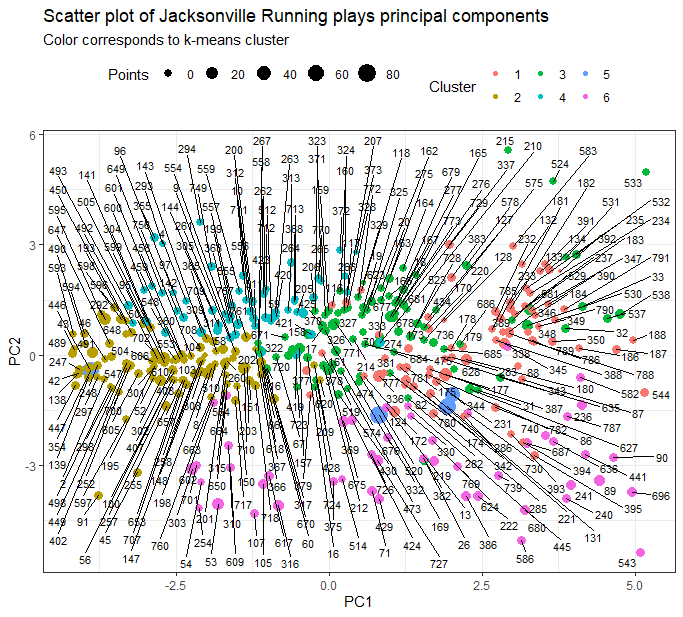
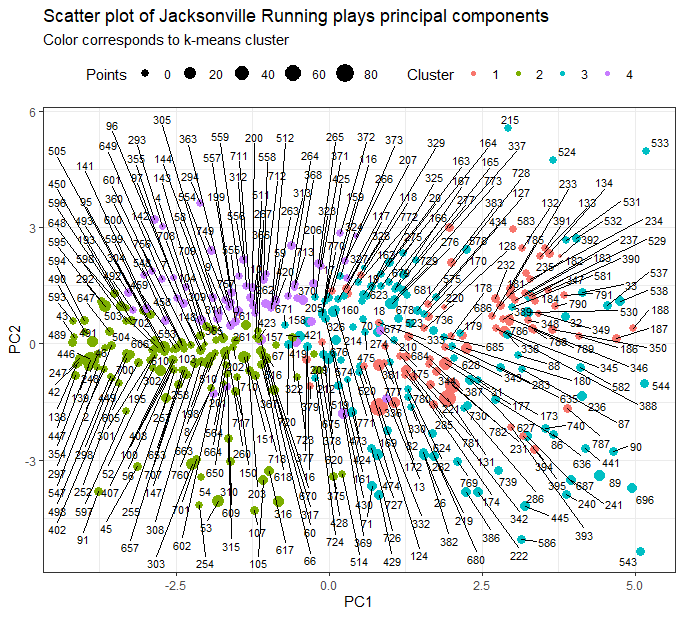
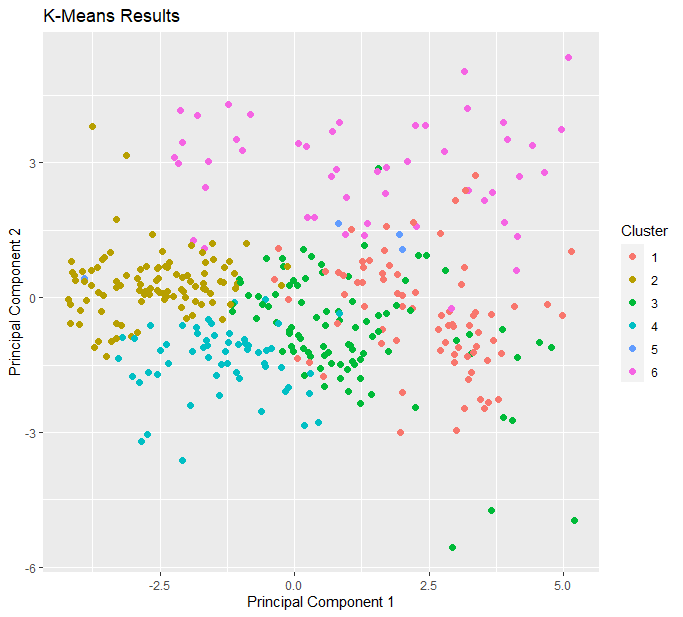
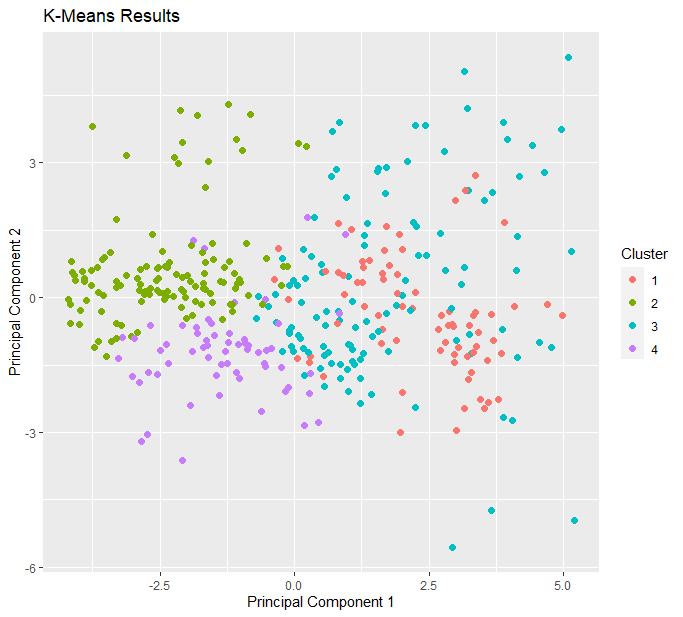
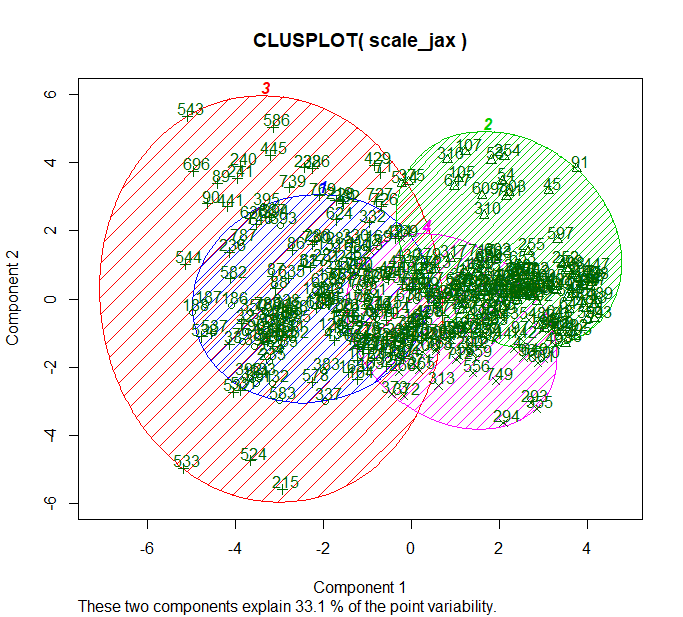
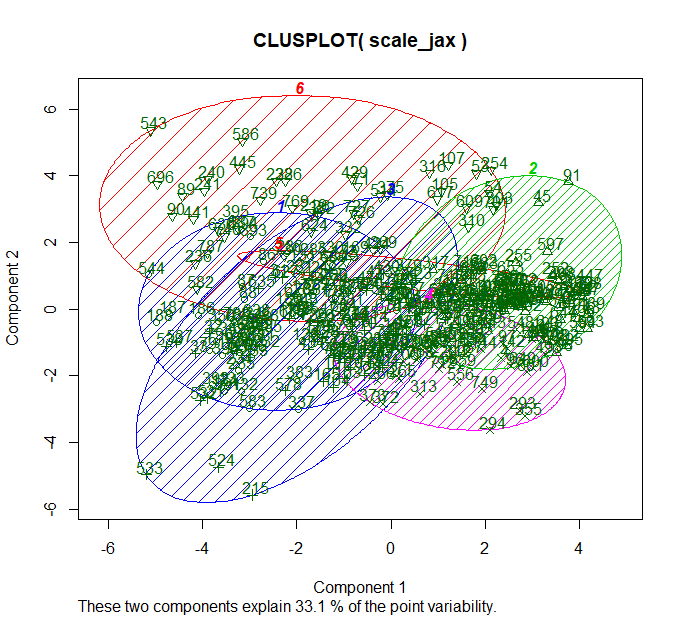
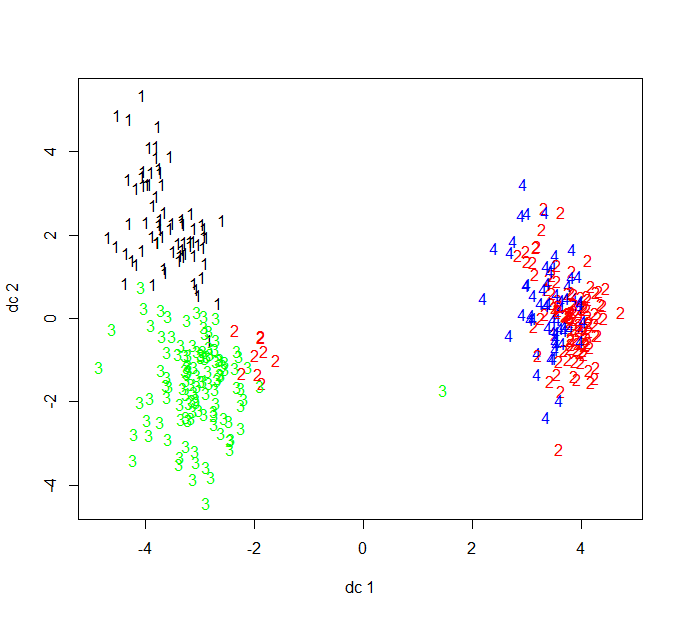
Joshua Harkness <https://github.com/darkhark/2019JAX_PlayByPlay> STA6704 Professor Aaron Smith 5/17/2020

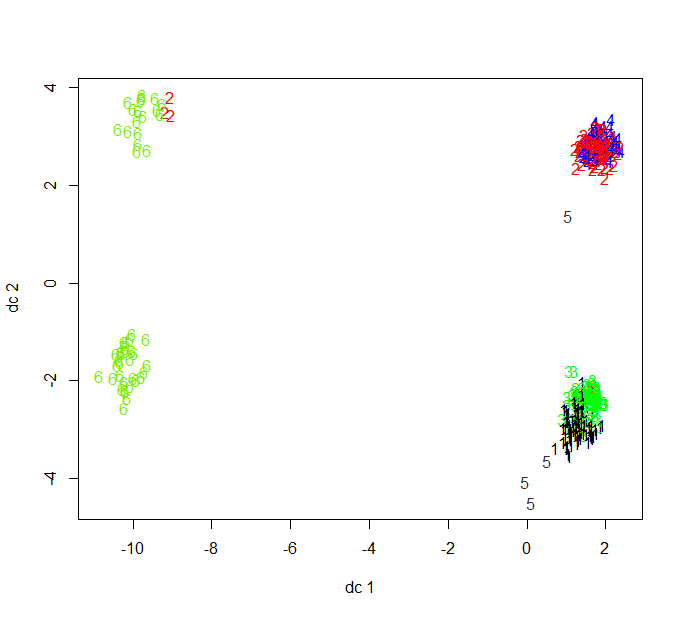


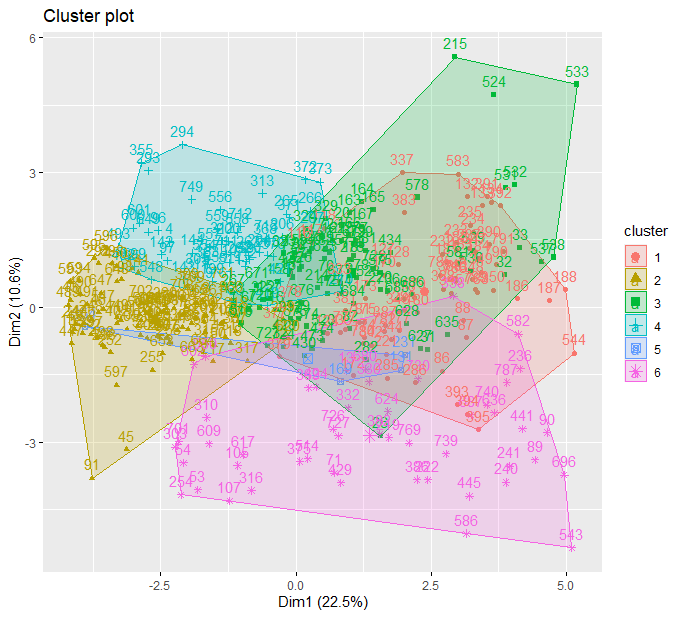
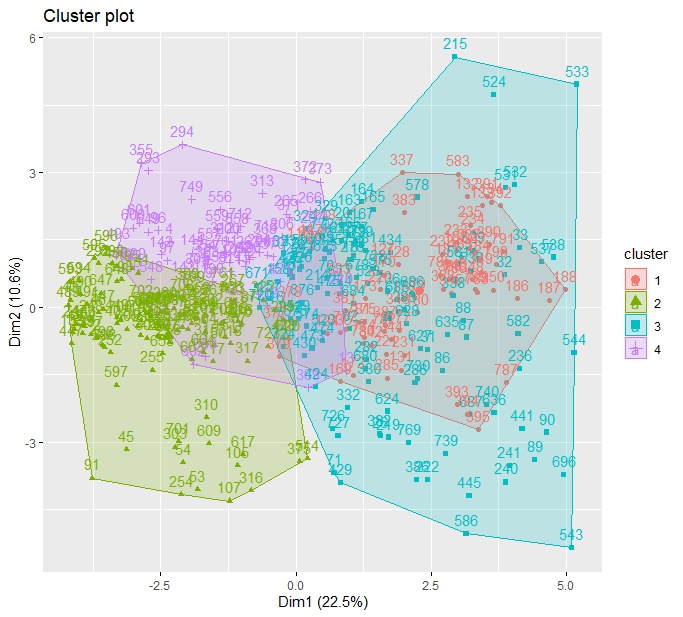


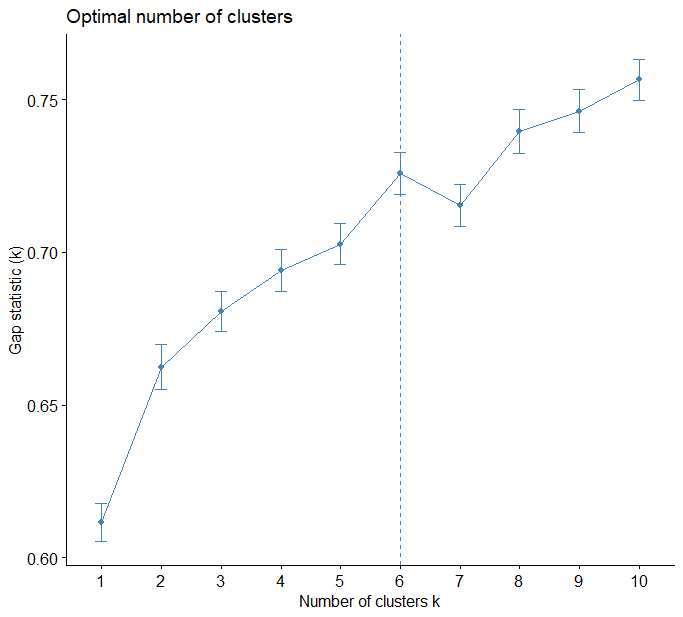
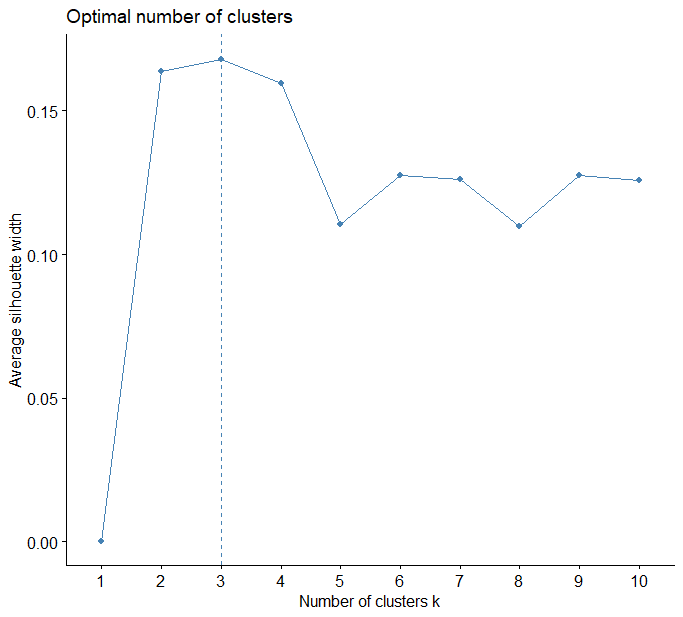
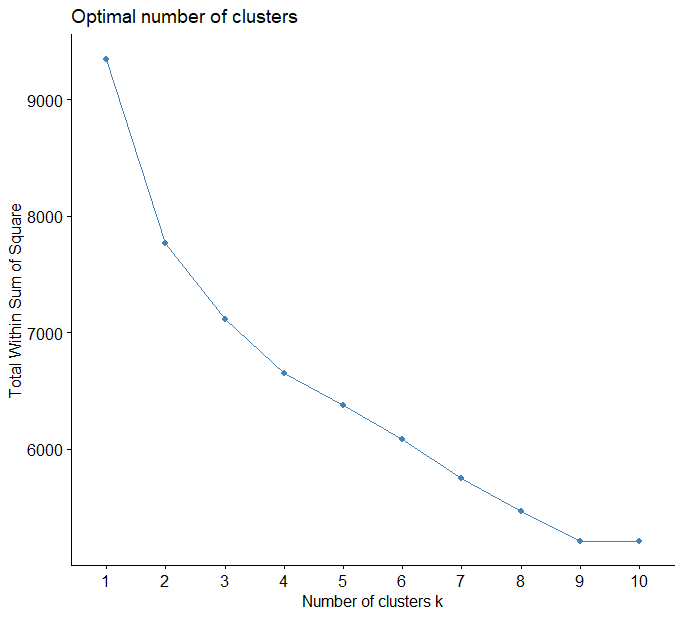


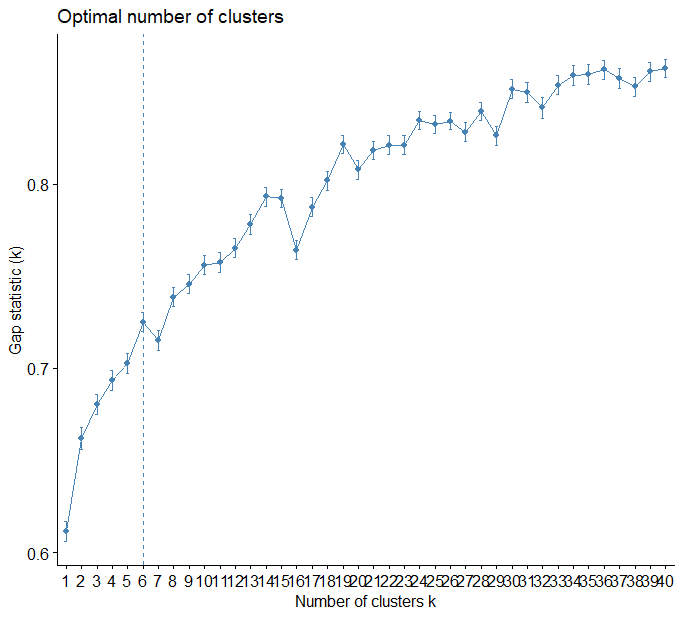




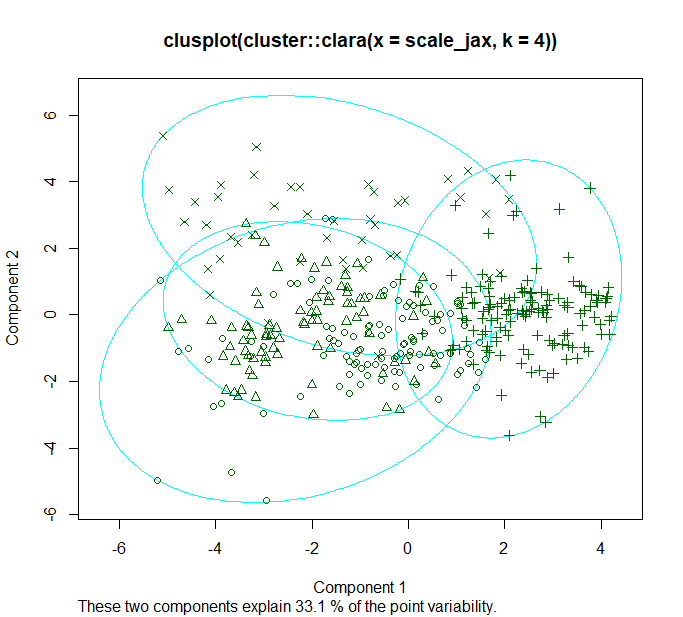


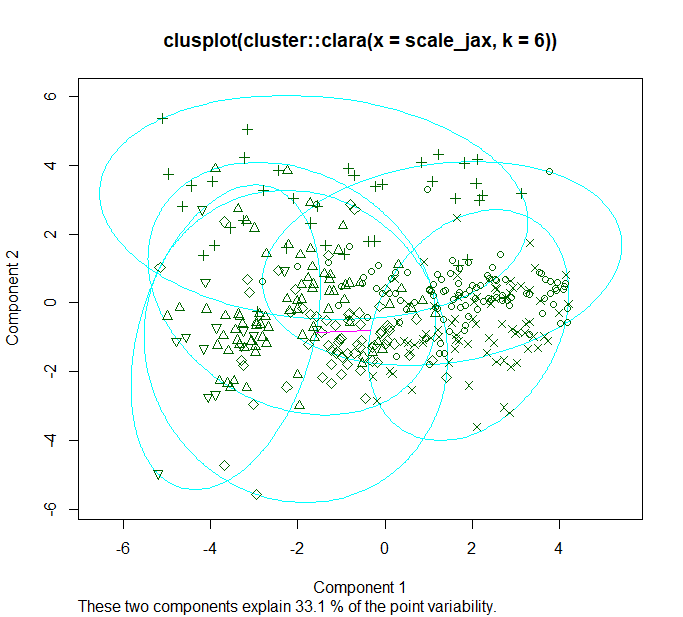


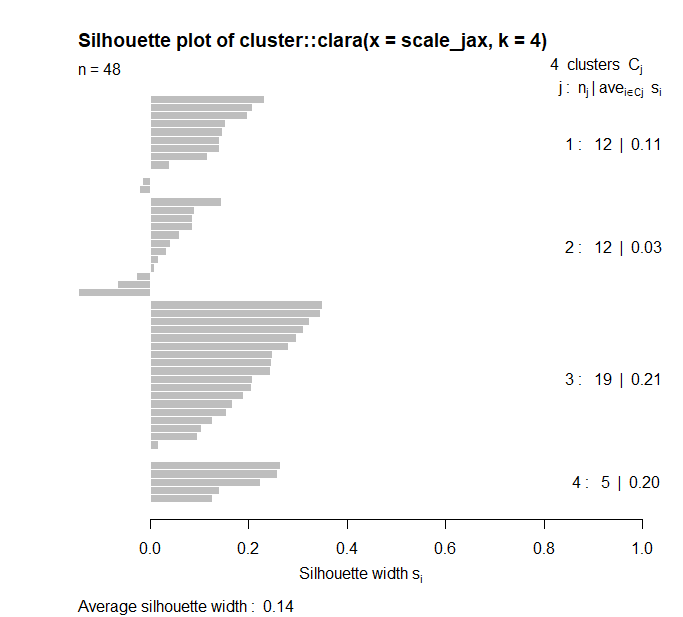


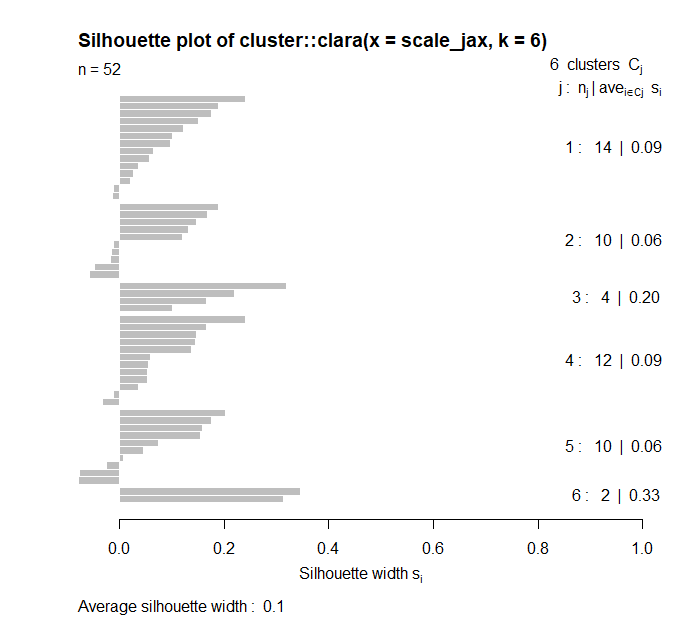


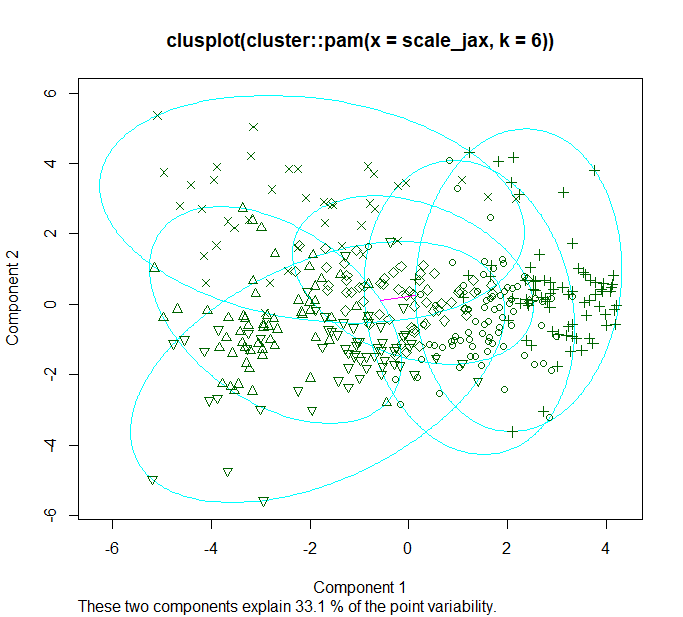
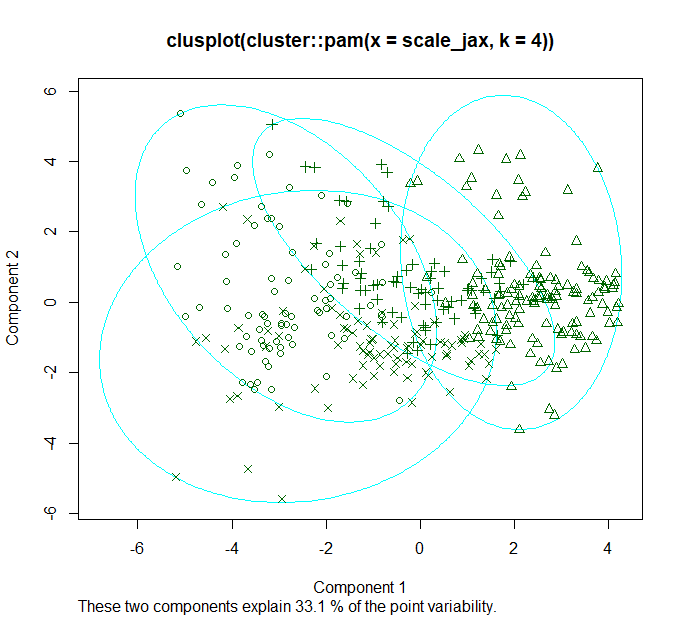
|  |  |
| --- | --- |
| $n  [1] 375  $cluster.number  [1] 6  $cluster.size  [1] 75 100 90 57 4 49  $min.cluster.size  [1] 4  $average.distance  [1] 5.786322 4.471316 5.937342 4.762156 5.941208 6.707538  $median.distance  [1] 5.756490 4.470899 5.331958 4.712757 5.881097 6.565785  $separation  [1] 2.069646 1.437844 1.437844 1.518093 3.493685 2.795147  $average.toother  [1] 7.307910 6.888010 6.960006 6.658411 10.157847 7.986869  $entropy  [1] 1.617567  $wb.ratio  [1] 0.7602157  $ch  [1] 39.64586 | $n  [1] 375  $cluster.number  [1] 4  $cluster.size  [1] 75 122 117 61  $min.cluster.size  [1] 61  $average.distance  [1] 5.891189 5.010590 6.891958 4.998280  $median.distance  [1] 5.644135 4.888192 6.729412 4.907953  $separation  [1] 1.743447 1.518093 1.743447 1.518093  $average.toother  [1] 7.336993 7.047929 7.305178 6.738753  $entropy  [1] 1.34602  $wb.ratio  [1] 0.8095848  $ch  [1] 43.59973 |

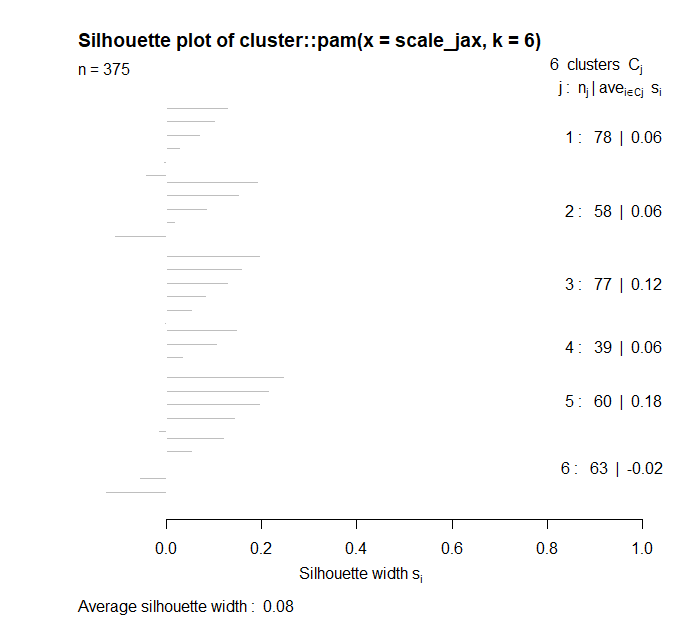


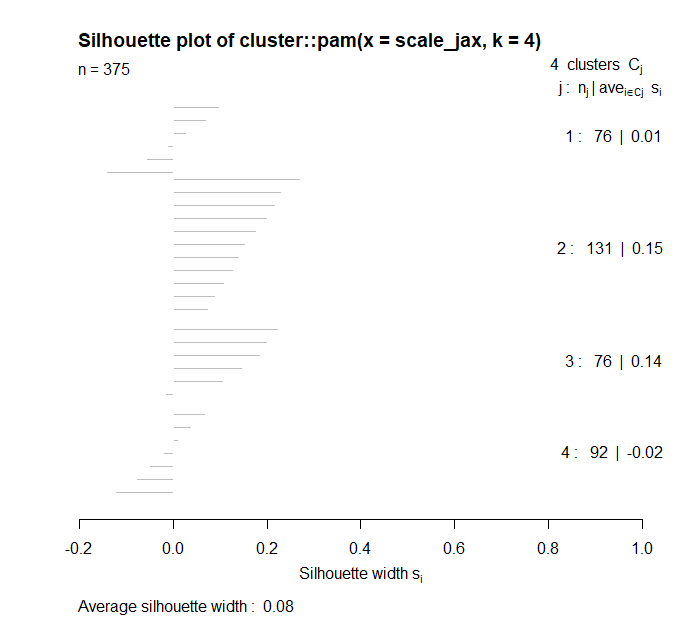




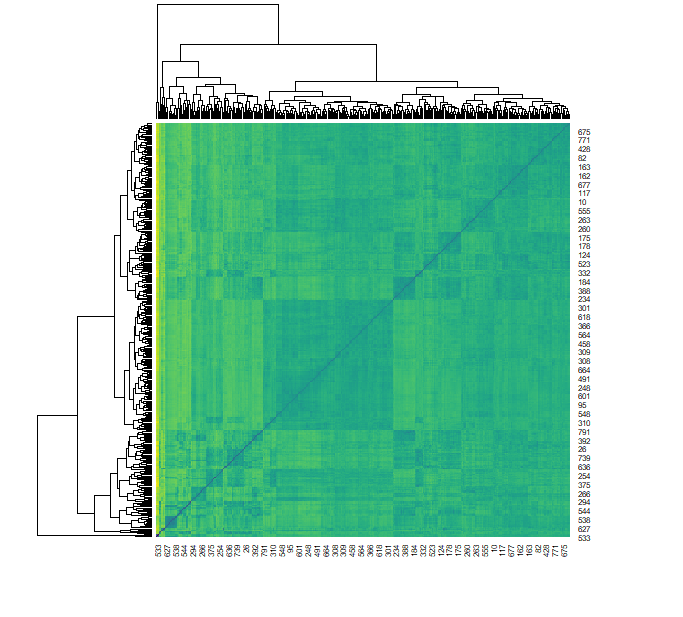
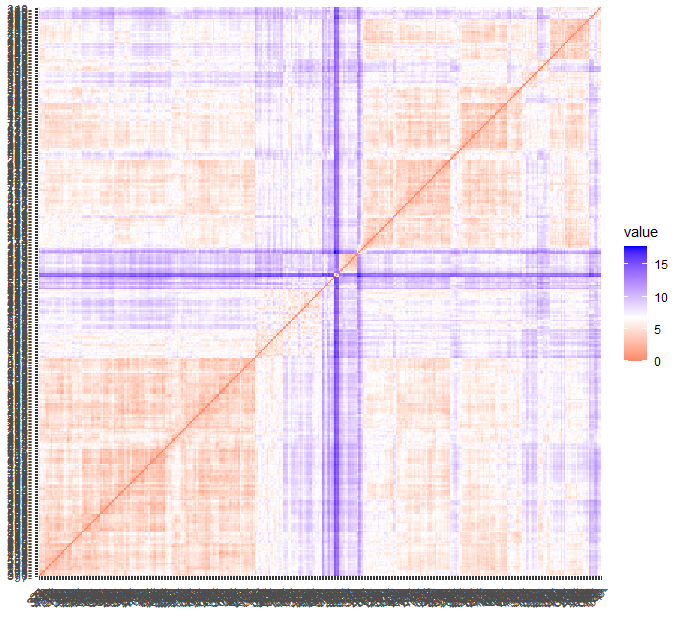


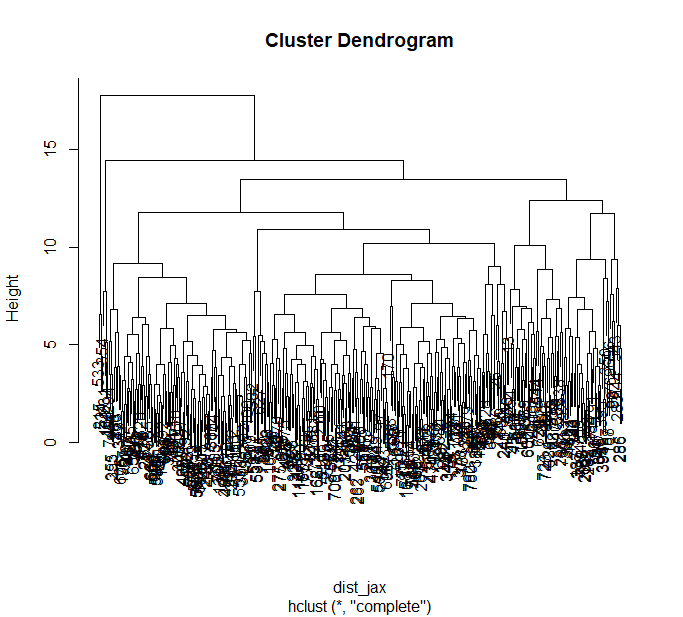


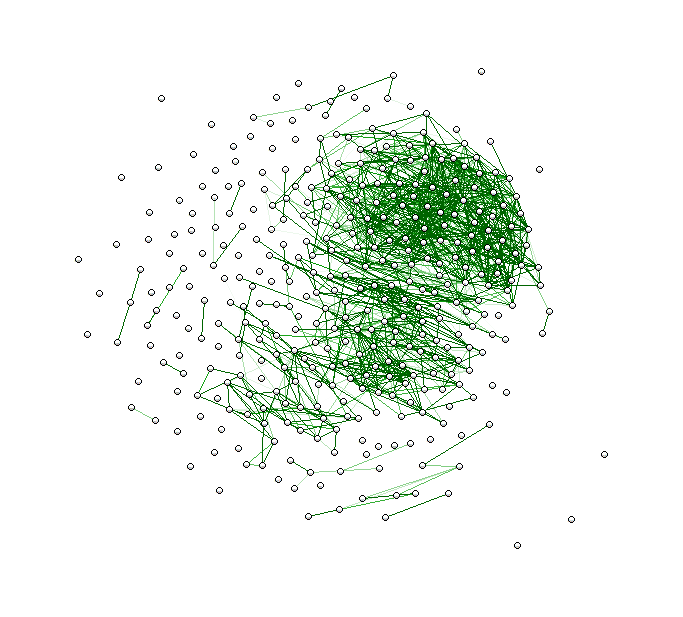


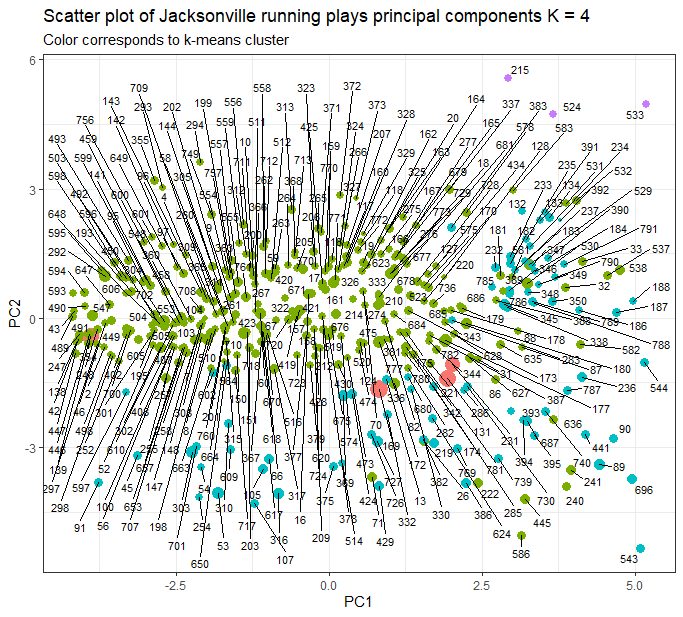
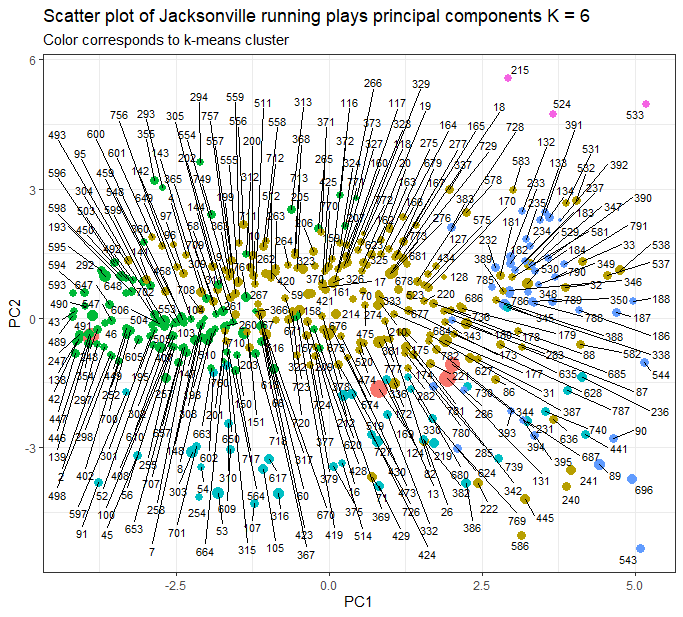


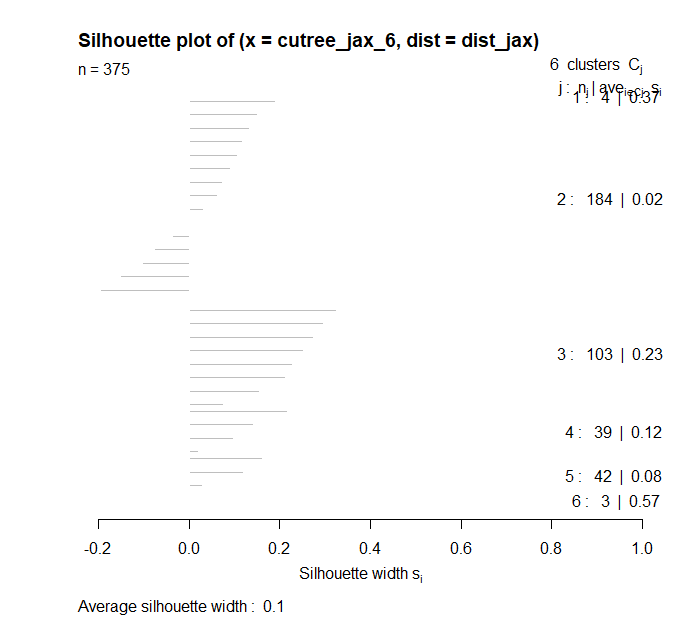
The above visualizations are based on k = 6 (left) and k = 4 (right). These two k values were based on the graphs provided on page four. The silhouette suggests a value of 3, but ideally, we would like more clusters for this data. Looking at the elbow method and confirming with the silhouette, it was safe to increase the cluster number to 4 without a significant dip in the silhouette’s value. The gap statistic suggested a higher cluster value of 6, so that was also taken into consideration. Throughout all the visualizations, there really does not seem to be any great clustering. Many of the clusters overlap one another and no strong distinguishing clusters appear. I believe the k-value at 4 appears to better for a few reasons though. The first page shows that the principal components clusters appear more segregated. On page two, the clusters are less congested, although one cluster completely absorbs another. In page five’s statistics, we can see that when k=6, one of the clusters only possesses 4 values, which is almost meaningless. Also, the average to other distances are larger in when k=4. Finally, when we look at the silhouettes produced by clara, a large data partitioning algorithm, there appears to be less values that do not belong to a cluster. Interestingly, pam, which partitions around medoids, argues that more values fit into the k=6. While both k values have weak arguments for either of them, k=4 appears to be the stronger.

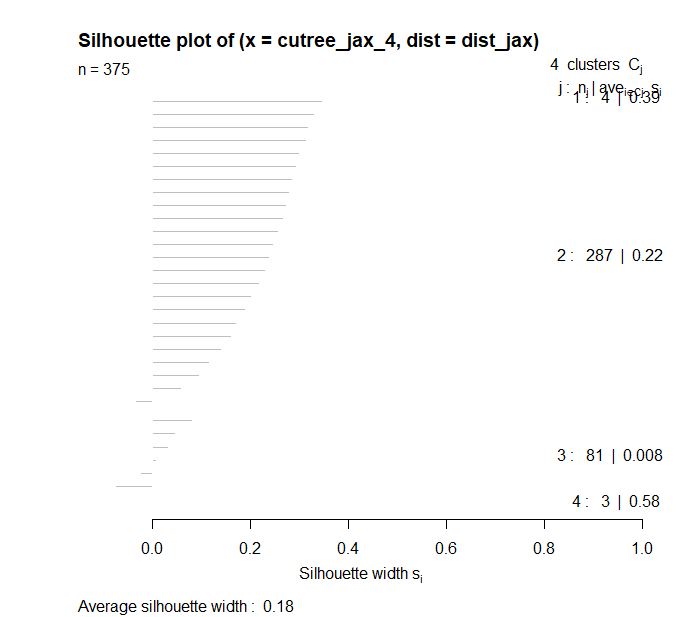


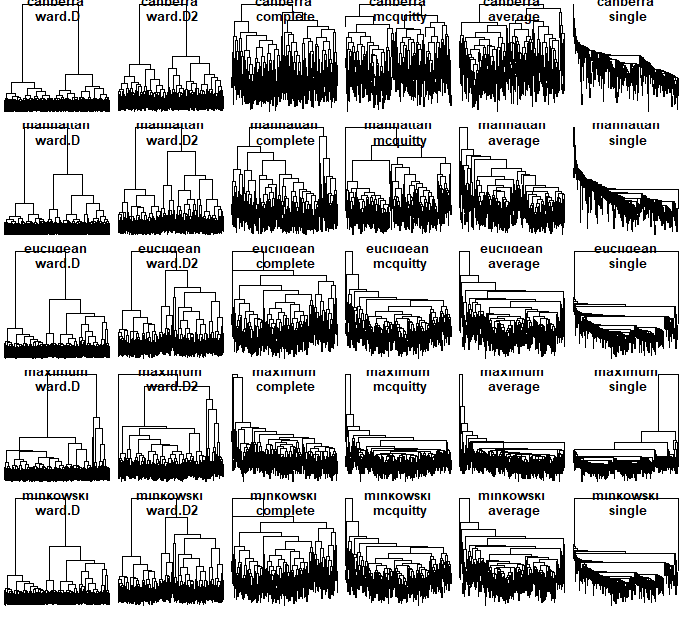


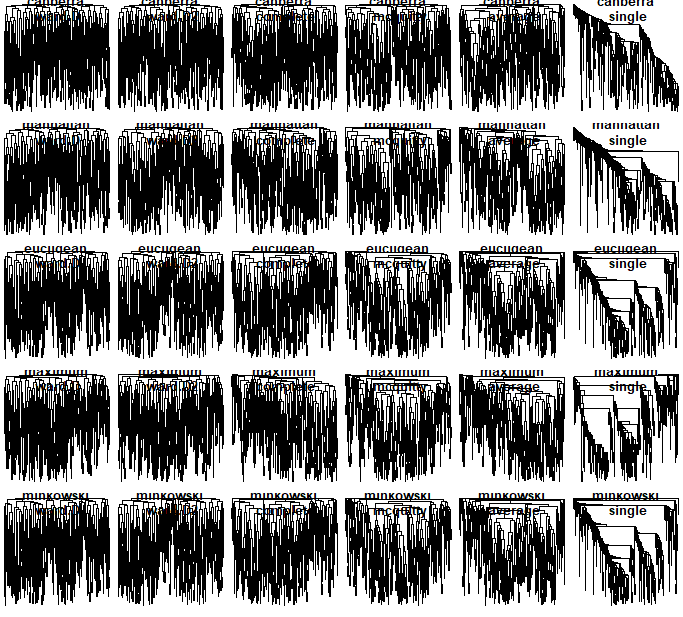


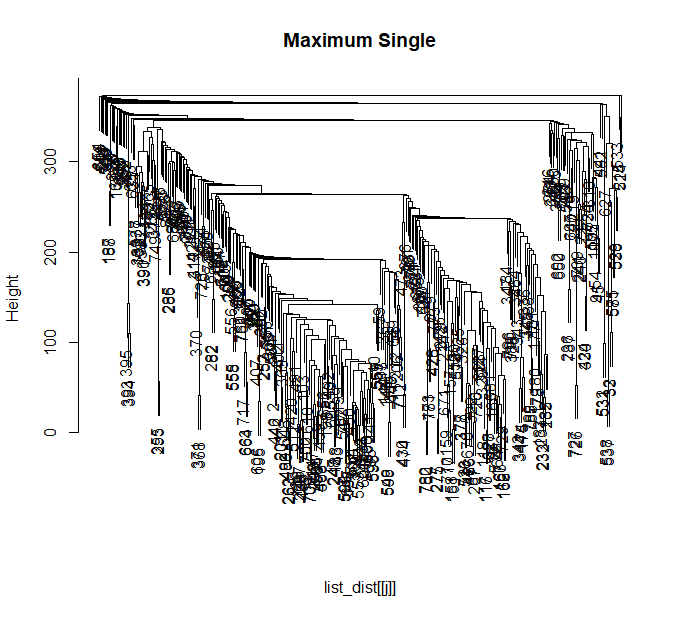
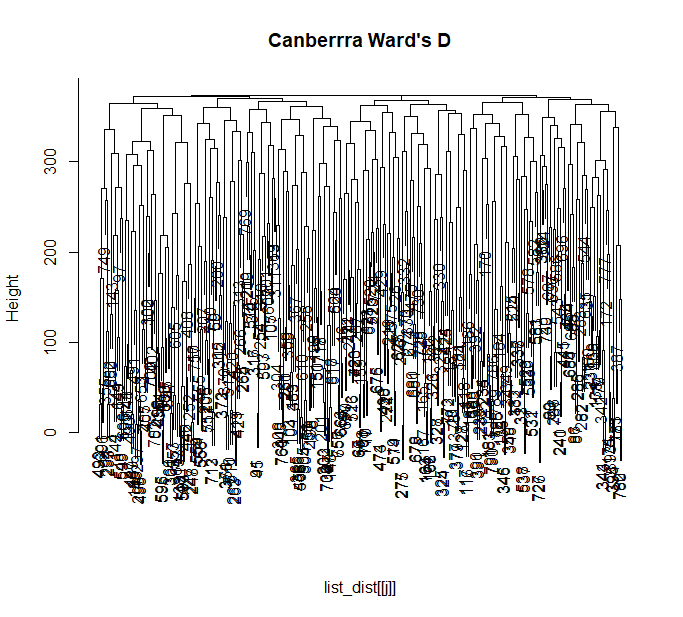


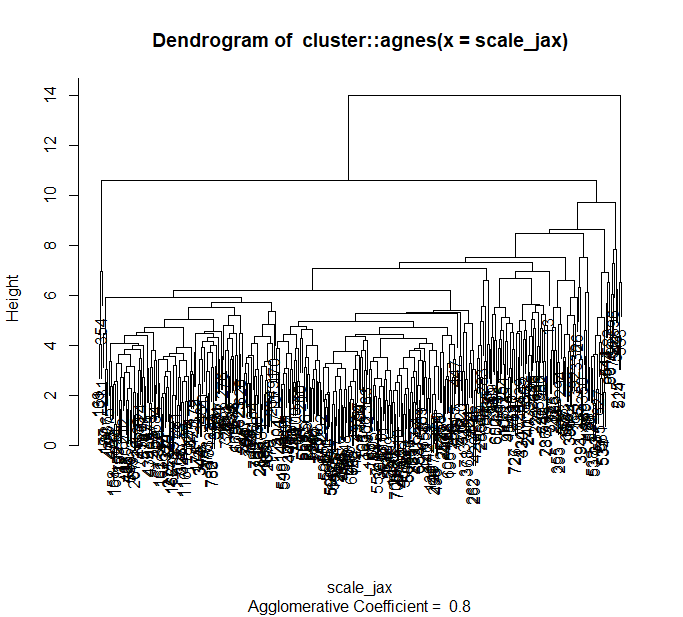
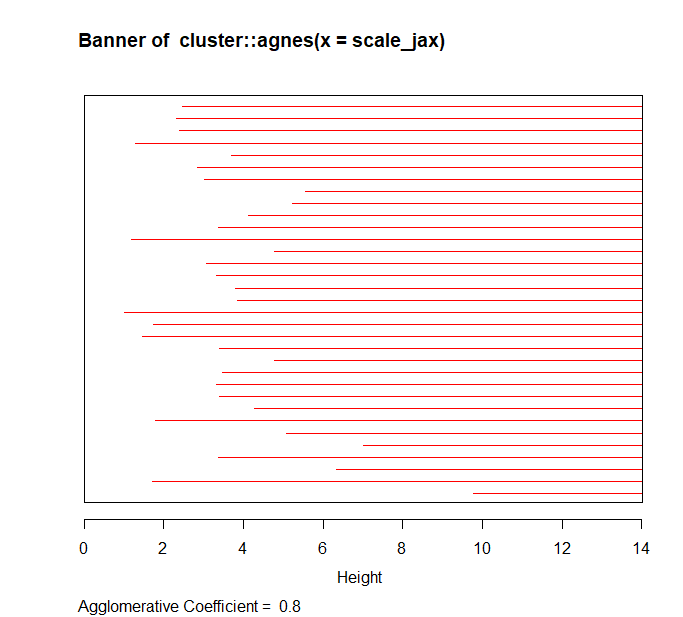


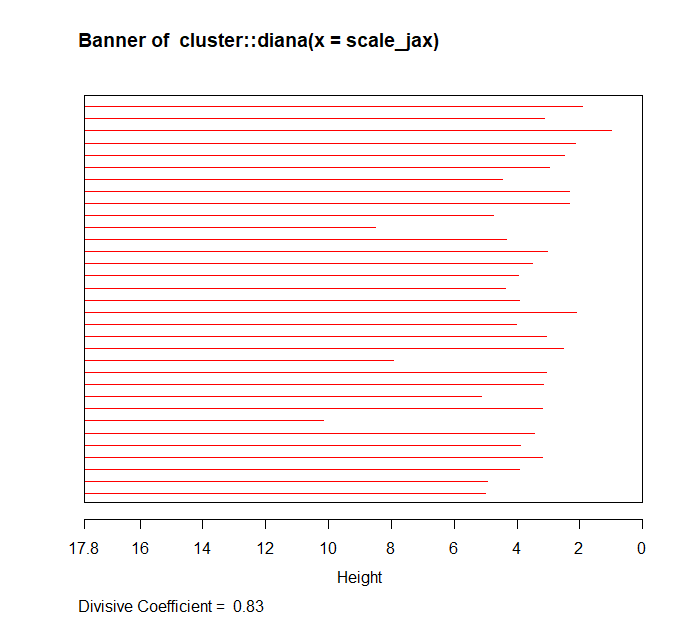


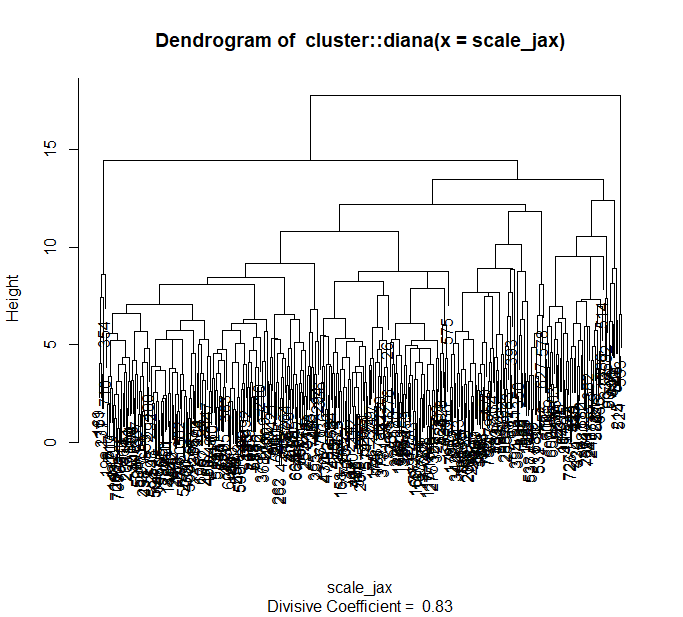












On page nine, the temperature maps indicate that some groupings do exist. While There does not appear to be any triangles, many blue squares are visible, which indicates those groups have small distances between one another. The cluster dendrogram on page nine was used to determine the k values on page 10. Two slices seemed more reasonable to me, which happened to align with the k values found from kmeans earlier (4 and 6). Using the silhouettes to the right of each scatter plot on page 10, I determined a k-value of four seemed more appropriate due to less values appearing out of their cluster. Using a matrix of different coefficients from different plots, the ward.D and Canberra algorithms returned the highest coefficient at .709, indicating that the plot would be good for determining clusters. The lowest coefficient in the matrix was in the single, Minkowski plot with a value of .557. The low coefficient indicates the plot would be best suited for outlier detection.