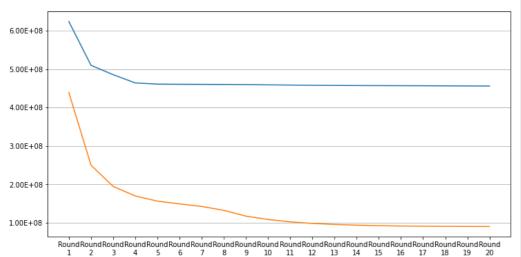
2021 Introduction to Massive Data Analysis

HW3 - K-means

107062117 李采蓉

- (a.) Euclidean distance
- 1. A plot of cost \(\) iteration for c1 and c2



2. Percentage improvement values and explanation.

For c1: 26.885383 %

For c2: **79.437750** %

在 improvement 上 c2 有明顯較好的表現,因為 c1 的 centroid 是隨機選取,平均散佈在所有資料點中,因此一開始選到的 centroid 有機會不能代表所有分群或與實際分群狀況差異較大,導致分類的結果較差。c2 則在一開始就選距離遠的點當 centroid,因此更有可能代表實際分佈,有較好的分群結果。

	c1	c2
Round 1	6.24E+08	4.39E+08
Round 2	5.10E+08	2.50E+08
Round 3	4.85E+08	1.94E+08
Round 4	4.64E+08	1.70E+08
Round 5	4.61E+08	1.56E+08
Round 6	4.61E+08	1.49E+08
Round 7	4.60E+08	1.43E+08
Round 8	4.60E+08	1.32E+08
Round 9	4.60E+08	1.17E+08
Round 10	4.59E+08	1.09E+08
Round 11	4.58E+08	1.02E+08
Round 12	4.58E+08	9.83E+07
Round 13	4.58E+08	9.56E+07
Round 14	4.57E+08	9.38E+07
Round 15	4.57E+08	9.24E+07
Round 16	4.57E+08	9.15E+07
Round 17	4.57E+08	9.10E+07
Round 18	4.56E+08	9.08E+07
Round 19	4.56E+08	9.05E+07
Round 20	4.56E+08	9.02E+07

3. The Distances for all pairs of centroids.

(1.) The Manhattan Distances for all pairs of centroids with c1

	1	2	3	4	5	6	7	8	9	10
1	0.00	728.92	3797.90	212.18	374.89	577.40	499.16	645.77	1731.06	406.70
2		0.00	3072.89	935.89	1100.83	1303.90	1225.35	1372.09	1005.29	490.93
3			0.00	4001.04	4170.30	4372.79	4294.95	4440.72	2513.42	3396.42
4				0.00	171.37	375.25	296.25	443.50	1934.09	609.75
5					0.00	204.52	125.60	272.93	2102.86	779.40
6						0.00	79.40	69.59	2306.38	983.02
7							0.00	147.87	2227.56	904.37
8								0.00	2374.55	1050.92
9									0.00	1327.58
10										0.00

(2.) The Euclidean Distances for all pairs of centroids with c1

	1	2	3	4	5	6	7	8	9	10
1	0.00	692.16	3490.26	205.75	346.72	512.61	444.73	566.20	1282.77	307.67
2		0.00	2798.80	897.66	1038.83	1204.08	1136.33	1257.45	669.89	412.08
3			0.00	3695.11	3836.91	4002.69	3934.87	4056.14	2294.58	3195.92
4				0.00	142.44	309.51	241.73	363.26	1474.95	504.63
5					0.00	167.15	99.55	220.90	1615.85	646.93
6						0.00	67.91	53.79	1782.20	814.08
7							0.00	121.63	1715.25	746.34
8								0.00	1835.64	867.82
9									0.00	975.32
10										0.00

(3.) The $\bf Manhattan\ Distances$ for all pairs of centroids with $\bf c2$

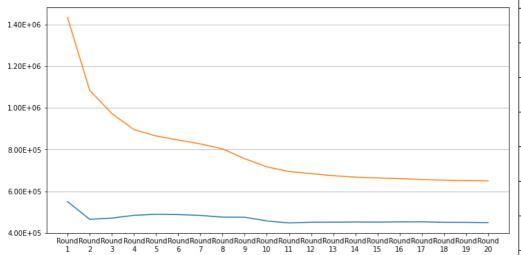
	1	2	3	4	5	6	7	8	9	10
1	0.00	15772.61	20215.65	9533.17	5604.20	3088.05	1311.04	471.27	2369.41	3349.66
2		0.00	16003.50	7219.20	10221.03	16105.35	14909.17	15434.46	13950.58	12776.88
3			0.00	10690.48	14613.55	17509.90	18912.61	19748.94	17851.81	16873.24
4				0.00	3935.29	8896.39	8228.36	9065.40	7168.73	6190.68
5					0.00	5893.07	4696.98	5221.25	3737.71	2564.17
6						0.00	1781.82	2619.81	2162.80	3337.75
7							0.00	840.72	1068.94	2137.79
8								0.00	1901.21	2883.73
9									0.00	1176.45
10										0.00

(4.) The Euclidean Distances for all pairs of centroids with c2

	1	2	3	4	5	6	7	8	9	10
1	0.00	15760.12	14110.83	9045.32	5567.68	1924.62	1100.86	402.89	2105.44	3169.00
2		0.00	11524.51	6743.88	10192.53	14455.12	14682.45	15362.42	13674.71	12597.04
3			0.00	9545.88	10883.38	12233.96	13208.00	13786.48	12508.96	11938.38
4				0.00	3494.22	7718.22	7957.78	8644.81	6947.82	5876.33
5					0.00	4404.56	4492.46	5169.94	3488.16	2407.92
6						0.00	1182.86	1615.79	1313.33	2153.77
7							0.00	698.49	1010.20	2085.46
8								0.00	1702.79	2768.61
9				_					0.00	1080.53
10				_						0.00

(b.) Manhattan distance

1. A plot of cost \ iteration for c1 and c2



2. Percentage improvement values and explanation.

For c1: **18.378954** %

For c2: **54.685694** %

在 improvement 上 c2 有明顯較好的表現,與先前所述相同,c2 選取距離遠的點當做 centroid,比 c1 的隨機選取更貼近實際可能的分群狀況。

	c1	c2
Round 1	5.50E+05	1.43E+06
Round 2	4.65E+05	1.08E+06
Round 3	4.71E+05	9.73E+05
Round 4	4.84E+05	8.96E+05
Round 5	4.89E+05	8.65E+05
Round 6	4.88E+05	8.46E+05
Round 7	4.84E+05	8.27E+05
Round 8	4.75E+05	8.04E+05
Round 9	4.75E+05	7.56E+05
Round 10	4.57E+05	7.17E+05
Round 11	4.47E+05	6.95E+05
Round 12	4.51E+05	6.84E+05
Round 13	4.51E+05	6.75E+05
Round 14	4.52E+05	6.67E+05
Round 15	4.52E+05	6.64E+05
Round 16	4.53E+05	6.60E+05
Round 17	4.53E+05	6.56E+05
Round 18	4.51E+05	6.53E+05
Round 19	4.50E+05	6.51E+05
Round 20	4.49E+05	6.50E+05

3. The Distances for all pairs of centroids.

(1.) The Manhattan Distances for all pairs of centroids with c1

	1	2	3	4	5	6	7	8	9	10
1	0.00	2341.02	11929.30	651.19	496.33	947.74	770.74	1056.80	1260.51	737.71
2		0.00	9597.44	2778.95	2830.14	3280.36	3104.29	3388.98	2380.46	1605.27
3			0.00	12323.29	12421.26	12871.48	12695.55	12979.13	10775.94	11196.79
4				0.00	335.95	558.47	382.46	667.53	1653.83	1379.17
5					0.00	452.86	276.33	561.85	1755.11	1226.66
6						0.00	177.59	110.22	2205.31	1677.67
7							0.00	287.43	2028.90	1500.99
8								0.00	2314.67	1786.81
9			_						0.00	1006.37
10										0.00

(2.) The Euclidean Distances for all pairs of centroids with c1

	1	2	3	4	5	6	7	8	9	10
1	0.00	2219.18	9948.04	528.70	413.37	827.72	681.03	917.13	832.15	729.06
2		0.00	7767.95	2734.05	2628.49	3044.48	2898.71	3133.46	1812.45	1491.36
3			0.00	10433.06	10361.37	10773.53	10626.49	10862.97	9340.28	9236.84
4				0.00	221.37	375.16	249.38	457.26	1156.58	1251.16
5					0.00	415.99	270.75	505.07	1171.96	1137.14
6						0.00	147.05	89.49	1529.46	1553.12
7							0.00	236.51	1391.55	1407.40
8								0.00	1613.56	1642.13
9									0.00	709.41
10										0.00

(3.) The Manhattan Distances for all pairs of centroids with c2

	1	2	3	4	5	6	7	8	9	10
1	0.00	15757.69	20200.26	9517.67	5588.85	3281.49	1430.21	602.95	2102.55	3211.46
2		0.00	16003.50	7219.20	10221.03	16325.27	14506.49	15335.96	14980.06	12922.93
3			0.00	10690.48	14613.55	17521.52	18775.12	19602.26	18111.89	16995.13
4				0.00	3935.29	9116.02	8090.51	8918.81	7771.22	6312.53
5					0.00	6110.83	4293.50	5123.07	4768.92	2710.06
6						0.00	1855.58	2682.57	1358.80	3413.04
7							0.00	833.43	674.83	1784.51
8								0.00	1500.82	2614.00
9									0.00	2062.25
10										0.00

(4.) The Euclidean Distances for all pairs of centroids with c2

	1	2	3	4	5	6	7	8	9	10
1	0.00	15747.23	14100.14	9032.33	5554.79	2006.70	1338.16	514.63	1571.24	3022.66
2		0.00	11524.51	6743.88	10192.53	14474.55	14412.06	15239.88	14328.23	12731.40
3			0.00	9545.88	10883.38	12167.79	13125.35	13684.61	12643.99	12006.39
4				0.00	3494.22	7742.63	7694.28	8521.20	7588.40	6009.82
5					0.00	4452.97	4219.76	5047.52	4167.64	2542.57
6						0.00	1405.11	1637.73	910.99	2124.26
7							0.00	827.84	566.55	1684.52
8								0.00	1081.38	2511.46
9									0.00	1649.39
10										0.00