ISOM 2600 Business Analytics

TOPIC 4: CLUSTERING WITH HIERARCHICAL METHOD

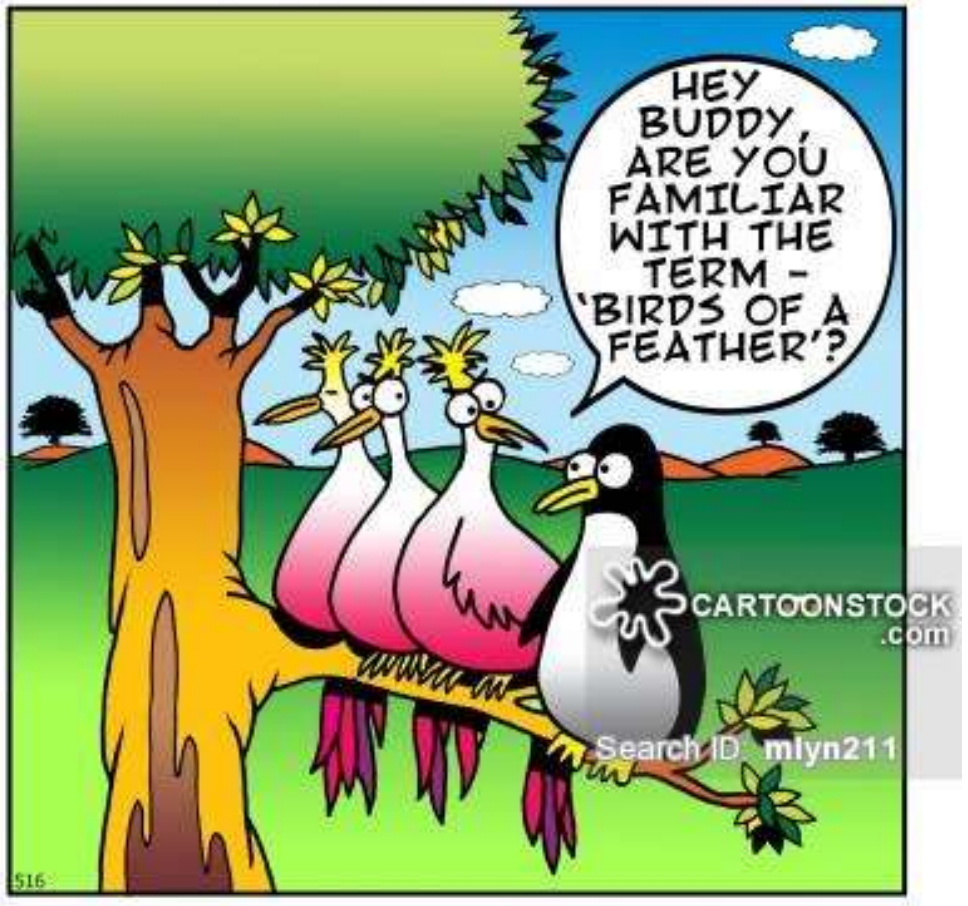
XUHU WAN

HKUST

JANUARY 15, 2022

Goals for this topic

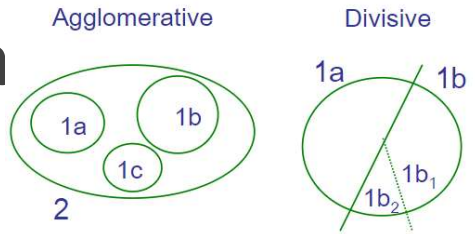
- Cluster data into different groups using Hierarchical Clustering.



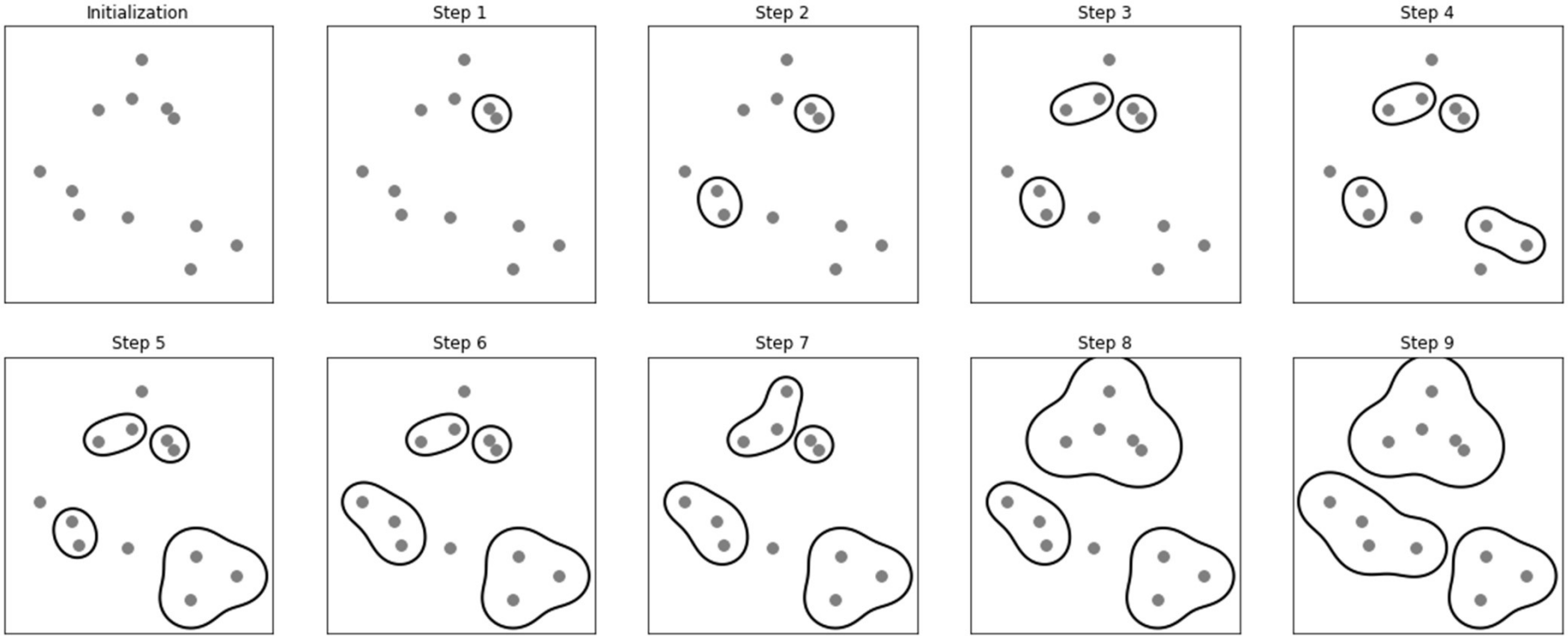
Births of a feather flock together

Hierarchical Clustering

Hierarchical methods can be either agglom
divisive.

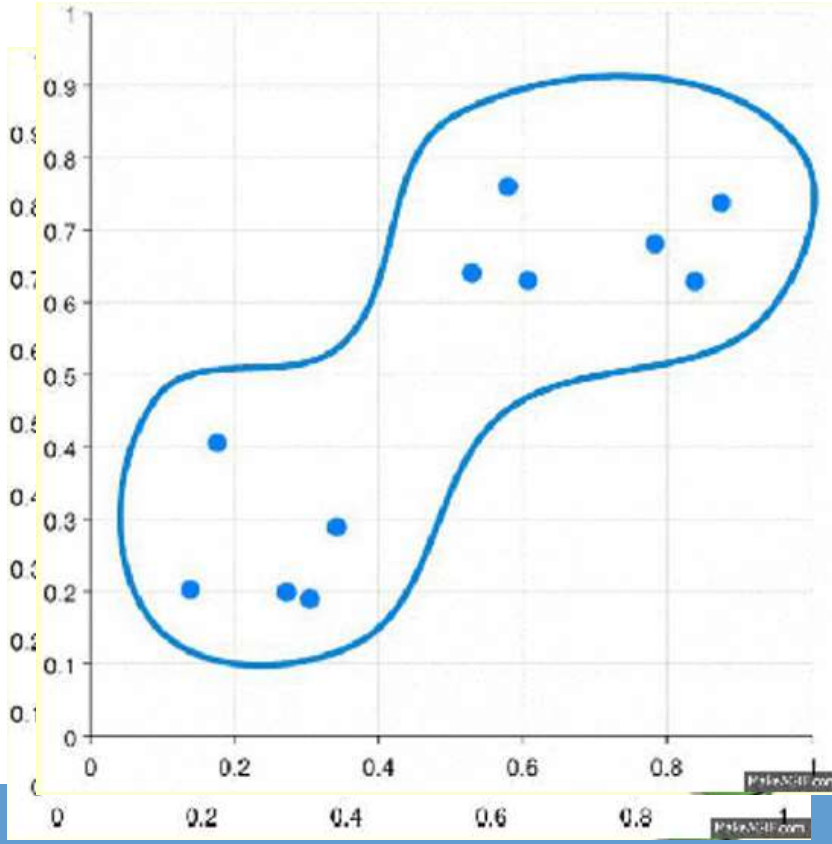


Most of the programs are of agglomerative type, and we only focus on it.



Hierarchical Clustering
(Continued)

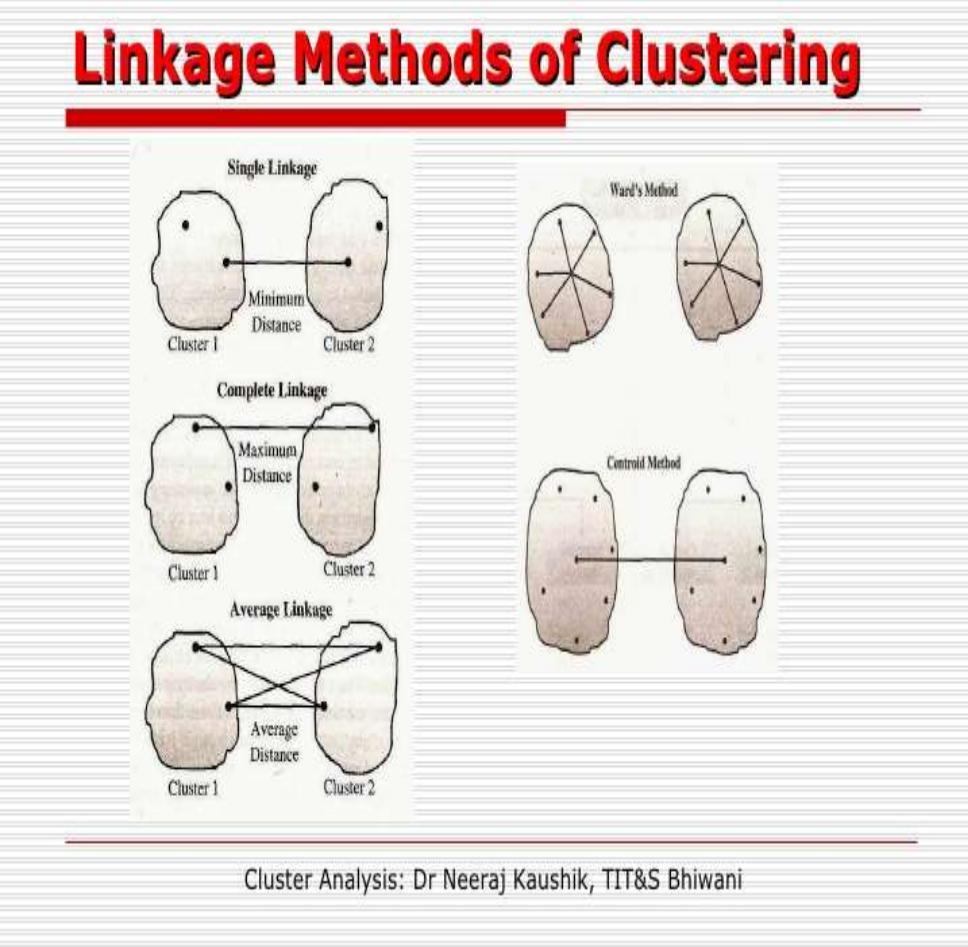
- How does agglomerative Hierarchical clustering work?
  1. Make each data point a single-point cluster → forms N clusters
  2. Take the two closest data points and make them one cluster → forms N-1 clusters
  3. Take the two closest clusters and make them one cluster → Forms N-2 clusters.
  4. Repeat step-3 until you are left with only one cluster



How to merge two clusters?

Linkage method

- **Single Linkage**: minimal inter-cluster distance
- **Complete Linkage**: maximal inter-cluster distance
- **Average Linkage**: mean inter-cluster distance
- **Centroid Linkage**: distance between the centroid for cluster A and cluster B



- **Ward Linkage**: it picks the two clusters such that the sum of squared deviation within all clusters increases the least. This often leads to clusters that are relatively equally sized.

- Mathematically, suppose there are K groups after each merge, then the sum of squared deviation within all clusters is defined as

$$\sum_{j=1}^{K} \sum_{i \in C_j} \left(X_{ij} - \bar{X}_{.j} \right)^2$$

Example Single Linkage

table	a	b	c	d	e
a	0	17	21	31	23
b	17	0	30	34	21
c	21	30	0	28	39
d	31	34	28	0	43
e	23	21	39	43	0

$D_1(a, b) = 17$

table	(a, b)	c	d	e
(a,b)	0			
c		0	28	39
d		28	0	43
e		39	43	0

$D_2((a, b), c) = (21 + 30)/2 = 25.5$
$D_2((a, b), d) = (31 + 34)/2 = 32.5$
$D_2((a, b), e) = (23 + 21)/2 = 22$

table	(a, b)	c	d	e
(a,b)	0	25.5	32.5	22
c	25.5	0	28	39
d	32.5	28	0	43
e	22	39	43	0

$D_2((a, b), e) = 22$

$D_3(((a, b), e), c) = (D_1(a, c) + D_1(b, c) + D_1(e, c))/3 = 30$
$D_3(((a, b), e), d) = (D_1(a, d) + D_1(b, d) + D_1(e, d))/3 = 29$

table	((a, b),e)	c	d
((a,b),e)	0		
c		0	28
d		28	0

$D_3(c, d) = 28$

table	((a, b),e)	c	d
((a,b),e)	0	30	29
c	30	0	28
d	29	28	0

table	((a, b),e)	(c,d)
((a,b),e)	0	
(c,d)		0

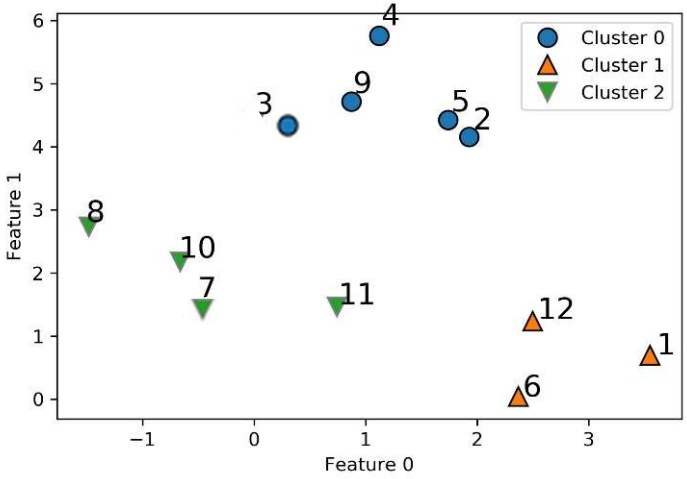
Hierarchical Clustering

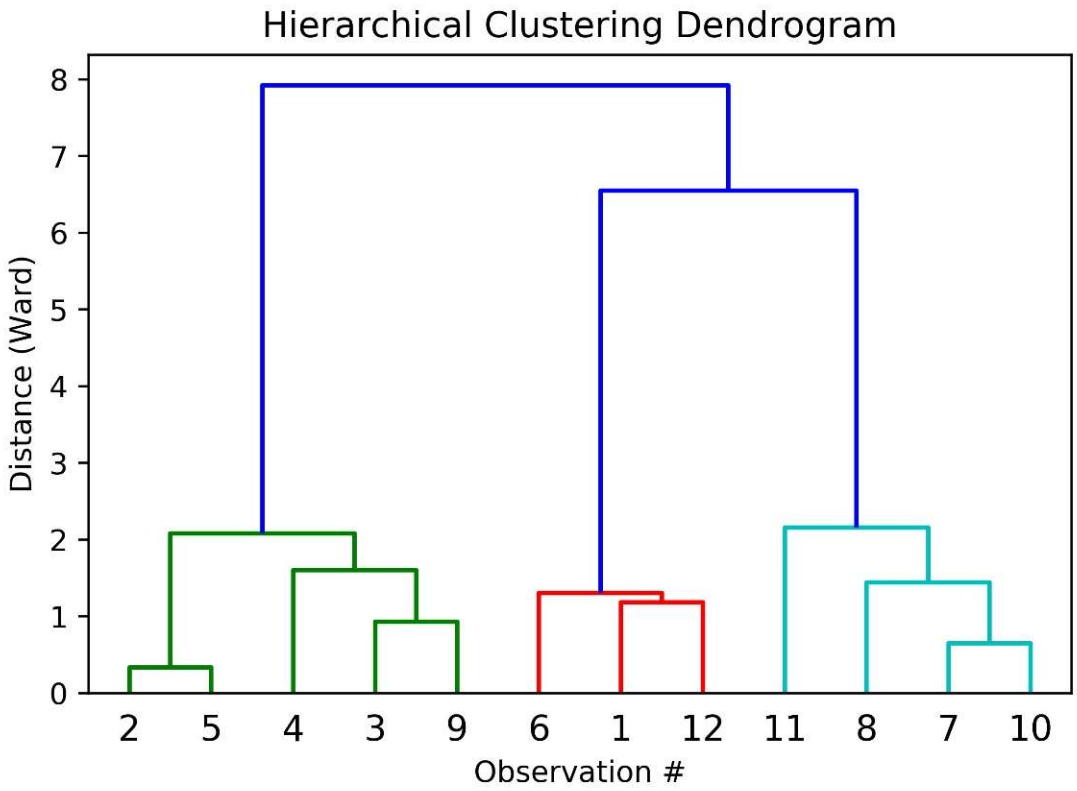
```
dendrogram(ward(X), labels=n)
```

- **Dendrogram** (/tree graph)
  - A tool to visualize hierarchical clustering.
  - The horizontal axis: lists the observations in a particular order.
  - The vertical axis: shows the successive steps or the distance between the centers of the clusters.

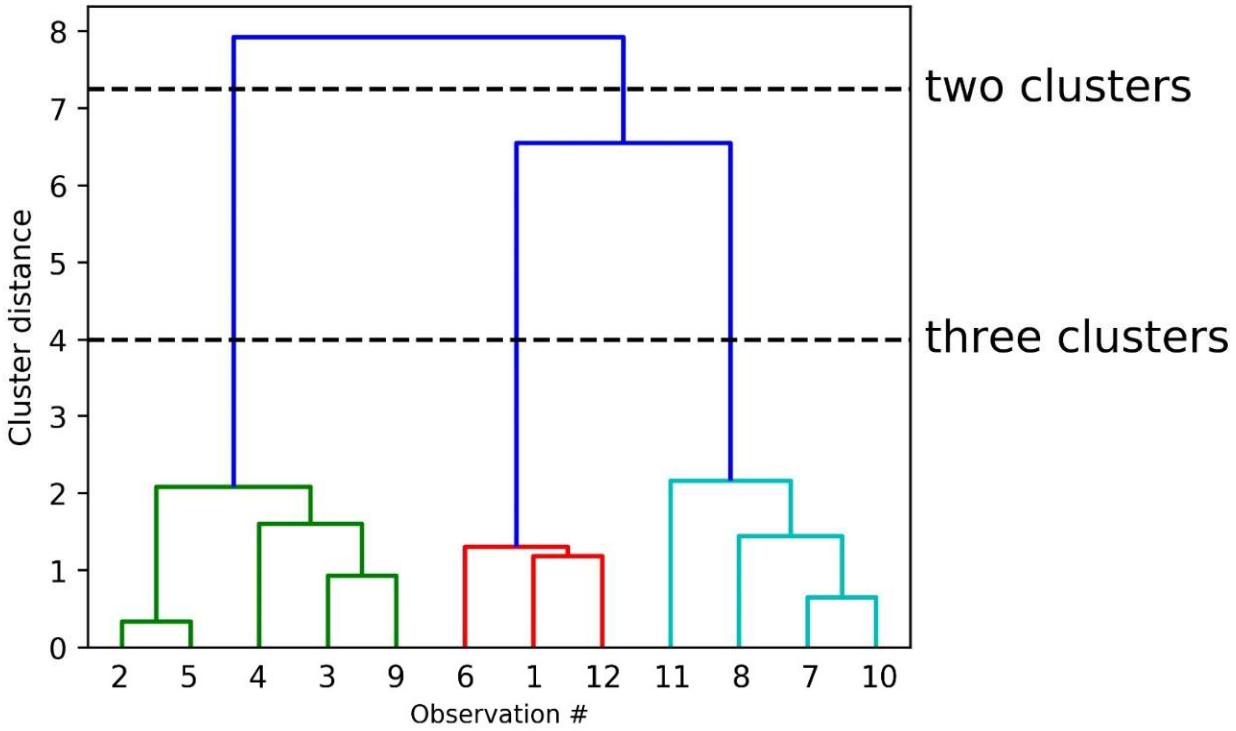
◦ A simulation study (Continued)

Dendrogram (Linkage=ward)





- How to choose the number of clusters in the dendrogram plot?



  - As a general rule, you cut the dendrogram with a horizontal line at the largest distance (jump) between two successive horizontal lines.
  - A large distance indicates that at least two very different observations exist, one from each of the two clusters just combined.
    - E.g. The largest increase in distance occurs when we combine the last three clusters.

How to determine the # of cluster in practice?

In practice: different numbers of clusters should be explored (say, starting with what a hierarchical clustering dendrogram may indicate), and the final choice should be made based on both statistical and qualitative criteria.

Example - Forbes Financial Data

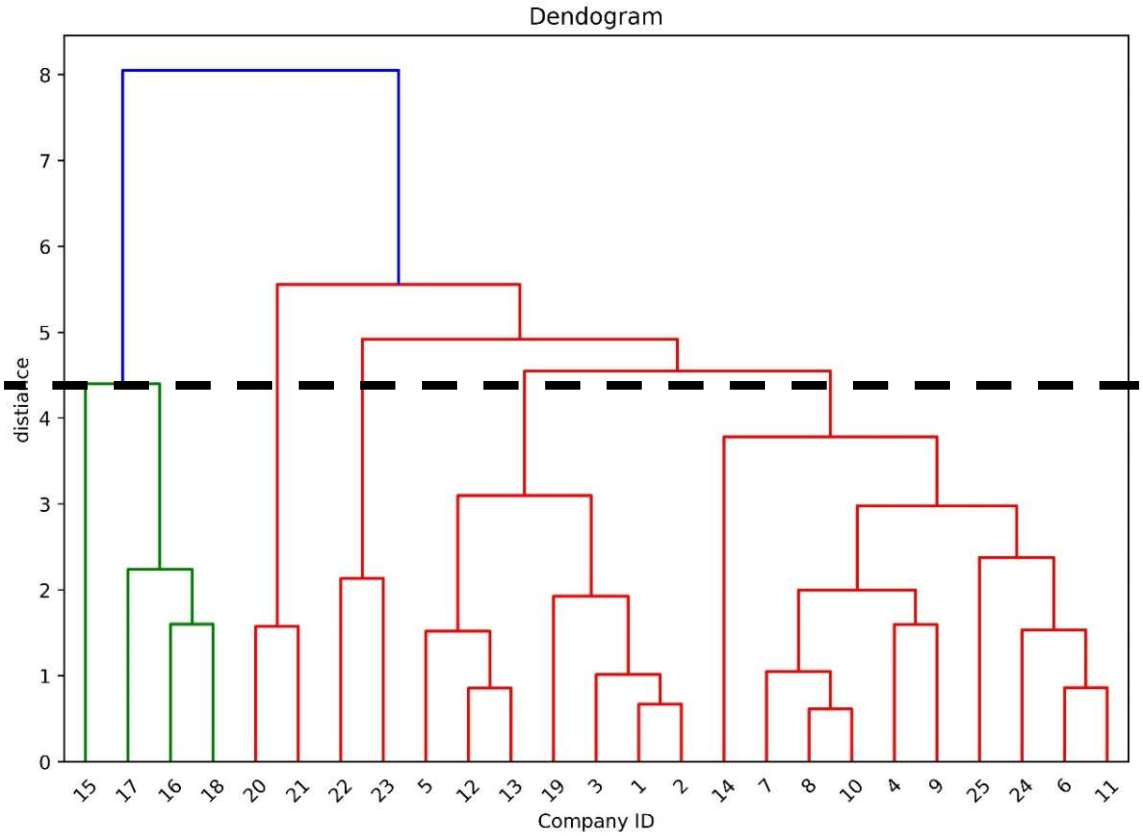
(Continued)

Now, we apply the Hierarchical clustering and K-means procedures to the financial performance data set.

The information on the type of company is not used to derive the clusters. However, the information will be used to interpret the results.

Example - Forbes Financial Data

Method 1: Hierarchical clustering




```
from sklearn.cluster import AgglomerativeClustering
plt.title("Dendogram")
Z = linkage(X_scaled, method='complete', metric='euclidean')
dendrogram(Z, labels = range(1, len(X_scaled) + 1))
plt.xlabel('Company ID')
plt.ylabel('distiance')
```

- We use complete linkage method with the Euclidian distance.
- The horizontal axis lists the observation numbers.
- The distances are measured being the maximal inter-cluster distance of the two clusters just joined.
- The distances are progressively increasing.

Example - Forbes Financial Data

```
# We choose 2 clusters
cluster = AgglomerativeClustering(n_clusters = 2, linkage = 'complete', affinity = 'euclidean').fit(X_d)
hier_labels = cluster.labels_
Forbes['hierarchical_label'] = hier_labels
print(hier_labels)
```

	TYPE	hierarchical_label
1	Chem	0
2	Chem	0
3	Chem	0
4	Chem	0
5	Chem	0
6	Chem	0
7	Chem	0
8	Chem	0
9	Chem	0
10	Chem	0
11	Chem	0
12	Chem	0
13	Chem	0
14	Chem	0
15	Heal	1
16	Heal	1
17	Heal	1
18	Heal	1
19	Heal	0
20	Groc	0
21	Groc	0
22	Groc	0
23	Groc	0
24	Groc	0
25	Groc	0

Analysis of the Hierarchical clustering result

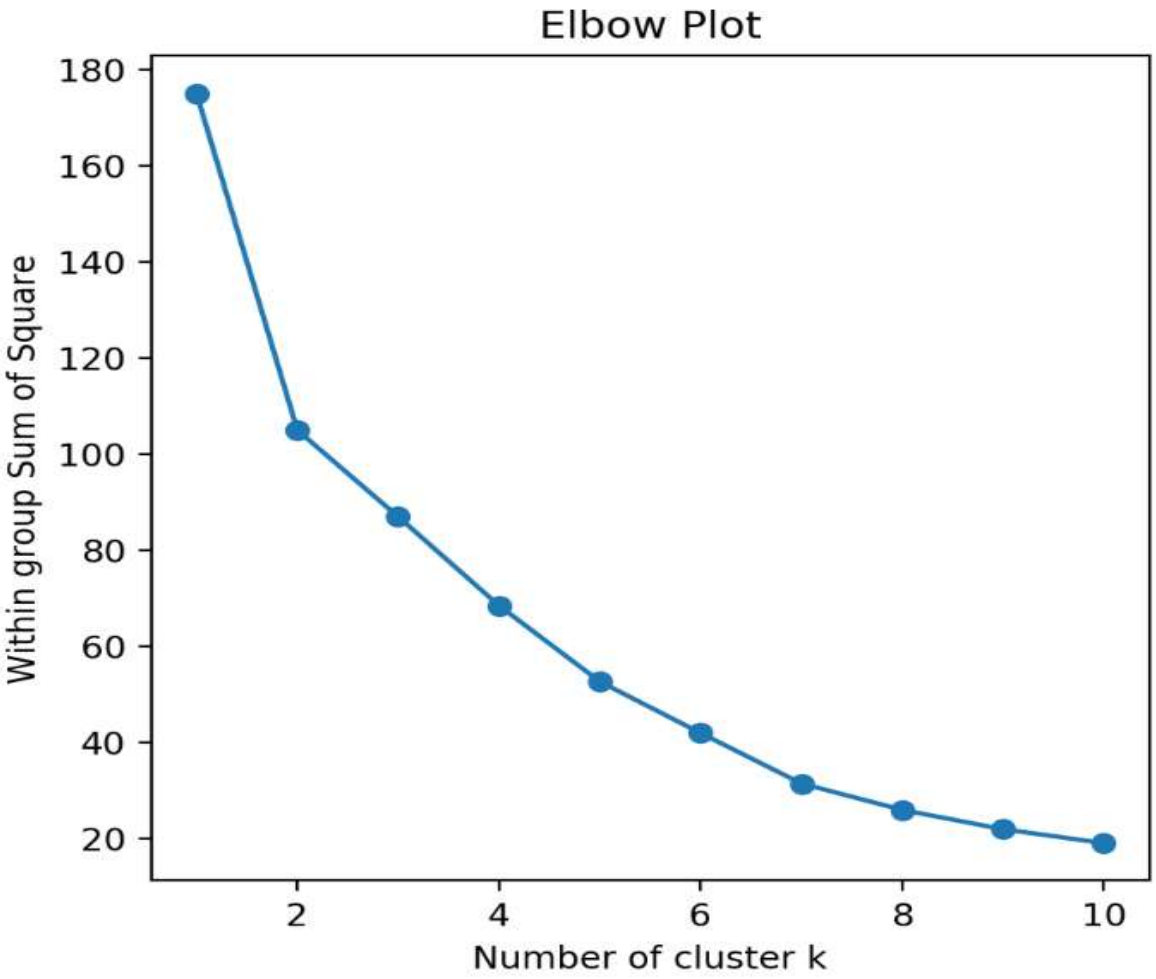
- Cluster 1: Pure hospital management firm.
- Cluster 0: Other firms

(Note: refer to Appendix 1 for other linkage method's result.)

Example - Forbes Financial Data

(Continued)

- Method 2: K-means




```
from sklearn.cluster import KMeans

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

```
plt.figure(figsize=(10,10))
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()
```

How to choose k? Use Elbow rule.

From the plot,
$k = 3$ is reasonable.

Interpret clusters by K-mean clustering

```
kmeans = KMeans(n_clusters=3, random_state=100)
kmeans.fit(X_scaled)
kmeans_labels = kmeans.predict(X_scaled)
Forbes['kmeans_label'] = kmeans_labels
```

```
Forbes.groupby('kmeans_label').mean()
```

kmeans_label	OBSNO	ROR5	DE	SALESGR5	EPS5	NPM1	PE	PAYOUTR1	hierarchical_label
0	16.5	8.875000	1.475000	32.100000	26.100000	5.850000	20.250000	0.289629	1.000000
1	10.0	13.957143	0.457143	15.871429	13.400000	5.228571	9.285714	0.463828	0.571429
2	13.5	9.150000	0.607143	12.892857	9.371429	3.400000	7.714286	0.413286	0.428571

```
Forbes.groupby('kmeans_label').std()
```

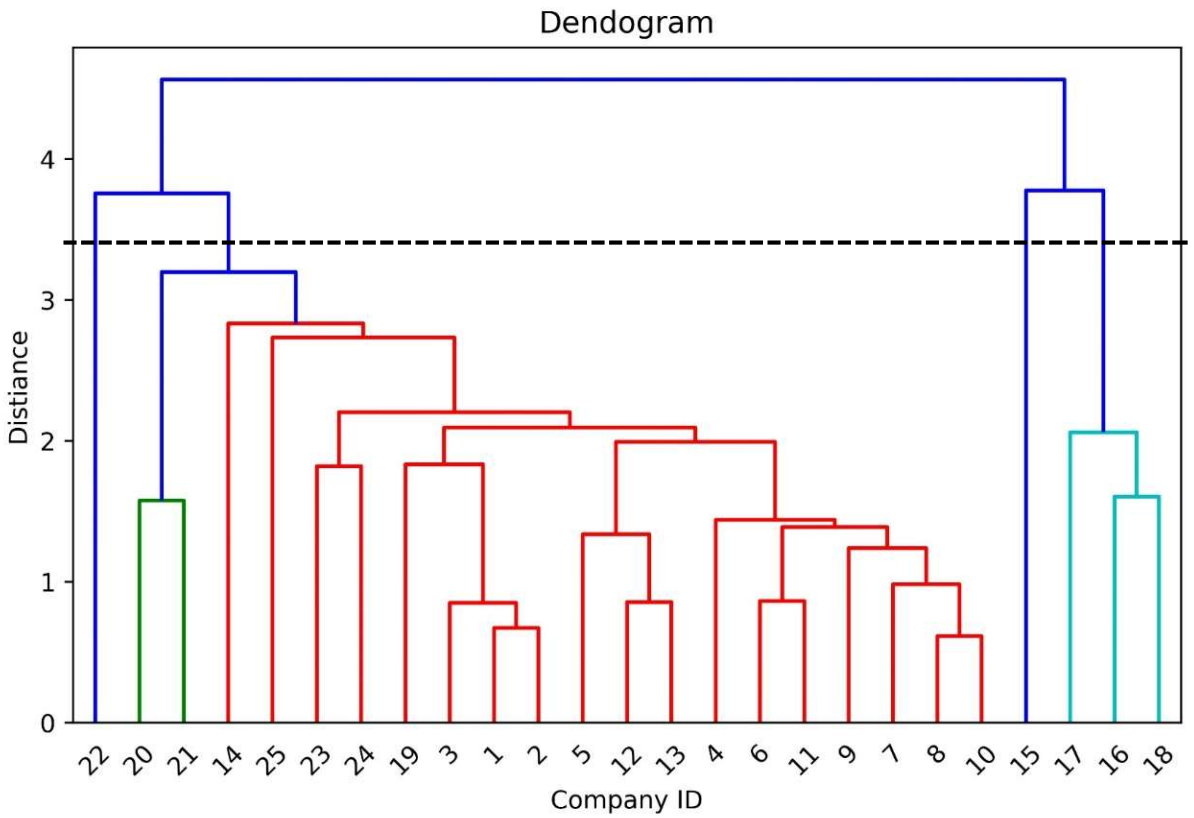
kmeans_label	OBSNO	ROR5	DE	SALESGR5	EPS5	NPM1	PE	PAYOUTR1	hierarchical_label
0	1.290994	0.340343	0.826136	8.589529	5.643285	0.741620	1.500000	0.095025	0.00000
1	9.416298	2.169376	0.237045	2.320714	2.136976	2.665655	2.138090	0.118959	0.97590
2	7.057457	1.233351	0.368916	2.735089	7.281740	1.957628	1.815683	0.118231	1.08941

Cluster 0: Growth company(/stock). (largest SALESGR5 and large PE).

Cluster 1: Income company(/stock). (large in ROR and Payout)

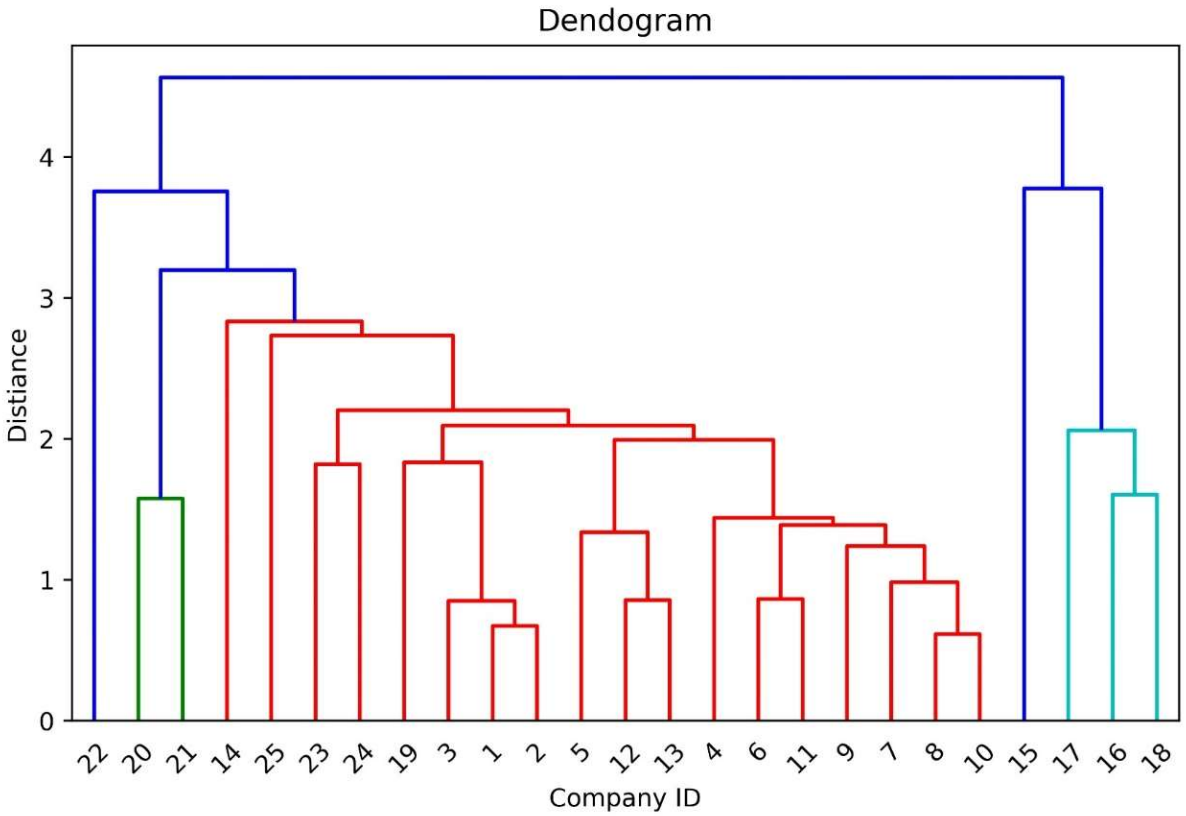
Cluster 2: Value company(/stock). (large Payout and lowest PE).

Result from Hierarchical clustering



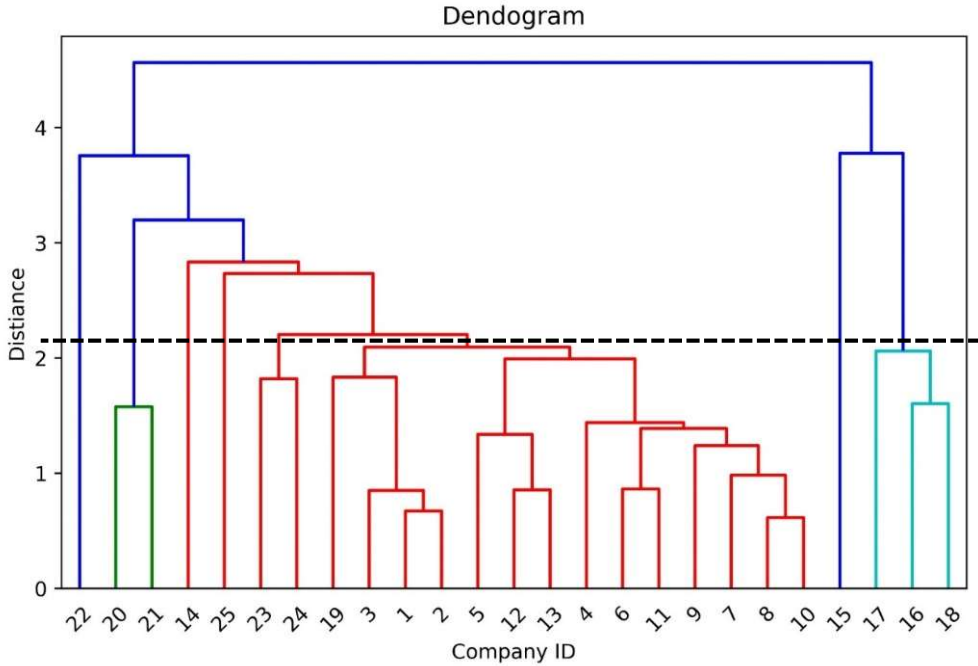
- We use centroid linkage method with the Euclidian distance.
- The horizontal axis lists the observation numbers.
- The distances are measured being the distance between the **centers** of the two clusters just joined.
- The distances are progressively increasing.

Analysis of the results



- Companies 8 and 10 are closest, so they form the first cluster.
- At the right end, 15, 16, 17, and 18 form a single cluster at the step in which there are two clusters.
- Company 22 stays by itself until there are only four clusters.

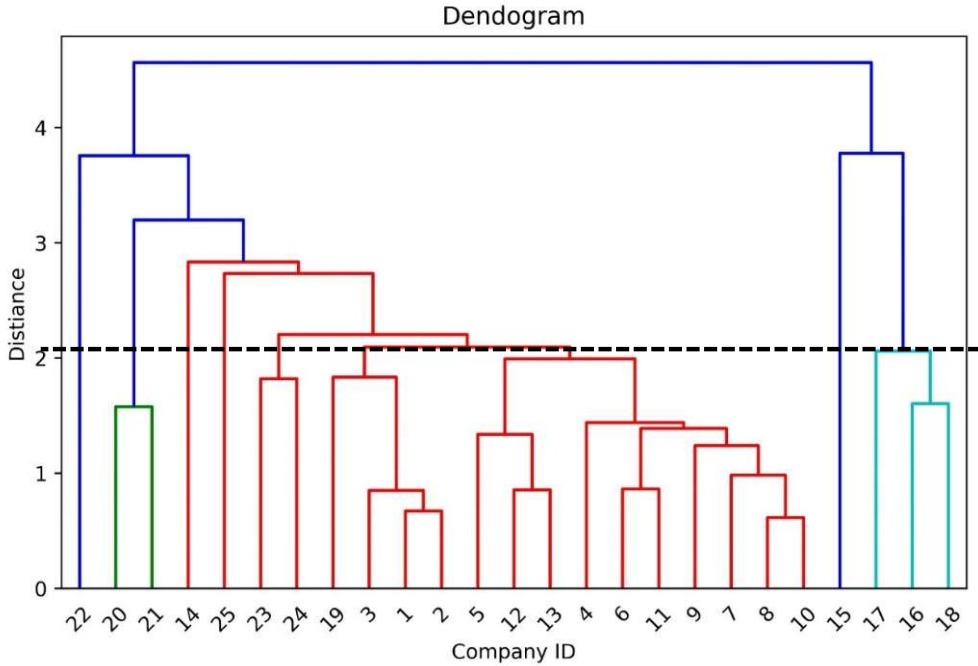
	TYPE	SYMBOL	OBSNO	ROR5	DE	SALESGR5	EPS5	NPM1	PE	PAYOUTR1
1	Chem	dia	1	13.0	0.7	20.2	15.5	7.2	9	0.426398
2	Chem	dow	2	13.0	0.7	17.2	12.7	7.3	8	0.380693
3	Chem	stf	3	13.0	0.4	14.5	15.1	7.9	8	0.406780
4	Chem	dd	4	12.2	0.2	12.9	11.1	5.4	9	0.568182
5	Chem	uk	5	10.0	0.4	13.6	8.0	6.7	5	0.324544
6	Chem	psm	6	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



Chemical companies

- 13 of the chemical companies, all except no. 14 (rci), are clustered together with only one non-chemical firm, company 19 (ahs), when the number of clusters is eight.
- The results are impressive when one considers that these are large diversified companies with varied emphasis, ranging from industrial chemicals to textiles to oil and gas production.

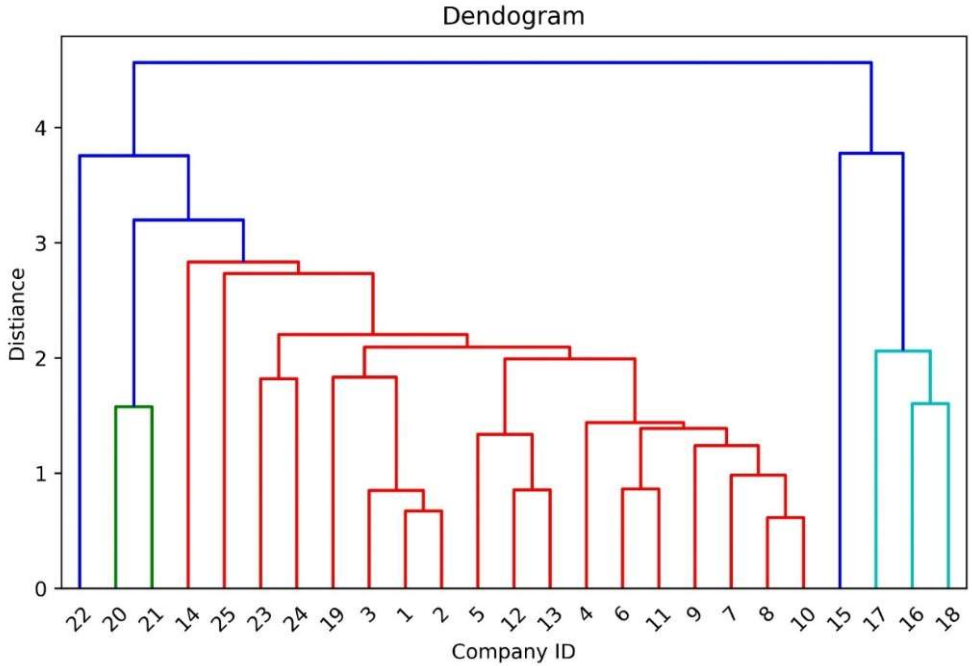
	TYPE	SYMBOL	OBSNO	ROR5	DE	SALESGR5	EPS5	NPM1	PE	PAYOUTR1
15	Heal	hum	15	9.2	2.7	39.8	34.4	5.8	21	0.390879
16	Heal	hca	16	8.9	0.9	27.8	23.5	6.7	22	0.161290
17	Heal	nme	17	8.4	1.2	38.7	24.6	4.9	19	0.303030
18	Heal	ami	18	9.0	1.1	22.1	21.9	6.0	19	0.303318
19	Heal	ahs	19	12.9	0.3	16.0	16.2	5.7	14	0.287500



Hospital management firms

- At the level of nine clusters, three of the four hospital management firms, companies 16, 17, and 18 (hca, nme, and ami) have been clustered together, and the other, company 15(hum), is added to that cluster before it is aggregated with any non-hospital management firms.
- From the data table, company 15 has clearly a different D/E value from others.
- The misfit in this health group is company 19, clustered with chemical firms.
- In fact, company 19 is a large, established supplies and equipment firm.

	TYPE	SYMBOL	OBSNO	ROR5	DE	SALESGR5	EPS5	NPM1	PE	PAYOUTR1
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



Grocery firms

- The grocery firms does not cluster tightly.

- From the table, they vary substantially on most variables. Company 22 is highly leveraged (high DE) and company 21 has low leverage (low DE) relative to others.
- Import disparities:
  - Three of them, companies 21, 24, and 25 (win, ka, sa), are three of the four largest United States grocery supermarket chains.
  - Two others, companies 20 and 22 (kls, sgl), have a diversified mix of grocery, drug, department, and other stores.
  - The remaining firm, company 23 (slc) concentrates on convenient stores (7-Eleven) and has a substantial franchising operation.
  - Thus, the six are quite different from each other.

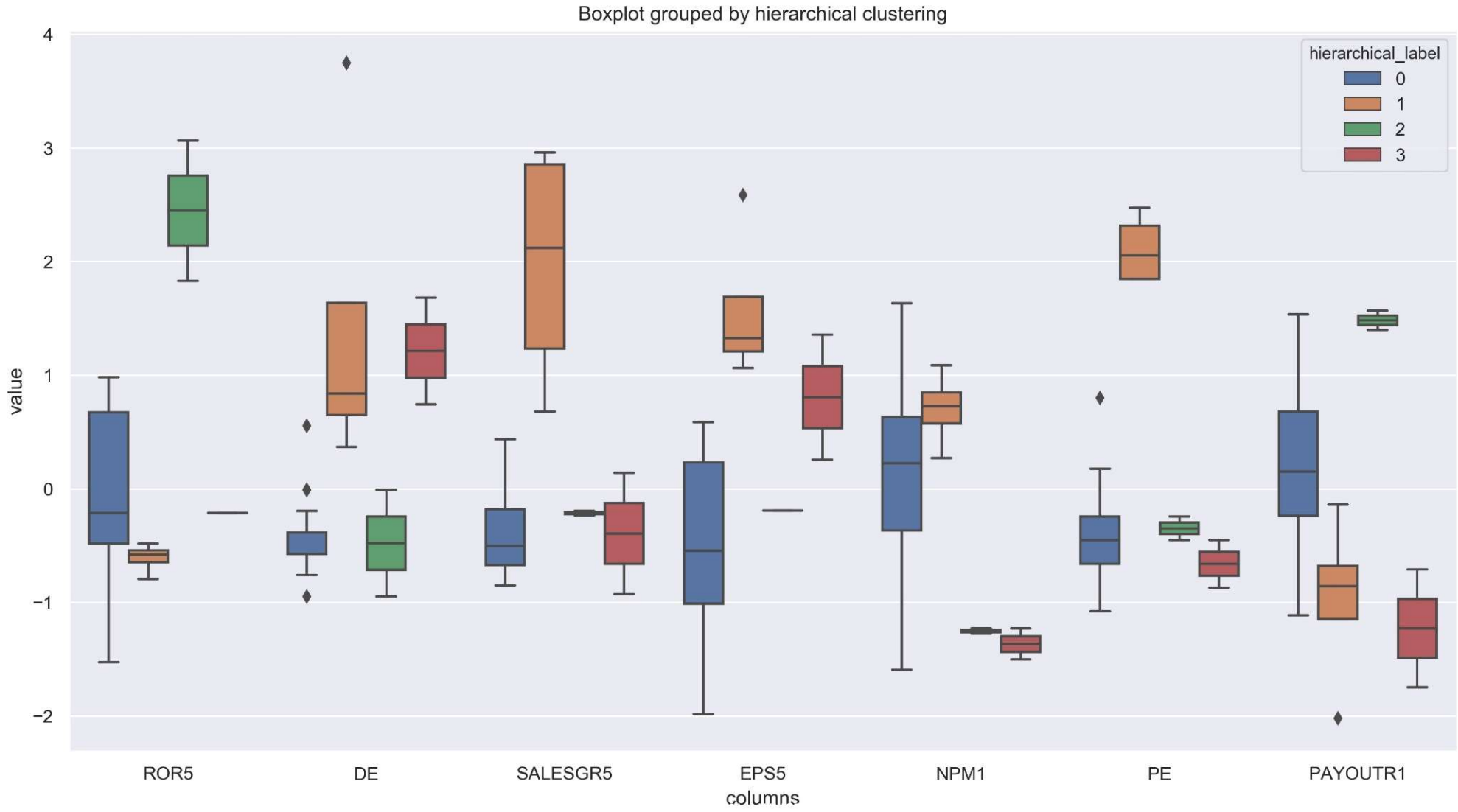
Analysis of the result

- Cluster 0: Chemical firm, and largest United States grocery supermarket chains and Healthcare company that established supplies and equipment firm.
- Cluster 1: Pure hospital management firm.
- Cluster 2: High return grocery firm
- Cluster 3: High debt grocery firm

	TYPE	hierarchical_label
1	Chem	0
2	Chem	0
3	Chem	0
4	Chem	0
5	Chem	0
6	Chem	0
7	Chem	0
8	Chem	0
9	Chem	0
10	Chem	0
11	Chem	0
12	Chem	0
13	Chem	0
14	Chem	0
15	Heal	1
16	Heal	1
17	Heal	1
18	Heal	1
19	Heal	0
20	Groc	2
21	Groc	2
22	Groc	3
23	Groc	3
24	Groc	0
25	Groc	0

ring

Interpret clusters by hierarchical clustering



Take away from Topic 4

Clustering Method

- Hierarchical clustering
  - Dendrogram
  - Define the distance between two clusters: Linkage method

Determine the # of cluster in practice

- Elbow plot [K-means]
- Dendrogram [Hierarchical clustering]