Chapter 14

Clustering Analysis of European Health Status Pre-Covid-19

Ümmü Şahin Şener[1]

Ersin Şener[2]

Abstract

The influence of genetic factors as well as the individual's living environment and consumption habits are very important for a healthy and long life. The consumption of unhealthy products (alcohol, cigarettes, etc.) and the occurrence of diseases such as obesity and diabetes as a result of an unhealthy diet are inevitable. The aim of this study is to perform a cluster analysis by individual health criteria to observe the level of preparedness of Europe for a pandemic using data from the period just before the Covid-19 pandemic, which was announced worldwide in March 2020. The cluster analysis is conducted using data on the health status of people living in Europe prior to the Covid-19 pandemic contained in the Global Health Report published by the World Health Organization (WHO) in November 2021. In assessing European health status, countries are analyzed according to four categories and clustered using the k-means method: 1) demographic characteristics; 2) alcohol and tobacco prevalence per capita; 3) the likelihood of dying from cardiovascular disease (CVD), cancer, diabetes and chronic respiratory disease (CRD), and 4) the prevalence of diabetes, tuberculosis, systolic blood pressure (SBP) and diastolic blood pressure (DBP), and obesity. Data from 38 countries with no missing observations are analyzed using data from the health report. The countries are divided into 2 clusters using the cluster dendrogram, the Calinski-Harabasz index and the elbow method. There are a total of 23 countries in cluster 1 and 15 countries in cluster 2. The Ward method and Euclidean distance are used for clustering. The k-means method is used to calculate the confusing matrix by bootstrap for 2 clusters. An accurate prediction is achieved with 94.04% success for cluster 1 and 91.7% success for cluster 2.

1 Asst. Prof., Department of Mathematics, Faculty of Art & Science, Kirklareli University, ummusahin@klu.edu.tr, 0000-0001-9055-8734

2 Corresponding author: Asst. Prof., Department of Mathematics, Faculty of Art & Science, Kirklareli University, ersinsener@klu.edu.tr, 0000-0002-5934-3652

https://doi.org/10.58830/ozgur.pub518.c2136

1. Introduction

Pandemics are inevitable in today's globalized world, and the Covid-19 outbreak we have experienced in the last 5 years is the biggest example of this. Personal measures have certainly had an impact in slowing the pandemic. But the main thing is that we are prepared for the global pandemic with both our individual chronic diseases (this shows our body resistance) and our use of unhealthy things (smoking, alcohol, etc.). The immune system of individuals with chronic diseases is weaker than that of healthy individuals(Nicholson, 2016). The immune system is also negatively affected by the use of things that are harmful to public health (smoking, alcohol, etc.). Countries' pre-Covid-19 health status will inevitably be a very effective indicator of their preparedness for the Covid-19 pandemic period. The indicator of readiness for a pandemic will not only consist of determining the measures to be taken during the epidemic period. The unfavorable health status of individuals before the pandemic is expected to trigger their resistance to the virus in a possible virus outbreak, that is, the mortality rate. With this perspective, a cluster analysis was conducted using data on the individual health status of Europe before Covid-19.

Cluster analysis is widely used in many different fields to discover natural groups in the data set and to act according to these groups. Clustering is used in many disciplines, e.g. in economics (Ada et al., 2024), healthcare (Rizvi et al., 2021), marketing (Saunders, 1980; Wedel & Kamakura, 2000), bioinformatics (Sudre et al., 2021) and image processing (Ren et al., 2024). Clustering plays an important role in a variety of areas, from decision support systems to customer segmentation (Tressa et al., 2024). Focusing the literature review on cluster analysis applications on the health status of countries with pre-Covid-19 data, some preliminary examples are found.

Rizvi et al. conducted a study on clustering of countries according to disease prevalence, health systems and environmental indicators for Covid-19 cases. The authors are used the k-means algorithm and utilized the elbow method to determine the number of clusters. In data sets consisting of 79 countries and 18 variables, they expressed countries in 4 clusters (Rizvi et al., 2021). Sudre et al. study, clusters of symptoms in Covid-19; the authors by using the Mc2PCA clustering algorithm for the time series of symptoms observed in Covid-19 cases, 6 different clusters were formed (Sudre et al., 2021). Kiaghadi et al. used cluster analysis to assess Covid-19 risk, vulnerability and prevalence of infection. The authors clustered individuals between the ages of 45-65 in Houston, TX, which they determined as the location, to be infected with Covid-19 considering environmental factors (Kiaghadi et

al., 2020). Several studies have been conducted on the observation of the relationship between some socio-demographic data of the countries and the Covid-19 data (Carrillo-Larco & Castillo-Cara, 2020; Farseev et al., 2020; Imtyaz et al., 2020; Porcheddu et al., 2020).

To observe the health status of Europe before Covid-19, the WHO shared health data as an indicator of the health status of countries in the health report (*World Health Statistics*, 2021) published in November 2021. The health report data published by WHO and the dataset shared by our world in data team (Mathieu et al., 2020) were combined.

In this study, it is aimed to cluster the health status of individuals living in European countries in the light of data on the health status of countries before the Covid-19 pandemic. The health status in our clustering study on the pre-Covid-19 health status of European countries was addressed through the data we obtained under 4 sub-categories. These sub-categories are 1- demographic characteristics; 2- alcohol per capita and tobacco prevalence; 3- probability of dying from any cardiovascular disease (CVD), cancer, diabetes, chronic respiratory disease (CRD), and 4- the prevalence of diabetes, tuberculosis, systolic blood pressure (SBP) and diastolic blood pressure (DBP) and obesity.

In determining Europe's level of preparedness for the Covid-19 pandemic through our data in 4 categories, firstly, the h-clustering method is used to observe the suitability of the data for clustering and a cluster dendrograms is obtained. Calinski-Harabasz index and elbow method is used to determine the number of clusters and silhouette scores are calculated. After the number of clusters is determined, clusters are formed by k-means method (Ward and Euclidean distances). Cluster prediction performance is calculated with bootstrap. Finally, the distributions of the clusters are visualized with Principal Component Analysis (PCA).

2. Material and Methods

In this section, the methods used to classify Europe's readiness for Covid-19 are mentioned.

2.1.Data Set and Variables

These data consist of measures of European Health Status Pre-Covid-19 for **4** categories in **38** European countries (Mathieu et al., 2020; *World Health Statistics*, 2021).The dataset used in the analysis is a compiled dataset. The websites from which the data is obtained, and the definitions of the variables are given in Table 1.

Table 1 Variable and Sources

Variables	Description of variables	Type	References
location	Geographical location	Primary data	(Mathieu et al., 2020)
median_age	Median age of the population,	Primary data	(Mathieu et al., 2020)
population_ density	Number of inhabitants per square kilometer	Primary data	(Mathieu et al., 2020)
life_expectancy	United Nations Development Programme (UNDP)	Primary data	(Mathieu et al., 2020)
human_dev_ index	Life expectancy at birth in 2019	Primary data	(Nations, n.d.)
total_alcohol_ per	Total alcohol per capita (≥ 15 years of age) consumption (liters of pure alcohol).	Comparable estimates	(*World Health Statistics*, 2021)
tabacco_ prevelance	Prevalence of tobacco use among people aged 15 years and over, (age-standardized) (%).	Comparable estimates	(WHO, 2019)
p_cvd_c_d_de_ rate	Probability (%) of dying between 30 and 70 years of age from any of CVD, cancer, diabetes and CRD.	Comparable estimates	(WHO, 2020b)
cardiovasc_ death_rate	Mortality rate from cardiovascular diseases (2017) (deaths per 100,000 people per year)	Primary data	(Mathieu et al., 2020)
diabetes_ prevalence	Prevalence of diabetes in the population aged 20-79 years [2017] (%).	Comparable estimates	(Mathieu et al., 2020)
tuberculos_ incidence	Tuberculosis incidence (per 100 000 population).	Comparable estimates	(WHO, 2020a)
SBP_DBP	Prevalence of high blood pressure (>140 mmHg SBP and/or >90 mmHg DBP) among persons aged >18 years [age-standardized].	Comparable estimates	(*Noncommunicable Diseases: Risk Factors*, 2017)
obesity_ prevalence	Age-standardized prevalence of obesity among adults (18+ years) (%).	Comparable estimates	(*Noncommunicable Diseases: Risk Factors*, 2017)

2.2.Methods

Cluster analysis (clustering) is a widely used method, especially in fields like data mining, pattern recognition and machine learning. This method of analysis aims to group observations or data points according to their similar characteristics. The basic stages of clustering analysis can be summarized as data pre-processing, selection of the appropriate clustering algorithm, evaluation and interpretation of the results.

2.2.1. The Phases of Clustering Analysis

Data Preprocessing: The first step in cluster analysis is to make the data suitable for analysis. This step may involve processing missing values in the data, normalizing the data if necessary and reducing the data size. Data preprocessing allows for clearer distinctions between clusters and improves the accuracy of the analysis process (Aggarwal & Reddy, 2013).

Selection of an Appropriate Clustering Algorithm: Different clustering algorithms vary according to different data types. Some commonly used algorithms are as follows: K-Means, Hierarchical Clustering and Density Based Clustering (DBSCAN). The K-Means algorithm is particularly effective for large datasets, but the algorithm requires the number of clusters to be determined in advance. Hierarchical Clustering offers better results in understanding the internal structure of the dataset but can be computationally expensive for large datasets (Tan et al., 2018).

Evaluation and Validation of Results: In evaluating the results of clustering analysis, the homogeneity of each cluster and the heterogeneity between clusters are analyzed. Some validation metrics commonly used in this phase are Silhouette score, Davies-Bouldin index and intra-cluster average Distances. Choosing the right evaluation metrics increases the significance and validity of the clustering result (Raschka & Mirjalili, 2017).

Interpretation and Visualization of Results: In the final stage of the analysis, the relationships between clusters are interpreted in terms of meaningful patterns and business knowledge. Visualizing cluster results, especially for multidimensional data, supports inferences and decision-making processes. Graphical methods and dimensionality reduction techniques (e.g. PCA or t-SNE) are frequently used at this stage (Jain & Dubes, 1988).

2.2.2. K-means

As one of the unsupervised learning methods, clustering analysis allows the observations to be found in an association on the basis of variables or to be categorized into subgroups to form more than one group. *k-means* clustering is a vector-based method that aims to partition n observations into k clusters, where each observation belongs to the cluster with the closest mean (cluster center's), which is taken as the prototype of a possible cluster. It minimizes intra-cluster variances (quadratic Euclidean distances) and optimizes quadratic errors.

With a set of observations X is as $(x_1, x_2, \ldots, x_n)$, each observation a l -dimensional real vector, *k*-means clustering is a method to partition the n observations into $k(\leq n)$ sets $K = \{K_1, K_2, \ldots, K_k\}$ in such a way that the within-cluster sum of squares (WSS) (i.e. the variance) is minimized. Formally, the objective is to find:

$$\arg_K \min \sum_{i=1}^{k} \sum_{x \in K_i} x - \mu_i^2 = \arg_K \min \sum_{x \in K_i} |K_i| \mathrm{Var} K_i \tag{1}$$

where μ_i is the mean is points in K_i (also called centroid of cluster).

$$\mu_i = \frac{1}{|K_i|} \sum_{x \in K_i} x \tag{2}$$

$|K_i|$ is the size of K_i, and . is the L^2 norm. This corresponds to the minimization of the pairwise squared deviations of points in the same cluster:

$$\arg_K \min \sum_{i=1}^{k} \frac{1}{|K_i|} \sum_{x,y \in K_i} x - y^2 \tag{3}$$

Equivalence follows from identity $|K_i| \sum_{x \in K_i} x - \mu_i^2 = \frac{1}{2} \sum_{x,y \in K_i} x - y^2$. As the total variance is constant, it is equivalent to minimizing the sum of squares of the deviations between points in different clusters (between-cluster sum of squares, BSS)(Kriegel et al., 2017).

2.2.3. Hierarchical clustering

Hierarchical clustering, one of the unsupervised learning methods, is used to perform data exploratory analysis. Hierarchical clustering roughly consists of creating a dendrogram shape that forms a tree structure by binary

merging from the data items (individual observations) stored in the leaves. The information stored in the nodes is combined in pairs to form a tree structure that minimizes the distance between two subclusters. Hierarchical clustering is considered in two categories: agglomerative hierarchical clustering and divisive hierarchical clustering (Nielsen, 2016). Basically, hierarchical clustering presents the observations in the data set in the form of a dendrogram that forms a tree structure that will branch up to a single subgroup (leaf) based on variables. In this visualization, it is an indication that clustering is possible in the data set in exploratory data analysis.

3. Clustering Analysis and Results

Cluster analysis is performed on our data set of 38 countries. There is no missing value on observations in the data set. R statistical programming language is used for all calculations related to clustering analysis. To observe the suitability of the data set for a possible clustering analysis, the hierarchical clustering method is applied and the dendrogram is drawn and presented in Figure 1.

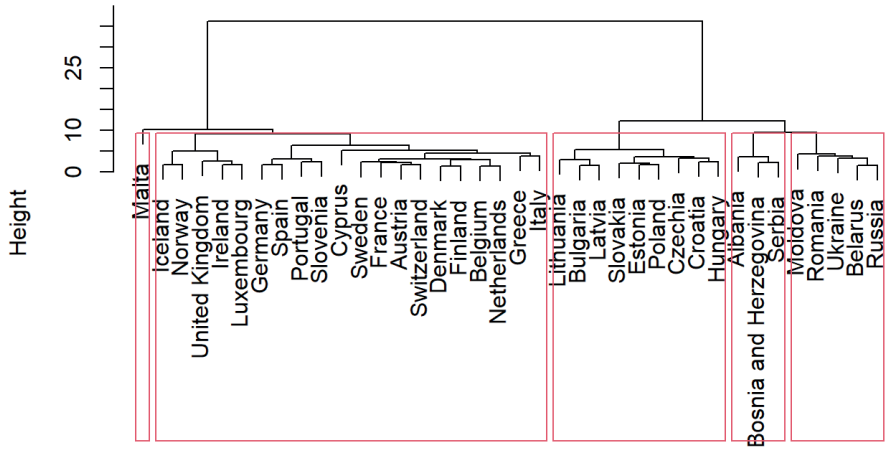

Figure 1 The countries' dendrogram of health status clustered by hierarchical clustering

The leaves of a tree are in the same cluster if there is a path between them. This dendrogram suggests that two sets can be a proper number, as shown in Figure 1. The exploratory data analysis with the hierarchical clustering method showed that the data set is suitable for clustering analysis. In clustering analysis, the most relevant question to be asked is "What is the optimal number of clusters for this data set?".

In determining the number of clusters, it is a common practice to obtain information and expert opinions in the field. Otherwise, calculations can be made with various methods and ***k***-values (number of clusters) several times. In this way, the optimum number of clusters can be reached. Thus, elbow method, which is a frequently used method for determining the optimum number of clusters, is used. The elbow method is to calculate the sum in total squares (WSS) for different values of ***k*** and look for an 'elbow' in the curve (Zumel & Mount, 2020). For the elbow method, the WSS calculated for 4 possible clusters is given in Figure 2.

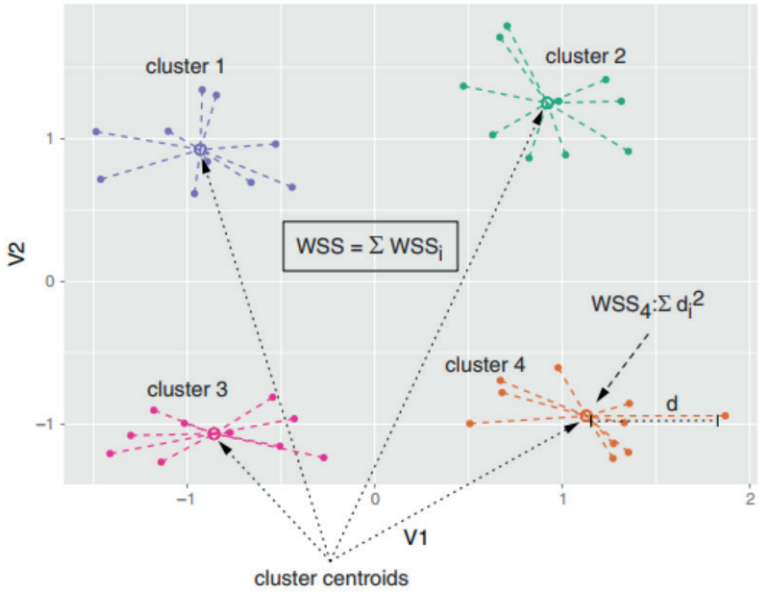

Figure 2. Cluster WSS and total WSS for set of four clusters (Zumel & Mount, 2020)

To find the number of clusters in the clustering of the health status of European countries, the WSS values from 2 to 10 are calculated and presented in Figure 3.

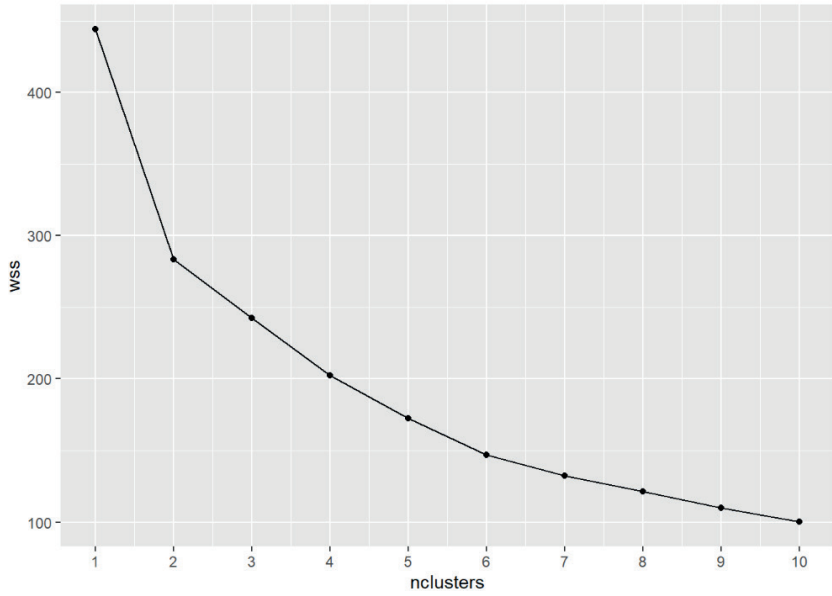

Figure 3. Cluster WSS and total WSS for set of four clusters

The elbow is a bit difficult to see in Figure 3. However, if one focuses our attention on the points where the WSS values fall, it can be said that the number of clusters of the elbow can be 2, 4 or 6. A further method for determining the number of clusters, the Calinski-Harabasz Index (CH), is another widely used measure for optimizing the number of clusters. To calculate the Calinski-Harabasz Index (or CH Index for short), a few terms have to be defined first. The Total Sum of Squares (TSS) of a point set is the sum of the squares of the distances of all points from the center of the data. For a given clustering with a total sum of squares, it is also possible to define the sum of squares (BSS) as follows:

$$BSS = TSS - WSS \tag{4}$$

The *WSS* measures the distance between clusters. A good clustering has a small *WSS* (all clusters should be distributed around their center) and a large *BSS*.

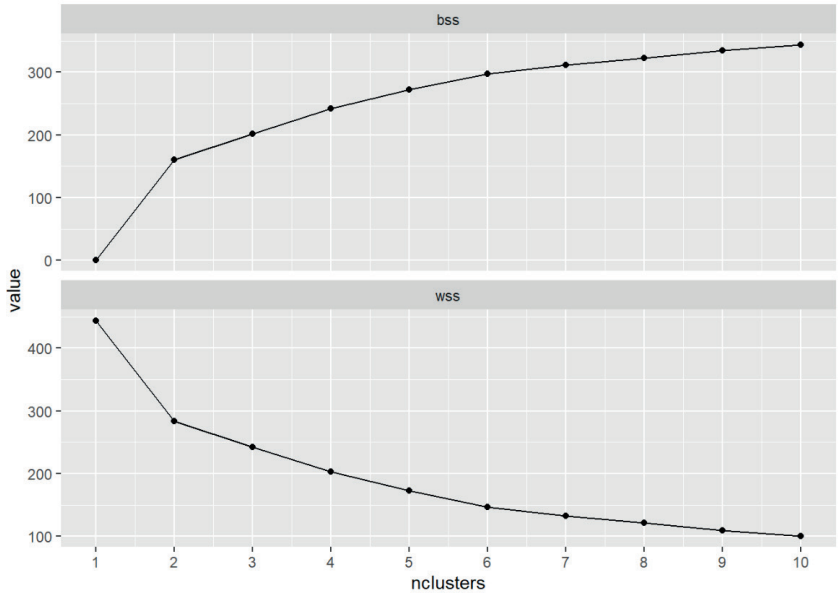

Figure 4. Cluster WSS and total WSS for set of four clusters

Figure 4 shows that BSS increases and WSS decreases as $\boldsymbol{k}$ increases. One expects to find a cluster with a good balance between BSS and WSS. Several measures of BSS and WSS need to be analysed to find such a cluster. The intra-cluster variance is given by W:

$$W = WSS / (n - k) \tag{5}$$

where n is the number of data points and k is the number of clusters. Cluster variance is given by B.

$$B = BSS / (k - 1) \tag{6}$$

Again, we can think of B as the average contribution to BSS from each cluster.

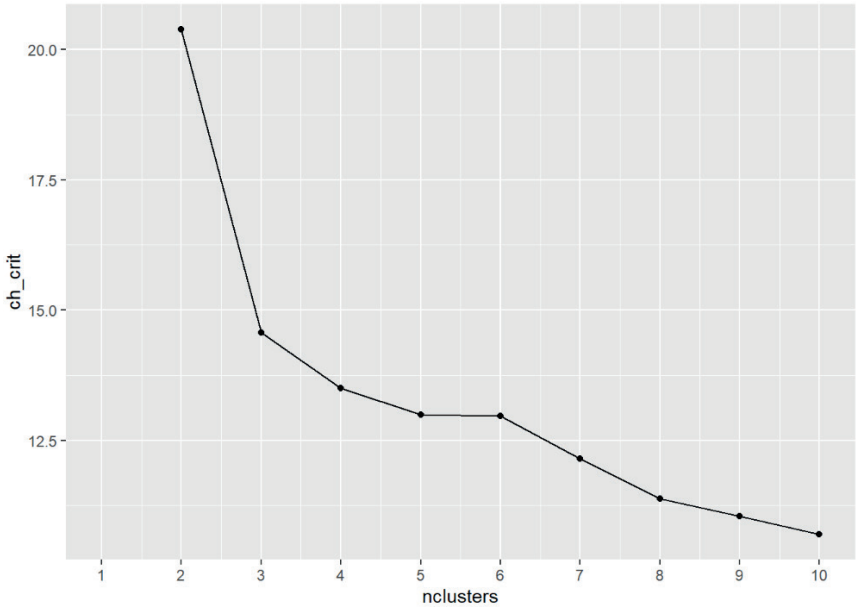

Figure 5. Calinski-Harabasz index of European Health Status clusters

CH criterion maximized at $k=2$, at $k=5$ is another local maximum in Figure 5. It has been observed that 2 and 5 clusters are more appropriate for ***ch index*** and 2 and 6 clusters for elbow. Another calculation scores frequently used to determine the number of clusters is the ***average silhouette width*** (**asw**) (Zumel & Mount, 2020). The results of Ch index and silhouette score in determining the number of clusters were calculated from $k=2 to 10$ and presented in Figure 6.

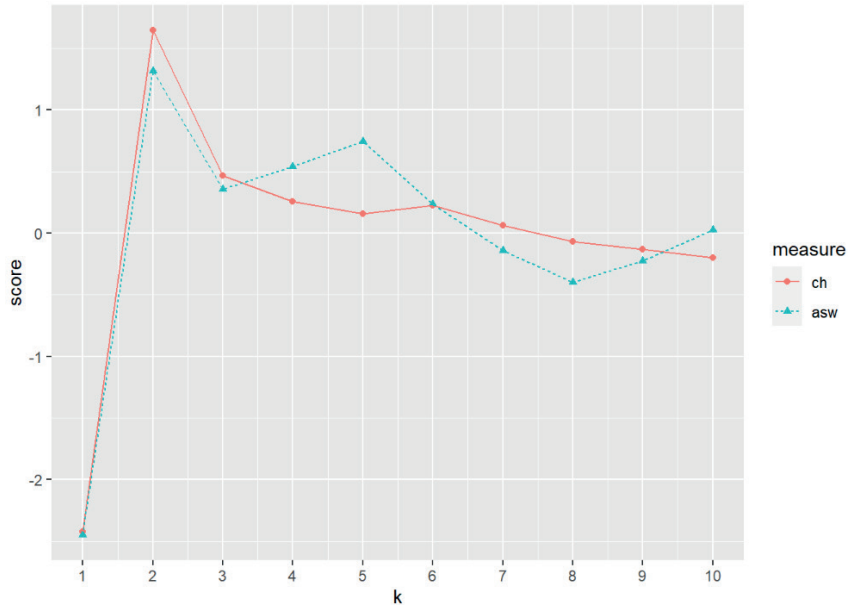

Figure 6. Ch index and asw score according to the number of clusters k=2 to 10

Both scores indicate that the optimum number of clusters for the data set is achieved for $k = 2$. The number of clusters for the k-means method is determined as $k = 2$. By applying the ***k-means*** method for $k = 2$, the data set consisting of 38 countries is divided into 2 clusters as 23 and 15 countries. The countries categorized into two clusters are given in Table 2 as cluster 1 and cluster 2.

Table 2. Cluster of European Health Status Pre-Covid-19

Cluster 1		**Cluster 2**	
Austria	Italy	Albania	Moldova
Belgium	Luxembourg	Belarus	Poland
Cyprus	Malta	Bosnia and	Romania
Czechia	Netherlands	Herzegovina	Russia
Denmark	Norway	Bulgaria	Serbia
Estonia	Portugal	Croatia	Slovakia
Finland	Slovenia	Hungary	Ukraine
France	Spain	Latvia	
Germany	Sweden	Lithuania	
Greece	Switzerland		
Iceland	United Kingdom		
Ireland			

It is expected that in clustering analysis, as well as the separation of the data set into clusters, the reasons for the separation of the variables belonging to the separated clusters should also reveal a logical insight. To this end, the cluster means of the variables in the data set are calculated and the cluster means of the variables for cluster 1 and cluster 2 are given in Table 3.

Table 3. Cluster averages based on variable

Variables	Cluster 1	Cluster 2
median_age	42.5435	41.7400
population_density	203.6304	77.0653
life_expectancy	81.9304	75.8173
human_dev_index	0.9213	0.8267
total_alcohol_percapita	10.9043	10.6533
tabacco_prevelance	25.8174	31.1667
p_cvd_c_d_death_rate	10.4000	20.4333
cardiovasc_death_rate	138.4935	362.1400
diabetes_prevalence	5.9422	6.9933
tuberculos_incidence	7.3043	32.1200
SBP_DBP	20.9348	29.1467
obesity_prevalence	22.4652	22.9000

It is observed that the median age of the clusters is quite close. It can be said that the population density is lower in cluster 2 and there is more of rural life. According to a rough analysis, it can be said that the individual health is worse in cluster 2 and some of the diseases are already worse than in cluster 1 at a level that will affect life expectancy.

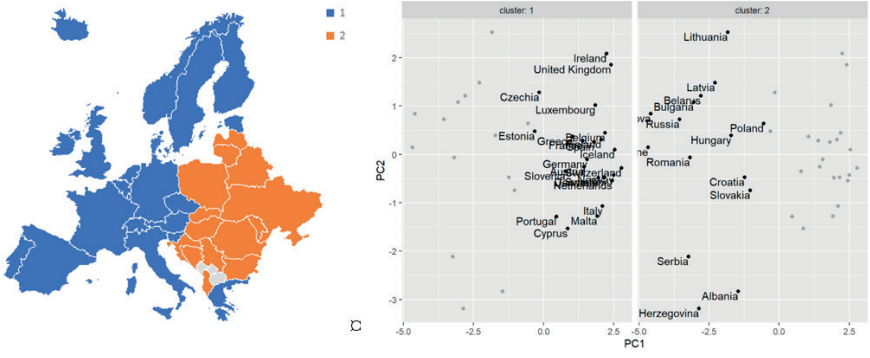

Figure 7. European health status map pre-Covid-19 and PCA clusters

The countries falling into clusters by PCA are visualized and given in Figure 7 in order to better observe the cluster decomposition formed in the clustering analysis. The countries colored (blue is cluster 1 and orange is cluster 2) according to the clusters on the map of Europe can be clearly seen in Figure 7. It is seen that the pre-Covid-19 health status of the eastern bloc of the European continent is worse than the western bloc.

The visual representation of the separation of countries in the clustering analysis is the cluster dendrogram given in Figure 8. It is clearly seen that the countries in the leaves in the tree's branching structure are divided into two different clusters.

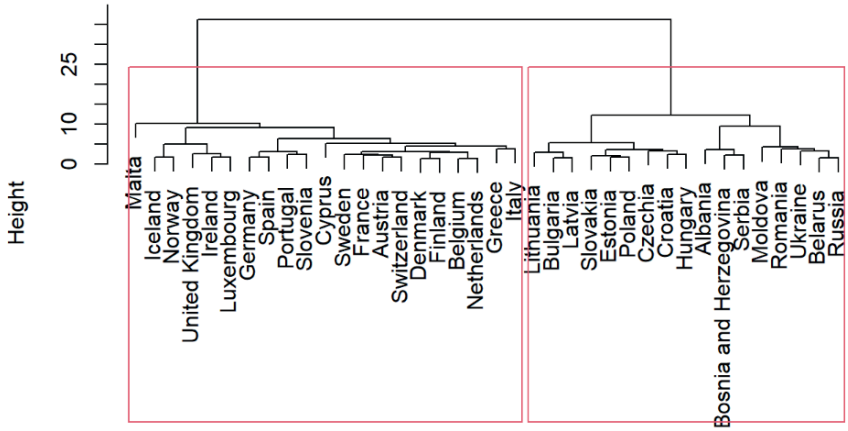

Figure 8. The countries' dendrogram of health status clustered by hierarchical clustering $k = 2$

The question that always comes to mind when evaluating clusters is 'Is a cluster "real"?' or 'Does the cluster represent the real structure in the data or is it an artefact of the clustering algorithm?'. One way to assess whether a cluster represents the true structure is to see if the cluster falls below reasonable variations in the data set. For this purpose, the performance of the cluster algorithm in predicting the real cluster with bootstrap on the dataset divided into two clusters with the k-means algorithm is measured and it is calculated as 0.9404 for cluster 1 and 0.9171 for cluster 2.

This shows that the stability of clustering is partly a function of the clustering algorithm and not only of the data. Also, the fact that both clustering algorithms discover the same clusters can be taken as an indication that 2 is the optimal number of clusters.

4. Conclusion

At the end of this chapter, topics such as estimating the appropriate number of clusters for a dataset, clustering a dataset using both hierarchical clustering and k-means, and evaluating the resulting clusters for pre-Covid-19 health status in Europe are covered. The aim of clustering is to discover or uncover similarities between subsets of the data. In a good clustering, the points in the same cluster should be more similar (closer) than the points in other clusters.

This study, looking at the clustering of pre-Covid-19 health status in Europe , shows that Europe is divided into 2 distinct clusters based on the variables in the dataset. The logical framework of this distinction is given in the results section. As a result, it can be seen that the cluster division is based on the health indicators of European countries in the pre-Covid-19 period and their preparedness for a possible pandemic. A disease or a pandemic can knock on the door at any time, but are we ready for a pandemic?

References

Ada, E., Sagnak, M., Mangla, S. K., & Kazancoglu, Y. (2024). A circular business cluster model for sustainable operations management. *International Journal of Logistics Research and Applications*, *27*(4), 493–511. https://doi.org/10.1080/13675567.2021.2008335

Aggarwal, C. C., & Reddy, C. K. (2013). *Data Clustering Algorithms and Applications* (1st ed.). CRC Press.

Carrillo-Larco, R. M., & Castillo-Cara, M. (2020). Using country-level variables to classify countries according to the number of confirmed COVID-19 cases: An unsupervised machine learning approach. *Wellcome Open Research*, *5*, 56. https://doi.org/10.12688/wellcomeopenres.15819.1

Farseev, A., Chu-Farseeva, Y.-Y., Yang, Q., & Loo, D. B. (2020). *Understanding Economic and Health Factors Impacting the Spread of COVID-19 Disease*. https://doi.org/10.1101/2020.04.10.20058222

Imtyaz, A., Abid Haleem, & Javaid, M. (2020). Analysing governmental response to the COVID-19 pandemic. *Journal of Oral Biology and Craniofacial Research*, *10*(4), 504–513. https://doi.org/10.1016/j.jobcr.2020.08.005

Jain, A. K., & Dubes, R. C. (1988). *Algorithms for clustering data*. Prentice-Hall, Inc.

Kiaghadi, A., Rifai, H. S., & Liaw, W. (2020). Assessing COVID-19 risk, vulnerability and infection prevalence in communities. *PLoS ONE*, *15*(10 October). https://doi.org/10.1371/journal.pone.0241166

Kriegel, H. P., Schubert, E., & Zimek, A. (2017). The (black) art of runtime evaluation: Are we comparing algorithms or implementations? *Knowledge and Information Systems*, *52*(2), 341–378. https://doi.org/10.1007/s10115-016-1004-2

Mathieu, E., Ritchie, H., Rodés-Guirao, L., Appel, C., Giattino, C., Hasell, J., Macdonald, B., Dattani, S., Beltekian, D., Ortiz-Ospina, E., & others. (2020). Coronavirus pandemic (COVID-19). *Our World in Data*.

Nations, U. (n.d.). Country Insights. *Human Development Reports*.

Nicholson, L. B. (2016). The immune system. *Essays in Biochemistry*, *60*(3), 275. https://doi.org/10.1042/EBC20160017

Nielsen, F. (2016). *Hierarchical Clustering* (pp. 195–211). https://doi.org/10.1007/978-3-319-21903-5_8

Noncommunicable diseases: Risk factors. (2017). https://www.who.int/data/gho/data/themes/topics/topic-details/GHO/ncd-risk-factors

Porcheddu, R., Serra, C., Kelvin, D., Kelvin, N., & Rubino, S. (2020). Similarity in Case Fatality Rates (CFR) of COVID-19/SARS-COV-2 in Italy and China. *Journal of Infection in Developing Countries*, *14*(2), 125–128. https://doi.org/10.3855/jidc.12600

Raschka, S., & Mirjalili, V. (2017). *Python Machine Learning* (1st ed.). Packt. https://www.packtpub.com/en-us/product/python-machine-lear-

ning-9781783555130?type=print&srsltid=AfmBOoq8veAbXuClFO-etHOJOPDZGKqCyJdNJiaZYEiZmhddiAQ8xUNXZ

Ren, Y., Pu, J., Yang, Z., Xu, J., Li, G., Pu, X., Yu, P. S., & He, L. (2024). Deep Clustering: A Comprehensive Survey. *IEEE Transactions on Neural Networks and Learning Systems*, 1–21. https://doi.org/10.1109/tnnls.2024.3403155

Rizvi, S. A., Umair, M., & Cheema, M. A. (2021). Clustering of countries for COVID-19 cases based on disease prevalence, health systems and environmental indicators. *Chaos, Solitons and Fractals*, *151*. https://doi.org/10.1016/j.chaos.2021.111240

Saunders, J. A. (1980). Cluster Analysis for Market Segmentation. *European Journal of Marketing*, *14*(7), 422–435. https://doi.org/10.1108/EUM0000000004918/FULL/XML

Sudre, C. H., Lee, K. A., Ni Lochlainn, M., Varsavsky, T., Murray, B., Graham, M. S., Menni, C., Modat, M., E Bowyer, R. C., Nguyen, L. H., Drew, D. A., Joshi, A. D., Ma, W., Guo, C.-G., Lo, C.-H., Ganesh, S., Buwe, A., Capdevila Pujol, J., Lavigne du Cadet, J., … Ourselin, S. (2021). Symptom clusters in COVID-19: A potential clinical prediction tool from the COVID Symptom Study app. In *Sci. Adv* (Vol. 7). https://www.science.org

Tan, P., Steinbach, M., Karpatne, A., & Kumar, V. (2018). *Introduction to Data Minning* (2nd ed.).

Tressa, N., Asha, V., Kumar, P., Shree, O., Uday Kiran, M., & Reddy, V. V. S. (2024). Customer-Based Market Segmentation Using Clustering in Data mining. *2nd International Conference on Intelligent Data Communication Technologies and Internet of Things, IDCIoT 2024*, 687–691. https://doi.org/10.1109/IDCIoT59759.2024.10467258

Wedel, M., & Kamakura, W. A. (2000). *Market Segmentation* (Vol. 8). Springer US. https://doi.org/10.1007/978-1-4615-4651-1

WHO. (2019, December 18). *WHO global report on trends in prevalence of tobacco use 2000-2025, third edition*. https://www.who.int/publications/i/item/who-global-report-on-trends-in-prevalence-of-tobacco-use-2000-2025-third-edition

WHO. (2020a). *GLOBAL TUBERCULOSIS REPORT 2020*. http://apps.who.int/bookorders.

WHO. (2020b). *Mortality and global health estimates*. https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/

World Health Statistics. (2021). https://www.who.int/data/gho/data/themes/topics/topic-details/GHO/world-health-statistics

Zumel, Nina., & Mount, John. (2020). *Practical data science with R* (2nd ed.). Manning Publications.