

Comparison of K-Means and Two-Step Cluster Analysis Methods for Clustering COVID-19 Data

Sawitree Pansayta¹, Wirapong Chansanam^{2*}

^{1,2}Department of Information Science, Faculty of Humanities and Social Sciences, Khon Kaen University, Khon Kaen, Thailand.

E-mail: wirach@kku.ac.th

Abstracts: This study compares the K-Means and two-step cluster analysis methods for clustering COVID-19 data. The dataset had 1,893,941 cumulative cases from January 2020 to October 2021. K-means clustering resulted in eight clusters, while two-step cluster analysis clustering resulted in three grouped cases by nationality, occupation, patient type, and risk group. These clusters were categorized based on age, gender, nationality, occupation, and region of infection. Group 1 had 5,883 workers infected in community settings, Group 2 had 7,420 foreign migrant workers infected in industrial settings or through direct contact with patients, and Group 3 had 6,870 cases of indirect transmission. The study recommends targeted interventions and continued monitoring and evaluation based on the clusters. The findings can help improve government policies, medical facilities, and treatment.

Keywords: K-Means, Two-step cluster analysis, Clustering, COVID-19, Thailand

1. INTRODUCTION

According to a publication by Abd-Alrazaq et al. [1], the White House enlisted the help of the global artificial intelligence (AI) community to combat COVID-19. The researchers focused on the challenge of combating misinformation during the pandemic, which led to the spread of the virus and unhealthy mask-wearing practices. Tasnim et al. [2] suggested the use of advanced data mining techniques like natural language processing to detect and remove non-scientific online information. Ayyoubzadeh et al. [3] emphasized that data mining can forecast the global expansion and trends of COVID-19. The authors used the LSTM model to analyze Google Trends data. To tackle COVID-19, Franch-Pardo et al. [4] recommended an interdisciplinary approach that includes data mining, web-based mapping, and spatiotemporal analysis. Li et al. [5] categorized COVID-19 news and user-generated content using linear regression and content analysis. Qin et al. [6] estimated new and confirmed cases of COVID-19 using social media search indexes (SMSI) for symptoms like dry cough and fever. Kumar [7] described how AI and modern technologies like ML and NLP can help fight COVID-19. Ren et al. [8] used publication mining to find links between diabetes and COVID-19 research. Huang et al. [9] performed data mining on 485 suspected COVID-19 patients from Sina Weibo to study the number of infected people seeking medical advice on the platform. They recommended a classification model for treatment. Sarker et al. [10] identified positive-tested COVID-19 patients on Twitter using a semi-automated filtering process.

The k-means algorithm is a widely used clustering method that minimizes grouping errors. However, its performance is highly dependent on its initial state or local search procedure [11]. The algorithm partitions the data into k clusters, which are mutually exclusive and recover the specified cluster index for all observations. K-means is particularly effective for large datasets, as it is generally more accurate than other clustering methods [12, 13]. It measures the cohesiveness of objects in a dataset by treating each observation as an object located in space. The intra-cluster distance between a point and its cluster center is a good indicator of how tightly knit a cluster is [14-16].

Two-step cluster analysis is a statistical method used to group objects or cases in a dataset based on their similarity [17]. It is a type of unsupervised machine learning that can be used for clustering analysis, which is the process of dividing a dataset into groups or clusters that share similar characteristics. Two-step cluster analysis is particularly useful for large datasets with many variables, as it can identify the most relevant variables and group objects based on those variables. This technique is commonly used in various fields such as healthcare [18, 19], tourism [20, 21], and transport [22, 23]. Two-Step Cluster Analysis is a type of clustering method used in data analysis. It is a technique that automatically categorizes a large number of cases into different groups or clusters based on their similarities and differences. The two-step cluster analysis approach involves two main stages. In the

first stage, the algorithm uses a pre-clustering technique called "hierarchical clustering" to group the observations into a small number of "seed" clusters [24]. This is done by calculating the distance between each observation and the existing clusters, and merging the closest clusters until the desired number of clusters is reached. In the second stage, the algorithm uses a statistical model called "logistic regression" to refine and optimize the clustering results [25]. It does this by assigning each remaining observation to the cluster that provides the best fit according to a set of criteria such as the Bayesian Information Criterion (BIC) [17] or the Akaike Information Criterion (AIC). Two-step cluster analysis has several advantages over other clustering methods. For one, it can handle large datasets with many variables efficiently. It also automatically selects the number of clusters and optimizes the cluster assignments, making it a powerful tool for exploratory data analysis. Overall, Two-Step Cluster Analysis is a useful and powerful tool for identifying groups or clusters in large and complex datasets, and it can provide valuable insights for decision-making and problem-solving in various fields.

This article explores how data mining can be used to analyze COVID-19 data using two different methods: K-Means and Two-step cluster analysis. Data mining is a common technique used to find patterns and correlations in large datasets, which can help in making better decisions. Scientists use data mining techniques to identify correlations between variables, such as individual decision-making and group strategies, to fight the pandemic.

2. MATERIALS AND METHODS

This research framework involves three stages: data acquisition and pre-processing, knowledge discovery using comparative COVID-19 data between k-means and Two-Step Cluster Analysis insights. At the end of the research, data interpretation and further implementation are provided. The study consists of five main steps.

To obtain data for the COVID-19 outbreak, daily new cases were sourced from the official open government data website (<https://data.go.th/dataset/covid-19-daily>) [26]. The sample size was 1,893,941 cases collected from January 2020 to October 2021. After data cleansing and extraction using a data mining program, 20,100 usable examples were obtained. The first study uses the k-means clustering method, which classifies a dataset into a specific number of clusters [27]. It groups COVID-19 datasets based on a known number of clusters that are inputted by the end-user. To identify the appropriate number of clusters, performance evaluation or cluster validity is used. The k-means clustering method faces challenges in determining the ideal number of clusters, and data analysts must rely on their experience to select a suitable value for "k". This study suggests using the "elbow method" as a commonly used approach for identifying the best number of clusters through a process called "cluster validation."

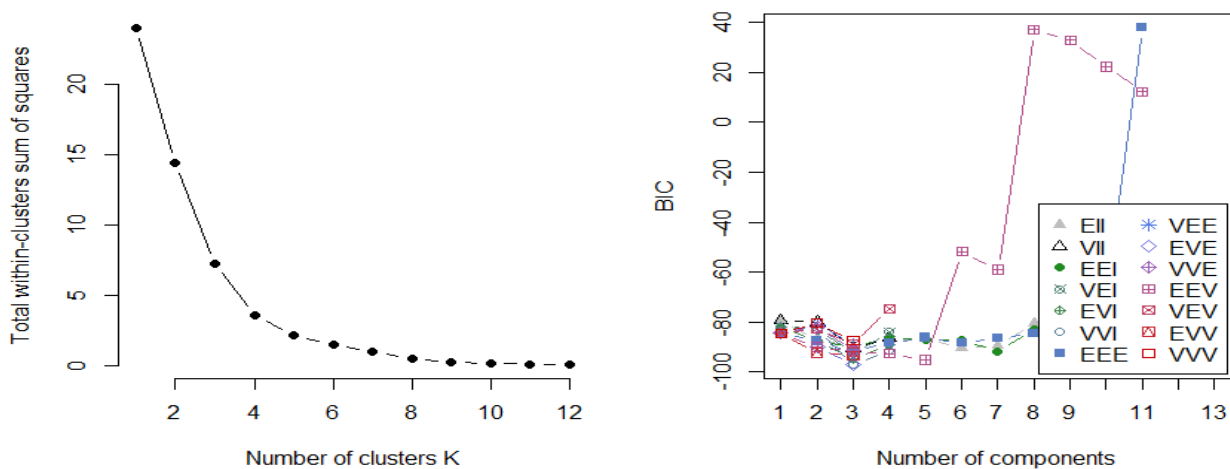
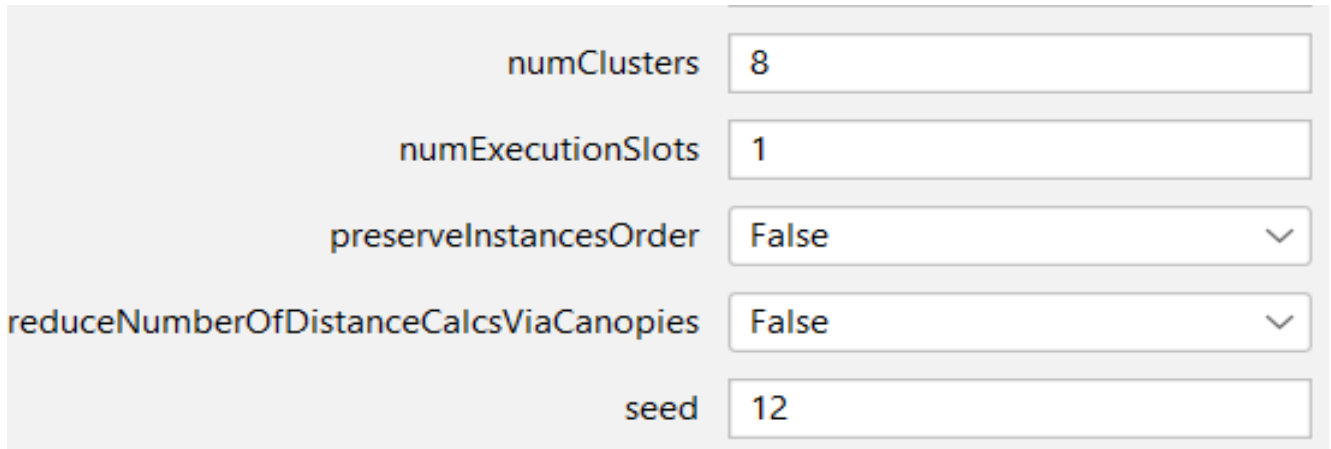


Figure 1. Elbow Plot of the Best Number of Clusters and a Bayesian Inference Criterion (BIC).

Based on the elbow plot in figure 1, it appears that $k = 4, 5, 6,$ and 7 are all valid options for the optimal number of clusters, with $k = 8$ being a strong possibility. Another method using Bayesian Inference Criterion (BIC) for k -means was also considered, which generated a probability for the Gaussian mix using the k -means model. The predicting Enhanced Vegetation Index (EVI) model was chosen for its accuracy, with 8 clusters being the most reasonable option. The k -means cluster method was used for the clustering process, and the outcomes were shown for $4, 5, 6, 7,$ and 8 clusters using WEKA (see Figure 2).



numClusters	8
numExecutionSlots	1
preserveInstancesOrder	False
reduceNumberOfDistanceCalcsViaCanopies	False
seed	12

Figure 2: Weka Clusters Simple K-Means Parameter Box.

In the second step of the analysis, the same dataset was analyzed using the Two-Step Cluster Analysis. IBM SPSS Statistics Version 28.0.10 (142) was used to clean and analyze the data. The selected variables were categorized as Categorical Variables. To determine the appropriate grouping, Schwarz's Bayesian Criterion (BIC) or Akaike Information Criterion (AIC) were used, but using the lowest BIC or AIC could result in too many groups. Therefore, the Ratio of Distance Measures was used to analyze the data. This measure indicates that the distance between each group is far apart [28]. The system then automatically selects the appropriate number of groups, as shown in figures 3 and 4.

Cluster Distribution

		N	% of Combined	% of Total
Cluster	1	5883	29.3%	29.3%
	2	7420	36.9%	36.9%
	3	6807	33.8%	33.8%
	Combined	20110	100.0%	100.0%
Total		20110		100.0%

Figure 3. Cluster Distribution using Schwarz's Bayesian Criterion (BIC) method.

Cluster Distribution

	N	% of Combined	% of Total
Cluster 1	5883	29.3%	29.3%
Cluster 2	7420	36.9%	36.9%
Cluster 3	6807	33.8%	33.8%
Combined	20110	100.0%	100.0%
Total	20110		100.0%

Figure 4. Cluster Distribution using Akaike Information Criterion (AIC) method.

For the clustering analysis, we selected variables that are important for grouping the data: nationality, occupation, risk group, and patient type. After the analysis, only these four variables remained. We changed the display from the model summary diagram to the clusters table (see Figure 5), which presents the clustering results and details within the clusters. We then analyzed the data in each cluster to summarize the results and name each cluster. This resulted in a model for categorizing COVID-19 patients in Thailand.

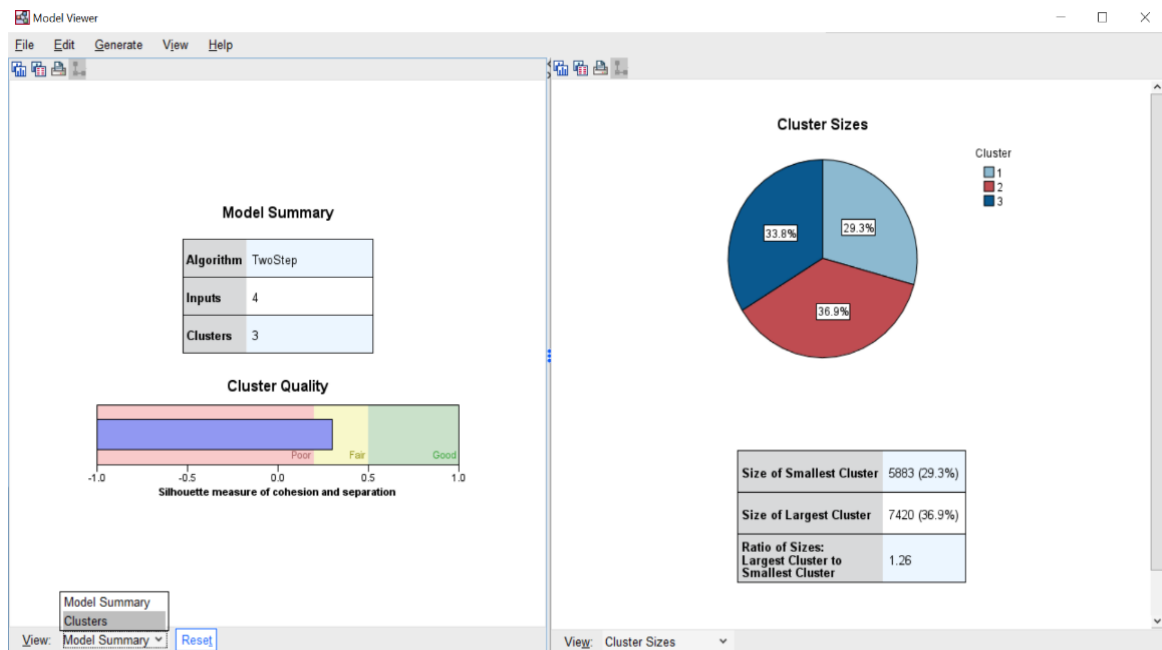


Figure 5. Model Viewer.

3. RESULTS

3.1. K-Means Techniques Result

We utilized each cluster’s centroid to show the characteristics of each type of cluster in Table 1. Thai COVID-19 patients in the third wave were mostly females with an average age of 34.72, working in the industrial sector, having close contact with a previously confirmed patient and touching an infected person, and living in central Thailand. Most of them recovered, and mortality was low. However, clustering with cluster=8 was unable to explain comorbidity and mortality due to COVID-19.

Table 1. Characteristics of Thai COVID-19 Patients using 8 Clusters.

Cluster	Characteristic	Percentage of member
0	ripple_2,female,27,foreigners,'industrial career',workplace,'check by themselves',central	17
1	ripple_3,female,54,thai,'Commerce and service careers','Close contact with a previous confirmed patient','touch an infected person',central	26
2	ripple_2,male,43,thai,'industrial career','Close contact with a previous confirmed patient','touch an infected person',central	11
3	ripple_1,male,24,thai,'not working',StateQuarantine,'Thai people come from abroad',central	10
4	ripple_4,female,22,thai,'Commerce and service careers','medical and public health personnel','medical personnel',north	4
5	ripple_3,male,22,thai,'not working','Close contact with a previous confirmed patient','touch an infected person',central	17
6	ripple_3,male,24,thai,'industrial career','Close contact with a previous confirmed patient','risk group survey',east	6
7	ripple_2,male,19,foreigners,'industrial career',workplace,'risk group survey',central	10

Table 1 shows the clustering results with six clusters, indicating that the majority of patients reside in central Thailand. Therefore, more intensive treatment is required for patients with these characteristics. The age group range of 19-27 years and 43-54 years or over is represented in eight clusters with similar characteristics. The clusters with the highest number of infected people are 1, 0, and 5, with patients who have come into contact with infected persons and checked themselves. Conversely, clusters 4 and 6 have the fewest infected people, consisting of those who have had close contact with a previously confirmed patient. It appears that people who maintain social distancing are less likely to get infected.

3.2. Two Step Clustering Analysis

This study analyzed data from 20,110 COVID-19 patients using Two Step Clustering Analysis with four variables: nationality, occupation, riskGroup, and patient_type. The analysis resulted in three groups: Group 1 had 5,883 confirmed cases (29.3%), Group 2 had 7,420 confirmed cases (36.9%), and Group 3 had 6,870 confirmed cases (33.8%), as shown in Figure 6. Further analysis showed that Group 2 had the highest number of cases, followed by Group 3 and then Group 1, as illustrated in Table 2.

Cluster Distribution

	N	% of Combined	% of Total
Cluster 1	5883	29.3%	29.3%
Cluster 2	7420	36.9%	36.9%
Cluster 3	6807	33.8%	33.8%
Combined	20110	100.0%	100.0%
Total	20110		100.0%

Model Summary

Algorithm	TwoStep
Inputs	4
Clusters	3

Cluster Quality

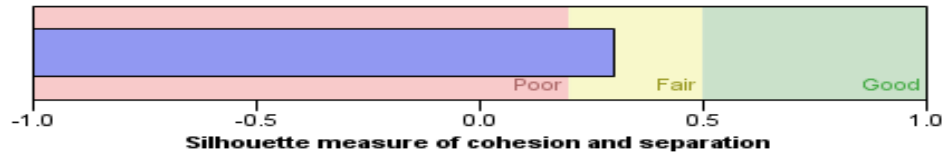


Figure 6. The Results of Two Step Clustering Analysis.

Table 2. Results of grouping.

Factor	Cluster 1	Cluster 2	Cluster 3
	5,883 Cases	7,420 Cases	6,807 Cases
Nationality	Thai (99.2%)	Foreigners (63.9%)	Thai (100%)
Occupation	Commerce and service careers (51.0%)	Industrial career (79.9%)	Not working (34.6%)
Patient_type	Thai people come from abroad (26.4%)	Risk group survey (58.4 %)	Touch an infected person (100%)
Risk group	Community (45.0%)	Workplace (66.7%)	Close contact with a previous confirmed patient (100%)

3.3. The Details of Data Analysis used Two Step Clustering Analysis

The COVID-19 patients in Thailand can be classified into three main groups. Group 1 includes Thai nationals who work in the trading and service industry and are at risk of being infected through activities such as visiting restaurants, boxing stadiums, and entertainment places. Group 2 comprises foreign migrant workers who work in the industrial sector and are likely to be infected through their workplaces. This is the largest cluster. Group 3 consists of individuals who were infected indirectly by close contact with patients from the first two groups. Overall, these three groups represent the major categories of COVID-19 patients in Thailand.

4. DISCUSSION AND CONCLUSION

This study compared K-Means and Two-step cluster analysis methods to analyze COVID-19 data in Thailand. The K-Means technique showed that demographic factors such as age, gender, nationality, career, behavioral risk, and region were related to the spread of the disease. The results from clustering consisted of eight groups. The onset of the disease was mainly in Bangkok and industrial areas, and adult workers were the main sources of new infections. Data analytics can help develop accurate prediction models for guiding treatment decisions and resource allocation, but data quality is crucial. The Two-step Cluster Analysis identified three patient groups: trade and service workers, foreign migrant workers in industry, and patients infected through contact with confirmed cases. Future research should analyze each group's characteristics, compare findings with other countries, and develop targeted interventions and prevention strategies. This model can aid in managing and planning for future outbreaks with similar characteristics.

REFERENCES

- [1] A. Abd-Alrazaq, D. Alhuwail, M. Househ, M. Hamdi, and Z. Shah, "Top concerns of tweeters during the COVID-19 pandemic: infoveillance study," *Journal of medical Internet research*, vol. 22, no. 4, p. e19016, 2020, doi: 10.2196/19016.
- [2] S. Tasnim, M. M. Hossain, and H. Mazumder, "Impact of rumors and misinformation on COVID-19 in social media," *Journal of preventive medicine and public health*, vol. 53, no. 3, pp. 171–174, 2020, doi: 10.3961/jpmph.20.094.
- [3] S. M. Ayyoubzadeh, S. M. Ayyoubzadeh, H. Zahedi, M. Ahmadi, and S. R. N. Kalhori, "Predicting COVID-19 incidence through analysis of google trends data in Iran: data mining and deep learning pilot study," *JMIR public health and surveillance*, vol. 6, no. 2, p. e18828, 2020, doi: 10.2196/18828.
- [4] I. Franch-Pardo, B. M. Napoletano, F. Rosete-Verges, and L. Billa, "Spatial analysis and GIS in the study of COVID-19. A review," *Science of the total environment*, vol. 739, p. 140033, 2020, doi:10.1016/j.scitotenv.2020.140033.
- [5] J. Li, Q. Xu, R. Cuomo, V. Purushothaman, and T. Mackey, "Data mining and content analysis of the Chinese social media platform Weibo during the early COVID-19 outbreak: retrospective observational infoveillance study," *JMIR Public Health and Surveillance*, vol. 6, no. 2, p. e18700, 2020, doi: 10.2196/18700.
- [6] L. Qin et al., "Prediction of number of cases of 2019 novel coronavirus (COVID-19) using social media search index," *International journal of environmental research and public health*, vol. 17, no. 7, p. 2365, 2020, doi.org/10.3390/ijerph17072365.
- [7] S. Kumar, "Monitoring novel corona virus (COVID-19) infections in India by cluster analysis," *Annals of Data Science*, vol. 7, no. 3, pp. 417–425, 2020, doi: 10.1007/s40745-020-00289-7.
- [8] X. Ren et al., "Identifying potential treatments of COVID-19 from Traditional Chinese Medicine (TCM) by using a data-driven approach," *Journal of Ethnopharmacology*, vol. 258, p. 112932, 2020, doi: 10.1016/j.jep.2020.112932.
- [9] C. Huang et al., "Mining the characteristics of COVID-19 patients in China: analysis of social media posts," *Journal of medical Internet research*, vol. 22, no. 5, p. e19087, 2020, doi: 10.2196/19087.
- [10] A. Sarker, S. Lakamana, W. Hogg-Bremer, A. Xie, M. A. Al-Garadi, and Y.-C. Yang, "Self-reported COVID-19 symptoms on Twitter: an analysis and a research resource," *Journal of the American Medical Informatics Association*, vol. 27, no. 8, pp. 1310–1315, 2020, doi: 10.1093/jamia/ocaa116.
- [11] A. Likas, N. Vlassis, and J. J. Verbeek, "The global k-means clustering algorithm," *Pattern recognition*, vol. 36, no. 2, pp. 451–461, 2003, doi: 10.1016/S0031-3203(02)00060-2.
- [12] I. G. Costa, F. de A. de Carvalho, and M. C. de Souto, "Comparative analysis of clustering methods for gene expression time course data," *Genetics and Molecular Biology*, vol. 27, pp. 623–631, 2004, doi: 10.1590/S1415-47572004000400025.
- [13] K. Koonsanit, C. Jaruskulchai, and A. Eiumnoh, "Determination of the initialization number of clusters in K-means clustering application using Co-occurrence statistics techniques for multispectral satellite imagery," *International Journal of Information and Electronics Engineering*, vol. 2, no. 5, pp. 785–789, 2012.
- [14] S. P. Adhau, R. M. Moharil, and P. G. Adhau, "K-Means clustering technique applied to availability of micro hydro power," *Sustainable Energy Technologies and Assessments*, vol. 8, pp. 191–201, 2014, doi: 10.1016/j.seta.2014.09.001.
- [15] G. B. Mufti, P. Bertrand, and E. L. Moubarki, "Determining the number of groups from measures of cluster stability," in *Proceedings of international symposium on applied stochastic models and data analysis*, 2005, pp. 17–20.
- [16] S. Ray and R. H. Turi, "Determination of number of clusters in k-means clustering and application in colour image segmentation," in *Proceedings of the 4th international conference on advances in pattern recognition and digital techniques*, 1999, vol. 137, p. 143.

- [17] M. J. Norušis, IBM SPSS statistics 19 advanced statistical procedures companion. prentice hall Upper Saddle River, NJ, 2012.
- [18] B. Griffin, K. A. Sherman, M. Jones, and P. Bayl-Smith, "The clustering of health behaviours in older Australians and its association with physical and psychological status, and sociodemographic indicators," *Annals of Behavioral Medicine*, vol. 48, no. 2, pp. 205–214, 2014.
- [19] D. J. McLernon, J. J. Powell, R. Jugdaohsingh, and H. M. Macdonald, "Do lifestyle choices explain the effect of alcohol on bone mineral density in women around menopause?," *The American journal of clinical nutrition*, vol. 95, no. 5, pp. 1261–1269, 2012.
- [20] C. H. Hsu, S. K. Kang, and T. Lam, "Reference group influences among Chinese travelers," *Journal of Travel Research*, vol. 44, no. 4, pp. 474–484, 2006.
- [21] A. Tkaczynski, S. R. Rundle-Thiele, and N. K. Prebensen, "Segmenting potential nature-based tourists based on temporal factors: The case of Norway," *Journal of Travel Research*, vol. 54, no. 2, pp. 251–265, 2015.
- [22] E. Cerin, E. Leslie, L. du Toit, N. Owen, and L. D. Frank, "Destinations that matter: associations with walking for transport," *Health & place*, vol. 13, no. 3, pp. 713–724, 2007.
- [23] H.-L. Chang and T.-H. Yeh, "Motorcyclist accident involvement by age, gender, and risky behaviors in Taipei, Taiwan," *Transportation research part F: traffic psychology and behaviour*, vol. 10, no. 2, pp. 109–122, 2007.
- [24] S. Okazaki, "Lessons learned from i-mode: What makes consumers click wireless banner ads?," *Computers in Human Behavior*, vol. 23, no. 3, pp. 1692–1719, 2007.
- [25] M. Norusis, SPSS 15.0 advanced statistical procedures companion. Prentice Hall Press, 2007.
- [26] Digital Government Development Agency. 2021. Thailand's daily COVID-19 information report. Retrieved July 20, 2022 from <https://data.go.th/dataset/covid-19-daily>
- [27] T. M. Kodinariya and P. R. Makwana, "Review on determining number of Cluster in K-Means Clustering," *International Journal*, vol. 1, no. 6, pp. 90–95, 2013.
- [28] J. Kertoasri, "TwoStep Cluster Analysis," p. 10, Sep. 2011. National Statistical Office, Handbook on Data Quality Assessment Methods and Tools. Bangkok: National Statistical Office, 2018.

DOI: <https://doi.org/10.15379/ijmst.v10i2.1203>

This is an open access article licensed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>), which permits unrestricted, non-commercial use, distribution and reproduction in any medium, provided the work is properly cited.