Clustering using K-Mean algorithm

K-means clustering is a clustering algorithm that aims to partition n observations into k clusters.

There are 3 steps:

Step 1:
Initialisation – K initial “means” (centroids) are generated at random

Step 2:
Assignment – K clusters are created by associating each observation with the nearest centroid

Step 3:
Update – The centroid of the clusters becomes the new mean

Assignment and Update are repeated iteratively until convergence

The end result is that the sum of squared errors is minimised between points and their respective centroids.

```
#-----------------------------------------------------------------
import copy
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
#-----------------------------------------------------------------

df = pd.DataFrame({
                 'x': [12, 20, 28, 18, 29, 33, 24, 45, 45, 52, 51, 52, 55, 53, 55, 61,
64, 69, 72],
               'y': [39, 36, 30, 52, 54, 46, 55, 59, 63, 70, 66, 63, 58, 23, 14, 8, 19,
7, 24]
               })

print("Step 1: Initialisation – K initial “means” (centroids) are generated at
random");

print("------------------------------------------");
print("Data set for training");
print("------------------------------------------");
print(df);
print("------------------------------------------");
np.random.seed(200)
k = 3
# centroids[i] = [x, y]
```

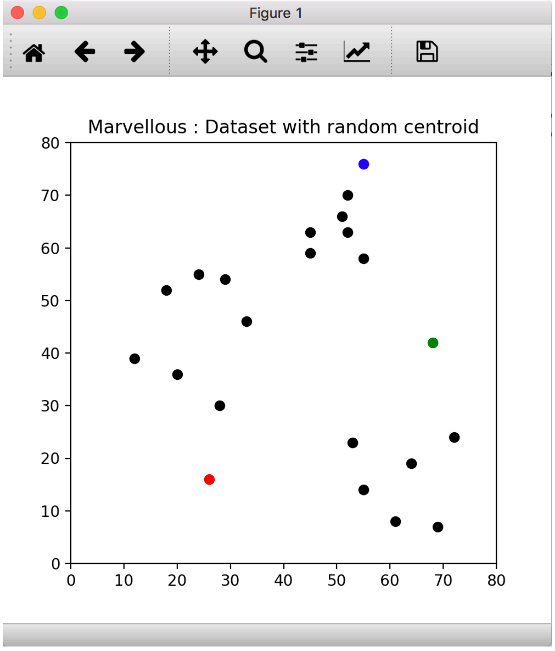
```
centroids = {
    i+1: [np.random.randint(0, 80), np.random.randint(0, 80)]
    for i in range(k)
}
print("-------------------------------------------------");
print("Random centroid generated");
print(centroids);
print("-------------------------------------------------");

fig = plt.figure(figsize=(5, 5))
plt.scatter(df['x'], df['y'], color='k')

colmap = {1: 'r', 2: 'g', 3: 'b'}
for i in centroids.keys():
    plt.scatter(*centroids[i], color=colmap[i])

plt.title("Marvellous : Dataset with random centroid");

plt.xlim(0, 80)
plt.ylim(0, 80)
plt.show()

#-------------------------------------------------------------------------------

# Assignment – K clusters are created by associating each observation with the
nearest centroid

def assignment(df, centroids):

    for i in centroids.keys():
        # sqrt((x1 - x2)^2 - (y1 - y2)^2)
        df['distance_from_{}'.format(i)] = (
                                    np.sqrt(
                                        (df['x'] - centroids[i][0]) ** 2
                                        + (df['y'] - centroids[i][1]) ** 2
                                        )
                                    )

          centroid_distance_cols = ['distance_from_{}'.format(i) for i in
centroids.keys()]

    df['closest'] = df.loc[:, centroid_distance_cols].idxmin(axis=1)

    df['closest'] = df['closest'].map(lambda x: int(x.lstrip('distance_from_')))

    df['color'] = df['closest'].map(lambda x: colmap[x])
    return df
```

```
print("Step 2 : Assignment – K clusters are created by associating each observation with the nearest centroid");

print("Before assignment dataset");
print(df)
df = assignment(df, centroids)

print("First centroid : Red");
print("Second centroid : Green");
print("Third centroid : Blue");

print("After assignment dataset");
print(df)

fig = plt.figure(figsize=(5, 5))
plt.scatter(df['x'], df['y'], color=df['color'], alpha=0.5, edgecolor='k')
for i in centroids.keys():
    plt.scatter(*centroids[i], color=colmap[i])
plt.xlim(0, 80)
plt.ylim(0, 80)
plt.title("Marvellous : Dataset with clustering & random centroid");
plt.show()

# ---------------------

old_centroids = copy.deepcopy(centroids)
print("Step 3:Update – The centroid of the clusters becomes the new mean Assignment and Update are repeated iteratively until convergence");

def update(k):
    print("Old values of centroids");
    print(k);

    for i in centroids.keys():
        centroids[i][0] = np.mean(df[df['closest'] == i]['x'])
        centroids[i][1] = np.mean(df[df['closest'] == i]['y'])

    print("New values of centroids");
    print(k);
    return k

centroids = update(centroids)

fig = plt.figure(figsize=(5, 5))
ax = plt.axes()
plt.scatter(df['x'], df['y'], color=df['color'], alpha=0.5, edgecolor='k')
for i in centroids.keys():
    plt.scatter(*centroids[i], color=colmap[i])
plt.xlim(0, 80)
plt.ylim(0, 80)

for i in old_centroids.keys():
    old_x = old_centroids[i][0]
    old_y = old_centroids[i][1]
    dx = (centroids[i][0] - old_centroids[i][0]) * 0.75
    dy = (centroids[i][1] - old_centroids[i][1]) * 0.75
      ax.arrow(old_x, old_y, dx, dy, head_width=2, head_length=3, fc=colmap[i],
ec=colmap[i])

plt.title("Marvellous : Dataset with clustering and updated centroids");
plt.show()

#--------------------

## Repeat Assigment Stage
print("Before assignment dataset");
print(df)
df = assignment(df, centroids)
print("After assignment dataset");
print(df)

# Plot results
fig = plt.figure(figsize=(5, 5))
plt.scatter(df['x'], df['y'], color=df['color'], alpha=0.5, edgecolor='k')
for i in centroids.keys():
    plt.scatter(*centroids[i], color=colmap[i])
plt.xlim(0, 80)
plt.ylim(0, 80)
plt.title("Marvellous : Dataset with clustering and updated centroids");
plt.show()

# Continue until all assigned categories don't change any more
while True:
    closest_centroids = df['closest'].copy(deep=True)
    centroids = update(centroids)
    print("Before assignment dataset");
    print(df)
    df = assignment(df, centroids)
    print("After assignment dataset");
    print(df)
    if closest_centroids.equals(df['closest']):
        break

print("Final values of centroids");
print(centroids);
```

```
fig = plt.figure(figsize=(5, 5))
plt.scatter(df['x'], df['y'], color=df['color'], alpha=0.5, edgecolor='k')
for i in centroids.keys():
    plt.scatter(*centroids[i], color=colmap[i])
plt.xlim(0, 80)
plt.ylim(0, 80)
plt.title("Marvellous : Final dataset with set centroids");
plt.show()
```

Output of above application :

```
Step 1: Initialisation – K initial “means” (centroids) are generated at random
------------------------------------------
Data set for training
------------------------------------------
     x   y
0   12  39
1   20  36
2   28  30
3   18  52
4   29  54
5   33  46
6   24  55
7   45  59
8   45  63
9   52  70
10  51  66
11  52  63
12  55  58
13  53  23
14  55  14
15  61   8
16  64  19
17  69   7
18  72  24
------------------------------------------
Random centroid generated
{1: [26, 16], 2: [68, 42], 3: [55, 76]}
```



Step 2 : Assignment – K clusters are created by associating each observation with the nearest centroid

Before assignment dataset

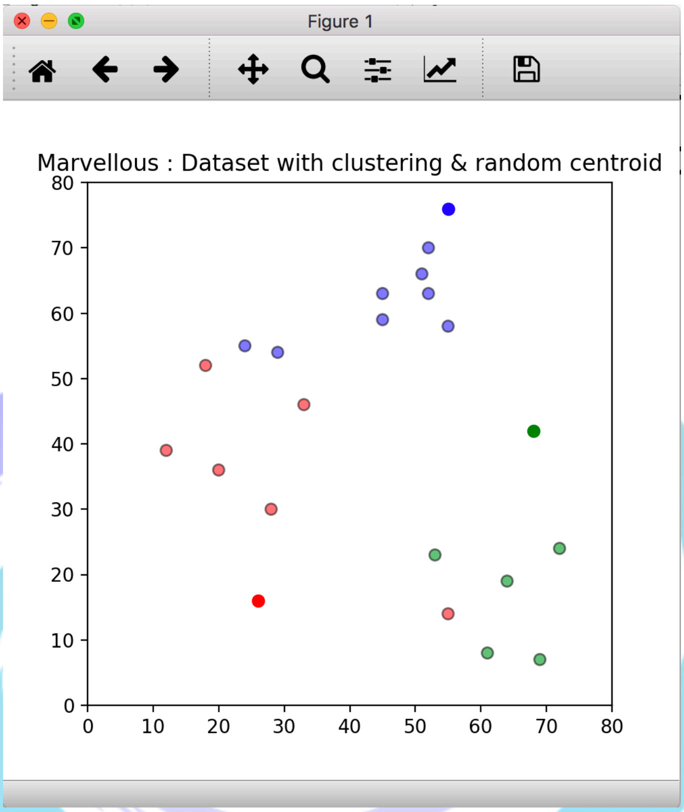
```
     x   y
0   12  39
1   20  36
2   28  30
3   18  52
4   29  54
5   33  46
6   24  55
7   45  59
8   45  63
9   52  70
10  51  66
11  52  63
12  55  58
13  53  23
14  55  14
15  61   8
16  64  19
17  69   7
18  72  24
```

First centroid : Red
Second centroid : Green
Third centroid : Blue

After assignment dataset

```
     x   y  distance_from_1  distance_from_2  distance_from_3  closest color
0   12  39        26.925824        56.080300        56.727418        1     r
1   20  36        20.880613        48.373546        53.150729        1     r
2   28  30        14.142136        41.761226        53.338541        1     r
3   18  52        36.878178        50.990195        44.102154        1     r
4   29  54        38.118237        40.804412        34.058773        3     b
5   33  46        30.805844        35.227830        37.202150        1     r
6   24  55        39.051248        45.880279        37.443290        3     b
7   45  59        47.010637        28.600699        19.723083        3     b
8   45  63        50.695167        31.144823        16.401219        3     b
9   52  70        59.933296        32.249031         6.708204        3     b
10  51  66        55.901699        29.410882        10.770330        3     b
11  52  63        53.712196        26.400758        13.341664        3     b
12  55  58        51.039201        20.615528        18.000000        3     b
13  53  23        27.892651        24.207437        53.037722        2     g
14  55  14        29.068884        30.870698        62.000000        1     r
15  61   8        35.902646        34.713110        68.264193        2     g
```

```
16  64  19       38.118237       23.345235       57.706152       2     g
17  69   7      43.931765       35.014283       70.405966       2     g
18  72  24       46.690470       18.439089       54.708317       2     g
```



Step 3:Update – The centroid of the clusters becomes the new mean Assignment and Update are repeated iteratively until convergence
Old values of centroids
{1: [26, 16], 2: [68, 42], 3: [55, 76]}

New values of centroids
{1: [27.666666666666668, 36.166666666666664], 2: [63.8, 16.2], 3: [44.125, 61.0]}

Note : New centroids are mean of the generated X and Y coordinates of clustering members. We can use additions of distances to decide the error rate.

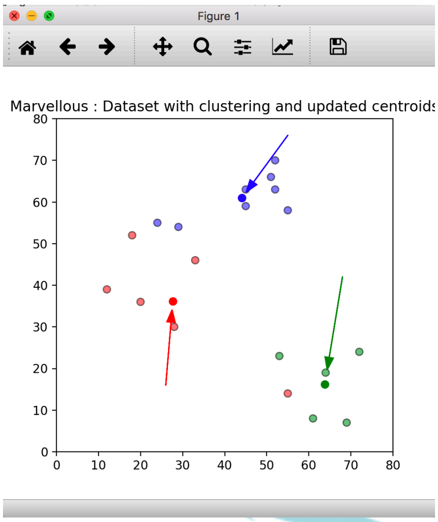
Example :
X for centroid 1 : (12+20+28+18+55) / 6 = 27.66
Y for centroid 2 : (39+36+30+52+14) / 6 = 36.16

Before assignment dataset

```
     x   y  distance_from_1  distance_from_2  distance_from_3  closest color
0   12  39        26.925824        56.080300        56.727418        1     r
1   20  36        20.880613        48.373546        53.150729        1     r
2   28  30        14.142136        41.761226        53.338541        1     r
3   18  52        36.878178        50.990195        44.102154        1     r
4   29  54        38.118237        40.804412        34.058773        3     b
5   33  46        30.805844        35.227830        37.202150        1     r
6   24  55        39.051248        45.880279        37.443290        3     b
7   45  59        47.010637        28.600699        19.723083        3     b
8   45  63        50.695167        31.144823        16.401219        3     b
9   52  70        59.933296        32.249031         6.708204        3     b
10  51  66        55.901699        29.410882        10.770330        3     b
11  52  63        53.712196        26.400758        13.341664        3     b
12  55  58        51.039201        20.615528        18.000000        3     b
13  53  23        27.892651        24.207437        53.037722        2     g
14  55  14        29.068884        30.870698        62.000000        1     r
15  61   8        35.902646        34.713110        68.264193        2     g
16  64  19        38.118237        23.345235        57.706152        2     g
17  69   7        43.931765        35.014283        70.405966        2     g
18  72  24        46.690470        18.439089        54.708317        2     g
```

After assignment dataset

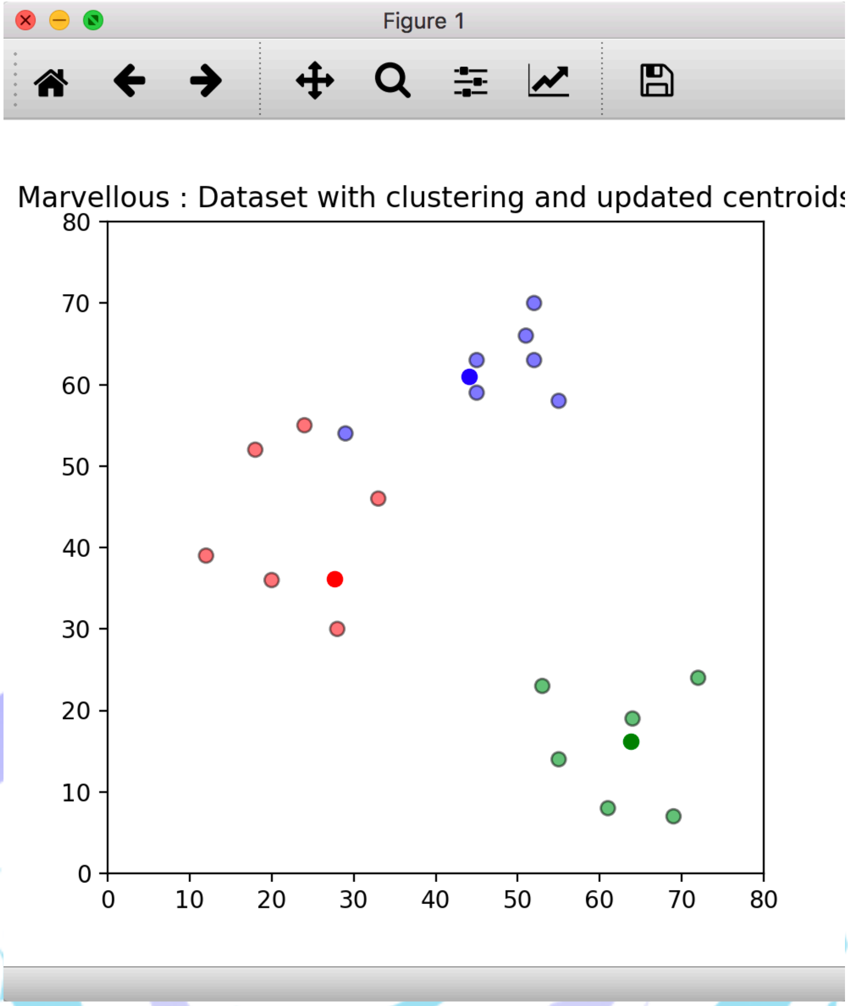
```
     x   y  distance_from_1  distance_from_2  distance_from_3  closest color
0   12  39        15.920811        56.595760        38.936045        1     r
1   20  36         7.668478        48.067453        34.742130        1     r
2   28  30         6.175669        38.367695        34.943034        1     r
3   18  52        18.550981        58.131575        27.631786        1     r
4   29  54        17.883108        51.379763        16.666302        3     b
5   33  46        11.186549        42.856505        18.675268        1     r
6   24  55        19.186946        55.583091        21.000372        1     r
7   45  59        28.667151        46.746979         2.183031        3     b
8   45  63        31.944831        50.434909         2.183031        3     b
9   52  70        41.674999        55.078853        11.958914        3     b
10  51  66        37.874427        51.418674         8.500919        3     b
11  52  63        36.223458        48.264687         8.125000        3     b
12  55  58        34.982932        42.716273        11.281207        3     b
13  53  23        28.550637        12.762445        39.022630        2     g
14  55  14        35.191934         9.070832        48.241742        2     g
15  61   8        43.640259         8.664872        55.621629        2     g
16  64  19        40.184643         2.807134        46.465209        2     g
17  69   7        50.587932        10.567876        59.453895        2     g
18  72  24        45.972516        11.317243        46.325108        2     g
```



```
Old values of centroids
{1: [27.666666666666668, 36.166666666666664], 2: [63.8, 16.2], 3: [44.125,
61.0]}

New values of centroids
{1: [22.5, 43.0], 2: [62.333333333333336, 15.833333333333334], 3: [47.0,
61.857142857142854]}

Before assignment dataset
     x   y  distance_from_1  distance_from_2  distance_from_3  closest color
0   12  39        15.920811        56.595760        38.936045        1     r
1   20  36         7.668478        48.067453        34.742130        1     r
2   28  30         6.175669        38.367695        34.943034        1     r
3   18  52        18.550981        58.131575        27.631786        1     r
4   29  54        17.883108        51.379763        16.666302        3     b
5   33  46        11.186549        42.856505        18.675268        1     r
6   24  55        19.186946        55.583091        21.000372        1     r
7   45  59        28.667151        46.746979         2.183031        3     b
8   45  63        31.944831        50.434909         2.183031        3     b
9   52  70        41.674999        55.078853        11.958914        3     b
10  51  66        37.874427        51.418674         8.500919        3     b
11  52  63        36.223458        48.264687         8.125000        3     b
12  55  58        34.982932        42.716273        11.281207        3     b
13  53  23        28.550637        12.762445        39.022630        2     g
14  55  14        35.191934         9.070832        48.241742        2     g
15  61   8        43.640259         8.664872        55.621629        2     g
16  64  19        40.184643         2.807134        46.465209        2     g
17  69   7        50.587932        10.567876        59.453895        2     g
18  72  24        45.972516        11.317243        46.325108        2     g
```

After assignment dataset

```
    x   y  distance_from_1  distance_from_2  distance_from_3  closest color
0   12  39       11.236103        55.408834        41.802500        1     r
1   20  36        7.433034        46.891423        37.384380        1     r
2   28  30       14.115594        37.141247        37.092823        1     r
3   18  52       10.062306        57.214266        30.629451        1     r
4   29  54       12.776932        50.673519        19.640130        1     r
5   33  46       10.920165        42.076980        21.152990        1     r
6   24  55       12.093387        54.803943        24.000425        1     r
7   45  59       27.608875        46.516723         3.487587        3     b
8   45  63       30.103986        50.250760         2.303502        3     b
9   52  70       39.990624        55.143500         9.555424        3     b
10  51  66       36.623080        51.430914         5.758756        3     b
11  52  63       35.640567        48.285321         5.128949        3     b
12  55  58       35.794553        42.799598         8.881303        3     b
13  53  23       36.472592        11.767422        39.317649        2     g
14  55  14       43.557433         7.559027        48.521193        2     g
15  61   8       52.031241         7.945998        55.647029        2     g
16  64  19       47.940067         3.578485        46.105690        2     g
17  69   7       58.806887        11.066717        59.104197        2     g
18  72  24       53.021222        12.654600        45.366984        2     g
Old values of centroids
{1: [22.5, 43.0], 2: [62.333333333333336, 15.833333333333334], 3: [47.0, 61.857142857142854]}

New values of centroids
{1: [23.428571428571427, 44.57142857142857], 2: [62.333333333333336, 15.833333333333334], 3: [50.0, 63.166666666666664]}
```



```
Before assignment dataset
     x   y  distance_from_1  distance_from_2  distance_from_3  closest color
0   12  39        11.236103        55.408834        41.802500        1     r
1   20  36         7.433034        46.891423        37.384380        1     r
2   28  30        14.115594        37.141247        37.092823        1     r
3   18  52        10.062306        57.214266        30.629451        1     r
4   29  54        12.776932        50.673519        19.640130        1     r
5   33  46        10.920165        42.076980        21.152990        1     r
6   24  55        12.093387        54.803943        24.000425        1     r
7   45  59        27.608875        46.516723         3.487587        3     b
8   45  63        30.103986        50.250760         2.303502        3     b
9   52  70        39.990624        55.143500         9.555424        3     b
10  51  66        36.623080        51.430914         5.758756        3     b
11  52  63        35.640567        48.285321         5.128949        3     b
12  55  58        35.794553        42.799598         8.881303        3     b
13  53  23        36.472592        11.767422        39.317649        2     g
14  55  14        43.557433         7.559027        48.521193        2     g
15  61   8        52.031241         7.945998        55.647029        2     g
16  64  19        47.940067         3.578485        46.105690        2     g
17  69   7        58.806887        11.066717        59.104197        2     g
18  72  24        53.021222        12.654600        45.366984        2     g
```

```
After assignment dataset
     x   y  distance_from_1  distance_from_2  distance_from_3  closest color
0   12  39        12.714286        55.408834        45.033629        1     r
1   20  36         9.231711        46.891423        40.472556        1     r
2   28  30        15.271689        37.141247        39.799846        1     r
3   18  52         9.200710        57.214266        33.892395        1     r
4   29  54        10.951656        50.673519        22.913485        1     r
5   33  46         9.677451        42.076980        24.159769        1     r
6   24  55        10.444215        54.803943        27.252421        1     r
7   45  59        25.952075        46.516723         6.508541        3     b
8   45  63        28.371443        50.250760         5.002777        3     b
9   52  70        38.248383        55.143500         7.120003        3     b
10  51  66        34.919441        51.430914         3.004626        3     b
11  52  63        33.999100        48.285321         2.006932        3     b
12  55  58        34.308623        42.799598         7.189885        3     b
13  53  23        36.603223        11.767422        40.278544        2     g
14  55  14        43.947325         7.559027        49.420250        2     g
15  61   8        52.431685         7.945998        56.252654        2     g
16  64  19        47.957677         3.578485        46.332434        2     g
17  69   7        59.062402        11.066717        59.293292        2     g
18  72  24        52.748150        12.654600        44.922464        2     g
Final values of centroids
{1: [23.428571428571427, 44.57142857142857], 2: [62.333333333333336, 15.833333333333334], 3: [50.0, 63.166666666666664]}
```

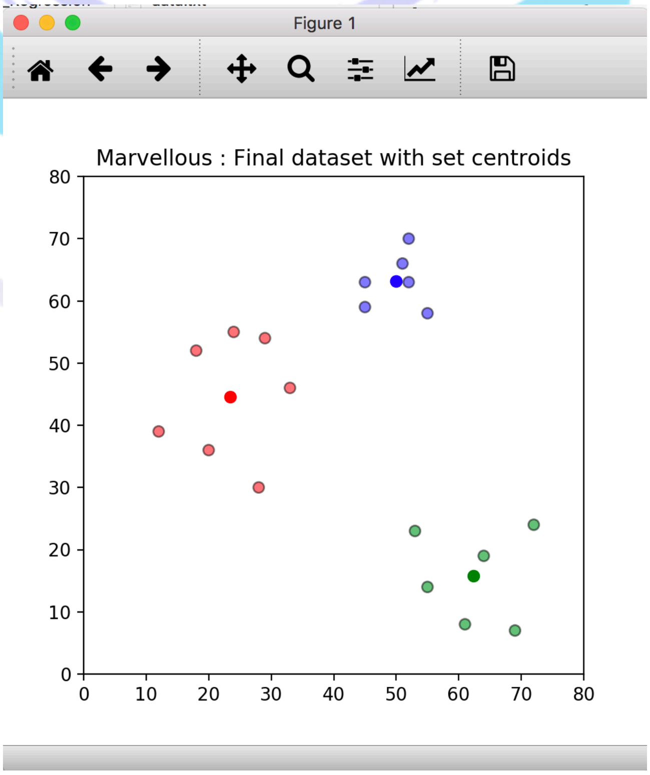


Figure 1
Marvellous : Final dataset with set centroids
80
70
60
50
40
30
20
10
0
0
10
20
30
40
50
60
70
80