STUDENT SEGMENTATION BASED ON COMMUTING BEHAVIOUR AND MENTAL HEALTH USING K-MEANS CLUSTERING

A. Poompavai
Assistant Professor, Department of Mathematics, Peri College of Arts and Science

Received:Sep 24, 2026 Accepted:Oct 21, 2026 Published:Oct 23, 2026

ABSTRACT

Understanding the impact of commuting on students' mental health and academic performance is crucial in higher education research. This study employs the K-means clustering technique to classify students based on their commuting behaviors and the resultant psychological and academic effects. An analyzed dataset includes variables concerning travel characteristics, stress levels, sleep patterns, and academic impact, rigorously preprocessed and standardized for accurate results. The study identifies the optimal number of clusters using the Elbow Method and Gap Statistic, revealing a three-cluster solution. The analysis results in three significant student segments: Moderate Stress Commuters, High Stress Commuters, and Low Stress/Short Distance Commuters. The High Stress Commuters exhibit notably elevated levels of anxiety, sleep deprivation, fatigue, and academic struggles, while the Low Stress group enjoys comparatively better mental health with minimal academic disruptions. The clustering's validity is validated through silhouette analysis, which points to a moderately acceptable cluster structure. Additionally, one-way ANOVA results underscore a statistically significant difference in stress levels across the clusters ($p < 0.001$), affirmed further by Tukey's post hoc test demonstrating significant pairwise differences among the groups. These findings stress the essential influence of commuting conditions on students' mental well-being and academic achievements. This research offers educational institutions critical insights for the development of tailored interventions, such as enhanced transportation support and mental health resources, aimed at improving student well-being and academic performance.

KEYWORDS

K-means clustering, student segmentation, commuting behavior, stress analysis, mental health, academic performance, ANOVA.

INTRODUCTION

In recent years, research in educational and public health domains has increasingly focused on the relationship between commuting behavior and student well-being. Frequent long commutes to educational institutions can lead to a range of negative outcomes for students, including physical fatigue, mental stress, and diminished academic performance. Factors contributing to these issues include long travel times, unreliable transportation options, and poor commuting conditions, all of which may result in sleep deprivation, anxiety, and a lack of motivation to engage academically. Mental health problems such as stress, anxiety, and exhaustion among students have escalated, highlighting the impact of various external determinants beyond academic pressures. As commuting has emerged as a significant yet often overlooked factor, it is critical to understand how daily travel challenges can disrupt sleep patterns and reduce the time available for studying and extracurricular pursuits, ultimately impairing the overall quality of life for students. Traditional statistical methods tend to analyze individual variables in isolation, potentially overlooking complex interrelationships. In contrast, advanced machine learning techniques—specifically K-means clustering—allow for the identification of underlying group structures by categorizing students with similar commuting experiences across multiple dimensions. This method facilitates a nuanced

understanding of diverse student experiences and helps pinpoint those at higher risk due to commuting-related stressors. The study aims to segment students according to their commuting behaviors and assess the consequent effects on mental health and academic performance, utilizing K-means clustering for analysis. By examining various factors such as travel conditions, stress levels, sleep deprivation, and academic challenges, the research seeks to discover meaningful patterns and classify students into distinct categories. Additionally, the study employs statistical validation methods, including silhouette analysis and analysis of variance (ANOVA), to evaluate the robustness and significance of the identified clusters. The anticipated findings are poised to inform policymakers and educational institutions, enabling them to implement targeted interventions. By recognizing students particularly vulnerable to the stresses of commuting, institutions can initiate supportive strategies such as enhancing transportation systems, offering flexible academic schedules, and providing mental health resources, ultimately promoting student well-being and fostering academic achievement.

REVIEW OF LITERATURE

Clustering has been a key technique within data analysis, stemming from early statistical classification methods established by foundational works such as Steinhaus (1956), and later developments formalized in K-means by MacQueen (1967) and Lloyd (1982). These contributions laid the groundwork for contemporary unsupervised learning algorithms by introducing concepts of partitioning data into groups based on similarity. The 1990s saw a rise in the significance of clustering methods, particularly in data mining, where distance measures like Euclidean distance became prevalent for categorizing similar observations. Crucial advancements in clustering validation were made through methodologies introduced by Rousseeuw (1987) with the Silhouette method and Tibshirani et al. (2001) with the Gap Statistic. These approaches continue to be essential for assessing cluster quality and determining optimal cluster counts, thereby enhancing the methodological underpinnings of clustering that bolster its reliability in empirical research.

In the current landscape, K-means clustering stands out as one of the most utilized unsupervised learning algorithms due to its straightforward nature and computational efficiency. A review by Ahmed et al. (2020) underscored its effectiveness in revealing underlying data structures, despite its limitations, such as sensitivity to its initial conditions and fixation on the predetermined number of clusters. Adamu et al. (2023) reinforced K-means' prominence across various fields, including healthcare, transportation, and education, owing to its scalability and ease of deployment.

The application of clustering techniques to student-related and behavioral datasets has surged, as highlighted in studies by Bahr (2010) and Kuh et al. (2000). These studies illustrated the efficacy of K-means clustering in forming student typologies based on behavioral metrics, thereby facilitating data-driven

decision-making in educational environments, marking a departure from conventional classification methodologies.

Recent research from 2023 to 2025 has extended K-means applications to more advanced contexts. Yadav and Sharma (2024) emphasized the role of clustering in managing substantial datasets, advocating K-means as a cornerstone technique in data mining and analytics. Kumar and Chambers (2024) raised concerns regarding clustering fidelity while advocating for robust validation techniques to yield reliable and interpretable outcomes. Clustering applications have also been explored extensively in transportation and behavioral research, with studies like Zhang et al. (2024) employing K-means to dissect travel patterns and effectively categorize distinct commuter groups, hence revealing insights into travel behavior. In a similar vein, Tian et al. (2023) examined educational settings with clustering to delineate elements influencing student productivity and satisfaction, underscoring its applicability in academic inquiry.

Furthermore, recent advancements have focused on enhancing the precision and adaptability of clustering techniques. Research by Xiao (2024) proposed refined K-means methodologies aimed at improving clustering consistency and reliability, particularly in the context of complex datasets filled with high dimensionality and noise, addressing significant limitations of classical K-means approaches.

METHODOLOGY

Data Collection and Pre-processing

The dataset utilized in this study encompasses various factors pertinent to students, specifically focusing on their commuting behavior, mental health indicators, and the resultant academic impacts. Major variables within this dataset include travel-related aspects, such as travel time and transportation reliability, alongside psychological elements like stress levels, anxiety, and sleep deprivation. The academic impact is also measured to evaluate the overall effect of these factors on student performance.

The data preprocessing stage undertaken in this analysis involved several key steps: firstly, missing values were addressed through listwise deletion to ensure a complete dataset. Following this, only the relevant numerical variables were selected for the purpose of clustering, which aids in identifying patterns within the data. Lastly, standardization of the variables was performed to ensure comparability across different measures, thereby enhancing the integrity and reliability of the analysis conducted.

The standardization was performed using Z-score normalization:

$$Z = \frac{X - \mu}{\sigma}$$

where X is the original value, μ is the mean, and σ is the standard deviation.

k-Means Clustering Algorithm

K-means clustering is an unsupervised machine learning method designed to group observations into K distinct clusters. The primary objective of this technique is to enhance the similarity of data points within

the same cluster while minimizing the similarity between different clusters. This partitioning allows for a more organized representation of data and is commonly utilized in various analytical applications.

The algorithm aims to minimize the Within-Cluster Sum of Squares (WCSS):

$$J = \sum_{k=1}^{K} \sum_{i \in Ck} \|x_i - \mu_k\|^2$$

where:

- K = number of clusters
- C_k= set of points in cluster k
- x_i = data point
- μ_k = centroid of cluster k

Algorithm Steps

The K-means algorithm follows these steps:

1. Initialize K cluster centroids randomly
2. Assign each data point to the nearest centroid using Euclidean distance:

$$d(X_i, \mu_k) = \sqrt{\sum_{j=1}^{k} (X_{ij} - \mu_{kj})^2}$$

3. Recompute cluster centroids:

$$\mu_k = \frac{1}{|Ck|} \sum_{i \in Ck} x_i$$

4. Repeat steps 2 and 3 until convergence (no change in cluster assignments)

Determination of Optimal Clusters

The optimal number of clusters was determined using:

Elbow Method

The total within-cluster variation is computed as:

$$WCSS = \sum_{k=1}^{K} \sum_{i \in Ck} (x_i - \mu_k)^2$$

The optimal K is identified where the reduction in WCSS begins to diminish (elbow point).

Gap Statistic

The Gap Statistic compares observed clustering with a reference null distribution:

$$Gap(K) = E_n^*[\log(W_k)] - \log(W_k)$$

where E_n^* is the expected value under a reference distribution.

Cluster Validation

The quality of clustering was evaluated using the Silhouette Coefficient:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

where:

- a(i) = average intra-cluster distance

- b(i) = minimum inter-cluster distance

The silhouette value ranges from -1 to 1, where higher values indicate better clustering.

Statistical Analysis (ANOVA)

To test whether stress levels differ significantly across clusters, a one-way ANOVA was performed.

The ANOVA model is given by:

$$Y_{ij}=\mu+\tau_i+\epsilon_{ij}$$

where:

- Yij = observed value
- μ = overall mean
- τ_i = effect of the i^{th} cluster
- ϵ_{ij} = random error term

The F-statistic is computed as:

$$F = \frac{\text{Between-grou variance}}{\text{Within-gro variance}}$$

Post Hoc Analysis

Tukey's Honestly Significant Difference (HSD) test was used to identify pairwise differences between clusters:

$$HSD = q.\sqrt{\frac{MS_{with}}{n}}$$

where:

- q = studentized range statistic
- MS_{within} = mean square within groups
- n = sample size per group

RESULTS AND DISCUSSION

Optimal Number of Clusters

The analysis of cluster optimization employed two methods: the Elbow Method and the Gap Statistic. The Elbow plot demonstrated a distinct bend at k = 3, indicating that a partition of the data into three clusters is most effective. This conclusion was validated by the Gap Statistic, which reinforced the suitability of utilizing three clusters for data segmentation.

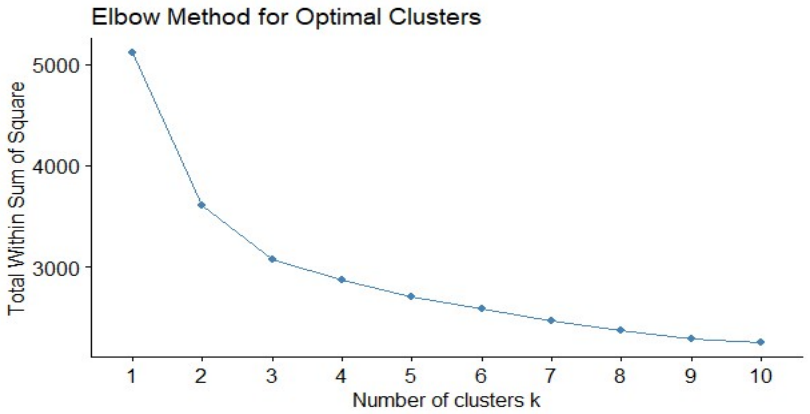


FIGURE 1. ELBOW METHOD FOR OPTIMAL CLUSTERS

The analysis of the within-cluster sum of squares (WSS) indicates a sharp decrease in values up to k = 3, after which the WSS stabilizes. This trend suggests that increasing the number of clusters beyond three does not lead to a meaningful enhancement in the model's performance, indicating that three clusters may be optimal.

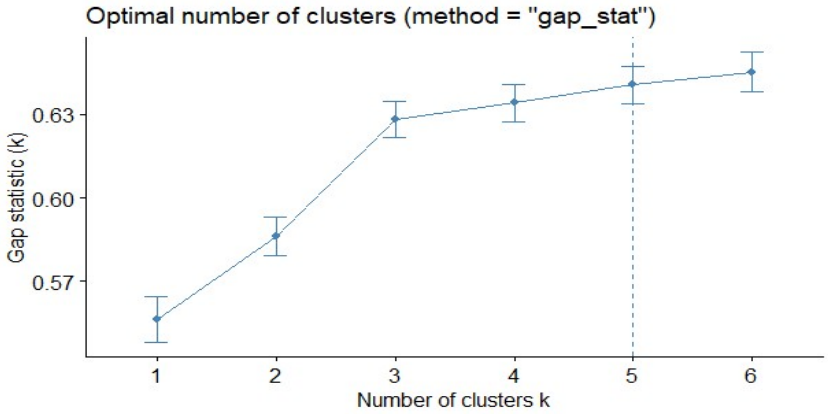


FIGURE 2: GAP STATISTIC FOR OPTIMAL CLUSTER SELECTION

The Gap Statistic method was utilized to determine the optimal number of clusters in a dataset, revealing that the gap value increases significantly until k = 3, after which it stabilizes. The presence of a vertical reference line at k = 3 signifies this point as the optimal number of clusters, as subsequent increases show minimal enhancement in clustering quality indicated by the overlapping error bars. Consequently, k = 3 was chosen for the K-means clustering analysis. Both the Elbow Method and the Gap Statistic lead to a consistent conclusion that three clusters offer the most suitable and interpretable segmentation of the data.

Cluster Formation using K-Means

K-means clustering was utilized with a specified number of clusters, k = 3, leading to the formation of three distinct groups of students. These groups were identified based on their commuting behaviors and

associated mental health factors, indicating a significant relationship between transportation methods and wellness in the student population.

TABLE 1: CLUSTER DISTRIBUTION

Cluster Label	Frequency
Moderate Stress Commuters	260
High Stress Commuters	151
Low Stress / Short Distance	101

The data indicates that within the student population, the predominant classification is Moderate Stress Commuters, who represent the largest segment. This is followed by those identified as High Stress Commuters, while the group characterized as Low Stress / Short Distance commuters is the smallest in number. This distribution highlights the varying levels of stress associated with commuting among students.

Cluster Characteristics

TABLE 2: CLUSTER-WISE MEAN VALUES (SELECTED VARIABLES)

Variable	Cluster 1 (Moderate)	Cluster 2 (High)	Cluster 3 (Low)
Travel Affect Academic	Moderate	High	Low
Difficulty	Moderate	High	Low
Late to College	Moderate	High	Low
Tiredness	Moderate	High	Low
Sleep Deprivation	Moderate	High	Low
Stress Level	Moderate	High	Low
Physical Health	Moderate	Poor	Better
Mental Fatigue	Moderate	High	Low
Anxiety	Moderate	High	Low
Motivation Decrease	Moderate	High	Low

Cluster analysis categorizes commuters into three distinct groups based on their stress levels and commuting experiences. Cluster 2, identified as High Stress Commuters, experiences the highest levels of stress, anxiety, fatigue, and academic challenges, likely due to long-distance travel coupled with poor transport conditions. In contrast, Cluster 1, the Moderate Stress Commuters, displays moderate levels of stress across all measured variables, reflecting the typical impact of commuting on this demographic. Finally, Cluster 3 represents Low Stress or Short Distance Commuters, characterized by the lowest stress levels, along with better physical and mental health, indicating that these commuters benefit from shorter travel times and minimal commuting burdens.

Cluster Visualization

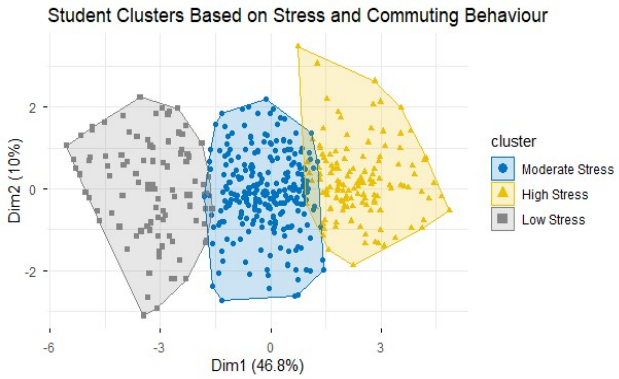


FIGURE 3: K-MEANS CLUSTERING OF STUDENTS BASED ON COMMUTING AND STRESS FACTORS

The figure presents the findings from a K-means clustering analysis that categorizes students according to their commuting habits and mental health indicators. The analysis yields the following key insights:

1. Cluster Separation: The identification of three distinct groups illustrates that students within each cluster display comparable commuting patterns and stress profiles, underscoring the effectiveness of the clustering process.

2. Cluster Density: One of the clusters exhibits a higher density, indicating that the students in this group possess very similar characteristics. In contrast, another cluster is more dispersed, reflecting a broader range of variability in stress levels or commuting behaviors among its members.

3. Inter-Cluster Distance: The significant distances observed between the clusters indicate substantial differences among the student groups. Specifically, one cluster corresponds to high-stress commuters, another to those with moderate stress, and the third to low-stress or short-distance commuters.

Overall, this clustering visualization enhances the understanding of the correlation between commuting behavior and the varying stress levels among students.

Silhouette Analysis

TABLE 3: SILHOUETTE SCORES

Cluster	Size	Average Silhouette Width
1	260	0.16
2	151	0.19
3	101	0.27

Cluster 3 demonstrates the strongest clustering performance with a silhouette score of 0.27, signifying a highly cohesive and well-separated grouping within the dataset. In contrast, Clusters 1 and 2 exhibit a moderate level of clustering quality, suggesting that while they provide some degree of separation, they do not achieve the same level of effectiveness as Cluster 3. Despite these results, the overall clustering structure

is deemed acceptable; however, there remains potential for improvement in order to enhance the clarity and distinctiveness of the clusters.

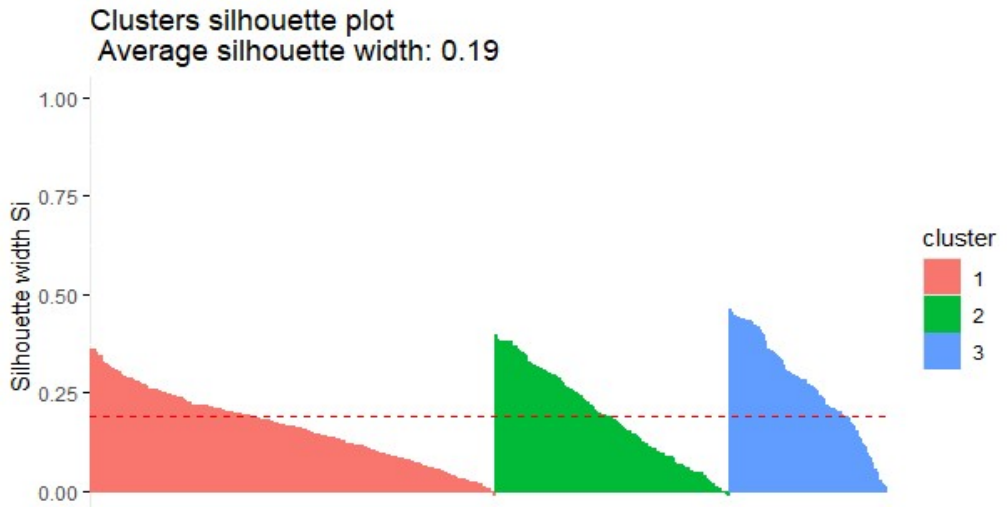


FIGURE 4: SILHOUETTE PLOT FOR CLUSTER VALIDATION

The silhouette plot is a diagnostic tool utilized to assess clustering quality by comparing how suitable each observation is within its assigned cluster relative to other clusters. The average silhouette width recorded is 0.19, reflecting a weak to moderate clustering structure, suggesting that while the clusters exhibit some degree of separation, they are not very distinct but remain interpretable. In detail, cluster interpretations reveal varying qualities among them.

Cluster 1 (Red) is categorized as having moderate quality, with silhouette values ranging from approximately 0 to 0.35. A significant number of observations are close to a value of 0, suggesting that many points are near the boundaries of this cluster, resulting in moderate overlap with other clusters. This indicates a less compact structure and implies mixed characteristics among the observations. **Cluster 2 (Green)** demonstrates weak to moderate quality, with silhouette values decreasing from around 0.40 to 0, indicating that several points hover near zero. This observation implies moderate cohesion within the cluster but signals poor separation from other clusters; consequently, it suggests that some students within this cluster may have been misclassified or possess similar features as students in other clusters.

Cluster 3 (Blue) is marked as the best-defined cluster, exhibiting the highest silhouette values, reaching up to approximately 0.45. The majority of points in this cluster are positioned above the average silhouette line (0.19), indicating that it is the most stable and well-segregated cluster and likely represents a distinct group, which may relate to low-stress students.

The red dashed line in the plot indicates the average silhouette width of 0.19. Clusters with silhouette values above this line signify good clustering, whereas those falling below it suggest weak clustering.

ANOVA Analysis

ANOVA was conducted to assess the significance of differences in stress levels across various clusters. This statistical test allows researchers to determine if the means of stress levels in different groups vary in a way that is not attributable to chance, thus providing insights into the relationship between stress and the defined clusters.

TABLE 4: ANOVA RESULTS

Source	Df	Sum Sq	Mean Sq	F value	p-value
Cluster	2	206.8	103.39	190	<2e-16 ***
Residual	509	276.9	0.54		

The p-value of less than 0.001 demonstrates a highly significant distinction in stress levels among the various clusters analyzed. This result validates the effectiveness of the clustering method in differentiating students based on their reported stress levels.

Post-Hoc Analysis (Tukey HSD)

TABLE 5: PAIRWISE CLUSTER COMPARISON

Comparison	Mean Difference	p-value
Cluster 2 – Cluster 1	+0.904	<0.001
Cluster 3 – Cluster 1	-0.927	<0.001
Cluster 3 – Cluster 2	-1.832	<0.001

All pairwise comparisons of stress levels among the clusters are statistically significant. Notably, Cluster 2 exhibits significantly higher stress levels compared to Cluster 1. Conversely, Cluster 3 demonstrates significantly lower stress levels when compared to both Cluster 1 and Cluster 2.

CONCLUSION

This study utilized K-means clustering to analyze student commuting behaviors and their correlation with mental health, uncovering significant patterns that illustrate the impact of travel on academic and psychological well-being. The study employed the Elbow Method and Gap Statistic to determine that three clusters formed the optimal number for stable and interpretable results. The analysis revealed the existence of three distinct groups among students: high stress commuters, moderate stress commuters, and low stress/short-distance commuters. Those in the high stress category reported heightened levels of fatigue, anxiety, sleep deprivation, and faced academic challenges, indicating that longer and more demanding commutes adversely affect mental health and academic performance. Conversely, students categorized as low stress exhibited improved physical and psychological well-being, likely attributable to shorter and less strenuous travel requirements. The silhouette analysis suggested a moderately effective clustering structure, with one cluster being well-defined while the others displayed some overlap. Nevertheless, these clusters effectively segmented the student population. Statistical validation through ANOVA indicated significant differences in stress across the groups ($p < 0.001$), with post-hoc analysis confirming that the clusters are

statistically distinguishable from each other, affirming the clustering's reliability. The overarching findings underscore the importance of commuting as a crucial factor influencing student well-being and academic achievement. The study advocates for measures to alleviate travel burdens, improve transportation reliability, and offer institutional support for students who commute long distances. Such interventions could notably enhance both student performance and mental health. Additionally, the study contributes significantly to existing research by showcasing that data-driven clustering approaches can pinpoint vulnerable student populations, thereby facilitating targeted interventions aimed at bolstering academic success and overall mental well-being. Looking ahead, future research avenues could involve exploring advanced clustering methodologies such as hierarchical clustering or DBSCAN, utilizing longitudinal data, and employing machine learning techniques to enhance the prediction of student risk levels more precisely.

FUTURE WORK

The paper suggests several future directions for research based on current findings and identified limitations. It highlights the importance of methodological enhancements and proposes the use of advanced algorithms, such as density-based clustering methods like DBSCAN or hierarchical clustering, to improve the handling of data overlaps revealed in silhouette analysis. Additionally, it recommends shifting towards predictive modeling, utilizing supervised machine learning techniques like Random Forests or Neural Networks to assess students' risk levels based on their commuting profiles at the beginning of each semester. The document also emphasizes the need for data expansion through longitudinal studies, which would allow researchers to track whether the stress associated with commuting accumulates over the years or if students manage to adapt over time. It encourages the inclusion of high-dimensional data in future research, exploring complex variables such as the specific modes of transport (e.g., bus versus train versus walking) and socio-economic factors that could impact commuting decisions.

Furthermore, the paper discusses the implication of findings for institutional interventions. It outlines a call for policy development within educational institutions to create tailored interventions, such as implementing flexible hybrid schedules for segments identified as experiencing "High Stress," and enhancing shuttle services. It also stresses the importance of resource allocation by advocating for dedicated mental health resources aimed specifically at long-distance commuters, which could help alleviate anxiety and fatigue related to their commuting experiences.

REFERENCES

1. MacQueen, J. (1967). *Some methods for classification and analysis of multivariate observations*. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1, 281–297.
2. Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53–65.
3. Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a dataset via the gap statistic. *Journal of the Royal Statistical Society: Series B, 63*(2), 411–423.
4. Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis* (5th ed.). Wiley.
5. Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.
6. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning* (2nd ed.). Springer.
7. World Health Organization (2022). *World mental health report: Transforming mental health for all*. WHO Press.
8. Evans, G. W. & Wener, R. E. (2006). Rail commuting duration and passenger stress. *Health Psychology, 25*(3), 408–412.
9. Stutzer, A. & Frey, B. S. (2008). Stress that doesn't pay: The commuting paradox. *Scandinavian Journal of Economics, 110*(2), 339–366.
10. Duffy, M. E. et al. (2019). Mental health and well-being among college students. *Journal of American College Health, 67*(5), 435–445.
11. Beiter, R. et al. (2015). The prevalence and correlates of depression, anxiety, and stress in a sample of college students. *Journal of Affective Disorders, 173*, 90–96.