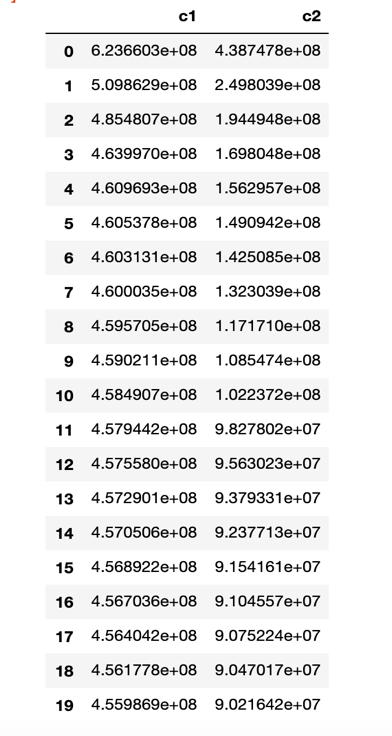
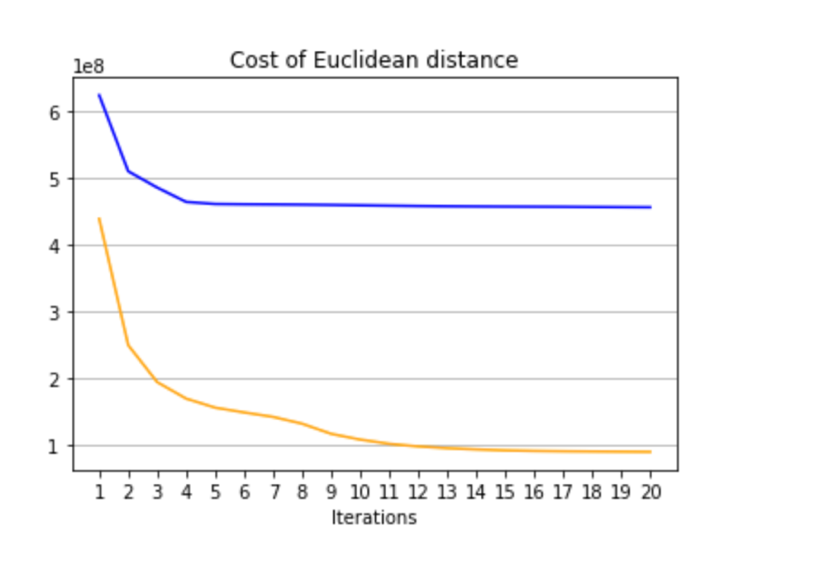
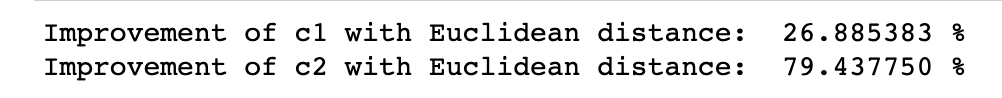
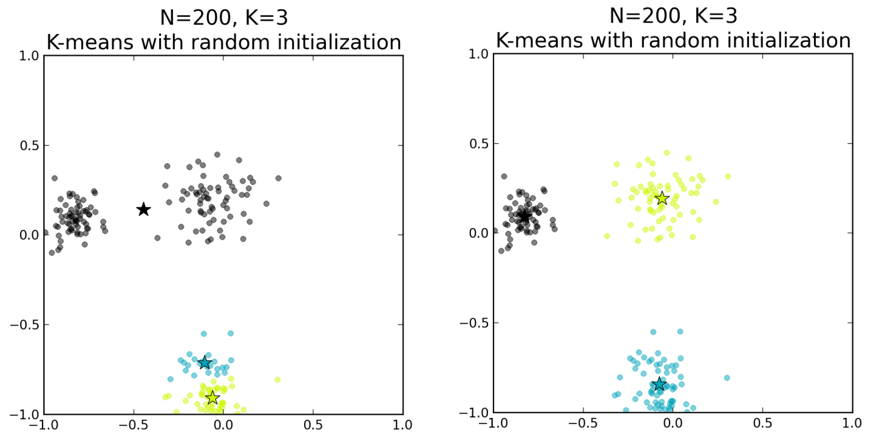
MDA\_HW3\_KMeans

107062103 王依婷

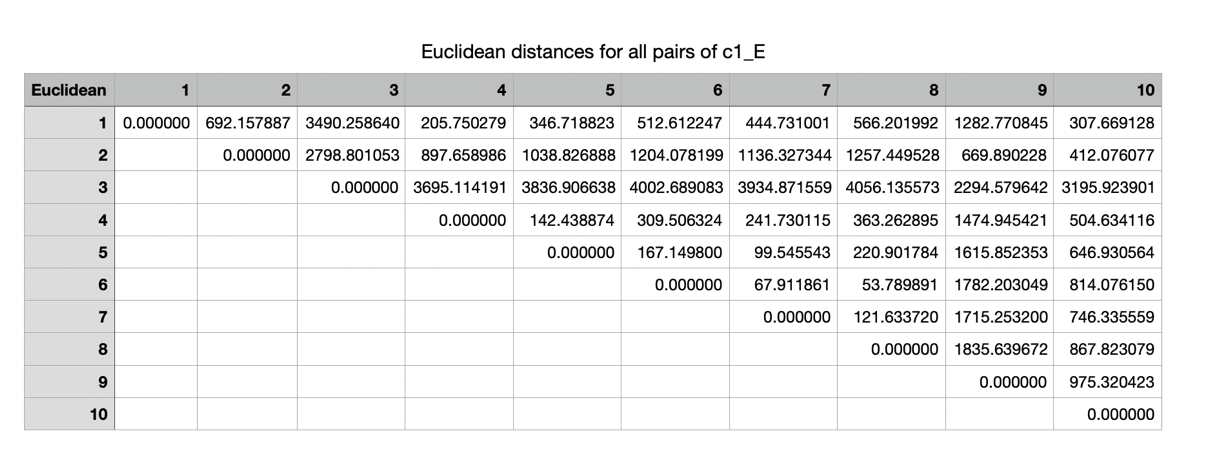
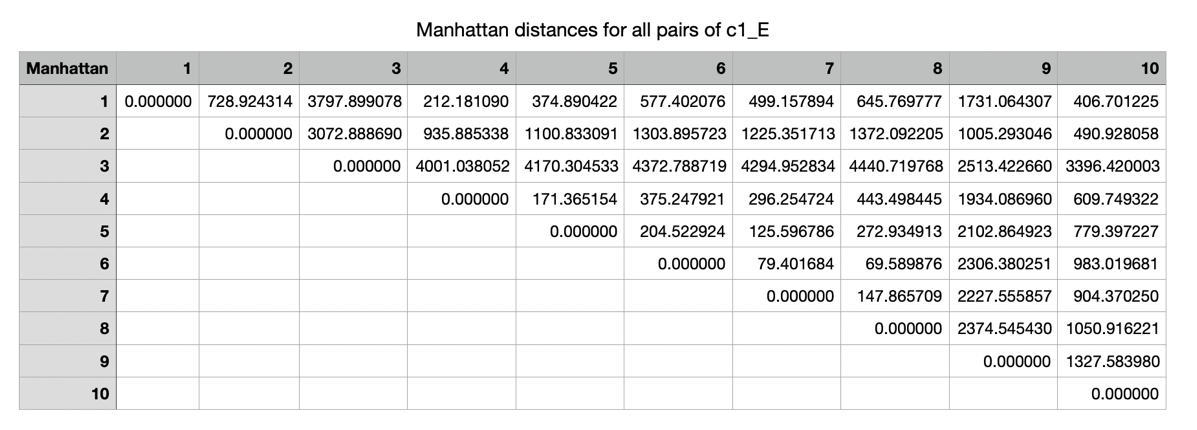
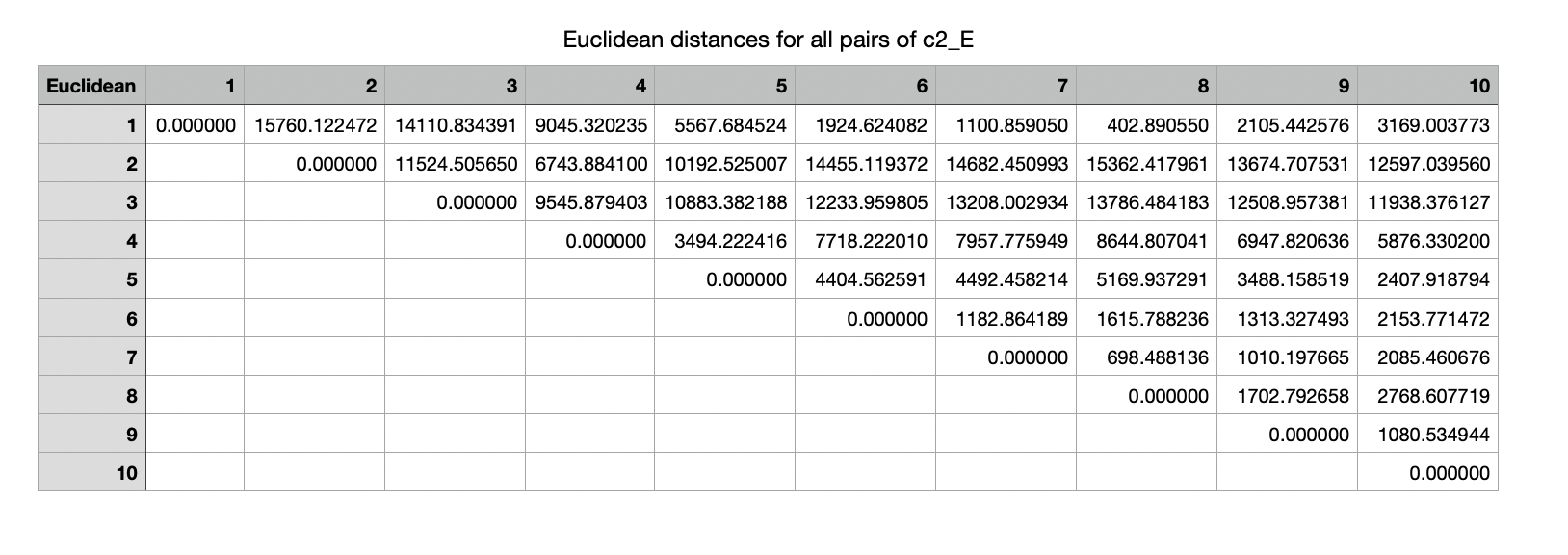
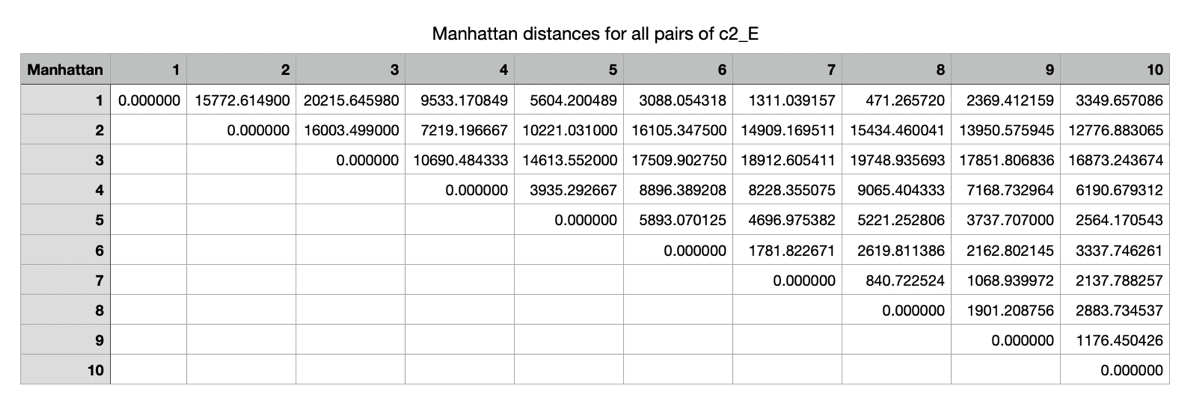
1. with Euclidean distance
2. A plot of cost、iteration for 2 initialization strategies(c1 and c2).



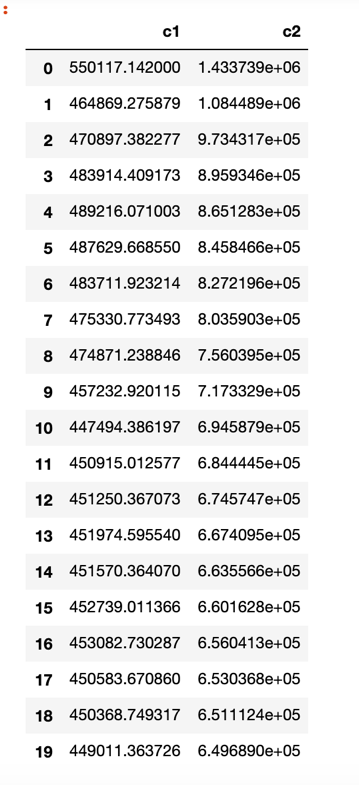
1. Percentage improvement values and explanation.

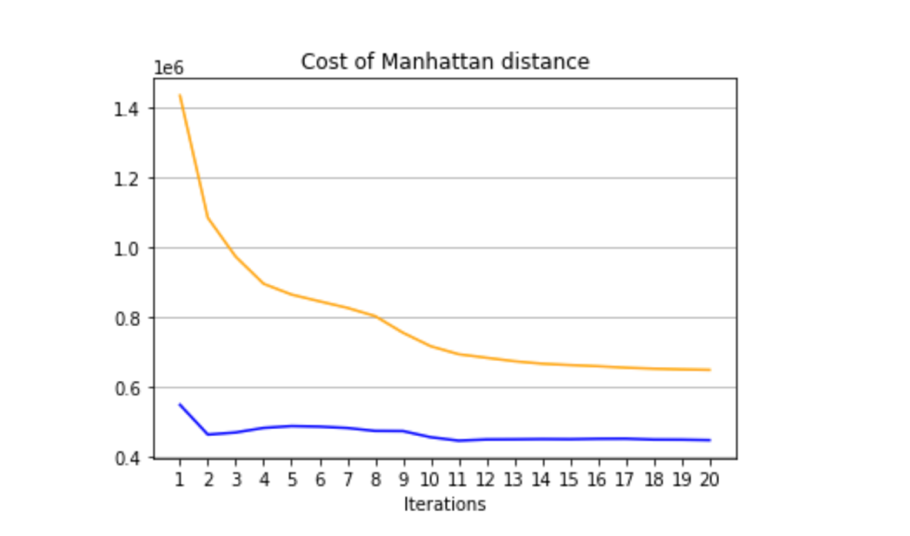
c1中的centroid是隨機挑選，所以c2相較於c1平均來說，centroid的分布比較分散在所有資料點中。而使用Euclidean distance，就是單純計算點與點之間的直線距離，如果centroid初始的位置不好，最後分完的結果可能會和實際的分布差異很大。所以在這種情況下使用c2能獲得比較好的結果。c1的cost都比較大也是因為如此。

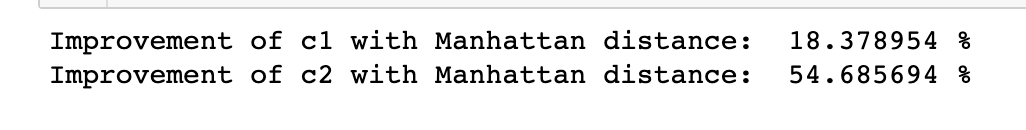
1. The Distances for all pairs of centroids.

* Euclidean distances for all pairs of centroids, with c1.
* Manhattan distances for all pairs of centroids, with c1
* Euclidean distances for all pairs of centroids, with c2.
* Manhattan distances for all pairs of centroids, with c2.

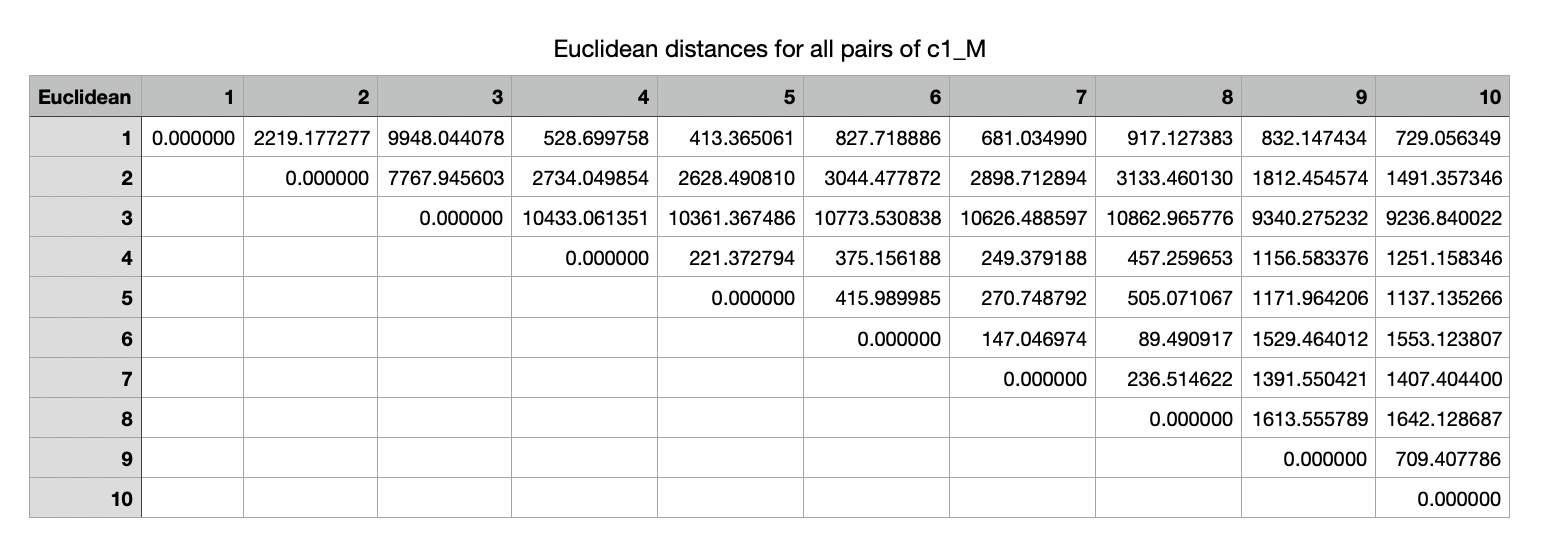
1. with Manhattan distance
2. A plot of cost、iteration for 2 initialization strategies(c1 and c2).

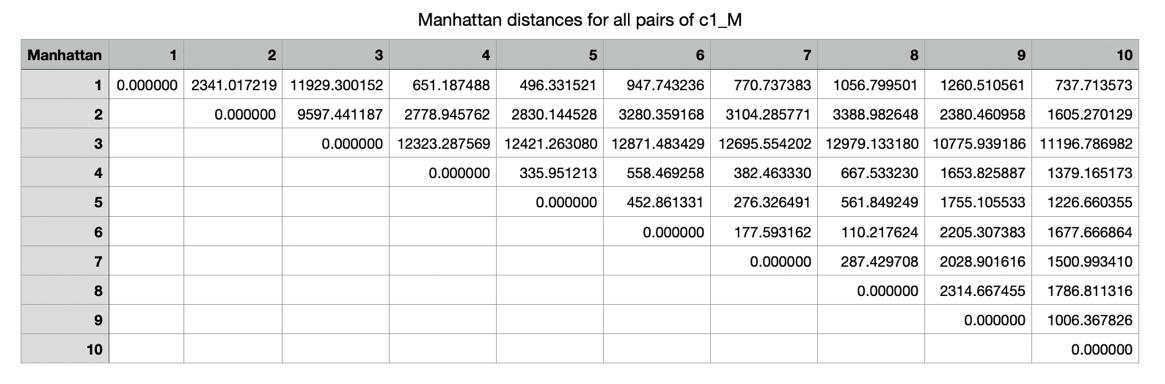
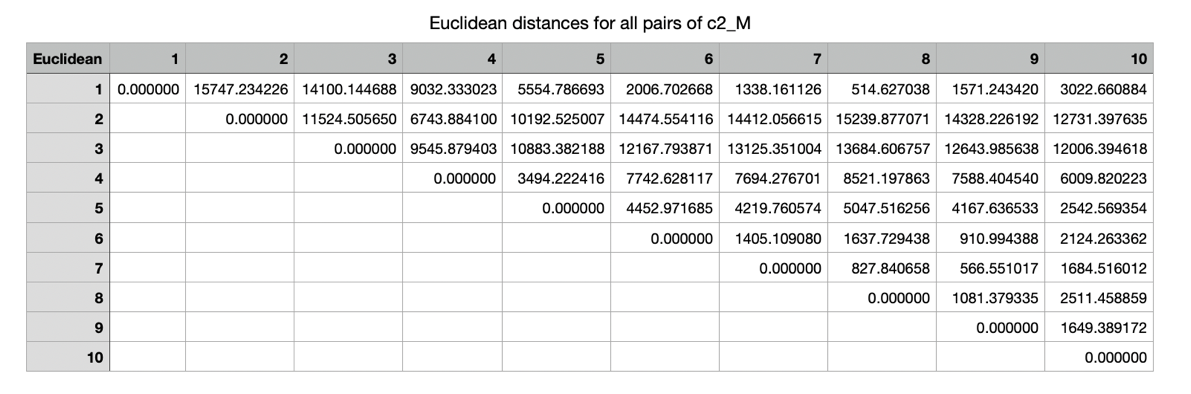
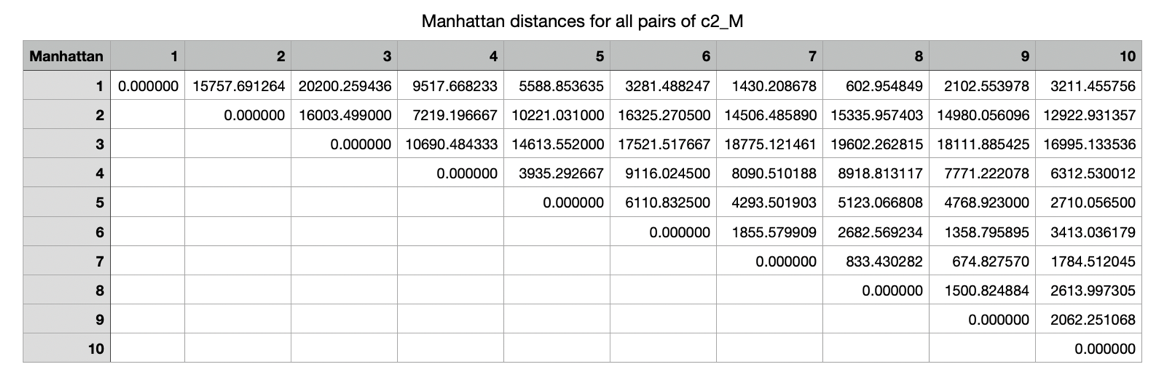




1. Percentage improvement values and explanation.

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。

1. The Distances for all pairs of centroids.

* Euclidean distances for all pairs of centroids, with c1.
* Manhattan distances for all pairs of centroids, with c1
* Euclidean distances for all pairs of centroids, with c2.
* Manhattan distances for all pairs of centroids, with c2.