Clustering and Forecasting Environmental Indicators

Name: Hamza Nasir
Student Number: 23121983
GitHub Repository: https://github.com/hamz-boop/hamzanasir4.git

Introduction:

Global economies now prioritize environmental sustainability, while data-driven decision-making relies on recognizing patterns in demographic and ecological indicators. This project includes a thorough examination of World Bank data from European countries with an emphasis on Germany.

- Agricultural land (sq. km)
- Population, total
- Forest area (sq. km)
- CO2 emissions (kt)
- Urban population

I want to discover correlations and use K-Means clustering alongside a logistic model to predict environmental trends like CO2 emissions. The analysis demonstrates the common steps of exploration, modeling, and interpretation that occur in professional data science projects.

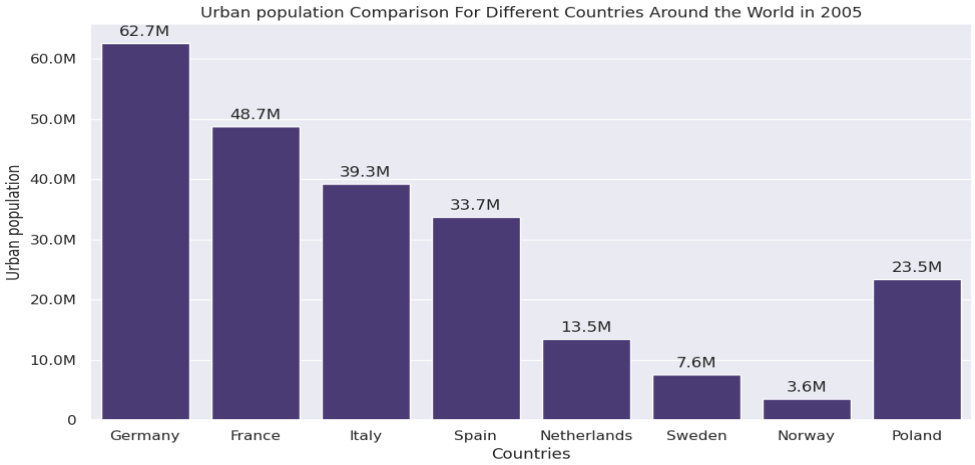
Data Preparation:

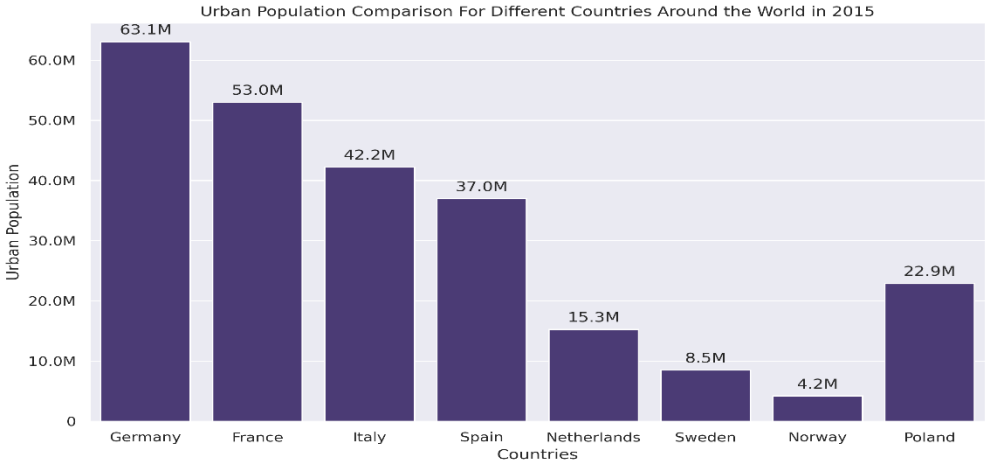
We acquired the dataset through the World Bank's platform. The dataset was refined by selecting specific indicators and countries, including Germany, France, Italy, Spain, the Netherlands, Sweden, Norway, and Poland, before removing unnecessary columns containing codes and metadata. The dataset transformation organized indicators so that years became rows and countries became columns. The dataset used mean imputation to handle missing values and Min-Max normalization to scale the data before applying clustering methods.

Visual Analysis:
Urban Population Comparison (Bar Charts)
The urban population of selected countries was represented by two bar charts for the years 2005 and 2015.

- Germany led the way in 2005 with 62.7 million urban residents, while France followed with a population of 48.7 million.
- Germany's urban population reached 63.1M by 2015, while France's urban population grew to 53.0 M.

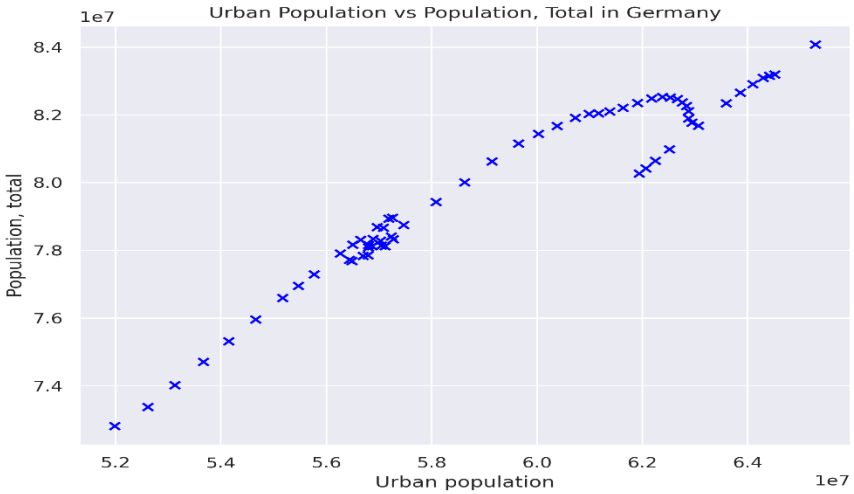
The observed trend reflects the growth of urban areas, which correlates with demographic expansion and migration movements.



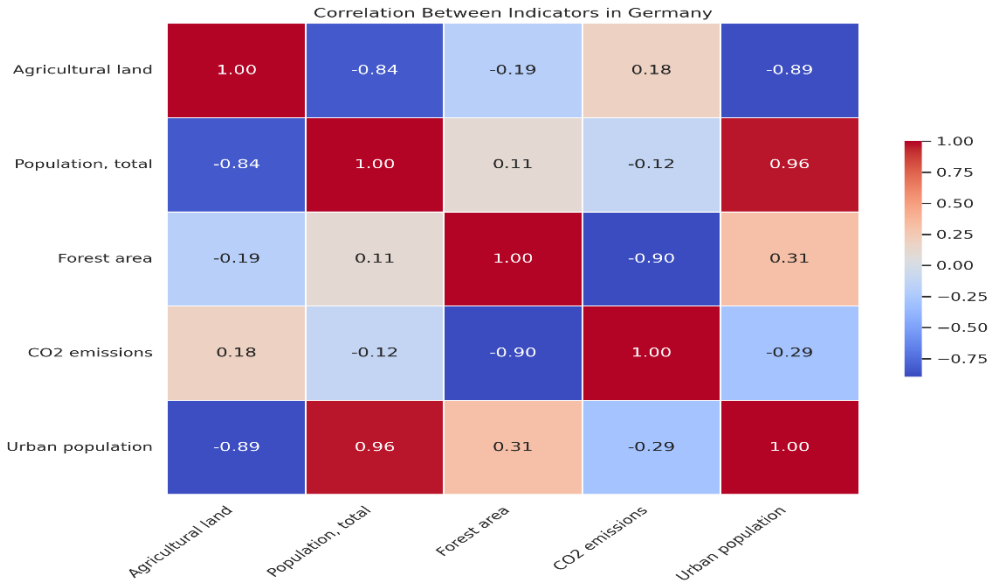


Urban vs Total Population (Scatter Plot)

Germany's scatter plot illustrates a strong positive association between urban population numbers and total population figures. The expansion of urban areas appears to be the main force behind the population changes in Germany.



Correlation Heatmap:



The heatmap reveals the correlations between different indicators in Germany.

- There is a robust positive relationship between total population numbers and urban population figures, with a correlation coefficient of +0.96.
- CO2 emissions decrease as forest area increases (-0.90).

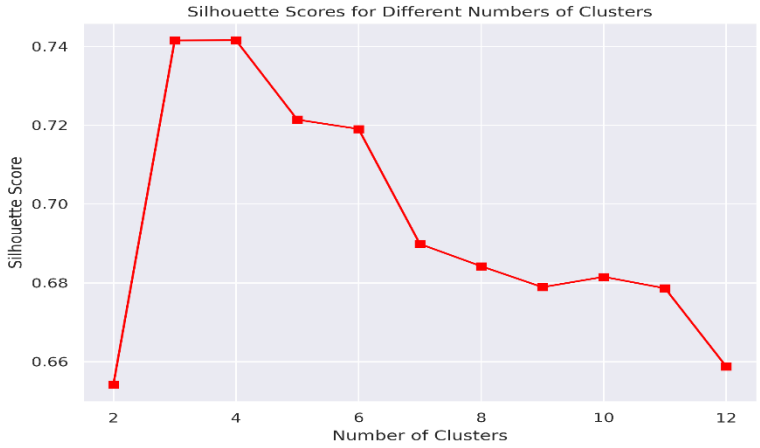
The extent of agricultural land demonstrates negative associations with population density and urban development levels. The analysis demonstrates how population expansion affects natural land conservation through various trade-offs.

Clustering Analysis: K-Means:

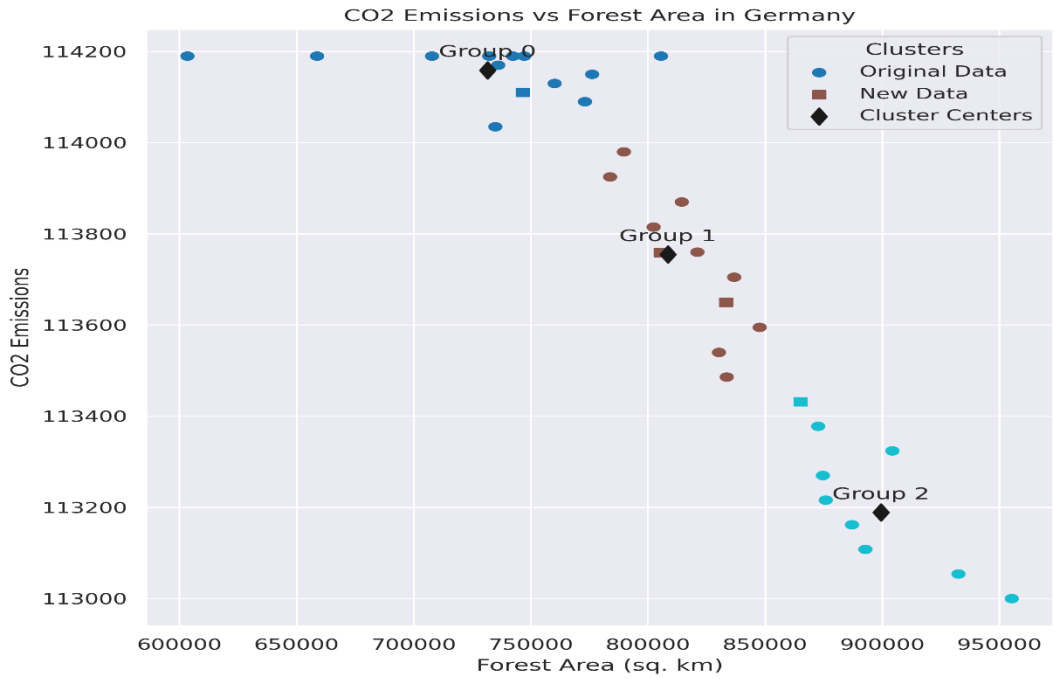
Using CO2 emissions and Forest area, K-Means clustering was applied to Germany's data.

Silhouette Analysis:

For two to twelve clusters, a silhouette score was calculated. Three clusters had the greatest score (~0.74), suggesting that there is the ideal amount for this dataset.



Cluster Visualization:



A scatter plot was generated to visualize the clusters:

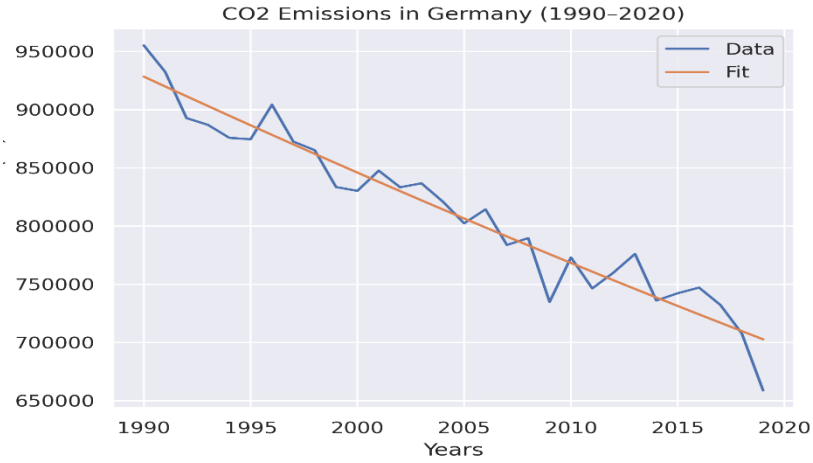
- Group 0: High emissions, low forest area.
- Group 1: Medium emissions, medium forest
- Group 2: Low emissions, high forest area

This pattern suggests distinct stages or policies in Germany's environmental strategy, with CO2 emissions dropping as forest area increases.

Forecasting: Logistic Model on CO2 Emissions:

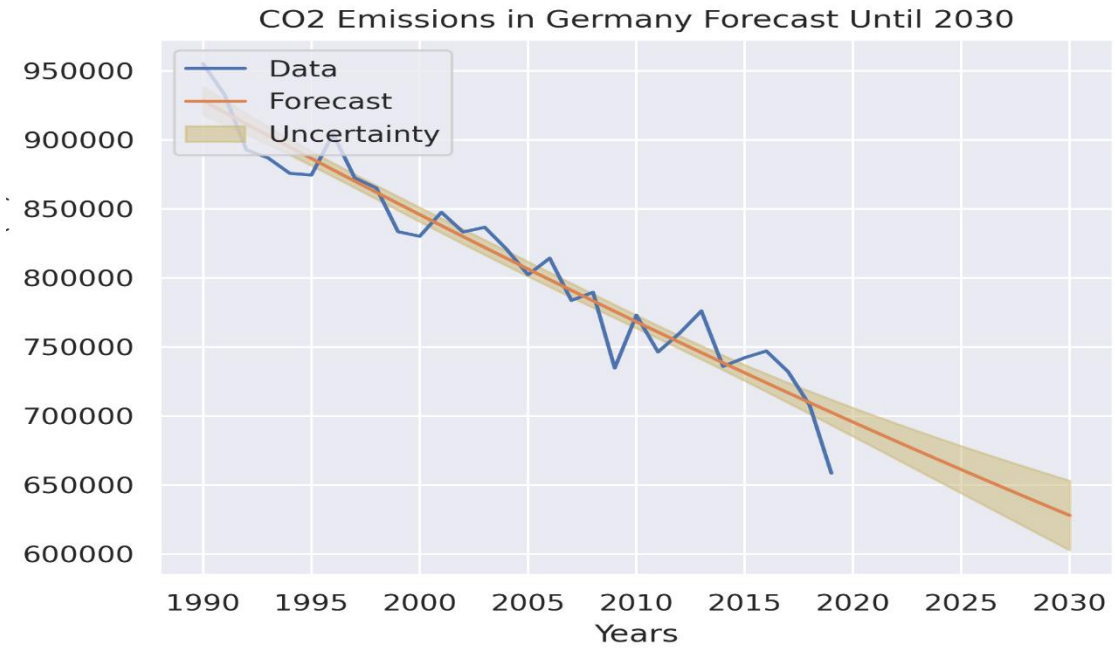
From 1990 to 2019, Germany's CO2 emissions were fitted to a logistic model. Germany's policy effectiveness in reducing emissions was reflected in the model's declining trend.

Fit Plot:



Forecast Plot (Until 2030):

According to projections, CO2 emissions should decline to approximately 650,000 kt by 2030. The model integrates uncertainty bounds to demonstrate its confidence levels. The projected reduction in emissions matches the country's targets for renewable energy growth and achieving carbon neutrality.



Conclusion:

The analysis uses clustering and logistic forecasting to effectively process environmental data.

- Urbanization has consistently increased across Europe.
- Urban population expansion shows a strong correlation with population growth.
- Forest conservation correlates with reduced CO2 emissions.
- Clustering methods identify significant changes in emission-forest area states throughout different periods.
- According to logistic forecasting projections, Germany is expected to maintain its emissions reduction trajectory until 2030.

The available insights enable policymakers and stakeholders to assess sustainability initiatives and make decisions grounded in data analysis.