GROUP WORK PROJECT # 1 MScFE 660: RISK MANAGEMENT
Group Number: 5638

FULL LEGAL NAME	LOCATION (COUNTRY)	EMAIL ADDRESS	MARK X FOR ANY NON-CONTRIBUTING MEMBER
Shailza Virmani	India	virmanishailza@gmail.com	
Zhe Zhang	Hong Kong	zhezhangcs@gmail.com	
Anubhav Mishra	India	anubhav0by0@gmail.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	**Shailza Virmani**
Team member 2	**Zhe Zhang**
Team member 3	**Anubhav Mishra**

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 2

(a) Problem Statement

The problem that the student's thesis attempts to solve is the accurate prediction of the behavior of energy markets, specifically the price of crude oil. The issue revolves around the complication and volatility of oil markets that is influenced by factors such as geopolitical events, economic indicators and market sentiment. The thesis focuses on developing a methodology that uses probabilistic graphical models [i.e. Bayesian networks] to effectively model the complicated relation between these factors and produce reliable prediction of future oil prices.

(b) Suitability of Bayesian Model

The Bayesian networks are well suited to address the challenge highlighted in the thesis for the following reasons:

- Bayesian networks can be leveraged to Probabilistic Graphical Models to predict the likelihood of one or more events at the same time. In this case, it could be the price of crude oil and the spot price
- The causal dependence of the various identified factors like the microeconomic factors, the macroeconomic and the geopolitical factors intertwined together could depict the dynamics of oil price
- The partial sets of causes of crude oil change can be identified such as crude oil supply shocks.
- The factors are interactive. For example, the war in the middle east might lead to demand shocks and unplanned supply disruptions might lead to supply shocks. Using Bayesian networks, we can evaluate those relationships better.
- Based on the thesis, we can discover the probability distribution across potential neural networks where minor changes to the conventional neural networks helps us fairly estimate the inference problem. We further identify the degree of uncertainty in our predictions caused by overfitting over small datasets.

(c) Advantages of the Methodology

- The creation of the Probabilistic Graphical Model(PGM) aids in a better analysis of the market microstructures, the macroeconomic factors without the guidance from experts in the oil markets. The future predictions using the current state of the market aid in better risk management. The possibility of higher alpha generation with the use of an automatic trading system. Unlike traditional models, the PGM Models leverage the Bayesian network which include the microeconomic, the macroeconomic and the geopolitical factors that are significant in determining the price of crude oil which evolve constantly with the changing market dynamics and thus shifting the outcome as a response which proves to be an advantage over traditional models.

- It helps analyze the structure of oil markets without any expert assistance, thus helping market participants, risk managers, and policy makers to have a better understanding of the structure of the energy market.

Step 3

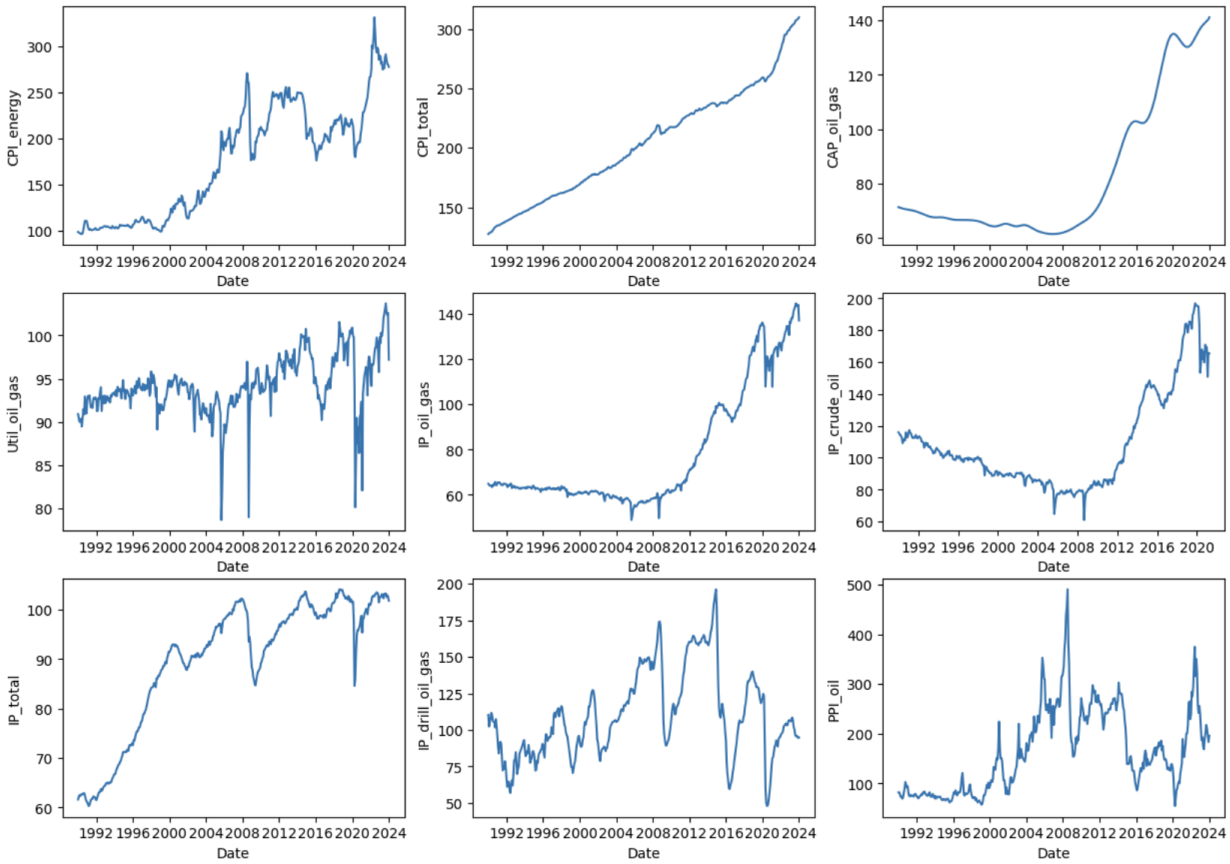
Macroeconomic Data

We collected 9 data sources from Fred including the CPI of the energy sector (CPIENGSL), total CPI index (CPIAUCSL), industry capacity of oil and gas extraction (CAPG211S), capacity utilization of oil and gas extraction (CAPUTLG211S), industrial production of oil and gas extraction (IPG211S), industrial production of crude oil (IPG211111CN), total industry production index (INDPRO), industrial production of drilling oil and gas wells (IPN213111N), and producer price index of oil and gas extraction sector (PCU211211). The plots of these 9 data sources are shown as follows.

We can find that these 9 data sources have a similar trend, that is a drop between 2014 and 2016 and a surge after that, and then a drop in 2020 due to the global pandemic.

GROUP WORK PROJECT # 1 MScFE 660: RISK MANAGEMENT
Group Number: 5638

Macroeconomic Data

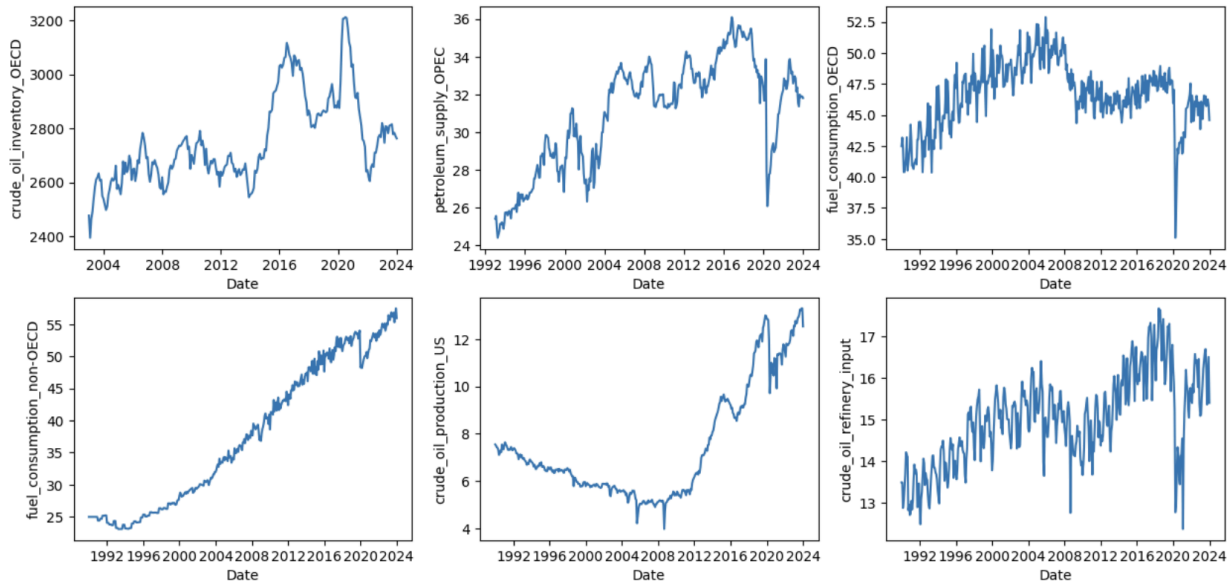


Microeconomic Data

Here, EIA API is used to get microeconomic data for the global oil market.

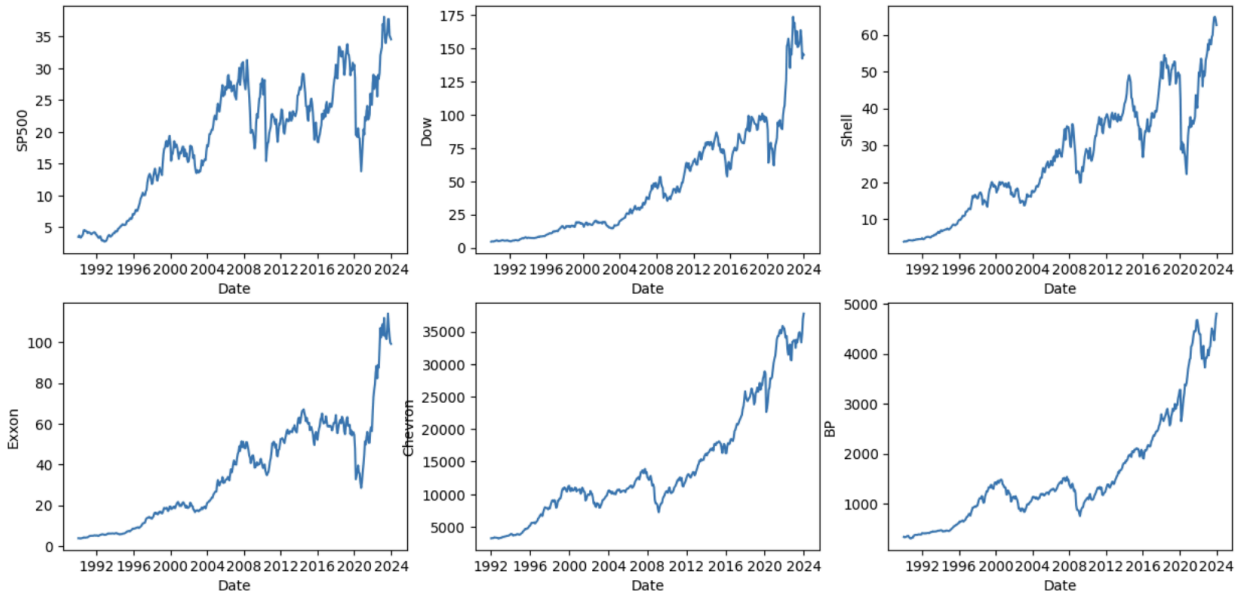
Monthly data for six important indicators is fetched - PASC_OECD_T3, PAPR_OPEC, PATC_OECD, PATC_NON_OECD, COPRPUS, and CORIPUS, which covers aspects like oil production, consumption and trade patterns.

Checked that data is consistent for analysis and organized it to float64 data type which helps in plotting and analysis and finally sorted the data to maintain the order which is crucial for time series analysis.



Financial Data

We collected the index price of S&P 500, Dow Jones and stock prices of four oil companies Shell PLC (SHEL), Exxon Mobil Corp. (XOM), Chevron Corp. (CVX) and BP PLC (BP) using yfinance API then resample the stock price and obtain the monthly mean price for each index and stock.



GROUP WORK PROJECT # 1 MScFE 660: RISK MANAGEMENT
Group Number: 5638

Step 4

Data Dictionary

Name	Ticker	Source	Frequency	Start Date	End Date
CPI_energy	CPIENGSL	Fred	Monthly	1990-01	2024-01
CPI_total	USACPALTT01CTGYM	Fred	Monthly	1990-01	2024-01
CAP_oil_gas	CAPG211S	Fred	Monthly	1990-01	2024-01
Util_oil_gas	CAPUTLG211S	Fred	Monthly	1990-01	2024-01
IP_oil_gas	IPG211S	Fred	Monthly	1990-01	2024-01
IP_crude_oil	IPG211111CN	Fred	Monthly	1990-01	2024-01
IP_total	INDPRO	Fred	Monthly	1990-01	2024-01
IP_drill_oil_gas	IPN213111N	Fred	Monthly	1990-01	2024-01
PPI_oil	PCU211211	Fred	Monthly	1990-01	2024-01
crude_oil_inventory_OECD	PASC_OECD_T3	EIA	Monthly	2003-01	2024-01
petroleum_supply_OPEC	PAPR_OPEC	EIA	Monthly	1993-01	2024-01
fuel_consumption_OECD	PATC_OECD	EIA	Monthly	1990-01	2024-01
fuel_consumption_non-OECD	PATC_NON_OECD	EIA	Monthly	1990-01	2024-01
crude_oil_production_US	COPRPUS	EIA	Monthly	1990-01	2024-01
crude_oil_refinery_input	CORIPUS	EIA	Monthly	1990-01	2024-01
SP500	^GSPC	yfinance	Monthly	1990-01	2024-01
Dow	^DJI	yfinance	Monthly	1990-01	2024-01
Shell	SHEL	yfinance	Monthly	1990-01	2024-01
Exxon	XOM	yfinance	Monthly	1990-01	2024-01
Chevron	CVX	yfinance	Monthly	1990-01	2024-01
BP	BP	yfinance	Monthly	1990-01	2024-01

Step 5

Data Cleaning

Outlier Detection

We utilized the clipping method to clear the outlier. We identify the value as an “extreme outlier” if it falls outside two standard deviations of the mean. To remove these outliers, we replaced them with the closest boundary value within the two standard deviations.

Bad data

We write a function to identify and delete the bad data such as duplicate rows from the DataFrame.

Missing values

We first interpolate the data using the linear method and then remove the rows where not all features have records, such as the rows before 2003-01.

Step 6

We selected the period between 2003-01 and 2024-01 because all features have records in this period, and it includes several regimes that help us analyze the energy markets. Additionally, we removed data points that are more than two standard deviations away from the mean value. We consider these values as extreme values that might introduce additional noise to our analysis; thus, they were removed.

Step 7

Here, we used various visualizations to understand the relationship and characteristics of the dataset. The findings from these visualizations are- we found out how the values are distributed by observing the histograms. It also helped us spot outliers and understand the data spread.

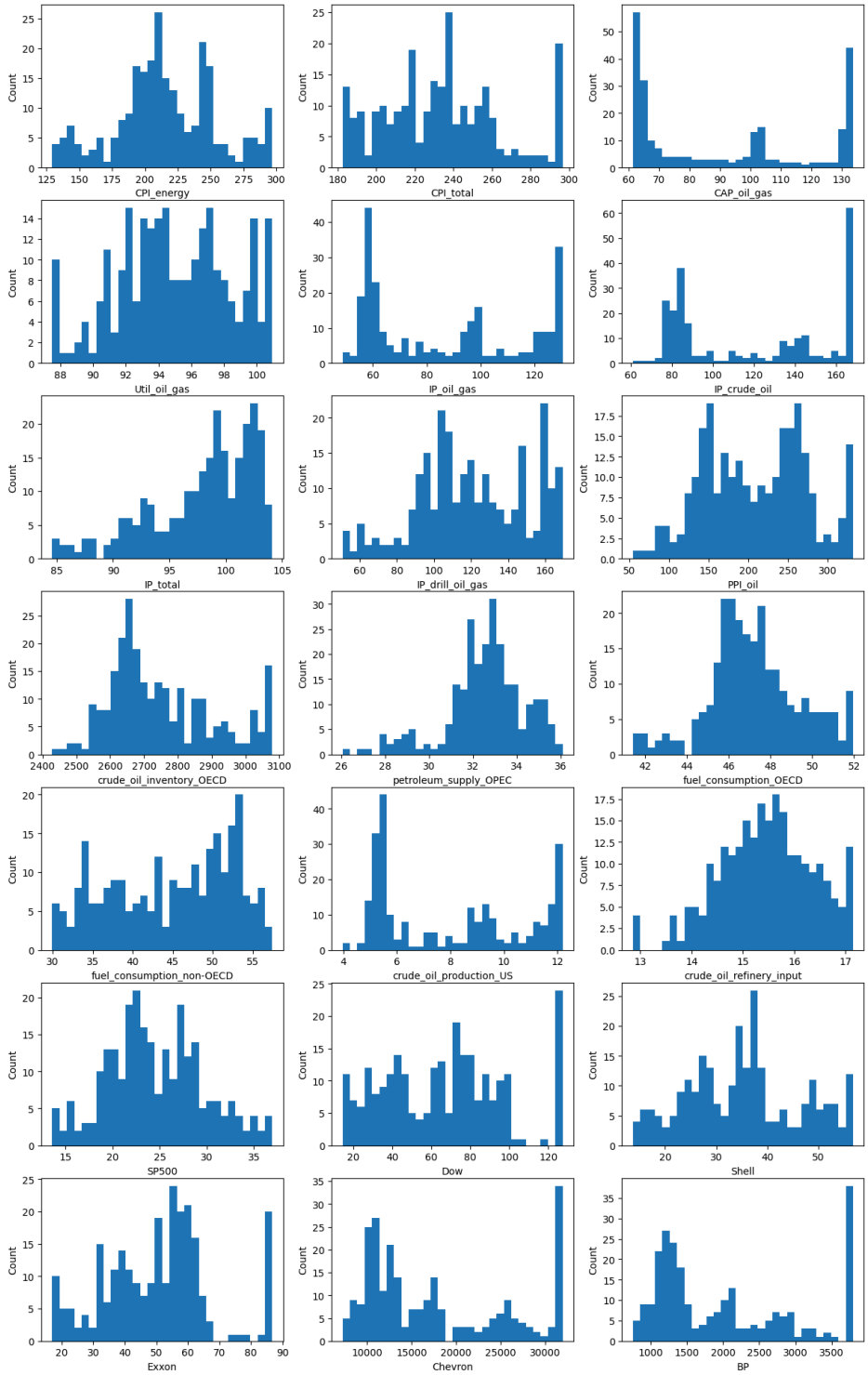
Trends and seasonal patterns[i.e. how variables change over time] was observed via time series plots.

Potential connections between the variables was highlighted by the heatmaps produced.

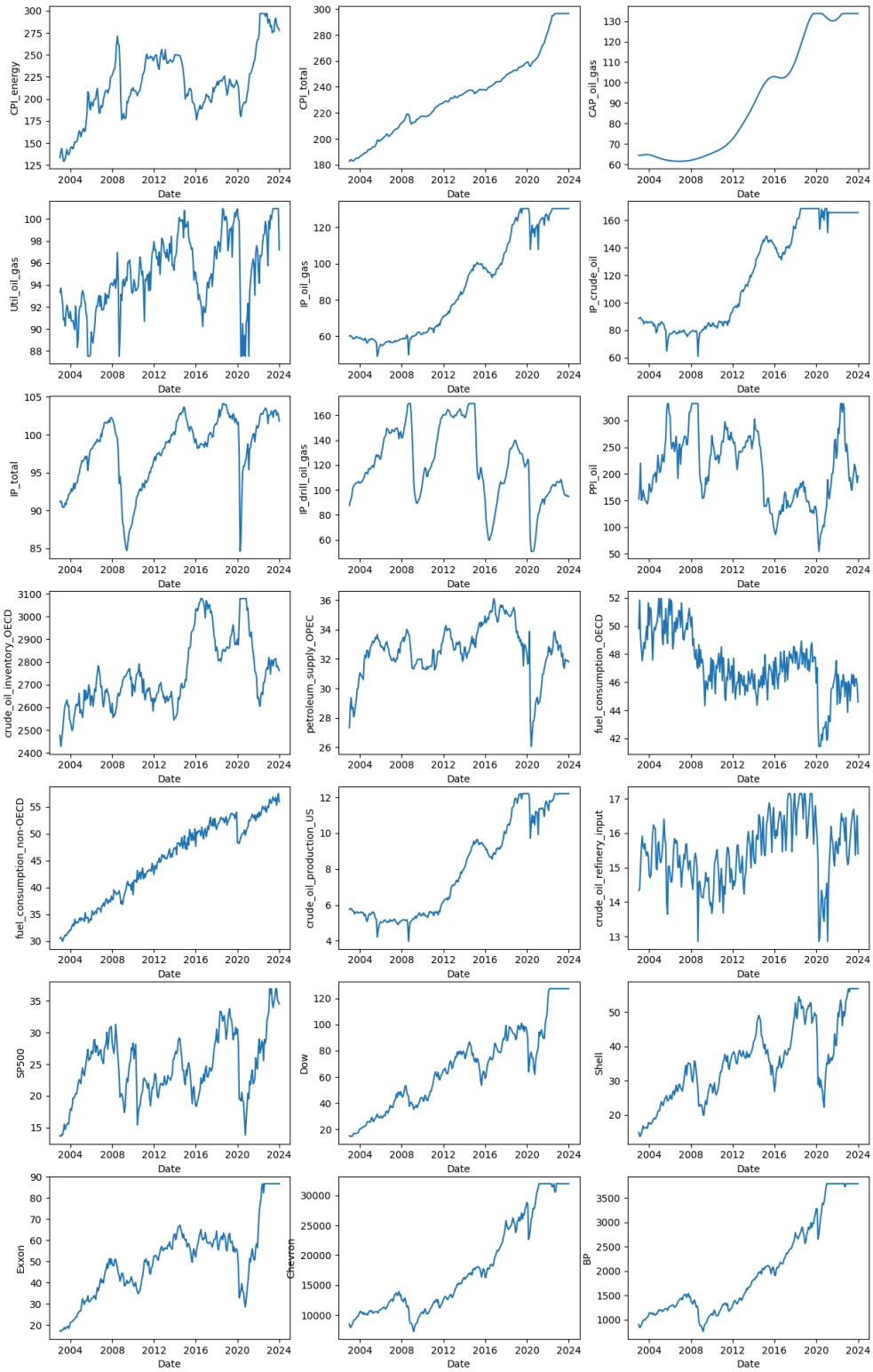
Using PCA, we visualized relationships between the lower-dimensional space which simplified the complex data.

Below are the plots for reference-

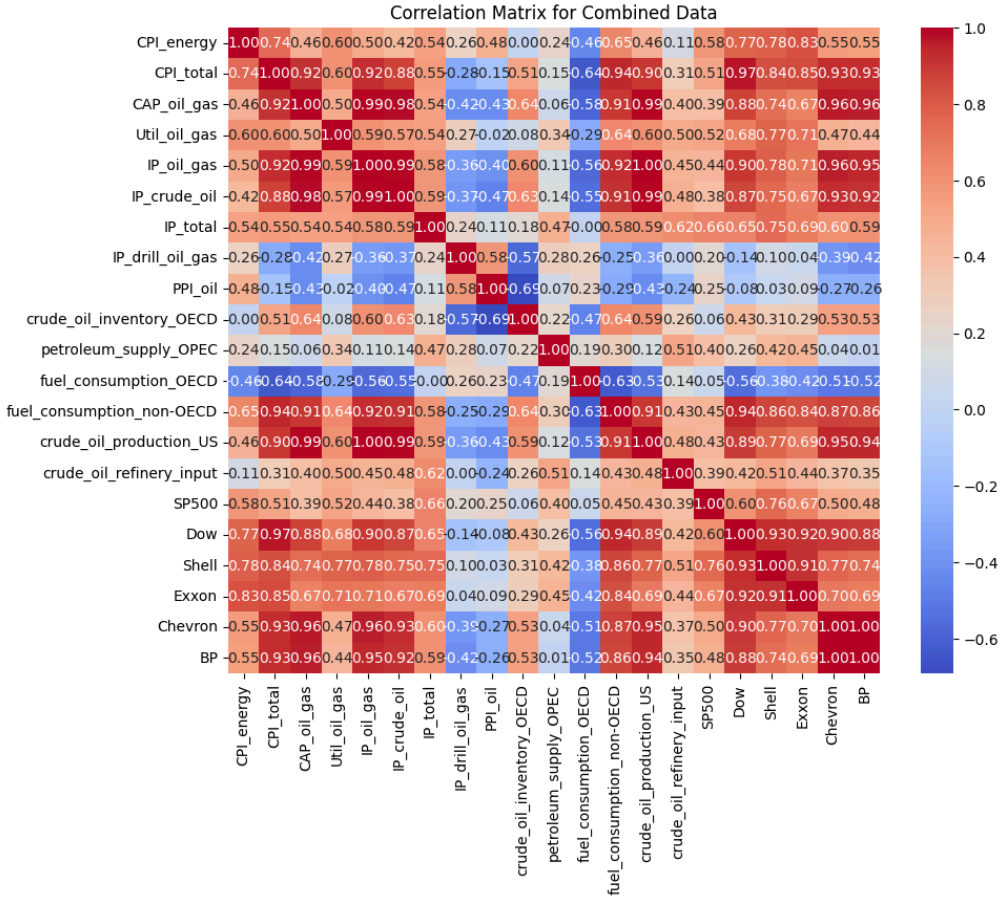
Distribution Plots

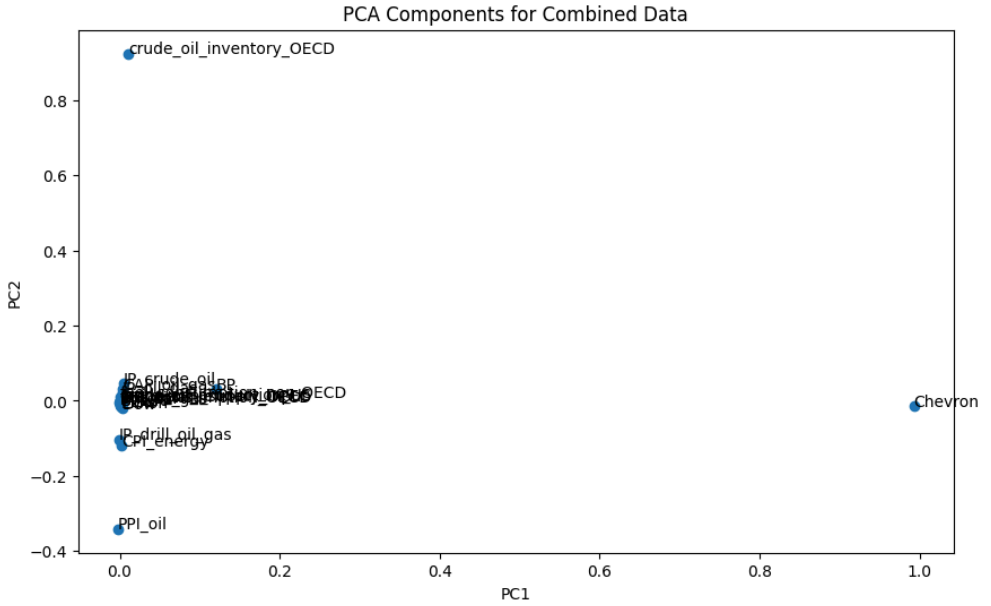


Time Series Plots



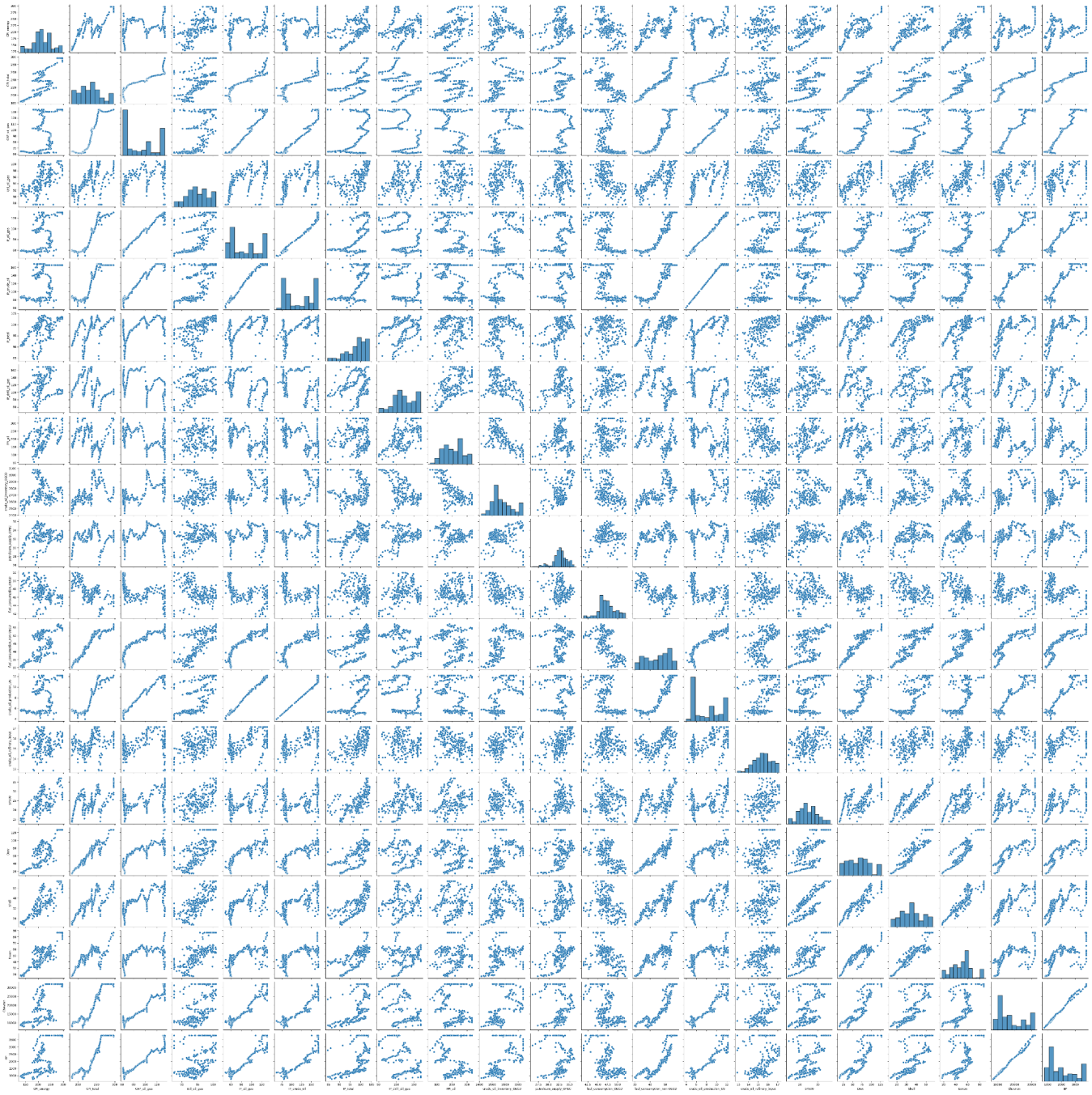
Multivariate Plots





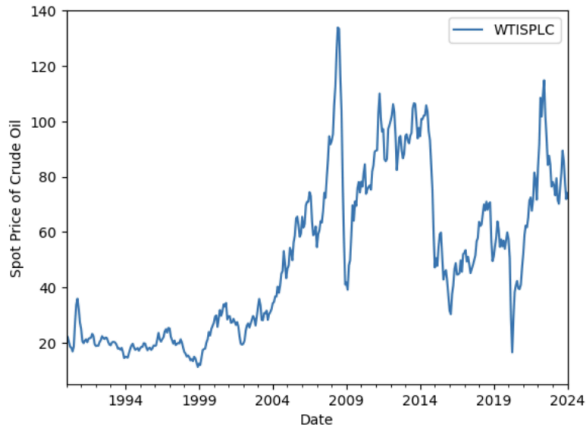
GROUP WORK PROJECT # 1 MScFE 660: RISK MANAGEMENT

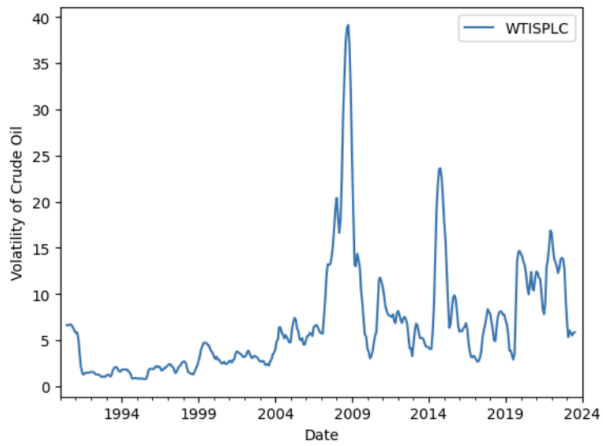
Group Number: 5638



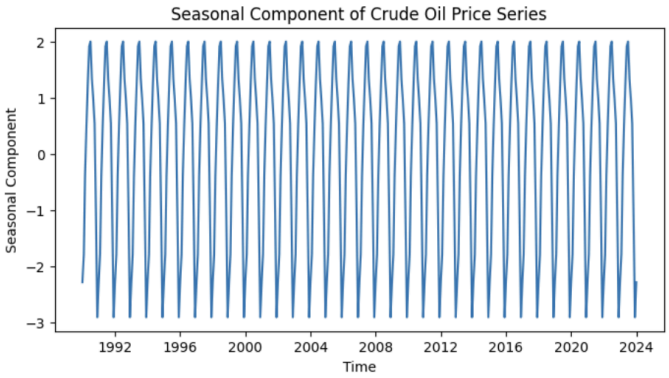
Step 8

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



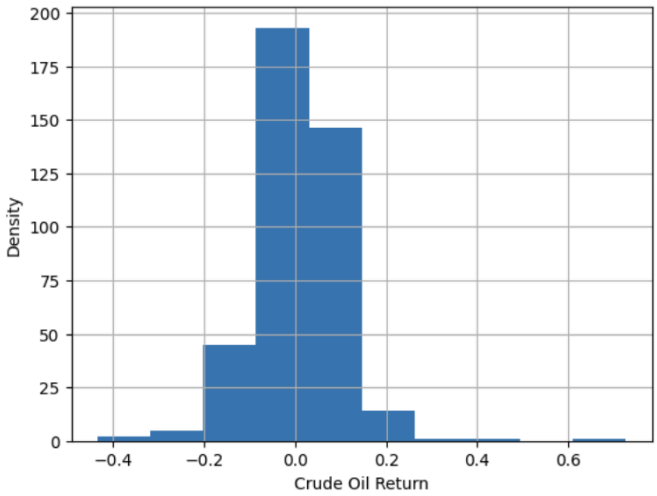


We plot the time series of crude oil spot prices and find that the oil price has greater volatility compared to other assets such as stocks and bonds. It can move sharply during certain periods, such as the years 2008-2010, and it shows sharp spikes during 2008 and 2022. The volatility figure also shows a pattern of volatility clustering.



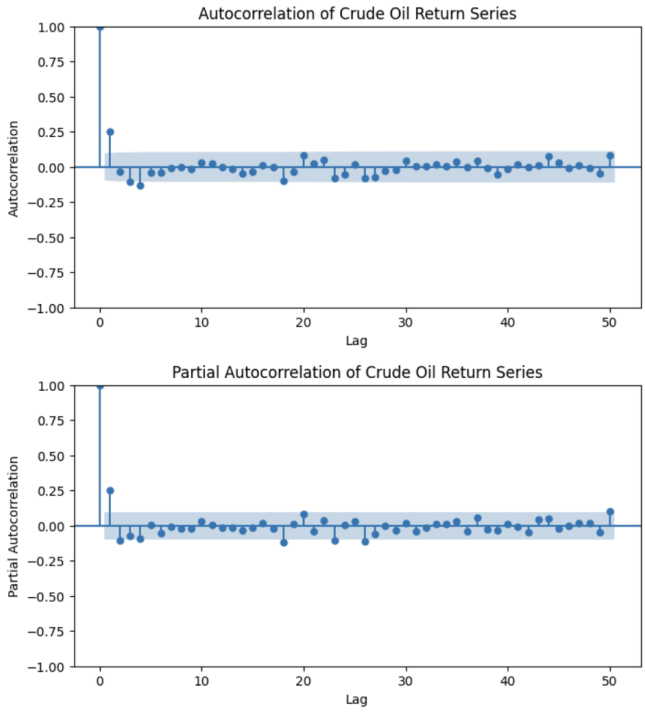
We also identify that there is seasonality within the crude oil price series, possibly due to the weather condition and seasonal demand (e.g. the heating in the winter).

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



We plot the distribution of crude oil return and find that it follows t-distribution.

(c) What types of autocorrelation do the oil return have?



We observe significant autocorrelation at lag 1, 3, and 4. Additionally, the partial autocorrelation analysis reveals significant partial autocorrelation at lag 1, 2, 4, 18, 20, 23, 26, and 50. It also reveals the presence of seasonal effects.

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

In the long run, the price of oil has seen a significant increase, likely due to the rising price levels. Besides, the oil price generally has had higher volatility in recent years, possibly due to the more interconnected markets and more frequent geopolitical events.

Step 9

Probabilistic Graphical Model: Belief networks and Markov Networks

Probabilistic Graphical Models (PGMs) serve as graphical depictions of joint probability distributions, utilizing interdependencies among various random variables. Imagine that PGMs are like maps for showing how different things are related. They help us understand how one thing can affect another (Belief networks) or how one state transfers to other states (Markov chain). The probabilistic graphical models help us make sense of how things work together and make predictions about what might happen next.

A belief network (also known as Bayesian network), is a graphical model that represents probabilistic relationships among variables. It uses nodes to represent variables and directed edges to show the dependencies between variables. Each node in the network represents a random variable, and the edges between nodes indicate probabilistic dependencies or causal relationships. Belief networks are used to model uncertainty and make predictions based on available evidence. They allow us to infer the probability of certain events occurring given the evidence we have.

A Markov chain, on the other hand, represents a stochastic process transitioning between a set of defined states without the causality relationship. It's commonly used to model scenarios where the future state depends solely on the current state and is independent of the sequence of past states, which is known as memorylessness or the Markov property.

In summary, although belief networks and Markov chains are both PGMs, they have three major differences. First of all, belief networks emphasize causal relationships and conditional dependencies between variables, while Markov chains focus on temporal relationships, modeling how a system evolves over time based on the current state. Second, nodes in a belief network represent variables, and directed edges between nodes indicate probabilistic dependencies or causal relationships, whereas the Markov chain is a sequence of states where each state depends only on the preceding state. Third, belief networks can model complex conditional dependencies and systems with complex interconnected variables, while Markov chains simply capture sequential dependencies.

Parameter Learning and Structure Learning

Parameter Learning entails the estimation of conditional probability distributions of the individual variables given that we have a set of samples and a Directed Acyclic Graph (DAG) which depicts the interdependencies of the variables. The use of graphical models causes the reduction in dimension of the estimation problem. Each of the components can be represented by a Conditional Probability Distribution (CPD) and can often be estimated separately.

Parameter Learning is of two main types:

1. Maximum Likelihood Estimation - We use the relative frequencies of the occurrences of the variables to represent the CPD
2. Bayesian Estimation - We use the prior conditional probability that express our beliefs about the variables before their observation

Structure Learning - It entails the estimation of a DAG and determination of arcs of the graph to capture the dependencies of the variables given a set of data samples. Structure Learning is challenging since it is computationally intensive and a brute force algorithm will not work. Following are the two approaches for Structure Learning:

1. Score-based approach - This approach assigns a score to each candidate belief network which quantifies how well a graph G represents the dataset D. The score is nothing but the posterior probability of the graph G given the dataset D and the objective of the approach is the maximization of the score. One such scoring parameter is the Bayesian Information Criterion. Given the exponential number of graphs possible given a set of variables, an exhaustive search would be unfeasible and thus improvisation of the graph search mechanism is inevitable which leads us to one of the approaches called the Hill Climbing Algorithm which changes the solution in an incremental manner by changing one variable at a time and accepting the change in the event of a higher score after beginning the procedure from a non-optimal solution
2. Constraint-based approach - It uses the independence test to determine a set of edge constraints for the graph and then identifies the optimal DAG which satisfies the requirement. The technique works well given prior knowledge of the structure and it requires a substantial amount of data to ensure testing of the power adequately. The solution becomes less dependable when the sample size is reduced.

Markov Chain and Markov Blanket

A Markov Chain is a graphical abstraction of a stochastic process that transitions between a finite number of states. Informally, this may be thought as, “What happens next depends on the state of affairs now”. The property of memorylessness is intrinsic to the Markovian statistic and is used to illustrate a phenomenon where the future is conditional on only the current state and independent of the past states. To elaborate the property a bit further, in the Markov Chain, regardless of how the process attained the current state, the potential future state is fixed and is dependent on the present state. The state space can be considered as all the possible values that the process could attain.

GROUP WORK PROJECT # 1 MScFE 660: RISK MANAGEMENT
Group Number: 5638

Random walks illustrate their utility in the field of mathematics and Finite state machines are used to describe Markov chains. While Markov chains are not restricted to any finite size, yet the majority of the implementations are focused on a finite number of states. Markov Chains find their usage in the field of mathematics, biology, economics, game theory and in statistical/information-theoretic contexts.

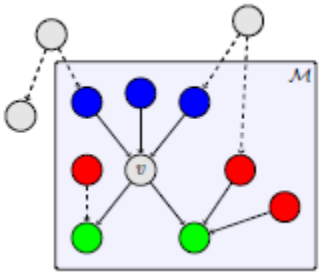
For a first order discrete time stochastic process, the Markov process illustrates the property when the future state in dependent on only the current state and not on any past state, while in a k-order discrete time stochastic process the future state is dependent on the past k states of the process. A discrete time Markov chain is formed by a countably infinite sequence in which the chain shifts state in discrete time steps. A continuous time Markov chain is otherwise a continuous time stochastic process. The Markov chain, as known as the Markov process, was named after Andrew Markov, a Russian mathematician. Markov chains find their application in many statistical models such as studying cruise control systems in motor vehicles, customer queues in airports, foreign exchange rates and animal population dynamics.

The Markov process deals with the states of a process and the change of states is called a transition whereby the probability of changing from one state to the other is called the transition probability. In a finite change of state, the transition probabilities are captured in a transition matrix which describes the states space, the transition and the initial state. By the rule of convention, all possible states and transitions are included in the definition of the process and for each state, there is always a next state and thus, the process never terminates.

The Markov blanket defines the limits of the systems which in other words implies that it is the boundary between the external and internal state. A Markov blanket is said to be minimal when we cannot drop any variable within the blanket without losing any valuable information. Mathematically speaking, the Markov blanket of a node v is the union of the parents of the node v, its children and the parents of its children. Every set of nodes is conditionally independent of v when conditioned on the Markov Blanket of v.

The diagram on the right illustrates a Markov blanket which isolates the inside and outside variables. It should be noted that only the necessary information to predict the behavior of the node and its descendants are present in the Markov blanket. If we compare the Markov blanket with a Markov chain, in the later phenomena, the probability of the next state is dependent on the current state which was established in the preceding event.

GROUP WORK PROJECT # 1 MScFE 660: RISK MANAGEMENT
Group Number: 5638



Step 10

Algorithm of Inferred Causality

Input: A dataset containing Xi variables where i=1,2,....m

Output: A completed, partially directed acyclic graph

1. **Determination of Markov Blankets**
   - 1.1. For each variable X_i, determine the Markov Blanket $B(X_i)$ such that $B(X_i)$ includes the parents of X_i, the children of X_i and the parents of the children of Xi. Capture the relationship of X_i and $B(X_i)$ in a python dictionary
   - 1.2. Check whether the Markov Blankets are symmetric viz. if Xi belongs to $B(X_j)$, then X_j belongs to $B(X_i)$ and we drop the asymmetric blanks as false positive
2. **Learning Neighbors**
   - 2.1. For a pair of variables X_i, X_j, search for a set SX_iX_j belonging to the set of all vertices V such that X_i is conditionally independent of X_j given SX_iX_j and X_i, X_j do not belong to SX_iX_j. If not found, we place an undirected arc between X_i and X_j. If we have the Markov blanket for X_i and X_j, then we search for the smaller of the Markov blanket of Xi except X_j or vice versa.
   - 2.2. Check whether N(x) is symmetric else correct asymmetries in Step 1.2
3. **Learning Arc direction**
   - 3.1. For each pair of non-adjacent variable X_i and X_j, with a common neighbor X_k such that X_k does not belong to SX_iX_j, set the direction of the arc from X_i -> X_k and X_j -> X_k, such that we obtain a v structure where the arc direct towards X_k
   - 3.2. If two variables are not adjacent and X_i-> X_k and X_k – X_j, then we set X_k -> X_j

References

1. “Markov Chain.” Wikipedia, Wikimedia Foundation, 8 Apr. 2024, en.wikipedia.org/wiki/Markov_chain#:~:text=A%20countably%20infinite%20sequence%2C%20in,a%20two%2Dstate%20Markov%20process.

2. Băncioiu, Camil. “Accelerating Causal Inference and Feature Selection Methods through G-Test Computation Reuse.'' 26 Sep. 2021.

3. Alvi, Danish A. Application of Probabilistic Graphical Models in Forecasting Crude Oil Price. 2018. University College London, Dissertation.
https://arxiv.org/abs/1804.10869