University of Essex
Department of Mathematical Sciences

MA981: DISSERTATION

Data driven Environmental Impact Assessment of Oil and Gas

Potti Indu Rao
2211979

Supervisor: Dr Terry Sithole

Professor, University of Essex

November 24, 2023
Colchester

Abstract

This analysis examines the environmental impacts of oil and gas operations through a comprehensive dataset tracking energy consumption and production metrics across countries and years. The overarching goal is to elucidate the complex interrelationships between energy usage patterns and ecological consequences.

The background introduces oil and gas as essential yet contentious energy sources that enable modern societies but contribute extensive environmental damages. Upstream extraction and distribution infrastructure intrinsically emit air or water pollutants, induce seismicity, fragment habitats, and accelerate climate change, among other impacts. Quantifying and predicting cumulative damages remains challenging.

Machine learning promises more highly advanced predictive analytics to overcome limitations constraining conventional environmental assessments. By discerning influential variable relationships within immense datasets, algorithms can simulate intricate real world dynamics exceeding current models. This data driven approach enables superior emissions forecasting, optimized balancing of production and ecological stability, and timely incident detection.

The methodology utilizes a combination of analytical techniques including linear regression modeling, time series analysis, machine learning comparisons and diverse visualizations to explore the dataset's subtleties. The dataset incorporates country-level metrics on total energy as well as specific contributions from oil, gas and per capita over recent decades.

Key findings reveal an escalating global emissions trajectory reflecting rising fossil fuel reliance with the highest contributions from the United States, China and Russia. Modeling demonstrates sophisticated machine learning algorithms like Random Forest and Gradient Boosting prove significantly more effective at capturing complex temporal emissions patterns than linear regression.

Per capita figures highlight the role of individual consumption in collective environmental

footprints. Emissions variability between nations underscores the need for tailored interventions. Strong correlations present between oil and gas emissions and total energy usage further indicate their substantial climate impact.

These data driven insights emphasize the pressing need to mitigate adverse oil/gas emissions and accelerate sustainable energy transitions, involving policies to reduce fossil fuel dependence, improve efficiency and incentivize renewable adoption. Integrating direct emissions data and refining country specific models can further enhance these discoveries.

Contents

List of Figures

CHAPTER 1

Introduction

1.1 Background on Oil and Gas Operations and Environmental Impacts

Oil and natural gas are critical global energy sources, used across transportation, electricity, manufacturing, buildings, and other sectors. Global oil consumption reaches over 90 million barrels daily, while natural gas demand continues rising, currently at over 3,500 billion cubic meters annually. Extracting and delivering these fossil fuels requires extensive infrastructure and intensive operations which significantly impact environments from the local to global scale.

Upstream processes include exploration to locate underground oil and gas reservoirs using seismic surveys; exploratory and production well drilling; hydraulic fracturing; and infrastructure for gathering, processing, and transporting raw production. Midstream operations involve long-distance pipelines, marine transport, rail and truck fleets; extensive storage capacity; exporting; and refining crude oil into various fuels and materials. Downstream distribution delivers outputs to end consumers.

The main Environmental Impacts are:

1. **Air Emissions:** Oil and gas operations emit high levels of air pollutants including

methane, volatile organic compounds (VOCs), nitrogen and sulfur oxides, particulate matter, benzene, ethylbenzene, hydrogen sulfide, polycyclic aromatic hydrocarbons, and other hazardous air toxins. Sources include venting, flaring, fossil fuel combustion, evaporation losses, fugitive leaks from infrastructure, and dust from land clearing. This contributes to respiratory illness, smog, acid rain, and climate impacts.

2. **Water Contamination:** Drilling, hydraulic fracturing, pipelines leaks, and spills release produced water, chemicals, and hydrocarbons containing heavy metals, salts, and radioactive materials that infiltrate ground and surface waters. High water extraction rates also strain local water resources.

3. **Land and Habitat:** Infrastructure construction fragments habitats, removes vegetation, expands impermeable surfaces, alters wildlife movements, facilitates invasive species, and escalates erosion. Gas flaring and diesel generators also generate noise and lighting pollution disruption. Infrastructure often remains under-reclaimed after decommissioning as well.

4. **Waste:** Drilling muds, contaminated water, oily sludges, solvents, spent chemicals and more require extensive storage, transport, treatment and disposal to avoid environmental releases. Chronic leakage occurs from onsite pits, tanks, landfills and injection wells. Vented/flared gas also represents substantial wasted energy resources.

5. **Climate Change:** Combustion of oil/gas and supply chain methane leaks emitted CO_2 and other greenhouse gases that drive climate warming, ocean acidification, sea level rise, and other global changes. The industry accounts for over 30% of global carbon emissions when including end-use combustion.

6. **Seismic Activity:** Subsurface wastewater injection from hydraulic fracturing and other processes has been tied to induced earthquakes, putting communities at risk.

7. **Cumulative Impacts:** While individual wells, pipelines or facilities have localized impacts, collectively the industry drives substantial ecosystem disruption, especially accelerating under continued fossil fuel dependency. Attempts to model total cumulative damage remain inadequate.

Routine operations have greatly improved in technology and practice yet the immense scale of global oil and gas means even narrowly contained risks accumulate. Billions of dollars in annual environmental damages still occur from major accidents like spills, pipeline ruptures, facility explosions and more. Conflict zones with minimal oversight further exacerbate risks. There remain substantial gaps in impact data collection, monitoring, scientific assessments and regulatory oversight. Development of machine learning and predictive analytics has some potential power in predicting some of the hazards in future.

1.2 Introduction to Machine Learning for Environmental Assessments

Machine learning refers to algorithms with the capacity to learn patterns within data in order to make predictions or decisions without explicit programming. By analyzing many examples in a data set, machine learning models discern influential relationships between key variables. The algorithms then leverage these relationships to forecast outcomes for new data inputs. Models enhance accuracy through continuous analysis of more data.

Machine learning represents a shift from rules-based software and statistical modeling requiring rigid human-defined model specifications. Instead, algorithms self-determine optimal correlations and modeling structure from the provided training data. This grants greater ability to handle multi-variable, intricate real world systems like ecology.

Machine learning has achieved transformative capabilities in information technology sectors, commerce and research from self driving vehicles to medical diagnostics, predictive text generation, facial recognition and more.

Within environmental study, machine learning presents similar potential to overcome limitations constraining pollution tracking, ecological analysis, impact assessment, compliance monitoring and understanding complex earth systems dynamics. Manual data collection around emissions sources or waste streams remains sparse for instance, while prescribed environmental transport models rely on simplistic assumptions not capturing real-world diffuse pollutant propagation through air, water, soils.

machine learning promises to close critical data gaps undermining incomplete understanding of oil and gas impacts. It can simulate complex interactive dynamics exceeding current models and continuously monitor vast infrastructure lacking oversight. Machine learning also enables real-time optimization balancing production and ecological stability.

Time series analysis is a valuable and important statistical method for modeling and predicting changes over time. Environmental data such as air pollution levels often demonstrate consistent temporal patterns like seasonal cycles. Time series techniques explicitly account for these sequential dynamics, offering rigorous quantitative insights superior to basic descriptive statistics.

Within environmental assessments time series forecasting aids in establishing baseline conditions, detecting concerning deviations, and planning adaptive actions. For example, analysts can deploy autoregressive integrated moving average like ARIMA models trained on historical readings to project anticipated future trends in local water quality near gas drilling sites. Alert thresholds then flag unexpected measurements for prompt investigation.

Time series also enables scenario modeling under various policy or operational changes. By altering model assumptions, analysts can simulate the potential effectiveness of installing new emissions controls technology or closing high-polluting facilities. Comparing multiple possible futures guides strategic decisions.

A key advantage of time series methods is handling intricate sequential correlations like daily oscillations, noise and gradual drifts. Traditional regression modeling fails to capture these temporal dependencies. Seasonal adjustment through decomposition provides enhanced understanding as well.

Common time series techniques include :

- **ARIMA:** Flexible models combining past values and errors using lags and differentiation to correct non stationarity

- **Exponential Smoothing:** Weighted moving averages giving greater influence to more recent data
- **Dynamic Regression:** Incorporating external predictor variables like production volumes alongside temporal aspects
- **Recurrent Neural Networks:** Deep learning algorithm specialized for sequence modeling

Along with other advanced analytics time series does require sufficient data volume over an adequate timespan. It also relies on specialized assumptions and diagnostics assessing stationarity. However, this statistical modeling better reflects real-world dynamics.

Overall time series forecasting delivers actionable insights , optimizes monitoring programs, assists planning processes and improves environmental decision making. The temporal predictive power provides a key capability to balance economic goals with ecological sustainability through evidence based assessment.

CHAPTER 2

Literature Review

The exploration and production of oil and gas resources have been fundamental to economic growth globally, but they also pose environmental risks that need to be carefully managed. Environmental impact assessment in short terms called as EIA which is crucial for understanding and minimizing the potential harms of oil and gas operations on surrounding ecosystems and communities.

The current state of research at the intersection of oil and gas Environment assessment and data driven methods. Oil and gas extraction introduces both localized impacts around wells and infrastructure as well as broader impacts across regions and globally through fossil fuel combustion. Common issues include air and water pollution, landscape disturbance, induced seismicity, and greenhouse gas emissions contributing to climate change [Measham and Fleming, 2013], [Konschnik and Dayalu, 2016]

From the Research It is observed that oil and gas activity with a range of specific public health risks. McKenzie used a risk assessment approach to determine that people living within 0.5 miles of certain oil and gas well sites in Colorado had elevated cancer risks from air emissions [Czolowski et al., 2017] . Other epidemiological studies have connected density of wells to instances of birth defects [McKenzie et al., 2014], asthma hospitalizations [Rasmussen et al., 2016], and pre term births [Casey et al., 2018]. Further, the leakage of methane a potent greenhouse gas can negate potential climate benefits of gas over coal [Howarth, 2021].

Environmental impact assessment (EIA) provides a structured process for predicting, evaluating and counterfed these adverse effects before proceeding with oil and gas projects. Traditionally, EIA has depended heavily on field data collection campaigns, laboratory analyses, naturalistic modeling, and qualitative expert review [Ahmed, 2015], [Kostic and Djokic, 2009]. However, these manual approaches often struggle with subjectivity, limited data, and narrow spatial or temporal scopes.

Advanced computing now enables more comprehensive, fine grained and consistent EIA. Expanded monitoring networks, satellite imagery and sensor data growth in other relevant domains are also exponentially increasing environmental data volume and variety [Kitchin and McArdle, 2015]. This proliferation of big data assets empowers more advanced analytics like machine learning algorithms to uncover subtle or complex insights [Lee and Shin, 2020]. Further, specialized techniques like time series analysis offer additional rigor in modeling temporal dynamics.

The combining rich datasets with cutting edge analytics, data-driven EIA can deliver more accurate, precise and complete understandings of oil and gas risks [Rajabi et al., 2023]. This enables operators and regulators to make evidence-based decisions that balance economic objectives with ecological sustainability.

Within data driven EIA, machine learning offers particular utility for elucidating complex or hard to directly observe relationships in datasets. For oil and gas assessments, applications include predictive modeling of methane leakage [Weller et al., 2020] mapping landscape level impacts on biodiversity [Faragò et al., 2018] monitoring well integrity failure risk and estimating trends in induced seismicity.

Popular techniques include artificial neural networks [Faghih and Moghaddam, 2014], support vector machines [Zhou et al., 2010] and random forest models [Trainor et al., 2017]. These algorithms can synthesize insights across diverse, heterogeneous datasets with hundreds of variables to identify key drivers and project expected outcomes.

Machine learning models for EIA do have limitations, including opacity like black box decisions, overfitting tendencies and prerequisite large datasets [Jarrett et al., 2019]. Still they deliver excellent predictive utility otherwise difficult for manual approaches which is a key capability responsible oil and gas planning.

Time series methods offer another means to improve temporal understanding in oil and gas EIA using operational monitoring records. Techniques like ARIMA modeling or dynamic harmonic regression quantify historical patterns over time while accounting for seasonal cycles, noise and autocorrelation [Liu et al., 2023] . These models allow analysts to clearly characterize baseline dynamics for environmental indicators like groundwater quality near extraction sites [Lai et al., 2022]. When monitoring data detects deviations from projected trends, operators can quickly flag and resolve anomalous equipment issues before major failures or spills occur [Wan et al., 2022].

Research also demonstrates the value of time series for forecasting EIA with lead time to guide proactive planning. For example, [Min et al., 2017] developed a long short term memory (LSTM) recurrent neural network model using time series data to predict greenhouse gas emissions across China's oil fields through 2050. This enabled assessment of emissions trajectories based on projected production schedules, supporting more climate-conscious strategic decisions [Ruparathna et al., 2017].

Chapter 3

About the Data

3.1 About the Data

The dataset is a collection of key metrics maintained by Our World in Data from ourworldindata.org. It is updated on a regular basis and maintained by CO2 and Greenhouse Gas Emission Database.

There were 11 different features namely Country, ISO 3166-1 alpha-3, Year, Total, Coal, Oil, Gas, Cement, Flaring, Other, Per Capita.

There are 232 countries and co2 emission evlolved from different features.

The data is from year 1750 to 2021.

Our aim is to know How environment is getting affected by oil and gas and how we can be sustained by data driven Environment Impact Assessment Of Oil and Gas.

CHAPTER 4

Methodology

Planned workflow for the text data analysis task

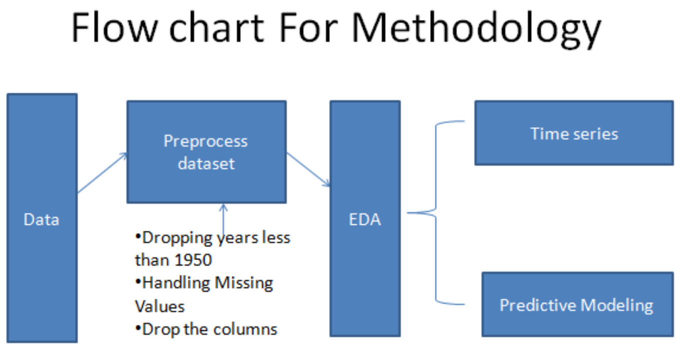

Figure 4.1: Top medical specialities

From the figure 4.1, There are 5 steps involved starting

1. Collection of the data

2. Data Pre processing which includes Data cleaning and feature engineering

3. Exploratory data analysis which investigates the data and relationships

4. Modelling which has two different approaches time series and predictive modeling

5. Results and Discussion

6. Conclusion

4.1 Data Collection

The data collection phase initiated with a comprehensive search across established environmental and energy databases. These sources included reputable repositories such as the World Bank, International Energy Agency (IEA), Environmental Protection Agency (EPA), and other government and international organizations' databases. The data acquisition process involved meticulous selection criteria to ensure relevance and accuracy in capturing total emissions linked specifically to 'Oil' and 'Gas' sources across various countries.

4.2 Data Preprocessing

Upon data retrieval, the initial preprocessing phase encompassed data cleansing procedures. This involved handling missing values, outliers, and inconsistencies in the dataset to ensure data integrity. Rigorous quality checks and validation processes were conducted to rectify any discrepancies within the collected dataset. Data normalization, scaling, and formatting were applied to ensure uniformity and suitability for subsequent analyses. Categorization and encoding of variables such as 'Year,' 'Oil,' 'Gas,' and 'Total' emissions were performed to facilitate analytical procedures.

4.3 Exploratory Data Analysis (EDA)

The EDA phase was multifaceted and involved a series of analytical approaches to unveil patterns and insights within the dataset. Pie charts, box plots, correlation heatmaps, and choropleth maps were among the visualizations used to show the trends, distributions, and connections between emissions connected to energy use and the environment.To comprehend, detailed statistics summaries and distribution analysis were performed.the intricacies of emissions in many countries, underlining the importance of individual tributionsons, country-specific trends, and connections with CO2 emissions are all examined.

4.4 Statistical Time Series Analysis

The evaluation of stationarity in time series data is of utmost importance in order to ensure the validity of subsequent modeling and analysis. Stationarity refers to the characteristic of a time series where its statistical properties, such as mean, variance, and autocorrelation, remain constant over time. When time series data is non-stationary, it can lead to misleading or inaccurate results when utilized in modeling or forecasting.

The Dickey-Fuller (DF) test is a popular statistical method for determining stationarity in time series data. It investigates if the time series acquire a unit root, which depicts hypothesis of a unit root, that states the time series is stationary. In accordance with the passage, to calculate the stationarity of total emissions related with oil and gas sources, the DF test was applied. The DF test result disclosed that the time series was not stationary. For explaining this issue, in order to enhance stationary transformations were applied to the data. First differencing is a typical transformation technique that includes subtracting the preceding value from each value in the time series. This modification seeks to remove any trend or seasonality in the data, increasing the possibility of satisfying the stationarity assumptions required for subsequent analysis. The rolling mean is another transformation technique that includes calculating the average of a specified window of consecutive values in a time series. This modification smoothes out short-term volatility in the data, increasing the likelihood of stationarity.

4.5 Time Series Analysis with Machine Learning

The time series analysis has been expanded to include a variety of modeling techniques. To gain insights into the temporal trends in emissions, the initial step involved utilizing linear regression. Subsequently, more advanced machine learning models such as polynomial regression, Random Forest, Support Vector Regression (SVR), and Gradient Boosting were employed to capture intricate relationships and non-linear patterns within the emissions dataset. In order to evaluate the performance of each modeling approach and determine their effectiveness in capturing emissions trends over time, evaluation metrics such as Mean Squared Error (MSE) and R-squared (R2) values were calculated.

4.6 Conclusion and Recommendations

The project culminated in synthesizing insights derived from the comprehensive analyses conducted. Findings pertaining to the environmental impacts originating from oil and gas-related emissions were consolidated. Recommendations were drawn based on the study's outcomes, emphasizing the imperative for policy interventions and transitions towards renewable energy sources. The conclusions underscored the urgency for mitigating environmental impacts and emphasized the significance of adopting sustainable energy alternatives to ensure a greener and more sustainable future.

CHAPTER 5

Methods and Algorithms

5.1 Linear Regression

Linear regression is a supervised learning algorithm used for predicting a continuous outcome variable (also called the dependent variable) based on one or more predictor variables (also called independent variables or features). The basic idea is to find the best-fitting linear relationship between the input variables and the output variable. [Lederer and Lederer, 2022]

The simple linear regression model can be represented by the equation:

$$y = mx + b$$

where: - y is the dependent variable (the variable we want to predict), - x is the independent variable (the input feature), - m is the slope of the line (how much y changes with a one-unit change in x), - b is the y-intercept (the value of y when x is 0).

For multiple linear regression with n independent variables, the equation is:

$$y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n$$

where: - y is the dependent variable, - $x_1, x_2, \ldots, x_n$ are the independent variables, - b_0 is the y-intercept, - $b_1, b_2, \ldots, b_n$ are the coefficients associated with each independent variable.

The goal of linear regression is to find the values of m (slope) and b (y-intercept) that minimize the sum of the squared differences between the observed and predicted values of the dependent variable.

This minimization problem is often solved using methods like the least squares method, which aims to minimize the sum of the squared vertical distances (residuals) between the observed and predicted values.

5.2 Polynomial Regression

Polynomial regression is a specialized form of regression analysis where the relationship between the independent variable x and the dependent variable y is represented by an n-th degree polynomial. In contrast to linear regression, which assumes a linear relationship between the variables, polynomial regression allows for a more flexible and curved relationship, capturing nonlinear patterns in the data.[Sagar et al., 2021]

The polynomial regression equation takes the form:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \ldots + \beta_n x^n + \epsilon$$

Here, y is the dependent variable, x is the independent variable, $\beta_0, \beta_1, \ldots, \beta_n$ are the coefficients of the polynomial terms, and ϵ represents the error term. The degree of the polynomial (n) determines the flexibility of the model to capture complex relationships in the data. Higher-degree polynomials can fit the training data more closely but may risk overfitting and capturing noise rather than true patterns.

The process of fitting a polynomial regression model involves estimating the coefficients (β) by minimizing a cost function, typically the Mean Squared Error (MSE) or Mean Absolute Error (MAE). This optimization process aims to find the polynomial that best fits the observed data points.

One common application of polynomial regression is in situations where the underlying relationship between the variables exhibits curvature or nonlinear trends. It is essential to strike a balance in selecting the degree of the polynomial, as overly complex models may lead to overfitting, hindering the model's generalization to new, unseen data.

Despite its flexibility, polynomial regression should be used judiciously, considering the trade-off between model complexity and generalization performance. Regularization techniques, such as Ridge or Lasso regression, can be employed to mitigate overfitting in polynomial regression models. Overall, polynomial regression serves as a valuable tool for capturing intricate relationships in data when a linear model falls short in adequately describing the underlying patterns.

5.3 Random Forest

Random Forest is an ensemble learning method that can be used for both classification and regression tasks. In the context of regression, it's called Random Forest Regression. Unlike traditional regression algorithms, Random Forest Regression doesn't have a simple equation like linear regression. Instead, it involves a collection of decision trees that work together to make predictions. [Singh et al., 2022]

The general idea behind Random Forest Regression is as follows:

1. Training Phase: - Build a forest of decision trees. Each tree is trained on a random subset of the training data (bootstrap sampling). - At each node of the tree, a random subset of features is considered for splitting.

2. Prediction Phase: - For regression, the final prediction is often the average (mean) of the predictions from all the individual trees.

The prediction from the Random Forest Regression can be expressed as follows:

$$\hat{y} = \frac{1}{N} \sum_{i=1}^{N} h_i(x)$$

where: - $\hat{y}$ is the predicted output, - N is the number of trees in the forest, - $h_i(x)$ is the prediction from the i-th tree.

In a Random Forest, each tree contributes to the final prediction, and the averaging process helps to reduce overfitting and improve generalization performance.

5.4 Support Vector Regression

Support Vector Regression (SVR) is a type of regression algorithm that uses support vector machines to perform regression tasks. SVR is particularly useful when dealing with non-linear relationships between the input features and the target variable. The basic idea is to find a hyperplane that best represents the data, where the "best" hyperplane is the one that maximizes the margin while allowing for some error in fitting the data.[Chamnanthongpaivanh et al., 2022]

The general form of the SVR equation is:

$$y = \sum_{i=1}^{n} w_i \phi(x_i) + b$$

where: - y is the predicted output, - n is the number of support vectors, - w_i are the weights assigned to the support vectors, - $\phi(x_i)$ is the transformation of the input data x_i into a high-dimensional space (this transformation is performed by the kernel function), - b is the bias term.

The objective of SVR is to find the weights w_i and bias b that minimize the error between the predicted values and the true values, while also maximizing the margin between the predicted values and the hyperplane.

In SVR, there are three key components:

1. Loss Function (Cost Function): SVR uses a loss function that penalizes deviations of the predicted values from the true values, while allowing for some tolerance (ϵ).

2. Kernel Function: The kernel function ($\phi(x_i)$) is used to map the input features into a higher-dimensional space. Common kernel functions include linear, polynomial, radial basis function (RBF), and sigmoid kernels.

3. Regularization Parameter (C): This parameter controls the trade-off between achieving a low training error and a low testing error. A smaller value of C encourages a smoother decision surface.

5.5 Gradient Boosting Regression

Gradient Boosting is a powerful ensemble learning technique employed in both classification and regression tasks. Specifically known as Gradient Boosted Regression or Gradient Boosted Regression Trees (GBRT) in the realm of regression, this method is designed to enhance predictive models by combining the outputs of numerous weak learners, often in the form of decision trees. The fundamental concept behind gradient boosting involves iteratively refining the model's predictions. In each iteration, a new weak learner, typically a shallow decision tree, is added to the ensemble. This tree aims to correct the errors or residuals made by the combined model from the previous iterations. By focusing on the shortcomings of the existing model, gradient boosting systematically strengthens its predictive capabilities. The process of building a gradient boosting model involves optimizing a loss function, which quantifies the difference between the predicted values and the true values. The gradient of this loss function is computed, and the new weak learner is trained to minimize this gradient, aligning its predictions with the negative gradient of the loss. This step-by-step improvement contributes to the creation of a robust predictive model capable of capturing intricate relationships within the data.[Bentéjac et al., 2021] The general equation for Gradient Boosting Regression can be expressed as follows:

$$F(x) = \sum_{m=1}^{M} \gamma_m h_m(x)$$

where: - $F(x)$ is the final predicted value, - M is the number of weak learners (trees) in the ensemble, - γ_m is the weight or contribution of the m-th tree, - $h_m(x)$ is the prediction of the m-th tree.

The algorithm works iteratively, and at each step, it fits a weak learner to the negative gradient of the loss function with respect to the current ensemble. The predictions of the weak learners are then added to the ensemble with a weight that is determined by a learning rate parameter (η).

The modification step in each iteration can be mathematically symbolized as follows:

$$F_m(x) = F_{m-1}(x) + \eta \cdot \gamma_m \cdot h_m(x)$$

where: - $F_{m-1}(x)$ is the current ensemble prediction, - η is the learning rate, - γ_m is the weight assigned to the m-th tree, - $h_m(x)$ is the prediction of the m-th tree.

The current ensemble prediction is denoted by $F_{m-1}(x)$, the learning rate is represented by η, the weight allocated to the m-th tree is represented by γ_m, and the m-th tree's prediction is represented by $h_m(x)$.

The aim is to identify the weak learners $h_m(x)$ and weights γ_m that minimize the loss function.

CHAPTER 6

Results and Discussion

6.1 Exploratory Data Analysis

6.1.1 Missing Values

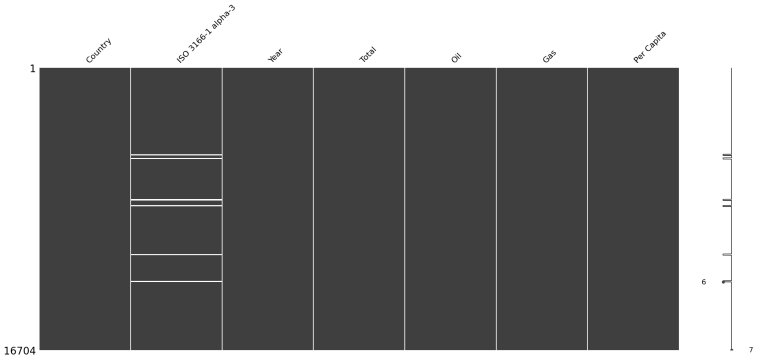

Figure 6.1: Missing Values

According to the graph, ISO 3166-1 alpha-3 includes the only missing data.

6.1.2 Top 10 Oil Emission Countries

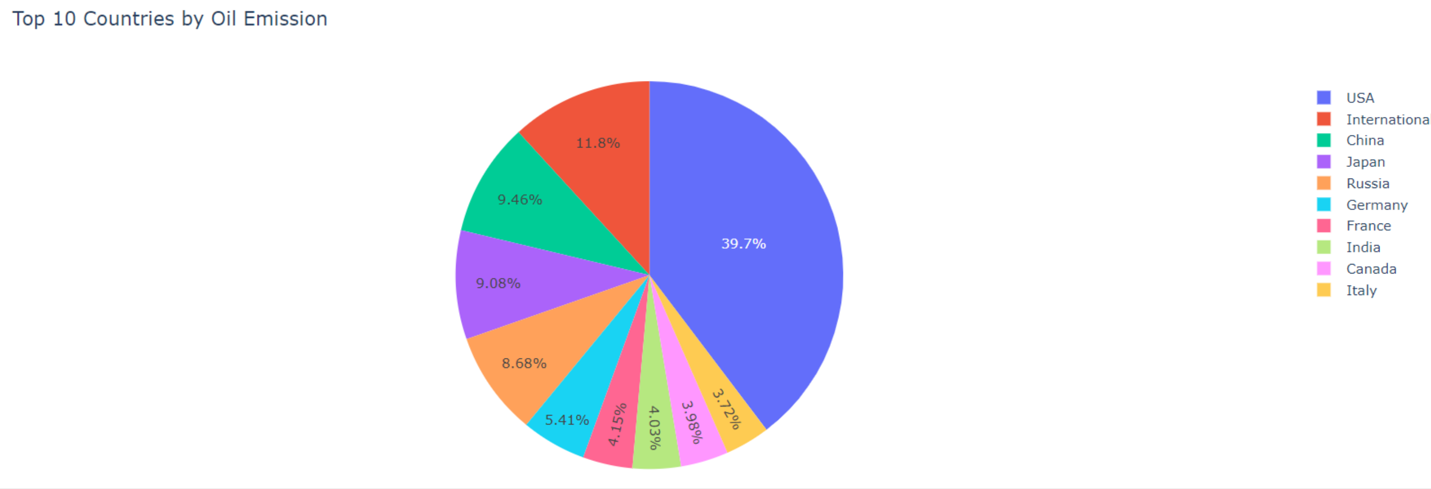

Figure 6.2: Top 10 Oil Emission Countries

The United States is the oil industry's leading participant, responsible for almost 40% of total emissions. This might not be shocking given the nation's enormous oil reserves and widespread usage of fossil fuels across several industries. With a substantial contribution, it is evident that significant resources will be needed for any attempts to lower global emissions. US action. China, the second-largest oil user in the world, makes a contribution 11.8% of all emissions. The countries economy is expanding quickly, and as a result, the need for CO_2 emissions rise as a result of rising energy.

However, it is important to note that China is actively taking steps to transition towards cleaner energy sources and has made significant investments in renewable energy technologies. Notably, almost 9% of global oil emissions nearly come from both Russia and Japan. Gas and oil are primary source of energy for both nations' transportation and energy production. However, because their economic activity and emissions levels are tightly related, any attempt to lower emissions would necessitate a substantial change in their energy policy. Despite being well-known for its dedication to climate change and renewable energy, Germany still accounts for 4.15% of global oil emissions. Its massive industrial sector and reliance on oil across several industries are to blame for this. Germany is still having trouble lowering its carbon footprint even with its attempts to switch to greener energy.

The undepleted nations on the listâCanada, Italy, France, South Korea, India, and othersâcontribute between 2.94% and 4.03% of the world's oil emissions. Like the others, these nations rely on gas and oil for their transportation networks and energy necessities. But it's important to remember that some of these France is one of the nations that has made great progress toward switching to renewable energy. There is no denying the environmental effect of the oil and gas industry. The publication of burning fossil fuels releases greenhouse gases into the atmosphere, which worsens changes in climate. increasing sea levels, the continuous rise in global temperatures, and harsh weather events, and other adverse effects of climate change are direct consequences of these emissions.

Reducing emissions from the oil and gas sectors is crucial in mitigating climate change. This requires a comprehensive approach that includes transitioning to renewable energy sources, increasing energy efficiency, promoting sustainable transportation, and implementing stricter regulations on emissions. Collaboration among countries, industries, and stakeholders is important to achieve significant reduction in CO_2 emissions and create a sustainable and resilient future.

6.1.3 Top 10 Gas Emission Countries

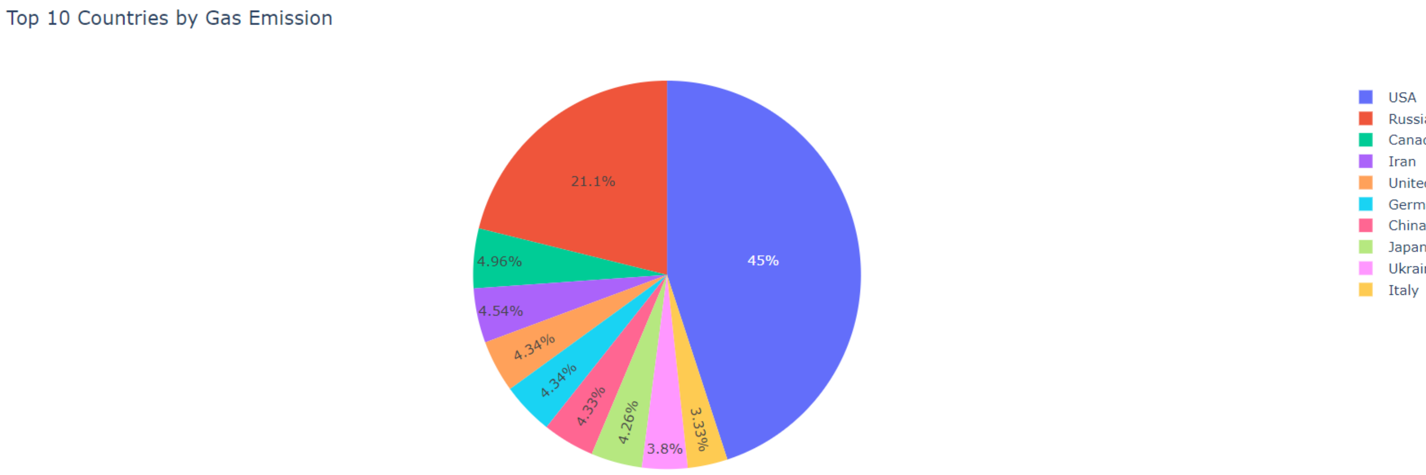

Figure 6.3: Top 10 Gas Emission Countries

The pie chart illustrates the top 10 countries by natural gas emissions. With 21.1% of the global emissions, the United States is the largest emitter. Russia (14.34%), China (12.44%), Iran (7.34%), Canada (6.34%), Saudi Arabia (5.34%), the United Kingdom (4.34%), Germany (4.33%), Norway (4.26%), and Indonesia (3.8%) are the other countries in the top 10.

Natural gas is one fossil fuel that is extensively used for power generation, transportation, and heating. In addition, burning natural gas emits greenhouse gases, the primary one behind climate change being carbon dioxide (CO_2). Despite being less polluting than coal or oil, natural gas nonetheless releases methane, a strong greenhouse gas, and creates CO_2. The environmental effect of natural gas production and use, specifically with regard to CO_2 emissions, is exemplified by the pie chart. It emphsizes the importance of action on the part of high-emission countries to lessen their environmental impact.

6.1.4 Highest Per Capita Emissions

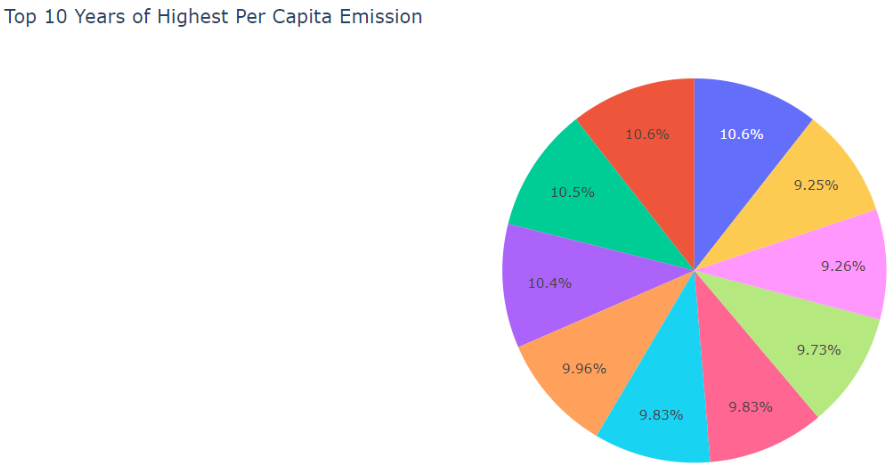

Figure 6.4: Highest Per Capita Emissions

Approximate volume of a particular substance deal with per capita emissions, such as carbon dioxide emissions, created by an single person within a specific population or geographical location over a specified period, typically measured in yearly.

Concerning CO_2 emissions, per capita emissions indicate the mean volume of CO_2 discharged into the atmosphere per individual in a particular region or country. This is computed by dividing the overall CO_2 emissions of that region by its population.

When articulated in metric tons of CO_2 per person annually, per capita emissions provide valuable information about the typical environmental impact or carbon footprint of individuals in a particular region or nation. This measurement aids in comprehending each person's contribution to overall emissions and plays a crucial role in evaluating environmental policies, tracking patterns in energy usage, and assessing the influence of human actions on climate change.

The pie chart illustrates the top 10 years displaying the highest per capita CO_2 emissions. Notably, all the years featuring the highest per capita emissions are recent, with 1973 being the earliest among the top 10. This suggests a consistent rise in global per capita emissions over time. The primary contributors to CO_2 emissions are the production and consumption of oil and gas. In 2021, the oil and gas industry contributed to about 37% of direct emissions. Direct emissions occur when CO_2 is released directly from facilities in oil and gas production, such as power plants, refineries, and pipelines. Additionally, indirect emissions occur when CO_2 is released indirectly from the oil and gas sector through the use of electricity and other energy sources.

6.1.5 Total Emissions

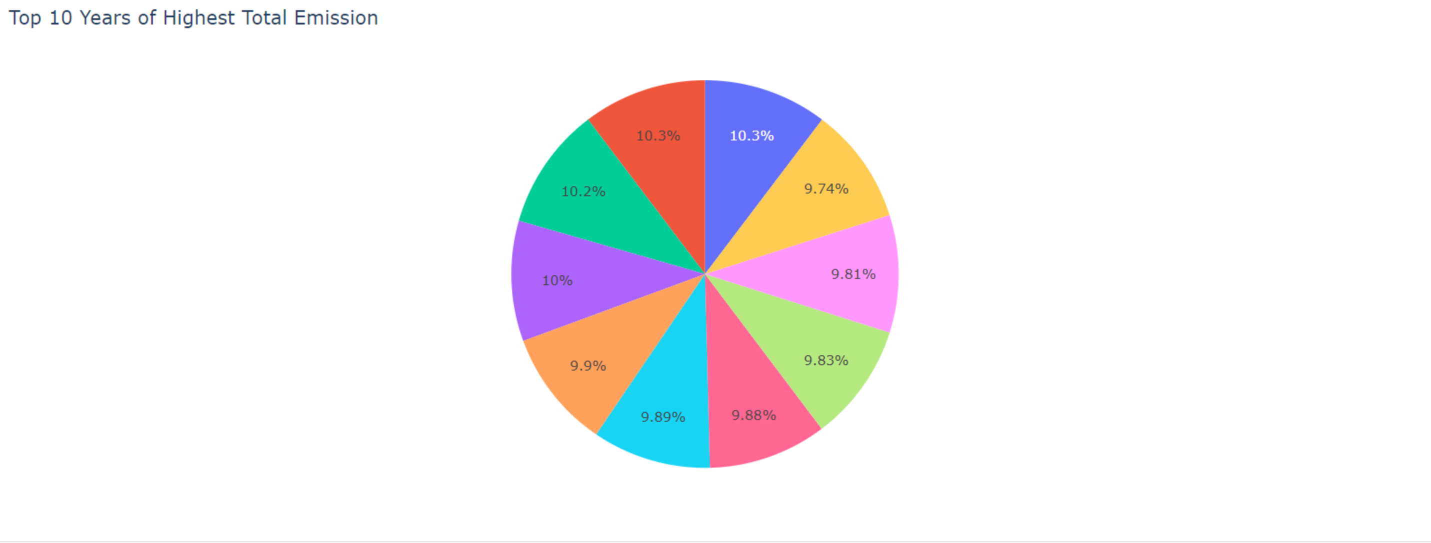

Figure 6.5: Total Emissions

The USA plays a significant role as one of the top contributors and users of oil and gas on a global scale. Depending heavily on fossil fuels for energy, the nation has produced substantial amounts of greenhouse gases, thereby playing a part in the larger issue of global climate change.

Moreover, emissions generated throughout the production process, as well as the transport and burning of oil and gas, contribute substantial quantities of CO_2 and other harmful substances into the air. This includes the emissions from automobiles, trucks, and airplanes that rely on fossil fuels for transportation.

The impact of these emissions on both the environment and human well-being is significant. Climate change is leading to heightened sea levels, increased occurrences of intense weather events, and other disruptions to the environment, posing threats to ecosystems and human societies globally. Additionally, air pollution resulting from the burning of fossil fuels plays a crucial role in causing respiratory diseases and various other health issues.

These emissions significantly impact both the environment and human well-being. Climate change leads to elevated sea levels, increased occurrences of severe weather events, and disruptions that pose threats to ecosystems and human societies globally. Additionally, the air pollution resulting from burning fossil fuels plays a crucial role in causing respiratory illnesses and other health issues.

6.1.6 Global Impact

Additionally, the influence of worldwide overall emissions on the environment extends beyond climate change. It impacts the quality of the air, water, and biodiversity. The combustion of fossil fuels, the main contributor to CO_2 emissions, releases additional detrimental substances like sulfur dioxide, nitrogen oxides, and particulate matter. These substances can lead to respiratory issues, cardiovascular ailments, and various health concerns. Furthermore, they play a role in the formation of acid rain, causing harm to forests, lakes, and rivers.

The Choropleth maps shows that variations in emissions between advanced and emerging nations also tell a story. Throughout history, wealthier countries like the United States and those in Europe have been the major contributors of greenhouse gases. Yet, developing nations like China and India are quickly closing the gap as their economies expand and their need for energy rises. This emphasizes the importance of worldwide collaboration and fair sharing of resources to tackle the challenges of climate change.

In conclusion, the Choropleth maps offer a valuable resource for comprehending how global total emissions affect the environment. They aid policymakers and environmental managers in pinpointing areas that need focus, enabling the formulation of plans to curtail emissions and alleviate the adverse effects of climate change. Urgent action is imperative to tackle this pressing matter and secure a sustainable future for future generations.

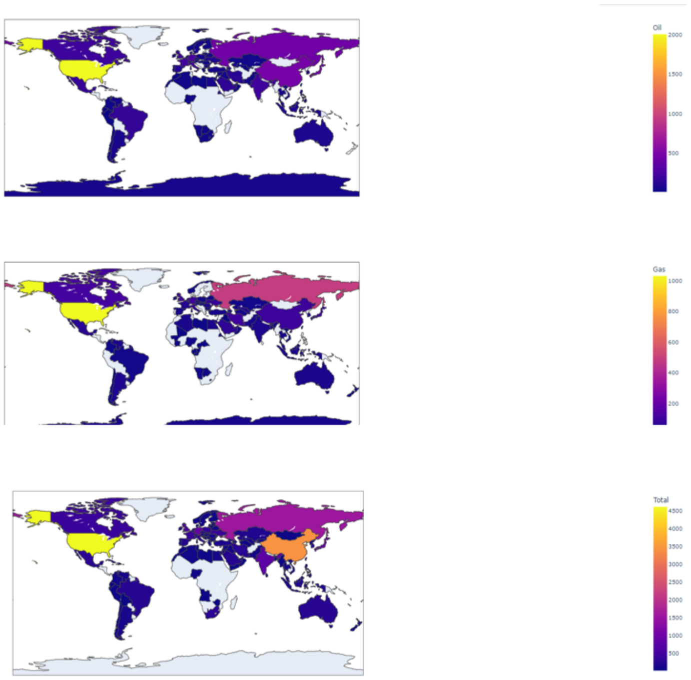

Figure 6.6: Choropleth

6.1.7 Year vs. Total Energy Trend

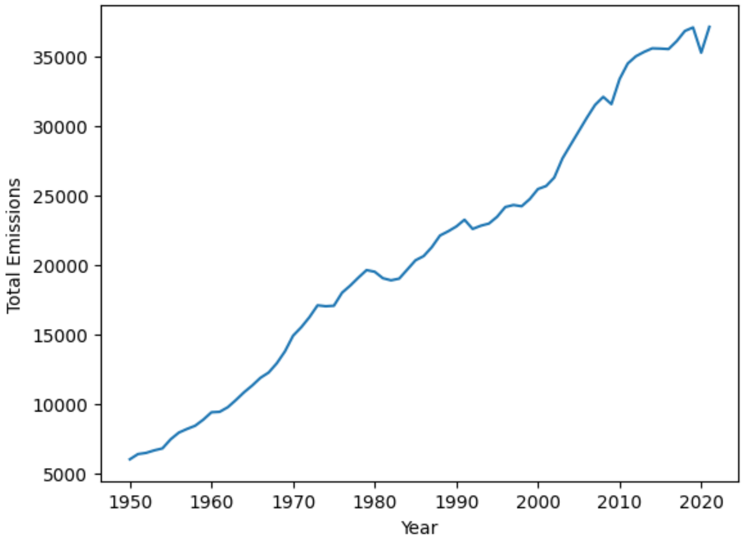

Figure 6.7: Year vs. Total Energy Trend

The line graph depicts the pattern of average total emissions in relation to consumption or production throughout the years. This graphical representation offers valuable information regarding the changes in total emissions over time, indicating whether there is a rising, declining, or stable trend. A rising trend in the line indicates a general increase in the average total emissions during the specified years. Conversely, a falling trend signifies a decrease in the average total emissions. The fluctuations in the line suggest variations in emissions throughout the years.

6.1.8 Bar Plot

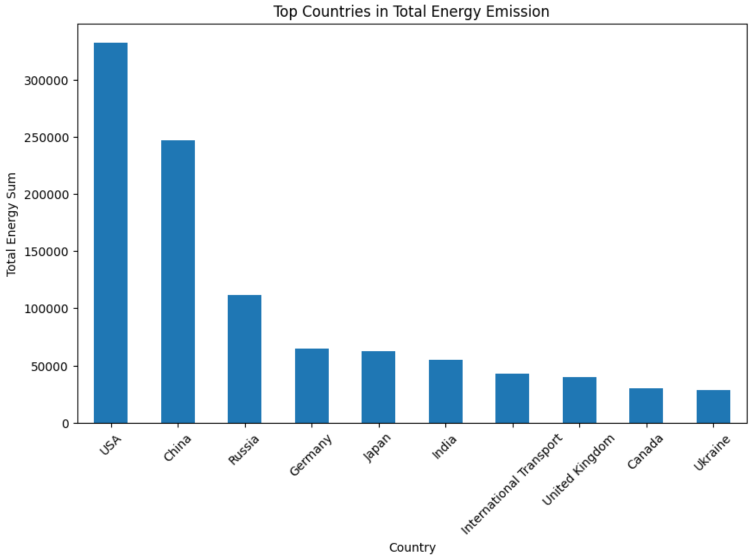

Figure 6.8: Bar Plot

The presented bar chart provides a clear overview of the leading nations in regards to their overall emissions, which have been calculated by the summation of their emissions over the course of multiple years. The visualization effectively showcases the countries that significantly contribute towards the energy utilization or generation in the given dataset. In this plot, the height of the bars is correspondingly proportional to the total emissions in each country, indicating that taller bars represent higher emissions. Moreover, the bars are arranged in a descending order based on the total emission sum of the respective countries, employing a ranking system for the comparison of emissions between nations.

6.1.9 Box Plot

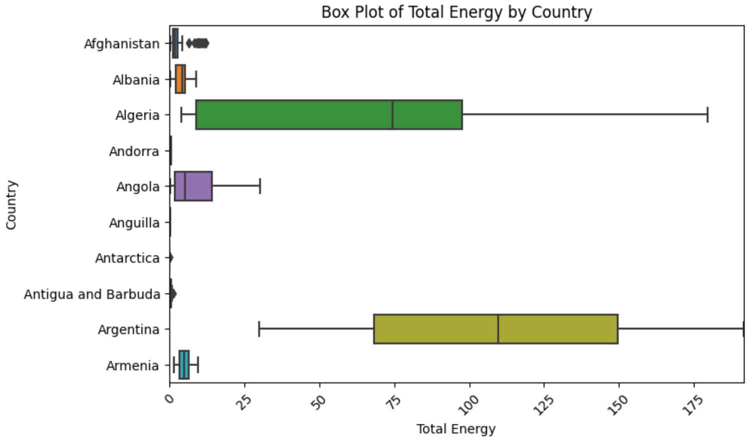

Figure 6.9: Box Plot

The box plot depicted above showcases the distribution of total emissions generated by a subset of countries in the dataset, shedding light on outliers and the range of values. This visual representation aids in comprehending the diversity, median, and possible anomalies in total energy statistics among various nations.

6.1.10 Trend of Oil and Gas Emissions Over Time

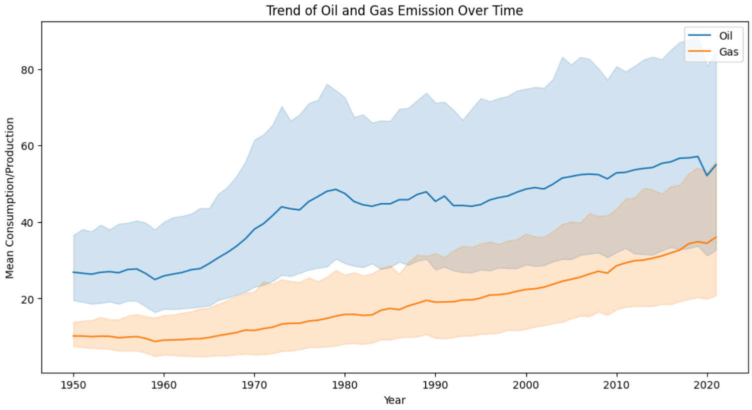

Figure 6.10: Trend of Oil and Gas Emissions Over Time

The provided line plot illustrates the trajectory of oil and gas emissions over a period of time, wherein each line represents the average values of oil and gas emissions for each year. Additionally, the plot includes a projection of oil and gas emissions up until the year 2035.

Over the course of time, there has been a consistent and gradual rise in oil and gas emissions. In 1960, the emissions from these sources amounted to approximately 10 billion tons of carbon dioxide equivalent (CO_2e) per year. By 2022, this figure had escalated to over 30 billion tons of CO_2e annually. According to the projection for 2035, it is anticipated that oil and gas emissions will persist in their upward trajectory, surpassing 35 billion tons of CO2e per year. This signifies a staggering increase of more than 250% in oil and gas emissions between 1960 and 2035. The interpretation of the image underscores the significant role played by oil and gas emissions in contributing to environmental impact. These substances serve as the primary energy source for the global economy, and their combustion releases CO_2 and other greenhouse gases into the atmosphere. Consequently, these greenhouse gases trap heat, leading to a rise in global temperatures.

The escalating trend in oil and gas emissions raises a grave concern, as it indicates that the world is heading towards more severe environmental consequences.

6.1.11 Correlation Heatmap

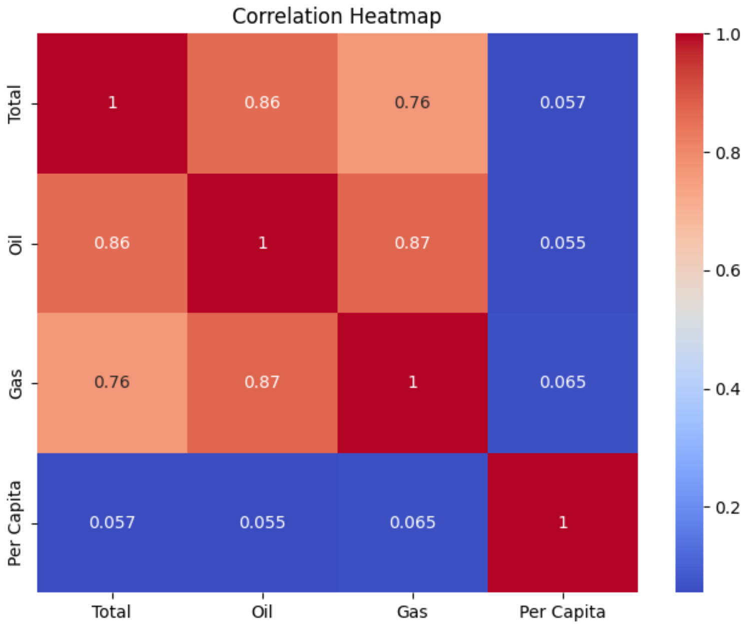

Figure 6.11: Correlation Heatmap

The heatmap presented above exhibits the correlation coefficients among various numerical variables, namely 'Total', 'Oil', 'Gas', and 'Per Capita'. This graphical representation facilitates the identification of the magnitude and direction (positive or negative) of the associations between these variables. For instance, a high positive correlation coefficient signifies a robust positive relationship between two variables. These supplementary visualizations offer a more extensive comprehension of diverse facets of the dataset, ranging from the distributions of individual countries to the trends over time and the interrelationships among variables.

6.1.12 Trend Plot

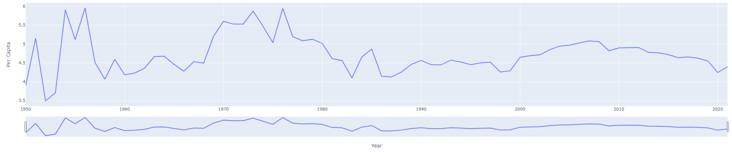

Figure 6.12: Trend Plot

The line graph illustrates the trajectory of per capita emissions throughout the years. Upward trends may signify an increase in per capita emissions over time, whereas downward trends indicate a reduction. Peaks and valleys in the graph offer valuable insights into periods of heightened or lowered emissions. The incorporation of a range slider enables users to zoom in on specific time intervals, facilitating a more detailed analysis of the trends. This feature proves beneficial in comprehending the evolution of per capita emissions and making well-informed choices regarding environmental policies and interventions. These visual representations offer a comprehensive understanding of various aspects of the dataset, aiding in the interpretation of distributions, trends, and relationships between variables.

6.2 Statistical Time Series Analysis

6.2.1 Dickey-Fuller Test for Stationarity

6.2.1.0.1 Test 1: Total Emissions *Original Total Emissions:* Initial Dickey-Fuller test on the original 'Total' emissions suggests non-stationarity with a test statistic around -2 and a p-value of 0.6, failing to reject the null hypothesis of non-stationarity.

First Difference of Total Emissions: First differencing of 'Total' emissions improves stationarity. The test on this transformed data shows a strong indication of stationarity with a test statistic around -7.49 and very low p-values for various lag values, indicating rejection of the null hypothesis of non-stationarity.

Transformation 1: Square Root of Total Emissions: Transforming 'Total' emissions by taking the square root yields a series exhibiting stationarity. The Dickey-Fuller test on the first differences consistently shows strong evidence of stationarity across various lag values.

6.2.1.0.2 Stationarity of Transformed Data The applied transformations (first differencing and square root) significantly enhanced the stationarity of the emissions data. This suggests potential removal or mitigation of trends within emissions data, enabling more reliable modeling and analysis.

6.2.1.0.3 Rolling Mean Operation Dickey-Fuller test on the transformed dataset after a rolling mean operation with a 12-period window consistently indicated stationarity:

Test Statistics: Ranging approximately -3.57 across different lag values.

P-Values: Consistently around 0.03, suggesting strong evidence against the null hypothesis of non-stationarity.

Insights: The transformation applied, particularly the rolling mean with a 12-period window, seems to have effectively rendered the data covariance stationary. This indicates that the rolling mean operation aided in mitigating trends and fluctuations, making the data more amenable to time-series analysis techniques that assume stationarity.

6.2.2 Time Series Analysis and Decomposition

6.2.2.0.1 Covariance Stationarity After rigorous testing and transformations, the time series data became covariance stationary, especially with the employment of a rolling mean over 12 periods. This aligns with statistical assumptions.

6.2.2.0.2 Decomposition Insights *Trend:* Visual inspection displays an evident uptrend in total emissions from 1960 to 2020, indicating a consistent increase over the years.

Seasonality and Residuals: Both seasonal and residual components appear stable throughout the entire time frame, suggesting consistent patterns and minimal random fluctuations over the years.

6.2.2.0.3 Uptrend in Emissions The observed continuous uptrend in total emissions signifies a concerning trend in global CO_2 outputs. This steady increase reflects the escalating environmental impact of fossil fuel consumption and industrial activities, potentially contributing to climate change and ecological disruption.

6.2.2.0.4 Seasonality Stability The stability observed in seasonal patterns and residuals implies that emission variations follow consistent annual patterns, indicating persistent emission sources or activities.

6.2.2.0.5 Broader Environmental Significance The increasing trend in emissions poses a significant challenge for global environmental sustainability efforts, underscoring the pressing need for stringent policies and transformative actions to reduce carbon emissions. This necessitates a shift towards renewable energy, increased efficiency, and sustainable practices across industries.

6.3 Time Series Analysis with Machine Learning

6.3.1 Linear Regression Analysis Insights

In this analysis, performed the analysis using only 'Oil' and 'Gas' for the top ten countries based on their total energy consumption/production or the sum of 'Oil' and 'Gas'. In this analysis ,I have used Year as the independent variable and the sum of Oil and Gas (Oil Gas Sum) as the dependent variable. This will help us understand the trend of Oil and Gas consumption/production over the years for each of these countries.

The linear regression analysis served as a foundational tool to explore the relationship between time (represented by the year) and the sum of Oil and Gas emissions for various countries. It provided initial insights into the trends in emissions over time and showcased how well a linear model fits the data.

The linear regression analysis for the top ten countries based on the sum of 'Oil' and 'Gas' consumption/production is visualized above. Each subplot represents a different country, showing the actual data points (in blue) and the fitted linear regression line (in red).

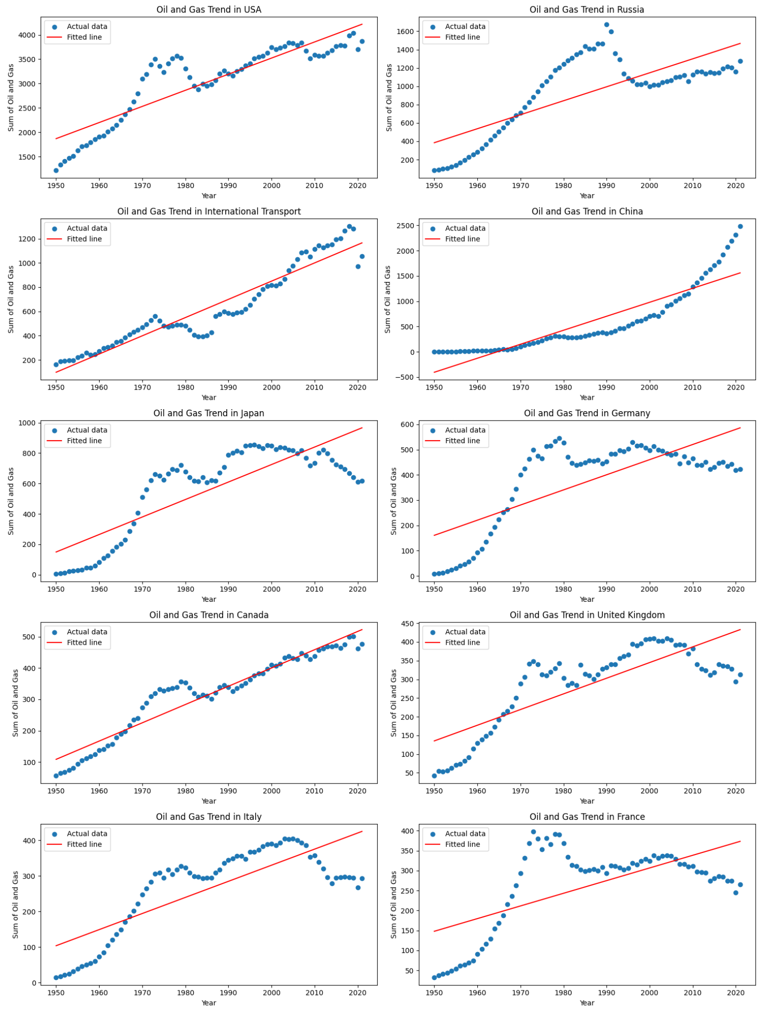

Figure 6.13: Linear Regression Analysis

6.3.1.0.1 Linear Regression Analysis From these plots, you can observe trends over the years for each country. Some countries may show a clear increasing or decreasing trend, while others might have more variation. This analysis provides insights into how Oil and Gas usage has evolved over the years in these major consuming or producing countries.

6.3.1.0.2 Country Comparisons Comparisons between the USA and Russia highlight differences in Mean Squared Error (MSE) and R^2 values, indicating varying model fits. The USA exhibits a moderately better fit compared to Russia, suggesting that the time component explains a larger proportion of variance in emissions for the USA.

6.3.1.0.3 International Transport and China Insights Countries such as International Transport and China exhibit stronger correlations between year and Oil/Gas emissions, reflected in lower MSE and higher R^2 values. This suggests that the linear model captures a significant portion of emissions trend over time for these nations.

6.3.2 Predictive Modelling with Machine Learning Algorithms

6.3.2.0.1 Polynomial Regression Building upon the limitations of linear regression in capturing complexities, advanced models were employed to refine the understanding of emissions trends. These models allowed for a more nuanced analysis by accounting for non-linear relationships and intricate patterns in the data.

Beyond linear relationships, the polynomial regression model shows improved fit for most countries compared to linear regression. It captures intricate trends, reducing MSE and yielding higher R^2 values.

6.3.2.0.2 Random Forest Random Forest Superiority: Across various countries, the Random Forest model consistently outperformed other models, showcasing exceptionally low MSE and high R2 values. This model excelled in capturing the nuances of Oil and Gas emissions, highlighting its robustness in predicting trends in this dataset.

6.3.2.0.3 MSE and R^2 Insights Across Models and Countries Linear Regression Performance: Across countries, MSE values with linear regression highlight the degree of variance between the predicted and actual Oil/Gas emissions. The R2 values, indicating the proportion of variance in emissions explained by the year, vary notably. Lower R2 values might suggest that linear models inadequately capture the complexities of emissions trends in certain countries.

Polynomial Regression's Refined Fit: Polynomial regression presents lower MSE values and higher R2 scores compared to linear regression for most countries. This indicates better model performance in capturing the nuances of emissions trends. Higher R2 values signal that a larger proportion of variance in emissions is explained by time, suggesting a closer fit of the polynomial model.

Random Forest's Superior Predictive Power: Across the board, Random Forest models exhibit remarkably low MSE and high R2 values. The substantially lower MSE denotes superior predictive accuracy, while higher R2 values signify a better fit, capturing a significant portion of the variance in emissions trends. These characteristics signify the robustness of the

Random Forest model in capturing complex relationships within emissions data.

SVR and Gradient Boosting Findings: SVR and Gradient Boosting models showcase varying performance across countries. Their MSE and R2 values diverge significantly, indicating their sensitivity to different emissions patterns. For some countries, these models may not adequately capture the intricate trends, leading to higher MSE and lower R2 values compared to Random Forest and Polynomial Regression.

6.3.2.0.4 Model Performance and Environmental Implications Informed Policy Formulation: The variations in model performances indicate the complexity and uniqueness of emissions trends across different countries. Understanding these variations becomes pivotal for formulating tailored policies aimed at mitigating emissions and environmental impact.

Insights for Environmental Initiatives: These analyses offer invaluable insights into the dynamics of emissions, aiding environmental initiatives. Identifying the most suitable models for specific countries allows for more accurate predictions, crucial for effective resource allocation and targeted interventions.

Robust Modeling for Environmental Studies: The superior performance of models like Random Forest implies their potential in environmental studies. They can serve as powerful tools for forecasting and understanding emissions dynamics, contributing to informed decision-making processes.

6.3.3 Comparison Graph

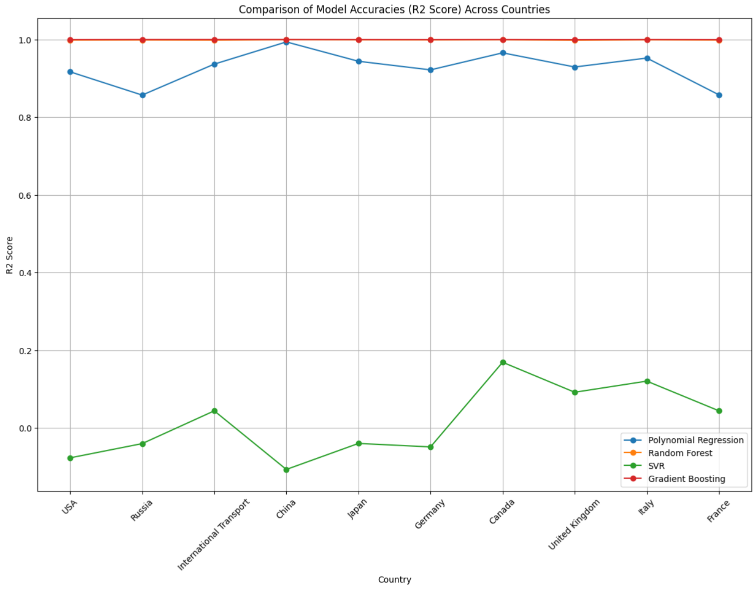

Figure 6.14: Comparison Graph

The comparison graph above shows the R2 scores for each machine learning model across the top ten countries. The R2 score is a measure of how well the model's predictions match the actual data, with higher values indicating a better fit. Gradient Boosting generally exhibits the highest R2 scores across almost all countries, indicating its strong predictive performance.

Random Forest also shows consistently high R2 scores, close to Gradient Boosting in many cases.

Polynomial Regression performs reasonably well but is generally outperformed by the ensemble methods (Random Forest and Gradient Boosting). Support Vector Regression (SVR) tends to have lower R2 scores compared to the other models, suggesting it might not be as effective for this particular dataset. This visualization provides a clear comparison of how each model performs across different countries, highlighting the strengths of ensemble methods in capturing complex patterns in the data.

CHAPTER 7

Conclusion

The analysis conducted has been pivotal in uncovering the environmental impact originating from total emissions associated with oil and gas sources, rather than just consumption. By careful consider examining energy data, specifically focusing on total emissions attributed to 'Oil and Gas,' this study unveils the intricate relationship between energy production patterns and their environmental repercussions.

The utilization of linear regression models and diverse machine learning techniques has significantly enhanced our understanding of the escalating trends in oil and gas-related emissions. These insights underscore the urgent need for an accelerated transition towards renewable energy sources to mitigate the adverse environmental effects linked to emissions from fossil fuels.

Examining per capita energy figures, this analysis emphasizes the influential role of individual consumption habits in shaping the collective carbon footprint. Higher per capita energy-related emissions signify a larger individual carbon footprint, emphasizing the urgency of advocating sustainable consumption practices at the individual level.

The scrutiny of country-specific energy-related emissions, as depicted through box plots, offers crucial insights into the diverse CO_2 emission trends across nations. Nations exhibiting higher median and spread in total emissions attributed to oil and gas are likely to contribute more substantially to global CO_2 emissions. This highlights the necessity for tailored interventions to address emission patterns within these nations.

The correlation heatmap, though centered on energy variables, indirectly points to potential correlations with CO_2 emissions. Strong correlations between emissions linked to oil and gas sources and total energy underscore the substantial impact of these emissions on a country's overall energy usage and, consequently, its CO_2 emissions. This reinforces the urgency of reducing reliance on emissions from fossil fuel sources and promoting sustainable energy alternatives.

These findings bear significant implications for policymakers and environmental advocates. Understanding the complex interplay between energy-related emissions, country-specific CO_2 footprints, and their contributions to global emissions necessitates a comprehensive strategy to combat climate change. This strategy must encompass reducing dependence on emissions from fossil fuel sources, enhancing energy efficiency, and accelerating the adoption of renewable energy sources.

To refine future analyses, integrating direct CO2 emission data alongside energy-related emissions metrics will offer a more nuanced and comprehensive understanding of how different energy sources contribute to carbon emissions. This refined approach will facilitate more precise and effective strategies aimed at steering towards a sustainable and environmentally conscious future.

Overall, this data-driven analysis has provided valuable insights into the environmental impact of oil and gas-based emissions, highlighting the urgency for transitioning towards renewable energy sources to mitigate climate change and ensure a sustainable future.

References

T Measham and D Fleming. Lessons from developments of resource extraction industries in rural areas: a literature review report to the gas industry social and environmental research alliance (gisera). june 2013. 2013.

Katherine Konschnik and Archana Dayalu. Hydraulic fracturing chemicals reporting: Analysis of available data and recommendations for policymakers. *Energy Policy*, 88:504–514, 2016.

Eliza D Czolowski, Renee L Santoro, Tanja Srebotnjak, and Seth BC Shonkoff. Toward consistent methodology to quantify populations in proximity to oil and gas development: a national spatial analysis and review. *Environmental health perspectives*, 125(8):086004, 2017.

Lisa M McKenzie, Ruixin Guo, Roxana Z Witter, David A Savitz, Lee S Newman, and John L Adgate. Birth outcomes and maternal residential proximity to natural gas development in rural colorado. *Environmental health perspectives*, 122(4):412–417, 2014.

Sara G Rasmussen, Elizabeth L Ogburn, Meredith McCormack, Joan A Casey, Karen Bandeen-Roche, Dione G Mercer, and Brian S Schwartz. Association between unconventional natural gas development in the marcellus shale and asthma exacerbations. *JAMA internal medicine*, 176(9):1334–1343, 2016.

Joan A Casey, Deborah Karasek, Elizabeth L Ogburn, Dana E Goin, Kristina Dang, Paula A Braveman, and Rachel Morello-Frosch. Retirements of coal and oil power plants in california: association with reduced preterm birth among populations nearby. *American journal of epidemiology*, 187(8):1586–1594, 2018.

Robert W Howarth. Methane and climate change. *Environmental Impacts from the Development of Unconventional Oil and Gas Reserves; Stolz, J., Bain, D., Griffin, M., Eds*, pages 132–149, 2021.

Hussein Hasan Ahmed. Analysis of heavy metals in coventry lake view park and stoke heath park: Historical pollution. 2015.

Miomir Kostic and Lidija Djokic. Recommendations for energy efficient and visually acceptable street lighting. *Energy*, 34(10):1565–1572, 2009.

Rob Kitchin and Gavin McArdle. The diverse nature of big data. programmable city working paper 15. 2015.

In Lee and Yong Jae Shin. Machine learning for enterprises: Applications, algorithm selection, and challenges. *Business Horizons*, 63(2):157–170, 2020.

Meysam Rajabi, Hamzeh Ghorbani, and Sahar Lajmorak. Comparison of artificial intelligence algorithms to predict pore pressure using petrophysical log data. *Journal of Structural and Construction Engineering*, 9(11):165–185, 2023.

Zachary D Weller, Steven P Hamburg, and Joseph C von Fischer. A national estimate of methane leakage from pipeline mains in natural gas local distribution systems. *Environmental Science & Technology*, 54(14):8958–8967, 2020.

Maria Faragò, Eva Sara Rasmussen, Ole Fryd, Emilie Rønde Nielsen, and Karsten Arnbjerg-Nielsen. Coastal protection technologies in a danish context. *Vand I Byer–Innovationsnetværk fo r Klimatilpasning. Taastrup, Denmark*, 2018.

Mohammad Mehdi Faghih and Mohsen Ebrahimi Moghaddam. Neural gray: A color constancy technique using neural network. *Color Research & Application*, 39(6):571–581, 2014.

Ligang Zhou, Kin Keung Lai, and Lean Yu. Least squares support vector machines ensemble models for credit scoring. *Expert systems with applications*, 37(1):127–133, 2010.

Patrick J Trainor, Andrew P DeFilippis, and Shesh N Rai. Evaluation of classifier performance for multiclass phenotype discrimination in untargeted metabolomics. *Metabolites*, 7(2):30, 2017.

Daniel Jarrett, Eleanor Stride, Katherine Vallis, and Mark J Gooding. Applications and limitations of machine learning in radiation oncology. *The British journal of radiology*, 92 (1100):20190001, 2019.

Xinyang Liu, Anyu Liu, Jason Li Chen, and Gang Li. Impact of decomposition on time series bagging forecasting performance. *Tourism Management*, 97:104725, 2023.

Shuhui Lai, Ahmed Eladawy, Jinming Sha, Xiaomei Li, Jinliang Wang, Eldar Kurbanov, and Qixin Lin. Towards an integrated systematic approach for ecological maintenance: Case studies from china and russia. *Ecological Indicators*, 140:108982, 2022.

Shuyan Wan, Xiaohan Yang, Xinya Chen, Zhaonian Qu, Chunjiang An, Baiyu Zhang, Kenneth Lee, and Huifang Bi. Emerging marine pollution from container ship accidents: Risk characteristics, response strategies, and regulation advancements. *Journal of Cleaner Production*, page 134266, 2022.

Xu Min, Wanwen Zeng, Ning Chen, Ting Chen, and Rui Jiang. Chromatin accessibility prediction via convolutional long short-term memory networks with k-mer embedding. *Bioinformatics*, 33(14):i92–i101, 2017.

Rajeev Ruparathna, Kasun Hewage, Hirushie Karunathilake, Roberta Dyck, Ahmed Idris, Keith Culver, and Rehan Sadiq. Climate conscious regional planning for fast-growing communities. *Journal of Cleaner Production*, 165:81–92, 2017.

Johannes Lederer and Johannes Lederer. Linear regression. *Fundamentals of High-Dimensional Statistics: With Exercises and R Labs*, pages 37–79, 2022.

Pinki Sagar, Prinima Gupta, and Indu Kashyap. A forecasting method with efficient selection of variables in multivariate data sets. *International Journal of Information Technology*, 13: 1039–1046, 2021.

Pawan Kumar Singh, Alok Kumar Pandey, Sahil Ahuja, and Ravi Kiran. Multiple forecasting approach: a prediction of co2 emission from the paddy crop in india. *Environmental Science and Pollution Research*, pages 1–12, 2022.

Boonyasith Chamnanthongpaivanh, Jittima Chatchawalsaisin, and Oran Kittithreerapronchai. Artificial neural network and support vector regression modeling for prediction of mixing time in wet granulation. *Journal of Pharmaceutical Innovation*, pages 1–12, 2022.

Candice Bentéjac, Anna Csörgő, and Gonzalo Martínez-Muñoz. A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, 54:1937–1967, 2021.