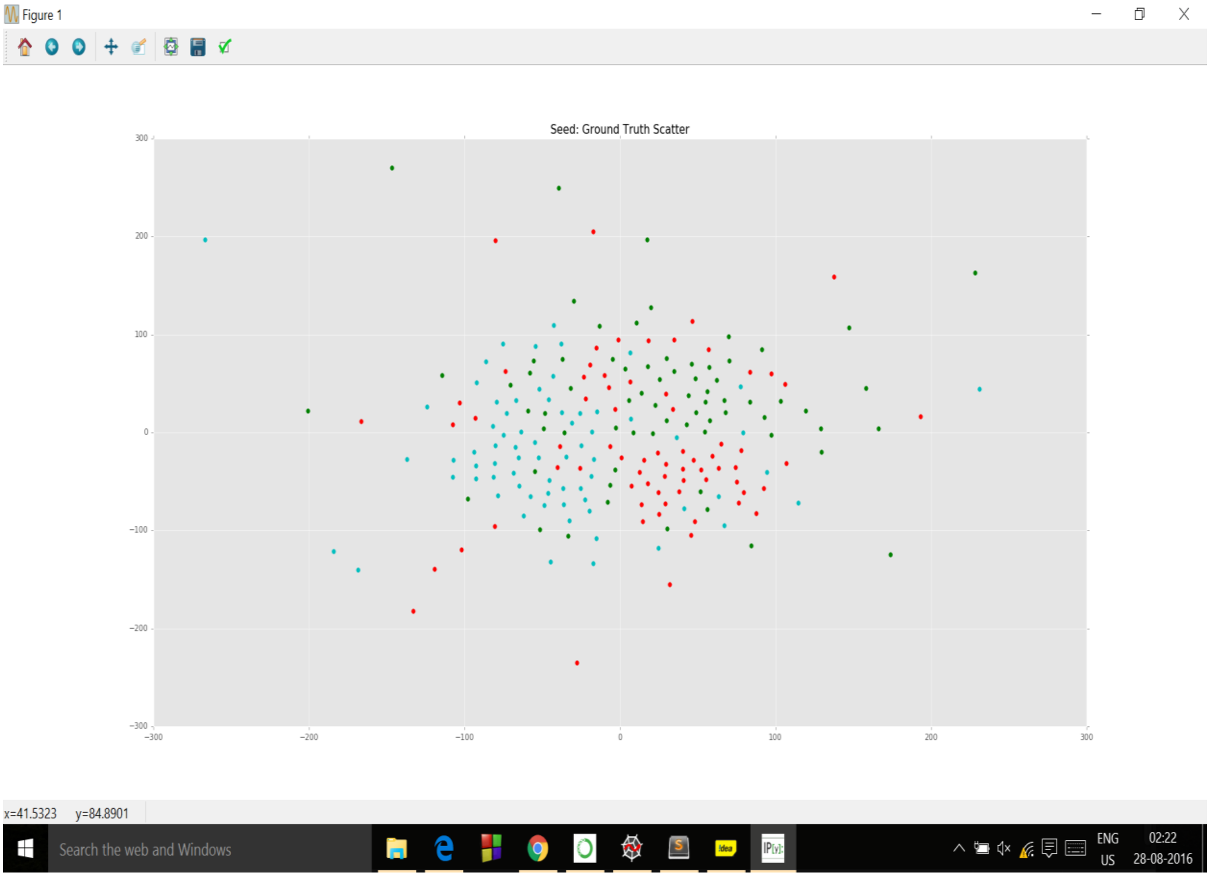
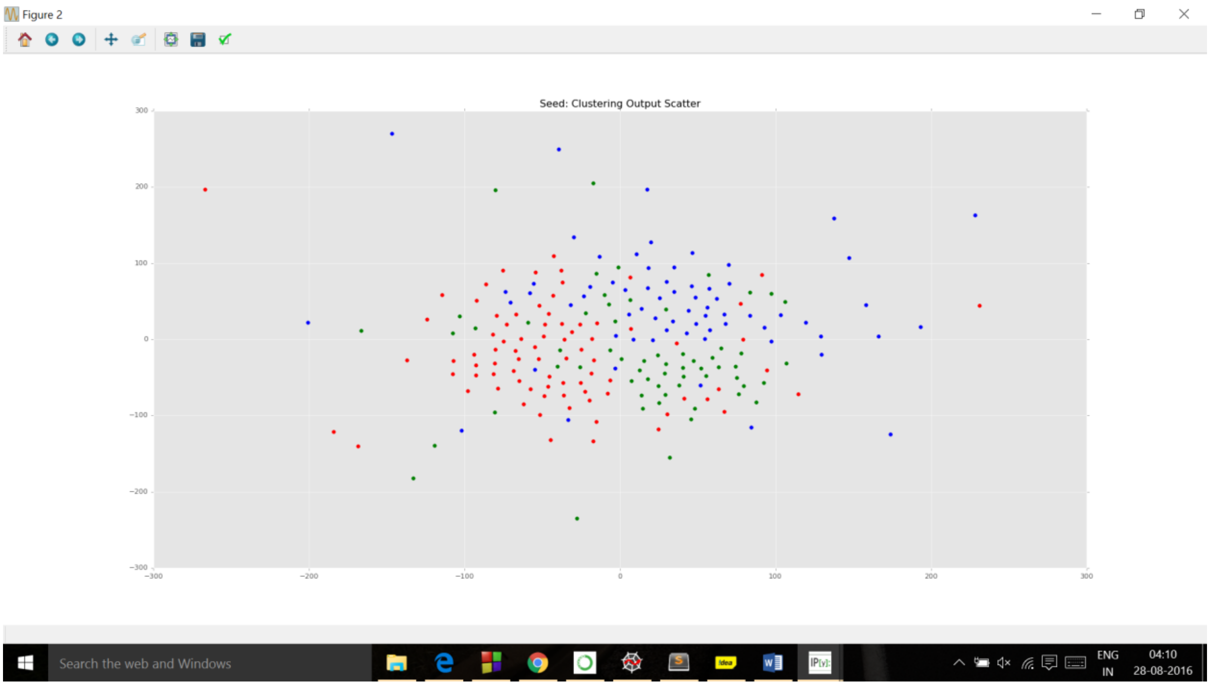
Report

1). Seeds_Dataset:

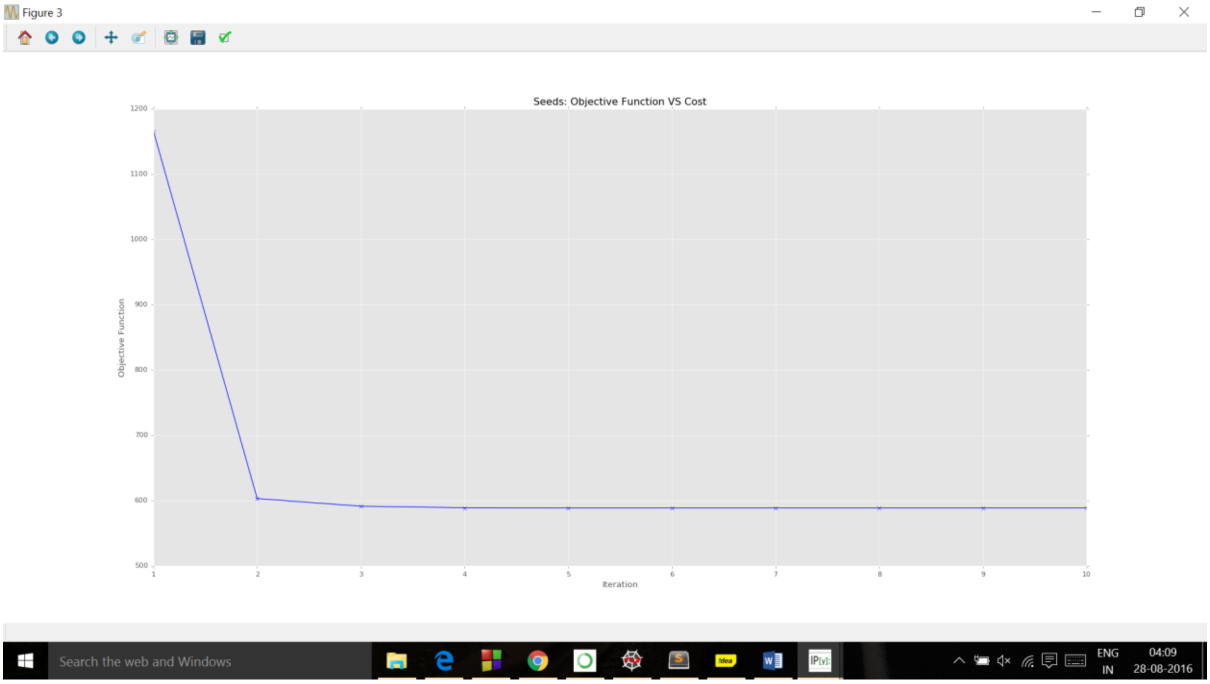
Ground Truth, K=3:



Using my_Kmeans, K=3:

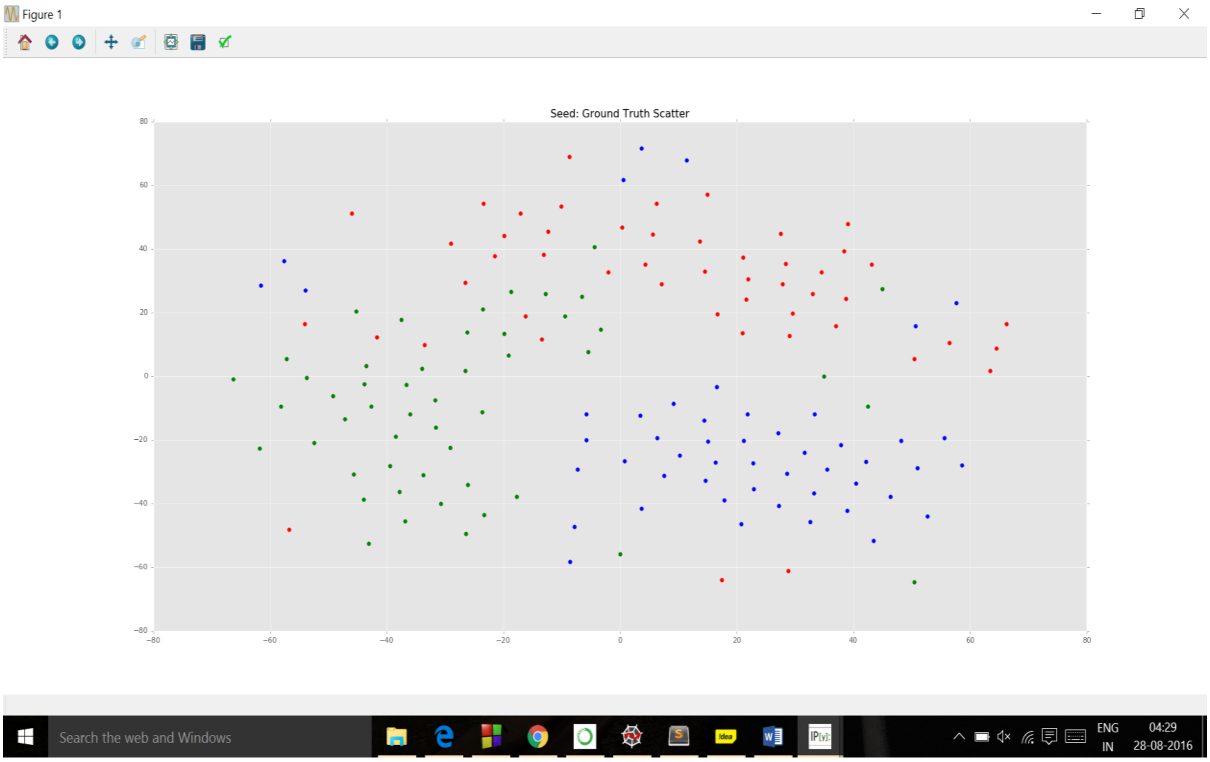


Objective Function Vs Cost:

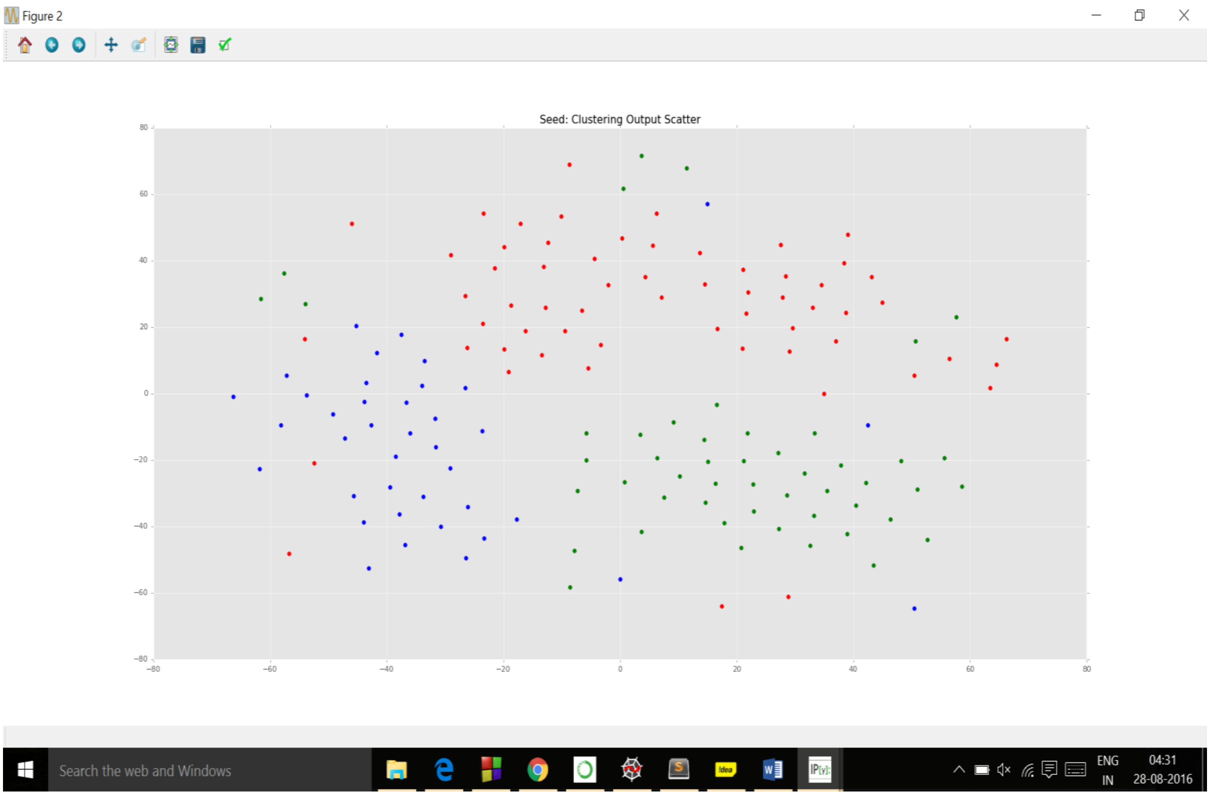


2). Iris Data:

Ground Truth, K=3:



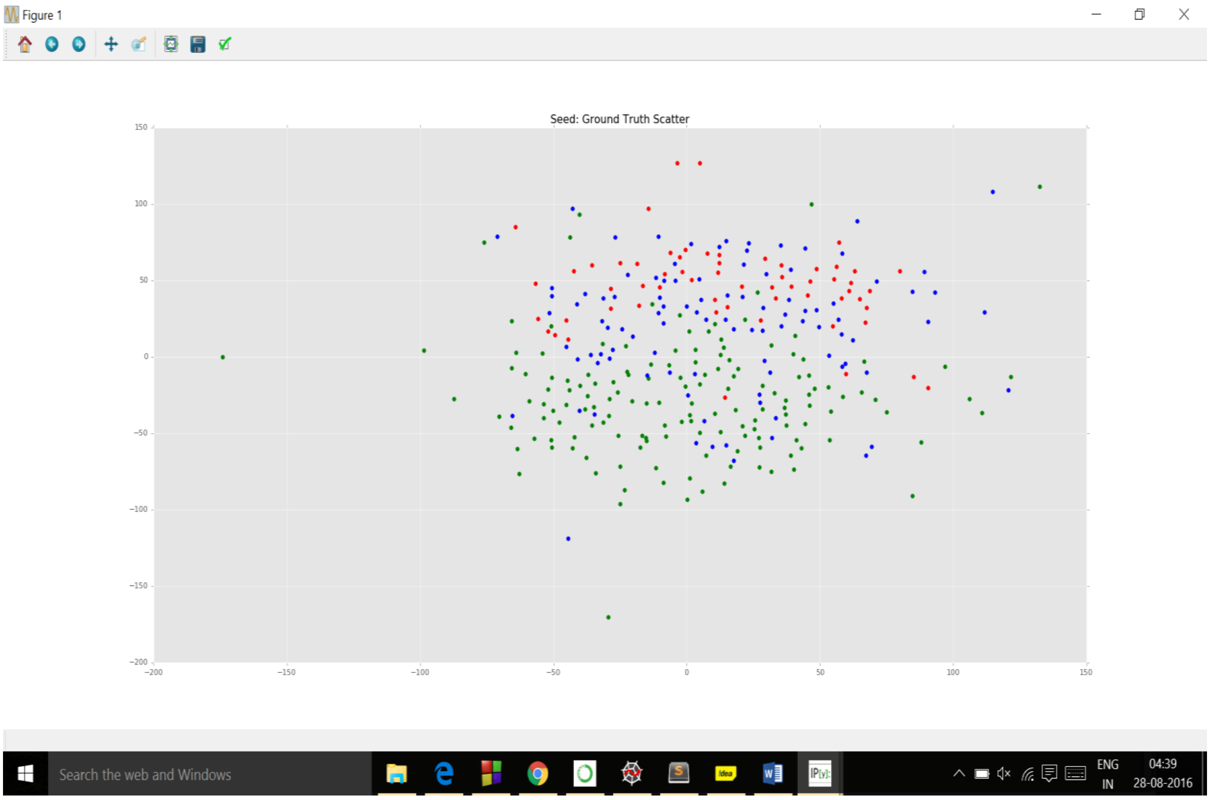
Using my_Kmeans, K=3:



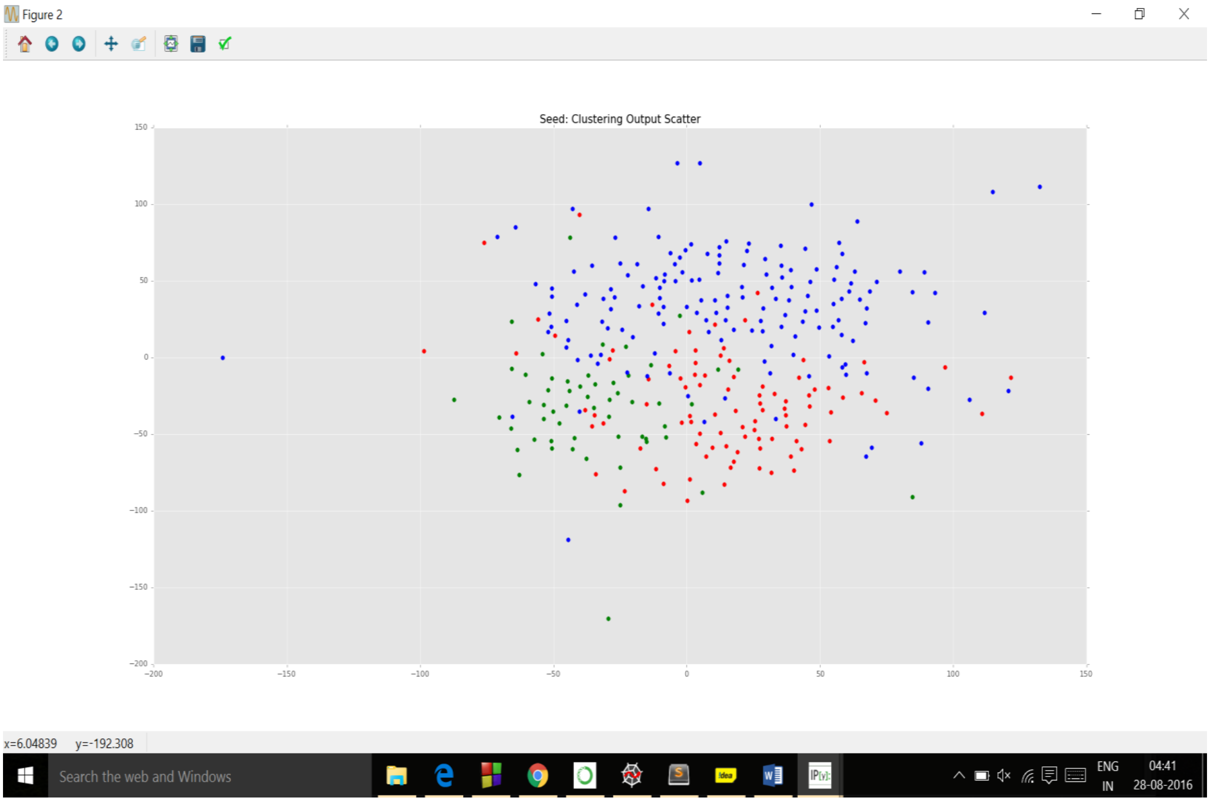
Objective Function Vs Cost:

Vertebral Column Data Set:

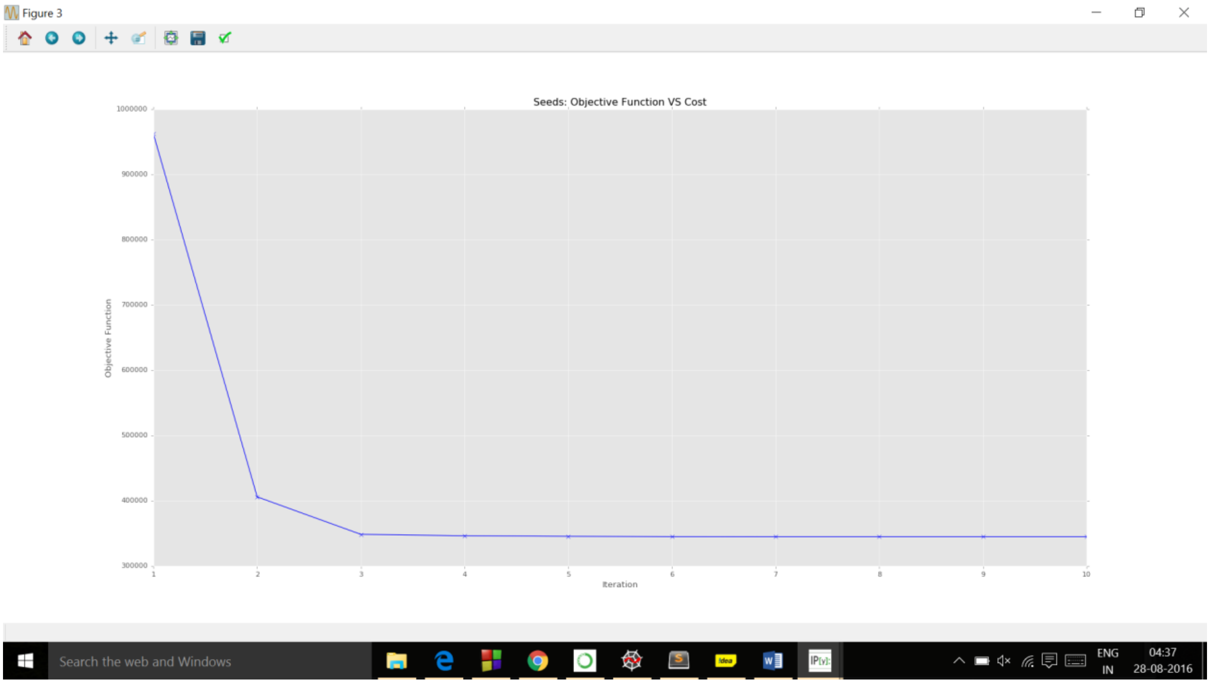
Ground Truth, K=3:



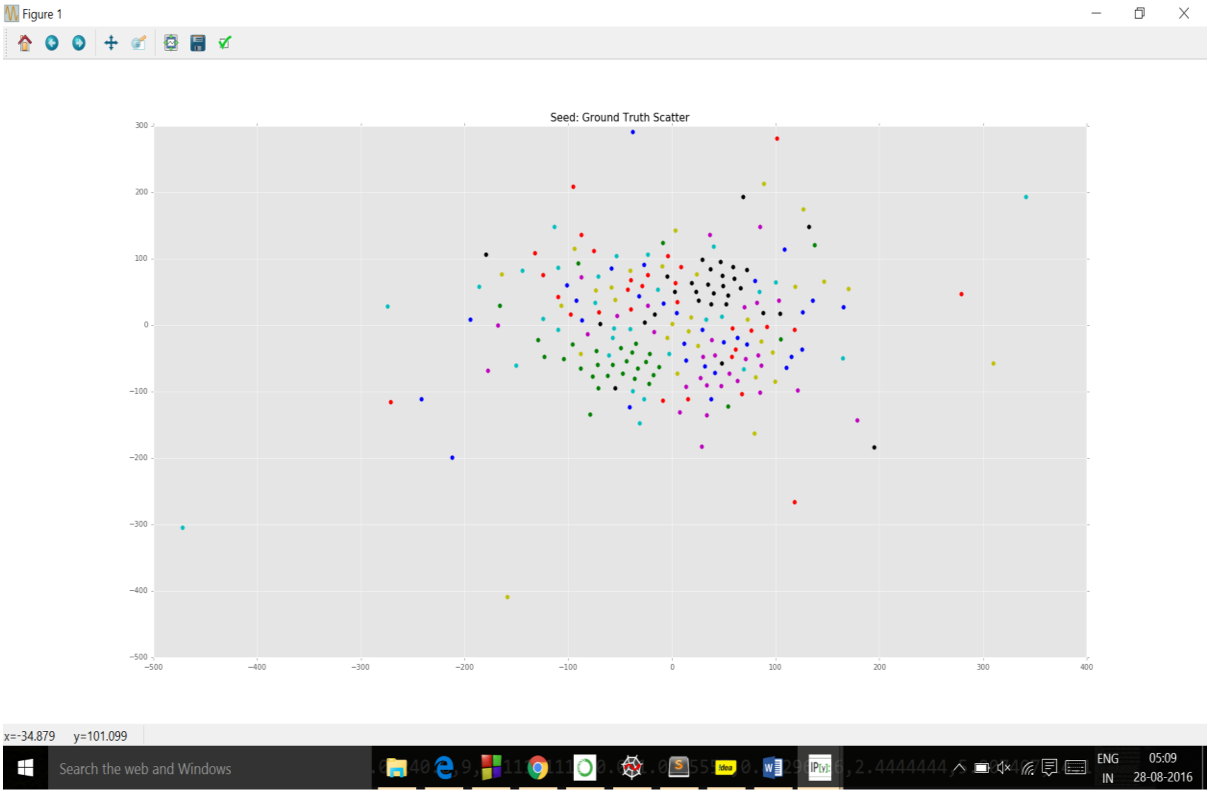
Using my_Kmeans, K=3:



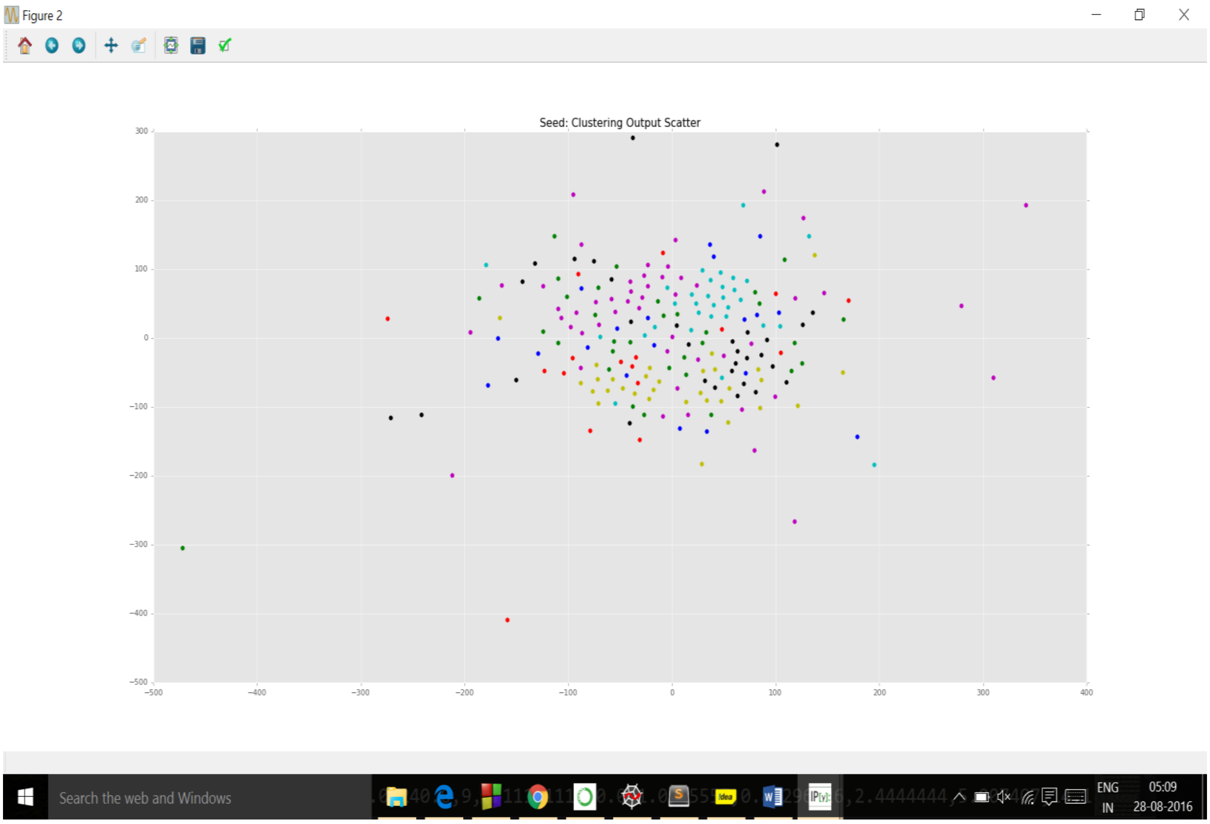
Objective Function Vs Cost:



4). Segmentation:



Using my_Kmeans, K=3:



Objective Function Vs Cost:

Quantitative estimation:

	K=2			
Dataset Name	NMI	AMI	RI	ARI
Iris	0.551633383976	0.427184431359	0.7291592617908408	0.4670219429349123
Segmentation	0.299947658256	0.149864066226	0.47980861244019135	0.10062760602331203
Seeds	0.497313968384	0.384529798943	0.6890179995443153	0.42049630739845095
Vertebral	0.419038134651	0.329669017523	0.637141664056791	0.28970200403426377

	K=true value			
Dataset Name	NMI	AMI	RI	ARI
Iris(K=3)	0.699469093732	0.695003444857	0.8734654818865344	0.7147364117956806
Segmentation (K=7)	0.523787103428	0.478248222137	0.8227842333105491	0.35471468279415935
Seeds(K=3)	0.705525783533	0.700682062285	0.8722624743677375	0.7122251532084475
Vertebral (K=3)	0.416465568369	0.404620310849	0.6744086021505377	0.3121816342056434

	K=12			
Dataset Name	NMI	AMI	RI	ARI
Iris	0.435507320166	0.284526014095	0.6269798657718121	0.22897944145539917
Segmentation	0.451604451108	0.386696963035	0.7115652768284348	0.2775441939853539
Seeds	0.276036934885	0.179330387355	0.5372020961494646	0.14678816433352465
Vertebral	0.36556497863	0.232772095463	0.6461384278108362	0.17096822739454343

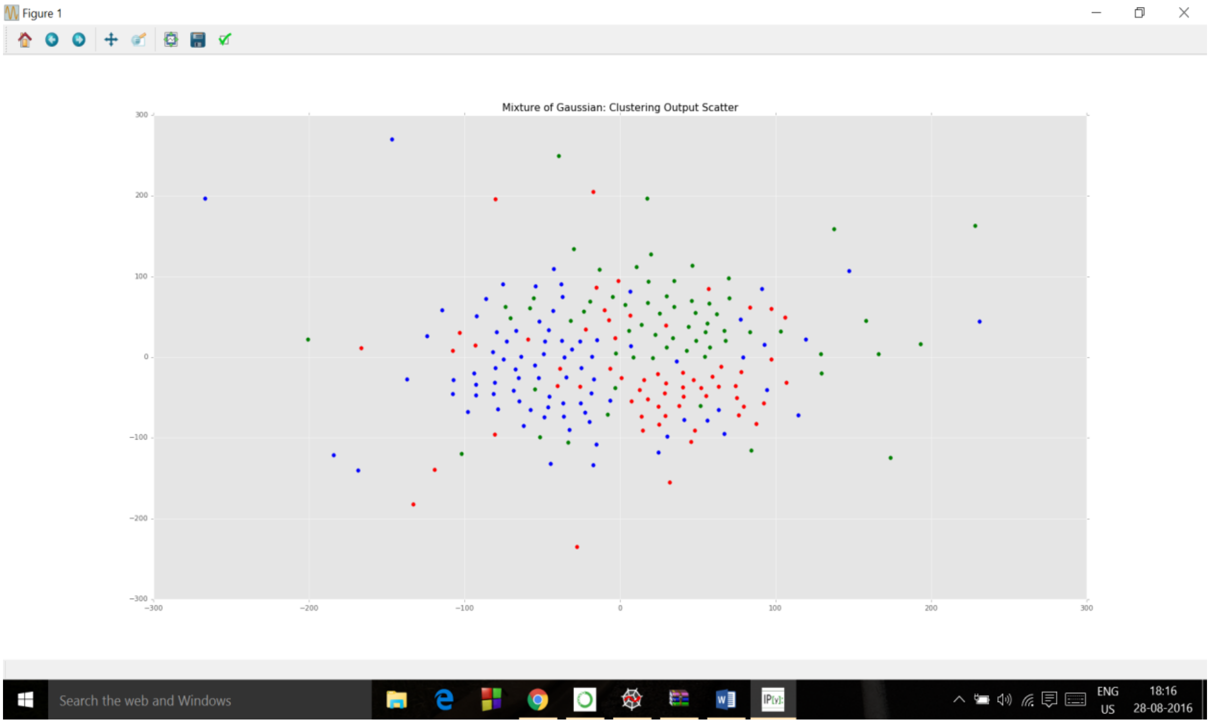
So, we observe that for Seeds Dataset, we see that if no of clusters is equal to 3(which is ground truth) then the result is best as all NMI = 0.705525783533, AMI = 0.695003444857, RI = 0.8734654818865344, and ARI = 0.7122251532084475 are nearer to 1. For K=2, the measures are not as good when K=3. For K=12 too measures decrease by great amount.
However cost decreases with increase in no of clusters.
We observe that for Iris Dataset, we see that if no of clusters is equal to 3(which is ground truth) then the result is best as all NMI = 0.699469093732, AMI = 0.700682062285, RI = 0.8722624743677375, and ARI = 0.35471468279415935 are nearer to 1. For K=2, the measures are not as good when K=3. For K=12 too measures decrease by great amount. So the best output is when K=3. Therefore our unsupervised algorithm is working correctly.
However cost decreases with increase in no of clusters.
We observe that for Vertebral Dataset, we see that if no of clusters is equal to 3(which is ground truth) then the result is best as all NMI = 0.416465568369, AMI = 0.404620310849, RI = 0.6744086021505377, and ARI = 0.3121816342056434 are nearer to 1. For K=2, the measures are not as good when K=3. For K=12 too measures decrease by great amount. So the best output is when K=3.
Cost decreases with increase in no of clusters.
We observe that for Segmentation Dataset, we see that if no of clusters is equal to 7(which is ground truth) then the result is best as all NMI = 0.523787103428, AMI = 0.478248222137,

RI = 0.8227842333105491, and ARI = 0.35471468279415935 are nearer to 1. For K=12, the measures are not as good when K=3. For K=2 too measures decrease by great amount. So the best output is when K=7. Here K=12 is better than K=2.
Cost decreases with increase in no of clusters.

Using the plots we see that cost function decreases with no of iterations. There is knee or elbow indicating the no of clusters that should be there.
Iris = 3
Segmentation = 7
Seeds = 3
Vertebral = 3
TSNE plots give visualisation of the dataset in 2D for my clustering and the ground truth available.

Bonus:

Seeds Dataset:



Chosen Model:

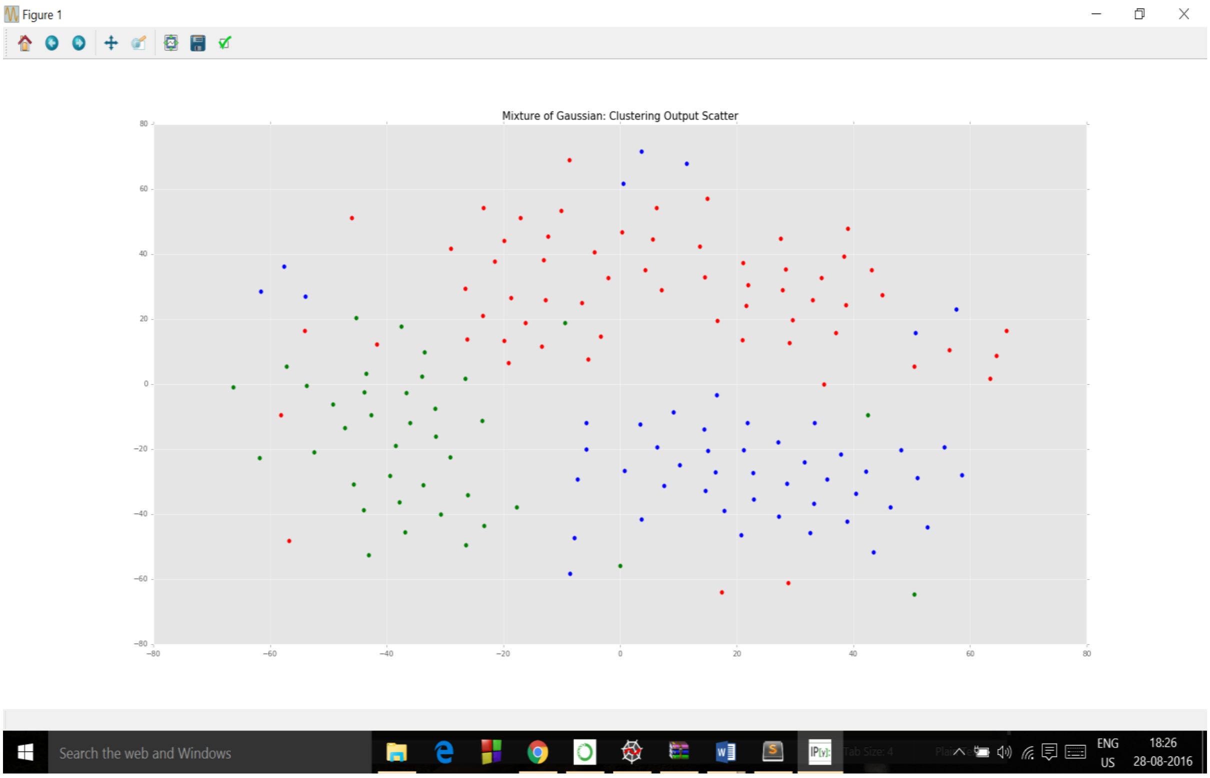
Means of Chosen Gaussians

[[18.66937846 16.27588423 0.8848712 6.20396821 3.7164789
3.57857812 6.06210976]
[11.98270165 13.28104861 0.85287051 5.2266245 2.8784518
4.39921997 5.06446933]
[14.8670757 14.55690736 0.88120926 5.59286773 3.30940102
2.90216388 5.21902277]]

Covariance of Chosen Gaussians

[[1.27252422e+00 2.37090021e-01 1.21802535e-03 4.88112117e-02
2.45890378e-02 1.49693979e+00 4.97759813e-02]
[5.54190680e-01 1.21770144e-01 1.49684531e-03 2.09728480e-02
2.35919253e-02 2.32028537e+00 4.04170300e-02]
[6.98121780e-01 1.70222665e-01 1.23957300e-03 3.37184088e-02
1.81923039e-02 1.60514840e+00 8.34464996e-02]]

Iris Datset:



Chosen Model:

Means of Chosen Gaussians

[[5.90845874 2.74067981 4.36295651 1.38571005]

 [5.006   3.418   1.464   0.244   ]

 [6.74388131 3.05099112 5.6461753   2.07166896]]

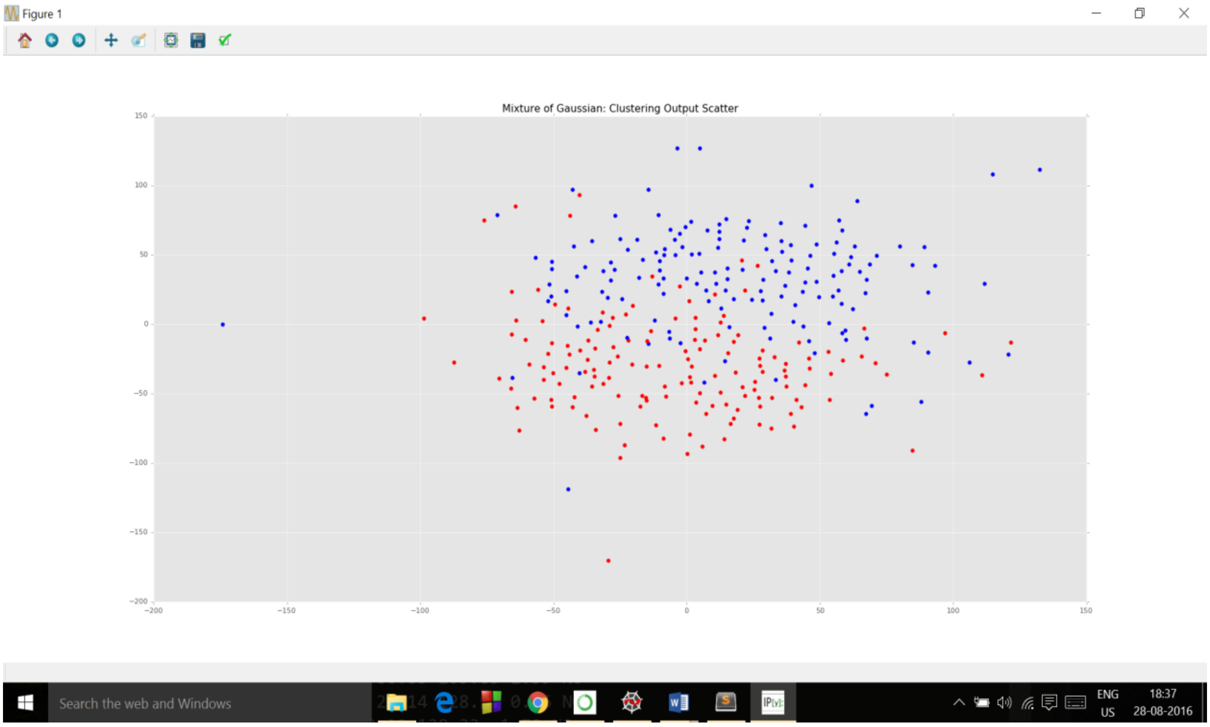
Covariance of Chosen Gaussians

[[0.23312203 0.08959638 0.2683054   0.06211418]

 [0.122764   0.143276   0.030504   0.012264]

 [0.30984681 0.08372304 0.28132848 0.06837876]]

Vertebral Dataset:



Chosen Model:

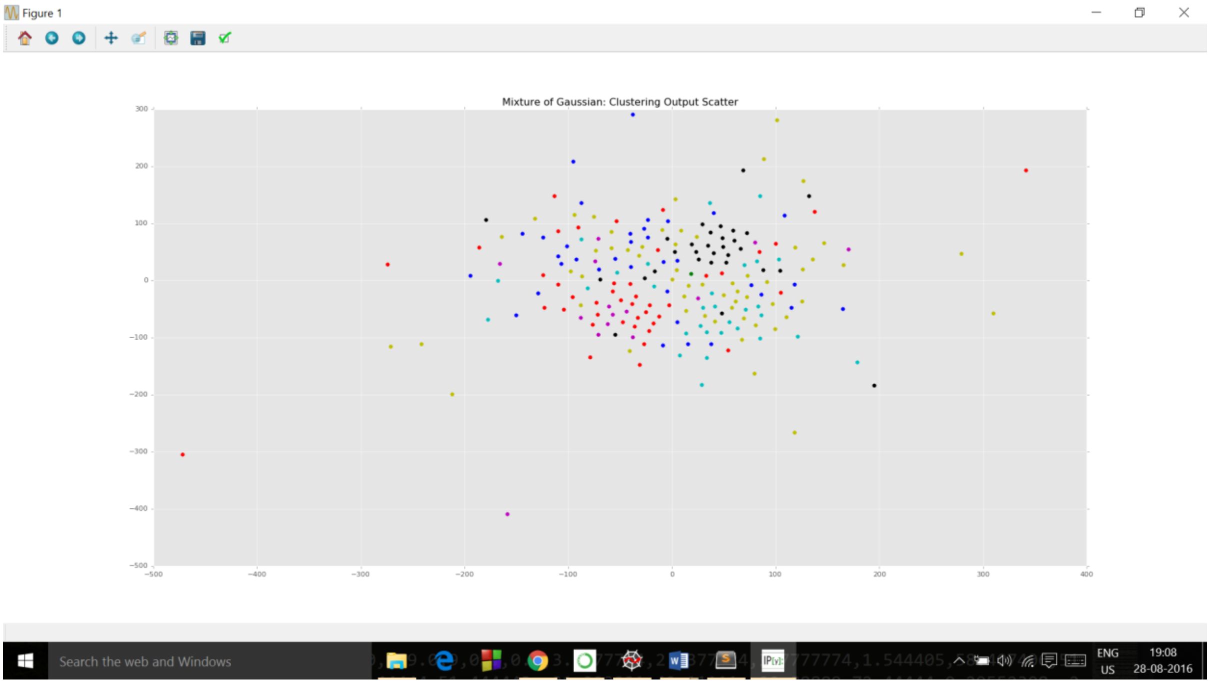
Means of Chosen Gaussians

[[73.6785054   21.68635771   65.7857022   51.99232126 114.17757307
   48.35546041]
 [47.57242441   13.48053297   38.34684875   34.09229623 121.59027654
    4.66969569]]

Covariance of Chosen Gaussians

[[171.9612352   121.1893693   223.5465201   136.91920255
   204.47123905 1781.72597945]
 [80.46465036   45.57483083   87.65323362   62.80330166
   122.44414715   93.07058232]]

Segmentation:



Chosen Model:

Means of Chosen Gaussians

[[1.50711402e+002 1.39401088e+002 9.00000000e+000 1.03490624e-002
2.06981248e-002 2.25924455e+000 1.67602513e+000 2.39076863e+000
1.59458307e+000 5.22574280e+001 4.68424768e+001 6.41936641e+001
4.57361438e+001 -1.62448534e+001 3.58087057e+001 -1.95638522e+001
6.41936641e+001 2.88697943e-001 -2.02123632e+000]
[9.46973363e+001 1.05084174e+002 9.00000000e+000 6.14728571e-003
3.16061655e-003 1.98009499e+000 2.98564928e+000 2.88989295e+000
1.09634699e+001 1.98006611e+001 1.71170195e+001 2.65558747e+001
1.57290893e+001 -8.05092430e+000 2.02656404e+001 -1.22147161e+001
2.65938022e+001 4.52572523e-001 -1.92472881e+000]
[8.00000000e+001 8.70000000e+001 9.00000000e+000 0.00000000e+000
1.11111110e-001 2.43888910e+001 5.72996400e+002 4.47222250e+001
1.38632920e+003 6.74444400e+001 5.87777800e+001 7.90000000e+001
6.45555600e+001 -2.60000000e+001 3.46666680e+001 -8.66666700e+000
7.90000000e+001 3.06281270e-001 -2.42212720e+000]

[1.16400000e+002 4.58666667e+001 9.00000000e+000 7.40740733e-003
1.26882315e-041 8.31481465e-001 5.80651252e-001 1.13703719e+000
7.98594144e-001 1.19069135e+002 1.08018520e+002 1.35174076e+002
1.14014814e+002 -3.31518515e+001 4.83148152e+001 -1.51629631e+001
1.35174076e+002 2.05043759e-001 -2.32403933e+000]
[1.45601520e+002 1.31037080e+002 9.00000000e+000 1.62040479e-009
1.60484162e-002 7.53791445e+000 2.53593156e+001 1.12265156e+001
3.52839063e+001 3.99673569e+001 3.49730875e+001 4.87388909e+001
3.61900921e+001 -1.49828093e+001 2.63146044e+001 -1.13317950e+001
4.87469150e+001 3.09945290e-001 -2.17889087e+000]
[1.20334697e+002 1.17709956e+002 9.00000000e+000 2.05469887e-003
0.00000000e+000 6.78506805e-001 4.10266185e-001 7.72232691e-001
5.14774206e-001 3.61992389e+000 2.95987049e+000 5.86106303e+000
2.03883817e+000 -1.98016015e+000 6.72341737e+000 -4.74325717e+000
6.01296926e+000 6.57349712e-001 -1.63595585e+000]
[1.30700000e+002 2.03500000e+002 9.00000000e+000 2.59259257e-002
3.84267826e-146 1.50740747e+000 1.97301749e+000 2.14259257e+000
2.06424624e+000 1.49777777e+001 1.19111111e+001 1.36148151e+001
1.94074074e+001 -9.20000004e+000 -4.08888889e+000 1.32888890e+001
1.94185185e+001 4.10565311e-001 2.28777013e+000]]

Covariance of Chosen Gaussians

[[4.54624097e+03 3.45094805e+03 1.00000000e-03 2.04279272e-03
4.02127500e-03 1.76452598e+00 1.58428430e+00 1.73892374e+00
6.45263507e-01 5.70341042e+01 4.18991989e+01 9.40952291e+01
4.51063503e+01 2.28925481e+01 5.73552305e+01 9.74129702e+00
9.40952291e+01 1.81411394e-03 6.73036142e-03]
[5.05076184e+03 1.19723163e+03 1.00000000e-03 1.64524262e-03
1.34119012e-03 1.46723412e+00 1.41735252e+01 1.13666690e+01
8.89335367e+02 3.21140830e+01 3.61048879e+01 4.79381199e+01
2.57594930e+01 5.21559379e+01 5.08607160e+01 1.81286031e+01
4.76838075e+01 1.55108623e-02 9.64653045e-02]

[1.00000001e-03 1.00000002e-03 1.00000000e-03 1.00000000e-03
1.00000000e-03 1.00000000e-03 1.00000076e-03 1.00000000e-03
1.00000419e-03 1.00000001e-03 1.00000001e-03 1.00000001e-03
1.00000001e-03 1.00000000e-03 1.00000000e-03 1.00000000e-03
1.00000001e-03 1.00000000e-03 1.00000000e-03]
[3.32384100e+03 6.33783222e+02 1.00000000e-03 1.76817557e-03
1.00000000e-03 9.71901056e-02 1.15068102e-01 3.72138419e-01
7.56114664e-01 1.71798027e+02 2.42755000e+02 8.61003559e+01
2.15382703e+02 7.15536583e+01 1.51813425e+02 3.62304108e+01
8.61003559e+01 5.01920496e-03 8.71115830e-03]
[5.79407665e+03 1.90551076e+03 1.00000000e-03 1.00000018e-03
2.52560603e-03 6.14175093e+01 4.59492166e+03 8.73639813e+01
2.57828301e+03 8.49576621e+01 8.24455438e+01 9.30190865e+01
8.43726523e+01 2.44034552e+01 1.33386613e+01 6.93985685e+00
9.30080557e+01 9.00200580e-03 3.55005471e-02]
[5.84555056e+03 8.87015724e+02 1.00000000e-03 1.22407808e-03
1.00000000e-03 2.25107038e-01 2.00195779e-01 3.56158562e-01
4.15989580e-01 6.50670188e+00 9.78273435e+00 1.25568747e+01
2.92906594e+00 1.97065921e+01 2.38845041e+01 8.14902607e+00
1.28711668e+01 7.55818274e-02 5.28314077e-01]
[6.28241100e+03 6.88851000e+02 1.00000000e-03 3.20850476e-03
1.00000000e-03 5.24813440e-01 8.96924457e+00 8.65543941e-01
2.91174028e+00 2.48370869e+01 1.55848677e+01 3.03053087e+01
3.30506579e+01 1.68161443e+01 1.06334281e+01 1.24196015e+01
3.28616456e+01 8.01947167e-03 7.11505032e-02]]

For the best ones the evaluation matrix is given as:

	K=true value			
Dataset Name	NMI	AMI	RI	ARI
Iris(K=3)	0.785696319068	0.775960558744	0.8922595078299776	0.7583384522539416
Segmentation (K=7)	0.637795903482	0.584569247381	0.8577352472089315	0.47746234113605246
Seeds(K=3)	0.680466444803	0.674569727235	0.8572795625427204	0.6789831632262734
Vertebral (K=2)	0.374600044759	0.304432764963	0.7062532623447124	0.4121249543788916

K=12

	K=12			
Dataset Name	NMI	AMI	RI	ARI
Iris	0.609022601286	0.39019008934	0.74917225950783	0.3015504303777993
Segmentation	0.621537743238	0.545109289615	0.8764183185235817	0.4380361285717118
Seeds	0.521165787651	0.336155683136	0.729915698336751	0.25096762529243416
Vertebral	0.345093876718	0.22102246453	0.6511118070779831	0.13605161089243753

K = 5

	K=5			
Dataset Name	NMI	AMI	RI	ARI
Iris	0.719474660815	0.608782530062	0.608782530062	0.6120507616294565
Segmentation	0.526159501559	0.416173746265	0.765185691501481	0.3313430482095195
Seeds	0.620185905124	0.510922579699	0.8076555023923445	0.519647490260244
Vertebral	0.373538450282	0.332284397365	0.6594007725232279	0.24646745423053043

So, we observe that for Seeds Dataset, we see that if no of clusters is equal to 3(which is ground truth) then the result is best as all NMI = 0.680466444803, AMI = 0.674569727235, RI = 0.8572795625427204, ARI = 0.6789831632262734 are nearer to 1. For K=5, the measures are not as good when K=3. For K=12 too measures decrease by great amount.
We observe that for Iris Dataset, we see that if no of clusters is equal to 3(which is ground truth) then the result is best as all NMI = 0.785696319068, AMI = 0.775960558744, RI =

0.8922595078299776, ARI = 0.7583384522539416 are nearer to 1. For K=5, the measures are not as good when K=3. For K=12 too measures decrease by great amount. So the best output is when K=3.
We observe that for Vertebral Dataset, we see that if no of clusters is equal to 2(which is ground truth) then the result is best as all NMI = 0.374600044759, AMI = 0.304432764963, RI = 0.7062532623447124, ARI = 0.4121249543788916 are nearer to 1. For K=5, the measures are not as good when K=2. For K=12 too measures decrease by great amount. So the best output is when K=2. So this means that abnormal DH and SL can be combined in 1, therby giving two clusters NO and abnormal ones.
We observe that for Segmentation Dataset, we see that if no of clusters is equal to 7(which is ground truth) then the result is best as all NMI = 0.637795903482, AMI = 0.584569247381, RI = 0.8577352472089315, AMI = 0.47746234113605246 are nearer to 1. For K=12, the measures are not as good when K=3. For K=5 too measures decrease by great amount. So the best output is when K=7.

Iris = 3
Segmentation = 7
Seeds = 3
Vertebral = 2
TSNE plots give visualisation of the dataset in 2D for my clustering and the ground truth available.
The purposed mean and variance gives the best result. The evaluation matrix is better in these cases.