Author: Mansi Joshi

World Covid 19 Cases – Cluster Analysis

Introduction

The purpose of this project is to apply cluster analysis to country-level COVID-19 data to examine the incidence of COVID-19. This will help us uncover country groups as far as COVID-19 crisis is concerned so as to serve a baseline for policy guidelines at the country level.

Dataset

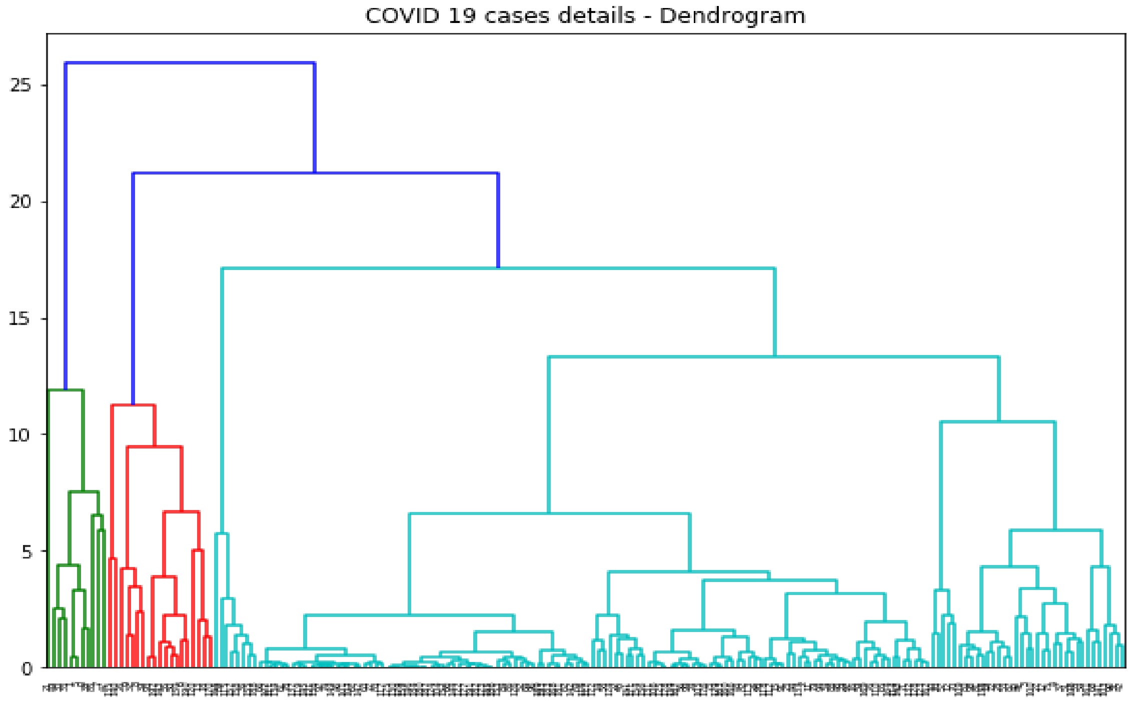
The data used is as of July 19, 2020, taken from https://www.worldometers.info/coronavirus/.

Determining Optimum Number of Clusters

Once data is scaled, we try to find the optimum number of clusters to perform further analysis.

There are 2 methods that have been used to determine optimum number of clusters:

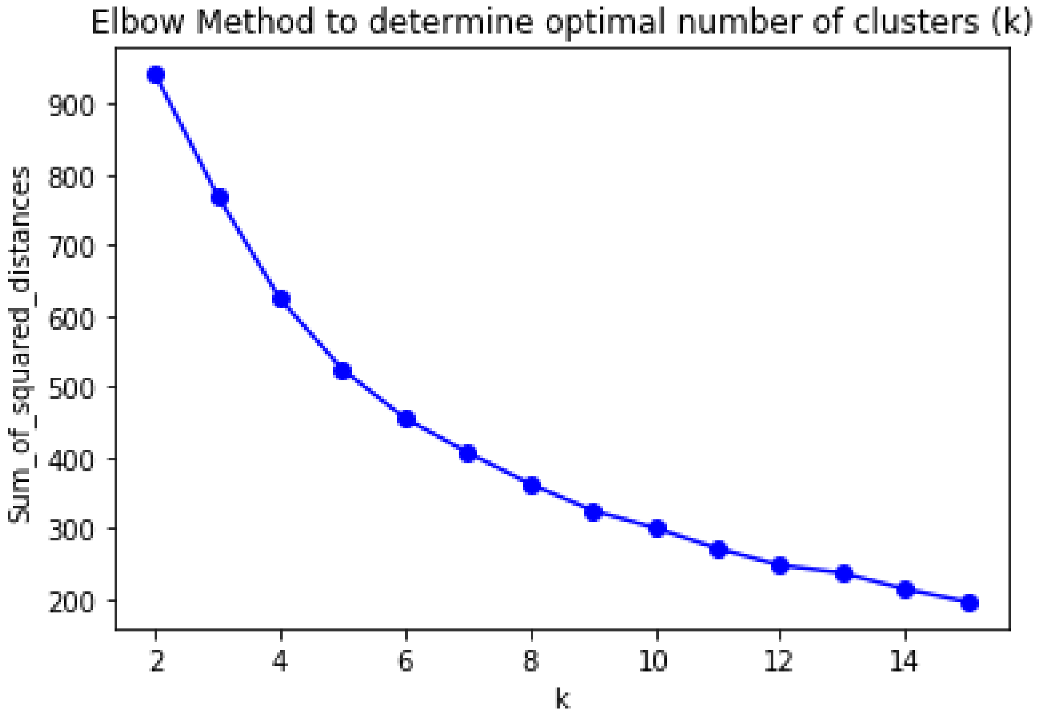
1. Using scipy.cluster.hierarchy:
   Using this hierarchy, we generate a dendrogram from scaled data, to obtain optimum number of cluster. The resultant dendrogram is as shown below:



From the figure, we see that the dendrogram is divided into multiple branches. Majorly, we see it is resulting into **3 clusters** – evident from the colors green, red and blue.

The blue cluster visibly has more members than the rest 2. We will be able to see this further in cluster memberships as well.

2. Elbow Method:
   In this method, we generate a scree plot, to determine optimum number of clusters.
   This method involves calculation of withinness sum of squares (wss), which is then used to plot the scaled data as a scree plot.
   The elbow curve, (where there is a visible curve on the plot), gives optimum number of clusters for the data.
   The resultant plot is shown below:



   Often, this plot does not give a clear interpretation of the optimum number of clusters, as seen in this plot. There is a slight curve around k = 5, k = 6, but is not visibly clear. Hence, we cannot rely on this method.

   Thus, we consider number of clusters as 3 (k = 3), from the Dendrogram.

Cluster Analysis

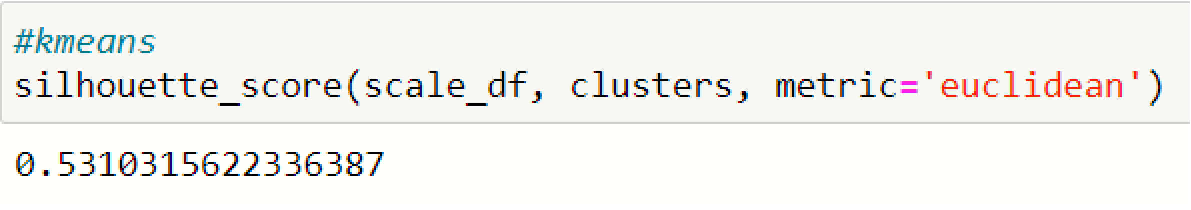
1. K-Means
   K-means method attempts to minimize the total squared error.
   In this method, each cluster is represented by the center of the cluster.
   Using k = 3, we use the K-Means method to perform cluster analysis.

2. Hierarchical (Agglomerative Clustering)
   The agglomerative clustering method in Hierarchical, uses a bottom up approach to form clusters.
   It uses the concept of dissimilarity matrix, which merges the members that have the least dissimilarity.

Evaluation of Cluster Solutions

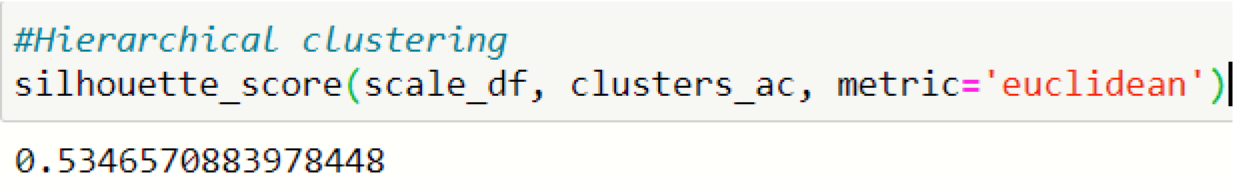
3 methods have been used to evaluate the strength of the above cluster solutions:

1. Silhouette Score
   The Silhouette Score is based on Silhouette coefficient, which is given by the mean of Silhouette coefficient of each sample. The higher the score in the positive range, the cluster distribution is considered better. The resultant score for the 2 clustering techniques is as follows:



```
#kmeans
silhouette_score(scale_df, clusters, metric='euclidean')
```

```
0.5310315622336387
```

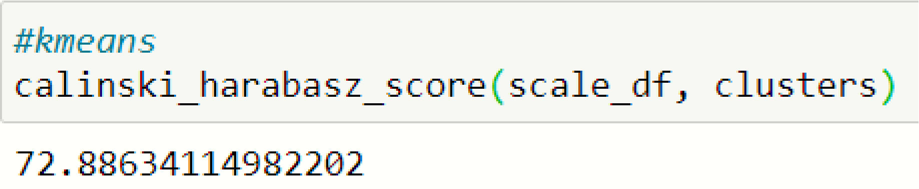


```
#Hierarchical clustering
silhouette_score(scale_df, clusters_ac, metric='euclidean')
```

```
0.5346570883978448
```

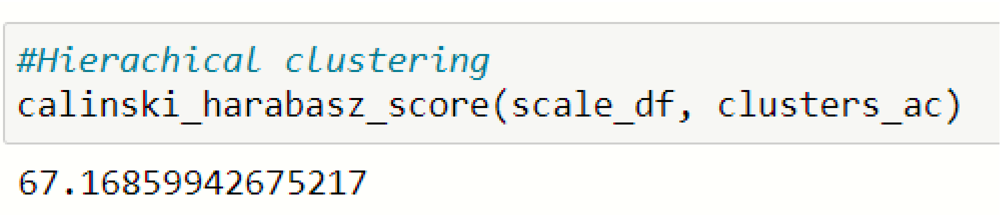
   Although the scores are similar, out of these, the score for Hierarchical method is better.

2. Calinski_harabasz_score
   This score is based on Calinski-Harabasz index (or Variance Ratio Criterion), and is determined by ratio of sum of between-clusters dispersion and inter-cluster dispersion for all clusters.
   There is no threshold or decision point, it is based on a comparative scale, the higher score depicts a better cluster distribution.



```
#kmeans
calinski_harabasz_score(scale_df, clusters)
```

```
72.88634114982202
```

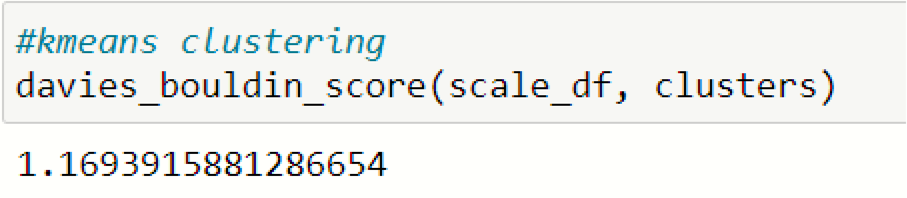


```
#Hierachical clustering
calinski_harabasz_score(scale_df, clusters_ac)

67.16859942675217
```

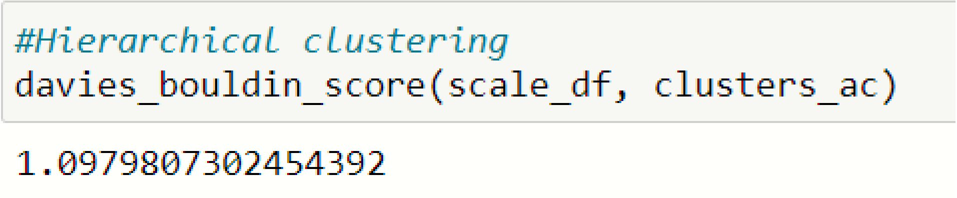
Out of these two, the score for K-Means method is better.

3. Davies_bouldin_score
   This score is based on Davies-Bouldin Index, which is based on the average ‘similarity’ between clusters.
   A lower score depicts a better distribution.
   The results are as follows:



```
#kmeans clustering
davies_bouldin_score(scale_df, clusters)

1.1693915881286654
```



```
#Hierarchical clustering
davies_bouldin_score(scale_df, clusters_ac)

1.0979807302454392
```

This method indicates the distribution with Hierarchical method is better.

In conclusion, we see that 2 out of these 3 evaluation techniques indicate that the distribution in Hierarchical clustering is better.

Hence, we select this Hierarchical (Agglomerative) technique for further analysis.

Cluster Means and Membership

We got the following cluster means (centers) using the Hierarchical clustering method:

Clust_mem	**0**	**1**	**2**
Mortality_rate	0.0224	0.0181	0.1175

Cases_per_ml	1562.945	16549.09	4727.358
Deaths_per_ml	33.7694	229.6818	425.3985
Recovered_per_ml	1087.758	12688.94	3178.19
Active_per_ml	441.4174	3630.467	1070.91
Critical_per_ml	5.2403	43.351	2.0925
Tests_per_ml	62077.74	107780.2	76792.53

Based on Cases per million and Active cases per million, we can consider following 3 categories:

- Cluster 0 – Least affected
- Cluster 1 – Highly affected
- Cluster 2 – Moderately affected

Observations and Conclusions

- ‘Moderately affected’ cluster’s mean of cases per million (4727.358) is quite close to ‘Least affected’ cluster, compared to cluster ‘Highly affected’.
- We could expect the mortality rate to show a similar pattern, but it is in fact, higher in Moderately affected cluster (0.1175).
- This could possibly be because, countries belonging to ‘Highly affected’ cluster, have high populations.
- Hence, mortality rate is recorded higher, even though number of deaths in comparison is lesser than the ‘Moderately affected’ cluster.
- ‘Moderately affected’ group has high mortality.
  This could be because, this cluster comprises European Countries (Spain, Italy, France, ect).
  In these countries, as per the demographic distribution, higher population comprises of older age group.
  Due to the nature of COVID 19, mortality would be higher in these countries, as it is higher fatality rate in older age groups.
- Tests per million conducted is less in ‘Least affected’ group.
  Lesser capacity of the healthcare system to be able to conduct tests in these countries could be one reason for this. This is a probable cause the cases being lesser in this cluster.
- In contrast, the reason for highest cases in ‘Highly affected’ group could be that the highest number of tests conducted, are from countries in this group (eg: USA, Qatar, Kuwait)

Policy suggestions

- <u>Moderately affected group:</u>

Countries in this group, have higher populations of older age group. Even though the recorded cases might be at a moderate level, the mortality rate is higher (eg: Yemen, Mexico, Netherlands).
Hence, provisions for higher age groups should be strengthened.
Households with senior citizens should be given provisions so that they do not need to step into risky environment.

- Highly affected group:
  Countries in this group have high populations.
  Even though number of deaths is lesser for countries in this cluster, the number of cases per million and active cases is highest in this group. (Eg: USA, Armenia)
  This group shows highest number of tests conducted.
  While this might be a comparatively ‘close-to-real’ count of cases, these countries must implement stricter ways to control the number of cases.
  Restricting not only international, but also domestic travel, can help reduce the exponential number of cases.
  Availing work from home options, restricting non-essential travel beyond a certain perimeter of the location of residence of people, is another beneficial method which can be implemented.
- Least affected group:
  Even though this group depicts least cases out of the 3 groups, many countries in this group have high populations (eg: India, China, Pakistan)
  Mainly, for this group, the tests conducted is the least, which could be the main reason of least reported cases.
  Focusing on increasing number of tests conducted would help get a clearer picture of the situation.
  One way to do that, could be drive-through tests, wherein the patient would not need to enter the clinic/hospital, but the test sample would be taken by health authorities in their vehicle. This would reduce risk and exposure within hospitals and health clinics.