เริ่มต้นด้วยการ import

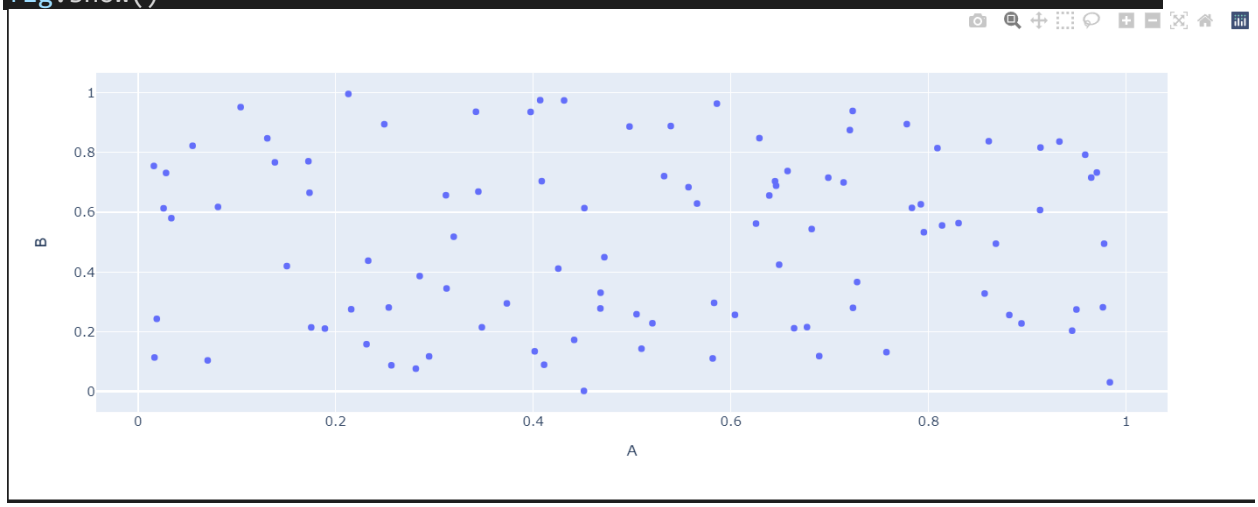
```
import numpy as np
import matplotlib.pyplot as plt
from sklearn import cluster
import pandas as pd
from scipy.cluster.hierarchy import dendrogram, linkage
from scipy.spatial import distance_matrix
```

```
data = pd.read_csv("data/data2Dset3.csv",header=None)
data.columns = ["A","B"]
data.head()
```

	A	B
0	0.028405	0.731484
1	0.471940	0.449512
2	0.970259	0.732859
3	0.070531	0.104092
4	0.539139	0.888368

2) Plot จุดข้อมูล data1

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



3) Data2Dset3 เริ่มด้วยการเขียนโปรแกรม plot จุดข้อมูลโดยใช้วิธี kmeans

K=1

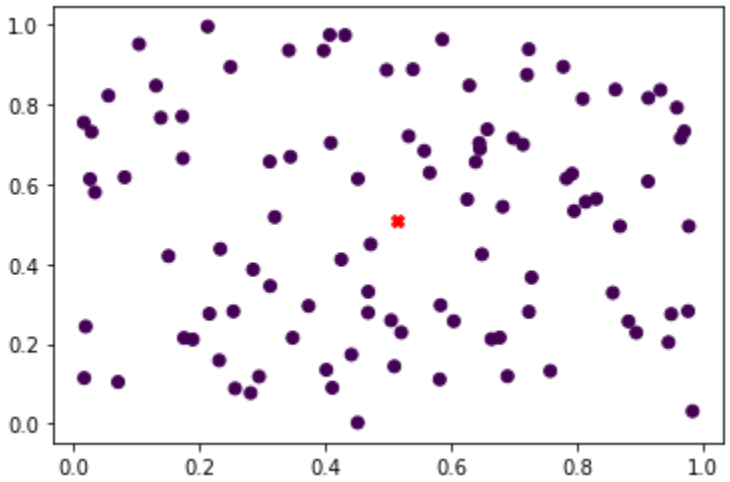
```
model_kmeans = cluster.KMeans(n_clusters=1, max_iter=50, random_state=1)
```

```
model_kmeans.fit(data)
data1['cluster_id'] = model_kmeans.labels_
centroids = model_kmeans.cluster_centers_
print(centroids)
```

```
[[0.51374147 0.51161133]]
```

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

ผลลัพธ์ K=1



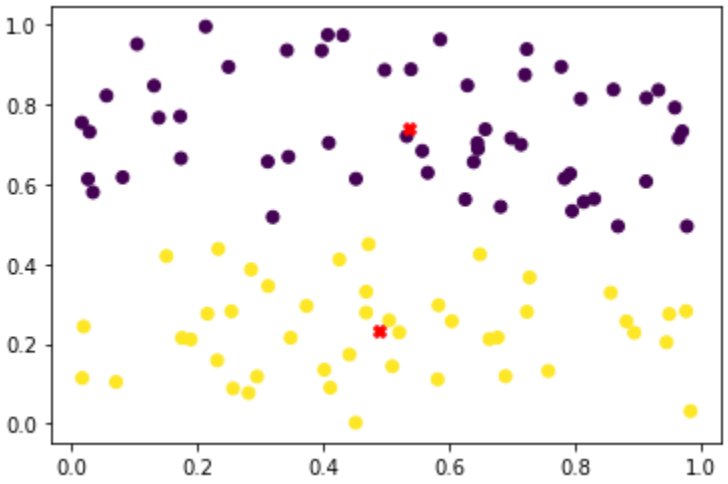
K=2

```
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)
```

```
[[0.53548368 0.73937577 0.        ]
 [0.48716765 0.23323258 0.        ]]
```

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

ผลลัพธ์ K=2



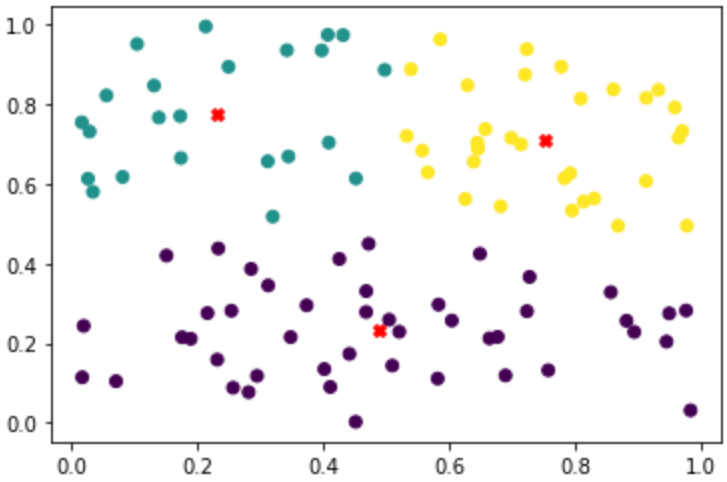
K=3

```
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)
```

```
[[4.87167648e-01 2.33232575e-01 1.00000000e+00]
 [2.31849764e-01 7.77406137e-01 1.11022302e-16]
 [7.53720560e-01 7.12041439e-01 2.77555756e-16]]
```

```
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

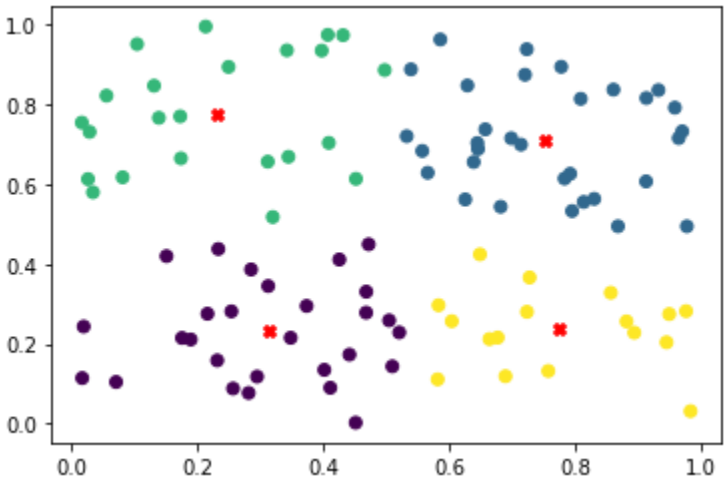


K=4

```
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)
 [[ 3.13529957e-01  2.31446910e-01 -1.11022302e-16]
  [ 7.53720560e-01  7.12041439e-01  2.00000000e+00]
  [ 2.31849764e-01  7.77406137e-01  1.00000000e+00]
  [ 7.73159140e-01  2.36173670e-01  3.33066907e-16]]


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

ผลลัพธ์K=4



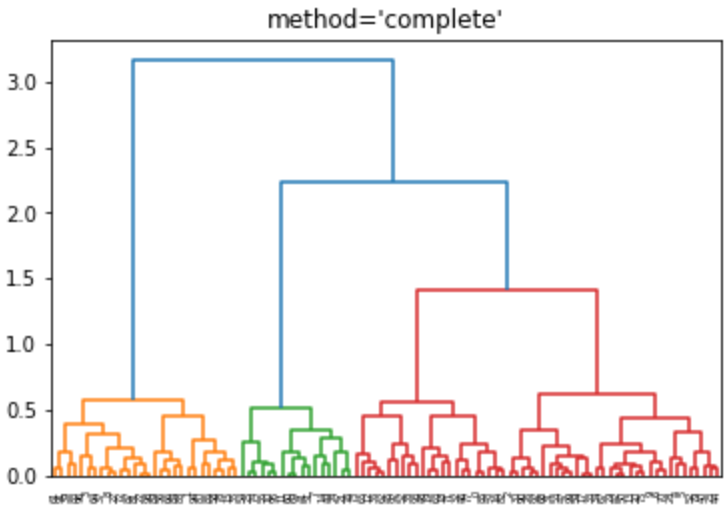
จากผลลัพธ์สรุปได้ว่า K=4 ดีที่สุดสำหรับ Data2Dset2 เพราะข้อมูลจะดูละเอียดอ่อนมากขึ้นถ้าเทียบกับKอื่นๆ ตัวข้อมูลจะถูกแบ่งเป็น 4กลุ่มใหญ่ๆอย่างเห็นได้ชัด

4) เขียนโปแกรมจัดกลุ่มชุดข้อมูลที่อ่านเข้ามา โดยใช้วิธี Hierarchical Clustering

4.1) ให้เลือกใช้method ที่ต่างกัน 3 แบบ แสดง dendrogram ที่ได้แต่ละแบบ

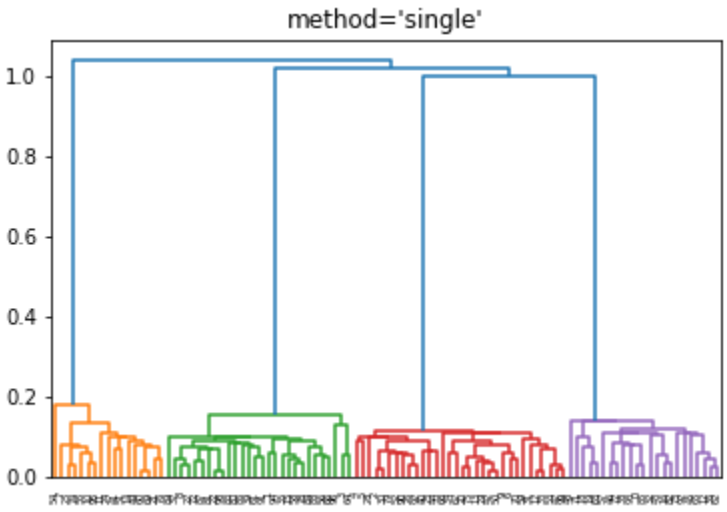
Complete:

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



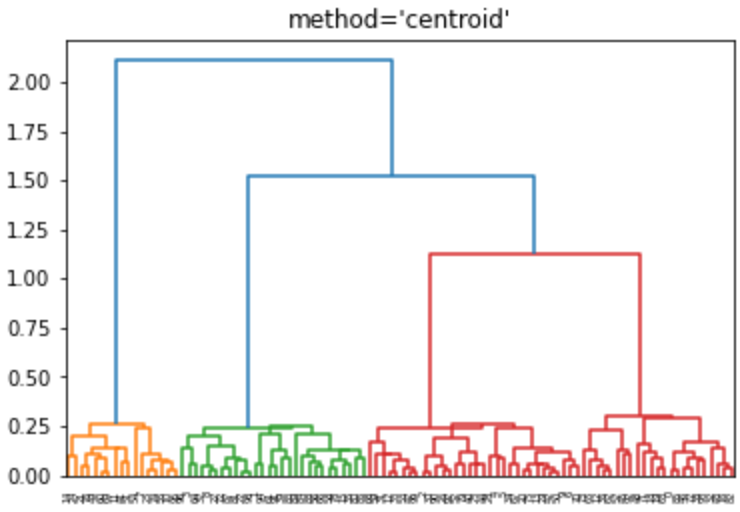
Single:

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



Centroid:

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



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

Complete: ค่า t ที่เหมาะที่สุดคือ 1

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

Single: ค่า t ที่เหมาะที่สุดคือ 1

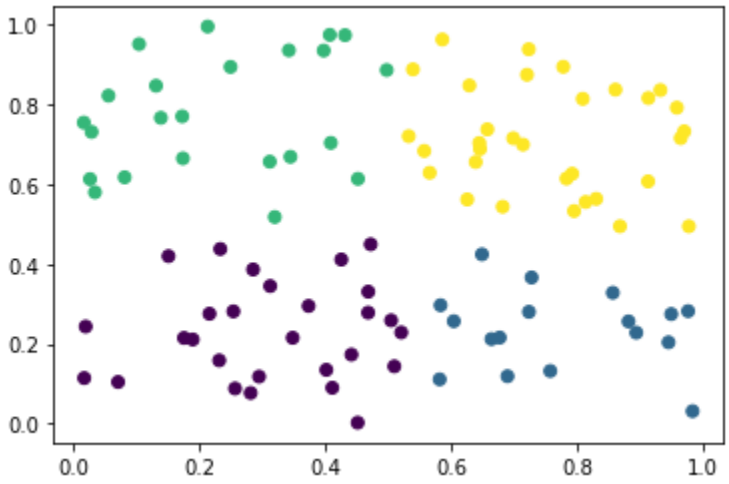
```
cluster_id = fcluster(linkage_data,t=1,criterion='distance')
plt.scatter(data["A"],data["B"],c=cluster_id)
plt.show()
```

Centroid: ค่า t ที่เหมาะที่สุดคือ 1

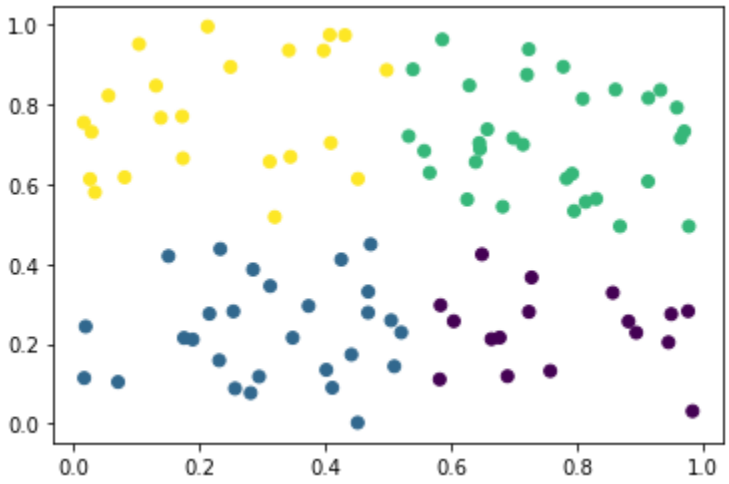
```
cluster_id = fcluster(linkage_data,t=1,criterion='distance')
plt.scatter(data["A"],data["B"],c=cluster_id)
plt.show()
```

4.3) Plot ผลการจัดกลุ่ม ที่ได้แต่ละแบบในข้อ 4.2)

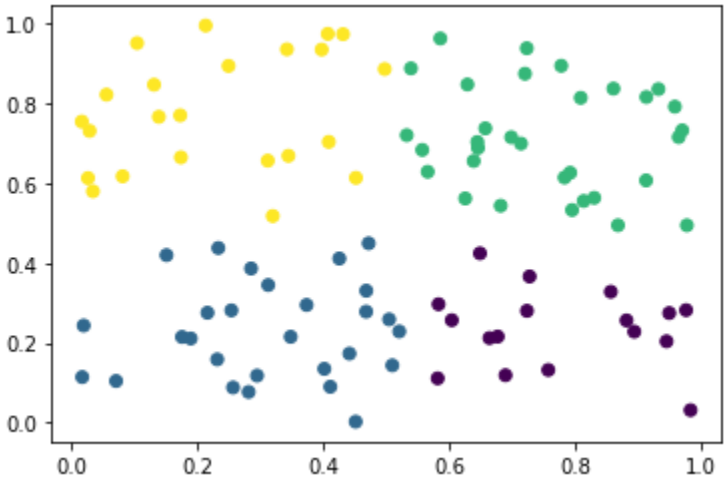
Complete:



Single:



Centroid:



6. เขียนบรรยายสรุปผลการทดลอง แสดงความคิดเห็น วิธีใด เหมาะกับ ชุดข้อมูลแบบไหน แต่ละวิธีมีข้อดี/ ข้อเสีย อย่างไร

- Data2DSet2 จากการทดลองพบว่า วิธี Hierarchical Clusterings เพราะจะมีความละเอียดมากกว่าและจากข้อมูลเราไม่สามารถแยกข้อมูลด้วยตาเปล่าได้ การใช้ kmeans จึงค่อนข้างยาก