

# Data Mining Assignment 2

-Dewangee Agrawal (2016034)

Assumptions for the code - The number of datapoints and clusters are input along with the datapoints.

## Question 1 :

- Code submitted - skeleton.py

## Question 2 :

- Code submitted -kmeans.py

## Question 3 :

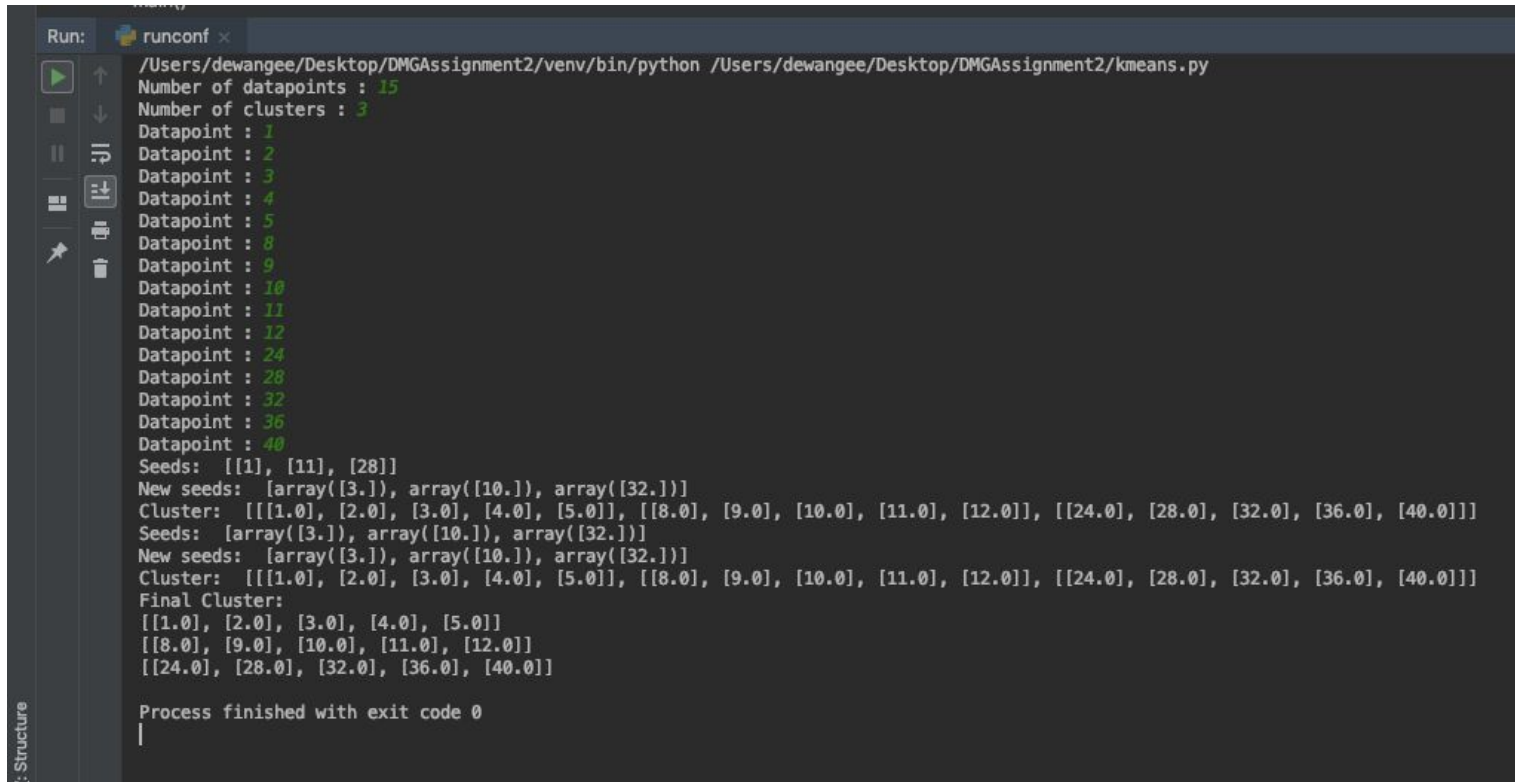
- Code submitted - kmedians.py

## Question 4 :

### Part a :

YES.

The code gives the same clusters as required.



```
Run: runconf x
/Users/dewangee/Desktop/DMGAssignment2/venv/bin/python /Users/dewangee/Desktop/DMGAssignment2/kmeans.py
Number of datapoints : 15
Number of clusters : 3
Datapoint : 1
Datapoint : 2
Datapoint : 3
Datapoint : 4
Datapoint : 5
Datapoint : 8
Datapoint : 9
Datapoint : 10
Datapoint : 11
Datapoint : 12
Datapoint : 24
Datapoint : 28
Datapoint : 32
Datapoint : 36
Datapoint : 40
Seeds: [[1], [11], [28]]
New seeds: [array([3.]), array([10.]), array([32.])]
Cluster: [[[1.0], [2.0], [3.0], [4.0], [5.0]], [[8.0], [9.0], [10.0], [11.0], [12.0]], [[24.0], [28.0], [32.0], [36.0], [40.0]]]
Seeds: [array([3.]), array([10.]), array([32.])]
New seeds: [array([3.]), array([10.]), array([32.])]
Cluster: [[[1.0], [2.0], [3.0], [4.0], [5.0]], [[8.0], [9.0], [10.0], [11.0], [12.0]], [[24.0], [28.0], [32.0], [36.0], [40.0]]]
Final Cluster:
[[1.0], [2.0], [3.0], [4.0], [5.0]]
[[8.0], [9.0], [10.0], [11.0], [12.0]]
[[24.0], [28.0], [32.0], [36.0], [40.0]]

Process finished with exit code 0
```

### Part b :

YES.

The code gives the same clusters as required.

```
Run: runconf x
/Users/dewangee/Desktop/DMGAssignment2/venv/bin/python /Users/dewangee/Desktop/DMGAssignment2/kmeans.py
Number of datapoints : 15
Number of clusters : 3
Datapoint : 1
Datapoint : 2
Datapoint : 3
Datapoint : 4
Datapoint : 5
Datapoint : 8
Datapoint : 9
Datapoint : 10
Datapoint : 11
Datapoint : 12
Datapoint : 24
Datapoint : 28
Datapoint : 32
Datapoint : 36
Datapoint : 40
Seeds: [[1], [2], [3]]
New seeds: [array([1.]), array([2.]), array([17.07692308])]
Cluster: [[1.0], [2.0], [3.0], [4.0], [5.0], [8.0], [9.0], [10.0], [11.0], [12.0], [24.0], [28.0], [32.0], [36.0], [40.0]]
Seeds: [array([1.]), array([2.]), array([17.07692308])]
New seeds: [array([1.]), array([5.16666667]), array([24.125])]
Cluster: [[1.0], [2.0], [3.0], [4.0], [5.0], [8.0], [9.0], [10.0], [11.0], [12.0], [24.0], [28.0], [32.0], [36.0], [40.0]]
Seeds: [array([1.]), array([5.16666667]), array([24.125])]
New seeds: [array([2.]), array([8.42857143]), array([32.])]
Cluster: [[1.0], [2.0], [3.0], [4.0], [5.0], [8.0], [9.0], [10.0], [11.0], [12.0], [24.0], [28.0], [32.0], [36.0], [40.0]]
Seeds: [array([2.]), array([8.42857143]), array([32.])]
New seeds: [array([3.]), array([10.]), array([32.])]
Cluster: [[1.0], [2.0], [3.0], [4.0], [5.0], [8.0], [9.0], [10.0], [11.0], [12.0], [24.0], [28.0], [32.0], [36.0], [40.0]]
Seeds: [array([3.]), array([10.]), array([32.])]
New seeds: [array([3.]), array([10.]), array([32.])]
Cluster: [[1.0], [2.0], [3.0], [4.0], [5.0], [8.0], [9.0], [10.0], [11.0], [12.0], [24.0], [28.0], [32.0], [36.0], [40.0]]
Final Cluster:
[[1.0], [2.0], [3.0], [4.0], [5.0]]
[[8.0], [9.0], [10.0], [11.0], [12.0]]
[[24.0], [28.0], [32.0], [36.0], [40.0]]
Process finished with exit code 0
```

### Part c :

In part a and b, we arrive at the same clusters but the number of iterations is more in the second case because of different initial seeds. Hence, to find the right clusters in less time and for the algorithm to converge faster, we should use appropriate initial seeds. We should choose the seeds to not be very close to each other and differ according the values of data points. For eg, in part b, the seeds are [1,2,3] whereas the values of datapoints goes over to 40. In part a, the seeds [1,11,28] are chosen according to similarity in data point values and convergence occurs faster.