1. Proportionality between correlation-based distance and Euclidean distance (In R file)
2. PVE Calculation
   1. In R file
   2. In R file
3. Hierarchical Clustering
   1. In R file
   2. After cutting the dendrogram, the clusters are as follows:
      1. Alabama, Alaska, Arizona, California, Delaware, Florida, Illinois, Louisiana, Maryland, Michigan, Mississippi, Nevada, South Carolina
      2. Arkansas, Colorado, Georgia, Massachusetts, Missouri, New Jersey, Oklahoma, Oregon, Rhode Island, Tennessee, Texas, Virginia, Washington, Wyoming
      3. Connecticut, Hawaii, Idaho, Indiana, Iowa, Kansas, Kentucky, Maine, Minnesota, Montana, Nebraska, New Hampshire, North Dakota, Ohio, Pennsylvania, South Dakota, Utah, Vermont, West Virginia, Wisconsin
   3. In R file
   4. Scaling the variables makes it such that there appear to be 4 main “clusters” instead of 3 if we cut from the same height as the unscaled dendrogram. The variables should probably be scaled in this case due to the varying units of measure used in each variable。
4. Simulation
   1. In R file
   2. In R file
   3. The K-means clustering with 3 clusters results in clusters that perfectly match the actual class labels.
   4. With 2 clusters, the algorithm merged the 2 clusters that had the least inter-cluster distance between each other into one.
   5. With 4 clusters, the algorithm essentially splits one of the clusters into 2 different ones of approximately equal size.
   6. Performing K-means clustering with 3 clusters on the first 2 principal components instead of the raw data still leads to clusters that perfectly match the actual class labels.
   7. Since scaling the variables to have standard deviation one greatly affects the Euclidean distances between the observations, the results of K-means clustering with scaled variables here leads to significantly poorer results than with their unscaled counterparts. The algorithm becomes unable to discern the originally clearly separated clusters after scaling due to the significantly altered formation of the observations.
5. Gene Testing
   1. In R file
   2. The genes appear to separate into 2 groups in the complete linkage dendrogram, while the average linkage dendrogram gives mixed results, and the single linkage dendrogram portrays the two groups (healthy and sick) as being far in distance but being linked by a single sub-branch that is quite far from the top. Overall, the complete and single linkages result in 2 clusters, while average provides 3 clusters.
   3. In order to determine which genes differ the most across the two groups, we are essentially searching for the genes that *vary* the most in the dataset from each other. For this reason, we may use principal components analysis, as each principal component attempts to describe the largest amount of variation left unexplained by any previous principal components. This means that the variability of any given gene in the dataset can be calculated by summing up the values in the loading vector of that gene. It is important to order the values only after taking their absolute values, as each principal component is unique up to a sign change, meaning that some genes’ values may be negative, and therefore be smaller, while having more weight due to having a larger distance from 0. We may see that the 5 genes with the most variability (as determined by the PCA loadings), and by extension the ones that differ the most, are the 865th, 6th, 911th, 428th, and 624th.