ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

«Διοίκηση στη Ναυτική Επιστήμη και Τεχνολογία»

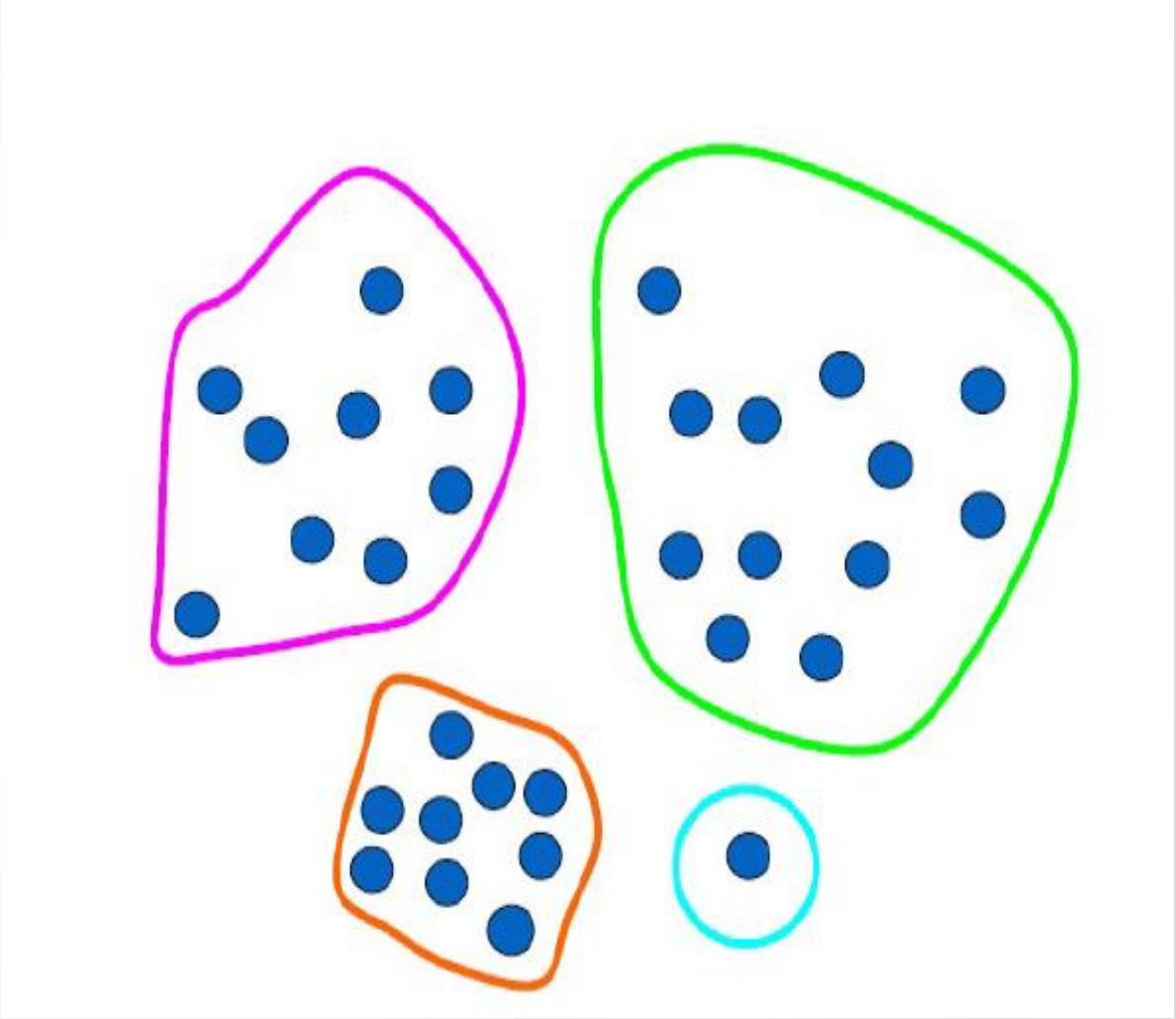
Image by David Mark from Pixabay

Διαχείριση μεγάλου όγκου δεδομένων στη Ναυτιλία και το Θαλάσσιο Περιβάλλον

Διδάσκοντες: Γ. Γαλάνης, Δ. Κουλουμπού

Συσταδοποίηση – Clustering

1. Εισαγωγή



Συσταδοποίηση – Clustering

- Η **συσταδοποίηση (clustering)** είναι μια περιγραφική μέθοδος.
- Έχοντας ένα σύνολο δεδομένων, στόχος της συσταδοποίησης είναι η δημιουργία συστάδων (clusters), δηλαδή ομάδων, οι οποίες θα περιέχουν όμοια ή παρεμφερή δείγματα.

Συσταδοποίηση – Clustering

- Στο **πρόβλημα της συσταδοποίησης** μας δίνεται ένα σύνολο n παρατηρήσεων, και για κάθε παρατήρηση έχουμε δεδομένα από N μεταβλητές $X_1, X_2, \dots, X_N$.

Συσταδοποίηση – Clustering

Παρατήρηση	X_1	X_2	X_3	….	X_N
1	x_{11}	x_{12}	x_{13}	…	x_{1N}
2	x_{21}	x_{22}	x_{23}	…	x_{2N}
3	x_{31}	x_{32}	x_{33}	…	x_{3N}
⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮
n	x_{n1}	x_{n2}	x_{n3}	…	x_{nN}

Συσταδοποίηση – Clustering

- Το σύνολο των δεδομένων δίνεται χωρίς τις αντίστοιχες κλάσεις ή ετικέτες και χρειαζόμαστε κάποιον αλγόριθμο, ο οποίος θα ομαδοποιήσει αυτόματα τα δεδομένα σε συστάδες.

Η Έννοια των Clusters

- Οι συστάδες (clusters) που δημιουργούνται θέλουμε να διαχωρίζουν ορθά τα δεδομένα.
- Αυτό πρακτικά σημαίνει ότι μια συστάδα θέλουμε να απαρτίζεται από αντικείμενα, όπου κάθε αντικείμενο είναι πιο «κοντά» σε κάθε άλλο αντικείμενο της ίδιας συστάδας απ’ ότι σε κάποιο άλλο αντικείμενο διαφορετικής συστάδας.

Η Έννοια των Clusters

- Ουσιαστικά αναζητείται ένα πεπερασμένο σύνολο κατηγοριών ή συστάδων, για να περιγράψει τα δεδομένα.
- Οι κατηγορίες μπορεί να είναι αμοιβαία αποκλειόμενες (να μην έχουν κοινά στοιχεία) και εξαντλητικές ή να έχουν μία πιο σύνθετη αναπαράσταση, όπως για παράδειγμα ιεραρχικές και επικαλυπτόμενες (να έχουν κοινά στοιχεία).

Συσταδοποίηση – Clustering

- Μια επιτυχημένη ανάλυση θα πρέπει να καταλήξει σε συστάδες (clusters) για τις οποίες οι παρατηρήσεις μέσα σε κάθε συστάδα θα είναι όσο γίνεται πιο ομοιογενείς αλλά παρατηρήσεις διαφορετικών ομάδων θα διαφέρουν όσο γίνεται περισσότερο.

Ανάλυση κατά Συστάδες

- Η ανάλυση κατά συστάδες είναι επομένως μια πολυμεταβλητή στατιστική μέθοδος της οποίας σκοπός είναι να κατατάξει τις υπάρχουσες τιμές σε ομάδες χρησιμοποιώντας την πληροφορία που υπάρχει σε διάφορες μεταβλητές.

Ανάλυση κατά Συστάδες

- Δύο βασικές έννοιες της συγκεκριμένης μεθόδου είναι η ***έννοια της απόστασης*** και η ***έννοια της ομοιότητας***.
- Πρόκειται για δύο αντίθετες, μεταξύ τους, έννοιες οι οποίες ουσιαστικά ποσοτικοποιούν την ποιοτική έννοια της ομοιογένειας και της ομοιότητας.

Ανάλυση κατά Συστάδες

- Παρατηρήσεις που μοιάζουν πολύ μεταξύ τους έχουν, δηλαδή, σχετικά όμοιες τιμές (ποιοτικές ή ποσοτικές), θα έχουν:

- **Μεγάλες τιμές αν καθορίσουμε κάποιο μέτρο ομοιότητας**
- **Μικρές τιμές αν καθορίσουμε κάποιο μέτρο το οποίο θα δηλώνει τη μεταξύ τους απόσταση.**

Clustering – Εφαρμογές

- Η ανάλυση συστάδων έχει εφαρμογές σε πολλές επιστήμες όπως η κοινωνιολογία, η βιολογία και η στατιστική.
- Η ανάλυση συστάδων εφαρμόζεται σε πολλούς τομείς της πληροφορικής όπως η αναγνώριση προτύπων, η εξόρυξη γνώσης, η ανάκτηση δεδομένων, η τεχνητή νοημοσύνη και η μηχανική μάθηση.

Clustering – Εφαρμογές

- Μερικές εφαρμογές της ανάλυση κατά συστάδων είναι οι εξής:
- **Για τη διερεύνηση των διαθέσιμων δεδομένων** και την προσπάθεια ανίχνευσης τυχόν μεταξύ τους ομοιοτήτων, σχέσεων, μεταβλητών με διακριτική ικανότητα κ.λ.π.

Clustering – Εφαρμογές

- **Για τη μείωση των διαστάσεων του προβλήματος.** Σε τεράστιες βάσεις δεδομένων συχνά περιέχεται μικρότερη από την αναμενόμενη πληροφορία, καθώς μπορεί να υπάρχουν επικαλύψεις μεταξύ της πληροφορίας που περιέχεται σε διάφορες μεταβλητές, στοιχεία χωρίς ιδιαίτερο ενδιαφέρον κ.λ.π.

Clustering – Εφαρμογές

- Ομαδοποιώντας τα δεδομένα, μας δίνεται η δυνατότητα να εστιάσουμε στις μεταβλητές οι οποίες παίζουν σημαντικό ρόλο στην ερμηνεία τους και να επικεντρωθούμε σ’ αυτές.

Clustering – Εφαρμογές

- **Για προβλέψεις νέων τιμών.** Έχοντας ήδη ομαδοποιήσει τα δεδομένα μας με την εφαρμογή της ανάλυσης κατά συστάδες, μας δίνεται η δυνατότητα να κατατάξουμε καινούριες διαθέσιμες παρατηρήσεις.

Clustering – Εφαρμογές

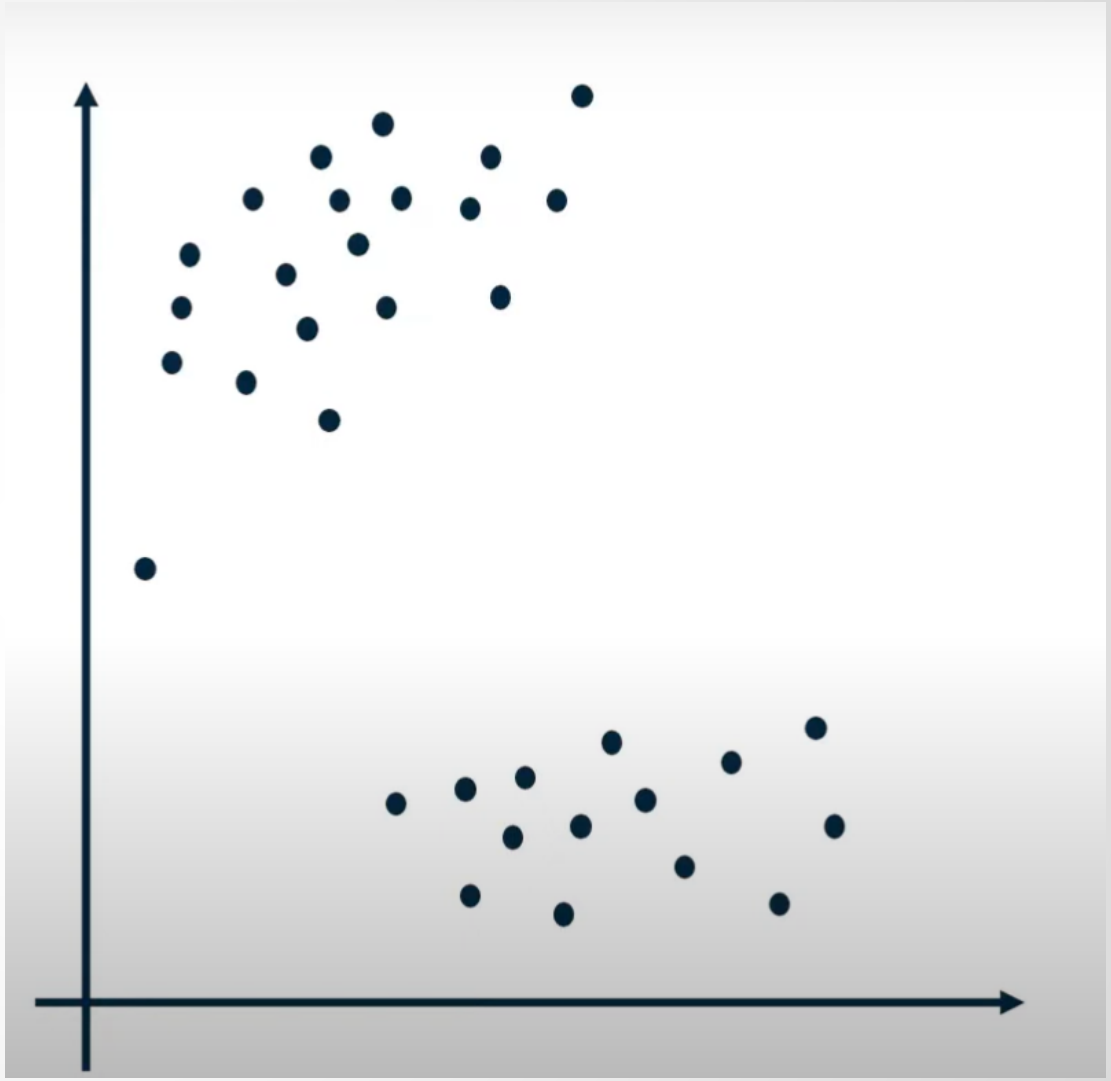
- Χαρακτηριστικό παράδειγμα της προαναφερθείσας εφαρμογής έχουμε στις τράπεζες οι οποίες κατατάσσουν τους πελάτες τους σε διάφορες κατηγορίες ανάλογα με κάποια διαθέσιμα χαρακτηριστικά και σύμφωνα με τη δυνατότητα που έχουν για αποπληρωμή κάποιου δανείου που έχουν πάρει.

Clustering – Εφαρμογές

- Το ενδιαφέρον της τράπεζας εστιάζεται, εκτός των άλλων, και στη σωστή κατάταξη νέων πελατών σε κάποια από τις ομάδες αυτές ώστε να αποφασίσει, με το λιγότερο δυνατό ρίσκο αν θα τους δώσει δάνειο ή όχι.

Συσταδοποίηση – Clustering

2. Ο Αλγόριθμος $k - Means$



Ο Αλγόριθμος k − Means

- Υποθέτουμε ότι διαθέτουμε ένα σύνολο n παρατηρήσεων, και για κάθε παρατήρηση έχουμε δεδομένα από N μεταβλητές $X_1, X_2, \dots, X_N$.

Ο Αλγόριθμος k − Means

- Στην συσταδοποίηση $k - means$, κάθε ομάδα αντιπροσωπεύεται από το κέντρο της (δηλαδή το κέντρο βάρους της) που αντιστοιχεί στο μέσο όρο των συνεταγμένων των σημείων που αντιστοιχούν στην ομάδα.

Ο Αλγόριθμος $k - \mathbf{Means}$

- Για να κατανοήσουμε καλύτερα των αλγόριθμο $k - means$ θα θεωρήσουμε την περίπτωση όπου για κάθε παρατήρηση έχουμε δεδομένα από δύο μεταβλητές.
- Η μεθοδολογία για τις πιο σύνθετες περιπτώσεις γενικεύεται ανάλογα.

Ο Αλγόριθμος $k-$ Means για $N=2$ Μεταβλητές

- Υποθέτουμε ότι έχουμε n παρατηρήσεις και για κάθε παρατήρηση έχουμε δεδομένα από δύο ποσοτικές μεταβλητές X, Y.

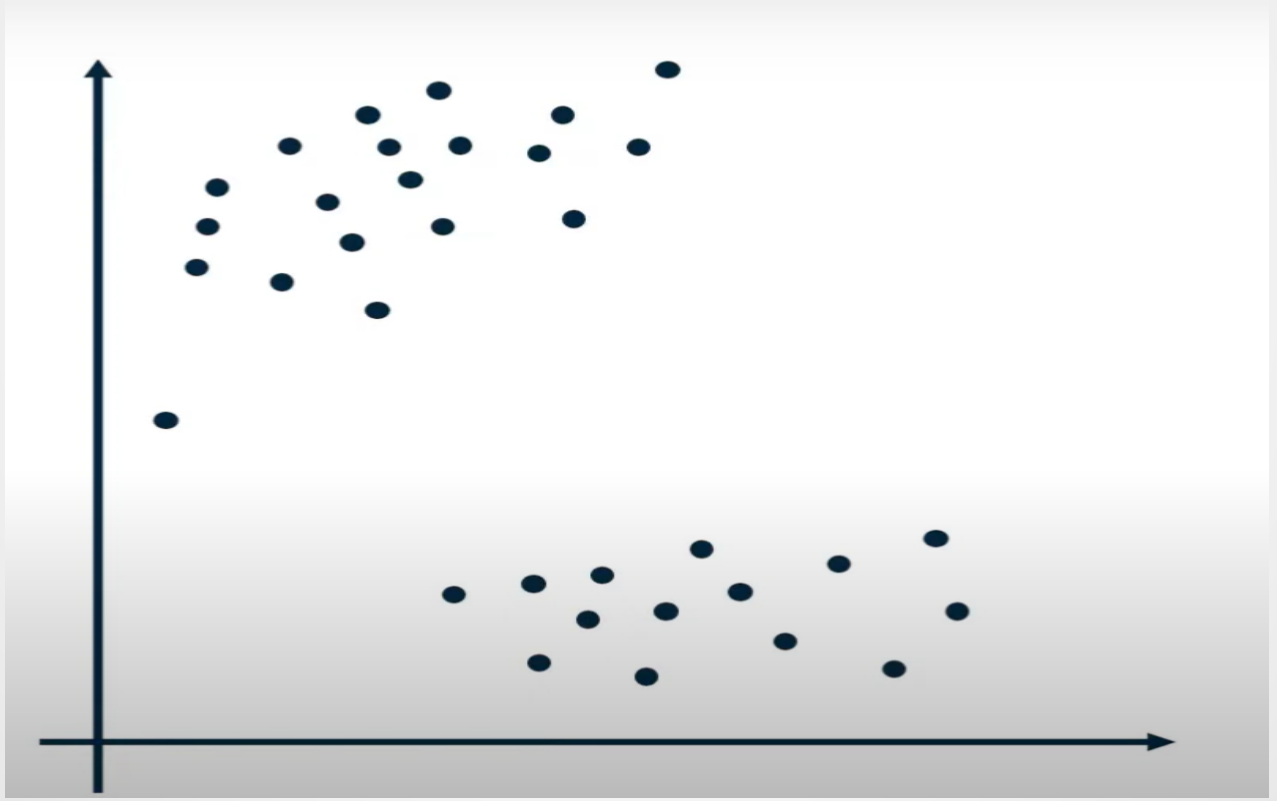
Ο Αλγόριθμος k − Means

- Επομένως σε κάθε i παρατήρηση μπορούμε να αντιστοιχήσουμε ένα διατεταγμένο ζεύγος $(x_i, y_i), i = 1,2, \dots, n$. Όπου
  - x_i είναι η τιμή της παρατήρησης i για την μεταβλητή X,
  - y_i είναι η τιμή της παρατήρησης i για την μεταβλητή Y.

Ο Αλγόριθμος k − Means

- Συνεπώς μπορούμε να αντιστοιχήσουμε κάθε μία από τις n παρατηρήσεις μας με ένα σημείο στο επίπεδο, όπως φαίνεται στο παρακάτω διάγραμμα διασποράς.

Ο Αλγόριθμος k − Means



Ο Αλγόριθμος $k - \mathbf{Means}$

Παρατήρηση

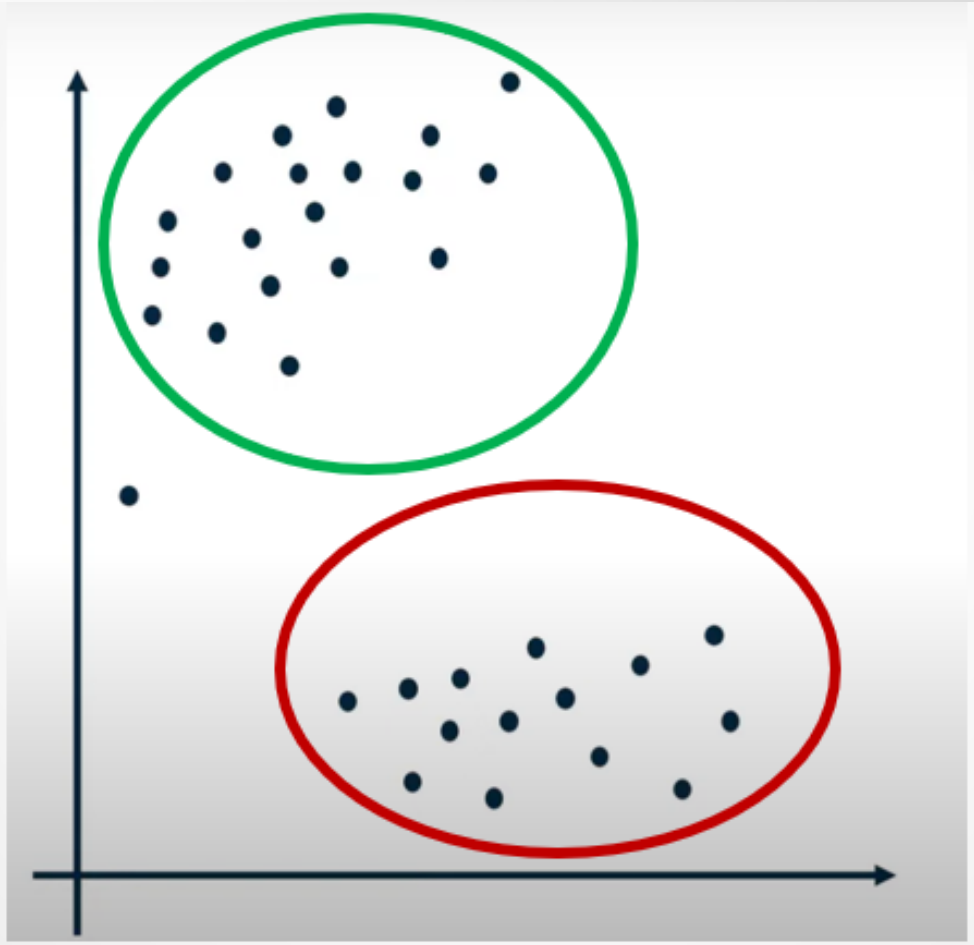
Όμοια αν για κάθε μία παρατήρηση έχουμε δεδομένα από $N \in \mathbb{N}$ διαφορετικές ποσοτικές μεταβλητές μπορούμε κάθε παρατήρηση να την αντιστοιχήσουμε με ένα σημείο σε ένα Ευκλείδειο χώρο διάστασης N.

Ο Αλγόριθμος k − Means, Περίπτωση $k = 2, N = 2$

- **Το k στην ονομασία του αλγόριθμου k − $means$ δηλώνει των αριθμό των συστάδων που θέλουμε ο αλγόριθμος να δημιουργήσει.**
- Ας δούμε την απλή περίπτωση που $k = 2$.

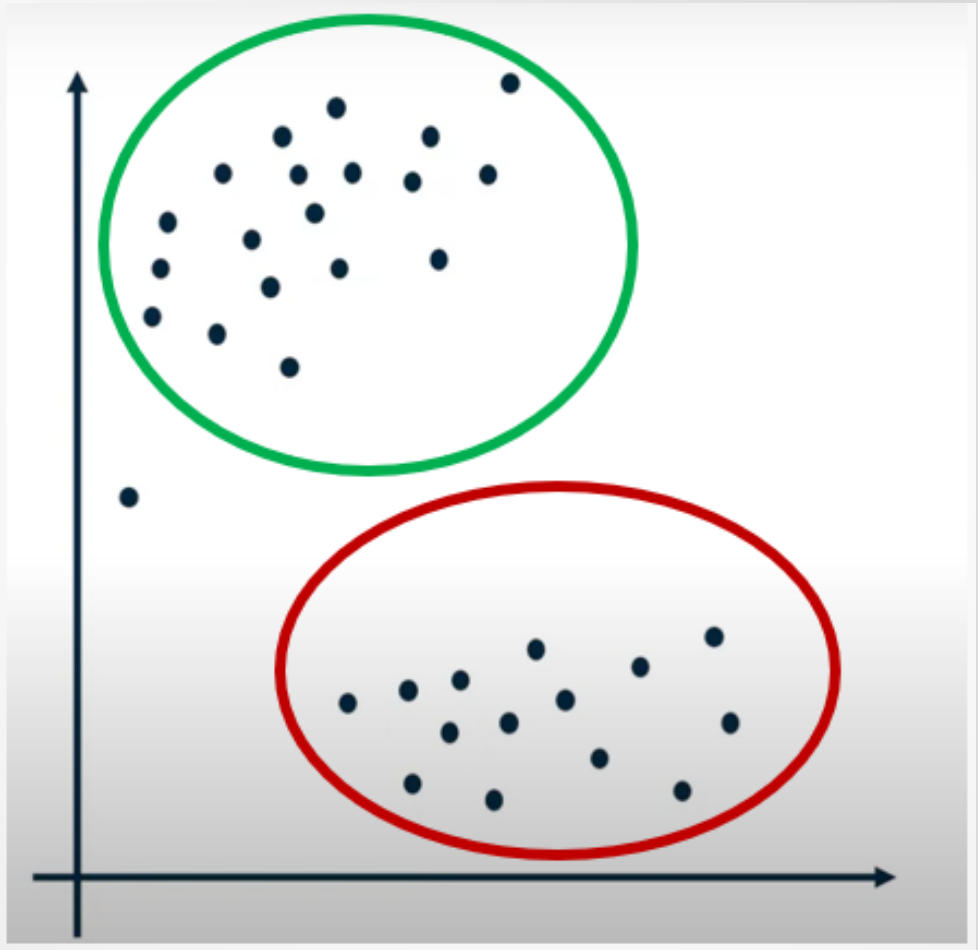
Ο Αλγόριθμος k − Means

- Στο συγκεκριμένο παράδειγμα, φαίνεται σχεδόν ξεκάθαρα ποιες θα είναι οι συστάδες.



Ο Αλγόριθμος $k - \mathbf{Means}$

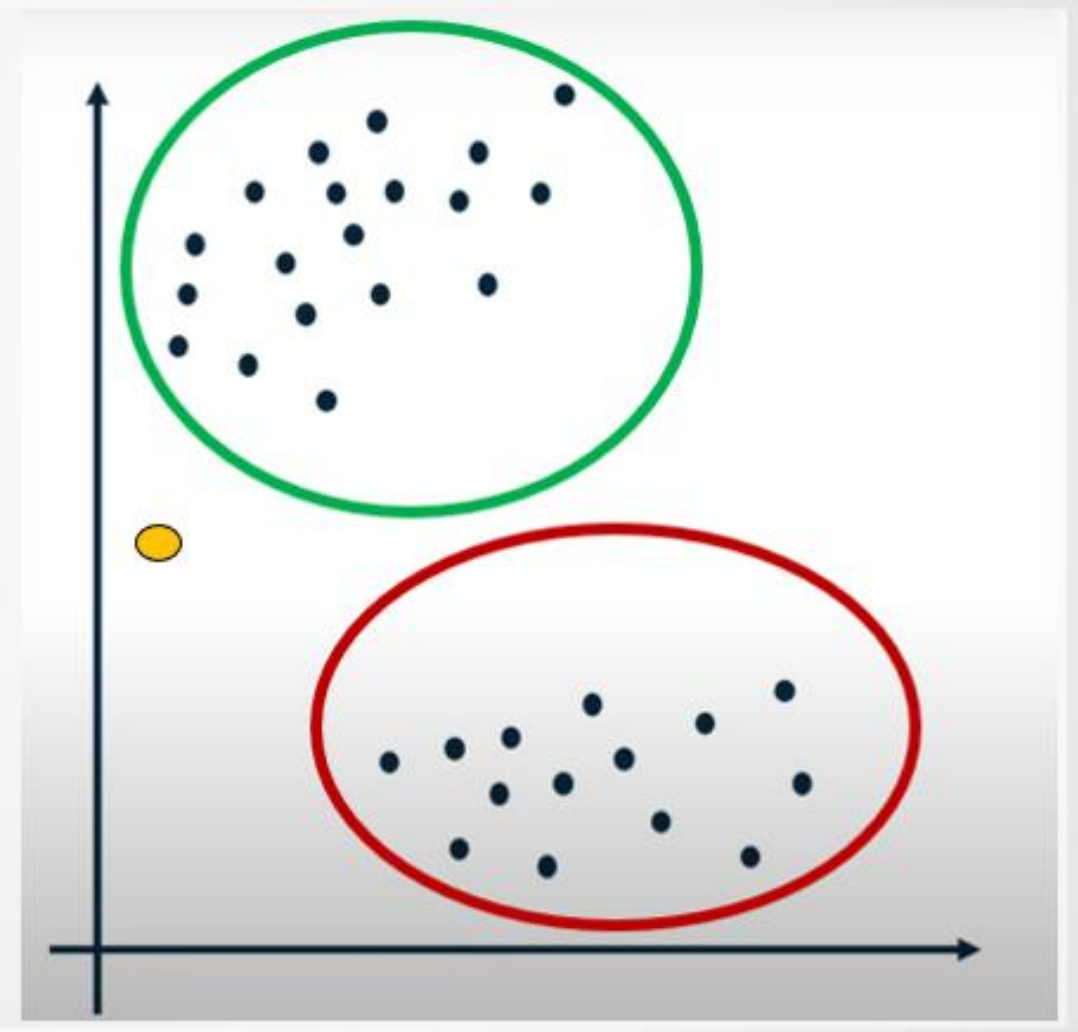
- Στα παραδείγματα της πραγματικής ζωής είναι σαφώς πιο πολύπλοκη η δημιουργία συστάδων.
- Τα δεδομένα είναι πολύ περισσότερα, όπως και οι μεταβλητές.
- Επομένως η δημιουργία ενός αλγόριθμου κρίνεται απαραίτητη.



Ο Αλγόριθμος k − Means

- Στο συγκεκριμένο παράδειγμα ενδιαφέρον έχει το σημείο που δεν έχει τοποθετηθεί σε κάποιο από τα «προφανή» clusters.
- Το cluster στο οποίο θα πρέπει να τοποθετηθεί δεν είναι φανερό. Αυτό το σημείο αποτελεί και ένα παράδειγμα που κάνει τον αλγόριθμο ιδιαίτερα ενδιαφέρον.

Ο Αλγόριθμος k − Means



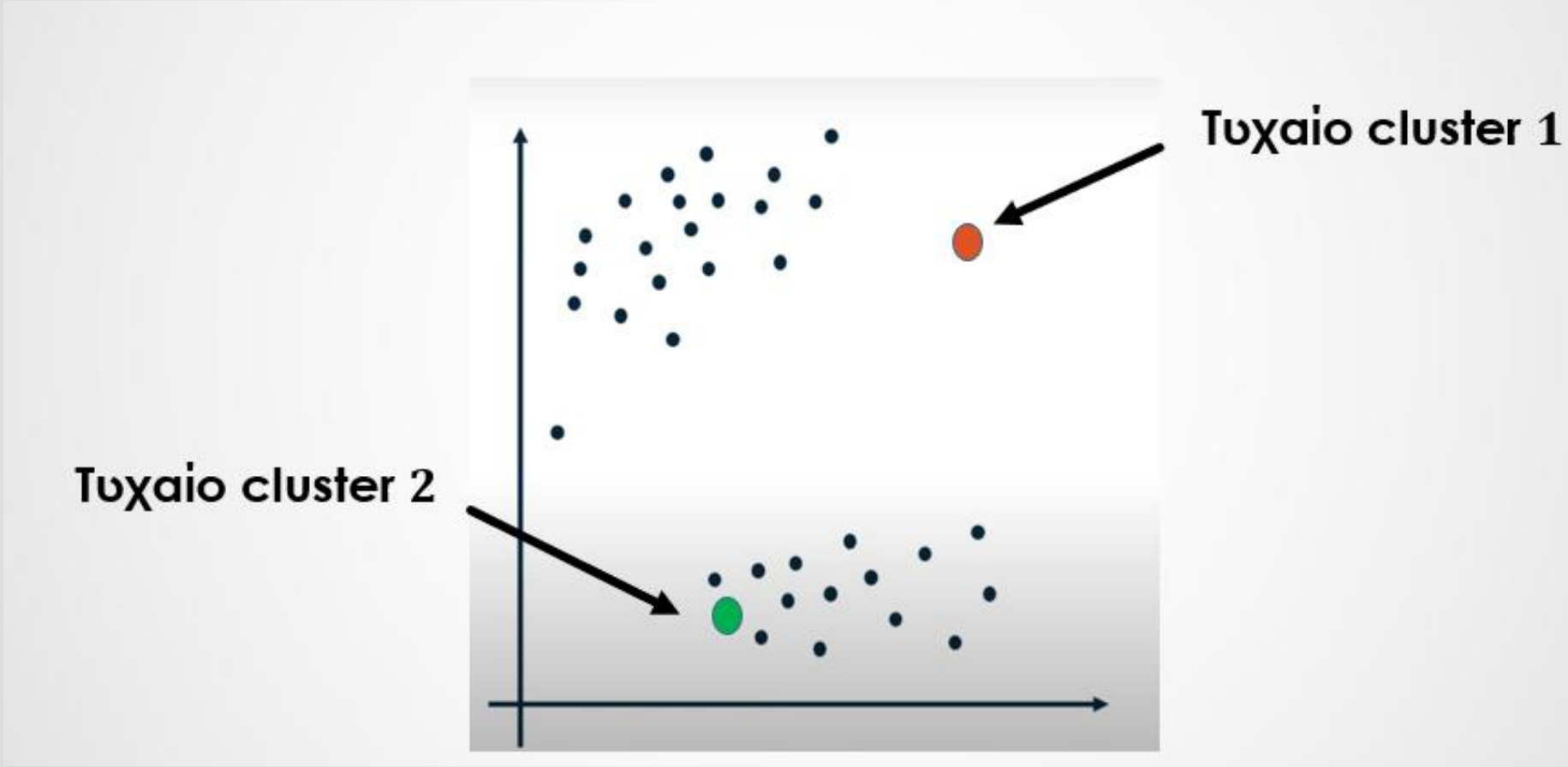
Επιλογή Τυχαίων Σημείων Clusters

- Αρχικά δεν γνωρίζουμε τίποτα για τα δεδομένα και το πως αυτά τοποθετούνται στον χώρο.

Βήμα 1 (Επιλογή τυχαίων σημείων clusters)

- Το πρώτο βήμα του αλγόριθμου $k - Means$, για $k = 2$, είναι η επιλογή δύο τυχαίων σημείων στο επίπεδο, όπως φαίνεται στο παρακάτω σχήμα.

Τοποθετώντας τα Δεδομένα σε Συστάδες



Επιλογή Τυχαίων Σημείων Clusters

Παρατήρηση

Τα τυχαία σημεία δεν αποτελούν κατά ανάγκη σημεία των δεδομένων του δείγματος.

Τοποθετώντας τα Δεδομένα σε Συστάδες

Βήμα 2 (Τοποθέτηση των δεδομένων σε συστάδες)

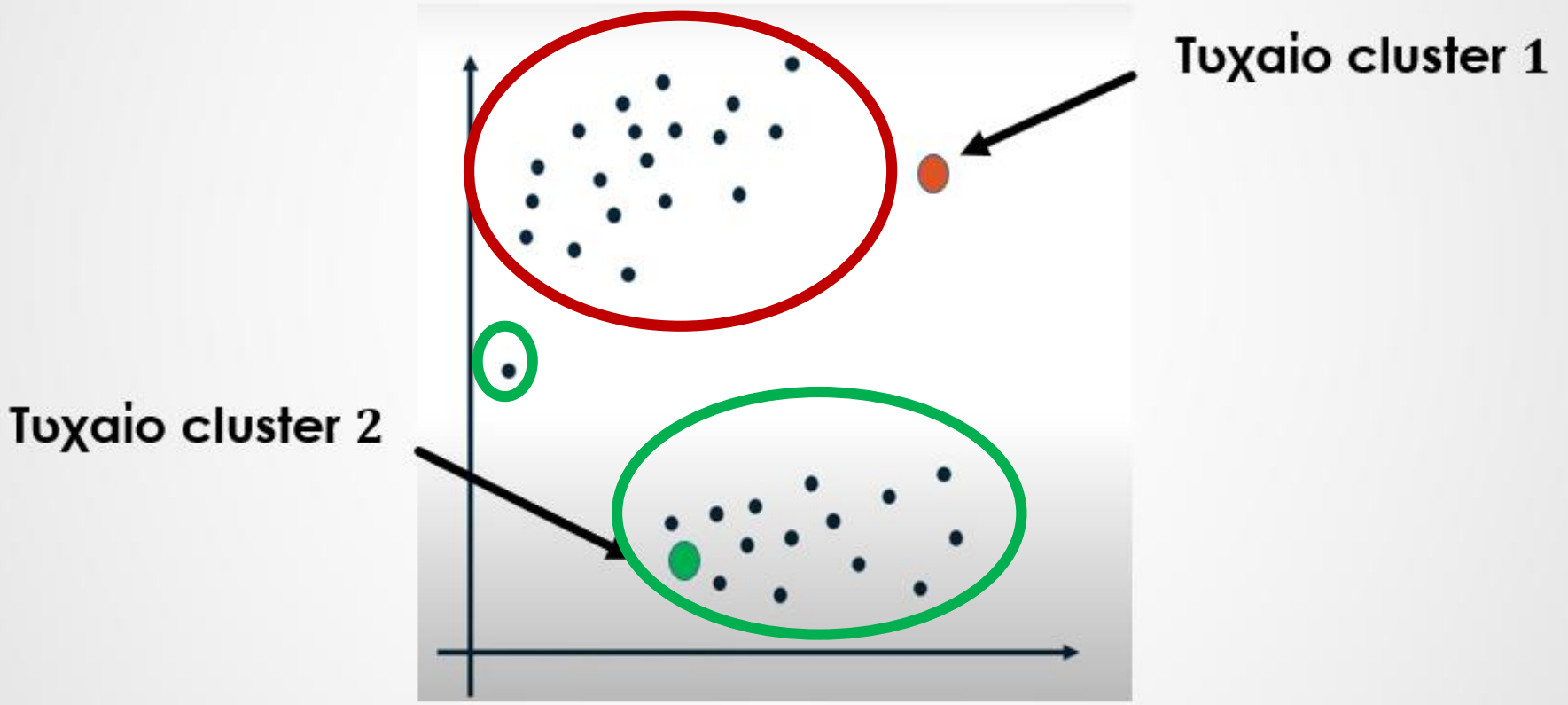
Τοποθετούμε τα δεδομένα του δείγματος σε συστάδες σύμφωνα με τον παρακάτω κανόνα.

Τοποθετώντας τα Δεδομένα σε Συστάδες

- **Όποιο σημείο βρίσκεται πιο κοντά στο σημείο «Cluster 1», θα ανήκει στην συστάδα 1.**
- **Όποιο σημείο βρίσκεται πιο κοντά στο σημείο «Cluster 2», θα ανήκει στην συστάδα 2.**

Την δημιουργία των συστάδων μετά το πρώτο βήμα του αλγόριθμου μπορούμε να την δούμε στο παρακάτω σχήμα.

Τοποθετώντας τα Δεδομένα σε Συστάδες



Η Έννοια της Απόστασης στον Αλγόριθμο των $k - Means$

- Ο αλγόριθμος $k - means$ χρησιμοποιεί την κυρίως την Ευκλείδεια έννοια της απόστασης, η οποία σε έναν χώρο N διάστασης ορίζεται ως εξής:

Η Έννοια της Απόστασης στον Αλγόριθμο των $k - Means$

Ορισμός

Έστω δύο σημεία $x = (x_1, x_2, \dots, x_N)$ και $y = (y_1, y_2, \dots, y_N)$ του $\mathbb{R}^N$.

Ορίζουμε απόσταση $d(x, y)$ των σημείων x, y τον μη αρνητικό αριθμό

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

Μετακίνηση των Clusters

Βήμα 3 (Μετακίνηση των Clusters)

- Στη συνέχεια για κάθε μία από τις συστάδες που έχουμε δημιουργήσει μετακινούμε το αρχικό τυχαίο cluster, στο σημείο που αποτελεί την **μέση τιμή (mean)** όλων των δεδομένων του δείγματος που ανήκουν στην συγκεκριμένη συστάδα.

Μετακίνηση των Clusters

- Ποιο συγκεκριμένα αν τα σημεία

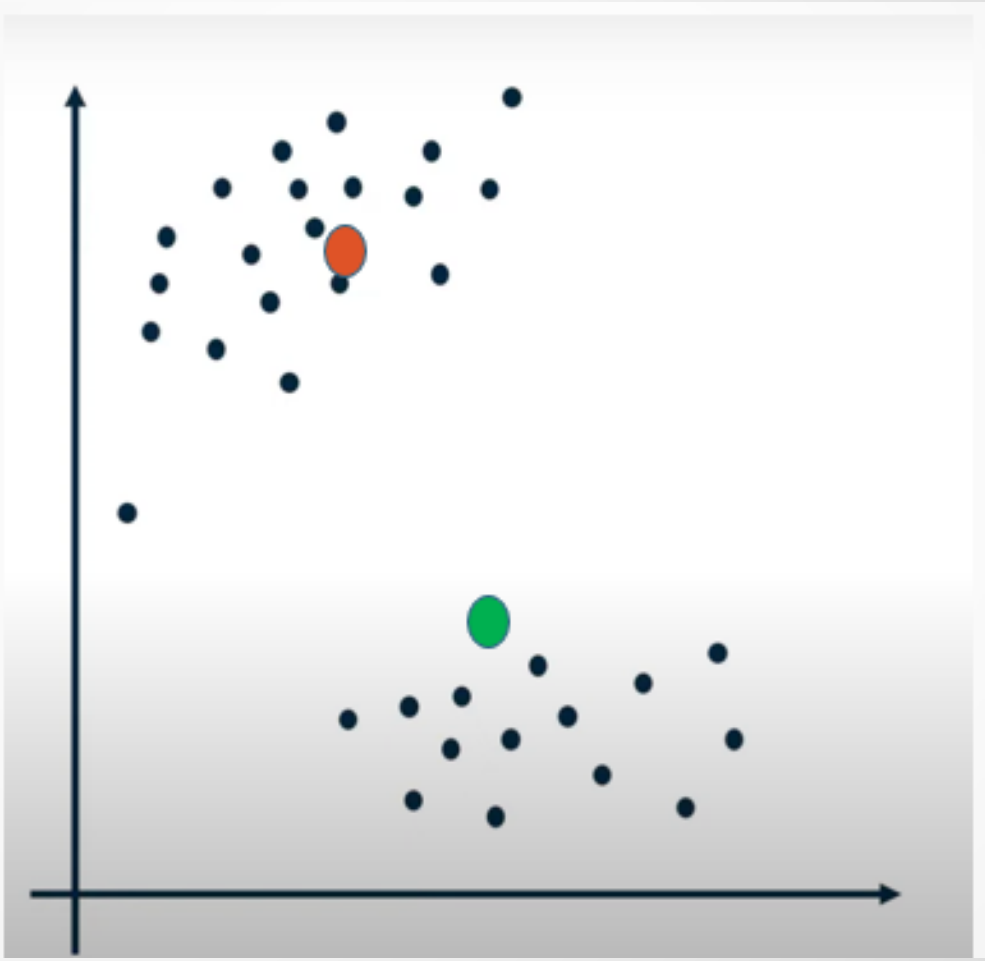
$$(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m),$$

Είναι τα σημεία του δείγματος που ανήκουν σε μία συστάδα, τότε το τυχαίο cluster αυτής της συστάδας θα μετατοπισθεί στο σημείο

$$(\overline{\boldsymbol{x}}, \overline{\boldsymbol{y}}) = \left(\frac{\boldsymbol{x_1 + x_2 + \cdots + x_m}}{\boldsymbol{m}}, \frac{\boldsymbol{y_1 + y_2 + \cdots + y_m}}{\boldsymbol{m}} \right).$$

Μετακίνηση των Clusters

- Τα δεδομένα του δείγματος και τα σημεία των clusters μετά την μετακίνηση τους.



Επανατοποθετώντας τα Δεδομένα σε Συστάδες

Βήμα 4 (Αναδιαμορφώνουμε τις συστάδες)

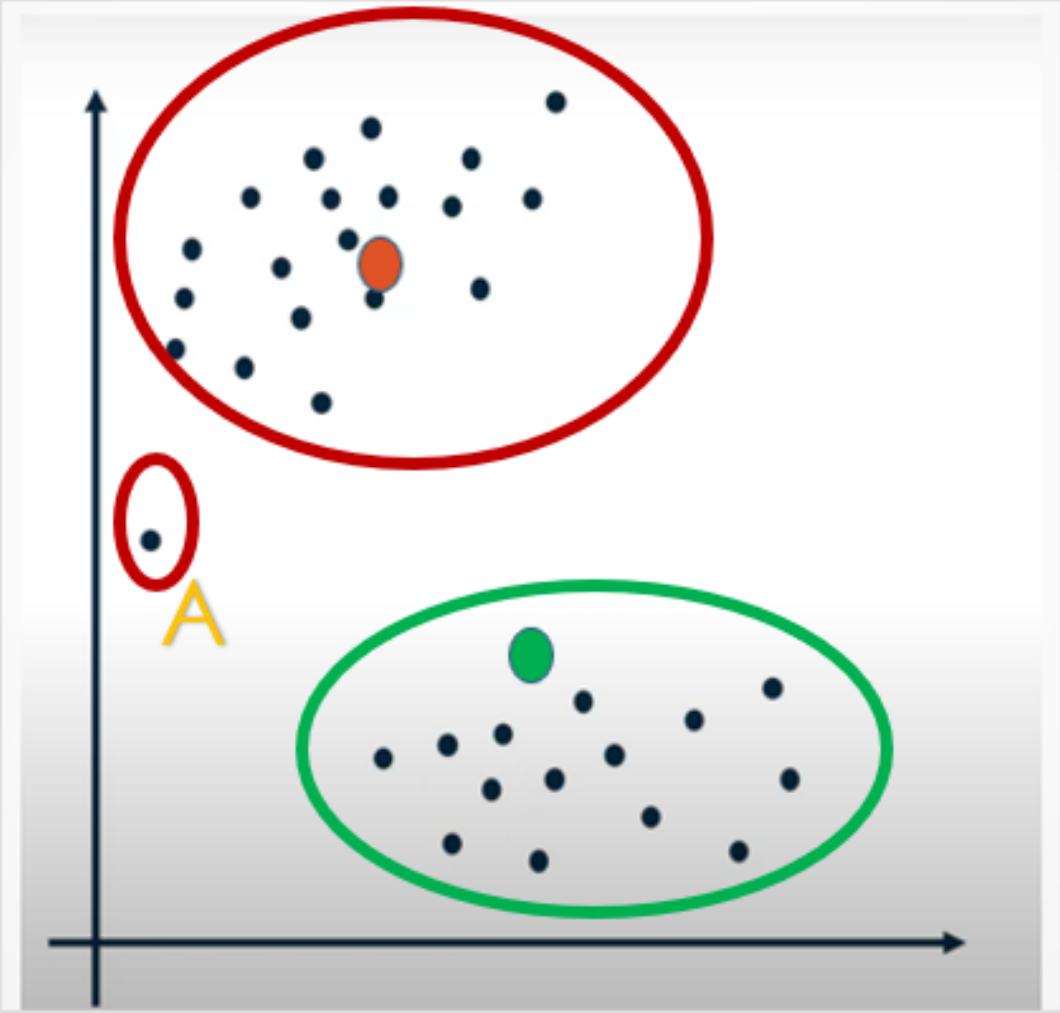
- Επαναλαμβάνουμε το βήμα 2, με τα νέα clusters.

➢ **Όποιο σημείο βρίσκεται πιο κοντά στο νέο σημείο «Cluster 1», θα ανήκει στην συστάδα 1.**

➢ **Όποιο σημείο βρίσκεται πιο κοντά στο νέο σημείο «Cluster 2», θα ανήκει στην συστάδα 2.**

Επανατοποθετώντας τα Δεδομένα σε Συστάδες

- Παρατηρείστε ότι η σύσταση των συστάδων έχει αλλάξει.
- Πλέον το σημείο A, ανήκει σε διαφορετική συστάδα από ότι στο προηγούμενο βήμα.



Επανάληψη Βημάτων

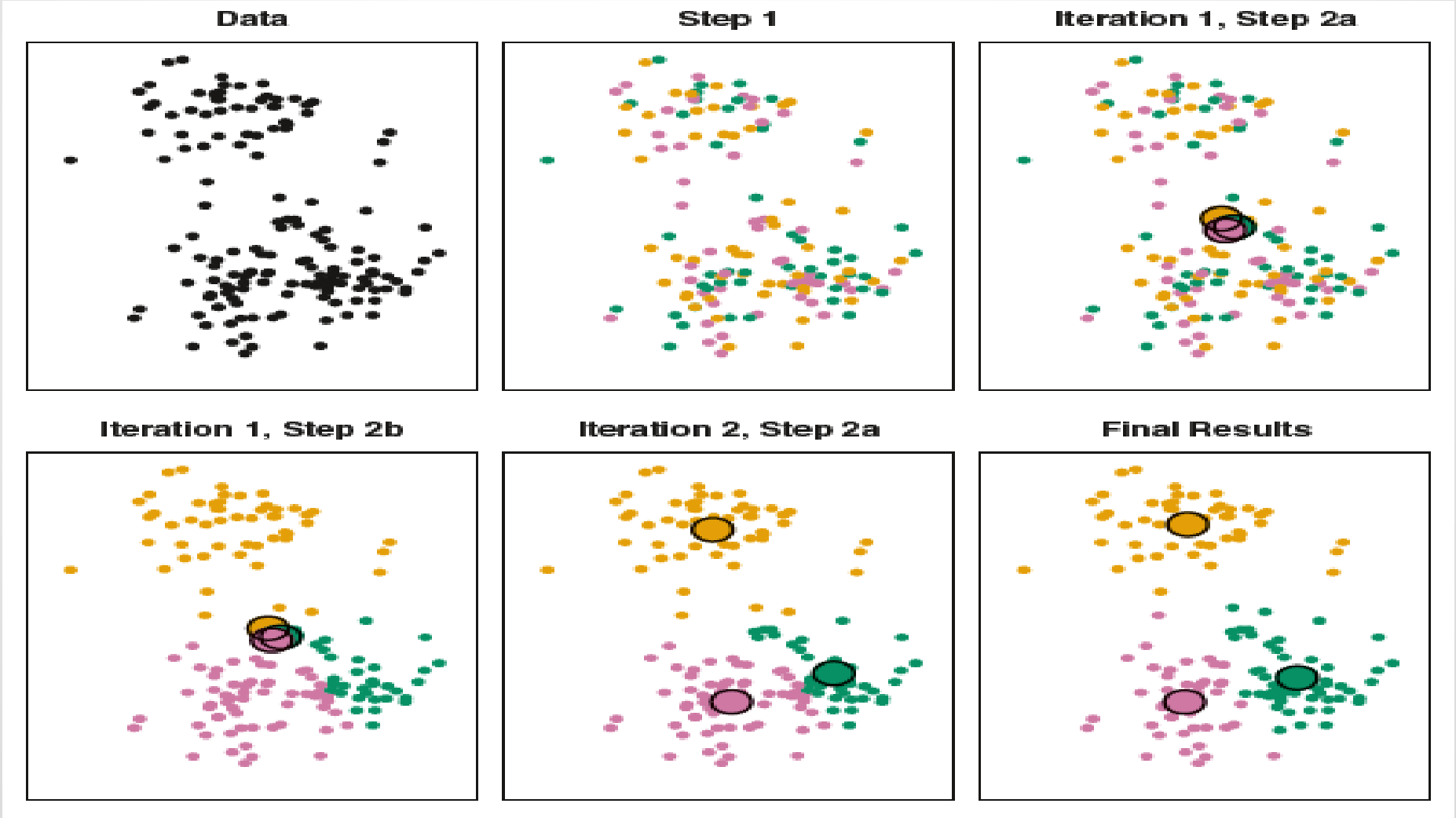
Βήμα 5:

- Επαναλαμβάνουμε τα βήματα 3,4 μέχρι το σημείο όπου η σύσταση των συστάδων να παραμείνει σταθερή.

Τερματισμός του Αλγόριθμού

- Ο αλγόριθμος τερματίζεται όταν δεν υπάρχει σημείο Των δεδομένων που επανατοποθετείται σε διαφορετική συστάδα.
- Με άλλα λόγια ο αλγόριθμός τερματίζεται όταν οι συστάδες παραμένουν σταθερές ως προς τη σύσταση τους.

Παράδειγμα Εξέλιξης Αλγόριθμου $k - Means$



Ποιοτική Μεταβλητή στον Αλγόριθμο των $k - Means$

- Στις περιπτώσεις που στα δεδομένα μας έχουμε ποιοτική μεταβλητή με μόνο δύο τιμές, τότε την ποσοτικοποιούμε με **δείκτρια μεταβλητή** ή **ψευδομεταβλητή** (dummy variable).

Ποιοτική Μεταβλητή στον Αλγόριθμο των $k - Means$

Δηλαδή την αντικαθιστούμε με την μεταβλητή η οποία έχει

- **Την τιμή 1 όταν ή συνθήκη ισχύει.**
- **Την τιμή 0 όταν η συνθήκη δεν ισχύει.**

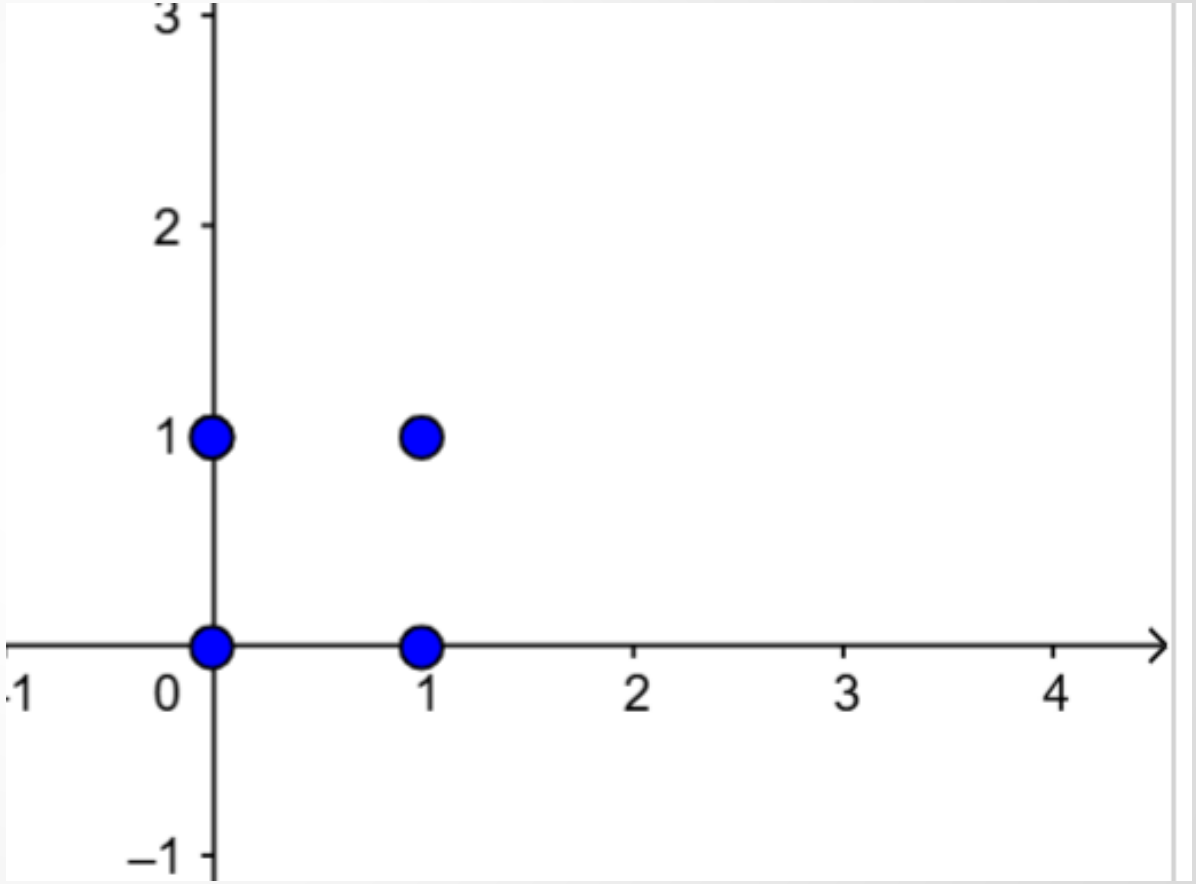
Ποιοτική Μεταβλητή στον Αλγόριθμο των $k - Means$

Σημείωση

Στις περιπτώσεις που στα δεδομένα μας υπάρχουν πολλές ποιοτικές μεταβλητές, ο αλγόριθμος

$k - means$ δεν θεωρείται ενδεδειγμένος, αφού σε περιπτώσεις ποιοτικών μεταβλητών στους άξονες j που αντιστοιχούν ποιοτικές μεταβλητές έχουμε τοποθέτηση σημείων μόνο στις θέσεις $x_j = 0$ και $x_j = 1$.

Ποιοτική Μεταβλητή στον Αλγόριθμο των $k - Means$



Βήματα Αλγόριθμου $k - Means$ – Σύνοψη

Στη γενική περίπτωση τα βήματα του αλγόριθμου $k - means$ μπορούν να συνοψιστούν ως εξής:

- **Βήμα 1**: Καθορίζουμε τον αριθμό των ομάδων (k) που θα δημιουργηθούν. (Ο καθορισμός γίνεται από τον ερευνητή).

- **Βήμα 2**: Επιλέγουμε (τυχαία) k αντικείμενα ως αρχικά κέντρα (βάρους) των ομάδων.

Βήματα Αλγόριθμου $k - Means$ – Σύνοψη

- **Βήμα 3**: Αντιστοιχούμε κάθε παρατήρηση στο πλησιέστερο κέντρο βάρους (με βάση συνήθως την ευκλείδεια απόσταση μεταξύ του αντικειμένου και του κέντρου βάρους). Με τον τρόπο αυτό δημιουργούμε τις αρχικές συστάδες.

- **Βήμα 4**: Για κάθε ομάδα, επαναϋπολογίζουμε το κέντρο της χρησιμοποιώντας τις νέες μέσες τιμές όλων των σημείων / δεδομένων που ανήκουν στη συστάδα.

Βήματα Αλγόριθμου $k - Means$ – Σύνοψη

- **Βήμα 5**: Επαναλαμβάνουμε τα βήματα 3 και 4 μέχρι να σταματήσουν οι ανακατανομές των σημείων στις ομάδες που δημιουργούνται ή να επιτευχθεί ένα μέγιστος (προεπιλεγμένος) αριθμός επαναλήψεων.

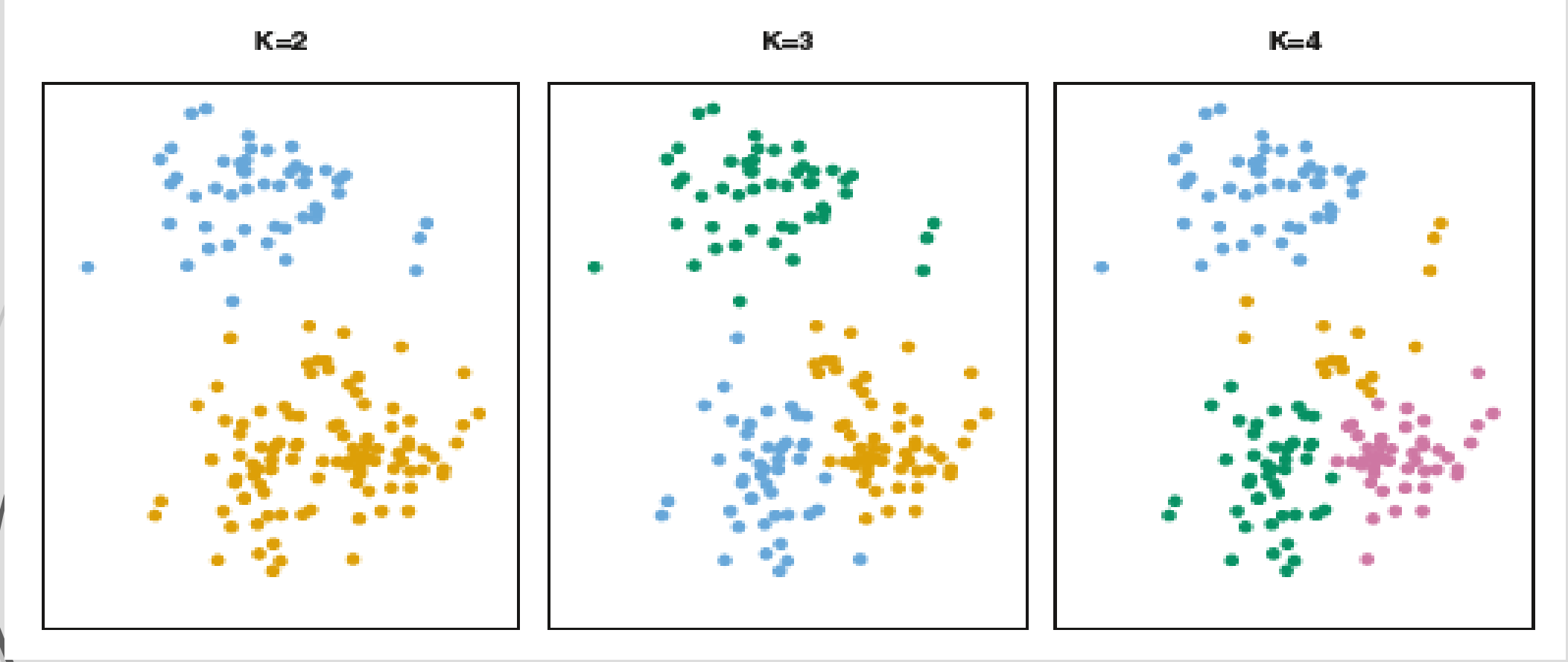
Βήματα Αλγόριθμου $k - Means$ – Σύνοψη

- **Τερματισμός Αλγόριθμου:**
- **Κριτήριο 1**: Έπειτα από συγκεκριμένο αριθμό επαναλήψεων όταν οι συστάδες παραμένουν σταθερές ως προς τη σύσταση τους.
- **Κριτήριο 2**: Όταν η μεγαλύτερη απόσταση ανάμεσα σε διαδοχικά κέντρα όλων των ομάδων γίνει 0.

την εφαρμογή Αλγόριθμου $k - Means$ – Σύνοψη

- Στο παρακάτω σχήμα παρουσιάζεται το αποτέλεσμα από την εφαρμογή του αλγορίθμου σε ένα σύνολο δεδομένων που περιέχει 150 παρατηρήσεις σε δύο διαστάσεις (δεδομένα δύο ποσοτικών μεταβλητών) χρησιμοποιώντας ενδεικτικά, τρείς διαφορετικές τιμές του k $(k = 2, k = 3, k = 4)$.

την εφαρμογή Αλγόριθμου $k - Means$ – Σύνοψη



Επιλογή Αριθμού Ομάδων

- Ένα ζήτημα που επηρεάζει τα αποτελέσματα της μεθόδου k-means είναι η επιλογή του αριθμού των ομάδων, δηλαδή ο προσδιορισμός της τιμής k.
- Έχουν προταθεί διάφορες μεθόδους για τον προσδιορισμός της κατάλληλης τιμής του k. Μία από αυτές είναι η μέθοδος του αγκώνα (elbow method).

Επιλογή Αριθμού Ομάδων

- Στη μέθοδο αυτή εξετάζεται η μεταβολή της διασποράς εντός των ομάδων καθώς αυξάνεται η τιμή του k κατά μία μονάδα κάθε φορά.

Παράδειγμα 1

Στον παρακάτω πίνακα έχουμε δεδομένα από 36 πελάτες ενός φυσικού καταστήματος.

Για κάθε έναν από τους πελάτες έχουμε δεδομένα για το φύλο του (1: Άντρας, 0: Γυναίκα), το έσοδα σε € που έχει το κατάστημα από τον πελάτη αυτόν (Total Purchase) καθώς και το ετήσιο εισόδημα του πελάτη (Yearly Revenue).

Παράδειγμα 1

	A	B	C	D	E
1					
2	Name	Total Purchase	Yearly Revenue	Gender	
3	A	5.000	100.000	1	
4	AA	1.700	25.000	2	
5	B	6.000	85.000	2	
6	BB	2.000	20.000	1	
7	C	7.000	105.000	2	
8	CC	3.000	30.000	1	
9	D	8.000	75.000	1	
10	DD	4.000	20.000	2	
11	E	9.000	110.000	1	
12	EE	5.000	43.000	2	
13	F	7.500	100.000	1	
14	FF	6.000	25.000	2	
15	G	6.500	105.000	2	
16	GG	7.000	30.000	2	
17	H	6.000	95.000	1	
18	HH	8.000	20.000	2	
19	I	6.500	110.000	1	
20	II	9.000	60.000	2	
21	J	7.000	100.000	2	
22	JJ	10.000	25.000	2	
23	K	7.500	105.000	2	
24	L	18.000	95.000	1	
25	M	16.000	110.000	2	
26	N	12.000	100.000	1	
27	O	11.000	105.000	1	
28	P	20.000	95.000	1	
29	Q	21.000	110.000	1	
30	R	22.000	100.000	1	
31	S	23.000	105.000	1	
32	T	24.000	95.000	2	
33	U	25.000	110.000	2	
34	V	26.000	100.000	2	
35	W	27.000	105.000	2	
36	X	28.000	95.000	2	
37	Y	29.000	110.000	1	
38	Z	30.000	100.000	2	
39					
40					

Παράδειγμα 1

Προκειμένου ο ιδιοκτήτης του φυσικού καταστήματος να οργανώσει αποτελεσματικότερα τις επιχειρηματικές του επενδύσεις θέλει να ομαδοποιήσει τους παραπάνω 36 πελάτες σε συστάδες ως προς τα έσοδα σε € που έχει το κατάστημα από τον πελάτη αυτόν (Total Purchase) και το ετήσιο εισόδημα του πελάτη (Yearly Revenue).

Παράδειγμα 1 – Υλοποίηση Αλγόριθμου $k - Means$ στο SPSS

Θα χρησιμοποιήσουμε την μέθοδο $k - means$ και θα δημιουργήσουμε $k = 2$ κλάσεις.

Η υλοποίηση θα γίνει μέσω του **SPSS.**

Εισαγωγή Δεδομένων από Αρχείο **Excel** στο **SPSS**

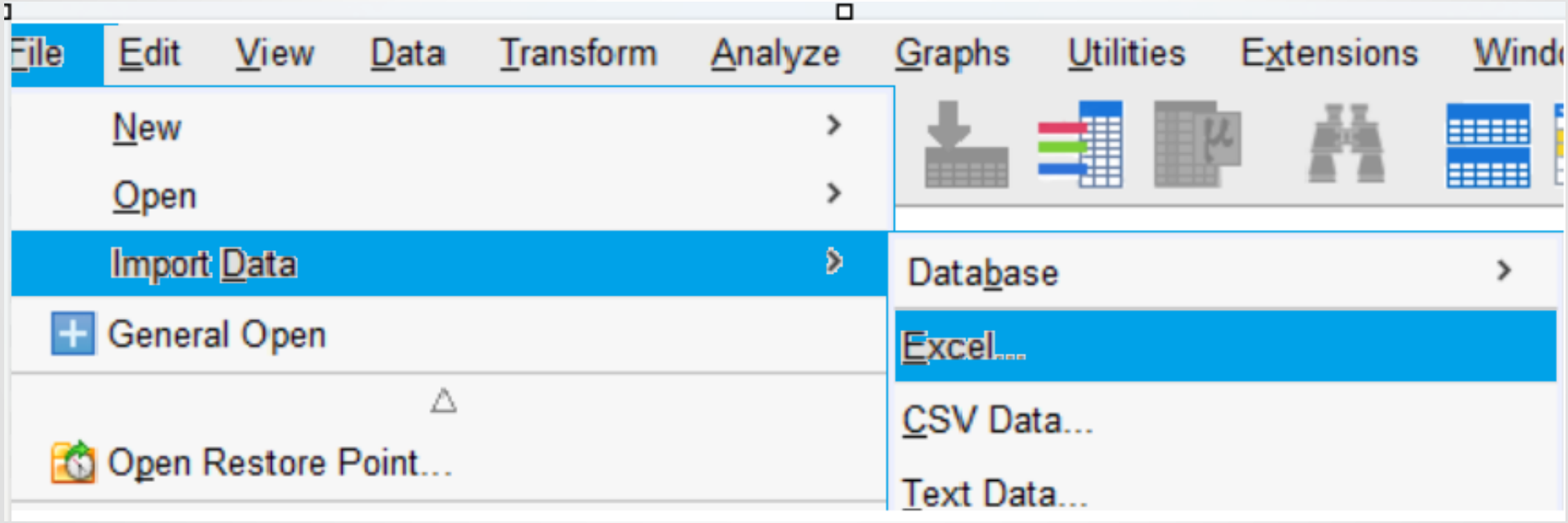
- Για εισαγωγή δεδομένων από αρχείο **Excel** στο **SPSS** ακολουθούμε στο SPSS τη διαδρομή

$$\boldsymbol{File \rightarrow Import\ Data \rightarrow Excel}$$

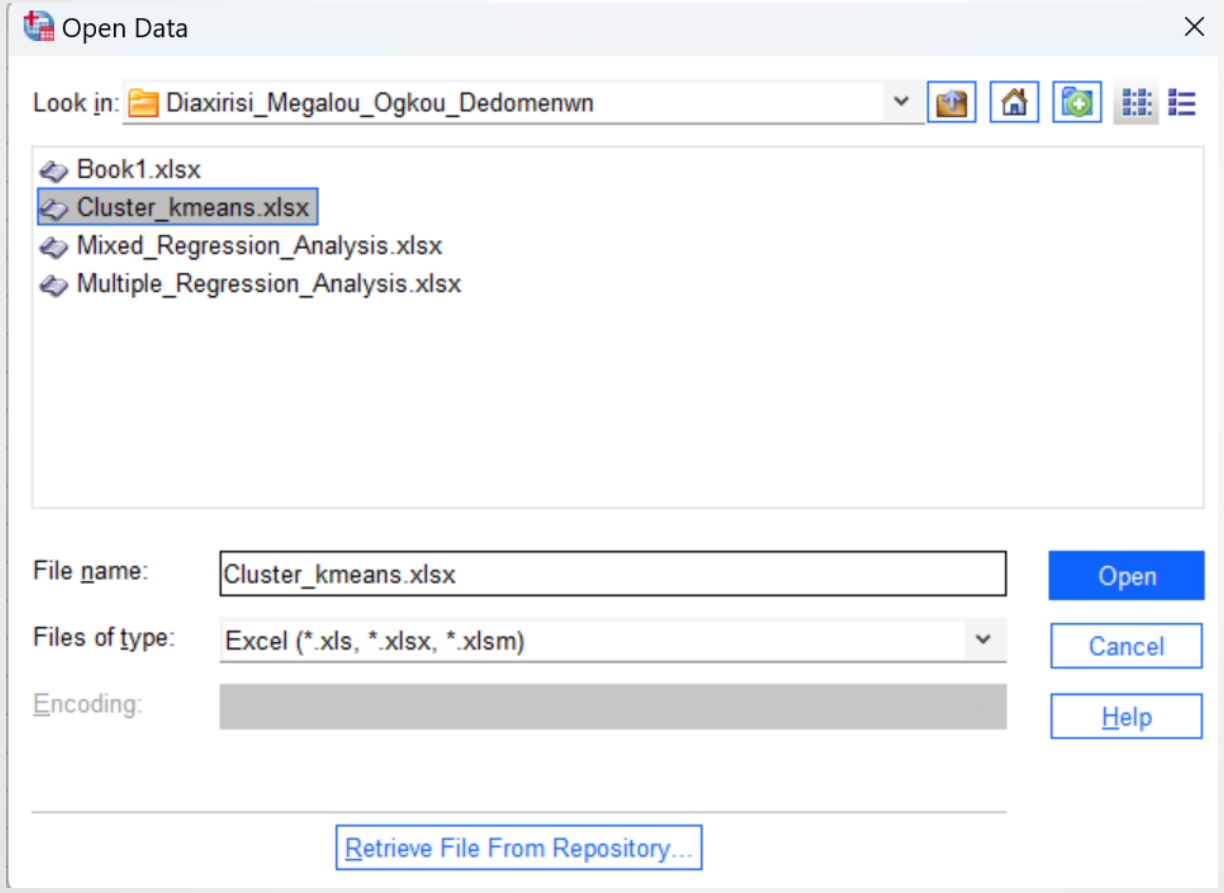
Όπως φαίνεται στο σχήμα.

Στη συνέχεια βρίσκουμε το αρχείο και επιλέγουμε **open**

Εισαγωγή Δεδομένων από Αρχείο **Excel** στο **SPSS**



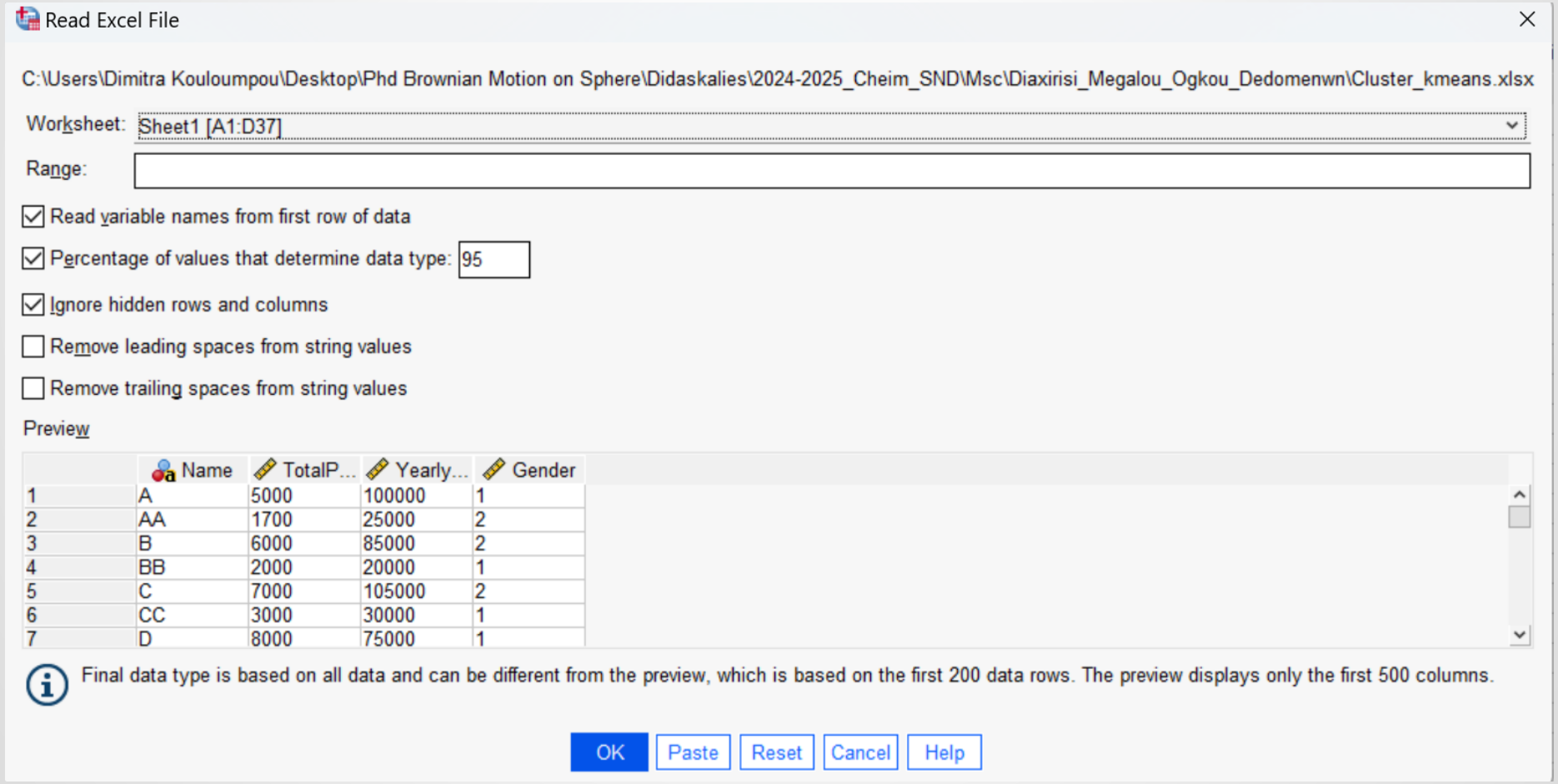
Εισαγωγή Δεδομένων από Αρχείο **Excel** στο **SPSS**



Εισαγωγή Δεδομένων από Αρχείο **Excel** στο **SPSS**

- Αν το αρχείο **excel** έχει πολλά φύλα εργασίας (worksheet) επιλέγουμε το κατάλληλο καθώς και το εύρος των στηλών και γραμμών που θέλουμε να εισάγουμε στο **SPSS**.

Εισαγωγή Δεδομένων από Αρχείο **Excel** στο **SPSS**



Υλοποίηση Αλγόριθμου $k - Means$ στο SPSS

- Θέλουμε να ομαδοποιήσουμε τους πελάτες του φυσικού καταστήματος με βάση τις τιμές των μεταβλητών **«Total Purchase»** και **«Yearly Revenue».**

- Θέλουμε να δημιουργήσουμε $k = 2$ **κλάσεις**.

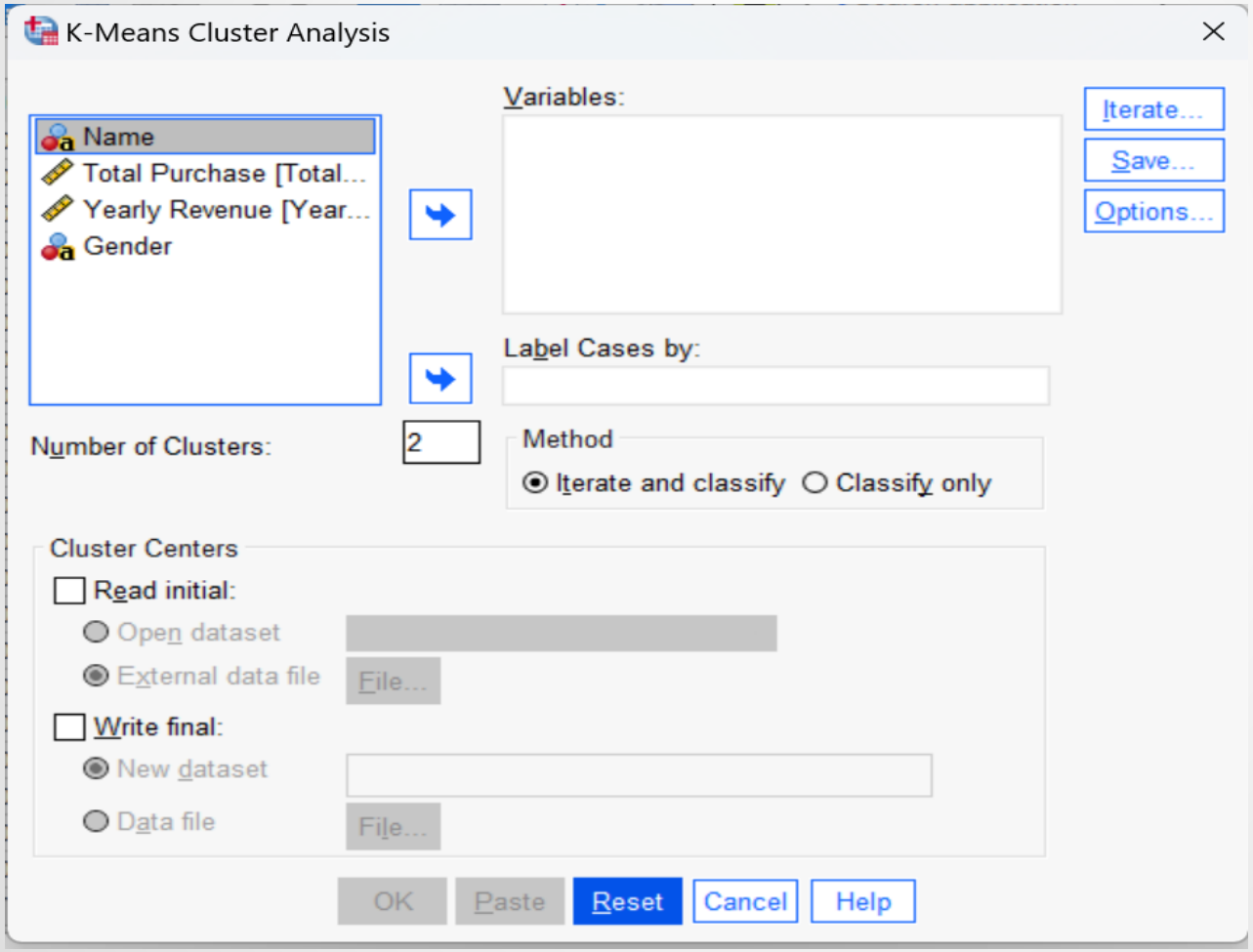
Υλοποίηση Αλγόριθμου $k - Means$ στο SPSS

- Η συσταδική ανάλυση με τον αλγόριθμο $k - means$ πραγματοποιείται ακολουθώντας την διαδρομή

$$\boldsymbol{Analyse \rightarrow Classify \rightarrow k - means\ Cluster}$$

Στη συνέχεια το SPSS ανοίγει το παρακάτω παράθυρο διαλόγου.

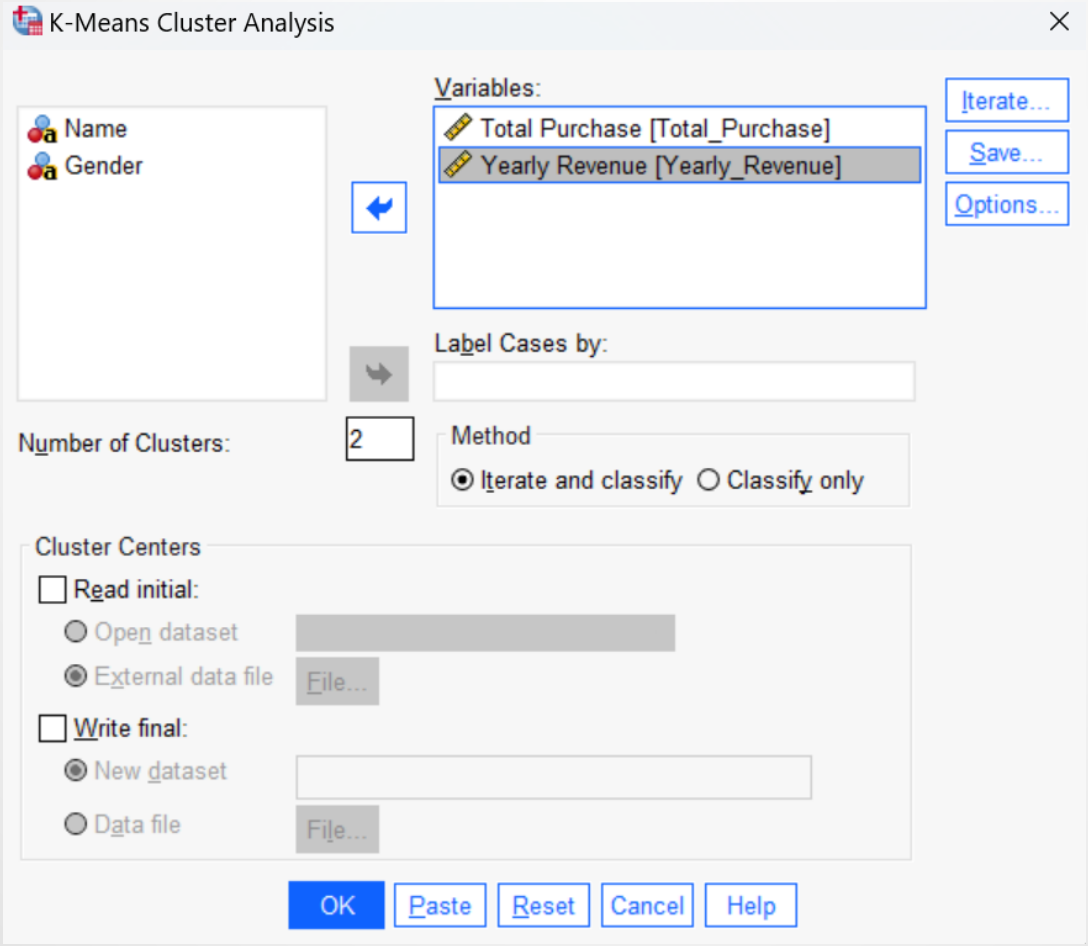
Υλοποίηση Αλγόριθμου $k - Means$ στο SPSS



Υλοποίηση Αλγόριθμου $k - Means$ στο SPSS

- Στο παράθυρο διαλόγου πρέπει να δηλώσουμε:
- Τις μεταβλητές βάση των οποίων θα πραγματοποιηθεί η συσταδική ανάλυση **(τις τοποθετούμε στο πλαίσιο Variables).**
- Τον αριθμό των ομάδων που θέλουμε να σχηματιστούν **(Το δηλώνουμε στο πλαίσιο *Number of Clusters*).**

Υλοποίηση Αλγόριθμου $k - Means$ στο SPSS



Υλοποίηση Αλγόριθμου $k - Means$ στο SPSS

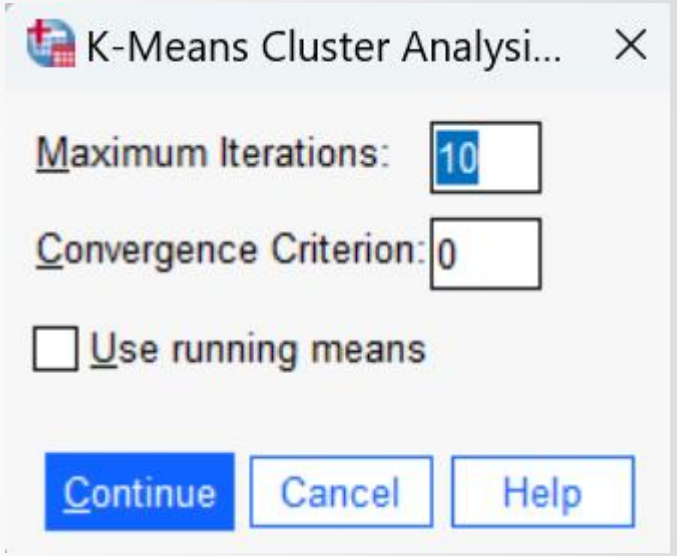
- Στο ίδιο παράθυρο διαλόγου και ΠΡΙΝ πατήσουμε ΟΚ διαμορφώνουμε κατάλληλες ρυθμίσεις από τις τρεις επιλογές

Iterate **Save** **Options**

Κάθε μία από τις παραπάνω επιλογές ανοίγει και ένα νέο πλαίσιο διαλόγου όπως θα δούμε στα παρακάτω σχήματα.

Υλοποίηση Αλγόριθμου $k - Means$ στο SPSS – Iterate

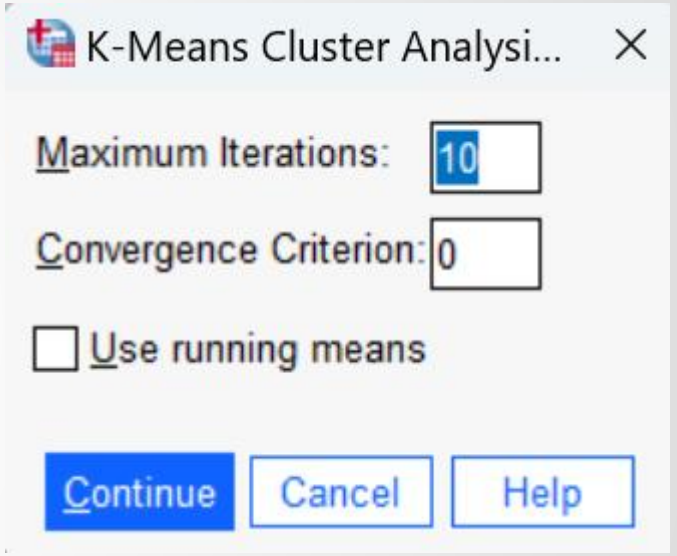
- Στο παράθυρο διαλόγου **Iterate** Επιλέγουμε τα κριτήρια τερματισμού του αλγορίθμου.



Υλοποίηση Αλγόριθμου $k - Means$ στο SPSS – Iterate

- Μπορούμε να επιλέξουμε να σταματήσει έπειτα από συγκεκριμένο αριθμό επαναλήψεων (π.χ. 10 στην περίπτωσή μας).

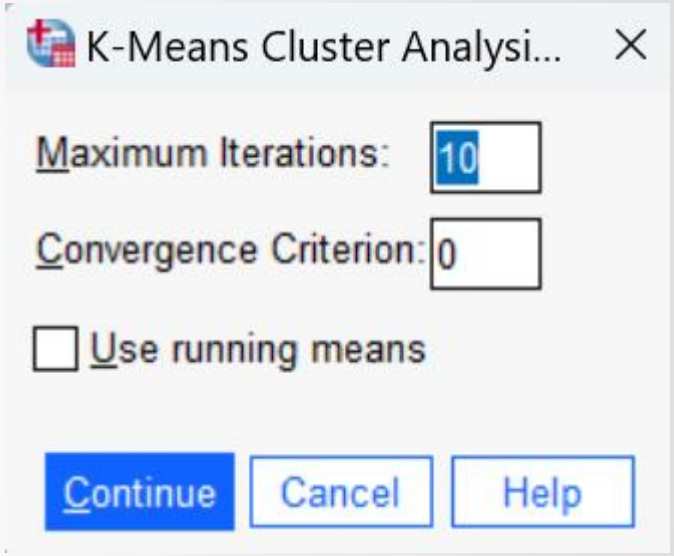
(Πρώτο κριτήριο).



Υλοποίηση Αλγόριθμου $k - Means$ στο SPSS – Iterate

- Είτε όταν η μεγαλύτερη απόσταση ανάμεσα σε διαδοχικά κέντρα όλων των ομάδων γίνει 0.

(Δεύτερο κριτήριο).



Παράδειγμα 1 – Υλοποίηση Αλγόριθμου $k - Means$ στο SPSS – Iterate

- Το δεύτερο κριτήριο αντιστοιχεί στην περίπτωση που οι ομάδες δεν αλλάξουν καθόλου μετά από μία επανάληψη.

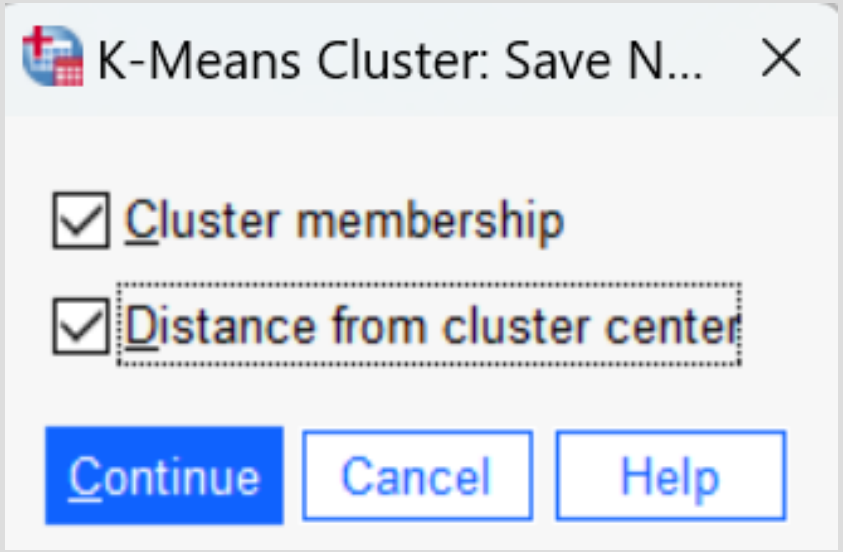
Παράδειγμα 1 – Υλοποίηση Αλγόριθμου $k - Means$ στο SPSS – Iterate

- Αν το δεύτερο κριτήριο επιτευχθεί πριν το πρώτο τότε η επαναληπτική διαδικασία σταματά πριν ολοκληρωθούν όλες οι επαναλήψεις, διαφορετικά σταματά όταν πραγματοποιηθεί ο μέγιστος αριθμός(10 στην προκειμένη περίπτωση) επαναλήψεων.

Αφού ολοκληρώσουμε τις ρυθμίσεις στο πλαίσιο διαλόγου **Iterate**, πατάμε **Continue.**

Παράδειγμα 1 – Υλοποίηση Αλγόριθμου $k - Means$ στο SPSS – Save

- Πατώντας το κουμπί επιλογών **Save** επιλέγουμε τις καινούριες μεταβλητές που θέλουμε να δημιουργήσουμε.



Παράδειγμα 1 – Υλοποίηση Αλγόριθμου $k - Means$ στο SPSS – Save

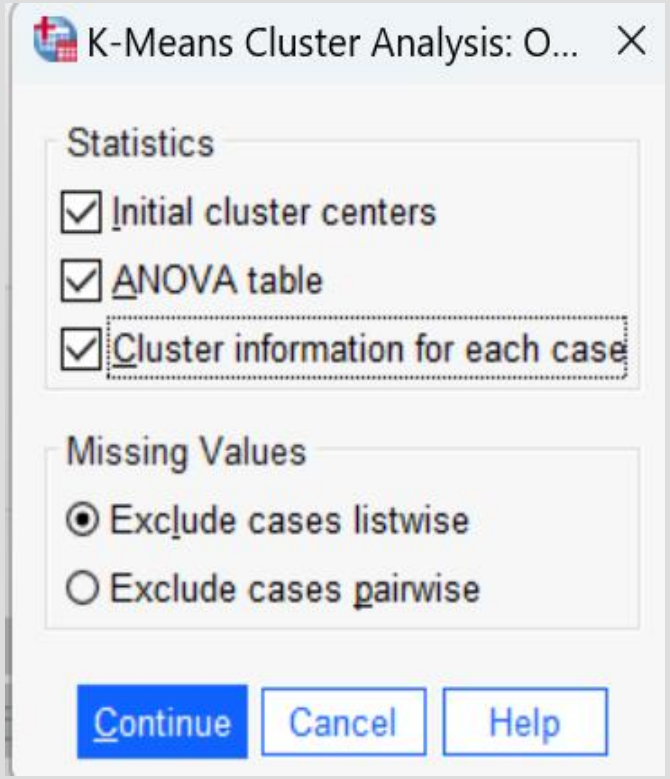
- Η επιλογή ***Cluster membership*** θα μας δημιουργήσει μια καινούρια στήλη όπου σε κάθε παρατήρηση θα δίνεται η τιμή της ομάδας που την κατατάξαμε.

Παράδειγμα 1 – Υλοποίηση Αλγόριθμου $k - Means$ στο SPSS – Save

- Αυτή η μεταβλητή είναι ιδιαίτερα χρήσιμη, όταν μετά την ανάλυση προσπαθήσουμε να δούμε τα χαρακτηριστικά κάθε ομάδας, αλλά και το αν κάποιες μεταβλητές προσφέρουν πληροφορία σχετικά με την ομαδοποίηση που κάναμε.

Παράδειγμα 1 – Υλοποίηση Αλγόριθμου $k - Means$ στο SPSS – Options

- Τέλος, πατώντας το κουμπί επιλογών **Options** μπορούμε να επιλέξουμε ποια αποτελέσματα θα εμφανιστούν στην οθόνη. Μπορούμε να επιλέξουμε να μας δώσει ο υπολογιστής:



Παράδειγμα 1 – Υλοποίηση Αλγόριθμου $k - Means$ στο SPSS – Options

- ***Initial cluster centers:*** Τα αρχικά κέντρα των ομάδων, η χρησιμότητά των οποίων έγκειται στο να ελέγξουμε κατά πόσο αλλάζουν τα αποτελέσματά σε διαδοχικές εκτελέσεις του αλγορίθμου,

Παράδειγμα 1 – Υλοποίηση Αλγόριθμου $k - Means$ στο SPSS – Options

- ***Cluster information for each case:*** Τα τελικά κέντρα των ομάδων και τις Ευκλείδειες αποστάσεις μεταξύ των ατόμων και των κέντρων των συστάδων που χρησιμοποιήθηκαν για την ταξινόμηση κάθε ατόμου.

Παράδειγμα 1 – Υλοποίηση Αλγόριθμου $k - Means$ στο SPSS – Options

- ***ANOVA table:*** Τους πίνακες ανάλυσης διακύμανσης για τις μεταβλητές που χρησιμοποιήσαμε, ώστε να βρούμε ποιες περιέχουν όντως πληροφορία για την ομαδοποίηση που κάναμε.

Παράδειγμα 1 – Υλοποίηση Αλγόριθμου $k - Means$ στο SPSS – Αποτελέσματα Output

Στον πίνακα **Cluster Membership**, βλέπουμε σε ποια ομάδα – cluster ανήκει κάθε εγγραφή καθώς και η απόσταση της από το κέντρο της αντίστοιχης κλάσης.

Cluster Membership

Case Number	Cluster	Distance
1	1	10735,455
2	2	6165,785
3	1	18673,820
4	2	10430,000
5	1	9584,643
6	2	2577,770
7	1	27077,169
8	2	9924,963
9	1	11246,367
10	2	13212,301
11	1	8248,543
12	2	4819,222
13	1	10040,303
14	2	1443,918
15	1	11378,962
16	2	10096,777
17	1	12891,560
18	2	30394,159
19	1	8745,328
20	2	6531,837
21	1	9133,623
22	1	6392,604
23	1	9043,697
24	1	3815,454
25	1	6190,874
26	1	7355,009
27	1	10481,669
28	1	6380,559
29	1	8349,344
30	1	10225,346
31	1	12974,087
32	1	10352,443
33	1	12007,209
34	1	13675,497
35	1	16086,903
36	1	14339,966

Παράδειγμα 1 – Υλοποίηση Αλγόριθμου $k - Means$ στο SPSS – Αποτελέσματα Output

Εδώ βλέπουμε τα κέντρα των κλάσεων στο αρχικό και τελικό βήμα.

Initial Cluster Centers

	Cluster	
	1	2
Total Purchase	29000	2000
Yearly Revenue	110000	20000

Final Cluster Centers

	Cluster	
	1	2
Total Purchase	15692	5570
Yearly Revenue	100962	29800

Παράδειγμα 1 – Υλοποίηση Αλγόριθμου $k - Means$ στο SPSS – Αποτελέσματα Output

Ευκλείδειες αποστάσεις μεταξύ των κέντρων των κλάσεων

Distances between Final Cluster Centers

Cluster	1	2
1		71877,852
2	71877,852	

Παράδειγμα 1 – Υλοποίηση Αλγόριθμου $k - Means$ στο SPSS – Αποτελέσματα Output

Το πλήθος των στοιχείων που έχει κάθε κλάση

Number of Cases in each Cluster

Cluster	1	26,000
	2	10,000
Valid		36,000
Missing		,000

Παράδειγμα 1 – Υλοποίηση Αλγόριθμου $k - Means$ στο SPSS – Αποτελέσματα Output

Ο αριθμός των επαναλήψεων που χρειάστηκε ο αλγόριθμος $k - means$.

Iteration History[a]

Iteration	Change in Cluster Centers 1	2
1	16086,903	10430,000
2	,000	,000

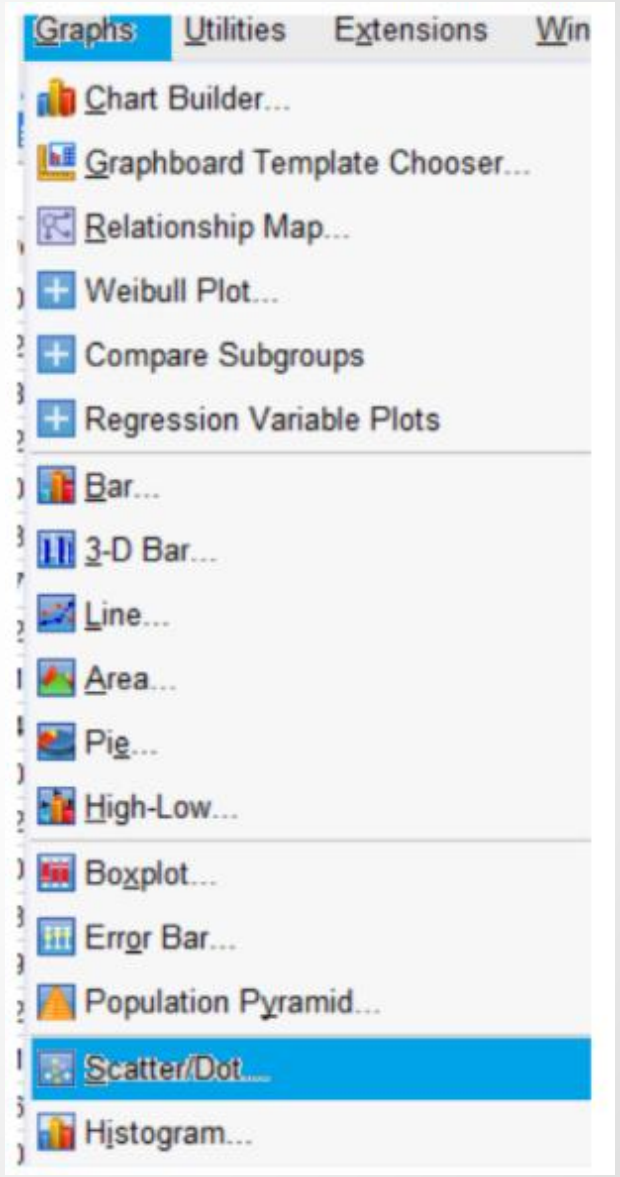
a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is ,000. The current iteration is 2. The minimum distance between initial centers is 93962,759.

Παράδειγμα 1 – Υλοποίηση Αλγόριθμου $k - Means$ στο SPSS – Γραφική Απεικόνιση

- Το διάγραμμα διασποράς με την συσταδοποίηση που δημιούργησε ο αλγόριθμος $k - means$ υλοποιείται μέσω της διαδρομής.

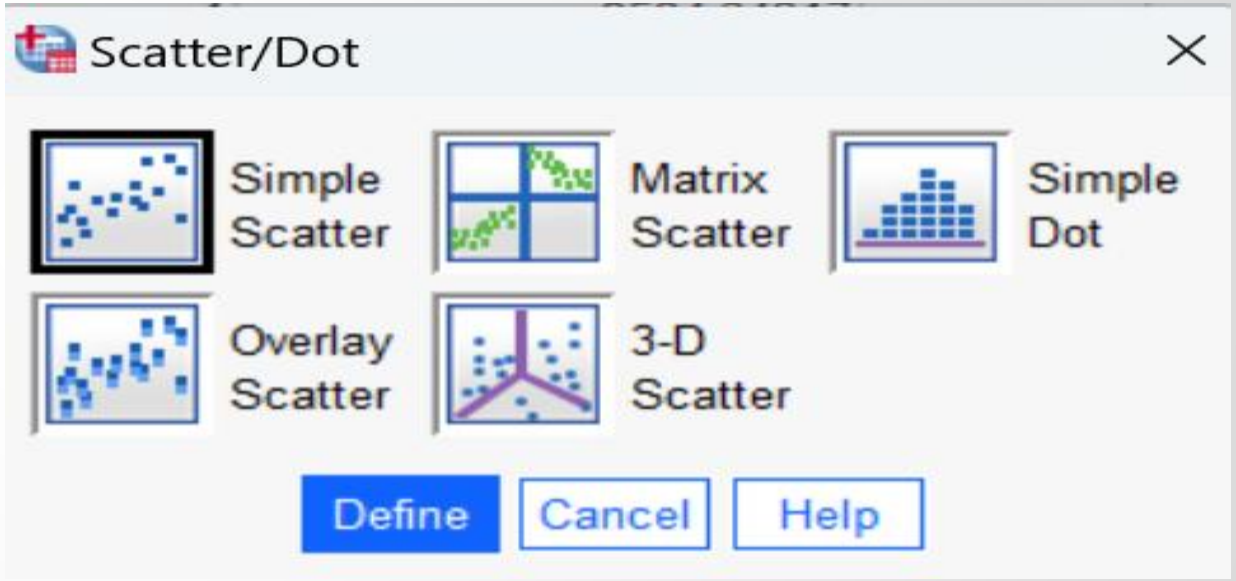
$$Graphs \rightarrow Scatter\ Plot$$

Παράδειγμα 1 – Υλοποίηση Αλγόριθμου k – $Means$ στο SPSS – Γραφική Αναπαράσταση



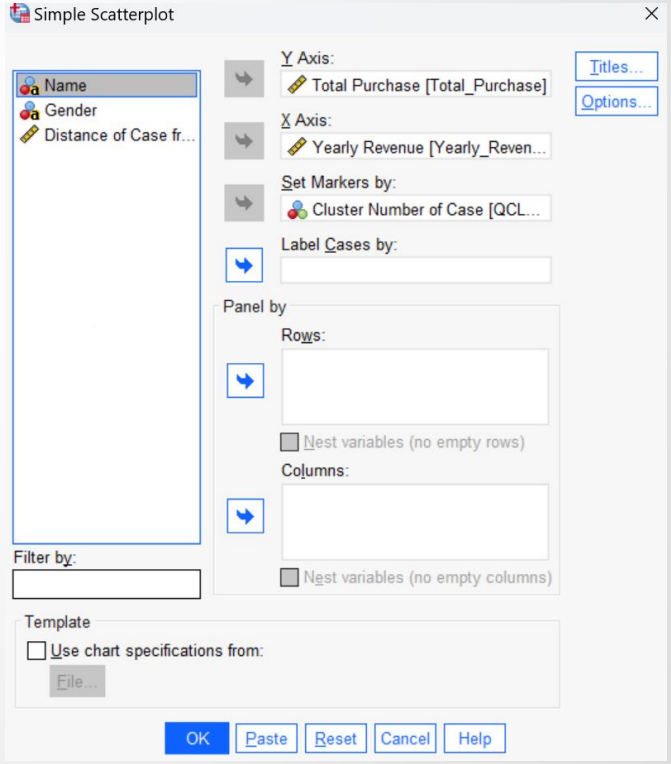
Παράδειγμα 1 – Υλοποίηση Αλγόριθμου $k-Means$ στο SPSS – Γραφική Αναπαράσταση

Επιλέγουμε Simple Scatter και μετά Define

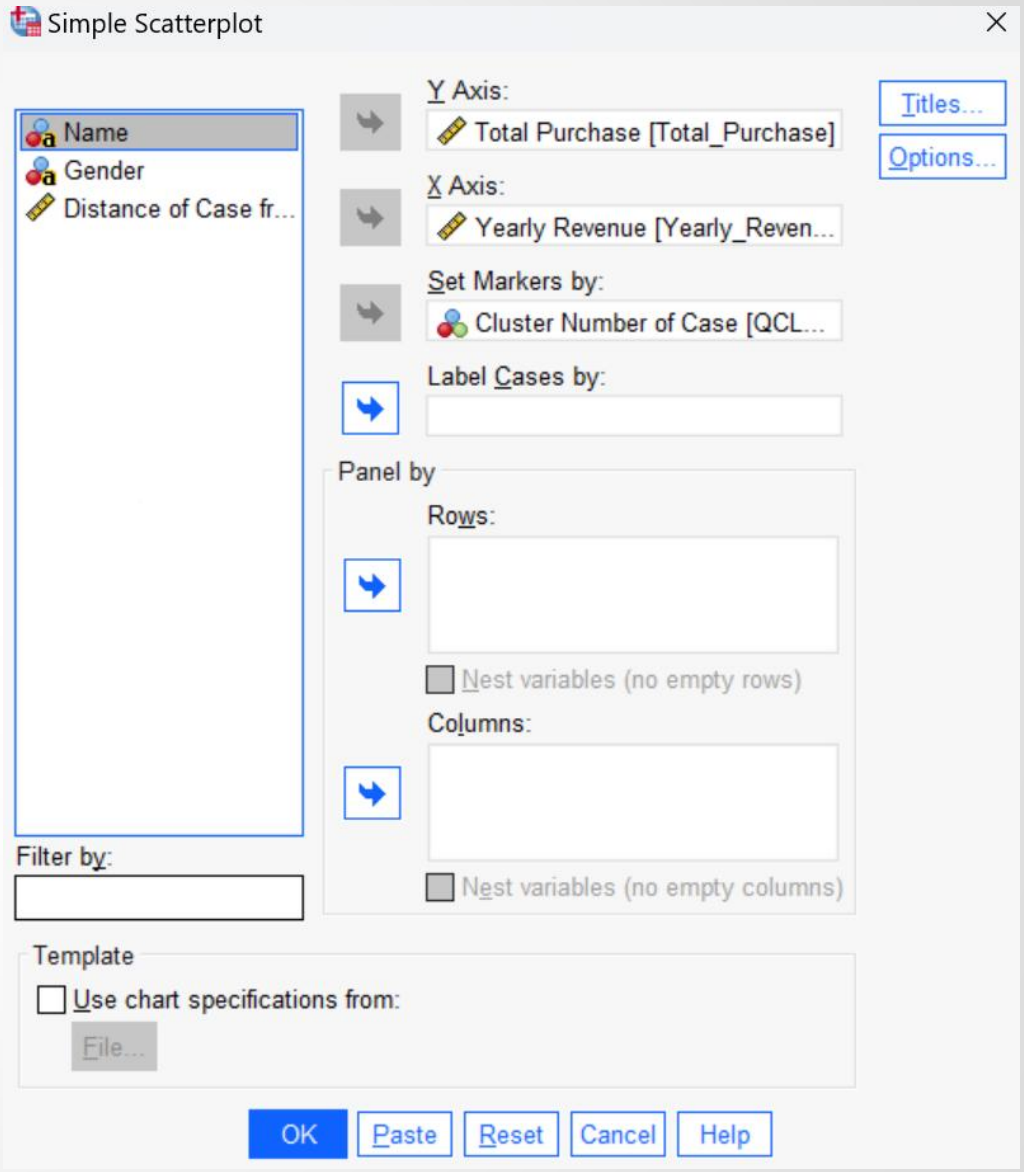


Παράδειγμα 1 – Υλοποίηση Αλγόριθμου k – $Means$ στο SPSS – Γραφική Αναπαράσταση

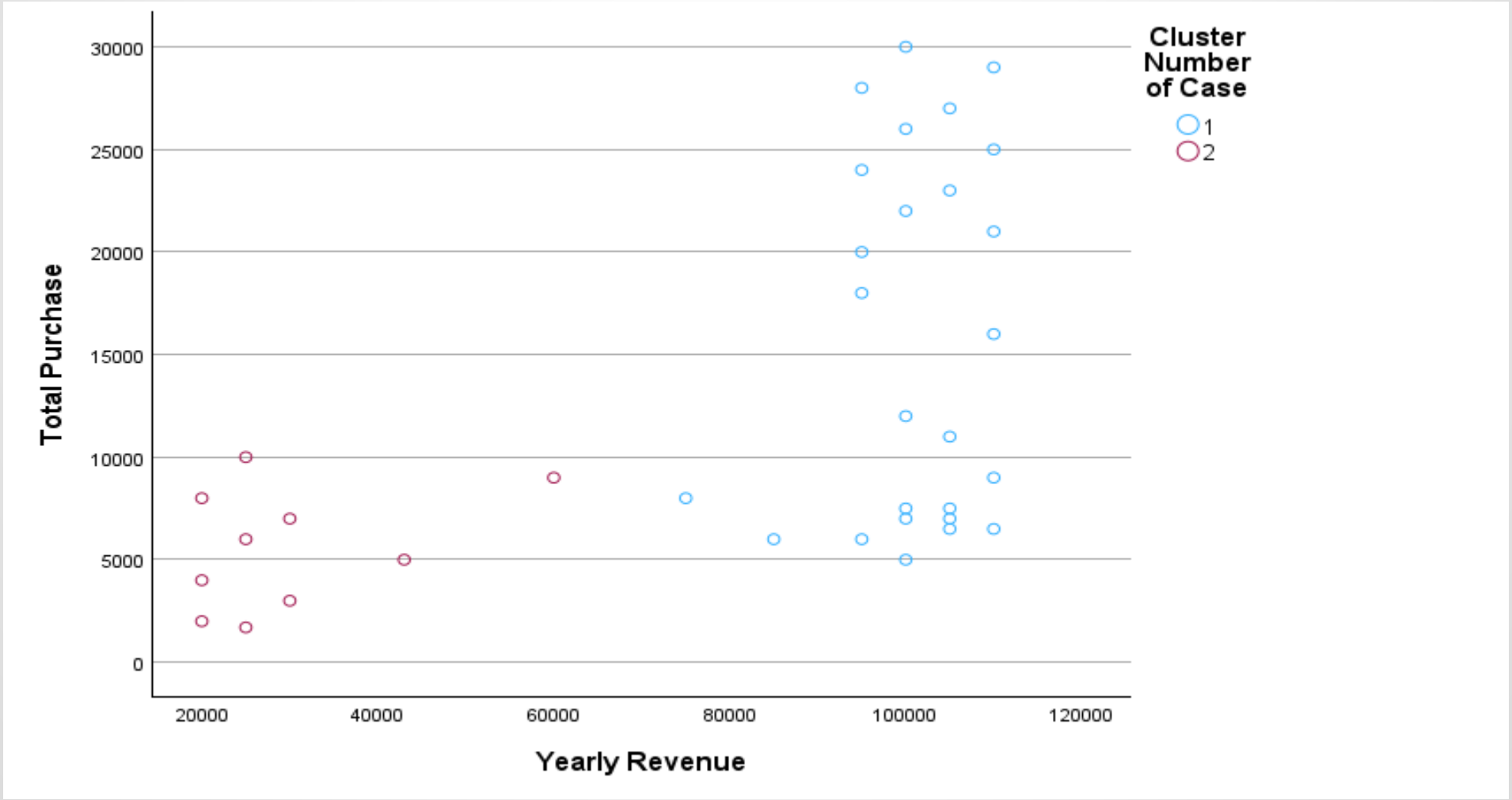
Τοποθετούμε τις κατάλληλες μεταβλητές στα κατάλληλα πλαίσια διαλόγου



Παράδειγμα 1 – Υλοποίηση Αλγόριθμου k – $Means$ στο SPSS – Γραφική Αναπαράσταση



Παράδειγμα 1 – Υλοποίηση Αλγόριθμου $k-Means$ στο SPSS – Γραφική Αναπαράσταση



Παράδειγμα 1 – Υλοποίηση Αλγόριθμου $k - Means$ Online

- Μέσω του διαδικτύου μπορούμε επίσης να εφαρμόσουμε τον αλγόριθμο k-means.

https://www.statskingdom.com/cluster-analysis.html

Μετατρέποντας τα Δεδομένα μας σε Ίδια Κλίμακα (Scaling)

- Εάν όλες οι διαστάσεις έχουν τις ίδιες μονάδες μέτρησης δεν χρειάζεται να μετατρέψουμε την κλίμακα των δεδομένων.
- Όταν όμως τα δεδομένα μας έχουν διαφορετικές μονάδες μέτρησης (π.χ. Kg και €) πρέπει να μετατρέψουμε την κλίμακα των δεδομένων μας.

Μετατρέποντας τα Δεδομένα μας σε Ίδια Κλίμακα (Scaling)

- Για παράδειγμα αν έχουμε ύψος και βάρος, θα προκύψουν διαφορετικά αποτελέσματα αν το ύψος μετρηθεί σε m από ότι σε mm. Αφού στη δεύτερη περίπτωση το ύψος είναι αυτό που επηρεάζει το αποτέλεσμα.
- Η μετατροπή των δεδομένων μας στην ίδια κλίμακα γίνεται με τη χρήση της παρακάτω πρότασης.

Μετατρέποντας τα Δεδομένα μας σε Ίδια Κλίμακα (Scaling)

Πρόταση:

Αν η τυχαία μεταβλητή X ακολουθεί την Κανονική κατανομή με παραμέτρους μ, σ^2 , δηλαδή

$X \sim N(\mu, \sigma^2)$ τότε η τυχαία μεταβλητή

$Z = \frac{X-\mu}{\sigma}$ ακολουθεί την τυποποιημένη κανονική κατανομή, δηλαδή

$$Z \sim N(0, 1)$$

Μετατρέποντας τα Δεδομένα μας σε Ίδια Κλίμακα (Scaling)

- Αν $x_1, x_2, \dots, x_n$ τα δεδομένα μιας μεταβλητής x τα οποία έχουν μέση τιμή $\bar{x}$ και τυπική απόκλιση s, τότε η **κανονικοποιημένη τιμή της x_i** ονομάζεται η

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

Μετατρέποντας τα Δεδομένα μας σε Ίδια Κλίμακα (Scaling)

- Αν τα $x_1, x_2, \dots, x_n$ ακολουθούν κανονική κατανομή, τότε σύμφωνα με την παραπάνω πρόταση τα $z_1, z_2, \dots, z_n$ ακολουθούν τυποποιημένη κανονική κατανομή.

Μετρικές Συναρτήσεις – Manhattan Distance

Manhattan Distance: Γνωστή και ως **απόσταση L_1**.

Υπολογίζει την απόσταση δύο σημείων με βάση το άθροισμα των απόλυτων διαφορών των συντεταγμένων τους.

- Για δύο σημεία $x = (x_1, x_2, \dots, x_n), y = (y_1, y_2, \dots, y_n) \in \mathbb{R}^n$ η Manhattan απόσταση ισούται με

$$\boldsymbol{d = \sum_{i=1}^{n} |x_i - y_i|}$$

Μετρικές Συναρτήσεις – Chebyshev Distance

Chebyshev Distance: ή **απόσταση $mini - max$.**
Υπολογίζει την μεγαλύτερη απόλυτη διαφορά μεταξύ των συντεταγμένων δύο σημείων.

Για δύο σημεία $x = (x_1, x_2, \dots, x_n), y = (y_1, y_2, \dots, y_n) \in \mathbb{R}^n$ η Chebyshev απόσταση ισούται με

$$d = \max_{i=1,2,\dots,n} |x_i - y_i|$$

Μετρικές Συναρτήσεις – Minkowski Distance τάξης p

Minkowski Distance τάξης p:

Η απόσταση Minkowski αποτελεί γενίκευση της ευκλείδειας απόστασης και της απόστασης Manhattan.

Για δύο σημεία $x = (x_1, x_2, \dots, x_n), y = (y_1, y_2, \dots, y_n) \in \mathbb{R}^n$ η Minkowski απόσταση τάξης p ισούται με

$$\boldsymbol{d = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}}}$$

Μετρικές Συναρτήσεις – Minkowski Distance τάξης p

- Για $p = 2$, η Minkowski απόσταση τάξης 2 ισοδυναμεί με την ευκλείδεια απόσταση.
- Για $p = 1$, η Minkowski απόσταση τάξης 1 ισοδυναμεί με την Manhattan απόσταση.

Μετρικές Συναρτήσεις – Απόσταση Hamming (Για κατηγορικές Μεταβλητές)

- Απόσταση Hamming: Χρησιμοποιείται για δυαδικές ή κατηγορικές μεταβλητές. Υπολογίζει τον αριθμό των θέσεων στις οποίες οι αντίστοιχες τιμές διαφέρουν

Επιλογή Απόστασης για την Συσταδοποίηση

- Σε κάθε εφαρμογή του αλγόριθμου, η επιλογή της μετρικής αποφασίζεται από τον ερευνητή με βάση τη φύση των υπό μελέτη δεδομένων και τον στόχο της ανάλυσης.

Εφαρμογή

Έστω τα σημεία $x = (1,3)$ και $y = (-2,5)$

- Η **Ευκλείδεια απόσταση** των x, y ισούται με

$$d_2 = \sqrt{(-2-1)^2 + (5-3)^2} = \sqrt{13}$$

- Η **Manhattan απόσταση** των x, y ισούται με

$$d = |-2-1| + |5-3| = 5$$

- Η **Chebyshev απόσταση** των x, y ισούται με

$$d = max\{|-2-1|, |5-3|\} = 3$$

Εφαρμογή

- Η **Minkowski απόσταση** τάξης p των x, y ισούται με

$$d_p = [|-2-1|^p + |5-3|^p]^{\frac{1}{p}}$$

Δηλαδή

$$d_p = [3^p + 2^p]^{\frac{1}{p}}$$

Μετρική Συνάρτηση

- Μία συνάρτηση $d: X \times X \longrightarrow \mathbb{R}$ ορίζεται ως συνάρτηση απόστασης (ή μετρική) όταν ικανοποιεί και τις τέσσερις ακόλουθες ιδιότητες, για κάθε $x, y, z \in X$

➢ **Μη αρνητικότητα**

$$\boldsymbol{d(x, y) \geq 0}$$

Μετρική Συνάρτηση

- **Ταυτοτική Ιδιότητα του Μηδενός**

$$d(x, y) = 0 \Leftrightarrow x = y$$

Δηλαδή, δύο διαφορετικά σημεία δεν έχουν ποτέ μηδενική απόσταση.

- **Συμμετρία**

$$d(x, y) = d(y, x)$$

Μετρική Συνάρτηση – Μετρικό Χώρος

Τριγωνική Ανισότητα

$$\boldsymbol{d(x,z) \leq d(x,y) + d(y,z)}$$

- Εάν μια συνάρτηση ικανοποιεί και τις τέσσερις παραπάνω ιδιότητες, τότε ονομάζεται μετρική και το σύνολο (X, d) ονομάζεται μετρικός χώρος.

Μετρικές Απόστασης στη Μέθοδο $k - means$

- Οι μετρικές που χρησιμοποιούνται ευρέως στη μέθοδο $k - means$ είναι αυτή της ευκλείδειας απόστασης. Άλλες μετρικές αποστάσεις στη μέθοδο περιλαμβάνουν:

Πλεονεκτήματα Μεθόδου $k - Means$

- Τα κυριότερα πλεονεκτήματα του αλγόριθμου $k - means$ είναι τα εξής:

- Είναι απλός και εύκολος στην εφαρμογή του.
- Χρειάζεται λιγότερο χρόνο και απαιτεί πολύ μικρότερο κόστος επεξεργασίας για μεγάλα σύνολα δεδομένων σε αντίθεση με τις περισσότερες από τις ιεραρχικές μεθόδους.
- Οι ομάδες που δημιουργεί είναι συμπαγείς, σφαιρικές και με σχεδόν ίσο αριθμό στοιχείων.

Μειονεκτήματα Μεθόδου $k - Means$

- Τα κύρια μειονεκτήματα του αλγόριθμου είναι τα εξής:
  - Δεν είναι καθόλου αποτελεσματικός για ποιοτικές μεταβλητές
  - Απαιτεί την επιλογή του αριθμού k των ομάδων πριν την εφαρμογή του

Μειονεκτήματα Μεθόδου $k - Means$

- Εξαρτάται από την επιλογή των αρχικών κεντροειδών. Διαφορετικά αρχικά κεντροειδή δίνουν, συνήθως διαφορετική συσταδοποιήση.
- Είναι ευαίσθητος στην ύπαρξη ακραίων τιμών. Στοιχεία με κάποια ή κάποιες ακραίες τιμές σε κάποιες από τις συντεταγμένες τους επηρεάζουν συνήθως των καθορισμών των νέων κεντροειδών και αυτό επηρεάζει τη σύσταση των ομάδων.

Μειονεκτήματα Μεθόδου $k - Means$

- Είναι ευαίσθητος στην κλίμακα μέτρησης των δεδομένων. Η μετατροπή των δεδομένων σε ίδια κλίμακα μπορεί να επηρεάσει σημαντικά τα αποτελέσματα.

Εφαρμογές της Μεθόδου $k - Means$

Η μέθοδος k-means βρίσκει εφαρμογή σε πολλούς τομείς όπως

- **Ανάλυση Δεδομένων και Εξόρυξη Δεδομένων:** Η $k - means$ χρησιμοποιείται ευρέως στην ανάλυση δεδομένων για την κατηγοριοποίηση μεγάλων συνόλων δεδομένων. Με τη χρήση της μεθόδου δίνεται η δυνατότητα να ανακαλυφθούν κρυφά πρότυπα και σχέσεων στα δεδομένα, επιτρέποντας στους αναλυτές να κατανοήσουν καλύτερα τις δομές των δεδομένων τους.

Εφαρμογές της Μεθόδου $k - Means$

- **Επεξεργασία Εικόνας:** Στην επεξεργασία εικόνας, η μέθοδος $k - means$ χρησιμοποιείται για την ομαδοποίηση χρωμάτων και τη συμπίεση εικόνας. Με την ομαδοποίηση των χρωμάτων σε κλάσεις, οι εικόνες μπορούν να αναπαρασταθούν με λιγότερους χρωματικούς τόνους, μειώνοντας έτσι το μέγεθος του αρχείου χωρίς σημαντική απώλεια ποιότητας.

Εφαρμογές της Μεθόδου $k - Means$

- **Μάρκετινγκ και Στόχευση Πελατών:** Στον τομέα του μάρκετινγκ, η k-means μπορεί να χρησιμοποιηθεί προκειμένου το αγοραστικό κενό να κατηγοριοποιηθεί με βάση χαρακτηριστικά όπως οι αγοραστικές συνήθειες, το ετήσιο εισόδημα, τα δημογραφικά στοιχεία, κ.α.. Αυτό δίνει τη δυνατότητα στις επιχειρήσεις να στοχεύουν συγκεκριμένες ομάδες πελατών με πιο εξατομικευμένες στρατηγικές μάρκετινγκ.

Εφαρμογές της Μεθόδου $k - Means$

- **Συντήρηση Σκαφών στο Πολεμικό Ναυτικό:** Στην διαχείριση των πόρων του Πολεμικού Ναυτικού, η μέθοδος $k - means$ μπορεί να εφαρμοστεί για την ομαδοποίηση σκαφών ανάλογα με τις ανάγκες συντήρησης τους ή την παρούσα κατάσταση τους. Αυτό επιτρέπει στους υπεύθυνους συντήρησης να κατανοήσουν ποια πλοία χρειάζονται άμεσες επισκευές ή αναβαθμίσεις και ποια είναι σε καλή κατάσταση.

Εφαρμογές της Μεθόδου $k - Means$

- **Χρηματοοικονομικά:** Στα χρηματοοικονομικά, η k-means εφαρμόζεται την ανάλυση κινδύνου και στη διαχείριση χαρτοφυλακίων. Οι επενδυτές ομαδοποιούν τα χρηματοοικονομικά προϊόντα και τους πελάτες με παρόμοια χαρακτηριστικά κινδύνου.

Εφαρμογές της Μεθόδου $k - Means$

- **Ανάλυση Δεδομένων από Συστήματα Ελέγχου Πλοίων:** Τα σύγχρονα πλοία διαθέτουν πληθώρα αισθητήρων που συλλέγουν δεδομένα σχετικά με τη λειτουργία τους. Η μέθοδος $k - means$ μπορεί να εφαρμοστεί για την ανάλυση αυτών των δεδομένων ώστε να εντοπιστούν ανωμαλίες ή προβλήματα στη λειτουργία των συστημάτων ελέγχου.

Η Μέθοδος $k - Means$ στην Διαχείριση Μεγάλου Όγκου Δεδομένων

Μια πολύ χαρακτηριστική εφαρμογή του $k - means$, που συνδέει διαχείριση μεγάλου όγκου δεδομένων και AI, είναι η εξής:

- Ομαδοποίηση χρηστών (user clustering) με βάση τη συμπεριφορά τους, ώστε να βελτιωθούν συστήματα προτάσεων (recommender systems).

Η Μέθοδος $k-Means$ στην Διαχείριση Μεγάλου Όγκου Δεδομένων

Έστω για παράδειγμα ότι έχουμε μια πλατφόρμα (π.χ. video streaming ή ηλεκτρονικό κατάστημα) με:

- Εκατομμύρια χρήστες,
- Δισεκατομμύρια events (clicks, views, αγορές, χρόνος παρακολούθησης, κ.λπ.).

Θέλουμε να τμηματοποιήσουμε τους χρήστες (customer segmentation) και να βελτιώσεις τα AI μοντέλα προτάσεων (recommendation / personalization).

Η Μέθοδος $k-Means$ στην Διαχείριση Μεγάλου Όγκου Δεδομένων

Βήμα 1:

Διαχείριση μεγάλου όγκου δεδομένων (Big Data Part)

- Συλλέγουμε log data ανά χρήστη: Αριθμός προβολών ανά κατηγορία, μέσος χρόνος παραμονής,συχνότητα login,αν αγοράζει προσφορές, premium κ.λπ.

Η Μέθοδος $k - Means$ στην Διαχείριση Μεγάλου Όγκου Δεδομένων

Βήμα 1:

Διαχείριση μεγάλου όγκου δεδομένων (Big Data Part)

- Μέσω της συσταδοποίησης χωρίζουμε τους εκατομμύρια χρήστες σε ομάδες όπως
- **Cluster 1:** "heavy users" (πολλές ώρες, premium)
- **Cluster 2:** "casual users"
- **Cluster 3:** "deal hunters" (ευαίσθητοι σε προσφορές)
- κ.λπ.

Η Μέθοδος $k-Means$ στην Διαχείριση Μεγάλου Όγκου Δεδομένων

Άρα το $k-means$ λειτουργεί ως εργαλείο συνοπτικής αναπαράστασης ενός τεράστιου dataset:

Η Μέθοδος $k - Means$ στην Διαχείριση Μεγάλου Όγκου Δεδομένων

Βήμα 2: Σύνδεση με AI / Machine Learning

Τα clusters που προκύπτουν χρησιμοποιούνται μέσα σε μοντέλα AI: Features για supervised μοντέλα

Για κάθε χρήστη προσθέτεις το "cluster id" ως επιπλέον χαρακτηριστικό σε μοντέλα:

- πρόβλεψης churn, (ποιοι πελάτες είναι πιθανό να αποχωρήσουν, να διακόψουν συνδρομή).
- πρόβλεψης πιθανότητας αγοράς,
- recommendation models (π.χ. learning-to-rank).

Η Μέθοδος $k-Means$ στην Διαχείριση Μεγάλου Όγκου Δεδομένων

Το cluster id συμπυκνώνει πληροφορία συμπεριφοράς που είναι δύσκολο να συλλάβει το μοντέλο μόνο του.

- Οι κεντροειδείς του $k-means$ (centroids) χρησιμοποιούνται ως "anchors" για γρήγορη αναζήτηση:

Η Μέθοδος $k-Means$ στην Διαχείριση Μεγάλου Όγκου Δεδομένων

- Πρώτα βρίσκεις το κοντινότερο centroid,μετά ψάχνεις μόνο μέσα στο αντίστοιχο cluster για παρόμοιους χρήστες/προϊόντα.
- Αυτό επιταχύνει σημαντικά την online λειτουργία του recommender, άρα είναι κρίσιμο για εφαρμογές AI σε παραγωγή.

Η Μέθοδος $k - Means$ στην Διαχείριση Μεγάλου Όγκου Δεδομένων

Τι κερδίζεις συνολικά:

- **Από πλευράς big data:**
- ✓ Συμπύκνωση τεράστιου όγκου δεδομένων σε λίγα "νοηματικά" clusters.
- ✓ Δυνατότητα κατανεμημένου υπολογισμού (map – reduce style $k - means$).

Η Μέθοδος $k - Means$ στην Διαχείριση Μεγάλου Όγκου Δεδομένων

- **Από πλευράς AI:**
- ✓ Καλύτερα χαρακτηριστικά εισόδου για supervised μοντέλα.
- ✓ Καλύτερη αρχικοποίηση / δομή στον χώρο των embeddings.
- ✓ Πιο γρήγορες και ποιοτικές προτάσεις σε recommender systems.

Βιβλιογραφία

Δ.Β Παναγιωτάκος, Γ. Κουρλαμπά
Εγχειρίδιο Χειρισμού του Στατιστικού Προγράμματος SPSS

ΣΤ. Δημητριάδης, Στατιστική Επιχειρήσεων με Εφαρμογές σε SPSS και LISREL, Εκδόσεις Κριτική, 2016

Verykios, V., Kagklis, V., & Stavropoulos, E. 2015. *Η επιστήμη των δεδομένων μέσα από τη γλώσσα R* [Undergraduate textbook]. Kallipos, Open Academic Editions. https://hdl.handle.net/11419/2972

Χαλικιάς Μ. Ποσοτική Ανάλυση και Στοιχεία Θεωρίας Αποφάσεων στη Διοίκηση και Οικονομία με Χρήση Λογισμικών EXCEL, ISALOS και SPSS, Εκδόσεις Broken Hill Publishers Ltd, 2021