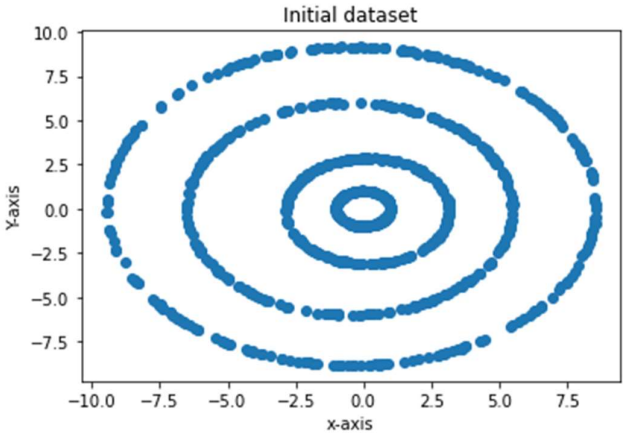
QUESTIONS

1. You are given a data set with 1000 data points each in R^2.

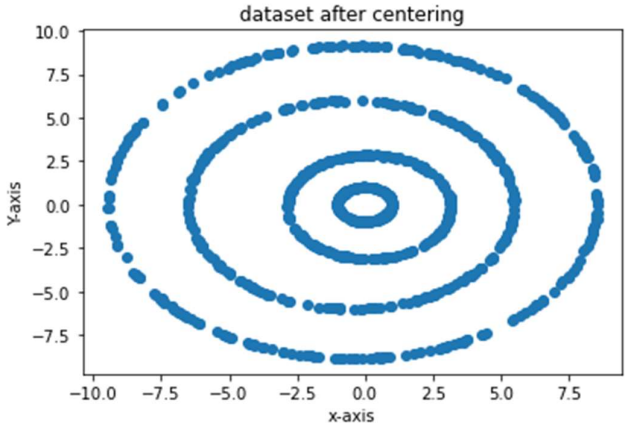
i. Write a piece of code to run the PCA algorithm on this data set. How much ofthe variance in the data set is explained by each of the principal components?

ANSWER-

Step 1 – Load the dataset



Step 2 – Mean centering (we do mean centering by subtracting the mean from all features.

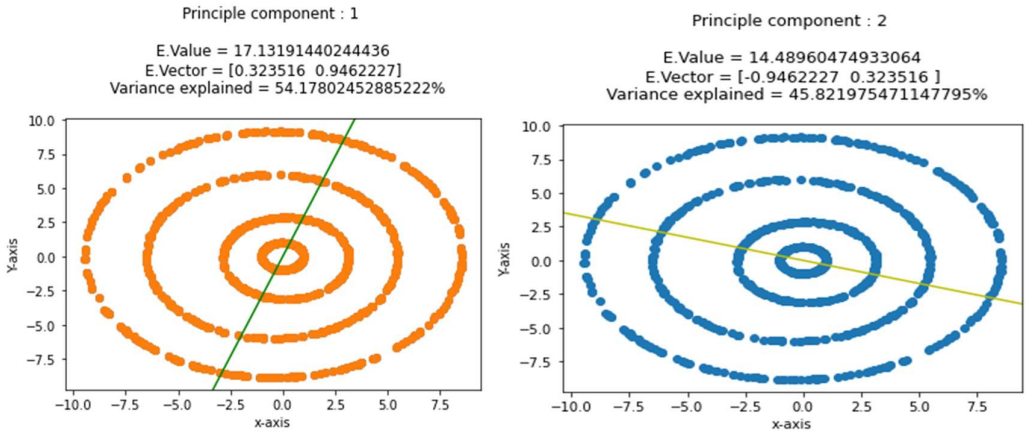


Step 3 – Compute the covariance matrix

Step 4 – Computer the Eigen value and eigenvectors of the covariance matrix

Eigen Value 1 = 17.1319144
Eigen Value 2 = 14.48960475
Eigen Vector 1 = [0.323516 0.9462227]
Eigen Vector 2= [-0.9462227 0.323516]

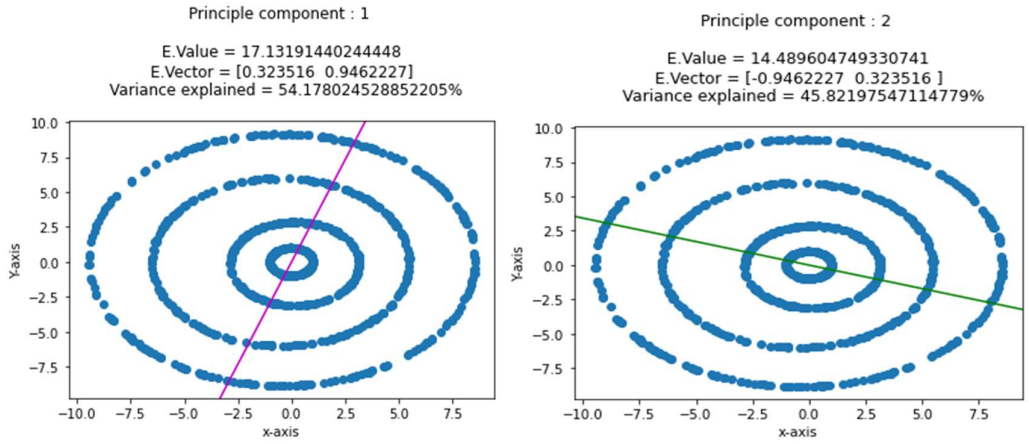
Step 5 – compute the variance explained and select k top eigenvectors



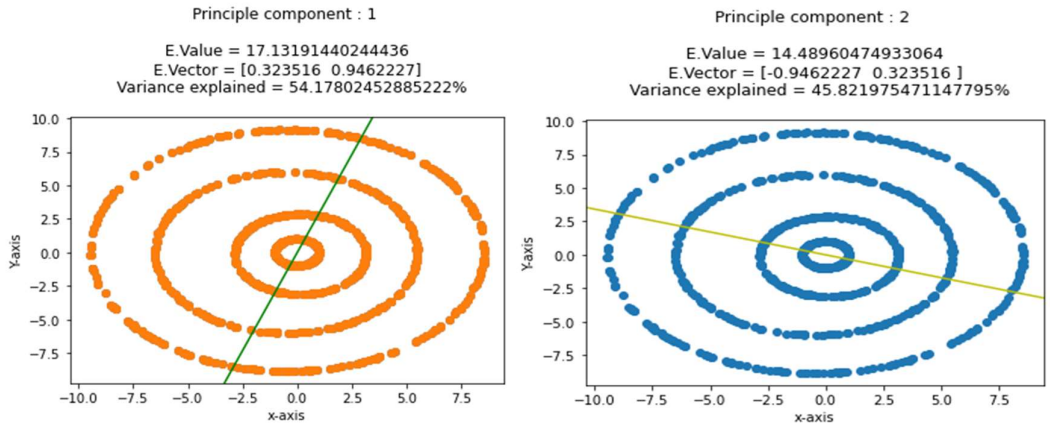
ii. Study the effect of running PCA without Centering the data set. What are your observations? Does Centering help?

ANSWER: Principal component analysis with centering does not help much in this data set because data is already centered(almost).

Principal components and variance explained "without" centering -



Principal component and variance explained "with" centering -



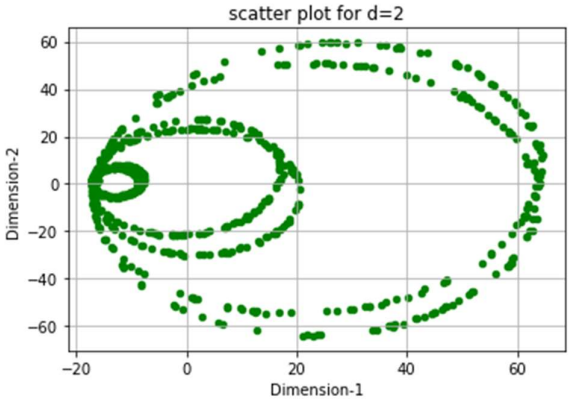
iii. Write a piece of code to implement the Kernel PCA algorithm on this dataset. Use the following kernels

Plot the projection of each point in the dataset onto the top-2 components foreach kernel.

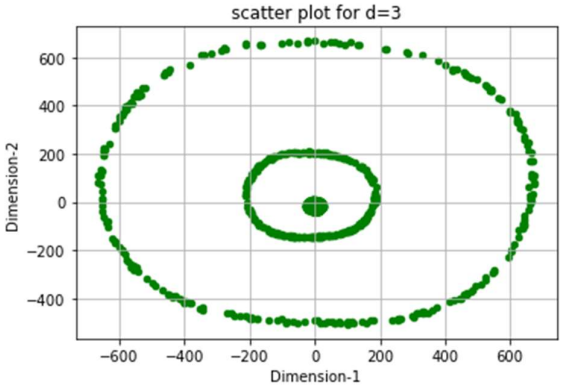
A. $\kappa(x, y) = (1 + x^T y)^d$ for $d = \{2, 3\}$

Ans.

Input : Degree of Polynomial(d) = 2



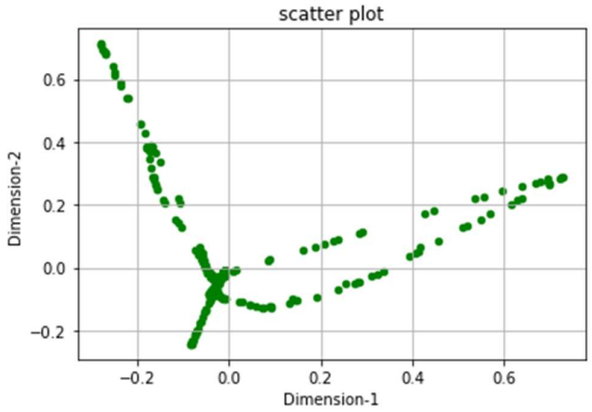
Input: Degree of Polynomial(d) = 3



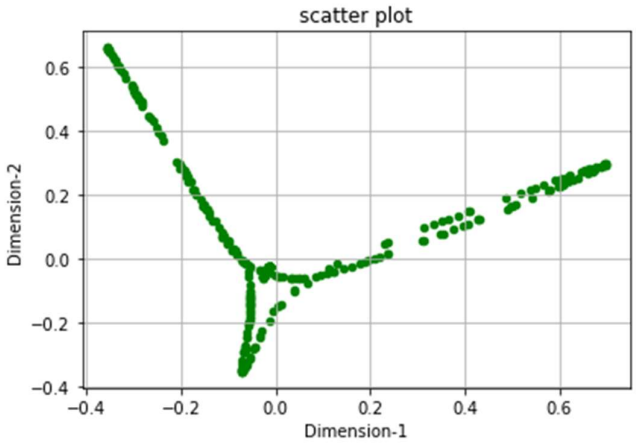
B. $\kappa(x, y) = \exp^{-(x-y)^T (x-y)}$ for $\sigma = \{0.1, 0.2, \ldots, 1\}$

Ans.

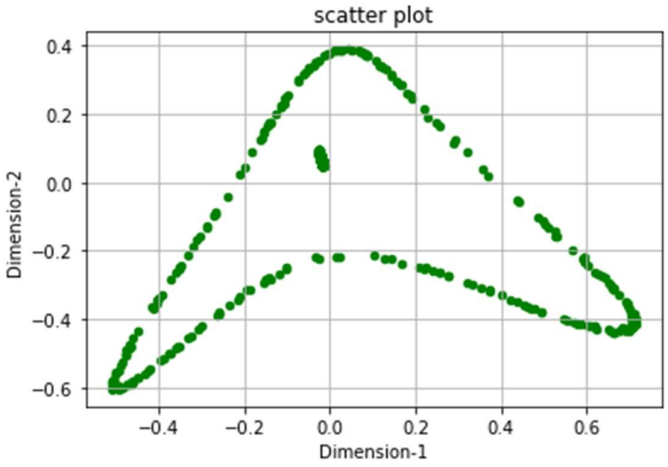
Input : Value of sigma = 0.1



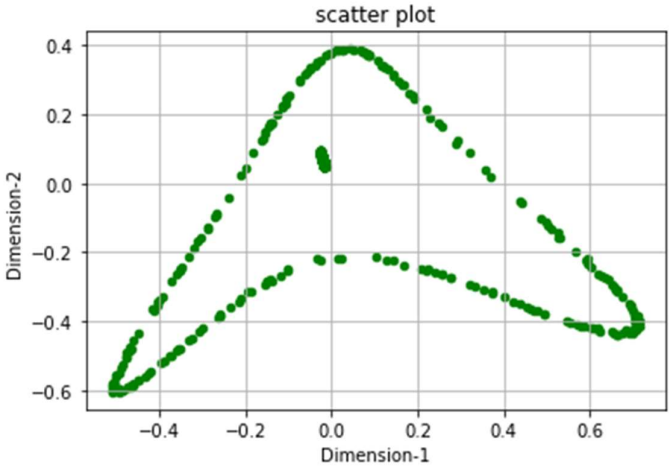
Input : Value of Sigma = 0.2



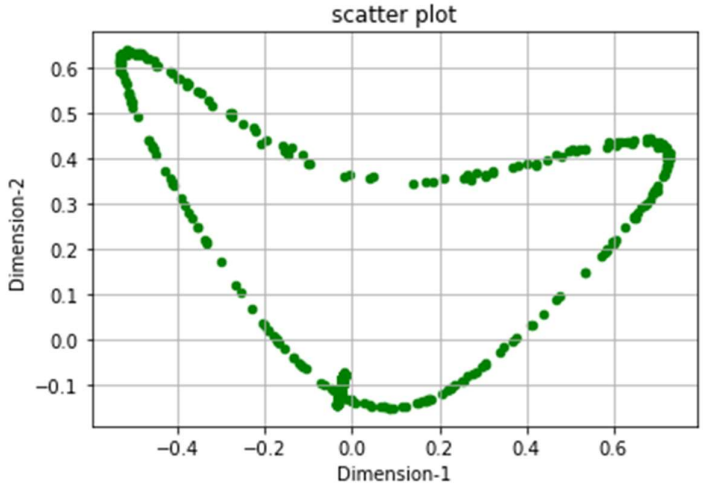
Input : Value of Sigma = 0.3



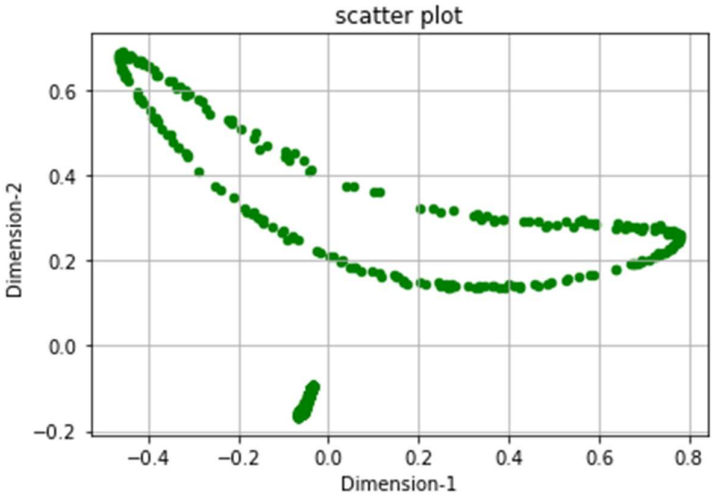
Input : Value of Sigma = 0.4



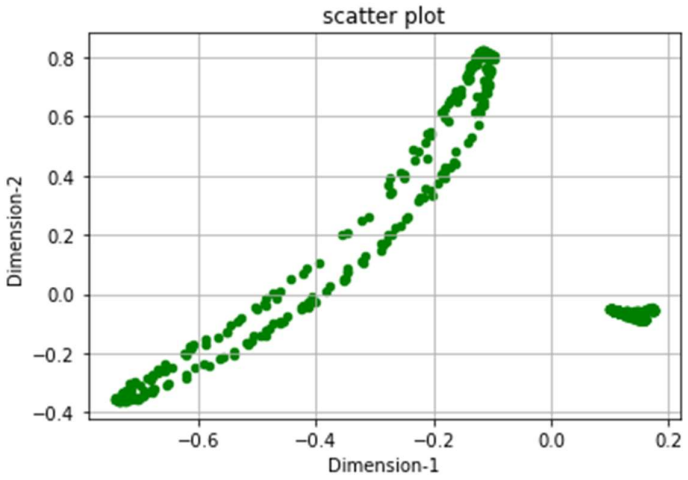
Input : Value of Sigma = 0.5



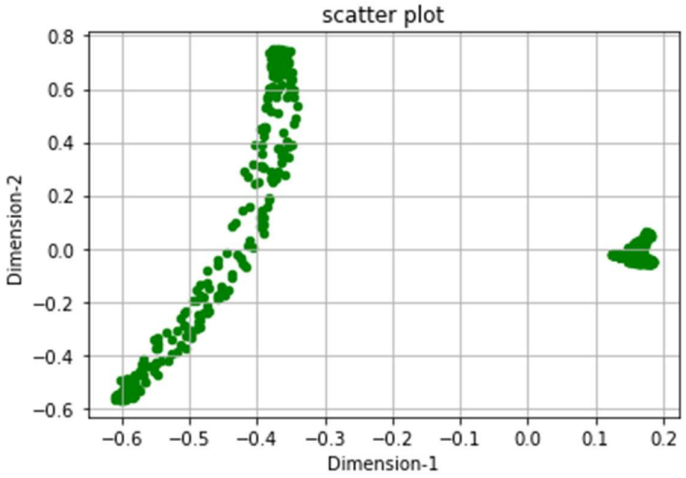
Input : Value of Sigma = 0.6



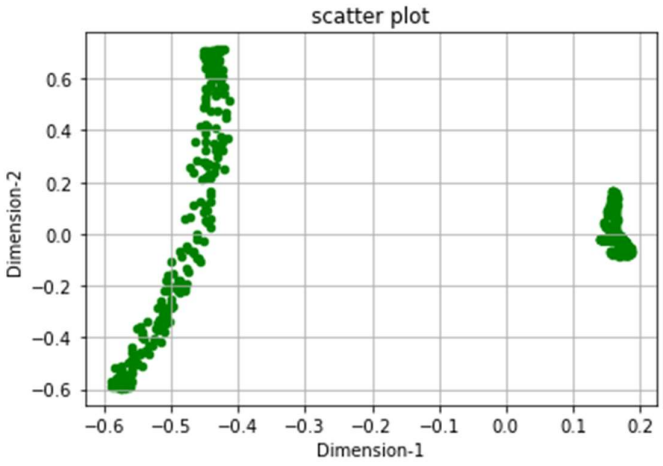
Input : Value of Sigma = 0.7



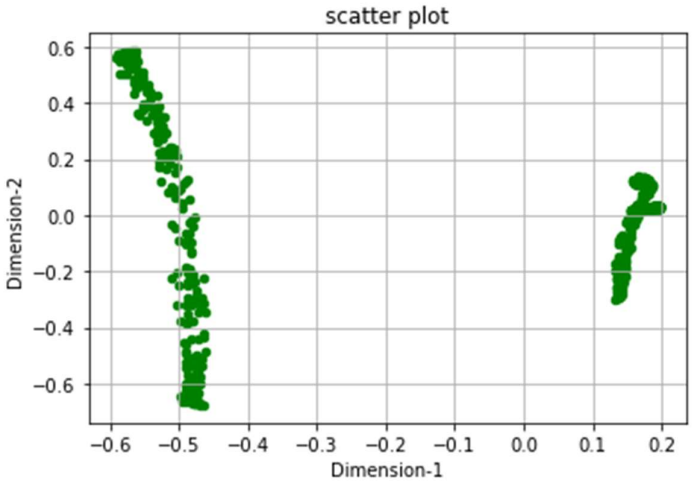
Input : Value of Sigma = 0.8



Input : Value of Sigma = 0.9



Input : Value of Sigma = 1.0



iv. Which Kernel do you think is best suited for this dataset and why?

Answer. The Polynomial Kernel is best suited for this dataset because the variance explained by the polynomial kernel is the highest. We need such a kernel that preserves the maximum features of the dataset. As we can see from the above plots, with kernel polynomial, our data is linearly seperable and has maximum variance.

2. You are given a data set with 1000 data points each in R^2.

i. Write a piece of code to run the algorithm studied in class for the K-means problem with $k = 4$. Try 5 different random initialization and plot the error function w.r.t iterations in each case. In each case, plot the clusters obtained in different colors.

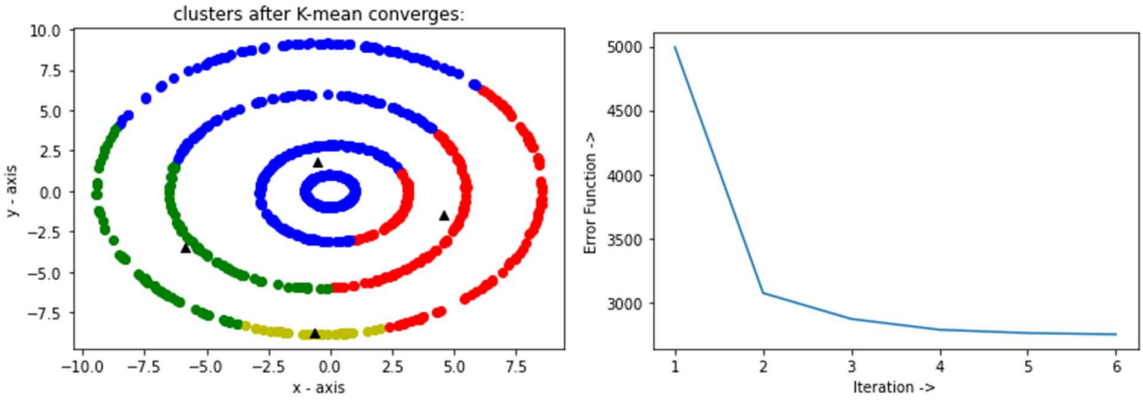
Ans.

Step 1: Define the number of clusters based on provided K value
Step 2: Select random K points or centroids (can differ from the input dataset)
Step 3: Form the K clusters by assigning each data point to their closest centroid
Step 4:Calculate the variance and define a new centroid of each cluster
Step 5:Repeat the process from the third step to reassign each datapoint to the new closest centroid of each cluster until the algorithm finds the best possible solution

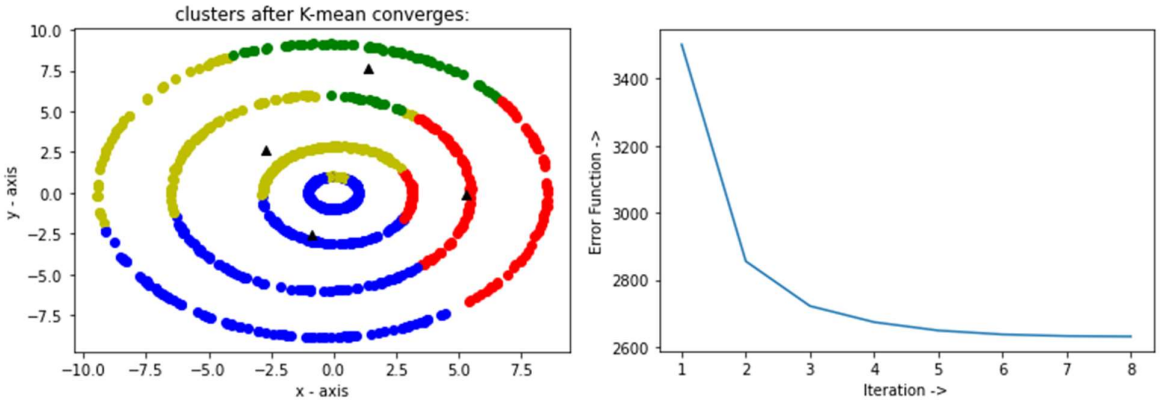
Input: Number of Clusters =4.

The output obtained are as follows –

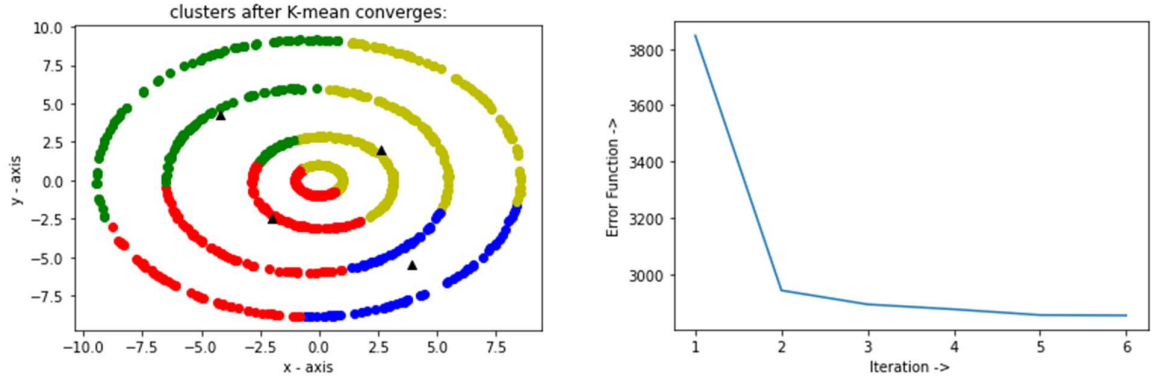
Initialization 1: Initial centroids: [array([1.6258, -5.6383]), array([-1.3719, -5.8996]), array([-1.7002, -2 .4755]), array([-0.93234, -8.8217])]



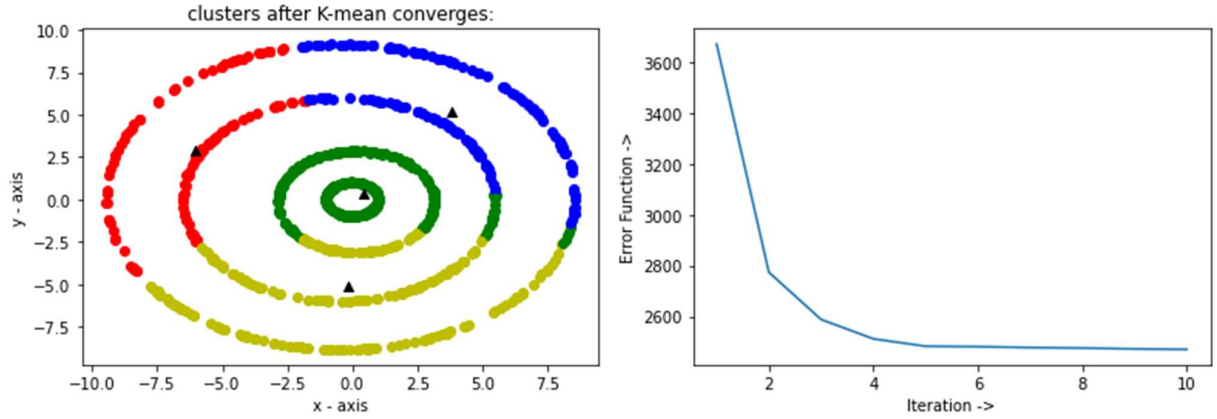
Intialization 2 : Initial centroids: [array([5.1696, 1.9047]), array([3.3005, 8.3096]), array([-0.096992, -0.97998]), array([0.59282, 2.8424])]



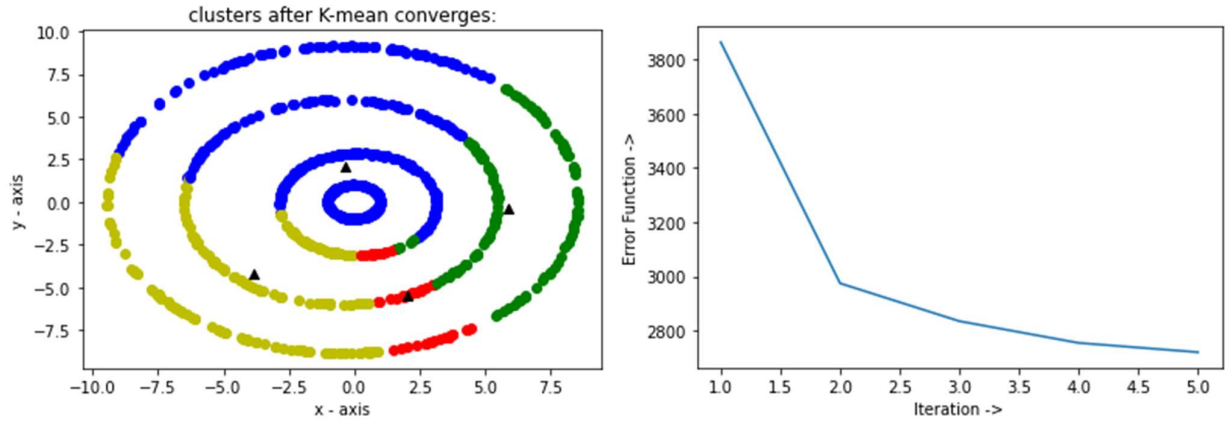
Initialization 3: Initial Centroids : [array([-0.029761, -3.107]), array([-8.4168, 4.4581]),
array([3.8245 -4.1391]), array([3.1157 , -0.42094])]



Initialization 4: Initial Centroids: [array([-4.5281, 8.1848]), array([-0.27349, -0.97976]),
array([-0.10709, 9.1726]), array([-0.029761, -3.107])]



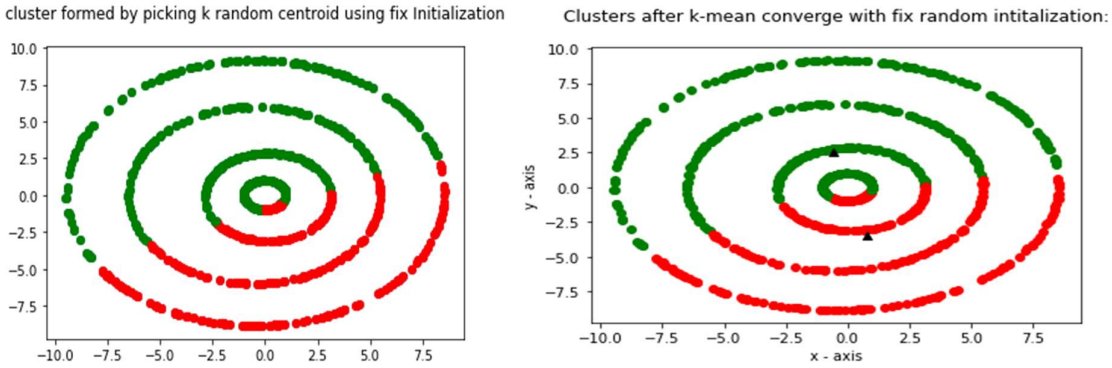
Initialization 5: Initial Centroids:[array([2.1372, -5.3692]), array([3.7787, -4.1209]),
array([0.98515, -0.12683]), array([-0.91776, -5.9734])]



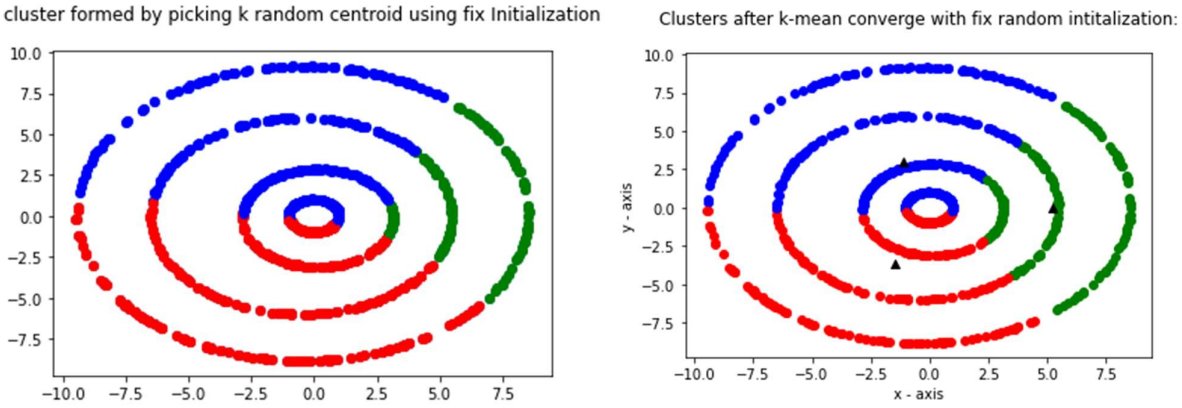
ii. Fix a random initialization. For $K = \{2, 3, 4, 5\}$, obtain cluster centers according to the K-means algorithm using the fixed initialization. For each value of K, plot the Voronoi regions associated with each cluster center. (You can assume the minimum and maximum value in the data set to be the range for each component of R^2).

Ans. The result obtained are as follows for k ={2,3,4,5}

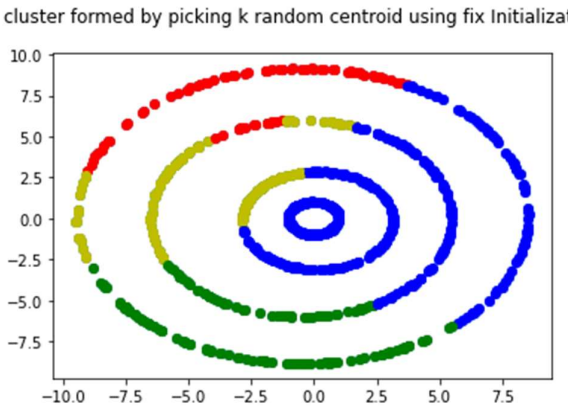
Input : Number of Clusters = 2

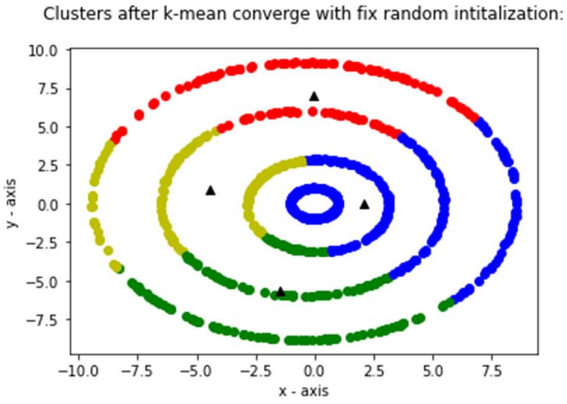


Input: Number of Clusters = 3

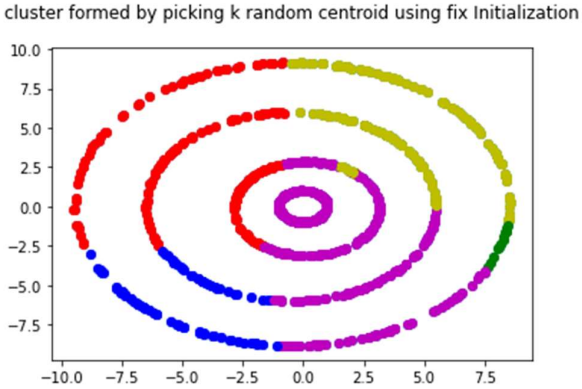


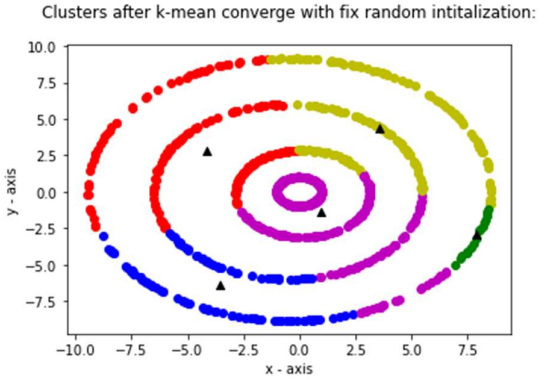
Input: Number of Clusters = 4





Input: Number of Clusters = 5



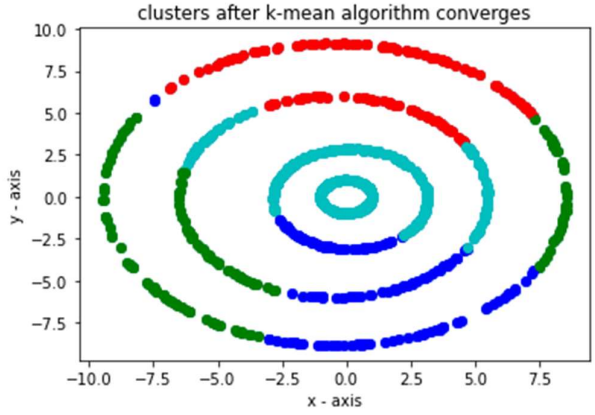


iii. Run the spectral clustering algorithm (spectral relaxation of K-means using Kernel-PCA) $k = 4$. Choose an appropriate kernel for this data set and plot the clusters obtained in different colors. Explain your choice of kernel based on the output you obtain.

Answer: Polynomial kernel with d=2 is better since it is almost forming a complete ring for inner cluster as well as the cluster just near it. Hence from visual analysis d = 2 is better.

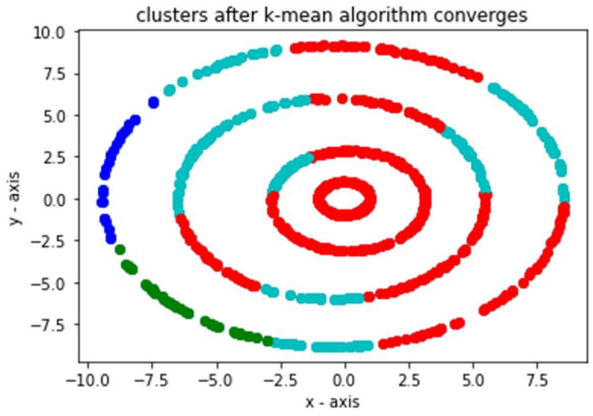
The output obtained using a polynomial kernel of degree= 2
Input:

```
Type of kernel function(kernel or gaussian):kernel
Number of Clusters:4
Degree of polynomial:2
```



Input:

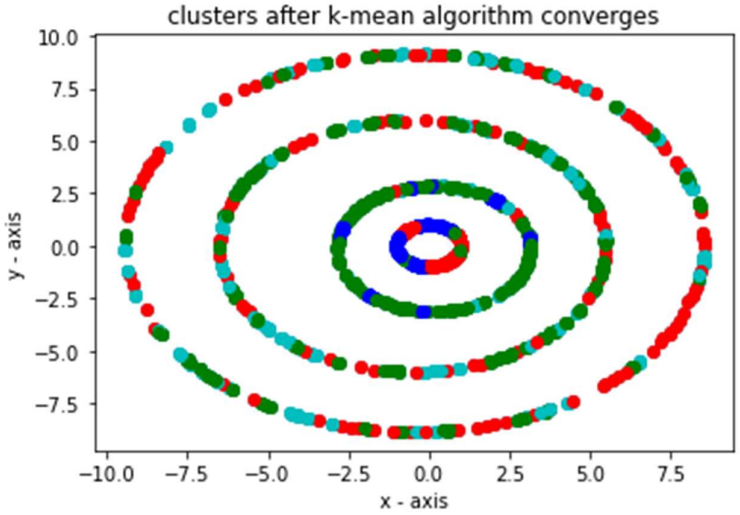
```
Type of kernel function(kernel or gaussian):kernel
Number of Clusters:4
Degree of polynomial:3
```



The output obtained using gaussian kernel for sigma = 0.1

Input:

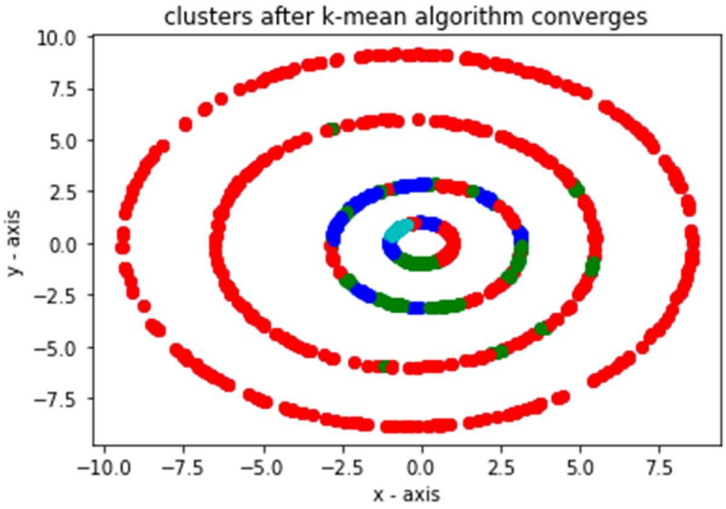
```
Type of kernel function(kernel or gaussian):gaussian
Number of Clusters:4
input value of sigma:0.1
```



The output obtained using gaussian kernel for sigma = 0.2

Input:

```
Type of kernel function(kernel or gaussian):gaussian
Number of Clusters:4
input value of sigma:0.2
```



iv. Instead of using the method suggested by spectral clustering to map eigenvectors to cluster assignments, use the following method: Assign data point i to cluster f whenever

$$f = \arg\max_{j=1,\ldots,k} i$$

where $vj \in Rn$ is the eigenvector of the Kernel matrix associated with the j-th largest eigenvalue

. How does this mapping perform for this dataset?. Explain your insights.

Answer. K- mean algorithm is better than this approach because the original k-mean improves after each iteration until it converges. In this approach, we only perform one iteration which is not suitable.

Input:

```
Type of kernel function(kernel or gaussian):kernel
Number of Componenets/clusters:4
Degree of polynomial:2
```