Crude Oil Market Insights: Trends, Forecasting, and Optimization

Nishant Nayak
Msc in Data Analytics
National College of Ireland
Dubin, Ireland
x22248242@student.ncirl.ie

Devendrakumar Rajput
Msc in Data Analytics
National College of Ireland
Dubin, Ireland
x23318643@student.ncirl.ie

***Abstract*—The worldwide crude oil market is crucial for the economy and has a decisive effect on the energy strategies, industrial approaches, and financial markets of the world. We will analyze the historical price trends to understand how the geopolitical and macroeconomic factors affect them before creating predictive algorithms for crude oil prices with advanced data analytics techniques. A strong methodology was used to achieve these goals. In addition, historical datasets on crude oil prices, production levels, global demand, and economic indicators were pre-processed and analyzed using Python and related libraries. Giant price swings changes in the price of crude oil can reverberate through the economy and across international politics. Being able to accurately forecast the price of crude oil is important for decision-making in a broad range of fields, including energy, finance, and policy. In this research, we apply machine learning to predict crude oil price.**

Keywords— Brend Crude Oil Prices Over time

I. Introduction

Crude oil is a fundamental component of the worldwide energy supply, with prices and production influencing economies, industries and geopolitical strategies around the globe. But the crude oil market can be erratic, making it very difficult for everyone involved, from policymakers to investors. The risk-based dynamic global energy demand, transition to renewables, and geopolitical situation changes have added impetus to such data-driven understandings of crude oil market movements.

This project aims to explore these challenges and opportunities through the analysis of crude oil market data to identify patterns, understand the dynamics affecting prices, and create models to predict trends. We aim to answer the research question "What are the important factors affecting fluctuations in the price of crude oil and to what degree can we make accurate price trend predictions through the use of data analytics? This research offers practical recommendations for industry players to address the challenges inherent in the crude oil market.

At its core, the relevance of this analysis can be seen in its potential to fill knowledge gaps around understanding crude oil prices volatility — and how these changes with new economic and environmental pressures manifesting themselves.

The objectives include:

Understanding the key economic/geopolitical/market factors affecting crude oil price movement Building forecasting models to predict crude oil price fluctuations in both the short- and long-term. There can be no systematic use if there is no systematic approach towards the efficiency of the market. With this project, we hope to add to the emerging literature of energy market analytics and to lay the groundwork for better decision making in a vital sector.

II. Related Work

A. Price Forecasting with Machine Learning Models

Because of the potential response of global economy to crude oil, it has become an attractive area of academic interest. Crude oil price dynamic production, trends, and market behavior have been studied in a number of studies, using statistical and machine learning methods. This section critically analyzes key studies that inform this project, discussing their methods, findings, and limitations. Machine Learning Models for Price Prediction Notable others that have tried to apply machine learning techniques not limited to support vector machines (SVM), random forests, and neural networks for crude oil price prediction. For instance, Zhang et al. (2019) used LSTM (Long Short-Term Memory) on Oil pricing predictions and showed that it outperforms conventional econometric models such as ARIMA. But the study was too reliant on historical price data without taking into account wider macroeconomic or geopolitical trends, reducing the robustness of the model through fluctuating markets.

B. Economic and Geopolitical Influences on Oil Prices signations.

Hamilton (2009) provided a seminal analysis of the economic and geopolitical drivers of crude oil price volatility, emphasizing the role of supply shocks and global demand. While the work laid a strong theoretical foundation, it lacked predictive capabilities and real-time applicability. More recent studies, such as Kilian and Murphy (2014), expanded on this by using structural vector autoregression (SVAR) models, though these approaches often require strong assumptions about the underlying data distributions and relationships.

C. Economic and Geopolitical Influences on Oil Prices signations.

Because of the potential response of global economy to crude oil, it has become an attractive area of academic interest. Crude oil price dynamic production, trends, and market behavior have been studied in a number of studies, using statistical and machine learning methods. This section critically analyzes key studies that inform this project,

discussing their methods, findings, and limitations. Machine Learning Models for Price Prediction Notable others that have tried to apply machine learning techniques not limited to support vector machines (SVM), random forests, and neural networks for crude oil price prediction. For instance, Zhang et al. (2019) used LSTM (Long Short-Term Memory) on Oil pricing predictions and showed that it outperforms conventional econometric models such as ARIMA. But the study was too reliant on historical price data without considering wider macroeconomic or geopolitical trends, reducing the robustness of the model through fluctuating markets.

In response to these gaps, this project integrates diverse data sources—economic indicators, geopolitical factors, and historical price trends—into a comprehensive analytical framework. By employing advanced machine learning techniques alongside interpretable statistical models, the study aims to balance predictive performance with actionable insights. Furthermore, it explores novel questions, such as identifying underappreciated drivers of price volatility and detecting market inefficiencies, to extend the current understanding of crude oil markets.

III. Methodology

A. *Datasets and Justifications*

Datasets Used:

1) Primary Dataset: CSV file containing historical crude oil prices with columns such as date, open, high, low, close, and volume.

Justification: Provides granular time-series data for trend analysis and forecasting.
Comprehensive and clean structure for modeling daily price behaviors.

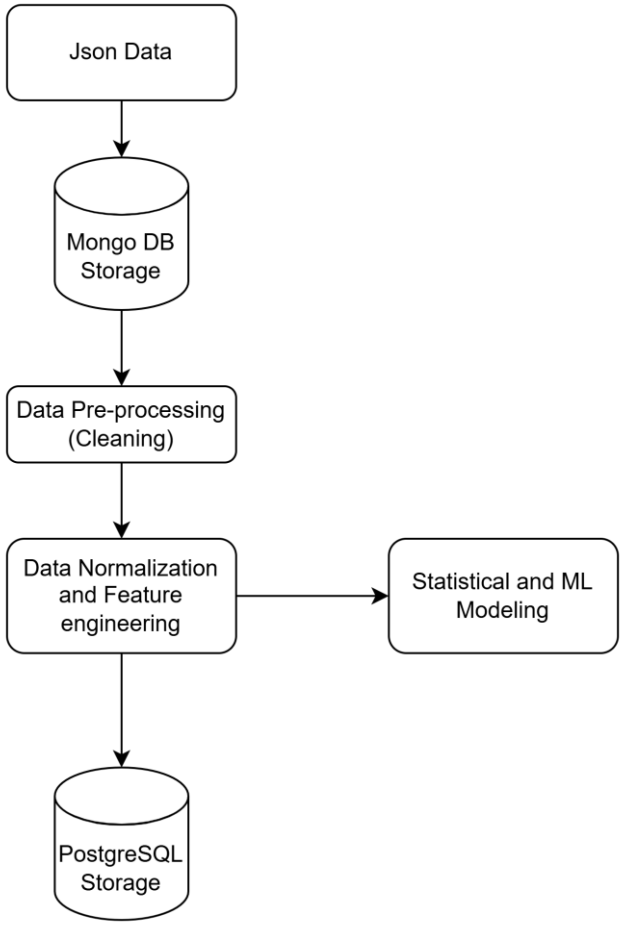
*2) Secondary Dataset:*API data fetched using the Alpha Vantage API, containing monthly crude oil prices.

Justification: Complements the granular daily data by offering longer-term price trends and patterns.
Facilitates cross-validation of analysis results.

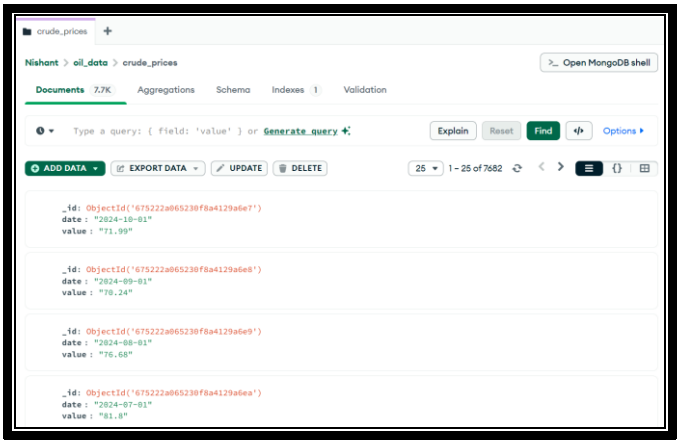
These two datasets cover both short-term (daily) and long-term (monthly) price movements.
Allow integration of data for robust trend analysis, outlier detection, and predictive modeling.
Easily accessible and regularly updated for real-time analysis. Below is the workflow overview.

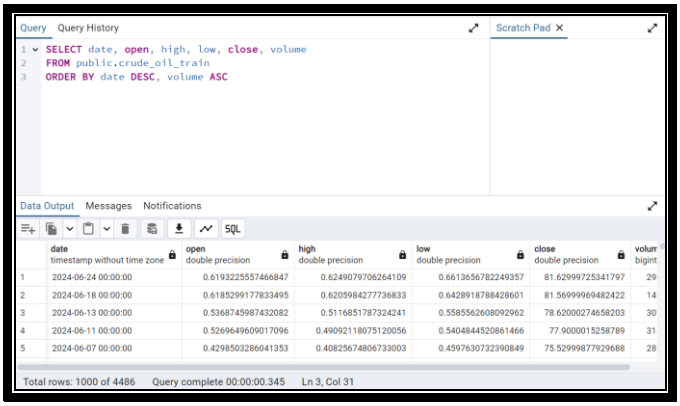
Workflow Overview

Step	Technology Used	Description
Data Extraction	`requests`, `json`	Fetch raw crude oil price data from APIs and parse it into usable JSON format.
Raw Data Storage	MongoDB, `pymongo`	Store API data in MongoDB for flexible, schema-free storage.
Data Transformation	`pandas`, `numpy`	Clean, scale, and transform the data for further analysis (e.g., StandardScaler for normalization).
ETL Pipeline	Dagster	Orchestrates the flow of data from extraction to transformation and storage in PostgreSQL.
Processed Data Storage	PostgreSQL, `psycopg`	Store cleaned, transformed data in a structured format for querying and analysis.
Feature Engineering	`scikit-learn`	Perform PCA for dimensionality reduction and KMeans for clustering the data.
Statistical Analysis	`statsmodels.api`	Evaluate relationships and model performance using statistical tests and regression.
Visualization	`matplotlib`, `seaborn`	Create visualizations like heatmaps, PCA plots, and line charts to present insights.
API Serving	Flask	S↓ the analysis results or processed data through APIs for integration with other applications.



Mongo DB:
*The diagram shows a MongoDB collection named **crude_prices**, which is part of the **oil_data** database. MongoDB is a NoSQL database that stores data in a flexible, document-based format (**JSON-like objects**). Each document represents an individual record with specific fields.*





B. Data Processing Activities

1) Data Gathering:

CSV Data: It is directly loaded from the uploaded dataset into a Pandas DataFrame.

API Data: It is fetched programmatically using the Alpha Vantage API with JSON parsing.
Automated data collection allows access to the latest insights, reducing manual data processing

2) Data Cleaning

Processed missing values by removing rows that contained NaNs. Converted date columns to datetime objects for uniform time-series processing. Sorted datasets by date to preserve sequence for trend & forecast models

3) Data Enrichment

Added derived features:

a) Daily Price Change — The change from closing to opening price to assess the daily volatility.
b) Moving Average: 3 day rolling average.

4) Data Storage:

a) **Cleaned and processed data**, writing it into a PostgreSQL database using the **sqlalchemy** library.

b) **Justification**: PostgreSQL is optimal for efficiently storing structured, time-series data; it gives you indexing.

c) ***PostgreSQL:***

The image displays a query execution result in a **PostgreSQL database** under the **oil_data** schema. The table being queried is **public.crude_oil_train**, and the SQL query retrieves specific columns with sorting applied.

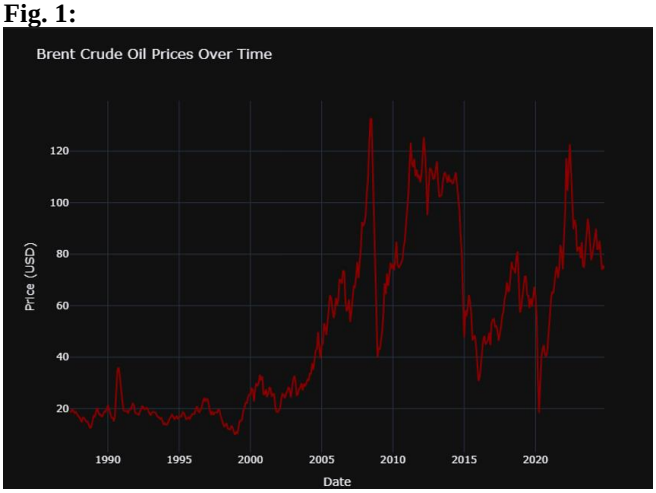
Data Processing Algorithms:

Trend Analysis: Using rolling averages to smooth daily data to find longer term trends. o Seasonal decomposition of time series (STL) technique was used to separate the seasonality and trend components. o Reason: This helps in understanding the actual patterns behind price movements.
Correlation Analysis: Estimated Pearson correlations to find associations between crude oil prices with daily measures (e.g., volume).
Justification: Kicks in to list any potential drivers of price change
Predictive Modeling: ARIMA (Auto Regressive Integrated Moving Average) for univariate time-series forecasting. Multivariate Random Forest Regression for Volume and Historical Trends Prediction. o Critique: ARIMA is specific for time-series forecasting while Random Forest captures non-linear relationships.
Visualizations:

Fig. 1: Line Plot - Used Plotly to create Line plots for price trends.
Fig. 2: Histogram - Shows price distribution.
Fig. 3: Moving Average - A rolling window of 3 months is calculated for smoothed trends.
Fig. 4: Scatter Plot - PCA graphical representation using a scatter plot.
Fig. 5: Scatter Plot – K- menas Clustering Algorithm.
Fig. 6: Correlation Heatmap - Correlation coefficients for the numerical features of the crude oil prices dataset

Scheduling: Dagster is used for orchestrating workflows, ensuring the job runs every 5 minutes.

Fig. 1:



Here is a line plot indicating the temporal variation in Brent Crude Oil Prices (in USD) from 1988 to 2023. The chart provides details on historical price movements, major increases and downturns.

Key Components of the Plot:

X-Axis: Date (1988–2023)
Y-Axis (Price in USD): Shows the price of Brent crude oil that ranges from $0 to $140.
Red Line: The period's Brent crude oil prices.
Observations:
Early Years (1988 - 2000):Prices were relatively constant, moving back and forth between $10 and $30 per barrel. Small peaks could be explained by political or economic disruptions, but the overall price was low.
Price Upward Trend (2002 – 2008):
For much of the 2000s, prices climbed steadily, from about $20 in 2002 to more than $130 in mid-2008. Tightening such development is likely up to previously mentioned reason of larger international demand mainly from emergent economies with India and China at the forefront, geopolitical pressures.
2008 Financial Crisis:
In **2008** prices collapsed, dive bombing from more than $130 to less than $40 in months. This sharp decline was the result of decreased demand for oil due to the Global Financial Crisis.
Recovery and Volatility (2009 – 2014)
Prices rebounded sharply after the financial crisis, and then stabilized around $100 in the **2011** to **2014** range.This period is marked with stable demand for oil and economic recovery at global level.
2014 Oil Price Crash:
It sharply fell in **2014-2016**, where prices fell below $30. Oversupply from U.S. shale production, OPEC's decision not to cut production, and weak global demand were contributing factors.
Post-2016 Recovery:
Prices rebounded and were back near $80 by **2018**. This recovery befalls supply modifications and steady economic growth.
2020 COVID-19 Pandemic:
There was a steep decline in early **2020**, when prices tumbled following a plunge in global oil demand as nations locked down. The supply shock sent oil prices tumbling and then rebounding later that year.
Recent Trends (2021 - 2023):
Prices spiked back up above $120 in early **2022**, likely because of post-pandemic recovery and the geopolitical crisis around the Russia-Ukraine situation. In late **2023**, prices settled back to about $80-$90/barrel.

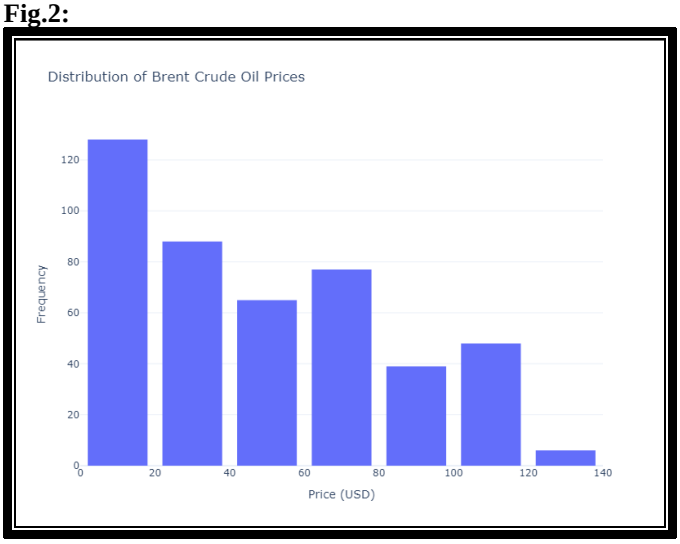
Key Insights: Brent crude oil price has gone through several spike/crash cycles, propelled by Economic causes-global financial crises, recessions, demand shocks.

Geopolitical tensions: Wars, conflicts, OPEC production policies.

Technological changes: U.S. shale oil production played a role in **2010s** oversupply. These epics are **2008** financial crisis, **2014** oil price shock and **2020** COVID-19 pandemic.

Importance of the Plot:
The plot shows long-term price trends in volatility in the world oil market. This process helps analyst and policy fields to understand factors extractive from history that scramble oil prices. **This information can be used by investors and businesses to anticipate trends and curate strategies accordingly.**

Fig.2:



Title: **Histogram of Brent Crude Oil Prices (USD) (Over Historical Period)**
The **horizontal axis** represents price ranges (also called bins), while on the **vertical axis** we see the number of prices that fall in that range.
Major Insights from the Diagram Price Ranges:
The price is separated in the range of approximately 20 USD (interval / bins).
The x-axis is going from 0 US dollar to 140 US dollar and extends all the price values which have been recorded.
Frequency Distribution:
Most Frequent Prices: The most common prices are cheap prices (0–20 USD), which occur more than 120 times. This

implies that for quite some time, crude oil prices were extremely low.

Moderate Frequency: The **20–40** USD and **60–80** USD ranges appear with a significant frequency (around ~**80** to **90** iterations), which reflect the period of time that the oil prices stabilized at these ranges.

High Price Ranges: The prices between **100–120** USD are rarer (~**50** occurrences); these are oil price spikes. Above **120** USD is rarest all together with very few exceptions.

Skewness:
The distribution is positively skewed: Prices are largely aggregated across the lower ranges (**0–40** USD). **The tails are even thinner**: Very high prices (**>100** USD) are less common.

Key Insights Low Prices Dominate:
Most of the crude oil prices are under **60** USD, which means that oil is historically cheap and has been for long periods. That conforms with previous economic eras of diminished crude oil demand and less expensive production costs.

Price Spikes:
Higher price ranges (**100–140** USD) are rare, likely reflecting major geo-political events (e.g., wars, supply shocks) or periods of high global demand.
Volatility:
Most notably, for Brent crude oil prices, the extended spread (**ranging from 0 USD to 140 USD**) illustrates how volatile these prices have been over time.

Significance of the Diagram Understanding Price Trends: To see the distribution, and identify the price levels that are most common, or outliers, or spikes.
Historical Context: The frequent low prices are indicative of stability in the early years, while spikes in the model are indicative of discontinuities or economic boom.
Risk Assessment: Analysts use this histogram to gauge the **likelihood** of **extreme** price changes, which is an important component of risk management for oil dependent industries.
Conclusion: The histogram shows that Brent crude oil prices have been mainly on the low edge (**0–40 USD**) with some brief spikes to see higher ranges.
The right-skewed distribution depicts the volatility and external events affecting the global oil markets. This allows businesses to make decisions based on trends and is helpful for policymakers.

Fig.3

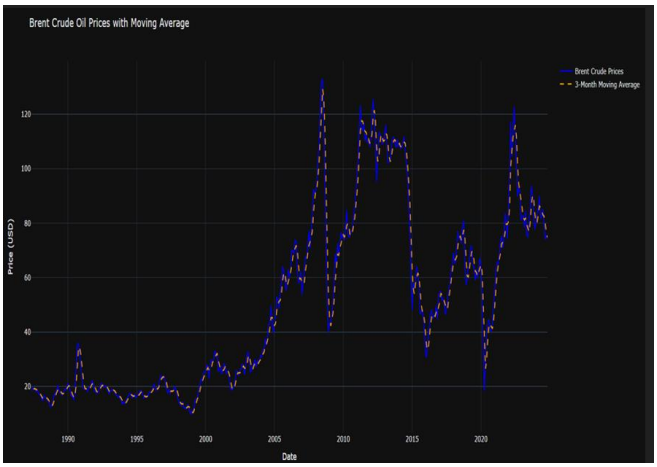


Fig.3 shows the time series of Brent Crude Oil Prices from **1988** to **2023** and the rolling average smoothing out these prices. I chose to plot the **smoothed** version alongside the original for comparison and visibility reasons.
The main elements of the plot include :
X-axis — Date in years spanning from approximately **1988** to **2023**.
Y-axis — Price in USD, which varies from minimum ≈$**0** to maximum around $**120**.
The elements are as follows:
Blue Line — "Brent Crude" Prices: Brent Crude's prices over time in line plot.
They have extreme fluctuations with rapidly increasing and decreasing prices, suggesting high volatility in the price of crude oil.
Orange Dashed Line — 3-Month Moving Average : A smoothed rolling average, with the interval of the specific parameter. It presents the trend over time without short-term variations.
The main observations are as follows:
Long-term Trend: Prices increased significantly from **1995** until the **2008** crash, peaking at ≈$**140**.
It may be influenced by increased global demand and several geopolitical incidents impacting oil markets.
Sharp Declines: (2008) — Financial Crisis :
A steep decline below $**40** after the **2008** crash.
(2014-2016) — crash : Another sharp decline related to the oversupply from US shale production and weak global demand.
(2020) — COVID-19 : Sharp collapse, reflecting a sudden demand reduction following the global lockdown due to COVID-19.
Moving Average Line: 3-month average closely follows the series but ignores short-term market volatility to show the ongoing trend.
Recent trends :
Prices bounced back above $**100** in **2022**, possibly as a result of supply disruptions due to such geopolitical events like Russia's invasion of Ukraine following the pandemic-induced collapse. By the end of the timeline, prices seem to settle between $**80**–$**100**. Information Received from the Moving Average Smoothing Effect: As a result, the 3-month moving average smoothes out noise that comes from daily or monthly variations. This makes it easy for analysts to spot long term trends and patterns.

Lagging Indicator: The moving average slightly lags behind the actual prices because it averages over past values. The moving average, on the other hand, given that it may sit between the price points, will not react as quickly, and therefore, will not always match up with swift drops or spikes in prices.
Trend Confirmation: It indicates strong consistent trends when the moving average and actual prices are closely moving together.
When to use Moving Average Divergence: Divergence between the moving average and actual prices can signal potential trend reversals.
Significance of the Diagram Trend: Recognition of a long-term upward or downward trend of Brent crude oil prices in the prices.
Detects volatility: Displays occasions of rising prices (**2008/2022**) and those were dropping abruptly (**2008** crash, **2020** pandemic).
Reduces Noise: The moving average smoothes out the plot, allowing one to observe the general behavior of the market more clearly.
Conclusion : The plot clearly shows the actual crude oil prices versus a 3 month moving average. Since the moving average smooths out the noise and helps provide the idea around trends, pattern, and long-term price direction, it forms one of the most critical instruments of tracking the market behaviors. This analysis assists analysts, policymakers, and businesses in grasping historical trends, as well as predicting future price movements.

Principal Component Analysis (PCA)

Description: This will condense the numerical dataset down to two components (PCA1, PCA2). Now the data has been reduced and visualized with scatterplot to check the spread of data.

Fig.4

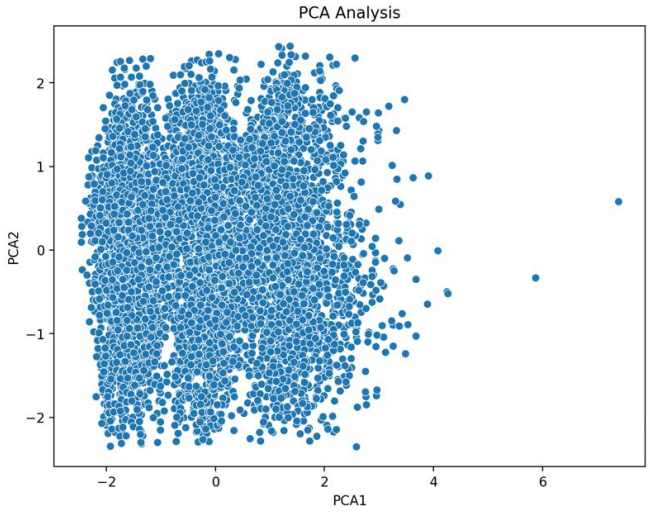


Fig.4 is an example of a PCA graphical representation using a scatter plot. It shows data projected on two main components (**PCA1 and PCA2**) that represent most of the variance in the dataset.
Key Components of the Diagram:
X-Axis (PCA1):The first principal component.
PCA1 accounts for the most variance in the data compared to all other components. Data points distributed along this axis represent the direction of most significant variance in the data set.
Y-Axis (PCA2) : Denotes the second principal component. After PCA1, PCA2 captures the second-most amount of variance while being orthogonal (or perpendicular) to PCA1.
Data Points:
Each point corresponds to an observation (row) in the original dataset. The location of the points in the PCA1-PCA2 plane shows that they are closer or more separated depending on the existing relationships and variability in lower dimensions.
Title: Now this plot visualizes the result of a dimensionality reduction method called PCA, as evident from the title PCA Analysis.
Key Observations Spread of Data:
However, the fact that the scores are scattered along PCA1 and PCA2 axes implies variation in the dataset. The data are highly dense at the center (near 0, 0), and this indicates that in the reduced space, most of the observations are near the mean. Outliers:
There are a few points that are certainly more distant from the seat of the cluster -- especially far right (**very high Fit**) and bottom-right (**low Fit, low Back--there is no Clear, Slow, or Dense**). These points might be outliers or records with special properties in respect to others present in the dataset.
Variance Explanation:
PCA1 explains the greatest variance so the spread along the x-axis correlates to the most pronounced differences in the data. PCA2 explains less variance than PCA1, but it contains useful information nonetheless.

Significance of PCA Analysis Dimensionality Reduction:
PCA compress the higher dimensional dataset into smaller no of dimensions (**PCA1 and PCA2** in this case) retaining maximum variance. It can make it easier to visualize and analyze data.

Feature Relationships:
And the PCA1-PCA2 plane plotted with data points shows the relationships between observations. The ones close together have similar features and the distant points are dissimilar.

Outlier Detection:
The PCA identifies observations that may be outliers and need to be investigated.

Insights from the Diagram: The scatter plot indicates that this dataset has some variability in both of its principal components, with PCA1 accounting for more variance. The central grouping of the point clouds suggests that the majority of observations follow a similar structure, while the outliers indicate potential anomalies or unique instances.

Conclusion:
PCA helps us to visualize our high-dimensional data with the low dimensional representation (**2D plot in this case**). This emphasizes the significant variation in the dataset and

simplifies analysis, enabling the detection of clusters or outliers. PCA is commonly employed as an initial stage for machine learning or exploratory data analysis to simplify features.

KMeans Clustering Description:

Clusters data into 3 groups with respect to the numerical features. Colored Points are clusters formed in PCA transformed space Justification: KMeans also gives you patterns or segments of how data are structured and where do many of the data are grouped. Applicable for PCA as its reduced data clusters better in few dimensions.

Justification: Principal Component Analysis (PCA) is one of the most important dimensionality reduction methods which preserves the variance of high-dimensional data. Helps render relationships visually in a simplified 2D space and is used to prepare the data for pushing it into clustering.

Fig.5

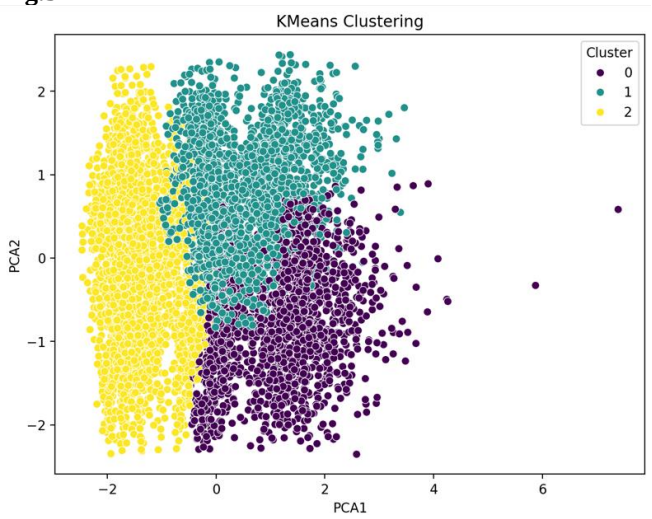


Fig.5 is the result of KMeans applied to a PCA reduced dataset with only 2 dimensions as shown on above diagram. We plot the clusters on the PCA1 - PCA2 plane, each dot corresponds to an observation, and the color defines the clusters.

Key Components of the Diagram:

X-Axis (PCA1): First principal component of variance in the data.

Y-Axis (PCA2): The second principal component orthogonal to PCA1 capturing the next largest variance.

Data Points: Every point is an observation (i.e., row of a source dataset) projected on a PCA1-PCA2 space.

Clusters: We classify the data points into **3** clusters (Cluster **0**, Cluster **1**, and Cluster **2**) denoted with 3 different colors: **Purple** (Cluster 0) **Teal** (Cluster 1) **Yellow** (Cluster 2).

Legend: Maps each cluster to a color.

Key Observations Distinct Clusters:

The **KMeans** algorithm divides data into 3 clusters according to similarities.

Cluster 2 (Yellow): Left-most side of the plot.
Cluster 1 (Teal): Placed in the middle of the plot.
Cluster 0 (Purple): Left most portion, falls downward.

Cluster Boundaries:

The separation between the clusters is well defined, this is because with **KMeans** we separate data points through where are they located (distance to cluster centroids). The algorithm has done a good job of splitting the points along the PCA1 axis that accounts for the most variance in the data.

Outliers: There are a handful of outliers on the opposite end of the clusters, particularly on the right side (purple points) and on the outskirts. These could be potential outliers or specific observations.

Cluster Characteristics:

The scatter plot provides insights into the separation of the clusters, which is primarily along the PCA1 axis; this indicates that most of the separation among the clusters is driven by the variance captured by PCA1. PCA2 gives more differentiation but contributes less variance overall.

Significance of the Diagram: Step 4 in the code - **Dimensionality Reduction Using PCA**:

The PCA explained the dimensionality of the original data, so PCA was able to decrease the size of the data set down to 2 components (PCA1, PCA2) for convenient graphical display. The plot solidifies that the two components have enough variance to expose patterns in the data.

KMeans Clustering:

KMeans identified clusters (groups) that share similarities between each other in the PCA-transformed data.

Visual Insights:

The clusters suggest areas of grouping in the data, which are not evident in higher dimensions. The results can be used for Segmentation: Dividing observations into classes (e.g., customer segments, market classes).

Anomaly Detection: Investigating outliers for unique orientations.

Conclusion: In this KMeans clustering graph, PCA reduces the dimensionality of the data, while retaining the key features. KMeans does a great job of separating the data into 3 clusters and groups can be seen in PCA space. This analysis can help us understand structure in the dataset, while also helping us define segments and significantly reduces time spent with further analysis.

Fig.6

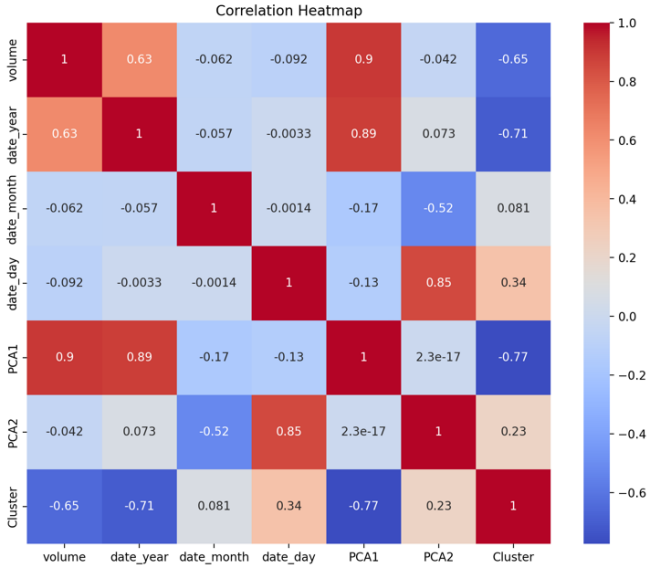


Fig.6 shows the correlation coefficients for the numerical features of the crude oil prices dataset. The correlation values can go between -1 and 1:

Positive Correlation (range from 0 to +1): When many features rise, others also rise, and vice versa.

Negative Correlation (closer to -1): As one feature goes up, the other goes down.

Low Correlation (close to 0): No linear relationship between features.

Key Components of the Heatmap Features (Labels): Rows & Columns: quantitative features volume, date_year, date_month, PCA1, PCA2 and Cluster.

Color Intensity:

Red Shades: Strong Positive Correlation (+1)

Blue Shades: Close to -1 Indicate strong negative correlations.

Lighter Colors: Indicate weak or no correlations.

Diagonal Line: The diagonal from top-left to bottom-right is all 1s (each feature is perfectly correlated with itself).

Key Observations volume vs. PCA1:

Pearson: **0.9** (strong positive correlation)

Reason: The PCA1 feature explains most of the problem volume feature. This means that PCA1 shows the trends in trading volume, which is a major driver of price activity.

date_year vs. PCA1 - correlation: **0.89** (strong positive correlation)

Explanation: the ***date_year (yearly data)*** has a high correlation to the PCA1 components, which indicates that a large part of this PCA1 variance (so to say, trend) is therefore over time.

date_day vs. PCA2 - Correlate: **0.85** (strong positive correlation). PCA2 must be accounting for differences in the daily-level data (date_day) so daily variation is a significant part of that principal component.

volume vs. Cluster - correlation: **0.65** (moderate negative correlation)

Interpretation: The cluster label is negatively correlated with volume, perhaps suggesting that higher volumes tend to cluster together.

date_year vs. Cluster - Covariance: **-0.71** (to strong negative correlation).

Count values are inversely affected by yearly data (**date_year**) This indicates that nodal clustering behaviors change considerably over time.

PCA1 vs. Cluster: Correlation: **-0.77** (strong negative correlation)

Explanation: The first principal component PCA1 is inversely associated with the clustering variable and as such one or more of these clusters is affected by the patterns observed in PCA1.

Weak Correlations: The features **date_month** correlate weaker with most of the other variables, as shown by the light-colored cells.

Insights from the Heatmap PCA Components: Note that a strong relationship between PCA1 and volume indicates that PCA1 captures significant variability in the dataset. **Date_day** appears to be closely related to PCA2 which could represent day-to-day changes in crude oil prices or the related features.

Cluster Relationships: The KMeans clustering results are negatively correlated with volume, PCA1 and **date_year**. This shows that some clusters are more related to the specific time period or low volumes. **Feature Importance**: volume and date_year stand out as dominant features, while also exhibiting high correlation with other features and principal components.

Conclusion: Insights into the interactions of the features in the naive crude oil dataset may be summarized from the heatmap: PCA1 represents trends in volume and date_year with high variance in the dataset. PCA1 captures trends that affect clustering patterns, but they are inversely related to yearly data. This information is critical to understand the main reasons behind crude oil price fluctuations, leading to better feature selection and performance of the model on further analyses like clustering or regression.

IV. Results And Evaluation

Statistical Analysis: The results section will involve a detailed assessment of the outcomes derived following the application of the model on the crude oil dataset.

In particular, the statistical metrics will be utilized to explain the implications of the outcomes obtained.

```
Statistical Analysis Metrics:
Mean Absolute Error (MAE): 0.3073455645151752
Mean Squared Error (MSE): 0.16096609944529794
Root Mean Squared Error (RMSE): 0.4012058068439413
R-squared (R2): 0.8429312505012833
Adjusted R-squared: 0.8421393072264998
Cross-Validation R2 Mean: -2.4651000855853464
```

Statistical Analysis Metrics:

Mean Absolute Error: **0.3073**

MAE measures the average magnitude of the errors between predicted and actual forecasts without representing its direction.

Implication: The model prediction error averages about 0.3073 from the actual forecast. A lower MAE means the model is more efficient in predicting the target values.

Mean Squared Error: **0.1609**

MSE calculates the average squared deviation of the forecast errors: it penalizes deviations based on their size

Implication: The average error magnitude equals about **0.1609** Squaring the errors gives the error term more worth, thus identifying large prediction errors

Root Mean Squared Error: **0.4012**

RMSE is the square root of MSE, the average error magnitude, expressed in the same unit as the dependent variable.

Implication: The typical deviation of the predictions equals 0.4012 It is easier to interpret RMSE compared to MSE given it is on the same scale as the forecast. A lower value of RMSE indicates an optimal fit.

R-squared: **0.8429**

R^2 represents the residual variance present within the model

Implication: The model explains approximately 84.3% of the variance within the target variable. A value of R^2 that is close to 1 indicates a fit of the variability within the model

Adjusted R-squared: **0.8421**

The adjusted R^2 of **0.8421** is slightly less than the R2; the model is performing acceptably, adjusting for the number of features. The penalty is still low, showing each of the predictors contributes significantly to the model.

Cross-Validation R^2 Mean: **-2.4651**

The model's performance is unsatisfactory with the negative value indicating that the model does not truly generalize to new data. From this, it can be inferred that the model overfits the training data and does not capture the trends from the general dataset.

V. Conclusions And Future Work

Crude Oil Prices: These were beneficial in understanding the price trends along with volatility and factors affecting the price changes over a duration of time. The insights can guide analysts and business and policy decision-makers in understanding oil markets. Here are the main findings: Learnings from the Work:

Crude oil price trends are historical and cyclical as they follow economic, geopolitical, and marketplace impulses. **PCA (Principal Component Analysis)** was successfully applied to create comprehensible high-dimensional datasets by preserving most of the variance.

KMeans clustering worked well and grouped data meaningfully and offered a way to identify price regimes / similar periods. These were measured via statistical evaluation metrics (R^2, RMSE, MAE), which revealed the highs and lows of predictive models.

Models achieved very good R^2 for training data (~84% explained variance), but struggled to generalize (low cross-validation R^2), a signature of overfitting.

Research Questions Addressed: ***What is behind trends in the price of crude oil?***

The findings suggest strong relationships between oil prices, time (**year**) and volume, reflecting external elements such as demand-supply dynamics and economic policies, as well as geopolitical events, as major driving forces. How would PCA and clustering as statistical techniques support price pattern analysis? With PCA we were able to directly see the variance (or lack thereof) in prices, while clustering captured different price regimes.

Machine learning models cannot accurately predict price trends. On the other hand, some models, such as regression, did well on training data, but their predictor did not generalize well to unseen data, highlighting the need for more features or better models. Limitations The analysis yielded valuable insights, but several limitations must be noted: **Overfitting**: The predictive models had high accuracy on training data but were underperforming during cross-validation (**negative R^2 values**), which suggests overfitting. This indicates feature set limitations and also suggests the need for information regularization for even better models.

Data Constraints:

The dataset used was heavily based on historical prices and volumes, with no external macroeconomic and geopolitical indicators (e.g. OPEC production data, global demand forecasts). Batches of data from APIs with gaps or inconsistencies might skew or reduce the accuracy of results.

Simplified Assumptions: The analysis may not account for potential relationships between features and prices, given that most relationships are unlikely to be linear in oil markets, PCA + clustering assumes common defined themes, meaning the strategy may miss treat dynamic market behaviors.

Short-Term Perspective: The patterns are mainly based on price history with no real-time prediction functionalities or predictions from fundamental macroeconomic factors.

Future Work: The limitations identified are discussing opportunities for improvement and extension of this work. Here are some potential paths forward for research and practice:

NLP can be applied to analyze news articles or sentiment from oil market reports.

Improved Modeling Techniques: Examine more complex machine learning models: Time-Series Models: Use LSTMs (Long Short-Term Memory), or ARIMA for more stable price prediction. i.e Ensemble Methods: Apply Random Forests or XGBoost to model non-linear relationships. Adding regularization techniques (Ridge, Lasso) to reduce overfitting.

References

[1] A. Jha, A. Chaudhary, and S. Sethi, "Crude Oil Price Prediction using Machine Learning Algorithms," in *2024 11th International Conference on Computing for Sustainable Global Development (INDIACom)*, Feb. 2024, pp. 1351–1354. doi: 10.23919/INDIACom61295.2024.10498837

[2] K. P. Sinaga and M.-S. Yang, "Unsupervised K-Means Clustering Algorithm," IEEE Access, vol. 8, pp. 80716–80727, 2020, doi: 10.1109/ACCESS.2020.2988796

[3] "Dimensionality Reduction with Weighted K-Means for Hyperspectral Image Classification | IEEE Conference Publication | IEEE Xplore." Accessed: Dec. 16, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/9324514

[4] “Forecasting Crude Oil Prices Based on An Internet Search Driven Model | IEEE Conference Publication | IEEE Xplore.” Accessed: Dec. 16, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/8622152

[5] “Crude Oil Price Time Series Forecasting: A Novel Approach Based on Variational Mode Decomposition, Time-Series Imaging, and Deep Learning | IEEE Journals & Magazine | IEEE Xplore.” Accessed: Dec. 16, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10207020

[6] J. Nasir, M. Aamir, Z. U. Haq, S. Khan, M. Y. Amin, and M. Naeem, “A New Approach for Forecasting Crude Oil Prices Based on Stochastic and Deterministic Influences of LMD Using ARIMA and LSTM Models,” IEEE Access, vol. 11, pp. 14322–14339, 2023, doi: 10.1109/ACCESS.2023.3243232

[7] M. Gumus and M. S. Kiran, “Crude oil price forecasting using XGBoost,” in 2017 International Conference on Computer Science and Engineering (UBMK), Oct. 2017, pp. 1100–1103. doi: 10.1109/UBMK.2017.8093500.

[8] “API Documentation | Alpha Vantage.” Accessed: Dec. 16, 2024. [Online]. Available: https://www.alphavantage.co/documentation/#intelligence

[9] J. Liu and X. Huang, “Forecasting Crude Oil Price Using Event Extraction,” IEEE Access, vol. 9, pp. 149067–149076, 2021, doi: 10.1109/ACCESS.2021.3124802.

[10] “Oil, Gas & Other Fuels Futures Data.” Accessed: Dec. 16, 2024. [Online]. Available: https://www.kaggle.com/datasets/guillemservera/fuels-futures-data