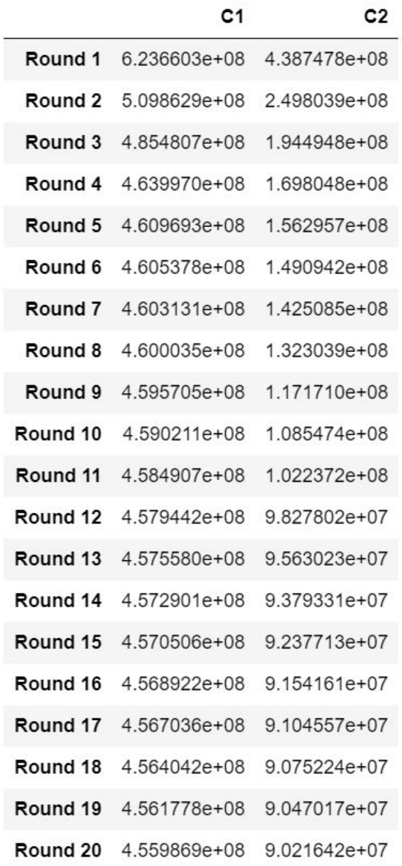
MDA HW3 KMeans Report

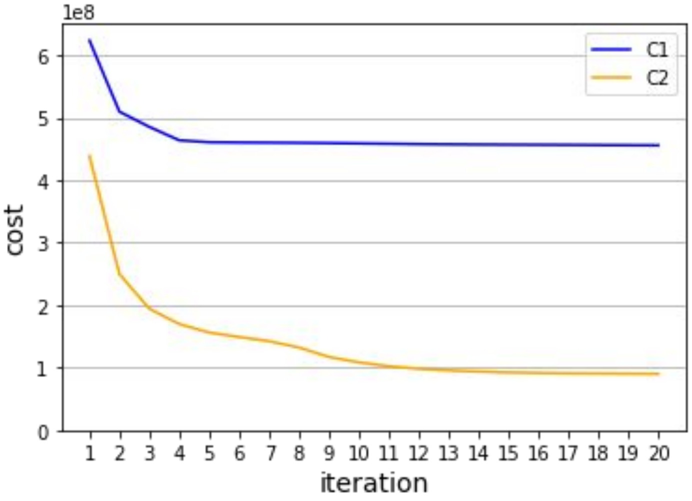
106062314 蔡政諺

(a) Euclidean distance

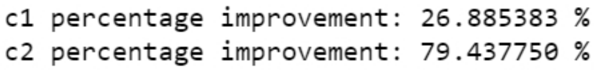
1. Plot of cost vs. iteration



	C1	C2
Round 1	6.236603e+08	4.387478e+08
Round 2	5.098629e+08	2.498039e+08
Round 3	4.854807e+08	1.944948e+08
Round 4	4.639970e+08	1.698048e+08
Round 5	4.609693e+08	1.562957e+08
Round 6	4.605378e+08	1.490942e+08
Round 7	4.603131e+08	1.425085e+08
Round 8	4.600035e+08	1.323039e+08
Round 9	4.595705e+08	1.171710e+08
Round 10	4.590211e+08	1.085474e+08
Round 11	4.584907e+08	1.022372e+08
Round 12	4.579442e+08	9.827802e+07
Round 13	4.575580e+08	9.563023e+07
Round 14	4.572901e+08	9.379331e+07
Round 15	4.570506e+08	9.237713e+07
Round 16	4.568922e+08	9.154161e+07
Round 17	4.567036e+08	9.104557e+07
Round 18	4.564042e+08	9.075224e+07
Round 19	4.561778e+08	9.047017e+07
Round 20	4.559869e+08	9.021642e+07

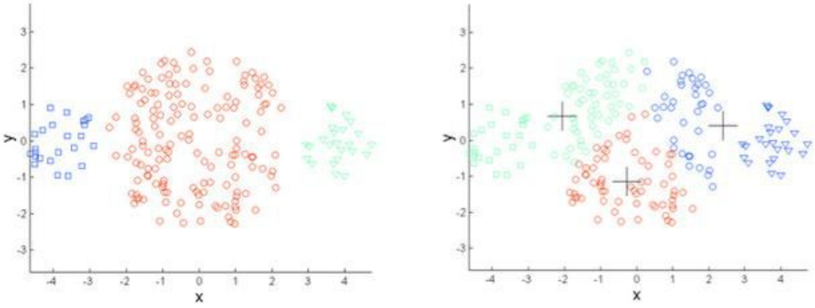


2. Percentage improvement and explanation



```
c1 percentage improvement: 26.885383 %
c2 percentage improvement: 79.437750 %
```

使用 Euclidean distance 做為 cost function 時，**C2 (as far as possible)的表現會比 C1 (random initialization)更好**。原因是 C2 的 centroids 會相較於 C1 更發散，因此 C1 的 centroids 會較 C2 更容易互搶本應屬於同一個 cluster 的 data points，導致最後收斂的時候，並沒有最 optimally 分割 k 個群，反而掉入一個 local minimum，如下圖。

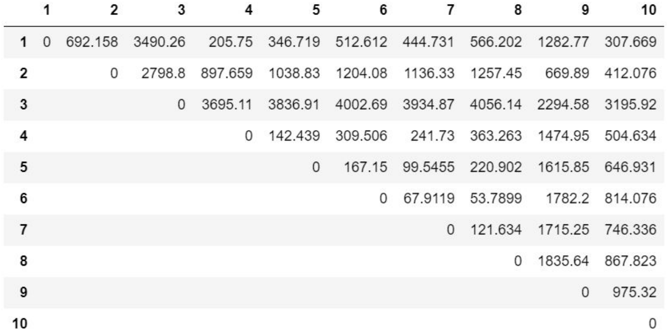


Original Points

K-means (k = 3)

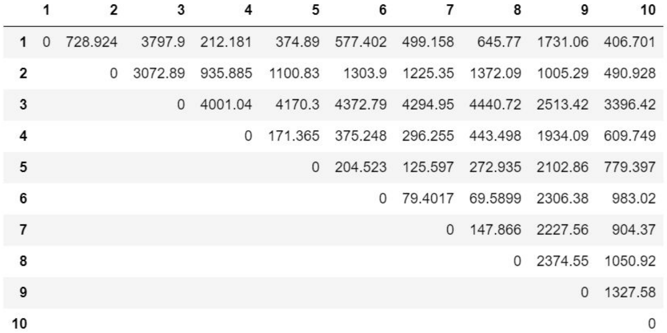
3. Distance of centroids

(i) Euclidean distances for centroids in c1



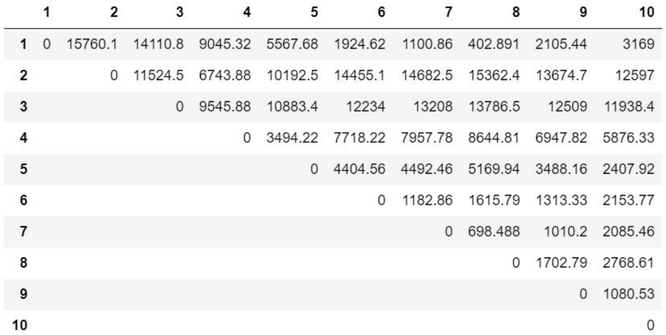
	1	2	3	4	5	6	7	8	9	10
1	0	692.158	3490.26	205.75	346.719	512.612	444.731	566.202	1282.77	307.669
2		0	2798.8	897.659	1038.83	1204.08	1136.33	1257.45	669.89	412.076
3			0	3695.11	3836.91	4002.69	3934.87	4056.14	2294.58	3195.92
4				0	142.439	309.506	241.73	363.263	1474.95	504.634
5					0	167.15	99.5455	220.902	1615.85	646.931
6						0	67.9119	53.7899	1782.2	814.076
7							0	121.634	1715.25	746.336
8								0	1835.64	867.823
9									0	975.32
10										0

(ii) Manhattan distances for centroids in c1



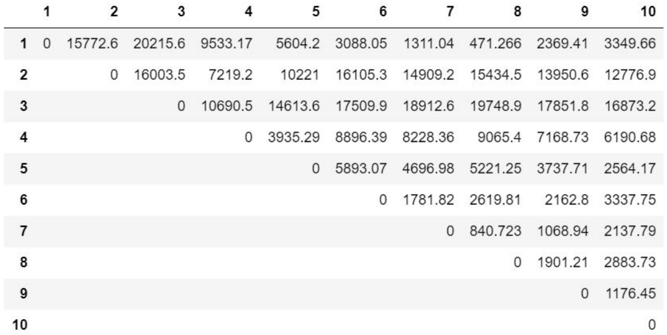
	1	2	3	4	5	6	7	8	9	10
1	0	728.924	3797.9	212.181	374.89	577.402	499.158	645.77	1731.06	406.701
2		0	3072.89	935.885	1100.83	1303.9	1225.35	1372.09	1005.29	490.928
3			0	4001.04	4170.3	4372.79	4294.95	4440.72	2513.42	3396.42
4				0	171.365	375.248	296.255	443.498	1934.09	609.749
5					0	204.523	125.597	272.935	2102.86	779.397
6						0	79.4017	69.5899	2306.38	983.02
7							0	147.866	2227.56	904.37
8								0	2374.55	1050.92
9									0	1327.58
10										0

(iii) Euclidean distances for centroids in c2



	1	2	3	4	5	6	7	8	9	10
1	0	15760.1	14110.8	9045.32	5567.68	1924.62	1100.86	402.891	2105.44	3169
2		0	11524.5	6743.88	10192.5	14455.1	14682.5	15362.4	13674.7	12597
3			0	9545.88	10883.4	12234	13208	13786.5	12509	11938.4
4				0	3494.22	7718.22	7957.78	8644.81	6947.82	5876.33
5					0	4404.56	4492.46	5169.94	3488.16	2407.92
6						0	1182.86	1615.79	1313.33	2153.77
7							0	698.488	1010.2	2085.46
8								0	1702.79	2768.61
9									0	1080.53
10										0

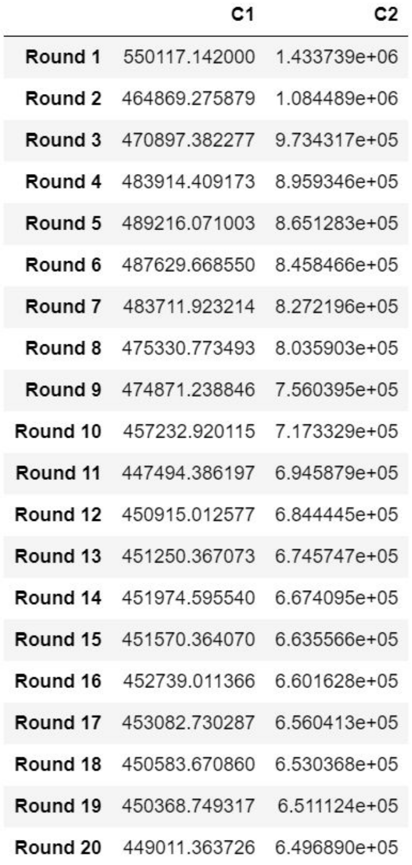
(iv) Manhattan distances for centroids in c2



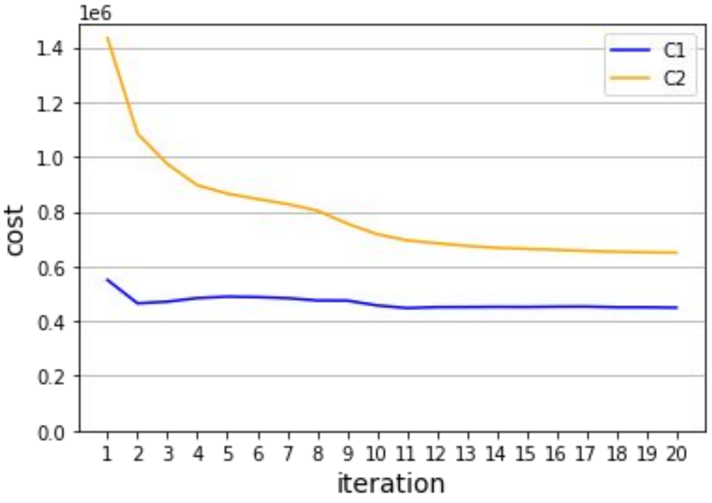
	1	2	3	4	5	6	7	8	9	10
1	0	15772.6	20215.6	9533.17	5604.2	3088.05	1311.04	471.266	2369.41	3349.66
2		0	16003.5	7219.2	10221	16105.3	14909.2	15434.5	13950.6	12776.9
3			0	10690.5	14613.6	17509.9	18912.6	19748.9	17851.8	16873.2
4				0	3935.29	8896.39	8228.36	9065.4	7168.73	6190.68
5					0	5893.07	4696.98	5221.25	3737.71	2564.17
6						0	1781.82	2619.81	2162.8	3337.75
7							0	840.723	1068.94	2137.79
8								0	1901.21	2883.73
9									0	1176.45
10										0

(b) Manhattan distance

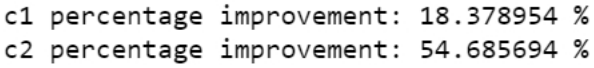
1. Plot of cost vs. iteration



	C1	C2
Round 1	550117.142000	1.433739e+06
Round 2	464869.275879	1.084489e+06
Round 3	470897.382277	9.734317e+05
Round 4	483914.409173	8.959346e+05
Round 5	489216.071003	8.651283e+05
Round 6	487629.668550	8.458466e+05
Round 7	483711.923214	8.272196e+05
Round 8	475330.773493	8.035903e+05
Round 9	474871.238846	7.560395e+05
Round 10	457232.920115	7.173329e+05
Round 11	447494.386197	6.945879e+05
Round 12	450915.012577	6.844445e+05
Round 13	451250.367073	6.745747e+05
Round 14	451974.595540	6.674095e+05
Round 15	451570.364070	6.635566e+05
Round 16	452739.011366	6.601628e+05
Round 17	453082.730287	6.560413e+05
Round 18	450583.670860	6.530368e+05
Round 19	450368.749317	6.511124e+05
Round 20	449011.363726	6.496890e+05



2. Percentage improvement and explanation

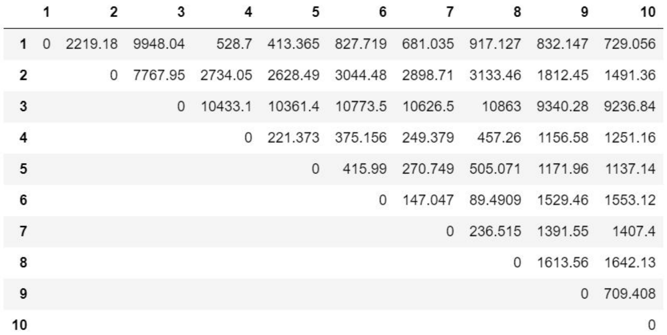


```
c1 percentage improvement: 18.378954 %
c2 percentage improvement: 54.685694 %
```

使用 Manhattan distance 做為 cost function 時，**C1 (random initialization)的表現會比 C2 (as far as possible)更好**。原因是相較於 Euclidean distance，Manhattan distance 對於離群值的敏感度較低，也就是若有一群較遠的 data points，centroids 不會積極地趨向這些 data points。但若使用 C2 進行初始化，centroid 會有較 C1 更高的機率被分配到離群的 cluster 中(因為要離群表示距離夠大)，於是 centroid 就有可能會被這些離群值卡住，陷入 local minimum。相對地，若使用 C1 進行初始化，各個 centroids 在 data points 之間理論上會是 uniformly distributed，且 centroids 並不會積極地去解決離群值的問題，而在 Manhattan distance 的度量下，這種情況可以讓大部分點的距離都變短，於是使 cost 極趨近於最佳解。

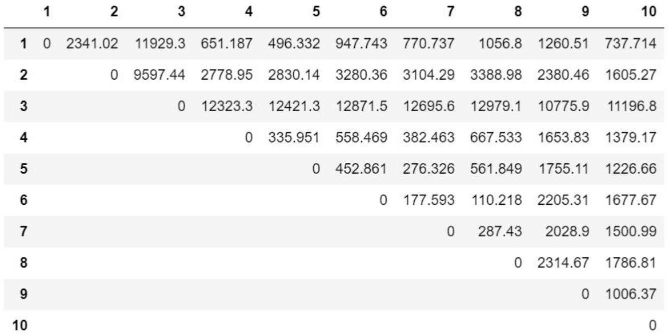
3. Distance of centroids

(i) Euclidean distances for centroids in c1



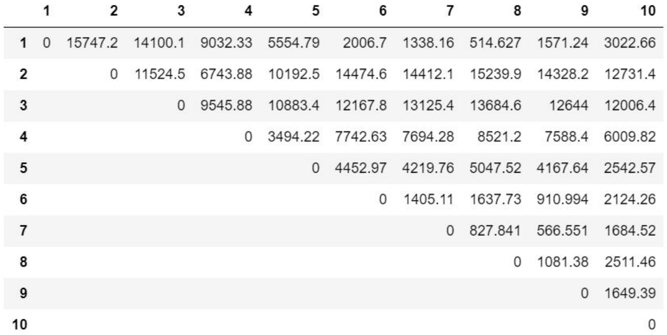
	1	2	3	4	5	6	7	8	9	10
1	0	2219.18	9948.04	528.7	413.365	827.719	681.035	917.127	832.147	729.056
2		0	7767.95	2734.05	2628.49	3044.48	2898.71	3133.46	1812.45	1491.36
3			0	10433.1	10361.4	10773.5	10626.5	10863	9340.28	9236.84
4				0	221.373	375.156	249.379	457.26	1156.58	1251.16
5					0	415.99	270.749	505.071	1171.96	1137.14
6						0	147.047	89.4909	1529.46	1553.12
7							0	236.515	1391.55	1407.4
8								0	1613.56	1642.13
9									0	709.408
10										0

(ii) Manhattan distances for centroids in c1



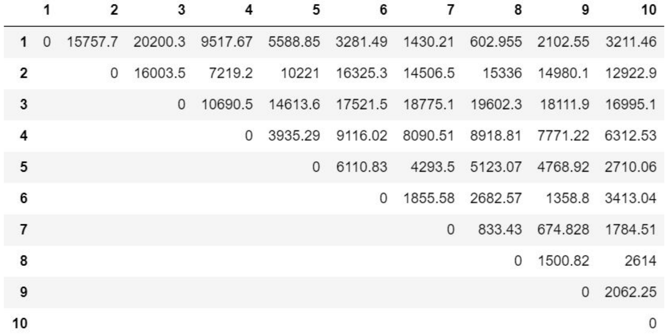
	1	2	3	4	5	6	7	8	9	10
1	0	2341.02	11929.3	651.187	496.332	947.743	770.737	1056.8	1260.51	737.714
2		0	9597.44	2778.95	2830.14	3280.36	3104.29	3388.98	2380.46	1605.27
3			0	12323.3	12421.3	12871.5	12695.6	12979.1	10775.9	11196.8
4				0	335.951	558.469	382.463	667.533	1653.83	1379.17
5					0	452.861	276.326	561.849	1755.11	1226.66
6						0	177.593	110.218	2205.31	1677.67
7							0	287.43	2028.9	1500.99
8								0	2314.67	1786.81
9									0	1006.37
10										0

(iii) Euclidean distances for centroids in c2



	1	2	3	4	5	6	7	8	9	10
1	0	15747.2	14100.1	9032.33	5554.79	2006.7	1338.16	514.627	1571.24	3022.66
2		0	11524.5	6743.88	10192.5	14474.6	14412.1	15239.9	14328.2	12731.4
3			0	9545.88	10883.4	12167.8	13125.4	13684.6	12644	12006.4
4				0	3494.22	7742.63	7694.28	8521.2	7588.4	6009.82
5					0	4452.97	4219.76	5047.52	4167.64	2542.57
6						0	1405.11	1637.73	910.994	2124.26
7							0	827.841	566.551	1684.52
8								0	1081.38	2511.46
9									0	1649.39
10										0

(iv) Manhattan distances for centroids in c2



	1	2	3	4	5	6	7	8	9	10
1	0	15757.7	20200.3	9517.67	5588.85	3281.49	1430.21	602.955	2102.55	3211.46
2		0	16003.5	7219.2	10221	16325.3	14506.5	15336	14980.1	12922.9
3			0	10690.5	14613.6	17521.5	18775.1	19602.3	18111.9	16995.1
4				0	3935.29	9116.02	8090.51	8918.81	7771.22	6312.53
5					0	6110.83	4293.5	5123.07	4768.92	2710.06
6						0	1855.58	2682.57	1358.8	3413.04
7							0	833.43	674.828	1784.51
8								0	1500.82	2614
9									0	2062.25
10										0