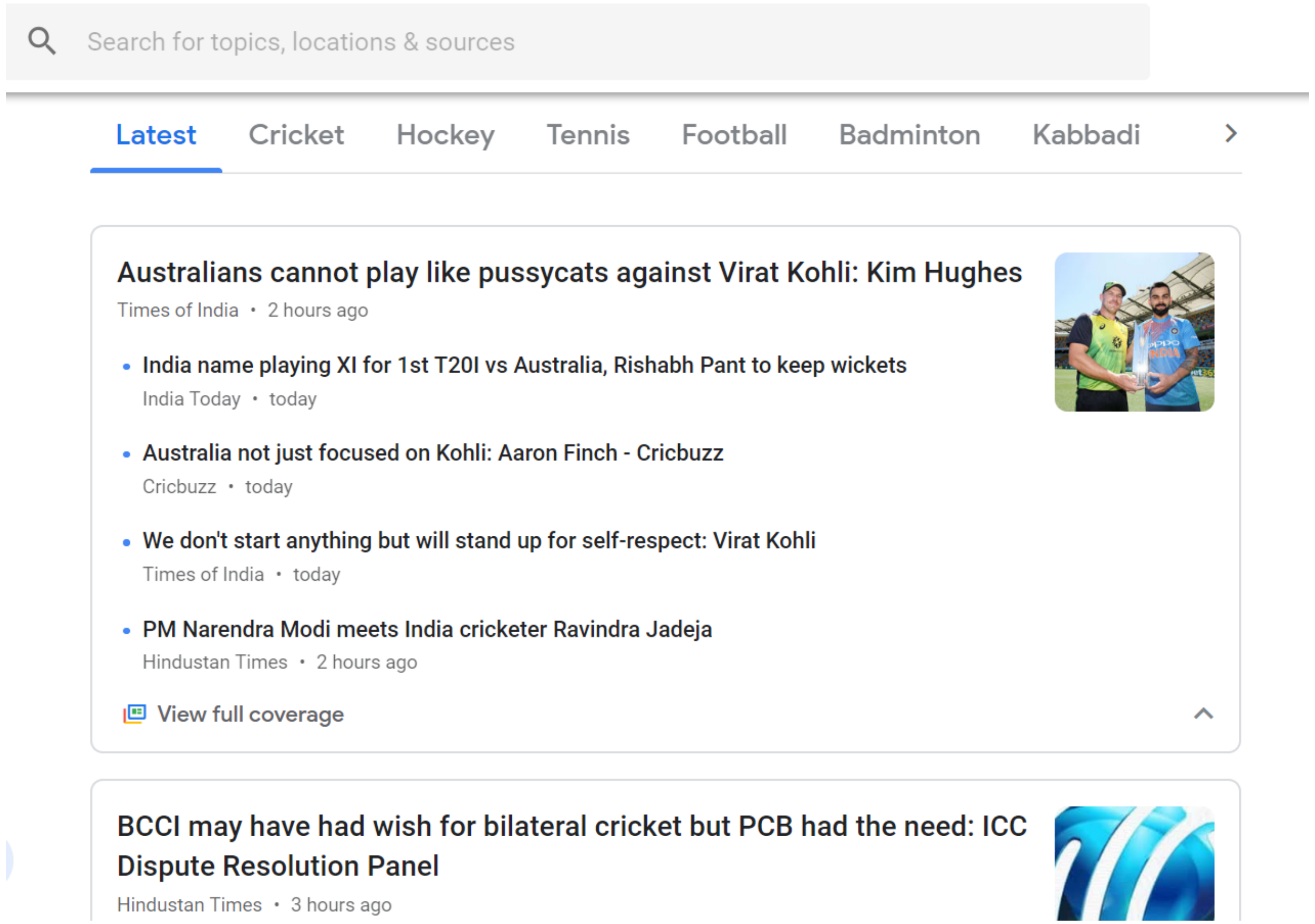
Unsupervised learning: basics

CLUSTER ANALYSIS IN PYTHON

Shaumik Daityari
Business Analyst

Everyday example: Google news

- How does Google News classify articles?
- Unsupervised Learning Algorithm: Clustering
- Match frequent terms in articles to find similarity



Australians cannot play like pussycats against Virat Kohli: Kim Hughes
Times of India • 2 hours ago

- India name playing XI for 1st T20I vs Australia, Rishabh Pant to keep wickets
  India Today • today
- Australia not just focused on Kohli: Aaron Finch - Cricbuzz
  Cricbuzz • today
- We don't start anything but will stand up for self-respect: Virat Kohli
  Times of India • today
- PM Narendra Modi meets India cricketer Ravindra Jadeja
  Hindustan Times • 2 hours ago

BCCI may have had wish for bilateral cricket but PCB had the need: ICC Dispute Resolution Panel
Hindustan Times • 3 hours ago

Labeled and unlabeled data

Data with no labels	• Point 1: (1, 2) • Point 2: (2, 2) • Point 3: (3, 1)
Data with labels	• Point 1: (1, 2), Label: Danger Zone • Point 2: (2, 2), Label: Normal Zone • Point 3: (3, 1), Label: Normal Zone

What is unsupervised learning?

- A group of machine learning algorithms that find patterns in data
- Data for algorithms has not been labeled, classified or characterized
- The objective of the algorithm is to interpret any structure in the data
- Common unsupervised learning algorithms: clustering, neural networks, anomaly detection

What is clustering?

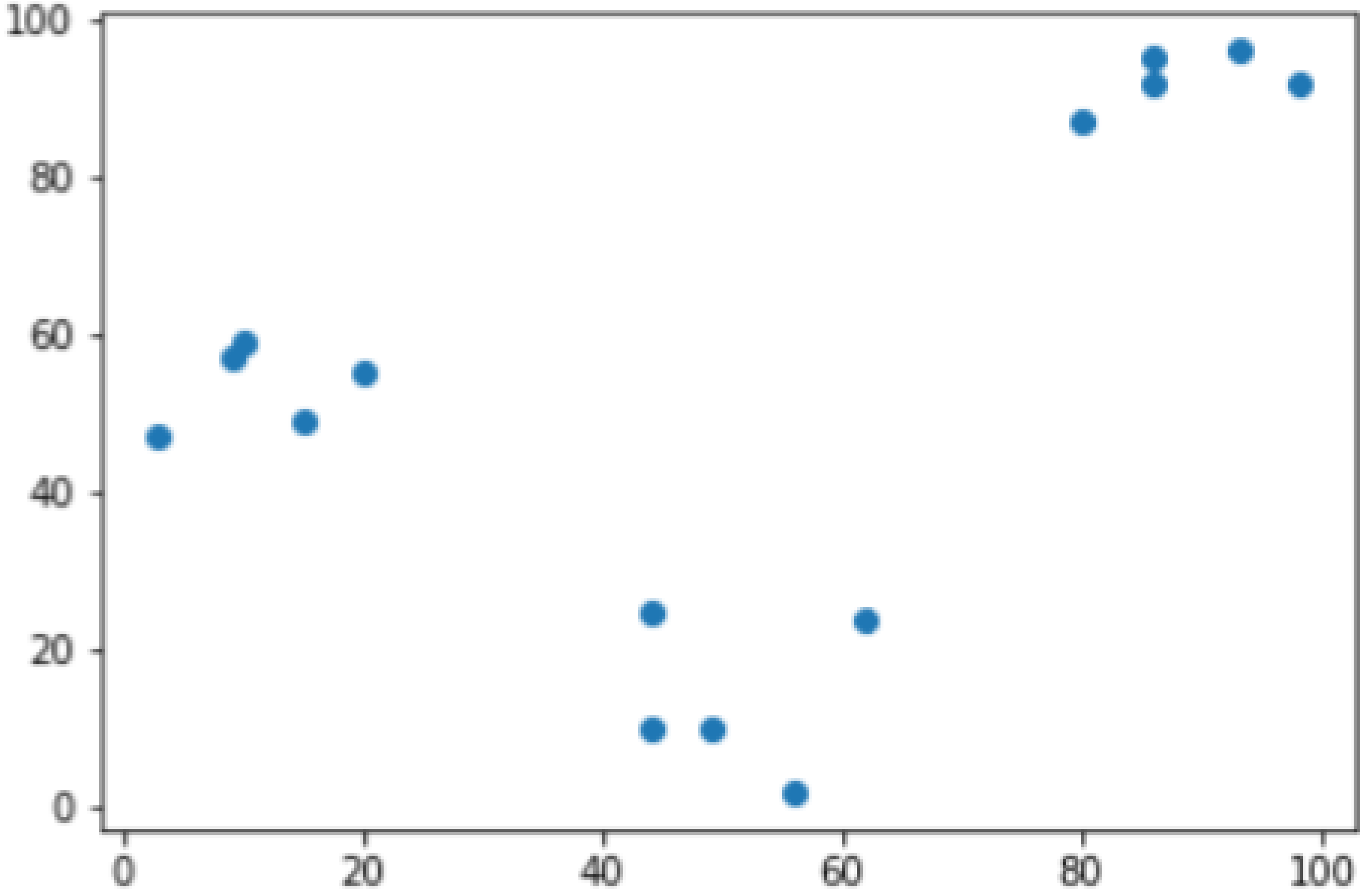
- The process of grouping items with similar characteristics
- Items in groups similar to each other than in other groups
- Example: distance between points on a 2D plane

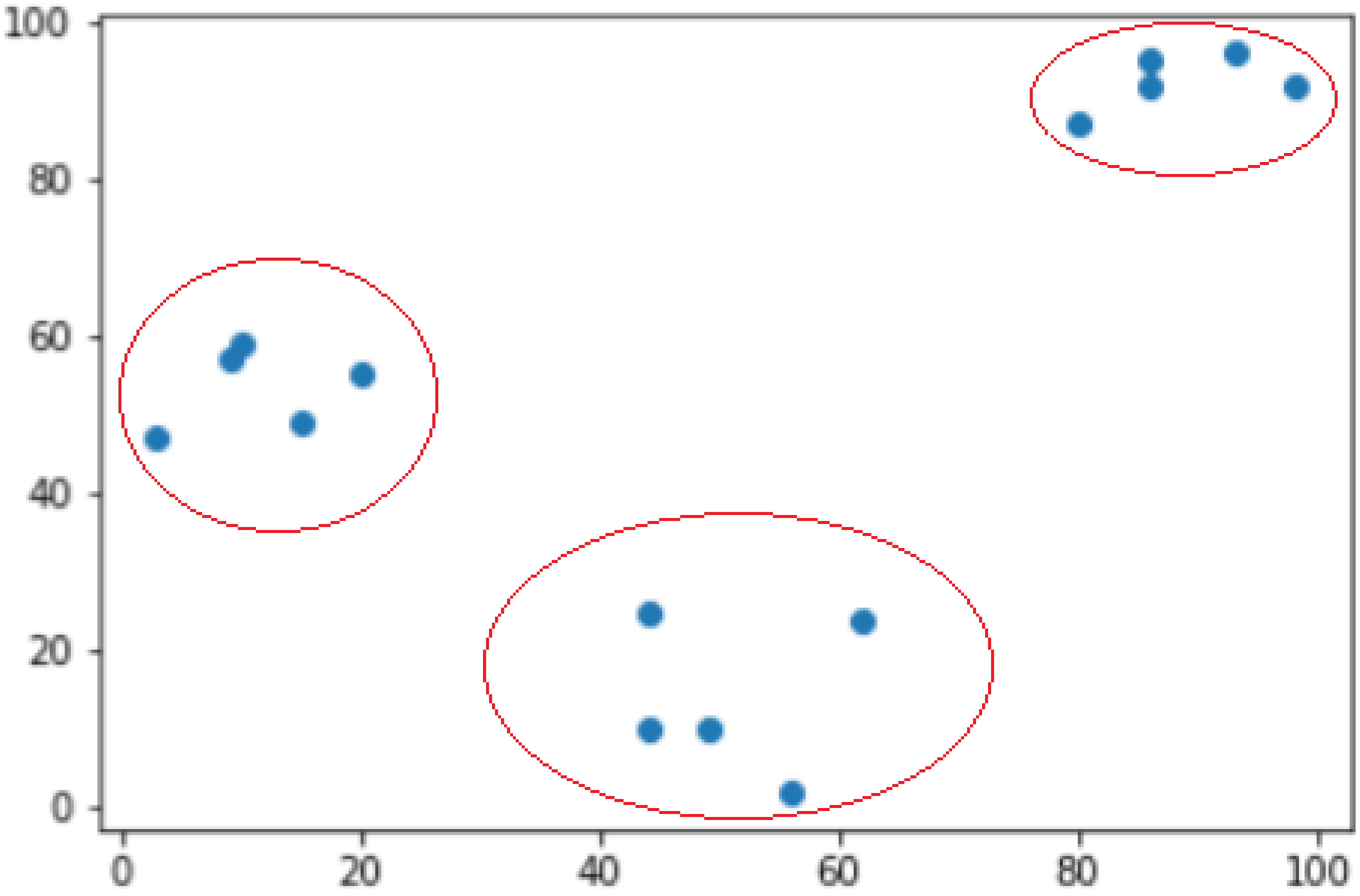
Plotting data for clustering - Pokemon sightings

```python
from matplotlib import pyplot as plt
```

```python
x_coordinates = [80, 93, 86, 98, 86, 9, 15, 3, 10, 20, 44, 56, 49, 62, 44]
y_coordinates = [87, 96, 95, 92, 92, 57, 49, 47, 59, 55, 25, 2, 10, 24, 10]
```

```python
plt.scatter(x_coordinates, y_coordinates)
plt.show()
```


100
80
60
40
20
0
0
20
40
60
80
100


100
80
60
40
20
0
0
20
40
60
80
100

Up next - some practice

CLUSTER ANALYSIS IN PYTHON

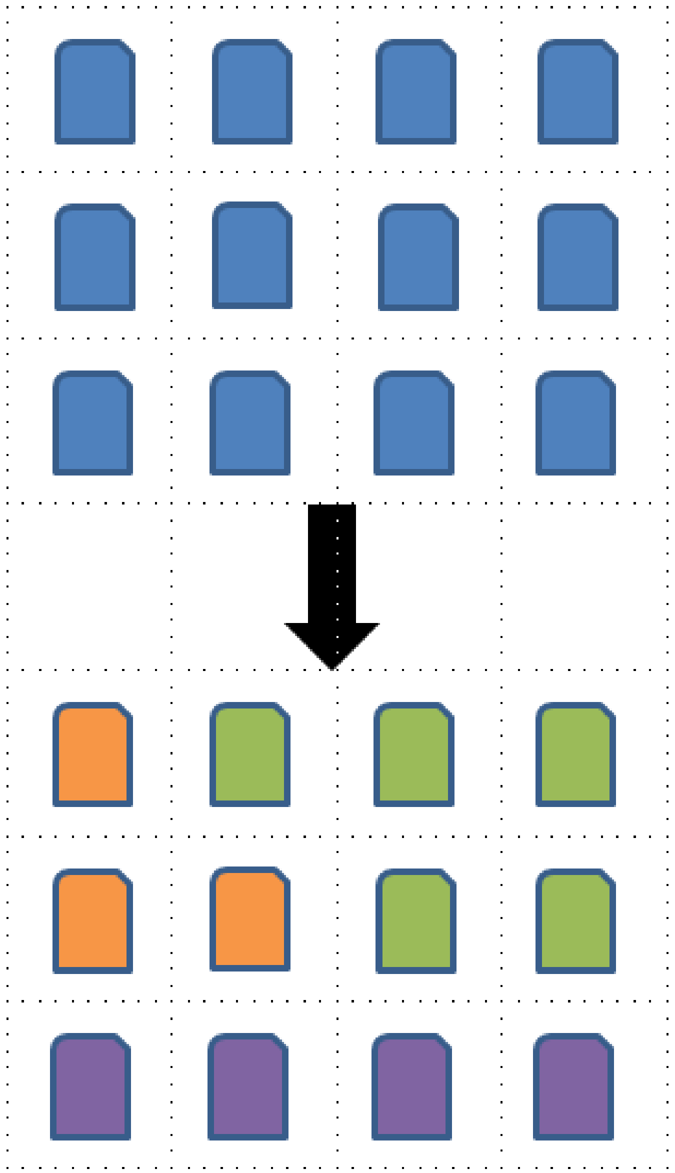
Basics of cluster analysis

CLUSTER ANALYSIS IN PYTHON

Shaumik Daityari
Business Analyst

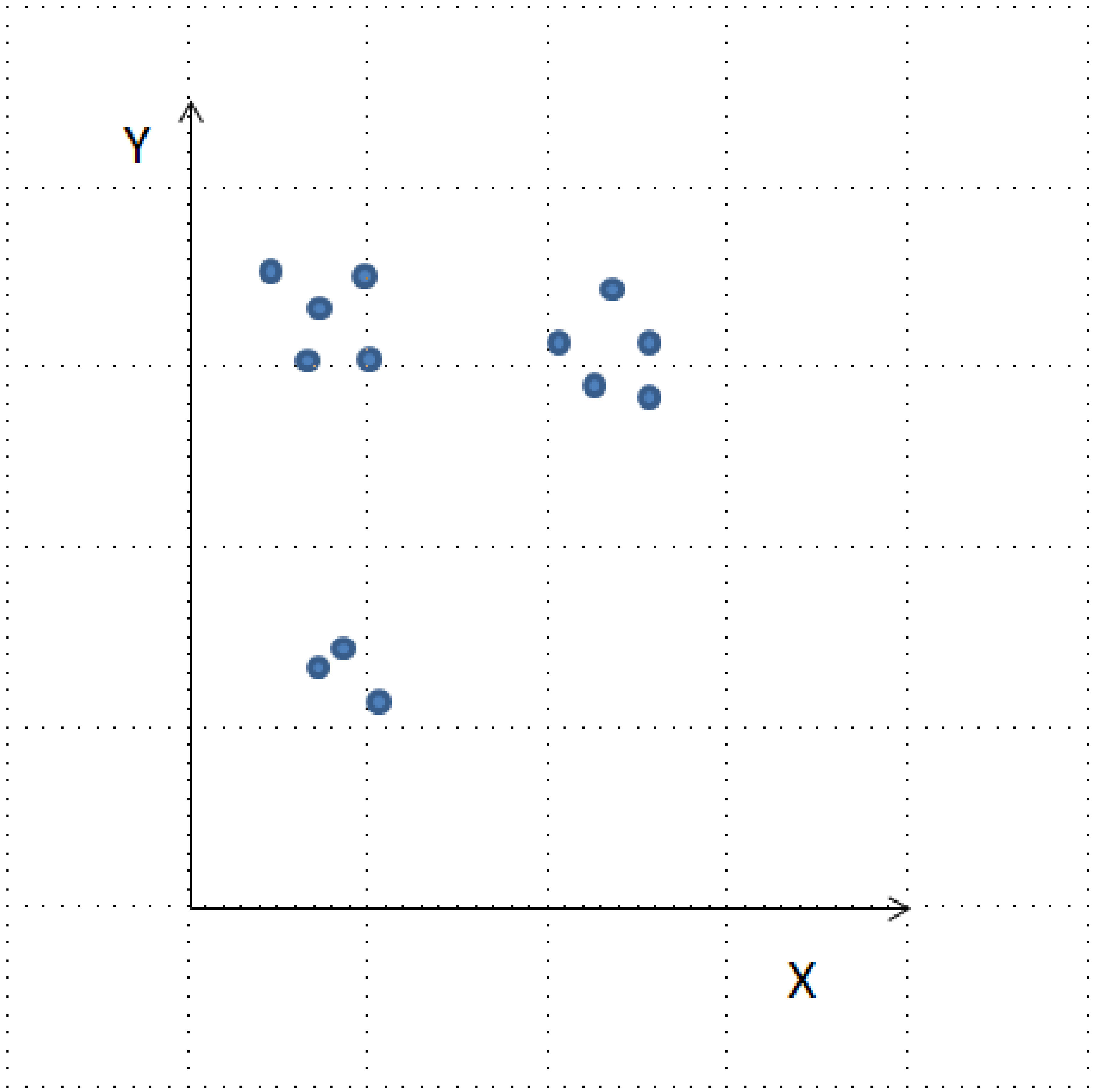
What is a cluster?

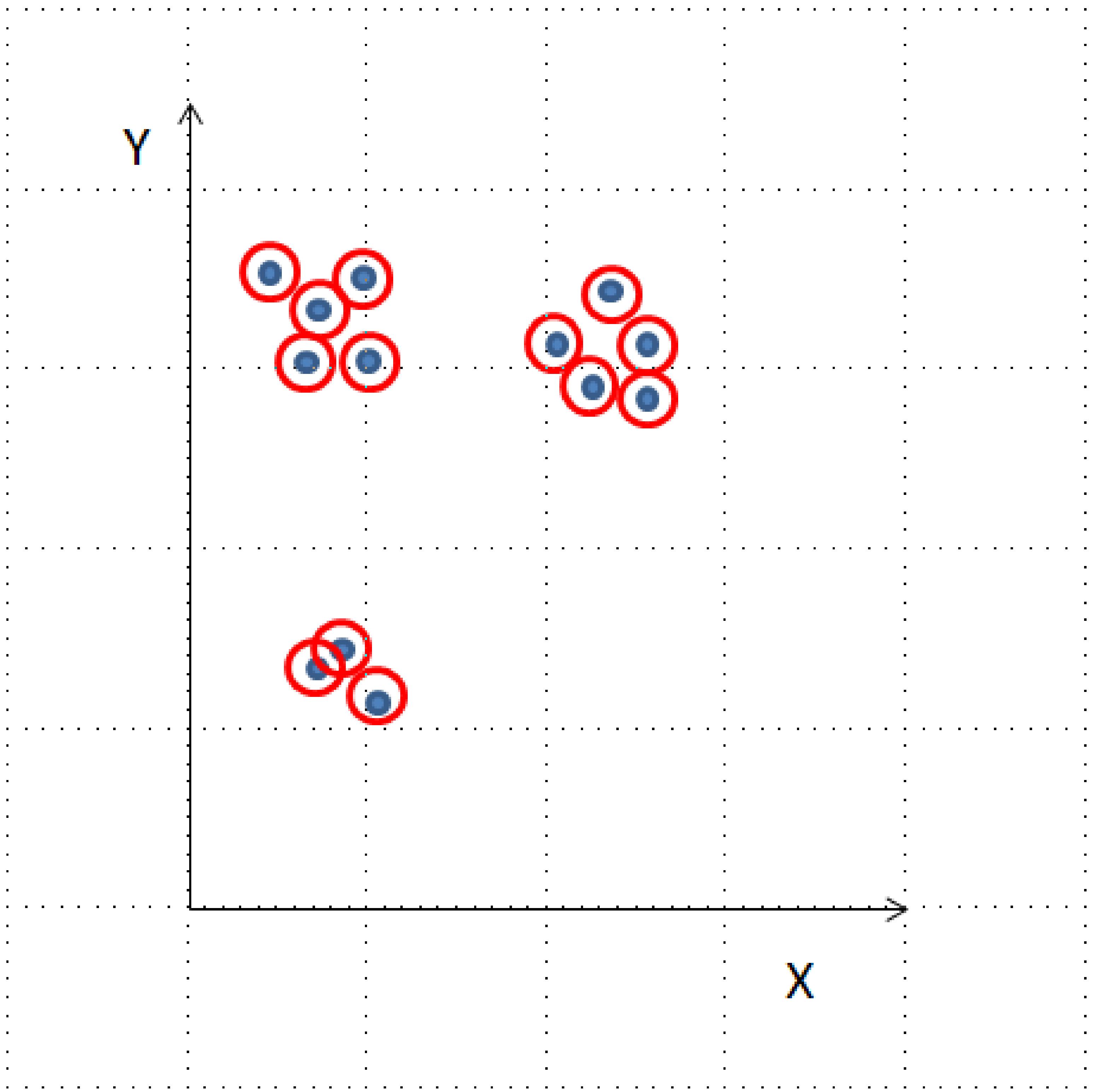
- A group of items with similar characteristics
- Google News: articles where similar words and word associations appear together
- Customer Segments

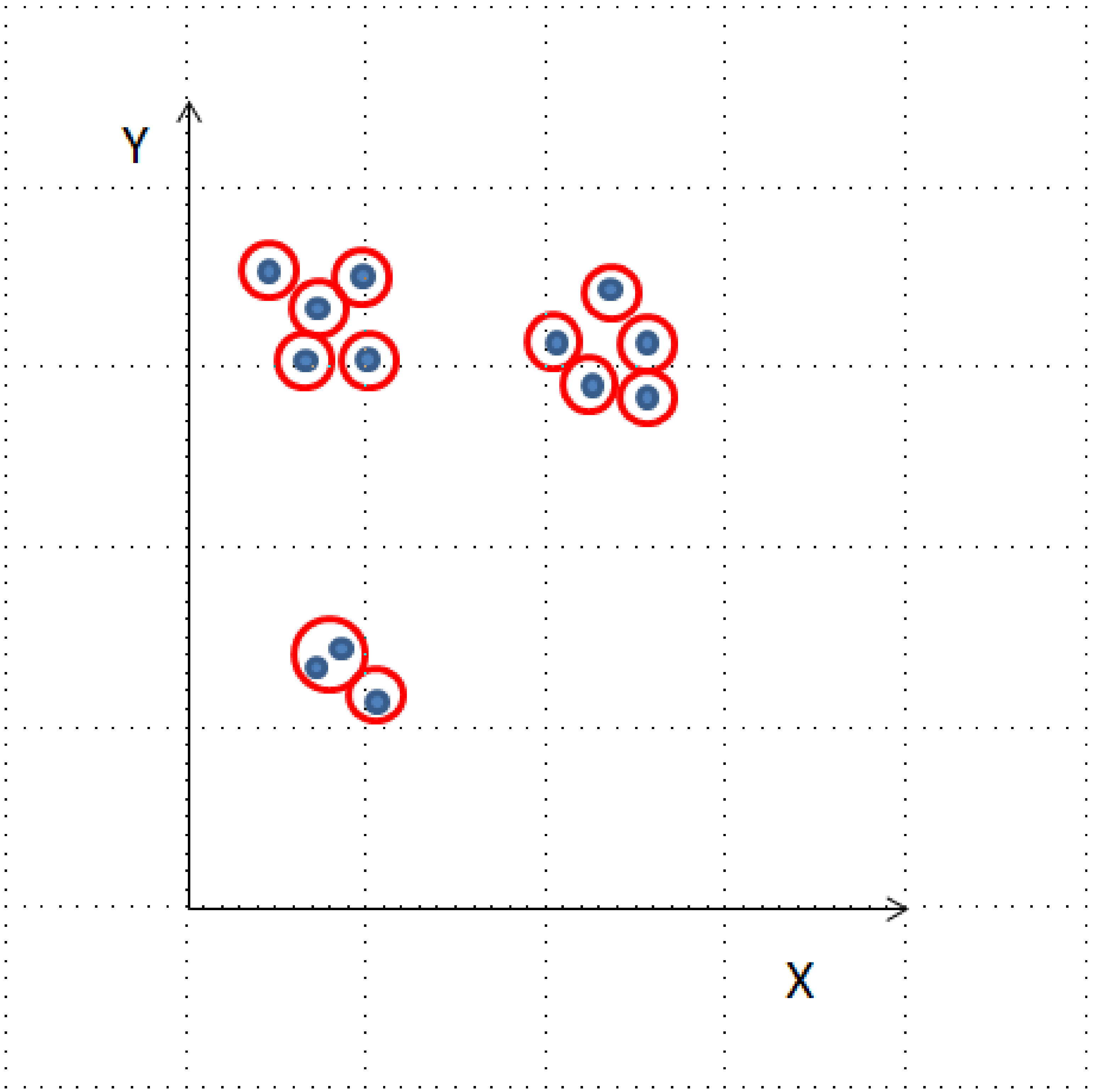


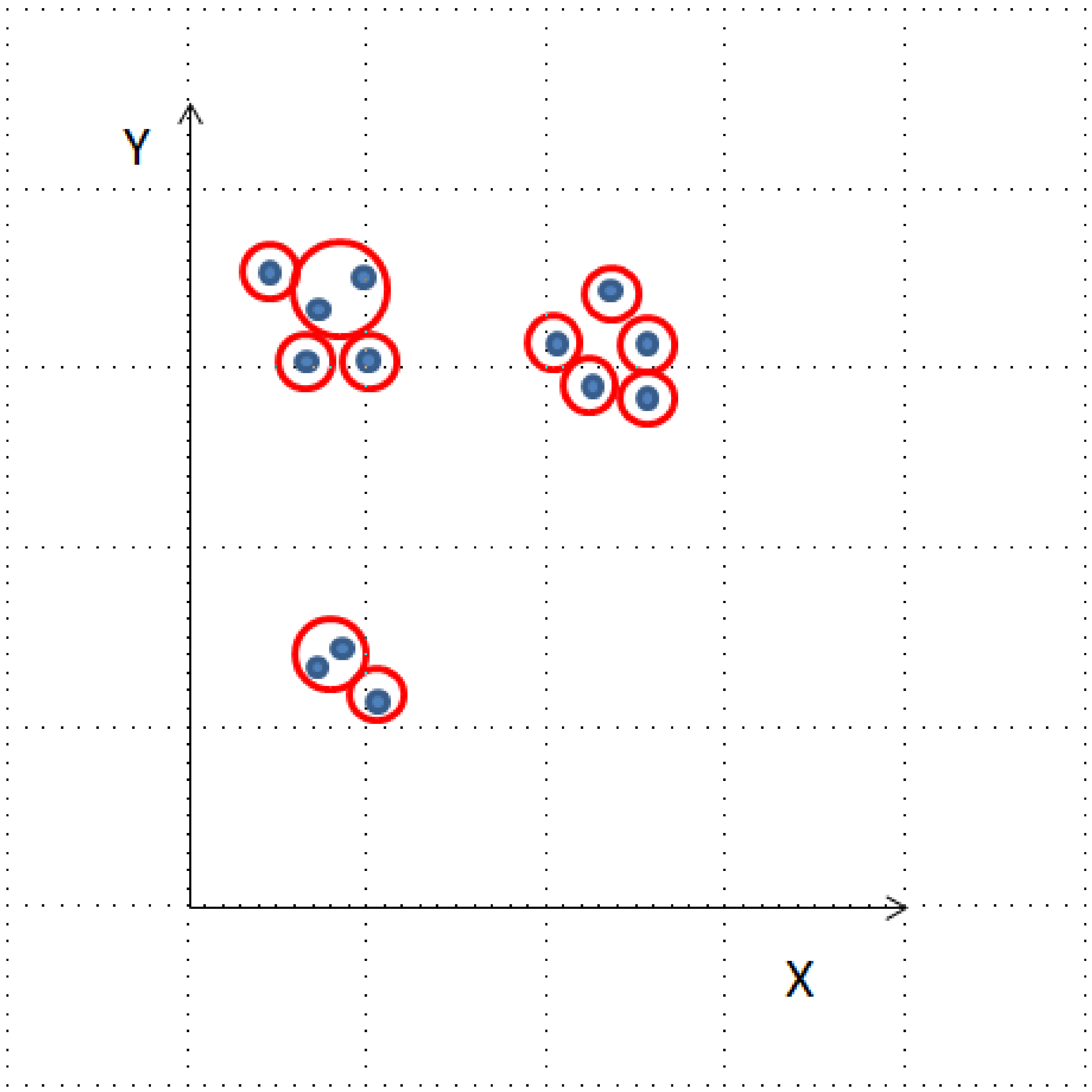
Clustering algorithms

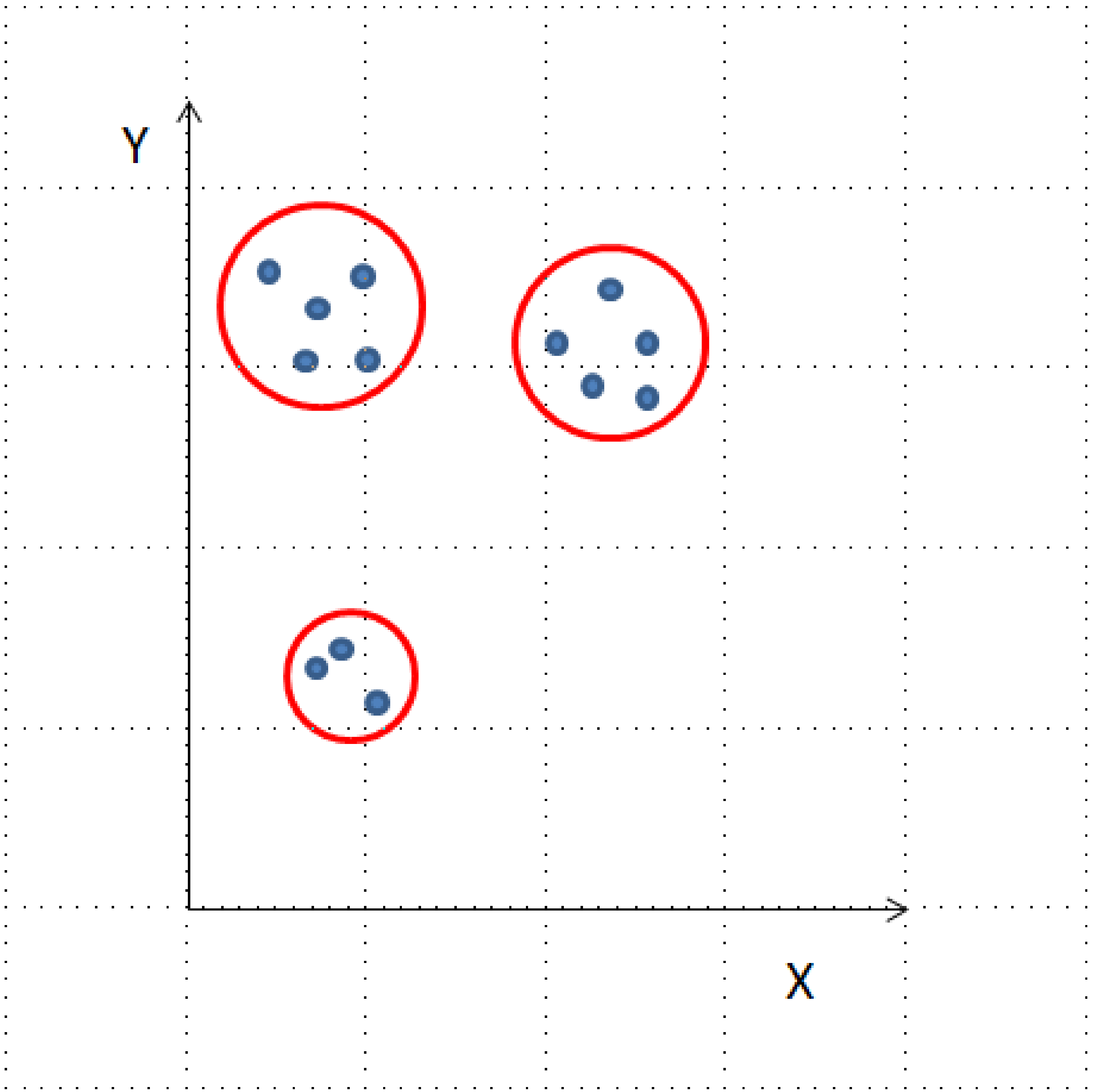
- Hierarchical clustering
- K means clustering
- Other clustering algorithms: DBSCAN, Gaussian Methods


Y
X


Y
X


Y
X


Y
X


Y
X

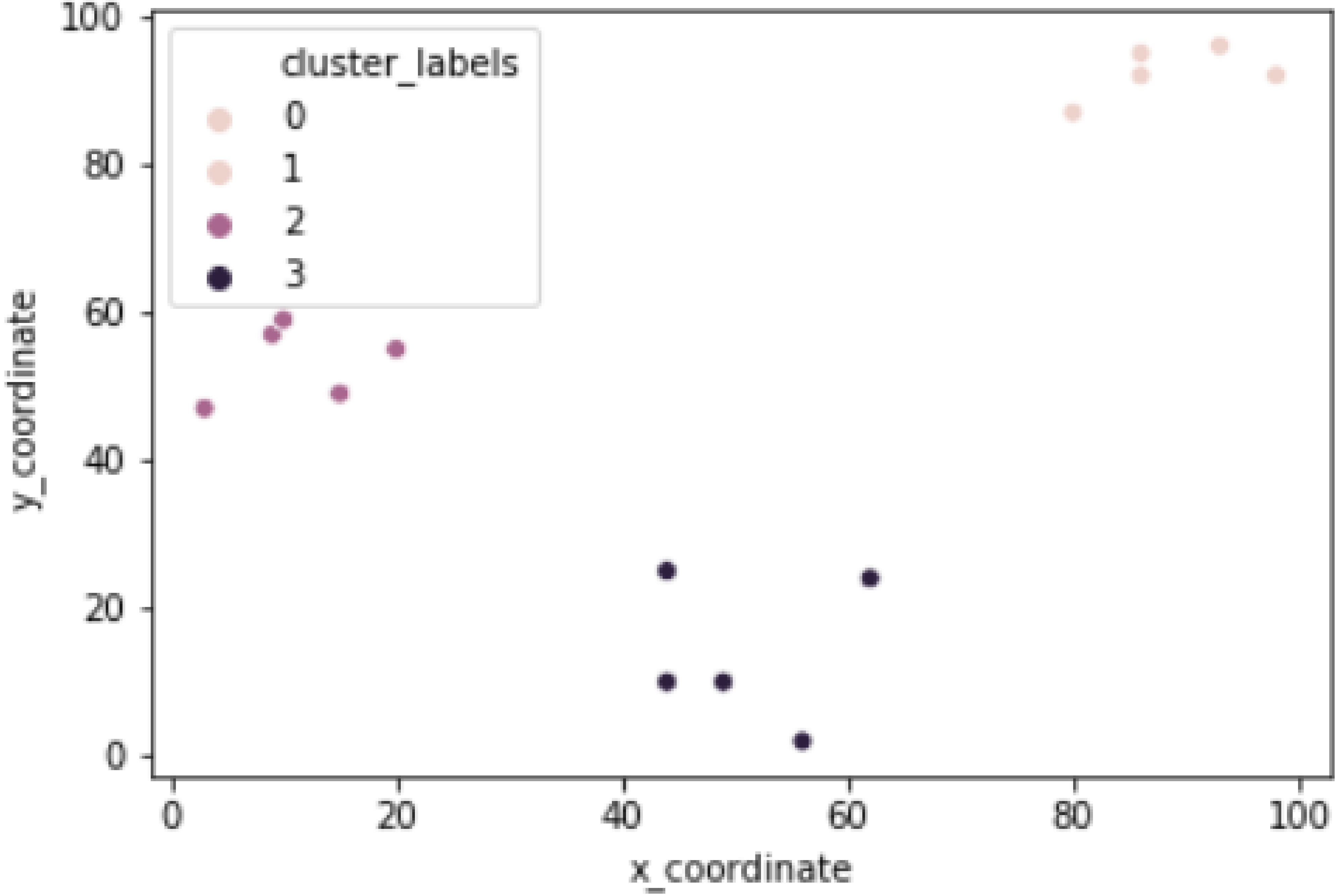
Hierarchical clustering in SciPy

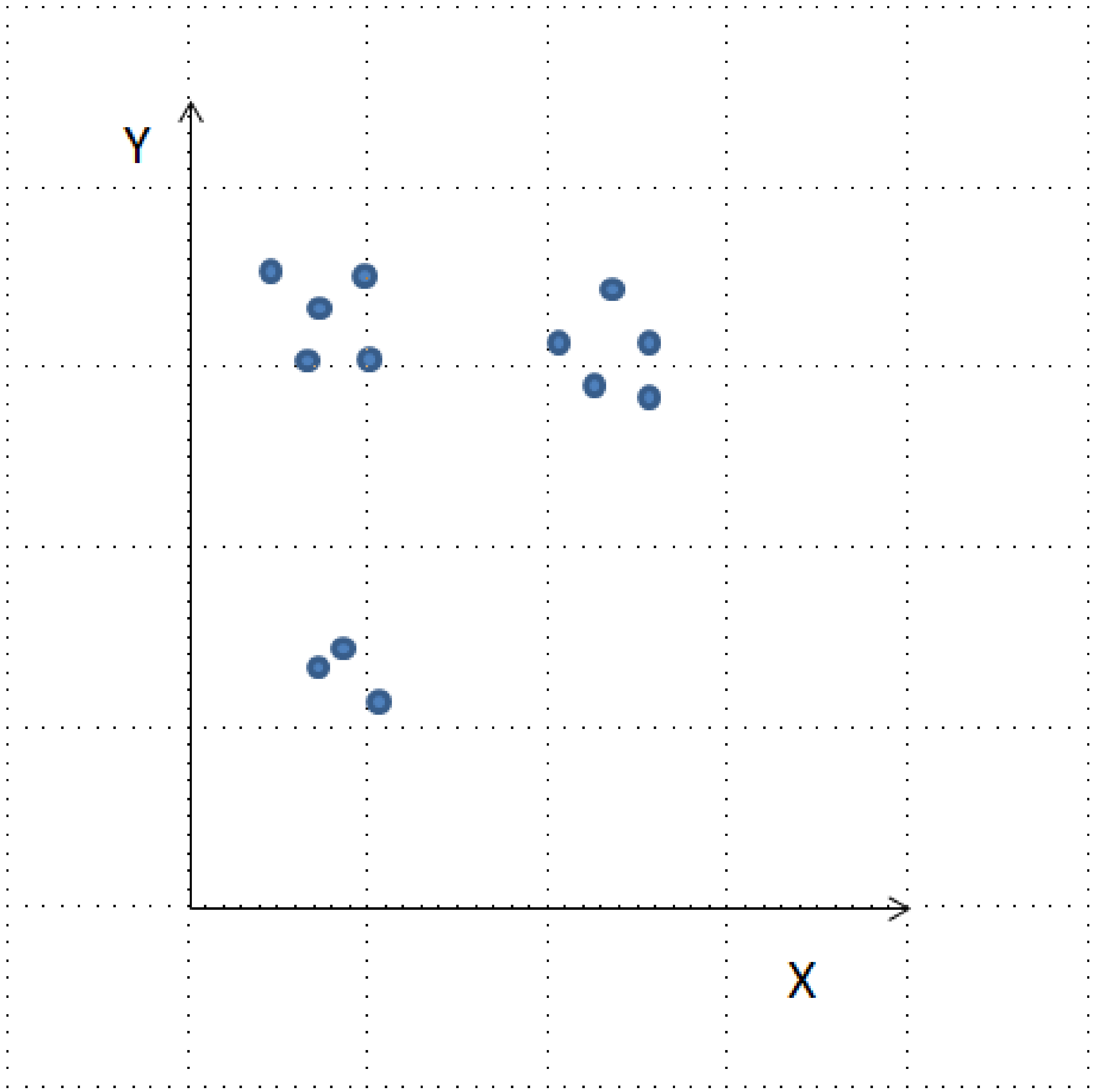
```
from scipy.cluster.hierarchy import linkage, fcluster
from matplotlib import pyplot as plt
import seaborn as sns, pandas as pd
```

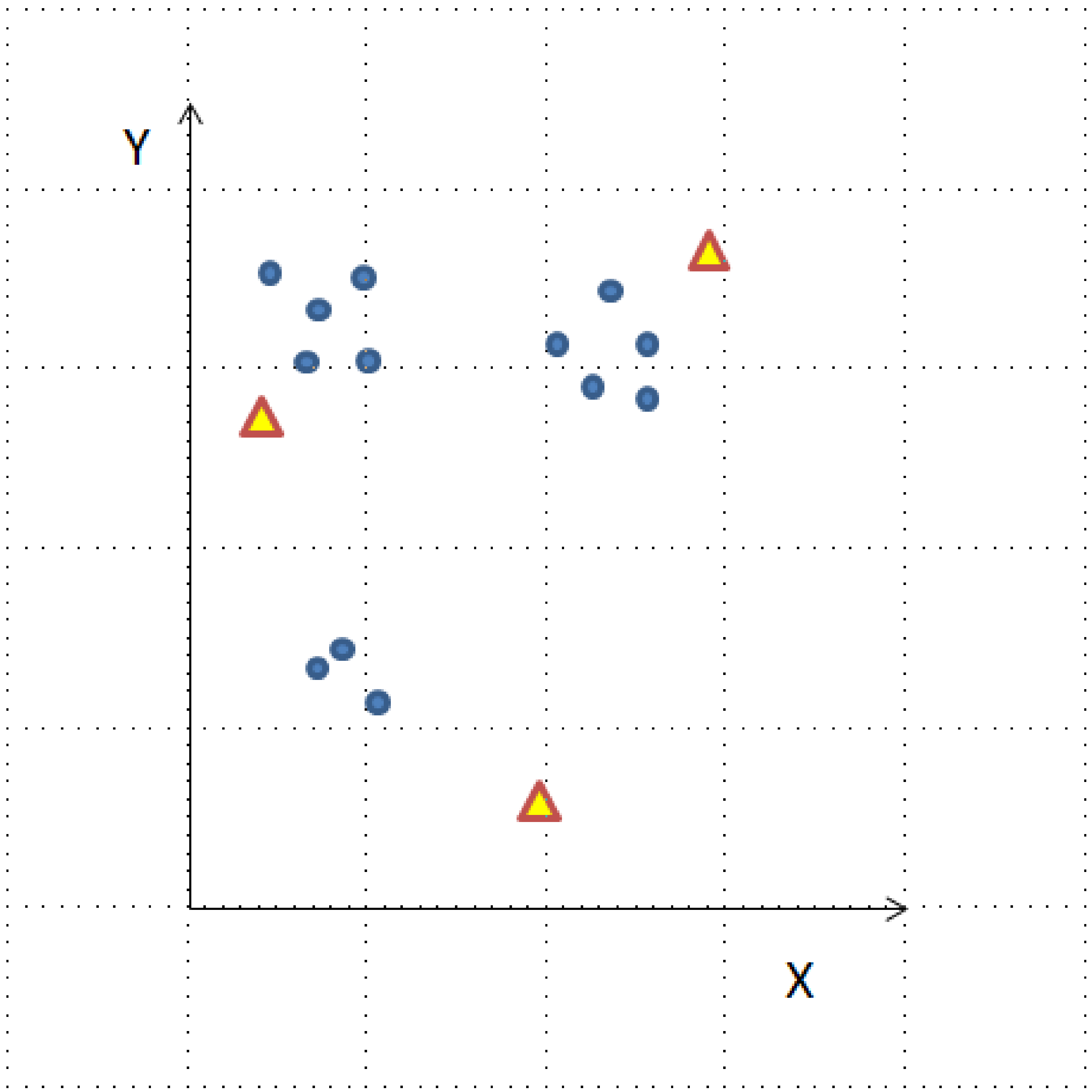
```
x_coordinates = [80.1, 93.1, 86.6, 98.5, 86.4, 9.5, 15.2, 3.4,
                 10.4, 20.3, 44.2, 56.8, 49.2, 62.5, 44.0]
y_coordinates = [87.2, 96.1, 95.6, 92.4, 92.4, 57.7, 49.4,
                 47.3, 59.1, 55.5, 25.6, 2.1, 10.9, 24.1, 10.3]

df = pd.DataFrame({'x_coordinate': x_coordinates,
                   'y_coordinate': y_coordinates})
```

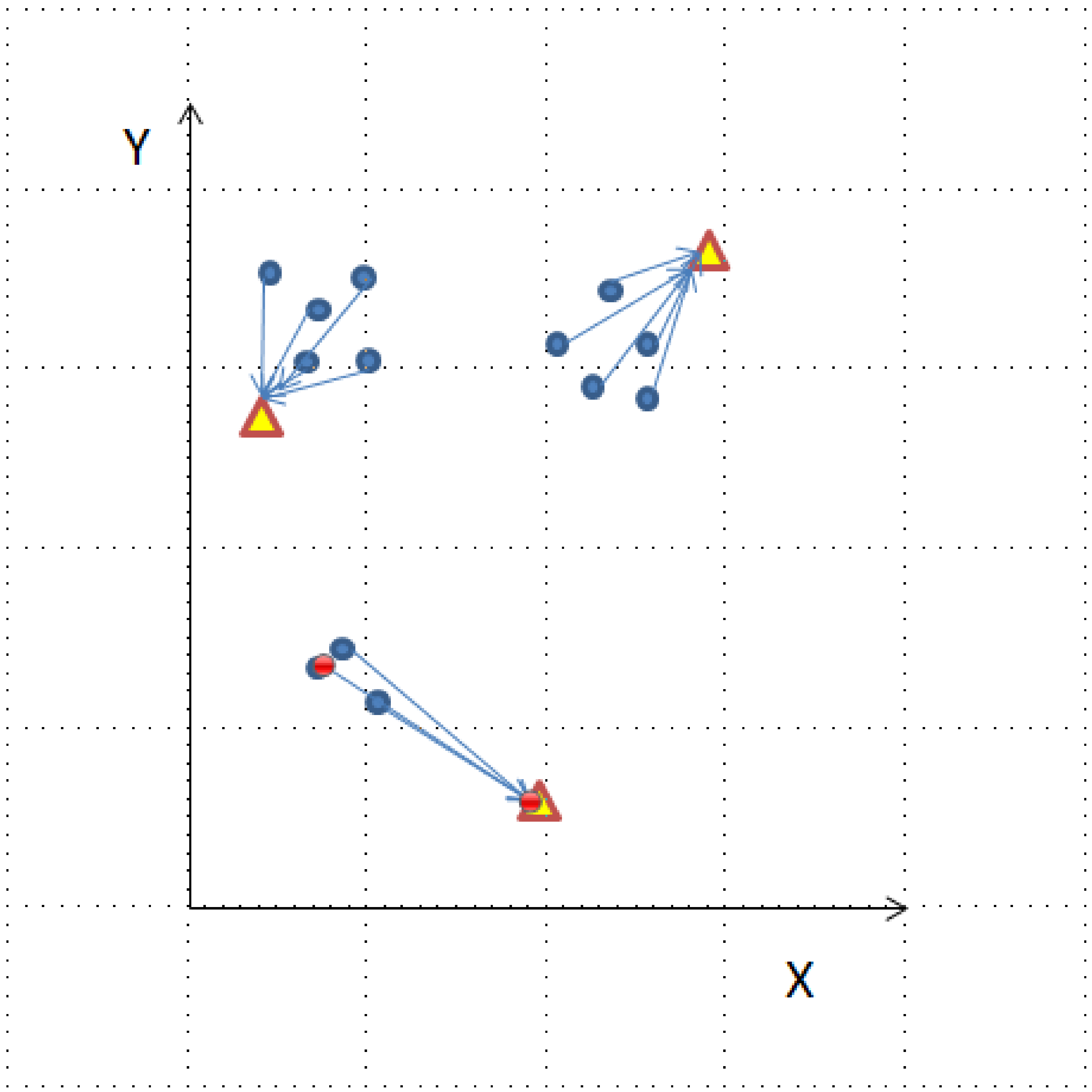
```
Z = linkage(df, 'ward')
df['cluster_labels'] = fcluster(Z, 3, criterion='maxclust')
```

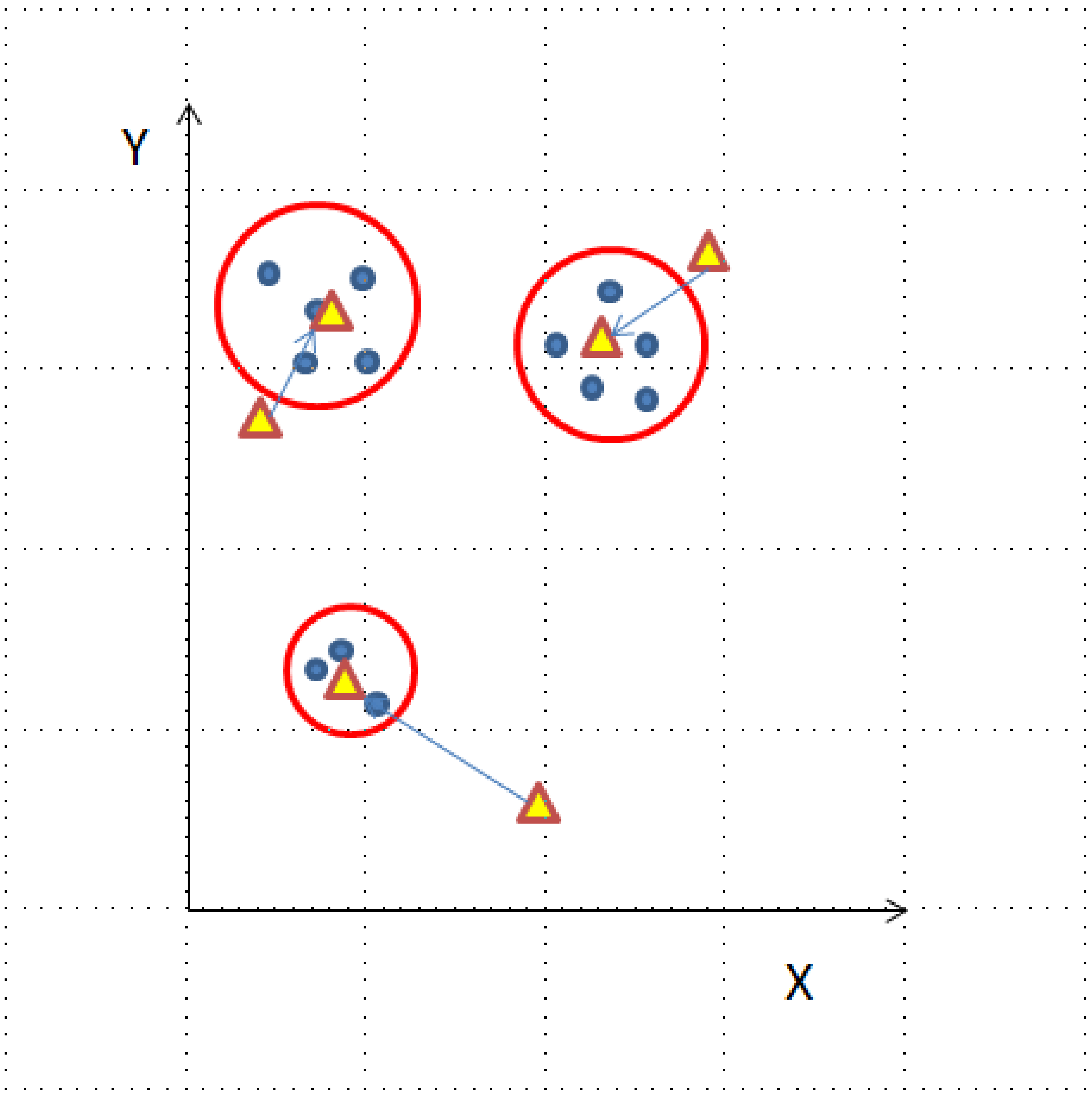
```
sns.scatterplot(x='x_coordinate', y='y_coordinate',
                hue='cluster_labels', data = df)
plt.show()
```


cluster_labels
0
1
2
3
y_coordinate
x_coordinate
0
20
40
60
80
100


Y
X


Y
X


Y
X


Y
X

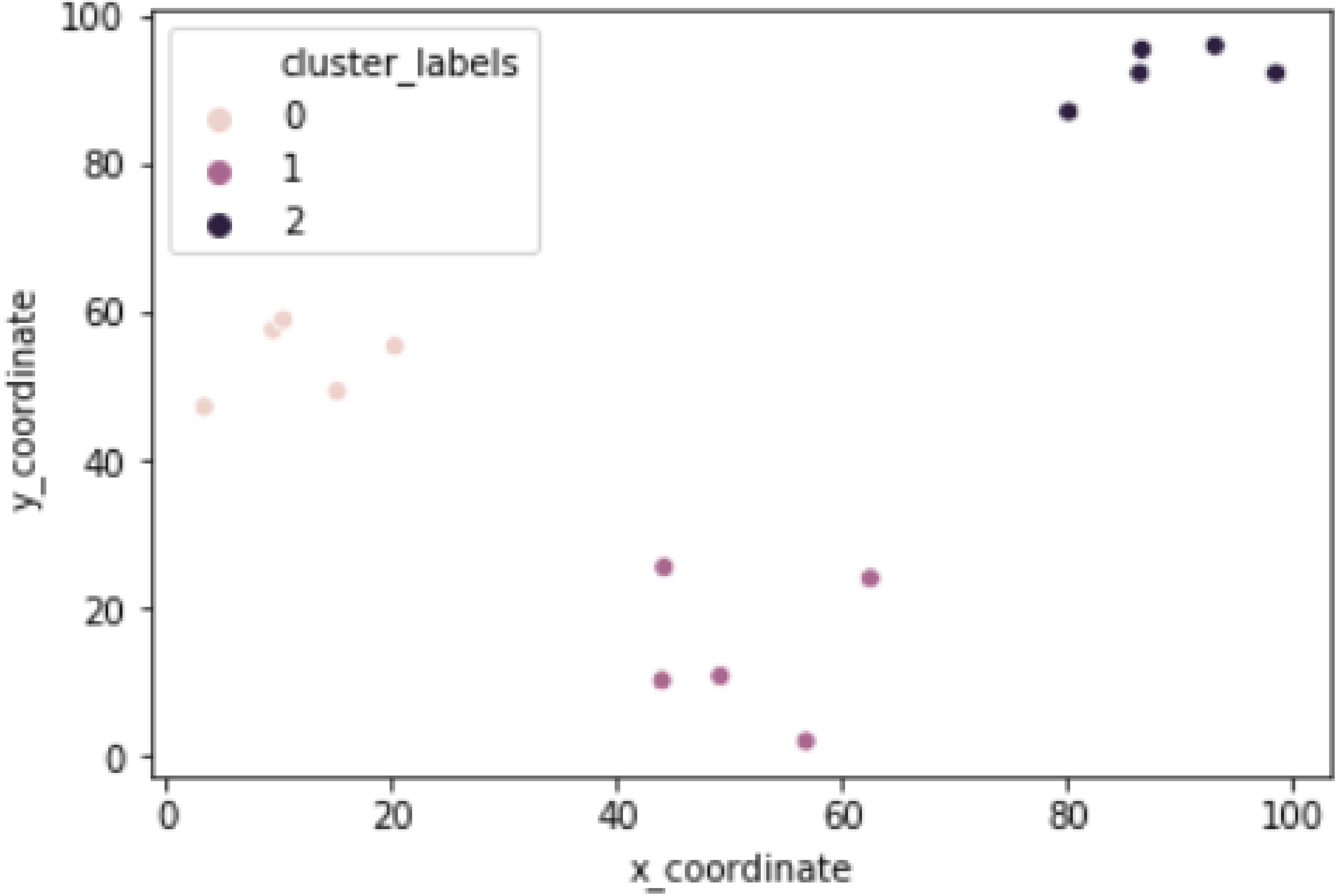
K-means clustering in SciPy

```
from scipy.cluster.vq import kmeans, vq
from matplotlib import pyplot as plt
import seaborn as sns, pandas as pd

import random
random.seed((1000,2000))
```

```
x_coordinates = [80.1, 93.1, 86.6, 98.5, 86.4, 9.5, 15.2, 3.4,
                 10.4, 20.3, 44.2, 56.8, 49.2, 62.5, 44.0]
y_coordinates = [87.2, 96.1, 95.6, 92.4, 92.4, 57.7, 49.4,
                 47.3, 59.1, 55.5, 25.6, 2.1, 10.9, 24.1, 10.3]

df = pd.DataFrame({'x_coordinate': x_coordinates, 'y_coordinate': y_coordinates})
```

```
centroids,_ = kmeans(df, 3)
df['cluster_labels'], _ = vq(df, centroids)
```

```
sns.scatterplot(x='x_coordinate', y='y_coordinate',
                hue='cluster_labels', data = df)
plt.show()
```


cluster_labels
0
1
2
y_coordinate
0
20
40
60
80
100
x_coordinate
0
20
40
60
80
100

Next up: hands-on exercises

CLUSTER ANALYSIS IN PYTHON

Data preparation for cluster analysis

CLUSTER ANALYSIS IN PYTHON

Shaumik Daityari
Business Analyst

datacamp

Why do we need to prepare data for clustering?

- Variables have incomparable units (product dimensions in cm, price in $)
- Variables with same units have vastly different scales and variances (expenditures on cereals, travel)
- Data in raw form may lead to bias in clustering
- Clusters may be heavily dependent on one variable
- Solution: normalization of individual variables

Normalization of data

Normalization: process of rescaling data to a standard deviation of 1

> x_new = x / std_dev(x)

```
from scipy.cluster.vq import whiten
```

```
data = [5, 1, 3, 3, 2, 3, 3, 8, 1, 2, 2, 3, 5]
```

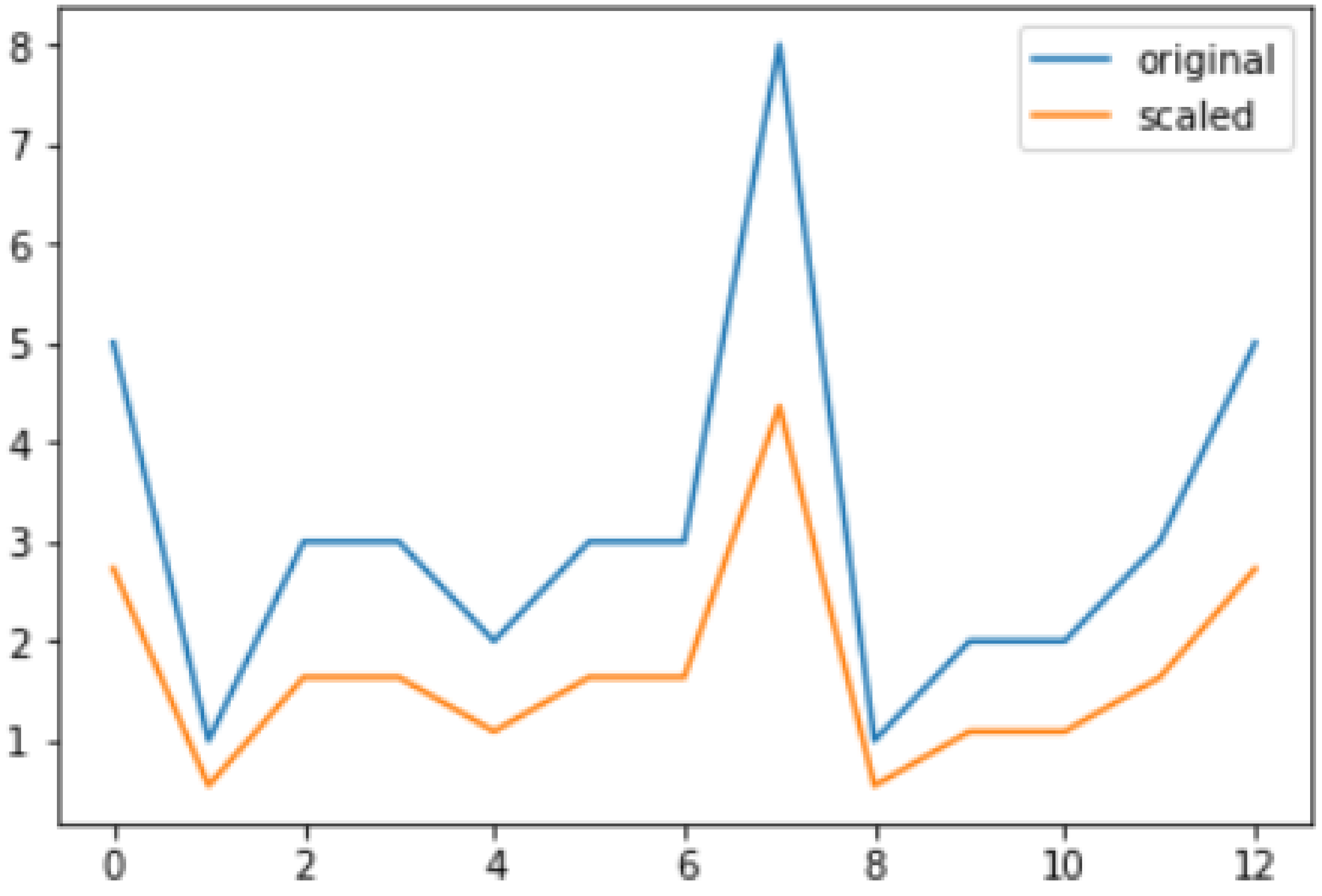
```
scaled_data = whiten(data)
print(scaled_data)
```

```
[2.73, 0.55, 1.64, 1.64, 1.09, 1.64, 1.64, 4.36, 0.55, 1.09, 1.09, 1.64, 2.73]
```

Illustration: normalization of data

```python
# Import plotting library
from matplotlib import pyplot as plt

# Initialize original, scaled data
plt.plot(data,
         label="original")
plt.plot(scaled_data,
         label="scaled")
```

```python
# Show legend and display plot
plt.legend()
plt.show()
```



Next up: some DIY exercises

CLUSTER ANALYSIS IN PYTHON