**3.0 Data Modelling**

In assignment 3, we explored detailed about K-Means clustering and Hierarchal clustering. K-means clustering algorithm is one of the most famous centroid models. It runs iteratively to find the local optima and assigns each data point to the nearest centroids based on Euclidean distance to form multiple clusters. Number of clusters need to know prior to the experiment to perform k-mean clustering. In hierarchical clustering, the data points which are in their own clusters are merged with the nearest clusters until we have just one cluster at the end. The algorithm terminates when only one cluster left. Hierarchical clusters can be represented by dendrograms. The number of clusters is determined by cutting the vertical lines in dendrogram without intersecting other clusters. We used Big Mart dataset, which consists of 8523 rows of items from different product types across different types of outlet, to compare the efficiency and effectiveness of both the clustering algorithms. In this research, we looked at two factors when making comparison between both hierarchical and k-means clustering. This include the number of k clusters determined by both methods as well as the speed of the algorithm when building their clustering model.

**3.1 Number of Clusters**

In hierarchical clustering, the number of clusters are determined by plotting the dendrogram chart and then pruning the branches into several clusters using the cutree() function. This method is more of a “trial-and-error” method and requires you to tabulate the table of averages for each feature column in order to determine if the clusters make sense.

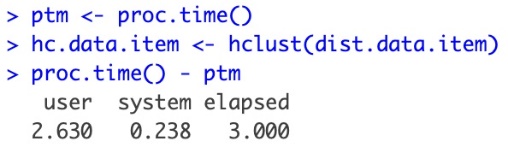
In k-means clustering, the number of clusters can be determined using a scree-plot, also known as the “Elbow method”. This method calculates the sum of squared errors (SSE) for each simulated iteration of a k value. When this graph is plotted, we look at the first significant “dent” or “kink”, which correlates to the number of optimal clusters. This method is less prone to error as compared to the visual identification of the optimal cut-off height of the dendrogram in hierarchical clustering.

**3.2 Speed of Algorithms**

In order to measure the speed of the algorithms when building the hierarchical and k-means model, we encapsulate the modelling code within the proc.time() function:

ptm <- proc.time()  
# MODELLING CODE  
proc.time() - ptm

This way, a timer is started before the code is executed and stopped when the code execution finishes. This will give us the elapsed time, which is the time taken for the code to run. When we model the hierarchical clustering via hclust(), it took 3 seconds to complete.

** **

In comparison, the k-means model only took 0.012 seconds to complete. This clearly shows that k- means clustering is comparatively faster than hierarchical clustering. This is as expected as hierarchical clustering starts by clustering each individual observation as its own separate clustering and merging them, while k-means uses the efficient mean-shifting centroid to determine the correct cluster.

As a conclusion, we found out that K-Means performs better than hierarchical clustering based on the computational time of the Big Mart dataset. K-Means clustering is computationally efficient even when the k value is bigger as compared to hierarchical clustering. Furthermore, K-Means clustering is easy to understand and implement. Thus, we prefer to use K-Means clustering which can compute faster on large dataset and at the same time produce quality results.