DOI: 10.54254/2755-2721/2026.BA36238

K-Means Clustering on Physical Activity Volume Groups and Associations with Health-Related Indicators in Gym Members

Qianqian Chen

School of International Education, Chengdu University of Technology, Chengdu, China
chenqianqian@student.zy.cdut.edu.cn

Abstract. It remains unclear whether health benefits differ meaningfully across PA volume levels among individuals who already exercise. To examine this, PA volume data from 787 gym members were analyzed using K-Means clustering. The optimal cluster count was identified using within-cluster sum of squares (WCSS), Silhouette Coefficient, and Davies-Bouldin Index, which indicated four PA volume groups: Low, Moderate, High, and Very High Activity. After the clustering step, one-way ANCOVA was applied to examine differences in body fat percentage and resting heart rate across the groups, with age serving as a covariate and gender taken as a fixed factor. Body fat percentage varied significantly by cluster, and the pattern was non-linear. Resting heart rate, by contrast, did not show visible group differences. These findings, taken together, underscore the utility of coupling clustering with adjusted group comparisons to identify PA volume subgroups, and may in turn guide more individualized exercise assessment and activity recommendations.

Keywords: physical activity volume, K-Means clustering, ANCOVA, health-related indicators

1. Introduction

Lack of physical activity ranks among the top contributors to preventable death, and the idea that "physical activity is medicine" has become so widespread that some have questioned whether too much exercise brings its own risks [1, 2]. Warburton and Bredin's work pointed to a curvilinear relationship between physical activity and health benefits, and an earlier study found that excellent fitness did not carry additional survival benefits beyond a good fitness level [1, 3]. Put differently, extra activity does not always mean proportionally better health.

However, most existing work has paid less attention to individuals who already exercise regularly, including gym members. Among active people, further increases in activity may not bring proportional gains in health-related indicators. This raises the question of whether different PA volume groups set apart health-related indicators in an already-active population.

Physical activity behaviors vary widely across individuals. Clustering can sort people into subgroups that share similar PA volume patterns, which opens a way to check for health-related differences between groups [4]. On that basis, K-Means clustering was applied to identify PA volume groups, and ANCOVA was then used to compare body fat percentage and resting heart rate across those groups. This approach brings together unsupervised clustering and adjusted group

DOI: 10.54254/2755-2721/2026.BA36238

comparisons with the goal of offering evidence that could inform individualized exercise assessment and practical activity recommendations for active individuals.

2. Method

2.1. Data source and indicator selection

The research uses data collected from gym members, which is downloaded from Kaggle's open datasets [5]. It records 973 gym members' exercise behaviors and physiological characteristics.

This study excluded participants with session durations of less than one hour to maintain consistency in activity volume estimation, producing a final analytical sample of 787 that contained 376 women and 411 men. The study calculated Physical Activity (PA) Volume with unit MET-hour by using the formula:

$$PA\ Volume = \sum_{i=1}^{n} (MET_i \times Duration_i) \tag{1}$$

The MET_i represents the Metabolic Equivalent (MET). Standard MET values were assigned using the 2024 Adult Compendium of Physical Activities [6]. Activities not listed in the Compendium were assigned the MET value of the nearest comparable activity, based on the recommended classification approach.

Table 1 presents the variables included in the final analysis after data processing. Body fat percentage serves as a key marker of metabolic and cardiovascular risk [7]. The resting heart rate has been recognized as a population-level biomarker of cardiorespiratory fitness [8]. The two health-related indicators link physical activity to metabolism and cardiovascular health. Besides, age is selected as a covariate, and gender is a fixed factor. At the same time, PA volume is used as the sole input feature for clustering.

Table 1. Indicator selector of study population

Indicator	Role	Statistics	Range
Age (years)	Covariate	39.0 ± 12.20	18-59
Gender	Fixed factor	-	-
Body Fat (%)	Health-related indictor	24.38 ± 6.58	10.00-35.00
RHR (bpm)	Health-related indictor	62.16 ± 7.31	50.00-74.00
PA Volume (MET-h)	Cluster input	7.94 ± 3.20	2.30-14.93

2.2. Method introduction

2.2.1. K-Means clustering

K-Means clustering was selected because it is computationally efficient, simple, and well suited for exploratory clustering analysis [4]. K-Means will assign observations to the nearest centroid in every iteration. Then, the centroid is updated using the mean of all observations until cluster membership stabilizes, as measured by the within-cluster sum of squares (WCSS) [9]. In this study,

DOI: 10.54254/2755-2721/2026.BA36238

K-Means clustering was performed using a single feature, Physical Activity Volume (PA, MET-hour). Therefore, this study did not require data standardization before clustering.

2.2.2. Model evaluation metrics

To determine the optimal number of clusters, three complementary metrics were computed from K = 2 to K = 8. The metrics contain WCSS, Silhouette Coefficient, and Davies-Bouldin (DB) Index. Lkotun et al have proved that Silhouette is one of the most reliable clustering validity indicators [9]. Convergent evidence from all three metrics identified K = 4 as the optimal solution, balancing cluster compactness and separation without over-partitioning the data.

2.2.3. Statistical analysis

All analyses were carried out in IBM SPSS Statistics. The study used one-way ANCOVA to compare body fat percentage and resting heart rate among PA volume groups. Gender was included as a fixed factor, and age as a covariate. Before ANCOVA, the homogeneity of variance assumption for each dependent variable was assessed using Levene's test. Estimated marginal means (EMMs) accompanied by 95% confidence intervals were calculated to describe adjusted group differences. Statistical significance was assessed using F-statistics and p-values, while partial eta squared (ηp^2) served as the effect size. Pairwise cluster differences were tested with Bonferroni-adjusted post hoc comparisons. Two-sided tests were used throughout, with significance set at $p < 0.05$.

3. Results

3.1. Optimal cluster number determination

The clustering performance for K values ranging from 2 to 8 is summarized in Table 2. As K increased, the reduction in WCSS became noticeably smaller after K = 4, indicating an elbow at this point. Although the highest Silhouette Coefficient was observed at K = 2, the value for K = 4 was only 0.016 lower, while the Davies–Bouldin Index showed little difference between the two solutions. Considering the overall performance of the three metrics and the ability to distinguish different PA volume levels, K = 4 was selected for the subsequent analyses.

Table 2. Model evaluation metrics

K	WCSS	Silhouette Score	DB_Index
2	2371.10	0.6547	0.4243
3	849.77	0.6339	0.4749
4	426.79	0.6387	0.4406
5	293.10	0.5957	0.4862
6	213.69	0.6025	0.4891
7	156.31	0.5643	0.5155
8	112.31	0.5746	0.4895

DOI: 10.54254/2755-2721/2026.BA36238

3.2. Cluster indicators and activity volume group characterization

Table 3 summarizes the composition of the four activity clusters. The MA group was the largest, occupying 36.2% of the sample, while VHA was the smallest at 12.1%. Mean PA volume increased progressively from 3.20 MET-h in LA to 12.90 MET-h in VHA. It shows clear separation in physical activity volume group across clusters.

Table 3. Cluster composition and PA volume group of four clusters

Cluster	n	%	Mean	SD
LA	190	24.1	3.20	0.65
MA	285	36.2	7.78	0.77
HA	217	27.6	10.12	0.67
VHA	95	12.1	12.90	0.92

Figure 1 shows the distribution of PA volume across four clusters. Each cluster has a clear separation from the others, which indicates that participants were successfully classified into different PA volume groups.

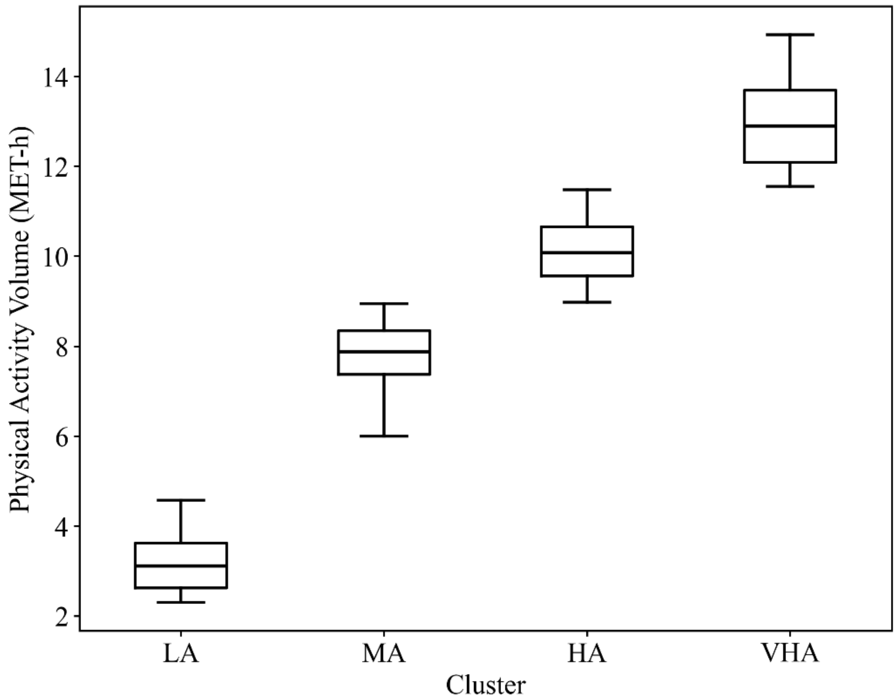
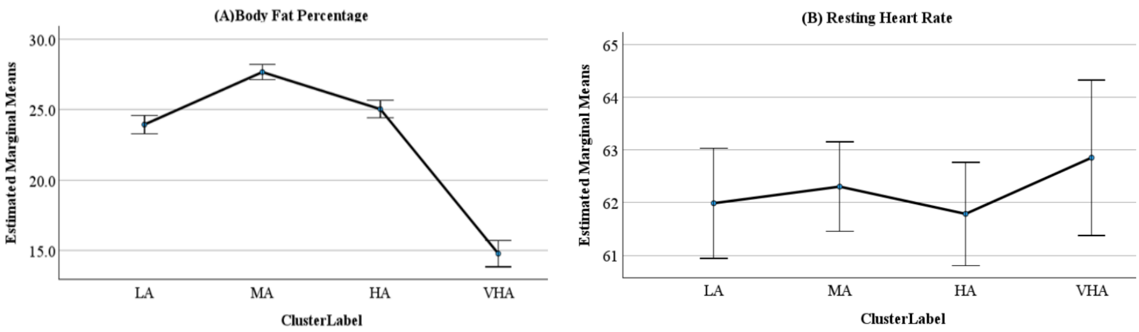


Figure 1. Boxplot of K-Means (photo/picture credit: original)

3.3. Health-related indicator comparison

To explore the associations between PA volume groups and health-related indicators, ANCOVA was used to compare body fat percentage and resting heart rate in four clusters. Age was included as a covariate, while gender was set as a fixed factor.

DOI: 10.54254/2755-2721/2026.BA36238



(a) Adjusted means of body fat percentage (b) Adjusted means of resting heart rate

Figure 2. Adjusted means of two health-related indicators (photo/picture credit: original)

As shown in Figure 2(a), adjusted body fat percentage demonstrated a nonlinear distribution between PA volume groups. The moderate activity (MA) cluster had the highest adjusted mean, followed by the high activity (HA) and low activity (LA) clusters. The very high activity cluster (VHA) was substantially lower than the other three groups. In contrast, resting heart rate showed limited variation among clusters in Figure 2(b). Although the VHA group had a slightly higher adjusted mean value than the other groups, the differences were not obvious.

Table 4 summarizes the ANCOVA results. Body fat percentage varied significantly across clusters. Bonferroni-adjusted comparisons showed that VHA was lower than all three other groups, and MA was higher than both LA and HA. For resting heart rate, neither the ANCOVA main effect nor the Bonferroni pairwise comparisons reached significance.

Table 4. ANCOVA results

Indicator	F (3,781)	p-value	ηp^2
Body Fat (%)	186.02	<0.001	0.418
RestingHeartRate(bpm)	0.61	0.608	0.002

3.4. Discussion

This study has identified four PA clusters among gym members. After adjusting for age and gender, significant differences in body fat percentage were observed among the four PA clusters, but not in resting heart rate.

The VHA group had the lowest adjusted body fat, while the MA group came in highest, surpassing even the LA group. This pattern lines up with Warburton and Bredin's observation that the relationship between exercise and health benefit is curvilinear, with limited additional gain beyond moderate volumes [3]. The body fat data also pushes that curvilinear relationship into the realm of body composition variation. Thomas et al. found that modest fat loss was largely attributable to low exercise doses combined with rising caloric intake. Yet, diet alone is unlikely to account for the whole picture [10]. Sleep debt, genetic makeup, and shifts in non-exercise activity can all throw off expected weight change [10]. On balance, the results are more likely to reflect a mix of confounders operating in concert rather than activity volume on their own.

The fact that resting heart rate did not differ meaningfully across PA volume groups hints that pushing activity volume higher does not move the needle on resting cardiovascular function, at least

DOI: 10.54254/2755-2721/2026.BA36238

not in an already-active population. This dovetails with evidence that moderate exercise intensity tends to offer the best return for cardiovascular health [11].

Several limitations need to be noted. One, clustering relied on a single feature, PA volume, which cannot capture the full picture of how people move. Later work could fold in additional activity dimensions like exercise intensity and frequency for richer profiles [1]. Two, only body fat percentage and resting heart rate were examined. Future research would benefit from adding measures such as blood pressure or blood lipids to broaden the physiological picture. Three, the cross-sectional design precludes causal inference, so longitudinal work that tracks how PA volume and health-related indicators shift together over time is needed to move beyond association. Four, all data were drawn from a single public Kaggle dataset, which may carry labeling noise or partial synthetic artifacts; pooling data across multiple independent sources would strengthen confidence in the findings.

4. Conclusion

Participants were divided into four PA volume groups through K-Means clustering, and ANCOVA was applied to compare body fat percentage and resting heart rate across these groups, with age and gender controlled throughout. Body fat percentage varied significantly by cluster, but resting heart rate showed no meaningful group differences after adjustment.

What these results point to is straightforward: doing more activity does not automatically bring more health benefit. Active people may do better by aiming for an appropriate volume rather than chasing ever-higher totals. Being able to identify activity-volume groups could also help gym members and their coaches with benchmark training outcomes against realistic reference points.

This study shows that coupling clustering with adjusted group comparisons can help unpack heterogeneity within an already-active population. It also lends support to individualized exercise assessment and tailored activity guidance. Moving forward, longitudinal designs, a broader set of health-related indicators, and data drawn from multiple independent sources would be needed to further probe how PA volume relates to health benefit.

References

[1] Franklin, B. A., Eijsvogels, T. M. H., Pandey, A., et al. (2022). Physical activity, cardiorespiratory fitness, and cardiovascular health: A clinical practice statement of the American Society for Preventive Cardiology Part II: Physical activity, cardiorespiratory fitness, minimum and goal intensities for exercise training, prescriptive methods, and special patient populations. *American Journal of Preventive Cardiology, 12*, Article 100425.

[2] Lee, I.-M., Shiroma, E. J., Lobelo, F., et al. (2012). Effect of physical inactivity on major non-communicable diseases worldwide: An analysis of burden of disease and life expectancy. *The Lancet, 380*(9838), 219–229.

[3] Warburton, D. E. R., & Bredin, S. S. D. (2017). Health benefits of physical activity: A systematic review of current systematic reviews. *Current Opinion in Cardiology, 32*(5), 541–556.

[4] Jain, A. K. (2008). Data clustering: 50 years beyond K-means. In *Machine Learning and Knowledge Discovery in Databases* (Lecture Notes in Computer Science, Vol. 5211, pp. 3–4). Springer.

[5] Khorasani, V. (2024). *Gym members exercise dataset* [Data set]. Kaggle. https://www.kaggle.com/datasets/valakhorasani/gym-members-exercise-dataset

[6] *Adult Compendium of Physical Activities*. (n.d.). Compendium of Physical Activities. Retrieved August 5, 2026, from https://pacompendium.com/adult-compendium/

[7] Huang, T., Feng, H., Xie, Z., et al. (2025). Effects of exercise on body fat percentage and cardiorespiratory fitness in sedentary adults: A systematic review and network meta-analysis. *Frontiers in Public Health, 13*, Article 1624562.

[8] Gonzales, T. I., Jeon, J. Y., Lindsay, T., et al. (2023). Resting heart rate is a population-level biomarker of cardiorespiratory fitness: The Fenland Study. *PLOS ONE, 18*(5), Article e0285272.

DOI: 10.54254/2755-2721/2026.BA36238

[9] Ikotun, A. M., Habyarimana, F., & Ezugwu, A. E. (2025). Benchmarking validity indices for evolutionary K-means clustering performance. *Scientific Reports, 15*(1), Article 8473.
[10] Thomas, D. M., Bouchard, C., Church, T., et al. (2012). Why do individuals not lose more weight from an exercise intervention at a defined dose? An energy balance analysis. *Obesity Reviews, 13*(10), 835–847.
[11] O'Keefe, J. H., O'Keefe, E. L., Eckert, R., et al. (2023). Training strategies to optimize cardiovascular durability and life expectancy. *Missouri Medicine, 120*(2), 155–162.