Part B

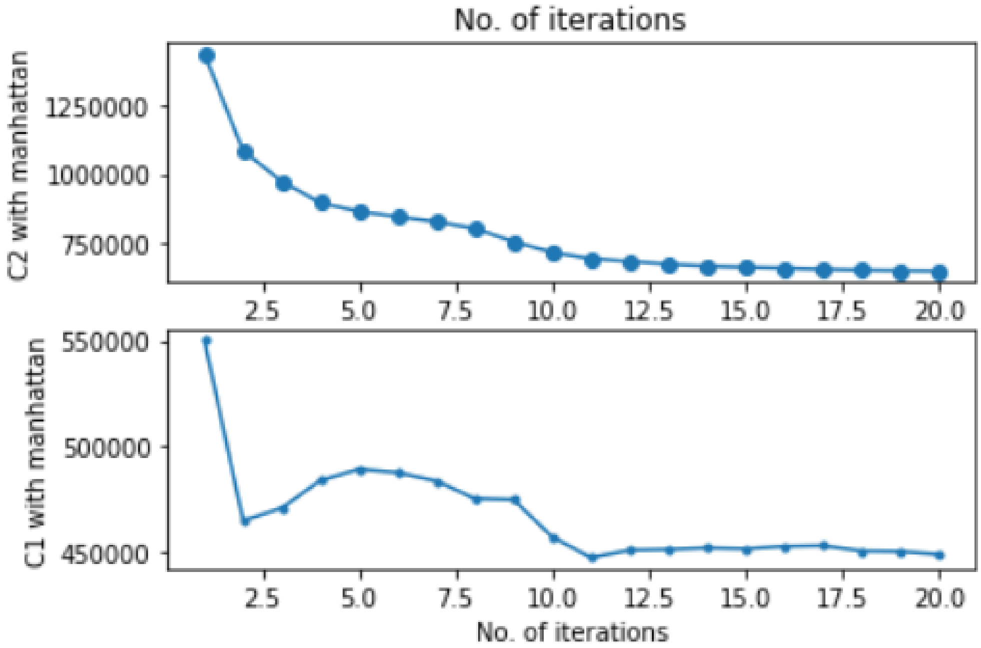
A. Run the *k*-means on data.txt using *c1.txt* and *c2.txt*. Generate a graph (line plot) where you plot the cost function *φ*(*i*) as a function of the number of iterations *i*=1..20 for *c1.txt* and also for *c2.txt*.

The graphs are given below

B. Random initialization of k-means using c1.txt better than initialization using c2.txt in terms of cost φ(i).

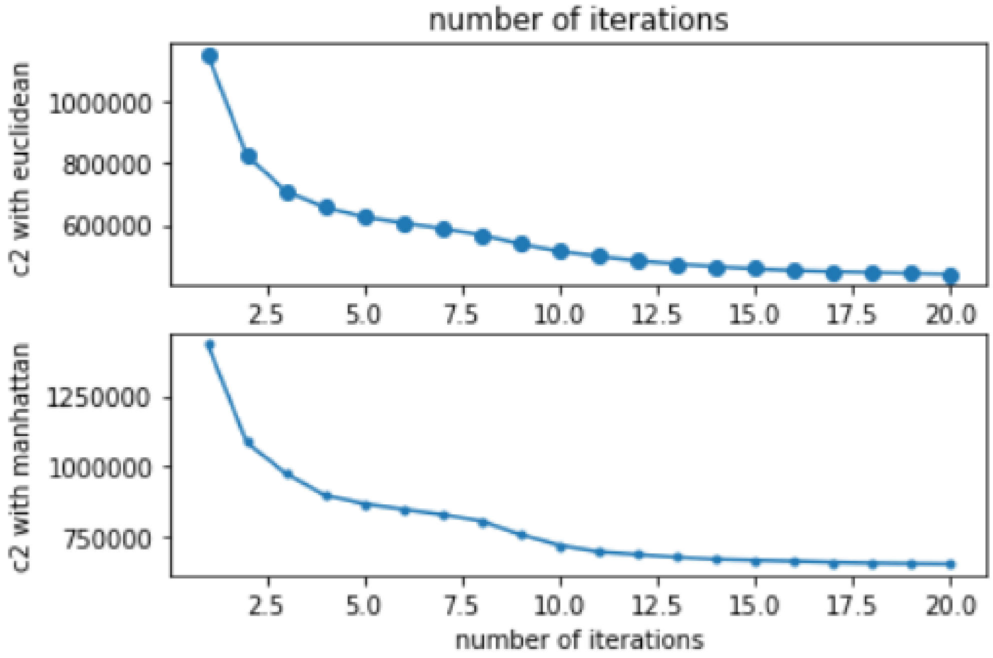
As we can see from the graph below where the centroids I got by using Manhattan distance major.

1. From 1st graph of C2 .text the clusters are started forming from 9th iteration without making so much changes because **good clusters do not overlap** and here Less over lapping, almost no over lapping.
2. Whereas C1 is not forming good clusters as there is overlap from 3rd and 11th iteration and clusters are segregating till 13th iteration with large distances.
3. **As compare to C2, cluster quality of C1 is not that good**



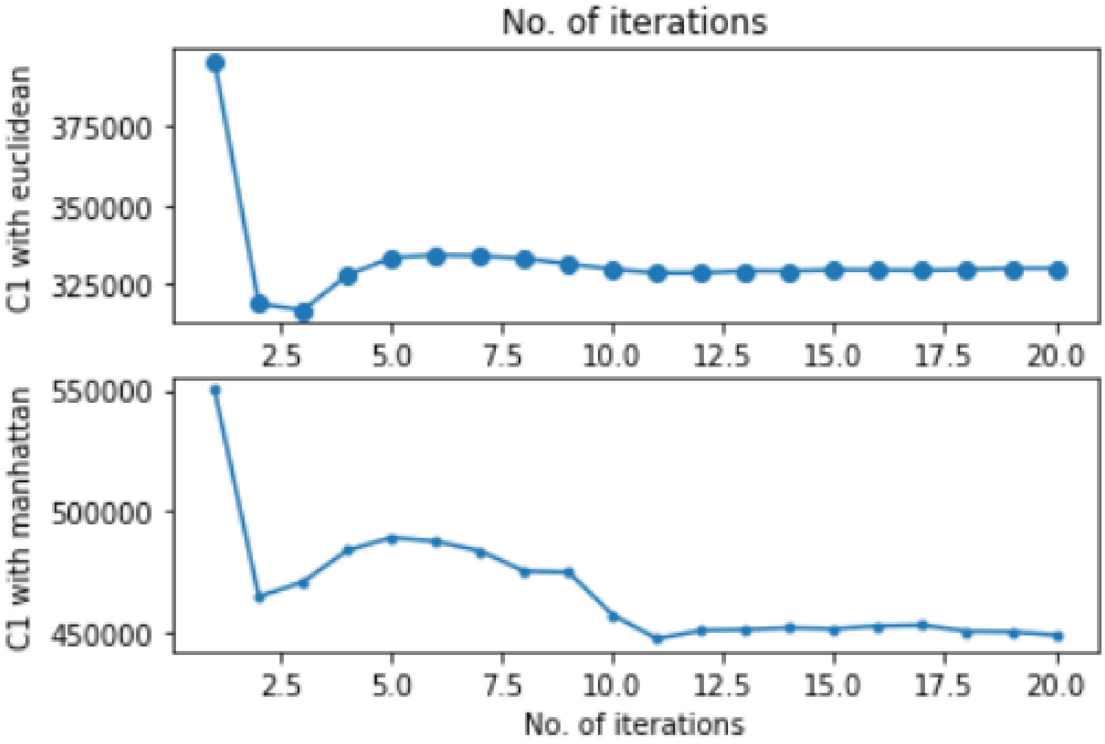
Comparison Manhattan and Euclidian W.r.t C1 and C2 in Data.txt:

By using both type of distance matrix we can see that C2 is creating pretty good clusters compare to C1



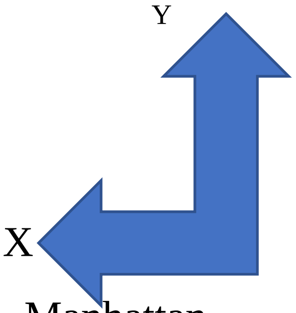
Inference:

1. Less over lapping, almost no over lapping in both type of distance functions
2. C1 improved by 9% after 8th iteration
3. **Actual Clusters segregate less so in this case the final set of clusters are "Really good"**



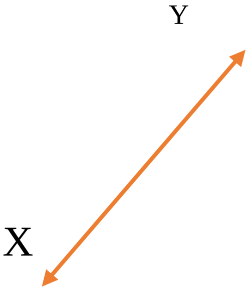
Inference:

1. Sudden fall and up stream, almost no over lapping in both type of distance functions
2. C1 improved after 11th iteration
3. **As compare to C2, C1cluster quality is not that good**



Manhattan

(uses taxicab geometry)



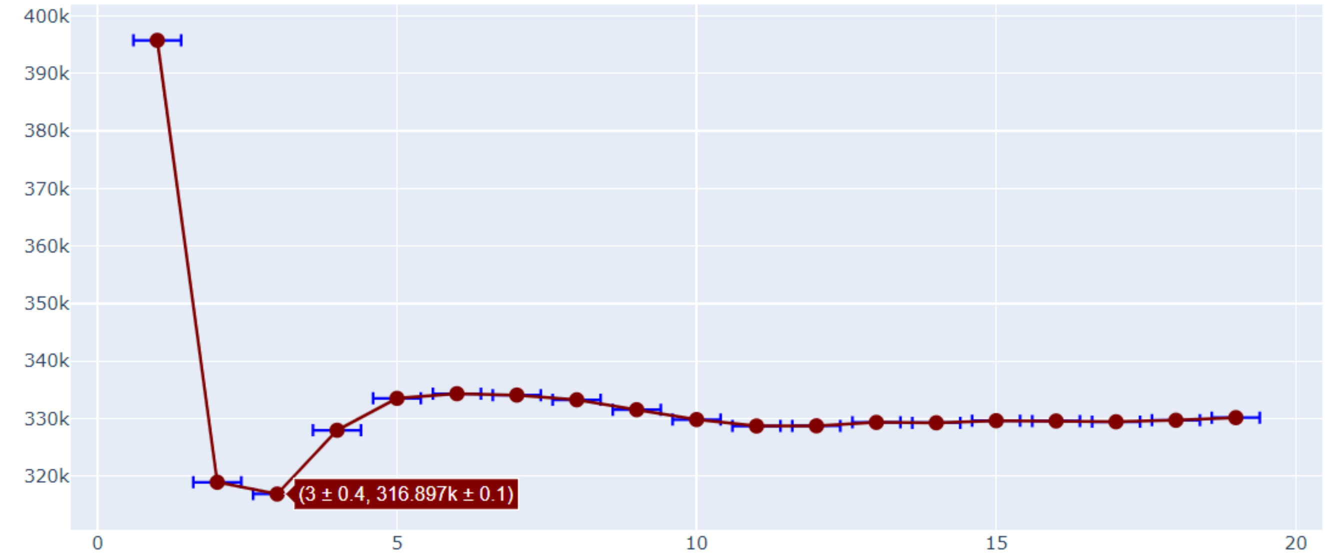
Euclidean

Part A

A.

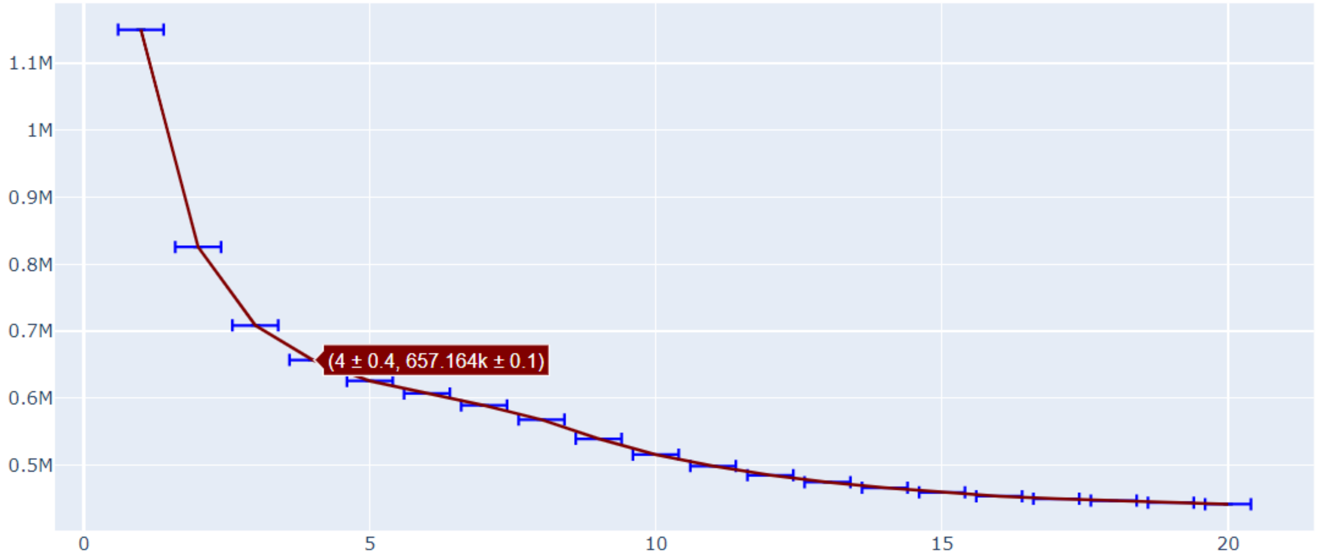
Euclidean

C1.txt



1. After 1st iteration sudden downfall we can see from 395749 to 318949
2. From 3rd to 9th iteration the values gradually increased and after that it remained same without making any disturbances between clusters.
3. These are creating good clusters after 11th iteration which means euclidean requires more number of iterations to cluster in a better way for C1.txt

C2



Here c2 text is forming clusters in a less discreminant way as after single downfall in 2nd iteration it slowly clustered and maintained through out rest of the iterations.

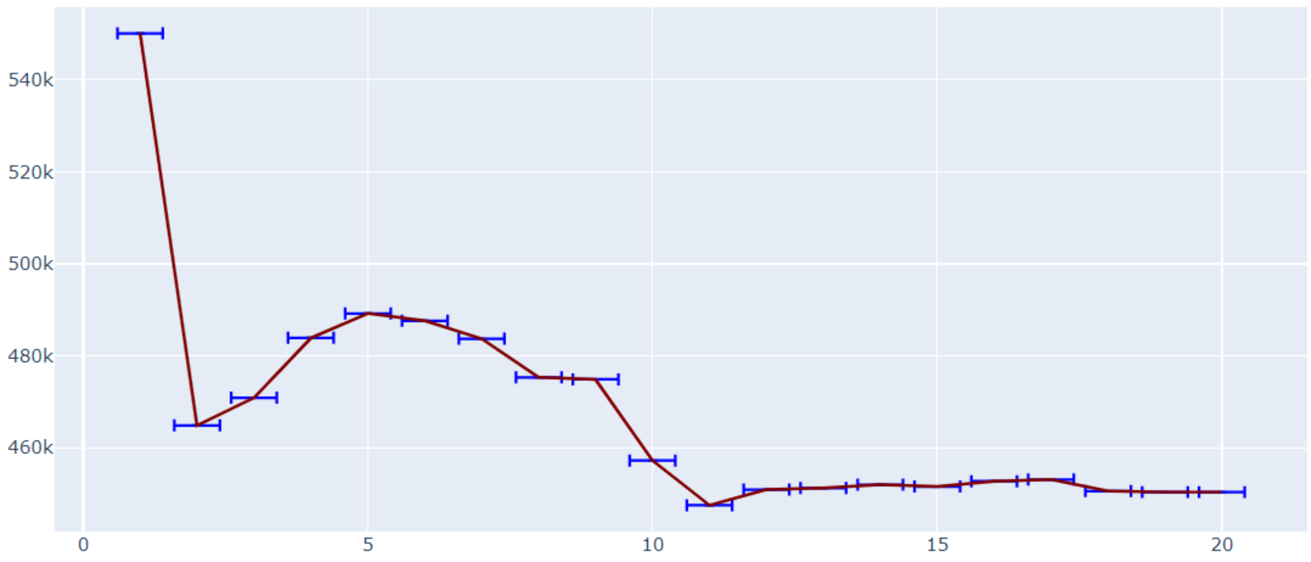
Euclidean metric is not the best one compared to manhattan. Reason may be:

1. if two points are close on most of the variables, but more differnt from each of them in such case, Euclidean distance will exagerate that discrepancy

2. But Manhattan distance will shrug it off, as it is more influenced by the closeness of the other variable.

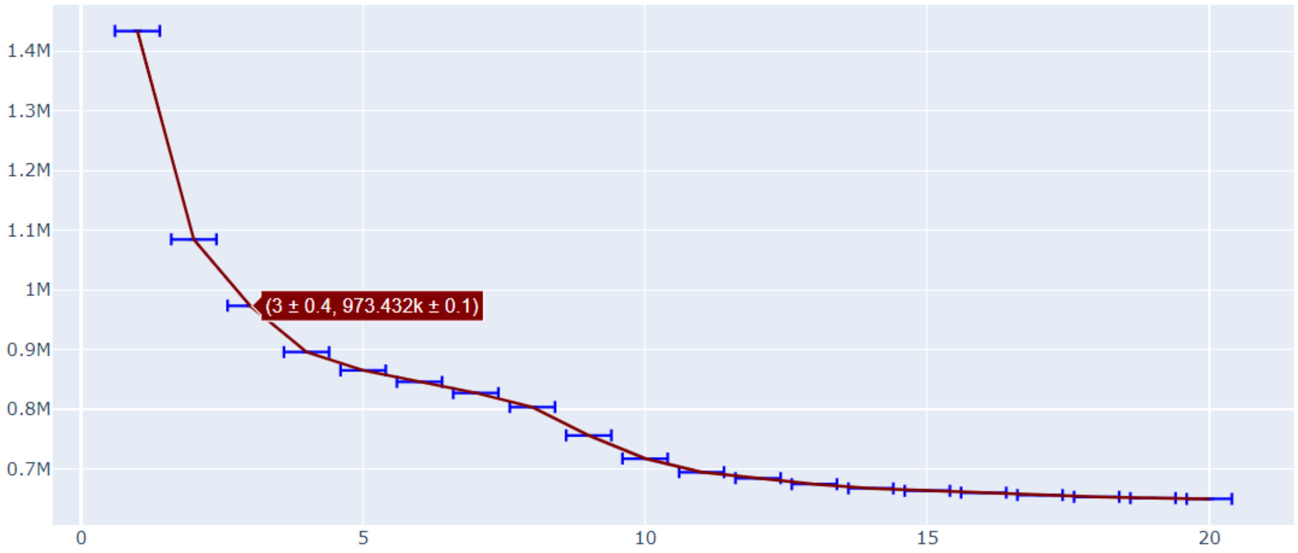
Manhattan

C1txt



1. After 1st iteration the cost value decreased till 3rd iteration.
2. After 3rd iteration it again increased till 9th iteration which means an unstability while creating clustering

C2 Txt



Uniform changes after 3rd iteration which is a good result compared to c1

Conclusion:

Random initialization of *k*-means using *c2.txt* better than initialization using *c1.txt* in terms of cost $\varphi(i)$ because from both of the graphs for 2 different distance matrix we saw a**ctual Clusters segregate less so in this case the final set of clusters by C2 is better than C1 as less overlapping is seen.**