SPSS Tutorial

AEB 37 / AE 802

Marketing Research Methods

Week 7

Cluster analysis

Lecture / Tutorial outline

- Cluster analysis
- Example of cluster analysis
- Work on the assignment

Cluster Analysis

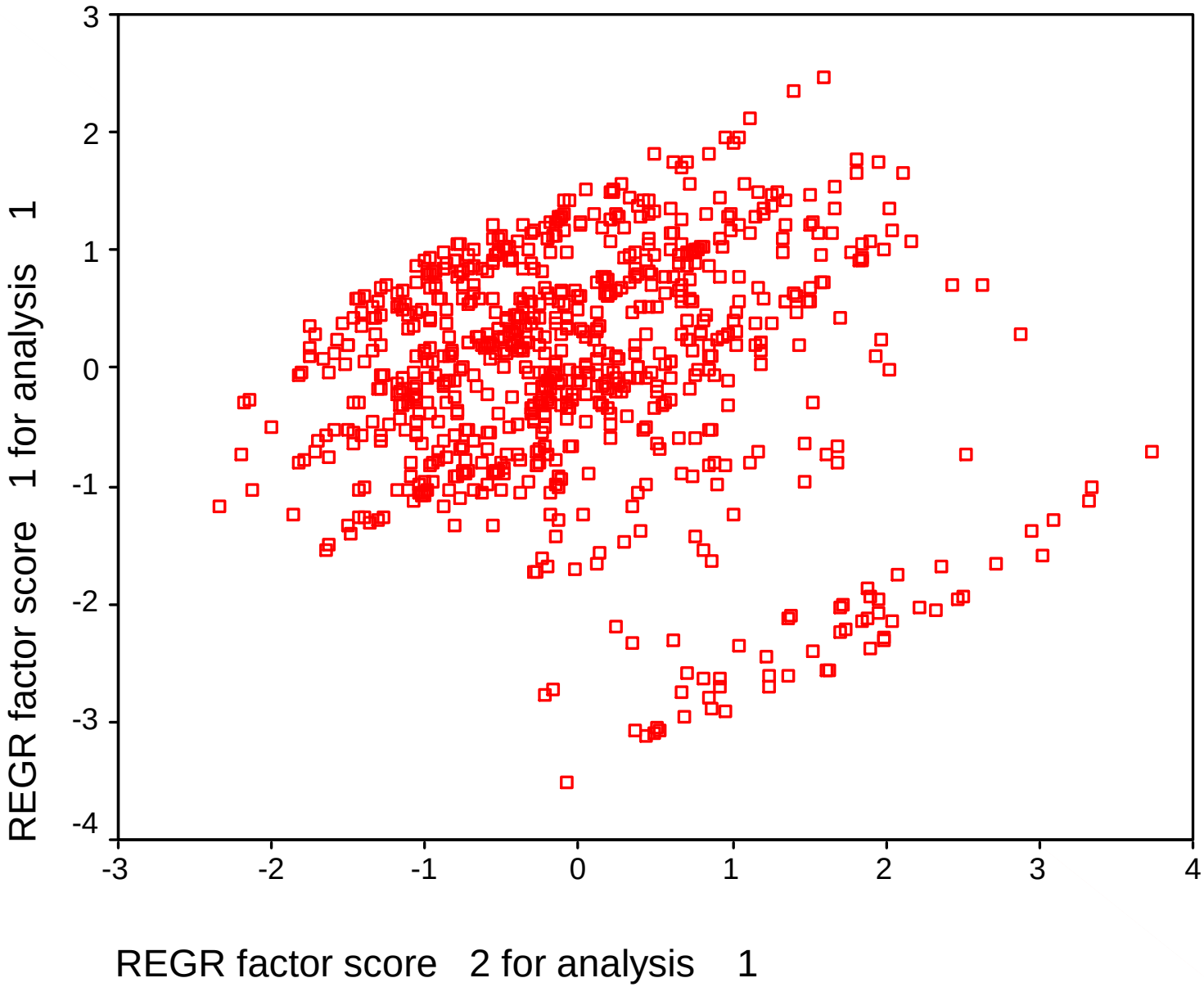
- It is a class of techniques used to classify cases into groups that are relatively homogeneous within themselves and heterogeneous between each other, on the basis of a defined set of variables. These groups are called **clusters**.

Cluster Analysis and marketing research

- **Market segmentation**. E.g. clustering of consumers according to their attribute preferences
- **Understanding buyers behaviours**. Consumers with similar behaviours/characteristics are clustered
- **Identifying new product opportunities.** Clusters of similar brands/products can help identifying competitors / market opportunities
- **Reducing data.** E.g. in preference mapping

Steps to conduct a Cluster Analysis

1. Select a **distance measure**
2. Select a **clustering algorithm**
3. Determine the **number of clusters**
4. Validate the analysis


REGR factor score 1 for analysis 1
REGR factor score 2 for analysis 1
3
2
1
0
-1
-2
-3
-4
-3
-2
-1
0
1
2
3
4

Defining distance: the Euclidean distance

$$D_{ij} = \sqrt{\sum_{k=1}^{n} \left(x_{ki} - x_{kj} \right)^2}$$

D_{ij} distance between cases *i* and *j*

x_{ki} value of variable X_k for case *j*

Problems:

- Different measures = different weights
- Correlation between variables (double counting)

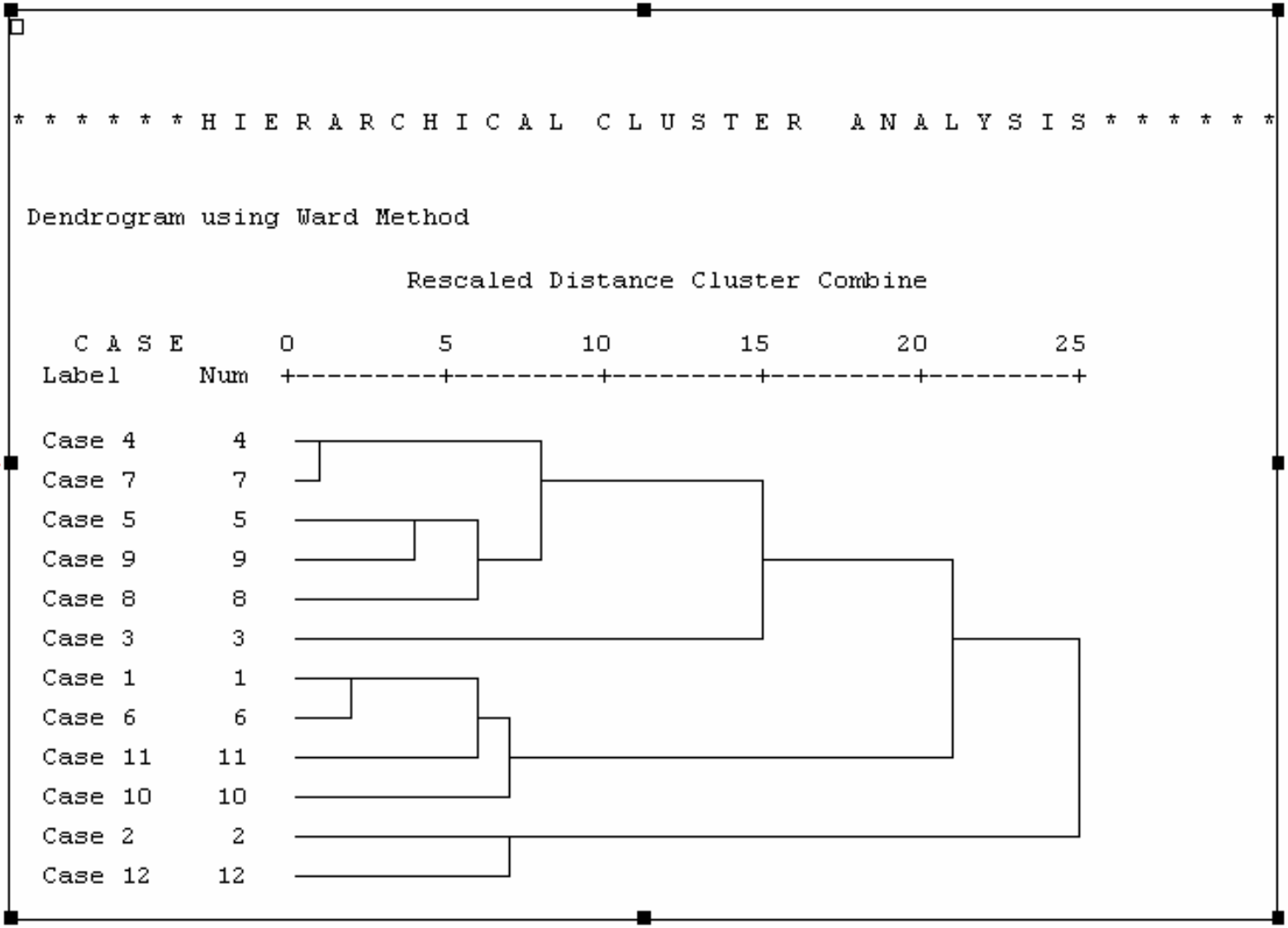
Solution: *Principal component analysis*

Clustering procedures

- **Hierarchical procedures**
  - **Agglomerative** (start from *n* clusters, to get to *1* cluster)
  - **Divisive** (start from *1* cluster, to get to *n* cluster)
- **Non hierarchical procedures**
  - **K-means clustering**

Agglomerative clustering

Dendrogram



Agglomerative clustering

- **Linkage methods**
  - Single linkage (minimum distance)
  - Complete linkage (maximum distance)
  - Average linkage
- **Ward's method**
  1. Compute sum of squared distances within clusters
  2. Aggregate clusters with the minimum increase in the overall sum of squares
- **Centroid method**
  - The distance between two clusters is defined as the difference between the centroids (cluster averages)

K-means clustering

1. The number *k* of cluster is fixed
2. An initial set of *k* “seeds” (aggregation centres) is provided
   - First *k* elements
   - Other seeds
3. Given a certain treshold, all units are assigned to the nearest cluster seed
4. New seeds are computed
5. Go back to step 3 until no reclassification is necessary

*Units can be reassigned in successive steps (**optimising partioning**)*

Hierarchical vs Non hierarchical methods

Hierarchical clustering

- No decision about the number of clusters
- Problems when data contain a high level of error
- Can be very slow
- Initial decision are more influential (one-step only)

Non hierarchical clustering

- Faster, more reliable
- Need to specify the number of clusters (arbitrary)
- Need to set the initial seeds (arbitrary)

Suggested approach

1. First perform a hierarchical method to define the number of clusters
2. Then use the *k*-means procedure to actually form the clusters

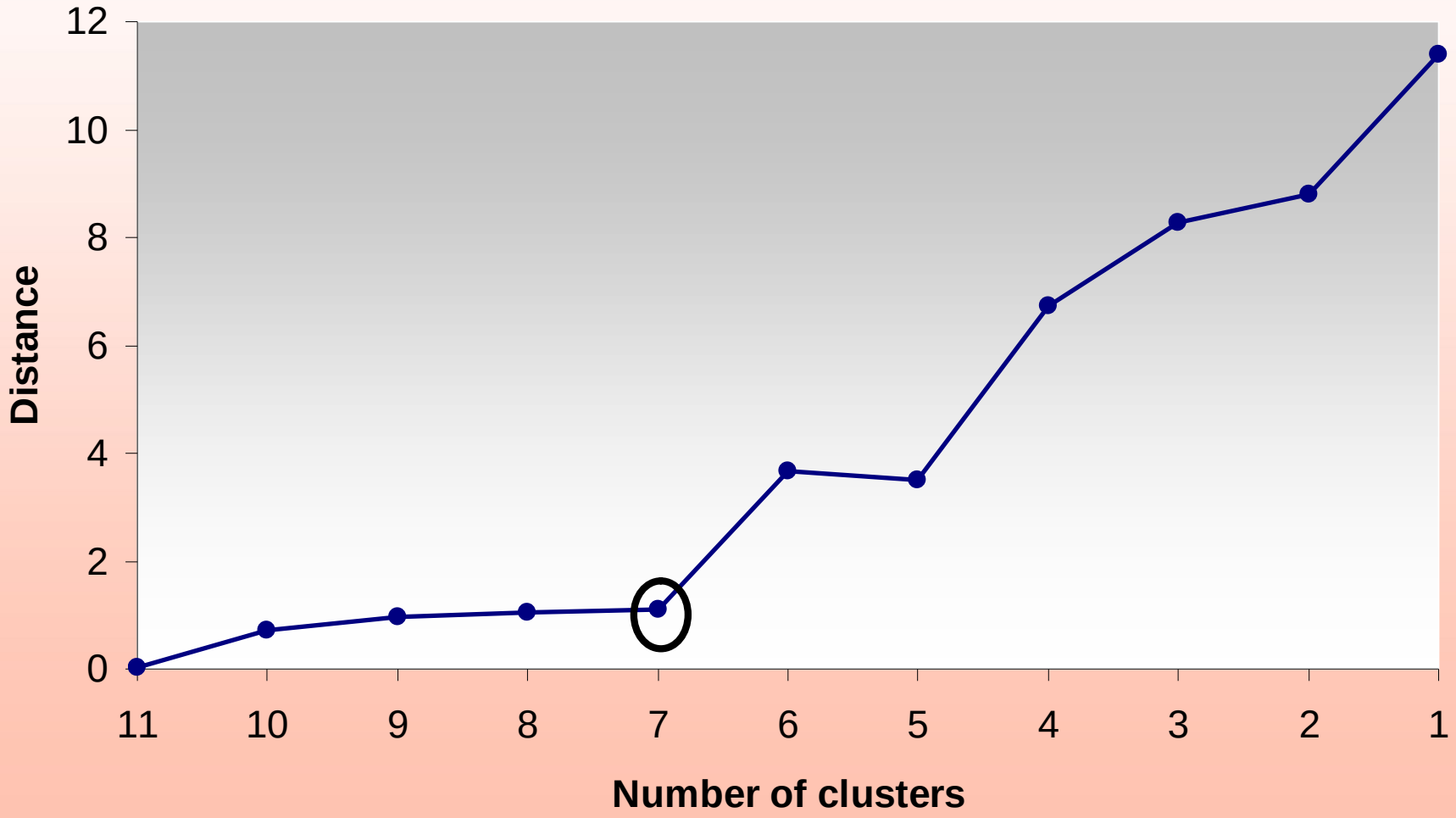
Defining the number of clusters: elbow rule (1)

Stage	Number of clusters
0	12
1	11
2	10
3	9
4	8
5	7
6	6
7	5
8	4
9	3
10	2
11	1

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	4	7	.015	0	0	4
2	6	10	.708	0	0	5
3	8	9	.974	0	0	4
4	4	8	1.042	1	3	6
5	1	6	1.100	0	2	7
6	4	5	3.680	4	0	7
7	1	4	3.492	5	6	8
8	1	11	6.744	7	0	9
9	1	2	8.276	8	0	10
10	1	12	8.787	9	0	11
11	1	3	11.403	10	0	0

Elbow rule (2): the scree diagram



Validating the analysis

- Impact of initial seeds / order of cases
- Impact of the selected method
- Consider the relevance of the chosen set of variables

SPSS Example

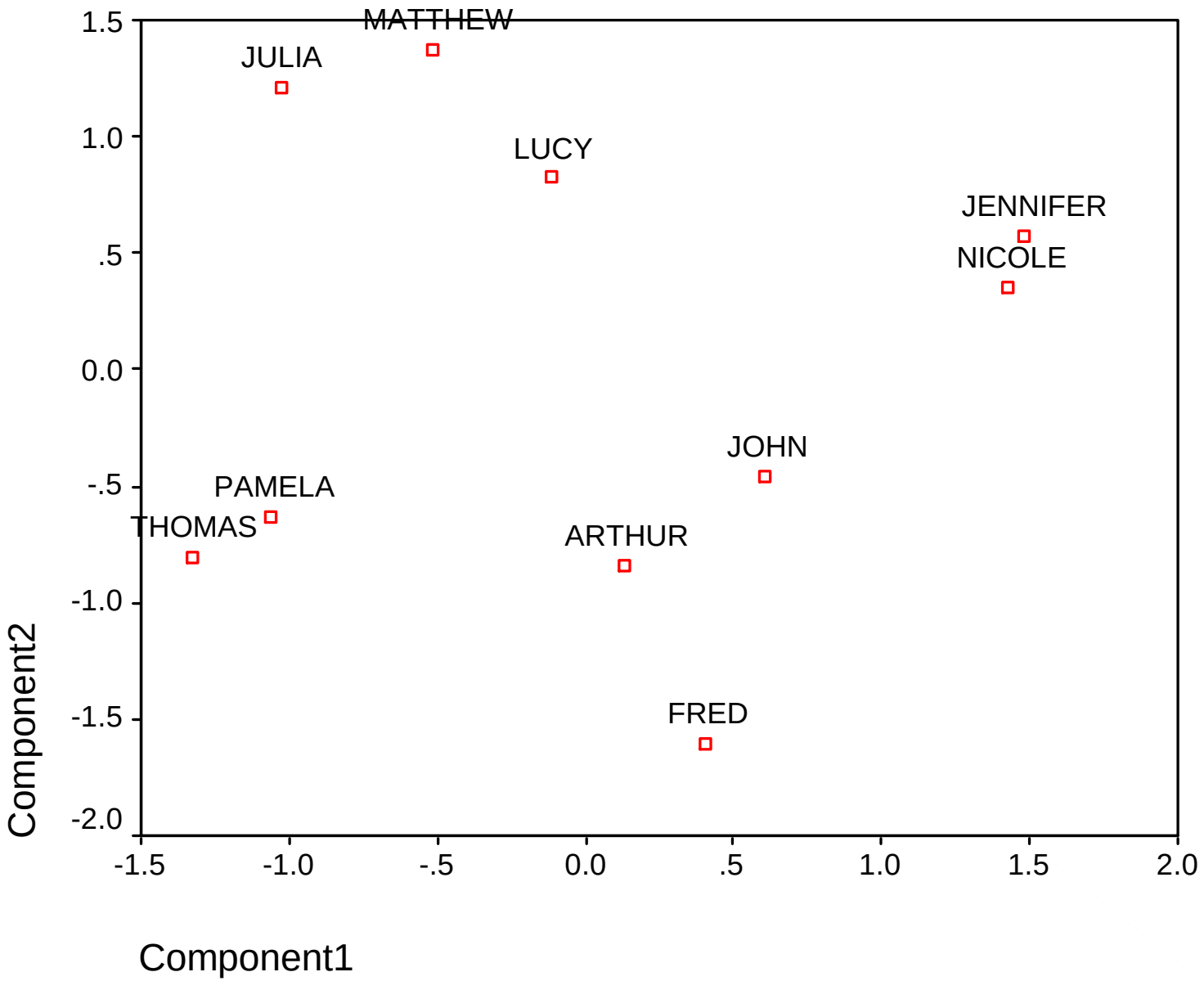
consumers.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities S-PLUS Window Help

Find

11 : case_lbl

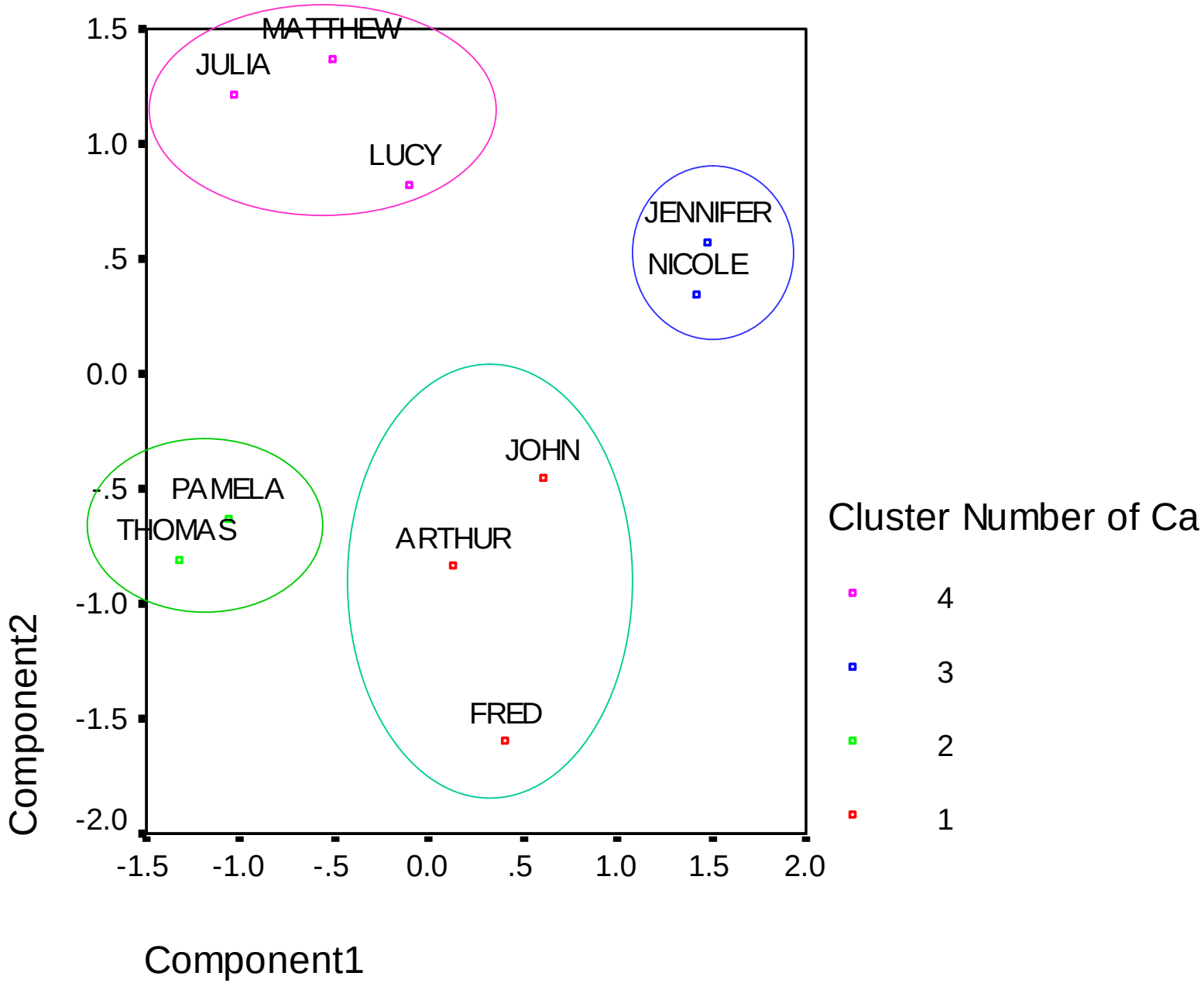
	case_lbl	type	red_sang	red_chia	white_al	red_lamb	compo1	compo2
1	JOHN	Consumer	8.00	6.00	2.00	3.00	.60305	-.45597
2	PAMELA	Consumer	2.00	2.00	2.00	2.00	-1.05945	-.62691
3	NICOLE	Consumer	10.00	7.00	2.00	5.00	1.42420	.35048
4	MATTHEW	Consumer	3.00	3.00	7.00	4.00	-.51418	1.37353
5	THOMAS	Consumer	1.00	1.00	1.00	2.00	-1.32988	-.80649
6	JENNIFER	Consumer	9.00	10.00	5.00	4.00	1.48258	.56904
7	ARTHUR	Consumer	6.00	6.00	2.00	2.00	.13076	-.83874
8	FRED	Consumer	8.00	7.00	1.00	1.00	.40411	-1.60087
9	JULIA	Consumer	2.00	2.00	8.00	3.00	-1.03048	1.21368
10	LUCY	Consumer	5.00	5.00	7.00	3.00	-.11071	.82227
11								


MATTHEW
JULIA
LUCY
JENNIFER
NICOLE
JOHN
PAMELA
THOMAS
ARTHUR
FRED
1.5
1.0
.5
0.0
-.5
-1.0
-1.5
-2.0
-1.5
-1.0
-.5
0.0
.5
1.0
1.5
2.0
Component2
Component1

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	3	6	.026	0	0	8
2	2	5	.078	0	0	7
3	4	9	.224	0	0	5
4	1	7	.409	0	0	6
5	4	10	.849	3	0	8
6	1	8	1.456	4	0	7
7	1	2	4.503	6	2	9
8	3	4	9.878	1	5	9
9	1	3	18.000	7	8	0

Number of clusters: 10 – 6 = 4


MATTHEW
JULIA
LUCY
JENNIFER
NICOLE
JOHN
PAMELA
THOMAS
ARTHUR
FRED
1.5
1.0
.5
0.0
-.5
-1.0
-1.5
-2.0
-1.5
-1.0
-.5
0.0
.5
1.0
1.5
2.0
Component2
Component1
Cluster Number of Ca
4
3
2
1

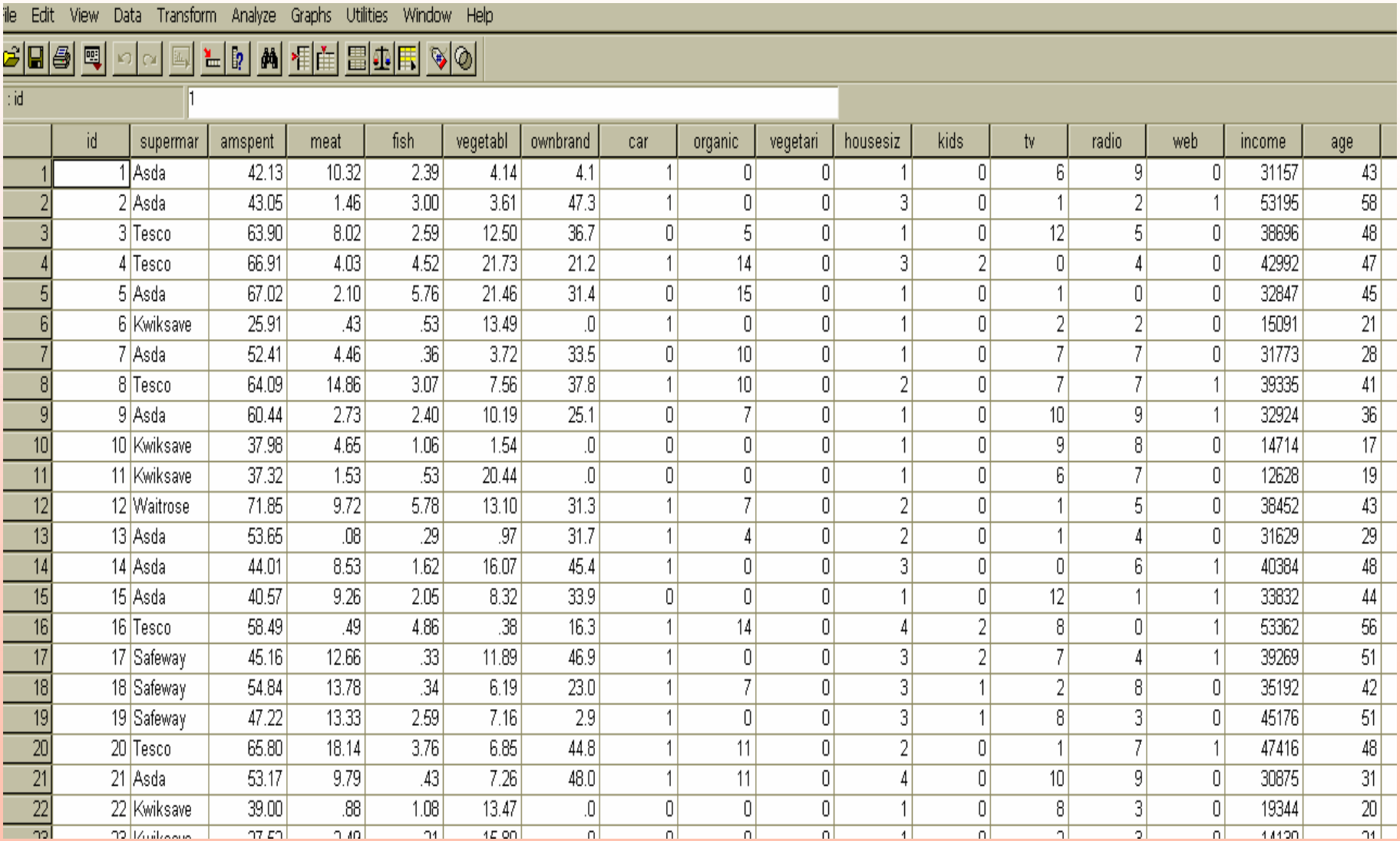
Open the dataset supermarkets.sav

From your N: directory (if you saved it there last time

Or download it from: http://www.rdg.ac.uk/~aes02mm/supermarket.sav

- Open it in SPSS

The supermarkets.sav dataset



ile Edit View Data Transform Analyze Graphs Utilities Window Help

: id 1

	id	supermar	amspent	meat	fish	vegetabl	ownbrand	car	organic	vegetari	housesiz	kids	tv	radio	web	income	age
1	1	Asda	42.13	10.32	2.39	4.14	4.1	1	0	0	1	0	6	9	0	31157	43
2	2	Asda	43.05	1.46	3.00	3.61	47.3	1	0	0	3	0	1	2	1	53195	58
3	3	Tesco	63.90	8.02	2.59	12.50	36.7	0	5	0	1	0	12	5	0	38696	48
4	4	Tesco	66.91	4.03	4.52	21.73	21.2	1	14	0	3	2	0	4	0	42992	47
5	5	Asda	67.02	2.10	5.76	21.46	31.4	0	15	0	1	0	1	0	0	32847	45
6	6	Kwiksave	25.91	.43	.53	13.49	.0	1	0	0	1	0	2	2	0	15091	21
7	7	Asda	52.41	4.46	.36	3.72	33.5	0	10	0	1	0	7	7	0	31773	28
8	8	Tesco	64.09	14.86	3.07	7.56	37.8	1	10	0	2	0	7	7	1	39335	41
9	9	Asda	60.44	2.73	2.40	10.19	25.1	0	7	0	1	0	10	9	1	32924	36
10	10	Kwiksave	37.98	4.65	1.06	1.54	.0	0	0	0	1	0	9	8	0	14714	17
11	11	Kwiksave	37.32	1.53	.53	20.44	.0	0	0	0	1	0	6	7	0	12628	19
12	12	Waitrose	71.85	9.72	5.78	13.10	31.3	1	7	0	2	0	1	5	0	38452	43
13	13	Asda	53.65	.08	.29	.97	31.7	1	4	0	2	0	1	4	0	31629	29
14	14	Asda	44.01	8.53	1.62	16.07	45.4	1	0	0	3	0	0	6	1	40384	48
15	15	Asda	40.57	9.26	2.05	8.32	33.9	0	0	0	1	0	12	1	1	33832	44
16	16	Tesco	58.49	.49	4.86	.38	16.3	1	14	0	4	2	8	0	1	53362	56
17	17	Safeway	45.16	12.66	.33	11.89	46.9	1	0	0	3	2	7	4	1	39269	51
18	18	Safeway	54.84	13.78	.34	6.19	23.0	1	7	0	3	1	2	8	0	35192	42
19	19	Safeway	47.22	13.33	2.59	7.16	2.9	1	0	0	3	1	8	3	0	45176	51
20	20	Tesco	65.80	18.14	3.76	6.85	44.8	1	11	0	2	0	1	7	1	47416	48
21	21	Asda	53.17	9.79	.43	7.26	48.0	1	11	0	4	0	10	9	0	30875	31
22	22	Kwiksave	39.00	.88	1.08	13.47	.0	0	0	0	1	0	8	3	0	19344	20

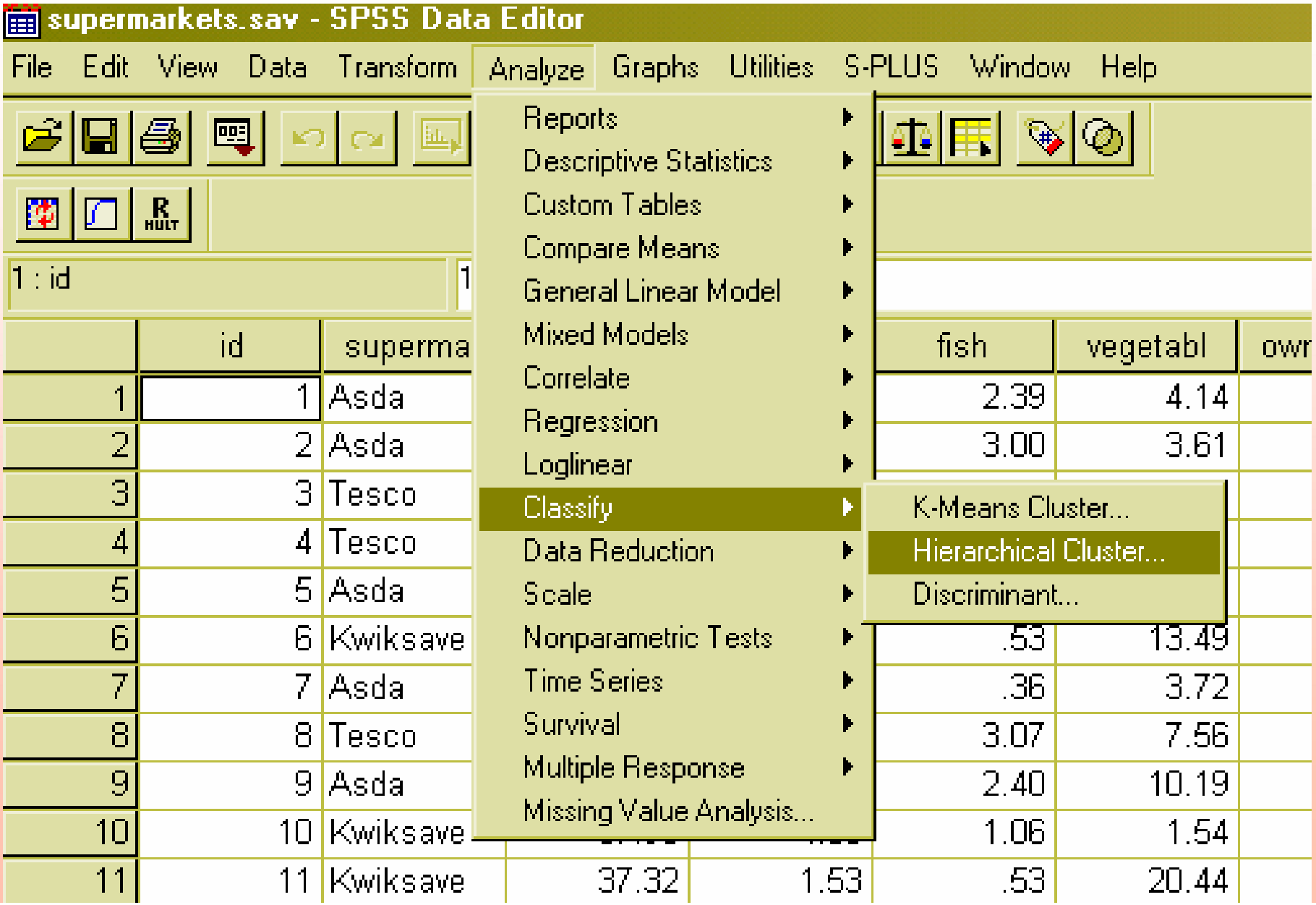
Run Principal Components Analysis and save scores

- Select the variables to perform the analysis
- Set the rule to extract principal components
- Give instruction to save the principal components as new variables

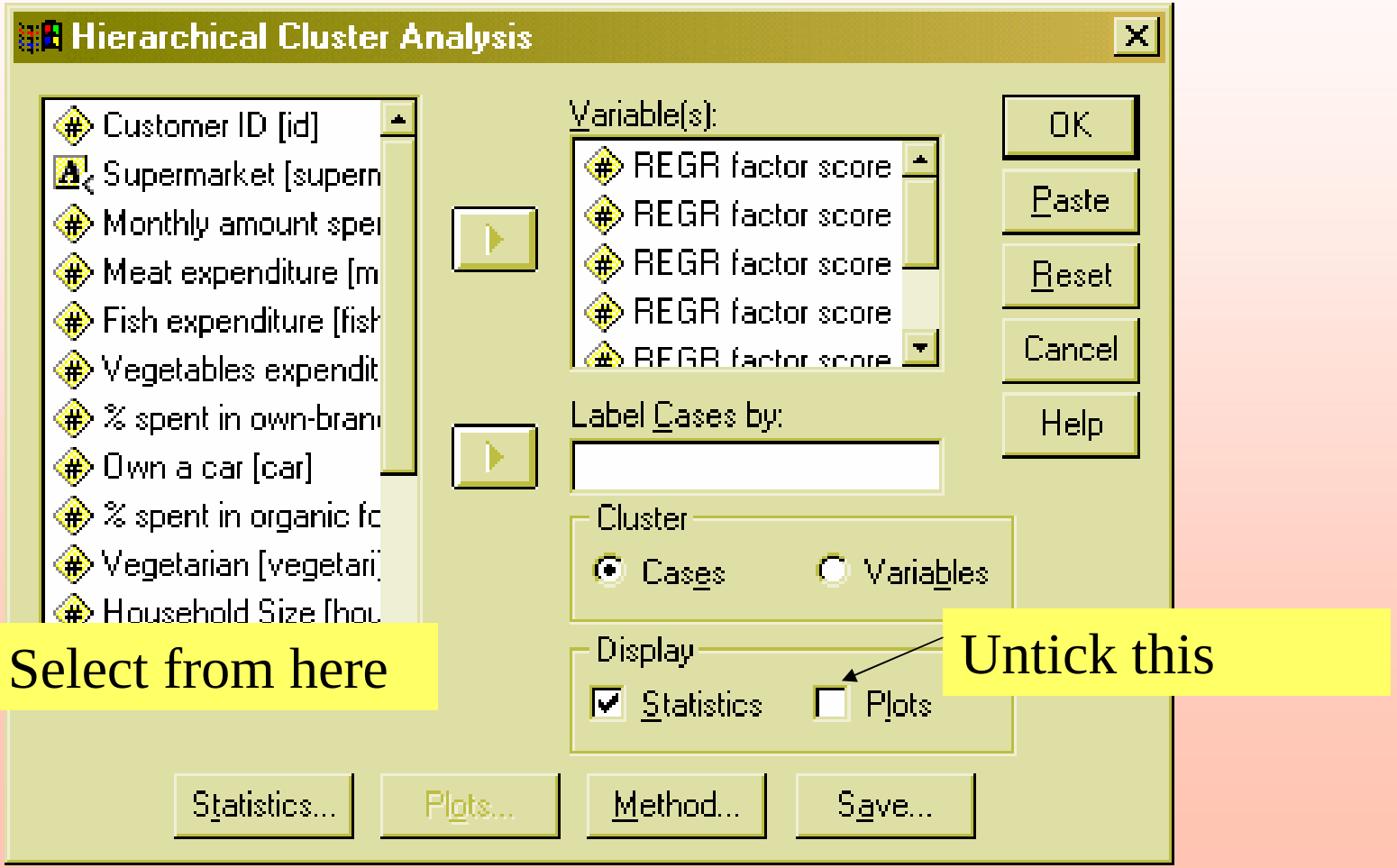
Cluster analysis: basic steps

- Apply Ward's methods on the principal components score
- Check the agglomeration schedule
- Decide the number of clusters
- Apply the *k*-means method

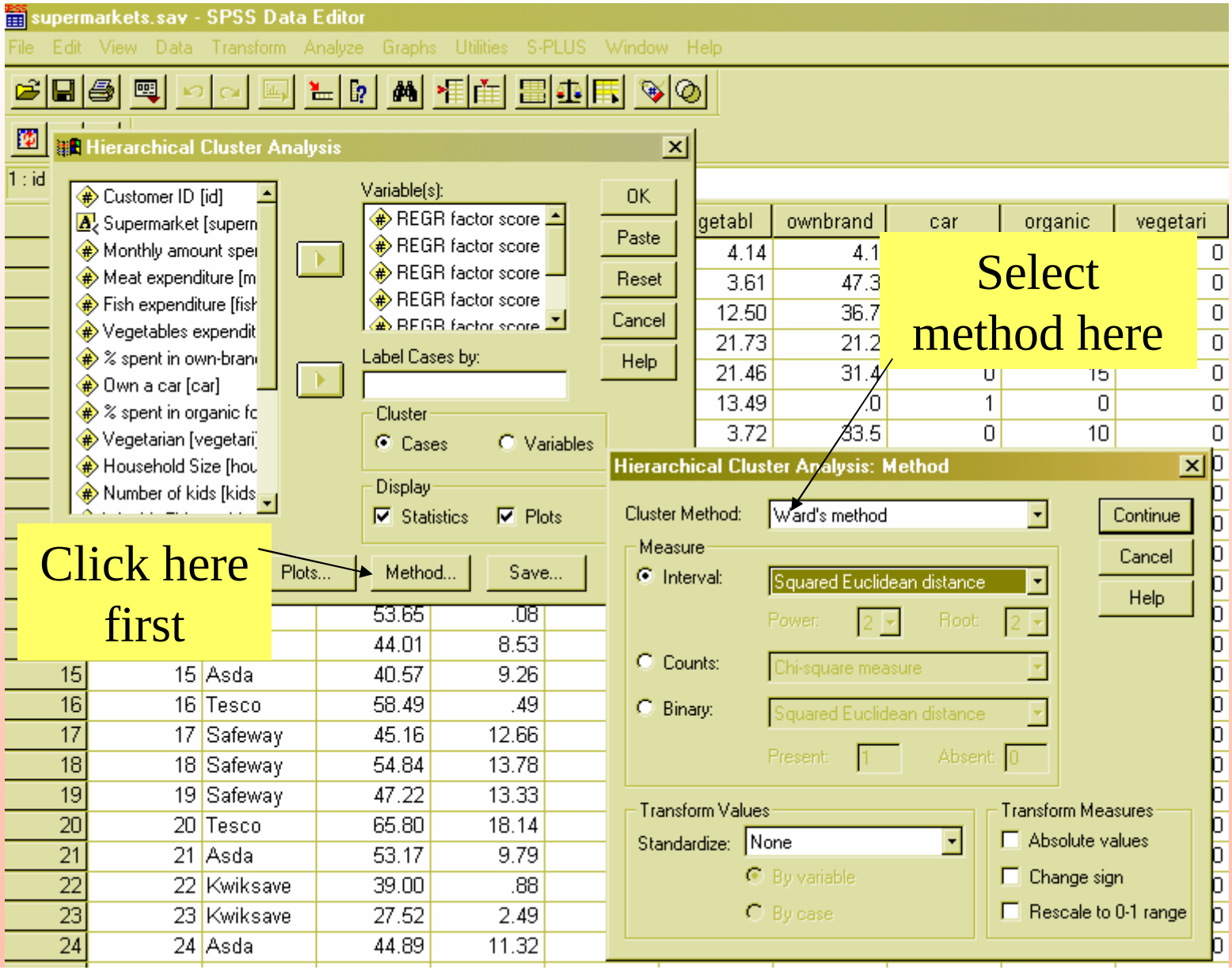
Analyse / Classify



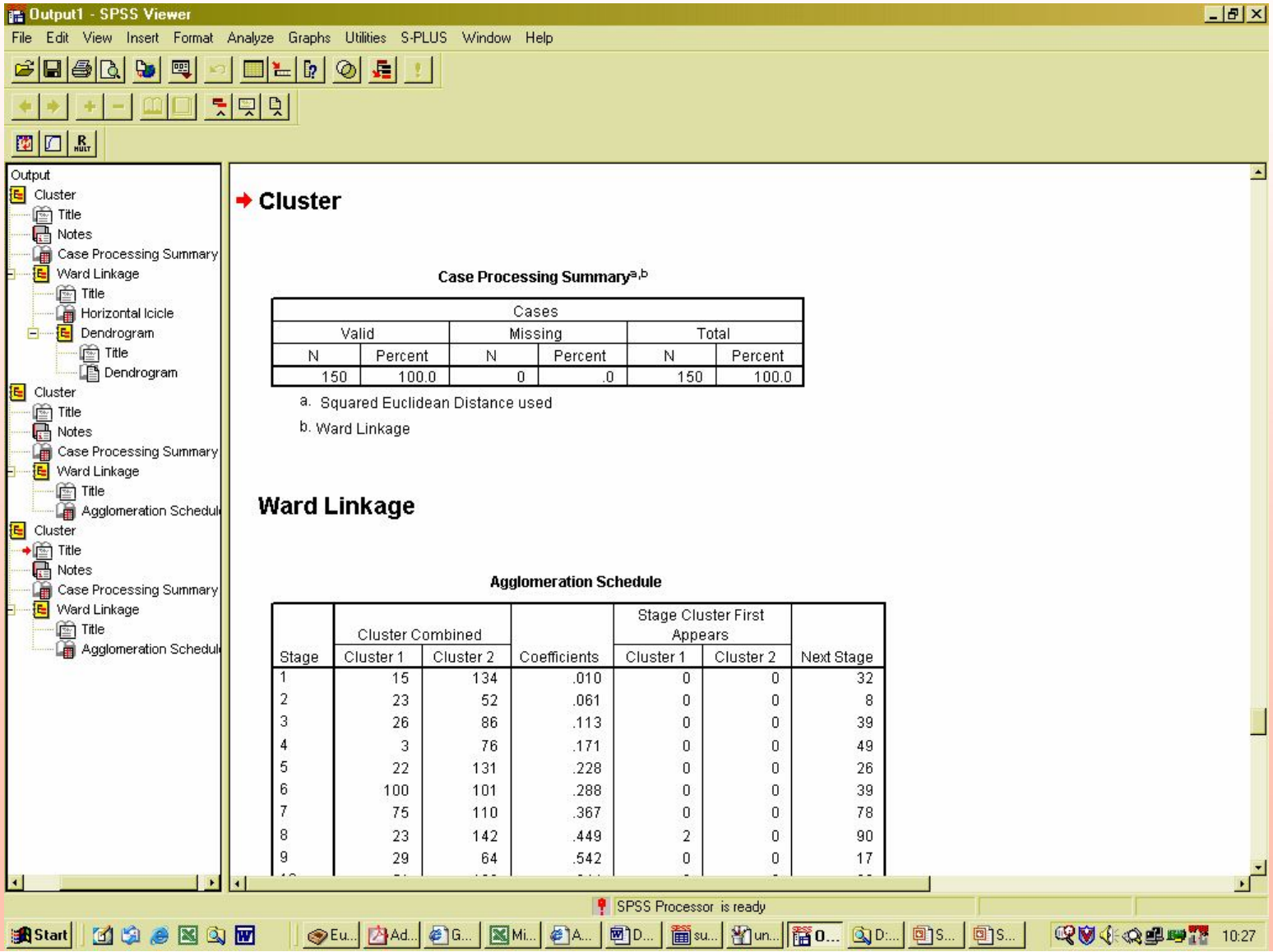
Select the component scores



Select Ward's algorithm



Output: Agglomeration schedule



Number of clusters

Identify the step where the “distance coefficients” makes a bigger jump

ng Summary

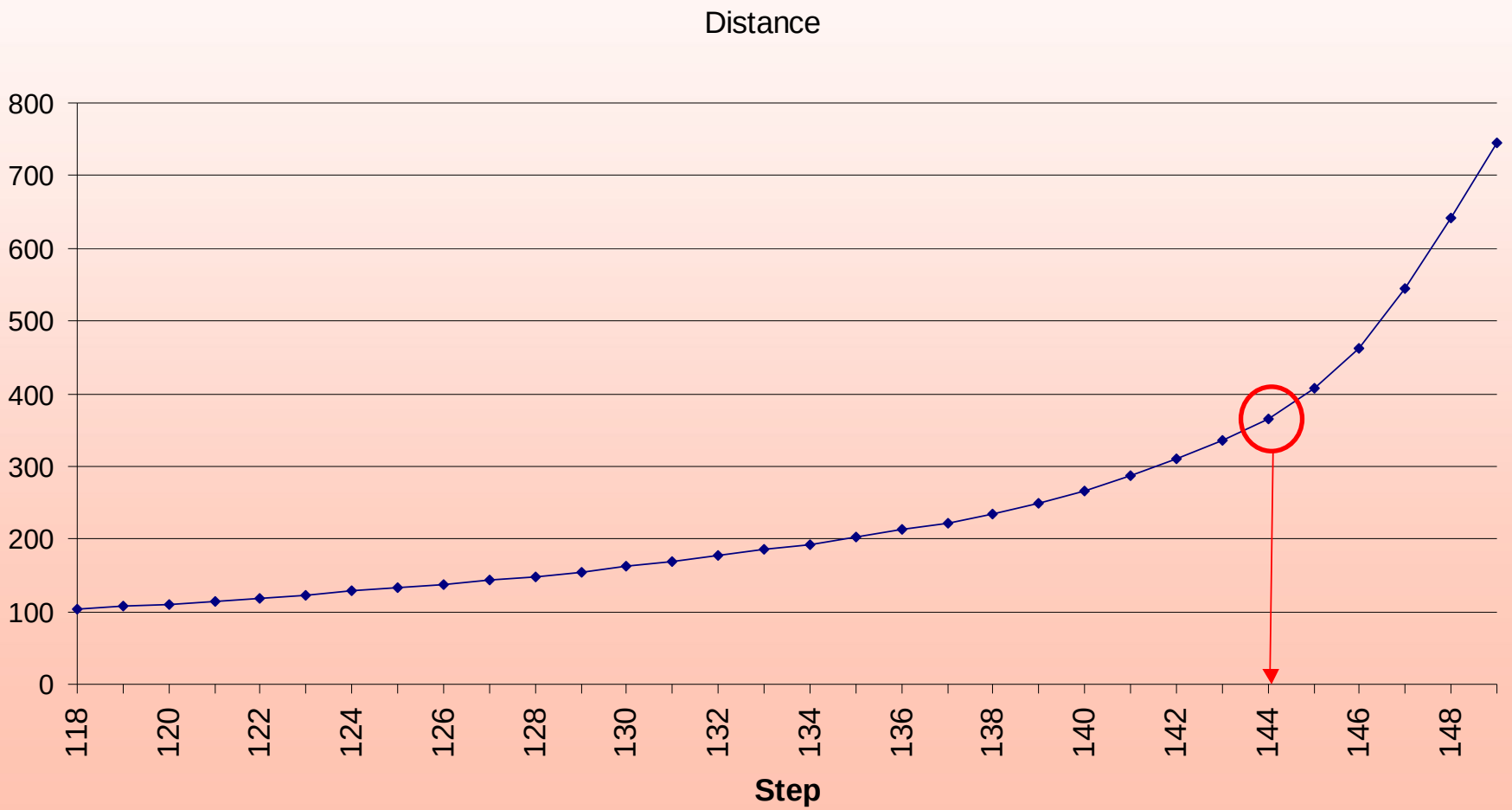
tion Schedul

ng Summary

tion Schedul

131	4	53	169.657	118	112	139
132	84	102	177.150	82	117	141
133	5	118	185.005	105	97	137
134	31	77	192.994	120	108	144
135	3	7	202.437	96	125	142
136	6	10	212.393	106	123	147
137	5	12	222.672	133	127	146
138	47	49	234.383	122	103	141
139	4	16	249.159	131	126	144
140	1	15	266.917	129	109	142
141	47	84	286.668	138	132	143
142	1	3	311.016	140	135	145
143	36	47	335.895	130	141	149
144	4	31	364.634	139	134	148
145	1	2	408.251	142	128	146
146	1	5	463.151	145	137	147
147	1	6	545.117	146	136	148
148	1	4	642.226	147	144	149
149	1	36	745.000	148	143	0

The scree diagram (Excel needed)



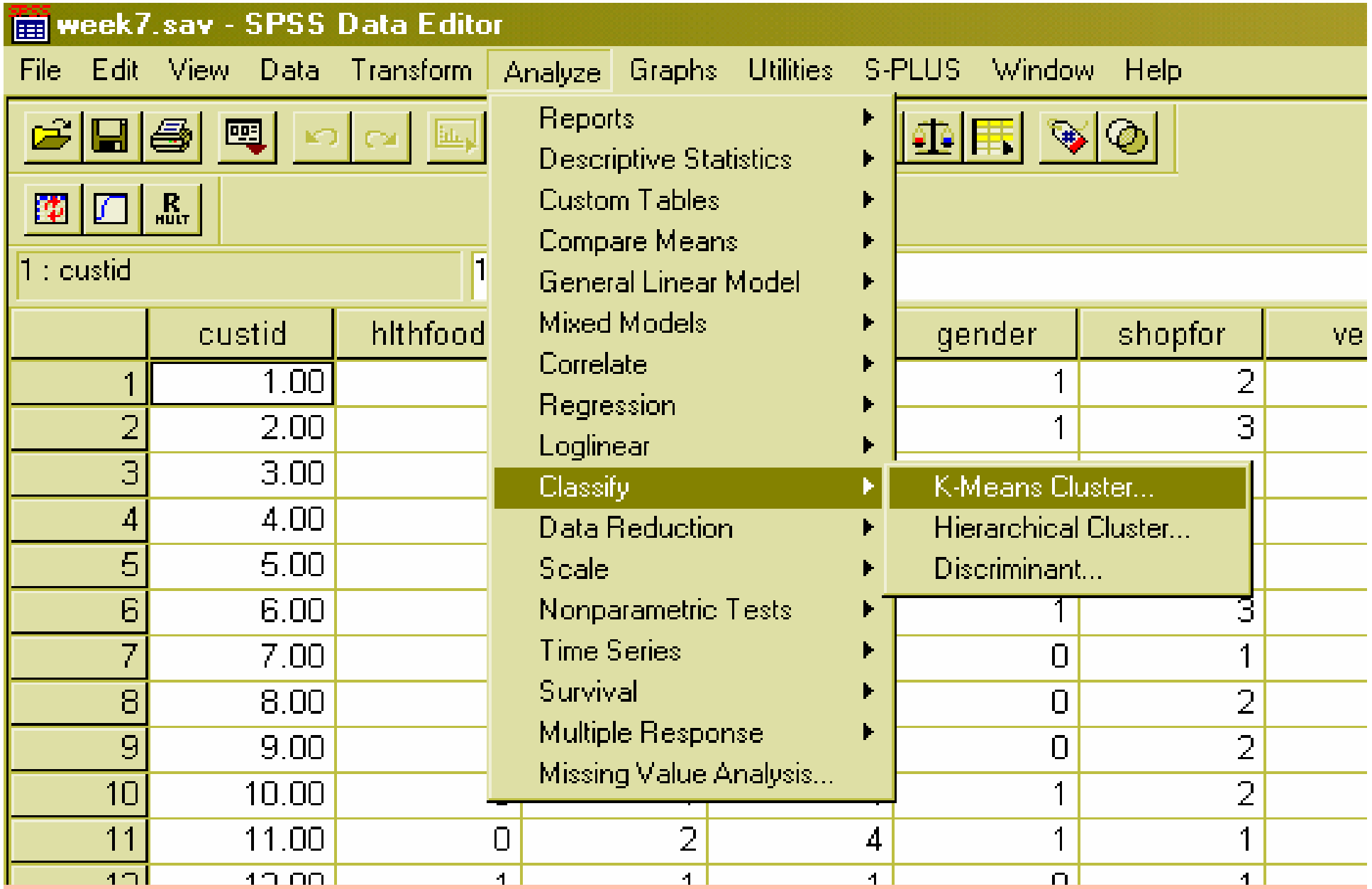
Number of clusters

Number of cases	150
Step of ‘elbow’	144
Number of clusters	6

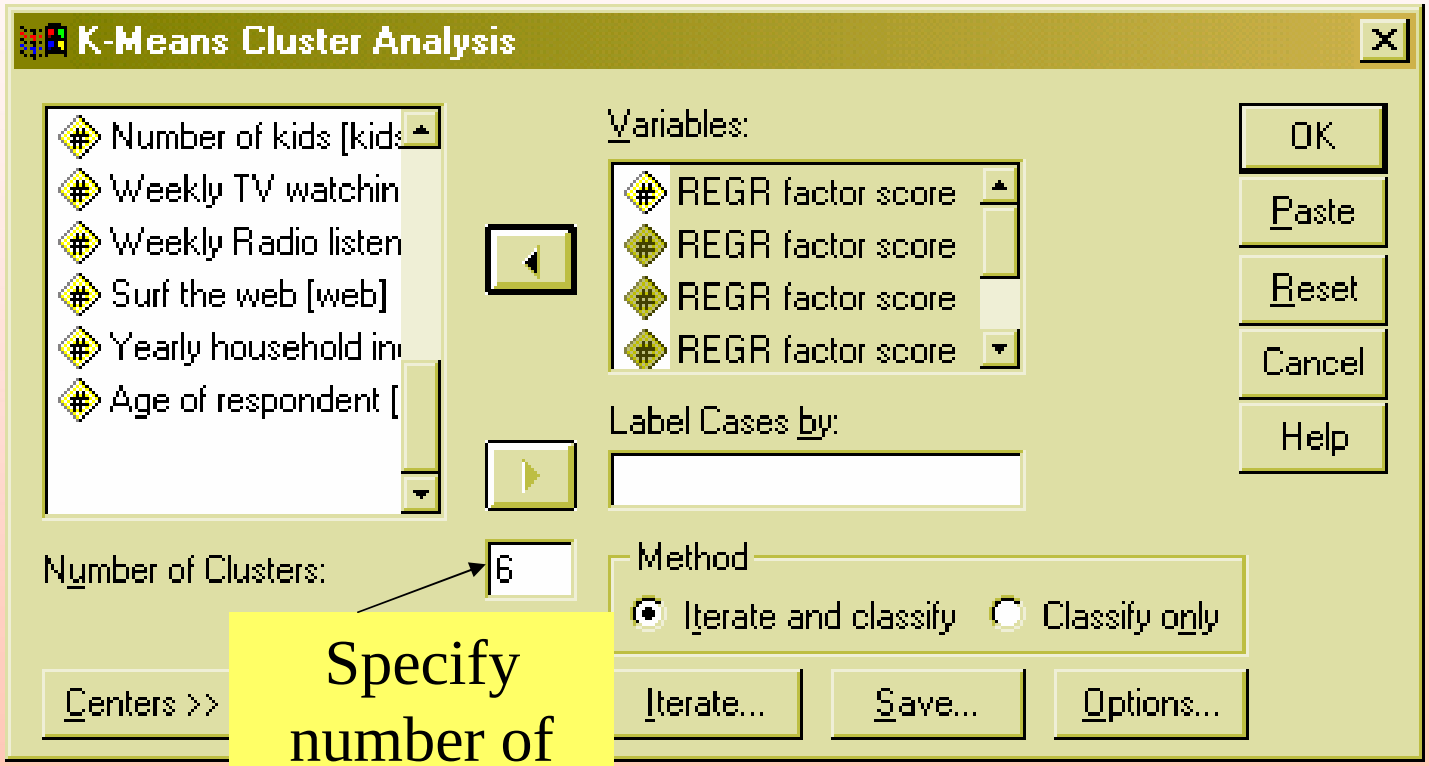
Now repeat the analysis

- Choose the *k*-means technique
- Set **6** as the number of clusters
- Save cluster number for each case
- Run the analysis

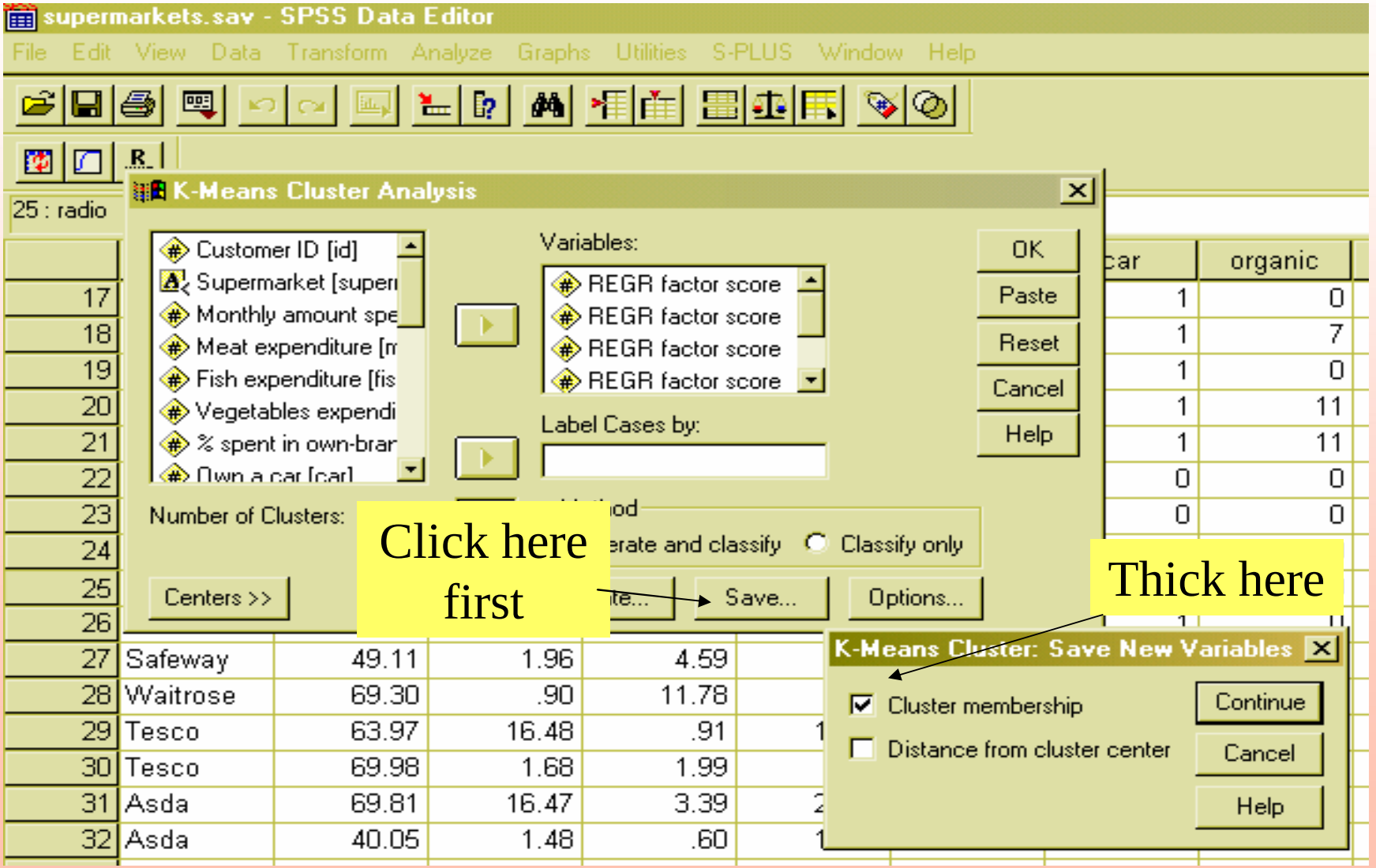
K-means


week7.sav - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities S-PLUS Window Help
Reports
Descriptive Statistics
Custom Tables
Compare Means
General Linear Model
Mixed Models
Correlate
Regression
Loglinear
Classify
Data Reduction
Scale
Nonparametric Tests
Time Series
Survival
Multiple Response
Missing Value Analysis...
K-Means Cluster...
Hierarchical Cluster...
Discriminant...
1 : custid
custid
hlthfood
gender
shopfor
1.00
2.00
3.00
4.00
5.00
6.00
7.00
8.00
9.00
10.00
11.00

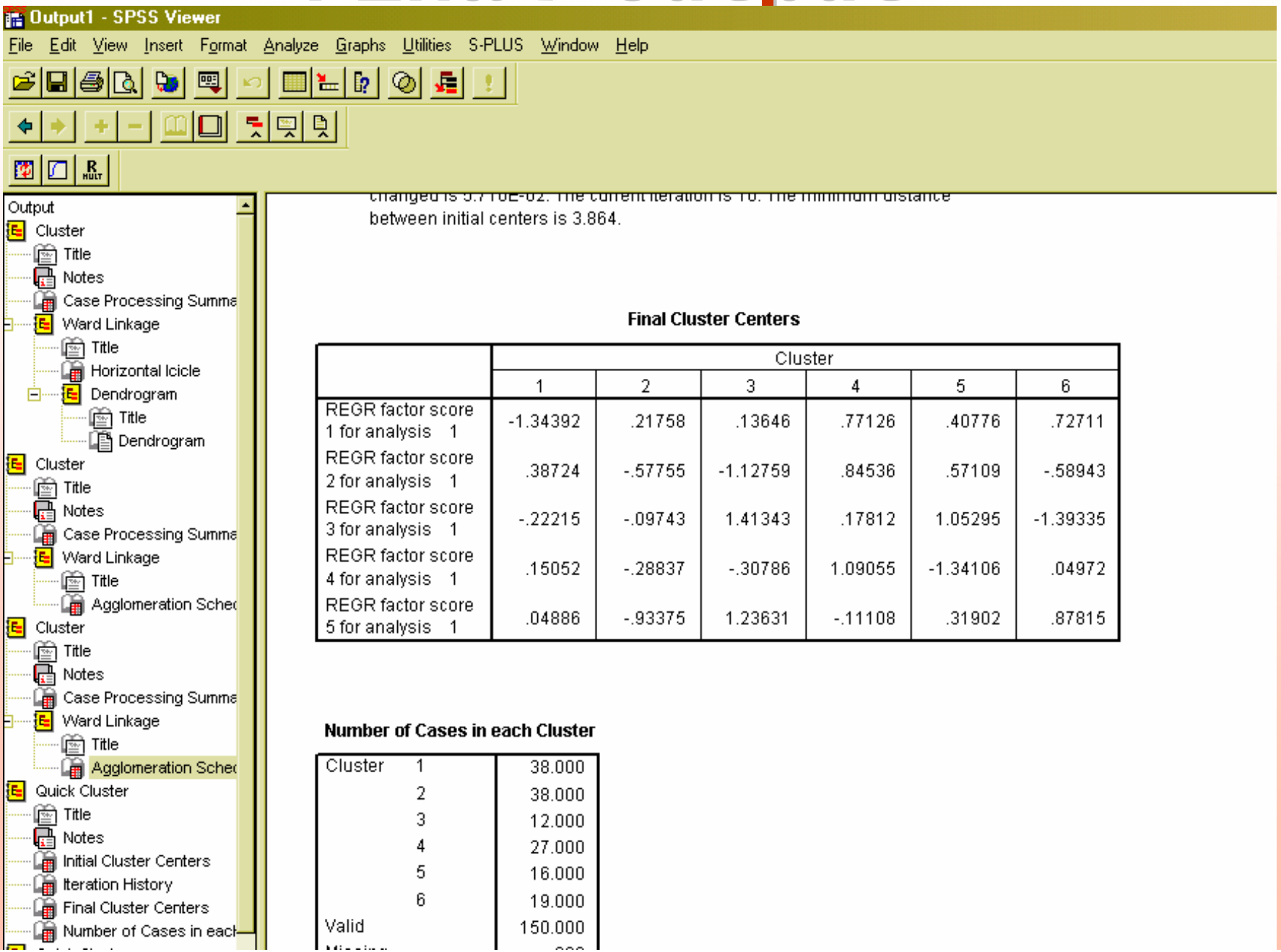
K-means dialog box



Save cluster membership



Final output



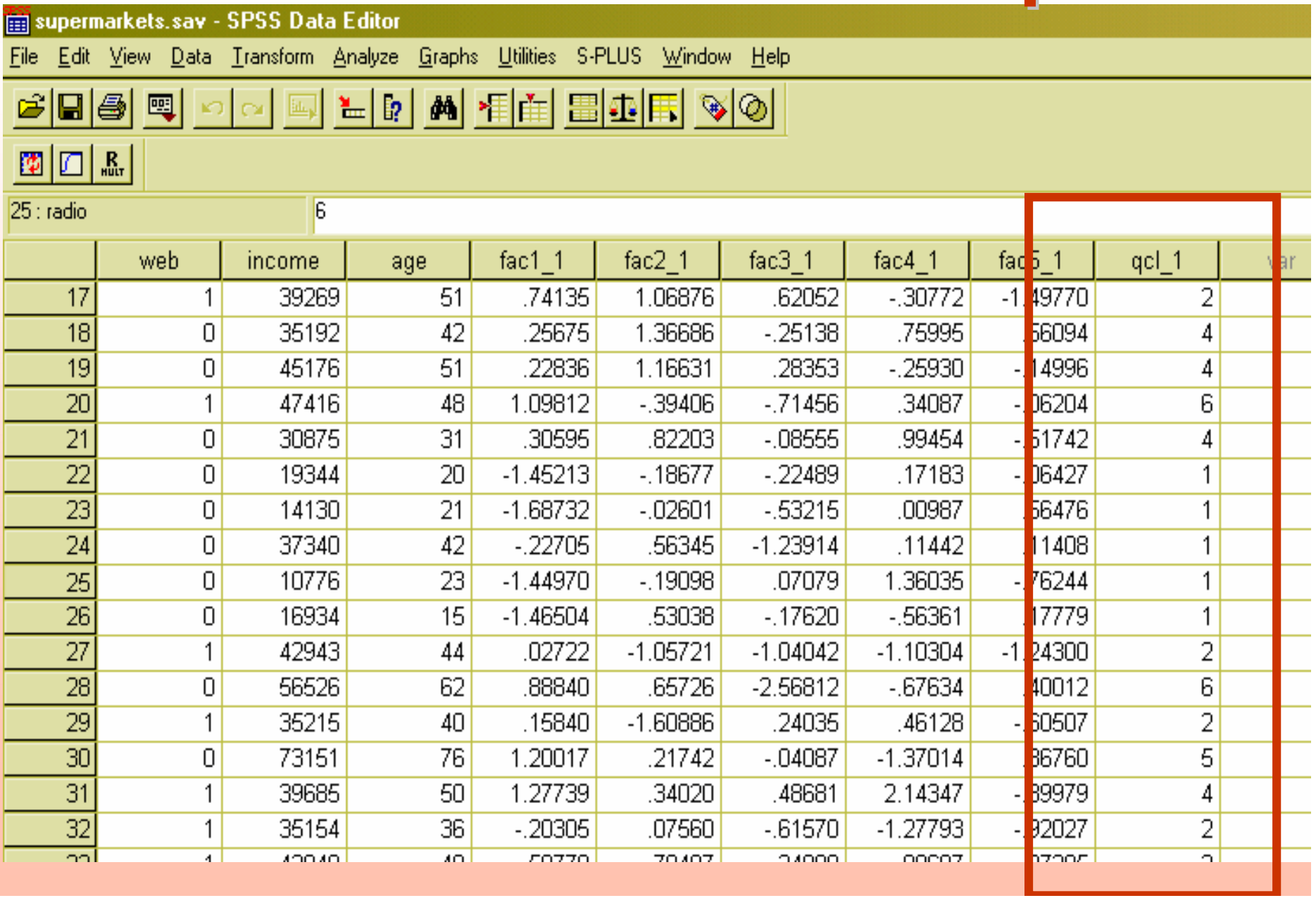
Final Cluster Centers

	Cluster					
	1	2	3	4	5	6
REGR factor score 1 for analysis 1	-1.34392	.21758	.13646	.77126	.40776	.72711
REGR factor score 2 for analysis 1	.38724	-.57755	-1.12759	.84536	.57109	-.58943
REGR factor score 3 for analysis 1	-.22215	-.09743	1.41343	.17812	1.05295	-1.39335
REGR factor score 4 for analysis 1	.15052	-.28837	-.30786	1.09055	-1.34106	.04972
REGR factor score 5 for analysis 1	.04886	-.93375	1.23631	-.11108	.31902	.87815

Number of Cases in each Cluster

Cluster	1	38.000
	2	38.000
	3	12.000
	4	27.000
	5	16.000
	6	19.000
Valid		150.000

Cluster membership



	web	income	age	fac1_1	fac2_1	fac3_1	fac4_1	fac5_1	qcl_1	var
17	1	39269	51	.74135	1.06876	.62052	-.30772	-1.49770	2	
18	0	35192	42	.25675	1.36686	-.25138	.75995	.56094	4	
19	0	45176	51	.22836	1.16631	.28353	-.25930	-.14996	4	
20	1	47416	48	1.09812	-.39406	-.71456	.34087	-.06204	6	
21	0	30875	31	.30595	.82203	-.08555	.99454	-.51742	4	
22	0	19344	20	-1.45213	-.18677	-.22489	.17183	-.06427	1	
23	0	14130	21	-1.68732	-.02601	-.53215	.00987	.56476	1	
24	0	37340	42	-.22705	.56345	-1.23914	.11442	.11408	1	
25	0	10776	23	-1.44970	-.19098	.07079	1.36035	-.76244	1	
26	0	16934	15	-1.46504	.53038	-.17620	-.56361	.17779	1	
27	1	42943	44	.02722	-1.05721	-1.04042	-1.10304	-1.24300	2	
28	0	56526	62	.88840	.65726	-2.56812	-.67634	.40012	6	
29	1	35215	40	.15840	-1.60886	.24035	.46128	-.60507	2	
30	0	73151	76	1.20017	.21742	-.04087	-1.37014	.86760	5	
31	1	39685	50	1.27739	.34020	.48681	2.14347	-.89979	4	
32	1	35154	36	-.20305	.07560	-.61570	-1.27793	-.92027	2	

Component meaning (tutorial week 5)

1. "Old Rich Big Spender"

2. Family shopper

3. Vegetarian TV lover

4. Organic radio listener

5. Vegetarian TV and web hater

Component Matrix[a]

	1				5
Monthly amount spent					.173
Meat expenditure			.347	.334	-5.95E-02
Fish expenditure	.525	-.206	-.475	-4.35E-02	
Vegetables expenditure	.192	-.345	-.127	.383	
% spent in own-brand product	.646	-.281	-.134	-.239	
Own a car	.536	.619	-.102	-.172	6.008E-02
% spent in organic food	.492	-.186	.190	.460	.342
Vegetarian	1.784E-02	-9.24E-02	.647	-.287	.507
Household Size	.649	.612	.135	-6.12E-02	-3.29E-03
Number of kids	.369	.663	.247	.184	1.694E-02
Weekly TV watching (hours)	.124	-9.53E-02	.462	.232	-.529
Weekly Radio listening (hours)	2.989E-02	.406	-.349	.559	-8.14E-02
Surf the web	.443	-.271	.182	-5.61E-02	-.465
Yearly household income	.908	-4.75E-02	-7.46E-02	-.197	-3.26E-02
Age of respondent	.891	-5.64E-02	-6.73E-02	-.228	6.942E-04

Extraction Method: Principal Component Analysis.

a. 5 components extracted.

Final Cluster Centers

	Cluster					
	1	2	3	4	5	6
REGR factor score 1 for analysis 1	-1.34392	.21758	.13646	.77126	.40776	.72711
REGR factor score 2 for analysis 1	.38724	-.57755	-1.12759	.84536	.57109	-.58943
REGR factor score 3 for analysis 1	-.22215	-.09743	1.41343	.17812	1.05295	-1.39335
REGR factor score 4 for analysis 1	.15052	-.28837	-.30786	1.09055	-1.34106	.04972
REGR factor score 5 for analysis 1	.04886	-.93375	1.23631	-.11108	.31902	.87815

Cluster interpretation through mean component values

- Cluster 1 is very far from profile 1 (-1.34) and more similar to profile 2 (0.38)
- Cluster 2 is very far from profile 5 (-0.93) and not particularly similar to any profile
- Cluster 3 is extremely similar to profiles 3 and 5 and very far from profile 2
- Cluster 4 is similar to profiles 2 and 4
- Cluster 5 is very similar to profile 3 and very far from profile 4
- Cluster 6 is very similar to profile 5 and very far from profile 3

Which cluster to target?

- Objective: target the organic consumer
- Which is the cluster that looks more “organic”?
- Compute the descriptive statistics on the original variables for that cluster

Representation of factors 1 and 4

(and cluster membership)