

Table of Contents

[Keywords 2](#_Toc144571887)

[Introduction 3](#_Toc144571888)

[Tools and Technologies 5](#_Toc144571889)

[Data Collection and Preparation 7](#_Toc144571890)

[Implementation 8](#_Toc144571891)

[Descriptive Analysis 8](#_Toc144571892)

[Regression Analysis 8](#_Toc144571893)

[Cluster Analysis 9](#_Toc144571894)

[Classification Analysis 10](#_Toc144571895)

[Discussion of Findings. 14](#_Toc144571896)

[Conclusion 23](#_Toc144571897)

[References 24](#_Toc144571898)

[Appendix 25](#_Toc144571899)

# Keywords

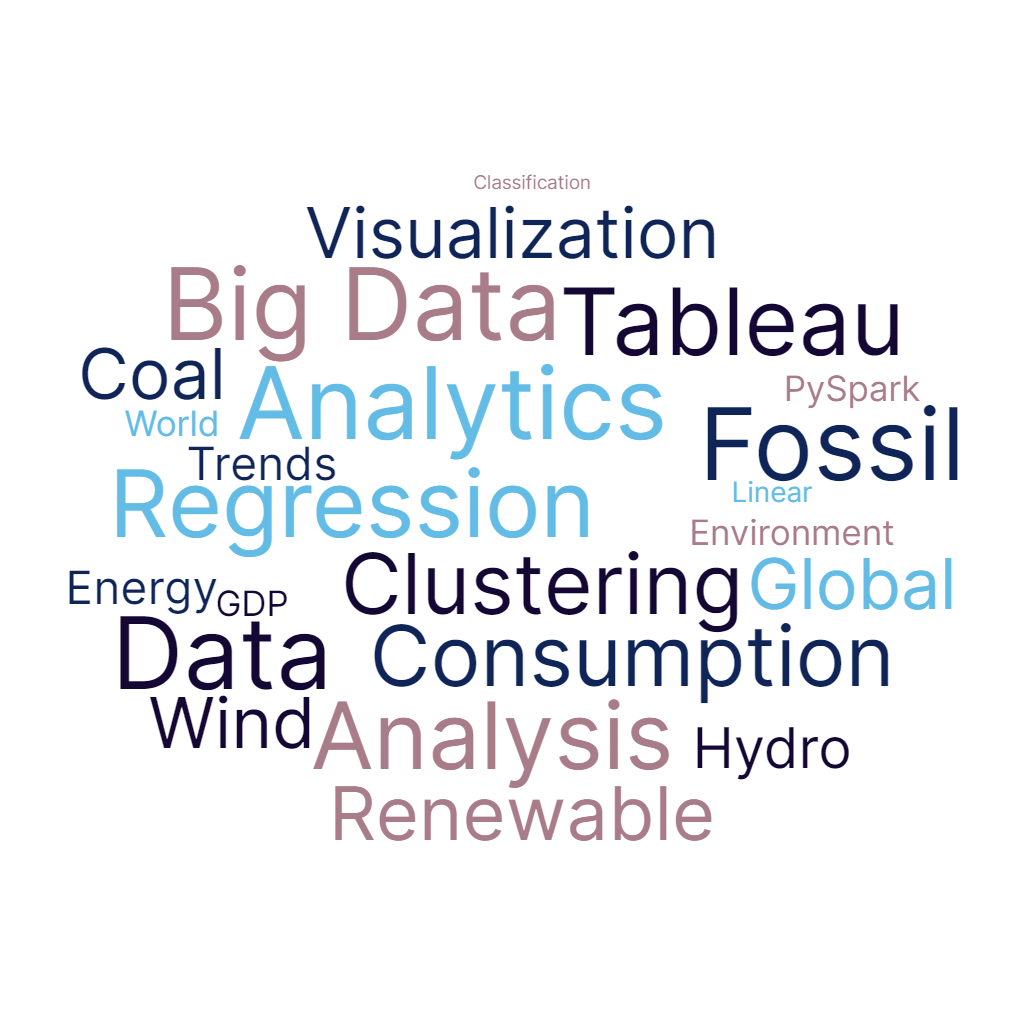
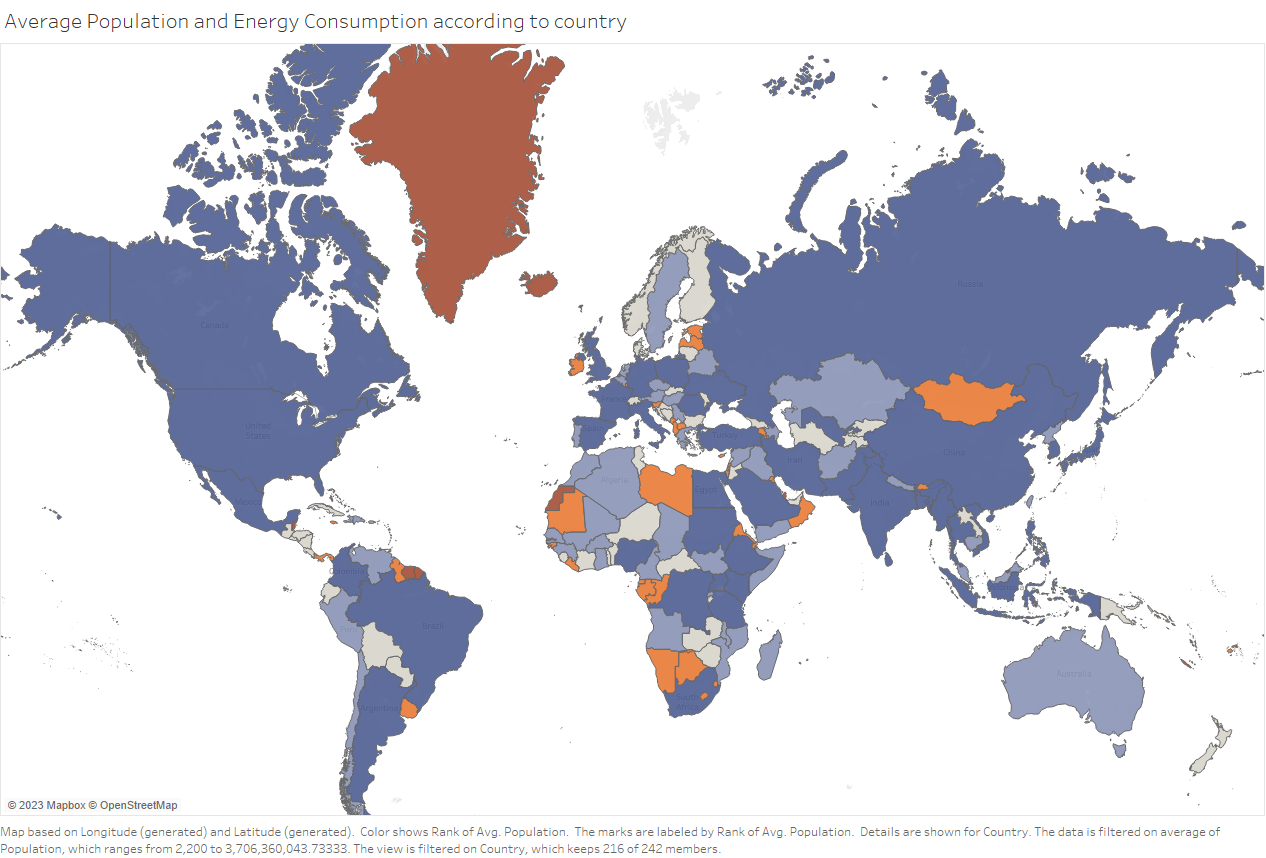


Figure 1: Keywords

# Introduction

Understanding the complex dynamics of global energy use and production is crucial in an era marked by the intersectional issues of increasing demand for energy, preservation of the environment, and economic growth. This study provides a detailed analysis of a large dataset gathered from "**World Energy Consumption**" providing a global perspective of energy-related measures spanning numerous nations and years. With the help of Tableau for visualization and PySpark for data analysis, this study provides vital new information about the ever-evolving environment of global energy trends.

Complete data analysis becomes an essential tool for policy makers, corporations, and researchers as governments work to strike a balance between the necessity of providing energy for economic progress and the critical need to mitigate climate change. The dataset selected for this study has a lot of data, including statistics on electricity generation across a variety of energy sources, annual fluctuations, consumption trends, and production levels. This multidimensional dataset acts as the foundation for a thorough study that uses cutting-edge methodologies to extract insightful data that helps with decision-making.



A comprehensive analytical approach is required because of the complex link between energy use, production, growth in the economy, and environmental sustainability. The core of this study's data analysis approach is PySpark, a potent tool for handling and analyzing massive data. By managing enormous datasets, PySpark makes it possible to identify useful patterns and correlations, enabling the detection of trends in energy consumption, changes in manufacturing techniques, and the development of energy portfolios across countries. Tableau makes it easier to create immersive visualizations that turn complex statistics into useful insights, complementing PySpark's abilities. Tableau brings forth the complex interplay between energy use and production variables by visually presenting the analytical results.

The next sections of this work are organized in sections, starting with a thorough explanation of the attributes of the dataset and the justification for its selection. The methods section then describes the procedures followed to preprocess the data and carry out various analysis, such as regression, clustering, and classification. Assessing energy patterns and their effects on the environment and the economy are the two main goals of these investigations. The results section displays a number of painstakingly designed Tableau dashboards that transform complex data sets into easily understandable representations, enabling viewers to quickly spot hidden correlations and insights.

# Tools and Technologies

This project's implementation relies on the use of PySpark and Tableau to carry out an extensive study of worldwide energy consumption and production trends. Two essential technologies are used in the implementation phase: **Tableau** to create interactive data visualization that visually depict the results of sophisticated analysis and **PySpark** for data processing and analysis.



Figure 2: Tools and Technology

A distributed processing system called PySpark is set up to take advantage of a cluster's computing capability. The comprehensive energy dataset can be handled effectively thanks to its intrinsic capacity to process enormous datasets in parallel. **Jupyter Notebook** is used to code the scripts with the help of anaconda to manage and run the packages used in the analysis.

# Data Collection and Preparation

**The Dataset:**

The "**World Energy Consumption**" dataset, which forms the basis of this implementation, contains a wealth of energy-related indicators spanning numerous nations and decades. This dataset includes information on a wide range of factors, such as annual percentage changes, production levels, consumption rates, and statistics on the amount of power generated using different types of fuels. Due to its comprehensiveness and multidimensional structure, PySpark and Tableau are the perfect tools for a thorough investigation.

*Datalink:* [*https://www.kaggle.com/datasets/pralabhpoudel/world-energy-consumption*](https://www.kaggle.com/datasets/pralabhpoudel/world-energy-consumption)

**Data Loading**: One the data is ready to be analyzed, the dataset is loaded into a PySpark Data Frame, enabling distributed processing.

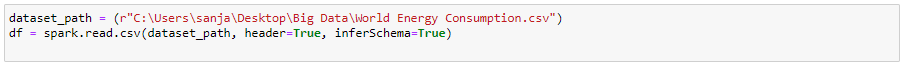
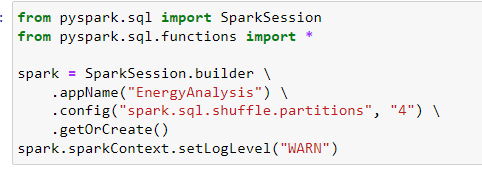


Figure 5:Loading Data in PySPark

**Processing**: Missing values are handled, and relevant columns are selected for analysis.



Figure 3: Handling Missing Values

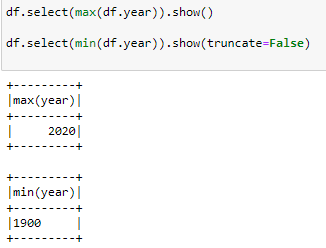
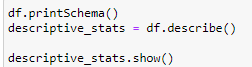
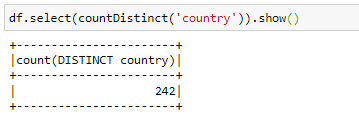


Figure 4: Selection of Relevant columns

# Implementation

Descriptive Analysis: To comprehend the fundamental properties, distribution, and correlations of the data, descriptive analysis entails summarizing and visualizing it. It offers perceptions into the primary patterns and variations present in a dataset, giving a clear understanding of the main characteristics of the data. For instance, descriptive analysis aids in identifying trends and patterns in energy consumption across many nations and epochs in the context of energy consumption and GDP growth analysis.

Basic statistics and data distribution are computed using PySpark's functions.



Regression Analysis: A statistical technique for simulating and quantifying correlations between variables is regression analysis. Because of the values of the other variables, we can anticipate the value of one variable. Regression analysis, which provides coefficients that characterize the degree and direction of this relationship, aids in assessing how changes in GDP growth effect energy use in the context of energy consumption and GDP growth. Modeling relationships between variables, such as Fossil Fuel consumption and GDP growth. Regression models are constructed using PySpark's `pyspark.ml.regression` module.

Modeling relationships between variables, such as Fossil Fuel consumption and GDP growth. Regression models are constructed using PySpark's `pyspark.ml.regression` module. The model was also evaluated using **Root Mean Squared Error (RMSE)** which was only 1.028166580495679e-11. Suggesting it was closer to zero. A very small **RMSE** indicates that the predicted values are very close to the observed values, suggesting that the model's predictions are highly accurate.

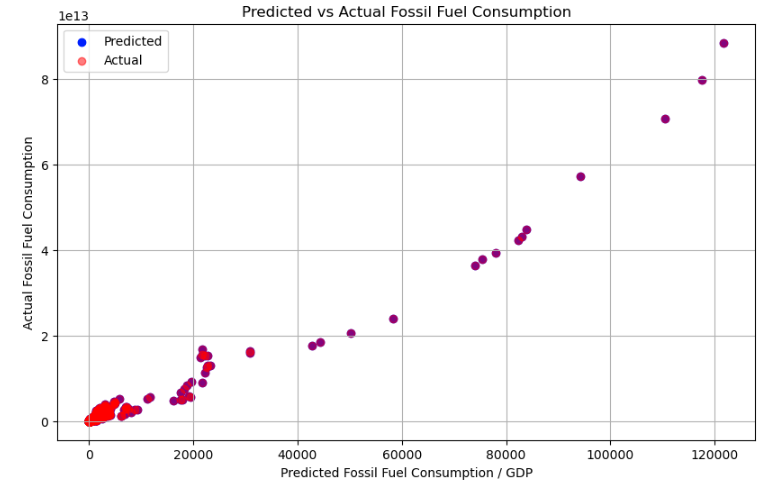


Figure 6: Linear Regression Plot of Predicted and Actual Values

The predicted values are denoted by blue color whereas the red color dots are the actual values.

Cluster Analysis: An unsupervised learning method called clustering analysis is used to put comparable data points collectively based on their properties. It assists in finding trends and resemblances in a dataset. When used to group countries with comparable patterns of energy production and consumption, clustering might provide information about local patterns or shared traits between nations. Clustering is performed using the `pyspark.ml.clustering` package. , clustering was also done using the average fossil fuel consumption and the average population of countries from 1990 to 2020.

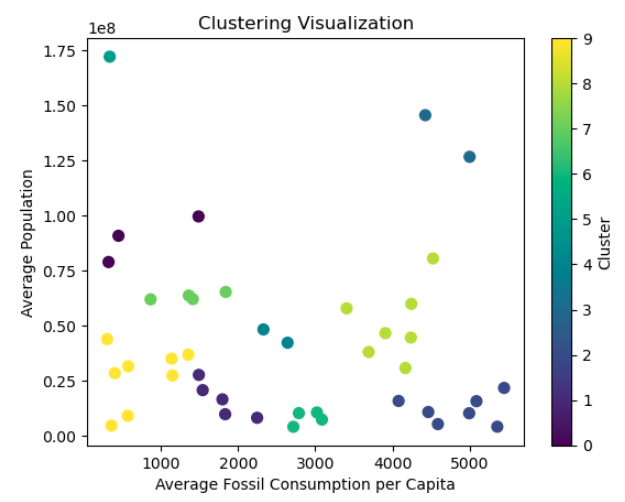


Figure 7: Clustering based on Avg Fossil Fuel Consumption and Population

Classification Analysis: Based on input features, classification modeling is used to forecast categorical outcomes or classes. When categorizing or labeling data points, it is useful. In the context of energy analysis, categorization can be used to forecast a nation's reliance on different energy sources (such as renewable, fossil fuels, and nuclear), based on past energy consumption patterns and other pertinent factors, enabling a greater knowledge of energy consumption patterns. The classification modules in PySpark are used to carry out classification tasks. The duty of classifying nations according to their main energy sources was established. Specifically, a criterion was used to divide countries into two groups based on their share of renewable energy sources. On the provided training dataset, a machine learning classification approach called Logistic Regression was used to train the model.

A variety of crucial classification variables, such as **accuracy, precision, recall, and F1-score** were used to thoroughly evaluate the model's performance. Notably, the model attained an impressive level of accuracy and had a well-balanced F1-score, highlighting its efficiency in correctly classifying nations on the basis of their energy sources. A representation of the confusion matrix was created in order to acquire more insight into the model's performance. This matrix made it easier to examine true positives, true negatives, false positives, and false negatives, giving a complete picture of the classification abilities of the model.

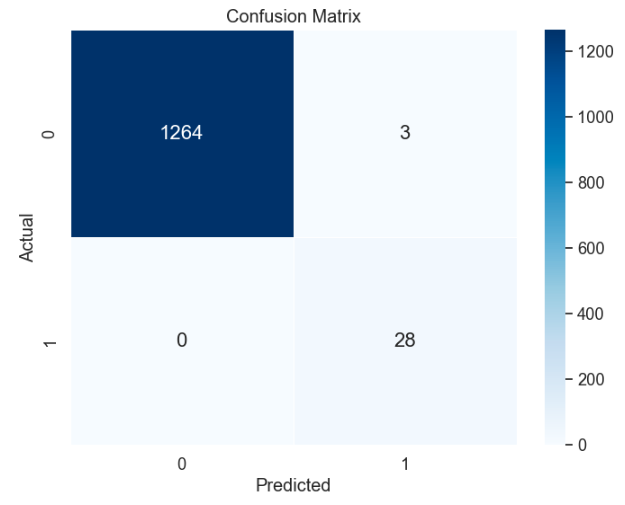


Figure 8: Confusion Matrix

Similar to that, PySpark was also used to classify nations according to their major energy use. Energy-related features were changed using feature engineering with the required scale. The classification of nations was based on factors like energy usage. For the purpose of evaluating performance, the dataset was divided into training and testing sets. With room for more complex models, a logistic regression model was initially used. Metrics for the categorization task-specific model evaluation comprised **accuracy**, **precision, recall, F1-score, and ROC-AUC.** To improve prediction performance, model tuning involved feature and hyperparameter improvement.

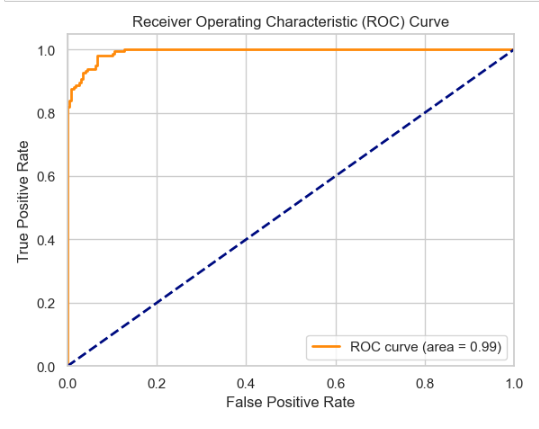


Figure 9: ROC Curve

**Data Transformation**: To get data ready for visualization tasks like aggregations, transformations, and filters need to be carried out.

**Visualization in Tableau**: The results of the analysis are transformed into dynamic Tableau Charts that display trends, patterns, and correlations.

Loading the data in tableau for visualization.

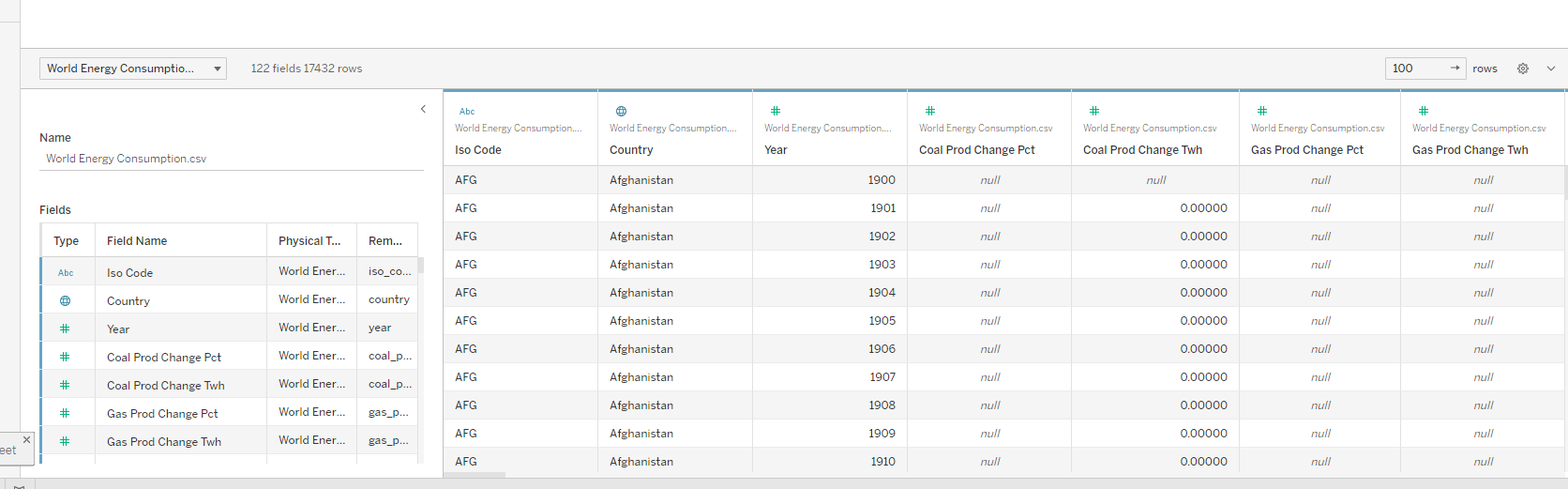
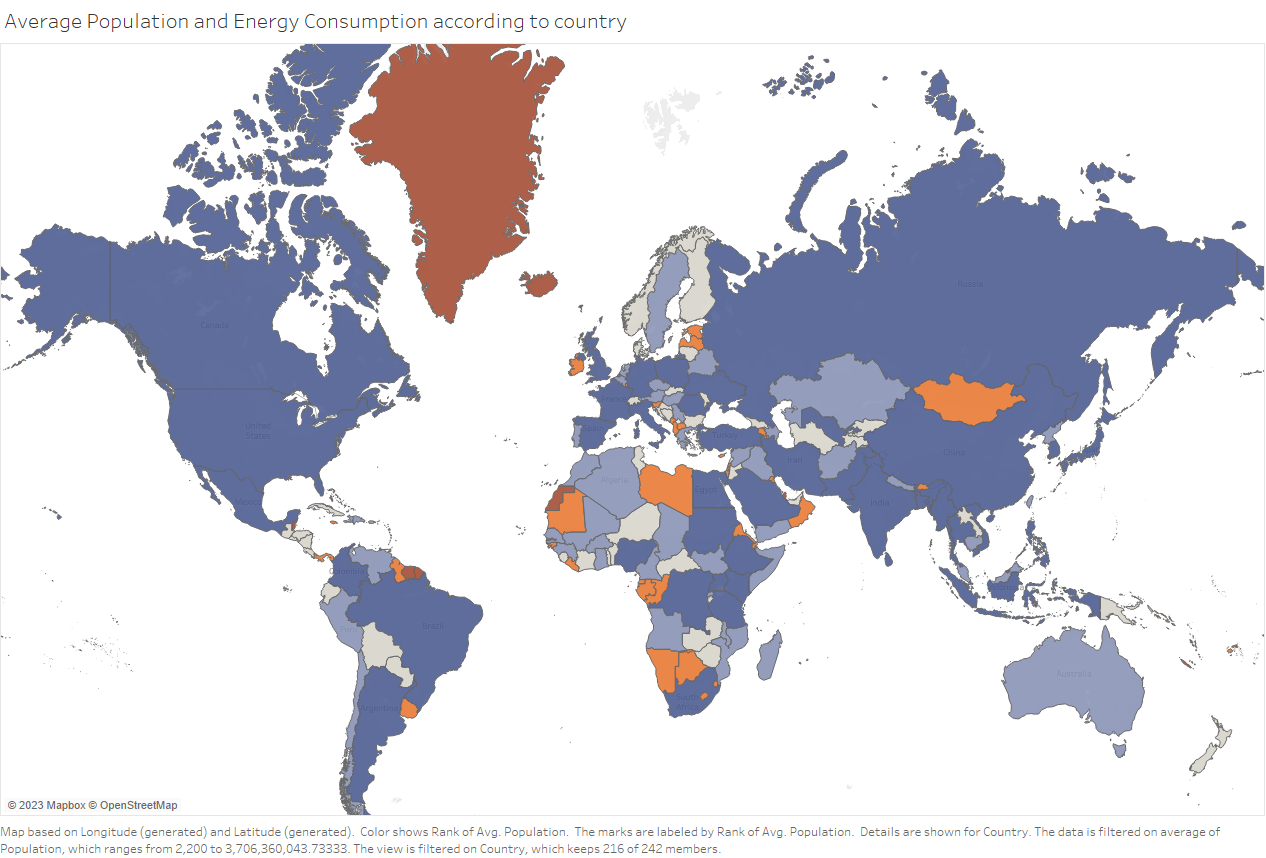


Figure 10: Data Loaded in Tableau



With the help of this implementation, we can undertake a thorough analysis of the patterns of global energy production and consumption. PySpark's parallel processing abilities are used to examine the enormous and complex dataset and extract valuable insights. Tableau then combines these facts into interactive dashboards that are visually appealing and help stakeholders quickly understand complex energy trends. By using an integrated approach, the implementation provides stakeholders with a way to comprehend how the global energy landscape is changing, to influence sustainable decision-making, and to contribute to a future of responsible energy use and preservation of the environment.

# Discussion of Findings.

Insightful results from the study of the extensive data on world energy production and consumption shed light on the intricate interactions among sources of energy, consumption trends, financial indicators, and environmental ramifications. This talk covers the major conclusions drawn from the data analysis performed with PySpark and the consequences for policy regarding energy, sustainability, and growth in the economy.

**Shift Towards Renewable Energy Sources**:

Recent statistics and trends make it clear that the world has been consuming more renewable energy. The largest rate of renewable energy consumption is in China consuming around 5000 terawatt-hours, which is closely followed by Europe with around 4500, the US also with 4500. This worldwide trend of nations shifting more and more to sustainable energy sources is reflected in the transition toward renewable energy, which is very significant.

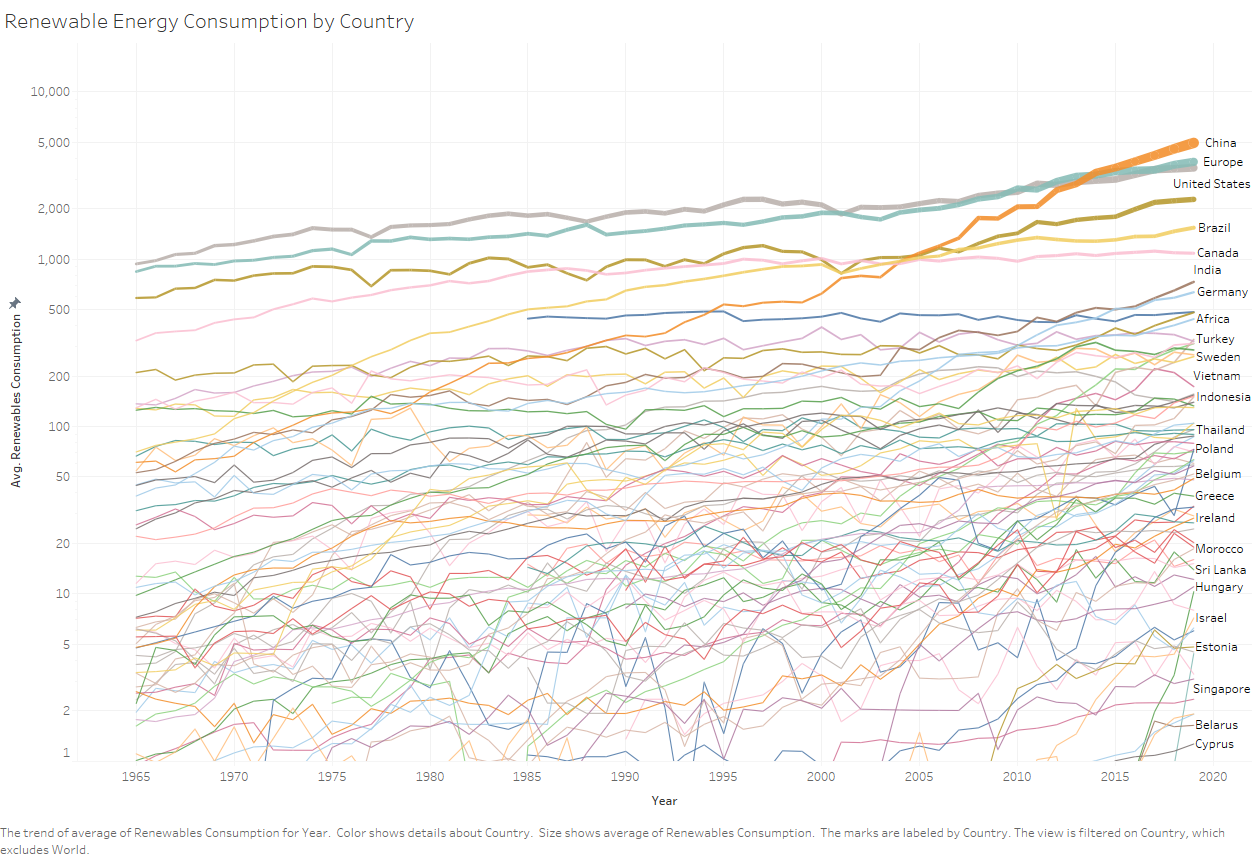


Figure 11: Renewable energy Trends from 1965-2020

The shift away from non-renewable energy sources is a positive development for both the environment and energy sustainability. It reflects a growing awareness of the environmental harm caused by non-renewables, such as carbon emissions and climate change. Moreover, the data indicates that renewable technologies have become more affordable and efficient, appealing to both developed and emerging economies. Globally, this transition reduces greenhouse gas emissions and enhances energy security by reducing fossil fuel reliance. It's a clear sign of the world's move towards a more sustainable energy future, expected to persist as renewable technologies advance and global climate efforts intensify.

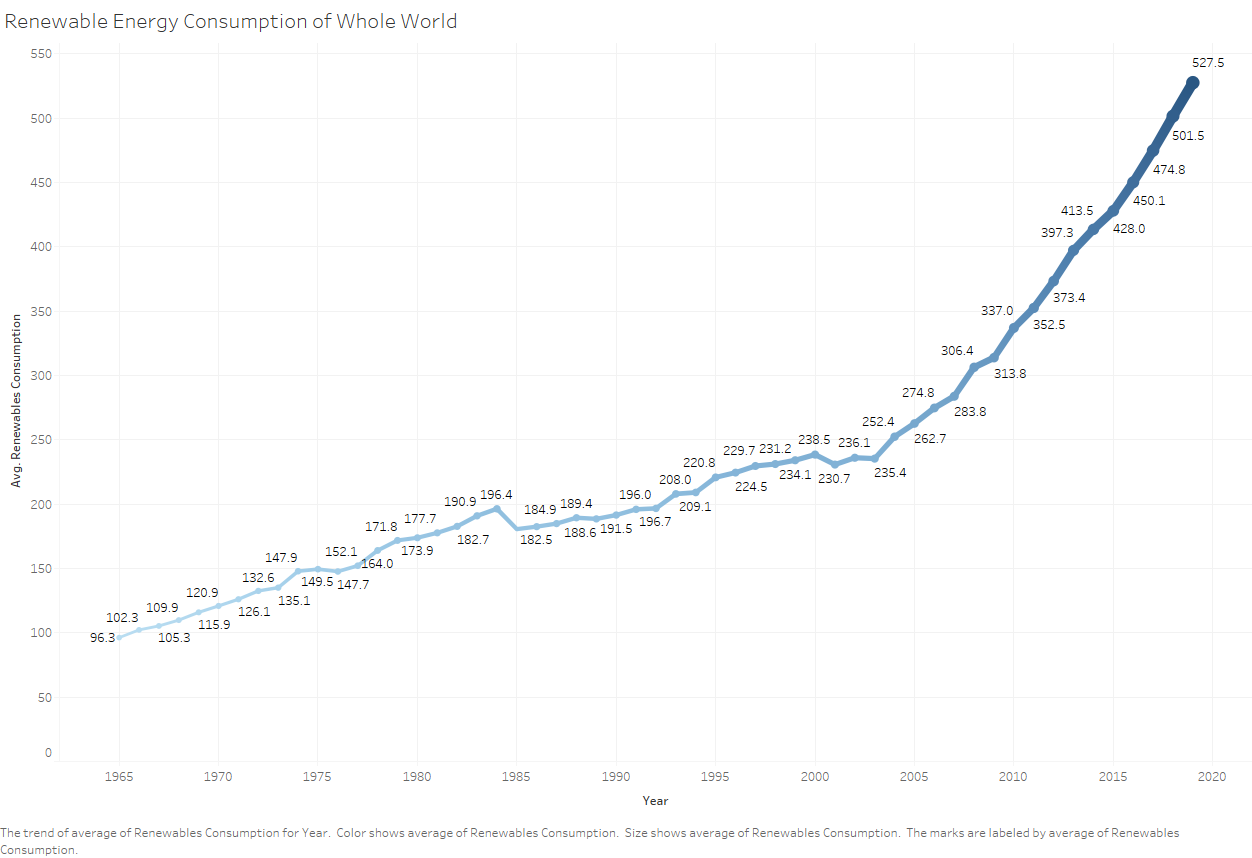


Figure 12: Rise in Renewable Energy Consumption

The per capita coal consumption for the entire population has displayed a notable decline over time. From 1990 onwards, there was a substantial decrease in coal consumption per person. However, there was a temporary rise in coal consumption in the early 2000s. Yet, recent data shows a clear reversal in this trend, with coal consumption per capita declining significantly from 2018 to 2019, marking a significant drop in recent years.

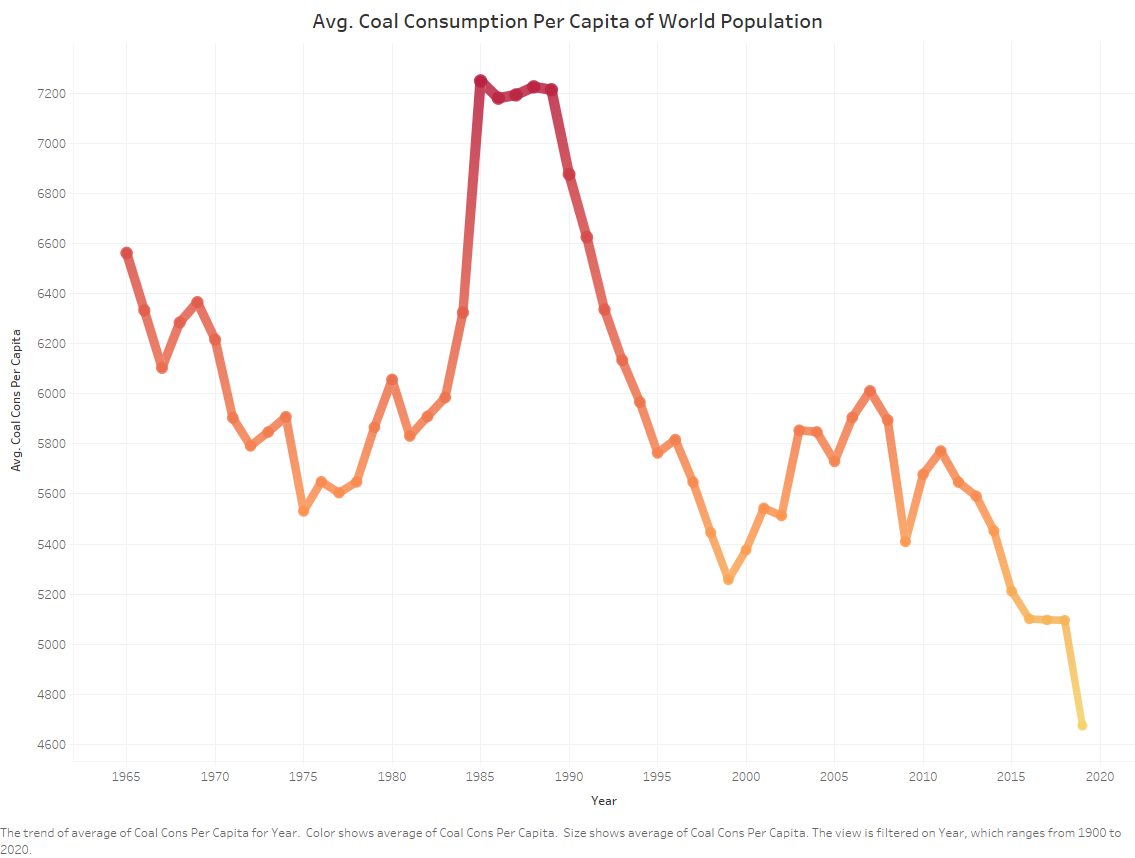


Figure 13: Decrement in Coal Consumption

This shift away from coal consumption per person is a noteworthy development with implications for both environmental sustainability and energy policy. It suggests a decrease in the reliance on coal, a fossil fuel associated with high carbon emissions and air pollution. The data underscores efforts to reduce coal consumption and transition to cleaner energy sources, aligning with global objectives to combat climate change and improve air quality. This shift also reflects changes in energy production and consumption patterns, indicating a move towards more sustainable and environmentally responsible energy choices.

**Energy Consumption and Economic Growth:**

The results of the regression analysis have unveiled a complex relationship between energy consumption and economic growth. It's clear that countries with higher energy consumption often demonstrate stronger economic performance, suggesting the vital role of energy in driving economic activities. However, this relationship is far from straightforward and depends significantly on the composition of the energy mix and the implementation of efficiency measures within each country.

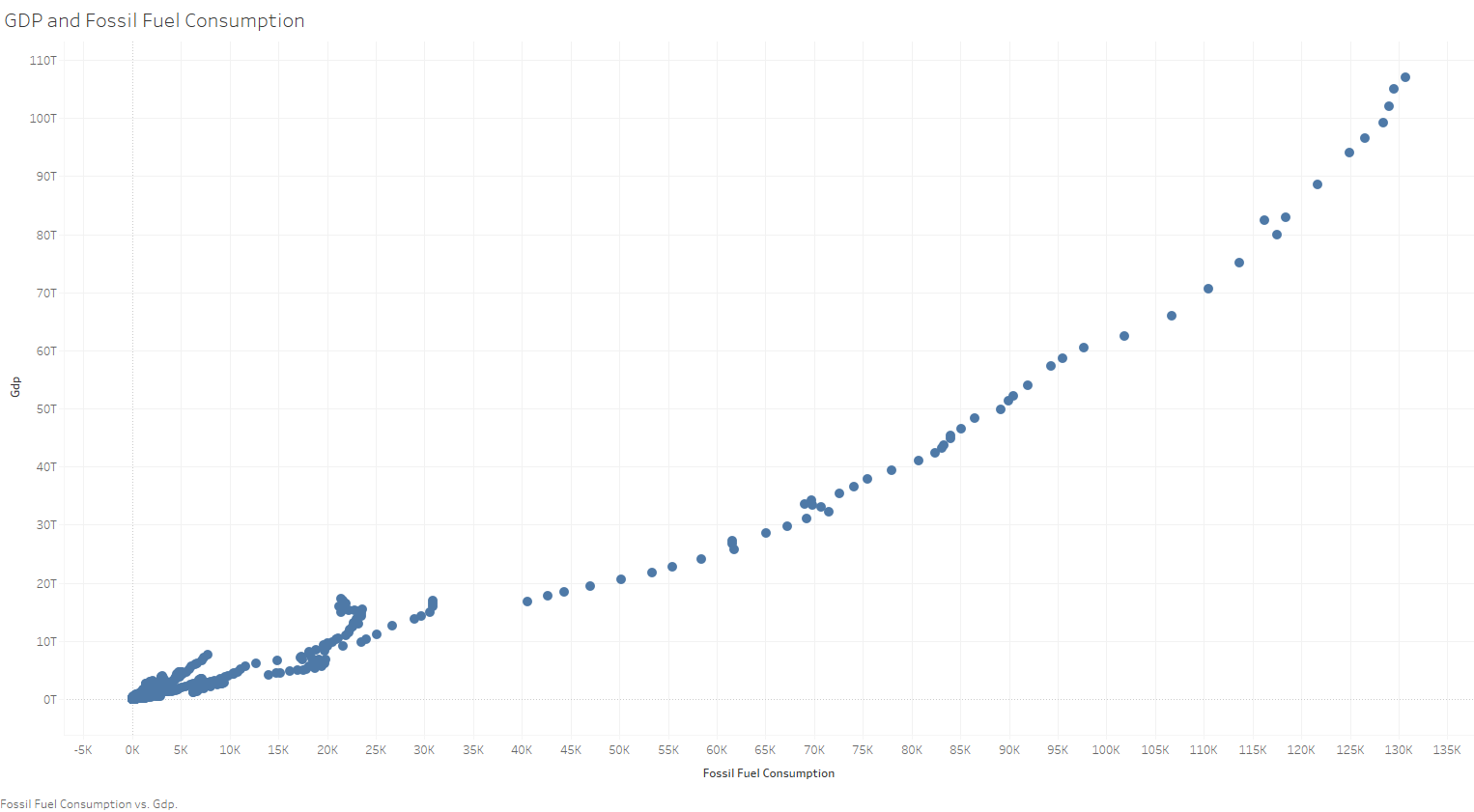


Figure 14: Linear Regression Showing Correlation between Fuel Consumption and GDP

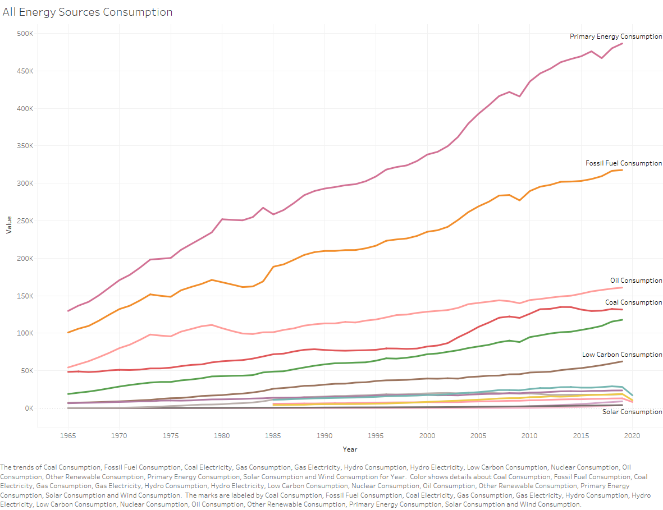


Figure 15: Trend Showing All Energy Sources Consumption

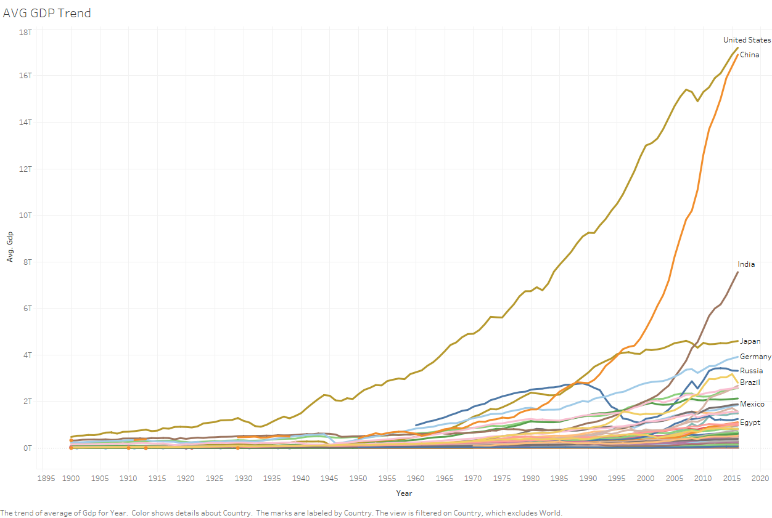


Figure 16: Average GDP of Countries from 1990

These findings underscore the critical importance of tailored energy efficiency initiatives. By optimizing energy use and investing in technologies that enhance energy efficiency, countries can effectively decouple energy consumption from economic growth. This not only contributes to economic sustainability but also mitigates the environmental impact associated with increased energy consumption. In essence, the results highlight the need for a balanced approach in energy policy and strategy, one that promotes both economic development and environmental sustainability by harnessing the benefits of energy while minimizing its negative effects.

**Clustering Patterns:**

Clustering is a technique used for grouping similar data points together based on their inherent characteristics or attributes. The goal of clustering is to identify patterns, similarities, or natural divisions within a dataset, where data points within the same cluster share more similarities with each other than with those in other clusters. Clustering can be valuable for exploring complex datasets, uncovering hidden structures, and gaining insights into the relationships between data points.

The clustering analysis, a fundamental technique in data science, played a pivotal role in our study as it unveiled distinct energy consumption and production patterns across various countries. This analytical approach allowed us to categorize countries into clusters based on similarities in their energy profiles. These clusters shed light on the profound influence of geographical, economic, and policy factors on a nation's energy choices. Notably, our analysis revealed that China stands out as both the largest producer and consumer of energy globally, with a substantial chrome production of approximately 4500 units and an energy consumption of about 9500 terawatt-hours.

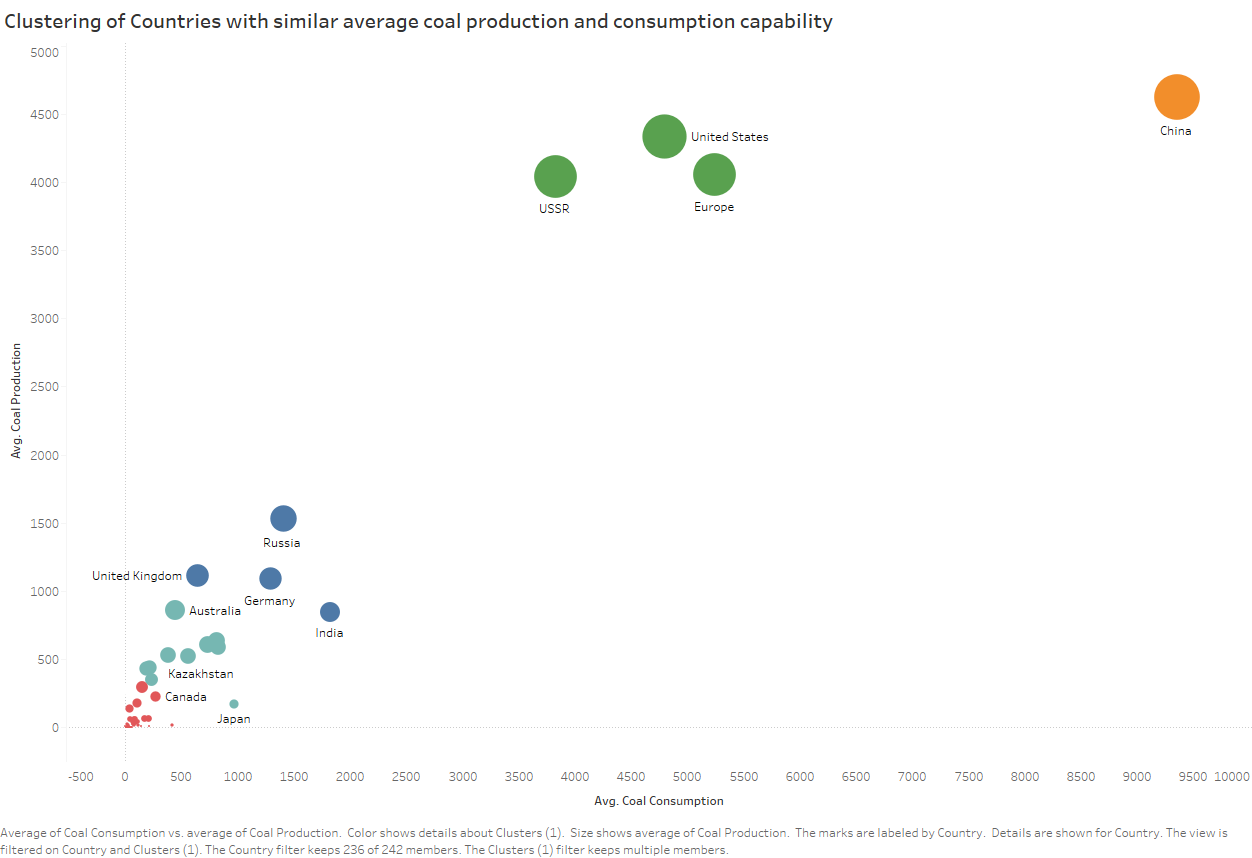


Figure 17: Clustering Based on average coal production and consumption

Furthermore, the clustering approach highlighted two other significant clusters. The first cluster includes the United States, Europe, and the USSR, suggesting that these countries share common patterns in energy consumption and production. In contrast, the third cluster comprises countries such as Russia, the UK, Germany, India, which exhibit similar energy consumption and production trends. These clusters offer valuable insights into how nations with comparable energy profiles can collaborate and learn from each other's strategies.

**Electricity Generation Trends:**

The visualizations of electricity generation trends have provided encouraging insights into the evolving landscape of energy production. Notably, there is a promising upward trajectory in the utilization of low-carbon electricity sources, including wind, solar, and nuclear power. This discovery strongly aligns with global initiatives aimed at diversifying energy sources and mitigating the impacts of greenhouse gas emissions. Particularly noteworthy is the prominence of nuclear power within the low-carbon energy mix. Nuclear power's consistent and stable contribution stands out as a crucial component in the ongoing transition towards cleaner and more sustainable energy systems. Its reliability plays a pivotal role in maintaining a steady energy supply while reducing the carbon footprint.

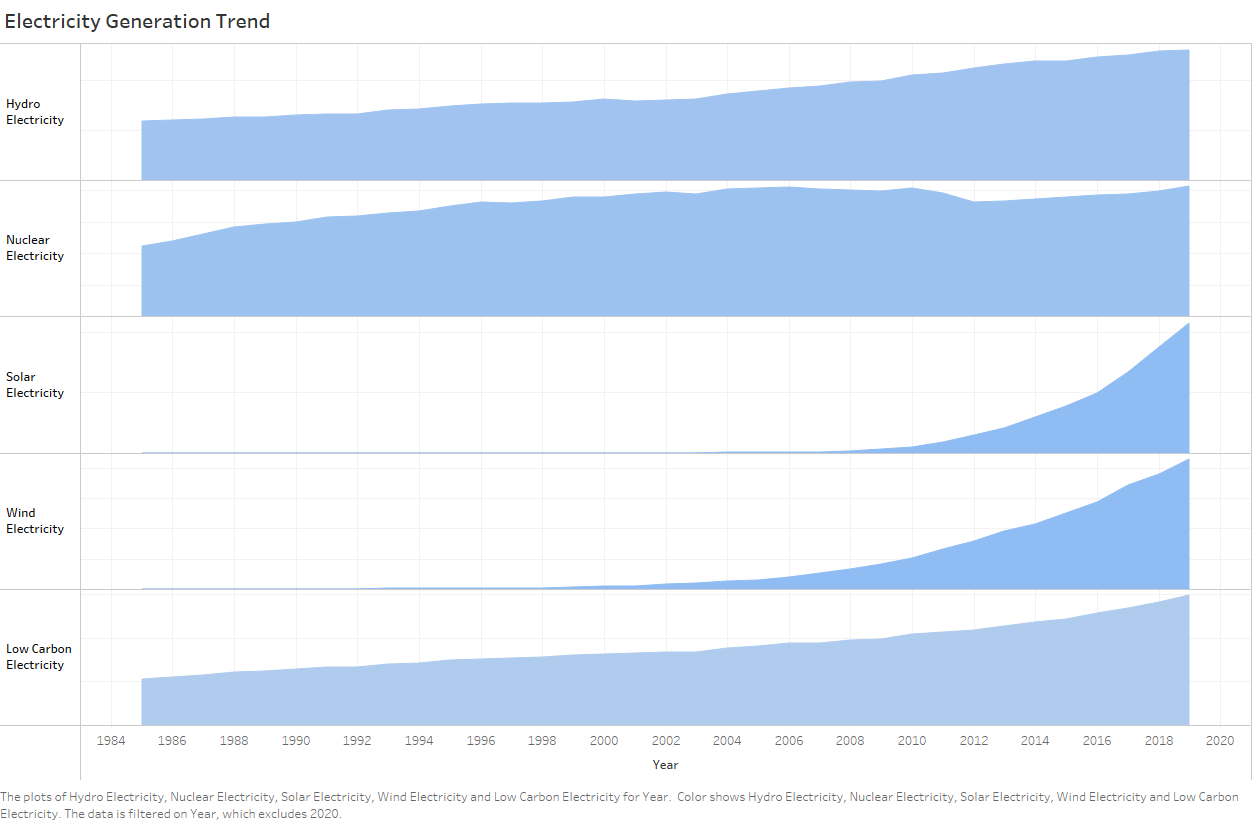


Figure 18: Electricity Generation From Different Sources

The chart further accentuates these trends by illustrating historical electricity source consumption patterns. Hydroelectricity and nuclear electricity have been consistently utilized by countries since 1985, experiencing gradual increases in consumption. However, it becomes particularly evident that electricity sources like solar, wind, and low-carbon electricity have been on a pronounced upward trajectory since around 2008. Notably, some sources of electricity, characterized as "good electricity," have seen a rapid rise since 2010. This data underscores a global commitment to shifting towards renewable energy sources such as hydro, nuclear, and wind, as well as low-carbon alternatives. It is a compelling indication that the world is progressively embracing cleaner and more sustainable energy options, marking a significant step forward in the quest for a greener and more environmentally responsible energy future.

**Pareto Analysis:**

The Pareto Principle, also known as the 80/20 rule, asserts that roughly 80% of outcomes stem from approximately 20% of inputs or causes. This principle, coined by economist Vilfredo Pareto, is widely applicable in economics, business, and life, highlighting the disproportionate impact of certain factors.

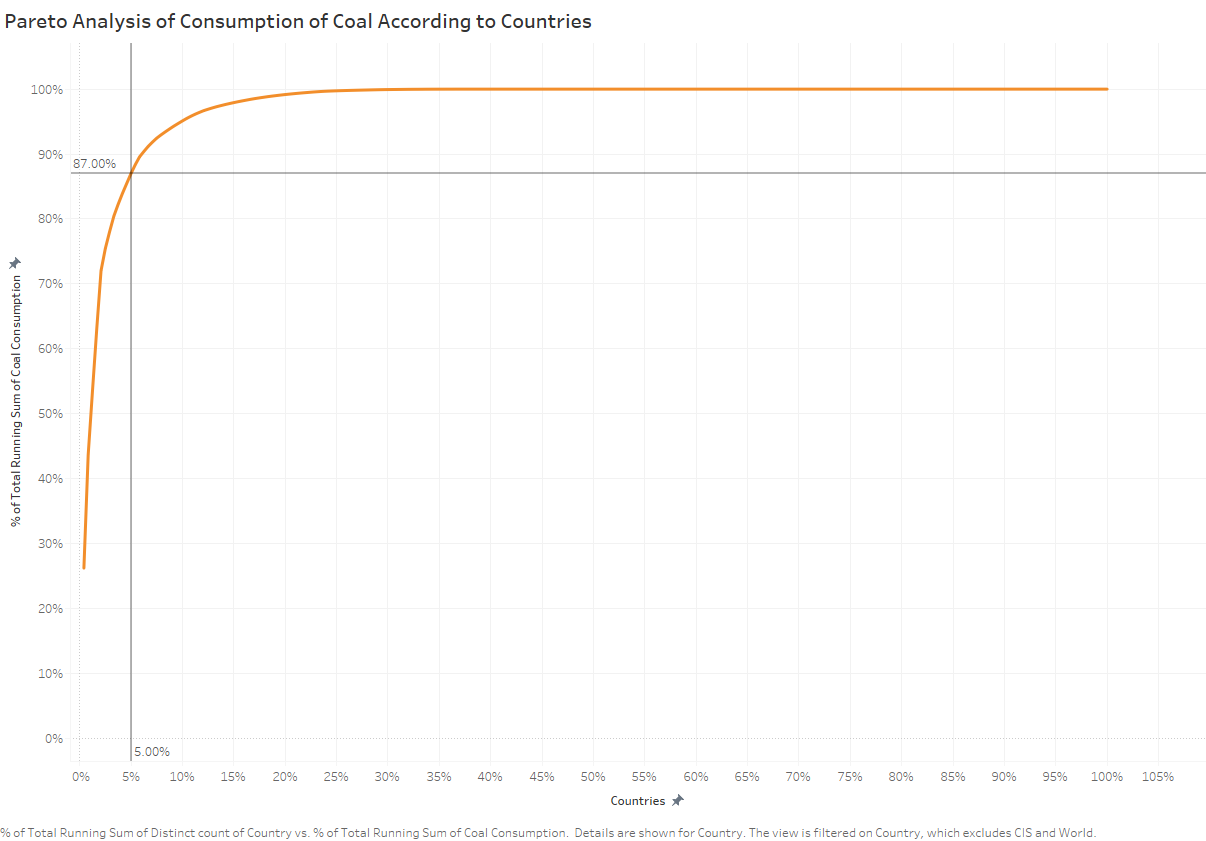


Figure 19: Pareto Analysis on Coal Consumption

Applying the Pareto Principle to global coal consumption reveals: just 5% of countries are responsible for a staggering 87% of the world's coal usage. This highlights the outsized influence of a handful of nations on global coal consumption and its environmental implications, given coal's contribution to climate change.

**KPI of Countries and Fossil Fuel Consumption from 2000 – 2020**

The calculated field used for the KPI

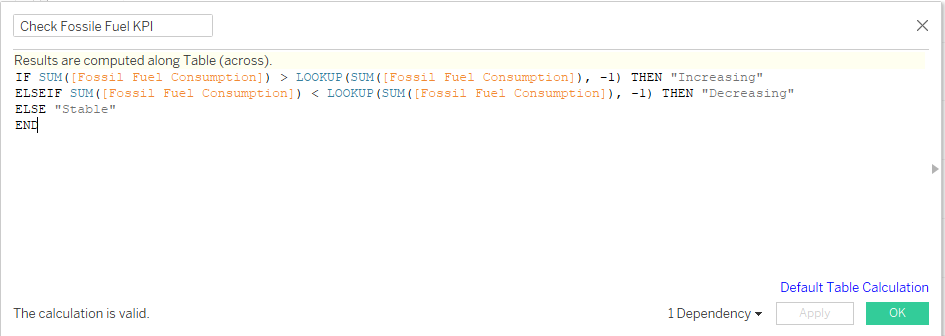


Figure 20: Calculated Field For KPI

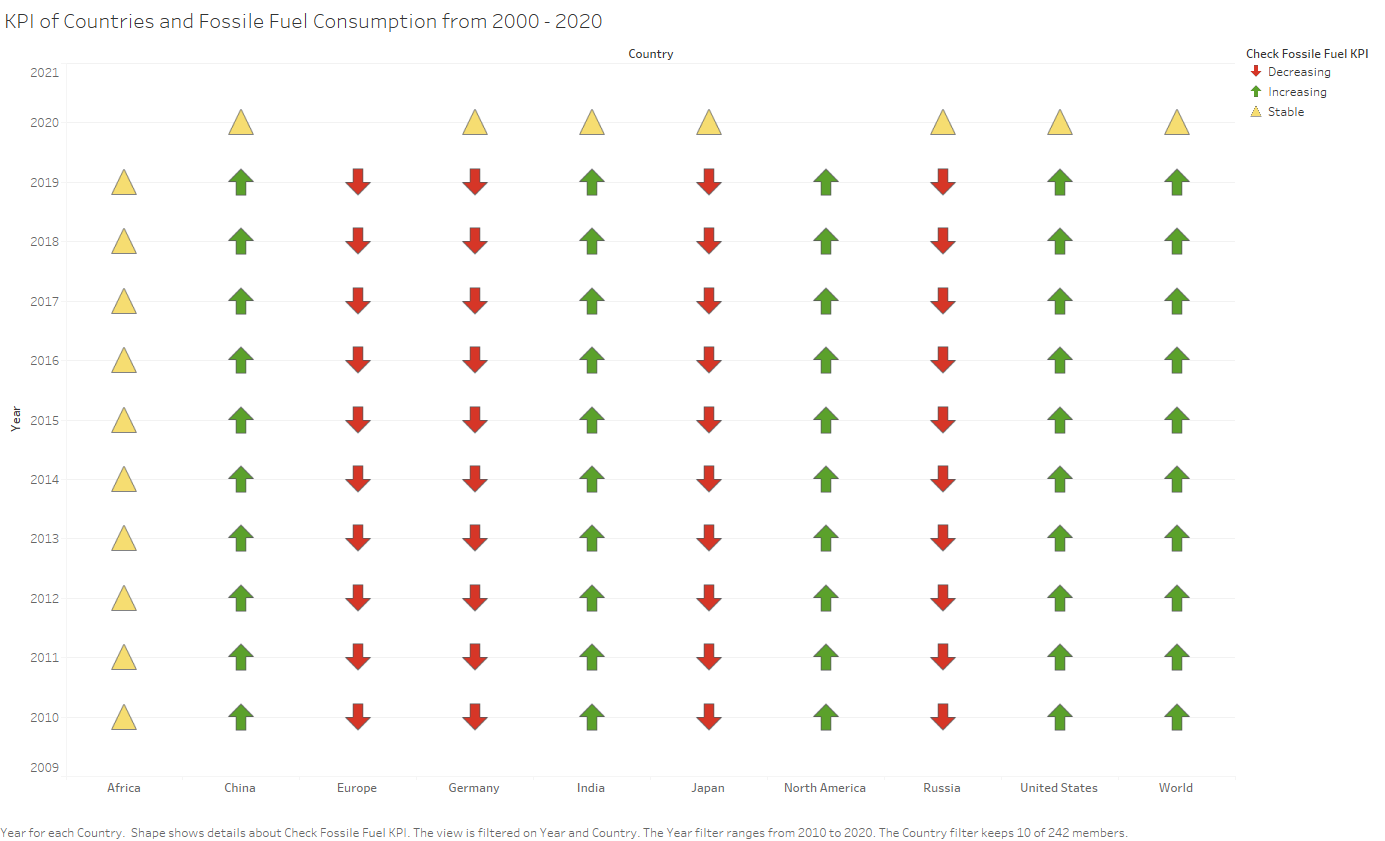


Figure 21: KPI Chart

This KPI chart shows the trend of countries from 2000 to 2010 of consumption of fossil fuel using the calculated field shown above.

# Conclusion

As this comprehensive analysis concludes, it becomes evident that advanced tools like PySpark and Tableau have facilitated a meticulous exploration of the intricate global energy landscape. Employing sophisticated techniques such as regression and clustering, this endeavor has shed light on the indispensable role of data-driven insights in shaping sustainable energy strategies on a global scale.

Among the noteworthy findings, a substantial shift towards renewable energy sources emerges, as indicated by the increasing prominence of wind, solar, and low-carbon alternatives in electricity generation. This observed trend underscores a collective commitment to reducing carbon emissions and embracing environmentally responsible energy solutions.

Furthermore, the regression analysis has unveiled the intricate relationship between energy consumption and economic growth. While a positive correlation between higher energy utilization and robust economic performance is apparent, the strength of this link is intricately tied to energy efficiency and the employed energy mix. This emphasizes the paramount importance of targeted energy efficiency measures.

Moreover, the clustering analysis has revealed diverse energy consumption patterns, influenced by geographical, economic, and policy factors. These insights hold the potential to guide tailored policy interventions and foster collaborative knowledge-sharing among countries with similar energy profiles.

Additionally, through the utilization of Tableau, a visually compelling trend towards low-carbon energy sources becomes evident. This trend aligns harmoniously with global endeavors to diversify energy sources and combat climate change. The notable prevalence of nuclear energy within the low-carbon mix underscores its pivotal role in the transition towards cleaner and more sustainable energy systems.

The implications of these insights extend to a variety of stakeholders, including policymakers, businesses, and investors. They serve as valuable guides for decision-makers to align their strategies with global trends. Looking forward, the integration of advanced analytical techniques and longer-term analysis offers the potential for even deeper insights. In conclusion, this analysis exemplifies the profound impact of data-driven exploration in shaping a sustainable and responsible energy future.

# References

Naveen. (2023a). How to install PySpark on Windows. *Spark by {Examples}*. https://sparkbyexamples.com/pyspark/how-to-install-and-run-pyspark-on-windows/

*Anaconda navigator#*. Anaconda Navigator - Anaconda documentation. (n.d.). https://docs.anaconda.com/free/navigator/index.html

*Descriptive analytics: Steps, techniques, use case, examples*. KnowledgeHut. (n.d.). https://www.knowledgehut.com/blog/data-science/descriptive-analytics

Pryke, B. (2023). How to use Jupyter Notebook: A beginner’s tutorial. *Dataquest*. https://www.dataquest.io/blog/jupyter-notebook-tutorial/

*Data Classification: Definition, Types, & Best Practices*. (n.d.). https://levity.ai/blog/data-classification-types-applications

*KPI calculations*. (n.d.). Tableau. https://www.tableau.com/drive/kpi-calculations

Naveen. (2023). PySpark withColumnRenamed to rename column on DataFrame. *Spark by {Examples}*. https://sparkbyexamples.com/pyspark/pyspark-rename-dataframe-column/

Blanchard, T. (2022, January 6). The Pareto principle — spending time and energy effectively as a data scientist. *Medium*. https://towardsdatascience.com/the-pareto-principle-spending-time-and-energy-effectively-as-a-data-scientist-8d48f7cc2b0f

# Appendix