Assignment 4

Made by,
Swati Verma
MT19073

Ques1:

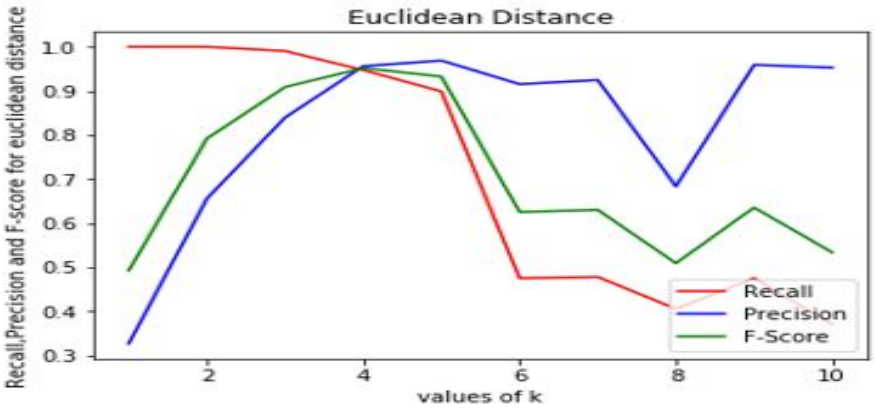
```
cluster=4
old_clist=initial_centers(cluster)
list_distance_euc= cal_euclidean_distance(listdata1,old_clist,cluster)
new_clist=next_centroid(list_distance_euc)
while(compare_centroids(new_clist,old_clist)!=True):
    old_clist=new_clist
    list_distance_euc= cal_euclidean_distance(listdata1,old_clist,cluster)
    new_clist=next_centroid(list_distance_euc)

print("Number of cluster is 4")
print("Number points in first cluster is ", len(list_distance_euc[0]))
print("Number points in second cluster is ", len(list_distance_euc[1]))
print("Number points in third cluster is ", len(list_distance_euc[2]))
print("Number points in fourth cluster is ", len(list_distance_euc[3]))
```

```
Number of cluster is 4
Number points in first cluster is  57
Number points in second cluster is  74
Number points in third cluster is  111
Number points in fourth cluster is  87
```

Ques 2:

Graph corresponding to euclidean distance:

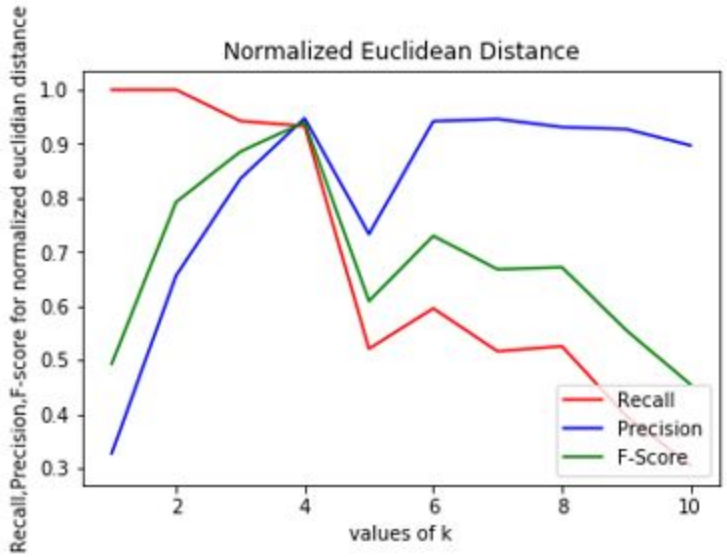
Precision, recall, and f-score is intersecting when number of clusters is 4.As our data set has four different categories so when number of clusters is 4 then it is giving the best result.



Ques 3:

Graph corresponding to normalized euclidean distance:

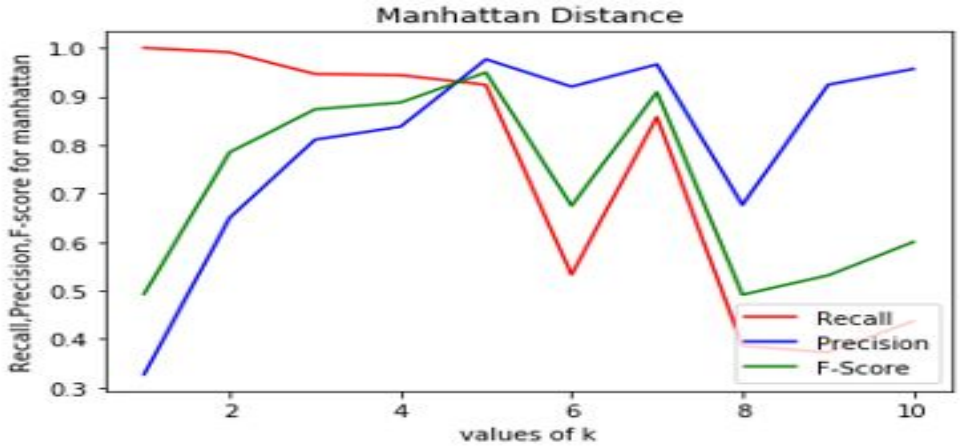
Precision, recall, and f-score are intersecting when the number of clusters is 4.As our data set to have four different categories so when a number of clusters is 4 then it is giving the best result for normalized Euclidean distance also.



Ques 4:

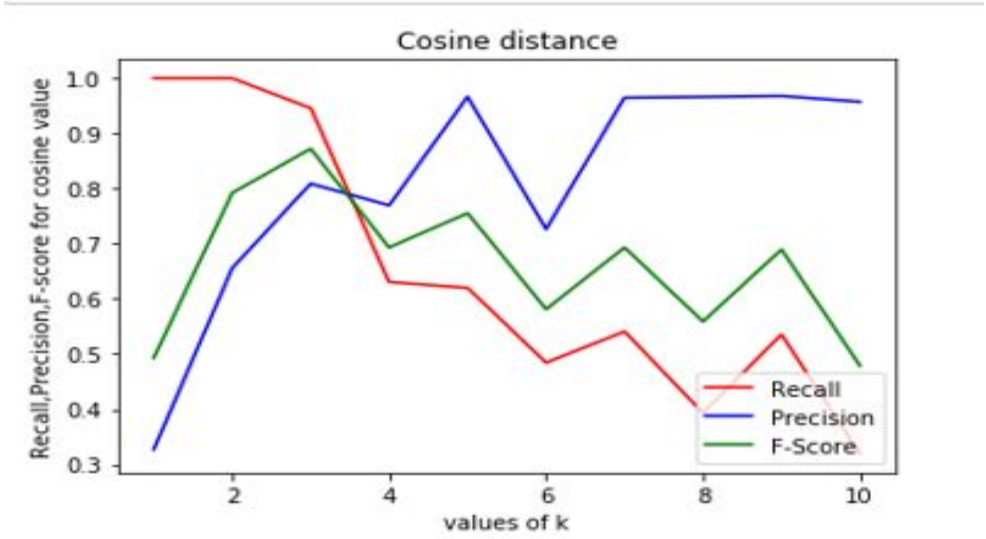
Graph corresponding to manhattan distance:

Precision, recall, and f-score are intersecting when the number of clusters is 4.As our data set to have four different categories so when a number of clusters are 4 then it is giving the best result for Manhattan distance also.



Ques 5:

Graph corresponding to cosine similarity:
Precision, recall, and f-score are intersecting when the number of clusters is 4.As our data set to have four different categories so when a number of clusters are 4 then it is giving the best result for cosine similarity also.



Ques 6:

→ One of the observations I have taken by seeing the graphs is that each time when we are plotting a new graph, the graph changes. This happens because we are taking initial centroid by using the function randint() which generates random numbers. All the values after that are depending on the values of initial centroid that's why each time we are running the algorithm the plot corresponding to recall, precision and f-score is changing.

→ Among all the clusterings we are getting the best values are coming for normalized values because when we normalize our data then the values of our data changes into common scales which gives better results because the data is no more scattered now.

→ One more important observation was that when the number of clusters is one, then each time we are getting recall as 1 because for number of clusters equals to one, all the data points will belong to the same cluster which will give false negative as 0 and as

we know that recall is equal to true positive divided by true positive plus false negative so when false-negative value is zero then recall will be equal to 1.