

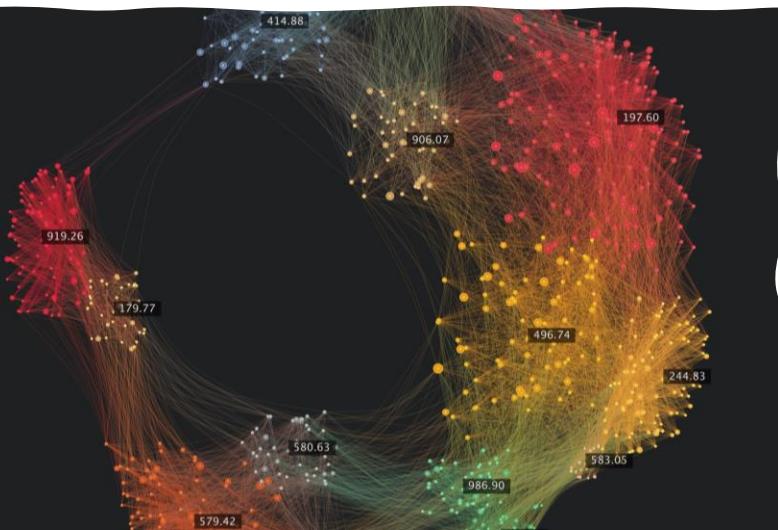
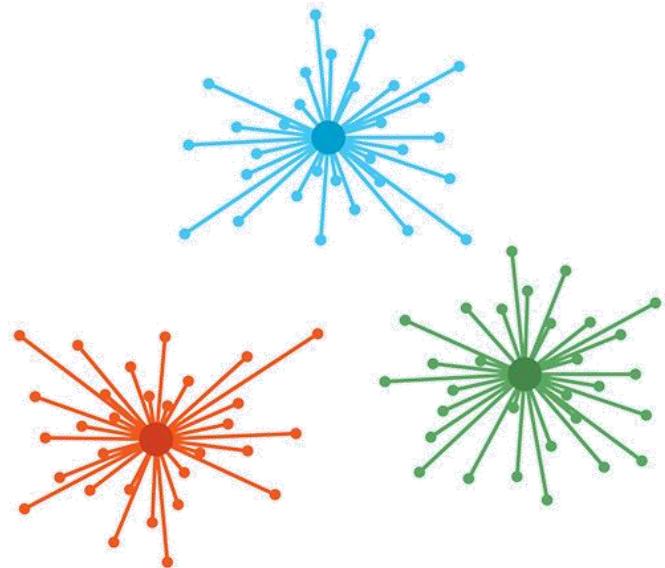
IDENTIFYING AND UNDERSTANDING GROUPS

Chapter 5

Clustering.

Analysing groups

Introduction



- For a given data set, ***not known beforehand***, what **groups of observations** the **entire data set** is composed of.
- Example: Look at **customer data** collected from a **particular store**
 - Is it possible to **identify and summarize classes** of **customers** **directly from the data**
 - Can you answer questions such as “*What types of customers visit the store?*”
 - **Clustering** groups data into sets of related observations or clusters, so that observations **within each group** are **more similar** to other observations **within the group** than to observations **within other groups**.

Overview.

- **Clustering** is a widely used and flexible approach to analyzing data, in which observations are automatically organized into groups.
- Clustering is **an unsupervised method** for grouping:
 - Groups are **not known in advance** and **a goal (a specific variable) is not used** to direct **how the grouping** is generated - **all variables are considered** in the analysis.
- **Clustering method** chosen to **subdivide the data into groups** applies an automated procedure to discover the groups based on some criteria and its solution is **extracted from patterns or structure existing in the data**.
- Many clustering methods exist.
- For clustering, there is **no way to measure accuracy** and **the solution** is judged by its “usefulness.”
 - Clustering is **open ended way to explore, understand, and formulate questions** about the **data** in exploratory data analysis.

TABLE 3.1 Portion of a Table Describing Various Animals

Name	Hair	Feathers	Eggs	Milk	Airborne	Aquatic	Predator	Toothed	Backbone	Breathes	Venomous	Fins	Legs	Tail
Aardvark	1	0	0	1	0	0	1	1	1	1	0	0	4	0
Bass	0	0	1	0	0	1	1	1	1	0	0	1	0	1
Boar	1	0	0	1	0	0	1	1	1	1	0	0	4	1
Buffalo	1	0	0	1	0	0	0	1	1	1	0	0	4	1
Carp	0	0	1	0	0	1	0	1	1	0	0	1	0	1
Catfish	0	0	1	0	0	1	1	1	1	0	0	1	0	1
Cheetah	1	0	0	1	0	0	1	1	1	1	0	0	4	1
Chicken	0	1	1	0	1	0	0	0	1	1	0	0	2	1
Clam	0	0	1	0	0	0	1	0	0	0	0	0	0	0

- **Cluster analysis** organizes the **data into groups** of.
- Using the **numeric variables** shown in Table 3.1, the data set can be clustered in a number of ways.



Mammals

platypus

- Animals organized into **4** groups representing the following general categories:
 - Mammals:** Four-legged animals that produce milk, the fish,
 - Fish and amphibians:** Aquatic animals
 - Invertebrates:** Six-legged animals with no backbone, and
 - Birds:** set of two-legged airborne animals with feathers.

Fish and amphibians

seasnake

pitviper
toad
tuatara
frog
newt

seahorse bass
haddock stingray
herring pike tuna
catfish piranha
carp sole

Figure 3.1 shows one possible grouping of this data

Invertebrates

scorpion

gnat
moth termite

honeybee

housefly flea

starfish crab
crayfish clam
seawasp octopus

dove flamingo
gull sparrow
crow wren kiwi
hawk duck vulture
lark chicken
penguin ostrich

Birds

Figure 3.1 Four groups of animals in the data set

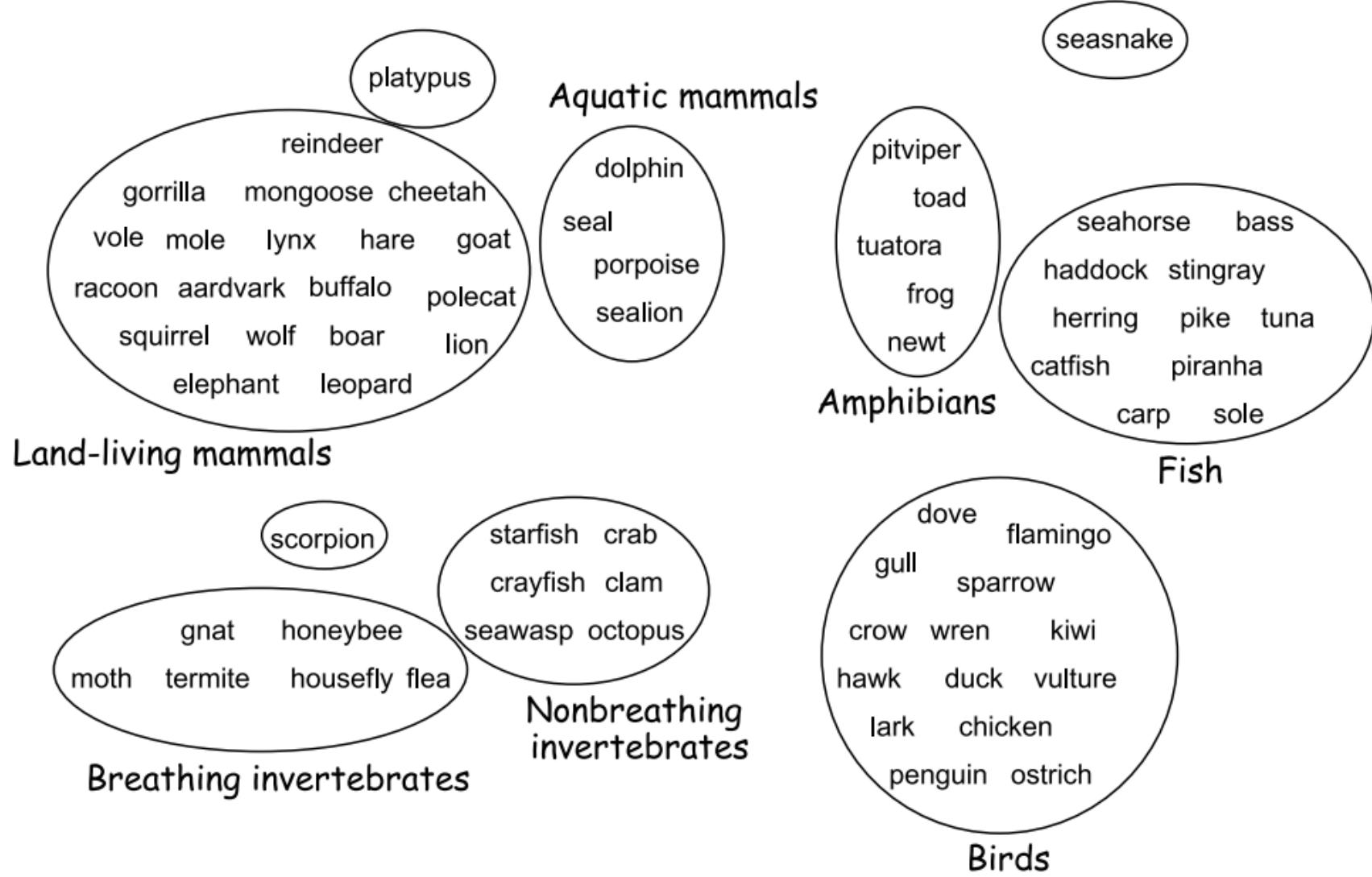


Figure 3.2 Ten groups generated from the animal data set

In Fig. 3.2,
Subsets of larger mammals:

- Land-living mammals group
- Aquatic mammals group
- Platypus

Platypus: Group of its own, since it is both aquatic and lays eggs.

Fish and amphibians and **invertebrates** groups have been divided into smaller groups.

Original group of birds remains the same since it already represented a set of similar animals.

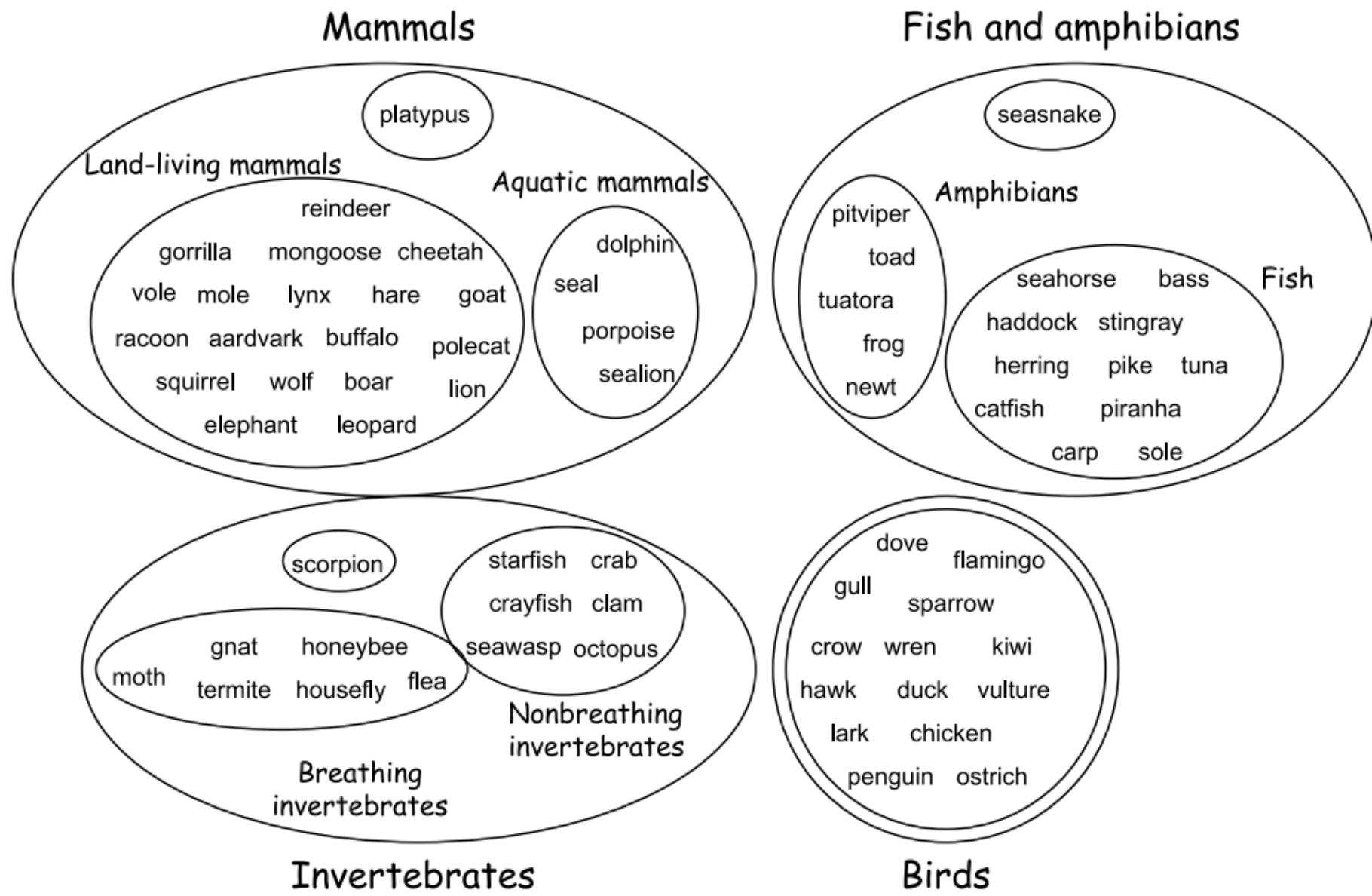


Figure 3.3 The two groupings superimposed

Zoo Data Set

Download: [Data Folder](#), [Data Set Description](#)

Abstract: Artificial, 7 classes of animals

Data Set Characteristics:	Multivariate	Number
Attribute Characteristics:	Categorical, Integer	Number
Associated Tasks:	Classification	Missing

Data Set Information:

A simple database containing 17 Boolean-valued attributes. The "type" attribute appears 2 instances of "frog" and one of "girl"!)

Class# -- Set of animals:

=====

1 -- (41) aardvark, antelope, bear, boar, buffalo, calf, cavy, cheetah, deer, dolphin, elephant, platypus, polecat, pony, porpoise, puma, pussycat, raccoon, reindeer, seal, sealion
2 -- (20) chicken, crow, dove, duck, flamingo, gull, hawk, kiwi, lark, ostrich, parakeet, penguin
3 -- (5) pitviper, seasnake, slowworm, tortoise, tuatara
4 -- (13) bass, carp, catfish, chub, dogfish, haddock, herring, pike, piranha, seahorse, shark
5 -- (4) frog, newt, toad
6 -- (8) flea, gnat, honeybee, housefly, ladybird, moth, termite, wasp
7 -- (10) clam, crab, crayfish, lobster, octopus, scorpion, seawasp, slug, starfish, worm

Attribute Information:

1. animal name: Unique for each instance
2. hair: Boolean
3. feathers: Boolean
4. eggs: Boolean
5. milk: Boolean
6. airborne: Boolean
7. aquatic: Boolean
8. predator: Boolean
9. toothed: Boolean
10. backbone: Boolean
11. breathes: Boolean
12. venomous: Boolean
13. fins: Boolean
14. legs: Numeric (set of values: {0,2,4,5,6,8})
15. tail: Boolean
16. domestic: Boolean
17. catsize: Boolean
18. type: Numeric (integer values in range [1,7])

Attribute Name	Role	Type	Description	Units	Missing Values
animal_name	Other	Categorical	N/A	N/A	False
hair	Feature	Categorical	N/A	N/A	False
feathers	Feature	Categorical	N/A	N/A	False
eggs	Feature	Categorical	N/A	N/A	False
breathes	Feature	Categorical	N/A	N/A	False
venomous	Feature	Categorical	N/A	N/A	False
domestic	Feature	Categorical	N/A	N/A	False
catsize	Feature	Categorical	N/A	N/A	False
type	Target	Categorical	N/A	N/A	False

aardvark,1,0,0,1,0,0,1,1,1,1,0,0,4,0,0,1,1
antelope,1,0,0,1,0,0,0,1,1,1,0,0,4,1,0,1,1
bass,0,0,1,0,0,1,1,1,0,0,1,0,1,0,0,4
bear,1,0,0,1,0,0,1,1,1,0,0,4,0,0,1,1
boar,1,0,0,1,0,0,1,1,1,0,0,4,1,0,1,1
buffalo,1,0,0,1,0,0,0,1,1,1,0,0,4,1,0,1,1
calf,1,0,0,1,0,0,0,1,1,1,0,0,4,1,1,1,1
carp,0,0,1,0,0,1,0,1,1,0,0,1,0,1,1,0,4
catfish,0,0,1,0,0,1,1,1,0,0,1,0,1,0,0,4
cavy,1,0,0,1,0,0,0,1,1,1,0,0,4,0,1,0,1
cheetah,1,0,0,1,0,0,1,1,1,1,0,0,4,1,0,1,1
chicken,0,1,1,0,1,0,0,0,1,1,0,0,2,1,1,0,2
chub,0,0,1,0,0,1,1,1,0,0,1,0,1,0,0,4
clam,0,0,1,0,0,0,1,0,0,0,0,0,0,0,0,0,7
crab,0,0,1,0,0,1,1,0,0,0,0,0,4,0,0,0,7
crayfish,0,0,1,0,0,1,1,0,0,0,0,0,6,0,0,0,7
crow,0,1,1,0,1,0,1,0,1,1,0,0,2,1,0,0,2
deer,1,0,0,1,0,0,0,1,1,1,0,0,4,1,0,1,1
dogfish,0,0,1,0,0,1,1,1,1,0,0,1,0,1,0,1,4
dolphin,0,0,0,1,0,1,1,1,1,0,1,0,1,0,1,1
dove,0,1,1,0,1,0,0,0,1,1,0,0,2,1,1,0,2
duck,0,1,1,0,1,1,0,0,1,1,0,0,2,1,0,0,2
elephant,1,0,0,1,0,0,0,1,1,1,0,0,4,1,0,1,1
flamingo,0,1,1,0,1,0,0,0,1,1,0,0,2,1,0,1,2
flea,0,0,1,0,0,0,0,0,1,0,0,6,0,0,0,6
frog,0,0,1,0,0,1,1,1,1,0,0,4,0,0,0,5
frog,0,0,1,0,0,1,1,1,1,1,0,4,0,0,0,5
fruitbat,1,0,0,1,1,0,0,1,1,1,0,0,2,1,0,0,1
giraffe,1,0,0,1,0,0,0,1,1,1,0,0,4,1,0,1,1

Class	Count	Mean (petal width (cm))
Iris-setosa	50	0.244
Iris-versicolor	50	1.33
Iris-virginica	50	2.03

FIGURE 5.1 Simple summary table showing how the mean petal width changes for the different classes of flowers.

Question: How a single variable such as petal width varies among different species as illustrated in Figure 5.1?

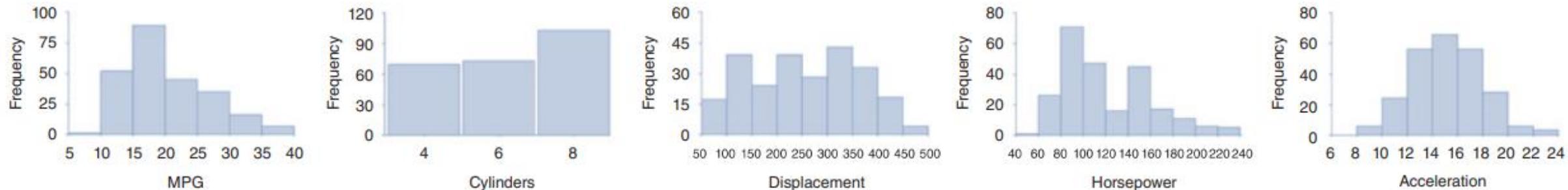
- Easily extended to help understand the relationships between groups and multiple variables, as illustrated in Figure 5.2 where **three predefined categories** are used to **group the observations**.
- **Summary information on multiple variables** is presented (using the mean value in the example).
- These tables can use summary statistics (e.g., mean, mode, median, and so on) in addition to graphs such as box plots that illustrate the subpopulations.

Class	Count	Mean (sepal length (cm))	Mean (sepal width (cm))	Mean (petal length (cm))	Mean (petal width (cm))
Iris-setosa	50	5.01	3.42	1.46	0.244
Iris-versicolor	50	5.94	2.77	4.26	1.33
Iris-virginica	50	6.59	2.97	5.55	2.03

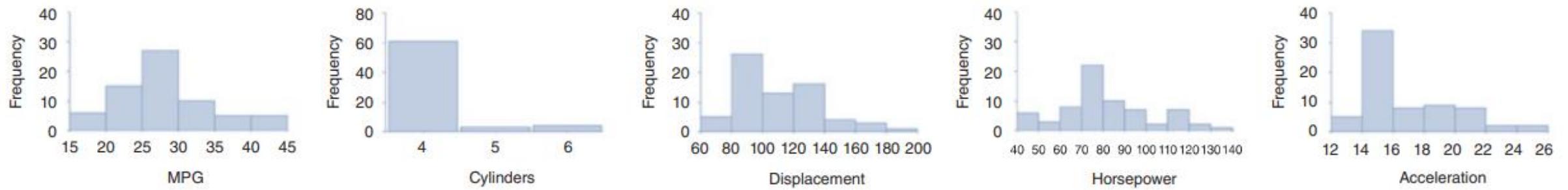
FIGURE 5.2 The use of a summary table to understand multiple variables for a series of groups.

A variety of graphs (e.g., histograms, box plots, and so on) for each group can also be shown in a table or grid format known as small multiples that allows comparison.

American cars



Asian cars



European cars

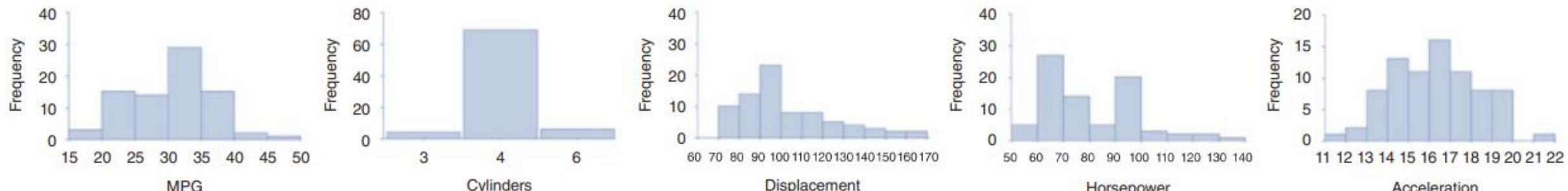


FIGURE 5.3 Matrix showing the frequency distribution for a common set of variables for three groups of cars—American, European, and Asian.

A variety of graphs (e.g., histograms, box plots, and so on) for each group can also be shown in a table or grid format known as small multiples that allows comparison.

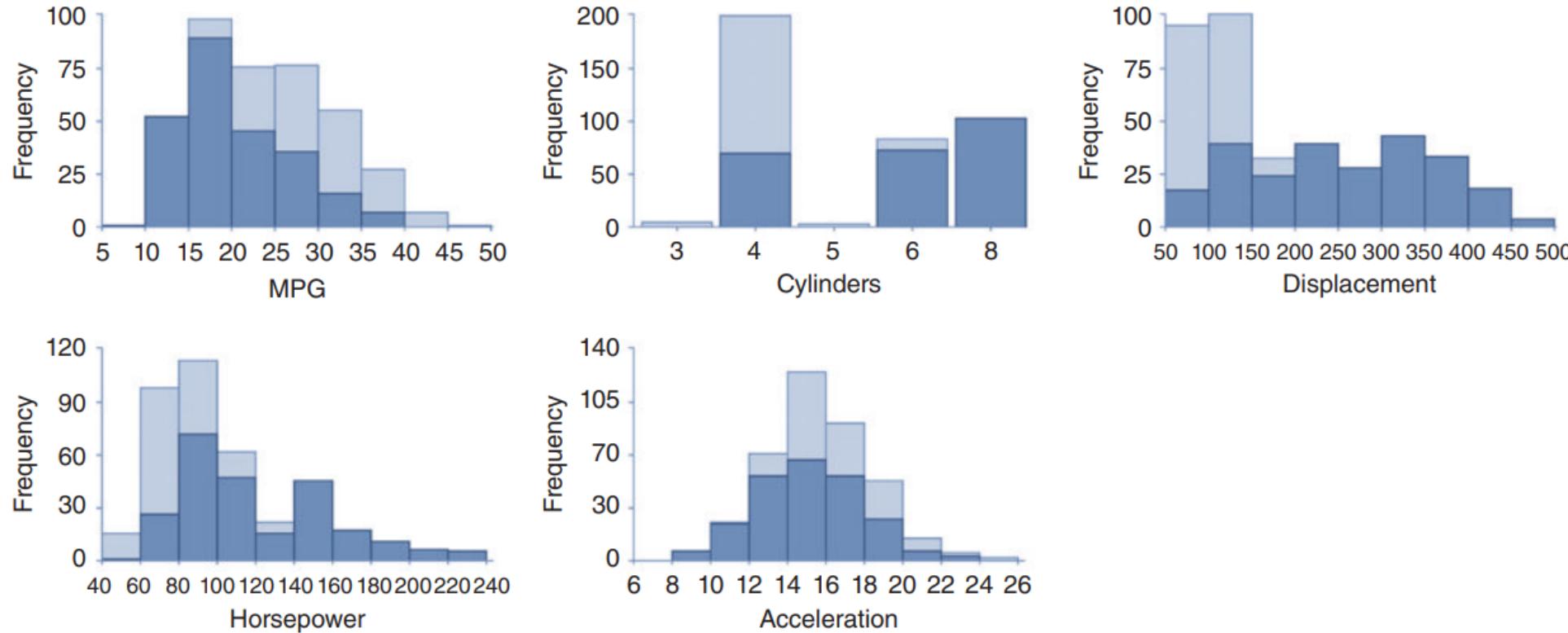


FIGURE 5.4 Highlighting a group of observations using shading.

Clustering.

To illustrate the process of clustering, **a set of observations** are shown on the scatterplot.

- Observations plotted using two hypothetical dimensions
- **Similarity** between observations is **proportional to the physical distance** between the observations.
- **Two clear regions** that can be considered as clusters:
 - Cluster A and Cluster B, since many of the observations are contained within these two regions on the scatterplot
- Clustering not only assists in **identifying groups of related observations**, can also **locate outliers** (observations that are not similar to others) since they fall into groups of their own

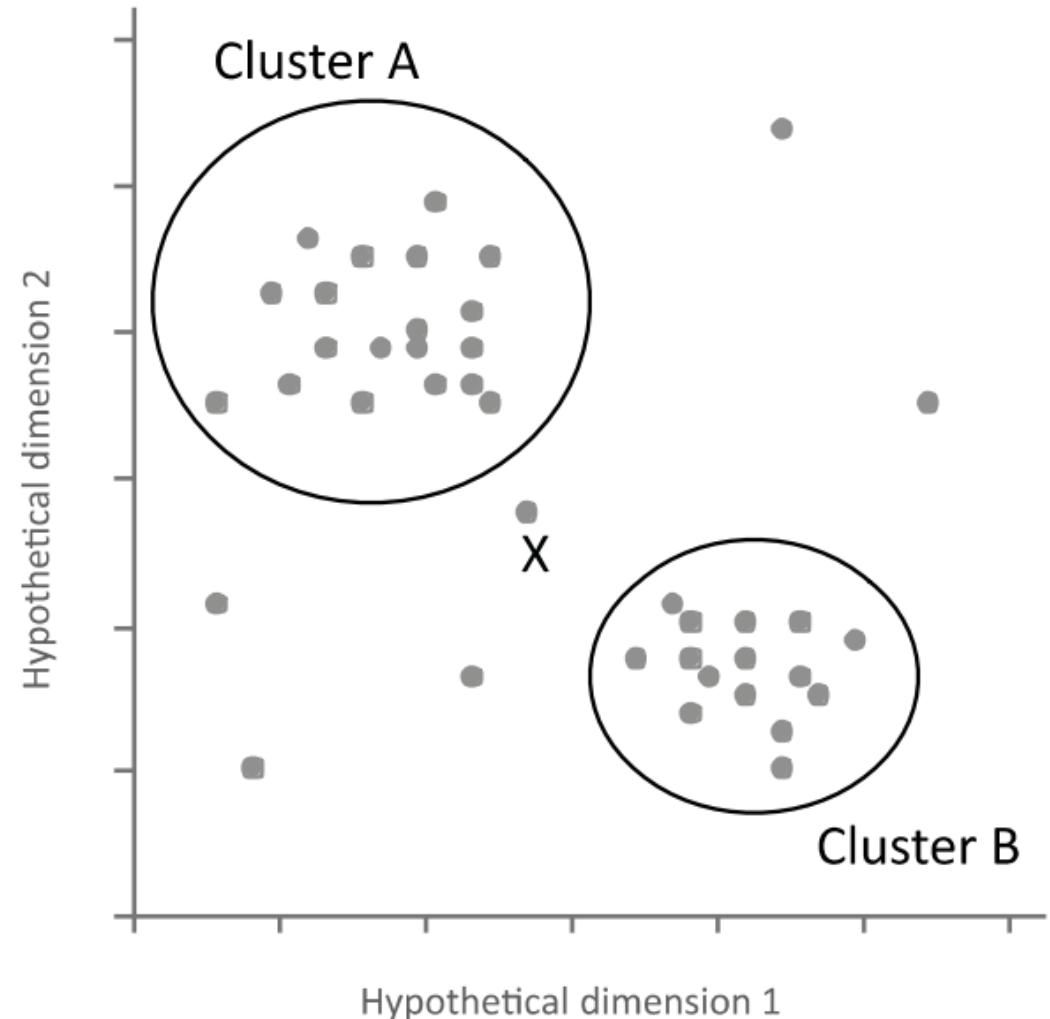


FIGURE 5.5 Illustration of clusters and outliers 15

Clustering.

- **Cluster analysis** requires **neither a prior understanding** of the data nor **assumptions** to be made about it.
- For example, it is **not necessary** that **variables** follow a **normal distribution**, nor is it necessary to **identify independent** and **response variables**
- **Care** should be taken to **select the appropriate variables**, including only those **relevant** to the **grouping exercise**.
- **Results** from a cluster analysis are **groups of observations** that **share common characteristics**.

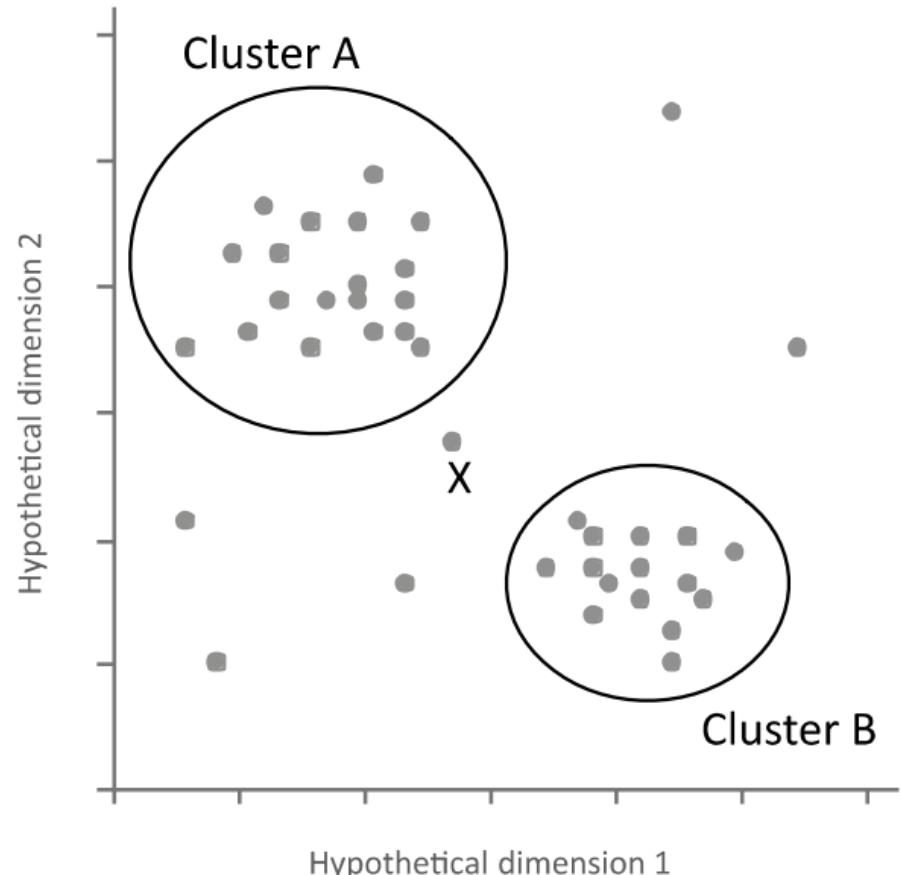


FIGURE 5.5 Illustration of clusters and outliers.



The way to get
started is to quit
talking and begin
doing.

Walt Disney

Clustering.

- Is a **flexible approach for grouping**.
- For example, based on the **criteria for clustering** the observations, observation **X** was **not** determined to be a **member of cluster A**.
- However, if a more **relaxed criterion** was used, **X may have been included in cluster A**.
- In Figure 5.5, there are **six observations that do not fall within cluster A or B**, observed to be **outliers**.

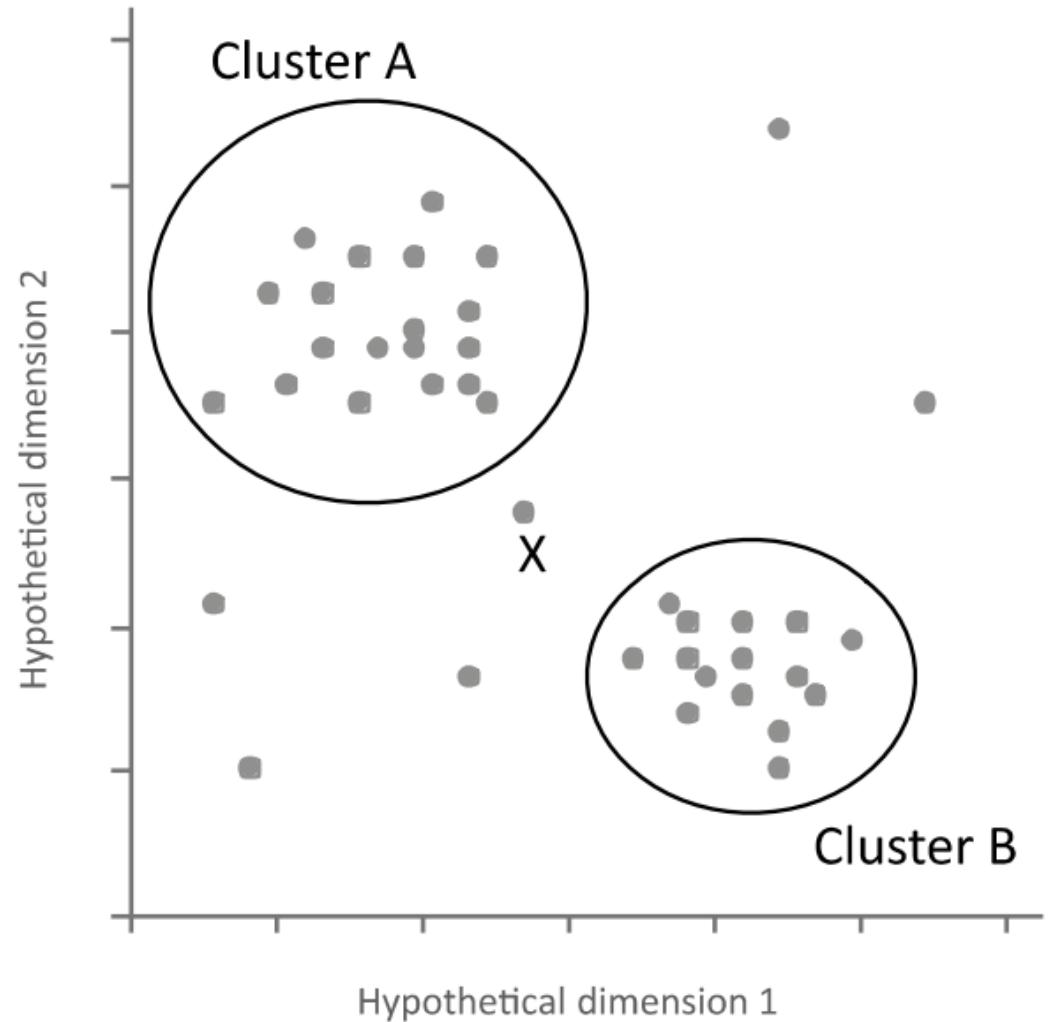


FIGURE 5.5 Illustration of clusters and outliers. 18

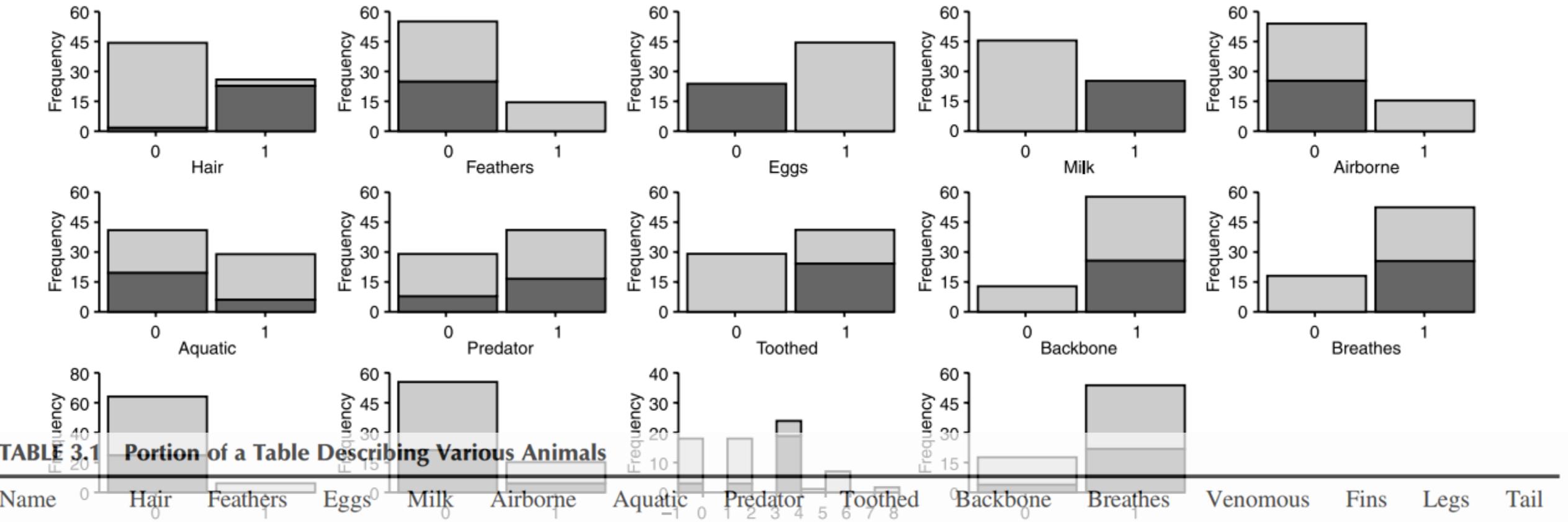


TABLE 3.1 Portion of a Table Describing Various Animals

Name	Hair	Feathers	Eggs	Milk	Airborne	Aquatic	Predator	Tooothed	Backbone	Breathes	Venomous	Fins	Legs	Tail
Aardvark	1	0	0	1	0	0	0	1	1	0	0	0	4	0
Bass	0	0	1	0	0	1	0	0	1	1	0	1	0	1
Boar	1	0	0	1	0	0	1	0	1	1	0	0	4	1
Buffalo	1	0	0	1	0	0	0	0	1	1	0	0	4	1
Carp	0	0	1	0	0	0	1	0	1	1	0	0	1	0
Catfish	0	0	1	0	0	0	1	1	1	0	0	0	1	0
Cheetah	1	0	0	1	0	0	0	1	1	1	0	0	4	1
Chicken	0	1	1	0	1	0	0	0	1	1	0	0	2	1
Clam	0	0	1	0	0	0	0	1	0	0	0	0	0	0

Figure 3.4 Histogram matrix for the animals in the mammals groups compared to the total data set

Clustering methods.

- Different types of **clustering methods** exist. Selection of a **method** influenced by:
 - **Complexity**
 - **Time** it takes to generate the clusters, and
 - **Number of observations** it can process.
- Certain methods require **number of clusters** to be generated **chosen prior** to the analysis; Other methods this **number determined** based on an **analysis** of the results.
- Clustering is **dependent** on the **data set being analyzed**:
 - Different data sets result in different groups of observations, even if data sets contain similar content.
 - Some cases: adding single new observation to existing set can have substantial affect on the groupings.
 - Ties in similarity scores can affect groupings, and reordering the observations may give different results.
- Number of **clustering methods**, the most common:

Hierarchical

Partitioned

Fuzzy

Hierarchical.

Organize observations hierarchically.

For example, **animal dataset** can be organized by **four high level** categories which are then subsequently organized into **sub-categories**,

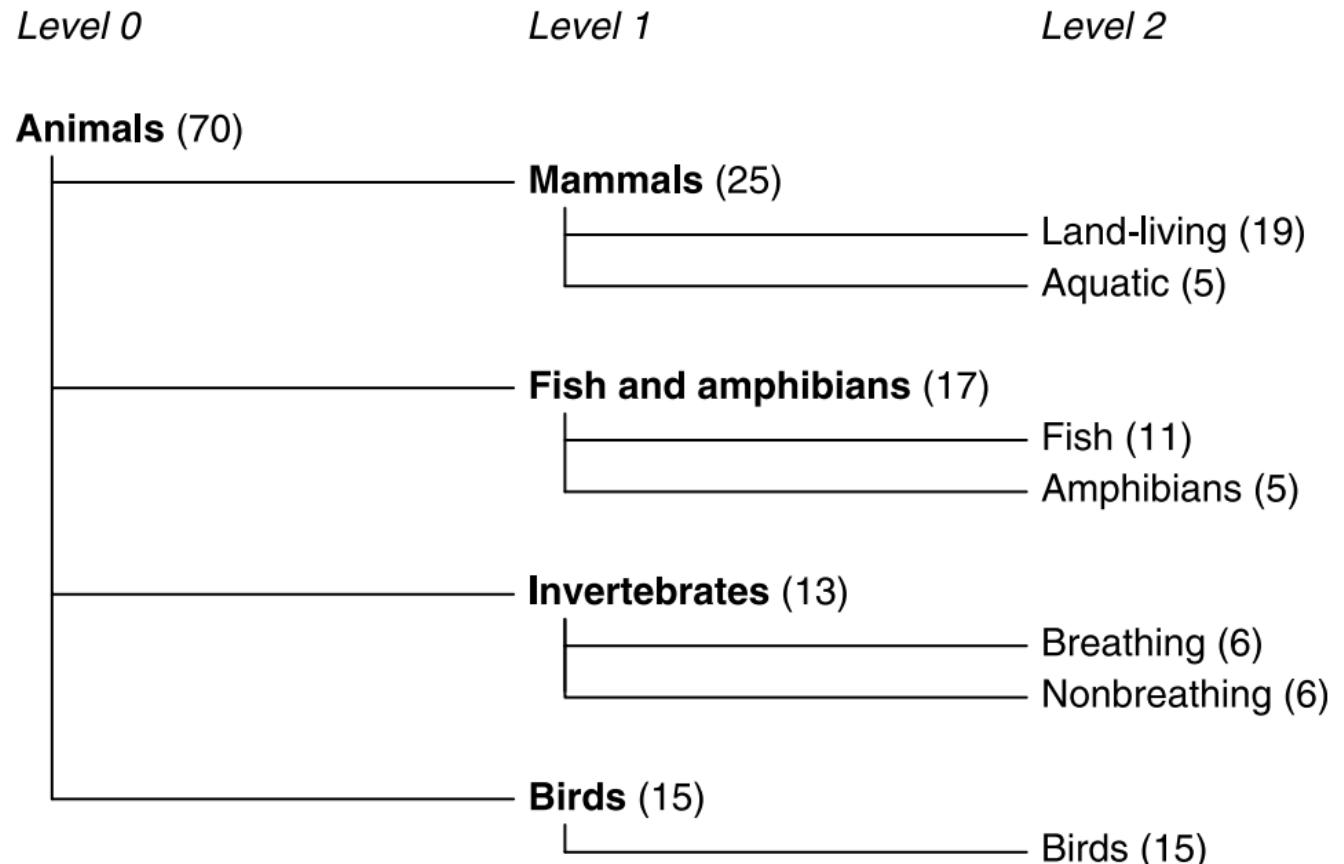


Figure 3.5 Hierarchical relationships for the animal data set

Partitioned.

Assigns observations to a single cluster and number of clusters is set prior to any cluster analysis.

For example, Fig. 3.1 generated as result of clustering animal data set where **number of clusters** was **preset** to **four**.

Does not provide a **flexible hierarchical organization**, is generally faster to compute and can handle larger data sets

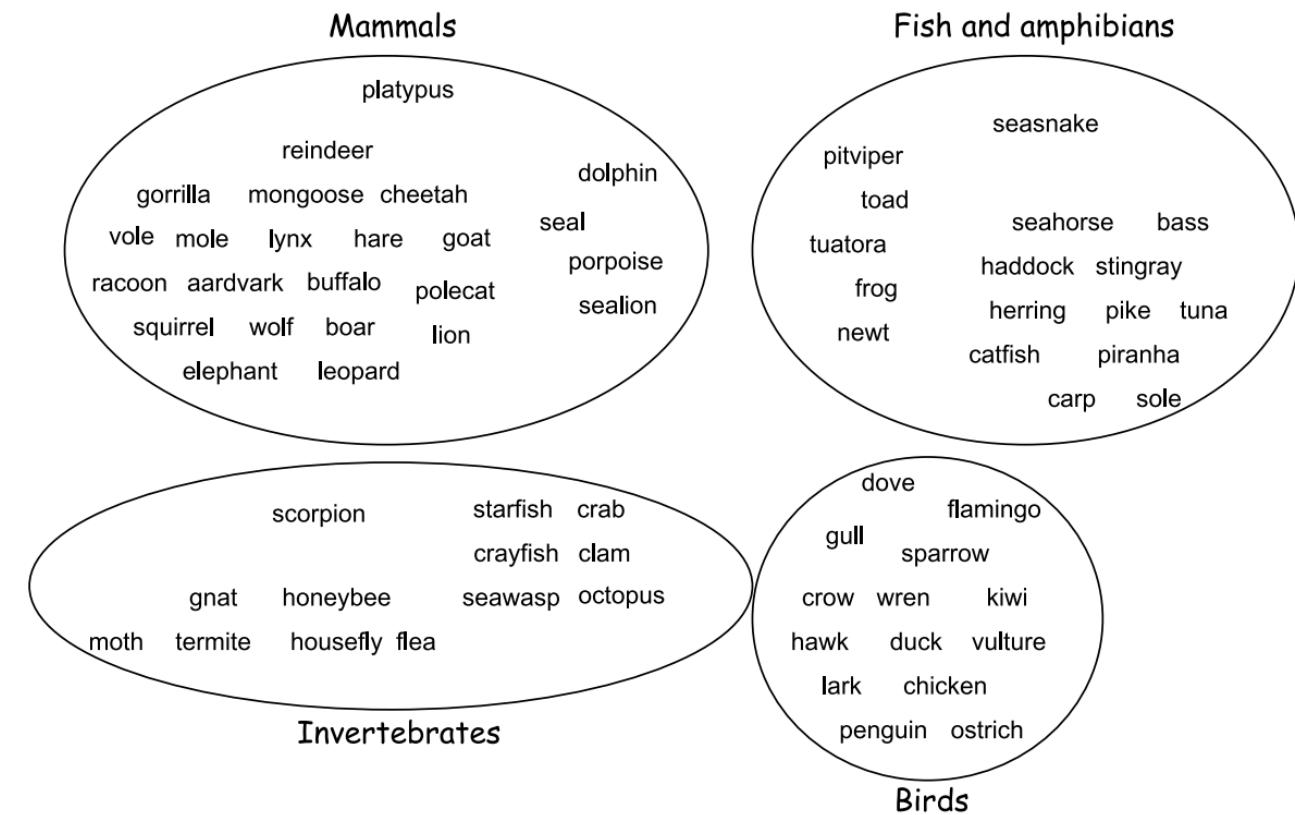


Figure 3.1 Four groups of animals in the data set

Distance measures.

- A **requirement** for any **clustering exercise** is **calculating the distance between two observations**.
- **All clustering approaches** require a **formal approach** to defining how **similar two observations** are to each other **as measured** by the **distance between them**.

Distance: Method of clustering to measure how similar observations are to each other

Distances.

To calculate similarity, **compute** the distance between observations.

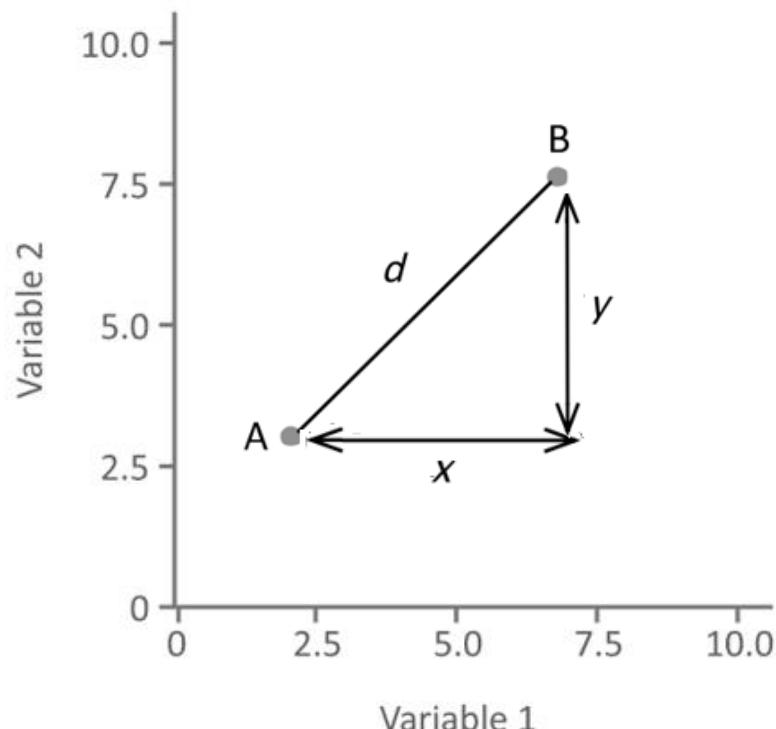


TABLE 5.1 Table Showing Two Observations A and B

Observation ID	Variable 1	Variable 2
A	2	3
B	7	8

Physical distance

FIGURE 5.6 Distance between two observations A and B.

TABLE 3.3 Information on Customers

Customer	Geographic	Income	Employer information	Age	Children	Gender	Electronics	Groceries	Clothes
A	Urban	85,000	Technology	31	0	Male	15	1	0
B	Suburbs	38,000	Services	43	3	Female	1	33	8
C	Suburbs	150,000	Legal	51	2	Female	5	0	0
D	Suburbs	32,000	Services	49	1	Female	2	29	9
E	Urban	104,000	Technology	36	0	Male	8	0	1

- Starting point for any cluster analysis is **data table or data matrix**.
- Example: Retail company gathered information about its customers: addresses, income, occupations, along with purchase categories information those customers made in the year.
- Table 3.3: Five customer entries. **Cluster analysis can only operate on numerical data**.
- Any nonnumeric values** should be **converted to numbers**.

TABLE 3.4 Information on Customers Converted to Numbers

Customer	Transformed values										
	Urban	Income	Technology	Services	Legal	Age	Children	Female	Electronics	Groceries	Clothes
A	1	85,000	1	0	0	31	0	0	15	1	0
B	0	38,000	0	1	0	43	3	1	1	33	8
C	0	150,000	0	0	1	51	2	1	5	0	0
D	0	32,000	0	1	0	49	1	1	2	29	9
E	1	104,000	1	0	0	36	0	0	8	0	1

Distances.

TABLE 3.5 Information on Customers, Normalized to the Range 0 to 1

Customer	Urban	Income	Technology	Services	Legal	Age	Children	Female	Electronics	Groceries	Clothes
A	1	0.449	1	0	0	0	0	0	1	0.03	0
B	0	0.051	0	1	0	0.6	1	1	0	1	0.89
C	0	1.000	0	0	1	1	0.67	1	0.29	0	0
D	0	0.000	0	1	0	0.9	0.33	1	0.071	0.88	1
E	1	0.610	1	0	0	0.25	0	0	0.5	0	0.11

- Common to **normalize the data** to standard ranges, such as **between zero and one** or using **z-score** (the number of standard deviations above or below the mean), to ensure a variable is not considered more important as a consequence of the range on which it is measured.
- Table 3.5: Variables: **income, age, children, electronics, groceries, and clothes** have all been normalized (range 0-1).
- After deciding which variables to use, a **method for determining the distance** must be selected. Many methods to choose from: **Euclidean, Manhattan**.

Distances.

- Distance measures generally **share certain properties**:
 1. Any distance measure **between two observations** ≥ 0
 2. **Distance** between observations A and B is the **same** as the distance between observations B and A.
 3. If **distance is zero**, there is **no difference** between the two observations, that is, the two observations have the **same values** for **all variables**.
 4. Distance between A and B ($d_{A,B}$) satisfies the following assumptions with respect to a third observation (C), based on the distance between A and C ($d_{A,C}$) and the distance between B and C ($d_{B,C}$):

$$d_{A,B} \leq d_{A,C} + d_{B,C}$$



After the break.

Distances.

- Clustering methods generally require a **distance matrix** or **dissimilarity matrix (D)** as input.
 - **Square n-by-n matrix**, where **n** is number of observations in the data set to be clustered.
- This distance matrix has the following format:

Diagonal values are all zero
distances since there **is no**
distance between two
identical observations

$$D = \begin{bmatrix} & 1 & 2 & 3 & \dots & n \\ 1 & 0 & d_{1,2} & d_{1,3} & \dots & d_{1,n} \\ 2 & d_{2,1} & 0 & d_{2,3} & \dots & d_{2,n} \\ 3 & d_{3,1} & d_{3,2} & 0 & \dots & d_{3,n} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ n & d_{n,1} & d_{n,2} & d_{n,3} & \dots & 0 \end{bmatrix}$$

Where:

$d_{i,j}$: Distance value between pairs of observations;

$d_{1,2}$: Distance between first and second observation..

1. Numeric Distance Measures.

Numeric Distance Measures.

- Distance measures “**compute a number**” for **any pair of observations**.
- These **measures compare the values for each of the variables** in the **two observations** and **compute a distance** based on some **function relating to the differences** between these values

TABLE 3.6 Six Numeric Variables Describing Five Customers

Customer	Income	Age	Children	Electronics	Groceries	Clothes
A	0.449	0	0	1	0.03	0
B	0.051	0.6	1	0	1	0.89
C	1.000	1	0.67	0.29	0	0
D	0.000	0.9	0.33	0.071	0.88	1
E	0.610	0.25	0	0.5	0	0.11

Table is limited to six normalized variables (income, age, children, electronic, groceries, and clothes)

Distances: Calculation

Euclidean.

- Distance between the **two observations** is calculated using **simple trigonometry**:

$$x = 7 - 2 = 5$$

$$y = 8 - 3 = 5$$

$$d = \sqrt{x^2 + y^2} = \sqrt{25 + 25} = 7.07$$

- **Euclidean distance** is one of most common distance functions.
- If a data set only has **two variables**, then Euclidean distance would **calculate the physical distance** between the two points plotted on a scatterplot.

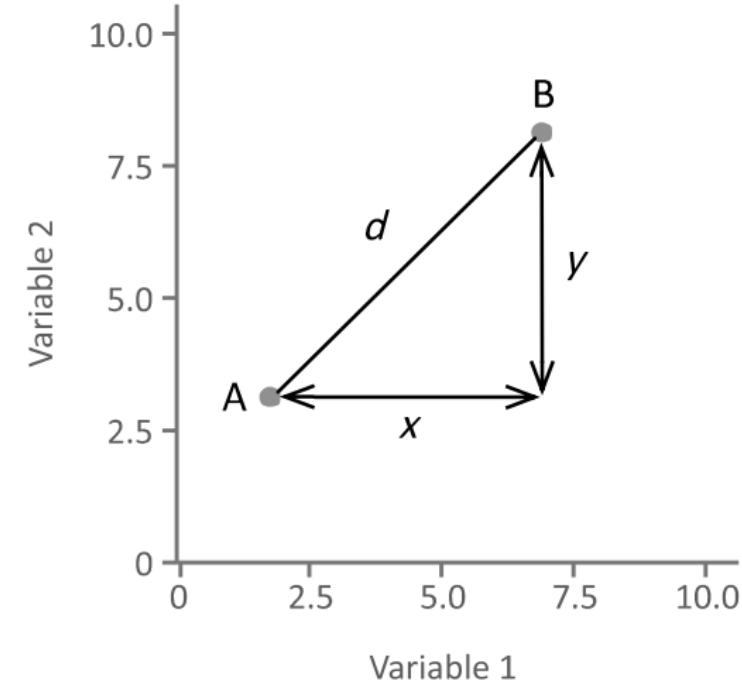


FIGURE 5.6 Distance between two observations A and B.

Euclidean Distance (d).

The following formula calculates the **Euclidean distance** between **two observations (p and q)**, measured over **n variables**:

$$d_{p,q} = \sqrt{\sum_{i=1}^n (p_i - q_i)^2}$$

TABLE 5.2 Three Observations with Values for Five Variables

ID	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5
A	0.7	0.8	0.4	0.5	0.2
B	0.6	0.8	0.5	0.4	0.2
C	0.8	0.9	0.7	0.8	0.9

The Euclidean distance between A and B is

$$d_{A-B} = \sqrt{(0.7 - 0.6)^2 + (0.8 - 0.8)^2 + (0.4 - 0.5)^2 + (0.5 - 0.4)^2 + (0.2 - 0.2)^2}$$

$$d_{A-B} = 0.17$$

The Euclidean distances between A and C is

$$d_{A-C} = \sqrt{(0.7 - 0.8)^2 + (0.8 - 0.9)^2 + (0.4 - 0.7)^2 + (0.5 - 0.8)^2 + (0.2 - 0.9)^2}$$

$$d_{A-C} = 0.83$$

The Euclidean distance between B and C is

$$d_{B-C} = \sqrt{(0.6 - 0.8)^2 + (0.8 - 0.9)^2 + (0.5 - 0.7)^2 + (0.4 - 0.8)^2 + (0.2 - 0.9)^2}$$

$$d_{B-C} = 0.86$$

The distance between **A and B** is **0.17**, whereas distance between **A and C** is **0.83**, which indicates that there is **more similarity between observations A and B than A and C. C is not closely related to either A or B.**

Euclidean Distance (d).

- Figure 5.7 where the **values for each variable** are plotted along the **horizontal axis** and the **height** of the bar measured against the **vertical axis** represents the **data value**.
- The **shape of histograms A and B** are **similar**, whereas the shape of histogram **C** is **not similar to A or B**.

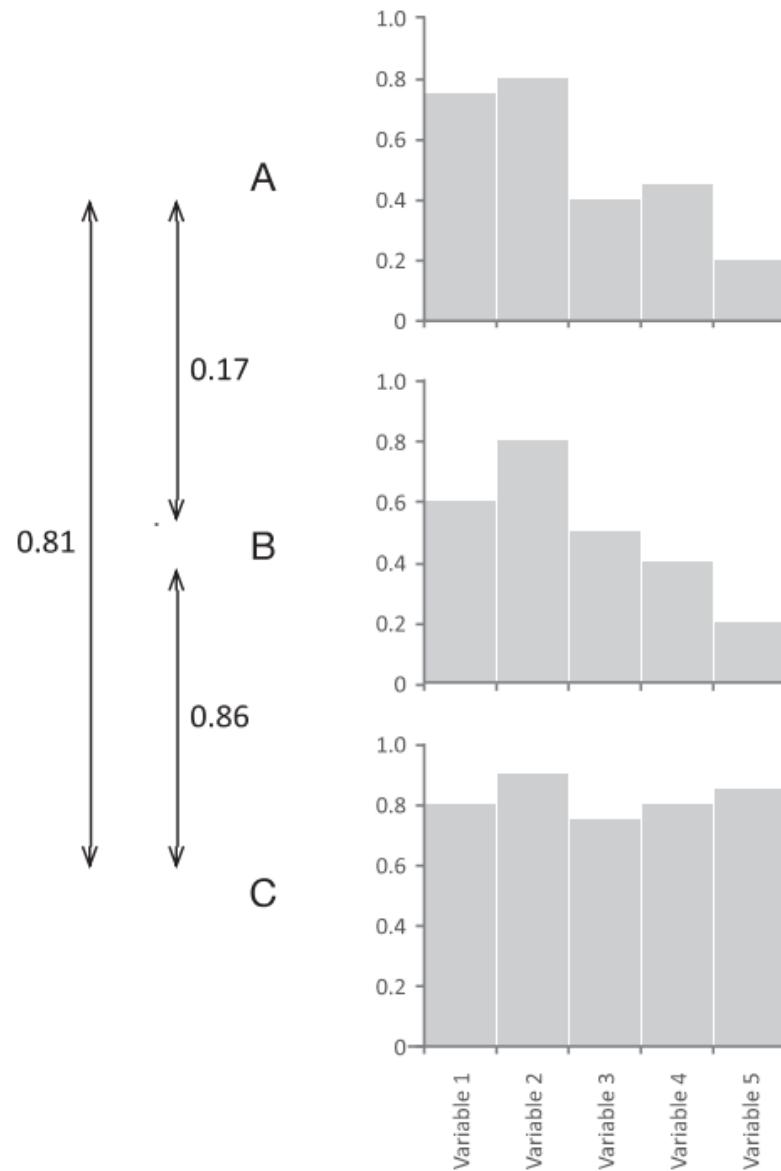


FIGURE 5.7 Distances between three observations: A–B, B–C, and A–C.

Euclidean Distance (d).

TABLE 3.6 Six Numeric Variables Describing Five Customers

Customer	Income	Age	Children	Electronics	Groceries	Clothes
A	0.449	0	0	1	0.03	0
B	0.051	0.6	1	0	1	0.89
C	1.000	1	0.67	0.29	0	0
D	0.000	0.9	0.33	0.071	0.88	1
E	0.610	0.25	0	0.5	0	0.11

The Euclidean distance between A and B is

$$d_{p,q} = \sqrt{\sum_{i=1}^n (p_i - q_i)^2}$$

$$d_{A,B} = \sqrt{(0.449 - 0.051)^2 + (0 - 0.6)^2 + (0 - 1)^2 + (1 - 0)^2 + (0.03 - 1)^2 + (0 - 0.89)^2}$$

$$d_{A,B} = 2.061$$

Square Euclidean.

TABLE 3.6 Six Numeric Variables Describing Five Customers

Customer	Income	Age	Children	Electronics	Groceries	Clothes
A	0.449	0	0	1	0.03	0
B	0.051	0.6	1	0	1	0.89
C	1.000	1	0.67	0.29	0	0
D	0.000	0.9	0.33	0.071	0.88	1
E	0.610	0.25	0	0.5	0	0.11

- The **square Euclidean** is the **sum of the squares** of the difference between the two observations, and it is calculated using the following formula:

$$d_{p,q} = \sum_{i=1}^n (p_i - q_i)^2$$

- To **calculate the square euclidean** distance between observations A and B using the data in Table 3.6, the following calculation is made:

$$d_{A,B} = (0.449 - 0.051)^2 + (0 - 0.6)^2 + (0 - 1)^2 + (1 - 0)^2$$

$$+ (0.03 - 1)^2 + (0 - 0.89)^2 \quad d_{A,B} = 4.249$$

Manhattan Distance.

TABLE 3.6 Six Numeric Variables Describing Five Customers

Customer	Income	Age	Children	Electronics	Groceries	Clothes
A	0.449	0	0	1	0.03	0
B	0.051	0.6	1	0	1	0.89
C	1.000	1	0.67	0.29	0	0
D	0.000	0.9	0.33	0.071	0.88	1
E	0.610	0.25	0	0.5	0	0.11

Manhattan distance, also called the **city block distance**, is the **sum of the absolute distances** between the variables, which is **always a positive value** representing the difference. The formula is

$$d_{p,q} = \sum_{i=1}^n |p_i - q_i|$$

$$\begin{aligned} d_{A,B} &= |0.449 - 0.051| + |0 - 0.6| + |0 - 1| + |1 - 0| \\ &\quad + |0.03 - 1| + |0 - 0.89| \end{aligned}$$

$$d_{A,B} = 4.856$$



Binary distance measures.

Binary Distance Measures.

- Distances calculated for **categorical values** because they can be easily converted to binary dummy variables.
- For each variable, the values in the two observations are compared to determine whether they are **the same or different**.

TABLE 3.8 Four Alternatives (A, B, C, and D) for Comparing Two Binary Values

		q_i	
p_i	1	1 A	0 B
	0	0 C	1 D

TABLE 3.7 Distances between all Pairs of Observations

		Euclidean	Square euclidean	Manhattan
	A,B	2.061	4.249	4.856
	A,C	1.502	2.258	2.962
	A,D	1.924	3.705	4.459
	A,E	0.593	0.351	1.052
	B,C	1.744	3.043	3.857
	B,D	0.754	0.569	1.321
	B,E	1.813	3.29	4.187
	C,D	1.714	2.939	3.526
	C,E	1.103	1.217	2.131
	D,E	1.628	2.651	3.789

Binary Distance Measures.

- **Similarity measures** explain how **alike** two **observations** are **to each other**, with **high similarity values** representing situations when the **two observations are alike**.
- This contrasts with **distance (or dissimilarity)** measures, where **low values** indicate the **observations are alike**.
- The similarity and distance calculations for binary variables are based on the **number of common and different values in the four situations** summarized in Table 3.8 for **two observations p and q, over all variables**

TABLE 3.8 Four Alternatives (A , B , C , and D) for Comparing Two Binary Values

		q_i	
	1	1	0
p_i	1	A	B
	0	C	D

- a:** Number of variables where the value for both p and q is one (A)
b: Number of variables where the value for p is one and the value for q is zero (B)
c: Number of variables where the value for p is zero and the value for q is one (C)
d: Number of variables where the value for both p and q is zero (D)

Binary Distance Measures.

- Take **five binary variables**: **urban, technology, services, legal, and female**.
- Using Table 3.9, the values for a, b, c, and d for customers B and C are calculated as:

TABLE 3.9 Five Customers with Data for Five Binary Variables

Customer	Urban	Technology	Services	Legal	Female
A	1	1	0	0	0
B	0	0	1	0	1
C	0	0	0	1	1
D	0	0	1	0	1
E	1	1	0	0	0

- a: 1 (since female is both one in observations B and C);
b: 1 (since services is one in observation B and zero in observation C);
c: 1 (since legal is zero in observation B and one in observation C);
d: 2 (since urban and technology are both zero in observations B and C).

Since these **four counts cover all possible situations**, summing a, b, c, and d should **equal the total number of variables** selected, which is five in this example.

Simple matching.

- Calculates **number of common ones or zeros as a proportion of all variables**
- Formula is used to calculate the **similarity coefficients ($s_{p,q}$)**

$$s_{p,q} = \frac{a + d}{a + b + c + d}$$

TABLE 3.9 Five Customers with Data for Five Binary Variables

Customer	Urban	Technology	Services	Legal	Female
A	1	1	0	0	0
B	0	0	1	0	1
C	0	0	0	1	1
D	0	0	1	0	1
E	1	1	0	0	0

The corresponding distance calculation is:

$$d_{p,q} = 1 - \frac{a + d}{a + b + c + d}$$

Using Table 3.9, the simple distance between B and C is: 0.6

$$d_{B,C} = 1 - \frac{1 + 2}{1 + 1 + 1 + 2} \quad d_{B,C} = 0.4$$

Jaccard.

- Calculates the proportion of common ones against the total number of values that are one in either or both observations.
- Five Customers with Data for Five Binary Variables Customer, Urban, Technology, Services, Legal and Female the method does not incorporate d (the number of values where the variables are both zero).

The corresponding distance calculation is:

Using Table 3.9, the simple distance between B and C is:

TABLE 3.9 Five Customers with Data for Five Binary Variables

Customer	Urban	Technology	Services	Legal	Female
A	1	1	0	0	0
B	0	0	1	0	1
C	0	0	0	1	1
D	0	0	1	0	1
E	1	1	0	0	0

Formula is used to calculate the similarity coefficients ($s_{p,q}$):

$$s_{p,q} = \frac{a}{a + b + c}$$

$$s_{B,C} = 0.33$$

$$d_{p,q} = \frac{b + c}{a + b + c}$$

$$d_{B,C} = \frac{1 + 1}{1 + 1 + 1} \quad d_{B,C} = 0.67$$

An aerial photograph of a city street at night, showing the intersection of two major roads. The scene is filled with long exposure light trails from vehicles, creating vibrant streaks of red, yellow, and blue across the dark asphalt. The surrounding environment includes modern skyscrapers with illuminated windows, streetlights, and a few small figures of people on the sidewalks. The overall atmosphere is dynamic and urban.

AFTER THE
BREAK.

Why grouping?

- **Finding hidden relationships: organize observations in different ways.**
 - Data set of retail transactions is grouped and these groups are used to find nontrivial associations, such as customers who purchase doormats often purchase umbrellas at the same time
- **Becoming familiar with the data: types of observations are present in the data.**
 - Database of medical records will be used to create a general model for predicting a number of medical conditions; data set is characterized by grouping the observations; both male and female patients with a variety of conditions; young female patients having flu.
- **Segmentation:** Techniques for grouping data may lead to divisions that simplify the data for analysis.

Grouping Approaches.

- Supervised and Unsupervised
- Types of variables
- Data set size limit
- Interpretable and actionable
- Overlapping group/mutually exclusive

Grouping Approaches.

- **Supervised versus unsupervised:**
- **One distinction** between the different methods is whether they **use the response variable to guide how the groups are generated**.
- Methods that **do not use any data to guide** how the groups are generated are called **unsupervised methods**, whereas methods that make use of the response variable to **guide group generation** are called **supervised methods**.
- For example,
 - A data set of cars could be grouped using an unsupervised method. The groups generated would be based on general classes of cars.
 - Alternatively, we could group the cars using car fuel efficiency to direct the grouping.

Good Clustering

- A **good clustering method** will produce high quality clusters with
 - High intra-class similarity (within groups)
 - Low inter-class similarity (between groups)
- Clustering methods (*unsupervised*) can be
 - **Hierarchical**
 - **Partitioning**
 - Others (Density based, grid based, model based)

Given dataset, no class label

Hierarchical decomposition
Divisive or bottom-up
CHAMELEON and BIRCH

Centroid based techniques
k-means
k-medoids
K: no. of partitions
PAM, CLARA, CLARANS

1. AGGLOMERATIVE HIERARCHICAL CLUSTERING

- Example of a hierarchical method for grouping observations.
- Uses a “**bottom-up**” approach to clustering as it **starts with each observation** and **progressively creates cluster** by merging **observations** together until all are a **member** of a **final single cluster**.
- The **major limitation** of agglomerative hierarchical clustering is that it is normally **limited to data sets with fewer than 10,000 observations** because the **computational cost** to generate the hierarchical tree can be **high**, especially for larger numbers of observations/expensive.

AGGLOMERATIVE HIERARCHICAL CLUSTERING

TABLE 5.4 Table of Observations to Cluster

	Name	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5	
1	A	7.9	8.6	4.4	5.0	2.5	
2	B	6.8	8.2	5.2	4.2	2.2	
3	C	8.7	9.6	7.5	8.9	9.8	
4	D	6.1	7.3	7.9	7.3	8.3	Measured over 5 variables
5	E	1.5	2.0	5.1	3.6	4.2	
6	F	3.7	4.3	5.4	3.3	5.8	Measured on same scale
7	G	7.2	8.5	8.6	6.7	6.1	
8	H	8.5	9.7	6.3	5.2	5.0	
9	I	2.0	3.4	5.8	6.1	5.6	
10	J	1.3	2.6	4.2	4.5	2.1	
11	K	3.4	2.9	6.5	5.9	7.4	
12	L	2.3	5.3	6.2	8.3	9.9	
13	M	3.8	5.5	4.6	6.7	3.3	
14	N	3.2	5.9	5.2	6.2	3.7	

STEPS for AGGLOMERATIVE HIERARCHICAL CLUSTERING

1. **Distance** between **all pairs** of observations is calculated.

Method for calculating the distance along with the variables to include in the calculation should be set **prior to clustering** (*Euclidean distance* across all continuous variables shown in Table 5.4.)

Distances between **all combinations of observations** are **summarized** in a **distance matrix**, as illustrated in Table 5.5.

- Diagonal values are excluded, since these pairs are of the same observation. It should be noted that a **distance matrix** is usually **symmetric** about the diagonal as the distance between, for example, A and B is the same as the distance between B and A.

STEPS for AGGLOMERATIVE HIERARCHICAL CLUSTERING

TABLE 5.6 Calculated Distances Between All Pairs of Observations

A	B	C	D	E	F	G	H	I	J	K	L	M	N
A	0.282	1.373	1.2	1.272	0.978	1.106	0.563	1.178	1.189	1.251	1.473	0.757	0.793
B	0.282	1.423	1.147	1.113	0.82	1.025	0.56	1.064	1.065	1.144	1.44	0.724	0.7
C	1.373	1.423		0.582	1.905	1.555	0.709	0.943	1.468	1.995	1.305	1.076	1.416
D	1.2	1.147	0.582		1.406	1.092	0.403	0.808	0.978	1.543	0.797	0.744	1.065
E	1.272	1.113	1.905	1.406		0.476	1.518	1.435	0.542	0.383	0.719	1.223	0.797
F	0.978	0.82	1.555	1.092	0.476		1.191	1.039	0.57	0.706	0.595	1.076	0.727
G	1.106	1.025	0.709	0.403	1.518	1.191		0.648	1.163	1.624	1.033	1.108	1.148
H	0.563	0.56	0.943	0.808	1.435	1.039	0.648		1.218	1.475	1.169	1.315	0.984
I	1.178	1.064	1.468	0.978	0.542	0.57	1.163	1.218		0.659	0.346	0.727	0.553
J	1.189	1.065	1.995	1.543	0.383	0.706	1.624	1.475	0.659		0.937	1.344	0.665
K	1.251	1.144	1.305	0.797	0.719	0.595	1.033	1.169	0.346	0.937		0.64	0.774
L	1.473	1.44	1.076	0.744	1.223	1.076	1.108	1.315	0.727	1.344	0.64		0.985
M	0.757	0.724	1.416	1.065	0.797	0.727	1.148	0.984	0.553	0.665	0.774	0.985	
N	0.793	0.7	1.378	0.974	0.727	0.624	1.051	0.937	0.458	0.659	0.683	0.919	0.196

STEPS for AGGLOMERATIVE HIERARCHICAL CLUSTERING

Two closest observations are identified (**M** and **N**) and are merged into a **single cluster**.

These **two observations** from now on will be considered a **single group**

TABLE 5.5 Distance Matrix Format

	A	B	C	D	...
A		$d_{A,B}$	$d_{A,C}$	$d_{A,D}$...
B	$d_{B,A}$		$d_{B,C}$	$d_{B,D}$...
C	$d_{C,A}$	$d_{C,B}$		$d_{C,D}$...
D	$d_{D,A}$	$d_{D,B}$	$d_{D,C}$...
...

2. All observations (minus M & N merged into a cluster) along with the newly created cluster are compared to **see which observation or cluster** should be joined into the next cluster.

Analyse individual observations and clusters, **so, a joining or linkage rule** is needed to determine the distance between an observation and a cluster of observations. This joining/linkage rule should be set prior to clustering.

JOINING OR LINKAGE RULE.

Average linkage

Distance between all members of cluster (a, b, and c) and observation under consideration (X) are determined.

Average is calculated.

Single linkage

Distance between all members of cluster (a, b, and c) and observation under consideration (X) are determined.

Smallest is selected.

Complete linkage

Distance between all members of cluster (a, b, and c) and observation under consideration X) are determined.

Highest is selected

Linkage rules for considering an observation and a cluster & two clusters.

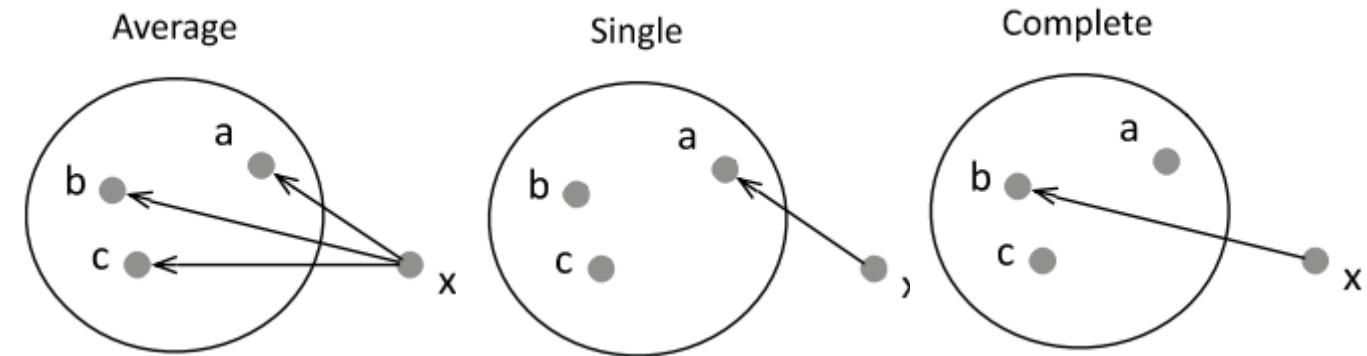
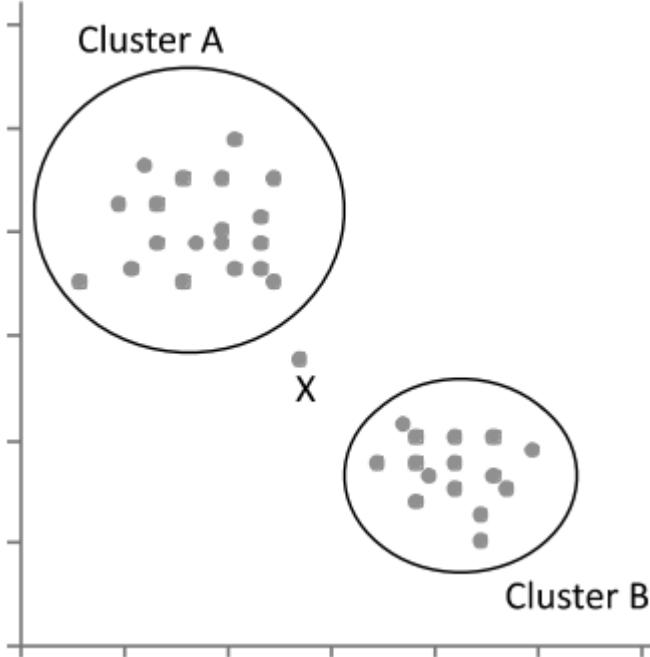


FIGURE 5.9 Different linkage rules for considering an observation and a cluster.

FIGURE 5.8 Comparing observation X with two clusters A and B.

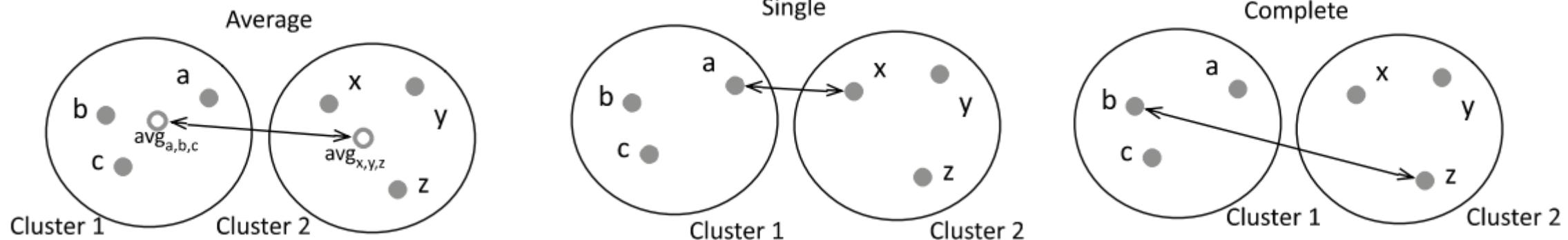


FIGURE 5.10 Different linkage rules for considering two clusters.

STEPS for AGGLOMERATIVE HIERARCHICAL CLUSTERING

- Distances between all combinations of groups and observations are considered and the **smallest distance** is selected.
 - Consider **distance between two clusters**, the **linkage/joining concept is extended to the joining of two clusters**, as illustrated in Figure 5.10.
3. Process of assessing all pairs of observations/clusters, then combining the pair with the smallest distance is repeated until there are no more clusters or observations to join together since only a single cluster remains.
- Figure 5.11 illustrates this process for some steps based on the observations shown in Table 5.6

AGGLOMERATIVE HIERARCHICAL CLUSTERING

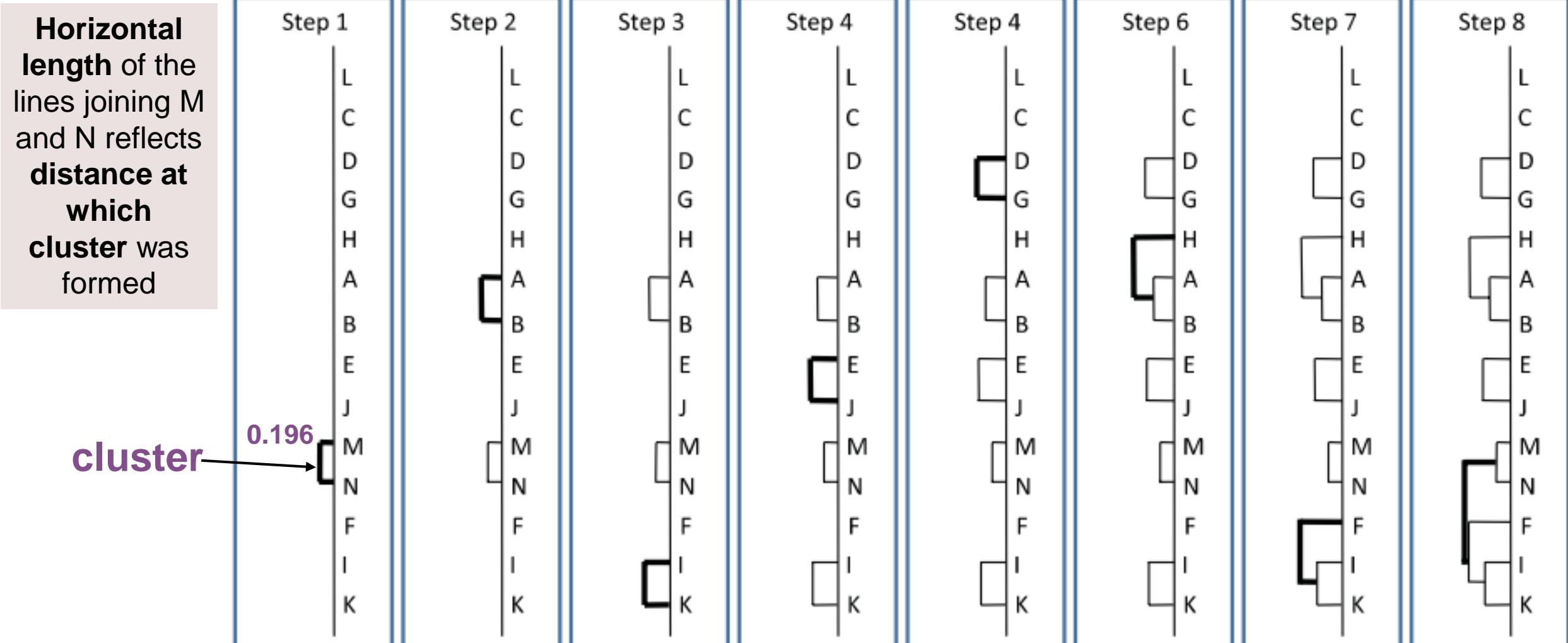


FIGURE 5.11 Steps 1 through 8 of the clustering process.

AGGLOMERATIVE HIERARCHICAL CLUSTERING

- Figure 5.12 shows **completed hierarchical clustering** for all 14 observations.
- When clustering completes, **a tree called a dendrogram** is generated showing **similarity between observations and clusters**.

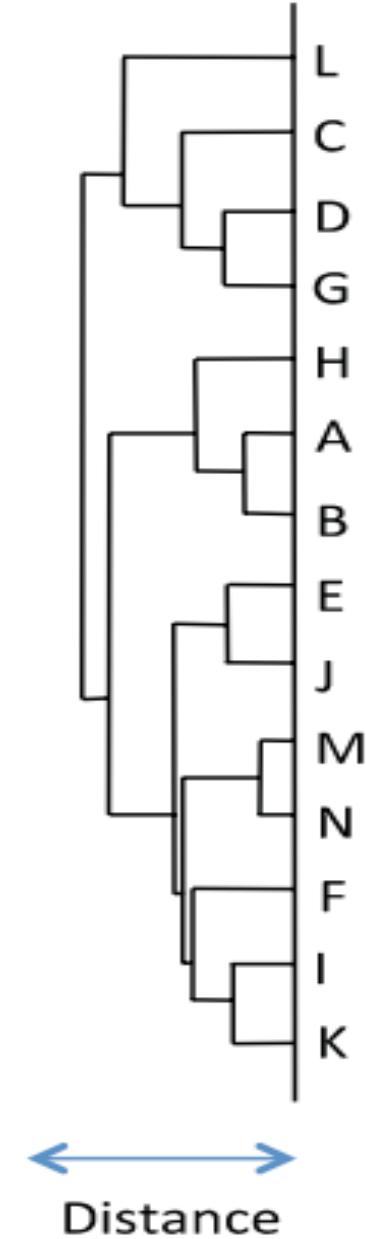


FIGURE 5.12 Completed hierarchical clustering for the 14 observations.

AGGLOMERATIVE HIERARCHICAL CLUSTERING

- To **divide a data set into a series of distinct clusters**, select a **distance** at which the clusters are to be created.
- Where this distance intersects with a horizontal line on the tree (shown by the circles), a **cluster is formed**.
- Three different distances (i , j , k) are used to divide the tree into clusters:
- Where this vertical line intersects with the tree **at distance i** : 2 clusters formed: $\{L,C,D,G\}$ and $\{H,A,B,E,J,M,N,F,I,K\}$;
- At distance j** : 4 clusters are formed: $\{L\}$, $\{C,D,G\}$, $\{H,A,B\}$, and $\{E,J,M,N,F,I,K\}$;
- At distance k** : 9 clusters are formed: $\{L\}$, $\{C\}$, $\{D,G\}$, $\{H\}$, $\{A,B\}$, $\{E,J\}$, $\{M,N\}$, $\{F\}$, and $\{I,K\}$.

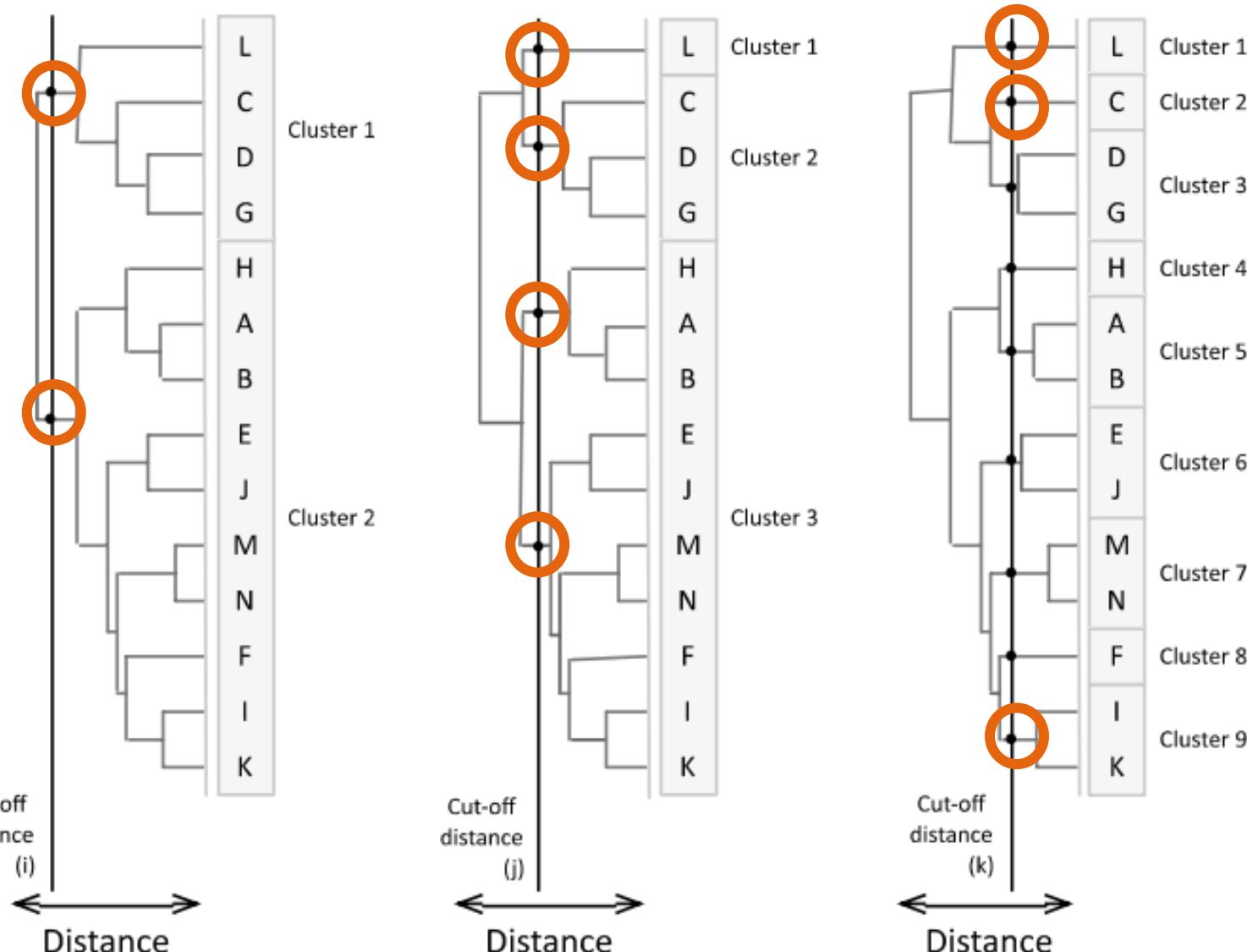


FIGURE 5.13 Cluster generation using three distance cut-offs.

AGGLOMERATIVE HIERARCHICAL CLUSTERING LINKAGE RULE choice

- Different joining/linkage rules **change** how the final hierarchical clustering is presented (Fig. 5.14)
- Since the **barrier for merging observations** and clusters is **lowest** with **single linkage approach**, the clustering dendrogram may contain **chains of clusters as well as clusters that are spread out**.
- **Barrier** to joining clusters is **highest** with **complete linkage**; it is possible that an observation is closer to observations in other clusters than the cluster to which it has been assigned.

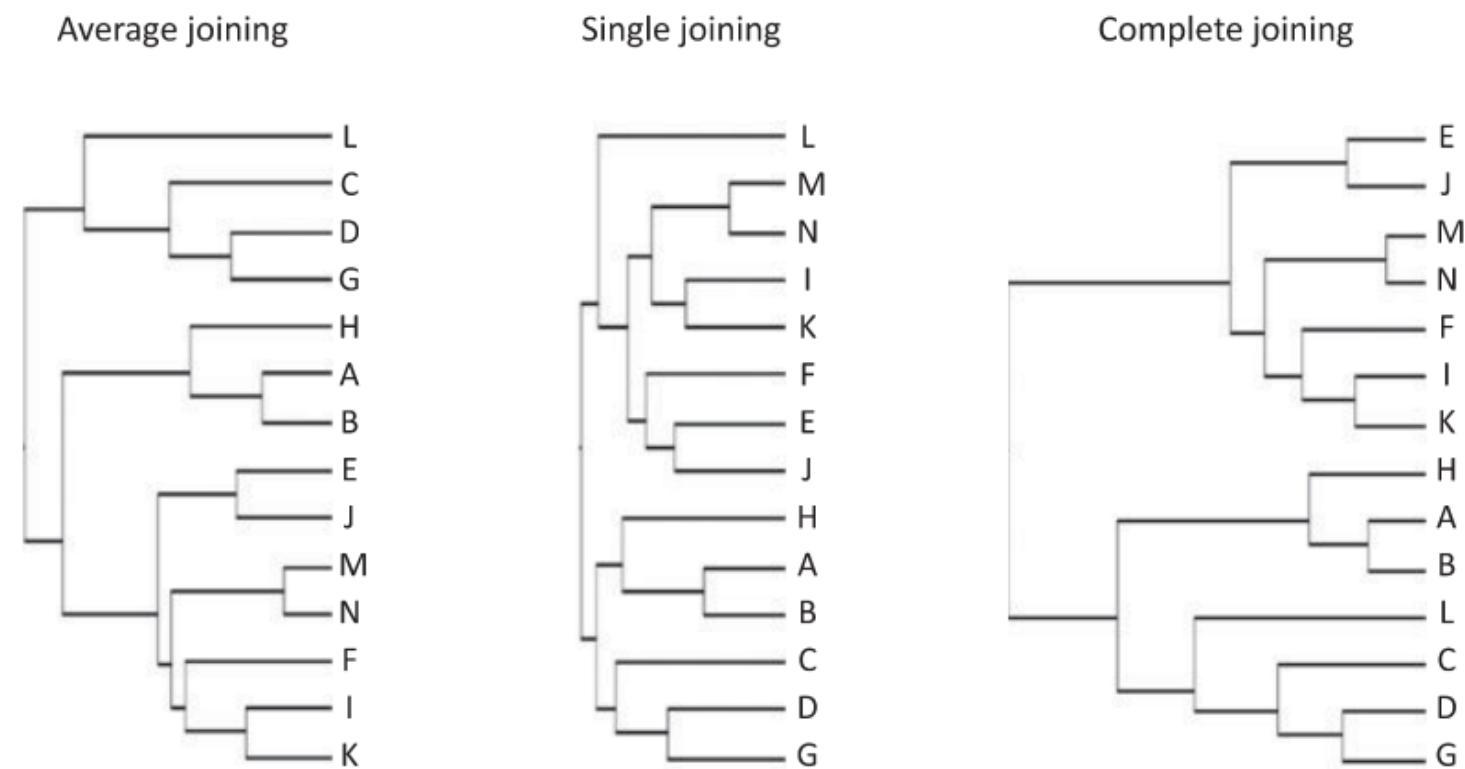


FIGURE 5.14 Different results using three different methods for joining clusters.

The **average linkage approach** **moderates** the tendencies of the single or complete linkage approaches

Example: AGGLOMERATIVE HIERARCHICAL CLUSTERING

Data set of 392 cars explored using hierarchical agglomerative clustering.

TABLE 5.7 Data Table Containing Automobile Observations

Car Name	MPG	Cylinders	Displacement	Horsepower	Weight	Acceleration	Model Year	Origin
Chevrolet Chevelle Malibu	18	8	307	130	3,504	12	70	American
Buick Skylark 320	15	8	350	165	3,693	11.5	70	American
Plymouth Satellite	18	8	318	150	3,436	11	70	American
Amc rebel sst	16	8	304	150	3,433	12	70	American
Ford Torino	17	8	302	140	3,449	10.5	70	American
Ford Galaxie 500	15	8	429	198	4,341	10	70	American

- This data set was **clustered using the Euclidean distance method and the complete linkage rule.**
- The following variables were used in the clustering: **Displacement, Horsepower, Acceleration, and MPG (miles per gallon)**

Example: AGGLOMERATIVE HIERARCHICAL CLUSTERING

TABLE 5.7 Data Table Containing Automobile Observations

Car Name	MPG	Cylinders	Displacement	Horsepower	Weight	Acceleration	Model Year	Origin
Chevrolet Chevelle Malibu	18	8	307	130	3,504	12	70	American
Buick Skylark 320	15	8	350	165	3,693	11.5	70	American
Plymouth Satellite	18	8	318	150	3,436	11	70	American
Amc rebel sst	16	8	304	150	3,433	12	70	American
Ford Torino	17	8	302	140	3,449	10.5	70	American
Ford Galaxie 500	15	8	429	198	4,341	10	70	American

Dendrogram of the generated clusters **shows relationships** between observations based on **similarity of the four selected variables**.

Each horizontal line at the right represents a single automobile and the order of the observations is related to how similar each car is to its neighbours

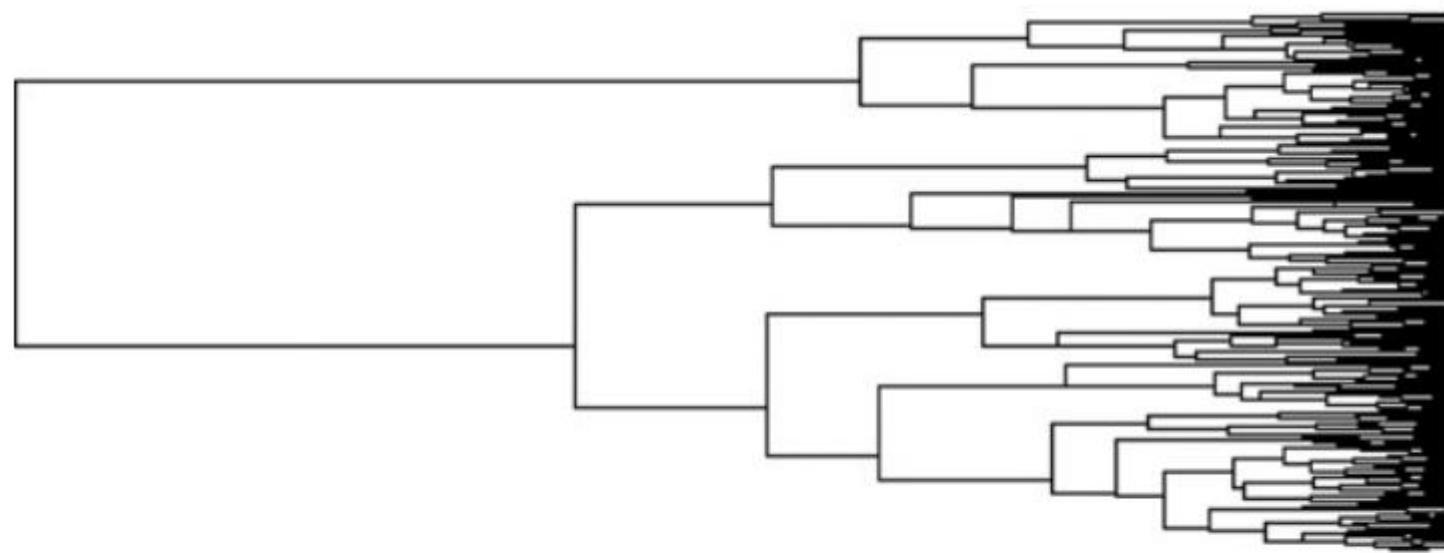
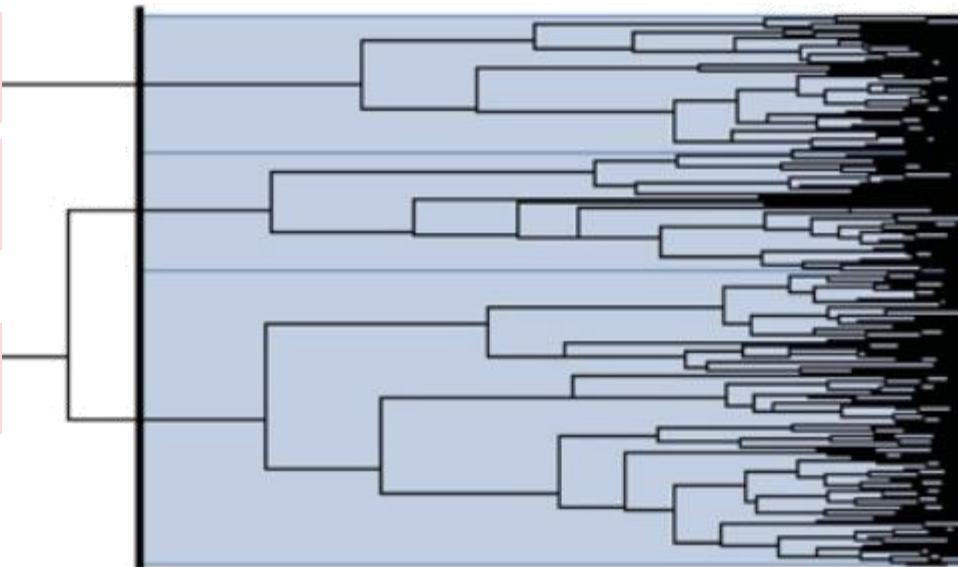


FIGURE 5.15 Hierarchical agglomerative clustering dendrogram generated for the automobile data set.

low fuel efficiency, low acceleration values, higher values for horsepower and displacement.

good fuel efficiency and acceleration as well as low horsepower and displacement

Average fuel efficiency and acceleration as well as few high values for displacement or horsepower

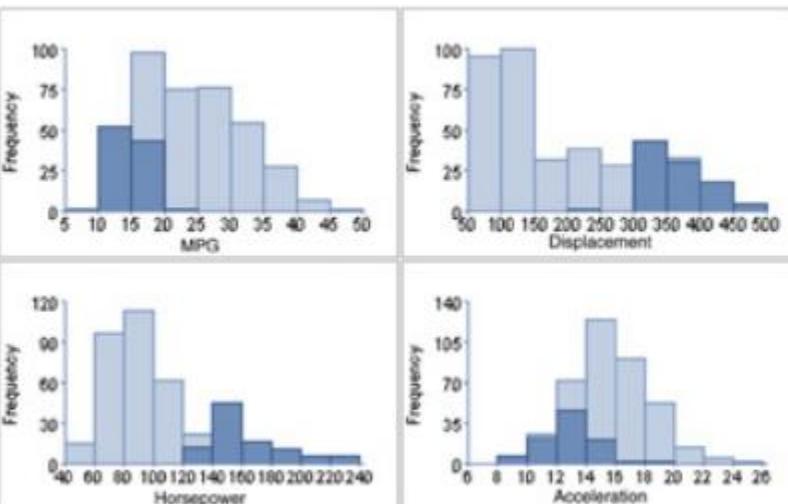


Cluster 1 (97 observations)

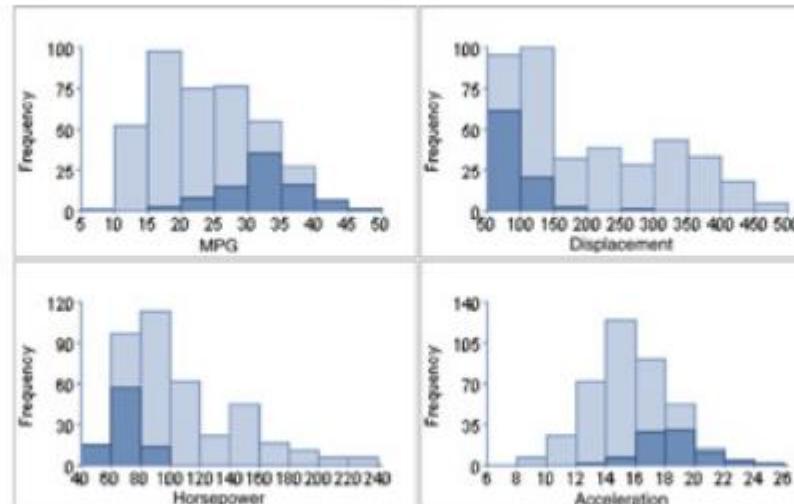
Cluster 2 (85 observations)

Cluster 3 (210 observations)

Cluster 1 (97 observations)



Cluster 2 (85 observations)



Cluster 3 (210 observations)

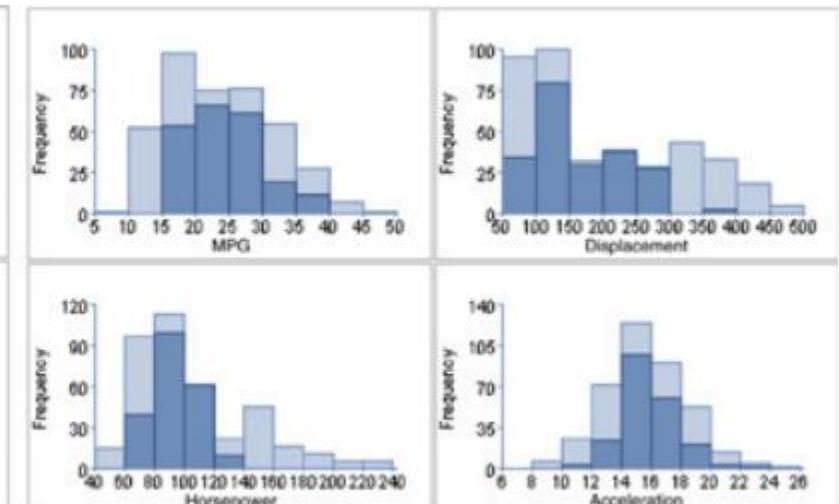


FIGURE 5.16 Automobile data set with three clusters identified.

Distance was set to create nine clusters

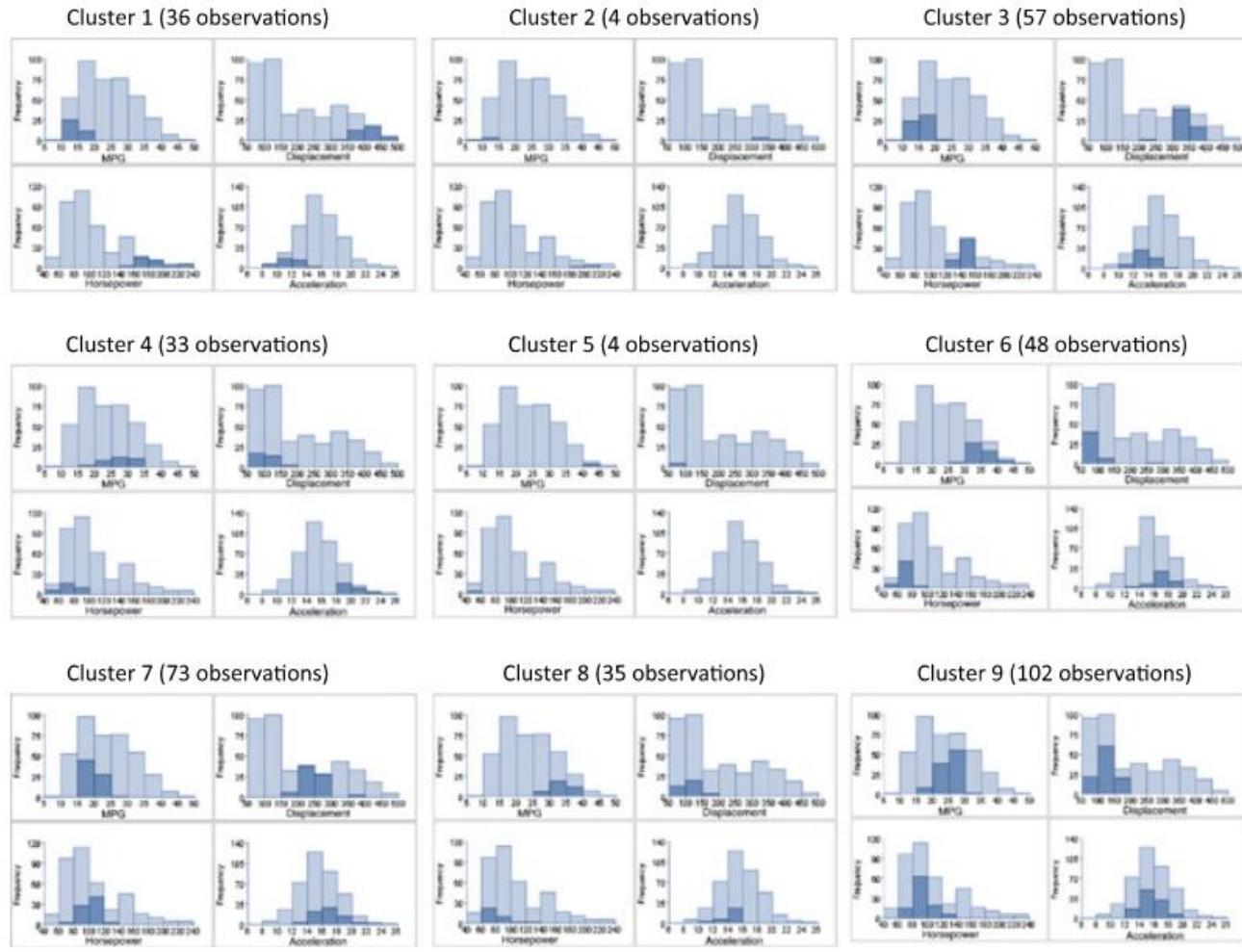


FIGURE 5.17 Automobile data set cluster and split into nine groups.

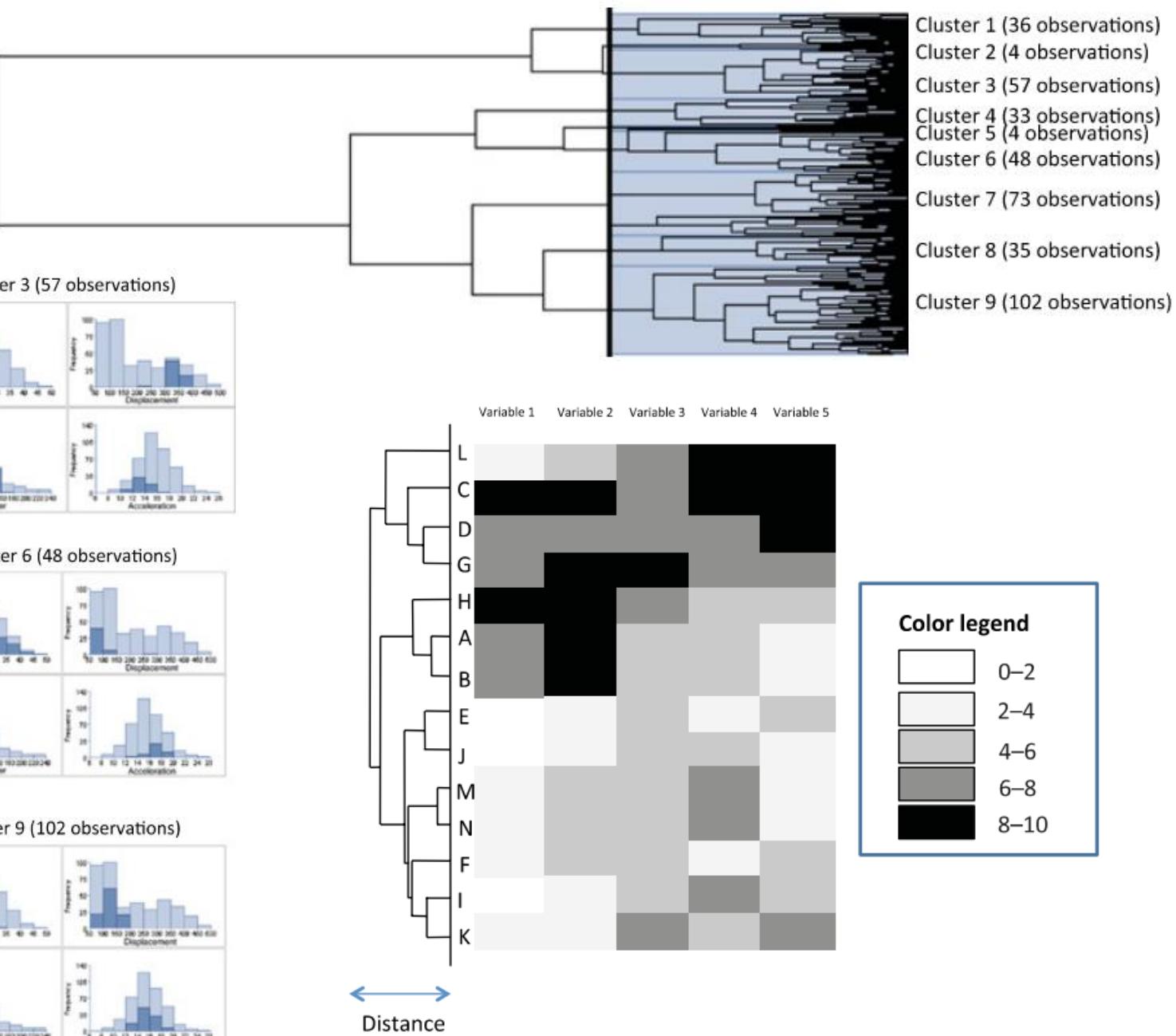
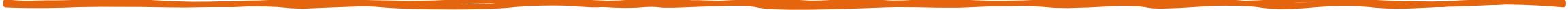


FIGURE 5.18 Clustering dendrogram coupled with a colored heatmap.

Enough for today.

Summary of steps for HAC.



1

Each observation corresponds to a single group or cluster.

2

In each iteration two most similar groups meet under some criterion of similarity between groups.

3

Two newly united groups form a single group

4

Iterate until forming a single group.

Example 2: Use HAC

Euclidean Distance

TABLE 3.14 Five Observations Measured Over Five Variables

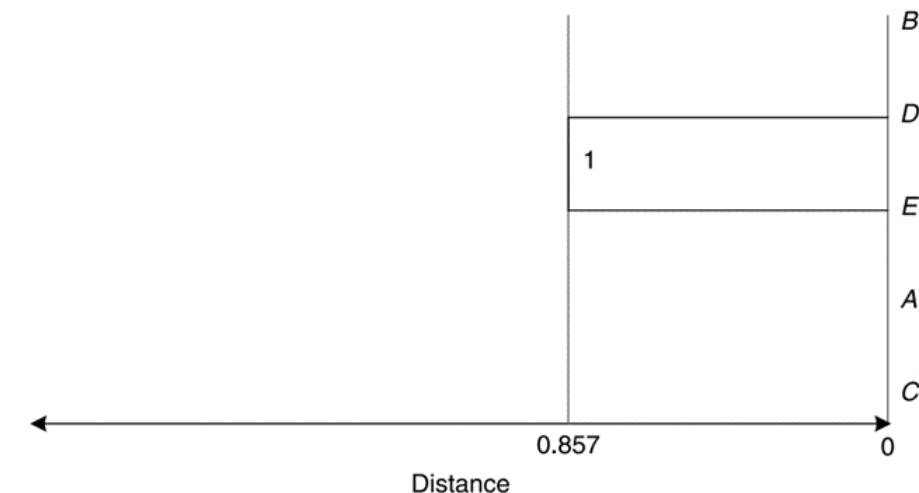
	Percentage body fat	Weight	Height	Chest	Abdomen
A	12.3	154.25	67.75	93.1	85.2
B	31.6	217	70	113.3	111.2
C	22.2	177.75	68.5	102	95
D	14.1	176	73	96.7	86.5
E	23.6	197	73.25	103.6	99.8

TABLE 3.16 Distance Matrix Using the Euclidean Distance

	{A}	{B}	{C}	{D}	{E}
{A}	0	2.041	0.871	1.036	1.547
{B}	2.041	0	1.185	1.768	1.022
{C}	0.871	1.185	0	1.011	0.941
{D}	1.036	1.768	1.011	0	0.857
{E}	1.547	1.022	0.941	0.857	0

TABLE 3.15 Normalized Data

	Percentage body fat	Weight	Height	Chest	Abdomen
A	0.0	0.0	0.0	0.0	0.0
B	1.0	1.0	0.409	1.0	1.0
C	0.513	0.375	0.136	0.441	0.377
D	0.00932	0.347	0.955	0.178	0.050
E	0.585	0.681	1.0	0.520	0.562



Single Linkage

Example 2: Use HAC

Single Linkage

TABLE 3.17 Distance Matrix for the Four Groups Using the Single Linkage Criteria

	{A}	{B}	{C}	{D,E}
{A}	0	2.041	0.871	1.036
{B}	2.041	0	1.185	1.022
{C}	0.871	1.185	0	0.941
{D,E}	1.036	1.022	0.941	0

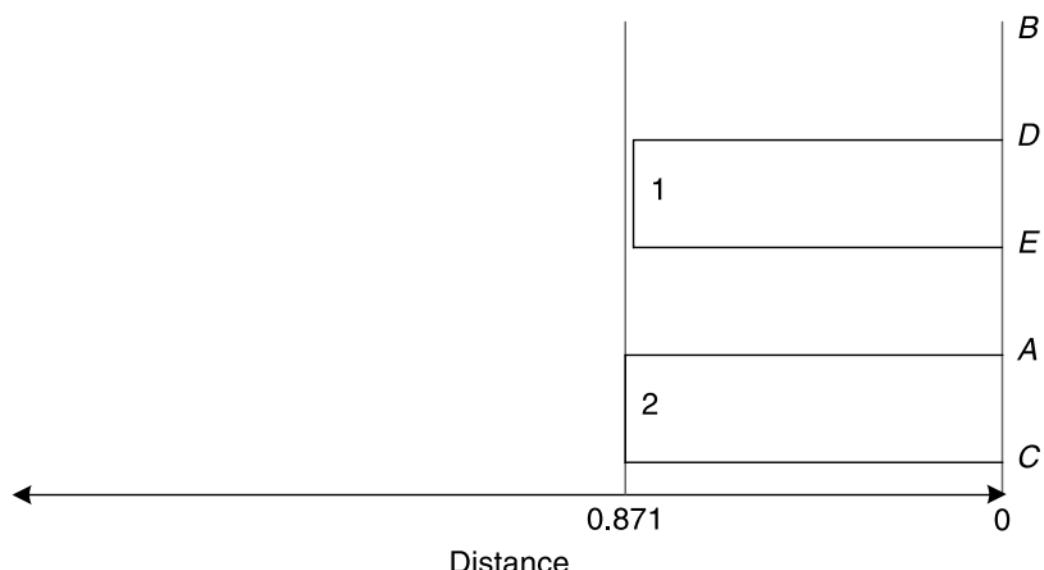
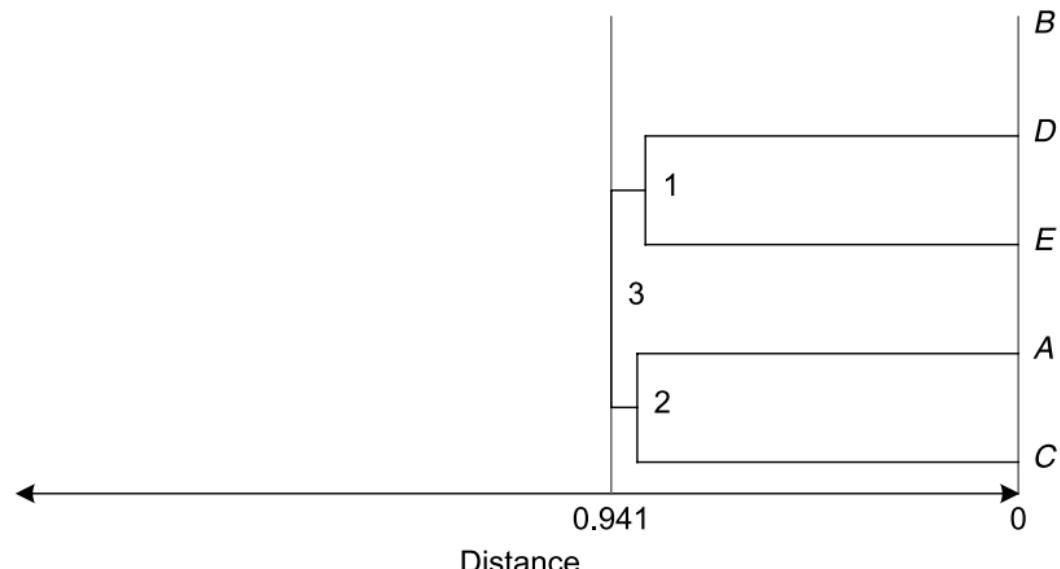


TABLE 3.18 Distance Matrix for the Three Groups Using the Single Linkage Criteria

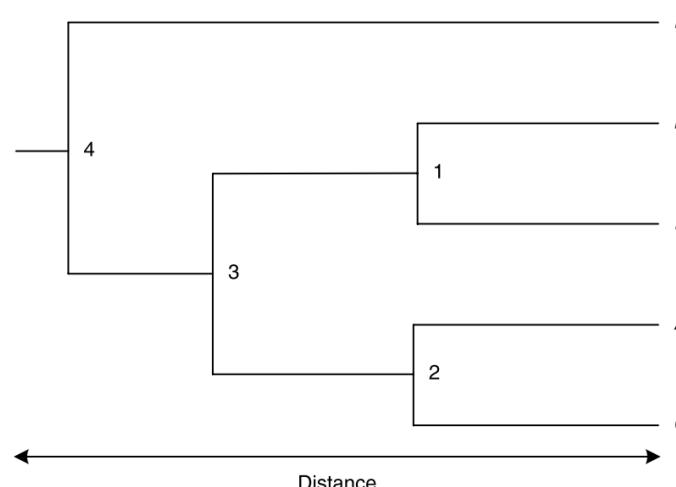
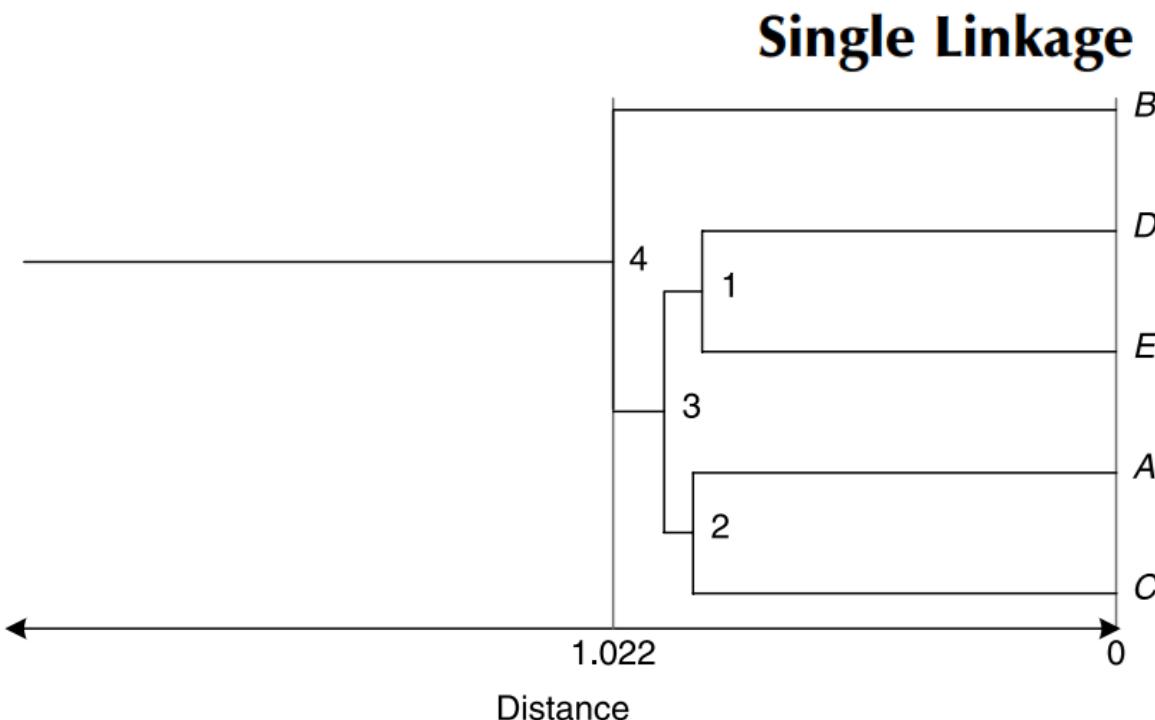
	{B}	{D,E}	{A,C}
{B}	0	1.022	1.185
{D,E}	1.022	0	0.941
{A,C}	1.185	0.941	0



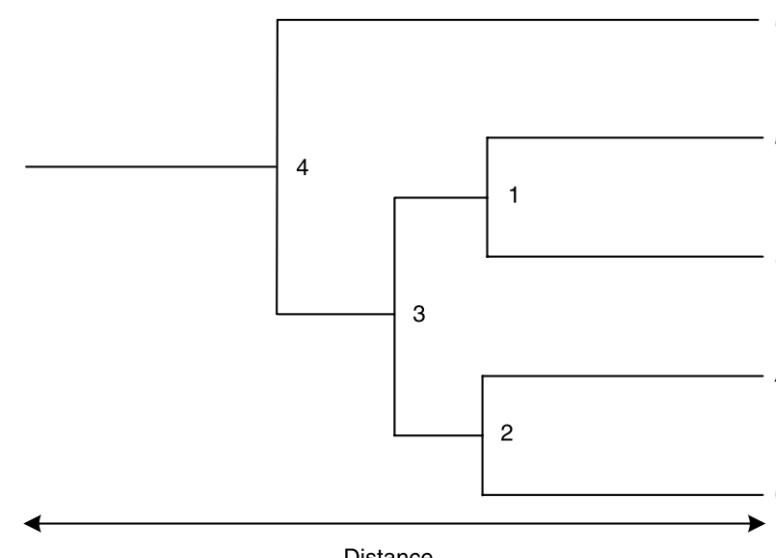
Example 2: Use HAC

TABLE 3.19 Distance Matrix for the Two Groups Using the Single Linkage Criteria

		{B}	{A,C,D,E}
{B}	0	1.022	
{A,C,D,E}	1.022	0	



complete linkage



average linkage



Partition Based Clustering

K-Means

k-Means Clustering

A non-hierarchical method for grouping a data set.

Groups data using a “top-down” approach

Starts with a “predefined number of clusters” and **assigns all** observations to **each of them**.

No overlaps in the groups

Each observation is assigned only to a single group.

Computationally faster and can handle greater numbers of observations

Limitations of *k*-Means Clustering

- **Number of groups (*k*)** must be specified before creating the clusters
 - This number is **not guaranteed** to provide the **best partitioning** of the observations.
- When a data set contains many **outliers**, k-means **may not create an optimal grouping**.
 - Sensitive to **noise** and **outliers** in the data
- **No hierarchical organization** is generated using k-means clustering and hence there is **no ordering** of the **individual observations**.
- Limited to use with **continuous variables** (calculate means)

k must be at least 2 and less than the total number of observations in the data set.

While collecting data, **humans** tend to make **mistakes** and **instruments** tend to be **inaccurate**, so collected data has some error bound to it. This **error** is referred to as **noise** in a dataset.

A photograph of a breakfast scene. In the foreground, a stack of folded newspapers sits on a dark wooden table. To the right, a white plate holds three golden-brown biscuits. The background is softly blurred, showing a window with a view of greenery and flowers.

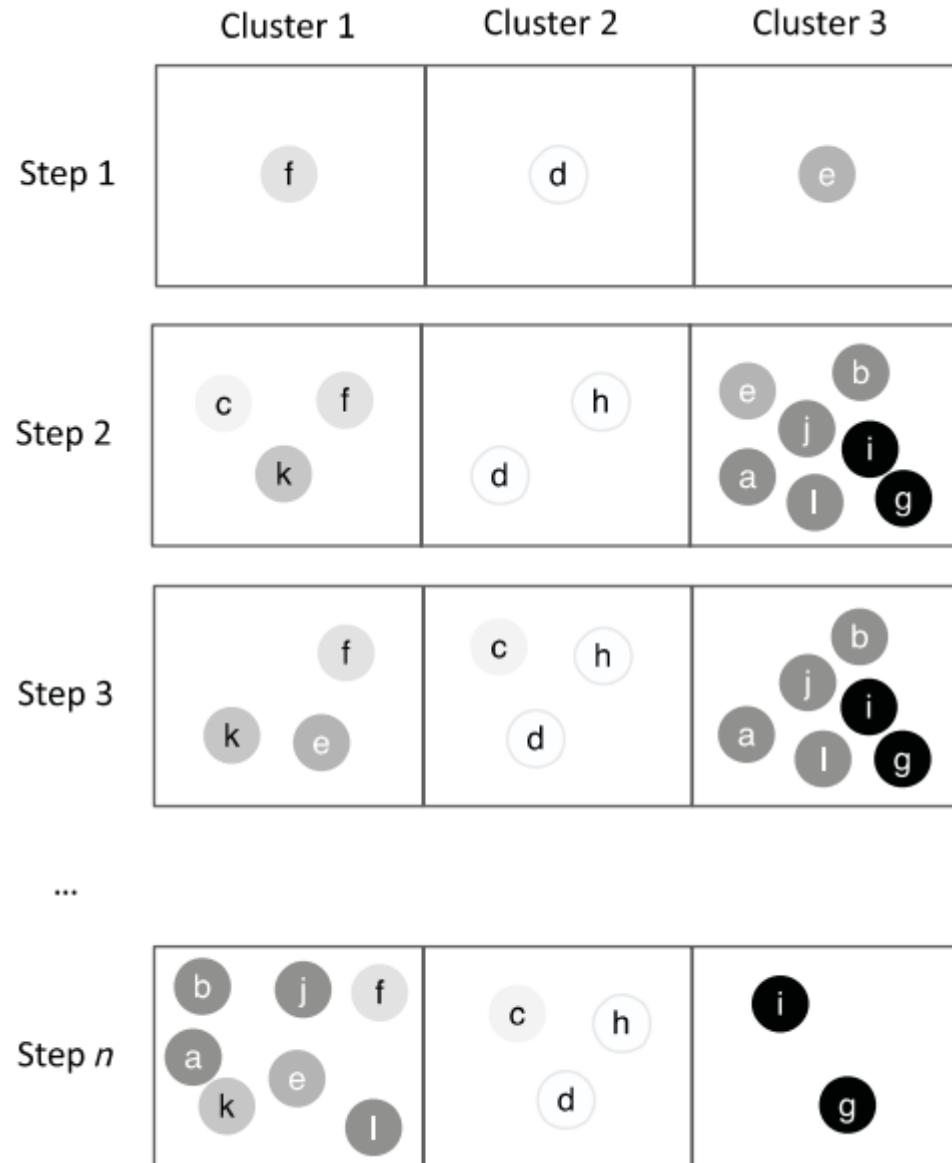
AFTER THE BREAK

Homework.

Given the following dataset, use HAC to create clusters. Depict the resultant dendrogram.

	X	y
O1	0.07	0.83
O2	0.85	0.14
O3	0.66	0.89
O4	0.49	0.64
O5	0.80	0.46

Illustrating conceptually k-means clustering.



- Defining value for **k** (number of groups to create)
- Allocates** an observation – usually selected randomly – to each of these **k** groups.
- All other observations are compared to each of the allocated observations and placed in the group to which they are most similar.
- Centre point** for each group is then calculated.
- Grouping process continues by determining the distance from all observations to these new group centres.
 - If an **observation is closer** to the **centre of another group**, it is **moved** to the group that it is closest to.
 - Centres** of its old and new groups are then **recalculated**.
- Process of **comparing** and **moving observations** where appropriate is **repeated** until **no observations** are **moved** after a **recalculation of the group's centre**.

12 hypothetical observations are used: a, b, c, d, e, f, g, h, i, j, k, and l.

***k*-means CLUSTERING**

Table 6.9. Table of observations to illustrate k-means clustering

Name	Variable 1	Variable 2	Variable 3	Variable 4	Variable
A	7	8	4	5	2
B	6	8	5	4	2
C	8	9	7	8	9
D	6	7	7	7	8
E	1	2	5	3	4
F	3	4	5	3	5
G	7	8	8	6	6
H	8	9	6	5	5
I	2	3	5	6	5
J	1	2	4	4	2
K	3	2	6	5	7
L	2	5	6	8	9
M	3	5	4	6	3
N	3	5	5	6	3

	Cluster 1	Cluster 2	Cluster 3
Step 1	D	K	M

Figure 6.22. Initial random assignment of three clusters

Choose k , here $k = 3$

Calculate the Euclidean distance of each observation with the initial cluster wise elements. Here, A with D, K, M and so on.

k-means **CLUSTERING**

Table 6.10. Observation distance to each cluster and cluster assignment

Name	Cluster 1 distance	Cluster 2 distance	Cluster 3 distance	Cluster assigned
A	7.1	9	5.2	3
B	7.1	8.5	4.9	3
C	3.2	9.4	9.5	1
E	9.3	4.2	4.9	2
F	6.9	3.6	3.9	2
G	2.8	7.6	7.1	1
H	4.7	8.8	7.1	1
I	6.8	2.8	3.2	2
J	10.2	5.8	4.2	3
L	4.8	4.8	6.7	1
N	6.6	5.2	1	3

	Cluster 1	Cluster 2	Cluster 3
Step 1	D	K	M

Figure 6.22. Initial random assignment of three clusters

Cluster 1		Cluster 2		Cluster 3	
C	G	E	F	A	B
D		K		M	
L	H	I		N	J

Figure 6.23. Initial assignment of all observations

k*- means CLUSTERING- *next step Find center

- **Center of each cluster** is calculated by taking the average value for each variable in the group
- For example, center of Cluster 1 is now:
 - **Variable 1: 6.2 Variable 2: 7.6 Variable 3: 6.8 Variable: 6.8, Variable 5:7.4**

Table 6.11. Calculation of cluster center

Name	Cluster 1				
	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5
C	8	9	7	8	9
D	6	7	7	7	8
G	7	8	8	6	6
H	8	9	6	5	5
L	2	5	6	8	9
Center (average)	6.2	7.6	6.8	6.8	7.4

k*- means CLUSTERING- *next step

- **Center of each cluster** is calculated by taking the average value for each variable in the group
- For example, center of Cluster 2 is now:
 - **Variable 1:2.25 Variable 2: 2.75 Variable 3: 5.25 Variable: 4.25 Variable 5:5.25**

Table 6.11. Calculation of cluster center

	Cluster 2				
Name	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5
E	1	2	5	3	4
F	3	4	5	3	5
I	2	3	5	6	5
K	3	2	6	5	7
Center (average)	2.25	2.75	5.25	4.25	5.25

k*- means CLUSTERING- *next step

- **Center of each cluster** is calculated by taking the average value for each variable in the group
- For example, center of Cluster 1 is now:
 - **Variable 1: 4 Variable 2: 5.6 Variable 3: 4.4 Variable: 5 Variable 5:2.4**

Table 6.11. Calculation of cluster center

	Cluster 3				
Name	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5
A	7	8	4	5	2
B	6	8	5	4	2
J	1	2	4	4	2
M	3	5	4	6	3
N	3	5	5	6	3
Center (average)	4	5.6	4.4	5	2.4

k*- means CLUSTERING- *next step

- Each **observation** is now compared to the **centers** of each cluster.
- For example, **A** is compared to the **center of Cluster 1, Cluster 2, and Cluster 3** using the Euclidean distance.
 - We have the following Euclidean distance:
 - From **A** to the center of **cluster 1**: 6.4
 - From **A** to the center of **cluster 2**: 7.9
 - From **A** to the center of **cluster 3**: 3.9
- Since **A is still closest to Cluster 3 it remains in Cluster 3**
- If an observation is moved, then **centers** for the **two clusters affected** are **recalculated**. Process of examining the observations and moving them as appropriate is **repeated until no further moves** are needed.

***k*-means CLUSTERING**

TABLE 5.8 Data Table Used in the *k*-Mean Clustering Example

Name	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5
A	7.9	8.6	4.4	5.0	2.5
B	6.8	8.2	5.2	4.2	2.2
C	8.7	9.6	7.5	8.9	9.8
D	6.1	7.3	7.9	7.3	8.3
E	1.5	2.0	5.1	3.6	4.2
F	3.7	4.3	5.4	3.3	5.8
G	7.2	8.5	8.6	6.7	6.1
H	8.5	9.7	6.3	5.2	5.0
I	2.0	3.4	5.8	6.1	5.6
J	1.3	2.6	4.2	4.5	2.1
K	3.4	2.9	6.5	5.9	7.4
L	2.3	5.3	6.2	8.3	9.9
M	3.8	5.5	4.6	6.7	3.3
N	3.2	5.9	5.2	6.2	3.7

Cluster 1 Cluster 2 Cluster 3

Step 1



FIGURE 5.20 Single observation randomly assigned to each cluster.

k-means CLUSTERING

TABLE 5.9 Euclidean Distances and Cluster Assignments

Name	Cluster 1	Cluster 2	Cluster 3	Cluster Assignment
A	1.178	1.106	1.2	2
B	1.064	0.025	1.147	2
C	1.468	0.709	0.582	3
E	0.542	1.518	1.406	1
F	0.57	1.191	1.092	1
H	1.218	0.648	0.808	2
J	0.659	1.624	1.543	1
K	0.346	1.033	0.797	1
L	0.727	1.108	0.744	1
M	0.553	1.148	1.065	1
N	0.458	1.051	0.974	1

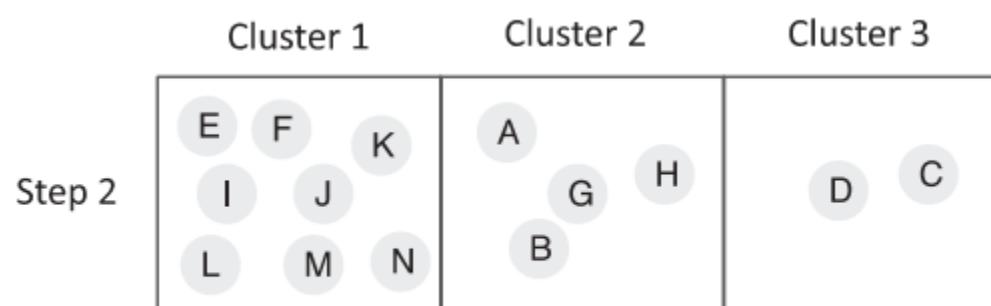


FIGURE 5.21 Initial assignment of other observations to each cluster.

k-means CLUSTERING

TABLE 3.14 Five Observations Measured Over Five Variables

	Percentage body fat	Weight	Height	Chest	Abdomen
A	12.3	154.25	67.75	93.1	85.2
B	31.6	217	70	113.3	111.2
C	22.2	177.75	68.5	102	95
D	14.1	176	73	96.7	86.5
E	23.6	197	73.25	103.6	99.8

- Example of a hierarchical method for grouping observations.
- Uses a “**bottom-up**” approach to clustering as it starts with each observation and progressively creates cluster by merging observations together until all are a member of a final single cluster

Another Example.

TABLE 3.28 Data Set of Wines

Malic acid	Ash	Alkalinity of ash	Magnesium	Total phenols	Nonflavanoids
14.23	1.71	2.43	15.6	127	3.06
13.2	1.78	2.14	11.2	100	2.76
13.16	2.36	2.67	18.6	101	3.24
14.37	1.95	2.5	16.8	113	3.49
13.24	2.59	2.87	21	118	2.69
14.2	1.76	2.45	15.2	112	3.39
14.39	1.87	2.45	14.6	96	2.52
14.06	2.15	2.61	17.6	121	2.51
14.83	1.64	2.17	14	97	2.98

Another Example.

Centroid values							
Cluster ID	Malic acid	Ash	Alkalinity of ash	Magnesium	Total phenols	Nonflavanoids	Count
Cluster 1	13.1	3.62	2.44	21.5	97.7	1.06	54
Cluster 2	12.2	1.7	2.24	20	94.8	1.98	64
Cluster 3	13.7	1.86	2.44	17.1	107	2.95	60

Figure 3.20 Summary of three clusters generated from k means clustering

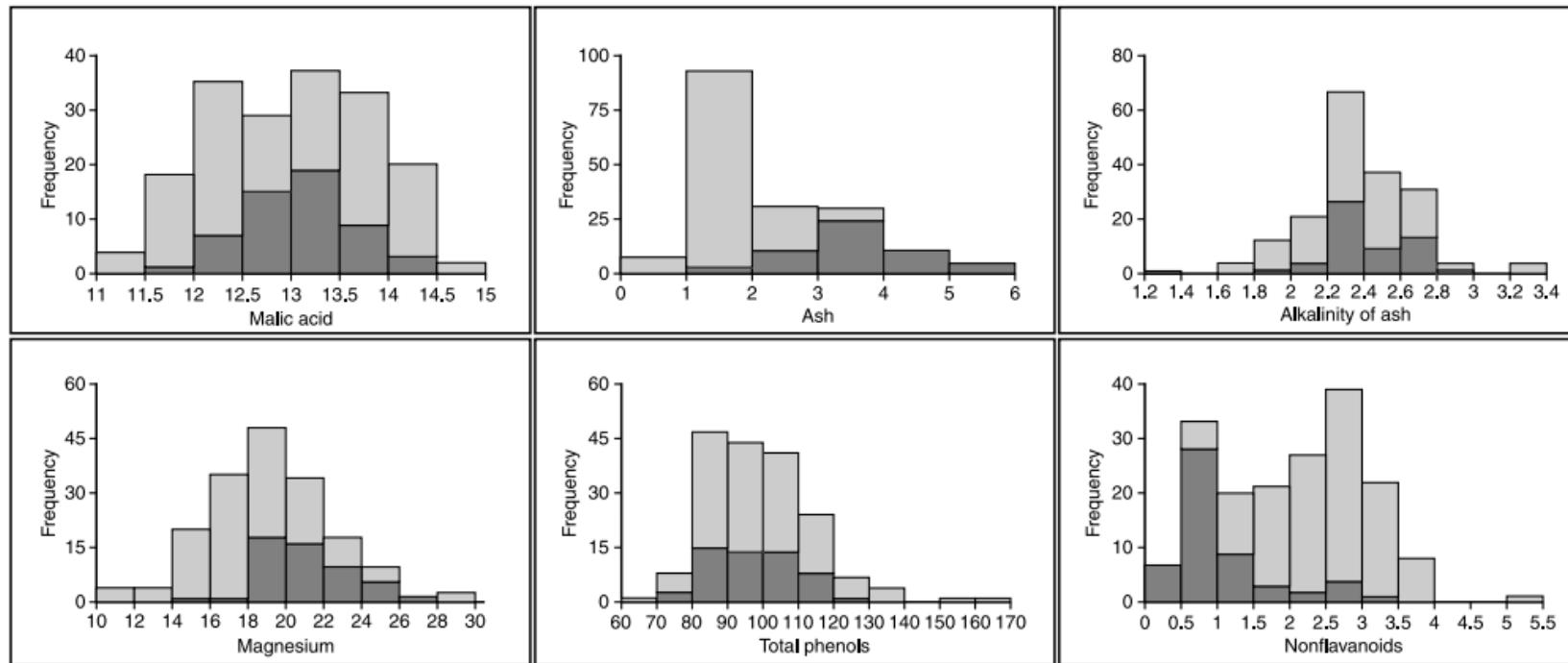


Figure 3.21 Summary of cluster 1 from the k means clustering

Source: <http://archive.ics.uci.edu/ml/datasets/Wine>

Miscellaneous Partitioned-Based Clustering.

- Two alternative **partitioned-based clustering** methods are:
 - **k-modes:** Categorical data, and it operates similarly to k-means. Where k-means uses a mean function to compute the cluster centroid, k-modes uses a mode function for this computation.
 - **k-medoids:** Does not use a mean centroid, but it instead uses a representative observation as the cluster centroid. This approach helps overcome issues with outliers that can skew results generated with the k-means method.

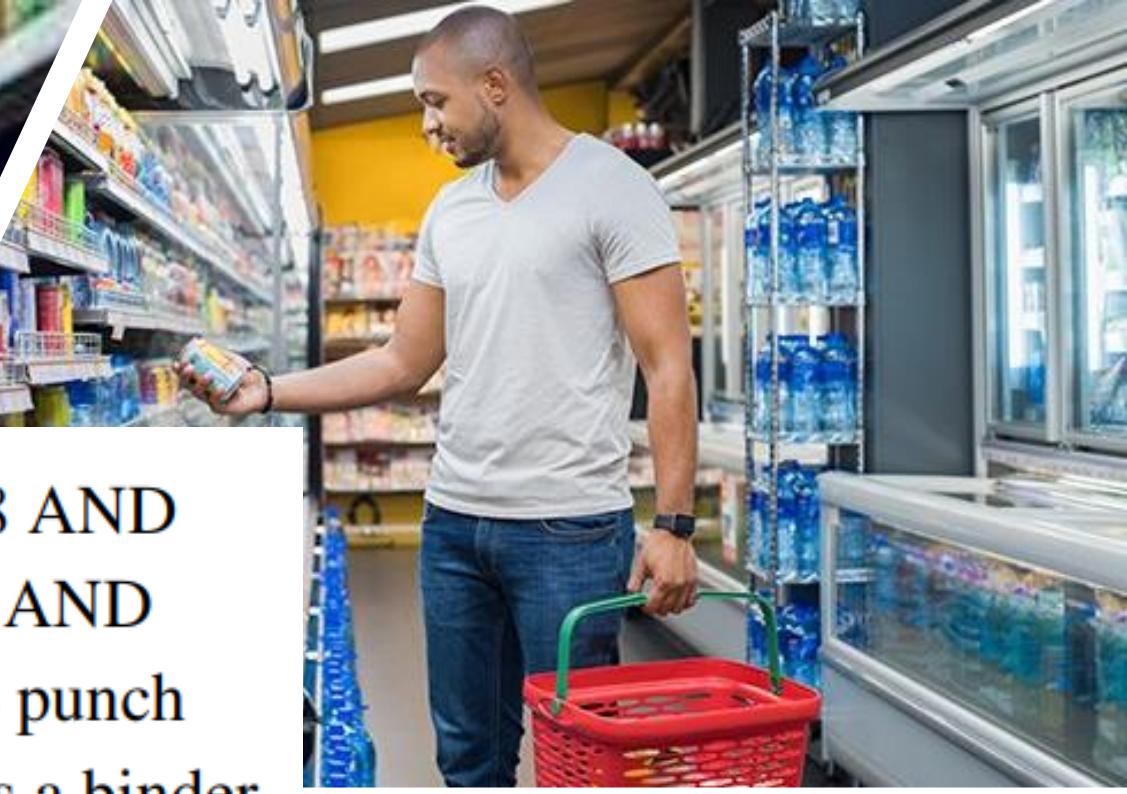
MBA.



What is Market Basket Analysis?

- Patterns
- Frequency Patterns
 - Appear frequently in a data set
- Frequent Item set
- First few applications of Data analytics & Machine Learning





IF the customer is age 18 AND
the customer buys paper AND
the customer buys a hole punch
THEN the customer buys a binder



Market Basket Analysis

- Used by retailers to **increase sales** by better understanding customer purchasing patterns.
- Analyzes customer buying habits by finding associations between different items that customers place in their shopping baskets
- Involves analyzing large data sets, such as purchase history, to reveal product groupings, as well as products that are likely to be purchased together.
- Leads to the discovery of associations and correlations among items in large transactional or relational data sets.

Market Basket Analysis

- “Which groups or sets of items are customers likely to purchase on a given trip to the store ?”
- Helps in business decision making process, such as:
 - Catalog design,
 - Cross-marketing
 - Customer shopping behavior analysis.
- Can be used to plan different store layouts.
 - Place them together
 - Place them far apart
 - Decide which item to put on sale.
- Patterns can be represented in the form of Association rules

‘Basket data’

IF the customer’s age is 18 AND
the customer buys paper AND
the customer buys a hole punch
THEN the customer buys a binder



Associative Rules. (advantages)

- **Easy to interpret:** The results are presented in the form of a rule that is easily understood.
- **Actionable:** It is possible to perform some sort of action based on the rule.
 - For example, the rule in the previous example allowed the retailer to market this combination of items differently.
- **Large data sets:** It is possible to use this technique with large numbers of observations

Association Rule Mining.

- Proposed by **Agrawal et al in 1993**.
- It is an important data mining model studied extensively by the database and data mining community.
- Assume all data are categorical.
- No good algorithm for numeric data.
- Initially used for **Market Basket Analysis** to find how items purchased by customers are related.

{Bread... } \Rightarrow {Milk} [sup = 5%, conf = 100%]

- Used to predict the **likelihood of products** being purchased together.
- Count the frequency of items that occur together, seeking to find associations that occur far more often than expected.

The model:data

- $I = \{i_1, i_2, \dots, i_m\}$: a set of items.
- Transaction t :
 - t a set of items, and $t \subseteq I$.
- Transaction Database T : a set of transactions $T = \{t_1, t_2, \dots, t_n\}$.

Transaction data

- **Market basket transactions:**

supermarket
data

$t_1: \{\text{bread, cheese, milk}\}$
 $t_2: \{\text{apple, eggs, salt, yogurt}\}$
... ...
 $t_n: \{\text{biscuit, eggs, milk}\}$

- **An item:** an item/article in a basket
- **I:** Set of all items sold in the store
- **Transaction:** Items purchased in a basket (may have TID (transaction ID))
- **Transactional dataset:** A set of transactions

Transaction data

- A **text document data set**.
- Each document is treated as a “**bag**” of keywords
 - doc1: Student, Teach, School
 - doc2: Student, School
 - doc3: Teach, School, City, Game
 - doc4: Baseball, Basketball
 - doc5: Basketball, Player, Spectator
 - doc6: Baseball, Coach, Game, Team
 - doc7: Basketball, Team, City, Gam

Set of documents

The rules.

- A transaction t contains X , a set of items (itemset) in I , if $X \subseteq t$.
- An **association rule** is an implication of the form:

$$X \Rightarrow Y, \text{ where } X, Y \subset I, \text{ and } X \cap Y = \emptyset$$

Antecedent Consequent

- An **itemset** is a **set of items**.
E.g., **X = {milk, bread, cereal}** is an itemset.
- A k -itemset is an itemset with k items.
E.g., **{milk, bread, cereal}** is a 3-itemset

Example.



Jeans	Shirt	Jacket	Shoes	Transactions
1	1	0	1	Jeans, Shirt ,Schuhe
0	0	0	1	Schuhe
0	1	0	1	Shirt, Schuhe
...
1	1	0	1	Jeans ,Shirt ,Schuhe
0	1	0	1	Shirt ,Schuhe

Example.



Lhs	Rhs	Frequency	Support	Confidence	Lift
Shirt	Shoes	8	0.42	0.8	1.27
Jeans, Shirt	Shoes	3	0.16	0.75	1.19
Shoes	Shirt	8	0.42	0.67	1.27
Jeans, Shoes	Shirt	3	0.16	0.6	1.14

Jeans	Shirt	Jacket	Shoes
1	1	0	1
0	0	0	1
0	1	0	1

Lhs	Rhs	Frequency	Support	Confidence	Share
Shirt	Shoes	8	0.42	0.8	1.27
Jeans, Shirt	Shoes	3	0.16	0.75	1.19
Shoes	Shirt	8	0.42	0.67	1.27
Jeans, Shoes	Shirt	3	0.16	0.6	1.14

0	1	0	1
0	1	0	1
0	0	1	1
0	0	1	0
0	1	1	0
1	1	1	0
1	0	0	1
1	0	0	1
1	1	0	1
1	1	0	1
0	1	0	1

Lhs	Rhs	Frequency	Support	Confidence	Lift
Shirt	Shoes	8	0.42	0.8	1.27

Rule

$Lhs \rightarrow Rhs$

$Shirt \rightarrow Shoes$

Frequency

$frq(Lhs, Rhs)$

$frq(Shirt, Shoes) = 8$

Support

$frq(Lhs, Rhs) / N$

$8 / 19 = 0.42$

Confidence

$frq(Lhs, Rhs) / frq(Lhs)$

$8 / 10 = 0.8$

Lift

$\frac{Support}{Support(Lhs) \times Support(Rhs)}$

$\frac{Support}{Support(Shirt) \times Support(Shoes)}$

$\frac{8 / 19}{10 / 19 \times 12 / 19} = 1.27$

The Apriori algorithm

- Probably the **best known algorithm**
- **Two steps:**
 1. Find all item sets that have minimum support (frequent item sets, also called large item sets).
 2. Use frequent item sets to **generate rules**.

Apriori Property: All non-empty subset of frequent itemset must be frequent.

If an itemset is infrequent, all its supersets will be infrequent.

The Apriori algorithm

- **Join Step:** This step generates $(K+1)$ itemset from K -itemsets by joining each item with itself.
- **Prune Step:** This step scans the count of each item in the database. If the candidate item does not meet minimum support, then it is regarded as infrequent and thus it is removed. This step is performed to reduce the size of the candidate itemsets.

Example 1: Apriori algorithm

Dataset T

TID	Items
T100	1, 3, 4
T200	2, 3, 5
T300	1, 2, 3, 5
T400	2, 5

Iteration 1 → $C_1: \{1\}:2, \{2\}:3, \{3\}:3, \{4\}:1, \{5\}:3$

→ $L_1: \{1\}:2, \{2\}:3, \{3\}:3, \{5\}:3$

→ $C_2: \{1,2\}, \{1,3\}, \{1,5\}, \{2,3\}, \{2,5\}, \{3,5\}$

Iteration 2 → $C_2: \{1,2\}:1, \{1,3\}:2, \{1,5\}:1, \{2,3\}:2,$
 $\{2,5\}:3, \{3,5\}:2$

→ $L_2: \{1,3\}:2, \{2,3\}:2, \{2,5\}:3, \{3,5\}:2$

→ $C_3: \{2, 3, 5\}$

Iteration 3 → $C_3: \{2, 3, 5\}:2 \rightarrow L_3: \{2, 3, 5\}$

A photograph of a breakfast setup on a wooden table. In the foreground, there's a white saucer and a white cup. To the left, a stack of folded newspapers sits next to a white mug. On the right, a white plate holds three golden-brown biscuits. The background is blurred, showing foliage and sunlight.

BREAK

Example 1: Apriori algorithm

Dataset T

Itemset	sup_count
I1,I2	4
I1,I3	4
I1,I4	1
I1,I5	2
I2,I3	4
I2,I4	2
I2,I5	2
I3,I4	0
I3,I5	1
I4,I5	0

Itemset	sup_count
I1	6
I2	7
I3	6
I4	2
I5	2

Itemset	sup_count
I1	6
I2	7
I3	6
I4	2
I5	2

Itemset	sup_count
I1,I2	4
I1,I3	4
I1,I5	2
I2,I3	4
I2,I4	2
I2,I5	2
I2,I5	2

Itemset	sup_count
I1,I2,I3	2
I1,I2,I5	2

Itemset	sup_count
I1,I2,I3	2
I1,I2,I5	2

minimum support count is 2
minimum confidence is 60%

TID	Items
T100	1, 2, 5
T200	2, 4
T300	2, 3
T400	1, 2, 4
T500	1, 3
T600	2, 3
T700	1, 3
T800	1, 2, 3, 5
T900	1, 2, 3

Example 1: Apriori algorithm

Dataset T

$L = \{\{I_1\}, \{I_2\}, \{I_3\}, \{I_4\}, \{I_5\}, \{I_1, I_2\}, \{I_1, I_3\}, \{I_1, I_5\}, \{I_2, I_3\}, \{I_2, I_4\}, \{I_2, I_5\}, \{I_1, I_2, I_3\}, \{I_1, I_2, I_5\}\}.$

Lets take $I = \{I_1, I_2, I_5\}$.

Its all non-empty subsets are $\{I_1, I_2\}, \{I_1, I_5\}, \{I_2, I_5\}, \{I_1\}, \{I_2\}, \{I_5\}$.

minimum support count is 2

minimum confidence is 60%

TID	Items
T100	1, 2, 5
T200	2, 4
T300	2, 3
T400	1, 2, 4
T500	1, 3
T600	2, 3
T700	1, 3
T800	1, 2, 3, 5
T900	1, 2, 3



Found **frequent item sets**, next what?

Generate STRONG association rules from the frequent item sets

Rules must satisfy minimum support and minimum confidence

$$\text{Confidence}(A \rightarrow B) = \text{Support_count}(A \cup B) / \text{Support_count}(A)$$

Rule generation.

$[I_1 \wedge I_2] \Rightarrow [I_3]$ //confidence = $\text{sup}(I_1 \wedge I_2 \wedge I_3) / \text{sup}(I_1 \wedge I_2) = 2/4 * 100 = 50\%$

$[I_1 \wedge I_3] \Rightarrow [I_2]$ //confidence = $\text{sup}(I_1 \wedge I_2 \wedge I_3) / \text{sup}(I_1 \wedge I_3) = 2/4 * 100 = 50\%$

$[I_2 \wedge I_3] \Rightarrow [I_1]$ //confidence = $\text{sup}(I_1 \wedge I_2 \wedge I_3) / \text{sup}(I_2 \wedge I_3) = 2/4 * 100 = 50\%$

$[I_1] \Rightarrow [I_2 \wedge I_3]$ //confidence = $\text{sup}(I_1 \wedge I_2 \wedge I_3) / \text{sup}(I_1) = 2/6 * 100 = 33\%$

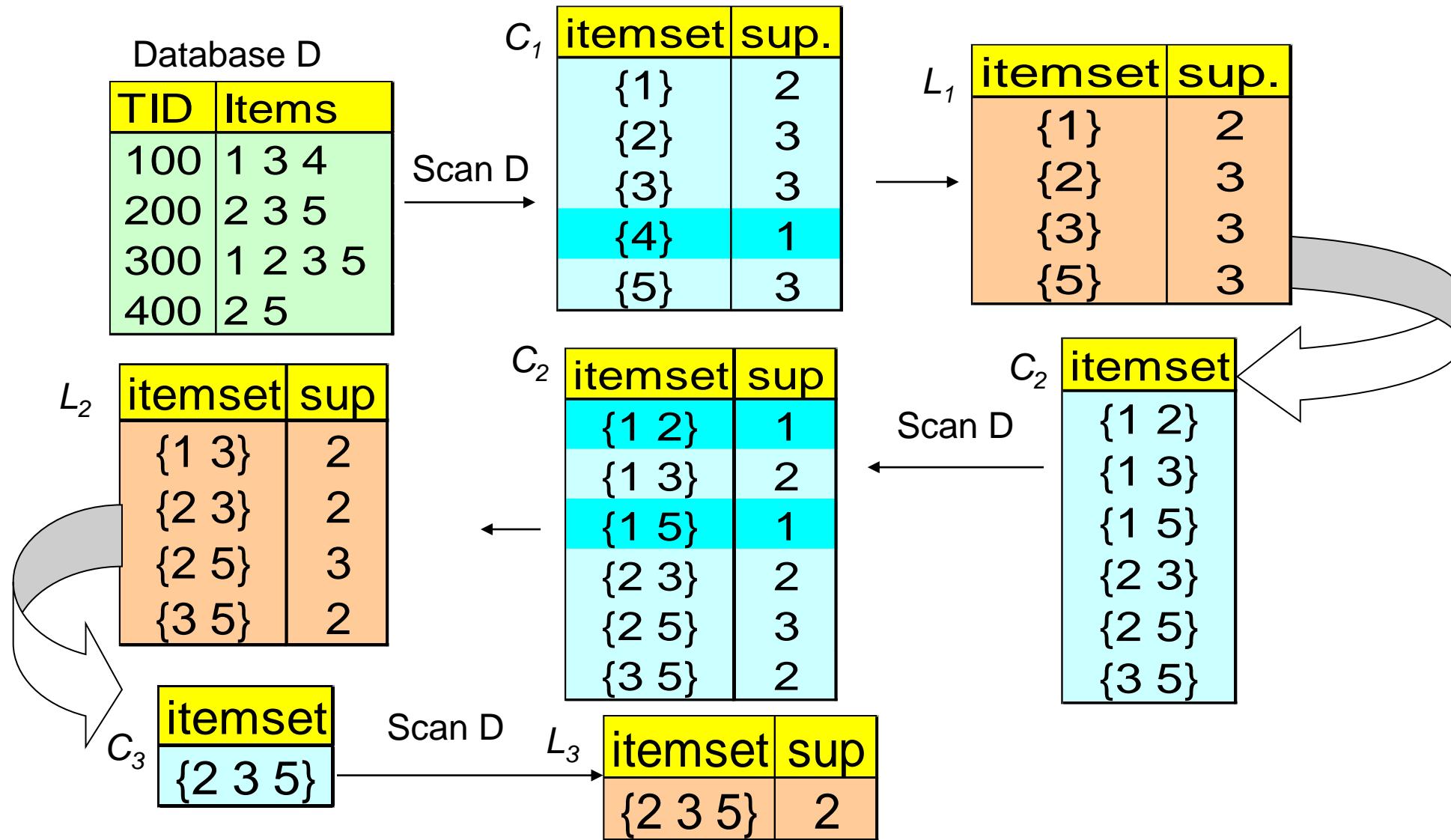
$[I_2] \Rightarrow [I_1 \wedge I_3]$ //confidence = $\text{sup}(I_1 \wedge I_2 \wedge I_3) / \text{sup}(I_2) = 2/7 * 100 = 28\%$

$[I_3] \Rightarrow [I_1 \wedge I_2]$ //confidence = $\text{sup}(I_1 \wedge I_2 \wedge I_3) / \text{sup}(I_3) = 2/6 * 100 = 33\%$

If minimum confidence is 50%, then first 3 rules can be considered as strong association rules.

How do we find rules for minimum confidence is 60% ?

The Apriori Algorithm – Example 2

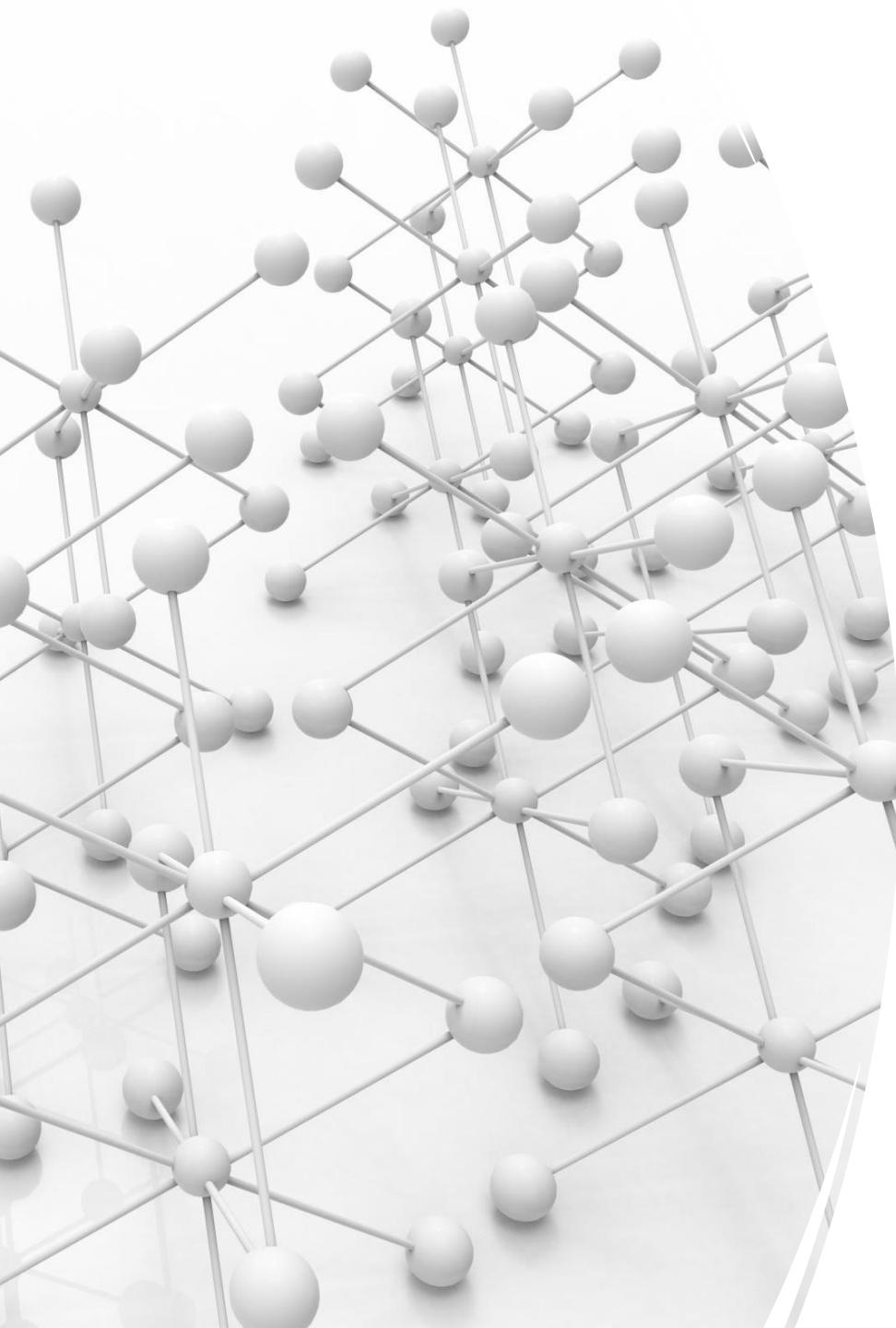


Improving Apriori's Efficiency

- **Hash-based itemset counting:** A k-itemset whose corresponding hashing bucket count is below the threshold cannot be frequent
- **Transaction reduction:** A transaction that does not contain any frequent k-itemset is useless in subsequent scans
- **Partitioning:** Any itemset that is potentially frequent in DB must be frequent in at least one of the partitions of DB
- **Sampling:** mining on a subset of given data, need a lower support threshold + a method to determine the completeness
- **Dynamic itemset counting:** add new candidate itemsets immediately (unlike Apriori) when all of their subsets are estimated to be frequent

A scenic landscape featuring a row of wind turbines on a hillside overlooking a body of water under a clear sky.

FP Growth Algorithm



FP Growth Algorithm: Association Rule Mining

- FP stands for Frequent Pattern Algorithm
- Discovers frequent item sets
- Encodes data set using a data structure called **FP Tree**
- **Extracts** frequent item sets directly from FP tree

Frequency Growth Algorithm.

Dataset T

Step 1: Count occurrences of individual items

Step 2: Filter out non-frequent items using minimum support

Step 3: Order the item sets based on individual occurrences

Step 4: Create the tree and add the transactions one by one

Minimum support: 2

TID	Items
T100	1, 2, 5
T200	2, 4
T300	2, 3
T400	1, 2, 4
T500	1, 3
T600	2, 3
T700	1, 3
T800	1, 2, 3, 5
T900	1, 2, 3

FP Tree Construction.

- Create root node labelled NULL
- Scan T, second time, items are processed in L order (sorted)
- Branch is created for each transaction with items having their support count
- Whenever same node created in another transaction, increment support count
- Use Item header table for ease of tree traversal

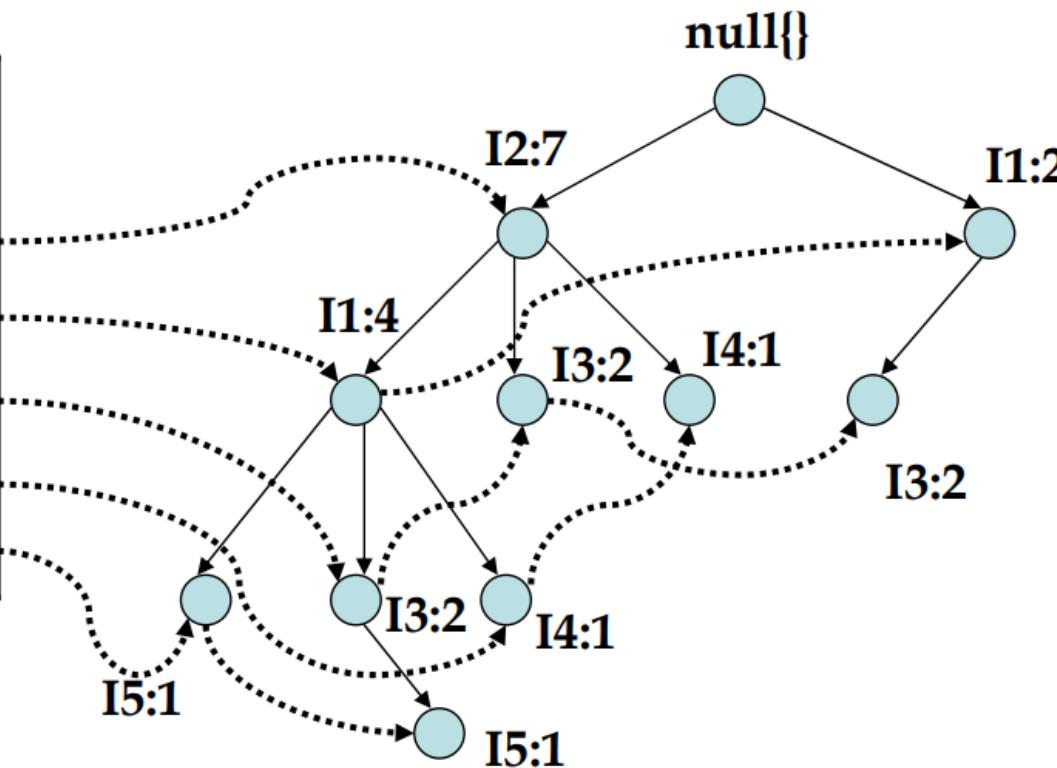
Item header table

Item Id	Sup Count
I2	7
I1	6
I3	6
I4	2
I5	2

Frequency Growth Tree.

Dataset T

Item Id	Sup Count	Node-link
I2	7	
I1	6	
I3	6	
I4	2	
I5	2	



An FP-Tree that registers compressed, frequent pattern information

TID	Items
T100	1, 2, 5
T200	2, 4
T300	2, 3
T400	1, 2, 4
T500	1, 3
T600	2, 3
T700	1, 3
T800	1, 2, 3, 5
T900	1, 2, 3

Mining FP tree by creating conditional pattern bases.

1. Start from each frequent length:1 pattern
2. Construct its conditional pattern base which consists of set of prefix paths in FP Tree co-occurring with suffix pattern
3. Construct its conditional FP tree and perform mining on the tree
4. Pattern growth is achieved by concatenation of suffix pattern with frequent patterns generated from a conditional FP tree
5. Union of all frequent patterns (generated by step 4) gives required frequent itemset.

Frequency Growth Tree.

Item	Conditional pattern base	Conditional FP-Tree	Frequent pattern generated
I5	$\{(I2 \mid I1: 1), (I2 \mid I1 \mid I3: 1)\}$	$\langle I2: 2, I1: 2 \rangle$	$I2 \mid I5: 2, I1 \mid I5: 2, I2 \mid I1 \mid I5: 2$
I4	$\{(I2 \mid I1: 1), (I2: 1)\}$	$\langle I2: 2 \rangle$	$I2 \mid I4: 2$
I3	$\{(I2 \mid I1: 1), (I2: 2), (I1: 2)\}$	$\langle I2: 4, I1: 2 \rangle, \langle I1: 2 \rangle$	$I2 \mid I3: 4, I1, I3: 2, I2 \mid I1 \mid I3: 2$
I2	$\{(I2: 4)\}$	$\langle I2: 4 \rangle$	$I2 \mid I1: 4$

Minimum support: 2

Generating strong Association rules.

- Same as explained with Apriori algorithm.

Find frequent item set & FP tree construction

Transaction ID	Items
1	{b, a}
2	{b, c, d}
3	{c, d, e, a}
4	{a, d, e}
5	{c, b, a}
6	{a, b, c, d}
7	{a}
8	{a, b, c}
9	{d, b, a}
10	{b, c, e}

Grouping by Combinations of Values.

- Consider a simple situation concerning a **shop that sells only cameras and televisions.**
- A data set of **31,612** sales transactions is used which contains **three variables: Customer ID, Gender, and Purchase.**

TABLE 5.12 Grouping by Different Value Combinations

Group Number	Count	Gender	Purchase
Group 1	16,099	Male	Camera or Television
Group 2	15,513	Female	Camera or Television
Group 3	16,106	Male or Female	Camera
Group 4	15,506	Male or Female	Television
Group 5	7,889	Male	Camera
Group 6	8,210	Male	Television
Group 7	8,217	Female	Camera
Group 8	7,296	Female	Television

Table 5.11 shows three rows from this table. Each row of the table contains an entry where values are assigned for each variable (i.e., **there are no missing values**).

Grouping of this set of 31,612 observations is done based on specific values for the variables Gender and Purchase.

- There are **eight ways** of grouping based on the values for the different categories of Gender and Purchase.
- For example, Group 5 has 7,889 observations where Gender is “male” and Purchase is “camera.”

Grouping by Combinations of Values.

If an **additional variable** is added to this data set, the number of possible groups will increase to 26 as shown in Table 5.13.

For example, if another variable **Income** which has two values – **above \$50K and below \$50K** is added to the table.

Increasing number of variables or **number of possible values for each variable** or both **increases number of groups**.

When the number of groups becomes too large, it becomes impractical to generate all combinations.

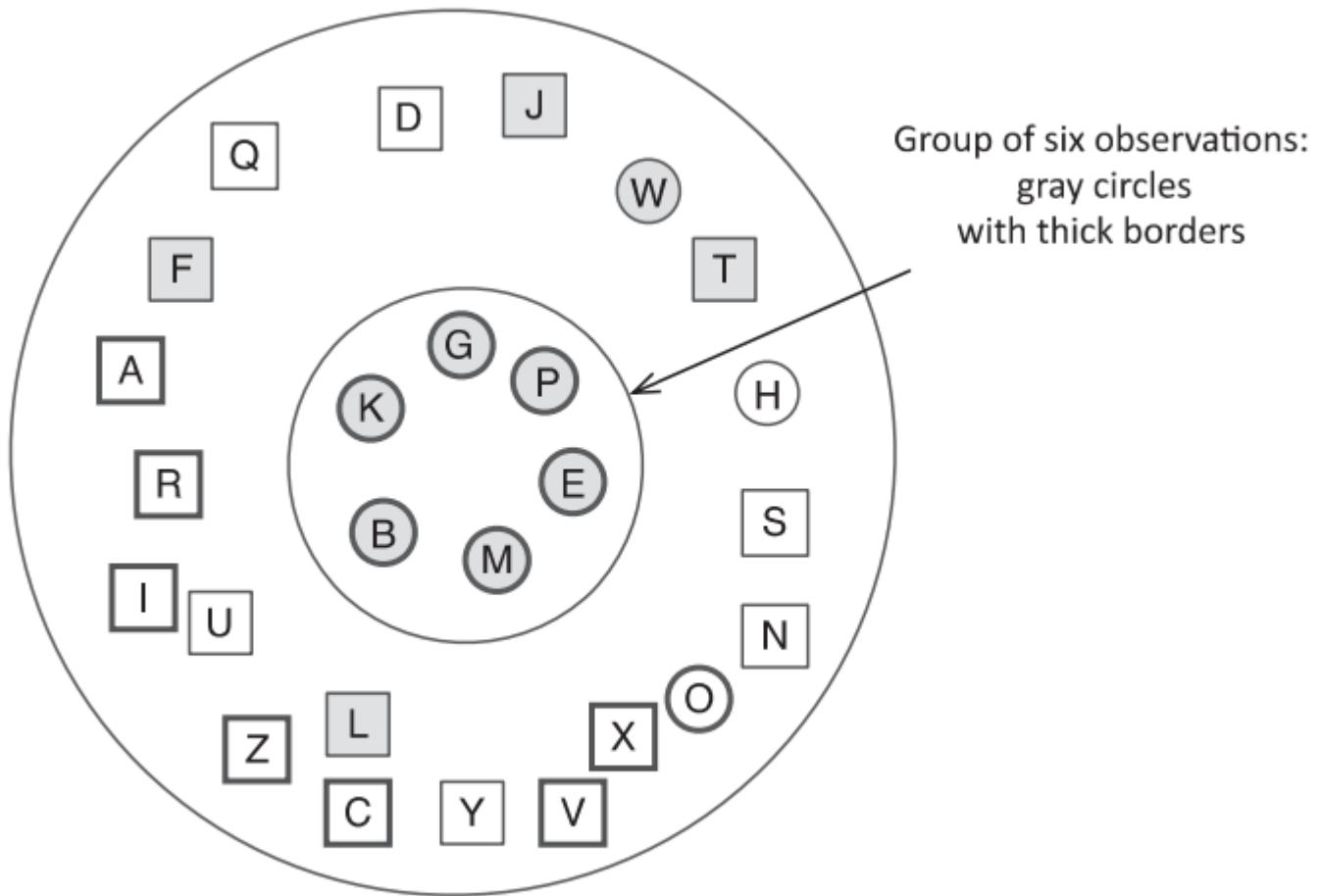
Techniques for generating the groups take advantage of this by requiring that groups **reach a certain size** before they are generated.

Care to be taken in **setting this cut-off value** since **rules** can only be created from generated groups.

TABLE 5.13 Table Showing Groups by Different Value Combinations

Group Number	Count	Gender	Purchase	Income
Group 1	16,099	Male	Camera or Television	Below \$50K or Above \$50K
Group 2	15,513	Female	Camera or Television	Below \$50K or Above \$50K
Group 3	16,106	Male or Female	Camera	Below \$50K or Above \$50K
Group 4	15,506	Male or Female	Television	Below \$50K or Above \$50K
Group 5	15,854	Male or Female	Camera or Television	Below \$50K
Group 6	15,758	Male or Female	Camera or Television	Above \$50K
Group 7	7,889	Male	Camera	Below \$50K or Above \$50K
Group 8	8,210	Male	Television	Below \$50K or Above \$50K
Subject matter knowledge and information generated from the data characterization phase will help in setting the cut-off value.				
Group 13	7,305	Female	Camera or Television	Below \$50K
Group 14	8,208	Female	Camera or Television	Above \$50K
Group 15	8,534	Male or Female	Camera	Below \$50K
Group 16	7,572	Male or Female	Camera	Above \$50K
Group 17	7,320	Male or Female	Television	Below \$50K
Group 18	8,186	Male or Female	Television	Above \$50K
Group 19	4,371	Male	Camera	Below \$50K
Group 20	3,518	Male	Camera	Above \$50K
Group 21	4,178	Male	Television	Below \$50K
Group 22	4,032	Male	Television	Above \$50K
Group 23	4,163	Female	Camera	Below \$50K
Group 24	4,054	Female	Camera	Above \$50K
Group 25	3,142	Female	Television	Below \$50K
Group 26	4,154	Female	Television	Above \$50K

Extracting and Assessing Rules.



Rule 1

IF *Color* = gray AND
Shape = circle
THEN *Border* = thick

Rule 2

IF *Border* = thick AND
Color = gray
THEN *Shape* = circle

Rule 3

IF *Border* = thick AND
Shape = circle
THEN *Color* = gray

FIGURE 5.26 Group of six observations (gray circles with thick borders).

Extracting and Assessing Rules

Rule 1

*IF Color = gray AND
Shape = circle*

THEN Border = thick

- **Support:** The support is a measure of the number of observations a rule maps onto. Its value is the proportion of the observations a rule selects out of all observations in the data set.
- **Confidence:** Each rule is divided into two parts: **antecedent** and **consequence**.
 - The confidence score is a measure for how predictable a rule is.
 - The confidence or predictability value is calculated using the support for the entire group divided by the support for all observations
 - **Confidence = group support / IF-part support**
 - Confidence values range from **no confidence (0)** to **high confidence (1)**.

Extracting and Assessing Rules

Rule 1

*IF Color = gray AND
Shape = circle*

THEN Border = thick

- Lift:

- Confidence value does not indicate the strength of the association between gray circles (IF-part) and thick borders (THEN-part).
- Lift score takes this into account.
- Lift is often described as the importance of the rule as it describes the association between the IF-part of the rule and the THEN-part of the rule.
- Calculated by dividing the confidence value by the support value across all observations of the THEN-part:
 - Lift = confidence / THEN-part support

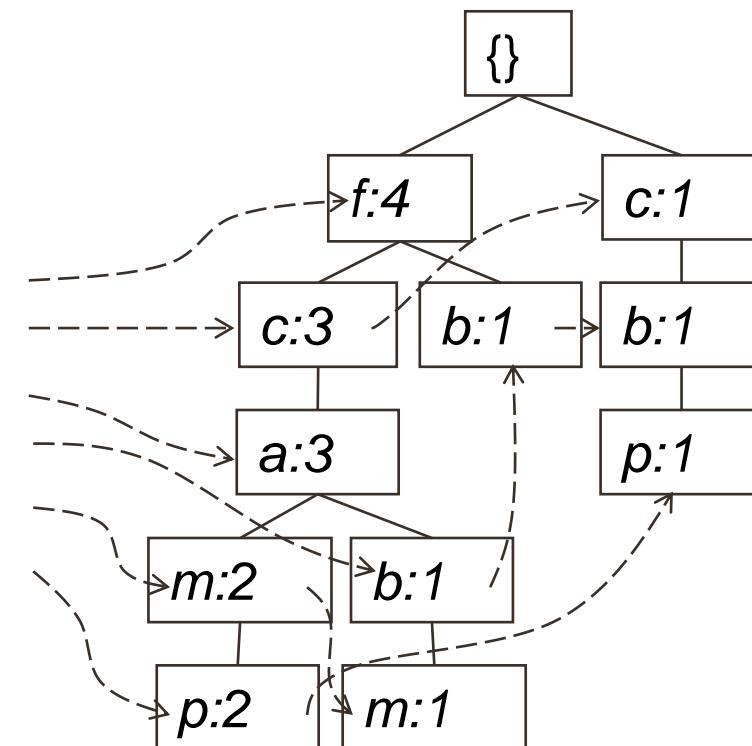
TAKE A
BREAK.

Construct FP-tree from a Transaction DB

TID	Items bought	(ordered) frequent items
100	{f, a, c, d, g, i, m, p}	{f, c, a, m, p}
200	{a, b, c, f, l, m, o}	{f, c, a, b, m}
300	{b, f, h, j, o}	{f, b}
400	{b, c, k, s, p}	{c, b, p}
500	{a, f, c, e, l, p, m, n}	{f, c, a, m, p}

min_support = 0.5

Header Table	
Item frequency head	
f	4
c	4
a	3
b	3
m	3
p	3



Benefits of the FP-tree Structure

- **Completeness:**

- never breaks a long pattern of any transaction
- preserves complete information for frequent pattern mining

- **Compactness**

- reduce irrelevant information—infrequent items are gone
- frequency descending ordering: more frequent items are more likely to be shared
- never be larger than the original database (if not count node-links and counts)

ID apples, beer, cheese, dates, eggs, fish, glue, honey, ice-cream

1	1	1		1			1	1	
2			1	1	1				
3		1	1			1			
4		1				1			1
5					1		1		
6						1			1
7	1			1				1	
8						1			1
9			1		1				
10		1					1		
11					1		1		
12	1								
13				1			1		
14				1			1		
15								1	1
16				1					
17	1					1			
18	1	1	1	1				1	
19	1	1		1			1	1	
20					1				