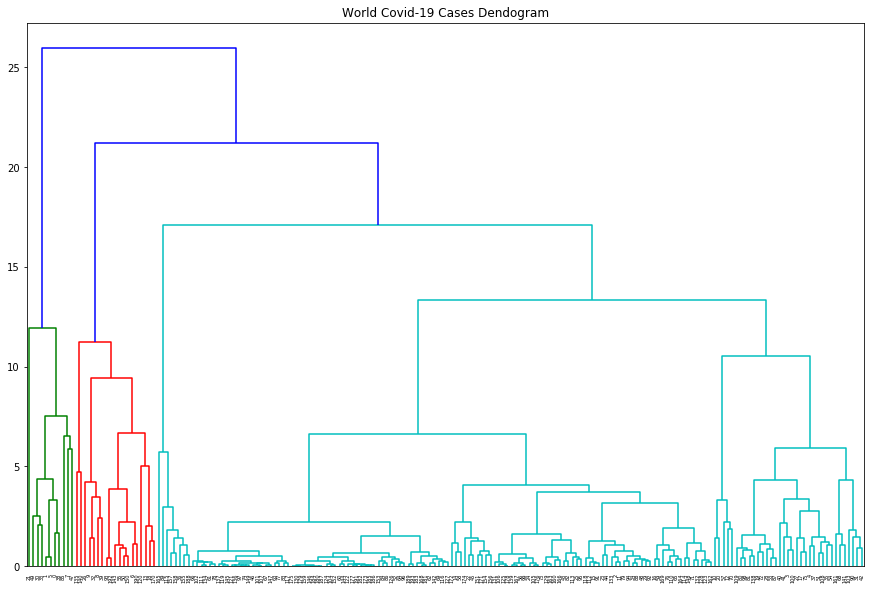
**Project Clustering**

# **Step 1 to 9 – See the Code File**

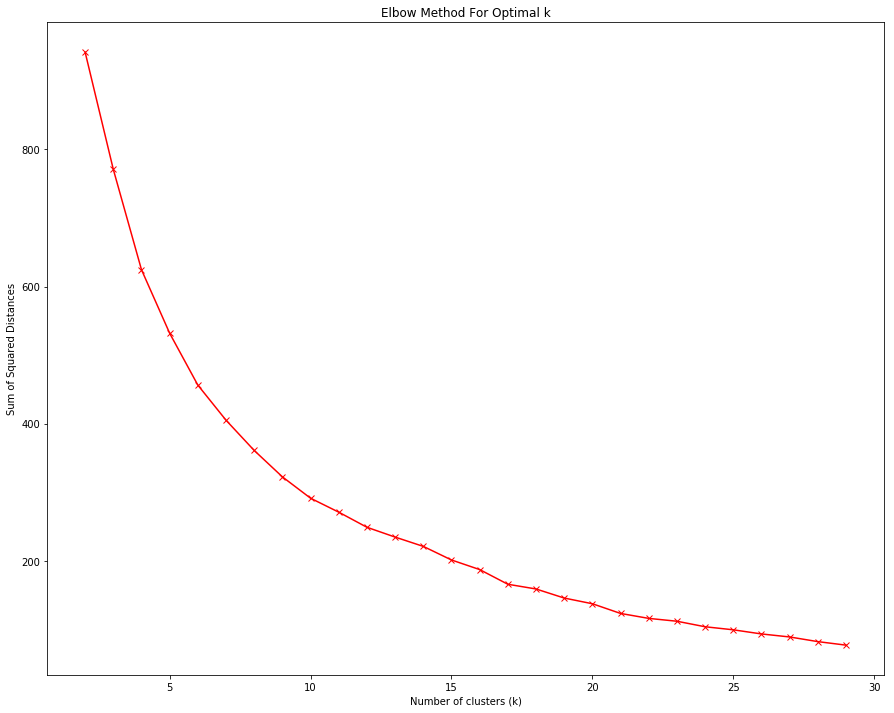
# **Step 10 – Results from parts 3, 4, 6, and 9**

*Determining the optimal number of clusters with Scipy.cluster. hierarchy*



**Figure 1 - Dendogram**

*Determining the Optimal Number of Clusters with Elbow Method*



**Figure 2 – Elbow curve**

*Evaluating the Cluster Performance*

|  |  |  |
| --- | --- | --- |
| Score Metric | Clustering Method | Value |
| Silhouette Coefficient | K-Means | 0.5668 |
| Silhouette Coefficient | Agglomerative Clustering | 0.5561 |
| Calinski-Harabasz | K-Means | 72.7305 |
| Calinski-Harabasz | Agglomerative Clustering | 68.7412 |
| Davies-Bouldin | K-Means | 1.1035 |
| Davies-Bouldin | Agglomerative Clustering | 0.9952 |

*Cluster centers (means) from unscaled data*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Cluster | Mortality Rate | Cases per Million | Deaths per Million | Recovered per Million | Active per Million | Critical per Million | Tests per Million |
| 0 | 2.41% | 1,495.30 | 33.45 | 1,050.99 | 410.86 | 4.19 | 61,838.31 |
| 1 | 11.88% | 5,438.45 | 491.31 | 3,507.77 | 1,376.59 | 3.52 | 83,004.08 |
| 2 | 1.86% | 14,957.23 | 212.88 | 11,447.57 | 3,296.78 | 48.40 | 99,538.69 |

# **Step 11 – Defining the names/labels for the clusters**

|  |  |  |
| --- | --- | --- |
| Cluster | Name/Label | Pattern |
| 0 | Low Risk of Contagion | Practically all indicators are the lowest |
| 1 | Moderate Risk of Contagion | Presents some low and high indicators, but in general it is on average |
| 2 | High Risk of Contagion | Practically all indicators are the highest |

# **Step 12 – Microsoft Excel file (attachment)**

# **Step 13 – Analysing the pattern of results**

1) It is interesting to note that cluster “2” has the lowest mortality rate and cluster “1” has the highest mortality rate and deaths by million. In my view, the cluster “2” should have the highest value in all features.

2) The mortality rate of cluster “1” seems to have some outliers because the difference for the other 2 classes is huge.

3) Why China and India do not belong to cluster ‘2’?

It is very strange that these two countries classified as “low risk” (cluster = ‘0’). If we compare the proportion of infected population with other very populous countries, we will see a huge discrepancy.

|  |  |  |  |
| --- | --- | --- | --- |
| Country | Total Cases | Total Deaths | % Population |
| China | 83,660 | 4,634 | 0.006 |
| India | 1,118,107 | 27,503 | 0.081 |
| Brazil | 2,098,389 | 79,488 | 0.988 |
| EUA | 3,893,949 | 143,263 | 1.179 |

Considering that neither China nor India has an excellent public health service, it is very weird these values. There are other similar cases, also in countries with “poor public health system” that I was expecting a high proportion, but it does not happen.

|  |  |  |  |
| --- | --- | --- | --- |
| Country | Total Cases | Total Deaths | % Population |
| Philippines | 67,456 | 1,831 | 0.062 |
| Venezuela | 11,891 | 112 | 0.042 |
| Nigeria | 36,663 | 789 | 0.018 |

4) Ecuador, Mexico in the same group with San Marino, Netherlands, Ireland, and the UK?

I did not understand why Mexico and Ecuador be in the same group with that European countries, since both countries do not have the same public health system level and the climate is much hotter.

5) No Africa country in the cluster ‘1’ or ‘2’!

This finding is the worst characteristic, because I cannot imagine that the contagion risk in USA, Qatar, or Brazil is much greater than in Zimbabwe, Mozambique, or Ethiopia. For me, it does not make sense to me.

In short, this dataset does not seem to have the reliability that I was expected, although the source is reliable. As the values are sent from each government, it seems like that some countries do not have interest to show their real situations, or even do not have means to get a better information. Therefore, it would be necessary more transparency and accuracy in these numbers to avoid inconsistent information and, consequently, wrong analysis and decisions.

# **Step 14 – Policy suggestions**

Assuming that the dataset has correct values, the suggestions for each cluster are:

**Low Risk of Contagion** (cluster ‘0’): as the group has a mix of countries with different culture, economic and financial situation, climate, and public health system, the suggestions must more general than to the other clusters.

* keep actions to maintain the social distancing.
* it can adopt a gradual opening gradual of its parks, schools, bars, restaurants, and other services.
* hospitals and clinics must remain alert and act on new cases.

**Moderate Risk of Contagion** (cluster ‘1’): as this group of countries has a good financial and economic situation, it could be implemented a social program to help people who have lost their jobs and do not have money enough to pay their expenses.

* intensify campaigns so that people are more aware of the risk.
* partial lockdown can be an action to prevent the spread of contagion.
* keep schools, bars, restaurants, parks closed or with limited access.
* public transport and government services should be organized to maintain services minimal to society.

**High Risk of Contagion** (cluster ‘2’): similar to “moderate risk”, a social program must be carried out to help the unemployed and their families, although in this case, the time will be longer than the “moderate”, due to the seriousness of the situation - some countries may not have enough budget for all period of the pandemic.

* energic actions must be taken, with lockdown and police actions.
* people should be oriented to stay at home and go out only in necessary circumstances.
* self-isolation in case of symptoms.
* schools, restaurants, malls, parks must be kept closed.
* temporary hospitals and clinics must be created to help the public health service and all healthcare professionals should be recruited to assist patients.

**Comparison of the strength of cluster solutions using the following metrics**

From Cluster Solutions applied, Hierarchical Approach (Agglomerative Clustering) presents a lower value in two out of the tree metrics applied. Therefore, Hierarchical Approach was the model utilized in the pattern examination over our dataset about COVID-19.

**Cluster Centers Means**

Countries in dataset were grouped into tree different clusters in accordance with a collection of data points that could define their similarities. These are the central points of each cluster obtained from our calculations:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Cluster Number | Cluster | Mortality Rate | Cases per million | Deaths per million | Recovered per million | Active per million | Critical per million | Tests per million |
| 0 | **Countries with lower rates of infection** | 0.022 | 1562.945 | 33.769 | 1087.758 | 441.417 | 5.24 | 62077.74 |
| 1 | **Countries with greater rates of infection, active cases and critical per million** | 0.018 | 16549.086 | 229.682 | 12688.937 | 3630.467 | 43.351 | 107780.19 |
| 2 | **Countries with higher mortality rate levels** | 0.118 | 4727.358 | 425.398 | 3178.19 | 1070.91 | 2.092 | 76792.529 |

**Result of the analysis**

Based on our results, we divided countries into tree groups.

The first cluster is Countries with lower rates of infection. In this group we can find lower number of cases per million, and subsequently, actives per million. The lowest number of cases comes from Papua New Guinea, with 1.78 cases per million. In this group, a lot of countries registered no active cases, such as Gibraltar, Dominica, Brunei, Greenland, and others. This cluster also holds the less number deaths per million, where countries like Gibraltar, Greenland, French Polynesia, Vietnam, and others registered no deaths so far.

What is surprising about this group is that we find here countries that hold the 4th and 5th position with highest number of total cases, such as Russia and South Africa. India is also in this group, while holding the 3rd position in highest number of total cases. That is because India is the second most populated country in the world, with over 1.3 billion people; therefore, its rates of infections are low, compared with its whole population.

The second cluster is Countries with greater rates of infection, active cases and critical per million which means that this group holds the highest rates of cases per millions, active cases per millions, critical cases per million and also recovered cases per million. Countries that hold the first and second position of global number of total cases, like USA and Brazil are inside this cluster, and together they represent almost 40% of total cases around the world so far. The country with greater number of Cases per million is Qatar, with 37,982 cases. Almost all cases were recovered in that country with 36,817 recovered cases, which represents almost 97% of recovery. So even being at the top of cases per millions, Qatar has a successful recovery rate. The number of Actives per millions is more expressive in French Guiana, with 7.846 people per million. About the critical cases per million, we have Chile leading the list, with over 92 per million. Finally, the cluster holds a greater number of tests per million, which is coherent, since we have countries with greater number of population and infected cases registered, like USA and Brazil.

The third cluster is Countries with higher mortality rate levels. This group presents countries with higher mortality rate and deaths per million. The country with greater mortality rate is Yemen, with 27.7%, while the country with greater number of deaths is San Marino, with over 1,237 deaths per million, meaning that 6% of infected people have died from COVID-19 so far in this country. We highlight in this list France, that from a total number of cases of 174,000 people, had over 30,000 deaths registered, representing over 17%. We also find here Italy, where from the 244,000 infected, over 35,000 have died. Belgium itself registered only nearly 64,000 cases, but registered a death number of nearly 10,000 people, which represents over 15% of COVID-19 cases in that country, while UK holds the same percentage of deaths.

**Policy Suggestions for Clusters**

Cluster 0 - Countries with lower rates of infection: Despite the fact that these countries hold the lowest rates of infection, they also hold the lowest values for Tests per millions and in this group, we find countries with a high number of populations, such as Japan, Russia, Pakistan, Indonesia, India and China. These countries together represent almost 45% of global population, so it is possible that the number of cases is low because there are unaware cases. Governments must increase their testing capability to cover a greater range of their population and bring a more reliable result over the number of infected.

Cluster 1 - Countries with greater rates of infection, active cases and critical per million: Since these countries hold greater number of cases per millions, they must invest in more policies to contain the spread of the virus, increasing lockdown measures and enforcing the closure of businesses and social gatherings in order to flatten the curve. Countries like USA and Brazil are in this cluster and, as said before, hold the position of first and second number of cases worldwide. These countries are also marked by a controversy governmental administration, that do not stimulate lockdown, prioritizing the continuity of “normality” in order to avoid an economical crisis. These countries must also ensure at a faster pace hospital facilities with appropriate Intensive Unit of Treatment (IUT) and healthcare professionals to provide an appropriate healthcare assistance to COVID-19 cases without collapsing the health system and leaving other types of cases without proper support.

Custer 2 - Countries with higher mortality rate levels:

Countries with higher death rates must invest in Intensive Unit of Treatment (IUT), to ensure that critical patients will have the appropriate treatment and a chance of survival. That’s the struggle Italy faced when it became the epicenter of the virus, hitting a mortality rate of over 14%. This country had difficulty to provide appropriate treatment for the infected, while the healthcare system collapsed.

Another guideline that is also important is to increase the number of testing. That is because some countries only test patients that go to the hospital and do not test those that do no present severe symptoms or even the asymptomatic ones. That is why in some cases the death rate can appear higher than in countries where testing is widespread such as the United Kingdom, or Yemen, that leads the rank of mortality rate, but actually registered only 4 tests for each million of its inhabitants.

**Conclusion**

It is a bit controversial to state the mortality rate or even the number of infected when it is not possible to ensure that the whole global population was tested. So the lack of a widespread, systematic testing in most countries is the main source of discrepancies in these rates. To get an accurate figure across a population, it is necessary to test not just symptomatic cases, but asymptomatic people too. Having that data would give an accurate picture of how the pandemic is affecting whole populations, not just the sick.

Another factor is that death rates may be underestimated when an infected patient dies due to an underlying condition such as asthma or another chronical disease. Not only that, but there are also those cases where a person probably died due to COVID-19 but never got the chance to be testes. And there is more likely to happen in places where the economy is instable or where population has not so much access to proper healthcare assistance. That might explain why countries like Tanzania, Algeria or Syria, have no tests registered at all.

So as a global guideline a proper lockdown and an effective testing is the way to flatten the curve.