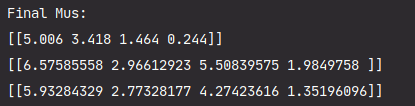
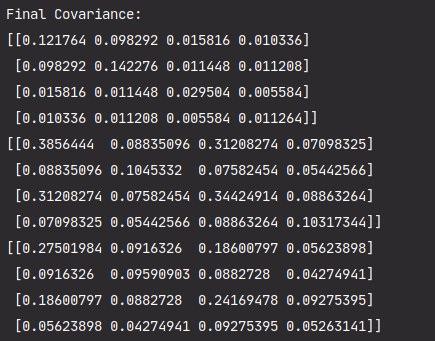
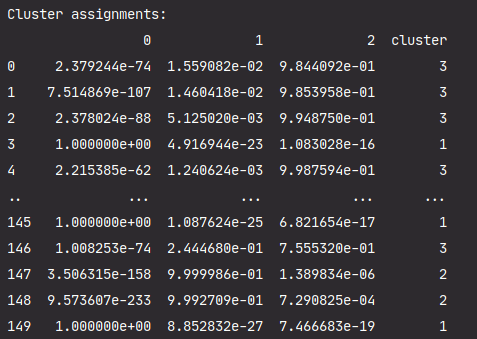
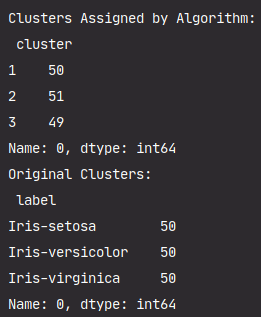
# 4.1 Exercieses: Clustering

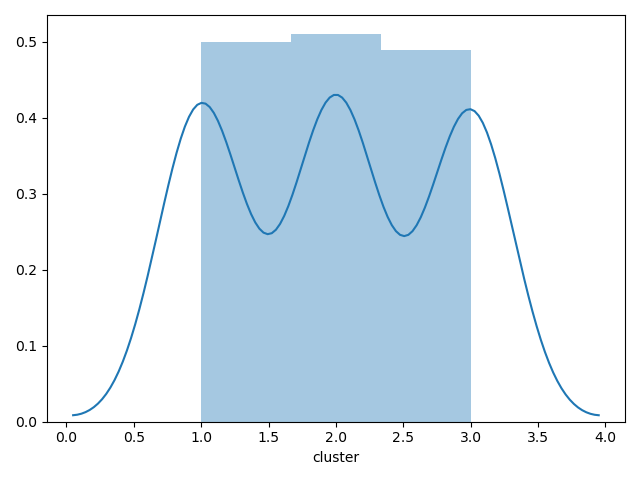
## 1. Expectation Maximization Clustering

1. 1.The final mean for each cluster  
   
2. The final covariance matrix for each cluster  
   
3. Number of iterations the EM algorithm took to converge.  
   
4. Final cluster assignment of all the points, where each point will be assigned to the cluster that yields the highest probability *P(Ci|xj)*
5. Final size of each cluster

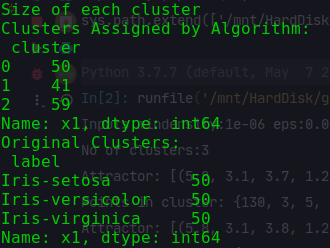
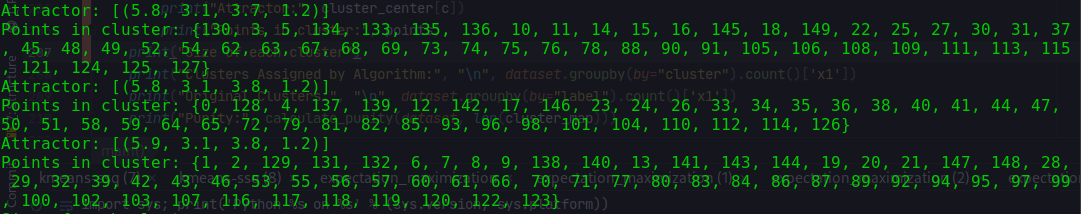


Purity score



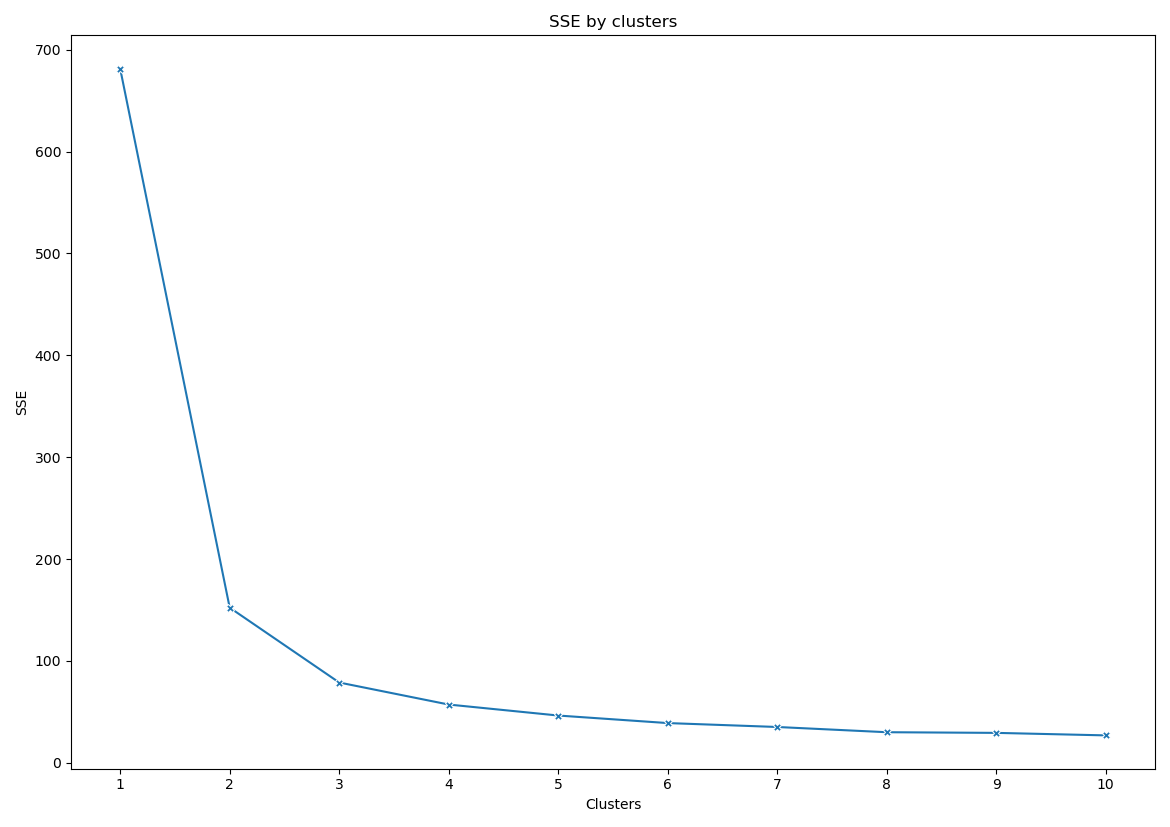


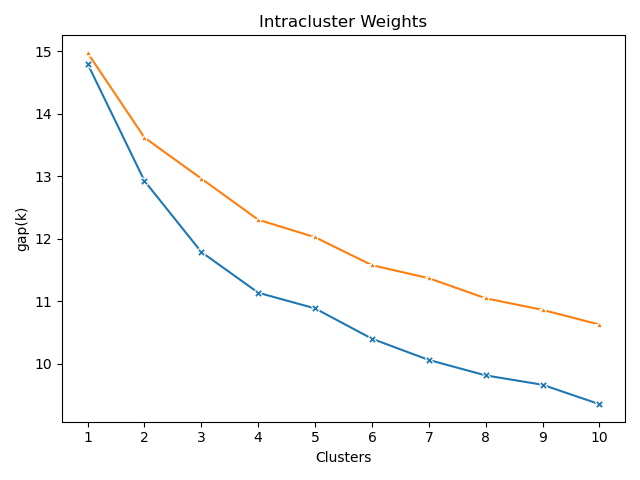
## 2. Density Based Clustering: DENCLUE

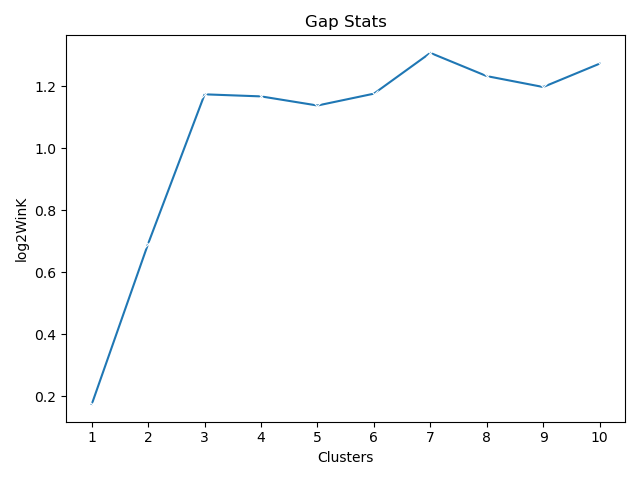
1. The number of clusters, and the size of each cluster.  
     
     
   
2. The density attractor, followed by the set of point in that cluster.  
   
3. Purity of the clustering, based on the true id.



## 3. Clustering

1. The SSQs computed fork values between 1 and 10 (inclusive). There should be one plot corresponding to the SSQs.  
     
   
2. The gap statistics computed fork values between 1 and 10 (inclusive). There should be two plots corresponding to the gap statistics.





1. Where did you estimate the elbow point to be (between what values of k)? What value of k was typically estimated as optimal by the gap statistic? To adequately answer this question, consider generating both measures several times, as there may be some amount of variation in the value of k that they each estimate as optimal.  
   Ans: I would go with the 3 clusters, as elbow plot shows the change is small from 4 clusters onwards. Gap statastics is showing optimal between 3 and 4.
2. How close are the estimates generated by the elbow point and gap statistic to the number of species of Iris represented in the dataset?  
   Ans: Both stats are very close to the number species i.e. 3.
3. Assuming we are trying to generate one cluster for each Iris species represented in the dataset, does one measure seem to be a consistently better criterion for choosing the value of k than the other? Why or why not?  
   Ans: As per my understanding elbow plot is more consistent with this dataset. as elbow is showing clearly change in slope after 3 clusters. Gap statastics keep fluctuating between 3 and 4.

## 4. Complete the following in your text:

MiningMassive Datasets Page 252 --- Exercise 7.2.2

Exercise 7.2.2 : How would the clustering of Example 7.2 change if we used for the distance between two clusters:

(a) The minimum of the distances between any two points, one from each cluster.

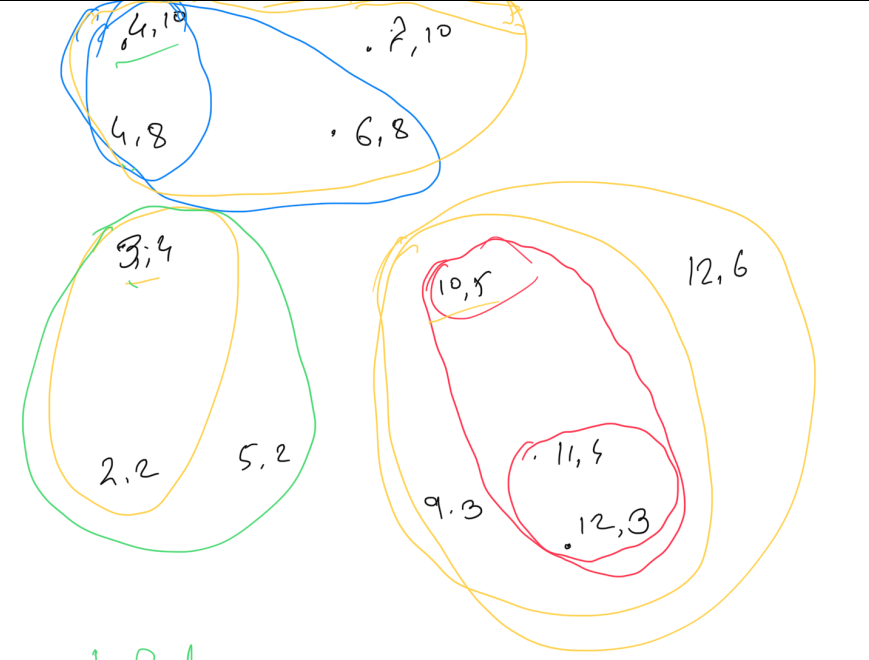
Cluster growth

1- Red

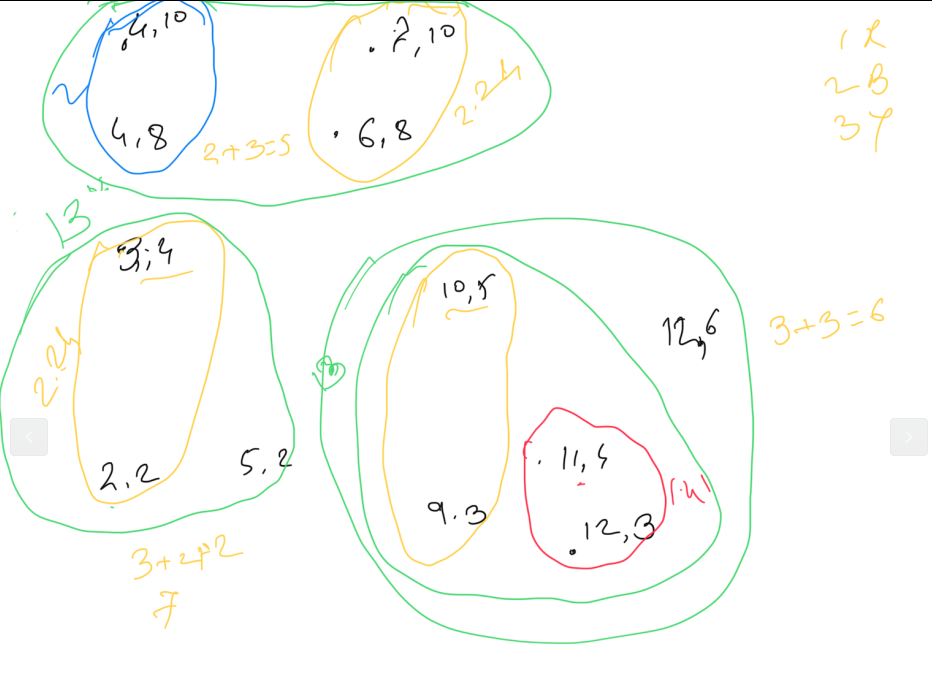
2 - Blue

3 - Yellow

4 - Green



(b) The average of the distances between pairs of points, one from each of the two clusters.



Mining Massive Datasets Page 260 --- Exercise 7.3.4

For the three clusters of Fig. 7.8:

(a) Compute the representation of the cluster as in the BFR Algorithm. That is, compute N , SUM , and SUMSQ .

Cluster 1 [[4,10], [7,10], [4,8], [6,8]]

N = 4

SUM = [21, 36]

SUMSQ = [117, 328]

Cluster 2 [[3,4], [2,2], [5,2]]

N = 3

SUM = [10, 8]

SUMSQ = [38, 24]

Cluster 3 [[10,5], [12,6], [11,4], [9,3], [12,3] ]

N = 5

SUM = [54, 21]

SUMSQ = [590, 95]

(b) Compute the variance and standard deviation of each cluster in each of

the two dimensions.

Variance = SUMSQi/N - (SUMi/N)2

Std. Dev = SQRT(Variance)

Cluster 1 [[4,10], [7,10], [4,8], [6,8]]

var1 = 117/4 - (21/4)2 = 29.25 - (5.25)2 = 29.25 - 27.56 = 1.68

Std Dev1 = 1.29

Var2 = 328/4 - (36/4)2 = 82 - 81 = 1

Std Dev2 = 1

Cluster 2 [[3,4], [2,2], [5,2]]

Var1 = 38/3 - (10/3)2  = 12.55 - 11.11 = 1.43

Std Dev1 = 1.19

Var2 = 24/3 - (8/3)2 = 8 - 7.11 = 0.88

Std Dev2 = 0.94

Cluster 3 [[10,5], [12,6], [11,4], [9,3], [12,3] ]

Var1 = 590/5 - (54/5)2 = 1.36

Std1 = 1.16

Var2 = 95/5 - (21/5)2 = 1.36

StdDev2= 1.16

Mining Massive Datasets Page 260 --- Exercise 7.3.5

Exercise 7.3.5 : Suppose a cluster of three-dimensional points has standard deviations of 2, 3, and 5, in the three dimensions, in that order. Compute the Mahalanobis distance between the origin (0, 0, 0) and the point (1, −3, 4).

std devs 2,3,5

Point 1 = (0,0,0)

Point 2 = (1, -3, 4)

Mahalanobis distance = SQRT((1/2)2+(-3/3)2+(4/5)2) = 1.37