Report

An iterative **K-means algorithm** has been implemented that tries to partition the dataset into K different non-overlapping sub groups . It tries to make the inter-cluster data points distance as smaller as possible.

Distance Measures :

1) Euclidian Distance

2) Manhattan Distance

3) Cosine Similarity

For each distance measures Precision, Recall and F score has been calculated for K=1 to 10.

Please find the data as below:

1) Euclidian Distance

```
Precision:
     [0.3230980499862675, 0.6476784731143329, 0.8302074523655539,
     0.48889835472019594, 0.851024697845507, 0.9028608724029152,
     0.8784333045941618, 0.8784704694598687, 0.9444038929440389,
     0.9134681288553804]
Recall
     [1.0, 1.0, 0.9457100759378896, 0.5203445540065738, 0.550663039782387,
     0.47036155502663496, 0.40417091692168194, 0.3944803354867959,
     0.43992972911708034, 0.3021081264875893]
F_Score
    [0.4883962303317142, 0.7861709474059387, 0.8842027180968024,
    0.5041315507727784, 0.6686622625928984, 0.6185029248481687,
    0.5536192509217931, 0.5444661712944857, 0.6002474290574499,
    0.45404991056979815]
```

1.1) Euclidian Distance with l2 normalization

```
Precision:
[0.32485729112610273, 0.6514047866805411, 0.833593141075604,
0.7590798187521784, 0.9733749540947484, 0.8635368310414294,
0.9384053964392348, 0.7100963552576456, 0.8851069741594888,
0.8738218303435695]
Recall
```

[1.0, 1.0, 0.9762665449566408, 0.6212345960748517, 0.9072911912368782,
0.5624714742126883, 0.48413966225467825, 0.4835120949338202,
0.36347558192606116, 0.3279324509356458]

F_Score

[0.49040344692518606, 0.7889098928796472, **0.8993062854740382**,
0.6832742446584883, 0.9391720309454911, 0.6812230091552945,
0.6387414850777163, 0.5752978311780879, 0.515327994823263,
0.47689371940595704]

2) Manhattan Distance

Precision:

[0.3230980499862675, 0.6476784731143329, 0.8109652734053373,
0.829349007211301, 0.9174352607307555, 0.7028666901160746,
0.8739756642662031, 0.6948466046190969, 0.9412466843501326,
0.914586639734526]

Recall

[1.0, 1.0, 0.9488835996826476, 0.9515470928255695, 0.5862518417771733,
0.4529638444973365, 0.39890060070270883, 0.456930749178284,
0.4021874645812082, 0.3592315538932336]

F_Score

[0.4883962303317142, 0.7861709474059387, **0.8745201472854046**,
0.8862556740209014, 0.7153723808865224, 0.5508994417258253,
0.5477821011673152, 0.5513162393162393, 0.5635670610656713,
0.5158481507100134]

3) Cosine Similarity

Precision:
[0.3230980499862675, 0.6212646404203306, 0.5427425419904738,
0.7869502523431867, 0.8019949382164657, 0.7040176301067977,
0.8577215189873417, 0.8770917484131564, 0.6994230957269971,
0.7101242167747609]

Recall
[1.0, 0.9649212286070498, 0.6134534738750992, 0.7422645358721524,
0.6105633004646945, 0.470701575427859, 0.5759945596735804,
0.4306925082171597, 0.40536098832596623, 0.36608863198458574]

F_Score
[0.4883962303317142, 0.755865314185515, 0.5759357292969062,
0.7639545056867891, 0.6933075933075933, 0.5641896481456323,
0.6891781936533766, 0.5777051423359051, 0.513256556524235,
0.48311707736603965]

Comparison among F Score Values .

Random data points are selected in order to calculate initial means values. Maximum F Score values are as follows:

Euclidian Distance	0.8842027180968024
Manhattan Distance	0.8745201472854046
Cosine Similarity	0.7639545056867891
Euclidian Distance with l2 normalization	0.8993062854740382

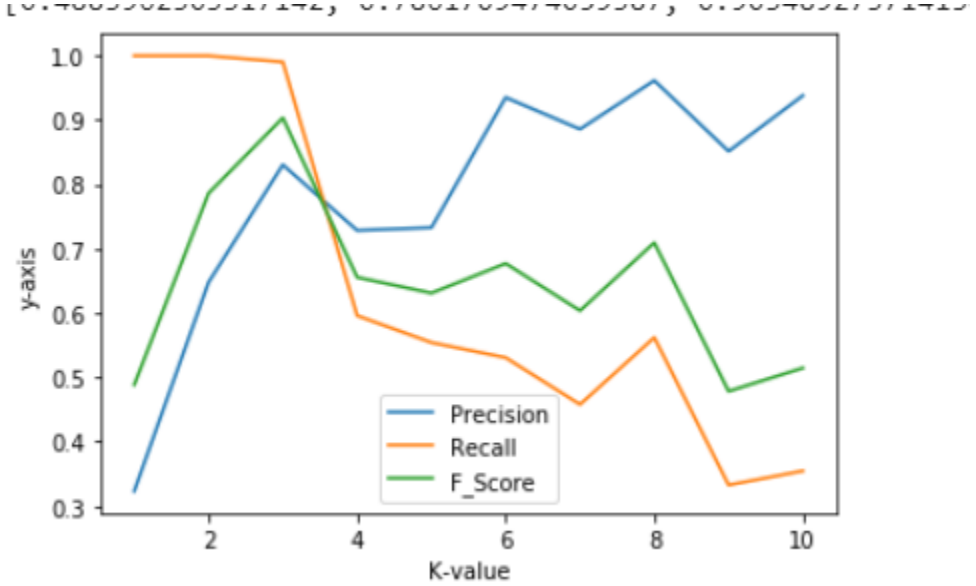
Question 6 : Best setting for K means algorithm **highly rely on the random initial points** selected for initial means value.

But on an average best set up is obtained by using Euclidian Distance (with and without normalization) **on the basis of F score value**.

I have executed the program several times but every time best F score values were obtained for Euclidian Distance(With and without normalization).

Graphs for various Distances :

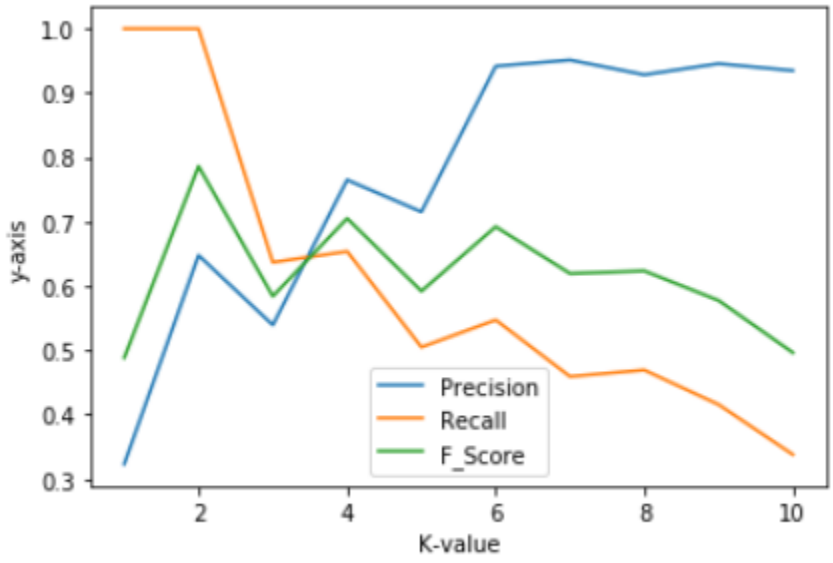
1) Euclidian Distance



Inferences:

- Highest value of Recall is 1 which keeps on decreasing as the value of K increases.
- Precision , Recall and F_Score values intersect near K=4
- Precision value increases with increase in the value of K on X axis.

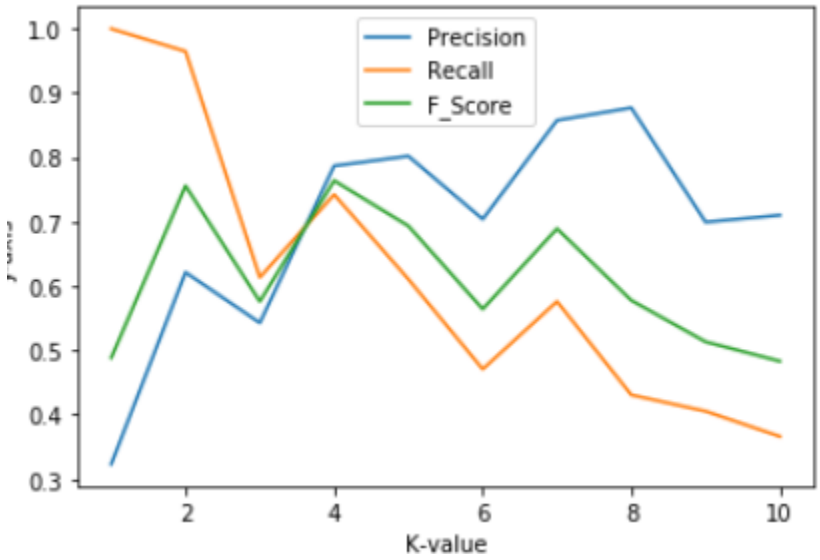
2) Manhattan Distance



Inferences :

- Precision , Recall and F_Score curves interest near K=4.
- Highest Value of F score is approx 0.8 for Manhattan Distance.
- Value of Recall reaches 1 at lower values of K and keeps on decreasing as K increases.

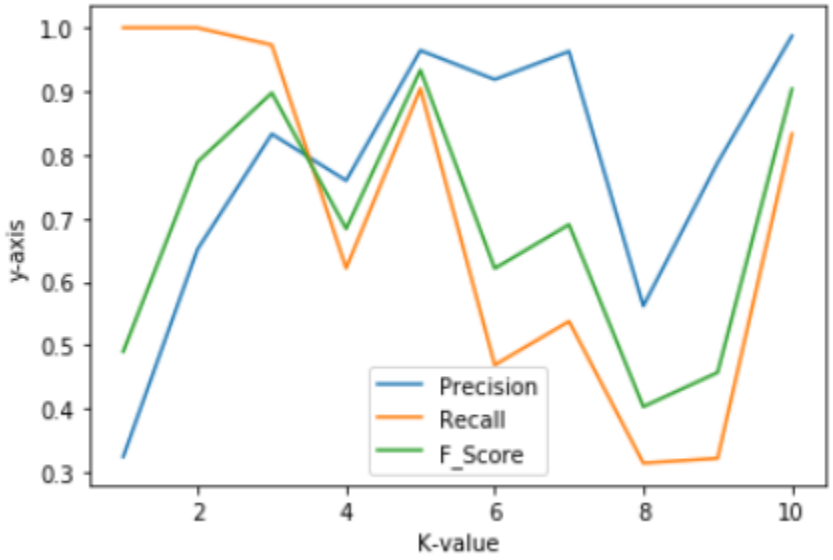
Cosine Distance



Inferences:

- Highest value of F_Score reaches near 0.8.
- Value of F_Score varies from (0.45,0.8)
- Precision , Recall and F_Score curves intersects near k=4

4) Normalized Euclidean Distance



Inferences :

- Precision ,Recall and F_Score curves shows random behavior at various values of k.
- All the three curves intersect near K=4.
- Highest value of F_Score is approx 0.9 near K=3.