UNIVERSITY OF WOLVERHAMPTON

HERALD COLLEGE KATHMANDU

REGRESSION ANALYSIS REPORT

Student ID: 2417739

Student Name: Preeyanshu Singh

Section: L5CG20

Module Leader: Siman Giri

Tutor: Bibek Kanal

Abstract:

Introduction:

This coursework aims at a machine learning model that predicts energy consumption on the bases of various economic and demographic indicators. The course will entitle its relation with the UN Sustainable Development Goal (UNSDG) Goal 7, Affordability and Availability of Clean Energy through analyzing energy consumption-effective factors and promoting improved management of energy consumption. The analysis founded on machine learning seeks to provide insights for the aid of policymakers and businesses in their decision-making process regarding the pattern of energy consumption.

The Energy Consumption Dataset contains a host of economic and demographic features. It is available in CSV format, where a number of inputs impact Energy Consumption. The target variable is Energy Consumption, continuous in nature, which we intend to predict using regression models. Ties between economic and energy use should serve as a basis for formulating sustainable energy policies.

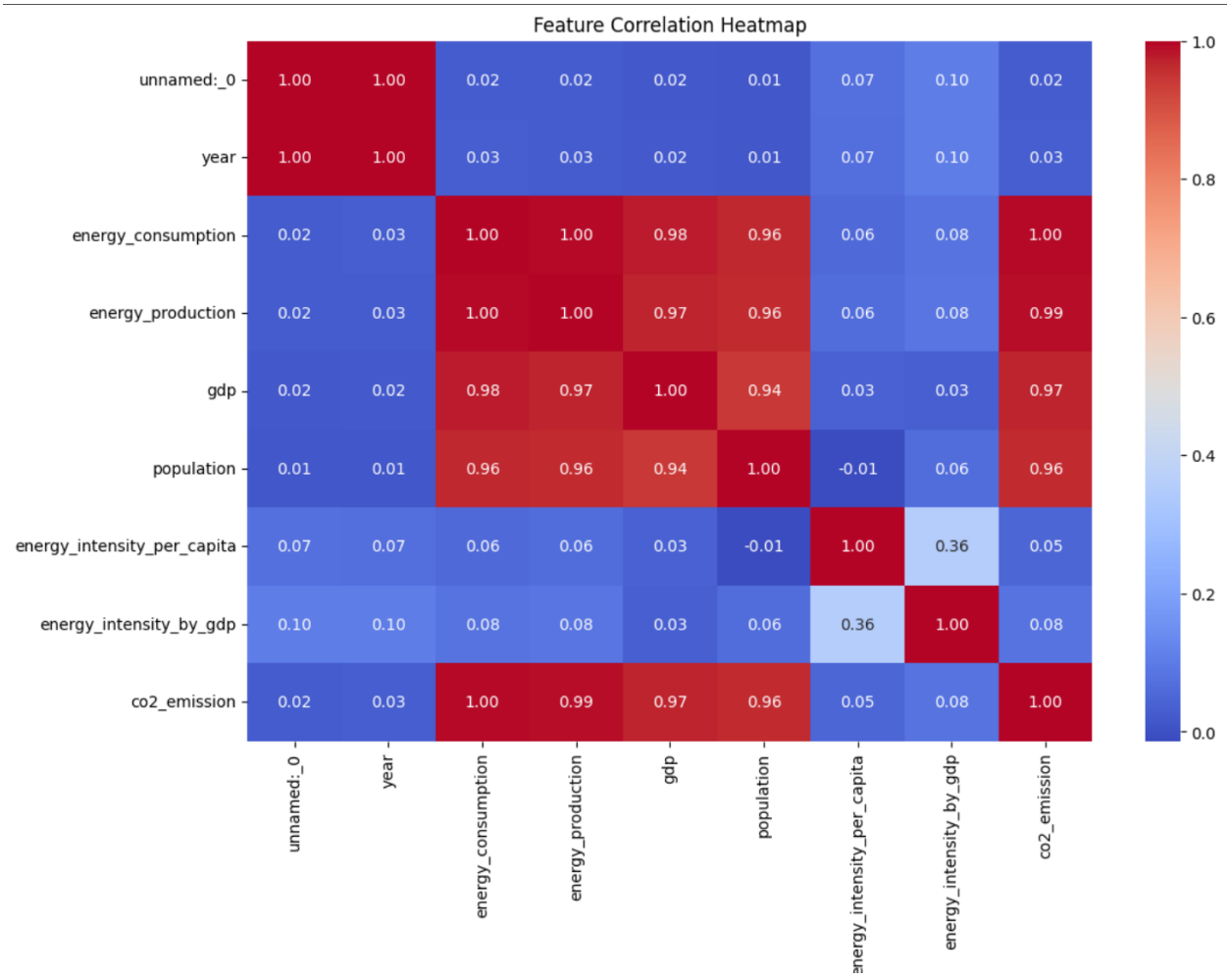
Methodology

Data Preprocessing & Exploratory Data Analysis (EDA):

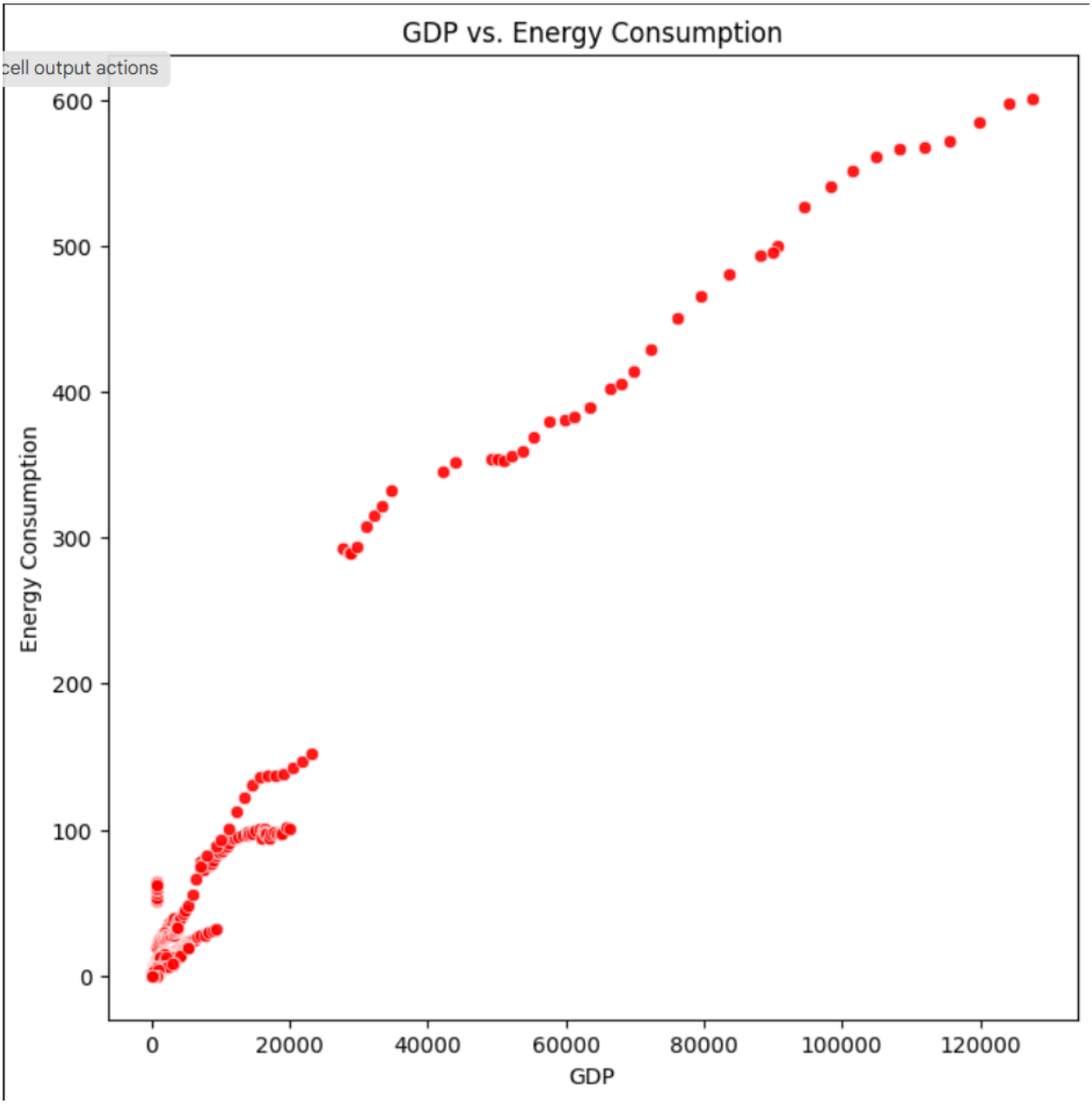
Date cleaning is important step to be done before building the model so various step was carried out in data cleaning stage in which missing values were checked and were handled properly, some of the categorical data were also encoded so it could be used while building models. Column names were standardized, and duplicate records were removed to maintain consistency in the dataset and furthermore, numerical features were normalized using StandardScaler to improve model efficiency.

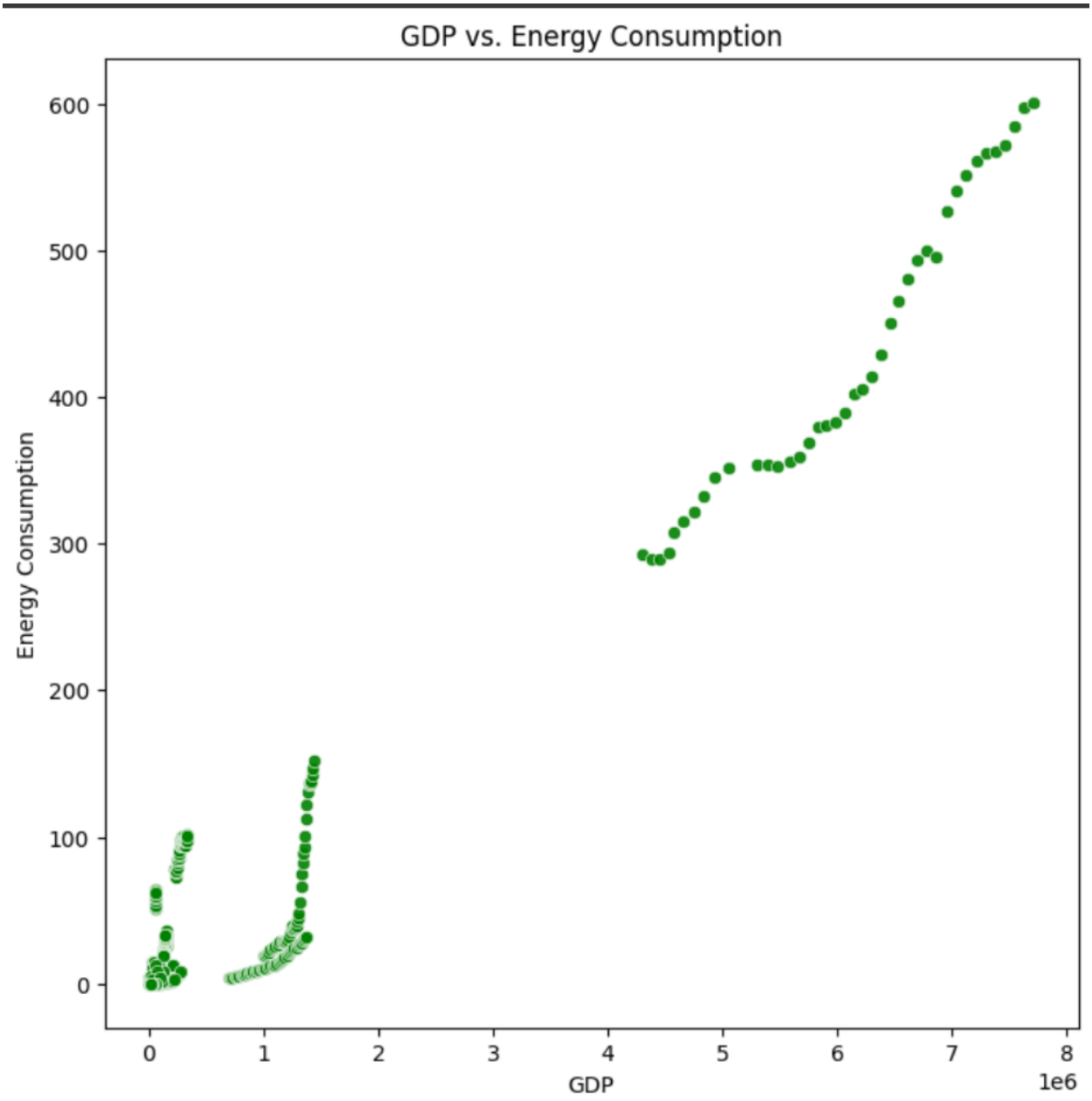
After data cleaning is completed to understand data properly various exploratory data analysis techniques were applied. Some of the visualization plots were used such as:

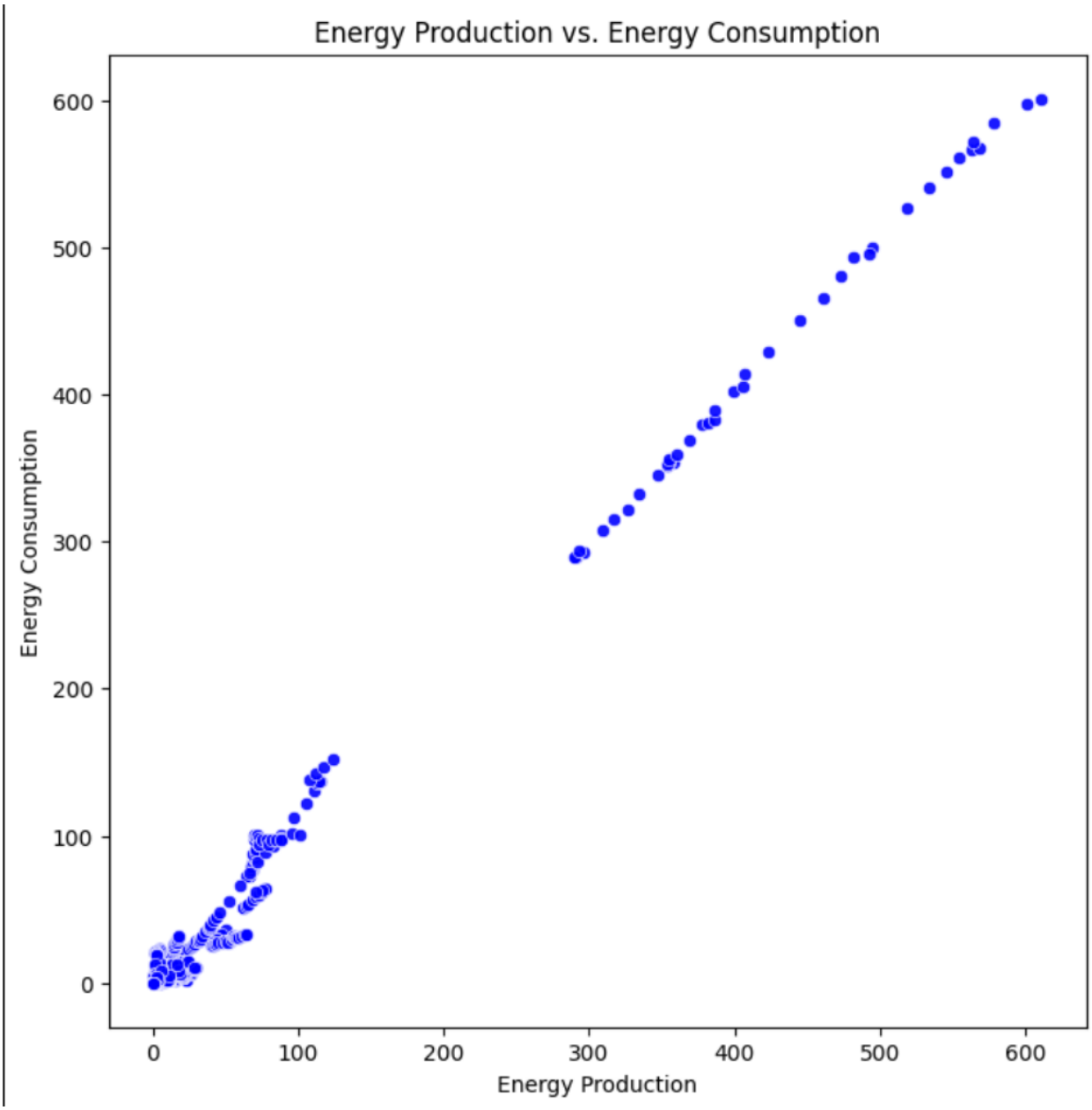
A correlation heatmap was utilized to determine the relationships between different features, revealing how strongly each feature correlates with energy consumption.

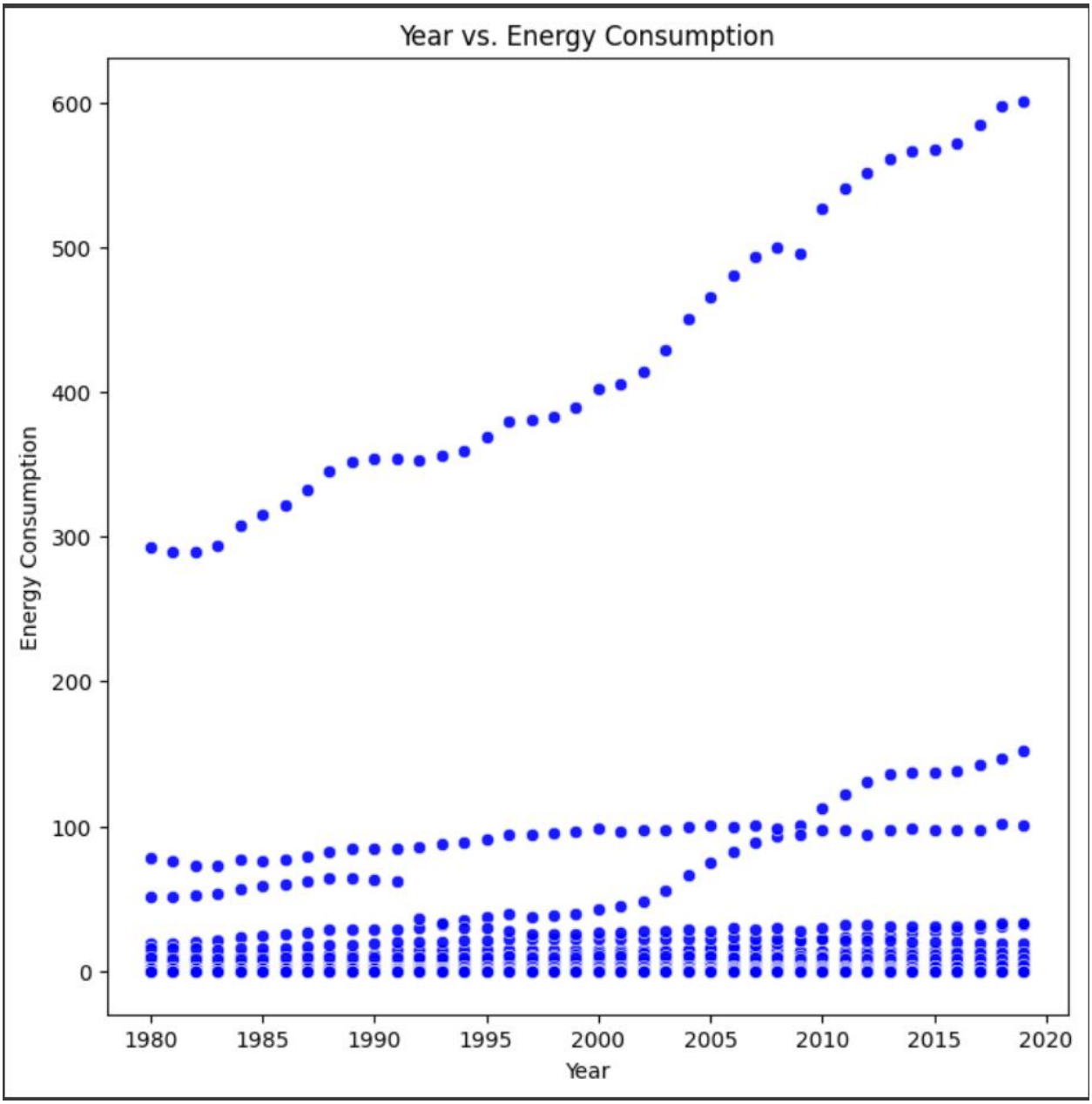


Scatterplots were plotted to visualize the interactions between GDP, population, and energy consumption, providing further insights into significant trends.




GDP vs. Energy Consumption
Energy Consumption
GDP
1e6
0
100
200
300
400
500
600
0
1
2
3
4
5
6
7
8


Energy Production vs. Energy Consumption
Energy Consumption
Energy Production
0
100
200
300
400
500
600


Year vs. Energy Consumption
Energy Consumption
Year
0
100
200
300
400
500
600
1980
1985
1990
1995
2000
2005
2010
2015
2020

Model Building:

In this task we are predicting energy consumptions and three models were implemented and evaluated. The first model was liner regression form scratch in which Gradient Descent as the optimization algorithm and Mean Squared Error (MSE) as the loss function were used. The first model also provided a fundamental understanding of how linear regression works. The second model was Linear Regression using Scikit-Learn in which optimized implementation with built-in functionalities are done. And third a Random Forest Regressor was trained to capture potential non-linear relationships between features. This ensemble learning method is known for its ability to improve predictive accuracy by aggregating multiple decision trees.

Hyperparameter Tuning & Feature Selection

Hyperparameter tuning was done by GridSearchCV for Ridge Regression to find the best alpha value, while for the Random Forest model, RandomizedSearchCV was used to select the best combination of hyperparameters, which included the number of estimators and maximum depth. These hyperparameter tuning methods were used in order to prevent overfitting and improve the generalization to unseen data.

Feature selection was necessary for selecting the most relevant factors that predict energy consumption. The two major techniques used in feature selection were the Recursive Feature Elimination to Ridge Regression and Random Forest Feature Importance for ranking and selecting the most important attributes. This would help in reducing computational complexity and improving model interpretability.

Model Evaluation

Performance Metrics:

The models were evaluated using the R-squared (R^2) metric, which measures how well the independent variables explain the variance in the target variable. The following results were observed:

- Linear Regression R-squared (Before Tuning): 0.9903
- Random Forest R-squared (Before Tuning): 0.9892
- Linear Regression R-squared (After Feature Selection & Tuning): [Insert Value]
- Random Forest R-squared (After Feature Selection & Tuning): [Insert Value]

Comparison:

The results emphasize the performance of the best model in predicting energy consumption. Besides, feature selection was performed by comparing, before and after selecting the top features, the performance of the models. Further, through residual error analysis using scatter plots, the goodness of fit for the models was determined in an attempt to see where each of them could be improved.

Conclusion

The best performance for the model was by Ridge Regression Model with an R-squared of 0.9905. Feature selection positively/negatively contributed to performance in terms of making the most influential predictors. Important factors influencing energy consumption included year, energy_production, and energy_intensity_per_capita, valuable insights for further energy management strategy.

This model can be used to predict energy consumption and manage the supply of energy consumption or manage the costs for the purpose.

Discussion

Model Performance:

The results suggest that the Random Forest model exhibited superior performance in predicting energy consumption. This finding reinforces the importance of selecting an appropriate modeling approach based on data characteristics and complexity.

Impact of Hyper-parameter Tuning and Feature Selection:

Applying hyperparameter tuning and feature selection significantly improved model performance. These techniques helped in reducing error rates and enhancing predictive accuracy by focusing on the most relevant features while optimizing model settings.

Interpretation of Results:

The study identified key predictors of energy consumption, offering valuable insights for optimizing energy efficiency and policy development. These predictors can be utilized to develop targeted strategies for energy conservation and demand management.

Limitations:

Despite its effectiveness, the model has certain limitations, including:

- Potential biases due to dataset constraints.
- Assumptions regarding feature relationships that may not fully capture real-world dynamics.

Suggestions for Future Research:

Further research could involve:

- Applying deep learning methodologies to enhance predictive capabilities.
- Expanding the dataset to include broader geographical and temporal scopes.
- Investigating the impact of emerging energy-efficient technologies on consumption patterns.