CLUSTER ANALYSIS

Chapter 8

The Plan

1. Cluster Analysis – What is it?
2. Design Issues
3. Forming Clusters
4. Beer - How Many Clusters?

Cluster Analysis – What is it?

Cluster analysis is the art of finding groups in data.

We want to form groups in such a way that objects in the same group are similar to each other, whereas objects in different groups are as dissimilar as possible.

Classification of animals, plants, minerals, diseases, stars, customers, insurance policies, regions, medicine, chemistry, history, etc.

Forming Groups

Factor Analysis

Variables

Cluster Analysis: Respondents	1	2	3	4	5
1	7	5	6	4	3
2	4	5	2	2	2
3	5	4	2	5	5
4	2	4	2	4	4
5	4	5	5	1	1
6	6	2	3	6	5
7	3	4	4	5	2
8	5	6	7	2	4
9	4	6	3	6	5

Objective

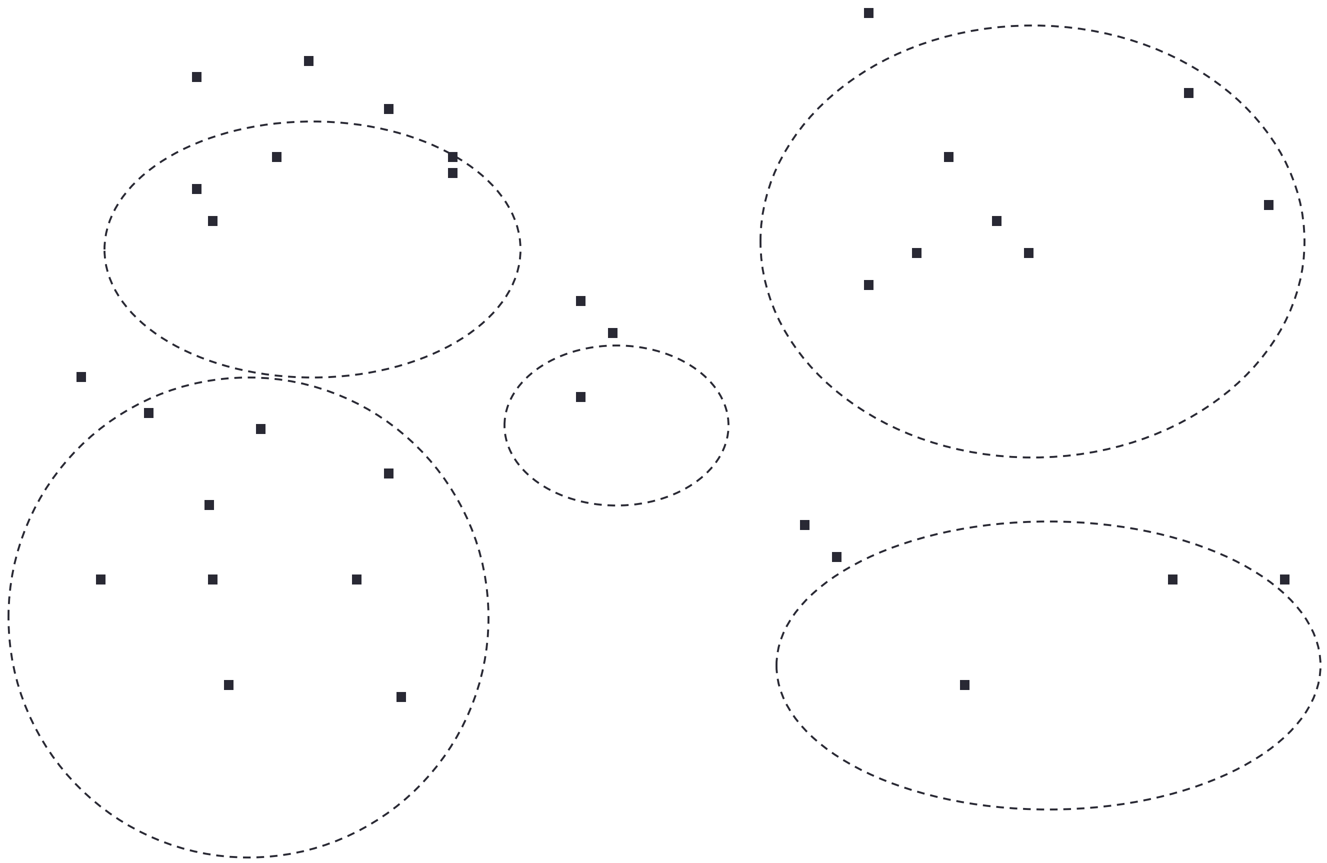
The objective is usually to uncover a structure that is already present in the data.

- E.g. to identify and characterize different market segments.
- E.g. Beer customers segmented by taste and brand.

It has no statistical basis upon which to draw inferences from a sample to a population.

Trade-off: # of Clusters vs. Homogeneity.

High # of clusters leads to high homogeneity within clusters, but parsimony is then compromised.



Design Issues

What should be done about:

- irrelevant variables?
- outliers?

- Should we standardize the variables?
- Representativeness of the sample?
- Multicollinearity?

Variable Selection

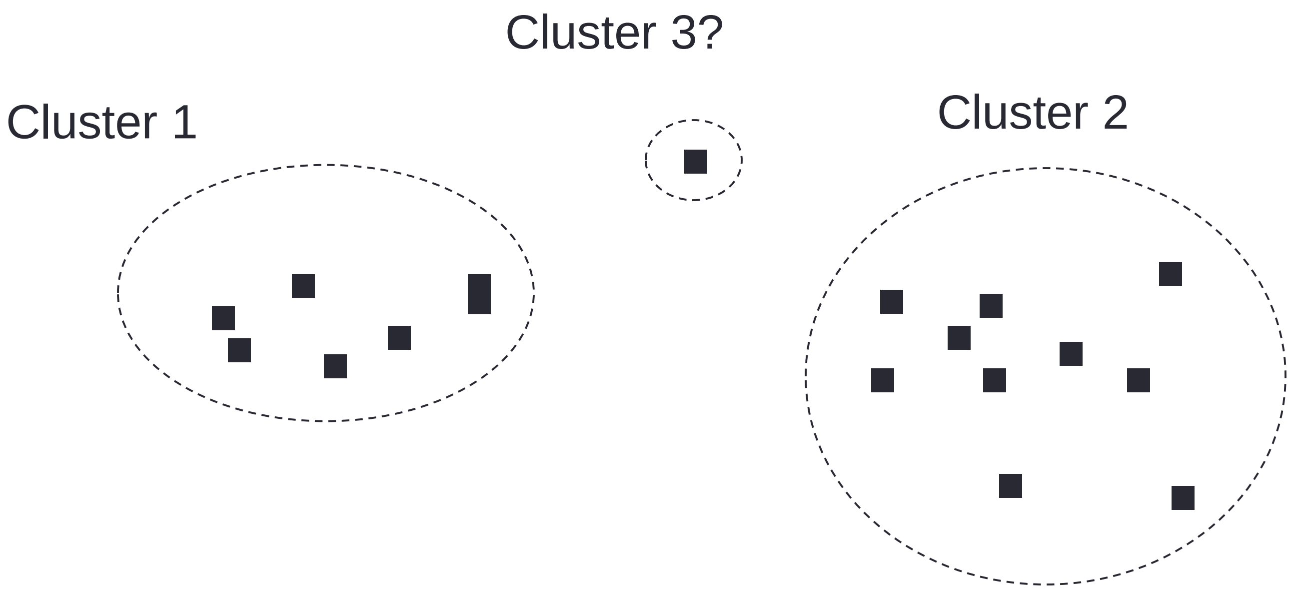
Important to select variables that:

1. Characterize the objects being clustered.
2. Relate specifically to the objectives of the cluster analysis.

The program doesn’t recognize meaningless variables, so they only contribute to clouding the results.

Outliers

Outliers have a substantial effect on the results, so they should be deleted if possible. Or, increase sample size to see if they are just a small cluster.



Standardizing the Variables

Units of measurement, like centimeters versus meters, have a substantial affect on how clusters are formed.

- Try to use variables that are measured on the same scale (e.g. 1-7).
- Consider standardizing the data before clustering.

BUT

- Standardizing removes the ”natural” weighting of the measurement unit, and can thus distort the results.

Sample

- With cluster analysis we are often try to make inferences about a population, thus having a representative sample is very important.

Multicollinearity (bad)

- Multicollinearity can introduce a hidden weighting of the variables.
  - The variables with high collinearity have a combined impact that influences the cluster method, thus biasing the results.

Forming Clusters

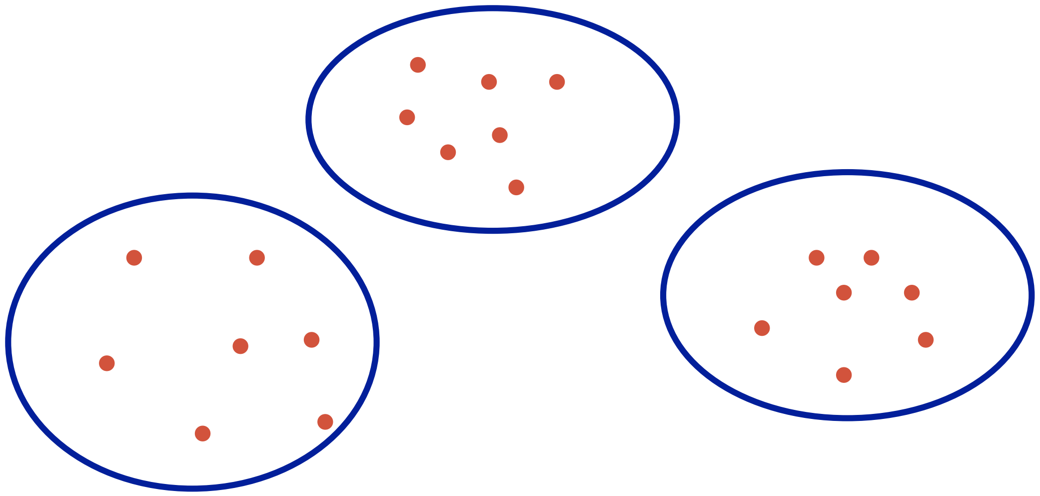
Issues:

- What method to use (clustering algorithms):
  - Partitioning around medoids
  - Agglomerative nesting
  - Divisive method
  - Monothetic algorithm
  - Fuzzy clustering

- How many clusters to form?

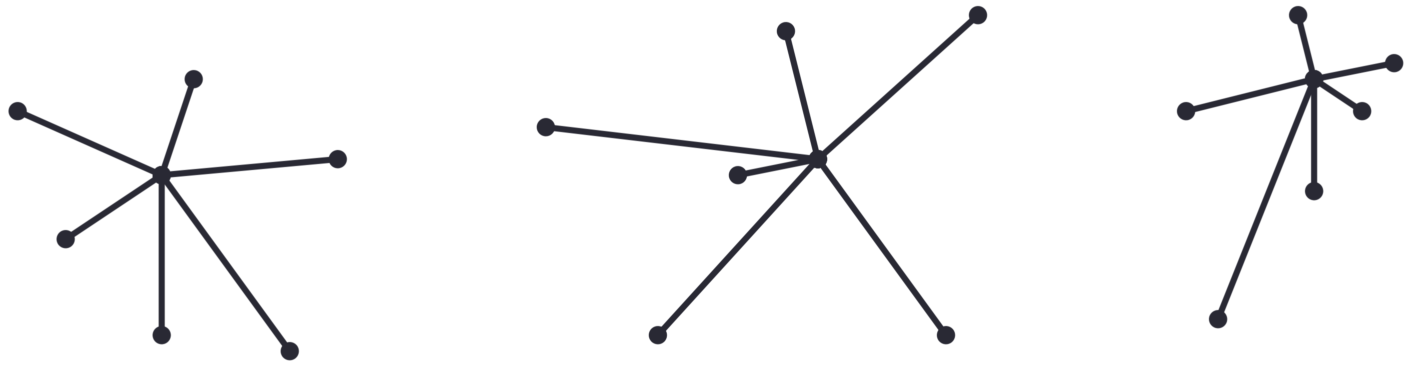
Partitioning Method

- A partitioning method constructs $k \leq n$ clusters.
- It classifies the data into k groups such that
  - each group must contain at least one object
  - each object must belong to only one group.



Partitioning Around Medoids

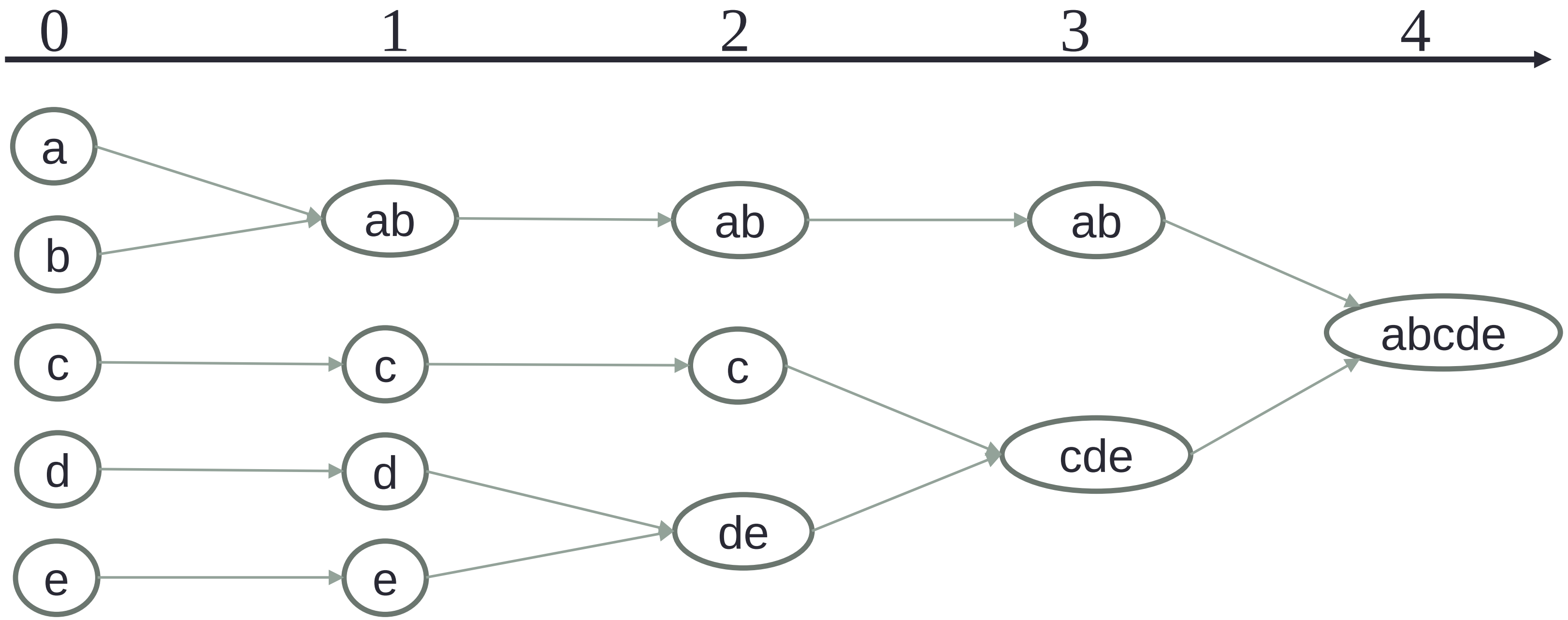
- The method selects k objects (so-called representative objects or medoids) in the data set.
- The corresponding k clusters are then found by assigning each remaining object to the nearest medoid.



Agglomerative (our example)

- Agglomerative methods start when all objects are apart.
- At each step two clusters (objects) are merged, until only one is left.
- Remember : Once an agglomerative method has joined two objects, they can never be separated.

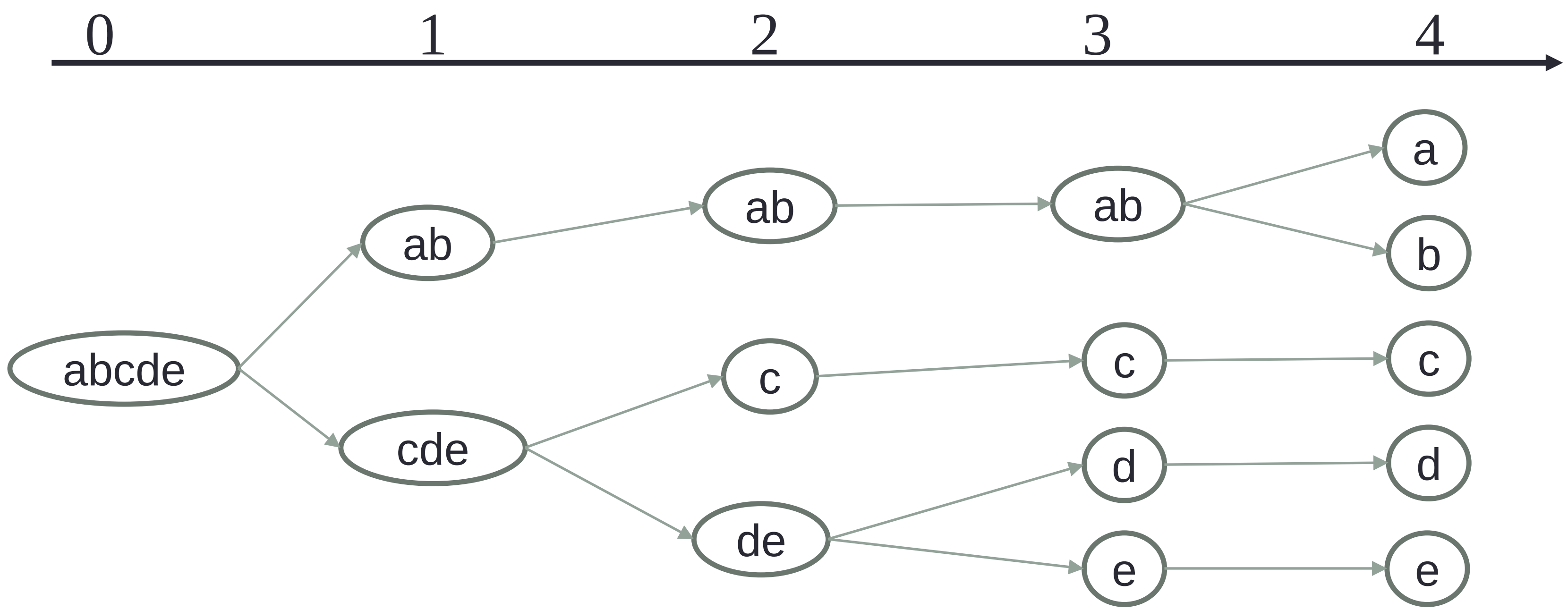
Agglomerative



Divisive

- Divisive methods start when all objects are together.
- At each step a cluster is split up, until each object is constituting one separate cluster.
- Remember : Once a divisive method has split up two objects, they can never be reunited.

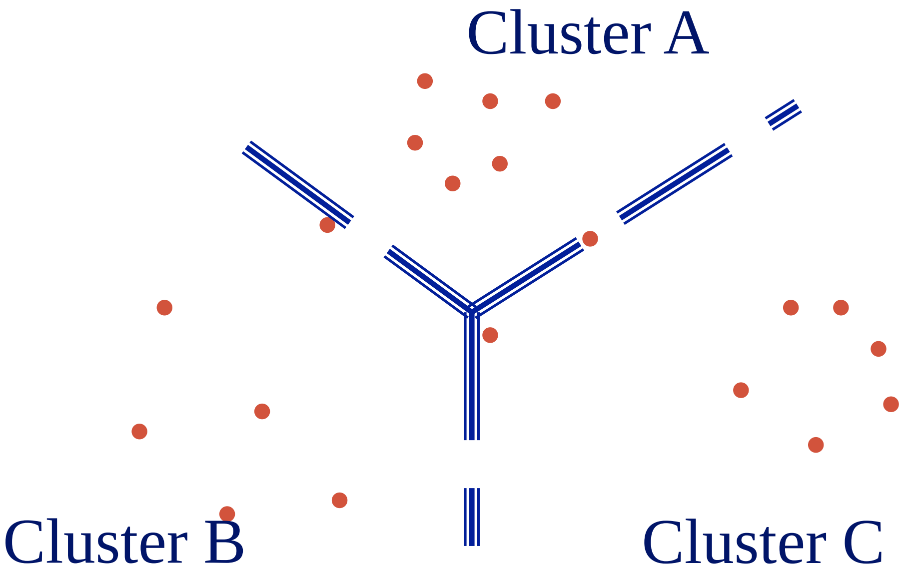
Divisive



Fuzzy Algorithm

The procedure assigns a membership coefficient to each object stating that object i belongs e.g.:

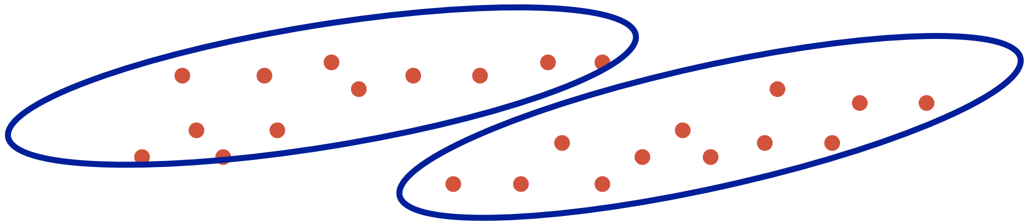
- 90 % to cluster A
- 7 % to cluster B
- 3 % to cluster C.



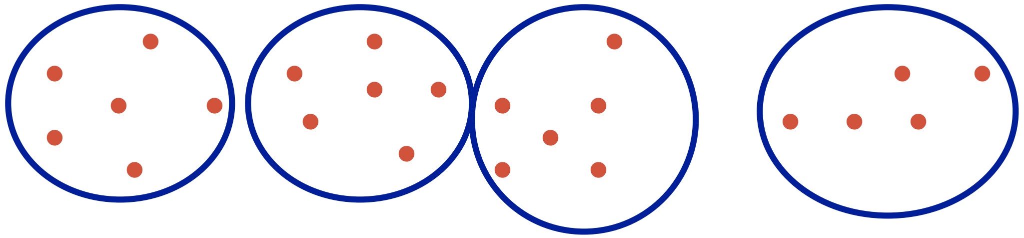
Dissimilarities Between Clusters

- The group average technique produces ball-shaped clusters.

- The nearest neighbor technique produces elongated clusters.

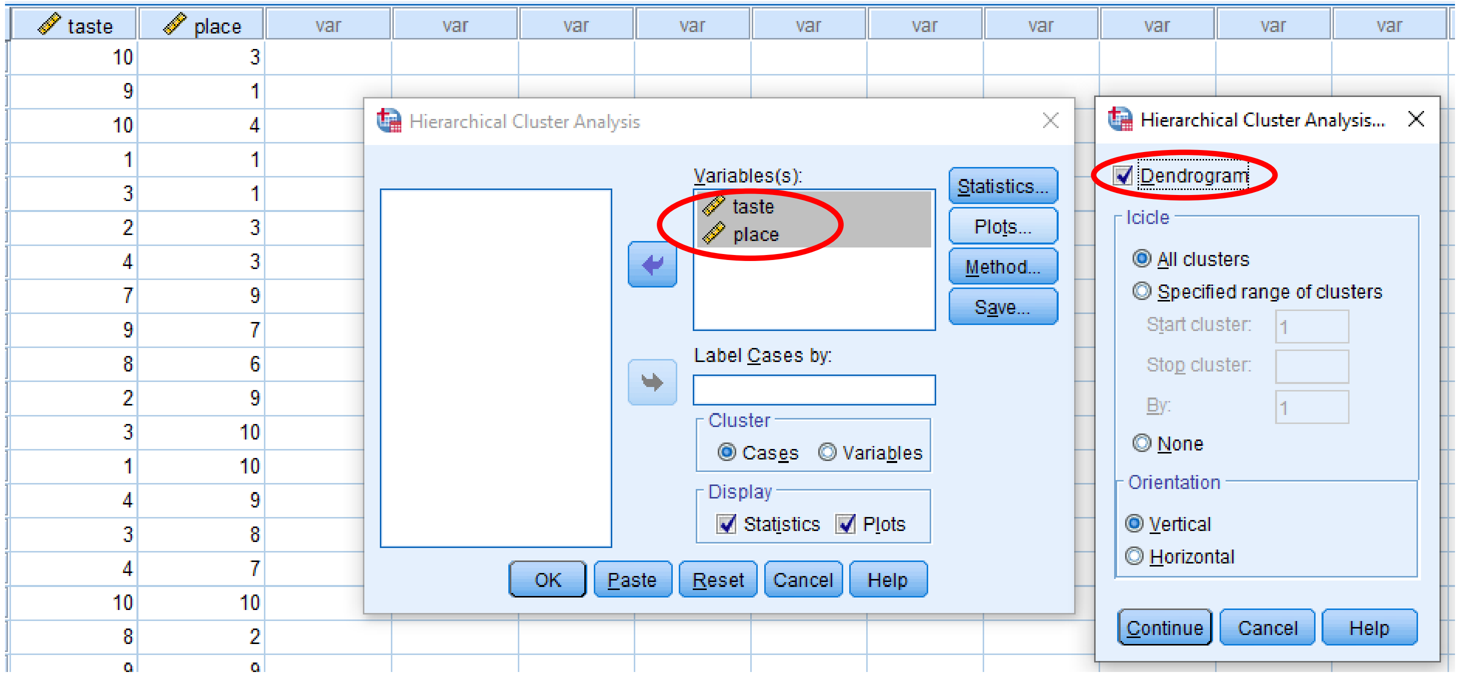
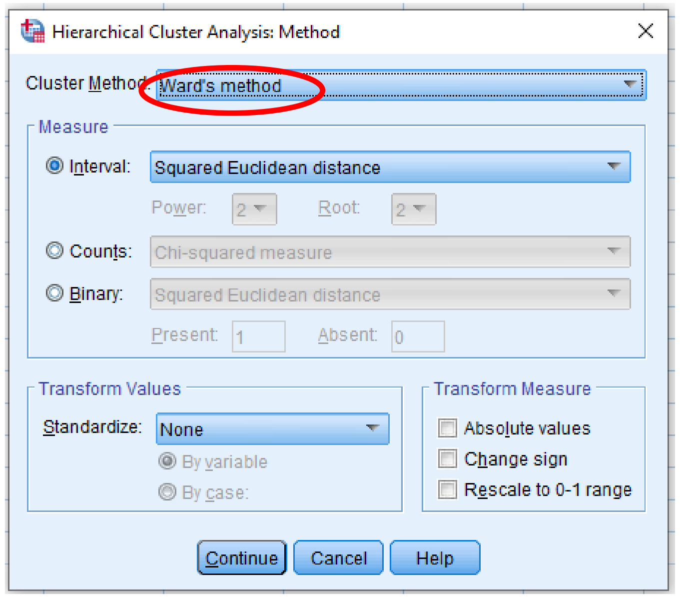
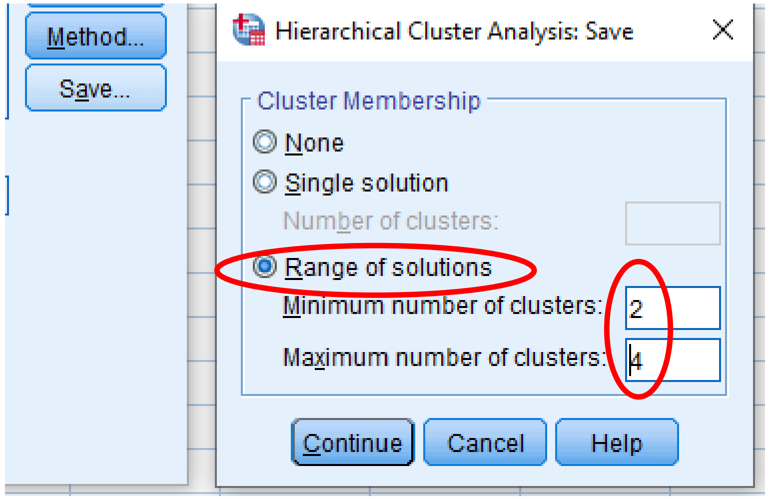


- The furthest neighbor technique produces compact clusters.



4. Beer - How Many Clusters?

- Dendogram – visual logic.
- Agglomeration coefficient – look for large increases.


taste
place
var
Hierarchical Cluster Analysis
Variables(s):
taste
place
Label Cases by:
Cluster
Cases
Variables
Display
Statistics
Plots
Statistics...
Plots...
Method...
Save...
OK
Paste
Reset
Cancel
Help
Hierarchical Cluster Analysis...
Dendrogram
Icicle
All clusters
Specified range of clusters
Start cluster:
Stop cluster:
By:
None
Orientation
Vertical
Horizontal
Continue
Cancel
Help

Hierarchical Cluster Analysis: Method
Cluster Method:
Ward's method
Measure
Interval:
Squared Euclidean distance
Power:
Root:
Counts:
Chi-squared measure
Binary:
Squared Euclidean distance
Present:
Absent:
Transform Values
Standardize:
None
By variable
By case:
Transform Measure
Absolute values
Change sign
Rescale to 0-1 range
Continue
Cancel
Help

Method...
Save...
Hierarchical Cluster Analysis: Save
Cluster Membership
None
Single solution
Number of clusters:
Range of solutions
Minimum number of clusters:
2
Maximum number of clusters:
4
Continue
Cancel
Help

It saved 4, 3, and 2 cluster solutions.

taste	place	CLU4_1	CLU3_1	CLU2_1
10	3	1	1	1
9	1	1	1	1
10	4	1	1	1
1	1	2	2	2
3	1	2	2	2
2	3	2	2	2
4	3	2	2	2
7	9	3	1	1
9	7	3	1	1
8	6	3	1	1
2	9	4	3	2
3	10	4	3	2
1	10	4	3	2
4	9	4	3	2
3	8	4	3	2
4	7	4	3	2
10	10	3	1	1
8	2	1	1	1
9	9	3	1	1
1	3	2	2	2

Ward Linkage

Agglomeration Schedule

	Cluster Combined			Stage Cluster First Appears		
Stage	Cluster 1	Cluster 2	Coefficients	Cluster 1	Cluster 2	Next Stage
1	6	20	,500	0	0	11
2	1	3	1,000	0	0	14
3	17	19	2,000	0	0	12
4	2	18	3,000	0	0	14
5	15	16	4,000	0	0	10
6	12	14	5,000	0	0	10
7	11	13	6,000	0	0	15
8	9	10	7,000	0	0	16
9	4	5	9,000	0	0	13
10	12	15	13,000	6	5	15
11	6	7	17,167	1	0	13
12	8	17	21,500	0	3	16
13	4	6	26,433	9	11	18
14	1	2	32,683	2	4	17
15	11	12	39,350	7	10	18
16	8	9	49,017	12	8	17
17	1	8	122,156	14	16	19
18	4	11	243,253	13	15	19
19	1	4	442,550	17	18	0

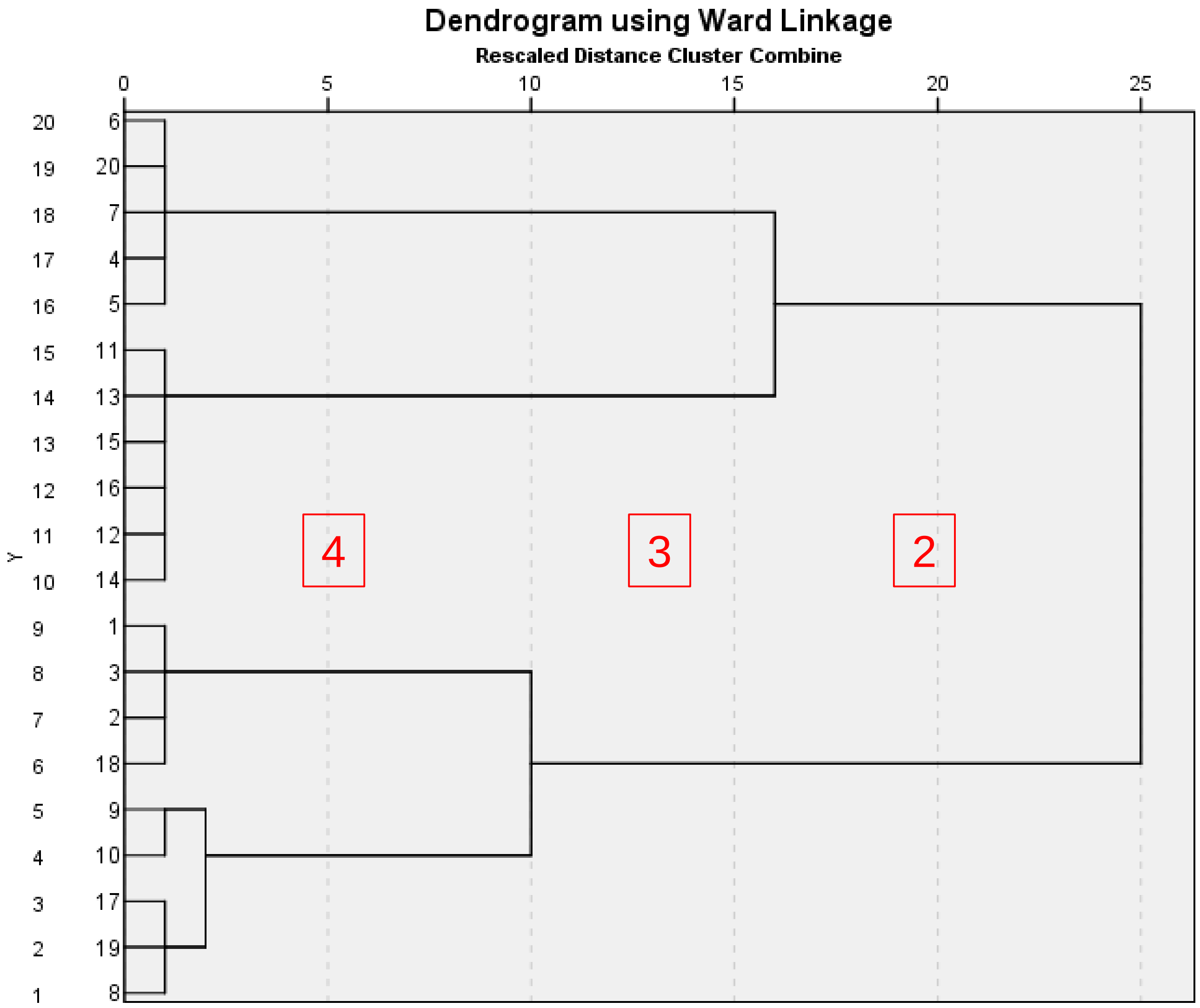
RoT: Stop forming clusters when the degree of change drops. In this case after 4 clusters.

1.18
1.25
2.49
1.99
1.81

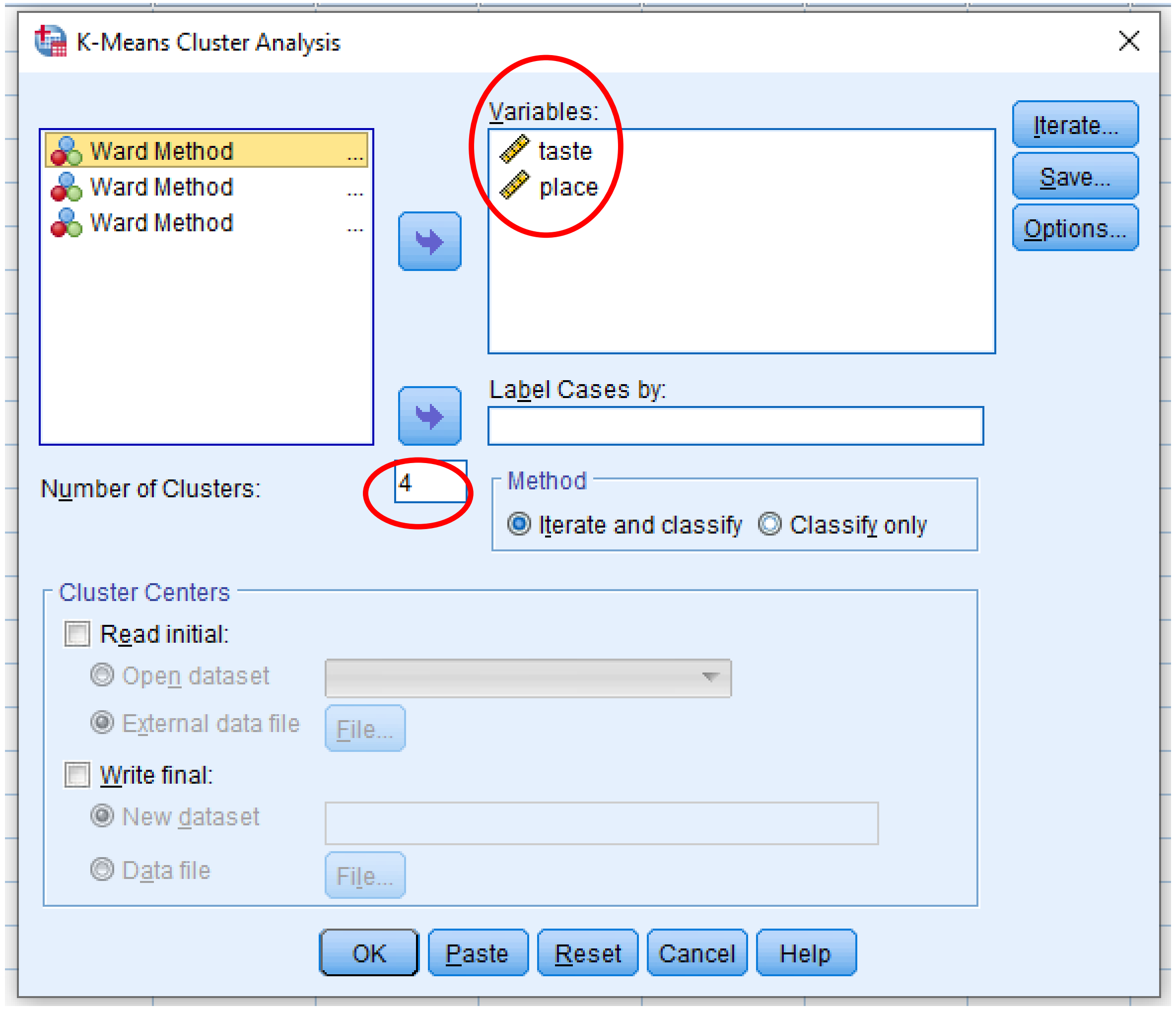
243/122 = 1.99

442/243 = 1.81

The length of the lines indicates the stability of the solution.



Now that we know how many clusters from Hierarchical Clustering, we move to K-means Clustering.



Final Cluster Centers

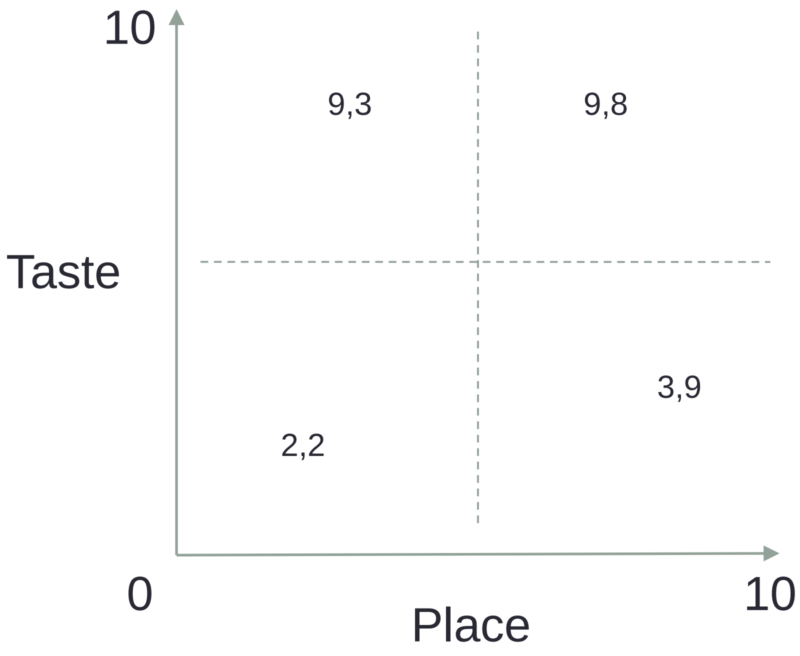
	Cluster 1	2	3	4
taste	2	9	3	9
place	2	3	9	8

Imagine the cluster centers in a two dimensional space (segments) below.

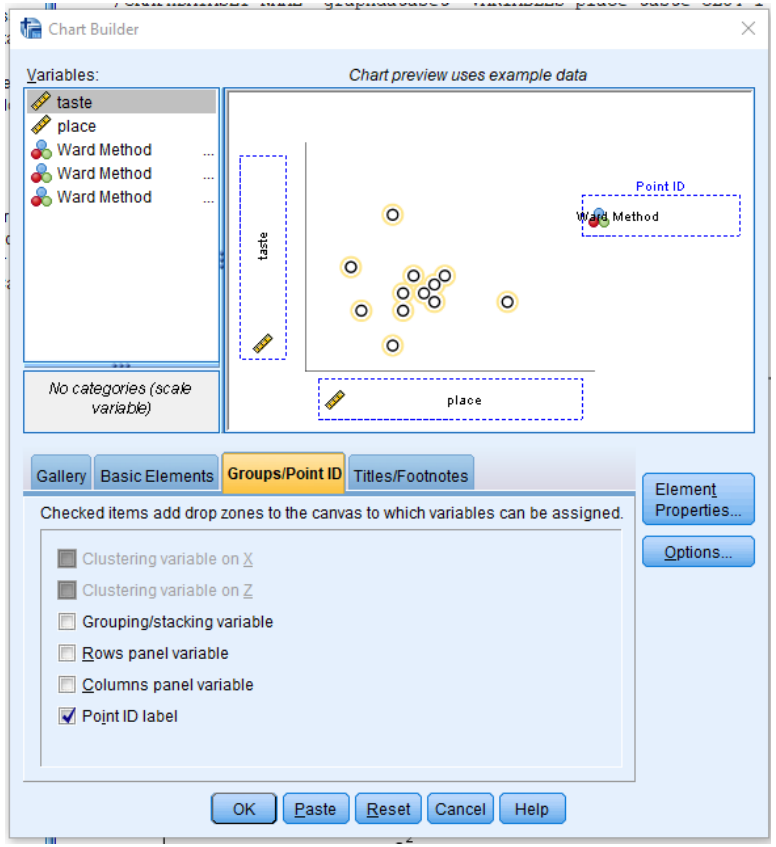
Number of Cases in each Cluster

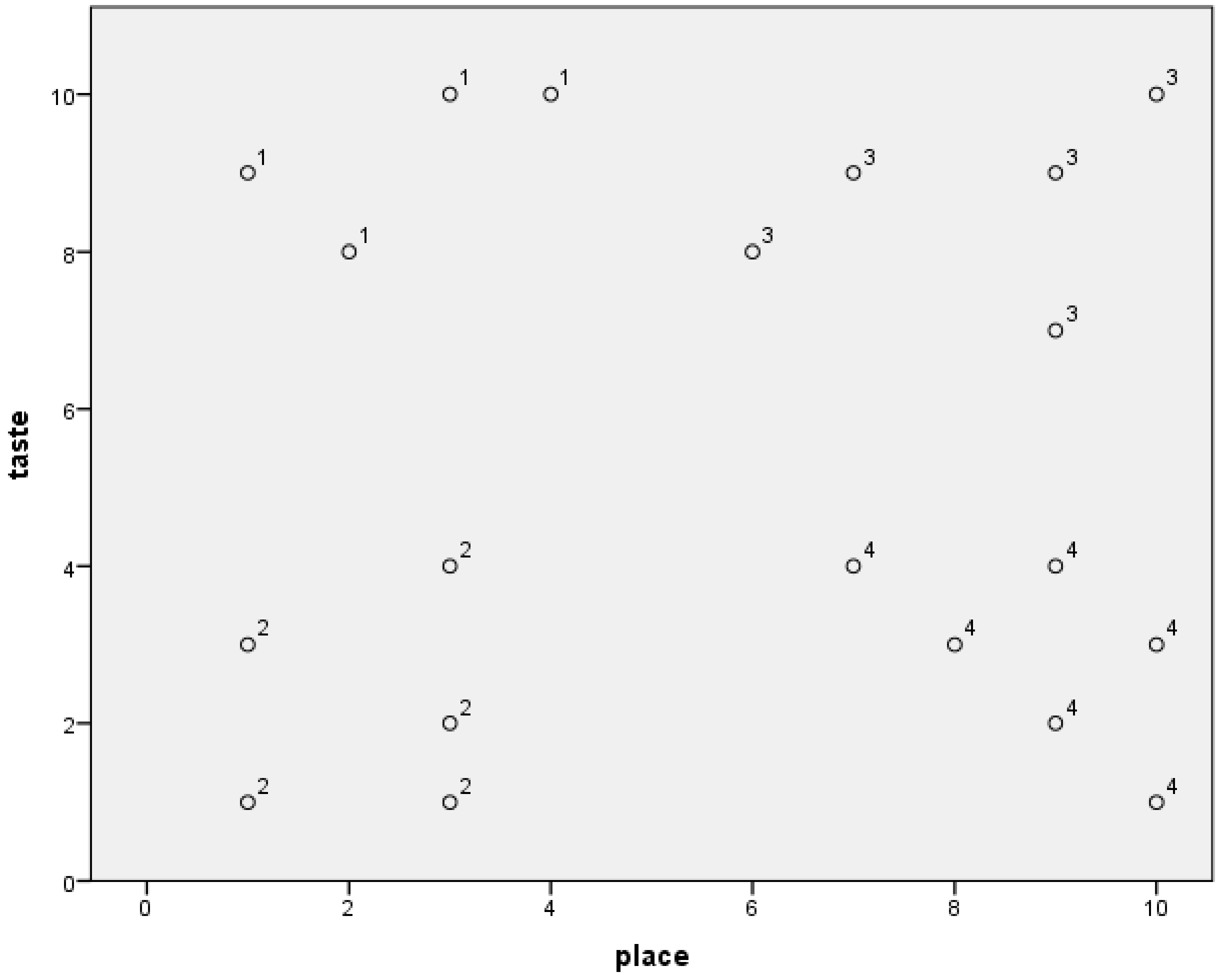
Cluster	1	5,000
	2	4,000
	3	6,000
	4	5,000
Valid		20,000
Missing		,000

The amount of data clustered around each center.



To create a graph in SPSS, *choose Graphs > Chart builder > Scatter, then drag and drop the Simple Scatter into the graph window. Drag and drop taste to the Y-axis and place to the X-axis. Choose Groups/Point ID, and tick Point ID Label. Drag and drop the Cluster Number variable to the Point Label Variable. Click OK.*




taste
place
0
2
4
6
8
10
0
2
4
6
8
10

General Procedure

1. Identify objectives.
2. Select variables.
3. Get raw data.
4. Decide what to do (if anything) about the raw data.
5. Define dissimilarities.
6. Define algorithm.
7. Evaluate the results.
8. Do it all again, slightly differently, to evaluate robustness!

Validation

- Split sample or second sample.

- Criterion or predictive validity:
- Take a variable (not used in the analysis) and measure it across clusters to see if the results come out as predicted.
- e.g. age with beer drinking habits.