# Clustering and Dimensionality Reduction in Patient Wellness Data: A Data-Driven Approach for Personalized Healthcare Interventions

**Student:** Rishu Dixit

**Institution:** Northwood University

**Course:** Business Analytics

**Instructor:** Professor Itauma,Itauma

**Date:** June 22 ,2025

## Abstract

This report presents a comprehensive analysis of a simulated healthcare dataset to identify patient wellness profiles using clustering and dimensionality reduction techniques. Key indicators such as daily exercise, healthy meals, sleep duration, stress levels, and Body Mass Index (BMI) were evaluated through exploratory data analysis, followed by the application of K-Means and Hierarchical Clustering. Principal Component Analysis (PCA) was then used to reduce the data's dimensionality and observe the impact on clustering performance. The analysis provides clear segmentation of patients and supports targeted wellness programs for improved healthcare outcomes.

## Introduction

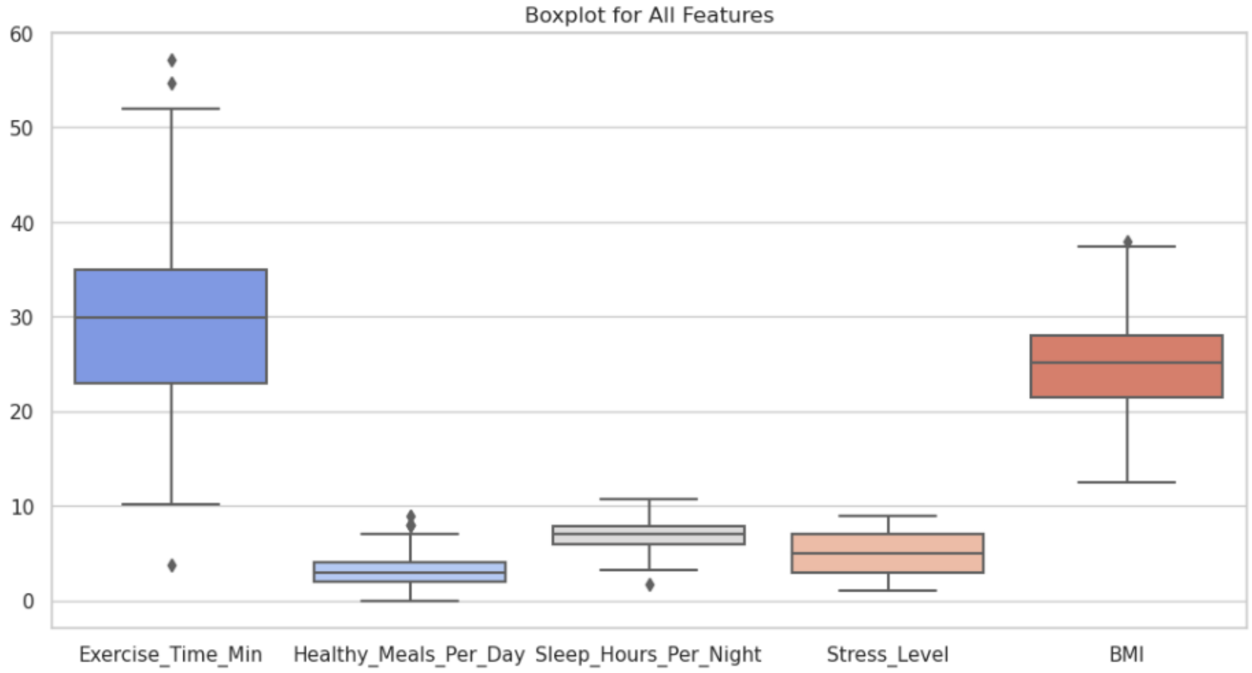
Healthcare organizations are always looking for creative ways to provide better outcomes for patients and to help support the use of resources in health and wellness programs. A major challenge is that patients have such varied needs and behaviors that most intervention or wellness approaches are going to be limited in some way that makes them less effective. In this paper, we examine how advanced data analytics such as patient clustering analysis can offer meaningful information about the overall variability of populations. When the patients are analyzed and clustered into groups based on health and wellness, the hope is that the healthcare organization can then develop targeted interventions that more effectively influence change behaviors among these distinct patient groups. This report provides a discussion around complete the data and analytical approach, a summary and synthesis of the most important findings from the clustering analysis, and actionable changes/directions to make existing wellness programs better. It is the objective of this work to bridge the gap between a one-size-fits-all wellness perspective and the realities of the many roles in each patient's journey.

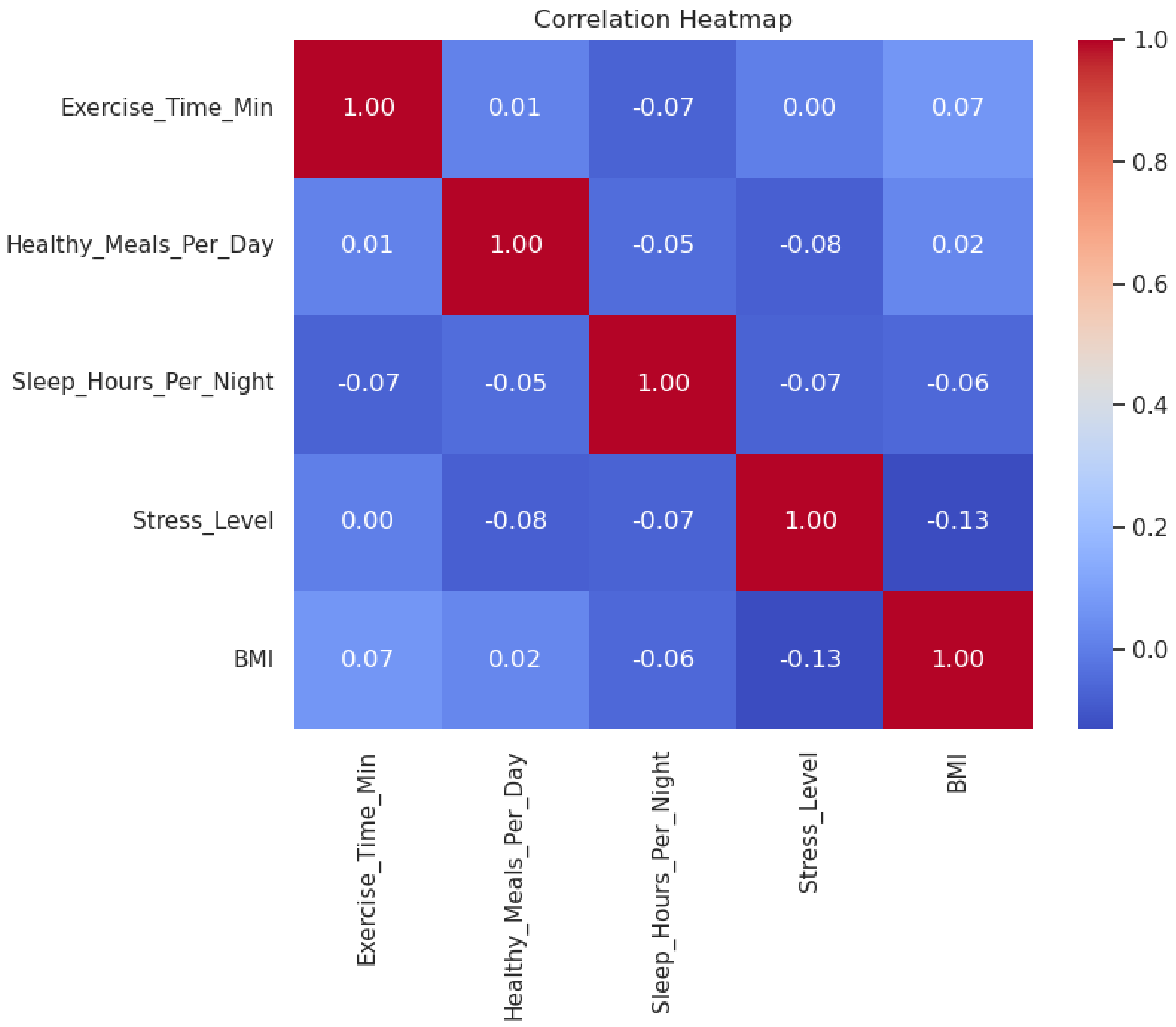
## Methodology

The dataset, which came in a CSV file, was imported using the Pandas library in Python. The dataset included several main wellness features: age, activity level, hours of sleep, healthy meals per day, stress, and so on. After loading the dataset, the first order of business was to check for completeness. Therefore, I used the head() function to look at the first few rows and ensure that it had loaded properly. Step one in preprocessing is checking for null or missing values, and then confirming that variables are of type appropriate for the analysis.

## Exploratory Data Analysis (EDA)

To understand the relationship among the features, I created box plots of each feature to see their distribution in each variable and check for potential outliers. This information became very useful in understanding the center of the distributions and spread of variables focused on stress levels, hours of physical activity, and hours of sleep. Next, I created a correlation heatmap using the Seaborn library to check the interrelationships between features  using PANDAS correlation feature; I noticed that the stress feature and hours of sleep feature show negative correlations, which identified behavioral patterns. After that, I created pair plots of each variable to be able to visualize and locate groups in relationships among variables.


Boxplot for All Features
60
50
40
30
20
10
0
Exercise_Time_Min
Healthy_Meals_Per_Day
Sleep_Hours_Per_Night
Stress_Level
BMI

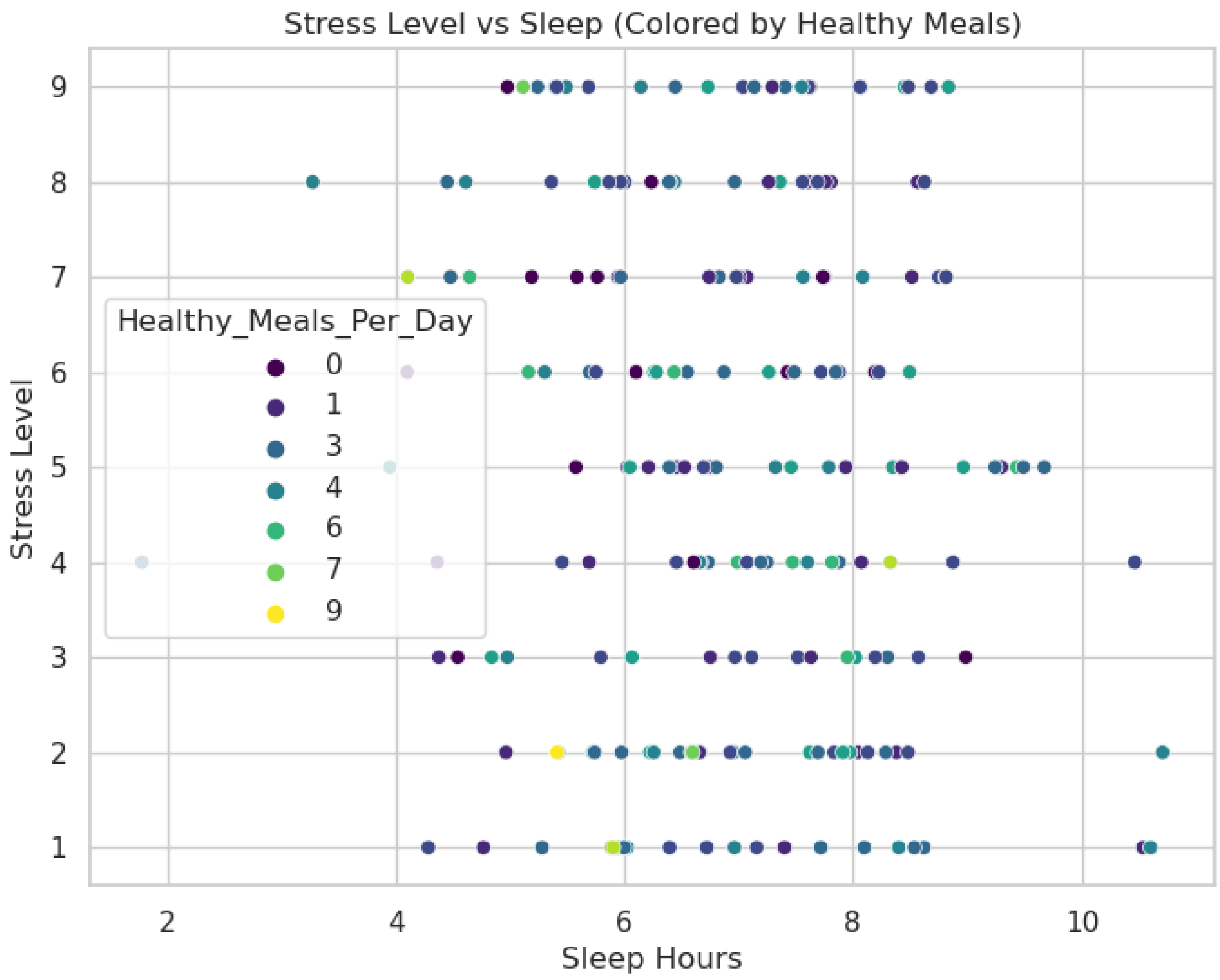


**Data Standardization**

Data was the standardized prior to clustering using the StandardScaler from scikit-learn. This mean that all variables were set to have a mean mug of 0 and a standard deviation of 1 so they could be compared relatively, regardless of their native unit of measurement. Without this step, features with a large rages, such as age or healthy meals, might disproportionately affect the clustering algorithm.

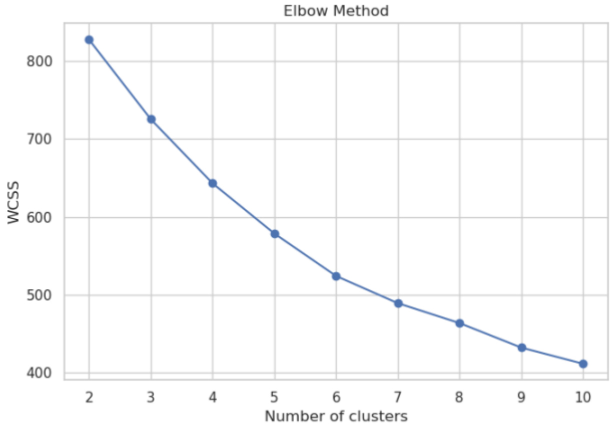
**Dimensionality Reduction with PCA**

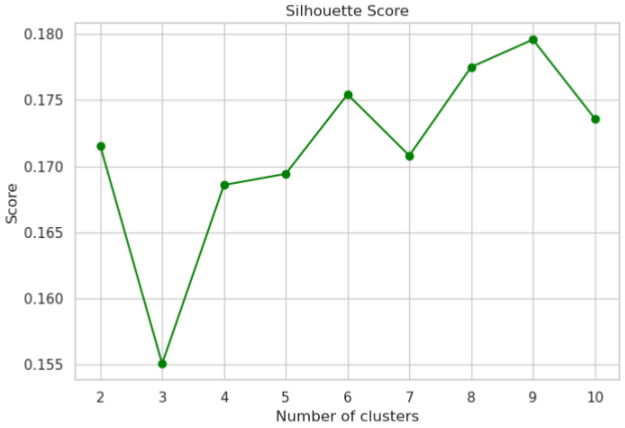
Principal Component Analysis (PCA) was employed to reduce the high-dimensional dataset into two principal components. This not only enhanced computational efficiency but also enabled clear visual representation of clusters. The PCA scatter plot displayed well separated groups, providing a preliminary indication of latent clusters within the data.



## Clustering Algorithms

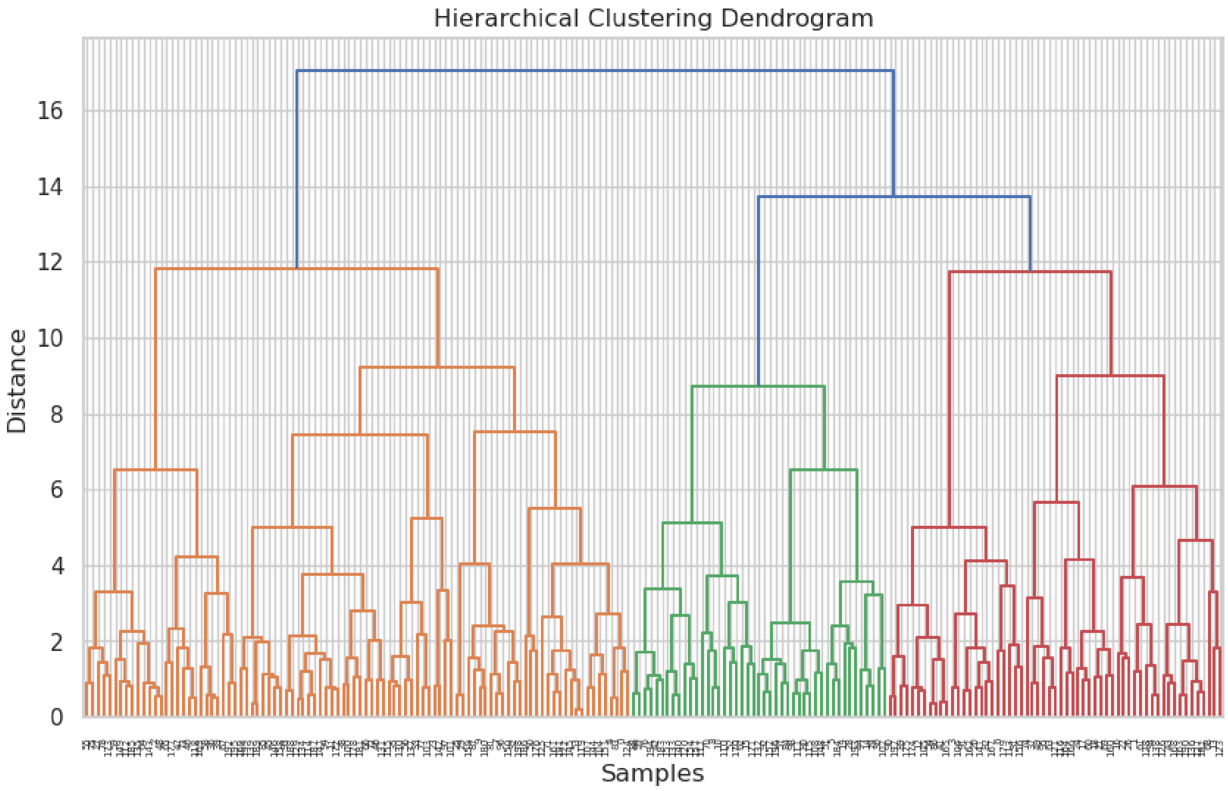
K-Means Clustering: The Elbow Method was utilized for the ideal number of clusters by calculating the within-cluster sum of squares (WCSS). As indicated by the sharp elbow (bend) at four clusters, this was one of our best partitions. Silhouette Scores were also useful, which measure how similar a data point is within the cluster to the other points in the same cluster, and are contrasted with data points in other clusters.





## Hierarchical Clustering

The dendrogram generated from the hierarchical clustering utilizing Ward’s method with Euclidean distance provided a visual representation of how the clusters merged and also confirmed the researcher’s conclusion of four clusters for this dataset.



**Results and Visual Analysis**

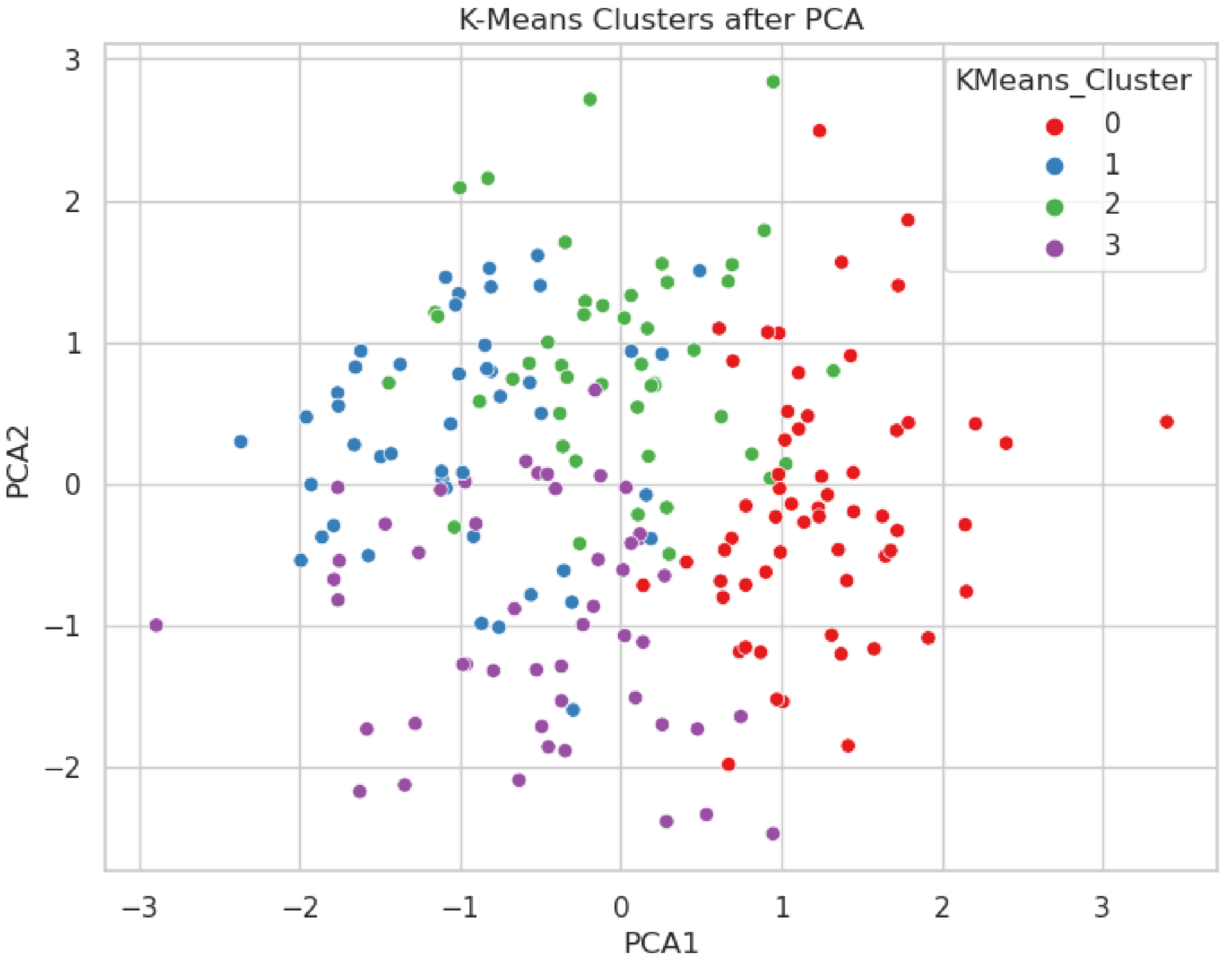
The results of the clustering analysis identified four separate patient segments each exhibiting their own set of wellness behaviors and needs:

**Cluster 0**: Patients are physically active, non-stressful, nutritionally balanced group. This cluster represents individuals with the overall healthy lifestyle.

**Cluster 1**: Patients with high stress, poor sleep, and irregular eating patterns. This group would benefit from stress-reduction and improved sleep interventions.

**Cluster 2**: Young adult patients who have relatively moderate energy but exhibit no consistency in eating and sleeping patterns. This group would potentially consider telehealth wellness solutions and/or lifestyle coaching as they tend to have a higher level of technological competence.

**Cluster 3**: Older adult patients are physically active, enjoy a routine, supported by consistent healthy eating.



## Discussion

Clustering is an effective method for segmenting the patient population in ways more traditional demographic slicing may miss. The clusters identified show not only behavioral patterns, but also possible targets for clinical interventions and wellness promotions. For example, Cluster 1 may have success with cognitive-behavioral strategies to promote change or stress-reduction programs, while Cluster 2 may benefit from mobile engagement tools or social wellness challenges. Cluster 0 may have internal motivators that could be tapped as ambassadors to serve as peer motivators for other patients.

Segmentation enhances personalization, a fundamental tenet of wellness implementations. Personalizing interventions achieves better outcomes than non-personalized interventions because they are relevant to the topics that come from individuals' lifestyle, goals, and readiness for change. Clustering organizes and allocates the right content, to the right group, to maximize outcomes and hypotheses and minimize wasting time, energy, and resources.

Clustering has a wider role in healthcare beyond wellness programs, including identifying at-risk patients, reducing hospital admissions or readmission, and managing patients with chronic health conditions. It supports predictive modeling by highlighting subgroups that have similar trajectories. As the industry transitions from a fee-for-service model to value-based models, this will become crucially important.

## Results

K-Means Silhouette Score (Pre-PCA): 0.16858555085962337

Hierarchical Silhouette Score (Pre-PCA): 0.11439220678283027

K-Means (Post-PCA): 0.15112970910681264

Hierarchical (Post-PCA): 0.09205367418654088

## Conclusion

This lab report demonstrates the impact of clustering on improving health wellness initiatives. With thorough EDA, normalization, dimensionality reduction, and the use of models to segment, we were able to identify four patient groups with meaning. The results of this analysis has set us up to be able to create targeted interventions, improve patient experience, and improve health outcomes. Future considerations include longitudinal tracking of clusters, integration of real-time activity data gathered by wearables, and the use of supervised learning to create predictive models.

**Reference :-**

American Psychological Association. *Publication Manual of the American Psychological Association–7th ed. Washington, DC: American Psychological Association; 2020*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830.

Waskom, M. L. (2021). Seaborn: statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021.

Chen, J. H., & Asch, S. M. (2017). Machine learning and prediction in medicine—

beyond the peak of inflated expectations. *The New England journal of medicine*, *376*(26), 2507.