Prediction of CO2 Emissions Using Global Sustainability Data

Group 09

Jahnavi Mishra

Akshat Jain

(857) 492 8422

(857) 334 6863

mishra.j@northeastern.edu

jain.akshat2.edu

Percentage of Effort Contributed by Student 1: 50%

Percentage of Effort Contributed by Student 2: 50%

Signature of Student 1: Jahnavi Mishra

Signature of Student 2: Akshat Jain

Submission Date: 12th April, 2024

Problem Setting:

Sustainable energy is defined as power sources that have a low environmental impact and are continuously available. This is crucial in the fight against climate change. Renewably abundant resources—such as solar, wind, hydro, and geothermal energy—harness nature's bounty without diminishing limited supplies, making them prime examples of sustainability. Sustainable energy, as opposed to fossil fuels, reduces greenhouse gas emissions, hence reducing the growing threat of global warming. Innovations in technology increase price and efficiency, which encourages the shift to greener options. Adopting sustainable energy promotes economic growth, energy independence, and ecosystem protection. A global commitment to creating a sustainable and environmentally friendly energy landscape for future generations is emerging as countries give priority to renewable efforts. In recent years, due to increase in climate change, it has become increasingly important to study and make advancements in the field of sustainable energy. Through this project, we aim to contribute to this goal.

Problem Definition:

We aim explore a detailed dataset encompassing sustainable energy metrics and relevant factors worldwide spanning the years 2000 to 2020. We will analyze critical elements like electricity accessibility, renewable energy usage, carbon emissions, energy efficiency, financial investments, and economic development. Additionally we can conduct cross-country comparisons, monitor advancements towards achieving Sustainable Development Goal 7, and acquire in-depth understanding of global energy consumption trends throughout the specified period.
This is a supervised learning problem.

Some possible questions that we aim to find the answers for are:

- Predict future energy usage, aid planning, and track SDG 7 progress.
- Forecast CO2 emissions, support climate strategies.
- Categorize regions for infrastructure development, understand sustainable energy's role.
- Monitor progress towards Goal 7, evaluate policy impact.
- Analyze access, density, and growth for equitable distribution.
- Identify intensive areas for environmental impact reduction.
- Identify regions for green investments based on capacity.
- Guide investors towards sustainable opportunities.

Data Source:

Global Data on Sustainable Energy (2000-2020) has been taken from Kaggle.
Link: https://www.kaggle.com/datasets/anshtanwar/global-data-on-sustainable-energy

Dataset Description:

This dataset has 21 columns and 3650 records.

The key features are as follows:

• Entity: The name of the country or region for which the data is reported.
• Year: The year for which the data is reported, ranging from 2000 to 2020. • Access to electricity (% of population): The percentage of population with access to electricity.
• Access to clean fuels for cooking (% of population): The percentage of the population with primary reliance on clean fuels.
• Renewable-electricity-generating-capacity-per-capita: Installed Renewable energy capacity per person • Financial flows to developing countries (US $): Aid and assistance from developed countries for clean energy projects.
• Renewable energy share in total final energy consumption (%): Percentage of renewable energy in final energy consumption.
• Electricity from fossil fuels (TWh): Electricity generated from fossil fuels (coal, oil, gas) in terawatt-hours.
• Electricity from nuclear (TWh): Electricity generated from nuclear power in terawatthours.
• Electricity from renewables (TWh): Electricity generated from renewable sources (hydro, solar, wind, etc.) in terawatt-hours.
• Low-carbon electricity (% electricity): Percentage of electricity from low-carbon sources (nuclear and renewables).
• Primary energy consumption per capita (kWh/person): Energy consumption per person in kilowatt-hours.
• Energy intensity level of primary energy (MJ/$2011 PPP GDP): Energy use per unit of GDP at purchasing power parity.
• Value_co2_emissions (metric tons per capita): Carbon dioxide emissions per person in metric tons.

• Renewables (% equivalent primary energy): Equivalent primary energy that is derived from renewable sources.
• GDP growth (annual %): Annual GDP growth rate based on constant local currency.
• GDP per capita: Gross domestic product per person.
• Density (P/Km2): Population density in persons per square kilometer.
• Land Area (Km2): Total land area in square kilometers.
• Latitude: Latitude of the country's centroid in decimal degrees.
• Longitude: Longitude of the country's centroid in decimal degrees

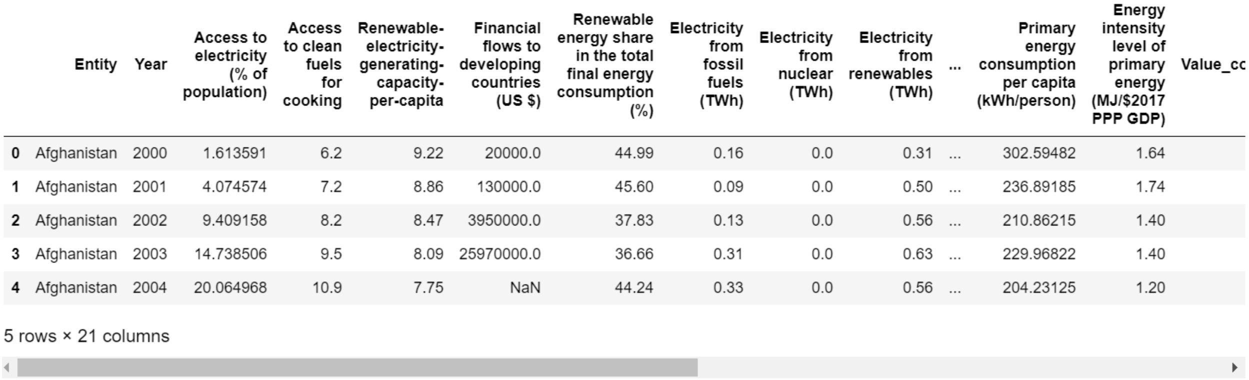
Data Exploration and Data Processing

1. Data Loading and Initial Inspection

We loaded the dataset, “global-data-on-sustainable-energy.csv”. Initial inspection helps to verify that the data has been loaded correctly and to identify any potential issues or inconsistencies.

We renamed a few columns for avoid key error due to escape sequences.

We checked the data types and the data statistics like mean, std, quartiles for all the columns.



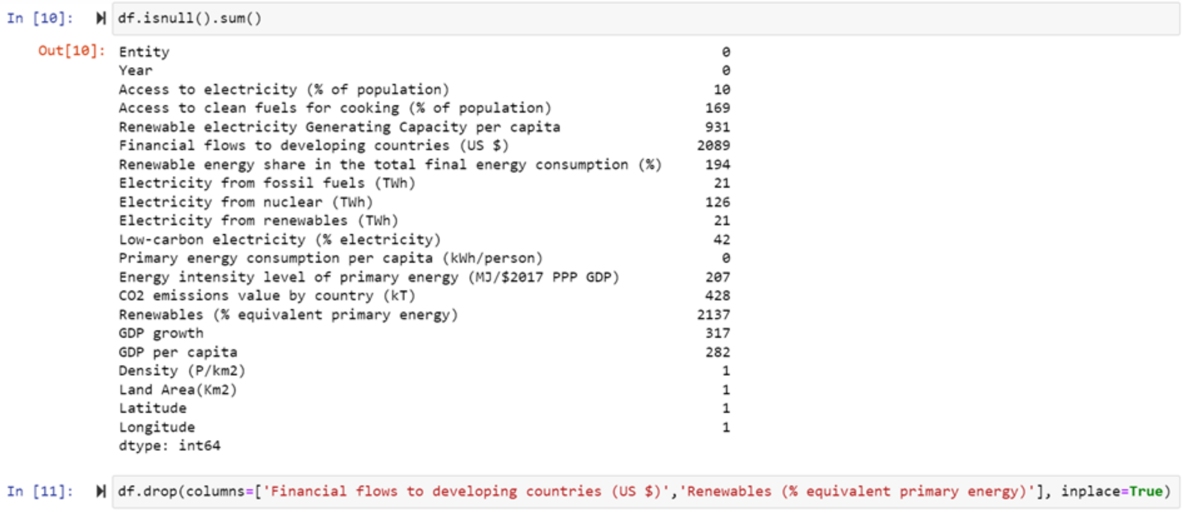
	Entity	Year	Access to electricity (% of population)	Access to clean fuels for cooking	Renewable-electricity-generating-capacity-per-capita	Financial flows to developing countries (US $)	Renewable energy share in the total final energy consumption (%)	Electricity from fossil fuels (TWh)	Electricity from nuclear (TWh)	Electricity from renewables (TWh)	...	Primary energy consumption per capita (kWh/person)	Energy intensity level of primary energy (MJ/$2017 PPP GDP)	Value_cc
0	Afghanistan	2000	1.613591	6.2	9.22	20000.0	44.99	0.16	0.0	0.31	...	302.59482	1.64	
1	Afghanistan	2001	4.074574	7.2	8.86	130000.0	45.60	0.09	0.0	0.50	...	236.89185	1.74	
2	Afghanistan	2002	9.409158	8.2	8.47	3950000.0	37.83	0.13	0.0	0.56	...	210.86215	1.40	
3	Afghanistan	2003	14.738506	9.5	8.09	25970000.0	36.66	0.31	0.0	0.63	...	229.96822	1.40	
4	Afghanistan	2004	20.064968	10.9	7.75	NaN	44.24	0.33	0.0	0.56	...	204.23125	1.20	

5 rows × 21 columns

2. Handling Missing Values

We checked for the missing values and dropped the 2 columns namely 'Financial flows to developing countries (US $)' and 'Renewables (% equivalent primary energy)' that had 2/3 of the values missing.

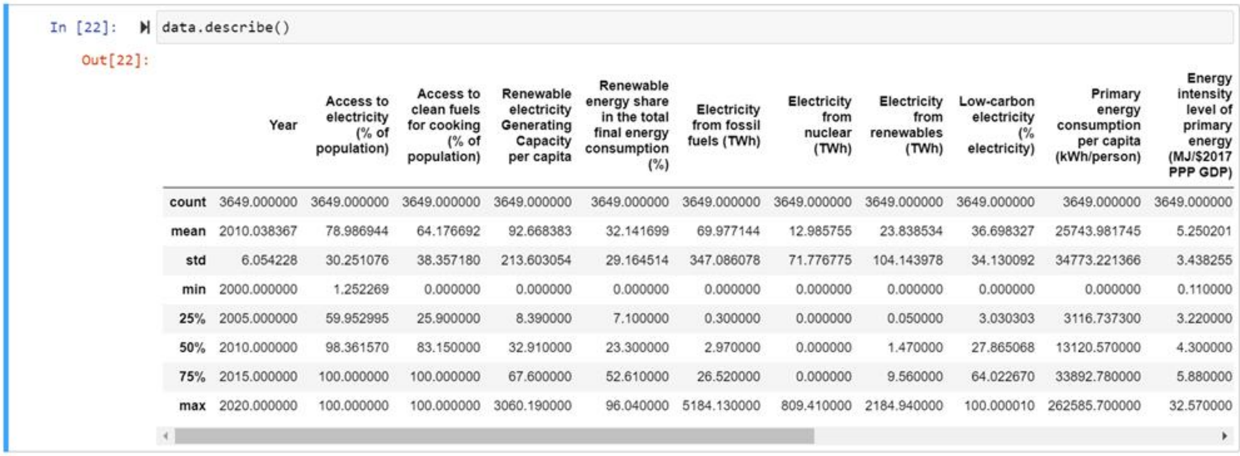
For the rest of the numeric values, we replaced the missing values with the median of the column.



```
In [10]:  df.isnull().sum()

Out[10]:  Entity                                                          0
          Year                                                            0
          Access to electricity (% of population)                        10
          Access to clean fuels for cooking (% of population)           169
          Renewable electricity Generating Capacity per capita          931
          Financial flows to developing countries (US $)               2089
          Renewable energy share in the total final energy consumption (%)   194
          Electricity from fossil fuels (TWh)                            21
          Electricity from nuclear (TWh)                                126
          Electricity from renewables (TWh)                              21
          Low-carbon electricity (% electricity)                         42
          Primary energy consumption per capita (kWh/person)              0
          Energy intensity level of primary energy (MJ/$2017 PPP GDP)   207
          CO2 emissions value by country (kT)                           428
          Renewables (% equivalent primary energy)                     2137
          GDP growth                                                    317
          GDP per capita                                                282
          Density (P/km2)                                                 1
          Land Area(Km2)                                                  1
          Latitude                                                        1
          Longitude                                                       1
          dtype: int64

In [11]:  df.drop(columns=['Financial flows to developing countries (US $)','Renewables (% equivalent primary energy)'], inplace=True)
```

3. Summary Statistics

We've displayed the summary statistics. Summary statistics offer insights into numerical features, such as mean, standard deviation, and quartile values, facilitating a better understanding of the data distribution.

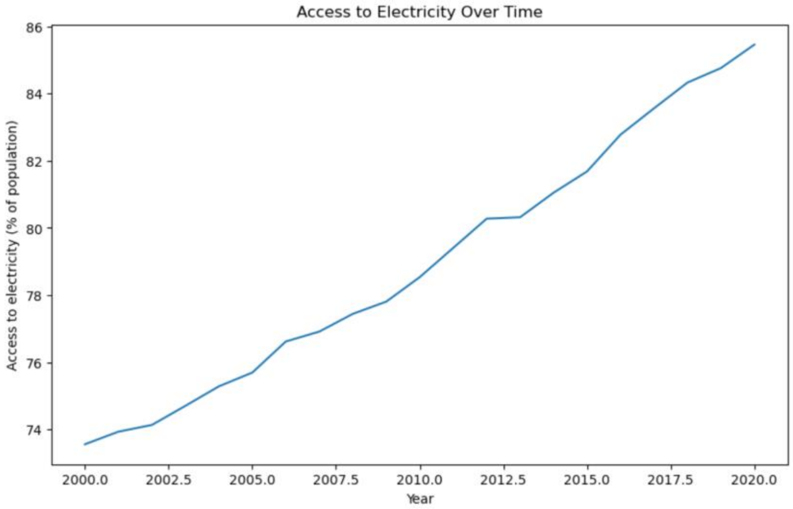


```
In [22]:  data.describe()

Out[22]:
```

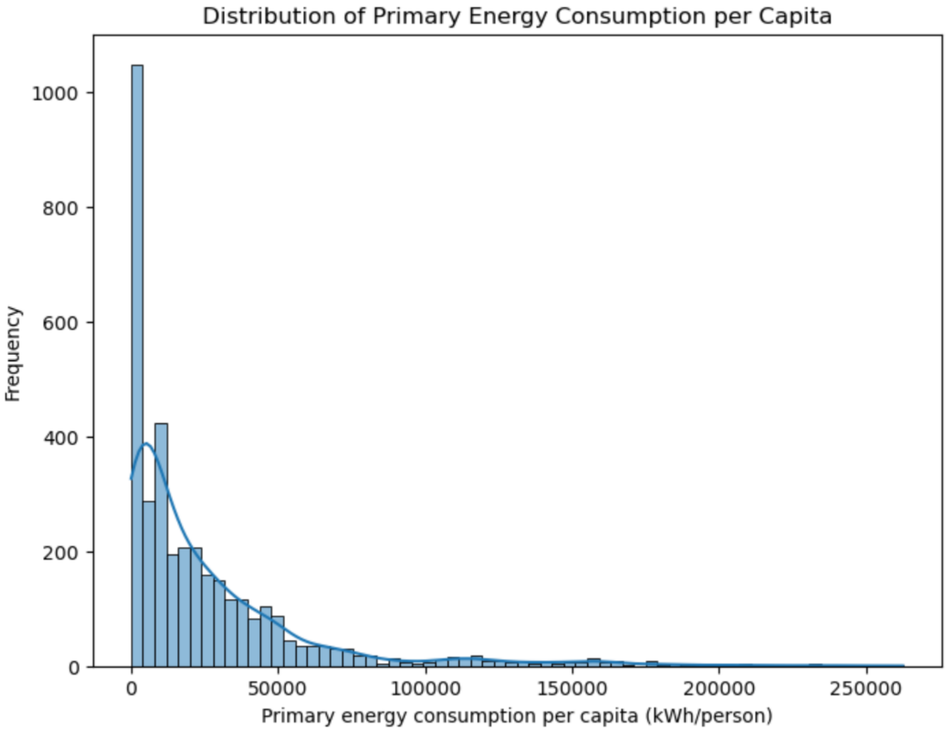
	Year	Access to electricity (% of population)	Access to clean fuels for cooking (% of population)	Renewable electricity Generating Capacity per capita	Renewable energy share in the total final energy consumption (%)	Electricity from fossil fuels (TWh)	Electricity from nuclear (TWh)	Electricity from renewables (TWh)	Low-carbon electricity (% electricity)	Primary energy consumption per capita (kWh/person)	Energy intensity level of primary energy (MJ/$2017 PPP GDP)
count	3649.000000	3649.000000	3649.000000	3649.000000	3649.000000	3649.000000	3649.000000	3649.000000	3649.000000	3649.000000	3649.000000
mean	2010.038367	78.986944	64.176692	92.668383	32.141699	69.977144	12.985755	23.838534	36.698327	25743.981745	5.250201
std	6.054228	30.251076	38.357180	213.603054	29.164514	347.086078	71.776775	104.143978	34.130092	34773.221366	3.438255
min	2000.000000	1.252269	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.110000
25%	2005.000000	59.952995	25.900000	8.390000	7.100000	0.300000	0.000000	0.050000	3.030303	3116.737300	3.220000
50%	2010.000000	98.361570	83.150000	32.910000	23.300000	2.970000	0.000000	1.470000	27.865068	13120.570000	4.300000
75%	2015.000000	100.000000	100.000000	67.600000	52.610000	26.520000	0.000000	9.560000	64.022670	33892.780000	5.880000
max	2020.000000	100.000000	100.000000	3060.190000	96.040000	5184.130000	809.410000	2184.940000	100.000010	262585.700000	32.570000

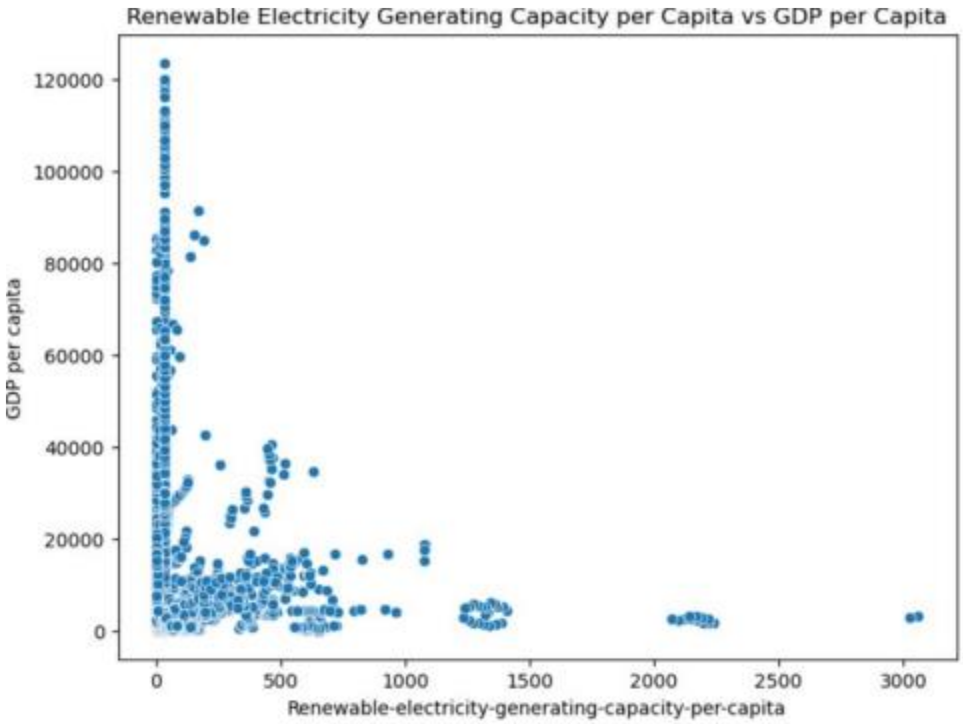
4. Data Visualization

Various graphs were made to explore the dataset to gain insights into the dataset.

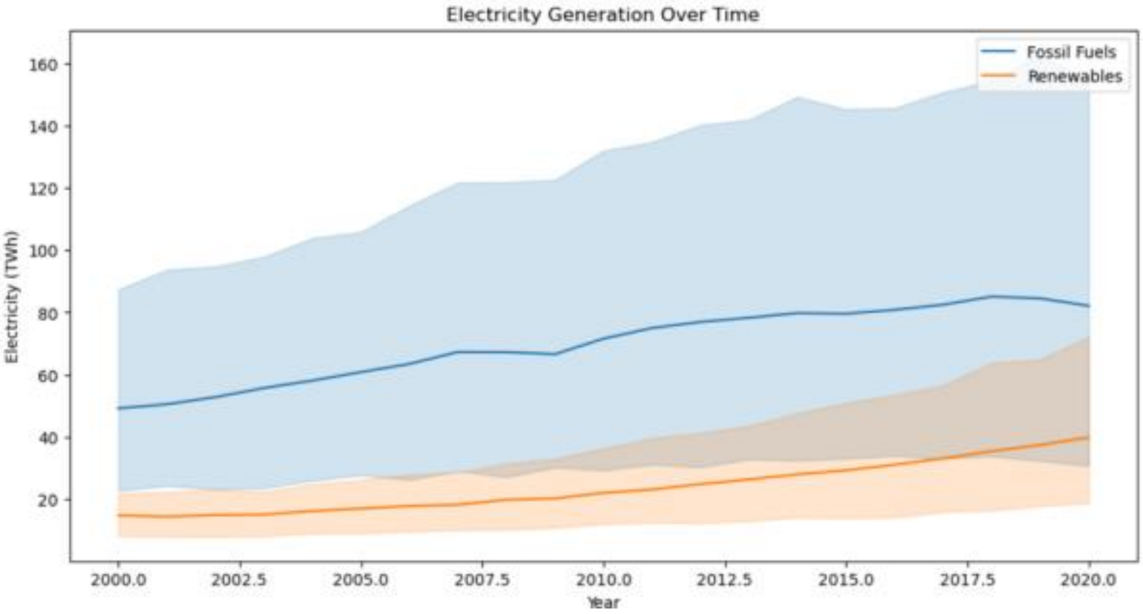


Percentage of Population with access to electricity has grown consistently over the last 2 decades, leading to more CO2 released in the atmosphere.

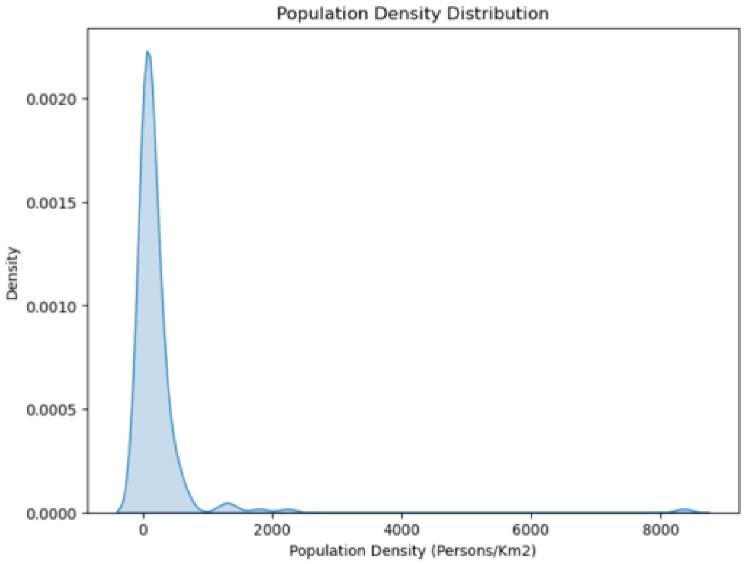


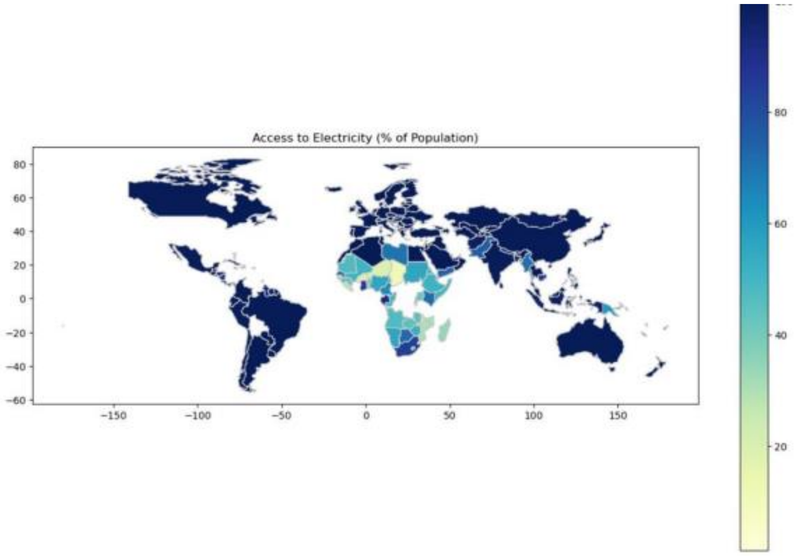


The countries with High GDP’s clearly have low Renewable electricity generation capacity per capita.

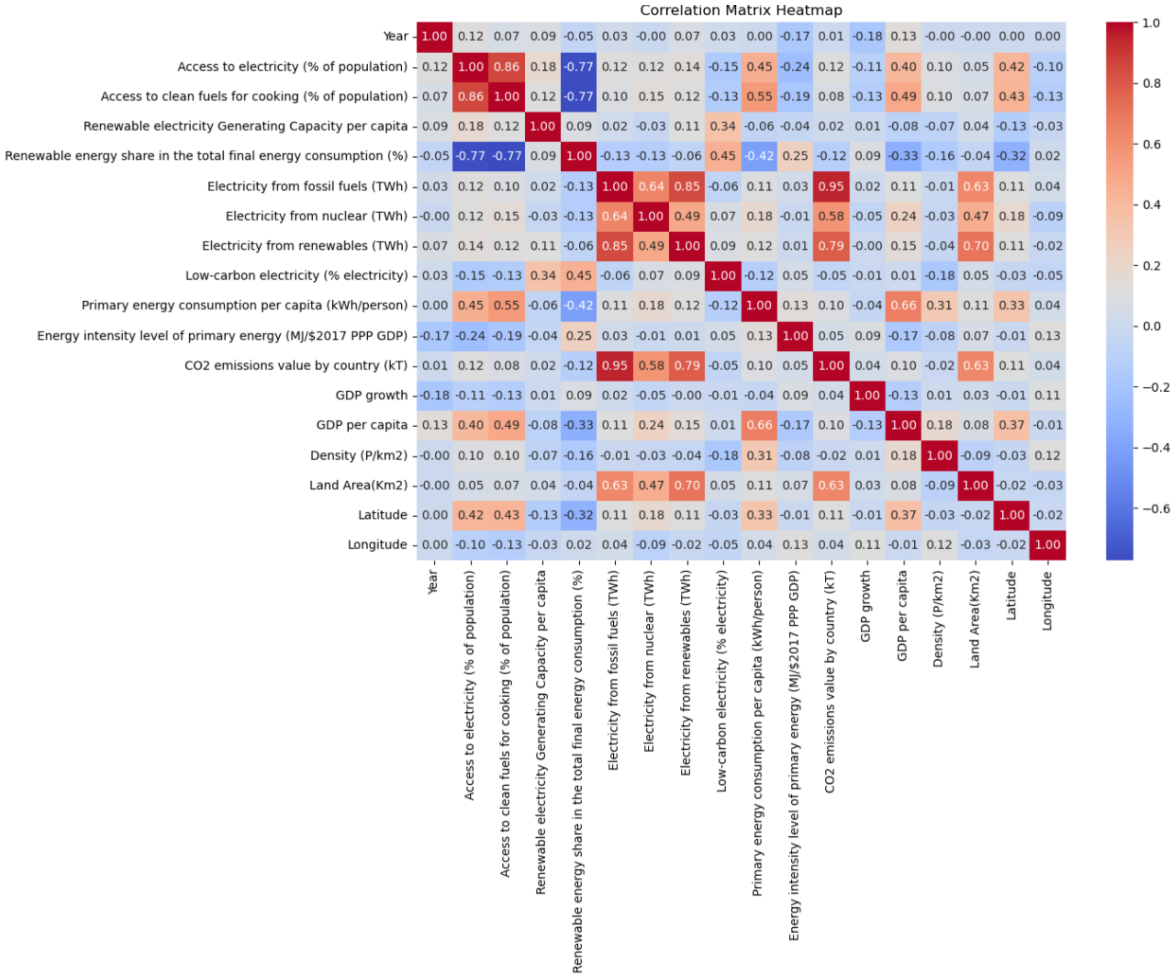


While the production of electricity from sustainable sources has seen consistent increase, it is still nowhere enough to tackle the growing issue caused by fossil fuel consumption.



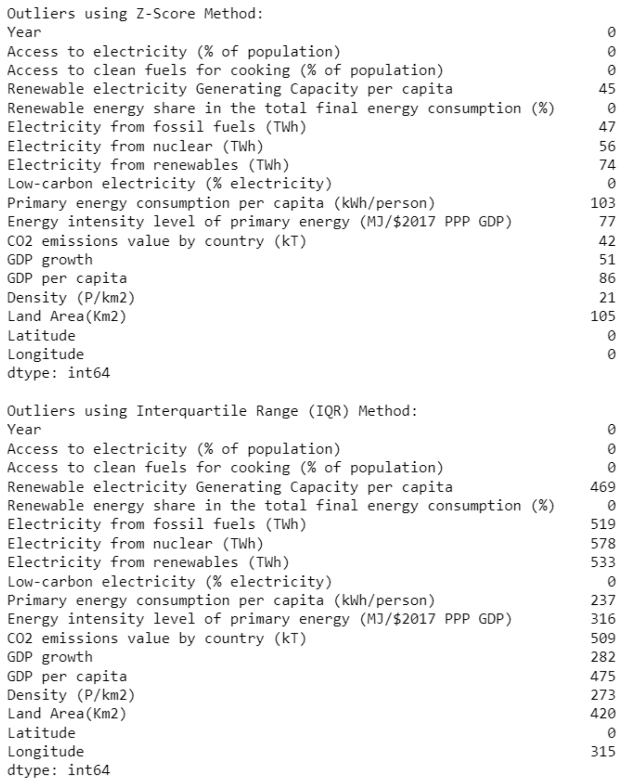


Almost all countries have a large percentage of their population that have access to Electricity (Except of-course the missing ones).



5. Outlier Analysis:

We conducted outlier analysis using Z score method and IQR method. From both we saw that there were some outliers in the dataset.



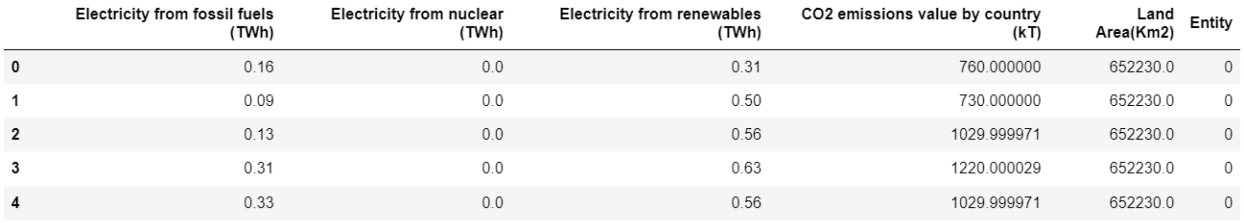
```
Outliers using Z-Score Method:
Year                                                               0
Access to electricity (% of population)                            0
Access to clean fuels for cooking (% of population)                0
Renewable electricity Generating Capacity per capita              45
Renewable energy share in the total final energy consumption (%)   0
Electricity from fossil fuels (TWh)                               47
Electricity from nuclear (TWh)                                    56
Electricity from renewables (TWh)                                 74
Low-carbon electricity (% electricity)                             0
Primary energy consumption per capita (kWh/person)               103
Energy intensity level of primary energy (MJ/$2017 PPP GDP)       77
CO2 emissions value by country (kT)                               42
GDP growth                                                        51
GDP per capita                                                    86
Density (P/km2)                                                   21
Land Area(Km2)                                                   105
Latitude                                                           0
Longitude                                                          0
dtype: int64

Outliers using Interquartile Range (IQR) Method:
Year                                                               0
Access to electricity (% of population)                            0
Access to clean fuels for cooking (% of population)                0
Renewable electricity Generating Capacity per capita             469
Renewable energy share in the total final energy consumption (%)   0
Electricity from fossil fuels (TWh)                              519
Electricity from nuclear (TWh)                                   578
Electricity from renewables (TWh)                                533
Low-carbon electricity (% electricity)                             0
Primary energy consumption per capita (kWh/person)               237
Energy intensity level of primary energy (MJ/$2017 PPP GDP)      316
CO2 emissions value by country (kT)                              509
GDP growth                                                       282
GDP per capita                                                   475
Density (P/km2)                                                  273
Land Area(Km2)                                                   420
Latitude                                                           0
Longitude                                                        315
dtype: int64
```

Since the number of outliers is very less compared to the total number of records, we decided to keep the outliers.

Data Mining Tasks:

1. Label Encoding

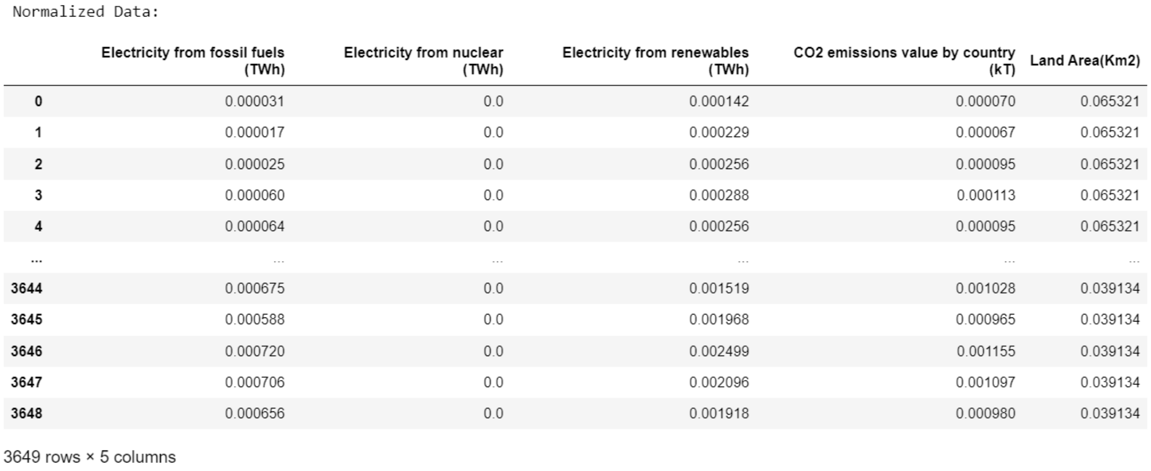
Label encoded the Entity variable which is a nominal variable using LabelEncoder() so that it could be used for further processing.



	Electricity from fossil fuels (TWh)	Electricity from nuclear (TWh)	Electricity from renewables (TWh)	CO2 emissions value by country (kT)	Land Area(Km2)	Entity
0	0.16	0.0	0.31	760.000000	652230.0	0
1	0.09	0.0	0.50	730.000000	652230.0	0
2	0.13	0.0	0.56	1029.999971	652230.0	0
3	0.31	0.0	0.63	1220.000029	652230.0	0
4	0.33	0.0	0.56	1029.999971	652230.0	0

2. Normalization

Using min-max scaler normalized the whole dataset, between a value from 0 to 1 so that all the variables are in same scale, and no variable has more impact that the others in terms of scale.



Normalized Data:

	Electricity from fossil fuels (TWh)	Electricity from nuclear (TWh)	Electricity from renewables (TWh)	CO2 emissions value by country (kT)	Land Area(Km2)
0	0.000031	0.0	0.000142	0.000070	0.065321
1	0.000017	0.0	0.000229	0.000067	0.065321
2	0.000025	0.0	0.000256	0.000095	0.065321
3	0.000060	0.0	0.000288	0.000113	0.065321
4	0.000064	0.0	0.000256	0.000095	0.065321
...	...	...	...	...	...
3644	0.000675	0.0	0.001519	0.001028	0.039134
3645	0.000588	0.0	0.001968	0.000965	0.039134
3646	0.000720	0.0	0.002499	0.001155	0.039134
3647	0.000706	0.0	0.002096	0.001097	0.039134
3648	0.000656	0.0	0.001918	0.000980	0.039134

3649 rows × 5 columns

3. Dimensionality Reduction: Variable Selection

- Using Correlation Analysis:

Calculated pearson correlation of all the numerical columns with the target variable 'CO2 emissions value by country (kT)'

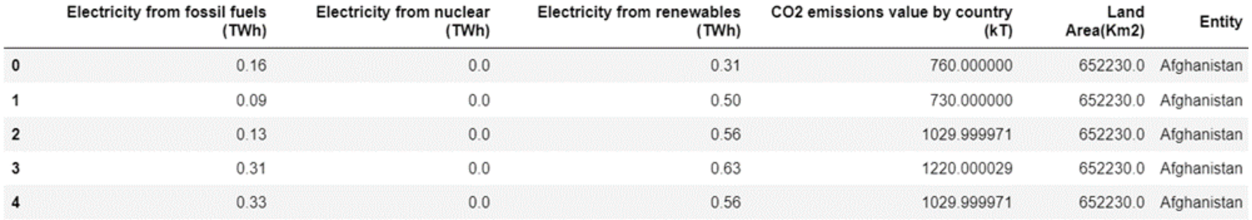
Here are the top 5 most positively and top 5 most negatively correlated features to the target variable.

```
 Top 5 Most Positively Correlated to the Target Variable:
Electricity from fossil fuels (TWh)        0.948371
Electricity from renewables (TWh)          0.785686
Land Area(Km2)                             0.634150
Electricity from nuclear (TWh)             0.583052
Access to electricity (% of population)    0.115784
Name: CO2 emissions value by country (kT), dtype: float64



Top 5 Most Negatively Correlated to the Target Variable:
Low-carbon electricity (% electricity)                 -0.046170
Density (P/km2)                                        -0.017081
```

```
Year                                                      0.006273
Renewable electricity Generating Capacity per capita     0.019933
Longitude                                                 0.039214
Name: CO2 emissions value by country (kT), dtype: float64
```

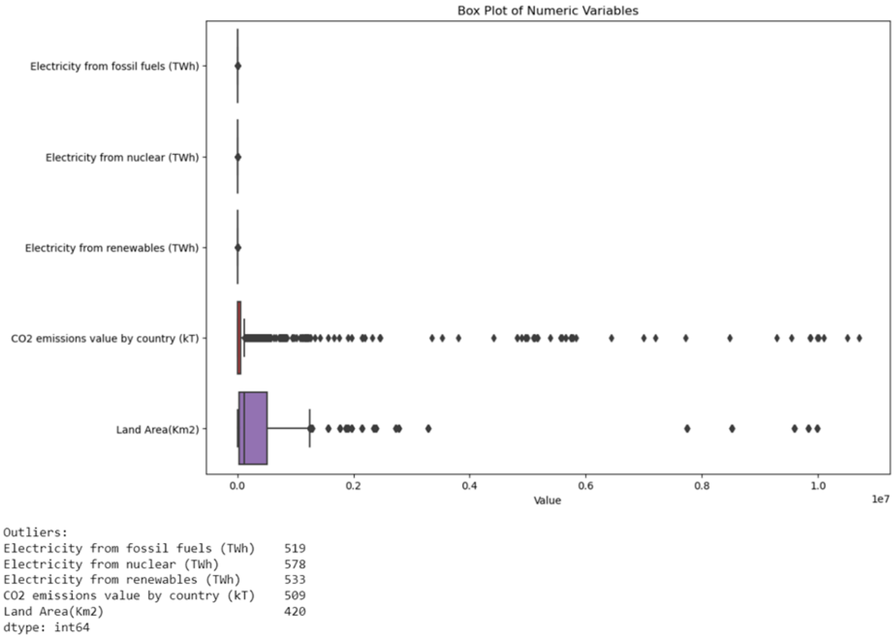
Based on this, we removed the elements which have a correlation < 0.5 with the target variable.

Hence we are left with the following features:



	Electricity from fossil fuels (TWh)	Electricity from nuclear (TWh)	Electricity from renewables (TWh)	CO2 emissions value by country (kT)	Land Area(Km2)	Entity
0	0.16	0.0	0.31	760.000000	652230.0	Afghanistan
1	0.09	0.0	0.50	730.000000	652230.0	Afghanistan
2	0.13	0.0	0.56	1029.999971	652230.0	Afghanistan
3	0.31	0.0	0.63	1220.000029	652230.0	Afghanistan
4	0.33	0.0	0.56	1029.999971	652230.0	Afghanistan

We also checked the outliers for the new dataframe:

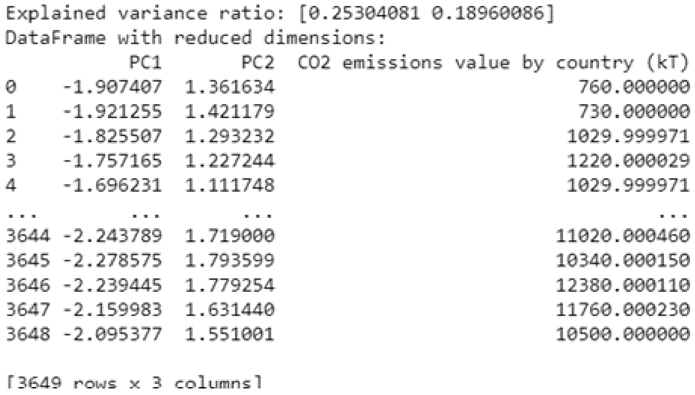


```
Outliers:
Electricity from fossil fuels (TWh)    519
Electricity from nuclear (TWh)         578
Electricity from renewables (TWh)      533
CO2 emissions value by country (kT)    509
Land Area(Km2)                         420
dtype: int64
```

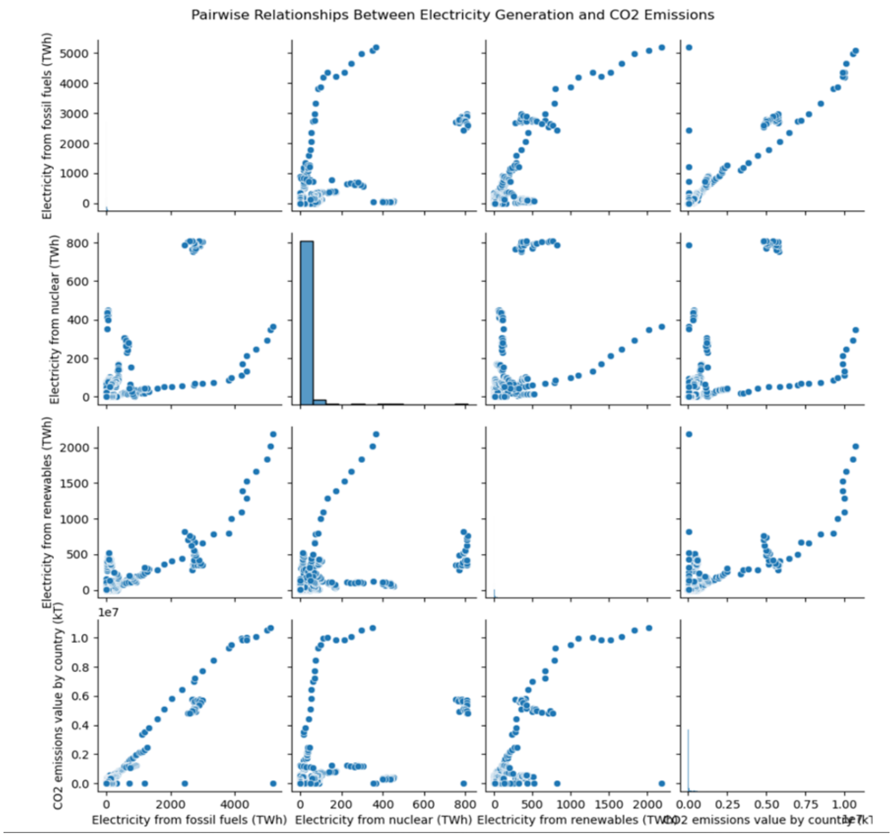
- Using Principal Component Analysis:

Applied PCA with two components (n_components=2) to the scaled numerical features. This reduces the dimensionality of the data while retaining most of its variance. We saw that the first 2 PC's do not capture much of the variance. Thus, we cannot remove any variable using PCA.

Then we conducted correlation analysis with our target variable and removed the columns that were not correlated to the target variable (CO2 emissions)



```
Explained variance ratio: [0.25304081 0.18960086]
DataFrame with reduced dimensions:
           PC1       PC2  CO2 emissions value by country (kT)
0    -1.907407  1.361634                            760.000000
1    -1.921255  1.421179                            730.000000
2    -1.825507  1.293232                           1029.999971
3    -1.757165  1.227244                           1220.000029
4    -1.696231  1.111748                           1029.999971
...        ...       ...                                   ...
3644 -2.243789  1.719000                          11020.000460
3645 -2.278575  1.793599                          10340.000150
3646 -2.239445  1.779254                          12380.000110
3647 -2.159983  1.631440                          11760.000230
3648 -2.095377  1.551001                          10500.000000

[3649 rows x 3 columns]
```



Data Mining Models:

Our target variable is CO2 emissions value by country (kT) which is a numerical column. All the predictors are numerical columns. On the basis of this, we will be experimenting on the following regression models.

1. Linear Regression:

Linear regression is simple and interpretable. It provides insights into the relationships between variables, as coefficients represent the change in the target variable for a one-unit change in the predictor variable. Additionally, inference on coefficients can provide statistical significance tests. However, linear regression assumes a linear relationship between variables, which might not always be the case. It can't capture complex nonlinear relationships in the data. Also, it's sensitive to outliers, multicollinearity, and heteroscedasticity.

Given the dataset's mix of numerical features like "Access to electricity (% of population)" and "GDP per capita" alongside the target variable "Value_co2_emissions (metric tons per capita)," linear regression can offer an initial understanding of the direct relationships between these factors and CO2 emissions. It would assume a linear relationship between each predictor variable and CO2 emissions.

2. Random Forest Regression:

Random forests handle nonlinear relationships well and are robust to overfitting due to the averaging effect of multiple trees. They can handle both numerical and categorical data without the need for extensive preprocessing. Additionally, they provide feature importance scores, which can help in understanding the relative importance of different variables. However, random forests can be computationally expensive and memory-intensive, especially for large datasets. They might not perform well on very sparse data or data with high dimensionality. Furthermore, random forests can struggle with extrapolation beyond the range of the training data.

With features like "Access to clean fuels for cooking (% of population)" and "Renewable energy share in total final energy consumption (%)," random forest regression can capture complex interactions between these factors and CO2 emissions. It can handle both numerical and categorical features effectively, providing insights into the importance of various predictors.

3. Gradient Boosting Regression (e.g., XGBoost, LightGBM):

Gradient boosting often provides superior predictive performance compared to random forests, especially when dealing with heterogeneous data. It handles missing data well and is less prone to overfitting. Moreover, it can capture complex interactions between variables.

However, gradient boosting can be sensitive to hyperparameters and requires careful tuning. Training time can be longer compared to random forests due to the sequential nature of tree building. Additionally, gradient boosting models can be harder to interpret compared to linear models or random forests.

Features such as "Electricity from renewables (TWh)" and "Financial flows to developing countries (US $)" might exhibit nonlinear relationships with CO2 emissions. Gradient boosting regression models can capture these nonlinearities, offering improved predictive performance compared to simpler models like linear regression.

4. K Nearest Neighbor Regression (KNN):

KNN is a non-parametric model, meaning it makes no assumptions about the underlying distribution of the data. This flexibility enables the model to capture complex patterns in the data without imposing rigid constraints. KNN is particularly effective at capturing local patterns in the data. It tends to perform well when the relationships between features and the target variable are non-linear and vary across different regions of the feature space. It does not assume independence between features, which can be beneficial when dealing with correlated or interdependent features in the dataset.

KNN has fewer hyperparameters to tune compared to some other regression models, simplifying the model selection process and reducing the risk of overfitting.

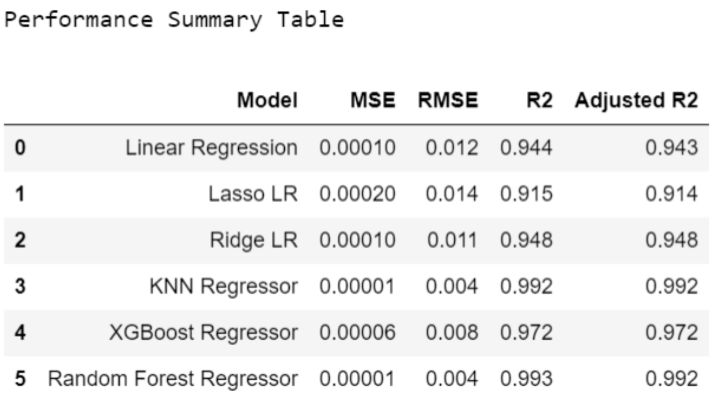
On the basis of outlier analysis that we conducted in the previous milestone, we have quite a few columns with outliers. KNN is relatively robust to outliers since it considers multiple neighbors when making predictions. Outliers may have less influence on the predictions compared to other regression models that rely on fitting a global function to the data.

Conclusion

Each of these models has its strengths and weaknesses, and the choice of model depends on factors such as the nature of the data, the complexity of the relationship between variables, computational resources, and interpretability requirements. It's often beneficial to experiment with multiple models and select the one that provides the best balance of predictive performance, interpretability, and computational efficiency for a particular task.

In summary, each model offers a unique approach to predicting CO2 emissions based on the dataset's features. Linear regression provides a simple baseline, while more complex models like random forests, gradient boosting. KNN offers robustness towards outliers and easier hyperparameter tuning. The choice of model would depend on considerations such as interpretability, computational resources, and the desired balance between predictive performance and model complexity. Experimentation with multiple models would help identify the most suitable approach for your specific prediction task.

Performance Evaluation



Performance Summary Table

	Model	MSE	RMSE	R2	Adjusted R2
0	Linear Regression	0.00010	0.012	0.944	0.943
1	Lasso LR	0.00020	0.014	0.915	0.914
2	Ridge LR	0.00010	0.011	0.948	0.948
3	KNN Regressor	0.00001	0.004	0.992	0.992
4	XGBoost Regressor	0.00006	0.008	0.972	0.972
5	Random Forest Regressor	0.00001	0.004	0.993	0.992

To determine the best model from the provided summary table, we can consider several factors such as:

1. Performance Metrics:

Identify which model has the lowest values for error metrics (MSE, RMSE) and highest values for goodness-of-fit metrics (R2, Adjusted R2).

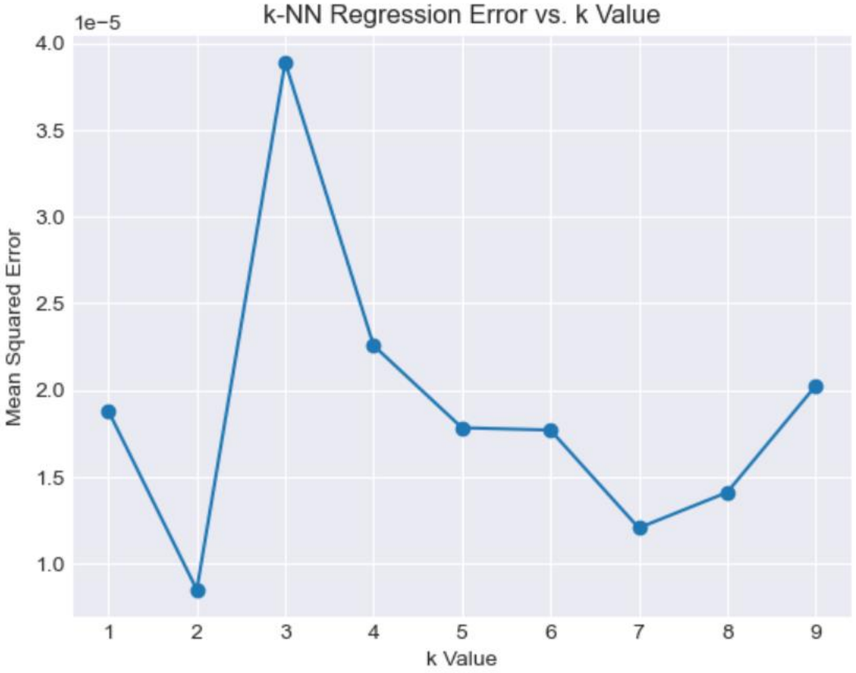
2. Trade-offs:

We need to consider the trade-offs between different metrics. For example, a model with slightly higher error metrics but better interpretability may be preferred over a more complex model with slightly lower error metrics.

3. Domain Knowledge:

Based on the above metrics:

- The KNN Regressor has the lowest MSE and RMSE, indicating the smallest errors in prediction.

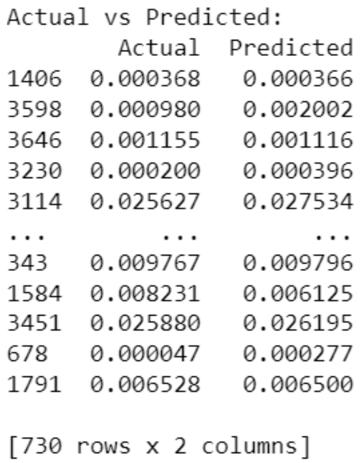


Used k=2 on the basis of this graph.

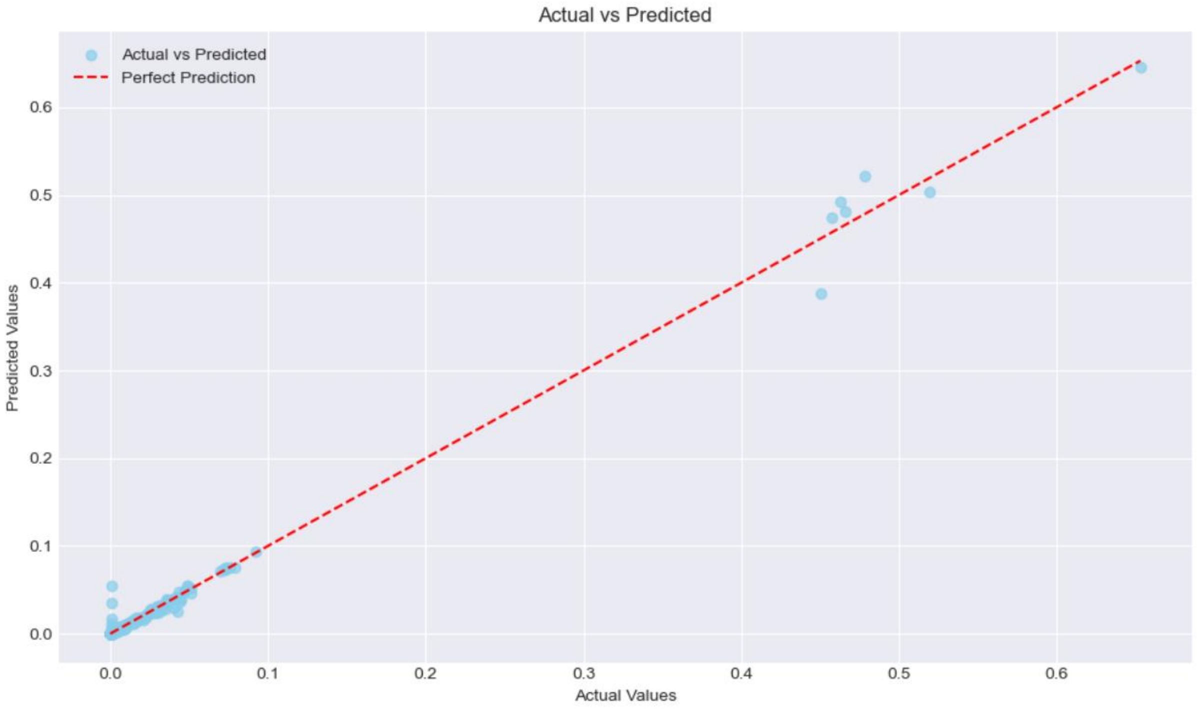
- The Random Forest Regressor has the highest R2 and Adjusted R2, indicating the best goodness of fit. Based on domain knowledge, we have concluded that Random Forest is the best model to predict the CO2 emissions by country. As this is a regression problem, we can not plot ROC curve and confusion matrix.

Project Results

These are the comparisons between the actual and predicted values of random forest model.



```
Actual vs Predicted:
        Actual  Predicted
1406  0.000368   0.000366
3598  0.000980   0.002002
3646  0.001155   0.001116
3230  0.000200   0.000396
3114  0.025627   0.027534
...        ...        ...
343   0.009767   0.009796
1584  0.008231   0.006125
3451  0.025880   0.026195
678   0.000047   0.000277
1791  0.006528   0.006500

[730 rows x 2 columns]
```



Conclusion

Implications for Sustainable Energy Research: The high performance of the Random Forest model suggests that it can effectively contribute to the study and advancement of sustainable energy. By accurately predicting outcomes related to sustainable energy initiatives, the model can assist researchers, policymakers, and stakeholders in making informed decisions to promote renewable energy adoption, combat climate change, and achieve environmental sustainability goals.

In conclusion, the Random Forest model demonstrates strong predictive capabilities and holds promise for supporting efforts aimed at advancing sustainable energy practices and mitigating the impacts of climate change.