ΣΤΑΤΙΣΤΙΚΗ ΚΑΙ ΘΕΩΡΙΑ ΑΠΟΦΑΣΕΩΝ

ΔΙΑΛΕΞΗ: ΑΝΑΛΥΣΗ ΚΑΤΑ ΣΥΣΤΑΔΕΣ

Διδάσκουσα: Ε. Γάκη, Επίκ. Καθηγήτρια

Περιεχόμενα Διάλεξης

Ανάλυση Κατά Συστάδες (Cluster Analysis)

- Σκοπός της Ανάλυσης Κατά Συστάδες
- Παραδείγματα Εφαρμογών
- Βασικές Έννοιες
- Προσεγγίσεις για την ομαδοποίηση των δεδομένων
- Επιπρόσθετοι Σκοποί της Ανάλυσης Κατά Συστάδες
- Η Απόσταση (Έννοια, Μέτρα)
- Προβλήματα που πρέπει να αντιμετωπιστούν
- Η Μέθοδος K-Means
- Εφαρμογή της Μεθόδου K-Means
- Ιεραρχική Ομαδοποίηση
- Εφαρμογή της Ιεραρχικής Ομαδοποίησης

Σκοπός Ανάλυσης Κατά Συστάδες

Σκοπός της Ανάλυσης Κατά Συστάδες είναι η κατάταξη σε ομάδες των υπαρχουσών παρατηρήσεων, χρησιμοποιώντας την πληροφορία που υπάρχει σε κάποιες μεταβλητές.

Εξετάζει πόσο όμοιες είναι κάποιες παρατηρήσεις ως προς κάποιον αριθμό μεταβλητών ώστε να δημιουργήσει ομάδες από παρατηρήσεις που μοιάζουν μεταξύ τους.

Μια επιτυχημένη ανάλυση θα πρέπει να καταλήξει σε ομάδες για τις οποίες οι παρατηρήσεις μέσα σε κάθε ομάδα να είναι όσο γίνεται πιο ομοιογενείς. Οι παρατηρήσεις διαφορετικών ομάδων θα πρέπει να διαφέρουν όσο γίνεται περισσότερο.

Παραδείγματα Εφαρμογών

- ✓ Στη Βιολογία: για την κατάταξη διαφορετικών ειδών ζώων σε ομάδες με βάση κάποια χαρακτηριστικά τους.
- ✓ Στο Μάρκετινγκ: για την ομαδοποίηση πελατών σύμφωνα με τα στοιχεία που υπάρχουν σχετικά με τις αγοραστικές συνήθειες και τα δημογραφικά χαρακτηριστικά τους.
- ✓ Στην Αρχαιολογία: για την κατάταξη των ευρημάτων μιας ανασκαφής σε ομάδες που π.χ. αντανακλούν διαφορετικές χρονικές περιόδους.
- ✓ Στην Πληροφορική: για τον εντοπισμό και την ομαδοποίηση της συμπεριφοράς των χρηστών Internet ανάλογα με τον τρόπο με τον οποίο σερφάρουν.

→ Πληθώρα εφαρμογών σε κάθε επιστήμη

Βασικές Έννοιες

- ✓ Η Έννοια της Απόστασης
- ✓ Η Έννοια της Ομοιότητας

➢Είναι δύο αντίθετες έννοιες: παρατηρήσεις που είναι όμοιες θα έχουν μεγάλη ομοιότητα και μικρή απόσταση.

➢Οι έννοιες αυτές ποσοτικοποιούν αυτό που στην καθημερινή γλώσσα εννοούν.

➢Μας επιτρέπουν να μετρήσουμε πόσο μοιάζουν οι παρατηρήσεις μεταξύ τους και επομένως να τις τοποθετήσουμε στην ίδια ομάδα.

Διάφορες Προσεγγίσεις για Ομαδοποίηση

- ✓ Ιεραρχικές Μέθοδοι

  Ξεκινάμε με κάθε παρατήρηση να είναι από μόνη της μια ομάδα. Σε κάθε βήμα ενώνουμε τις 2 παρατηρήσεις που έχουν την μικρότερη απόσταση. Αν 2 παρατηρήσεις έχουν ενωθεί σε προηγούμενο βήμα, ενώνουμε μια προϋπάρχουσα ομάδα με μια παρατήρηση μέχρι να φτιάξουμε μια ομάδα. Κοιτώντας τα αποτελέσματα, διαλέγουμε πόσες ομάδες τελικά προκύπτουν.

- ✓ K-Means

  Ο αριθμός των ομάδων είναι γνωστός από πριν. Με έναν επαναληπτικό αλγόριθμο μοιράζουμε τις παρατηρήσεις στις ομάδες ανάλογα με το ποια ομάδα είναι πιο κοντά στην παρατήρηση.

Οι δύο αυτές μέθοδοι βασίζονται σε αλγοριθμικές λύσεις.

- ✓ Στατιστικές Μέθοδοι

  Ξεκινώντας από κάποιες υποθέσεις κατατάσσουμε τις παρατηρήσεις. Έχουν αρκετά υπολογιστικά προβλήματα και γι' αυτόν δεν προσφέρονται από πολλά στατιστικά πακέτα που χρησιμοποιούνται στην πράξη.

Επιπρόσθετοι Σκοποί

- ✓ Απόκτηση γνώσης σχετικά με τα δεδομένα.
- ✓ Διερεύνηση σχέσεων στα δεδομένα.
- ✓ Μείωση των διαστάσεων του προβλήματος.
- ✓ Δημιουργία και έλεγχο υποθέσεων σχετικά με τα δεδομένα.
- ✓ Πρόβλεψη καινούργιων τιμών.

Η Απόσταση

Έννοια της Απόστασης

Σκοπός της Απόστασης είναι να μετρήσει πόσο απέχουν δύο παρατηρήσεις, να ποσοτικοποιήσει δηλαδή αν μοιάζουν ή όχι οι παρατηρήσεις.

Έστω δύο μεταβλητές, ηλικία και βάρος και δύο παρατηρήσεις $y=(y_1,y_2)$ και $x=(x_1,x_2)$. Η απόσταση ανάμεσα στις δύο παρατηρήσεις δίνεται από την ευκλείδεια απόσταση:

$$d(x,y)=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2}$$

Και γενικά για p μεταβλητές:

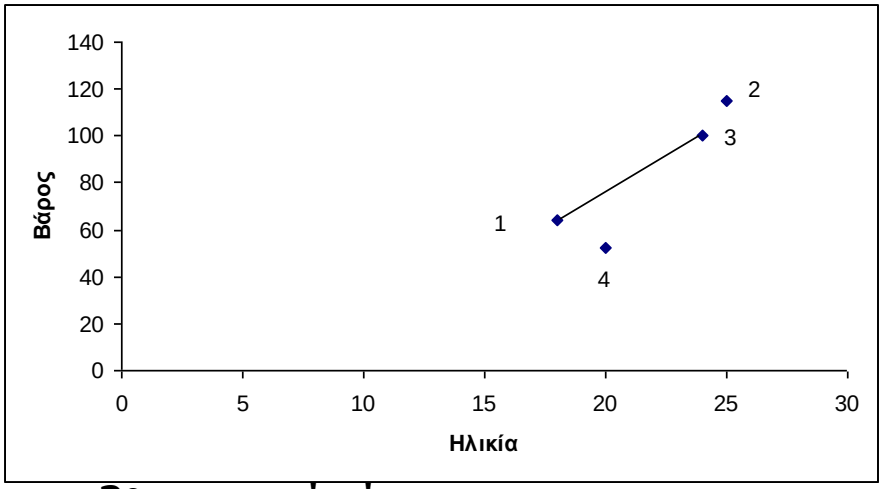
$$d(x,y)=\sqrt{\sum_{i=1}^{p}(x_i-y_i)^2}$$

Η Απόσταση

Έννοια της Απόστασης

Έστω 4 φοιτητές για τους οποίους ξέρουμε την ηλικία και το βάρος:

Φοιτητής	Ηλικία	Βάρος
1	18	64
2	25	115
3	24	100
4	20	52



Η απόσταση ανάμεσα στον 1ο και τον 3ο φοιτητή είναι:

$$\sqrt{(18-24)^2+(64-100)^2}=\sqrt{1332}$$

Η Ευκλείδια Απόσταση δεν φαίνεται να είναι καλό μέτρο απόστασης

Παρατηρούμε ότι:

-Η κλίμακα είναι διαφορετική στις παρατηρήσεις.

-Η απόσταση καθορίζεται σε ένα μεγάλο βαθμό από το βάρος

Η Απόσταση

Έννοια της Απόστασης

Θα μιλάμε για ΑΠΟΣΤΑΣΗ όταν έχουμε μια συνάρτηση που μετρά το πόσο απέχουν (διαφέρουν) μεταξύ τους δύο παρατηρήσεις.

Υπάρχουν διάφορα μέτρα απόστασης. Τα μέτρα απόστασης χωρίζονται σε ομάδες ανάλογα με το είδος των δεδομένων στα οποία μπορούν να εφαρμοστούν.

Η επιλογή του μέτρου απόστασης εξαρτάται από τη φύση των δεδομένων, τη μέθοδο που θα χρησιμοποιήσουμε αλλά και τον σκοπό της ανάλυσης.

Προβλήματα της Ανάλυσης κατά Συστάδες

- Όποια μέθοδος και αν επιλεγεί, ο ερευνητής μπορεί να λειτουργήσει υποκειμενικά σε πολλά σημεία, με αποτέλεσμα από τα ίδια δεδομένα να προκύπτουν αντικρουόμενα συμπεράσματα.

- Όταν στα δεδομένα υπάρχουν ομοιογενείς ομάδες τότε οποιαδήποτε μέθοδος και αν χρησιμοποιηθεί θα καταφέρει να τις αναγνωρίσει. Επομένως, οι αντιφατικές λύσεις είναι μια ένδειξη ότι δεν υπάρχει η κατάλληλη δομή στα δεδομένα.

Προβλήματα της Ανάλυσης κατά Συστάδες

Ποιες μεταβλητές πρέπει να χρησιμοποιηθούν?

-Δεν υπάρχει κάποιος τρόπος που να οδηγεί στην επιλογή των μεταβλητών.

-Εάν δεν υπάρχει εμπειρία ή θεωρητικός λόγος για την επιλογή συγκεκριμένων μεταβλητών τις χρησιμοποιούμε όλες.

-Εναλλακτικά επιλέγουμε εκείνες που μπορούν να δημιουργήσουν ομοιογενείς ομάδες.

-Εκ των υστέρων μπορούμε να δούμε τις «αδιάφορες» για την ανάλυση μεταβλητές, να τις αφαιρέσουμε και να ξαναομαδοποιήσουμε τα δεδομένα.

-Ο μετασχηματισμός των δεδομένων για την τυποποίηση των μεταβλητών μπορεί να οδηγήσει σε χάσιμο πληροφορίας.

Προβλήματα της Ανάλυσης κατά Συστάδες

Ποια απόσταση / ομοιότητα να χρησιμοποιήσουμε ?

-Η επιλογή της απόστασης έχει να κάνει με τη μέθοδο που θα χρησιμοποιήσουμε και τον τύπο των δεδομένων.

-Ο σκοπός της ανάλυσης και επιμέρους χαρακτηριστικά είναι επίσης σημαντικά.

Πόσες ομάδες θα φτιάξουμε ?

-Οποιοσδήποτε λογικός αριθμός μπορεί να χρησιμοποιηθεί.

-Ο προσδιορισμός εξαρτάται και από τη μορφή των δεδομένων.

Προβλήματα της Ανάλυσης κατά Συστάδες

Ποια μέθοδο θα χρησιμοποιήσουμε ?

-Οι ιεραρχικές μέθοδοι δεν είναι καλό να χρησιμοποιούνται για μεγάλο πλήθος δεδομένων γιατί απαιτούν πολύ χρόνο και υπολογιστική ισχύ. Επιπλέον δημιουργούνται ομάδες με ανομοιογενές μέγεθος.

-Η μέθοδος K-means αποφεύγει αυτά τα προβλήματα, εξαρτάται πολύ από τις αρχικές τιμές που θα χρησιμοποιήσουμε.

Η Μέθοδος K-Means

Ο αλγόριθμος K-Means ανήκει σε μια κατηγορία αλγορίθμων γνωστούς ως αλγόριθμοι διαμέρισης (partitioning algorithms).

Στην ουσία διαμερίζουν το πολυεπίπεδο που δημιουργούν τα δεδομένα σε περιοχές και αντιστοιχούν μια περιοχή σε κάθε ομάδα.

Η Μέθοδος K-Means

Ο Αλγόριθμος

Η μέθοδος θεωρεί ότι ο αριθμός των ομάδων που θα προκύψουν είναι γνωστός εκ των προτέρων.

Άρση περιορισμού:

-Είτε τρέχουμε τον αλγόριθμο με διαφορετικές επιλογές ως προς το πλήθος των ομάδων

-Είτε πρέπει με κάποιον τρόπο να έχουμε καταλήξει στον αριθμό των ομάδων

Ο αλγόριθμος δουλεύει ικανοποιητικά για μεγάλα σετ δεδομένων (Quick Clustering).

Συνήθως χρησιμοποιείται η ευκλείδεια απόσταση. Για την χρησιμοποίηση άλλης απόστασης απαιτούνται ειδικοί μετασχηματισμοί στα δεδομένα.

Η Μέθοδος K-Means

Ο Αλγόριθμος

Βήματα Αλγορίθμου:

1. Βρίσκουμε τα αρχικά κέντρα.
2. Κατατάσσουμε κάθε παρατήρηση στην ομάδα της οποίας το κέντρο έχει τη μικρότερη απόσταση από την παρατήρηση.
3. Από τις παρατηρήσεις που είναι μέσα στην ομάδα υπολογίζουμε τα νέα κέντρα.
4. Αν τα νέα κέντρα δεν διαφέρουν από τα παλιά σταματάμε αλλιώς πηγαίνουμε πίσω στο βήμα 2.

Η Μέθοδος K-Means

Χαρακτηριστικά του Αλγορίθμου

- Είναι ιδιαίτερα γρήγορος. Επομένως είναι χρήσιμος για μεγάλα σετ δεδομένων και δεν χρειάζεται μεγάλη υπολογιστική ισχύ.
- Ελαχιστοποιεί το άθροισμα των τετραγωνικών αποστάσεων των παρατηρήσεων από τα κέντρα των ομάδων που ανήκουν. Η λύση έχει ομάδες με ίδιο αριθμό παρατηρήσεων.
- Μειονέκτημα: εξαρτάται από τις αρχικές τιμές, οι οποίες μπορεί να οδηγήσουν σε διαφορετική ομαδοποίηση. Λύση σε αυτό βρίσκεται εάν τρέξουμε τον αλγόριθμο με διαφορετικές αρχικές τιμές ώστε να είμαστε σίγουροι ότι δεν παγιδεύτηκε σε κάποια μη βέλτιστη λύση.
- Πρόβλημα αποτελεί η επιλογή του αριθμού των ομάδων. Μια τακτική είναι η ομαδοποίηση με διαφορετικό αριθμό ομάδων και στο τέλος η επιλογή της ομάδας που είναι η βέλτιστη.

Η Μέθοδος K-Means

Χαρακτηριστικά του Αλγορίθμου

Χρήσιμες Στρατηγικές:

- ✓ Η επιλογή των αρχικών κέντρων πρέπει να γίνεται ώστε αυτά να είναι όσο πιο μακριά μεταξύ τους.
- ✓ Για την αποφυγή μεγάλου αριθμού ομαδοποιήσεων μελετάμε τη λύση που έχουμε, προσπαθώντας να ενώσουμε ή να διαλύσουμε ομάδες.
- ✓ Η βέλτιστη λύση είναι σπάνιο να επιτευχθεί με μια μόνο επιλογή αριθμού ομάδων. Δοκιμάζουμε διάφορες επιλογές και χρησιμοποιούμε και τη διαίσθησή μας ώστε να επιτύχουμε καλύτερη ομαδοποίηση.

Η Μέθοδος K-Means

Χαρακτηριστικά του Αλγορίθμου

- ✓ Η δυναμική του αλγόριθμου είναι ότι με λίγες επαναλήψεις πλησιάζει κοντά στην τελική λύση. Επομένως δεν είναι απαραίτητος μεγάλος αριθμός επαναλήψεων.
- ✓ Η μέθοδος βασίζεται στην ευκλείδεια απόσταση. Μπορούν να χρησιμοποιηθεί όμως κάθε είδους απόσταση. Πρόβλημα αποτελεί ο ορισμός του μέσου της ομάδας σε μη συνεχή δεδομένα.
  - Σε κατηγορικά δεδομένα με κατάταξη μπορούμε να χρησιμοποιήσουμε το διάνυσμα των διαμέσων.
  - Σε ονομαστικά δεδομένα μπορούμε να χρησιμοποιήσουμε την κορυφή (επικρατούσα τιμή).
  - Σε μικτού τύπου δεδομένα το κέντρο μπορεί να αποτελείται από τις κορυφές των κατηγορικών μεταβλητών και τους μέσους των συνεχών.

Η Μέθοδος K-Means

Ο Αλγόριθμος

Η μέθοδος θεωρεί ότι ο αριθμός των ομάδων που θα προκύψουν είναι γνωστός εκ των προτέρων.

Βήματα Αλγορίθμου:

1. Βρίσκουμε τα αρχικά κέντρα.
2. Κατατάσσουμε κάθε παρατήρηση στην ομάδα της οποίας το κέντρο έχει τη μικρότερη απόσταση από την παρατήρηση.
3. Από τις παρατηρήσεις που είναι μέσα στην ομάδα υπολογίζουμε τα νέα κέντρα.
4. Αν τα νέα κέντρα δεν διαφέρουν από τα παλιά σταματάμε αλλιώς πηγαίνουμε πίσω στο βήμα 2.

Εφαρμογή της Μεθόδου K-Means

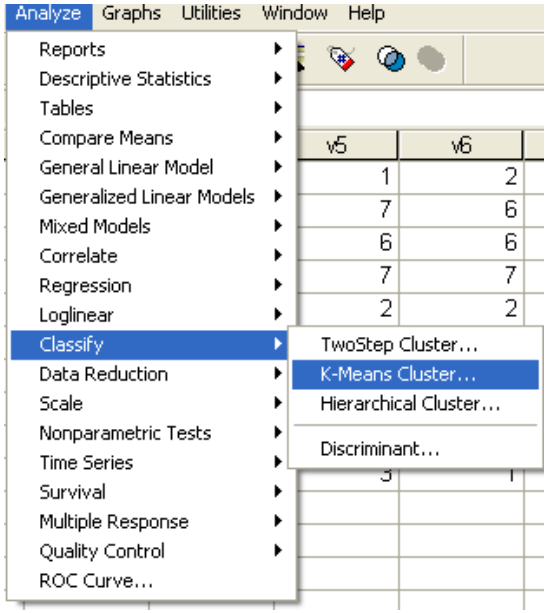
Περιγραφή προβλήματος

Καταγράφεται η στάση 10 καταναλωτών απέναντι στην άθληση σε γυμναστήρια. Εκφράζεται ο βαθμός συμφωνίας ή διαφωνίας τους σε μια επταβάθμια κλίμακα στις εξής απόψεις:

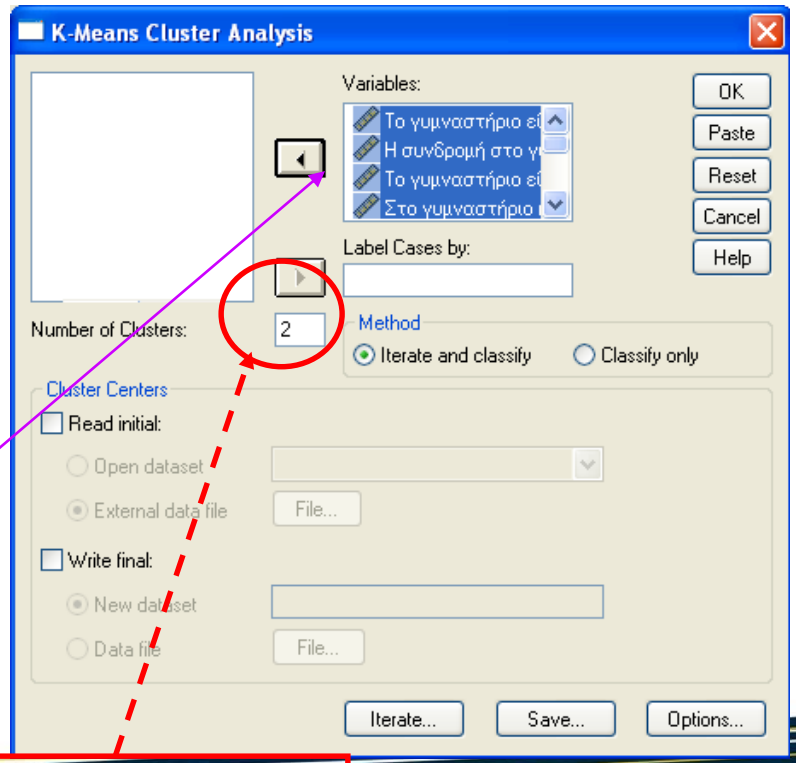
-Το γυμναστήριο είναι διασκέδαση (v1)
-Η συνδρομή στο γυμναστήριο είναι χαμένα χρήματα (v2)
-Το γυμναστήριο είναι καλό για την υγεία μου (v3)
-Στο γυμναστήριο περνάω όμορφα (v4)
-Δεν μου αρέσουν τα γυμναστήρια (v5)
-Υπάρχουν καλύτεροι τρόποι άθλησης από το γυμναστήριο (v6)

v1	v2	v3	v4	v5	v6
7	2	6	7	1	2
3	7	2	3	7	6
2	6	3	2	6	6
1	7	2	2	7	7
6	1	7	7	2	2
6	2	6	5	3	2
2	6	1	1	7	7
7	3	7	7	4	3
7	2	6	6	2	2
7	1	6	7	3	1

Εφαρμογή της Μεθόδου K-Means



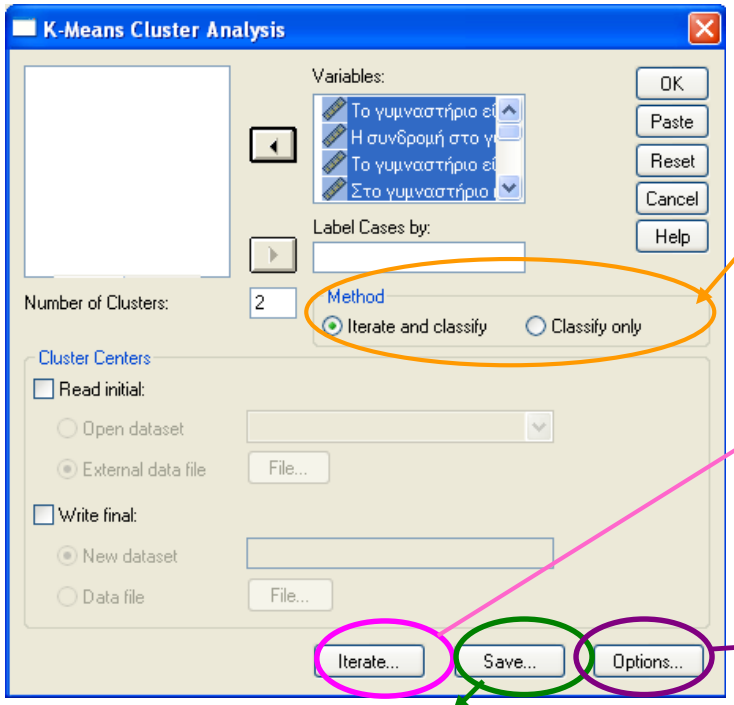
Υπενθυμίζουμε ότι η μέθοδος χρησιμοποιείται όταν ο αριθμός των ομάδων είναι γνωστός.



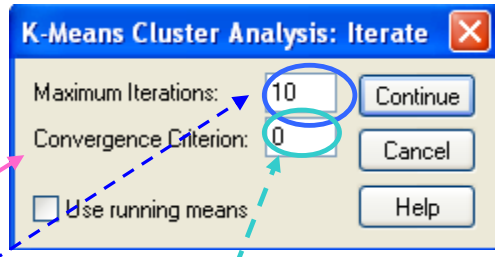
Μεταφέρουμε τις μεταβλητές για τις οποίες θα εφαρμοστεί η μέθοδος.

Ο αριθμός ομάδων είναι προεπιλεγμένος

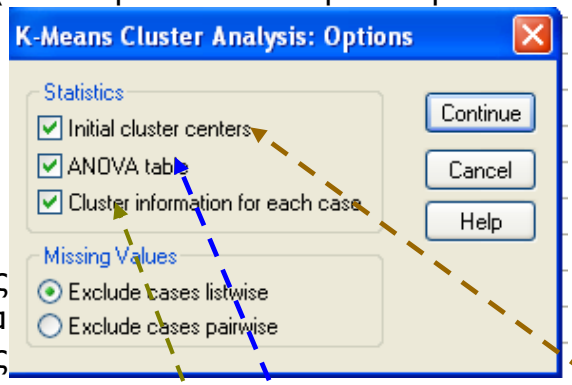
Εφαρμογή της Μεθόδου K-Means

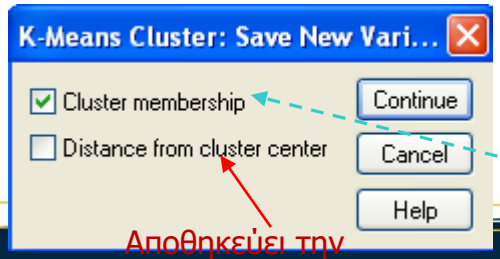


Επιλέγουμε τη μέθοδο που θα χρησιμοποιήσουμε (εάν τα αρχικά κέντρα θα ξαναυπολογιστούν ή όχι).



Επιλέγουμε τα κριτήρια τερματισμού του αλγορίθμου. Μπορούμε να επιλέξουμε είτε να σταματήσει μετά από ένα συγκεκριμένο **αριθμό**, είτε όταν η μεγαλύτερη απόσταση ανάμεσα σε διαδοχικά κέντρα όλων των ομάδων γίνει **0**.





Αποθηκεύει την ευκλείδεια απόσταση

Επιλέγουμε τις καινούργιες μεταβλητές που θέλουμε να δημιουργήσουμε. Επιλέγοντας Cluster Membership δημιουργείται μια νέα στήλη όπου σε κάθε παρατήρηση θα δίνεται η τιμή της ομάδας που την κατατάξαμε.

Επιλέγουμε να εμφανιστούν τα αρχικά κέντρα, ο πίνακας ANOVA ώστε να δούμε ποιες έχουν πληροφορία για την ομαδοποίηση που κάναμε και πληροφορίες για κάθε παρατήρηση

Εφαρμογή της Μεθόδου K-Means

Initial Cluster Centers

	Cluster	
	1	2
Το γυμναστήριο είναι διασκέδαση	7	1
Η συνδρομή στο γυμναστήριο είναι χαμένα χρήματα	1	7
Το γυμναστήριο είναι καλό για την υγεία μου	6	2
Στο γυμναστήριο περνάω όμορφα	7	2
Δεν μου αρέσουν τα γυμναστήρια	3	7
Υπάρχουν καλύτεροι τρόποι άθλησης από το γυμναστήριο	1	7

Περιέχει τα αρχικά κέντρα των ομάδων, αυτά από όπου ξεκινά ο αλγόριθμος. Οι τιμές στον πίνακα αποτελούν τους μέσους κάθε μεταβλητής μέσα σε κάθε αρχικό κέντρο.

Iteration History[a]

	Change in Cluster Centers	
Iteration	1	2
1	1,555	1,250
2	,000	,000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is ,000. The current iteration is 2. The minimum distance between initial centers is 12,845.

Περιέχει πληροφορίες για το πώς μετακινείται ο αλγόριθμος σε κάθε επανάληψη. Η τιμή **1,555** είναι η απόσταση ανάμεσα στο κέντρο της ομάδας στην τρέχουσα επανάληψη με το κέντρο της ομάδας κατά την προηγούμενη. Όταν η απόσταση **μηδενιστεί** σταματά ο αλγόριθμός.

Εφαρμογή της Μεθόδου K-Means

Cluster Membership

Case Number	Cluster	Distance
1	1	1,658
2	2	1,601
3	2	1,436
4	2	1,250
5	1	1,443
6	1	1,756
7	2	1,601
8	1	2,327
9	1	,866
10	1	1,555

Δείχνει την τοποθέτηση των παρατηρήσεων, δηλαδή την ομάδα στην οποία ανήκει ο καθένας από τους 10 καταναλωτές. Η στήλη Distance δείχνει την απόσταση της κάθε παρατήρησης από τα αρχικά κέντρα.

Final Cluster Centers

	Cluster	
	1	2
Το γυμναστήριο είναι διασκέδαση	7	2
Η συνδρομή στο γυμναστήριο είναι χαμένα χρήματα	2	7
Το γυμναστήριο είναι καλό για την υγεία μου	6	2
Στο γυμναστήριο περνάω όμορφα	7	2
Δεν μου αρέσουν τα γυμναστήρια	3	7
Υπάρχουν καλύτεροι τρόποι άθλησης από το γυμναστήριο	2	7

Περιέχει τα κέντρα των ομάδων που βρέθηκαν αφού σταμάτησε ο αλγόριθμος.

Distances between Final Cluster Centers

Cluster	1	2
1		10,995
2	10,995	

Περιέχει την ευκλείδεια απόσταση μεταξύ των τελικών κέντρων των ομάδων

Εφαρμογή της Μεθόδου K-Means

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Το γυμναστήριο είναι διασκέδαση	52,267	1	,417	8	125,440	,000
Η συνδρομή στο γυμναστήριο είναι χαμένα χρήματα	52,267	1	,479	8	109,078	,000
Το γυμναστήριο είναι καλό για την υγεία μου	45,067	1	,417	8	108,160	,000
Στο γυμναστήριο περνάω όμορφα	48,600	1	,688	8	70,691	,000
Δεν μου αρέσουν τα γυμναστήρια	43,350	1	,781	8	55,488	,000
Υπάρχουν καλύτεροι τρόποι άθλησης από το γυμναστήριο	48,600	1	,375	8	129,600	,000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Υψηλές τιμές F υποδηλώνουν μεταβλητές σημαντικές για τον διαχωρισμό των ομάδων. Χαμηλές τιμές F (κοντά στο 1,0) υποδηλώνουν μεταβλητές που δεν είναι στατιστικά σημαντικές

Ο πίνακας περιέχει την Ανάλυση Διακύμανσης για το αν διαφέρουν οι μέσες τιμές ανάμεσα στις ομάδες. Μεταβλητές με καλή ικανότητα να ξεχωρίζουν τις παρατηρήσεις πρέπει να είναι στατιστικά σημαντικές.

ΠΡΟΣΟΧΗ: έχει περιγραφικό σκοπό για να συγκρίνουμε μεταβλητές μεταξύ τους, καθώς ο αλγόριθμος έχει σχεδιαστεί για να μεγιστοποιεί την ελεγχοσυνάρτηση F και επομένως η χρήση του είναι μάλλον ενδεικτική.

Εφαρμογή της Μεθόδου K-Means

Number of Cases in each Cluster

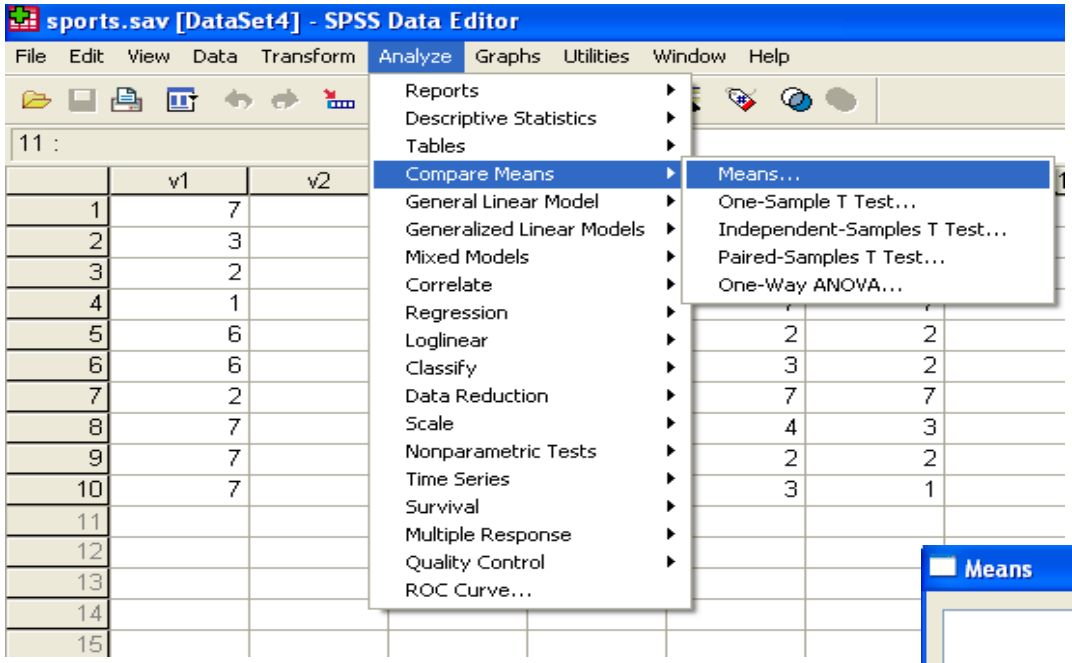
Cluster	1	6,000
	2	4,000
Valid		10,000
Missing		,000

Παρουσιάζει πόσες παρατηρήσεις έχει κάθε ομάδα.
Εδώ, η 1η ομάδα έχει 6 και η 2η 4 παρατηρήσεις.

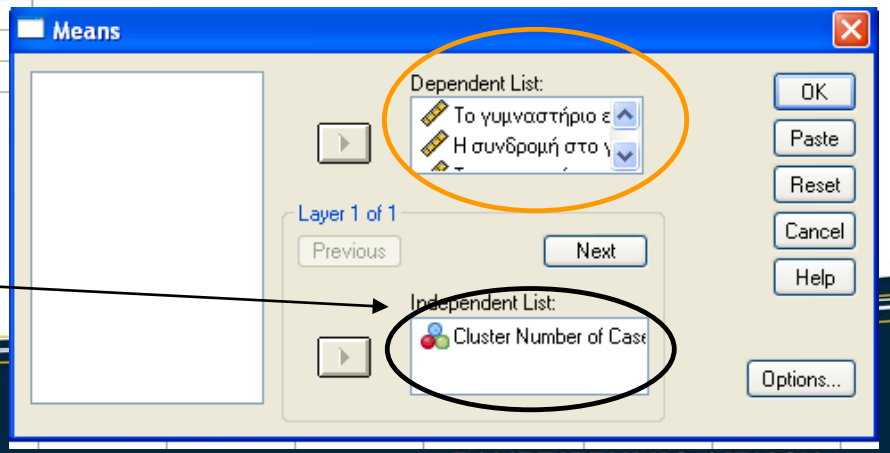
v1	v2	v3	v4	v5	v6	QCL_1
7	2	6	7	1	2	1
3	7	2	3	7	6	2
2	6	3	2	6	6	2
1	7	2	2	7	7	2
6	1	7	7	2	2	1
6	2	6	5	3	2	1
2	6	1	1	7	7	2
7	3	7	7	4	3	1
7	2	6	6	2	2	1
7	1	6	7	3	1	1

Έχει δημιουργηθεί μια νέα μεταβλητή, η οποία λαμβάνει (στη συγκεκριμένη περίπτωση) τις τιμές 1: ανήκει στην 1η Ομάδα και 2: ανήκει στην 2η Ομάδα.

Εφαρμογή της Μεθόδου K-Means



Οι αρχικές μεταβλητές



Η νέα μεταβλητή

Εφαρμογή της Μεθόδου K-Means

Report

Mean

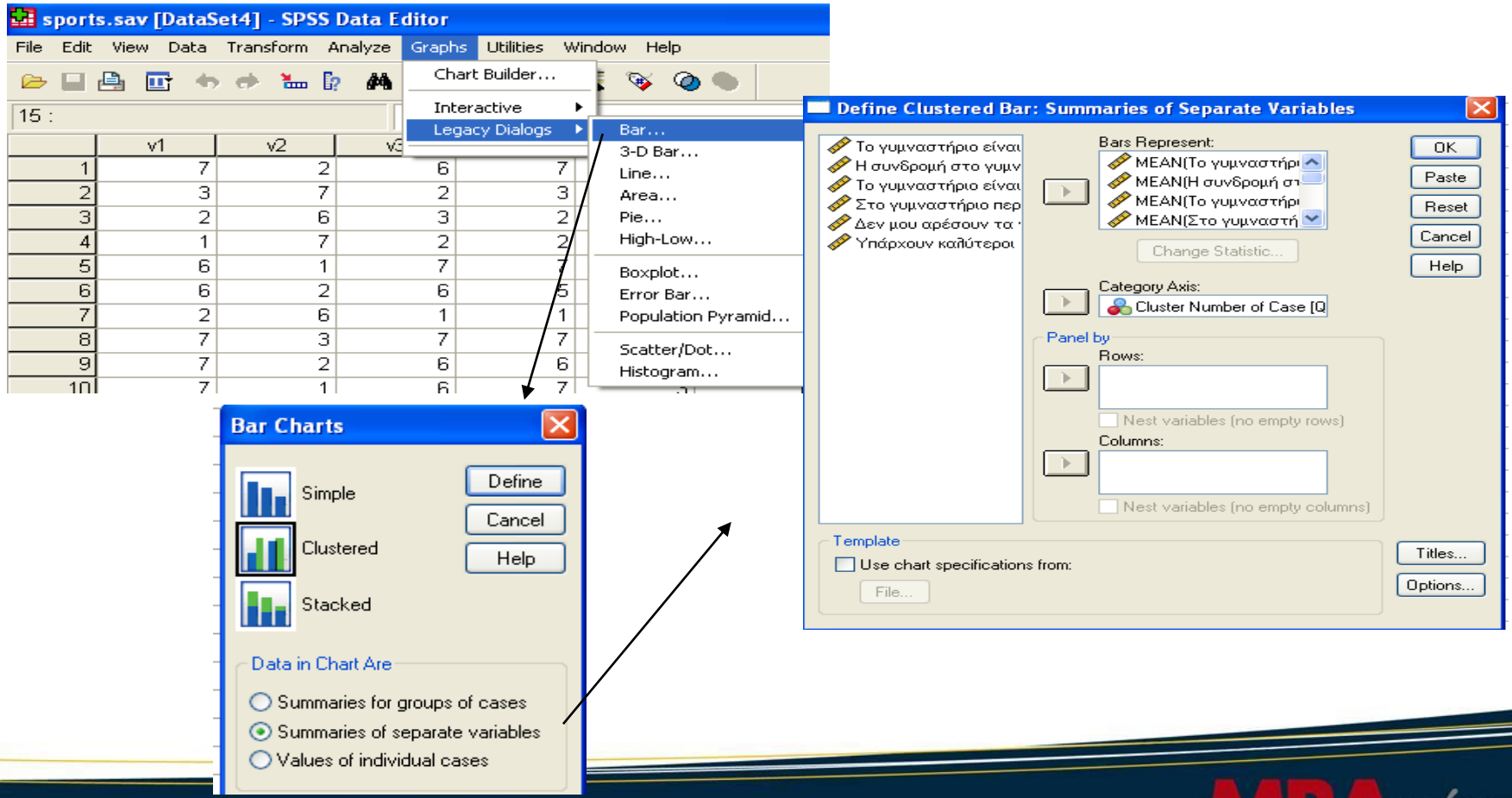
Cluster Number of Case	Το γυμναστήριο είναι διασκέδαση	Η συνδρομή στο γυμναστήριο είναι χαμένα χρήματα	Το γυμναστήριο είναι καλό για την υγεία μου	Στο γυμναστήριο περνάω όμορφα	Δεν μου αρέσουν τα γυμναστήρια	Υπάρχουν καλύτεροι τρόποι άθλησης από το γυμναστήριο
1	6,67	1,83	6,33	6,50	2,50	2,00
2	2,00	6,50	2,00	2,00	6,75	6,50
Total	4,80	3,70	4,60	4,70	4,20	3,80

Παρέχονται πληροφορίες για τους μέσους της κάθε μεταβλητής.

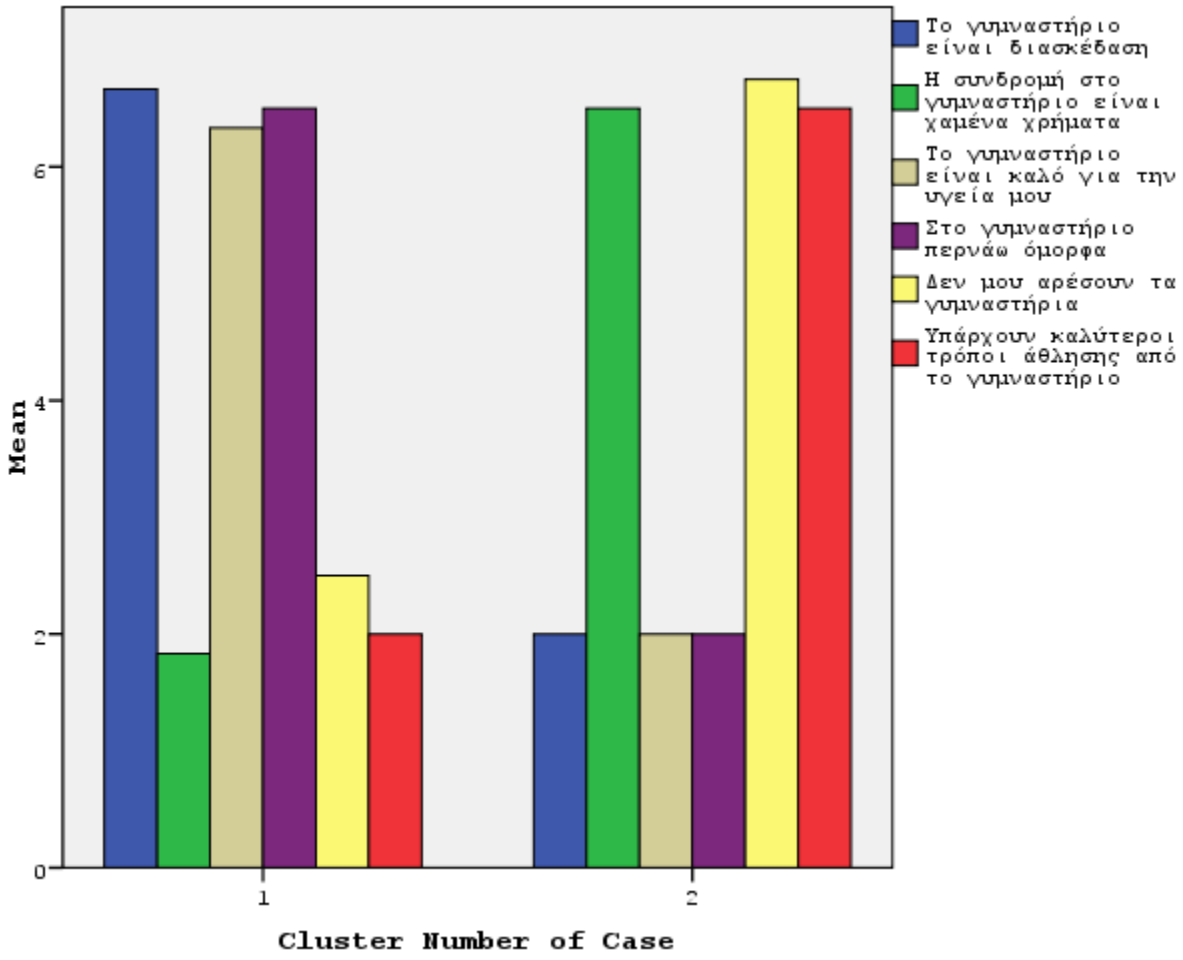
Η 1η Ομάδα εμπεριέχει υψηλές τιμές μέσων για τις μεταβλητές v1, v3, v4. Άρα οι καταναλωτές που ανήκουν στην 1η ομάδα έχουν κοινό χαρακτηριστικό τη θετική στάση απέναντι στο γυμναστήριο. Η ομάδα αυτή θα μπορούσε να ονομαστεί «Αθλητικοί Καταναλωτές».

Η 2η Ομάδα εμπεριέχει υψηλές τιμές μέσων για τις μεταβλητές v2, v5,v6. Άρα οι καταναλωτές που ανήκουν στην 2η ομάδα έχουν αρνητική στάση απέναντι στο γυμναστήριο. Η ομάδα αυτή θα μπορούσε να ονομαστεί «Αποστάτες Καταναλωτές».

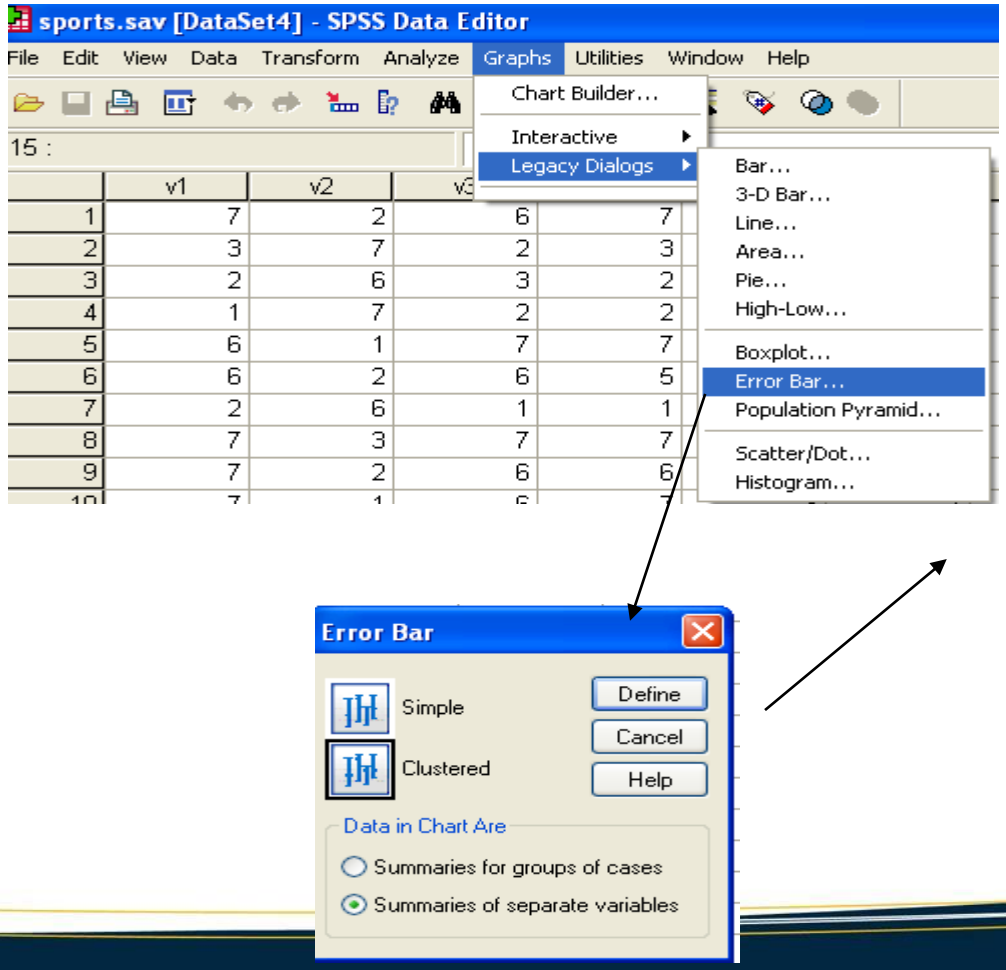
Εφαρμογή της Μεθόδου K-Means



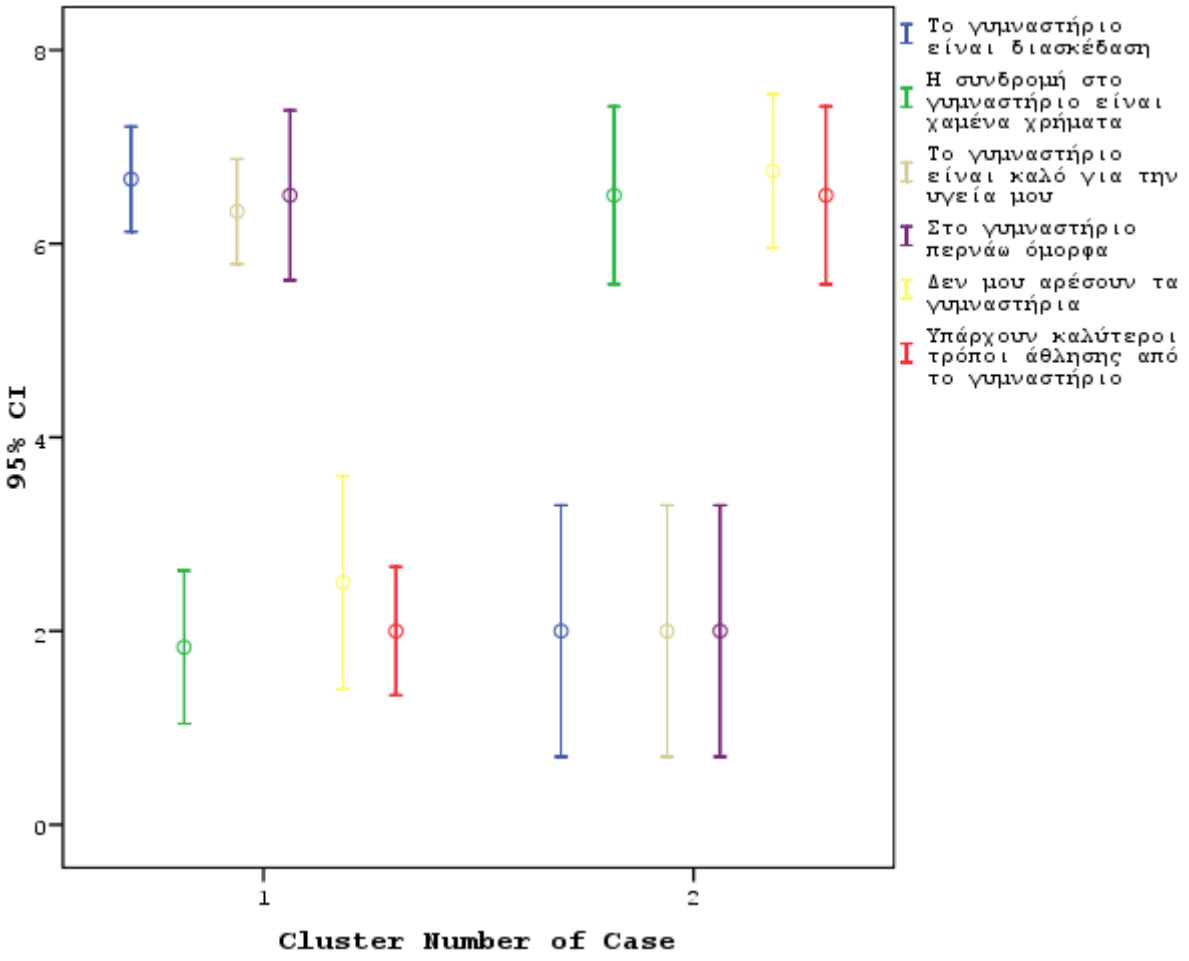
Εφαρμογή της Μεθόδου K-Means



Εφαρμογή της Μεθόδου K-Means



Εφαρμογή της Μεθόδου K-Means



Ιεραρχική Ομαδοποίηση

- Ο αριθμός των ομάδων δεν είναι γνωστός από πριν.
- Οι μέθοδοι λειτουργούν ιεραρχικά. Ξεκινούν χρησιμοποιώντας κάθε παρατήρηση σαν μια ομάδα και σε κάθε βήμα ενώνουν σε ομάδες τις παρατηρήσεις που βρίσκονται πιο κοντά.
- Οι ιεραρχικοί αλγόριθμοι δουλεύουν είτε προς τα εμπρός είτε προς τα πίσω.
  - ✓Κάποιοι αλγόριθμοι (αλγόριθμοι divisive) ξεκινούν με όλες τις παρατηρήσεις σε μια ομάδα. Η παρατήρηση που βρίσκεται πιο μακριά από τις υπόλοιπες φεύγει από την ομάδα και σχηματίζει μια καινούργια ομάδα από μόνη της. Βρίσκουμε τη δεύτερη πιο απομακρυσμένη παρατήρηση, η οποία μπορεί να σχηματίζει μια ομάδα από μόνη της ή να πάει στην νέα ομάδα.
  - ✓Οι αλγόριθμοι agglomerative ξεκινούν με κάθε παρατήρηση ως μια ομάδα και ενώνουν στη συνέχεια ομάδες που είναι πιο κοντινές. Είναι οι πιο διαδεδομένοι αλγόριθμοι.

Οι ιεραρχικές μέθοδοι χρειάζονται πολύ χρόνο και χώρο και γι’ αυτό είναι ασύμφορες για μεγάλα σετ δεδομένων.

Ιεραρχική Ομαδοποίηση

Επιλογή Μεθόδου

Nearest Neighbour (or singe linkage): Η μέθοδος του κοντινότερου γείτονα υπολογίζει την απόσταση ανάμεσα σε δύο ομάδες ως τη μικρότερη απόσταση από μια παρατήρηση μέσα στην μια ομάδα με κάποια παρατήρηση στην άλλη ομάδα.

Συν: Έχει κάποιες χρήσιμες μαθηματικές ιδιότητες.

Μείον: Παράγει ομάδες που δεν είναι συμπαγείς και συνήθως δημιουργεί μερικές πολύ μεγάλες ομάδες και κάποιες πολύ μικρές.

Furthest Neighbour (or complete linkage): Η μέθοδος του μακρύτερου γείτονα υπολογίζει την απόσταση ανάμεσα σε δυο ομάδες ως τη μεγαλύτερη απόσταση από μια παρατήρηση μέσα στη μια ομάδα με κάποια παρατήρηση στην άλλη ομάδα.

Συν: Οι ομάδες που δημιουργούνται είναι συνήθως συμπαγείς.

Μείον: Αποτυγχάνει να δημιουργήσει κάποιες μικρές μα πολύ συμπαγείς ομάδες.

Ιεραρχική Ομαδοποίηση

Επιλογή Μεθόδου

Average within groups: Η απόσταση είναι ο μέσος όλων των αποστάσεων που προκύπτουν όταν ενώσουμε τις δύο ομάδες.

Centroid: Η απόσταση υπολογίζεται ως η απόσταση των κέντρων των ομάδων.

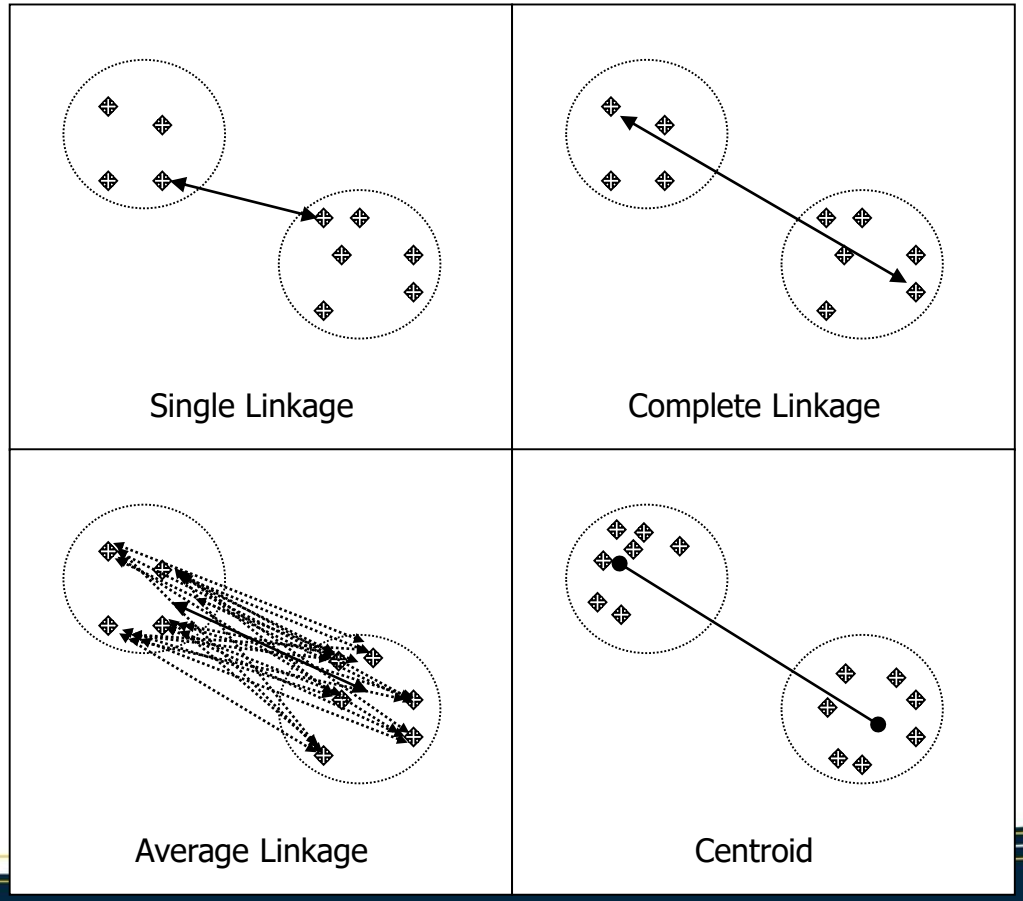
Συν: Έχει μερικές καλές ιδιότητες και παράγει συμπαγείς και ελλειπτικές ομάδες.

Ward Method: Είναι σχεδιασμένη να ελαχιστοποιεί τη διακύμανση μέσα στις ομάδες. Για κάθε παρατήρηση μπορούμε να υπολογίσουμε την απόστασή της από το κέντρο της ομάδας. Αν αθροίσουμε για όλες τις ομάδες έχουμε το συνολικό άθροισμα. Αυτό το άθροισμα είναι αρχικά 0. Σε κάθε βήμα ενώνουμε τις ομάδες οι οποίες οδηγούν στη μικρότερη αύξηση του συνολικού αθροίσματος αποστάσεων.

Συν: Έχει μερικές καλές ιδιότητες και συνήθως δημιουργεί ομάδες με παρόμοιο αριθμό παρατηρήσεων.

Ιεραρχική Ομαδοποίηση

Επιλογή Μεθόδου



Ιεραρχική Ομαδοποίηση

Επιλογή Μεθόδου

- Για όλες τις μεθόδους χρειαζόμαστε έναν πίνακα αποστάσεων.
- Για τη μέθοδο centroid χρειαζόμαστε το κέντρο κάθε ομάδας. Αν τα στοιχεία μας δεν είναι συνεχή το κέντρο δεν μπορεί να είναι απλώς οι μέσοι των μεταβλητών και σε αυτή την περίπτωση χρησιμοποιούμε την κορυφή ή τη διάμεσο.
- Παρόμοια προβλήματα έχει και η μέθοδος ward.

Καλύτερη επίδοση: Μέθοδος Ward και Average between groups.
Χειρότερη επίδοση: Μέθοδος κοντινότερου γείτονα

Ιεραρχική Ομαδοποίηση

Επιλογή Μεθόδου

✓Δεν είναι ξεκάθαρο ποια μέθοδος είναι η καλύτερη.
✓Αν οι ομάδες είναι αρκετά διαφορετικές μεταξύ τους, κάθε μέθοδος θα βρει τη σωστή ομαδοποίηση.
✓Κάθε μέθοδος δουλεύει καλύτερα με συγκεκριμένη μορφή δεδομένων.
✓Μέθοδοι που βασίζονται σε τετραγωνικές αποστάσεις (αλγόριθμος K-means, μέθοδος Ward) τείνουν να βρίσκουν ομάδες με περίπου ίδια διακύμανση
✓Οι περισσότερες μέθοδοι αποτυγχάνουν να βρουν ομάδες με περίεργα σχήματα. Σε αυτή την περίπτωση η μέθοδος του κοντινότερου γείτονα είναι πιο αποδοτική.
✓Για μερικές από τις μεθόδους δεν είναι απαραίτητο να υπάρχει μια αύξηση σε κάθε βήμα της απόστασης. Η απόσταση των ομάδων που συγχωνεύονται δεν πρέπει απαραίτητα να είναι αύξουσα.

Ιεραρχική Ομαδοποίηση

Χαρακτηριστικά Αλγορίθμου

Μειονεκτήματα της ιεραρχικής ομαδοποίησης

- Δε συμφέρει από άποψη υπολογιστικού φόρτου για μεγάλα σετ δεδομένων.
- Ομάδες που φτιάχνονται σε αρχικά βήματα δεν μπορούν να χωρίσουν και επομένως οι παρατηρήσεις που ενώνονται σε αρχικά βήματα μένουν για πάντα μαζί.
- Η μέθοδος εφοδιάζει με μια ποικιλία λύσεων, μια για κάθε διαφορετικό αριθμό ομάδων. Συνεπώς απαιτείται ένα κριτήριο για να επιλεγεί η τελική λύση.

Εφαρμογή της Ιεραρχικής Ομαδοποίησης

Περιγραφή προβλήματος

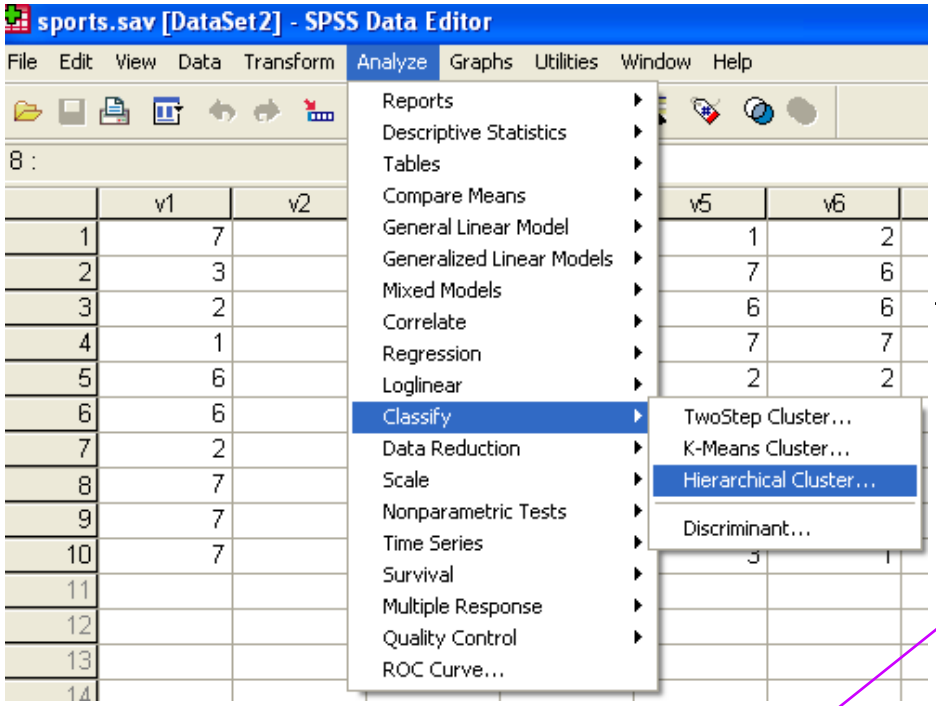
Καταγράφεται η στάση 10 καταναλωτών απέναντι στην άθληση σε γυμναστήρια. Εκφράζεται ο βαθμός συμφωνίας ή διαφωνίας τους σε μια επταβάθμια κλίμακα στις εξής απόψεις:

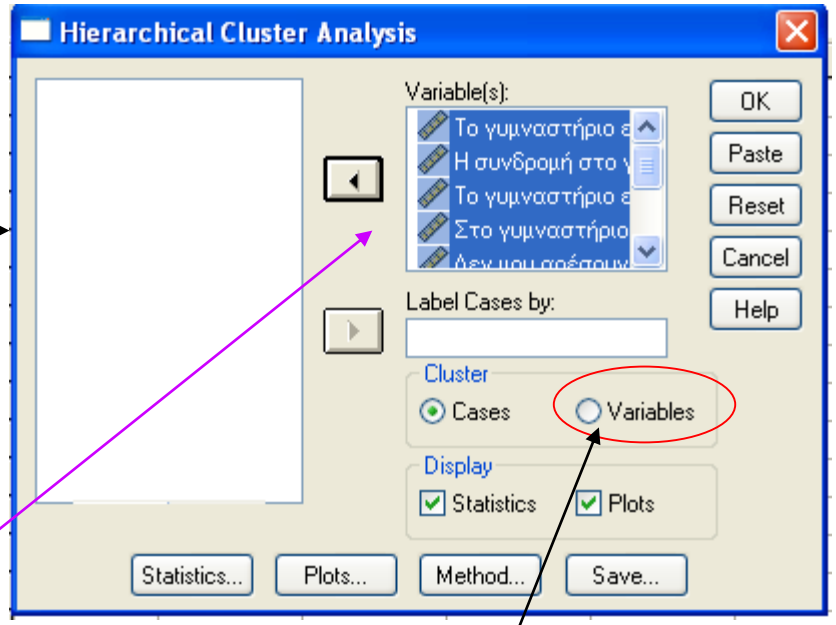
-Το γυμναστήριο είναι διασκέδαση (v1)
-Η συνδρομή στο γυμναστήριο είναι χαμένα χρήματα (v2)
-Το γυμναστήριο είναι καλό για την υγεία μου (v3)
-Στο γυμναστήριο περνάω όμορφα (v4)
-Δεν μου αρέσουν τα γυμναστήρια (v5)
-Υπάρχουν καλύτεροι τρόποι άθλησης από το γυμναστήριο (v6)

v1	v2	v3	v4	v5	v6
7	2	6	7	1	2
3	7	2	3	7	6
2	6	3	2	6	6
1	7	2	2	7	7
6	1	7	7	2	2
6	2	6	5	3	2
2	6	1	1	7	7
7	3	7	7	4	3
7	2	6	6	2	2
7	1	6	7	3	1

Εφαρμογή της Ιεραρχικής Ομαδοποίησης

Υπενθυμίζουμε ότι δεν απαιτείται να είναι γνωστός ο αριθμός των ομάδων.

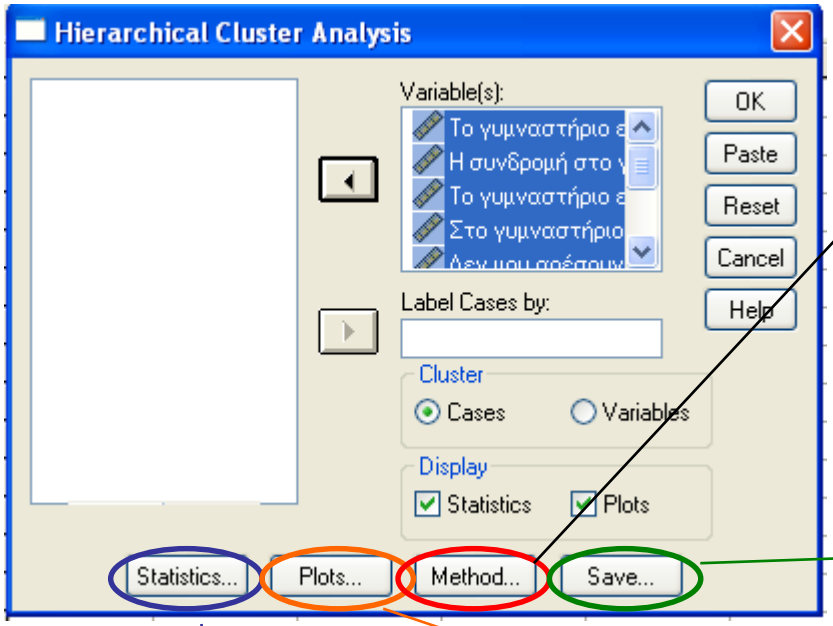




Μεταφέρουμε τις μεταβλητές για τις οποίες θα εφαρμοστεί η μέθοδος.

Δίνεται η δυνατότητα να ομαδοποιηθούν και οι μεταβλητές. Χρειάζεται ΠΡΟΣΟΧΗ και καλό είναι να αποφεύγεται αυτή η επιλογή.

Εφαρμογή της Ιεραρχικής Ομαδοποίησης



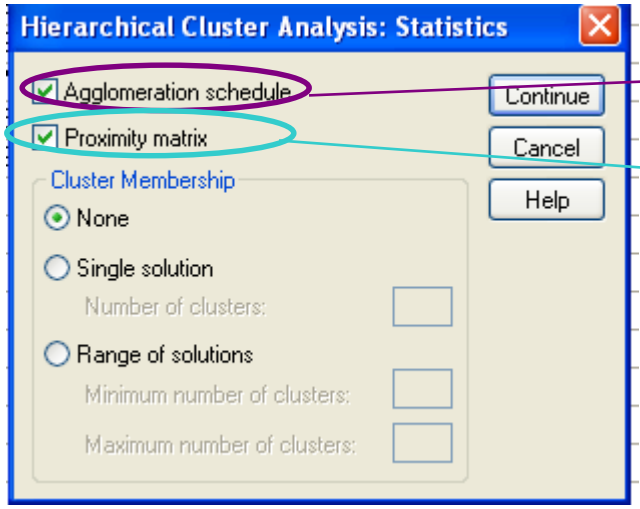
Επιλέγουμε τη μέθοδο και την απόσταση που θα χρησιμοποιήσουμε.

Επιλέγουμε να αποθηκευτούν οι μεταβλητές που θα δείχνουν τη λύση ή τις λύσεις.

Παρέχει επιλογές για τις πληροφορίες που θέλουμε να εμφανιστούν.

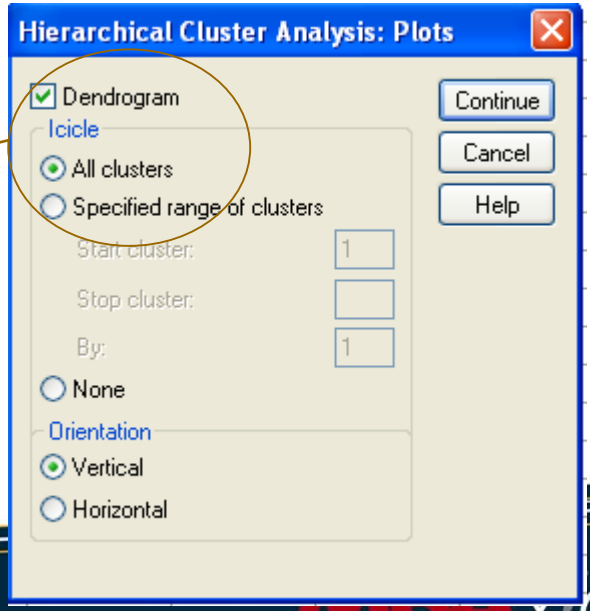
Επιλέγουμε τα γραφήματα που θέλουμε να εμφανιστούν.

Εφαρμογή της Ιεραρχικής Ομαδοποίησης



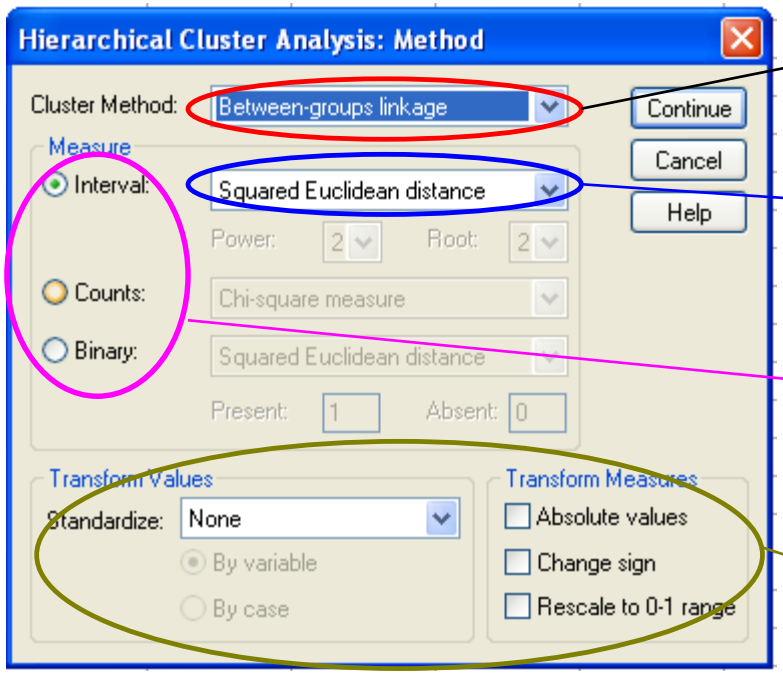
Εμφανίζονται κάποιες ποσότητες που είναι χρήσιμες για να βρούμε τον αριθμό των ομάδων που θα κρατήσουμε.

Εμφανίζεται ο πίνακας αποστάσεων όλων των παρατηρήσεων



Είναι γραφήματα που μπορούν να μας δώσουν γραφικά τη σειρά με την οποία οι παρατηρήσεις ενώνονται για να δημιουργήσουν ομάδες.

Εφαρμογή της Ιεραρχικής Ομαδοποίησης



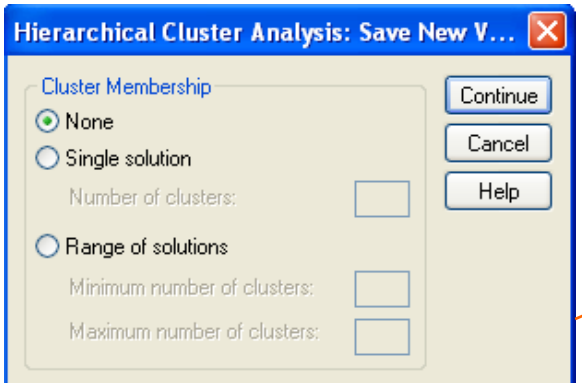
Επιλέγουμε τη μέθοδο που θα χρησιμοποιήσουμε.

Επιλέγουμε το μέτρο απόστασης που θα χρησιμοποιήσουμε.

Τα μέτρα αποστάσεων είναι ομαδοποιημένα ανάλογα με το είδος των δεδομένων.

Επιλέγουμε μετασχηματισμούς των δεδομένων ώστε να μεγαλώσουν οι δυνατές επιλογές.

Εφαρμογή της Ιεραρχικής Ομαδοποίησης



Επιλέγουμε να δημιουργήσουμε μεταβλητές που να δείχνουν που ανήκει κάθε παρατήρηση.

Εφαρμογή της Ιεραρχικής Ομαδοποίησης

Proximity Matrix

Case	Squared Euclidean Distance									
	1	2	3	4	5	6	7	8	9	10
1	,000	125,000	116,000	163,000	4,000	9,000	163,000	12,000	2,000	6,000
2	125,000	,000	5,000	6,000	127,000	86,000	8,000	91,000	107,000	125,000
3	116,000	5,000	,000	5,000	114,000	75,000	7,000	88,000	98,000	118,000
4	163,000	6,000	5,000	,000	161,000	116,000	4,000	127,000	143,000	165,000
5	4,000	127,000	114,000	161,000	,000	7,000	163,000	10,000	4,000	4,000
6	9,000	86,000	75,000	116,000	7,000	,000	114,000	9,000	3,000	7,000
7	163,000	8,000	7,000	4,000	163,000	114,000	,000	131,000	141,000	163,000
8	12,000	91,000	88,000	127,000	10,000	9,000	131,000	,000	8,000	10,000
9	2,000	107,000	98,000	143,000	4,000	3,000	141,000	8,000	,000	4,000
10	6,000	125,000	118,000	165,000	4,000	7,000	163,000	10,000	4,000	,000

This is a dissimilarity matrix

Περιέχει τις αποστάσεις όλων των παρατηρήσεων. Αναφέρεται το μέτρο απόστασης (ή ομοιότητας) που χρησιμοποιείται.

Εφαρμογή της Ιεραρχικής Ομαδοποίησης

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	1	9	2,000	0	0	4
2	5	10	4,000	0	0	4
3	4	7	4,000	0	0	7
4	1	5	4,500	1	2	6
5	2	3	5,000	0	0	7
6	1	6	6,500	4	0	8
7	2	4	6,500	5	3	9
8	1	8	9,800	6	0	9
9	1	2	125,833	8	7	0

Όταν ενώνονται δύο παρατηρήσεις διατηρείται η μικρότερη από τις δύο τιμές.

Ο πίνακας παρουσιάζει τις μέσες αποστάσεις μεταξύ των ομάδων για κάθε μεταβλητή.

Δείχνει τη σειρά με την οποία έγινε η ομαδοποίηση.

Για παράδειγμα, η μικρότερη απόσταση ήταν μεταξύ της παρατήρησης 1 και 9, οι οποίες δημιούργησαν στο πρώτο στάδιο μια ομάδα. Μετά το 4ο στάδιο υπάρχουν οι ομάδες {1,9,5,10} και {4,7}, ενώ οι υπόλοιπες παρατηρήσεις είναι μόνες τους από μια ομάδα.

Στη στήλη coefficients φαίνεται η τιμή της απόστασης ανάμεσα στις παρατηρήσεις που ενώθηκαν. Σε κάθε στάδιο η τιμή της μεγαλώνει (το αντίθετο θα συνέβαινε αν επιλέγαμε μέτρο ομοιότητας). Σταματάμε όταν η τιμή γίνεται ξαφνικά πολύ μεγάλη.

Εφαρμογή της Ιεραρχικής Ομαδοποίησης

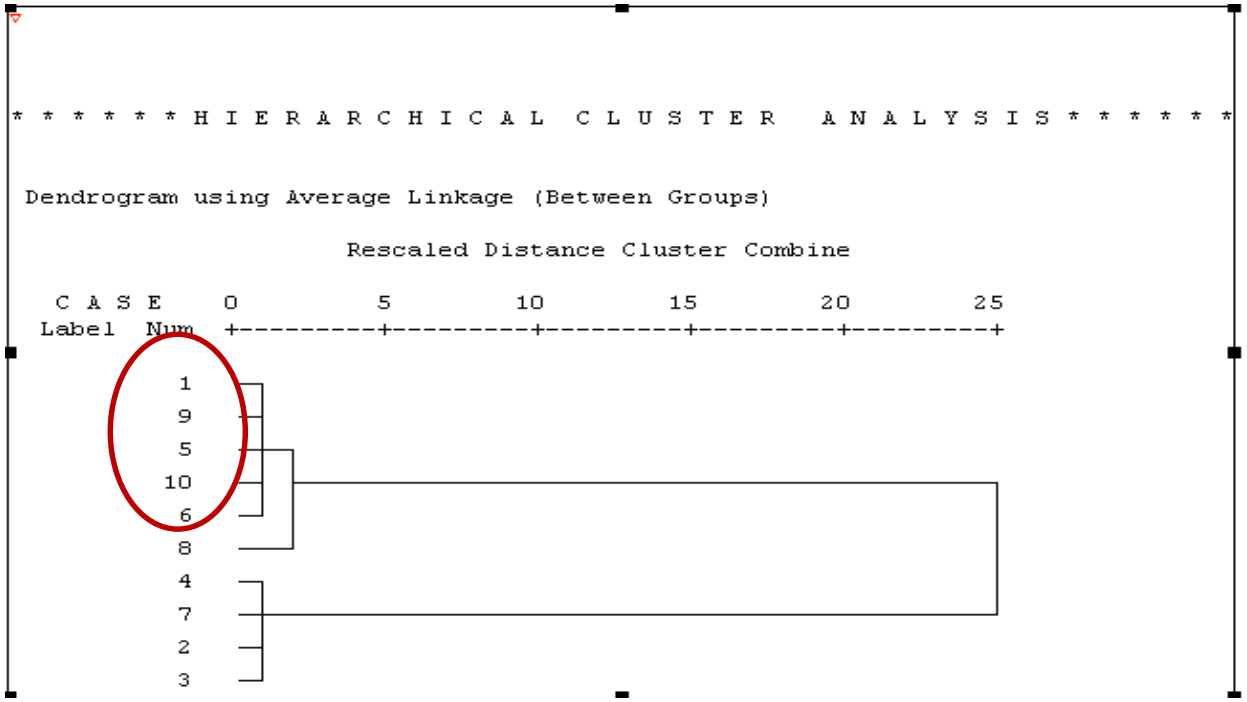
Vertical Icicle

Number of clusters	Case																		
	7		4		3		2		8		6		10		5		9		1
1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
2	X	X	X	X	X	X	X		X	X	X	X	X	X	X	X	X	X	X
3	X	X	X	X	X	X	X		X		X	X	X	X	X	X	X	X	X
4	X	X	X		X	X	X		X		X	X	X	X	X	X	X	X	X
5	X	X	X		X	X	X		X		X		X	X	X	X	X	X	X
6	X	X	X		X		X		X		X		X	X	X	X	X	X	X
7	X	X	X		X		X		X		X		X	X	X		X	X	X
8	X		X		X		X		X		X		X	X	X		X	X	X
9	X		X		X		X		X		X		X		X		X	X	X

Το παραπάνω διάγραμμα δείχνει πως γίνεται η διαδικασία.

Δεν έχει καλά γραφικά και προτειμάται το δενδρόγραμμα.

Εφαρμογή της Ιεραρχικής Ομαδοποίησης



Το δενδρόγραμμα δείχνει τις δύο ομάδες που δημιουργήθηκαν με την ιεραρχική ομαδοποίηση, αλλά και τις υποομάδες των παρατηρήσεων που εμφανίστηκαν κατά τη διαδικασία της μεθόδου. Για παράδειγμα, στο 1ο στάδιο της ανάλυσης οι καταναλωτές 1,9,5,10,6 αποτέλεσαν μια υποομάδα. Στο 2ο στάδιο, οι καταναλωτές 4,7,2,3 και οι 5,8 δημιούργησαν δύο υποομάδες. Οι δύο τελικές ομάδες αποτελούνται από τους καταναλωτές 1,9,5,10,6,8 (1η ομάδα) και 4,7,2,3 (2η ομάδα).

Σύγκριση των Δύο Μεθόδων

- Η μέθοδος K-Means και η Ιεραρχική Ομαδοποίηση κατέληξαν στα ίδια συμπεράσματα.

- Συνήθως χρησιμοποιούνται **συνδυαστικά**. Αρχικά χρησιμοποιείται η ιεραρχική ομαδοποίηση για να εντοπιστεί ο αριθμός των ομάδων και στη συνέχει η μέθοδος K-Means.

- Ο συνδυασμός των μεθόδων δεν είναι απαραίτητος όταν ο αριθμός των ομάδων είναι γνωστός.

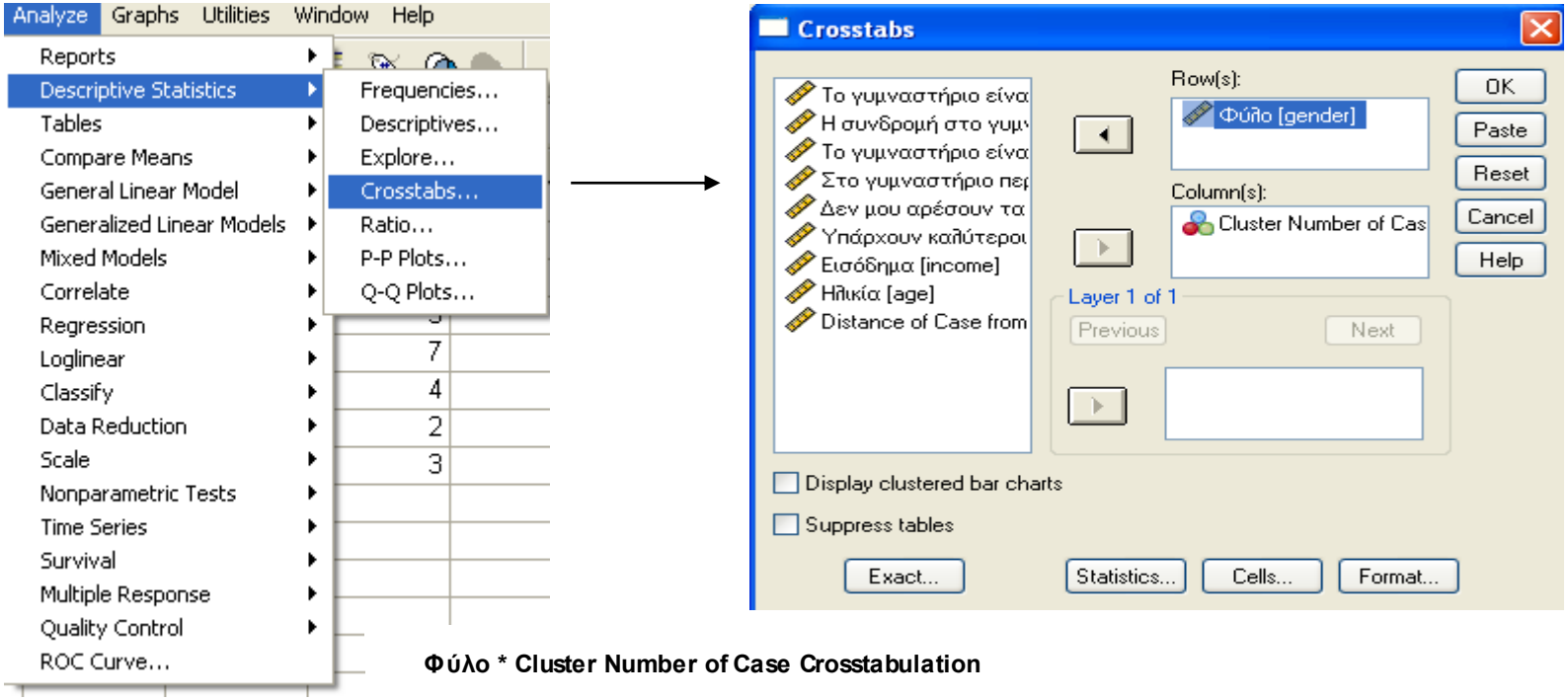
Εφαρμογή της Ανάλυσης κατά Συστάδες

Στο προηγούμενο παράδειγμα προσθέτουμε τις μεταβλητές Ηλικία, Φύλο (1: άνδρας, 2: γυναίκα) και Εισόδημα (1: λιγότερο από 1000 ευρώ, 2: 1000-2000 ευρώ, 3: πάνω από 2000 ευρώ).

v1	v2	v3	v4	v5	v6	gender	income	age	QCL_1
2	6	3	2	6	6	2	2	50	2
1	7	2	2	7	7	2	3	66	2
6	1	7	7	2	2	1	1	19	1
6	2	6	5	3	2	1	2	22	1
2	6	1	1	7	7	2	3	67	2
7	3	7	7	4	3	1	2	38	1
7	2	6	6	2	2	1	1	50	1
7	1	6	7	3	1	1	1	52	1

Ξανατρέχουμε τη μέθοδο K-Means ώστε να σχηματιστούν οι ομάδες και στη συνέχεια συγκρίνουμε τη νέα μεταβλητή QCL_1 με τις μεταβλητές Ηλικία, Φύλο και Εισόδημα

Εφαρμογή της Ανάλυσης κατά Συστάδες

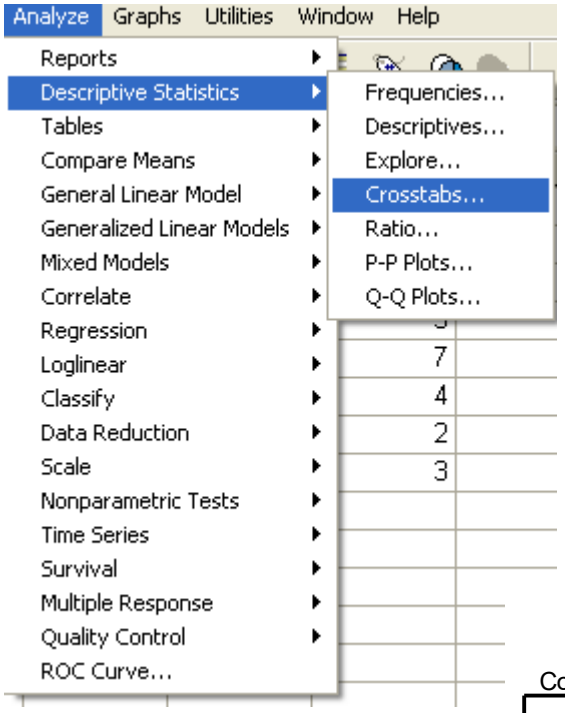


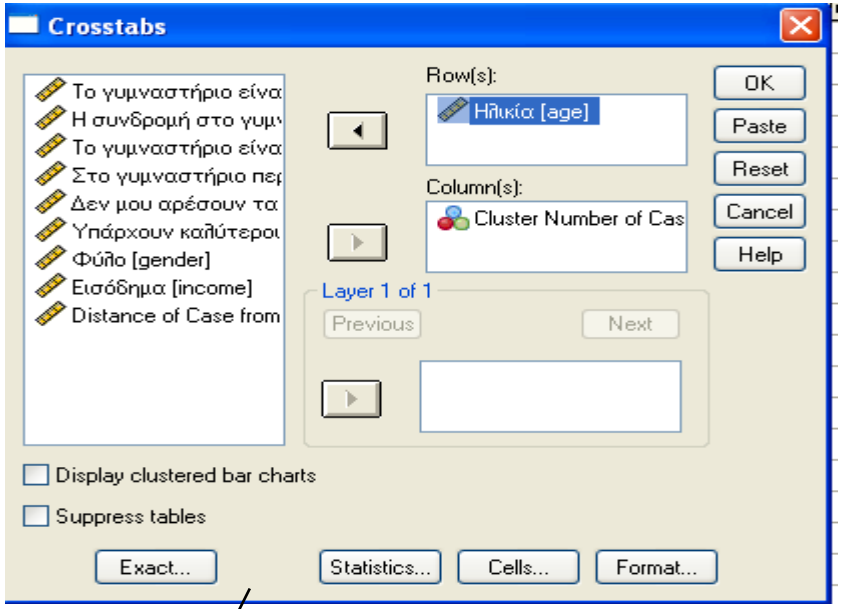
Φύλο * Cluster Number of Case Crosstabulation

Count

		Cluster Number of Case		
		1	2	Total
Φύλο	άνδρας	6	1	7
	γυναίκα	0	3	3
Total		6	4	10

Εφαρμογή της Ανάλυσης κατά Συστάδες



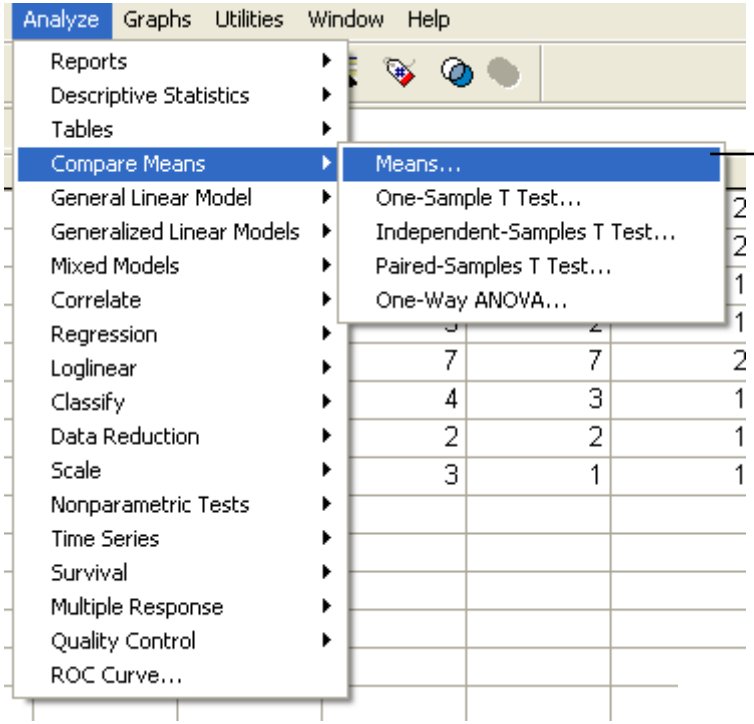


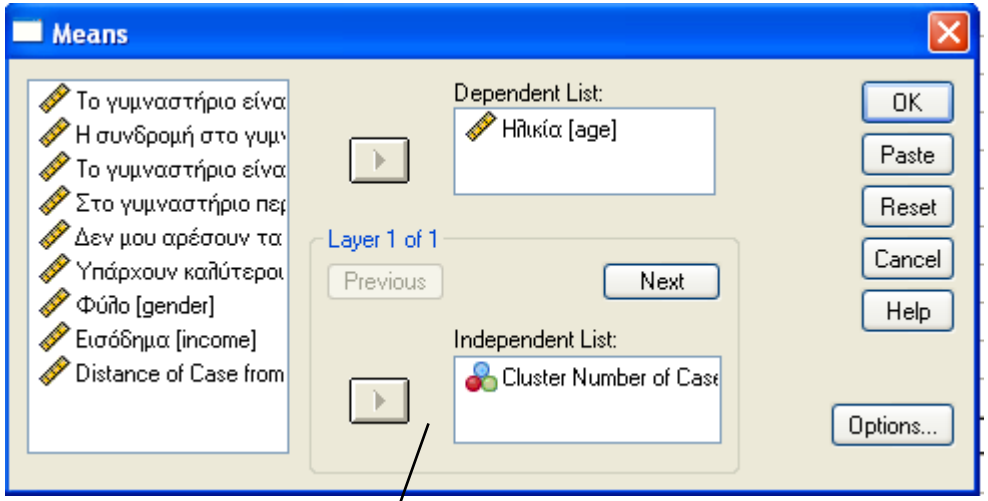
Εισόδημα * Cluster Number of Case Crosstabulation

Count

		Cluster Number of Case		
		1	2	Total
Εισόδημα	λιγότερα από 1000 ευρώ	4	0	4
	1000-2000 ευρώ	2	2	4
	πάνω από 2000 ευρώ	0	2	2
Total		6	4	10

Εφαρμογή της Ανάλυσης κατά Συστάδες





Report

Ηλικία

Cluster Number of Case	Mean	N	Std. Deviation
1	34,50	6	14,335
2	53,25	4	17,347
Total	42,00	10	17,556

Εφαρμογή της Ανάλυσης κατά Συστάδες

Από τους πίνακες φαίνεται ότι η ομάδα 1 (θυμίζουμε ότι είναι η ομάδα των «αθλητικών καταναλωτών») απαρτίζεται από άτομα μικρότερης ηλικίας (μέσος όρος 34,5 έτη), άνδρες με χαμηλό εισόδημα (κάτω από 2000 ευρώ.

Αντίθετα οι ομάδα 2 («αποστάτες καταναλωτές») είναι γυναίκες, μεγαλύτερης ηλικίας (μέσος όρος 53,25 έτη) με υψηλό εισόδημα (πάνω από 2000 ευρώ).