Name: Mohammad Osama Ataullah
Roll No. : MT21127
Name: Sehban Fazili
Roll No. : MT21143

DMG Assignment - 3

1) Three modelling techniques used by us are as follows:
    a) K-Means Clustering
    b) Agglomerative Clustering
    c) Spectral Clustering
    d) Gaussian Mixture Modelling

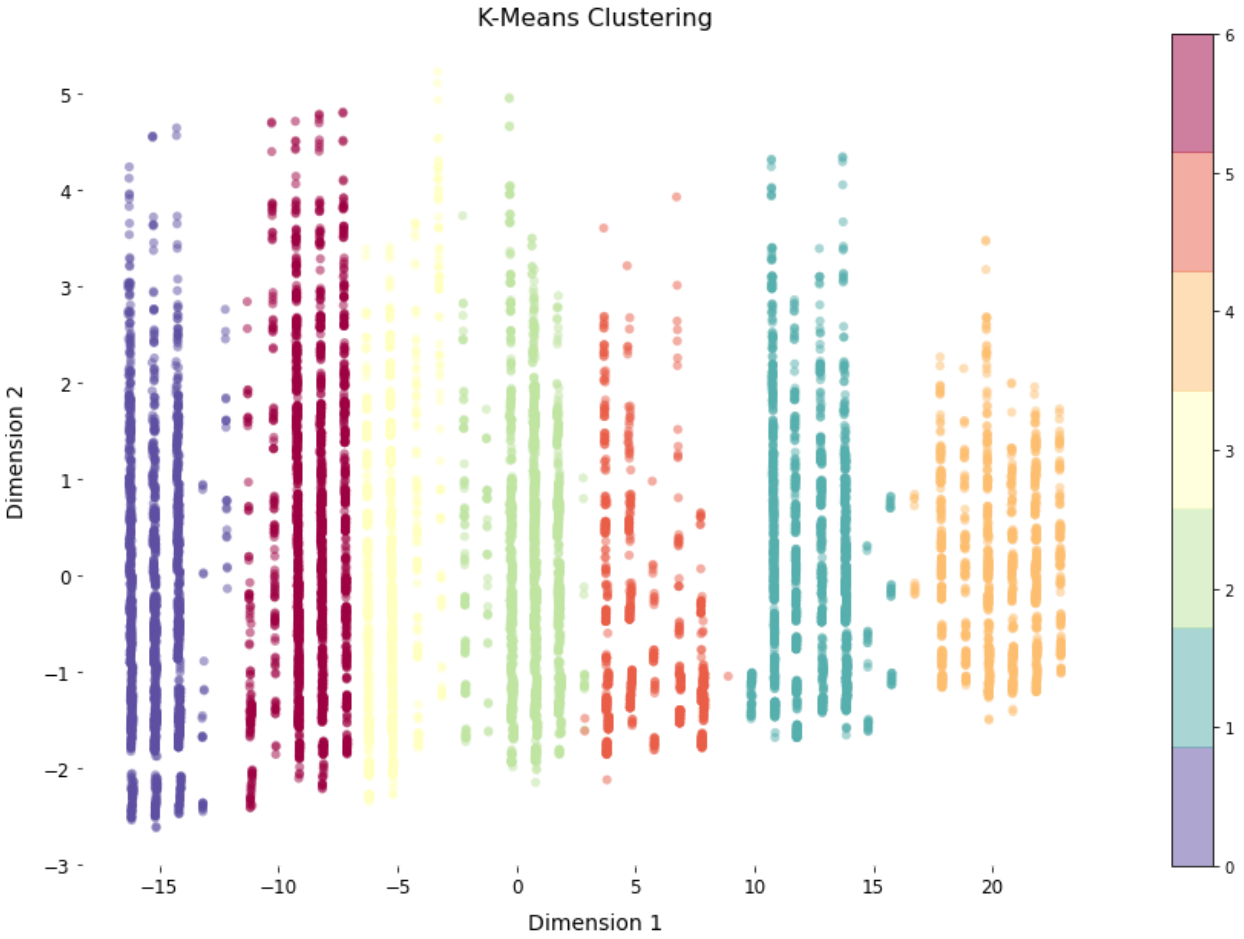
(i) Centroid/representative object/prototype of each cluster for every model.
(ii) Visualization of the clusters.
(iii) Compare your cluster distribution with the true label count.

The above three questions are shown for each clustering techniques in the following para.

a) K-Means Clustering
    - Centroids:

	Elevation	Aspect	Slope	Hillshade_9am	Hillshade_Noon	Horizontal_Distance_To_Hydrology	Vertical_Distance_To_Hydrology	Horizontal_Distance_To_Fire_Points	Soil_Type	Wilderness	ClusterID
0	1.554033	1.639777	1.321157	0.030949	8.117707e-03	1.743278	1.526636	1.333841	37.976662	0.975647	0
1	1.176357	1.610950	1.281977	0.081395	1.550388e-02	1.257752	1.470930	1.078488	10.386143	1.816376	1
2	0.393574	1.609438	1.356928	0.032129	3.514056e-03	1.412651	1.263554	1.288655	22.187249	1.097892	2
3	0.650888	1.581657	1.297633	0.065089	1.597633e-02	1.530178	1.461538	1.161538	28.192308	0.244970	3
4	1.480474	1.437238	1.304045	0.041841	3.486750e-03	0.923989	1.464435	1.037657	2.552999	2.601116	4
5	0.842538	1.547591	1.213866	0.009401	-6.418477e-17	0.820212	1.030552	1.088132	17.340776	0.943596	5
6	0.702367	1.579623	1.334136	0.051344	9.626955e-03	1.849178	1.630566	1.314079	31.438428	1.801444	6

- Cluster Visualization:



- Cluster Distribution
  Original Distribution of Classes

```
target
1    4112
2    4711
3    1056
4     175
5     597
6     686
7    1158
dtype: int64
```

  After K-Means

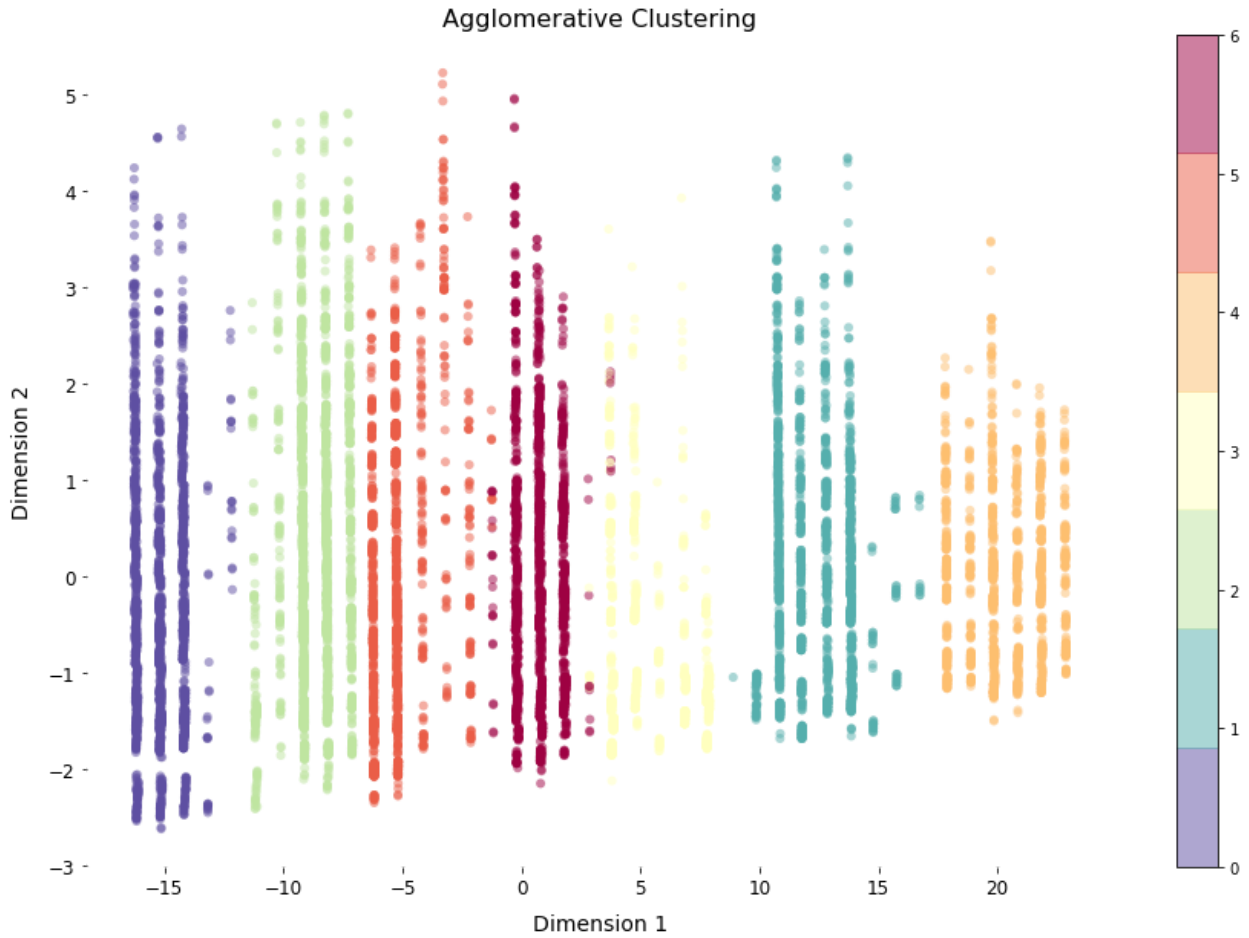
		count
clusters	labels	
0	1	1002
	2	155
	7	814
1	1	318
	2	973
	3	329
	4	35
	5	156
	6	250
	7	3
2	1	879
	2	959
	5	53
	6	20
	7	81
3	1	578
	2	890
	5	149
	7	73
4	1	11
	2	237
	3	665
	4	127
	5	62
	6	327
	7	5
5	1	306
	2	372
	3	48
	4	13
	5	64
	6	46
	7	2
6	1	1018
	2	1125
	3	14
	5	113
	6	43
	7	180

b) Agglomerative Clustering

- Centroids:

	Elevation	Aspect	Slope	Hillshade_9am	Hillshade_Noon	Horizontal_Distance_To_Hydrology	Vertical_Distance_To_Hydrology	Horizontal_Distance_To_Fire_Points	Soil_Type	Wilderness	ClusterID
0	1.554033	1.639777	1.321157	0.030949	0.008118	1.743278	1.526636	1.333841	37.976662	0.975647	0
1	1.170039	1.610308	1.280347	0.080925	0.015414	1.257707	1.468208	1.077553	10.364644	1.807322	1
2	0.702367	1.579623	1.334136	0.051344	0.009627	1.849178	1.630566	1.314079	31.438428	1.801444	2
3	0.812941	1.542353	1.214118	0.009412	0.000000	0.792941	1.022353	1.088235	17.352941	0.949412	3
4	1.491918	1.436402	1.306395	0.042164	0.003514	0.920590	1.468025	1.038651	2.526353	2.621223	4
5	0.652101	1.563585	1.299160	0.061625	0.015126	1.535014	1.448739	1.178711	28.017367	0.333333	5
6	0.392725	1.630469	1.358461	0.033737	0.003690	1.414866	1.269373	1.278861	22.047443	1.053769	6

- Cluster Visualization:



- Cluster Distribution
  Original Distribution of Classes

```
target
1    4112
2    4711
3    1056
4     175
5     597
6     686
7    1158
dtype: int64
```

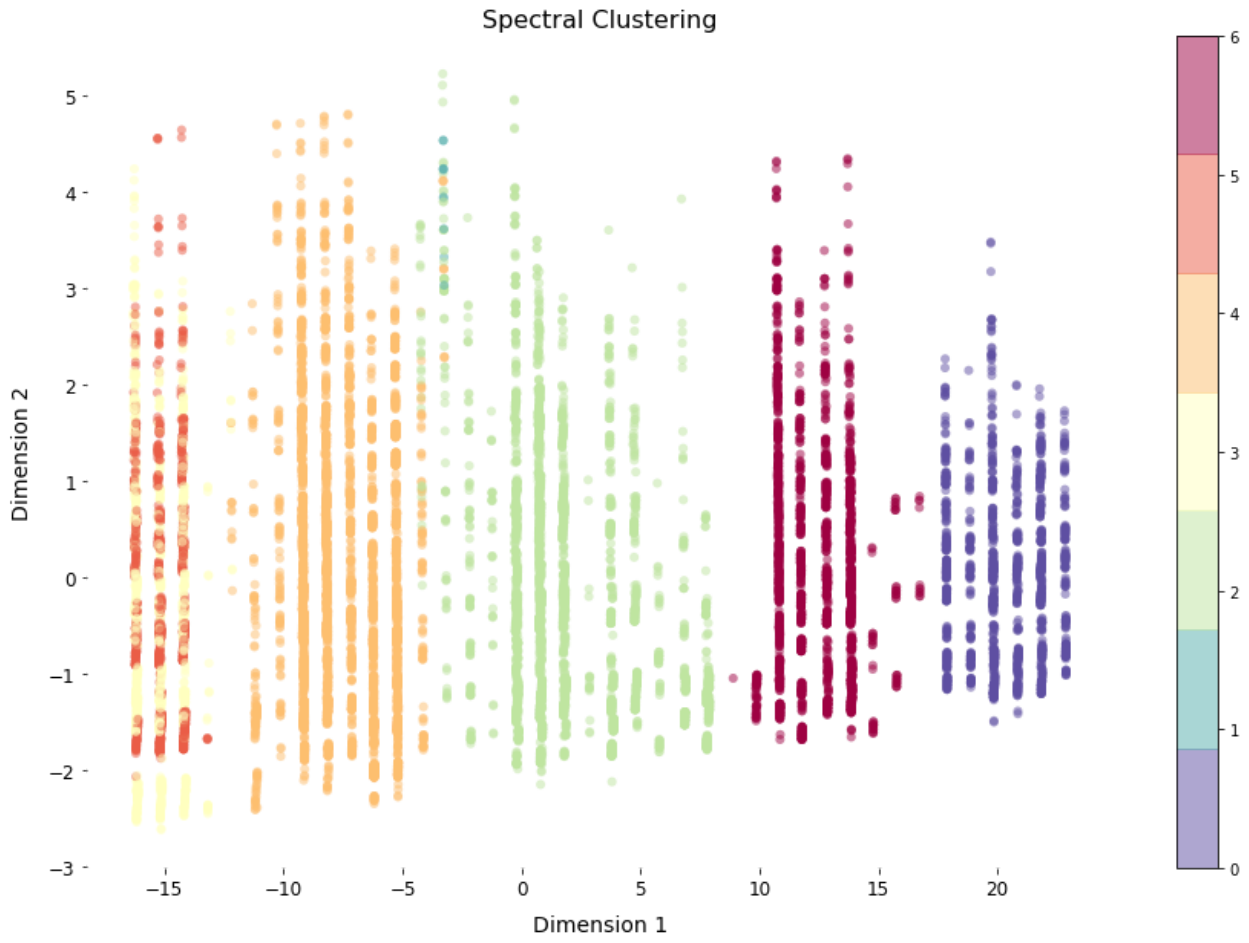
After Agglomerative Clustering

clusters	labels	count
0	1	1002
	2	155
	7	814
1	1	318
	2	984
	3	329
	4	35
	5	156
	6	251
	7	3
2	1	1018
	2	1125
	3	14
	5	113
	6	43
	7	180
3	1	303
	2	374
	3	48
	4	13
	5	64
	6	45
	7	3
4	1	11
	2	226
	3	665
	4	127
	5	62
	6	327
	7	5
5	1	588
	2	963
	5	161
	7	73
6	1	872
	2	884
	5	41
	6	20
	7	80

c) Spectral Clustering

- Centroids:

	Elevation	Aspect	Slope	Hillshade_9am	Hillshade_Noon	Horizontal_Distance_To_Hydrology	Vertical_Distance_To_Hydrology	Horizontal_Distance_To_Fire_Points	Soil_Type	Wilderness	ClusterID
0	1.491918	1.436402	1.306395	0.042164	0.003514	0.920590	1.468025	1.038651	2.526353	2.621223	0
1	3.000000	0.000000	1.800000	0.000000	0.000000	5.700000	3.000000	2.000000	26.000000	2.000000	1
2	0.510725	1.589717	1.324821	0.029282	0.002383	1.293156	1.241743	1.228124	20.918284	1.081716	2
3	3.000000	1.583741	1.298727	0.028404	0.008815	1.953967	1.628795	1.371205	38.076396	1.048972	3
4	0.689900	1.583761	1.311079	0.054781	0.012472	1.681585	1.533382	1.250917	30.239423	1.153583	4
5	0.000000	1.707135	1.343983	0.034079	0.007455	1.510117	1.413206	1.297125	37.903088	0.883919	5
6	1.170039	1.610308	1.280347	0.080925	0.015414	1.257707	1.468208	1.077553	10.364644	1.807322	6



- Cluster Distribution
  Original Distribution of Classes

```
target
1    4112
2    4711
3    1056
4     175
5     597
6     686
7    1158
dtype: int64
```

After Spectral Clustering

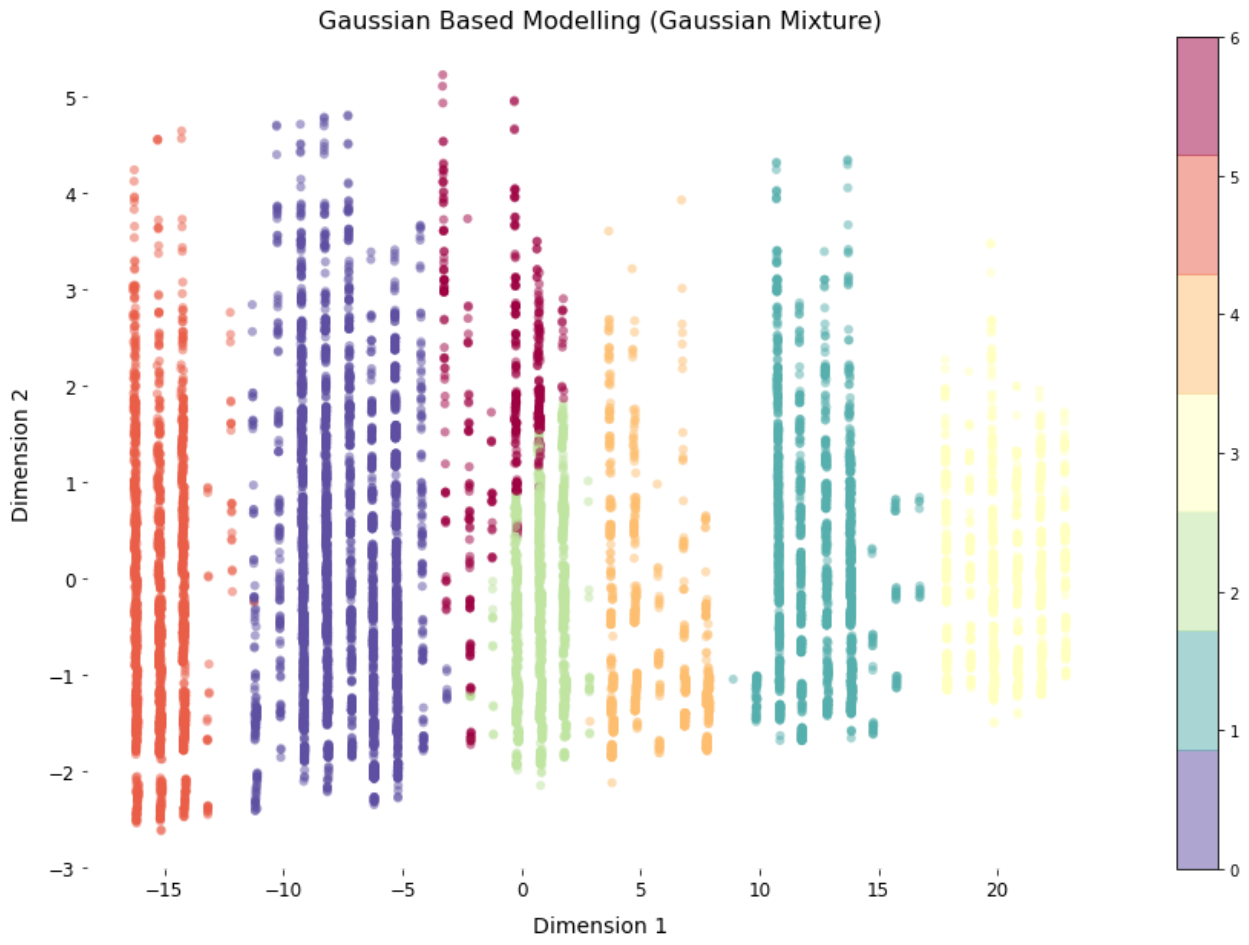
clusters	labels	count
0	1	11
	2	226
	3	665
	4	127
	5	62
	6	327
	7	5
1	2	5
	7	5
2	1	1219
	2	1391
	3	48
	4	13
	5	117
	6	65
	7	84
3	1	456
	2	30
	7	535
4	1	1565
	2	1958
	3	14
	5	262
	6	43
	7	247
5	1	543
	2	117
	7	279
6	1	318
	2	984
	3	329
	4	35
	5	156
	6	251
	7	3

d) Gaussian Mixture

- Centroid

	Elevation	Aspect	Slope	Hillshade_9am	Hillshade_Noon	Horizontal_Distance_To_Hydrology	Vertical_Distance_To_Hydrology	Horizontal_Distance_To_Fire_Points	Soil_Type	Wilderness	ClusterID
0	0.663322	1.580730	1.311281	0.056900	1.252151e-02	1.680474	1.542217	1.246350	30.175197	1.161460	0
1	1.171389	1.610710	1.280721	0.081032	1.543478e-02	1.257947	1.468831	1.077682	10.368797	1.809104	1
2	0.433819	1.623121	1.363100	0.039088	4.559748e-03	0.963181	1.154157	1.282947	21.978306	0.980088	2
3	1.489481	1.436289	1.305894	0.042096	3.507963e-03	0.921173	1.467261	1.038550	2.532027	2.616941	3
4	0.840643	1.547927	1.213722	0.009704	4.798667e-09	0.819587	1.030449	1.087959	17.344790	0.944193	4
5	1.575380	1.638384	1.319281	0.030453	7.987597e-03	1.739030	1.525830	1.335425	37.912893	0.969285	5
6	0.329588	1.565596	1.402810	0.018729	1.113480e-05	3.063011	1.777951	1.335717	23.339173	1.563111	6

- Cluster Visualization



- Cluster Distribution
  Original Distribution of Classes

```
target
1    4112
2    4711
3    1056
4     175
5     597
6     686
7    1158
dtype: int64
```

After Gaussian Based Modelling (Gaussian Mixture)

clusters	labels	count
0	1	1570
	2	1970
	3	14
	5	262
	6	43
	7	243
1	1	318
	2	984
	3	329
	4	35
	5	156
	6	251
	7	3
2	1	705
	2	738
	5	41
	6	19
	7	71
3	1	11
	2	226
	3	665
	4	127
	5	62
	6	327
	7	5
4	1	305
	2	372
	3	48
	4	13
	5	64
	6	45
	7	2
5	1	1003
	2	155
	7	815
6	1	200
	2	266
	5	12
	6	1
	7	19

(iv) Compare the cluster formation of the gaussian based method with the other three clustering methods and report your observations on the results.
Observations:

- The size of clusters is different in Gaussian Mixture than in the other three clustering techniques.
- In spectral clustering some data points are overlapping in more than one cluster while Gaussian clustering defines well defined clusters.
- Gaussian mixture clustering gives a similar train f1 score on normalized and unnormalized input data while spectral clustering f1 score for train decreases considerably from 57% to 41%.
- Unlike other models, the Gaussian Mixture model could predict the probability of how a data point belongs to each of the clusters.
- Unlike other models, the Gaussian model worked faster even on large numbers of samples.

	F1 Score	***K Means Clustering***	***Spectral Clustering***	***Gaussian Mixture Modelling***
On Normalized Input Data	Train	57	41	57
	Test	46	52	46
On Unnormalized Input Data	Train	56	57	57
	Test	36	39.4	38.5

2)

- We split our dataset in the training set and validation set using the train test function provided by model selection module from sklearn.
- Then we tested all the models on train and test set and evaluated there f1 score.

F1 Score	K Means Clustering	Agglomerative Clustering	Spectral Clustering	Gaussian Mixture Modelling
Train	56.86	53.61	41.75	47.91
Test	56.82	32.02	23.80	30.82

- We reached the conclusion that Kmeans gives the highest accuracy compared to the other three.
- Therefore, we selected Kmeans as our best model, and implemented it in our predict function.

Learning:

- Learned how to implement different types of clustering algorithms.
- Learned how to evaluate different clustering algorithms using measures such as f1 score and find the best algorithm for a dataset.
- Got to know that there is no clear accuracy checker in clustering algorithms.
- Learned how to visualise clusters obtained using any clustering technique.
- Learned different techniques to get the centroid of different clustering algorithms.