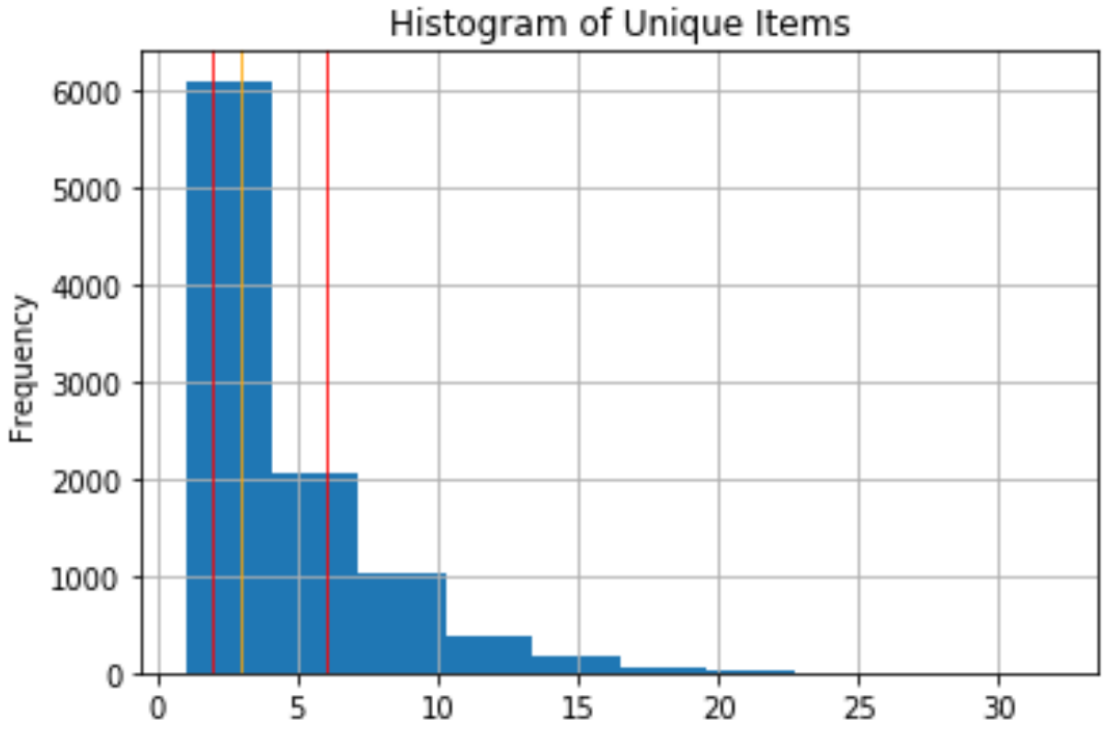
CS 584-04: Machine Learning

Fall 2019: Assignment 2

Amitdeb Prasad Bhattacharya A20402789

Question 1

a) (10 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?




```
Median: 3.0, 25th Percentile: 2.0, 75th Percentile: 6.0
```

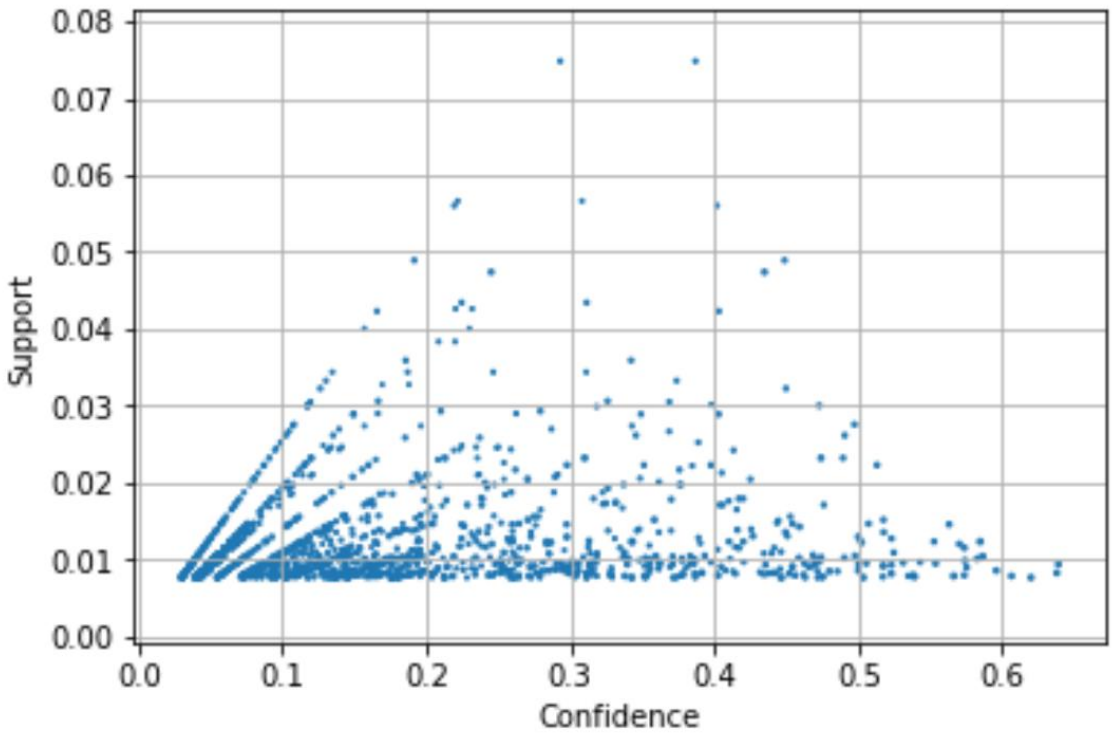
b) (10 points) If you are interested in the *k*-itemsets which can be found in the market baskets of at least seventy five (75) customers. How many itemsets can you find? Also, what is the largest *k* value among your itemsets?

```
Total Item-sets Found: 524
The highest k-value is : 4
```

c) (10 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. Also, you **do not** need to show those rules.

```
Total Association rules found: 1228
```

d) (10 points) Graph the Support metrics on the vertical axis against the Confidence metrics on the horizontal axis for the rules you found in (c). Please use the Lift metrics to indicate the size of the marker.

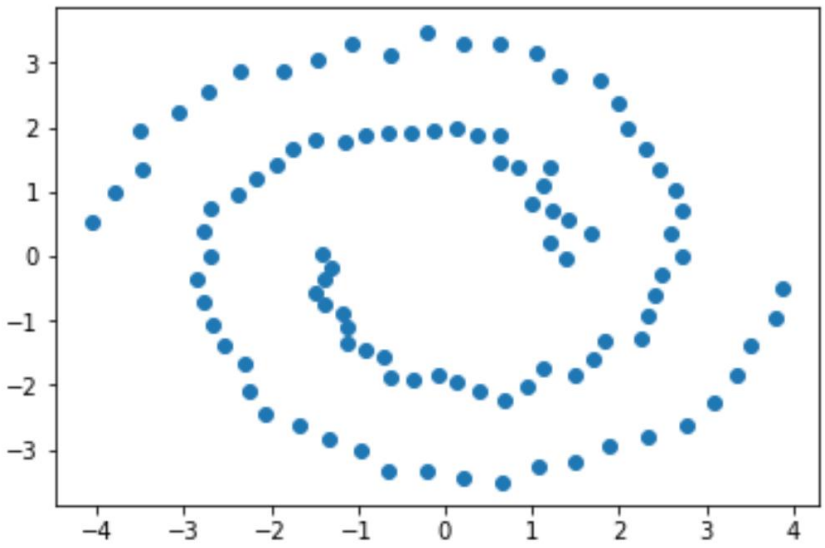


e) (10 points) List the rules whose Confidence metrics are at least 60%. Please include their Support and Lift metrics.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(butter, root vegetables)	(whole milk)	0.012913	0.255516	0.008236	0.637795	2.496107	0.004936	2.055423
1	(butter, yogurt)	(whole milk)	0.014642	0.255516	0.009354	0.638889	2.500387	0.005613	2.061648
2	(yogurt, root vegetables, other vegetables)	(whole milk)	0.012913	0.255516	0.007829	0.606299	2.372842	0.004530	1.890989
3	(yogurt, tropical fruit, other vegetables)	(whole milk)	0.012303	0.255516	0.007626	0.619835	2.425816	0.004482	1.958317

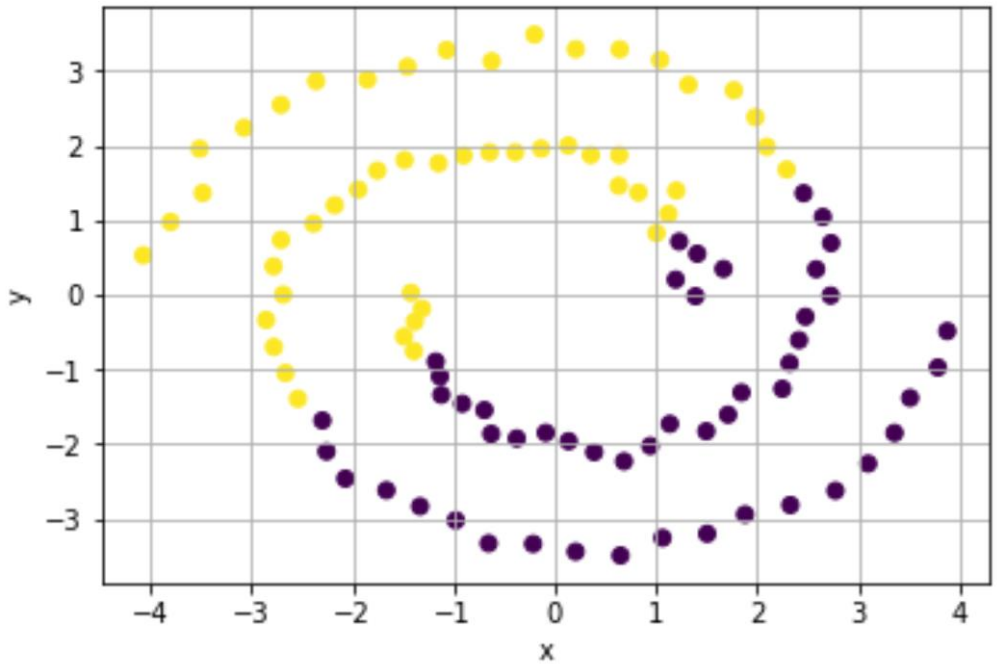
Question 2

a) (10 points) Generate a scatterplot of y (vertical axis) versus x (horizontal axis). How many clusters will you say by visual inspection?



We can see a spiral structure with 2 cluster.

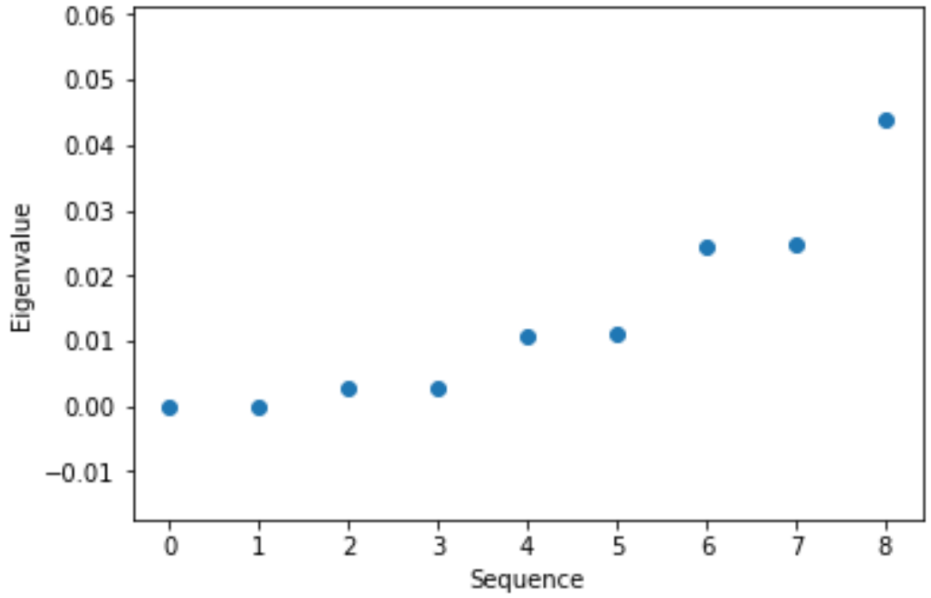
b) (10 points) Apply the K-mean algorithm directly using your number of clusters that you think in (a). Regenerate the scatterplot using the K-mean cluster identifier to control the color scheme?



c) (10 points) Apply the nearest neighbor algorithm using the Euclidean distance. How many nearest neighbors will you use? Remember that you may need to try a couple of values first and use the eigenvalue plot to validate your choice.

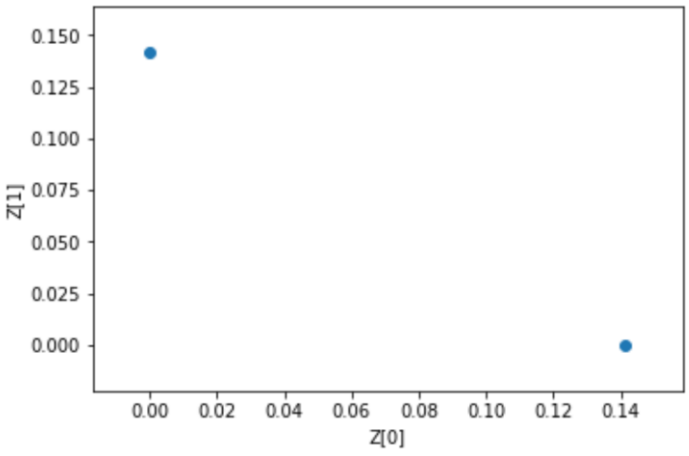
We will use 3 nearest neighbors.

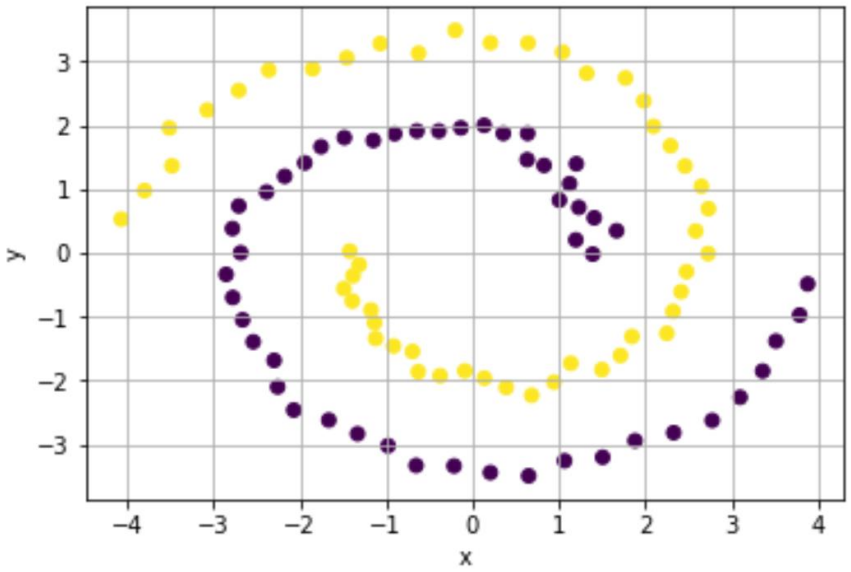
d) (10 points) Retrieve the first <u>two</u> eigenvectors that correspond to the first two smallest eigenvalues. Display up to ten decimal places the means and the standard deviation of these two eigenvectors. Also, plot the first eigenvector on the horizontal axis and the second eigenvector on the vertical axis.



The above graph confirms that the three nearest neighbor's solution is appropriate.

e) (10 points) Apply the K-mean algorithm on your first <u>two</u> eigenvectors that correspond to the first two smallest eigenvalues. Regenerate the scatterplot using the K-mean cluster identifier to control the color scheme?




3
2
1
0
−1
−2
−3
y
−4
−3
−2
−1
0
1
2
3
4
x