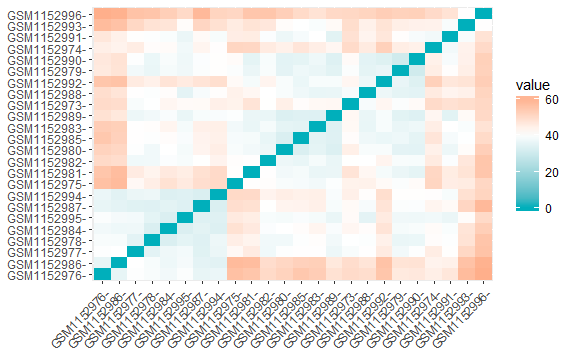
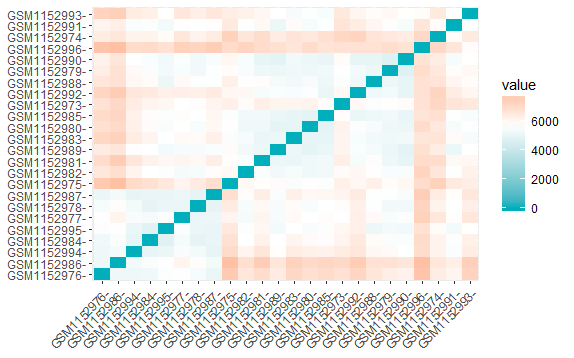
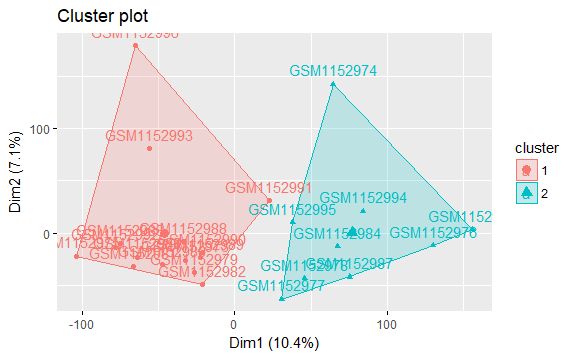
K-Means write up

Unsupervised learning methods were also tried on the data. In particular K-means was tried, to see if the clusters would form around the disease without knowing who actually had the disease. An important metric for K-means clustering is the distance. Figure 1 shows Euclidean distance between all points while Figure 2 shows the Manhattan distance. From these distance plots, one can see that some patients had more similar gene expression rates than others. Kmeans clustering was done using the Eucliean distance, with 50 random starts, taking the best of the 50. Attempting to split the data into 2 clusters produced clusters of size 9 and 15, however the patients in each clusters had a similar rate of the disease compared to the original cluster. PCA was done to make the data view able and Figure – shows the 2 clusters. Expanding the search to look from 2 to 8 clusters, the presence of the disease did not seem to be hugely impactful. Using 6 or more clusters, the patents who did not have the disease seemed to be clustered more; however, the disease still did not seem to be a significant factor. Table – shows the results of clusters with various sizes. Here patients who did not have the disease are denoted as ‘1’, while those who did are ‘0’.

  
Illustration 1: Eucliean Distance

  
Illustration 2: Manhattan Distance



[1] "Number of Cluster : 2"

Actual: 0 0 0 0 0 1 0 1 1[1] "Cluster: 1 mean: 0.333333333333333 "

Actual: 0 0 0 0 0 0 0 0 1 0 1 0 1 0 1[1] "Cluster: 2 mean: 0.266666666666667 "

[1] "==========================="

[1] "Number of Cluster : 3"

Actual: 0 0 0 1[1] "Cluster: 1 mean: 0.25 "

Actual: 0 0 0 0 1 0 1 1[1] "Cluster: 2 mean: 0.375 "

Actual: 0 0 0 0 0 0 0 0 1 0 1 1[1] "Cluster: 3 mean: 0.25 "

[1] "==========================="

[1] "Number of Cluster : 4"

Actual: 0 0[1] "Cluster: 1 mean: 0 "

Actual: 0 1[1] "Cluster: 2 mean: 0.5 "

Actual: 0 0 0 0 0 0 0 0 1 0 1 1[1] "Cluster: 3 mean: 0.25 "

Actual: 0 0 0 0 1 0 1 1[1] "Cluster: 4 mean: 0.375 "

[1] "==========================="

[1] "Number of Cluster : 5"

Actual: 0 0 0 0 0 0 0[1] "Cluster: 1 mean: 0 "

Actual: 0 1[1] "Cluster: 2 mean: 0.5 "

Actual: 0 0 0 0 1 0 1 1[1] "Cluster: 3 mean: 0.375 "

Actual: 0 0[1] "Cluster: 4 mean: 0 "

Actual: 0 0 1 1 1[1] "Cluster: 5 mean: 0.6 "

[1] "==========================="

[1] "Number of Cluster : 6"

Actual: 0 1[1] "Cluster: 1 mean: 0.5 "

Actual: 0[1] "Cluster: 2 mean: 0 "

Actual: 0 0 0 0 1 0 1 1[1] "Cluster: 3 mean: 0.375 "

Actual: 0 1 1 1[1] "Cluster: 4 mean: 0.75 "

Actual: 0 0 0 0 0 0 0[1] "Cluster: 5 mean: 0 "

Actual: 0 0[1] "Cluster: 6 mean: 0 "

[1] "==========================="

[1] "Number of Cluster : 7"

Actual: 0 1[1] "Cluster: 1 mean: 0.5 "

Actual: 0 0 0 0 1 1[1] "Cluster: 2 mean: 0.333333333333333 "

Actual: 0 0[1] "Cluster: 3 mean: 0 "

Actual: 0 0 0 0 0 0 0[1] "Cluster: 4 mean: 0 "

Actual: 0 1 1 1[1] "Cluster: 5 mean: 0.75 "

Actual: 0 1[1] "Cluster: 6 mean: 0.5 "

Actual: 0[1] "Cluster: 7 mean: 0 "

[1] "==========================="

[1] "Number of Cluster : 8"

Actual: 0 0 0 0 1 1[1] "Cluster: 1 mean: 0.333333333333333 "

Actual: 0 1[1] "Cluster: 2 mean: 0.5 "

Actual: 0 1 1 1[1] "Cluster: 3 mean: 0.75 "

Actual: 0 1[1] "Cluster: 4 mean: 0.5 "

Actual: 0[1] "Cluster: 5 mean: 0 "

Actual: 0[1] "Cluster: 6 mean: 0 "

Actual: 0 0 0 0 0 0[1] "Cluster: 7 mean: 0 "

Actual: 0 0[1] "Cluster: 8 mean: 0 "

[1] "==========================="