STAT 542 Assignment 5 kakilai2

Question 1

I have used the R package *hmm.discnp* to fit hidden Markov model. For each of the sequence, AIC is used to select the best model among the HMM formed with states 1, 2, and 3.

By definition, AIC = -2 logL + 2p whereas for HMM, p = m^2 + km – 1, m is the number of states, and k is the number of parameters.

The output in R is given below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sequence | 1 Hidden State | 2 Hidden States | | 3 Hidden States | |
|  | AIC | AIC | % Increase | AIC | % Increase |
| 1 | 992.0096 | **976.3713** | **-1.5764** | 980.4732 | 0.4201 |
| 2 | **1101.103** | 1110.292 | 0.8346 | 1124.236 | 1.2559 |
| 3 | 1098.827 | 1039.076 | -5.438 | **1040.261** | **0.1149** |

For each of the states, the percentage increase of AIC over that of 1 state less is calculated to give a clearer view in the AIC change. My best guess of the number of states of each sequence is as follows:

For Sequence 1, AIC is lowest at k = 2, so it is likely that sequence 1 has 2 hidden states.

For Sequence 2, AIC is lowest at k = 1, so it is likely that sequence 2 has 1 hidden state.

For Sequence 3, AIC is lowest at k = 2 , and AIC of k = 3 is very close to that of k = 2 (just 0.11% increase). Therefore, it is possible that sequence 3 has 2 or 3 hidden states. In this case, I have already guessed sequence 1 has 2 hidden states, hence I would guess sequence 3 has 3 states here.

For sequence 1,

Transition matrix A of the hidden chain :

and the emission matrix B :

For sequence 2,

Transition matrix A does not exist because there is no transition.

and the emission matrix B :

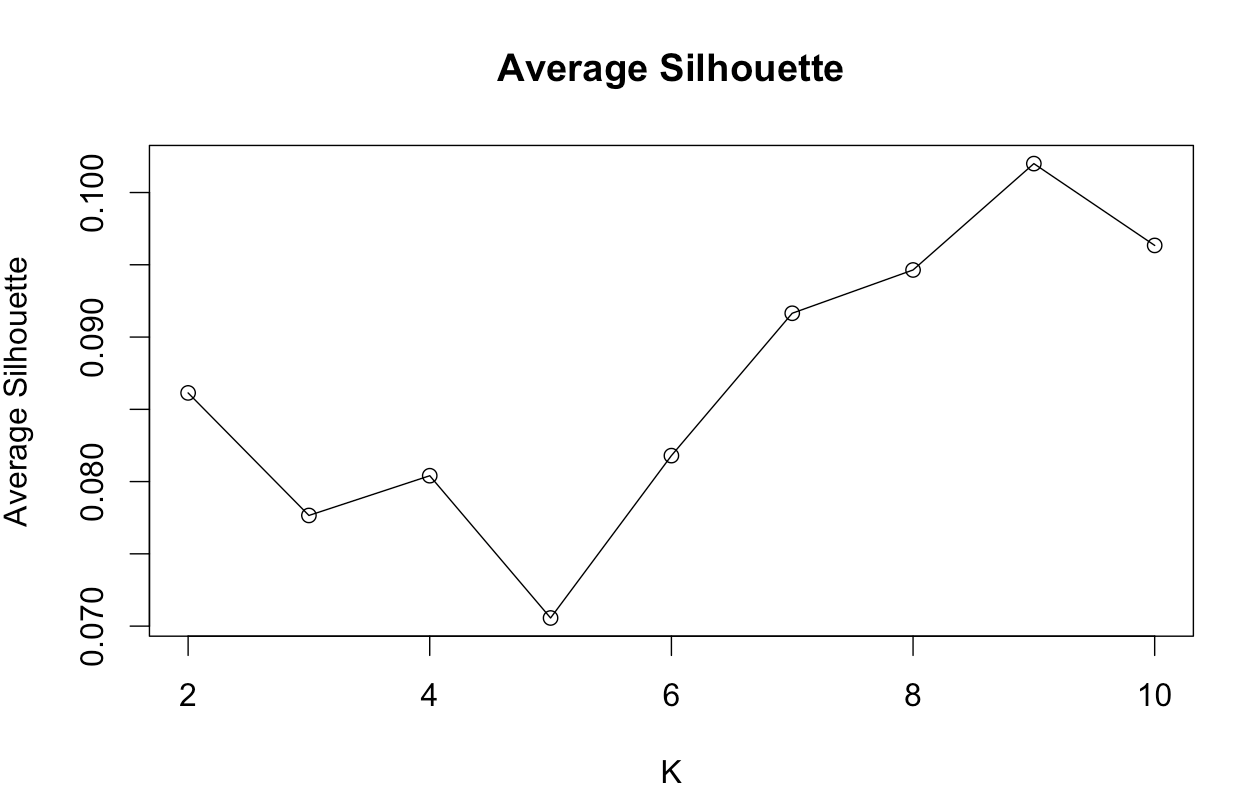
For sequence 3,

Transition matrix A of the hidden chain:

and the emission matrix B :

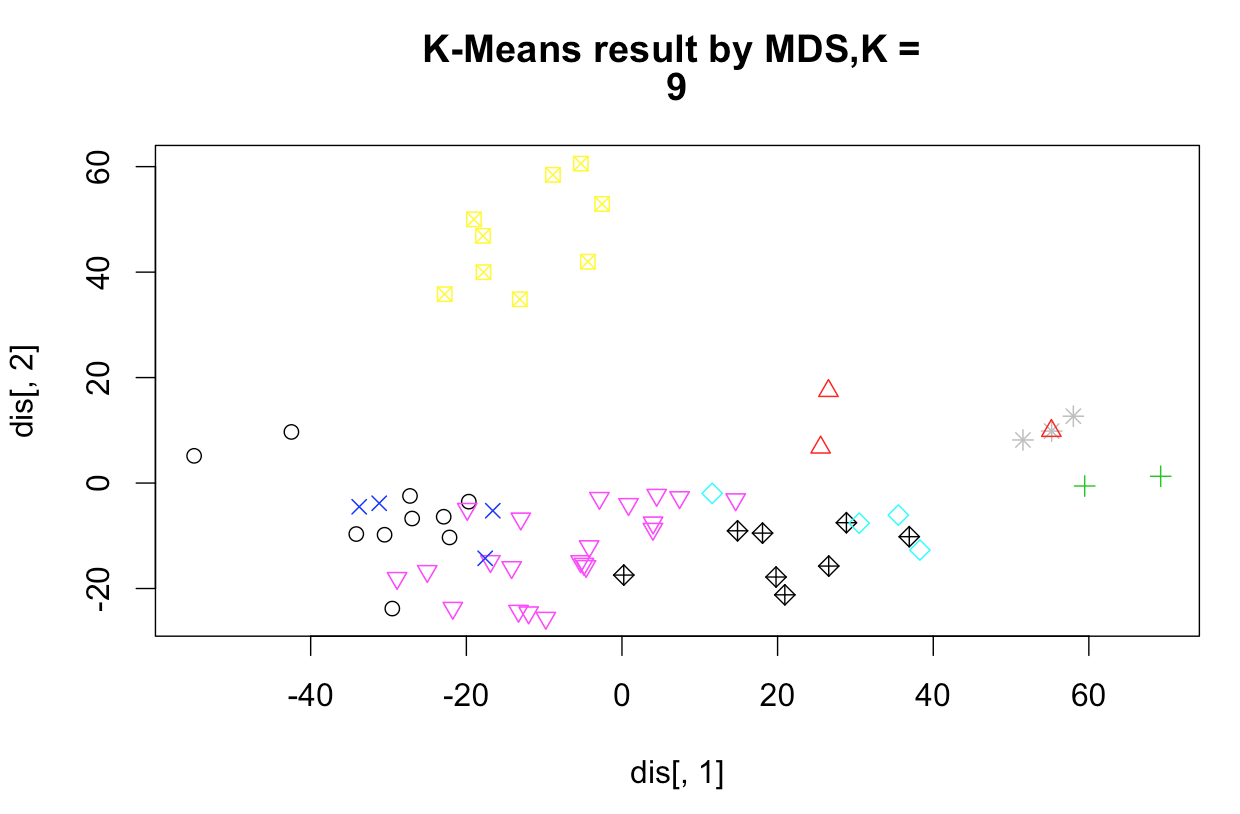
Question 2

1. Using K-means clustering, The plot of average silhouettes versus K = 2, . . . , 10 is shown as below:

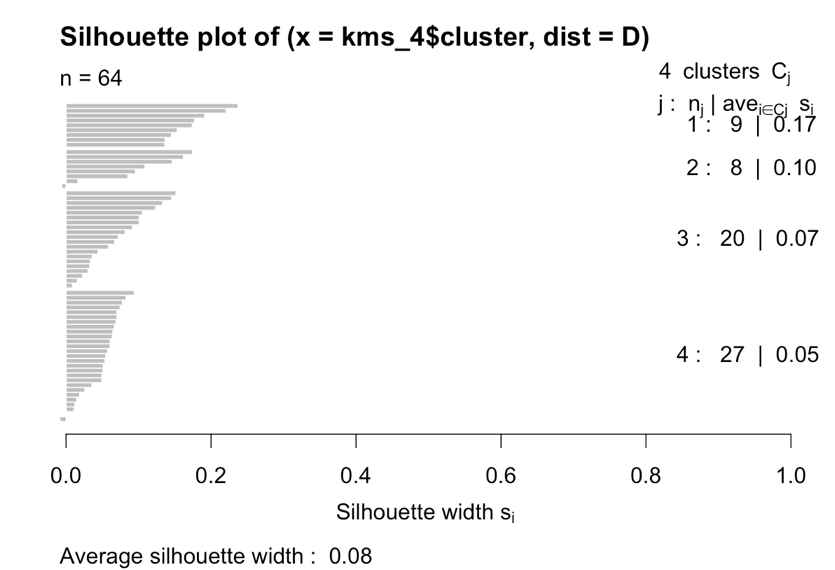


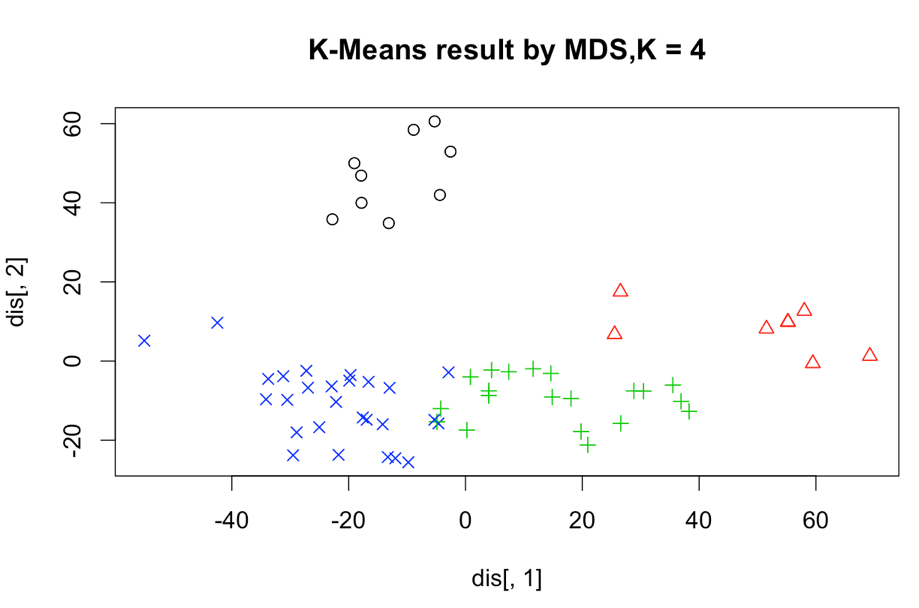
Based on silhouette statistics, the number of clusters should be 9 which give maximum silhouettes coefficient. Below is the plot of silhouettes and clustering result using MDS to display the data with K = 9:





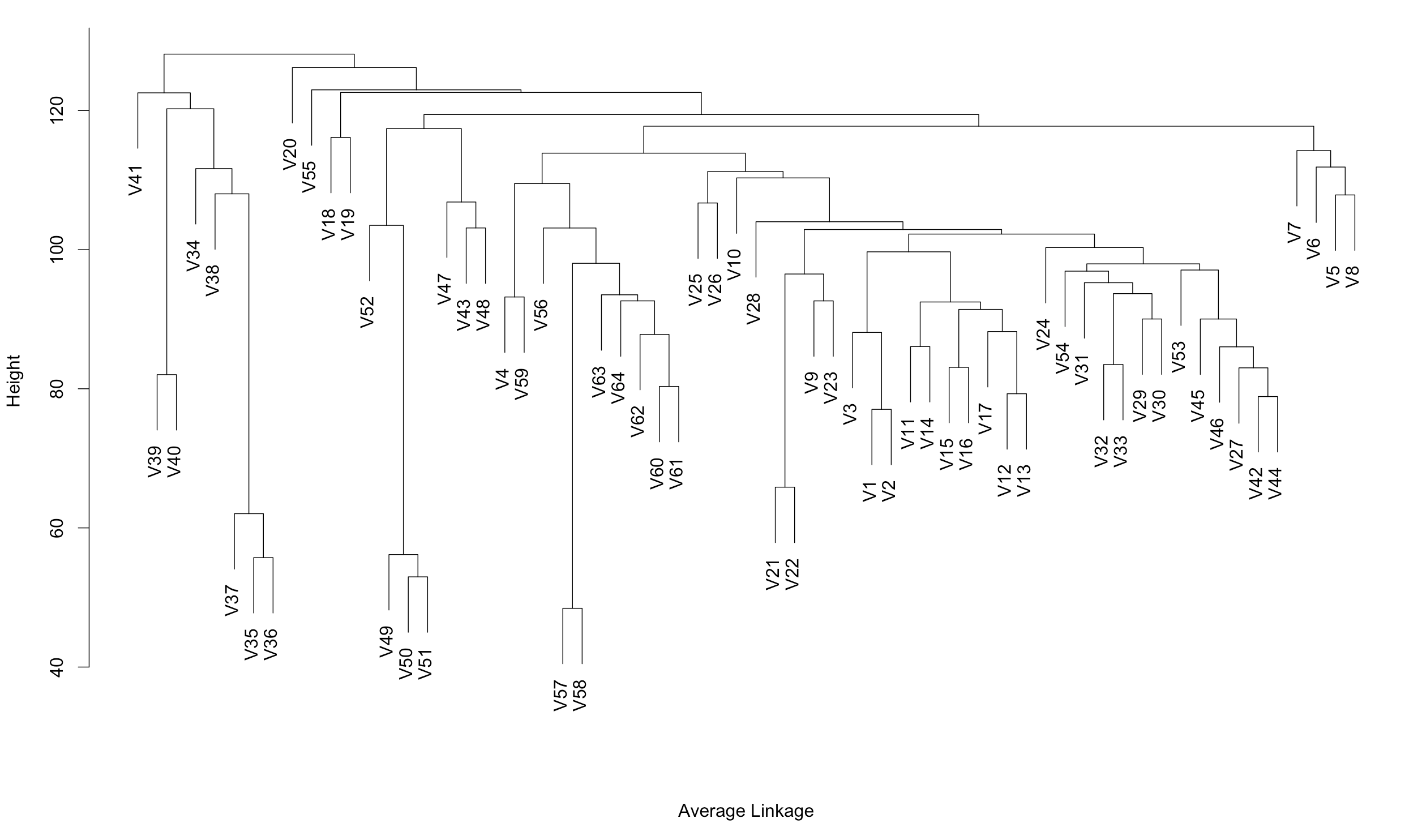
To keep the number of cluster low, cluster number = 4 could also be considered since it gives a local maximum in the plot of average Silhouette. The respective plot of silhouettes and clustering result is also shown below:



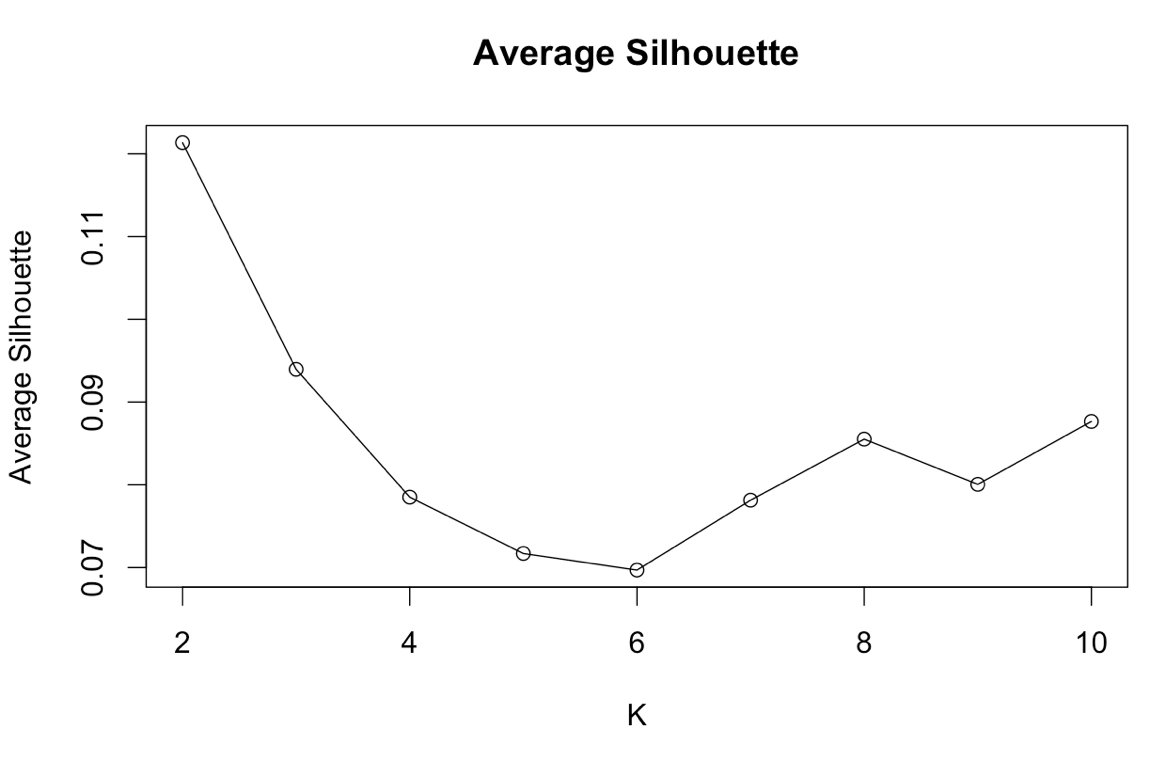


From the about plot, we can see that the clustering result is nice with K\* = 4.

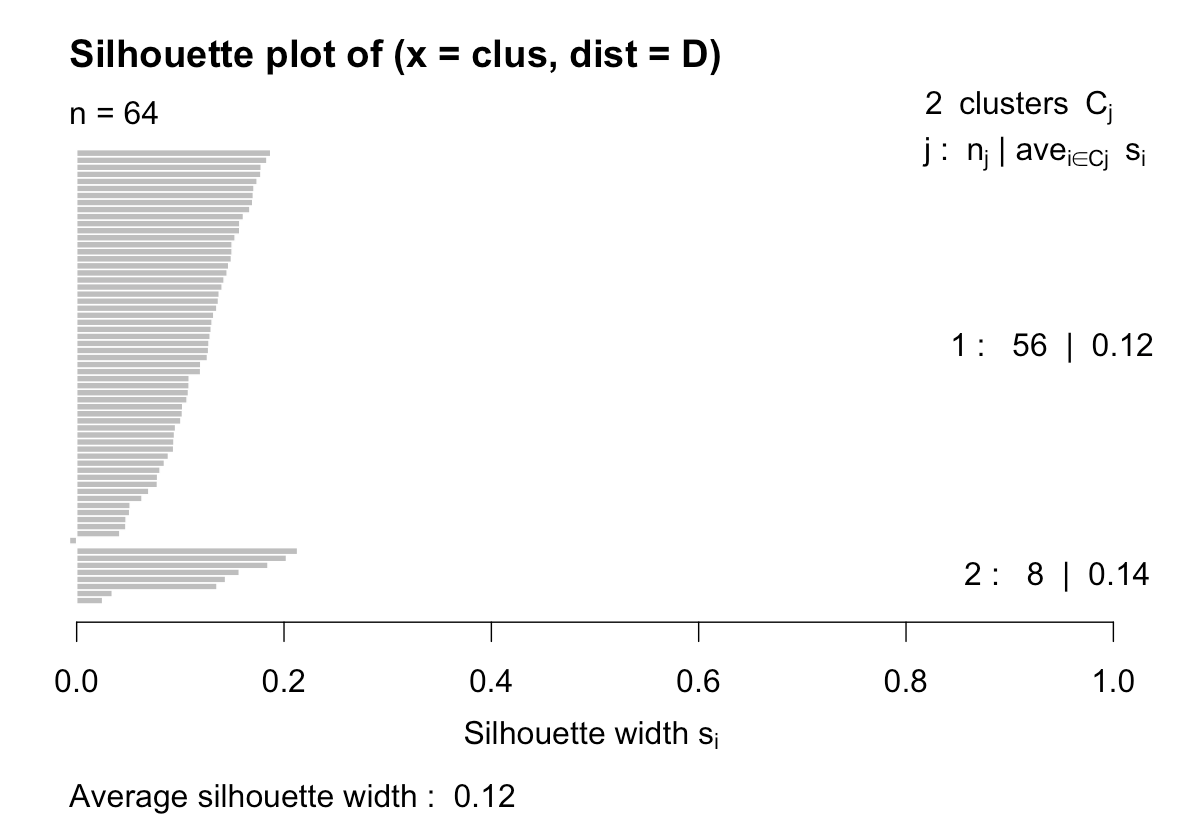
1. Repeating part (a) with average linkage hierarchical clustering, a dendrogram is generated as below:

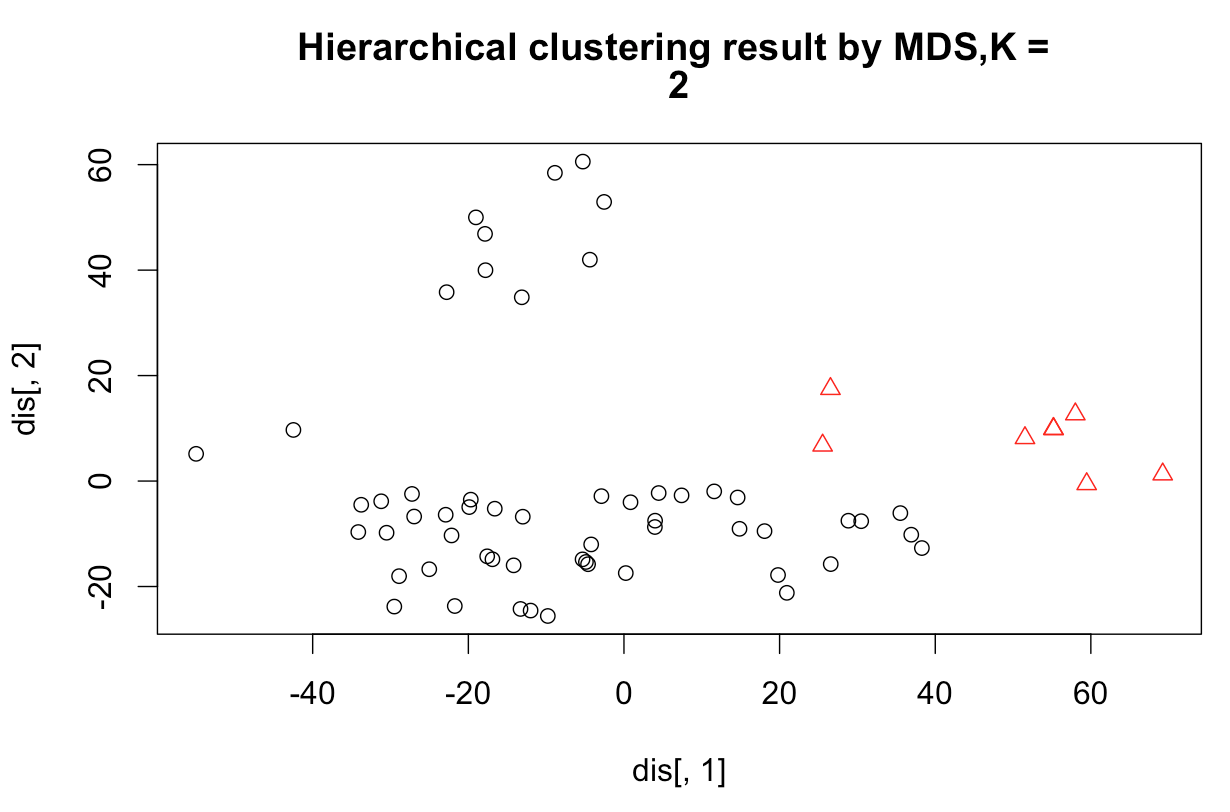


The plot of average silhouettes versus K = 2,…, 10 is shown below:

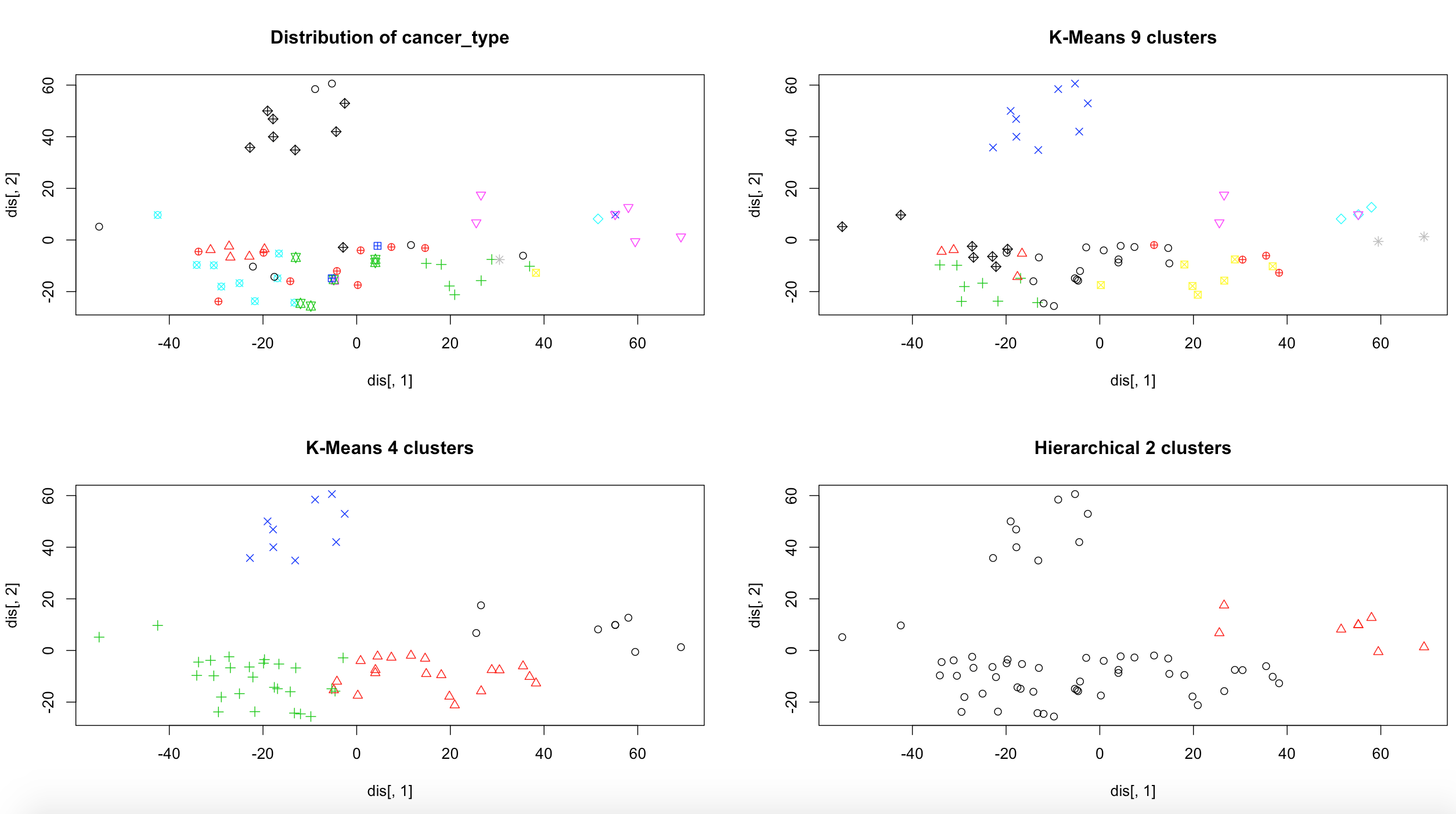


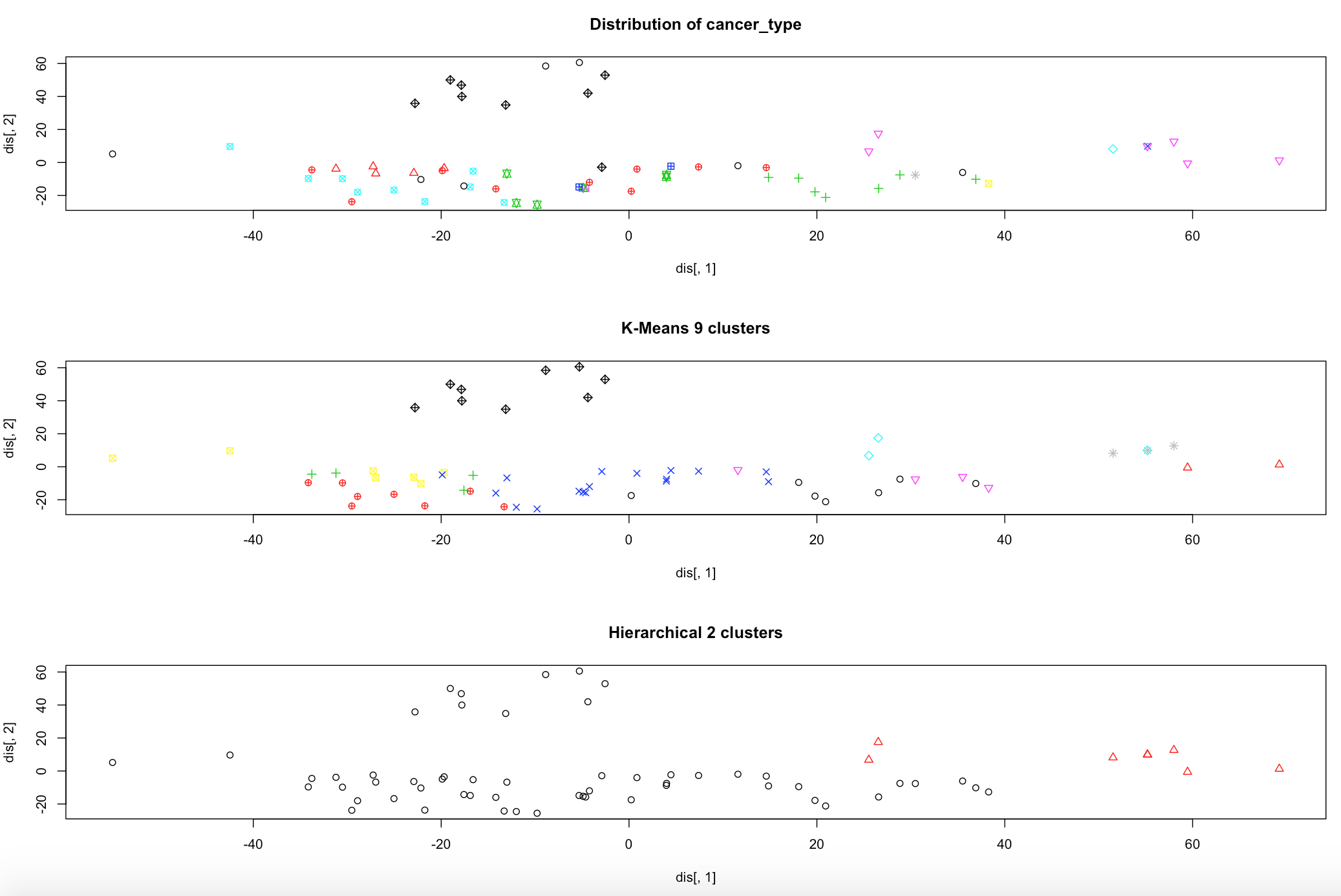
Based on silhouette statistics, the number of clusters should be 2 which give maximum silhouettes coefficient. Below is the plot of silhouettes and clustering result using MDS to display the data with K = 2:





C) To find out if the clustering results match the cancer type. I first compare the distribution of the original data type to the different clsuters. As shown below:



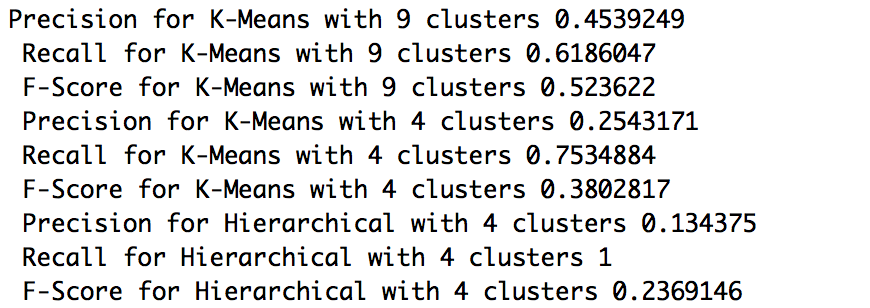


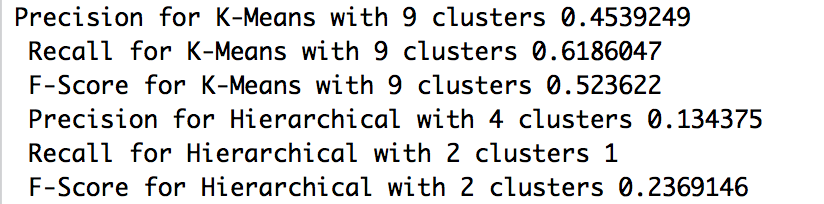
Comment:

In the first plot, different types of cancers are represented by different colors, while the second and third plot shows the clustering results of K-means. It looks like the clustering result match the cancer type quite well. This is because it has roughly identified a number of distinct cancer types, which includes the cancer types annotated in black, light blue and red. The third plot shows the clustering result of Hierarchical with average linkage. With only 2 clusters formed, it is only able to distinct the cancer type in pink color on the right end of the first plot. While all cancer types were clustered in the same group. Therefore, we cannot say this that this kind of clustering match cancer types well.

To have a more objective analysis, the precision, recall and F-score of the clustering results was computed. These measures estimate whether the points as being in the same cluster was correct with respect to the underlying true categories. Precision is the fraction of points correctly put in the same cluster, recall is the fraction of actual pairs that are identified. F-measure is the harmonic mean of precision and recall.

In this case, since we would like to determine if the clustering result match cancer types, we assume the cancer types as the underlying true categories to get the precision, recall and F-score. The result from R is as below:





Comment: