CS 434 Assignment 4

John Miller and Brandon Lee

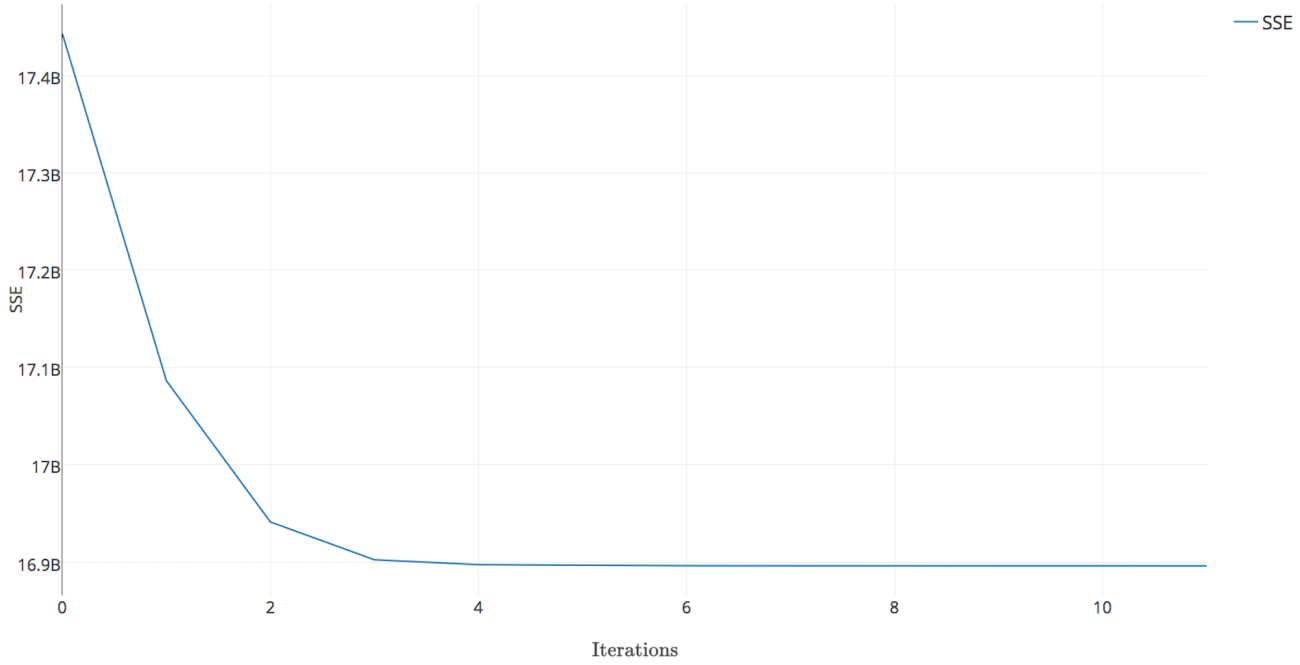
May 20, 2017

Part I: Non-hierarchical clustering - K-Means algorithm

Implement K-means algorithm

We ran our k-means algorithm 20 times with a k values of 2 and found that on average it converged after 16 iterations with an average final SSE of 16916817395.85. This run was fairly typical, although

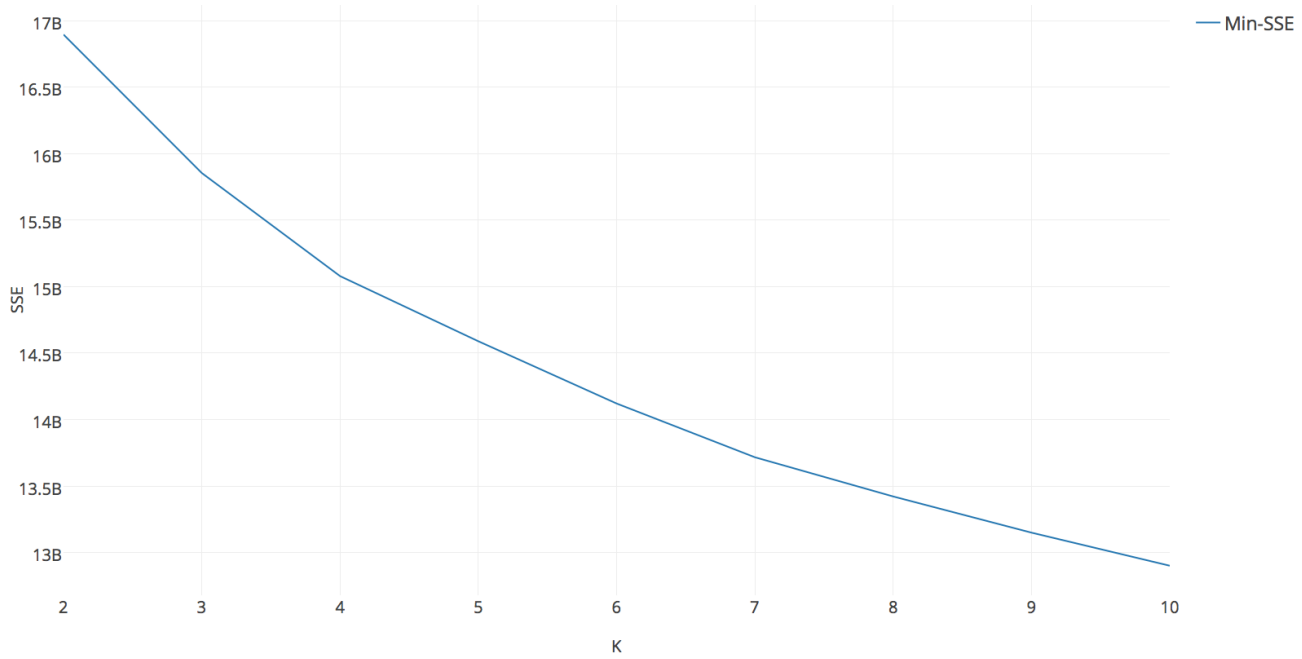
Figure 1: K-Means SSE vs. Iterations



there were some runs that had a slightly different pattern. This pattern still followed the same decreasing trend, but had a small plateau after about half the total iterations.

K-means with varying K

Figure 2: Varying K values vs. SSE

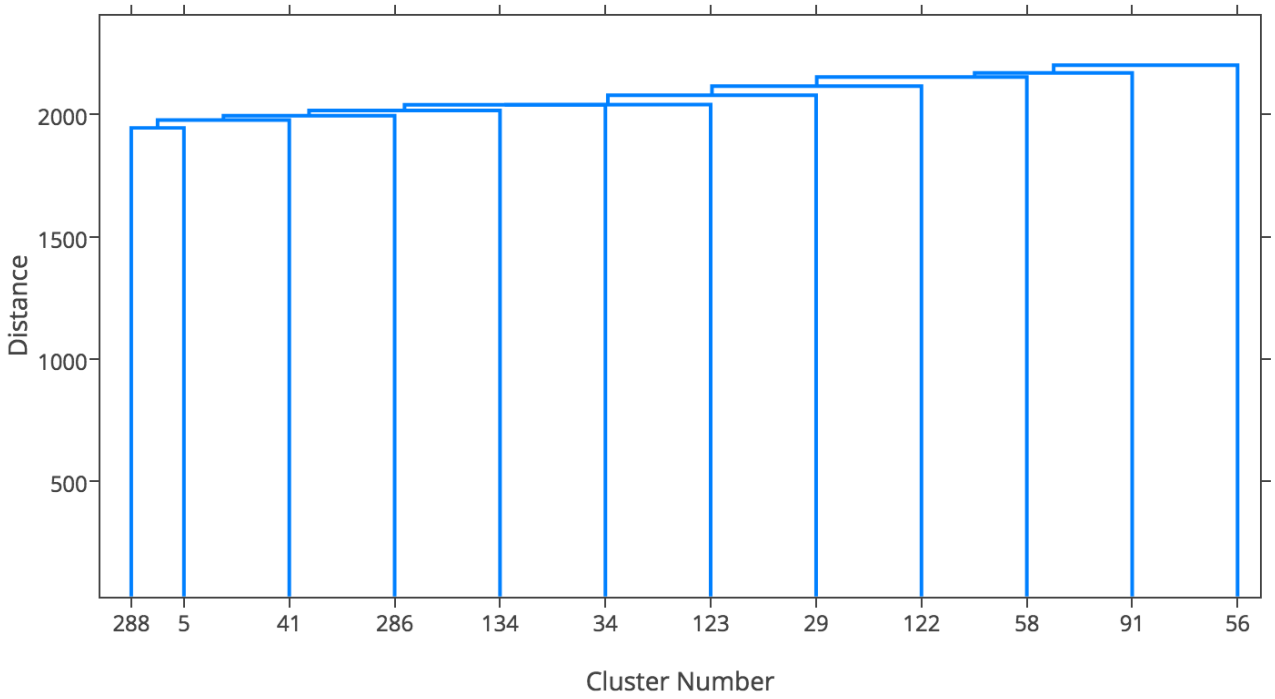


We ran our algorithm 10 times for each k=2...10 and tracked the minimum SSE achieved for each k. Each increase of k led to a smaller total SSE. Based on our results we believe k=10 would be the best choice for this dataset. Although there is a "knee" at the k=3 mark which means k=2 could be a better option.

Part 2: Hierarchical agglomerative clustering (HAC)

Single Link

Figure 3: Single Link HAC

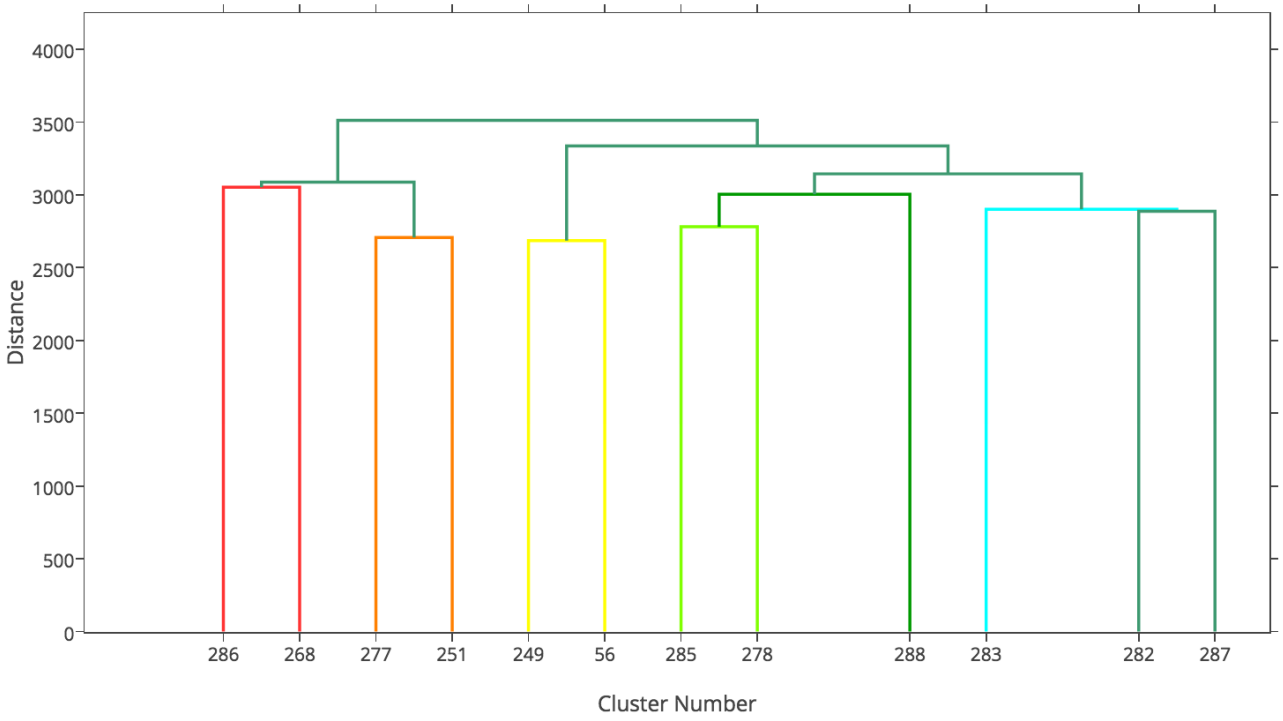


Cluster1	Cluster2	Formed	Height	Dist	C1 Size	C2 Size
288	5	289	11	1944.21269413	138	1
289	41	290	10	1976.6580382	139	1
290	286	291	9	1993.71562666	140	2
291	134	292	8	2015.57014266	142	1
292	34	293	7	2038.74127834	143	1
293	123	294	6	2039.14565443	144	1
294	29	295	5	2077.67490238	145	1
295	122	296	4	2114.781549	146	1
296	58	297	3	2152.28994329	147	1
297	91	298	2	2168.80035965	148	1
298	56	299	1	2200.34133716	149	1

From just the dendrogram we had a hard time determining the number of clusters that were found. But looking at the cluster sizes it looks like there is one very large cluster at the end and all the other clusters are of size one except for one which is size two.

Complete Link

Figure 4: Complete Link HAC



Cluster1	Cluster2	Formed	Height	Dist	C1 Size	C2 Size
249	56	289	11	2684.97094956	3	1
277	251	290	10	2706.01662966	16	6
285	278	291	9	2780.74036904	10	14
287	282	292	8	2887.18513435	25	16
292	283	293	7	2900.29601937	41	40
291	288	294	6	3003.82722539	24	8
286	268	295	5	3051.97755562	7	4
295	290	296	4	3086.77938959	11	22
294	293	297	3	3144.01781165	32	81
297	289	298	2	3335.61103848	113	4
298	296	299	1	3512.02135529	117	33

From just the dendrogram we had a hard time determining the number of clusters that were found. Again looking at the clusters listed in the table there appear to be 12 unique clusters of various sizes when we started tracking cluster merges.