COVID-19 Statistical Clustering

IAN ST. JOHN, George Mason University

In this paper, k-means clustering will be used to see if there is a correlation between different countries data regarding the COVID-19 pandemic, and where the United States places giving different clustering methods and metrics. The data set provided by Our World in Data [3] uses 46 features to classify COVID-19 data from 193 countries between the dates of 1/22/2020 to 12/1/2020.

1 INTRODUCTION

COVID-19 has killed more than one and a half million people worldwide at the time of writing. Being able to analyze the infection and categorize different outcomes could lead to better handling of pandemics in the future. While this research is not going to be able to provide hard statistical information on the spread of the virus due to the methodology and due to the fact that the pandemic is still going on, it may provide a jumping off point for future research and investigations into the handling of COVID-19 by various countries.

1.1 Purpose

Different countries and localities all around the world have had different responses to the virus; some nations were very strict and had heavy lock downs, others were more relaxed with their public health policies. The purpose of this paper is not to determine whether any particular country had a better response to the virus, only to categorize the countries together based on various statistics collected.

1.2 Clarity

This paper will also focus on the United States of America as its country of interest and will attempt to remain apolitical. The data was collected by an organization based out of Oxford University [1] with ties to the Bill and Melinda Gates Foundation. [2] The data collected has been deemed by the author to be trustworthy and representative of data known at the time of writing.

2 DEFINITION

The goal of this paper is to be able to rank similarity between COVID-19 statistics of various countries based on features defined in Table 1. The statistics of each day will be categorized and a running total between all of the countries will be kept. In the end a lookup table for each country will be produced that contains the similarity score between that country and other countries. The days being checked will be January 23rd, 2020 to November 30th, 2020.

2.1 K Means Clustering

The method for determining similarity will be K-Means Clustering, the description of which is as follows [4]:

If we have a set of observations X size n and each observation has d features, then the algorithm attempts to partition the n observations into k sets S all while minimizing the sum of squares error, which is also know as the variance.

$$\sum_{i=1}^{k} \sum_{x \in S_i} \| x - \mu_i \|^2 \tag{1}$$

2.2 Distance Measure

Choice of the distance measurement used in the k-means clustering algorithm does have a large impact on how it functions. To play it safe and to be transparent, this paper will utilize euclidean distance as its measurement. The formula is as follows:

$$d(p,q) = \sqrt{(p_1 + q_1)^2 + (p_2 + q_2)^2 + \cdots + (p_n + q_n)^2} \tag{2}$$

where d is the distance function, p and q are both positions in multi-dimensional space with p_n being the nth dimensional coordinate.

2.3 Language Choice

The programming language used in this paper will be Python 3.8.5 due to the number of statistical packages already developed for the language and the ability to read through large amounts of data quickly. The source code for this project will be included with this paper.

3 PROCEDURE

In this section the way the program was constructed and the choices made when processing the data will be described.

3.1 Data Preprocessing

The OWID COVID-19 dataset contains over 60,000 entries with 50 parameters per entry, some of these are unnecessary so preprocessing of the data will be required to have a more concise and usable dataset.

Classifications Typically, a dataset that is used in data mining will contain many features that describe the data and a classification per row/entry that tells the algorithm what that data represents. In the OWID Covid-19 dataset, there are several classification columns/features such as:

- **iso_code** (ISO 3166-1 alpha-3) A code from the International Organization for Standardization that represents names of countries and their subdivisions.
- **continent** Continent of the geographical location.
- **location** Geographical location.
- **date** Date of observation.

This paper is only concerned with geographical location and the date of observation, so the first two features listed above can be removed from the dataset.

Feature Type Features in a dataset can have different types and it is important to ensure that the type of data being used in calculations is valid/works with the algorithms being used. In this case only one feature is problematic, tests_units. This feature is provided by the local government and is a string that explains how their

testing occurs/what units they use. In some studies this information might be useful, but for a brief overview like in this paper, this non-numerical data can cause issues so it will be removed from the dataset.

Data per Date The work that is going to be done will require the data to be separated by date into smaller datasets, where this could be done by hand, it is more efficient to write a program to process out the previously mentioned features and to split the data into separate day based .csv files. The pseudo-code to do so follows:

Listing 1. Preprocessing pseudo-code

```
def preprocess(inFilePath):
    file = read_file(inFilePath)

    headers = file[0]
    data = file[1:]
    exclude = {0, 1, 33}
    features = [include not in exclude]
    data = [[row[i] for i in features] for row in data]

    days = [list of days (1/23-11/30)]
    dayDict = {day : [all rows from day] for day in days}
    for day in days:
        outFilePath = str(day) + ".csv"
        outData = dayDict.get(day)
        with open(outFilePath, 'w') as outFile:
            csvWriter = csv.writer(outFile)
            csvWriter.writerow(headers)
            csvWriter.writerows(outData)
```

The above code has been condensed to fit in the paper, but should provide an understanding of how preprocessing occurred.

3.2 Determining Features

After the preprocessing is done, there should be 313 different .csv files that represent each day in the original dataset and there should be 45 remaining features to use. Given the usage of k-means clustering to attempt to determine similarity between different countries, it could be enticing to use all 45 features to have the most accurate result. Sometimes is data mining, more does not always mean better. Sometimes, noisy data can exist, and in the case of the OWID COVID-19 dataset, not all features are documented equally.

Automatic Feature Determination Depending on the rigor of the study, the individuals doing the research might choose to use statistical analysis to determine what features should and should not be used for classification. This can be done in many different ways, the simplest of which is to find *NULL* data in each feature, and determine which feature has the most; meaning, find the feature that is lacking in the most data.

Given the OWID dataset, and how the data was collected (through government agencies around the world) even *NULL* data in this set could be useful for determining similarity. A country not posting it's testing statistics is similar in it's handling of the virus to another country that also will not post it's statistics. Given this knowledge, the features were broken into 8 feature sets that were chose manually based on the writers interest in how the countries compared given those metrics.

Manually Chosen Feature Sets The manually selected feature sets are as follows:

(1) {7, 10, 12, 14, 16, 24, 27, 28, 29}
(2) {7, 10, 24}
(3) {7, 10, 14}
(4) {7, 14, 16}
(5) {24, 28, 29}
(6) {12, 27, 29}
(7) {7, 10, 14, 16}
(8) {7, 14, 12, 24, 29}

The feature sets shown above contain numbers that represent the feature ids that can be looked up in Table 1. To save time, only feature set three and five will be discussed in length.

Feature Set Three The names of the features in this set are as follows:

- New Cases per Million
- New Deaths per Million
- ICU Patients per Million

These features were chosen specifically because of how they represent the direct consequences of the virus and because they are scaled with population. More than just population size/density are contributing factors to how countries handled/are handling COVID-19, however if data that is more independent of variables (such as population size) can be selected, then the outcome of the analysis should be more representative of reality.

Feature Set Five The names of the features in this set are as follows:

- New Tests per Thousand
- Tests per Case
- Stringency Index

These features were chose due to how they represent the countries internal efforts to handle the COVID-19 pandemic. Specifically the stringency index of each nation which is defined as a "composite measure based on 9 response indicators including school closures, workplace closures, and travel bans, rescaled to a value from 0 to 100 (100 = strictest response)." [3]

3.3 Determining K Values

A big part of k-means clustering is picking a good representative value for k. The value of k should never be smaller than n, the number of features in the feature set, so a minimum threshold has been found 3. Set k to be too large and meaningful connections could be lost in the noise. As a minimum, the writer has decided that a cluster of nations should contain no less than approximately 20 nations, so a upper boundary of k has been found at 10. Which k value should be chosen out of them?

It was said early that more is not always better, however when it comes to the number of tests, sometimes more can be useful. Finding a good value of k makes sense when the task is direct classification, not exploration. Due to the nature of this project, there is no way to tune k to find certain results, that would be disingenuous. So instead, testing should occur on all the previously mentioned feature sets and also on all values of k between 3 and 10. This can be done with automation.

3.4 Automated Testing

Automated testing can allow a algorithm to be run on a set of parameters, and allow the user to step away while the machine works on the problem. That is what will be done here. There are 8 feature sets and 8 possible values for *k*, meaning that if it only takes 200 milliseconds to use k-means clustering on a single day dataset, then it would take over an hour to go through all permutations. This would be impossible to do by hand, and tedious to run manually.

So with good planning ahead of time, a system can be set up to allow automated testing and to organize all of the data neatly. The pseudo-code that runs the automated testing is as follows:

Listing 2. Automation pseudo-code

```
for k in k_set:
    for feature in feature_set:
        for day in days:
            day_data = getPreprocessedData(day)
            clusters = k_cluster(day_data, features, k)
            outFilePath = k+"/"+f+"/"+str(day)+".csv"
            with open(outFilePath, 'w') as outFile:
                csvWriter = csv.writer(outFile)
                csvWriter.writerow(data)
```

getPreprocesseData() is a helper method used to load the correct days dataset and *k_cluster()* is the actual implementation of the k-means clustering algorithm. Again the example above is reduced to save space.

3.5 Using the Data

Now that the data has been clustered and neatly organized, it needs to be used. The initial idea for this project was to use the features to plot the nations per day in 3D space and show how the clusters moved around with an animation. This fell through for a number of reasons.

Firstly, the cluster indexes change per day. China and South Korea might be clustered together one day with a cluster index of three, but the next day might be clustered together with a cluster index of two. Is it the same cluster? Possibly, but it depends on how far the centroid of a cluster on one day is from the centroid on the other day. Trying to figure this out causes problems.

Next, plotting them restricts the features to *n<=3* and considering our *k* value minimum is three, it does restrict what possible feature sets could be used.

Finally, it is not practical to show information. Over 200 small colored dots moving on a screen might look cool, but it is hard to say concrete information based on an interesting animation. Also it might be difficult to determine countries, depending on the plot type. So a different use of the data was found.

3.6 Comparing Nations

With a small adjustment to the code, it can be made to track the number of times a country shared a cluster with a different country. This can be outputted after all the days in a k/feature_set run have been completed. This will allow readers to pick a k value, then a feature set, and finally pick a nation. From there they can view, based on the prior selected parameters, statistically how similar that country performed compared to others.

Python Dictionaries This was done by adding a python dictionary to the above code after the two initial for loops. This dictionary would hold a nations name as a key, and the value itself would be another dictionary, whose keys were the names of the other nations, and the values are a counter for the number of times both of the nations appeared in the same cluster. Using pythonic comprehension, this could be done in few lines of code.

4 RESULTS

The results of everything mentioned above is now available to look at and will be included with this report. It is too large to add in on it's own. However, certain connections can already be made.

4.1 United States Compared

When *k=10* and the features *New Cases per Million*, *ICU Patients per Million*, and *Hospital Patients per Million* the following were the top five similar nations:

(1) Costa Rica (223 matches)
(2) Brazil (191 matches)
(3) Gabon (191 matches)
(4) Belize (187 matches)
(5) Cape Verde (186 matches)

However when *k=10* and the feature were all the features listed in feature set one, the results were as follows:

(1) Panama (213 matches)
(2) Colombia (211 matches)
(3) Argentina (205 matches)
(4) Brazil (204 matches)
(5) Bosnia and Herzegovina (202 matches)

This shows that by changing the *k* value and the features used, the argument could be made that the United States was similar statistically to different nations. This could be used to push various narratives. For a better example, only one feature or *k* value should change between comparisons, that would show what feature/*k* value is more volatile.

Feature sets two and three will now be compared using the same *k* value of six to establish how volatile the clustering is. For feature set two, which contains the features *New Cases per Million*, *New Deaths per Million*, *New Tests per Thousand*:

(1) Panama (155 matches)
(2) Kuwait (141 matches)
(3) Brazil (136 matches)
(4) Moldova (124 matches)
(5) Armenia (116 matches)

And now feature set three which contains the features *New Cases per Million*, *New Deaths per Million*, *ICU Patients per Million*:

(1) United Kingdom (162 matches)
(2) France (144 matches)
(3) Luxembourg (127 matches)
(4) Portugal (126 matches)
(5) Italy (120 matches)

This could imply that when it comes to testing, cases, and deaths, the USA is more similar to Panama; but when ICU patients are prioritized, European nations like the UK and France are more similar. However, it should be noted that the other features still confound the results to an extant and that other factors such as ICU reporting in the US and Panama might be similar, but all that might mean is that the United States is reporting low ICU patients and Panama might not be reporting high quantities of ICU patients, or vice-versa.

5 CONCLUSION

Correlation does not equal causation, and the previous examples of how the United States compared to other nations is not qualifying how good the US, or other nations handled the pandemic. The purpose of this research/project was to create a tool/dataset that could be used as a jumping off point for further research and investigations. In this regard, the research has been successful. A clustering tool was made, and data was categorized. Conclusions are up to the reader.

5.1 Possible Improvements

UX Creating a CLI/GUI/Web tool that could allow an end user to sift through the data generated in this project quickly would definitely increase it's real world usability and the author will look into this in the future.

Efficiency This project used mainly built in Python libraries and relied heavily on pythonic comprehension. There is now doubt that run time efficiencies could be gained if more time was spent programming.

Automation Having this project hook up automatically to the OWID database daily, and rerun itself would increase it's relevancy.

REFERENCES

[1] 2020. About Our World in Data. https://ourworldindata.org/about. (Dec. 2020).
[2] 2020. Our World In Data Bias. https://mediabiasfactcheck.com/our-world-in-data/. (Dec. 2020).
[3] 2020. Owid/Covid-19-Data. https://github.com/owid/covid-19-data/blob/master/public/data/owid-covid-data.csv. (Dec. 2020).
[4] Hans-Peter Kriegel, Erich Schubert, and Arthur Zimek. 2016. The (black) art of runtime evaluation: Are we comparing algorithms or implementations? *Knowledge and Information Systems* 52, 2 (Oct. 2016), 341–378. https://doi.org/10.1007/s10115-016-1004-2

A VIDEO PRESENTATION URL

https://youtu.be/xXXvtKGmkuc

B REFERENCE TABLES

Table 1. Feature Info

Feature ID	Feature Name
0	Total Cases
1	New Cases
2	New Cases Smoothed
3	Total Deaths
4	New Deaths
5	New Deaths Smoothed
6	Total Cases Per Million
7	New Cases Per Million
8	New Cases Smoothed Per Million
9	Total Deaths Per Million
10	New Deaths Per Million
11	New Deaths Smoothed Per Million
12	Reproduction Rate
13	Icu Patients
14	Icu Patients Per Million
15	Hosp Patients
16	Hosp Patients Per Million
17	Weekly Icu Admissions
18	Weekly Icu Admissions Per Million
19	Weekly Hosp Admissions
20	Weekly Hosp Admissions Per Million
21	New Tests
22	Total Tests
23	Total Tests Per Thousand
24	New Tests Per Thousand
25	New Tests Smoothed
26	New Tests Smoothed Per Thousand
27	Positive Rate
28	Tests Per Case
29	Stringency Index
30	Population
31	Population Density
32	Median Age
33	Aged 65 Older
34	Aged 70 Older
35	Gdp Per Capita
36	Extreme Poverty
37	Cardiovasc Death Rate
38	Diabetes Prevalence
39	Female Smokers
40	Male Smokers
41	Handwashing Facilities
42	Hospital Beds Per Thousand
43	Life Expectancy
44	Human Development Index