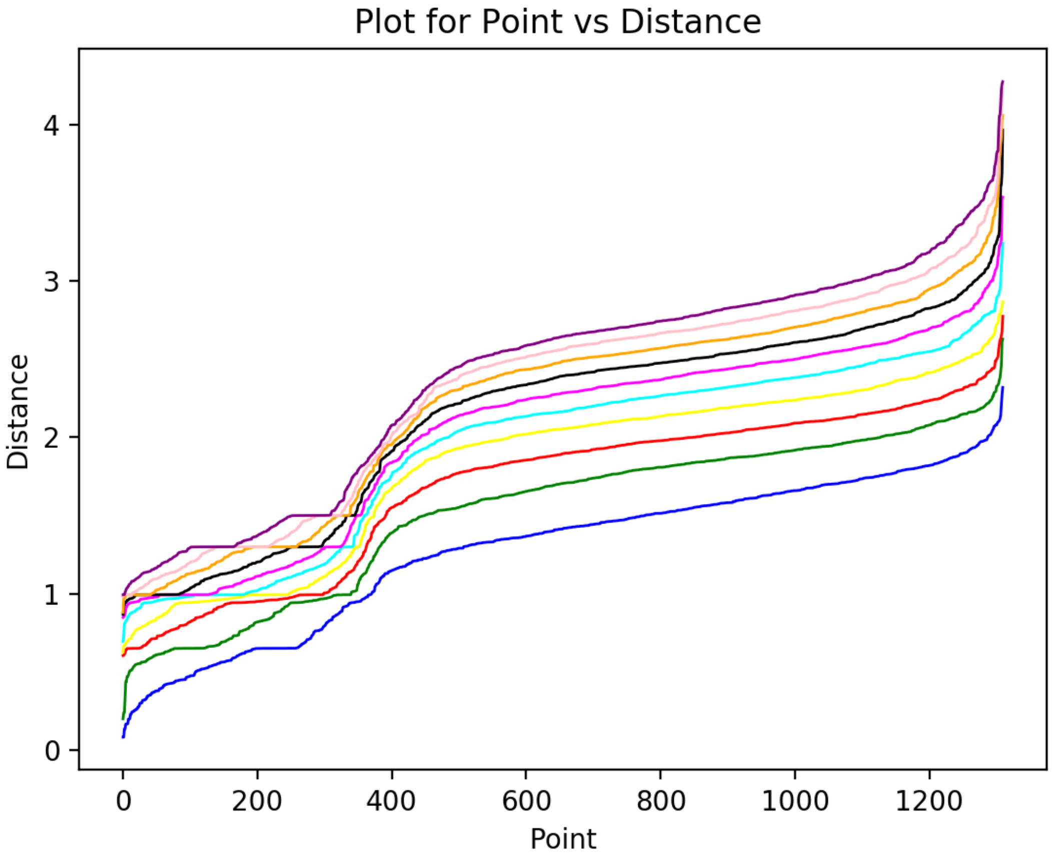
Pratishta Rao, Srikant Lakshminarayan

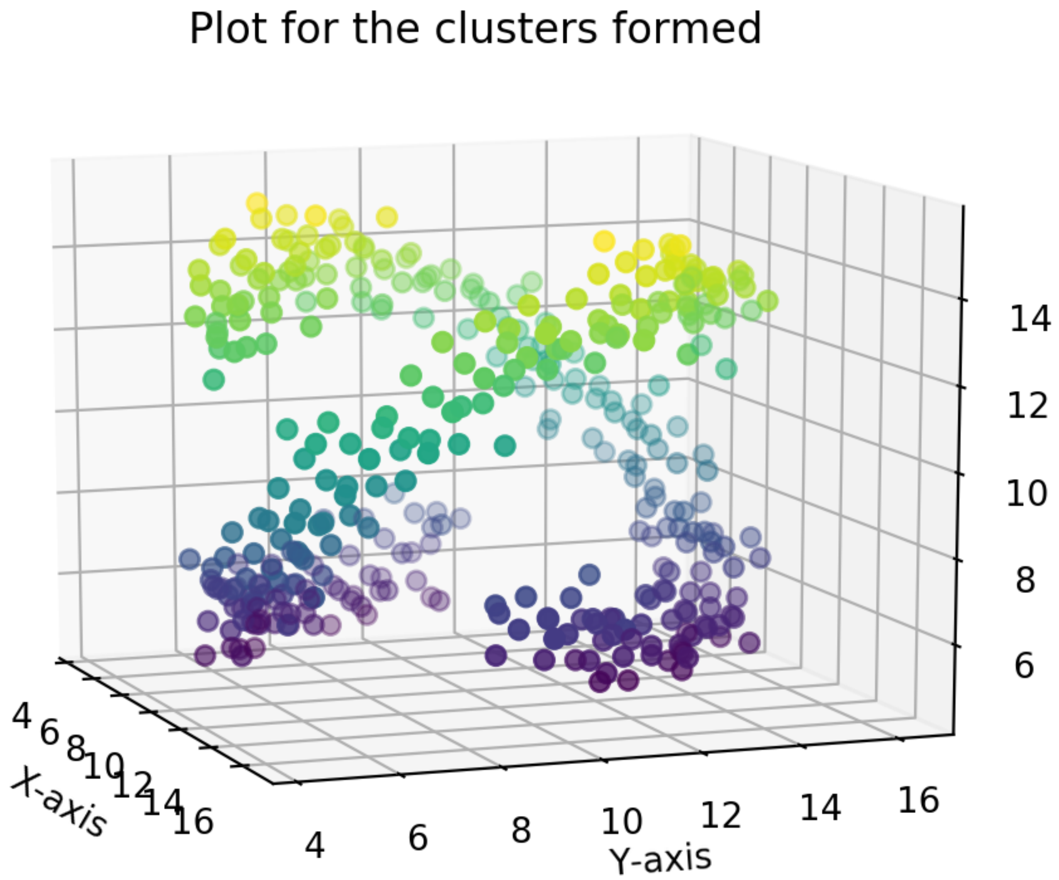
Big Data Analytics - Home Work 10

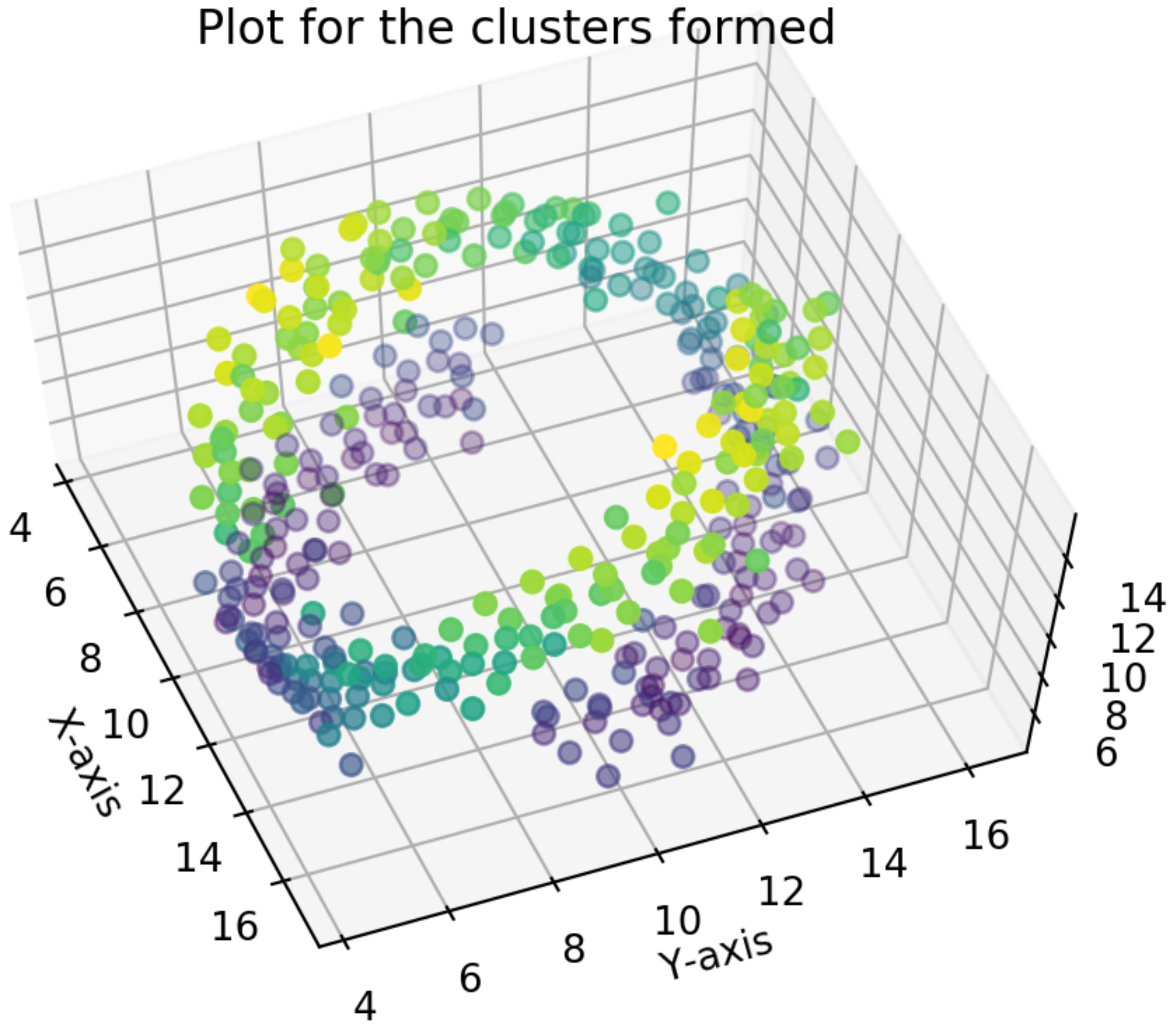
1. Plot obtained is as follows:

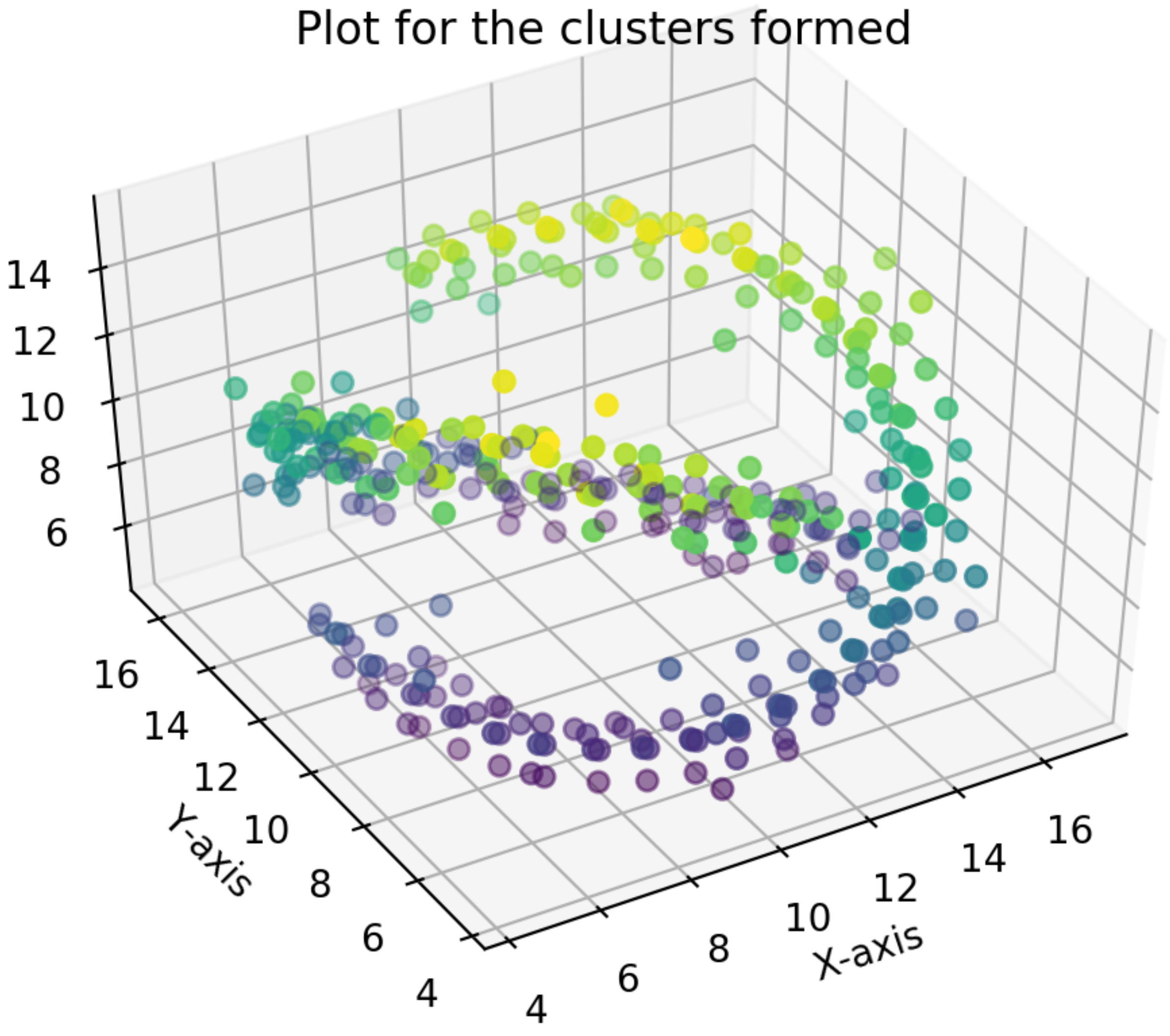


2. Generate a graph to show the resulting clusters.

<u>Solution:</u>




Plot for the clusters formed
X-axis
4
6
8
10
12
14
16
Y-axis
4
6
8
10
12
14
16
6
8
10
12
14


Plot for the clusters formed
14
12
10
8
6
16
14
12
10
8
6
4
Y-axis
4
6
8
10
12
14
16
X-axis

3. Describe what parameter value you decided to use for eps.

Solution:

From the graph above, the knee point is just above 1.3.

4. What distance metric did you use? However, in the real world you would use something that described the distance between people, or data records or events especially for social network analysis.

Solution:

For the distance metric euclidean distance is used. For the real world we would ideally use either Manhattan distance or Haversine’s

5. Using DBscan, identify the number of clusters. How many clusters did you identify? What evidence do you have to support this?

Solution:

Using DBscan we identified two clusters. The graph displaying the clusters obtained after using DBscan is shown in question 2.

6. Sort the clusters from smallest to largest:

a. How many data points were in it?

b. What was the center of mass of this cluster?

Solution:

for epsilon=1.3
minpoints=8

Clusters	Data points	Center of mass
Cluster 1	190	(9.1, 11.3, 9.82)
Cluster 2	192	(11.28, 9.03, 9.72)

7. Also, report an estimate of the number of noise points overall in the data set.

Solution:

The number of noise points overall in the data set is 928

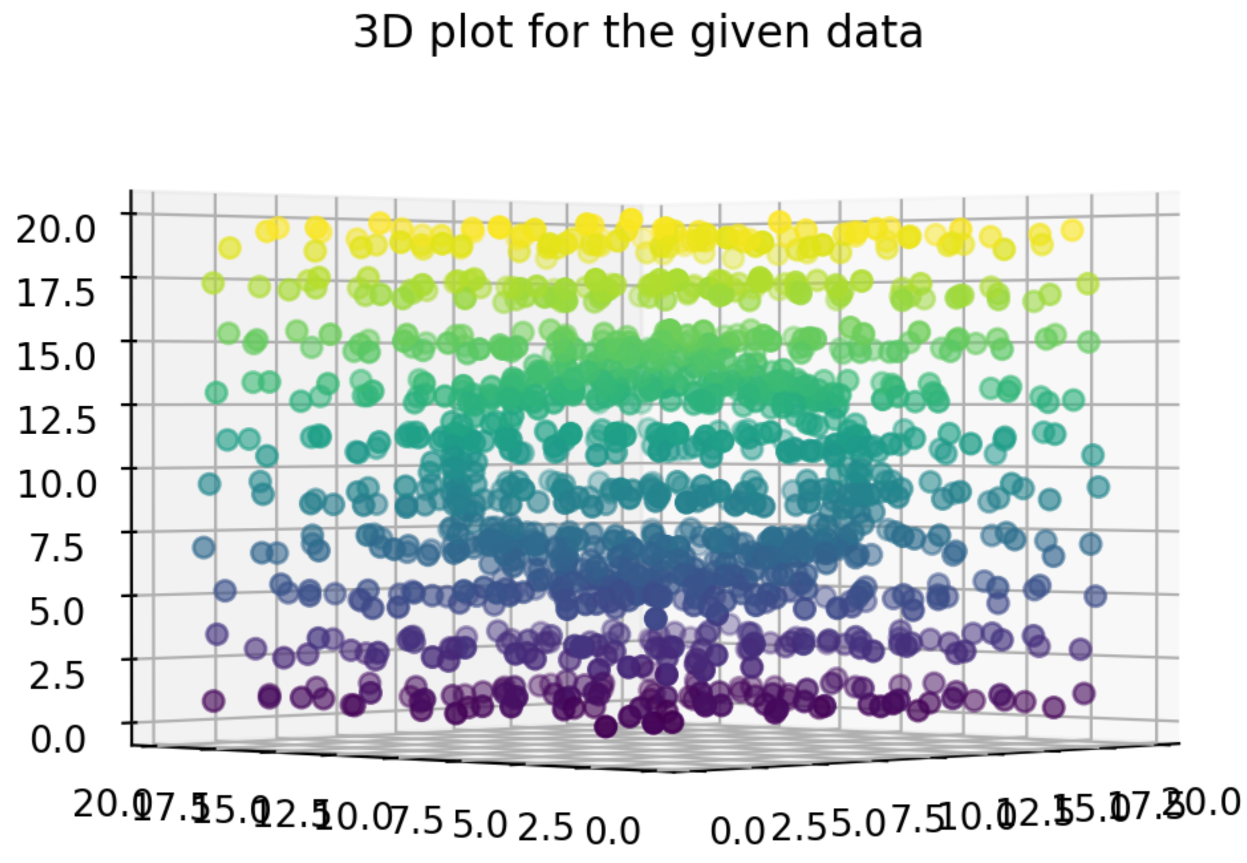
8. Describe what you learned from this exercise, how hard it was for you to implement this, compare and contrast this to other algorithms you have implemented was it harder or longer show any graphs or data visualization you would like to show, provide evidence of learning.

Solution:

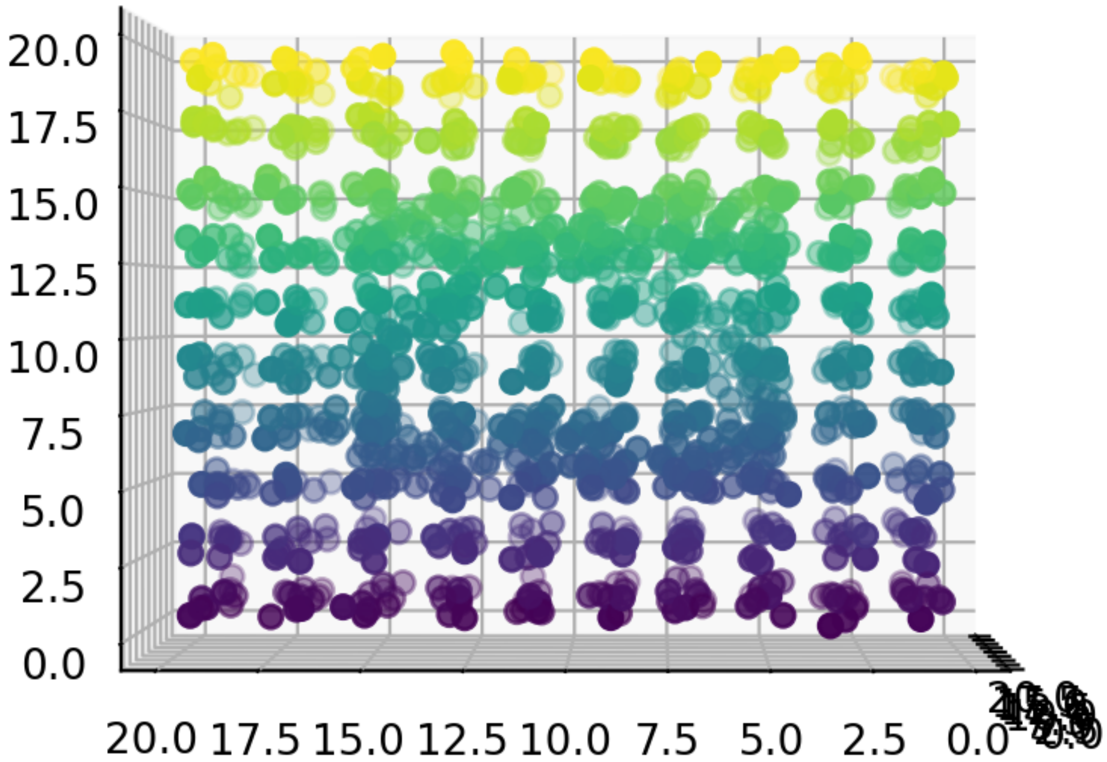
We learnt the implementation DBscan and how it can be used in outlier detection. Because of the visualization that we did, we understood it better.

The harder part was finding the epsilon and min points by plotting the graphs. All the distances were computed and stored which was then sorted and plotted where the inflection or knee point was observed. The DB-scan algorithm was also tried with different changes in values of epsilon and min points to which it resulted in two clusters. DB-scan was easier than k-means and is better as it finds non-globular clusters

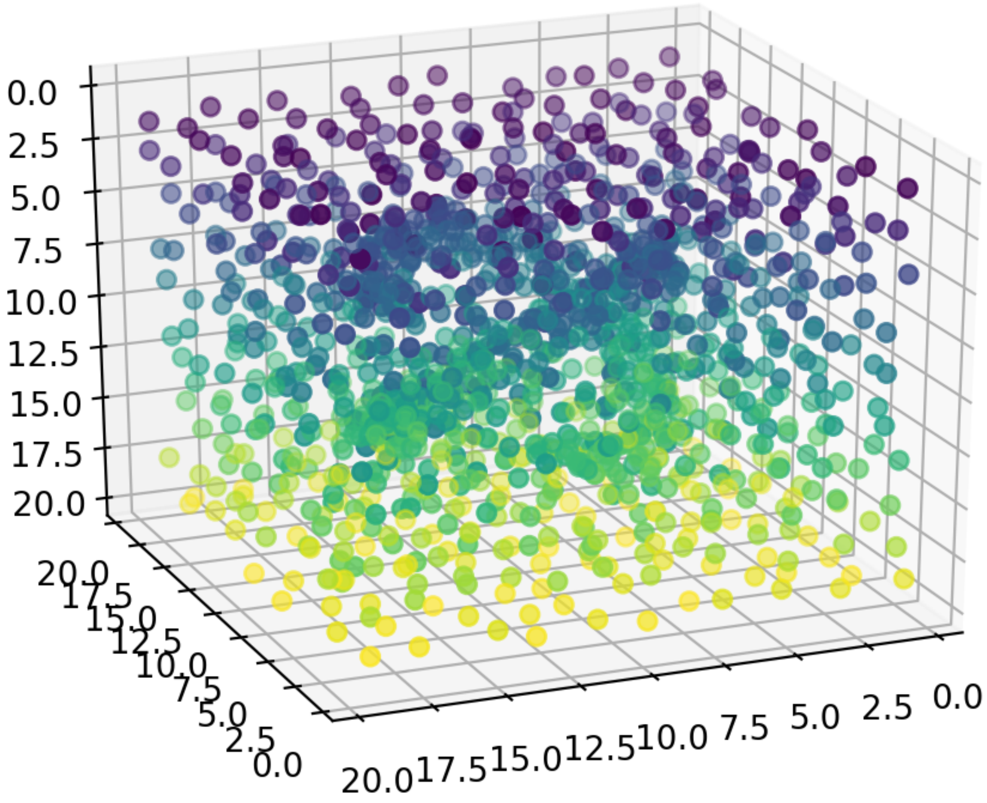
The following are 3D and 2D plots before the application of DBscan on these points:

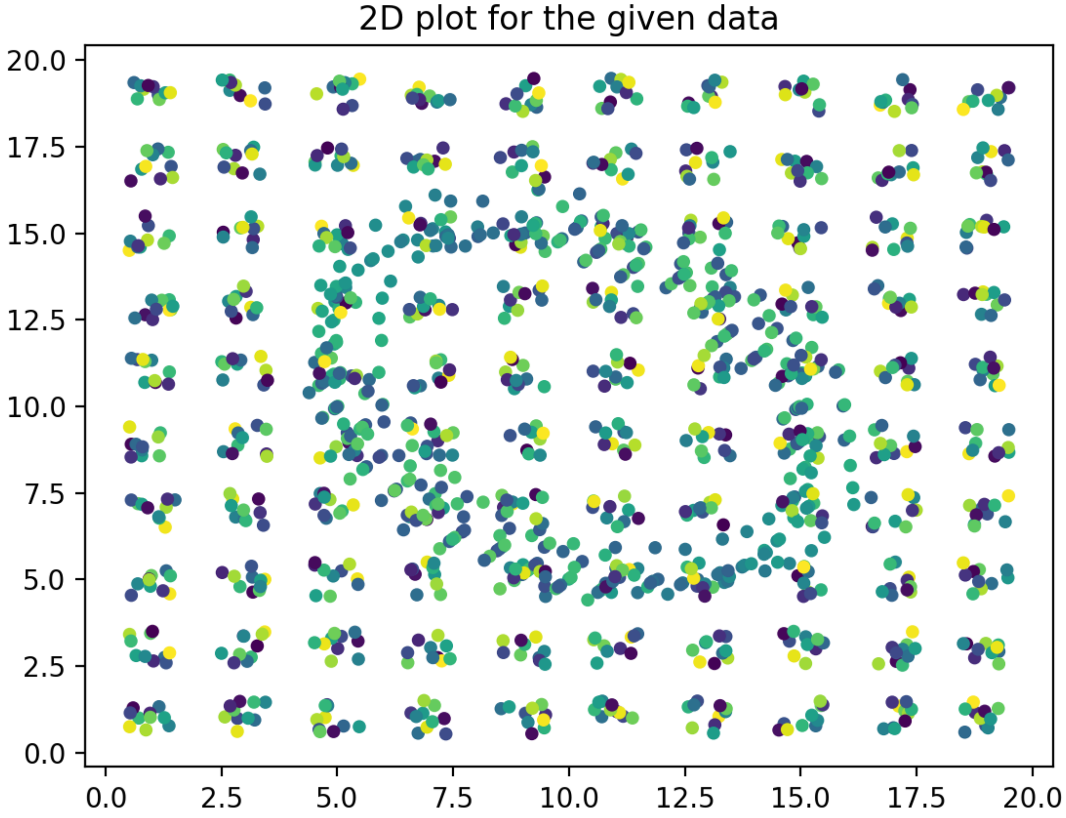

3D plot for the given data

3D plot for the given data



3D plot for the given data




2D plot for the given data
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0