Department of Financial Mathematics

Final Year Project
Dissertation of Bachelor of Science in Financial Mathematics

Unsupervised Learning methods and its Applications on the spread of Covid-19: A country level clustering study

无监督学习聚类方法及其在新型冠状病毒肺炎研究中的应用

Name: Yeyu Wang
ID Number: 1715617

Supervisor: **Dr. Mu He**

5th May 2021

Abstract

Outbreak of Covid-19 in 2020 has brought a disaster to countries around the world. This project aims to apply unsupervised learning techniques into the study of Covid-19. Three methodologies including Principal Component Analysis (PCA), K-means clustering and hierarchical clustering are applied to cluster 27 most aggreesively affected countries based on data in areas ranging from economy, medical level, geographical features to Covid-19 situation. The results demonstrate that: 1. PCA and K-means clustering and hierarchical clustering suit this dataset well. 2. Both hierarchical clustering and K-means clustering group countries into 5 clusters and they are mainly clustered based on geographic locations. One thing need to be noted is that USA is put in a separate cluster by using both of these clustering techniques and Mexico is put in a separate cluster on the results of K-means. The analyses and discussions of these results might give health sector officials, policy makers and economy experts some insights.
Keywords: Principal Component Analysis (PCA), K-means clustering, Hierarchical clustering, Covid-19

摘要

新冠肺炎疫情的爆发使得全球陷入灾难。这个项目的主要研究目的是将无监督学习技术运用到新冠疫情的研究并得出一些结论。这篇文章针对四个方面的 13 个不同种类的数据，采用包括主成分分析（PCA），K-means 聚类以及层次聚类方法对 27 个受新冠疫情影响严重的国家进行聚类。结论表明：1. 对于本数据集 PCA 能有效的实现降维，同时两种聚类方法都能得出不错的结果。2. 两种聚类结果基本都是基于地理位置给国家聚类。需要注意的是，美国在两类方法中都被单独归到了一类，墨西哥则在 K-means 的结果里被单独归为一类。聚类结果的观察和分析也能为各个国家的卫生部官员，政策制定者和经济学家们提供想法和建议。
关键词: 主成分分析 (PCA), K-means 聚类, 层次聚类, 新型冠状病毒肺炎

Contents

1 Introduction

War, famine and plague have always been the 'sword of Damocles' since the day mankind were born. The Spanish flu, prevailing from 1919-1920, infected nearly 500 million people and caused an estimated 20 million deaths. Also, Smallpox made its debut in 3rd century B.C. and killed 300 million people in the 20th century. Countless doctors had made efforts to find a cure during their lifetime before it was eliminated in 1980. Throughout history, except the two famous infectious diseases mentioned above, there still exists many plagues that took countless people' s life including numerous medical personnel. Thus, some people concluded that the history of human beings is mainly about the history of fighting with plagues, and it seems to be even more sensible after the outbreak of Covid-19.

The coronavirus, outbroke since January 2020, has now swept almost every country in the world. As of 26th April 2021, totally 146841882 cases of Covid-19 have been confirmed including 3104743 deaths [1]. According to the website, the majority of these cases happened in Americas and Europe. However, even the countries that have been struck less aggressively by Covid-19 are still under a lot of stress because the disease brought huge impacts and great challenges to not just some countries but the entire human society.

1.1 The impacts on health care and medical system

The high volume of confirmed cases brought huge burden to the functioning of existing medical system. Doctors, nurses and other medical staff in public hospitals are overloaded with more tests and diagnoses, leaving them physically and mentally drained. Small-sized community hospitals are to some extent caught in a dilemma: They want to take care of more patients and ease the burden on public hospitals, but limited by lack of good doctors and equipment, they can not afford treatments for large number of people. Besides, due to the fact that the majority of attention has been paid to Covid-19, patients with other disease sometimes get neglected and their optimal timing of treatment maybe missed. Lastly, Covid-19 is reckoned to bring negative effects on people's mental health. Owing to some lockdown policy and remote work requirements, many people feel distressed, emotional exhaustion or lonely because they have less social support. Based on data from the Centers for Disease Control and Prevention, the percentage among surveyed adults in U.S. who feels like to have symptoms of anxiety and depression increased from 19% in 2019 to 42% in 2020 [2].

1.2 The impacts on world economy

Most countries experienced economy recession during the past 2020, leading to the -15%- -5% real GDP growth [3]. The downturn, according to IMF, is the worst since the Great Depression of the 1930s. Many people lost their jobs or businesses because of this recession. The unemployment rate even peaked up to 14.8% in U.S., and in accordance with a survey of over 5800 small U.S. business, 41.3% of them were temporarily closed because of Covid-19. Even though the economic situation is heating up, many people are still living in misery, and countries hit hard by Covid-19 like UK or Italy may take lots of time to recover.

1.3 The impacts on food systems

Farmers have been unable to enter markets to sell products or buy resources due to border closures and trade restrictions, which disrupted domestic and foreign food supply chains and the balance of food system. World Food Programme (WFP) forecasted that the number of people in food crisis or worse will reach 265 million because of Covid-19 [4].

1.4 Current solutions and researches

Even though the situation of Covid-19 is still serious today, people have been trying to fight with it and made some progress. Vaccines have been developed and used, scholars in different areas have been doing researches and try to find solutions or strategies. Zarikas et.al [5] used Hierarchical clustering and used Euclidean distance between time series to cluster countries around the world with respect to active cases and population and land area. Fang & Ding [6] used Principal Component Analysis (PCA) and K-means to cluster countries in China with respect to indexes like economic development, geographic characteristics and disease situation. Azarafza et al.[7] used KDE and K-means to estimate infection pattern between provinces and concluded that travelling is the main factor for the outbreak of Covid-19 in Iran and Tehran city is responsible for the spread of Covid-19 in Iran. Mahmoudi et al.[8] used fuzzy clustering method to cluster high risk countries and the results showed that the spread in Spain and Italy is similar but it is different from other countries. Wang et al. [9] examined the spatiotemporal clustering of Covid-19 in the United States during the first 16 epi-weeks by using Spatial and Space-Time Scan Statistics. They found that people in metropolitan and counties near core airports have higher risks of getting Covid-19. In Zhang and Li's paper [10], they improved classical SIR epidemic model by separating people into four clusters and established differential equations for each cluster. They then used Runge Kutta method to simulate the relationship between the proportion of each cluster and time by setting different contact rate. Zhu et al. [11] improved SIR model by introducing the infection coefficient, which is a liner function of time t. Lin et.al. [12] used K-means to cluster 162 countries into 6 clusters and set confirm, dead, dead rate, heal, heal rate as cluster centers. They also used ARIMA (1,1,1) to make a 10-steps prediction. The accuracy can even reach to 99%. In the research of Leichtweis et al.[13], they found that there exists a linear relationship between reproduction number and temperature, they also discovered that solar radiation has a negative relationship with the dissemination of Covid-19.

Motivated by the former researchers and based on the fact that few researchers have tried to use knowledge of unsupervised learning to achieve some meaningful results, we wish to use these techniques to discover some useful information behind these data.

The main contribution of this paper is to use knowledge of unsupervised learning, mainly clustering and dimensionality reduction techniques, to get common features of 27 most affected countries based on data in areas ranging from economy, medical level, geographical features to Covid-19 situation. The results might be useful for health sector officials, policy makers and economy experts. In the next section, literatures related to clustering, K-means, Hierarchical clustering and dimension reduction will be reviewed. Section 3 presents principles of methodologies used in this project. Section 4 graphically displays and explains data and the results of three techniques mentioned above. Finally, the conclusion will be made and future research suggestions based on limitations of this project will be mentioned.

2 Literature review

2.1 Overview of machine learning

In early stages of data analysis, due to the lack of computing tools and epoch-making knowledge, this area is not a fashion. However, in recent years, with the upsurge of machine learning and deep learning based on using modern computers to deal with big data, researches in the field of data analysis are advancing at an unprecedented speed. Most frequently discussed knowledge of machine learning can be divided into three types: supervised learning, unsupervised learning and semi-supervised learning. Among many techniques, the core of supervised learning is classification and the core of unsupervised learning is clustering.

The inspiration of classification and clustering is that people are born with the habit to keep similar things together and set different things apart. For example, if the owner of a fruit store laid in a new stock of fruits, he will intuitively put apples with apples and oranges with oranges. Meanwhile he will set these two kinds of fruits apart. He arranges them in this way because apples, oranges are both like a label, and we just put things with the same label together. This process is kind like classification. However, in real life, we do not really know everything and thus we can not tag everything with a label. As a result, scientists developed another technique: clustering. Clustering is to put things without labels but are similar together. For example, in many fruit shops, owners set customers who buys fruits with red skin together and who buys green skin into another cluster. This is because they have something in common: the fruits they bought have the same skin. By doing so, if one customer come again, owner can recommend the customer with something he might need and make more profits.

Compared with classification, clustering is more difficult because there is no such clue as the label in classification to guide us to group data. However, clustering can get results without training. Thus, we tend to use this technique if we have no prior information and knowledge on this dataset.

2.2 Overview of clustering

Everitt et al. [14] commented in their book that there is no entirely clear definition and no need to define clear definitions on the term clustering. However, in the book of Jain & Dubes [15], they mentioned that instances in the same cluster should be as similar as possible and instances in different clusters should be as different as possible. Meanwhile, the similarity and dissimilarity measurements should be clear and make sense. Also, a simple definition is defined in the article of Saxena et al. [16], they defined clustering as a technique so that objects are grouped based on some similarity inherent among them. The definition is simple. Meanwhile, it is actually the ultimate goal of clustering: dividing an unlabeled dataset into a set of natural data structure, i.e.: things grouped in the identical group are more related and similar to each other [17]. Based on the measure of similarity, clustering can achieve many goals. For example, it can help group chaotic and erratic data, identify specific distributions from them [18], reduce the quantity of data and to better understand data structure by discovering the internal pattern and feature space hidden behind them [19].

Clustering can be divided into many different types. For example, hierarchical clustering, including Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH), Clustering Using Representatives (CURE), Robust Clustering Using Links (ROCK) and Chameleon, which will all be

introduced in the section 2.4. The advantage of hierarchical clustering is that it can be applied on the data set with any shape. The disadvantage of it is that the time complexity is relatively high [20]. Partitional clustering, including classical algorithms like K-means, K-medoids, Partitioning around medoid (PAM). Besides, Frey et al. [21] presented a new partitional catering algorithm: Affinity Propagation clustering (AP). It is efficient and the computational time is only 1% of other algorithms when dealing with large-scale data. The advantages of it are that the time complexity is low while the computing efficiency is high. The disadvantages of this type are that it can not be used to deal with non-convex data and it is sensitive to the outliers in the data. Also, to set the number of clusters before using is another thing that needs to be carefully considered. Density-based clustering, including algorithms like DBSCAN algorithm, SNN algorithm, OPTICS. The core of them is that data in the region with high density can be grouped into the same cluster. For DBSCN, it is an algorithm whose pre-set radius neighborhood of each point in one cluster should have at least a pre-set number of points. Density-based clustering algorithm is suitable for dealing with large data sets with noise or with arbitrary shape. However, this algorithm will output bad clustering results if the data space's density is not even [22]. Fuzzy theory was first applied by Ruspini, who proposed Fuzzy C-means (FCM), in 1969 into clustering analysis. Relatively high accuracy of this type of clustering comes with relatively low scalability. Also, like partitional clustering, this algorithm might fall into local optimal. Grid-based clustering, which changes original data space into a grid structure so that users can compute the statistical information of points in the grid to clustering, is another clustering technique. Classic grid-based clustering algorithms including STING and CLIQUE. The advantages of grid-based clustering are that it has low time complexity and a relative high scalability. However, the result of this clustering is sensitive to the mesh size. Some researchers also developed clustering techniques based on graph theory. For example: CLICK [23] and MST-based clustering. The idea is to treat data points as nodes and the relationship among data points as edge. Nevertheless, if the graph complexity increase, the time complexity will increase dramatically.

The most discussed problems between researchers are how to determine the number of clusters and how to assess the similarity between points. The solutions of these problems are also the key factors to determine the quality of clustering. Rezaee mentioned that the optimal k should be in the interval of $(1, \sqrt{n})$. One simple rule of thumb is to set the number of $K = \sqrt{\frac{n}{2}}$, while n is the number of objects. Another rule of thumb is called 'Elbow criterion'. The criterion states that if adding another cluster does not significantly decrease total within cluster sum of squares, then it should be the number of clusters you choose. Ball & Hall [24] came up with a new algorithm: Iterative Self-organizing Data Analysis Techniques Algorithm (ISODATA). This algorithm dose not set a fixed number of clusters, instead, the number changes all the time by merging and splitting clusters, which makes the result more accurate. Besides, there still exists other criteria such as akaike information criterion (AIC), bayesian information criterion (BIC) or deviance information criterion (DIC) to determine the number of clusters.

If we are dealing with continuous data, most of the time we will use Euclidean Distance to measure the distance between data points, and from which the cluster obtained tends to be spherical. Sometimes we will use Minkowski Distance or Mahalanobis Distance, but it takes lots of time to compute (Wang et al., 2012). Banerjee et al. [25] proposed an idea of using Bregman divergences as distance

measurement, and it has advantages like linear computational complexity and can be applied to mixed data types.

Each clustering algorithm, given a data set, will always crate a segmentation regardless of whether it is true or not. Therefore, although clustering analysis can be a subjective process, it is meaningful and necessary to develop some evaluation criteria to test the validity and accuracy of it. Since the clustering results follow the principle that 'within cluster similarity should be big, between cluster similarity should be small', many evaluation criteria are also based on this principle. Usually, evaluation criteria can be divided into three types: External criterion, internal criterion and relative criterion. External criterion is to compare the results of clustering with some prespecified structure, which are obtained from some prior knowledge or experience. Internal criterion is to use the information gained from dataset, like CoPhenetic Correlation Coefficient, to assess the clustering structure. Relative criterion is to compare different clustering structures and choose the best one. In Jin's paper [26], he commented that external criterion is better than relative criterion, and relative criterion is better than internal criterion.

The applications of clustering are wide. In marketing area, clustering can be used to group customers with similar shopping behaviors. In biology area, researchers will use this technique to cluster plants and animals based on their features. Some researchers in the area of bioinformatics used clustering and made some breakthrough. Herwig et al. [27] created algorithms based on K-means to group a set of 2029 human cDNA clones. Tavazoie et al. [28] used K-means to partition 3000 genes into 30 clusters. In the area of pattern recognition, clustering is mainly used in voice recognition and character recognition. In the area of machine learning, it is used in image segmentation and machine vision. Lastly, clustering can be applied to agriculture. Agglomerative clustering was implemented in precision agriculture, which raised productivity [29].

2.3 K-means clustering

K-means clustering was first proposed by Macqueen [30] and has been widely used ever since. The algorithm of it will be introduced detailly in the methodology part. However, K-means still have shortcomings and many researchers have been exploring and improving it.

Bradley and Fayyad [31] presented a refined initial starting condition to let the algorithms like K-means converge to a better result. Pelleg and Moore [32] invented X-means to accelerate the process of each iterative. Huang [33] offered a new algorithm: K-modes, which can be used to deal with categorical data. Instead of using the mean of each cluster as the center point, this algorithm uses mode of each cluster as center point. Also, Euclidean distance is replaced by Hamming distance. Sun et al. [34] applied refined initial points into K-modes and found that the result is more precise. Ding and He [35] introduced nearest neighbor consistency into clustering and the clustering accuracy is improved. In Yang et al.'s paper [36], they constructed a distance cost function to get the optimal number of clusters. Likas et al. [37] proposed a global k-means algorithm. Different from randomly choosing initial values, this algorithm attempts to add one new cluster center at each stage. Gu [38] used subtractive clustering to determine the center of clustering and combine this with Locality Sensitive K-Means to improve the original algorithm. Shi and Xu [39] combined genetic algorithm with K-means algorithm so that the accuracy is improved.

2.4 Hierarchical clustering

Hierarchical clustering is one of the most frequently used method among all clustering techniques. The algorithm of it will be introduced in the methodology part. However, just like K-means, this method still has shortcomings and many researchers have tried to improve it. BIRCH was invented by Zhang, Ramakrishnan and Livny in 1996 [40]. They designed a new way to store the summaries of original data while reducing the storage space, and called it the clustering feature tree (CF tree). Each node of the CF tree represents a subcluster. After building the CF tree, hierarchical clustering can be applied to the summaries. In order to group data into different shape of clusters, Guha et al. [41] created a new algorithm: CURE. This algorithm uses some representative points to replace the center point of a cluster. Although CURE and BIRCH both can deal with outliers well and the accuracy of CURE is better, the computational complexity of CURE is higher and it can not handle categorical data. In 1999, Guha and others introduced ROCK algorithm [42], which is an improvement of CURE and can handle categorical data. This algorithm does not use any distance function. CHAMELEON is an algorithm proposed by Karypis in 1999 [43]. This algorithm uses dynamic modeling in the process of hierarchical clustering. It can deal with numerical data or categorical data, but it can not be applied to high dimensions. In 2007, Gelbard [44] presented a new algorithm: Binary-positive algorithm. It transforms original data into binary data and the measurement of similarity is only conducted between the objects whose value equals to 1. This algorithm has a relative high accuracy and robustness.

2.5 Dimension reduction

In real life, people usually deal with data in three-dimensional space. Therefore, when we encounter data whose dimension is higher than three, it is difficult for us to immediately identify the characteristics and structure of them. In fact, the dimensions of high-dimensional data range from tens to hundreds, thus traditional clustering methods are time-consuming and the result is not very ideal. To solve this problem, researchers presented technique called 'Dimension reduction'. Dimension reduction is to use linear or non-linear transformation to embed data from high dimensional space into low dimensional space so as to eliminate the influence of irrelevant dimensions and retain meaningful properties of original data. Classical methods including principal component analysis (PCA), linear discriminant analysis (LDA) which deals with linear structure. ISOMAP, local linear embedding (LLE). Kernel PCA are typical manifold learning algorithms, which are used to deal with non-linear structure [45].

2.6 Clustering on the age of big data

With the explosive growth of data in all industries, big data has been a popular concept in recent years. A subjective but simple definition of big data is defined as follows: Any data that can not be loaded into your computer's working memory is big data [46]. Typical characteristics of big data is 4V: big in Volume, big in Variety, high in Velocity of processing and important in Value. Thus, due to the 4v mentioned above, meaningful information might be hidden behind massive and messy data and we need to use new techniques to find them out. Zhao et al. [47] proposed parallel K-means algorithm based on MapReduce. The performance of this algorithm is good based on three evaluation criteria: Speedup,

sizeup and scaleup. Havens et al. examined the effectiveness of three clustering techniques which aims to extend FCM to big data. The results showed that random sampling plus extension FCM, bit-reduced FCM and approximate kernel FCM are all good techniques for approximating FCM when dealing with big data. Wu [48] combined K-means with genetic algorithm based on differential evolution algorithm, which effectively improved the convergence speed. Lastly, some researchers redivided clustering skills into many types based on knowledge of big data. In Zhou & Zhang's article [49], they divided big data clustering techniques into three types: Distributed clustering, parallel clustering and high-dimensional clustering and the first two algorithms are collectively called multi-machine clustering.

3 Methodology

3.1 Principle Component Analysis

Before we apply clustering methods to our data, we should first preprocess our dataset by Principle Component Analysis (PCA). PCA is a mathematical technique to reduce the dimension of dataset. Generally speaking, it uses linear combination to reduce the dimesion of dataset with m observations $X = [X^{(1)}, X^{(2)}, X^{(3)}...X^{(m)}]$ from n to k

Algorithm

Suppose we have a dataset with n dimension $X = [X^{(1)}, X^{(2)}, X^{(3)}...X^{(m)}]$. First we need to scale data by Zero-centering:

$$X^{(i)} = X^{(i)} - \frac{1}{m}\sum_{i=1}^{m} X^{(i)}$$

Also, we define the covariance between data $X^{(1)}$ and $X^{(2)}$:

$$Cov(X^{(1)}, X^{(2)}) = \frac{\sum_{i=1}^{n}\left(X_i^{(1)} - \bar{X}^{(1)}\right)\left(X_i^{(2)} - \bar{X}^{(2)}\right)}{(n-1)}.$$

Then we define the covariance matrix of our data:

$$S = \begin{bmatrix} Cov(X^{(1)}, X^{(1)}) & Cov(X^{(1)}, X^{(2)}) & \cdots & Cov(X^{(1)}, X^{(m)}) \\ Cov(X^{(2)}, X^{(1)}) & Cov(X^{(2)}, X^{(2)}) & \cdots & Cov(X^{(2)}, X^{(m)}) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(X^{(m)}, X^{(1)}) & Cov(X^{(m)}, X^{(2)}) & \cdots & Cov(X^{(m)}, X^{(m)}) \end{bmatrix} = \frac{1}{m} X X^T$$

After getting this covariance matrix, we can solve its eigenvector through Singular Value Decomposition(SVD).

We know that $X_{m\times n} = U_{m\times m}\Sigma_{m\times n}V^T_{n\times n}$, where $U_{m\times m}$ and $V_{n\times n}$ are orthogonal matrix, $\Sigma_{m\times n}$ is diagonal matrix. Therefore, $X^T X = \left(U\Sigma V^T\right)^T U\Sigma V^T = V\Sigma^2 V^T$, and we can get the eigenvectors and eignvalues we want through this process. Normally, we would discard the remaining eigenvalues and their corresponding eigenvectors after we get a 85% contribution rate.

Contribution rate is defined as

$$\frac{\lambda_i}{\sum_{i=1}^{n}\lambda_i}$$

Since we have found k eigenvalues and their eigenvectors $W = [W^{(1)}, W^{(2)}, W^{(3)}...W^{(k)}]$, we can get a new dataset with K-dimension $D = [Z^{(1)}, Z^{(2)}, Z^{(3)}...Z^{(m)}]$ and $Z^{(i)} = W^T X^{(i)}$.

By using this algorithm, we can successfully reduce the dimension of dataset from n to k

3.2 K-means clustering

After preprocessing data, we can assign the ith observation which has no labels into the Kth cluster ($K = C(i)$) and each observation is assigned to one and one cluster only. The number of cluster is pre-defined and the center of a cluster should be the mean of observation in that cluster.

The number of cluster K is up to yourself, but here we introduce one efficient way to choose K: Using Gap statistic. First, we define dissimilarity. In K-means, squared Euclidean distance

$$d\left(x_i, x_{i'}\right) = d_{ii'} = \sum_{j=1}^{p}\left(x_{ij} - x_{i'j}\right)^2 = \left\|x_i - x_{i'}\right\|^2$$

is chosen as the dissimilarity measure. Then, we define loss function.

Total scatter T can be defined as:

$$T = \frac{1}{2}\sum_{i=1}^{N}\sum_{i'=1}^{N} d_{ii'} = \frac{1}{2}\sum_{k=1}^{K}\sum_{C(i)=k}\left(\sum_{C(i')=k} d_{ii'} + \sum_{C(i')\neq k} d_{ii'}\right)$$

This can be written as:

$$\frac{1}{2}\sum_{k=1}^{K}\sum_{C(i)=k}\sum_{C(i')=k} d_{ii'} + \frac{1}{2}\sum_{k=1}^{K}\sum_{C(i)=k}\sum_{C(i')\neq k} d_{ii'}$$

The first part is called within cluster scatter $W(C)$ and the second part is called between cluster scatter $B(C)$. Therefore, $T = W(C) + B(C)$. We know that for a given dataset, the total scatter is a constant. Thus, $W(C)$ can be written as $T - B(C)$. From this equation, we can see that minimize $W(C)$ is to maximize $B(C)$. It is fine that you choose either minimize $W(C)$ or maximize $B(C)$ as your goal, but Clustering itself means that the distance between the observations in a cluster should be small. Thus, We define W as our loss function:

$$W = \frac{1}{2}\sum_{k=1}^{K}\sum_{C(i)=k}\sum_{(i')=k} d\left(x_i, x_{i'}\right)$$

Now we can introduce Gap Statistic:

$$Gap_n(k) = E_n^*\left\{\log\left(W_k\right)\right\} - \log\left(W_k\right)$$

$E_n^*\left\{\log\left(W_k\right)\right\}$ is the expectation of a sample under a reference distribution with size n, normally this reference distribution can be chosen as uniform distribution. The maximum number of $Gap_n(k)$ is the best number of cluster K.

Before finally introducing algorithm, there is one transformation of loss function W need to be done. Loss function W can be written as:

$$W = \sum_{k=1}^{K} N_k \sum_{C(i)=k}\left\|x_i - \bar{x}_k\right\|^2$$

This $\bar{x}_k = (\bar{x}_{1k}, \ldots, \bar{x}_{pk})$ is the mean vector for the Kth cluster, and

$$N_k = \sum_{i=1}^{N} I(C(i) = k)$$

Our loss function can be explained as to minimize the dissimilarity from the observations in this cluster to the cluster centroid.

Algorithm

First, we need to randomly select K points as the center of K clusters.

$$m_k = (m_1, \ldots, m_k)$$

Second, we calculate the Euclidean distance between each observation and each centroid and we assign each observation to a cluster which contains the closest centroid to this observation.

$$C_i = \left\{ x : \|x - m_i\|^2 \leq \|x - m_j\|^2 \right\}$$

Here, C_i means the ith cluster, x are observations, m_i is the center of the ith cluster, $1 \leq j \leq k$ Third, we update the centroid for each cluster.

$$m_i = \frac{1}{N\left(C_i\right)} \sum_{x \in C_i} x$$

$N(C_i)$ is the number of observations in the ith cluster, x is one observation in the ith cluster. By doing so, we can get the new centroid in this cluster, then we reallocate each observation to one cluster by doing step2. We keep iterating step2 and step3 until the observations in each cluster do not change. The process for clustering is actually the process to minimize loss function.

3.3 Hierarchical Clustering

Except K-means, Hierarchical clustering is another approach to cluster data. At the lowest level of Hierarchical clustering there exist many clusters, each observation is one cluster, while at the top of Hierarchical clustering, there exist only one cluster which contains all observations. There are two types of Hierarchical clustering, one is agglomerative clustering and this is a bottom-up method, meaning that this approach start from bottom and merge two clusters with the smallest intergroup dissmilarity into one cluster. This process ends until there exist only one cluster. Another approach is divisive clustering and this is a top-down method, meaning that this approach start from the top and split one cluster into two cluster with the largest between-group dissimilarity. This process ends until each cluster contains only one observation. Normally we will use agglomerative clustering.

One way to intuitively express agglomerative clustering is to plot a dendrogram in which each node represent one cluster and the top node represents total data. One monotonicity property of this agglomerative clustering is that with the merging of clusters, the intergroup dissimilarity is increasing monotonously. Because of this property, the height of each node is proportional to the intergroup dissimilarity [50].

Before we introduce the algorithm of Agglomerative Clustering, we have to first introduce a measure of dissimilarity between two clusters. We know that at the bottom of agglomerative clustering, each observation is a cluster, we can just use the Euclidean distance to measure it. Euclidean distance is defined as:

$$\sqrt{(a_i - b_i)^2}$$

But for dissimilarity between clusters, we should use other approaches and there are three ways to measure.

First method is called Single linkage agglomerative clustering.

Suppose A and B are two clusters. Single linkage agglomerative clustering takes the distance of two closest observations that each from separate cluster as its value.

$$d_{Single}(A, B) = \min_{\substack{i \in A \\ i' \in B}} d_{ii'}$$

Second method is called Complete linkage agglomerative clustering.

Complete linkage agglomerative clustering takes the distance of two furthest observations that each from different cluster as its value.

$$d_{Complete}(A, B) = \max_{\substack{i \in A \\ i' \in B}} d_{ii'}$$

Third method is called Group Average clustering.

Group Average clustering takes the average distance between two clusters as its value,

$$d_{Average}(A, B) = \frac{1}{N(C_A)N(C_B)} \sum_{i \in A} \sum_{i' \in B} d_{ii'}$$

Where $N(C_A)$ and $N(C_B)$ are the number of observations in each cluster. Usually we will use this method.

Algorithm for agglomerative clustering

First, we calculate the Euclidean distance between each observation and combine the two closest observations as one new cluster.

Second, we use one of three methods mentioned above to calculate distance between observations that are outside the cluster and the cluster. We also calculate the distance between outside cluster observations too. We compare all of these results and merge the closest two observations (clusters) into a new cluster.

Iterating setp2 until all the data have been merged into one cluster.

The number of clusters and the method to calculate distance between clusters can be determined by yourself.

Another method is called ward method. The idea is to use ESS (Error of sum of squares) as a measurement to minimize the loss of information in each merging. ESS is defined as follows:

$$ESS(C) = \sum_{\mathbf{x} \in C} (\mathbf{x} - m_{\mathbf{x}})^T (\mathbf{x} - m_{\mathbf{x}})$$

, where C is one cluster, $m_{\mathbf{x}}$ is the mean of this cluster. This method merges two clusters which have the smallest ESS at one time until all points have been clustered into one cluster.

4 Results and discussion

All the techniques mentioned in the previous part will be implemented in this part. The process is to first load in data and then use PCA to reduce the dimension of data. After that, two clustering techniques: Hierarchical clustering and K-means clustering will be used to see how these countries are grouped. Meanwhile some explanations and conjectures based on the results will be made. R studio is used to generate and visualize results.

4.1 Data description

27 countries with a large number of infections have been filtered through website: Worldometer as the target for clustering.

See Table 1 for details of data definition and source

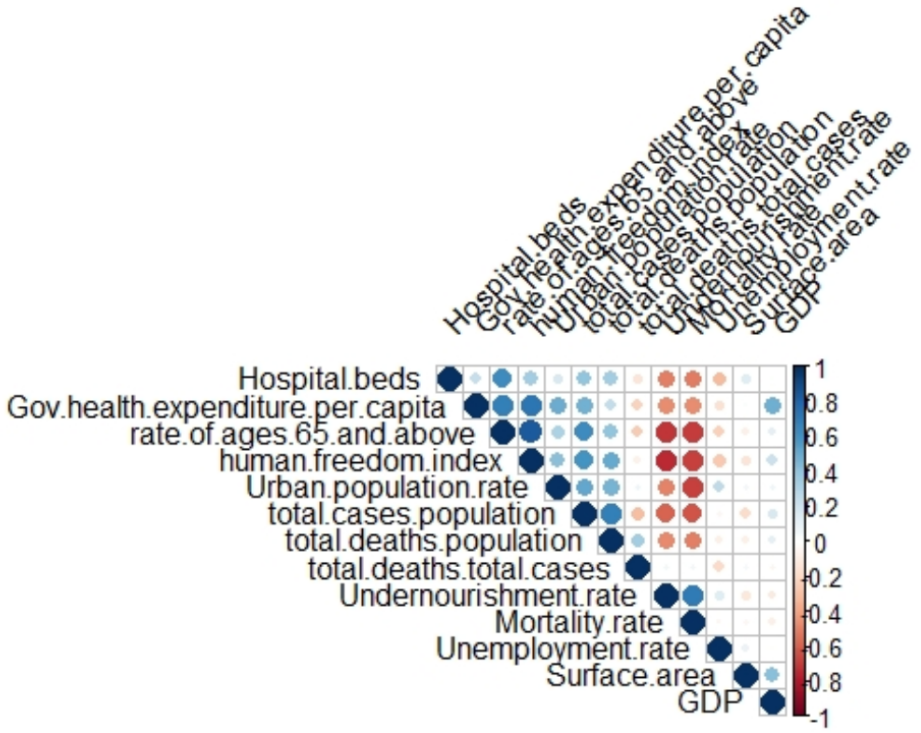
Table 1: Details of data

	Definition	Source	Unit
Unemployment rate	The share of the labor force that is without work but available for and seeking employment. Definitions of labor force and unemployment differ by country.	WorldBank	% of total labor force
Undernourishment rate	Population below minimum level of dietary energy consumption (also referred to as prevalence of undernourishment) shows the percentage of the population whose food intake is insufficient to meet dietary energy requirements continuously. Data showing as 5 may signify a prevalence of undernourishment below 5%.	WorldBank	% of population
Hospital beds	Hospital beds include inpatient beds available in public, private, general, and specialized hospitals and rehabilitation centers. In most cases beds for both acute and chronic care are included.	WorldBank	per 1000 people
Surface area	A country's total area, including areas under inland bodies of water and some coastal waterways.	WorldBank	sq. km
Urban population rate	Urban population refers to people living in urban areas as defined by national statistical offices.	WorldBank	% of total population
Mortality rate	Neonatal mortality rate is the number of neonates dying before reaching 28 days of age	WorldBank	per 1,000 live births
GDP	GDP at purchaser's prices is the sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products.	WorldBank	current US$
Gov health expenditure per capita	Public expenditure on health from domestic sources per capita	WorldBank	current US$
rate of ages 65 and above	Total population 65 years of age or older	WorldBank	Number of people
human freedom index	The Human Freedom Index presents the state of human freedom in the world based on a broad measure that encompasses personal, civil, and economic freedom	Cato Institute and the Fraser Institute	a scale of 0 to 10
total cases/population	number of total cases/population	worldometers.info	percentage
total deaths/total cases	number of total deaths/number of total cases	worldometers.info	percentage
total deaths/population	number of total deaths/population	worldometers.info	percentage

Table 2: Correlation table

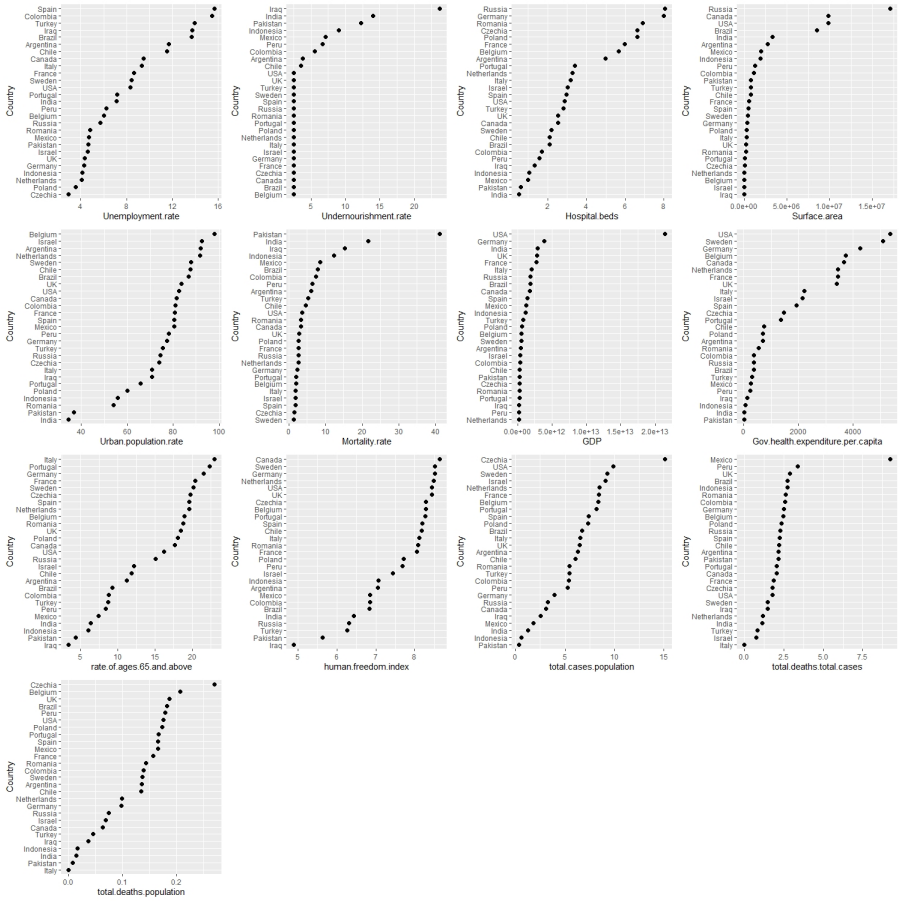
Unemployment rate	1	0.13	-0.32	0.08	0.25	-0.03	0	-0.16	-0.23	-0.27	-0.05	-0.19	-0.09
Undernourishment rate	0.13	1	-0.51	-0.13	-0.49	0.7	-0.12	-0.46	-0.71	-0.75	-0.58	0.05	-0.47
Hospital beds	-0.32	-0.51	1	0.13	0.16	-0.51	0	0.24	0.63	0.35	0.39	-0.13	0.35
Surface area	0.08	-0.13	0.13	1	0.05	-0.04	0.41	0.02	-0.08	-0.14	-0.19	0.05	-0.09
Urban population rate	0.25	-0.49	0.16	0.05	1	-0.68	0.06	0.49	0.31	0.42	0.52	0.07	0.46
Mortality rate	-0.03	0.7	-0.51	-0.04	-0.68	1	-0.08	-0.47	-0.7	-0.68	-0.63	0.05	-0.51
GDP	0	-0.12	0	0.41	0.06	-0.08	1	0.49	0.11	0.21	0.17	-0.05	0.11
Gov health expenditure per capita	-0.16	-0.46	0.24	0.02	0.49	-0.47	0.49	1	0.68	0.72	0.48	-0.22	0.24
rate of ages 65 and above	-0.23	-0.71	0.63	-0.08	0.31	-0.7	0.11	0.68	1	0.84	0.63	-0.25	0.39
human freedom index	-0.27	-0.75	0.35	-0.14	0.42	-0.68	0.21	0.72	0.84	1	0.6	-0.08	0.5
total cases/population	-0.05	-0.58	0.39	-0.19	0.52	-0.63	0.17	0.48	0.63	0.6	1	-0.3	0.68
total deaths/total cases	-0.19	0.05	-0.13	0.05	0.07	0.05	-0.05	-0.22	-0.25	-0.08	-0.3	1	0.34
total deaths/population	-0.09	-0.47	0.35	-0.09	0.46	-0.51	0.11	0.24	0.39	0.5	0.68	0.34	1

Figure 1: Correlation plot



From the correlation table and correlatioin plot, some basic descriptions of correlations can be drawn. Undernourishment rate and mortality rate are almost negative correlated to all other variables. While hospital beds, urban population rate are mostly positive correlated to other variables. GDP and surface area does not show much correlations with other variables. On the contrary, correlations between Human freedom index & rate of ages 65 and above and other variables are relatively high.

Figure 2: Rank plot



From the rank plot, we can see that there exist some outstanding countries in the data of some variables. For instance, Iraq is much higher than other countries in the undernourishment rate, Russia has a much bigger surface area. The mortality rate in Pakistan even reaches to 40%. GDP of USA is far ahead then other countries. Czechia has a high total cases/population rate and Mexico has a high rate of total deaths/total cases.

4.2 Data processing and results

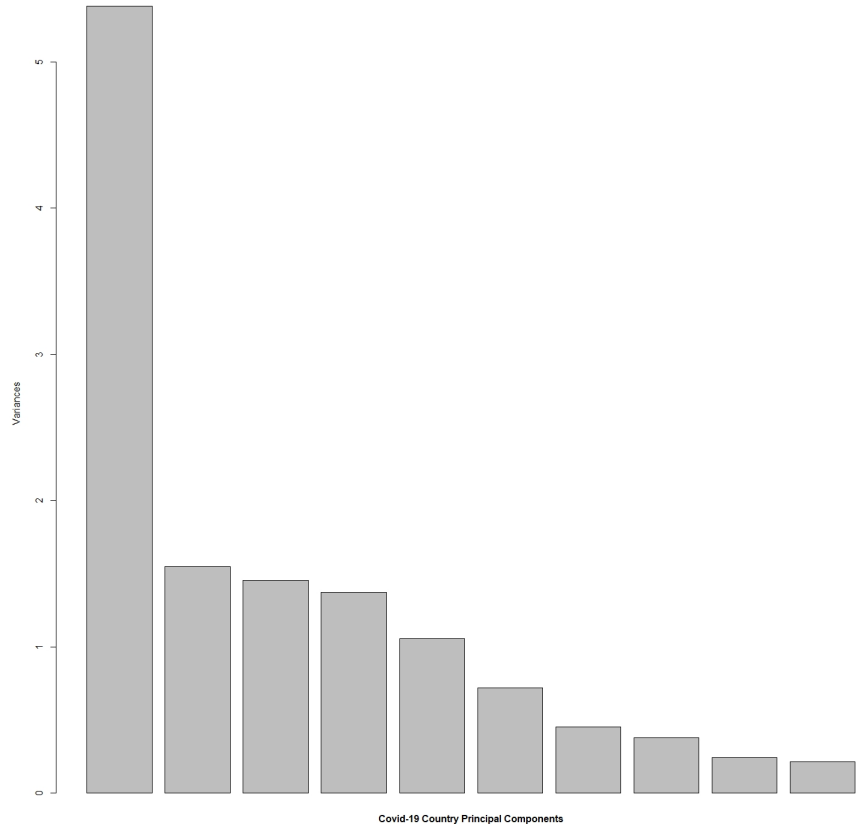
4.2.1 PCA

In sum, 13 principal components are obtained and shown in the following table.

Table 3: PCA proportion table

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13
Standard deviation	2.31961345	1.24452433	1.20728469	1.17118765	1.02867662	0.84789749	0.67376902	0.61593992	0.49619039	0.46325244	0.31977102	0.20341598	0.16261646
Proportion of Variance	0.41389	0.11914	0.11212	0.10551	0.0814	0.0553	0.03492	0.02918	0.01894	0.01651	0.00787	0.00318	0.00203
Cumulative Proportion	0.41389	0.53303	0.64515	0.75067	0.83206	0.88737	0.92229	0.95147	0.97041	0.98692	0.99478	0.99797	1

Figure 3: Variance plot



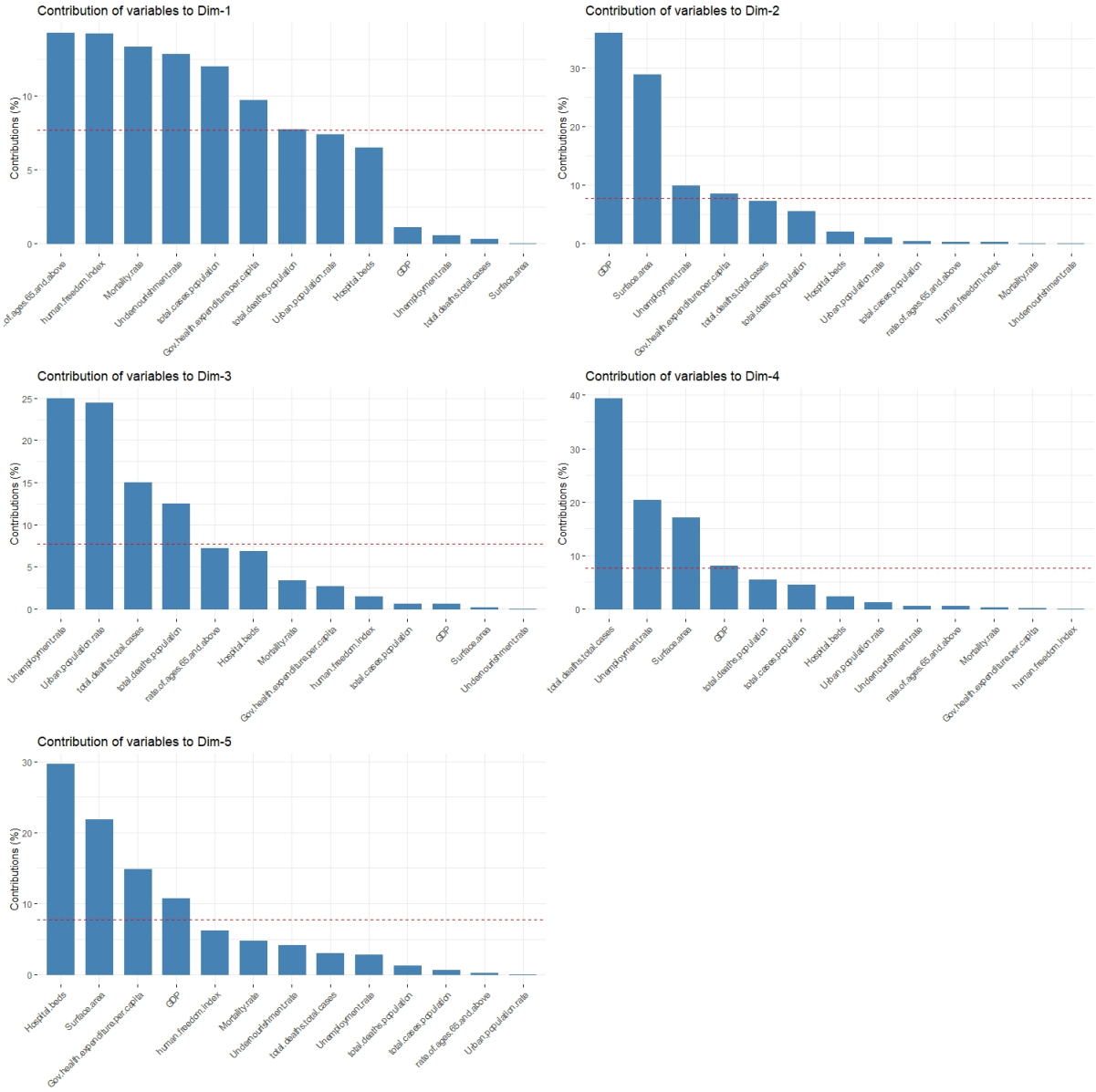
As illustrated in table 3, first five principal components explain 83% of overall variability and figure 3 indicated that most variance are contained in the first five components. Meanwhile, each component's contribution is bigger than 0.077 i.e.: $\frac{1}{13}$, meaning that the information contained in each component is not less then each variable before transformation. Thus, first five components are chosen for further analysis.

The coefficients of five components are listed in table 4. In PC1, undernourishment rate & mortality rate work against rate of ages 65 and above & human freedom index. If a country scores high in PC1, this means that this is a developed, free country that has a good medical system. PC2 can be explained as 'Geography index': The coefficients of surface area and GDP are relatively large and negative. Also, if the country has a large surface area it is more likely to have a large GDP. Thus, if a country scores low on PC2, this means that this country has a good geographical location. PC3 has negative and large coefficients of unemployment rate and urban population rate. In PC4, unemployment rate and total deaths/total cases are against each other. In PC5, the coefficient of surface area and hospitals beds are large and negative. Also, the contribution of each variable in each dim is plotted on figure 4.

Table 4: Coefficients table

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13
Unemployment rate	-0.072	-0.314	-0.499	0.451	-0.167	0.033	-0.336	0.469	-0.195	0.083	-0.019	0.007	-0.191
Undernourishment rate	-0.359	-0.01	0.014	0.079	0.204	0.196	0.47	0.341	-0.01	-0.609	0.091	-0.239	-0.104
Hospital beds	0.255	0.143	0.262	-0.154	-0.544	0.206	0.325	0.422	-0.053	0.305	0.2	-0.071	-0.243
Surface area	-0.006	-0.538	-0.045	-0.413	-0.467	-0.032	-0.045	-0.294	-0.271	-0.382	-0.032	-0.025	-0.1
Urban population rate	0.272	-0.103	-0.495	0.111	-0.005	-0.263	0.531	-0.214	-0.029	0.192	0.111	-0.41	0.208
Mortality rate	-0.365	-0.012	0.185	-0.055	0.218	0.179	-0.049	-0.137	-0.661	0.393	0.02	-0.368	-0.035
GDP	0.104	-0.599	0.077	-0.284	0.327	0.335	-0.058	0.208	0.401	0.246	0.022	-0.184	0.148
Gov health expenditure per capita	0.312	-0.291	0.165	0.036	0.386	-0.235	0.329	0.142	-0.382	0.033	-0.083	0.525	-0.163
rate of ages 65 and above	0.378	0.053	0.269	0.073	-0.049	-0.123	-0.169	0.297	-0.215	-0.23	-0.47	-0.357	0.442
human freedom index	0.378	0.047	0.121	-0.016	0.248	-0.243	-0.343	-0.014	-0.021	-0.206	0.531	-0.349	-0.395
total cases/population	0.347	0.065	-0.077	0.211	0.079	0.527	0.06	-0.333	0.035	-0.068	-0.439	-0.112	-0.464
total deaths/total cases	-0.057	0.269	-0.387	-0.627	0.173	-0.225	-0.043	0.264	0.002	0.066	-0.363	-0.077	-0.295
total deaths/population	0.278	0.236	-0.353	-0.234	0.112	0.496	-0.1	0.044	-0.306	-0.163	0.322	0.236	0.37

Figure 4: Contribution plot



Based on the score of each country on the first five components (table 5) and biplots on first four components (figure 5), some conclusions can be made. USA is relatively high on PC1 and low on

PC2, thus it is on the bottom of biplot 1. UK, Netherlands, Germany and France have a close value on both PC1 and PC2, therefore they are likely to be in the same cluster. Russia & Canada, who are close on dim1 with Romania and Poland, are separated on PC2. Pakistan, India and Italy are high on PC3 and USA, Russia and Mexico are low on PC4. Besides, countries like Poland, Romina and Germany who are European countries and are close on biplot 2 might be in the same cluster.

Table 5: Score table

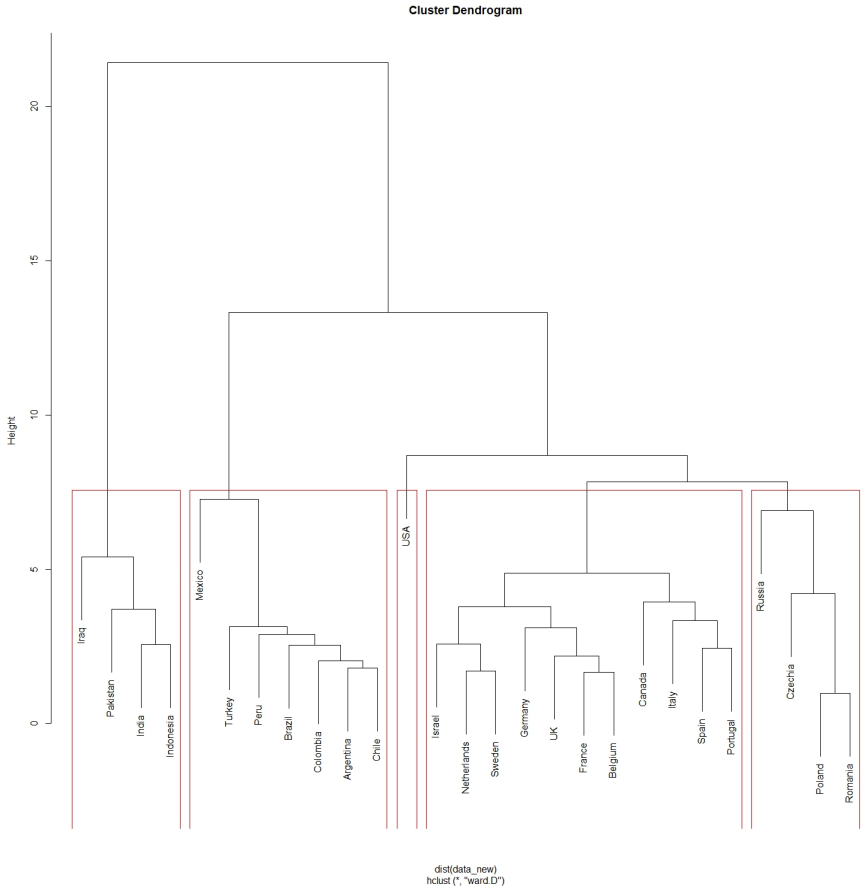
	PC1	PC2	PC3	PC4	PC5
USA	2.60027891	-4.3586009	0.16751309	-1.6812174	1.85538014
India	-4.4307588	-0.6379982	1.69834659	0.08111109	0.4991862
Brazil	-0.4859058	-0.9550347	-2.1774076	-0.2637271	-0.9536302
Russian	-0.2793895	-1.4640579	0.62090656	-2.0863399	-3.6240125
UK	1.65797504	0.44878819	-0.0108926	-0.5375453	1.25068083
France	2.01632524	0.07538703	0.32240618	0.33673426	-0.0108001
Spain	1.20339666	-0.1403181	-1.2273959	1.16711901	0.01554523
Italy	0.74667756	-0.5988239	1.49510587	1.76703243	-0.1997849
Turkey	-1.3716309	-0.5198176	-0.6744533	1.63916793	-1.017053
Germany	1.97788812	0.13373047	1.65913688	-0.6877271	-0.0486901
Colombia	-1.2166263	-0.1083057	-1.9441181	0.93913222	-0.2128537
Argentina	-0.0246603	0.01033795	-1.2177316	0.43940061	-1.0528761
Mexico	-1.8722768	1.61147981	-2.3587298	-3.3577385	1.16209728
Poland	0.93100549	1.54896644	1.06993154	-0.6597213	-0.7189938
Indonesia	-3.1079613	0.23934311	0.90349729	-0.5967693	0.39268494
Peru	-0.77238	0.96240302	-1.0509383	-0.5948229	0.52973632
Czech	2.91135492	1.88545657	0.36736236	-0.1442462	-0.1511686
The Netherlands	1.85252039	0.33085162	0.6693532	0.77176683	0.62028933
Canada	0.72003841	-1.71654	0.21956466	-0.3238559	-0.381567
Chile	0.11400835	0.21330681	-1.2417652	0.76289755	0.09352616
Portugal	1.29567412	0.98893435	0.44514036	0.33545841	0.10234382
Romania	0.62835105	1.51598389	1.32720806	-0.6633	-0.8484928
Israel	0.73328278	0.19351115	0.20647292	1.04989056	0.08845278
Belgium	2.5273926	0.75869241	-0.3798227	-0.0249396	0.30879331
Iraq	-4.8633981	-0.4973346	-0.8575694	1.60585442	0.18448304
Sweden	2.20921031	-0.2328379	0.00505996	1.09311274	1.13214852
Pakistan	-5.7003921	0.31249662	1.96381884	-0.3667277	0.98457484

Figure 5: PCA biplot

4.2.2 Hierarchical clustering

Four types of linkage methods have been applied and ward.D linkage seems to give the best and most reasonable result.

Figure 6: Hierarchical clustering using ward linkage



This tree is cut into 5 clusters. Figure 7 shows the detail of each cluster. Table 6 lists the mean of each cluster on different variables (data are scaled). USA is grouped into one cluster separately, which might because USA is way much better than other countries in many aspects. For example, low unemployment rate, high GDP and government health expenditure per capita, large surface area. However, it is also high on total cases/population rate. Cluster 2 mainly contains developing countries in Asia because it has high undernourishment rate, low GDP and high mortality rate. Cluster 3 mainly consist of countries in South America and cluster 4 is composed of countries in Europe. Other countries are clustered in cluster 5 and are mainly developed countries.

Figure 7: Results of hierarchical clustering

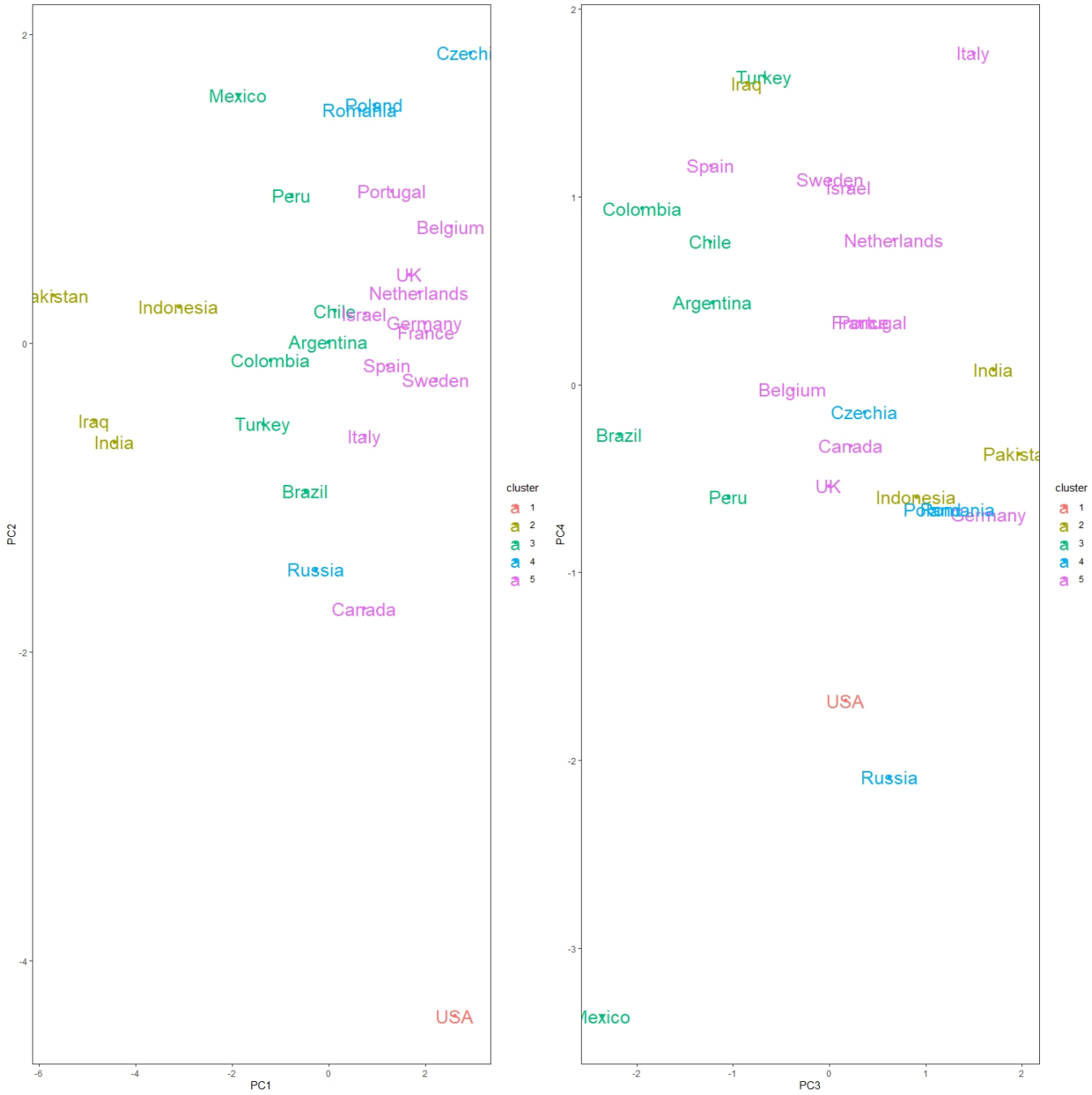
USA	India	Brazil	Russia	UK	France
1	2	3	4	5	5
Spain	Italy	Turkey	Germany	Colombia	Argentina
5	5	3	5	3	3
Mexico	Poland	Indonesia	Peru	Czechia	Netherlands
3	4	2	3	4	5
Canada	Chile	Portugal	Romania	Israel	Belgium
5	3	5	4	5	5
Iraq	Sweden	Pakistan			
2	5	2			

Table 6: Mean of each cluster

	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
Unemployment.rate	0.095	-0.135	0.783	-0.93	-0.12
Undernourishment.rate	-0.477	2.025	-0.066	-0.477	-0.477
Hospital.beds	-0.262	-1.135	-0.501	1.571	0.184
Surface.area	1.814	-0.203	0.028	0.506	-0.293
Urban.population.rate	0.45	-1.645	0.491	-0.618	0.47
Mortality.rate	-0.334	1.922	0.014	-0.468	-0.508
GDP	4.841	-0.184	-0.28	-0.281	-0.093
Gov.health.expenditure.per.capita	2.136	-0.974	-0.757	-0.554	0.843
rate.of.ages.65.and.above	0.3	-1.52	-0.812	0.583	0.83
human.freedom.index	0.902	-1.531	-0.438	0.063	0.731
total.cases.population	1.177	-1.468	-0.218	0.544	0.368
total.deaths.total.cases	-0.297	-0.245	0.663	0.006	-0.308
total.deaths.population	0.801	-1.446	0.286	0.644	0.037

The projection of clusters on different principal components is illustrated on figure 8. On the left of this figure, cluster 2, 3 and 5 are separated vividly on PC1 and USA is still at the bottom. However, on the right of this figure, the separation of clusters is not that obvious. Cluster 3 is partly separated from cluster 5, while cluster 2, 4 and 5 are to some extent mixed on the right of PC3.

Figure 8: Projection of clusters on PCA using hierarchical clustering



4.2.3 K-means clustering

Before getting results, the first thing is to set the number of clusters. Based on 23 indices (figure 9), the optimal number should be 3, however, is not realistic to group 27 countries into just 2 or 3 clusters. Therefore, 5 seems to be a good choice.

Figure 9: Optimal number of clusters of K-means

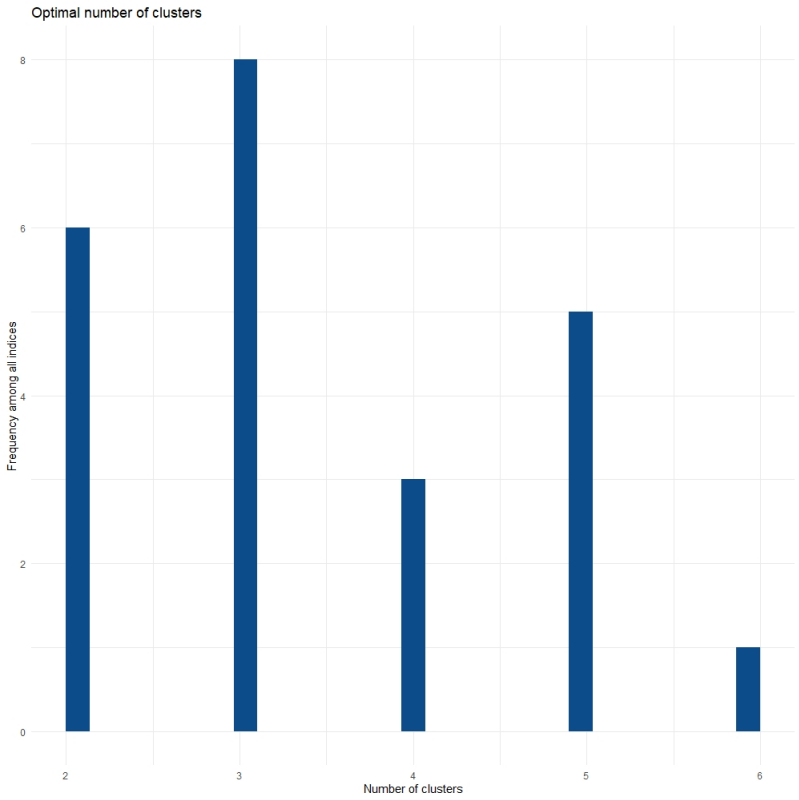


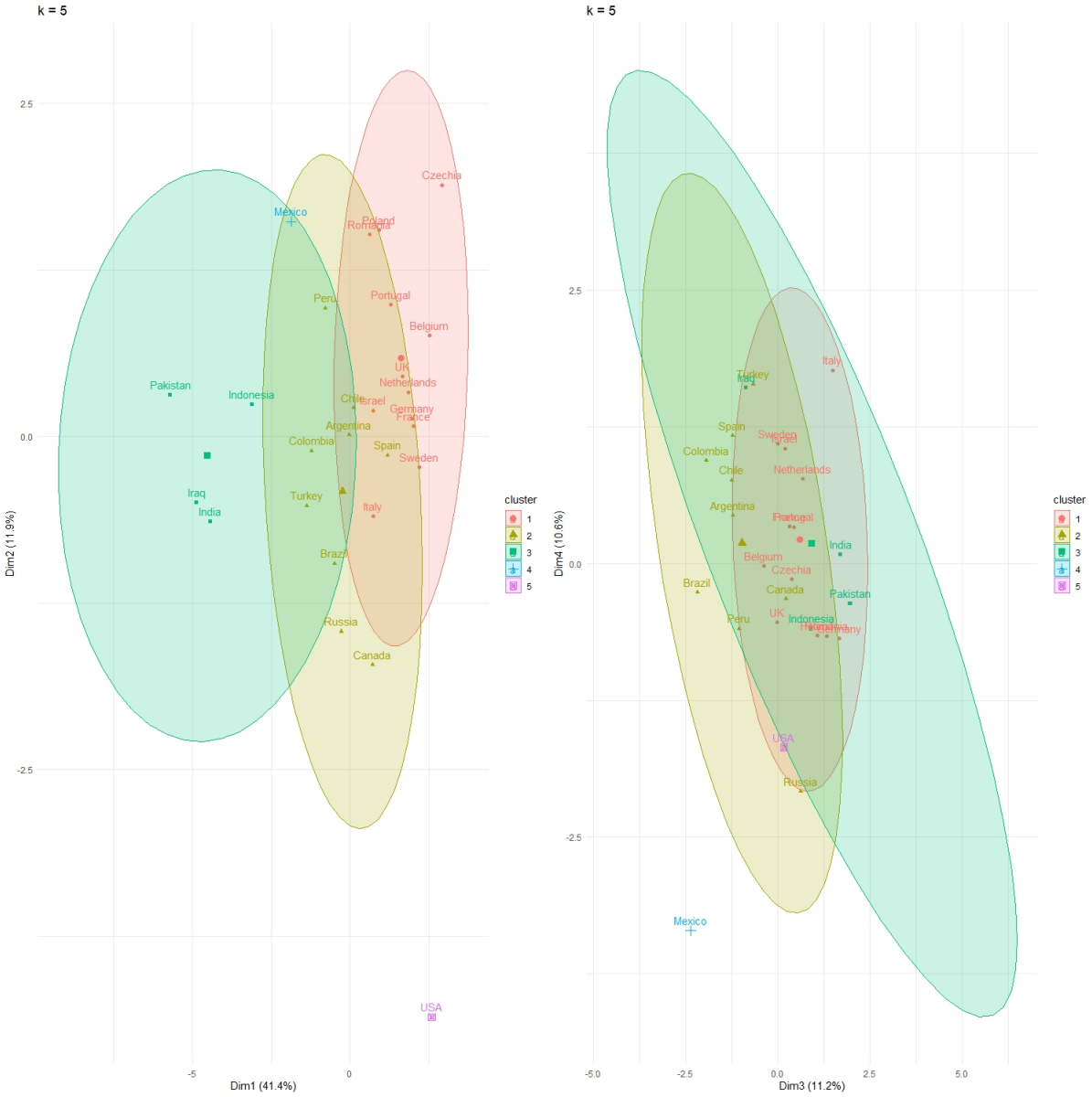
Table 7 listed the countries in each cluster. Compare with hierarchical clustering, K-means clustering put more countries together thus making the results more unbalanced. 12 countries are grouped in cluster 1 and 9 countries are contained in cluster 2. Besides, different from hierarchical clustering, USA and Mexico are each put in a separate cluster. One thing in common is that India, Indonesia, Iraq and Pakistan are grouped together in one cluster both by K-means clustering and hierarchical clustering. The ratio of between sums of squares/ total sum of squares is 61.1%, indicating a good clustering result.

Table 7: results of K-means clustering

Country	Cluster
USA	1
Mexico	2
Brazil	3
Russia	3
Spain	3
Turkey	3
Colombia	3
Argentina	3
Peru	3
Canada	3
Chile	3
UK	4
France	4
Italy	4
Germany	4
Poland	4
Czechia	4
Netherlands	4
Portugal	4
Romania	4
Israel	4
Belgium	4
Sweden	4
India	5
Indonesia	5
Iraq	5
Pakistan	5

Left of figure 10 shows good separation between cluster 1 and cluster 3 on the dimension of PC1. However, clusters are mixed on the right of figure 10. Little separation can be found on both PC3 and PC4.

Figure 10: Projection of clusters on PCA using K-means clustering



Lastly, a table of cluster centers based on K-means clustering can be found on table 8.

Table 8: Centers of each cluster

	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
Unemployment.rate	0.095	-0.817	0.899	-0.569	-0.135
Undernourishment.rate	-0.477	0.462	-0.262	-0.477	2.025
Hospital.beds	-0.262	-1.087	-0.115	0.577	-1.135
Surface.area	1.814	-0.092	0.583	-0.513	-0.203
Urban.population.rate	0.45	0.322	0.421	0.169	-1.645
Mortality.rate	-0.334	0.249	-0.173	-0.504	1.922
GDP	4.841	-0.148	-0.219	-0.165	-0.184
Gov.health.expenditure.per.capita	2.136	-0.851	-0.441	0.548	-0.974
rate.of.ages.65.and.above	0.3	-1.132	-0.338	0.829	-1.52
human.freedom.index	0.902	-0.69	-0.205	0.647	-1.531
total.cases.population	1.177	-1.282	-0.17	0.626	-1.468
total.deaths.total.cases	-0.297	4.418	0.01	-0.269	-0.245
total.deaths.population	0.801	0.645	0.06	0.317	-1.446

4.3 Discussion

Our research findings focus on two aspects: First is to see how these aggressively affected countries are grouped and what are the features in different groups. Second is to observe whether there exists some unusual countries and to explain why they are not grouped together.

4.3.1 Clusters based on hierarchical clustering

The result of hierarchical clustering suggests that 27 countries are clustered mainly based on geographical location, i.e., based on which continent the country is on. This result is reasonable because countries closed or border to each other are more likely to infect each other. Also, countries in the same continent are more likely to have the same thinking mode and use the same methods to deal with Covid-19. For example, countries in cluster 2, which are in Asia, have a smaller total cases/ population ratio compare with other countries. This mainly because the freedom index of Asian countries is not high and people will take government's advice to stay at home. While countries like UK, Germany, they have high freedom index and people might ignore the government's suggestions which recommends them to stay at home. They tend to chase freedom and might still have many social activities. Therefore, the percentage of people infected by Covid-19 is high in these countries.

Features of each cluster are listed as follows: Cluster 1 is USA. Cluster 2 is relatively high on undernourishment rate, mortality rate and low on number of hospital beds, urban population rate and total cases/population ratio. Cluster 3 is high on unemployment rate, urban population rate and total deaths/ total cases ratio. Cluster 4 is low on unemployment rate and high on number of hospital beds. Cluster 5 is low on surface area, mortality rate and total deaths/total cases ratio while high on rate of ages 65 and above.

4.3.2 Clustering based on K-means clustering

The results of K-means are a bit different from the result of hierarchical clustering, but still make sense. Most of the countries are also clustered based on geographical location. What is different is that K-means combines Czechia, Poland and Romania into cluster 5 obtained from hierarchical clustering. This is logical because these three countries are all European countries and most countries in cluster 5 also locate in Europe. Besides, K-means moves Russia, Spain and Canda into cluster 3 obtained from hierarchical clustering. This is reasonable to some extent because they are very close on PC1 with other countries in cluster 3.

Features of each cluster are listed as follows: Cluster 1 is low on undernourishment rate, surface area, mortality rate and high on number of hospital beds, human freedom index, ratio of total cases/population. Cluster 2 is Mexico. Cluster 3 is high on unemployment rate, surface area and low on GDP. Cluster 4 is high on undernourishment rate, mortality rate, government expenditure per capita low on number of hospital beds and ratio of total cases/population. Cluster 5 is USA.

4.3.3 Unusual countries and their features

Even though USA is a developed country with good medical system and locates in North America, it is listed alone in the results of both of these two clustering. This is logical because USA has a high freedom index therefore people might still gather together even though the situation of Covid-19 is bad, which finally resulted in a high ratio of total cases/ population. However, USA is very high on GDP, government health expenditure per capita, suggesting a good medical level and economy development. Thus, the total deaths/total cases ratio is low. These extreme data led USA to be put in a group alone.

Mexico is listed alone in the results of K-means clustering. This might result from a very high ratio of total death/total cases compared with other countries and a relative low number of hospital beds, which reflects a relatively low medical level.

5 Conclusion

5.1 Summary

In this project, we aim to group 27 countries who are aggressively affected by Covid-19 based on data in areas ranging from economy, medical level, geographical features to Covid-19 situation. Three techniques are performed to accomplish this goal, namely PCA, hierarchical clustering and K-means clustering. Both hierarchical clustering and K-means clustering grouped countries into 5 clusters and they are mainly clustered based on geographic location. However, cluster structures of hierarchical clustering are more balanced than K-means clustering. One thing needs to be noted is that both two clustering techniques put USA in a separate cluster and Mexico is put in a separate cluster on the results of K-means. Projections of clusters grouped by hierarchical clustering and K-means clustering separate well on the biplots of different principal components. This indicate that PCA and these two clustering techniques suits this dataset well.

5.2 Contributions

The contribution of this project is to apply unsupervised learning techniques to the study of Covid-19. Based on the fact that few researchers have used these techniques in this area, therefore, our results might lay a foundation for future research. Also, some discussions have been written, which might provide policy makers, economy experts, officials of health ministry some insights.

5.3 Limitations and recommendations for future research

The first limitation of this project is that the methodologies used are simple. Only PCA and two basic clustering techniques are used. Also, the number of clusters of K-means is decided by myself, making it very subjective. Future researchers can apply more clustering techniques and try to find a good result. Besides, they can apply a more objective way to decide the number of clusters.

The second limitation is that the volume of this dataset is not very large, containing only 13 variables and 27 countries. Future researchers can obtain more types of data not just limited in these four areas mentioned above. Also, in each area, other types of data can be found and used. For example, in the area of Covid-19 situation, researchers can use data like basic reproduction number of infection (R0) of each country. More data may generate a more reliable result and a more reasonable reason might be discovered from it.

References

[1] "Who coronavirus (covid-19) dashboard." [Online]. Available: https://covid19.who.int/.

[2] "Global impact of the covid-19 pandemic: 1 year on." [Online]. Available: https://www.medicalnewstoday.com/articles/global-impact-of-the-covid-19-pandemic-1-year-on#Reaching-for-new-coping-strategies.

[3] "Coronavirus: How the pandemic has changed the world economy." [Online]. Available: https://www.bbc.com/news/business-51706225.

[4] J. Schmidhuber, "Covid-19: from a global health crisis to a global food crisis?," *FAO Food Outlook*, no. June, pp. 63–71, 2020.

[5] V. Zarikas, S. G. Poulopoulos, Z. Gareiou, and E. Zervas, "Clustering analysis of countries using the covid-19 cases dataset," *Data in brief*, vol. 31, p. 105787, 2020.

[6] C. Fang and S. Ding, "China's city division based on principal component analysis and cluster analysis under covid-19 (in chinese)," *Statistics And Management*, vol. v.35;No.272, no. 03, pp. 50–54, 2020.

[7] M. Azarafza, M. Azarafza, and H. Akgun, "Clustering method for spread pattern analysis of corona-virus (covid-19) infection in iran," *medRxiv*, 2020.

[8] M. R. Mahmoudi, D. Baleanu, Z. Mansor, B. A. Tuan, and K.-H. Pho, "Fuzzy clustering method to compare the spread rate of covid-19 in the high risks countries," *Chaos, Solitons & Fractals*, vol. 140, p. 110230, 2020.

[9] Y. Wang, Y. Liu, J. Struthers, and M. Lian, "Spatiotemporal characteristics of the covid-19 epidemic in the united states," *Clinical infectious diseases*, vol. 72, no. 4, pp. 643–651, 2021.

[10] Y. Zhang and J. Li, "Prediction and analysis of propagation of novel coronavirus pneumonia epidemic based on sir model," *Journal of AnHui University of technology*, pp. 94–101, 1 2020.

[11] R. Zhu, S. Tang, T. Liu, Y. Guo, S. Dong, Y. Cheng, and T. Yang, "Covid ¯ 19 epidemic prediction based on improved sir model and the impact of prevention and control on epidemic development," *Journal of Liaoning Normal University(Natural Science Edition)*, pp. 33–38, 5 2020.

[12] D. Lin, X. Jin, W. Liu, Z. Huang, and X. Huang, "Prediction and analysis of new coronavirus epidemic situation," *Journal of Heilongjiang University of Technology*, vol. v.20, no. 09, pp. 120–125, 2020.

[13] B. G. Leichtweis, L. de Faria Silva, F. L. da Silva, and L. A. Peternelli, "How the global health security index and environment factor influence the spread of covid-19: A country level analysis," *One Health*, vol. 12, p. 100235, 2021.

[14] B. Everitt, *Cluster Analysis. [electronic resource].* Wiley series in probability and statistics, Wiley, 2011.

[15] A. K. Jain and R. C. Dubes, *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.

[16] A. Saxena, M. Prasad, A. Gupta, N. Bharill, O. P. Patel, A. Tiwari, M. J. Er, W. Ding, and C.-T. Lin, "A review of clustering techniques and developments," *Neurocomputing*, vol. 267, pp. 664–681, 2017.

[17] A. K. Mann and N. Kaur, "Review paper on clustering techniques," *Global Journal of Computer Science and Technology*, 2013.

[18] J. Zhang and D. Zhang, "survey of classical clustering methods (in chinese)," *Fujian computer*, pp. 84–85, 7 2017.

[19] R. Zhang, Y. Chen, M. Zhang, and K. Meng, "Overviewing of visual analysis approaches forclustering high-dimensional data," *Journal of Graphics*, pp. 44–56, 2 2020.

[20] D. Xu and Y. Tian, "A comprehensive survey of clustering algorithms," *Annals of Data Science*, vol. 2, no. 2, pp. 165–193, 2015.

[21] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *science*, vol. 315, no. 5814, pp. 972–976, 2007.

[22] T. S. Madhulatha, "An overview on clustering methods," *arXiv preprint arXiv:1205.1117*, 2012.

[23] R. Sharan and R. Shamir, "Click: a clustering algorithm with applications to gene expression analysis," in *Proc Int Conf Intell Syst Mol Biol*, vol. 8, p. 16, 2000.

[24] G. H. Ball and D. J. Hall, "Isodata, a novel method of data analysis and pattern classification," tech. rep., Stanford research inst Menlo Park CA, 1965.

[25] A. Banerjee, S. Merugu, I. S. Dhillon, J. Ghosh, and J. Lafferty, "Clustering with bregman divergences.," *Journal of machine learning research*, vol. 6, no. 10, 2005.

[26] J. Jin, "Review of clustering method," *Computer Science*, pp. 288–293, 12 2014.

[27] R. Herwig, A. J. Poustka, C. Müller, C. Bull, H. Lehrach, and J. O'Brien, "Large-scale clustering of cdna-fingerprinting data," *Genome research*, vol. 9, no. 11, pp. 1093–1105, 1999.

[28] S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church, "Systematic determination of genetic network architecture," *Nature genetics*, vol. 22, no. 3, pp. 281–285, 1999.

[29] J. Vijaywargiya, "Review: Unsupervised clustering, its approaches and applications," tech. rep., 12 2018.

[30] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, pp. 281–297, Oakland, CA, USA, 1967.

[31] P. S. Bradley and U. M. Fayyad, "Refining initial points for k-means clustering.," in *ICML*, vol. 98, pp. 91–99, Citeseer, 1998.

[32] D. Pelleg, A. W. Moore, *et al.*, "X-means: Extending k-means with efficient estimation of the number of clusters.," in *Icml*, vol. 1, pp. 727–734, 2000.

[33] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data mining and knowledge discovery*, vol. 2, no. 3, pp. 283–304, 1998.

[34] Y. Sun, Q. Zhu, and Z. Chen, "An iterative initial-points refinement algorithm for categorical data clustering," *Pattern Recognition Letters*, vol. 23, no. 7, pp. 875–884, 2002.

[35] C. Ding and X. He, "K-nearest-neighbor consistency in data clustering: incorporating local information into global optimization," in *Proceedings of the 2004 ACM symposium on Applied computing*, pp. 584–589, 2004.

[36] S. Yang, Y. Li, X. Hu, and R. Pan, "Optimization study on k value of k—means algorithm," *System Engineering Theory and Practice*, pp. 97–101, 2 2006.

[37] A. Likas, N. Vlassis, and J. J. Verbeek, "The global k-means clustering algorithm," *Pattern recognition*, vol. 36, no. 2, pp. 451–461, 2003.

[38] L. Gu, "A novel locality sensitive k-means clustering algorithm based on subtractive clustering," in *2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, pp. 836–839, IEEE, 2016.

[39] H. Shi and M. Xu, "A data classification method using genetic algorithm and k-means algorithm with optimizing initial cluster center," in *2018 IEEE International Conference on Computer and Communication Engineering Technology (CCET)*, pp. 224–228, IEEE, 2018.

[40] T. Zhang, R. Ramakrishnan, and M. Livny, "Birch: an efficient data clustering method for very large databases," *ACM sigmod record*, vol. 25, no. 2, pp. 103–114, 1996.

[41] S. Guha, R. Rastogi, and K. Shim, "Cure: An efficient clustering algorithm for large databases," *ACM Sigmod record*, vol. 27, no. 2, pp. 73–84, 1998.

[42] S. Guha, R. Rastogi, and K. Shim, "Rock: A robust clustering algorithm for categorical attributes," *Information systems*, vol. 25, no. 5, pp. 345–366, 2000.

[43] G. Karypis, E.-H. Han, and V. Kumar, "Chameleon: Hierarchical clustering using dynamic modeling," *Computer*, vol. 32, no. 8, pp. 68–75, 1999.

[44] R. Gelbard, O. Goldman, and I. Spiegler, "Investigating diversity of clustering methods: An empirical comparison," *Data & Knowledge Engineering*, vol. 63, no. 1, pp. 155–166, 2007.

[45] M. Jing, "A summary of dimension reduction for high dimensional data," *Joumal of Xi' an UniverSity(Natural Science Edition)*, pp. 48–52, 12 2014.

[46] T. C. Havens, J. C. Bezdek, C. Leckie, L. O. Hall, and M. Palaniswami, "Fuzzy c-means algorithms for very large data," *IEEE Transactions on Fuzzy Systems*, vol. 20, no. 6, pp. 1130–1146, 2012.

[47] W. Zhao, H. Ma, and Q. He, "Parallel k-means clustering based on mapreduce," in *IEEE international conference on cloud computing*, pp. 674–679, Springer, 2009.

[48] H. Wu, "Improvement of unsupervised k means clustering algorithm under background of big data mining," *Modern Electronics Technique*, pp. 118–121, 10 2020.

[49] Y. Zhang and Y. Zhou, "Review of clustering algorithms," *CODEN JYIID*, pp. 1869–1882, 7 2019.

[50] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer Science and Business Media, 2009.