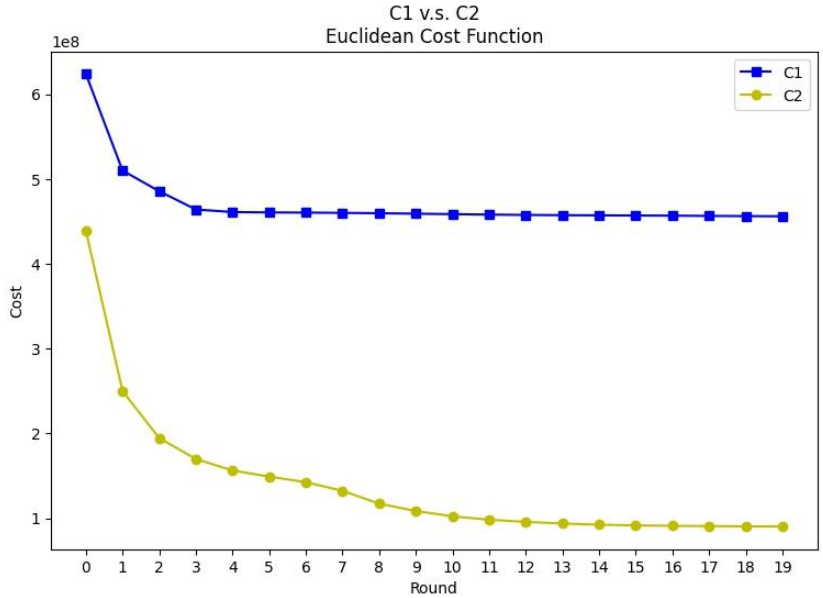
k-means on MapReduce

105072123 黃海茵

(a) Euclidean

	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



- Percentage improvement：C1 是 26.885%，C2 是 79.438%。
  C2 各點相距較遠，所以 data 在最初的距離會較 C1 來的大，因此 C2 的 percentage improvement 會較大。使用 Euclidean 時，是計算平方和開根號，也就是兩點間的距離，所以 centroid 相距較遠，cluster 分布會較平均，每個點到 centroid 的距離也會越小，所以 C2 的 cost 會小於 C1。

- C1 Euclidean

Euclidean	1	2	3	4	5	6	7	8	9	10
1	0	692.1579	3490.259	205.7503	346.7188	512.6122	444.731	566.202	1282.771	307.6691
2		0	2798.801	897.659	1038.827	1204.078	1136.327	1257.45	669.8902	412.0761
3			0	3695.114	3836.907	4002.689	3934.872	4056.136	2294.58	3195.924
4				0	142.4389	309.5063	241.7301	363.2629	1474.945	504.6341
5					0	167.1498	99.54554	220.9018	1615.852	646.9306
6						0	67.91186	53.78989	1782.203	814.0762
7							0	121.6337	1715.253	746.3356
8								0	1835.64	867.8231
9									0	975.3204
10										0

- C1 Manhattan

Manhattan	1	2	3	4	5	6	7	8	9	10
1	0	728.9243	3797.899	212.1811	374.8904	577.4021	499.1579	645.7698	1731.064	406.7012
2		0	3072.889	935.8853	1100.833	1303.896	1225.352	1372.092	1005.293	490.9281
3			0	4001.038	4170.305	4372.789	4294.953	4440.72	2513.423	3396.42
4				0	171.3652	375.2479	296.2547	443.4984	1934.087	609.7493
5					0	204.5229	125.5968	272.9349	2102.865	779.3972
6						0	79.40168	69.58988	2306.38	983.0197
7							0	147.8657	2227.556	904.3703
8								0	2374.545	1050.916
9									0	1327.584
10										0

- C2 Euclidean

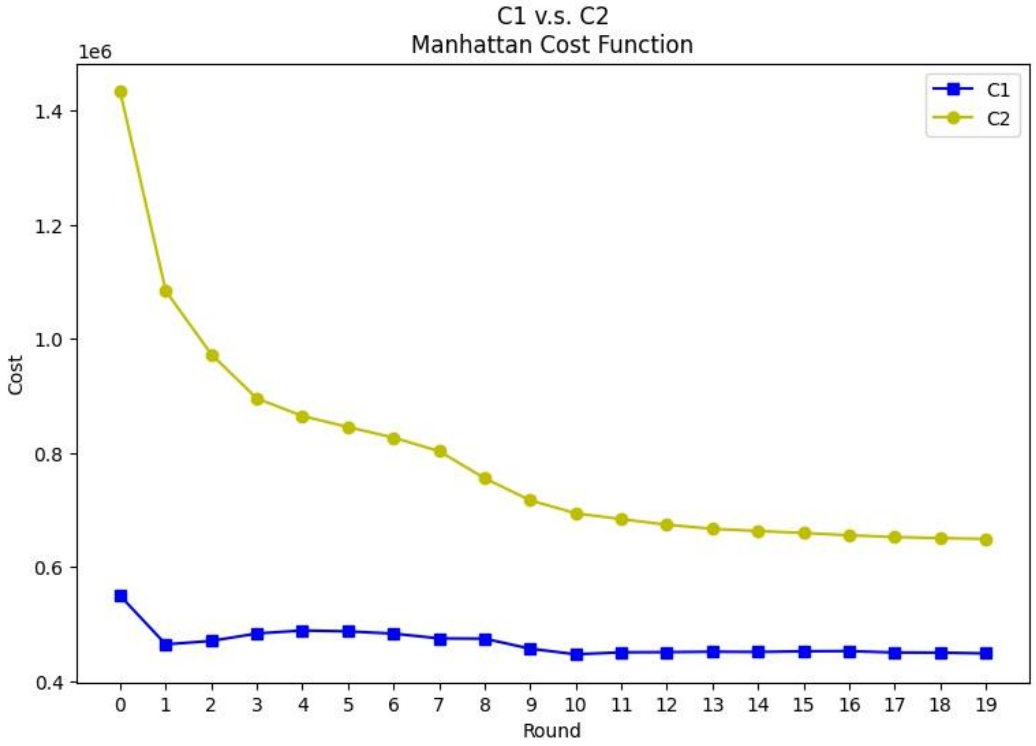
Euclidean	1	2	3	4	5	6	7	8	9	10
1	0	15760.12	14110.83	9045.32	5567.685	1924.624	1100.859	402.8905	2105.443	3169.004
2		0	11524.51	6743.884	10192.53	14455.12	14682.45	15362.42	13674.71	12597.04
3			0	9545.879	10883.38	12233.96	13208	13786.48	12508.96	11938.38
4				0	3494.222	7718.222	7957.776	8644.807	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	698.4881	1010.198	2085.461
8								0	1702.793	2768.608
9									0	1080.535
10										0

- C2 Manhattan

Manhattan	1	2	3	4	5	6	7	8	9	10
1	0	15772.61	20215.65	9533.171	5604.2	3088.054	1311.039	471.2657	2369.412	3349.657
2		0	16003.5	7219.197	10221.03	16105.35	14909.17	15434.46	13950.58	12776.88
3			0	10690.48	14613.55	17509.9	18912.61	19748.94	17851.81	16873.24
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	840.7225	1068.94	2137.788
8								0	1901.209	2883.735
9									0	1176.45
10										0

(b) Manhattan

	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



- Percentage improvement：C1 是 18.379%，C2 是 54.686%。
  同(a)小題，C2 各點相距較遠，所以 data 在最初的距離會較 C1 來的大，因此 C2 的 percentage improvement 會較大。使用 Manhattan 時，是計算每個維度距離的總和，而非兩點間的距離，所以不一定會比較小。

- C1 Euclidean

Euclidean	1	2	3	4	5	6	7	8	9	10
1	0	2219.177	9948.044	528.6998	413.3651	827.7189	681.035	917.1274	832.1474	729.0563
2		0	7767.946	2734.05	2628.491	3044.478	2898.713	3133.46	1812.455	1491.357
3			0	10433.06	10361.37	10773.53	10626.49	10862.97	9340.275	9236.84
4				0	221.3728	375.1562	249.3792	457.2597	1156.583	1251.158
5					0	415.99	270.7488	505.0711	1171.964	1137.135
6						0	147.047	89.49092	1529.464	1553.124
7							0	236.5146	1391.55	1407.404
8								0	1613.556	1642.129
9									0	709.4078
10										0

- C1 Manhattan

Manhattan	1	2	3	4	5	6	7	8	9	10
1	0	2341.017	11929.3	651.1875	496.3315	947.7432	770.7374	1056.8	1260.511	737.7136
2		0	9597.441	2778.946	2830.145	3280.359	3104.286	3388.983	2380.461	1605.27
3			0	12323.29	12421.26	12871.48	12695.55	12979.13	10775.94	11196.79
4				0	335.9512	558.4693	382.4633	667.5332	1653.826	1379.165
5					0	452.8613	276.3265	561.8492	1755.106	1226.66
6						0	177.5932	110.2176	2205.307	1677.667
7							0	287.4297	2028.902	1500.993
8								0	2314.667	1786.811
9									0	1006.368
10										0

- C2 Euclidean

Euclidean	1	2	3	4	5	6	7	8	9	10
1	0	15747.23	14100.14	9032.333	5554.787	2006.703	1338.161	514.627	1571.243	3022.661
2		0	11524.51	6743.884	10192.53	14474.55	14412.06	15239.88	14328.23	12731.4
3			0	9545.879	10883.38	12167.79	13125.35	13684.61	12643.99	12006.39
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	910.9944	2124.263
7							0	827.8407	566.551	1684.516
8								0	1081.379	2511.459
9									0	1649.389
10										0

- C2 Manhattan

Manhattan	1	2	3	4	5	6	7	8	9	10
1	0	15757.69	20200.26	9517.668	5588.854	3281.488	1430.209	602.9548	2102.554	3211.456
2		0	16003.5	7219.197	10221.03	16325.27	14506.49	15335.96	14980.06	12922.93
3			0	10690.48	14613.55	17521.52	18775.12	19602.26	18111.89	16995.13
4				0	3935.293	9116.025	8090.51	8918.813	7771.222	6312.53
5					0	6110.833	4293.502	5123.067	4768.923	2710.056
6						0	1855.58	2682.569	1358.796	3413.036
7							0	833.4303	674.8276	1784.512
8								0	1500.825	2613.997
9									0	2062.251
10										0