Clustering Countries based on COVID-19 Confirmed Cases and Indicators

Arpit Parwal, Sudeeptha Mouni Ganji, Vanessa Tan, Yeon-Soo Park
University of Southern California
INF 552 Machine Learning for Data Science - Final Project

Abstract

The COVID-19 outbreak was characterized as a pandemic by the WHO on March 11, 2020. As the new epidemic Coronavirus Disease 2019 (COVID-19) spreads, there has been a lot of research for better understanding on the effects of the virus. We can help the global community and researchers to better understand the disease by sharing insights derived from datasets which are available in Kaggle. We used the publicly available COVID-19 related datasets from sources like Johns Hopkins, United Nations and many others. Since data clustering is one of the useful unsupervised learning that groups data into clusters, we decided to use clustering methods, K-means and GMM, to group similar countries together so that we get insight into finding the similar attributes and hidden patterns of COVID-19 clusters.

Introduction and Related Work

We are attempting to analyse how we could effectively group different countries based on various features like health conditions of people, nourishment levels, testing done, access to healthcare etc. by using 2 different clustering algorithms, K-Means and Expectation Maximization using Gaussian Mixture Models. We are also trying to find out which factors are affecting Covid-19 cases the most and analyse how the top 5 factors are affecting the Covid-19 cases in different countries.

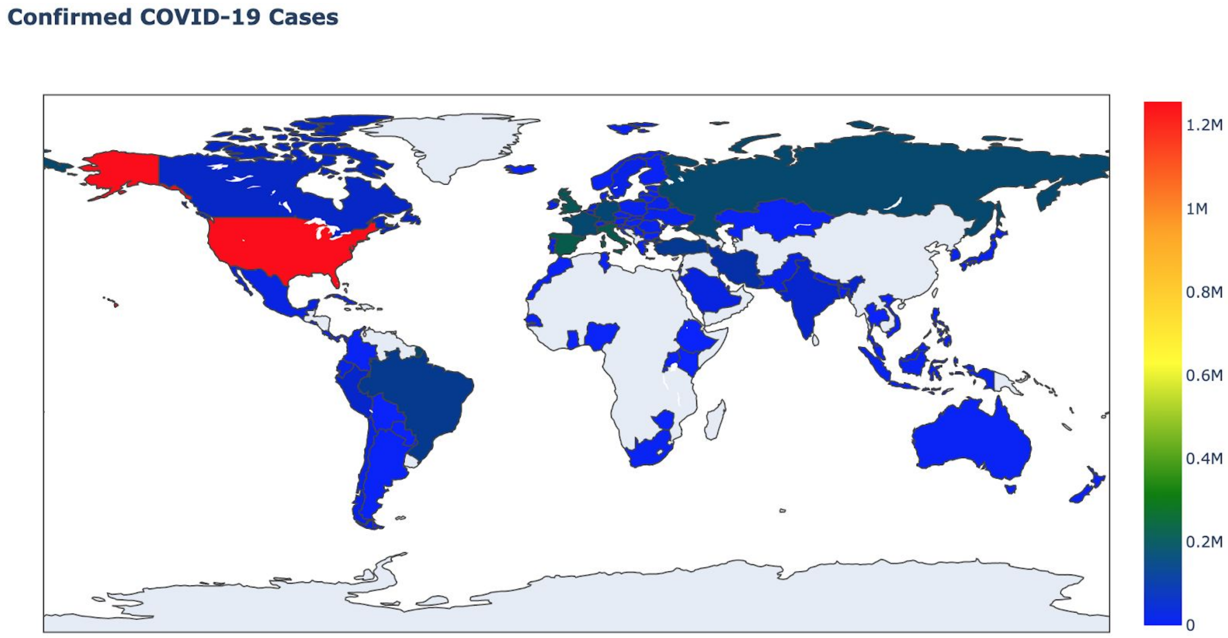


Figure 1: Map of confirmed COVID-19 Cases in different countries

The figure above shows countries grouped with respect to the number of COVID-19 cases. It can be analysed that most of the countries are in low range. It also shows the USA having the highest number of confirmed COVID-19 cases. However, clustering countries only on the basis of confirmed COVID-19 cases will be a naive approach. No conclusive and preventive measures can be concluded from such clustering. In order to address the problem of clustering countries such that it gives important information about their characteristics can help us to develop a strategy ahead of time that will help us prevent sudden outbreaks of such pandemics in the future. For instance, if our experiments find a country characteristically similar to the United States of America, we can learn from the pros and cons of the country and use it to make better policies and decisions in the other similarly grouped countries. This will help us to evade expected dangers and help policy makers design better policies and practices to keep the impact of this pandemic to the minimum.

Algorithms

Clustering is a data analysis algorithm which is used to group related data based on different factors. There are different clustering algorithms like K-Means clustering, Mean-Shift clustering, Expectation Maximization using Gaussian Mixtures, Density - based spatial clustering etc.

We have used the K-Means algorithm and Expectation Maximization using Gaussian Mixture Models for our analysis.

K-Means:

The K-Means algorithm is an iterative algorithm that tries to partition the dataset into a K-number of clusters based on data points belonging to distinct non-overlapping sub-groups. It assigns data points to a cluster such the sum of the squared distance between the data points and the cluster's centroid is minimum.The k-means algorithm works as follows:

1. Pick k centroids at random.
2. Assign each data point to the nearest centroid
3. Consider for each centroid, the data point assigned to it and recompute the centroid.
4. Repeat step 2& 3 until convergence.

The k-means algorithm has a time complexity O(n) as we are just iteratively computing the distances between points and group centers, thereby rendering it to run quickly. However, K-means is considered naive as it uses the mean value for the cluster center and the data points are circularly distributed. Additionally, it classifies every data point to belong in one cluster only. We have used sklearn's k-means library implementation in our experiment to cluster the various countries. We utilized the K-Means++ method to initialize the centroids since it tends to improve the clusters and converge more rapidly. This can be more computationally costly relative to random centroid initialization though. Random centroid initialization can be problematic though since we can get different clusters every time. With K-Means++ initialization, it specifies a procedure to initialize the clusters with before moving forward with the standard K-Means clustering algorithm. We have used sklearn's K-Means implementation in our experiment to cluster the various countries on different factors.

Expectation Maximization using GMM:

In GMMs we assume that the data points are Gaussian distributed and hence we have two parameters to describe the shape of the clusters, namely, the mean and standard deviation. As a result, they give us more flexibility than K-Means. To find the parameters for the Gaussian of each cluster, we use an optimization algorithm called Expectation-Maximization. The algorithm works as follows:

1. Select the number of clusters and randomly initialize Gaussian Distribution parameters for each cluster.

2. Compute the probability that each data point belongs to a particular cluster.
3. Based on these probabilities, we compute a new set of parameters for each Gaussian distribution so that we maximize the probabilities of data points within the cluster.
4. Repeat step 2&3 until convergence.

As we are using standard deviation as one of the parameters, GMMs give more flexibility in terms of cluster covariance than K-Means allowing the clusters to take on any ellipse shape unlike K-Means which allow only circular shapes. It also allows the data point to have a mixed membership by saying that a data point can belong X-percent to one class, Y-percent to another class etc. We have used sklearn's gmm implementation in our experiment to cluster the various countries on different factors.

Mutual information regression:

Regression allows us to estimate the relationship between a dependent variable and one or more independent variables whereas mutual information of two random variables is the measure of mutual dependence between the two variables. Mutual information when used with regression refers to the amount of information obtained about one variable through observing the other variable. We have used sklearn's mutual_info_regression on different features to understand their impact on the total number of Covid cases. The function relies on nonparametric methods based on entropy estimation from k-nearest neighbors distances. The result is then plotted on a bar chart so that we get the indicators for the Covid cases in decreasing order of impact.

Datasets

Figure 2. Description of dataset attributes.

file	Description	# of factors
Full-list-total-tests-for-covid-19.csv:	This dataset contains the country names, the country codes,the total cases in the country and date they were updated on. This is available on Kaggle now, but we used an updated version of this dataset from HDX website where Kaggle gets this data from.	4
Johns-hopkins-covid-19-05-08.csv	This dataset contains the daily case report: Country_Region, Last_Update,Lat,Long_,ISO3,Confirmed,Deaths,Recovered,Active,Incident_Rate,People_Tested,People_Hospitalized,Mortality_Rate, UID.	15
Inform-COVID-indicators.csv	This dataset contains health indicators for different countries like population density, basic sanitation services, the prevalence of malnourishment, current health expenditure per capita, access to	3

	healthcare, physician density, the population in urban areas, maternal mortality rate, etc. It was available on Kaggle, but since contributors have changed the format of the dataset, now we no longer access this file and need to combine different multiple .csv files.	
World_Population.csv	This dataset contains population by country.	1
Education_index.csv	This dataset contains both Education index and HDI index by country. HDI index (or HDI Rank) is calculated by the average achievement in key dimensions of human development: a long and healthy life, being knowledgeable and having a decent standard of living.	2
Multidimensional_poverty_index.csv	This dataset contains a poverty index by country.	1

Data Merging

Data merging is the process of combining more than two data sets into a single data set. This process is useful and important when we have raw data stored in different multiple files that we want to analyze all. Since we used different 6 data sets, the labels and data points were different. So, we cleaned all the datasets and combined them into a single csv file called **features_combined.csv**.

This is a 3-step process.

1. We set the country name as ID which is unique in all files so that we can append data in an efficient way and reduce the redundancy.
2. We used the look-up table and code to append the data. Since we combined different datasets, there were a few instances with missing values. We throw away the rows where the instances are in so that this task could increase the trust of the result accuracy.
3. We merged all data sets in to new .csv file (**features_combined.csv**)

Experimental Analysis

Elbow Method

We have initially used the elbow method in order to determine the best number of clusters to create so that the clustering is efficiently done. In this method, we find the sum of squared distances of samples to their closest cluster center, also called inertia. We do this by iteratively running the k-means algorithm with varying the number of clusters each time.

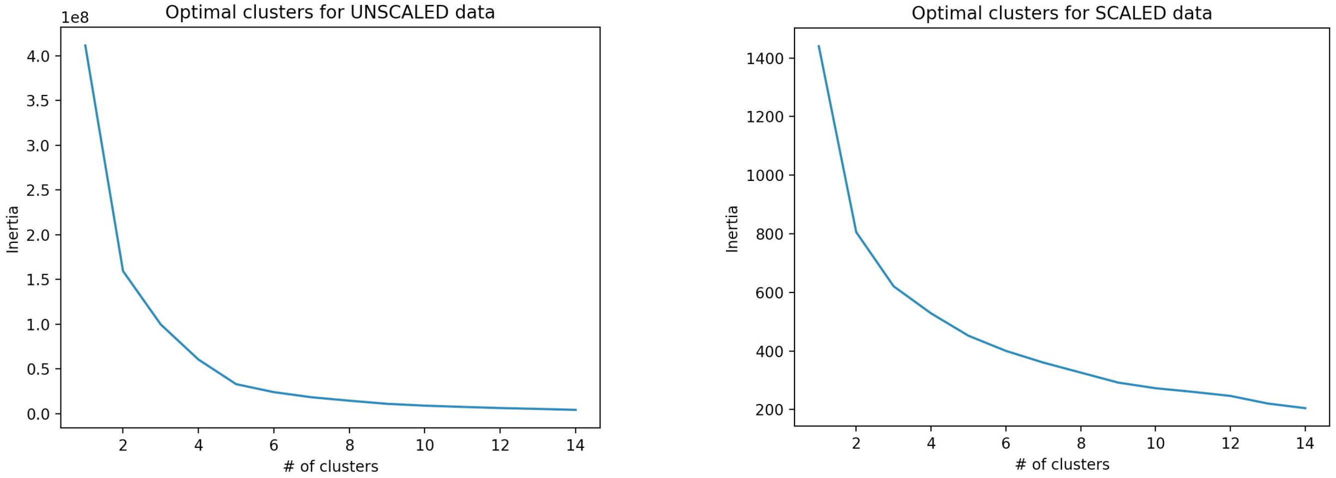


Figure 3 & 4: Elbow curves for unscaled and scaled data

We select the value of k at the elbow ie the point after which the inertia starts decreasing in a linear fashion. Based on the graph generated above, we determined 5 clusters to be the optimal number of clusters with our data.

<u>Finding top 5 factors affecting Covid cases:</u>

In order to visualize our clusters, we wanted to see which factors had the most correlation with each country's ratio of confirmed cases. We have used sklearn's mutual_info_regression to measure the dependency between these features and the number of cases across the world. On plotting the output, we got the following bar graph:

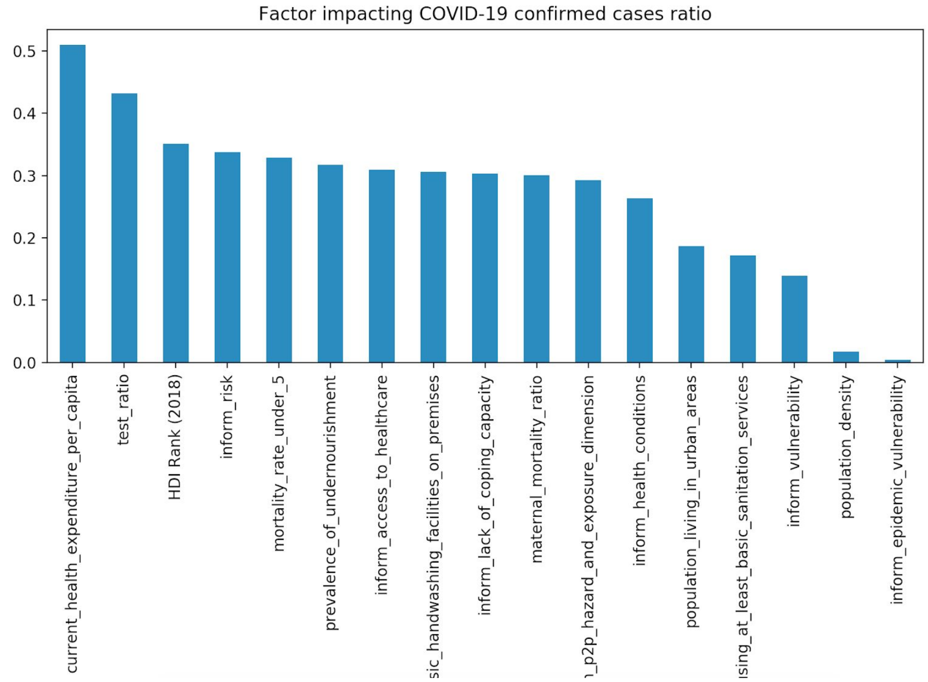


Figure 5: Mutual information regression bar chart of confirmed cases ratio against other factors

Since the current health expenditure per capita was the highest indicator of the confirmed cases ratio, we visualized our clusters with these two factors.

We used the following features from the features_combined.csv file for clustering using k-means and GMM:
'confirmed_ratio', 'test_ratio', 'HDI Rank(2018)', 'inform_risk', 'inform_p2p_hazard_and_exposure_dimension', 'population_density', 'population_living_in_urban_areas', 'proportion_of_population_with_basic_handwashing_facilities_on_premises', 'people_using_at_least_basic_sanitation_services', 'inform_vulnerability', 'inform_health_conditions', 'inform_epidemic_vulnerability', 'mortality_rate_under_5'.
We wanted to see how countries would be clustered based upon all these different features.

K-Means clustering without scaling:
We created the K-Means clusters for the countries based on all the features listed above. The plot below shows the visualization of the clusters with the unscaled data of current health expenditure per capita vs. the ratio of confirmed cases.

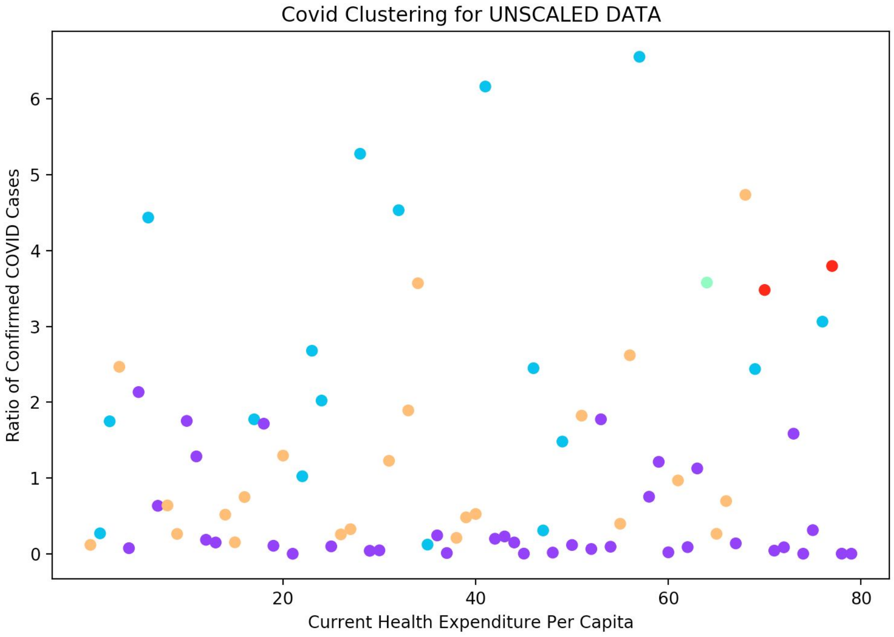


Figure 6: Clustering of unscaled data for the ratio of confirmed cases vs current health expenditure per capita

The clustering groups for the unscaled data are as follows:
Group 0 : ['Bangladesh', 'Belarus', 'Bolivia', 'Canada', 'Chile', 'Colombia', 'Costa Rica', 'Ecuador', 'El Salvador', 'Ethiopia', 'Ghana', 'India', 'Indonesia', 'Kazakhstan', 'Kenya', 'Malaysia', 'Mexico', 'Morocco', 'Nepal', 'Nigeria', 'Pakistan', 'Paraguay', 'Peru', 'Philippines', 'Romania', 'Russia', 'Rwanda', 'Senegal', 'Serbia', 'South Africa', 'Thailand', 'Tunisia', 'Turkey', 'Uganda', 'Ukraine', 'Vietnam', 'Zimbabwe']

Group 1 : ['Australia', 'Austria', 'Belgium', 'Denmark', 'Finland', 'France', 'Germany', 'Iceland', 'Ireland', 'Japan', 'Luxembourg', 'Netherlands', 'New Zealand', 'Norway', 'Qatar', 'Sweden', 'United Kingdom']

Group 2 : ['Singapore']

Group 3 : ['Argentina', 'Bahrain', 'Brazil', 'Bulgaria', 'Croatia', 'Cuba', 'Czechia', 'Estonia', 'Greece', 'Hungary', 'Iran', 'Israel', 'Italy', 'Korea, South', 'Latvia', 'Lithuania', 'Panama', 'Poland', 'Portugal', 'Saudi Arabia', 'Slovakia', 'Slovenia', 'Spain']

Group 4 : ['Switzerland', 'United States of America']

From looking at the visualization of the different clusters based on the health expenditure per capita and ratio of confirmed cases, we observed that the clusters weren't very informative. This was because the factors had a lot of variation in the magnitude of their data. We realized that this difference in magnitude also affected the way each of the factors contributed to the clustering of each group of countries. Factors with higher magnitudes had more weight on the distance calculation of clusters than factors with lower magnitudes.

<u>K-Means clustering with scaling:</u>

Since K-Means is a distance based algorithm, we had to scale the data to the same magnitude. We used sklearn's StandardScaler to scale all the features and performed kmeans again. The newly generated cluster groups were more well defined and evenly distributed. We also plotted the clusters with current health expenditure per capita vs. the ratio of confirmed cases:

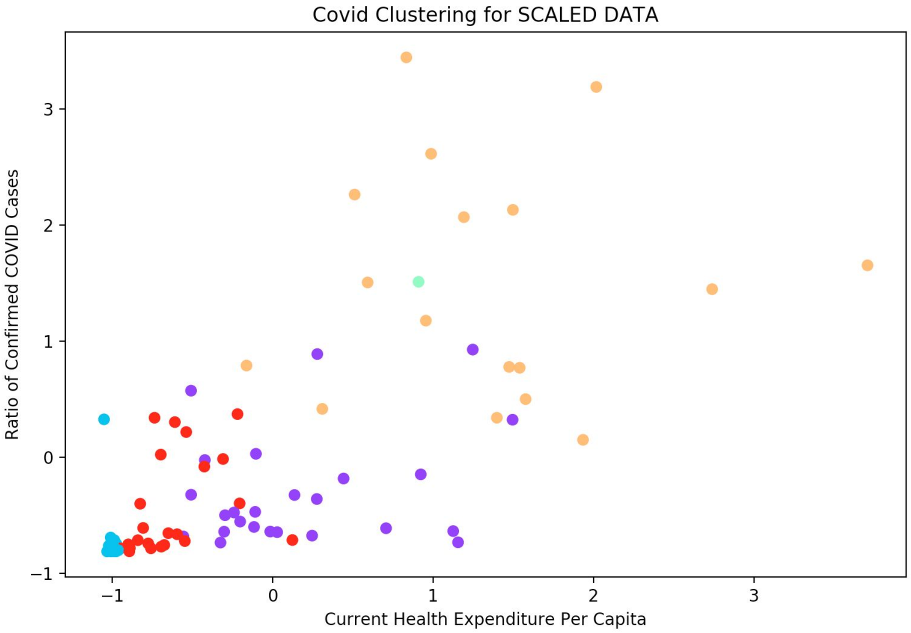


Figure 7: Clustering of scaled data for the ratio of confirmed cases vs current health expenditure per capita

The clustering groups for the scaled data from all the features are:

Group 0 : ['Argentina', 'Australia', 'Austria', 'Belarus', 'Bulgaria', 'Croatia', 'Czechia', 'Estonia', 'Finland', 'France', 'Greece', 'Hungary', 'Japan', 'Korea, South', 'Latvia', 'Lithuania', 'Malaysia', 'New Zealand', 'Poland', 'Portugal', 'Romania', 'Russia', 'Saudi Arabia', 'Slovakia', 'Slovenia']

Group 1 : ['Bangladesh', 'Canada', 'Colombia', 'Costa Rica', 'Ethiopia', 'Ghana', 'Kenya', 'Nepal', 'Nigeria', 'Pakistan', 'Rwanda', 'Senegal', 'Uganda', 'Zimbabwe']

Group 2 : ['Singapore']

Group 3 : ['Bahrain', 'Belgium', 'Denmark', 'Germany', 'Iceland', 'Ireland', 'Israel', 'Italy', 'Luxembourg', 'Netherlands', 'Norway', 'Qatar', 'Spain', 'Sweden', 'Switzerland', 'United Kingdom', 'United States of America']

Group 4 : ['Bolivia', 'Brazil', 'Chile', 'Cuba', 'Ecuador', 'El Salvador', 'India', 'Indonesia', 'Iran', 'Kazakhstan', 'Mexico', 'Morocco', 'Panama', 'Paraguay', 'Peru', 'Philippines', 'Serbia', 'South Africa', 'Thailand', 'Tunisia', 'Turkey', 'Ukraine', 'Vietnam']

We can see that after scaling all of the factors so that they had more similar magnitudes, the factors were more evenly weighted according to their correlation with the ratio of confirmed cases. The clusters based on current health expenditure per capita and ratio of confirmed cases show a pretty clear distinction. The blue cluster in the plot above shows how countries that have lower ratios of confirmed cases have lower spending on current health expenditure per capita. Similarly with the orange cluster, higher ratios of confirmed cases correlate to higher health expenditure per capita.

K-Means clustering for top 5 factors with scaling:

From looking at the bar chart of the information regression above, we can see the top 5 factors that indicated a correlation with the ratio of confirmed cases were: current health expenditure per capita, test ratio, HDI, informed risk, and mortality rate under 5.

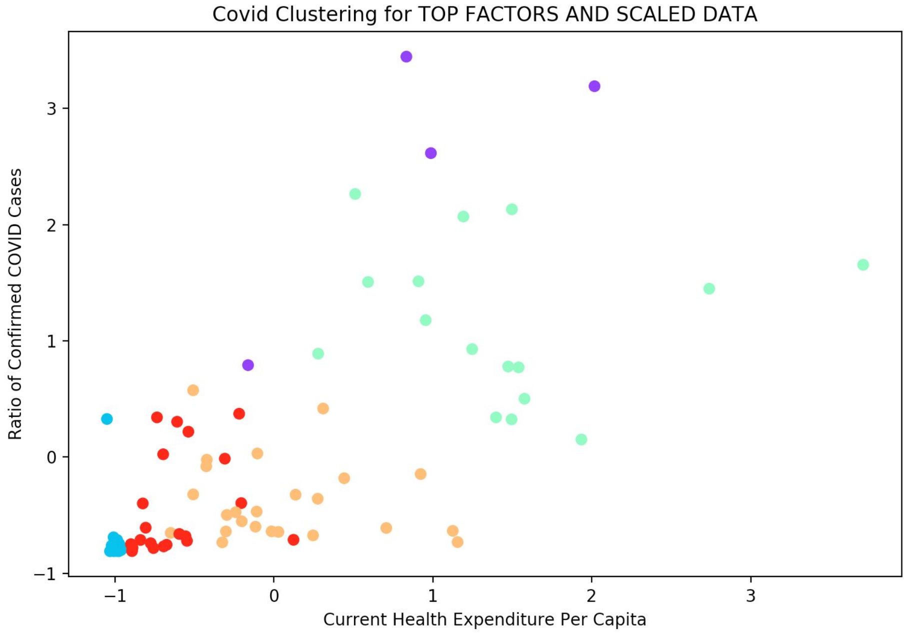


Figure 8: Clustering of scaled data with the top indicator factors for the ratio of confirmed cases vs current health expenditure per capita

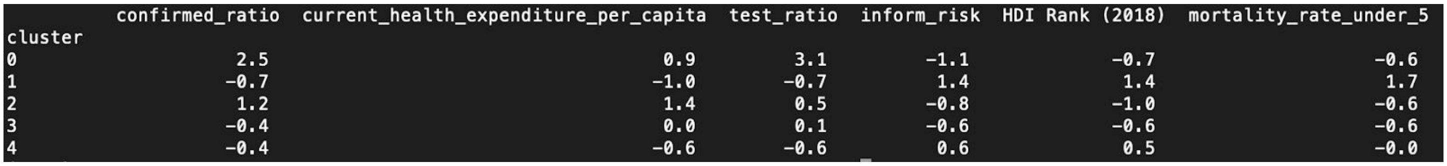
Here are the results of clustering with just the top 5 indicating factors:
Group 0 : ['Bahrain', 'Iceland', 'Luxembourg', 'Qatar']

Group 1 : ['Bangladesh', 'Canada', 'Colombia', 'Costa Rica', 'Ethiopia', 'Ghana', 'India', 'Kenya', 'Nepal', 'Nigeria', 'Pakistan', 'Rwanda', 'Senegal', 'Uganda', 'Zimbabwe']

Group 2 : ['Austria', 'Belgium', 'Denmark', 'France', 'Germany', 'Ireland', 'Italy', 'Netherlands', 'Norway', 'Portugal', 'Singapore', 'Spain', 'Sweden', 'Switzerland', 'United Kingdom', 'United States of America']

Group 3 : ['Argentina', 'Australia', 'Belarus', 'Bulgaria', 'Croatia', 'Czechia', 'Estonia', 'Finland', 'Greece', 'Hungary', 'Israel', 'Japan', 'Kazakhstan', 'Korea, South', 'Latvia', 'Lithuania', 'New Zealand', 'Poland', 'Romania', 'Russia', 'Saudi Arabia', 'Serbia', 'Slovakia', 'Slovenia']

Group 4 : ['Bolivia', 'Brazil', 'Chile', 'Cuba', 'Ecuador', 'El Salvador', 'Indonesia', 'Iran', 'Malaysia', 'Mexico', 'Morocco', 'Panama', 'Paraguay', 'Peru', 'Philippines', 'South Africa', 'Thailand', 'Tunisia', 'Turkey', 'Ukraine', 'Vietnam']



```
         confirmed_ratio  current_health_expenditure_per_capita  test_ratio  inform_risk  HDI Rank (2018)  mortality_rate_under_5
cluster
0                    2.5                                    0.9         3.1         -1.1             -0.7                    -0.6
1                   -0.7                                   -1.0        -0.7          1.4              1.4                     1.7
2                    1.2                                    1.4         0.5         -0.8             -1.0                    -0.6
3                   -0.4                                    0.0         0.1         -0.6             -0.6                    -0.6
4                   -0.4                                   -0.6        -0.6          0.6              0.5                    -0.0
```

Figure 9: Average values of the top 5 factors for each cluster

Based on the results obtained we were able to derive the following about the different cluster groups:

- Countries in Group 0 represent countries with highest confirmed cases ratio, high current per capita health expenditure, highest test ratio, lowest information risk, low human development index and lowest mortality rate under 5. While the low information risk could have contributed to an increase in cases, the high current per capita health expenditure and high test ratio could be because of the high number of confirmed cases in these countries.
- Countries in Group 1 represent countries with lowest confirmed cases ratio, lowest current per capita health expenditure, lowest test ratio, highest information risk, highest human development index rank and highest mortality rate under 5. The high information risk could have contributed to a lower number of confirmed cases resulting in lower test ratio.
- Countries in Group 2 represent countries with high confirmed cases ratio, highest current per capita health expenditure, moderate test ratio, lowest information risk, lowest human development index rank and lowest mortality rate under 5. The low information risk could have contributed to an increase in cases resulting in an more than average increase in testing ratio.
- Countries in Group 3 represent countries with low confirmed test cases ratio, moderate current per capita health expenditure, moderate test ratio, low information risk, low human development index rank and lowest mortality rate under 5.
- Countries in Group 4 represent countries with low confirmed test cases ratio, low current per capita health expenditure, low test ratio, high information risk, high low human development index rank and low mortality rate under 5.

Groups 0 and 2 were characterised by a very high confirmed cases ratio but had the lowest information risk and lowest human development index ranks despite having highest current per capita health expenditures. Group 1 had the lowest confirmed cases ratio despite having lowest current per capita health expenditure by having highest information risk. Groups 3 and 4 were similar in their confirmed cases ratio but Group 3 had a higher health expenditure and testing ratio but quite lower information risk, human development index and mortality rate under 5 compared to Group 4.

Baseline Comparison

In addition to the K-Means algorithm we also tried experimenting with another baseline. K-Means is actually a special case of GMM in which each cluster's covariance along all dimensions approaches 0. This implies that a point will get assigned only to the cluster closest to it. With GMM, each cluster can have unconstrained covariance structure.

Comparisons of clustering on scaled data with highest contributing factors:

We set up a similar environment and used the same refined data considering only the top factors that contributed to the number of deaths due to COVID-19 to run it with GMM algorithm. This will help us to analyse and compare results from both K-Means and GMM.

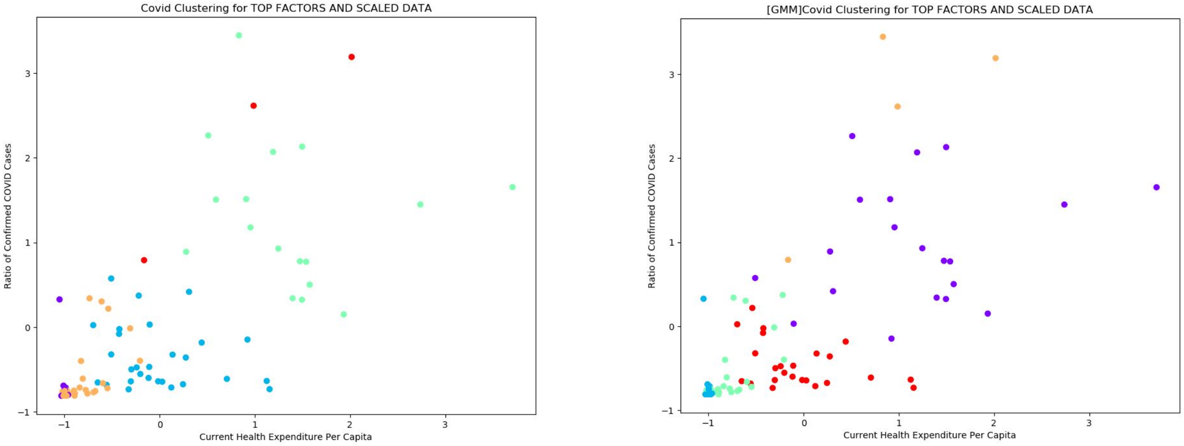


Figure 10 & 11: Plotting of clusters generated by using K-Means(left) and GMM(right)

Collating results on World map:

Using GMM algorithms to cluster countries, we found out that it produced identical results apart from a few exceptions.

For example countries like Qatar and the United States were added to clusters of countries like Bahrain, Iceland and Luxembourg. Clustering countries like the USA and Bahrain together is odd.

Furthermore, clustering using GMM will help us in the verification of the clusters we already have.

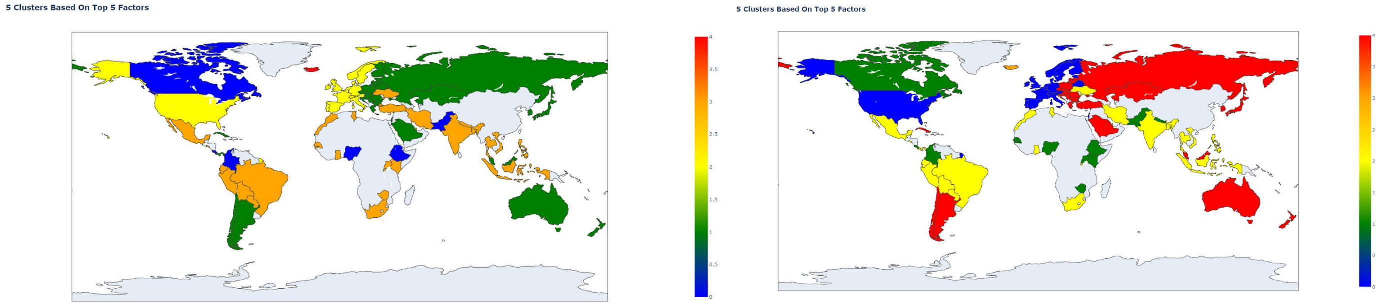

5 Clusters Based On Top 5 Factors
5 Clusters Based On Top 5 Factors

Figure 12 & 13: World Map Visualisation of clustering using K-Means (left) and GMM (right)

<u>GMM Clusters:</u>

Similarly in Group 1 countries like Canada and Pakistan are grouped together. As we can see, clustering in GMM is somewhere incorrect when checked with real facts and data, we can evidently conclude that K means will provide us with better and case specialized clustering of countries. Apart from a few exceptions, GMM gave us similar results as K means. This indicates that the data can be clustered using KMeans and no further exploration of baselines are required.

The final analysis of the group as found by GMM algorithm are as follows:

Group 0 : ['Argentina', 'Brazil', 'Bulgaria', 'Chile', 'Croatia', 'Cuba', 'Ecuador', 'Greece', 'Hungary', 'Iran', 'Kazakhstan', 'Korea, South', 'Malaysia', 'Panama', 'Peru', 'Poland', 'Romania', 'Russia', 'Saudi Arabia', 'Serbia', 'Slovakia', 'Turkey']

Group 1 : ['Canada', 'Costa Rica', 'Nigeria', 'Pakistan']

Group 2 : ['Bahrain', 'Iceland', 'Luxembourg', 'Qatar', 'United States of America']

Group 3 : ['Bangladesh', 'Bolivia', 'Colombia', 'El Salvador', 'Ethiopia', 'Ghana', 'India', 'Indonesia', 'Kenya', 'Mexico', 'Morocco', 'Nepal', 'Paraguay', 'Philippines', 'Rwanda', 'Senegal', 'South Africa', 'Thailand', 'Tunisia', 'Uganda', 'Ukraine', 'Vietnam', 'Zimbabwe']

Group 4 : ['Australia', 'Austria', 'Belarus', 'Belgium', 'Czechia', 'Denmark', 'Estonia', 'Finland', 'France', 'Germany', 'Ireland', 'Israel', 'Italy', 'Japan', 'Latvia', 'Lithuania', 'Netherlands', 'New Zealand', 'Norway', 'Portugal', 'Singapore', 'Slovenia', 'Spain', 'Sweden', 'Switzerland', 'United Kingdom']

Conclusion

This experiment identified the top factors that indicated a higher ratio of confirmed COVID-19 cases: current health expenditure per capita, test ratio, HDI, informed risk, and mortality rate under 5. We showed the necessity of scaling the data to similar magnitudes in order to give the different factors an even weight in clustering. By looking at the elbow curve of the sum of squared differences between data points and their clusters, we were able to find 5 clusters to be the optimal number of clusters for our data. The mutual information regression of the ratio of confirmed cases against the other factors allowed us to identify current health expenditure per capita and the ratio of confirmed cases as the best factors to visualize our clusters on. We utilized K-Means clustering by scaling the top factors in order to visually group countries and analyze the similarities between them. We were able to recognize the patterns among the groups by finding the averages of these factors to compare the clusters. Furthermore, we used GMM clustering as an additional baseline and found related clustering results, conclusive enough that further exploration of dissimilar baselines and features can help in more explicit and generalised clusters of countries.

Future scope

During our current research, the data from few of the countries was unavailable as the number of cases are changing rapidly and on a daily basis. Once the complete data is available, we could use the clusters to implement policies in different countries under a single group so as to help control the current outbreak and prevent such outbreaks in the future.

1. **Generalisation of the result to other relatable diseases**
   There can be a lot of research done on how we can use the data and clusters generated to predict the effect of other relatable diseases like COVID-19. Investigating the response of the clustering result of COVID-19 to other viruses like MERS, SERS can help us theorize association of countries that can have a similar response to an outbreak
2. **Exploring other baselines and other features**
   K-Means algorithm has provided us with satisfactory results. The project also explored other baselines like GMM and found similar result.However, since K-Means is a special case of GMM, there exists a scope of exploring other baselines such as DBSCAN, Spectral and Mean shift algorithms.
3. **Clustering size K**
   Clustering algorithms can be improved by using an optimal number of clusters. Elbow curve is a very useful method when we set the number of K, but if data is too clustered, it can be hard to find a clear elbow. So, we can try a different method for finding optimal K, when we build the model.

References

[1] Imad Dabbura, (Sep 17, 2018), *K-means Clustering: Algorithm, Applications, Evaluations, Methos, and Drawbacks.* Retrieved from https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation- methods-and-drawbacks-aa03e644b48a

[2] M Syakur-B Khotimah-E Rochman-B Satoto - IOP Conference Series: Materials Science and Engineering - 2018, *Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster*

[3] M. A. Sulaiman and J. Labadin, "Feature selection with mutual information for regression problems," *2015 9th International Conference on IT in Asia (CITA)*, Kota Samarahan, 2015, pp. 1-6, doi: 10.1109/CITA.2015.7349826.

[4] The Human Data Exchange, The Covid-19 Tests Performed by Country 2020 [Data file]. Retrieved from https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases

[5] Human Development Report, Human Development Data (1990-2018) [Data file]. Retrieved from http://hdr.undp.org/en/data

[6] COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University, JHU CSSE COVID-19 Dataset [Data file]. Retrieved from https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data

[7] Kaggle, Total Covid-19 Tests Performed By Country [Data file]. Retrieved from https://www.kaggle.com/roche-data-science-coalition/uncover#total-covid-19-tests-performed-by-country.csv

[8] Scikit Learn, *K-means Clustering API reference,* Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.mutual_info_regression.html

[9] P. Trebuňa, J. Halčinová, M. Fil'o and J. Markovič, "The importance of normalization and standardization in the process of clustering," *2014 IEEE 12th International Symposium on Applied Machine Intelligence and Informatics (SAMI)*, Herl'any, 2014, pp. 381-385, doi: 10.1109/SAMI.2014.6822444.
[10] M. Hamouz, J. Kittler, J. -. Kamarainen, P. Paalanen and H. Kalviainen, "Affine-invariant face detection and localization using GMM-based feature detector and enhanced appearance model," Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings., Seoul, South Korea, 2004, pp. 67-72, doi: 10.1109/AFGR.2004.1301510.
[11] United Nations, World Population Data 2020 [Data file]. Retrieved from https://worldpopulationreview.com.

List of Figures

Individual Contribution

Student Name / Class	**Contribution**
Arpit Parwal aparwal@usc.edu USC ID: 5382582950 Tuesday	● GMM Algorithm ● Data Collection ● Cluster Averages ● Report Writing
Sudeeptha Mouni Ganji sganji@usc.edu USC ID: 2942771049 Monday	● GMM Algorithm ● Elbow curve visualization ● Data Analysis ● Report Writing

Vanessa Tan tanvanes@usc.edu USC ID: 4233243951 Monday	● K-Means Algorithm ● Scaling dataset ● Mutual information regression ● Report Writing
Yeon-Soo Park **yeonsoop@usc.edu** **USC ID: 1240112911** **Tuesday**	● K-Means Algorithm ● Combining datasets ● Data visualization ● Report Writing