GROUP WORK PROJECT # _1__ MScFE 660: RISK MANAGEMENT
GROUP NUMBER: ________9039____

FULL LEGAL NAME	LOCATION (COUNTRY)	EMAIL ADDRESS	MARK X FOR ANY NON-CONTRIBUTING MEMBER
Christopher Enny Ofikwu	Nigeria	c.ofikwu@outlook.com	
Bharat Swami	India	bharatswami1299@gmail.com	
Chengjie Wang	China	cjay-wang@hotmail.com	

Statement of integrity: By typing the names of all group members in the text boxes below, you confirm that the assignment submitted is original work produced by the group (excluding any non-contributing members identified with an "X" above).	
Team member 1	**Christopher Enny Ofikwu**
Team member 2	**Bharat Swami**
Team member 3	**Chengjie Wang**

Use the box below to explain any attempts to reach out to a non-contributing member. Type (N/A) if all members contributed. **Note:** You may be required to provide proof of your outreach to non-contributing members upon request.
N/A

STEP 1 & 2

STUDENT A: THE PROBLEM THE THESIS ATTEMPTS TO SOLVE

This research addresses three related issues associated with crude oil price prediction because these economic indicators strongly affect global financial stability and market dynamics. Traded models such as GARCH incorporate only limited factors that influence oil prices since they depend on mathematical rules about error variance while ignoring essential factors including macroeconomic variables (GDP growth and CPI) and microeconomic variables (supply and demand) and geopolitical variables (OPEC policies) (page 7). Inaccurate oil price forecasting becomes a challenge for financial professionals who depend on dependable market predictions because this modeling restriction produces reduced predictability (page 7). Research based on the 2008 Financial Crisis demonstrates that economic slowdowns related to oil prices required wider factor integration since it happened before the crisis (page 7).

Using expert knowledge to build oil market structure models proves inefficient along with being prone to errors because of abundant available data and changing market conditions (page 8). Real-time forecasting becomes problematic in a fast-moving market due to the long implementation time and incapable adjustment to recent data changes (page 8).

The current models are deficient in strengthened validation systems for meeting financial market deployment requirements during market volatility and financial distress (page 2). The implementation of high-stakes trading decisions requires rigorous testing including historical backtesting with stress testing, otherwise financial losses become a significant risk (page 10). Alvi presents a PGM-based approach in his thesis as a solution to address these problems with the goal of advancing accuracy in addition to implementing automatic structure discovery and forecast validation for real-world deployment.

STUDENT B: WHY ARE BAYESIAN NETWORKS WELL SUITED TO SOLVE IT?

Tracing Causal Relationships: Bayesian Networks represent variable dependencies which enables their use to solve the first issue of handling diverse factors. A BN represents a directed acyclic graph which uses edges to illustrate direct dependencies (page 15). Through this approach the model factors joint probability distributions as:

$$p(x_1, \ldots, x_D) = \prod_{i=1}^{D} p(x_i|\mathrm{pa}(x_i))$$

The technology allows scientists to establish paths that show how oil supply combined with demand variables and geopolitical situations control price fluctuations (page 3). BNs handle broader inference tasks by integrating physical and macroeconomic market data through Bayes' Theorem for unobservables discovery from observed evidence (page 9 and page 14). The 'Rain-Sprinkler-Wet Grass' scenario on page 16 demonstrates the inference ability of BNs for extracting oil price changes from diverse factors.

Process automation through learning relationships directly from data enables BNs to solve the problem of manual structure construction and its connected errors. The thesis implements Hill Climbing Search algorithm to enable "expert-free" Bayesian Network learning while the model builds its oil market structure in a dynamic manner (page 9 and page 43). Using the combination of graph and probability theories (page 12) the automation system learns patterns without human supervision while adapting to new market or data conditions (e.g., OPEC policy adjustments).

The third problem involving validation and deployment can be effectively solved by the robust inference and testing capabilities of BNs. The researchers employ Bayes' Theorem from page 14 to perform efficient posterior probability calculations which support updated economic prediction methods (page 71). The research validates the model's performance by using HMMs to discretize data while conducting inference testing and error evaluation (42.86% error rate on test data distributed on page 69) and economic distress stress tests (page 3). The model functions effectively for trading purposes through its combination with HMMs for analyzing time series data (page 71).

STUDENT C: What Are the Advantages of Using This Methodology for This Problem?

A holistic modeling design is enabled through BNs because they integrate across supply components from OPEC and non-OPEC producers together with demand segments from OECD and non-OECD consumers with diverse macroeconomic indicators such as CPI and GDP in one unified framework (page 9). The global-macro strategy defined on page 3 demonstrates superior performance than GARCH-type models because it captures these essential interactions to yield more precise forecasts that remain true to actual worldwide market dynamics according to page 71. Predicting recessions alongside evaluating market sentiment requires the knowledge provided by this analysis (page 7).

A feature of BNs lies in their automated structure learning capacity which reduces dependency on expert knowledge while conserving time and eliminating errors (page 8). The thesis explicates that the method needs only selected datasets for its operation but lacks specific technical requirements beyond this (page 71) thus enabling flexibility across massive datasets (e.g., EIA, FRED) and market condition changes (page 71).

The methodology demonstrates strong robustness through its effective backtesting and stress testing (page 10) before deployment in highly volatile financial markets, which becomes a vital advantage. The model's rate of error (67.86% for validation and 42.86% for testing on pages 67-69) and trading simulation results on page 69 demonstrate how this system provides practical reliability which boosts trader confidence for making decisions and minimizes financial risks.

A new approach of combining HMM-discretized time-series with BNs provides both original concepts and flexible capabilities for temporal exploration and forecasting improvement (page 71). New techniques including reinforcement learning (page 72) can be integrated through this model's flexible design to enhance its long-term utility which leads to potential alpha returns in commodity trading (page 72).

Beyond their applications in trading BNs generate structural market dynamics data (page 71) which supports policymakers in developing efficient energy policies that combat economic declines and decrease fossil fuel usages (page 3). The dual capability of this approach increases its worth when used in both economic situations and environmental applications.

STEP 3 student A is the macroeconomic / geopolitical specialist :

Data Collection

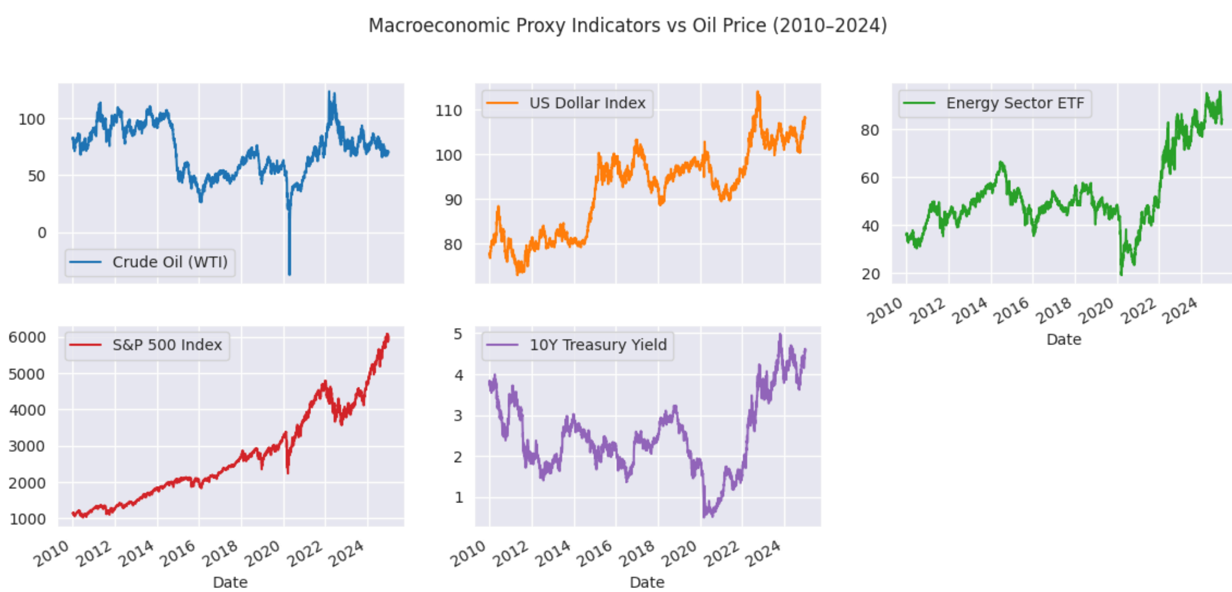
For this analysis, we collected data from the following key sources:

1. **Crude Oil Prices (WTI)**: Sourced from Yahoo Finance using the symbol CL=F, representing the benchmark for oil pricing.
2. **Macroeconomic Indicators**: These include data points such as the S&P 500 Index (^GSPC), US Dollar Index (DX-Y.NYB), and the 10-year US Treasury Yield (^TNX).
3. **Energy Sector Performance**: The performance of the energy sector, as represented by the Energy Sector ETF (XLE), reflects investor confidence in energy companies, which is closely linked to oil price trends.

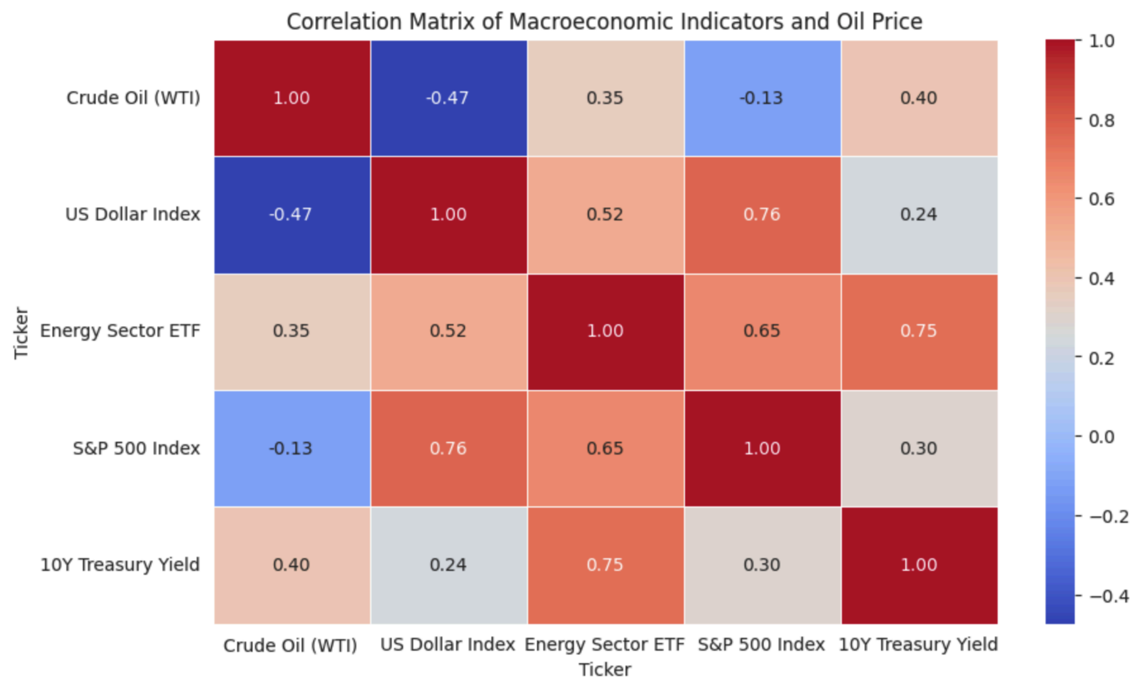
Data Visualization

We used matplotlib and seaborn to visualize the relationships between crude oil prices and the macroeconomic proxies. The data was plotted in a 2x3 subplot grid, with each plot showing the time series of the respective variables. The visualizations revealed important insights into the co-movements between oil prices and the economic factors:

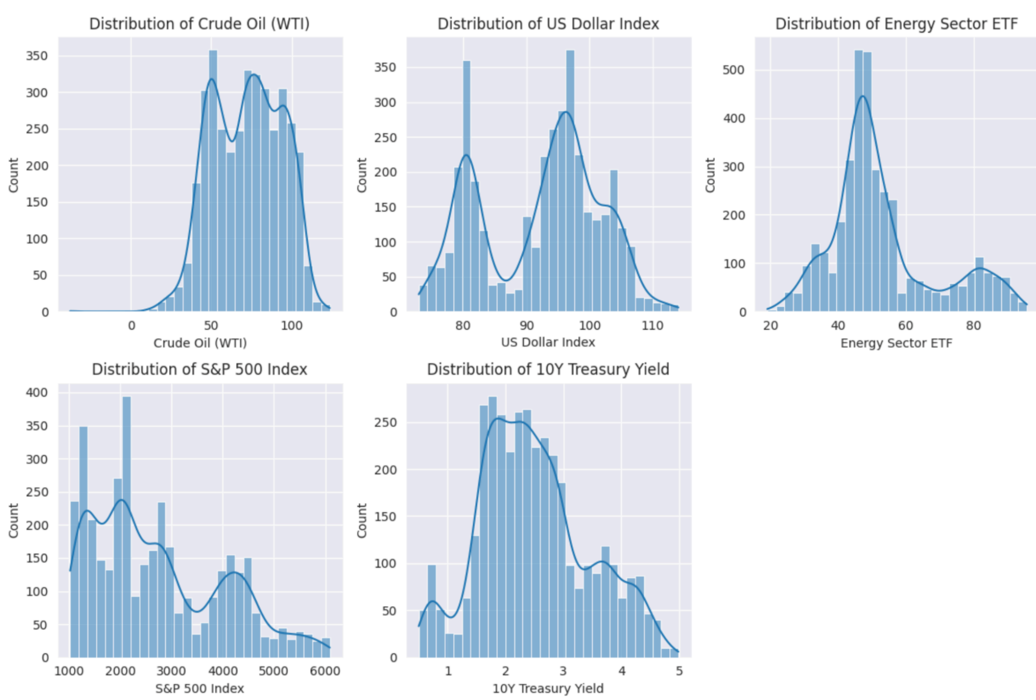
- **Crude Oil Prices (WTI)**: Exhibited significant volatility, reflecting geopolitical events and market reactions to supply and demand shocks.
- **S&P 500 Index**: Showed a positive correlation with crude oil prices, as both tend to rise during periods of economic growth.
- **US Dollar Index**: Displayed an inverse relationship with oil prices, consistent with the general trend that a stronger dollar tends to lower the price of oil.
- **10Y US Treasury Yield**: Demonstrated a potential influence on oil prices, as changes in interest rates can affect investment flows into energy markets.
- **Energy Sector ETF (XLE)**: Followed similar trends to crude oil prices, confirming that the energy sector's performance is closely tied to oil price movements.



The time series plot for each indicators vs Oil Price



Correlation Matrix Heatmap for tickers



Distribution Plot for each ticker

STEP 3 Student B : Microeconomic Data

Microeconomic data refers to data related to individual companies or participants in the overall environment of our problem. And in our case we can say that companies or participants which are responsible for converting crude oil from existing natural state to usable state. For a very higher level view we can define our environment as following map-

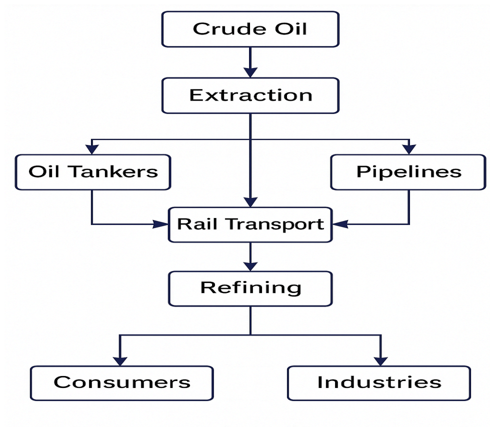


Fig: Map showing crude oil industry environment.

As we can see from above map we are defining the system in 6 sections namely-

1. Extraction: These include companies which are responsible for finding and extraction the crude oil which is present in nature.
2. Transportation: These participants are responsible for transporting the non-refined crude oil to companies or reservoirs which are responsible for refining the product.
3. Refining: They are responsible for refining the crude oil into diesel, jet fuel, gasoline, etc.
4. Marking and Distribution: They are responsible for distribution of these refined products to companies which use or make products from these.
5. Industrial Users: Participants which use these refined products and use them to produce different consumer or petrochemical products for other industrial uses.
6. Service Providers: Participants or Companies which take advantage of vulnerable parts of the whole system and provide services to all other sections.

Right now, we do not know how much each section affects the price of crude oil. But by the end of GWP3 we want to know these dependencies.

5 Example categories of microeconomics data for a company. These categories are namely-

1. Production: Daily Output, CapEx, rig count
2. Financial: Revenue, Net income, Cost/barrel
3. Inventory: Ending inventory, storage capacity
4. Risk: Volatility, short interest, shutdowns
5. Logistics: Export volume, shipping delays

For more complicated and better results we should be using these features for companies but gathering micro economical data for last 20 years is not possible unless one has paid subscription for very know trusted financial data provider.

Now for above reason we are going to use SEC website for microeconomic data. We can find the annual and quarter data for different companies in the 10-K forms (which are required by companies to submit to SEC). This also give this data authenticity since it is submitted to SEC by individual companies.

The main problem we tackle here is all the relevant data are in forms for a particular quarter or fiscal year and we need to extract these data using scraper.

These forms contains lots of different types of information related to companies and there internal structure, we need to exact only the useful information or features which are present in almost form for different companies.

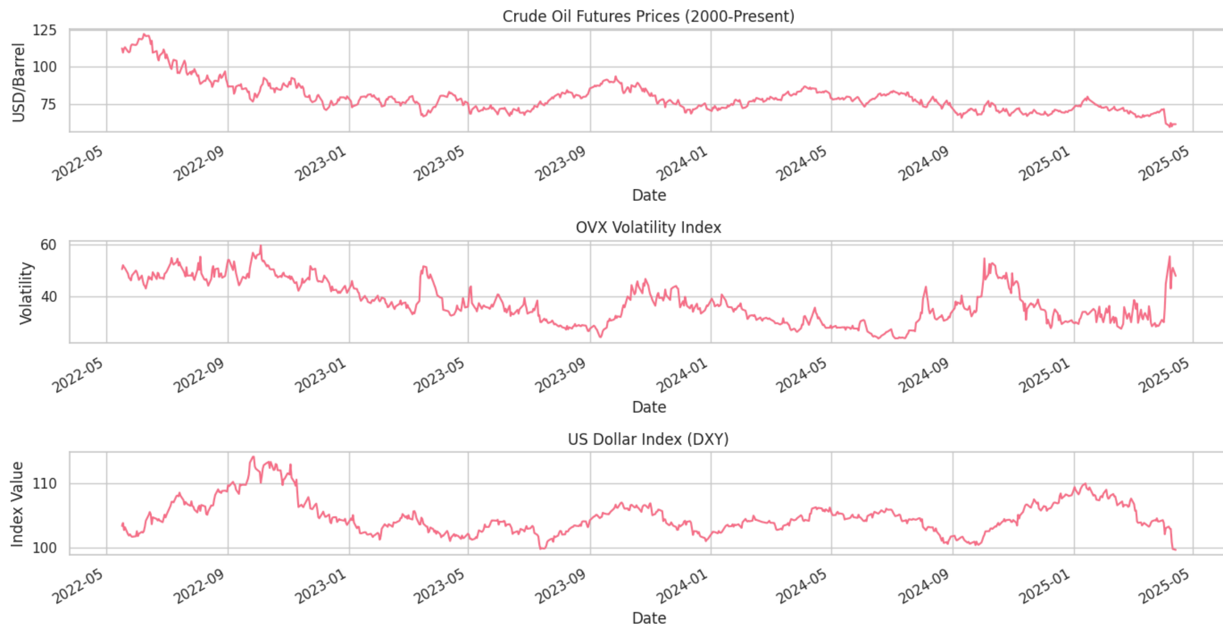
After scrapping and analysing the data we have came up with the conclusion of including below features (with reason why they maybe reasonable to include in our data set.

STEP 3 Student C Financial Data

c) Student C (Financial specialist)

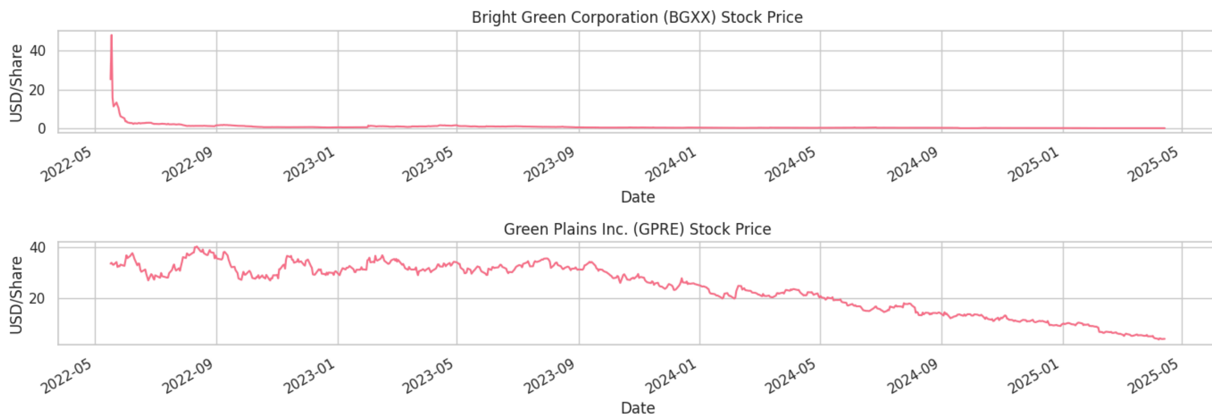
The financial specialist focused on collecting financial data on securities markets for crude oil futures, Bright Green Corporation stock prices, Green Plain Inc stock price and US Dollar Index (DXY). We looked at the energy sector which included stocks from green energy.

The time series of prices and returns extracted for this data from yahoo finance can be graphically visualized as follows:



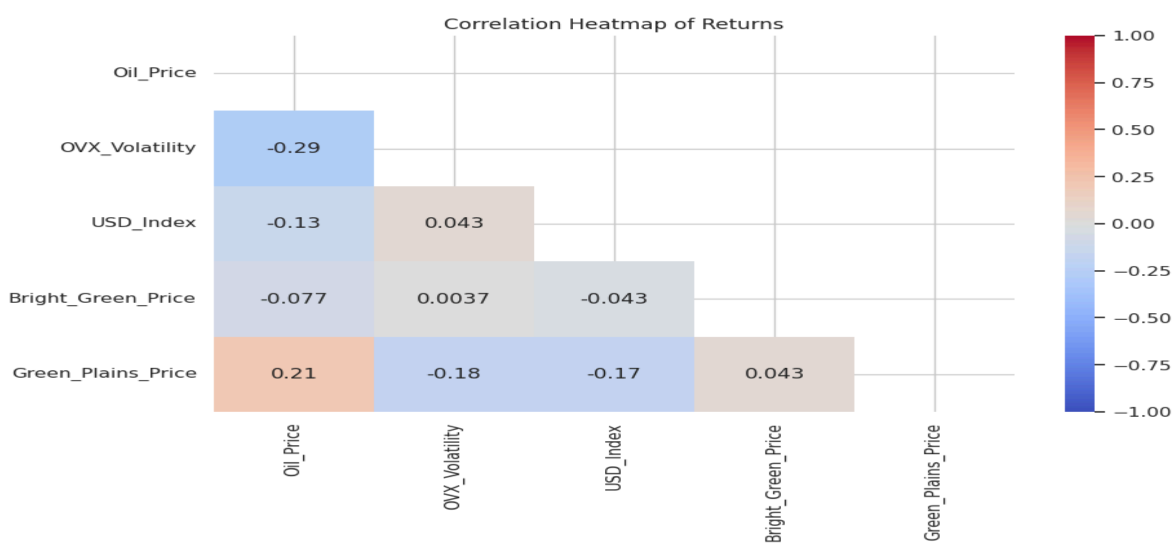
The time series charts show important fluctuation patterns between crude oil prices and both market volatility (OVX) and US Dollar Index (DXY) movements from mid-2022 to early 2025. The crude oil prices appear in recurring patterns yet experience heightened volatility during periods of geopolitical developments such as the Russia-Ukraine conflict in 2022–2023 alongside OPEC+ supply changes during the 2023–2024 period. Market volatility measurements through the OVX Index experience strong upward movements that correspond to events including supply chain restrictions and changes in market demands after the pandemic. The DXY index shows an increasing strength starting in 2023 because Federal Reserve monetary policies tighten while investors show rising levels of risk aversion globally. The relationship between dollar exchange value and oil prices seems suppressed at various times due to structural conditions especially restricted oil stockpiles worldwide and sustained energy needs across emerging economies.

The forthcoming market period of 2024–2025 presents investors with a complex challenge because of merging high volatility indicators (OVX) and stable dollar exchange value measures (DXY). The oil industry depends heavily on geopolitical events alongside energy transition moves yet the strong US dollar can reduce purchasing power in foreign markets. The analysis of these indices must include a complete view: persistent high volatility in oil prices together with a powerful dollar may reveal systemic economic challenges that may require risk reduction strategies. Public authorities need to find stability between fighting inflation through interest rate adjustments which affect the DXY and maintaining energy security from unstable crude markets . Future industry trends underline the necessity of adaptable frameworks that institutions need to implement for navigating their interdependent financial planning and energy policy domains.



Both Bright Green Corporation (BGXX) and Green Plains Inc. (GPRE) followed separate stock price paths since May 2022 through May 2025. BGXX stock faced an immediate downturn in late 2022 after its peak value reached $40 per share before settling at near zero for the rest of 2022 and the entire 2025 period which indicates weak investor confidence potentially stemming from cannabis sector issues or operational problems. The renewable fuels industry player GPRE showed more market fluctuations from $30 per share in 2022 through variable trading between $20-$40 until it declined to under $10 levels by 2025. Both companies maintain a steady decline in stock price which suggests the existence of sector-related or macroeconomic market forces that affect their performance throughout this timeframe.

Correlation Heatmap



The return correlation heatmap demonstrates subtle relationships between all included variables. The correlation analysis indicates between Green Plains Price (GPRE) and Oil Price shows an 0.21 value which demonstrates an increasing oil price benefits GPRE because the company operates in the renewable fuels sector and higher oil prices encourage biofuel consumption. Oil Price data indicates an opposite relationship (-0.29) with OVX Volatility since both variables move in opposite directions even though such patterns frequently appear in commodity markets.

STEP 4 DICTIONARY

Variable	Description	Source	Freq	StartDate	End-Date	Ticker
Crude_Oil_Futures	Front-month Crude Oil Futures Contract (WTI)	Yahoo Finance	Daily	2000-01-01	2025-04-15	CL=F
OVX_Volatility	CBOE Crude Oil Volatility Index	Yahoo Finance	Daily	2000-01-01	2025-04-15	^OVX
USD_Exchange_Rate	US Dollar Index (DXY) measuring USD strength	Yahoo Finance	Daily	2000-01-01	2025-04-15	DX-Y.NYB
Bright_Green	Bright Green Corporation stock price	Yahoo Finance	Daily	2000-01-01	2025-04-15	BGXX
Green_Plains	Green Plains Inc. stock price	Yahoo Finance	Daily	2000-01-01	2025-04-15	GPRE

Microeconomic Data

All the data is downloaded from the 10-K from the SEC website.

Features / Property / Index	Reason
Revenues	Captures the scale of operations (demand proxy)
NetIncomeLoss	Profitability, affects reinvestment & stability
CashAndCashEquivalentsAtCarryingValue	Liquidity snapshot
NetCashProvidedByUsedInOperatingActivities	Core operating health
NetCashProvidedByUsedInInvestingActivities	Capital expenditures
NetCashProvidedByUsedInFinancingActivities	Debt/equity financing mix
Assets & LiabilitiesAndStockholdersEquity	Total balance sheet size
StockholdersEquity	Residual claims value
PropertyPlantAndEquipmentNet	Capital intensity of the business
DepreciationDepletionAndAmortization	Asset aging, CAPEX required
InterestExpense	Debt burden
LongTermDebt	Leverage structure
LongTermDebtNoncurrent	
LongTermDebtCurrent	
RetainedEarningsAccumulatedDeficit	Reinvestment vs. payout behavior
IncomeTaxExpenseBenefit	Tax efficiency, effective rate
OperatingLeaseLiability / OperatingLeaseRightOfUseAsset	Obligations from leasing (important post-IFRS 16)

And Now below are the different companies we are choosing in different sections for our analysis.

1. Integrated Oil & Gas: These companies involve tasks like production, reffing the natural component, distribution, etc

Ticker	Company Name
XOM	ExxonMobil
SHEL	Shell
TTE	TotalEnergies
BP	BP
CVX	Chevron
EQNR	Equinor
E	Eni
OXY	Occidental Petroleum
CVE	Cenovus Energy
PCCYE	PetroChina

2. Refining and Marketing:

Ticker	Company Name
MPC	Marathon Petroleum
PSX	Phillips 66
VLO	Valero Energy
DINO	HF Sinclair
DK	Delek US Holdings

PBF	PBF Energy
SUN	Sunoco

3. E&P (Exploration and Production):

Ticker	Company Name
COP	ConocoPhillips
DVN	Devon Energy
EOG	EOG Resources
EC	Ecopetrol
WDS	Woodside Energy

4. Oilfield Services and Equipment

Ticker	Company Name
SLB	Schlumberger
BKR	Baker Hughes
HAL	Halliburton
TS	Tenaris
PPSI	Pioneer Power Solutions

5. Midstream (Transportation and Pipelines)

Ticker	Company Name
ET	Energy Transfer
EPD	Enterprise Products Partners

PAA	Plains All American Pipeline
WKC	Westlake Chemical (midstream/chemicals)
ENB	Enbridge
OKE	ONEOK
TRGP	Targa Resources
KMI	Kinder Morgan
CQp	Cheniere Energy Partners
TRP	TC Energy

6. LNG and Gas focused

Ticker	Company Name
LNG	Cheniere Energy

7. Energy Holding , Big Companies in Oil, etc

Ticker	Company Name
IEP	Icahn Enterprises
GLP	Global Partners LP
CEG	Constellation Energy
IDOY	Idemitsu Kosan CO

Now we are stating the reason why above companies or organisations are included into the list. So we have taken the largest 100 companies in the Oil and gas sector revenue wise, and then divided into the respective subsections. And also data for these companies are also easily available on SEC website.

STEP 5:

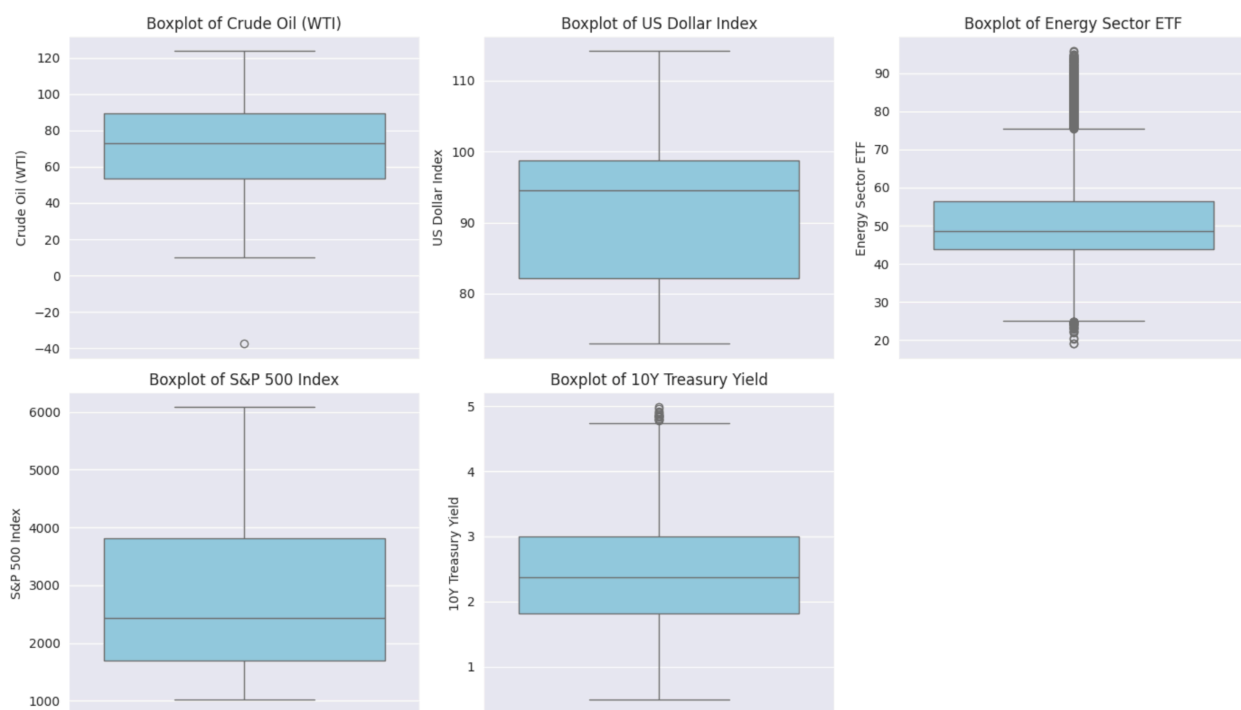
Missing Values

S/N	Reason	Variable	Count	Percentage
	Outlier	**Bright Green Price**	**48**	**6.58%**
	Outlier	**Oil Price**	**39**	**5.35%**
	Outlier	**USD Index**	**26**	**3.57%**
	Unfixable Missing	**All**	**45**	**6.17%**

Extreme outlier

In this step, I mainly focused on identifying and handling outliers in the dataset using statistical methods. Outliers are values that significantly deviate from other data points and could distort the results of statistical analysis or machine learning models. Therefore, I used boxplots, z-scores, and the 3 standard deviation rule to detect and replace the outliers. Here is a summary of the steps I took:

1. Visualizing Outliers with Boxplots: First, I visualized the distribution of each variable using boxplots. Boxplots can clearly show the data spread and help identify extreme values (outliers) that fall outside the whiskers of the box, which typically represent the 1.5 * Interquartile Range (IQR).
2. Identifying Outliers using Z-Scores: Values with a z-score greater than 3 (or less than -3) were considered outliers and flagged for further analysis. I displayed the rows containing outliers to identify the extreme values that could distort the dataset.
3. Replacing Outliers Using the 3 Standard Deviation Rule: After identifying the outliers, I chose to replace them using the 3 standard deviation rule. For each variable (in this case, the Energy Sector ETF), I calculated the mean and standard deviation, and defined the upper and lower bounds as mean ± 3 * standard deviation. Any data points that fell outside this range were considered outliers and were replaced with values within the bounds of mean ± 3 * standard deviation. This method ensures that extreme values are corrected without distorting the overall distribution of the data.
4. Resulting Dataset: After replacing the outliers, the dataset was cleaned, and the values were adjusted to a reasonable range. This cleaning process ensures that extreme values no longer have an undue influence on subsequent analyses and model training.



Boxplot for each ticker

Bad Data

Now the missing and wrong data in microeconomic data:

- We have to completely trust our data because we have extract the data from official SEC website and if those data have any wrong value then its filled companies fault to submitting the wrong data.
- Now missing data, we donn't have missing data since all the microeconomic data is either quarterly or annually based and again companies are required to submit these to SEC on regular basis.
- Now the most important part of data preprocessing of microeconomic data from SEC is that we have multiple duplicate data for same time period we just need to remove those, and which we have done in the attached python notebook.

STEP 6

We have combined our all the different section data here into one dataset. This dataset contains-

- Macroeconomic data: Data which are related to microeconomic factors related to crude oil ecosystem, which are most extracted oil countries, countries GDP and other rate related to OPEC and NOn-OPEC countries, etc
- MicroEconomic data: Here we gather data related to individual members of crude oil ecosystem all largest revenue companies related to oil (directly or indirectly). This includes NetIncomeLoss, etc features listed in step 3 microeconomic table.
- Financial data: These include all the exchange traded derivatives of crude oil like oil futures etc.

Note: We may include the stock price of individual companies from Microeconomic data in Financial data, this will be done in future for further improvement of the our prediction model.

Now we have gathered the Microeconomic and Macroeconomic data on annual and quarterly basis while our financial data are in daily frequency. we are going to create a model which will take data in this frequency and changes it along the way of process. This intensive computation is needed because all the economica data doesn't make any sense for daily frequency, for example, Countries GDP are quarterly or annually dase not of daily frequency based, while stock price for option based on crude oil are on daily frequency.

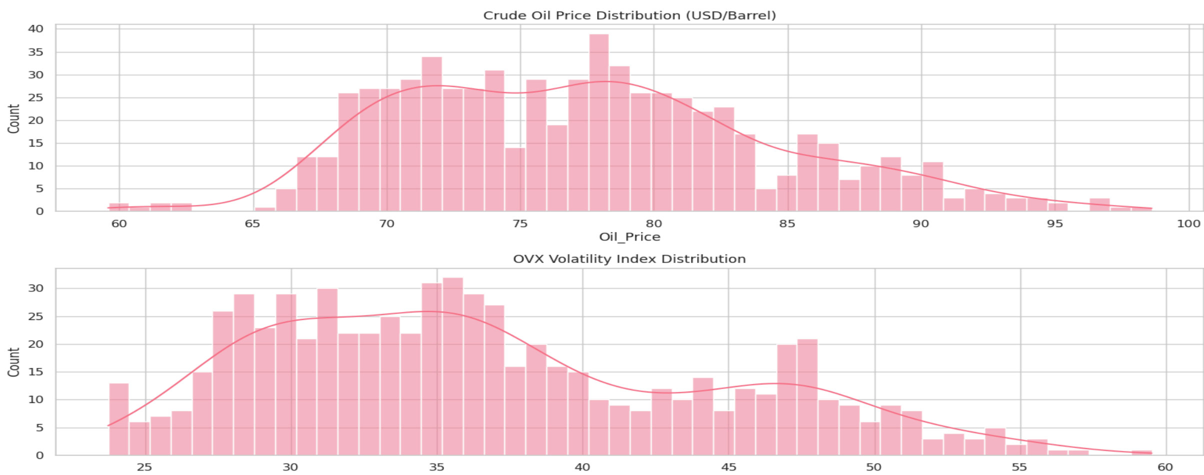
Now this section concerns about removing or discarding some particular data. Below are some points we took as notes during discussion and based on these we discarded some data, these are-

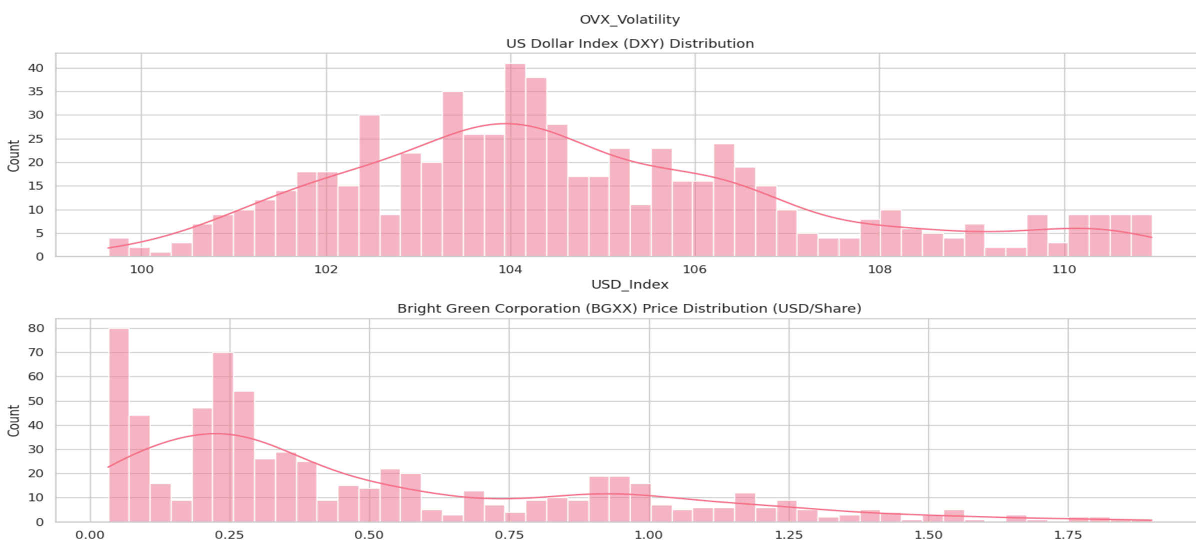
- If we want we can create a database of all the macroeconomic data related to OPEC and Non-OPEC countries but we don't need this heavy amount of data, only the top five to seven companies are sufficient enough.
- We have discarded half the companies from the 100 largest revenue crude oil companies when we categorize them into sections and went forward with 40 companies from 6

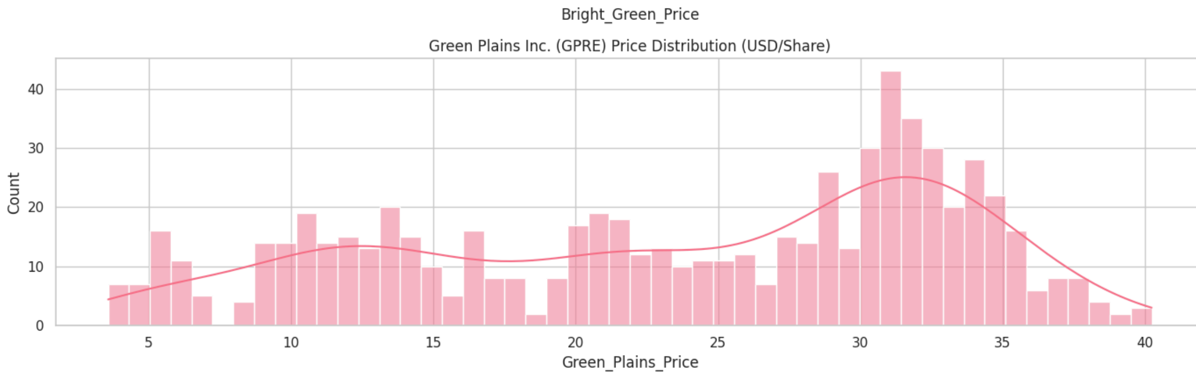
subsections of crude oil.

- Similarly we are not going to use all the exchange traded crude oil derivatives for our project, we are only considering the more correlated and volume generating derivative with crude oil and gas.

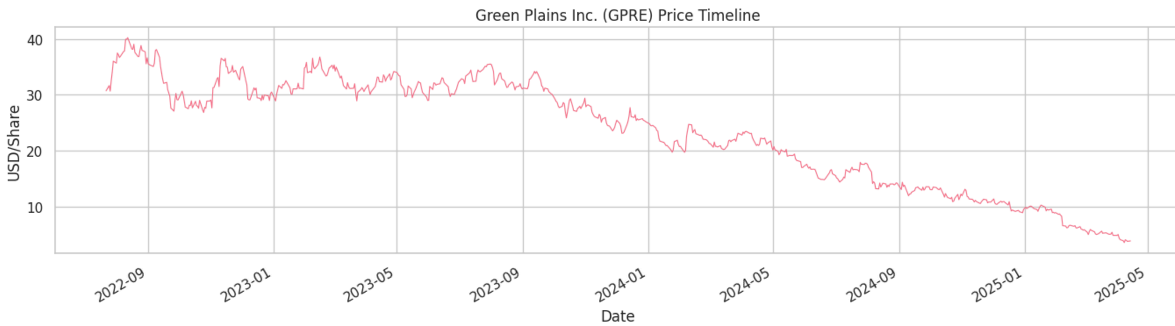
STEP 7



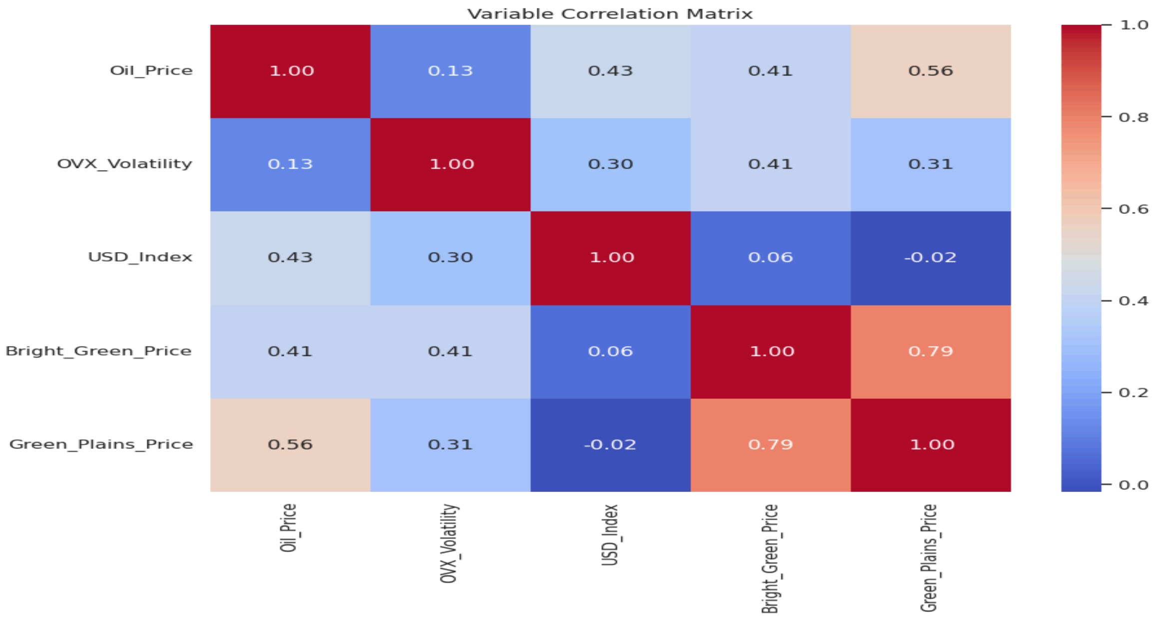




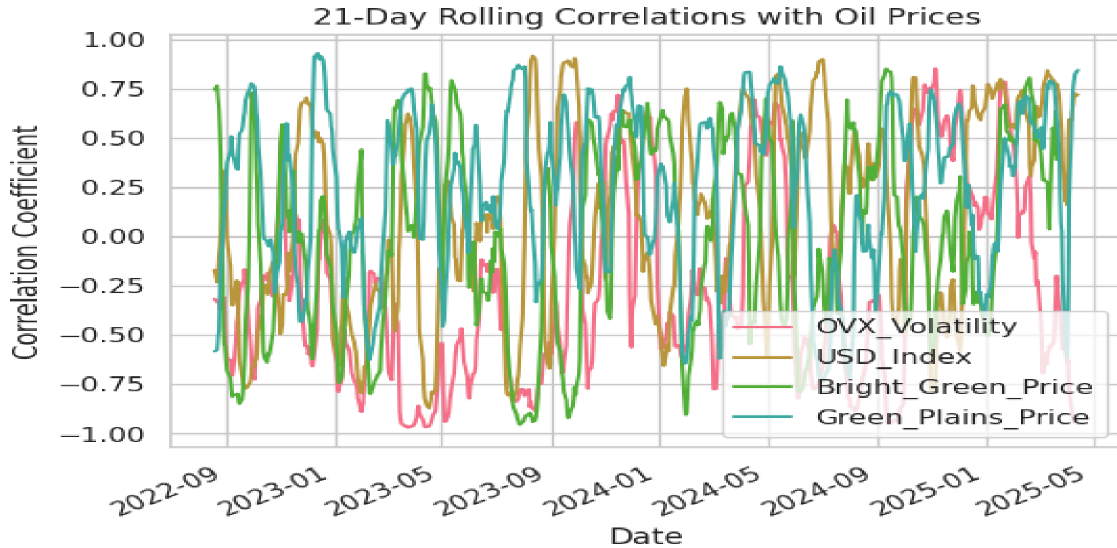
TIME SERIES PLOTS




```
Generating Multivariate Analysis...
Correlation matrix saved
```



```
Rolling correlations saved
<Figure size 1400x600 with 0 Axes>
```



STEP 8

(a) What makes oil prices look different from other asset prices? E.g., spikes, clustered volatility, seasonality, etc.

The behavior of oil prices shows distinctive features which make them separate from standard financial instruments including stocks and bonds. The prices of oil experience quick dramatic increases because of geopolitical events which rarely occur in other market sectors. The supply decisions of OPEC coupled with hostilities in oil-producing regions such as the Middle East cause prices to increase rapidly. Temporary high market uncertainty persists after specific disruptive events due to market imbalances combined with speculative market activities. The price stability of equity markets post earnings reports does not apply to oil as it requires physical supply chains to adjust before market stability occurs. The oil market exhibits forecastable yet powerful price shifts during seasonal periods when people use more heating oil during cold months and drive during hot months although most financial instruments lack this feature.

The physical limitations of oil storage as well as its distribution process create additional distinct characteristics. The physical nature of oil distinguishes it from both financial instruments like stocks and currencies because it takes up precious space in storage. The fuel storage capacity approaching its limit during the 2020 COVID-19 market collapse caused prices to drop to negative values which no financial instrument alone could achieve. Centralized oil production regulation through organizations like OPEC controls crude oil markets in contrast to the decentralization of equities and bonds markets. The dual role of oil as a global economic indicator and its limited storage capacity joins forces with production control mechanisms to create market dynamics that make oil prices behave exceptionally complex and reactive compared to other asset classes.

(b) What types of distributions do oil returns have?

Negative Skewness: The left tail which covers negative returns up to -8% extends beyond the right tail which encompasses positive returns that reach only +6% thus demonstrating greater occurrence of larger negative returns. The data shows that oil markets experience more frequent substantial price drops rather than rises which demonstrates their susceptibility to price crashes.

Leptokurtosis (Fat Tails): Most daily returns cluster near 0% while the distribution demonstrates important mass in both tails. Long heavy tails in oil price behaviors reveal its tendency for large fluctuations which is confirmed by the existence of extreme price movements between +4% and -4% despite their lower occurrence frequency.

Volatility Clustering: The high level of returns concentrating on zero value together with excessive outliers indicates a pattern of alternating volatile and calm periods that aligns with the expectations from GARCH volatility clustering.

(c) What type of autocorrelation do the oil returns have?

Recent price movements show strong influence on near-future price effects (initial autocorrelation ≈0.8) because market momentum and delayed supply-demand responses interact. The pattern of decreasing autocorrelation with time shows evidence of long memory structures that distinguish price variability from standard prices and returns as well as stationary price levels. The pattern in the graph demonstrates characteristics of undifferenced data or squared returns rather than daily returns. To create models for this data set price stationarization through differencing must be applied alongside volatility-specific frameworks (e.g. GARCH) to identify risk-based autocorrelations instead of return-based autocorrelations while accounting for potential model specification breakdowns and structural changes in the data.

(d) What other stylized facts can you say about oil prices?

Oil market prices generate an array of stable market periods along with turbulent movements since geopolitical conflicts cooperate with supply-and-demand market forces to trigger price changes. Price increases after supply disruptions exceed price decreases from oversupply situations because production factories need time to adapt their output. Oil market price fluctuations generate distribution patterns characterized by heavy tails known as leptokurtic returns and negative skewness brought about by typical crash patterns with abnormal surge behavior. The market storage characteristics that develop in futures curves with contango or backwardation patterns coincide with volatility patterns which exhibit long-term memory throughout the system. Worldwide economic cycles along with the DXY connect inversely to petroleum prices which display market sensitivities to both economic cycles. The economy experiences price growth during expansion phases and price reduction is typical for recessions. Shale technology together with the energy transition has altered oil price behavior patterns thus creating distinctions between oil as a financial asset.

STEP 9

Student A: Probabilistic Graphical models distinguish belief networks & Markov networks.

PGMs exist as two distinct subfields namely Belief Networks (Bayesian Networks) and Markov Networks (Markov Random Fields). They differ in their path connections and how they model dependent factors. DAGs in Belief Networks represent causal relationships and nodes connect to CPDs as conditional probability distributions for asymmetric relationships. The dissertation of Alvi employs Hidden Markov Models (HMMs) which belong to the Bayesian Network family for modeling crude oil price regimes through directed state transitions and emissions. Markov Networks implement undirected graphs to express symmetric dependencies through potential functions over cliques for spatial or non-causal scenarios yet they do not function well with Alvi's temporal HMM approach.

Student B: Parameter learning and distinguishes it from structure learning.

Parameter Learning: The process of parameter learning generates estimates for conditional probability distributions and parameters that belong to each node within a predefined Bayesian network structure. Each node in a Bayesian network represents a variable while the network connections demonstrate variable conditional dependencies. The CPD of a node establishes its value probabilities when its parent nodes exhibit specific values. According to Alvi (2018) the representation of oil price probability depends on GDP growth and geopolitical events as defined in the CPD model.

Methods:

MLE uses a process that selects parameters that result in the highest possibility of observing the data samples. The CPD demonstrates an oil price rise occurs 70% of the time according to historical data where GDP exhibits growth.

The Bayesian Estimation process brings expert judgments about oil price patterns into analysis while it modifies these beliefs based on actual measurements. The technique becomes effective when dealing with scarce data to reduce the risk of model fitting problems.

This parameter learning approach determines probabilities from the node "Oil Price" whose parent nodes consist of "GDP" and "OPEC Production" by generating results such as P(Oil Price = High | GDP = High, OPEC Production = Low).

Structure Learning: Structure learning identifies the network graph structure of Bayesian networks by determining variable dependencies. There exists a question which this process helps to resolve about which graph edges need to be included. The process becomes harder because it requires solutions from many potential graph configurations to discover the one that fits the data.

Methods:

Statistical tests especially chi-squared (e.g., chi-squared) determine conditional independence relations through Constraint-Based methods. The graph requires no edge to connect two variables if they remain independent when considering selected other variables. PC algorithm represents one of the frequently used examples.

Each potential graph receives a score evaluation through this approach using criteria such as Bayesian Information Criterion (BIC) then selects the configuration achieving the highest score. The big number of potential graphs necessitates heuristic search in this procedure.

Hybrid Approaches: Combine constraint-based and score-based methods for efficiency.

Student C: Markov chains and Markov blankets.

The stochastic process known as the Markov chain responds to the Markov property where the future state transition probability depends solely on present state conditions without needing the order of previous states. Formally, for a sequence of states X1, X2 ………Xn:

P(Xn+1=x|X1, X2, ……..Xn) = P(Xn+1=x|Xn)

The key components in the Markov chain are; Space state, transition matrix and steady-state distribution. The properties are such that; (a) All state communicates (Irreducibility), (b) No Cyclical patterns in transition (Aperiodicity), and (c) Guarantees convergence (Ergodicity).

Markov Blankets: A Bayesian network or probabilistic graphical model defines the Markov blanket of node X as a minimal set of nodes which establishes conditional independence between X and every other node in the network. It consists of (1) Parent which is the direct cause of X; (2) Children which represents the direct effect of X and (3) Spouses which is the other children of Xs children. It is mathematically represented as: FOr a node X, its Markov Blanket MB(X) satisfies: P(X| all other nodes) = P(X|MB(X)).

STEP 10

Algorithm 1: Inferred Causality for Causal Graph Discovery

Input:
- D: Dataset with variables V = {X1, X2, ..., Xn} (e.g., Oil_Price, OVX_Volatility, etc.)
- alpha: Significance level for conditional independence tests (e.g., 0.05)

Output:
- G: Directed Acyclic Graph (DAG) representing causal relationships

// Initialize an undirected graph with all variables fully connected
G ← Undirected graph with vertices V and edges between all pairs of variables

```
Phase 1: Learn Markov Blankets
Function LearnMarkovBlankets(D, V, alpha):
    For each variable Xi in V:
        MB(Xi) ← Empty set  // Markov blanket of Xi
        For each variable Xj in V \ {Xi}:
            // Test conditional independence of Xi and Xj given subsets S of V \ {Xi, Xj}
            For each subset S in V \ {Xi, Xj}:
                If Xi is conditionally independent of Xj given S (p-value > alpha):
                    // Xi and Xj are not directly dependent given S
                    Remove Xj from potential Markov blanket of Xi
                    Break
                Else:
                    Add Xj to MB(Xi)  // Xj is part of Xi's Markov blanket
    Return MB = {MB(X1), MB(X2), ..., MB(Xn)}

// Phase 2: Learn Neighbors (Skeleton of the Graph)
Function LearnNeighbors(G, MB, D, alpha):
    For each pair of variables (Xi, Xj) in V:
        If edge (Xi, Xj) exists in G:
            // Use Markov blankets to limit the conditioning set
            S ← MB(Xi) \ {Xj}  // Conditioning set: Markov blanket of Xi excluding Xj
            For each subset S' of S:
                If Xi is conditionally independent of Xj given S' (p-value > alpha):
                    Remove edge (Xi, Xj) from G
                    Record S' as the separating set for (Xi, Xj)
                    Break
    Return G  // G is now the skeleton (undirected graph)

// Phase 3: Learn Arc Directions (Orient Edges)
Function LearnArcDirections(G, D, alpha):
    // Step 1: Apply v-structure rule to orient edges
    For each triple of variables (Xi, Xj, Xk) in G where Xi–Xj–Xk and Xi not connected to Xk:
        If Xj is not in the separating set of (Xi, Xk):
            Orient edges as Xi → Xj ← Xk  // v-structure
            Mark (Xi, Xj) and (Xj, Xk) as directed

    // Step 2: Apply orientation rules to propagate directions
    While there are undirected edges in G:
        // Rule 1: Avoid creating new v-structures
        For each triple (Xi, Xj, Xk) where Xi → Xj – Xk and Xi not connected to Xk:
            Orient Xj → Xk
```

```
            Mark (Xj, Xk) as directed

        // Rule 2: Avoid cycles
        For each triple (Xi, Xj, Xk) where Xi → Xj and Xj – Xk and Xk → Xi:
            Orient Xj → Xk
            Mark (Xj, Xk) as directed

        // Rule 3: Directed path consistency
        For each triple (Xi, Xj, Xk) where Xi → Xj and Xi – Xk and Xk → Xj:
            Orient Xi → Xk
            Mark (Xi, Xk) as directed

    Return G  // G is now a DAG

// Main Algorithm
MB ← LearnMarkovBlankets(D, V, alpha)
G ← LearnNeighbors(G, MB, D, alpha)
G ← LearnArcDirections(G, D, alpha)

Return G
```

REFERENCES:

1. Alvi, Danish A. Application of Probabilistic Graphical Models in Forecasting Crude
2. Oil Price. 2018. University College London, Dissertation. https://arxiv.org/abs/1804.10869
3. Koller, Daphne, and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.
4. Ross, Sheldon M. *Introduction to Probability Models*. 12th ed., Academic Press, 2019.
5. Russell, Stuart, and Peter Norvig. *Artificial Intelligence: A Modern Approach*. 4th ed., Pearson, 2020.
6. EIA Database, crude oil
7. SEC Website
8. SEC Filling Scraper, Adam Getbags