Clustering

K-Means Clustering

Divisive Methods

> “...tending to cause disagreements that separate people into opposing groups…”

This uses nonparametric algorithms:

- the number and nature of the parameters are flexible and not fixed in advance
- the number of clusters unknown

Divisive Methods

We split the data into a small number of clusters

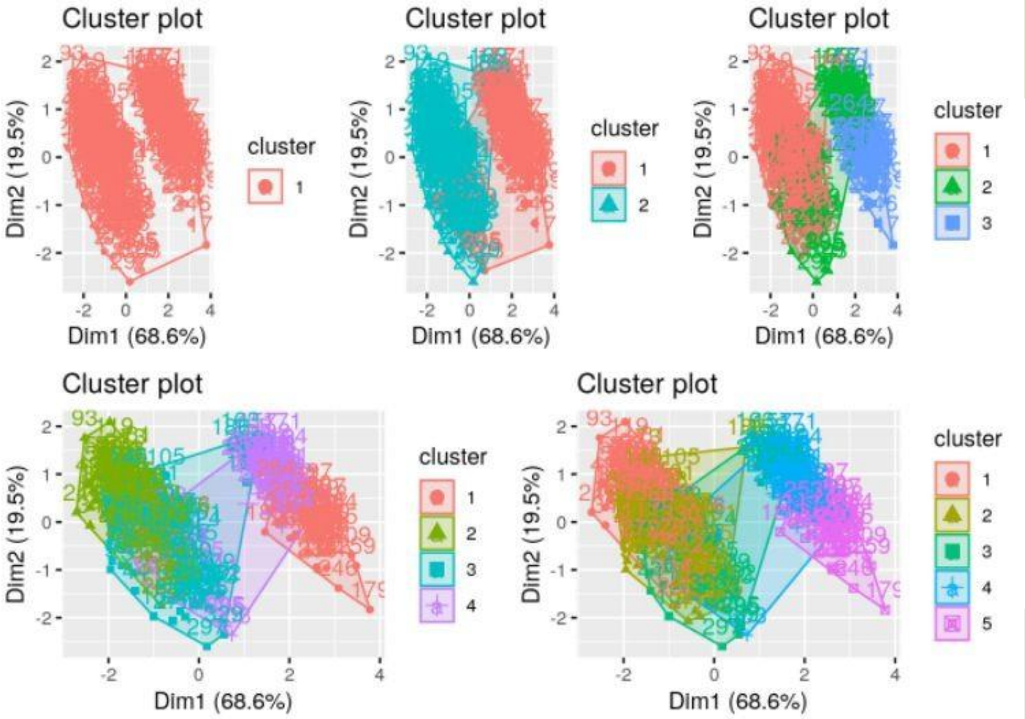
1. an initial allocation of seeds (randomly picked initial cluster centres)
2. allocation of points to closest centre
3. reallocation of each point (centre adjusted)
4. Repeat (until no changes or iterations max)

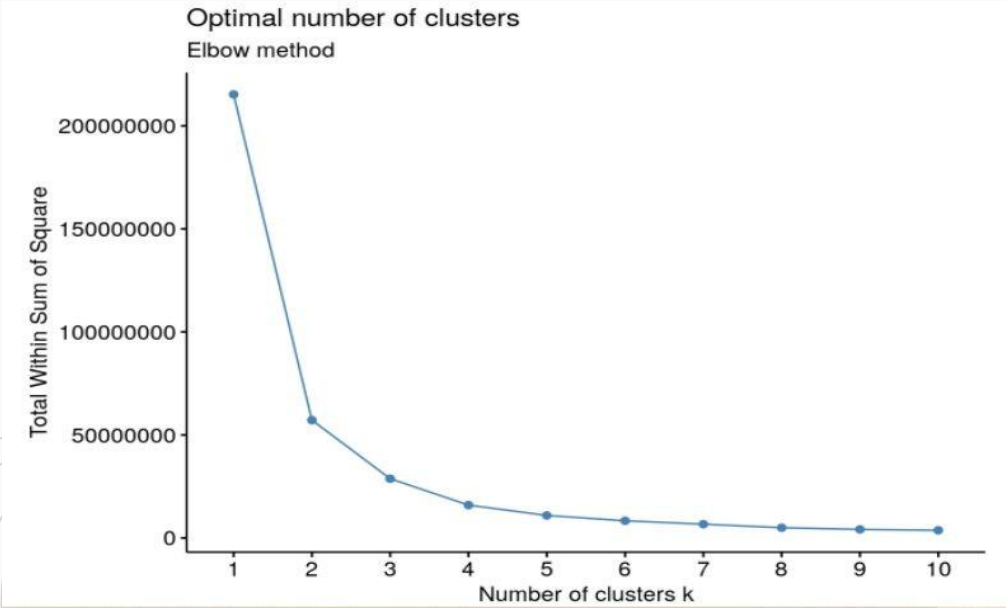
How to Choose the Number of Clusters

1. Plot it and see…
2. Elbow Method
3. Silhouette Method
4. Gap Method
5. Compare and Choose!

1. Plot it and see...

```
library(factoextra                    )
)                                     )
p  <- fviz_cluster(k1, data =         )
1                             df)
p  <- fviz_cluster(k2, data =
2                             df)
p  <- fviz_cluster(k3, data = df
3
p  <- fviz_cluster(k4, data = df
4
p  <- fviz_cluster(k5, data = df
5
```



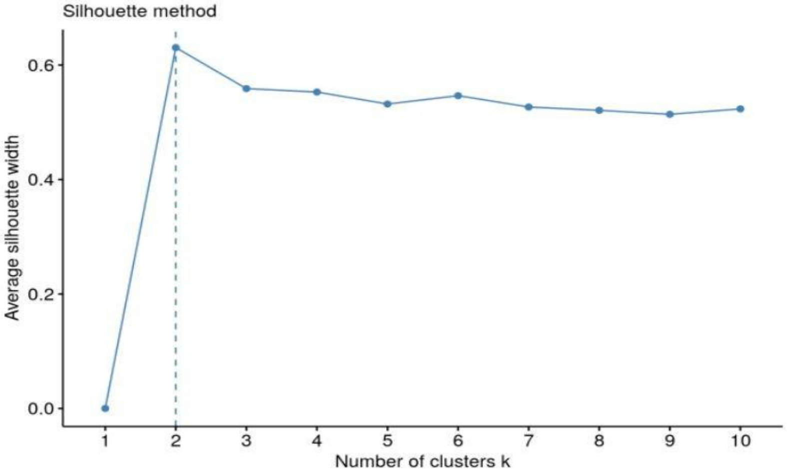


2. Elbow Method…

Looking for when the *marginal total within sum of squares* for an additional cluster begins to decrease at a linear rate

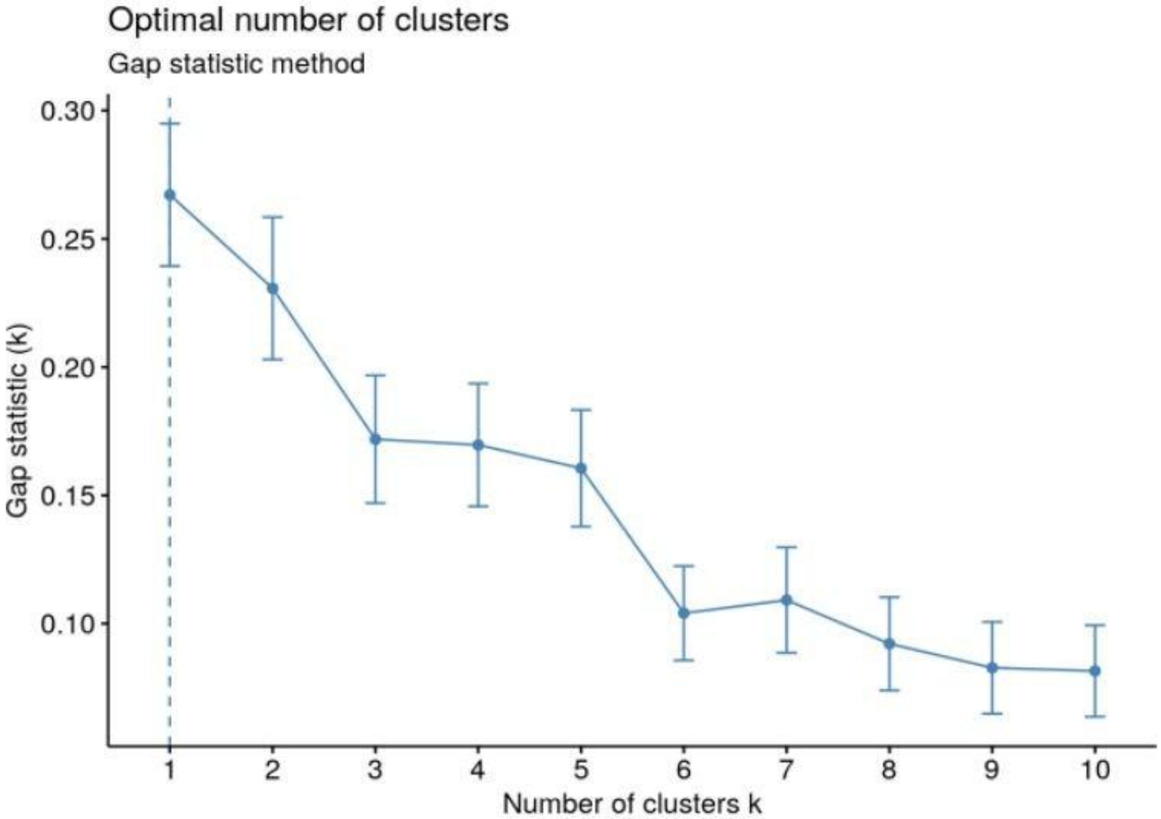
3. Silhouette Method

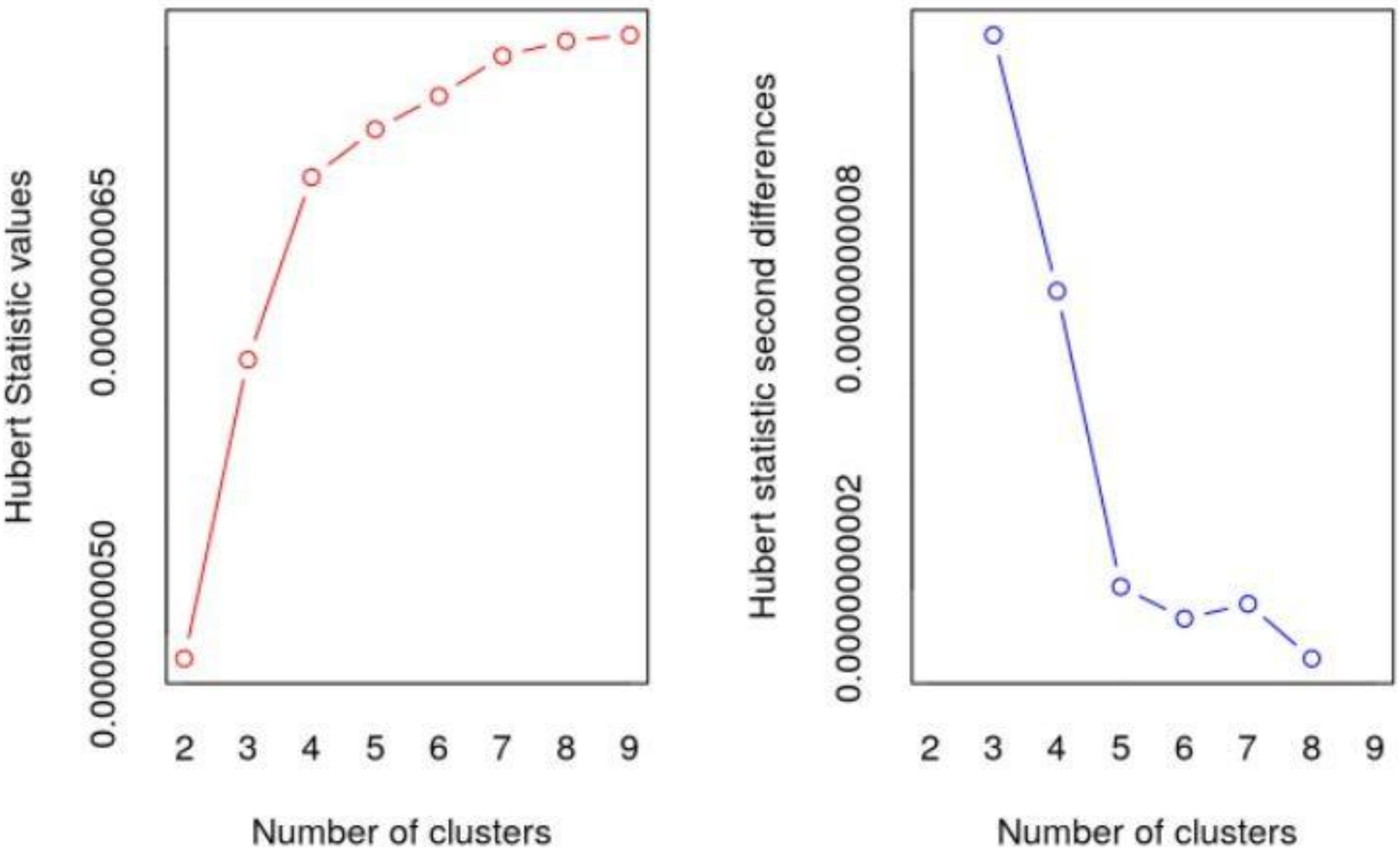
Looks at the mean within distance and the mean nearest-cluster between distance for each sample

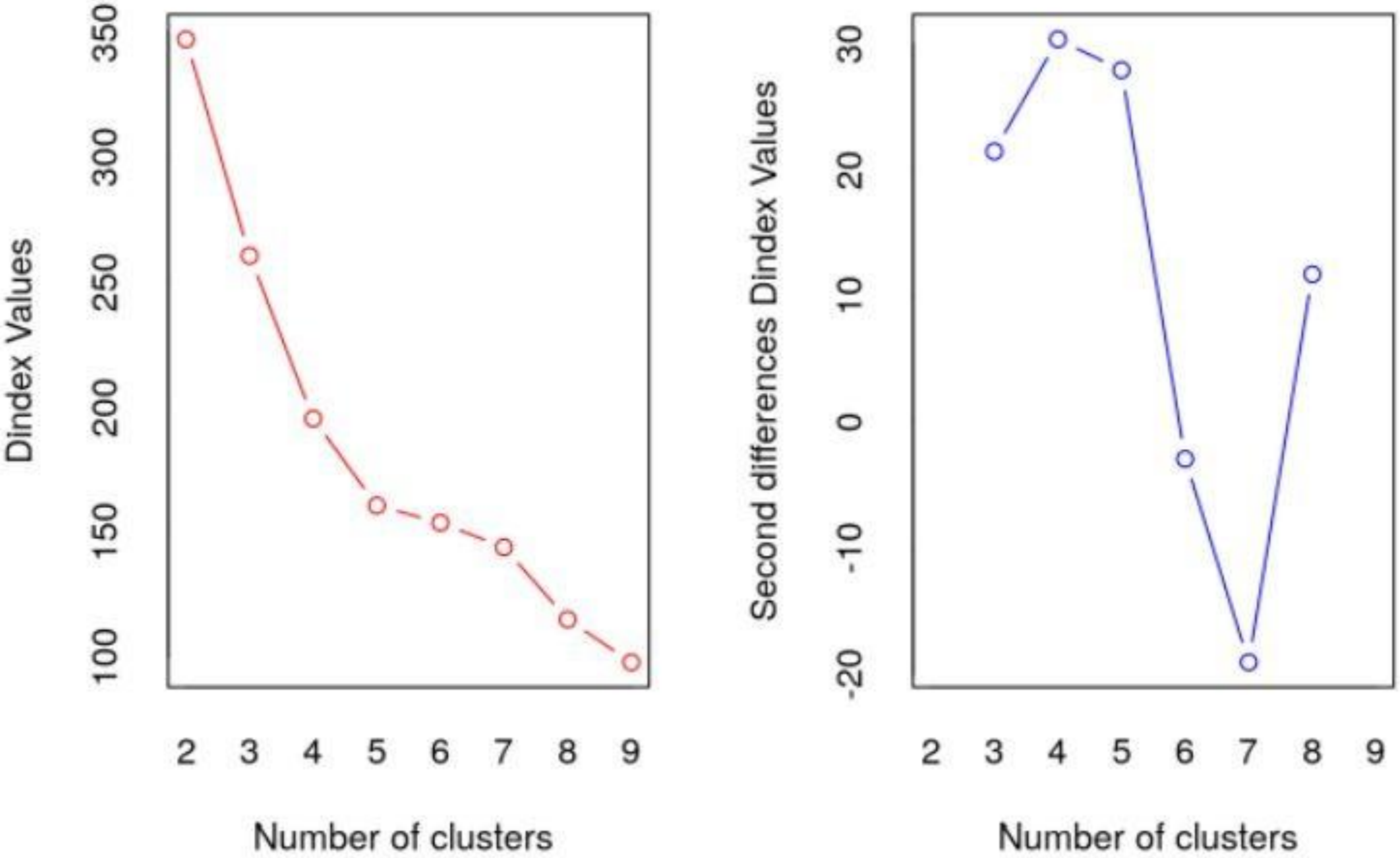


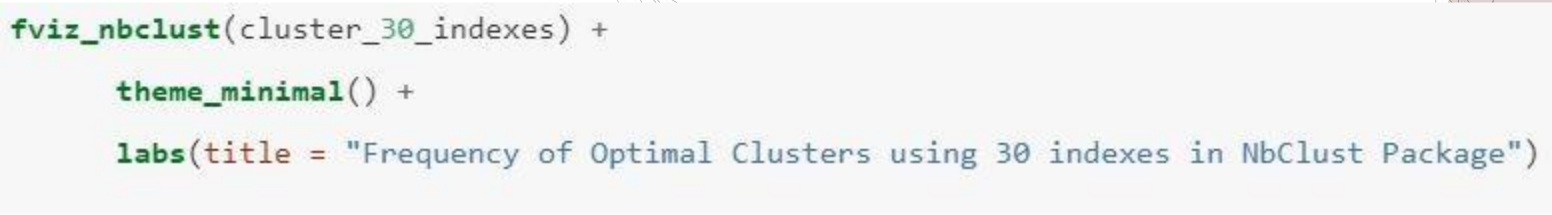
4. Gap Method

Looking for which organisation of clusters gives the biggest ‘gap’

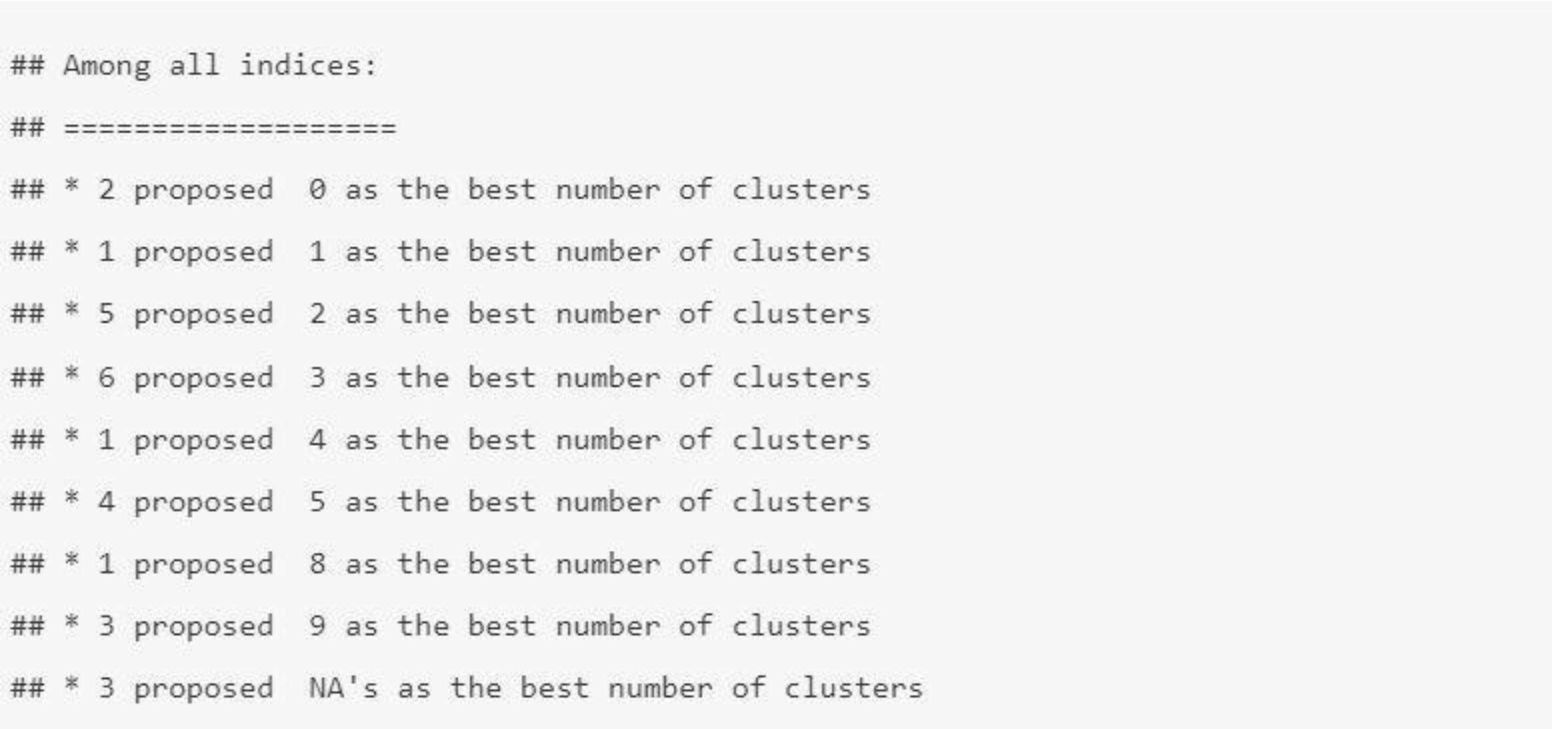



Hubert Statistic values
0.0000000050
0.0000000065
2
3
4
5
6
7
8
9
Number of clusters
Hubert statistic second differences
0.0000000002
0.0000000008
2
3
4
5
6
7
8
9
Number of clusters


Dindex Values
350
300
250
200
150
100
2
3
4
5
6
7
8
9
Number of clusters
Second differences Dindex Values
30
20
10
0
-10
-20
2
3
4
5
6
7
8
9
Number of clusters



```r
fviz_nbclust(cluster_30_indexes) +
      theme_minimal() +
      labs(title = "Frequency of Optimal Clusters using 30 indexes in NbClust Package")
```



```
## Among all indices:
## ===================
## * 2 proposed  0 as the best number of clusters
## * 1 proposed  1 as the best number of clusters
## * 5 proposed  2 as the best number of clusters
## * 6 proposed  3 as the best number of clusters
## * 1 proposed  4 as the best number of clusters
## * 4 proposed  5 as the best number of clusters
## * 1 proposed  8 as the best number of clusters
## * 3 proposed  9 as the best number of clusters
## * 3 proposed  NA's as the best number of clusters
```

Hierarchical Agglomerati ve Clustering

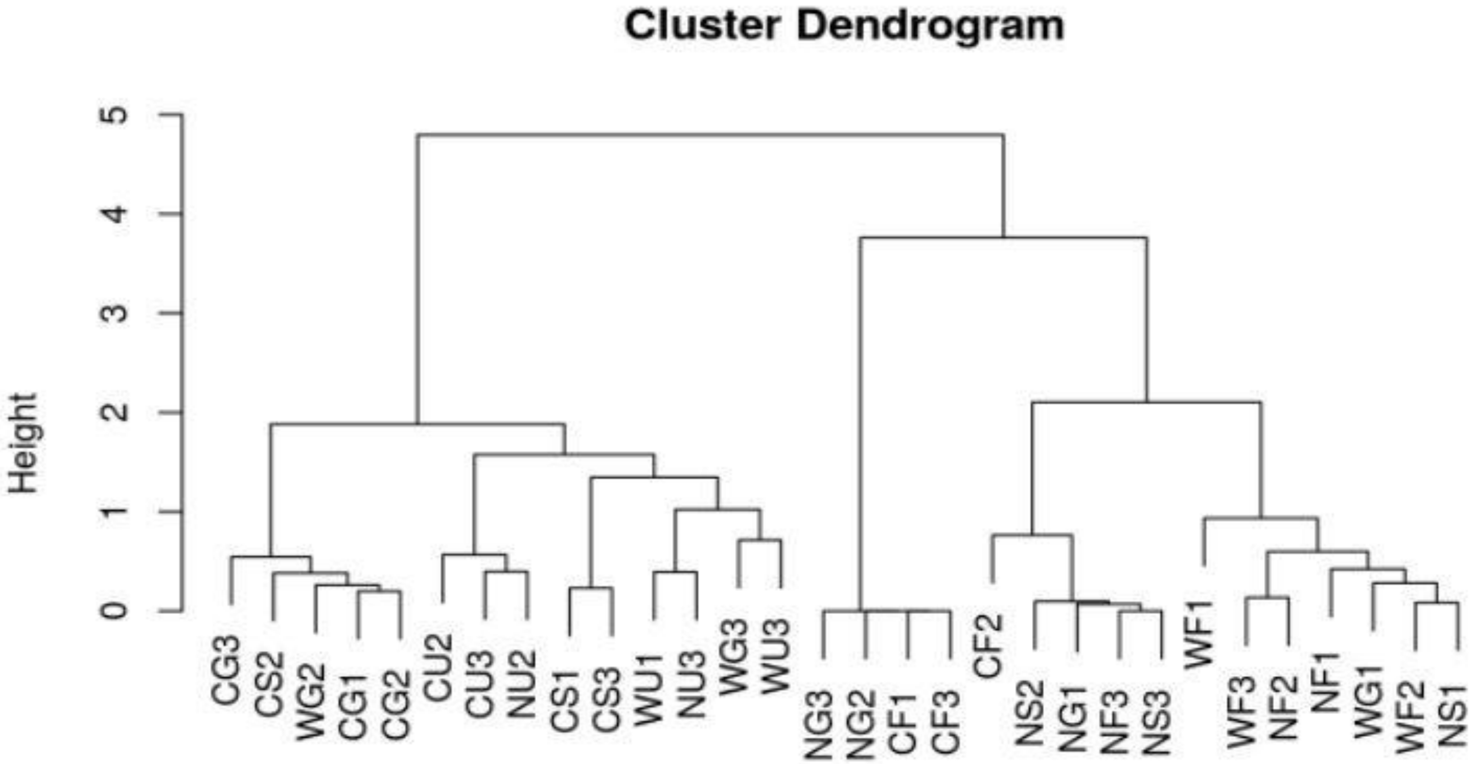
Hierarchical Agglomerative Clustering

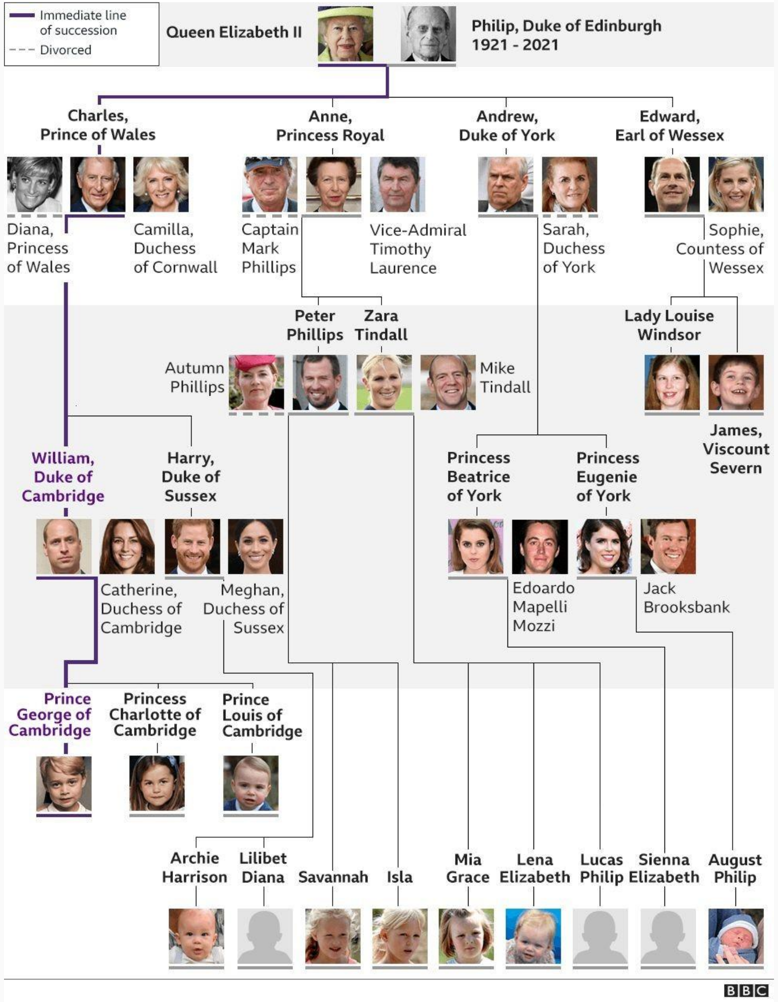
Hierarchical ⇒ Ranking, ordered

Agglomeration ⇒ "...a large group of many different

things collected or brought together..."

https://dictionary.cambridge.org/dictionary/english/agglomeration


Cluster Dendrogram
Height
0
1
2
3
4
5
CG3
CS2
WG2
CG1
CG2
CU2
CU3
NU2
CS1
CS3
WU1
NU3
WG3
WU3
NG3
NG2
CF1
CF3
CF2
NS2
NG1
NF3
NS3
WF1
WF3
NF2
NF1
WG1
WF2
NS1



https://www.bbc.com/news/ uk-23272491

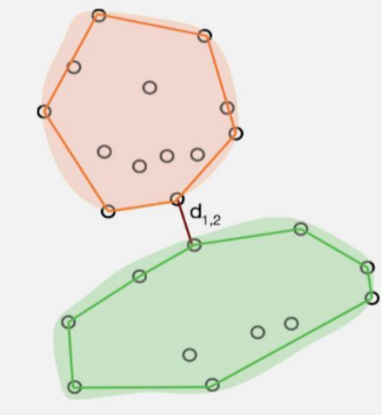
- Fuse zero distance observations
- Fuse until two clusters are close enough to fuse
- Fuse until all the clusters have been fused into one big one

Nearest-Neighbour (single linkage) Clustering

The smallest distance between any two points in the two clusters

Chaining often leads to uninformative dendrograms

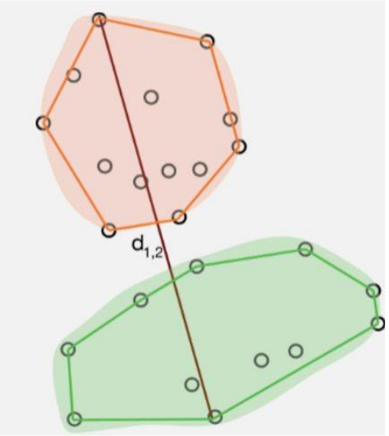
Adaptive for well separated data; robust to the choice of measure



Farthest-Neighbour (complete linkage) Clustering

The largest distance between any two points in the two clusters

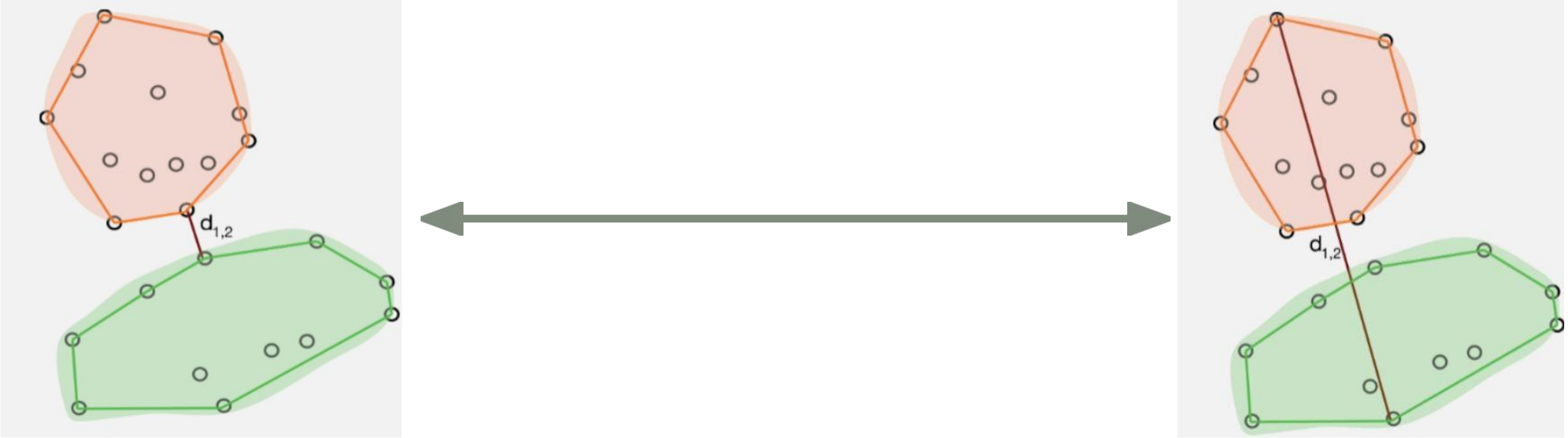
Can be sensitive to tied distances



Clusters are often compact, spherical and well defined; robust to a certain amount of measurement error and choice of distance

Group Average Linkage (UPGMA)

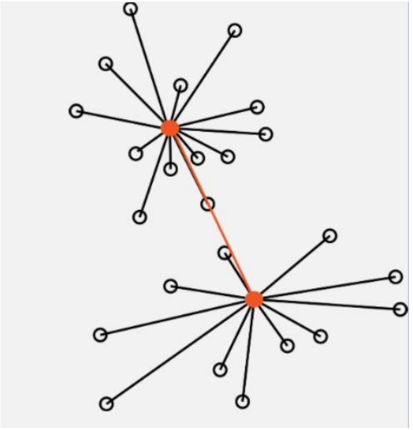
The distance between two clusters is the average of the distances between the members of the two groups



Ward's Method

Minimises the increase in total sum of squared distances within the clusters

A bad start can mean it will never reach the global optimum for a given number of clusters; tends to form spherical clusters of equal size



It is not allowed to swap points between clusters

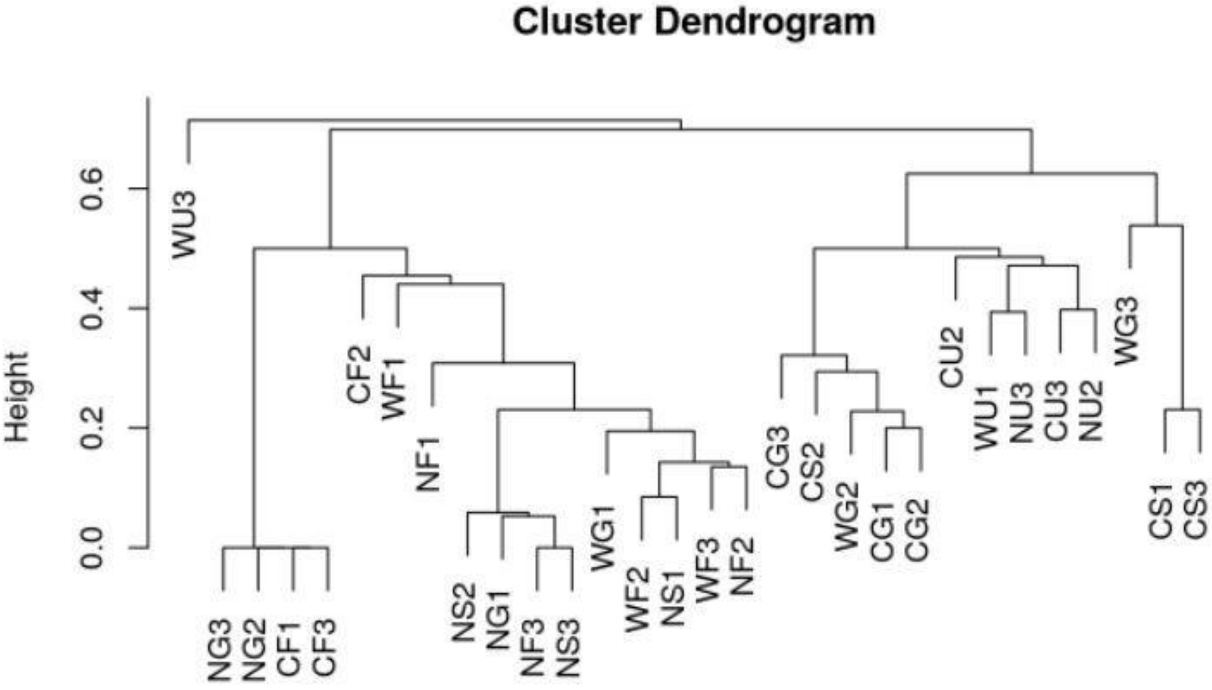
Method	Pros	Cons
Single linkage	number of clusters	comb-like trees.
Complete linkage	compact clusters	one obs. can alter groups
Average linkage	similar size and variance	not robust
Centroid	robust to outliers	smaller number of clusters
Ward	minimising an inertia	clusters small if high variability

[code along `ants`, script below]

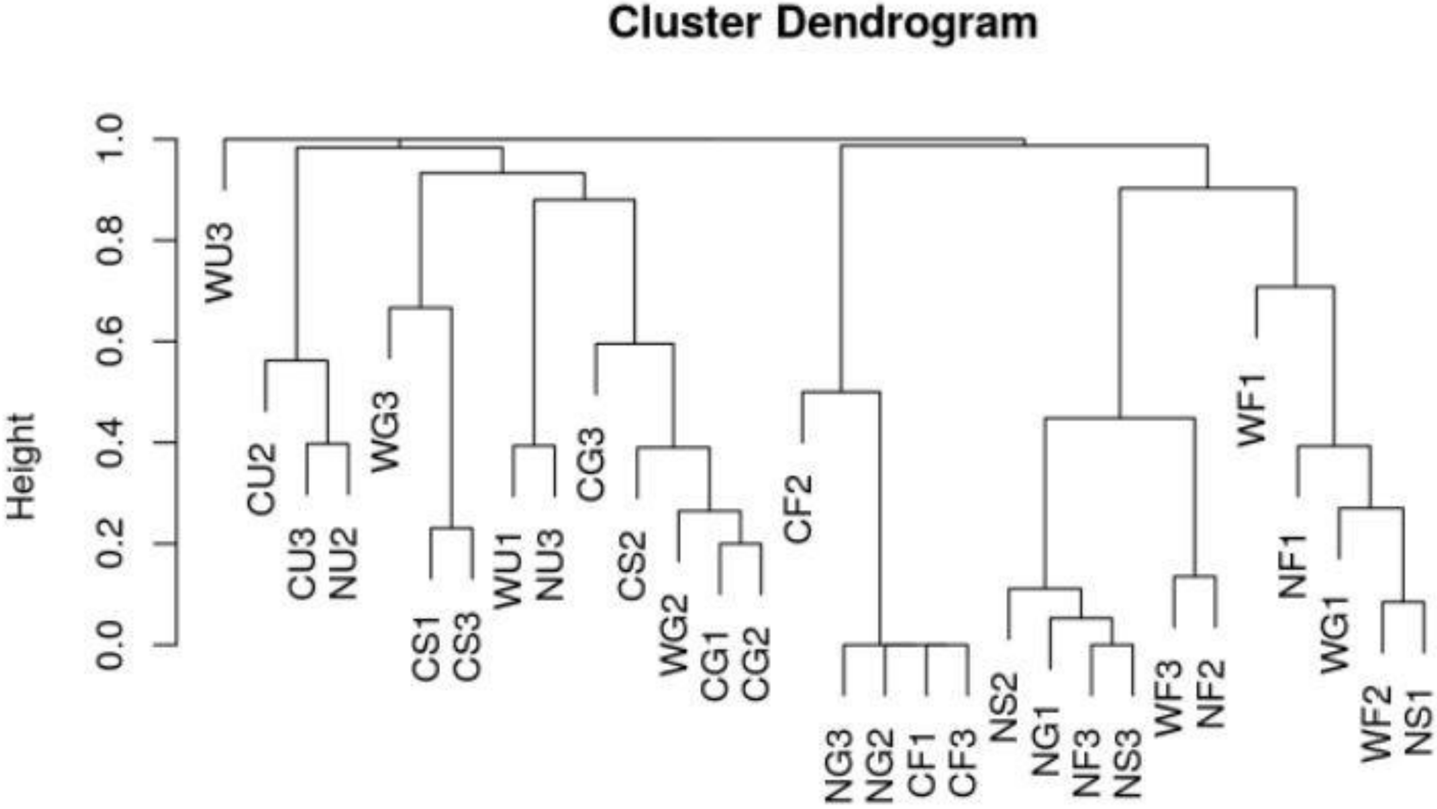
Data were collected on the distribution of ant species at 30 sites across the Auckland region using pitfall traps. Twenty pitfall traps at each site were left open for ten days and the number of individuals captured counted for the four most abundant species: *Nylanderia spp*, *Pheidole rugosula*, *Tetramorium grassii*, and *Pachycondyla sp*.

```
glimpse(ants)
## Rows: 30
## Columns: 8
## $ Location <chr> "West", "West", "West", "West", "West", "West", "West", "West…
## $ Habitat  <chr> "Forest", "Grass", "Urban", "Forest", "Grass", "Forest", "Gra…
## $ Month    <dbl> 1, 1, 1, 2, 2, 3, 3, 3, 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 3, 1, 1…
## $ Site     <chr> "WF1", "WG1", "WU1", "WF2", "WG2", "WF3", "WG3", "WU3", "CF1"…
## $ Nyl      <dbl> 0, 0, 3, 0, 5, 0, 0, 0, 0, 0, 1, 0, 0, 22, 15, 0, 0, 10, 2, 0…
## $ Phe      <dbl> 0, 2, 7, 0, 0, 0, 3, 1, 0, 3, 0, 0, 7, 109, 1, 0, 13, 47, 0, …
## $ Tet      <dbl> 0, 7, 0, 0, 25, 0, 2, 0, 0, 22, 5, 0, 30, 54, 35, 0, 14, 7, 4…
## $ Pac      <dbl> 157, 37, 0, 31, 0, 21, 1, 0, 1, 2, 1, 3, 1, 0, 4, 1, 2, 0, 0,…
```

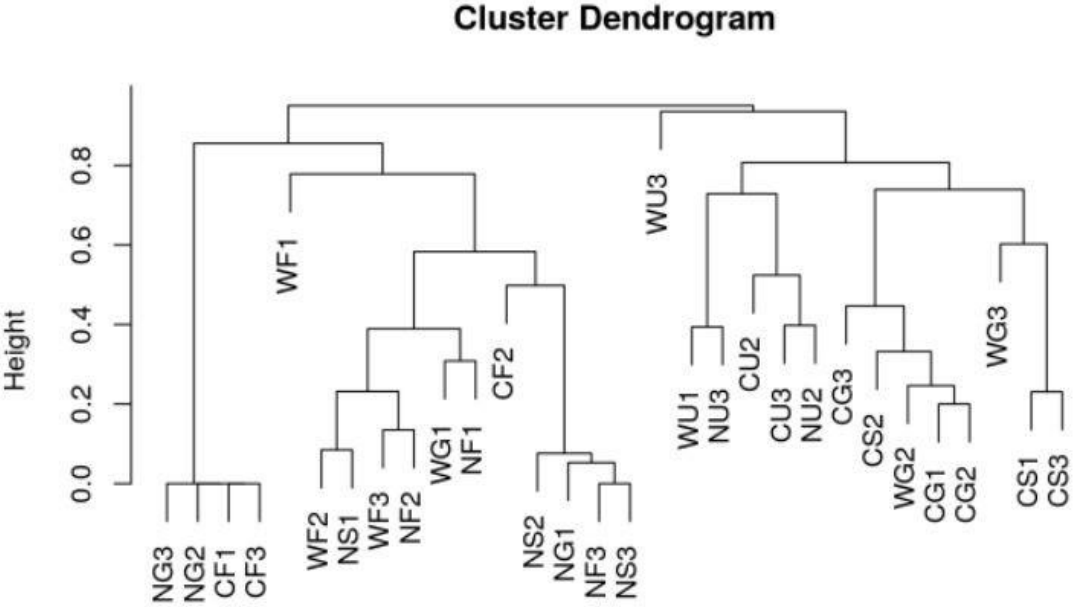
Nearest-Neighbour Clustering



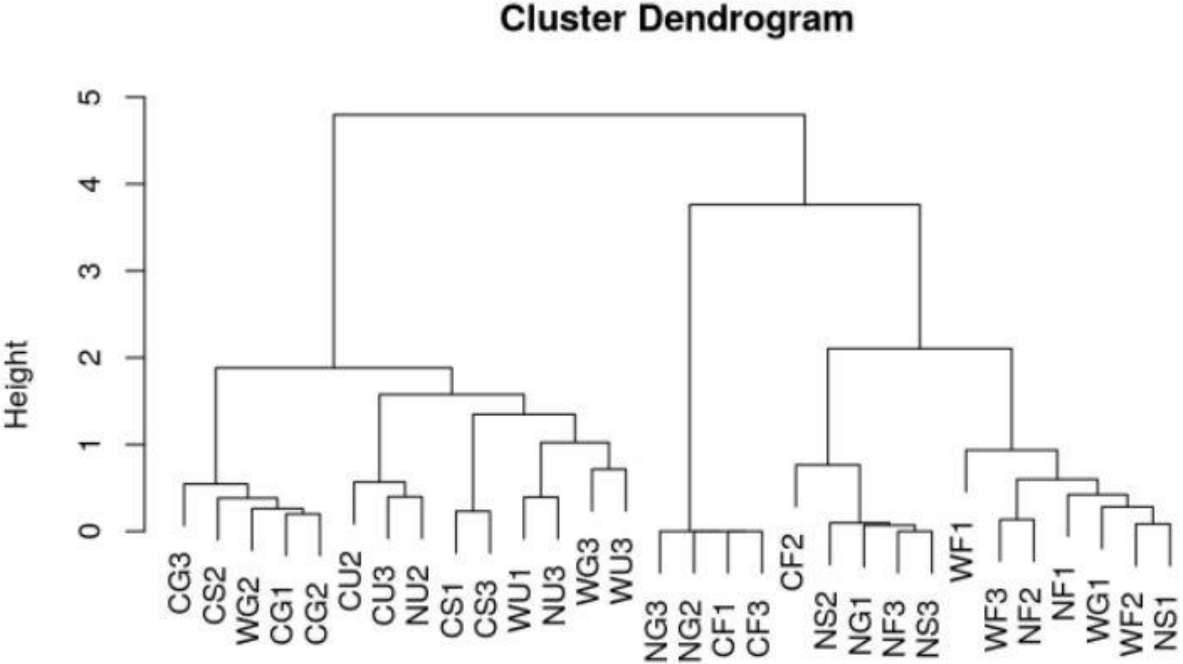
Farthest-Neighbour Clustering



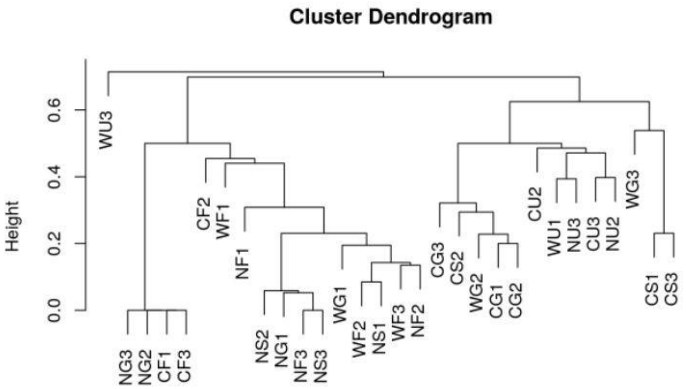
Group Average Linkage (UPGMA)



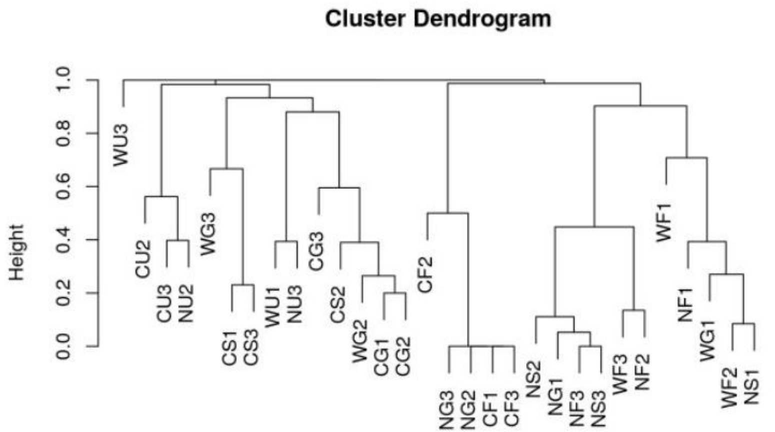
Ward's Method



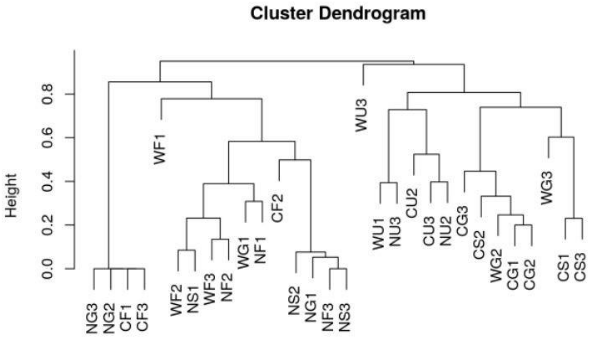
Nearest Neighbour



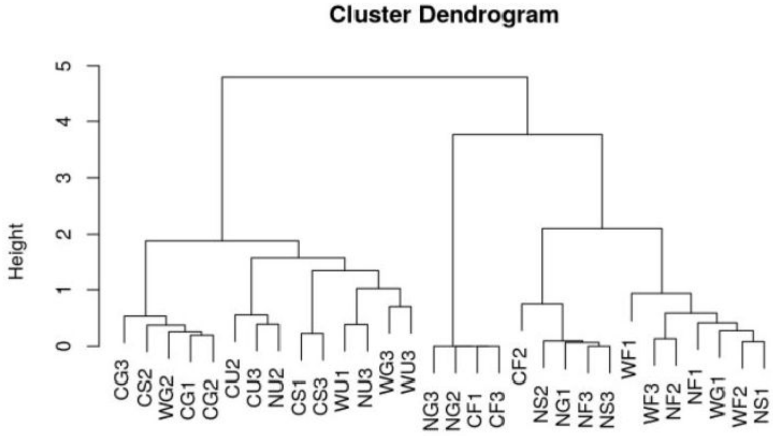
Furthest Neighbour



Average Linkage



Ward's Method



In your groups, discuss the cluster/cutree plot and the prompt at the end of your script.