MONASH University

ETX2250/ETF5922: Data Visualization and Analytics

Hierarchical clustering

Lecturer: *Lauren Kennedy*

Department of Econometrics and Business Statistics

lauren.kennedy1@monash.edu

Week 7

What is cluster analysis?

- Cluster: collection of data objects
- Grouping similar objects in a cluster
- No pre-defined classes means we are doing unsupervised classification

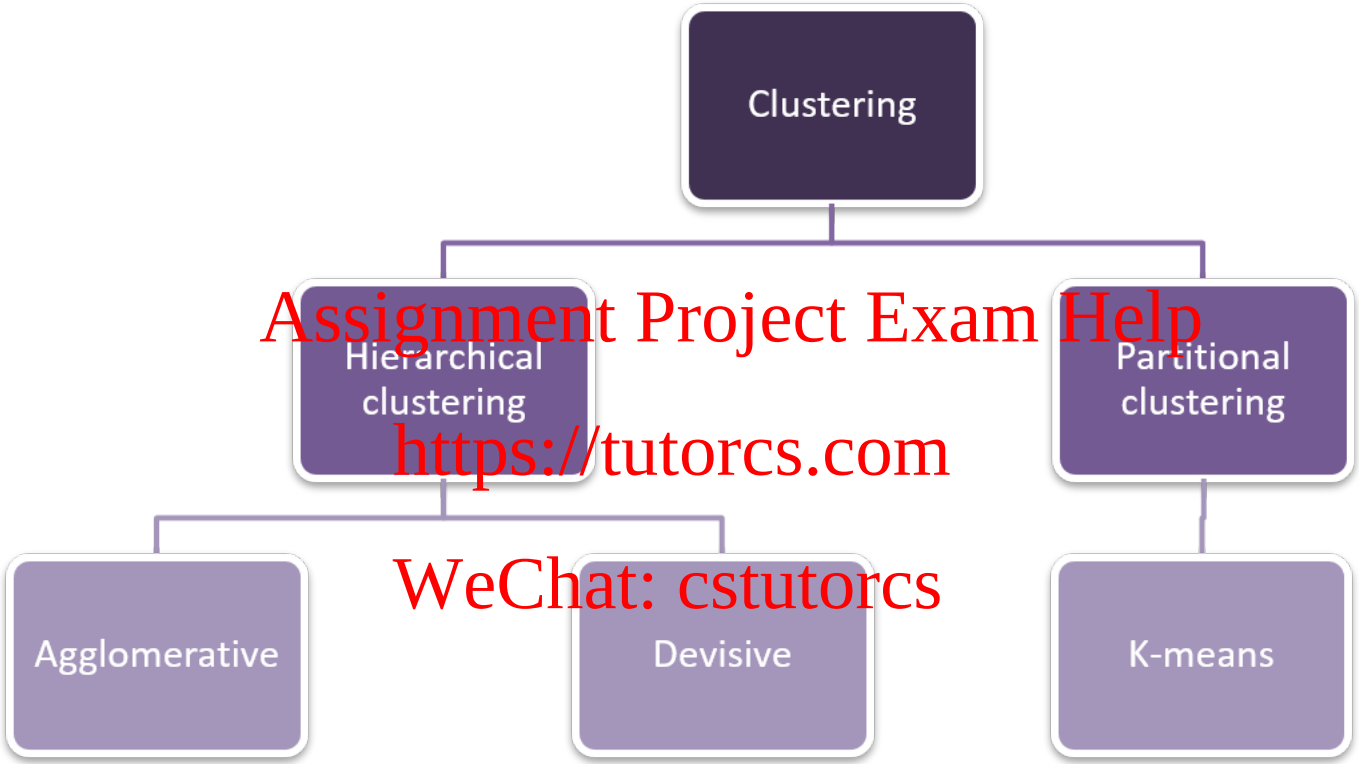
For analysis of data

- Cluster analysis – grouping a set of data objects into cluster
- In what contexts would you want to complete a cluster analysis?

Cluster analysis applications:

- Market segmentation – targeted marketing
- Market structure analysis – identify groups of similar products according to competitive measures of similarity
- Finance – develop a balanced portfolio eg. Volatility, beta, vega, returns over different periods
- Land use – identify are of similar land use in Earth observation database
- Medical eg. Types of tumour
- Similar responses on a survey
- Dimension reduction

Clustering methods



Hierachical Clustering

Agglomerative

1. Group each observation into cluster of one
2. Merge most similar clusters
3. Merge those a but further apart
4. Sequence of nested clusters

Devisive

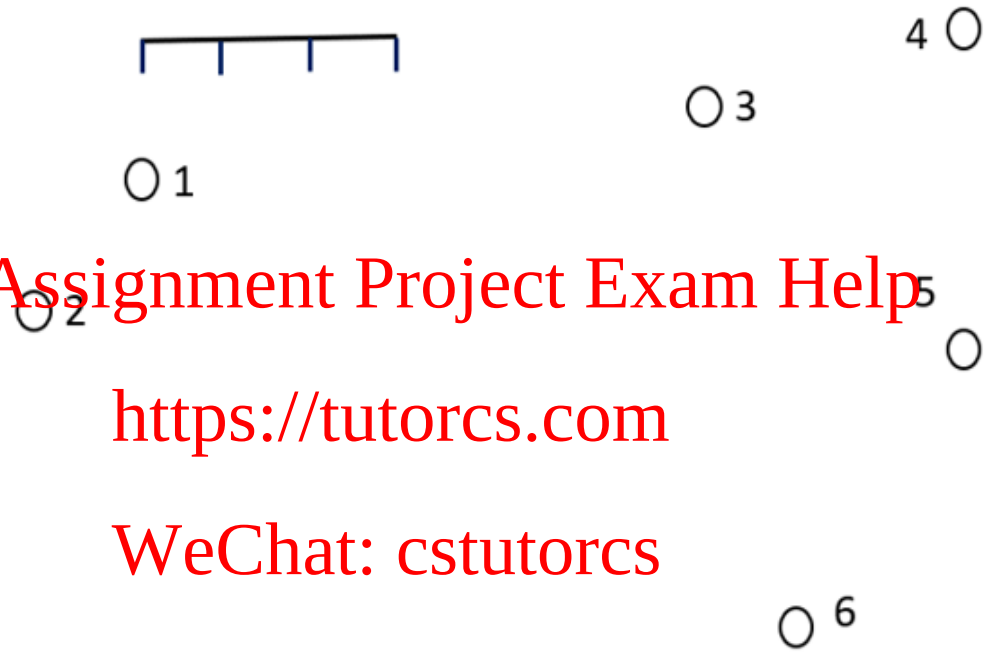
1. Start with one large cluster (all observations)
2. Recursively split into smaller clusters

Partitional clustering

K-means

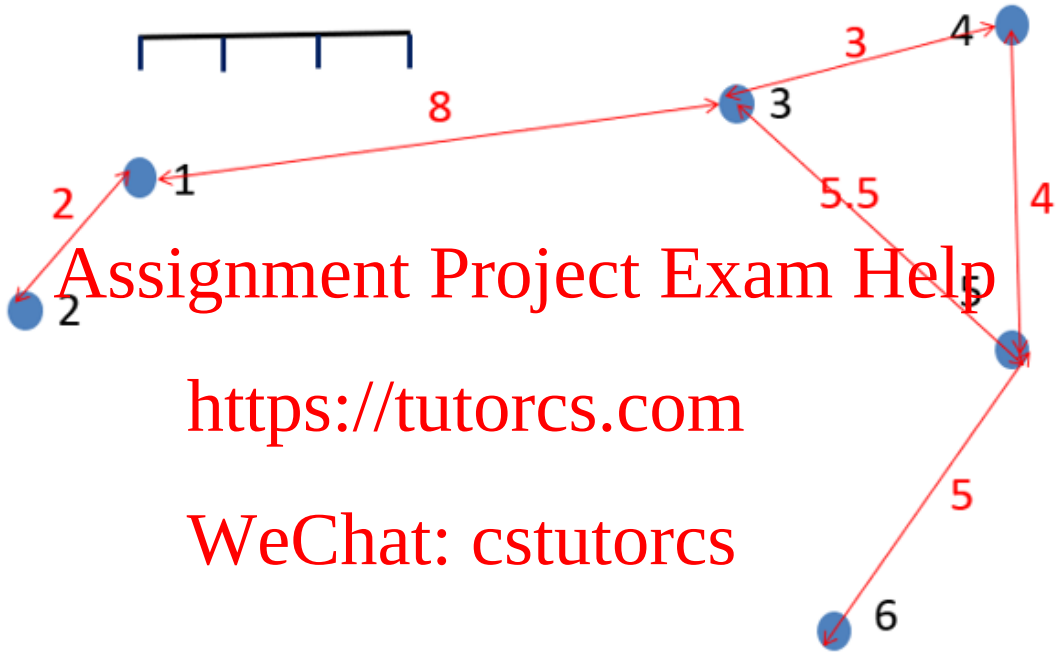
1. Random assignment of points in the midst of cluster
2. Repeatedly re-assign points cluster
3. Minimise within-cluster distances
4. Maximise between-cluster distances

An example



Distances between points in red; Names of points in black

An example

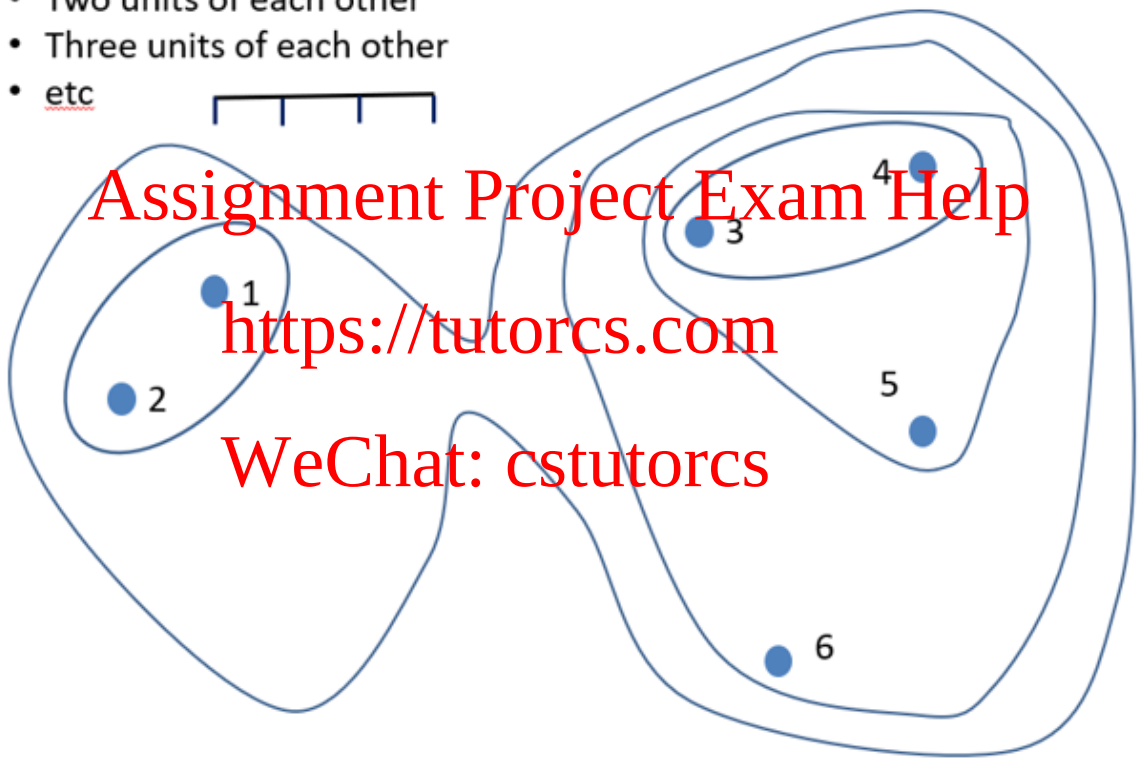


Distances between points in red; Names of points in black

An example

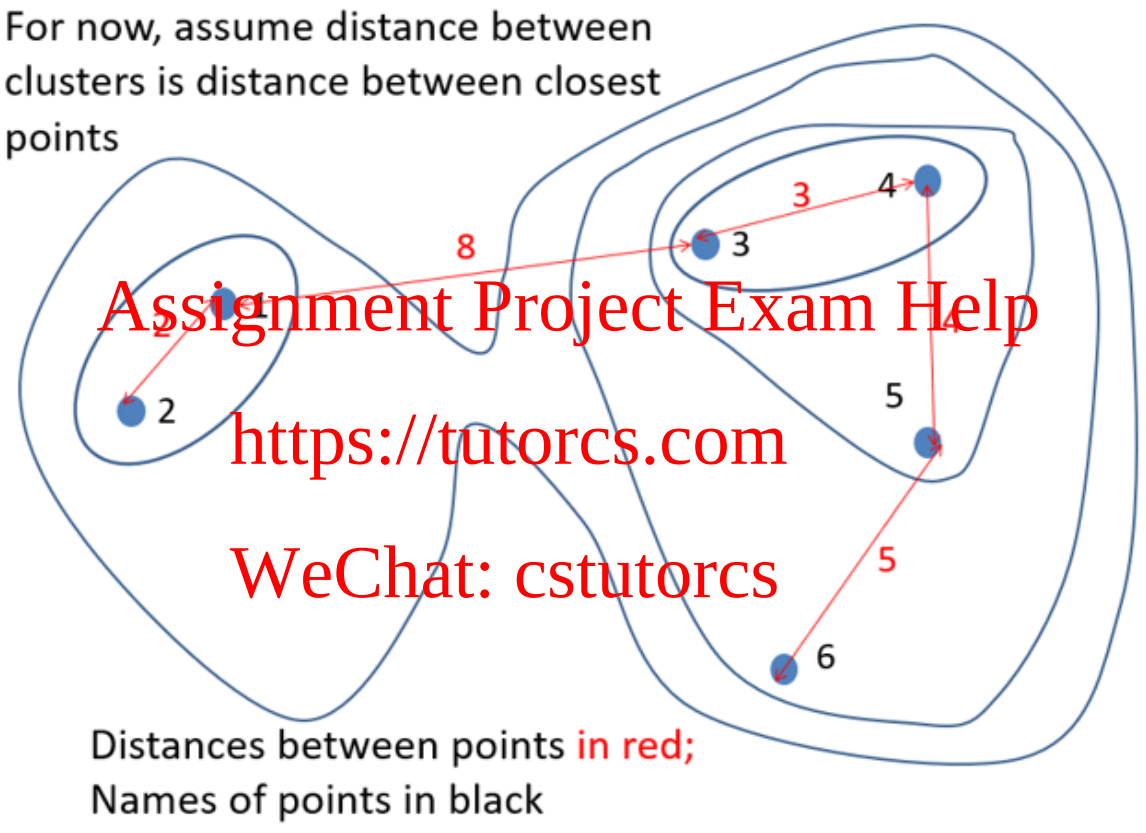
Hierarchical clustering: Suppose in order to be in a cluster, points have to be within:

- Two units of each other
- Three units of each other
- etc

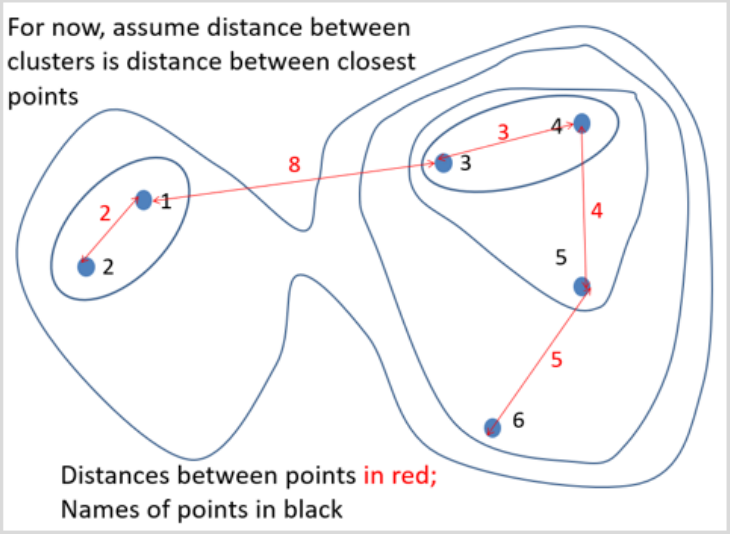


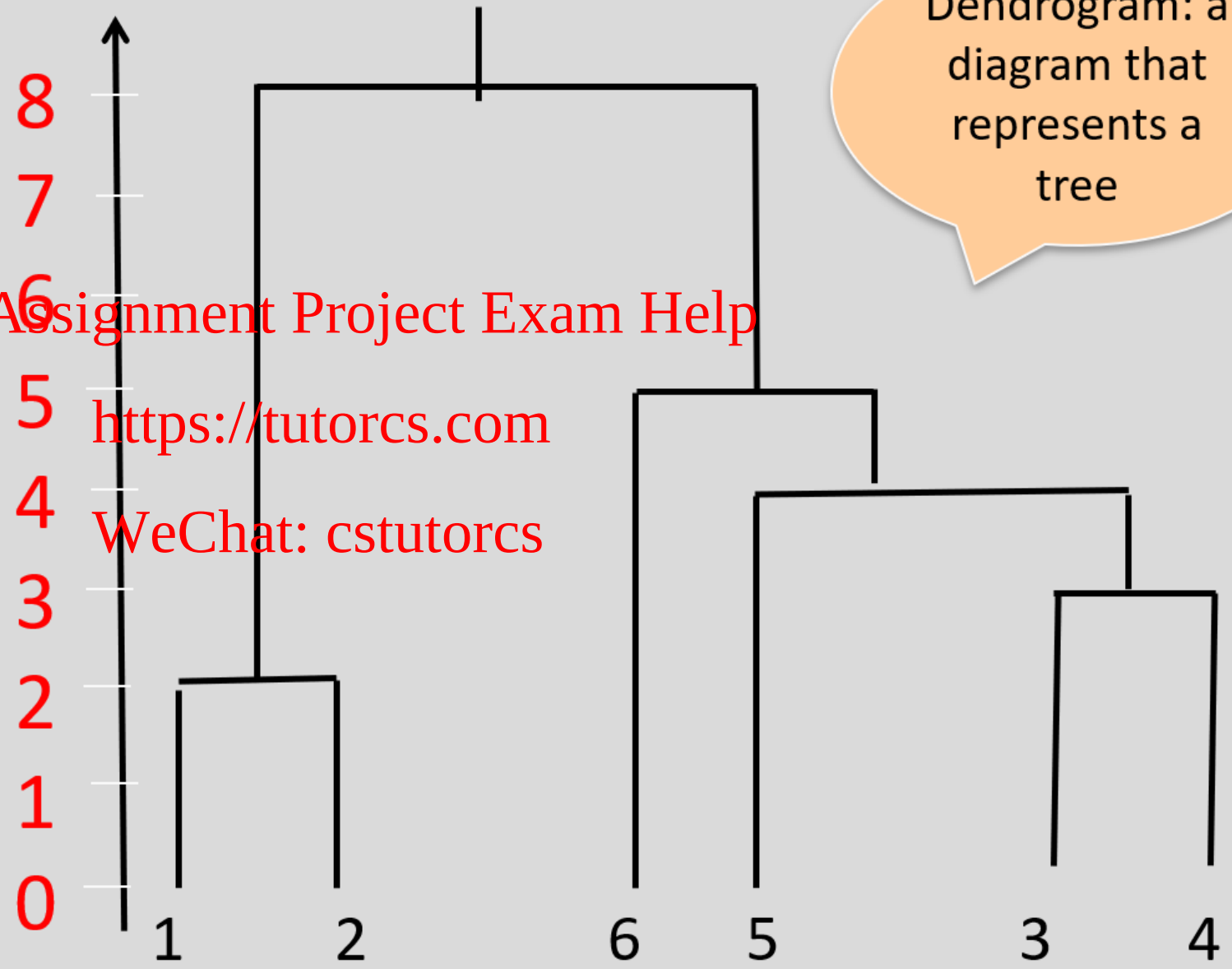
Distances between points in red; Names of points in black

An example



MONASH University



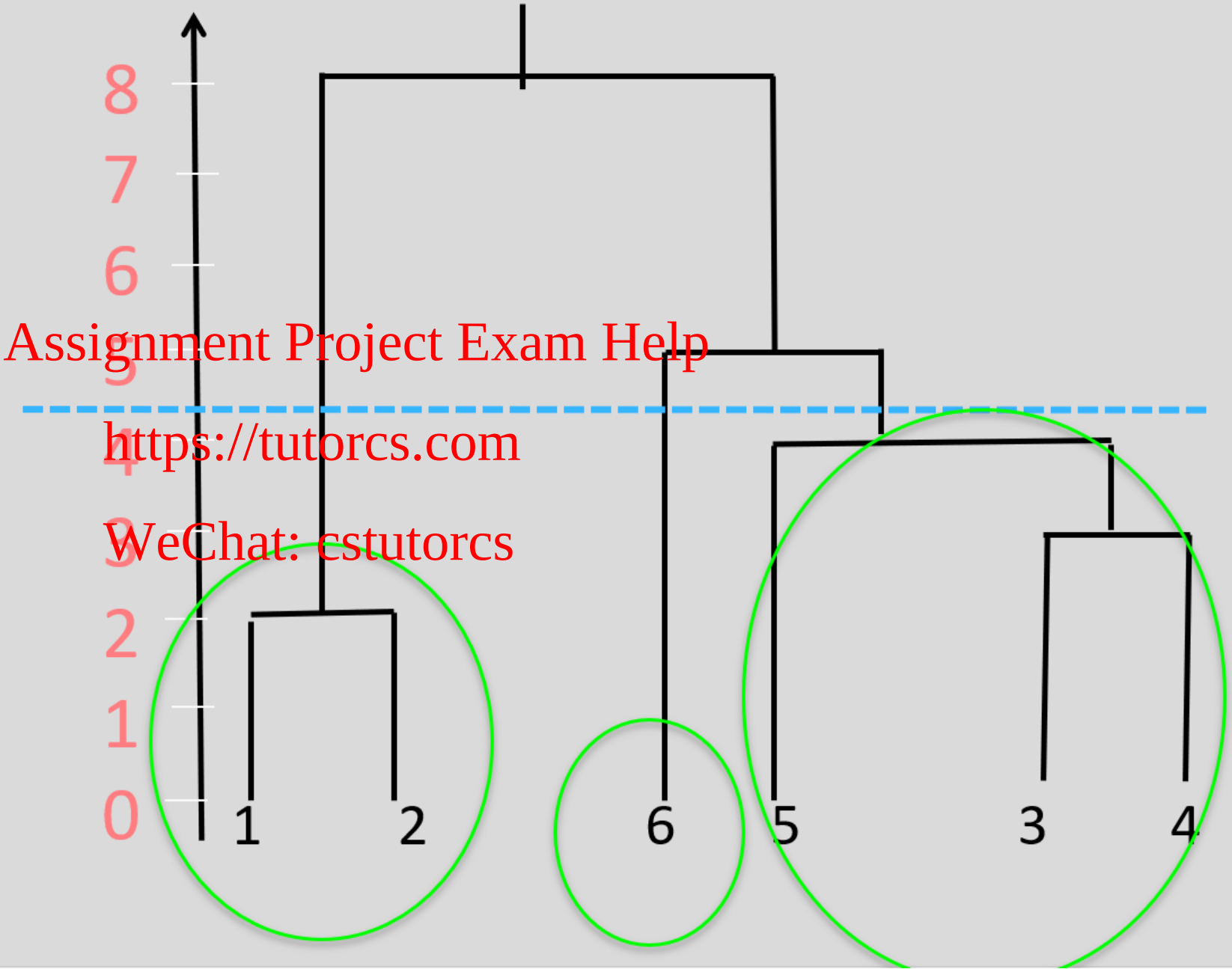


Dendrogram: a diagram that represents a tree

MONASH University

If this distance is just above 4, we create three clusters.

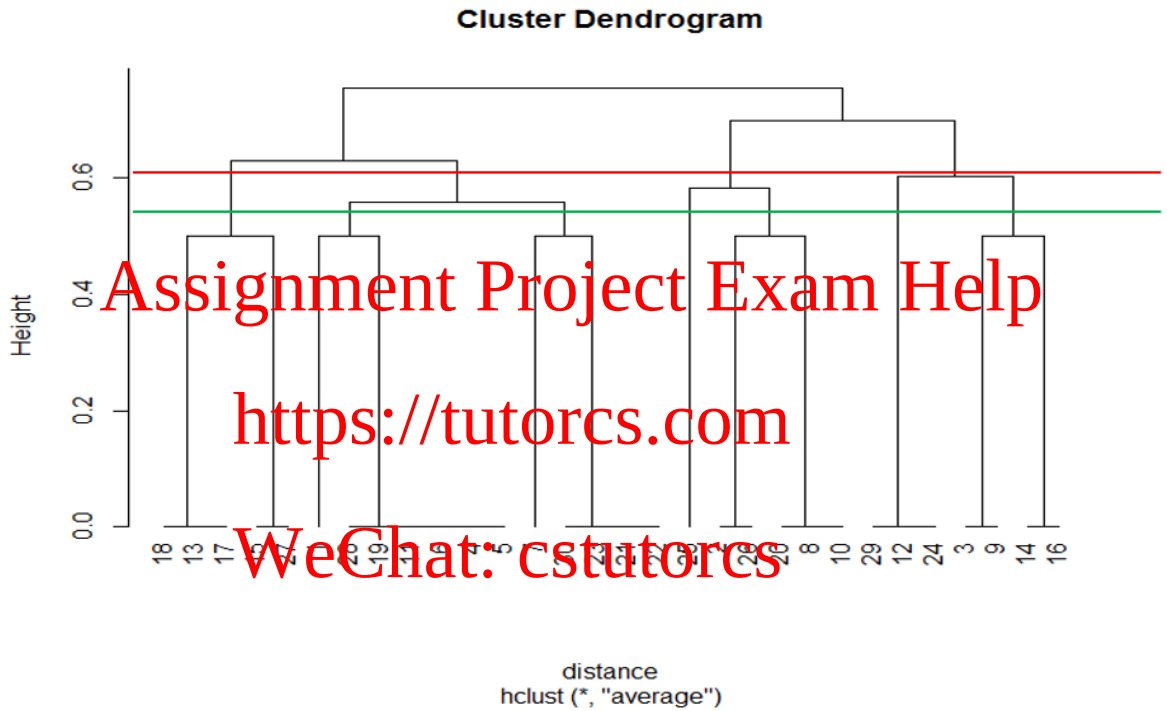
Name the members of the clusters



Dendrogram

- Visually, see what the possible numbers of clusters are by introducing a horizontal line
- Exercise: Based on the dendrogram on the next slide, if there are to be 7 clusters, list the ID numbers corresponding to the cluster containing the ID 10
- In the Hierarchical clustering tutorial, we use R to select the clusters and then investigate cluster properties

Example



Red line = 4 clusters Green line = 7 clusters

Distance

We need to decide

- How to define the distance between two data points
  - Recall that the data may be of different types
- How to define the distance between two clusters
  - This will depend on the distance between two points, but once that distance is decided there are still many alternative ways of defining distance between clusters

Distance

- In the example, we used
  - Euclidean distance (the usual shortest distance between two points)
  - between-cluster distance was taken to be the distance between the two points, one in each cluster, closest together
- When you have a data frame whose columns represent many different kinds of variable, defining distance between data points takes some thought.

Distance

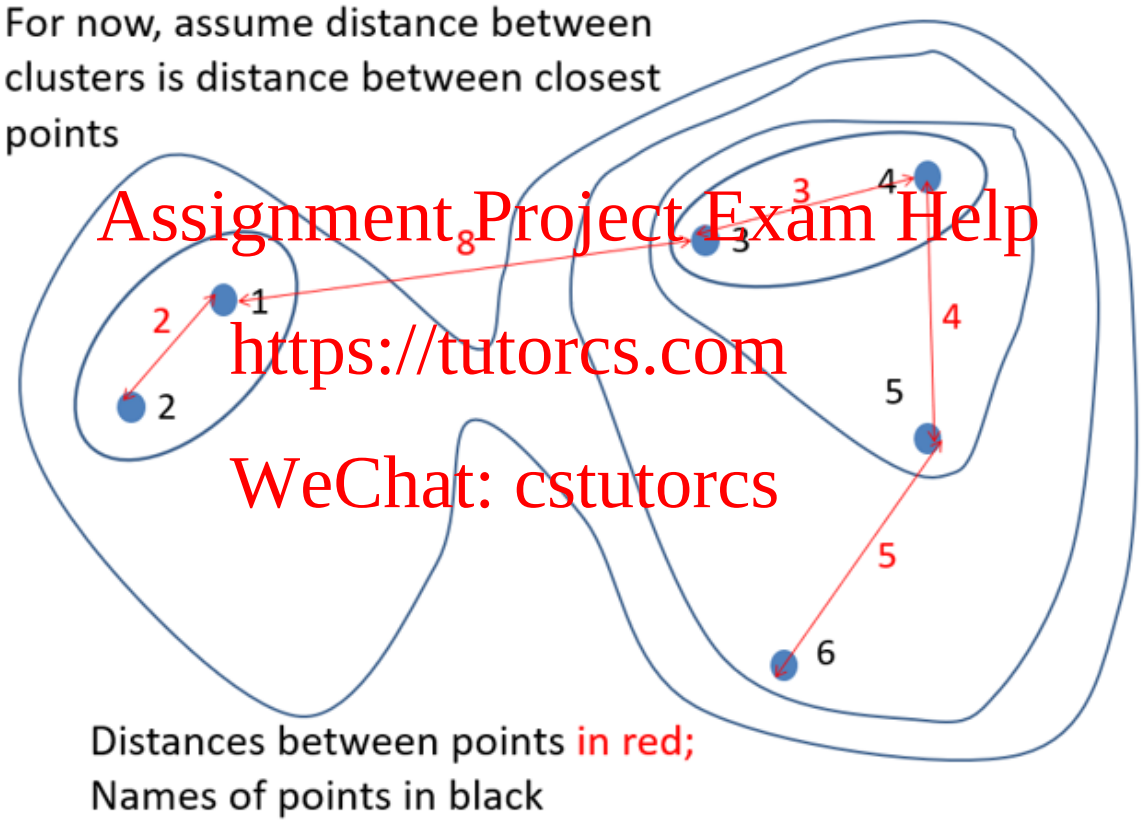
- Need to decide:
  - What are “distances” between points (=cases=rows)?
- And for the hierarchical method, in addition
  - What are “distances” between clusters?
- Distance is a dissimilarity measure (larger distance, less similar)
- For categorical variables, see how many attributes match: Matching is a similarity measure (more matching, more similar)
- Where a similarity measure s has been used, a corresponding “distance” d is defined to feed into the cluster algorithm:

$$d = \sqrt{1 - s}$$

Distance between clusters

Distance between clusters

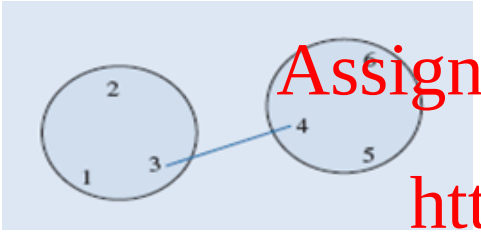
- In this diagram, we assume distance between clusters is distance between the closest points.



Distance between clusters

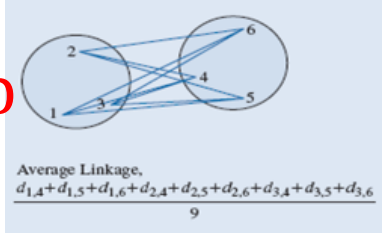
Single Linkage

Shortest distance between clusters



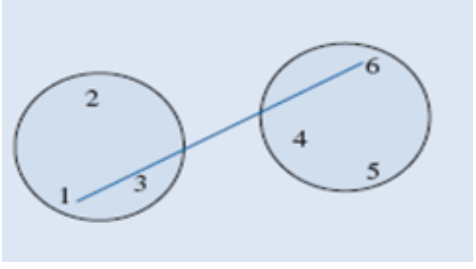
Average linkage

Average distance between all linkage



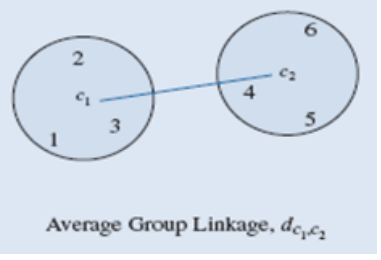
Complete linkage

Longest distance between clusters



Centroid linkage

Average distance between clusters



Distance Matrix to Dendrogram

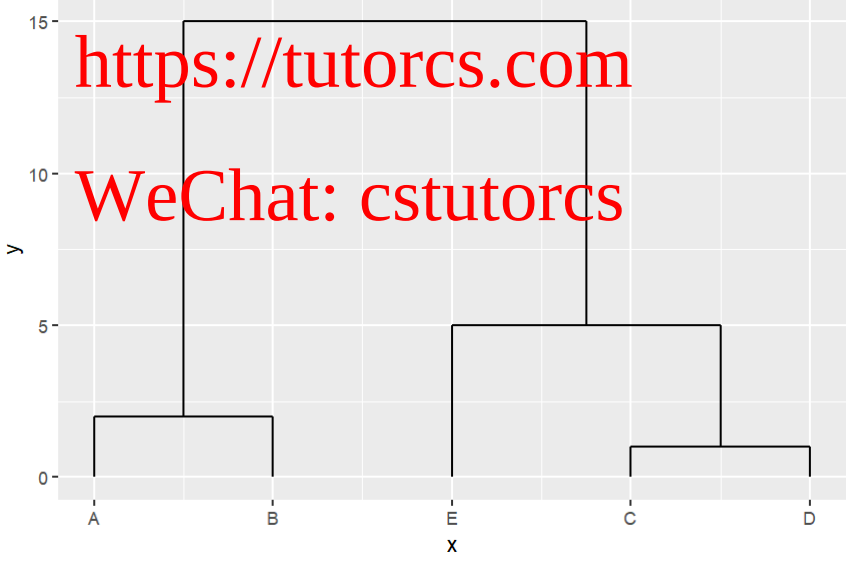
	A	B	C	D	E
A	0	2	8	10	13
B	2	0	10	15	14
C	8	10	0	1	5
D	10	15	1	0	4
E	13	14	5	4	0

Given these distances, create a dendrogram for:

1. Complete linkage
2. Single linkage

Complete linkage

```
library(ggplot2)
library(ggdendro)
library(ade4)
hc_complete<-hclust(as.dist(dist_matrix), method = "complete")
ggdendrogram(hc_complete)+ theme_gray()
```

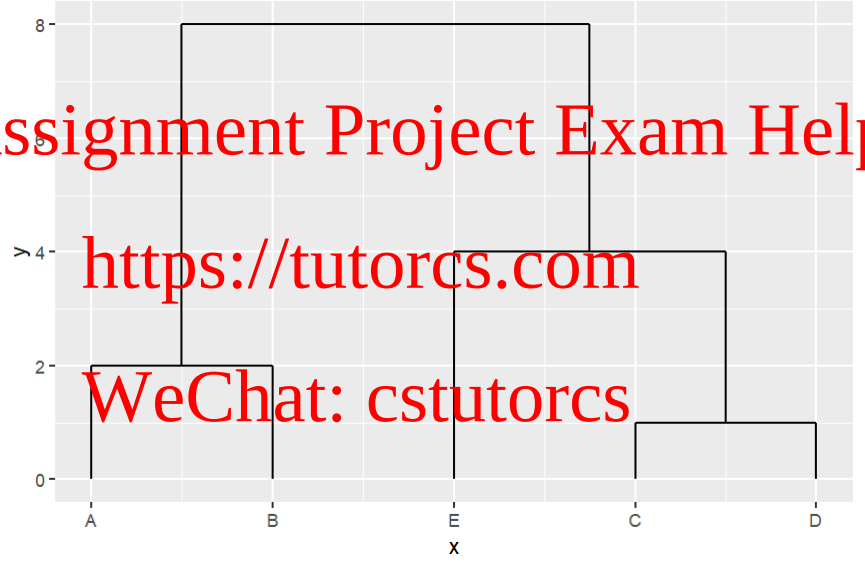



Breaking down the steps

- Let's start with the smallest distance:
  - C-D (dist 1): This is first cluster, recompute the similarities by considering C and D as one observation = 1
  - A-B (dist 2): is the second cluster, recompute the similarities = 2
- For the next step, look for the larger distance between D-E or C-E. Which is longer? (D-E = 4), (C-D = 5).
- For complete linkage, look for maximum distance between cluster A new cluster C-D-E = 5
- Now look for maximum distance between A-B and C-D-E A-B to C-D-E = 15

Single linkage

```
hc_single<-hclust(as.dist(dist_matrix), method = "single")
ggdendrogram(hc_single)+ theme_gray()
```



Breaking down the steps

- Let's start with the smallest distance:
  - C-D (dist 1): This is first cluster, recompute the similarities by considering C and D as one observation = 1
  - A-B (dist 2): is the second cluster, recompute the similarities = 2
- For the next step, look for the shorter distance between D-E or C-E. Which is shorter? (D-E = 4), (C-D = 5).
- For single linkage, look for minimum distance between cluster
  - A new cluster C-D-E = 4
- Now look for minimum distance between A-B and C-D-E
  - A-B to C-D-E = 8

Centroid

- A centroid is the average observation of a cluster
- That is, the average value for each variable across the points in the cluster
- The centroid doesn't have to be an actual observation

Hierachical Cluster: Pros and Cons

Hierarchical clustering methods differ in the method of representing similarity between clusters, each with advantages and disadvantages

Single linkage

- Most versatile algorithm
- Poorly delineated cluster structures within data can produce unacceptable snakelike chairs for clusters

Complete linkage

- Eliminates the chaining problem
- Only considers outermost observations in a cluster, so can be very impacted by outliers

Hierachical Cluster: Pros and Cons

Average linkage

- Based on average similarity of all individuals in a cluster and tends to generate clusters with small within-cluster variation
- Less affected by outliers

Centroid linkage

- Measures distance between cluster centroids
- Less affected by outliers

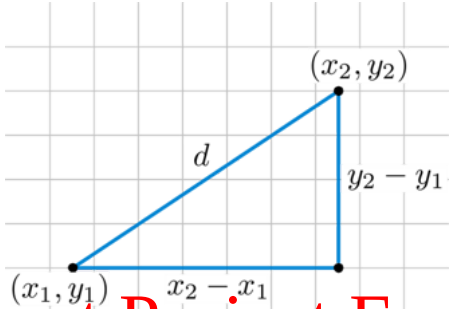
Different types of distance

Numerical distance

For numerical variables, dissimilarity is measured by:

- Euclidean distance
- Manhattan distance
- Many others

Numerical distance



Euclidean distance becomes smaller as observations become more similar with respect to their variable values. Same is true for Manhattan distances. Euclidean distance

$$d_E = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Manhattan distance

$$d_M = |x_2 - x_1| + |y_2 - y_1|$$

Euclidean distance

- Most common method to measure the distance between observations, when observations include continuous variables.
- Let observations $u = (Age_u, Income_u, Kids_u)$ and $v = (Age_v, Income_v, Kids_v)$ each comprise measurements of 3 variables.

The Euclidean distance between observations between u and v is

$$d_{u,v} = \sqrt{(Age_u - Age_v)^2 + (Income_u - Income_v)^2 + (Kids_u - Kids_v)^2}$$

Manhattan Distance

Manhattan distance: Distance between observations that include continuous variables

- Let observations u and v each comprise measurements of the variables Age, Income, and Kids.
- The Manhattan distance between observations u and v is

$$d_{u,v} = |Age_u - Age_v| + |Income_u - Income_v| + |Kids_u - Kids_v|$$

- Also need to standardize variables for Manhattan distance
- Manhattan distance is more robust to outliers

Matching (Dis)Similarity between observations

- We've discussed using distance measures for numeric variables.
- How can we discuss similarity measures for nominal or binary variables?
- For binary variables we consider two (there are many others!)
  - Matching coefficient
  - Jaccard's coefficient

Matching coefficient

Simplest overlap measure of similarity between two observations

- Find the proportion of variables with matching values
- Used when clustering observations solely on the basis of categorical variables encoded as 0 or 1
- Matching coefficient

$$S_M = \frac{\texttt{N variables matching value for obs}u \texttt{ and } v}{\texttt{total number of variables}}$$

Weakeness of matching coefficient

- If two observations both have a 0 entry for a categorical variable, this is counted as a sign of similarity between the two observations
- However, matching 0 entries do not necessarily imply similarity
- If 1 means “Owns a Toyota Land Cruiser” and 0 means “Does not own a Toyota Land Cruiser” then two observations sharing a zero does not indicate significant similarity

Jaccard's coefficient: A similarity measure that does not count matching zero entries

$$S_J = \frac{\texttt{N vars with matching nonzero value for obs } u \texttt{ and } v}{(\texttt{total N vars}) - (\texttt{N vars with matching zero values for obs} u \texttt{ and } v)}$$

Your turn!

- For u and v as follows, what is the similarity using
  - Matching
  - Jaccard

u=	0	0	0	0	0	1	1	1	1	1
v=	0	0	0	0	1	1	0	0	0	0

Example: Know Thy Customer

Example

- KTC is a Financial Advising Company
- Wishes to perform market segmentation of its clients. Thus each client is a case or “point”
- Variables: Age, Gender, Annual Income, Marital Status, Number of Children, Car Loan or not, Home Mortgage or not
- What type is each variable?
- How should similarity between cases be measured in relation to values of
  - Numerical variables?
  - Categorical variables?

Example

- Corresponding to each customer, KTC has a vector of measurements on seven customer variables: Age, Female, Income, Married, Children, Car Loan, Mortgage

Age	Female	Income	Married	Children	CarLoan	Mortgage
...	...	...	...	...	...	...
61	1	57881	1	2	0	1
27	0	55539	1	0	0	1

- 61-year-old female with an annual income of $57,881, married with two children, no car loan and but a mortgage.
- 27-year-old male with annual income of $55,539, married, no children, no car loan but mortgage

Start by considering only the binary variables: Female, Married, Car Loan, Mortgage

Similarity for Categorical Variables

Female	Married	CarLoan	Mortgage	case
1	1	0	1	u
0	1	0	1	v
1	1	1	1	w

Calculate the similarity between

- u and v
- v and w
- w and u

Using Jaccard and Matching

Similarity for categorical variables

- Binary variables in KTC data:
  - Gender
  - Married or not
  - Car loan or not
  - Home mortgage or not
- Given these variables, would you use Matching or Jaccard?

Example

- Corresponding to each customer, KTC has a vector of measurements on seven customer variables: Age, Female, Income, Married, Children, Car Loan, Mortgage

Age	Female	Income	Married	Children	CarLoan	Mortgage
...	...	...	...	...	...	...
61	1	57881	1	2	0	1
27	0	55539	1	0	0	1

Similarity for contionuous variables?

Age	Income	Children
61	57881	2
27	55539	0

$$d_{1,2} = \sqrt{(61-27)^2 + (57881-55539)^2 + (2-0)^2}$$

$$d_{1,2} = \sqrt{1156 + 5484964 + 4}$$

$$d_{1,2} = \sqrt{5486124}$$

$$d_{1,2} = 2342$$

What is the problem with this?

Similarity for contionuous variables?

Age	Income	Children
61	57881	2
27	55539	0

What about Manhattan Distance?

$$d_{1,2} = |61 - 27| + |57881 - 55539| + |2 - 0|$$

$$d_{1,2} = 34 + 2342 + 2$$

$$d_{1,2} = 2371$$

The difference of scale are less, but the variables age and number of children are still swamped by income.

Scaling distances

- Euclidean and Manhattan distance are highly influenced by the scale on which variables are measured.
  - Therefore, it is common to standardize the units of each variable of each observation u;
- Example: $Income_u$, the value of variable Income in observation u, is replaced with its z-score.
- The conversion to z-scores also makes it easier to identify outlier measurements, which can distort the Euclidean distance between observations.
- Recall: What are z-scores?

Z- scores

-Consider values of a variable $x = x_1, x_2, \ldots, x_n$

- Suppose the mean and standard deviation of this set of value is given by $\bar{x}$ and s
- Then the Z- score corresponding to x_i is given by

$$z_i = \frac{x_i - \bar{x}}{s}$$

- If certain variables are known to be more important, deliberately choose unequal weighting

Variables of Mixed Types

- A database may contain all types of variables
- A weighted formula is used to combine the effects of numerical and categorical variables [See the gower metric in the command daisy in the package cluster]
- We focus on cluster analysis using just one type of variable

Limitations of Hierarchical Clustering

- Need to calculate and store an nxn distance matrix, infeasible for large datasets
- Low stability: change a few records and a very different result may be obtained
- Influenced by choice of metric, especially when using “average”
- Sensitive to outliers
- Only one pass through the data: misallocations can never be fixed

Interpreting, Profiling, & Validating Clusters

- The clusters centroid, a mean profile of the cluster on each clustering variable, is particularly useful in the interpretation stage
  - Interpretation involves examining the distinguishing characteristics of each cluster's profile and identifying substantial differences between clusters
  - If there is not substantial difference between clusters, then other cluster solutions should be examined
  - The cluster centroid should also be assessed for correspondence with the researcher's prior expectations based on theory or practical experience

Interpreting, Profiling, & Validating Clusters

- Validation is essential in cluster analysis since the clusters are descriptive of structure and require additional support for their relevance:
  - Cross-validation empirically validates a cluster solution by creating two sub-samples (randomly splitting the sample) and then comparing the two cluster solutions for consistency with respect to number of clusters and the cluster profiles
  - Validation is also achieved by examining differences on variables not included in the cluster analysis but for which there is a theoretical and relevant reason to expect variation across the clusters

How to in R

An example using the Palmer Penguins

- R packages : ggdendro, ade4

```
library(tidytuesdayR)
penguins<- tidytuesdayR::tt_load(2020, week = 31)$penguins

##
##     Downloading file 1 of 2: `penguins.csv`
##     Downloading file 2 of 2: `penguins_raw.csv`

head(penguins)

## # A tibble: 6 x 8
##   species island bill_length_mm bill_depth_mm flipper_length_~ body_mass_g se
##   <chr>   <chr>           <dbl>         <dbl>            <dbl>       <dbl> <c
## 1 Adelie  Torge~           39.1          18.7              181        3750 ma
## 2 Adelie  Torge~           39.5          17.4              186        3800 fe
## 3 Adelie  Torge~           40.3          18                195        3250 fe
## 4 Adelie  Torge~           NA            NA                 NA          NA <N
```

An example using the Palmer Penguins

- There are two nominal variables. Let's turn them into binary variables

```
library(tidyverse)
penguins <- penguins %>%
  mutate(torgersen_yn = ifelse(island == "Torgersen",1,0),
         female_yn = ifelse(sex == "female",1,0))
```

- Let's also only select the first 50 observations

```
penguins_min <- penguins[1:50,]
```

Focus on binary variables

- Calculate the distance (method 1 is Jaccard, 2 is Matching). Which would be best to use in our case?

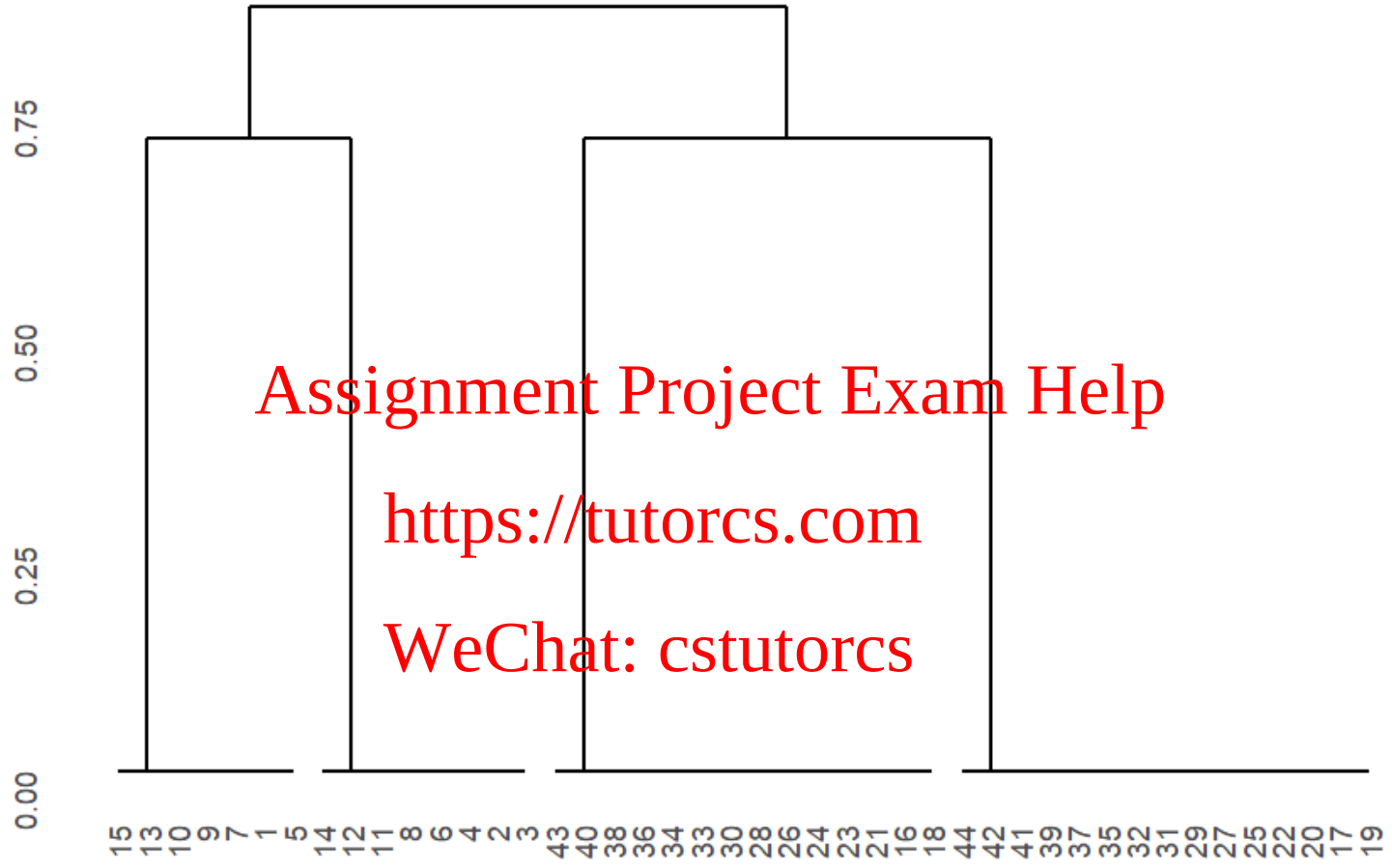
```
library(ade4)
penguins_binary <- penguins_min %>%
    select("torgersen_yn","female_yn")%>%
    drop_na() # Removes any NA entries
distance_binary <- dist.binary(penguins_binary, method = 2)
```


- Calculate clusters using average linkage (other options are possible!)

```
cluster_binary<-hclust(distance_binary, method = "average")
```

Create a dendrogram

```
library(ggdendro)
ggdendrogram(cluster_binary)
```



Calculate clusters

- Specify the number of clusters we want (k) or the height at which to cut h, and add to the data frame

```
penguins_binary <- penguins_binary %>%
  mutate(cut1_binary = cutree(cluster_binary, k= 3))
head(penguins_binary)
```



```
## # A tibble: 6 x 3
##   torgersen_yn female_yn cut1_binary
##          <dbl>     <dbl>       <int>
## 1            1         0           1
## 2            1         1           1
## 3            1         1           1
## 4            1         1           1
## 5            1         0           1
## 6            1         1           1
```

Focus on continuous variables

- Use bill length and body weight for the cluster
- Use `scale` function to scale the variables

```
penguins_cont <- penguins_min %>%
    select("bill_length_mm","body_mass_g")%>%
    drop_na() %>% # Removes any NA entries
    scale() %>%
  data.frame()
distance_cont <- dist(penguins_cont, method = "euclidean")
```

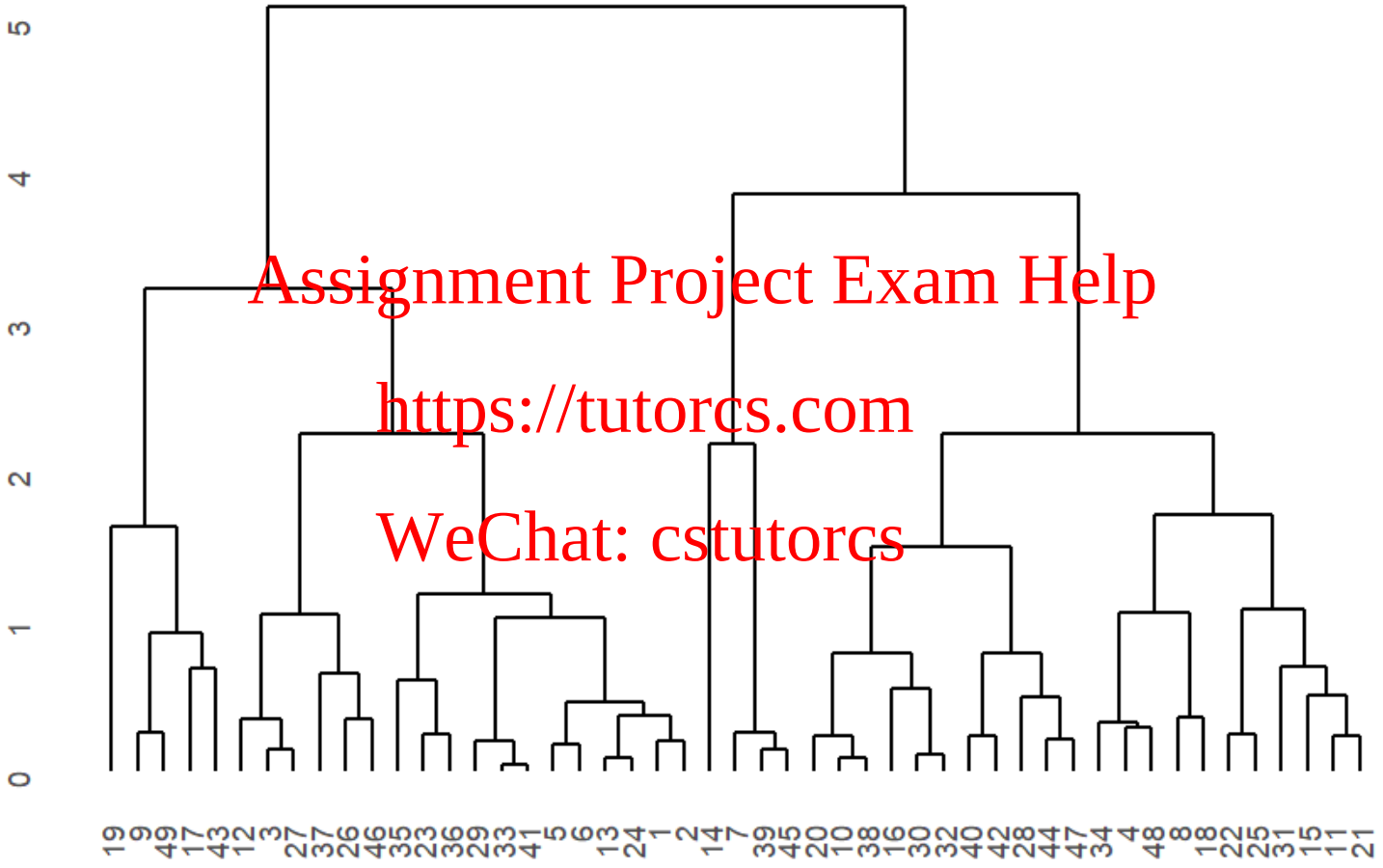
Hierachical clustering

- Run complete clustering

```
cluster_cont <- hclust(distance_cont, method = "complete")
```

- Plot the dendrogram

```
ggdendrogram(cluster_cont)
```



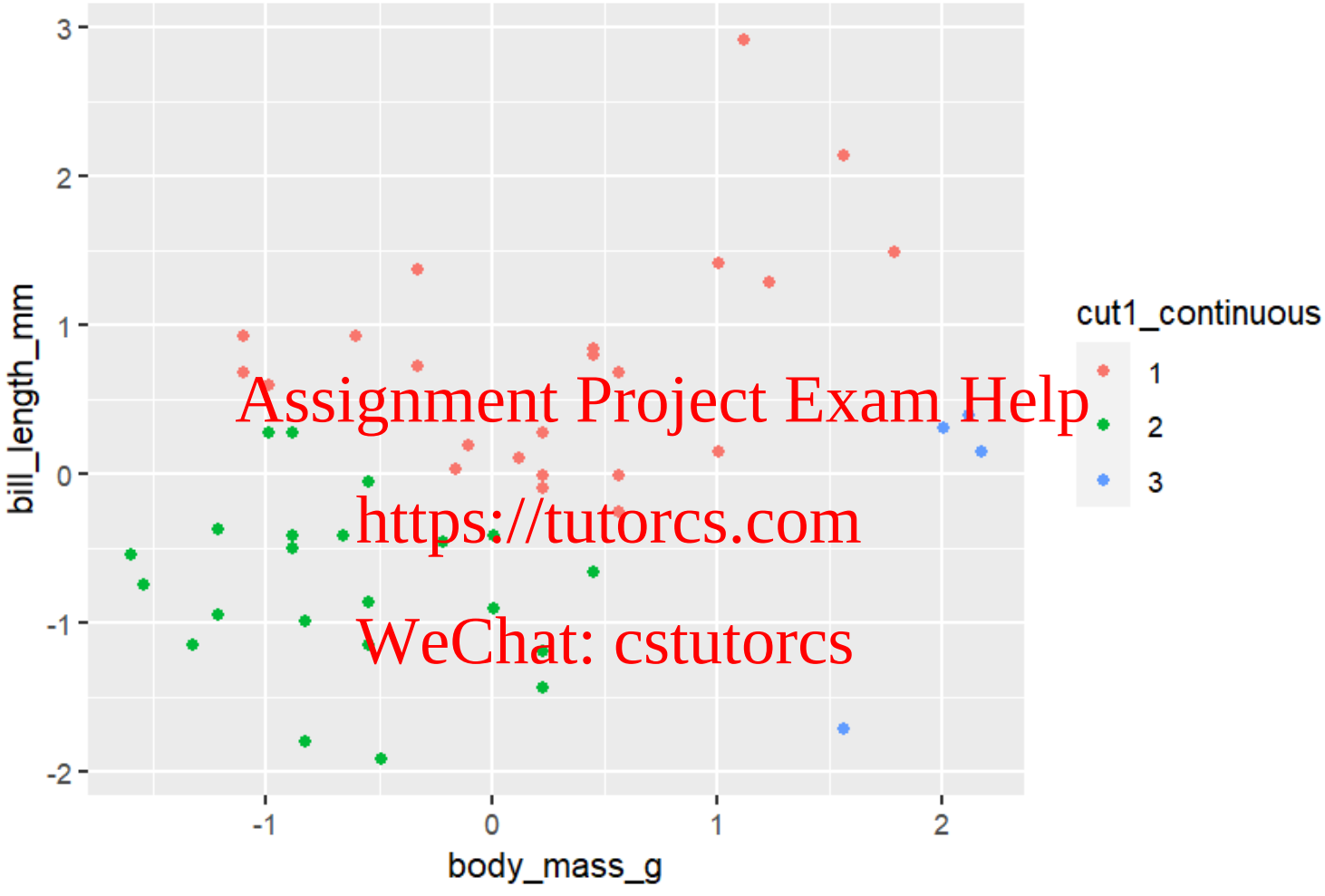
What about the clusters?

- Choose 3 clusters.

```
penguins_cont <- penguins_cont %>%
  mutate(cut1_continuous = factor(cutree(cluster_cont, k = 3)))
```

Plot the clusters

```
library(ggplot2)
ggplot(penguins_cont, aes(x=body_mass_g, y= bill_length_mm,
                          colour = cut1_continuous))+
  geom_point()
```



A cautionary example

- What should we call the different categories?
  - Larger animal, larger bill = call this group "pelicans"
  - Larger animal, small bill = call this group "chickens"
  - Smaller animal, small bill = call this group "sparrows"
- But the group names don't change the birds from being all penguins.
- Where else have we seen the group names being overinterpreted?
  - e.g., personality tests look at which items correlate strongly together to form different personality "types". An example of this is the Eagle - Dove - Owl - Peacock test.

Summary

- There are two types of distances to pay attention to in hierarchical clustering.
  - The distance between points
  - The distance between clusters
- The choice of how to measure these distances has implications for the shape and potential validity of the clusters.
- Validation and interpretation is very important, and must be completed carefully!

Lecturer: *Lauren Kennedy*

Department of Econometrics and Business Statistics

lauren.kennedy1@monash.edu

Week 7