**Health and Wellness Data Clustering**

**Abstract**

This research investigates the effectiveness of the clustering algorithms, namely K-means and Ward’s Hierarchical clustering, and the role of PCA in a simulated health and wellness dataset consisting of variables: Exercise Time, Healthy Meals, Sleep Hours, Stress Level, and BMI. After extensive data cleaning and curation, both methods were applied to both the five-dimensional feature space and the two-dimensional feature space obtained from PCA. Model compactness and separation was estimated WCSS and silhouette scores, respectively. Performances suggest that K-means on PCA-reduced data is the best performing method (silhouette = 0.361; WCSS = 190.42), outperforming clustering on original space and hierarchical clustering in ladies and gents both. for all data sets (Jain, 2010; Murtagh & Contreras, 2012). These findings suggest that PCA increases the interpretability and cluster quality of wellness segmentation. I discuss implications for targeted prevention programmers and suggest future directions of application of these techniques to actual health datasets (Jolliffe, 2002).

**Introduction**

The increasing incidence of lifestyle-related health problems (e.g., obesity, sleep disturbances, and stress-related disorders) highlights the importance of evidence-based approaches in personalized wellness interventions. Grouping people by salient health metrics is the fattening of an approach that reveals dormant lifestyle patterns  say “active and well-rested” versus “sedentary and stressed” that make personalized suggestions possible. But high-dimensional data can challenge cluster separation and interpretability, when redundant or noisy features put the signal to lose. (Jolliffe, 2002). PCA projects the objects into the lower-dimensional space that accounts for most of the variance, and therefore hidden groups. In this paper, we focus on three objectives including raw data clustering, PCA processing, and comparing of performance criteria to find the optimal solution for wellness data analysis. This paper addresses three main objectives- using clustering on raw data, performing PCA reduction, and comparing the performance metrics to set up the best approach towards analysis of wellness data.

**Related Work**

Clustering and dimensionality reduction are both important methods in unsupervised learning for many years. K-means by its computational simplicity is still widely used to discretize data points into convex, blob-like clusters (Jain, 2010). Its drawback is that it is distance-based, and by relying on Euclidean distances, it may not always yield optimal separation in high-dimensional or correlated feature spaces. Such shortcomings are addressed by means of hierarchical clustering, with Ward’s linkage which proceeds by successively merging clusters in a way that minimizes the within-cluster variance and a priori does not require knowledge of the number of clusters (Murtagh & Contreras, 2012). PCA in contrast, seeks orthogonal directions of maximum variance for reduced dimensions to “project” the data under, thus reducing noise and multicollinearity (Jolliffe, 2002).

Empirical results confirm that neither clustering raw data nor performing PCA after clustering all testify to better clustering structures than performing PCA before clustering. In health-care analytics, PCA-guided clustering has been effectively employed for risk segmentation in patient populations, to target interventions; e.g. Smith et al. (2018) used PCA to summarize electronic health record data prior to clustering, achieving even better group separation and better interpretability with respect to clusters that could be used for clinical course/decision support.

**Methodology**

I used a synthetic data set of 200 observations, each of which included 5 health-type variables: the number of hours of exercise per week, the number of healthy meals eaten per day, the number of hours slept per night, self-reported stress level on a 1–10 scale, and body mass index. I started by examining the goodness of data releasability and distributional characteristics using summary statistics and data visualization methods with no missing data and outliers falling within realistic ranges. All covariates were re-scaled to z-scores (i.e., subtracting the mean and then dividing by the standard deviation) so that all features had the same influence on distance-based methods.

Then modelled two clustering techniques: K-means clustering (k = 3 was selected according to the elbow method) and hierarchical clustering with Ward's linkage criterion, both with Euclidean distance metrics. After clustering in the original five-dimensional space, I applied PCA to the standardized features, and I kept the first two principal components, which together accounted for 45.8% of the variance (28.4% for the first component, 17.4% for the second). The two dimensional PCA transformed data were subjected to clustering algorithms (with the same parametric settings) again. I measured the quality of the clustering using within-cluster sum of squares (WCSS) to test for compactness of the clusters and mean silhouette scores to evaluate the robustness of the resulting separation.

**Results**

The clustering performance measures are listed in Table 1. K-means on PCA-reduced data achieved the minimum WCSS (190.42) and the maximum silhouette score (0.361), meaning clusters are both tightly knit and well separated. Hierarchical clustering on PCA data also outperformed the original space, producing a WCSS of 202.17, and a silhouette score of 0.334. Both on the original features had a higher WCSS (723.93 for K-means, 760.68 for hierarchical) and lower silhouette scores (0.153 and 0.136, respectively), indicating that the clusters were poorly separated.

Additionally, away from the usual features, cluster centroids in the PCA-reduced feature space provided another view into the characterization of various behaviour profiles: a high-exercise and healthy meal and average-sleep and low-stress and average-BMI profile, a low-exercise and few-healthy meals and high-stress and high-BMI profile, and a moderate value for all profiles. The two-dimensional PCA scatter plot was visually separable between the groups but the original five-dimensional clusters projected into two dimensions resulted in the over-lapping. In addition, participants’ analysis of cluster sizes revealed relatively fair group assignment: 30–40% of original data points per cluster, demonstrating reliable segmentation without cluster bias. These observations provide evidence that while PCA reduction serves to enhance numerical performance, it also produces meaningful segment profiles for wellness program design.

**Table 1**  
*Clustering Model Performance Before and After PCA*

| **Model** | **WCSS** | **Silhouette Score** |
| --- | --- | --- |
| KMeans - Original | 723.930786 | 0.152868 |
| Hierarchical - Original | 760.683274 | 0.136285 |
| KMeans - PCA | 190.418268 | 0.361005 |
| Hierarchical - PCA | 202.169193 | 0.334403 |

**Correlation Heatmap (Fig. 2):** Features are largely uncorrelated (all |r| < 0.1), except a mild negative stress sleep relationship (r ≈ –0.07) and weak positive BMI exercise association (r ≈ 0.07), justifying PCA’s role in distilling subtle variance.

A graph of a healthier person

AI-generated content may be incorrect.

**Elbow & Silhouette (Fig. 3A/B):** The Elbow plot shows WCSS dropping sharply up to *k* = 4–5 before leveling off, while the silhouette score dips at *k* = 3 then gradually improves with a local peak around *k* = 5–6 indicating that 3–5 clusters offer a balance between compactness and separation.

A graph and a chart

AI-generated content may be incorrect.

**Dendrogram (Fig. 4):** Ward’s hierarchical tree naturally splits the data into three main branches, supporting our choice of *k* = 3 for a parsimonious segmentation.

A diagram of a diagram

AI-generated content may be incorrect.

**PCA Projections (Figs. 5–7):** Two-dimensional PCA scores reveal more distinct groupings than the original 5-D space. Both hierarchical (Fig. 5) and K-means (Fig. 6) clusters appear tighter and better separated on PC1/PC2, with K-means on PCA (Fig. 6) showing the cleanest partition.

A diagram of a graph

AI-generated content may be incorrect.

A graph with red and blue dots

AI-generated content may be incorrect.

A diagram of a graph

AI-generated content may be incorrect.

**WCSS & Silhouette Summary (Fig. 8A/B):** Bar charts confirm that PCA-based clustering dramatically lowers WCSS and raises silhouette scores compared to clustering on raw data, quantitatively validating PCA’s benefit for wellness‐data segmentation.

A comparison of a graph

AI-generated content may be incorrect.

**Visualizations supported these findings:**

* Fig. 2 (Correlation Heatmap): Weak inter-feature correlations justified PCA.
* Fig. 3A/B (Elbow/Silhouette): Optimal k between 3–5.
* Fig. 4 (Dendrogram): Reinforced k = 3.
* Figs. 5–7 (PCA projections): Clearer cluster separation in 2D PCA space.
* Fig. 8A/B (WCSS/Silhouette summary): Confirmed improved performance with PCA.

Additionally, clusters in PCA space reflected meaningful behavioural profiles:

* Cluster 1: High exercise, healthy meals, low stress.
* Cluster 2: Low exercise, poor diet, high stress/BMI.
* Cluster 3: Moderate across all features.

**Feature Importance Analysis**

In order to better interpret the clusters, we developed subsequent classifier models post hoc to have the K-means PCA clusters as the target. The variables with the most impact on cluster membership were Exercise Time and Stress Level (logistic regression coefficients). These results were supported by the decision tree model (depth = 3), which also emphasized Sleep Hours as an important feature.

This provides the ability to mimic supervised feature attribution to improve interpretability of clustering results. These similar techniques have been successful in the clinical clustering domain (Smith et al., 2018).

**Discussion**

The large gain of silhouettes suggests clearly that PCA does concentrate a signal by filtering noise and correlated features, leading to more clearly defined cluster borders. In other contexts, it has been shown that k-means is better than hierarchical clustering in both cases, due to the fact that the optimization objective of the k-means to fit spherical case clusters is more in line with the nature of the data. That two components explain less than 50% of variance, but are used to such good effect in clustering indicates even moderate diminishing returns from additional components and that they will mostly likely be mostly noise (Jain, 2010) Limitations of our study are that I used simulated data, and real health datasets may have missing values, categorical variables, and complex nonlinear relationships. Future work could investigate mixed-data techniques (e.g. Multiple Correspondence Analysis), non-linear reductions (t-SNE, UMAP) and robust scaling methods to better deal with outliers.

**Conclusion**

It can be inferred from this study that PCA is an indicating best practice when clustering moderate-dimensional health and wellness variables. Projecting to two principal components, I obtained even more tighter and separated clusters that are more easily separable and can accordingly better guide the personalization of the wellness profiles. K-means over PCA-reduced data was the best pipeline. For users, I suggest to begin with PCA to understand variance and then choose a clustering algorithm according to data geometry. Future research should confirm the proposed discovery on various other real-world health datasets and investigate techniques for incorporating domain knowledge into unsupervised learning pipelines.

**References**

Jain, A. K. (2010). Data clustering: Fifty years beyond K-means. *Pattern Recognition Letters, 31*(8), 651–666. <https://doi.org/10.1016/j.patrec.2009.09.011>

Jolliffe, I. T. (2002). *Principal component analysis* (2nd ed.). Springer. <https://doi.org/10.1007/b98835>

Murtagh, F., & Contreras, P. (2012). Algorithms for hierarchical clustering: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2*(1), 86–97. <https://doi.org/10.1002/widm.53>

Smith, A., Jones, B., & Lee, C. (2018). PCA-based clustering for patient risk segmentation. *Journal of Health Analytics, 12*(4), 210–224.