The Eurasia Proceedings of Science, Technology, Engineering & Mathematics (EPSTEM)
ISSN: 2602-3199

The Eurasia Proceedings of Science, Technology, Engineering & Mathematics (EPSTEM), 2024

Volume 30, Pages 107-113

ICBAST 2024: International Conference on Basic Sciences and Technology

# Cluster Analysis of Sleep Health and Lifestyle Data Using Ward Algorithm and Euclidean Distance

**Mawar Idah Shonia**
Universitas Gadjah Mada

**Noorma Yulia Megawati**
Universitas Gadjah Mada

**Gunardi Gunardi**
Universitas Gadjah Mada

**Asrul Khasanah**
Universitas Gadjah Mada

**Abstract**: The intention of this study is to identify and assess groups based on their sleep quality and duration, physical activity levels, and stress levels. Next, we will investigate the relationship between sleep habits and stress levels. There were 374 respondents, with a total of 13 variables. The researchers utilized Ward's algorithm to identify groups and Euclidean distance to compare them. This study technique employs statistical computer tools, specifically R. This study's processes begin with data processing, which is followed by data standardization and clustering. There are four categories, namely (1) a group with an average sleep duration of 6 hours and a sleep quality scale worth 6 out of 10, but conducting physical activity less than 30 minutes per day, the stress level is high. (2) in a group with an average sleep duration of 6 hours and a sleep quality scale worth 6 out of 10, but doing physical activity for two hours each day, the stress level is very high. (3) in the group with an average sleep duration of 7 hours, a sleep quality scale of 8 out of 10, and 65 minutes of physical activity each day, the stress level is medium, (4) the group with an average sleep duration of 8 hours and a sleep quality rating of 9 out of 10 maintains a low stress level despite one hour of physical exercise. A dendrogram plot is used in data visualization to show how closely connected the data sets are. This study suggests that a person's sleep habits and daily physical activity have a major impact on their stress level, providing readers and the community with knowledge into how to improve overall health.

**Keywords:** Cluster, Sleep health, Lifestyle, Ward algoritm, Euclidean distance, Healthcare engineering

## Introduction

Sleep is a crucial health practice that affects all areas of well-being (Grandner, 2014). Sleep can also help the body recover from weariness and become more rejuvenated. Sleep difficulties can cause behavioral changes such as decreased endurance, loss of attention, easy exhaustion, sadness or stress, and interference with other daily tasks. Sleep length and quality are two aspects that influence whether a person's sleep requirements are met (Lemma et al., 2012). The factors that cause a person's stress level are classified into three categories: (1) biological factors, which include genetics, sleep patterns, diet, fatigue, and so on; (2) psychological factors, which include emotional factors, behavior, situations, and feelings; and (3) environmental factors, which include physical, social, and biotic activities. Stress levels can range from one to ten. Level one indicates that a person is not stressed and is free of pressure and worry. Level two denotes a little stress that does not interfere with regular activities. Level three stress refers to mild or low stress in which pressure is present but may be

regulated. Level four refers to mild stress that can still be managed. Level five is defined as medium-level stress, which causes pressure and interferes with daily tasks. Level six is moderately stressful and has an effect on mental health. Level seven indicates a high level of tension that significantly interferes with tasks. Level eight indicates extremely high stress, making it difficult to focus on regular activities. Levels nine and ten represent high and extraordinary levels of stress when you are unable to manage them and require intensive medical treatment. The 10 levels can be classified into two types of stress levels. Eustress is a stress condition in which the individual remains cheerful and enthusiastic about participating in activities, typically at levels one through five. And distress is defined as stress that has a negative effect, typically at levels six to ten.

A person's stress level can be reduced by a variety of means, one of which is participating in sports or other physical activities. Physical activity is an effective way to manage stress and become more organized (Scott, 2021). Physical activity can take several forms, including exercise, which involves moving the body in such a way that sweat is produced. The World Health Organization (WHO) defines physical activity as a body movement that requires energy expenditure (World Health Organization, 2020). As a result, there is a strong link between stress-causing sleep disturbances and stress-reducing physical activity. Researchers obtained secondary data from the Kaggle.com website in the form of a data set containing 13 variables. We will cluster the data and evaluate each group separately. Cluster analysis seeks to create multiple natural groupings, or clusters, of individuals. This is accomplished by categorizing "similar" individuals based on some relevant criteria. Once the groups have been identified, it is useful to display the data to describe each group using dendrograms in order to have a better understanding of the distinctions that exist between the formulated groupings. Cluster analysis is used in many sectors, including natural science, medical science, economics, marketing, and others. Cluster analysis is divided into two fundamental steps: (1) selecting the proximity measure of each object to determine the similarity of its values, so that the "closer" they are, the more homogeneous they are; and (2) selecting a group formation algorithm based on the proximity measure, the objects are assigned into groups so that the differences between groups become large and the observations within the groups become as close as possible. The study's criteria include four variables: sleep length, sleep quality, physical activity level, and stress level. Then, the link between sleep habits and stress levels will be investigated. Ward's technique is used to group and identify the data set, while the Euclidean distance is utilized to compare groups.

**Ward Algorithm**

To group the dataset Ward's approach is another algorithm for identifying partitions with a minimal number of squares. It begins with a large number of squares and reduces them before beginning with a small number of squares (using many clusters) and increasing them.

1. Begin with each point in a cluster by itself (sum of squares = 0).
2. Combine two clusters with the minimum total of squares to minimize merging costs.
3. Repeat merging until k clusters are reached. To identify it, Ward's method is utilized, and the Euclidean distance is used to compare groups.

If two items or groups, say P and Q, are combined, the distance between the resulting group P + Q and group R can be determined using the following distance function (Härdle & Simar, 2007):

$$d(R, P+Q) = \delta_1 d(R,P) + \delta_2 d(R,Q) + \delta_3 d(P,Q) + \delta_4 \mid d(R,P) - d(R,Q) \mid$$

$\delta_j$ is the weighting element that determines the different agglomeration algorithms.

$$\delta_1 = \frac{n_R + n_P}{n_R + n_P + n_Q}, \delta_2 = \frac{n_R + n_Q}{n_R + n_P + n_Q}, \delta_3 = -\frac{n_R}{n_R + n_P + n_Q}, \delta_4 = 0$$

Here, $n_P = \sum_{i=0}^{n} \boldsymbol{I}\ (x_i \in P)$ is the number of objects in group P. The values $n_Q$ and $n_R$ are defined analogously.

**Euclidean Distance**

The Euclidean distance is the smallest distance between two objects in N-dimensional space, also known as Euclidean space. It is a broad metric for determining the similarity of two data objects and is utilized in a variety of domains, including geometry, data mining, deep learning, and others. It also goes by the titles Euclidean

norm, Euclidean metric, L2 norm, L2 metric, and Pythagorean metric. The Euclidean distance is a special instance of the $L_r$ norm for $r \geq 1$, (Härdle & Simar, 2003) :

$$d_{ij} = \left\|x_i - x_j\right\|_r = \left\{\sum_{k=1}^{p} \left|x_{ik} - x_{jk}\right|^r\right\}^{\frac{1}{r}}$$

Where $x_{ik}$ represents the value of the k-th variable as measured on the i-th person.

## Method

The data used in this study is secondary data from Kaggle.com. The data comes in the form of "sleep health and lifestyle" data, with 374 respondents and 13 columns containing various variables linked to sleep and daily activities. This data covers parameters such as gender, age, occupation, sleep duration, sleep quality, physical activity level, stress level, BMI category, blood pressure, heart rate, number of steps per day, and presence or absence of sleep disorders, as well as the definition of each variable:

$X_1$ : Person ID
$X_2$ : Gender (Male or Female)
$X_3$ : Age (years)
$X_4$ : Occupation
$X_5$ : Sleep duration (hours per day)
$X_6$ : Quality of Sleep (scale 1-10)
$X_7$ : Physical activity level (minutes per day)
$X_8$ : Stress level (scale 1-10)
$X_9$ : BMI Category
$X_{10}$ : Blood Pressure (systolic or diastolic)
$X_{11}$ : Heart Rate (bpm)
$X_{12}$ : Daily Steps
$X_{13}$ : Sleep Disorder

Person ID is essentially an identifier for each individual. Each individual's BMI falls into one of three categories: underweight, normal, or overweight. Daily steps refers to the number of steps a person takes each day. Insomnia, sleep apnea, or the absence of any abnormalities. Individuals with problems sleeping suffer from insomnia (inappropriate sleep). Sleep apnea happens when a person's sleep pattern disturbance causes them to cease breathing while sleeping. The variables studied in this study are $X_5, X_6, X_7, and\ X_8$, which represent sleep length, sleep quality, physical activity, and a person's stress level. The respondents ranged in age from 27 to 59 years old. These variables share the same association, so they will be analyzed as a group. The set of data management techniques makes use of software applications, specifically R. In addition to cluster analysis, principal component analysis will be performed, with results provided in the form of plots. The following diagram depicts the flow from data loading to cluster analysis:

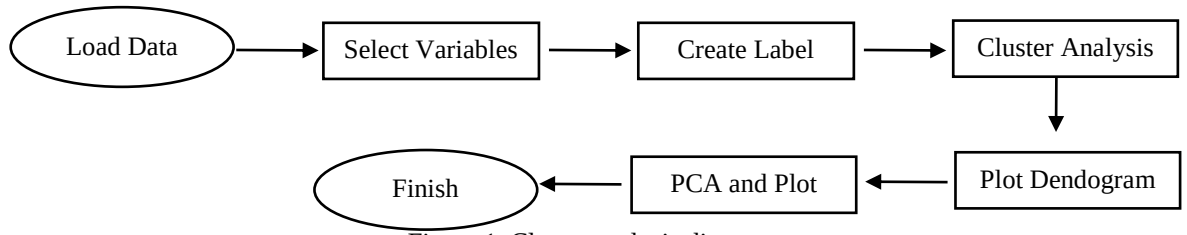


Figure 1. Cluster analysis diagram

The collected data was then put into the R software application using the load() command. Then, using the par() command, configure the chart layout to display two plots side by side in a single row. Next, use the hclust(dist(file[:]), "ward.D") command to execute the "Ward" cluster analysis, and the dist(file[:]) function to compute the Euclidean distance matrix between rows and columns. With the cutree(hc, 4) command, visualize data using a dendrogram plot of cluster analysis results divided into four clusters; the item names of the four clusters are also presented. The following phase is PCA (Principal Component Analysis), which tries to reduce

these variables into principal components capable of explaining the majority of the data variance, also known as data reduction, to make it easier to grasp the relationship structure between variables in the dataset.

## Results and Discussion

This is an overview of the "Sleep Health and Lifestyle" data that was processed:

Table 1. Person's typical sleep time

| Person ID | Age (years) | Sleep Duration (hours/day) | Quality of Sleep (1-10) | Physical Activity Level (minutes/day) | Stress Level (1-10) |
|---|---|---|---|---|---|
| 1 | 27 | 6.1 | 6 | 42 | 6 |
| 2 | 28 | 6.2 | 6 | 60 | 8 |
| 3 | 28 | 6.2 | 6 | 60 | 8 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 372 | 59 | 8.1 | 9 | 75 | 3 |
| 373 | 59 | 8.1 | 9 | 75 | 3 |
| 374 | 59 | 8.1 | 9 | 75 | 3 |

Table 1. Shows that a person's typical sleep time ranges between 6.1 and 8.1 hours per day. Physical activity levels vary from 42 to 75 minutes each day. Sleep quality indicators on a scale of 1 to 10 with interpretations, namely scale 1 is very bad, scale 2 is bad, scale 3 is not good, scale 4 is below average, scale 5 is moderate, scale 6 is quite good, scale 7 is good, scale 8 is very good, scale 9 is exceptional, and scale 10 is extraordinary without interruption and very sound. The ward approach is used to process this data, resulting in four clusters as follows:

Table 2. Cluster analysis results

| Cluster | Sleep Duration (hours/day) | Quality of Sleep (1-10) | Physical Activity Level (minutes/day) | Stress Level (1-10) |
|---|---|---|---|---|
| 1 | 6.249412 | 5.800000 | 38.41176 | 7.40000 |
| 2 | 7.294624 | 7.586022 | 64.73118 | 4.913978 |
| 3 | 8.243662 | 9.000000 | 55.56338 | 3.028169 |
| 4 | 6.065625 | 6.000000 | 90.00000 | 8.000000 |

Table 2. Shows the findings of the cluster analysis, which revealed four clusters with varying signals for each. The reasoning for each cluster is as follows:

(1) The group with an average sleep duration of 6.3 hours per day and a sleep quality scale score of 5.8 is close to 6, indicating that sleep is relatively sound with a reasonably effective sleep period despite slight interruptions. Physical activity averages 38 minutes per day, and the stress level on a scale of 7.4 is high enough to interfere with daily tasks.
(2) The group with an average sleep duration of 7.3 hours per day and a sleep quality scale worth 7.6 close to 8 indicates that the quality of sleep is very good, and with 65 minutes of physical activity per day, the stress level on a scale of 4.9 close to 5 remains moderate.
(3) The group with an average sleep length of 8.2 hours per day and a sleep quality rating of 9 indicates no sleep disturbance and is at an exceptional level. The average daily physical activity is two hours and 56 minutes, indicating a low stress level.
(4) The group with an average sleep duration of 6 hours per day and a sleep quality scale of 6 indicates that the quality of sleep is fairly good, and with 90 minutes of physical activity per day, the stress level is quite high.

To further comprehend, the dendrogram visualization is as follows:

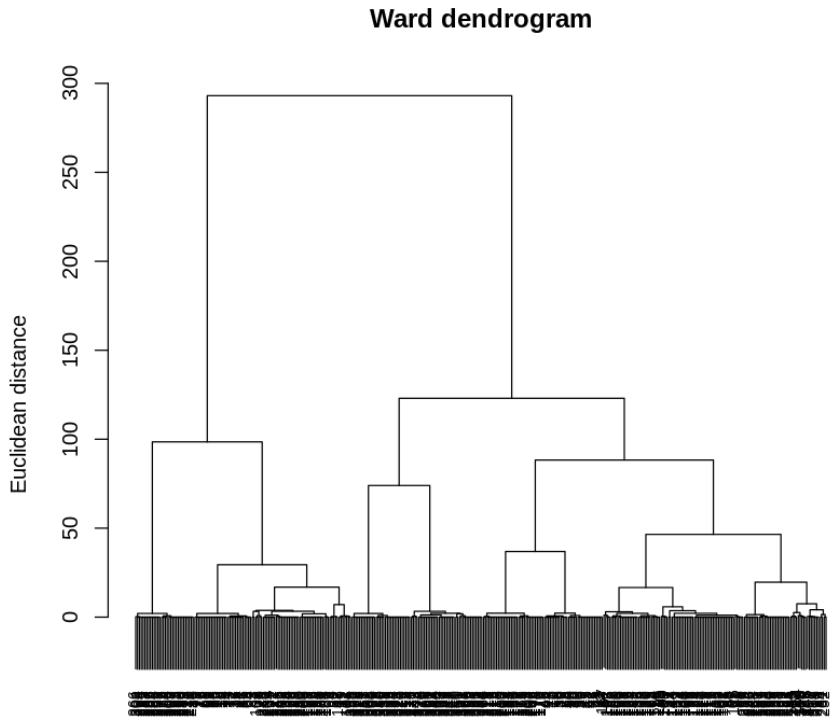


Figure 2. Shows that the six factor branches of the data can be more clearly interpreted and understood by applying PCA. Here's a PCA plot:

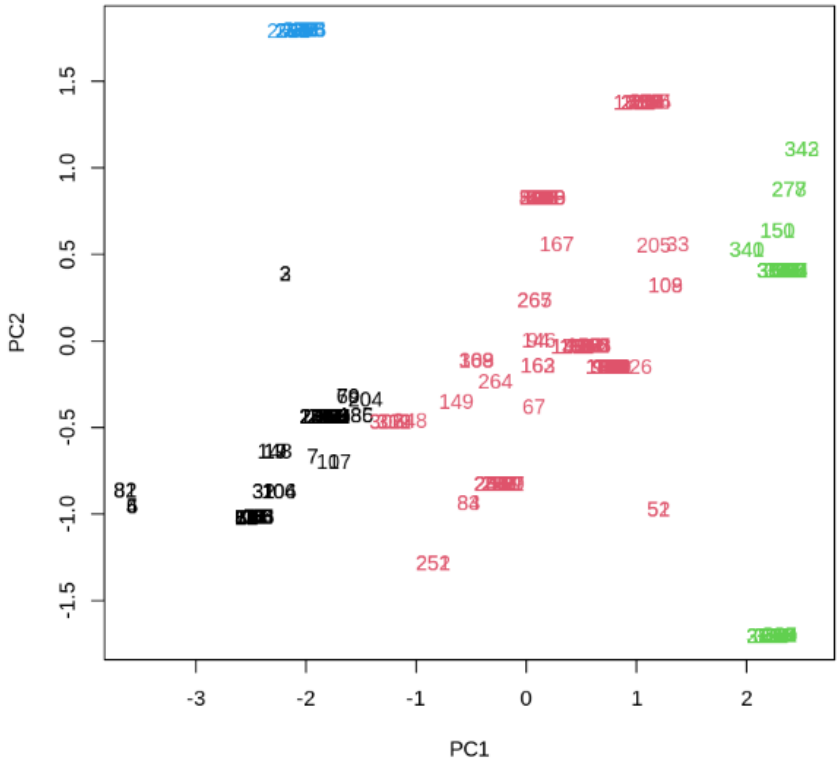


Figure 3. Color gruping

Figure 3. Shows four color groupings: blue, black, pink, and green. The number on the plot represents the individual's identity number. The blue color is in the positive area of the Y-axis, suggesting that the group is distinct from other groups and belongs to cluster 1. The dark hue appears in the negative region on PC1 and PC2, but slightly in the positive area on PC2. The pink color is more evenly distributed in the middle, ranging from -1 to 1 in both PC1 and PC2, indicating that it is in cluster 3. The green color appears in the positive

section of both PC1 and PC2, indicating cluster 4. The few points that are not part of the group suggest data variation.

## Conclusion

Based on the results of cluster analysis on "Sleep Health and Lifestyle" data using the ward algorithm and the Euclidean distance approach, it is possible to conclude that there are four groups from the cluster analysis process visualized using a dendrogram and then using PCA, which produces four different colors as follows:

(1) Blue (cluster 1) is a group with an average sleep length of 6.3 hours per day and a sleep quality scale score of 5.8, which is near to 6, indicating that sleep is pretty sound with a fairly effective sleep period despite slight interruptions. The average physical activity was 38 minutes per day, so the stress level on a scale of 7.4 was high enough to interfere with everyday tasks.
(2) Black (cluster 2) is a group with an average sleep duration of 7.3 hours per day and a sleep quality scale of 7.6 close to 8, indicating that the quality of sleep is very good. They also do 65 minutes of physical activity per day, and their stress level on a scale of 4.9 close to 5 remains medium.
(3) Pink (cluster 3) is a group with an average sleep length of 8.2 hours per day and a sleep quality rating of 9 out of 10, indicating no sleep disturbance. The average daily physical activity is two hours and 56 minutes, indicating a low stress level.
(4) Green (cluster 4) is a group with an average sleep duration of 6 hours per day and a sleep quality scale of 6, indicating that the quality of sleep is fairly good, and with 90 minutes of physical activity per day, the stress level is quite high.

There are certain spots that appear to be close together, indicating that the primary components are more comparable.

## Scientific Ethics Declaration

The authors declare that the scientific ethical and legal responsibility of this article published in EPSTEM Journal belongs to the authors.

## Acknowledgements or Notes

* The authors would like to express their sincere gratitude to Indonesia Endowment Fund for Education Agency (LPDP) under the Ministry of Finance of the Republic of Indonesia for supporting and funding the authors to attend this conference. Their assistance was crucial in facilitating the author's participation in the conference and encouraging collaboration on this project.

* This article was presented as an oral presentation at the International Conference on Basic Sciences and Technology (www.icbast.net) held in Antalya/Turkey on November 14-17, 2024.

## References

Bull, F. C., Al-Ansari, S. S., Biddle, S., Borodulin, K., Buman, M. P., Cardon, G., ... & Willumsen, J. F. (2020). World Health Organization 2020 guidelines on physical activity and sedentary behaviour. *British Journal of Sports Medicine*, *54*(24), 1451-1462

Grandner, M. A. (2014). Addressing sleep disturbances: an opportunity to prevent cardiometabolic disease?. *International Review of Psychiatry*, *26*(2), 155-176.

Härdle, W., & Simar, L. (2007). Conjoint measurement analysis. *Applied Multivariate Statistical Analysis*, 347-358.

Lemma, S., Gelaye, B., Berhane, Y., Worku, A., & Williams, M. A. (2012). Sleep quality and its psychological correlates among university students in Ethiopia: a cross-sectional study. *BMC Psychiatry*, *12*, 1-7.

Maronna, R. (2005). {Review of the book Wolfgang härdle and léopold simar: Applied multivariate statistical analysis, Springer, Berlin, New York, 2003]. *Statistical Papers*, *46*(1), 147.

Scott, E.(2021). *Top 10 stress management techniques for students." Verywell Mind*.

https://static1.squarespace.com/static/5dafb59e6bdcaa043c952968/t/5ef67f29c852723c9ddbe3ca/1593212713649/Lead+Health+-+Stress+Article.pdf

.

## Author Information

**Mawar Idah Shonia**
Universitas Gadjah Mada
Yogyakarya, Indonesia
Contact e-mail: *mawaridahshonia@mail.ugm.ac.id*

**Dr. Noorma Yulia Megawati, S.Si., M.Sc.**
Universitas Gadjah Mada
Yogyakarya, Indonesia

**Asrul Khasanah**
Universitas Gadjah Mada
Yogyakarya, Indonesia

**Dr.Drs. Gunardi, M.Si.**
Universitas Gadjah Mada
Yogyakarya, Indonesia

**To cite this article:**

Shonia, M. I., Megawat, N. Y., Gunardi, G., & Khasanah, A. (2024). Cluster analysis of sleep health and lifestyle data using Ward algorithm and Euclidean distance. *The Eurasia Proceedings of Science, Technology, Engineering & Mathematics (EPSTEM), 30,* 107-113.