Analyzing Crude Oil Price Influence on Refined Fuel Markets

Trent Cooper II, A

Western Governors University

Table of Contents

[A. Project Highlights 4](#_Toc204396383)

[B. Project Execution 4](#_Toc204396384)

[C. Data Collection Process 5](#_Toc204396385)

[C.1 Advantages and Limitations of Data Set 6](#_Toc204396386)

[D. Data Extraction and Preparation 7](#_Toc204396387)

[E. Data Analysis Process 7](#_Toc204396388)

[E.1 Data Analysis Methods 7](#_Toc204396389)

[E.2 Advantages and Limitations of Tools and Techniques 8](#_Toc204396390)

[E.3 Application of Analytical Methods 8](#_Toc204396391)

[F Data Analysis Results 9](#_Toc204396392)

[F.1 Statistical Significance 9](#_Toc204396393)

[F.2 Practical Significance 11](#_Toc204396394)

[F.3 Overall Success 11](#_Toc204396395)

[G. Conclusion 12](#_Toc204396396)

[G.1 Summary of Conclusions 12](#_Toc204396397)

[G.2 Effective Storytelling 12](#_Toc204396398)

[G.3 Recommended Courses of Action 13](#_Toc204396399)

[References 14](#_Toc204396400)

# A. Project Highlights

This project aimed to explore whether U.S. crude oil prices significantly influence refined fuel prices such as diesel, gasoline, jet fuel, and propane. The analysis involved examining price data from the U.S. Energy Information Administration (EIA), transforming it into a usable format, and visualizing patterns to determine the relationship between crude oil and downstream fuel markets.

The project followed the CRISP-DM framework, progressing through understanding the business problem, analyzing and preparing the dataset, modeling relationships using regression, and visualizing results using Tableau.

**The scope of the project** is limited to analyzing the historical relationship between U.S. crude oil prices and six refined fuel types: conventional gasoline, RBOB regular gasoline, ultra-low sulfur diesel, No. 2 heating oil, kerosene-type jet fuel, and propane. The dataset spans monthly average prices from January 1986 through 2025, sourced exclusively from the U.S. Energy Information Administration (EIA).

This analysis focuses on U.S. domestic market data and excludes external factors such as global oil benchmarks (e.g., Brent crude), inflation, geopolitical conflict, or economic indicators. The project also does not attempt to forecast future prices but instead relies on regression modeling and lagged correlation techniques to evaluate historical patterns and relationships.

Python was used to process the dataset and build a multiple linear regression model. Tableau was used to create several visualizations, including a lagged correlation heatmap and grouped bar chart of regression coefficients. Excel was briefly used to preview the original structure of the dataset.

The structured approach supported the creation of a final dashboard and report that clearly display the influence of crude oil prices on refined fuel prices and assist with fuel-related decision-making.

# B. Project Execution

The project was completed using the CRISP-DM methodology as originally planned. Data was collected from the U.S. Energy Information Administration and consisted of monthly prices for crude oil and seven refined fuels, spanning from 1986 to 2025. The data was structured across multiple sheets, which were reviewed in Excel and processed in Python.

The data preparation stage involved reshaping the data to create a unified structure suitable for regression analysis. After preparing the dataset, exploratory analysis was conducted to examine trends and potential relationships. Lagged correlations were calculated to determine how crude oil prices affect refined fuels over time. This analysis helped justify the inclusion of lag variables in the model.

A multiple linear regression model was used to evaluate the statistical impact of crude oil prices on the prices of each refined fuel. Tableau was used to create interactive visualizations, including a heatmap of lagged correlation values and a grouped bar chart showing regression coefficients. The dashboard was published to Tableau Public and used to support the final analysis.

No major deviations occurred from the original project plan. The modeling, analysis, and deliverables were completed within the expected timeline and aligned with the initial proposal.

| **Milestone** | **Start Date** | **End Date** | **Description** |
| --- | --- | --- | --- |
| Data Exploration | July 10, 2025 | July 11, 2025 | Reviewed EIA dataset and confirmed relevant columns |
| Data Preparation | July 11, 2025 | July 12, 2025 | Structured dataset for regression and correlation |
| Modeling and Analysis | July 12, 2025 | July 13, 2025 | Performed lag analysis and regression modeling |
| Dashboard Development | July 13, 2025 | July 14, 2025 | Created Tableau visuals and designed dashboard layout |
| Final Report and README | July 14, 2025 | July 16, 2025 | Wrote summary, organized deliverables for submission |

# C. Data Collection Process

The dataset used for this project was sourced from the U.S. Energy Information Administration (EIA). It included monthly average prices from 1986 to 2025 for U.S. crude oil and seven refined fuels. The data was stored in an Excel workbook, with each fuel type on a separate sheet. Excel was used briefly to examine the file structure, but all data preparation was performed using Python.

Each sheet was read into a structured format using pandas, and fuel types were labeled based on their corresponding sheet. Dates were converted to a consistent datetime format, and price columns were renamed for clarity. The data was concatenated into a single DataFrame to allow for time series analysis and modeling.

No major cleaning was required, as the dataset was already complete and well-formatted. Minor formatting adjustments were made to ensure uniformity across fuel types. No imputation was necessary, as missing values were negligible and did not impact the analysis.

The prepared dataset was used for both lag correlation analysis and regression modeling. The data was also exported into Tableau to create the dashboard visuals.

The data used in this project is governed by the publicly available standards set by the U.S. Energy Information Administration (EIA). As a federal agency, the EIA provides high-quality, transparent data that is freely accessible and does not require licensing or special permissions, making it ideal for academic and public research.

No personally identifiable information (PII) was present in the dataset, and the project involved no data collection from individuals. All data was stored securely on a password-protected local device and used exclusively for academic purposes. The dataset was not modified beyond basic formatting and structural adjustments necessary for analysis.

This project complied with ethical, legal, and regulatory standards associated with public data use. No proprietary or sensitive information was used, and all data handling decisions prioritized data integrity, security, and responsible use throughout the course of the analysis.

## C.1 Advantages and Limitations of Data Set

The primary advantage of the dataset is its **length and completeness**. Spanning nearly four decades (1986–2025), the data provides a robust foundation for analyzing long-term trends and performing time series analysis. It includes consistent monthly price records for U.S. crude oil and multiple refined fuels, which supports correlation and regression modeling.

Another advantage is the dataset’s **credibility and accessibility**. It was sourced from the U.S. Energy Information Administration (EIA), a reputable and publicly funded agency. The data is well-organized, clearly labeled, and free from personal or proprietary information, making it easy to use in academic and professional settings.

However, the dataset also has limitations. It only covers **U.S. domestic pricing**, so the results may not generalize to international fuel markets. Additionally, while it captures pricing behavior, it lacks contextual variables such as supply chain disruptions, regulatory changes, or geopolitical events that may influence price fluctuations.

Despite these limitations, the dataset was suitable for addressing the research question and produced reliable, interpretable results.

# D. Data Extraction and Preparation

The dataset was extracted from the U.S. Energy Information Administration (EIA) and included monthly average prices for U.S. crude oil and seven refined fuel types from 1986 through 2025. The data was provided in Excel format, with each fuel type organized on a separate worksheet.

Each sheet contained one or more columns representing regional prices (e.g., New York Harbor, Gulf Coast, Los Angeles). To ensure consistency across all fuel types, the **New York Harbor price** was selected when available. If New York Harbor was not present, the **only available regional column** for that fuel was used.

The data was loaded into Python using pandas for processing. Columns were renamed for clarity, dates were converted to a standard datetime format, and all fuel types were merged into a single dataset. The resulting structure allowed for streamlined correlation analysis and regression modeling.

No significant data cleaning was required. The dataset was already complete and well-organized. Minor formatting adjustments were made to standardize column names and handle any irregularities in sheet labeling. The final dataset was saved and exported for use in Tableau and regression analysis.

# E. Data Analysis Process

## E.1 Data Analysis Methods

The project used three primary methods of analysis: descriptive analytics, lagged correlation testing, and multiple linear regression.

* **Descriptive Analytics**: Line charts were created to display annual average prices for crude oil and refined fuels. This helped visualize long-term pricing trends and contextualize the relationship between input (crude) and output (fuels).
* **Lagged Correlation Analysis**: This diagnostic method was used to determine whether changes in crude oil prices preceded changes in refined fuel prices. Correlations were calculated at multiple monthly lags (0–12 months), and the results were displayed in a heatmap. This method was appropriate because it allowed us to detect delayed market responses.
* **Multiple Linear Regression**: This predictive method modeled refined fuel prices as a function of crude oil price. It was used to support the hypothesis that crude oil prices significantly influence refined fuel prices. The model produced coefficients and p-values, helping assess both the strength and significance of the relationships.

These methods were appropriate for the project because they supported both **exploratory** and **inferential** goals: identifying patterns and quantifying relationships.

## E.2 Advantages and Limitations of Tools and Techniques

This project used several tools and techniques to conduct the analysis, each offering specific advantages while also presenting certain limitations. Python, along with its associated libraries such as pandas, numpy, and scikit learn, was the primary programming environment used to perform regression modeling and prepare lagged correlation data. Python’s flexibility and extensive library ecosystem made it well-suited for exploratory data analysis and statistical testing. However, it does require careful management of code and dependencies, especially when integrating multiple steps like data cleaning, modeling, and evaluation in a single workflow.

Jupyter Notebook served as the interface for executing Python code. Its interactive nature allowed for iterative testing and easy visualization of intermediate results, which was helpful during model development. One limitation, however, is that notebooks can become disorganized when outputs are not run in a linear order, potentially leading to confusion or reproducibility issues if not well-documented.

Tableau was used to create interactive visualizations and the final dashboard. It offered intuitive drag-and-drop functionality and strong visual storytelling capabilities, which helped present results to a non-technical audience. Tableau’s main limitation is its lack of advanced statistical modeling features, which is why it was used for visualization rather than analysis.

Excel was used only during the early phase of the project to inspect the raw data file and understand the structure of the sheets. While it is a convenient tool for quick viewing, it was not used for transformation or analysis due to its limitations in handling complex time-series or statistical tasks.

Each of these tools contributed meaningfully to different stages of the analysis. Python and Jupyter supported modeling and testing, Tableau enhanced communication and exploration of results, and Excel helped with initial data inspection. While limitations exist in each platform, using them together provided a balanced and effective analytical environment.

## E.3 Application of Analytical Methods

The data analysis process followed a structured, step-by-step approach to ensure reproducibility and accuracy. First, the dataset was loaded into Python and cleaned for consistency. Monthly average prices for crude oil and each refined fuel were extracted, and the date column was standardized to ensure proper alignment across all time series. The data was then reshaped from wide format to long format for easier analysis.

For the lagged correlation analysis, the crude oil price series was lagged by up to 12 months, and Pearson correlation coefficients were calculated between the lagged crude oil values and each refined fuel type. The resulting correlations and associated p-values were compiled to identify which lags had the strongest statistical relationship. To evaluate significance, an alpha level of 0.05 was used as the threshold for rejecting the null hypothesis, which stated that no correlation existed between the crude oil and refined fuel prices at each lag.

For regression modeling, a multiple linear regression was applied with crude oil price and its lagged values as features, and each refined fuel price as the response variable. The model’s assumptions — including linearity, independence, and normality of residuals — were verified through diagnostic plots and metrics such as R² and mean squared error (MSE). Each model was evaluated to determine whether crude oil prices had a meaningful predictive relationship with refined fuel prices.

These methods ensured the hypothesis could be tested from both a correlation-based and model-based perspective. All code was executed in Jupyter Notebook, and the output was validated visually and numerically before being summarized in Tableau for final presentation.

 **Lagged Correlation Analysis**:

* Crude oil price and each refined fuel price series were aligned by date.
* For each fuel, correlation coefficients were calculated between crude oil prices and fuel prices at lags ranging from 0 to 12 months.
* Correlation results were visualized using a Tableau heatmap.
* Assumption: Variables are continuous and time-aligned. This was verified during preprocessing.

 **Multiple Linear Regression**:

* The dataset was merged into a wide format, with fuel prices and crude oil prices by month.
* Each fuel type was modeled individually using crude oil price as the predictor variable.
* The scikit-learn LinearRegression() function was used to fit the model.
* The regression outputs (coefficients and p-values) were reviewed to ensure statistical significance.
* Assumptions: Linearity, independence, and minimal multicollinearity were checked by examining residuals and correlation matrices

# F Data Analysis Results

## F.1 Statistical Significance

A multiple linear regression model was used to evaluate whether crude oil prices significantly influence refined fuel prices. Each refined fuel was modeled using current crude oil price as the primary predictor. In addition, lagged correlation analysis was performed to determine the lead-lag strength of the relationship between crude oil and refined fuel prices.

**Regression Model Results (Lag 0)**

| **Fuel Type** | **Coefficient** | **R² Score** | **MSE** |
| --- | --- | --- | --- |
| Conventional Gasoline | 0.0258 | 0.8411 | 0.0548 |
| Jet Fuel | 0.0273 | 0.8907 | 0.0546 |
| No. 2 Heating Oil | 0.0271 | 0.8250 | 0.0899 |
| Propane | 0.0151 | 0.6208 | 0.0460 |
| RBOB Regular Gasoline | 0.0279 | 0.7877 | 0.0799 |
| ULS Diesel | 0.0277 | 0.8186 | 0.0975 |

* **Model Type**: Supervised regression
* **Algorithm Used**: Multiple linear regression (scikit-learn)
* **Performance Metrics**: Coefficient (effect size), R² score (explained variance), MSE (model error)
* **Benchmark**: R² ≥ 0.70 and positive coefficient indicates strong model performance

All refined fuels showed **positive coefficients**, indicating a direct relationship between crude oil and refined fuel prices. Jet Fuel had the strongest overall model fit (R² = 0.8907), followed closely by Conventional Gasoline and ULS Diesel. Propane showed the weakest fit (R² = 0.6208), suggesting its price is influenced by additional factors beyond crude oil.

**Lagged Correlation Test Results**

Lagged correlation was calculated between crude oil and refined fuel prices over 0–12 months. The **strongest absolute correlation** for each fuel is shown below:

| **Fuel Type** | **Max Correlation (r)** | **Lag (Months)** | **p-value** | **Significant (α = 0.05)** |
| --- | --- | --- | --- | --- |
| Conventional Gasoline | 0.982 | 0 | < 0.0001 | Yes |
| Jet Fuel | 0.981 | 0 | < 1 × 10⁻³⁰⁰ | Yes |
| No. 2 Heating Oil | 0.977 | 0 | < 1 × 10⁻³¹⁰ | Yes |
| Propane | 0.888 | 0 | < 1 × 10⁻¹³⁰ | Yes |
| RBOB Regular Gasoline | 0.909 | 0 | < 1 × 10⁻¹⁰⁰ | Yes |
| ULS Diesel | 0.920 | 0 | < 1 × 10⁻⁹⁰ | Yes |

All fuels showed strong, statistically significant correlations at lag 0, confirming that the current-month crude oil price has a measurable and immediate impact on refined fuel prices. These findings align with the regression model results and support the project’s hypothesis.

**Conclusion**

Both the regression and correlation results show strong, statistically significant relationships between crude oil prices and refined fuel prices. This provides sufficient evidence to **reject the null hypothesis** and support the claim that fluctuations in crude oil prices significantly influence refined fuel markets.

## F.2 Practical Significance

While statistical significance confirms that a relationship exists between crude oil prices and refined fuel prices, practical significance evaluates how meaningful that relationship is in real-world terms.

The regression results showed that changes in crude oil prices had a **strong and consistent effect** on refined fuel prices across most fuel types:

* **Jet Fuel**, **ULS Diesel**, and **Conventional Gasoline** had high R² scores (above 0.84), indicating that crude oil prices account for a large portion of their price variability.
* **Propane**, though influenced by crude oil (R² = 0.62), is likely affected by other variables such as weather seasonality, storage levels, and alternative production inputs.

The positive coefficients from the regression models confirm a **direct economic relationship**: as crude oil prices increase, refined fuel prices tend to rise. This insight has several practical applications:

**Example Application:**

A transportation logistics company could use this model to anticipate fuel cost changes based on crude oil market trends. If crude oil prices spike, they could:

* Adjust delivery pricing,
* Pre-purchase contracts at fixed fuel rates, or
* Shift to more fuel-efficient routing in advance.

**Decision Impact:**

Knowing the strength and immediacy of the crude oil–fuel relationship allows industry stakeholders to:

* Forecast budgetary needs,
* Inform supply chain timing, and
* Develop hedging strategies to reduce exposure to price volatility.

The findings are not only statistically sound but **actionable**, particularly for sectors like aviation, freight shipping, and agriculture, where fuel costs form a significant portion of operating expenses. The project thus achieves a high degree of practical significance, aligning data insights with operational decision-making.

## F.3 Overall Success

This project successfully met its original objective: to evaluate whether fluctuations in U.S. crude oil prices influence refined fuel prices in a statistically and practically meaningful way. The project integrated a multistep analysis process that included:

* **Multiple linear regression modeling**, which revealed strong, positive relationships between crude oil and refined fuel prices, with R² values exceeding 0.84 for most fuels.
* **Lagged correlation testing**, which confirmed immediate (lag 0) and statistically significant correlations between crude oil and all refined fuels (p-values < 0.0001).
* **Visualization and dashboard tools**, which effectively communicated the trends, relationships, and insights derived from the data.

In terms of success criteria defined in Task 2 (Section B6), the project met all key benchmarks:

* R² scores exceeded the target of 0.70 for all fuels.
* The model coefficients aligned with economic expectations (i.e., positive association).
* Stakeholder decisions could be directly supported by these findings (see F.2).

The analysis was completed within the defined timeline, used open and reliable data from the U.S. Energy Information Administration, and followed sound methodological practices. Based on these outcomes, the project can be considered both **statistically valid and practically useful**, demonstrating a successful execution of a real-world data analytics problem.

# G. Conclusion

## G.1 Summary of Conclusions

This project explored the relationship between U.S. crude oil prices and refined fuel prices through statistical modeling and correlation analysis. The results consistently demonstrated that:

* Crude oil prices have a strong, positive, and immediate influence on refined fuel prices.
* Regression models achieved high R² values, confirming that crude oil accounts for the majority of variance in refined fuel pricing.
* Lagged correlation testing showed statistically significant relationships at lag 0 for all fuels, with extremely small p-values, rejecting the null hypothesis in every case.

These findings supported the project hypothesis and aligned with real-world economic expectations. The results confirm that fluctuations in crude oil pricing can be used to anticipate changes in refined fuel markets, validating the model’s usefulness for decision-making.

## G.2 Effective Storytelling

The findings were visually communicated through a series of targeted visualizations created in **Tableau**:

1. **Line chart of annual crude oil prices**, giving stakeholders an intuitive view of long-term market trends.
2. **Line chart of refined fuel prices**, highlighting differences and patterns across fuel types.
3. **Grouped bar chart showing regression coefficients**, offering a side-by-side comparison of crude oil’s influence on each fuel.
4. **Heat map of lagged correlation results**, identifying the strongest lead-lag relationships and their statistical significance.

Together, these visuals created a coherent narrative from raw data to actionable insight, supporting transparency and understanding for technical and non-technical audiences alike.

## G.3 Recommended Courses of Action

**Recommendation 1:**  
Organizations in fuel-dependent industries (e.g., logistics, aviation, agriculture) should **monitor crude oil market trends in real-time** as a leading indicator for refined fuel costs. This could improve operational budgeting and risk forecasting.

**Recommendation 2:**  
Firms should consider **developing or adopting dynamic pricing models** that incorporate crude oil movements, possibly extending this work with machine learning to forecast future prices based on macroeconomic and geopolitical inputs.

These recommendations are rooted directly in the analysis, which confirms that crude oil price movements offer meaningful predictive power for downstream fuel markets.

 

# References

Asche, F., Osmundsen, P., & Tveterås, R. (2001). *An analysis of crude oil and refined product prices* (SNF Report No. 31/01). Institute for Research in Economics and Business Administration. <https://openaccess.nhh.no/nhh-xmlui/bitstream/handle/11250/164625/R31_01.pdf>

Brown, S. P. A., & Yücel, M. K. (1998). Crude oil and gasoline prices: An asymmetric relationship? *Federal Reserve Bank of Dallas Economic Review*, Q1, 2–11. https://www.dallasfed.org/research/er/1998/er9801a

U.S. Energy Information Administration. (2024). *Annual Energy Outlook 2024*. <https://www.eia.gov/outlooks/aeo/>