Ian Battin: ibattin@calpoly.edu
Derek Lance: dlance@calpoly.edu

CSC 466 - Lab 4 Report

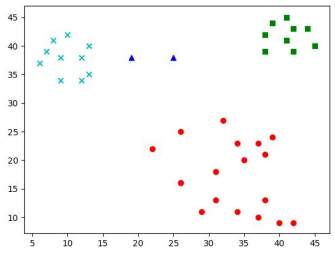
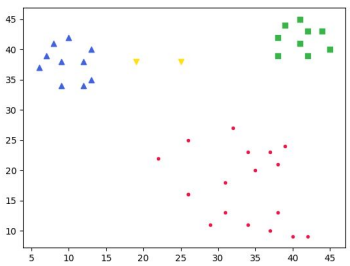
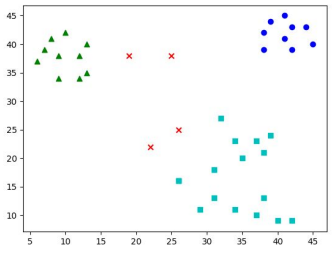
Study Design

We chose to design all of our algorithms in a similar way, using centroid for our clusters in all 3 cases to produce predictable results for tuning our parameters. This seems to work just fine for most datasets. We chose normal euclidean distance with kmeans and dbscan but went with squared euclidean distance for hclustering as it gave finer control over the threshold selection. Aside from these changes, the methods were implemented in the standard way.

Results

4clusters.csv

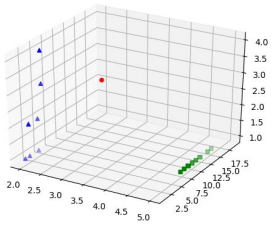
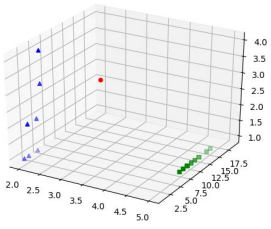
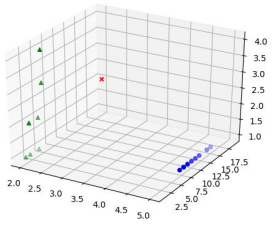
Method	Kmeans	hclustering	dbscan
Params	Num Clusters: 4	Threshold: 140	Epsilon: 6.3, MinClusters: 2
Clusters	Cluster 0: Center: 33.166666666666664, 17.27777777777778 Max Dist. to Center: 12.124101086698913 Min Dist. to Center: 2.2838671990104937 Avg Dist. to Center: 7.579450143514094 SSE: 1164.1111111111109 18 Points: Cluster 1: Center: 22.0, 38.0 Max Dist. to Center: 3.0 Min Dist. to Center: 3.0 Avg Dist. to Center: 3.0 SSE: 18.0 2 Points:	Cluster: 0 Center: [33.166666666666664, 17.27777777777778] Max Dist. to Center: 146.99382716049377 Min Dist. to Center: 5.2160493827160375 Avg Dist. to Center: 64.67283950617283 SSE: 106334.09670781894 18 Points: Cluster: 1 Center: [41.111111111111114, 41.77777777777778] Max Dist. to Center: 18.283950617283928 Min Dist. to Center: 0.6172839506172859 Avg Dist. to Center: 9.604938271604938	Cluster 0: Center: 41.111111111111114, 41.77777777777778 Max Dist. to Center: 4.275973645531965 Min Dist. to Center: 0.7856742013183874 Avg Dist. to Center: 2.9117006199139848 Sum Squared Error: 86.44444444444443 9 Points: Cluster 1: Center: 9.9, 37.8 Max Dist. to Center: 4.3416586692184795 Min Dist. to Center: 0.9219544457292896 Avg Dist. to Center:

	Cluster 2: Center: 41.111111111111114, 41.77777777777778 Max Dist. to Center: 4.275973645531965 Min Dist. to Center: 0.7856742013183874 Avg Dist. to Center: 2.9117006199139848 SSE: 86.44444444444446 9 Points: Cluster 3: Center: 9.9, 37.8 Max Dist. to Center: 4.3416586692184795 Min Dist. to Center: 0.9219544457292896 Avg Dist. to Center: 3.4299301473944324 SSE: 128.5 10 Points:	SSE: 1102.625514403292 9 Points: Cluster: 2 Center: [22.0, 38.0] Max Dist. to Center: 9.0 Min Dist. to Center: 9.0 Avg Dist. to Center: 9.0 SSE: 162.0 2 Points: Cluster: 3 Center: [9.9, 37.8] Max Dist. to Center: 18.849999999999977 Min Dist. to Center: 0.8500000000000018 Avg Dist. to Center: 12.85 SSE: 1973.3049999999996 10 Points:	3.4299301473944324 Sum Squared Error: 128.5 10 Points: Cluster 2: Center: 34.3125, 16.5 Max Dist. to Center: 10.751635050074942 Min Dist. to Center: 3.5668832683450686 Avg Dist. to Center: 7.066141146631838 Sum Squared Error: 875.4375 16 Points: Outliers: Percentage of data: 10.26 4 Points: 19.0, 38.0 25.0, 38.0 26.0, 25.0 22.0, 22.0
			

Comments: All methods did pretty good with this dataset as can be human verified by the graphs. DBSCAN performed slightly worse arguably labeling incorrect outliers, howerver, it could go either way.

AccidentsSet01.csv

Method	Kmeans	hclustering	dbscan
Params	Num Clusters: 3	Threshold: 30	Epsilon: 6.3, MinClusters: 2

Clusters	Cluster 0: Center: 2.0, 19.0, 2.0 Max Dist. to Center: 0.0 Min Dist. to Center: 0.0 Avg Dist. to Center: 0.0 SSE: 0.0 1 Points: Cluster 1: Center: 2.0, 3.125, 2.0 Max Dist. to Center: 2.7414640249326636 Min Dist. to Center: 0.875 Avg Dist. to Center: 1.6467460440475656 SSE: 24.875 8 Points: Cluster 2: Center: 5.0, 10.3, 1.0 Max Dist. to Center: 4.699999999999999 Min Dist. to Center: 0.3000000000000007 Avg Dist. to Center: 2.16 SSE: 64.1 10 Points:	Cluster 0: Center: 2.0, 19.0, 2.0 Max Dist. to Center: 0.0 Min Dist. to Center: 0.0 Avg Dist. to Center: 0.0 SSE: 0.0 1 Points: Cluster 1: Center: 2.0, 3.125, 2.0 Max Dist. to Center: 2.7414640249326636 Min Dist. to Center: 0.875 Avg Dist. to Center: 1.6467460440475656 SSE: 24.875 8 Points: Cluster 2: Center: 5.0, 10.3, 1.0 Max Dist. to Center: 4.699999999999999 Min Dist. to Center: 0.3000000000000007 Avg Dist. to Center: 2.16 SSE: 64.1 10 Points:	Cluster 0: Center: 5.0, 10.3, 1.0 Max Dist. to Center: 4.699999999999999 Min Dist. to Center: 0.3000000000000007 Avg Dist. to Center: 2.16 Sum Squared Error: 64.10000000000002 10 Points: Cluster 1: Center: 2.0, 3.125, 2.0 Max Dist. to Center: 2.7414640249326636 Min Dist. to Center: 0.875 Avg Dist. to Center: 1.6467460440475654 Sum Squared Error: 24.875 8 Points: Outliers: Percentage of data: 5.26 1 Points: 2.0, 19.0, 2.0 [[2.0, 19.0, 2.0]]
			

Comments: All methods did equally perfect as can be human verified.

AccidentsSet02.csv

Method	Kmeans	hclustering	dbscan

Params	Num Clusters: 6	Threshold: 10	Epsilon: 2, MinClusters: 3
Clusters	Cluster 0: Center: 4.0, 25.0, 0.0, 2.0, 70.0, 2.0, 0.0 Max Dist. to Center: 0.0 Min Dist. to Center: 0.0 Avg Dist. to Center: 0.0 SSE: 0.0 1 Points: Cluster 1: Center: 1.4545454545454546, 2.0, 0.45454545454545453, 7.454545454545454, 35.0, 1.0, 0.0 Max Dist. to Center: 2.3897162596536208 Min Dist. to Center: 0.8430562268632456 Avg Dist. to Center: 1.7781890443345616 SSE: 36.18181818181818 11 Points: Cluster 2: Center: 1.1, 1.5, 0.0, 2.0, 70.0, 1.0, 0.3 Max Dist. to Center: 1.7748239349298847 Min Dist. to Center: 0.5916079783099616 Avg Dist. to Center: 0.8795848709726043 SSE: 9.499999999999998 10 Points: Cluster 3: Center: 1.4166666666666667, 3.0833333333333335, 0.9166666666666666, 4.0, 45.0, 1.1666666666666667, 0.4166666666666667 Max Dist. to Center: 4.801620096962644	Cluster: 0 Center: [1.4583333333333333, 3.5, 0.5, 4.291666666666667, 39.166666666666664, 1.0, 0.375] Max Dist. to Center: 53.796874999999986 Min Dist. to Center: 18.296874999999982 Avg Dist. to Center: 30.32812499999999 SSE: 24101.359157986128 24 Points: Cluster: 1 Center: [1.0, 3.0, 2.0, 4.0, 45.0, 2.0, 0.0] Max Dist. to Center: 0.0 Min Dist. to Center: 0.0 Avg Dist. to Center: 0.0 SSE: 0.0 1 Points: Cluster: 2 Center: [1.2, 1.6, 0.6, 9.0, 35.0, 1.0, 0.0] Max Dist. to Center: 1.3600000000000003 Min Dist. to Center: 0.35999999999999993 Avg Dist. to Center: 0.6399999999999999 SSE: 2.8160000000000007 5 Points: Cluster: 3 Center: [5.0, 6.0, 0.0, 4.0, 45.0, 2.0, 0.0] Max Dist. to Center: 0.0 Min Dist. to Center: 0.0 Avg Dist. to Center: 0.0 SSE: 0.0 1 Points:	Cluster 0: Center: 1.875, 4.625, 0.125, 3.5, 35.0, 1.125, 0.375 Max Dist. to Center: 1.8915932438026946 Min Dist. to Center: 1.0383279828647594 Avg Dist. to Center: 1.4958402482240243 Sum Squared Error: 18.375 8 Points: Cluster 1: Center: 1.1, 2.4, 1.1, 4.0, 45.0, 1.1, 0.5 Max Dist. to Center: 1.5620499351813308 Min Dist. to Center: 0.6633249580710799 Avg Dist. to Center: 0.9175337680246306 Sum Squared Error: 9.6 10 Points: Cluster 2: Center: 1.4545454545454546, 2.0, 0.45454545454545453, 7.454545454545454, 35.0, 1.0, 0.0 Max Dist. to Center: 2.3897162596536208 Min Dist. to Center: 0.8430562268632456 Avg Dist. to Center: 1.7781890443345616 Sum Squared Error: 36.18181818181818 11 Points: Cluster 3: Center: 1.1, 1.5, 0.0, 2.0, 70.0, 1.0, 0.3 Max Dist. to Center: 1.7748239349298847

	Min Dist. to Center: 0.6236095644623237 Avg Dist. to Center: 1.7548938881153708 SSE: 55.33333333333333 12 Points: Cluster 4: Center: 2.75, 7.0, 0.0, 2.0, 70.0, 1.75, 0.5 Max Dist. to Center: 3.856812155135378 Min Dist. to Center: 2.207940216581962 Avg Dist. to Center: 2.7697339070949396 SSE: 32.5 4 Points: Cluster 5: Center: 1.7272727272727273, 5.090909090909091, 0.5454545454545454, 3.3636363636363638, 35.0, 1.3636363636363635, 0.45454545454545453 Max Dist. to Center: 4.684244737450274 Min Dist. to Center: 1.1022141502711038 Avg Dist. to Center: 2.2131405215660482 SSE: 67.63636363636364 11 Points:	Cluster: 4 Center: [3.0, 4.0, 0.0, 4.0, 35.0, 2.0, 0.0] Max Dist. to Center: 0.0 Min Dist. to Center: 0.0 Avg Dist. to Center: 0.0 SSE: 0.0 1 Points: Cluster: 5 Center: [1.25, 2.3333333333333335, 0.0, 2.0, 70.0, 1.0, 0.25] Max Dist. to Center: 32.23611111111111 Min Dist. to Center: 0.23611111111111122 Avg Dist. to Center: 4.930555555555556 SSE: 1179.2152777777774 12 Points: Cluster: 6 Center: [1.0, 4.0, 0.0, 4.0, 35.0, 3.0, 1.0] Max Dist. to Center: 0.0 Min Dist. to Center: 0.0 Avg Dist. to Center: 0.0 SSE: 0.0 1 Points: Cluster: 7 Center: [1.0, 6.0, 5.0, 3.0, 35.0, 2.0, 0.0] Max Dist. to Center: 0.0 Min Dist. to Center: 0.0 Avg Dist. to Center: 0.0 SSE: 0.0 1 Points: Cluster: 8 Center: [4.0, 25.0, 0.0, 2.0, 70.0, 2.0, 0.0] Max Dist. to Center: 0.0 Min Dist. to Center: 0.0 Avg Dist. to Center: 0.0	Min Dist. to Center: 0.5916079783099616 Avg Dist. to Center: 0.8795848709726043 Sum Squared Error: 9.499999999999998 10 Points: Outliers: Percentage of data: 20.41 10 Points: 1.0, 4.0, 0.0, 4.0, 35.0, 3.0, 1.0 1.0, 6.0, 5.0, 3.0, 35.0, 2.0, 0.0 2.0, 9.0, 0.0, 2.0, 35.0, 1.0, 1.0 1.0, 7.0, 0.0, 4.0, 45.0, 1.0, 0.0 5.0, 6.0, 0.0, 4.0, 45.0, 2.0, 0.0 6.0, 9.0, 0.0, 2.0, 70.0, 2.0, 0.0 4.0, 25.0, 0.0, 2.0, 70.0, 2.0, 0.0 1.0, 6.0, 0.0, 2.0, 70.0, 3.0, 2.0 1.0, 8.0, 0.0, 2.0, 70.0, 1.0, 0.0 3.0, 5.0, 0.0, 2.0, 70.0, 1.0, 0.0

		SSE: 0.0 1 Points: Cluster: 9 Center: [6.0, 9.0, 0.0, 2.0, 70.0, 2.0, 0.0] Max Dist. to Center: 0.0 Min Dist. to Center: 0.0 Avg Dist. to Center: 0.0 SSE: 0.0 1 Points: Cluster: 10 Center: [1.0, 6.0, 0.0, 2.0, 70.0, 3.0, 2.0] Max Dist. to Center: 0.0 Min Dist. to Center: 0.0 Avg Dist. to Center: 0.0 SSE: 0.0 1 Points: 11 clusters total	

Comments: Too many points to visualize. However, with this dataset we can see that kmeans and dbscan did the best, while hclustering failed to distinguish outliers.

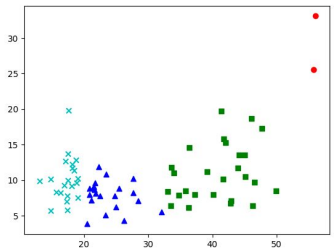
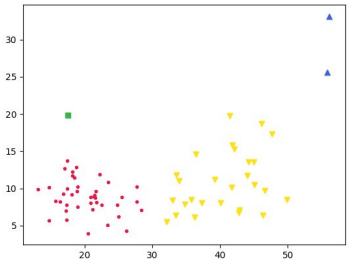
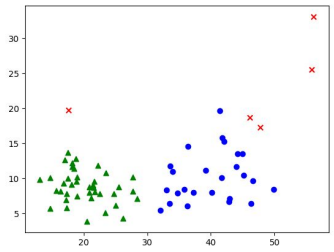
AccidentsSet03.csv

Couldn't be read by parse.csv when all others could. Dataset file must be errored in some way

Birth_death_rate.csv

Method	Kmeans	hclustering	dbscan
Params	Num Clusters: 4	Threshold: 100	Epsilon: 4, MinClusters: 5

Clusters	Cluster 0: Center: 55.95, 29.35 Max Dist. to Center: 3.752998800959041 Min Dist. to Center: 3.752998800959041 Avg Dist. to Center: 3.752998800959041 SSE: 28.17 2 Points: Cluster 1: Center: 23.838095238095235, 7.904761904761905 Max Dist. to Center: 8.604763222389025 Min Dist. to Center: 0.9675928005439097 Avg Dist. to Center: 3.2654021226656957 SSE: 278.5190476190477 21 Points: Cluster 2: Center: 40.830769230769235, 11.015384615384615 Max Dist. to Center: 9.411594259681811 Min Dist. to Center: 1.2623355830603658 Avg Dist. to Center: 5.822690183319602 SSE: 1013.1692307692308 26 Points: Cluster 3: Center: 17.247619047619047, 10.138095238095238 Max Dist. to Center: 9.668328498955693 Min Dist. to Center: 0.2876915707998713 Avg Dist. to Center: 2.8522554024647997	Cluster: 0 Center: [20.327499999999997, 8.84] Max Dist. to Center: 68.19285625000003 Min Dist. to Center: 0.32935625000000174 Avg Dist. to Center: 18.97389375 SSE: 26981.58962264531 40 Points: Cluster: 1 Center: [17.6, 19.8] Max Dist. to Center: 0.0 Min Dist. to Center: 0.0 Avg Dist. to Center: 0.0 SSE: 0.0 1 Points: Cluster: 2 Center: [40.507407407407406, 10.81111111111111] Max Dist. to Center: 98.89240054869678 Min Dist. to Center: 1.8605486968449827 Avg Dist. to Center: 41.328340192043896 SSE: 68774.94081347356 27 Points: Cluster: 3 Center: [55.95, 29.35] Max Dist. to Center: 14.085 Min Dist. to Center: 14.084999999999999 Avg Dist. to Center: 14.085 SSE: 396.77445 2 Points: 4 clusters total	Cluster 0: Center: 39.996, 10.235999999999999 Max Dist. to Center: 10.054994380903448 Min Dist. to Center: 1.250164789137816 Avg Dist. to Center: 5.646862030760941 Sum Squared Error: 913.7072 25 Points: Cluster 1: Center: 20.327499999999997, 8.84 Max Dist. to Center: 8.25789659961906 Min Dist. to Center: 0.5738956786733994 Avg Dist. to Center: 3.8522611792393584 Sum Squared Error: 758.95575 40 Points: Outliers: Percentage of data: 7.14 5 Points: 55.8, 25.6 56.1, 33.1 46.1, 18.7 47.7, 17.3 17.6, 19.8

	SSE: 245.9819047619048 21 Points:		
			

Comments: All methods did reasonably well. We'd argue hclustering performed best here as it was able to distinguish between the two large clusters as well as the outlier without splitting any clusters like kmeans did. DBSCAN marked some points outliers that didn't seem too far out.

economy.csv

Method	Kmeans	hclustering	dbscan
Params	Num Clusters: 10	Threshold: 20	Epsilon: 3, MinClusters: 2
Clusters	Cluster 0: Center: 14.0, 14.0, 11.0, 16.0, 17.0, 17.0, 20.0, 16.0, 12.0, 15.0 Max Dist. to Center: 0.0 Min Dist. to Center: 0.0 Avg Dist. to Center: 0.0 SSE: 0.0 1 Points: Cluster 1: Center: 8.0, 7.0, 6.0, 5.0, 7.0, 9.0, 10.0, 10.0, 8.0, 8.0 Max Dist. to Center: 0.0 Min Dist. to Center: 0.0 Avg Dist. to Center: 0.0 SSE: 0.0 1 Points: Cluster 2:	Cluster: 0 Center: [10.25, 9.25, 8.833333333333334, 9.75, 9.666666666666666, 11.75, 13.0, 13.75, 12.25, 12.166666666666666] Max Dist. to Center: 13.20833333333333 Min Dist. to Center: 1.0416666666666667 Avg Dist. to Center: 7.291666666666667 SSE: 901.548611111111 12 Points: Cluster: 1 Center: [9.5, 9.0, 8.5, 8.5, 8.5, 9.5, 10.0, 11.0, 10.0, 10.5] Max Dist. to Center: 3.5 Min Dist. to Center: 3.5	Cluster 0: Center: 10.166666666666666, 9.166666666666666, 9.0, 9.833333333333334, 10.0, 11.833333333333334, 12.833333333333334, 13.166666666666666, 12.0, 12.0 Max Dist. to Center: 3.1885210782848317 Min Dist. to Center: 0.4082482904638615 Avg Dist. to Center: 2.070401157509524 Sum Squared Error: 31.0 6 Points: Cluster 1: Center: 8.5, 6.5, 5.5, 7.75,

	Center: 13.0, 12.0, 11.0, 12.0, 12.0, 14.0, 18.0, 21.0, 18.0, 17.0 Max Dist. to Center: 0.0 Min Dist. to Center: 0.0 Avg Dist. to Center: 0.0 SSE: 0.0 1 Points: Cluster 3: Center: 9.0, 4.0, 4.0, 6.0, 8.0, 10.0, 10.0, 10.0, 9.0, 15.0 Max Dist. to Center: 0.0 Min Dist. to Center: 0.0 Avg Dist. to Center: 0.0 SSE: 0.0 1 Points: Cluster 4: Center: 10.75, 9.25, 8.75, 9.0, 8.75, 10.0, 10.5, 11.0, 9.5, 10.5 Max Dist. to Center: 3.391164991562634 Min Dist. to Center: 2.449489742783178 Avg Dist. to Center: 2.8046479039720986 SSE: 32.0 4 Points: Cluster 5: Center: 13.5, 12.5, 13.0, 12.5, 12.5, 13.5, 14.5, 14.5, 13.5, 13.5 Max Dist. to Center: 1.8027756377319946 Min Dist. to Center: 1.8027756377319946 Avg Dist. to Center: 1.8027756377319946 SSE: 6.499999999999999 2 Points: Cluster 6: Center: 8.0, 6.0, 5.0, 6.0, 6.0, 9.0, 11.0, 10.0, 8.0, 9.0	Avg Dist. to Center: 3.5 SSE: 24.5 2 Points: Cluster: 2 Center: [8.5, 6.5, 5.5, 7.75, 7.75, 10.25, 12.5, 14.25, 12.0, 12.0] Max Dist. to Center: 7.25 Min Dist. to Center: 1.25 Avg Dist. to Center: 4.25 SSE: 94.75 4 Points: Cluster: 3 Center: [13.0, 10.0, 9.0, 9.0, 9.0, 10.0, 10.0, 10.0, 8.0, 9.0] Max Dist. to Center: 0.0 Min Dist. to Center: 0.0 Avg Dist. to Center: 0.0 SSE: 0.0 1 Points: Cluster: 4 Center: [8.0, 6.5, 5.5, 5.5, 6.5, 9.0, 10.5, 10.0, 8.0, 8.5] Max Dist. to Center: 1.5 Min Dist. to Center: 1.5 Avg Dist. to Center: 1.5 SSE: 4.5 2 Points: Cluster: 5 Center: [13.5, 12.5, 13.0, 12.5, 12.5, 13.5, 14.5, 14.5, 13.5, 13.5] Max Dist. to Center: 3.25 Min Dist. to Center: 3.25 Avg Dist. to Center: 3.25 SSE: 21.125 2 Points: Cluster: 6 Center: [14.0, 14.0, 11.0, 16.0, 17.0, 17.0, 20.0, 16.0, 12.0, 15.0] Max Dist. to Center: 0.0	7.75, 10.25, 12.5, 14.25, 12.0, 12.0 Max Dist. to Center: 2.692582403567252 Min Dist. to Center: 1.118033988749895 Avg Dist. to Center: 1.9667111372878017 Sum Squared Error: 17.0 4 Points: Outliers: Percentage of data: 58.33 14 Points: 14.0, 14.0, 11.0, 16.0, 17.0, 17.0, 20.0, 16.0, 12.0, 15.0 13.0, 10.0, 9.0, 10.0, 10.0, 11.0, 14.0, 15.0, 13.0, 12.0 8.0, 7.0, 6.0, 5.0, 7.0, 9.0, 10.0, 10.0, 8.0, 8.0 13.0, 10.0, 9.0, 9.0, 9.0, 10.0, 10.0, 10.0, 8.0, 9.0 9.0, 4.0, 4.0, 6.0, 8.0, 10.0, 10.0, 10.0, 9.0, 15.0 13.0, 12.0, 11.0, 12.0, 12.0, 14.0, 18.0, 21.0, 18.0, 17.0 9.0, 9.0, 10.0, 9.0, 9.0, 10.0, 11.0, 15.0, 13.0, 12.0 9.0, 9.0, 9.0, 9.0, 9.0, 10.0, 11.0, 11.0, 11.0, 11.0 13.0, 13.0, 14.0, 13.0, 13.0, 13.0, 14.0, 14.0, 14.0, 14.0 8.0, 6.0, 5.0, 6.0, 6.0, 9.0, 11.0, 10.0, 8.0, 9.0 9.0, 8.0, 7.0, 9.0, 8.0, 12.0, 13.0, 13.0, 12.0, 13.0 10.0, 9.0, 8.0, 8.0, 8.0, 9.0, 9.0, 11.0, 9.0, 10.0 11.0, 11.0, 9.0, 10.0, 9.0, 13.0, 14.0, 16.0, 13.0, 13.0 14.0, 12.0, 12.0, 12.0, 12.0, 14.0, 15.0, 15.0, 13.0, 13.0

	Max Dist. to Center: 0.0 Min Dist. to Center: 0.0 Avg Dist. to Center: 0.0 SSE: 0.0 1 Points: Cluster 7: Center: 9.0, 8.0, 8.0, 8.75, 8.75, 11.25, 12.5, 14.5, 12.25, 12.0 Max Dist. to Center: 3.122498999199199 Min Dist. to Center: 2.3979157616563596 Avg Dist. to Center: 2.8591236359887935 SSE: 33.0 4 Points: Cluster 8: Center: 10.666666666666666, 9.833333333333334, 9.166666666666666, 10.0, 10.0, 11.833333333333334, 13.166666666666666, 13.833333333333334, 12.5, 12.166666666666666 Max Dist. to Center: 3.1928740101113355 Min Dist. to Center: 1.481365736219265 Avg Dist. to Center: 2.359119809065246 SSE: 35.833333333333336 6 Points: Cluster 9: Center: 8.666666666666666, 6.333333333333333, 5.0, 7.666666666666667, 7.666666666666667, 10.333333333333334, 12.666666666666666, 14.0, 12.333333333333334, 12.333333333333334 Max Dist. to Center:	Min Dist. to Center: 0.0 Avg Dist. to Center: 0.0 SSE: 0.0 1 Points: Cluster: 7 Center: [13.0, 12.0, 11.0, 12.0, 12.0, 14.0, 18.0, 21.0, 18.0, 17.0] Max Dist. to Center: 0.0 Min Dist. to Center: 0.0 Avg Dist. to Center: 0.0 SSE: 0.0 1 Points:	

	2.134374745810949 Min Dist. to Center: 1.1055415967851343 Avg Dist. to Center: 1.7085114752534034 SSE: 9.333333333333332 3 Points:		

Comments: From our testing, this dataset seems very scattered with only a few real clusters mixed in with lots of outliers. Kmeans and hclustering both seem to support this in their high SSE rates and lots of individual clusters which is corroborated by DBSCANs high outlier rate.

iris.csv

Method	Kmeans	hclustering	dbscan
Params	Num Clusters: 3	Threshold: 1.45	Epsilon: 3, MinClusters: 2
Clusters	Cluster 0: Center: 6.853846153846153, 3.076923076923076, 5.715384615384614, 2.053846153846153 Max Dist. to Center: 1.5515596276348889 Min Dist. to Center: 0.2394520371734348 Avg Dist. to Center: 0.7318458783535899 SSE: 25.413846153846155 39 Points: Cluster 1: Center: 5.005999999999999, 3.418, 1.4639999999999997, 0.24399999999999994 Max Dist. to Center: 1.239351443296049 Min Dist. to Center: 0.05993329625508689	Cluster: 0 Center: [5.005999999999999, 3.418, 1.464, 0.24399999999999986] Max Dist. to Center: 1.535992 Min Dist. to Center: 0.0035919999999999893 Avg Dist. to Center: 0.304808 SSE: 10.342247990400006 50 Points: Cluster: 1 Center: [6.926666666666668, 3.0999999999999996, 5.8533333333333335, 2.173333333333334] Max Dist. to Center: 1.9782666666666697 Min Dist. to Center: 0.044266666666666926 Avg Dist. to Center: 0.6370666666666668	Cluster 0: Center: 5.016326530612244, 3.4408163265306126, 1.4673469387755103, 0.24285714285714274 Max Dist. to Center: 1.1887820793371973 Min Dist. to Center: 0.06953765321388092 Avg Dist. to Center: 0.46919585594428453 Sum Squared Error: 13.673061224489793 49 Points: Cluster 1: Center: 6.220238095238092, 2.8964285714285714, 4.847619047619048, 1.6714285714285715 Max Dist. to Center: 1.8143427375115766

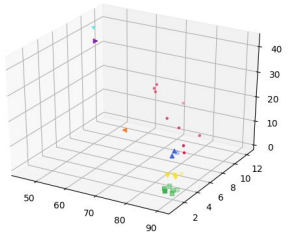
	Avg Dist. to Center: 0.4841322496689399 SSE: 15.240400000000003 50 Points: Cluster 2: Center: 5.88360655737705, 2.740983606557377, 4.388524590163935, 1.4344262295081966 Max Dist. to Center: 1.6468010735564589 Min Dist. to Center: 0.23571239518479165 Avg Dist. to Center: 0.7311084910642592 SSE: 38.29081967213114 61 Points:	SSE: 21.89615804444444 30 Points: Cluster: 2 Center: [5.9771428571428595, 2.7742857142857145, 4.499999999999999, 1.4628571428571426] Max Dist. to Center: 3.2262775510204103 Min Dist. to Center: 0.017706122448979435 Avg Dist. to Center: 0.7208653061224487 SSE: 67.19258414927117 70 Points:	Min Dist. to Center: 0.1688378980787914 Avg Dist. to Center: 0.8773287480916279 Sum Squared Error: 77.1854761904762 84 Points: Cluster 2: Center: 5.0, 2.3, 3.275, 1.025 Max Dist. to Center: 0.3758324094593226 Min Dist. to Center: 0.03535533905932725 Avg Dist. to Center: 0.2298114911361295 Sum Squared Error: 0.2949999999999998 4 Points: Outliers: Percentage of data: 8.67 13 Points: 4.5, 2.3, 1.3, 0.3 6.2, 2.2, 4.5, 1.5 6.3, 2.3, 4.4, 1.3 7.6, 3.0, 6.6, 2.1 4.9, 2.5, 4.5, 1.7 6.7, 2.5, 5.8, 1.8 7.2, 3.6, 6.1, 2.5 7.7, 3.8, 6.7, 2.2 7.7, 2.6, 6.9, 2.3 7.7, 2.8, 6.7, 2.0 7.9, 3.8, 6.4, 2.0 6.1, 2.6, 5.6, 1.4 7.7, 3.0, 6.1, 2.3

Comments: We knew the true clusters for this given this dataset. That meant kmeans was easy as we could just plug in 3 clusters and get them back and it did a great job. Using this info, hclustering was adjusted and was able to also get three very similar clusters. DBSCAN had a much harder time and struggled to generate more than 2 clusters without lots of outliers. This leads us to believe this is a very dense grouping without distinct cluster borders.

mammal_milk.csv

Method	Kmeans	hclustering	dbscan
Params	Num Clusters: 8	Threshold: 20	Epsilon: 3, MinClusters: 2
Clusters	Cluster 0: Center: 46.4, 9.7, 42.0, 0.0, 0.85 Max Dist. to Center: 0.0 Min Dist. to Center: 0.0 Avg Dist. to Center: 0.0 SSE: 0.0 1 Points: Cluster 1: Center: 86.825, 3.8000000000000003, 3.2750000000000004, 5.3500000000000005, 0.7775000000000001 Max Dist. to Center: 1.9036166236929117 Min Dist. to Center: 0.42925080081463024 Avg Dist. to Center: 1.3131731936925362 SSE: 8.351075000000002 4 Points: Cluster 2: Center: 44.9, 10.6, 34.9, 0.9, 0.53 Max Dist. to Center: 0.0 Min Dist. to Center: 0.0 Avg Dist. to Center: 0.0 SSE: 0.0 1 Points: Cluster 3: Center: 89.7, 2.125, 1.725, 6.25, 0.35 Max Dist. to Center: 1.642452434623296 Min Dist. to Center: 0.8070006195784445 Avg Dist. to Center:	Cluster: 0 Center: [73.54444444444444, 9.733333333333333, 12.777777777777779, 2.855555555555556, 1.25] Max Dist. to Center: 152.2836111111112 Min Dist. to Center: 1.6269444444444288 Avg Dist. to Center: 88.26722222222222 SSE: 99315.60767268516 9 Points: Cluster: 1 Center: [89.61666666666666, 1.75, 2.4833333333333334, 5.8999999999999995, 0.29] Max Dist. to Center: 8.289155555555572 Min Dist. to Center: 0.8687888888888936 Avg Dist. to Center: 3.2450777777777753 SSE: 99.61823383333348 6 Points: Cluster: 2 Center: [86.825, 3.8, 3.275, 5.3500000000000005, 0.7775000000000001] Max Dist. to Center: 3.6237562499999996 Min Dist. to Center: 0.18425625000000137 Avg Dist. to Center: 2.0877687500000004 SSE: 25.875549965781246 4 Points: Cluster: 3	Cluster 0: Center: 90.13333333333333, 2.1, 1.4000000000000001, 6.2, 0.4066666666666667 Max Dist. to Center: 0.9509585807080256 Min Dist. to Center: 0.43338461235053555 Avg Dist. to Center: 0.7366827596544202 Sum Squared Error: 1.7739333333333334 3 Points: Cluster 1: Center: 82.125, 6.300000000000001, 6.324999999999999, 4.5, 0.93 Max Dist. to Center: 1.806972606322518 Min Dist. to Center: 0.7427314454094439 Avg Dist. to Center: 1.2585998081677703 Sum Squared Error: 7.226800000000004 4 Points: Cluster 2: Center: 86.825, 3.8000000000000003, 3.2750000000000004, 5.3500000000000005, 0.7775000000000001 Max Dist. to Center: 1.9036166236929117 Min Dist. to Center: 0.42925080081463024 Avg Dist. to Center: 1.3131731936925362

	1.1084705175453013 SSE: 5.370799999999978 4 Points: Cluster 4: Center: 89.45, 1.0, 4.0, 5.2, 0.16999999999999998 Max Dist. to Center: 1.3990711204224053 Min Dist. to Center: 1.3990711204224053 Avg Dist. to Center: 1.3990711204224053 SSE: 3.9148000000000094 2 Points: Cluster 5: Center: 65.16666666666667, 10.733333333333334, 20.400000000000002, 2.2333333333333334, 1.2166666666666666 Max Dist. to Center: 1.2018504251546631 Min Dist. to Center: 0.5002777006601271 Avg Dist. to Center: 0.9484482783747646 SSE: 3.001666666666678 3 Points: Cluster 6: Center: 81.18571428571428, 7.42857142857143, 6.8999999999999995, 4.014285714285714, 0.9314285714285715 Max Dist. to Center: 5.935769124725188 Min Dist. to Center: 1.5280700112179693 Avg Dist. to Center: 2.6702085191443983 SSE: 63.5349142857143 7 Points: Cluster 7:	Center: [81.89999999999999, 6.033333333333334, 6.733333333333334, 4.766666666666667, 0.8733333333333334] Max Dist. to Center: 1.4320444444444445 Min Dist. to Center: 0.31467777777778083 Avg Dist. to Center: 0.9577555555555559 SSE: 3.418875779999998 3 Points: Cluster: 4 Center: [70.7, 3.6, 17.6, 5.6, 0.63] Max Dist. to Center: 0.0 Min Dist. to Center: 0.0 Avg Dist. to Center: 0.0 SSE: 0.0 1 Points: Cluster: 5 Center: [44.9, 10.6, 34.9, 0.9, 0.53] Max Dist. to Center: 0.0 Min Dist. to Center: 0.0 Avg Dist. to Center: 0.0 SSE: 0.0 1 Points: Cluster: 6 Center: [46.4, 9.7, 42.0, 0.0, 0.85] Max Dist. to Center: 0.0 Min Dist. to Center: 0.0 Avg Dist. to Center: 0.0 SSE: 0.0 1 Points:	Sum Squared Error: 8.351075000000002 4 Points: Cluster 3: Center: 65.16666666666667, 10.733333333333334, 20.400000000000002, 2.233333333333333, 1.2166666666666666 Max Dist. to Center: 1.201850425154663 Min Dist. to Center: 0.5002777006601273 Avg Dist. to Center: 0.9484482783747646 Sum Squared Error: 3.0016666666666785 3 Points: Outliers: Percentage of data: 44.0 11 Points: 88.5, 1.4, 3.5, 6.0, 0.24 88.4, 2.2, 2.7, 6.4, 0.18 90.4, 0.6, 4.5, 4.4, 0.1 81.9, 7.4, 7.2, 2.7, 0.85 81.6, 10.1, 6.3, 4.4, 0.75 76.3, 9.3, 9.5, 3.0, 1.2 70.7, 3.6, 17.6, 5.6, 0.63 71.3, 12.3, 13.1, 1.9, 2.3 72.5, 9.2, 12.6, 3.3, 1.4 46.4, 9.7, 42.0, 0.0, 0.85 44.9, 10.6, 34.9, 0.9, 0.53

	Center: 71.5, 8.366666666666667, 14.433333333333335, 3.6, 1.4433333333333334 Max Dist. to Center: 6.168500628191587 Min Dist. to Center: 2.268795568872026 Avg Dist. to Center: 4.336781350051774 SSE: 64.1106 3 Points:		
			

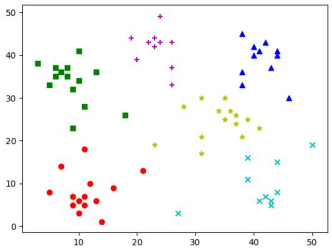
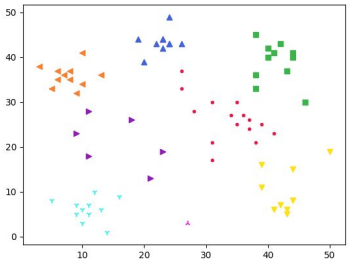
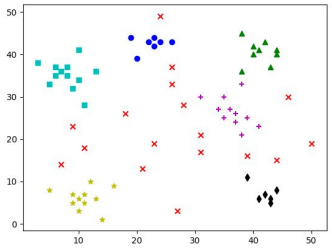
Comments: Looking at this dataset, you can see theres around 15-20 potential categories corresponding to each animal, however, only one data point for each in many cases. This is not much to work with. You can see a graph for hclustering which shows 3 dimension of the 5 and you can see it seemed to do a decent job clustering into different groups with 2 or 3 larger ones showing similarities in certain animal groups. This is reflected in both kmeans and DBSCANs outputs, showing a few clusters with several points but mostly single point clusters.

many_clusters.csv

Method	Kmeans	hclustering	dbscan
Params	Num Clusters: 6	Threshold: 200	Epsilon: 5, MinClusters: 4

Clusters	Cluster 0: Center: 11.357142857142858, 8.0 Max Dist. to Center: 10.862075946961106 Min Dist. to Center: 1.0618620533798935 Avg Dist. to Center: 4.97272032239714 SSE: 467.2142857142857 14 Points: Cluster 1: Center: 41.27272727272727, 38.90909090909091 Max Dist. to Center: 10.08558418062448 Min Dist. to Center: 1.676280810416891 Avg Dist. to Center: 4.396794888519616 SSE: 277.09090909090907 11 Points: Cluster 2: Center: 8.785714285714286, 33.642857142857146 Max Dist. to Center: 11.971479713494581 Min Dist. to Center: 1.2657175104763807 Avg Dist. to Center: 5.028802899674039 SSE: 495.57142857142856 14 Points: Cluster 3: Center: 41.2, 9.6 Max Dist. to Center: 15.658863304850708 Min Dist. to Center: 2.607680962081062 Avg Dist. to Center: 6.250851785588283 SSE: 572.0 10 Points:	Cluster: 0 Center: [33.666666666666664, 26.266666666666666] Max Dist. to Center: 173.98222222222222 Min Dist. to Center: 0.6488888888888918 Avg Dist. to Center: 43.75111111111111 SSE: 60550.14577777777 15 Points: Cluster: 1 Center: [41.27272727272727, 38.90909090909091] Max Dist. to Center: 101.71900826446276 Min Dist. to Center: 2.8099173553719083 Avg Dist. to Center: 25.190082644628102 SSE: 15790.518407212614 11 Points: Cluster: 2 Center: [42.77777777777778, 10.333333333333334] Max Dist. to Center: 127.27160493827158 Min Dist. to Center: 6.938271604938272 Avg Dist. to Center: 33.28395061728395 SSE: 20940.600823045264 9 Points: Cluster: 3 Center: [22.8, 43.4] Max Dist. to Center: 32.80000000000002 Min Dist. to Center: 0.4000000000000014 Avg Dist. to Center: 9.2 SSE: 2152.960000000001 10 Points: Cluster: 4	Cluster 0: Center: 22.428571428571427, 42.57142857142857 Max Dist. to Center: 4.3189189879517045 Min Dist. to Center: 0.6060915267313268 Avg Dist. to Center: 2.3159871090114765 Sum Squared Error: 51.42857142857144 7 Points: Cluster 1: Center: 41.111111111111114, 40.55555555555556 Max Dist. to Center: 5.516529594358407 Min Dist. to Center: 0.4581228472908504 Avg Dist. to Center: 2.99514435732394 Sum Squared Error: 105.11111111111109 9 Points: Cluster 2: Center: 8.0, 35.166666666666664 Max Dist. to Center: 7.769241347204441 Min Dist. to Center: 0.1666666666666643 Avg Dist. to Center: 3.509204319867297 Sum Squared Error: 203.66666666666666 12 Points: Cluster 3: Center: 36.45454545454545, 26.454545454545453 Max Dist. to Center: 6.725429722975157 Min Dist. to Center:

	Cluster 4: Center: 23.3, 41.7 Max Dist. to Center: 9.109335870413389 Min Dist. to Center: 0.424264068711927 Avg Dist. to Center: 4.005879958716357 SSE: 228.20000000000005 10 Points: Cluster 5: Center: 34.0, 24.5 Max Dist. to Center: 12.298373876248844 Min Dist. to Center: 1.118033988749895 Avg Dist. to Center: 5.32151913461238 SSE: 499.5 14 Points:	Center: [7.7272727272727275, 35.81818181818182] Max Dist. to Center: 32.01652892561981 Min Dist. to Center: 0.5619834710743798 Avg Dist. to Center: 12.528925619834713 SSE: 3140.900075131479 11 Points: Cluster: 5 Center: [15.5, 21.166666666666668] Max Dist. to Center: 96.94444444444446 Min Dist. to Center: 29.6111111111111 Avg Dist. to Center: 55.055555555555564 SSE: 21467.944444444445 6 Points: Cluster: 6 Center: [10.909090909090908, 6.090909090909091] Max Dist. to Center: 38.56198347107437 Min Dist. to Center: 0.8347107438016514 Avg Dist. to Center: 13.801652892561982 SSE: 4371.643876784373 11 Points: Cluster: 7 Center: [27.0, 3.0] Max Dist. to Center: 0.0 Min Dist. to Center: 0.0 Avg Dist. to Center: 0.0 SSE: 0.0 1 Points:	0.710022697809696 Avg Dist. to Center: 3.6253942371119194 Sum Squared Error: 193.45454545454544 11 Points: Cluster 4: Center: 10.909090909090908, 6.090909090909091 Max Dist. to Center: 6.20982958470475 Min Dist. to Center: 0.9136250564655347 Avg Dist. to Center: 3.14885882918028 Sum Squared Error: 151.8181818181818 11 Points: Cluster 5: Center: 42.0, 7.166666666666667 Max Dist. to Center: 4.867693955503411 Min Dist. to Center: 0.16666666666666696 Avg Dist. to Center: 2.1100853808578495 Sum Squared Error: 38.83333333333333 6 Points: Outliers: Percentage of data: 23.29 17 Points: 24.0, 49.0 26.0, 37.0 26.0, 33.0 46.0, 30.0 28.0, 28.0 18.0, 26.0 9.0, 23.0 31.0, 21.0 23.0, 19.0 50.0, 19.0

			11.0, 18.0 31.0, 17.0 39.0, 16.0 44.0, 15.0 7.0, 14.0 21.0, 13.0 27.0, 3.0
			

Comments: This dataset would be tough for a human to draw the lines between clusters. All algorithms did a good job but you can easily see the differences in technique in the graphs. Kmeans focused on the dense parts being the center and added the outliers to the closest cluster whereas hclustering treated the purple triangle cluster a bit differently and centered the cluster in the middle. This is also a great dataset to see DBSCANs strengths as it identifies the dense clusters and treats all the other questionable points as outliers.

planets.csv

Method	Kmeans	hclustering	dbscan
Params	Num Clusters: 9	Threshold: 600	Epsilon: 10, MinClusters: 1
Clusters	Cluster 0: Center: 339.625, 16.067, 2.7335 Max Dist. to Center: 0.0 Min Dist. to Center: 0.0 Avg Dist. to Center: 0.0 SSE: 0.0 1 Points: Cluster 1:	Cluster: 0 Center: [120.49914285714286, 4.413285714285714, 2.2864428571428577] Max Dist. to Center: 1575.7574232124489 Min Dist. to Center: 91.75265683673456 Avg Dist. to Center: 685.119696000408	Cluster 0: Center: 132.1965, 3.7872500000000002, 2.293875 Max Dist. to Center: 4.409794476857743 Min Dist. to Center: 0.9273509735396841 Avg Dist. to Center: 2.3100167918366497

	Center: 34.2, 12.5, 2.82 Max Dist. to Center: 0.0 Min Dist. to Center: 0.0 Avg Dist. to Center: 0.0 SSE: 0.0 1 Points: Cluster 2: Center: 55.144, 4.542, 3.0343 Max Dist. to Center: 0.0 Min Dist. to Center: 0.0 Avg Dist. to Center: 0.0 SSE: 0.0 1 Points: Cluster 3: Center: 338.333, 16.773, 2.7465 Max Dist. to Center: 0.0 Min Dist. to Center: 0.0 Avg Dist. to Center: 0.0 SSE: 0.0 1 Points: Cluster 4: Center: 126.48833333333333, 3.4135000000000004, 2.5851166666666665 Max Dist. to Center: 11.455597315783614 Min Dist. to Center: 3.857429091340786 Avg Dist. to Center: 7.779701268893809 SSE: 421.93789020166616 6 Points: Cluster 5: Center: 64.75, 5.2, 2.8 Max Dist. to Center: 4.875715332133328 Min Dist. to Center: 4.875715332133321 Avg Dist. to Center: 4.875715332133325 SSE: 47.54519999999995	SSE: 6201408.178859425 7 Points: Cluster: 1 Center: [164.1, 10.0, 1.93] Max Dist. to Center: 0.0 Min Dist. to Center: 0.0 Avg Dist. to Center: 0.0 SSE: 0.0 1 Points: Cluster: 2 Center: [115.072, 2.666, 3.1676] Max Dist. to Center: 0.0 Min Dist. to Center: 0.0 Avg Dist. to Center: 0.0 SSE: 0.0 2 Points: Cluster: 3 Center: [89.9, 2.1, 3.35] Max Dist. to Center: 0.0 Min Dist. to Center: 0.0 Avg Dist. to Center: 0.0 SSE: 0.0 1 Points: Cluster: 4 Center: [194.6, 1.8, 3.02] Max Dist. to Center: 0.0 Min Dist. to Center: 0.0 Avg Dist. to Center: 0.0 SSE: 0.0 1 Points: Cluster: 5 Center: [65.68599999999999, 5.3855, 2.883575] Max Dist. to Center: 155.582676030625 Min Dist. to Center: 15.794719530625011 Avg Dist. to Center: 79.20770809187496 SSE: 38097.86920782941 4 Points:	Sum Squared Error: 28.2457606175 4 Points: Cluster 1: Center: 338.97900000000004, 16.42, 2.74 Max Dist. to Center: 0.736184250035287 Min Dist. to Center: 0.7361842500352388 Avg Dist. to Center: 0.7361842500352629 Sum Squared Error: 1.0839344999999647 2 Points: Cluster 2: Center: 73.46457142857143, 4.698, 2.751985714285714 Max Dist. to Center: 18.3234105675037 Min Dist. to Center: 3.865007371775327 Avg Dist. to Center: 10.31079115176507 Sum Squared Error: 946.4240339628575 7 Points: Cluster 3: Center: 115.072, 2.666, 3.1676 Max Dist. to Center: 0.0 Min Dist. to Center: 0.0 Avg Dist. to Center: 0.0 Sum Squared Error: 0.0 2 Points: Outliers: Percentage of data: 21.05 4 Points: 153.1, 6.5, 2.45 194.6, 1.8, 3.02 164.1, 10.0, 1.93 34.2, 12.5, 2.82

	2 Points: Cluster 6: Center: 194.6, 1.8, 3.02 Max Dist. to Center: 0.0 Min Dist. to Center: 0.0 Avg Dist. to Center: 0.0 SSE: 0.0 1 Points: Cluster 7: Center: 82.402, 4.486, 2.6574 Max Dist. to Center: 7.898904655710187 Min Dist. to Center: 1.670329979375332 Avg Dist. to Center: 4.009874327276631 SSE: 91.00925032000013 4 Points: Cluster 8: Center: 158.6, 8.25, 2.19 Max Dist. to Center: 5.777551384453452 Min Dist. to Center: 5.777551384453452 Avg Dist. to Center: 5.777551384453452 SSE: 66.7602 2 Points:	Cluster: 6 Center: [34.2, 12.5, 2.82] Max Dist. to Center: 0.0 Min Dist. to Center: 0.0 Avg Dist. to Center: 0.0 SSE: 0.0 1 Points: Cluster: 7 Center: [338.97900000000004, 16.42, 2.74] Max Dist. to Center: 0.541967250000018 Min Dist. to Center: 0.541967249999947 Avg Dist. to Center: 0.5419672499999826 SSE: 0.5874570001450872 2 Points:	[[153.1, 6.5, 2.45], [194.6, 1.8, 3.02], [164.1, 10.0, 1.93], [34.2, 12.5, 2.82]]

Comments: This dataset is strange, we can only guess that it is the locations of planets in space. Due to this, there aren't large clusters and there are many "outlier" planets. Even the clusters with 3 or 4 points have a very high SSE due to the distances being very large here. Our methods seem to tell us that there are 2 or 3 larger clusters with multiple planets, and maybe 2 or 3 more with two planets. The rest seem to be lonely.

Analysis:

Overall our findings were different for every dataset and we learned that it’s way easier to determine the accuracy of your methods when you can easily visualize them. Some of the trends we saw were that kmeans and hierarchal clustering often produced similar results or similar amounts of clusters but in slightly different configurations. Kmeans seemed to focus the center on the densest areas where hclustering put the center in empty space surrounded by a wall of coordinates. DBSCAN on the other hand did a much better job at showing the important clusters and throwing out all of the noise as outliers. In the end, kmeans was best for finding clusters when you knew how many you wanted roughly, hclustering was best for circular or ring shaped clusters, and DBSCAN was best at finding dense clusters surrounded by noise.