**Points required:**

1] Samples from data set:

There is folder called samples in the project folder.

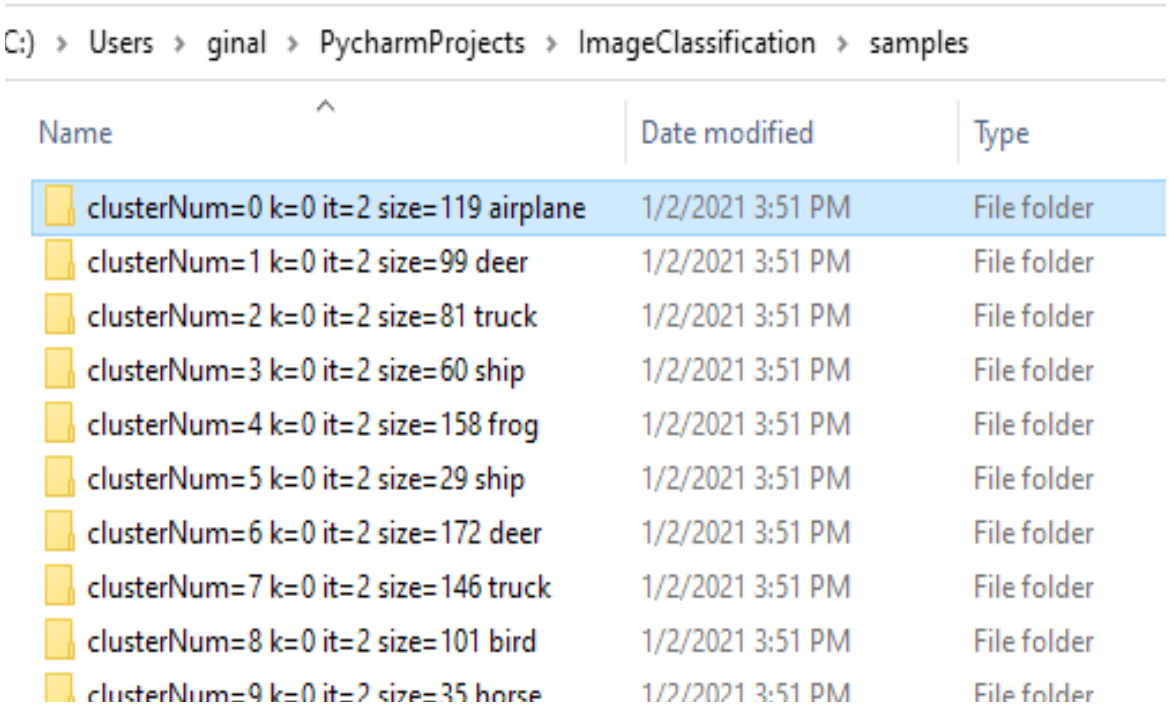
The folder contains sub folders, each sub folder indicates cluster.

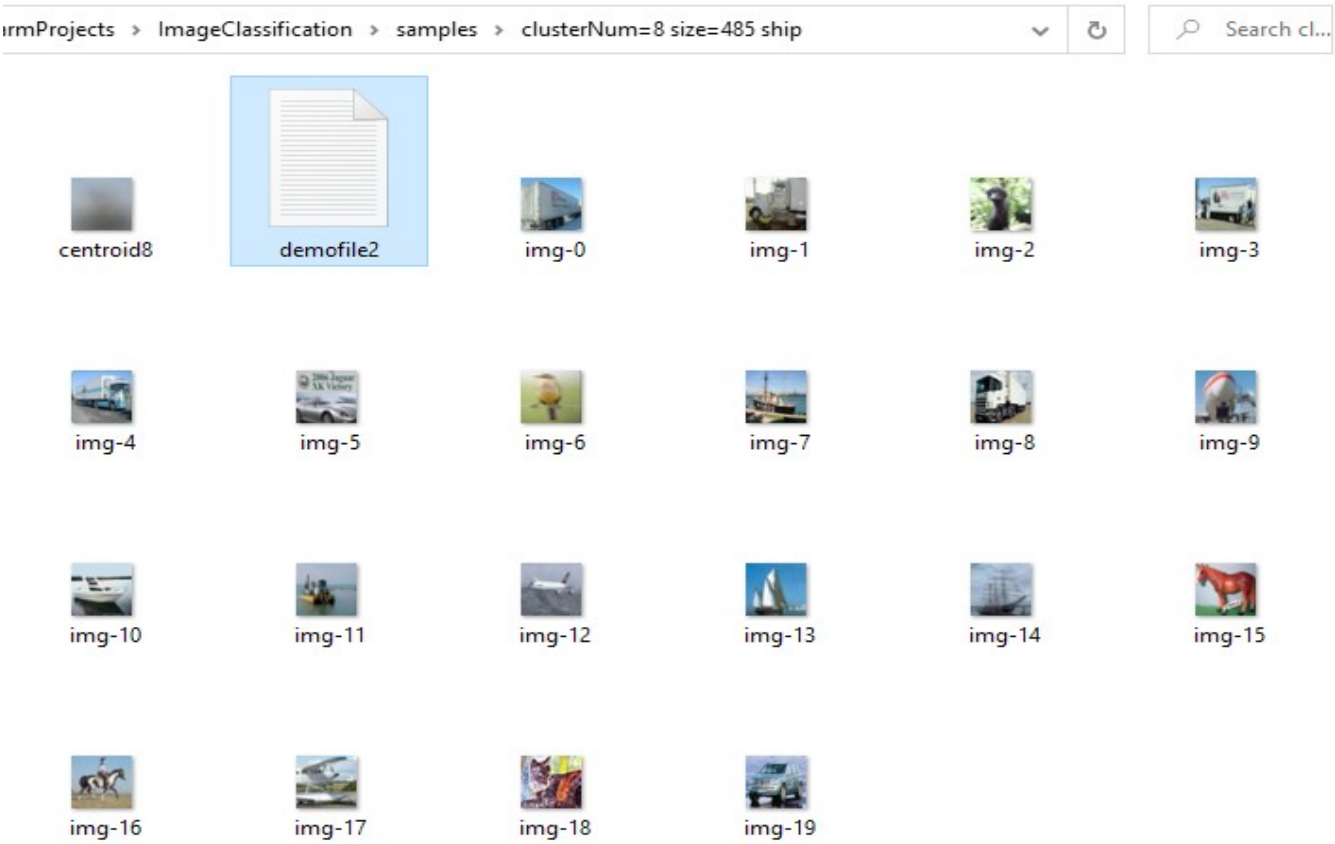
The name of each sub folder (cluster) is : clusterNumber, sizeOfTheCluster, theDominantCategory.

Inside each cluster there are:

-20 samples from the data set included in this cluster.
-The centeroid of this cluster.
-Text file contains number of k, iterations and accuracy of this cluster.

# Screenshots :



demofile2 - Notepad

File Edit Format View Help

```
accuracy is: 35.876288659793815%
 number of k clusters are: 10
 number of iterations are: 3
```

The following results are from one batch only.

```
74
75      self.data1=loadData.unpickle('./data_batch_1')
76      self.data2=loadData.unpickle('./data_batch_2')
77      self.data3=loadData.unpickle('./data_batch_3')
78      self.data4=loadData.unpickle('./data_batch_4')
79      self.data5=loadData.unpickle('./data_batch_5')
80      self.test=loadData.unpickle('./test_batch')
81      #choose the batches you want
82      self.dataAll=[self.data1,self.data2,self.data3,self.data4,self.data5]
83      matrix=self.prepare_data2()
84
85      #self.dataAll = loadData.unpickle('./data_batch_1')
86      #matrix = self.prepare_data()
87
88      #choose number of clusters you want and iterations
89      self.k = 2
90      self.it = 2
```

In line 83 put the batches you want like the picture I put the 5 batches.
In line 89 choose k.
In line 90 choose number of iterations.

2] <u>Different values for k and accuracy:</u>

-For k=6 and iterations =4 :
accuracy for cluster 1 =16.532040926225093%
accuracy for cluster 2 =25.634057971014492%
accuracy for cluster 3 =18.49568434032059%
accuracy for cluster 4 =18.02785095972902%
accuracy for cluster 5 =29.303442754203363%
accuracy for cluster 6 =17.670416942422236%
average accuracy =20.94%

-For k=7 and iterations =4 :
accuracy for cluster 1 =17.332513829133376%
accuracy for cluster 2 =31.18148599269184%
accuracy for cluster 3 =19.27319922128488%
accuracy for cluster 4 =18.692372170997487%
accuracy for cluster 5 =29.198966408268735%
accuracy for cluster 6 =18.338323353293415%
accuracy for cluster 7 =16.578014184397162%
average accuracy =21.5%

-For k=10 and iterations =4 :
accuracy for cluster 1 =17.72253408179631%
accuracy for cluster 2 =28.716216216216218%
accuracy for cluster 3 =26.079869600651996%
accuracy for cluster 4 =19.141588554514062%
accuracy for cluster 5 =32.04930662557781%
accuracy for cluster 6 =17.49663526244953%
accuracy for cluster 7 =16.021505376344088%
accuracy for cluster 8 =17.139001349527664%
accuracy for cluster 9 =29.991431019708656%
accuracy for cluster 10 =21.86115214180207%
average accuracy =22.8%

-For k=10 and iterations =12 :
accuracy for cluster 1 =21.234309623430963%
accuracy for cluster 2 =27.0392749244713%
accuracy for cluster 3 =26.522101751459548%
accuracy for cluster 4 =19.98892580287929%
accuracy for cluster 5 =25.503355704697988%
accuracy for cluster 6 =19.930675909878683%
accuracy for cluster 7 =17.451205510907002%
accuracy for cluster 8 =17.664670658682635%
accuracy for cluster 9 =27.553648068669528%
accuracy for cluster 10=24.160206718346252%
average accuracy =22.8%

comment : as they are 10 classes, at k=10 gave us better accuracy.

3] <u>Different k / restarts:</u>

Different restarts didn’t widely affect the results (small effect).

At k=10 .. iterations = 4 random restarts:

-restart 1: k=10
accuracy for cluster 1 =22.095671981776764%
accuracy for cluster 2 =28.34008097165992%
accuracy for cluster 3 =15.959595959596%
accuracy for cluster 4 =30.609597924773023%
accuracy for cluster 5 =20.101195952161913%
accuracy for cluster 6 =19.24643584521385%
accuracy for cluster 7 =17.382617382617383%
accuracy for cluster 8 =30.244530244530246%
accuracy for cluster 9 =13.617021276595745%
accuracy for cluster 10=27.479892761394 1%
average accuracy =22.5%

-restart 2: k=10
accuracy for cluster 1 =36.73184357541899%
accuracy for cluster 2 =32.240437158469945%
accuracy for cluster 3 =18.56508875739645%
accuracy for cluster 4 =26.62721893491124%
accuracy for cluster 5 =14.761904761904763%
accuracy for cluster 6 =15.706393054459353%
accuracy for cluster 7 =20.553935860058 31%
accuracy for cluster 8 =17.206132879045995%
accuracy for cluster 9 =32.016210739615%
accuracy for cluster 10=19.57255343082115%
average accuracy =23.3%

-restart 3: k=9
accuracy for cluster 1 =20.574712643678 16%
accuracy for cluster 2 =31.818181818181817%
accuracy for cluster 3 =15.934065934065934%
accuracy for cluster 4 =33.111111111111114%
accuracy for cluster 5 =20.334448160535118%
accuracy for cluster 6 =16.422018348623855%
accuracy for cluster 7 =29.824561403508 77%

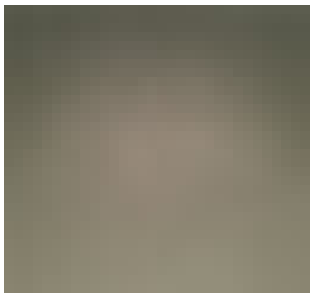
accuracy for cluster 8 =19.47791164586346%
accuracy for cluster 9 =18.40607210626186%
average accuracy =22.8%

-restart 4: k=9
accuracy for cluster 1 =29.80700500357398%
accuracy for cluster 2 =32.65306122448898%
accuracy for cluster 3 =20.06745362563238%
accuracy for cluster 4 =21.77514792899408 4%
accuracy for cluster 5 =27.07182320441989%
accuracy for cluster 6 =17.02508960573476 7%
accuracy for cluster 7 =15.61743341404358 4%
accuracy for cluster 8 =15.55360281195079%
accuracy for cluster 9 =22.77542372881356%
average accuracy =22.4%

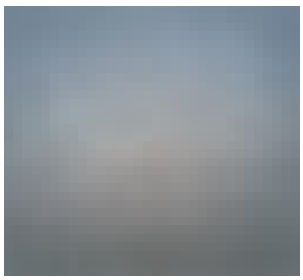
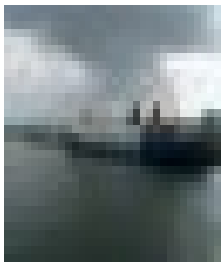
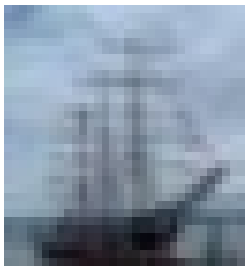
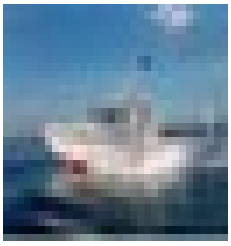
comment: changing the restarts has effect on results.

4],5] <u>Mean images and representative images:</u>

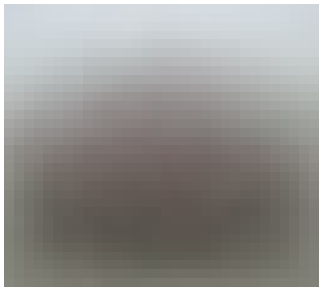
-1]Mean Image: -Its representative images:

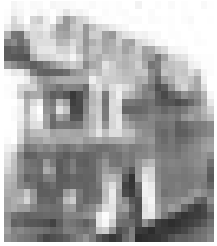
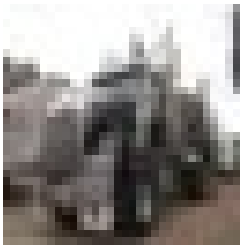



-2]Mean Image: -Its representative images:

-3]Mean Image: -Its representative images:



-4]Mean Image:  -Its representative images:

-5]Mean Image:  -Its representative images:

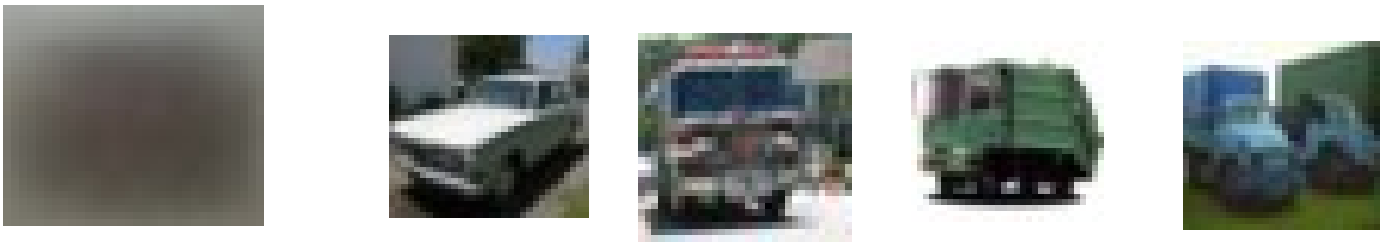
-6]Mean Image:  -Its representative images:

-7]Mean Image:  -Its representative images:

-8]Mean Image:  -Its representative images:

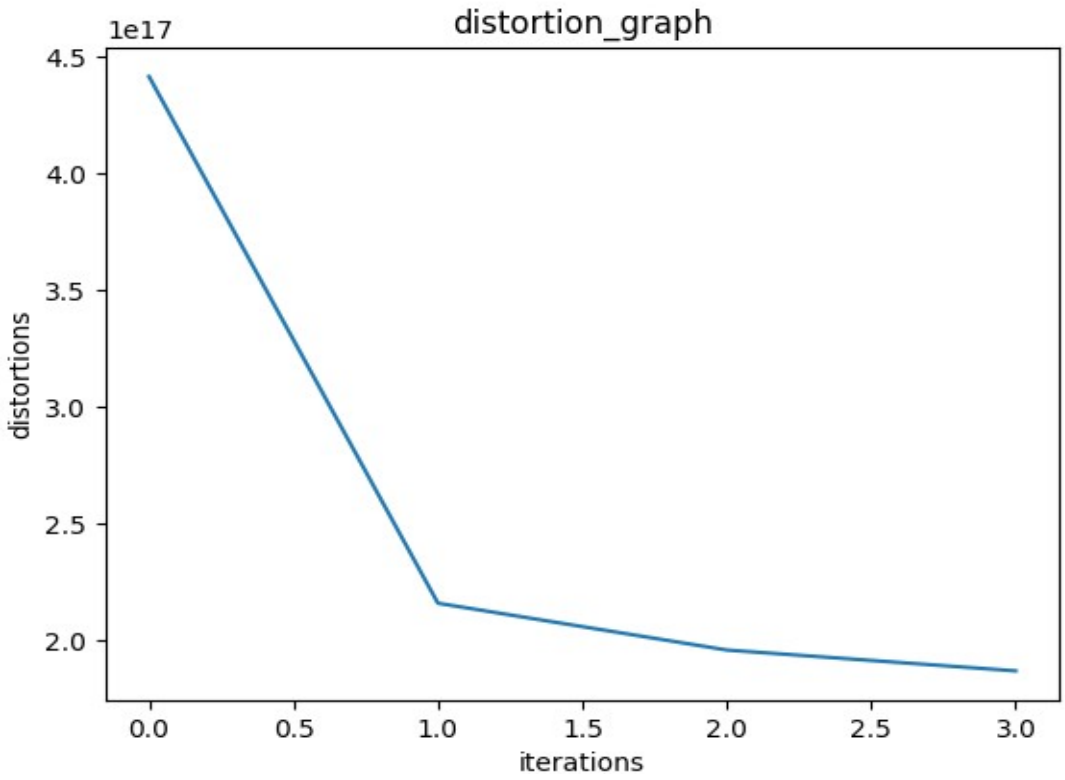
-9]Mean Image:  -Its representative images:

-10]Mean Image:  -Its representative images:

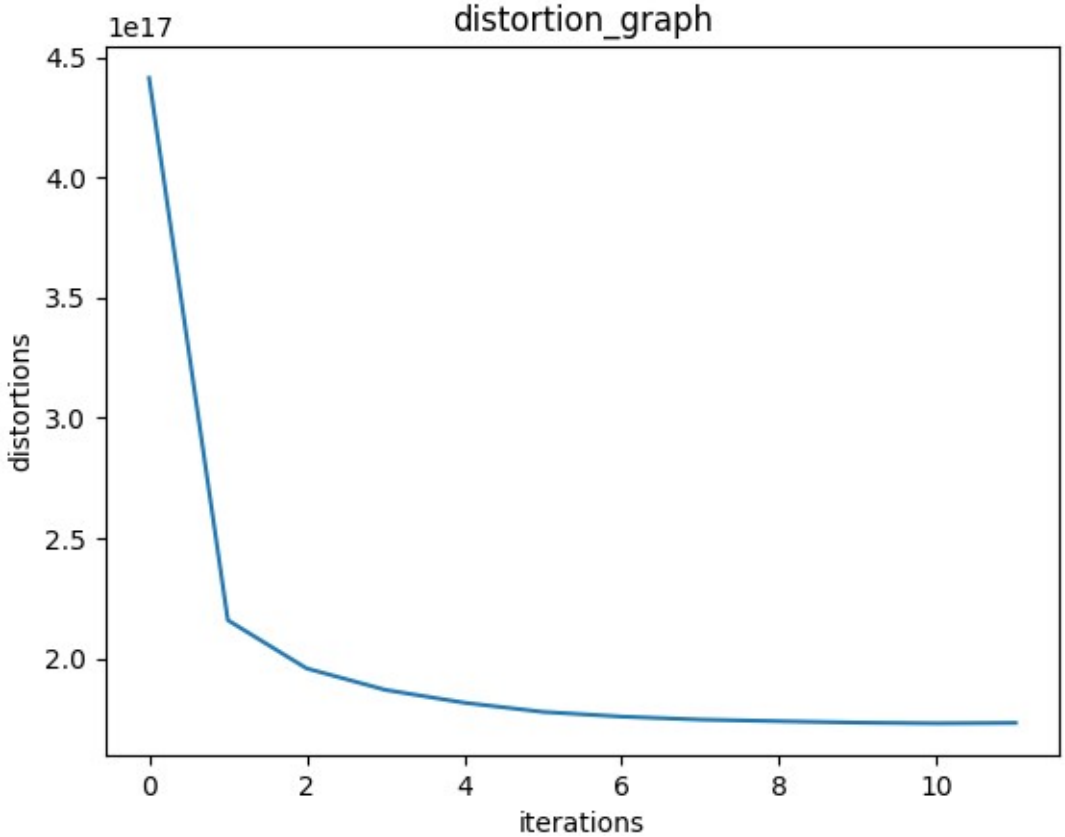


6] <u>Plotting of distortion measure :</u>

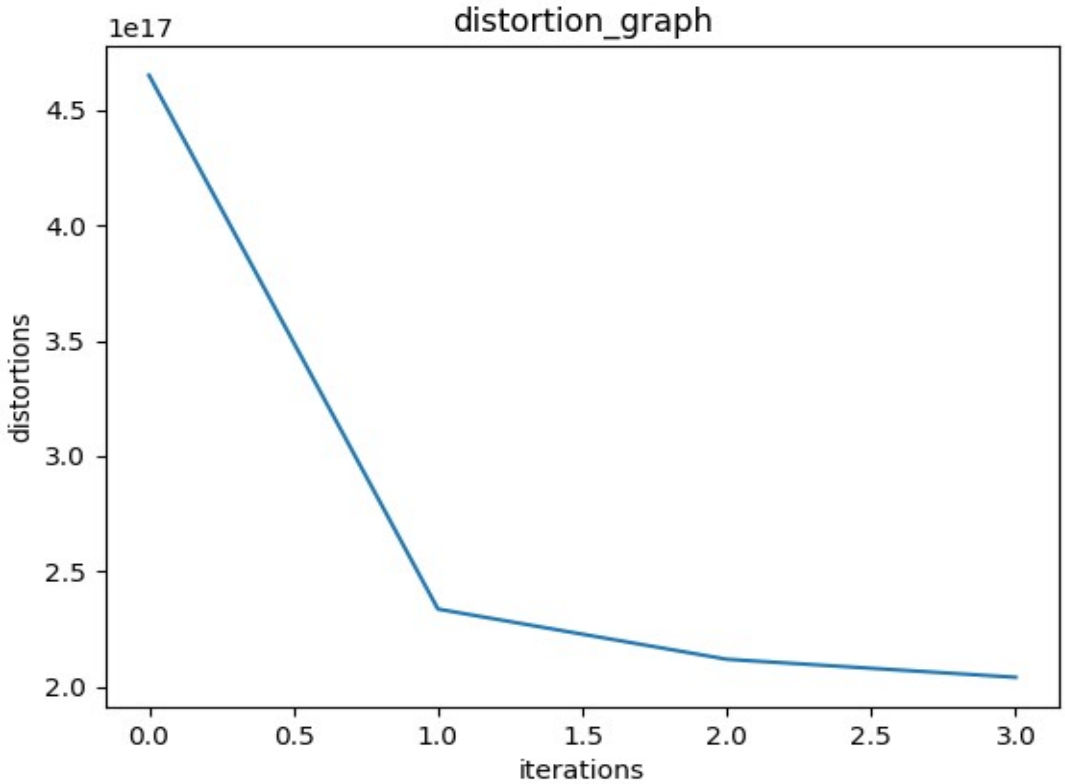
-At k=10, iterations =4 (0,1,2,3):



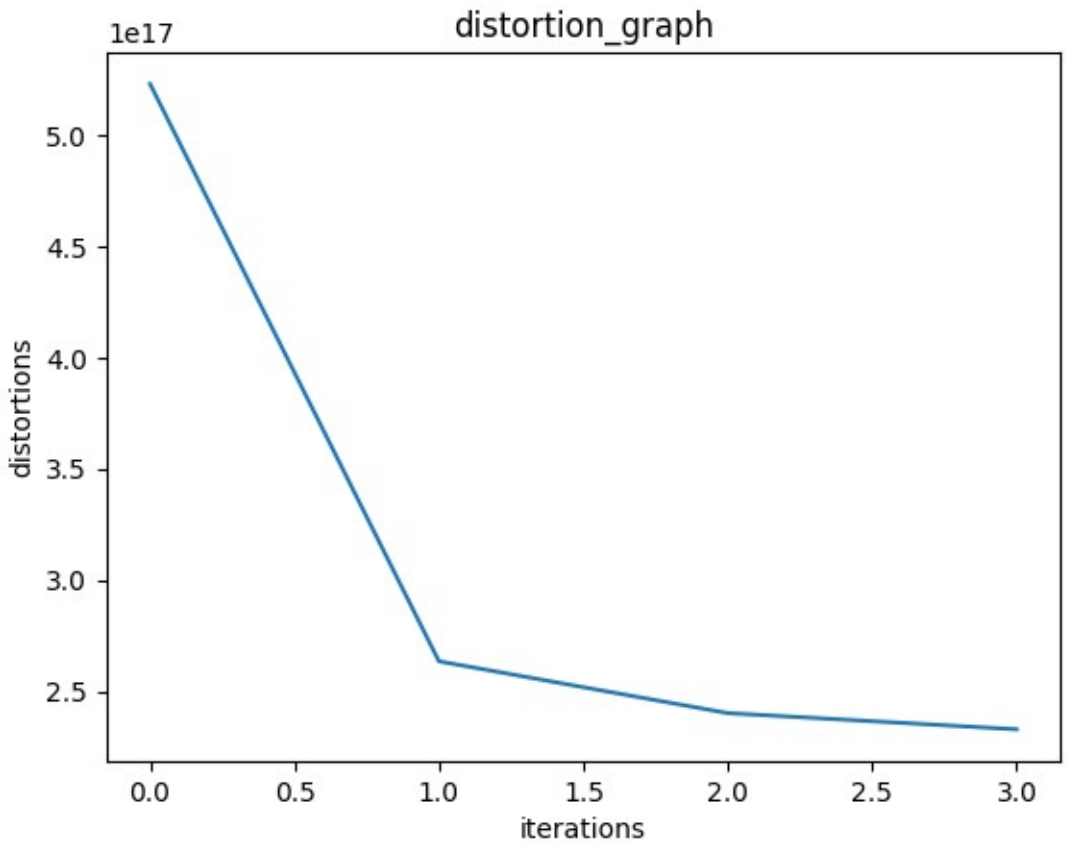
k=10,iterations=12



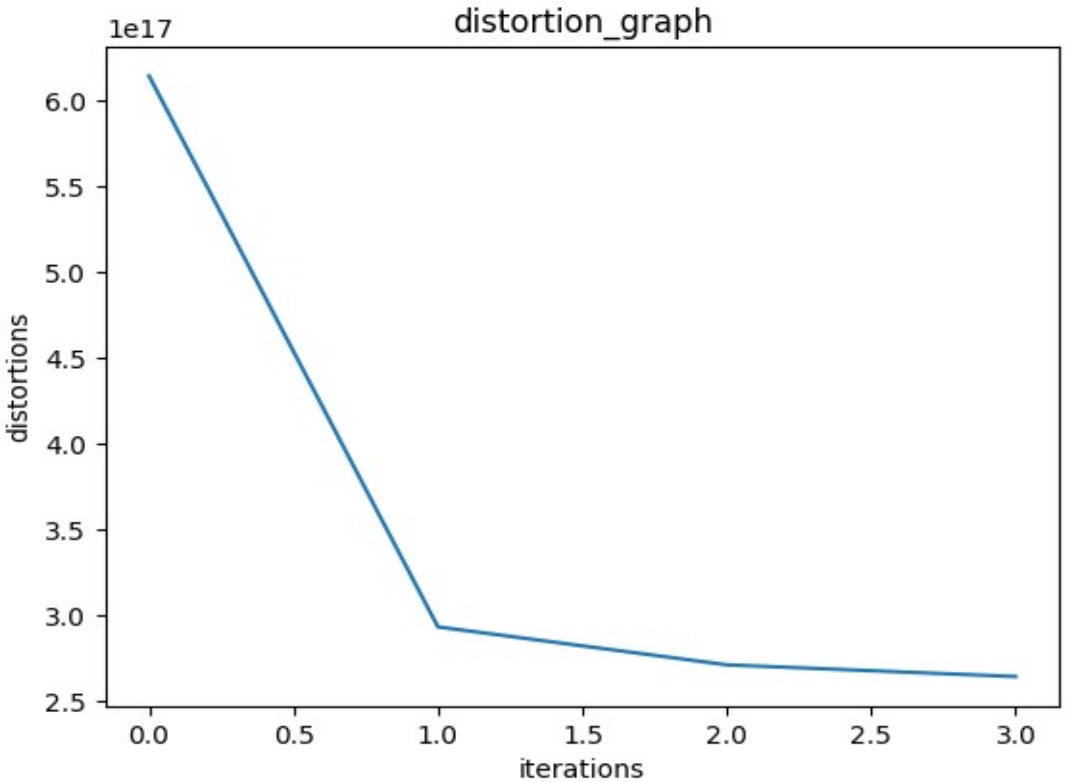

-At k=9, iterations =4



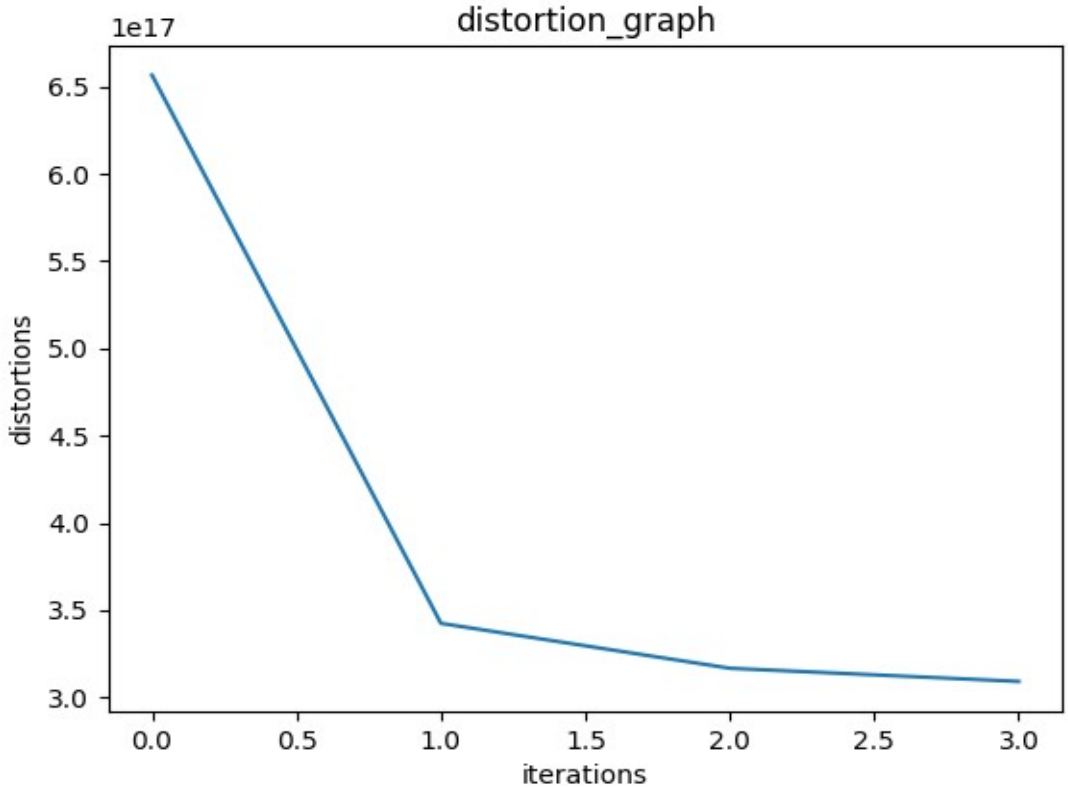
-At k=8, iterations =4:



-At k=7, iterations =4:




-At k=6, iterations=4:



although different values for k with different restart but it gives the same result that distortion measure always decreases.

Method of distortion:

reference:

$$\sum\left(Data\ X_i\ -\ Centroid\ X\right)^2 + \left(Data\ Y_i\ -\ Centroid\ Y\right)^2 \ldots$$

https://avidml.wordpress.com/2016/10/29/easily-understand-k-means-clustering/