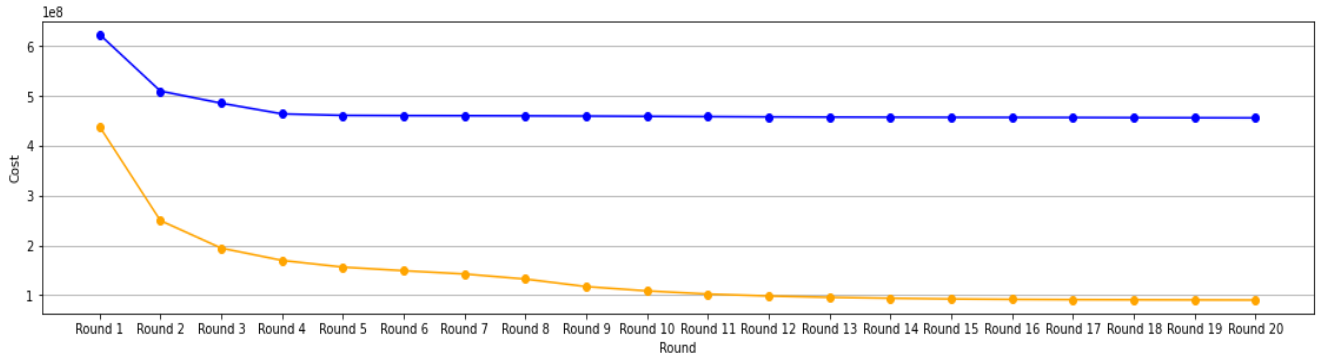
HW3_e

姓名: 湯睿哲

學號: 107062333

(a)

1. A plot of cost vs. iteration for 2 initialization strategies(c1 and c2) for (a)
   p.s. 1e8 代表 E+08
   c1: 藍線
   c2: 橘線



2. Percentage improvement values and explanation
   After 20 iterations:
   C1 improves by about 26.885%
   C2 improves by about 79.438%

   Explanation:
   From the graph and the percentage of C1 and C2, we can conclude that C2 is better than C1 because the initial 10 centroids of C2 is far apart to each other. As a result, the less overlap will happen when we process re-computing clusters iteratively, and that means the true clusters will be less split. And then, we will gain better final clusters sets within 20 iterations.

The Euclidean Distances for all pairs of centroids, with C1 and C2

C1:

	A	B	C	D	E	F	G	H	I	J	K
1	Euclidean	1	2	3	4	5	6	7	8	9	10
2	1	0.000	692.158	3490.259	205.750	346.719	512.612	444.731	566.202	1282.771	307.669
3	2		0.000	2798.801	897.659	1038.827	1204.078	1136.327	1257.450	669.890	412.076
4	3			0.000	3695.114	3836.907	4002.689	3934.872	4056.136	2294.580	3195.924
5	4				0.000	142.439	309.506	241.730	363.263	1474.945	504.634
6	5					0.000	167.150	99.546	220.902	1615.852	646.931
7	6						0.000	67.912	53.790	1782.203	814.076
8	7							0.000	121.634	1715.253	746.336
9	8								0.000	1835.640	867.823
10	9									0.000	975.320
11	10										0.000

C2:

	A	B	C	D	E	F	G	H	I	J	K
1	Euclidean	1	2	3	4	5	6	7	8	9	10
2	1	0.000	15760.122	14110.834	9045.320	5567.685	1924.624	1100.859	402.891	2105.443	3169.004
3	2		0.000	11524.506	6743.884	10192.525	14455.119	14682.451	15362.418	13674.708	12597.040
4	3			0.000	9545.879	10883.382	12233.960	13208.003	13786.484	12508.957	11938.376
5	4				0.000	3494.222	7718.222	7957.776	8644.807	6947.821	5876.330
6	5					0.000	4404.563	4492.458	5169.937	3488.159	2407.919
7	6						0.000	1182.864	1615.788	1313.327	2153.771
8	7							0.000	698.488	1010.198	2085.461
9	8								0.000	1702.793	2768.608
10	9									0.000	1080.535
11	10										0.000

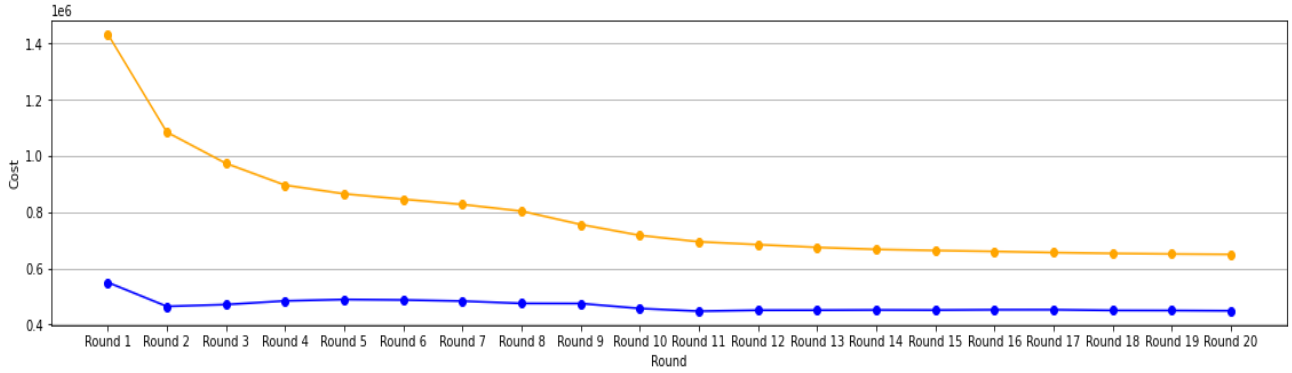
(b)

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

   p.s. 1e6 代表 E+06

   c1: 藍線

   c2: 橘線



2. Percentage improvement values and explanation

After 20 iterations:

C1 improves by about 18.394%

C2 improves by about 54.686%

Explanation:

from the percentage of C1 and C2, we just can conclude that C2 has better rate of reducing since its initial distribution. However, in terms of cost, C1 is better than C2 because C2's "far apart distribution" may use Euclidean distance metric to determine, which means it isn't necessarily far apart for Manhattan distance metric.

So, we can't say the far apart point in Euclidean distance metric, is also the far apart in Manhattan distance metric

The Manhattan Distances for all pairs of centroids, with C1 and C2

C1:

	A	B	C	D	E	F	G	H	I	J	K
1	Manhattar	1	2	3	4	5	6	7	8	9	10
2	1	0.000	2336.037	11924.080	652.125	499.653	952.467	775.805	1061.783	1255.347	732.704
3	2		0.000	9597.441	2777.742	2828.265	3279.855	3104.116	3388.744	2380.461	1605.270
4	3			0.000	12322.898	12419.434	12870.996	12695.389	12978.891	10775.939	11196.787
5	4				0.000	335.664	558.134	382.469	667.483	1653.445	1376.942
6	5					0.000	454.244	278.003	563.475	1753.198	1224.764
7	6						0.000	177.255	110.470	2204.800	1677.155
8	7							0.000	287.351	2028.739	1500.840
9	8								0.000	2314.427	1786.572
10	9									0.000	1006.368
11	10										0.000

C2:

	A	B	C	D	E	F	G	H	I	J	K
1	Manhattar	1	2	3	4	5	6	7	8	9	10
2	1	0.000	15757.691	20200.259	9517.668	5588.854	3281.488	1430.209	602.955	2102.554	3211.456
3	2		0.000	16003.499	7219.197	10221.031	16325.271	14506.486	15335.957	14980.056	12922.931
4	3			0.000	10690.484	14613.552	17521.518	18775.121	19602.263	18111.885	16995.134
5	4				0.000	3935.293	9116.024	8090.510	8918.813	7771.222	6312.530
6	5					0.000	6110.833	4293.502	5123.067	4768.923	2710.056
7	6						0.000	1855.580	2682.569	1358.796	3413.036
8	7							0.000	833.430	674.828	1784.512
9	8								0.000	1500.825	2613.997
10	9									0.000	2062.251
11	10										0.000

(c) Map/Reduce Explanation: 於 Kmeans.ipynb 的 markdon 裡