เริ่มต้นด้วยการ import

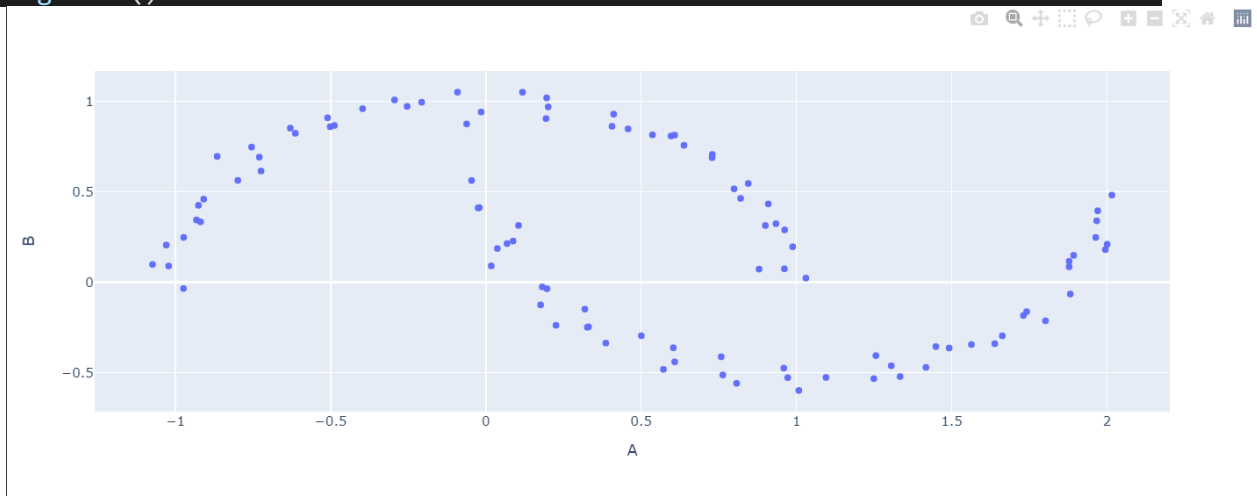
```
import numpy as np
import matplotlib.pyplot as plt
from sklearn import cluster
import pandas as pd
from scipy.cluster.hierarchy import dendrogram, linkage
from scipy.spatial import distance_matrix
```

```
data = pd.read_csv("data/data2Dset2.csv")
data.columns = ["A","B"]
data.head()
```

	A	B
0	1.967099	0.339064
1	0.762843	-0.513650
2	-1.029709	0.205156
3	0.637710	0.756872
4	2.000786	0.209418

2) Plot จุดข้อมูล data1

```
1) import plotly.express as px
2) fig = px.scatter(data, x="A", y="B")
3) fig.show()
```



3) Data2Dset2 เริ่มด้วยการเขียนโปรแกรม plot จุดข้อมูลโดยใช้วิธี kmeans

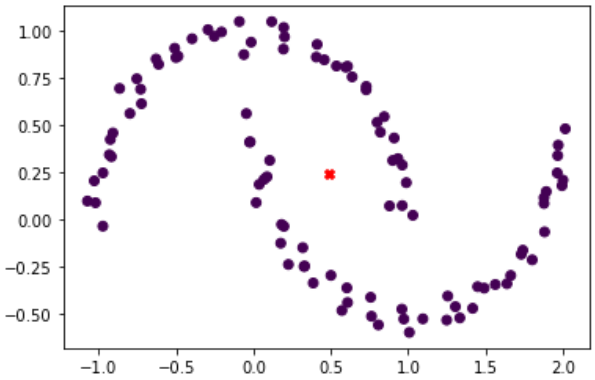
K=1

```
model_kmeans = cluster.KMeans(n_clusters=1, max_iter=50, random_state=1)
model_kmeans.fit(data)
data1['cluster_id'] = model_kmeans.labels_
centroids = model_kmeans.cluster_centers_
print(centroids)
```

```
[[0.49237282 0.24273502 0.54      ]]
```

```
plt.scatter(data['A'],data['B'], c=data['cluster_id'])
plt.scatter(centroids[:,0],centroids[:,1],marker='X',c='r')
plt.show()
```

ผลลัพธ์ K=1



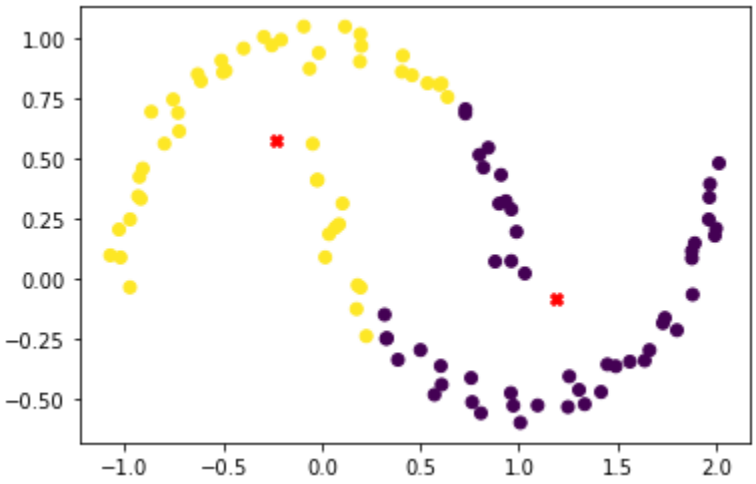
K=2

```
model_kmeans = cluster.KMeans(n_clusters=2, max_iter=50, random_state=1)
model_kmeans.fit(data)
data1['cluster_id'] = model_kmeans.labels_
centroids = model_kmeans.cluster_centers_
print(centroids)
```

```
[[ 1.26403634 -0.10476722  0.5       ]
 [-0.16497018  0.53875545  1.37037037]]
```

```
plt.scatter(data['A'],data['B'], c=data1['cluster_id'])
plt.scatter(centroids[:,0],centroids[:,1],marker='X',c='r')
plt.show()
```

ผลลัพธ์ K=2



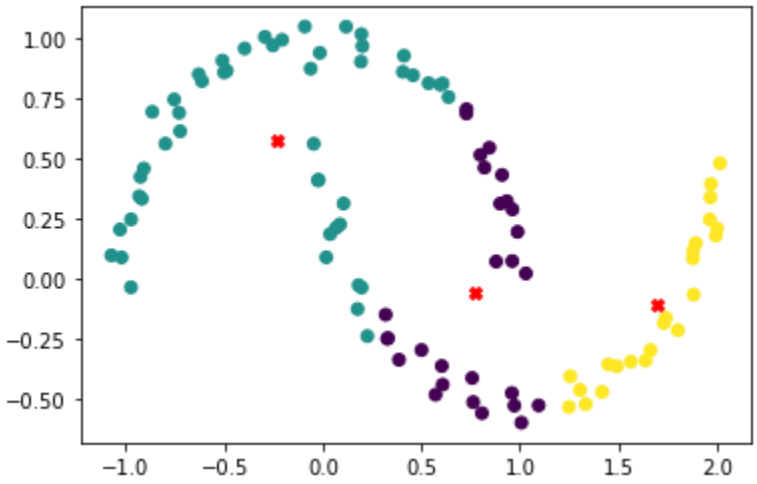
K=3

```
model_kmeans = cluster.KMeans(n_clusters=3, max_iter=50, random_state=1)
model_kmeans.fit(data)
data['cluster_id'] = model_kmeans.labels_
centroids = model_kmeans.cluster_centers_
print(centroids)
```

```
[[ 1.69940443 -0.10991557  2.        ]
 [ 0.45486035  0.28170959  0.52631579]
 [-0.78880298  0.5372057   3.        ]]
```

```
plt.scatter(data['A'],data['B'], c=data['cluster_id'])
plt.scatter(centroids[:,0],centroids[:,1],marker='X',c='r')
plt.show()
```

ผลลัพธ์ K=3



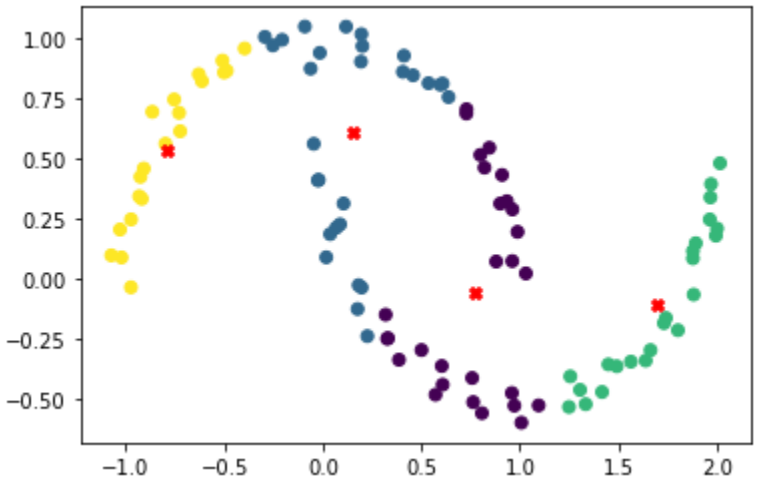
K=4

```
model_kmeans = cluster.KMeans(n_clusters=4, max_iter=50, random_state=1)
model_kmeans.fit(data)
data['cluster_id'] = model_kmeans.labels_
centroids = model_kmeans.cluster_centers_
print(centroids)
```

```
[[ 0.67963398 -0.18109054  0.        ]
 [ 0.25256409  0.6982297   0.        ]
 [ 1.69940443 -0.10991557  0.        ]
 [-0.78880298  0.5372057   0.        ]]
```

```
plt.scatter(data['A'],data['B'], c=data['cluster_id'])
plt.scatter(centroids[:,0],centroids[:,1],marker='X',c='r')
plt.show()
```

ผลลัพธ์K=4



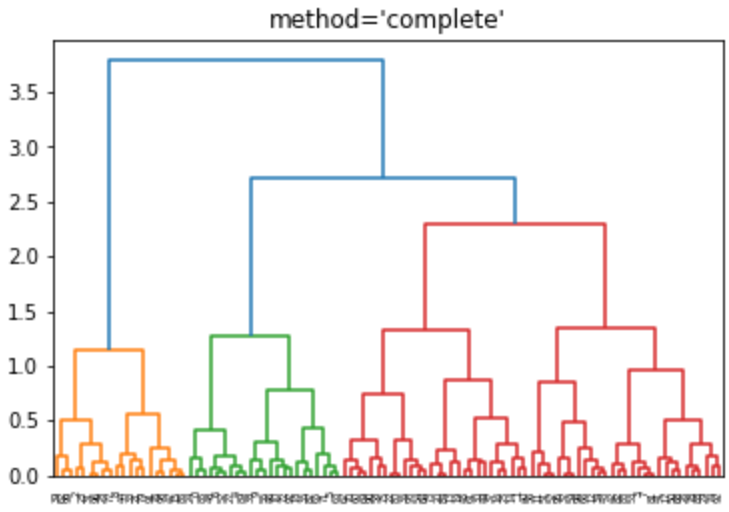
จากผลลัพธ์สรุปได้ว่า K=4 ดีที่สุดสำหรับ Data2Dset2 เพราะข้อมูลจะดูละเอียดอ่อนมากขึ้นถ้าเทียบกับKอื่นๆ ตัวข้อมูลจะถูกแบ่งเป็น4กลุ่มซึ่งจะแยกกันไปตามช่วงประมาณ1

4) เขียนโปแกรมจัดกลุ่มชุดข้อมูลที่อ่านเข้ามา โดยใช้วิธี Hierarchical Clustering

4.1) ให้เลือกใช้method ที่ต่างกัน 3 แบบ แสดง dendrogram ที่ได้แต่ละแบบ

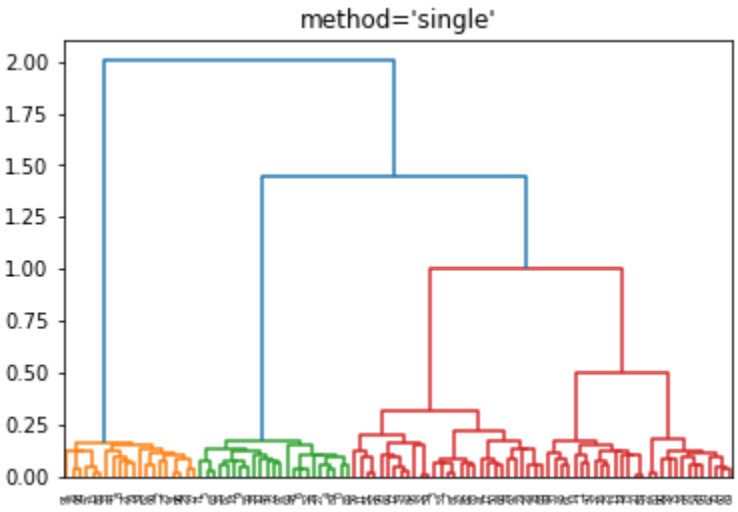
Complete:

```
linkage_data = linkage(data, method='complete' , metric='euclidean')
dendrogram(linkage_data)
plt.title("method='complete'")
plt.show()
```

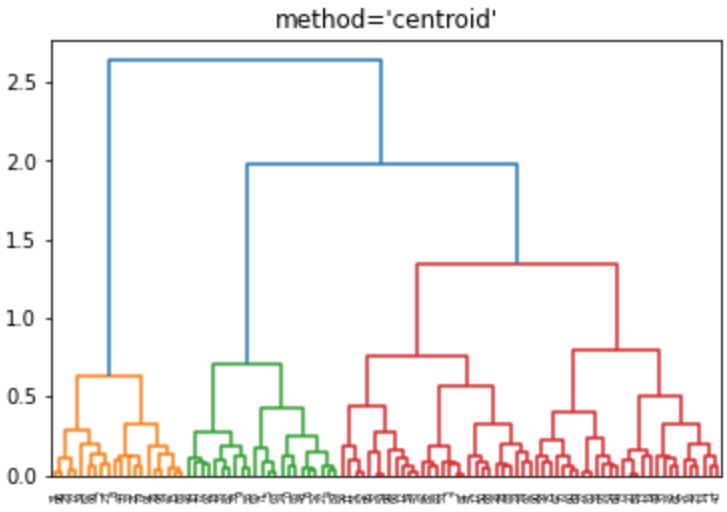


Single:

```
linkage_data = linkage(data, method='single' , metric='euclidean')
dendrogram(linkage_data)
plt.title("method='single'")
plt.show()
```



Centroid:



4.2) เลือก cut-off โดยกำหนด criterion='distance' และให้นักศึกษาเลือกระบุค่า t ที่คิดว่า เหมาะสม สำหรับแต่ละ dendrogram ที่ได้ในข้อ 4.1)

Complete: ค่า t ที่เหมาะที่สุดคือ 2

```
cluster_id = fcluster(linkage_data,t=2,criterion='distance')
plt.scatter(data["A"],data["B"],c=cluster_id)
plt.show()
```

Single: ค่า t ที่เหมาะที่สุดคือ 2

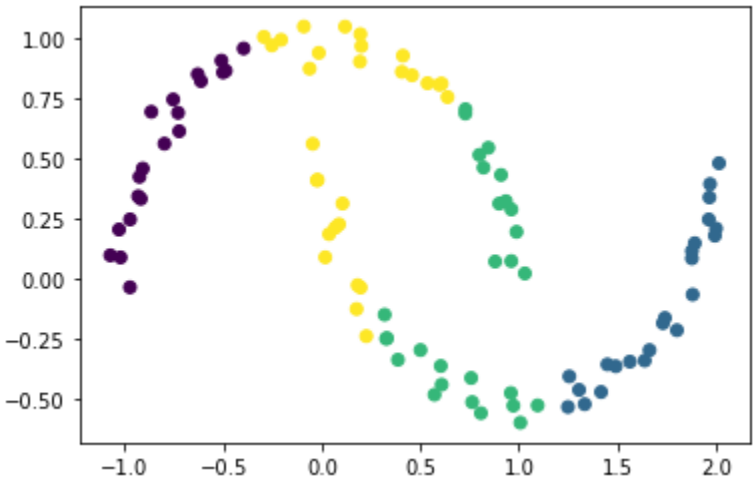
```
cluster_id = fcluster(linkage_data,t=1,criterion='distance')
plt.scatter(data["A"],data["B"],c=cluster_id)
plt.show()
```

Centroid: ค่า t ที่เหมาะที่สุดคือ 0.8

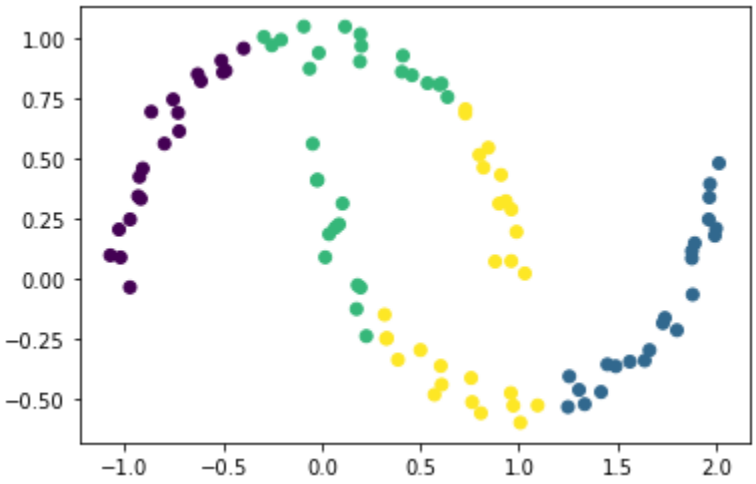
```
cluster_id = fcluster(linkage_data,t=0.8,criterion='distance')
plt.scatter(data["A"],data["B"],c=cluster_id)
plt.show()
```

4.3) Plot ผลการจัดกลุ่ม ที่ได้แต่ละแบบในข้อ 4.2)

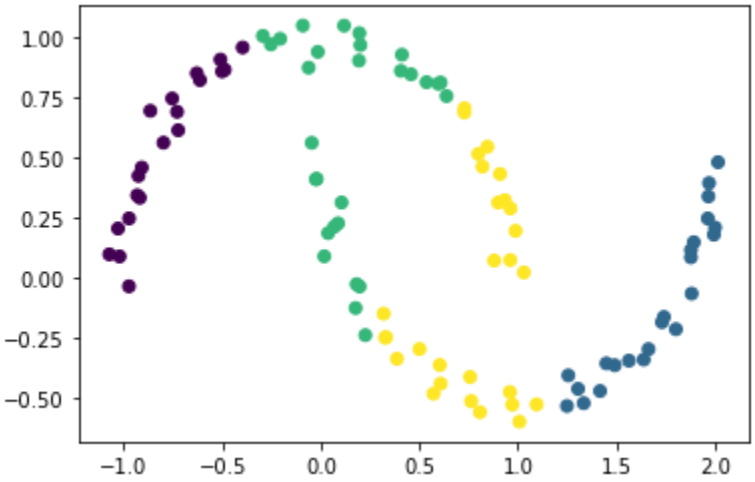
Complete:



Single:



Centroid:



6. เขียนบรรยายสรุปผลการทดลอง แสดงความคิดเห็น วิธีใด เหมาะกับ ชุดข้อมูลแบบไหน แต่ละวิธีมีข้อดี/ ข้อเสียอย่างไร

- Data2DSet2 จากการทดลองพบว่า วิธี Hierarchical Clusterings เพราะ ตัวข้อมูลมีความชิดกันมากจึงไม่สามารถกำหนดค่า K-mean ได้เนื่องจากข้อมูลมีความใกล้กันมาก