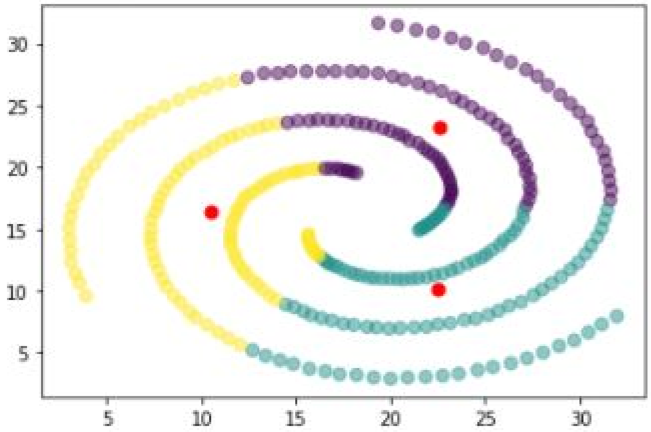
Data mining assignment 2

The assignment is done by using jupyter notebook and many python libraries. This document will include the results from the notebook and the explanations.

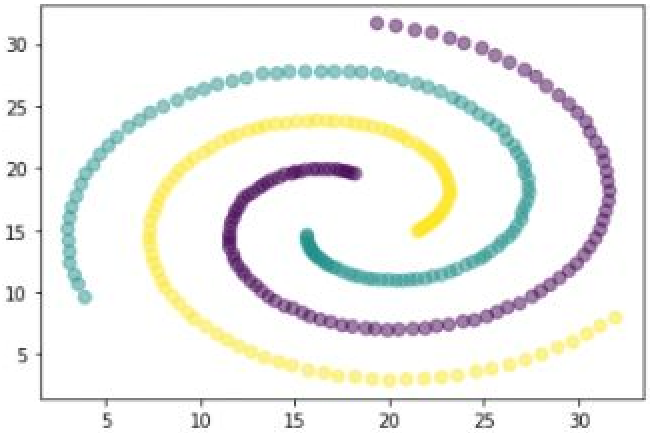
1. Started with pruning the data. The file spiral.txt file was not readable as is so replaced the tabs in it with commas and space so that it could be read with the function:
   data = pd.read_csv('spiral.txt', header=None)

   First, we do the clustering with k-means and get the following plot:



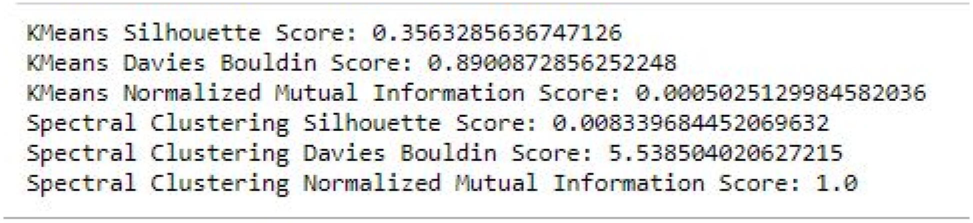
   Here you can see the different groups in different colors and the centroids for the clusters are marked with red.

   Then we do the spectral clustering and get this plot:



   Here the different clusters are displayed in different colors.

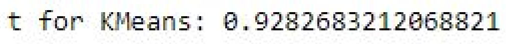
   Then we compute the different quality matrices. The resulting output is:



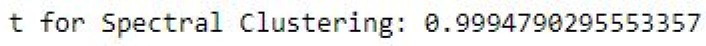
```
KMeans Silhouette Score: 0.3563285636747126
KMeans Davies Bouldin Score: 0.8900872856252248
KMeans Normalized Mutual Information Score: 0.0005025129984582036
Spectral Clustering Silhouette Score: 0.008339684452069632
Spectral Clustering Davies Bouldin Score: 5.538504020627215
Spectral Clustering Normalized Mutual Information Score: 1.0
```

According to the Silhouette and Davis Bouldin Score, the k-means algorithm is the better one because its Silhouette score is closer to one and the Bouldin score is closer to 0. If we look at the normalized mutual information score the k-means algorithm is really poor while the spectral cluster algorithm gets the perfect score. So as a conclusion, both algorithms have their perks but by looking at the graphs and the normalized mutual information score, then the spectral clustering does a better job.

2. A. Started with computing t in the notebook and got the following:



```
t for KMeans: 0.9282683212068821
```



```
t for Spectral Clustering: 0.9994790295553357
```

By looking the algorithm:

$$\tau = \frac{1}{n} \sum_{i=1}^{n} \frac{\sum_{j \neq i} c_{ij} K_{ij}}{\sum_{j \neq i} K_{ij}},$$

We can make the conclusion that the maximal value for this quality measuring algorithm is 1 and therefore the spectral clustering method is the better algorithm. This method goes against the results form the Silhouette and Davies-Bouldin score, but as also concluded in the first exercise they were not very suitable for this data, and therefore the t value is more accurate than them.

B. It uses Euclidean distance which can be unreliable. Also choosing different sigmas can give very different results and therefore the score can be quite misleading.

C. Dunn index, because it’s based on the data itself which would be useful in this case.

3. A. Calvluating the values by using the lift formula:

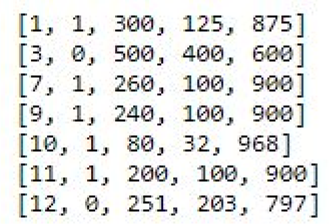
$$lift = \frac{P(A \cap B)}{P(A) * P(B)}$$

After this, the row number 2 and 8 are under 1 and they get pruned out.

B. Using the mutual information formula given in the assignment to prune the data:

$$MI = \log \frac{P(\mathbf{X}C)^{P(\mathbf{X}C)} P(\mathbf{X}\neg C)^{P(\mathbf{X}\neg C)} P(\neg\mathbf{X}C)^{P(\neg\mathbf{X}C)} P(\neg\mathbf{X}\neg C)^{P(\neg\mathbf{X}\neg C)}}{P(\mathbf{X})^{P(\mathbf{X})} P(\neg\mathbf{X})^{P(\neg\mathbf{X})} P(C)^{P(C)} P(\neg C)^{P(\neg C)}}$$

After pruning we are left with the following rules:



```
[1, 1, 300, 125, 875]
[3, 0, 500, 400, 600]
[7, 1, 260, 100, 900]
[9, 1, 240, 100, 900]
[10, 1, 80, 32, 968]
[11, 1, 200, 100, 900]
[12, 0, 251, 203, 797]
```