Machine Learning

K-Means

Assignment 1 Exercise 1

Nama: Laela Citra Asih

NIM: 1301144300

Kelas : IF-38-Gab02

[Pick the date]

1. Deskripsi Kasus

K-Means adalah salah satu metode *clustering* . K-Means termasuk kedalam algoritma *unsupervised learning* dimana algoritma ini menerima inputan data tanpa data label. Proses *clustering* data dengan K-Means dilakukan dengan cara (1) memilih K buah *centroid* , (2) mengelompokan data sehingga terbentuk K *cluster* yang telah ditentukan, (3) mengupdate *centroid* , (4) mengulangi langkah 2 dan 3 hingga *centroid* tidak berubah.

Pada tugas 4 ini, akan dilakukan implementasi metode K-Means untuk data set Aggregation.csv

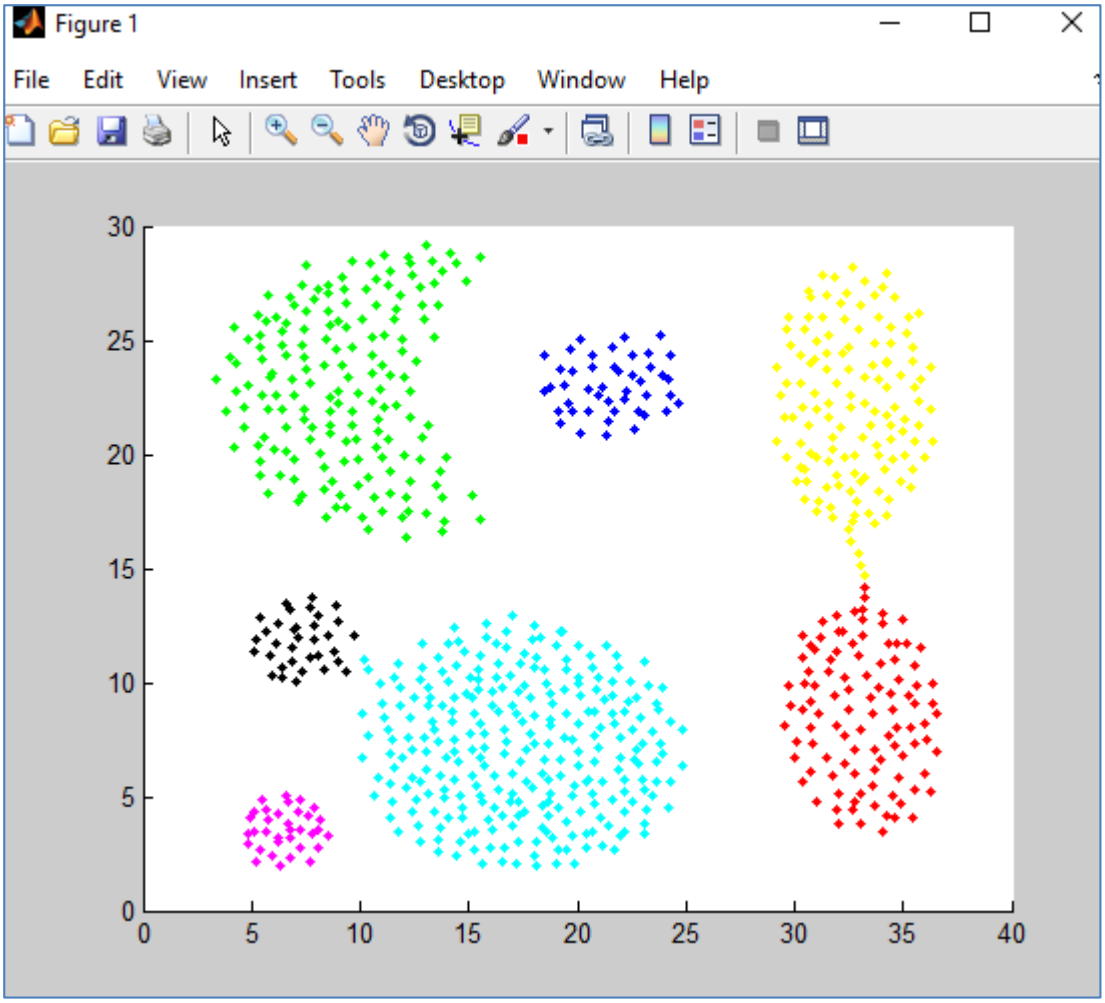
2. Implementasi K-Means

a. Load Dataset

Untuk menampilkan data menggunakan scatter plot dengan warna yang berbeda untuk setiap kelas data, maka ditulislah baris program sebagai berikut.

```
for i=1: length(a(:,1));
    if a(i,3)==1
        scatter(a(i,1),a(i,2),'.b');hold on;
    elseif a(i,3)==2
        scatter(a(i,1),a(i,2),'.g');hold on;
    elseif a(i,3)==3
        scatter(a(i,1),a(i,2),'.r');hold on;
    elseif a(i,3)==4
        scatter(a(i,1),a(i,2),'.c');hold on;
    elseif a(i,3)==5
        scatter(a(i,1),a(i,2),'.m');hold on;
    elseif a(i,3)==6
        scatter(a(i,1),a(i,2),'.y');hold on;
    elseif a(i,3)==7
        scatter(a(i,1),a(i,2),'.k');hold on;
    end
end
```

Setiap data akan ditampilkan pada scatter plot dengan memanfaatkan atribut pertama dan atribut kedua data, setiap kelas data akan dibedakan dengan warna yang berbeda. Setelah dijalankan, dengan menggunakan data set Aggregation.csv, data tersebut direpresentasikan sebagai berikut.



b. K-Means

i. Fungsi K-Means

Berikut baris program untuk melakukan *clustering* menggunakan metode K-Means.

```
function [ finalCentroids result ] = KMeans( datas ,
centroids)

c = centroids;
[m n] = size(datas);
d = zeros(length(datas(:,1)),max(datas(:,3))+1);
a = ones(max(datas(:,3)),n-1);
c = centroids;

while (a~= c)
    a = c;
    for j=1:length(datas(:,1))
        for i=1:max(datas(:,3))
            d(j,i) = norm(datas(j,1:2)-
c(i,:))*norm(datas(j,1:2)-c(i,:));
        end
        [num] = min(d(j,1:7));
        [x y] = ind2sub(size(d(j,1:7)),find(d(j,1:7)==num));
```

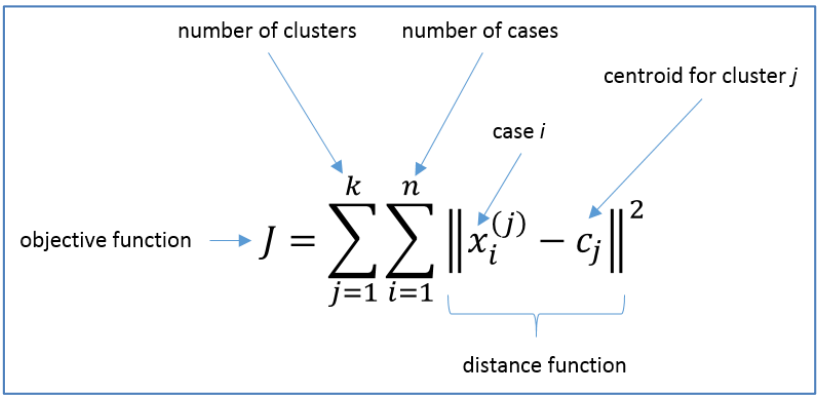
```
        d(j,max(datas(:,3))+1)=y;
    end
    for i=1:max(datas(:,3))
        selectedrows = find(d(:,max(datas(:,3))+1)== i);
        nA = datas(selectedrows,:);
        c(i,:)= 1/length(nA)*sum(nA(:,1:2));
    end
end

finalCentroids = c;
r = datas;
r(:,3) = d(:,max(datas(:,3))+1);
result = r;
end
```

Inputan pada fungsi KMeans tersebut ialah matriks data dan *centroid* awal. Pertama-tama dilakukan inisialisasi matriks-matriks yang diperlukan untuk proses *clustering.* Kemudian untuk *clustering* dilakukan proses perhitungan jarak (*dissimilarity*) setiap data dengan setiap *centroid.* Dari setiap perhitungan tersebut disimpan kedalam suatu matriks (matriks d) kemudian dicari nilai minimal dari jarak data kepada setiap *centroid*. Nilai minimal data dengan satu centroid tersebutlah menentukan *cluster* untuk data tersebut. Setelah seluruh data tercluster, nilai centroid diupdate dengan rata-rata data pada setiap *cluster* tersebut. Kemudian, kembali dilakukan perhitungan jarak setiap data dengan setiap *centroid* dan pemberian *cluster* data hingga nilai centroid tidak berubah lagi.

ii. Fungsi Sum Square Error (SSE)

Sum Square Error (SSE) merupakan fungsi objektif yang digunakan untuk mengukur performansi KMeans. Secara matematis SSE dirumuskan sebagai berikut.



Sesuai rumus tersebut, berikut function SSE yang dapat digunakan untuk menghitung performansi KMeans.

```
function  nSSE  = SSE( datas , finalcentroids)
```

```
[m n] = size(datas);
s=0;

for i=1:max(datas(:,n))
        selectedrows = find(datas(:,n)== i);
        nA = datas(selectedrows,:);
        for j=1:length(nA)
            s = norm(nA(j,1:2)-
finalcentroids(i,:))*norm(nA(j,1:2)-finalcentroids(i,:))+s;
        end
end

nSSE = s;
end
```

Untuk menghitung SSE diperlukan dataset dan centroid akhir dari proses clustering data. Untuk setiap cluster akan dilakukan perhitungan jarak terhadap centroidnya.

c. K-Means dengan Random Selected Data

Untuk melakukan clustering data menggunakan random data sebagai centroid awal maka digunakan baris program berikut untuk mendapatkan data yang akan digunakan sebagai centroid awal.

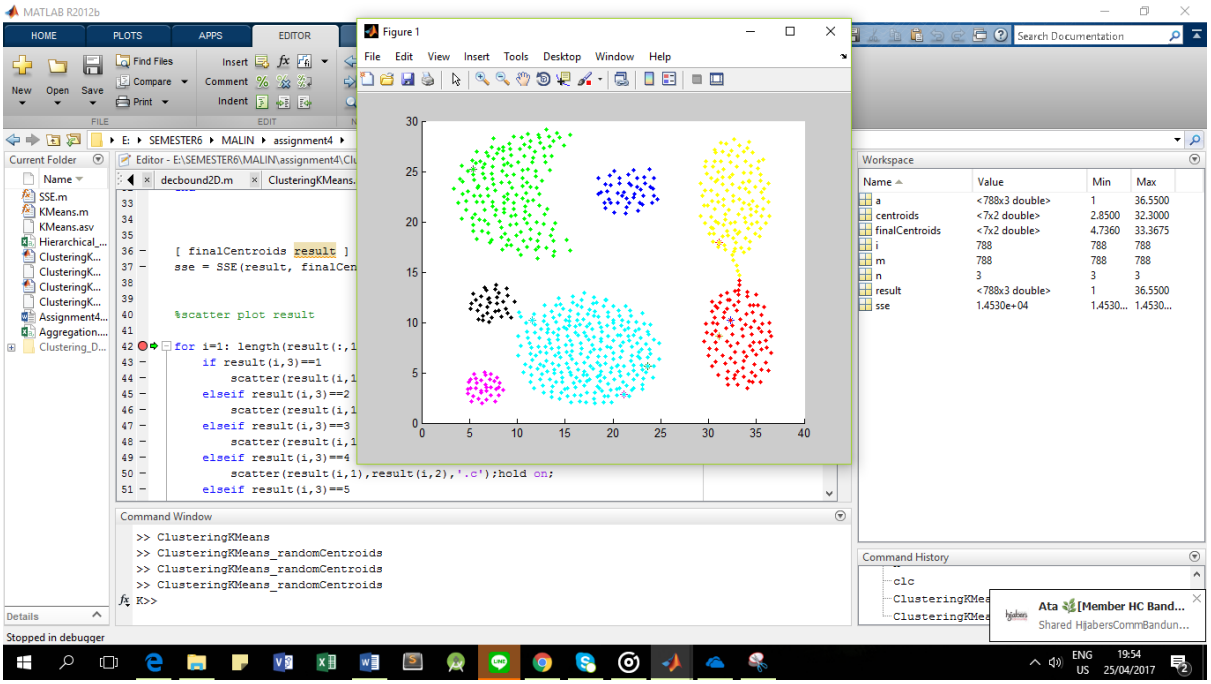
```
%centroid awal
for i=1:max(a(:,3))
    centroids(i,:) = a(randi(m),1:2);
end
```

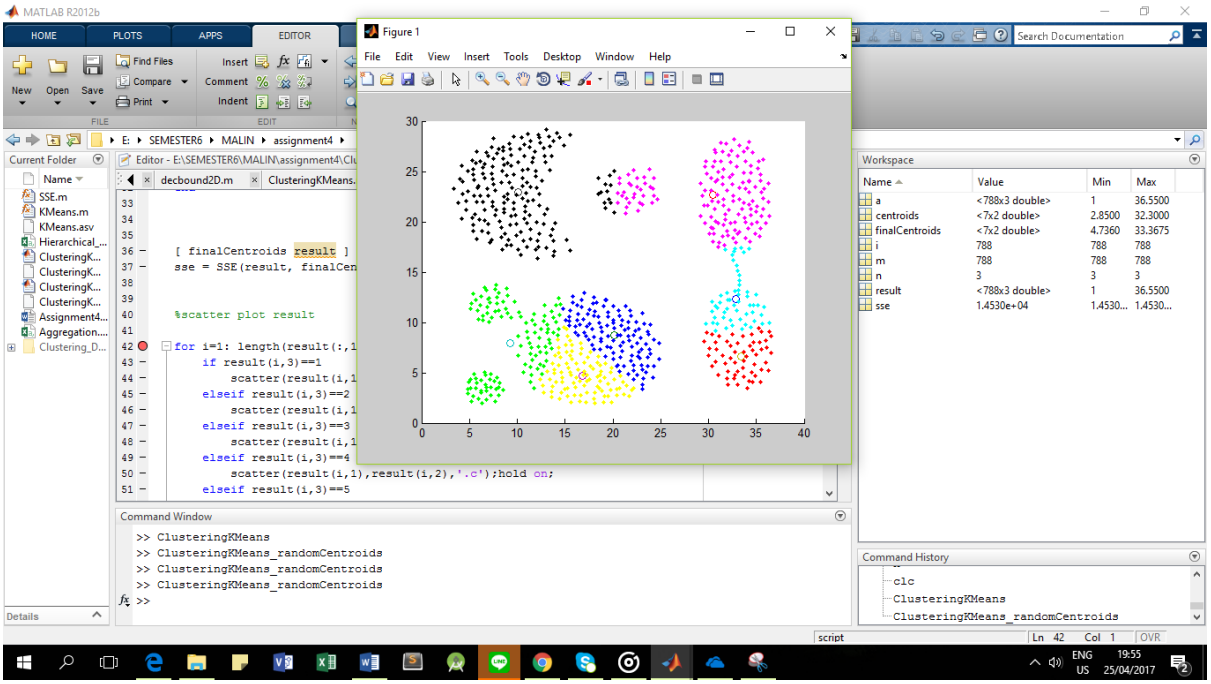
Dengan menambahkan baris program berikut, maka data centroid tersebut ditampilkan dengan symbol * .

```
if (i<=7)
        scatter(centroids(i,1),centroids(i,2),'*');hold on;
    end
```

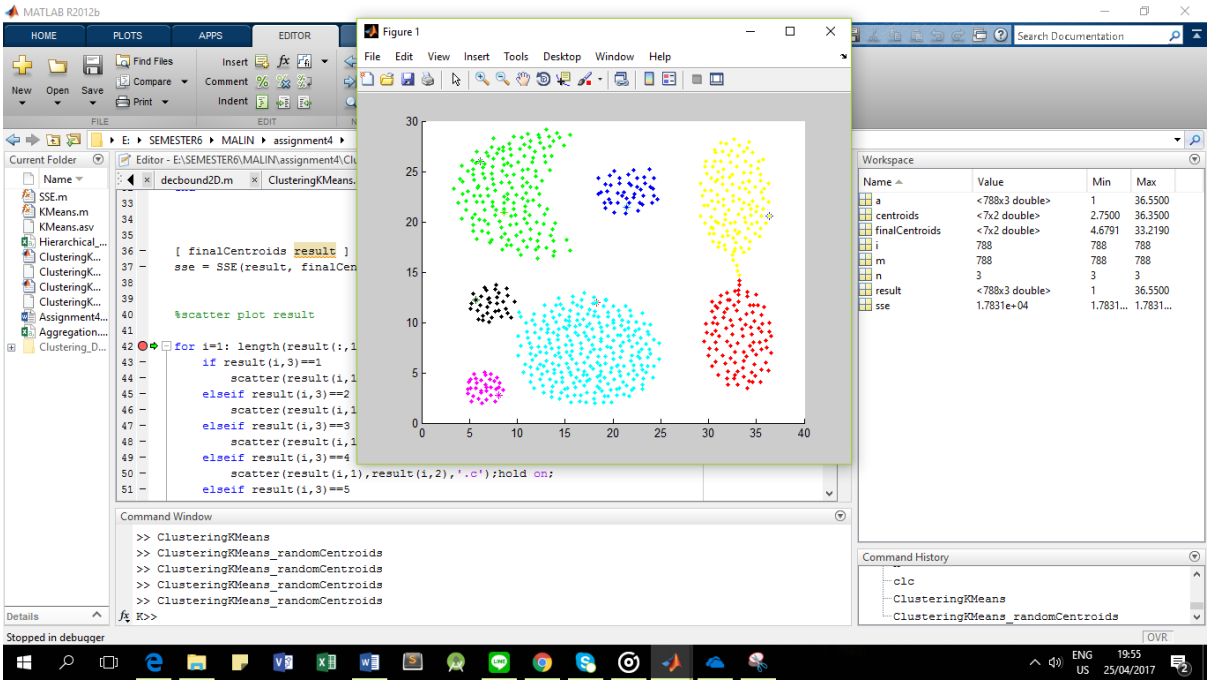
Berikut 5 kali hasil running dengan random centroid

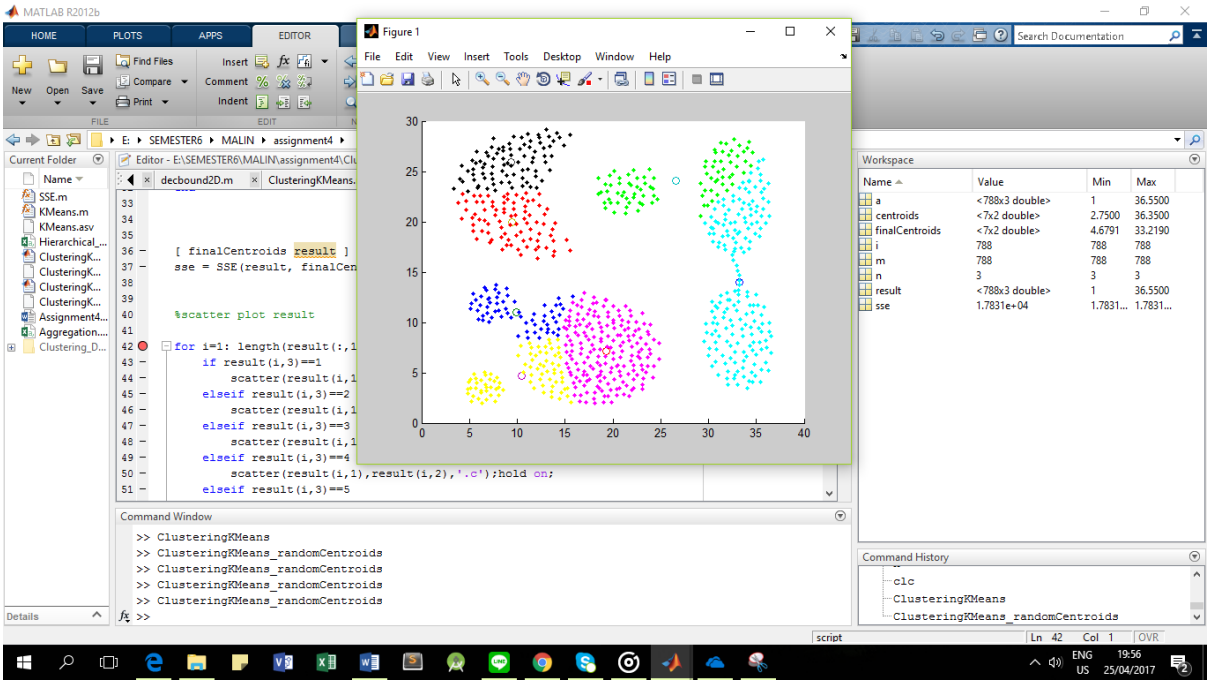
1.


MATLAB R2012b
Figure 1
Command Window
>> ClusteringKMeans
>> ClusteringKMeans_randomCentroids
>> ClusteringKMeans_randomCentroids
>> ClusteringKMeans_randomCentroids
Workspace
Command History
Stopped in debugger

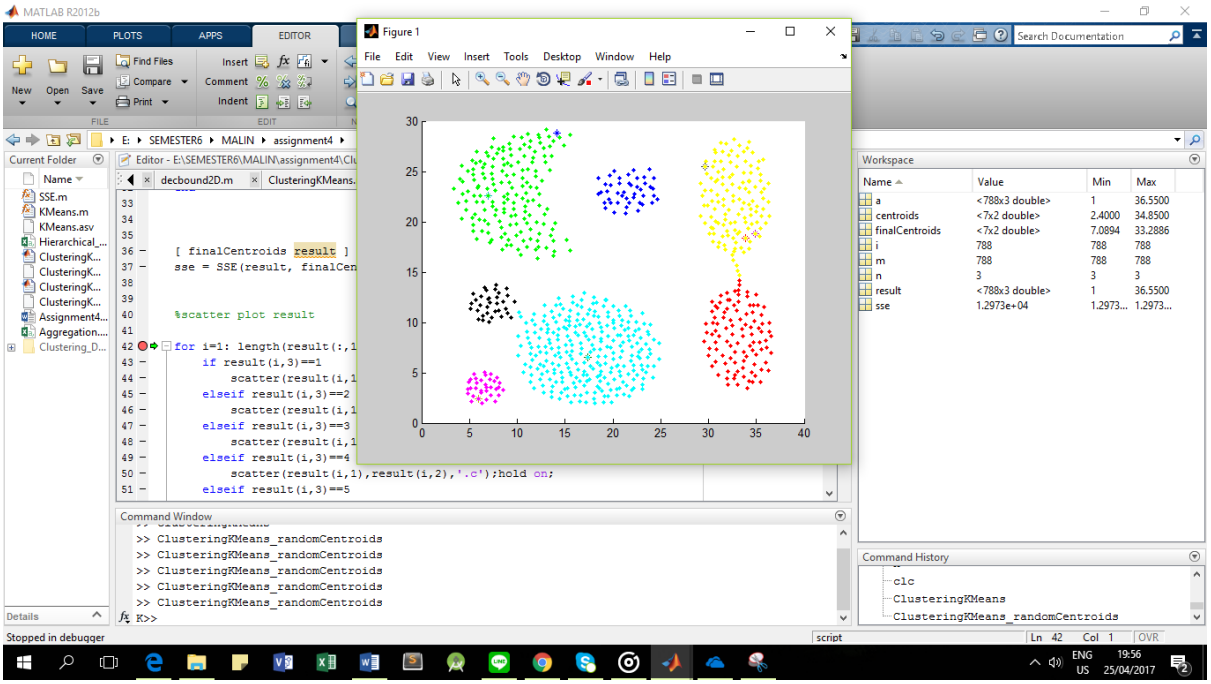

MATLAB R2012b
Figure 1
Command Window
>> ClusteringKMeans
>> ClusteringKMeans_randomCentroids
>> ClusteringKMeans_randomCentroids
>> ClusteringKMeans_randomCentroids
Workspace
Command History

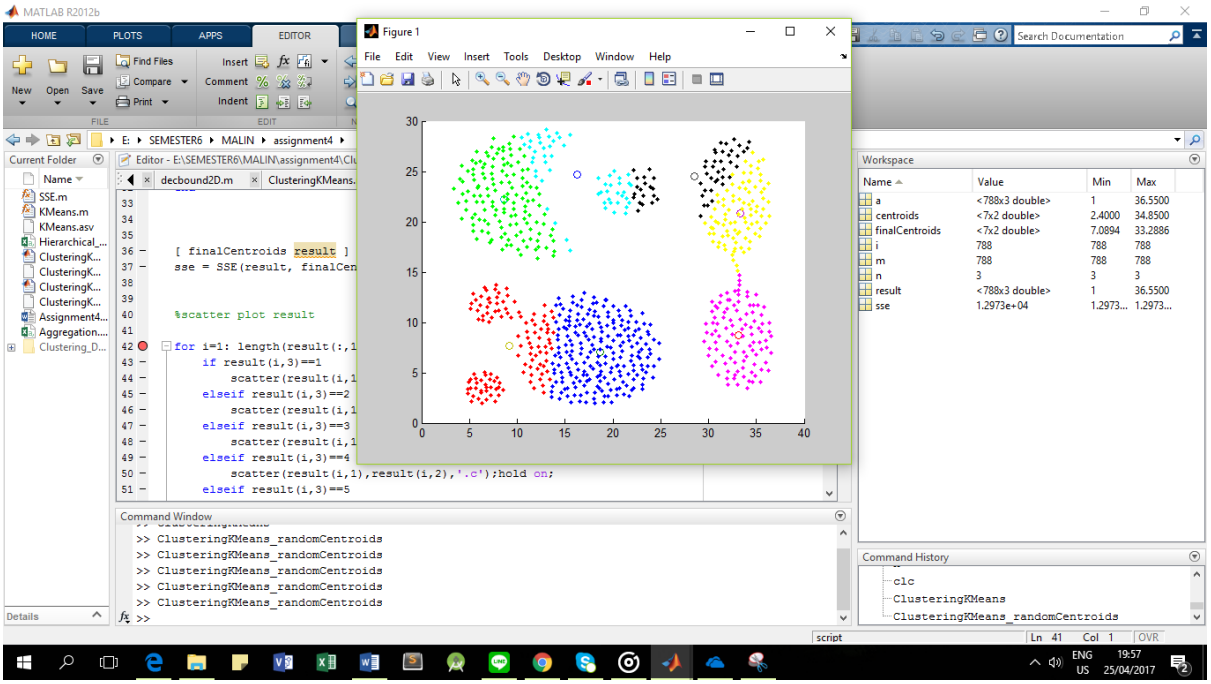
2.


MATLAB R2012b
Figure 1
File Edit View Insert Tools Desktop Window Help
Command Window
>> ClusteringKMeans
>> ClusteringKMeans_randomCentroids
>> ClusteringKMeans_randomCentroids
>> ClusteringKMeans_randomCentroids
>> ClusteringKMeans_randomCentroids
Workspace
Command History
clc
ClusteringKMeans
ClusteringKMeans_randomCentroids
Stopped in debugger

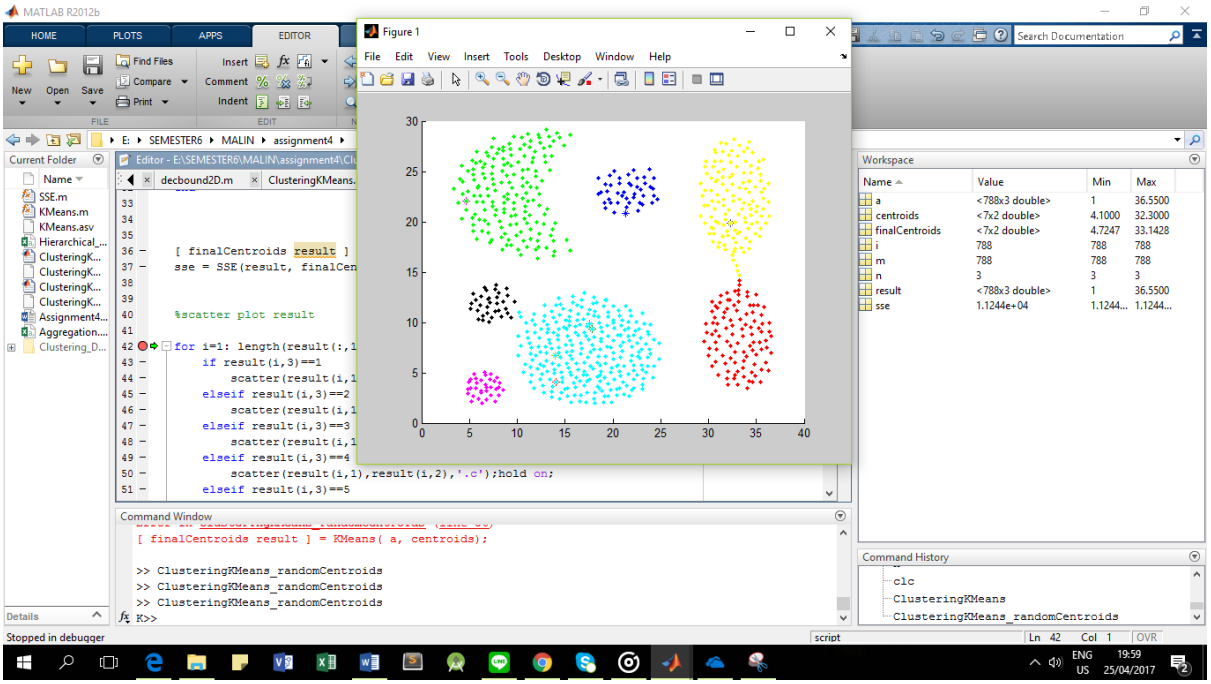

MATLAB R2012b
Figure 1
File Edit View Insert Tools Desktop Window Help
Command Window
>> ClusteringKMeans
>> ClusteringKMeans_randomCentroids
>> ClusteringKMeans_randomCentroids
>> ClusteringKMeans_randomCentroids
>> ClusteringKMeans_randomCentroids
Workspace
Command History
clc
ClusteringKMeans
ClusteringKMeans_randomCentroids
script

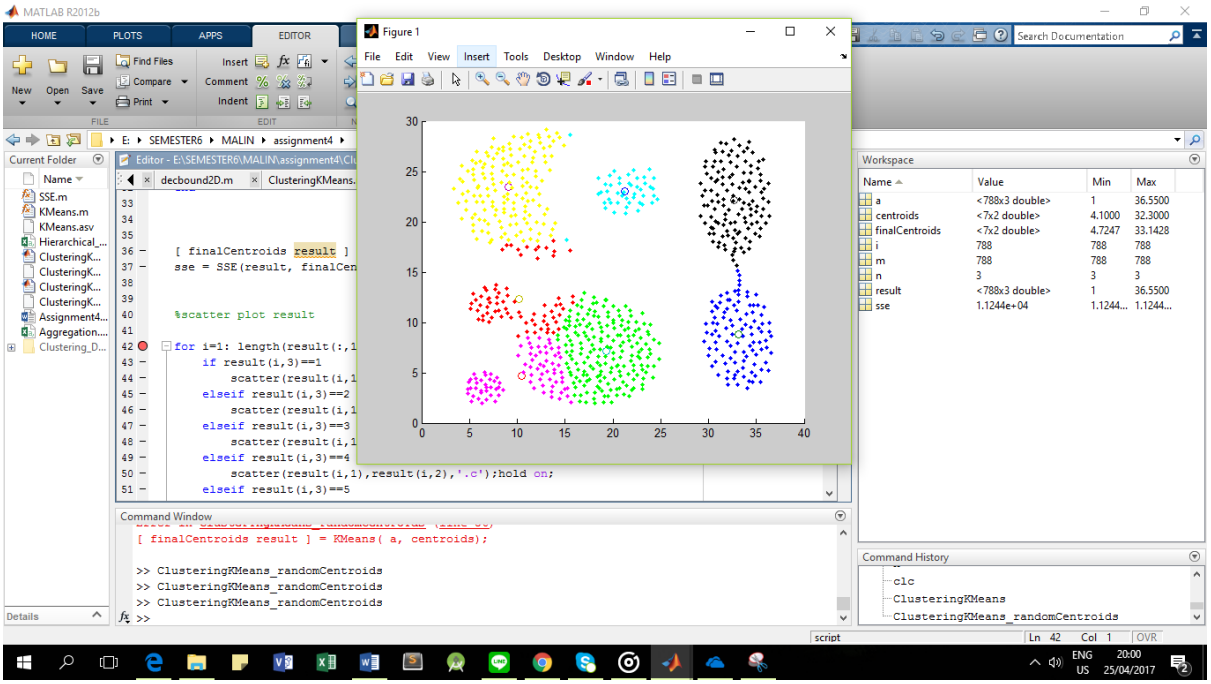
3.


MATLAB R2012b
Figure 1
File Edit View Insert Tools Desktop Window Help
Workspace
Command Window
>> ClusteringKMeans_randomCentroids
Command History
clc
ClusteringKMeans
ClusteringKMeans_randomCentroids

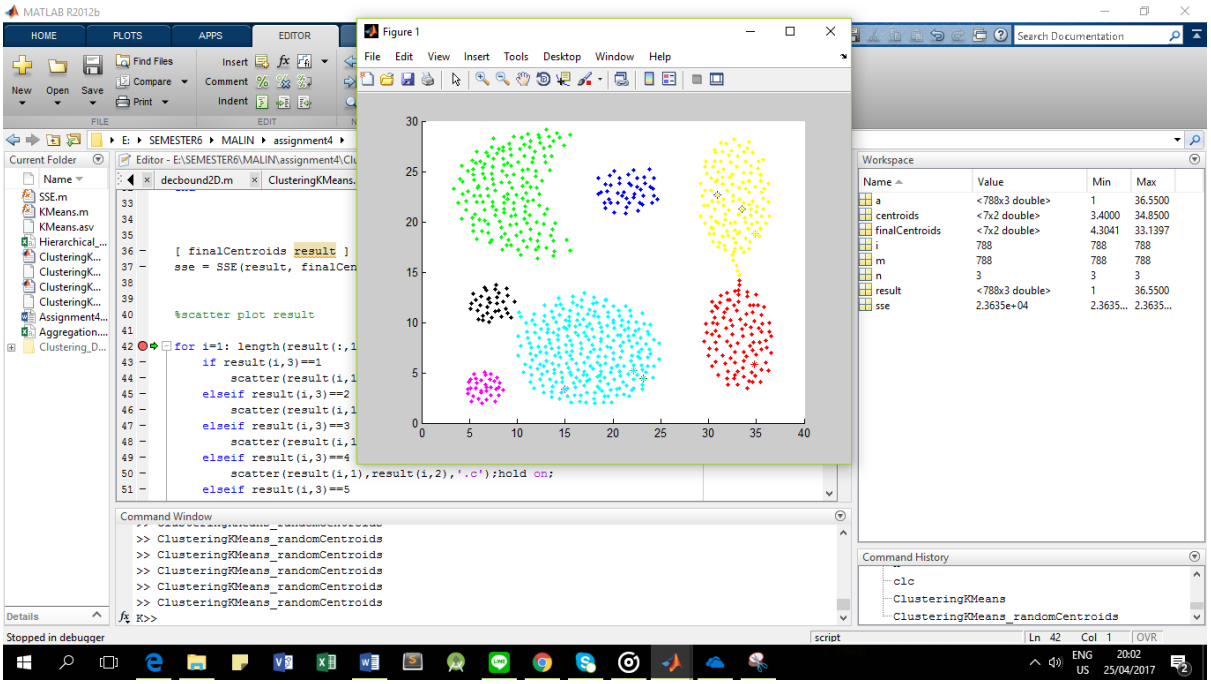

MATLAB R2012b
Figure 1
File Edit View Insert Tools Desktop Window Help
Workspace
Command Window
>> ClusteringKMeans_randomCentroids
Command History
clc
ClusteringKMeans
ClusteringKMeans_randomCentroids

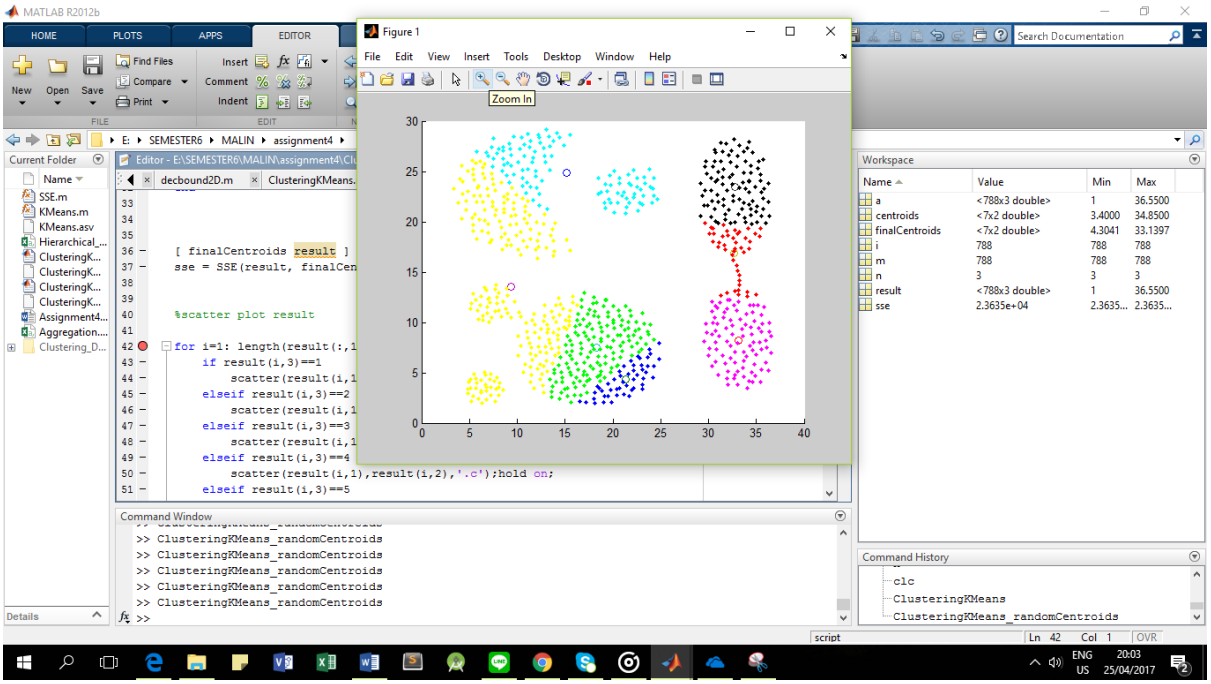
4.


MATLAB R2012b
Figure 1
File Edit View Insert Tools Desktop Window Help
Command Window
>> ClusteringKMeans_randomCentroids
>> ClusteringKMeans_randomCentroids
>> ClusteringKMeans_randomCentroids


MATLAB R2012b
Figure 1
File Edit View Insert Tools Desktop Window Help
Command Window
>> ClusteringKMeans_randomCentroids
>> ClusteringKMeans_randomCentroids
>> ClusteringKMeans_randomCentroids

5.



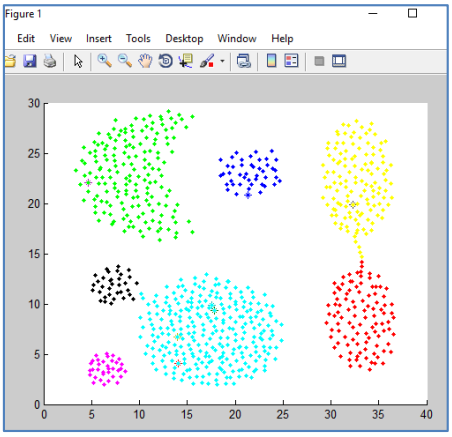


Dengan merunning program sebanyak lima kali didapatkan hasil terbaik pada saat running keempat dengan SSE yang didapatkan yakni 1.1244e+04

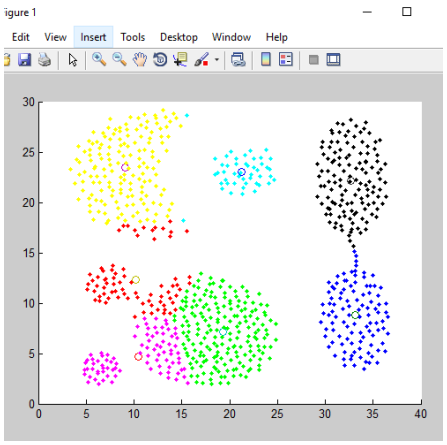
d. Clustering Result (K-Means with Random Selected Data)

Salah satu hasil clustering yang didapatkan dengan menggunakan random data ialah sebagai berikut.

- Visual data set berdasarkan label kelas dan centroid awal (dengan symbol ‘x’).



- Hasil clustering data dan centroid akhir dengan symbol ‘o’



e. K-Cluster dan K-Classes

Dengan melakukan random data untuk digunakan sebagai centroid awal maka kita tidak dapat menentukan suatu data yang dipilih akan dijadikan kluster berapa, yang jelas akan terpilih data acak sebanyak jumlah kelas, akhirnya data akan terkelompokkan berdasarkan centroid terdekatnya, namun dengan pewarnaan yang tidak sesuai dengan yang diinginkan. Seperti scatter plot pada jawaban poin d, warna hitam pada data asal merupakan pewarnaan untuk data dengan berlabel 1 namun dikarenakan random nilai, pada saat inisialisasi random centroid cluster 1, data yang di jadikan centroid 1 merupakan data kelompok 6. Sehingga pada hasil clustering dikarenakan centroid 1 awal merupakan data pada data label 6, maka hasil clustering dihasilkan data dengan warna hitam.

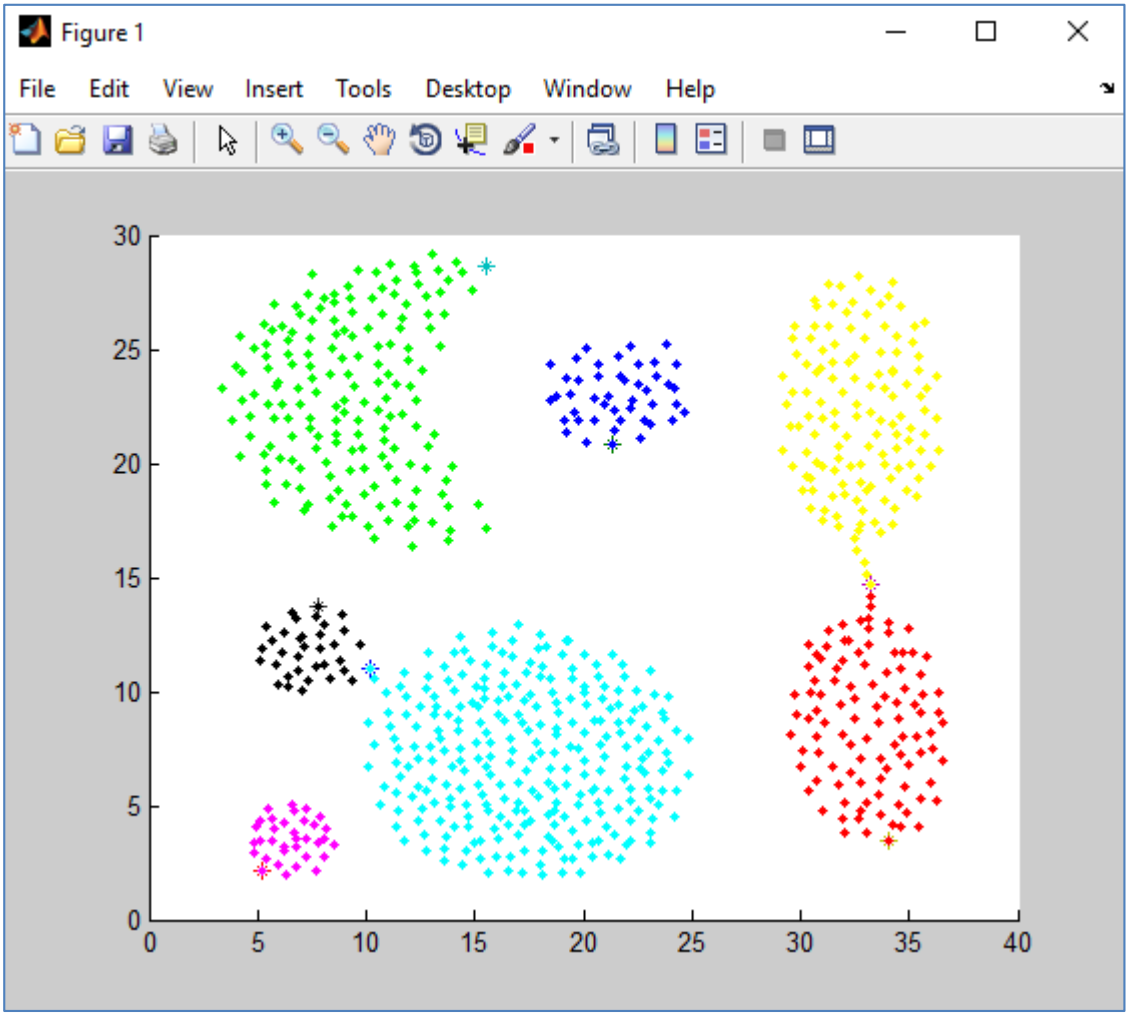
Clustering akan sesuai dengan kelas apabila centroid yang diambil merupakan mepresentasi dari data pada kelas tersebut. Namun apabila centroid yang digunakan merupakan centroid yang berada pada kelas yang sama ataupun terlalu dekat dengan kelas yang berbeda maka hal tersebut dapat menyebabkan ketidaksesuaian cluster data dengan kelas data (terlihat pada running untuk kelima kalinya pada data random, jawaban poin c).

f. K-Means dengan initial centroid yang merupakan satu data dari setiap kelas

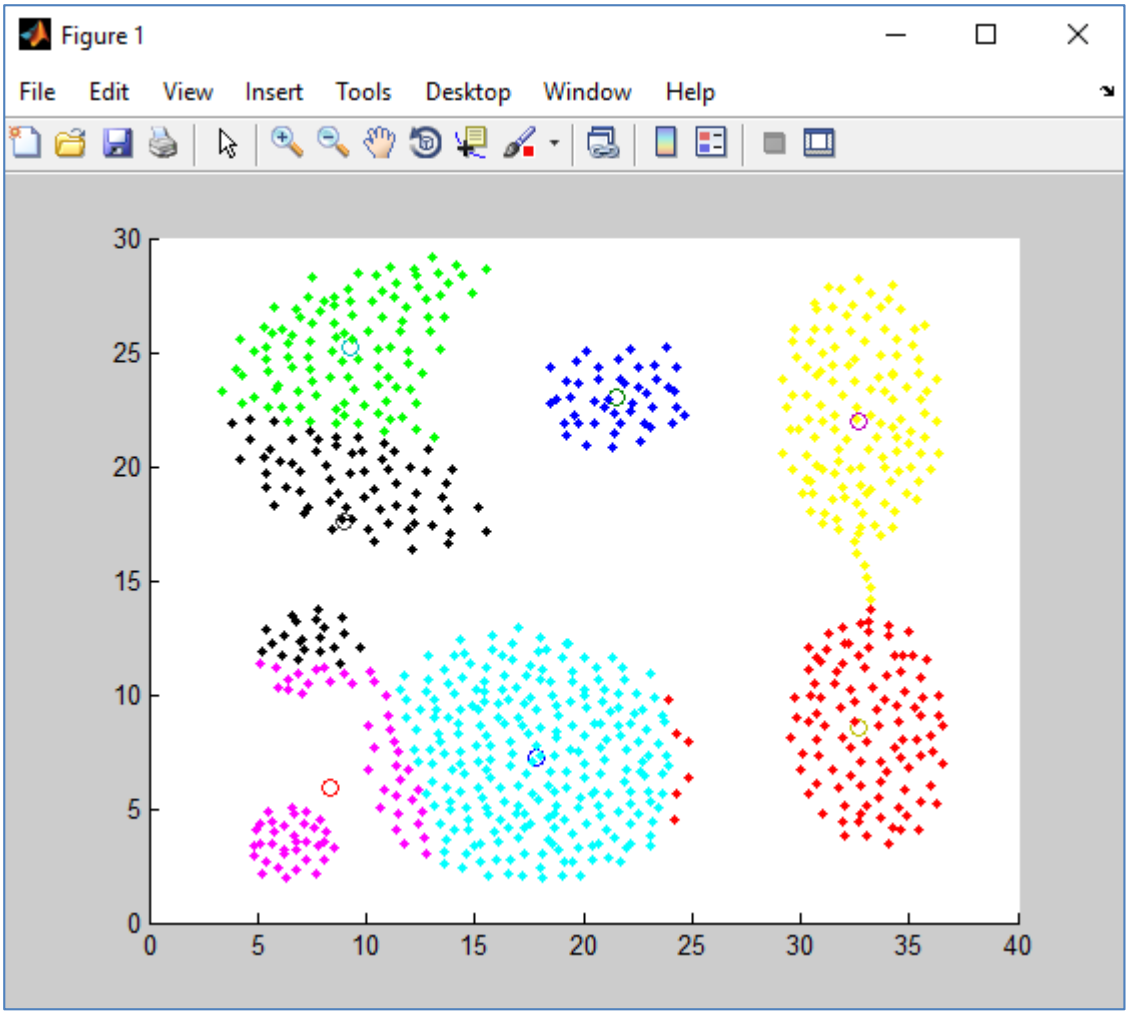
Kali ini akan dilakukan clustering dengan memilih satu data dari setiap kelas untuk dijadikan centroid awal, centroid 1 merupakan data dengan label 1 , dst. hingga centroid 7.

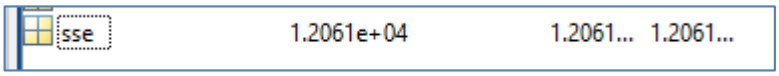
```
%centroid awal
centroids = zeros(max(a(:,3)),n-1);
 centroids(1,:) = a(710,1:2);
 centroids(2,:) = a(1,1:2);
 centroids(3,:) = a(478,1:2);
 centroids(4,:) = a(205,1:2);
 centroids(5,:) = a(755,1:2);
 centroids(6,:) = a(580,1:2);
 centroids(7,:) = a(171,1:2);
```

- Visualisasi data berdasarkan kelas dan centroid awal yang digunakan.



- Hasil clusetering data

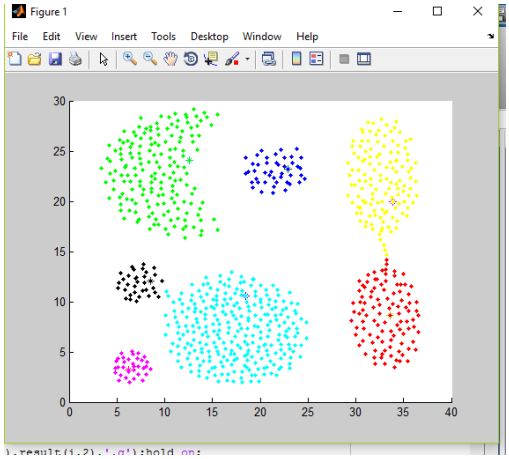




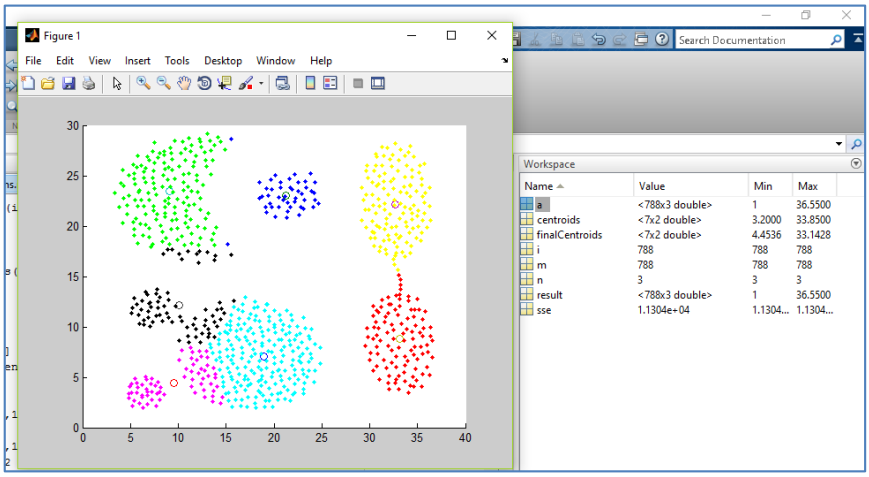
Running program dengan centroid yang berbeda

```
%centroid awal
centroids = zeros(max(a(:,3)),n-1);
 centroids(1,:) = a(743,1:2);
 centroids(2,:) = a(29,1:2);
 centroids(3,:) = a(525,1:2);
 centroids(4,:) = a(421,1:2);
 centroids(5,:) = a(765,1:2);
 centroids(6,:) = a(643,1:2);
 centroids(7,:) = a(189,1:2);
```

- Visualisasi data berdasarkan kelas dan centroid awal yang digunakan.



- Hasil clusetering data



g. K-Cluster dan K-Classes (ii)

Dengan melakukan memilih secara manual data dari setiap kelas yang akan dijadikan centroid awal akan memudahkan untuk mendapatkan kelompok data yang dibutuhkan. Tentunya dengan pewarnaan yang dapat disesuaikan dengan label kelas.

Clustering akan sesuai dengan kelas apabila centroid yang diambil merupakan merepresentasi dari data pada kelas tersebut. Namun apabila centroid yang digunakan merupakan centroid yang berada pada kelas yang sama namun terlalu dekat dengan kelas yang berbeda maka hal tersebut dapat menyebabkan ketidaksesuaian cluster data dengan kelas/kelompok data.

h. Clustering Result

Hasil clustering data sangat bergantung pada inisialisasi centroid data dan representasi data. Semakin centroid awal merepresentasi karakteristik dari keseluruhan data pada setiap cluster semakin kecil nilai error dan semakin baik hasil clustering yang didapatkan.