Task 1

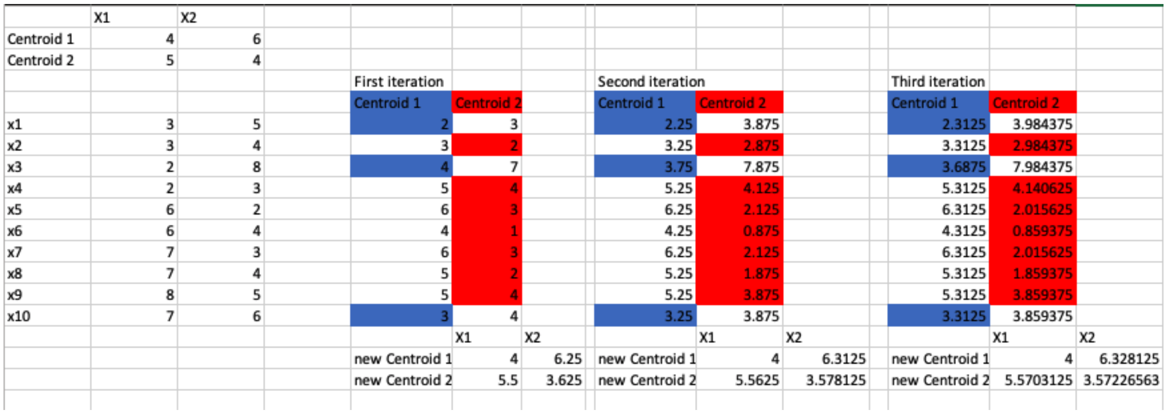
Suppose we have 10 college football teams X1 to X10. We want to cluster them into 2 groups. For each football team, we have two features: One is # wins in Season 2016, and the other is # wins in Season 2017.

(1) Initialize with two centroids, (4, 6) and (5, 4). Use Manhattan distance as the distance metric. First, perform one iteration of the K-means algorithm and report the coordinates of the resulting centroids. Second, please use K-Means to find two clusters.

First iteration:

Using centroid 1 (4, 6)	Using centroid 2 (5, 4)
X1: $\lvert 4-3\rvert + \lvert 6-5\rvert = 2$	X1: $\lvert 5-3\rvert + \lvert 4-5\rvert = 3$
X2: $\lvert 4-3\rvert + \lvert 6-4\rvert = 3$	X2: $\lvert 5-3\rvert + \lvert 4-4\rvert = 2$
X3: $\lvert 4-2\rvert + \lvert 6-8\rvert = 4$	X3: $\lvert 5-2\rvert + \lvert 4-8\rvert = 7$
X4: $\lvert 4-2\rvert + \lvert 6-3\rvert = 5$	X4: $\lvert 5-2\rvert + \lvert 4-3\rvert = 4$
X5: $\lvert 4-6\rvert + \lvert 6-2\rvert = 6$	X5: $\lvert 5-6\rvert + \lvert 4-2\rvert = 3$
X6: $\lvert 4-6\rvert + \lvert 6-4\rvert = 4$	X6: $\lvert 5-6\rvert + \lvert 4-4\rvert = 1$
X7: $\lvert 4-7\rvert + \lvert 6-3\rvert = 6$	X7: $\lvert 5-7\rvert + \lvert 4-3\rvert = 3$
X8: $\lvert 4-7\rvert + \lvert 6-4\rvert = 5$	X8: $\lvert 5-7\rvert + \lvert 4-4\rvert = 2$
X9: $\lvert 4-8\rvert + \lvert 6-5\rvert = 5$	X9: $\lvert 5-8\rvert + \lvert 4-5\rvert = 4$
X10: $\lvert 4-7\rvert + \lvert 6-6\rvert = 3$	X10: $\lvert 5-7\rvert + \lvert 4-6\rvert = 4$

New Centroid 1 = (4, 6.25)
New Centroid 2 = (5.5, 3.625)



	X1	X2										
Centroid 1	4	6										
Centroid 2	5	4										
				First iteration			Second iteration			Third iteration		
				Centroid 1	Centroid 2		Centroid 1	Centroid 2		Centroid 1	Centroid 2	
x1	3	5		2	3		2.25	3.875		2.3125	3.984375	
x2	3	4		3	2		3.25	2.875		3.3125	2.984375	
x3	2	8		4	7		3.75	7.875		3.6875	7.984375	
x4	2	3		5	4		5.25	4.125		5.3125	4.140625	
x5	6	2		6	3		6.25	2.125		6.3125	2.015625	
x6	6	4		4	1		4.25	0.875		4.3125	0.859375	
x7	7	3		6	3		6.25	2.125		6.3125	2.015625	
x8	7	4		5	2		5.25	1.875		5.3125	1.859375	
x9	8	5		5	4		5.25	3.875		5.3125	3.859375	
x10	7	6		3	4		3.25	3.875		3.3125	3.859375	
					X1	X2		X1	X2		X1	X2
				new Centroid 1	4	6.25	new Centroid 1	4	6.3125	new Centroid 1	4	6.328125
				new Centroid 2	5.5	3.625	new Centroid 2	5.5625	3.578125	new Centroid 2	5.5703125	3.57226563

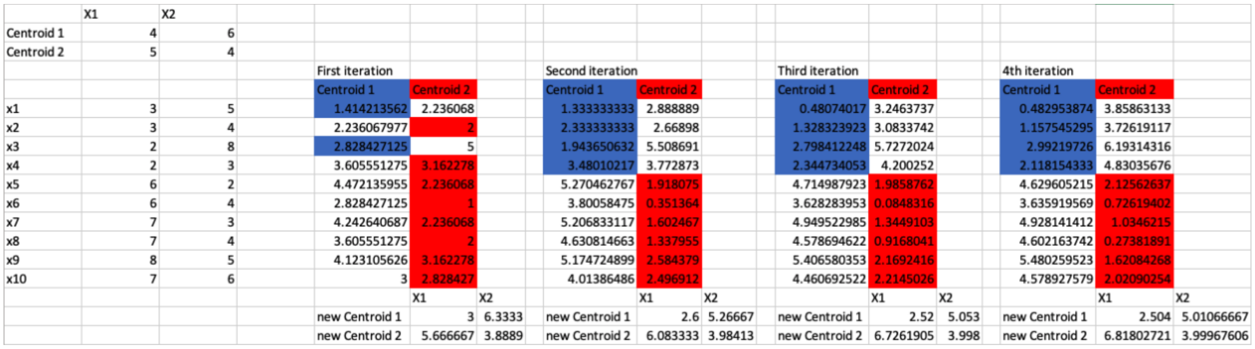
After k-means, cluster centroids seemed to stabilize around centroid 1 = (4, 6.33) and centroid 2 = (5.57, 3.57).

(2) Initialize with two centroids, (4, 6) and (5, 4). Use Euclidean distance as the distance metric. First, perform one iteration of the K-means algorithm and report the coordinates of the resulting centroids. Second, please use K-Means to find two clusters.

First iteration:

Using centroid 1 (4, 6)	Using centroid 2 (5,4)
X1: $\sqrt{(3-4)^2+(5-6)^2}=1.414$	X1: $\sqrt{(3-5)^2+(5-4)^2}=2.23$
X2: $\sqrt{(3-4)^2+(4-6)^2}=2.23$	X2: $\sqrt{(3-5)^2+(4-4)^2}=2$
X3:$\sqrt{(2-4)^2+(8-6)^2}=2.82$	X3: $\sqrt{(2-5)^2+(8-4)^2}=5$
X4: $\sqrt{(2-4)^2+(3-6)^2}=3.60$	X4: $\sqrt{(2-5)^2+(3-4)^2}=3.16$
X5: $\sqrt{(6-4)^2+(2-6)^2}=4.47$	X5: $\sqrt{(6-5)^2+(2-4)^2}=2.23$
X6: $\sqrt{(6-4)^2+(4-6)^2}=2.82$	X6: $\sqrt{(6-5)^2+(4-4)^2}=1$
X7: $\sqrt{(7-4)^2+(3-6)^2}=4.24$	X7: $\sqrt{(7-5)^2+(3-4)^2}=2.23$
X8:$\sqrt{(7-4)^2+(4-6)^2}=3.60$	X8: $\sqrt{(7-5)^2+(4-4)^2}=\ 2$
X9: $\sqrt{(8-4)^2+(5-6)^2}=4.123$	X9: $\sqrt{(8-5)^2+(5-4)^2}=3.16$
X10: $\sqrt{(7-4)^2+(6-6)^2}=3$	X10: $\sqrt{(7-5)^2+(6-4)^2}=2.82$

New Centroid 1 = (3, 6.333)
New Centroid 2 = (5.667, 3.889)



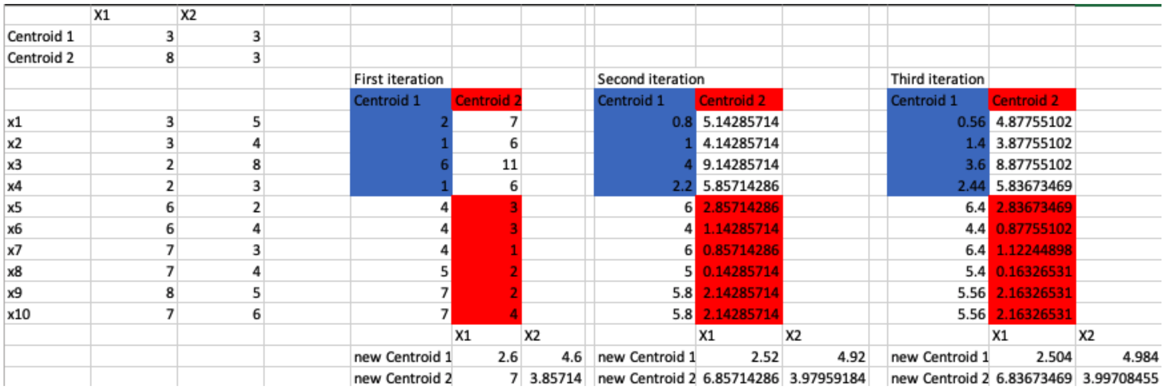
	X1	X2												
Centroid 1	4	6												
Centroid 2	5	4												
			First iteration			Second iteration			Third iteration			4th iteration		
			Centroid 1	Centroid 2		Centroid 1	Centroid 2		Centroid 1	Centroid 2		Centroid 1	Centroid 2	
x1	3	5	1.414213562	2.236068		1.333333333	2.888889		0.48074017	3.2463737		0.482953874	3.85863133	
x2	3	4	2.236067977	2		2.333333333	2.66898		1.328323923	3.0833742		1.157545295	3.72619117	
x3	2	8	2.828427125	5		1.943650632	5.508691		2.798412248	5.7272024		2.99219726	6.19314316	
x4	2	3	3.605551275	3.162278		3.48010217	3.772873		2.344734053	4.200252		2.118154333	4.83035676	
x5	6	2	4.472135955	2.236068		5.270462767	1.918075		4.714987923	1.9858762		4.629605215	2.12562637	
x6	6	4	2.828427125	1		3.80058475	0.351364		3.628283953	0.0848316		3.635919569	0.72619402	
x7	7	3	4.242640687	2.236068		5.206833117	1.602467		4.949522985	1.3449103		4.928141412	1.0346215	
x8	7	4	3.605551275	2		4.630814663	1.337955		4.578694622	0.9168041		4.602163742	0.27381891	
x9	8	5	4.123105626	3.162278		5.174724899	2.584379		5.406580353	2.1692416		5.480259523	1.62084268	
x10	7	6	3	2.828427		4.01386486	2.496912		4.460692522	2.2145026		4.578927579	2.02090254	
				X1	X2		X1	X2		X1	X2		X1	X2
			new Centroid 1	3	6.3333	new Centroid 1	2.6	5.26667	new Centroid 1	2.52	5.053	new Centroid 1	2.504	5.01066667
			new Centroid 2	5.666667	3.8889	new Centroid 2	6.083333	3.98413	new Centroid 2	6.7261905	3.998	new Centroid 2	6.81802721	3.99967606

After k-means, cluster centroids seemed to stabilize around centroid 1 = (2.504, 5.01) and centroid 2 = (6.818, 3.99).

(3) Initialize with two centroids, (3, 3) and (8, 3). Use Manhattan distance as the distance metric. First, perform one iteration of the K-means algorithm and report the coordinates of the resulting centroids. Second, please use K-Means to find two clusters.

Using the same excel sheets as above and replacing the points, after the first iteration:

New Centroid 1 = (2.6, 4.6)
New Centroid 2 = (7, 3.86)



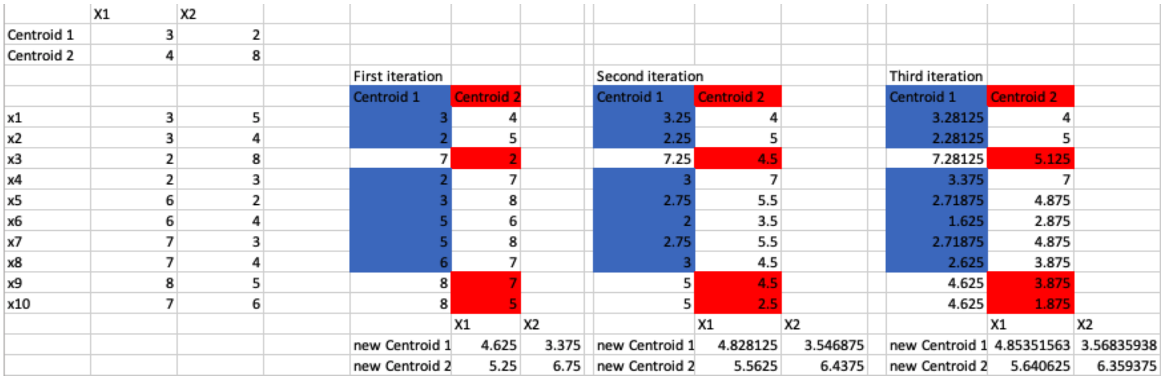
	X1	X2										
Centroid 1	3	3										
Centroid 2	8	3										
				First iteration			Second iteration			Third iteration		
				Centroid 1	Centroid 2		Centroid 1	Centroid 2		Centroid 1	Centroid 2	
x1	3	5		2	7		0.8	5.14285714		0.56	4.87755102	
x2	3	4		1	6		1	4.14285714		1.4	3.87755102	
x3	2	8		6	11		4	9.14285714		3.6	8.87755102	
x4	2	3		1	6		2.2	5.85714286		2.44	5.83673469	
x5	6	2		4	3		6	2.85714286		6.4	2.83673469	
x6	6	4		4	3		4	1.14285714		4.4	0.87755102	
x7	7	3		4	1		6	0.85714286		6.4	1.12244898	
x8	7	4		5	2		5	0.14285714		5.4	0.16326531	
x9	8	5		7	2		5.8	2.14285714		5.56	2.16326531	
x10	7	6		7	4		5.8	2.14285714		5.56	2.16326531	
					X1	X2		X1	X2		X1	X2
				new Centroid 1	2.6	4.6	new Centroid 1	2.52	4.92	new Centroid 1	2.504	4.984
				new Centroid 2	7	3.85714	new Centroid 2	6.85714286	3.97959184	new Centroid 2	6.83673469	3.99708455

After k-means, cluster centroids seemed to stabilize around centroid 1 = (2.504, 4.984) and centroid 2 = (6.83, 3.99).

(4) Initialize with two centroids, (3, 2) and (4, 8). Use Manhattan distance as the distance metric. First, perform one iteration of the K-means algorithm and report the coordinates of the resulting centroids. Second, please use K-Means to find two clusters.

Using the same excel sheets as above and replacing the points, after the first iteration:

New Centroid 1 = (4.625, 3.375)
New Centroid 2 = (5.25, 6.75)



	X1	X2										
Centroid 1	3	2										
Centroid 2	4	8										
				First iteration			Second iteration			Third iteration		
				Centroid 1	Centroid 2		Centroid 1	Centroid 2		Centroid 1	Centroid 2	
x1	3	5		3	4		3.25	4		3.28125	4	
x2	3	4		2	5		2.25	5		2.28125	5	
x3	2	8		7	2		7.25	4.5		7.28125	5.125	
x4	2	3		2	7		3	7		3.375	7	
x5	6	2		3	8		2.75	5.5		2.71875	4.875	
x6	6	4		5	6		2	3.5		1.625	2.875	
x7	7	3		5	8		2.75	5.5		2.71875	4.875	
x8	7	4		6	7		3	4.5		2.625	3.875	
x9	8	5		8	7		5	4.5		4.625	3.875	
x10	7	6		8	5		5	2.5		4.625	1.875	
					X1	X2		X1	X2		X1	X2
				new Centroid 1	4.625	3.375	new Centroid 1	4.828125	3.546875	new Centroid 1	4.85351563	3.56835938
				new Centroid 2	5.25	6.75	new Centroid 2	5.5625	6.4375	new Centroid 2	5.640625	6.359375

After k-means, cluster centroids seemed to stabilize around centroid 1 = (4.85, 3.568) and centroid 2 = (5.461, 6.359).

Task 2: K-Means Clustering with Real World Dataset

Q1: Run K-means clustering with Euclidean, Cosine and Jarcard similarity. Specify K=the number of categorical values of y(the variable of label). Compare the SSEs of Euclidean-K-means Cosine-K-means, Jarcard-K-means. Which method is better?

After running an iteration of the k-means algorithm:

```
Euclidean SSE: 86.96678921568629
Cosine SSE: 94.05184848484849
Jaccard SSE: 588.7835618052625
```

Q2: Compare the accuracies of Euclidean-K-means Cosine-K-means, Jarcard-K-means. First, label each cluster with the label of the highest votes. Later, compute the accuracy of the K-means with respect to the three similarity metrics. Which metric is better?

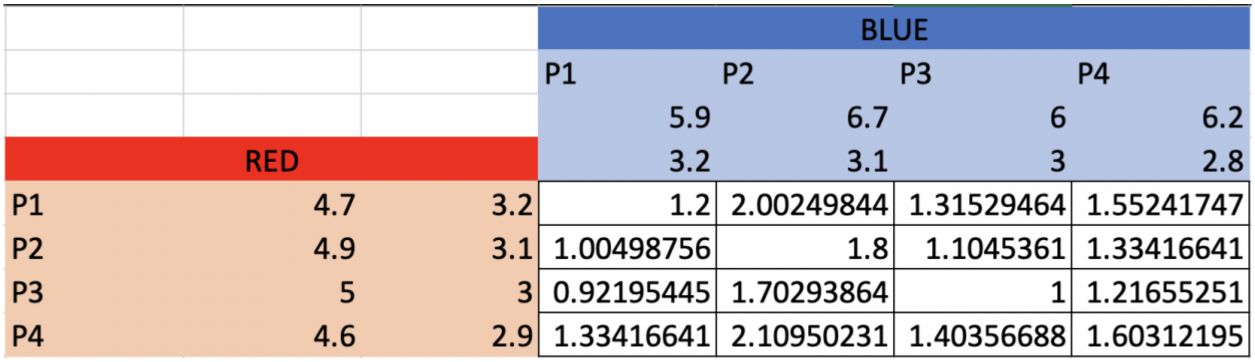
After running various iterations of each, it seems as the Euclidean had better accuracy but the Jaccard accuracy had less variance.

Q3: Which of Euclidean-K-means, Cosine-K-means, Jarcard-K-means requires more iterations and times?
Cosine seemed to be the algorithm that required more iteration to receive a good SSE.

Q4: Compare the SSEs of Euclidean-K-means Cosine-K-means, Jarcard-K-means with respect to the following three terminating conditions:
• when there is no change in centroid position
• when the SSE value increases in the next iteration
• when the maximum preset value (100) of iteration is complete
Which method requires more time or more iterations?

Out of all 3 methods, the maximum preset value took the most iterations. The other stop conditions would stop much before.

Task 3:
There are two clusters A (red) and B (blue), each has four members and plotted in Figure. The coordinates of each member are labeled in the figure. Compute the distance between two clusters using Euclidean distance.



			BLUE			
			P1	P2	P3	P4
			5.9	6.7	6	6.2
RED			3.2	3.1	3	2.8
P1	4.7	3.2	1.2	2.00249844	1.31529464	1.55241747
P2	4.9	3.1	1.00498756	1.8	1.1045361	1.33416641
P3	5	3	0.92195445	1.70293864	1	1.21655251
P4	4.6	2.9	1.33416641	2.10950231	1.40356688	1.60312195

A. What is the distance between the two farthest members? (round to four decimal places here, and next 2 problems)

The distance between farthest members [p1: (4.6, 2.9), p3: (6.7, 3.1)] is 2.1095.

B. What is the distance between the two closest members?
The distance between the two closest members is [p3: (5,3), p1: (5.9, 3.2)] is 0.9219.

C. What is the average distance between all pairs?
The average distance is 1.4129.

D, Discuss which distance (A, B, C) is more robust to noises in this case?
Average linkage is often the best for robustness against outliers/noise.

GITHUB LINK FOR TASK 2: https://github.com/CamachoBry/CAP5610-ML/blob/main/Homeworks/HW4/HW4-Task2.ipynb

Additional Questions:
•Approximately how many hours did you spend on this assignment?
8 hours

•Which aspects of this assignment did you find most challenging? Were there any significant stumbling blocks?
Coding the kmeans algorithm from scratch.

•Which aspects of this assignment did you like? Is there anything you would have changed?
I liked performing the algorithm step by step to see the inner workings.