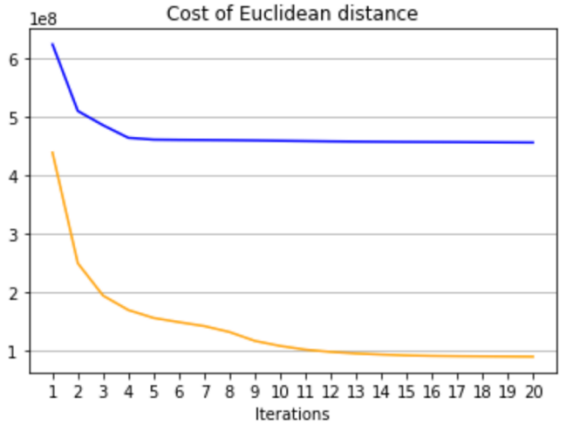
MDA_HW3_KMeans

107062103 王依婷

(a) with Euclidean distance

1. A plot of cost、iteration for 2 initialization strategies(c1 and c2).

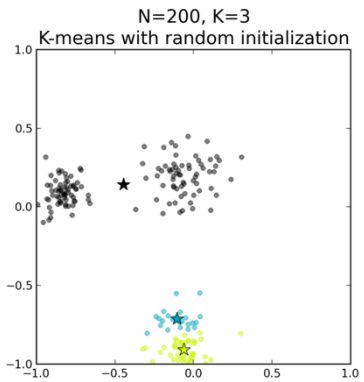


	c1	c2
0	6.236603e+08	4.387478e+08
1	5.098629e+08	2.498039e+08
2	4.854807e+08	1.944948e+08
3	4.639970e+08	1.698048e+08
4	4.609693e+08	1.562957e+08
5	4.605378e+08	1.490942e+08
6	4.603131e+08	1.425085e+08
7	4.600035e+08	1.323039e+08
8	4.595705e+08	1.171710e+08
9	4.590211e+08	1.085474e+08
10	4.584907e+08	1.022372e+08
11	4.579442e+08	9.827802e+07
12	4.575580e+08	9.563023e+07
13	4.572901e+08	9.379331e+07
14	4.570506e+08	9.237713e+07
15	4.568922e+08	9.154161e+07
16	4.567036e+08	9.104557e+07
17	4.564042e+08	9.075224e+07
18	4.561778e+08	9.047017e+07
19	4.559869e+08	9.021642e+07

2. Percentage improvement values and explanation.

```
Improvement of c1 with Euclidean distance:  26.885383 %
Improvement of c2 with Euclidean distance:  79.437750 %
```

c1 中的 centroid 是隨機挑選，所以 c2 相較於 c1 平均來說，centroid 的分布比較分散在所有資料點中。而使用 Euclidean distance，就是單純計算點與點之間的直線距離，如果 centroid 初始的位置不好，最後分完的結果可能會和實際的分布差異很大。所以在這種情況下使用 c2 能獲得比較好的結果。c1 的 cost 都比較大也是因為如此。



3. The Distances for all pairs of centroids.

- Euclidean distances for all pairs of centroids, with c1.

Euclidean distances for all pairs of c1_E

Euclidean	1	2	3	4	5	6	7	8	9	10
1	0.000000	692.157887	3490.258640	205.750279	346.718823	512.612247	444.731001	566.201992	1282.770845	307.669128
2		0.000000	2798.801053	897.658986	1038.826888	1204.078199	1136.327344	1257.449528	669.890228	412.076077
3			0.000000	3695.114191	3836.906638	4002.689083	3934.871559	4056.135573	2294.579642	3195.923901
4				0.000000	142.438874	309.506324	241.730115	363.262895	1474.945421	504.634116
5					0.000000	167.149800	99.545543	220.901784	1615.852353	646.930564
6						0.000000	67.911861	53.789891	1782.203049	814.076150
7							0.000000	121.633720	1715.253200	746.335559
8								0.000000	1835.639672	867.823079
9									0.000000	975.320423
10										0.000000

- Manhattan distances for all pairs of centroids, with c1

Manhattan distances for all pairs of c1_E

Manhattan	1	2	3	4	5	6	7	8	9	10
1	0.000000	728.924314	3797.899078	212.181090	374.890422	577.402076	499.157894	645.769777	1731.064307	406.701225
2		0.000000	3072.888690	935.885338	1100.833091	1303.895723	1225.351713	1372.092205	1005.293046	490.928058
3			0.000000	4001.038052	4170.304533	4372.788719	4294.952834	4440.719768	2513.422660	3396.420003
4				0.000000	171.365154	375.247921	296.254724	443.498445	1934.086960	609.749322
5					0.000000	204.522924	125.596786	272.934913	2102.864923	779.397227
6						0.000000	79.401684	69.589876	2306.380251	983.019681
7							0.000000	147.865709	2227.555857	904.370250
8								0.000000	2374.545430	1050.916221
9									0.000000	1327.583980
10										0.000000

- Euclidean distances for all pairs of centroids, with c2.

Euclidean distances for all pairs of c2_E

Euclidean	1	2	3	4	5	6	7	8	9	10
1	0.000000	15760.122472	14110.834391	9045.320235	5567.684524	1924.624082	1100.859050	402.890550	2105.442576	3169.003773
2		0.000000	11524.505650	6743.884100	10192.525007	14455.119372	14682.450993	15362.417961	13674.707531	12597.039560
3			0.000000	9545.879403	10883.382188	12233.959805	13208.002934	13786.484183	12508.957381	11938.376127
4				0.000000	3494.222416	7718.222010	7957.775949	8644.807041	6947.820636	5876.330200
5					0.000000	4404.562591	4492.458214	5169.937291	3488.158519	2407.918794
6						0.000000	1182.864189	1615.788236	1313.327493	2153.771472
7							0.000000	698.488136	1010.197665	2085.460676
8								0.000000	1702.792658	2768.607719
9									0.000000	1080.534944
10										0.000000

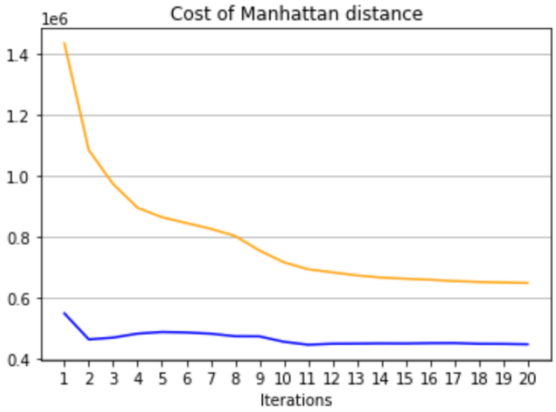
- Manhattan distances for all pairs of centroids, with c2.

Manhattan distances for all pairs of c2_E

Manhattan	1	2	3	4	5	6	7	8	9	10
1	0.000000	15772.614900	20215.645980	9533.170849	5604.200489	3088.054318	1311.039157	471.265720	2369.412159	3349.657086
2		0.000000	16003.499000	7219.196667	10221.031000	16105.347500	14909.169511	15434.460041	13950.575945	12776.883065
3			0.000000	10690.484333	14613.552000	17509.902750	18912.605411	19748.935693	17851.806836	16873.243674
4				0.000000	3935.292667	8896.389208	8228.355075	9065.404333	7168.732964	6190.679312
5					0.000000	5893.070125	4696.975382	5221.252806	3737.707000	2564.170543
6						0.000000	1781.822671	2619.811386	2162.802145	3337.746261
7							0.000000	840.722524	1068.939972	2137.788257
8								0.000000	1901.208756	2883.734537
9									0.000000	1176.450426
10										0.000000

(b) with Manhattan distance

1. A plot of cost、iteration for 2 initialization strategies(c1 and c2).



	c1	c2
0	550117.142000	1.433739e+06
1	464869.275879	1.084489e+06
2	470897.382277	9.734317e+05
3	483914.409173	8.959346e+05
4	489216.071003	8.651283e+05
5	487629.668550	8.458466e+05
6	483711.923214	8.272196e+05
7	475330.773493	8.035903e+05
8	474871.238846	7.560395e+05
9	457232.920115	7.173329e+05
10	447494.386197	6.945879e+05
11	450915.012577	6.844445e+05
12	451250.367073	6.745747e+05
13	451974.595540	6.674095e+05
14	451570.364070	6.635566e+05
15	452739.011366	6.601628e+05
16	453082.730287	6.560413e+05
17	450583.670860	6.530368e+05
18	450368.749317	6.511124e+05
19	449011.363726	6.496890e+05

2. Percentage improvement values and explanation.

```
Improvement of c1 with Manhattan distance:  18.378954 %
Improvement of c2 with Manhattan distance:  54.685694 %
```

可以發現用 Manhattan distance 計算時，c2 的 improvement 也比 c1 的好，原因和用 Euclidean distance 計算時一樣，因為初始時的 centroid 分佈較分散，重疊到的機會較少。但是在 c2 的情況下，整體的 cost 都比 c1 還要來得高，因為當初在產生 c2 時是以一般計算距離的方式算出來的，也就是 Euclidean distance，但是這樣的 centroid 不一定在 Euclidean distance 的計算之下也是如此。

3. The Distances for all pairs of centroids.

- Euclidean distances for all pairs of centroids, with c1.

Euclidean distances for all pairs of c1_M

Euclidean	1	2	3	4	5	6	7	8	9	10
1	0.000000	2219.177277	9948.044078	528.699758	413.365061	827.718886	681.034990	917.127383	832.147434	729.056349
2		0.000000	7767.945603	2734.049854	2628.490810	3044.477872	2898.712894	3133.460130	1812.454574	1491.357346
3			0.000000	10433.061351	10361.367486	10773.530838	10626.488597	10862.965776	9340.275232	9236.840022
4				0.000000	221.372794	375.156188	249.379188	457.259653	1156.583376	1251.158346
5					0.000000	415.989985	270.748792	505.071067	1171.964206	1137.135266
6						0.000000	147.046974	89.490917	1529.464012	1553.123807
7							0.000000	236.514622	1391.550421	1407.404400
8								0.000000	1613.555789	1642.128687
9									0.000000	709.407786
10										0.000000

- Manhattan distances for all pairs of centroids, with c1

Manhattan distances for all pairs of c1_M

Manhattan	1	2	3	4	5	6	7	8	9	10
1	0.000000	2341.017219	11929.300152	651.187488	496.331521	947.743236	770.737383	1056.799501	1260.510561	737.713573
2		0.000000	9597.441187	2778.945762	2830.144528	3280.359168	3104.285771	3388.982648	2380.460958	1605.270129
3			0.000000	12323.287569	12421.263080	12871.483429	12695.554202	12979.133180	10775.939186	11196.786982
4				0.000000	335.951213	558.469258	382.463330	667.533230	1653.825887	1379.165173
5					0.000000	452.861331	276.326491	561.849249	1755.105533	1226.660355
6						0.000000	177.593162	110.217624	2205.307383	1677.666864
7							0.000000	287.429708	2028.901616	1500.993410
8								0.000000	2314.667455	1786.811316
9									0.000000	1006.367826
10										0.000000

- Euclidean distances for all pairs of centroids, with c2.

Euclidean distances for all pairs of c2_M

Euclidean	1	2	3	4	5	6	7	8	9	10
1	0.000000	15747.234226	14100.144688	9032.333023	5554.786693	2006.702668	1338.161126	514.627038	1571.243420	3022.660884
2		0.000000	11524.505650	6743.884100	10192.525007	14474.554116	14412.056615	15239.877071	14328.226192	12731.397635
3			0.000000	9545.879403	10883.382188	12167.793871	13125.351004	13684.606757	12643.985638	12006.394618
4				0.000000	3494.222416	7742.628117	7694.276701	8521.197863	7588.404540	6009.820223
5					0.000000	4452.971685	4219.760574	5047.516256	4167.636533	2542.569354
6						0.000000	1405.109080	1637.729438	910.994388	2124.263362
7							0.000000	827.840658	566.551017	1684.516012
8								0.000000	1081.379335	2511.458859
9									0.000000	1649.389172
10										0.000000

- Manhattan distances for all pairs of centroids, with c2.

Manhattan distances for all pairs of c2_M

Manhattan	1	2	3	4	5	6	7	8	9	10
1	0.000000	15757.691264	20200.259436	9517.668233	5588.853635	3281.488247	1430.208678	602.954849	2102.553978	3211.455756
2		0.000000	16003.499000	7219.196667	10221.031000	16325.270500	14506.485890	15335.957403	14980.056096	12922.931357
3			0.000000	10690.484333	14613.552000	17521.517667	18775.121461	19602.262815	18111.885425	16995.133536
4				0.000000	3935.292667	9116.024500	8090.510188	8918.813117	7771.222078	6312.530012
5					0.000000	6110.832500	4293.501903	5123.066808	4768.923000	2710.056500
6						0.000000	1855.579909	2682.569234	1358.795895	3413.036179
7							0.000000	833.430282	674.827570	1784.512045
8								0.000000	1500.824884	2613.997305
9									0.000000	2062.251068
10										0.000000