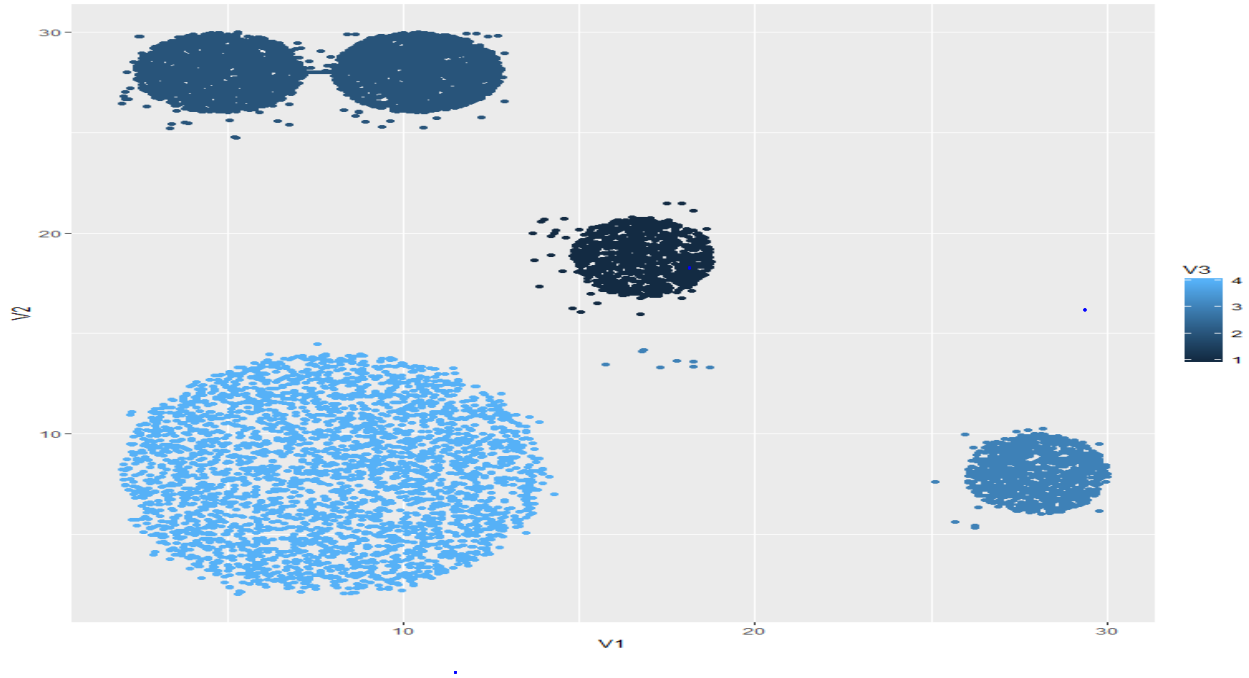
**Step1:**

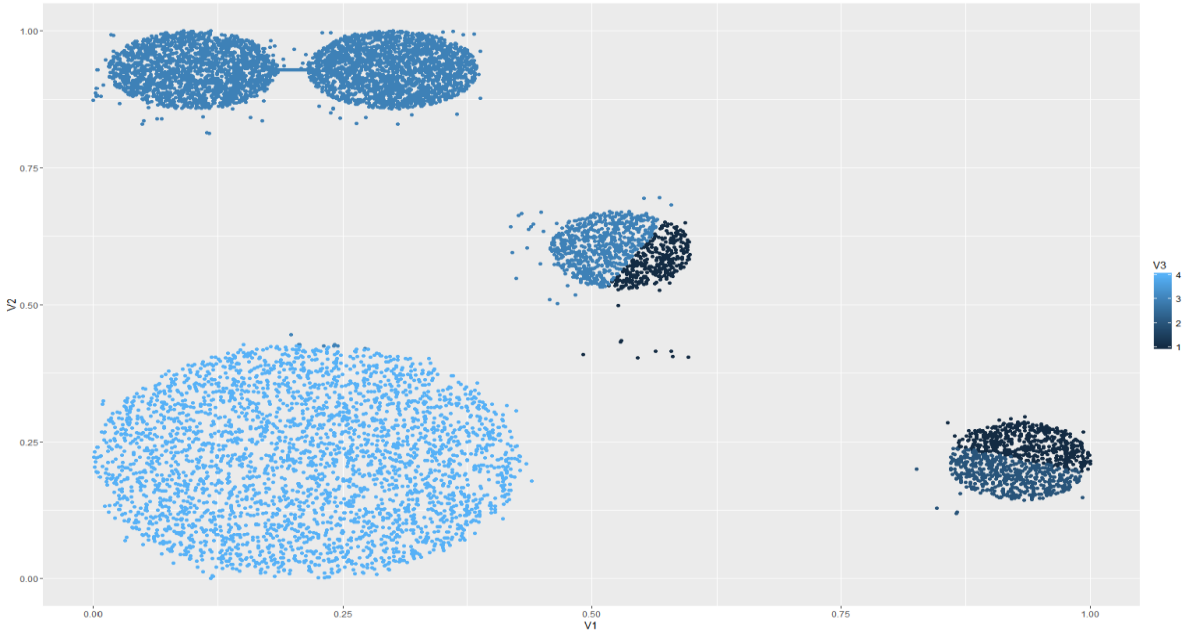
The script file is “KMeans.py”

The output data is the “step1.txt”, “step2a.txt”, “step2b.txt” in the result folder.

The ground truth graph is “truth.png”



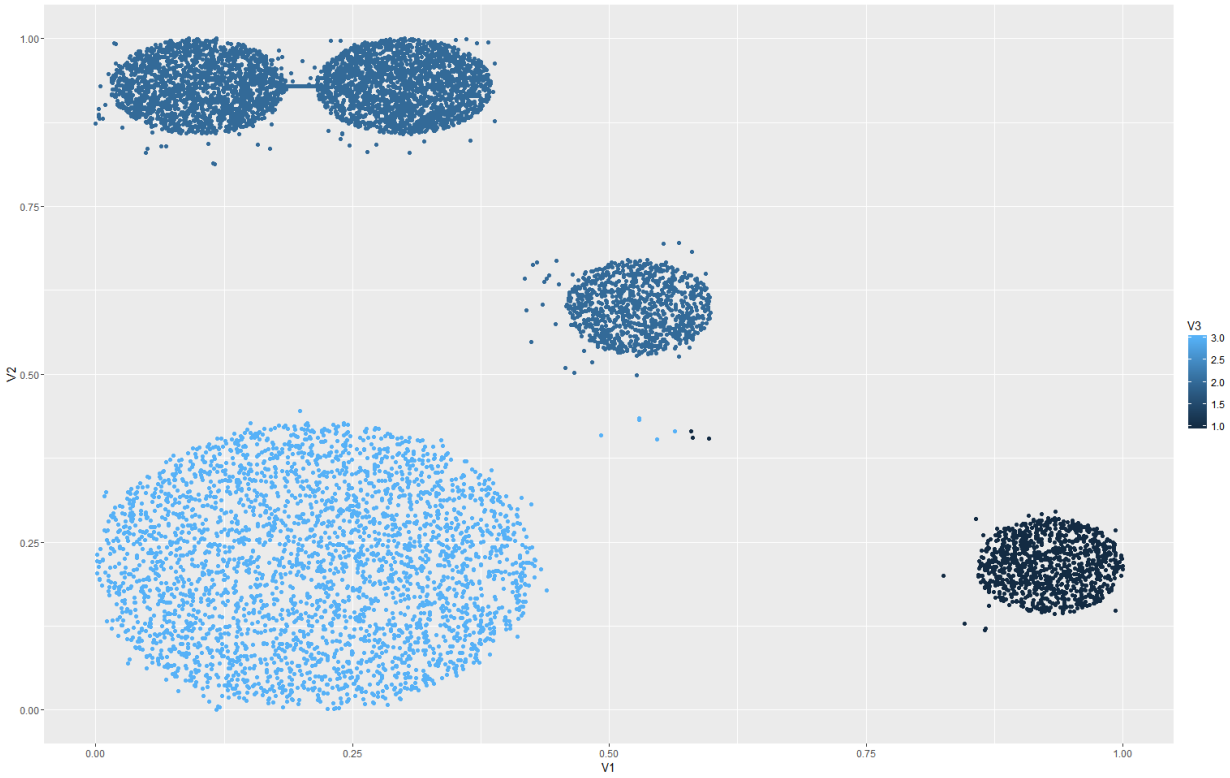

The step1 graph is:

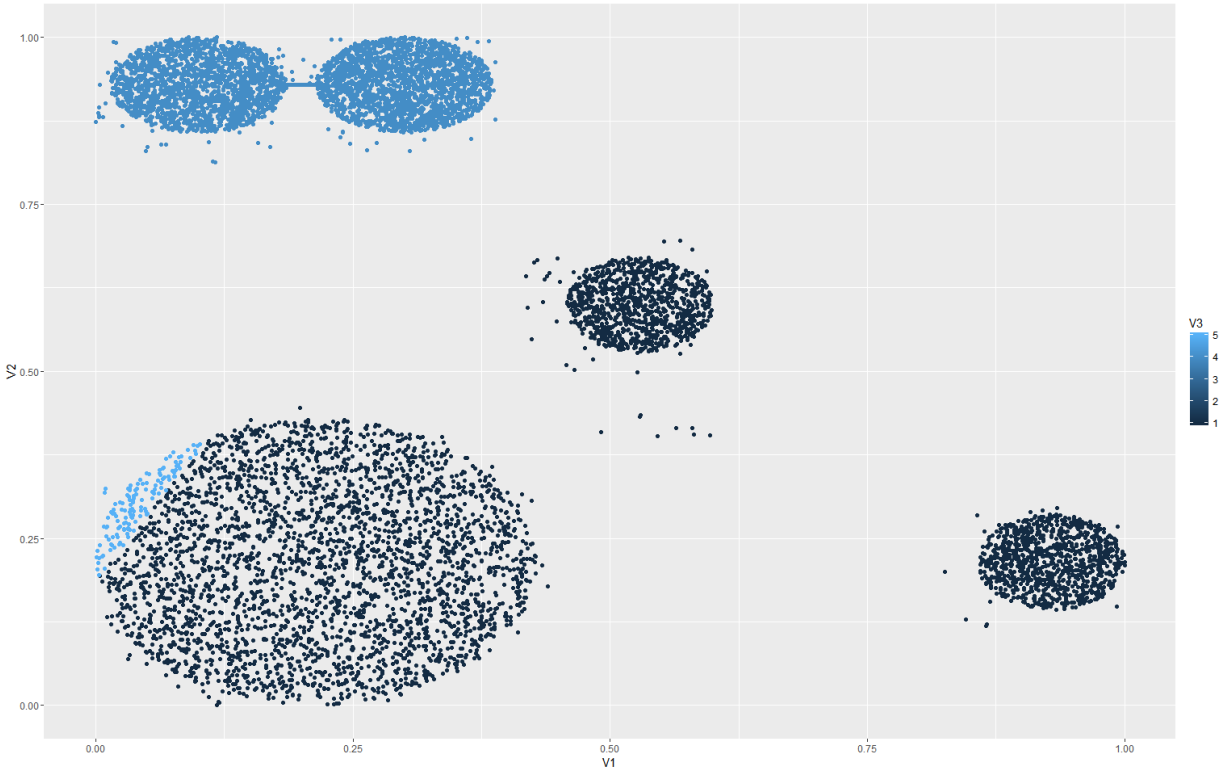



The cluster result is not well preformed. We can find there should be 5 clusters due to the sight of graph, so we need to increase the k for the cluster.

**Step2:**

The output file is in “step2a.txt” and “step2b.txt” in the result folder.



Compare different k, we can find when k = 3, the cluster fit the best by intuition.

A general way to choose the parameters k is use b-cubed to calculated different parameters, when the best recall and precision found, the parameter is the best.

**Step3**

The script file is named “bcubed.py” in the code folder.

| K | Precision | Recall |
|---|---|---|
| 4 | 0.901731160604 | 0.940585227286 |
| 3 | 0.900091637983 | 0.999423627702 |
| 5 | 0.839841452681 | 1.0 |

Since we want both high precision and recall, one method to combine it together is F-measure: F = 2*(P*R)/(P+R), then get the optimization. For instance, for this problem, we can find when k = 3, the F score is 0.9471.