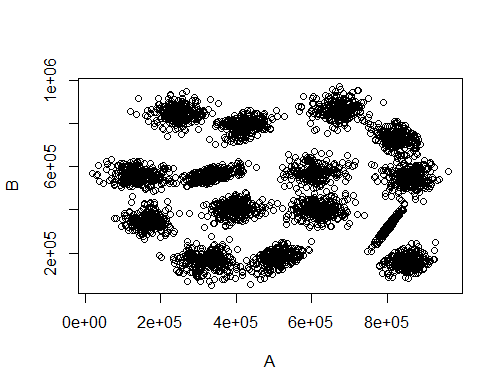
**Analysis of cluster shapes**

**Introduction**

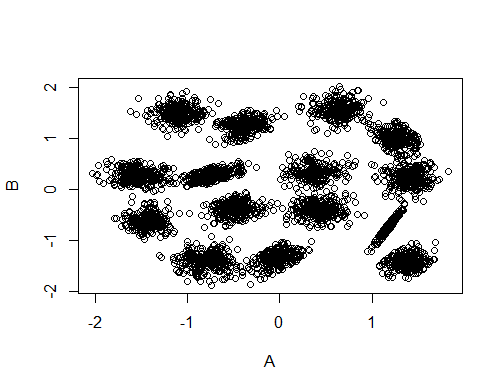
The dataset is Synthetic 2-d data, where N = 5000 vectors and k = 15 Gaussian clusters. I would like to analyse whether these 15 Gaussian clusters can be put into less clusters.

**Data Cleaning**

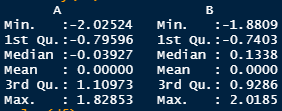
Initially, I started data cleaning to get the dataset into columns with names to allow the analysis to determine whether the number of clusters can be reduced. The code can be found under “code.R” in the GitHub repository.

**Analysis of Cluster points**

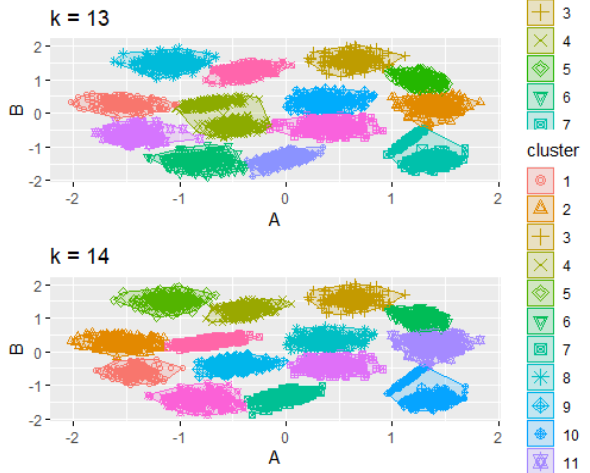
Initially we plot the graph, the graph is displayed below. We see that the numbers are very high, so I will need to scale the plot to make them more reasonable.



Now that the plot and summary of the plot is more reasonable, we are now ready to analyse whether clusters can merge.



Now that it is scaled and plotted, it shows that it has meaningful clusters in the data. We now need to see whether we could reduce the number of clusters using k-means algorithm.



In k = 13 we see more interconnected cluster points than compared to k = 14. However, k = 14 still has multiple interconnections. This means k = 15 cluster points is best for this dataset.