CSC 466 Lab 4 Report

Authors:
Sophia Chung, spchung@calpoly.edu
Andrew Kerr, adkerr@calpoly.edu

Introduction

The goal of this study is to analyze and compare the results of three clustering algorithms: k-means clustering, agglomerative hierarchical clustering, and DBSCAN. These algorithms will be run through five distinct data sets:

- 4clusters: 39 data points with 2 features (known 4 clusters)
- planets: 19 data points with 3 features
- iris: 150 data points with 4 features (known 3 clusters)
- mammal_milk: 25 data points with 5 features
- AccidentsSet03: 62 data points with 5 features

Implementations

k-Means

The goal behind k-means clustering, like other clustering algorithms, is to group similar data points together and assign them to clusters. The "k" in k-means stands for the number of clusters the algorithm will create, which is specified by the user. This algorithm has a few simple steps:

1. Choose initial cluster centroids randomly from the data points or through k-Means++
    a. Additionally, the user may decided to standardize the data using min-max standardization
2. Assign each data point to the cluster with the closest centroid
    a. The distance metric, selected by the user, is:
        i. Euclidean distance,
        ii. Manhattan distance, or
        iii. Cosine similarity
3. Recalculate the centroids of each cluster by taking the mean of all data points assigned to the cluster
4. Repeat steps 2 and 3 until a stoppage condition is met:
    a. The cluster centroids do not change from the previous iterations centroids
    b. The cluster centroids do not change enough from the previous iterations centroids
        i. The amount of change, epsilon or stoppage threshold, is specified by the user

Agglomerative Hierarchical Clustering

Hierarchical clustering aims to organize data points into a hierarchy of clusters. Unlike k-Means, this algorithm does not create disjoint clusters, but seeks to organize the data points by similarity. Agglomerative hierarchical clustering in particular works as follows:

1. Treat each data point as its own cluster
2. Calculate a distance matrix between all clusters
    a. The distance metric is selected by the user can be:
        i. Euclidean distance,
        ii. Manhattan distance, or
        iii. Cosine similarity
3. Merge the two most similar clusters together
    a. Update the values of the distance matrix using single-link method; the distance between two clusters is the distance between the two closest points in the clusters
4. Repeat step 3 until a single cluster remains

The result of this process is a dendrogram, a tree-like structure where each leaf is a single data point and each edge contains the distance the two child clusters were merged at. This allows us to "cut" the tree at a specific threshold, resulting in clusters as or more similar than the threshold value.

DBSCAN

DBSCAN finds core points in regions of high density and recursively expands clusters from them. DBSCAN allows for easy outlier detection as outliers (or noise points) are the remaining data points that do not get placed into clusters. This algorithm works as follows:

1. The distance metric, selected by the user, is:
    a. Euclidean distance,
    b. Manhattan distance, or
    c. Cosine similarity
2. Find core points
    a. Calculate a matrix of pairwise distances between the points
    b. Find the epsilon-neighborhood of each data point *d*; an epsilon-neighborhood contains all data points within some user-specified distance epsilon from *d*
    c. If the epsilon-neighborhood contains at least a certain number of data points (min_points), also specified by the user, then *d* is a core point
3. Starting with an arbitrarily but consistently chosen core point, we construct our clusters
    a. Data points in the core point's epsilon-neighborhood are added to the cluster

b. Add points in the epsilon-neighborhoods of other core points now in the cluster as well

c. Repeat until the current cluster cannot be expanded further, then repeat step 3 until all core points have been classified

Note: The DBSCAN clusters are numbered starting with 1 instead of 0. All data points are initialized as “belonging” to cluster 0. Once the algorithm is run then, as a result, cluster 0 denotes outliers.

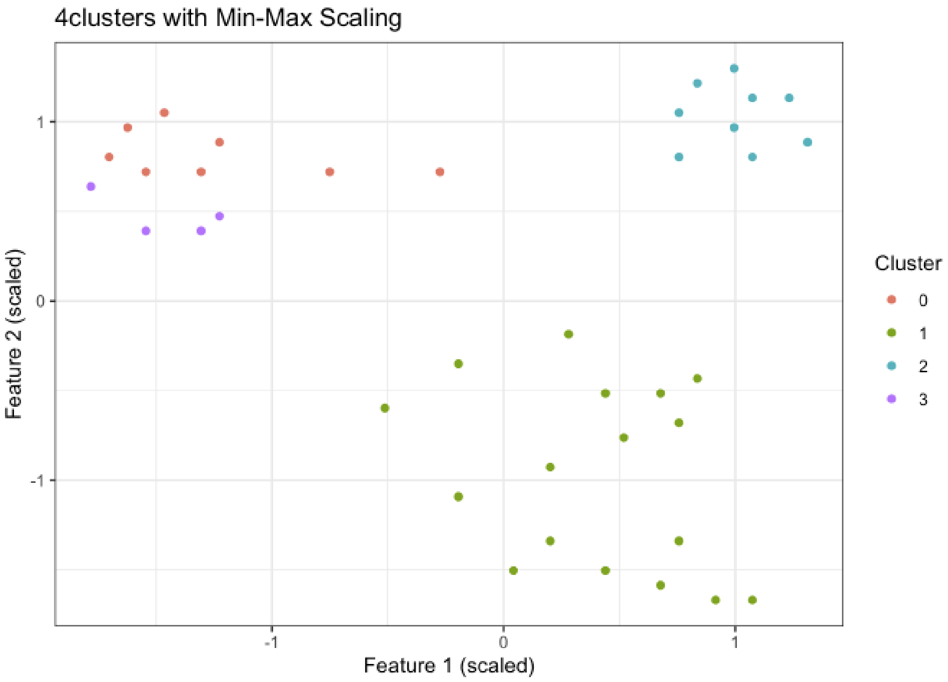
Results

4clusters

k-Means

Knowing that our data has 4 clusters, we set k equal to 4 and tuned the stoppage threshold. We wanted to increase the stoppage threshold as much as possible to make the clusters accurate, but without compromising metrics like average distance to the center and SSE for a cluster.

- k = 4
- Stoppage threshold = 0.2

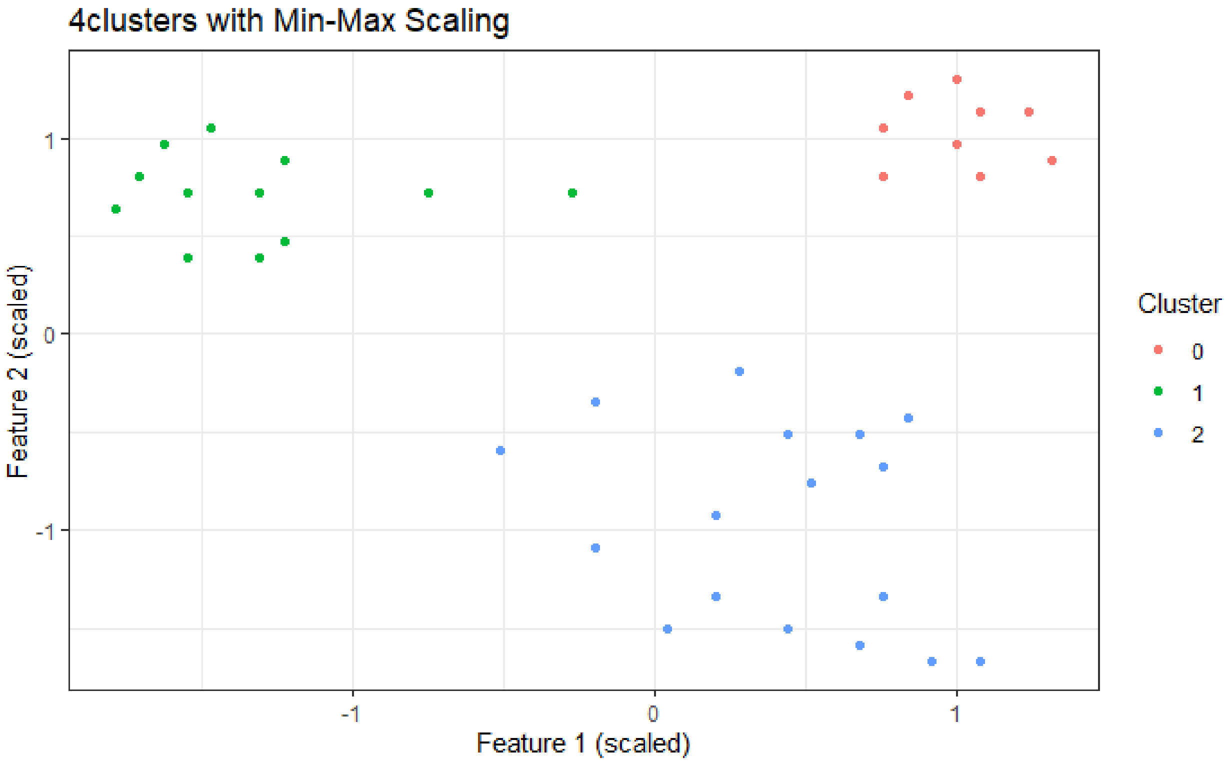


Cluster	Center	Avg Dist. to Center	SSE for Cluster
0	(0.1905, 0.8452)	0.1332	0.1908
1	(0.7163, 0.2049)	0.2059	0.8542
2	(0.8769, 0.8694)	0.089	0.0812
3	(0.1496, 0.6759)	0.0825	0.0418

Agglomerative Hierarchical Clustering

We tested thresholds between 0.1 and 0.3. With a threshold of 0 there were 11 clusters, while thresholds of 0.2 and 0.3 resulted in 3 clusters. Going up to a threshold of 0.35 resulted in 2 clusters, and 0.4 only 1 cluster. Since one of the clusters is very close to another, or could even be seen as two outlier points, we determined that the best this algorithm could do was catch 3 of the 4 clusters.

- Threshold: any value between 0.2 and 0.3

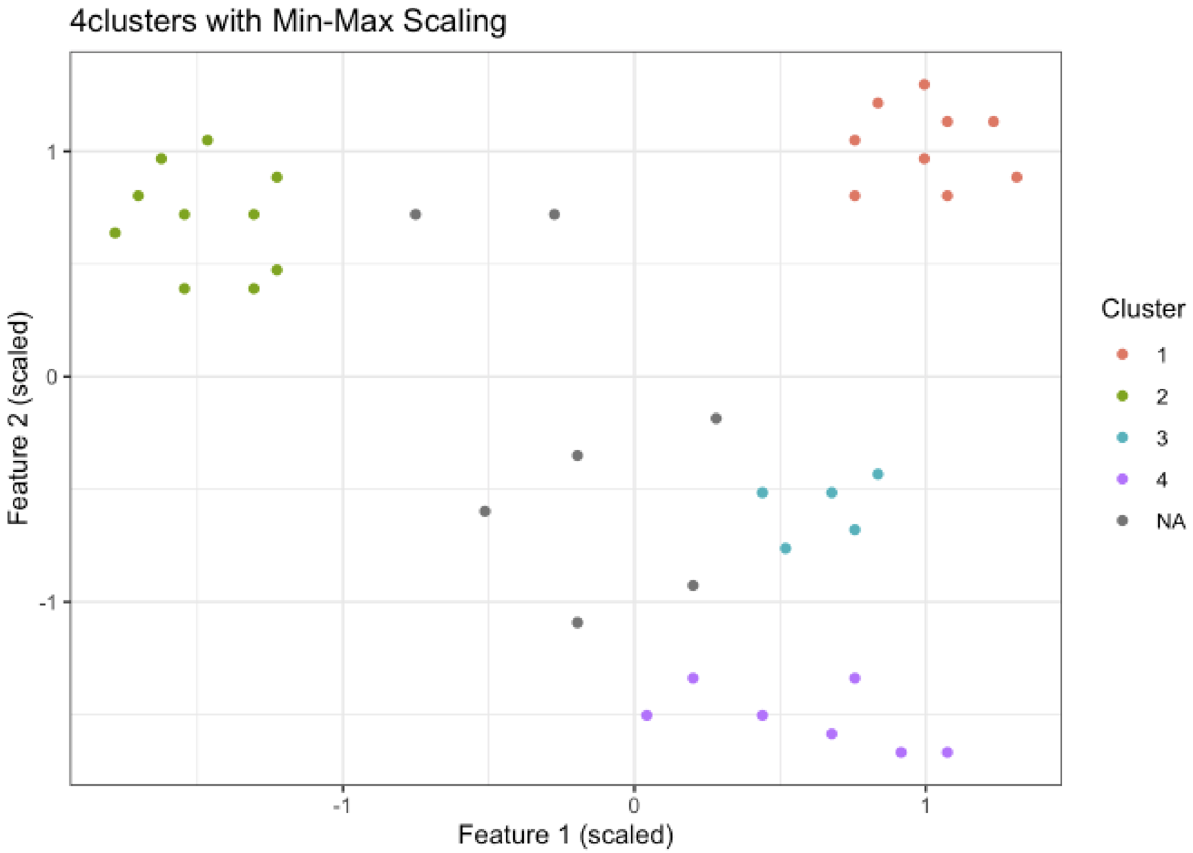


Cluster	Center	Avg Dist. to Center	SSE for Cluster
0	(0.9003, 0.9105)	0.0776	0.0611
1	(0.1517, 0.8009)	0.1272	0.2650
2	(0.6966, 0.2299)	0.2033	0.8360

DBSCAN

For this data set, we started with an epsilon of 0.1 and min_points of 3. Although we found 4 clusters as desired, we were left with too many noise points. To decrease the amount of noise points, we decreased min_points. Doing so made our algorithm less strict in defining core points and thus resulted in fewer noise points. Here are our final parameters:

- Epsilon: 0.1
- Min points: 2



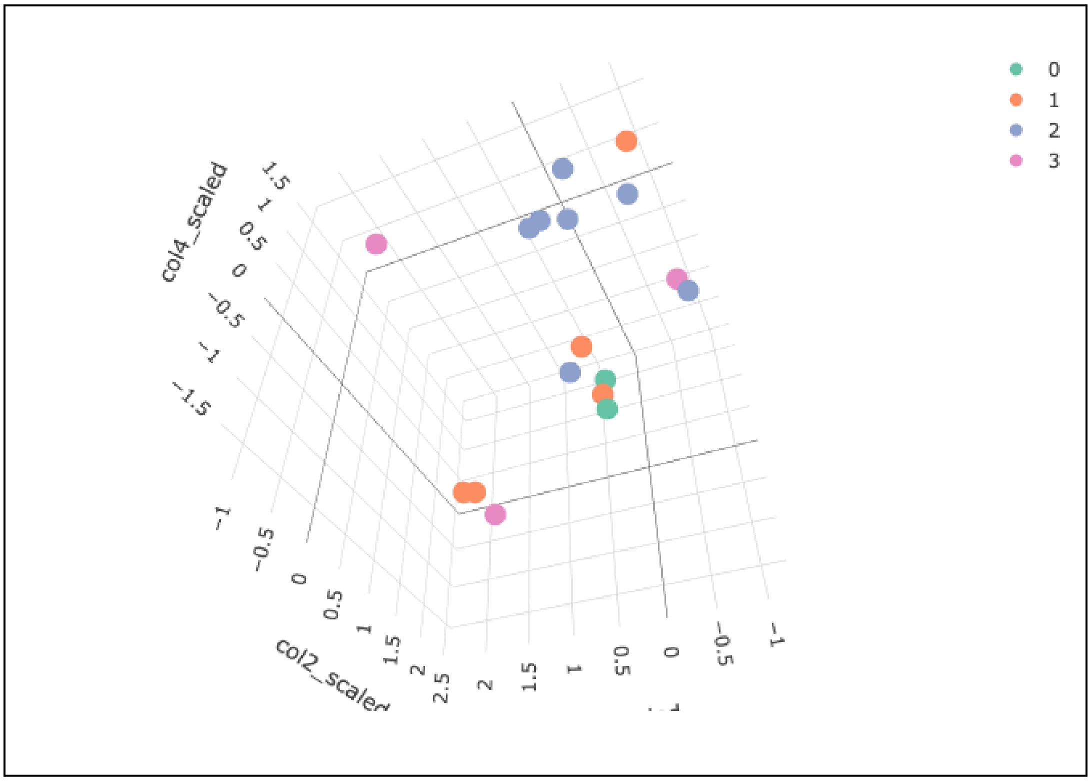
Cluster	Center	Avg Dist. to Center	SSE for Cluster
1	(0.9003, 0.9105)	0.0776	0.0611
2	(0.1, 0.8)	0.0918	0.0927
3	(0.7846, 0.3667)	0.0593	0.0196
4	(0.7656, 0.0516)	0.1092	0.1017

Planets

k-Means

We arbitrarily chose a k-value of 4 and similarly tuned our stoppage threshold. We realized that any stoppage threshold larger than 0.4 did not change any of our cluster metrics, which was unique to only this data set.

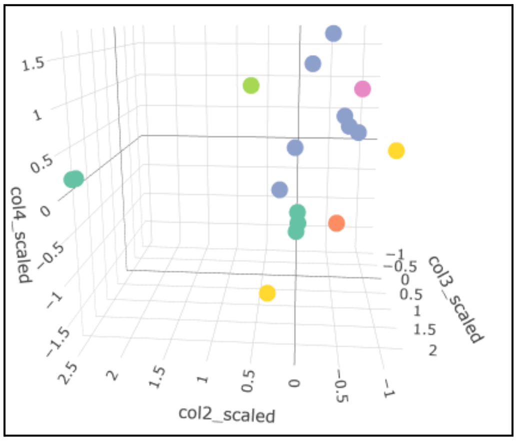
- k = 4
- Stoppage threshold = 0.4



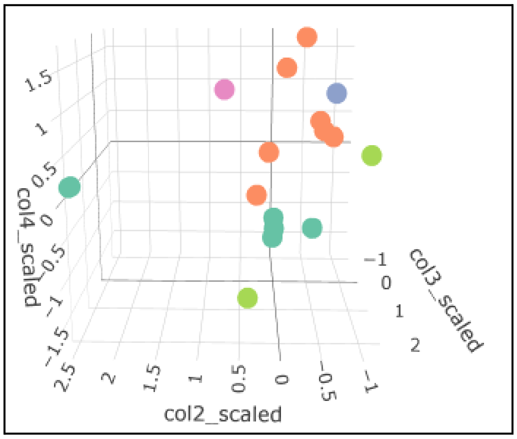
Cluster	Center	Avg Dist. to Center	SSE for Cluster
0	(0.9958, 1.0, 0.575)	0.0229	0.0021
1	(0.1824, 0.0697, 1.0)	0.285	0.7898
2	(0.4253, 0.5706, 0.0)	0.3675	1.1106
3	(0.0, 0.7291, 0.6268)	0.2815	0.3571

Agglomerative Hierarchical Clustering

For this data set, we tested with a threshold of 0.3 and a threshold of 0.35.



Threshold = 0.30



Threshold = 0.35

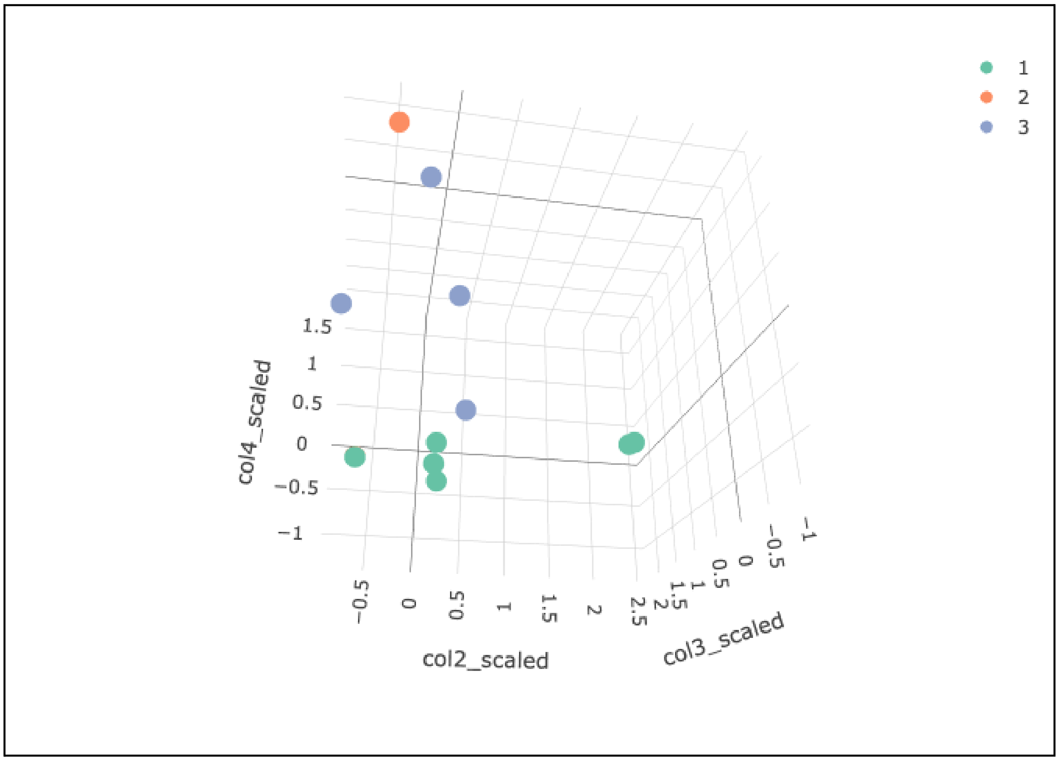
Any smaller of a threshold yielded too many clusters, and any larger of a threshold yielded only 1 or 2 clusters. From the 3D-scatter plots above we noticed that the 2 points to the left, which seem to be outliers, were always in the cluster with the majority of the data points, but never their own cluster.

- Threshold: 0.35

Cluster	Center	Avg Dist. to Center	SSE for Cluster
0	(0.2826, 0.2164, 0.2510)	0.1505	0.2056
1	(0.2062, 0.1805, 0.7746)	0.2058	0.3751
2	(0.4253, 0.5706, 0.0)	0	0
3	(0.0, 0.7291, 0.6268)	0	0
4	(0.9979, 0.9776, 0.5704)	0.0229	0.0011

DBSCAN

We started by testing with an epsilon of 0.1 and min_points of 2. This gave us 1 tiny cluster with most points ending up as noise points. We increased epsilon to 0.2 to broaden our neighborhoods, and this gave us 3 clusters with some noise points. We tried increasing our epsilon once more to 0.3, which resulted in 2 clusters and even less noise points. We decided to go with an epsilon of 0.2 because it resulted in more clusters with overall smaller average distances to the center.



- Epsilon: 0.2
- Min points: 2

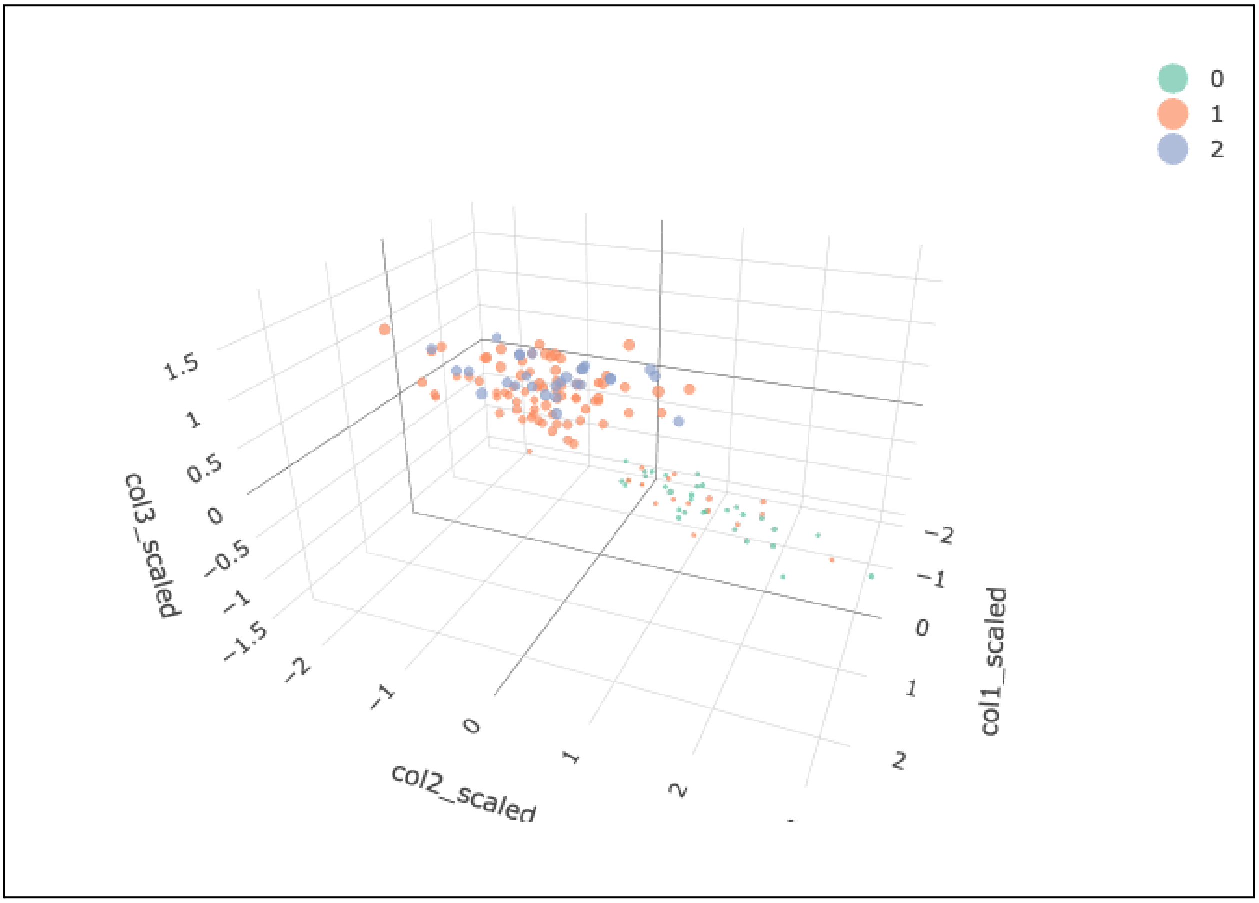
Cluster	Center	Avg Dist. to Center	SSE for Cluster
1	(0.2743, 0.2525, 0.2142)	0.1099	0.0912
2	(0.2373, 0.0937, 0.9144)	0.0697	0.01639
3	(0.1031, 0.278, 0.6715)	0.088	0.0329

Iris

k-Means

Knowing that our data has 3 clusters, we set k equal to 3 and similarly tuned the stoppage threshold.

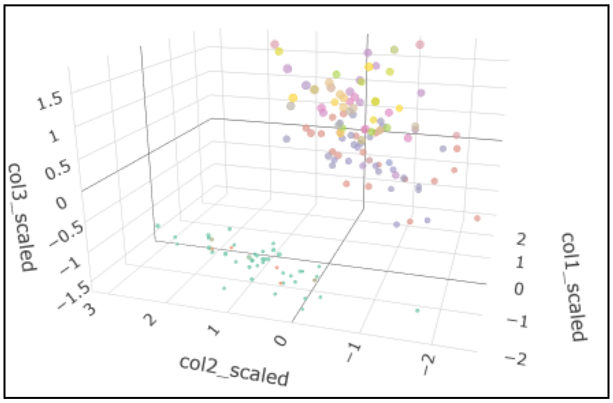
- k = 3
- Stoppage threshold = 0.4



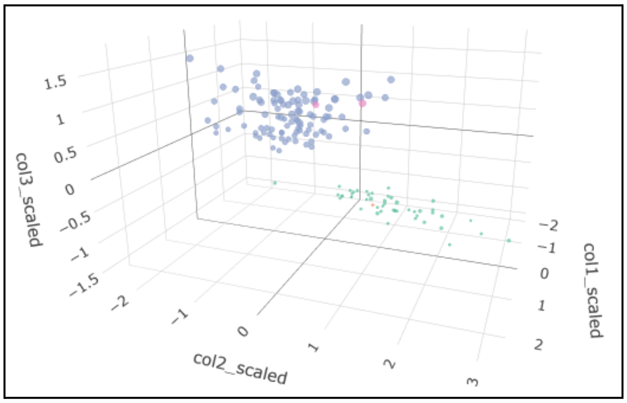
Cluster	Center	Avg Dist. to Center	SSE for Cluster
0	(0.2246, 0.4116, 0.2303, 0.2019)	0.3111	3.3876
1	(0.6036, 0.398, 0.7125, 0.7169)	0.2735	7.8362
2	(0.2996, 0.7827, 0.086, 0.0744)	0.1649	0.8768

Agglomerative Hierarchical Clustering

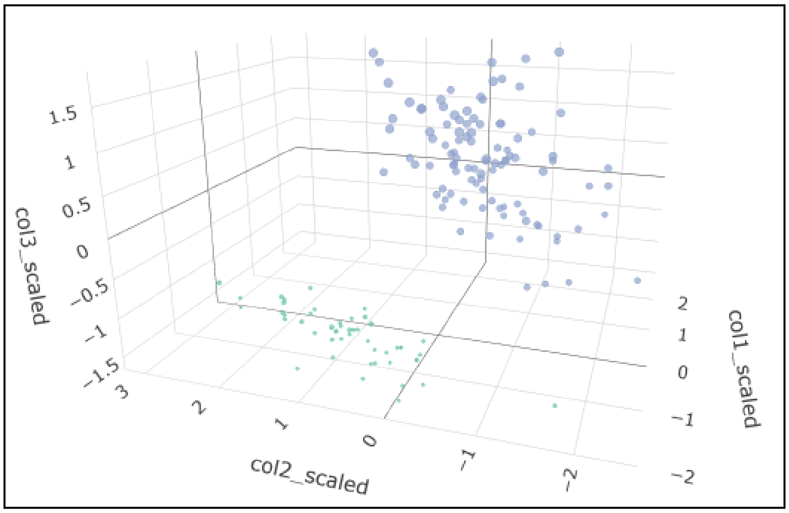
Using a threshold of 0.26, we saw 2 clusters. The first cluster captured all setosa and the second cluster captured all versicolor and virginica. A slightly lower threshold created a 1 data point cluster of setosa, while the other two clusters stayed the same. Lowering the threshold further continued to add clusters of size 1, 2 or 3 data points.



Threshold = 0.10 (42 clusters)



Threshold = 0.20 (4 clusters)



Threshold = 0.26 (2 clusters)

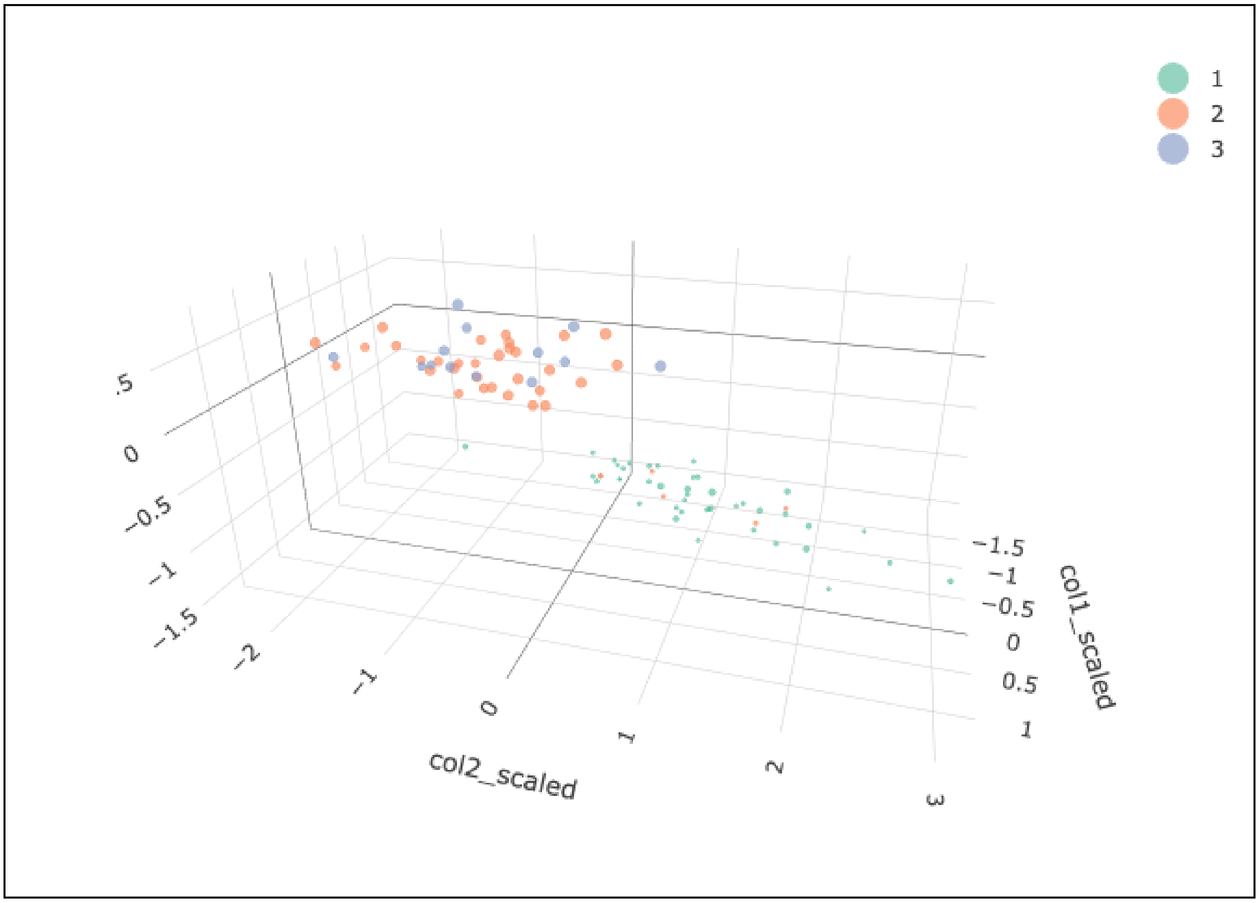
We decided that the best case here is 2 clusters.

- Threshold: 0.26

Cluster	Center	Avg Dist. to Center	SSE for Cluster
0	(0.1961, 0.5908, 0.0786, 0.0600)	0.1650	1.8450
1	(0.5450, 0.3633, 0.6620, 0.6567)	0.2898	10.2987

DBSCAN

We initially tested with an epsilon of 0.1 and min_points of 2, which gave us 10 clusters and lots of noise points. For this data set in particular, given that we have the ground truth, we wanted to ideally end up with no noise points and all points correctly classified. We increased epsilon to 0.12 and min_points to 6, with the intention of merging clusters and making our algorithm more strict in defining core points. This gave us 3 clusters as we wanted, but still with some noise points. Looking over our clusters, we see that out of the points that are classified, only 1 point is misclassified as versicolor instead of virginica. We were not able to simultaneously maintain 3 clusters while decreasing the amount of noise points, so we settled on these parameters as best.



- Epsilon: 0.12
- Min points: 6

Cluster	Center	Avg Dist. to Center	SSE for Cluster
1	(0.1858, 0.5731, 0.0802, 0.0602)	0.1398	1.0565
2	(0.4776, 0.3449, 0.5598, 0.5081)	0.1672	1.1899
3	(0.6774, 0.4647, 0.7692, 0.8878)	0.0972	0.1312

Mammal Milk

k-Means

- k = 4
- Stoppage threshold = 0.4

Cluster	Center	Avg Dist. to Center	SSE for Cluster
0	(0.8819, 0.3134, 0.0938, 0.7768, 0.253)	0.2634	1.1793
1	(0.1568, 0.8433, 0.7732, 0.1208, 0.2924)	0.2446	0.1251
2	(0.5552, 0.8359, 0.3424, 0.3855, 0.6545)	0.2652	0.4853
3	(0.9066, 0.2906, 0.1183, 0.5145, 0.1705)	0.3802	0.4341

Agglomerative Hierarchical Clustering

A threshold of 0.60 and greater resulted in 1 cluster, while a threshold of 0.45 resulted in 3 clusters with sizes of 1, 2, and 22. Decreasing the threshold further resulted in more clusters of size 1, so we decided that this is the best we will achieve.

- Threshold: 0.45

Cluster	Center	Avg Dist. to Center	SSE for Cluster
0	(0.8034, 0.4254, 0.1615, 0.6621, 0.3068)	0.4083	4.5974
1	(0.5802, 1.0, 0.2951, 0.2754, 1.0)	0	0
2	(0.0165, 0.8162, 0.9134, 0.0652, 0.2682)	0.1371	0.0376

DBSCAN

We first tested with an epsilon of 0.2 and min_points of 2. This gave us 2 clusters with a lot of remaining noise points. We decided that for this data set, we would rather have broad groups to get a general idea than a few dense groups with lots of remaining noise points. So, we increased our epsilon to 0.3 to reduce noise. This gave us 1 cluster with 17 data points and another with only 3 data points. Although this is not ideal, we decided this is the best we could do. We were not able to simultaneously increase the number of groups and reduce noise.

- Epsilon: 0.12
- Min points: 6

Cluster	Center	Avg Dist. to Center	SSE for Cluster
1	(0.8803, 0.3348, 0.0941, 0.7144, 0.2759)	0.3154	2.1157
2	(0.4454, 0.8661, 0.4732, 0.3237, 0.5076)	0.1306	0.0573

AccidentsSet03

k-Means

We arbitrarily chose a k-value of 5 and similarly tuned our stoppage threshold.

- k = 5
- Stoppage threshold = 0.6

Cluster	Center	Avg Dist. to Center	SSE for Cluster
0	(0.0741, 0.5, 0.0, 0.0, 0.2222)	0.2206	0.5185
1	(0.0, 0.0, 1.0, 0.75, 0.5)	0.559	0.625
2	(0.3056, 0.0, 1.0, 0.0417, 0.1944)	0.3074	1.4699
3	(0.0, 0.5238, 1.0, 0.0, 0.2381)	0.216	1.381
4	(0.0741, 0.0, 0.0, 0.0556, 0.2407)	0.2335	1.3148

Agglomerative Hierarchical Clustering

Large thresholds of 0.5+ resulted in 2 to 5 clusters, indicating that there are multiple outlier data points. However, decreasing to a threshold of 0.30 yields 17 clusters, most with only 1 data point. At a threshold of 0.49, we have 9 clusters. At a threshold of 0.50, we have 5 clusters. Since the additional clusters at 0.49 only had 1 data point each, we decided to go with 0.50.

- Threshold: 0.50

Cluster	Center	Avg Dist. to Center	SSE for Cluster
0	(0.0741, 0.1667, 0.0, 0.0370, 0.2346)	0.3324	3.3539
1	(0.0824, 0.3548, 1.0, 0.0, 0.2258)	0.3449	4.3301
2	(0.0556, 0.0 ,1.0 ,0.5 ,0.0)	0.0556	0.0062
3	(1.0, 0.0, 1.0, 0.0, 0.3333)	0	0
4	(0.0, 0.0, 1.0, 1.0, 1.0)	0	0

DBSCAN

Starting with an epsilon of 0.2 and min_points of 2, we found 7 clusters but had a good amount of noise points remaining. Using the same logic of wanting fewer noise points that we used in tuning for the Mammal Milk data set, we increased both epsilon and min_points to 0.4 and 3, respectively. This gave us 4 clusters with only a few noise points.

- Epsilon: 0.4
- Min points: 3

Cluster	Center	Avg Dist. to Center	SSE for Cluster
1	(0.0833, 0.0, 0.0, 0.0, 0.2083)	0.1928	0.6512
2	(0.2556, 0.0, 1.0, 0.0, 0.2)	0.236	0.6383
3	(0.0741, 0.5, 0.0, 0.0, 0.2222)	0.2206	0.5185
4	(0.0, 0.5, 1.0, 0.0, 0.2167)	0.195	0.95

Analysis

For ease of hyperparameter tuning and comparison of cluster metrics, we used Euclidean distance and min-max standardization for all our models. For all our k-Means models, we chose our initial cluster centroids through k-Means++. For all our agglomerative hierarchical clustering models, we used single-linkage methods.

The following is which method we determined worked best for each data set:

- 4clusters: k-Means
    - We ruled out DBSCAN first because it incorrectly classified too many points as noise points. K-Means was selected over hierarchical since when given k = 3, we achieved a similar result as hierarchical. Therefore, k-Means allows for more flexibility when using this data set.
- Planets: Agglomerative Hierarchical Clustering
    - We first ruled out DBSCAN because it classified too many points as noise points. K-Means failed to sufficiently separate the clusters, although each cluster has a low SSE and average distance to the center. Thus, hierarchical classifies this data best. We believe that hierarchical achieves what DBSCAN achieved without considering noise points.
- Iris: k-Means
    - Similar to Planets, DBSCAN was ruled out because it classified roughly 40% of the points as noise points. Although hierarchical perfectly classified all setosa points, it classified the versicolor and virginica together into a single cluster. Since k-Means instead classified a majority of setosa points into their own cluster and attempted to split up versicolor and virginica, we decided to select k-Means. Additionally, we believe that k-Means would far outperform hierarchical when using random initial centroid selection enough times, or selecting the initial centroids manually, over using k-Means++.
- Mammal Milk: k-Means
    - Hierarchical was ruled out since its results contained a cluster of one point; after examining the point, it is most likely an outlier since it contains the max values in 2 of the 5 features. Meanwhile, DBSCAN was ruled out for resulting in only 2 clusters, one cluster containing a majority of the points and with 20% of the points being labeled as outliers. Thus, k-Means was selected since it provides more insight into the data by resulting in more clusters, with smaller SSE values on average than either hierarchical or DBSCAN could achieve.
- AccidentsSet03: DBSCAN
    - Unlike with the other data sets, this data set worked well with DBSCAN. This leads us to believe that this is a messy data set, considering that DBSCAN

efficiently handles outliers. Only about 11% of the data was classified as noise points, and the 4 clusters that were created had low SSE's. Both hierarchical and k-Means left us with larger or inconsistent SSE's.

Appendix

4clusters

k-Means

Output for python3 kmeans.py data/4clusters.csv 4 1 1 1 0.2

```
Initial Centroid: K-means++
Distance Metric: Euclidean Distance
Standardization: Min-Max
Stoppage Threshold: 0.2
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Cluster 0:
Center: 0.1904761904761905,0.8452380952380951,
Max Dist. to Center: 0.2993452024507171
Min Dist. to Center: 0.019305724421052523
Avg Dist. to Center: 0.1332029124611435
SSE for Cluster: 0.19077193665105746

8 Points:
10,42,
8,41,
13,40,
…
12,38,
19,38,
25,38,
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Cluster 1:
Center: 0.7163461538461538 46154,0.2048611111111111,
37,10,
40,9,
42,9,
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Cluster 2:
Center: 0.8769230769230768,0.8694444444444445,
Max Dist. to Center: 0.13215721278769066
Min Dist. to Center: 0.028264150883128964
Avg Dist. to Center: 0.08902598518500386
SSE for Cluster: 0.08119808788078021

9 Points:
41,45,
39,44,
42,43,
…
45,40,
38,39,
42,39,
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Cluster 3:
Center: 0.14957264957264957,0.6759259259259259,
Max Dist. to Center: 0.1809579432515583
Min Dist. to Center: 0.019005219464567402
Avg Dist. to Center: 0.08251397596950677
SSE for Cluster: 0.04176609767777858

4 Points:
6,37,
13,35,
9,34,
12,34,
```

Agglomerative Hierarchical Clustering

Output for python3 hclustering.py data/4clusters.csv 0.3 1 1

```
Distance Metric: Euclidean Distance
Standardization: Min-Max
Threshold: 0.3
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Cluster 0:
```

```
Center: 0.9002849002849005,0.9104938271604939,
Max Dist. to Center: 0.11127332903982966
Min Dist. to Center: 0.021791975012685087
Avg Dist. to Center: 0.07761594060660193
SSE for Cluster: 0.06112064837136061

9 Points:
41,45,
39,44,
42,43,
…
42,39,
45,40,
38,39,
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Cluster 1:
Center: 0.1517094017094017,0.8009259259259259,
Max Dist. to Center: 0.3355020293765956
Min Dist. to Center: 0.005098939125098443
Avg Dist. to Center: 0.1272052372541887
SSE for Cluster: 0.26497400589280934

12 Points:
10,42,
8,41,
7,39,
…
9,34,
19,38,
25,38,
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Cluster 2:
Center: 0.6965811965811965,0.22993827160493824,
Max Dist. to Center: 0.3227567549556373
Min Dist. to Center: 0.05906684941050807
Avg Dist. to Center: 0.20328539731696782
SSE for Cluster: 0.8358548733370671

18 Points:
32,27,
39,24,
37,23,
…
26,16,
26,25,
22,22,
```

DBSCAN

```
Output for python3 dbscan.py data/4clusters.csv 0.1 2 1 1
Distance Metric: Euclidean Distance
Standardization: Min-Max
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Cluster 1:
Center: 0.9002849002849003,0.9104938271604939,
Max Dist. to Center: 0.11127332903982987
Min Dist. to Center: 0.02179197501268506
Avg Dist. to Center: 0.07761594060660192
SSE for Cluster: 0.061120648371360616

9 Points:
41,45,
39,44,
42,43,
…
45,40,
38,39,
42,39,
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Cluster 2:
Center: 0.09999999999999999,0.8,
Max Dist. to Center: 0.11849634421645952
Min Dist. to Center: 0.02373622919145095
Avg Dist. to Center: 0.09178732329118308
SSE for Cluster: 0.09265651252830741

10 Points:
10,42,
8,41,
13,40,
…
```

```
13,35,
9,34,
12,34,
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Cluster 3:
Center: 0.7846153846153847,0.36666666666666664,
Max Dist. to Center: 0.07929049280033958
Min Dist. to Center: 0.024474907800472366
Avg Dist. to Center: 0.059326035807227416
SSE for Cluster: 0.019641683103221556

5 Points:
39,24,
34,23,
37,23,
38,21,
35,20,
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Cluster 4:
Center: 0.7655677655677656,0.05158730158730159,
Max Dist. to Center: 0.17586895076677767
Min Dist. to Center: 0.037757377523971174
Avg Dist. to Center: 0.10918360430496417
SSE for Cluster: 0.1016705279525792

7 Points:
31,13,
38,13,
29,11,
34,11,
37,10,
40,9,
42,9,
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
8 Outliers (20.51% of data):
19,38,
25,38,
32,27,
26,25,
22,22,
31,18,
26,16,
26,16,
```

Planets

k-Means

Output for python3 kmeans.py data/planets.csv 4 1 1 1 0.4

```
Initial Centroid: K-means++
Distance Metric: Euclidean Distance
Standardization: Min-Max
Stoppage Threshold: 0.4
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Cluster 0:
Center: 0.9957698289269052,1.0,0.5750000000000001,
Max Dist. to Center: 0.04588210581640994
Min Dist. to Center: 0.0
Avg Dist. to Center: 0.02294105290820497
SSE for Cluster: 0.002105167634148239

2 Points:
1940YL,1953NH,
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Cluster 1:
Center: 0.18236883031840878,0.06973942813668929,1.0,
Max Dist. to Center: 0.5533665190668355
Min Dist. to Center: 0.0
Avg Dist. to Center: 0.28501791273652294
SSE for Cluster: 0.7898059093423289

7 Points:
1929EC,1948RO,1951AM,1938DL,1931DQ,1952DA,1948RB,
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Cluster 2:
Center: 0.42530899566178265,0.5705953211183669,0.0,
Max Dist. to Center: 0.4736163485248907
Min Dist. to Center: 0.0
Avg Dist. to Center: 0.367485474946094
SSE for Cluster: 1.110603173173588
```

```
7 Points:
1935RF,1941FD,1955QT,1930SY,1949HM,1951AX,1948RH,
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Cluster 3:
Center: 0.0,0.7290940214290242,0.6267605633802815,
Max Dist. to Center: 0.4397591780361217
Min Dist. to Center: 0.0
Avg Dist. to Center: 0.28147402724032894
SSE for Cluster: 0.35714020028567184

3 Points:
1924TZ,1936AB,1948TG,
```

Agglomerative Hierarchical Clustering

Output for python3 hclustering.py data/planets.csv 0.35 1 1

```
Distance Metric: Euclidean Distance
Standardization: Min-Max
Threshold: 0.35
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Cluster 0:
Center: 0.282554286182018,0.21640053980128796,0.2510160965794769,
Max Dist. to Center: 0.31310211025381035
Min Dist. to Center: 0.04312193326472973
Avg Dist. to Center: 0.15045190385442958
SSE for Cluster: 0.20562039719175348

7 Points:
1935RF,1941FD,1955QT,1930SY,1949HM,1951AX,1938DL,
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Cluster 1:
Center: 0.20618318736187277,0.18048247004374562,0.7746038732394367,
Max Dist. to Center: 0.34444138814140596
Min Dist. to Center: 0.13578093973293837
Avg Dist. to Center: 0.205757773601162
SSE for Cluster: 0.375094083809689

8 Points:
1929EC,1951AM,1924TZ,1931DQ,1936AB,1952DA,1948RB,
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Cluster 2:
```

```
Center: 0.42530899566178265,0.5705953211183669,0.0,
Max Dist. to Center: 0.0
Min Dist. to Center: 0.0
Avg Dist. to Center: 0.0
SSE for Cluster: 0.0

1 Points:
1948RH,
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Cluster 3:
Center: 0.0,0.7290940214290242,0.6267605633802815,
Max Dist. to Center: 0.0
Min Dist. to Center: 0.0
Avg Dist. to Center: 0.0
SSE for Cluster: 0.0

1 Points:
1948TG,
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Cluster 4:
Center: 0.9978849144634526,0.9776199835161352,0.5704225352112675,
Max Dist. to Center: 0.022941052908205015
Min Dist. to Center: 0.022941052908204924
Avg Dist. to Center: 0.02294105290820497
SSE for Cluster: 0.0010525838170741195

2 Points:
1940YL,1953NH,
```

DBSCAN

```
Output for python3 dbscan.py data/planets.csv 0.2 2 1 1
Distance Metric: Euclidean Distance
Standardization: Min-Max
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Cluster 1:
Center: 0.27431393413549426,0.25246729643483595,0.21421361502347422,
Max Dist. to Center: 0.21349359035517287
Min Dist. to Center: 0.04953181470368161
Avg Dist. to Center: 0.10988336485924359
SSE for Cluster: 0.09124864383879944
```

```
6 Points:
1935RF,1941FD,1955QT,1930SY,1949HM,1951AX,
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Cluster 2:
Center: 0.2373130337507844,0.09366216530357785,0.9143661971830986,
Max Dist. to Center: 0.10451942895585611
Min Dist. to Center: 0.052259714477928015
Avg Dist. to Center: 0.06967961930390405
SSE for Cluster: 0.01638646654388737

3 Points:
1929EC,1948RO,1951AM,
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Cluster 3:
Center: 0.1030891380862732,0.2780384200849553,0.671531690140845,
Max Dist. to Center: 0.12376513432920966
Min Dist. to Center: 0.06883059113168678
Avg Dist. to Center: 0.08795892122119711
SSE for Cluster: 0.032869042611735705

4 Points:
1924TZ,1931DQ,1936AB,1952DA,
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
6 Outliers (31.58% of data):
1940YL,1953NH,1938DL,1948RB,1948RH,1948TG,
```

Iris

k-Means

Output for python3 kmeans.py data/iris.csv 3 1 1 1 0.4

```
Initial Centroid: K-means++
Distance Metric: Euclidean Distance
Standardization: Min-Max
Stoppage Threshold: 0.4
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Cluster 0:
Center:
0.22461685823754796,0.41163793103448293,0.23027469316189358,0.201867816091
95406,
Max Dist. to Center: 0.48760169259751773
Min Dist. to Center: 0.1749677793539595
Avg Dist. to Center: 0.31113070261773057
SSE for Cluster: 3.387649180455373

33 Points:
4.9,3.0,1.4,0.2,Iris-setosa,
4.7,3.2,1.3,0.2,Iris-setosa,
4.6,3.1,1.5,0.2,Iris-setosa,
…
5.5,2.4,3.7,1.0,Iris-versicolor,
5.0,2.3,3.3,1.0,Iris-versicolor,
5.1,2.5,3.0,1.1,Iris-versicolor,
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Cluster 1:
Center:
0.6036324786324787,0.39797008547008533,0.7125162972620597,0.71688034188034
19,
Max Dist. to Center: 0.5739209326652599
Min Dist. to Center: 0.05475525918369429
Avg Dist. to Center: 0.2735477092163871
SSE for Cluster: 7.8362332624210795

89 Points:
7.0,3.2,4.7,1.4,Iris-versicolor,
6.4,3.2,4.5,1.5,Iris-versicolor,
6.9,3.1,4.9,1.5,Iris-versicolor,
…
6.5,3.0,5.2,2.0,Iris-virginica,
6.2,3.4,5.4,2.3,Iris-virginica,
5.9,3.0,5.1,1.8,Iris-virginica,
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Cluster 2:
Center:
0.2996031746031747,0.7827380952380952,0.08595641646489104,0.07440476190476
19,
Max Dist. to Center: 0.2948633603308806
Min Dist. to Center: 0.06118423730222713
Avg Dist. to Center: 0.16487214212806509
SSE for Cluster: 0.8767954927403849
```

```
28 Points:
5.1,3.5,1.4,0.2,Iris-setosa,
5.0,3.6,1.4,0.2,Iris-setosa,
5.4,3.9,1.7,0.4,Iris-setosa,
…
5.1,3.8,1.9,0.4,Iris-setosa,
5.1,3.8,1.6,0.2,Iris-setosa,
5.3,3.7,1.5,0.2,Iris-setosa,
```

Agglomerative Hierarchical Clustering

Output for python3 hclustering.py data/iris.csv 0.26 1 1

```
Distance Metric: Euclidean Distance
Standardization: Min-Max
Threshold: 0.26
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Cluster 0:
Center:
0.19611111111111111,0.5908333333333333,0.07864406779661015,0.06000000000000
001,
Max Dist. to Center: 0.48792791949563047
Min Dist. to Center: 0.02079349825613116
Avg Dist. to Center: 0.16501838332955018
SSE for Cluster: 1.8449596453498882

50 Points:
5.1,3.5,1.4,0.2,Iris-setosa,
5.2,3.5,1.5,0.2,Iris-setosa,
5.1,3.5,1.4,0.3,Iris-setosa,
…
5.5,4.2,1.4,0.2,Iris-setosa,
5.7,4.4,1.5,0.4,Iris-setosa,
4.5,2.3,1.3,0.3,Iris-setosa,
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Cluster 1:
Center:
0.5449999999999999,0.3633333333333333,0.6620338983050847,0.656666666666666
7,
Max Dist. to Center: 0.6702191020020616
Min Dist. to Center: 0.0647211726780973
```

```
Avg Dist. to Center: 0.2898391155312589
SSE for Cluster: 10.298728636229832

100 Points:
7.0,3.2,4.7,1.4,Iris-versicolor,
6.9,3.1,4.9,1.5,Iris-versicolor,
6.7,3.1,4.7,1.5,Iris-versicolor,
…
4.9,2.5,4.5,1.7,Iris-virginica,
7.7,3.8,6.7,2.2,Iris-virginica,
7.9,3.8,6.4,2.0,Iris-virginica,
```

DBSCAN

Output for python3 dbscan.py data/iris.csv 0.12 6 1 1

```
Distance Metric: Euclidean Distance
Standardization: Min-Max
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Cluster 1:
Center:
0.1858024691358025,0.5731481481481482,0.08022598870056495,0.06018518518518519,
Max Dist. to Center: 0.27300633134723207
Min Dist. to Center: 0.023276288787640143
Avg Dist. to Center: 0.13979211991336504
SSE for Cluster: 1.056540135672821

45 Points:
5.1,3.5,1.4,0.2,Iris-setosa,
4.9,3.0,1.4,0.2,Iris-setosa,
4.7,3.2,1.3,0.2,Iris-setosa,
…
4.6,3.2,1.4,0.2,Iris-setosa,
5.3,3.7,1.5,0.2,Iris-setosa,
5.0,3.3,1.4,0.2,Iris-setosa,
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Cluster 2:
Center:
0.4776234567901235,0.3449074074074073,0.5597928436911487,0.5081018518518517,
Max Dist. to Center: 0.32233841495850696
```

```
Min Dist. to Center: 0.05773970428666684
Avg Dist. to Center: 0.16724165282045994
SSE for Cluster: 1.1898523664891205

36 Points:
7.0,3.2,4.7,1.4,Iris-versicolor,
6.4,3.2,4.5,1.5,Iris-versicolor,
6.9,3.1,4.9,1.5,Iris-versicolor,
…
6.2,2.9,4.3,1.3,Iris-versicolor,
5.7,2.8,4.1,1.3,Iris-versicolor,
6.3,2.8,5.1,1.5,Iris-virginica,
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Cluster 3:
Center:
0.6773504273504274,0.4647435897435897,0.7692307692307693,0.887820512820512
8,
Max Dist. to Center: 0.1391587992480278
Min Dist. to Center: 0.06955734748780618
Avg Dist. to Center: 0.09717151261330216
SSE for Cluster: 0.1311913141697446

13 Points:
7.1,3.0,5.9,2.1,Iris-virginica,
6.5,3.0,5.8,2.2,Iris-virginica,
6.8,3.0,5.5,2.1,Iris-virginica,
…
6.7,3.3,5.7,2.5,Iris-virginica,
6.7,3.0,5.2,2.3,Iris-virginica,
6.5,3.0,5.2,2.0,Iris-virginica,
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
56 Outliers (37.33% of data):
5.8,4.0,1.2,0.2,Iris-setosa,
5.7,4.4,1.5,0.4,Iris-setosa,
5.2,4.1,1.5,0.1,Iris-setosa,
…
6.3,2.5,5.0,1.9,Iris-virginica,
6.2,3.4,5.4,2.3,Iris-virginica,
5.9,3.0,5.1,1.8,Iris-virginica,
```

Mammal Milk

k-Means

Output for python3 kmeans.py data/mammal_milk.csv 4 1 1 1 0.4

```
Initial Centroid: K-means++
Distance Metric: Euclidean Distance
Standardization: Min-Max
Stoppage Threshold: 0.4
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Cluster 0:
Center:
0.881904761904762,0.31339031339031337,0.0938211382113821,0.7768115942028987,0.2530303030303031,
Max Dist. to Center: 0.5259875692512498
Min Dist. to Center: 0.0953847227073042
Avg Dist. to Center: 0.2634220925493868
SSE for Cluster: 1.1792628441495474

14 Points:
Horse,Orangutan,Monkey,Donkey,Camel,Bison,Buffalo,Cat,Fox,Llama,Mule,Zebra,Sheep,Elephant.
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Cluster 1:
Center:
0.15677655677655675,0.8433048433048432,0.773170731707317,0.12077294685990338,0.2924242424242425,
Max Dist. to Center: 0.29666647380947364
Min Dist. to Center: 0.19257328365940457
Avg Dist. to Center: 0.2446198787344391
SSE for Cluster: 0.1250954662619126

2 Points:
Seal,Dolphin,
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Cluster 2:
Center:
0.5551648351648351,0.8358974358974359,0.3424390243902439,0.38550724637681155,0.6545454545454545,
Max Dist. to Center: 0.40274734373631843
Min Dist. to Center: 0.1606620178354097
```

```
Avg Dist. to Center: 0.265201472301722
SSE for Cluster: 0.4852687371153969

6 Points:
Dog,Rabbit,Rat,Deer,Reindeer,Whale,
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Cluster 3:
Center:
0.9065934065934066,0.2905982905982906,0.11829268292682926,0.51449275362318
85,0.17045454545454547,
Max Dist. to Center: 0.3964064225209095
Min Dist. to Center: 0.37213770255791545
Avg Dist. to Center: 0.3802272758789134
SSE for Cluster: 0.434110991145993

3 Points:
Hippo,Guinea Pig,Pig,
```

Agglomerative Hierarchical Clustering

```
Output for python3 hclustering.py data/mammal_milk.csv 0.45 1 1
Distance Metric: Euclidean Distance
Standardization: Min-Max
Threshold: 0.45
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Cluster 0:
Center:
0.8033966033966035,0.4254079254079253,0.161529933481153,0.6620553359683794,
0.3068181818181819,
Max Dist. to Center: 0.807875667517807
Min Dist. to Center: 0.03712080851960147
Avg Dist. to Center: 0.40830724594925416
SSE for Cluster: 4.597388433650213

22 Points:
Horse,Orangutan,Monkey,DonkeyMule,Camel,ZebraBison,Llama,Buffalo,Sheep,Fox,
Guinea Pig,Pig,Hippo,Dog,Rat,Deer,Reindeer,Whale,Cat,Elephant,
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Cluster 1:
Center: 0.58021978021978,1.0,0.2951219512195122,0.27536231884057966,1.0,
Max Dist. to Center: 0.0
```

```
Min Dist. to Center: 0.0
Avg Dist. to Center: 0.0
SSE for Cluster: 0.0

1 Points:
Rabbit,
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Cluster 2:
Center:
0.01648351648351648,0.8162393162393162,0.9134146341463414,0.06521739130434
782,0.2681818181818182,
Max Dist. to Center: 0.13707875897962535
Min Dist. to Center: 0.1370787589796252
Avg Dist. to Center: 0.1370787589796253
SSE for Cluster: 0.0375811723267884

2 Points:
Seal,Dolphin,
```

DBSCAN

```
Output for python3 dbscan.py data/mammal_milk.csv 0.3 3 1 1
Distance Metric: Euclidean Distance
Standardization: Min-Max
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Cluster 1:
Center:
0.8802844214608919,0.33484162895927605,0.09411764705882353,0.7144075021312
873,0.27593582887700535,
Max Dist. to Center: 0.652449225674806
Min Dist. to Center: 0.11326928635512187
Avg Dist. to Center: 0.3154271761436702
SSE for Cluster: 2.1157163349964887

17 Points:
Horse,Orangutan,Monkey,Donkey,Hippo,Camel,Bison,Buffalo,
Guinea Pig,Fox,Llama,Mule,Pig,Zebra,Sheep,Dog,Rat,
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Cluster 2:
```

```
Center:
0.44542124542124534,0.866096866096866,0.47317073170731705,0.32367149758454106,0.5075757575757577,
Max Dist. to Center: 0.19398585942323926
Min Dist. to Center: 0.09228796447036185
Avg Dist. to Center: 0.13059286045740084
SSE for Cluster: 0.057278835892874025

3 Points:
Deer,Reindeer,Whale.
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
5 Outliers (20.00% of data):
Cat,Elephant,Rabbit,Seal,Dolphin,
```

AccidentsSet03

k-Means

```
Output for python3 kmeans.py data/AccidentsSet03.csv 5 1 1 1 0.6
Initial Centroid: K-means++
Distance Metric: Euclidean Distance
Standardization: Min-Max
Stoppage Threshold: 0.6
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Cluster 0:
Center: 0.07407407407407408,0.5,0.0,0.0,0.2222222222222222,
Max Dist. to Center: 0.4459849843997146
Min Dist. to Center: 0.11712139482105108
Avg Dist. to Center: 0.2205750753161601
SSE for Cluster: 0.5185185185185184

9 Points:
3.00, 1.00, 2.00, 1.00, 0.00,
3.00, 1.00, 2.00, 1.00, 1.00,
1.00, 1.00, 2.00, 1.00, 1.00,
…
1.00, 1.00, 2.00, 1.00, 0.00,
1.00, 1.00, 2.00, 1.00, 0.00,
1.00, 1.00, 2.00, 1.00, 0.00,
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Cluster 1:
```

```
Center: 0.0,0.0,1.0,0.75,0.5,
Max Dist. to Center: 0.5590169943749475
Min Dist. to Center: 0.5590169943749475
Avg Dist. to Center: 0.5590169943749475
SSE for Cluster: 0.6250000000000001

2 Points:
1.00, 0.00, 4.00, 2.00, 0.00,
1.00, 0.00, 4.00, 3.00, 3.00,
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Cluster 2:
Center: 0.30555555555555564,0.0,1.0,0.041666666666666664,0.19444444444444442,
Max Dist. to Center: 0.7094218216178962
Min Dist. to Center: 0.16724436914989305
Avg Dist. to Center: 0.30741064725105854
SSE for Cluster: 1.4699074074074072

12 Points:
5.00, 0.00, 4.00, 1.00, 1.00,
5.00, 0.00, 4.00, 1.00, 0.00,
4.00, 0.00, 4.00, 1.00, 0.00,
…
1.00, 0.00, 4.00, 1.00, 1.00,
3.00, 0.00, 4.00, 1.00, 0.00,
3.00, 0.00, 4.00, 1.00, 2.00,
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Cluster 3:
Center: 0.0,0.5238095238095238,1.0,0.0,0.23809523809523805,
Max Dist. to Center: 0.6406487641463672
Min Dist. to Center: 0.09816918156232528
Avg Dist. to Center: 0.21600860778847428
SSE for Cluster: 1.3809523809523814

21 Points:
1.00, 1.00, 4.00, 1.00, 0.00,
1.00, 1.00, 4.00, 1.00, 1.00,
1.00, 1.00, 4.00, 1.00, 2.00,
…
1.00, 1.00, 4.00, 1.00, 1.00,
1.00, 1.00, 4.00, 1.00, 0.00,
1.00, 1.00, 4.00, 1.00, 1.00,
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Cluster 4:
Center: 0.07407407407407407,0.0,0.0,0.05555555555555555,0.24074074074074073,
Max Dist. to Center: 0.620024779362341
Min Dist. to Center: 0.13094570021973104
Avg Dist. to Center: 0.2335435183406332
SSE for Cluster: 1.3148148148148144

18 Points:
3.00, 0.00, 2.00, 1.00, 0.00,
2.00, 0.00, 2.00, 1.00, 0.00,
3.00, 0.00, 2.00, 1.00, 1.00,
…
1.00, 0.00, 2.00, 2.00, 1.00,
1.00, 0.00, 2.00, 1.00, 1.00,
1.00, 0.00, 2.00, 1.00, 1.00,
```

Agglomerative Hierarchical Clustering

Output for python3 hclustering.py data/AccidentsSet03.csv 0.50 1 1

```
Distance Metric: Euclidean Distance
Standardization: Min-Max
Threshold: 0.5
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Cluster 0:
Center:
0.07407407407407407,0.16666666666666666,0.0,0.037037037037037035,0.2345679012345679,
Max Dist. to Center: 0.6590210879714692
Min Dist. to Center: 0.21069195266513593
Avg Dist. to Center: 0.3324031356518222
SSE for Cluster: 3.353909465020576

27 Points:
3.00, 0.00, 2.00, 1.00, 0.00,
2.00, 0.00, 2.00, 1.00, 0.00,
2.00, 0.00, 2.00, 1.00, 0.00,
…
2.00, 1.00, 2.00, 1.00, 2.00,
1.00, 0.00, 2.00, 2.00, 2.00,
1.00, 0.00, 2.00, 2.00, 1.00,
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Cluster 1:
Center: 0.08243727598566307,0.3548387096774194,1.0,0.0,0.22580645161290322,
Max Dist. to Center: 0.7857396033164448
Min Dist. to Center: 0.19856921960004847
Avg Dist. to Center: 0.3448632545320441
SSE for Cluster: 4.330147351652728

31 Points:
5.00, 0.00, 4.00, 1.00, 1.00,
3.00, 0.00, 4.00, 1.00, 1.00,
3.00, 0.00, 4.00, 1.00, 1.00,
…
1.00, 1.00, 4.00, 1.00, 2.00,
1.00, 1.00, 4.00, 1.00, 2.00,
1.00, 2.00, 4.00, 1.00, 2.00,
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Cluster 2:
Center: 0.05555555555555555,0.0,1.0,0.5,0.0,
Max Dist. to Center: 0.05555555555555555
Min Dist. to Center: 0.05555555555555555
Avg Dist. to Center: 0.05555555555555555
SSE for Cluster: 0.006172839506172839

2 Points:
1.00, 0.00, 4.00, 2.00, 0.00,
2.00, 0.00, 4.00, 2.00, 0.00,
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Cluster 3:
Center: 1.0,0.0,1.0,0.0,0.3333333333333333,
Max Dist. to Center: 0.0
Min Dist. to Center: 0.0
Avg Dist. to Center: 0.0
SSE for Cluster: 0.0

1 Points:
10.00, 0.00, 4.00, 1.00, 1.00,
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Cluster 4:
```

```
Center: 0.0,0.0,1.0,1.0,1.0,
Max Dist. to Center: 0.0
Min Dist. to Center: 0.0
Avg Dist. to Center: 0.0
SSE for Cluster: 0.0

1 Points:
1.00, 0.00, 4.00, 3.00, 3.00,
```

DBSCAN

```
Output for python3 dbscan.py data/AccidentsSet03.csv 0.4 3 1 1
Distance Metric: Euclidean Distance
Standardization: Min-Max
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Cluster 1:
Center: 0.08333333333333333,0.0,0.0,0.0,0.20833333333333331,
Max Dist. to Center: 0.3254270698294439
Min Dist. to Center: 0.1502313031443329
Avg Dist. to Center: 0.19280707399271518
SSE for Cluster: 0.6512345679012346

16 Points:
3.00, 0.00, 2.00, 1.00, 0.00,
2.00, 0.00, 2.00, 1.00, 0.00,
3.00, 0.00, 2.00, 1.00, 1.00,
…
1.00, 0.00, 2.00, 1.00, 1.00,
1.00, 0.00, 2.00, 1.00, 1.00,
1.00, 0.00, 2.00, 1.00, 1.00,
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Cluster 2:
Center: 0.2555555555555556,0.0,1.0,0.0,0.2,
Max Dist. to Center: 0.46785562825393995
Min Dist. to Center: 0.13743685418725535
Avg Dist. to Center: 0.23601493708909213
SSE for Cluster: 0.6382716049382715

10 Points:
5.00, 0.00, 4.00, 1.00, 1.00,
5.00, 0.00, 4.00, 1.00, 0.00,
4.00, 0.00, 4.00, 1.00, 0.00,
…
1.00, 0.00, 4.00, 1.00, 1.00,
3.00, 0.00, 4.00, 1.00, 0.00,
3.00, 0.00, 4.00, 1.00, 2.00,
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Cluster 3:
Center: 0.07407407407407408,0.5,0.0,0.0,0.2222222222222222,
Max Dist. to Center: 0.4459849843997146
Min Dist. to Center: 0.11712139482105108
Avg Dist. to Center: 0.2205750753161601
SSE for Cluster: 0.5185185185185184

9 Points:
3.00, 1.00, 2.00, 1.00, 0.00,
3.00, 1.00, 2.00, 1.00, 1.00,
1.00, 1.00, 2.00, 1.00, 1.00,
…
1.00, 1.00, 2.00, 1.00, 0.00,
1.00, 1.00, 2.00, 1.00, 0.00,
1.00, 1.00, 2.00, 1.00, 0.00,
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Cluster 4:
Center: 0.0,0.5,1.0,0.0,0.21666666666666665,
Max Dist. to Center: 0.44999999999999996
Min Dist. to Center: 0.11666666666666667
Avg Dist. to Center: 0.195
SSE for Cluster: 0.95

20 Points:
1.00, 1.00, 4.00, 1.00, 0.00,
1.00, 1.00, 4.00, 1.00, 1.00,
1.00, 1.00, 4.00, 1.00, 2.00,
…
1.00, 1.00, 4.00, 1.00, 1.00,
1.00, 1.00, 4.00, 1.00, 0.00,
1.00, 1.00, 4.00, 1.00, 1.00,
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
7 Outliers (11.29% of data):
10.00, 0.00, 4.00, 1.00, 1.00,
1.00, 0.00, 4.00, 2.00, 0.00,
2.00, 0.00, 4.00, 2.00, 0.00,
1.00, 0.00, 4.00, 3.00, 3.00,
1.00, 0.00, 2.00, 2.00, 2.00,
1.00, 0.00, 2.00, 2.00, 1.00,
1.00, 2.00, 4.00, 1.00, 2.00,
```