CS 584-04: Machine Learning

Fall 2018 Assignment 2

Suhas Sreenivas (A20423132)

Question 1 (20 points)

Suppose a market basket can possibly contain these seven items: A, B, C, D, E, F, and G.

a) (1 point) What is the number of possible itemsets?
127

b) (3 points) List all the possible 1-itemsets.
('A')
('B')
('C')
('D')
('E')
('F')
('G')

c) (3 points) List all the possible 2-itemsets.
('A', 'B')
('A', 'C')
('A', 'D')
('A', 'E')
('A', 'F')
('A', 'G')
('B', 'C')
('B', 'D')
('B', 'E')
('B', 'F')
('B', 'G')
('C', 'D')
('C', 'E')
('C', 'F')
('C', 'G')
('D', 'E')
('D', 'F')
('D', 'G')
('E', 'F')
('E', 'G')
('F', 'G')

d) (3 points) List all the possible 3-itemsets.

('A', 'B', 'C')
('A', 'B', 'D')
('A', 'B', 'E')
('A', 'B', 'F')
('A', 'B', 'G')
('A', 'C', 'D')
('A', 'C', 'E')
('A', 'C', 'F')
('A', 'C', 'G')
('A', 'D', 'E')
('A', 'D', 'F')
('A', 'D', 'G')
('A', 'E', 'F')
('A', 'E', 'G')
('A', 'F', 'G')
('B', 'C', 'D')
('B', 'C', 'E')
('B', 'C', 'F')
('B', 'C', 'G')
('B', 'D', 'E')
('B', 'D', 'F')
('B', 'D', 'G')
('B', 'E', 'F')
('B', 'E', 'G')
('B', 'F', 'G')
('C', 'D', 'E')
('C', 'D', 'F')
('C', 'D', 'G')
('C', 'E', 'F')
('C', 'E', 'G')
('C', 'F', 'G')
('D', 'E', 'F')
('D', 'E', 'G')
('D', 'F', 'G')
('E', 'F', 'G')

e) (3 points) List all the possible 4-itemsets.

('A', 'B', 'C', 'D')
('A', 'B', 'C', 'E')

```
('A', 'B', 'C', 'F')
('A', 'B', 'C', 'G')
('A', 'B', 'D', 'E')
('A', 'B', 'D', 'F')
('A', 'B', 'D', 'G')
('A', 'B', 'E', 'F')
('A', 'B', 'E', 'G')
('A', 'B', 'F', 'G')
('A', 'C', 'D', 'E')
('A', 'C', 'D', 'F')
('A', 'C', 'D', 'G')
('A', 'C', 'E', 'F')
('A', 'C', 'E', 'G')
('A', 'C', 'F', 'G')
('A', 'D', 'E', 'F')
('A', 'D', 'E', 'G')
('A', 'D', 'F', 'G')
('A', 'E', 'F', 'G')
('B', 'C', 'D', 'E')
('B', 'C', 'D', 'F')
('B', 'C', 'D', 'G')
('B', 'C', 'E', 'F')
('B', 'C', 'E', 'G')
('B', 'C', 'F', 'G')
('B', 'D', 'E', 'F')
('B', 'D', 'E', 'G')
('B', 'D', 'F', 'G')
('B', 'E', 'F', 'G')
('C', 'D', 'E', 'F')
('C', 'D', 'E', 'G')
('C', 'D', 'F', 'G')
('C', 'E', 'F', 'G')
('D', 'E', 'F', 'G')
```

f) (3 points) List all the possible 5-itemsets.

```
('A', 'B', 'C', 'D', 'E')
('A', 'B', 'C', 'D', 'F')
('A', 'B', 'C', 'D', 'G')
('A', 'B', 'C', 'E', 'F')
('A', 'B', 'C', 'E', 'G')
('A', 'B', 'C', 'F', 'G')
('A', 'B', 'D', 'E', 'F')
('A', 'B', 'D', 'E', 'G')
```

('A', 'B', 'D', 'F', 'G')
('A', 'B', 'E', 'F', 'G')
('A', 'C', 'D', 'E', 'F')
('A', 'C', 'D', 'E', 'G')
('A', 'C', 'D', 'F', 'G')
('A', 'C', 'E', 'F', 'G')
('A', 'D', 'E', 'F', 'G')
('B', 'C', 'D', 'E', 'F')
('B', 'C', 'D', 'E', 'G')
('B', 'C', 'D', 'F', 'G')
('B', 'C', 'E', 'F', 'G')
('B', 'D', 'E', 'F', 'G')
('C', 'D', 'E', 'F', 'G')

g) (3 points) List all the possible 6-itemsets.
('A', 'B', 'C', 'D', 'E', 'F')
('A', 'B', 'C', 'D', 'E', 'G')
('A', 'B', 'C', 'D', 'F', 'G')
('A', 'B', 'C', 'E', 'F', 'G')
('A', 'B', 'D', 'E', 'F', 'G')
('A', 'C', 'D', 'E', 'F', 'G')
('B', 'C', 'D', 'E', 'F', 'G')

h) (1 point) List all the possible 7-itemsets.
('A', 'B', 'C', 'D', 'E', 'F', 'G')

Question 2 (30 points)

The file Groceries.csv contains market basket data. The variables are:

1. Customer: Customer Identifier
2. Item: Name of Product Purchased

The data is already sorted in ascending order by Customer and then by Item. Also, all the items bought by each customer are all distinct.

After you have imported the CSV file, please discover association rules using this dataset.

a) (2 points) How many customers in this market basket data?

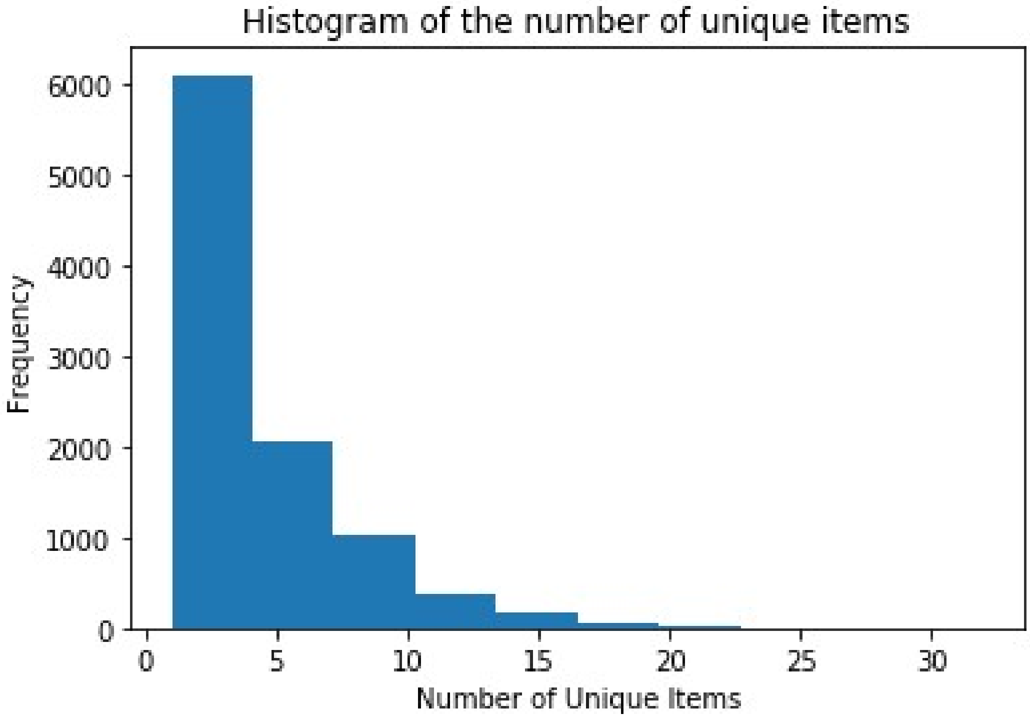
ANS: 9835 customers are in this market basket data.

b) (2 points) How many unique items in the market basket across all customers?

ANS: There are 169 unique items in the market basket across all customers.

c) (5 points) Create a dataset which contains the number of distinct items in each customer's market basket. Draw a histogram of the number of unique items. What are the median, the 25th percentile and the 75th percentile in this histogram?

ANS:



25th percentile: 2.0
Median: 3.0
75th percentile: 6.0

d) (5 points) Find out the *k*-itemsets which appeared in the market baskets of at least seventy five (75) customers. How many itemsets have you found? Also, what is the highest *k* value in your itemsets?

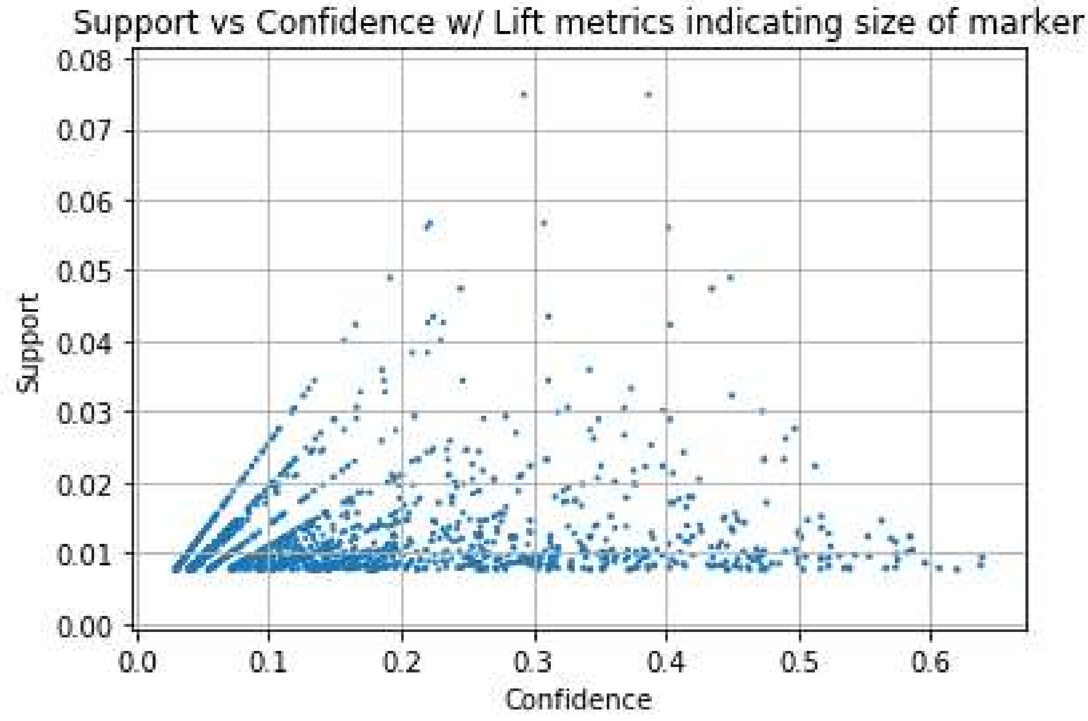
ANS: I have found 524 itemsets using minimum support as 0.0076258261311. The highest k value in the itemsets is 4.

e) (5 points) Find out the association rules whose Confidence metrics are at least 1%. How many association rules have you found? Please be reminded that a rule must have a non-empty antecedent and a non-empty consequent.

ANS: I have found 1228 association rules whose confidence metrics are at least 1%.

f) (5 points) Graph the Support metrics on the vertical axis against the Confidence metrics on the horizontal axis for the rules you found in (e). Please use the Lift metrics to indicate the size of the marker.

ANS:



g) (5 points) List the rules whose Confidence metrics are at least 60%. Please include their Support and Lift metrics.

ANS:

Antecedents	Consequents	Support	Lift
{'root vegetables', 'butter'}	{'whole milk'}	0.008235892221657347	2.4961068585089814
{'butter', 'yogurt'}	{'whole milk'}	0.009354346720894764	2.500386877127824
{'root vegetables', 'other vegetables', 'yogurt'}	{'whole milk'}	0.007829181494661922	2.3728423222863158
{'other vegetables', 'tropical fruit', 'yogurt'}	{'whole milk'}	0.007625826131164209	2.4258155114068

h) (1 point) What similarities do you find among the consequents that appeared in (g)?

ANS: All the associations rules that appeared in (g) have ‘whole milk’ as the consequent.

Question 3 (20 points)

You are asked to write a Python program to calculate the Elbow value and the Silhouette value. For this question, you will use the CARS.CSV dataset to test your program. Here are the specifications for performing the respective analyses.

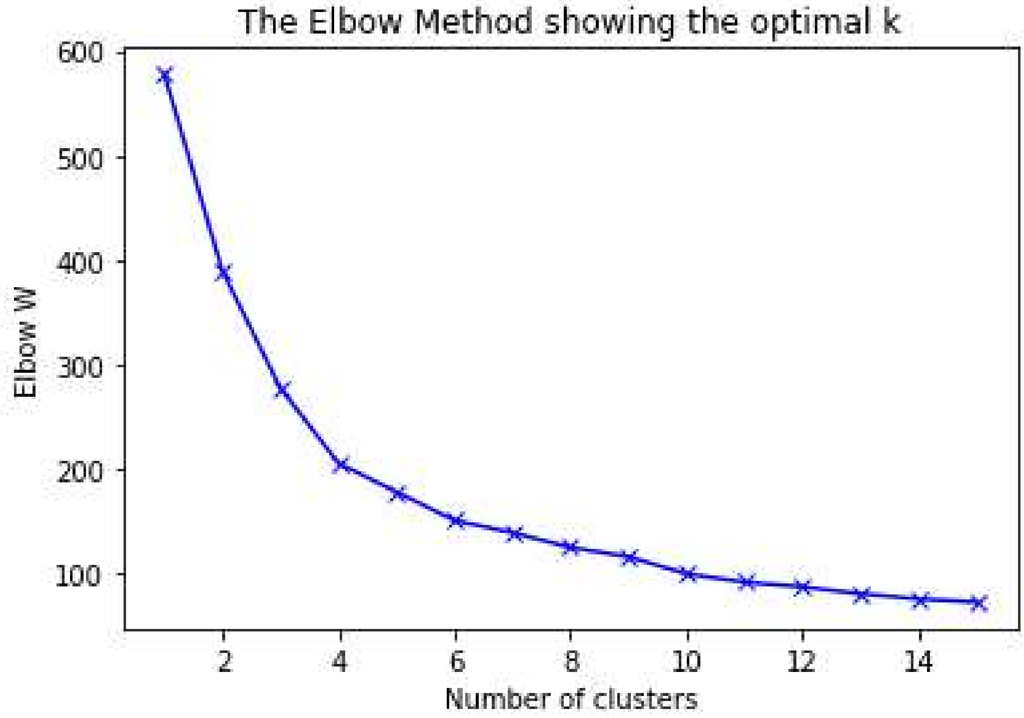
Clustering

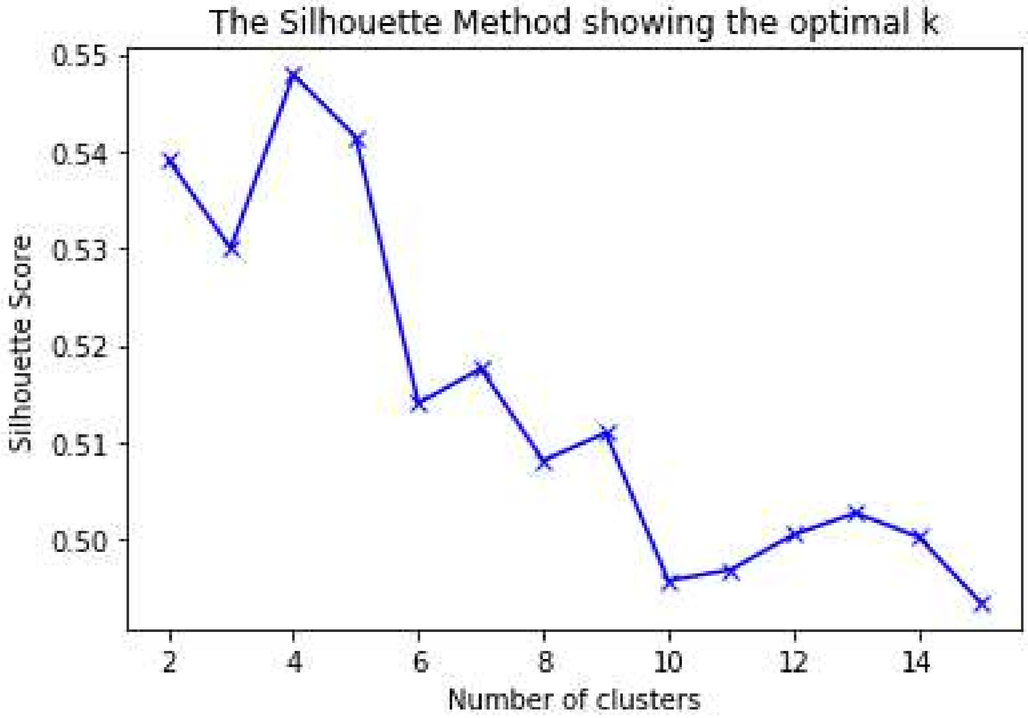
- The input interval variables are Horsepower and Weight
- The distance metric is Euclidean
- The maximum number of clusters is 15
- Consider the `silhouette_score` function for calculating the Silhouette value.

Please answer the following questions.

a) (15 points) List the Elbow values and the Silhouette values for your 1-cluster to 15-cluster solutions.

ANS:





```
Elbow values:
578.0051999326791
389.6723458473717
277.43171444086204
205.72734275832073
178.03905753772636
150.9631083711226
138.7572922450605
124.8642378076019
115.24546424393026
101.12009330650325
89.89871786415668
87.24115284810297
82.51092928267919
75.62114859897046
72.48232258923436


Silhouette values:
0.5391245600025193
0.5299743944459789
0.5478843834241527
0.5414715519866451
0.5143405374281904
0.5172293099648427
```

```
0.5120431409735351
0.5098045330963802
0.49704708282911375
0.499451872394465
0.5008049855056597
0.4952718099335287
0.49871286219973787
0.49284972698730944
```

b) (5 points) Based on the Elbow values and the Silhouette values, what do you suggest for the number of clusters?

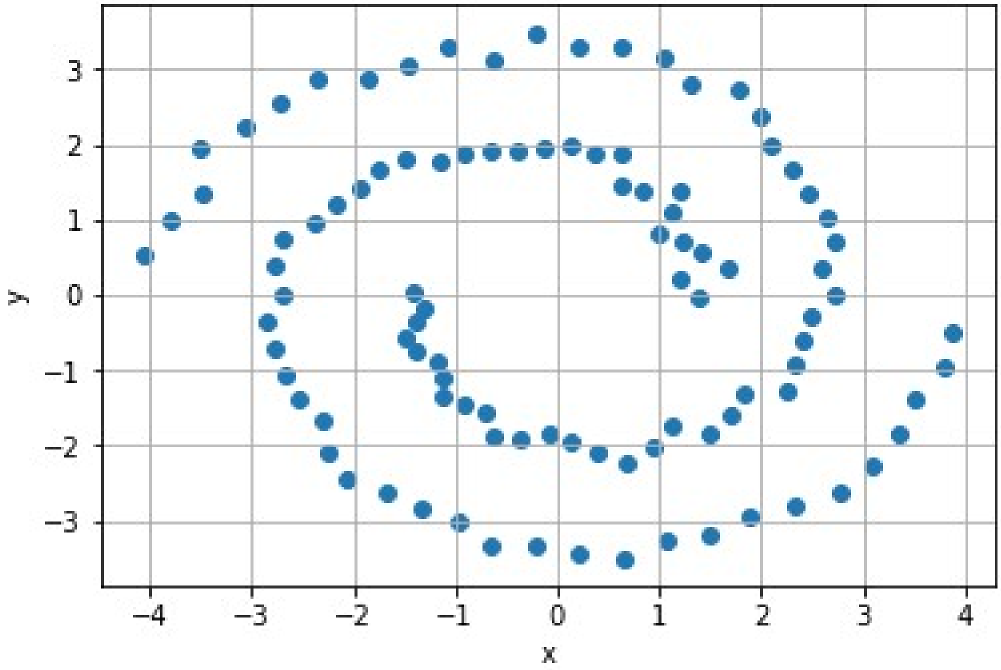
Based on the Elbow values, 3 or 4 clusters would seem appropriate. However, when we look at the Silhouette values, the graph peaks at 4 and hence, I would suggest 4 for the number of clusters.

Question 4 (30 points)

Apply the Spectral Clustering method to the Spiral.csv. Your input fields are x and y.

a) (5 points) Generate a scatterplot of y (vertical axis) versus x (horizontal axis). How many clusters will you say by visual inspection?

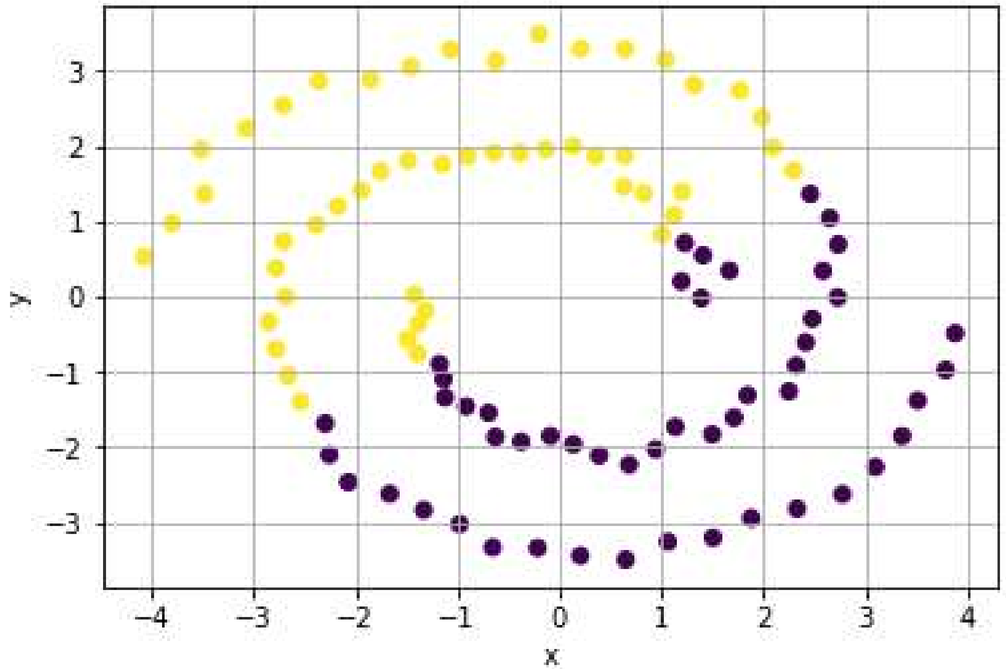
ANS:



By visual inspection, there are 2 clusters.

b) (5 points) Apply the K-mean algorithm directly using your number of clusters (in a). Regenerate the scatterplot using the K-mean cluster identifier to control the color scheme?

ANS:

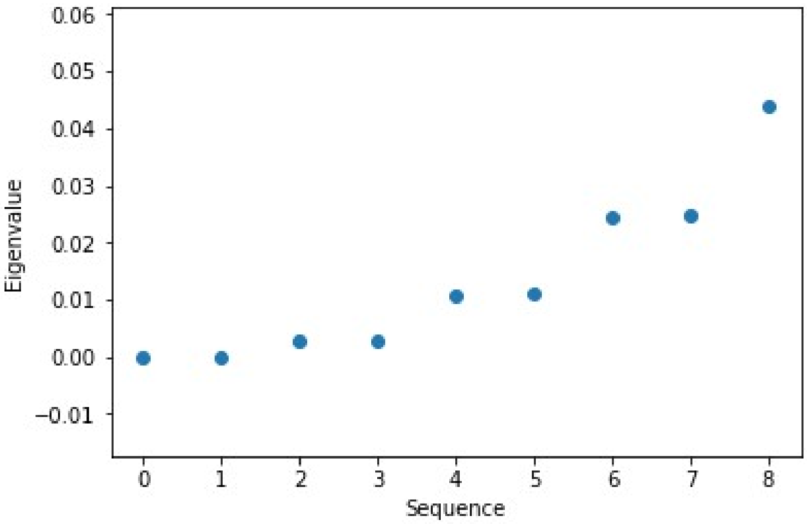


c) (5 points) Apply the nearest neighbor algorithm using the Euclidean distance. How many nearest neighbors will you use?

ANS: For the nearest neighbor algorithm, I will use 3 neighbors.

d) (5 points) Generate the sequence plot of the first nine eigenvalues, starting from the smallest eigenvalues. Based on this graph, do you think your number of nearest neighbors (in a) is appropriate?

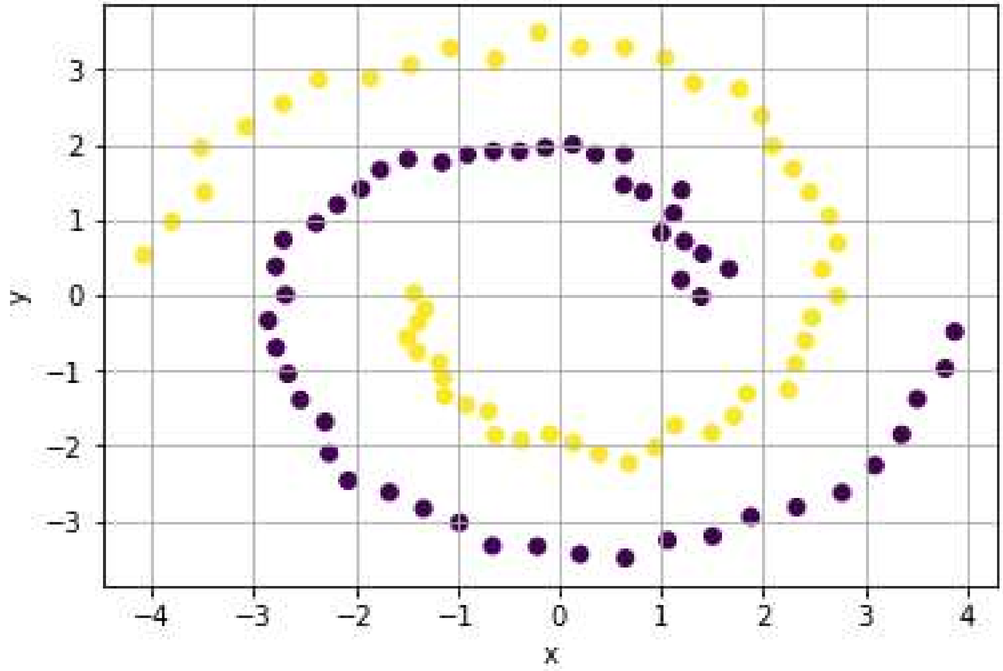
ANS:



No, because the graph suggests 3 neighbors, based on the jump in values from the graph. As we can see from further results, using 3 neighbors gives us the correct solution.

e) (5 points) Apply the K-mean algorithm on your first two eigenvectors that correspond to the first two smallest eigenvalues. Regenerate the scatterplot using the K-mean cluster identifier to control the color scheme?

ANS:



f) (5 points) Comment on your spectral clustering results?
It looks like the spectral clustering method works for this dataset and is appropriate because even though the data is connected, it is not compact. It separates the data into 2 parts in the same way we split it through visual inspection.