Peter Walsh

**Assignment 3 Write Up**

B. My selected clustering algorithm was hierarchical clustering. Hierarchical clustering is a method of cluster analysis that seeks to build a hierarchy of clusters. It operates by either merging smaller clusters into larger ones (agglomerative approach) or by dividing larger clusters into smaller ones (divisive approach). This technique is particularly useful for exploratory data analysis, where the goal is to identify natural groupings within the data.

C. To run hierarchical clustering, I used the hclust function in r. The hclust function is used to perform hierarchical clustering on a set of observations based on a distance matrix. It employs various linkage methods (such as complete, average, or single linkage) to group the observations into clusters, producing a hierarchical tree structure known as a dendrogram. This function allows users to visualize and analyze the relationships between clusters, facilitating the exploration of data patterns and groupings.

D. This method required me to select number of clusters, so I used silhouette analysis. Silhouette analysis is a technique used to evaluate the quality of clustering by measuring how similar an object is to its own cluster compared to other clusters. The silhouette score ranges from -1 to +1, where a score close to +1 indicates that the object is well-clustered and far from neighboring clusters, while a score near 0 suggests that it is on the border between two clusters. This analysis helps determine the optimal number of clusters and assess the effectiveness of the clustering method used. It ended up concluding that 2 was the optimal number of clusters.

E. Surprisingly, at each different number of gene, I found that 2 was the optimal amount of clusters. The highest silhouette score (0.99) resulted from the 10,000 gene level, indicating that 2 clusters resulted in a very well clustered data set. I could not implement the chi-squared test to test this.

A screenshot of a computer

Description automatically generated

Here is my alluvial diagram:

A graph with numbers and lines

Description automatically generated