Assignment2

name: Yuqi Zhou
ID: A20423555

Question 1:
(a). The number of possible itemsets: 127

(b). All the possible 1-itemsets:
[('A'), ('B'), ('C'), ('D'), ('E'), ('F'), (‘G’)]

(c). 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). 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). 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). 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). 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')]

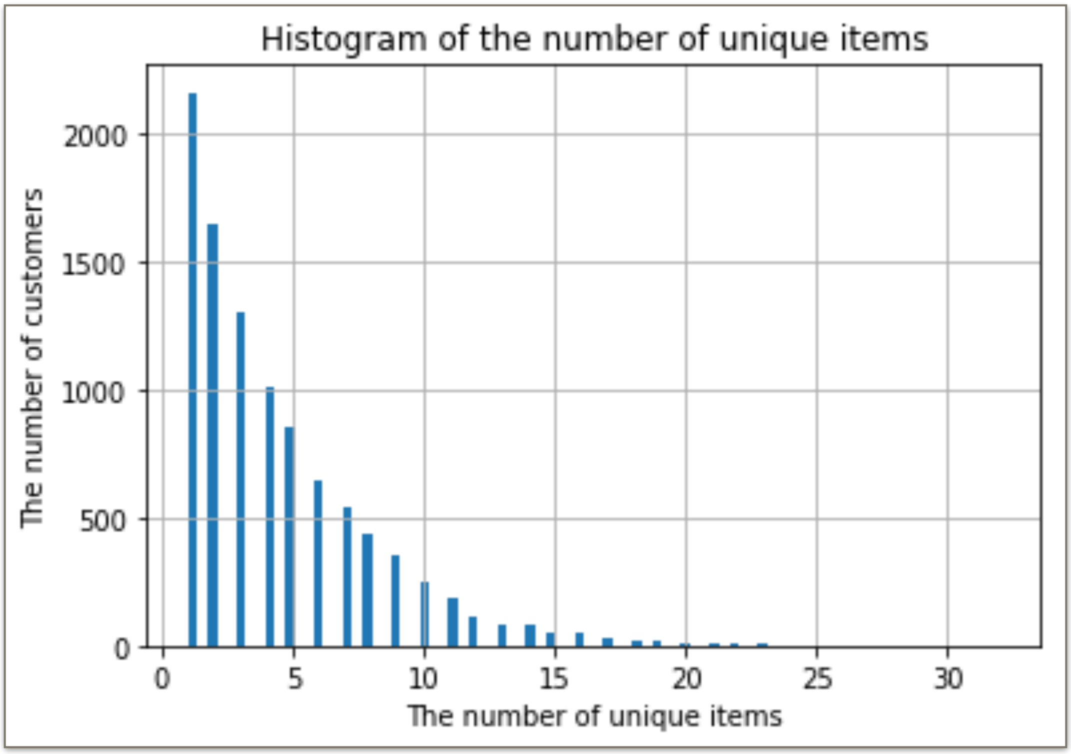
(h). All the possible 7-itemsets:
[('A', 'B', 'C', 'D', 'E', 'F', ‘G')]

Question 2:

(a). The number of customers in this market basket data: 9835

(b). The number of unique items in the market basket data: 169

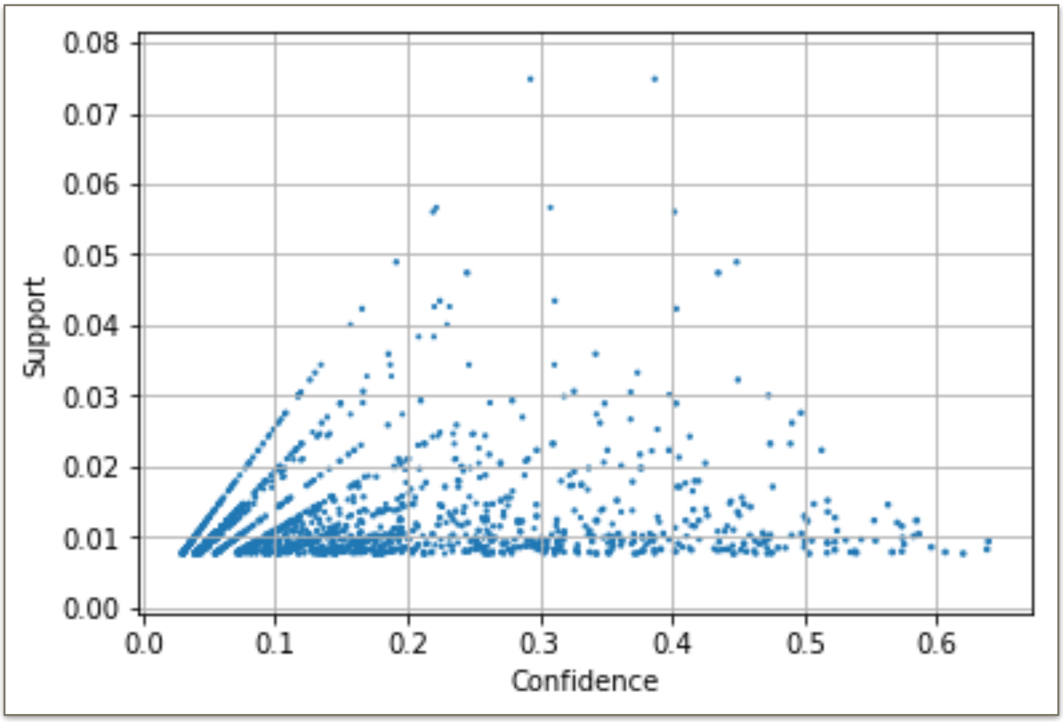
(c). The histogram of the number of unique items:



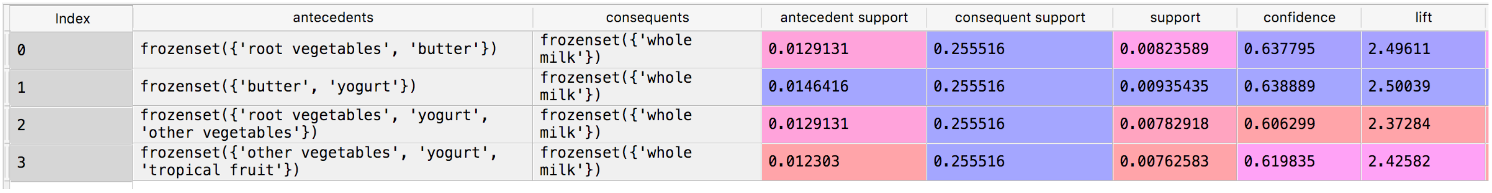
Median: 3.0
The 25th percentile: 2.0
The 75th percentile: 6.0

(d). The number of item sets have been found: 524
The highest k value: 4

(e). The number of association rules have found: 1228 (f). The graph:



(g). The rules whose Confidence metrics are at least 60%.



Index	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
0	frozenset({'root vegetables', 'butter'})	frozenset({'whole milk'})	0.0129131	0.255516	0.00823589	0.637795	2.49611
1	frozenset({'butter', 'yogurt'})	frozenset({'whole milk'})	0.0146416	0.255516	0.00935435	0.638889	2.50039
2	frozenset({'root vegetables', 'yogurt', 'other vegetables'})	frozenset({'whole milk'})	0.0129131	0.255516	0.00782918	0.606299	2.37284
3	frozenset({'other vegetables', 'yogurt', 'tropical fruit'})	frozenset({'whole milk'})	0.012303	0.255516	0.00762583	0.619835	2.42582

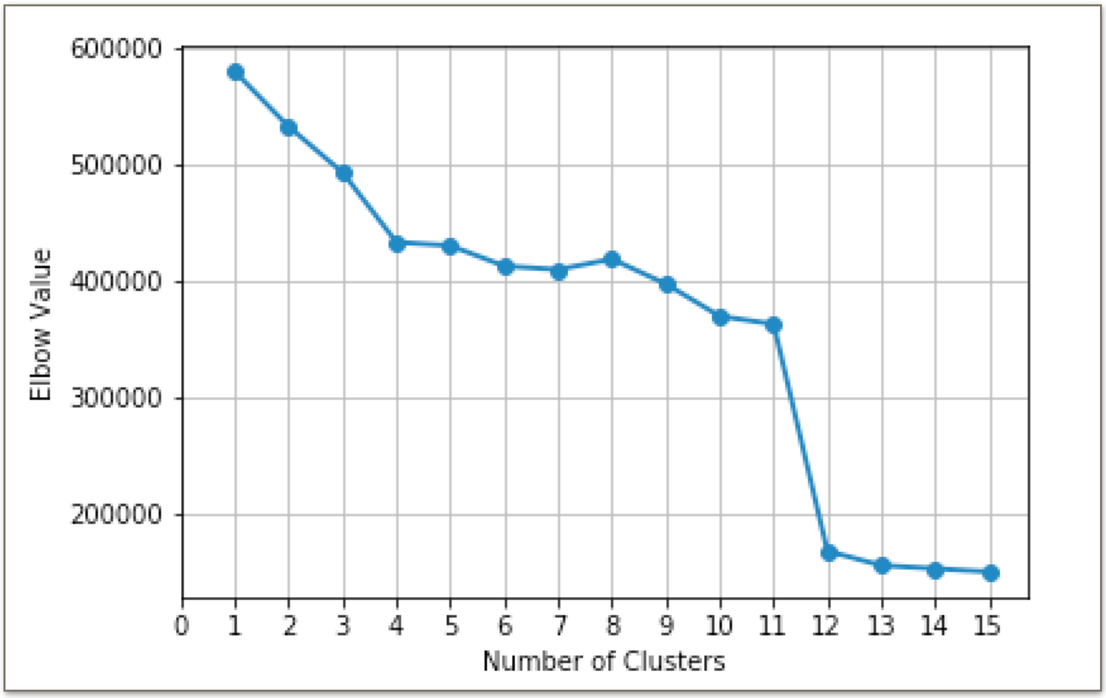
(h). All the consequents that appeared in (g) are {‘whole milk’}

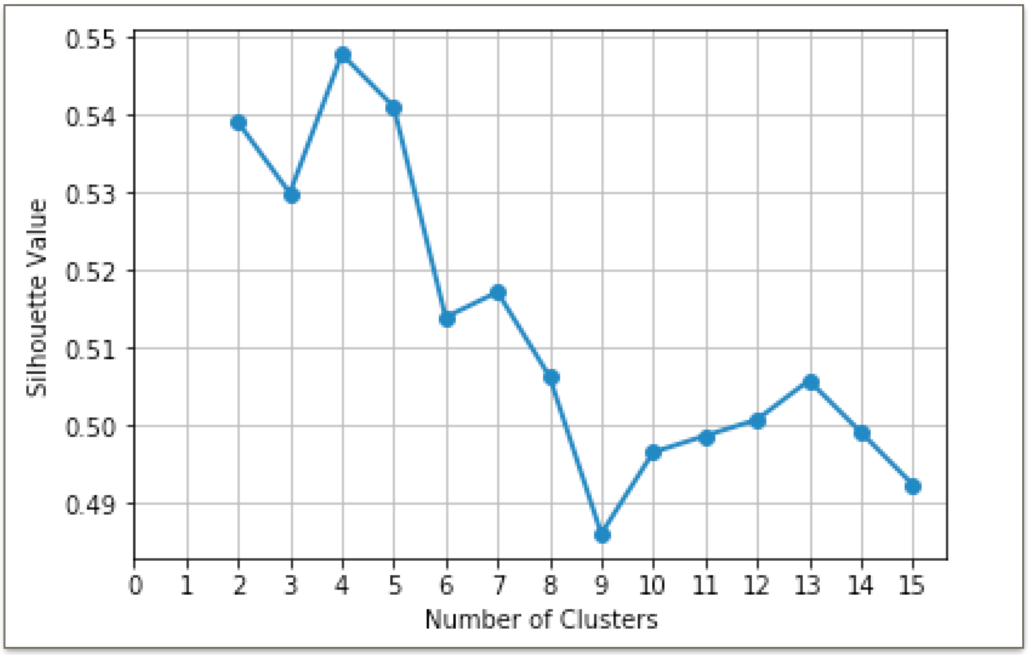
Question 3:

(a). The elbow values and the Silhouette values(for 1-cluster to 15-cluster solution):

N Clusters	Elbow Value	Silhouette Value:
1	579857.9543	nan
2	532455.2722	0.5391
3	493218.0813	0.5300
4	433215.8150	0.5479
5	430290.4574	0.5411
6	412804.9312	0.5140
7	409729.7423	0.5172
8	418744.2477	0.5064
9	397493.5317	0.4861
10	369702.7050	0.4966
11	362959.0026	0.4987
12	168058.0920	0.5008
13	155749.4156	0.5059
14	153006.5541	0.4992
15	150220.8996	0.4925

(b).

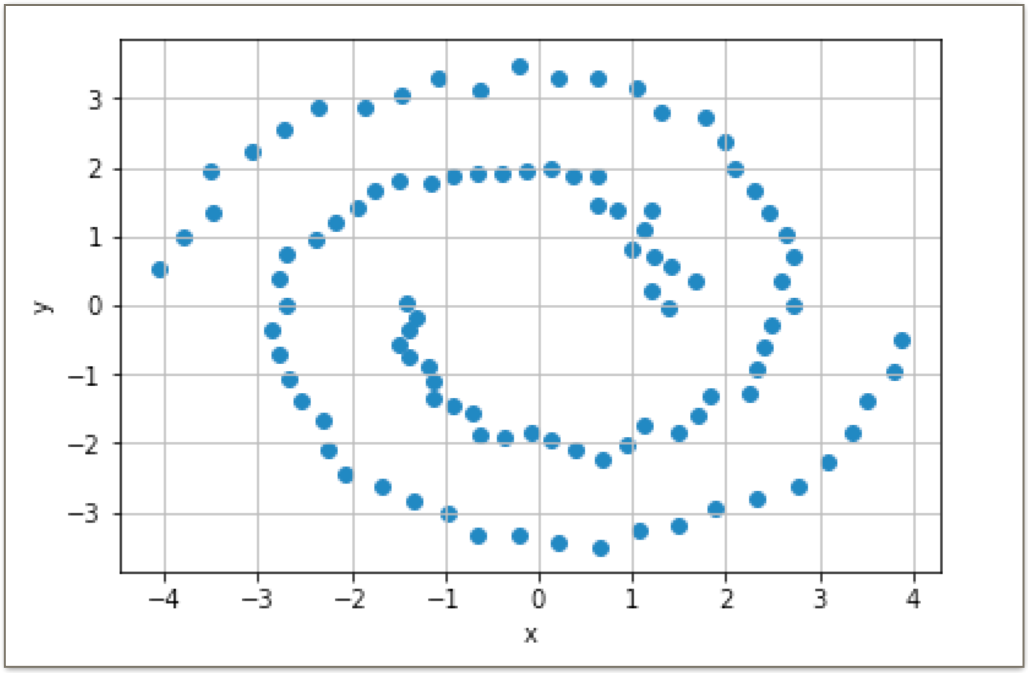




Based on the Elbow values, the Silhouette values and biggest acceleration value(182592.2342) suggest number of clusters is 13

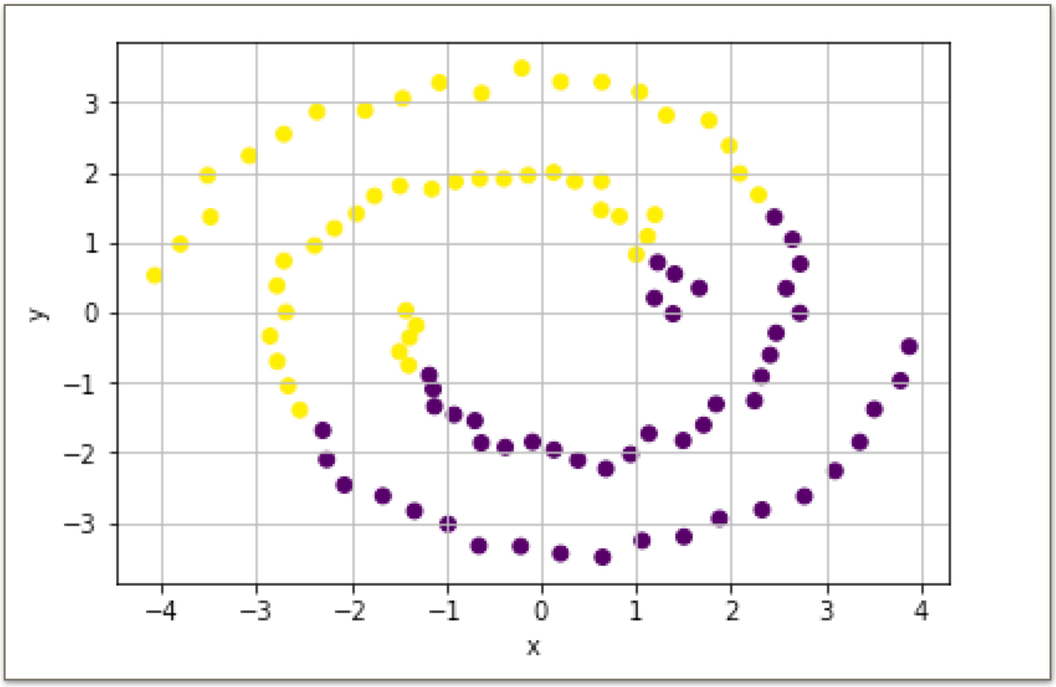
N Clusters	Slop	Acceleration:
1	0.0000	0.0000
2	-47402.6821	0.0000
3	-39237.1909	8165.4912
4	-60002.2663	-20765.0754
5	-2925.3575	57076.9088
6	-17485.5262	-14560.1687
7	-3075.1889	14410.3373
8	9014.5054	12089.6943
9	-21250.7160	-30265.2214
10	-27790.8268	-6540.1108
11	-6743.7023	21047.1244
12	-194900.9106	-188157.2083
13	-12308.6764	182592.2342
14	-2742.8615	9565.8150
15	-2785.6545	-42.7931

Question 4:
(a). The scatterplot of y versus x:



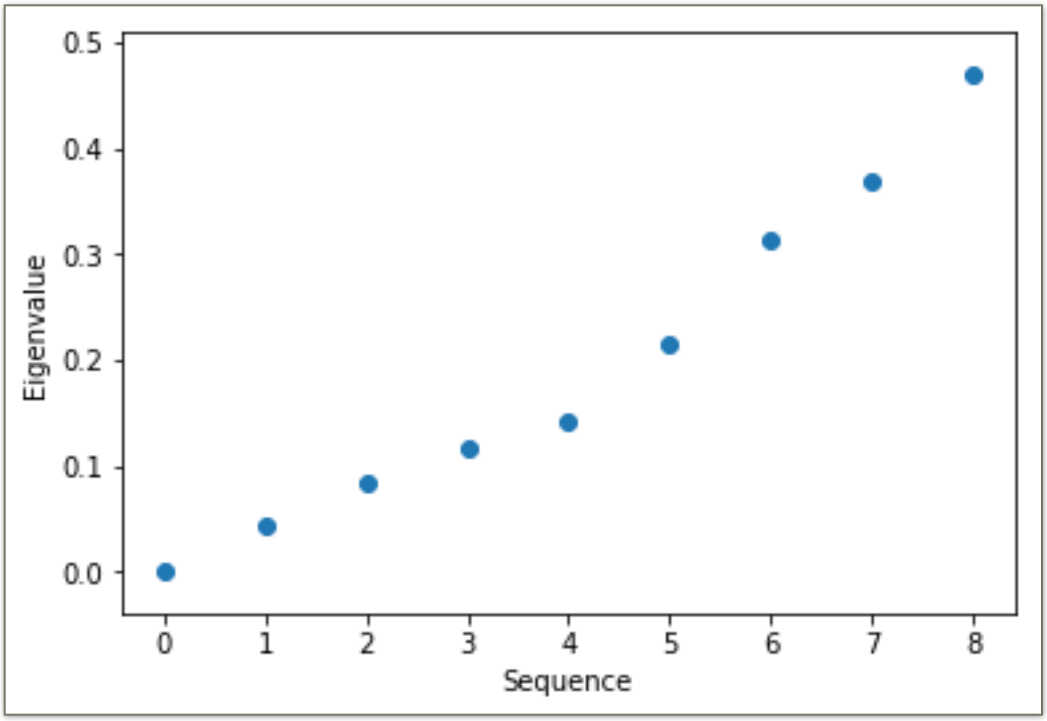
By visual inspection, there are 2 clusters.

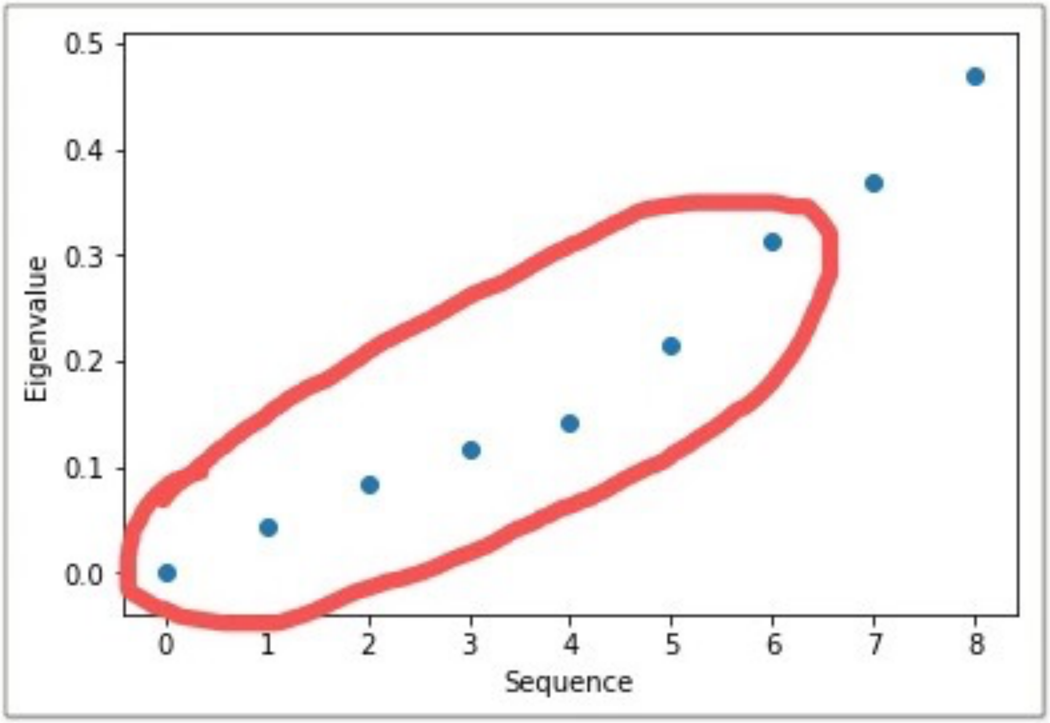
(b) Apply the K-mean algorithm using 2 of clusters
Regenerated scatterplot(different clusters are identified by different colors):



(c). 8 nearest neighbors will be used

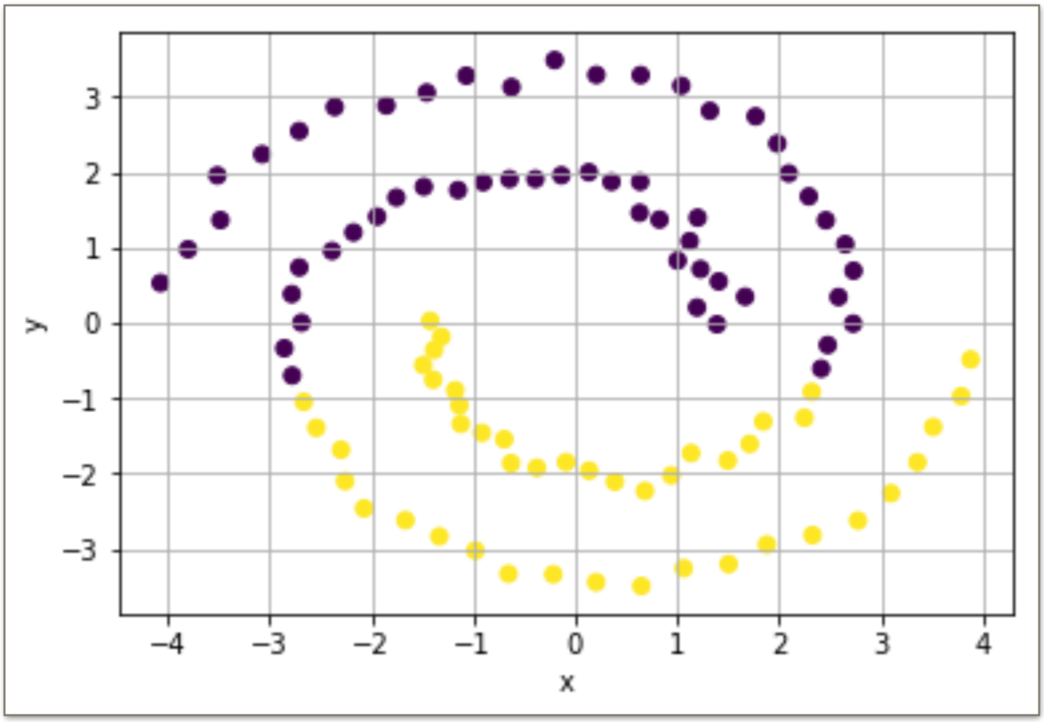
(d). The sequence plot of the first nine eigenvalues:





There is an obvious jump from 5 to 6.
The graph shows that the seven nearest neighbors solution is more appropriate.

(e). Apply the K-mean algorithm on the first two eigenvectors that correspond to the first two smallest eigenvalues. The regenerated scatterplot:



(f). The actual result doesn’t confirm to the expected result. This method works not so good on this dataset.