Homework 2

1.

a) What is the k generated by the algorithm?

The k is generated by Weka's EM implementation, when the parameter numClusters = -1 is 4.

Cluster output:

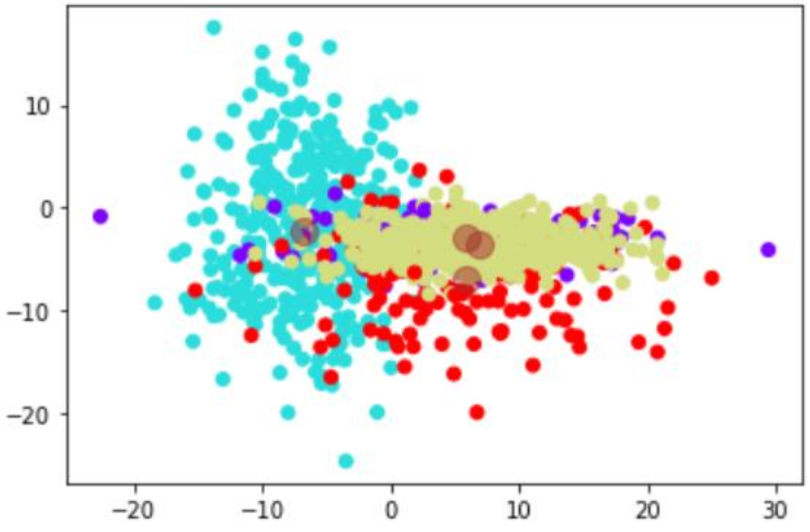
```
=== Run information ===

Scheme:       weka.clusterers.EM -I 100 -N -1 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter 1.0E-6 -M 1.0E-6 -K
10 -num-slots 1 -S 100
Relation:     Clustering
Instances:    977
Attributes:   10
              ï»¿cluster1
              cluster2
              cluster3
              cluster4
              cluster5
              cluster6
              cluster7
              cluster8
              cluster9
              cluster10
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


EM
==

Number of clusters selected by cross validation: 4
Number of iterations performed: 3


             Cluster
Attribute         0      1      2      3
             (0.11) (0.37) (0.13) (0.39)
========================================
ï»¿cluster1
  mean        5.8857-6.7841 5.8965 6.8552
  std. dev.   7.9289 4.0281 7.4119 6.0446

cluster2
  mean        -3.054-2.2253-7.1166-3.5652
  std. dev.   1.6282 7.2249 4.5713 1.8383

cluster3
  mean        3.5811 7.5501-6.3953 3.5377
  std. dev.   4.7259 3.1528 5.4642 7.1487

cluster4
  mean        2.2937 2.9934 9.2059-1.6068
  std. dev.   2.2476 5.2465 2.2937  5.027

cluster5
  mean       -4.3358 8.3075-2.8772-2.1415
  std. dev.   2.6792 7.7108 8.7461 1.0928

cluster6
  mean        8.1935 6.5091-6.5188-7.7354
  std. dev.   2.9051 4.0865 3.4286 6.0973

cluster7
  mean        8.5604 -4.898-3.2343 9.0502
  std. dev.   2.2228 1.2075 2.6843 1.8343

cluster8
  mean       -1.4552-7.0473 6.1973 -5.437
  std. dev.   7.2029 7.9383 3.2952 1.4725

cluster9
  mean         5.964-1.2951-4.5518-0.2541
  std. dev.   5.7789 2.3108 3.3909 3.3342

cluster10
  mean        9.6579-6.2533 3.7672-4.7881
  std. dev.   2.3239 4.7581 1.8076 2.4935



Time taken to build model (full training data) : 3.06 seconds
```

```
=== Model and evaluation on training set ===

Clustered Instances

0      105 ( 11%)
1      359 ( 37%)
2      128 ( 13%)
3      385 ( 39%)


Log likelihood: -28.10108
```

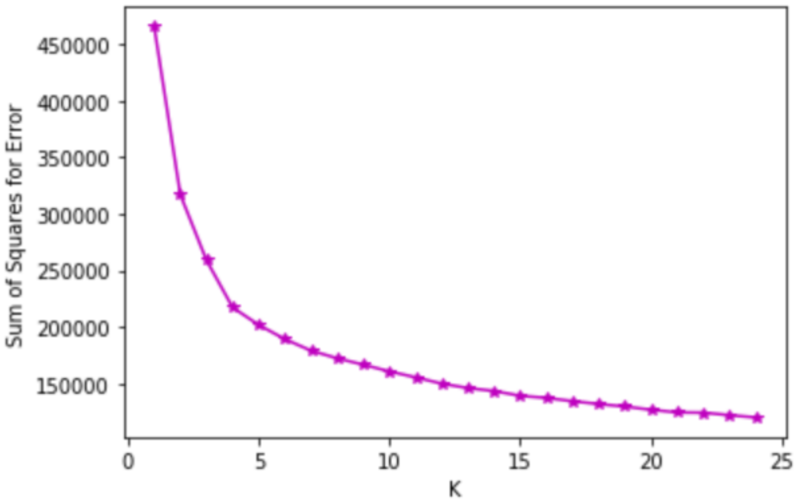
b) If you clustered the data using k-means, does the EM k value match the one you chose? Why or why not?

Yes, the EM k value matches with the data clustered using k-means. K value for the k-means and EM k value is 4.
The value of k for k-means and EM implementation is same as the only difference is the high accuracy with the results and high speed. The k-means is lower accuracy and lower speed in comparison with the EM implementation.

c) Include the output from your k-means implementation.



In the above graph, scatter plot for k = 4 for k-means cluster which is here shown by marker.



In the above plot, the elbow of the graph is at 4 and that is shown by the number of k in k means.