CS 349 HW9 Write Up

Ori Zur

OPTION 1

GitHub Repo: https://github.com/ozur1/CS349_HW9

For this assignment, I chose to explore the Coronavirus data. I first used the other experiments as a template in order to pull data from 'time_series_covid19_confirmed_global.csv.' The data in this file contained the to-date total of confirmed COVID-19 cases for every day since January 22, 2020 for each of the 193 countries in the database.

Next, I ran the array of case values associated with each country through a function called 'choose_best_degree'. This function tests the accuracy of regression curves of degree 1 through degree 9 and returns the degree that produced a regression curve with the smallest mean squared error to the actual data. Using scipy.stats.mode, I was able to determine that fitting a polynomial of degree 7 returned the smallest error. Knowing this, I ran the data of each country through 'get_regression_coefficients,' passing in 7 for the degree parameter. This function uses np.polyfit with the given degree and returns the resulting coefficients. By doing this for each country, I now was able to represent each country's data with an array of 8 values, the coefficients and constant term of the regression curve.

Next, I ran all of the coefficient data through sklearn's KMeans clustering algorithm, passing in 5 for the number of clusters. I wanted to see if any pattern would emerge by the clustering of these countries. The following is the results of this experiment, which shows which countries were clustered together:

Color Key:

North America / Central America: Green

Caribbean: Light Blue

South America: Dark Yellow

Europe: Yellow

Africa: Red

Middle East / North Africa / Central Asia: Pink

East / Southeast Asia: Gray

Oceania and other: Teal

Cluster 0:

Burma, Burundi, Cambodia, Comoros, Finland, Germany, Guinea-Bissau, India, Malawi, Mali, Marshall Islands, Philippines, Russia, Sao Tome and Principe, Sierra Leone, Solomon Islands, South Sudan, Sri Lanka, Tajikistan, UAE, Yemen

Cluster 1:

Afghanistan, Algeria, Argentina, Armenia, Austria, Azerbaijan, Bahrain, Brazil, Chile, Costa Rica, Croatia, Czechia, Denmark, Diamond Princess, Dominican Republic, Ecuador, Egypt, Estonia, Georgia, Greece, Iceland, Indonesia, Iran, Iraq, Ireland, Israel, Kuwait, Latvia, Lebanon, Lithuania, Luxembourg, Mexico, Monaco, Morocco, New Zealand, North Macedonia, Norway,

Oman, Pakistan, Portugal, Qatar, Romania, San Marino, Saudi Arabia, Senegal, Slovenia, South Africa, Switzerland, West Bank and Gaza

Cluster 2:

China, Japan, Korea (south), Malaysia, Nepal, Singapore, Taiwan, Thailand, United States, Vietnam

Cluster 3:

Angola, Bahamas, Barbados, Belize, Benin, Botswana, Cabo Verde, Central African Republic, Chad, Congo, Djibouti, Dominica, El Salvador, Equatorial Guinea, Eritrea, Eswatini, Fiji, Gabon, Ghana, Grenada, Guatemala, Guinea, Haiti, Kyrgyzstan, Laos, Lesotho, Liberia, Libya, MS Zaandam, Madagascar, Mauritania, Mauritius, Montenegro, Mozambique, Nicaragua, Niger, Papua New Guinea, Saint Kitts and Nevis, Samoa, Somalia, Sudan, Sweden, Syria, Tanzania, Timor-Leste, Uganda, UK, Uzbekistan, Vanuatu, Zambia, Zimbabwe

Cluster 4:

Albania, Andorra, Antigua and Barbuda, Australia, Bangladesh, Barbados, Belgium, Bhutan, Bolivia, Bosnia and Herzegovina, Brunei, Bulgaria, Burkina Faso, Cameroon, Canada, Colombia, Congo, Cote d'Ivoire, Cuba, Cyprus, Ethiopia, France, Gambia, Guyana, Holy See, Honduras, Hungary, Italy, Jamaica, Jordan, Kazakhstan, Kenya, Kiribati, Kosovo, Liechtenstein, Maldives, Malta, Micronesia, Moldova, Mongolia, Namibia, Netherlands, Panama, Paraguay, Peru, Poland,

Rwanda, Saint Lucia, Saint Vincent and the Grenadines, Serbia, Seychelles, Slovakia, Spain, Suriname, Togo, Trinidad and Tobago, Tunisia, Turkey, Ukraine, Uruguay, Venezuela

Observations:

Cluster 2 was the easiest to see a pattern, as it contained only East and Southeast Asian countries, with the Unites States being the one exception. The majority of African countries were found in Cluster 3, though some were also in Cluster 4. Egypt and Morocco were in Cluster 1, but so were many Middle Eastern countries, suggestion that maybe an additional cluster would have separated the region of Middle East and North Africa. European and South African countries were split between Cluster 1 and Cluster 4, with no discernable difference in population or region of those continents. Caribbean countries were split between Clusters 3 and 4, with the exception of the Dominican Republic. Countries in Oceania and other Pacific islands were spread throughout every cluster.