Lab 3: Clustering

Vittorio Dinovi - vdinovi@calpoly.edu

Billy Gottenstrater - wgottens@calpoly.edu

Design

1. K-Means

- Distance measures calculated using *Manhattan Distance*.
- Numpy was used for efficient matrix computations
- K cluster centroids are selected at the beginning from the set of points to ensure that there will be no empty clusters
- Clusters are calculated by finding the closest centroid to each point.
- New centroids are calculated using the Numpy's average function on all of the points in each cluster
- The algorithm stops when the difference between consecutive, or every other centroid locations is below a threshold supplied by the user

2. Hierarchical Clustering

- Distance measures calculated using *Manhattan Distance*.
- Numpy was used for efficient distance matrix computations.
- The distance matrix is calculated *once* at the start, and selectively updated during each merge.
- A flat list of tree Nodes is iteratively collapsed until one remains which is used as the root of the dendrogram.
- The dendrogram is written to a file in JSON format.
- All distance measures have been implemented: Single, Complete, Average, Centroid, and Wards.
    - The original data is not read by Single, Complete, and Average
    - Centroid and Average must read the original data in order to create centroids
- Plotted number-of-clusters against avg. variance in dist-to-center among clusters for all thresholds for each link method to determine optimal threshold and link method
    - Looking for plateaus in number-of-clusters following by large spikes
    - Looking for small change in avg. variance in distance-to-center among clusters
    - Compare plots for each link method to look for notable intervals

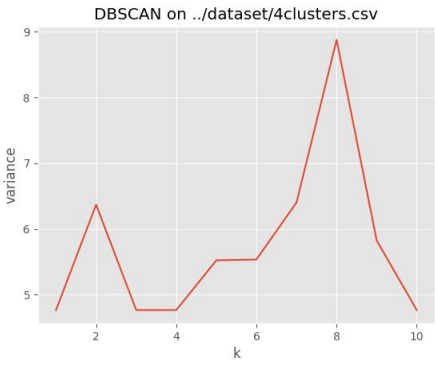
3. DBSCAN

- Distance measures are calculated using *Manhattan Distance*.
- Numpy was used for efficient matrix computations
- The user provides an epsilon value and a numpoints value.
- First, the core points are calculated. A point is a core point if it has at least *numPoints* points in the epsilon neighborhood if it.
- We initialize a cluster value to 1.
- Then, for each core point, we do the following:
    - Check if it has a label already. If it does, go to the next core point
    - If it doesn't have a label, label it with the current cluster, and perform the recursive function dbconnected on it.
    - Dbconnected will find and label all points in the epsilon neighborhood of a core point, and any core points encountered in the process.
    - Increment cluster value.
- Now, we have labels for all points. Using this info and the list of core points we had earlier, we have the following information:
    - **Core points**: the core points we calculated at the beginning
    - **Border points**: points with a cluster label that aren't core points
    - **Outlier points**: points with no cluster label
    - **Clusters**: core and border points with the same labels

Results

4 Clusters

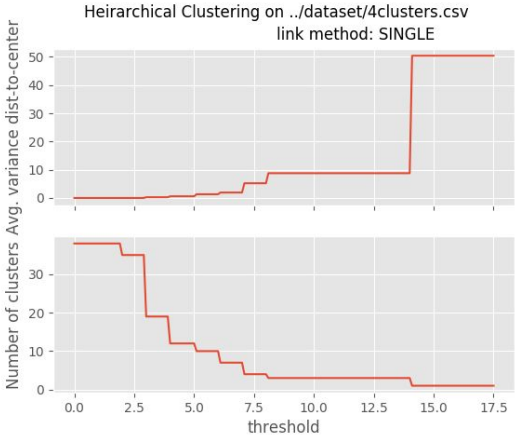
K-Means

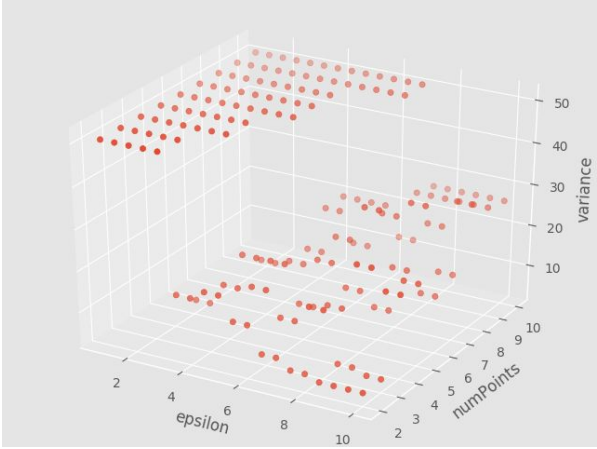


4 clusters

Cluster0
Center: (38.2, 10.4)
Max Dist: 5.2
Min Dist: 1.6
Avg Dist: 3.52
Dist-Variance: 1.7536
Size: 5
Sample: [(42, 9), (40, 9), (38, 13), (37, 10), (34, 11)]

Cluster1
Center: (24.4286, 39.5238)
Max Dist: 23.0476
Min Dist: 2.0952
Avg Dist: 16.8073
Dist-Variance: 23.9625
Size: 21
Sample: [(13, 40), (45, 40), (25, 38), (13, 35), (8, 41)]

	Cluster2 Center: (27.2857, 17.2857) Max Dist: 10.0 Min Dist: 2.5714 Avg Dist: 6.3673 Dist-Variance: 8.2907 Size: 7 Sample: [(29, 11), (26, 16), (31, 18), (26, 25), (26, 16)] Cluster3 Center: (35.8333, 23.0) Max Dist: 7.8333 Min Dist: 1.1667 Avg Dist: 3.8333 Dist-Variance: 4.5556 Size: 6 Sample: [(35, 20), (39, 24), (32, 27), (38, 21), (37, 23)]
Hierarchical  **link-method: SINGLE** **threshold: 10** **3 clusters**	Cluster0 Center: (41.1111, 41.7778) Max Dist: 5.8889 Min Dist: 0.8889 Avg Dist: 3.7037 Dist-Variance: 2.214 Size: 9 Sample: [(39, 44), (41, 45), (45, 40), (38, 42), (38, 39)] Cluster1 Center: (11.9167, 37.8333) Max Dist: 13.25 Min Dist: 0.25 Avg Dist: 5.6389 Dist-Variance: 9.3904 Size: 12 Sample: [(8, 41), (9, 38), (12, 38), (13, 35), (9, 34)] Cluster2 Center: (33.1667, 17.2778) Max Dist: 17.1111 Min Dist: 2.8889 Avg Dist: 9.9815 Dist-Variance: 14.7219 Size: 18 Sample: [(37, 10), (37, 23), (39, 24), (34, 23), (34, 11)]
DBSCAN	Cluster0 Center: (41.1111, 41.7778) Max Dist: 5.8889 Min Dist: 0.8889 Avg Dist: 3.7037 Dist-Variance: 2.214 Size: 9 Sample: [(42, 39), (44, 43), (42, 43), (41, 45), (38, 42)] Cluster1 Center: (9.5, 37.0) Max Dist: 5.5 Min Dist: 1.5 Avg Dist: 4.125



5 clusters found
Best Variance = 4.530975, achieved with e=8.5, numPoints=2

Dist-Variance: 1.7344
Size: 8
Sample: [(13, 35), (7, 39), (12, 38), (12, 34), (6, 37)]

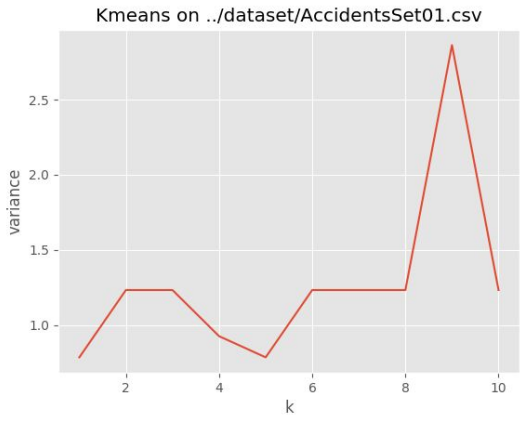
Cluster2
Center: (36.6, 22.2)
Max Dist: 4.2
Min Dist: 1.2
Avg Dist: 3.04
Dist-Variance: 1.1264
Size: 5
Sample: [(39, 24), (37, 23), (35, 20), (34, 23), (38, 21)]

Cluster3
Center: (38.5, 9.5)
Max Dist: 2.0
Min Dist: 2.0
Avg Dist: 2.0
Dist-Variance: 0.0
Size: 2
Sample: [(37, 10), (40, 9)]

Cluster4
Center: (26.9333, 22.6)
Max Dist: 36.3333
Min Dist: 3.3333
Avg Dist: 16.3822
Dist-Variance: 94.1312
Size: 15
Sample: [(13, 40), (26, 25), (19, 38), (31, 18), (10, 42)]

Accidents1

K-Means

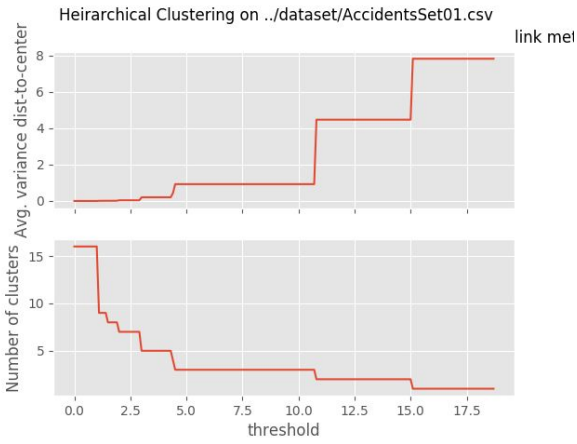


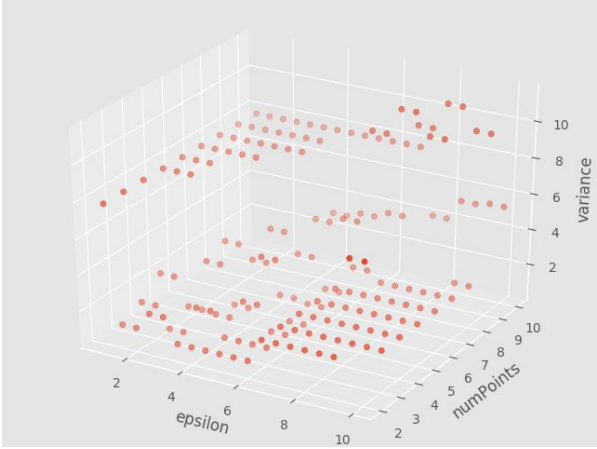
5 Clusters

Cluster0
Center: (5.0, 7.6667, 1.0)
Max Dist: 0.6667
Min Dist: 0.3333
Avg Dist: 0.4444
Dist-Variance: 0.0247
Size: 3
Sample: [(5, 8, 1), (5, 8, 1), (5, 7, 1)]

Cluster1
Center: (2.0, 19.0, 2.0)
Max Dist: 0.0
Min Dist: 0.0
Avg Dist: 0.0
Dist-Variance: 0.0
Size: 1
Sample: [(2, 19, 2)]

Cluster2
Center: (2.0, 3.125, 2.0)
Max Dist: 3.875

	Min Dist: 0.875 Avg Dist: 2.125 Dist-Variance: 1.0312 Size: 8 Sample: [(2, 4, 1), (2, 5, 4), (2, 2, 2), (2, 2, 2), (2, 2, 1)] Cluster3 Center: (5.0, 13.6667, 1.0) Max Dist: 1.6667 Min Dist: 0.3333 Avg Dist: 1.1111 Dist-Variance: 0.321 Size: 3 Sample: [(5, 15, 1), (5, 14, 1), (5, 12, 1)] Cluster4 Center: (5.0, 9.75, 1.0) Max Dist: 1.25 Min Dist: 0.25 Avg Dist: 0.75 Dist-Variance: 0.125 Size: 4 Sample: [(5, 9, 1), (5, 10, 1), (5, 9, 1), (5, 11, 1)]
Hierarchical  **link-method: AVERAGE** **threshold: 5** **3 clusters**	Cluster0 Center: (5.0, 10.3, 1.0) Max Dist: 4.7 Min Dist: 0.3 Avg Dist: 2.16 Dist-Variance: 1.7444 Size: 10 Sample: [(5, 11, 1), (5, 15, 1), (5, 8, 1), (5, 7, 1), (5, 9, 1)] Cluster1 Center: (2.0, 19.0, 2.0) Max Dist: 0.0 Min Dist: 0.0 Avg Dist: 0.0 Dist-Variance: 0.0 Size: 1 Sample: [(2, 19, 2)] Cluster2 Center: (2.0, 3.125, 2.0) Max Dist: 3.875 Min Dist: 0.875 Avg Dist: 2.125 Dist-Variance: 1.0312 Size: 8 Sample: [(2, 2, 1), (2, 2, 2), (2, 1, 1), (2, 5, 4), (2, 4, 2)]
DBSCAN	Cluster0 Center: (5.0, 10.3, 1.0) Max Dist: 4.7 Min Dist: 0.3 Avg Dist: 2.16



3 Clusters found
Best Variance = 0.6839, achieved with e=2.5, numPoints=3

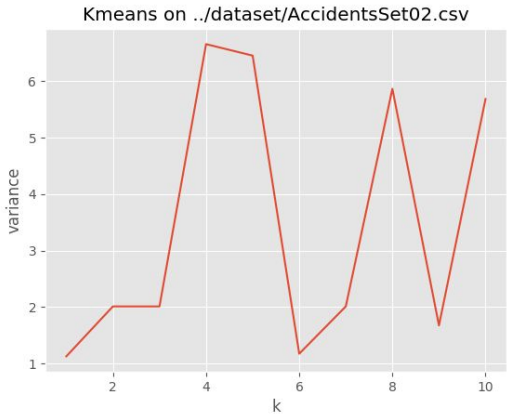
Dist-Variance: 1.7444
Size: 10
Sample: [(5, 8, 1), (5, 9, 1), (5, 9, 1), (5, 11, 1), (5, 12, 1)]

Cluster1
Center: (2.0, 19.0, 2.0)
Max Dist: 0.0
Min Dist: 0.0
Avg Dist: 0.0
Dist-Variance: 0.0
Size: 1
Sample: [(2, 19, 2)]

Cluster2
Center: (2.0, 3.125, 2.0)
Max Dist: 3.875
Min Dist: 0.875
Avg Dist: 2.125
Dist-Variance: 1.0312
Size: 8
Sample: [(2, 5, 3), (2, 2, 2), (2, 5, 4), (2, 2, 2), (2, 1, 1)]

Accidents2

K-Means



6 Clusters

Cluster0
Center: (1.4167, 3.0833, 0.9167, 4.0, 45.0, 1.1667, 0.4167)
Max Dist: 8.6667
Min Dist: 1.1667
Avg Dist: 3.0556
Dist-Variance: 4.1497
Size: 12
Sample: [(1.0, 3.0, 1.0, 4.0, 45.0, 1.0, 0.0), (2.0, 2.0, 0.0, 4.0, 45.0, 1.0, 1.0), (1.0, 2.0, 1.0, 4.0, 45.0, 1.0, 1.0), (5.0, 6.0, 0.0, 4.0, 45.0, 2.0, 0.0), (1.0, 3.0, 1.0, 4.0, 45.0, 1.0, 1.0)]

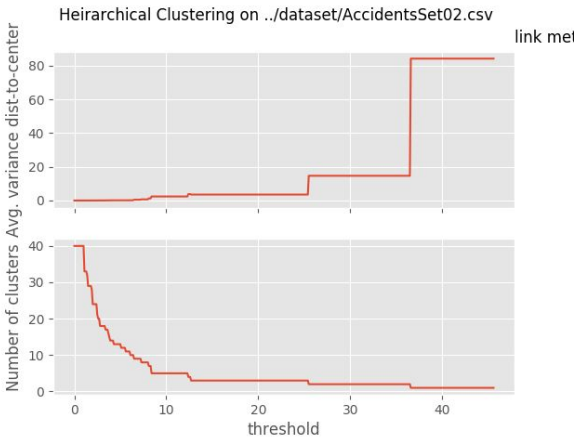
Cluster1
Center: (1.7333, 4.5333, 0.0, 2.0, 70.0, 1.2667, 0.3333)
Max Dist: 23.8
Min Dist: 2.3333
Avg Dist: 6.08
Dist-Variance: 25.004
Size: 15
Sample: [(1.0, 3.0, 0.0, 2.0, 70.0, 1.0, 0.0), (1.0, 6.0, 0.0, 2.0, 70.0, 3.0, 2.0), (1.0, 2.0, 0.0, 2.0, 70.0, 1.0, 0.0), (1.0, 1.0, 0.0, 2.0, 70.0, 1.0, 0.0), (1.0, 2.0, 0.0, 2.0, 70.0, 1.0, 2.0)]

Cluster2
Center: (2.125, 3.5, 0.0, 4.75, 35.0, 1.375, 0.375)
Max Dist: 4.625
Min Dist: 2.375
Avg Dist: 3.25
Dist-Variance: 0.4062
Size: 8
Sample: [(1.0, 4.0, 0.0, 4.0, 35.0, 3.0, 1.0), (2.0, 4.0, 0.0, 4.0, 35.0, 1.0, 1.0), (2.0, 2.0, 0.0, 6.0, 35.0, 1.0, 0.0), (3.0, 4.0, 0.0, 4.0, 35.0, 2.0, 0.0), (1.0, 4.0, 0.0, 4.0, 35.0, 1.0, 0.0)]

Cluster3
Center: (1.5, 6.0, 1.0, 2.8333, 35.0, 1.1667, 0.3333)
Max Dist: 6.1667
Min Dist: 2.1667
Avg Dist: 4.1111
Dist-Variance: 2.4969
Size: 6
Sample: [(1.0, 4.0, 0.0, 2.0, 35.0, 1.0, 0.0), (2.0, 6.0, 0.0, 3.0, 35.0, 1.0, 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), (2.0, 6.0, 0.0, 4.0, 35.0, 1.0, 0.0)]

Cluster4
Center: (1.125, 1.75, 0.625, 8.0, 35.0, 1.0, 0.0)
Max Dist: 3.75
Min Dist: 1.75
Avg Dist: 2.5
Dist-Variance: 0.6719
Size: 8
Sample: [(2.0, 1.0, 0.0, 9.0, 35.0, 1.0, 0.0), (1.0, 3.0, 1.0, 6.0, 35.0, 1.0, 0.0), (1.0, 2.0, 1.0, 9.0, 35.0, 1.0, 0.0), (1.0, 1.0, 0.0, 6.0, 35.0, 1.0, 0.0), (1.0, 1.0, 0.0, 9.0, 35.0, 1.0, 0.0)]

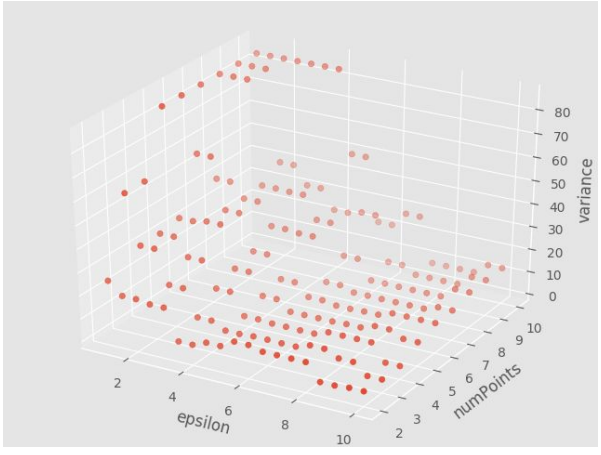
Hierarchical



link-method: CENTROID
threshold: 10

Cluster0
Center: (1.5909, 3.5455, 0.5, 5.4091, 35.0, 1.1818, 0.2273)
Max Dist: 11.0
Min Dist: 2.4545
Avg Dist: 5.5826
Dist-Variance: 4.8833
Size: 22
Sample: [(2.0, 4.0, 0.0, 4.0, 35.0, 1.0, 1.0), (3.0, 4.0, 0.0, 4.0, 35.0, 1.0, 1.0), (2.0, 9.0, 0.0, 2.0, 35.0, 1.0, 1.0), (1.0, 2.0, 1.0, 7.0, 35.0, 1.0, 0.0), (1.0, 4.0, 0.0, 4.0, 35.0, 3.0, 1.0)]

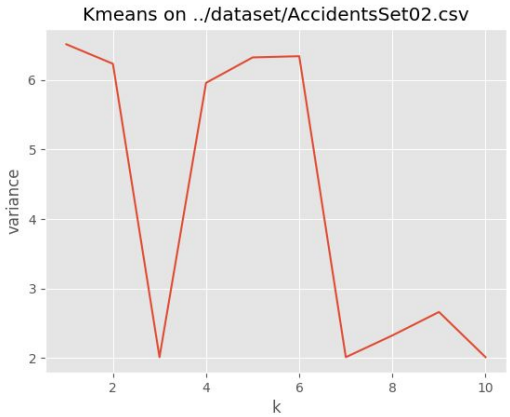
Cluster1
Center: (1.4167, 3.0833, 0.9167, 4.0, 45.0, 1.1667, 0.4167)
Max Dist: 8.6667
Min Dist: 1.1667
Avg Dist: 3.0556
Dist-Variance: 4.1497
Size: 12
Sample: [(5.0, 6.0, 0.0, 4.0, 45.0, 2.0, 0.0), (1.0, 2.0, 1.0, 4.0, 45.0, 1.0, 0.0), (1.0, 2.0, 1.0, 4.0, 45.0, 1.0,

5 clusters	0.0), (1.0, 3.0, 1.0, 4.0, 45.0, 1.0, 1.0), (1.0, 3.0, 2.0, 4.0, 45.0, 2.0, 0.0)] Cluster2 Center: (1.2308, 2.6154, 0.0, 2.0, 70.0, 1.1538, 0.3846) Max Dist: 7.0769 Min Dist: 1.1538 Avg Dist: 3.0414 Dist-Variance: 2.9828 Size: 13 Sample: [(2.0, 2.0, 0.0, 2.0, 70.0, 1.0, 0.0), (1.0, 1.0, 0.0, 2.0, 70.0, 1.0, 0.0), (1.0, 1.0, 0.0, 2.0, 70.0, 1.0, 1.0), (1.0, 1.0, 0.0, 2.0, 70.0, 1.0, 0.0), (1.0, 3.0, 0.0, 2.0, 70.0, 1.0, 0.0)] Cluster3 Center: (6.0, 9.0, 0.0, 2.0, 70.0, 2.0, 0.0) Max Dist: 0.0 Min Dist: 0.0 Avg Dist: 0.0 Dist-Variance: 0.0 Size: 1 Sample: [(6.0, 9.0, 0.0, 2.0, 70.0, 2.0, 0.0)] Cluster4 Center: (4.0, 25.0, 0.0, 2.0, 70.0, 2.0, 0.0) Max Dist: 0.0 Min Dist: 0.0 Avg Dist: 0.0 Dist-Variance: 0.0 Size: 1 Sample: [(4.0, 25.0, 0.0, 2.0, 70.0, 2.0, 0.0)]
DBSCAN  **Best Variance = 3, achieved with e=9.5, numPoints=3**	Cluster0 Center: (1.5909, 3.5455, 0.5, 5.4091, 35.0, 1.1818, 0.2273) Max Dist: 11.0 Min Dist: 2.4545 Avg Dist: 5.5826 Dist-Variance: 4.8833 Size: 22 Sample: [(3.0, 4.0, 0.0, 4.0, 35.0, 1.0, 1.0), (1.0, 4.0, 0.0, 2.0, 35.0, 1.0, 0.0), (1.0, 1.0, 0.0, 9.0, 35.0, 1.0, 0.0), (2.0, 4.0, 0.0, 4.0, 35.0, 1.0, 1.0), (2.0, 6.0, 0.0, 3.0, 35.0, 1.0, 1.0)] Cluster1 Center: (1.4167, 3.0833, 0.9167, 4.0, 45.0, 1.1667, 0.4167) Max Dist: 8.6667 Min Dist: 1.1667 Avg Dist: 3.0556 Dist-Variance: 4.1497 Size: 12 Sample: [(1.0, 7.0, 0.0, 4.0, 45.0, 1.0, 0.0), (1.0, 2.0, 1.0, 4.0, 45.0, 1.0, 0.0), (2.0, 2.0, 0.0, 4.0, 45.0, 1.0, 1.0), (1.0, 2.0, 1.0, 4.0, 45.0, 1.0, 1.0), (1.0, 3.0, 2.0, 4.0, 45.0, 1.0, 0.0)]

	Cluster2 Center: (1.2308, 2.6154, 0.0, 2.0, 70.0, 1.1538, 0.3846) Max Dist: 7.0769 Min Dist: 1.1538 Avg Dist: 3.0414 Dist-Variance: 2.9828 Size: 13 Sample: [(1.0, 8.0, 0.0, 2.0, 70.0, 1.0, 0.0), (1.0, 1.0, 0.0, 2.0, 70.0, 1.0, 0.0), (1.0, 1.0, 0.0, 2.0, 70.0, 1.0, 0.0), (1.0, 3.0, 0.0, 2.0, 70.0, 1.0, 0.0), (1.0, 2.0, 0.0, 2.0, 70.0, 1.0, 2.0)] Cluster3 Center: (5.0, 17.0, 0.0, 2.0, 70.0, 2.0, 0.0) Max Dist: 9.0 Min Dist: 9.0 Avg Dist: 9.0 Dist-Variance: 0.0 Size: 2 Sample: [(4.0, 25.0, 0.0, 2.0, 70.0, 2.0, 0.0), (6.0, 9.0, 0.0, 2.0, 70.0, 2.0, 0.0)]

Accidents3

K-Means



3 Clusters found

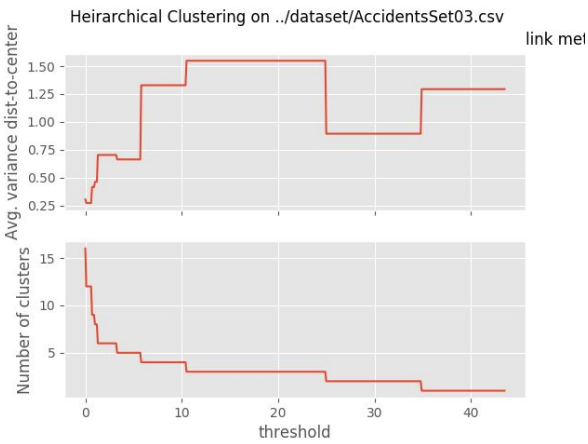
Cluster0
Center: (1.7333, 4.5333, 0.0, 2.0, 70.0, 1.2667, 0.3333)
Max Dist: 23.8
Min Dist: 2.3333
Avg Dist: 6.08
Dist-Variance: 25.004
Size: 15
Sample: [(1.0, 1.0, 0.0, 2.0, 70.0, 1.0, 0.0), (1.0, 1.0, 0.0, 2.0, 70.0, 1.0, 1.0), (1.0, 2.0, 0.0, 2.0, 70.0, 1.0, 0.0), (4.0, 25.0, 0.0, 2.0, 70.0, 2.0, 0.0), (1.0, 8.0, 0.0, 2.0, 70.0, 1.0, 0.0)]

Cluster1
Center: (1.4167, 3.0833, 0.9167, 4.0, 45.0, 1.1667, 0.4167)
Max Dist: 8.6667
Min Dist: 1.1667
Avg Dist: 3.0556
Dist-Variance: 4.1497
Size: 12
Sample: [(1.0, 3.0, 1.0, 4.0, 45.0, 1.0, 0.0), (1.0, 2.0, 1.0, 4.0, 45.0, 1.0, 1.0), (1.0, 7.0, 0.0, 4.0, 45.0, 1.0, 0.0), (1.0, 2.0, 1.0, 4.0, 45.0, 1.0, 1.0), (2.0, 2.0, 0.0, 4.0, 45.0, 1.0, 1.0)]

Cluster2
Center: (1.5909, 3.5455, 0.5, 5.4091, 35.0, 1.1818, 0.2273)
Max Dist: 11.0
Min Dist: 2.4545
Avg Dist: 5.5826
Dist-Variance: 4.8833
Size: 22
Sample: [(1.0, 2.0, 1.0, 9.0, 35.0, 1.0, 0.0), (1.0, 4.0, 0.0, 4.0, 35.0, 3.0, 1.0), (1.0, 1.0, 0.0, 6.0, 35.0, 1.0, 0.0), (1.0, 2.0, 1.0, 9.0, 35.0, 1.0, 0.0), (3.0, 3.0, 0.0, 6.0, 35.0, 1.0, 0.0)]

Hierarchical

Cluster0
Center: (1.5, 0.5, 2.0, 1.25, 0.75)
Max Dist: 3.0



link-method: WARDS
threshold: 2

6 clusters

Min Dist: 1.5
Avg Dist: 2.0
Dist-Variance: 0.375
Size: 4
Sample: [(3.0, 0.0, 2.0, 1.0, 0.0), (1.0, 1.0, 2.0, 1.0, 1.0), (1.0, 1.0, 2.0, 1.0, 1.0), (1.0, 0.0, 2.0, 2.0, 1.0)]

Cluster1
Center: (1.6471, 0.6471, 3.2941, 1.1176, 0.2941)
Max Dist: 5.1765
Min Dist: 2.1176
Avg Dist: 2.9412
Dist-Variance: 1.1187
Size: 17
Sample: [(1.0, 1.0, 4.0, 1.0, 0.0), (1.0, 1.0, 4.0, 1.0, 0.0), (1.0, 1.0, 4.0, 1.0, 0.0), (1.0, 0.0, 4.0, 1.0, 1.0), (1.0, 1.0, 4.0, 1.0, 0.0)]

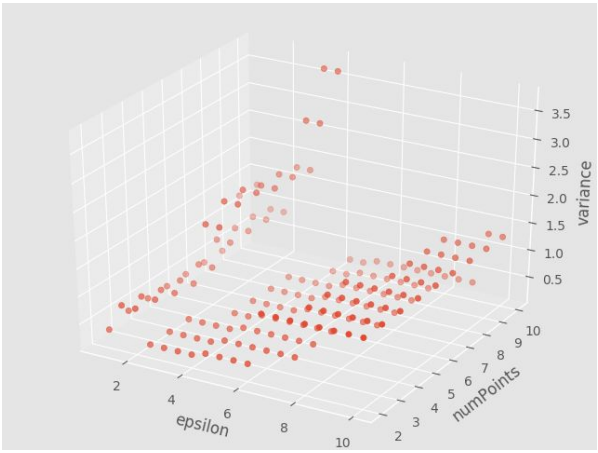
Cluster2
Center: (1.6667, 0.7333, 3.7333, 1.0667, 0.8)
Max Dist: 3.6
Min Dist: 1.4667
Avg Dist: 2.2044
Dist-Variance: 0.8341
Size: 15
Sample: [(2.0, 1.0, 2.0, 1.0, 2.0), (1.0, 1.0, 4.0, 1.0, 1.0), (1.0, 1.0, 4.0, 1.0, 1.0), (3.0, 1.0, 2.0, 1.0, 1.0), (1.0, 1.0, 4.0, 1.0, 1.0)]

Cluster3
Center: (2.7, 0.4, 4.0, 1.2, 1.4)
Max Dist: 5.5
Min Dist: 1.3
Avg Dist: 3.08
Dist-Variance: 1.7316
Size: 10
Sample: [(5.0, 0.0, 4.0, 1.0, 0.0), (4.0, 0.0, 4.0, 1.0, 0.0), (1.0, 1.0, 4.0, 1.0, 2.0), (1.0, 1.0, 4.0, 1.0, 2.0), (5.0, 0.0, 4.0, 1.0, 1.0)]

Cluster4
Center: (10.0, 0.0, 4.0, 1.0, 1.0)
Max Dist: 0.0
Min Dist: 0.0
Avg Dist: 0.0
Dist-Variance: 0.0
Size: 1
Sample: [(10.0, 0.0, 4.0, 1.0, 1.0)]

Cluster5
Center: (1.1333, 0.2, 2.0, 1.0, 0.6)
Max Dist: 1.6667
Min Dist: 0.7333
Avg Dist: 1.0311
Dist-Variance: 0.1581
Size: 15
Sample: [(1.0, 1.0, 2.0, 1.0, 0.0), (1.0, 0.0, 2.0, 1.0, 1.0), (1.0, 0.0, 2.0, 1.0, 1.0), (1.0, 1.0, 2.0, 1.0, 0.0), (1.0, 0.0, 2.0, 1.0, 1.0)]

DBSCAN



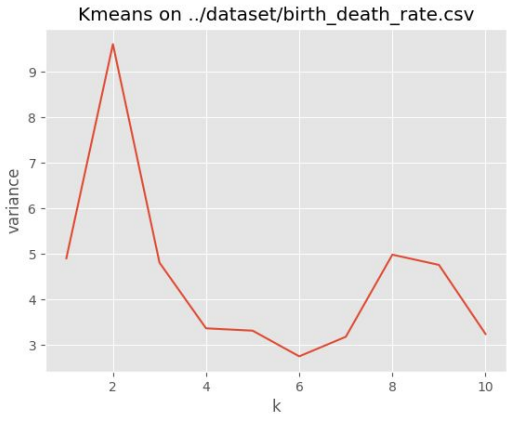
2 Clusters found
Best Variance = 0.2543, achieved with e=2.5, numPoints=2

Cluster0
Center: (1.7, 0.5167, 3.1, 1.0667, 0.6667)
Max Dist: 5.45
Min Dist: 2.2833
Avg Dist: 3.1422
Dist-Variance: 0.5086
Size: 60
Sample: [(5.0, 0.0, 4.0, 1.0, 0.0), (1.0, 1.0, 4.0, 1.0, 0.0), (1.0, 1.0, 4.0, 1.0, 1.0), (3.0, 0.0, 4.0, 1.0, 2.0), (1.0, 1.0, 2.0, 1.0, 0.0)]

Cluster1
Center: (5.5, 0.0, 4.0, 2.0, 2.0)
Max Dist: 6.5
Min Dist: 6.5
Avg Dist: 6.5
Dist-Variance: 0.0
Size: 2
Sample: [(1.0, 0.0, 4.0, 3.0, 3.0), (10.0, 0.0, 4.0, 1.0, 1.0)]

Birth-Death Rate

K-Means

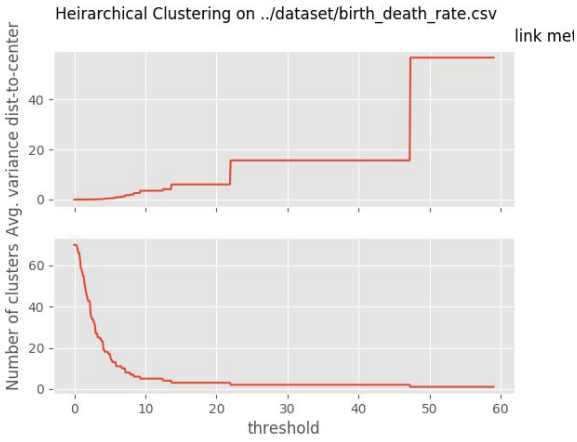


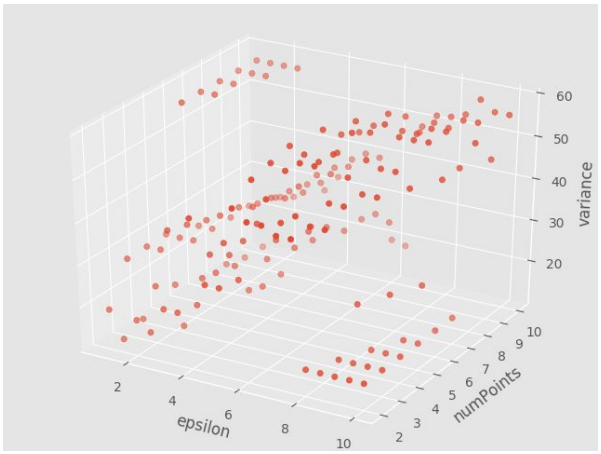
6 Clusters found

Cluster0
Center: (34.68, 8.82)
Max Dist: 7.5
Min Dist: 1.04
Avg Dist: 3.628
Dist-Variance: 3.5294
Size: 10
Sample: [(33.5, 6.4), (32.1, 5.5), (37.3, 8.0), (36.4, 14.6), (34.8, 7.9)]

Cluster1
Center: (17.4429, 11.6143)
Max Dist: 8.3429
Min Dist: 0.8429
Avg Dist: 3.0122
Dist-Variance: 3.9317
Size: 14
Sample: [(18.1, 9.2), (18.9, 9.6), (14.8, 10.1), (19.0, 10.2), (17.5, 13.7)]

Cluster2
Center: (17.3125, 6.775)
Max Dist: 6.0625
Min Dist: 0.2375
Avg Dist: 2.4938
Dist-Variance: 2.9281

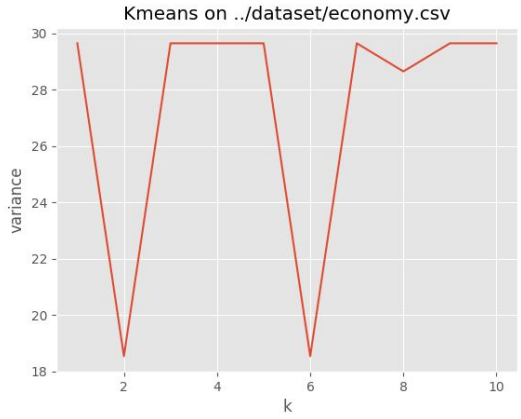
	Size: 8 Sample: [(15.7, 8.3), (16.4, 8.2), (17.4, 7.8), (20.5, 3.9), (19.0, 7.5)] Cluster3 Center: (49.42, 22.88) Max Dist: 16.9 Min Dist: 7.3 Avg Dist: 10.4 Dist-Variance: 12.52 Size: 5 Sample: [(46.1, 18.7), (56.1, 33.1), (41.4, 19.7), (55.8, 25.6), (47.7, 17.3)] Cluster4 Center: (23.5789, 8.2421) Max Dist: 6.5632 Min Dist: 1.5211 Avg Dist: 3.4598 Dist-Variance: 2.0287 Size: 19 Sample: [(20.9, 8.8), (21.8, 8.1), (23.4, 5.1), (25.0, 6.2), (23.5, 10.8)] Cluster5 Center: (43.6929, 10.5714) Max Dist: 8.2786 Min Dist: 1.4357 Avg Dist: 4.6888 Dist-Variance: 4.0591 Size: 14 Sample: [(44.0, 11.7), (42.9, 7.1), (41.8, 15.8), (44.2, 13.5), (41.7, 10.1)]
Hierarchical  **link-method: CENTROID** **threshold: 10** **5 clusters**	Cluster0 Center: (34.68, 8.82) Max Dist: 7.5 Min Dist: 1.04 Avg Dist: 3.628 Dist-Variance: 3.5294 Size: 10 Sample: [(36.4, 14.6), (33.5, 6.4), (33.6, 11.8), (34.8, 7.9), (34.0, 11.0)] Cluster1 Center: (43.9353, 11.9824) Max Dist: 10.2529 Min Dist: 0.3471 Avg Dist: 5.8125 Dist-Variance: 7.8075 Size: 17 Sample: [(40.1, 8.0), (41.4, 19.7), (46.3, 6.4), (45.0, 13.5), (49.9, 8.5)] Cluster2 Center: (20.3275, 8.84) Max Dist: 10.4125 Min Dist: 0.6125 Avg Dist: 4.8645

	Dist-Variance: 6.2663 Size: 40 Sample: [(15.7, 8.3), (19.0, 7.5), (23.5, 10.8), (28.4, 7.1), (17.1, 12.7)] Cluster3 Center: (17.6, 19.8) Max Dist: 0.0 Min Dist: 0.0 Avg Dist: 0.0 Dist-Variance: 0.0 Size: 1 Sample: [(17.6, 19.8)] Cluster4 Center: (55.95, 29.35) Max Dist: 3.9 Min Dist: 3.9 Avg Dist: 3.9 Dist-Variance: 0.0 Size: 2 Sample: [(56.1, 33.1), (55.8, 25.6)]
DBSCAN  **5 clusters found.** **Best Variance = 12.48706, achieved with e=1.5, numPoints=2**	Cluster0 Center: (21.4, 8.7429) Max Dist: 1.2429 Min Dist: 0.1571 Avg Dist: 0.6939 Dist-Variance: 0.172 Size: 7 Sample: [(21.4, 8.9), (21.7, 9.6), (21.6, 8.7), (21.5, 9.1), (21.8, 8.1)] Cluster1 Center: (17.0333, 7.6667) Max Dist: 1.1667 Min Dist: 0.5 Avg Dist: 0.8667 Dist-Variance: 0.0763 Size: 3 Sample: [(16.4, 8.2), (17.3, 7.0), (17.4, 7.8)] Cluster2 Center: (18.3, 11.7667) Max Dist: 0.5667 Min Dist: 0.1667 Avg Dist: 0.4222 Dist-Variance: 0.0328 Size: 3 Sample: [(18.5, 11.4), (18.2, 12.2), (18.2, 11.7)] Cluster3 Center: (17.85, 9.525) Max Dist: 1.175 Min Dist: 0.575 Avg Dist: 0.925 Dist-Variance: 0.0587 Size: 4 Sample: [(18.1, 9.2), (18.9, 9.6), (16.9, 9.3), (17.5, 10.0)]

Cluster4
Center: (32.2472, 10.6868)
Max Dist: 46.266
Min Dist: 2.066
Avg Dist: 14.013
Dist-Variance: 62.0955
Size: 53
Sample: [(18.8, 12.8), (44.0, 11.7), (47.7, 17.3), (37.3, 8.0), (33.5, 6.4)]

Economy

K-Means



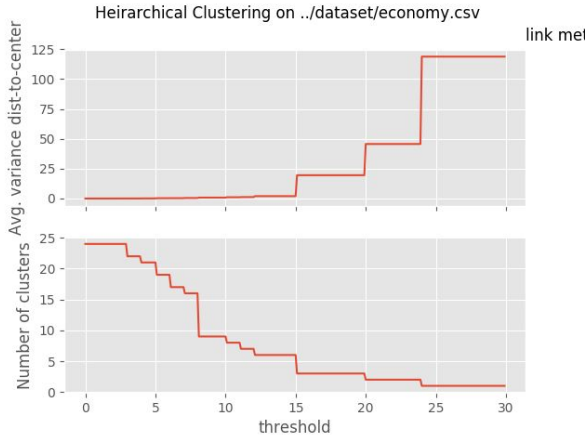
6 clusters

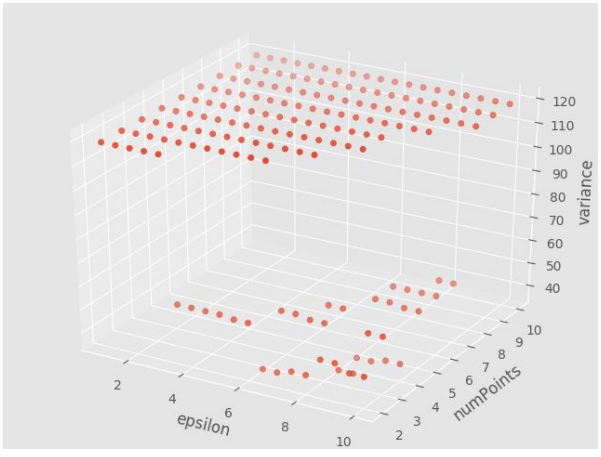
Cluster0
Center: (13.5, 12.75, 12.0, 13.25, 13.5, 14.5, 16.75, 16.5, 14.25, 14.75)
Max Dist: 17.75
Min Dist: 10.75
Avg Dist: 14.25
Dist-Variance: 10.625
Size: 4
Sample: [(13, 12, 11, 12, 12, 14, 18, 21, 18, 17), (14, 12, 12, 12, 12, 14, 15, 15, 13, 13), (13, 13, 14, 13, 13, 13, 14, 14, 14, 14), (14, 14, 11, 16, 17, 17, 20, 16, 12, 15)]

Cluster1
Center: (10.6667, 9.3333, 8.6667, 8.6667, 8.6667, 9.6667, 10.0, 10.6667, 9.3333, 10.0)
Max Dist: 7.3333
Min Dist: 5.3333
Avg Dist: 6.6667
Dist-Variance: 0.8889
Size: 3
Sample: [(13, 10, 9, 9, 9, 10, 10, 10, 8, 9), (10, 9, 8, 8, 8, 9, 9, 11, 9, 10), (9, 9, 9, 9, 9, 10, 11, 11, 11, 11)]

Cluster2
Center: (10.4444, 9.4444, 9.1111, 9.7778, 9.7778, 11.6667, 12.8889, 13.8889, 12.3333, 12.1111)
Max Dist: 9.7778
Min Dist: 3.2222
Avg Dist: 6.9877
Dist-Variance: 4.624
Size: 9
Sample: [(10, 9, 9, 10, 10, 12, 13, 13, 12, 12), (10, 9, 8, 10, 10, 12, 14, 14, 12, 12), (10, 10, 10, 10, 11, 11, 12, 12, 13, 12), (9, 9, 10, 9, 9, 10, 11, 15, 13, 12), (13, 10, 9, 10, 10, 11, 14, 15, 13, 12)]

Cluster3
Center: (8.0, 6.5, 5.5, 5.5, 6.5, 9.0, 10.5, 10.0, 8.0, 8.5)
Max Dist: 3.0
Min Dist: 3.0
Avg Dist: 3.0

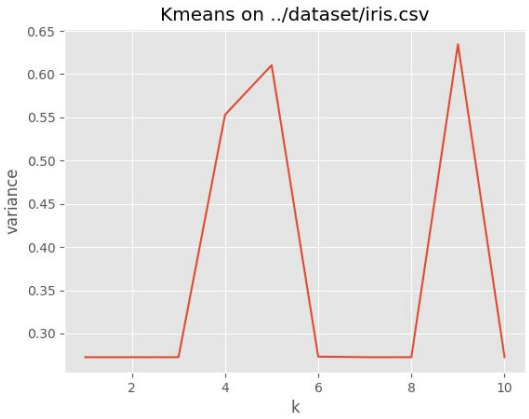
	Dist-Variance: 0.0 Size: 2 Sample: [(8, 6, 5, 6, 6, 9, 11, 10, 8, 9), (8, 7, 6, 5, 7, 9, 10, 10, 8, 8)] Cluster4 Center: (9.0, 5.0, 4.0, 6.5, 7.5, 10.5, 11.0, 11.5, 10.5, 14.0) Max Dist: 7.5 Min Dist: 7.5 Avg Dist: 7.5 Dist-Variance: 0.0 Size: 2 Sample: [(9, 6, 4, 7, 7, 11, 12, 13, 12, 13), (9, 4, 4, 6, 8, 10, 10, 10, 9, 15)] Cluster5 Center: (8.5, 7.0, 6.25, 8.25, 8.0, 10.5, 12.75, 14.25, 12.0, 12.0) Max Dist: 7.0 Min Dist: 3.0 Avg Dist: 5.0 Dist-Variance: 2.125 Size: 4 Sample: [(8, 7, 7, 8, 8, 10, 12, 15, 11, 11), (9, 7, 5, 8, 8, 10, 13, 14, 12, 12), (8, 6, 6, 8, 8, 10, 13, 15, 13, 12), (9, 8, 7, 9, 8, 12, 13, 13, 12, 13)]
Hierarchical  **link-method: SINGLE** **threshold: 14** **6 clusters**	Cluster0 Center: (9.9412, 8.6471, 8.0588, 9.0588, 9.0, 11.0, 12.2941, 13.3529, 11.7059, 11.7647) Max Dist: 18.5294 Min Dist: 5.8824 Avg Dist: 10.4637 Dist-Variance: 11.9145 Size: 17 Sample: [(10, 10, 10, 10, 10, 12, 12, 13, 12, 12), (8, 7, 7, 8, 8, 10, 12, 15, 11, 11), (10, 9, 9, 10, 10, 12, 13, 13, 12, 12), (11, 11, 9, 10, 9, 13, 14, 16, 13, 13), (13, 10, 9, 10, 10, 11, 14, 15, 13, 12)] Cluster1 Center: (13.5, 12.5, 13.0, 12.5, 12.5, 13.5, 14.5, 14.5, 13.5, 13.5) Max Dist: 5.5 Min Dist: 5.5 Avg Dist: 5.5 Dist-Variance: 0.0 Size: 2 Sample: [(14, 12, 12, 12, 12, 14, 15, 15, 13, 13), (13, 13, 14, 13, 13, 13, 14, 14, 14, 14)] Cluster2 Center: (8.0, 6.5, 5.5, 5.5, 6.5, 9.0, 10.5, 10.0, 8.0, 8.5) Max Dist: 3.0 Min Dist: 3.0 Avg Dist: 3.0 Dist-Variance: 0.0

	Size: 2 Sample: [(8, 6, 5, 6, 6, 9, 11, 10, 8, 9), (8, 7, 6, 5, 7, 9, 10, 10, 8, 8)] Cluster3 Center: (9.0, 4.0, 4.0, 6.0, 8.0, 10.0, 10.0, 10.0, 9.0, 15.0) Max Dist: 0.0 Min Dist: 0.0 Avg Dist: 0.0 Dist-Variance: 0.0 Size: 1 Sample: [(9, 4, 4, 6, 8, 10, 10, 10, 9, 15)] Cluster4 Center: (13.0, 12.0, 11.0, 12.0, 12.0, 14.0, 18.0, 21.0, 18.0, 17.0) Max Dist: 0.0 Min Dist: 0.0 Avg Dist: 0.0 Dist-Variance: 0.0 Size: 1 Sample: [(13, 12, 11, 12, 12, 14, 18, 21, 18, 17)] Cluster5 Center: (14.0, 14.0, 11.0, 16.0, 17.0, 17.0, 20.0, 16.0, 12.0, 15.0) Max Dist: 0.0 Min Dist: 0.0 Avg Dist: 0.0 Dist-Variance: 0.0 Size: 1 Sample: [(14, 14, 11, 16, 17, 17, 20, 16, 12, 15)]
DBSCAN  **3 clusters found.** **Best Variance = 38.30786666666666, achieved with e=8.5, numPoints=4**	Cluster0 Center: (10.25, 9.25, 9.0, 10.0, 9.75, 11.75, 12.75, 13.0, 11.5, 12.0) Max Dist: 5.75 Min Dist: 1.75 Avg Dist: 4.0 Dist-Variance: 2.1875 Size: 4 Sample: [(10, 9, 9, 10, 10, 12, 13, 13, 12, 12), (10, 9, 8, 10, 10, 12, 14, 14, 12, 12), (11, 9, 9, 10, 9, 11, 12, 12, 10, 12), (10, 10, 10, 10, 10, 12, 12, 13, 12, 12)] Cluster1 Center: (10.4211, 9.0, 8.3684, 9.2632, 9.5263, 11.3684, 12.7895, 13.5263, 11.7368, 12.1579) Max Dist: 43.8421 Min Dist: 8.0526 Avg Dist: 19.662 Dist-Variance: 112.7361 Size: 19 Sample: [(11, 11, 9, 10, 9, 13, 14, 16, 13, 13), (8, 6, 6, 8, 8, 10, 13, 15, 13, 12), (9, 9, 9, 9, 9, 10, 11, 11, 11, 11), (9, 6, 4, 7, 7, 11, 12, 13, 12, 13), (10, 9, 8, 8, 8, 9, 9, 11, 9, 10)] Cluster2

	Center: (9.0, 7.0, 5.0, 8.0, 8.0, 10.0, 13.0, 14.0, 12.0, 12.0) Max Dist: 0.0 Min Dist: 0.0 Avg Dist: 0.0 Dist-Variance: 0.0 Size: 1 Sample: [(9, 7, 5, 8, 8, 10, 13, 14, 12, 12)]

Iris

K-Means



3 clusters

Cluster0
Center: (6.8703, 3.0865, 5.7459, 2.0892)
Max Dist: 2.6811
Min Dist: 0.4
Avg Dist: 1.2066
Dist-Variance: 0.3339
Size: 37
Sample: [(6.9, 3.1, 5.4, 2.1), (7.9, 3.8, 6.4, 2.0), (6.5, 3.0, 5.8, 2.2), (7.1, 3.0, 5.9, 2.1), (6.7, 2.5, 5.8, 1.8)]

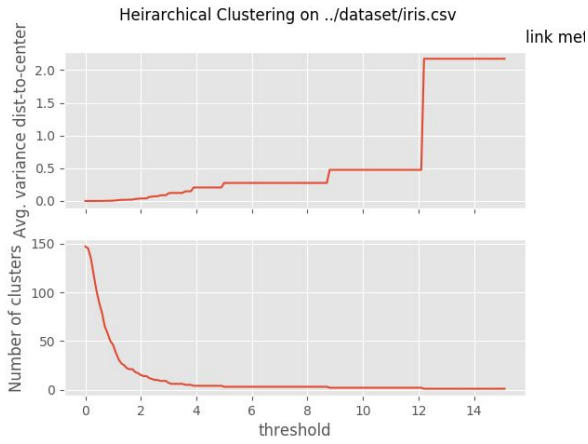
Cluster1
Center: (5.006, 3.418, 1.464, 0.244)
Max Dist: 1.868
Min Dist: 0.104
Avg Dist: 0.7743
Dist-Variance: 0.1799
Size: 50
Sample: [(5.4, 3.9, 1.7, 0.4), (4.6, 3.6, 1.0, 0.2), (5.0, 3.4, 1.6, 0.4), (4.3, 3.0, 1.1, 0.1), (5.4, 3.7, 1.5, 0.2)]

Cluster2
Center: (5.9048, 2.746, 4.4127, 1.4333)
Max Dist: 2.9968
Min Dist: 0.4032
Avg Dist: 1.2527
Dist-Variance: 0.3041
Size: 63
Sample: [(5.0, 2.0, 3.5, 1.0), (5.8, 2.7, 3.9, 1.2), (6.6, 3.0, 4.4, 1.4), (5.9, 3.0, 4.2, 1.5), (6.2, 2.9, 4.3, 1.3)]

Hierarchical

Cluster0
Center: (5.006, 3.418, 1.464, 0.244)
Max Dist: 1.868
Min Dist: 0.104
Avg Dist: 0.7743
Dist-Variance: 0.1799
Size: 50
Sample: [(4.4, 3.2, 1.3, 0.2), (5.1, 3.5, 1.4, 0.3), (4.8, 3.4, 1.6, 0.2), (4.7, 3.2, 1.3, 0.2), (4.9, 3.1, 1.5, 0.1)]

Cluster1
Center: (5.9394, 2.7545, 4.4424, 1.4455)
Max Dist: 3.0818



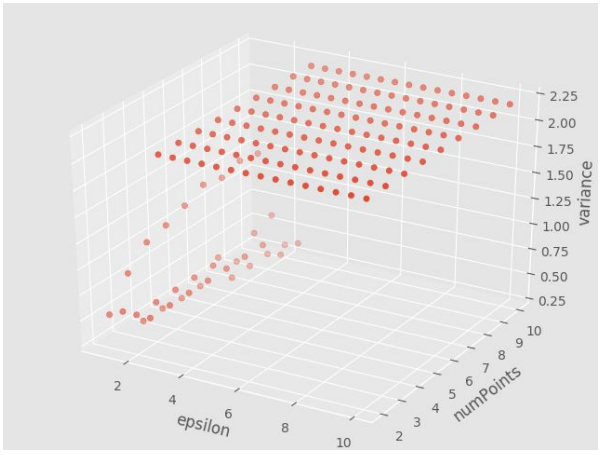
link-method: SINGLE
threshold: 6

3 clusters

Min Dist: 0.3182
Avg Dist: 1.2789
Dist-Variance: 0.3257
Size: 66
Sample: [(6.4, 3.2, 4.5, 1.5), (6.0, 2.7, 5.1, 1.6), (6.0, 2.2, 5.0, 1.5), (5.5, 2.3, 4.0, 1.3), (5.0, 2.3, 3.3, 1.0)]

Cluster2
Center: (6.8882, 3.1, 5.8059, 2.1235)
Max Dist: 2.5824
Min Dist: 0.3941
Avg Dist: 1.1896
Dist-Variance: 0.3213
Size: 34
Sample: [(7.2, 3.0, 5.8, 1.6), (6.4, 2.8, 5.6, 2.1), (6.2, 3.4, 5.4, 2.3), (6.5, 3.0, 5.5, 1.8), (6.3, 3.3, 6.0, 2.5)]

DBSCAN



Best Variance = 0.33516666666666667, achieved with e=2.0, numPoints=10

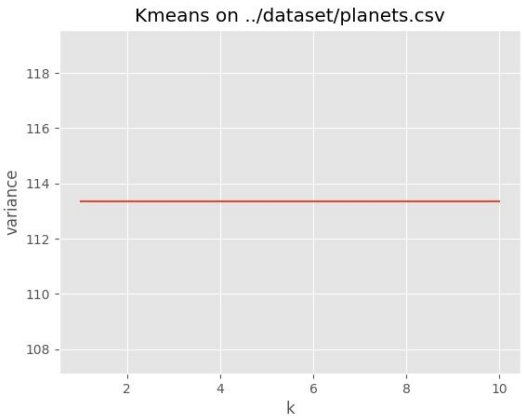
Cluster0
Center: (5.006, 3.418, 1.464, 0.244)
Max Dist: 1.868
Min Dist: 0.104
Avg Dist: 0.7743
Dist-Variance: 0.1799
Size: 50
Sample: [(5.0, 3.2, 1.2, 0.2), (5.1, 3.5, 1.4, 0.2), (5.4, 3.9, 1.3, 0.4), (5.1, 3.3, 1.7, 0.5), (4.6, 3.1, 1.5, 0.2)]

Cluster1
Center: (6.2, 2.8562, 4.8323, 1.6573)
Max Dist: 4.0458
Min Dist: 0.2313
Avg Dist: 1.707
Dist-Variance: 0.7801
Size: 96
Sample: [(6.8, 3.2, 5.9, 2.3), (5.7, 2.8, 4.1, 1.3), (6.3, 2.3, 4.4, 1.3), (6.7, 3.3, 5.7, 2.1), (6.5, 3.2, 5.1, 2.0)]

Cluster2
Center: (7.75, 3.25, 6.675, 2.125)
Max Dist: 1.1
Min Dist: 0.65
Avg Dist: 0.8875
Dist-Variance: 0.0455
Size: 4
Sample: [(7.7, 2.8, 6.7, 2.0), (7.7, 3.8, 6.7, 2.2), (7.9, 3.8, 6.4, 2.0), (7.7, 2.6, 6.9, 2.3)]

Planets

K-Means



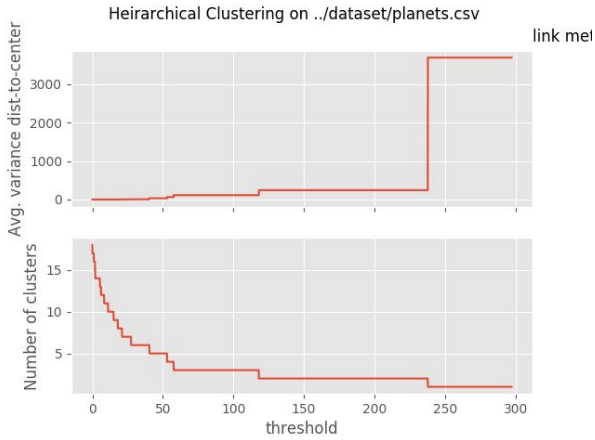
No good metric, but 3 clusters can be found

Cluster0
Center: (141.1922, 4.309, 2.5456)
Max Dist: 56.3911
Min Dist: 8.9556
Avg Dist: 22.0222
Dist-Variance: 213.3289
Size: 9
Sample: [(135.6, 1.0, 2.6), (194.6, 1.8, 3.02), (132.2, 4.7, 2.13), (130.916, 4.659, 2.2562), (115.072, 2.6660000000000004, 3.1676)]

Cluster1
Center: (68.5565, 5.6732, 2.7605)
Max Dist: 41.2428
Min Dist: 2.0663
Avg Dist: 16.3368
Dist-Variance: 126.6804
Size: 8
Sample: [(34.2, 12.5, 2.82), (89.9, 2.1, 3.35), (59.9, 5.7, 2.79), (78.1, 6.6, 2.9), (80.804, 4.622, 2.189)]

Cluster2
Center: (338.979, 16.42, 2.74)
Max Dist: 1.0055
Min Dist: 1.0055
Avg Dist: 1.0055
Dist-Variance: 0.0
Size: 2
Sample: [(338.333, 16.773, 2.7465), (339.625, 16.067, 2.7335)]

Hierarchical



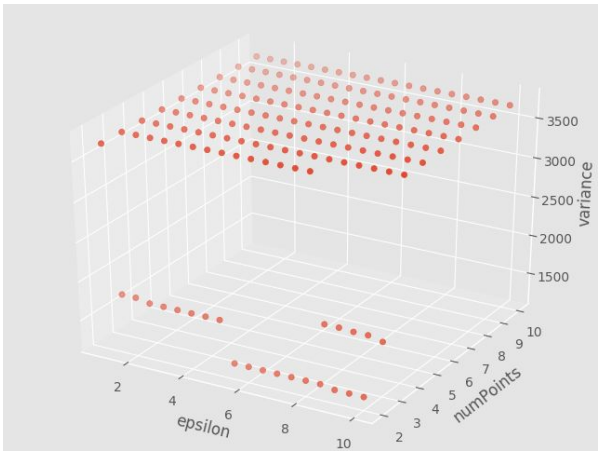
link-method: AVERAGE
threshold: 100

3 clusters

Cluster0
Center: (141.1922, 4.309, 2.5456)
Max Dist: 56.3911
Min Dist: 8.9556
Avg Dist: 22.0222
Dist-Variance: 213.3289
Size: 9
Sample: [(135.6, 1.0, 2.6), (115.072, 2.6660000000000004, 3.1676), (194.6, 1.8, 3.02), (130.916, 4.659, 2.2562), (164.1, 10.0, 1.93)]

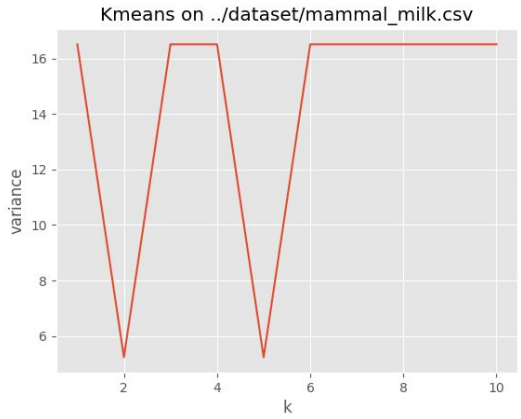
Cluster1
Center: (68.5565, 5.6732, 2.7605)
Max Dist: 41.2428
Min Dist: 2.0663
Avg Dist: 16.3368
Dist-Variance: 126.6804
Size: 8
Sample: [(69.6, 4.7, 2.81), (34.2, 12.5, 2.82), (89.9, 2.1, 3.35), (55.144, 4.542, 3.0343), (59.9, 5.7, 2.79)]

Cluster2
Center: (338.979, 16.42, 2.74)
Max Dist: 1.0055
Min Dist: 1.0055

	Avg Dist: 1.0055 Dist-Variance: 0.0 Size: 2 Sample: [(338.333, 16.773, 2.7465), (339.625, 16.067, 2.7335)]
DBSCAN  **3 clusters** **Best Variance = 1265.3577666666667, achieved with e=9.0, numPoints=2**	Cluster0 Center: (132.1965, 3.7872, 2.2939) Max Dist: 6.4969 Min Dist: 1.0801 Avg Dist: 3.2502 Dist-Variance: 4.0936 Size: 4 Sample: [(132.2, 4.7, 2.13), (135.6, 1.0, 2.6), (130.916, 4.659, 2.2562), (130.07, 4.79, 2.1893)] Cluster1 Center: (144.0538, 7.1678, 2.835) Max Dist: 204.5718 Min Dist: 10.099 Avg Dist: 82.9537 Dist-Variance: 3791.2618 Size: 12 Sample: [(55.144, 4.542, 3.0343), (339.625, 16.067, 2.7335), (69.6, 4.7, 2.81), (338.333, 16.773, 2.7465), (34.2, 12.5, 2.82)] Cluster2 Center: (79.9027, 5.2813, 2.4265) Max Dist: 3.5948 Min Dist: 1.7966 Avg Dist: 2.3965 Dist-Variance: 0.7179 Size: 3 Sample: [(80.804, 4.622, 2.1906), (80.804, 4.622, 2.189), (78.1, 6.6, 2.9)]

Mammal Milk

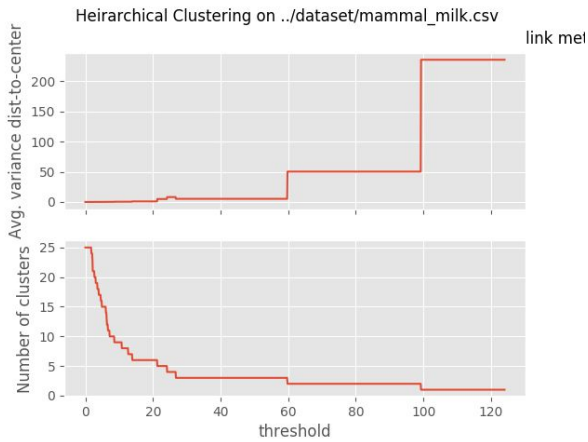
K-Means	Cluster0 Center: (85.4882, 4.5706, 4.4882, 4.9941, 0.6688) Max Dist: 21.4547 Min Dist: 2.9312 Avg Dist: 9.1152 Dist-Variance: 17.6757 Size: 17 Sample: [(82.8, 7.1, 5.1, 3.7, 1.1), (90.1, 2.6, 1.0, 6.9, 0.35), (88.5, 1.4, 3.5, 6.0, 0.24), (81.6, 10.1, 6.3, 4.4, 0.75), (86.9, 4.8, 1.7, 5.7, 0.9)] Cluster1 Center: (62.6625, 9.7, 22.675, 2.3, 1.17) Max Dist: 38.2075 Min Dist: 5.9425 Avg Dist: 19.6988



2 clusters found

Dist-Variance: 133.2257
Size: 8
Sample: [(71.3, 12.3, 13.1, 1.9, 2.3), (65.9, 10.4, 19.7, 2.6, 1.4), (70.7, 3.6, 17.6, 5.6, 0.63), (64.8, 11.1, 21.2, 1.6, 0.85), (44.9, 10.6, 34.9, 0.9, 0.53)]

Hierarchical



link-method: COMPLETE
threshold: 40

3 clusters

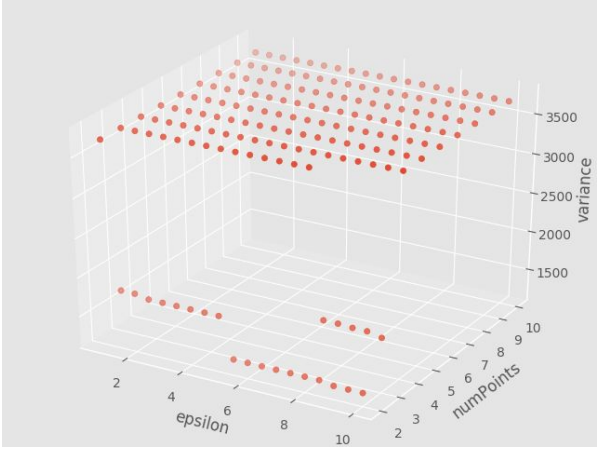
Cluster0
Center: (86.0625, 4.275, 4.175, 5.1188, 0.6356)
Max Dist: 13.2456
Min Dist: 2.2831
Avg Dist: 8.1717
Dist-Variance: 10.8026
Size: 16
Sample: [(81.6, 6.6, 5.9, 4.9, 0.93), (86.2, 3.0, 4.8, 5.3, 0.7), (81.9, 7.4, 7.2, 2.7, 0.85), (81.6, 10.1, 6.3, 4.4, 0.75), (88.4, 2.2, 2.7, 6.4, 0.18)]

Cluster1
Center: (69.4714, 9.5143, 16.2857, 2.9286, 1.3114)
Max Dist: 14.0114
Min Dist: 7.4886
Avg Dist: 10.6808
Dist-Variance: 4.902
Size: 7
Sample: [(70.7, 3.6, 17.6, 5.6, 0.63), (64.8, 10.7, 20.3, 2.5, 1.4), (65.9, 10.4, 19.7, 2.6, 1.4), (72.5, 9.2, 12.6, 3.3, 1.4), (71.3, 12.3, 13.1, 1.9, 2.3)]

Cluster2
Center: (45.65, 10.15, 38.45, 0.45, 0.69)
Max Dist: 5.36
Min Dist: 5.36
Avg Dist: 5.36
Dist-Variance: 0.0
Size: 2
Sample: [(46.4, 9.7, 42.0, 0.0, 0.85), (44.9, 10.6, 34.9, 0.9, 0.53)]

DBSCAN

Cluster0
Center: (132.1965, 3.7872, 2.2939)
Max Dist: 6.4969
Min Dist: 1.0801
Avg Dist: 3.2502
Dist-Variance: 4.0936
Size: 4
Sample: [(130.07, 4.79, 2.1893), (130.916, 4.659,



3 clusters found
Best Variance = 1265.3577666666667, achieved with e=9.0, numPoints=2

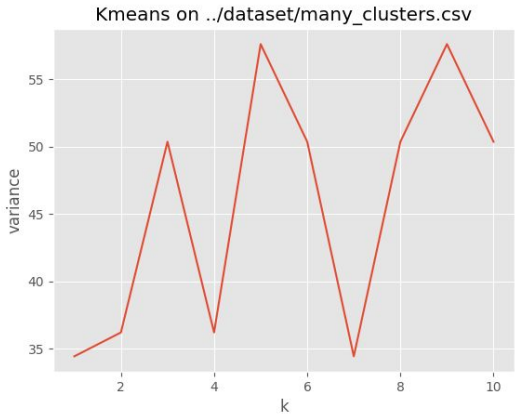
2.2562), (132.2, 4.7, 2.13), (135.6, 1.0, 2.6)]

Cluster1
Center: (144.0538, 7.1678, 2.835)
Max Dist: 204.5718
Min Dist: 10.099
Avg Dist: 82.9537
Dist-Variance: 3791.2618
Size: 12
Sample: [(89.9, 2.1, 3.35), (69.6, 4.7, 2.81), (338.333, 16.773, 2.7465), (339.625, 16.067, 2.7335), (164.1, 10.0, 1.93)]

Cluster2
Center: (79.9027, 5.2813, 2.4265)
Max Dist: 3.5948
Min Dist: 1.7966
Avg Dist: 2.3965
Dist-Variance: 0.7179
Size: 3
Sample: [(80.804, 4.622, 2.189), (80.804, 4.622, 2.1906), (78.1, 6.6, 2.9)]

Many Clusters

K-Means



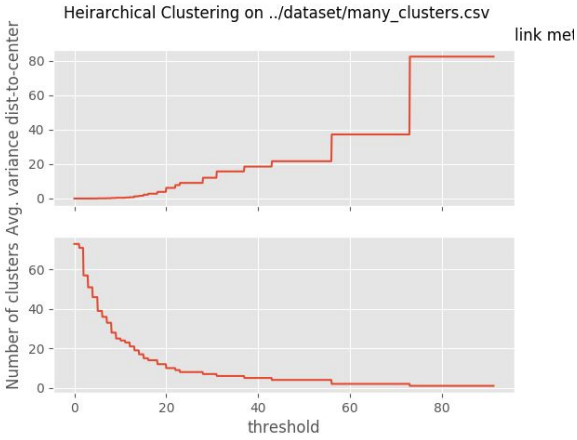
7 clusters

Cluster0
Center: (8.0769, 34.2308)
Max Dist: 12.1538
Min Dist: 0.8462
Avg Dist: 5.3373
Dist-Variance: 10.879
Size: 13
Sample: [(5, 33), (3, 38), (11, 28), (9, 32), (8, 37)]

Cluster1
Center: (42.7778, 10.3333)
Max Dist: 15.8889
Min Dist: 3.5556
Avg Dist: 6.6173
Dist-Variance: 13.3885
Size: 9
Sample: [(44, 15), (43, 6), (43, 5), (41, 6), (39, 16)]

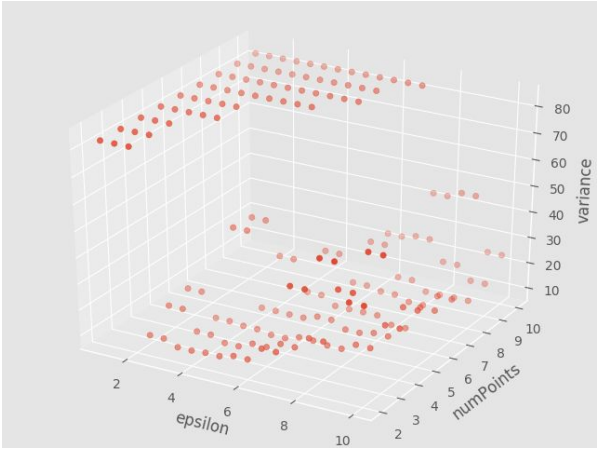
Cluster2
Center: (31.0, 42.25)
Max Dist: 15.25
Min Dist: 5.75
Avg Dist: 11.0
Dist-Variance: 6.8125
Size: 16
Sample: [(44, 40), (40, 42), (26, 37), (23, 44), (23, 42)]

Cluster3
Center: (21.3333, 8.3333)
Max Dist: 11.0
Min Dist: 5.0

	Avg Dist: 7.3333 Dist-Variance: 6.8889 Size: 3 Sample: [(27, 3), (21, 13), (16, 9)] Cluster4 Center: (37.7143, 28.1429) Max Dist: 14.1429 Min Dist: 2.8571 Avg Dist: 6.5918 Dist-Variance: 8.9529 Size: 14 Sample: [(39, 25), (35, 30), (31, 30), (38, 21), (34, 27)] Cluster5 Center: (26.1667, 24.0) Max Dist: 11.8333 Min Dist: 5.8333 Avg Dist: 8.8333 Dist-Variance: 3.5556 Size: 6 Sample: [(18, 26), (23, 19), (31, 17), (26, 33), (31, 21)] Cluster6 Center: (10.1667, 7.5) Max Dist: 11.3333 Min Dist: 1.3333 Avg Dist: 5.1667 Dist-Variance: 10.9722 Size: 12 Sample: [(5, 8), (9, 5), (13, 6), (12, 10), (11, 5)]
Hierarchical  **link-method: COMPLETE** **threshold: 50** **4 clusters**	Cluster0 Center: (33.5294, 34.3824) Max Dist: 24.1471 Min Dist: 5.8529 Avg Dist: 14.4429 Dist-Variance: 23.7955 Size: 34 Sample: [(34, 27), (31, 21), (38, 21), (41, 41), (23, 42)] Cluster1 Center: (8.9333, 32.6) Max Dist: 16.6667 Min Dist: 0.6667 Avg Dist: 7.4044 Dist-Variance: 19.2044 Size: 15 Sample: [(11, 28), (9, 23), (3, 38), (8, 35), (9, 32)] Cluster2 Center: (13.2, 7.7333) Max Dist: 21.0667 Min Dist: 1.9333 Avg Dist: 8.2133 Dist-Variance: 30.3278 Size: 15 Sample: [(11, 5), (16, 9), (10, 3), (23, 19), (5, 8)]

Cluster3
Center: (42.7778, 10.3333)
Max Dist: 15.8889
Min Dist: 3.5556
Avg Dist: 6.6173
Dist-Variance: 13.3885
Size: 9
Sample: [(44, 8), (42, 7), (39, 16), (41, 6), (43, 5)]

DBSCAN



Best Variance = 9.993928571428572, achieved with e=4.5, numPoints=3

Cluster0
Center: (23.6, 43.0)
Max Dist: 2.4
Min Dist: 0.4
Avg Dist: 1.52
Dist-Variance: 0.4096
Size: 5
Sample: [(24, 43), (23, 42), (23, 44), (26, 43), (22, 43)]

Cluster1
Center: (41.8333, 41.1667)
Max Dist: 3.3333
Min Dist: 1.0
Avg Dist: 2.3889
Dist-Variance: 0.571
Size: 6
Sample: [(42, 43), (44, 40), (44, 41), (41, 41), (40, 40)]

Cluster2
Center: (7.0, 36.0)
Max Dist: 2.0
Min Dist: 0.0
Avg Dist: 1.6
Dist-Variance: 0.64
Size: 5
Sample: [(8, 37), (8, 35), (6, 35), (7, 36), (6, 37)]

Cluster3
Center: (36.1429, 26.2857)
Max Dist: 4.8571
Min Dist: 0.8571
Avg Dist: 2.7755
Dist-Variance: 1.8301
Size: 7
Sample: [(39, 25), (34, 27), (36, 27), (37, 24), (35, 25)]

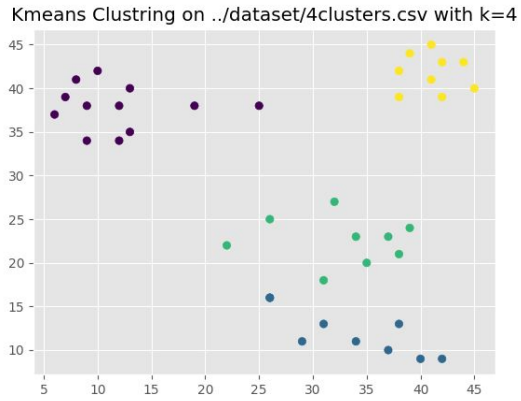
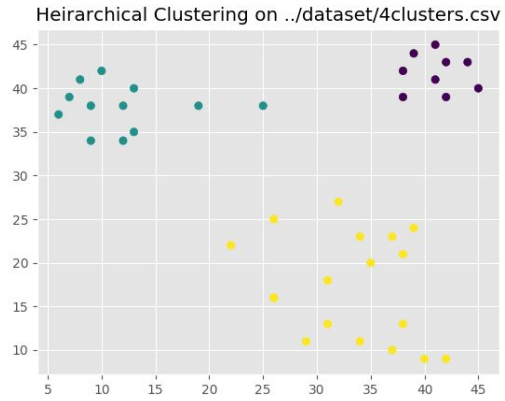
Cluster4
Center: (42.6, 6.4)
Max Dist: 3.0
Min Dist: 0.8
Avg Dist: 1.76
Dist-Variance: 0.5664
Size: 5
Sample: [(43, 5), (43, 6), (44, 8), (41, 6), (42, 7)]

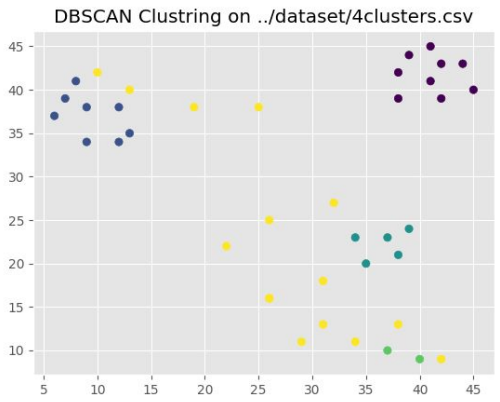
Cluster5
Center: (10.4286, 5.5714)
Max Dist: 3.0
Min Dist: 0.8571
Avg Dist: 2.1224
Dist-Variance: 0.6672

	Size: 7 Sample: [(9, 7), (10, 3), (10, 6), (11, 7), (11, 5)] Cluster6 Center: (24.0526, 25.5263) Max Dist: 36.5789 Min Dist: 6.4211 Avg Dist: 22.0803 Dist-Variance: 65.2732 Size: 38 Sample: [(9, 32), (26, 37), (26, 33), (11, 28), (7, 14)]

Visualization

4 Clusters

4 Clusters	
	

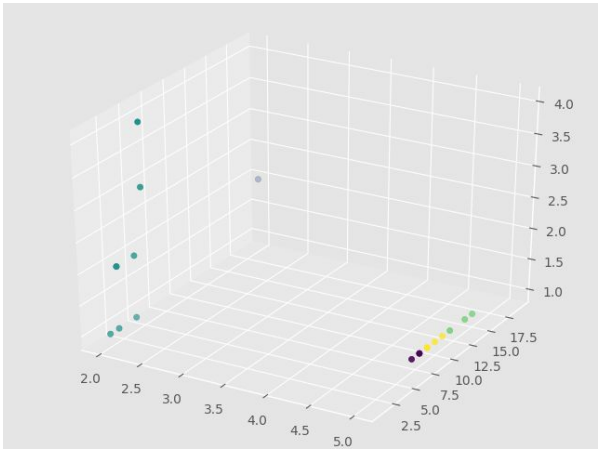


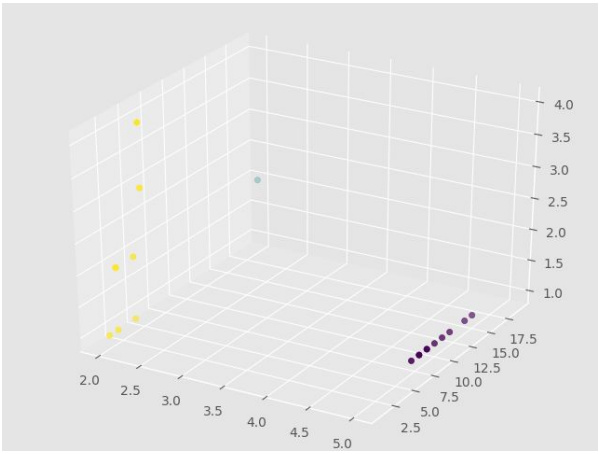
K-means:
K = 4 -> 4 clusters

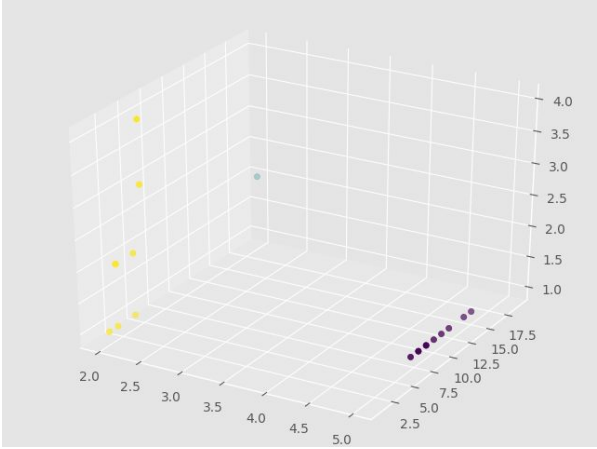
Hierarchical:
single, threshold=10.0 -> 3 clusters

DBscan:
e=8.5, numPoints=2 -> 5 clusters

Accidents1





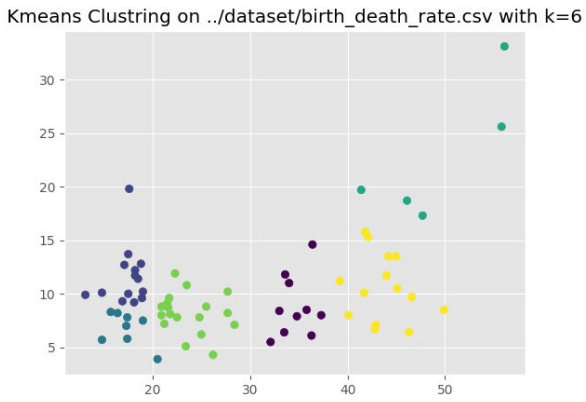


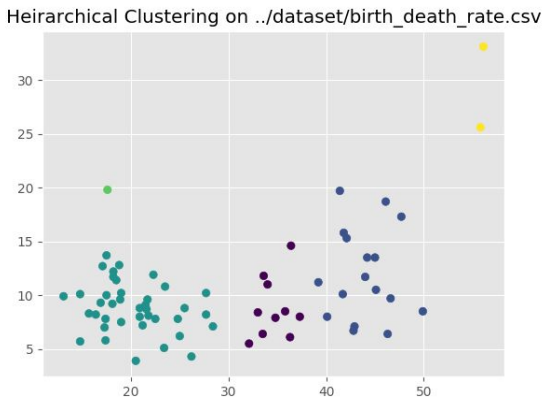
K-means:
K = 5 -> 5 clusters

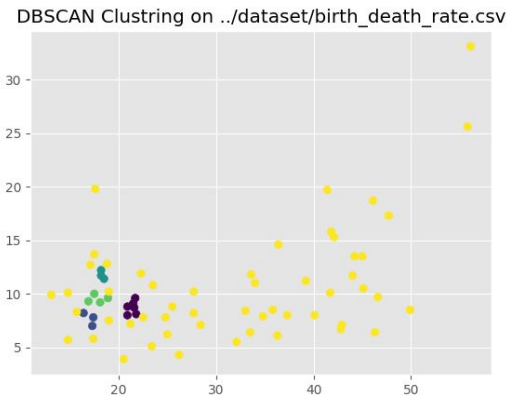
Hierarchical:
average, threshold=5.0 -> 5 clusters

DBscan:
e = 2.5, numpoints = 3 -> 3 clusters

Birth-Death Rate





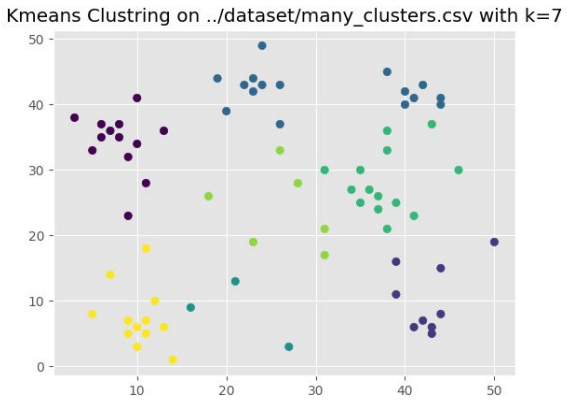


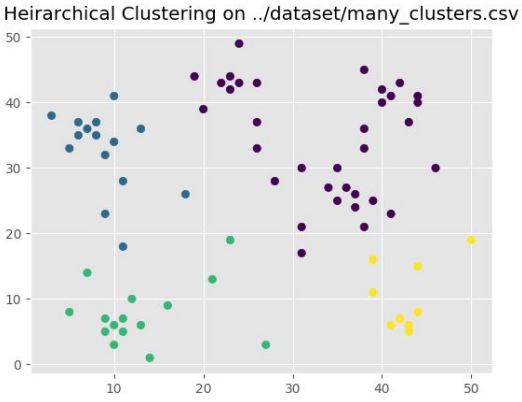
K-means:
K = 6 -> 6 clusters

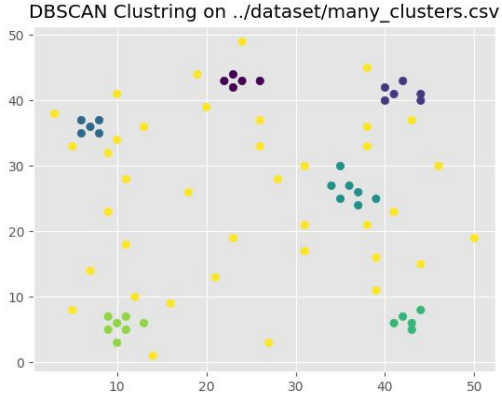
Hierarchical:
centroid, threshold=10.0 -> 5 clusters

DBscan:
e=1.5, numPoints=2 -> 5 clusters

Many Clusters





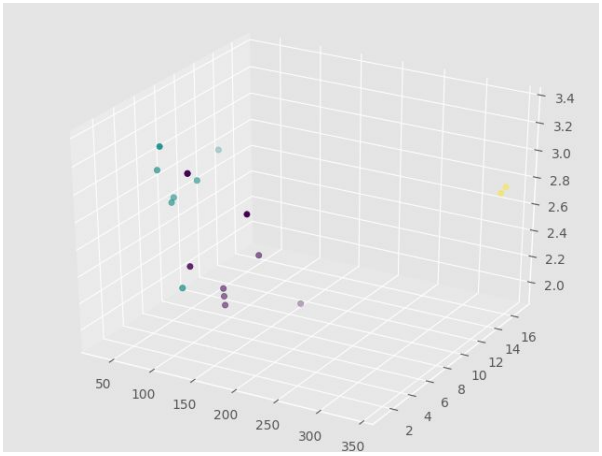


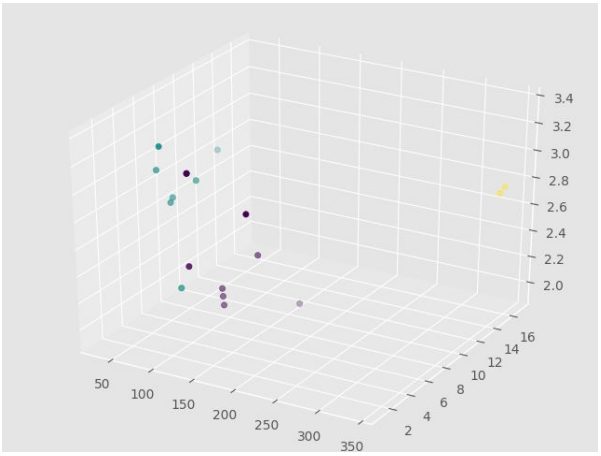
K-means:
K = 7 -> 7 clusters

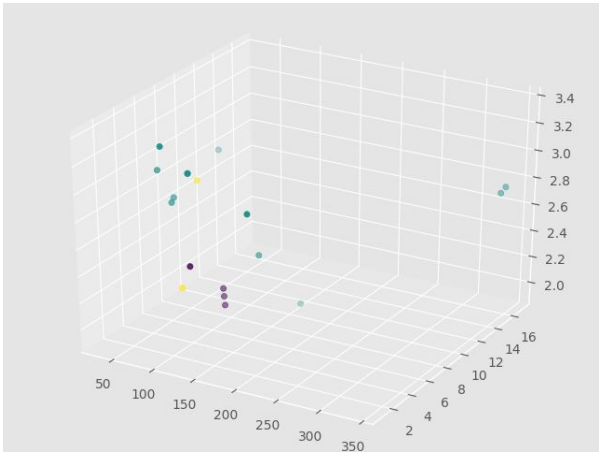
Hierarchical:
complete, threshold=50.0 -> 4 clusters

DBscan:
e=4.5, numPoints=3 -> 7 clusters

Planets







K-means:
K = 3 -> 3 clusters

Hierarchical:
average, threshold=100.0 -> 3 clusters

DBscan:
e=9.0, numPoints=2 -> 3 clusters

Discussion

Kmeans is definitely the least consistent of the clustering algorithms due to randomly picking centroids at the beginning. Given more time to optimize, we would devise a better, more consistent method of choosing initial centroids. We thought it was cool how DBSCAN can sometimes find clusters through noisy data, as in the case of many clusters. Kmeans can perform well if the initial clusters are chosen well.

We used variance to calculate how good the parameters passed to the functions were, but we think that it probably isn't the best for DBSCAN, since some clusters can be very spread out.

Analysis

Some datasets are difficult to tell how well a clustering algorithm did, because it's difficult for the human eye to see clusters. But for the ones that have 2 or 3 dimensions, we would say that the best algorithm for each is as follows:
***4 Clusters** - K-means / hierarchical*
***Accidents1** - DBSCAN / hierarchical*
***Birth-Death -** Kmeans*
***Many Clusters** - DBSCAN*
***Planets** - DBSCAN (tough call)*

All in all, K-Means tends to work better with wide clusters with not very perceptible differences between them. DBSCAN, as we know works for oddly-shaped clusters, but we also learned that it works well when there's a lot of noise, and the real clusters are more densely packed than the noise. Finally, hierarchical clustering seems to work well most of the time, but is less specialized than K-means or DBSCAN. It would be a good choice if you have no idea what the dataset is going to look like.