in gasoline prices (fig 1). Lagged correlation test (fig 3b) reinforce this relationship for other contributing factors. Additionally, we performed **a Granger causality test**, in which all test statistics yielded p-values below 0.001, strongly rejecting the null hypothesis and providing statistical support for a causal relationship. We further discussed this in Sec 3.4

3.2 Ripple Effects Across the Energy Supply Chain

To better understand how fuel prices affect various companies within the energy industry, we performed a detailed correlation analysis between stock prices and fuel prices, including both gasoline and Brent crude oil, at daily and weekly frequencies. We specifically investigated whether companies at different positions in the energy supply chain, upstream (exploration), midstream (transport) and downstream (refining and retail) respond differently to fuel price fluctuations.

Assigning Supply Chain Roles

To analyze how gasoline price changes affect companies differently, we categorized energy stocks into four roles along the oil and gas value chain—**Upstream**, **Midstream**, **Downstream**, and **Integrated**. Classification was based on business focus and operations.

- **Upstream (Exploration and Production)**: Focused on extracting crude oil and natural gas. Examples: `EOG`, `DVN`, `FANG`, `OXY`.
- **Midstream (Transportation and Storage)**: Handle pipelines and distribution infrastructure. Examples: `KMI`, `WMB`, `EPD`.
- **Downstream (Refining and Marketing)**: Convert crude into fuels and petrochemicals for retail. Examples: `PSX`, `MPC`, `VLO`.
- **Integrated**: Operate across multiple segments. Examples: `XOM`, `CVX`, `BP`, `E`.

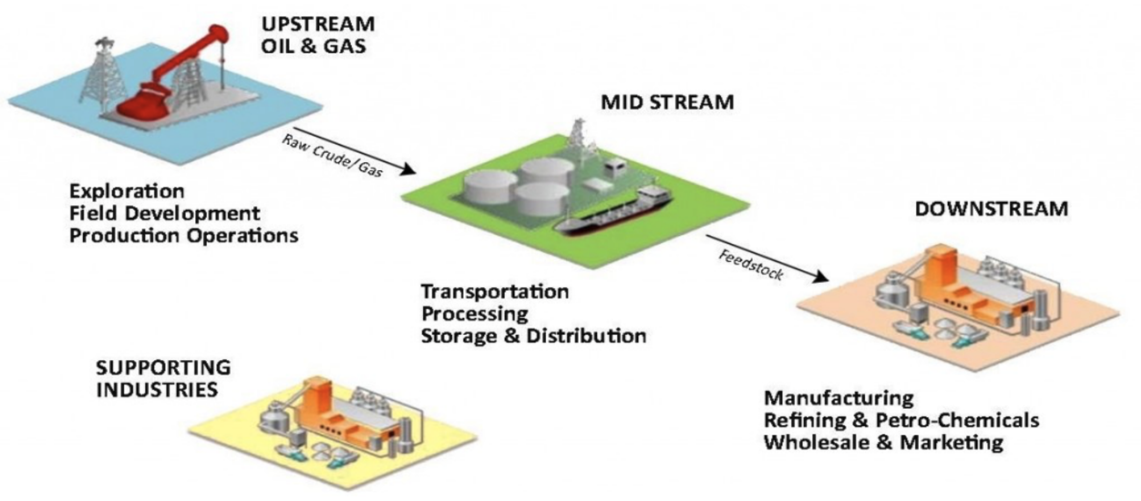
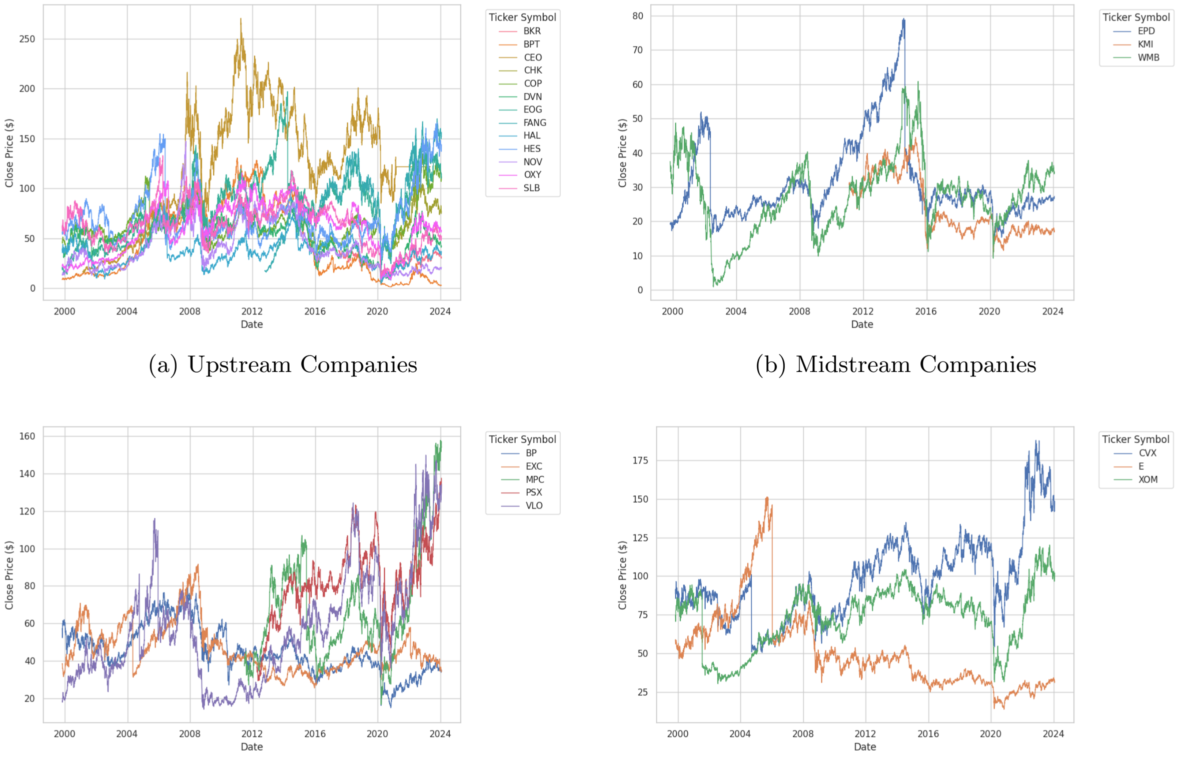


Figure 6: Diagram of the gasoline supply chain [2]

Other companies such as `EXC` (utility), `BPT` (royalty trust), and ETFs like `DIA`, `SPY` were excluded from supply chain analysis due to their different nature.

This segmentation enables structured comparisons of correlation patterns and highlights how companies across the chain exhibit varying sensitivities to fuel price movements.

To visually assess differences in market behavior between supply chain roles, we plotted historical stock price trends for companies within each category (see Figure 7); for overall trends in different segments of the energy supply chain, we plot representative stocks from each category (see Figure 8).



(a) Upstream Companies

(b) Midstream Companies

(c) Downstream Companies

(d) Integrated Oil Majors

Figure 7: Stock Price Trends by Supply Chain Segment (2000–2024)

Correlation Between Gasoline Prices and Stock Performance

To investigate the relationship between gasoline prices and energy stock performance, we conducted a comprehensive set of correlation analyses using both daily and weekly data. Our aim was to quantify how closely stock movements track fluctuations in gasoline prices across multiple time frames and return formats.

Specifically, we tested several correlation approaches across different data granularities:

1. **Daily Correlation — Log Returns:** For each stock and gasoline price series, we calculate the log return:

$$r_t = \ln\left(\frac{P_t}{P_{t-1}}\right)$$

where P_t is the close price on day t. The correlation is computed between the log return of each stock and the gasoline price's log return on a daily basis.

2. **Daily Correlation — Volatility Correlation:** We compute daily price volatility as the

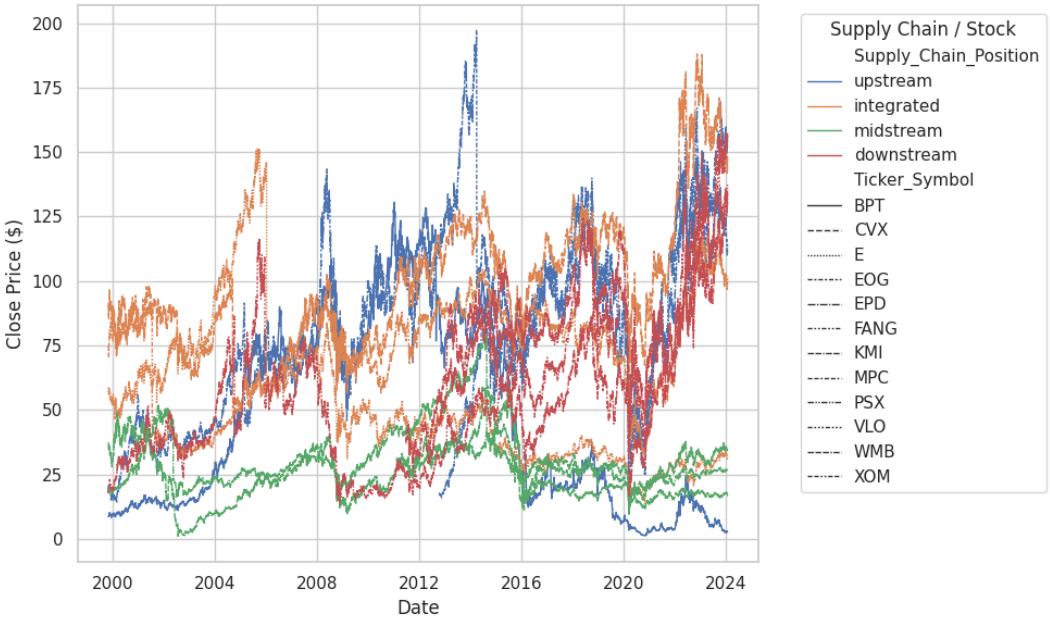


Figure 8: Selected Stock Price Trends by Supply Chain Category (2000–2024)

difference between high and low prices:

$$\text{Volatility}_t = \text{High}_t - \text{Low}_t$$

and correlate this with gasoline's log return to assess volatility co-movement.

3. **Daily Correlation — Rolling Volatility Correlation:** To smooth short-term fluctuations, we compute rolling volatility of log returns over a moving window (e.g., 5 days):

$$\text{RollingVol}_t = \sqrt{\frac{1}{n}\sum_{i=0}^{n-1}(r_{t-i} - \bar{r})^2}$$

and correlate this rolling stock volatility with gasoline log return.

4. **Daily Correlation — Normalized Log Returns:**

To account for scale differences and varying volatility across stocks and time, we additionally computed correlations between gasoline log returns and **normalized stock log returns**, where both series were transformed using **z-score standardization**. This correlation measure provides further insight into how aligned a company's relative return movement is with national gasoline price changes, independent of its baseline price level or volatility.

5. **Weekly Correlation — Log Return and Feature Correlation:** In addition to close prices, we evaluate:

- Weekly log returns

- Weekly open-close gap: $\text{Gap}_t = \ln(\text{Close}_t) - \ln(\text{Open}_t)$
- Weekly variance and volatility

All metrics are correlated against weekly gasoline log return to assess different aspects of weekly stock behavior.

6. **Interpolation Handling:** Since gasoline prices are originally reported at a weekly frequency, we apply **linear interpolation** to estimate missing values on intermediate trading days. This allows alignment with daily stock data and ensures valid calculation of daily log return correlations.

Across the correlation heatmaps (see Figure 9 and Figure 10), we observe that correlations between gasoline price fluctuations and stock-level metrics are generally modest in magnitude. This suggests that while some energy companies exhibit weak statistical relationships with gasoline dynamics, overall sensitivity appears limited across most features tested.

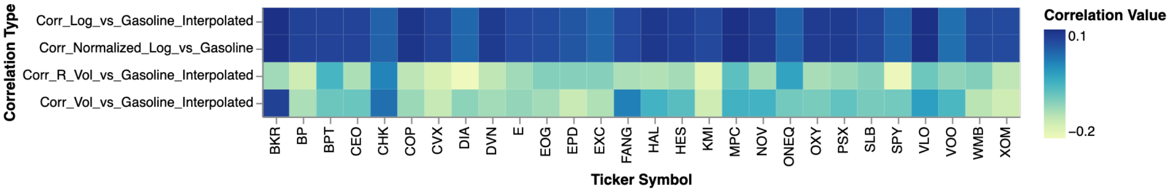


Figure 9: Daily Correlation: Stock-level Features vs. Interpolated Daily Gasoline Log Returns.

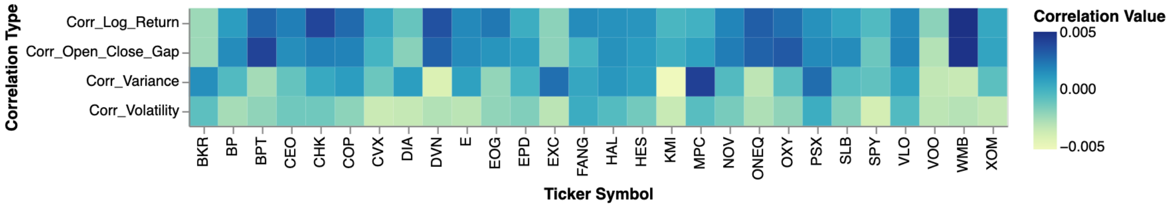


Figure 10: Weekly Correlation: Stock-level Features vs. Weekly Gasoline Prices Log Returns.

Specifically, daily correlations between gasoline log returns and stock returns or volatilities remain close to zero for most tickers, indicating that short-term gasoline price changes do not strongly co-move with equity performance. Similarly, weekly correlations involving open-close gaps, variance, or volatility reveal little consistent alignment with gasoline prices. Despite testing multiple correlation metrics using log returns, daily volatility, and rolling volatility, most results exhibit weak correlations with gasoline prices. These findings suggest that short-term gasoline price changes may not be the dominant driver of equity movement in this sector.

To explore whether absolute price levels might better capture firm-level sensitivity, we next examined direct correlations between normalized stock prices and gasoline prices.

To complement the return-based analysis, we also examined correlations using z-score normalized price levels. Figures 11 and 12 reveal which companies are most strongly aligned with fuel price changes at both weekly and daily levels. This approach mitigates scale differences and focuses on co-movement patterns over time.

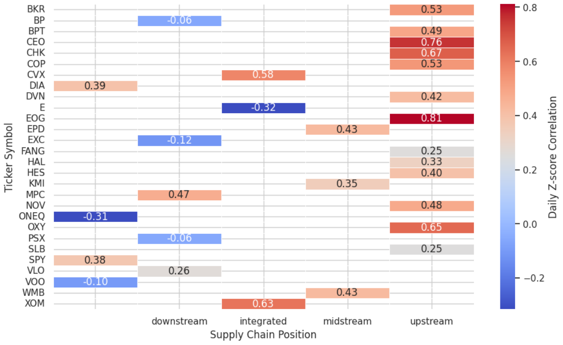


Figure 11: Daily Correlation: Normalized Stock Price vs. Gasoline Price

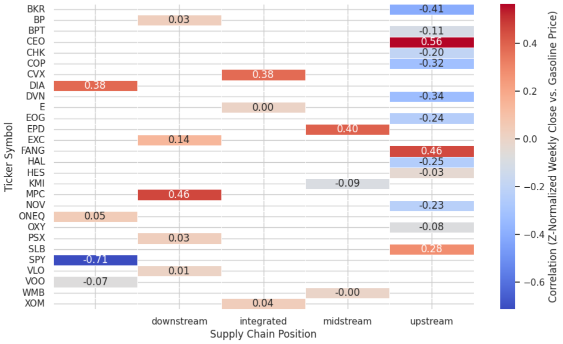


Figure 12: Weekly Correlation: Normalized Stock Price vs. Gasoline Price

Comparison with Crude Oil Price Given the limited responsiveness to gasoline price changes in log-return-based correlations, and the stronger alignment observed in normalized price correlations, we extend our analysis to Brent crude oil—an upstream benchmark more directly linked to input costs.

Brent prices are likely to impact exploration and production firms more directly than gasoline, which is subject to refining, distribution, and regulatory factors.

Figure 13 and 14 illustrate daily and weekly correlations between z-score normalized stock prices and Brent crude oil, reinforcing the differentiated price sensitivity along the supply chain.

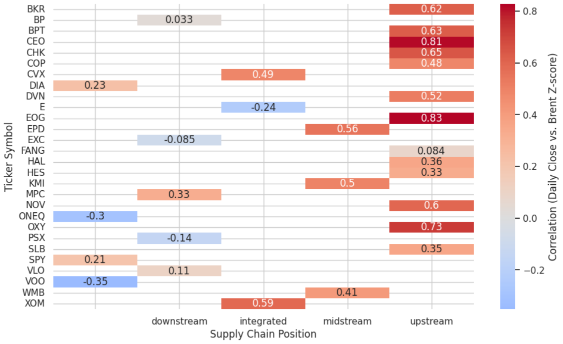


Figure 13: Daily Correlation: Normalized Stock Price vs. Brent Crude Oil Z-score

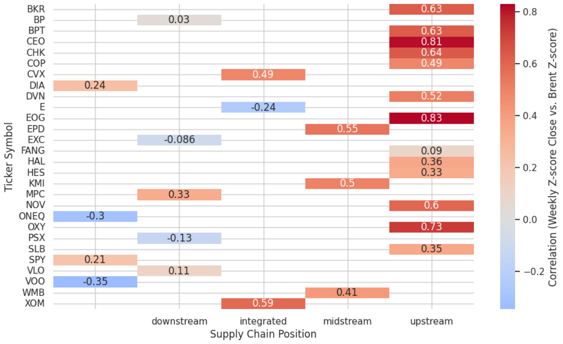


Figure 14: Weekly Correlation: Normalized Stock Price vs. Brent Crude Oil Z-score

Since stock log returns exhibit stronger and more consistent correlations with Brent crude oil than with gasoline, the remaining analyses in this study will use Brent prices as the primary benchmark for testing relationships and sensitivity patterns.

Differentiating Responses Across Supply Chain Positions

While the previous sections examine average sensitivity across all companies, this section focuses on heterogeneity in response based on supply chain roles. Specifically, we compare correlation strengths across upstream, midstream, downstream, and integrated firms to assess whether certain business models are more exposed to fuel price fluctuations.

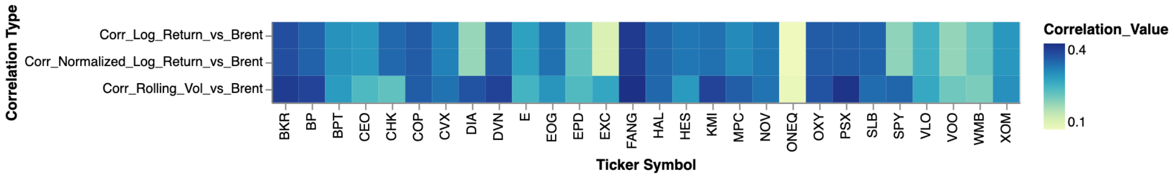


Figure 15: Daily Correlation: Stock Log Returns and Volatility vs. Brent Crude Oil

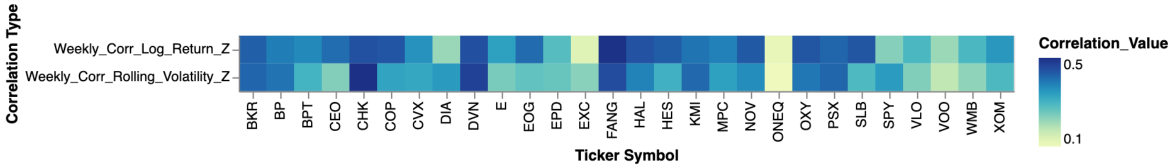


Figure 16: Weekly Correlation heatmap: Stock Log Return and Volatility vs. Brent Crude Oil (aggregated)

As shown in Figures 11–16, upstream firms such as CHK, CEO, and EOG exhibit higher alignment with Brent log returns and volatility, reaffirming their exposure to crude oil price fluctuations. In contrast, downstream and midstream companies generally show weaker or more dispersed correlation patterns.

These patterns highlight the greater exposure of exploration and production companies to commodity price shocks. In contrast, midstream firms—operating under more stable fee-based contracts—tend to show weaker or inconsistent correlations. Integrated firms show mixed patterns, reflecting diversified operations.

Together, these findings highlight the importance of supply chain segmentation when modeling energy stock behavior or designing price-sensitive investment strategies.

Beyond Correlation: Testing Causality with Granger Analysis While our correlation analysis revealed that certain upstream companies (e.g., CEO, EOG, CHK) exhibit stronger alignment with crude oil prices, correlation alone does not imply causality. To test whether changes in Brent crude oil prices actually precede and help explain future movements in stock returns, we conducted Granger causality tests.

Figure 17 presents a heatmap of p-values for the Granger causality test (Brent $\rightarrow$ stock log returns) across 1 to 3 lag days. Only results with statistically significant p-values ($p < 0.05$) are shown.

Several observations emerge:

- **Causality is selective**: Most firms do not exhibit significant Granger causality, underscoring that correlation does not imply predictability.
- **Different from correlation leaders**: Some highly correlated firms (e.g., CEO, EOG) are not Granger-causal, while others like PSX, MPC, SPY, and WMB are, suggesting predictive signals may emerge from different mechanisms.
- **Cross-supply-chain impact**: Granger-sensitive firms span downstream, midstream, and market proxies, indicating broader market transmission rather than upstream exclusivity.

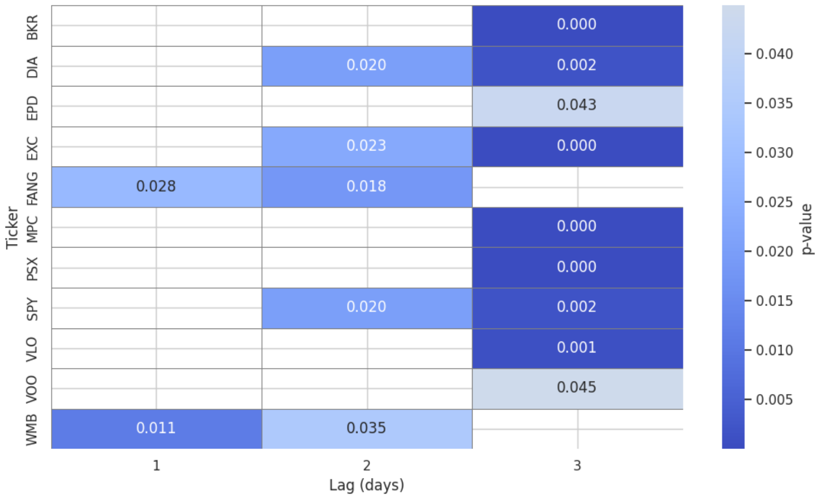


Figure 17: Significant Granger Causality: Brent → Stock Returns (p-value < 0.05)

Overall, Granger tests provide complementary insight by identifying firms whose returns may be predictably influenced by prior oil price changes—useful for modeling or hedging strategies.

Clustering Energy Stocks by Price Behavior

To go beyond the traditional classification of energy companies into upstream, midstream, and downstream, we applied an unsupervised clustering approach to group companies based on their empirical sensitivity to oil price fluctuations. The goal is to uncover latent groups such as oil-sensitive versus oil-insensitive stocks, which may not align perfectly with supply chain labels.

Feature Construction We constructed correlation-based feature vectors for each company using the following indicators:

- **Corr_Log_Return_vs_Brent**: Daily log return correlation with Brent crude oil.
- **Corr_Rolling_Vol_vs_Brent**: Daily rolling return volatility correlation with Brent.
- **Weekly_Corr_Log_Return_Z**: Weekly normalized log return correlation with Brent.
- **Weekly_Corr_Zscore_vs_Brent_Zscore**: Weekly correlation between Z-scored stock prices and Brent prices.

These metrics were selected for both statistical signal strength and interpretability. Gasoline-related features were excluded due to their weaker and inconsistent correlation patterns observed in earlier sections.

Clustering Methodology To further explore among energy companies in their sensitivity to oil prices, we apply **unsupervised clustering** based on multiple correlation-based features—including daily and weekly co-movement with Brent crude oil prices. By embedding these features into a lower-dimensional space using PCA and applying KMeans clustering across varying values of k, we identify consistent separation patterns. The silhouette scores suggest that a 2-cluster solution provides the most distinct separation ($Silhouette = 0.51$), with additional clusters gradually reducing separation quality.

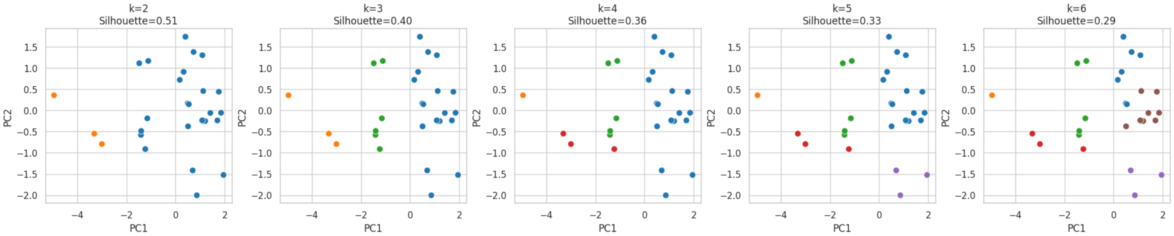


Figure 18: PCA Visualization of Clusters by k

Cluster Interpretation We observe the following patterns:

- **Cluster 0 (blue points)** typically captures firms with consistently high correlation across log return, rolling volatility, and weekly normalized movements—highlighting those with strong exposure to crude oil price dynamics. Many upstream producers fall into this group.

- **Cluster 1 and others**, in contrast, tend to exhibit weaker or more mixed correlations. These likely represent downstream or diversified firms whose revenues are less sensitive to crude price volatility.

Insights and Implications This clustering analysis provides an alternative lens for segmenting energy stocks. While some upstream firms indeed fall into highly oil-sensitive clusters, others do not—highlighting intra-category variation. Additionally, several downstream or integrated firms show unexpectedly high correlation values, defying simple supply chain-based assumptions.

Clustering Companies by Feature Importance from Return Prediction Models

To identify distinct response patterns across companies, we train return prediction models and compare their feature importance profiles. We use Brent **crude oil price**, **gasoline prices** and the **component make up percentage of gasoline** (e.g., crude oil share, taxes) as input features to predict weekly stock log returns (see formula above).

We apply XGBoost 2 regression models to estimate feature importances, and then cluster companies based on these profiles to reveal shared sensitivities.

As a preliminary experiment, we combine all stocks into a single dataset with a categorical stock identifier. After standardizing inputs and outputs, the combined model achieves a test-set MSE of approximately 0.39 (RMSE $\approx$ 0.62), suggesting moderate predictive performance. In contrast, per-stock models show signs of overfitting—test MSEs average around 1.0, while training errors are much lower—indicating that more data per company is needed to stabilize individual predictions.

Finally, we visualize the relationships among stocks by embedding their feature-importance profiles using UMAP, see Figure 19. The resulting projection reveals that stocks we previously labeled as the same type tend to lie close together in this low-dimensional space. This clustering pattern underscores the potential of our approach for grouping companies by sensitivity profile—and we expect these groupings to become even more pronounced as predictive accuracy improves.

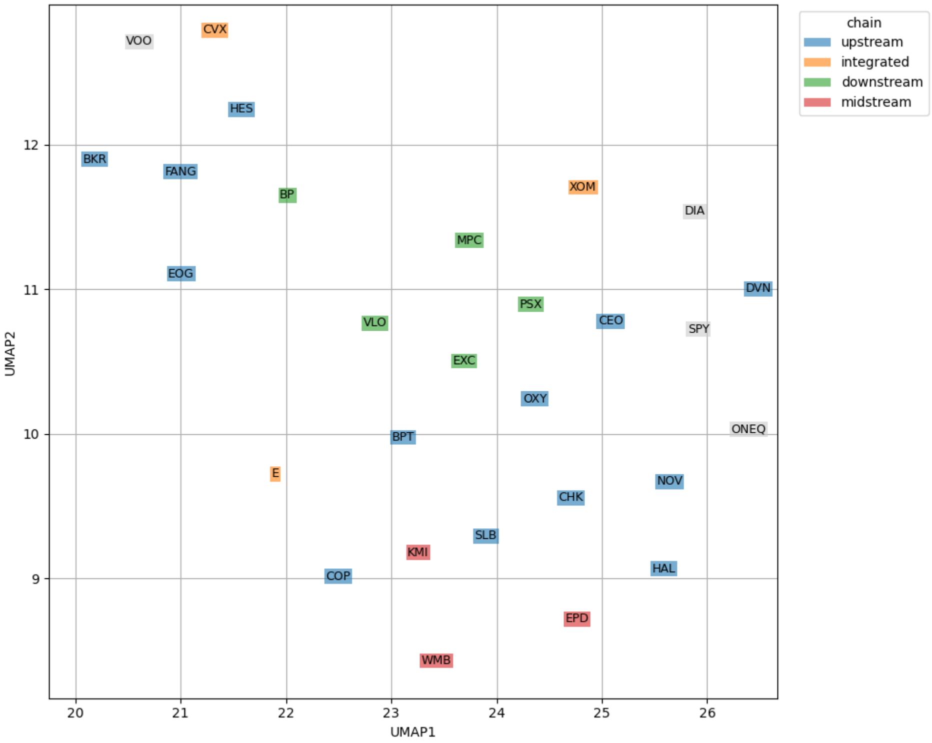


Figure 19: UMAP Projection of Stocks by Feature-Importance Profiles

3.3 Modeling EV and Automobile Manufacturing Indices

To assess how Gasoline and Brent crude oil prices influence sector-specific equity returns, we first conducted Pearson-correlation and 30-day rolling-correlation tests fig 20 between energy-price changes and our sector-return indices. We then developed two gradient-boosted decision-tree models for the daily return indices of electric-vehicle manufacturers (**EVidx**) and traditional automobile manufacturers (**Autoidx**).

Data Collection and Preprocessing

Our analysis covers the period from September 2018 through December 2023. We sourced closing prices for constituent stocks from Yfinance[1], constructing equal–weighted daily indices for both EV manufacturers (e.g., TSLA, NIO, RIVN, LCID) and traditional automakers (e.g., GM, F, MBGAF,

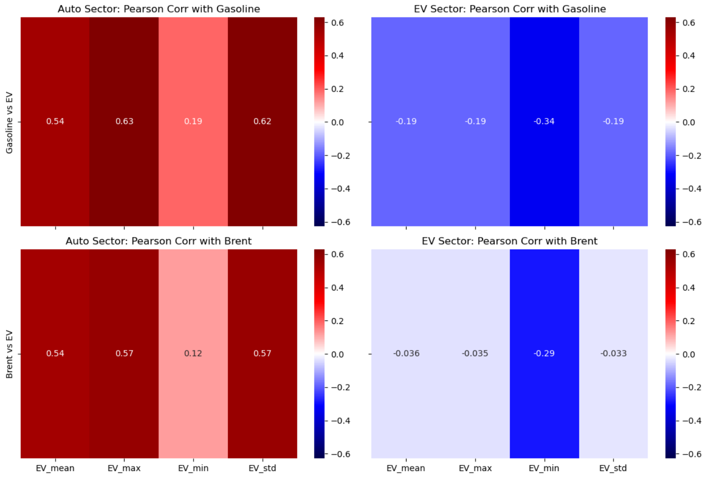


Figure 20: Impact of Brute and Gasoline on EV and Auto Indices.

TM). Daily percent–change returns were computed as

$$r_t = \frac{P_t - P_{t-1}}{P_{t-1}}$$

for trading days. For Brent crude oil and Gasoline prices obtained spot values, which included non–trading–day observations. These series were linearly interpolated to fill gaps and then re–indexed to the equity trading calendar so that each commodity return corresponded exactly to a stock–market date. All resulting series (EVidx, Autoidx, Brent, Gas) were standardized to zero mean and unit variance.

Feature Engineering

Given our objective to isolate the impact of oil and gasoline, the models relied exclusively on energy related features. Specifically, we included the daily percent–change returns of both Brent crude and Gasoline, along with their 1–5–day lags, as well as rolling volatility proxies defined as the 5–day standard deviation of each percent–change series. No additional macroeconomic or equity–index features were incorporated, ensuring that any predictive performance derived solely from energy price dynamics.

Model Specification and Hyperparameter Tuning

We utilise `HistGradientBoostingRegressor`, optimizing its hyperparameters via 5–fold, time–aware cross–validation. Specifically, we tuned the learning rate (0.01, 0.05, or 0.1), the maximum tree depth (3 or 5), the number of boosting iterations (`max_iter` = 200, 300, or 500), the minimum samples per leaf (10, 20, or 50), and the subsample fraction (0.7 or 1.0). The optimal configuration for **EVidx** was a learning rate of 0.01, max depth of 3, 300 iterations, 20 samples per leaf, and full subsampling. The **Autoidx** model instead performed best with a learning rate of 0.1, the same depth and iterations, 20 samples per leaf, and a subsample fraction of 0.7.

Hold–Out Evaluation and Ablation

We evaluated performance on a hold–out set covering September 2018–December 2023. Including interpolated oil and gas percent–change features, the EVidx model achieved an RMSE of 0.00507; omitting those features reduced RMSE slightly to 0.00438, indicating limited sensitivity of EV returns to daily energy–price moves. By contrast, the Autoidx model recorded an RMSE of 0.00112 (MAE = 0.00033, $R^2 = 0.994$) with full features, but RMSE rose to 0.00301 (MAE = 0.00053, $R^2 = 0.985$) when energy inputs were removed, underscoring the critical role of oil and gas percent–change dynamics in forecasting traditional auto returns.

Hold–Out Evaluation and Ablation

On the hold–out set (September 2018–December 2023), including the interpolated oil and gasoline return features had only a marginal effect on the EVidx model's predictive accuracy, whereas the Autoidx model saw a substantial degradation in performance once those energy inputs were removed. This contrast highlights that daily energy–price dynamics are far more informative for forecasting traditional auto returns than for EV returns.

Table 4: Model Performance on Hold–Out Set: Full vs. Ablated Features

	EVidx		**Autoidx**	
	Full Model	Ablated	Full Model	Ablated
RMSE	0.00507	0.00501	0.00112	0.00301
MAE	0.00226	0.00501	0.00033	0.00053

SHAP-Value Decomposition

To quantify and compare the marginal contributions of Brent versus gasoline features, we applied SHAP (SHapley Additive exPlanations) to each boosted model. For a given prediction $\hat{y}_i$, SHAP decomposes:

$$\hat{y}_i = \phi_0 + \sum_{j=1}^{M} \phi_j^{(i)},$$

where ϕ_0 is the model's expected value and $\phi_j^{(i)}$ the contribution of feature j on day i. Aggregating mean absolute SHAP values across our test set yielded:

- **EVidx**: Brent total mean —SHAP— = 2.82×10^{-5}, Gas total mean —SHAP— = 1.35×10^{-4}
- **Autoidx**: Brent total mean —SHAP— = 2.81×10^{-4}, Gas total mean —SHAP— = 9.93×10^{-5}

These results show that gasoline percent–change features dominate the EVidx model (gas impact 4.8× greater than Brent), whereas oil percent–change features dominate the Autoidx model (oil impact 2.8× greater than gas).

Interpretation

The SHAP analysis confirms our ablation findings: EV-sector returns derive more signal from gasoline dynamics than crude, while traditional auto returns are more sensitive to crude-oil shocks. This aligns with industry economic structures, where EV margins are less directly tied to fuel input costs than those of legacy automakers.

3.4 Limitations

1. **Data Size**: The dataset is relatively small, particularly after segmenting by countries and stocks (about 2K timestamps each). Additionally, inconsistencies in data frequency (e.g., weekly, monthly, and daily) introduce obstacles to accurate resampling and alignment.

2. **Modeling Constraints**: The XGBoost model we applied is prone to overfitting on small datasets, especially since we don't have a clear baseline to interpret the MSE. Even though we normalize the data for better explainability, the generalizability and robustness of the results remain in question.

3. **Simplifying Assumptions**: Our analysis assumes that gasoline prices are primarily driven by their component costs, while stock prices could be well predicted by gasoline prices and relevant factors in their makeup percentages. This simplification overlooks socio-technical contributors and also unpredictable abnormal periods like the 2008 financial crisis and the COVID-19 pandemic. Future work could consider such macroeconomic factors for a more comprehensive understanding.

4. **Causality Limitation**: Most of our models do not directly establish causality. To address this, more advanced methods such as Vector Autoregression (VAR) or difference-in-differences models would be required. Although we attempted these approaches, the results were not statistically significant. Therefore, our findings should be interpreted as indicating correlation rather than definitive causal relationships.

References

[1] Yahoo finance. `https://finance.yahoo.com`. Data accessed May 18, 2025.

[2] Stein Analytics. Oil - stein analytics, n.d. Accessed: 2025-05-18.