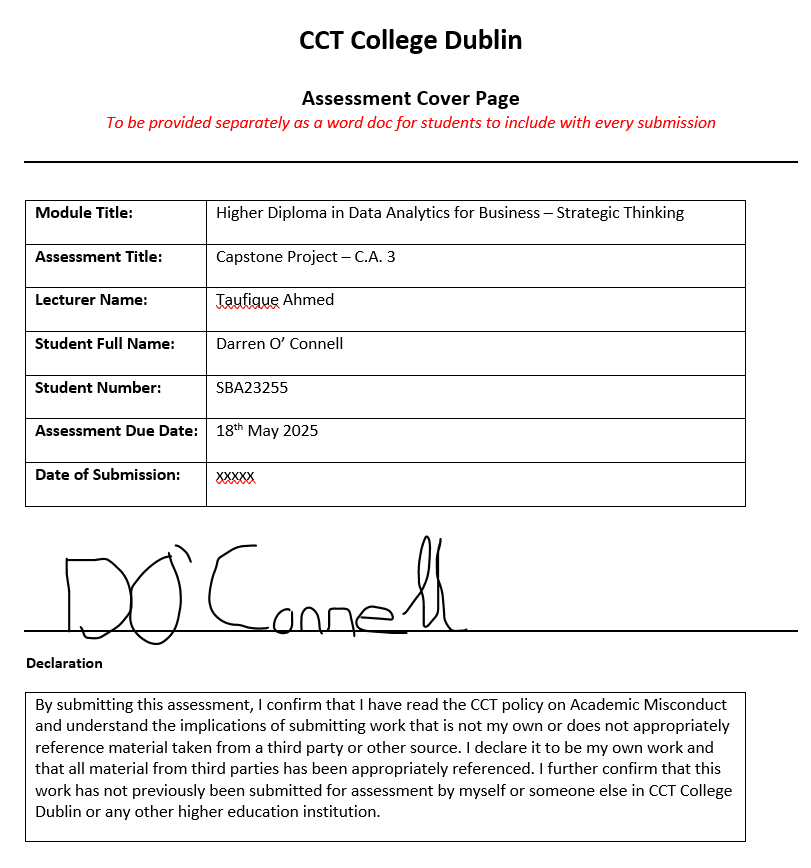
|  |
| --- |
| FoSSIL FUEL PRICES, INTEREST RATES, METAL PRICES, AND ECONOMIC ACTIVITY IN THE UNITED STATES  Continuous Assessment 3 |
| Darren O’ Connell SBA23255  Email |



A blue background with white text

Description automatically generated



A document with text and numbers

AI-generated content may be incorrect.

# Fossil Fuel Prices, Interest Rates, Metal Prices, and Economic Activity in the United States

## Abstract

This study explored time series data extracted from the FRED repository using the CRISP-DM methodology. Initial exploratory analysis revealed strong correlations between crude oil prices per barrel and LNG prices per million metric BTU respectively with manufacturers’ sales in the United States. Manufacturers’ sales were used as a proxy barometer of economic activity. Stationarity tests (e.g., ADF, KPSS) confirmed the non-stationary nature of the data, rendering it unsuitable for direct forecasting. Subsequent detrending by differences led to significantly weaker correlations between independent variables and the target variable. It is apparent that seasonal fluctuations in commodity prices and other economic factors exerted an undeniable impact on economic activity. Ergo, subsequent modelling will employ non-stationary data to capture underlying patterns and trend dynamics not inherent in the detrended data. This modelling strategy should lead to more robust and reliable predictive models.

**Keywords**

CRISP-DM, FRED, ADF, KPSS, stationarity, detrending, commodity prices, economic activity, United States

**Github link:**

<https://github.com/Darren-SBA23255-CCT-L8/Strategic_Thinking_HCI_S2>

## Table of Contents

Table of Figures x

Project Management Flowchart x

Business Understanding xx

Data Sources xx

Data Understanding xx

Data Preparation xx

Further EDA xx

Modelling xx

Evaluation xx

Deployment xx

Challenges xx

Conclusions xx

References xx

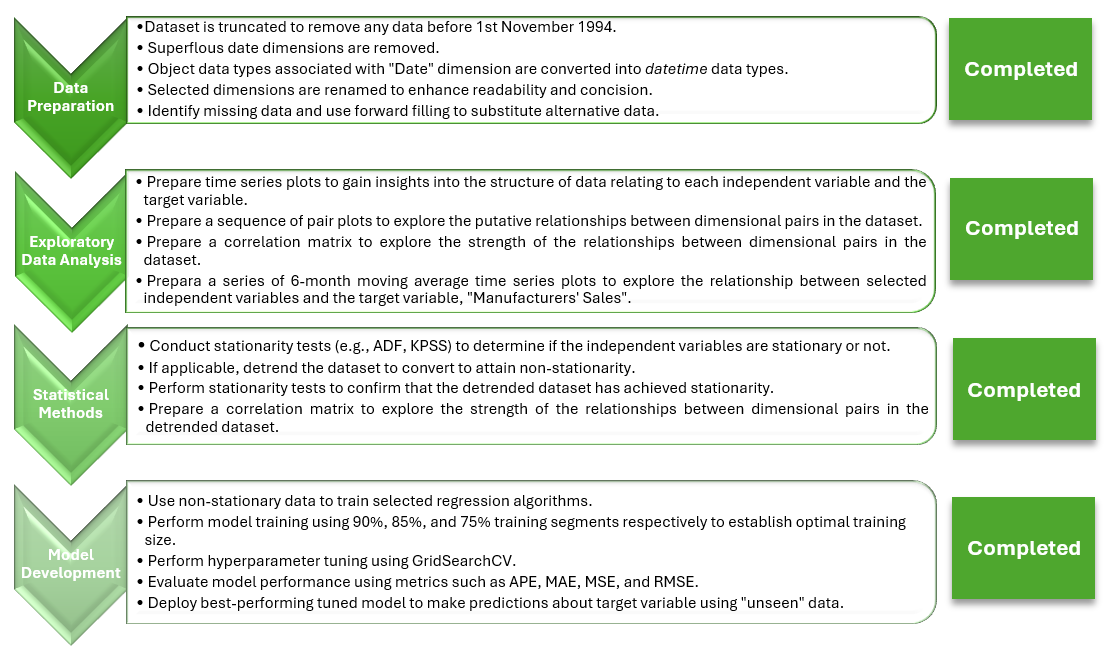
## Table of Figures

Figure 1. A preview of the imported dataset, "Consolidated FRED Source Dataset.csv" is shown in this figure.

Figure 2. This figure shows a list of data types associated with each dimension in the dataset.

[…]

## Project Management Flowchart



## Business Understanding

[Text goes here …]

The objective of this study is to elucidate any causal linkages between global fossil fuel prices, interest rates, and economic activity in the United States using the CRISP-DM methodology (Martínez-Plumed *et al.,* 2021). Manufacturers’ sales will be used as a proxy barometer of economic activity.

## Data Sources

Aggregate time series data from the Federal Reserve Bank of St. Louis are used to investigate if variations in global crude oil prices, LNG prices, and interest rates exert an appreciable effect on manufacturers’ sales in the United States. The data is primarily extracted from FRED, an online database comprising hundreds of thousands of economic time series aggregated from national, international, public, and private sources. FRED is maintained by the Research Department at the Federal Reserve Bank of St. Louis, Missouri, United States.

The following list shows the datasets that are used in this study:

1. Crude Oil Prices: Brent – Europe

<https://fred.stlouisfed.org/series/DCOILBRENTEU>

1. Crude Oil Prices: West Texas Intermediate (WTI) to Cushing, Oklahoma

<https://fred.stlouisfed.org/series/DCOILWTICO>

1. Global Price of LNG, Asia

<https://fred.stlouisfed.org/series/PNGASJPUSDM>

1. Commercial Bank Interest Rate on Credit Card Plans

<https://fred.stlouisfed.org/series/TERMCBCCALLNS#0>

1. Flexible Price Consumer Price Index - excluding food and energy

<https://fred.stlouisfed.org/series/FLEXCPIM679SFRBATL>

1. Export Price Index (End Use): Nonmonetary Gold

<https://fred.stlouisfed.org/series/IQ12260>

1. Global price of Iron Ore

<https://fred.stlouisfed.org/series/PIORECRUSDM>

1. Global price of Copper

<https://fred.stlouisfed.org/series/PCOPPUSDM>

1. US Manufacturers' Sales

<https://fred.stlouisfed.org/series/MNFCTRSMNSA>

The frequency of the time series data is monthly for ease of analysis. Non-seasonally adjusted (NSA) data is used to ensure fair comparisons between disparate datasets. Seasonally adjusted data is not used to avoid introducing biases into the datasets (Doppelt, 2023).

The relevant time series data from each of the datasets are consolidated in a Microsoft Excel spreadsheet to create the final dataset, “Consolidated FRED Source Dataset.csv”. This dataset is imported into a Jupyter notebook, “SBA23255\_L8\_HDipDB\_Strat\_CA2\_DOConnell” for subsequent exploratory data analysis (EDA) and data preparation. A preview of the consolidated dataset is shown in **Figure 1.**

A screenshot of a computer

Description automatically generated

Figure 1. A preview of the imported dataset, "Consolidated FRED Source Dataset.csv" is shown in this figure.

## Data Understanding

The dataset comprises 14 dimensions distributed amongst float and object data types. There are 689 observations in the untreated dataset.

A screenshot of a computer

Description automatically generated

Figure 2. This figure shows a list of data types associated with each dimension in the dataset.

The scope of the data series applicable to each dimension is shown in the following statement:

**Flexible Price CPI (% change):**

01-02-1967 to 01-06-2024

**US crude oil prices per barrel (USD $):**

01-01-1986 to 01-06-2024

**Brent crude oil prices (Europe) per barrel (USD $):**

01-05-1987 to 01-06-2024

**Global LNG price per million metric British thermal unit (BTU):**

01-01-1992 to 01-06-2024

**US manufacturers’ sales (millions, USD $):**

01-01-1992 to 01-06-2024

**US credit card rates (%):**

01-11-1994 to 01-06-2024

Since the data series associated with each dimension does not begin at the same date (e.g., 01-02-1967), there is a significant quantity of missing values. In order to minimise the adverse influence of missing values on data completeness and quality, the scope of the dataset is truncated from 1st November 1994 to 1st June 2024. This step is carried out during the data preparation stage.

## Data Preparation

During the data preparation stage, the following sequential steps were taken to improve data consistency and quality:

* The dataset is truncated to remove any data prior to 1st November 1994.
* Superfluous date dimensions are removed from the dataset.
* Object data types associated with “Date” dimension are transformed into datetime data types.
* Selected dimensions are renamed to render the corresponding labels concise and more readable. For example, the dimension “Actual Manufacturer's Sales (USD, millions)” is converted into “Manu Sales (USD, millions)”.

The Python code pertaining to the preceding data preparation steps is shown in **Figure 2.**

A screenshot of a computer

Description automatically generated

Figure 3. The Python code shown in this figure is used to restrict the date range of the dataset, remove extraneous dimensions, and rename selected dimensions.

The next step entails the application of the “find\_missing\_values” function to the truncated dataset. There are 237 missing values within the dataset relating to the dimension, “US credit card rates (%)”. No other missing data is reported. The “pad” (or forward fill) method is used to propagate the last valid observation forward to the next valid observation. Accordingly, missing values within the “US credit card rates (%)” dimension are filled. The Python code relating to the preceding steps are shown in **Figure 4.**

A screenshot of a computer code

Description automatically generated

Figure 4. The “find\_missing\_values” function is used to find dimensions with missing values in the dataset. The only dimension with missing data is US Credit Card Rates (%). The “fillna” method is subsequently used to replace missing data by “forward filling”.

A screenshot of a computer

Description automatically generated

Figure 5. A preview of the truncated dataset is shown in this figure. The data in the truncated dataset spans the range from 1st November 1994 to 1st June 2024 (inclusive).

## Further exploratory data analysis

Using the truncated dataset shown in **Figure 5,** further exploratory analysis is performed on the remaining 7 dimensions and 356 observations. The objective of this analysis is to use graphical representations of the data to uncover underlying patterns and trends, spot anomalies, and posit hypotheses.

[More text goes here …]

## Modelling

Non-stationary time series data will be used to train the following four models:

* Gradient Boosting Regressor
* Random Forest Regressor
* Support Vector Regressor
* XG Boost Regressor

The ideal test split (10%, 15%, or 25%) will also be identified. Hyperparameter tuning will be performed using GridSearchCV. The tuned models will then be used to predict the target variable, US manufacturers’ sales. The findings relating to model development will be presented in a separate report.

Whilst the ARIMA algorithm is commonly used for analysing and forecasting time series data (Siami-Namini *et al.,* 2018), modelling will not be conducted with this algorithm. Indeed, time series forecasting is not the objective of this study.

## Evaluation

To evaluate model accuracy and performance, a wide range of different evaluation metrics will be used:

* Mean Absolute Error (MAE)
* Mean Absolute Squared Error (MSE)
* Root Mean Squared Error (RMSE)
* Absolute Percentage Error (APE)

Using these metrics as a guide, the best-performing tuned model will be used to make predictions using “unseen” internal data. The findings relating to model evaluation will be presented in a separate report.

[More text goes here …]

## Deployment

Once the best-performing model has been identified, predictions using real-world external data will be performed. Model accuracy and reliability will then be assessed by comparing predictions with actual data relating to US manufacturers’ sales. A graphical representation of the average percentage error between predicted and actual data will be presented in a separate report.

[More text goes here …]

## Challenges

Thus far, no significant and insurmountable challenges were experienced during the project.

## Conclusions

[Text goes here …]

## References

Baidya, R., Lee, S.W. (2024). Addressing the non-stationarity and complexity of time series data for long-term forecasts. Appl. Sci. 14(11), 4436.

<https://doi.org/10.3390/app14114436>

[More text goes here …]

US Federal Government, 2020. The Economic Benefits of Oil and Gas. US Department of Energy.

<https://www.energy.gov/articles/economic-impact-oil-and-gas#:~:text=The%20oil%20and%20gas%20industry,and%20ensures%20our%20energy%20security>

## Word Count

Excluding Cover Titles, Abstract, Table of Contents, Table of Figures, and References: XXXX approximately.

**Github link:**

<https://github.com/Darren-SBA23255-CCT-L8/Strategic_Thinking_HCI_S2>