Γιώργος Νικολαίδης
ge12039
ΣΕΜΦΕ
Cluster Analysis in Machine Learning
1/2/2018

Cluster Analysis

Ο όρος clustering αναφέρεται σε ένα σύνολο τεχνικών για την εύρεση υποομάδων ή συμπλεγμάτων (clusters) σε ένα σύνολο απο δεδομένα . Όταν τοποθετούμε τις παρατηρήσεις ενός συνόλου δεδομένων σε clusters , επιθυμούμε να τα χωρίσουμε σε ξεχωριστά γκρούπ έτσι ώστε οι παρατηρήσεις που βρίσκονται στο ίδιο γκρούπ να είναι παρόμοιες μεταξύ τους , ενώ οι παρατηρήσεις που βρίσκονται σε διαφορετικά γκρούπ να είναι αρκετά διαφορετικές μεταξύ τους . Προκειμένου να γίνει πιο συγκεκριμένο θα πρέπει να ορίσουμε πότε 2 η περισσότερες παρατηρήσεις είναι παρόμοιες και πότε διαφορετικές . Προκείμενου να μπορούμε να ορίσουμε το παραπάνω θα πρέπει κάθε φορά να έχουμε επίγνωση των δεδομένων που μελετάμε . Όταν όλες οι παρατηρήσεις είναι παραγματικοί αριθμοί δηλαδή ποσοτικές μεταβλητές τότε μπορούμε να χρησιμοποιήσουμε ως δείκτη ομοιογένειας των παρατηρήσεων την ευκλείδεια απόσταση κάτι το οποίο όμως δεν μπορούμε να κάνουμε στην περίπτωση που μία απο τις μεταβλητές μας είναι κατηγορική.

Για παράδειγμα τώρα έστω ότι έχουμε ένα σύνολο απο n παρατηρήσεις , με p χαρακτηριστικά η κάθε μία. Οι n παρατηρήσεις θα μπορούσαν να είναι δείγματα ιστών για ασθενείς με καρκινό του μαστού , και τα p χαρακτηριστικά να είναι διάφορες μετρήσεις για κάθε ένα δείγμα ιστού. Τα χαρακτηριστικά αυτά θα μπορούσαν να είναι κλινικές μετρήσεις όπως το στάδιο ή το μέγεθος του καρκίνου ή μετρήσεις γονιδιακής έκφρασης. Είναι πολύ λογικό να θεωρήσουμε ότι ίσως υπάρχει ένας είδος ανομοιογένειας μεταξύ των n δειγμάτων ιστών. Για παράδειγμα ίσως υπάρχουν μερικά άγνωστα είδη καρκίνου του μαστού. Σε αυτή την περίπτωση θα μπορούσαμε επομένως να χρησιμοποιήσουμε cluster ανάλυση προκειμένου να εντοπίσουμε τέτοια πιθανά είδη. Το παραπάνω πρόβλημα ανήκει στην κατηγορία των unsupervised (χωρίς επιτήρηση) προβλημάτων. Μία άλλη εφαρμογή του του cluster analysis θα μπορούσε να είναι στον τομέα του marketing. Ίσως έχουμε πρόσβαση σε ένα μεγάλο σύνολο από μετρήσεις όπως πχ μέσο εισόδημα ανά νοικοκυριό , απασχόληση , απόσταση απο την κοντινότερη αστική περιοχή για ένα μεγάλο σύνολο ανθρώπων. Γνωρίζοντας τώρα αυτές τις πληροφορίες στόχος μας θα ήταν να χωρίσουμε τους ανθρώπους σε διάφορες ομάδες σχετικά με το πόσο επιρεπής είναι σε ένα συγκεκριμένο είδος διαφήμισης ή το πόσο πιθανό είναι να αγοράσουν ένα συγκεκριμένο προϊόν.

Παρατηρούμε επομένως ότι οι εφαρμογές του clustering είναι σε πολλούς τομείς και γι αυτό το λόγο έχουν αναπτυχθεί πολλές μέθοδοι. Οι μέθοδοι αυτοί χωρίζονται σε δύο βαισικές κατηγορίες : **1) Ιεραρχικές (hierarchical) , 2) Μη Ιεραρχικές (nonhierarchical)**

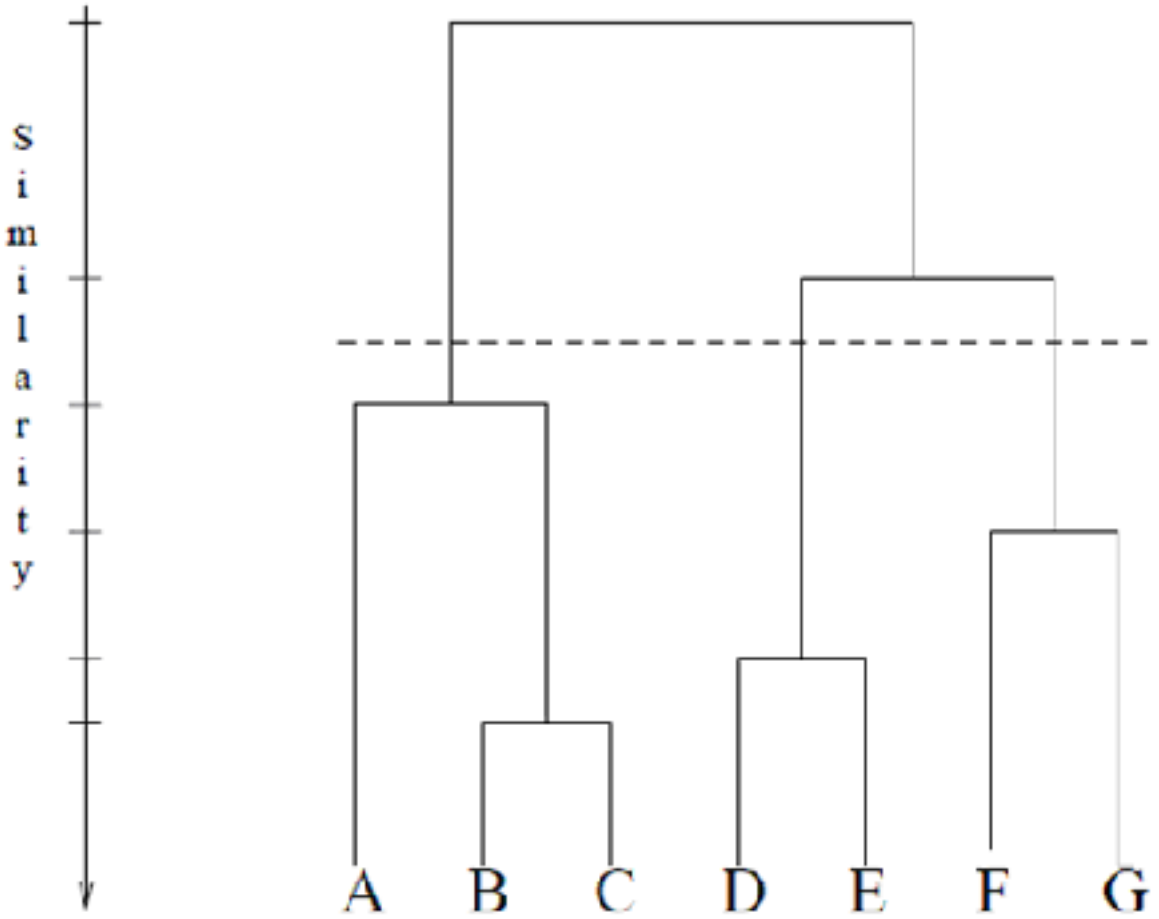
1) Hierarchical clustering

Σε αυτή την κατηγορία ανήκουν μέθοδοι που δημιουργούν μία ιεραρχία συμπλεγμάτων και συνήθως όταν τις εφαρμόζουμε δεν γνωρίζουμε απο πριν σε πόσα clusters θα χωρίσουμε τελικά τις παρατηρήσεις μας. Επιπλέον μπορούμε να χωρίσουμε τις μεθόδους της κατηγορίας αυτής σε 2 επιμέρους κατηγορίες :

i) Στις μεθόδους που ξεκινάνε θεωρώντας ότι κάθε παρατήρηση ανήκει σε ένα cluster που περιέχει μονό τον εαυτό της . Δηλαδή άμα έχουμε n παρατηρήσεις τότε θα ξεκινήσουμε με n clusters τα οποία στη συνέχεια ενώνονται.

ii) Στις μεθόδους που ξεκινάνε θεωρώντας ότι κάθε παρατήρηση ανήκει στο ίδιο cluster και έτσι αρχικά θα έχουμε ένα cluster το οποίο στη συνέχεια θα διαιρείται σε κομμάτια.

Τα αποτελέσματα των παραπάνω μεθόδων συνήθως παρουσιάζονται μέσω ενώς δεντροδιαγράμματος όπως αυτό της παρακάτω εικόνας. Στον οριζόντιο άξονα έχουμε τις παρατηρήσεις και στον κατακόρυφο την ποσότητα ως προς την οποία μετράμε την ομοιογένεια των παρατηρήσεών μας. Στη συνέχεια θέτουμε ένα όριο ως προς το οποίο θεωρούμε ότι απο εκεί και πάνω 2 παρατηρήσεις δεν θεωρούνται ομοιογενείς και φέρνουμε μία παράλληλη γραμμή όπως αυτή του σχήματος καθορίζοντας με αυτό τον τρόπο σε πόσα clusters θα χωριστούν τελικά οι παρατηρήσεις μας καθώς και με ποιο τρόπο. Για παράδειγμα στην παρακάτω εικόνα βλέπουμε ότι θα χωριστούν σε 3 clusters .

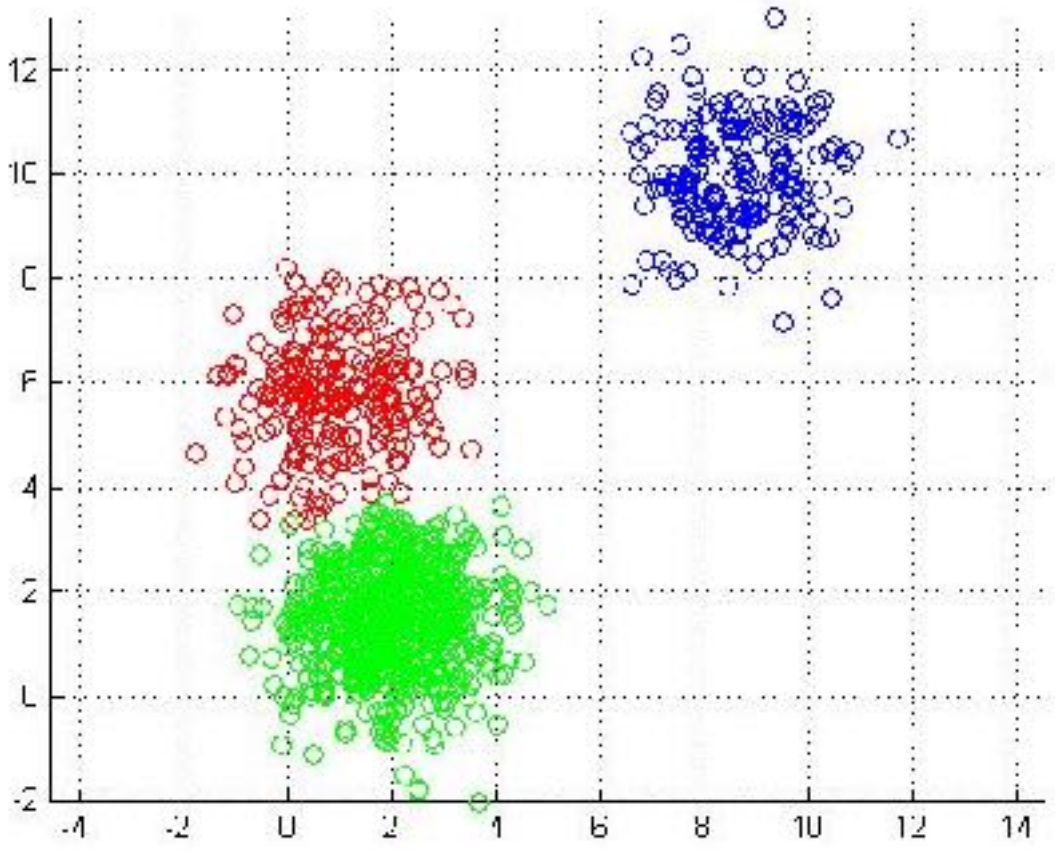


2) Nonhierarchical clustering

Στην κατηγορία αυτή ανήκουν μέθοδοι που χωρίζουν τα δεδομένα σε k κομμάτια όπου κάθε ένα από αυτά αποτελεί ένα cluster . Σε αντίθεση με την προηγούμενη κατηγορία εδώ γνωρίζουμε εκ των προτέρων τον αριθμό των clusters κάτι το οποίο όμως δεν είναι πάντα εφικτό και γι αυτό τον λόγο έχουν αναπτυχθεί διάφοροι μέθοδοι για την αντιμετώπιση αυτού του προβλήματος όπως θα δούμε και στην συνέχεια . Ο χωρισμός των παρατηρήσεων γίνεται ελαχιστοποιώντας ή μεγιστοποιώντας συνήθως κάποια τιμή. Σε αυτή την κατηγορία ανήκει και ο αλγόριθμος K-means clustering με τον οποίο θα ασχοληθούμε στη συνέχεια .

K-Means Clustering

Η μέθοδος αυτή αποτελεί μία απλή και κομψή προσέγγιση στο πώς μπορούμε να χωρίσουμε τα δεδομένα μας σε K διακριτά , μη επικαλυπτόμενα συμπλέγματα. Βασική παραδοχή του είναι ότι κάθε παρατήρηση αντιστοιχεί σε ένα σημείο του p-διάστατου χώρου όπου p ο αριθμός των μεταβλητών για κάθε παρατήρηση.Για να εκτελέσουμε την μέθοδο αυτή θα πρέπει πρώτα να προσδιορίσουμε τον επιθυμητό αριθμό συμπλεγμάτων K. Στην συνέχεια ο αλγόριθμος θα τοποθετήσει κάθε παρατήρηση σε ακριβώς ένα απο τα K συμπλέγματα . Για παράδειγμα άν οι παρατηρήσεις μας έχουν μόνο 2 μεταβλητές η κάθε μία και για K=3 θα μπορούσε να προκύψει κάτι τέτοιο



Κάθε χρώμα αποτελεί και ένα διαφορετικό σύμπλεγμα . Όταν οι παρατηρήσεις μας όμως έχουν παραπάνω απο 2 μεταβλητές δεν θα μπορούμε να τις παρουσιάσουμε εύκολα με γραφικό τρόπο καθώς τα σχήματα που θα προκύψουν θα είναι σε διαστάσεις μεγαλύτερες του 2 .

Ο αλγόριθμος βασίζεται σε ένα απλό μαθηματικό πρόβλημα. Ξεκινάμε ορίζοντας μερικούς συμβολισμούς . Έστω $C_1, C_2, C_3, \ldots, C\kappa$ σύνολα που περιέχουν τους δείκτες των παρατηρήσεων σε κάθε σύμπλεγμα. Αυτά τα σύνολα θα ικανοποιούν 2 ιδιότητες :

1. $C_1 \bigcup C_2 \bigcup C_3 \bigcup \ldots . \bigcup C_K = [1, \ldots . . . n]$. Με άλλα λόγια κάθε παρατήρηση ανήκει σε τουλάχιστον ένα απο τα K συμπλέγματα .

2. $C_k \bigcap Ck' = \emptyset$ για κάθε $k \neq k'$. Με άλλα λόγια τα συμπλέγματα δεν επικαλύπτονται δηλαδή καμοία παρατήρηση δεν ανήκει σε παραπάνω απο ένα σύμπλεγμα.

Για παράδειγμα αν η ι-οστή παρατήρηση είναι στο K-οστό σύμπλεγμα τότε $i \in C_K$. Η κεντρική ιδέα πίσω απο τον αλγόριθμο K-means clustering είναι ότι ένας χωρισμός σε συμπλέγματα θα θεωρείται καλός αν η μεταβολή , απόκλιση των παρατηρήσεων σε κάθε σύμπλεγμα είναι η μικρότερη δυνατή. Η τιμή αυτή θα συμβολίζεται $W(C_K)$ όπου K ο αντίστοιχος δείκτης για κάθε σύμπλεγμα. Επομένως έχουμε να λύσουμε το πρόβλημα ελαχιστοποίησης της ποσότητας $\sum_{k=1}^{K} W($. (1)

Με άλλα λόγια θέλουμε να χωρίσουμε τις παρατηρήσεις μας σε K συμπλέγματα έτσι ώστε η συνολική μεταβλητότητα που προκύπτει απο το άθροισμα των επιμέρους τιμών της εντός κάθε συμπλέγματος , να είναι η μικρότερη δυνατή. Προκειμένου όμως να μπορέσουμε να λύσουμε το παραπάνω πρόβλημα θα πρέπει αρχικά να ορίσουμε την ποσότητα $W(C_k)$ για κάθε σύμπλεγμα. Υπάρχουν αρκετοί τρόποι να την ορίσουμε όμως η πιο συνηθισμένη επιλογή περιλαμβάνει την Ευκλείδεια απόσταση :

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \quad (2)$$

όπου $|C_k|$ είναι ο αριθμός των παρατηρήσεων στο k-οστό σύμπλεγμα . Με άλλα λόγια η μεταβολή, απόκλιση για το k-οστό σύμπλεγμα είναι το άθροισμα όλων των ανά 2 Ευκλείδειων αποστάσεων μεταξύ των παρατηρήσεων στο k-οστό σύμπλεγμα δια το συνολικό αριθμό παρατηρήσεων στο k-οστό σύμπλεγμα. Επομένως κάνοντας αντικατάσταση την σχέση 2 στην σχέση 1 προκύπτει το άθροισμα που θέλουμε να ελαχιστοποιήσουμε . Στην συνέχεια χρειαζόμαστε έναν αλγόριθμο για να χωρίσει τις παρατηρήσεις με τέτοιο τρόπο σε k συμπλέγματα ώστε το παραπάνω άθροισμα να ελαχιστοποιηθεί. Αυτό στην πραγματικότητα αποτελεί ένα αρκετά δύσκολο πρόβλημα για να λυθεί με ακρίβεια καθώς υπάρχουν σχεδόν K^n τρόποι για να χωρίσεις n παρατηρήσεις σε K συμπλέγματα. Το παραπάνω είναι ένας αρκετά μεγάλος αριθμός εκτός αν το K και το n είναι αρκετά μικρά. Για το σκοπό αυτό θα εξετάσουμε τον παρακάτω αλγόριθμο ο οποίος παρέχει μια αρκετά καλή λύση στο παραπάνω πρόβλημα βελτιστοποίησης.

Αλγόριθμος K-Means Clustering

1) Αρχικά βάζουμε τυχαία έναν αριθμό απο το 1 έως το K σε κάθε μία απο τις παρατηρήσεις . Ο αριθμός αυτός θα αποτελεί το αρχικό σύμπλεγμα στο οποίο θα βρίσκεται η παρατήρηση .
2) Κάνουμε τις παρακάτω διαδικασίες μέχρι να μην έχουμε κάποια αλλαγή σε οποιοδήποτε σύμπλεγμα:
   (a) Για κάθε ένα απο τα K συμπλέγματα υπολογίζουμε το γεωμετρικό του κέντρο . Για το k-οστό σύμπλεγμα το γεωμετρικό κέντρο αποτελεί στην ουσία τον αριθμητικό μέσο των p χαρακτηριστικών των παρατηρήσεων που βρίσκονται στο σύμπλεγμα αυτό.
   (b) Τοποθετούμε κάθε παρατηρήση στο σύμπλεγμα του οποίου η απόσταση απο το γεωμετρικό του κέντρο είναι η μικρότερη. (Για την έννοια της απόστασης χρησιμοποιούμε την Ευκλείδεια απόσταση που ορίσαμε νωρίτερα).

Ο παραπάνω αλγόριθμος ελαχιστοποιεί το άθροισμα που προκύπτει απο τον συνδιασμό των σχέσεων (1) και (2) που ορίσαμε νωρίτερα . Για να καταλάβουμε γιατι πάμε να δούμε τα παρακάτω:

$$\frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2 = 2\sum_{i\in C_k}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{kj})^2 \quad (3)$$

όπου $\bar{x}_{kj}=\frac{1}{|C_k|}\sum_{i\in C_k}x_{ij}$ είναι ο μέσος για το χαρακτηριστικό j στο σύμπλεγμα C_k.

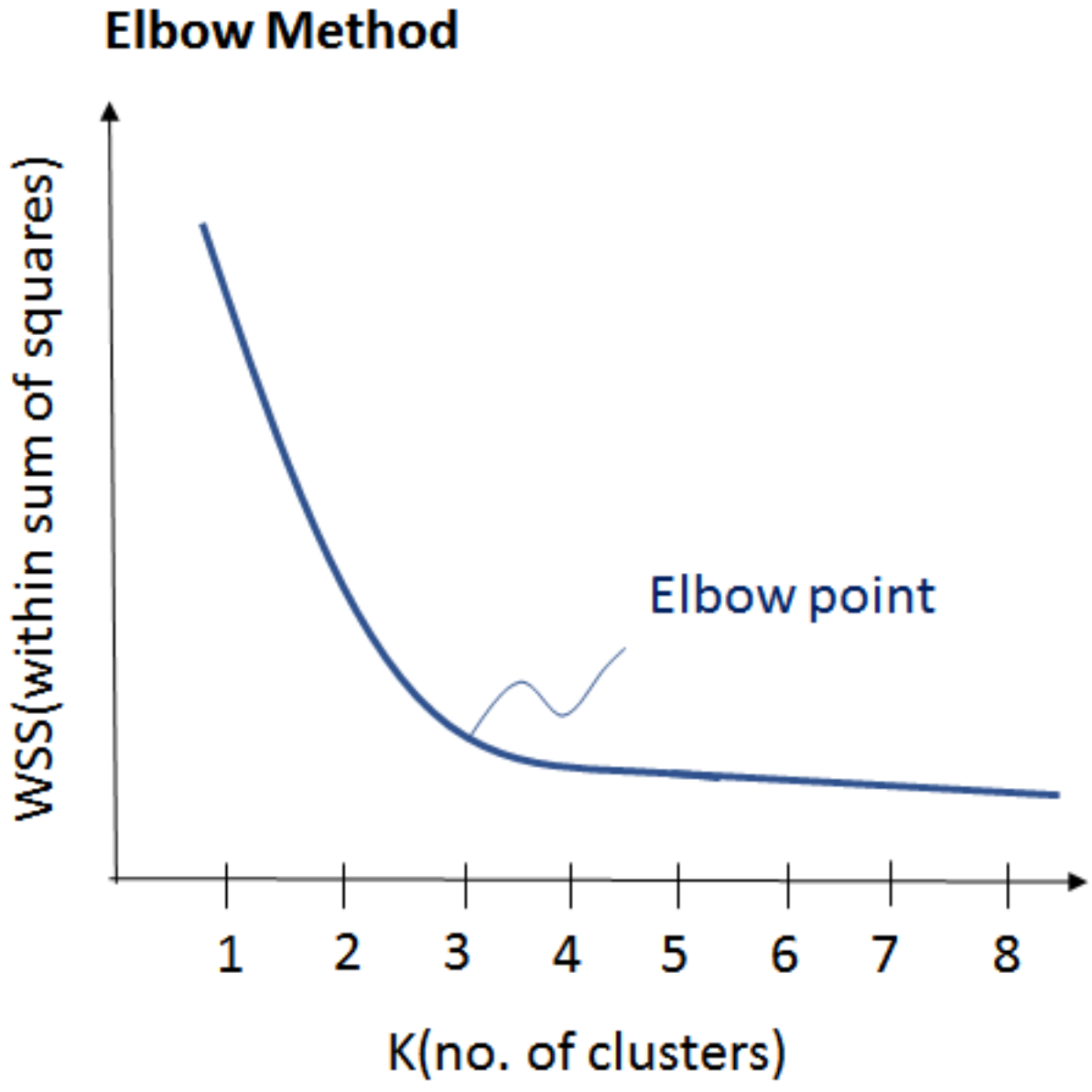
Στο βήμα 2(a) , οι μέσοι του συμπλέγματος για κάθε χαρακτηριστικό είναι οι σταθερές που ελαχιστοποιούν το άθροισμα των τετραγωνικών αποκλίσεων και στο βήμα 2(b) εναποθέτωντας τις παρατηρήσεις μπορεί μόνο να βελτιώσει την ποσότητα (3). Αυτό σημαίνει ότι καθώς ο αλγόριθμος τρέχει , η ομαδοποίηση που θα προκύπτει θα βελτιώνεται συνεχώς μέχρι να μην μπορεί να αλλάξει το αποτέλεσμα. Όταν συμβεί αυτό θα έχουμε φτάσει σε ένα τοπικό βέλτιστο. Επειδή το αποτέλεσμα του αλγορίθμου θα είναι ένα τοπικό ελάχιστο και όχι ολικό , τα αποτελέσματα που θα προκύψουν θα εξαρτώνται απο την αρχική τυχαία ομαδοποίηση της κάθε παρατήρησης στο βήμα 1 του αλγορίθμου. Γι αυτό το λόγο είναι σημαντικό να τρέξουμε τον αλγόριθμο πολλές φορές με διαφορετικές αρχικές ομαδοποιήσεις. Στη συνέχεια επιλέγουμε την καλύτερη λύση δηλαδή εκείνη που μας δίνει τη μικρότερη τιμή του αθροίσματος που προκύπτει από τις σχέσεις (1) και (2).

Όπως έχουμε δεί για να εκτελέσουμε τον παραπάνω αλγόριθμο θα πρέπει να αποφασίσουμε πόσα συμπλέγματα θεωρούμε ότι υπάρχουν στα δεδομένα μας. Το πρόβλημα της επιλογής του αριθμού των συμπλεγμάτων αποτελεί απο μόνο του ένα αρκετά δύσκολο πρόβλημα . Υπάρχουν αρκετοί τρόποι για να δώσουμε μία εκτίμηση για το K. Θα αναφερθούμε σε 3 απο αυτούς : **1) Elbow method** , **2)Average silhouette method** , **3)Gap statistic method**

1) Elbow Method

Η βασική ιδέα του K-means clustering είναι ο χωρισμός των παρατηρήσεων σε συμπλέγματα έτσι ώστε η συνολική εσω-συμπλεγματική απόκλιση (η αλλίως WSS) να είναι η μικρότερη δυνατή . Η τιμή αυτή μας λέει πόσο κατάλληλος είναι ο διαχωρισμός που έχουμε κάνει και θέλουμε να είναι μικρή.
Η Elbow Method κοιτάει την τιμή WSS σαν μία συνάρτηση του αριθμού των συμπλεγμάτων K. Επομένως η τιμή που θα πάρει τελικά το K θα είναι αυτή που αν προσθέσουμε άλλο ένα σύμπλεγμα στο K η τιμή του WSS θα βελτιωθεί ελάχιστα.
Ο κατάλληλος αριθμός συμπλεγμάτων μπορεί να οριστεί ως εξής:

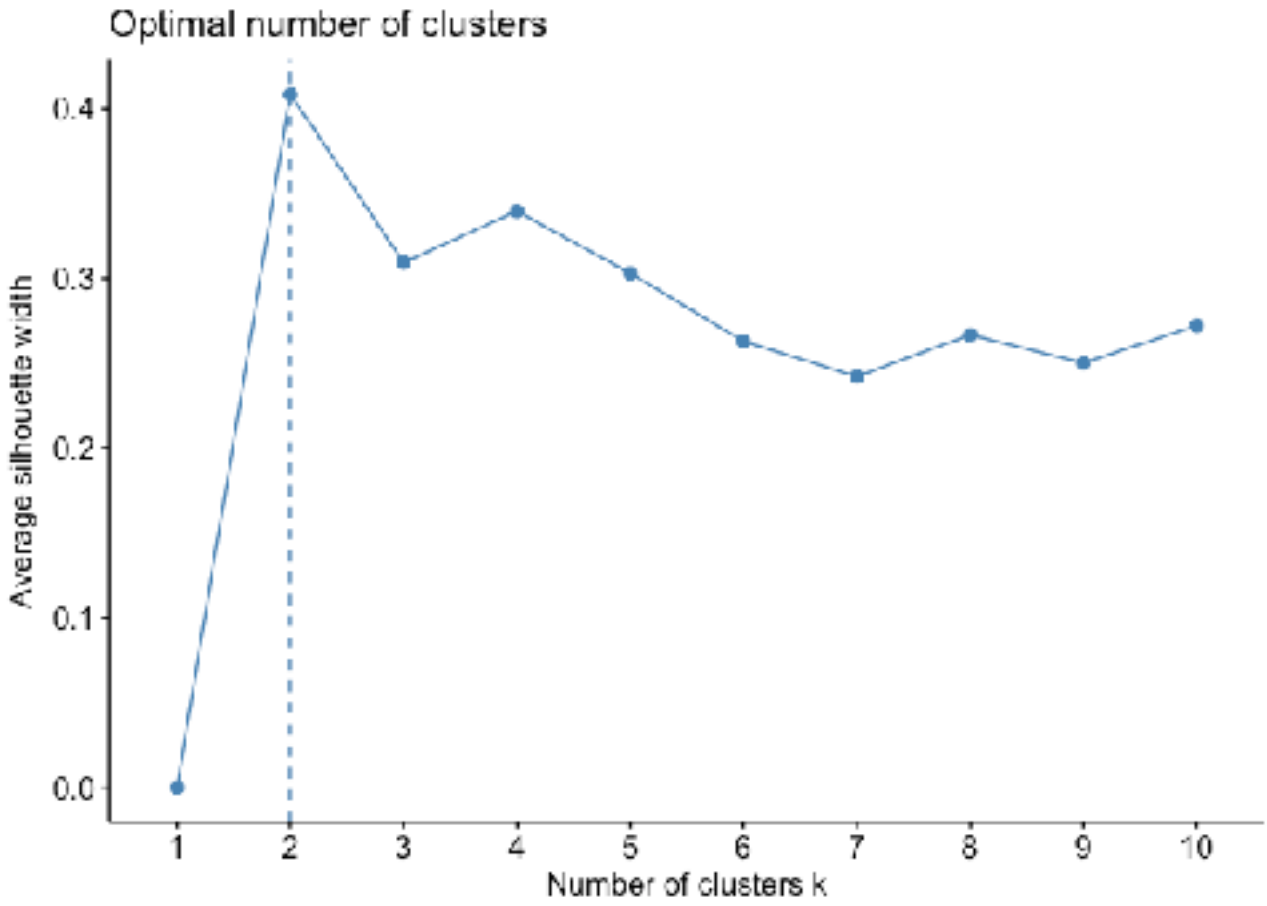
1) Τρέχουμε τον αλγόριθμο για διαφορετικές τιμές του k. Για παράδειγμα για τιμές απο το 1 έως το 10.
2) Για κάθε k υπολογίζουμε την τιμή WSS (απόσταση κάθε παρατήρησης απο το αντίστοιχο κέντρο του συμπλέγματος που ανήκει)
3) Κάνουμε την γραφική παράσταση WSS-k
4) Το σημείο στην καμπύλη που κάνει απότομη στροφή και μοιάζει με αγκώνα συνήθως θεωρείται σαν ένδειξη υποψήφιου αριθμού k για τον αριθμό των συμπλεγμάτων.



2) Average silhouette method

Η μέθοδος αυτή μετράει το πόσο καλή ομαδοποίηση των παρατηρήσεων έχει προκύψει. Αυτό γίνεται υπολογίζοντας την μέση silhouette η οποία όσο πιο μεγάλη είναι τόσο πιο καλή ομαδοποίηση των παρατηρήσεων έχουμε. Η μέθοδος αυτή υπολογίζει την παραπάνω τιμή για διάδορες τιμές του k . Η κατάλληλη τιμή για το k είναι αυτή που μεγιστοποιεί την μέση silhouette τιμή. Πληροφορίες για το πώς την υπολογίζουμε θα δούμε στην συνέχεια.
Ο αλγόριθμος είναι παρόμοιος με αυτόν της προηγούμενης μεθόδου (elbow method)

1) Τρέχουμε τον αλγόριθμο (k means clustering) για διάφορες τιμές του k πχ από το 1 μέχρι το 10
2) Για κάθε τιμή του k υπολογίζουμε την μέση silhouette των παρατηρήσεων (avg.sil)
3) Κάνουμε την γραφική παράσταση avg.sil με τον αριθμό των συμπλεγμάτων k
4) Το σημείο στο οποίο εμφανίζεται μέγιστο θεωρείται ως ο ιδανικός αριθμός συμπλεγμάτων k



3) Gap Statistic Method

Η Gap Statistic μέθοδος συγκρίνει την συνολική εσω-συμπλεγματική απόκλιση για διαφορετικές τιμές του k με τις αναμενόμενες τιμές τους υπό μηδενική κατανομή αναφοράς των δεδομένων. Η ιδανική τιμή για τον αριθμό των συμπλεγμάτων θα είναι αυτή που μεγιστοποιεί την στατιστική συνάρτηση gap. Αυτό σημαίνει ότι η δομή της ομαδοποίησης διαφέρει αρκετά απο την τυχαία ομοιόμορφη κατανομή των σημείων.
Ο αλγόριθμος λειτουργεί ως εξής:

1) Τοποθετούμε τις παρατηρήσεις σε συμπλέγματα , για διάφορες τιμές του k = 1,….,k_{max} και υπολογίζουμε την αντίστοιχη συνολική εσω-συμπλεγματική απόκλιση W_k
2) Παράγουμε Β σύνολα δεδομένων με μία τυχαία ομοιόμορφη κατανομή. Ομαδοποιούμε τα δεδομένα αυτά σε συμπλέγματα για κάθε μία απο τιμές του k = 1,….,k_{max} και υπολογίζουμε την αντίστοιχη συνολική εσω-συμπλεγματική απόκλιση W_{kb}.
3) Υπολογίζουμε την εκτιμημένη τιμή της στατιστικής συνάρτησης gap ως την απόκλιση της τιμής W_k απο την αναμενόμενη τιμή της W_{kb} υπό την μηδενική υπόθεση : $Gap(k) = \frac{1}{B}\sum_{b=1}^{B} log(W_{kb}) - log(W_k)$. Υπολογίζουμε επίσης την τυπική απόκλιση της συνάρτησης.
4) Επιλέγουμε την μικρότερη τιμή k για την οποία ισχύει $Gap(k) \geq Gap(k+1) - S_{k+1}$ όπου S_{k+1} η τυπική απόκλιση της συνάρτησης για k+1

Αξιολόγηση της λύσης

Κάθε φορά που γίνεται χρήση του παραπάνω αλγορίθμου σε ένα σύνολο δεδομένων θα προκύψει κάποια ομαδοποίηση. Επομένως αποτελεί βασική απαίτηση να μπορούμε να ελέγξουμε αν η ομαδοποίηση αυτή αναπαριστά αληθινές υποομάδες στα δεδομένα ή προκύπτουν απο την ομαδοποίηση του θορύβου που μπορεί να υπάρχει στα δεδομένα.
Η αξιολόγηση των αποτελεσμάτων αποτελεί δύσκολη διαδικασία και μερικές φορές τόσο δύσκολη όσο η ίδια η ομαδοποίηση. Μερικές προσεγγίσεις είναι :
α) **Εσωτερική (internal)** αξιολόγηση όπου η ομαδοποίηση συνοψίζεται σε μία μόνο ποιοτική ποσότητα
β) **Εξωτερική (external)** αξιολόγηση όπου η ομαδοποίηση που προέκυψε συγκρίνεται με μία υπάρχουσα ομαδοποίηση που ξέρουμε ότι ισχύει
γ) **Χειροκίνητη (manual)** αξιολόγηση απο κάποιον ειδικό
δ) **Έμμεση (indirect)** αξιολόγηση , αξιολογώντας την χρησιμότητα της ομαδοποίησης στην προοριζόμενη εφαρμογή της .

Εσωτερική αξιολόγηση

Όταν το αποτέλεσμα μίας ομαδοποίησης αξιολογείται με βάση τα δεδομένα με τα οποία έγινε η ομαδοποίηση , τότε πρόκειται για εσωτερική αξιολόγηση. Οι μέθοδοι που ανήκουν στην κατηγορία αυτή συνήθως θεωρούν ως καλύτερο έναν αλγόριθμο που παράγει συμπλέγματα με μεγάλη ομοιογένεια στο εσωτερικό ενός συμπλέγματος και χαμηλή ομοιογένεια μεταξύ των συμπλεγμάτων. Το παραπάνω έχει ως αποτέλεσμα τέτοιου είδους μέθοδοι αξιολόγησης να είναι πιο ευνοϊκές απέναντι σε μεθόδους που χρησιμοποιούν το ίδιο μοντέλο ομαδοποίησης. Για παράδειγμα ο αλγόριθμος k-means βελτιστοποιεί την απόσταση μεταξύ αντικειμένων συνεπώς κάποιο κριτήριο εσωτερικής αξιολόγησης που βασίζεται στην απόσταση είναι λογικό να αξιολογήσει το αποτέλεσμα ενός τέτοιου αλγορίθμου ως πολύ καλό. Επομένως η χρήση της εσωτερικής αξιολόγησης θα ταίριαζε καλύτερα στο να αποκτήσουμε μία εικόνα για καταστάσεις όπου ένας αλγόριθμος έχει καλύτερη επίδοση απο έναν άλλο κάτι το οποίο όμως δεν σημαίνει απαραίτητα ότι παράγει πιο έγκυρα αποτελέσματα.

Οι παρακάτω μέθοδοι μπορούν να χρησιμοποιηθούν για την αξιολόγηση των αποτελεσμάτων διαφόρων αλγορίθμων ομαδοποίησης βασιζόμενοι σε εσωτερικά κριτήρια:

- **Δείκτης Davies-Bouldin**

Ο δείκτης Davies-Bouldin μπορεί να υπολογιστεί μέσω της επόμενης φόρμουλας:

$$DB = \frac{1}{n}\sum_{i=1}^{n} max_{j\neq i}(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)})$$

όπου n ο αριθμός των συμπλεγμάτων , c_x είναι το κέντρο του συμπλέγματος x , σ_x είναι ο μέσος όρος των αποστάσεων όλων των στοιχείων στο σύμπλεγμα x απο το κέντρο c_x και $d(c_i, c_j)$ είναι η απόσταση μεταξύ των κέντρων c_i και c_j . Οι αλγόριθμοι επομένως που παράγουν ομαδοποιήσεις όπου οι αποστάσεις στο εσωτερικό κάθε συμπλέγματος είναι μικρές ενώ οι αποστάσεις μεταξύ των συμπλεγμάτων μεγάλες (πχ k-means) θα έχουν μικρή τιμή στο δείκτη Davies-Bouldin. Όσο πιο μικρή η τιμή του παραπάνω δείκτη τόσο πιο καλά είναι τα αποτελέσματα του αλγορίθμου με βάση το παραπάνω κριτήριο.

- **Δείκτης Dunn**

Ο δείκτης Dunn στοχεύει στο να αναγνωρίσει πυκνά και καλά διαχωρισμένα συμπλέγματα. Ορίζεται σαν τον λόγο ανάμεσα στην ελάχιστη απόσταση μεταξύ συμπλεγμάτων και της μέγιστης απόστασης στο εσωτερικό ενός συμπλέγματος. Ο παραπάνω δείκτης υπολογίζεται ως εξής:

$$D = \frac{min_{1\leq i<j\leq n} d(i,j)}{max_{1\leq k\leq n} d'(k)}$$

όπου d(i,j) είναι απόσταση μεταξύ των συμπλεγμάτων i και j, $d'(k)$ είναι η απόσταση στο εσωτερικό του συμπλέγματος k. Ο τρόπος με τον οποίο υπολογίζουμε τις παραπάνω αποστάσεις δεν είναι απαραίτητα συγκεκριμένος . Για παράδειγμα για την απόσταση d(i,j) θα μπορούσαμε να χρησιμοποιήσουμε την απόσταση των κέντρων των δύο συμπλεγμάτων ενώ για την απόσταση $d'(k)$ να χρησιμοποιήσουμε την μέγιστη απόσταση μεταξύ οποιουδήποτε ζεύγους παρατηρήσεων στο σύμπλεγμα k. Τέλος θεωρούμε ότι όσο μεγαλύτερος είναι ο δείκτης αυτός τόσο καλύτερη θεωρούμε ότι είναι η ομαδοποίηση που προέκυψε.

- **Συντελεστής Silhouette**

Το παραπάνω συντελεστή είδαμε ότι μπορούμε να τον χρησιμοποιήσουμε και για να επιλέξουμε την τιμή k για τον αριθμό των συμπλεγμάτων που θα χρησιμοποιήσουμε.
Για κάθε παρατήρηση i ο συντελεστής αυτός υπολογίζεται ως εξής:

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))}$$

όπου $a(i)$ η μέση απόσταση της i παρατήρησης απο όλες τις παρατηρήσεις στο ίδιο σύμπλεγμα. Όσο πιο μικρή είναι η τιμή $a(i)$ τόσο πιο καλά θεωρούμε ότι έχει τοποθετηθεί η παρατήρηση i. Στη συνέχεια ορίζουμε την μέση ανομοιομορφία ενός σημείου i με ένα σύμπλεγμα c ως την μέση τιμή της απόστασης του i απ όλες τις παρατηρήσεις στο σύμπλεγμα c. Υπολογίζουμε την παραπάνω τιμή για κάθε σύμπλεγμα στο οποίο δεν ανήκει η παρατήρηση i και ορίζουμε την μικρότερη απο αυτές ως $b(i)$. Το σύμπλεγμα απο το οποίο προέκυψε η τιμή $b(i)$ ονομάζεται γειτονικό του i. Επιπλέον ο συντελεστής $s(i)$ παίρνει τιμές απο το -1 στο 1 και όσο πιο κοντά στο 1 είναι τόσο πιο καλά τοποθετημένη είναι η παρατήρηση i. Υπολογίζοντας την τιμή αυτή επομένως για κάθε παρατήρηση και παίρνοντας την μέση τιμή τους ανάλογα με το πόσο κοντά στο 1 βρίσκεται μπορούμε να κρίνουμε αν η ομαδοποίηση που προέκυψε είναι καλή η όχι.

Εξωτερική Αξιολόγηση

Στην εξωτερική αξιολόγηση, τα αποτελέσματα της ομαδοποίησης αξιολογούνται με βάση δεδομένα τα οποία δεν χρησιμοποιήθηκαν για την ομαδοποίηση αλλά γνωρίζουμε σε ποιά ομάδα ανήκουν. Τα παραπάνω δεδομένα αποτελούν συνήθως σύνολα δεδομένων τα οποία έχουν τοποθετηθεί σε ομάδες και έχουν δημιουργηθεί απο ειδικούς. Μπορούμε επομένως να ελέγξουμε πόσο κοντά είναι η ομαδοποίηση που προέκυψε με την προκαθορισμένη ομαδοποίηση αυτών των παρατηρήσεων. Διάφορα μέτρα ποιότητας για την αξιολόγηση των αποτελεσμάτων ενός αλγορίθμου , που βασίζονται στην παραπάνω ιδέα είναι τα εξής:

- **Purity**

Η συνάρτηση purity είναι ένα μέτρο του βαθμού στον οποίο συμπλέγματα περιέχουν μία μόνο κατηγορία.Ο υπολογισμός του μπορεί να θεωρηθεί ως εξής: Για κάθε σύμπλεγμα,

μετρήστε τον αριθμό των σημείων-δεδομένων από την πιο κοινή κατηγορία στην εν λόγω ομάδα. Τώρα πάρτε το άθροισμα πάνω από όλα τα συμπλέγματα και διαιρέστε με το συνολικό αριθμό των σημείων-δεδομένων.
Δεδομένου ενός συνόλου M από συμπλέγματα και ενός συνόλου D από κατηγορίες όπου και τα δύο ομαδοποιούν N σημεία-δεδομένα ,το παραπάνω μέτρο μπορεί να οριστεί ως εξής:

$$\frac{1}{N}\sum_{m\in M} max_{d\in D}|m\cap d|$$

Αυτό το μέτρο δεν επιφέρει ποινή στην κατοχή πολλών συμπλεγμάτων. Συγκεκριμένα θα μπορούσε να πάρει την τιμή 1 αν βάζαμε κάθε παρατηρήση σε διαφορετικό σύμπλεγμα.

- **Δείκτης Rand**

Ο δείκτης Rand υπολογίζει πόσο όμοια είναι η ομαδοποίηση που προέκυψε απο τον αλγόριθμο με αυτή που μας έχει δοθεί. Ο παραπάνω δείκτης θα μπορούσε να θεωρηθεί και ως το ποσοστό των ορθών αποφάσεων που έχουν παρθεί απο τον αλγόριθμο. Υπολογίζεται ως εξής:

$$RI = \frac{TP + TN}{TP + FP + FN + TN}$$

όπου TP είναι ο αριθμός των αληθινά θετικών , TN ο αριθμός των αληθινά αρνητικών , FP είναι ο αριθμός των ψευδώς θετικών και FN ο αριθμός των ψευδώς αρνητικών. Ένα θέμα του παραπάνω δείκτη είναι ότι τα FP και FN έχουν το ίδιο βάρος. Το παραπάνω μπορεί να αποτελέσει ένα μη επιθυμητό χαρακτηριστικό σε μερικές εφαρμογές ομαδοποίησης. Ο παρακάτω δείκτης λαμβάνει υπόψιν του το προηγούμενο πρόβλημα.

- **Δείκτης F**

Ο δείκτης F μπορεί να χρησιμοποιηθεί για να εξισορροπηθεί η συμβολή των ψευδών αρνητικών με τη στάθμιση της ανάκλησης (**recall**) μέσω μιας παραμέτρου $\beta \geq 0$.
Ας ορίσουμε την ακρίβεια (**precision**) και την ανάκληση (και τα δύο εξωτερικά μέτρα αξιολόγησης απο μόνα τους) ως εξής:

$$P = \frac{TP}{TP + FP} \qquad R = \frac{TP}{TP + FN}$$

όπου P η ακρίβεια και R η ανάκληση. Στην συνέχεια υπολογίζουμε την τιμή F ως εξής:

$$F_\beta = \frac{(\beta^2 + 1)\cdot P \cdot R}{\beta^2 \cdot P + R}$$

Παρατηρούμε ότι όταν το β = 0 , $F_0 = P$. Με άλλα λόγια , η ανάκληση (recall) δεν έχει καμία επίδραση στο μέτρο F όταν β = 0 και αυξάνοντας το β αυξάνεται και το βάρος της ανάκλησης στη τελική τιμή του F. Επίσης παρατηρούμε ότι το TN δεν λαμβάνεται υπόψη και μπορεί να ποικίλει απο το 0 προς τα επάνω χωρίς περιορισμό.

- **Δείκτης Jaccard**

Ο δείκτης Jaccard χρησιμοποιείται για να ποσοτικοποιήσει την ομοιότητα μεταξύ δύο συνόλων δεδομένων.Ο δείκτης Jaccard παίρνει μια τιμή μεταξύ 0 και 1. Ένας δείκτης 1 σημαίνει ότι τα δύο σύνολα δεδομένων είναι πανομοιότυπα και ένας δείκτης 0 δηλώνει ότι τα σύνολα δεδομένων δεν έχουν κοινά στοιχεία. Ο δείκτης Jaccard ορίζεται ως εξής:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{TP}{TP + FP + FN}$$

Αυτός είναι ο αριθμός των μοναδικών στοιχείων που είναι κοινά και στα δύο σύνολα διαιρούμενο με το συνολικό αριθμό των μοναδικών στοιχείων και στα δύο σύνολα.Το TN δεν λαμβάνεται υπόψη και μπορεί να ποικίλει από 0 προς τα επάνω χωρίς περιορισμό.

- **Δείκτης Dice**

Το συμμετρικό μέτρο Dice διπλασιάζει το βάρος για το TP ενώ αγνοεί το TN και είναι ισοδύναμο με το F_1 - τον δείκτη F με β = 1.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{2TP}{2TP + FP + FN}$$

- **Δείκτης Fowlkes-Mallows**

Ο δείκτης Fowlkes-Mallows υπολογίζει την ομοιότητα μεταξύ των συμπλεγμάτων που επιστρέφονται από τον αλγόριθμο ομαδοποίησης και της προκαθορισμένης ομαδοποίησης. Όσο μεγαλύτερη είναι η τιμή του δείκτη Fowlkes-Mallows, τόσο πιο παρόμοια είναι τα παραπάνω. Μπορεί να υπολογιστεί χρησιμοποιώντας τον ακόλουθο τύπο:

$$FM = \sqrt{\frac{TP}{TP + FP} \cdot \frac{TP}{TP + FN}}$$

Ο δείκτης FM είναι ο γεωμετρικός μέσος όρος της ακρίβειας και της ανάκλησης
P και R και είναι επίσης γνωστός ως το μέτρο G, ενώ το μέτρο F είναι ο αρμονικός μέσος όρος τους.

- **Confusion matrix**

Στον τομέα του machine learning και ειδικότερα του προβλήματος της στατιστικής ταξινόμησης, ένας confusion matrix, επίσης γνωστός ως error matrix, είναι μια ειδική διάταξη πίνακα που επιτρέπει την απεικόνιση της απόδοσης ενός αλγορίθμου.Κάθε σειρά του πίνακα αντιπροσωπεύει τις περιπτώσεις σε μια προβλεπόμενη κλάση ενώ κάθε στήλη αντιπροσωπεύει τις περιπτώσεις σε μια πραγματική τάξη (ή αντίστροφα).Το όνομα πηγάζει από το γεγονός ότι καθιστά εύκολο να δούμε αν το σύστημα προκαλεί σύγχυση σε δύο κατηγορίες. Η μορφή του παραπάνω πίνακα φαίνεται απο το παρακάτω παράδειγμα:

		Truth				
	Asphalt	Concrete	Grass	Tree	Building	Total
Asphalt	2386	4	0	1	4	2394
Concrete	0	332	0	0	1	333
Grass	0	1	908	8	0	917
Tree	0	0	0	1084	9	1093
Building	12	0	0	6	2053	2071
Total	2397	337	908	1099	2067	8008

Predicted

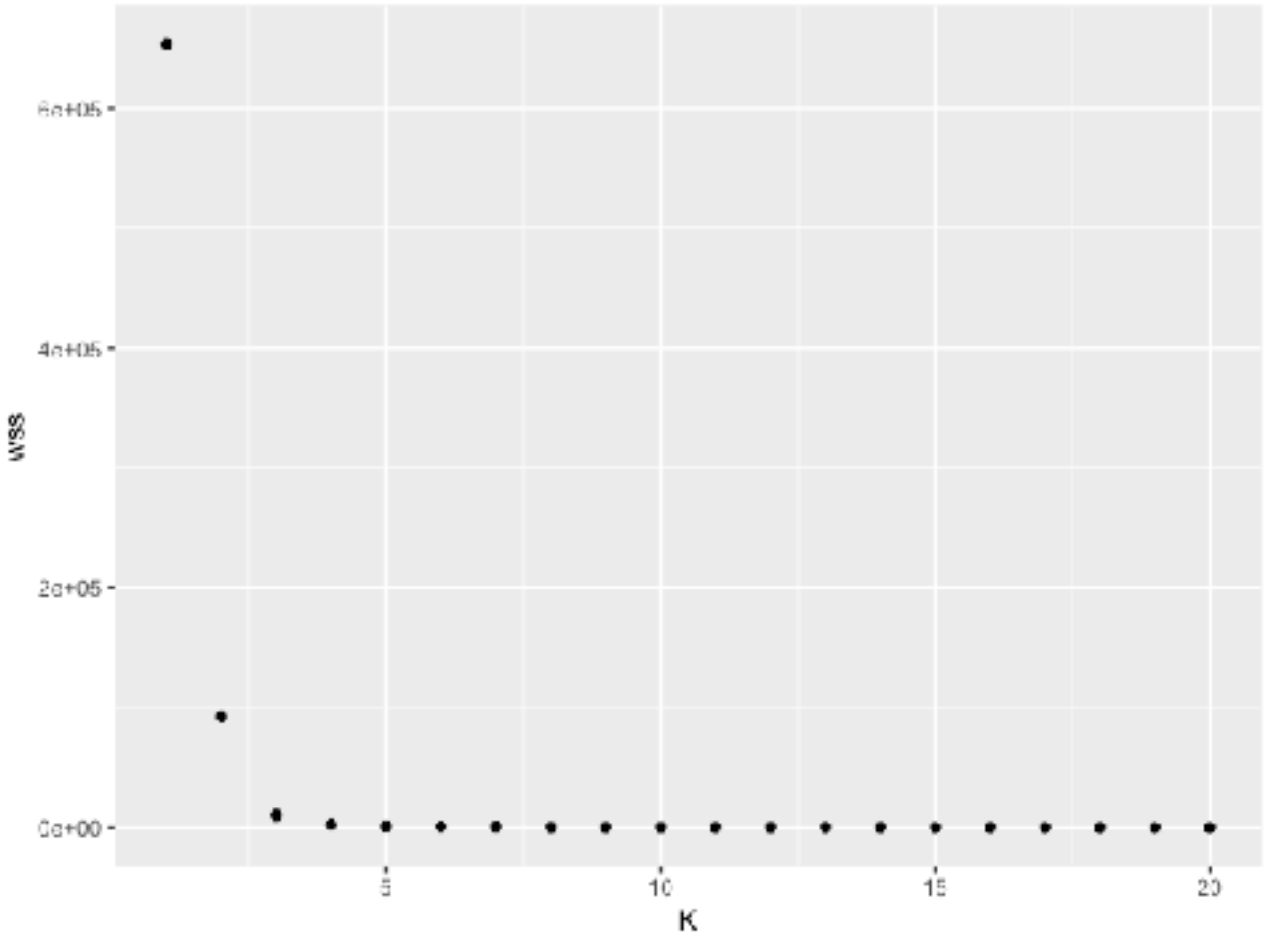
Στην παραπάνω εικόνα οι γραμμές αναφέρονται στην ομαδοποίηση του αλγορίθμου και οι στήλες στην προκαθορισμένη ομαδοποίηση των δεδομένων. Αν ο πίνακας που προκύπτει είναι σχεδόν διαγώνιος (η μπορεί να γίνει με μετάθεση στηλών)όπως αυτός του παραδείγματος σημαίνει ότι ο αλγόριθμος έχει δώσει σχετικά καλά αποτελέσματα.
Στη συνέχεια θα περάσουμε στην εφαρμογή του αλγορίθμου k-means σε ένα παράδειγμα προκειμένου να εφαρμόσουμε τα όσα αναφέραμε παραπάνω

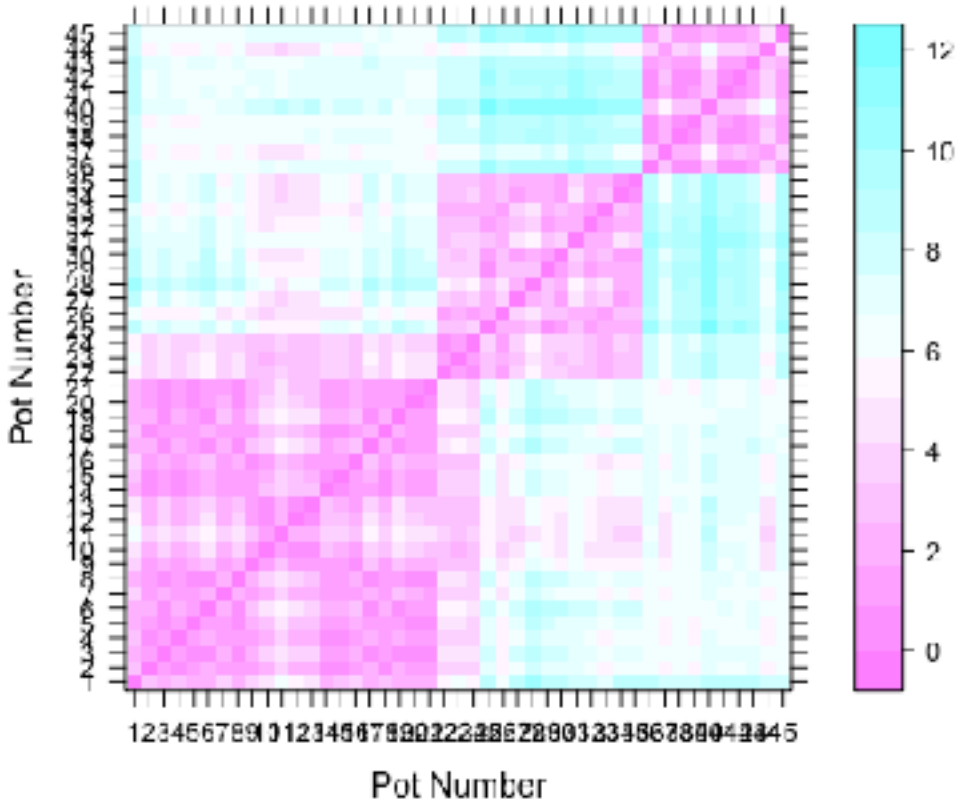
ΕΦΑΡΜΟΓΗ 1

Για την ανάλυση των δεδομένων θα κάνουμε χρήση της γλώσσας R.
Τα δεδομένα που θα χρησιμοποιήσουμε αρχικά δίνουν τη χημική σύνθεση 48 δειγμάτων ρωμαϊκής-βρετανικής κεραμικής, που προσδιορίζονται με φασματοφωτομετρία ατομικής απορρόφησης, για εννέα οξείδια (Tubb et al., 1980). Για τα δεδομένα αυτά, μας ενδιαφέρει για το κατά πόσο, με βάση τις χημικές τους συνθέσεις, οι γλάστρες μπορούν να χωριστούν σε ξεχωριστές ομάδες. Ενδεικτικά μερικά απο τα δεδομένα μας θα είναι της μορφής:

	Al2O3	Fe2O3	MgO	CaO	Na2O	K2O	TiO2	MnO	BaO
1	18.8	9.52	2.00	0.79	0.40	3.20	1.01	0.077	0.015
2	16.9	7.33	1.65	0.84	0.40	3.05	0.99	0.067	0.018
3	18.2	7.64	1.82	0.77	0.40	3.07	0.98	0.087	0.014
4	16.9	7.29	1.56	0.76	0.40	3.05	1.00	0.063	0.019
5	17.8	7.24	1.83	0.92	0.43	3.12	0.93	0.061	0.019
6	18.8	7.45	2.06	0.87	0.25	3.26	0.98	0.072	0.017

Αρχικά θα πρέπει να προσδιορίσουμε τον αριθμό k για τον αριθμό των συμπλεγμάτων που θα χρησιμοποιήσουμε. Κάνοντας χρήση του elbow method και απο το παρακάτω διάγραμμα έχουμε:


wss
6e+05
4e+05
2e+05
0e+00
5
10
15
20
K


Pot Number
Pot Number
12
10
8
6
4
2
0

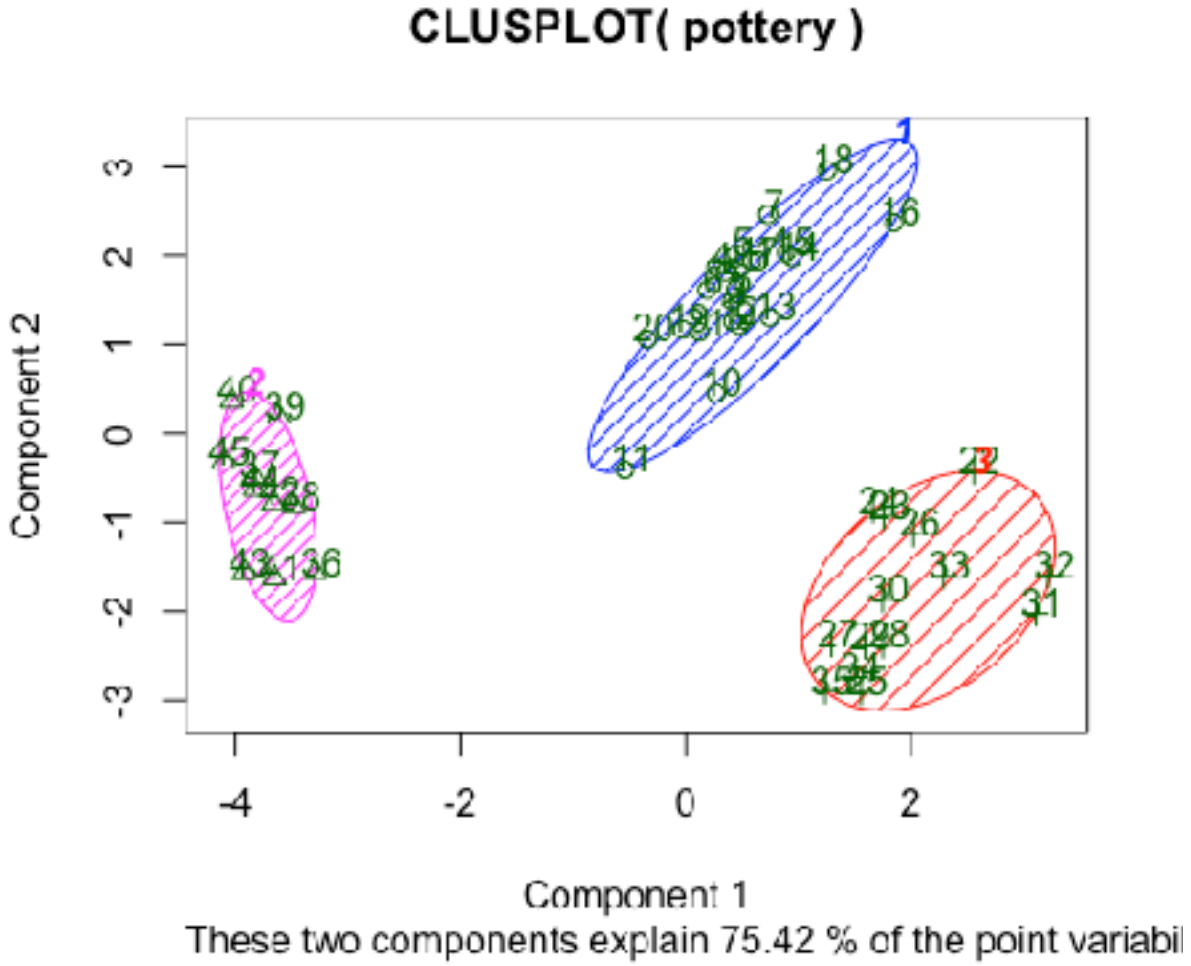
Παρατηρούμε ότι για K>3 η K>4 δεν έχουμε μεγάλη αλλαγή στο WSS επομένως θα επιλέξουμε μία απο αυτές τις τιμές. Για την επιλογή της κατάλληλης τιμής μπορούμε να κατασκευάσουμε και τον πίνακα ανομοιογένειας ο οποίος έχει τις ευκλείδειες αποστάσεις μεταξύ των παρατηρήσεων μας.

Απο το παραπάνω βλέπουμε ότι έχουμε 3 συμπλέγματα (όσο πιο ρόζ είναι το χρώμα τόσο μικρότερη είναι η απόσταση). Επομένως απο τα δύο παραπάνω επιλέγουμε K=3
Εφόσον έχουμε επιλέξει το K μένει να τρέξουμε τον αλγόριθμο K means.
Η ομαδοποίηση που προέκυψε μας δίνει τις εξής πληροφορίες:

Cluster means:

	Al2O3	Fe2O3	MgO	CaO	Na2O	K2O	TiO2	MnO	BaO
1	17.750	1.612	0.640	0.039	0.051	2.021	1.020	0.003	0.016
2	12.436	6.208	4.778	0.214	0.226	4.188	0.683	0.118	0.016
3	16.919	7.429	1.842	0.939	0.346	3.103	0.938	0.071	0.017

Επειδή οι παρατήσεις μας έχουν 9 χαρακτηριστικά δεν είναι εύκολο να δούμε τα συμπλέγματα στο επίπεδο . Παρ΄ όλα αυτά μπορούμε να επιλέξουμε 2 χαρακτηριστικά που επηρεάζουν περισσότερο την ομαδοποίηση των δεδομένων μας και να κάνουμε την γραφική παράσταση που ακολουθεί:



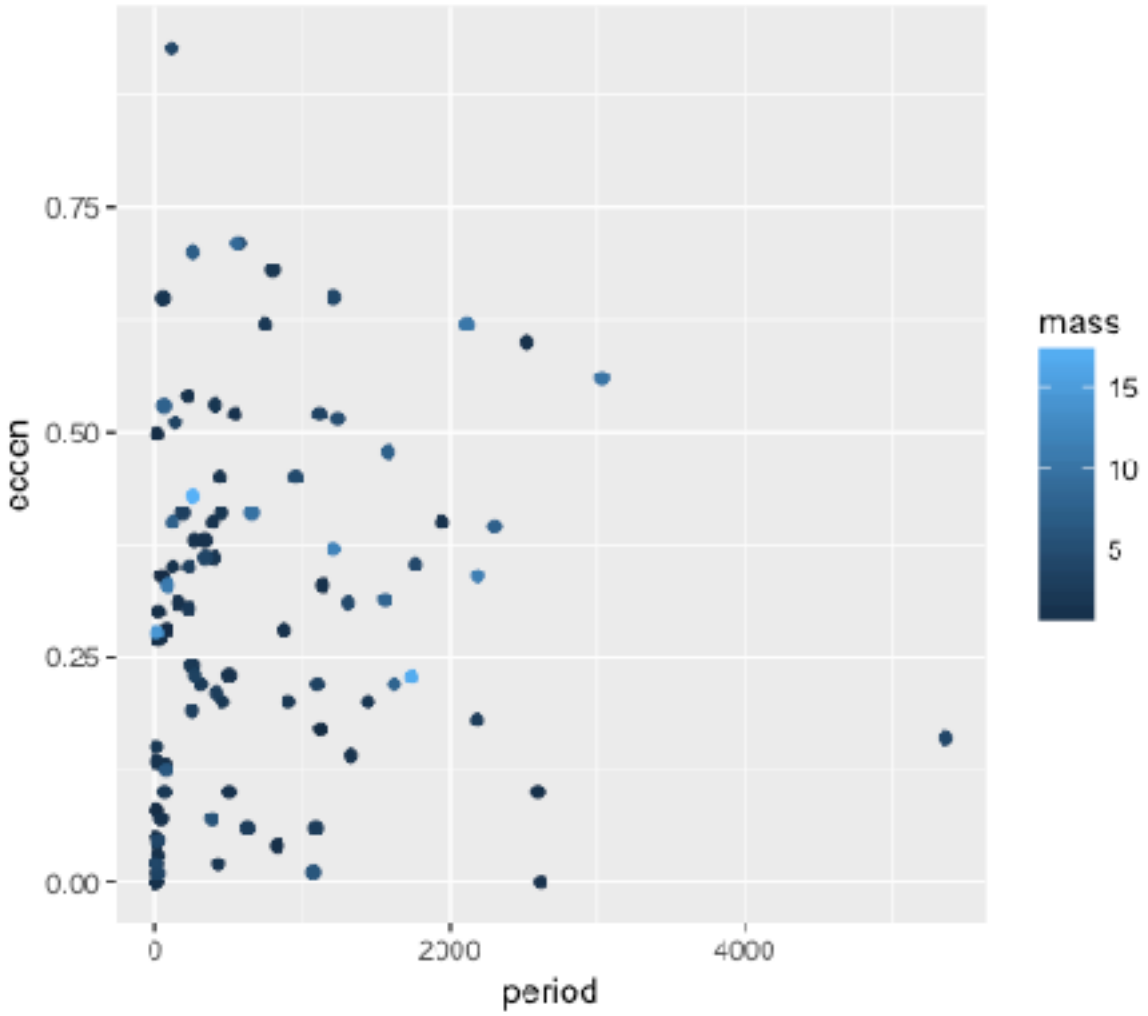
Τα δύο χαρακτηριστικά που επιλέχθηκαν εκφράζουν κατά 74.42% την μεταβλητότητα των παρατηρήσεων μας δηλαδή δεν εκφράζουν πλήρως τα δεδομένα μας . Παρ ‘ όλα αυτά για τα συγκεκριμένα χαρακτηριστικά είναι εμφανής η ομαδοποίηση των παρατηρήσεων που προέκυψε απο τον αλγόριθμο K means στην οποία τα 3 σύνολα-συμπλέγματα δεν τέμνονται και έχουμε σαφή διαχωρισμό των παρατηρήσεων .Η παραπάνω γραφική παρουσίαση των συμπλεγμάτων γίνεται ακόμα πιο δύσκολη και ανακριβής σε μεγαλύτερες διαστάσεις καθώς κάθε χαρακτηριστικό επηρεάζει σε μικρότερο βαθμό την μεταβλητότητα των παρατηρήσεων με αποτέλεσμα να μην μπορούμε να επιλέξουμε μόνο 2 χαρακτηριστικά. Επιπλέον παρατηρώντας τους μέσους για κάθε σύμπλεγμα βλέπουμε ότι δεν διαφέρουν ιδιαίτερα ως προς το χαρακτηριστικό BaO .

ΕΦΑΡΜΟΓΗ 2

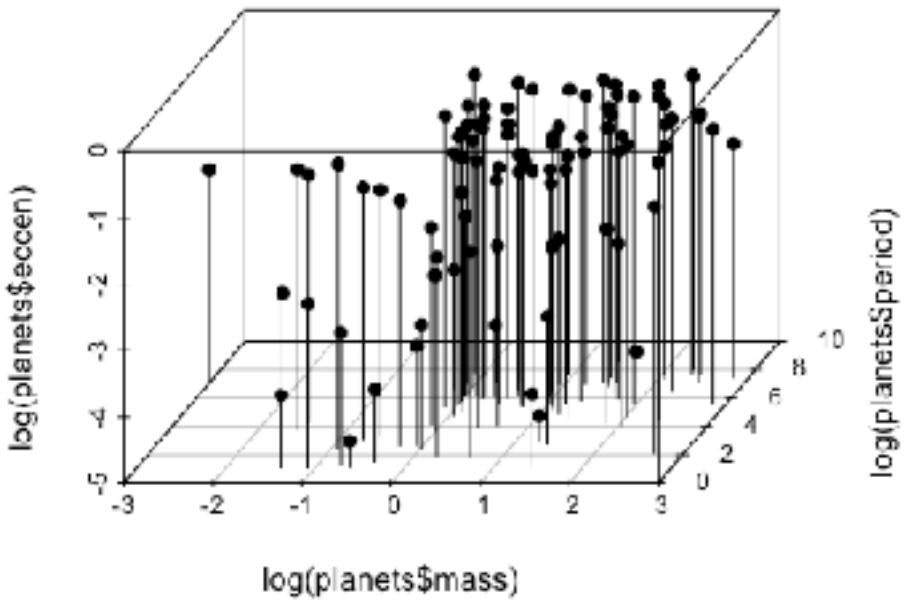
Οι εξωπλανήτες είναι πλανήτες εκτός του Ηλιακού Συστήματος.Ο πρώτος τέτοιος πλανήτης ανακαλύφθηκε το 1995 από τον Mayor και τον Queloz (1995). Ο πλανήτης, παρόμοιος στη μάζα με τον Δία, βρέθηκε σε τροχιά γύρω από ένα σχετικά συνηθισμένο αστέρι, 51 Πήγασος. Στην ενδιάμεση περίοδο έχουν ανακαλυφθεί πάνω από εκατό εξωπλανήτες, σχεδόν όλοι ανιχνεύονται έμμεσα, χρησιμοποιώντας τη βαρυτική επιρροή που ασκούν στα αντίστοιχα κεντρικά αστέρια. Ένας συναρπαστικός απολογισμός των εξωπλανητών και η ανακάλυψή τους δίνονται στους Mayor and Frei (2003).
Από τις ιδιότητες των εξωπλανητών που βρέθηκαν μέχρι τώρα φαίνεται ότι η θεωρία της πλανητικής ανάπτυξης που κατασκευάστηκε για τους πλανήτες του Ηλιακού Συστήματος ίσως χρειαστεί να αναδιατυπωθεί. Οι εξωπλανήτες δεν είναι καθόλου όπως οι εννέα τοπικοί πλανήτες που γνωρίζουμε τόσο καλά. Ένα πρώτο βήμα στη διαδικασία κατανόησης των εξωπλανητών μπορεί να είναι να προσπαθήσουμε να τους ταξινομήσουμε σε σχέση με τις γνωστές ιδιότητές τους. Για το σκοπό αυτό έχουμε τα παρακάτω δεδομένα (εμφανίζονται μερικά απο αυτά) :

	mass	period	eccen
1	0.120	4.950	0.00
2	0.197	3.971	0.00
3	0.210	44.28	0.34
4	0.220	75.80	0.28
5	0.230	6.403	0.08
6	0.250	3.024	0.02

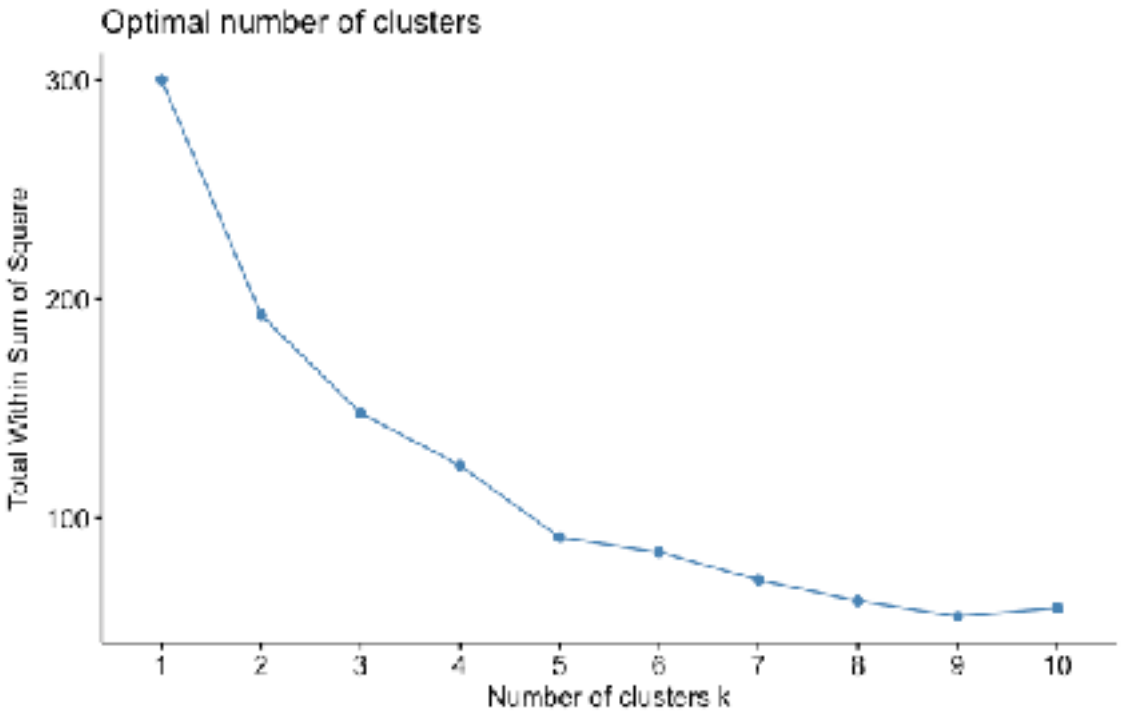
Στα δεδομένα έχουμε την μάζα (σε μάζα του Δία) την περίοδο (στις ημέρες της γης) και την εκκεντρότητα (εξάντληση) των εξωπλανητών που ανακαλύφθηκαν μέχρι τον Οκτώβριο του 2002. Αρχικά θα εξετετάσουμε τα δεδομένα μας μέσω γραφικών παραστάσεων προκειμένου να εξετάσουμε αν υπάρχει κάποιος προφανής διαχωρισμός τους σε ομάδες.



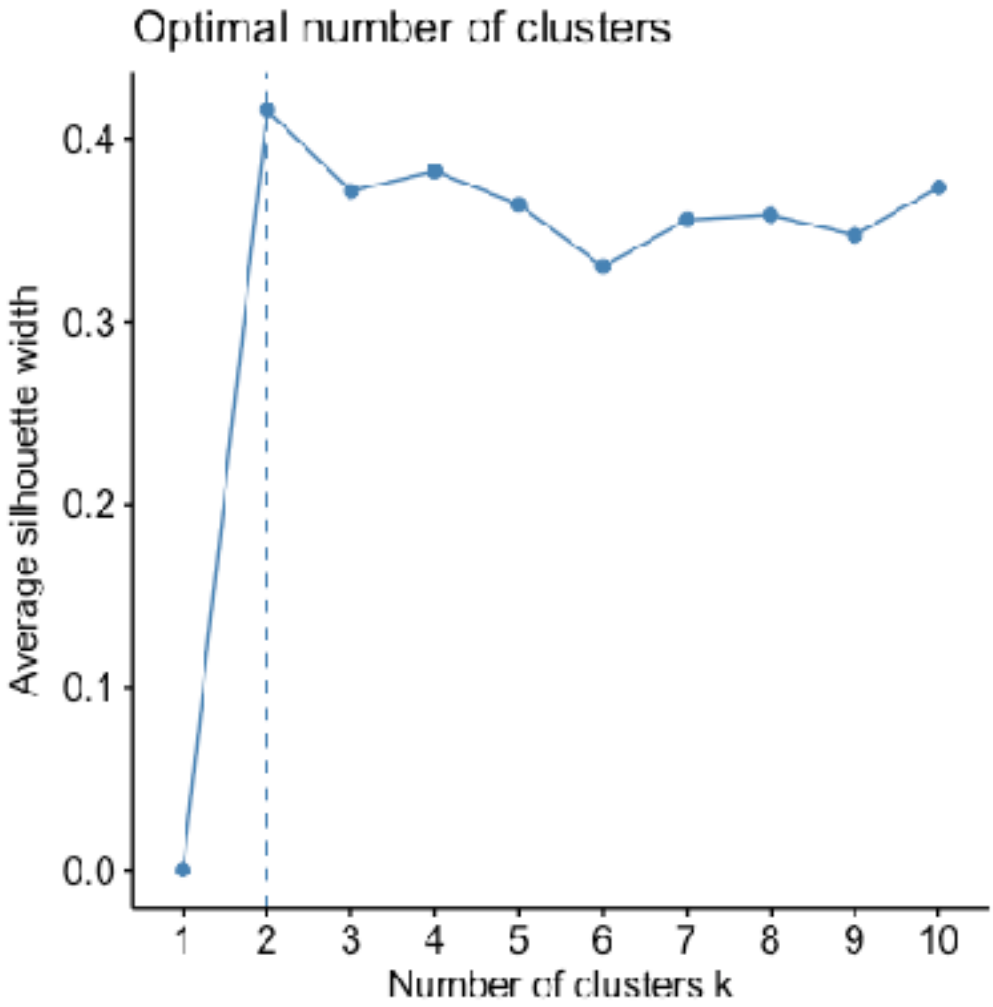
Απο το παραπάνω διάγραμμα παρατηρούμε ότι υπάρχει ένα σύνολο απο εξωπλανήτες με μικρή περίοδο . Επίσης παρατηρούμε ότι καθώς αυξάνεται η περίοδος, η εκκεντρότητα έχει μία τάση προς μείωση. Παρ' όλα αυτά δεν μπορούμε να διακρίνουμε κάποιο εμφανή διαχωρισμό των παρατηρήσεων σε συμπλέγματα.



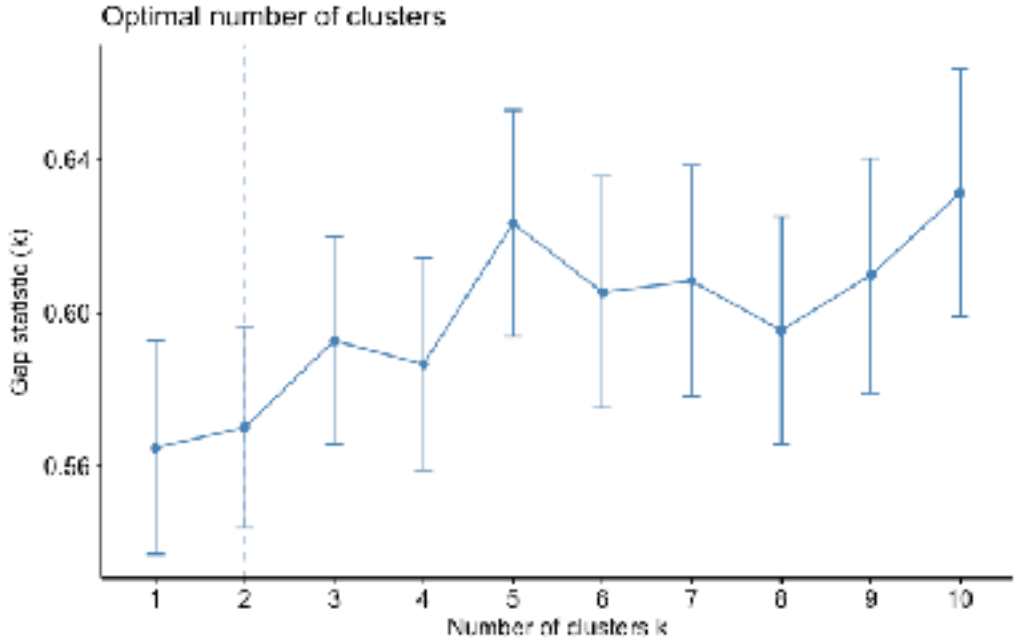
Το παραπάνω διάγραμμα περιέχει και τα 3 μεγέθη σε λογαριθμική κλίμακα . Ούτε απο αυτό το διάγραμμα μπορούμε να εντοπίσουμε κάποια εμφανή ομαδοποίηση των παρατηρήσεων. Επομένως στη συνέχεια πρέπει να καθορίσουμε τον αριθμό K όπως και στο προηγούμενο παράδειγμα. Επειδή τα δεδομένα μας είναι σε διαφορετικές κλίμακες γεγονός που μπορεί να επηρεάσει τον αλγόριθμο K means καθώς βασίζεται στην έννοια της απόστασης , θα πρέπει πρώτα να τυποποιηθούν. Στη συνέχεια κάνοντας χρήση του elbow method και απο το παρακάτω διάγραμμα έχουμε:



Απο το παραπάνω διάγραμμα βλέπουμε ότι δεν είναι ξεκάθαρο το σημείο το οποίο πρέπει να επιλέξουμε καθώς υπάρχουν πολλά υποψήφια σημεία πχ K=5 , K=7. Γι αυτό το λόγο θα χρησιμοποιήσουμε την average silhouette method προκειμένου να αποσαφηνίσουμε τα αποτελέσματα.



Παρατηρούμε ότι η συγκεκριμένη μέθοδος μας δίνει ότι ο ιδανικός αριθμός K για τα συμπλέγματα που πρέπει να χρησιμοποιήσουμε είναι K=2 καθώς στο σημείο αυτό έχουμε μεγιστοποίηση της αντίστοιχης συνάρτησης. Τέλος μπορούμε να κάνουμε χρήση και της Gap statistic μεθόδου απο την οποία προκύπτει το εξής διάγραμμα:

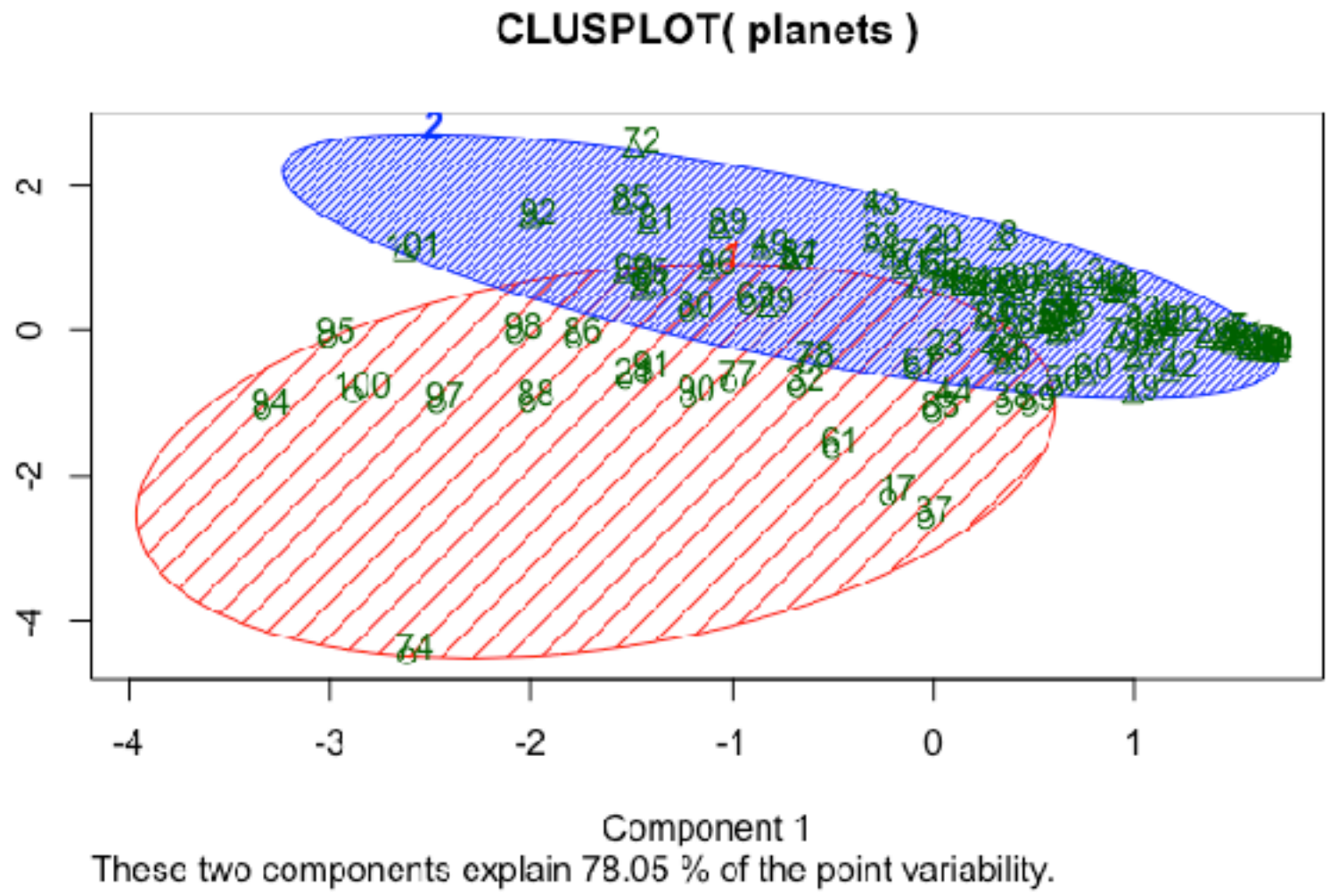


Παρατηρούμε ότι και αυτή η μέθοδος μας δίνει ως ιδανική τιμή το K=2. Λαμβάνοντας υπόψην επομένως και τις 3 παραπάνω μεθόδους θα τρέξουμε τον αλγόριθμο K means για K=2 και θα προκύψουν τα εξής αποτελέσματα:

cluster means

	mass	period	eccen
1	5.125	1832.356	0.313
2	2.671	241.163	0.270

Απο τους μέσους διακρίνουμε ότι η ομάδα 1 περιέχει εξωπλανήτες που έχουν 5 φορές το μέγεθος του Δία και αρκετά μεγάλη περίοδο . Η ομάδα 2 περιλαμβάνει εξωπλανήτες με μικρότερη μάζα και μικρότερη περίοδο απο αυτή στην ομάδα 1 ενώ οι εκεντρότητες είναι σχετικά κοντά . Περαιτέρω εξήγηση των αποτελεσμάτων απαιτεί σαφώς μια λεπτομερή γνώση της αστρονομίας. Επιπλέον στην ομάδα Α θα έχουμε 27 παρατηρήσεις ενώ στην ομάδα Β 74 την κατανομή των οποίων μπορούμε να δούμε και απο το παρακάτω γράφημα:



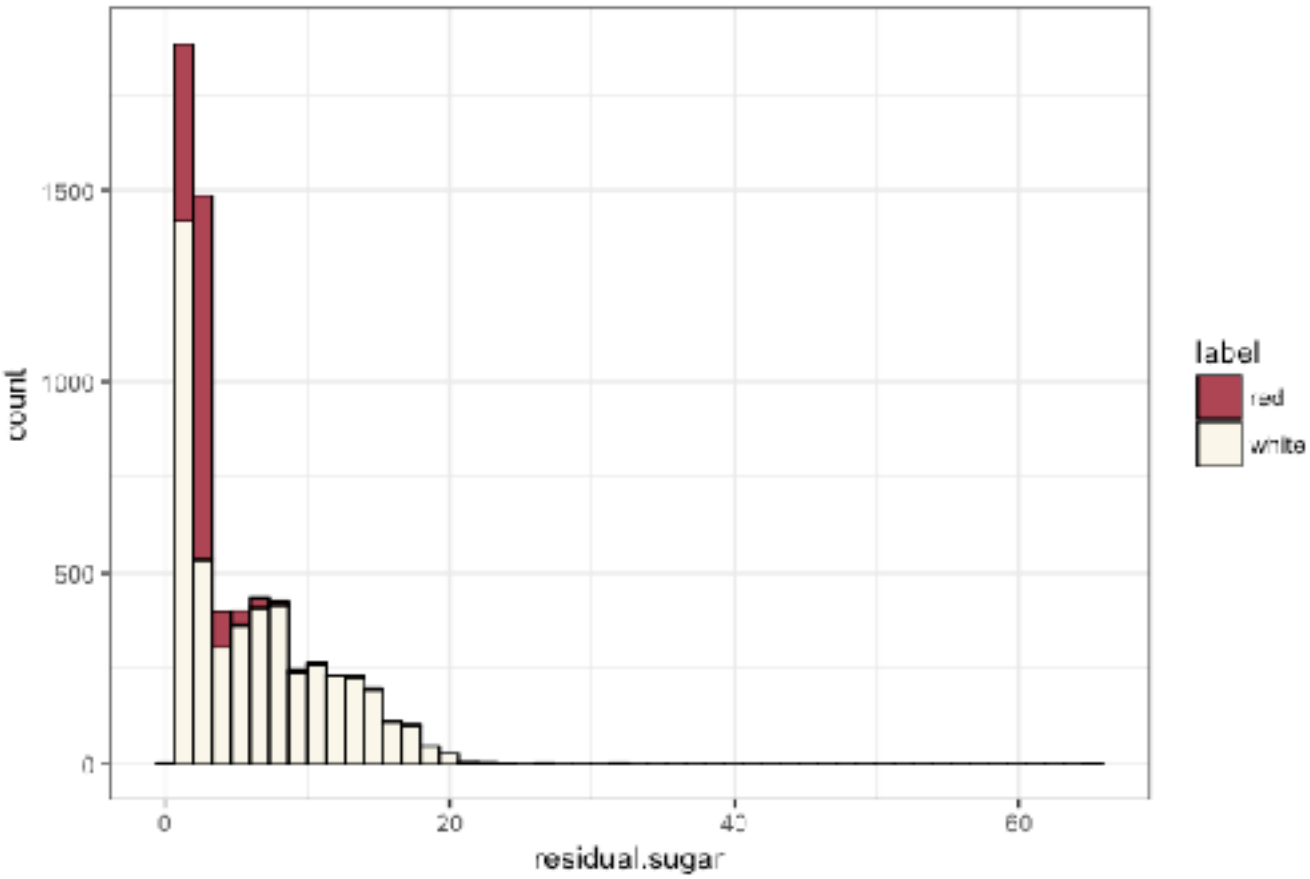
Η γραφική έγινε με βάση 2 χαρακτηριστικά όπως και πριν τα όποια όπως βλέπουμε επηρεάζουν κατά 78.05% την μεταβλητότητα των παρατηρήσεων. Σε αντίθεση με την προήγουμενη εφαρμογή βλέπουμε ότι τα 2 σύνολα τέμνονται και ο διαχωρισμός των παρατηρήσεων δεν είναι τόσο ξεκάθαρος όπως πριν.

ΕΦΑΡΜΟΓΗ 3

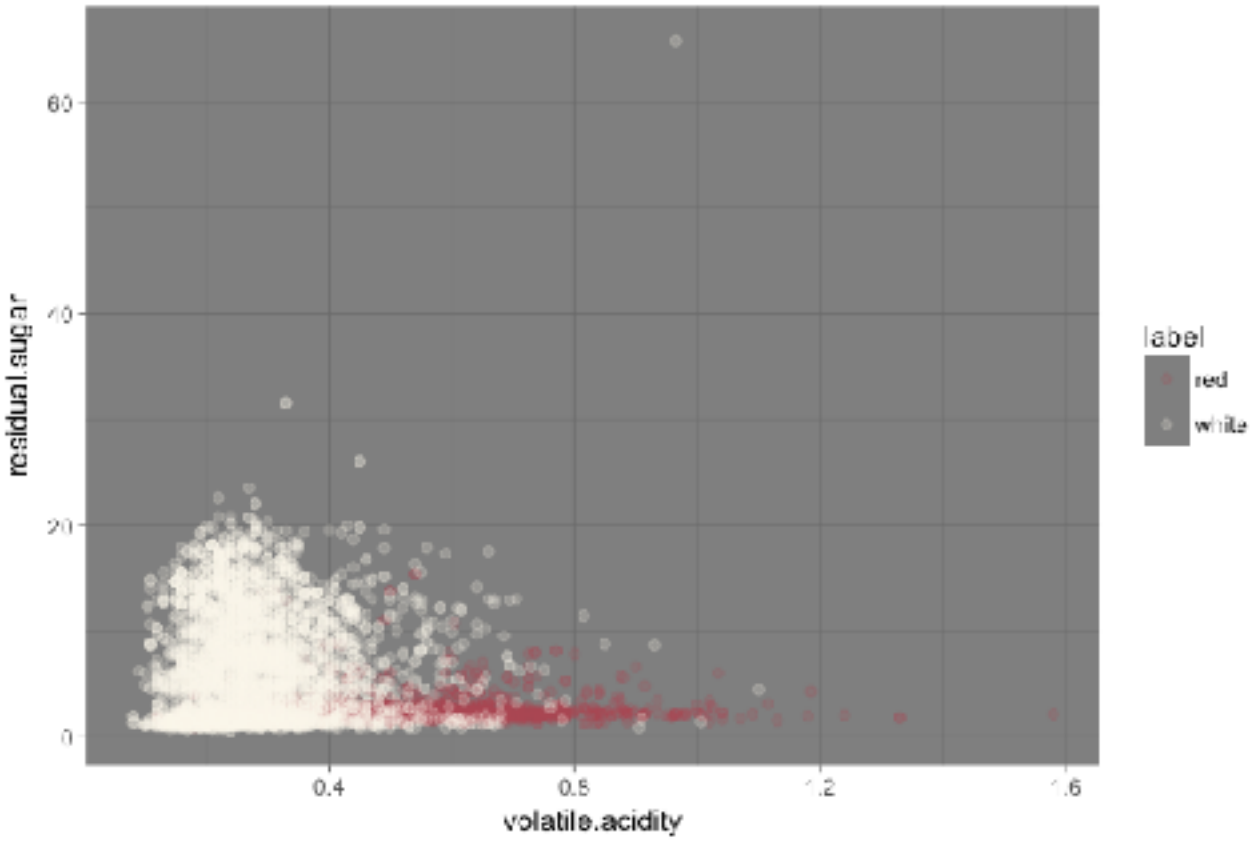
Σκόπος της παρακάτω εφαρμογής θα είναι η αξιολόγηση των αποτελεσμάτων του αλγορίθμου K means καθώς θα βασιστεί σε δεδομένα για τα οποία γνωρίζουμε εκ των προτέρων την κατηγορία στην οποία ανήκουν προκειμένου να συγκρίνουμε τα αποτελέσματα με την πραγματική ομαδοποίηση. Το σύνολο δεδομένων που θα χρησιμοποιήσουμε δημιουργήθηκε χρησιμοποιώντας δείγματα κόκκινου και λευκού κρασιού. Οι είσοδοι περιλαμβάνουν αντικειμενικές δοκιμές (π.χ. τιμές PH) και η έξοδος βασίζεται σε αισθητήρια δεδομένα (διάμεσος των τριών τουλάχιστον αξιολογήσεων των εμπειρογνωμόνων στον τομέα του οίνου). Κάθε εμπειρογνώμονας βαθμολόγησε την ποιότητα του κρασιού μεταξύ 0 (πολύ κακή) και 10 (πολύ εξαιρετική). Τα δεδομένα θα είναι της μορφής:

fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides	free.sulfur.dioxide	total.sulfur.dioxide	density	pH	sulphates	alcohol	quality	label
7.4	0.7	0.00	1.90	0.076	11	34	0.99	3.51	0.56	9.4	5	red

Παρατηρούμε ότι έχουμε στην κατοχή μας αν η αντίστοιχη μέτρηση προέρχεται απο κόκκινο η λευκό κρασί κάτι το οποίο όμως δεν θα το λάβουμε υπόψην μας για την εφαρμογή του αλγορίθμου k means. Αρχικά θα ρίξουμε μια ματιά στα δεδομένα μας μέσω γραφικών παραστάσεων . Η γραφική παρουσίαση των παρατηρήσεων είναι αρκετά σημαντική καθώς μας δίνει μία ιδέα σχετικά με τα χαρακτηριστικά στα οποία διαφέρουν η ταυτίζονται και κατα συνέπεια επηρεάζουν την μεταβλητότητά τους. Για παράδειγμα απο το παρακάτω ιστόγραμμα παρατηρούμε ότι τα δείγματα που αφορούν το κόκκινο κρασί έχουν μικρές τιμές για την μεταβλητή residual.sugar και απο κάποιο σημείο και μετά όλες οι παρατηρήσεις αφορούν άσπρο κρασί.Αναμένουμε επομένως ο μέσος της μίας ομάδας να έχει πιο μικρή τιμή στο συγκεκριμένο χαρακτηριστικό. Θα πρέπει να λάβουμε υπόψην μας το γεγονός ότι οι παρατηρήσεις που αφορούν λευκό κρασί είναι 4898 ενώ κόκκινο 1599 κάτι το οποίο θα κάνει πιο δύσκολη στην συνέχεια την προσπάθεια του αλγορίθμου να εντοπίσει ένα κόκκινο κρασί.



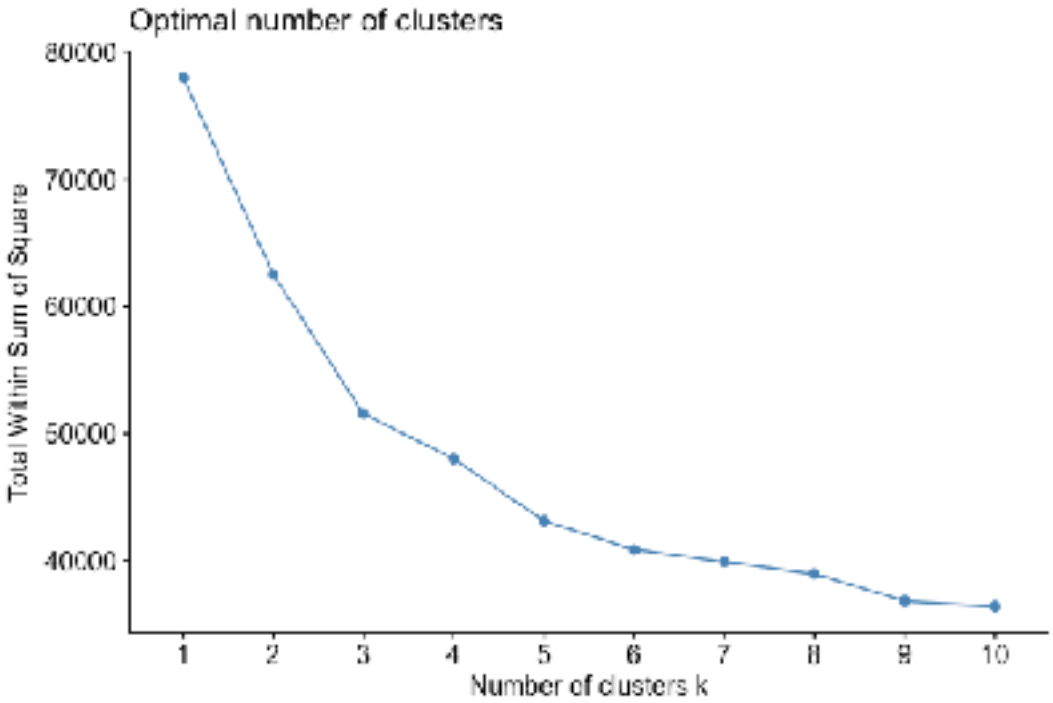
Απο το παρακάτω γράφημα μπορούμε να διακρίνουμε ότι τα περισσότερα λευκά κρασιά παίρνουν τιμές μικρότερες περίπου του 0.7 όσον αφορά την μεταβλητή volatile.acidity



γεγονός που πρόκειται να αξιοποιήσει ο αλγόριθμος για τον διαχωρισμό των

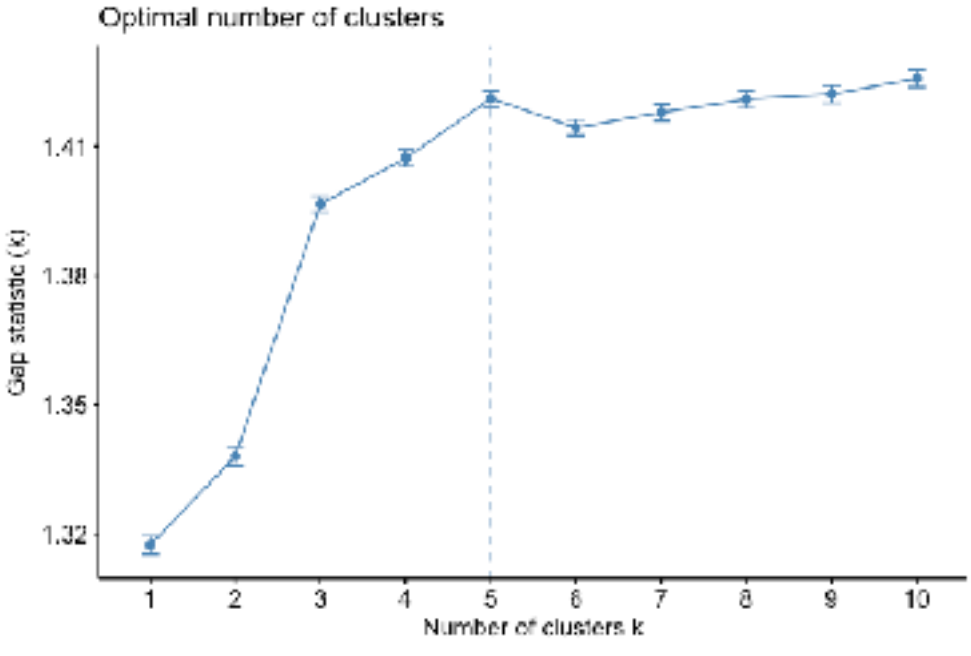
παρατηρήσεων σε ομάδες. Επιπλέον η διαφοροποίηση αυτή αναμένεται να υπάρχει και στους αντίστοιχους μέσους.
Για τον αριθμό των συμπλεγμάτων γνωρίζουμε εκ των προτέρων ότι τα δεδομένα μας προέρχονται απο 2 ομάδες. Είναι ιδιαίτερα ενδιαφέρον όμως να δούμε τι θας μα δείξουν οι μέθοδοι που χρησιμοποιήσαμε στις προηγούμενες εφαρμογές (έχει προηγηθεί τυποποίηση των παρατηρήσεων)

- **ELBOW METHOD**

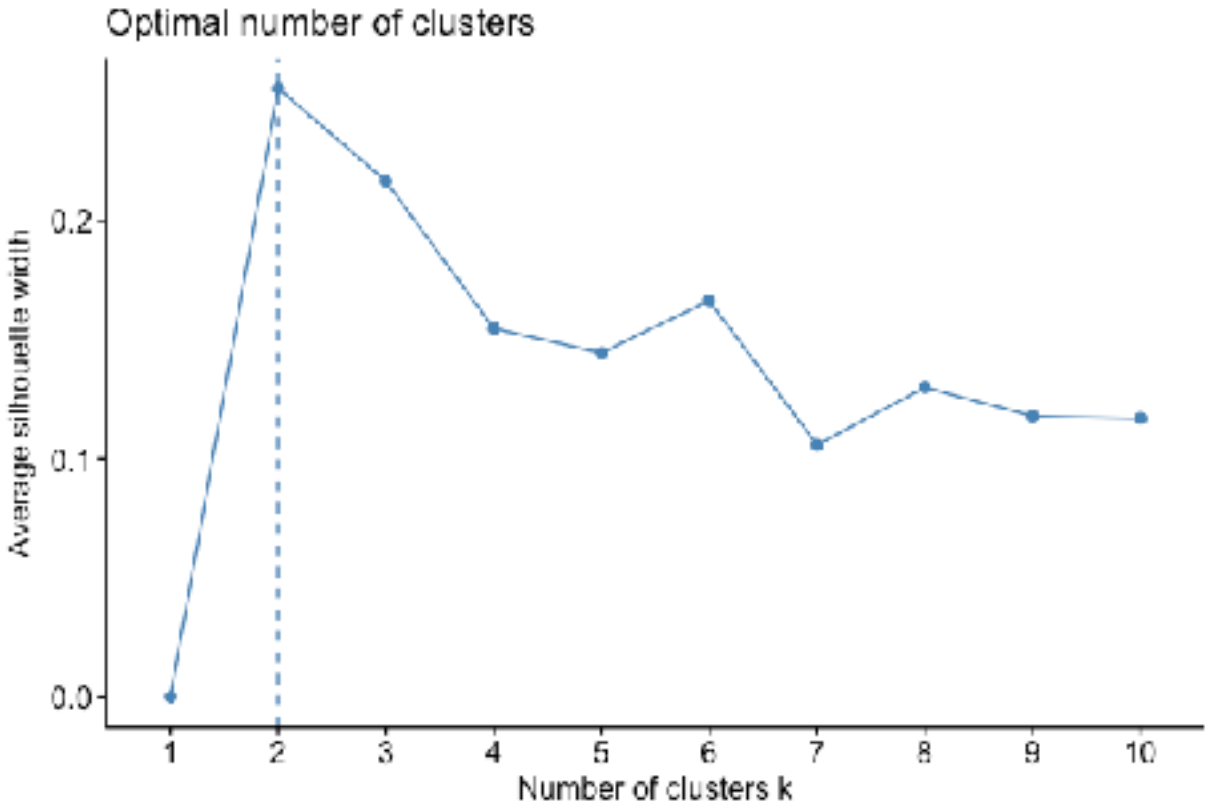


Απο το παραπάνω γράφημα θα μπορούσαμε να πούμε ότι προκύπτει το K=5 σαν ιδανική τιμή για τον αριθμό των συμπλεγμάτων κάτι με το οποίο συμφωνεί και η gap statistic method όπως βλέπουμε απο το επόμενο γράφημα.

- **GAP STATISTIC METHOD**



- **AVERAGE SILHOUETTE**



Η συγκεκριμένη μέθοδος βλέπουμε ότι μας έδωσε K=2 κάτι το οποίο ταυτίζεται και με την πραγματικότητα.Ο λόγος για τον οποίο τα αποτελέσματα δεν συμπίπτουν είναι διότι όπως είδαμε απο τα αρχικά γραφήματα υπήρχαν πολλές παρατηρήσεις για λευκά κρασιά που είχαν παρόμοια χαρακτηριστικά με αυτά των κόκκινων κρασιών το οποίο θα μπορούσε να χαρακτηριστεί και ως θόρυβος. Επίσης στην πραγματικότητα δεν υπάρχουν μόνο κόκκινα και λευκά κρασιά αλλά και άλλες ενδιάμεσες κατηγορίες όπως πχ ροζέ. Επομένως ο χωρισμός των παρατηρήσεων σε 2 μόνο κατηγορίες δεδομένης της χημικής τους σύστασης έχει ως δεδομένο την ύπαρξη θορύβου. Για την αντιμετώπιση παρόμοιων προβλημάτων, γνώση του αντικειμένου με το οποίο σχετίζονται οι παρατηρήσεις θα μπορούσε να δώσει μία αποτελεσματική λύση στην σωστή επιλογή του αριθμού του αριθμού K. Εφόσον γνωρίζουμε όμως εκ των προτέρων ότι K=2 , θα τρέξουμε τον αλγόριθμο για αυτή την τιμή.

	1	2
fixed.acidity	6.905	7.623
volatile.acidity	0.287	0.409
citric.acid	0.340	0.291
residual.sugar	7.245	3.076
chlorides	0.049	0.066
free.sulfur.dioxide	39.756	18.399
total.sulfur.dioxide	155.692	63.263
density	0.995	0.995
pH	3.191	3.255
sulphates	0.500	0.572
alcohol	10.259	10.797
quality	5.824	5.811

Απο τους μέσους βλέπουμε τις διαφοροποιήσεις για τις μεταβλητές residual.sugar, citric.acid που είχαμε αναφέρει και πριν καθώς επίσης και μεγαλή διαφοροποίηση στις μεταβλητές free.sulfur.dioxide , total.sulfur.dioxide. Επομένως λόγο της μεγάλης τιμής για το residual.sugar μπορούμε να πούμε ότι η ομάδα 1 είναι το κόκκινο κρασί ενώ η ομάδα 2 είναι το λευκό. Στη συνέχεια θα εξετάσουμε πόσο κοντά είναι τα αποτελέσματα του αλγορίθμου με την ομαδοποίηση την οποία γνωρίζουμε ότι ισχύει για τα δεδομένα μας.

	1	2
red	1515	84
white	1310	3588

Το παραπάνω πινακάκι αποτελεί ένα confusion matrix και παρατηρούμε ότι ο αλγόριθμος k means εντόπισε σχετικά καλά τα κόκκινα κρασιά , ενώ με τα λευκά ένα μεγάλο μέρος τον παρατηρήσεων τα τοποθέτησε στην κατηγορία με κόκκινα. Αυτό οφείλεται στην ύπαρξη θορύβου όπως αναφέραμε πιο πρίν αλλά και στο γεγονός ότι μπορεί μερικά λευκά κρασιά να έχουν την ίδια χημική σύσταση με κάποιο κόκκινο.
Συγκεκριμένα έχουμε TP=1515 , FP=1310 , FN=84 , TN=3588
Με βάση τις παραπάνω τιμές μπορούμε τώρα να υπολογίσουμε κάποιους απο τους δείκτες που αναφέραμε νωρίτερα.
Rand index = 0.785 ο οποίος μας δείχνει το ποσοστό των σωστών αποφάσεων δηλαδή 78.5 %
Fowlkes-Mallows index =0.713
Παρατηρούμε ότι ο δείκτης είναι σχετικά κοντά στο 1 γεγονός που δηλώνει ομοιότητα των δύο ομαδοποιήσεων όχι όμως σε αρκετά μεγάλο βαθμό.
Jaccard index = 0.52
Ο δείκτης αυτός βλέπουμε ότι είναι μικρότερος γεγονός που οφείλεται στην μεγάλη τιμή του FP παρουσιάζοντας μια όχι και τόσο καλή ομαδοποίηση.
Παρατηρούμε επομένως ότι οι δείκτες δεν είναι ανάγκη να συμφωνούν καθώς ο κάθε δείκτης έχει διαφορετική βαρύτητα για κάθε τιμή (FP,FN,TP,TN). Ανάλογα την αντίστοιχη εφαρμογή και το τι μας ενδιαφέρει κάθε φορά επιλέγουμε και τους αντίστοιχους δείκτες για αξιολόγηση στο τέλος. Επίσης παρατηρούμε οτι στο δεδομένο παράδειγμα ο αλγόριθμος K-means δεν βοηθάει αρκετά στον εντοπισμό του αριθμού K για τα συμπλέγματα αλλά ούτε και στο που θα ανήκει μία μελλοντική παρατήρηση λόγο της μεγάλης τιμής FP.