DATA MINING IEE 520

Ronak Vikam 1213203327
Harshil Shah 1213348082
Viren Bhanushali 1213232850
Jay Bhanushali 1213436781

Q 1.

To measure the cluster quality, we can use Sum of Squared Error (SSE). As given, the number of clusters (K= 2,3,4,5,10), following is the SSE for all the cluster.

SSE [2] = 7.187

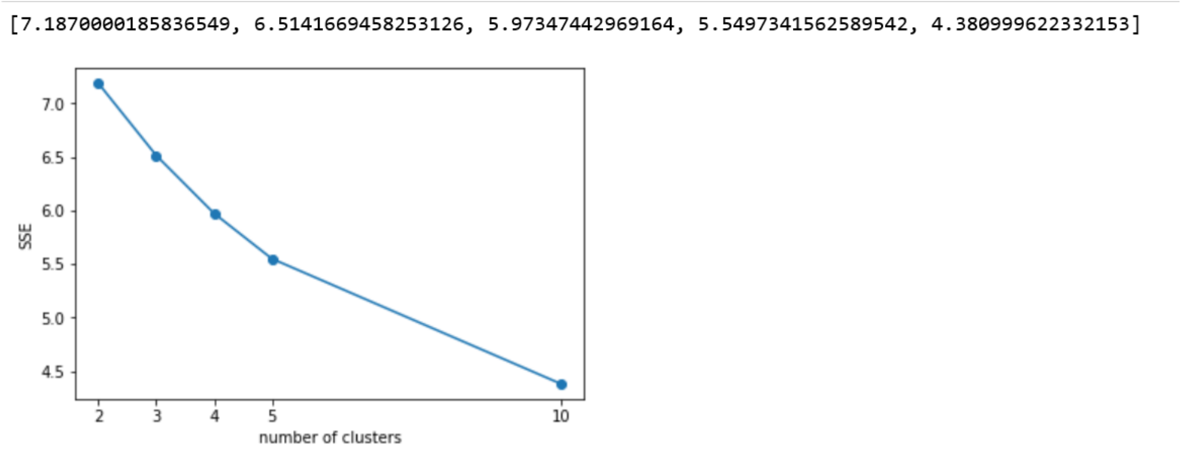
SSE [3] = 6.514

SSE [4] = 5.97

SSE [5] = 5.549

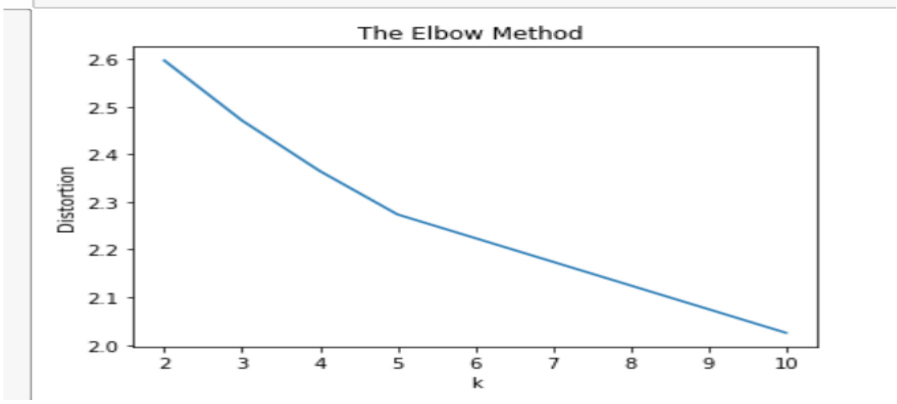
SSE [10] = 4.380

Plot of SSE against number of clusters



```
[7.1870000185836549, 6.5141669458253126, 5.97347442969164, 5.5497341562589542, 4.380999622332153]
```

Elbow Method



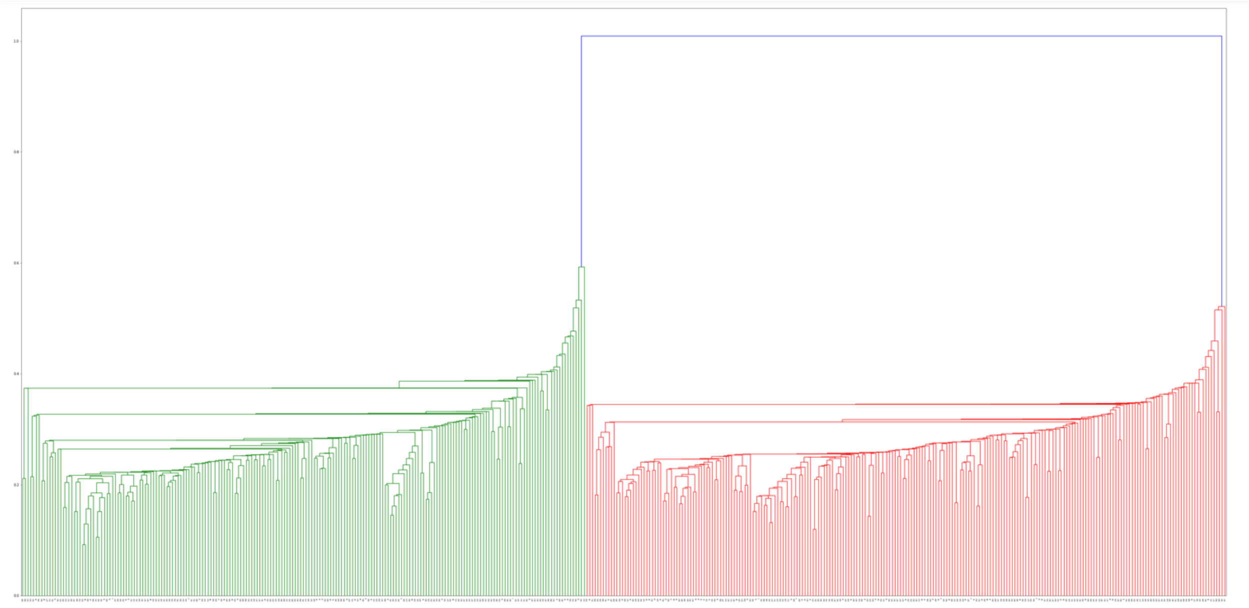
From the Elbow method we can confirm, that there are 10 clusters which will be able to explain the data better.

We can see that data is mixture, of categorical and continuous data, we can use Log Likelihood criteria to evaluate the no of clusters

Q 2.

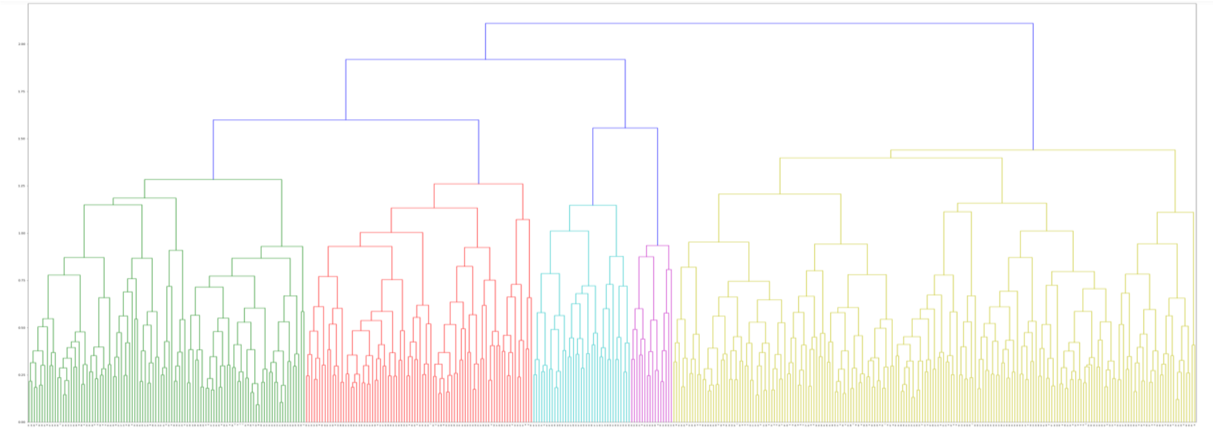
Hierarchical clustering can be done with different type of linkages such as single, complete, average and ward’s distance,

Single:



Colors of the clusters: Red Green(2 clusters)

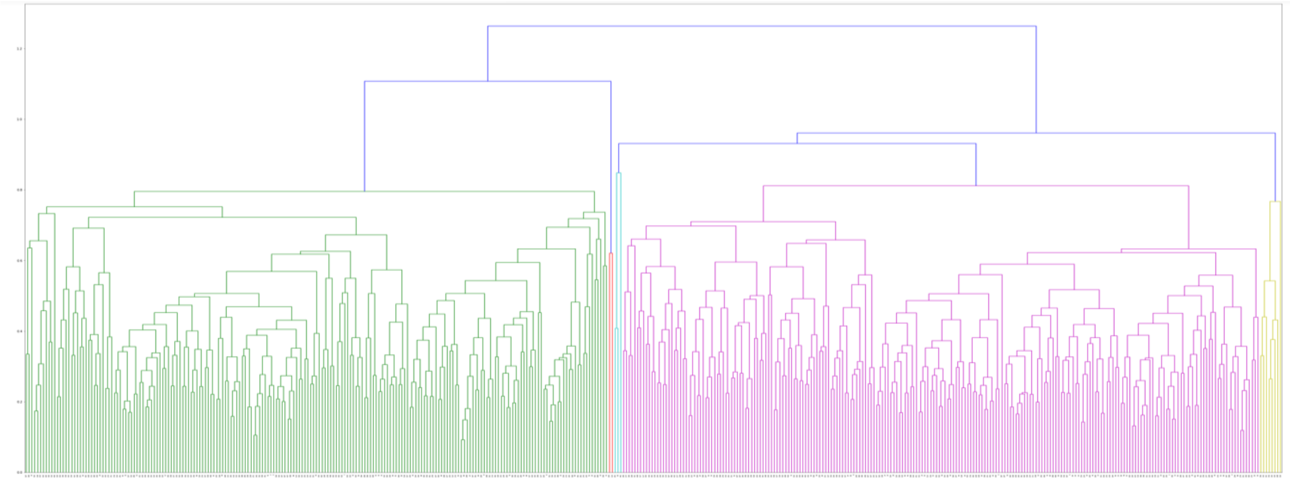
Complete:



Color of the clusters:

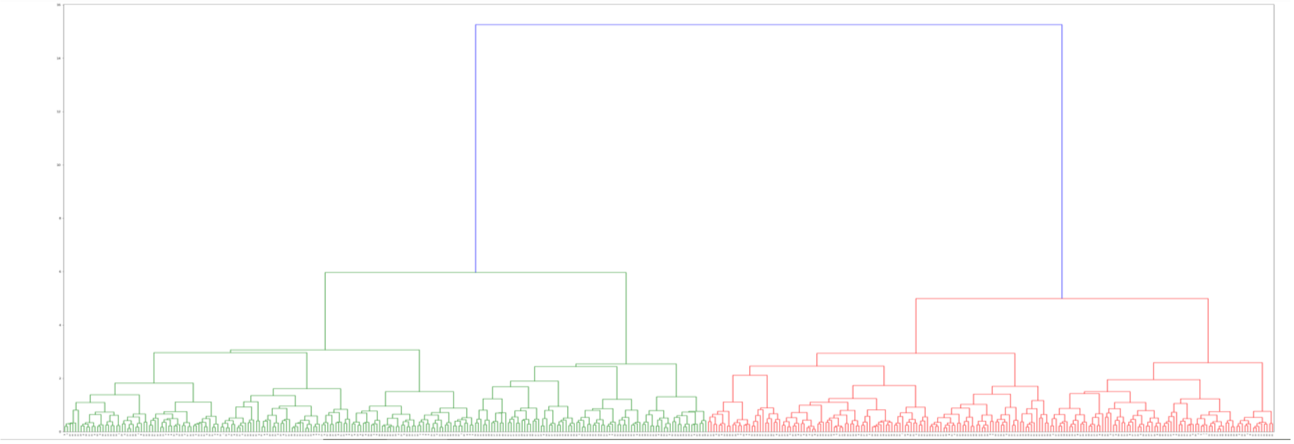
Green, Red, Blue, Magenta, Yellow

Average



Colors of the clusters: Green, Red, Blue, Pink, Light Green

Ward’s Distance



Colors of the clusters: Green, Red

To see which linkage performs the best, we can use the Cophenetic correlation coefficient for the linkages.

Linkages	Cophenetic Coefficient
Single	0.841
Complete	0.75
Average	0.8653
Ward’s Distance	0.8456

From the above table, we can see that that the Average Linkage has the highest coefficient which means that it has 7 clear clusters that can be differentiated from each other.