[4.1-PART-01; 4.1-PART-02]

1. **Cosine similarity: measure distance between two records**
2. **Jaccard distance**
3. **Distance Measures For Attributes**
4. **Measuring Distance Between Two Clusters- Hierarchical Clustering**

Compiled By: Dr. Nilamadhab Mishra [(PhD- CSIE) Taiwan]

4.1-PART-01

Cluster analysis

What is Cluster Analysis?

- **Cluster: A collection of data objects**
  - similar (or related) to one another within the same group.
  - dissimilar (or unrelated) to the objects in other groups.
- **Cluster analysis (or clustering, data segmentation, …)**
  - Finding similarities between data according to the characteristics found in the data and grouping similar data objects into clusters.

Cont..

- Unsupervised learning: no predefined classes (i.e., *learning by observations* vs. learning by examples: supervised)
- Typical applications
  - As a stand-alone tool to get insight into data distribution
  - As a preprocessing step for other algorithms

Clustering for Data Understanding and Applications

- **Biology:** taxonomy of living things: kingdom, class, order, family, genus and species
- **Information retrieval:** document clustering
- **Land use:** Identification of areas of similar land use in an earth observation database
- **Marketing:** Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs

Cont..

- **City-planning:** Identifying groups of houses according to their house type, value, and geographical location
- **Earth-quake studies:** Observed earth quake epicenters should be clustered along continent faults
- **Climate:** understanding earth climate, find patterns of atmospheric and ocean
- **Economic Science:** market research

Clustering as a Preprocessing Tool (Utility)

- Summarization:
  - Preprocessing for regression, PCA, classification, and association analysis
- Compression:
  - Image processing: vector quantization
- Finding K-nearest Neighbors
  - Localizing search to one or a small number of clusters
- Outlier detection
  - Outliers are often viewed as those “far away” from any cluster

Quality: What Is Good Clustering?

- A good clustering method will produce high quality clusters
  - high intra-class similarity: cohesive within clusters
  - low inter-class similarity: distinctive between clusters
- The quality of a clustering method depends on
  - the similarity measure used by the method
  - its implementation, and
  - Its ability to discover some or all of the hidden patterns

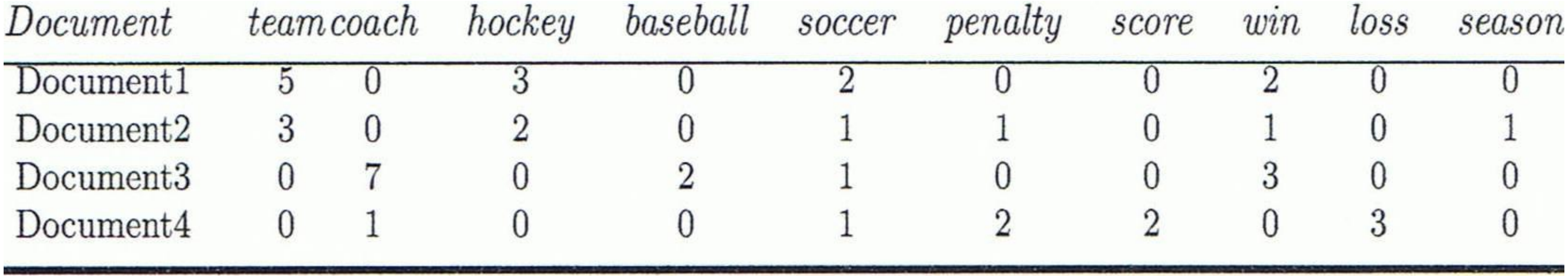
Cosine similarity:

- Let’s imagine that you need to determine how similar two documents or corpus of text are.
- Which distance metrics will you use?
- The answer is cosine similarity.
- In order to calculate it, we need to measure the cosine of the angle between two vectors. Then, cosine similarity returns the normalized dot product of them.

$$\mathbf{A} \cdot \mathbf{B} = \|\mathbf{A}\| \, \|\mathbf{B}\| \cos \theta$$

Cosine Similarity in Data mining Apps

- A **document** can be represented by thousands of attributes, each recording the *frequency* of a particular word (such as keywords) or phrase in the document.



Document	*team*	*coach*	*hockey*	*baseball*	*soccer*	*penalty*	*score*	*win*	*loss*	*season*
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	2	1	0	0	3	0	0
Document4	0	1	0	0	1	2	2	0	3	0

- Other vector objects: gene features in micro-arrays, …
- **Applications: information retrieval, biologic taxonomy, gene feature mapping, ...**
- Cosine measure: If d_1 and d_2 are two vectors (e.g., term-frequency vectors), then

$$\cos(d_1, d_2) = (d_1 \bullet d_2) / ||d_1||\ ||d_2||,$$

where $\bullet$ indicates vector dot product, $||d||$: the length of vector d

Example: Cosine Similarity

- $\cos(d_1, d_2) = (d_1 \bullet d_2) / ||d_1||\ ||d_2||$,
  where • indicates vector dot product, $||d|$: the length of vector d

- Ex: Find the **similarity** between documents 1 and 2.

$d_1 =$ (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)
$d_2 =$ (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)

$d_1 \bullet d_2 = 5*3+0*0+3*2+0*0+2*1+0*1+0*1+2*1+0*0+0*1 = 25$

$||d_1|| = (5*5+0*0+3*3+0*0+2*2+0*0+0*0+2*2+0*0+0*0)^{0.5}=(42)^{0.5} = 6.481$

$||d_2|| = (3*3+0*0+2*2+0*0+1*1+1*1+0*0+1*1+0*0+1*1)^{0.5}=(17)^{0.5} = 4.12$

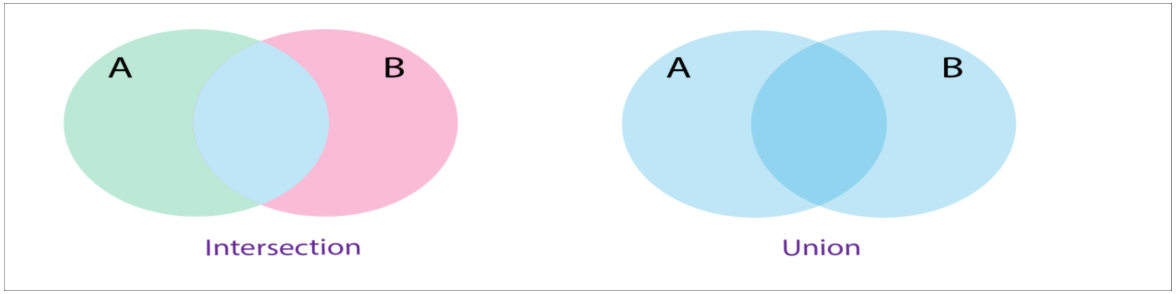
$\cos(d_1, d_2) = 0.94$

Jaccard similarity/Jaccard distance

- Similarity is computed through distance.
- A set is an unordered collection of objects.
- So for example, {1, 2, 3, 4} is equal to {2, 4, 3, 1}.
- We can calculate its cardinality (represented as |set|) which is no other thing than the number of elements contained in the set.

Cont..

- Let’s say we have two sets of objects, A and B. We wonder how many elements they have in common. This is called Intersection. It is represented mathematically as A ∩ B.
- Maybe, we want to get all items regardless of the set they belong to. This is called Union. It is represented mathematically as A U B.



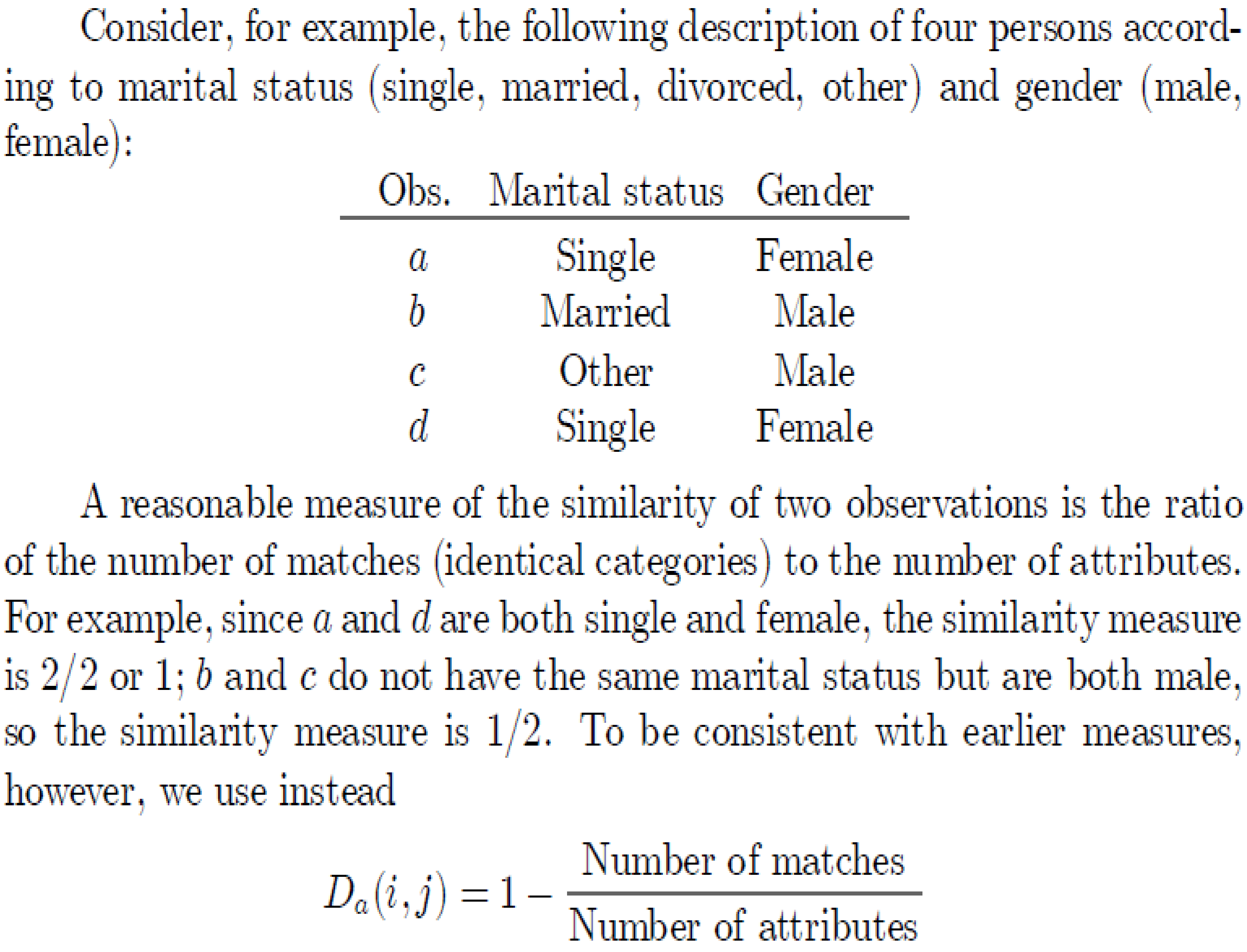
Cont..

- How does this relate to Jaccard similarity?
- Jaccard similarity is defined as the cardinality of the intersection of defined sets divided by the cardinality of the union of them.
- It can only be applied to finite sample sets.

Jaccard similarity = $|A \cap B| / |A \cup B|$

- Imagine we have the set A = {“flower”, “dog”, “cat”, 1, 3} and B = {“flower”, “cat”, “boat”}. Then, $A \cap B = 2$ and $A \cup B = 6$. As a result, the Jaccard similarity is 2/6 = 0.333.

DISTANCE MEASURES FOR ATTRIBUTES



Consider, for example, the following description of four persons according to marital status (single, married, divorced, other) and gender (male, female):

Obs.	Marital status	Gender
a	Single	Female
b	Married	Male
c	Other	Male
d	Single	Female

A reasonable measure of the similarity of two observations is the ratio of the number of matches (identical categories) to the number of attributes. For example, since a and d are both single and female, the similarity measure is 2/2 or 1; b and c do not have the same marital status but are both male, so the similarity measure is 1/2. To be consistent with earlier measures, however, we use instead

$$D_a(i,j) = 1 - \frac{\text{Number of matches}}{\text{Number of attributes}}$$

[A-13]: PRACTICE FOR YOU

1. A document can be represented by thousands of attributes, each recording the frequency of a particular word (such as keywords) or phrase in the document.

Formulate a scenario by Considering at least 6 documents and 12 keywords and keywords frequency may be assigned randomly. Compute the similarity among those documents.

2. Formulate a real-world scenario to use Jaccard similarity and compute the similarity between two object sets through the scenario.

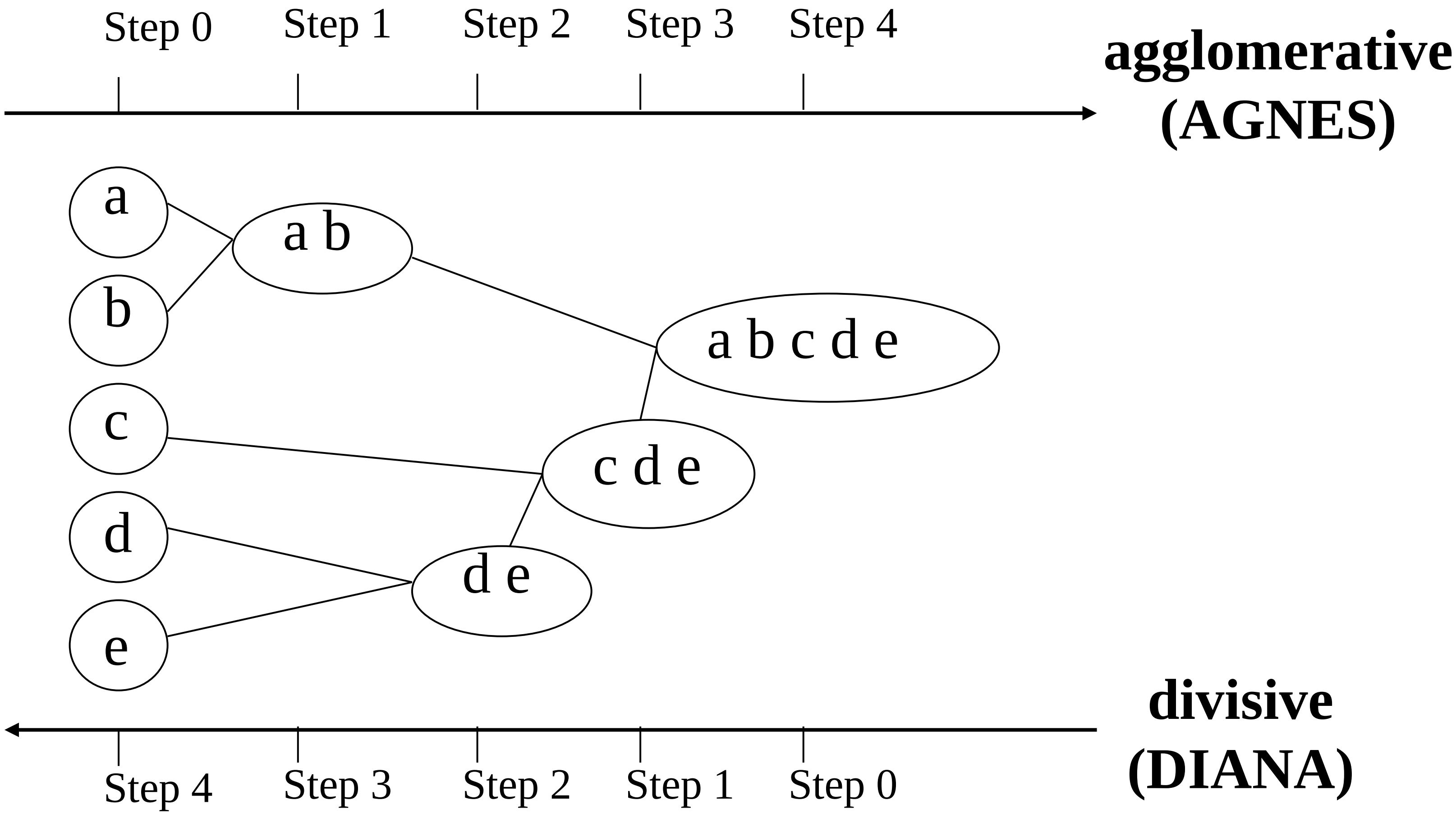
3. Consider a relation having 5 attributes and 6 records and compute the distance between those records or instances.

4.1-PART-02

Hierarchical Clustering

- Create a hierarchical decomposition of the set of data (or objects) using some criterion.
- Use distance matrix as clustering criteria. This method does not require the number of clusters k as an input, but needs a termination condition

Cont..


Step 0
Step 1
Step 2
Step 3
Step 4
agglomerative
(AGNES)
a
b
c
d
e
a b
d e
c d e
a b c d e
divisive
(DIANA)
Step 4
Step 3
Step 2
Step 1
Step 0

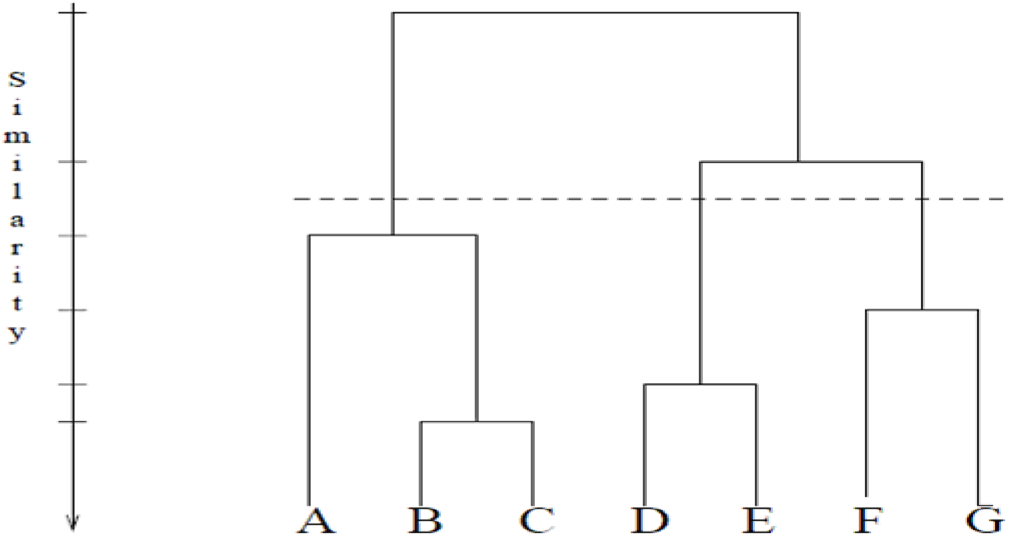
Hierarchical Clustering (Cont'd)

Two main types of hierarchical clustering

- **Agglomerative**:
  - ✓ Start with the points as individual clusters
  - ✓ At each step, merge the closest pair of clusters until only one cluster (or *k* clusters) left
- **Divisive**:
  - ✓ Start with one, all-inclusive cluster
  - ✓ At each step, split a cluster until each cluster contains a point (or there are *k* clusters)
- Traditional hierarchical algorithms use a *similarity* or *distance* matrix

Hierarchical Clustering (agglomerative)

- we assign each object (data point) to a separate cluster.
- Then compute the distance (similarity) between each of the clusters and join the two most similar clusters.
- A tree like diagram that records the sequences of merges or splits ----**DENDOGRAM**

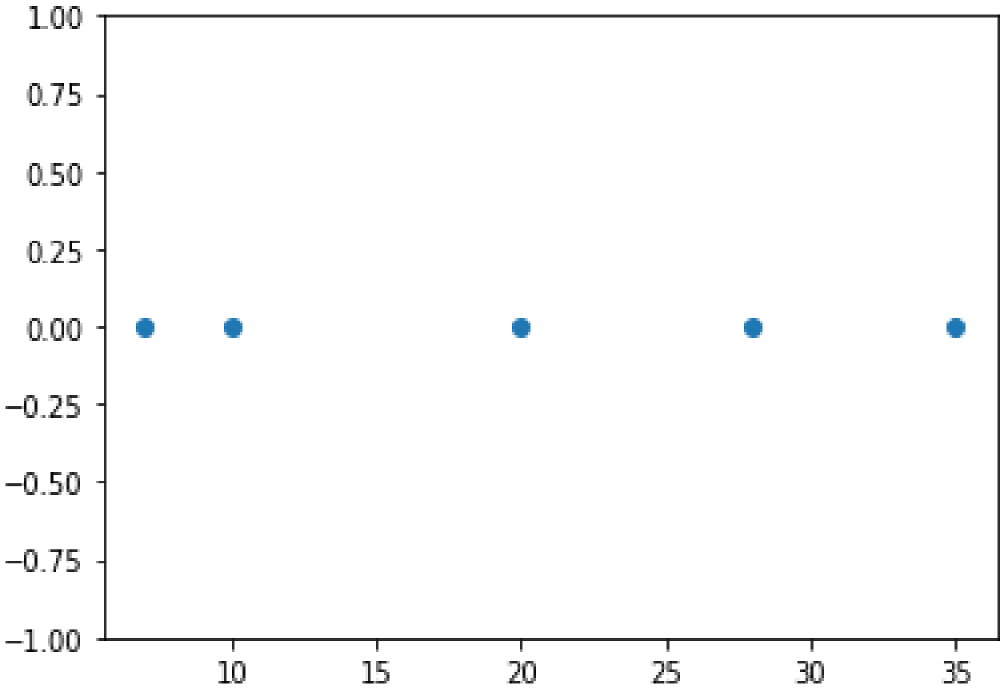


Example....problem

Consider one dimensional data set {7,10,20,28,35}, perform hierarchical clustering and plot the dendogram to visualize it.

Observation :

First, let’s the visualize the data.



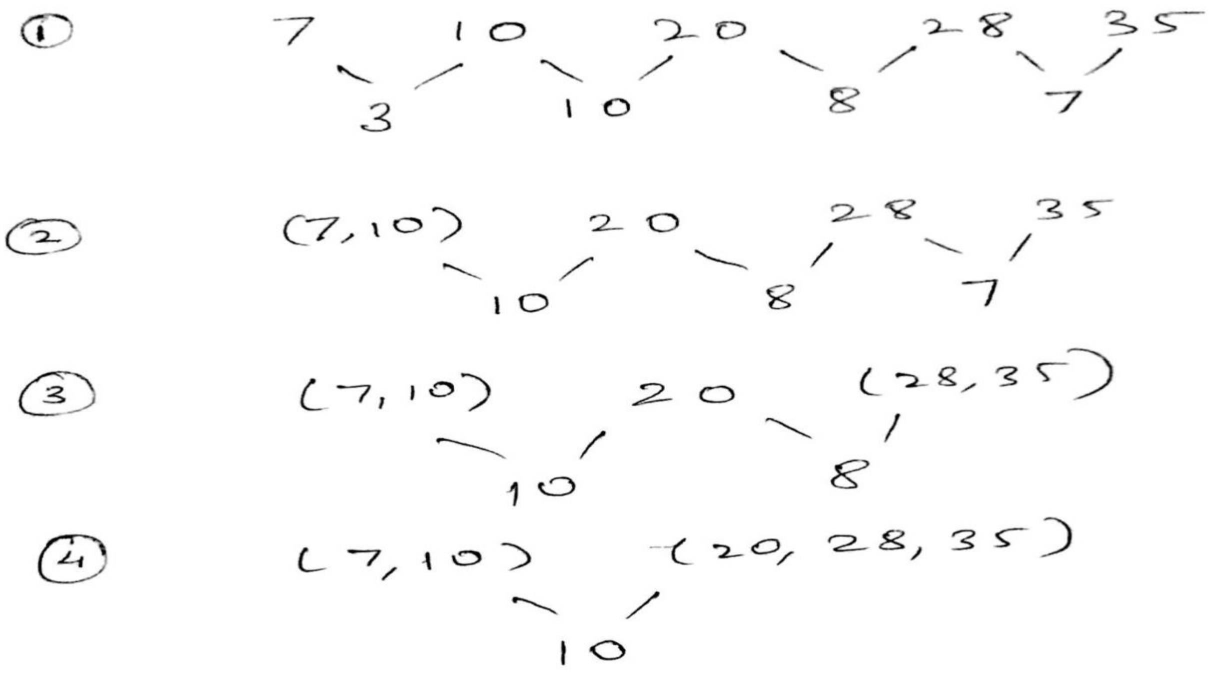
Observing the plot above, we can intuitively conclude that:

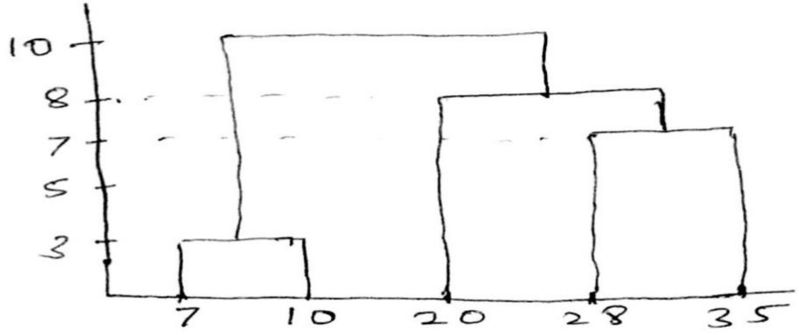
- The first two points (7 and 10) are close to each other and should be in the same cluster
- Also, the last two points (28 and 35) are close to each other and should be in the same cluster
- Cluster of the center point (20) is not easy to conclude.

Solution :

- Let’s solve the problem by hand using both the types of agglomerative hierarchical clustering :
- Single Linkage : In single link hierarchical clustering, we merge in each step the two clusters, whose two closest members have the smallest distance.

Single Linkage





(Dendogram)

Dendrogram

Solution analysis

Using single linkage two clusters are formed :

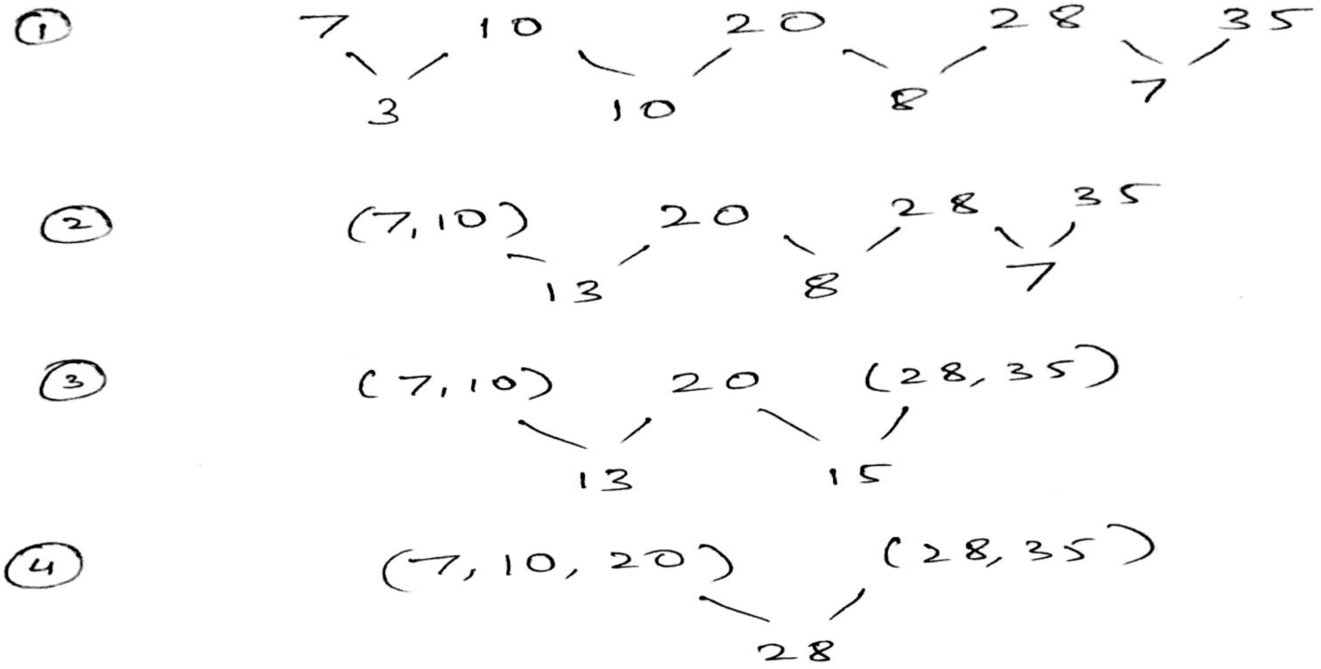
Cluster 1 : (7,10)

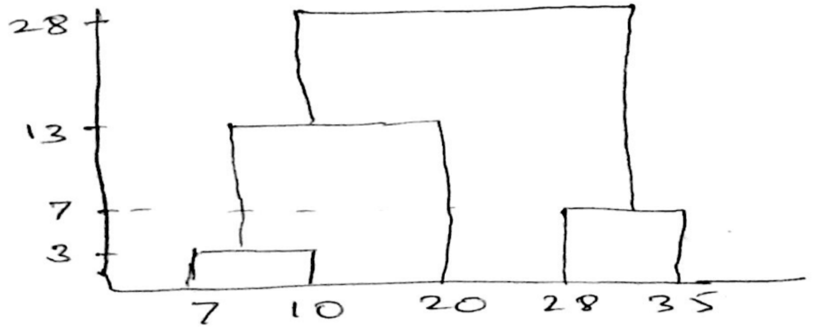
Cluster 2 : (20,28,35)

2. Complete Linkage :

In complete link hierarchical clustering, we merge in the members of the clusters in each step, which provide the smallest maximum pairwise distance.

Complete Linkage





(Dendogram)

Dendrogram

Analysis

Using complete linkage two clusters are formed

:

Cluster 1 : (7,10,20)

Cluster 2 : (28,35)

Conclusion :

- Hierarchical clustering is mostly used when the application requires a hierarchy, e.g creation of a taxonomy.
- However, they are expensive in terms of their computational and storage requirements.

Performing Hierarchical clustering on Dataset [R-Implements]

- ❖ Complete-linkage (farthest neighbor)
- ✓ distance is measured between the farthest pair of observations in two clusters.
- ❖ Single –linkage (nearest neighbor)
- ✓ distance is measured between the nearest pair of observations in two clusters.

Cont..

- ❖ Average-linkage – (Average distance)

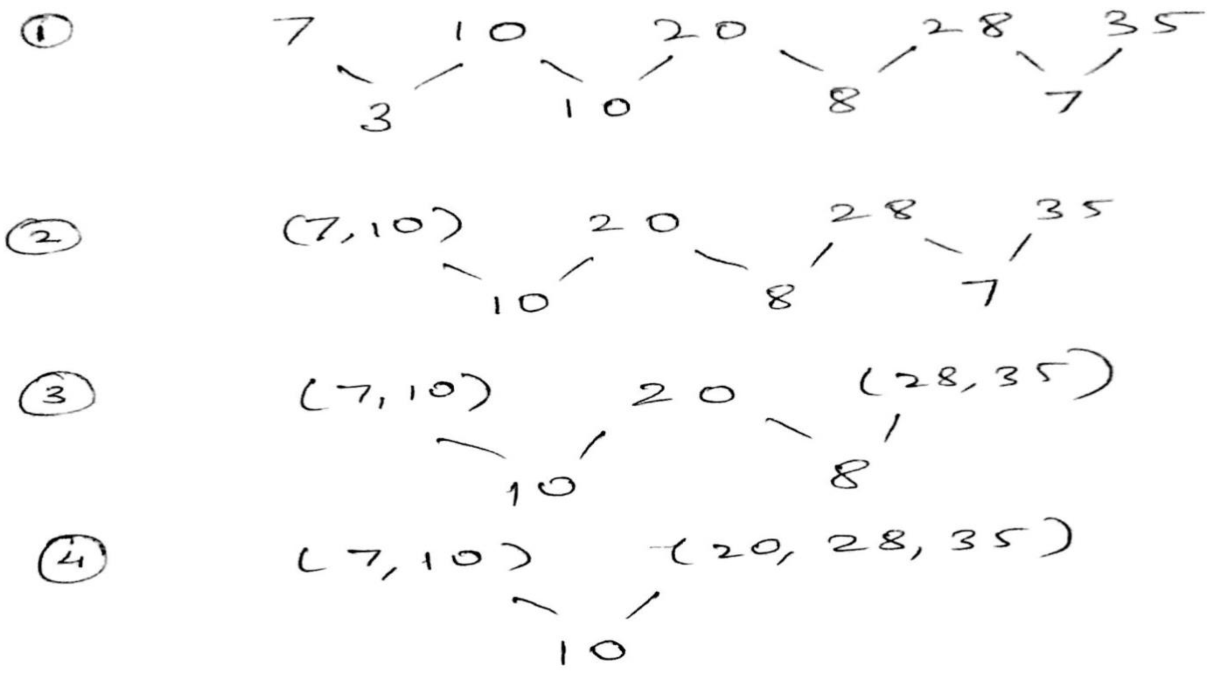
- ✓ Average-linkage is where the distance between each pair of observations in each cluster are added up and divided by the number of pairs to get an average inter-cluster distance.

Ex- C1 : (7,10)

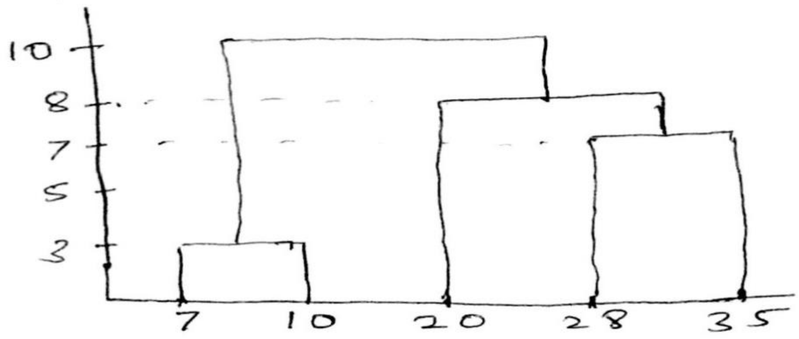
C2: 20

Distance = [(20-7)+ (20-10)]/2 = (13+10)/2 = 11.5

Single Linkage



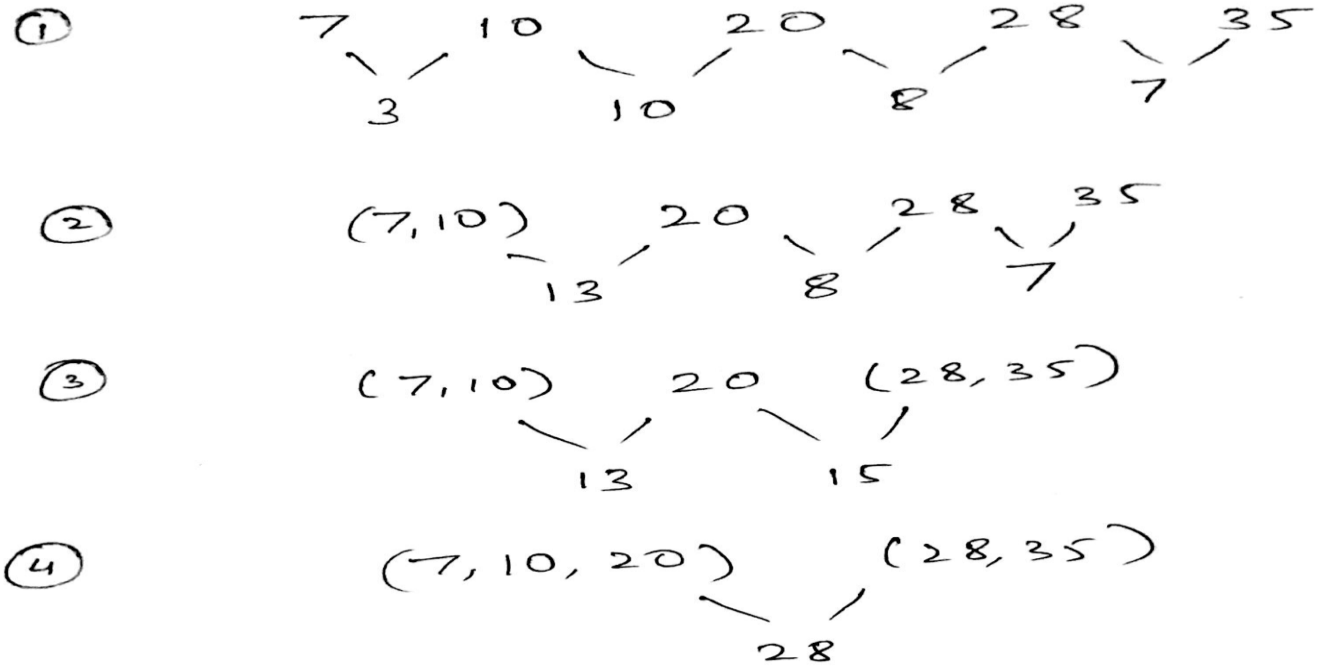
REVIEW

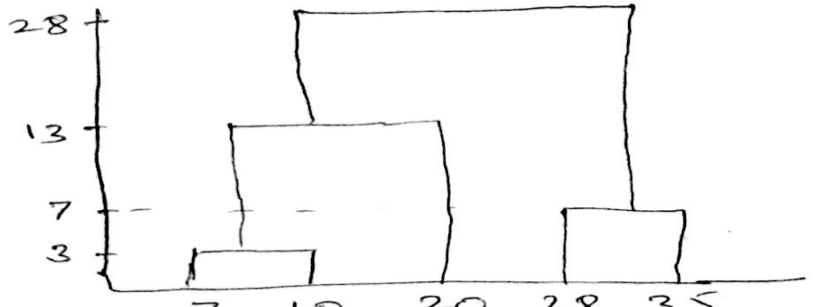


(Dendogram)

Dendrogram

Complete Linkage





(Dendogram)

REVIEW

Dendrogram

R-implements using Average-linkage Hierarchical clustering

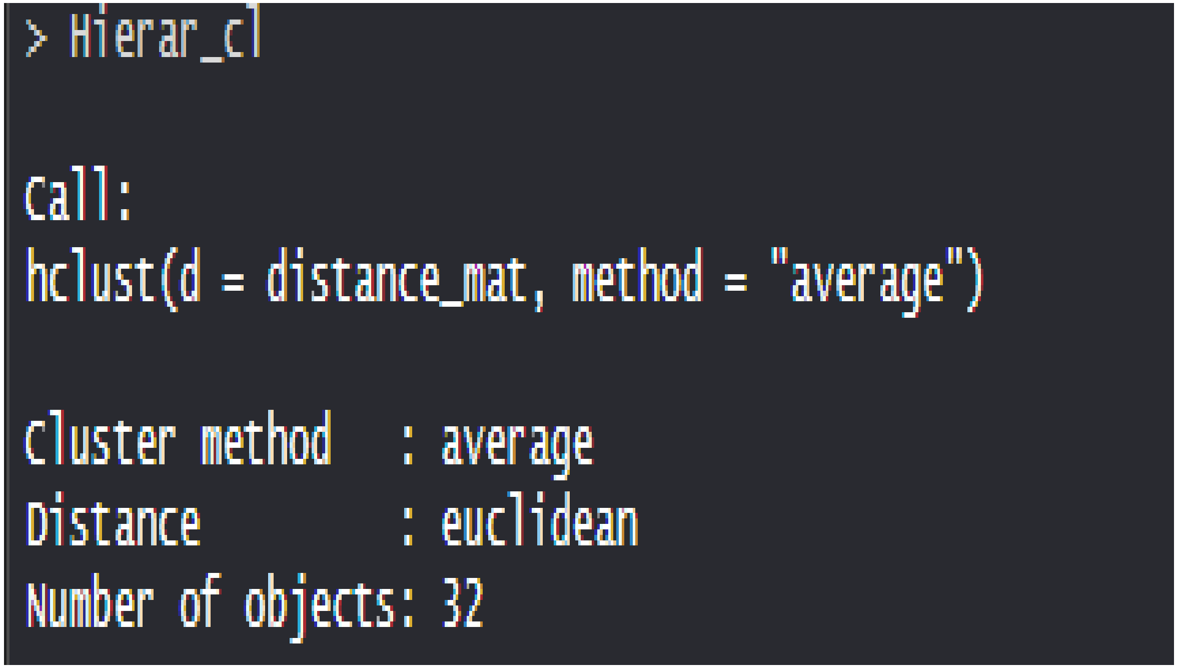
- ❖ hclust() is pre-installed in stats package when R is installed.

```
# Installing the package
install.packages("dplyr")

# Loading package
library(dplyr)

# Summary of dataset in package
head(mtcars)
```

```
# Finding distance matrix
distance_mat <- dist(mtcars, method = 'euclidean')
distance_mat

# Fitting Hierarchical clustering Model
# to training dataset
set.seed(240)  # Setting seed
Hierar_cl <- hclust(distance_mat, method = "average")
Hierar_cl

# Plotting dendrogram
plot(Hierar_cl)

# Choosing no. of clusters
# Cutting tree by height
abline(h = 110, col = "green")

# Cutting tree by no. of clusters
fit <- cutree(Hierar_cl, k = 3 )
fit

table(fit)
rect.hclust(Hierar_cl, k = 3, border = "green")
```

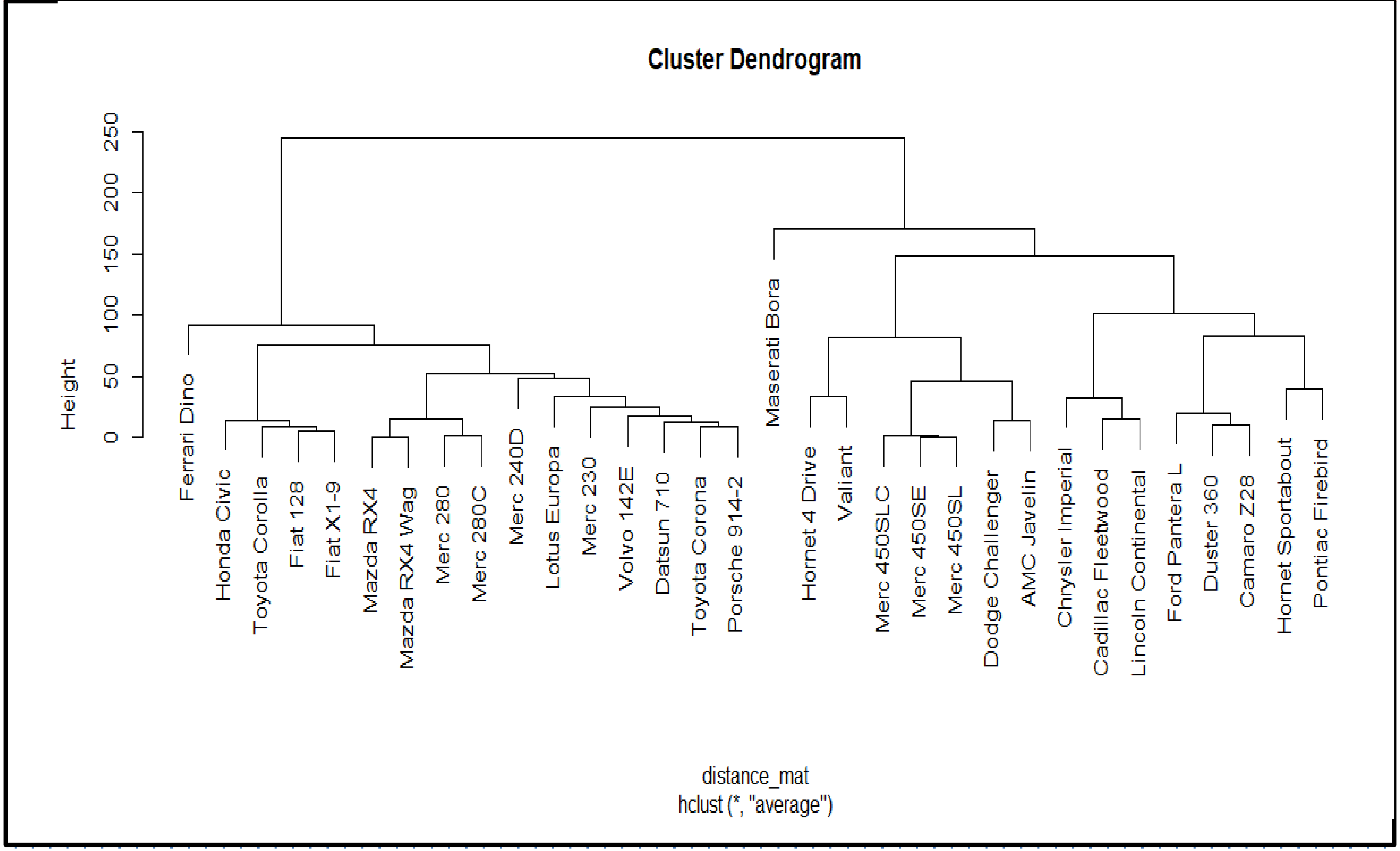
Model Hierar_cl:

In the model, the cluster method is average, distance is Euclidean and no. of objects are 32.



```
> Hierar_cl

Call:
hclust(d = distance_mat, method = "average")

Cluster method   : average
Distance         : euclidean
Number of objects: 32
```

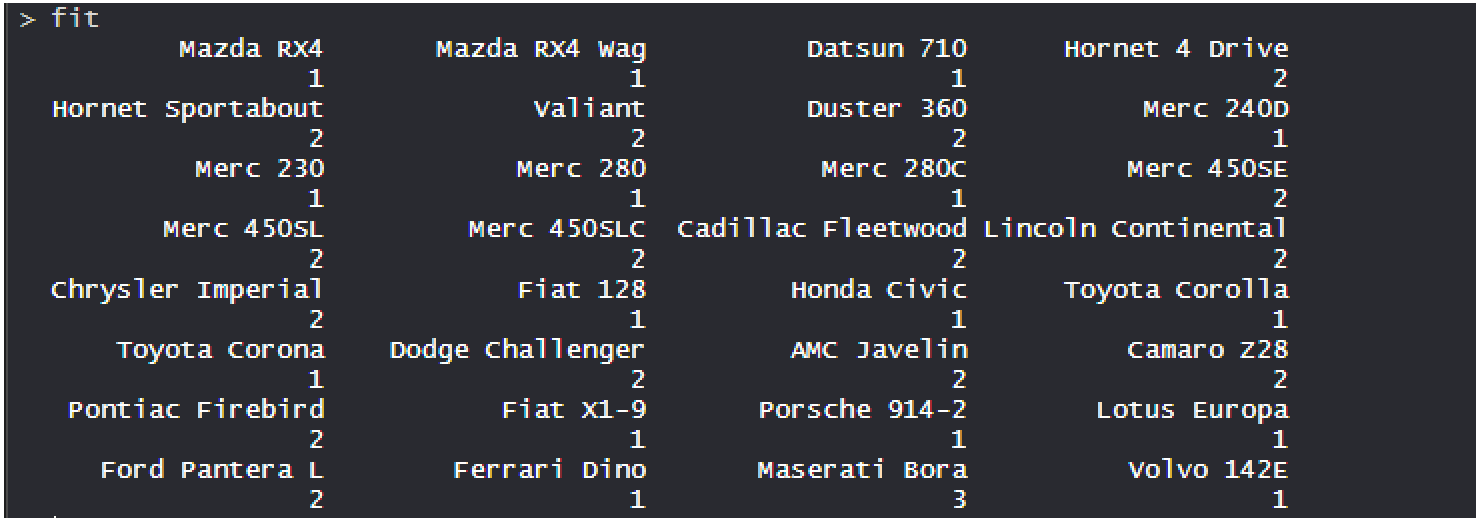
Plot dendrogram:



The plot dendrogram is shown with x-axis as distance matrix and y-axis as height.

Cutted tree:

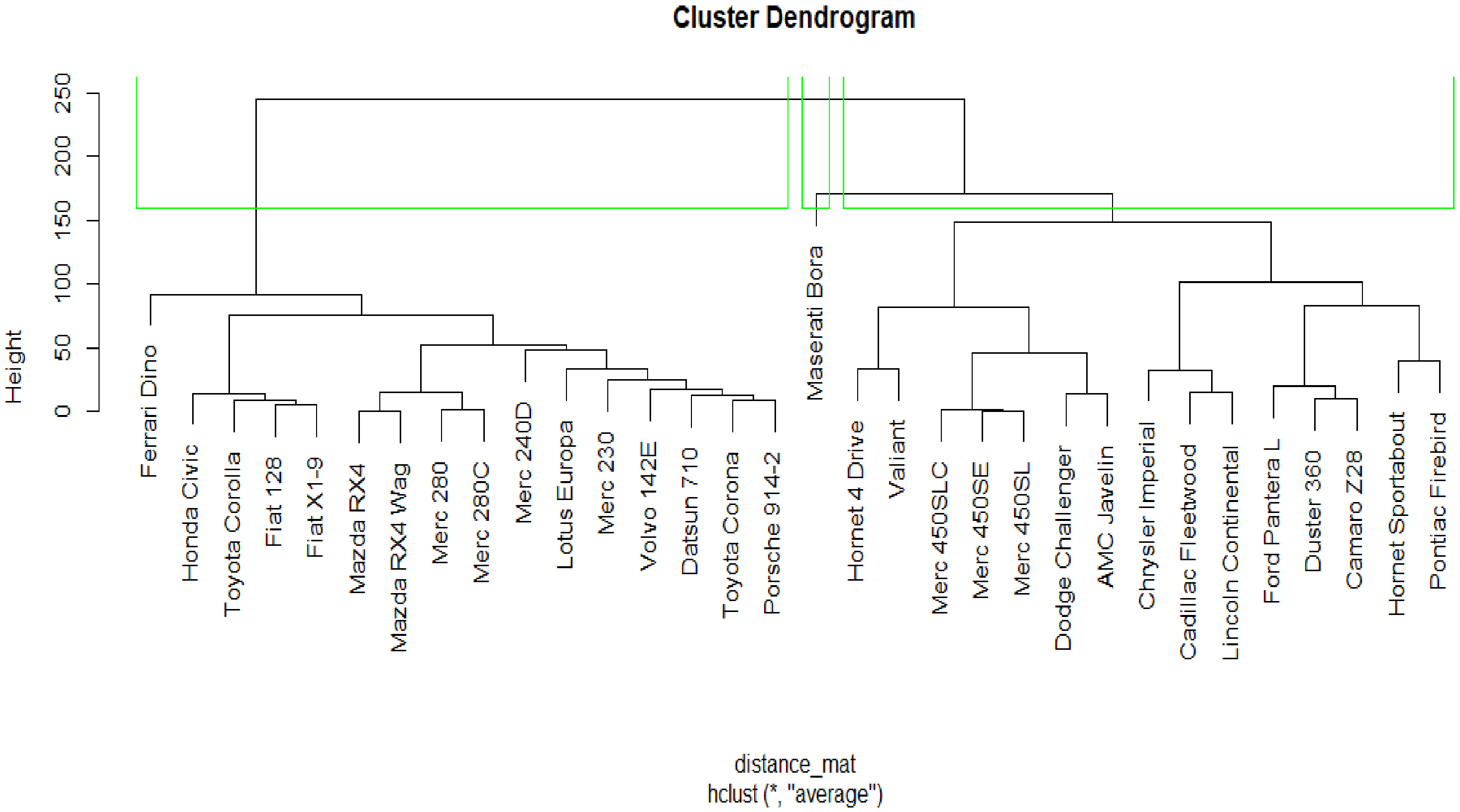
Tree is cut where k = 3 and each category represents its number of clusters.



```
> fit
          Mazda RX4       Mazda RX4 Wag          Datsun 710      Hornet 4 Drive 
                  1                   1                   1                   2 
  Hornet Sportabout             Valiant          Duster 360           Merc 240D 
                  2                   2                   2                   1 
           Merc 230            Merc 280           Merc 280C          Merc 450SE 
                  1                   1                   1                   2 
         Merc 450SL         Merc 450SLC  Cadillac Fleetwood Lincoln Continental 
                  2                   2                   2                   2 
  Chrysler Imperial            Fiat 128         Honda Civic      Toyota Corolla 
                  2                   1                   1                   1 
      Toyota Corona    Dodge Challenger         AMC Javelin          Camaro Z28 
                  1                   2                   2                   2 
   Pontiac Firebird           Fiat X1-9       Porsche 914-2        Lotus Europa 
                  2                   1                   1                   1 
     Ford Pantera L        Ferrari Dino       Maserati Bora          Volvo 142E 
                  2                   1                   3                   1 
```

Plotting dendrogram after cutting:

The plot denotes dendrogram after being cut. The green lines show the number of clusters



A-14: LAB--07

Consider a dataset without any class level and implement the hierarchical clustering algorithm.

Try to analyze and understand the results.

TASK FOR YOU [A15]

Consider one dimensional data set {8,11,21,29,36}, perform hierarchical clustering to decide the clusters using both single linkage and complete linkage analysis and plot the respective dendrogram to visualize it.

Give some key inferences to distinguish between those two analysis.

Cheers For the Great Patience!

Query Please?