**Data Preparation, Feature Engineering and Model Exploration**

## Project Title: Analyzing the Impact of Renewable Energy Adoption on Economic Recovery of Developing Nations

## 1. Overview

Data preparation is a crucial step in any machine learning project that transforms raw data into a clean, structured format ready for analysis. It involves several processes such as data cleaning, handling missing values, and dealing with inconsistencies. This stage ensures that all variables are accurate and relevant by removing duplicates, correcting errors, and filtering out unnecessary data. Another key aspect of data preparation is normalization or standardization, where numerical features are scaled to a consistent range to prevent bias in algorithms sensitive to data magnitude. Additionally, categorical variables are often encoded into numerical formats to be effectively used by machine learning models. Effective data preparation lays the groundwork for the model to learn patterns accurately and efficiently.

Feature engineering is the process of refining existing features or creating new ones to improve the model’s predictive power. This step involves selecting the most relevant variables to enhance model performance and reduce computational complexity. By transforming or combining existing features, new insights can be captured—such as calculating energy intensity by dividing energy consumption by GDP, thus offering a more insightful feature. Additionally, feature engineering leverages domain-specific knowledge to create meaningful features that help models capture complex relationships that raw data might not fully express. For instance, in a renewable energy study, the proportion of renewable energy adoption compared to total energy consumption can be a key feature.

## 2. Data Collection

The data for this research project has been collected from two primary sources:

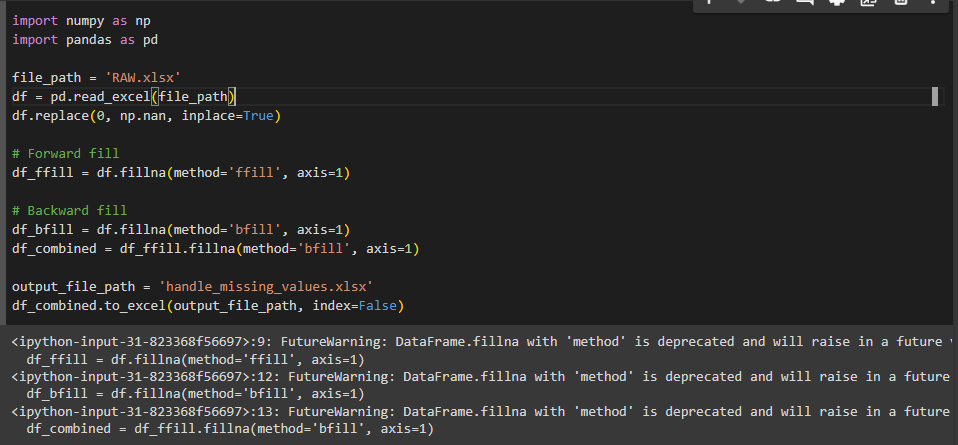
1. World Development Indicators (WDI) from the World Bank[1],

2. UNDP Data Futures Platform[2]

The data extracted from the World Development Indicators is in an Excel format (`.xlsx`). The dataset is organized in a structured format, with rows representing different years and columns representing various indicators relevant to the study.

## 3. Data Cleaning

The data on downloading had 37 columns with different indicators of the country. So, it was reduced to 20 relevant parameters covering per capita electricity cost, the percentage of renewable energy in total energy consumption, the cost of electricity imports and exports, and both domestic and foreign debt levels. The data was arranged according to the increasing order of time with first rows will represent the data of countries for that period of time and again repeated. The data was reshaped such that each country has its data from 1990 to 2023 in consecutive rows. The reshaped data was used for the data cleaning, using forward and backward filling. This approach is used to handle the missing data of time series which satisfies our case.



## 4. Exploratory Data Analysis

CO2 Emissions:

* Temporal Trends: A general decline in CO2 emissions was observed over time, suggesting increased efficiency or a shift towards cleaner energy sources.
* Regional Variations: Countries like Armenia, Kyrgyz Republic, and Uzbekistan showed initial high emissions followed by a significant decrease. In contrast, many least developed countries maintained low or negligible emissions throughout the period.
* Country-Specific Patterns: Kazakhstan and Moldova exhibited relatively high CO2 emissions, likely due to their industrial activities.

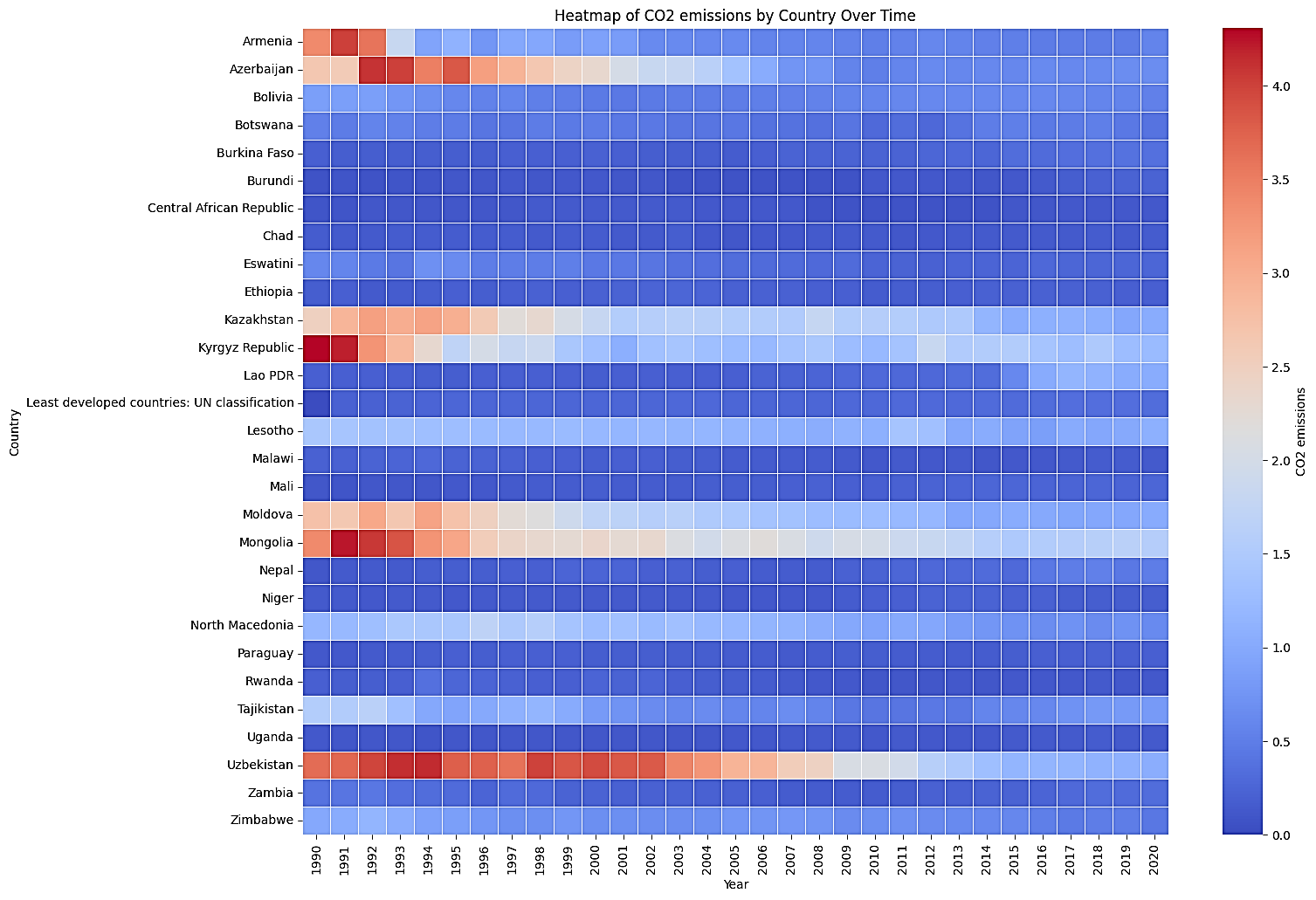


Figure 1. CO2 Emissions Over Time Period

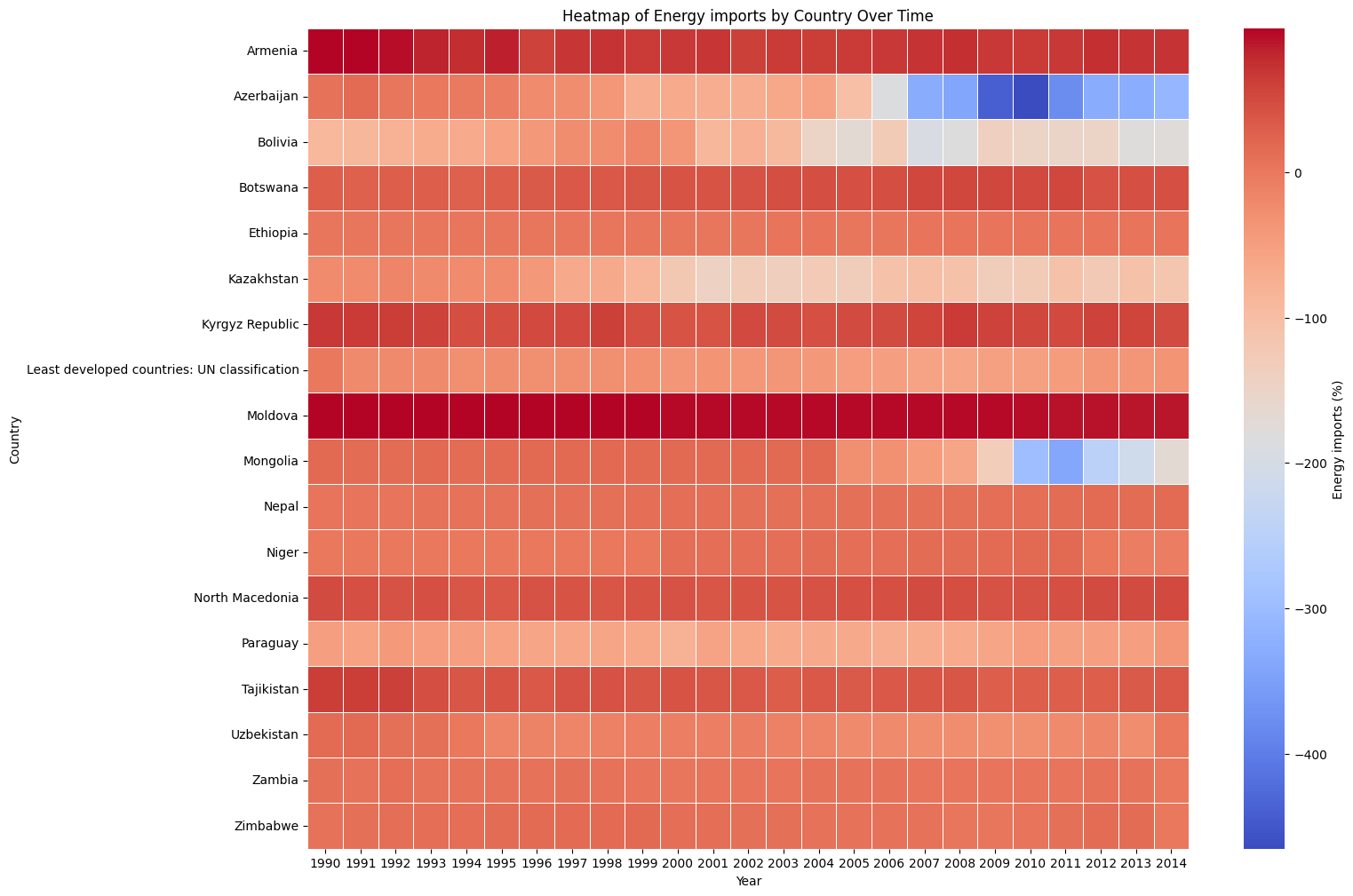


Figure 2. Energy Imports Over Time Period

Energy Imports:

* Import-Export Dynamics: Several countries transitioned from net energy importers to exporters or significantly reduced their energy dependence.
* Temporal Shifts: The colour changes in the heatmap indicate shifts in energy trade balances over time, potentially reflecting policy changes or resource development.
* Regional Variations: Moldova and Uzbekistan had notable negative energy imports, suggesting they might have been net energy exporters in the earlier years.

These findings provide a foundation for further exploration into the underlying factors driving these trends. Potential areas of investigation include:

* Policy Analysis: Examining specific policies implemented to reduce CO2 emissions or promote energy efficiency.
* Technological Advancements: Assessing the adoption of cleaner technologies and their impact on emissions.
* Economic Factors: Analysing economic shifts, industrialization patterns, and their correlation with energy consumption and emissions.
* Resource Availability: Investigating the availability and development of domestic energy resources.

## 5. Feature Engineering

The features were selected based on 20 relevant parameters covering per capita electricity cost, the percentage of renewable energy in total energy consumption, the cost of electricity imports and exports, and both domestic and foreign debt levels. These indicators are directly related to the project's objective of investigating the relationship between renewable energy adoption and economic recovery.

* Per Capita Electricity Cost: This indicator helps in understanding the economic burden of energy costs on individuals and its impact on overall economic growth.
* Renewable Energy Percentage: This reflects the extent to which renewable energy is being integrated into the national energy mix, which is central to the study's focus.
* Electricity Trade Costs: The costs associated with importing and exporting electricity are crucial for assessing the economic viability of energy strategies in developing nations.

## 6. Data Transformation

Feature scaling was performed on the numerical features of the dataset to ensure that no single feature dominated others due to its scale. This process is essential as many machine learning algorithms are sensitive to the scale of the input features. In this particular study, the standardization procedure was employed. This procedure transforms all features to have a mean of zero and a standard deviation of one, effectively bringing all features onto the same scale. This transformation was achieved using the StandardScaler function, a component of the scikit-learn library. The application of feature scaling via standardization ensures that all features contribute equally to the final prediction.

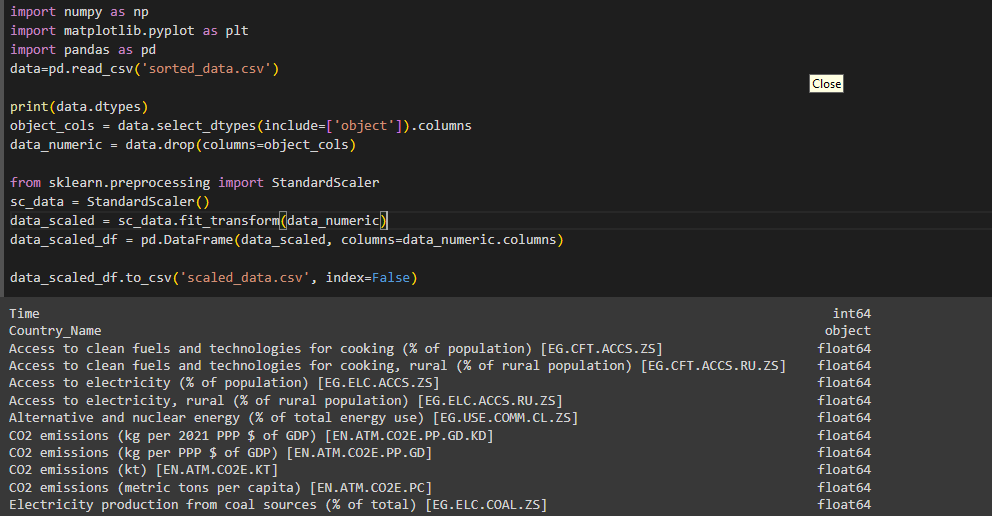


Figure 3. Feature Scaling

While for the categorical variable that is Country Name and Time in this context was one hot encoded to make the feature into a numerical format that can be used by algorithms. The Country names and time will now be converted to column matrix with unique series of vector assigned to each country which will allow machine learning algorithm to identify that particular Country and time.

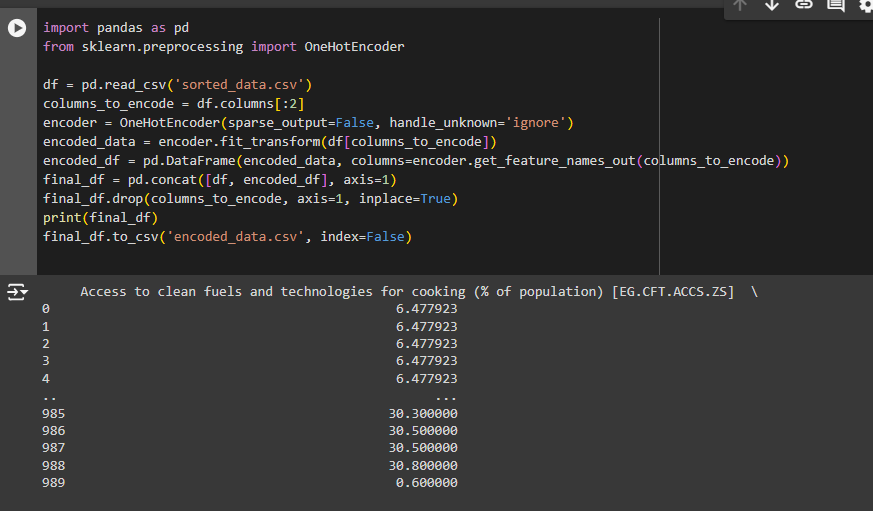


Figure 4.One hot encoding

## 7. Model Exploration

### 7.1 Model Selection

To analyze the impact of renewable energy adoption on economic recovery and social protection in developing nations, several machine learning algorithms will be implemented. These algorithms are suitable for handling complex, non-linear relationships and making accurate predictions based on diverse datasets.

1. Support Vector Regression (SVR):

Support Vector Regression uses an insensitive tube to minimize error shown in equation 1[10]. This tube around the regression line has a width of Epsilon, measured vertically. Points falling inside the tube are considered to be within acceptable error range. Points outside the tube can result in different outcomes as they are not considered in the model.

(1)

Where, is slack variable below the tube.

is slack variable above the tube.

For a non-linear dataset a kernel trick known as gaussian radial based function (rbf)[11] was employed in order to linearize the data is shown by equation 2[11]

Purpose: SVR is used for modeling and predicting continuous outcomes where the relationship between variables may be non-linear. It finds the best-fitting hyperplane that minimizes prediction errors, making it suitable for complex economic data where traditional linear models might be inadequate.

Use Case: SVR will be used to predict economic indicators such as GDP growth or employment rates based on variables like renewable energy adoption rates, energy prices, and policy changes.

2. K-Nearest Neighbor (KNN):

The K-NN[10] algorithm is used for classification and regression tasks. It works by finding the k closest data points to a new input and predicting its label based on the majority class or average value of its neighbors. New data point is recognized based on its proximity to existing data. In this algorithm, k number of neighbors were chosen based on Euclidean distance.

Purpose: KNN is a non-parametric algorithm used for classification and regression tasks. It predicts the value of a new data point based on the average of its nearest neighbors, making it useful for datasets where similar patterns or clusters exist.

Use Case: KNN will identify patterns in economic recovery by comparing countries with similar renewable energy adoption levels. It can also be used to forecast economic outcomes based on historical data from comparable nations.

3. Random Forest:

Random Forest Regression[10] is ensemble form of learning that combines the power of multiple decision trees to generate more accurate and robust predictions. By combining the outputs of many decision trees, Random Forest Regression reduces the risk of overfitting and improves the stability and reliability of your models. In this method, K data points from the training set are selected in order to build the decision trees and process is repeated to create multiple decision trees. Now these multiple regression decision trees are used to predict the value of y for a new data point and taking the average of the predictions from all the trees, improving prediction accuracy.

Purpose: Random Forest is an ensemble learning method that combines multiple decision trees to improve prediction accuracy and reduce overfitting. It is effective for handling large datasets with many features, providing robust predictions by considering various possible scenarios.

Use Case: Random Forest will model complex interactions between multiple economic indicators and renewable energy metrics. It will help identify the most significant factors influencing economic recovery and provide insights into the potential outcomes of different policy decisions.

## 7.2 Model Training

Scikit Learn Library is used in order to build all the models in the research. For the SVR model, rbf function was employed to linearize the non-linear data set and SVR function was induced from the scikit learn library. For the RF model building, the estimators or no of decision trees are set to 20 while the limiting condition is made to not split subset with data less than 5. Then, RandomForestRegressor function is induced for the model for training the model with training sets. For the KNN model, the value of K is taken as 20 which are said neighbors which we want to evaluate. After this Euclidean geometry is used to calculate the distance between the predicted point and neighbors. To build the model, kNeighborRegression Function is induced through the sci-kit library and the model is trained using training sets.

### 7.3 Model Evaluation

Adjusted R2 value method was employed in order to evaluate the machine learning models. Adjusted R-squared is a statistical measure that compares the fit of a model to the average of all possible models. It is a modified version of R-squared that accounts for the number of predictors in the model. A higher adjusted R-squared indicates a better fit of the model to the data[13]. It also penalizes the addition of unnecessary variables to the model. Adjusted R-squared values can range from 0 to 1, with values closer to 1 indicating a better fit. As the model contains more independent variables, it affects the totals sum of squares in regression analysis. Adjusted r squared accounts for unnecessary variables and avoids artificially inflating R square. The governing equation in calculating adjusted R2 value depends upon the R- squared value which is calculated as shown in equation 3[13]

(3)

where, represents the square of error made between regression line and the predicted dataset.

represents the square of the error made between the actual data and average data of the dataset.

Adjusted R2 value method can be expressed mathematically by equation 4[13].

(4)

where, K is the number of independent variables

N is the sample size

Therefore, Adjusted R2 value method ensures adding variables only brings substantial improvement to the model.

### 7.4 Code Implementation