Exploring the Effects of Lifestyle and Health Factors on Obesity

Julio Amaya - 1923105, Jinglin Chen (Tony) - 2383599, Zachary Clark - 2308238, Subhan Mehmood - 2213304

Introduction:

According to the CDC, the prevalence of obesity in the US is over 41.9%, where more than 2 in 5 US adults are obese. $173 billion was spent on medical expenditures for obesity in 2019.
The dataset we have chosen is called Estimation of Obesity Levels Based on Eating Habits and Physical Condition. It is from the UC Irvine Machine Learning Repository, which is also available on Kaggle. The dataset includes estimates of obesity levels in Mexico, Peru, and Colombia, based on individuals' eating habits and physical conditions. The dataset contains 2,111 instances and 16 features with an additional variable attributing to the obesity level, totaling 17 features. It includes categorical, binary, continuous, and integer feature types. The subject area of this dataset is health and medicine.
We have chosen this particular dataset, focusing on obesity, because we want to understand the relationship between diet, weight, and lifestyle choices that contribute to the development of obesity.
Our main question regarding our unsupervised learning method is: Can we group individuals into distinct clusters based on lifestyle and health-related attributes that align with known categories of obesity risk?

Methodology:

We performed unsupervised learning on our dataset and implemented K-Means and Hierarchical Clustering.
The explicit formula for the main optimization criterion for K-Means Clustering is the goal of this method. This formula is given by:

$$\text{minimize}_{C_1,\ldots,C_K} \left\{ \sum_{k=1}^{K} W(C_k) \right\}$$

In Hierarchical Clustering, the clusters are formed by successively merging or splitting groups of observations based on specific linkages that define how distances between clusters are calculated. Single Linkage merges clusters based on the minimum distance between points. Complete Linkage merges clusters based on the maximum distance between points. Average Linkage merges clusters based on the average distance between all pairs of points. Centroid Linkage merges clusters by minimizing the distance between the centroids of points.
K-Means Clustering allowed us to group individuals by minimizing within-cluster variance. It enabled us to control the number of clusters, allowing us to create a varied number of clusters based on our understanding of the dataset.
Hierarchical Clustering allowed us to build a hierarchy or ranking of clusters without specifying the number of clusters beforehand and provided a good visual of the data via dendrograms.
However, K-means and Hierarchical Clustering have downsides. K-Means finds local optima, not global optima, and the final solution depends on the random initial assignments made at the beginning. Hierarchical Clustering can be very slow for a large number of observations, which was the case for our dataset. The results also largely depend on the chosen linkage.

Additionally, we compared the performance of the two approaches using the Silhouette Coefficient and found which model performed better in clustering the dataset. We also utilized cluster validation techniques, including internal validation through the Silhouette coefficient, external validation via post hoc analysis by using external labels to assess the performance of the clustering methods, as well as relative validation using gap statistics, to determine the optimal number of clusters.

Data Analysis:

Exploratory Data Analysis (Subhan and Julio):
We started our project by checking the quality of our dataset, and we found that it was clean, so no preprocessing was required. Also, based on the dataset size and the attributes of our features, we opted to use the full dataset for our unsupervised learning methods.

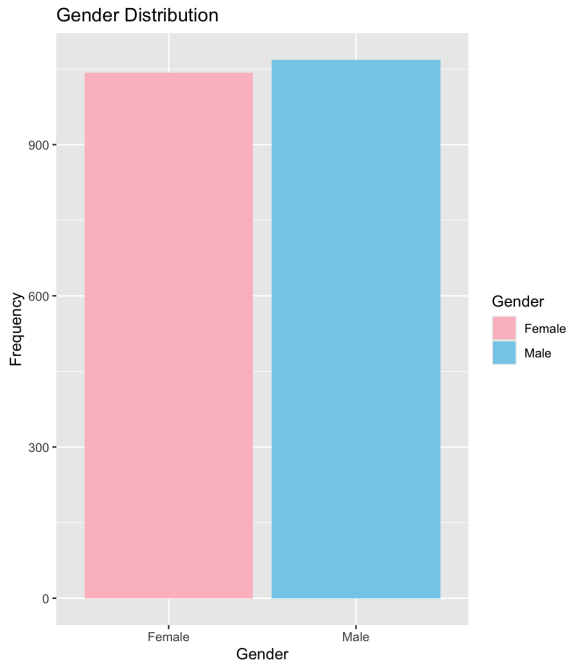
Next, we reviewed the summary of the dataset as a whole to identify any patterns or trends.

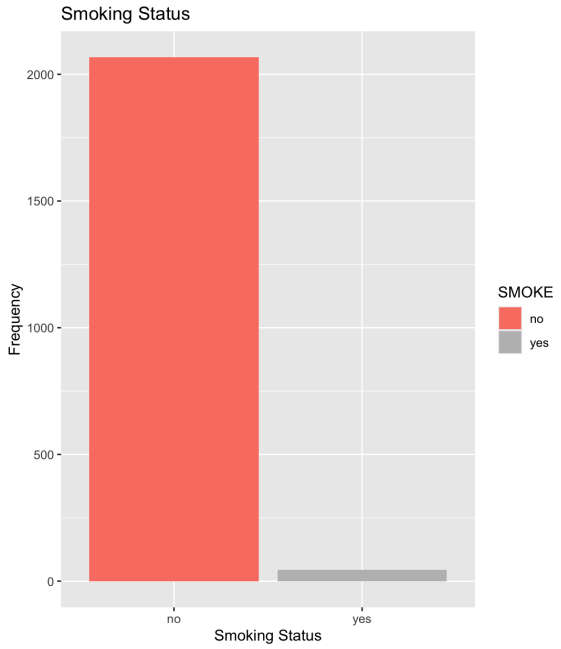
```
> summary(Obesity)
    Gender              Age            Height          Weight       family_history_with_overweight     FAVC               FCVC            NCP
 Length:2111        Min.   :14.00   Min.   :1.450   Min.   : 39.00   Length:2111                    Length:2111        Min.   :1.000   Min.   :1.000
 Class :character   1st Qu.:19.95   1st Qu.:1.630   1st Qu.: 65.47   Class :character               Class :character   1st Qu.:2.000   1st Qu.:2.659
 Mode  :character   Median :22.78   Median :1.700   Median : 83.00   Mode  :character               Mode  :character   Median :2.386   Median :3.000
                    Mean   :24.31   Mean   :1.702   Mean   : 86.59                                                    Mean   :2.419   Mean   :2.686
                    3rd Qu.:26.00   3rd Qu.:1.768   3rd Qu.:107.43                                                    3rd Qu.:3.000   3rd Qu.:3.000
                    Max.   :61.00   Max.   :1.980   Max.   :173.00                                                    Max.   :3.000   Max.   :4.000
     CAEC              SMOKE               CH2O            SCC                 FAF              TUE              CALC              MTRANS
 Length:2111        Length:2111        Min.   :1.000   Length:2111        Min.   :0.0000   Min.   :0.0000   Length:2111        Length:2111
 Class :character   Class :character   1st Qu.:1.585   Class :character   1st Qu.:0.1245   1st Qu.:0.0000   Class :character   Class :character
 Mode  :character   Mode  :character   Median :2.000   Mode  :character   Median :1.0000   Median :0.6253   Mode  :character   Mode  :character
                                       Mean   :2.008                      Mean   :1.0103   Mean   :0.6579
                                       3rd Qu.:2.477                      3rd Qu.:1.6667   3rd Qu.:1.0000
                                       Max.   :3.000                      Max.   :3.0000   Max.   :2.0000

  NObeyesdad
 Length:2111
 Class :character
 Mode  :character
```

Upon analysis, we can notice that the Age, Height, and Weight features have a diverse distribution. The age range is from 14 to 61, with a mean of 24.31, indicating a young adult population. Weight ranges from 39 to 173 kg, incorporating a wide span of body types. Health and dietary variables like FCVC, NCP, and CH20 appear to have a reasonable spread, judging from their means in relation to the minimum and maximum values. The mean of 1.0103 for FAF indicates that the individuals studied generally engaged in some form of weekly physical activity.

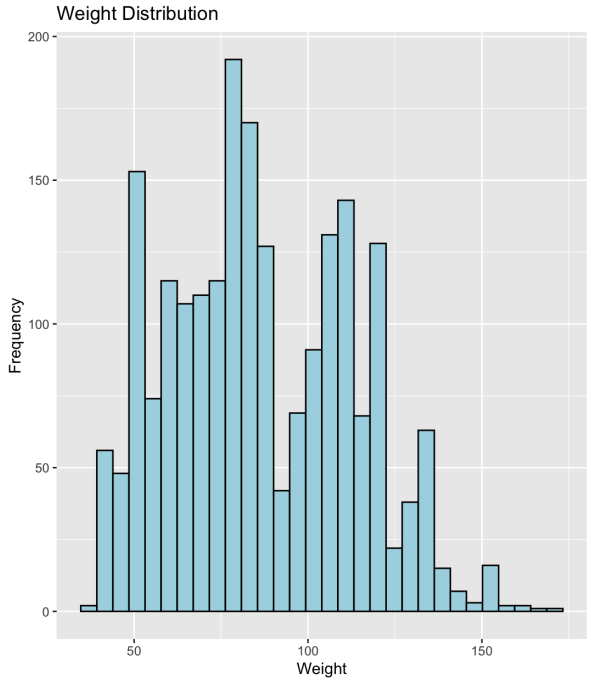
After, we performed some Exploratory Data Analysis to gauge the aspects of the dataset, where we first plotted bar charts of some of the features.

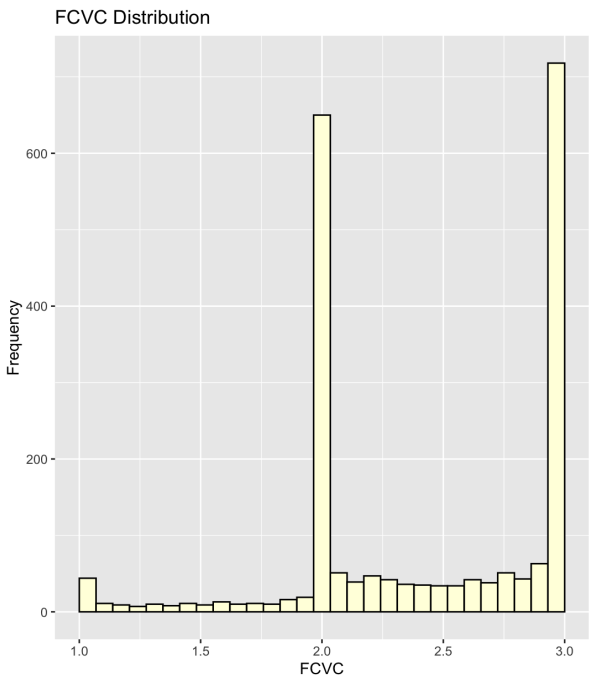




Here, for the Gender Distribution bar chart, we can see that the dataset has an almost equal distribution of males and females, but there are slightly more males who were studied.
From the Smoking Status bar chart, we can see that there were significantly more individuals who did not smoke as opposed to only a few who were smokers.

Next, we plotted histograms of some features to visualize their distributions.



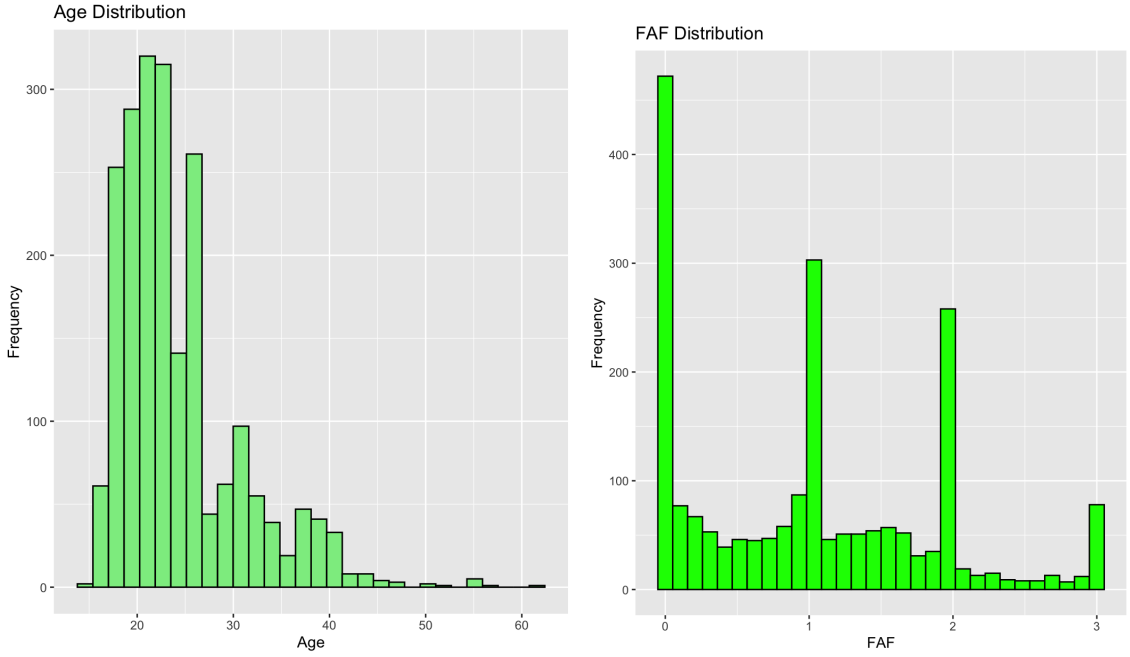


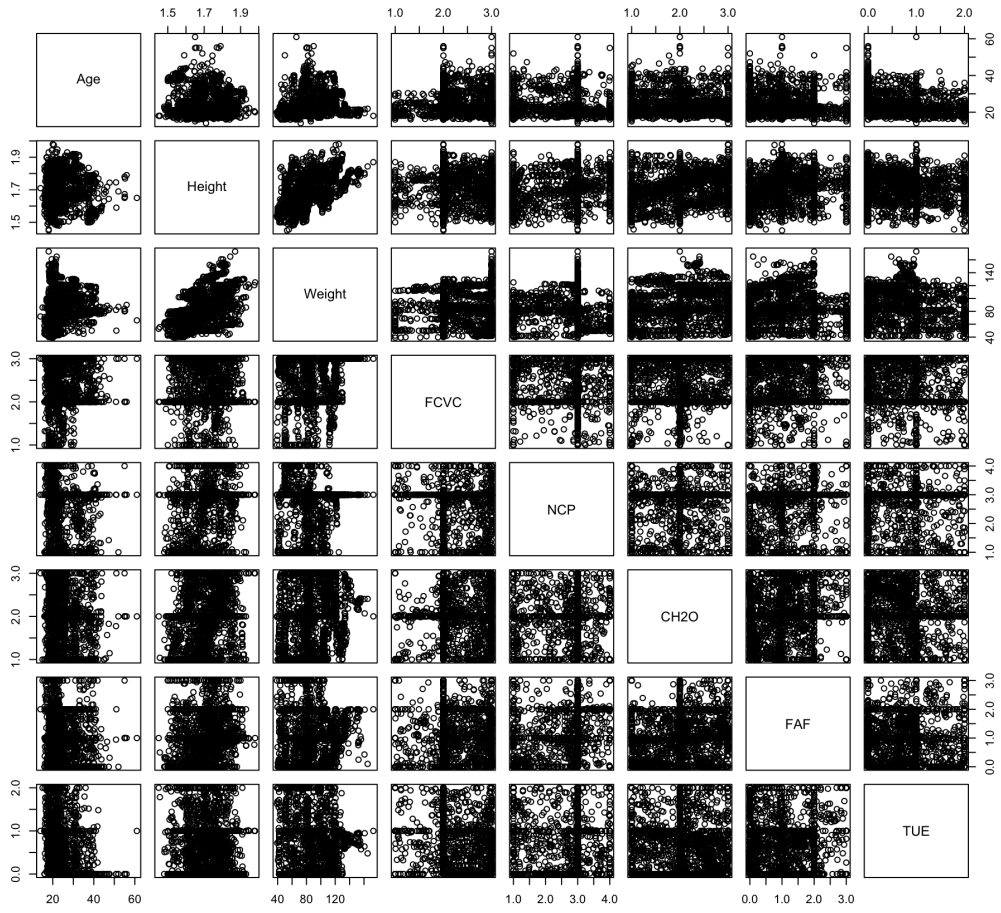
The Weight Distribution histogram again shows us how the weights of individuals are spread out, with a slight right skew. There are fewer individuals in the higher weight ranges, beyond 130 kg,

and multiple peaks are observed, specifically around 65-70 kg and 110-130 kg, suggesting that these weights are common.
The FCVC Distribution, which indicates how often individuals consume vegetables, shows values ranging from 1, indicating low frequency, to 3, indicating high frequency. There are two distinct peaks at 2.0 and 3.0, indicating a bimodal distribution, which suggests that individuals had a moderate to high frequency of consuming vegetables. The highest peak is at 3.0, suggesting that a large number of individuals frequently consumed vegetables. The earlier values seem to be less common, indicating that individuals who infrequently consumed vegetables were few.

The Age Distribution histogram reinforces our previous statements about the individuals studied, who are mainly from the young adult population, with most individuals ranging from 20 to 25 years old. The distribution is also right-skewed, further indicating that most individuals are younger and few are older. Also, the tail is long for this distribution, which could mean that the older population was underrepresented in this dataset.
The FAF Distribution histogram, representing the frequency of physical activity in hours per week, has three distinct peaks at 0, 1, and 2 hours, indicating that many individuals reported working out most for these specific hours, respectively. The highest peak is at 0 hours, showing that a large proportion of individuals did not engage in physical activity.

Both the Age Distribution and FAF Distribution histograms are shown below.



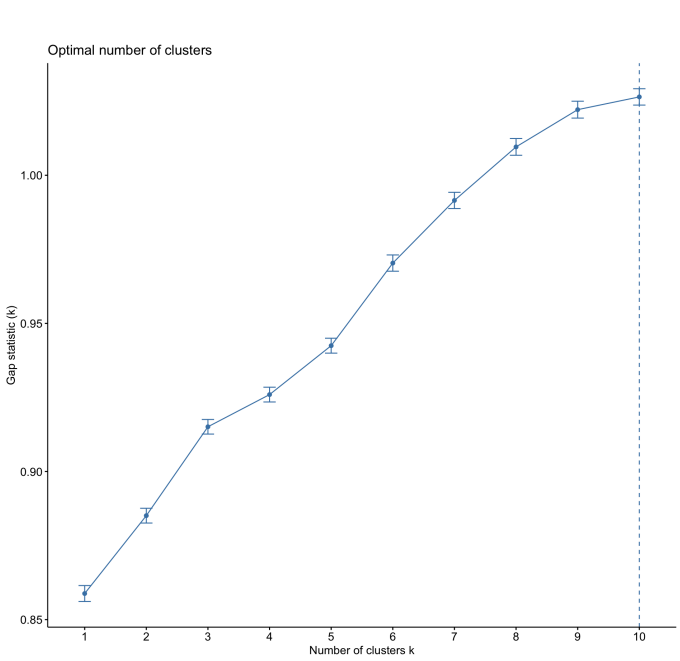


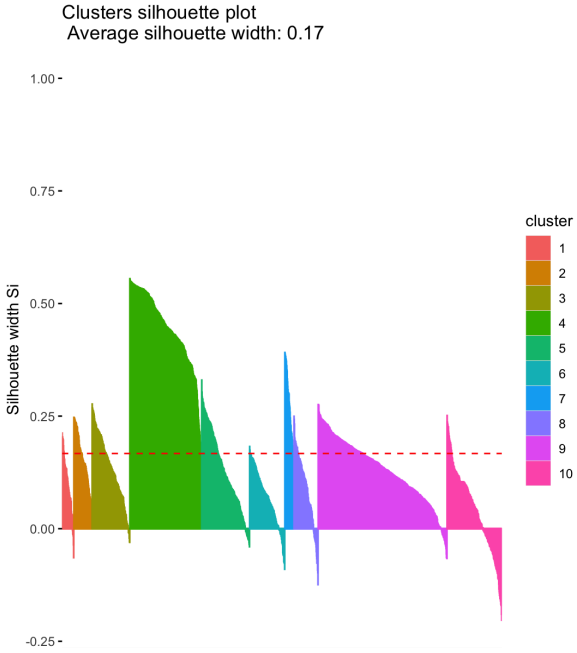
We also plotted a scatterplot matrix of all the features in our dataset, which was convoluted due to its large size. However, we can see that Height and Weight have a somewhat positive and linear relationship. Age and Weight have a slight positive trend but are somewhat scattered.

K-Means Clustering (Subhan and Julio):

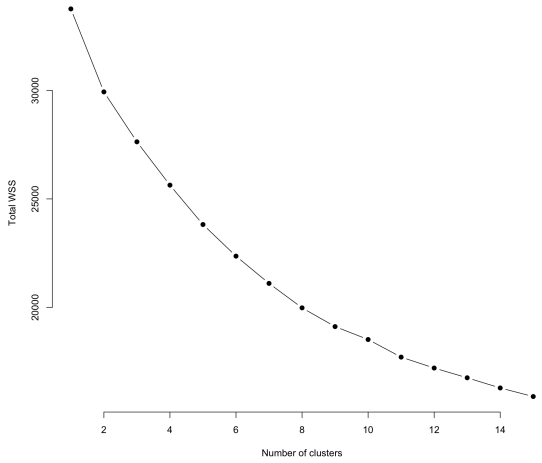
Following this, we implemented our first unsupervised learning model, K-Means Clustering. However, we first converted the categorical variables in this dataset into a numeric form using ifelse, as.numeric, and as.integer statements. Additionally, we considered all features from the dataset to be included within the models as they were all useful for their attributes in terms of health and diet habits, and connected to obesity in some way, but we excluded the feature, NObeyesdad, which is the categorical variable indicating the obesity level. Since this problem concerns unsupervised learning, we do not consider the response variable, nor do we need it. We also scaled our data, as this was necessary because both K-Means and Hierarchical Clustering use Euclidean Distance to measure similarity. If one feature had larger values compared to the others, it would be treated as more important due to its large numbers, which is not appropriate in this case. To give all features equal weight, we scaled the data for each variable, setting its mean to 0 and standard deviation to 1, so that each feature contributes equally to our analysis. Lastly, we selected all our numeric features and then applied the K-Means Clustering algorithm.

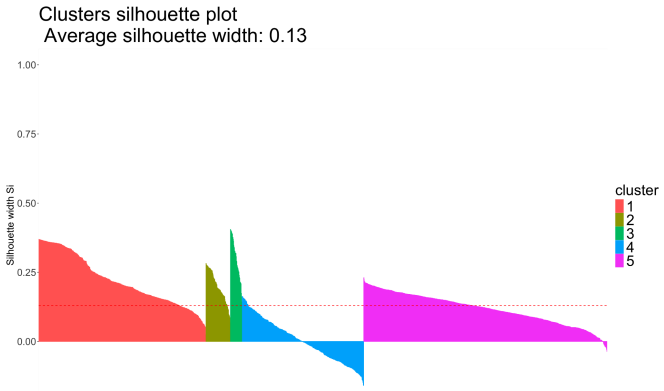
To implement K-Means Clustering, we first needed to find the optimal K for this dataset. We achieved this by deploying three methods to find the optimal K for each: the gap statistic, the elbow method, and the silhouette coefficient, and then selected the best option. A seed was set for each method for reproducibility.





From our code, each of the three methods yielded its own optimal K value. Hence, we performed further analysis by examining the WSS, silhouette scores, and visuals for each approach and agreed that K = 10 was the optimal K value via the gap statistic. The gap statistic and its corresponding silhouette plot are shown above. The silhouette plot had an average silhouette width of 0.17, indicating that the overall cluster separation was weak. However, some clusters were well-defined here, with high silhouette values, while others showed near-zero or negative silhouette widths, suggesting that some clusters were not well-defined. Moreover, for K = 10, Cluster 4 achieved the highest silhouette width of 0.4446, indicating that this cluster contained the most cohesive points among the others. K = 10 achieved a total WSS of 18,518.39 and a between-WSS of 15,241.61, which are good measures. The overall average silhouette width of K = 10 was 0.1669, suggesting that the overall clustering is weak to moderate. Also, K = 10 achieved a Between_SS/Total_SS of 45.1%, meaning that the clustering explains about 45.1% of the data's variability, which is higher compared to the other two methods; hence, this is another reason we chose K = 10 as our optimal K.

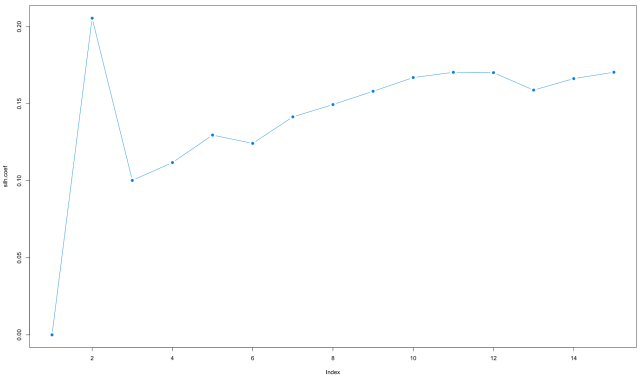


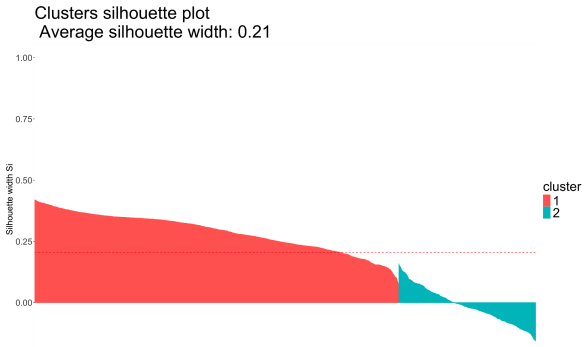


As for the statistics from the elbow method, which yielded an optimal K of 5, it achieved a total WSS of 23,819.48, which is higher than K = 10, and a between WSS of 9,940.517, lower than K = 10, indicating a less clear separation between clusters. It yielded an average silhouette width of 0.1296 and a Between_SS/Total_SS of 29.4%, both metrics less than K = 10. Therefore, it was clear that the optimal K from the elbow method was less effective than K = 10, and we did not consider it as a viable option for optimal K, even though the total WSS was higher, the other statistics were not advantageous for K = 5, and K = 10 had better measures. Additionally, the silhouette plot shows an average silhouette width of 0.13, indicating poor clustering structure. Compared to K = 10, the plot for K = 5 shows more low or negative silhouette widths, indicating poor cohesion among the clusters, which reinforces the idea that K = 5 is an inappropriate parameter here.

The optimal K of 2, as determined by the silhouette coefficient, achieved a total WSS of 29,934.54, higher than K = 10, and a between-WSS of 3,825.462, significantly lower than K = 10 and K = 5, indicating less clear separations between clusters. It also yielded an average silhouette width of 0.2054, which was higher than K = 10, but a Between_SS/Total_SS of 11.3%, which is much lower than both the other K values. Hence, even though K = 2 had a better total WSS and average silhouete width, the other statistics, in its entirety, did not seem better than the metrics of K = 10, and the slight increase in average silhouette width for K = 2 compared to K = 10 was not substantial for us to choose K = 2 as our optimal value. Moreover, the silhouette plot shows an average silhouette width of 0.21, indicating a moderate clustering structure. Cluster 1 shows relatively decent cohesion and separation, but Cluster 2 is not good, indicating that some points are improperly assigned.

The relevant plots and visuals for K = 2 are shown below.





Therefore, we implemented K-Means Clustering for K = 10. Figure 1 in the Appendix shows the clustering of our chosen optimal K value of 10 for this dataset.

From here, we can already see that clustering individuals into distinct clusters based on lifestyle and health-related attributes is not favorable and is challenging, especially given the statistics we achieved for our K-Means Clustering. Trying to group such individuals based on these features does not yield acceptable results, and it seems we are unable to do so.

Hierarchical Clustering (Tony and Zach):

For our second method of unsupervised learning, we implemented Hierarchical Clustering. We used 10 as our optimal K value, similar to K-Means Clustering, and implemented Hierarchical Clustering for each linkage: Single, Complete, Average, and Centroid. For each linkage, we also cut the dendrogram at K = 10 to define the number of clusters we wanted to extract. Through our analysis, we found the Single Linkage to be the best as it gave a better silhouette score compared to the other Linkages. The Single Linkage achieved a silhouette of 0.3976, while Complete yielded 0.0787, Average yielded 0.275628, and Centroid yielded 0.3281, the second highest and closest to Single. However, for a dataset like this, Single Linkage tends to generally perform worse than the other linkages; hence, it was peculiar that Single Linkage did so well, so through our consensus, we chose a different direction for our Hierarchical Clustering as we deemed this iteration of the Hierarchical Clustering to be unreliable.

To remedy this, instead of using K = 10 as our optimal K value, we chose K = 2, which, from previous analysis, was not a terrible parameter, even though we chose K = 10 in the end. Therefore, we implemented Hierarchical Clustering again, this time, with 2 as our optimal K value for each linkage, and cut the dendrogram at K = 2 to define the number of clusters we wanted to extract. Through our implementation, we found the Centroid Linkage to be the best as it gave the best silhouette score out of all the linkage methods. The Centroid Linkage achieved a silhouette of 0.4850, while Single yielded 0.4347, Complete yielded 0.4347, and Average yielded 0.4313. Here, we felt better with our results since the Centroid Linkage performed the best and is a linkage that generally does better for a dataset like this.

Figure 2 in the Appendix shows the dendrogram we achieved with Centroid Linkage, and it is indeed convoluted, granted that we used the full dataset for our Hierarchical Clustering. We can see that the dendrogram is dense at the lower height levels, suggesting that many low-distance

clusters were merged. In contrast, the higher height levels are sparse, indicating that few, widely spaced clusters were merged. This dendrogram can complicate the recognition of cohesive and well-separated clusters, as it shows a dataset with dense local grouping but a more distributed overall structure.

The silhouette plot we achieved for the Centroid Linkage Hierarchical Clustering (see Figure 3 in the Appendix) has an average silhouette width of 0.49, suggesting that the clustering is moderate to good and that the points are reasonably well within their own clusters and separated from other clusters. Here, Cluster 1 is where the majority of the points belong, while Cluster 2 is smaller and has observations with low or negative silhouette widths, meaning they do not fit well into their assigned cluster or might belong to a neighboring cluster instead.

Moreover, we implemented Hierarchical Clustering once more with K = 2 again, but this time, we only took into account the first 350 observations from the dataset instead of all the observations. This is primarily to achieve a more appropriate and clear dendrogram. Here, again, we cut the dendrogram at K = 2 to define the number of clusters we wanted to extract, but this time, we subsetted the dataset by taking only the first 350 observations and used Hierarchical Clustering on this new partitioned dataset. From the results, we found the Centroid Linkage to be the best-performing among the linkages as it achieved the highest silhouette score of 0.3399, while Single Linkage yielded 0.3306, Complete Linkage yielded 0.3306, and Average Linkage achieved 0.3306. Single, Complete, and Average all had the same values for their silhouette, while Centroid had a different one.

Figure 4 in the Appendix shows the dendrogram we achieved with Centroid Linkage. This time, the dendrogram is not as chaotic, and the merging of clusters can be seen clearly. There are many small and dense merges at low-level heights, where observations are merged at somewhat small distances, and fewer, broader merges at higher-level heights, suggesting that broader groupings exist among the data. This shows that it is relatively challenging to analyze well-separated and cohesive clusters at higher levels.

The silhouette plot we achieved for the Centroid Linkage Hierarchical Clustering (see Figure 5 in the Appendix) has an average silhouette width of 0.34, suggesting that the clustering quality is relatively weak to moderate and that most observations are assigned to their respective clusters fairly well. But, there are some observations that result in low or negative silhouette widths, indicating poor clustering. Cluster 1 contains the majority of the points, while Cluster 2 is smaller and contains points with low or negative silhouette values, which suggests that these observations are misclassified or belong to a neighboring cluster instead. This silhouette plot is similar to the plot for the Centroid Linkage Hierarchical Clustering of K = 2 with the full dataset.

Overall, our optimal K-Means Clustering Model with K = 10 achieved a silhouette score of 0.1669, our optimal Hierarchical Clustering Model with K = 2 and Centroid Linkage on the full dataset achieved a silhouette score of 0.4850, and our optimal Hierarchical Clustering Model with K = 2 and Centroid Linkage on a subset of the data achieved a silhouette score of 0.3399, which shows that the Hierarchical Clustering Model was better for this dataset in terms of unsupervised learning, in both cases, even though 0.4850, which is the highest of the two hierarchical methods, is moderate in its measure.

We also implemented External Cluster Validation by comparing our clustering results for both methods against the true labels with the Adjusted Rand Index. K-Means Clustering yielded a value of 0.2473, indicating moderate alignment with the true obesity labels and that the clusters captured some underlying structure. Hierarchical Clustering with the full dataset yielded a value of 6.9444e-06, and Hierarchical Clustering with the subset of the data yielded a value of 0.0061, both indicating a very weak alignment, basically random groupings in relation to the actual obesity categories.
It is very clear from this point on that grouping individuals into distinct clusters based on lifestyle and health-related attributes that align with known categories of obesity risk is impractical and problematic in this scenario. Finding good underlying structures and clusters based on these features is not possible and results in poor outcomes, suggesting that we cannot group individuals based on these entities as other factors at play make it difficult to do so and there could be other components needed to group such individuals, such as knowledge of the environment or genetics.

Conclusion:
To answer our main question that we posed in the beginning. Yes, to some extent, we can group individuals into distinct clusters based on lifestyle and health-related attributes that align with known categories of obesity risk. The K-Means Clustering with K = 10 produced a moderate structure, indicating that some clusters based on lifestyle and health exist, as shown by the silhouette score. This approach also demonstrated moderate agreement with the true obesity categories, as measured by the Adjusted Rand Index. As for the two Hierarchical Clustering models, the silhouette scores, which were higher than those of the K-Means Clustering, indicated moderate separation. However, the Adjusted Rand Index was very low for both scenarios, meaning that the clusters did not match the true obesity categories. As such, through our extensive analysis, although the K-Means Clustering aligned better with known obesity risk categories, the clusters did not accurately recreate the true obesity groupings, especially since the metrics of these models were not strong. This could be due to other unobserved factors, such as genetics, that may have also contributed to obesity, of which we were not aware. Hence, while we could cluster certain attributes, they only aligned with known obesity risk categories partially and not well.
Additionally, one main difficulty we faced in our data analysis was choosing the optimal K value for K-Means Clustering, as the three methods we implemented — gap statistics, the elbow method, and the silhouette coefficient — yielded different K values. Therefore, it was challenging to make sure we chose a good value for K to use in our models. Another difficulty we encountered was the low silhouette coefficients for our clustering models, which we found made it tough for our analysis and answering our main question.
Furthermore, a possible procedure to improve the data analysis would be to implement PCA before implementing the clustering models, so that we could apply dimensional reduction and reduce noise as well as reveal the underlying structure. Another possible procedure we could have practiced would be to engineer features so that we had more meaningful ones and fewer to worry about. For example, we could have created a new variable called BMI, taking into account the Weight and Height features, instead of using them separately.
In closing, though our models were not the best in our analysis, they answered our question to an extent. If we had implemented better procedures, we would have probably obtained better results and analysis.

References

Centers for Disease Control and Prevention. "Adult Obesity Facts." *Centers for Disease Control and Prevention*, 6 July 2023, https://www.cdc.gov/obesity/adult-obesity-facts/index.html.

UCI Machine Learning Repository. "Estimation of Obesity Levels Based on Eating Habits and Physical Condition." *UCI Machine Learning Repository*, 2019, https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition.

Appendix

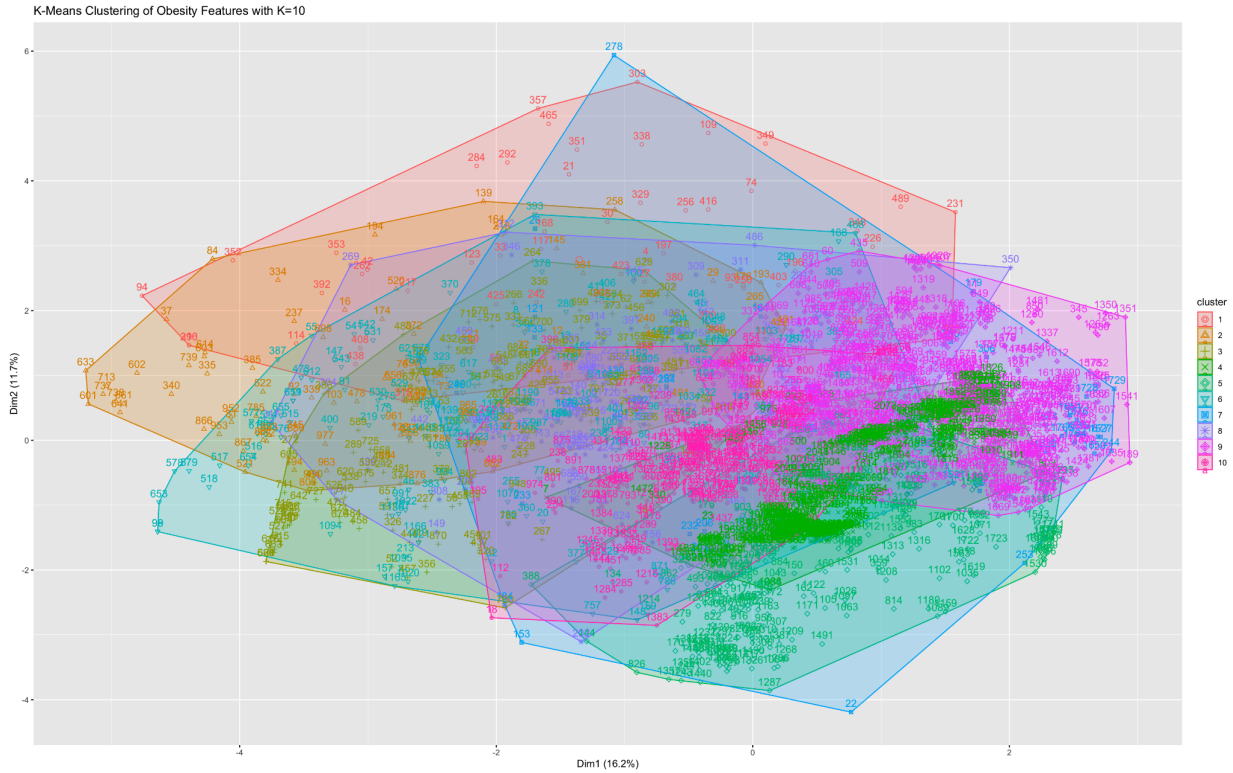


Figure 1: K-Means Clustering for K = 10

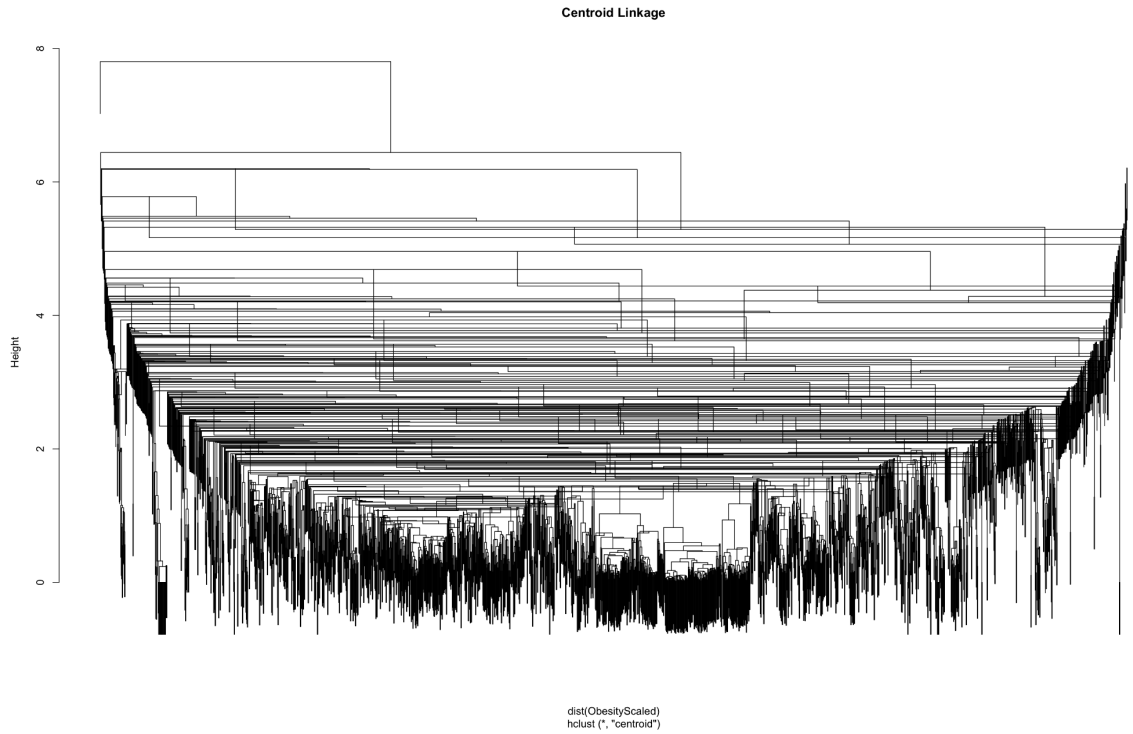


Figure 2: Dendrogram for Centroid Linkage Hierarchical Clustering with K = 2 and Full Dataset

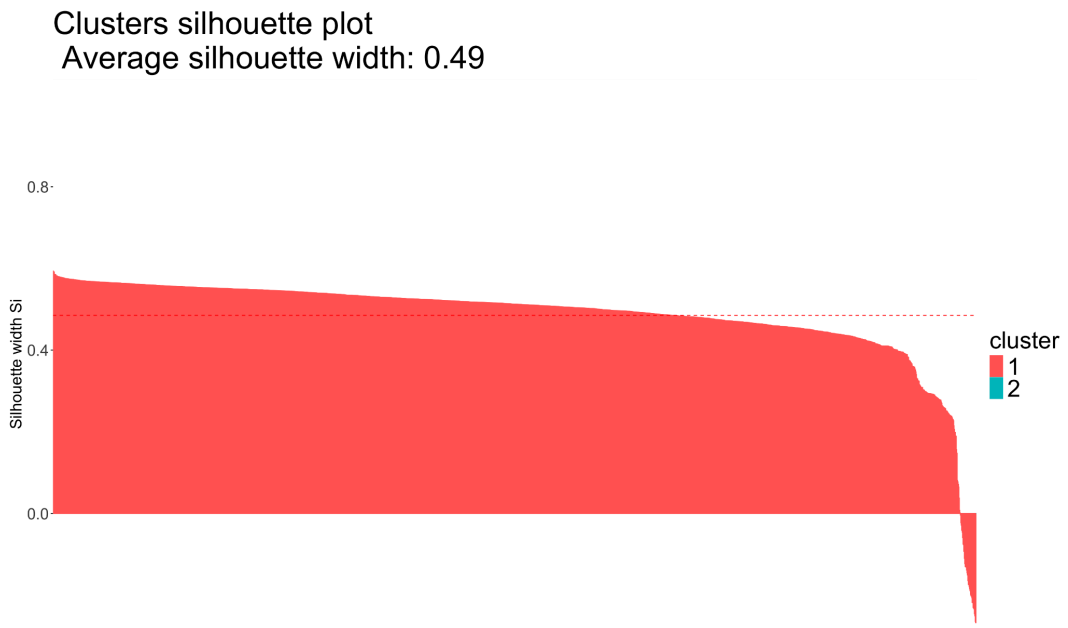


Figure 3: Silhouette Plot for Centroid Linkage Hierarchical Clustering with K = 2 and Full Dataset

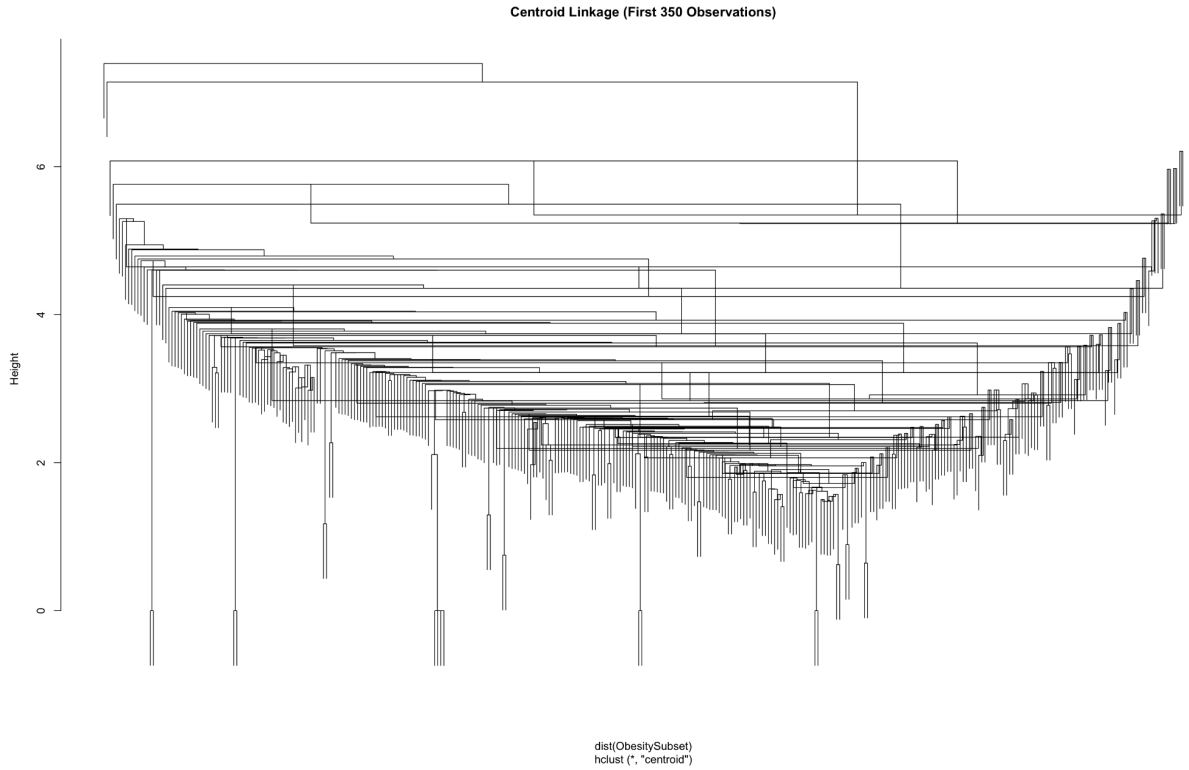


Figure 4: Dendrogram for Centroid Linkage Hierarchical Clustering with K = 2 and Subsetted Dataset

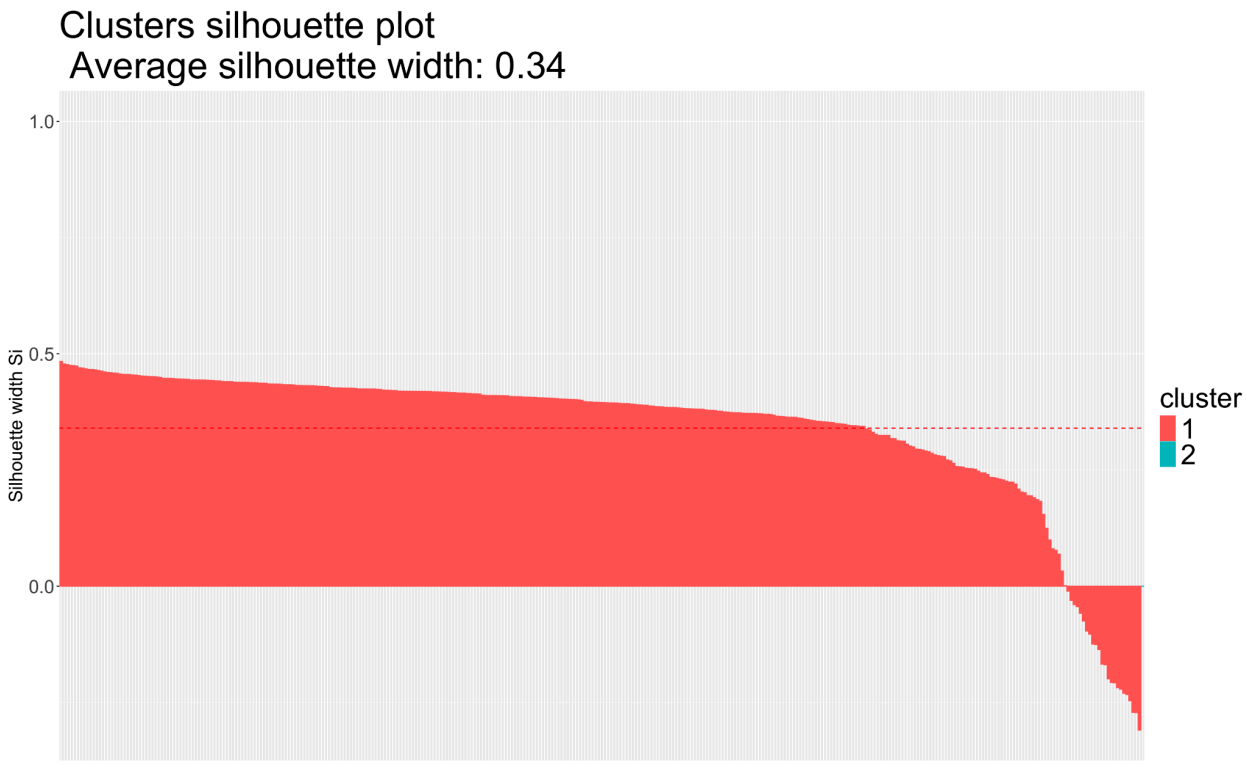


Figure 5: Silhouette Plot for Centroid Linkage Hierarchical Clustering with K = 2 and Subsetted Dataset