Covid Data Analysis and Visualization

Bikramaditya Subedi, Rajeshwor Niroula

Department of Computer Science and Engineering, Kathmandu University , Nepal

bs02281417@student.ku.edu.np
rn02094316@student.ku.edu.np

ABSTRACT

The project is aimed for implementation of various stages of data process and tools to draw any meaningful insights while relating Covid cases with certain factors of the Human Development Index (HDI). Through unsupervised learning methods and data visualization we have discovered that certain human development factors might have contributed to a larger death rate in some countries signifying the importance of these factors in crisis handling.

Keywords: WHO, HDI, Unsupervised learning, Choropleth, Dendrogram

I. INTRODUCTION

Ever since Covid-19 outbreak, there has been numerous data analysis done over the spread of the disease with time. In this analysis we are interested in finding any interesting relations if it exists between Covid-19 outbreak and factors determining HDI using different algorithms. Orange platform is used for data analysis and visualization. Two datasets are taken, WHO Covid time survey set and HDI datasets.

II. OBJECTIVES

The objectives of the data analysis are:

a) Evaluate the country's health sector's effectiveness by using casualties and number of Physicians using Unsupervised learning.
b) Analyze if casualties relate better with any other attributes than the number of physicians.
c) Visualize findings.

III. PROCEDURE

The data needs to be preprocessed , transformed before analyzing. The different procedures the data go through are extremely important and the quality of analysis depends on it.

A. Preprocessing and Transformation

Since we are interested in the total casualties all we need from covid survey data set is the mortality rate, which is not given in the dataset, so we used feature constructor to calculate mortality by selecting data till date 2022-01-01 as:

Mortality (per 10000) = (cumulative death/cumulative cases) *10000

Once the mortality is calculated for every country, we can remove undesired attributes and any instances of missing mortality values. To merge HDI and Covid dataset for every country we had to ensure both datasets have the same metadata designations, for that we edited the domain of Covid dataset to match country names of HDI dataset. Once the data is merged any country data not mentioned in either data set was cleaned.

B. Unsupervised Learning and Evaluation

For the evaluation of the relation between number of Physicians and Mortality, Scatter Plot is plotted from the data we have.

Although the data labeling looks cluttered, we can instantly identify some countries that have performed exceptionally badly in handling the situation with Yemen having the worst result. Even though it has a low number of Physicians, their mortality is more than expected as compared to a lot of countries with about the same number of physicians.

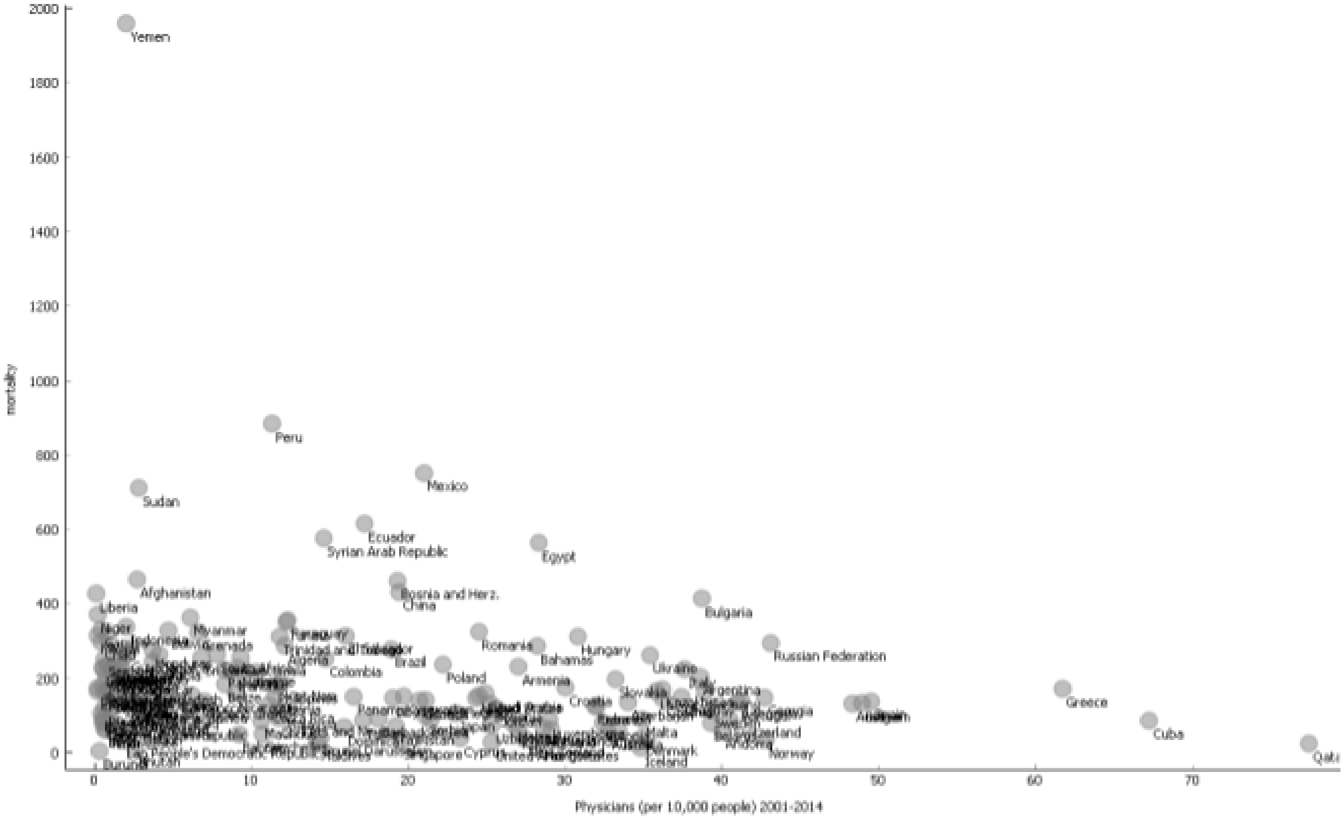
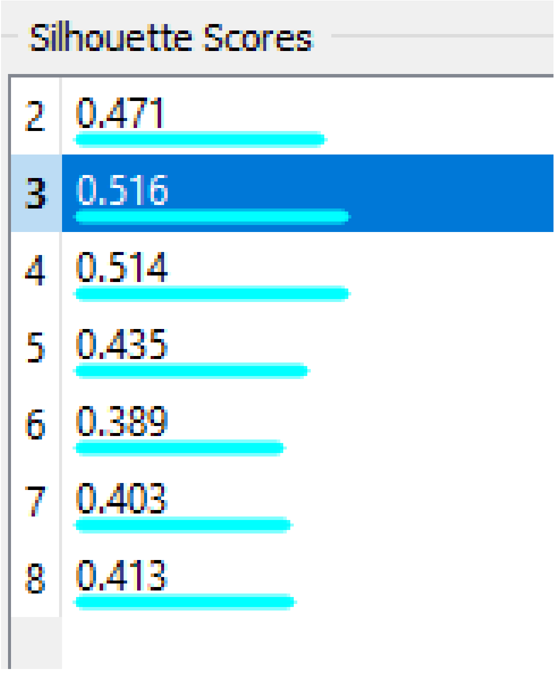


Fig. 1:Physicians VS Mortality Scatter plot

Similarly, Peru, Mexico, Sudan etc., also have poor performance despite adequate numbers of physicians this implies that the country despite having adequate health care providers is either ill equipped or the professionals themselves are not well qualified. Qatar, Greece and Cuba seem to have the greatest number of physicians, however their crisis response cannot be considered adequate because there are too many countries with mortality below 100 despites having less than one tenth of the number of physicians of highest counting countries. However, the mortality of these countries might also be because of the nature of the outbreak. Initially despite having medical infrastructure, professionals might not have known how to handle the situation. In Order to have a generalized view of countries health care performances clustering can be done.

C. K-Means Algorithm

Although Silhouette Scores suggest clustering into 3 groups to be the best, we have made 4 clusters since the scores don't differ by significant margin and only three clusters seem too generalized.To better visualize the clusters Choropleth map can be used where 4 clusters are represented by 4 colors as shown in Fig 3.


Silhouette Scores

2	0.471
3	0.516
4	0.514
5	0.435
6	0.389
7	0.403
8	0.413

Fig.2 :Physicians and Mortality Silhouette Scores

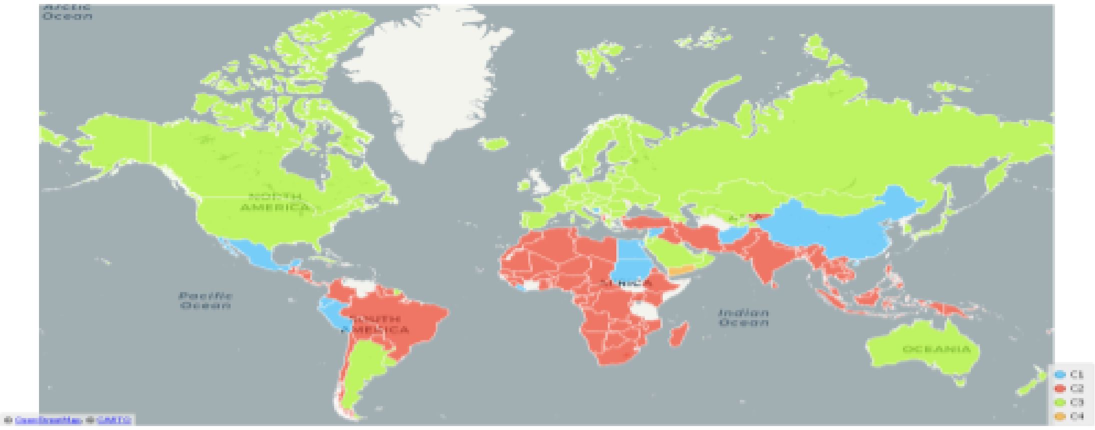


Fig. 3: K-means: Physicians and Mortality: Choropleth

Investigating we can see that Nepal, India, Turkey, and most African countries fall within the red cluster. These countries seem to have performed relatively well given their smaller medical infrastructure. Surprisingly countries like Brundin and Bhutan despite poor infrastructure have done exceptionally well, even better than Qatar and Greece, which could make them an outlier and requires further investigation. This result might also be because of the country's strict isolation from the rest of the world. Yemen as expected is an outlier and makes its own single data cluster with terrible crisis handling. Cluster blue with Egypt, Mexico, Peru etc tends to the group of bad performers since they have adequate infrastructure but still couldn't handle the crisis well. This could be because of administrative delay or severe lack of equipment or harshly physicians are not professional enough.

Let's see Fig. 4 which shows the scatter plot of Physicians and Mortality where the K-means algorithm is applied. It‘s interesting that China also falls within this category despite having a large economy and amped up infrastructure during covid. We all heard the news on how China built a hospital in 6 days. So why is it in the blue cluster? Well, this could be because it was the first country to face the covid outbreak, but it could also be because of herd immunity. Some populations tend to be more vulnerable towards certain diseases. The green cluster mostly comprises developed nations or developing nations with higher physician count, even though their infrastructure could have handled the crisis they were too slow to respond mostly because of political reasons. The US also falls within this category. Since the outbreak people were too arrogant to follow basic safety protocols like social distancing and wearing masks because they argued it violated their freedom while the government kept pushing the idea that covid was just a hoax.

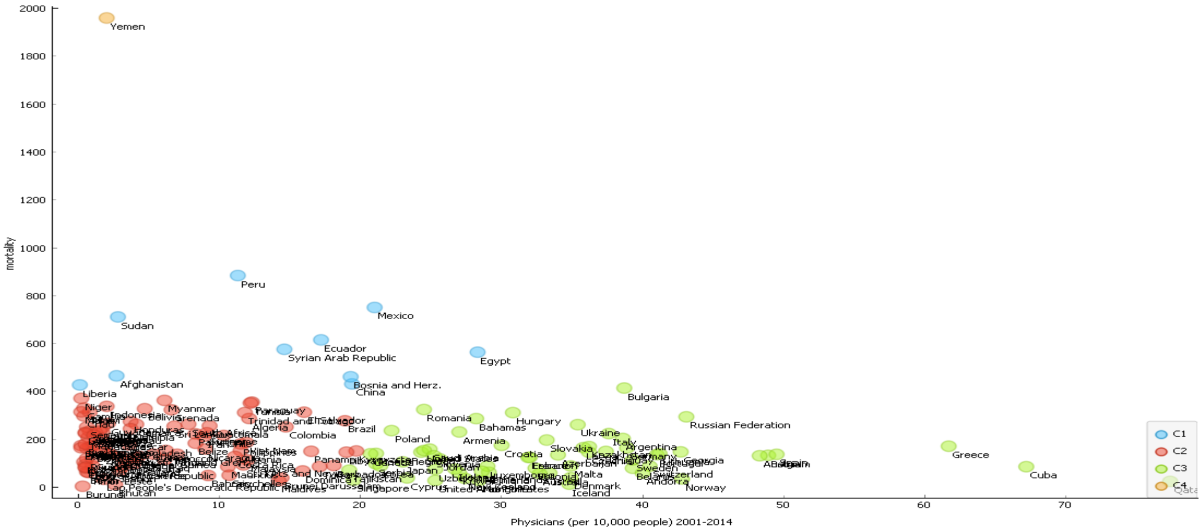


Fig. 4 Physicians Vs Mortality: Scatter Plot(K-means)

D. Hierarchical Clustering

Let's evaluate the same data with another form of clustering and see how it groups these countries. To make the comparison more relatable we have used the same 4 clusters divided using a hierarchical method and plotted the result in choropleth with 4 distinct colors. But first we need to choose a linkage method. After running tests with complete, single, average and ward's methods and visualizing the result in dendrogram it becomes more obvious that clusters with ward's method are well exposed and simplified compared to other methods.

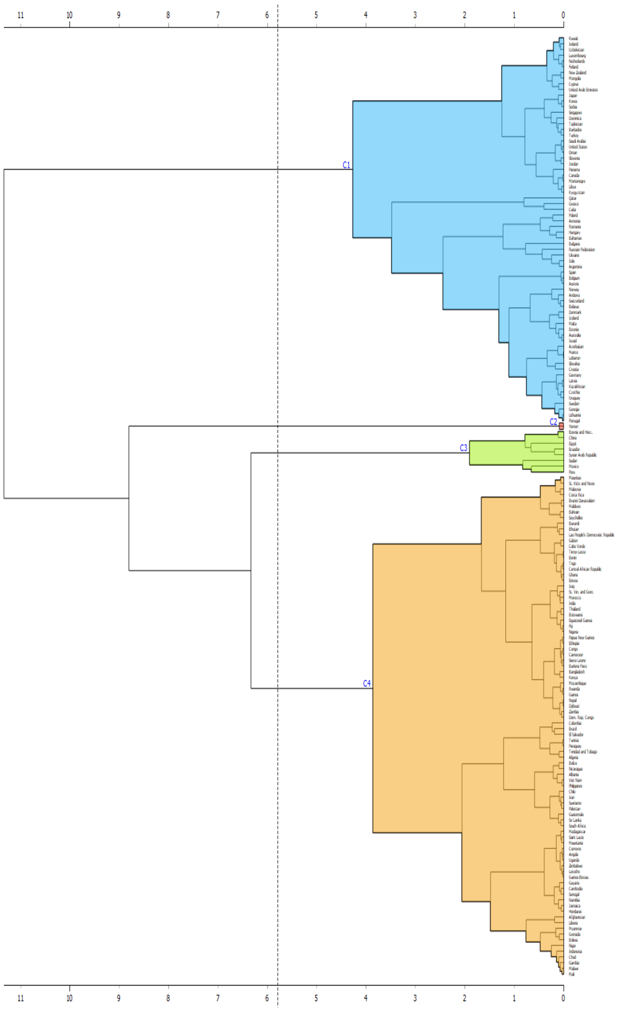


Fig. 5 Ward's method: Physicians and Mortality: Dendrogram

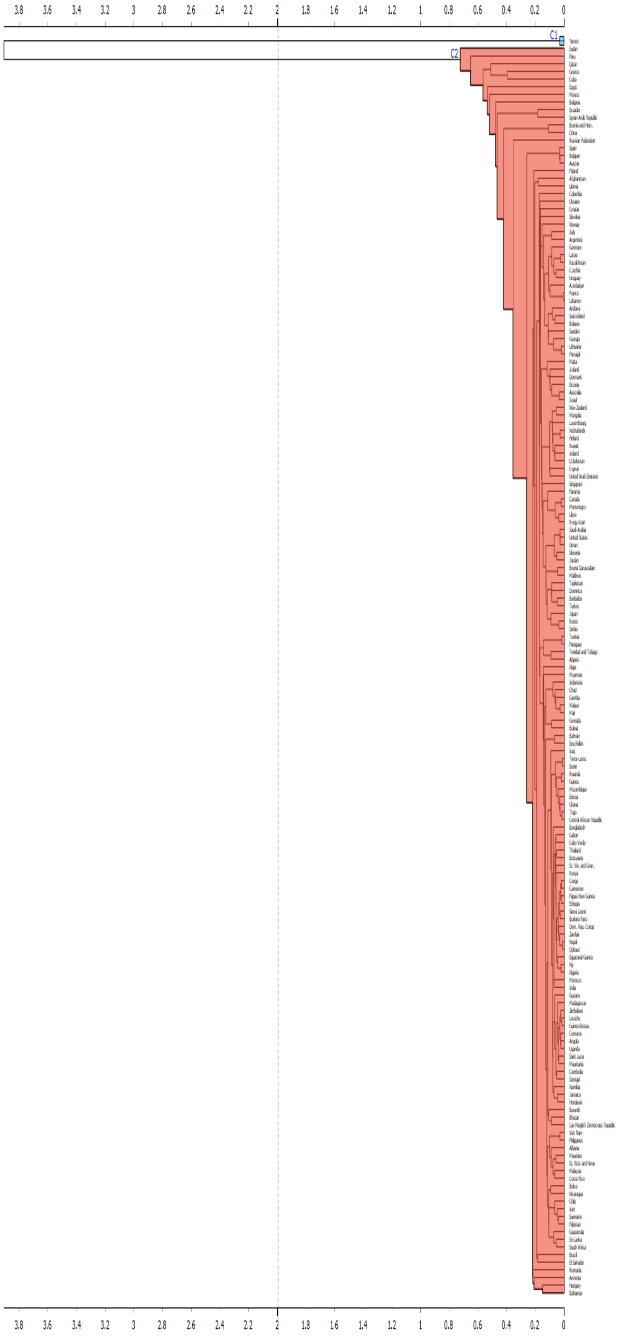


Fig. 6 Single link: Physicians and Mortality: Dendrogram

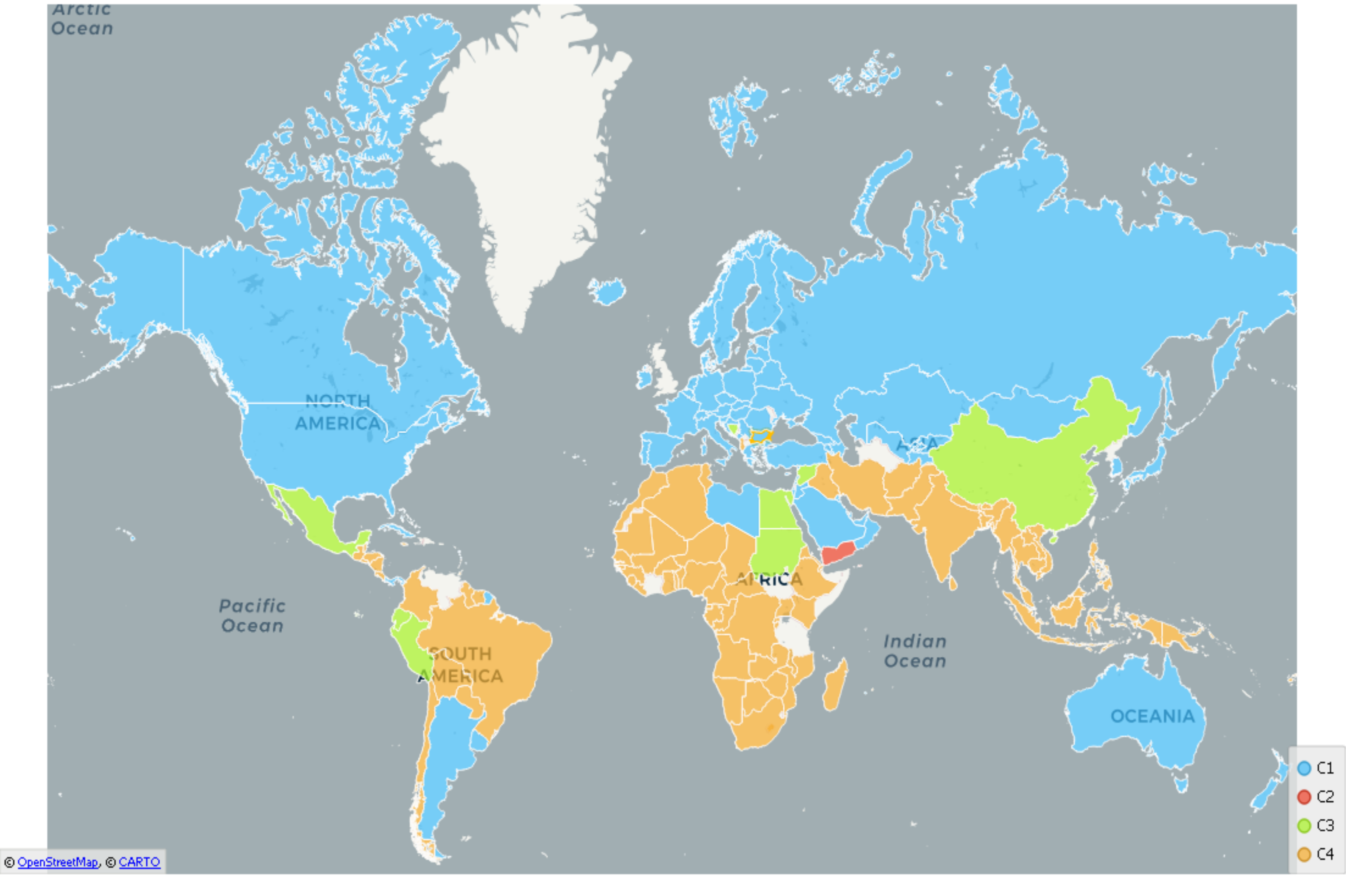


Fig. 7 Ward's Method: Physicians vs Mortality: Choropleth

We can see the result using hierarchical clustering isn't far different from K-means. Meaning of these clusters also resembles that of k-means but there are some differences. The differences are more prevalent towards the edge intersection of clusters. Ward's method seems to be putting countries with mortality 400 like Liberia and Afghanistan in orange clusters that are relatively better performing countries which were placed by k-means with poor performing countries. Similarly, countries like Libya, Panama, Dominica etc are placed in the groups with relatively developed countries or countries with a high number of physicians. In this scenario k-means clustering seems more appropriate as it works better with non-globular data compared with hierarchical clustering

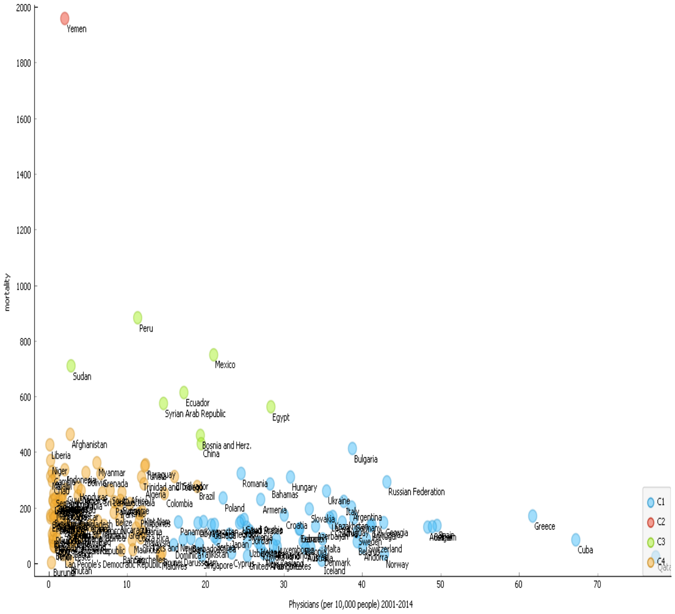


Fig. 8 Ward's Method: Physicians vs Mortality: Scatter Plot

E. Outlier Detection

Outlier detection can provide insight to interesting occurrences in the dataset. LOF has been used for the detection and the results are as expected by above analysis. It seems the outliers are countries that have performed exceptionally good (Burundi), exceptionally bad (Yemen) and unexpectedly bad like China, Peru etc as we discussed.

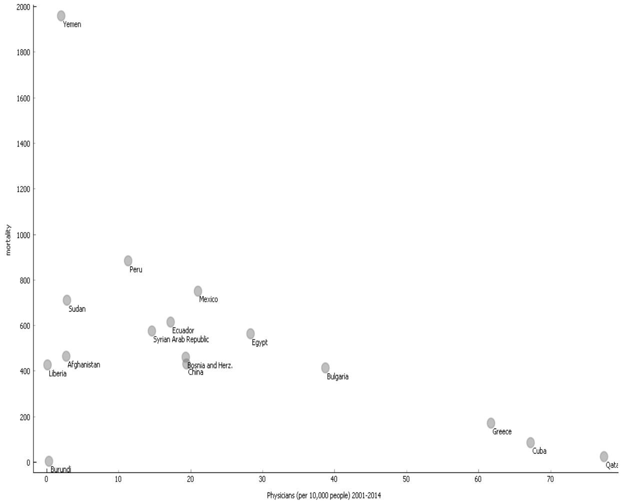


Fig. 9 Outliers: Physicians and Mortality

F. Correlation

Since the number of physicians available might be an obvious parameter to assess the mortality rate but there could be other factors relating to HDI that might have affected the mortality due to covid. To find it we can do correlation analysis between mortality and all 52 features in the HDI data set. Turns out unemployment in youth correlates with mortality by +0.265. However, it's difficult to draw a straight relation between youth unemployment and death by covid so we drew correlation between unemployed youth and rest of features and found youth unemployment is correlated with inequality in income by +0.288. We can speculate that income inequality might have led to unfair health care distribution and thus increase in mortality.

Scatter plot shows that the increasing income inequality does affect mortality however it is not consistent, that is not every country's mortality has been affected by increasing income inequality.

G. K-means

Clustering might give an insight to any interesting patterns relating to income inequality and mortality. Even though

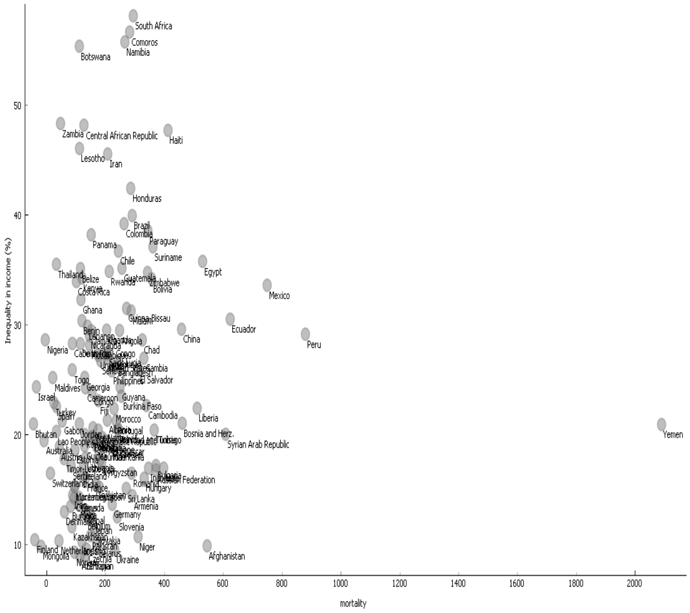
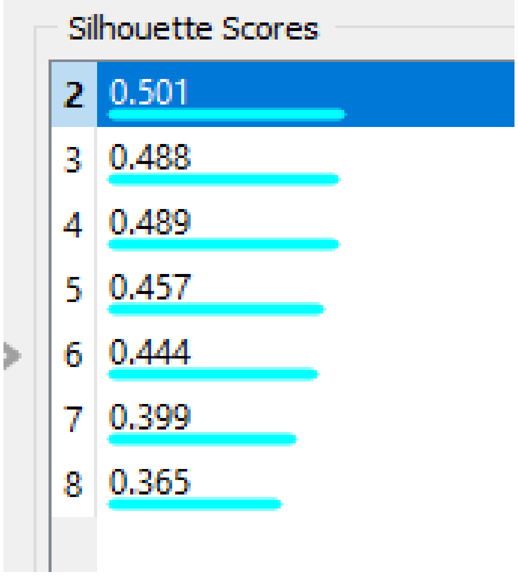


Fig. 10 Income Inequality and Mortality: Scatter Plot

Silhouette Scores suggest clusters of 2 is better but the score between 2 and 4 clusters are not significantly different hence 4 clusters have been chosen to ensure the results are not too generalized.



Silhouette Scores

2	0.501
3	0.488
4	0.489
5	0.457
6	0.444
7	0.399
8	0.365

Fig. 11 K-means: Income inequality and Mortality:Silhouette Scores

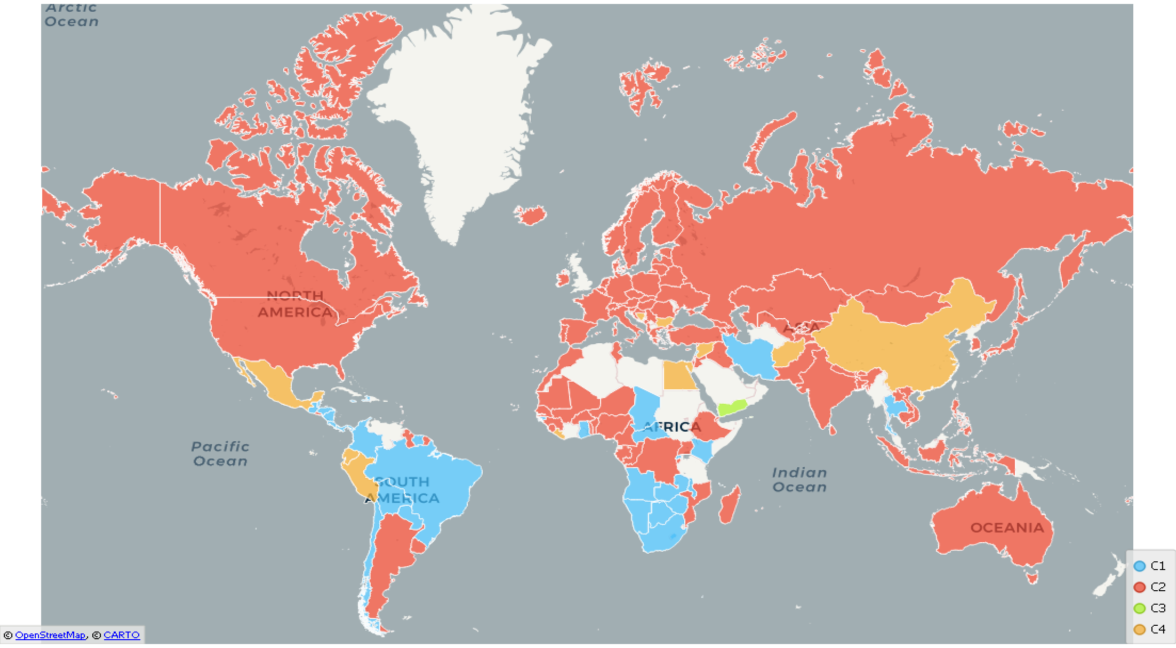


Fig. 12 K-means: Income inequality and Mortality: Choropleth

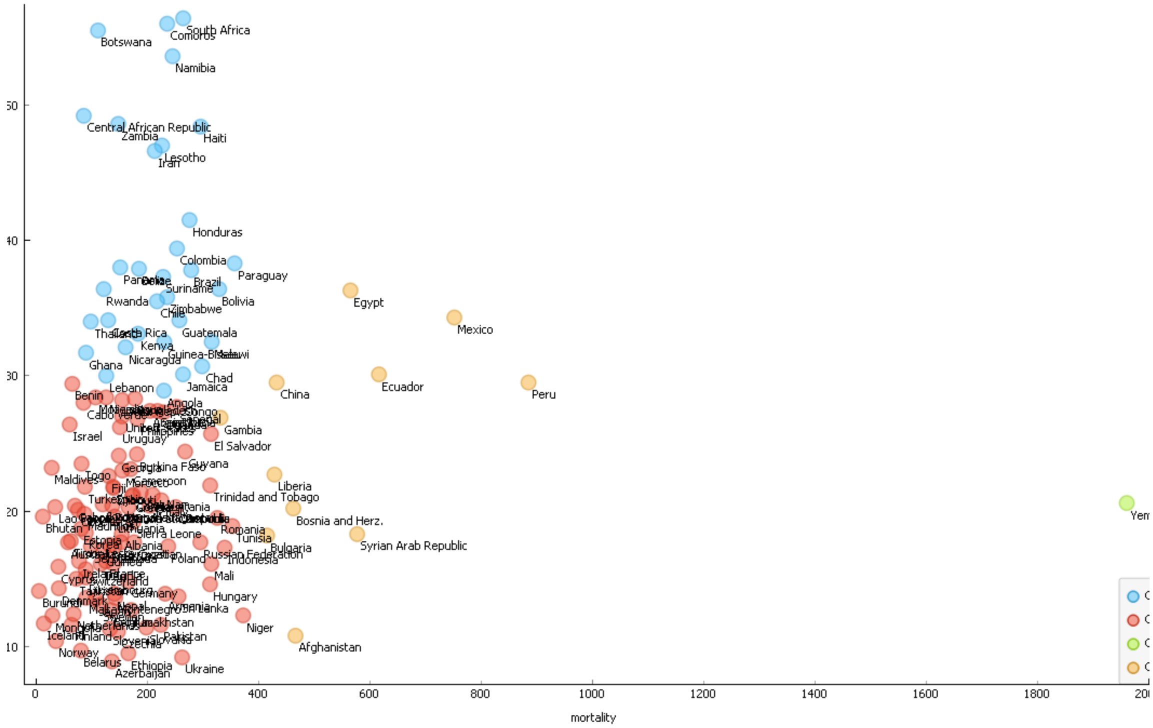


Fig. 13 K-means: Income inequality and Mortality:Scatter Plot

We can see countries with relatively lower inequalities and mortality have been placed in the red cluster and Yemen as an outlier forms its own single point cluster. These clusters are not so interesting, what's interesting is the blue and orange clusters. Orange cluster consists of countries whose mortality has been affected by income inequality while mortality of countries in blue clusters seems to have been completely unaffected by income inequality. Although Afghanistan, Bulgaria etc have more in common with red clusters, K-means seems to have placed them under orange clusters. Before we draw any conclusion let's use another algorithm to see more desirable clustering.

H. Hierarchical Clustering

First, we need to choose a linkage method. After running test with complete, single, average and ward's method and visualizing the result in dendrogram it becomes more obvious that clusters with ward's method are well exposed and simplified however the clustering result is almost identical to that of K-means which isn't helpful, on the other hand complete linkage seems to solve the issue we faced in k-means clustering.

In Fig.14 ,choropleth we can see not much has changed compared to k-means however countries like Afghanistan, Bulgaria etc have now been placed in red cluster while China, Egypt, Mexico etc have been placed under green cluster and what's interesting is that these countries are the ones with poor covid performance despite having adequate numbers of physicians as discovered early on. Since these countries also have relatively higher inequality given the size of their economy, we can make a safe assumption that economic inequality might have prevented people from accessing the proper physicians in the time of crisis and hence increasing the mortality.

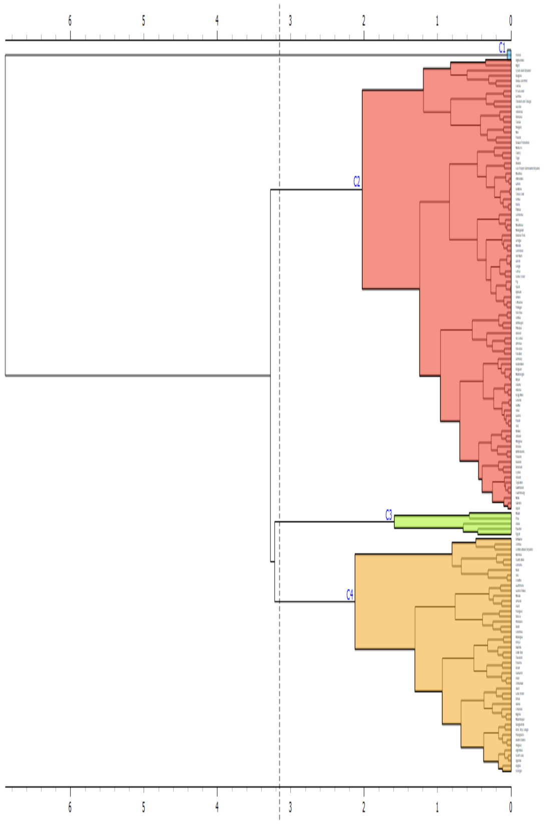


Fig. 15 Complete link: Income inequality and Mortality: Dendrogram

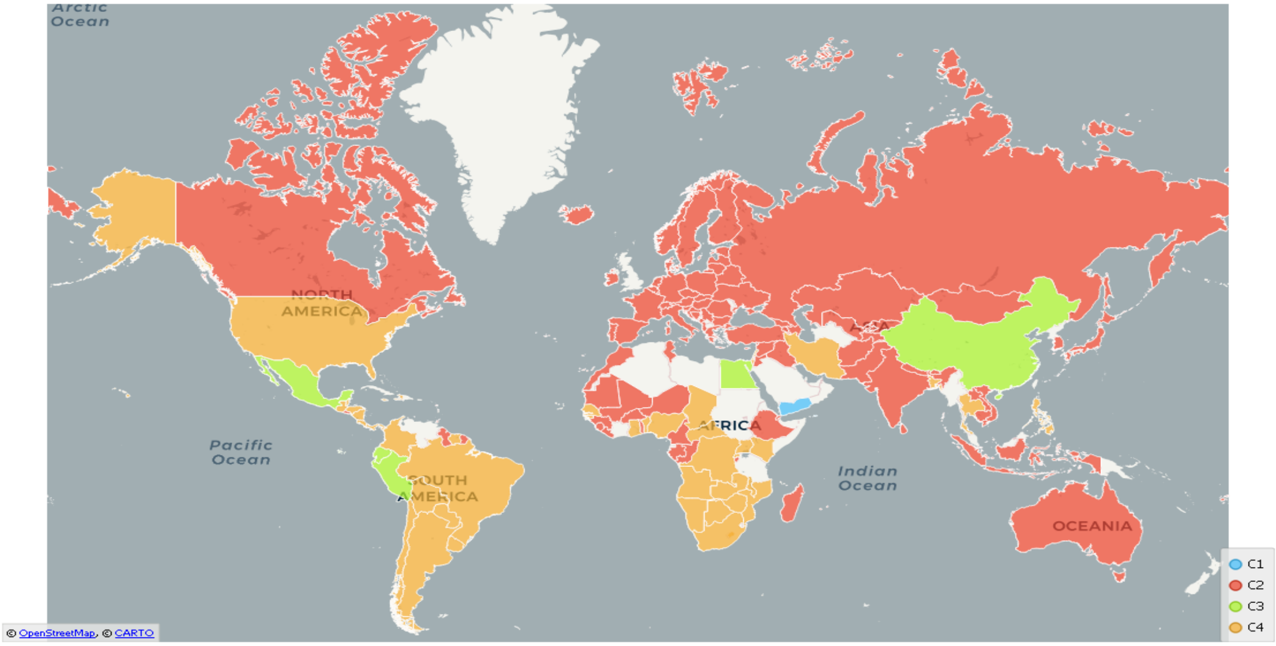


Fig. 14 Complete link: Income inequality and Mortality: Choropleth

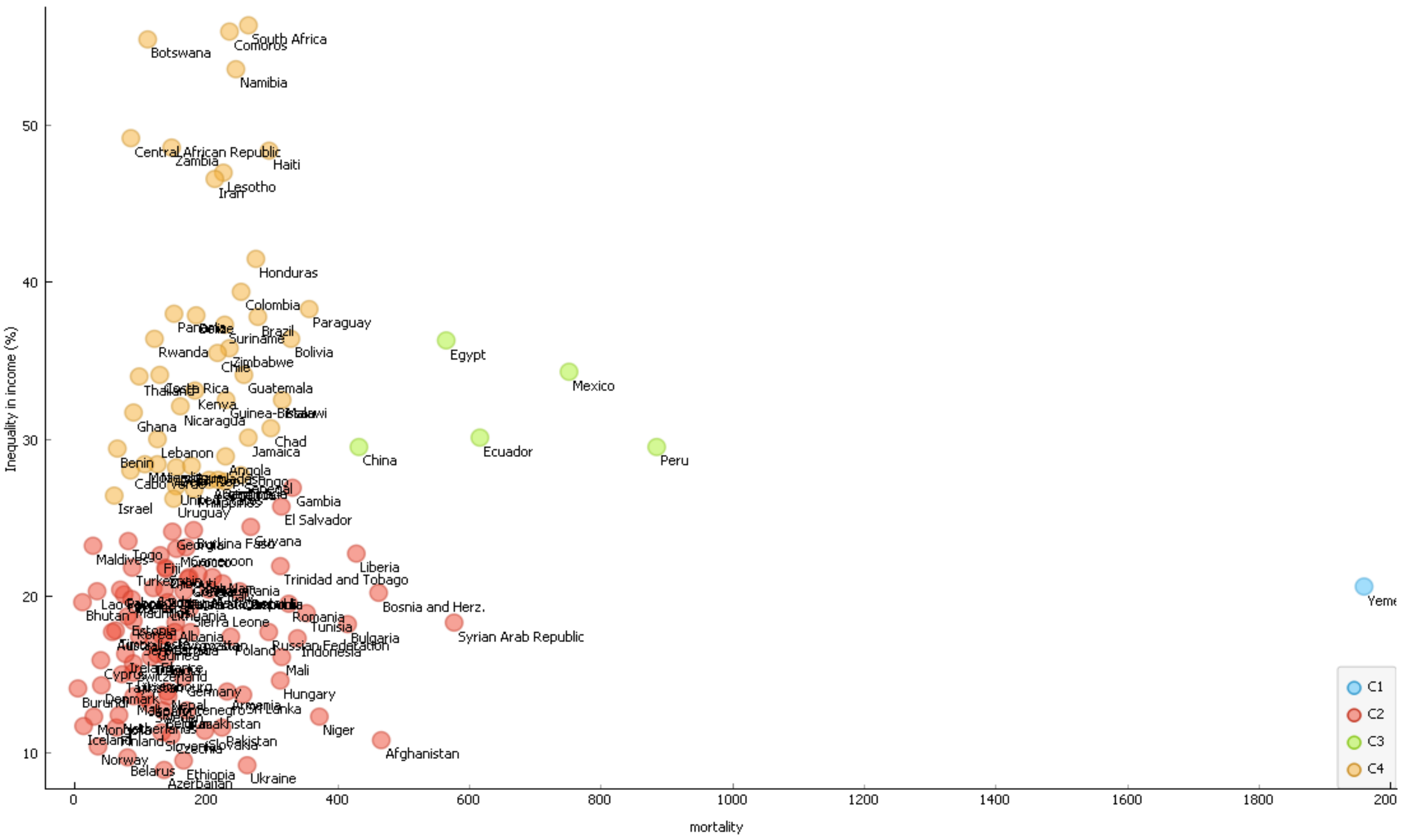


Fig 16: Complete-link:Income inequality and Mortality: Scatterplot

I. Outliers

Upon further investigating the data for outliers our initial assumption about green clusters seems to be solidified, their mortality has been severely affected by economic inequality compared to other countries. We can also see that Bhutan, Burundi and Iceland seem to be completely unaffected by increasing economic inequality which further solidifies our initial assumption from very first clustering analysis that it might be because these countries are extremely isolated from the rest of the world.

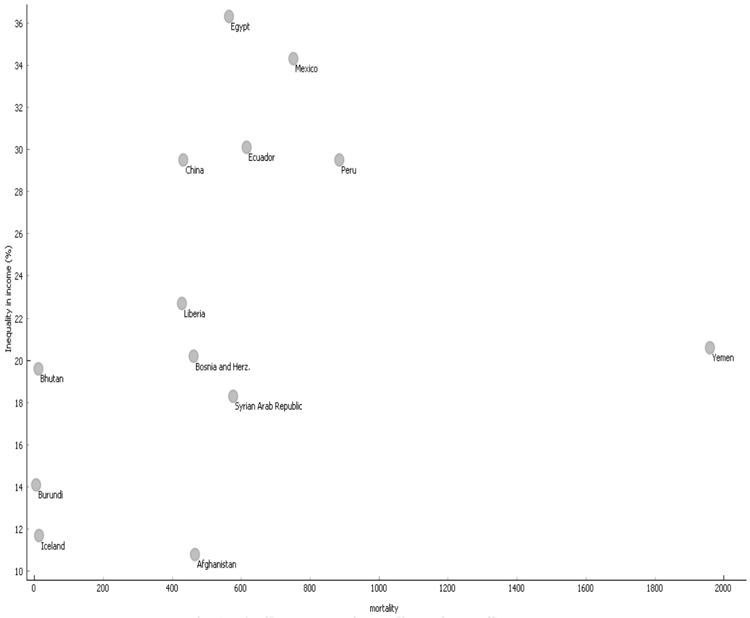


Fig. 17: Outliers: Income inequality and Mortality

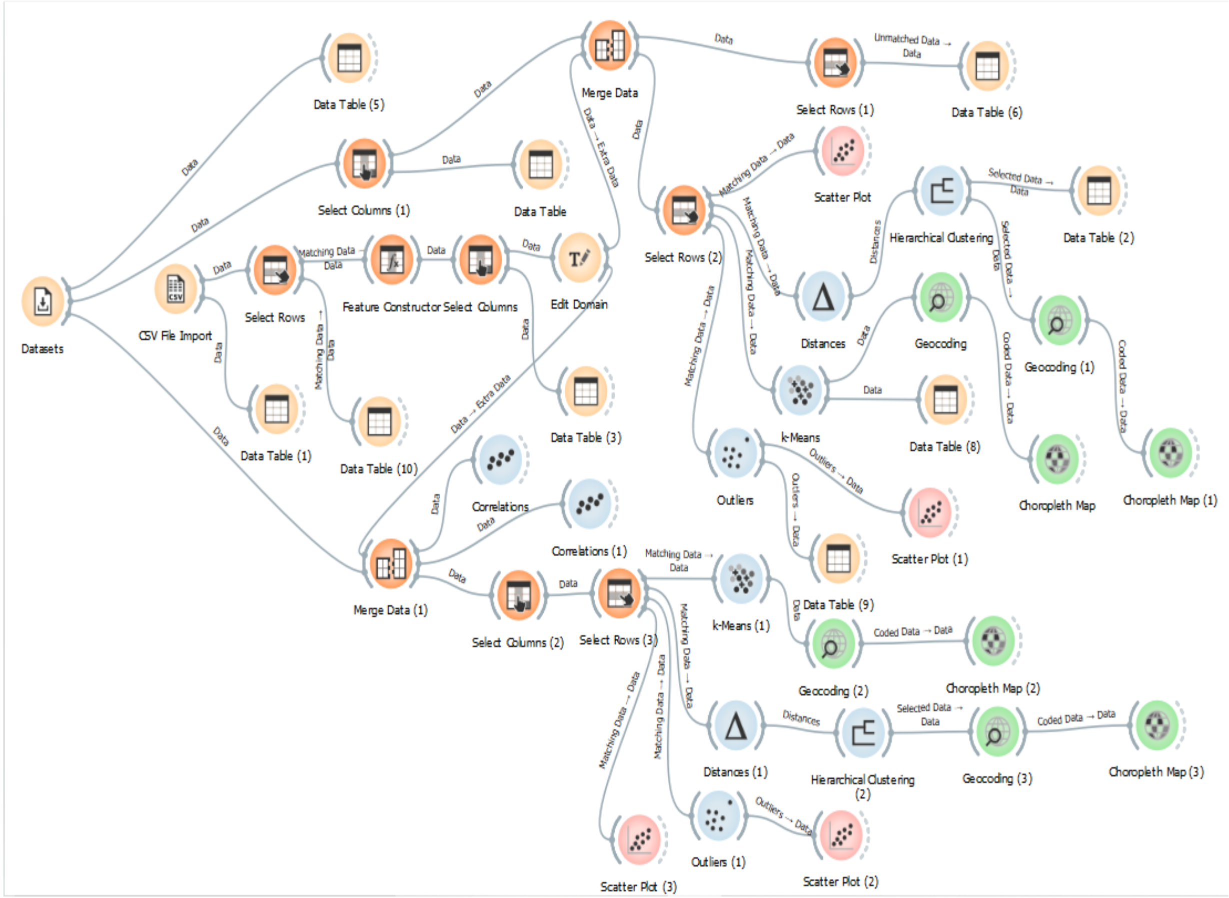


Fig. 18: Data Pipeline in Orange

IV. CONCLUSIONS

From above analysis we can conclude that unsupervised methods for clustering, correlations, distance measurement and outliers' detection prove useful tools for data analysis. As far as preference over clustering algorithm is concerned, the general rule of thumb is that K-means proves better for spherical data and hierarchical is better for globular data, however we saw from above analysis it depends on the types of clusters that one wishes to make and we wouldn’t know until we have tried most options

ACKNOWLEDGEMENT

We are thankful to the Department of Computer Science and Engineering, Kathmandu University. Also thankful to Assistant Professor Rajani Chulyadyo for providing us the opportunity to do this project and helping throughout the project morally and technically.

REFERENCES

[1]. Han, Jiawei, et al. *:Data Mining: Concepts and Techniques:Data Mining: Concepts and Techniques*. Third ed., Elsevier Science, 2011.

[2 .] Witten, Ian H., et al. *Data Mining: Practical Machine Learning Tools and Techniques*. Third ed., Morgan Kaufmann, 2011.