**GROUP WORK PROJECT #** 1
**GROUP NUMBER:** 6442

MScFE 660: RISK MANAGEMENT

| FULL LEGAL NAME | LOCATION (COUNTRY) | EMAIL ADDRESS | MARK X FOR ANY NON-CONTRIBUTING MEMBER |
|---|---|---|---|
| Marin Stoyanov | Bulgaria | azonealerts@gmx.com | |
| Prabhdeep Kaur | India | prabhdeep089kaur@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** | Marin Stoyanov |
| **Team member 2** | Prabhdeep Kaur |
| **Team member 3** | |

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

# Application of Bayesian Networks in understanding the problem of underlying factors driving crude oil prices

Mini-Capstone Project

By
Marin Stoyanov
Prabhdeep Kaur

## Introduction

The thesis investigates the combination of Bayesian networks and risk management to solve the problem of underlying factors driving crude oil prices, that is, to understand all the factors that may be affecting the prices of crude oil. Bayesian Networks, a type of Probabilistic Graphical Model, have the ability to model probabilistic relationships among variables and help in analyzing the factors influencing the prices of oil.

## Problem Formulation

- The problem that the thesis attempts to solve:

The problem that the thesis attempts to solve is the problem of underlying factors driving crude oil prices. Apart from predicting prices, it is also necessary to understand how crude oil prices react to hypothetical scenarios like geopolitical tensions or sudden changes in supply. Therefore, the thesis attempts to solve this problem by performing stress tests and evaluating the robustness of the model under extreme conditions.

- Why Bayesian Network is well suited:

Bayesian networks are suitable for solving the problem of underlying factors driving crude oil prices because it can combine a lot of variables to find the dependencies between them and successfully extrapolate the data thanks to the learned knowledge. This method applies probabilistic reasoning and can be dynamically adapted towards an ever changing world. Bayesian networks models casual relationships between variables which allows for effective stress testing by simulating the impact of different hypothetical situations on crude oil prices.

- Advantages of using Bayesian Network;

The main advantage is that it lets us introduce prior knowledge that could be either expert-based or extracted from other models and put it into the decision process. As compared to traditional time-series models that might overlook the complex interactions like the underlying factors driving crude oil prices, bayesian networks provide more accurate and reliable forecasts.

## Data Collection

The thesis employs the data on macroeconomic and financial indicators to construct Bayesian Network under the assumption that there's no microeconomic data.

The Macroeconomic data includes the variables:

| | Ticker | Description | Frequency | Source | Start Date | End Date |
|---|---|---|---|---|---|---|
| 0 | WTISPLC | Canadian dollar to US dollar exchange rate | Monthly | FRED | 1919-01-01 | 2024-06-01 |
| 1 | CPIENGSL | Consumer Price Index for All Urban Consumers: ... | Monthly | FRED | 1919-01-01 | 2024-06-01 |
| 2 | CAPG211S | Industrial Capacity: Mining: Oil and Gas Extra... | Monthly | FRED | 1919-01-01 | 2024-06-01 |
| 3 | CAPUTLG211S | Capacity Utilization: Mining: Oil and Gas Extr... | Monthly | FRED | 1919-01-01 | 2024-06-01 |
| 4 | IPG211S | Industrial Production Index: Mining: Oil and G... | Monthly | FRED | 1919-01-01 | 2024-06-01 |
| 5 | INDPRO | Industrial Production: Total Index | Monthly | FRED | 1919-01-01 | 2024-06-01 |
| 6 | IPN213111N | Industrial Production: Total Index | Monthly | FRED | 1919-01-01 | 2024-06-01 |
| 7 | PCU211211 | Producer Price Index: Mining: Oil and Gas Extr... | Monthly | FRED | 1919-01-01 | 2024-06-01 |

Table 1. Macroeconomic data frequencies and source

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| **WTISPLC** | 942.0 | 27.816605 | 29.284847 | 1.1700 | 3.0000 | 18.3125 | 39.34000 | 133.9300 |
| **CPIENGSL** | 810.0 | 114.829889 | 80.335992 | 21.3000 | 31.7500 | 101.5000 | 192.18225 | 331.7380 |
| **CAPG211S** | 629.0 | 81.421457 | 21.937600 | 61.4822 | 66.6233 | 76.3514 | 81.75120 | 149.2870 |
| **CAPUTLG211S** | 629.0 | 93.202929 | 3.239208 | 78.6360 | 91.5201 | 93.1402 | 94.95650 | 101.5441 |
| **IPG211S** | 629.0 | 76.173594 | 22.164265 | 48.8141 | 62.4444 | 68.5466 | 78.22200 | 144.4211 |
| **INDPRO** | 1265.0 | 45.891390 | 34.813005 | 3.6827 | 13.7629 | 39.1057 | 84.15640 | 104.1038 |
| **IPN213111N** | 629.0 | 130.404341 | 53.132521 | 47.9947 | 94.0882 | 113.9683 | 156.98800 | 334.6264 |
| **PCU211211** | 462.0 | 154.771381 | 83.664228 | 54.6000 | 77.3750 | 138.1500 | 221.97500 | 490.4000 |

Table 2. Macroeconomic data description

The Financial data includes the variables:

| | Ticker | Description | Frequency | Source | Start Date | End Date |
|---|---|---|---|---|---|---|
| 0 | DEXCAUS | Canadian dollar to US dollar exchange rate | Daily | FRED | 1971-01-04 | 2024-07-11 |
| 1 | VIXCLS | CBOE Volatility Index | Daily | FRED | 1971-01-04 | 2024-07-11 |
| 2 | DCOILWTICO | WTI Crude oil futures | Daily | FRED | 1971-01-04 | 2024-07-11 |
| 3 | DCOILBRENTEU | Brent crude oil futures | Daily | FRED | 1971-01-04 | 2024-07-11 |
| 4 | SP500-45 | S&P 500 Energy sector index | Daily | FRED | 1971-01-04 | 2024-07-11 |
| 5 | SP500 | S&P500 Index | Daily | FRED | 1971-01-04 | 2024-07-11 |

Table 3. Financial Data frequencies and source

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| **DEXCAUS** | 13425.0 | 1.229735 | 0.160415 | 0.9168 | 1.1035 | 1.2353 | 1.3474 | 1.6128 |
| **VIXCLS** | 8711.0 | 19.491915 | 7.879984 | 9.1400 | 13.7950 | 17.6300 | 22.8700 | 82.6900 |
| **DCOILWTICO** | 9700.0 | 47.291728 | 29.700010 | -36.9800 | 20.1500 | 39.2800 | 70.9425 | 145.3100 |
| **DCOILBRENTEU** | 9423.0 | 49.936190 | 32.927056 | 9.1000 | 19.2800 | 42.7200 | 74.2300 | 143.9500 |
| **SP500** | 2516.0 | 3177.629996 | 979.254402 | 1829.0800 | 2268.9750 | 2890.6550 | 4090.4225 | 5633.9100 |

Table 4. Financial data description

## Step - 5 & Step - 6

To get a more robust dataset for further analysis, both the Macroeconomic and Financial data are cleaned of extreme outliers and missing values.

We have filtered all the data and in order to get rid of the outliers we applied the 3 standard deviation outlier detection method and after this we dropped the data points that are outside of the 3 standard deviation metric. After this we checked for null data and we gladly found out that there is none so no further imputation is needed.

Plots of outliers can be observed below:

For Macroeconomic data,

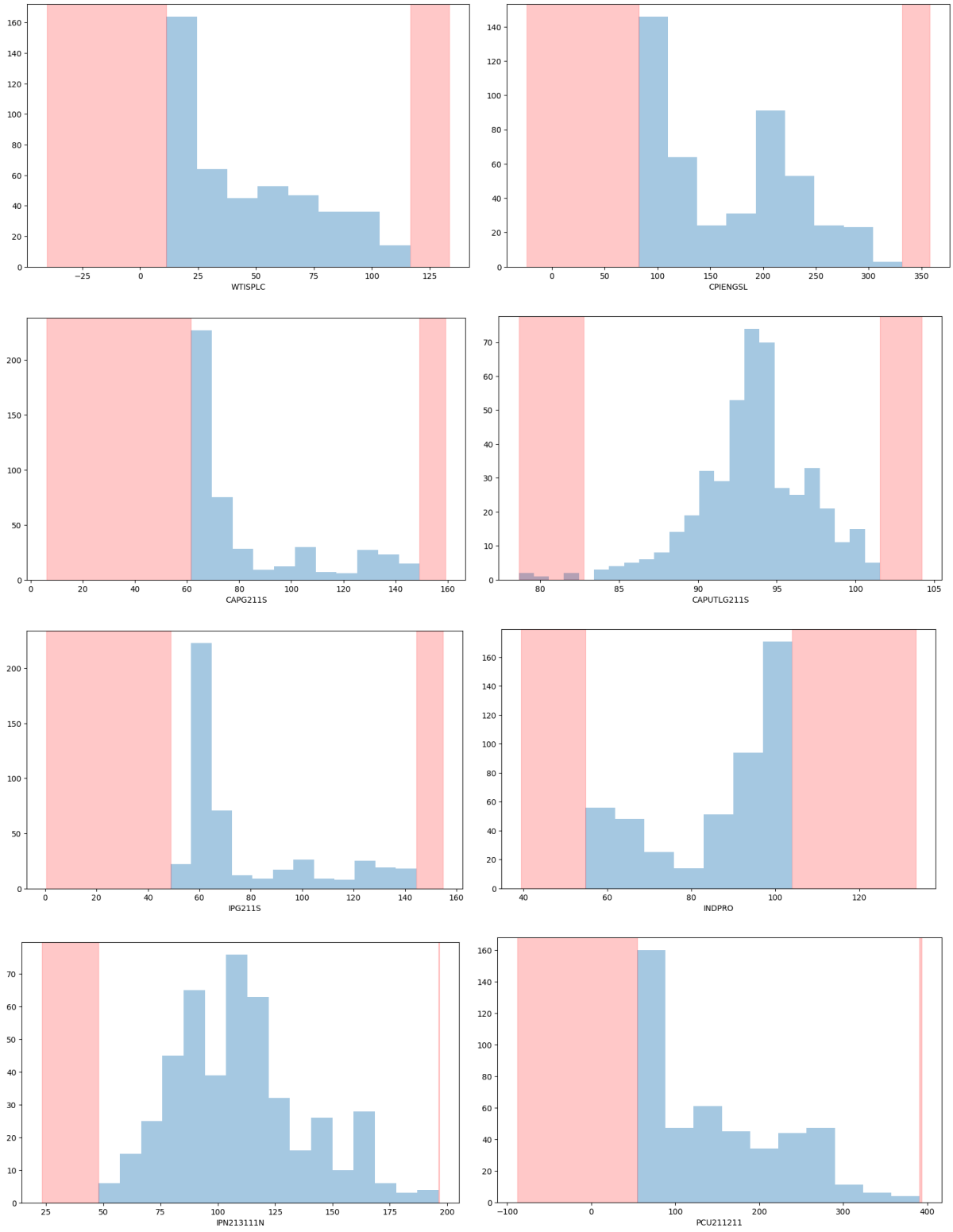



Fig.1. Outlier detection in Macroeconomic data

For Financial data,

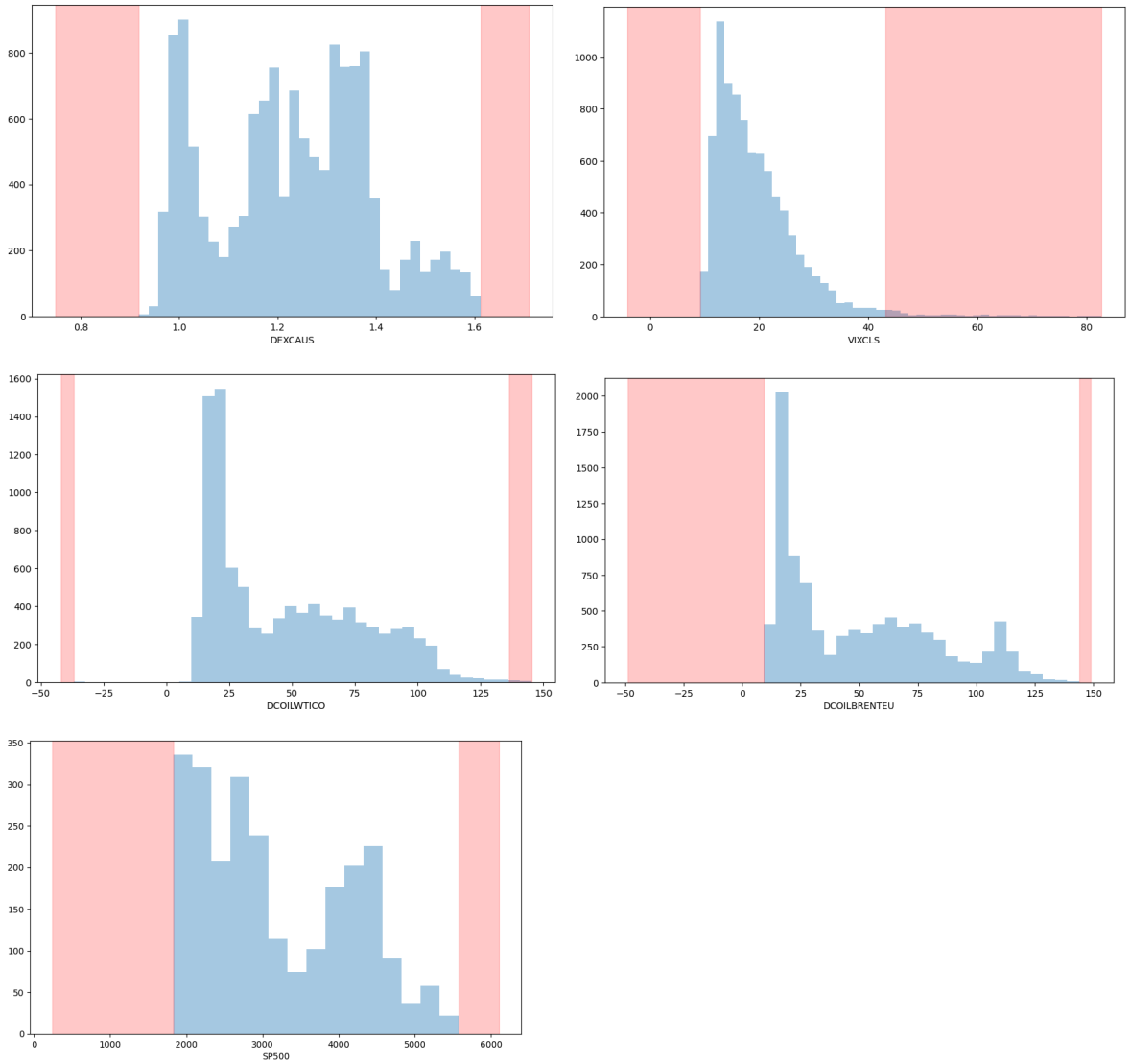



Fig.2. Outlier detection in Financial data

After removing outliers from the Financial data, no missing values were observed in the financial data and for macroeconomic data, the missing values were dropped.

Exploratory data analysis is performed on the data and plots are observed.

The Distributional plots for Macroeconomic data is as follows-

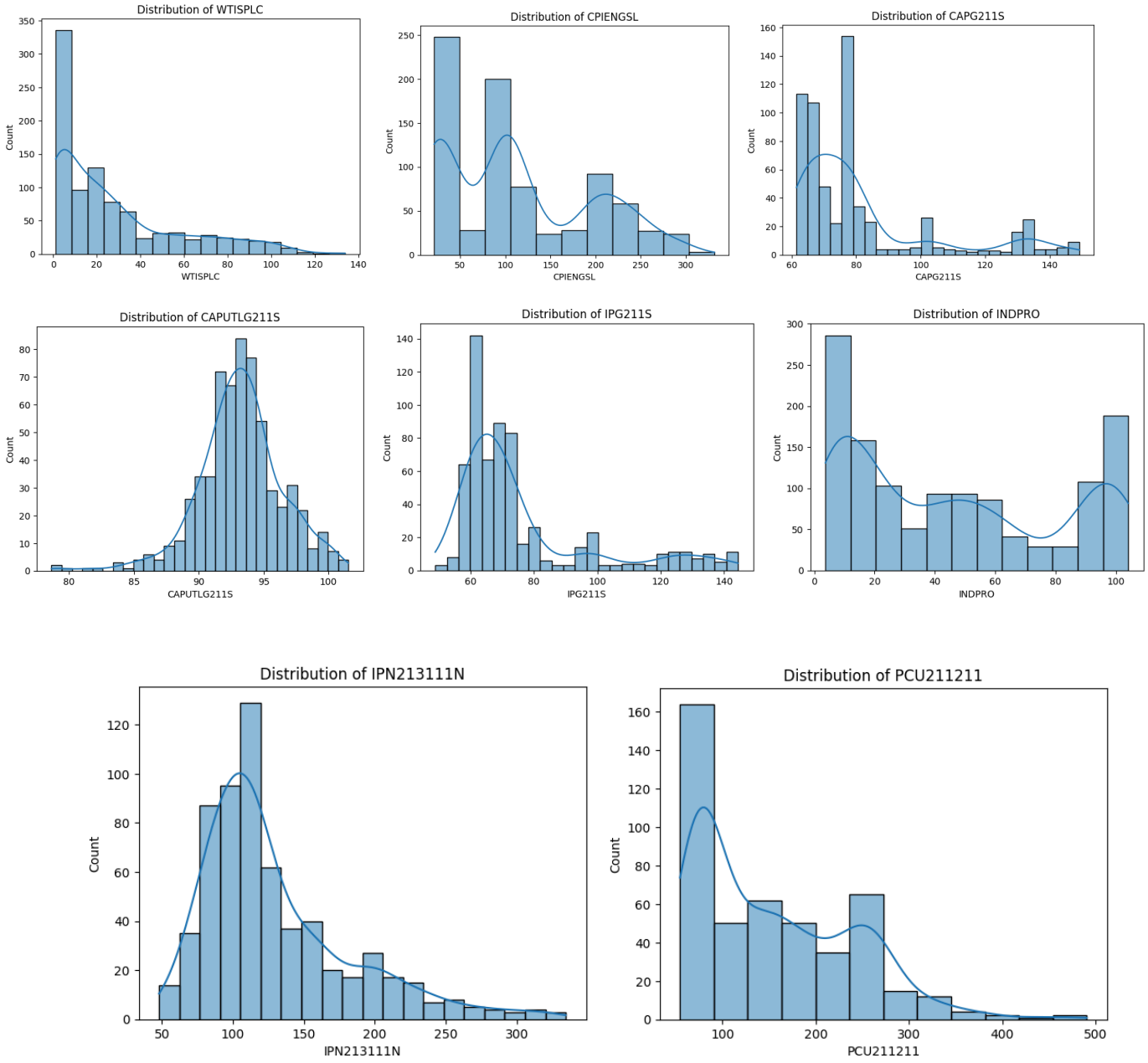



Fig.3. Distributional plots for Macroeconomic data

From all the plots above, we observe that the distribution of WTISPLC: Spot Crude Oil Price: West Texas Intermediate is positively skewed on the right side. The distribution of CPIENGSL, CAPG211S, IPG211S, INDPRO and PCU211211 is multimodal, indicating the presence of multiple subgroups within the data. CAPUTLG211S follows the normal distribution and IPN213111N follows the right skewed distribution.

The scatter plot matrix for the Macroeconomic data is-

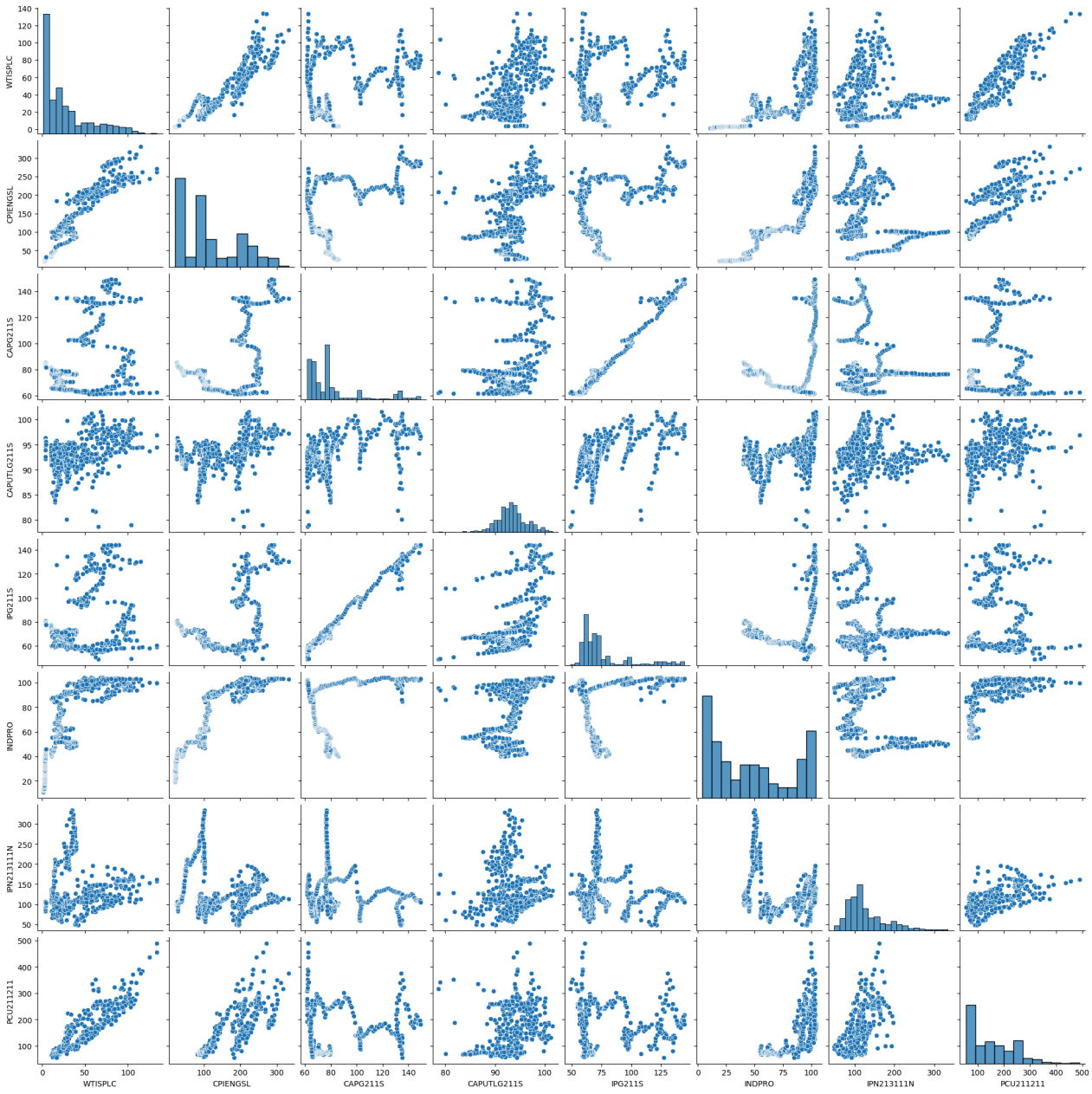



Fig.4. Scatter plot matrix for the Macroeconomic data

The correlation matrix for the Macroeconomic data is-

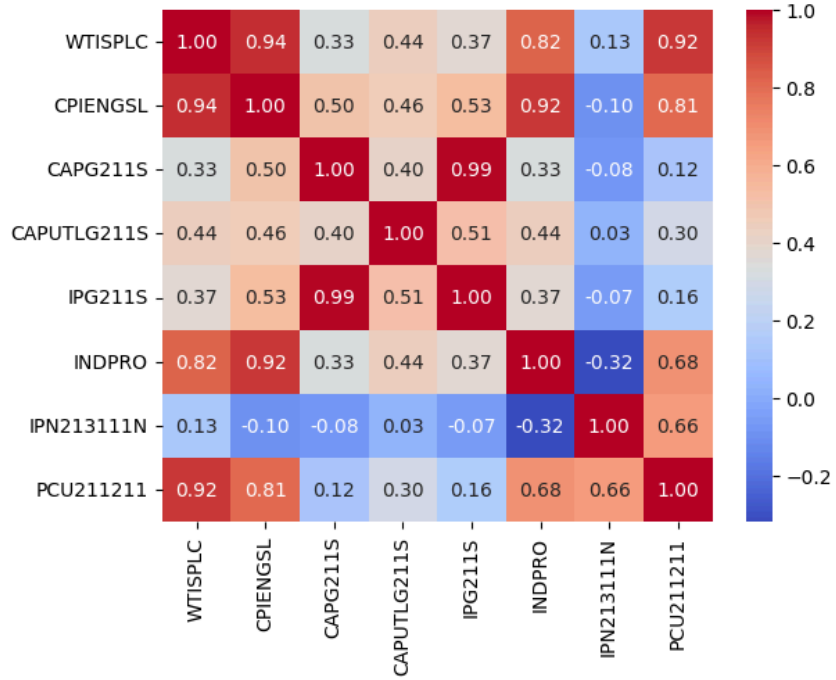



Fig.5. correlation matrix for the Macroeconomic data

Above plot shows that the correlation between CAPG211S and IPG211S is highest with 0.99 indicating a strong positive relationship while the correlation between INDPRO and IPN213111N is lowest with -0.32 which suggests a strong negative relation among them.

Time series plots of Macroeconomic data is as follows-

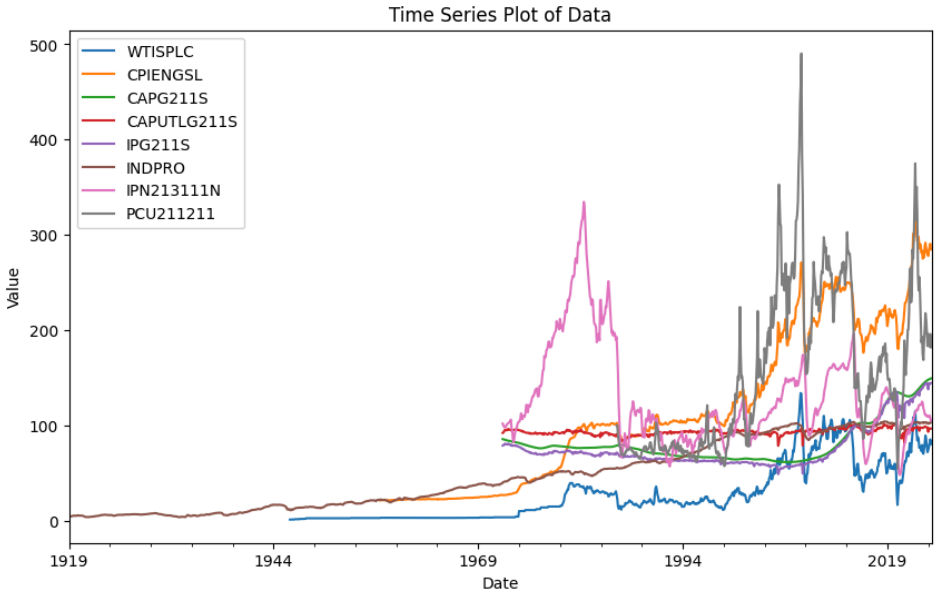



Fig. 6. Time series plots of all the Macroeconomic data

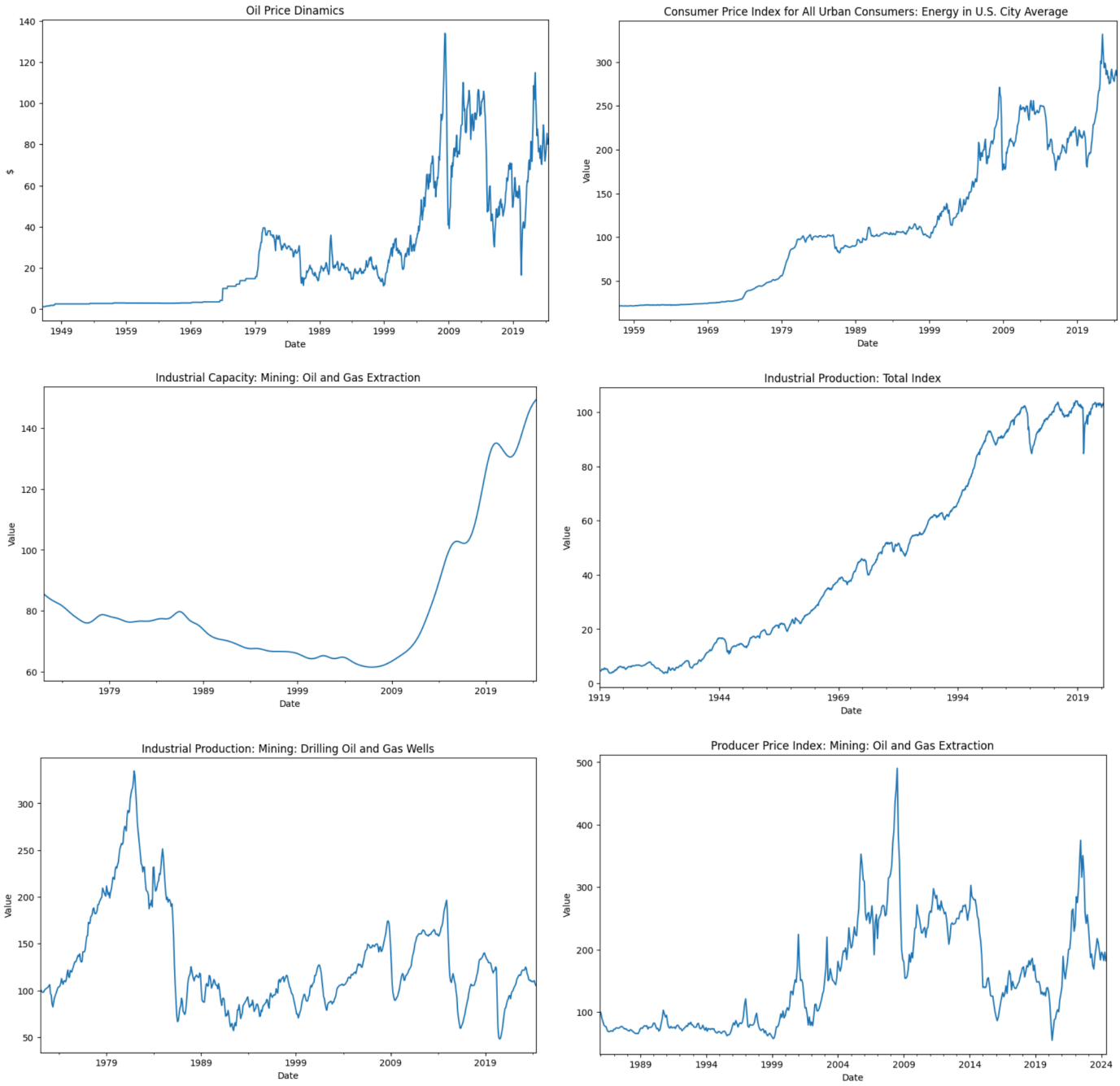



Fig. 6. Individual time series plots of the Macroeconomic data

Here we can see that the consumer City Average prices are in an up trend with a ranging period from 2008 till 2020 and after that steep climb for 2022 when the war between Russia and Ukraine began.

Obviously the Industrial Capacity has risen a lot which is quite normal because of the USA and world economy book and the need of energy for transportation, production and other industries.

As we can see the industrial production has been halted for the major crisis period of 2008 (the World Financial Crisis) and in 2020 (the Covid19 Crisis)

At first sight it looks like the mining and drilling for new wells in a down trend and this may be because of the renewable energy boom that is happening right now.

The distributional plots for Financial data is as follows-

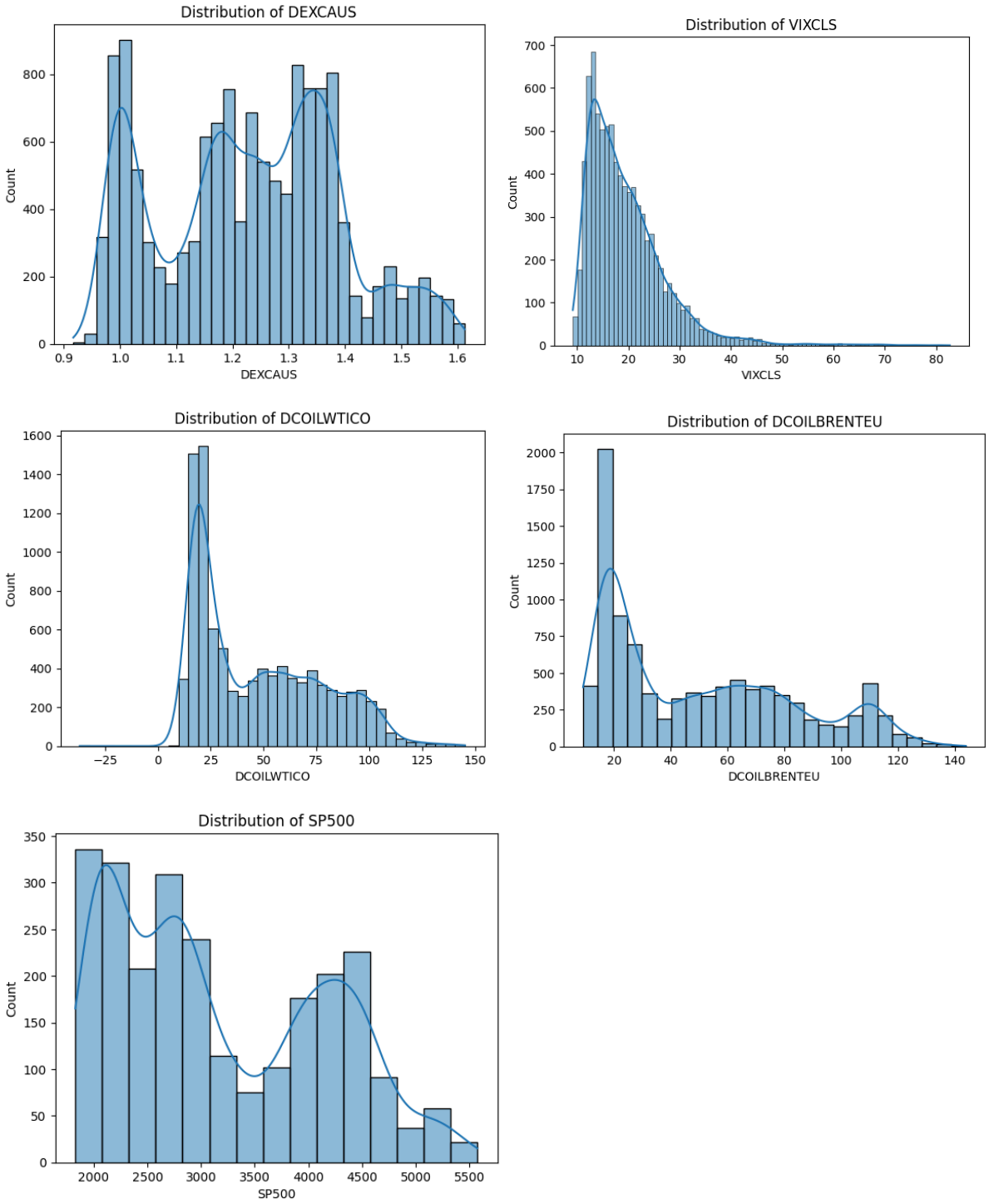



Fig. 7. Individual distribution plots of the Financial data

The distribution of VIXCLS is positively skewed on the right side while the distribution of DEXCAUS, DCOILWTICO, DCOILBRENTEU and SP500 is multimodal indicating the presence of multiple subgroups within the data.

The scatter plot for the Financial data is-

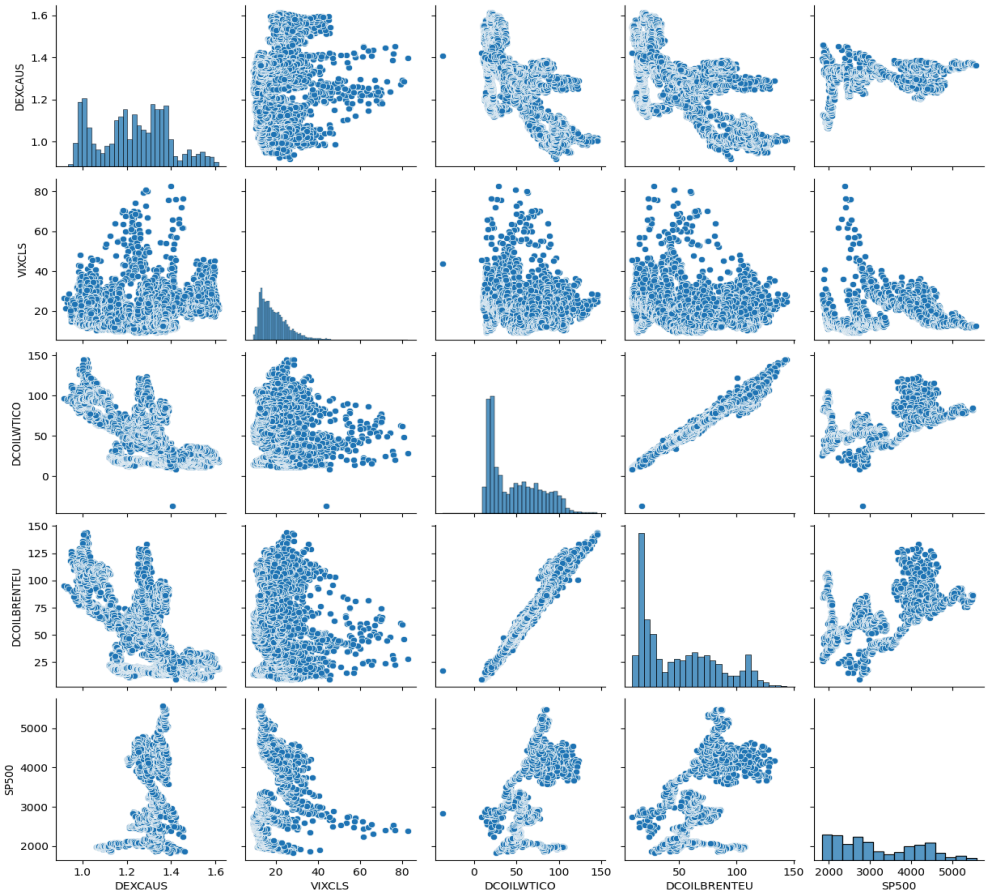



Fig. 8. Scatterplot matrix of the Financial data

The correlation matrix for the Financial data is-

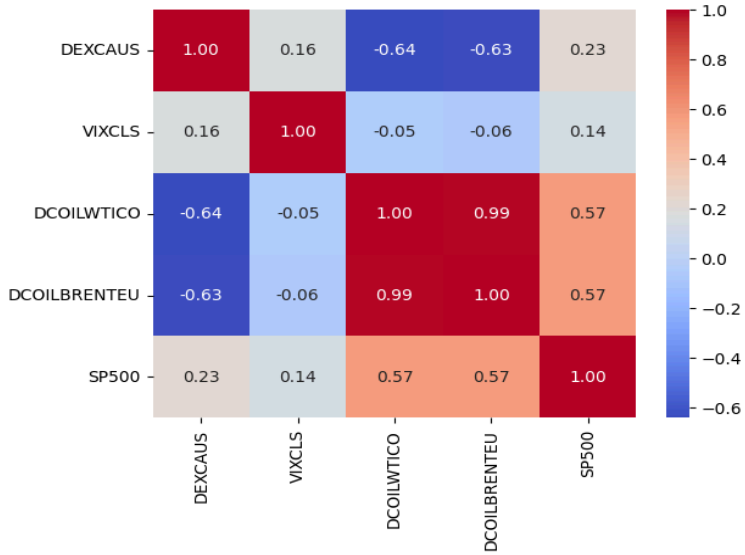



Fig. 9. Correlation matrix of the financial data

Above plot shows that the correlation between DCOILBRENTEU and DCOILWTICO is highest with 0.99 indicating a strong positive relationship while the correlation between DCOILWTICO and DEXCAUS is lowest with -0.64 which suggests a strong negative relation among them.

Time series plots of Financial data is as follows-

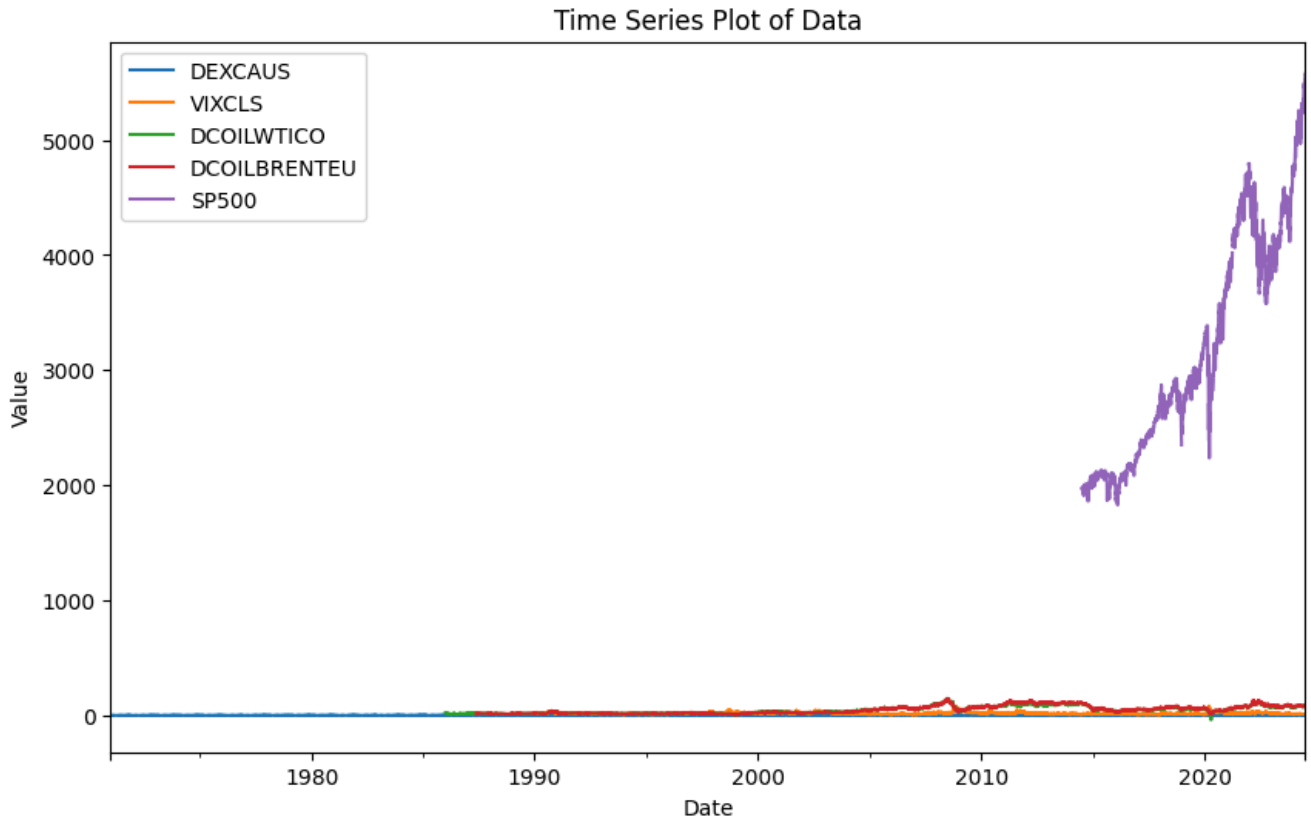



Fig. 10. Time series plot of the financial data

Since the chart of all the financial data in one plot is not very good visible we are plotting all the financial data time series individually so that we could get some better representation and be able to see some better pictures of the dynamics to every single variable

Below plot shows the canadian dollar to US dollar exchange rate over years.

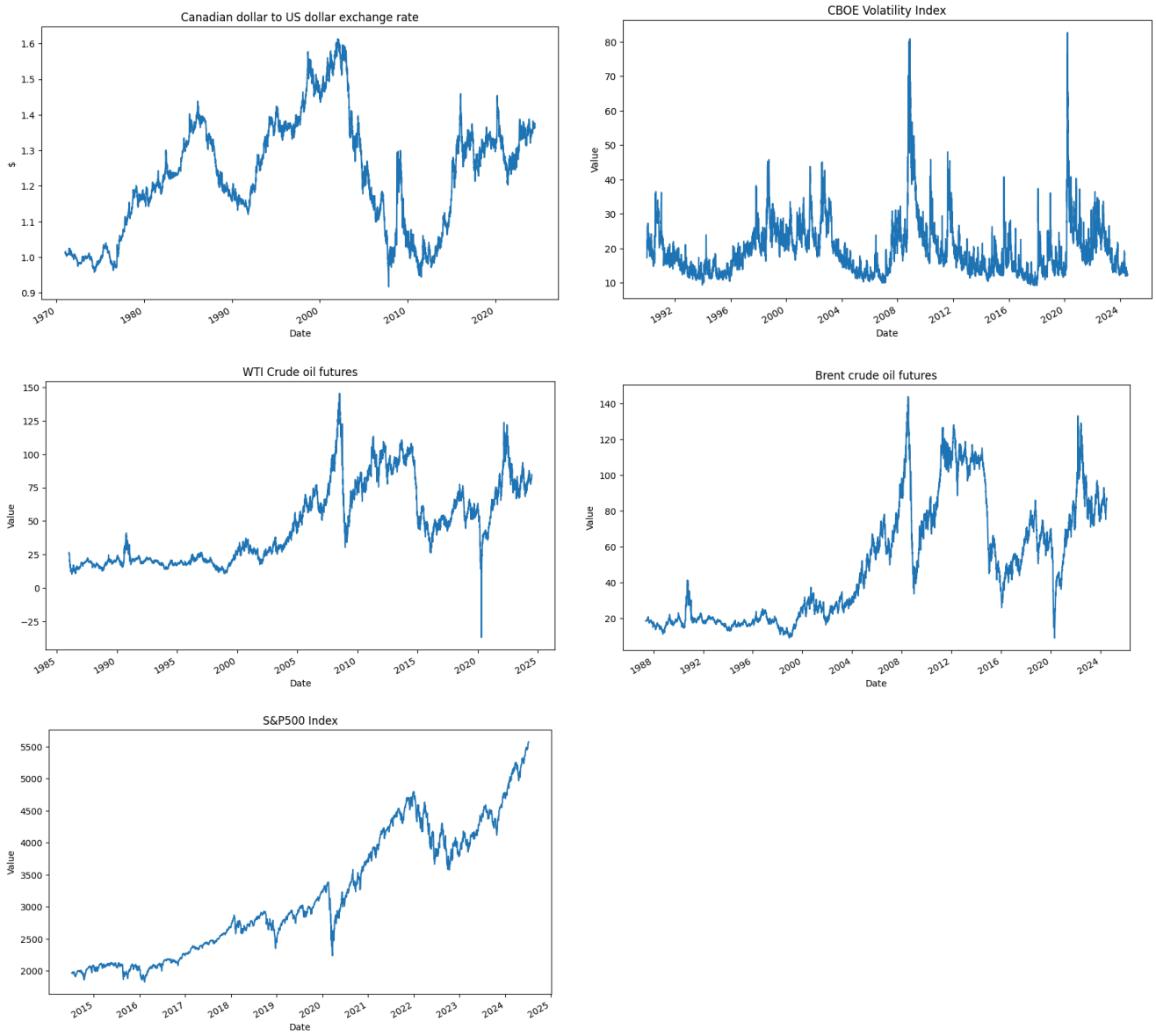



Fig. 11. Individual Time series plot of the financial data

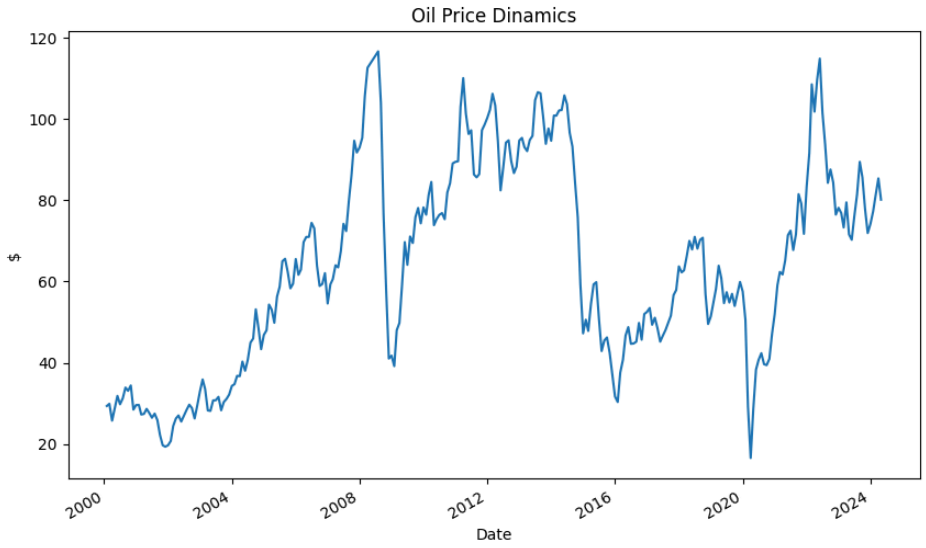



Fig. 12. Oil price dynamics plot for the period 2000 - 2024

We see that for the period 2000 - 2024 Oil price is locked in a very big range from $20 till $140 with big spikes and steep declines as well. Usually the big spikes are at periods of booming economy and afterwards followed by crises that are known with low consumption of oil so it is quite normal for prices to normalize and fall down. Such booming economic periods are 2002 - 2008, 2010 - 2012, 2020 - 2023 and they are followed by crisis (2008 - 2009 The Great Financial Crisis, 2015-2016, 2019-2020 Covid19) and step decline in demand for oil.

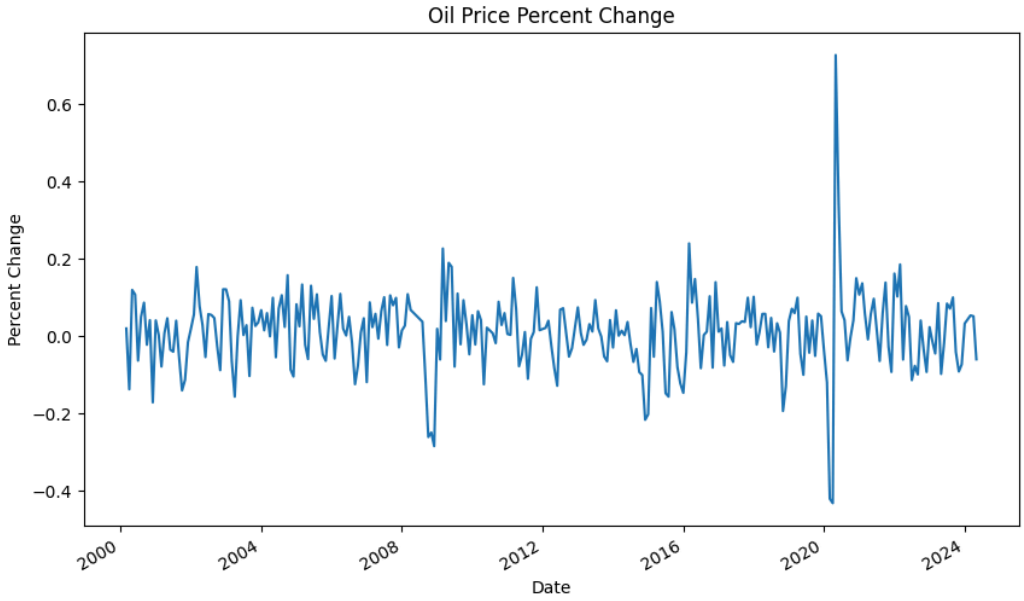
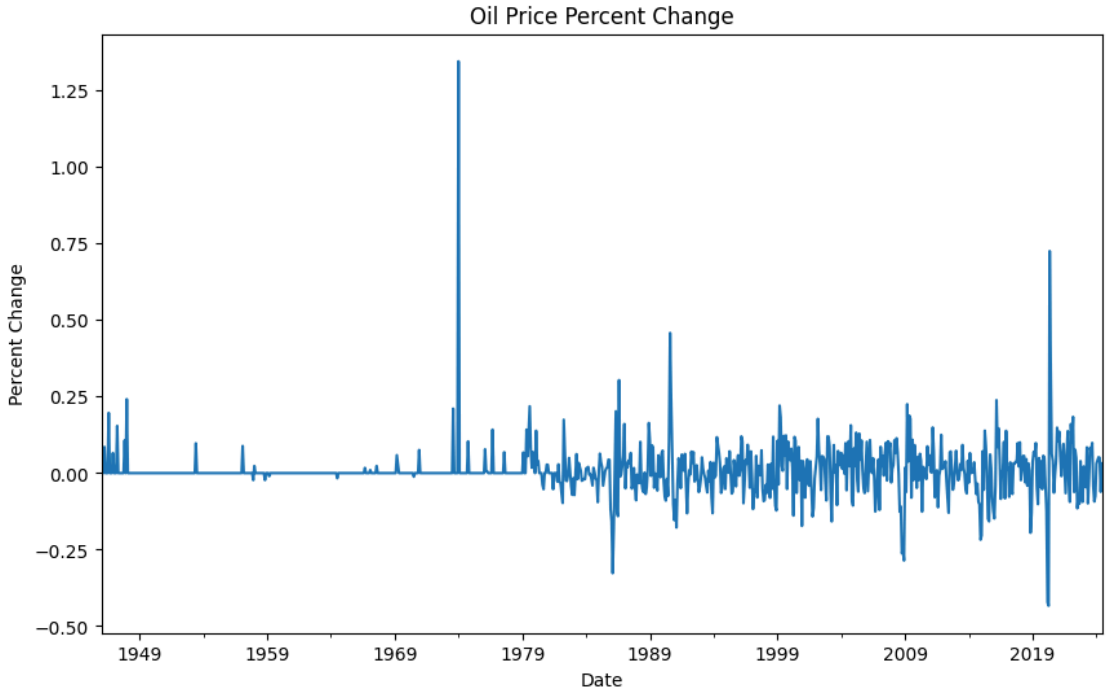



Fig. 13. Oil price returns chart

We can say that the plot of Oil prices returns for the period 2000 - 2024 look quite normal (but we have to check it with some statistical methods) with the biggest spike around Covid19. Let's look at the bigger picture: all the data




The entire Oil returns data show a more ugly picture with a lot more spikes. Lets see the ACF and Pacf plots for more insights-

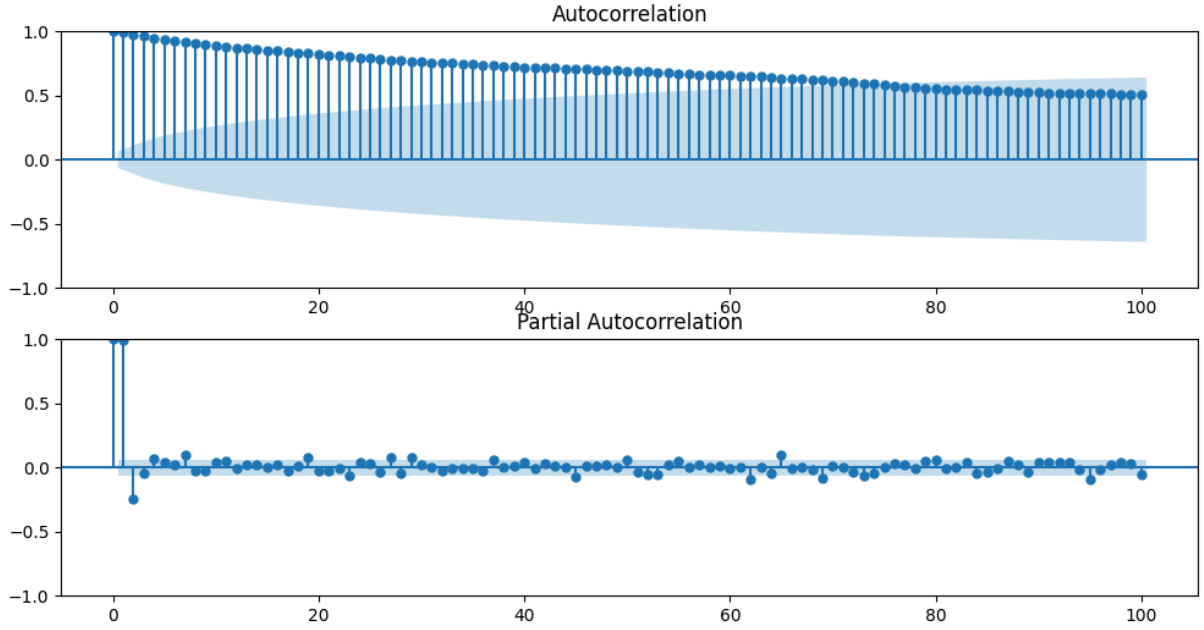



Fig. 14. ACF and PACF plots

From the ACF plot we can see that there is a trend in the Oil price, which is quite normal as everything with time goes up thanks to inflation (a totally normal process). The interesting thing is the PACF where we can see that there is a seasonality pattern where we can see that at the beginning of the year there is a downward spike (during the northern globe is winter and since there is most of the population located and the fact that it is winter suggests that there is also a lower demand for Oil) and also during the summer of the year (I am talking about the northern globe seasons ) then an upward spike in the Oil consumption because it is summer time and a lot of traveling is happening at this time of the year.

| | |
|---|---|
| Results of ADF Test: | |
| Test Statistic | -1.633727e+01 |
| P-value | 3.020024e-29 |
| #Lags Used | 3.000000e+00 |
| Number of Observations Used | 9.360000e+02 |
| Critical Value (1%) | -3.437356e+00 |
| Critical Value (5%) | -2.864633e+00 |
| Critical Value (10%) | -2.568417e+00 |
| dtype: float64 | |

Table 5. ADF test summary

Here the ADF test's p-value of 3.020962e-29 proves that it is very unlikely that the data has a unit root so the data is stationary.

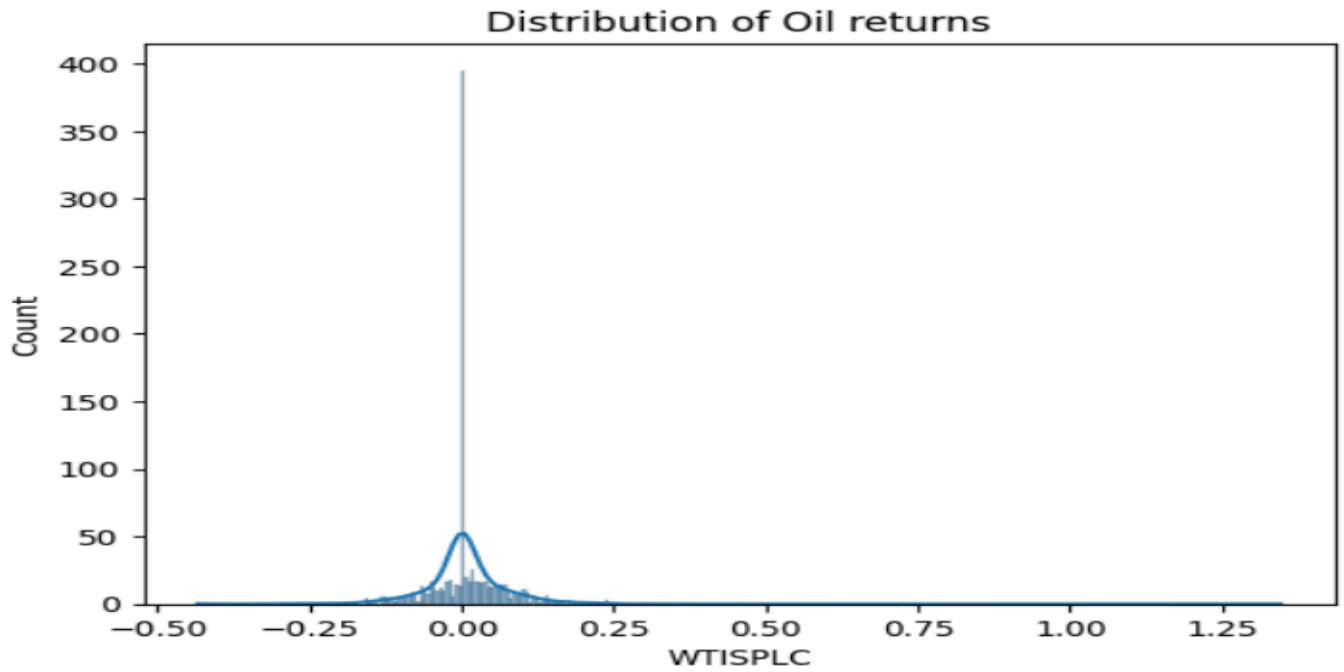



Fig. 15. SP500 distribution plot

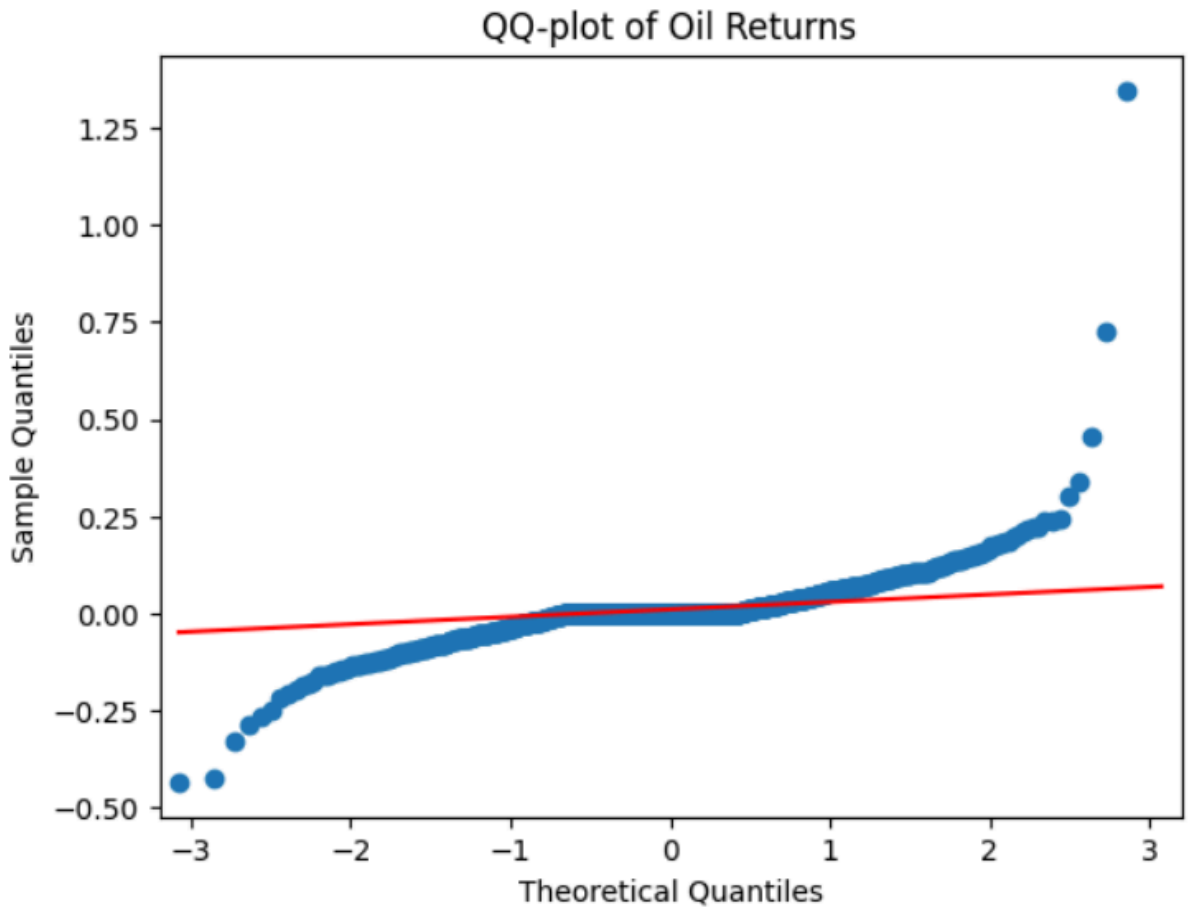



Fig. 16. QQ-plot of Oil Returns

Let's do some exploration about whether the Oil returns data is normally distributed and we are doing this by a QQ-plot that shows that the data has big tails and this suggests that overall the data may look like it is normally distributed by the bell shaped histogram plot. Actually it is not because of the fat tails. For better proof of this hypothesis we will do the statistical test for normality check called Shapiro-Wilk test for normality. This will undeniably prove or deny the hypothesis mentioned above.

```
Shapiro-Wilk Test:
Statistic: 0.8222251534461975
P-value: 4.956003613292808e-31
```

Table 6. Shapiro Wilk test results

Yet again a p-value of 4.912296382340533e-31 for the Shapiro-Wilk normality test proves that the data is not normally distributed. With such a small p-value we cannot accept the null hypothesis (which suggests that the data is normally distributed) so in this case we fail to accept it and we can say that we have

statistically proved that the data is NOT normally distributed. Something else is very interesting and it is the statistic= 0.822 for similarity to a normal distribution. This suggests that the data is close to a normal distribution (the bell shape) but the fat tails fails this thesis.

| | lb_stat | lb_pvalue |
|---|---|---|
| 1 | 928.818987 | 5.333691e-204 |
| 2 | 1834.309316 | 0.000000e+00 |
| 3 | 2713.712826 | 0.000000e+00 |
| 4 | 3568.678890 | 0.000000e+00 |
| 5 | 4401.499486 | 0.000000e+00 |
| 6 | 5214.266084 | 0.000000e+00 |
| 7 | 6010.910499 | 0.000000e+00 |
| 8 | 6792.604186 | 0.000000e+00 |
| 9 | 7559.300054 | 0.000000e+00 |
| 10 | 8311.736911 | 0.000000e+00 |

Table 7. Ljung-Box Test results

With the Ljung-Box Test we check for autocorrelation and prove that there is such a lag of 1. (The lag 1 p-value of the Ljung-Box Test is: 6.392185e-204 which is a lot lower than 0.05 ). This is also visible in the PACF plot but it is good to have it statistically proven.

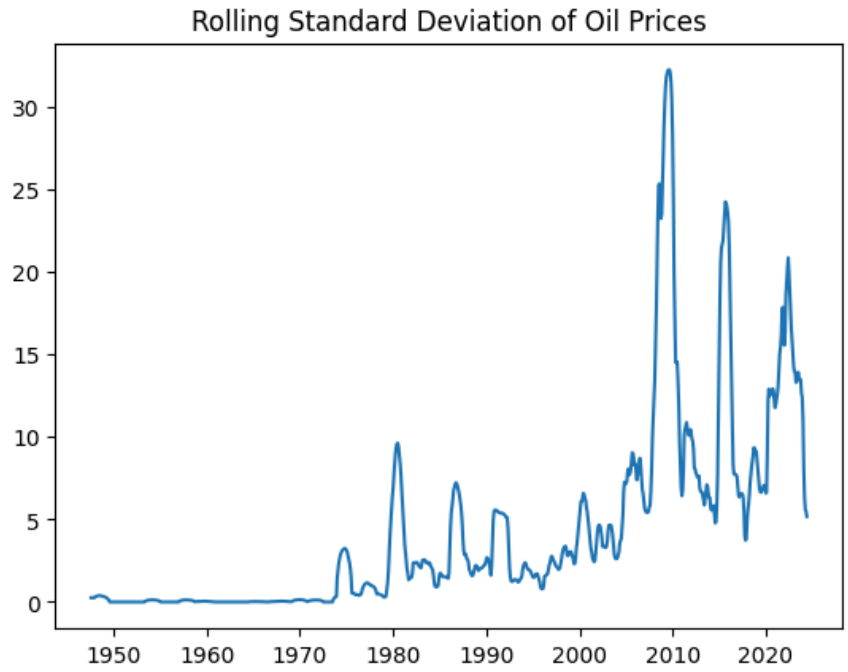



Fig. 17. Volatility clustering

It is quite normal for Oil prices to have big spikes in volatility (especially in war time periods) but also a big trough are quite normal at times of crisis where the demand and usage of oil is much lower than normally like in the 2008 Financial crisis and the Covid19 as well. So with the plot of the Rolling Standard Deviation of Oil Prices we can see exactly that: volatility clustering around periods of boom and bust (economy progress and crisis).

## Probabilistic Graphic Models:

The two major types of Probabilistic graphical models (PGM) are Markov Chain (Network) and Bayesian networks. They both are graphical representations used to model the joint probability distributions and thanks to this you can see whether or not there is dependence or independence among the given variables. The graphs are built from nodes and edges that connect the nodes. This structure is very informative and can be used for reasoning under uncertainty and visualization of the decision making process. How and when the information arrives and how it changes the progress path.

**Bayesian Networks** are visualized by Directed Acyclic Graphs (DAGs) which have directed edges that connect the nodes and show a direction of the information flow. They show a relationship between the nodes which can bring some kind of conditional dependency. This means that each node's behavior depends only on its parents. [1], [3]

**Markov Networks** are visualized by undirected graphs which lack directional edges. Also they are represented by the so-called transition matrix where some values for probabilities are located and they show what is the probability to go from one state to another which may or may not happen. When talking for Markov Chain we have to mention the Markov Property that says something like this: “Future depends only on today’s state and is independent of its past” and this means we can ignore what have happened in the past history and make decisions only on the state that we are located right now (the present state). The main differences with Bayesian Network is that there is a lack of direction towards the data flows (only probabilities from the transition matrix) and also Markov Networks can catch cyclic dependensies that otherwise Bayesian networks cannot. While Bayesian Networks rely and emphasize on directional relations and influence, Markov Chains do not, they handle the undirected dependencies. Also they both work with conditional dependencies but their main “blueprints'' are different. [1],[2]

## Markov Chains and Markov Blankets:

In the study and application of probabilistic models, Markov Chains and Markov Blankets serve as foundational elements. They provide a good mechanism for dealing with uncertainty and complexity. Both

Markov chains and Markov blankets help in efficient modeling, prediction and decision making for probabilistic dependencies. They also help in building more accurate and interpretable models which leads to better insights. [4]

Markov chains are an example of stochastic models that describe the transition from one state to another in a chain-like process. Markov chains have one of the key features of markov property which tells that the probability of transitioning to a particular state does not depend on the sequence of the events that preceded it but it depends on the current state. The transition matrix of a Markov chain is represented by the transition probability of the likelihood of moving from one state to another. Discrete-time markov chain and continuous time markov chain are generally two types of the markov chains. Markov chains can be generally used for modeling stock prices. [6]

A Markov blanket of a node is a set of nodes in the Bayesian network which protects the markov blanket from the rest of the network. The node becomes conditionally independent of the other nodes in the network when the Markov blanket becomes known. The node of a markov blanket consists of its parents, co-parents and children. In Bayesian networks, the Markov blanket is a crucial concept because it proves to be helpful in simplifying the computation of conditional probabilities. It is also used in machine learning for feature selection where for predicting the target variable it is required to find the subset of features that are most relevant. Markov blankets are used in Bayesian networks with probabilistic graphical models for diagnosis, prediction and decision making and helps in improving the performance of the model and reduces complexity. [7]

In probabilistic graphical models, both Markov chains and Markov blankets are crucial concepts that help in analysis of the stochastic processes. For providing a framework for modeling systems of memoryless property that evolves our time, markov chains are helpful in which future states depend upon the current state. On the other hand, Markov blankets make the computations more tractable and simplify the Bayesian network.

**Pseudocode for Casual Difference Algorithm:**

**1) Import and DataPrep**
Import data and do EDA
Clean the data by removing outliers, handling missing data etc.

**2) Graph Structure Learning**
Plot and Graph the data and learn the structural dependencies
Create DAG graph and find the causal relationship using constraint-based and score-based methods.

Find the Markov Blanket for each variable.
Check if they are symmetric and if not drop it

**3) Causal Effect Estimation**

Find the causal effects by the learned DAG
Check if variables are independent and if not create an undirected arc between them
Check if they are symmetric and if not correct them like in step 2 we did check for asymmetries

**4) Model Selection and Validation**

Set direction to adjacent and non adjacent variables and create a directional path where possible

**5) Interpretation and Reporting**

Create a report where the causal effects and relationships are shown. Summarize the insights found

## References:

1. Danish A. Alvi, Application of Probabilistic Graphical Models in Forecasting Crude Oil Price, Department of Computer Science, University College London, Spring 2018
2. Nir Friedman Daphne Koller. Probabilistic Graphical Models: Principles and Techniques. 1st ed. Adaptive Computation and Machine Learning series. The MIT Press, 2009. isbn: 0262013193,9780262013192.
3. Xin-She Yang, Introduction to Algorithms for Data Mining and Machine Learning, Academic Press, 2019, Pages 19-43, ISBN 9780128172162, https://doi.org/10.1016/B978-0-12-817216-2.00009-0 , https://www.sciencedirect.com/science/article/pii/B9780128172162000090
4. Wouter Duivesteijn. Markov Chains and Hidden Markov Models. 2006.
5. Data source : https://fred.stlouisfed.org
6. Richard E. Neapolitan, Xia Jiang, Probabilistic Methods for Financial and Marketing Informatics, Probabilistic Methods for Financial and Marketing Informatics, 2007 , https://www.sciencedirect.com/topics/computer-science/markov-blanket
7. Jean-Philippe Pellet and André Elisseeff. “Using Markov blankets for causal structure learning”. In: Journal of Machine Learning Research 9.Jul (2008), pp. 1295–1342.