| FULL LEGAL NAME | **LOCATION (COUNTRY)** | **EMAIL ADDRESS** | **MARK X FOR ANY NON-CONTRIBUTING MEMBER** |
| --- | --- | --- | --- |
| Amit Sadu Kesarkar | India | akesarkar81@gmail.com |  |
| Kumar Shantanu | Poland | kumarshan25@gmail.com |  |
| Vinay Harsh | India | vhc7769@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** | **Amit Sadu Kesarkar** |
| **Team member 2** | **Kumar Shantanu** |
| **Team member 3** | **Vinay Harsh** |

| 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**

**Problem and Overview of Bayesian networks**

**Problem Statement:**

Crude oil prices are pivotal indicators in the global economy, prompting extensive efforts from governments and businesses to forecast their future trends. However, predicting these prices remains a challenging endeavor. Traditional methods such as linear regressions and econometrics are frequently used, but alternative approaches, including structural models and computer-guided analytics, are also explored. Despite these efforts, a consensus on the optimal forecasting approach is still lacking (Ross).

Part of the problem in predicting oil prices, therefore, lies in its sensitivity to disruptions in the global demand and supply of oil. A whole host of factors can affect these prices: geopolitical tensions, economic growth patterns, technological advances, and environmental concerns. This complex interplay is such that changes in supply and demand dynamics can be so sudden and unpredictable that oil prices become erratic (Perry). This is because any economic disruption will lead to magnified effects in the current interconnected global economy. Geopolitical events, consumer behavior changes, or a technological breakthrough in the energy sector will instantly and massively reflect the oil price change. A complex web of these different facts predicts oil prices complex, like a jigsaw puzzle.

Bayesian models are known for their superior predictive accuracy compared to conventional time-series analysis. They are commonly applied in forecasting GDP, inflation, consumer prices, and exchange rates. The Bayesian normal multiple regression model with informative priors is particularly relevant here, incorporating insights from the current and anticipated state of the oil market (Lee & Huh). Bayesian networks are ideally suited for long-term oil price forecasting due to their ability to capture complex relationships and uncertainties inherent in economic and market dynamics. Long-term forecasting requires a methodology capable of handling intricate interdependencies among numerous variables and adapting to evolving conditions over extended periods.

Graphical representations, such as Bayesian networks, offer a valuable framework for visualizing underlying probability structures and developing innovative models. These networks elucidate the relationships among variables, effectively handling complex probability problems. An essential attribute of Bayesian networks is their capacity to identify and incorporate relevant features into the decision-making context. This ensures a comprehensive exploration of all pertinent elements when resolving a problem. Bayesian networks provide greater flexibility and scalability compared to alternative network structures and learning methods. Updating a Bayesian network with new data is straightforward, requiring minimal adjustments. Additionally, the graphical representation of Bayesian networks is easy for both humans and computers to understand, unlike other networks such as neural networks, which pose challenges for human comprehension (Ohri).

With the help of Bayesian networks, integrated with far-reaching sets of data on geopolitical developments, economic growth indicators, and technological and environmental issues, among others, a robust model for long-term forecasting of oil prices can be developed. The chances are high that this may capture complex interdependencies and adjust to changing conditions in a way that is more accurate and insightful than available methodologies do.

The specific problem that the thesis addresses is the need for more accurate and reliable long-term forecasting of crude oil prices. Bayesian networks can be applied in modeling these relationships between different factors that lead to the cost of oil, as it is complex and uncertain. Traditional methods simply cannot handle the dynamic and interdependent nature of these factors; hence, Bayesian networks are the ideal choice for long-term forecasting.

Among the advantages of Bayesian networks is that they make it possible to include prior knowledge and update predictions with new data as it becomes available. This adaptability is an essential point in the forecasting of oil prices because, within months, new geopolitical events or technological improvements may easily change the state of market conditions. In addition, since Bayesian networks are graphical, relationships between variables can be visualized, leading to improved human and computational analysis.

This work aims to create a complete framework for the multiple complexities that arise when trying to forecast the price of crude oil and, therefore, to offer something practical in decision-making within the world economy.

**Suitability of bayesian networks for oil price forecasting**

1. Dealing with uncertain and complex relationships: Estimating the price of crude oil requires considering a wide range of interrelated macroeconomic, geopolitical, and market variables. When modeling complex structures with several variables and uncertain interactions, Bayesian networks perform exceptionally well. According to the thesis, the complex relationship of global economic factors determines the price of crude oil. This covers the quantity of oil produced by nations like those in OPEC, the quantities of oil used by the developed economies (OECD), and the continuous political and economic developments that take place globally. Bayesian networks are perfect for depicting the complexities of the oil market because they can capture these complicated interactions and interdependence amongst the variables within a graphical structure.
2. Understanding the market structure: One of the primary issues addressed in the thesis is understanding the workings of the oil markets. Bayesian networks are capable of learning a model's structure and parameters from its data. The paper explores two primary methodologies for structure learning using Bayesian networks:
   1. Constraint-based learning: This method uses statistical tests to discover conditional independencies amongst variables and then builds a network that fulfills the constraints.
   2. Score-based learning: This approach examines a set of alternative graph topologies to discover the one that best matches the data based on a scoring system.
3. Addressing inaccurate, missing, or incomplete economic and financial information often results in noise. Given that Bayesian networks can execute probabilistic inference with imperfect knowledge, they are resilient to noisy or incomplete data. Due to this, they are especially well-suited for practical uses in financial markets, where data quality might fluctuate.
4. Including domain information: Domain expertise can be incorporated into the model using Bayesian networks. This means that domain knowledge from market specialists, energy strategists, and economists can be added to the prior probability or network structure in light of oil price predictions. Expert insight and data-driven learning together have the potential to produce models that are more precise and understandable.
5. Probabilistic estimations: Bayesian networks, as opposed to deterministic models, produce probabilistic estimates, which are more valuable for making decisions in unpredictable contexts like financial markets. The goal of the thesis is to present an accurate crude oil price estimate, and belief networks can offer probability distributions and confidence intervals for future oil prices in addition to point estimates.
6. Causality: Analyzing the nuances of the oil market requires the ability to depict causal linkages between variables, something Bayesian networks can provide. The aim of establishing a probabilistic graphical model to illustrate the motion of the oil market and establish the causal link between these many variables is addressed in the paper. In the energy industry, this causal reasoning capacity can help with making policies and strategic choices by enabling better interpretation of models.
7. Resilience towards fresh data: When new data is accessible, Bayesian networks may be adjusted effectively. This is especially crucial in the oil industry, which is changing swiftly and is susceptible to price changes due to market trends, geopolitical events, and new economic data. Bayesian networks are well-suited for keeping forecasts current because of their capacity to integrate new information and change probability.
8. Managing numerous time scales: A variety of time-scale elements, ranging from short-term supply interruptions to long-term changes in the economy, have an impact on the oil market. Bayesian networks can describe temporal relationships and produce multi-quarter predictions, especially when paired with methods like the Dynamic Bayesian Networks or the Hidden Markov Models.
9. Research and application: A further issue described in the paper is the research and exploitation of existing data and learned structures for forecasting oil market behavior. Belief networks are appropriate for this purpose because they can effectively execute inference on the structure, enabling both variable exploration and prediction.
10. Model credibility: The paper goes on to discuss ways to validate the developed model's performance and dependability. Bayesian networks offer a variety of model validation techniques, like cross-validation, sensitivity analyses, and posterior predictive checks. These strategies can aid in determining the of the model and credibility before implementation in financial markets.
11. In contrast to certain black box neural network algorithms, Bayesian networks show the correlations between variables graphically. This interpretability is critical for financial applications, whereby understanding the logic behind forecasts is sometimes as essential as the estimates themselves. The paper seeks to gain insight into the operations of the oil markets and the belief networks aid this purpose by offering perspectives on the market's structure and behavior.
12. Nonlinear connections amongst the variables are common in oil pricing. Bayesian networks may capture these nonlinear relationships using conditional probability tables and continuous distributions, resulting in a more accurate picture of the oil market's complex relations.
13. With the Bayesian networks, variables may be easily manipulated to perform hypothetical studies and stress tests. This is consistent with the paper's purpose of replicating economic hardship scenarios to test the model's dependability. Analysts can investigate how alternative scenarios may affect oil prices and evaluate how the model performs under various situations by modifying the values or probability of specific variables. The paper addresses using data from a variety of sources, such as the Energy Information Administration (EIA) and the Federal Reserve Economic Data (FRED). Bayesian networks may successfully integrate information from numerous sources, merging disparate datasets to create one unified model of the oil market. Due to the many factors impacting the price of oil, the issue naturally requires data that is highly dimensional. Bayesian networks can effectively produce and infer highly dimensional probability distributions, which makes them ideal for this challenging forecasting challenge.

**Step 4**

**Oil Price Related Data**

| **series\_id** | **description** | **unit** | **n\_observations** |
| --- | --- | --- | --- |
| PAPR\_NONOPEC | Total non-OPEC Liquids Petroleum Production | million barrels per day | 396 |
| PAPR\_OPEC | Total OPEC Petroleum Supply | million barrels per day | 396 |
| PATC\_OECD | Liquid Fuels Consumption, Total OECD | million barrels per day | 432 |
| PATC\_NON\_OECD | Liquid Fuels Consumption, Total non-OECD | million barrels per day | 432 |
| COPRPUS | U.S. Crude Oil Production | million barrels per day | 432 |
| CORIPUS | Crude OIl Refinery Input | million barrels per day | 432 |
| PASC\_OECD\_T3 | OECD End-of-period Commercial Crude Oil and Other Liquids Inventory | million barrels, end-of-period | 276 |
| COPS\_OPEC | OPEC Total Spare Crude Oil Production Capacity | million barrels per day | 396 |
| COPC\_OPEC | OPEC Total Crude Oil Production Capacity | million barrels per day | 432 |
| T3\_STCHANGE\_OOECD | Net Inventory Withdrawals, Total Other OECD Crude Oil and Other Liquids | million barrels per day | 275 |
| T3\_STCHANGE\_NOECD | Net Inventory Withdrawals, Total Non-OECD Crude Oil and Other Liquids | million barrels per day | 275 |

**Data Dictionary**

| **series\_id** | **description** | **unit** | **n\_observations** |
| --- | --- | --- | --- |
| **Global Inflation** | **Global Inflation** |  | **289** |
| **U.S. Unemployment** | **U.S. Unemployment** |  | **289** |
| **U.S. Interest Rate** | **U.S. Interest Rate** |  | **289** |
| **EU Unemployment** | **EU Unemployment** |  | **80** |
| **EU Interest Rate** | **EU Interest Rate** |  | **289** |
| **Current Account Balance** | **Current Account Balance** |  | **50** |
| **Imports of Goods and Services** | **Imports of Goods and Services** |  | **50** |
| **Exports of Goods and Services** | **Exports of Goods and Services** |  | **50** |
| **External Debt, Total** | **External Debt, Total** |  | **50** |
| **Gross Domestic Product, Current Prices** | **Gross Domestic Product, Current Prices** |  | **50** |
| **Crude Oil Prices** | **Crude Oil Prices** |  | **50** |

**Geopolitical Data**

| **Country** | **Political Stability Index** | **Country** | **Political Stability Index** | **Country** | **Political Stability Index** | **Country** | **Political Stability Index** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Iran | 24 | Iraq | 24 | Kuwait | 24 | Qatar | 24 |
| Saudi Arabia | 24 | United Arab Emirates | 24 | United States | 24 | Russia | 24 |
| Canada | 24 | China | 24 | Mexico | 24 | Venezuela | 24 |
| Brazil | 24 | Nigeria | 24 | Netherlands | 24 | Norway | 24 |
| Kazakhstan | 24 | Angola | 24 | Algeria | 24 | Colombia | 24 |
| Gabon | 24 | Oman | 24 | Egypt | 24 | Equatorial Guinea | 24 |
| Liberia | 24 | Lebanon | 24 | Saint Lucia | 24 | Libya | 24 |
| Malaysia | 24 | Syria | 24 | Sudan | 24 | Turkmenistan | 24 |
| Trinidad and Tobago | 24 | Tunisia | 24 | Uzbekistan | 24 | Yemen | 24 |

**Step 5**

**Extreme Outlier Treatment**

[STUDENT A TO WRITE]

**Bad Data Treatment**

[STUDENT A TO WRITE]

**Treatment of Missing Values**

Some datasets primarily the political stability data was collected on a yearly basis. Therefore, to address this limitation, we adopted the linear interpolation methodology to transform the dataset from a yearly resolution to a daily resolution.

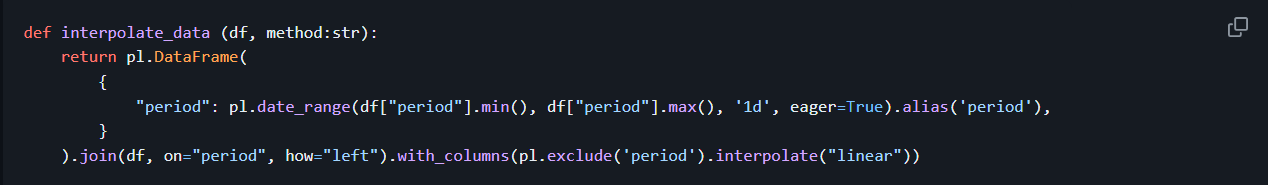
Mathematically, linear interpolation can be expressed as follows:

Where: - and are the known data points, is the point at which the value is to be estimated, is the interpolated value at .

In the context of our study, represents the date for which the value is being interpolated, and represents the corresponding data value.

The interpolation process ensures a smooth transition between data points, effectively approximating the daily values while maintaining the overall trend and seasonality inherent in the original yearly data. The result is a finely granulated dataset suitable for detailed analysis and modeling.

The following code achieved this:



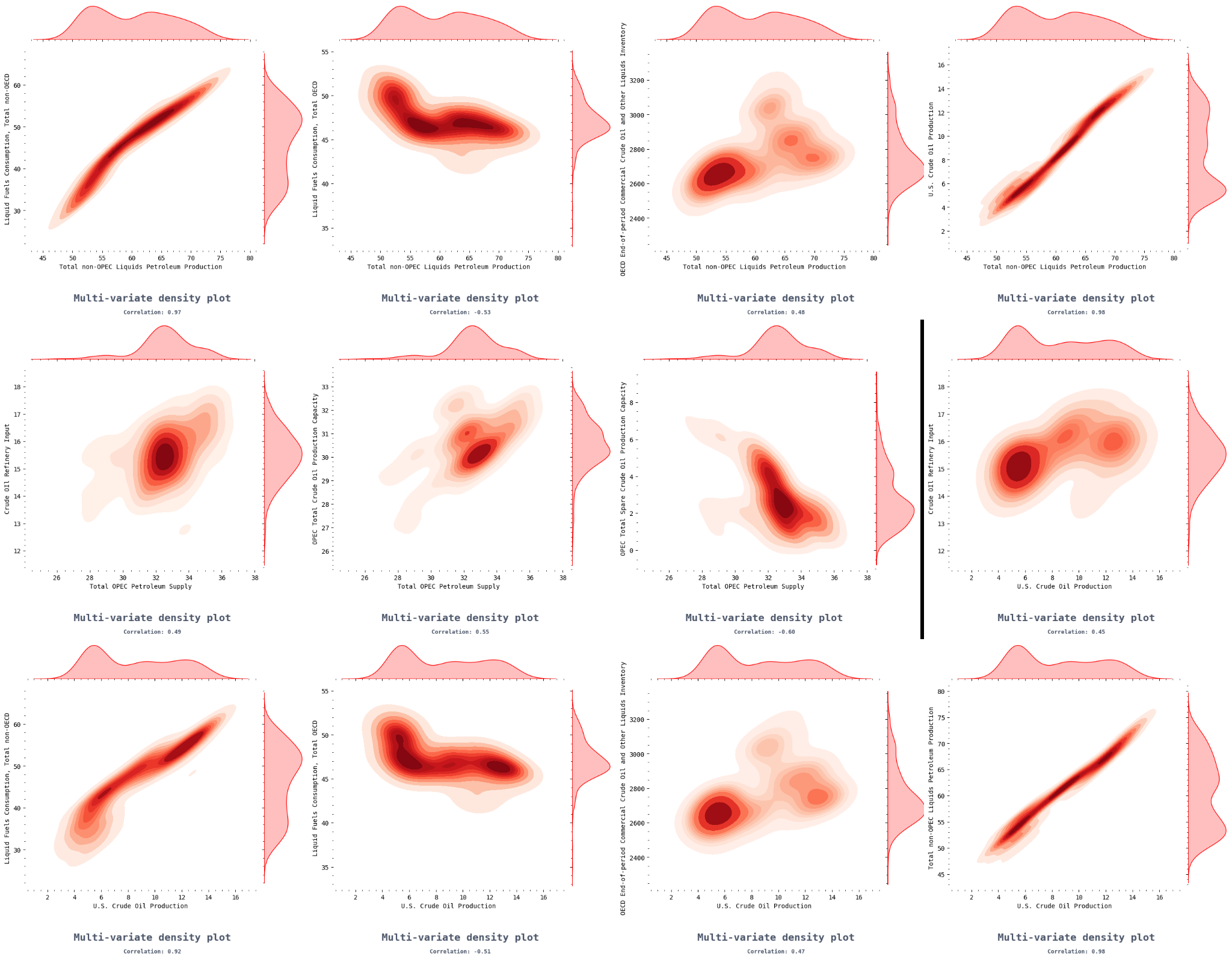
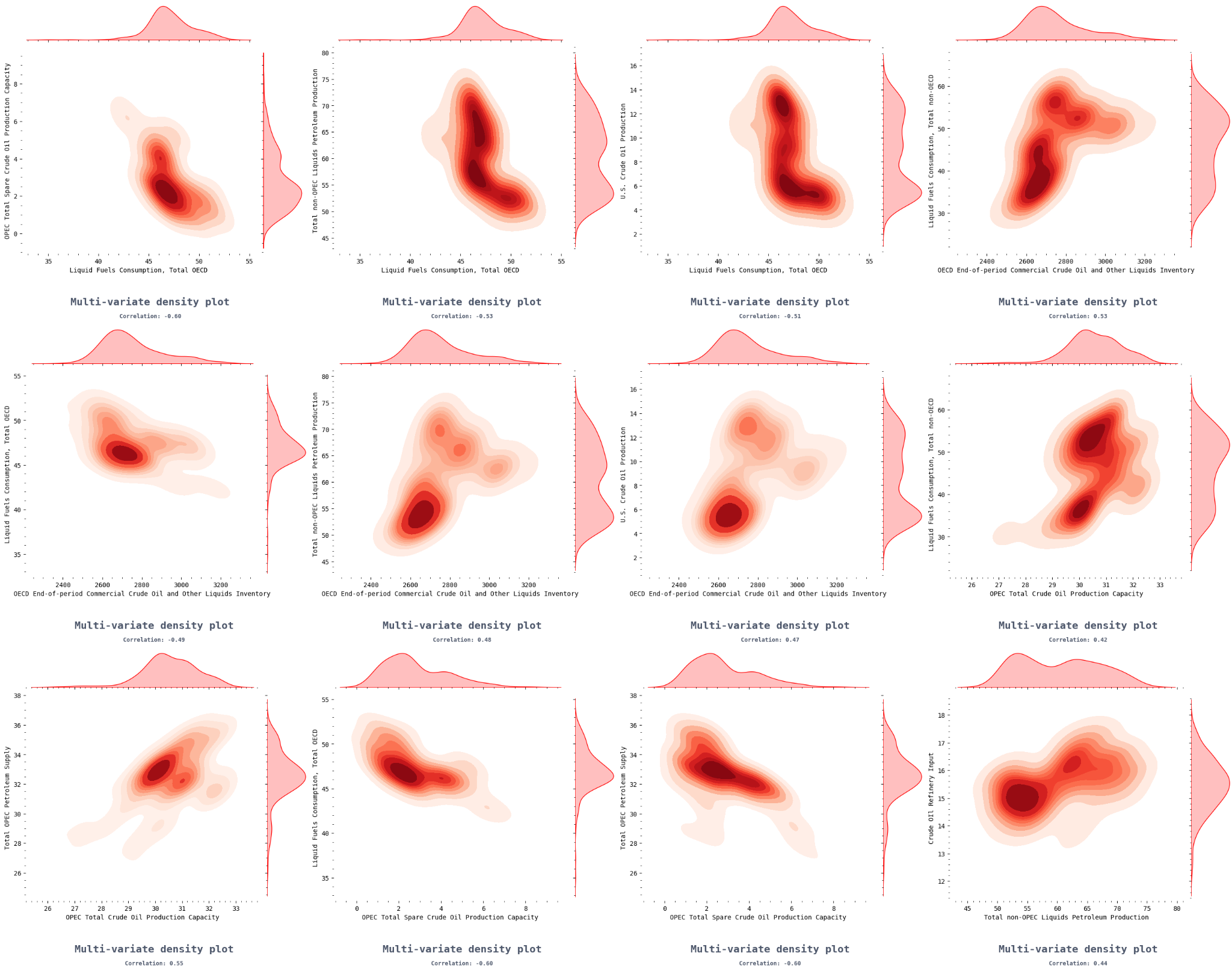
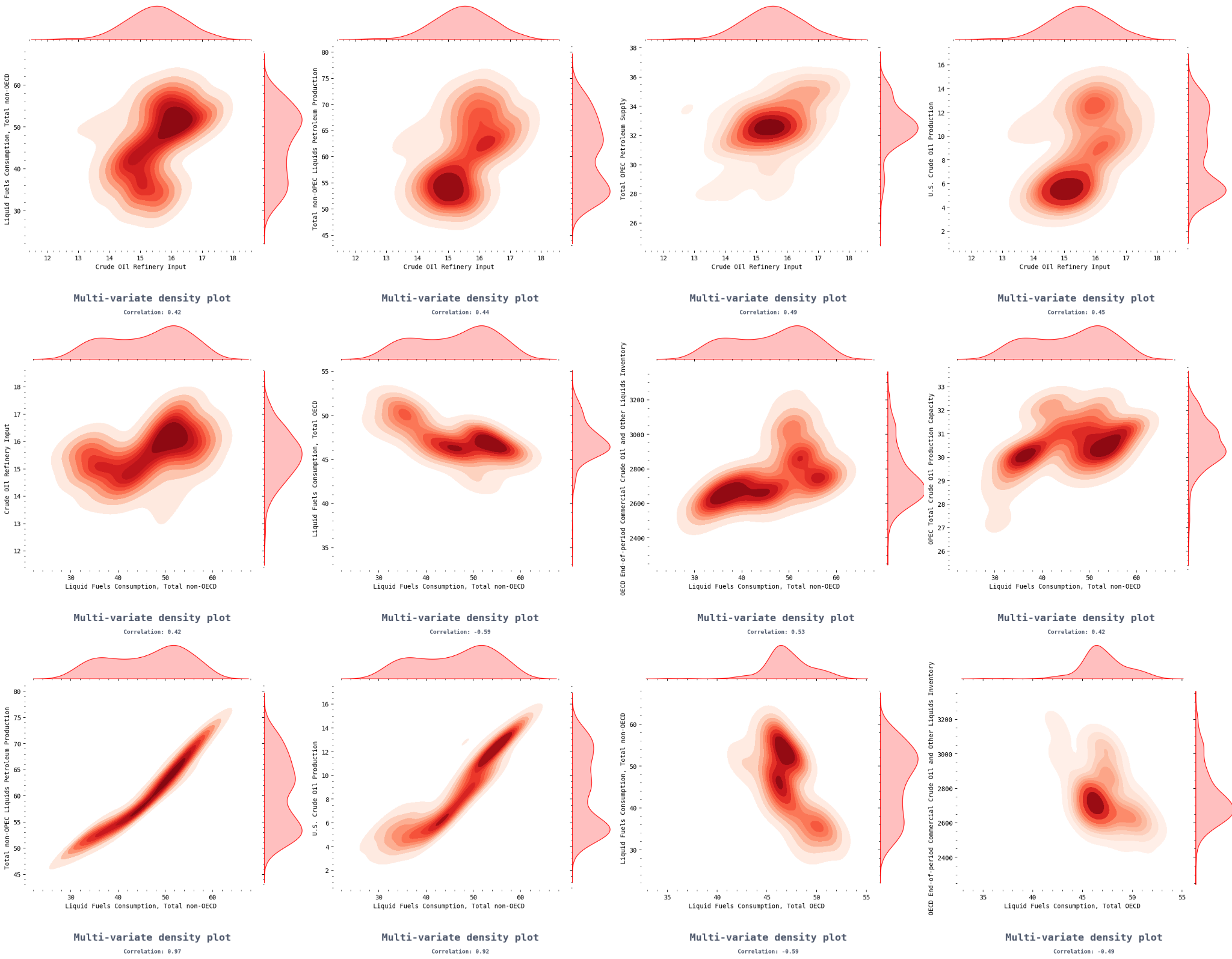
**Step 6**

**Step 7**

**Data Visualisation**

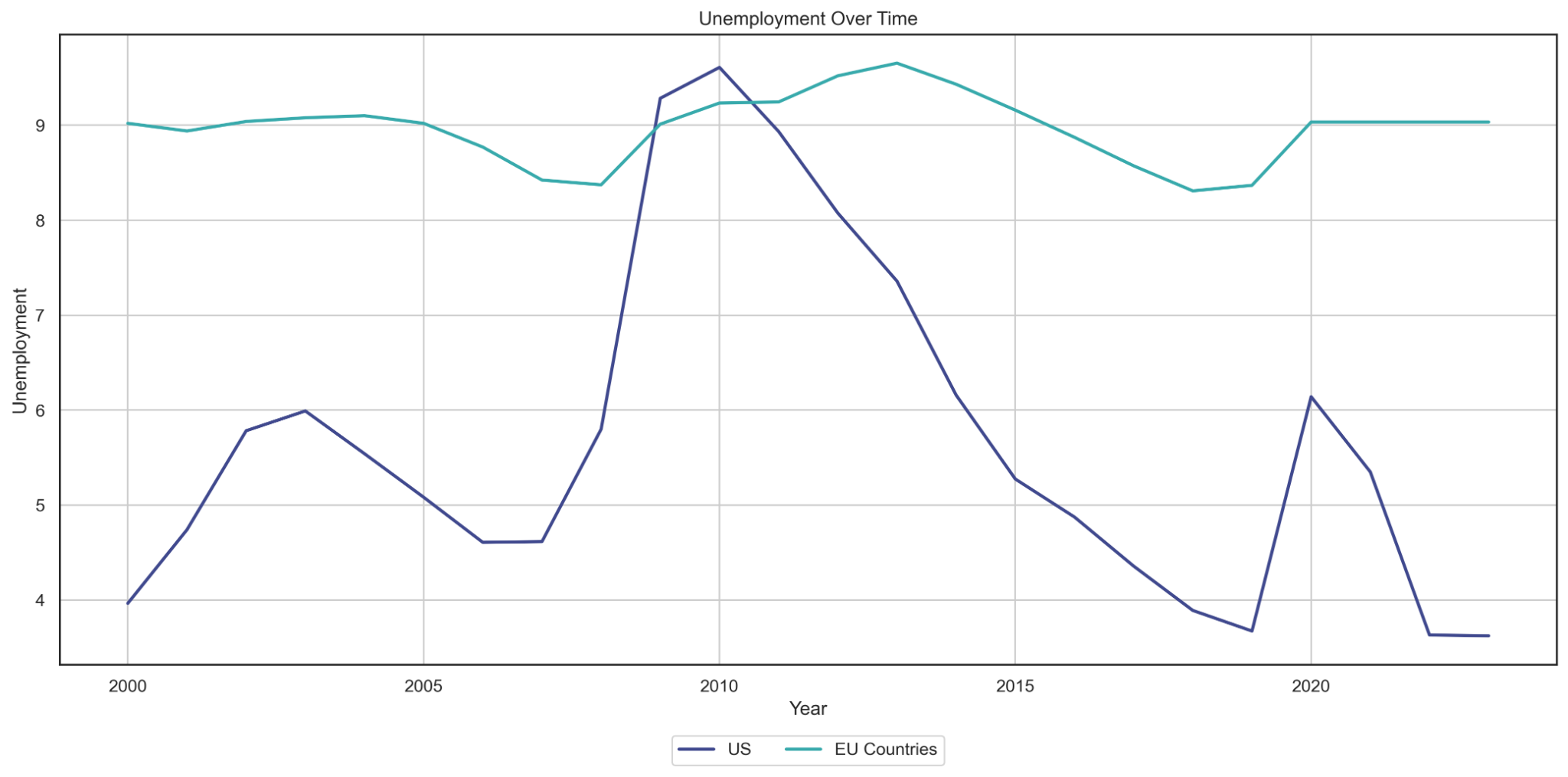
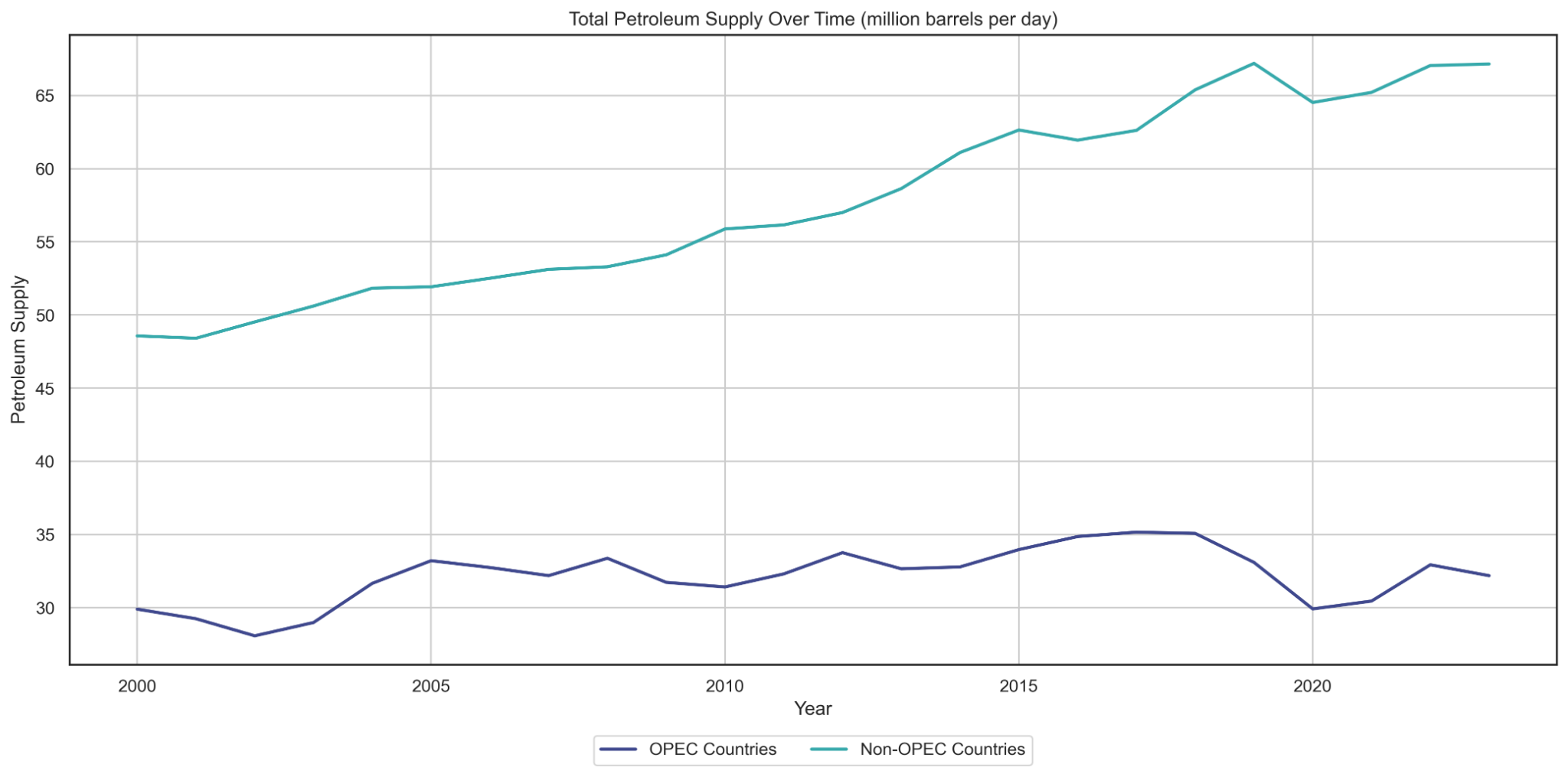
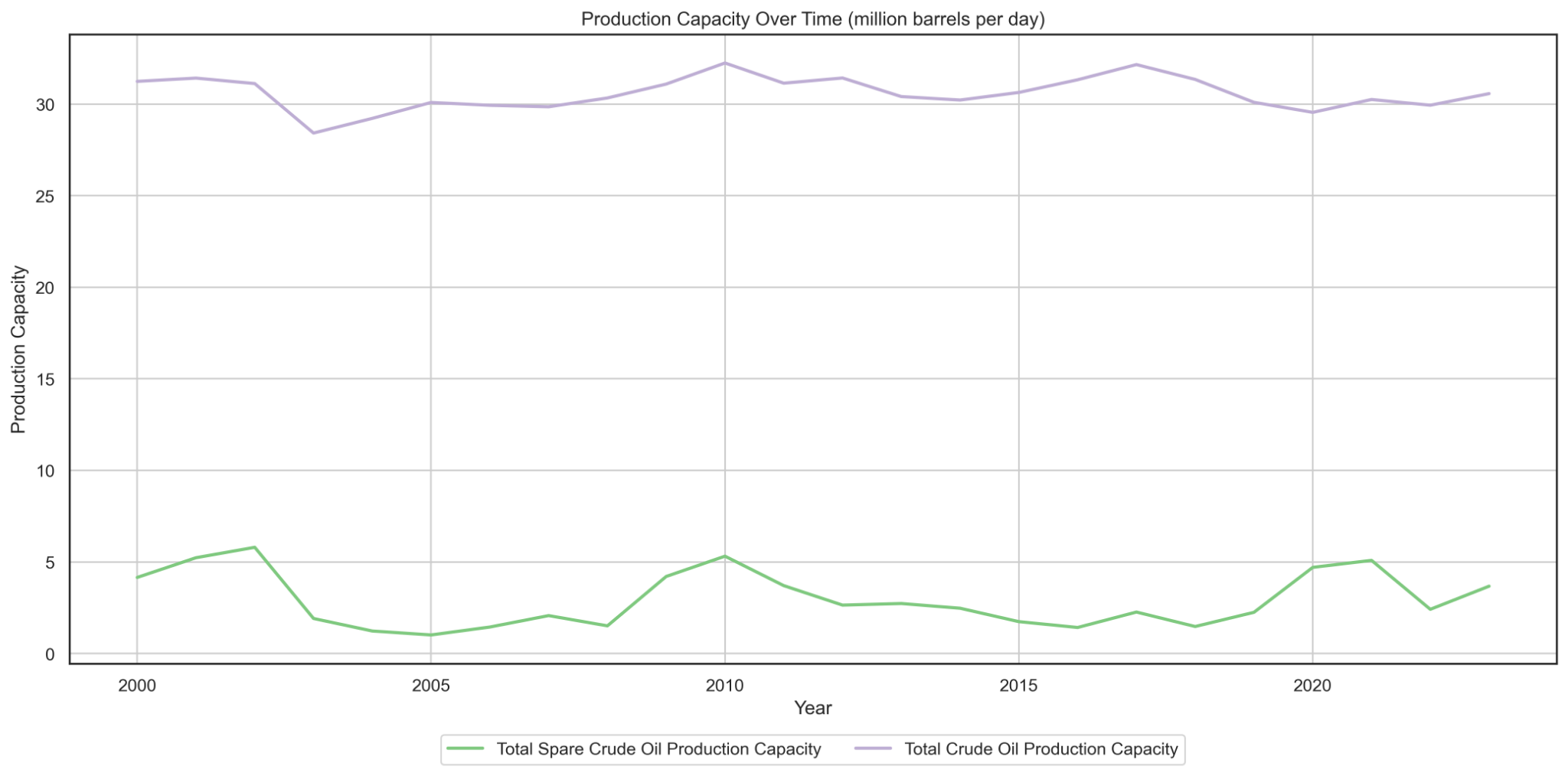
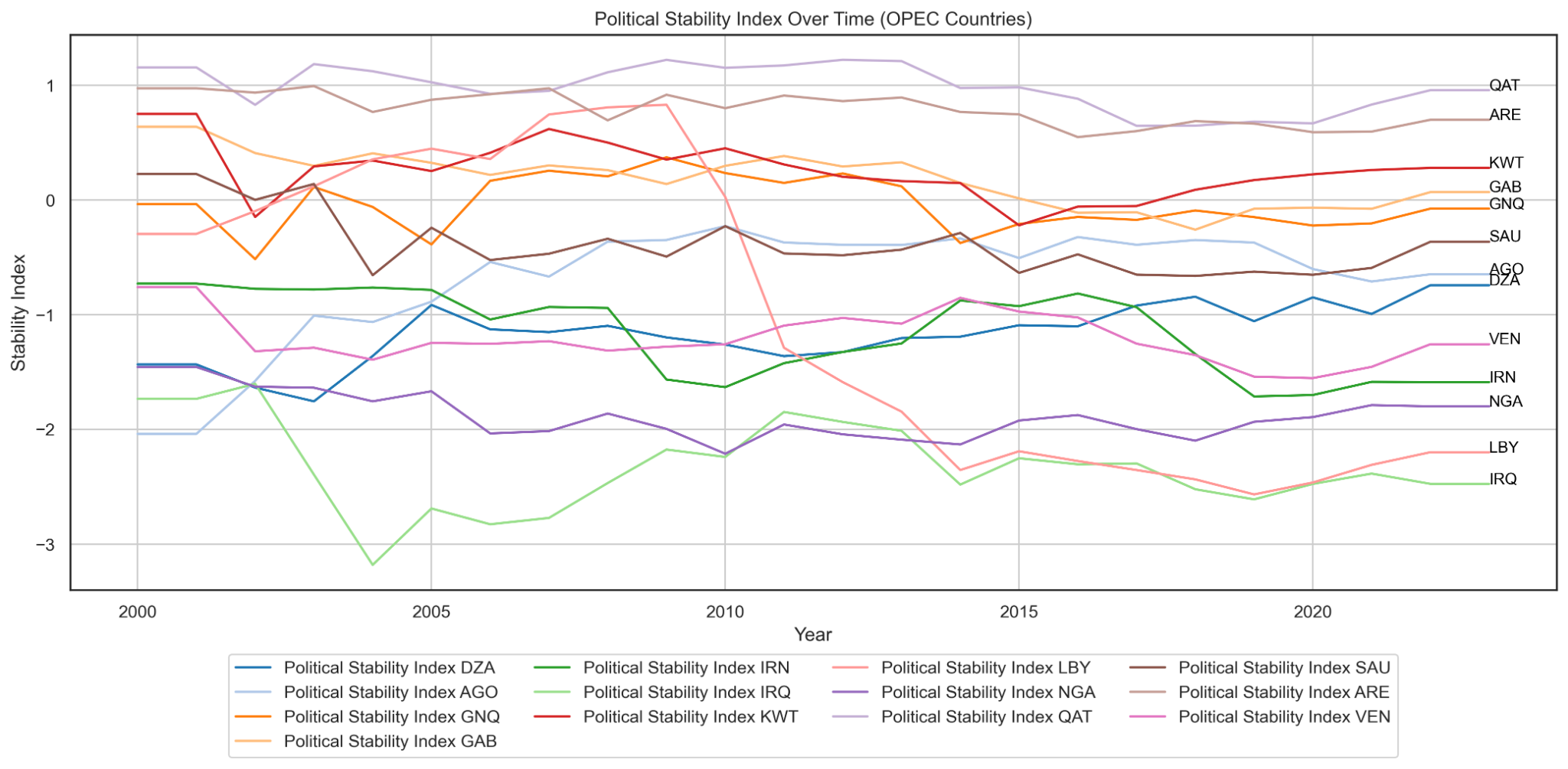
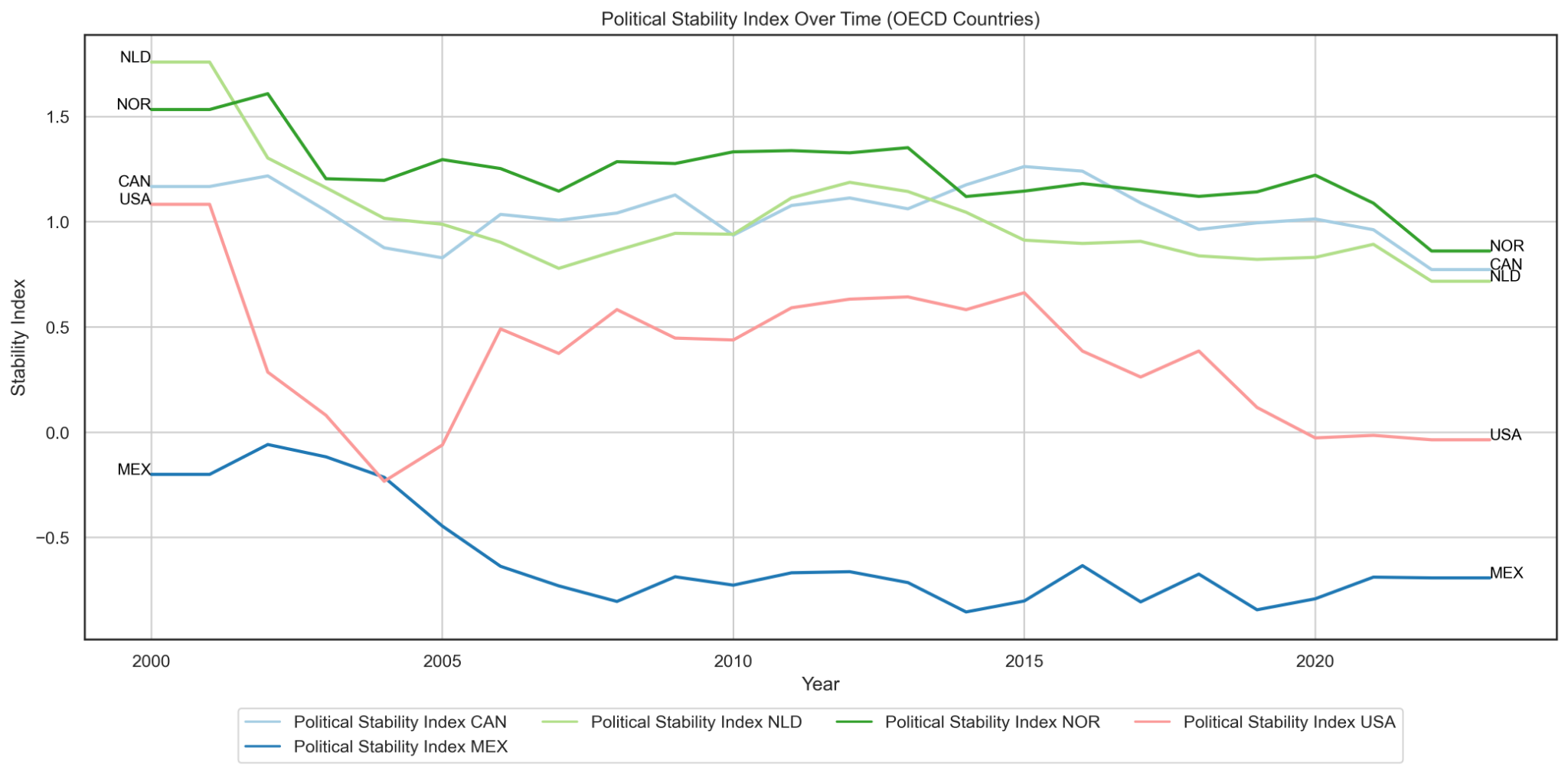
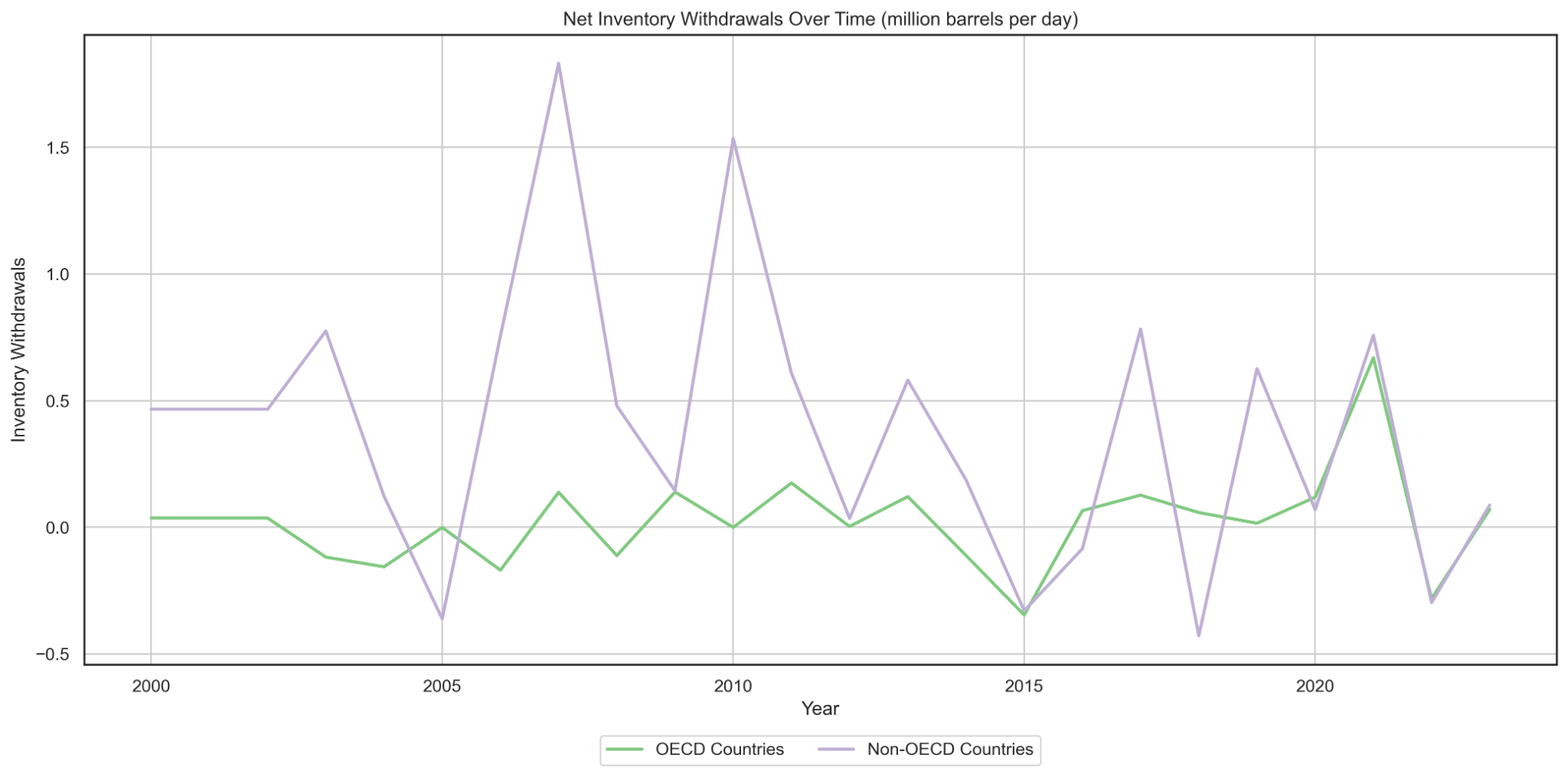
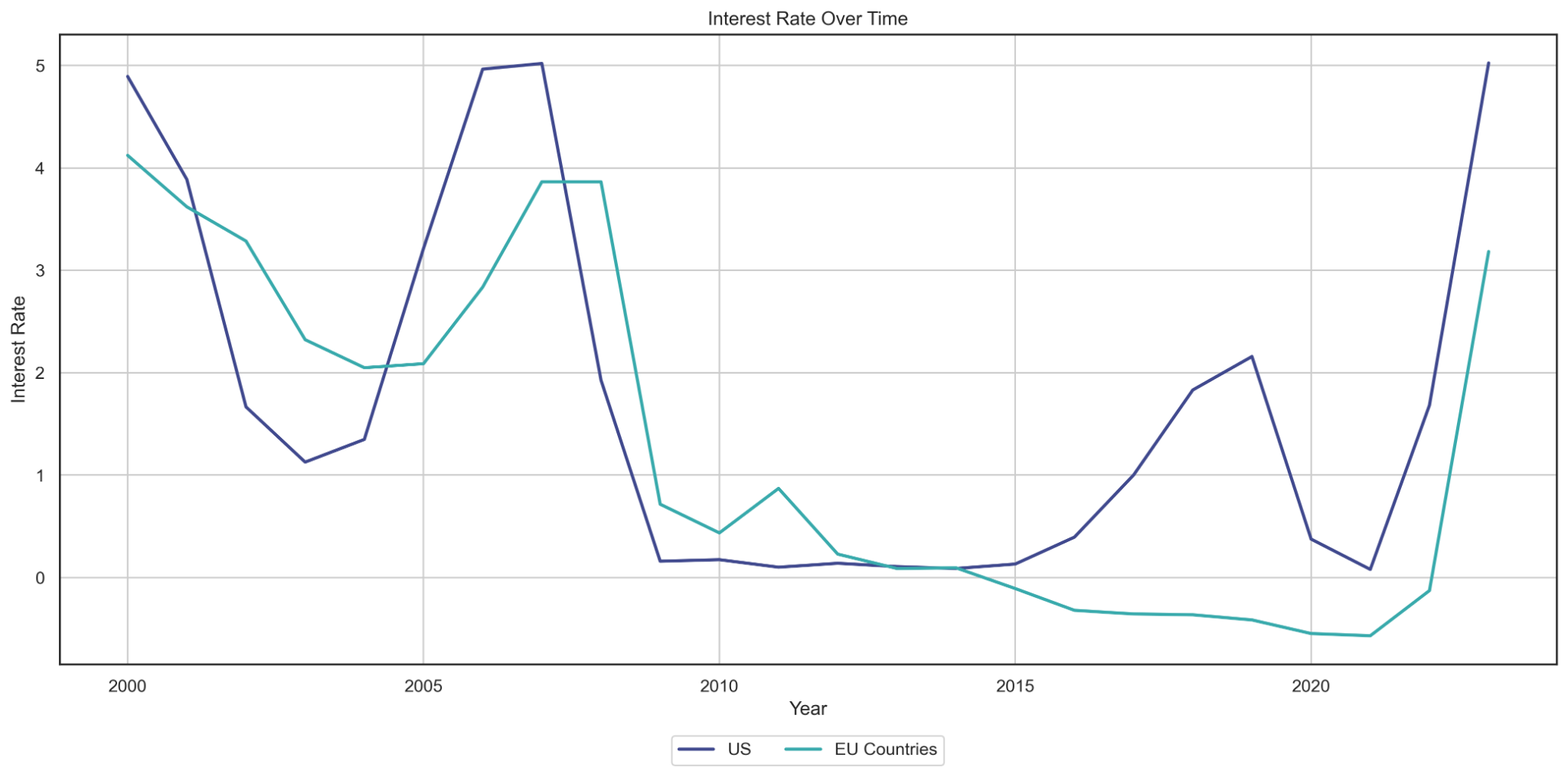
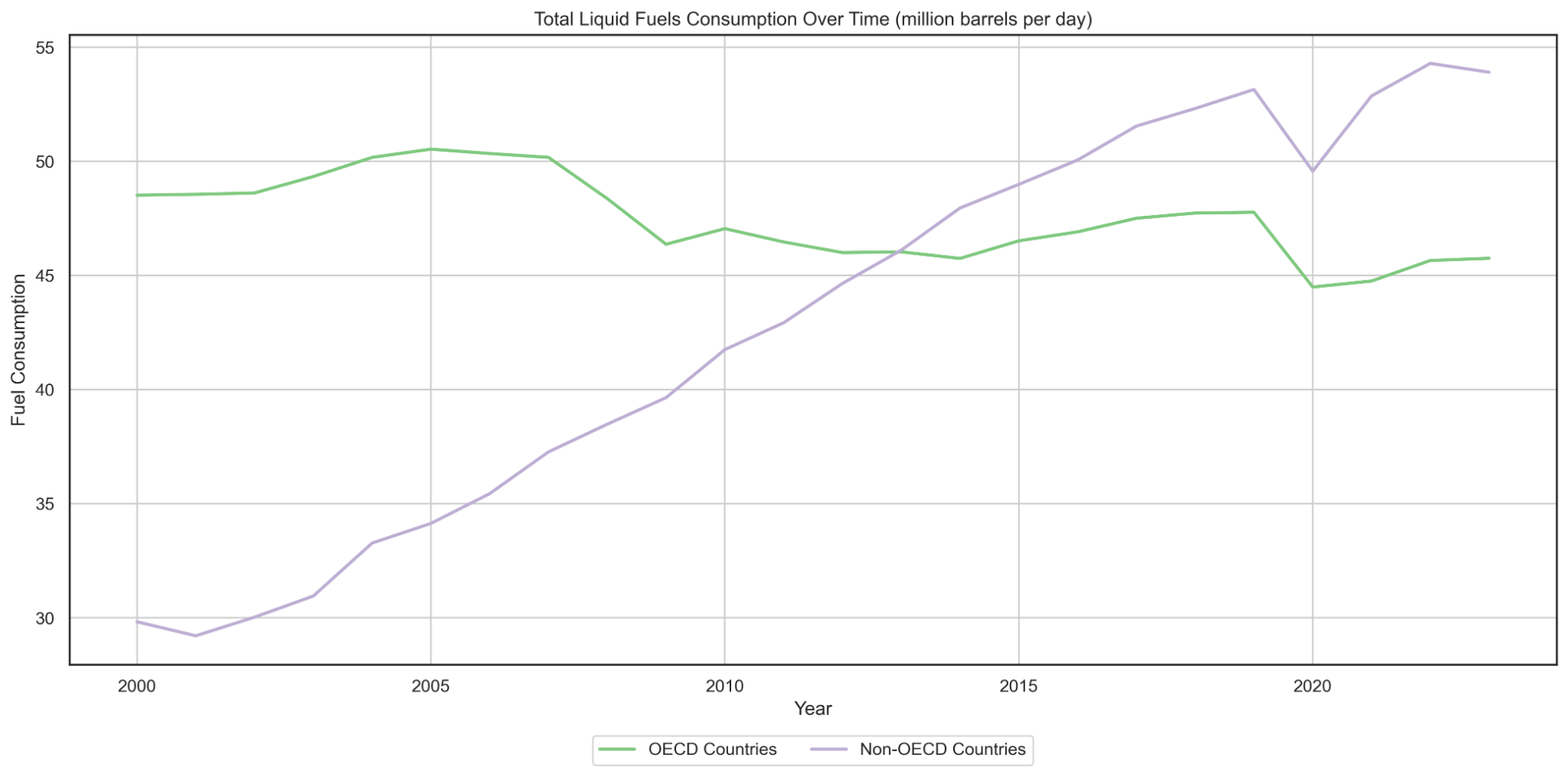
a. Distributional and Multivariate Kernel Density Plots

In our study, we subsetted the highly correlated variables to enhance the clarity and interpretability of the multivariate density plots. The plots are presented as follows:



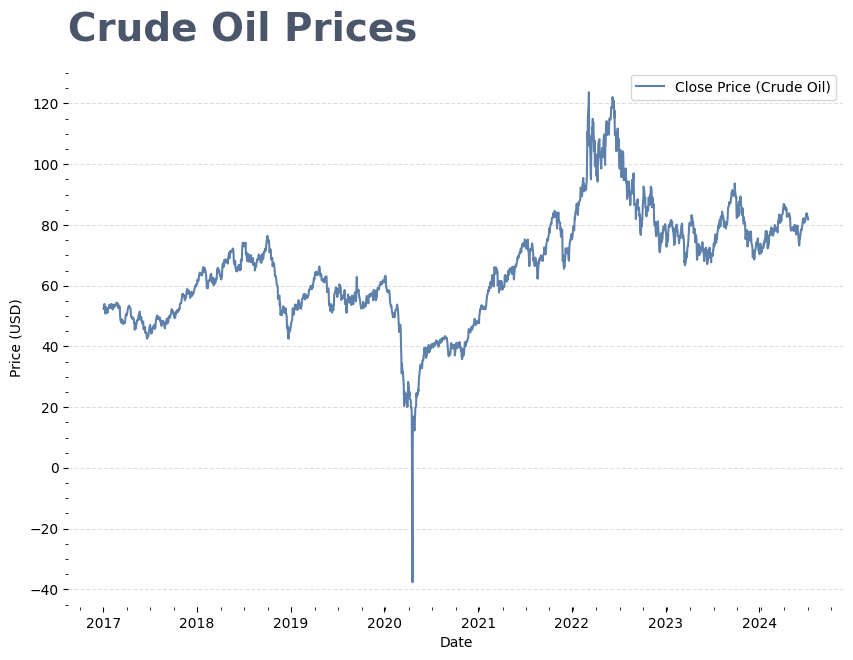
Time Series Plots

The time series plots are shown as follows:

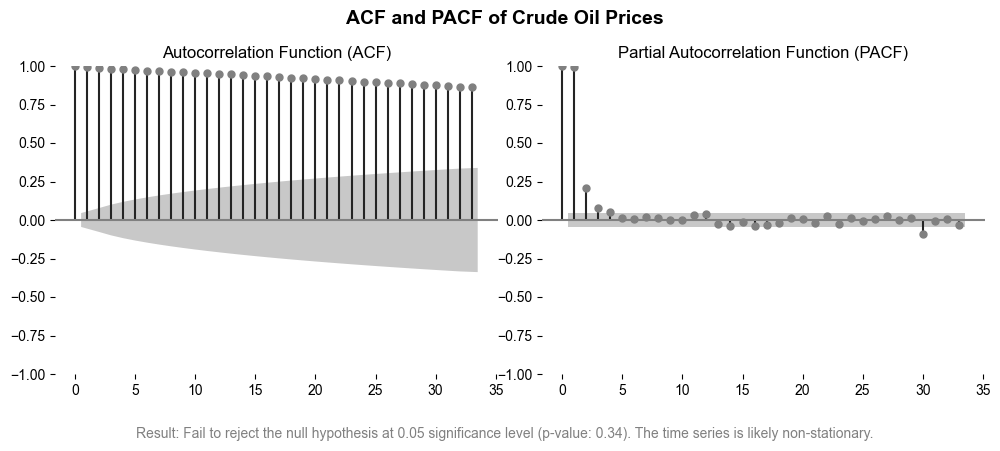


**Step 8**

**Oil Prices**

****

Oil prices exhibit significant volatility clustering, a phenomenon characterized by periods of high and low volatility.



The PACF of oil prices reveals that significant autocorrelations taper off after lag 3, indicating that each oil price return is directly influenced by its immediate three preceding returns (when controlled for preceding lags), with minimal direct effect beyond this horizon.

**Step 9**

**Model**

**Probabilistic Graphical Models**

**[STUDENT A TO WRITE]**

**Parameter Learning**

Two key objectives in Bayesian networks are parameter and structure learning. Parameter learning involves determining numerical values that describe the connections between variables in a specific network structure. On the other hand, structure learning is concerned with discovering the network's structure or determining which factors are closely linked.

Parameter learning refers to calculating conditional probability distributions (CPDs) of variables within a Bayesian network. With a network structure, parameter learning aims at determining the probabilities that influence the relations of parent and child nodes.

Assuming a Bayesian network structure that has two nodes: (parent node) and (child node). Conditional probability defines the link between and . If has two possible states and , and likewise has two possible states and , parameter learning requires estimating the probabilities:

|  |  |  |  |
| --- | --- | --- | --- |

These probabilities may be calculated using techniques like maximum likelihood estimation (MLE) and Bayesian estimation.

Maximum Likelihood Estimation (MLE):

MLE includes determining the value of parameters which maximize the likelihood for the data. For example, if observed data displays how frequently and take on particular states, those findings can be used to compute the probabilities which maximize the probability of the data considering the network structure.

Bayesian Estimation:

Bayesian estimate uses prior understanding of the parameters and then updates it depending on the observed data. This method includes historical distributions with the probability of data to generate posterior distributions for all the parameters.

Assume a network in which a node affects nodes & . The conditional probabilities and have to be determined. Using past financial data, parameter learning approaches like as MLE may estimate these probabilities while assessing the likelihood of various stock prices and trade volumes based on market trends.

**Structured Learning**

Structure learning attempts to discover the network structure itself and precisely what variables are closely connected by edges. The technique is considerably complicated than parameter learning since it requires searching through a wide array of network structures to discover the one which reflects the relationships in data most accurately.

Techniques for Structure Learning:

1. Score-Based: These approaches use a scoring system to assess network structures, balancing model fit with complexity. The Bayesian Information Criterion (BIC) along with the Akaike Information Criterion (AIC) are two often used scoring functions. The aim is to identify the structure that results in maximizing the score.
2. Constraint-Based: They make use of conditional independence tests to evaluate network structure. These tests determine if 2 variables have been conditionally independent based on a set of other factors. The network is built by connecting variables that are determined as conditionally dependent.

Differences

* Parameter Learning:
  + Estimates numerical CPD values for a particular network structure.
  + Less computationally demanding since it employs a defined structure to estimate probability.
  + Needs a defined network structure for determining parameters.
* Structure Learning:
  + Determines the network's structure by recognizing which factors are closely linked.
  + Computationally demanding, requiring a search across a space of potential network structures that expands exponentially with the number of variables.
  + Can be conducted with no prior knowledge of network structure, however, domain knowledge is typically useful for limiting the search space.

Parameter learning seeks to estimate the CPDs from a network structure, whereas structure learning seeks to reveal the network structure. Both processes are required for the creation of accurate and comprehensible models, with parameter learning giving numeric correlations and structure learning establishing the network's structure. Understanding and using these concepts enables experts to use Bayesian networks to simulate complicated connections.

**Markov Chains and Blankets**

**Step 10**

**References**