Variance
Journal of Statistics and Its Applications
P-ISSN: 2685-8738 E-ISSN: 2685-872X

Variance: Journal of Statistics and Its Applications
Volume 5 Nomor 2 | October 2023
10.30598/variancevol5iss2page131-138

WORKFORCE GROUPING IN COMPLETING PROJECTS WITH INTERN WORK ACTIVITY LOG DATA USING K-MEANS CLUSTERING

Brandon Anggawidjaja[1], Faizah Sari[2], Ahmad Fuad Zainuddin[3*]
[1,2,3]*Business Mathematics, School of Applied* STEM, Universitas Prasetiya Mulya
Jl. BSD Raya Utama, Kav. Edutown No. I.1, BSD City, Tangerang 15339, Banten, Indonesia

E-mail Corresponding Author: *ahmadfuadzain@gmail.com*

Abstract: *This study consists of an attempt to optimize the K-Means Clustering Algorithm and calculate the Full Time Equivalent (FTE) of each cluster based on the intern's daily work log data. The optimization will be done by using some of K-Means Clustering's validation methods to estimate the best K clusters of the data. The validation methods that will be used to optimize the algorithm are the Elbow Criterion Method and the Silhouette Score Index. The initial k cluster will be formed and evaluated using Davies Bouldin Index analysis. The divided clusters are supposed to be classified by the rate of complexity of each project. The calculated FTE will be used to estimate the workload for the current workforce. This estimation is hoped to help companies decide in their hiring decision.*

Keywords**:** *Complexity Project, FTE Calculation, K-Means Clustering, Workload.*

1. INTRODUCTION

There are a lot of changes that have to be done for companies to adjust to the COVID-19 pandemic. One of the changes that are quite impactful is the work system shift from offline (at the office) to online (from home). There are a few ways that a company implement this policy, and one of them is an online log report to track worker's start and end of work hours. Within the log report in several companies are listed the details regarding the activity or work stages and its durations. The change in the work system caused many companies to reevaluate the effectiveness and efficiency of their current number of employees. With the newly implemented work system, workers' productivity varies, and thus the company evaluation caused many industries to reduce their current employees according to their productiveness [1–4].

This study aims to help the company calculate the needed workforce to finish a planned project and to help decide on whether the company needs to hire new employees to lower the workload of the existing team of employees. This study can be separated into three stages, with the first stage where the log data activity will be grouped by the total hours worked on a project by an existing team of employees, which also include the average duration by workers and by tasks. The first stage will be done using the K-Means Algorithm with the help of Silhouette Score Analysis and Elbow Graph to help determine the initial K as well as the Davies-Bouldin Index to evaluate the created group [5–16]. The second stage is calculating the current workload of each group of projects that has been grouped. The second stage will be done by using the standard full-time equivalent formula [17]. The last stage will include the result from the first and second stages as the basic logic for creating a simple interface that can be used to produce recommendations regarding the workforce needed to finish future projects within the targeted time. The third stage interface will be made using a Python programming language in the Jupyter Notebook application.

2. METHODOLOGY

2.1. Data

The data that is used for this study is based on a company division that is in charge of analyzing the data of its employees as well as their customers. This study specifically focuses on the data analytics of the said company. The data consists of qualitative data such as job hierarchy, stage of analysis, name of the project, the name of the

 Submitted: 2023-09-29 *Accepted: 2023-10-27*

worker, and details for finding during any stage of analysis. The quantitative data are the date of work and duration of work that is done for each stage of analysis. This study was focused on workers within the intern level of the job hierarchy and the duration of the observation will be of one year. The data is collected by directly asking the supervisor of the interns that is working in the company.

2.2. Observed Variables

This study used the total duration of work, as well as the average duration, based on the number of workers and stages of analysis that are needed to finish each project.

2.3. Method of Analysis

This study consists of 2 analysis steps and 1 interface-building step. The first step will include grouping the observed data and labeling the grouped data based on its characteristics. The second step is about calculating the workload condition of each group. The third step will conclude the result gained from the first and second steps to create a simple interface that can give recommendations for planned projects in the future. The analysis will be conducted using Jupyter Notebook.

K-means clustering is one of the most used algorithms to group data into clusters based on similar characteristics. This is a non-hierarchical clustering method with its main procedure determining the initial centroid, calculating the nearest data points to the centroid, and calculating the new centroid based on the existing groups until the centroid stays constant [5–7]. The distance can be measured using the (1) equation. The proposed method in this study is to help approximate the optimal number of clusters by using the result comparison of Elbow Graph Analysis and Silhouette Score Analysis. Elbow Graph Analysis consists of calculating the smallest number of Within Cluster Sum of Squared Error (WCSS) for each number of clusters [8–12]. The equation to calculate the WCSS can be found in equation (2). Silhouette Score Analysis calculates the difference in distance of a data point within its cluster and the nearest data points of other clusters, which then is divided by the max value of either distance as shown in equation (3) [13–15]. Two of the most likely optimal number of clusters will then be evaluated using the Davies-Bouldin Index method, which formula can be seen in equation (4). Davies-Bouldin Index is used to make sure the distance between cluster is the farthest while the distance between data points in a cluster stay minimal [16].

$$D_{(i,j)} = \sqrt{(x_{1i} - C_{1j})^2 + (x_{2i} - C_{2j})^2 + \cdots + (x_{ki} - C_{kj})^2} \quad (1)$$

$D_{(i,j)}$: distance between data point i to centroid j

x_{ki} : data point i on the k data dimension

C_{kj} : centroid j on the k data dimension

$$SSE = \sum_{k=1}^{K} \sum_{x_i \in S_k} \left\| X_{k,i} - C_{ik} \right\|_2^2 \quad (2)$$

K : the maximum number of clusters

k : the number of clusters that is formed

X_{ki} : data point i on the k-th cluster

C_k : centroid on the k-th cluster

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i)} \quad (3)$$

$s(i)$: silhouette index/silhouette score

i : index of the data point

$a(i)$: average distance between a data point with another data point in the same cluster

$b(i)$: average distance between the nearest data point of a cluster and another data point in a different cluster

$$DB \text{ Index } = \frac{1}{k}\sum_{i=1}^{k} \max_{i \neq j} \left(\frac{S_i + S_j}{d(C_i, C_j)}\right) \quad (4)$$

k : the number of clusters that is formed

S_i : average distance between every data point with the centroid of cluster i

S_j : average distance between every data point with the centroid of cluster j

$d(C_i, C_j)$: distance between centroid in cluster i and cluster j

Full Time Equivalent (FTE) is a common method to calculate the workload that an employee has based on their work hours divided by the standard work hours stated by the government and/ or company policy [17]. The equation used to calculate the FTE can be seen in equation (5). Both work hours will be based on the same number of days.

$$\text{FTE } = \frac{\text{Actual Work Hour}}{\text{Total Work Hour Based on Company Policy}} \quad (5)$$

The interface will be built like a simple calculator with the main input based on details about planned projects.

3. RESULT AND DISCUSSION

Within the data that is used in this study, there are over 4000 rows of data. Within the duration of 1 year, the number of projects that have been completed or at least finalized to be continued in the future is only 16 projects. The log data includes daily discussion and daily progress report, which is excluded from the total duration and number of task types. The number of workers may vary between projects because the internship system consists of 2 or more interns working during the 3-months contract period. Within the 16 projects, there are some that have 5 interns working on the same projects but only have 3 interns at a time. The other 2 are different interns who also worked on the projects in the past. The grouping will be based on the characteristics of projects that are in the data. The data will be grouped, and all projects with a majority of data missing will be excluded from the analysis. The grouped and cleaned data are shown in Table 1.

Table 1. Grouped Data by Project Name

Project Name	Total Duration	Duration by Worker	Duration by Task
10	7.5	7.5	7.5
34	50.5	50.5	16.8
24	122.5	61.3	30.6
27	172.0	172.0	43.0
33	239.5	119.8	119.8
2	43.5	10.9	6.2
12	104.0	34.7	17.2
9	309.5	103.2	44.2
22	324.3	108.1	54.0
5	262.3	262.3	37.5
17	339.0	169.5	56.5
29	421.9	211.0	60.3
32	502.8	251.4	71.8
23	529.8	269.9	75.7
13	523.5	104.7	74.8
6	1266.2	253.2	180.9

3.1. Clustering Grouped Project Data

The first step in K-Means Clustering is determining the optimal number of clusters labelled with k.

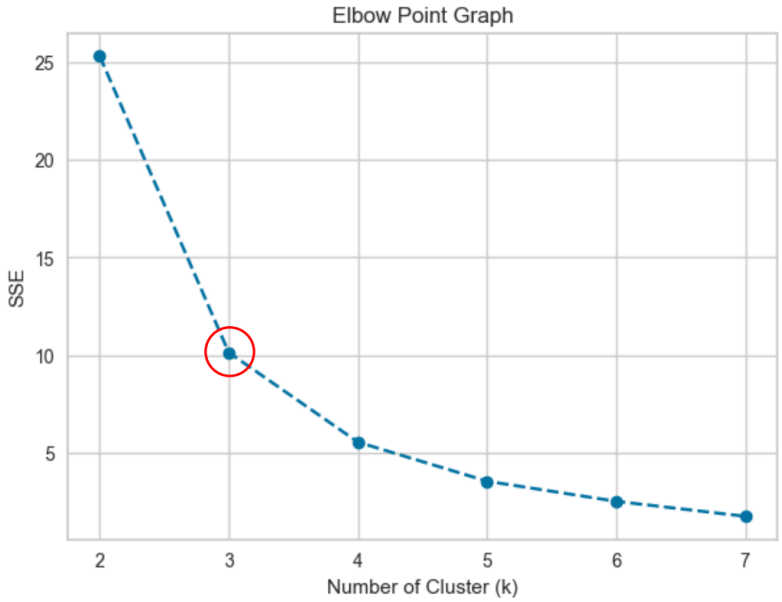

Figure 1. Elbow Graph with Elbow Point Highlighted in Red Circle

The methods used for determining the best number of clusters in this study are the Elbow Graph Method and Silhouette Score Analysis. Above is shown the Elbow Graph with the elbow point highlighted in a red circle. The elbow point is on k equals 3 which means the optimal number of clusters based on Elbow Criterion are 3 clusters/ groups.

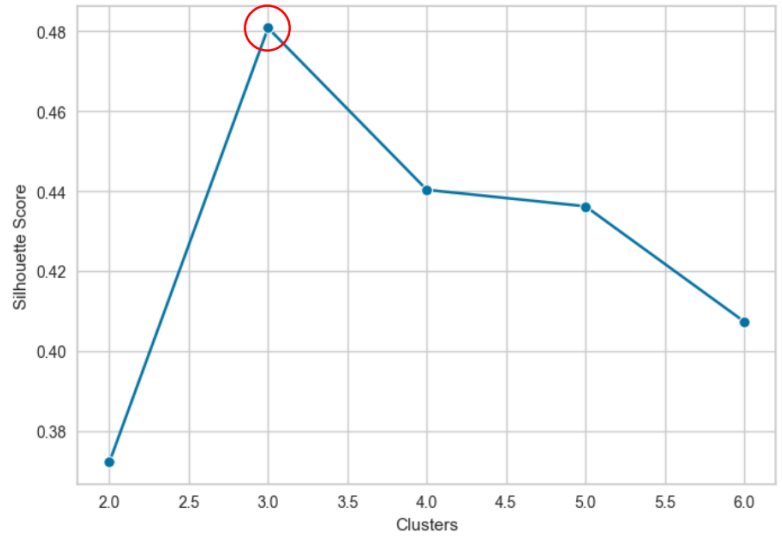

Figure 2. Silhouette Score Graph with Best Score Highlighted in Red Circle

Silhouette Scores are determined by finding the number of clusters with the highest score. Above, the Silhouette Score is plotted in a graph with the best score highlighted in a red circle. The best score is for k equals 3, which means that the optimal number of clusters is 3 clusters. The second-best score is for k equals 4, which means that the optimal number of clusters is 4 clusters.

Since both the Elbow Criterion and Silhouette Score pointed out that the optimal number of clusters is 3, the data will be clustered into 3 groups and 4 groups to be evaluated with the last step in this algorithm. The next and last step in the clustering algorithm is to evaluate the formed cluster using the Davies-Bouldin Index.

Table 2. Davies-Bouldin Index Score

Number of Cluster	Davies-Bouldin Index
3	0.34513578
4	0.47847030

The result of Davies-Bouldin Index calculations can be seen in Table 2. The result suggests that the optimal number of clusters for the project data is with k equal to 3.

Table 3. Grouped Projects Characteristics

Detail	Level 1	Level 2	Level 3
Average Work Days	14	30	47
Minimum Worker	1	1	3
Maximum Worker	2	4	5
Least Work Duration	7.5	43.5	523.5
Most Work Duration	239.5	529.8	1266.2
Median			
Number of Worker	1.5	2.5	4
Total Duration	123.5	286.65	894.85
Average			
Duration per Worker	82.2	157.3	179
Daily Work Duration	8.4	11.8	20.3
Estimated Range			
Work Duration	118.75	237.5	387.5

This concludes the clustering stage in this study. The characteristics of each level will be shown in Table 3. The clustering is then classified into levels of complexity, with level 1 as the lowest complexity and level 3 as the highest. Note that in the characteristics that are noted in Table 3, we can find an average daily work duration that is more than 10 hours. This data is interpreted as the number of overtime work that may be done outside work hours or work days by the workers.

3.2. Calculating Workload by Project Level of Complexity

Workload can be calculated by dividing the number of work hours done over a period of time by a worker divided by the standard work duration that is established by the company in the same period of time. The total duration in the data will be divided by the company work hours, which is 8 hours a day.

Table 4. Calculated FTE based on Level

Project Name	Total Duration	FTE
10	7.5	0.200
34	50.5	0.337
24	122.5	0.817
27	172.0	0.860
33	239.5	1.198
2	43.5	0.580
12	104.0	0.347
9	309.5	0.774
22	324.3	0.811
5	262.3	0.874
17	339.0	0.848
29	421.9	1.055
32	502.8	1.257
23	529.8	1.324
13	523.5	0.997
6	1266.2	2.026

The Calculated FTE is adjusted based on the average characteristics in the group of complexity levels. The classification that can be found in the results is underload for projects with an FTE score lower than 0.8, close to

normal for projects with FTE score between 0.8 and 0.99, normal workload for projects with FTE Score between 0.99 and 1.28, and overload for projects with FTE score higher than 1.28 [17].

3.3. Interface Building

The interface will be made so that the user can input a few details about the planned project. The details needed to give recommendations regarding the workforce needed are the Project name that is useful for identifying the current calculated project, the start and end date of the planned project, and the estimated level of complexity that the project has. The interface will then calculate the minimum number of workers needed and record the data in the same window of the interface. Figure 3 shows the example of input and output that can be produced by the built interface.

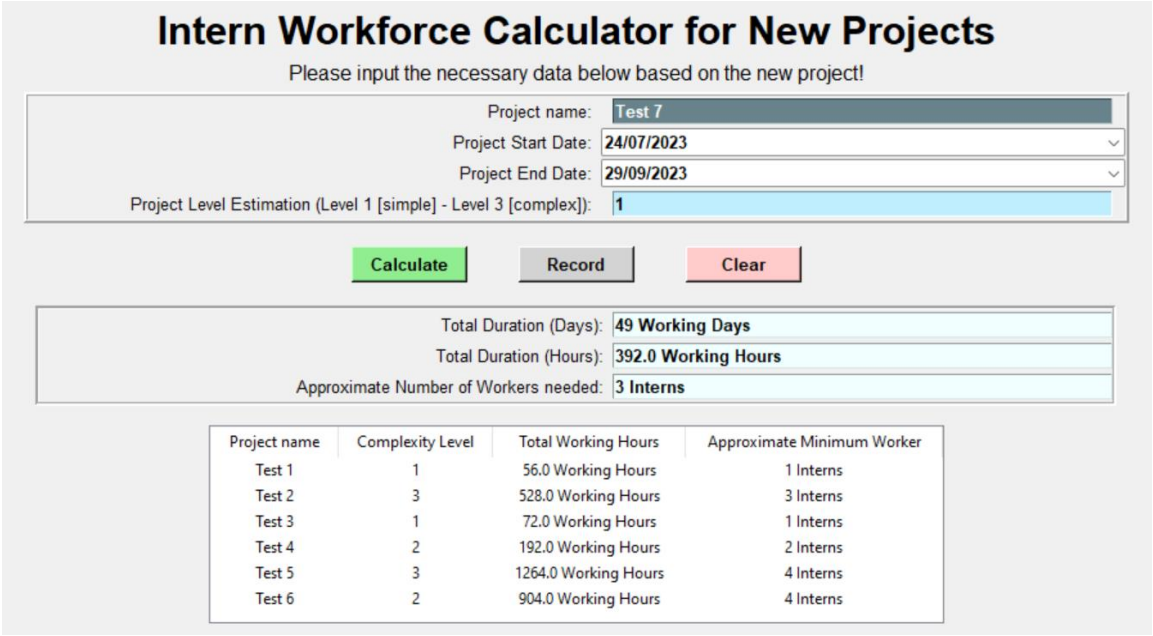

Figure 3. Calculator Interface

For example, in Figure 3, we can see that the test 7 project has a target to finish in 49 working days or approximately 392 working hours with 3 interns working on the project. This calculation is based on the project level estimation and project completion target that is inputted above.

4. CONCLUSION

Based on the result in the first step of the analysis, the optimal number of clusters is 3, which is gained by factoring the result of the Elbow Graph and Silhouette Score and then evaluated by the Davies-Bouldin Index. The resulting group consists of 3 types of projects, ranging from projects with low complexity to projects with high complexity. Each of the grouped data has its own characteristics, and their FTE is specified accordingly. This result is then successfully compacted into a simple interface that can be given planned project details such as the date of the start of the projects and the targeted duration of the project, as well as the estimated complexity of the project, which then will recommend whether the current workforce is enough to tackle the said project.

ACKNOWLEDGMENTS

This work was supported by the Universitas Prasetiya Mulya, Tangerang Selatan, Indonesia.

DAFTAR PUSTAKA

[1] Tan. Sue-Ann, "Covid-19 drove unprecedented drop of 196,400 in S'pore employment; services hardest-hit: MTI report," The Straits Times, 24-Nov-2021. [Online]. Available: https://www.straitstimes.com/singapore/jobs/covid19-drove-unprecedented-drop-of-196400-in-singapore-employment-serviceshardest. [Accessed: 08-Mar-2023]

[2] Choudhury. Prithwiraj, "Our work-from-anywhere future," Harvard Business Review, 07-Jul-2021. [Online]. Available: https://hbr.org/2020/11/our-work-from-anywhere-future. [Accessed: 08-Mar-2023].

[3] Ellul, A., Erel, I. and Rajan, U. (2020) "The COVID-19 pandemic crisis and Corporate Finance," The Review of Corporate Finance Studies, 9(3), pp. 421–429. Available at: https://doi.org/10.1093/rcfs/cfaa016 (Accessed: December 28, 2022).

[4] Wajiga, H.., Ndaghu, J. T. (2017) "Significance of Manpower Planning for Effective Utilization of Human Resources in an Organization: A Conceptual Approach," International Journal of Business and Management Invention, Vol. 6 (Issue 8), pp. 16 – 22. Available at: https://www.ijbmi.org/papers/Vol(6)8/Version-1/D0608011622.pdf.

[5] Purba, W., Tamba, S. and Saragih, J. (2018) "The effect of mining data k-means clustering toward students profile model drop out potential," Journal of Physics: Conference Series, 1007, p. 012049. Available at: https://doi.org/10.1088/1742-6596/1007/1/012049.

[6] T. Omar, A. Alzahrani, and M. Zohdy, "Clustering approach for analyzing the student's efficiency and performance based on Data," Journal of Data Analysis and Information Processing, vol. 08, no. 03, pp. 171–182, 2020. doi:10.4236/jdaip.2020.83010

[7] Chayo, P.W. and Sudarmana, L. (2021) "A Comparison of K-Means and Agglomerative Clustering for Users Segmentation based on Question Answerer Reputation in Brainly Platform," ELINVO (Electronics, Informatics, and Vocational Education), 6(2), pp. 166–173.

[8] Aldino, A.A., et al. (2021) "Implementation of K-means algorithm for clustering corn planting feasibility area in South Lampung Regency," Journal of Physics: Conference Series, 1751(1), p. 012038. Available at: https://doi.org/10.1088/1742-6596/1751/1/012038.

[9] Syakur, M.A. et al. (2018) "Integration K-means Clustering method and elbow method for identification of the Best Customer Profile Cluster," IOP Conference Series: Materials Science and Engineering, 336, p. 012017. Available at: https://doi.org/10.1088/1757-899x/336/1/012017.

[10] Hasugian, P. M., Sinaga, B., Manurung, J., & Al Hashim, S. A. (2021). Best cluster optimization with combination of K-means algorithm and elbow method towards rice production status determination. International Journal of Artificial Intelligence Research, 5(1). https://doi.org/10.29099/ijair.v6i1.232

[11] Umargono, E., Suseno, J. E., & Vincensius Gunawan, S. K. (2020). K-means clustering optimization using the elbow method and early centroid determination based on mean and median formula. Proceedings of the 2nd International Seminar on Science and Technology (ISSTEC 2019). https://doi.org/10.2991/assehr.k.201010.019

[12] M. Cui, "Introduction to the K-Means Clustering Algorithm Based on the Elbow Method," Accounting, Auditing and Finance, pp. 5–8, 2020. doi:10.23977/accaf.2020.010102

[13] Rasid Mamat, A. et al. (2018) "Silhouette index for determining optimal Kmeans clustering on images in different color models," International Journal of Engineering & Technology, 7(2.14), p. 105. Available at: https://doi.org/10.14419/ijet.v7i2.14.11464.

[14] K. R. Shahapure and C. Nicholas, “Cluster quality analysis using Silhouette score,” 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), 2020. doi:10.1109/dsaa49011.2020.00096

[15] G. Erda, C. Gunawan, and Z. Erda, “Grouping of poverty in Indonesia using K-means with silhouette coefficient,” Parameter: Journal of Statistics, vol. 3, no. 1, pp. 1–6, 2023. doi:10.22487/27765660.2023.v3.i1.16435

[16] Jumadi Dehotman Sitompul, B., Salim Sitompul, O. and Sihombing, P. (2019) “Enhancement clustering evaluation result of Davies-Bouldin index with determining initial centroid of K-means algorithm,” Journal of Physics: Conference Series, 1235(1), p. 012015. Available at: https://doi.org/10.1088/1742-6596/1235/1/012015.

[17] Y. Rachmuddin, D. S. Dewi, and R. S. Dewi, “Workload analysis using modified full time equivalent (M-FTE) and NASA-TLX methods to optimize engineer headcount in the Engineering Services Department,” IOP Conference Series: Materials Science and Engineering, vol. 1072, no. 1, p. 012036, 2021. doi:10.1088/1757-899x/1072/1/012036