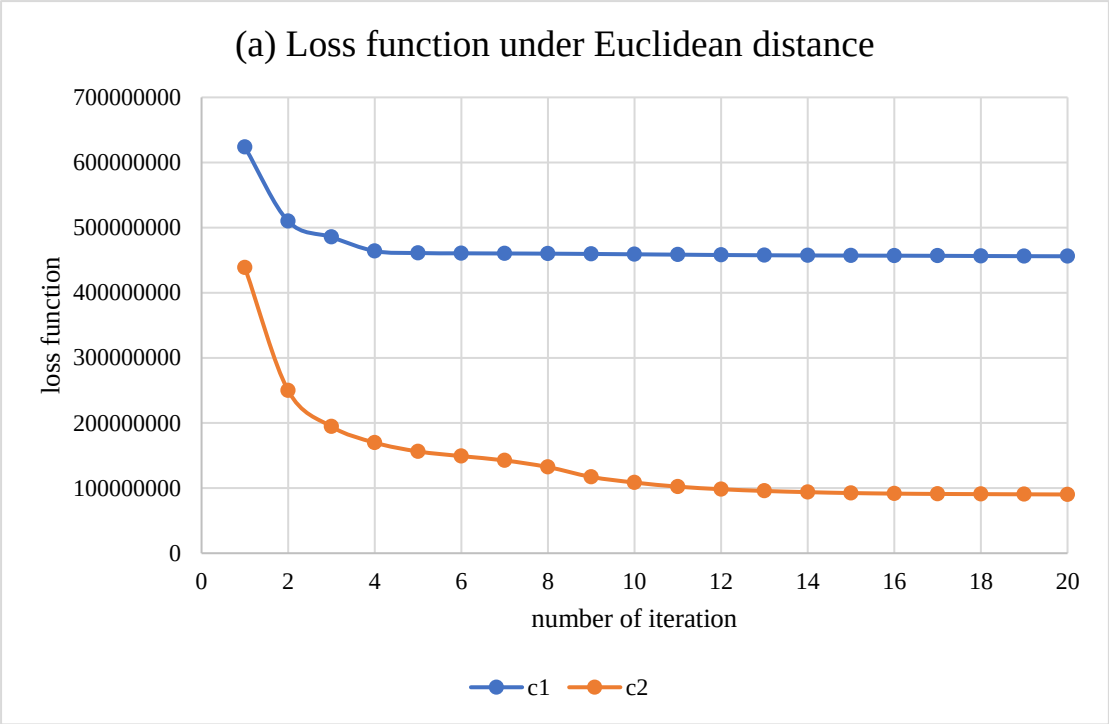
巨量資料分析導論 HW3:Kmeans

106061218 李丞恩

(a) Exploring initialization strategies with Euclidean distance

Cost v.s. iteration

Round		c1	c2
1	Round1	6.24E+08	4.39E+08
2	Round2	5.1E+08	2.5E+08
3	Round3	4.85E+08	1.94E+08
4	Round4	4.64E+08	1.7E+08
5	Round5	4.61E+08	1.56E+08
6	Round6	4.61E+08	1.49E+08
7	Round7	4.6E+08	1.43E+08
8	Round8	4.6E+08	1.32E+08
9	Round9	4.6E+08	1.17E+08
10	Round10	4.59E+08	1.09E+08
11	Round11	4.58E+08	1.02E+08
12	Round12	4.58E+08	98278016
13	Round13	4.58E+08	95630226
14	Round14	4.57E+08	93793314
15	Round15	4.57E+08	92377132
16	Round16	4.57E+08	91541606
17	Round17	4.57E+08	91045574
18	Round18	4.56E+08	90752240
19	Round19	4.56E+08	90470170
20	Round20	4.56E+08	90216416



Percentage improvement values: c1=0.268, c2=0.794

C2 的 Percentage improvement values 較佳，這是由於如果 initial centroid 隨機選取（如同 C1）時，可能會選到太靠近的 centroid，使目標函數容易收斂局部最小值而非全域最小值。

C1 內所有 centroid 的 Euclidean 距離

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

C2 內所有 centroid 的 Euclidean 距離

0	402.8905	15760.12	2105.443	14110.83	5567.685	1100.859	9045.32	3169.004	1924.624
	0	15362.42	1702.793	13786.48	5169.937	698.4881	8644.807	2768.608	1615.788
		0	13674.71	11524.51	10192.53	14682.45	6743.884	12597.04	14455.12
			0	12508.96	3488.159	1010.198	6947.821	1080.535	1313.327
				0	10883.38	13208	9545.879	11938.38	12233.96
					0	4492.458	3494.222	2407.919	4404.563
						0	7957.776	2085.461	1182.864
							0	5876.33	7718.222
								0	2153.771
									0

C1 內所有 centroid 的 Manhattan 距離

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

C2 內所有 centroid 的 Manhattan 距離

0	471.2657	15772.61	2369.412	20215.65	5604.2	1311.039	9533.171	3349.657	3088.054
	0	15434.46	1901.209	19748.94	5221.253	840.7225	9065.404	2883.735	2619.811
		0	13950.58	16003.5	10221.03	14909.17	7219.197	12776.88	16105.35
			0	17851.81	3737.707	1068.94	7168.733	1176.45	2162.802
				0	14613.55	18912.61	10690.48	16873.24	17509.9
					0	4696.975	3935.293	2564.171	5893.07
						0	8228.355	2137.788	1781.823
							0	6190.679	8896.389
								0	3337.746
									0

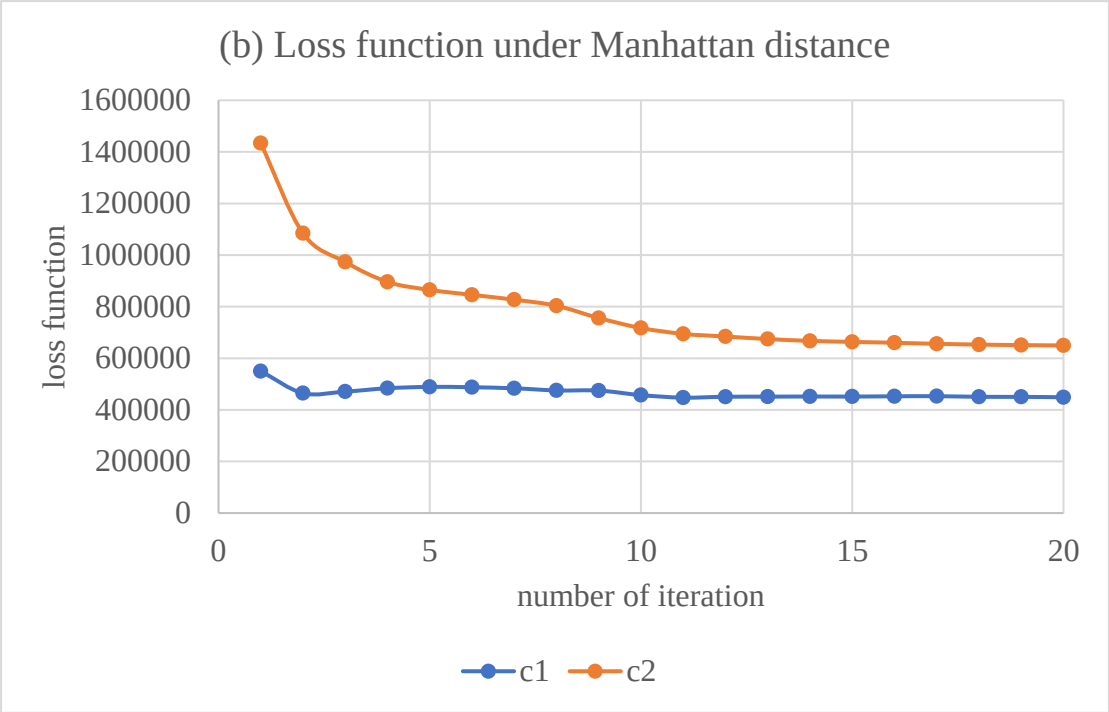
(b) Exploring initialization strategies with Manhattan distance

Cost v.s. iteration

	c1	c2
Round1	550117.1	1433739
Round2	464829.3	1084489
Round3	470934.2	973431.7
Round4	483874.8	895934.6
Round5	489234.2	865128.3
Round6	487664.7	845846.6
Round7	483718.7	827219.6
Round8	475337.9	803590.3
Round9	474872	756039.5
Round10	457244.8	717332.9
Round11	447493.2	694587.9
Round12	450891.8	684444.5
Round13	451232.6	674574.7
Round14	451860.1	667409.5
Round15	451567.2	663556.6
Round16	452710.1	660162.8
Round17	453078.2	656041.3
Round18	450646.1	653036.8
Round19	450420	651112.4
Round20	449009.6	649689

Percentage improvement values: c1=0.183, c2=0.546

同樣地，c2 的 Percentage improvement values 會較佳，理由與(a)小題相同。



C1 內所有 centroid 的 Euclidean 距離

0	236.5146	249.3792	147.047	270.1715	10626.49	2898.713	1391.55	680.1481	1407.404
	0	457.2597	89.49092	504.4996	10862.97	3133.46	1613.556	916.2348	1642.129
		0	375.1562	221.0468	10433.06	2734.05	1156.583	527.9658	1251.158
			0	415.4169	10773.53	3044.478	1529.464	826.8276	1553.124
				0	10361.92	2629.064	1172.382	413.0144	1137.709
					0	7767.946	9340.275	9948.957	9236.84
						0	1812.455	2220.037	1491.357
							0	832.9918	709.4078
								0	729.8969
									0

C2 內所有 centroid 的 Euclidean 距離

0	514.627	15747.23	1571.243	14100.14	5554.787	1338.161	9032.333	3022.661	2006.703
	0	15239.88	1081.379	13684.61	5047.516	827.8407	8521.198	2511.459	1637.729
		0	14328.23	11524.51	10192.53	14412.06	6743.884	12731.4	14474.55
			0	12643.99	4167.637	566.551	7588.405	1649.389	910.9944
				0	10883.38	13125.35	9545.879	12006.39	12167.79
					0	4219.761	3494.222	2542.569	4452.972
						0	7694.277	1684.516	1405.109

							0	6009.82	7742.628
								0	2124.263
									0

C1 內所有 centroid 的 Manhattan 距離

0	287.4297	382.4633	177.5932	275.7966	12695.55	3104.286	2028.902	769.4283	1500.993
	0	667.5332	110.2176	561.3195	12979.13	3388.983	2314.667	1055.49	1786.811
		0	558.4693	335.3379	12323.29	2778.946	1653.826	650.8044	1379.165
			0	452.3291	12871.48	3280.359	2205.307	946.434	1677.667
				0	12421.8	2830.684	1755.65	495.5646	1227.2
					0	9597.441	10775.94	11930.59	11196.79
						0	2380.461	2342.246	1605.27
							0	1261.792	1006.368
								0	738.9508
									0

C2 內所有 centroid 的 Manhattan 距離

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