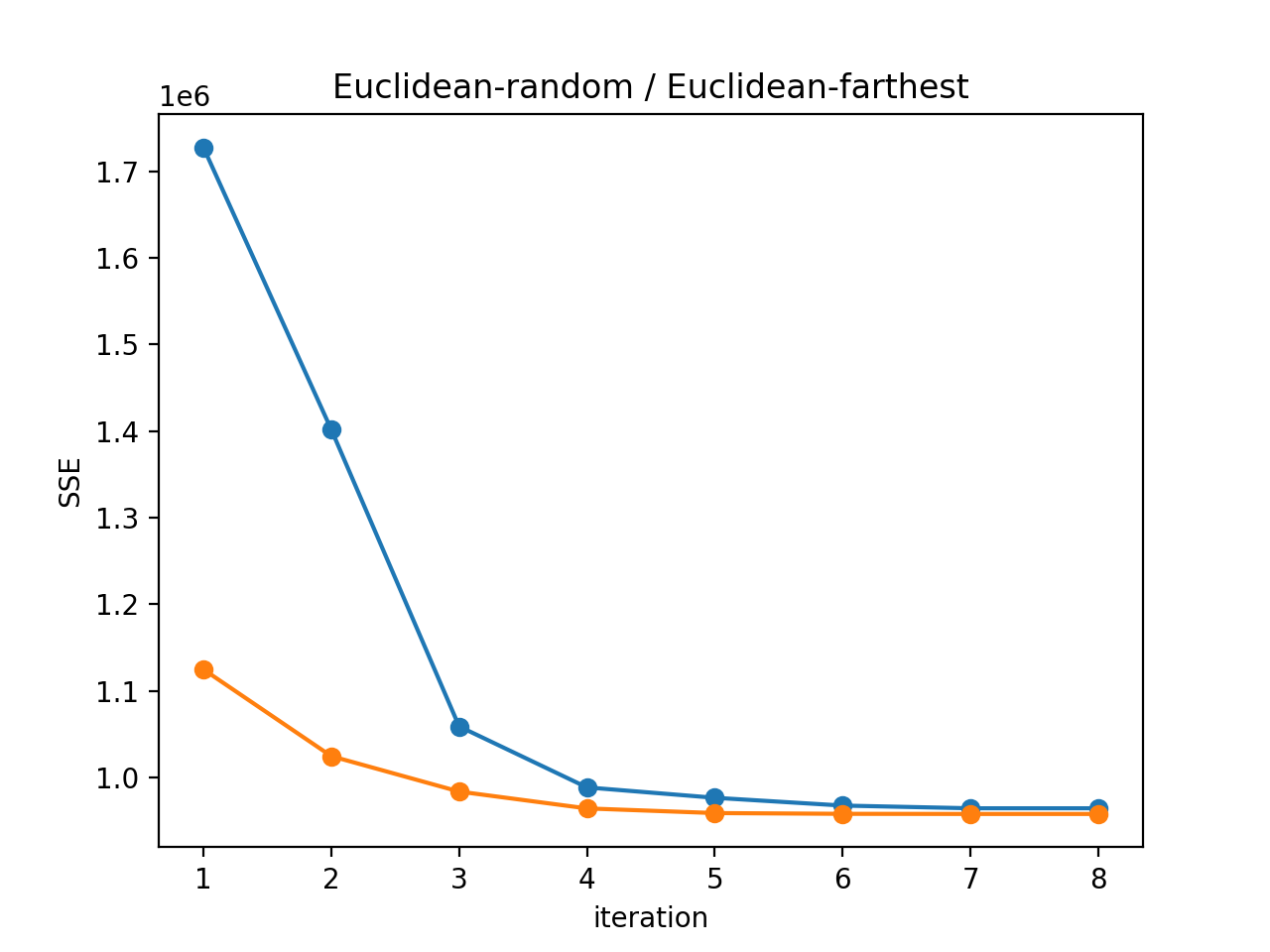
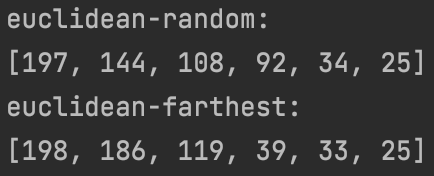
Clustering: K-means

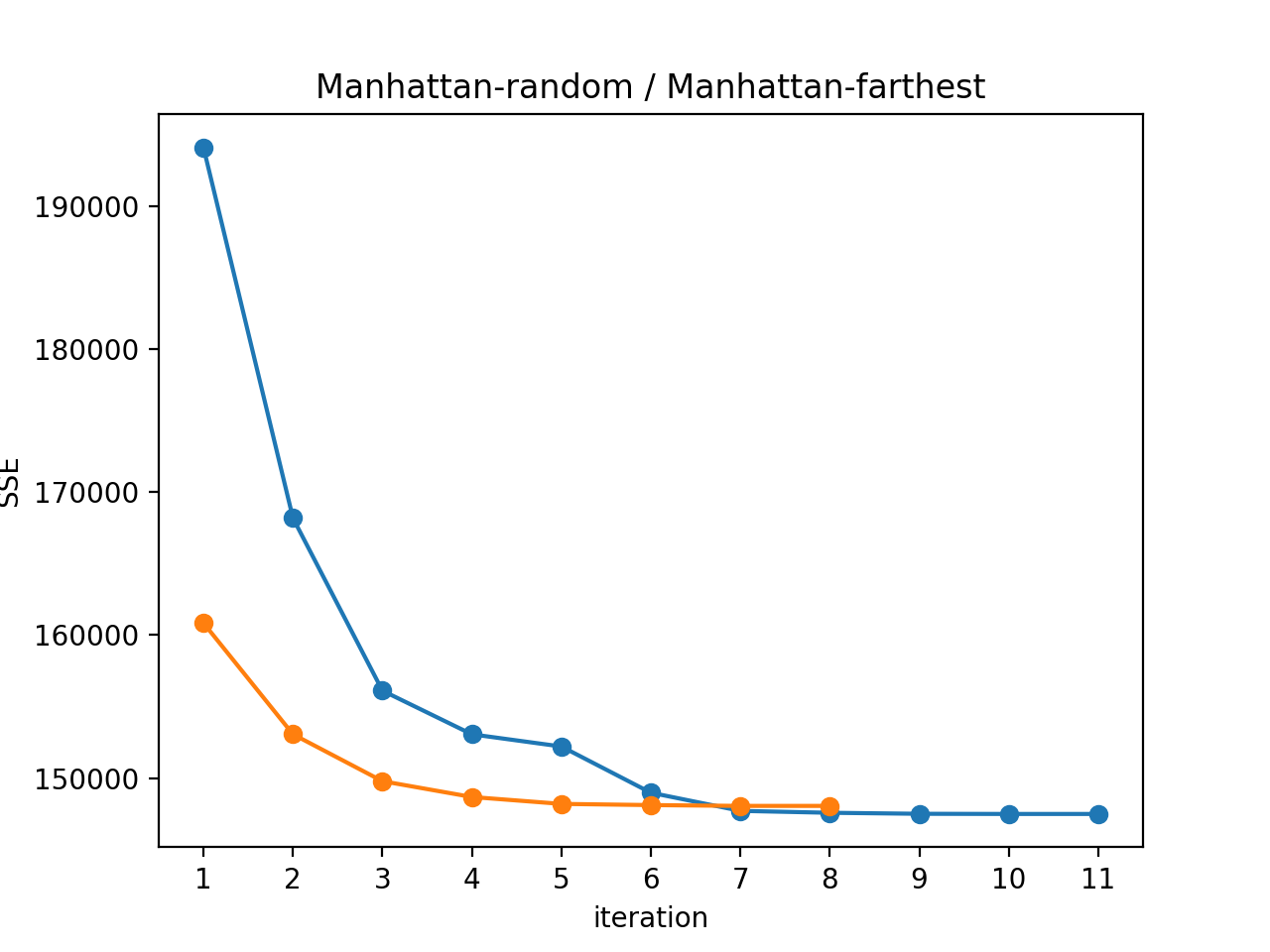
a-2



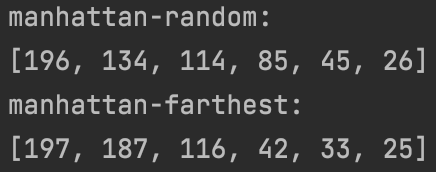
a-3.



b-2.



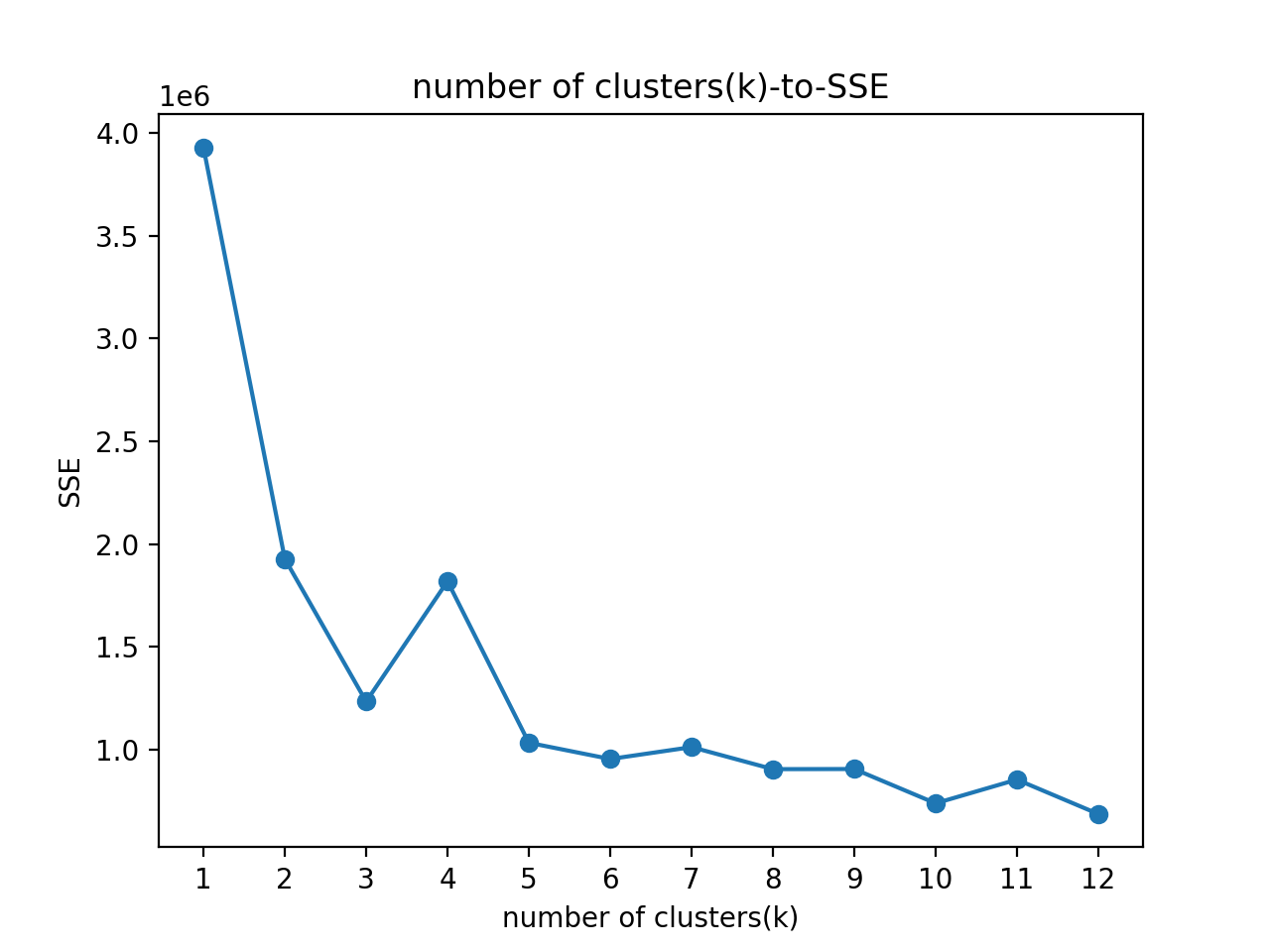
b-3.



c.

使用manhattan distance在一開始有較小的SSE，且相較於euclidean收斂的更快，且我的tolerance設0.01，manhattan farthest的迭代數量又更少，跑得更快了。

d.



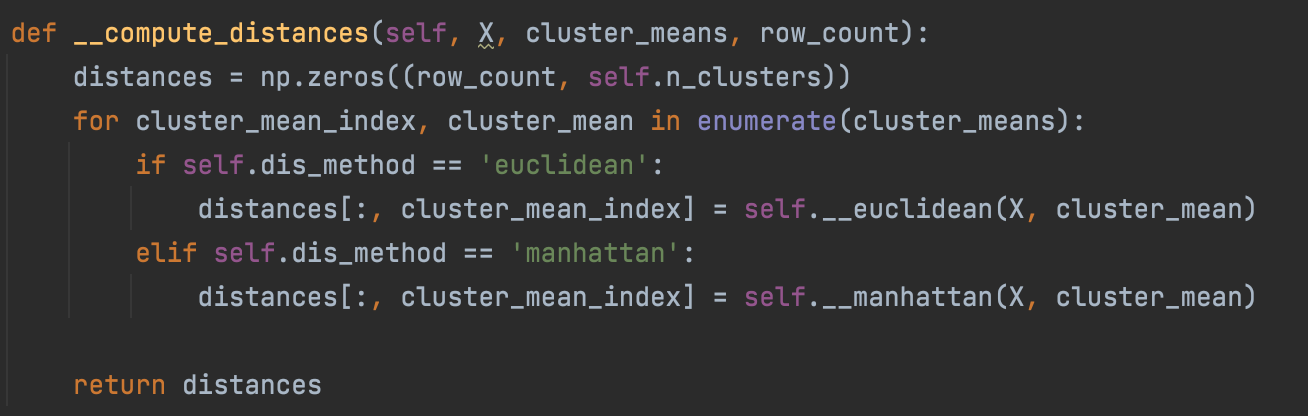
e.

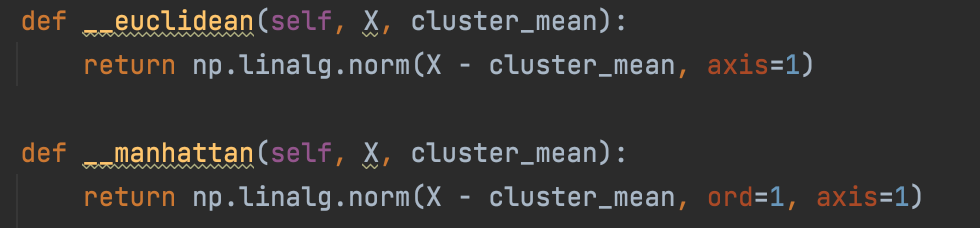
NO，若cluster的數量等於資料的數量，那肯定會有最小的SSE但是沒有意義，所以cluster不能大到無法表示資料的分群，且以d圖中顯示11比10有更高的SSE，所以在某些時候較大的cluster數量不一定會比較小的cluster數量好，所以，在可以有效表示資料的分群下取適量的cluster數量比較好。

f.

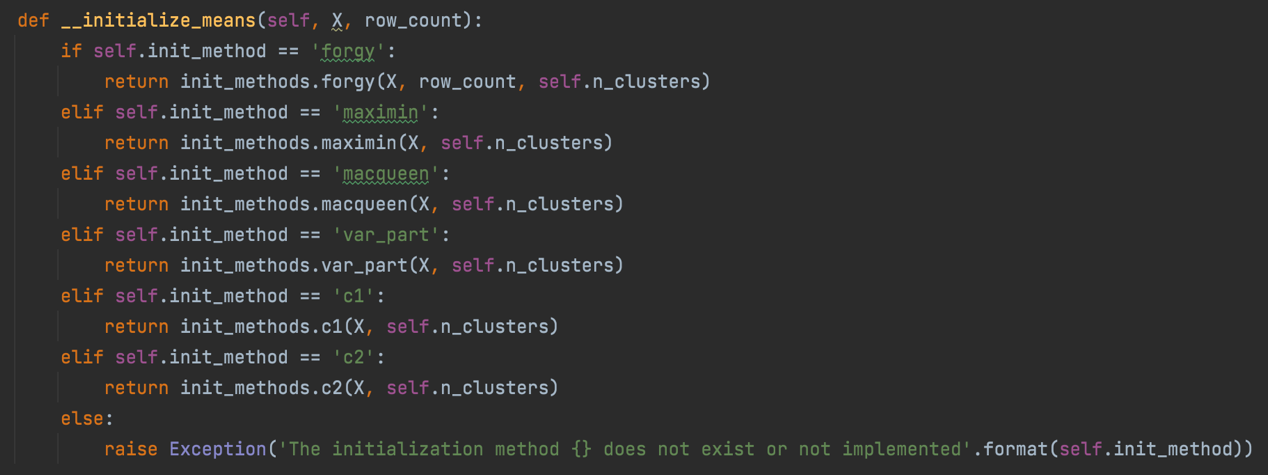
用github上的kmeans source code來改

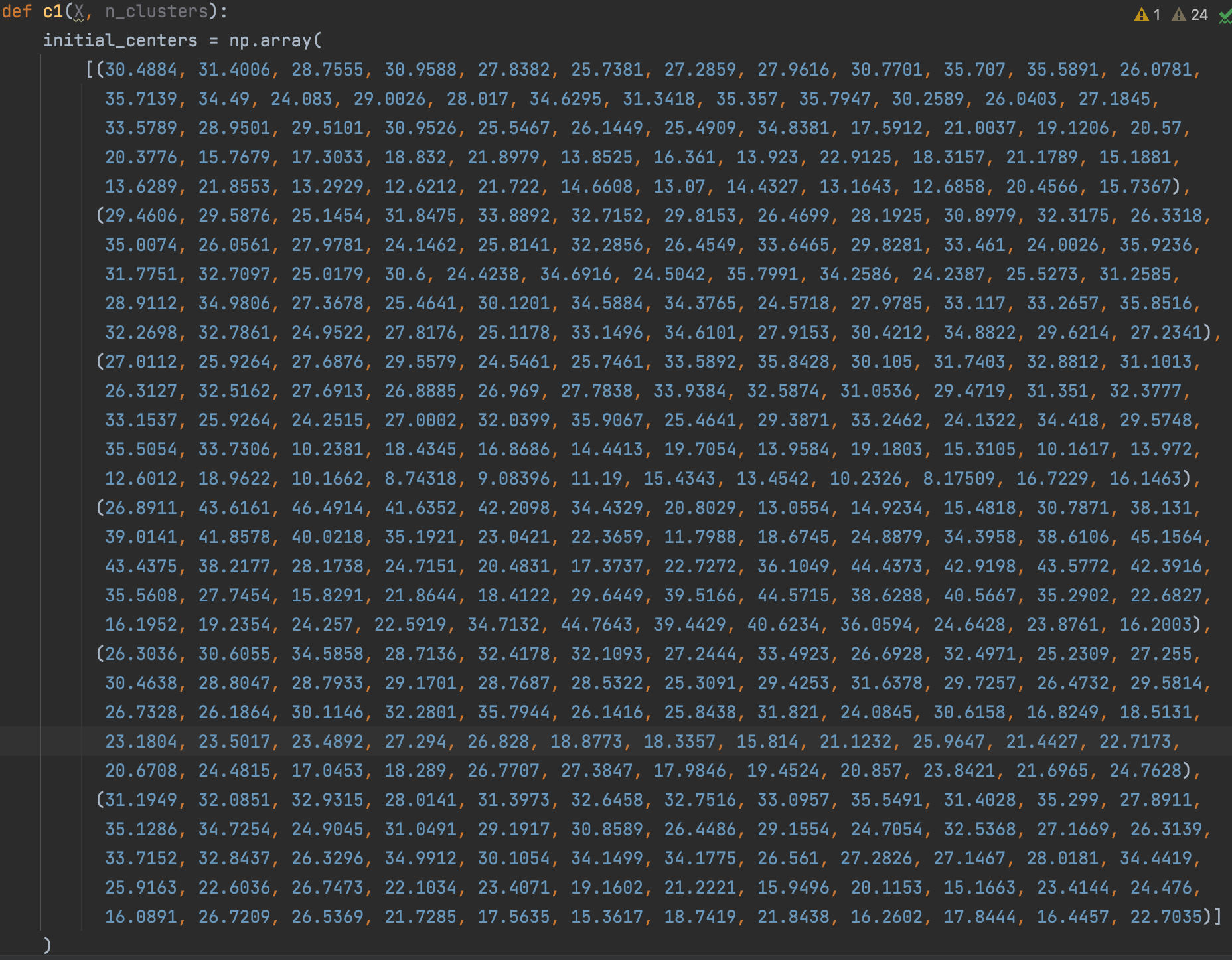
source code只有euclidean distance所以自己再加入manhattan distance

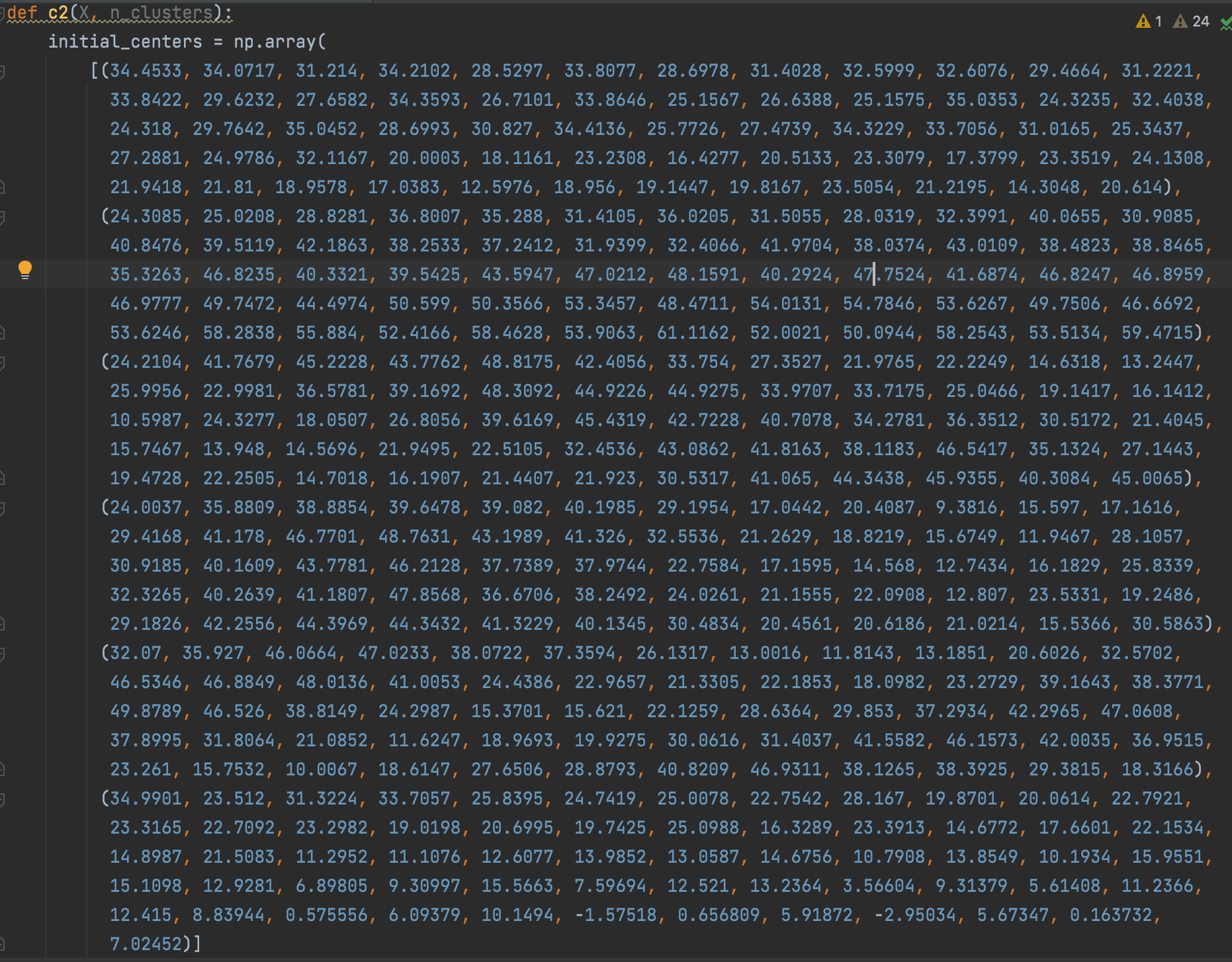




加入initial centroid直接複製貼上







在計算cluster質心的地方加入SSE的計算，同時計算cluster所有object的數量，一併return

