The University of Southampton

Academic Year 2020/2021

Faculty of Social, Human and Mathematical Sciences

Mathematical Sciences

MATH3092 Mathematical project

Performing unsupervised learning with clustering on value added by economic activity

I am aware of the requirements of good academic practice and the potential penalties for any breaches.
I confirm that this dissertation is all my own work.

Contents

Abstract

This report aims to perform clustering methods from unsupervised learning called Principle component analysis (PCA) and K-means clustering on value added by economic activity dataset and observe the results. This will be done by statistical programming in R by writing codes and using multiple R functions.

The dataset is pre-process and the metrics are plot against each other for observation, in which indicates 28 plots representing similar findings. PCA is then applied for dimensionality reduction resulting in reducing the plots into a 2-dimension graph represented by Principle component 1 and 2. An elbow plot is then performed to indicate the number of cluster necessary and lastly, K-means clustering is then implement on the 2-d plot of PC1 and PC2 for data segmentation.

In conclusion, after performing clustering algorithms the dataset is dimensionally reduced and then segmented in to 4 clusters and is represented by a 2-dimension graph.

1: Introduction

Economic is part of our everyday lives and an effort to measure it has been done through several metrics. This report will focus on the clustering of all countries who are members of the united nation(UN) by plotting an economic metric data set called value added by economic activity and the implementation will be done by using statistical programming to perform unsupervised learning methods.

Unsupervised learning is a powerful tool and is a type of machine learning in which can be used to identify patterns or segment data and its property to find similarities and differences in data indicates a highly useful tool in data visualization, data pre-processing, cross-selling strategies, customer segmentation, and image recognition *IBM Cloud Education (2020).*

This report will be focusing on the implementing of two unsupervised learning methods:

- Principal Component Analysis or PCA a dimensionality reduction method.
- K -mean Clustering a partitioning algorithm.

After that the results will be examine and discuss about the results, then a conclusion will be drawn out in the end. This report will only focus on the implementation of the algorithms on the dataset.

2. Background

2.1 Definitions and the data

Value added by economic activity is an economical metric which indicates the growth of money of given country by capturing the change in money using 8 different metrics.

The data in this report is from UNstats and represents the year 2019. The dataset contains the value added by 8 economic metrics from 243 countries and regions which are members of the United Nations (UN). The 8 metrics includes:

- Agriculture, hunting, forestry, fishing
- Mining, Manufacturing, Utilities
- Manufacturing
- Construction
- Wholesale, retail trade, restaurants and hotels
- Transport, storage and communication
- Other Activities
- Total Value Added

For deeper definitions and how each metrics are calculated can be found in the UNstats website (see in reference). The currency in this dataset is measured in US dollars and is set at 2015 price to avoid complications. Unsupervised learning will be performed on this dataset to observe on how the algorithms capture these 8 metrics.

2.2 Principal Components Analysis

PCA is a dimensionality reduction method, which after being performed most of the information in the data is still preserved. The method requires the use of linear transformation to create a new data representation, resulting a set of principal components.

Given a ***n x p*** data set of **X** we compute the first PC by assuming the variables in **X** has been centered to have mean zero.

$$z_{i1} = \phi_{11}x_{i1} + \phi_{21}x_{i2} + . . .+ \phi_{p_1}x_{ip}$$

(Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, (2013), page 375)

Next, we look for the linear combination of the sample feature values of the form that has largest sample variance, subject to the constraint that

$$\sum_{j=1}^{p} {\phi_{j_1}}^2 = 1.$$

Hence, the first principal component loading vector solves the optimization problem.

$$\underset{\phi_{11},\ldots,\phi_{P_1}}{\text{maximize}} \{\frac{1}{n}\sum_{i=1}^{n}(\sum_{j=1}^{p}\phi_{j_1}x_{ij})^2\}$$

The second principal component also finds the maximum variance, it is not correlated to the first principal component, and has an eigenvector that is perpendicular, or orthogonal, to the first component.

$$Z_{i2} = \phi_{12}x_{i1} + \phi_{22}x_{i2} + \ldots + \phi_{p2}x_{ip}$$

(Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, (2013), page 376)

$z_{11},\ldots,z_{n1}$ and $z_{12},\ldots,z_{n2}$ are the scores of the components. Now we can plot the PC against each other to produce a lower dimensional view of the data. This process repeats based on the number of dimensions, where a next principal component is the direction orthogonal to the prior components with the most variance. In this report PCA will use to reduce the dimension of the dataset and will be perform by prcomp() function in R.

2.2.1 Deciding how many principal components to use

There is no definite answer to this question. However, it is understanded that we normally decide the number of PCAs by examining a scree plot which is done by computing and plotting The Proportion of Variance Explained

$$\frac{\sum_{i=1}^{n}(\sum_{j=1}^{p}\phi_{jm}x_{ij})^2}{\sum_{j=1}^{p}\sum_{i=1}^{n}x_{ij}^2}$$

(Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, (2013), page 383)

against the principal component. We choose the smallest number of PCs that are required in order to explain a sizable amount of the variation in the data. This is done by looking (human observation) for a point at which the proportion of variance from each PCs drops of the scree plot this point is called an elbow. R programming will be performed in this report to show the scree plot.

2.3 K-means clustering

The aim of clustering is to organize a collection of unlabeled data into clusters such that items within a cluster are more similar to each other than they are to items in the other clusters. It is done so by using Clustering algorithms.

K-means is a type of Partitional Clustering method which decomposes a dataset into a partition which consist of K disjoint clusters.

Let $C_1,\ldots,C_K$ be sets containing the indices of the observations in each cluster. These sets satisfy two properties:

- $C_1 \cup C_2 \cup \ldots \cup C_K = \{1,\ldots,n\}$.
- $C_k \cap C_k = \emptyset$ for all $k \neq k'$

(Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, (2013), page 386)

K-means clustering will perform partitioning of a matrix of **M** datapoints and **n** dimensions into K clusters and the reason behind it is to minimize the within-cluster sum of squares *(Hartigan & Wong, 1979).* The following step is to maximize the division of the between sum of squares(BSS) with the total sum of squares(TSS). In addition, 1 should be the objective of the ratio BSS/TSS to get close to.

For the mathematical definition: Let **X** = {x_i}, i = 1,…,**n** be a set of points in the dimension which is going to be cluster into a set of k clusters. Also, let the mean of the cluster k be define as μ_k and the squared error between μ_k and C_k be define as:

$$\mathrm{J}(C_k) = \sum_{x_i \in c_k} |x_i - u_k|^2$$

The main objective here is to minimize the sum of square errors in all cluster K:

$$\mathrm{J}(C_k) = \sum_{\mathrm{k}=1}^{\mathrm{K}} \sum_{x_i \in c_k} |x_i - u_k|^2$$

(Jain, 2009).

K-means Clustering algorithm are describe as the following:

1. Initial clusters are randomly assigning the number of 1 to K to the observations.
2. Iterate until the changing in cluster assignments terminates:
    - Centroid of the cluster is computed for each of the K clusters. The kth cluster centroid is the vector of the p feature means for the observations in the kth cluster.
    - The following observations will be assigned to the closest cluster centroid (where closest is defined using Euclidean distance).

In this report the algorithm will use to partition the data after it is dimensionally reduce by PCA and will perform by kmeans() function in R. For all the commands see in R-code chapter.

2.3.1 Deciding number of clusters

The process of choosing the number K in K-means is done by subjection. However, there is a tool to help indicate the optimal number of K and in this report it will be done by using the Elbow method. It is done by plotting an Elbow plot; the within-cluster sum of squares (WSS) against the number of clusters. The variability of the observations within each cluster is measured by the within-cluster sum of squares and is represent by an Elbow plot. An observation on where the point of graph drops drastically is evaluate and is picked to be the number of K. In this report the elbow method will be done by R programming.

3. The implementation of unsupervised learning on the data set

In this section, the implementation of unsupervised learning on the dataset is done through R programming and the function set.seed() in R is also used to ensure that the results can be replicate. Full details of the code can be found in the R code section.

3.1 Data pre-processing

The view() function is apply to the dataset and from the observation the first 3 columns which are Country.Area, Year and Unit are not necessary for usage. Hence, elimination of these columns is performed. NA values are put to 0 by hand and due to the very low amount of NA, this will not affect the overall calculations in this report. As the data are measured on the same scale, standardization is not needed.

3.2 Plotting the metrics against each other

For observation, the 8 metrics describe in section 2.2 is plotted against each other:

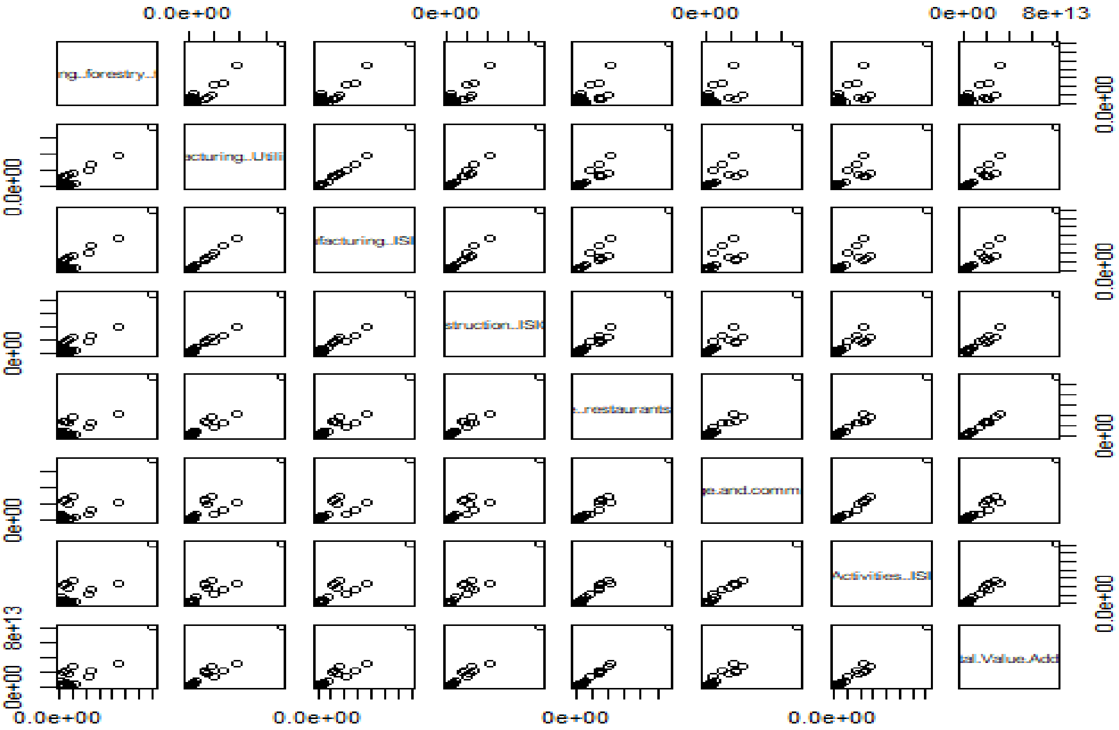

Figure 1 Value added by economy metric plotted against each other

There are 2 halves of 28 plots where indicates the same thing and the plots shows a trend of correlation and most are not scattered. As these plots indicates similar results, a better representation of these results is necessary. Hence, a dimensional reduction is the next step.

3.3 Dimensionality reduction by PCA

A function called prcomp() is applied to the dataset to perform PCA and a Scree plot is also implement to show the number of necessary PC. See Figure 4 for the Scree plot:

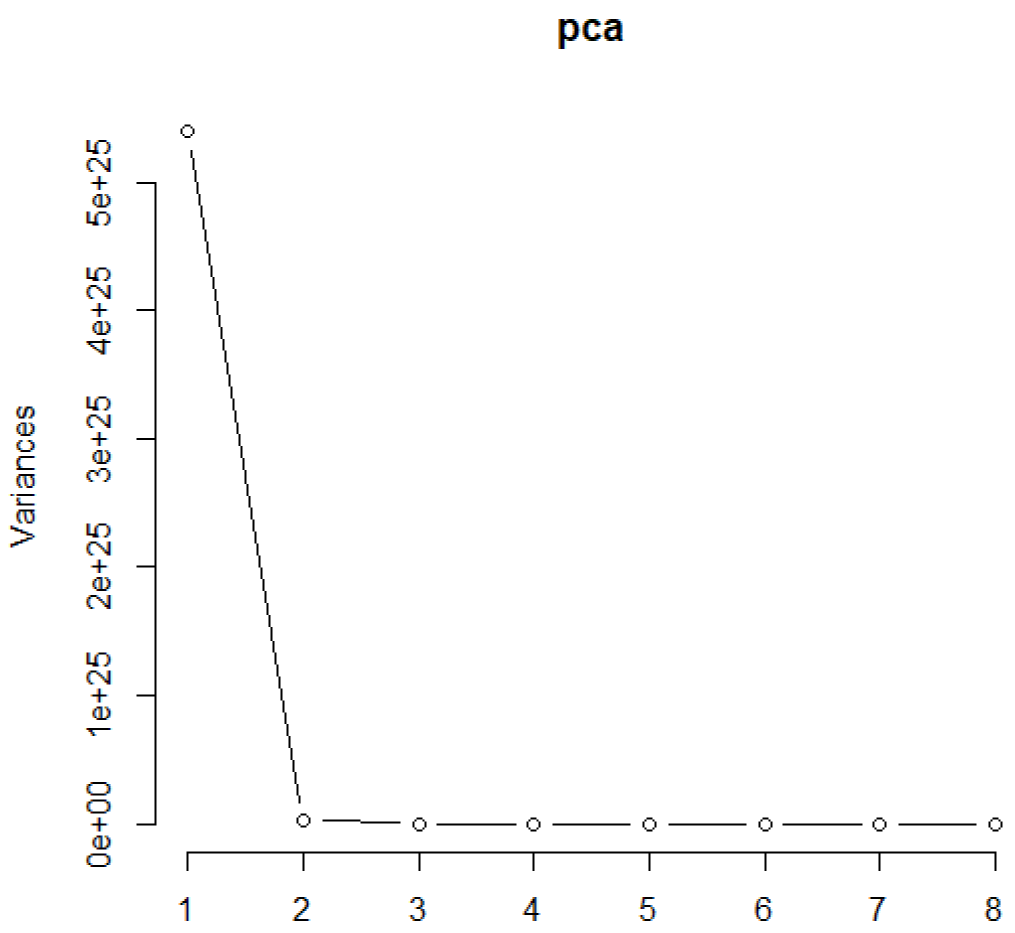

Figure 2 indicates the scree plot

As shown, the variances drop drastically after the 2nd PC. Hence, PC1 and PC2 will captures most of what the observation represents. For a detailed version of the coding part please refer to the R-code section.

PC1 and PC2 will be represented by a 2-d plot shown in figure 3:

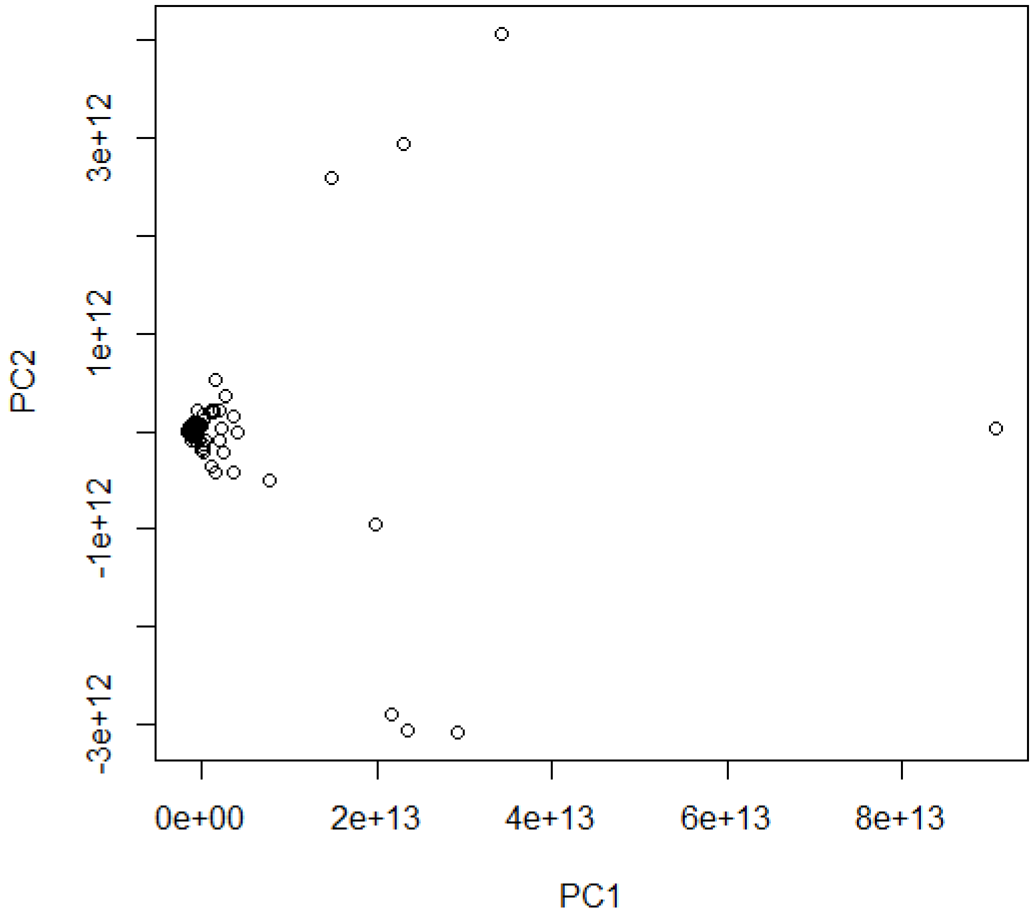

Figure 3 represents the plot of PC1 and PC2

PCA has reduce the dimension of Figure 2 from 28 two-dimensional to only 1 plot which also represents the data without losing much of the data as shown in figure 4.

3.4 Applying K-means clustering on PCA

To perform K-means, defining the number of clusters is an essential first step. As explain in section 2.3.1 an elbow method is performed to observe the number of K. See figure 4 for elbow plot:

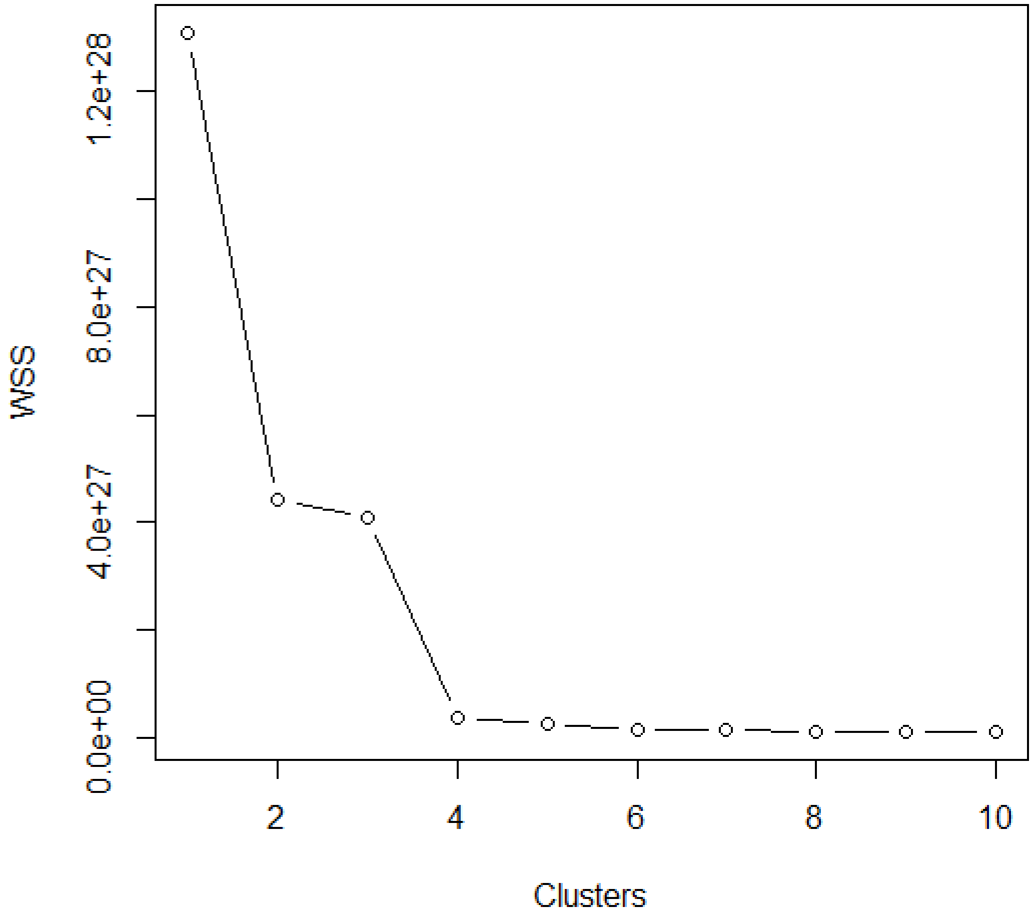

Figure 4 Elbow plot- shows the plot of WSS against the number of clusters

From observing the plot its shows that the point where the WSS drops the drastically is at the 4th cluster. Hence, the number of cluster K is 4 in this case. For the code used to produce this plot please refer to the R-code section.

Data segmentation by K-means is implement on the plot of PC1 and PC2 by kmeans() function in R. Also as shown in the figure 4, the number of cluster K is 4.

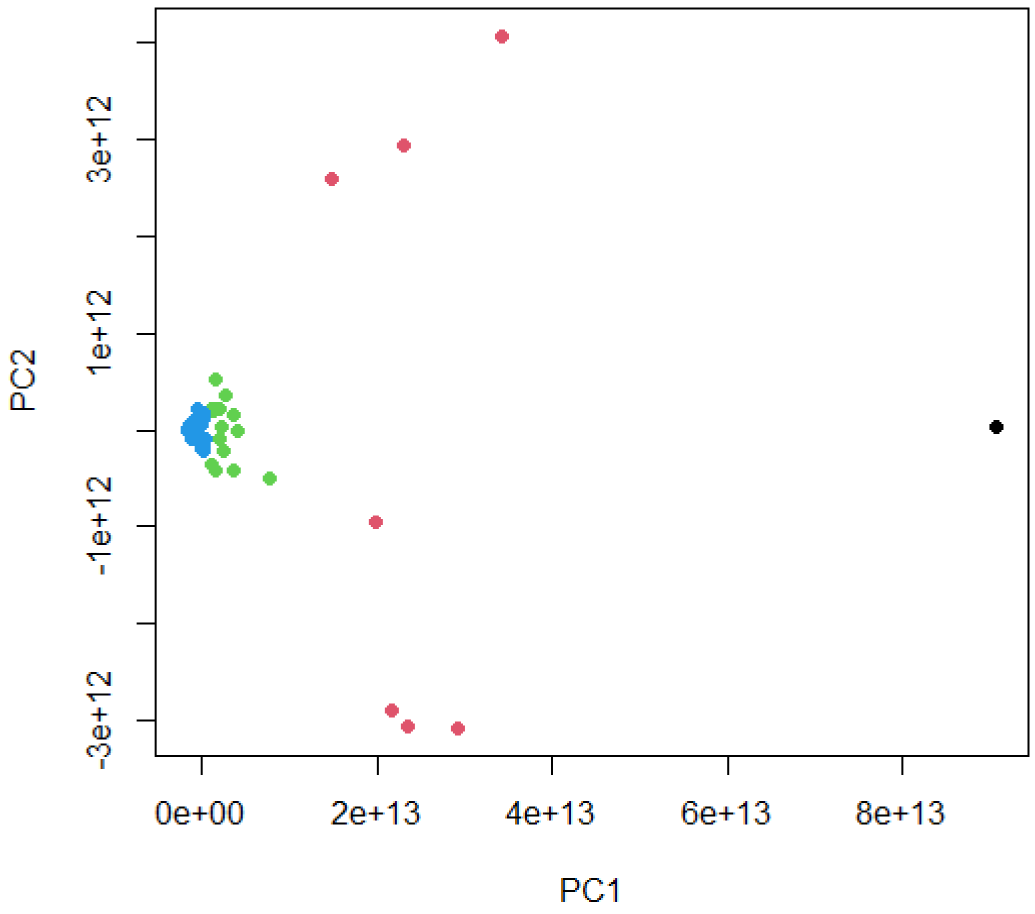

Figure 6 shows the implementation of K-means clustering algorithm on the plot of PC1 and PC2

To see which country belongs to which cluster and the centre of each cluster can be called by R command (see in r-code section). The implementation of Unsupervised learning method on the value added by economic activity is now completed. The dimension is reduced by PCA and segmented by K-means in which is all is performed by R programming. The size of the 4 clusters are as following:

- 7 (pink)
- 1 (black)
- 219 (blue)
- 15 (green)

4: Discussion and Conclusion

4.1 Discussions

The objective of this report has been completed as unsupervised learning is used to capture on how the 8 metrics perform relative to the 243 counties. We have use PCA to reduce the dimension shown on figure 3 and a partitioning is done by K-means show figure 6.

However, from the implementation of unsupervised learning; observations have been made and it indicates that this dataset might not have enough information to perform unsupervised learning to its full potential as shown in figure 1 in section 3.2. The data is not scattered but rather mostly group in the beginning and some of the plots shows a trend of correlation. Although the data might not contain enough information to fulfill the algorithm's potential but it still shows the property of both PCA and K-means which are dimensionality reduction and data segmentation.

4.2 Conclusion

In this report, we have complete the implementation of unsupervised learning on the dataset as it reduced the dimension of the 28 plots of the metrics from value added by economic activity of 243 countries using Principle component analysis and after the data is segment on it by using K-means clustering algorithm. The countries in the dataset are segmented into 4 clusters.

R- code used

```
setwd("<path>")


data.va  <- read.csv("Results.csv" , header = TRUE)


set.seed(1) #This is to produce the same results everytime it is being
re-done.


#Data pre-processing

View(data.va)


value.added <- data.va[, -c(1,2,3)]


plot(value.added)


#PCA

pca <- prcomp(value.added)


#Scree plot

plot(pca, type = "l")


#PCA plot


pca$x


summary(pca)
```

```
plot(pca$x[,1:2])

#elbow
k <- list()
for(i in 1:10){
  k[[i]] <- kmeans(pca$x[,1:2], i)
}

WSS <- list()
for(i in 1:10){
  WSS[[i]] <- k[[i]]$tot.withinss
}

plot(1:10, WSS, type = "b",
     ylab = "WSS", xlab = "Clusters")

#K-means on PCA
pca.kmeans <- kmeans(pca$x[,1:2], 4)
plot(pca$x[,1:2], col = pca.kmeans$cluster, pch = 19)
pca.kmeans$cluster #To see how country members are assign
pca.kmeans$centers #To see the centres of each cluster respect to PC’s
```

Bibliography

[1] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, (2013), An Introduction to Statistical Learning : with Applications in R, New York :Springer.

[2] Derek Greene, Padraig Cunningham, Rudolf Mayer, (2008), Applied and Computational Mechanics, Lecture notes.

[3] IBM Cloud Education (2020), Unsupervised Learning, IBM Cloud Education, viewed 12 January 2021 < https://www.ibm.com/cloud/learn/unsupervised-learning>

[4] Jolliffe IT, Cadima J. ,(2016), Principal component analysis: a review and recent developments.

[5] United Nations Statistics Division, 2019, Value Added by Economic Activity, at constant 2015 prices - US Dollars, < https://unstats.un.org/unsd/snaama/basic>

[6] United Nations Statistics Division, 2020, Glossary,

<https://unstats.un.org/unsd/snaama/Metadata/Glossary# >

[7] Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988) The New S Language. Wadsworth & Brooks/Cole.

[8] Mardia, K. V., J. T. Kent, and J. M. Bibby (1979) Multivariate Analysis, London: Academic Press.

[9] Venables, W. N. and B. D. Ripley (2002) Modern Applied Statistics with S, Springer-Verlag.

[10] *Forgy, E. W. (1965). Cluster analysis of multivariate data: efficiency vs interpretability of classifications. Biometrics,* ***21****, 768–769.*

[11] *Hartigan, J. A. and Wong, M. A. (1979). Algorithm AS 136: A K-means clustering algorithm. Applied Statistics,* ***28****, 100–108. doi: 10.2307/2346830.*

[12] *Lloyd, S. P. (1957, 1982). Least squares quantization in PCM. Technical Note, Bell Laboratories. Published in 1982 in IEEE Transactions on Information Theory,* ***28****, 128–137.*

[13] *MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, eds L. M. Le Cam & J. Neyman,* ***1****, pp. 281–297. Berkeley, CA: University of California Press.*

[14] *Jain, A. K. (2009). Data clustering: 50 years beyond K-means. Pattern Recognition Letters 31(8), 651- 666.*