Data Mining Project 2 Report

Thang Le

Data Visualization

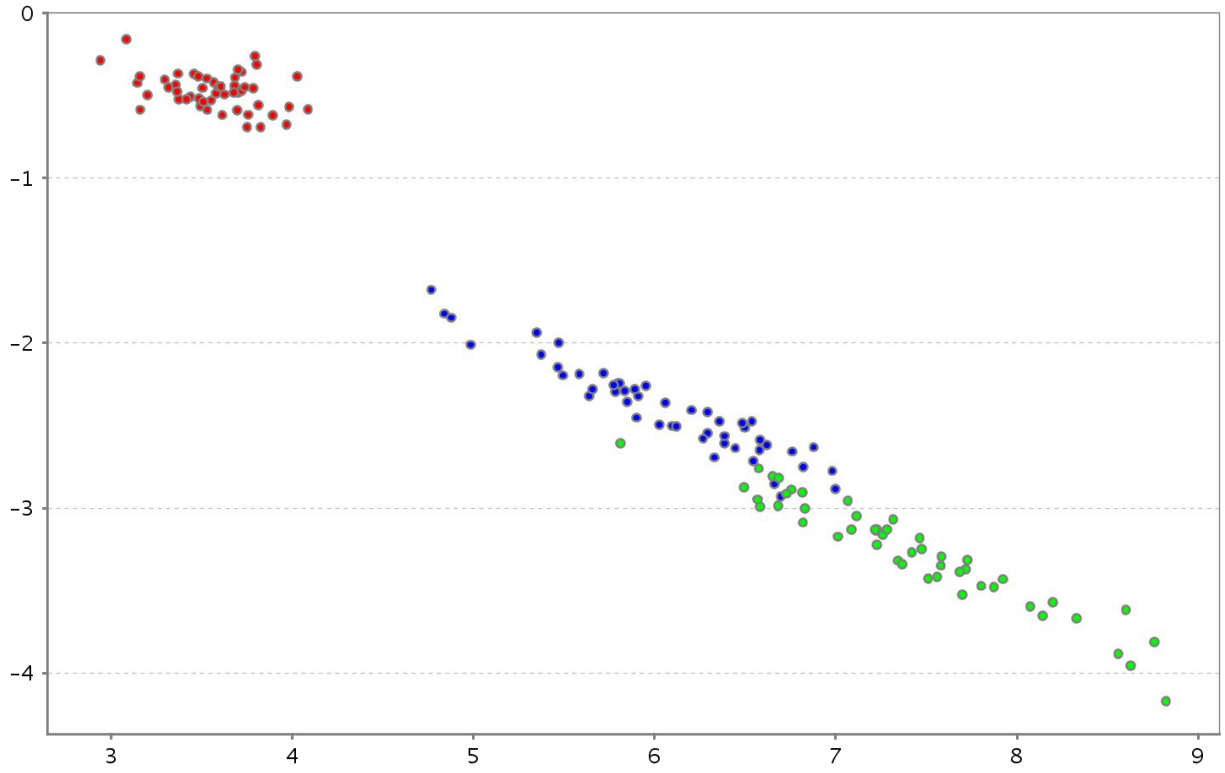

Figure 1: Plot of Iris dataset after Reducing dimension to 2 by using SVD

For the most part, the data of different classes are visually separated. Data of the red class are entirely separated from blue and green. Blue and green data are mostly separated except the bottom part of blue and top part of green where the data starts to overlap.

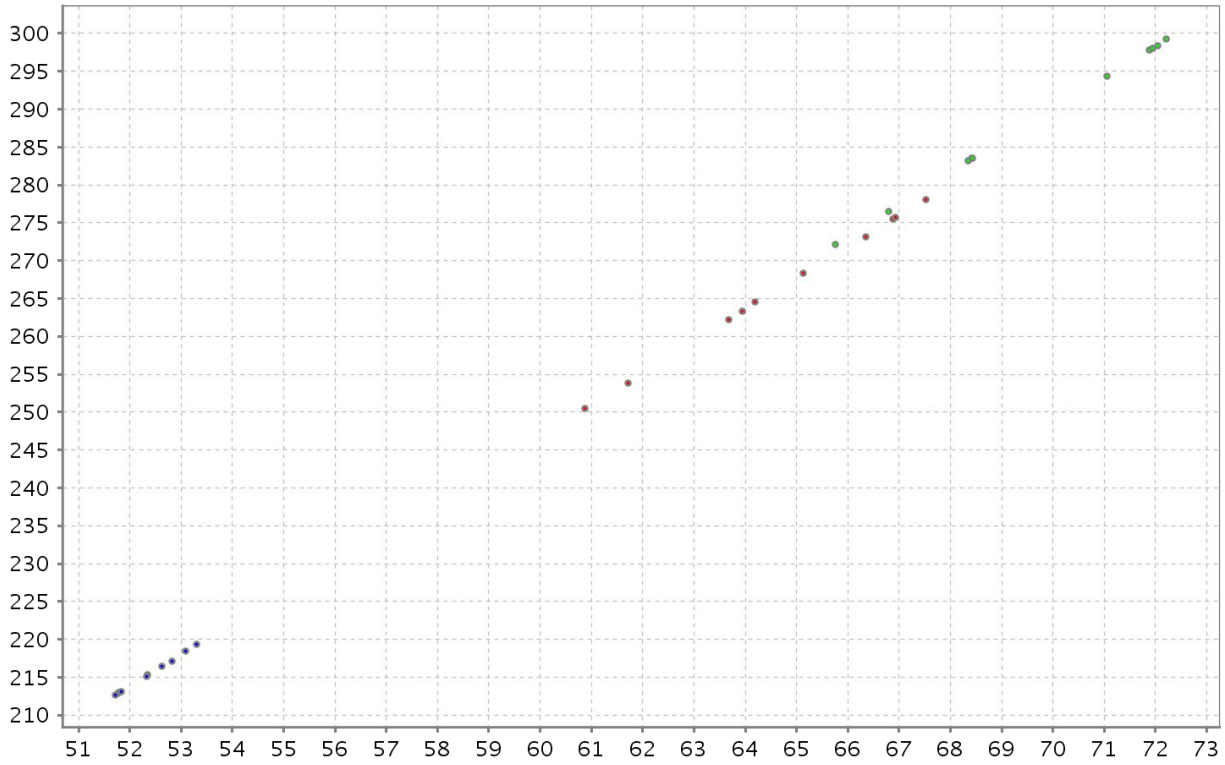

Figure 2: Plot of ATT dataset after Reducing dimension to 2 by using SVD

The blue data are well grouped together, but the red data are a bit spread out and are overlapping with some of green's data.

K-Means

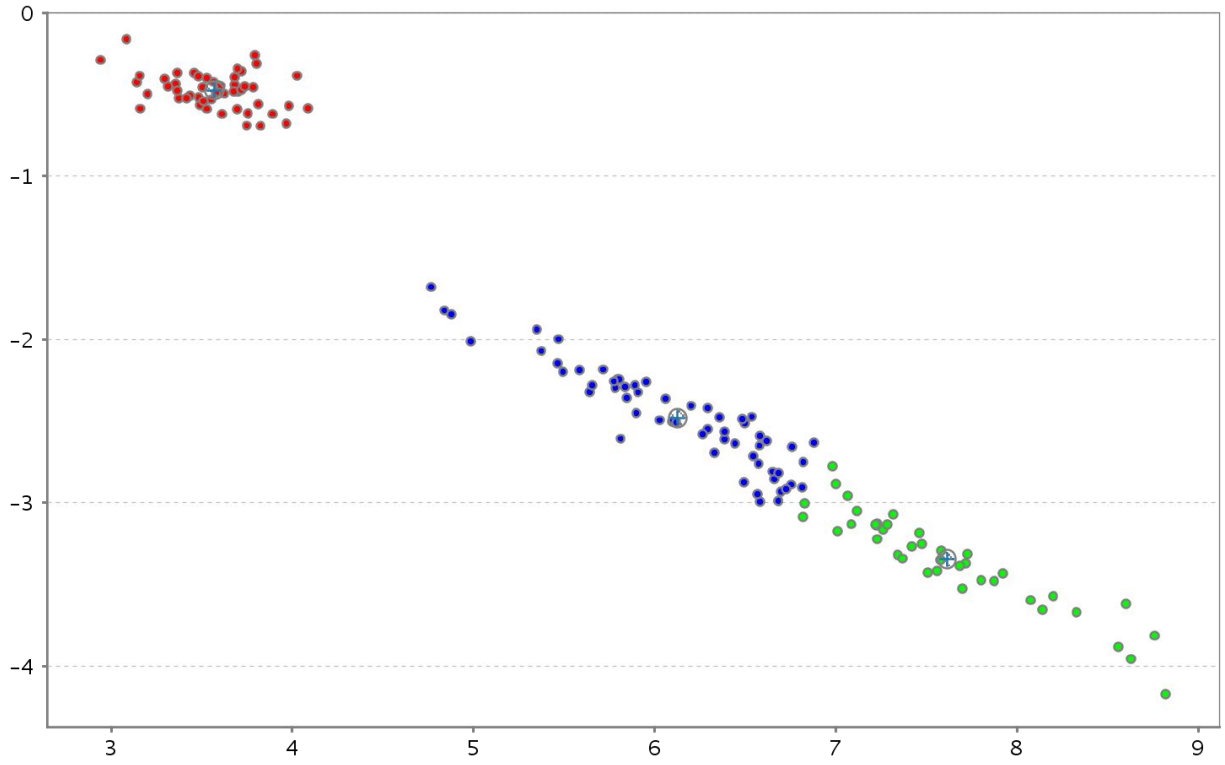

Figure 3: Plot of iris dataset produced by K-Means where k=3

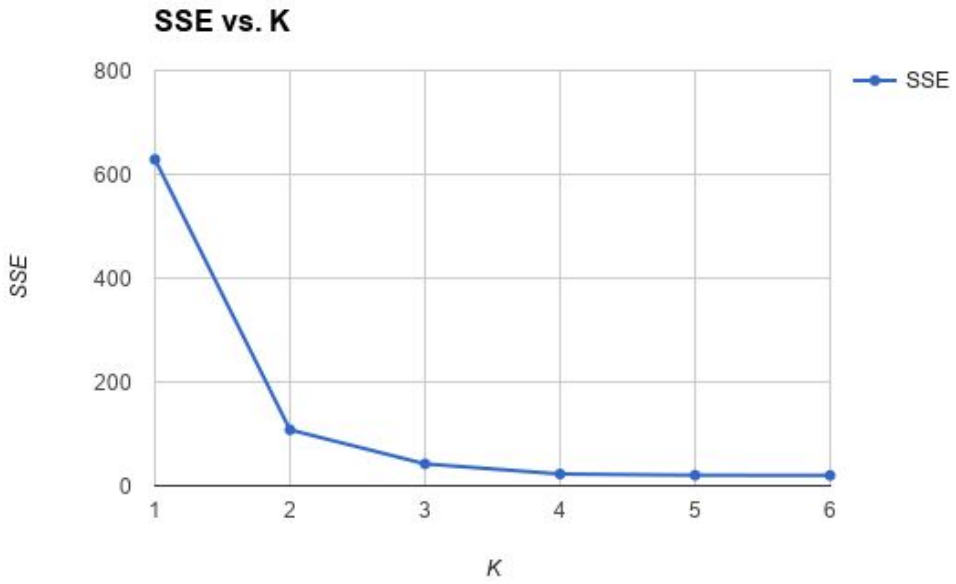

Figure 4: Iris dataset for SSE vs K

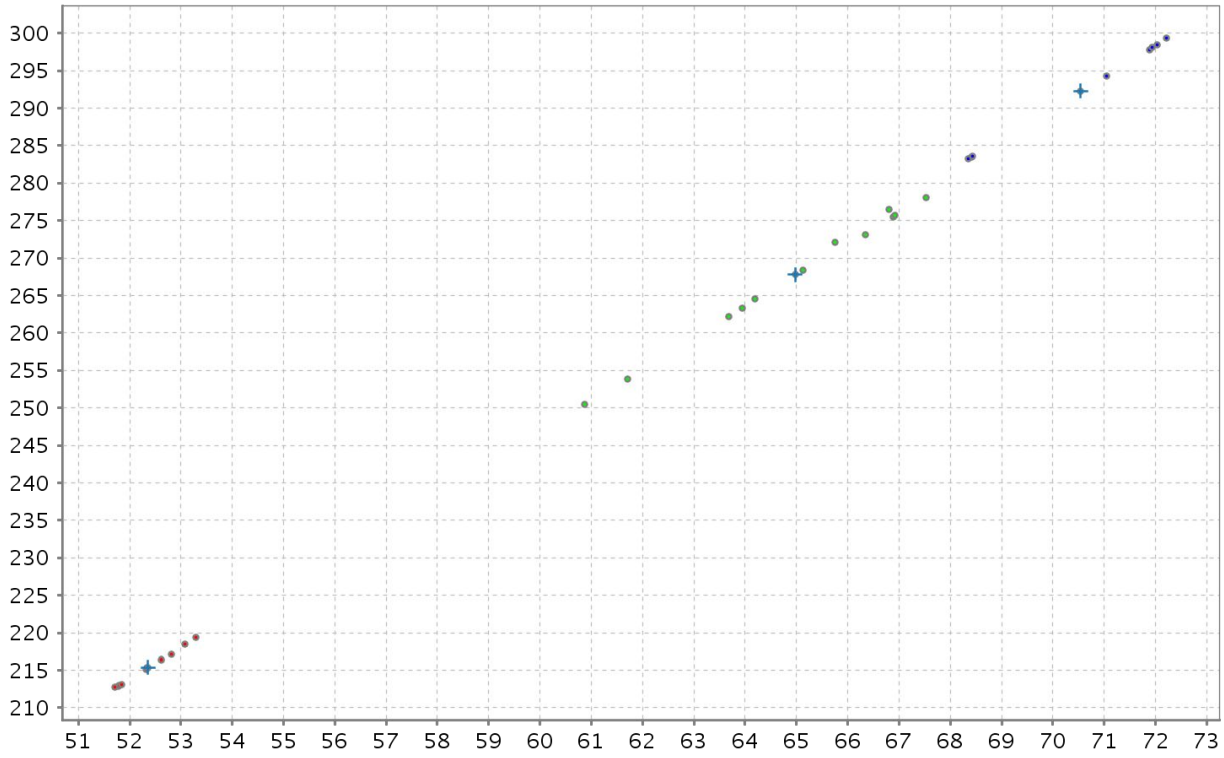

Figure 5: Plot of ATT dataset produced by K-Means where k=3

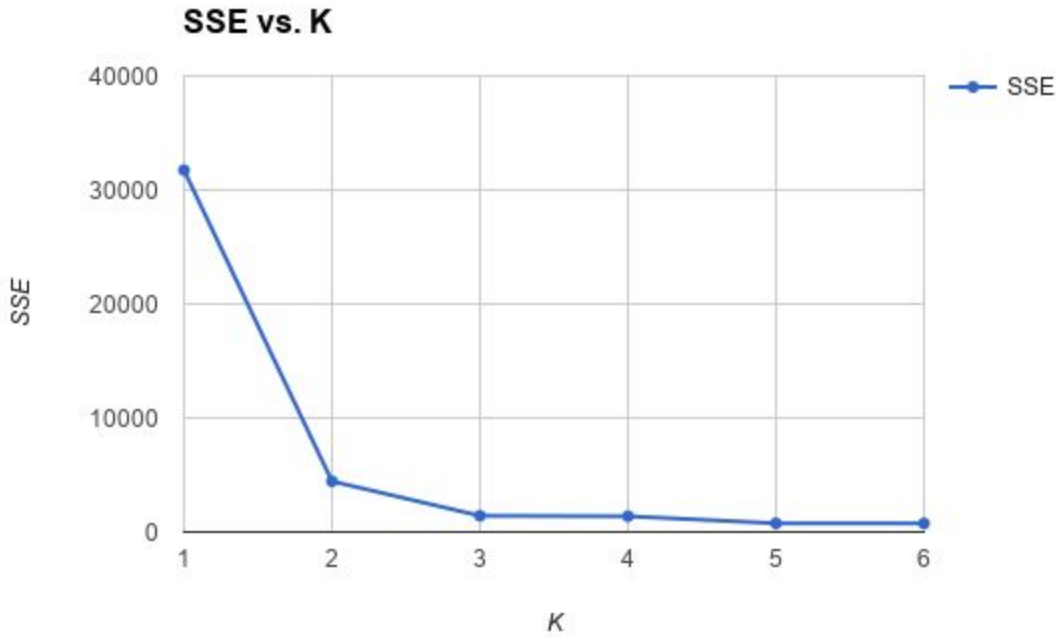

Figure 6: ATT dataset for SSE vs K

To evaluate the K-Means results, I computed the sum of squares error (SSE). I can use the SSE to help choose the right *k* value by plotting the SSE for different *k* values and look for where the SSE starts to “flatten out”. As seen in figures 4 and 6, the SSE starts to “flatten out” at k=3 which corresponds to the actual number of classes in both datasets.

DBSCAN

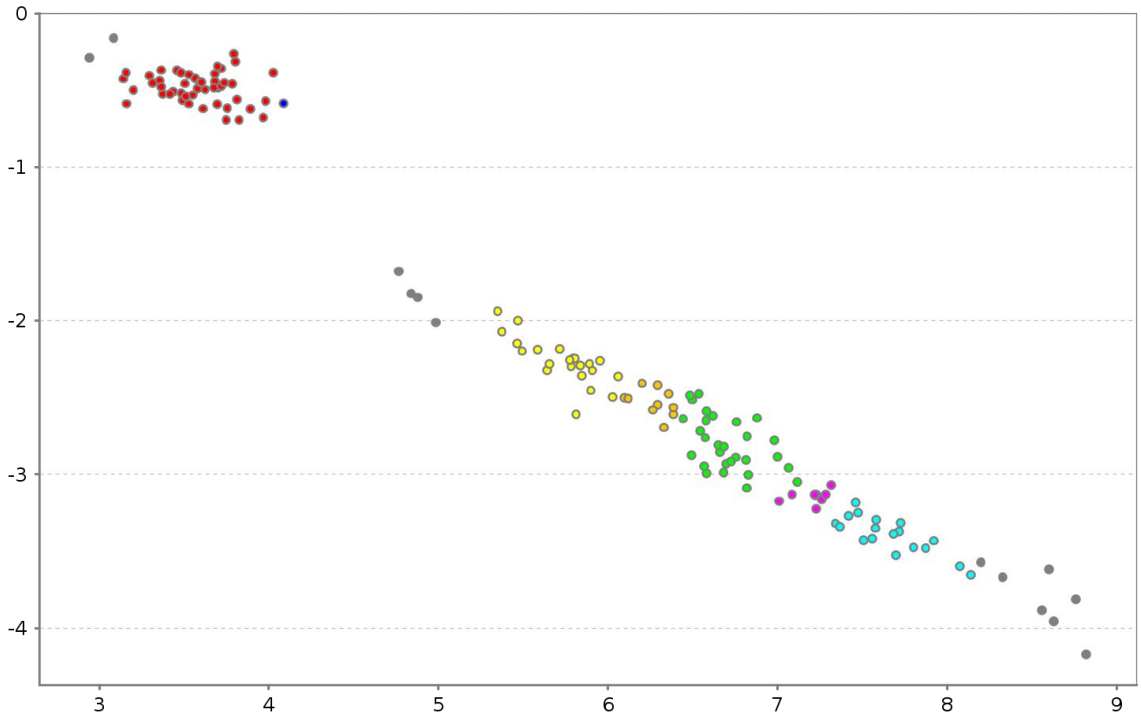

Figure 7: Plot of iris dataset produced by DBSCAN where eps=.4835, min_pts=15

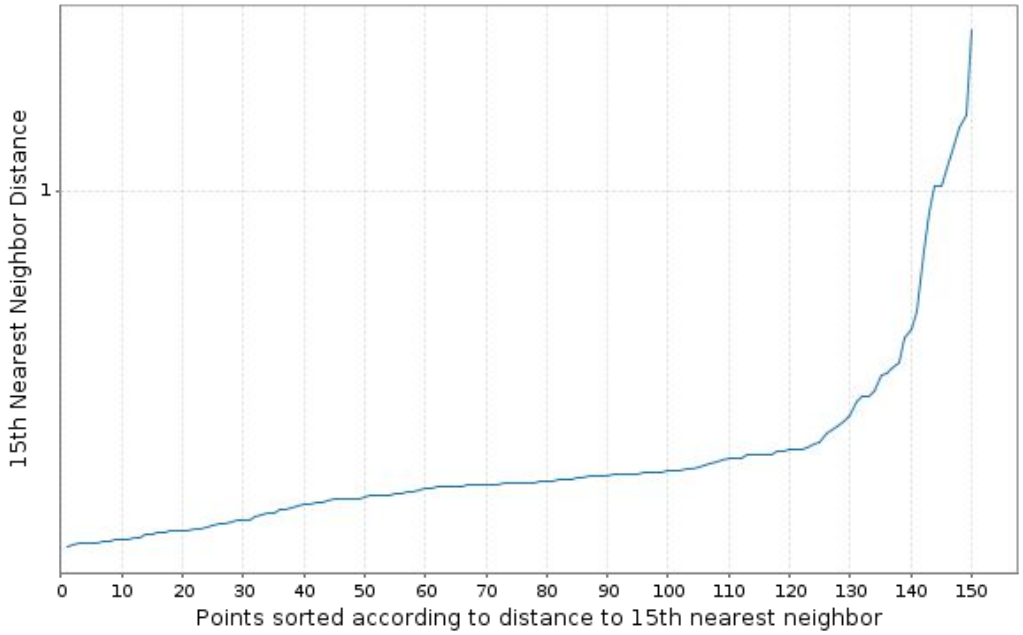

Figure 8: Plot of iris dataset for distances to the 15th nearest neighbor

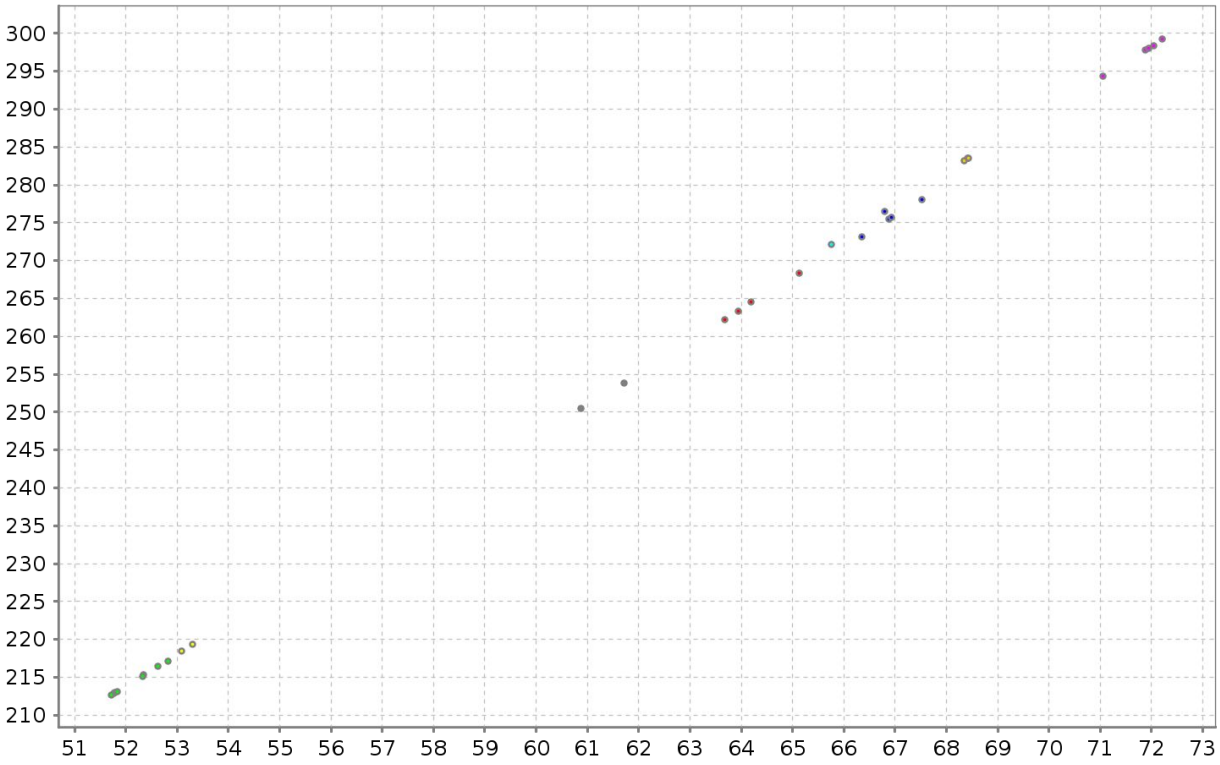

Figure 9: Plot of ATT dataset produced by DBSCAN where eps=5.2139, min_pts=3

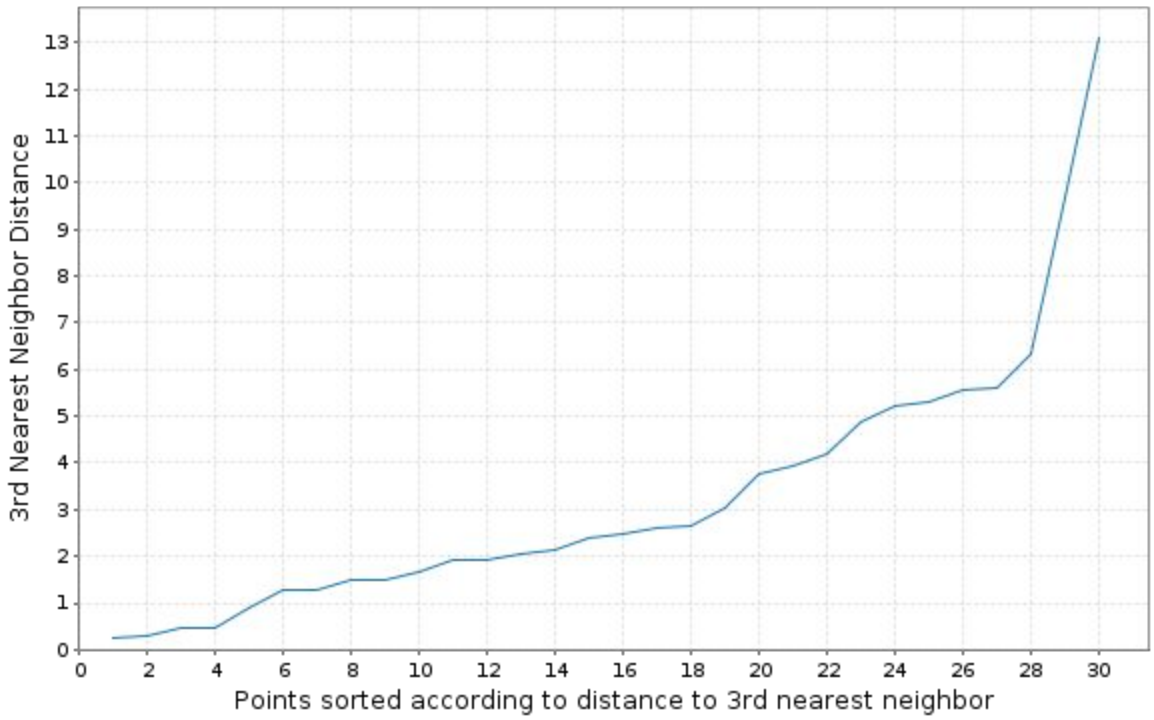

Figure 10: Plot of ATT dataset for distances to the 3rd nearest neighbor

To evaluate the DBSCAN results, I computed the purity of clusters. To pick eps and min_pts, I first chose a min_pts by visually inspecting figures 1 and 2. Then I plotted a k-distance graph with k=min_pts (figures 8 and 10) and look for an “elbow” in those graphs to pick the eps value.

- For the iris dataset, with eps=.4835 and min_pts=15, I got a purity of 0.853.
- For the iris dataset, with eps=5.2139 and min_pts=3, I got a purity of 0.9667.

Acknowledgement

The Breeze numerical processing library was used to do SVD and plot the clustering results.
Source: https://github.com/scalanlp/breeze