Clustering and fitting

NAME: BHASKAR CHERUKU

STUDENT ID: 22003064

MODULE: APPLIED DATA SCIENCE-I (7PAM2000-0105-2022)

GITHUB REPOSITORY LINK: https://github.com/bc22aba/applied_data_science/tree/main/assignment-3

INTRODUCTION:

Analysis Related To Public Data From The World Bank, And Specifically Country-by-country Indicators Related To Climate Change, Datalink. For The Visualization Perspective The Additional Relevant Indicators Are Used For GDP Per Capita (Additional Datalink). From The Data Retrieved From The Internet Find The Cluster Visualization For The Relevant Features.

DATA EXPLORATION:

Data Was Downloaded From The Mentioned Datalink, Dataset Contains The Total 66 Features And 266 Samples Are There, From The Year 1960 To 2021 Data Was Shared. Mainly Data Is Having The Country Name, Code, Indicator And Different Year CO2 Emission Data. Emission Data Has Been Observed During Different Year And Country, Idea Is To Explore The Data Clustering.

VISUALIZATION:

K-means Clustering algorithm applied to check the Co2 emission data total 4 different clusters are formed.

Since the dataset is containing information about the different year wise carbon emission for the particular Country Per capita. Carbon emission values are in terms of metric tons.

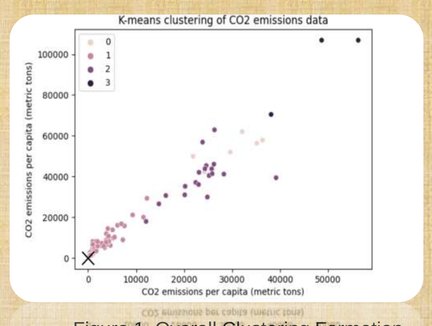

Figure 1. Overall Clustering Formation

Figure 1, display the cluster formation of the using the k-means algorithm, From the figure we can observe that there are total Four clusters has been

Figure 2. CO2 emissions per capita during 1960 t0 2020

It has been observed that during each and every year the co2 emissions value is increasing.

In the figure 2, we have tried to Observe the carbon emission per capita in the united states. Emission values are increasing from 1960 to 2020. So one can conclude that over a period of time, carbon emission values are increasing. Since emission Values are increasing, graph is linearly increasing.

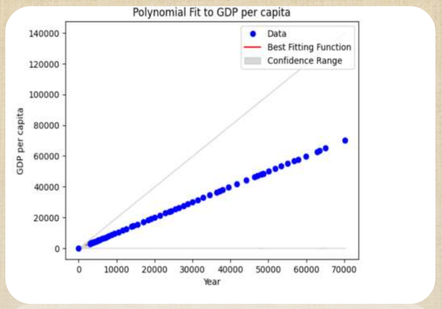

Figure 3. Polynomial Fit to GDP per capita

Figure 3, using the unsupervised Machine learning approach build The polynomial graph to fit the data from the various year to gdp per Capita. In the graph, data distribution, Confidence range and best fitting Function lines are displayed.

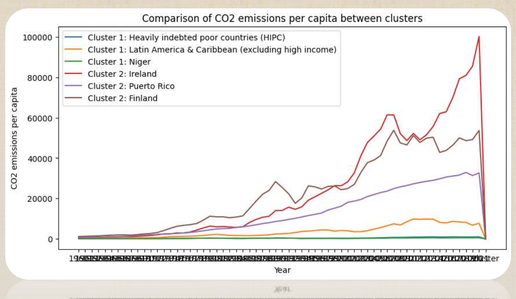

Figure 4. Different comparison of co2 emissions per capita between two different clusters country and based on the country cluster comparisons are made.

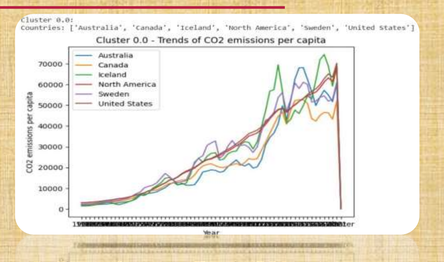

Figure 5: Cluster 0 country Trends CO2 emissions per capita

Here, data Distribution is distributed in a linearly Increasing manner, so line fitting function will be able to fit line properly. After getting the cluster Formation and country Distribution according to Cluster wise, for the Visualization purpose. Sample country from

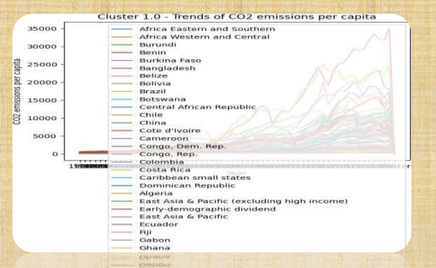

Figure 6: Cluster 1 country Trends CO2 emissions per capita

Cluster 1:{heavily indebted county, Latin America, and Niger} and from Cluster 2: {Ireland, Puerto Rica, and Finland} was considered.

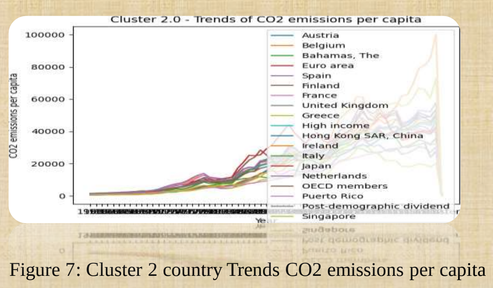

Figure 7: Cluster 2 country Trends CO2 emissions per capita

Figure 9. Indicate cluster Number and Country details For the same Mentioned.

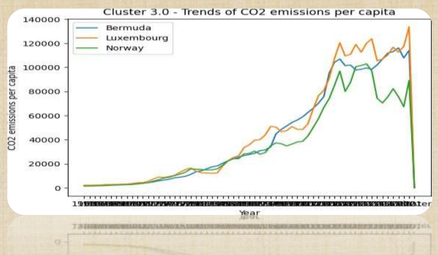

Figure 8: Cluster 3 country Trends CO2 emissions per capita

Cluster Number	Country Name
0	Countries: ['Australia', 'Canada', 'Iceland', 'North America', 'Sweden', 'United States']
1	Countries: ['Africa Eastern and Southern', 'Africa Western and Central', 'Burundi', 'Benin', 'Burkina Faso', 'Bangladesh', 'Belize', 'Bolivia', 'Brazil', 'Botswana', 'Central African Republic', 'Chile', 'China', "Cote d'Ivoire", 'Cameroon', 'Congo, Dem. Rep.', 'Congo, Rep.', 'Colombia', 'Costa Rica', 'Caribbean small states', 'Dominican Republic', 'Algeria', 'East Asia & Pacific (excluding high income)', 'Early-demographic dividend', 'East Asia & Pacific', 'Ecuador', 'Fiji', 'Gabon', 'Ghana', 'Guatemala', 'Guyana', 'Honduras', 'Heavily indebted poor countries (HIPC)', 'Haiti', 'IBRD only', 'IDA & IBRD total', 'IDA total', 'IDA blend', 'IDA only', 'India', 'Jamaica', 'Kenya', 'St. Kitts and Nevis', 'Korea, Rep.', 'Latin America & Caribbean (excluding high income)', 'Latin America & Caribbean', 'Low income', 'Sri Lanka', 'Lower middle income', 'Low & middle income', 'Lesotho', 'Late-demographic dividend', 'Morocco', 'Madagascar', 'Mexico', 'Middle income', 'Malawi', 'Malaysia', 'Niger', 'Nigeria', 'Nicaragua', 'Nepal', 'Pakistan', 'Panama', 'Peru', 'Philippines', 'Papua New Guinea', 'Pre-demographic dividend', 'Portugal', 'Rwanda', 'South Asia', 'Sudan', 'Senegal', 'Sierra Leone', 'Sub-Saharan Africa (excluding high income)', 'Sub-Saharan Africa', 'Suriname', 'Eswatini', 'Seychelles', 'Chad', 'East Asia & Pacific (IDA & IBRD countries)', 'Togo', 'Thailand', 'Latin America & the Caribbean (IDA & IBRD countries)', 'South Asia (IDA & IBRD)', 'Sub-Saharan Africa (IDA & IBRD countries)', 'Trinidad and Tobago', 'Turkiye', 'Uganda', 'Upper middle income', 'Uruguay', 'St. Vincent and the Grenadines', 'World', 'South Africa', 'Zambia', 'Zimbabwe']
2	Countries: ['Austria', 'Belgium', 'Bahamas, The', 'Euro area', 'Spain', 'Finland', 'France', 'United Kingdom', 'Greece', 'High income', 'Hong Kong SAR, China', 'Ireland', 'Italy', 'Japan', 'Netherlands', 'OECD members', 'Puerto Rico', 'Post-demographic dividend', 'Singapore']
3	Countries: ['Bermuda', 'Luxembourg', 'Norway']

Figure 9 . Cluster number and country details in table format.