A Dynamic Clustering of COVID-19-Affected Countries

Eric Lin, Eric Young

April 20, 2020

Problem Statement

In late December 2019, COVID-19 emerged in Wuhan, China as a novel coronavirus that causes symptoms of fever, cough, and shortness of breath. In the matter of a few months, this highly contagious disease has made its way across the borders of 210 countries, infecting more than 2.3 million people and taking away more than 160,000 lives. Since COVID-19 was first detected in the United States in January 2020, it has spread to at least 750,000 people around the country. In response to government-issued regional lockdowns and social distancing policies, American universities followed the norm to transition on-campus in-person courses to off-campus online courses in order to slow the spread of COVID-19. In particular, Duke University quickly restricted campus access and ordered students to travel back home before an inevitable daunting virus outbreak. As a top higher education destination for international students, Duke University's student body is represented by 51 different countries (Figure 1). With new cases and deaths surging around the world while international students return back home, this paper aims to utilize the k-means clustering algorithm to group the 51 countries represented by Duke students into characteristic clusters based on COVID-19 statistics and health indicators. This information can help the Duke University administration closely monitor the potential trajectory of the virus, evaluate the risks its international students might face, and allocate appropriate resources to them if applicable. Given that more than a fifth of Duke University's student population come from countries outside of the United States, this in-depth statistical analysis can also assist the university in making future decisions regarding the status of the Fall 2020 semester.

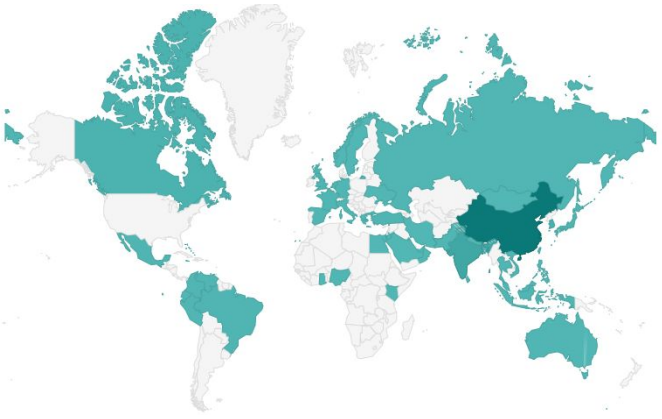

Figure 1: Duke University international students by home country

Summary Statistics and Data Analysis

Attribute data regarding COVID-19 statistics and health indicators were independently retrieved from sources including the WHO, World Bank, and Johns Hopkins. Subsequently, the data was compiled into a machine-readable comma-separated values (csv) file for data preprocessing in Python 3. The final.csv file (Figure 2) contains the following feature attributes:

A	B	C	D	E	F	G
Country Name	Total Cases / 1M pop	Deaths / 1M pop	Tests / 1M pop	Hospital Beds / 1K pop	UHC Service Coverage Index	Smoking Prevalence (% of total adult population)
Australia	260	3	16931	3.9	87	14.7
Bahamas	153	23	250	3.1	75	11.5
Bangladesh	18	0.6	162	0.8	48	23
Brazil	182	12	296	2.4	79	13.9
Canada	929	42	14555	2.7	89	14.3
China	57	3	1000	3.56	79	25.6
Colombia	75	4	1233	1.5	76	9
Ecuador	537	27	1803	1.6	77	7.1
Egypt	31	2	537	1.7	68	25.2
France	2342	302	7103	6.6	78	32.7
Germany	1740	55	20629	8.2	83	30.6
Ghana	34	0.3	2207	4.8	75	43.4
Greece	214	11	5113	4.8	75	43.4
Hong Kong	137	0.5	17579	4.9	89	10
India	13	0.4	291	0.7	55	11.5
Indonesia	25	2	154	0.6	57	39.4
Iran	979	61	4068	1.7	72	11
Israel	1577	20	27763	3.2	82	25.2
Italy	2960	391	22436	3.6	82	23.7
Jamaica	66	2	611	1.8	65	16.8
Japan	85	2	892	13.4	83	22.1
Jordan	41	0.7	2842	1.8	76	42.2
Kenya	5	0.3	246	1.4	55	10.7
Kuwait	467	2	4000	2	76	22.5
Malaysia	168	3	3210	1.8	73	21.5
Mexico	64	5	384	1.7	76	14
Mongolia	10	0	474	5.8	62	25.6
Nepal	1	0	1015	0.3	48	22.8
Netherlands	1906	215	9041	4.7	86	25.8
Nigeria	3	0.1	35	0.5	42	5.8
Norway	1310	30	26224	4.3	87	20.2
Oman	276	1	2500	1.8	69	11.1

Figure 2: final.csv

A. Country Name:
This attribute shows the names of each of the 51 countries represented by Duke students in alphabetical order.

I. COVID-19 Statistics

B. <u>Total Cases Per 1 Million Population</u>
This attribute shows the extent of the spread of COVID-19 in each of the countries examined. Note that the "Per 1 Million Population" serves to normalize the disproportionately larger number of cases countries with high populations detect. By dividing the total number of cases a country has with the number of millions of the country's population, an objective comparison among COVID-19's spread in different countries can be made. This normalization approach is also applied in attributes C and D.

Mean: 610.0784313725491
Min: 1 (Nepal)
Max: 4282 (Spain)
Median: 168.0
Standard Deviation: 901.6953714354576

C. <u>Total Deaths Per 1 Million Population</u>
This attribute shows the lethality of COVID-19 in each of the countries examined. The number of deaths caused by COVID-19 can be traced back to the quality of healthcare, effectiveness of government policies, and population dynamics of a given country.

Mean: 43.94705882352942
Min: 0.0 (Mongolia)
Max: 446.0 (Spain)
Median: 3.0
Standard Deviation: 98.75323842382396

D. Total Tests Per 1 Million Population

This attribute shows the extent of COVID-19 testing in each of the countries examined. Given the asymptomatic nature of a sizable portion of those infected with COVID-19, more testing can mean more cases detected in the short term, but lower cases detected in the long run as those diagnosed with the disease will likely be quarantined until they recover. Note that countries that are more vulnerable to COVID-19's spread and already have more cases are incentivized to test more people due to the exponential spread of the disease. Conversely, countries that are less affected by COVID-19 will likely test less.

Mean: 7810.823529411765
Min: 35 (Nigeria)
Max: 77550 (United Arab Emirates)
Median: 2842.0
Standard Deviation: 12368.859016134633

II. Health Indicators

E. Hospital Beds Per 1 Thousand Population

This attribute shows the availability of hospital beds for potential patients for each of the countries examined. The exponential spread of COVID-19 puts a significant strain on hospital resources. Countries with more hospital beds will likely have a comparatively easier time treating patients as they won't need to choose which patients to admit when there are not enough hospital beds.

Mean: 3.293333333333333
Min: 0.3 (Nepal)
Max: 13.4 (Japan)
Median: 2.4
Standard Deviation: 2.787666581088969

F. UHC Service Coverage Index

This attribute shows the coverage of essential health services (reproductive, maternal, newborn and child health, infectious diseases, non-communicable diseases and service capacity and access) for each of the countries examined. Ranging from 0 to 100, this index score serves as an indicator of a country's healthcare quality.

Mean: 73.90196078431373
Min: 42 (Nigeria)
Max: 89 (Canada)
Median: 76.0
Standard Deviation: 11.84914144645383

G. Smoking Prevalence (Percentage of Total Population)
This attribute shows the percentage of a country's adult population (both genders above age 15) that identify as tobacco smokers. This is relevant to our analysis because research suggests that smokers are more prone to serious complications if infected with COVID-19. This leads experts to speculate about the potential correlation between smoking prevalence and COVID-19 lethality rate.

Mean: 21.67254901960784
Min: 4.8 (Peru)
Max: 43.4 (Ghana)
Median: 22.1
Standard Deviation: 9.491483869844517

We next examine some notable relationships that arise between certain attributes.

1. Total Deaths Per 1 Million Population v.s. Total Cases Per 1 Million Population (Figure 3)

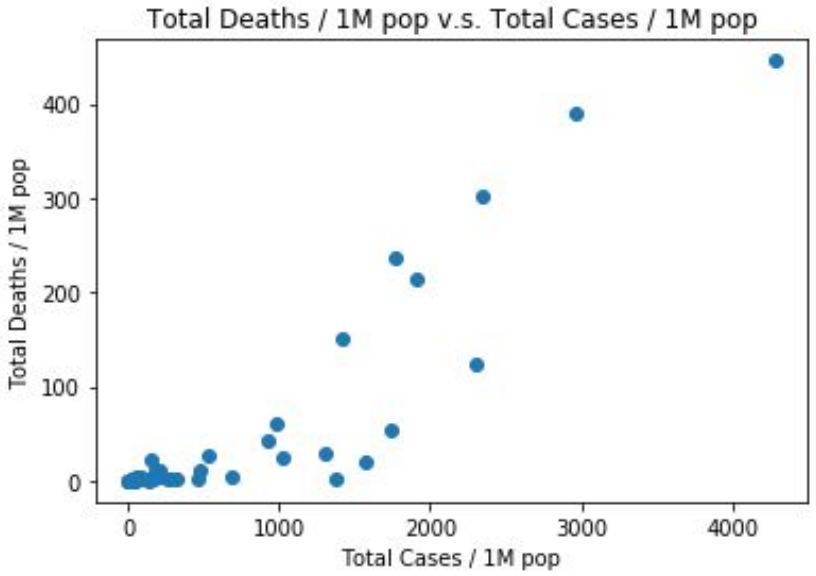

Figure 3: R^2 = 0.8021951570197311

Quite intuitively, a larger number of cases increases the likelihood of a larger number of deaths. This positive correlation is affirmed by the high R^2 value above.

2. Total Cases Per 1 Million Population v.s. Total Tests Per 1 Million Population (Figure 4)

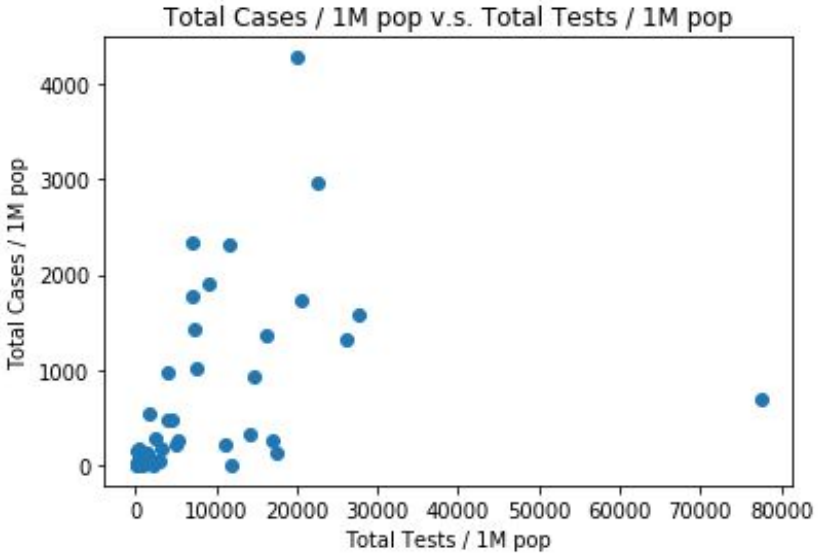

Figure 4

Excluding one outlier, the overall positive correlation of this relationship confirms our reasoning that an increase in tests will increase the number of cases considering the asymptomatic nature of many infected with COVID-19.

3. Total Deaths Per 1M Population v.s. Smoking Percentage (Figure 5)

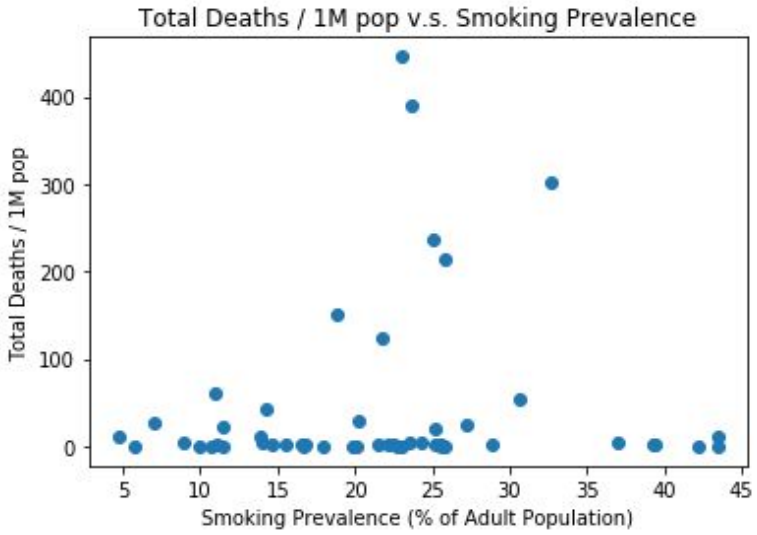

Figure 5

While many data points don't show any trends, a portion of the data shows a strong linear trend, which exhibits the higher detriments COVID-19 may have on smokers. However, it is important to note that correlation does not imply causation as lurking variables may pose extraneous influences on the relationship.

Methods

I. Theory

We employed the k-means clustering algorithm to efficiently group the 51 countries into multiple clusters with distinct characteristics. The general idea of the k-means clustering algorithm is to partition points into k-clusters c1, c2, …, ck and find the corresponding k center points z1, z2, …, zk in order to minimize the cost function, $\sum_{i=1}^{k} \sum_{x \in ci} ||x - zi||^2$, where x represents each data point.

More precisely, the k-means clustering algorithm (Figure 6) follows the following 3 iterative steps:

1. Initialization:
   To start the algorithm, randomly select k data points from the dataset as initial centroids. Note that centroids are the centers of clusters. In this case, since we have no information indicating the location of cluster centers, we arbitrarily define them.
2. Cluster Assignment:
   Calculate the Euclidean distances of each data point from each centroid, assigning all the data points that are closest to a centroid to a cluster.
3. Centroid Movement:
   Calculate the new centroid of each cluster, which is the mean of all data points in the cluster. Repeat steps 2 and 3 until the algorithm converges to acquire the final clustering assignments.

Lastly, we used the elbow method, which runs the k-means clustering algorithm with different numbers of clusters and finds the within-cluster sum of squares (measures within-cluster variance), to determine the ideal number of clusters to choose.

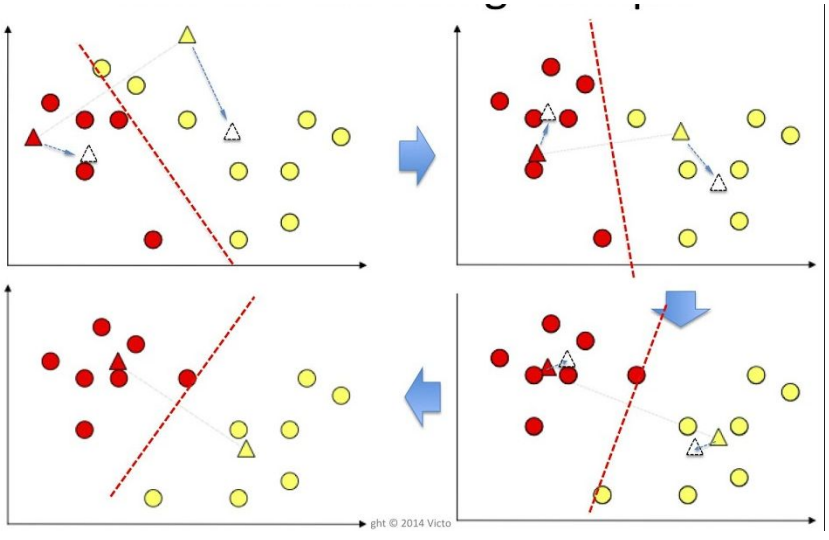

Figure 6: A step-by-step illustration of the k-means clustering algorithm

II. Python 3 Implementation

We applied the k-means clustering algorithm in Python 3, implemented by the scikit-learn package.

1. Data Preprocessing

Upon data mining, we normalized the data (rescaled numeric values into the range 0 and 1) to ensure that every attribute contributed equally to determining the distance between points. We chose normalization over standardization because our input attributes rely on the magnitude of values instead of the distribution of Gaussian processes.

```
# implement k-means clustering
from sklearn import preprocessing
# initialize the 6D list of attributes
attributes = []
for i in range(len(cases)):
    attributes.append([cases[i], deaths[i], tests[i], beds[i], UHC[i], smoking[i]])
# normalization
normalized_attributes = preprocessing.normalize(attributes)
```

2. Selecting the Ideal Number of Clusters (Elbow Plot)

To select the ideal number of clusters for the k-means clustering algorithm to find, we use the elbow method and notice that "elbow of the curve" is around 5 clusters (Figure 7). To minimize the within-cluster sum of squares error while drawing useful conclusions, we choose 5 clusters for our algorithm.

```
# Elbow Plot
from sklearn.cluster import KMeans
Error =[]
for i in range(1, 11):
    kmeans = KMeans(n_clusters = i).fit(normalized_attributes)
    kmeans.fit(normalized_attributes)
    Error.append(kmeans.inertia_)
plt.plot(range(1, 11), Error)
plt.title('Elbow Plot')
plt.xlabel('No of clusters')
plt.ylabel('Error')
plt.show()
```

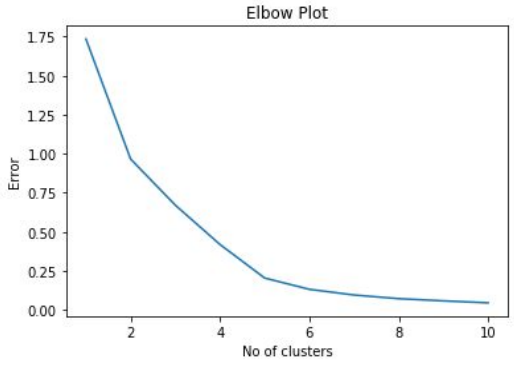

Figure 7: Note that the elbow of the curve is around 5 clusters

3. Implementing the K-Means Clustering Algorithm

We proceed to implement the k-means clustering algorithm with 5 clusters and a convergence criterion of 0.001. Note that the convergence criterion, which determines when iteration ceases, represents a proportion of the minimum distance between initial cluster centers.

```
119 # apply kmeans with k = 5 and a convergence criterion of 0.001
120 kmeans5 = KMeans(n_clusters=5, tol=0.001)
121 clusters = kmeans5.fit_predict(normalized_attributes)
```

Results

Cluster Assignments:
[0 2 1 2 0 0 0 4 0 4 0 0 0 0 1 1 4 0 0 0 0 0 1 0 0 1 0 0 4 3 0 0 0 0 0 0 0
 0 0 4 1 4 0 0 0 0 0 4 4 0 0]

Group 0:

Countries	Summary Statistics	Shared Characteristics
Australia Canada China Colombia Egypt Germany Ghana Greece Hong Kong Israel Italy Jamaica Japan Jordan Kuwait Malaysia Mongolia Nepal Norway Oman Pakistan Peru Philippines Russia Saudi Arabia Singapore South Korea	1. Total Cases / 1M pop Mean: 443.79411764705884 SD: 653.0611978985522 2. Total Deaths / 1M pop Mean: 18.664705882352937 SD: 66.0027890781137 3. Total Tests / 1 M pop Mean: 9651.970588235294 SD: 14304.365912063133 4. Hospital Beds / 1K pop Mean: 3.7370588235294115 SD: 3.1412479062156966 5. UHC Service Coverage Index Mean: 75.38235294117646 SD: 10.324199743613098 6. Smoking Prevalence Mean: 23.073529411764707 SD: 9.409204155366714	- Moderate # of cases - Moderate # of deaths - High # of tests - High # of hospital beds - Moderate UHC - High smoking prevalence

Taiwan Thailand Turkey Ukraine United Arab Emirates Venezuela Vietnam		

Group 1:

Countries	Summary Statistics	Shared Characteristics
Bangladesh India Indonesia Kenya Mexico Sri Lanka	1. Total Cases / 1M pop Mean: 23.166666666666668 SD: 19.21298750556219 2. Total Deaths / 1M pop Mean: 1.4333333333333336 SD: 1.7016332024133625 3. Total Tests / 1 M pop Mean: 256.5 SD: 80.68818583493703 4. Hospital Beds / 1K pop Mean: 1.4666666666666668 SD: 1.032257504480135 5. UHC Service Coverage Index Mean: 59.5 SD: 9.069178573608527 6. Smoking Prevalence Mean: 20.733333333333334 SD: 10.087230651780608	- Low # of cases - Low # of deaths - Moderate # of tests - Low # of hospital beds - Low UHC - High smoking prevalence

Group 2:

Countries	Summary Statistics	Shared Characteristics

Nigeria	1. Total Cases / 1M pop Mean: 3.0 SD: 0.0 2. Total Deaths / 1M pop Mean: 0.1 SD: 0.0 3. Total Tests / 1 M pop Mean: 35.0 SD: 0.0 4. Hospital Beds / 1K pop Mean: 0.5 SD: 0.0 5. UHC Service Coverage Index Mean: 42.0 SD: 0.0 6. Smoking Prevalence Mean: 5.8 SD: 0.0	- Low # of cases - Low # of deaths - Low # of tests - Low # of hospital beds - Low UHC - Low smoking prevalence

Group 3:

Countries	Summary Statistics	Shared Characteristics
Ecuador France Iran Netherlands Spain Sweden United Kingdom United States	1. Total Cases / 1M pop Mean: 1943.5 SD: 1058.9625819640655 2. Total Deaths / 1M pop Mean: 195.375 SD: 127.39008742833957 3. Total Tests / 1 M pop Mean: 8508.125 SD: 5121.334504733605 4. Hospital Beds / 1K pop Mean: 3.2625 SD: 1.544293932514144 5. UHC Service Coverage	- High # of cases - High # of deaths - High # of tests - High # of hospital beds - High UHC - High smoking prevalence

	Index Mean: 81.625 SD: 5.02338282435253 6. Smoking Prevalence Mean: 20.65 SD: 7.726901060580497	

Group 4:

Countries	Summary Statistics	Shared Characteristics
Bahamas Brazil	1. Total Cases / 1M pop Mean: 167.5 SD: 14.5 2. Total Deaths / 1M pop Mean: 17.5 SD: 5.5 3. Total Tests / 1 M pop Mean: 273.0 SD: 23.0 4. Hospital Beds / 1K pop Mean: 2.75 SD: 0.3500000000000001 5. UHC Service Coverage Index Mean: 77.0 SD: 2.0 6. Smoking Prevalence Mean: 12.7 SD: 1.2000000000000002	- Moderate # of cases - Moderate # of deaths - Moderate # of tests - Moderate # of hospital beds - High UHC - Moderate smoking prevalence

Conclusions

Upon mining, preprocessing, and analyzing (summary statistics, variable relationships) a compiled COVID-19 dataset containing 6 numeric attributes (total cases / 1M population, total deaths / 1M population, total tests / 1M population, total beds / 1K population, UHC service coverage index, smoking prevalence) corresponding to each of 51 countries, we were able to successfully implement the k-means clustering algorithm to group the countries Duke students come from into 5 distinct clusters..

References

"Coronavirus Cases:" *Worldometer*, www.worldometers.info/coronavirus/#countries.

D, F. "Clustering Using K-Means Algorithm." *Medium*, Towards Data Science, 19 Dec. 2017, towardsdatascience.com/clustering-using-k-means-algorithm-81da00f156f6.

"Indicator Metadata Registry Details." *World Health Organization*, World Health Organization, www.who.int/data/gho/indicator-metadata-registry/imr-details/4834.

"International Students by Country Duke University." *College Factual*, 7 Feb. 2020, www.collegefactual.com/colleges/duke-university/student-life/international/chart-international.html.