Giorgio Mendoza

CS539-F23-F02

Dr. Sethi

Milestone 3: Add in some Machine Learning

Introduction: This milestone report offers insights and preliminary findings in response to the research question, “Is there a statistically significant correlation between the production of plastic waste and key socio-economic factors such as GDP and population density across various countries and regions?” It describes the analytical steps taken, presents the initial results, and outlines future directions for a more expansive analysis. The current phase of the study focuses on two datasets, allowing for an initial understanding of fundamental processes such as dataset merging and K-means clustering. This initial work prepares us to add more factors, such as population density, to our final analysis to get a better understanding.

Summary of Process:

Identification of Business Problem:

Our primary goal was to explore the relationship between the Gross Domestic Product (GDP) of European Union countries and their generation of plastic packaging waste per capita. By focusing on this correlation, we aim to understand the potential environmental impact associated with economic growth within the EU.

Data Cleaning and Preparation:

I started the project by sourcing two datasets: one detailing the GDP of various countries and the other documenting per capita plastic packaging waste generation. The data was cleaned to include only EU member states, ensuring relevance and consistency. Furthermore, the temporal scope was narrowed down to the years 2000 to 2018 to manage data volume and computational demands.

During the initial stages of data preprocessing, I tackled missing values through the implementation of the K-Nearest Neighbors (KNN) imputation method, which provided a robust means of estimating missing GDP and plastic waste data for newer EU members.

Year	2000	2001	2002	2003	2004	2005	\
Code							
IRL	38806.5000	40966.3320	43012.816	44372.758	47028.863	49223.383	
LUX	50063.8240	50527.6640	51709.734	51717.030	52624.164	53262.094	
NLD	37899.9500	38636.2230	38653.125	38803.957	39682.375	40679.490	
DNK	39021.1760	39425.8630	39709.370	39983.145	41178.562	42264.630	

Figure 1. Snippet of GDP dataset

	Code	2000	2001	2002	2003	2004	2005	2006	2007	\
0	IRL	44.840	44.890	45.090	56.130	51.990	52.410	61.750	54.030	
1	LUX	21.870	21.890	21.810	39.500	48.190	47.930	46.880	52.580	
2	EST	27.718	27.982	28.902	29.388	21.260	23.290	26.850	27.850	
4	DEU	21.780	22.950	25.130	25.090	27.330	28.710	31.460	32.140	
5	PRT	27.790	29.280	31.190	31.550	32.860	33.860	35.860	35.890	

Figure 2. Snippet of Plastic Generation dataset w/ KNN imputation method

Exploratory Data Analysis (EDA) and Insights:

After data cleansing, exploratory data analysis was conducted to get insights from the datasets. I deployed K-means clustering to categorize countries based on their GDP and plastic waste generation metrics independently. This enabled the identification of patterns and outliers within each dataset.

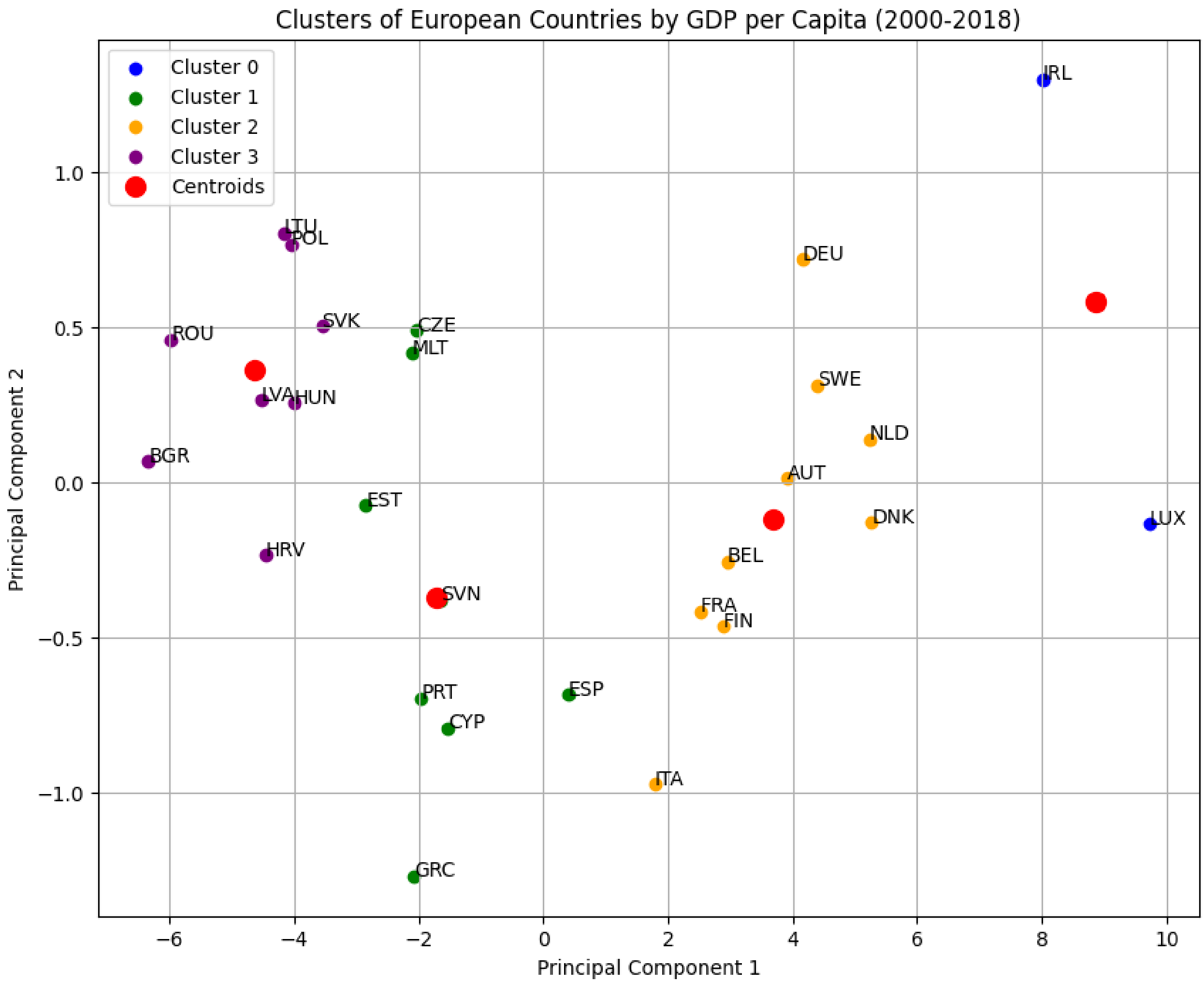


Figure 3. Cluster of European Countries by GDP

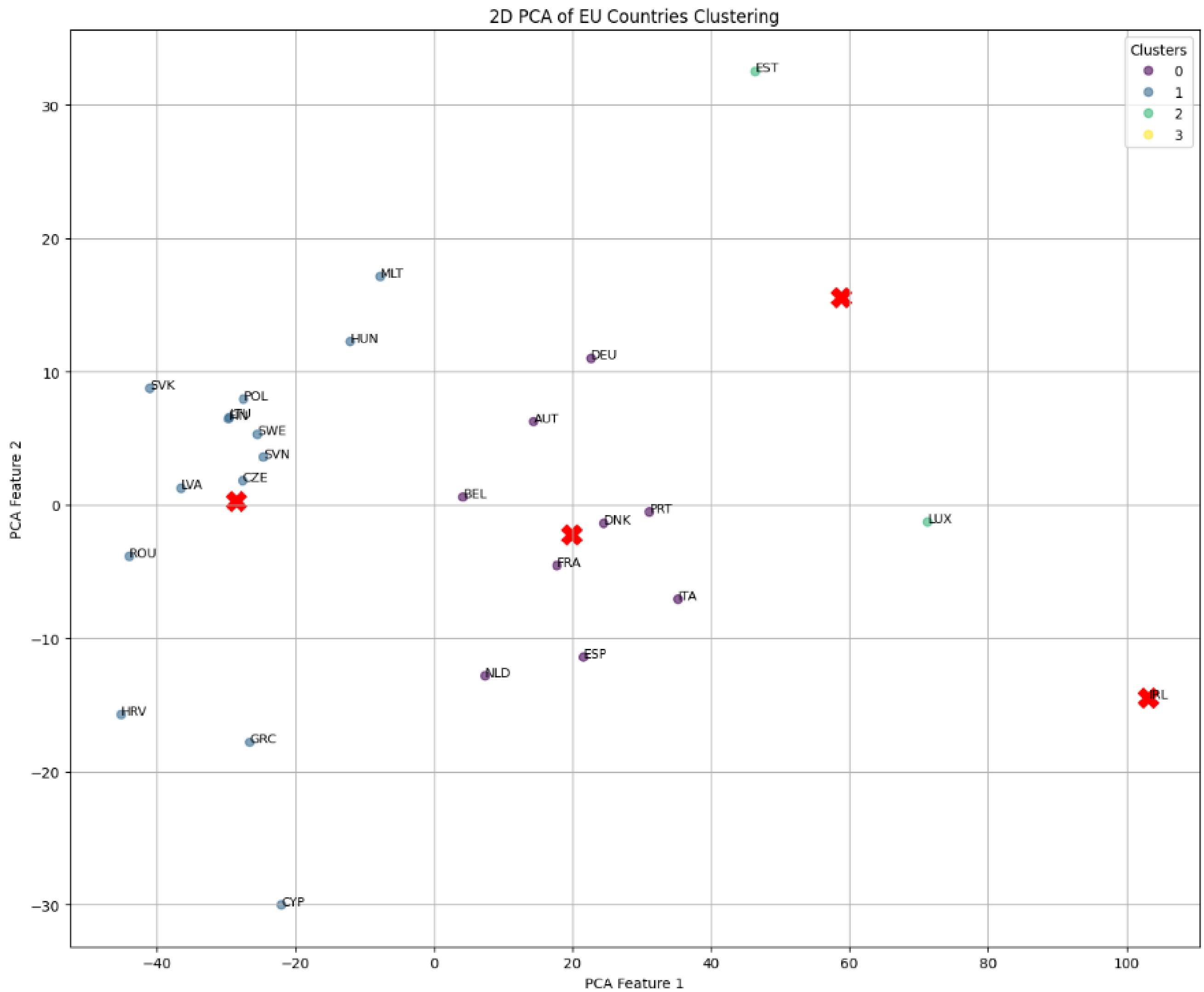


Figure 4. Cluster of Plastic Generation of EU countries

Moreover, I obtained additional details which included mean, median, maximum, and minimum values for GDP and plastic waste metrics for each country. These statistics were further visualized using bar charts to visualize the data distribution and trends across different EU nations.

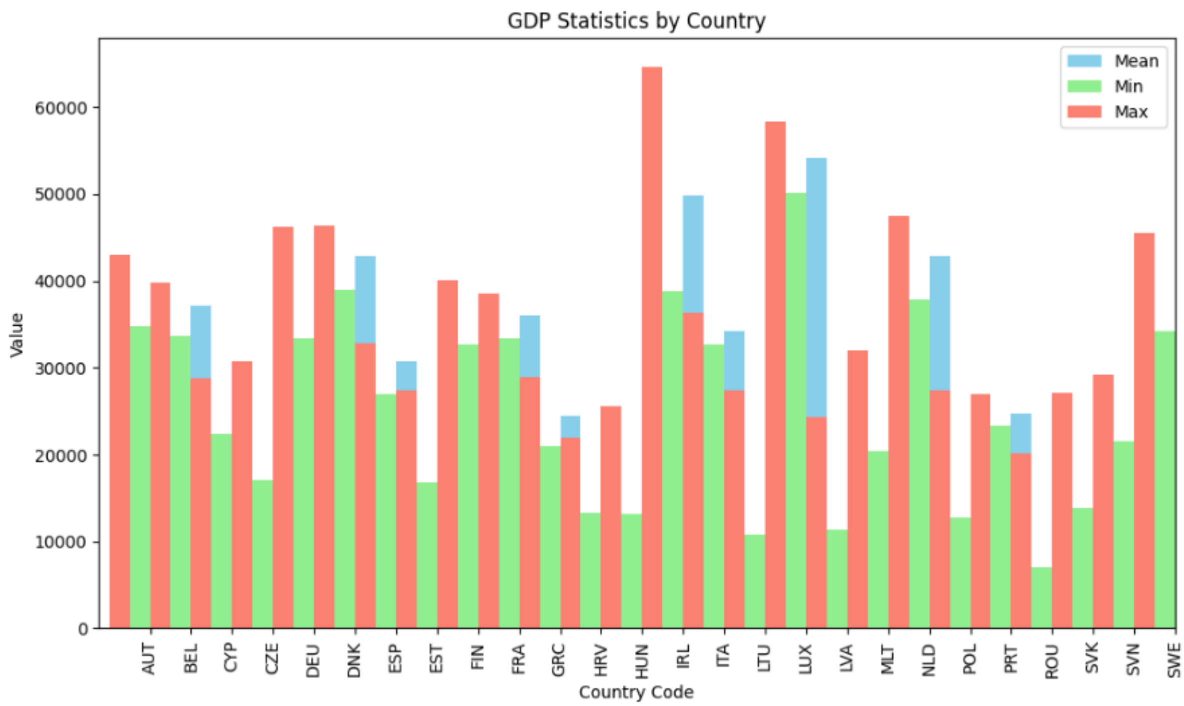


Figure 5. GDP Statistics by EU Country

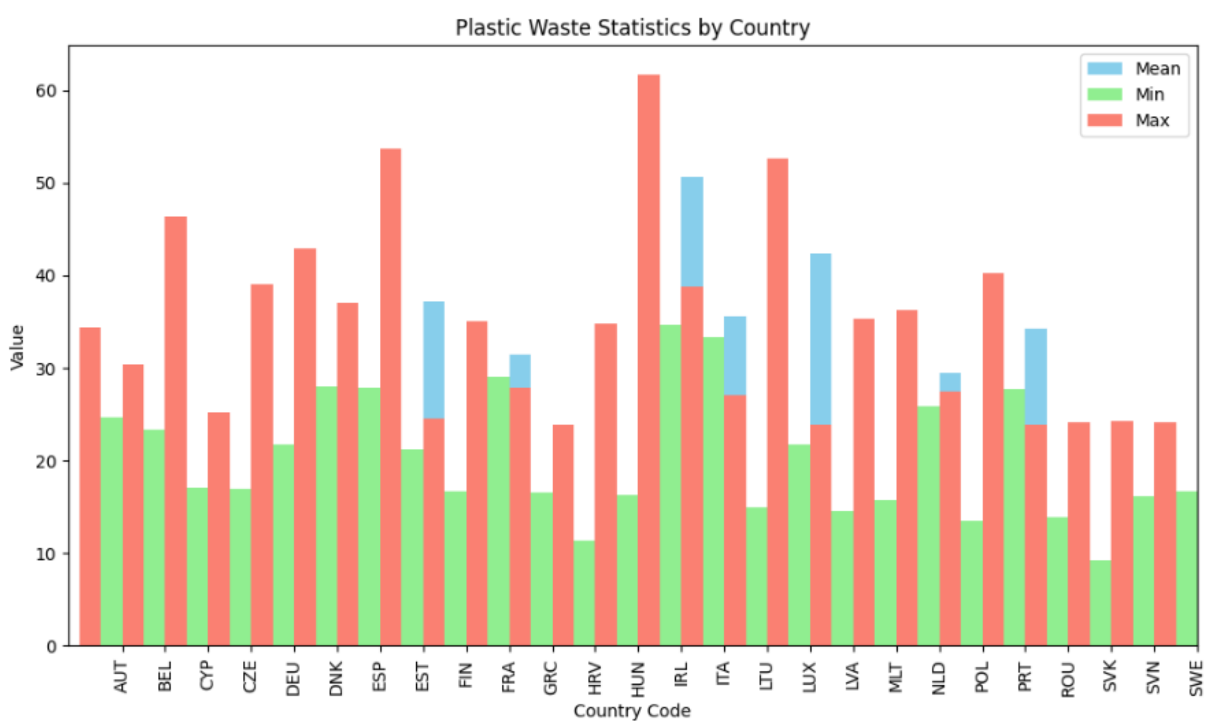


Figure 6. Plastic Waste Statistics Statistics by EU Country

Data Integration and Correlation Analysis:

Following EDA, I merged the two datasets, ensuring alignment on 'Country Code' and 'Year' fields. This merged dataset was then used to get a correlation matrix, which revealed a positive correlation between higher GDP and increased plastic waste generation, suggesting that economic affluence is a likely driver of waste production.

	Code	GDP_2000	GDP_2001	GDP_2002	GDP_2003	GDP_2004	GDP_2005
0	IRL	38806.5000	40966.3320	43012.816	44372.758	47028.863	49223.383
1	LUX	50063.8240	50527.6640	51709.734	51717.030	52624.164	53262.094
2	NLD	37899.9500	38636.2230	38653.125	38803.957	39682.375	40679.490
3	DNK	39021.1760	39425.8630	39709.370	39983.145	41178.562	42264.630
4	DEU	33367.2850	34260.2900	34590.930	34716.440	35528.715	36205.574
5	SWE	34202.6050	34666.6640	35569.773	36435.754	38016.062	39258.992

Figure 7. Snippet of merged dataset part 1

	Plastic_Waste_2000	Plastic_Waste_2001	Plastic_Waste_2002	\
0	44.840	44.890	45.090	
1	21.870	21.890	21.810	
2	28.760	30.290	32.820	
3	29.440	28.020	29.250	
4	21.780	22.950	25.130	
5	16.730	17.930	18.740	

Figure 8. Snippet of merged dataset part 2

Cluster Analysis:

In the final step, I applied K-means clustering to the merged dataset. This analysis corroborated the initial findings, illustrating that countries with similar economic profiles tend to show parallel patterns in plastic waste generation.

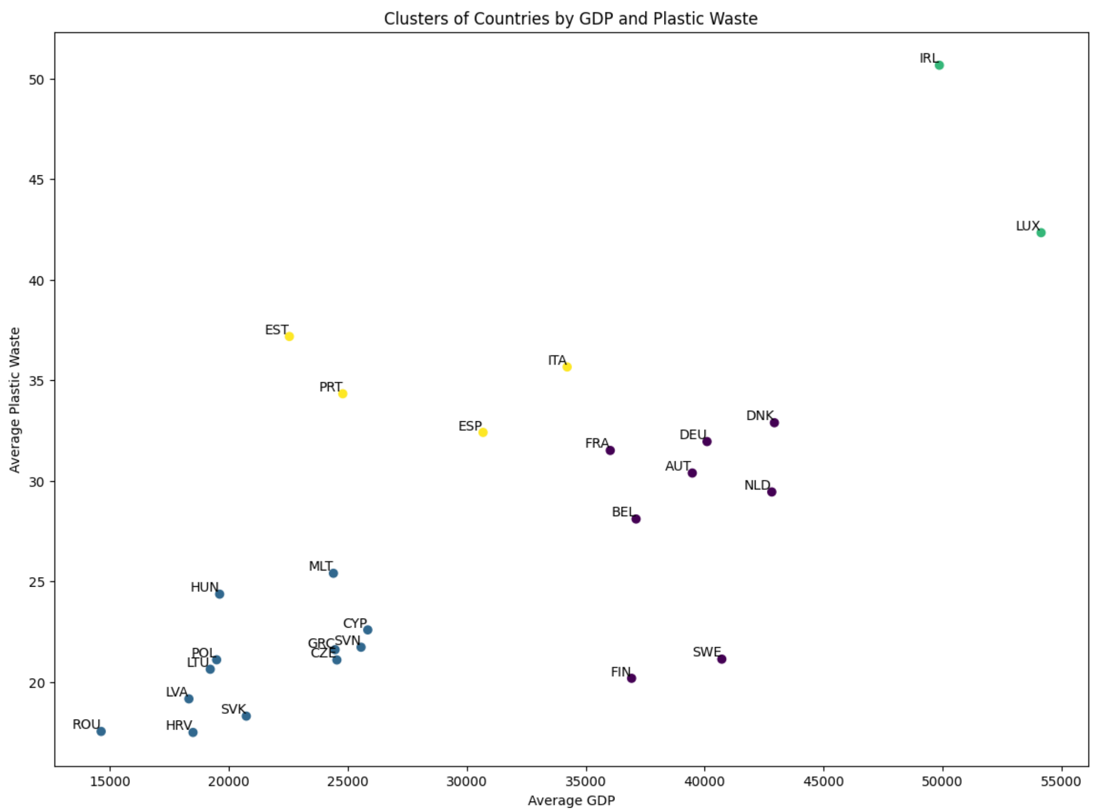
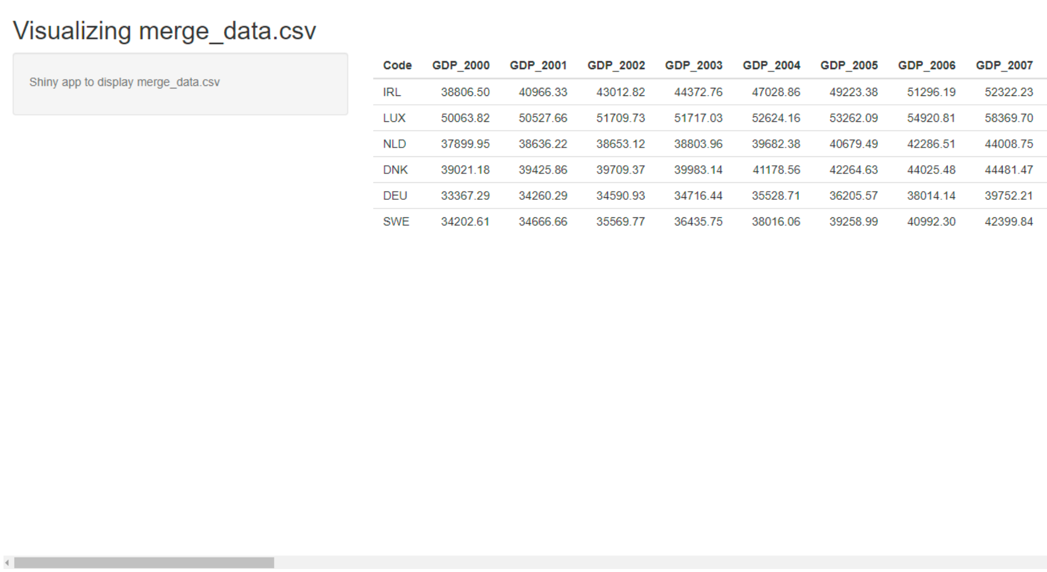


Figure 9. Cluster of Countries by GDP and Plastic Waste

Shiny App:

I also prepared a very basic Shiny app to display the data of the merged dataset. I will try to add more features in the future, I just wanted to know the basics.



Visualizing merge_data.csv

Shiny app to display merge_data.csv

Code	GDP_2000	GDP_2001	GDP_2002	GDP_2003	GDP_2004	GDP_2005	GDP_2006	GDP_2007
IRL	38806.50	40966.33	43012.82	44372.76	47028.86	49223.38	51296.19	52322.23
LUX	50063.82	50527.66	51709.73	51717.03	52624.16	53262.09	54920.81	58369.70
NLD	37899.95	38636.22	38653.12	38803.96	39682.38	40679.49	42286.51	44008.75
DNK	39021.18	39425.86	39709.37	39983.14	41178.56	42264.63	44025.48	44481.47
DEU	33367.29	34260.29	34590.93	34716.44	35528.71	36205.57	38014.14	39752.21
SWE	34202.61	34666.66	35569.77	36435.75	38016.06	39258.99	40992.30	42399.84

Figure 10. Shiny app displaying merged dataset

Research Question:

Research Question:

Is there a statistically significant correlation between the production of plastic waste and key socio-economic factors, such as Gross Domestic Product (GDP) and population density, across various countries and regions within the European Union?

Machine Learning Perspective:

This research question seeks to understand the nature and strength of the relationship between socio-economic factors and environmental outcomes, which can be classified under the umbrella of regression analysis in machine learning. In this context, the objective is to find out whether predictable patterns exist between independent predictor variables (socio-economic factors) and a dependent response variable (plastic waste production).

Predictor Variables:

Gross Domestic Product (GDP) per capita: This economic indicator represents the economic productivity and affluence of a nation, serving as a predictor for consumption patterns that may influence waste generation.

Population Density: As an indicator of population distribution and urbanization, population density can affect waste management practices and the efficiency of waste collection and recycling infrastructure.

Response Variable:

Plastic Waste Generation per capita: This is the primary response variable, representing the environmental impact of the studied socio-economic factors.

Machine Learning Question:

Our study uses K-means clustering, an unsupervised learning technique, to investigate potential patterns linking GDP per capita and plastic waste generation per capita. We aim to determine if countries can be grouped into clusters that reveal a notable correlation between economic status and environmental impact. This approach will highlight whether a higher GDP per capita aligns with greater plastic waste generation, focusing on understanding data patterns rather than predicting outcomes.

Analysis Approach:

Feature Engineering:

From the raw datasets, we extracted GDP and plastic waste generation data and aligned them by country codes and years, creating a structured time-series format. We addressed missing values by applying the KNN imputation method, which infers missing data within the context of neighboring points. We also performed standardization to ensure uniformity and comparability across features, enhancing the integrity of our subsequent clustering analysis.

Modeling:

Our primary modeling technique has been K-means clustering, which we used to discern natural groupings in the data based on economic indicators and waste production levels. The clusters formed provide insights into the patterns of GDP and plastic waste generation across different countries. Moving forward, we may explore hierarchical clustering for more nuanced groupings or PCA for dimensionality reduction to visualize high-dimensional clustering.

Preliminary Results:

Performance Measures:

For the clustering analysis, the Elbow Method was employed to determine the optimal number of clusters. This method is effective in a clustering context as it evaluates the within-cluster sum of squares (WCSS), which helps in finding the k-value where the marginal gain in explained variance starts to decrease.

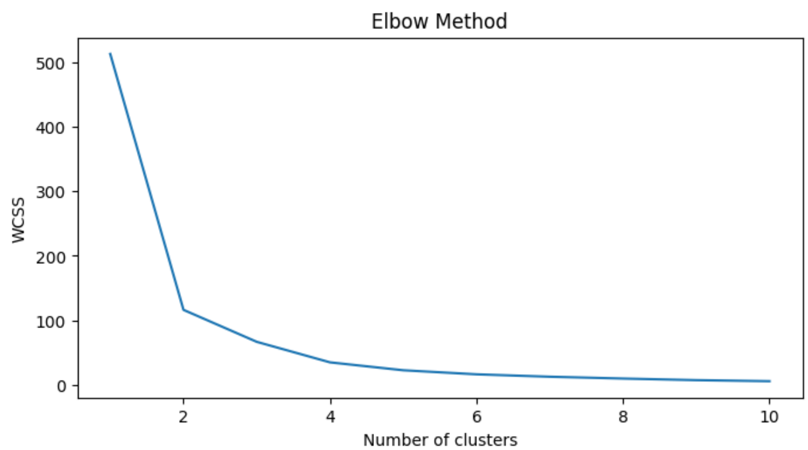


Figure 11. Elbow method plot

Results:

The correlation analysis indicated a positive pattern between GDP per capita and plastic waste generation per capita, suggesting that countries with higher economic output tend to have higher levels of plastic waste. The Elbow Method was used to inform the choice of the number of clusters for K-means, leading to a selection that balanced detail with generalization.

Tuning:

The K-means clustering model was tuned by varying the number of clusters and observing the Elbow plot. The 'elbow' point represents a diminishing return on the WCSS and was chosen as the cut-off for the number of clusters, ensuring an efficient yet insightful clustering solution.

Visualization:

To depict the clustering results, scatter plots were generated to show the distribution of countries based on GDP per capita and plastic waste generation, with different clusters marked by distinct colors. Bar charts were also created to demonstrate the economic and environmental profiles of the clusters.

Further Considerations:

The scope of the study will also be expanded to encompass more socio-economic factors, such as population size, recycling rates, and energy consumption, to provide a more holistic understanding of the factors influencing plastic waste generation.

Final Analysis Plan:

Scale: Data from 2000 to 2018, EU countries, potentially expanding globally. Additional socio-economic factors to be included.

Model Complexity: Beyond K-means, exploring hierarchical clustering, Random Forests, or PCA for nuanced insights.

Tools: R, Python, and scikit-learn, with a focus on efficient computation to manage longer processing times due to increased data and model complexity.

Potential Obstacles:

Data Completeness: Addressing missing data with robust imputation methods to avoid bias.

Model Overfitting: Implementing cross-validation and regularization to ensure model generalizability.

Computational Resources: Balancing model complexity with available computational power.

Interpretability: Ensuring complex model outputs remain understandable for stakeholders.

Research Focus: Keeping the analysis aligned with the core research questions to provide clear, relevant insights.

Appendix

```
library(shiny)
```

```r
library(readr) # for read_csv
#UI for application
ui <- fluidPage(
    # Application title
  titlePanel("Visualizing merge_data.csv"),
    # Sidebar with a simple input
  sidebarLayout(
    sidebarPanel(
      helpText("Shiny app to display merge_data.csv")
    ),
        # Show a table output in the main panel
    mainPanel(
      tableOutput("dataTable")
    )
  )
)
# Define server logic
server <- function(input, output) {
  # Reactive expression to read the data
  merged_data <- reactive({
    req(file.exists("C:\\Users\\Gio\\Downloads\\shiny\\merged_data.csv"))
    read_csv("C:\\Users\\Gio\\Downloads\\shiny\\merged_data.csv")
  })

  # Output the table
  output$dataTable <- renderTable({
    req(merged_data())
    head(merged_data())
  })
}
# Run the application
shinyApp(ui = ui, server = server)
```