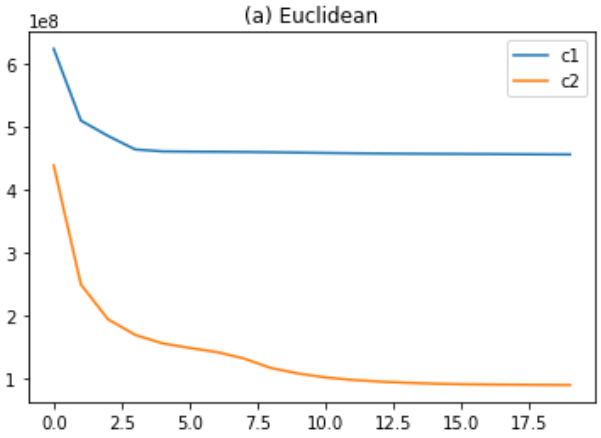
Introduction to Massive Data Analysis

HW3 - K-means Report

106070038 杜葳葳

(a) Initialization strategies with **Euclidean** distance

- A plot of cost vs iteration



	C1	C2
Round 1	623660345	438747790
Round 2	509862908.3	249803933.6
Round 3	485480682	194494814.4
Round 4	463997011.7	169804841.5
Round 5	460969266.6	156295748.8
Round 6	460537848	149094208.1
Round 7	460313099.7	142508531.6
Round 8	460003523.9	132303869.4
Round 9	459570539.3	117170969.8
Round 10	459021103.3	108547377.2
Round 11	458490656.2	102237203.3
Round 12	457944232.6	98278015.75
Round 13	457558005.2	95630226.12
Round 14	457290136.4	93793314.05
Round 15	457050555.1	92377131.97
Round 16	456892235.6	91541606.25
Round 17	456703630.7	91045573.83
Round 18	456404203	90752240.1
Round 19	456177800.5	90470170.18
Round 20	455986871	90216416.18

- Percentage improvement values and explanation
  - Euclidean
    - C1: 26.885%
    - C2: **79.438%**
  - Explanation
    - c2 的表現比 c1 佳，**初始選擇距離較遠的點作為 cluster 的中心相較隨機選取來的理想**，推測如果是隨機選取（c1），可能會選到過於靠近的點，導致在過程中掉入區域最佳解、而非全域的最佳解。

- The Euclidean and Manhattan Distances for all pairs of centroids, with 2 initialization strategies.

- Manhattan – c1

Manhattan	1	2	3	4	5	6	7	8	9	10
1	0	728.924	3797.899	212.181	374.89	577.402	499.158	645.77	1731.064	406.701
2		0	3072.889	935.885	1100.833	1303.896	1225.352	1372.092	1005.293	490.928
3			0	4001.038	4170.305	4372.789	4294.953	4440.72	2513.423	3396.42
4				0	171.365	375.248	296.255	443.498	1934.087	609.749
5					0	204.523	125.597	272.935	2102.865	779.397
6						0	79.402	69.59	2306.38	983.02
7							0	147.866	2227.556	904.37
8								0	2374.545	1050.916
9									0	1327.584
10										0

- Euclidean – c1

Euclidean	1	2	3	4	5	6	7	8	9	10
1	0	692.158	3490.259	205.75	346.719	512.612	444.731	566.202	1282.771	307.669
2		0	2798.801	897.659	1038.827	1204.078	1136.327	1257.45	669.89	412.076
3			0	3695.114	3836.907	4002.689	3934.872	4056.136	2294.58	3195.924
4				0	142.439	309.506	241.73	363.263	1474.945	504.634
5					0	167.15	99.546	220.902	1615.852	646.931
6						0	67.912	53.79	1782.203	814.076
7							0	121.634	1715.253	746.336
8								0	1835.64	867.823
9									0	975.32
10										0

- Manhattan – c2

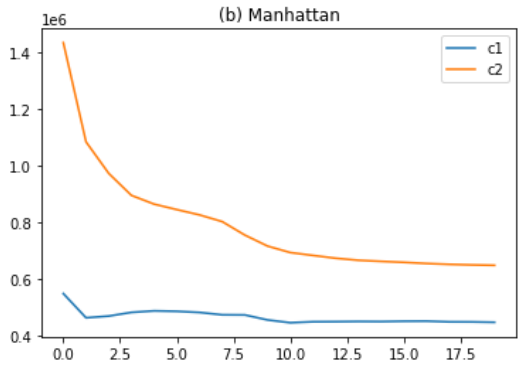
Manhattan	1	2	3	4	5	6	7	8	9	10
1	0	15772.615	20215.646	9533.171	5604.2	3088.054	1311.039	471.266	2369.412	3349.657
2		0	16003.499	7219.197	10221.031	16105.348	14909.17	15434.46	13950.576	12776.883
3			0	10690.484	14613.552	17509.903	18912.605	19748.936	17851.807	16873.244
4				0	3935.293	8896.389	8228.355	9065.404	7168.733	6190.679
5					0	5893.07	4696.975	5221.253	3737.707	2564.171
6						0	1781.823	2619.811	2162.802	3337.746
7							0	840.723	1068.94	2137.788
8								0	1901.209	2883.735
9									0	1176.45
10										0

- Euclidean – c2

Euclidean	1	2	3	4	5	6	7	8	9	10
1	0	15760.122	14110.834	9045.32	5567.685	1924.624	1100.859	402.891	2105.443	3169.004
2		0	11524.506	6743.884	10192.525	14455.119	14682.451	15362.418	13674.708	12597.04
3			0	9545.879	10883.382	12233.96	13208.003	13786.484	12508.957	11938.376
4				0	3494.222	7718.222	7957.776	8644.807	6947.821	5876.33
5					0	4404.563	4492.458	5169.937	3488.159	2407.919
6						0	1182.864	1615.788	1313.327	2153.771
7							0	698.488	1010.198	2085.461
8								0	1702.793	2768.608
9									0	1080.535
10										0

(b) Initialization strategies with **Manhattan** distance

- A plot of cost vs iteration



	C1	C2
Round 1	550117.142	1433739.31
Round 2	464869.276	1084488.777
Round 3	470897.382	973431.715
Round 4	483914.409	895934.593
Round 5	489216.071	865128.335
Round 6	487629.669	845846.647
Round 7	483711.923	827219.583
Round 8	475330.773	803590.346
Round 9	474871.239	756039.517
Round 10	457232.92	717332.903
Round 11	447494.386	694587.925
Round 12	450915.013	684444.502
Round 13	451250.367	674574.748
Round 14	451974.596	667409.47
Round 15	451570.364	663556.628
Round 16	452739.011	660162.777
Round 17	453082.73	656041.322
Round 18	450583.671	653036.754
Round 19	450368.749	651112.426
Round 20	449011.364	649689.013

- Percentage improvement values and explanation
  - Manhattan
    - C1: 18.379%
    - C2: 54.686%
  - Explanation
    - C1 和 C2 相比，與(a)相同，跑多次 iteration 後 C2 皆比 C1 來的好，初始時隨

機選取容易誤掉到區域最佳解。

- 與(a)綜合比較，我認為**用 Euclidean 的結果較好**，C1 和 C2 的 improvement percentage 皆比較高，**推測是 Manhattan 在高維資料（尤其此次作業為 58 維）的處理上不太理想，用 Euclidean 處理較合適**。

- The Euclidean and Manhattan Distances for all pairs of centroids, with 2 initialization strategies.
  - Manhattan – c1

Manhattan	1	2	3	4	5	6	7	8	9	10
1	0	2341.017	11929.3	651.187	496.332	947.743	770.737	1056.8	1260.511	737.714
2		0	9597.441	2778.946	2830.145	3280.359	3104.286	3388.983	2380.461	1605.27
3			0	12323.288	12421.263	12871.483	12695.554	12979.133	10775.939	11196.787
4				0	335.951	558.469	382.463	667.533	1653.826	1379.165
5					0	452.861	276.326	561.849	1755.106	1226.66
6						0	177.593	110.218	2205.307	1677.667
7							0	287.43	2028.902	1500.993
8								0	2314.667	1786.811
9									0	1006.368
10										0

  - Euclidean – c1

Euclidean	1	2	3	4	5	6	7	8	9	10
1	0	2219.177	9948.044	528.7	413.365	827.719	681.035	917.127	832.147	729.056
2		0	7767.946	2734.05	2628.491	3044.478	2898.713	3133.46	1812.455	1491.357
3			0	10433.061	10361.367	10773.531	10626.489	10862.966	9340.275	9236.84
4				0	221.373	375.156	249.379	457.26	1156.583	1251.158
5					0	415.99	270.749	505.071	1171.964	1137.135
6						0	147.047	89.491	1529.464	1553.124
7							0	236.515	1391.55	1407.404
8								0	1613.556	1642.129

9									0	709.408
10										0

- Manhattan – c2

Manhattan	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	0	15757.691	20200.259	9517.668	5588.854	3281.488	1430.209	602.955	2102.554	3211.456
2		0	16003.499	7219.197	10221.031	16325.27	14506.486	15335.957	14980.056	12922.931
3			0	10690.484	14613.552	17521.518	18775.121	19602.263	18111.885	16995.134
4				0	3935.293	9116.024	8090.51	8918.813	7771.222	6312.53
5					0	6110.832	4293.502	5123.067	4768.923	2710.056
6						0	1855.58	2682.569	1358.796	3413.036
7							0	833.43	674.828	1784.512
8								0	1500.825	2613.997
9									0	2062.251
10										0

- Euclidean – c2

Euclidean	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	0	15747.234	14100.145	9032.333	5554.787	2006.703	1338.161	514.627	1571.243	3022.661
2		0	11524.506	6743.884	10192.525	14474.554	14412.057	15239.877	14328.226	12731.398
3			0	9545.879	10883.382	12167.794	13125.351	13684.607	12643.986	12006.395
4				0	3494.222	7742.628	7694.277	8521.198	7588.405	6009.82
5					0	4452.972	4219.761	5047.516	4167.637	2542.569
6						0	1405.109	1637.729	910.994	2124.263
7							0	827.841	566.551	1684.516
8								0	1081.379	2511.459
9									0	1649.389
10										0