

# ML (K-Means Clustering Algorithm)

# Roll No.: 19103014

# Name: Amritpal Singh

# Branch: CSE

# Group: G1

**CODE :**

import pandas as pd

import numpy as np

import random

import matplotlib.pyplot as plt

import sys

def get\_centroids(K,X,m):

centroids = {}

for i in range(K):

rand=random.randint(0,m-1)

centroids['c'+str(i+1)] = list(X[rand])

return centroids

def calc\_distance(X1, X2):

return (sum((X1 - X2)\*\*2))\*\*0.5

def get\_values(K,n\_iter,X, centroids,m):

output={}

for key,value in centroids.items():

output[key] = []

for row in range(m):

min\_dis = sys.maxsize

min\_val = 'c1'

for key,value in centroids.items():

dis = calc\_distance(X[row],value)

if(dis<min\_dis):

min\_dis = dis

min\_val = key

output[min\_val].append(row)

Y={}

for i in range(n\_iter):

centroids = {}

for key,values in output.items():

x=0;y=0

for j in values:

x+=X[j][0]

y+=X[j][1]

centroids[key] = [x/len(values),y/len(values)]

output={}

for key,value in centroids.items():

output[key] = []

for row in range(m):

min\_dis = sys.maxsize

min\_val = 'c1'

for key,value in centroids.items():

dis = calc\_distance(X[row],value)

if(dis<min\_dis):

min\_dis = dis

min\_val = key

output[min\_val].append(row)

return output, centroids

def k\_means(K,n\_iter,X):

m=X.shape[0]

n=X.shape[1]

centroids = get\_centroids(K,X,m)

output, centroids = get\_values(K,n\_iter,X, centroids,m)

final = []

for key, values in output.items():

li=[]

li1=[]

li2=[]

for i in values:

li1.append(X[i][0])

li2.append(X[i][1])

li.append(li1)

li.append(li2)

final.append(li)

final\_c = []

li1=[]

li2=[]

for key, values in centroids.items():

li1.append(values[0])

li2.append(values[1])

final\_c.append(li1)

final\_c.append(li2)

return final, final\_c

dataset=pd.read\_csv('Mall\_Customers.csv')

dataset.describe()

dataset=dataset.values

X=dataset[:,3:]

K=5

final, final\_c = k\_means(K,100,X)

plt.scatter(X[:,0],X[:,1],c='black',label='unclustered data')

plt.xlabel('Annual Income')

plt.ylabel('Spending Score')

plt.legend()

plt.title('Actual data points of Mall database')

plt.show()

color=['red','blue','green','cyan','magenta']

labels=['cluster1','cluster2','cluster3','cluster4','cluster5']

for k in range(K):

plt.scatter(final[k][0],final[k][1],c=color[k],label=labels[k])

plt.scatter(final\_c[0],final\_c[1],s=300,c='yellow',label='Centroids')

plt.xlabel('Annual Income')

plt.ylabel('Spending Score')

plt.legend()

plt.title('Final clustering for mall database')

plt.show()

dataset=pd.read\_csv('iris.csv')

dataset.describe()

dataset=dataset.values

X=dataset[:,:2]

K=3

final, final\_c = k\_means(3,50,X)

plt.scatter(X[:,0],X[:,1],c='black',label='unclustered data')

plt.xlabel('sepal length')

plt.ylabel('sepal width')

plt.legend()

plt.title('Actual data points of iris database')

plt.show()

color=['red','blue','green','cyan','magenta']

labels=['cluster1','cluster2','cluster3','cluster4','cluster5']

for k in range(K):

plt.scatter(final[k][0],final[k][1],c=color[k],label=labels[k])

plt.scatter(final\_c[0],final\_c[1],s=300,c='yellow',label='Centroids')

plt.xlabel('sepal length')

plt.ylabel('sepal width')

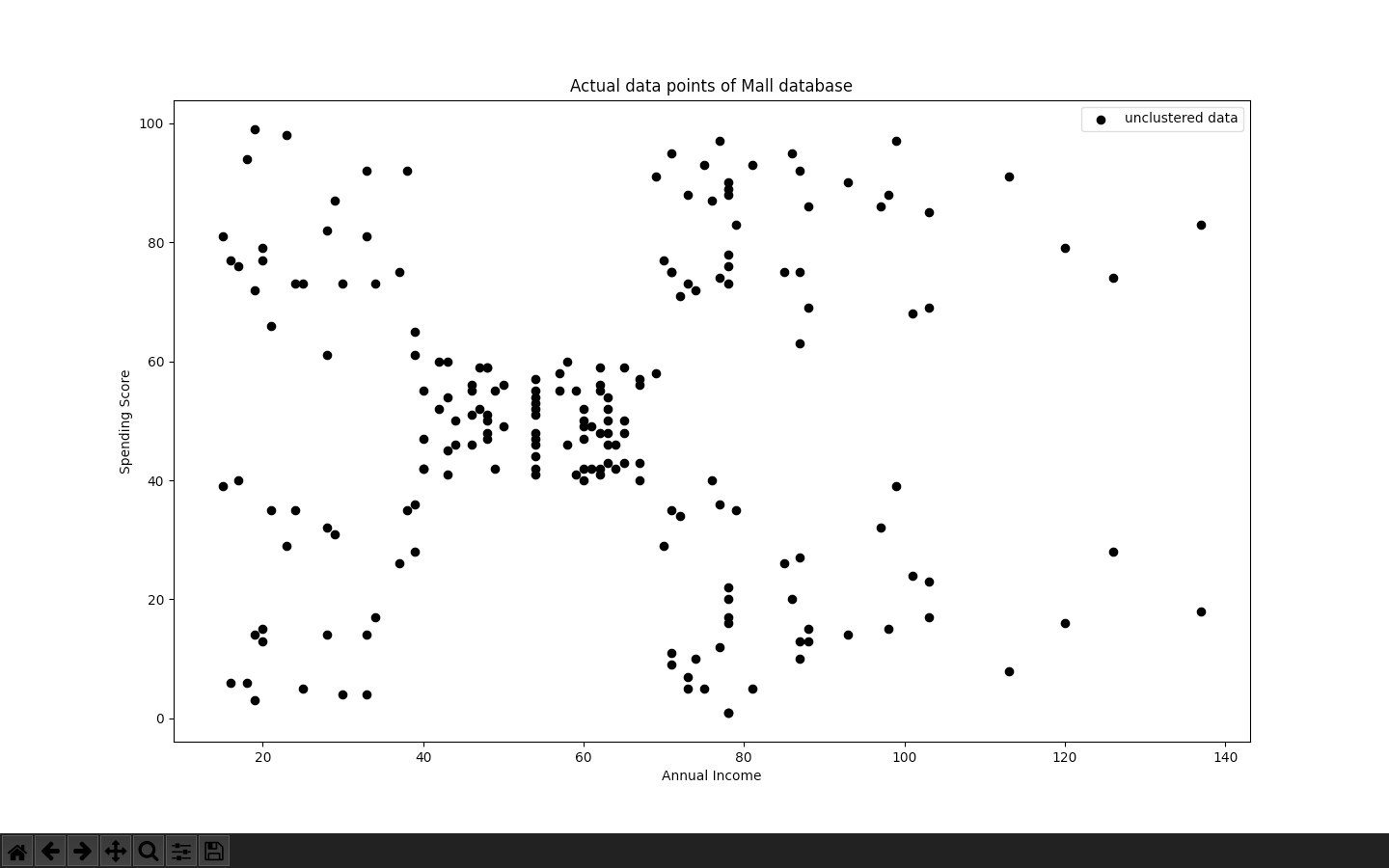
plt.legend()

plt.title('Final clustering for iris database')

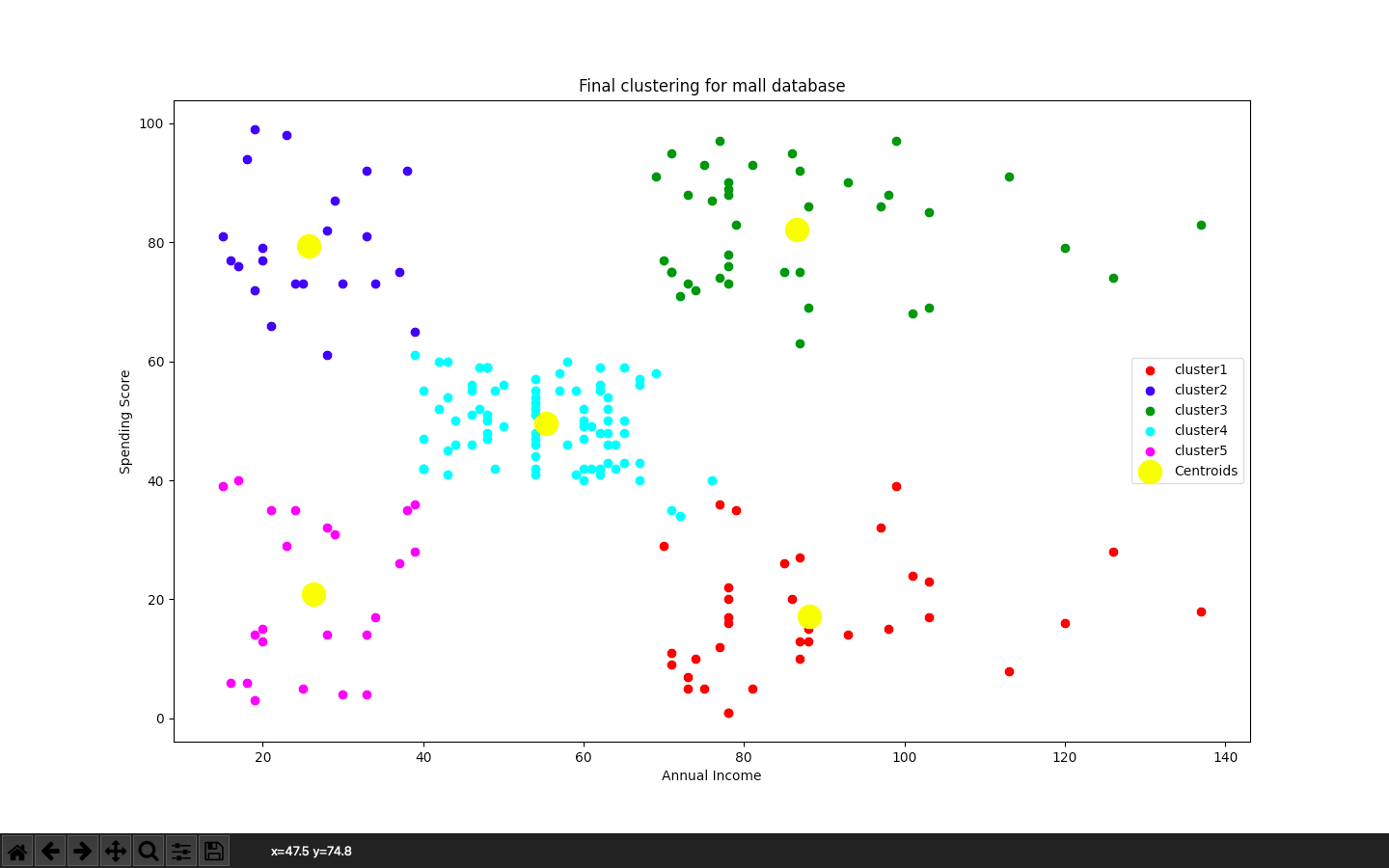
plt.show()

**MALL\_CUSTOMERS.CSV**

Actual data points of Mall database

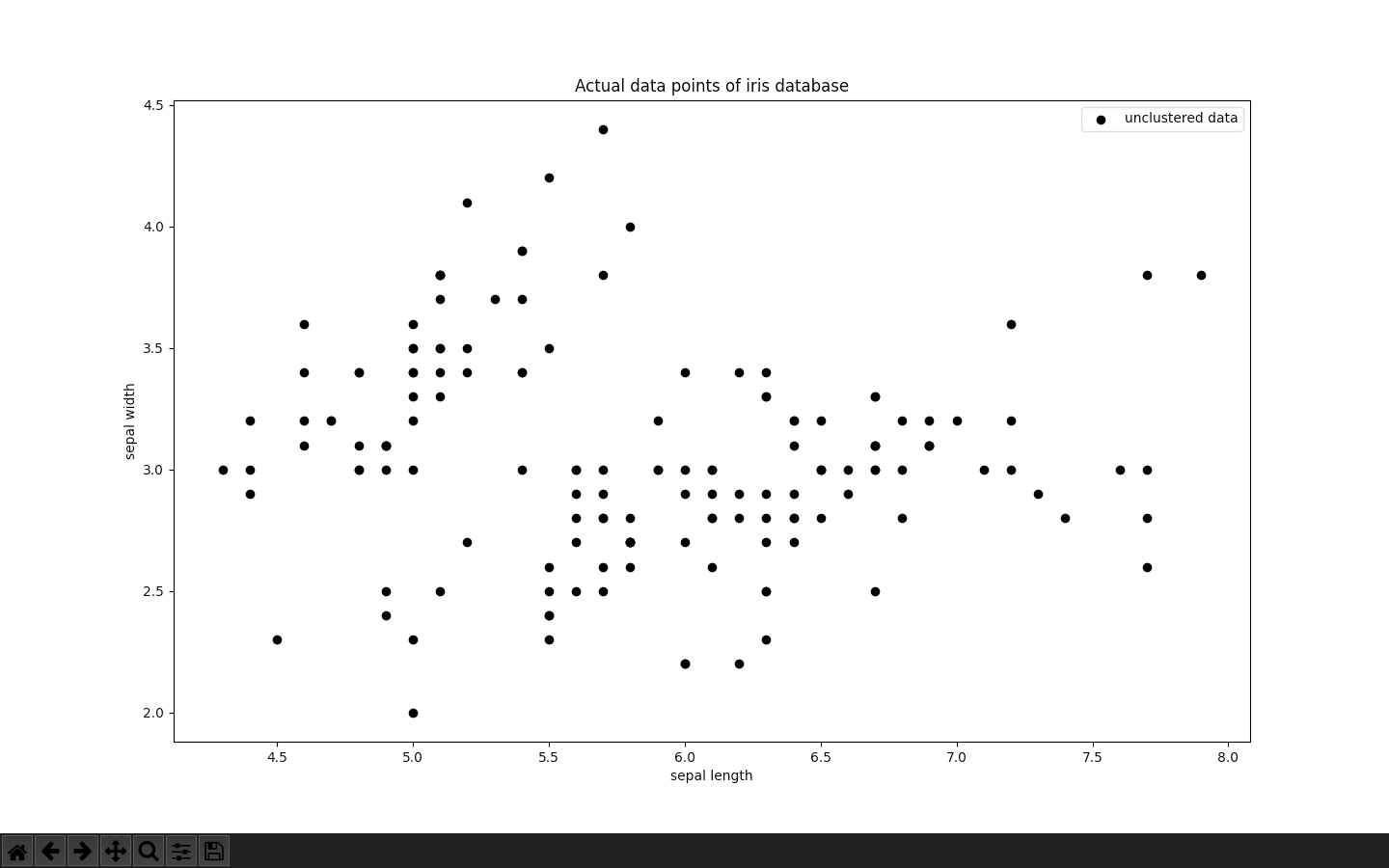
****

'Final clustering for mall database'

****

**IRIS.CSV**

Actual data points of iris database



Final clustering for iris database

