Giorgio Mendoza

CS539-F23-F02

Dr. Sethi

Analyzing the Correlation Between GDP, Population, and Plastic Waste Generation in the EU

Overview: This project investigates the potential relationship between plastic waste production and key socio-economic indicators, such as Gross Domestic Product (GDP) and population density, across various countries from the European Union.

Summary: This study examines the link between socio-economic factors and plastic waste in EU countries, merging data from 2000 to 2018 on GDP, population, and waste production. Employing PCA and K-means clustering, it uncovers notable correlations and patterns, suggesting how economic and demographic factors impact the environment.

Project Description:

Motivation: The project is driven by the critical challenge of addressing plastic pollution, a significant environmental concern that affects ecosystems worldwide. Amidst escalating plastic waste production and its harmful consequences, it’s important to discern the socio-economic catalysts to formulate effective waste management policies.

Expected Discovery: Key objectives include: 1) Assessing the correlation between GDP per capita and plastic waste generation; 2) Exploring the impact of population density on plastic waste output; 3) Employing PCA and K-means clustering to reveal patterns and clusters in EU countries. The goal is to provide insights that can inform targeted environmental policies within the EU.

Datasets: Data Sources and Preparation: This study combines three datasets: Eurostat's plastic packaging waste data (2000-2020), OurWorldInData.com's global GDP per capita records (since 1820), and the

World Bank's global population statistics. Each dataset was preprocessed for consistency and timeframe alignment. The initial phase involved merging the GDP and plastic waste datasets for preliminary analysis, followed by integrating global population data. This process enabled the examination of various socio-economic and environmental factors within the EU through advanced ML techniques.

Methodological Overview:

- Data Collection and Preparation
- Clustering Analysis
- Principal Component Analysis
- Hierarchical Clustering
- Correlation Analysis
- Data Visualization
- Final Data Aggregation

Summary of Process in Detail:

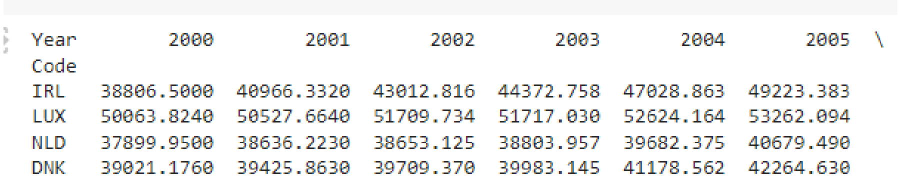
Identification of Business Problem:

The primary goal is to explore the relationship between the Gross Domestic Product of EU 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:

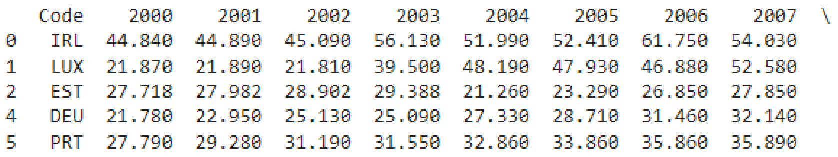
I started the project by sourcing two datasets: one detailing the GDP of various countries and the other documenting per capita plastic packaging waste generation. The data was cleaned to include only EU member states, ensuring relevance and consistency. Furthermore, the temporal scope was narrowed down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

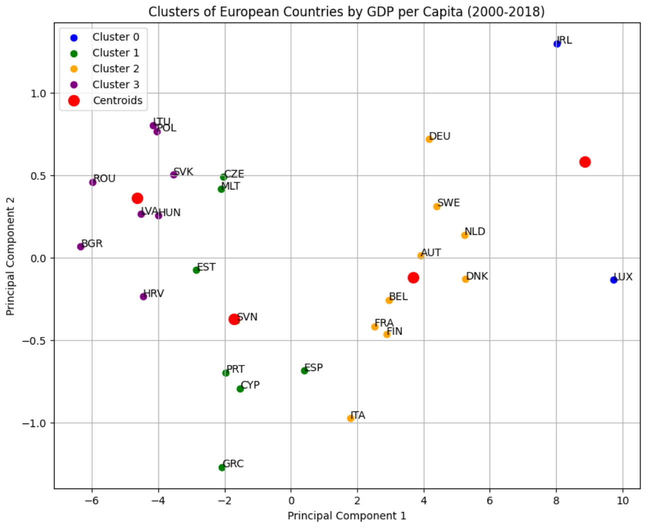


Figure 3. Cluster of European Countries by GDP

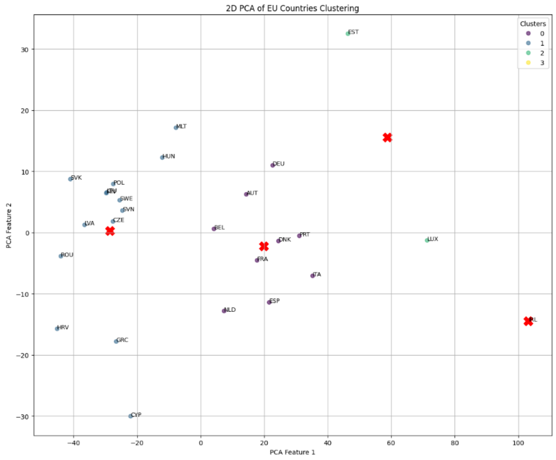


Figure 4. Cluster of Plastic Generation of EU countries

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.

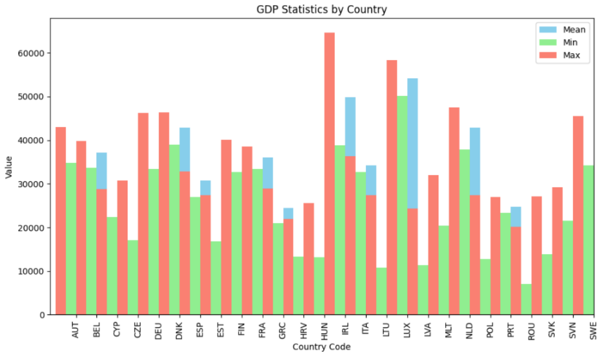


Figure 5. GDP Statistics by EU Country

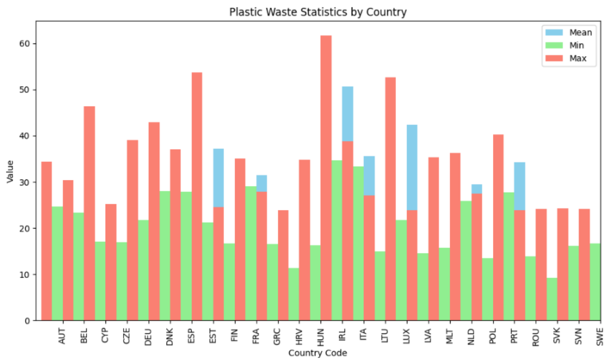
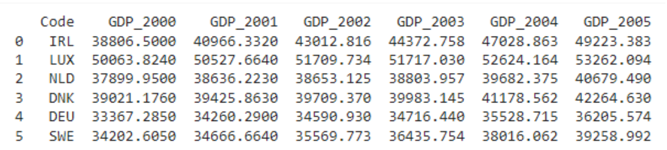


Figure 6. Plastic Waste Statistics Statistics by EU Country

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

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.

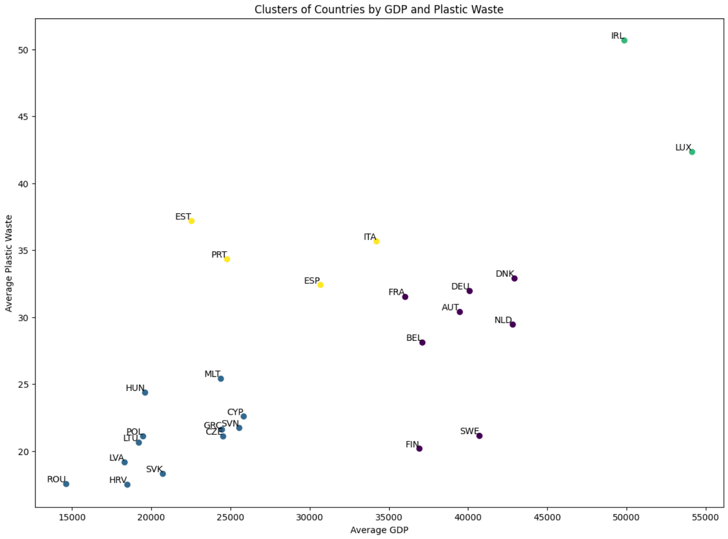


Figure 9. Cluster of Countries by GDP and Plastic Waste

Conclusion:

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.

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.

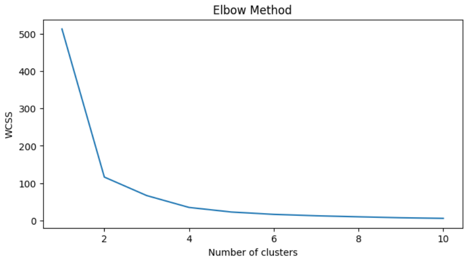
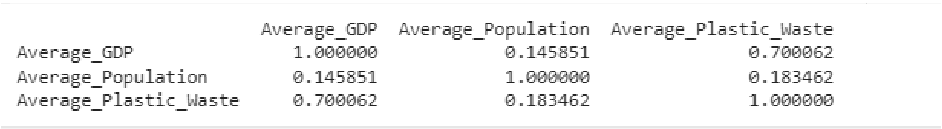


Figure 11. Elbow method plot

Correlation Matrix: The correlation matrix reveals significant insights into the relationships between GDP, population, and plastic waste production among European Union countries. A strong positive correlation (0.700) between average GDP and average plastic waste production suggests that higher GDP per capita is associated with increased plastic waste generation. This indicates that economic prosperity in the EU may be a significant factor in plastic waste production. In contrast, the correlation between average population and plastic waste (0.183) is relatively weaker, suggesting that population size has a less pronounced impact on plastic waste production compared to GDP. These findings highlight the importance of considering economic factors in strategies aimed at managing and reducing plastic waste in the EU.



```
                        Average_GDP  Average_Population  Average_Plastic_Waste
Average_GDP                1.000000            0.145851               0.700062
Average_Population         0.145851            1.000000               0.183462
Average_Plastic_Waste      0.700062            0.183462               1.000000
```

Figure 12. Correlation Matrix

Hierarchical Clustering Dendrogram: The dendrogram produced from hierarchical clustering provides a visual representation of the relationships between the selected EU countries based on their socio-economic indicators and plastic waste generation. Countries are grouped based on the similarity of their data profiles, with the 'distance' on the y-axis indicating the dissimilarity between clusters. For instance, we observe that countries like Germany (DEU) and Spain (ESP) form distinct clusters early on, suggesting their unique socio-economic or environmental patterns compared to other EU nations. In contrast, clusters containing countries like Belgium (BEL), Estonia (EST), and France (FRA) are linked at a shorter distance, indicating a closer resemblance in the factors being analyzed. This hierarchical approach not only categorizes countries into distinct groups but also hierarchically orders these groups based on their similarity, offering nuanced insights into how these nations compare in the context of GDP, population, and plastic waste metrics.

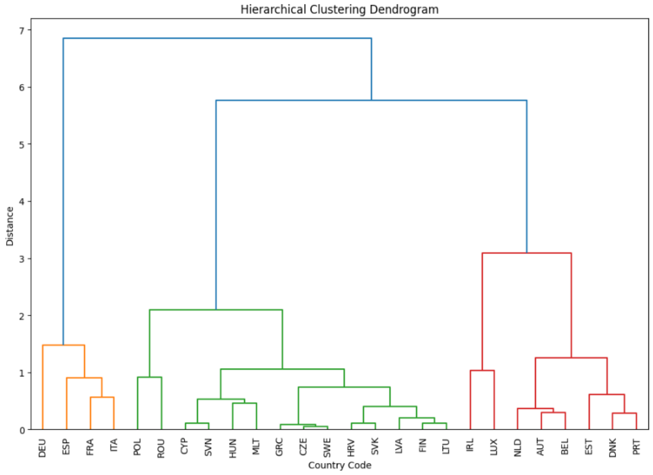
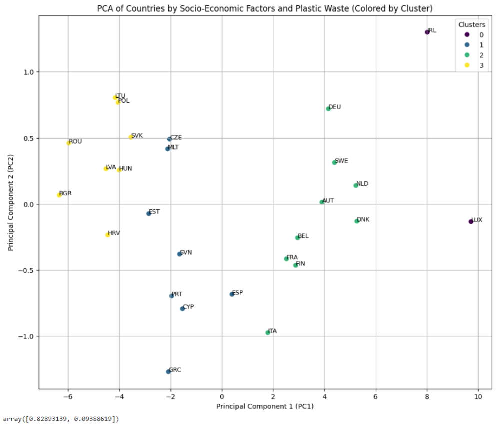


Figure 13. Hierarchical Clustering Dendrogram

PCA Analysis: The PCA scatter plot illustrates the clustering of EU countries based on socio-economic factors and plastic waste generation, with the results color-coded by cluster. This visualization, derived from the PCA captures the variance in the dataset with the first two principal components accounting for approximately 82.89% and 9.39% of the variance respectively. The plot highlights clear groupings of countries, indicating that nations with similar GDP and plastic waste profiles tend to cluster together. For example, the distinct positioning of countries like Germany (DEU) indicates its unique standing in terms of economic and environmental measures. The clustering also suggests potential patterns in the way socio-economic status relates to environmental impact, with wealthier countries possibly generating more plastic waste. These insights underscore the complex interplay between economic development and environmental health, pointing to the need for tailored strategies that address the nuances of plastic waste management in the context of varying economic realities.



```
array([0.82893139, 0.09388619])
```

Figure 14. PCA of Countries by Socio-Economic Factors & Plastic Waste

The conclusion of this study highlights the interconnectedness of GDP, population, and plastic waste generation in EU countries, emphasizing the need for policy adaptation and further research. A notable positive correlation between GDP per capita and plastic waste generation indicates that economic growth could lead to increased waste, urging a rethink of waste management strategies. K-means clustering, using the Elbow Method, and hierarchical clustering have identified clear groupings and detailed similarities among EU countries based on their economic and environmental profiles. These methods, along with PCA analysis, have visually demonstrated the variance and clustering of EU nations, underscoring the complexities and relationships between socio-economic factors and environmental impact.

This multidimensional analysis offers a granular view of the EU's environmental and economic landscape, with the clustering and correlation findings pointing towards a deeper connection between a nation's wealth and its environmental footprint. These results call for a sophisticated approach to environmental policy, one that takes into account the complexities of each nation's economic makeup.

Future research could address the limitations of this EU-focused analysis on economic growth and plastic waste, including exploring global data variations and incorporating broader socio-economic factors for a more comprehensive understanding and impactful policy development.