Julianna Perucci
Professor Kontothanassis
DS 210 Final Project
December 12, 2024

Clustering and Adjacency Graph Analysis of Global Health and Environmental Data

Background:

In this project, I used a CSV file titled "Life Expectancy and Socio-Economic (World Bank)" which contains country-specific data on disability-adjusted life years (DALYs) due to communicable and non-communicable diseases, as well as CO2 emission data. My code first cleans and pre-processes the data by ensuring unique countries by filtering out duplicate or incomplete data. Then, I wrote code to K-means cluster my data by grouping countries into 5 clusters based on DALYs from communicable and non-communicable diseases and CO2 emissions. Finally, I wrote code to build an adjacency graph to connect countries based on a 0.5 distance threshold. This project provides insight into identifying vulnerable regions that may require health interventions or face environmental risks. The results could be useful to policy-makers or researchers in deciding where resources should be allocated, whether or not a country is aligning with Sustainable Development Goals, or to answer research questions such as "Do poorer countries tend to have higher DALYs?" or "Do wealthier countries tend to emit more CO2?"

Code Explanation:

First, I imported libraries relevant to the CSV cleaning module I created, random sampling, HashSet data structures, and handling errors. Then, I defined the struct "Country" which contains the name of the country, the communicable and non-communicable disease metrics, the CO2 emissions, and the assigned cluster. On a separate module, I created my load_and_clean_data function, which reads the CSV file, removes invalid or duplicate entries, and creates "Country" objects for rows with valid data. Next, I created an initialize_centroids function, which randomly selects certain countries, k, as the initial centroids. Then, I created an assign_clusters function, which uses Euclidean distances, which is defined later, to assign each country to the nearest centroid. As a result, I created an update_centroids function, which updates the centroid positions based on the mean of the points assigned to each cluster. Next, I defined a euclidean_distance function, which computes the Euclidean distance ($\sqrt{(x1-x2)^2 + (y1-y2)^2}$) between two countries. Moving forward, I defined a kmeans function that implements the algorithm or k-means with a maximum amount of iterations. Then, I defined a build_graph function which builds an adjacency graph in which countries are connected if their Euclidean distance is below a specified threshold. On the main function, I loaded the data from "life expectancy.csv", applied k-means clustering with k equal to 5 clusters, built an adjacency graph

with a Euclidean distance threshold equal to 0.5, and printed my results. I also ran four tests for load_and_clean_data, euclidean_distance, kmeans, and build_graph, all of which had passed.

Output Analysis:
The output of the adjacency graph demonstrated some connections between countries. For example, Angola has a Euclidean distance less than 0.5 from Dominica, Sri Lanka, and Vietnam. The United Arab Emirates has a Euclidean distance of less than 0.5 from Burkina Faso, Equatorial Guinea, Mozambique, Pakistan, and Papua New Guinea. Also, Zambia has a Euclidean distance of less than 0.5 to Botswana, Comoros, Guatemala, and the Solomon Islands. These countries are just a few, there are plenty of other countries, as displayed below, with connections by small distances to others on the graph. The more connections a country has, the more representative those connected countries are of certain countries. In the output, if a country does not have a connection, it is displayed as ->[]. The results of the K-means clustering groups countries based on similar health and environmental factors. Cluster 3 includes countries like Angola, UAE, and Botswana, to name a few, which means these countries share higher DALYs from communicable diseases or similar CO2 emissions. Cluster 4 includes countries like Canada, Belgium, the United States of America, and Germany, which represent countries with lower DALYs and CO2 emission rates. Cluster 0 includes countries like China, India, and Niger. This cluster may represent countries with mixed data that don't exactly align with any other countries. Cluster 1 includes countries like Iraq, Morocco, and Saudi Arabia. This cluster has geographically diverse nations, some of which may exhibit higher communicable DALYs, and others representing higher non-communicable DALYs. Overall, this cluster represents nations that don't necessarily fit into the higher risk cluster, but also not the higher developed cluster, as many countries in this cluster are notable in the improvements that could be made to their healthcare systems. Cluster 2 includes countries like Ireland, Costa Rica, Portugal, and Australia, to name a few, all of which are economically stable countries, with strong healthcare systems. These countries are likely to have higher DALYs from non-communicable diseases or environmental vulnerabilities.

Output:
Loaded and cleaned 172 unique countries.
Graph connections:
Angola -> ["Dominica", "Sri Lanka", "Vietnam"]
Albania -> []
Andorra -> []
United Arab Emirates -> ["Burkina Faso", "Equatorial Guinea", "Mozambique", "Pakistan", "Papua New Guinea"]
Argentina -> []
Armenia -> []
Antigua and Barbuda -> ["Grenada", "Nicaragua", "Seychelles", "Uganda", "Uzbekistan"]

Australia -> []
Austria -> []
Azerbaijan -> []
Burundi -> []
Belgium -> []
Benin -> ["Mauritania"]
Burkina Faso -> ["United Arab Emirates", "Gabon", "Mozambique", "Pakistan", "Sudan"]
Bangladesh -> []
Bulgaria -> []
Bahrain -> []
Bosnia and Herzegovina -> []
Belarus -> []
Belize -> []
Bolivia -> []
Brazil -> []
Barbados -> []
Bhutan -> []
Botswana -> ["Comoros", "Guatemala", "Solomon Islands", "Zambia"]
Central African Republic -> []
Canada -> []
Switzerland -> []
Chile -> []
China -> []
Cote d'Ivoire -> []
Cameroon -> ["Tajikistan"]
Colombia -> []
Comoros -> ["Botswana", "Solomon Islands", "Zambia"]
Costa Rica -> []
Cuba -> []
Cyprus -> []
Germany -> []
Djibouti -> []
Dominica -> ["Angola", "Sri Lanka", "Uganda", "Vietnam"]
Denmark -> []
Dominican Republic -> ["Guinea"]
Algeria -> []
Ecuador -> []
Eritrea -> []
Spain -> []
Estonia -> []

Ethiopia -> []
Finland -> []
Fiji -> []
France -> []
Gabon -> ["Burkina Faso", "Kazakhstan", "Liberia", "Mozambique", "Pakistan", "Sudan"]
United Kingdom -> []
Georgia -> []
Ghana -> []
Guinea -> ["Dominican Republic", "Tajikistan"]
Guinea-Bissau -> []
Equatorial Guinea -> ["United Arab Emirates", "Papua New Guinea"]
Greece -> []
Grenada -> ["Antigua and Barbuda", "Guatemala", "Nicaragua", "Seychelles", "Uzbekistan"]
Greenland -> []
Guatemala -> ["Botswana", "Grenada", "Zambia"]
Guyana -> []
Honduras -> []
Croatia -> []
Haiti -> ["Kazakhstan", "Liberia", "Sri Lanka", "Sudan"]
Hungary -> []
Indonesia -> []
India -> []
Ireland -> []
Iraq -> []
Iceland -> []
Israel -> []
Italy -> []
Jamaica -> ["Kenya"]
Jordan -> []
Japan -> []
Kazakhstan -> ["Gabon", "Haiti", "Liberia", "Sudan"]
Kenya -> ["Jamaica", "Eswatini"]
Cambodia -> []
Kiribati -> []
Kuwait -> []
Lebanon -> []
Liberia -> ["Gabon", "Haiti", "Kazakhstan", "Sudan"]
Libya -> []
Sri Lanka -> ["Angola", "Dominica", "Haiti", "Vietnam"]
Lesotho -> []

Lithuania -> []
Luxembourg -> []
Latvia -> []
Morocco -> []
Monaco -> []
Moldova -> []
Madagascar -> []
Maldives -> ["Turkmenistan"]
Mexico -> []
Marshall Islands -> []
North Macedonia -> []
Mali -> ["Niger"]
Malta -> []
Myanmar -> []
Mongolia -> []
Mozambique -> ["United Arab Emirates", "Burkina Faso", "Gabon", "Pakistan", "Sudan"]
Mauritania -> ["Benin"]
Mauritius -> []
Malawi -> []
Malaysia -> []
Namibia -> []
Niger -> ["Mali"]
Nigeria -> []
Nicaragua -> ["Antigua and Barbuda", "Grenada", "Seychelles", "Uganda", "Uzbekistan"]
Netherlands -> []
Norway -> []
Nepal -> []
Nauru -> []
New Zealand -> []
Oman -> []
Pakistan -> ["United Arab Emirates", "Burkina Faso", "Gabon", "Mozambique", "Papua New Guinea"]
Panama -> []
Peru -> []
Philippines -> []
Palau -> []
Papua New Guinea -> ["United Arab Emirates", "Equatorial Guinea", "Pakistan"]
Poland -> []
Puerto Rico -> []
Portugal -> []

Paraguay -> []
Qatar -> []
Romania -> []
Rwanda -> []
Saudi Arabia -> []
Sudan -> ["Burkina Faso", "Gabon", "Haiti", "Kazakhstan", "Liberia", "Mozambique"]
Senegal -> []
Singapore -> []
Solomon Islands -> ["Botswana", "Comoros", "Zambia"]
Sierra Leone -> []
El Salvador -> []
San Marino -> []
Somalia -> []
Serbia -> []
Sao Tome and Principe -> []
Suriname -> []
Slovenia -> []
Sweden -> []
Eswatini -> ["Kenya"]
Seychelles -> ["Antigua and Barbuda", "Grenada", "Nicaragua", "Uganda", "Uzbekistan"]
Chad -> []
Togo -> []
Thailand -> []
Tajikistan -> ["Cameroon", "Guinea"]
Turkmenistan -> ["Maldives"]
Tonga -> []
Trinidad and Tobago -> []
Tunisia -> []
Tuvalu -> []
Tanzania -> []
Uganda -> ["Antigua and Barbuda", "Dominica", "Nicaragua", "Seychelles"]
Ukraine -> []
Uruguay -> []
United States -> []
Uzbekistan -> ["Antigua and Barbuda", "Grenada", "Nicaragua", "Seychelles"]
Vietnam -> ["Angola", "Dominica", "Sri Lanka"]
Vanuatu -> []
Samoa -> []
South Africa -> []
Zambia -> ["Botswana", "Comoros", "Guatemala", "Solomon Islands"]

Zimbabwe -> []
Afghanistan -> []
Bermuda -> []
American Samoa -> []
Montenegro -> []
South Sudan -> []

Clustered Results:
Angola - Cluster: 3
Albania - Cluster: 2
Andorra - Cluster: 1
United Arab Emirates - Cluster: 3
Argentina - Cluster: 2
Armenia - Cluster: 2
Antigua and Barbuda - Cluster: 3
Australia - Cluster: 2
Austria - Cluster: 4
Azerbaijan - Cluster: 1
Burundi - Cluster: 3
Belgium - Cluster: 4
Benin - Cluster: 3
Burkina Faso - Cluster: 3
Bangladesh - Cluster: 1
Bulgaria - Cluster: 2
Bahrain - Cluster: 4
Bosnia and Herzegovina - Cluster: 0
Belarus - Cluster: 4
Belize - Cluster: 3
Bolivia - Cluster: 1
Brazil - Cluster: 1
Barbados - Cluster: 3
Bhutan - Cluster: 2
Botswana - Cluster: 3
Central African Republic - Cluster: 1
Canada - Cluster: 4
Switzerland - Cluster: 4
Chile - Cluster: 2
China - Cluster: 0
Cote d'Ivoire - Cluster: 3
Cameroon - Cluster: 3

Colombia - Cluster: 0
Comoros - Cluster: 3
Costa Rica - Cluster: 2
Cuba - Cluster: 2
Cyprus - Cluster: 4
Germany - Cluster: 4
Djibouti - Cluster: 1
Dominica - Cluster: 3
Denmark - Cluster: 4
Dominican Republic - Cluster: 3
Algeria - Cluster: 1
Ecuador - Cluster: 2
Eritrea - Cluster: 3
Spain - Cluster: 4
Estonia - Cluster: 4
Ethiopia - Cluster: 3
Finland - Cluster: 4
Fiji - Cluster: 3
France - Cluster: 4
Gabon - Cluster: 3
United Kingdom - Cluster: 4
Georgia - Cluster: 2
Ghana - Cluster: 3
Guinea - Cluster: 3
Guinea-Bissau - Cluster: 3
Equatorial Guinea - Cluster: 3
Greece - Cluster: 4
Grenada - Cluster: 3
Greenland - Cluster: 4
Guatemala - Cluster: 3
Guyana - Cluster: 3
Honduras - Cluster: 2
Croatia - Cluster: 4
Haiti - Cluster: 3
Hungary - Cluster: 2
Indonesia - Cluster: 3
India - Cluster: 0
Ireland - Cluster: 2
Iraq - Cluster: 1
Iceland - Cluster: 2

Israel - Cluster: 4
Italy - Cluster: 4
Jamaica - Cluster: 3
Jordan - Cluster: 4
Japan - Cluster: 4
Kazakhstan - Cluster: 3
Kenya - Cluster: 3
Cambodia - Cluster: 3
Kiribati - Cluster: 0
Kuwait - Cluster: 4
Lebanon - Cluster: 0
Liberia - Cluster: 3
Libya - Cluster: 1
Sri Lanka - Cluster: 3
Lesotho - Cluster: 0
Lithuania - Cluster: 4
Luxembourg - Cluster: 4
Latvia - Cluster: 2
Morocco - Cluster: 1
Monaco - Cluster: 4
Moldova - Cluster: 3
Madagascar - Cluster: 3
Maldives - Cluster: 3
Mexico - Cluster: 0
Marshall Islands - Cluster: 3
North Macedonia - Cluster: 0
Mali - Cluster: 0
Malta - Cluster: 4
Myanmar - Cluster: 2
Mongolia - Cluster: 1
Mozambique - Cluster: 3
Mauritania - Cluster: 3
Mauritius - Cluster: 3
Malawi - Cluster: 1
Malaysia - Cluster: 2
Namibia - Cluster: 3
Niger - Cluster: 0
Nigeria - Cluster: 1
Nicaragua - Cluster: 3
Netherlands - Cluster: 4

Norway - Cluster: 2
Nepal - Cluster: 0
Nauru - Cluster: 3
New Zealand - Cluster: 4
Oman - Cluster: 3
Pakistan - Cluster: 3
Panama - Cluster: 3
Peru - Cluster: 0
Philippines - Cluster: 2
Palau - Cluster: 3
Papua New Guinea - Cluster: 3
Poland - Cluster: 4
Puerto Rico - Cluster: 1
Portugal - Cluster: 2
Paraguay - Cluster: 2
Qatar - Cluster: 4
Romania - Cluster: 2
Rwanda - Cluster: 3
Saudi Arabia - Cluster: 1
Sudan - Cluster: 3
Senegal - Cluster: 0
Singapore - Cluster: 4
Solomon Islands - Cluster: 3
Sierra Leone - Cluster: 0
El Salvador - Cluster: 3
San Marino - Cluster: 4
Somalia - Cluster: 0
Serbia - Cluster: 1
Sao Tome and Principe - Cluster: 0
Suriname - Cluster: 1
Slovenia - Cluster: 1
Sweden - Cluster: 4
Eswatini - Cluster: 3
Seychelles - Cluster: 3
Chad - Cluster: 0
Togo - Cluster: 3
Thailand - Cluster: 1
Tajikistan - Cluster: 3
Turkmenistan - Cluster: 3
Tonga - Cluster: 1

Trinidad and Tobago - Cluster: 3
Tunisia - Cluster: 2
Tuvalu - Cluster: 0
Tanzania - Cluster: 3
Uganda - Cluster: 3
Ukraine - Cluster: 2
Uruguay - Cluster: 3
United States - Cluster: 4
Uzbekistan - Cluster: 3
Vietnam - Cluster: 3
Vanuatu - Cluster: 3
Samoa - Cluster: 2
South Africa - Cluster: 3
Zambia - Cluster: 3
Zimbabwe - Cluster: 1
Afghanistan - Cluster: 3
Bermuda - Cluster: 3
American Samoa - Cluster: 3
Montenegro - Cluster: 1
South Sudan - Cluster: 3