Vanderbilt Data Science Coding Challenge Write-Up - Maya Poghosyan

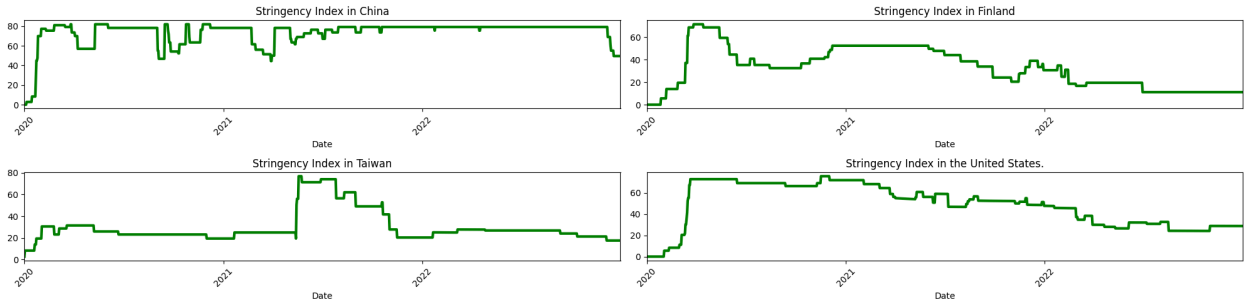
Over the course of multiple years where the COVID-19 pandemic was a prominent issue, different countries took different approaches: some more rigorously enforced quarantining, some provided more economic stimulus, and others dedicated more government resources to communities. Now that we are largely past the COVID-19 pandemic and back to normal functionality, we have a lot to learn for future outbreak management by studying these approaches and their results.

When retroactively analyzing different approaches to handling COVID-19 and their impact on infection and death rates, it could be helpful to *cluster* different countries into groups that took similar measures over time.

I predicted that the general pattern would be countries clustering around relative levels of economic development, with countries that don't have the political structure and economic strength to enforce containment policy or provide economic support to fall into different clusters than those that do. I'd expect some countries, such as China, New Zealand, and Australia, that upheld high levels of stringency for longer to be their own cluster.

I accessed a CSV database from Oxford that quantified government responses to COVID-19. Pandas was used to load in the CSV data into a DataFrame. The DataFrame had a row for each of 185 countries at each date across the 3 years recorded and columns for several different features. Some were numerical and others were flag variables, meaning they only had a few discrete values. These features were then used by the dataset authors to compute four key numerical features: StringencyIndex_Average, GovernmentResponseIndex_Average, ContainmentHealthIndex_Average, and EconomicSupportIndex. I decided to go with just these features as clustering is more accurate with fewer dimensions, meaning I dropped all the columns from the dataframe except for those containing the aforementioned 4 features, the date, and the 3-letter country code.

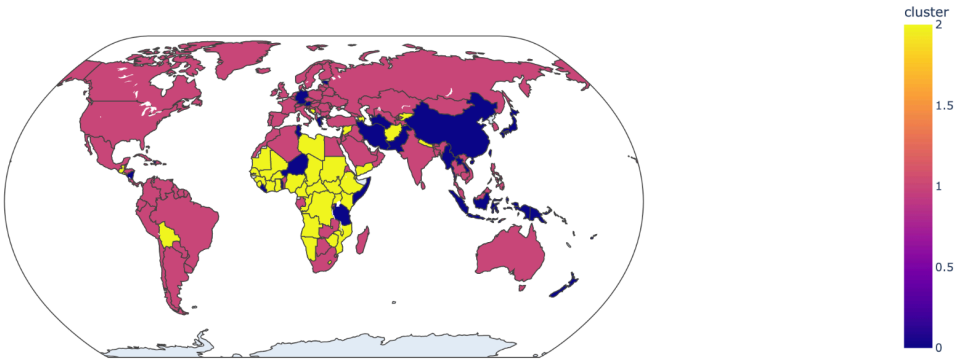
Originally, I considered averaging the indicators over the time frame and performing normal KMeans clustering. However, I realized that the actual nature of the trends over time also might impact which cluster a country falls into: did the level of these measures suddenly peak during the height of the outbreak only to relax them after, or did a country have strict measures for long durations? Therefore, I decided to challenge myself a little more and extract time series for each measure for each country across the time span of the pandemic. Here are the timeseries of the Stringency Index for four countries plotted out:



After creating a 3D array of the dimensions (number of countries, number of features per country, number of timestamps per timeseries), I used TSLearn to do time-series K-means clustering. Originally, I wanted to use Dynamic Time Warping (DTW) as the distance metric.

After running the code to perform clustering took a very long time to execute, I decided to simplify the process by using a Euclidean distance metric — this works because the time-series data is aligned in time and has the same length for every country.

The clustering was successful and outputted an array of length 185 consisting of 0s, 1s, and 2s that each corresponded to the respective country. To help visualize which country was placed into which cluster, I used Plotly to create a choropleth map that showed each map and its corresponding cluster through a color scheme, as seen below:



Because clustering is unsupervised, it is fundamentally a little less intuitive to evaluate it. However, Some things make sense — multiple of the countries with the strictest COVID restrictions (New Zealand, China, Indonesia) ended up in cluster 3. Many first-world countries such as the North American and European countries are in cluster 1. Finally, many countries in Africa are in cluster 2 — the presence of a pattern across countries that are close together in geography and human development indexes suggests a coherence to the clustering.

As for outside sources, I pulled the dataset from Oxford Covid-19 Government Response Tracker. I referenced documentation from TSLearn, Plotly Express, and I looked at the GeoPandas docs before ultimately switching to Plotly!