ISOM 2600 Business Analytics

TOPIC 3: CLUSTERING WITH KMEANS

XUHU WAN

HKUST

JANUARY 15, 2022

Goals for this topic

- Cluster data into different groups using K-means.

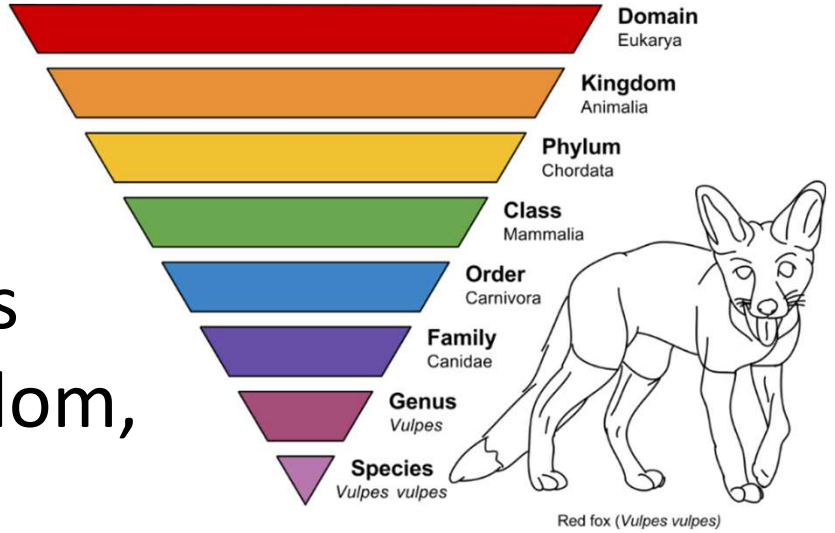
Births of a feather flock together

Introduction

Clustering is the task of partitioning the dataset into groups, called *clusters*.

The goal is to split up the data in such a way that points **within a single cluster** are very **similar** and points in **different clusters** are **different**.

- Application
  - Taxonomy:
    - The most general cluster is domain, followed by kingdom, phylum, etc.



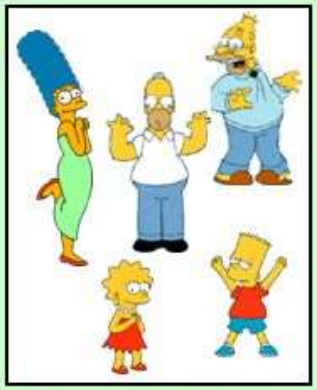
- Marketing:
  - Consumers can be clustered on the basis of their choice of purchases.

- Medicine:
  - Assign patients to specific diagnostic categories on the basis of their presenting symptoms and signs.

- Anthropology:
  - Classify stone tools, shards, or fossil remains by the civilization that produced them.

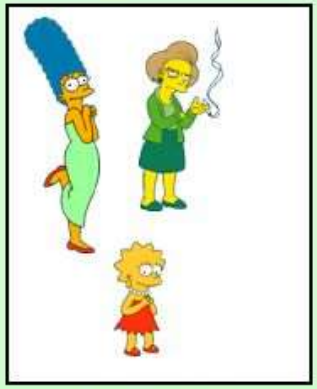
What is a natural grouping among these objects?

Clustering is subjective



Simpson's Family

School Employees



Females

Males

Some Naive Methods for Clustering

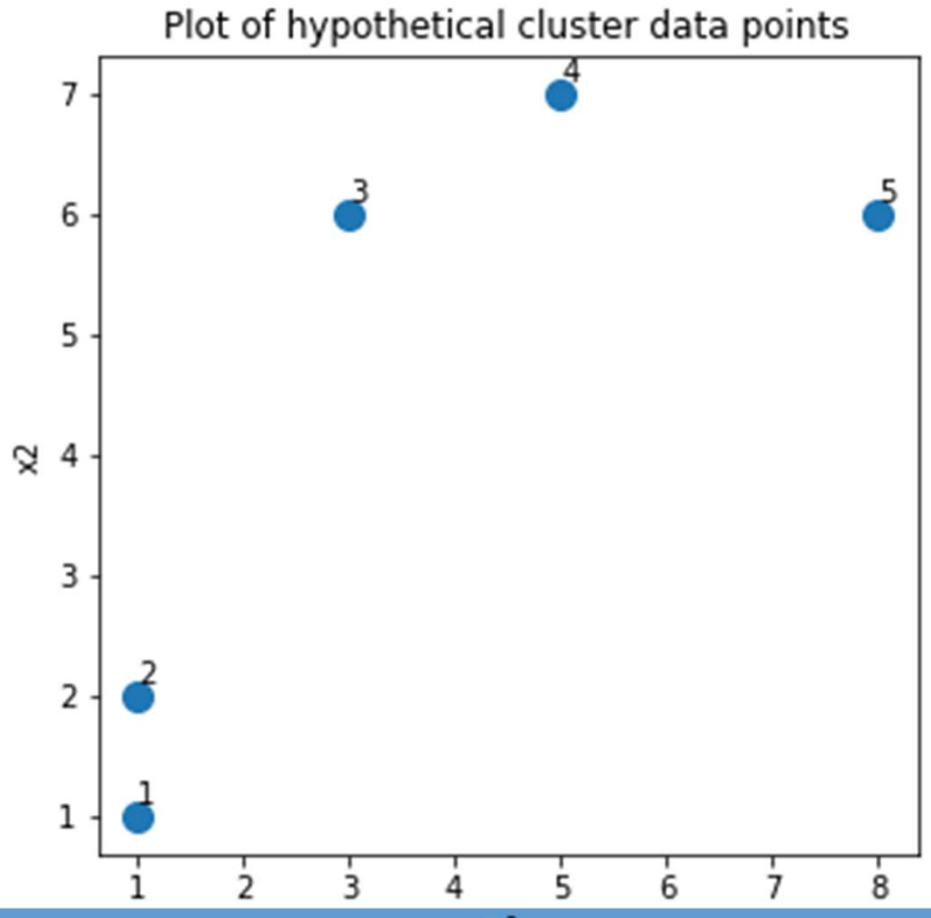
Scatter Plot (/Diagram)

Profile Plot (/Diagram)

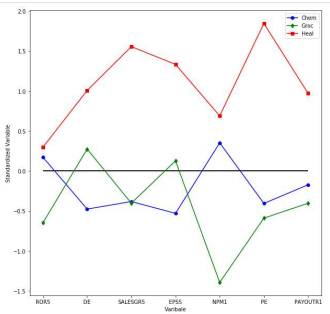
Scatter Plot (/Diagram)

Only used for two variables

- E.g. A hypothetical data set



Profile Plot (/Diagram)



A helpful technique for a **moderate** number of variables

To plot a profile of an individual case in the sample, the investigator customarily first standardizes the data by subtracting the mean and dividing by the standard deviation for each variable

However, the step is omitted by some researchers, especially if the units of the measurement of the variables are comparable

Example - Forbes Financial Data

The financial performance data from January 1981 issue of *Forbes*

Table shows the data for 25 companies from three industries:

- chemical companies (the first 14)
- health care companies (15-19)
- Supermarket companies (20-25)

	TYPE	SYMBOL	OBSNO	ROR5	DE	SALESGR5	EPS5	NPM1	PE	PAYOUTR1
1	Chem	dia	1	13.0	0.7	20.2	15.5	7.2	9	0.426398
2	Chem	dow	2	13.0	0.7	17.2	12.7	7.3	8	0.380693
3	Chem	stf	3	13.0	0.4	14.5	15.1	7.9	8	0.406780
4	Chem	dd	4	12.2	0.2	12.9	11.1	5.4	9	0.568182
5	Chem	uk	5	10.0	0.4	13.6	8.0	6.7	5	0.324544
6	Chem	psm	6	9.8	0.5	12.1	14.5	3.8	6	0.510808
7	Chem	gra	7	9.9	0.5	10.2	7.0	4.8	10	0.378913
8	Chem	hpc	8	10.3	0.3	11.4	8.7	4.5	9	0.481928
9	Chem	mtc	9	9.5	0.4	13.5	5.9	3.5	11	0.573248
10	Chem	acy	10	9.9	0.4	12.1	4.2	4.6	9	0.490798
11	Chem	cz	11	7.9	0.4	10.8	16.0	3.4	7	0.489130
12	Chem	ald	12	7.3	0.6	15.4	4.9	5.1	7	0.272277
13	Chem	rom	13	7.8	0.4	11.0	3.0	5.6	7	0.315646
14	Chem	rei	14	6.5	0.4	18.7	-3.1	1.3	10	0.384000
15	Heal	hum	15	9.2	2.7	39.8	34.4	5.8	21	0.390879
16	Heal	hca	16	8.9	0.9	27.8	23.5	6.7	22	0.161290
17	Heal	nme	17	8.4	1.2	38.7	24.6	4.9	19	0.303030
18	Heal	ami	18	9.0	1.1	22.1	21.9	6.0	19	0.303318
19	Heal	ahs	19	12.9	0.3	16.0	16.2	5.7	14	0.287500
20	Groc	lks	20	15.2	0.7	15.3	11.6	1.5	8	0.598930
21	Groc	win	21	18.4	0.2	15.0	11.6	1.6	9	0.578313
22	Groc	sgl	22	9.9	1.6	9.6	24.3	1.0	6	0.194946
23	Groc	slc	23	9.9	1.1	17.9	15.3	1.6	8	0.321070
24	Groc	kr	24	10.2	0.5	12.6	18.0	0.9	6	0.453731
25	Groc	sa	25	9.2	1.0	11.6	4.5	0.8	7	0.594966

Background of these companies

- Among the chemical companies, all of the large diversified firms were selected.
- From the major supermarket chains, the top six rated for return on equity were included.
- In the health care industry, four of the five companies included were those connected with hospital management; the remaining company involves hospital supplies and equipment (company 19).

Details of the variables

- P/E: price-to-earnings ratio, which is the price of one share of common stock divided by the earnings per share for the last year. The ratio shows the dollar amount investors are willing to pay for the stock per dollar of current earnings of the company.
- ROR5: percent rate of return on total capital (invested capital plus debt) averaged over the past five years.
- D/E: debt-to-equity (invested capital) ratio for the last year. This ratio indicates the extent to which management is using borrowed funds to operate the company.
- SALESGR5: percent annual compound growth rate of sales, computed from the most recent five years compared with the previous five years.

- EPS5: percent annual compound growth in earnings per share, computed from the most recent five years compared with the previous five years.
- NPM1: percent net profit margin, which is the net profits divided by the net sales for the past year, expressed as a percentage.
- PAYOUTR1: annual dividend divided by the latest 12-month earnings per share. This value represents the proportion of earnings paid out to shareholders rather than to operate and expand the company.

Data preprocessing

- Increase the index by 1 to be consistent with the company #
- Standardized the variables

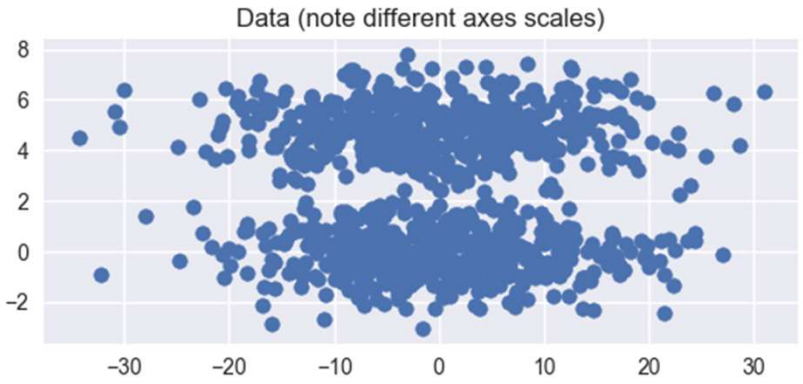
```
Forbes = pd.read_csv('Cluster.csv')
Forbes.index += 1
Forbes.head()
```

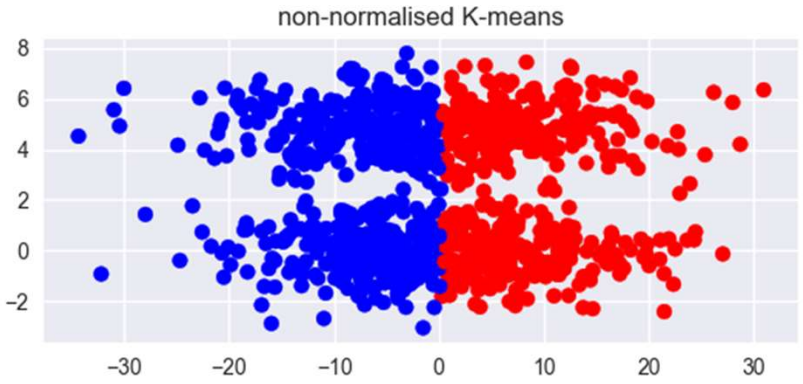
	TYPE	SYMBOL	OBSNO	ROR5	DE	SALESGR5	EPS5	NPM1	PE	PAYOUTR1
1	Chem	dia	1	13.0	0.7	20.2	15.5	7.2	9	0.426398
2	Chem	dow	2	13.0	0.7	17.2	12.7	7.3	8	0.380693
3	Chem	stf	3	13.0	0.4	14.5	15.1	7.9	8	0.406780
4	Chem	dd	4	12.2	0.2	12.9	11.1	5.4	9	0.568182
5	Chem	uk	5	10.0	0.4	13.6	8.0	6.7	5	0.324544

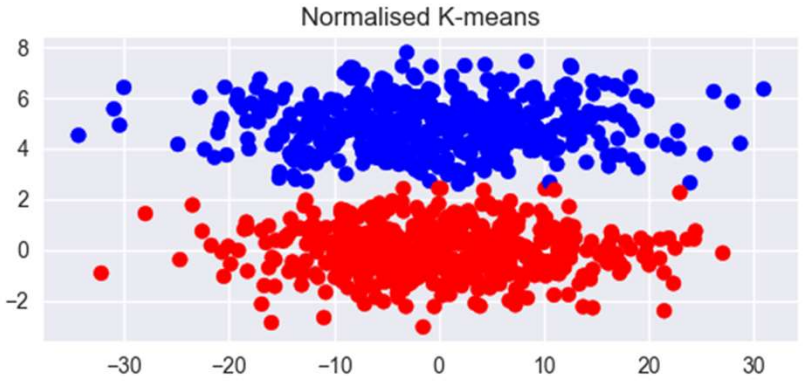
```
# standardization of the continuous variables
X = Forbes.iloc[:, 3:].values
# import the function StandardScaler
from sklearn.preprocessing import StandardScaler
# fit_transform means to fit the model by the data and then transform
# the data by the fitted model
X_scaled = StandardScaler().fit_transform(X)
```

	ROR5	DE	SALESGR5	EPS5	NPM1	PE	PAYOUTR1
1	0.982788	-0.007511	0.437915	0.283215	1.315877	-0.242222	0.153676
2	0.982788	-0.007511	0.051519	-0.058008	1.361314	-0.451035	-0.221020
3	0.982788	-0.570864	-0.296236	0.234469	1.633941	-0.451035	-0.007155
4	0.674221	-0.946433	-0.502314	-0.252993	0.497997	-0.242222	1.316042
5	-0.174341	-0.570864	-0.412155	-0.630776	1.088688	-1.077472	-0.681338

Why Standardization/Normalization is important in clustering?



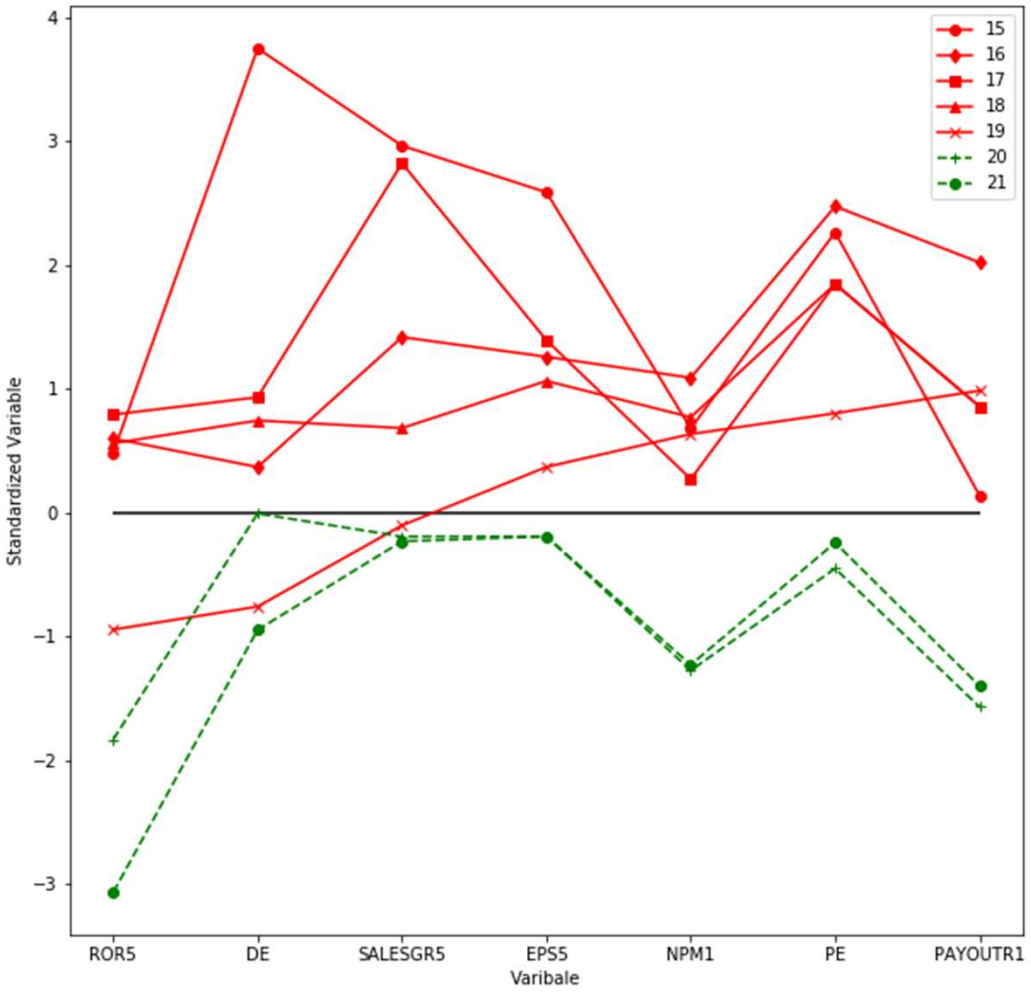




All clustering methods will group data points with the same principal

1. Individuals of the same group will be similar
2. Individuals from different groups should have big differences

In the second diagram, the data is not standardized so clustering method will split vertically because it can improve the similarity in the same group comparing to the groups horizontally cut.



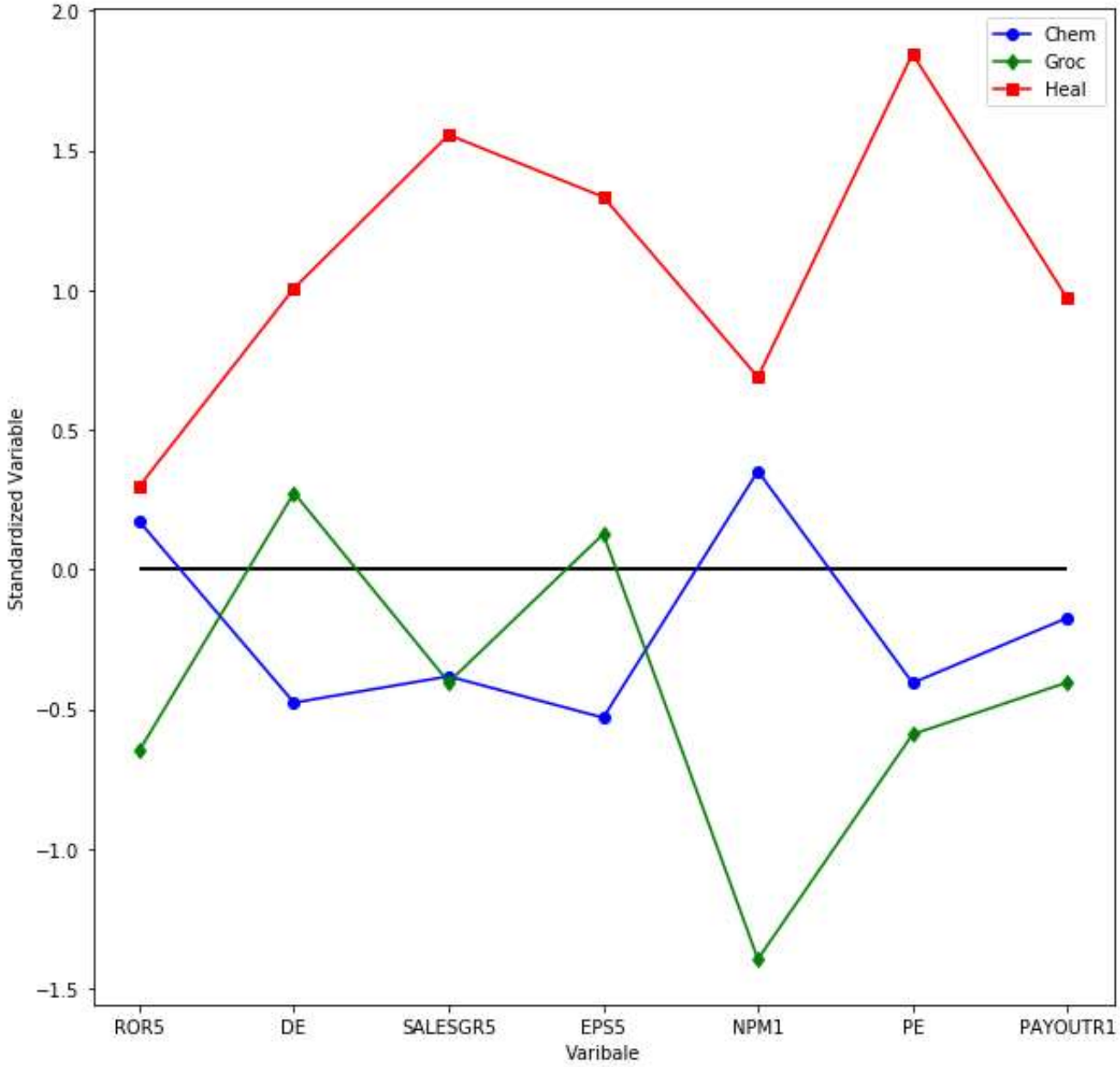
Conclusions based on the profile diagram

- Companies 20 and 21 are very similar.
- Companies 15, 16, 17 and 18 are very similar.
- Company 19 stands alone.

The clusters are consistent with the type of companies, especially noting that company 19 deals with hospital supplies.

Now we visualize the mean profile for these three types of companies.

TYPE	ROR5	DE	SALESGR5	EPS5	NPM1	PE	PAYOUTR1
Chem	0.171586	-0.476972	-0.382715	-0.530672	0.351948	-0.406289	-0.173453
Groc	-0.648507	0.274165	-0.403568	0.126821	-1.395241	-0.590243	-0.404496
Heal	0.297768	1.006524	1.555885	1.333696	0.688836	1.845902	0.971065


Chem
Groc
Heal
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
Standardized Variable
ROR5
DE
SALESGR5
EPS5
NPM1
PE
PAYOUTR1
Varibale

Two Simple Clustering Methods

For a large dataset, analytical methods such as K-means and Hierarchical clustering are necessary.

K-means Clustering

Hierarchical Clustering

Distance Measures

Both methods require defining some measure of **closeness** or **similarity** of two observations.

- Distance measure: the smaller the closer of the two data point.
- Similarity measure: the larger the closer of the two data point.

The converse of similarity is **distance**.

Some commonly used distance

- Euclidian distance
- Manhattan distance
- Correlation based measure

Some commonly used distance

1. Euclidian distance:

- The square root of the sum of the squared differences between coordinates of each variable for two observations:

- $Distance = \sqrt{\sum_{j=1}^{p}(a_j - b_j)^2}$

$for\ \boldsymbol{a} = (a_1, \ldots, a_p)$
$and\ \boldsymbol{b} = (b_1, \ldots, b_p).$

- **2. Manhattan distance (/city block distance)** (Optional):
  - Distance = $\sum_{i=1}^{p}|a_i - b_i|$, for $\boldsymbol{a} = (a_1, \dots, a_p)$ and $\boldsymbol{b} = (b_1, \dots, b_p)$.

3. Correlation based measure (Optional):

- Distance = $1 - r_{ab}$, where r_{ab} is the Pearson sample correlation between point $\boldsymbol{a} = (a_1, \ldots, a_p)$ and point $\boldsymbol{b} = (b_1, \ldots, b_p)$

$$r_{ab} = \sum_{i=1}^{p}(a_i - \bar{a})(b_i - \bar{b}) \Big/ \sqrt{\sum_{i=1}^{p}(a_i - \bar{a})^2 \sum_{i=1}^{p}(b_i - \bar{b})^2}.$$

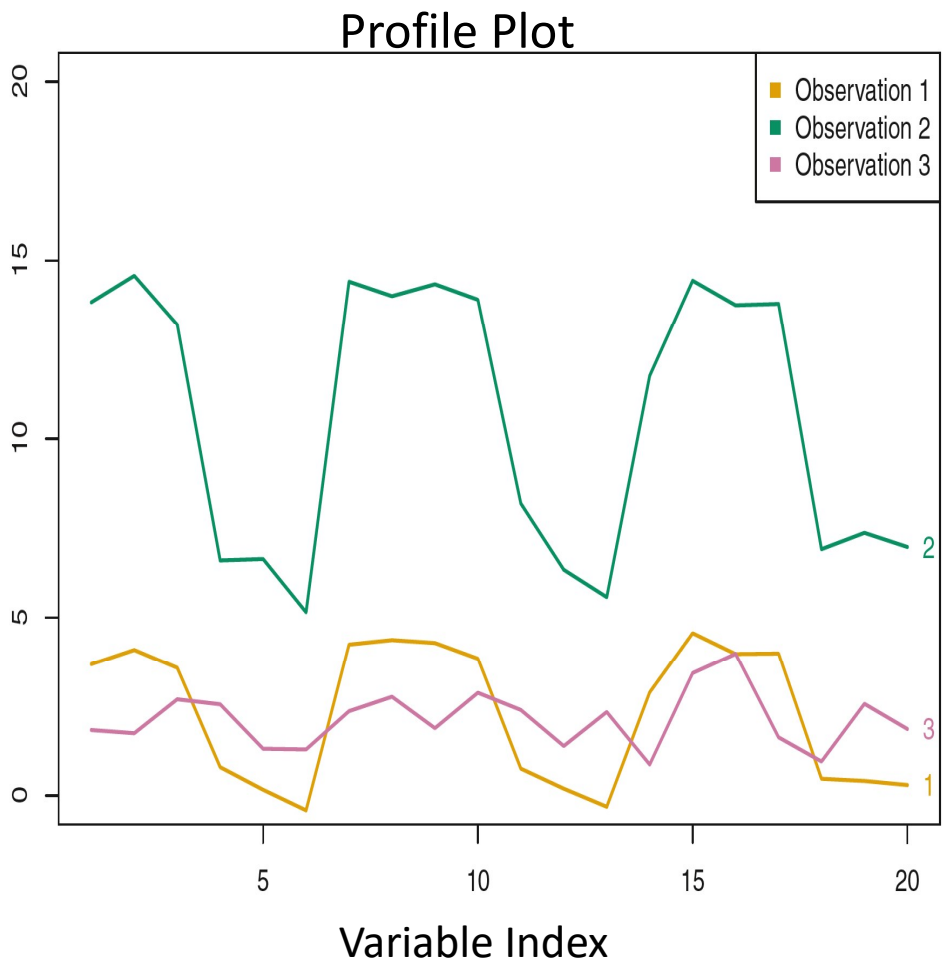
In most situations, different distance measures lead to different clusters.

When the variables have different units, it is advisable to standardize the data before computing the distances.

Issue 1: distance versus correlation

From Euclidean ‘distance’, Observations 1 and 3 are closed

Based on ‘correlation’ based distance, observations 1 and 2 are closed.



Issue 2: standardization

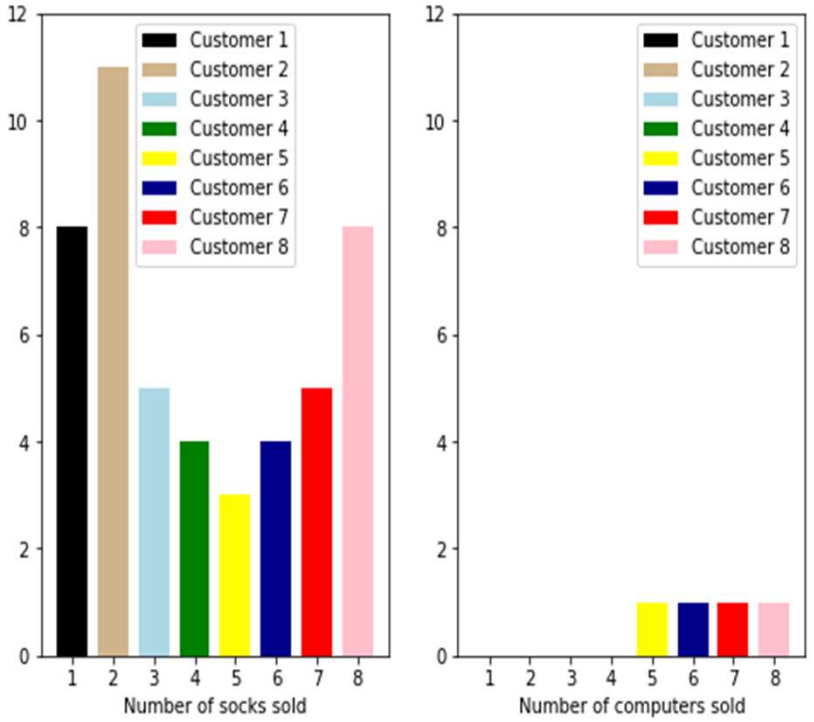
An electronic online retailer sells two items: socks and computers.

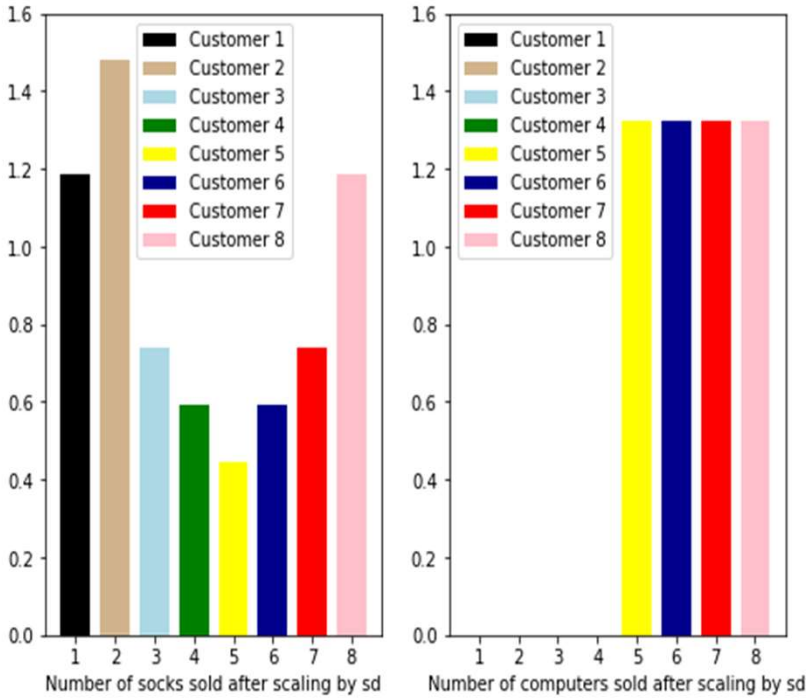
In the data set, we have n=8 customers on two variables.

The choice of scaling on the variables has dramatic effect on the clustering result (here we use Euclidean distance).

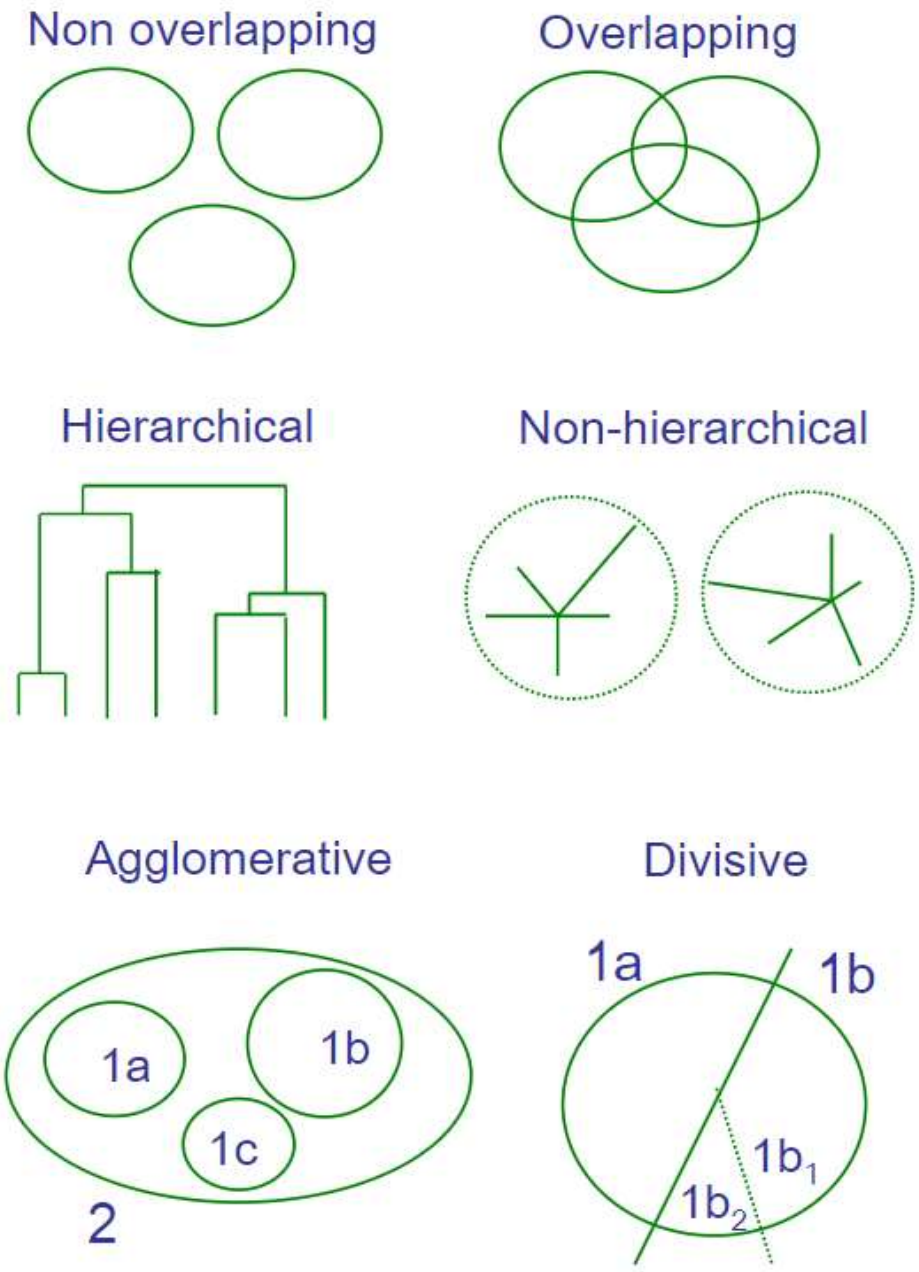
Left: the y-axis represents the number of pairs of socks, and computers sold

Right: the same data is shown, after scaling each variable by its standard deviation

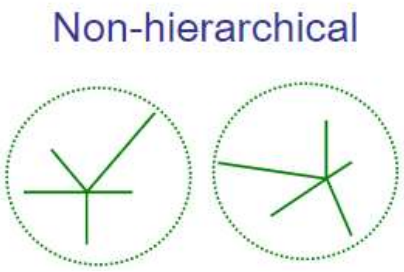




Different Type of Clustering



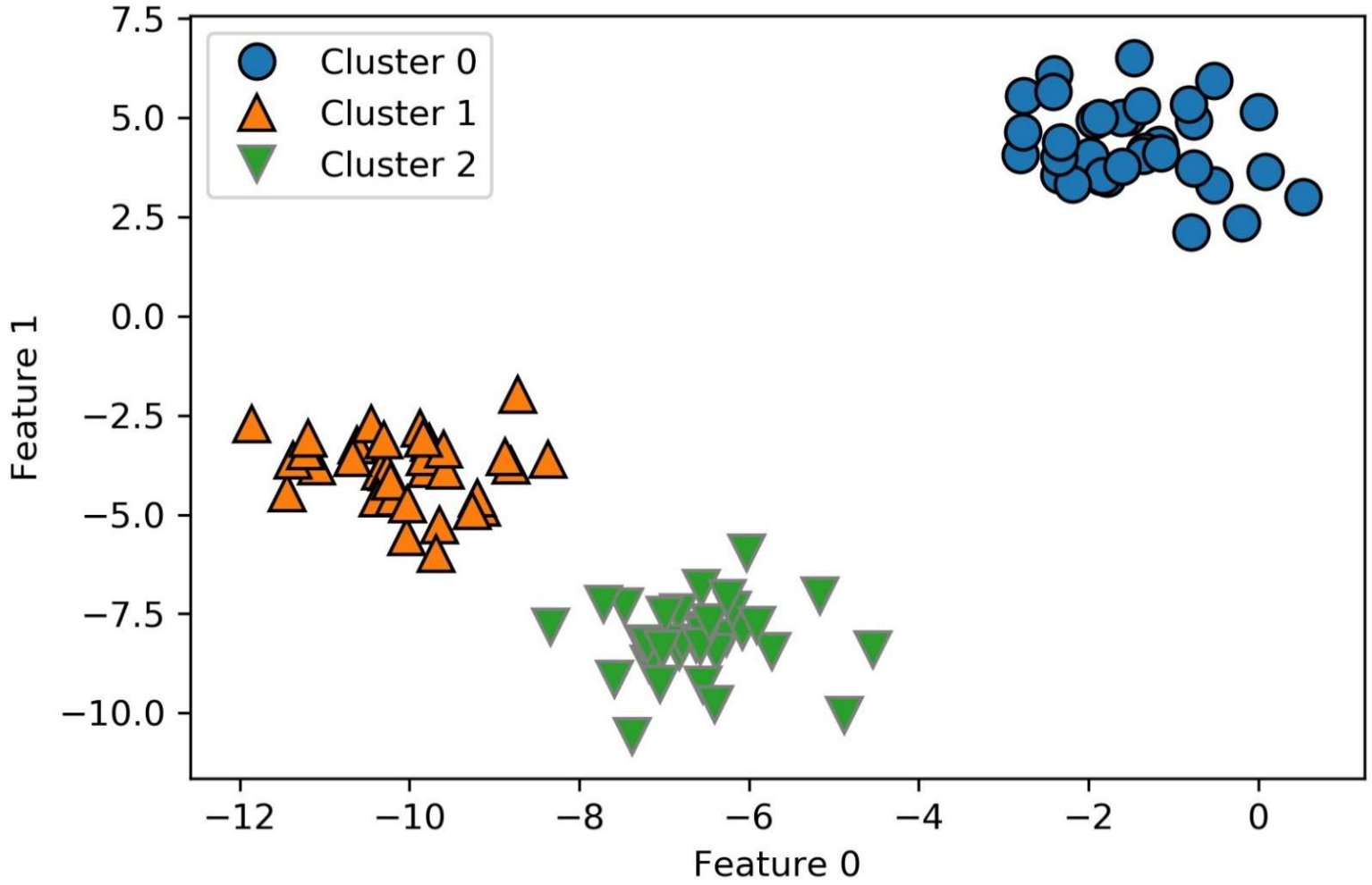
K-means



K-means clustering is a popular nonhierarchical clustering technique. For a specified number of clusters K, the basic algorithm proceeds in the following steps.

- 1. Divide the data into K initial clusters. The number of these clusters may be specified by the user or may be selected by some criteria.
- 2. Calculate the **centroids**, i.e. means, of the K clusters.
- 3. For a given case(/data point), calculate its distance to each centroid. If the case is closest to the centroid of its own cluster, leave it in that cluster; otherwise, reassign it to the cluster whose centroid is closest to it.
- 4. Repeat step 3 for each case.
- 5. Repeat steps 2, 3, and 4 until no cases are reassigned.

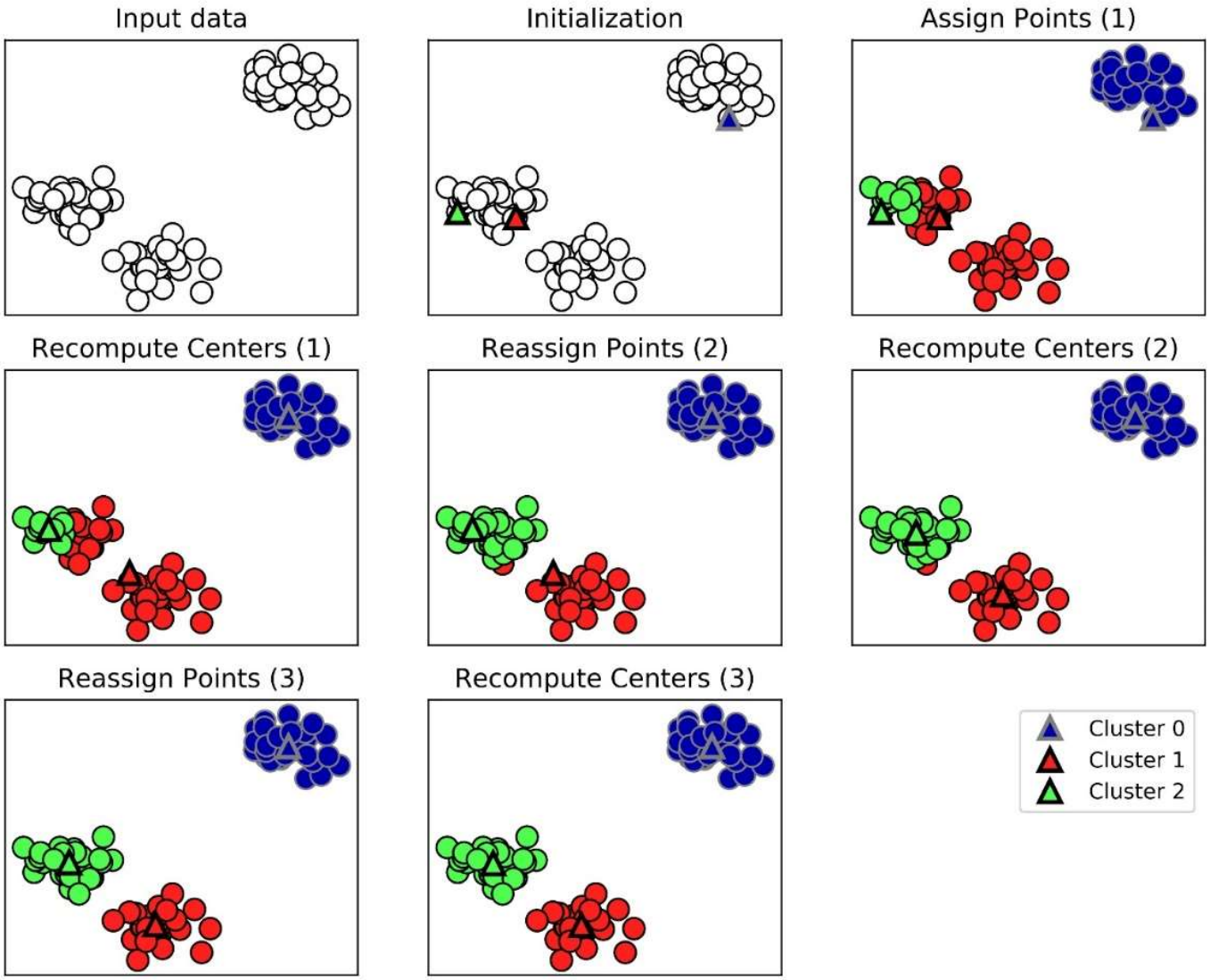
K-means - A simulation study



K-means - A simulation study

(continued)

- Apply K-means with K=3

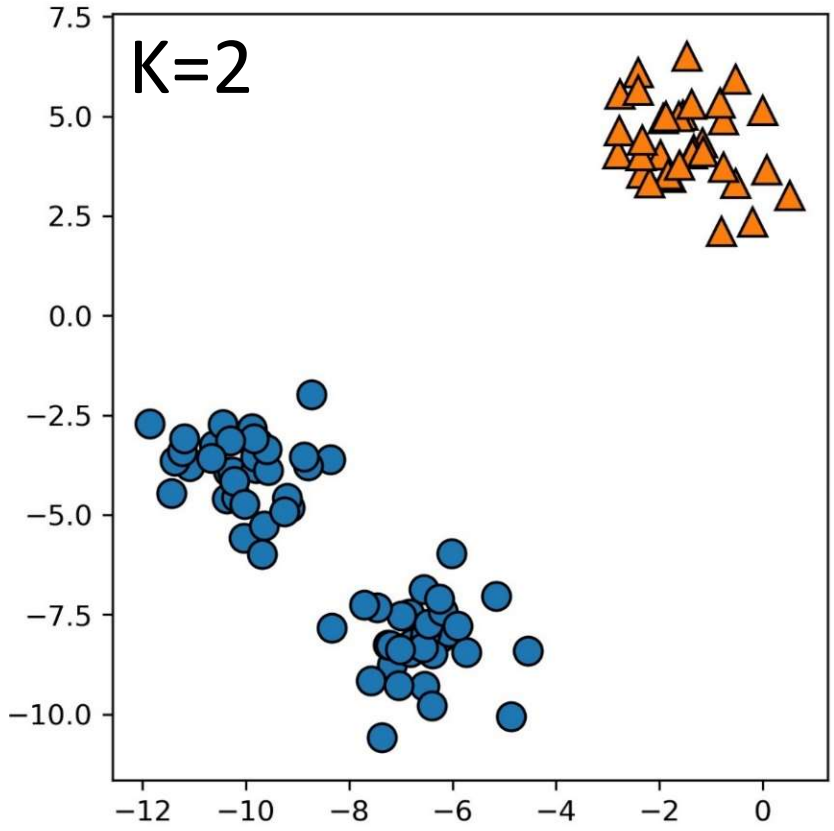


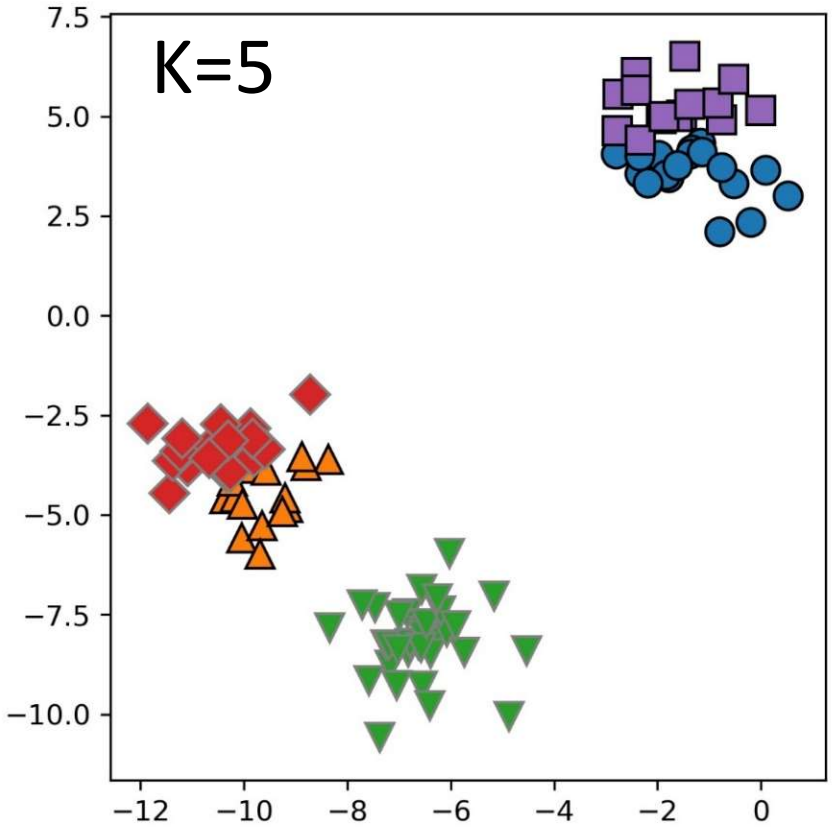
K-means - A simulation study

(continued)

Misspecification of the K value

```
kmeans = KMeans(n_clusters=2)
kmeans.fit(X)
assignments = kmeans.labels_
mglearn.discrete_scatter(X[:, 0], X[:, 1], assignments, ax=axes[0])
```





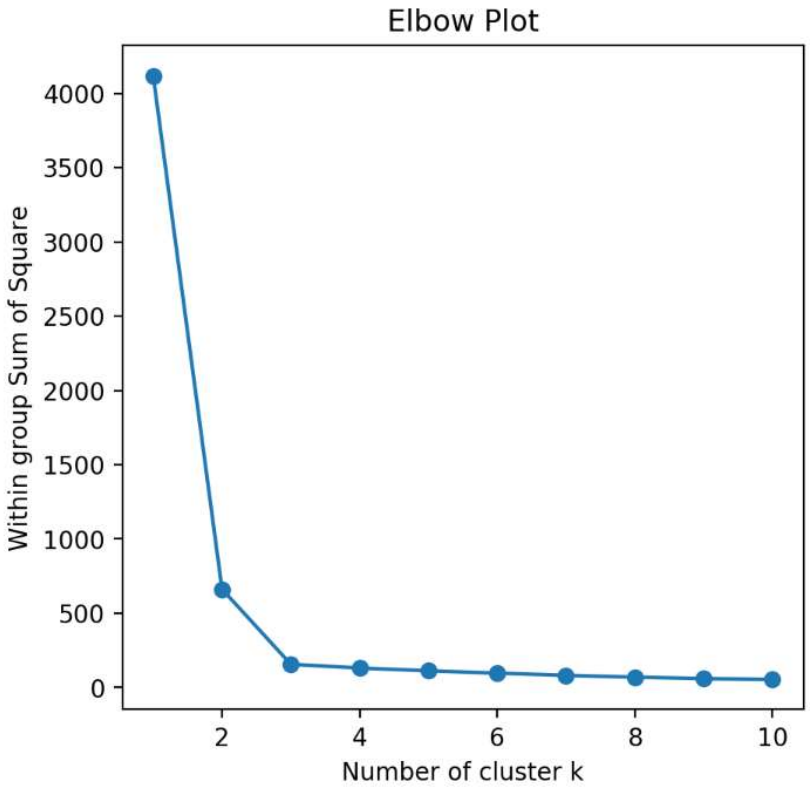
How to Choose K?
- Elbow Plot

1. Apply clustering algorithm (e.g., k-means clustering) for different values of k. For instance, by varying k from 1 to 10 clusters.

2. For each k, calculate the total within-cluster sum of square (WSS).

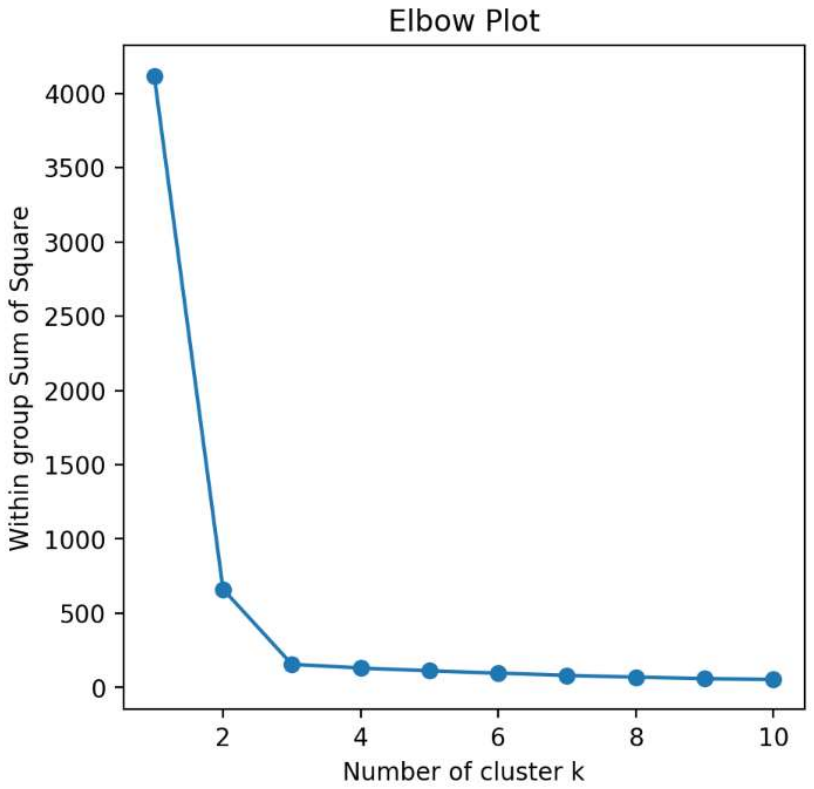
3. Plot the curve of total WSS according to the number of clusters k.

4. The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters.



$$Total\ WSS = \sum_{j=1}^{k} \sum_{x_i \in C_j} \left(d(x_i, \mu_j)\right)^2$$

```
wss= []
for i in range(1, 11):
    kmeans = KMeans(n_clusters = i)
    kmeans.fit(X)
    wss.append(kmeans.inertia_)
```




```
plt.figure(figsize=(5,5),dpi=200)
plt.plot(range(1, 11), wss, '-o')
plt.title('Elbow Plot')
plt.xlabel('Number of cluster k')
plt.ylabel('Within group Sum of Square')
plt.show()
```

Take away from Topic 3

Clustering Method

- Profile diagram/plot
- K-means
  - Elbow plot

Distance measure

- Euclidian distance
- Manhattan distance
- Correlation based distance