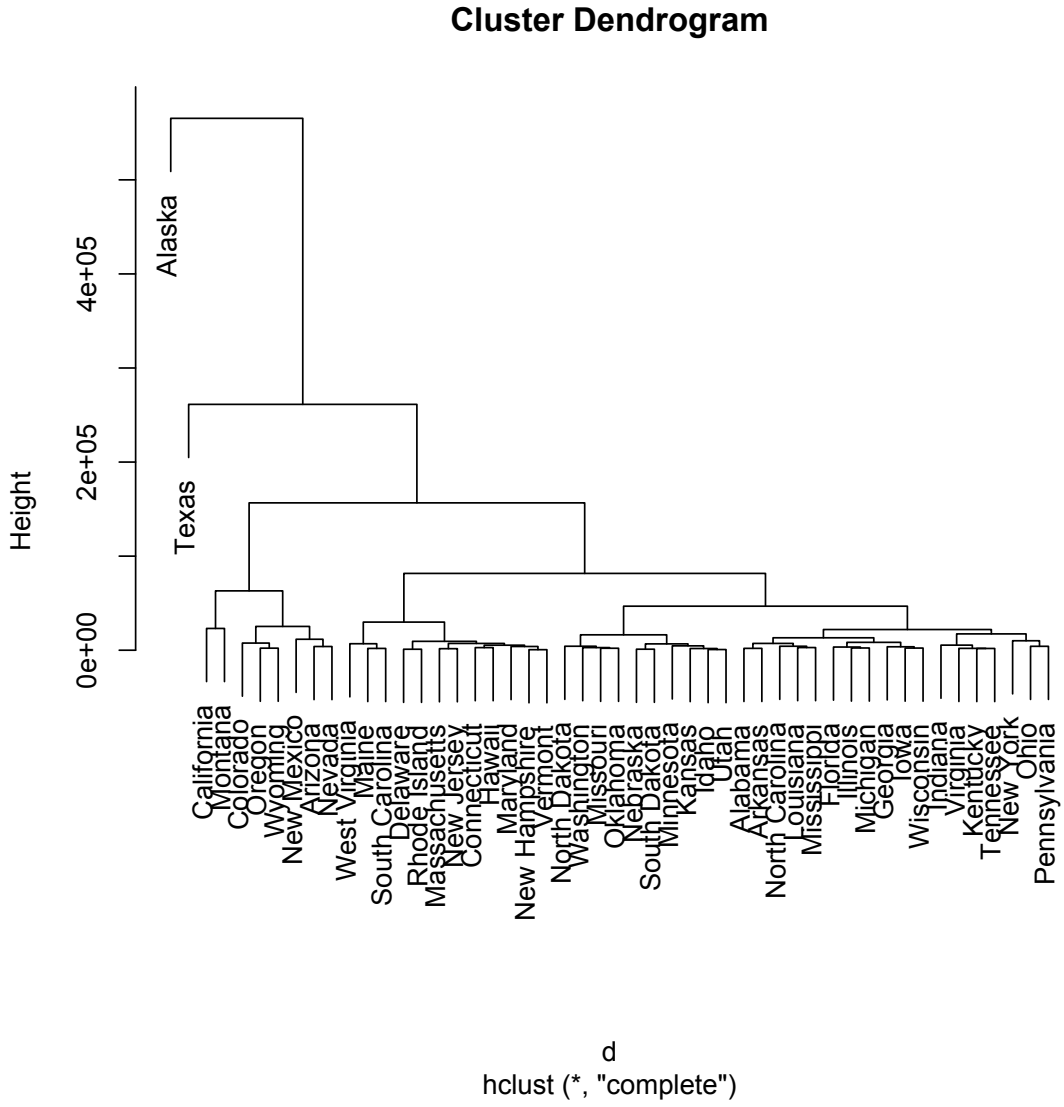
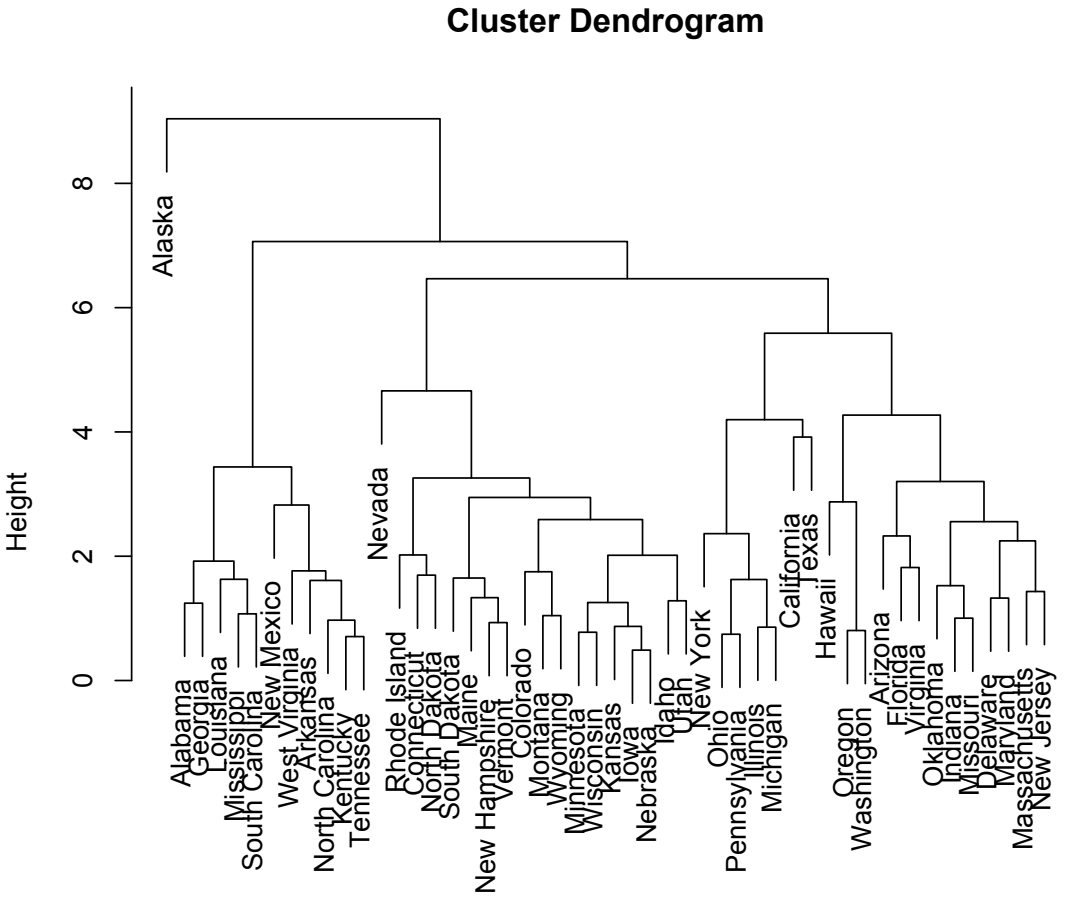
Agglomerative Hierarchical Clustering

1. Load the dataset
2. Use hierarchical clustering to cluster the data on all attributes and produce a dendrogram

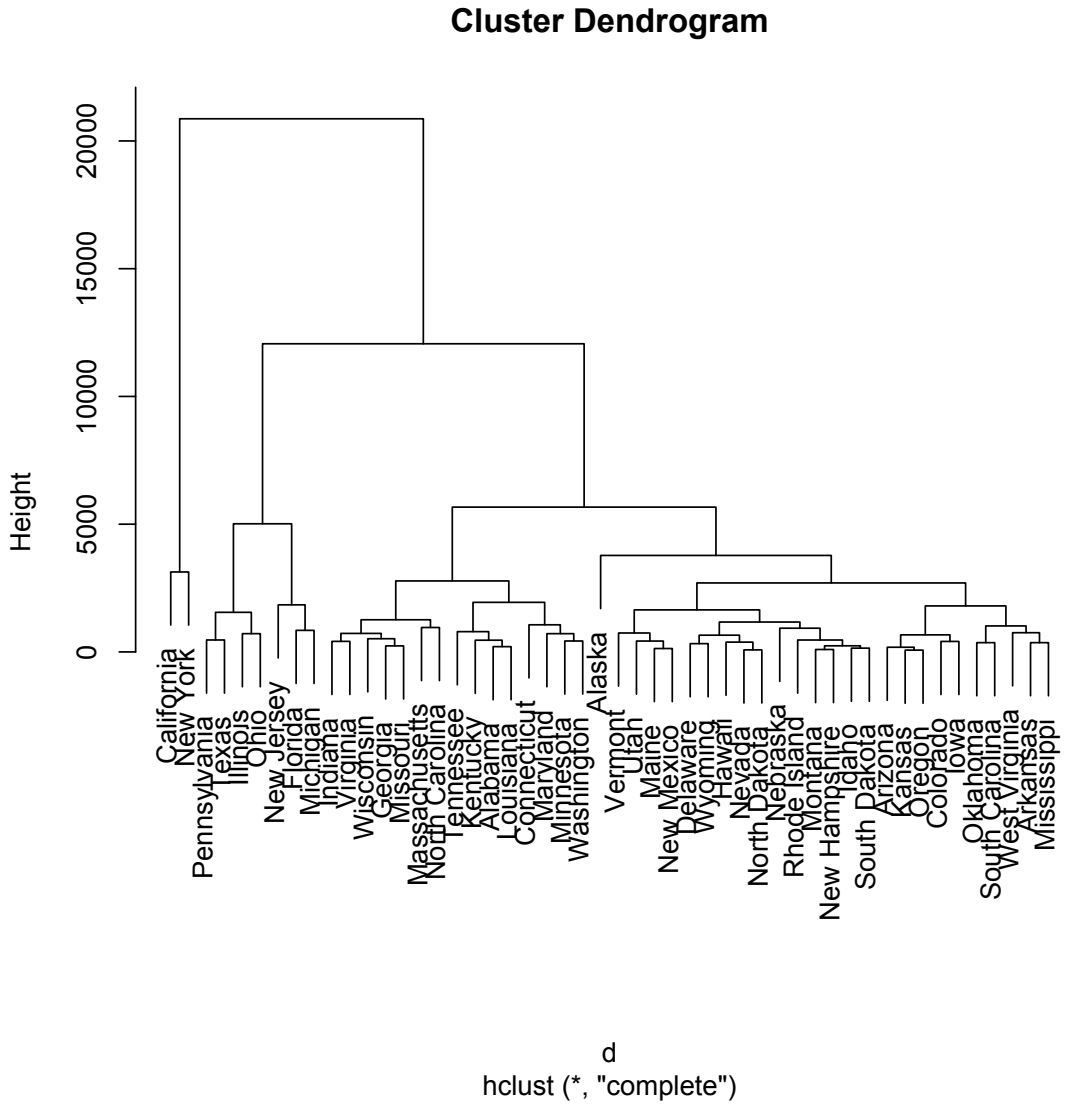


3. Repeat the previous item with a normalized dataset and note any differences

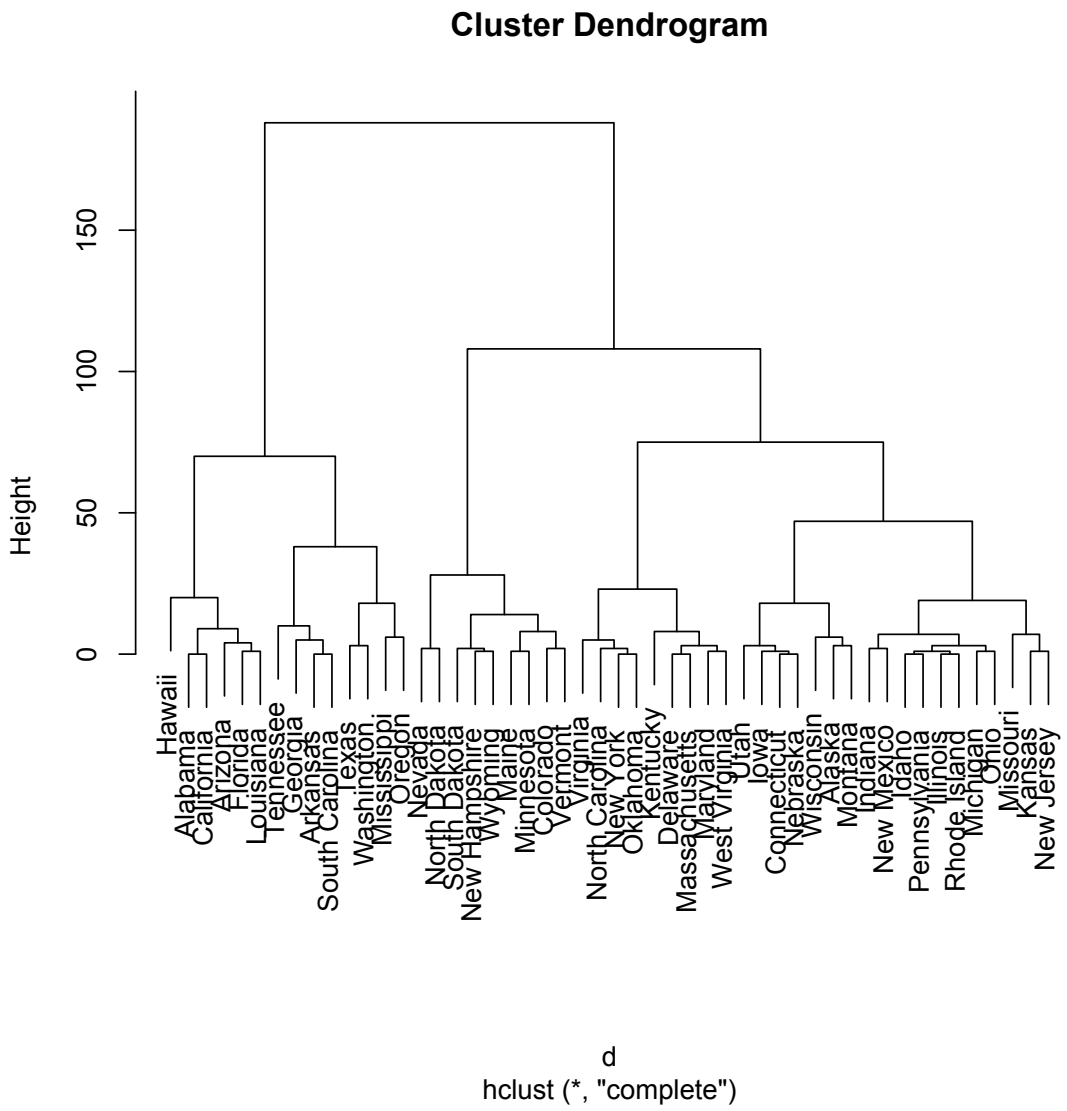


d
hclust (*, "complete")

4. Remove "Area" from the attributes and re-cluster (and note any differences)



5. Cluster only on the Frost attribute and observe the results

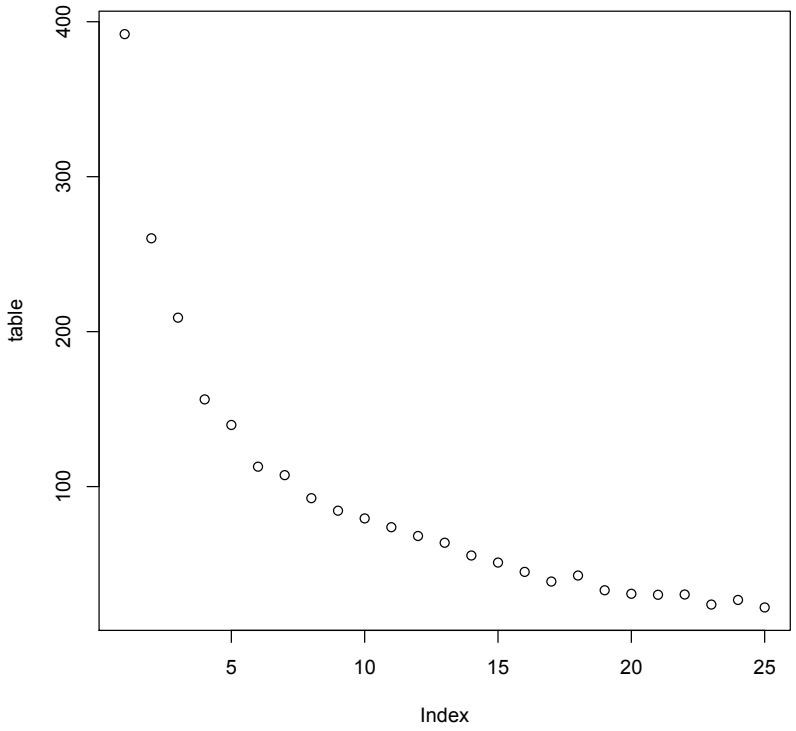


Using *k*-means

1. Make sure to use a normalized version of the dataset.
2. Using k-means, cluster the data into 3 clusters. Note the size of each cluster and the mean values. Do you have any insight into why they were divided this way?

   Some of the clusters seem pretty obvious. One cluster has pretty much all of the southern states. Another one mostly contains states with large cities and large populations. The third cluster seems to have a lot of states that have a smaller population.

3. Using a for loop, repeat the clustering process for k = 1 to 25, and plot the total within-cluster sum of squares error for each k-value.



4. Evaluate the plot from the previous item, and choose an appropriate k-value using the "elbow method" mentioned in your reading. Then re-cluster a single time using that k-value. Use this clustering for the remaining questions.
5. List the states in each cluster.

Cluster 1 – Delaware, Illinois, Indiana, Maryland, Michigan, Missouri, Nevada, New Jersey, Ohio, Pennsylvania, Virginia

Cluster 2 – Colorado, Connecticut, Idaho, Iowa, Kansas, Maine, Massachusetts, Minnesota, Nebraska, New Hampshire, North Dakota, Rhode Island, South Dakota, Utah, Vermont, Wisconsin, Wyoming

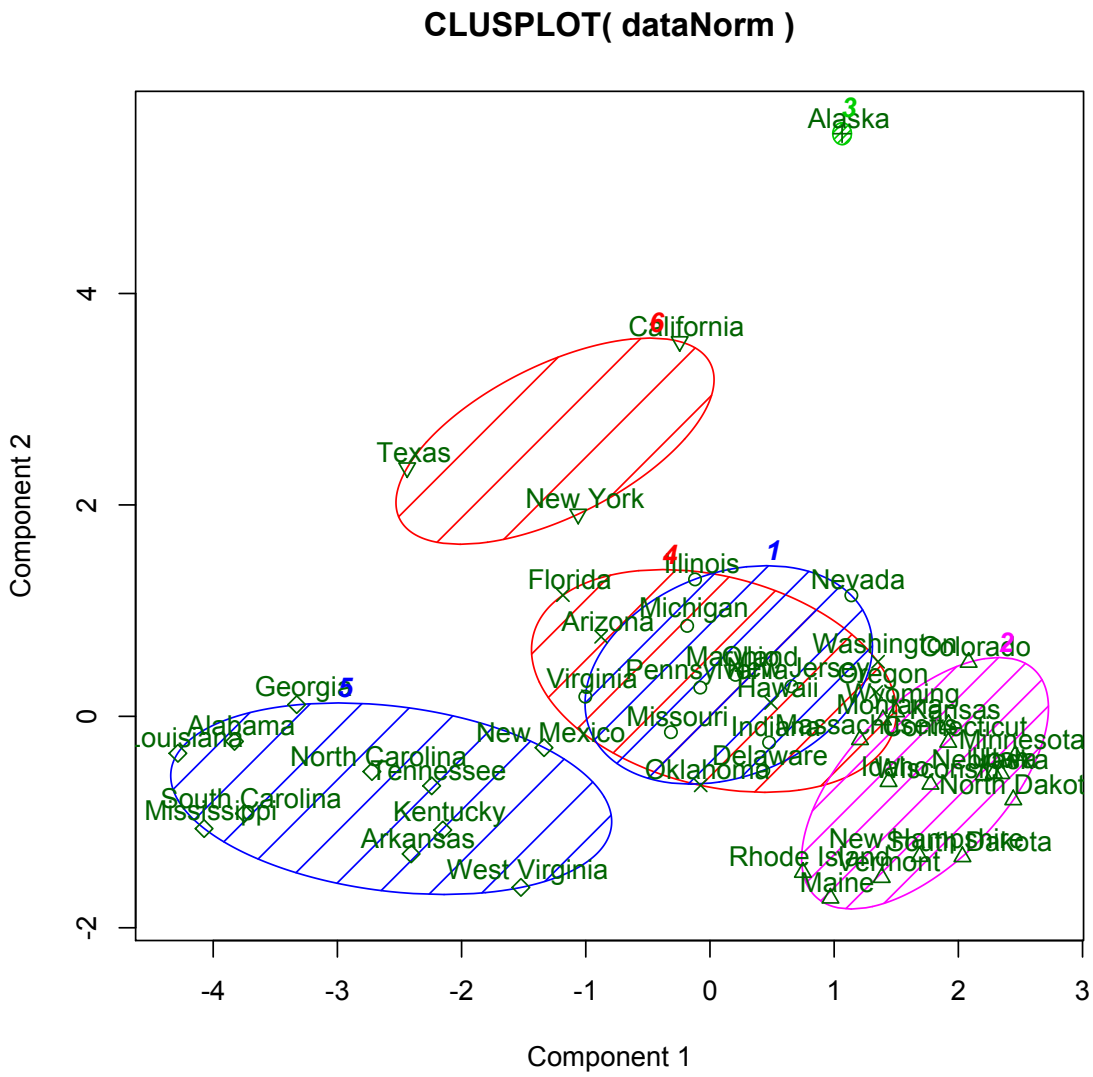
Cluster 3 – Alaska

Cluster 4 – Arizona, Florida, Hawaii, Oklahoma, Oregon, Washington

Cluster 5 – Alabama, Arkansas, Georgia, Louisiana, Kentucky, Mississippi, North Carolina, South Carolina, Tennessee, West Virginia

Cluster 6 – California, New York, Texas

6. Use "clusplot" to plot a 2D representation of the clustering.



7. Analyze the centers of each of these clusters. Can you identify any insight into this clustering?

   I think the biggest thing was that I realized I might have picked too big of a k-value. Groups 12, and 4 all seem pretty cluttered. When I reduced the clustering down to k=5 the graph looked much nicer.

My assignment is (D) Meets Requirements. I was able to do all of the assigned tasks, but I did not do anything extra.