Max Darling

MGT 665

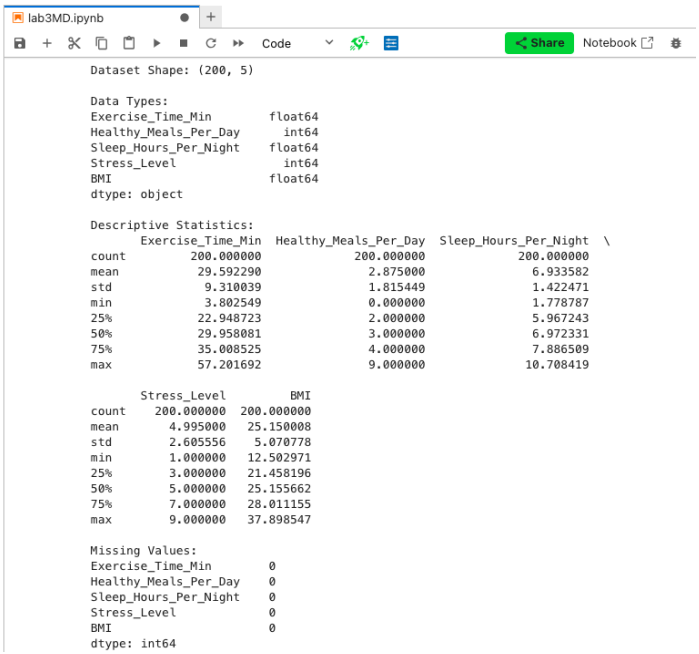
Lab 3

Using health and wellness data, I will be clustering the data to provide a deeper insight on how the healthcare organization could use these results to produce more efficient health interventions. First, I will use k-means clustering and hierarchical clustering with the data and forming an evaluation score. Then a principal component analysis (PCA) will be used to reduce the complexity of the data with a cumulative variance of over 70%. Lastly, I will then perform k-means clustering and hierarchical clustering with the PCA data and compare the results, using the silhouette score and within cluster sum of squares (WCSS)

Providing someone with a specific health plan is a difficult task. Everyones body reacts and performs differently, to different treatment, plans, and processes to improve their health. With the main objective of being able to sort an incoming patient needing health intervention based on healthy meals per day, BMI, stress level, sleep hours, and exercise time, having accurate groupings can help doctors set their patient on a path to a healthy life.

The health data provided, gives statistics on patients in categories listed in the paragraph above. To dive deeper into the descriptive statistics of the data look at Figure 1.

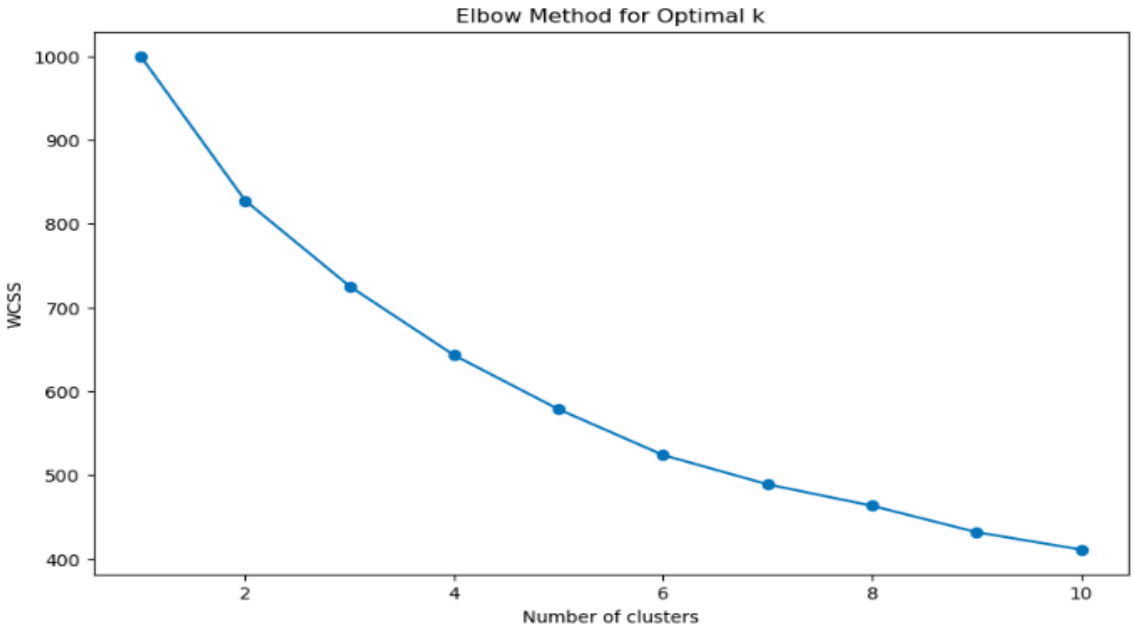
Figure 1



Within Figure 1, there are no strings as data types, so standardizing the data will not be a problem. Under descriptive statistics there is the mean, standard deviation, and quartiles of all of the data categories. There is also no missing data within this data set.

With little preprocessing steps needed, after standardizing the data, I figured out my K using the elbow method in Figure 2.

Figure 2 (Elbow Method Graph)



After analyzing the graph looking for when the lines start to level out, I chose 4 as my K value. Next, was creating a before and after PCA, K-means and hierarchical clustering models, that were evaluated by the silhouette score and WCSS.

Figure 3 (K-means before PCA)

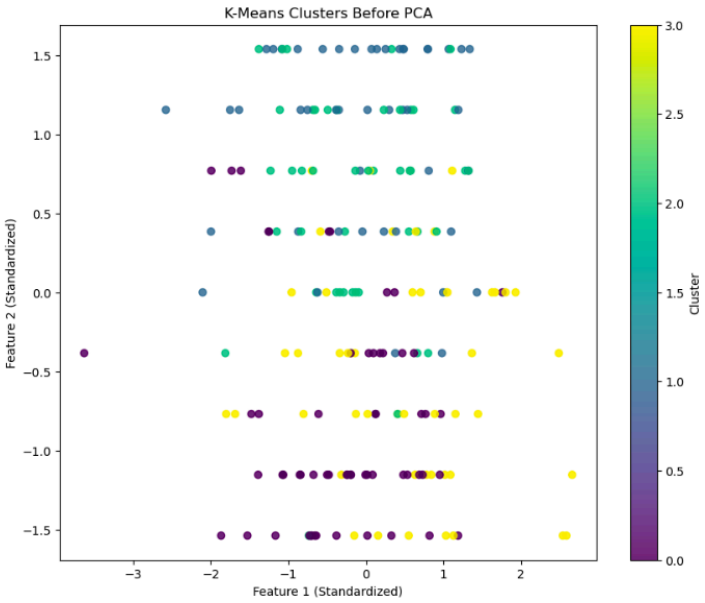


Figure 4 (Hierarchical before PCA)

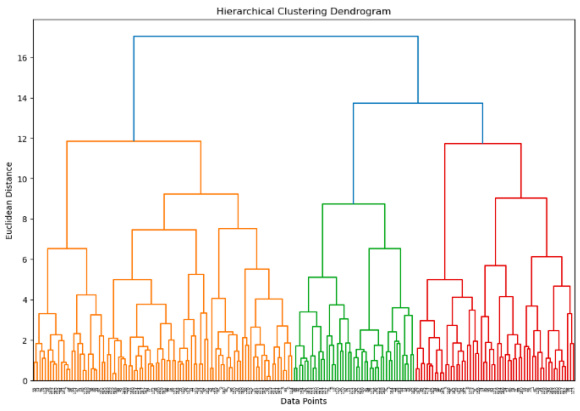


Figure 4 shows the dendrogram of the Hierarchical Clustering. Looking at Figure 3, the bottom half includes the majority of the yellow and purple dots, and the top half includes most of the green and blue dots. There still seem to be many data points intermixing with the sides of colors, meaning there is most likely noise within the data. With keeping a cumulative explained variance above 70% and PCA being a trustworthy route to reduce noise within clustering data, I will present the differences and then compare the improved results.

Figure 5 (PCA details)

```
Number of components selected: 4
Explained variance ratio: [0.23691549 0.22082517 0.19828377 0.18362786]
Cumulative explained variance: 0.84
```

Figure 5 shows that the machine found that 4 components were needed that equaled a cumulative explained variance of 0.84. This means that 84% of the variance was used within the data, keeping out noise, and not overfitting with a percentage too high.

Figure 6 (K-means after PCA)

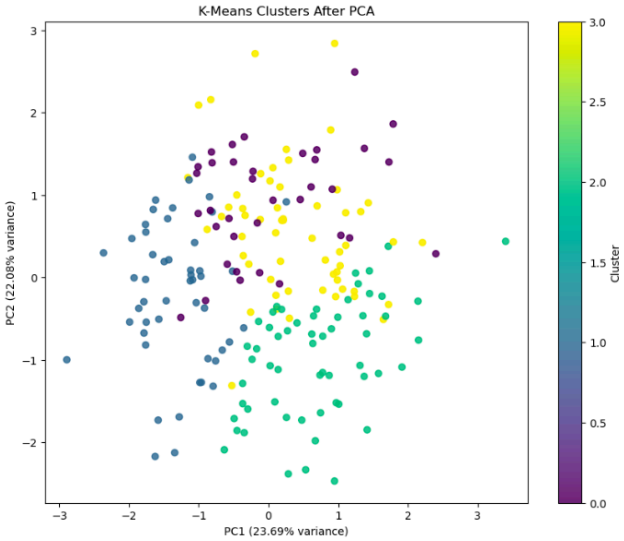


Figure 7 (Hierarchical after PCA)

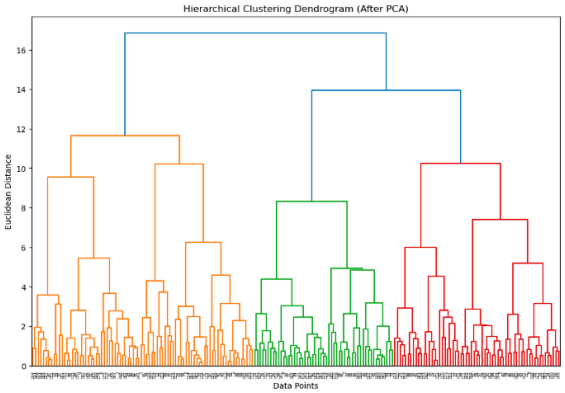


Figure 7 shows the changes made to the Hierarchical Clustering dendrogram. Figure 6 shows the K-means changes after PCA, which has made a more clear line between the clusters. Although there is still some mixing between the colors, the separation is more distinguishable from the K-means plot before PCA. Let's take a closer look into the specific numbers of our evaluation scores of WCSS and the Silhouette score.

Figure 8 (Stats and Visuals of Before vs. After PCA)

```
Clustering Performance Comparison (Before vs After PCA):
                          Metric  Before PCA   After PCA  Difference
0       Silhouette Score (K-Means)    0.168586    0.203051    0.034466
1                  WCSS (K-Means)  643.142085  488.980794 -154.161291
2  Silhouette Score (Hierarchical)    0.114392    0.162919    0.048527
```

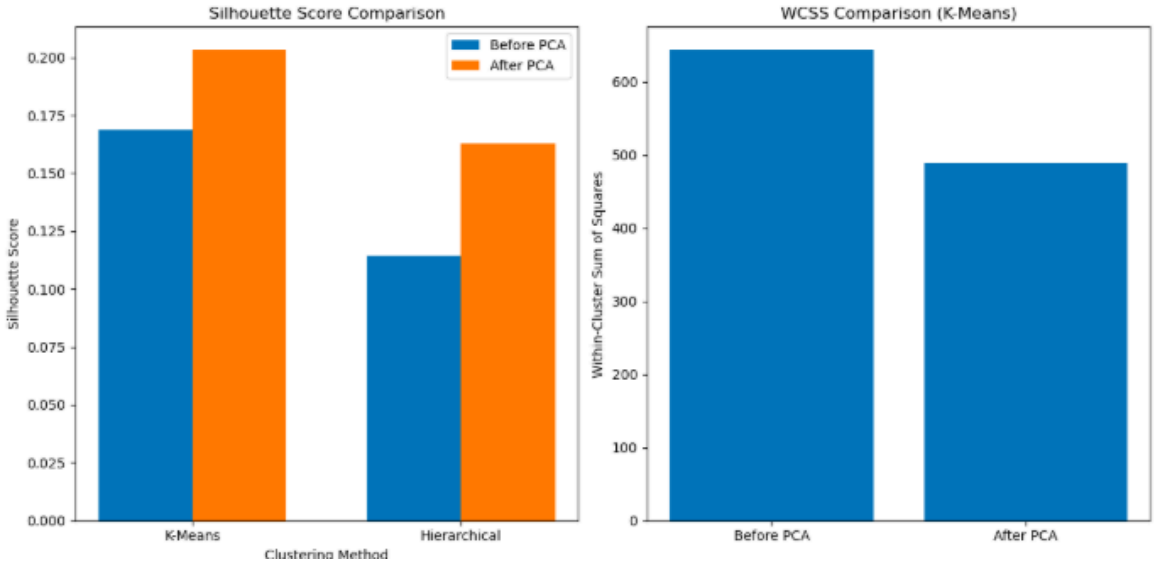


Figure 8 shows two visuals, on the left is the difference in Silhouette score for both models, before and after PCA. On the right is the WCSS before and after PCA for K-means. Lastly, the rows of data above is the specific number difference in the models evaluation scores before and after PCA. A higher Silhouette score means the clusters are better defined, which is clear that PCA did outperform the original K-means. After PCA, the K-means model had an increased Silhouette score of 0.034, while the Hierarchical Clustering score went up 0.049. WCSS is an error evaluation metric, so lower the number the better. Within the K-means model, it scored a lower WCSS score after PCA.

With these results presenting the dimensionality reduction method of PCA, improves the clusters within the data, making this useful for Doctors in categorizing patients. Using the new improved clusters, Doctors can use this to get a stronger basis on what type of health intervention is needed for that specific patient. Depending on the cluster the patient is place in, Doctors can see what they have done in the past with patients of the same cluster and act in an efficient manner to restore the patient to full health.