Chelsea Murray – 113481968 – Homework 4 – Data Mining Fall 2024

1. The chat transcript can be found here: https://chatgpt.com/share/6729a718-d024-800f-a2d2-42bd08fe6cf0

User: Assume I am looking for information at the intelligence level of a professional in the field of data mining. Tell me about three real-world companies/organizations from three distinct sectors—software technology, telecommunications, and transportation—that have implemented data clustering algorithms in their businesses. Include detailed descriptions of how those companies/organizations have applied data clustering algorithms in their operations.

ChatGPT: Certainly! Here's an overview of how companies in software technology, telecommunications, and transportation apply data clustering algorithms to enhance their operations.

1. Software Technology: Google

Application of Clustering: Google uses data clustering extensively across its services, particularly in organizing search results, personalizing recommendations, and improving user experience. One significant area is **image recognition and search**. Google's algorithms cluster images with similar visual content, allowing the company to categorize vast amounts of images, videos, and other visual data. This clustering supports Google Photos in organizing user photos by people, places, or activities.

How It Works: By applying clustering algorithms (such as k-means and hierarchical clustering), Google groups similar visual data based on pixel patterns, color distributions, or shape similarities. The clustered data helps the search algorithm quickly identify relevant images for users' search queries, improving the speed and accuracy of visual search results. Moreover, clustering aids in refining Google's recommendation systems for YouTube and Play Store by grouping users based on shared preferences, enabling the platform to suggest more personalized content.

2. Telecommunications: AT&T

Application of Clustering: AT&T uses data clustering to optimize its network performance and improve customer experience. Clustering algorithms help AT&T analyze patterns in user data across different geographic regions and identify areas experiencing high demand or recurring issues. By grouping these regions based on customer usage patterns and signal strength, AT&T can allocate network resources more effectively.

How It Works: Clustering algorithms such as DBSCAN (Density-Based Spatial Clustering of Applications with Noise) and k-means are employed to analyze location-based data, call drop frequencies, and network traffic. These clusters reveal zones of high usage or network congestion, helping AT&T to prioritize infrastructure upgrades or reconfigure network settings dynamically. For instance, AT&T might detect clusters of high data traffic in metropolitan areas during peak hours, prompting a temporary reallocation of bandwidth to prevent network slowdowns.

3. Transportation: Uber

Application of Clustering: Uber applies clustering algorithms for **demand forecasting** and **dynamic pricing**. By clustering locations based on demand patterns, Uber can identify "hot zones" where ride requests are likely to spike, especially during events, rush hours, or holidays. Clustering is also used to optimize driver availability in high-demand areas.

How It Works: Uber leverages k-means clustering and density-based clustering to group similar trip data by factors like pickup locations, drop-off points, and times of high demand. This data helps Uber forecast demand patterns in specific clusters, adjusting pricing through "surge pricing" when demand in these areas exceeds driver availability. Clustering also assists Uber in mapping out ideal "standby" zones for drivers to minimize wait times and optimize routes, enhancing both driver efficiency and customer satisfaction.

ChatGPT's response is vaguely plausible for Google, although it's much more likely that Google uses a convolutional neural network or other deep learning implementation given their available resources with regards to training data and computing power, and the shift in focus to AI as a corporate buzzword in recent years (as well as the rapidly-increasing presence of generative AI features in the public release of Google Photos on Pixel phones in particular). It's much more likely that Google would use clustering algorithms for the purposes of advertising and tailored search results based on similarities in user preferences and interests—in fact, it is possible to view and modify the interest "buckets" Google uses to identify your potential receptiveness to certain advertisements. This clustering would likely not be hierarchical in nature in a broad sense, since areas of interest and demographics may overlap, but could assume an hierarchical structure on a more specific level (ex. "interested in beauty products > interested in makeup > interested in Brand X's makeup products").

The response for AT&T makes more sense, as clustering "hotspots" for certain activities and issues would help the company with resource allocation based on actual activity. The emphasis on DBSCAN and k-means makes sense due to the geographic nature of the data with regards to the actual physical locations of AT&T customers, and the density of AT&T phones surrounding fixed tower locations. It would also be interesting to note if these results (particularly in identifying core points in DBSCAN and centroids in k-means) ever influence the company's decisions about geographical locations for new cell towers, as understanding the distribution of clusters of users could inform the future workload of a given tower.

ChatGPT's response about Uber also makes sense, for similar reasons to AT&T. Understanding the distribution of Uber traffic would indeed provide more accuracy in predicting availabilities, wait times, and where drivers would best be served to wait on ride assignments. I find it interesting that the model did not explicitly name DBSCAN this time, despite referencing density-based clustering, although that may have just been a linguistic quirk. Either way, it's a similar response to the AT&T response, and I almost wish the model had picked something like public transportation to discuss with regards to understanding, for example, the passenger load at certain times on certain bus lines to determine where extra buses or stops may be needed, or to evaluate whether bus routes are cost-effective at their current timing/workload, or to determine how to price advertising spaces inside the vehicles.

2. Given the table:

X	Y
4.38	-15.96
0.69	2.50
4.69	36.75
4.70	-7.83
4.38	34.43
0.47	-2.50
4.07	33.80
5.01	-16.59

2.1) Here is the application of bisecting K-means to cluster the data using K=4, number of trials=2, and choosing the cluster with the largest SSE for splitting at each step.

Cluster Step 1: Initialize cluster containing all points.

The first cluster contains every point in the table.

Cluster	Contents
A	{(4.38, -15.96), (0.69, 2.5), (4.69, 36.75), (4.7, -7.83), (4.38, 34.43), (0.47, -2.5), (4.07, 33.8), (5.01, -16.59)}

Cluster Step 2: Split into two on the "all points" cluster using K-means and find the bisection with the smallest SSE.

At the second step, the cluster is bisected twice (according to the number of trials specified). The first trial will use the first two points as centroids; the second trial will use the last two points as centroids.

Trial 1: K-Means Clustering around (4.38, -15.96) [Centroid 1] and (0.69, 2.5) [Centroid 2]

Trial 1 Round 1: (4.38, -15.96) [Centroid 1] and (0.69, 2.5) [Centroid 2]

Point	Distance to Centroid 1	Distance to Centroid 2	Closest Centroid
(4.38, -15.96)	0	$\text{sqrt}((0.69-4.38)^2+(2.5-(-15.96))^2)$ $=\text{sqrt}(13.6161+340.7716)$	Centroid 1

		=18.8252	
(0.69, 2.5)	sqrt((4.38-0.69)^2+(-15.96-2.5)^2) =sqrt(13.6161+340.7716) =18.8252	0	Centroid 2
(4.69, 36.75)	sqrt((4.38-4.69)^2+(-15.96-36.75)^2) =sqrt(0.0961+2778.3441) = 52.7109	sqrt((0.69-4.69)^2+(2.5-36.75)^2) =sqrt(16+1173.0625) = 34.4828	Centroid 2
(4.7, -7.83)	sqrt((4.38-4.7)^2+(-15.96-(-7.83))^2) =sqrt(0.1024+66.0969) = 8.1363	sqrt((0.69-4.7)^2+(2.5-(-7.83))^2) =sqrt(16.0801+106.7089) = 11.0810	Centroid 1
(4.38, 34.43)	sqrt((4.38-4.38)^2+(-15.96-34.43)^2) =sqrt(0+2539.1521) = 50.39	sqrt((0.69-4.38)^2+(2.5-34.43)^2) =sqrt(13.6161+1019.5249) = 32.1425	Centroid 2
(0.47, -2.5)	sqrt((4.38-0.47)^2+(-15.96-(-2.5))^2) =sqrt(15.2881+181.1716) = 14.0164	sqrt((0.69-0.47)^2+(2.5-(-2.5))^2) =sqrt(0.0484+25) = 5.0048	Centroid 2
(4.07, 33.8)	sqrt((4.38-4.07)^2+(-15.96-33.8)^2) =sqrt(0.0961+2476.0576) = 49.7610	sqrt((0.69-4.07)^2+(2.5-33.8)^2) =sqrt(11.4244+979.69) = 31.4820	Centroid 2
(5.01, -16.59)	sqrt((4.38-5.01)^2+(-15.96-(-16.59))^2) =sqrt(0.3969+0.3969) = 0.8910	sqrt((0.69-5.01)^2+(2.5-(-16.59))^2) =sqrt(18.6624+364.4281) = 19.5727	Centroid 1

New centroids = (mean(x), mean(y)) for each set: (4.6967, -13.46) and (2.86, 20.996)

Trial 1 Round 2: (4.6967, -13.46) [Centroid 1] and (2.86, 20.996) [Centroid 2]

Point	**Distance to Centroid 1**	**Distance to Centroid 2**	**Closest Centroid**
(4.38, -15.96)	sqrt((4.6967-4.38)^2+(-13.46-(-15.96))^2) =sqrt(0.1003+6.25) = 2.5200	sqrt((2.86-4.38)^2+(20.996-(-15.96))^2) =sqrt(2.3104+ 1365.7459) = 36.9872	Centroid 1 (unchanged)
(0.69, 2.5)	sqrt((4.6967-0.69)^2+(-13.46-2.5)^2) =sqrt(16.0536+254.7216) = 16.4552	sqrt((2.86-0.69)^2+(20.996-2.5)^2) =sqrt(4.7089+342.1020) = 18.6229	Centroid 1 (changed)
(4.69, 36.75)	sqrt((4.6967-4.69)^2+(-13.46-36.75)^2) =sqrt(0.00004+2521.0441) = 50.2100	sqrt((2.86-4.69)^2+(20.996-36.75)^2) =sqrt(3.3489+248.1885) = 15.8599	Centroid 2 (unchanged)
(4.7, -7.83)	sqrt((4.6967-4.7)^2+(-13.46-(-7.83))^2) =sqrt(0.00001+31.6969) = 5.6300	sqrt((2.86-4.7)^2+(20.996-(-7.83))^2) =sqrt(3.3856+830.9383) = 28.8847	Centroid 1 (unchanged)
(4.38, 34.43)	sqrt((4.6967-4.38)^2+(-13.46-34.43)^2) =sqrt(0.1003+2293.4521) = 47.8910	sqrt((2.86-4.38)^2+(20.996-34.43)^2) =sqrt(2.3104+180.4724) = 13.5197	Centroid 2 (unchanged)
(0.47, -2.5)	sqrt((4.6967-0.47)^2+(-13.46-(-2.5))^2) =sqrt(17.8650+120.1216) = 11.7468	sqrt((2.86-0.47)^2+(20.996-(-2.5))^2) =sqrt(5.7121+552.0620) = 23.6172	Centroid 1 (changed)
(4.07, 33.8)	sqrt((4.6967-4.07)^2+(-13.46-33.8)^2)	sqrt((2.86-4.07)^2+(20.996-33.8)^2)	Centroid 2

	=sqrt(0.3928+2233.5076) = 47.2642	=sqrt(1.4641+163.9424) = 12.8610	(unchanged)
(5.01, -16.59)	sqrt((4.6967-5.01)^2+(-13.46-(-16.59))^2) =sqrt(0.0982+9.7969) = 3.1456	sqrt((2.86-5.01)^2+(20.996-(-16.59))^2) =sqrt(4.6225+1412.7074) = 37.6474	Centroid 1 (unchanged)

New centroids = (mean(x), mean(y)) for each set: (3.05, -8.076) and (4.38, 34.9933)

Trial 1 Round 3: (3.05, -8.076) [Centroid 1] and (4.38, 34.9933) [Centroid 2]

Point	**Distance to Centroid 1**	**Distance to Centroid 2**	**Closest Centroid**
(4.38, -15.96)	sqrt((3.05-4.38)^2+(-8.076-(-15.96))^2) =sqrt(1.7689+62.1575) = 7.9954	sqrt((4.38-4.38)^2+(34.9933-(-15.96))^2) =sqrt(0+2596.2388) = 50.9533	Centroid 1 (unchanged)
(0.69, 2.5)	sqrt((3.05-0.69)^2+(-8.076-2.5)^2) =sqrt(5.5696+111.8518) = 10.8361	sqrt((4.38-0.69)^2+(34.9933-2.5)^2) =sqrt(13.6161+1055.8145) = 32.7022	Centroid 1 (unchanged)
(4.69, 36.75)	sqrt((3.05-4.69)^2+(-8.076-36.75)^2) =sqrt(2.6896+2009.3703) = 44.8560	sqrt((4.38-4.69)^2+(34.9933-36.75)^2) =sqrt(0.0961+3.0860) = 1.7838	Centroid 2 (unchanged)
(4.7, -7.83)	sqrt((3.05-4.7)^2+(-8.076-(-7.83))^2) =sqrt(2.7225+0.0605) = 1.6682	sqrt((4.38-4.7)^2+(34.9933-(-7.83))^2) =sqrt(0.1024+1833.8350) = 42.8245	Centroid 1 (unchanged)
(4.38, 34.43)	sqrt((3.05-4.38)^2+(-8.076-34.43)^2) =sqrt(1.7689+1806.7600) = 42.5268	sqrt((4.38-4.38)^2+(34.9933-34.43)^2) =sqrt(0+0.3173) = 0.5633	Centroid 2 (unchanged)
(0.47, -2.5)	sqrt((3.05-0.47)^2+(-8.076-(-2.5))^2) =sqrt(6.6564+31.0918) = 6.1440	sqrt((4.38-0.47)^2+(34.9933-(-2.5))^2) =sqrt(15.2881+1405.7475) = 37.6966	Centroid 1 (unchanged)
(4.07, 33.8)	sqrt((3.05-4.07)^2+(-8.076-33.8)^2) =sqrt(1.0404+1753.5994) = 41.8884	sqrt((4.38-4.07)^2+(34.9933-33.8)^2) =sqrt(0.0961+1.4240) = 1.5201	Centroid 2 (unchanged)
(5.01, -16.59)	sqrt((3.05-5.01)^2+(-8.076-(-16.59))^2) =sqrt(3.8416+72.4882) = 8.7367	sqrt((4.38-5.01)^2+(34.9933-(-16.59))^2) =sqrt(0.3969+2660.8368) = 51.5871	Centroid 1 (unchanged)

Since none of the points changed clusters, the centroids do not adjust and this is the final split. SSE can be found by squaring the Euclidean distances from each point to the respective centroid and then adding them together, resulting in the following bisection:

Cluster	**Centroid**	**Contents**	**SSE**
AA	(3.05, -8.076)	{(4.38, -15.96), (0.69, 2.5), (4.7, -7.83), (0.47, -2.5), (5.01, -16.59)}	298.2090
AB	(4.38, 34.9933)	{(4.69, 36.75), (4.38, 34.43), (4.07, 33.8)}	5.8100
		Total SSE	304.0190

SSE calculations

Cluster AA: $7.9954^2+10.8361^2+1.6682^2+6.1440^2+8.7367^2=298.2090$
Cluster AB: $1.7838^2+0.5633^2+1.5201^2=5.8100$

Trial 2: K-Means Clustering around (4.07, 33.8) [Centroid 1] and (5.01, -16.59) [Centroid 2]

Trial 2 Round 1: (4.07, 33.8) [Centroid 1] and (5.01, -16.59) [Centroid 2]

Point	**Distance to Centroid 1**	**Distance to Centroid 2**	**Closest Centroid**
(4.38, -15.96)	$\text{sqrt}((4.07-4.38)^2+(33.8-(-15.96))^2)$ =sqrt(0.0961+2476.0576) = 49.7610	$\text{sqrt}((5.01-4.38)^2+(-16.59-(-15.96))^2)$ =sqrt(0.3969+0.3969) = 0.8910	Centroid 2
(0.69, 2.5)	$\text{sqrt}((4.07-0.69)^2+(33.8-2.5)^2)$ =sqrt(11.4244+979.69) = 31.4820	$\text{sqrt}((5.01-0.69)^2+(-16.59-2.5)^2)$ =sqrt(18.6624+364.4281) = 19.5727	Centroid 2
(4.69, 36.75)	$\text{sqrt}((4.07-4.69)^2+(33.8-36.75)^2)$ =sqrt(0.3844+8.7025) = 3.0144	$\text{sqrt}((5.01-4.69)^2+(-16.59-36.75)^2)$ =sqrt(0.1024+2879.3956) = 53.6610	Centroid 1
(4.7, -7.83)	$\text{sqrt}((4.07-4.7)^2+(33.8-(-7.83))^2)$ =sqrt(0.3969+1733.0569) = 41.6348	$\text{sqrt}((5.01-4.7)^2+(-16.59-(-7.83))^2)$ =sqrt(0.0961+596.3364) = 24.4220	Centroid 2
(4.38, 34.43)	$\text{sqrt}((4.07-4.38)^2+(33.8-34.43)^2)$ =sqrt(0.0961+0.3969) = 0.7021	$\text{sqrt}((5.01-4.38)^2+(-16.59-34.43)^2)$ =sqrt(0.3969+2603.0404) = 51.0239	Centroid 1
(0.47, -2.5)	$\text{sqrt}((4.07-0.47)^2+(33.8-(-2.5))^2)$ =sqrt(12.96+1317.69) = 36.4781	$\text{sqrt}((5.01-0.47)^2+(-16.59-(-2.5))^2)$ =sqrt(20.6116+198.5281) = 14.8033	Centroid 2
(4.07, 33.8)	0	$\text{sqrt}((5.01-4.07)^2+(-16.59-33.8)^2)$ =sqrt(0.8836+2539.1521) = 50.3988	Centroid 1
(5.01, -16.59)	$\text{sqrt}((4.07-5.01)^2+(33.8-(-16.59))^2)$ =sqrt(0.8836+2539.1521) = 50.3988	0	Centroid 2

In Trial 2, the same cluster pattern that Trial 1 ended on surfaces immediately. Therefore, the new centroids and resulting clusters and SSE are already known, as stated above, although Centroids 1 and 2 are swapped. For the sake of showing work, I will include this conclusion:

New centroids = (mean(x), mean(y)) for each set: (4.38, 34.9933) and (3.05, -8.076)

Trial 2 Round 2: (4.38, 34.9933) [Centroid 1] and (3.05, -8.076) [Centroid 2]

Point	**Distance to Centroid 1**	**Distance to Centroid 2**	**Closest Centroid**
(4.38, -15.96)	$\text{sqrt}((4.38-4.38)^2+(34.9933-(-15.96))^2)$ =sqrt(0+2596.2388) = 50.9533	$\text{sqrt}((3.05-4.38)^2+(-8.076-(-15.96))^2)$ =sqrt(1.7689+62.1575) = 7.9954	Centroid 2 (unchanged)

(0.69, 2.5)	sqrt((4.38-0.69)^2+(34.9933-2.5)^2) =sqrt(13.6161+1055.8145) = 32.7022	sqrt((3.05-0.69)^2+(-8.076-2.5)^2) =sqrt(5.5696+111.8518) = 10.8361	Centroid 2 (unchanged)
(4.69, 36.75)	sqrt((4.38-4.69)^2+(34.9933-36.75)^2) =sqrt(0.0961+3.0860) = 1.7838	sqrt((3.05-4.69)^2+(-8.076-36.75)^2) =sqrt(2.6896+2009.3703) = 44.8560	Centroid 1 (unchanged)
(4.7, -7.83)	sqrt((4.38-4.7)^2+(34.9933-(-7.83))^2) =sqrt(0.1024+1833.8350) = 42.8245	sqrt((3.05-4.7)^2+(-8.076-(-7.83))^2) =sqrt(2.7225+0.0605) = 1.6682	Centroid 2 (unchanged)
(4.38, 34.43)	sqrt((4.38-4.38)^2+(34.9933-34.43)^2) =sqrt(0+0.3173) = 0.5633	sqrt((3.05-4.38)^2+(-8.076-34.43)^2) =sqrt(1.7689+1806.7600) = 42.5268	Centroid 1 (unchanged)
(0.47, -2.5)	sqrt((4.38-0.47)^2+(34.9933-(-2.5))^2) =sqrt(15.2881+1405.7475) = 37.6966	sqrt((3.05-0.47)^2+(-8.076-(-2.5))^2) =sqrt(6.6564+31.0918) = 6.1440	Centroid 2 (unchanged)
(4.07, 33.8)	sqrt((4.38-4.07)^2+(34.9933-33.8)^2) =sqrt(0.0961+1.4240) = 1.5201	sqrt((3.05-4.07)^2+(-8.076-33.8)^2) =sqrt(1.0404+1753.5994) = 41.8884	Centroid 1 (unchanged)
(5.01, -16.59)	sqrt((4.38-5.01)^2+(34.9933-(-16.59))^2) =sqrt(0.3969+2660.8368) = 51.5871	sqrt((3.05-5.01)^2+(-8.076-(-16.59))^2) =sqrt(3.8416+72.4882) = 8.7367	Centroid 2 (unchanged)

Since none of the points changed clusters, this bisection concludes in the following state:

Cluster	Centroid	Contents	SSE
AC	(4.38, 34.9933)	{(4.69, 36.75), (4.38, 34.43), (4.07, 33.8)}	5.8100
AD	(3.05, -8.076)	{(4.38, -15.96), (0.69, 2.5), (4.7, -7.83), (0.47, -2.5), (5.01, -16.59)}	298.2090
		Total SSE	304.0190

SSE calculations

Cluster AC: $1.7838^2+0.5633^2+1.5201^2=5.8100$
Cluster AD: $7.9954^2+10.8361^2+1.6682^2+6.1440^2+8.7367^2=298.2090$

Selecting a bisection

With identical SSEs and clusters, the difference between each set of clusters generated is non-existent apart from name. Arbitrarily, we'll pick the first one generated.

Cluster Step 3: Split into two on the highest-SSE cluster, using two trials to find the lowest-SSE bisection.

The current set of clusters is as follows:

Cluster	Centroid	Contents	SSE
AA	(3.05, -8.076)	{(4.38, -15.96), (0.69, 2.5), (4.7, -7.83), (0.47, -2.5),	298.2090

		(5.01, -16.59)}	
AB	(4.38, 34.9933)	{(4.69, 36.75), (4.38, 34.43), (4.07, 33.8)}	5.8100
		Total SSE	304.0190

Since Cluster AA has the highest SSE, it will be bisected using 2 trials of K-means to generate the next pair of clusters. These trials will again use the first 2 and last 2 points as centroids.

Trial 1: K-Means Clustering around (4.38, -15.96) [Centroid 1] and (0.69, 2.5) [Centroid 2]

Trial 1 Round 1: (4.38, -15.96) [Centroid 1] and (0.69, 2.5) [Centroid 2]

Point	Distance to Centroid 1	Distance to Centroid 2	Closest Centroid
(4.38, -15.96)	0	sqrt($(0.69-4.38)^2+(2.5-(-15.96))^2$) =sqrt(13.6161+340.7716) =18.8252	Centroid 1
(0.69, 2.5)	sqrt($(4.38-0.69)^2+(-15.96-2.5)^2$) =sqrt(13.6161+340.7716) =18.8252	0	Centroid 2
(4.7, -7.83)	sqrt($(4.38-4.7)^2+(-15.96-(-7.83))^2$) =sqrt(0.1024+66.0969) = 8.1363	sqrt($(0.69-4.7)^2+(2.5-(-7.83))^2$) =sqrt(16.0801+106.7089) = 11.0810	Centroid 1
(0.47, -2.5)	sqrt($(4.38-0.47)^2+(-15.96-(-2.5))^2$) =sqrt(15.2881+181.1716) = 14.0164	sqrt($(0.69-0.47)^2+(2.5-(-2.5))^2$) =sqrt(0.0484+25) = 5.0048	Centroid 2
(5.01, -16.59)	sqrt($(4.38-5.01)^2+(-15.96-(-16.59))^2$) =sqrt(0.3969+0.3969) = 0.8910	sqrt($(0.69-5.01)^2+(2.5-(-16.59))^2$) =sqrt(18.6624+364.4281) = 19.5727	Centroid 1

New centroids = (mean(x), mean(y)) for each set: (4.6967, -13.46) and (0.58, 0)

Trial 1 Round 2: (4.6967, -13.46) [Centroid 1] and (0.58, 0) [Centroid 2]

Point	Distance to Centroid 1	Distance to Centroid 2	Closest Centroid
(4.38, -15.96)	sqrt($(4.6967-4.38)^2+(-13.46-(-15.96))^2$) =sqrt(0.1003+6.25) = 2.5200	sqrt($(0.58-4.38)^2+(0-(-15.96))^2$) =sqrt(14.44+254.7216) = 16.4061	Centroid 1 (unchanged)
(0.69, 2.5)	sqrt($(4.6967-0.69)^2+(-13.46-2.5)^2$) =sqrt(16.0536+254.7216) = 16.4552	sqrt($(0.58-0.69)^2+(0-2.5)^2$) =sqrt(0.0121+6.25) = 2.5024	Centroid 2 (unchanged)
(4.7, -7.83)	sqrt($(4.6967-4.7)^2+(-13.46-(-7.83))^2$) =sqrt(0.00001+31.6969) = 5.6300	sqrt($(0.58-4.7)^2+(0-(-7.83))^2$) =sqrt(16.9744+61.3089) = 8.8478	Centroid 1 (unchanged)
(0.47, -2.5)	sqrt($(4.6967-0.47)^2+(-13.46-(-2.5))^2$) =sqrt(17.8650+120.1216) = 11.7468	sqrt($(0.58-0.47)^2+(0-(-2.5))^2$) =sqrt(0.0121+6.25) = 2.5024	Centroid 2 (unchanged)
(5.01, -16.59)	sqrt($(4.6967-5.01)^2+(-13.46-(-16.59))^2$) =sqrt(0.0982+9.7969)	sqrt($(0.58-5.01)^2+(0-(-16.59))^2$) =sqrt(19.6249+275.2281)	Centroid 1 (unchanged)

	= 3.1456	= 17.1713	

Since none of the points changed clusters, the centroids do not adjust and this is the final split.

Cluster	**Centroid**	**Contents**	**SSE**
AAA	(4.6967, -13.46)	{(4.38, -15.96), (4.7, -7.83), (5.01, -16.59)}	47.9421
AAB	(0.58, 0)	{(0.69, 2.5), (0.47, -2.5)}	12.5240
		Total SSE	60.4661

SSE calculations

Cluster AAA: $2.5200^2+5.6300^2+3.1456^2=47.9421$
Cluster AAB: $2.5024^2+2.5024^2=12.5240$

Trial 2: K-Means Clustering around (0.47, -2.5) [Centroid 1] and (5.01, -16.59) [Centroid 2]

Trial 2 Round 1: (0.47, -2.5) [Centroid 1] and (5.01, -16.59) [Centroid 2]

Point	**Distance to Centroid 1**	**Distance to Centroid 2**	**Closest Centroid**
(4.38, -15.96)	$\text{sqrt}((0.47-4.38)^2+(-2.5-(-15.96))^2)$ =sqrt(15.2881+181.1716) = 14.0164	$\text{sqrt}((5.01-4.38)^2+(-16.59-(-15.96))^2)$ =sqrt(0.3969+0.3969) = 0.8910	Centroid 2
(0.69, 2.5)	$\text{sqrt}((0.47-0.69)^2+(-2.5-2.5)^2)$ =sqrt(0.0484+25) = 5.0048	$\text{sqrt}((5.01-0.69)^2+(-16.59-2.5)^2)$ =sqrt(18.6624+364.4281) = 19.5727	Centroid 1
(4.7, -7.83)	$\text{sqrt}((0.47-4.7)^2+(-2.5-(-7.83))^2)$ =sqrt(17.8929+28.4089) = 6.8045	$\text{sqrt}((5.01-4.7)^2+(-16.59-(-7.83))^2)$ =sqrt(0.0961+596.3364) = 24.4220	Centroid 1
(0.47, -2.5)	0	$\text{sqrt}((5.01-0.47)^2+(-16.59-(-2.5))^2)$ =sqrt(20.6116+198.5281) = 14.8033	Centroid 1
(5.01, -16.59)	$\text{sqrt}((0.47-5.01)^2+(-2.5-(-16.59))^2)$ =sqrt(20.6116+198.5281) = 14.8034	0	Centroid 2

New centroids = (mean(x), mean(y)) for each set: (1.9533, -2.61) and (4.695, -16.275)

Trial 2 Round 2: (1.9533, -2.61) [Centroid 1] and (4.695, -16.275) [Centroid 2]

Point	**Distance to Centroid 1**	**Distance to Centroid 2**	**Closest Centroid**
(4.38, -15.96)	$\text{sqrt}((1.9533-4.38)^2+(-2.61-(-15.96))^2)$ =sqrt(5.8889+178.2225) = 13.5688	$\text{sqrt}((4.695-4.38)^2+(-16.275-(-15.96))^2)$ =sqrt(0.0992+0.0992) = 0.4454	Centroid 2 (unchanged)
(0.69, 2.5)	$\text{sqrt}((1.9533-0.69)^2+(-2.61-2.5)^2)$ =sqrt(1.5959+26.1121) = 5.2638	$\text{sqrt}((4.695-0.69)^2+(-16.275-2.5)^2)$ =sqrt(16.0400+352.5006) = 19.1974	Centroid 1 (unchanged)
(4.7, -7.83)	$\text{sqrt}((1.9533-4.7)^2+(-2.61-(-7.83))^2)$	$\text{sqrt}((4.695-4.7)^2+(-16.275-(-7.83))^2)$	Centroid 1

	=sqrt(7.5444+27.2484) = 5.8985	=sqrt(0.00003+71.3180) = 8.4450	(unchanged)
(0.47, -2.5)	sqrt($(1.9533-0.47)^2+(-2.61-(-2.5))^2$) =sqrt(2.2002+0.0121) =1.4874	sqrt($(4.695-0.47)^2+(-16.275-(-2.5))^2$) =sqrt(17.8506+189.7506) = 14.4084	Centroid 1 (unchanged)
(5.01, -16.59)	sqrt($(1.9533-5.01)^2+(-2.61-(-16.59))^2$) =sqrt(9.3434+195.4404) = 14.3103	sqrt($(4.695-5.01)^2+(-16.275-(-16.59))^2$) =sqrt(0.0992+0.0992) = 0.4454	Centroid 2 (unchanged)

Since none of the points changed clusters, the centroids do not adjust and this is the final split.

Cluster	Centroid	Contents	SSE
AAC	(1.9533, -2.61)	{(0.69, 2.5), (4.7, -7.83), (0.47, -2.5)}	64.7123
AAD	(4.695, -16.275)	{(4.38, -15.96), (5.01, -16.59)}	0.3968
		Total SSE	65.1091

SSE calculations

Cluster AAC: $5.2638^2+5.8985^2+1.4874^2= 64.7123$
Cluster AAD: $0.4454^2+0.4454^2= 0.3968$

Selecting a bisection

Comparing the two trial results:

Cluster	Centroid	Contents	SSE
AAA	(4.6967, -13.46)	{(4.38, -15.96), (4.7, -7.83), (5.01, -16.59)}	47.9421
AAB	(0.58, 0)	{(0.69, 2.5), (0.47, -2.5)}	12.5240
		Total SSE	60.4661
AAC	(1.9533, -2.61)	{(0.69, 2.5), (4.7, -7.83), (0.47, -2.5)}	64.7123
AAD	(4.695, -16.275)	{(4.38, -15.96), (5.01, -16.59)}	0.3968
		Total SSE	65.1091

With a lower SSE, the first trial is the more optimal bisection, and will be retained as the cluster set generated for this iteration.

Cluster Step 3: Split into two on the highest-SSE cluster, using two trials to find the lowest-SSE bisection.

The current set of clusters is as follows:

Cluster	Centroid	Contents	SSE
AAA	(4.6967, -13.46)	{(4.38, -15.96), (4.7, -7.83), (5.01, -16.59)}	47.9421
AAB	(0.58, 0)	{(0.69, 2.5), (0.47, -2.5)}	12.5240
AB	(4.38, 34.9933)	{(4.69, 36.75), (4.38, 34.43), (4.07, 33.8)}	5.8100
		Total SSE	66.2761

Since Cluster AAA has the highest SSE, it will be bisected using 2 trials of K-means to generate the next pair of clusters. These trials will again use the first 2 and last 2 points as centroids (since this set only contains 3 points, the trials will have 1 centroid in common).

Trial 1: K-Means Clustering around (4.38, -15.96) [Centroid 1] and (4.7, -7.83) [Centroid 2]

Trial 1 Round 1: (4.38, -15.96) [Centroid 1] and (4.7, -7.83) [Centroid 2]

Point	Distance to Centroid 1	Distance to Centroid 2	Closest Centroid
(4.38, -15.96)	0	sqrt((4.7-4.38)^2+(-7.83-(-15.96))^2) =sqrt(0.1024+66.0969) = 8.1363	Centroid 1
(4.7, -7.83)	sqrt((4.38-4.7)^2+(-15.96-(-7.83))^2) =sqrt(0.1024+66.0969) = 8.1363	0	Centroid 2
(5.01, -16.59)	sqrt((4.38-5.01)^2+(-15.96-(-16.59))^2) =sqrt(0.3969+0.3969) = 0.8910	sqrt((4.7-5.01)^2+(-7.83-(-16.59))^2) =sqrt(0.0961+76.7376) = 8.7655	Centroid 1

New centroids = (mean(x), mean(y)) for each set: (4.695, -16.275) and (4.7, -7.83)

Trial 1 Round 2: (4.695, -16.275) [Centroid 1] and (4.7, -7.83) [Centroid 2]

Point	Distance to Centroid 1	Distance to Centroid 2	Closest Centroid
(4.38, -15.96)	sqrt((4.695-4.38)^2+(-16.275-(-15.96))^2) =sqrt(0.0992+0.0992) = 0.4454	sqrt((4.7-4.38)^2+(-7.83-(-15.96))^2) =sqrt(0.1024+66.0969) = 8.1363	Centroid 1 (unchanged)
(4.7, -7.83)	sqrt((4.695-4.7)^2+(-16.275-(-7.83))^2) =sqrt(0.00003+71.3180) = 8.4450	0	Centroid 2 (unchanged)
(5.01, -16.59)	sqrt((4.695-5.01)^2+(-16.275-(-16.59))^2) =sqrt(0.0992+0.0992) = 0.4454	sqrt((4.7-5.01)^2+(-7.83-(-16.59))^2) =sqrt(0.0961+76.7376) = 8.7655	Centroid 1 (unchanged)

Since none of the points changed clusters, the centroids do not adjust and this is the final split.

Cluster	Centroid	Contents	SSE
AAAA	(4.695, -16.275)	{(4.38, -15.96), (5.01, -16.59)}	0.3968
AAAB	(4.7, -7.83)	{(4.7, -7.83)}	0
		Total SSE	0.3968

SSE calculations

Cluster AAAA: $0.4454^2+0.4454^2= 0.3968$
Cluster AAAB: $0^2 = 0$

Trial 2: K-Means Clustering around (4.7, -7.83) [Centroid 1] and (5.01, -16.59) [Centroid 2]

Trial 2 Round 1: (4.7, -7.83) [Centroid 1] and (5.01, -16.59) [Centroid 2]

Point	Distance to Centroid 1	Distance to Centroid 2	Closest Centroid
(4.38, -15.96)	sqrt((4.7-4.38)^2+(-7.83-(-15.96))^2) =sqrt(0.1024+66.0969) = 8.1363	sqrt((5.01-4.38)^2+(-16.59-(-15.96))^2) =sqrt(0.3969+0.3969) = 0.8910	Centroid 2
(4.7, -7.83)	0	sqrt((5.01-4.7)^2+(-16.59-(-7.83))^2) =sqrt(0.0961+596.3364) = 24.4220	Centroid 1
(5.01, -16.59)	sqrt((4.7-5.01)^2+(-7.83-(-16.59))^2) =sqrt(0.0961+76.7376) = 8.7655	0	Centroid 2

Much like with the second trial for the first bisection, the centroid split is the same (just labeled differently), so the math is identical (but mirrored) for the final centroid outcome and SSE.

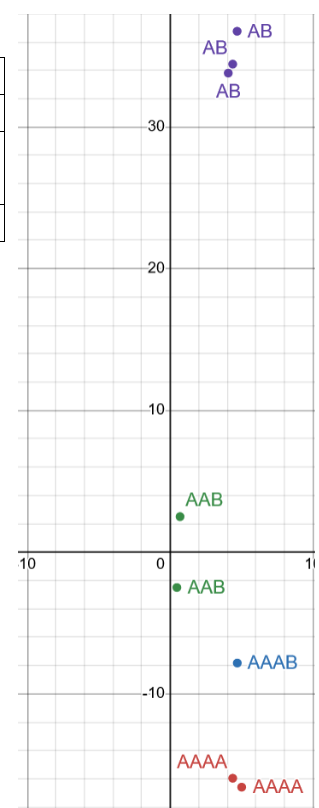
Trial 1 Round 2: (4.695, -16.275) [Centroid 1] and (4.7, -7.83) [Centroid 2]

Point	Distance to Centroid 1	Distance to Centroid 2	Closest Centroid
(4.38, -15.96)	sqrt((4.7-4.38)^2+(-7.83-(-15.96))^2) =sqrt(0.1024+66.0969) = 8.1363	sqrt((4.695-4.38)^2+(-16.275-(-15.96))^2) =sqrt(0.0992+0.0992) = 0.4454	Centroid 2 (unchanged)
(4.7, -7.83)	0	sqrt((4.695-4.7)^2+(-16.275-(-7.83))^2) =sqrt(0.00003+71.3180) = 8.4450	Centroid 1 (unchanged)
(5.01, -16.59)	sqrt((4.7-5.01)^2+(-7.83-(-16.59))^2) =sqrt(0.0961+76.7376) = 8.7655	sqrt((4.695-5.01)^2+(-16.275-(-16.59))^2) =sqrt(0.0992+0.0992) = 0.4454	Centroid 2 (unchanged)

Since none of the points changed clusters, the centroids do not adjust and this is the final split.

Cluster	Centroid	Contents	SSE
AAAC	(4.7, -7.83)	{(4.7, -7.83)}	0
AAAD	(4.695, -16.275)	{(4.38, -15.96), (5.01, -16.59)}	0.3968
		Total SSE	0.3968

Arbitrarily, with an identical SSE, the first bisection will be chosen and added to the cluster set.



Final Result

With 4 clusters generated, this is the final set of clusters:

Cluster	Centroid	Contents	SSE
AAAA	(4.695, -16.275)	{(4.38, -15.96), (5.01, -16.59)}	0.3968
AAAB	(4.7, -7.83)	{(4.7, -7.83)}	0
AAB	(0.58, 0)	{(0.69, 2.5), (0.47, -2.5)}	12.5240
AB	(4.38, 34.9933)	{(4.69, 36.75), (4.38, 34.43), (4.07, 33.8)}	5.8100
		Total SSE	18.7308

On the right is the cluster set on a plot generated using the Desmos graphing calculator. Predictably, the cluster with the highest SSE has the greatest distance between points.

2.2)

Step 1: Create the similarity matrix

Here is the similarity matrix for the points in the set, using the formula 1/(1+dist(a, b)) as the measure of similarity with dist referring to Euclidean distance. The points themselves are in color coded cells to indicate which set they belong to, and since detailed calculations were requested, the math is shown for every matrix entry (calculations on top "half" of matrix and values-only on bottom "half" for readability:

	(4.38, -15.96)	**(5.01, -16.59)**	**(4.7, -7.83)**	**(0.69, 2.5)**	**(0.47, -2.5)**	**(4.69, 36.75)**	**(4.38, 34.43)**	**(4.07, 33.8)**
(4.38, -15.96)	1	1/(1+sqrt((4.38-5.01)^2+(-15.96-(-16.59))^2)) =1/(1+sqrt((-0.63)^2+(0.63)^2)) =1/(1+sqrt(0.3969 +0.3969) =1/(1+sqrt(0.7938)) =1/(1+0.8910) =1/1.8910 =0.5288	1/(1+sqrt((4.38-4.7)^2+(-15.96-(-7.83))^2)) =1/(1+sqrt((-0.32)^2+(-8.13)^2)) =1/(1+sqrt(0.1024 +66.0969) =1/(1+sqrt(66.1993)) =1/(1+8.1363) =1/9.1363 =0.1095	1/(1+sqrt((4.38-0.69)^2+(-15.96-2.5)^2)) =1/(1+sqrt((3.69)^2+(-18.46)^2)) =1/(1+sqrt(13.6161 +340.7716)) =1/(1+sqrt(354.3877)) =1/(1+18.8252) =1/19.8252 =0.0504	1/(1+sqrt((4.38-0.47)^2+(-15.96-(-2.5))^2)) =1/(1+sqrt((3.91)^2+(-13.46)^2)) =1/(1+sqrt(15.2881 +181.1716) =1/(1+sqrt(196.4597)) =1/(1+14.0164) =1/15.0164 =0.0666	1/(1+sqrt((4.38-4.69)^2+(-15.96-36.75)^2)) =1/(1+sqrt((-0.31)^2+(-52.71)^2)) =1/(1+sqrt(0.0961 +2778.344) =1/(1+sqrt(2778.44)) =1/(1+52.7109) =1/53.7109 =0.0186	1/(1+sqrt((4.38-4.38)^2+(-15.96-34.43)^2)) =1/(1+sqrt((0)^2+(-50.39)^2)) =1/(1+sqrt(0 +2539.152) =1/(1+sqrt(2539.152)) =1/(1+50.39) =1/51.39 =0.0195	1/(1+sqrt((4.38-4.07)^2+(-15.96-33.8)^2)) =1/(1+sqrt((0.31)^2+(-49.76)^2)) =1/(1+sqrt(0.0961 +2476.058) =1/(1+sqrt(2476.154)) =1/(1+49.7610) =1/50.7610 =0.0197
(5.01, -16.59)	0.5288	1	1/(1+sqrt((5.01-4.7)^2+(-16.59-(-7.83))^2)) =1/(1+sqrt((0.31)^2+(-8.76)^2)) =1/(1+sqrt(0.0961 +76.7376) =1/(1+sqrt(76.8337)) =1/(1+8.7655) =1/9.7655 =0.1024	1/(1+sqrt((5.01-0.69)^2+(-16.59-2.5)^2)) =1/(1+sqrt((4.32)^2+(-19.09)^2)) =1/(1+sqrt(18.6624 +364.4281) =1/(1+sqrt(383.0905)) =1/(1+19.5727) =1/20.5727 =0.0486	1/(1+sqrt((5.01-0.47)^2+(-16.59-(-2.5))^2)) =1/(1+sqrt((4.54)^2+(-14.09)^2)) =1/(1+sqrt(20.6116 +198.5281) =1/(1+sqrt(219.1397)) =1/(1+14.8034) =1/15.8034 =0.0633	1/(1+sqrt((5.01-4.69)^2+(-16.59-36.75)^2)) =1/(1+sqrt((0.32)^2+(-53.34)^2)) =1/(1+sqrt(0.1024 +2845.1556) =1/(1+sqrt(2845.258)) =1/(1+53.3410) =1/54.3410 =0.0184	1/(1+sqrt((5.01-4.38)^2+(-16.59-34.43)^2)) =1/(1+sqrt((0.63)^2+(-51.02)^2)) =1/(1+sqrt(0.3969 +2603.04) =1/(1+sqrt(2603.4373)) =1/(1+51.0239) =1/52.0239 =0.0192	1/(1+sqrt((5.01-4.07)^2+(-16.59-33.8)^2)) =1/(1+sqrt((0.94)^2+(-50.39)^2)) =1/(1+sqrt(0.8836 +2539.1521) =1/(1+sqrt(2540.0357)) =1/(1+50.3988) =1/51.3988 =0.0195
(4.7, -7.83)	0.1095	0.1024	1	1/(1+sqrt((4.7-0.69)^2+(-7.83-2.5)^2)) =1/(1+sqrt((4.01)^2+(-10.33)^2)) =1/(1+sqrt(16.0801 +106.7089) =1/(1+sqrt(122.789)) =1/(1+11.0810) =1/12.0810	1/(1+sqrt((4.7-0.47)^2+(-7.83-(-2.5))^2)) =1/(1+sqrt((4.23)^2+(-5.33)^2)) =1/(1+sqrt(17.8929 +28.4089) =1/(1+sqrt(46.3018)) =1/(1+6.8045) =1/7.8045	1/(1+sqrt((4.7-4.69)^2+(-7.83-36.75)^2)) =1/(1+sqrt((0.01)^2+(-44.58)^2)) =1/(1+sqrt(.00004 +1987.3765) =1/(1+sqrt(1987.3765)) =1/(1+44.5800) =1/45.5800	1/(1+sqrt((4.7-4.38)^2+(-7.83-34.43)^2)) =1/(1+sqrt((0.32)^2+(-42.26)^2)) =1/(1+sqrt(0.1024 +1785.9076) =1/(1+sqrt(1786.01)) =1/(1+42.2612) =1/43.2612	1/(1+sqrt((4.7-4.07)^2+(-7.83-33.8)^2)) =1/(1+sqrt((0.63)^2+(-41.63)^2)) =1/(1+sqrt(0.3969 +1733.0569) =1/(1+sqrt(1733.4538)) =1/(1+41.6348) =1/42.6348

				=0.0828	=0.1281	=0.0219	=0.0231	=0.0235
(0.69, 2.5)	0.0504	0.0486	0.0828	1	1/(1+sqrt((0.69-0.47)^2+(2.5-(-2.5))^2)) =1/(1+sqrt((0.22)^2+(5)^2)) =1/(1+sqrt(0.0484+25) =1/(1+sqrt(25.0484)) =1/(1+5.0048) =1/6.0048 =0.1665	1/(1+sqrt((0.69-4.69)^2+(2.5-36.75)^2)) =1/(1+sqrt((-4)^2+(-34.25)^2)) =1/(1+sqrt(16 +1173.0625) =1/(1+sqrt(1189.0625)) =1/(1+34.4828) =1/35.4828 =0.0282	1/(1+sqrt((0.69-4.38)^2+(2.5-34.43)^2)) =1/(1+sqrt((-3.69)^2+(-31.93)^2)) =1/(1+sqrt(13.6161 +1019.5249) =1/(1+sqrt(1033.141)) =1/(1+32.1425) =1/33.1425 =0.0302	1/(1+sqrt((0.69-4.07)^2+(2.5-33.8)^2)) =1/(1+sqrt((-3.38)^2+(-31.3)^2)) =1/(1+sqrt(11.4244 +979.69) =1/(1+sqrt(991.1144)) =1/(1+31.4820) =1/32.4820 =0.0308
(0.47, -2.5)	0.0665	0.0633	0.1281	0.1665	1	1/(1+sqrt((0.47-4.69)^2+(-2.5-36.75)^2)) =1/(1+sqrt((-4.22)^2+(-39.25)^2)) =1/(1+sqrt(17.8084 +1540.5625) =1/(1+sqrt(1558.3709)) =1/(1+39.4762) =1/40.4762 =0.0247	1/(1+sqrt((0.47-4.38)^2+(-2.5-34.43)^2)) =1/(1+sqrt((-3.91)^2+(-36.93)^2)) =1/(1+sqrt(15.2881 +1363.8249) =1/(1+sqrt(1379.113)) =1/(1+37.1364) =1/38.1364 =0.0262	1/(1+sqrt((0.47-4.07)^2+(-2.5-33.8)^2)) =1/(1+sqrt((-3.6)^2+(-36.3)^2)) =1/(1+sqrt(12.96 +1317.69) =1/(1+sqrt(1330.65)) =1/(1+36.4780) =1/37.4780 =0.0267
(4.69, 36.75)	0.0186	0.0184	0.0219	0.0282	0.0247	1	1/(1+sqrt((4.69-4.38)^2+(36.75-34.43)^2)) =1/(1+sqrt((0.31)^2 +(2.32)^2)) =1/(1+sqrt(0.0961 +5.3824) =1/(1+sqrt(5.4785)) =1/(1+2.3406) =1/3.3406 =0.2993	1/(1+sqrt((4.69-4.07)^2+(36.75-33.8)^2)) =1/(1+sqrt((0.62)^2 +(2.95)^2)) =1/(1+sqrt(0.3844 +8.7025) =1/(1+sqrt(9.0869)) =1/(1+3.0144) =1/4.0144 =0.2491
(4.38, 34.43)	0.0195	0.0192	0.0231	0.0302	0.0262	0.2993	1	1/(1+sqrt((4.38-4.07)^2+(34.43-33.8)^2)) =1/(1+sqrt((0.31)^2 +(0.63)^2)) =1/(1+sqrt(0.0961 +0.3969) =1/(1+sqrt(0.493)) =1/(1+0.7021) =1/1.7021 =0.5875
(4.07, 33.8)	0.0197	0.0195	0.0235	0.0308	0.0267	0.2491	0.5875	1

Step 2: Determine the two closest points, merge them, and update the matrix

Based on the matrix above, the closest two points are (4.38, 34.43) and (4.07, 33.8) (similarity score of 0.5875). Merging them produces the following set of clusters/points:

Cluster/Point	**Contents**
A	{(4.38, 34.43), (4.07, 33.8)}
(4.38, -15.96)	{(4.38, -15.96)}
(5.01, -16.59)	{(5.01, -16.59)}
(0.69, 2.5)	{(0.69, 2.5)}
(4.69, 36.75)	{(4.69, 36.75)}
(4.7, -7.83)	{(4.7, -7.83)}
(0.47, -2.5)	{(0.47, -2.5)}

Updating the matrix and combining the two points in cluster A (keeping only the lowest similarity score between the two merged points, since MAX wants the greatest distance between clusters) produces the following:

	(4.38, -15.96)	**(5.01, -16.59)**	**(4.7, -7.83)**	**(0.69, 2.5)**	**(0.47, -2.5)**	**(4.69, 36.75)**	**A**
(4.38, -15.96)	1	0.5288	0.1095	0.0504	0.0666	0.0186	0.0195<0.0197
(5.01, -16.59)	0.5288	1	0.1024	0.0486	0.0633	0.0184	0.0192<0.0195
(4.7, -7.83)	0.1095	0.1024	1	0.0828	0.1281	0.0219	0.0231<0.0235
(0.69, 2.5)	0.0504	0.0486	0.0828	1	0.1665	0.0282	0.0302<0.0308
(0.47, -2.5)	0.0665	0.0633	0.1281	0.1665	1	0.0247	0.0262<0.0267
(4.69, 36.75)	0.0186	0.0184	0.0219	0.0282	0.0247	1	0.2491<0.2993
A	0.0195	0.0192	0.0231	0.0302	0.0262	0.2491	1

Step 3: Determine the two closest points/clusters, merge them, and update the matrix

Based on the updated matrix, the closest two points/clusters are (4.38, -15.96) and (5.01, -16.59) (similarity score of 0.5288). Merging them produces the following set of clusters/points:

Cluster/Point	**Contents**
A	{(4.38, 34.43), (4.07, 33.8)}
B	{(4.38, -15.96), (5.01, -16.59)}
(0.69, 2.5)	{(0.69, 2.5)}
(4.69, 36.75)	{(4.69, 36.75)}
(4.7, -7.83)	{(4.7, -7.83)}
(0.47, -2.5)	{(0.47, -2.5)}

Updating the matrix and combining the two points in cluster B produces the following:

	B	**(4.7, -7.83)**	**(0.69, 2.5)**	**(0.47, -2.5)**	**(4.69, 36.75)**	**A**
B	1	0.1024<0.1095	0.0486<0.0504	0.0633<0.0666	0.0184<0.0186	0.0192<0.0195
(4.7, -7.83)	0.1024	1	0.0828	0.1281	0.0219	0.0231
(0.69, 2.5)	0.0486	0.0828	1	0.1665	0.0282	0.0302
(0.47, -2.5)	0.0633	0.1281	0.1665	1	0.0247	0.0262
(4.69, 36.75)	0.0184	0.0219	0.0282	0.0247	1	0.2491
A	0.0192	0.0231	0.0302	0.0262	0.2491	1

Step 4: Determine the two closest points/clusters, merge them, and update the matrix

Based on the updated matrix, the closest two points/clusters are (4.69, 36.75) and Cluster A (similarity score of 0.2491). Merging them produces the following set of clusters/points:

Cluster/Point	Contents
AC	{{(4.38, 34.43), (4.07, 33.8)}, (4.69, 36.75)}
B	{(4.38, -15.96), (5.01, -16.59)}
(0.69, 2.5)	{(0.69, 2.5)}
(4.7, -7.83)	{(4.7, -7.83)}
(0.47, -2.5)	{(0.47, -2.5)}

And updating the matrix looks like this:

	B	**(4.7, -7.83)**	**(0.69, 2.5)**	**(0.47, -2.5)**	**AC**
B	1	0.1024	0.0486	0.0633	0.0184<0.0192
(4.7, -7.83)	0.1024	1	0.0828	0.1281	0.0219<0.0231
(0.69, 2.5)	0.0486	0.0828	1	0.1665	0.0282<0.0302
(0.47, -2.5)	0.0633	0.1281	0.1665	1	0.0247<0.0262
AC	0.0184	0.0219	0.0282	0.0247	1

Step 5: Determine the two closest points/clusters, merge them, and update the matrix

Based on the updated matrix, the closest two points/clusters are (0.69, 2.5) and (0.47, -2.5) (similarity score of 0.1665). Merging them produces the following set of clusters/points:

Cluster/Point	Contents
AC	{{(4.38, 34.43), (4.07, 33.8)}, (4.69, 36.75)}
B	{(4.38, -15.96), (5.01, -16.59)}
D	{(0.69, 2.5), (0.47, -2.5)}
(4.7, -7.83)	{(4.7, -7.83)}

And updating the matrix looks like this:

	B	**(4.7, -7.83)**	**D**	**AC**
B	1	0.1024	0.0486<0.0633	0.0184
(4.7, -7.83)	0.1024	1	0.0828<0.1281	0.0219
D	0.0486	0.0828	1	0.0247
AC	0.0184	0.0219	0.0247<0.0282	1

Step 6: Determine the two closest points/clusters, merge them, and update the matrix

Based on the updated matrix, the closest two points/clusters are (4.7, -7.83) and Cluster B (similarity score of 0.1024). Merging them produces the following set of clusters/points:

Cluster/Point	Contents
AC	{{(4.38, 34.43), (4.07, 33.8)}, (4.69, 36.75)}
BE	{{(4.38, -15.96), (5.01, -16.59)}, (4.7, -7.83)}
D	{(0.69, 2.5), (0.47, -2.5)}

And updating the matrix looks like this:

	BE	**D**	**AC**
BE	1	0.0486<0.0828	0.0184<0.0219
D	0.0486	1	0.0247
AC	0.0184	0.0247	1

Step 7: Determine the two closest points/clusters, merge them, and update the matrix

Based on the updated matrix, the closest two points/clusters are Cluster BE and Cluster D (similarity score of 0.0486). Merging them produces the following set of clusters/points:

Cluster/Point	Contents
AC	{{(4.38, 34.43), (4.07, 33.8)}, (4.69, 36.75)}
BED	{{{(4.38, -15.96), (5.01, -16.59)}, (4.7, -7.83)}, {(0.69, 2.5), (0.47, -2.5)}}

And updating the matrix looks like this:

	BED	**AC**
BED	1	0.0184<0.0247
AC	0.0184	1

Step 8: Merge the last two clusters and update the matrix

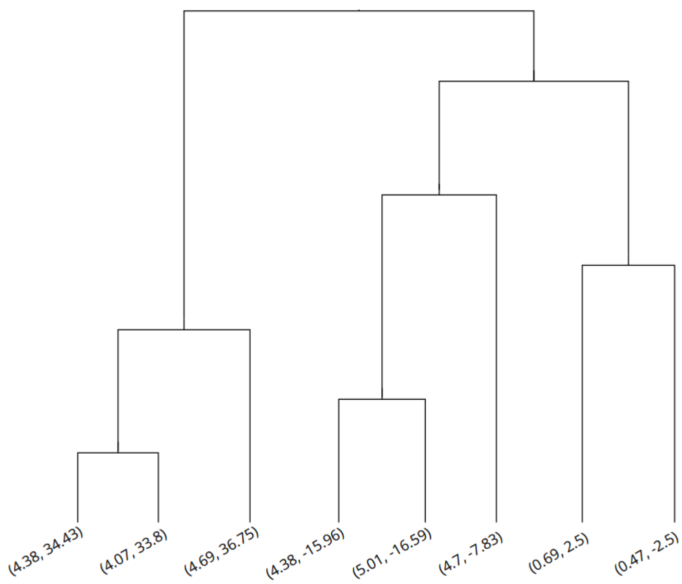
Since only two clusters remain, there is only one way to merge.

Cluster/Point	Contents
ACBED	{{{(4.38, 34.43), (4.07, 33.8)}, (4.69, 36.75)}, {{{(4.38, -15.96), (5.01, -16.59)}, (4.7, -7.83)}, {(0.69, 2.5), (0.47, -2.5)}}}

The updated matrix is one value:

	ACBED
ACBED	1

Dendogram



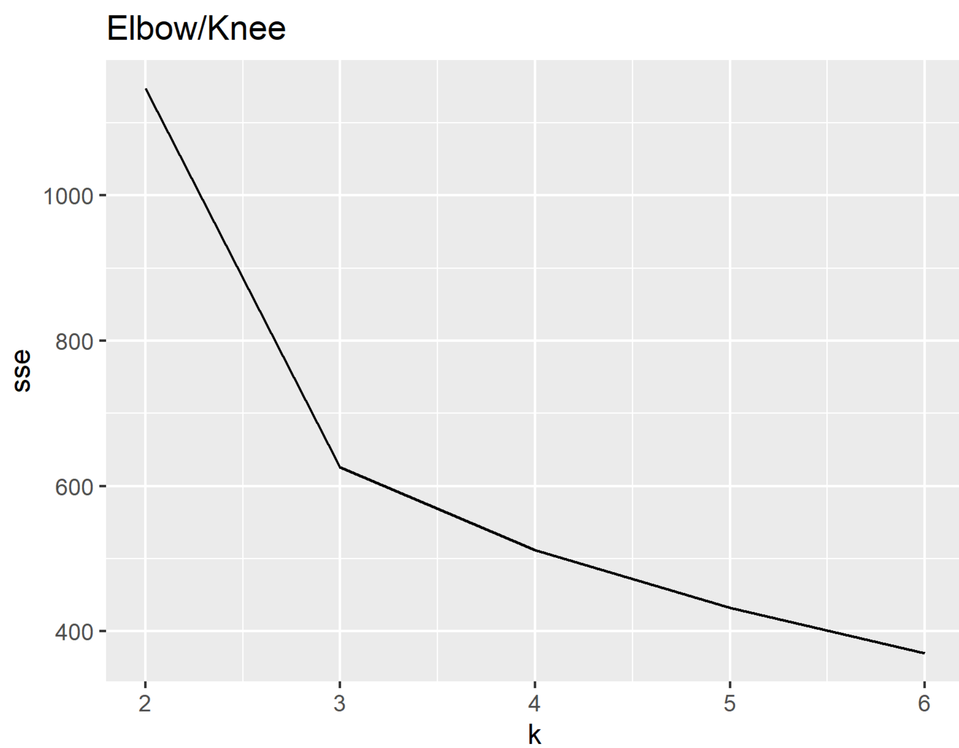
The height of the bracket indicates when the points were clustered (i.e. the lowest brackets were clustered first).

3. Here is the output from code that is written:

```
> printresult(2, k2)
K = 2
Total SSE: 1147.248
Cluster 1
SSE: 425.5769
Centers for 8 columns in Cluster 1 : 17.94238 15.95333 0.8842095 6.086417 3.641488
3.422846 5.90419 1.821429
Items in this cluster: 1 5 9 10 18 23 26 32 36 37 38 44 47 52 59 71 72 73 74 75 76 77 78
79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105
106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127
128 129 130 131 132 133 134 135 137 138 139 140

Cluster 2
SSE: 721.6711
Centers for 8 columns in Cluster 2 : 12.78429 13.62992 0.8621913 5.323278 3.003349
3.885104 5.077325 2.119048
Items in this cluster: 2 3 4 6 7 8 11 12 13 14 15 16 17 19 20 21 22 24 25 27 28 29 30 31
33 34 35 39 40 41 42 43 45 46 48 49 50 51 53 54 55 56 57 58 60 61 62 63 64 65 66 67 68 69
70 136 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160
161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182
183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202 203 204
205 206 207 208 209 210

> printresult(3, k3)
K = 3
Total SSE: 625.72
Cluster 1
SSE: 185.0922
Centers for 8 columns in Cluster 1 : 18.7218 16.29738 0.8850869 6.208934 3.722672 3.60359
6.066098 1.983607
Items in this cluster: 38 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91
92 93 94 95 96 97 98 99 100 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116
117 118 119 120 121 122 124 126 127 128 129 130 131 132 137
```

```
Cluster 2
SSE: 217.4687
Centers for 8 columns in Cluster 2 : 11.90907 13.25027 0.8515493 5.222333 2.865093 4.722187 5.09304 2.866667
Items in this cluster: 20 40 61 62 63 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207 208 209 210

Cluster 3
SSE: 223.1591
Centers for 8 columns in Cluster 3 : 14.63203 14.45324 0.8790973 5.561784 3.274892 2.744043 5.184932 1.135135
Items in this cluster: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 39 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 64 65 66 67 68 69 70 101 123 125 133 134 135 136 138 139 140

> printresult(4, k4)
K = 4
Total SSE: 512.3364
Cluster 1
SSE: 75.15799
Centers for 8 columns in Cluster 1 : 16.39333 15.30455 0.8792273 5.856788 3.46503 3.685182 5.663212 1.666667
Items in this cluster: 5 9 10 11 26 32 36 37 38 44 52 71 72 73 75 76 77 80 81 96 101 108 123 125 130 133 134 135 136 137 138 139 140

Cluster 2
SSE: 118.2116
Centers for 8 columns in Cluster 2 : 19.15104 16.46917 0.8870896 6.268854 3.772937 3.460417 6.12725 2
Items in this cluster: 74 78 79 82 83 84 85 86 87 88 89 90 91 92 93 94 95 97 98 99 100 102 103 104 105 106 107 109 110 111 112 113 114 115 116 117 118 119 120 121 122 124 126 127 128 129 131 132

Cluster 3
SSE: 168.4539
Centers for 8 columns in Cluster 3 : 11.89232 13.25696 0.8494203 5.233652 2.855565 4.897275 5.119217 2.971014
Items in this cluster: 40 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200 201 203 204 205 206 207 208 209 210

Cluster 4
SSE: 150.5129
Centers for 8 columns in Cluster 4 : 13.953 14.11917 0.878415 5.44485 3.1971 2.523653 5.024583 1.066667
Items in this cluster: 1 2 3 4 6 7 8 12 13 14 15 16 17 18 19 20 21 22 23 24 25 27 28 29 30 31 33 34 35 39 41 42 43 45 46 47 48 49 50 51 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 180 202

> printresult(5, k5)
K = 5
Total SSE: 432.4703
Cluster 1
SSE: 118.2116
Centers for 8 columns in Cluster 1 : 19.15104 16.46917 0.8870896 6.268854 3.772937 3.460417 6.12725 2
Items in this cluster: 74 78 79 82 83 84 85 86 87 88 89 90 91 92 93 94 95 97 98 99 100 102 103 104 105 106 107 109 110 111 112 113 114 115 116 117 118 119 120 121 122 124 126 127 128 129 131 132

Cluster 2
SSE: 70.87234
Centers for 8 columns in Cluster 2 : 12.7769 13.57207 0.8710655 5.266759 3.028207 2.76321 4.932138 1.551724
Items in this cluster: 13 14 15 17 20 24 27 28 30 31 41 42 43 60 61 62 63 64 65 66 70 147 149 161 166 180 199 200 202

Cluster 3
SSE: 72.7256
Centers for 8 columns in Cluster 3 : 14.89522 14.57826 0.8804435 5.609304 3.311174 2.354807 5.200435 1.043478
Items in this cluster: 1 2 3 4 5 6 7 8 9 10 12 16 18 19 21 22 23 25 26 29 32 33 34 35 36 37 39 45 46 47 48 49 50 51 53 54 55 56 57 58 59 67 68 69 138 139

Cluster 4
SSE: 49.45387
Centers for 8 columns in Cluster 4 : 16.4352 15.3156 0.880056 5.84796 3.47416 4.32636 5.68664 1.8
```

```
Items in this cluster: 11 38 40 44 52 71 72 73 75 76 77 80 81 96 101 108 123 125 130 133 134 135 136 137 140

Cluster 5
SSE: 121.2068
Centers for 8 columns in Cluster 5 : 11.80871 13.22339 0.84785 5.227806 2.842258 5.069823 5.115629 3
Items in this cluster: 141 142 143 144 145 146 148 150 151 152 153 154 155 156 157 158 159 160 162 163 164 165 167 168 169 170 171 172 173 174 175 176 177 178 179 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 201 203 204 205 206 207 208 209 210

> printresult(6, k6)
K = 6
Total SSE: 370.2333
Cluster 1
SSE: 70.87234
Centers for 8 columns in Cluster 1 : 12.7769 13.57207 0.8710655 5.266759 3.028207 2.76321 4.932138 1.551724
Items in this cluster: 13 14 15 17 20 24 27 28 30 31 41 42 43 60 61 62 63 64 65 66 70 147 149 161 166 180 199 200 202

Cluster 2
SSE: 49.45387
Centers for 8 columns in Cluster 2 : 16.4352 15.3156 0.880056 5.84796 3.47416 4.32636 5.68664 1.8
Items in this cluster: 11 38 40 44 52 71 72 73 75 76 77 80 81 96 101 108 123 125 130 133 134 135 136 137 140

Cluster 3
SSE: 25.81297
Centers for 8 columns in Cluster 3 : 19.58333 16.646 0.8877267 6.315867 3.835067 5.081533 6.1444 2
Items in this cluster: 78 79 82 83 89 90 94 95 103 114 115 117 121 126 127

Cluster 4
SSE: 72.7256
Centers for 8 columns in Cluster 4 : 14.89522 14.57826 0.8804435 5.609304 3.311174 2.354807 5.200435 1.043478
Items in this cluster: 1 2 3 4 5 6 7 8 9 10 12 16 18 19 21 22 23 25 26 29 32 33 34 35 36 37 39 45 46 47 48 49 50 51 53 54 55 56 57 58 59 67 68 69 138 139

Cluster 5
SSE: 30.16164
Centers for 8 columns in Cluster 5 : 18.95455 16.38879 0.8868 6.247485 3.744697 2.723545 6.119455 2
Items in this cluster: 74 84 85 86 87 88 91 92 93 97 98 99 100 102 104 105 106 107 109 110 111 112 113 116 118 119 120 122 124 128 129 131 132

Cluster 6
SSE: 121.2068
Centers for 8 columns in Cluster 6 : 11.80871 13.22339 0.84785 5.227806 2.842258 5.069823 5.115629 3
Items in this cluster: 141 142 143 144 145 146 148 150 151 152 153 154 155 156 157 158 159 160 162 163 164 165 167 168 169 170 171 172 173 174 175 176 177 178 179 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 201 203 204 205 206 207 208 209 210
```



Code:

```
#Written for Data Mining Fall 2024 by Chelsea Murray

#Please update working directory as necessary for your machine.

setwd("C:/Users/Chelsea/Documents/FA24/Data Mining/Homework/homework 4")


dpath = paste(getwd(), "/seeds/seeds_dataset.txt", sep="")

df<-read.table(dpath, header=FALSE, sep="")


#3.2: K-Means and Elbow/Knee


#Cluster with k=2 through 6

k2 = kmeans(df, 2)
```

```
k3 = kmeans(df, 3)
k4 = kmeans(df, 4)
k5 = kmeans(df, 5)
k6 = kmeans(df, 6)

#Function for printing results
#k = number of clusters (ex. k=2)
#res = kmeans result (ex. res=k2)
printresult<-function(k, res){
  cat("K =", k, "\n")
  #Total SSE for all clusters
  cat("Total SSE:", res$tot.withinss, "\n")
  #convert center indexes to a data frame
  kc = data.frame(seq(1, 8*k, by=k))
  for (i in 2:k){
    col = toString(i)
    kc[col] <- data.frame(seq(i, 8*k, by=k))
  }
  for (i in 1:k) {
    cat("Cluster", i, "\n")
    cat("SSE:", res$withinss[i], "\n")
    cat("Centers for 8 columns in Cluster", i, ":", res$centers[kc[,i]], "\n")
    cat("Items in this cluster:", which(res$cluster == i), "\n\n")
  }
```

```
}

#K=x results
printresult(2, k2)
printresult(3, k3)
printresult(4, k4)
printresult(5, k5)
printresult(6, k6)

#Elbow/knee
library(ggplot2)
sse <- c(k2$tot.withinss, k3$tot.withinss, k4$tot.withinss, k5$tot.withinss,
        k6$tot.withinss)
k <- 2:6
elb <- data.frame(k, sse)
ggplot(elb, aes(x=k, y=sse)) + geom_line() +ggtitle("Elbow/Knee")
ggsave("elbow.png")

#Based on the output, k=3 seems to be an adequate K-value, since the decrease
#in SSE drops sharply at first but "evens out" at the "elbow" of k=3.


#3.3: Bisecting K-means
bisect_km <- function(k, clusters, sses, means=NULL){
```

```
#find cluster with highest sse
high = 0
ind = 0
for(i in 1:length(clusters)){
  if (sses[i] > high){
    ind = i
    high = sses[i]
  }
}
#set current cluster for bisection to highest sse cluster
data = clusters[ind]

#3 trials
t1 = kmeans(data, 2)
t2 = kmeans(data, 2)
t3 = kmeans(data, 2)
#determine ideal trial
kept = t1
if(t1$tot.withinss > t2$tot.withinss){
  if(t2$tot.withinss > t3$tot.withinss){
    kept = t3
  }
  else {
    kept = t2
```

```
  }
}
else{
  if(t1$tot.withinss > t3$tot.withinss){
    kept = t3
  }
  else {
    kept = t1
  }
}
#for identified ideal trial, split the data into the 2 clusters
an <- which(kept$cluster == 1)
bn <- which(kept$cluster == 2)
a <- data[-bn]
b <- data[-an]
#update sse vector
sses[ind] = kept$withinss[1]
sses.append(kept$withinss[2])
#update cluster set
clusters[ind] = a
clusters.append(b)

if(k > 1){
  bisect_km(k-1, clusters, sses, kmres)
```

```
  }
  else{
    cat("K =", k, "\n")
    #Total SSE for all clusters
    cat("Total SSE:", sum(sses), "\n")
    for (i in 1:k) {
      cat("Cluster", i, "\n")
      cat("SSE:", sses[i], "\n")
    }
  }
}

cluster = list(df)
sses = list(0)

bisect_km(k, cluster, sses)
```