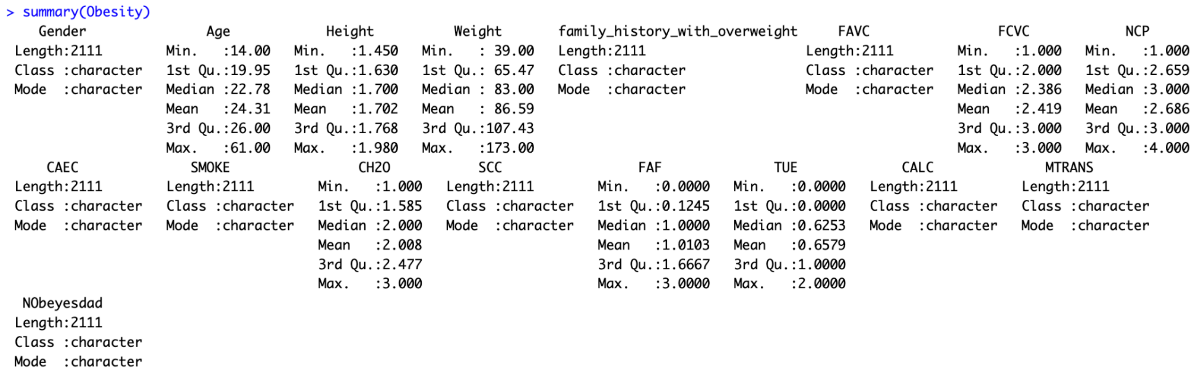
Obesity Final Report

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.
As such, 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. Moreover, we chose this dataset because it contains a range of categorical and numerical variables, as well as a sufficient number of variables and observations, which would allow us to implement various models for unsupervised learning. 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.
K-Means Clustering allowed us to group individuals by minimizing within-cluster variance, which worked well with our dataset. 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, provided a good visual of the data via dendrograms, and worked well with our dataset.
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.
Overall, K-Means was faster and allowed us to control the number of clusters, while Hierarchical Clustering did not require a pre-specified K value and provided a distinct visual representation of the data through dendrograms.
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:
We first started off our project by checking the quality of our dataset, where we found that our data was clean and no preprocessing was required. Also, based on the dataset size and the attributes of our features, we chose to go with the full dataset for our unsupervised learning methods.

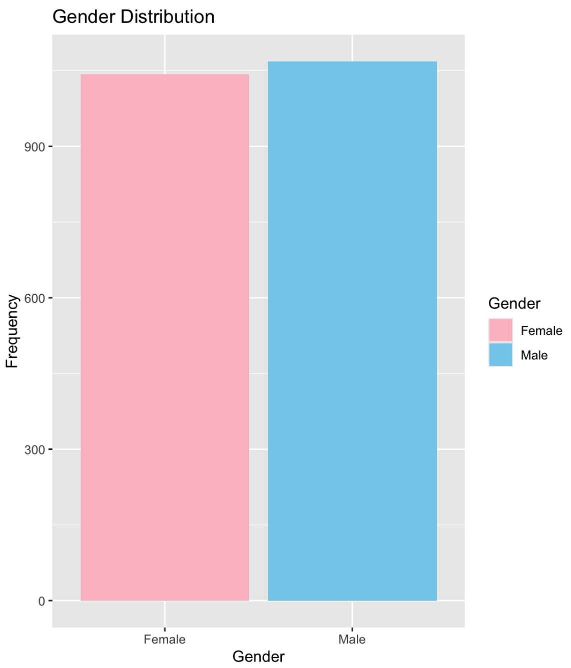
Next, we looked at the summary of the dataset as a whole to understand any patterns or trends.

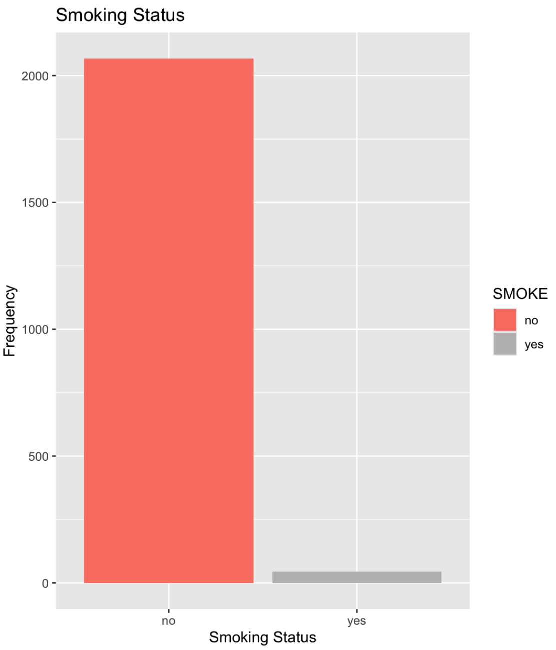


```
> summary(Obesity)
    Gender               Age            Height          Weight       family_history_with_overweight     FAVC                FCVC            NCP
 Length:2111        Min.   :14.00   Min.   :1.450   Min.   : 39.00   Length:2111                    Length:2111        Min.   :1.000   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. Age ranges from 14 to 61, with a mean of 24.31, suggesting 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 seem to have a reasonable spread judging from their mean in relation to the minimum and maximum values. The mean of 1.0103 for FAF indicates that individuals who were studied generally engaged in some form of weekly physical activity.

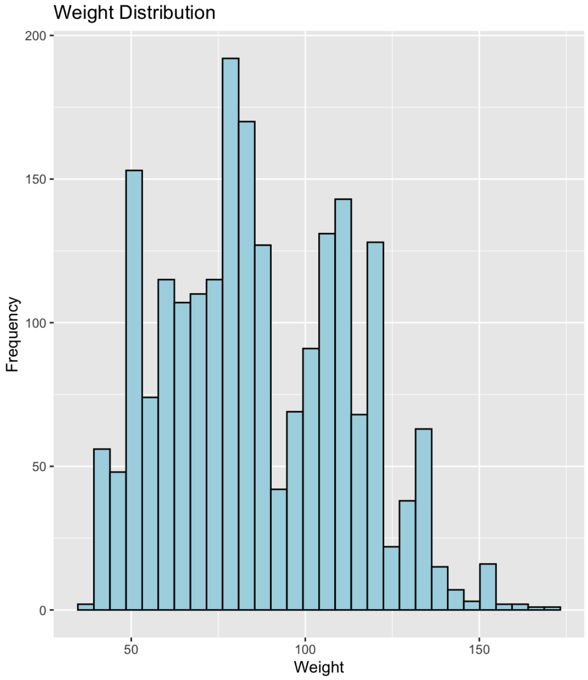
After, we performed some Exploratory Data Analysis to gauge the aspects of the dataset, where we first plotted bar charts of some of the features.

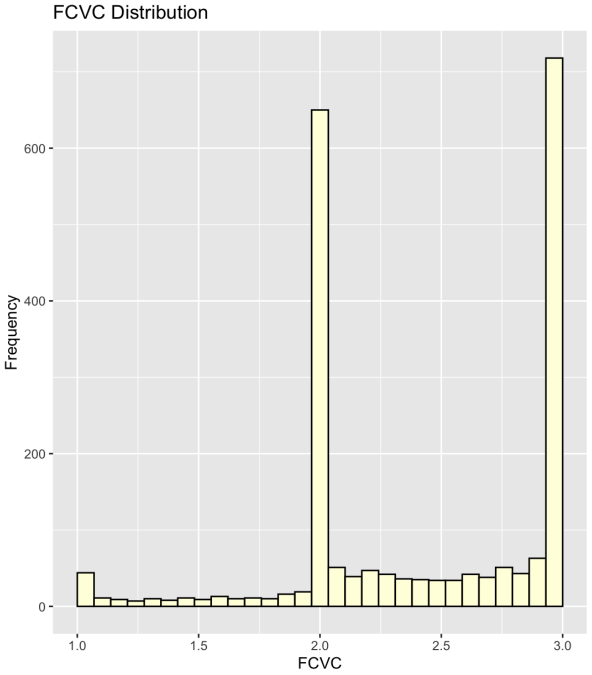




Here, for the Gender Distribution bar chart, we can notice that the dataset contains almost an equal distribution of males and females, but there are slightly more males who were studied. From the Smoking Status bar chart, we can see that there were significantly more individuals who did not smoke as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

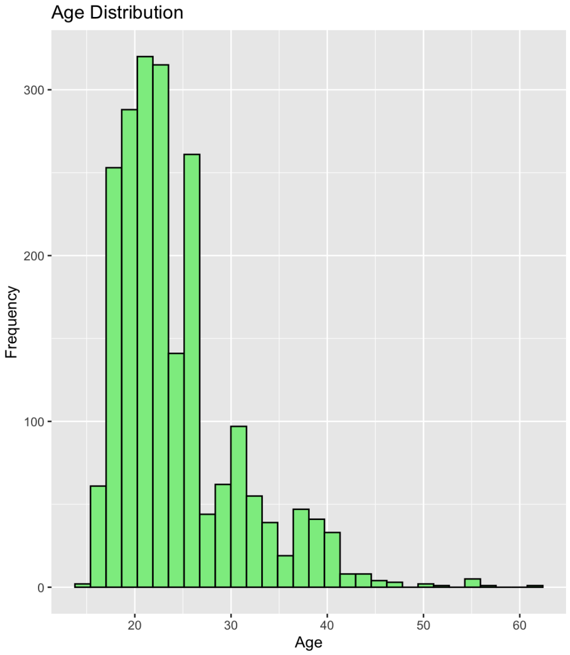
Next, we plotted histograms of some of the features to view their distributions.

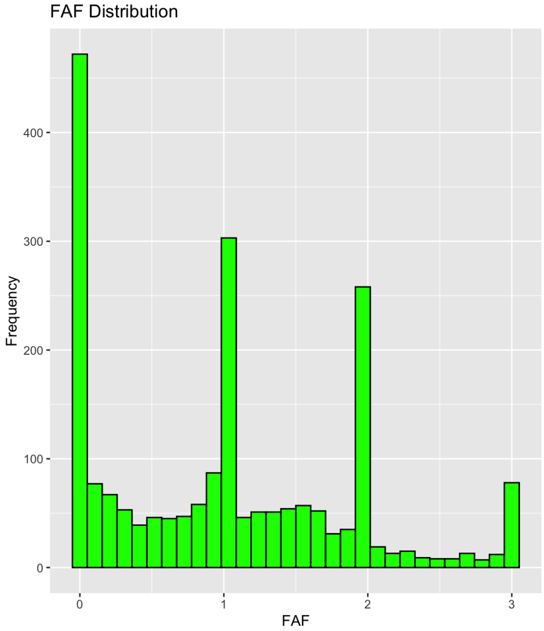




The Weight Distribution histogram again shows us how the weights of individuals are spread out among individuals with slight right-skew. There are fewer individuals in the higher ranges, beyond 130 kg, and there are multiple peaks, specifically around 65-70 kg and 90-100 kg, suggesting these weights are common.

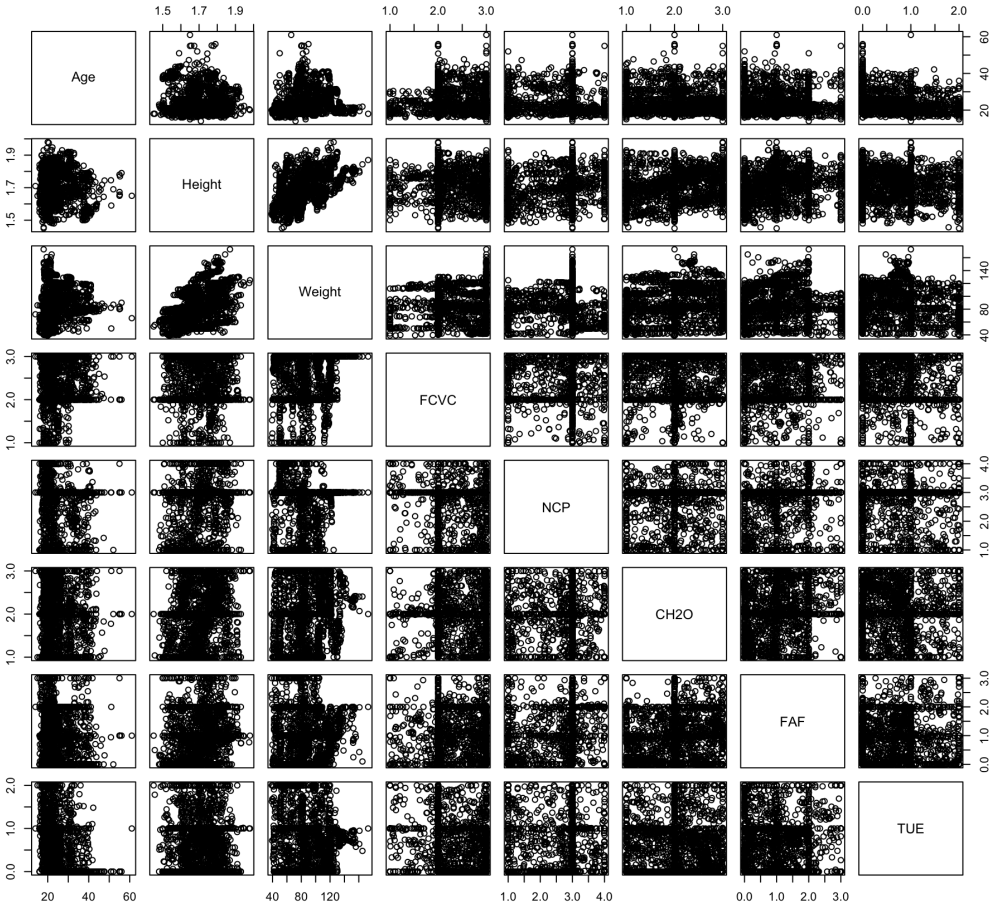
The FCVC Distribution, which is how often individuals consume vegetables, shows us that the values range from 1, indicating low frequency, and 3, indicating high frequency. There are two distinct peaks at 2.0 and 3.0, showing a bimodal distribution, suggesting that individuals had a moderate to high frequency in their consumption of vegetables, with the highest peak being at 3.0, indicating 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 being mostly from the young adult population, with most individuals ranging from 20-25 years. The distribution is right-skewed as well, further showing 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.

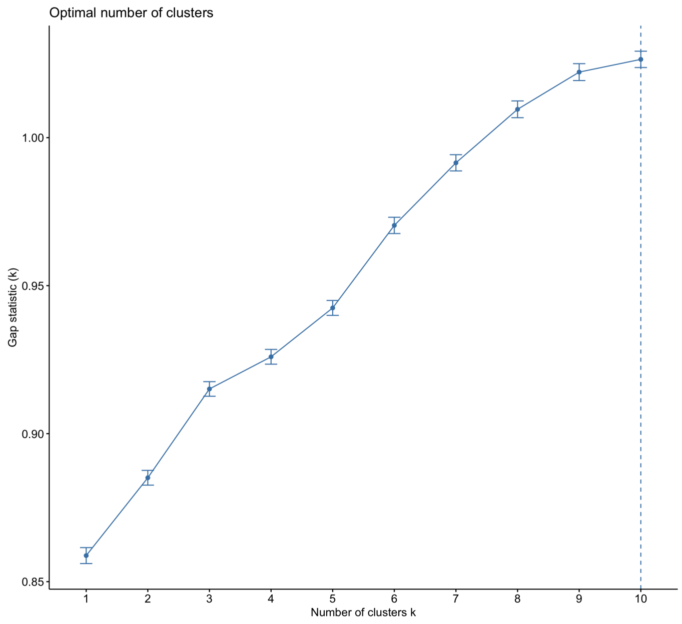
The FAF Distribution histogram, representing the physical activity frequency in hours per week, has three distinct peaks, at 0, 1, and 2 hours, indicating that many individuals stated working out most for these hours, respectively. The highest peak is at 0 hours, showing that a large proportion of individuals did not engage in physical activity.

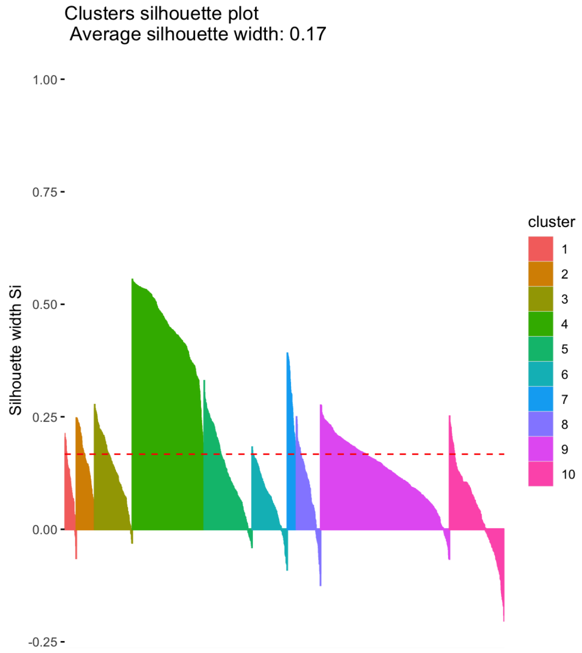


We also plotted a scatterplot matrix of all the features in our dataset, which was convoluted, granted that our dataset was very large, but we can see that Height and Weight have a somewhat positive and linear relationship. Age and Weight have a slight positive trend but are somewhat scattered.

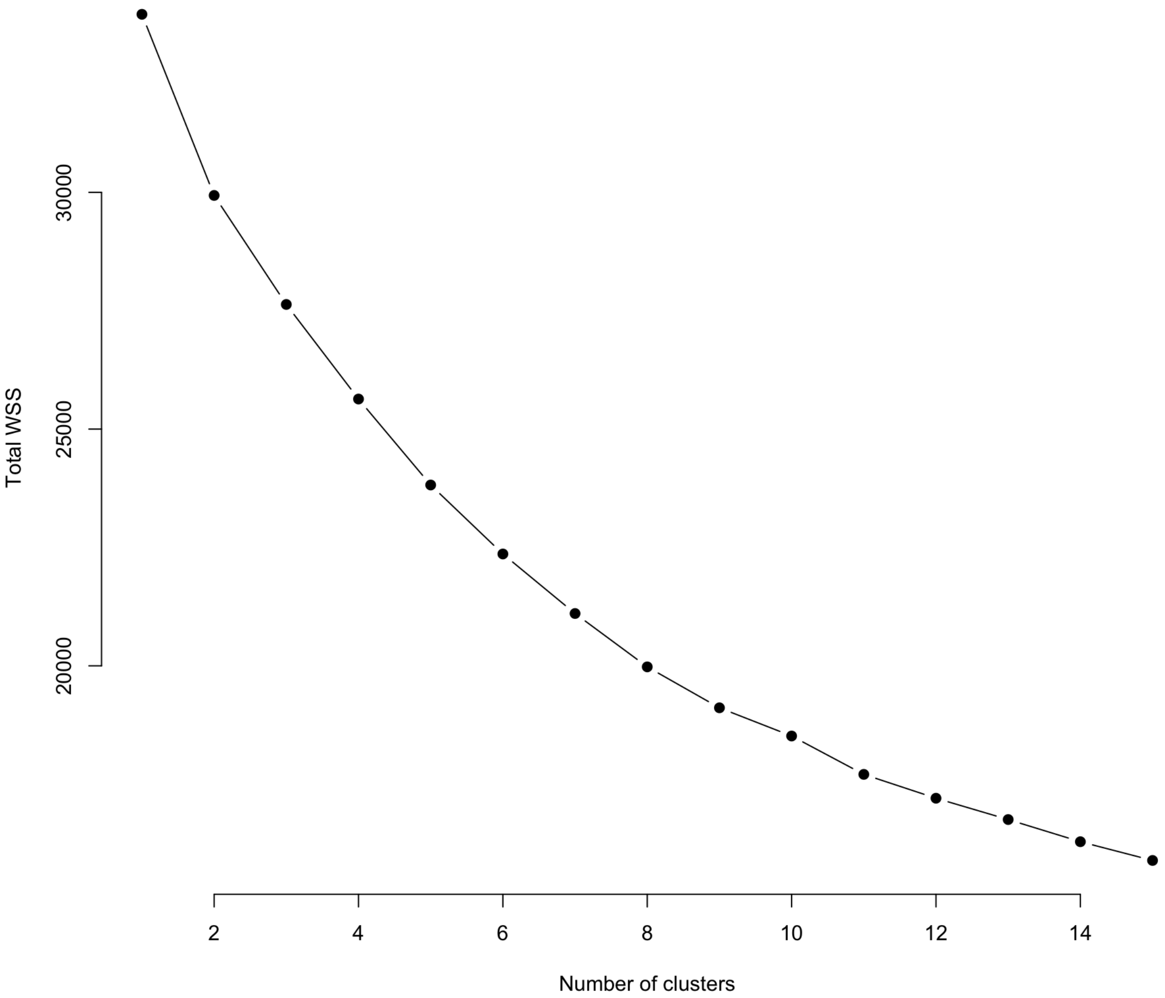
Following this, we implemented our first unsupervised learning model, K-Means Clustering. But, we first converted the categorical variables within this dataset into numeric form via ifelse, as.numeric, and as.integer statements. Additionally, we considered all features from the dataset to be included 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. We also scaled our data, as this was necessary because both K-Means and Hierarchical Clustering use Euclidean Distance to measure similarity, and if one feature had larger values compared to the others, that one feature would be treated as more important than the others because of its large numbers, which is not appropriate here. So, to give all the features equal weight, we scaled the data for each variable, giving them a mean of 0 and a standard deviation of 1 so that each feature contributes equally to our analysis. Lastly, we selected all our numeric features and then applied K-Means Clustering.

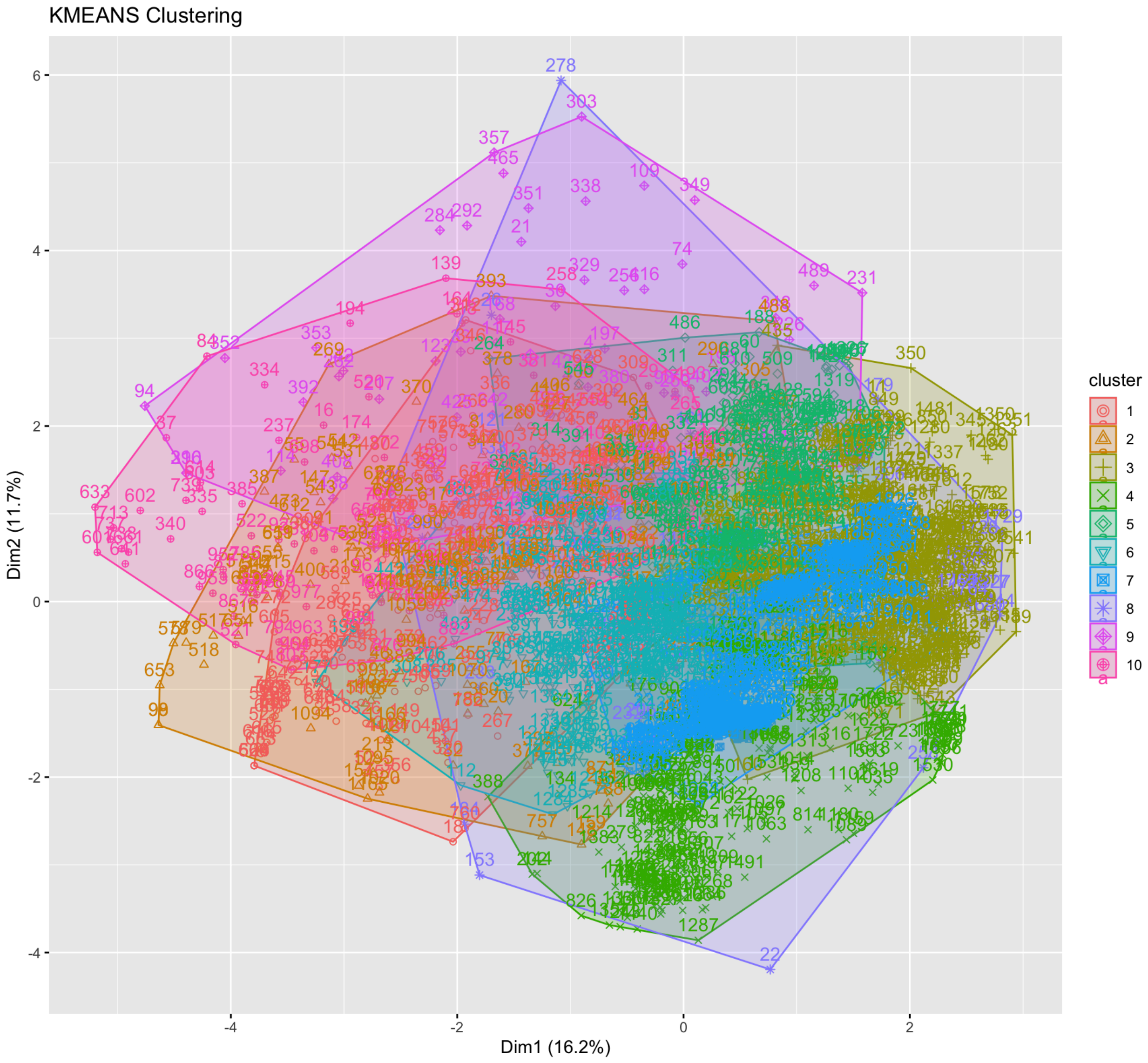
To proceed to implement K-Means Clustering, we had to first find the optimal K for this dataset. We did this by deploying three methods to find the optimal K for each: gap statistic, elbow method, and silhouette coefficient, then selecting the best option. A seed was set for each method for reproducibility.

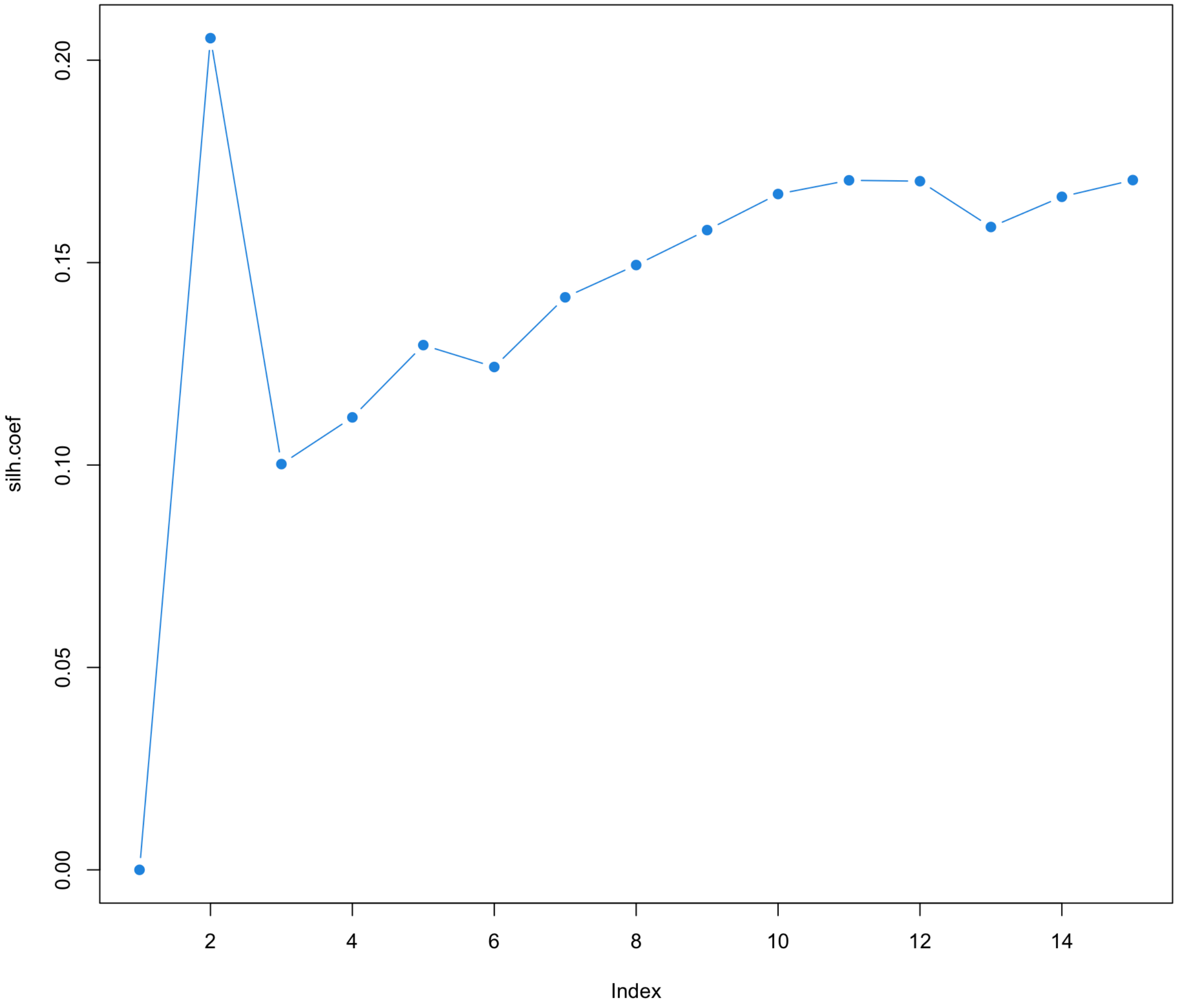




From our code, among the three methods, each method gave its own optimal K value. Hence, we performed further analysis by looking at the silhouette scores and visuals for each approach and agreed upon K = 10 being the optimal K value.


Total WSS
30000
25000
20000
2
4
6
8
10
12
14
Number of clusters


KMEANS Clustering
Dim2 (11.7%)
6
4
2
0
-2
-4
Dim1 (16.2%)
-4
-2
0
2
cluster
1
2
3
4
5
6
7
8
9
10
a


silh.coef
0.00
0.05
0.10
0.15
0.20
2
4
6
8
10
12
14
Index