

Χρήση PCA, hierarchical clustering και k-means clustering

Σε δεδομένα κρασιών της Ιταλίας

Εργασίας για το μάθημα Δ2 Στατιστική Ανάλυση Δικτύων

Φοιτητές: Γεωργούλης Φίλιππος

Φουδούλη Αναστασία

Καθηγήτρια: Κολυβά-Μαχαίρα Φωτεινή

Περιεχόμενα

[1. Δεδομένα 3](#_Toc508148956)

[1.2. Βασικά στατιστικά χαρακτηριστικά των δεδομένων 4](#_Toc508148957)

[2. Principal Component Analysis 4](#_Toc508148958)

[3. Ανάλυση κατά συστάδες 11](#_Toc508148959)

[3.1. Συσταδοποίηση με τη μέθοδο k-means 12](#_Toc508148960)

[3.2 Συσταδοποίηση με Ιεραρχική μέθοδο. 17](#_Toc508148961)

### Δεδομένα

Για την παρούσα εργασία επιλέξαμε το wine dataset, το οποίο περιέχει τα αποτελέσματα χημικών αναλύσεων σε κρασιά που καλλιεργούνται σε μία συγκεκριμένη περιοχή της Ιταλίας όμως προέρονται από τρία διαφορετικά οινοποιεία. Το πλήθος των δειγμάτων ειναι 178 με αποτελέσματα 13 χημικών αναλύσεων για το καθένα.

Τα δεδομένα αντλήθηκαν από το UCI Machine Learning Repository (διεύθυνση: https://archive.ics.uci.edu/ml/datasets/wine)

Συγκεκριμένα έχουμε τα παρακάτω δεδομένα:

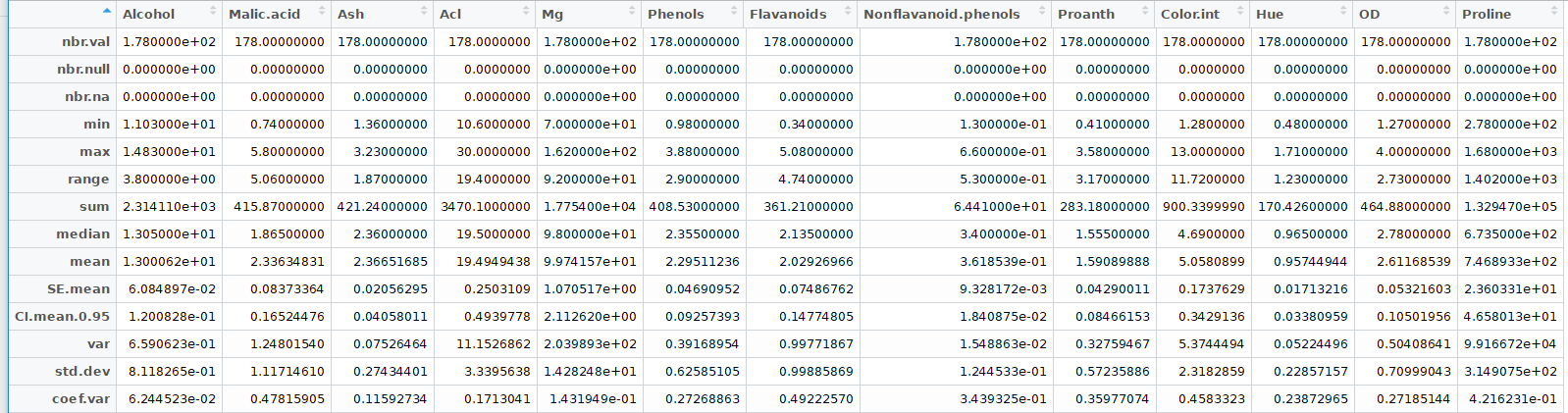
|  |  |
| --- | --- |
| Alcohol | Περιεκτικότητα σε αλκοόλ |
| Malic | Περιεκτηκότητα σε μηλικό οξύ |
| Ash | Περιεκτικότητα σε τέφρα |
| Alcalinity | Αλκαλικότητα τέφρας |
| Magnesium | Περιεκτικότητα σε μαγνήσιο |
| Phenols | Συνολικές φαινόλες στο κρασί |
| Flavanoids | Περιεκτικότητα σε φλαβανοειδή |
| Nonflavanoids | Περιεκτικότητα σε μη φλαβανοειδείς φαινόλες |
| Proanthocyanins | Περιεκτικότητα σε προανθοκυανινες |
| Color | Ένταση χρώματος |
| Hue | Απόχρωση |
| Dilution | Διάλυση D280 / OD315 των αραιωμένων οίνων |
| Proline | Περιεκτικότητα προλίνης |

Υποσημείωση1: Επιπλέον στα δεδομένα περιέχεται ο τύπος κάθε κρασιού, που έχει μετατραπεί απο ποιοτική σε κατηγορική μεταβλητή και υπάρχουν 59 κρασιά τύπου1, 71 τύπου2 και 48 τύπου3. Για την ανάλυση που θα κάνουμε δεν θα ληφθεί υπόψαν στα δεδομένα.

Υποσημείωση2: δεν υπάρχουν missing values στα δεδομένα

### Βασικά στατιστικά χαρακτηριστικά των δεδομένων

Χρησιμοποιήθηκε η βιβλιοθήκη pastecs της R για το παρακάτω διάγραμμα. Επιστρέφει το πλήθος των παρατηρήσεων που υπολογίστηκαν για κάθε μεταβλητή, αν υπάρχουν NA ή μηδενικές τιμές, την ελάχιστη και τη μέγιστη τιμή της παρατήρησης κάθε μεταβλητής, καθώς και το εύρος. Επιστρέφει ακόμα τη διάμεσο, τη μέση τιμή, τη διασπορά και την τυπική απόκλιση κάθε μεταβλητής κλπ.



Γίνεται λοιπόν εμφανές το πρόβλημα ότι η μονάδα μέτρησης κάθε ενός χαρακτηριστικού είναι διαφορετική. Αυτό μπορεί εύκολα να δημιουργήσει προβήματα σε προβλήματα ταξινόμησης αλλά και σε μεθόδους παραγοντικής ανάλυσης, καθώς θα δωθεί μεγαλύτερο βάρος σε χαρακτηριστικά με μεγάλες τιμές χωρίς όμως αυτό να είναι πάντοτε ορθό.

### Principal Component Analysis

Στη συνέχεια πραγματοποιήσαμε ανάλυση σε κύριες συνιστώσες. Η μέθοδος αυτή έχει στόχο την δημιουργία γραμμικών συνδυασμών των αρχικών μεταβλητών ώστε οι γραμμικοί αυτοί συνδυασμοί να είναι και ασυσχέτιστοι μεταξύ τους και να περιέχουν το μεγαλύτερο μέρος της διακύμανσης των αρχικών μεταβήτών.

Έστω ένας τετραγωνικός πίνακας Α, τότε η φασματική ανάλυση του Α είναι Α=ΡΛΡ’ όπου Ρ ο ορθογώνιος πίνακας των κανονικοποιημένων ιδιοδιανυσμάτων και Λ ο διαγώνιος πίνακας των ιδιοτιμών. Θεωρώντας τώρα ότι ο Α είναι ο πίνακας συνδιασποράς ενός συνόλου δεδομένων, τότε με κατάλληλο μετασχηματισμό Ρ’ΑΡ=Λ μπορεί να μετατραπεί σε διαγώνιο διαγώνιο πίνακα συνδιασποράς. Με βάση αυτή την ανάλυση είναι εύκολο να πραγματοποιηθεί ανάλυση σε κύριες συνιστώσες.

Έστω ότι το σύνολο δεδομένων μας αποτελείται από κ μεταβλητές. Άρα θέλουμε να κατασκευάσουμε κ συνιστώσες που είναι ο γραμμικός μετασχηματισμός των μεταβλητών αυτών. Δηλαδή Υ=ΑΧ, όπου Υ το διάνυσμα των συνιστωσών, Χ το διάνυσμα των μεταβλητών και Α ο πίνακας των γραμμικών συνδυασμών. Δηλαδή για να βρεθούν οι συνιστώσες αρκεί να βρεθεί ο πίνακας Α ο οποίος είναι ο πίνακας των κανονικοποιημένων ιδιοτιμών.

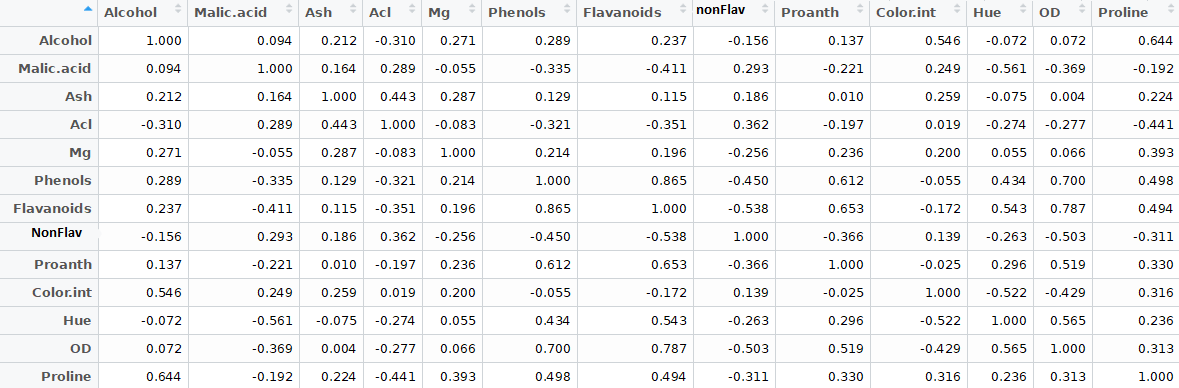
Ακολουθείται λοιπόν η εξής διαδικασία:

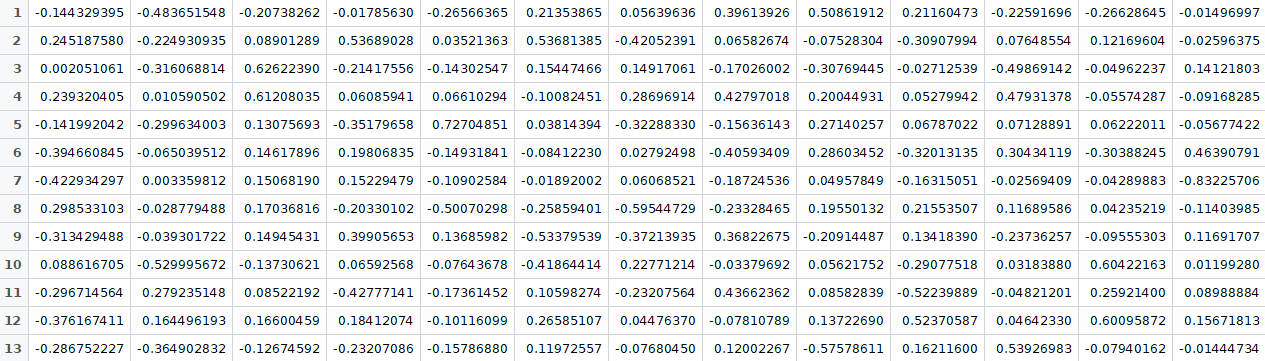
* Ευρεση ιδιοτιμών και ιδιοδιανυσμάτων του πίνακα Σ, που μπορεί να είναι είτε πίνακας συναδιακύμασνης είτε πίνακας συνδιασποράς
* Η μεγαλύτερη ιδιοτιμή και το ιδιοδιάνυσμά της αντιστοιχούν στη πρώτη κύρια συνιστώσα, η δεύτερη μεγαλύτερη στη δεύτερη κλπ.
* Η διακύμανση κάθε συνιστώσας ισούται με την ιδιοτιμή που της αντιστοιχεί Var(Yj)=λj
* Ο πίνακας διακύμανσης των κυρίων συνιστωσών είναι ο διαγώνιος πίνακας των ιδιοτιμών Λ
* Η συνολική διακύμανση αρχικών μεταβλητών και συνιστωσών ταυτίζονται (tr(Σ)=tr(Λ))
* Η ποσότητα λj/Σλj ερμηνεύεται ως το ποσοστό της συνολικής διακύμασης που εξηγεί η j-συνιστώσα

Άρα λοιπόν, διατηρώντας όλες τις συνιστώσες, δεν χάνω πληροφορία, ενώ κρατώντας τις πρώτες εξηγώ μεγάλο ποσοστό της πληροφορίας αυτής, αφού οι μεγαλύτερες ιδιοτιμές αντιστοιχούν στις πρώτες συνιστώσες και ερμηνεύονται ως η διακύμανση της κάθε συνιστώσας.

Μία πρώτη επιλογή που κληθήκαμε να κάνουμε, ήταν ο πίνακας στον οποίο θα εφαρμόζαμε την ανάλυση. Επειδή οι τάξεις μεγάθων των μεταβλητών διαφέρουν, τα δεδομένα δεν είναι μετρημένα στην ίδια μονάδα μέτρησης καθώς και η διασπορά αυτών διαφέρει και κυμένεται απο έως και επιλέξαμε τον πίνακα συσχετίσεων, επειδή οι συσχετίσεις δεν αλλάζουν όταν αλλάζει η μονάδα μέτρησης ή κλίμακας.

* Υπολογισμός Pearson correlation matrix των δεδομένων

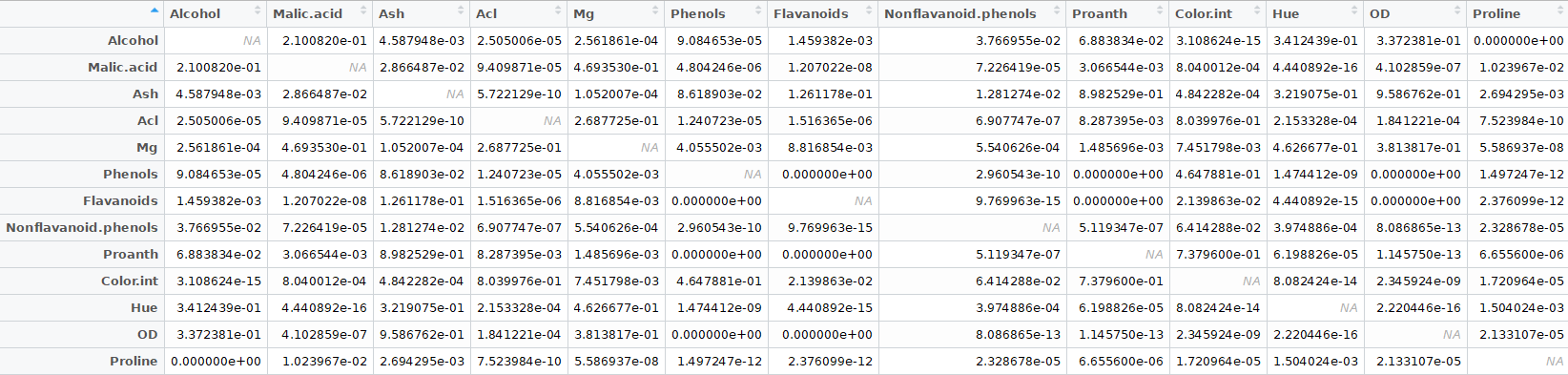
Ο πίνακας που προέκυψε είναι ο παρακάτω:



Με αντίστοιχο πίνακα p\_values:

Σχεδόν όλα τα pvalues παραπάνω είναι πάρα πολύ μικρά, που ερμηνεύεται ως απόρριψη της μηδενικής υπόθεσης ότι δηλαδή οι πραγματικές συσχετίσεις των μεταβλητών είναι 0. Άρα θεωρούμε ότι οι συσχετίσεις είναι πραγματικές και στατιστικά σημαντικές, και αφού υπαρχουν και οι μεταβλητές δεν είναι ασυσχέτιστες έχει νόημα η ανάλυση σε κύριες συνιστώσες.

Επιπλέον εφαρμόσαμε δύο ακόμα τεστ.



* Υπολογισμός Batlett's sphericity test και KMO index

Πρακτικά και τα δύο τεστ ερμηνεύουν την ύπαρξη σημαντικών συσχετίσεων έμμεσων και μη, ως λόγο πραγματοποίσησης ανάλυσης σε κύριες συνιστώσες.

Αν οι συσχετίσεις δεν είναι επαρκώς μεγάλες, αυτό σημαίνει ότι δεν υπάρχει κάποιος λόγος να γίνει ανάλυση σε κύριες συνιστώσες, καθώς συνήθως κάθε συνιστώσα θα ερμηνεύει μία μεταβλητή και όχι γραμμικό συνδυασμό αυτών.

* Bartlett’s sphericity test

Χρησιμοποιήθηκε το τέστ από την βιβλιοθήκη psych της R

Τα αποτελέσματα ήταν τα παρακάτω:

$chisq

[1] 1317.181

$p.value

[1] 2.468617e-224

$df

[1] 78

H p.value του τέστ είναι πολύ μικρή και <0,05 άρα μπορόυμε χωρίς σημαντικό σφάλμα να απορρίψουμε την μηδενική υπόθεση, οτι δηλαδή ο πίνακας συσχετίσεων είναι ο μοναδιαίος και άρα δεν υπαρχουν συσχετίσεις. Όμως παρόλα αυτά γνωρίζουμε ότι το Bartlett’s test for sphericity τείνει να είναι στατιστικά σημαντικό όταν ο λόγος >5 και στα συγκεκριμένα δεδομένα το λόγος αυτός προκείπτει 13.6923 οπότε δεν εμπιστευόμαστε απόλυτα αυτό το συγκεκριμένο τέστ.

* ΚΜΟ index

Πάλι χρησιμποποιήθηκε εντολή από τη βιβλιοθήκη psych

Τα αποτελέσματα:

Kaiser-Meyer-Olkin factor adequacy

Call: KMO(r = R)

Overall MSA = 0.78

Παρατηρούμε ότι η τιμή του δείκτη είναι 0,78 η οποία είναι εξαιρετική για να περάσουμε σε ανάλυση σε κύριες συνιστώσες. Σε συνδυασμό με τα δύο προηγούμενα αποτελέσματα έχει νόημα η ανάλυση.

* Υπολογισμός ιδιοτιμών του πίνακα συσχετίσεων

Σε φθίνουσα σειρά είναι οι παρακάτω:

4.7059 2.4970 1.4461 0.9190 0.8532 0.6417 0.5510 0.3485 0.2889 0.2509 0.2258 0.1688 0.1034

* Και των αντίστοιχων ιδιοδιανυσμάτων

Δηλαδή η πρώτη συνιστώσα είναι:

*...* με Var()==4.7059 κλπ

* Επιλογή αριθμού συνιστωσών που θα διατηρηθούν

Μια ακόμη δύσκολη επιλογή που κριθήκαμε να κάνουμε ήταν αυτή του πλήθους των συνιστωσών έπρεπε να διατηρήσουμε. Με βάση

* Το κριτήριο του Kaiser

Eπιλέγουμε συνιστώσες όσες και οι ιδιοτιμές που είναι μεγαλυτερες από τη μέση τιμή των ιδιοτιμών; . Στο πίνακα συσχετίσεων επειδή =1 αρκέι να παρουμε μόνο τις ιδιοτιμές που είναι μεγαλύτερες της μονάδας. Σύμωνα με αυτό το κριτήριο, αρκεί να διατηρήσουμε 3 συνιστώσες.

* Ποσοστό συνολικής διακύμανσης που εξηγούν οι συνιστώσες

Αρκεί να βάλουμε ένα κατώφλι, πχ 85% και διαλέξουμε αρκιβώς τόσες συνιστώσες ώστε αθροιστικά να εξηγούν μεγαλύτερο ποσοστό από αυτό το κατώφλι.

Η κάθε συνιστώσα μας εξηγεί

36.20 19.21 11.12 7.07 6.56 4.94 4.24 2.68 2.22 1.93 1.74 1.30 0.80

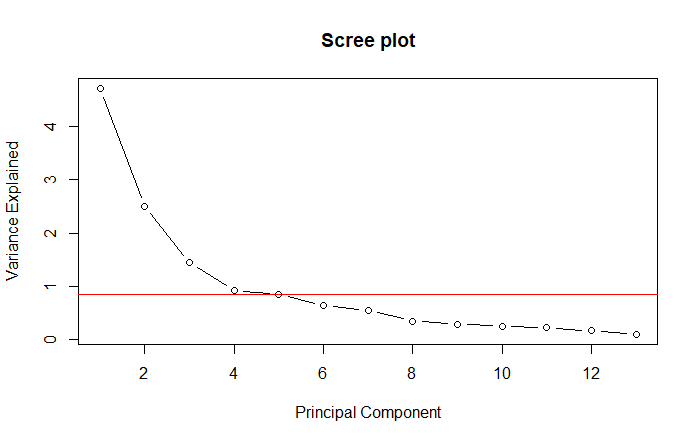
ποσοστό της ματαβλητότητας. Επομένως αθροιστικά εξηγεί

36.20 55.41 66.53 73.60 80.16 85.10 89.34 92.02 94.24 96.17 97.91 99.20 100.00

ποσοστό της μεταβλητότητας. Επομένως με βάση αυτό το κριτήριο, με κατώφλι 85% αρκεί να κρατήσουμε 6 συνιστώσες.

* Scree plot

Το scree plot είναι ένα γράφημα το οποίο έχει στον οριζόντιο άξονα τη σειρά και στον κάθετο την τιμή κάθε ιδιοτιμής. Επιλέγουμε τόσες συνιστώσες μέχρι το γράφημα να αρχίσει να αλλάζει κλίση. Το scree plot που προέκυψε από τα δεδομένα μας είναι παρακάτω και σύμωνα με αυτό η κλίση φαίνεται να αλλάζει μετά την 4 συνιστώσα, άρα θα διατηρούσαμε 4.



Το πλήθος των συνιστωσών διαφέρει ανάλογα με την επιλογή του κριτηρίου. Εμείς αποφασίσαμε να κρατήσουμε τις 6 πρώτες συνιστώσες, καθώς θέλαμε να διατηρήσουμε όσο γίνεται μεγαλύτερο κομμάτι της διακύμανσης, πέφτοντας όμως σημαντικά σε διαστάσης. Δεν κρατήσαμε 3 ή 4 συνιστώσες όπως προκύπτουν από το κριτήριο του Kaiser και το scree plot αντίστοιχα, γιατί στην πρώτη περίπτωση χάνεται 33.47% και στη δεύτερη 26.4% της συνολικής διακύμανσης.

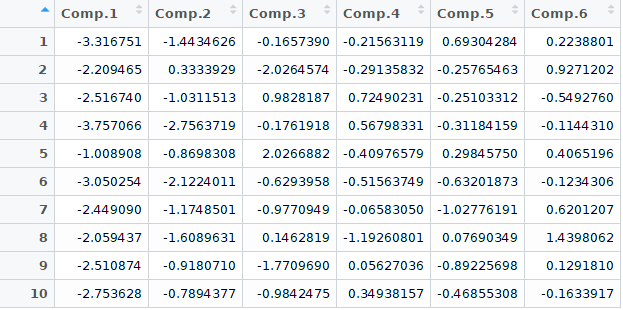
* Eύρεση κυρίων συνιστωσών

Οι κύριες συνιστώσες λοιπόν που διατηρήσαμε είναι οι παρακάτω:

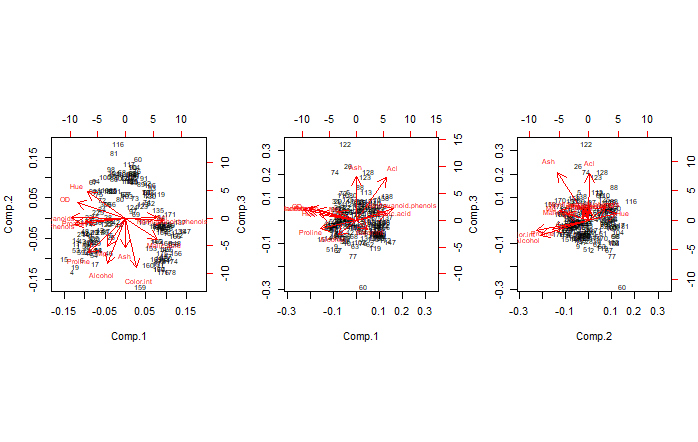
PCA loadings

|  |
| --- |
| Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6  Alcohol -0.144 -0.484 -0.207 -0.266 0.214  Malic.acid 0.245 -0.225 0.537 0.537  Ash -0.316 0.626 -0.214 -0.143 0.154  Acl 0.239 0.612 -0.101  Mg -0.142 -0.300 0.131 -0.352 0.727  Phenols -0.395 0.146 0.198 -0.149  Flavanoids -0.423 0.151 0.152 -0.109  Nonflavanoid.phenols 0.299 0.170 -0.203 -0.501 -0.259  Proanth -0.313 0.149 0.399 0.137 -0.534  Color.int -0.530 -0.137 -0.419  Hue -0.297 0.279 -0.428 -0.174 0.106  OD -0.376 0.164 0.166 0.184 -0.101 0.266  Proline -0.287 -0.365 -0.127 -0.232 -0.158 0.120 |

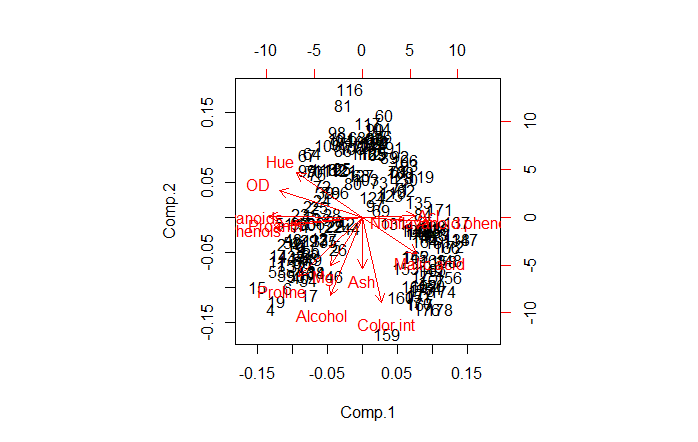
PCA scores, για τις πρώτες 10 παρατηρήσεις



Ένα επιπλέον χρήσιμο εργαλείο για να δούμε την ομαδοποίηση των δεδομένων είναι τα biplots με άξονες τις πρώτες κύριες συνιστώσες. Στα biplots μπορούμε να δούμε σε μορφή διανύσματος κάθε μεταβλητή και να δούμε πόσο «κοντά» είναι η μία στην άλλη. Επιπροσθέτως μπορούμε να παρατηρήσουμε πόσο καλυτερη διασπορά στο χώρο έχουν οι παρατηρήσεις στο πρώτο γράφημα με άξονες τις δύο πρώτες συνιστώσες και πόσο οι παρατηρήσεις αρχίζουν να καλύπτουν η μία την άλλη στα δύο επόμενα γραφήματα.

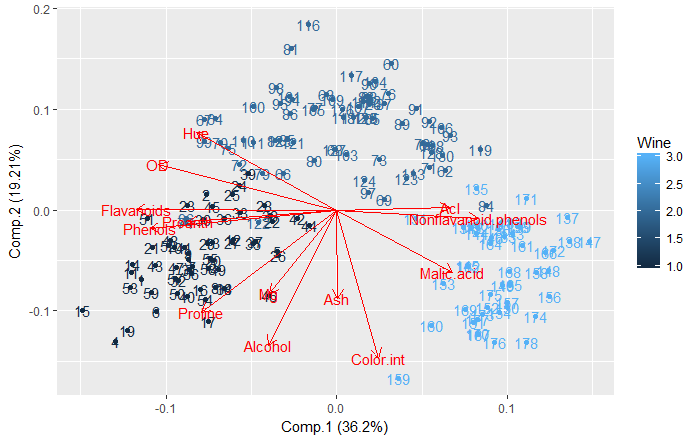


55.39% 47.31% 30.33%



Παρατηρώντας λοιπόν το πρώτο γράφημα, μπορούμε να δούμε ότι τα φλαβανοειδή, τα μη φλαβανοειδή, οι φαινόλες, η προάνθη καθώς και το αλκοόλ έχουν μεγάλη γραμμική συσχέτιση μεταξύ τους.

Αξίζει να παρατηρηθεί ότι επειδή τα δεδομένα που επιλέξαμε είναι ήδη χωρισμένα σε 3 διαφορετικά είδη κρασιού, το rotation που έγινε με τις κύριες συνιστώσες έκανε και ένα άτυπο clustering, όπως φαίνεται στο παρακάτω γράφημα.



### Ανάλυση κατά συστάδες

Η ανάλυση κατά συστάδες είναι µια µέθοδος που σκοπό έχει να κατατάξει σε οµάδες τις υπάρχουσες παρατηρήσεις χρησιµοποιώντας την πληροφορία που υπάρχει σε κάποιες µεταβλητές. Με άλλα λόγια η ανάλυση κατά συστάδες εξετάζει πόσο όµοιες είναι κάποιες παρατηρήσεις ως προς κάποιον αριθµό µεταβλητών µε σκοπό 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 αλλά παρατηρήσεις διαφορετικών οµάδων να διαφέρουν όσο γίνεται περισσότερο.

Παρακάτω δουλέψαμε με δύο μεθόδους:

* Τον αλγόριθμο K-means, που με έναν επαναληπτικό αλγόριθμο προσπαθεί να τοποθετήσει κάθε παρατήρηση σε ομάδες ανάλογα με το πια ομάδα είναι πιο κοντά στην παρατήρηση
* Και την ιεραρχική μέθοδο, όπου κάθε παρατήρηση αρχικά αποτελεί μόνη της μία ομάδα, στη συνέχεια 2 παρατηρήσεις με μικρή απόσταση εννόνονται. Αν δύο παρατηρήσεις εννοθούν, εννόνουμε μία προυπάρχουσα ομάδα με μία παρατήρηση μέχρι να φτιάξουμε ομάδα κοκ

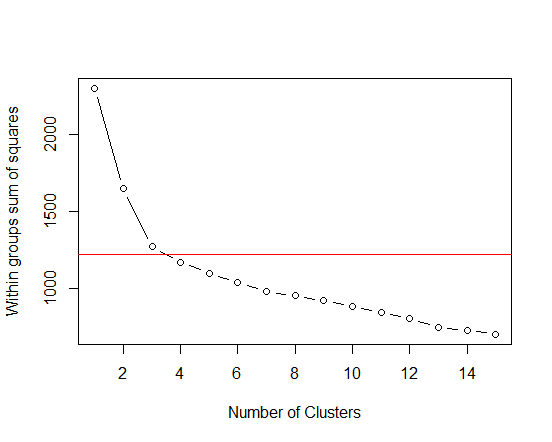
### 3.1. Συσταδοποίηση με τη μέθοδο k-means

Η περιγραφή του αλγορίθμου είναι η παρακάτω:

* Επιλέγουμε το πλήθος των κέντρων (centroids)
* Η παρατήρηση κατατάσσεται σε μία ομάδα αν η απόστασή της από το κέντρο της συγκεκριμένης ομάδας είναι μικρότερη από την απόστασή της από κάθε άλλο κέντρο
* Μετά την κατάταξη όλων των παρατηρήσεων, υπολογίζονται νέα κέντρα (:= διάνυσμα των μέσων)
* Επανάληψη διαδικασίας μέχρι να μην υπαρχουν διαφορές ανάμεσα σε δύο διαδοχικές επαναλήψεις

Στην R επιλέξαμε να χρησιμοποιήσουμε τον αλγόριθμο Hartigan-Wong. Κάθε παρατήρηση τοποθετείται στην ομάδα που ελαχιστοποιεί το SS(k)= , όπου k είναι ο αριθμός των κλάσεων η τιμής της j-μεταβλητής της i-παρατήρησης η τιμή της ξ-μεταβλητής για την k-κλάση

Επειδή όταν πριγράφαμε τα δεδομένα παρατηρήσαμε οτι οι τιμές των διαφόρων μεταβλητών διαφέρουν σημαντικά (τιμές της τάξης του για φαινόλες και του για προλίνη), πρωτού προχωρήσουμε στον αλγόριθμο, έγινε αρχικά κανονικοποίηση των δεδομένων έτσι ώστε αυτά να ακολουθούν την κανονική κατανομή Ν(0,1).

Ο αλγοριθμός προυποθέτει να επιλέξουμε το πλήθος των κέντρων, γι αυτό κατασκευάσαμε το παρακάτω γράφημα που μας δίνει το άθροισμα των τετραγώνων των αποστάσεων από το κέντρο κάθε κλάσης όπου έχει τοποθετηθεί σε σχεση με το πλήθος των κλάσεων.

Παρατηρείται απότομη πτώση καθώς πάμε από 1 σε 3 κλάσεις, ενώ στη συνέχεια η πτώση της τιμής του αθροίσματος γίνεται πιο ομαλά. Αυτό αποτελεί μία καλή ένδειξη για να επιλέξουμε 3 κλάσεις.

Επιπλέον χρησιμοποιήθηκε η βιβλιοθήκη NbClust της R που προσφέρει ένα πακέτο 30 κριτηρίων για βέλτιστη επιλογή πλήθους clusters. Προέκυψε το αποτέλεσμα:

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* Among all indices:

\* 4 proposed 2 as the best number of clusters

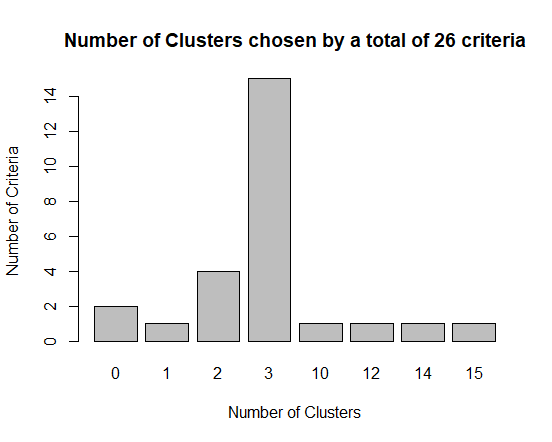
\* 15 proposed 3 as the best number of clusters

\* 1 proposed 10 as the best number of clusters

\* 1 proposed 12 as the best number of clusters

\* 1 proposed 14 as the best number of clusters

\* 1 proposed 15 as the best number of clusters



\*\*\*\*\* Conclusion \*\*\*\*\*

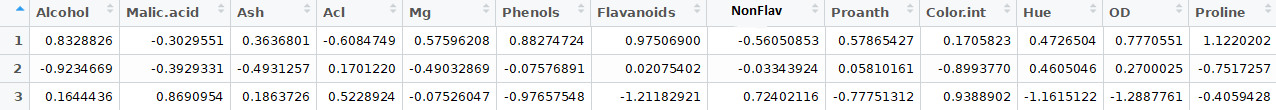
\* According to the majority rule, the best number of clusters is 3

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

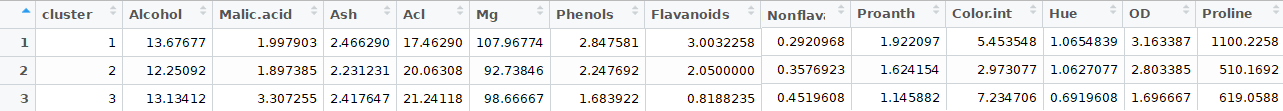
Σύμφωνα με όλα τα παραπάνω αποφασίσαμε ότι ο βέλτιστος αριθμός ομάδων είναι 3. Επιπλέον αυτό συμφωνεί με την αρχική μας γνώση για ύπαρξη 3 τύπων κρασιού στα δεδομένα μας. Γι αυτό εφαρμόσαμε τον αλγόριθμο στα δεδομένα που έχουμε αφαιρέσει την ιδιότητα τύπος κρασιού και στη συνέχεια ελέγξαμε αν διέκρινε σωστά τις ομάδες. Τέλος σαν απόσταση επιλέξαμε την ευκλείδια.

Ο αλγόριθμος σταμάτησε μετά από 3 επαναλήψεις και τοποθέτησε 62, 65 και 51 παρατηρήσεις σε κάθε μία από τις κλάσεις.

Τα κέντρα ήταν:



Επειδή τα παραπάνω κέντρα είναι υπολογισμένα στα κανονικοποιημένα δεδομένα έγινε μετασχηματισμός τους και τα κέντρα στα αρχικά δεδομένα είναι:



Έγινε πολύ καλη τοποθέτηση των δεδομένων σε ομάδες. Σε σχέση με τια πραγματικά γκρουπ των δεδομένων δίνεται το παρακάτω cross\_tab

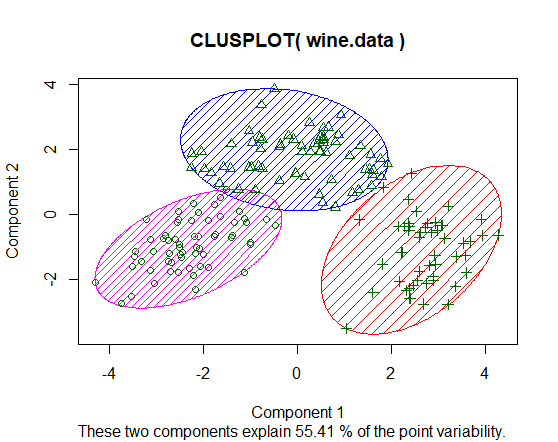
1 2 3

1 59 0 0

2 3 65 3

3 0 0 48

Όλα τα κρασιά τύπου 2 τοποθετήθηκαν στην ίδια κλάση, ενώ μόνο τρία κρασιά τύπου 1 και τρία κρασιά τύπου 2 τοποθετήθηκαν σε λάθος ομάδα. Συνολικά δηλαδή το 89,75 των δεδομένων τοποθετήθηκε στη σωστή κλάση.

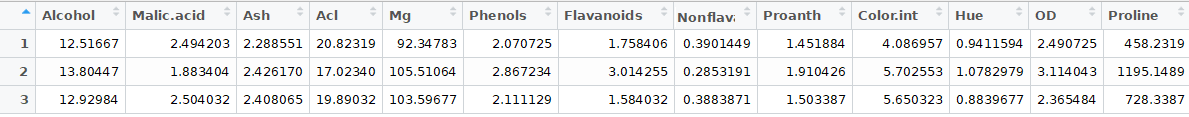


Στο γράφημα φαίνονται τα δεδομένα όπως τοποθετήθηκαν σε κάθε κλάση. Το γράφημα έχει γίνει με άξονες τους δύο πρώτους principal components που εξηγούν το 55,41% της μεταβλητότητας των αρχικών δεδομένων καθώς μπορεί ν δώσει ένα καλύτερο οπτικό αποτέλεσμα.

Για να συγκριθούν τα αποτελέσματα τρέξαμε επιπλέον τον αλγόριθμο στα δεδομένα πριν την κανονικοποίηση. Τα αποτελέσμα ήταν σημαντικά χειρότερα και δίνονται παρακάτω:

Τοποθέτησε 69, 47 και 62 παρατηρήσεις σε κάθε κλάση

Τα κέντρα ήταν :



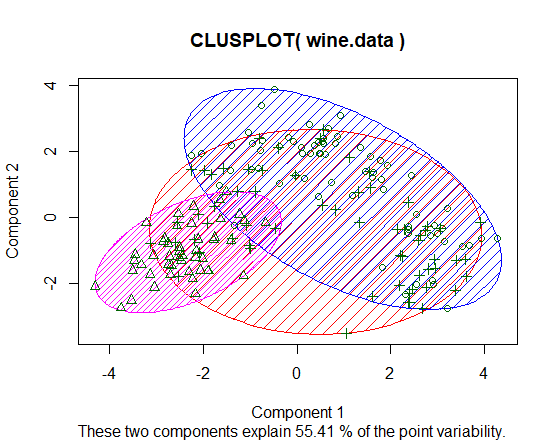
1 2 3

1 0 46 13

2 50 1 20

3 19 0 29 το cross-tab και 37.11% η επιτυχής τοποθέτηση

Τέλος με το παρακάτω διάγραμμα φαίνεται και οπτικά η αποτυχία σωστής τοποθέτησης.



Υποσημείωση: Επειδή τα κέντρα επιλέγονται τυχαία σε κάθε εκτέλεση του αλγορίθμου, τα αποτελέσματα κάθε εκτέλεσης μπορούν να διαφέρουν κάθε φορα.

### 3.2 Συσταδοποίηση με Ιεραρχική μέθοδο.

Στην Ιεραρχική ομαδοποίηση ο αριθμός των ομάδων δεν είναι γνωστός. Ο αλγόριθμος ξεκινάει δημιουργώντας αρχικά μία ομάδα για κάθε παρατήρηση και στη συνέχεια ενώνει παρατηρήσεις που βρίσκονται πιο κοντά. Τέλος κάθε τέτοιος αλγόριθμος δουλεύει σε έναν πίνακα αποστάσεων (δηλαδή τις αποστάσεις όλων των παρατηρήσεων από τις υπόλοιπες).

Συνοπτικά τα βήματα του αλγορίθμου είναι:

* Δημιουργία πίνακα αποστάσεων για όλες τις ομάδες
* Ένωση δύο παρατηρήσεων με την μικρότερηαπόσταση

Δηλαδή δημιουργείται μία ομάδα με τις παρατηρήσεις που είναι πιο κοντά.

Αν η μικρότερη απόσταση αφορά μία ήδη δημιουργηθείσα ομάδα και μία παρατήρηση απλά βάζουμε αυτή την παρατήρηση στην ομάδα.

Αν αφορά δύο ομάδες τις εννώνοουμε.

* Τέλος αλγορίθμου όταν όλες οι παρατηρήσεις τοποθετηθούν σε μία ομάδα.

Παρατηρούμε ότι η επιλογή της απόστασης είναι πάρα πολύ σημαντική για την «επιτυχία» του αλγορίθμου. Δοκιμάσαμε τις παρακάτω αποστάσεις:

* Euclidean

(p,q):=

* Chebyshev/ Maximum distance

*:=* i=1,2,..,n

* Manhattan

==

* Canberra

(p,q)=

* Minkowski

(p,q)=, για p=3,4,5

Ακόμα σημαντικό ρόλο παίζει η επιλογή της μεθόδου συσσωμάτωσης (τρόπος υπολογισμού απόστασης μεταξύ δύο ομάδων). Έγινε ανά δύο συνδυασμός των παραπάνω αποστάσεων με τις παρακάτω linkage μεθόδους:

* Nearest neighbor ή single linkage

Υπολογίζει την απόσταση ανάμεσα σε δύο ομάδες ως τη μικρότερη απόσταση από μία παρατήρηση μέσα σε μία ομάδα με κάποια παρατήρηση στην άλλη ομάδα.

* Furthest neighbor ή complete linkage

Υπολογίζει την απόσταση ανάμεσα σε δύο ομάδες ως τη μεγαλύτερη απόσταση από μία παρατήρηση στην μία ομάδα με κάποια παρατήρηση στην άλλη ομάδα.

* Average between groups

Υπολογίζει την απόσταση ανάμεσα σε δύο ομάδες ως το μέσο της απόστασης ανάμεσα σε ολές τις αποστάσεις μια ομάδας με τα στοιχεία της άλλης

* Average within groups

Υπολογίζει την απόσταση ανάμεσα σε δύο ομάδες ως το μέσο όρο όλων των αποστάσεων που προκύπτουν όταν ενώσουμε τις δύο ομάδες (δηλαδή υπολογίζω και αποστάσεις within group).

* Centroid

Υπολογίζει την απόσταση ανάμεσα σε δύο ομάδες ως την απόσταση των κέντρων των ομάδων

* Ward.D

Είναι αρκετά διαφορετική από τις προηγουμενες μεθόδους καθώς προσπαθεί να ελαχιστοποιήσει την διακύμανση μέσα στις ομάδες. Για κάθε παρατήρηση υπολογίζεται η απόστασή της από το κέντρο της ομάδας. Αθροίζοντας για ολές τις ομάδες παίρνουμε το συνολικό άθροισμα. Σε κάθε βήμα ενώνει ομάδες που οδηγούν στη μικρότερη αύξηση αυτού του αθροίσματος.

* Average

Υπολογίζει την απόσταση ανάμεσα σε δύο ομάδες ως ένα σταθμισμένο μέσο. Η στάθμιση γίνεται μέσω του πλήθους των παρατηρήσεων που περιέχονται σε κάθε κλάση.Εννώσει τις ομάδες με τη μικρότερη απόσταση. Αν παραδείγματος χάρην υπάρχουν NP παρατηρήσεις στη Ρ κλάση και ΝQ στη Q τότε αν εννωθούν προκύπτει

d(P+Q,R)=NP\*d(P,R)/(NP+NQ)+NQ\*d(Q,R)/(NP+NQ)

* Mcquitty

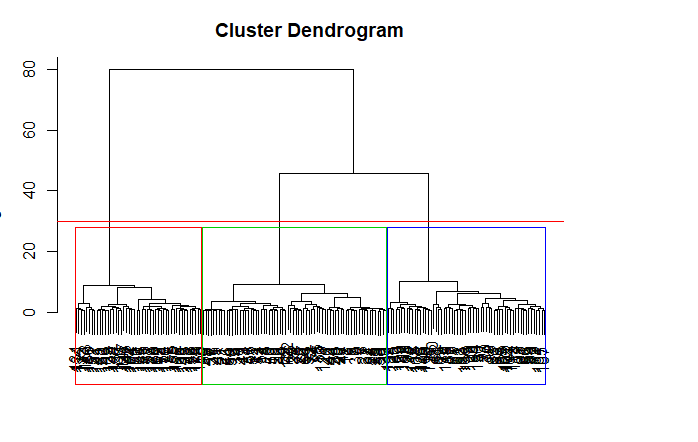
Αποτελεί απλοποίηση της παραπάνω μεθόδου. Υπολογίζει την απόσταση ανάμεσα σε δύο ομάδες ως ένα μη σταθμισμένο μέσο των προηγούμενων αποστάσεών τους.

d(P+Q,R)=0,5(d(P,R)+d(Q,R))

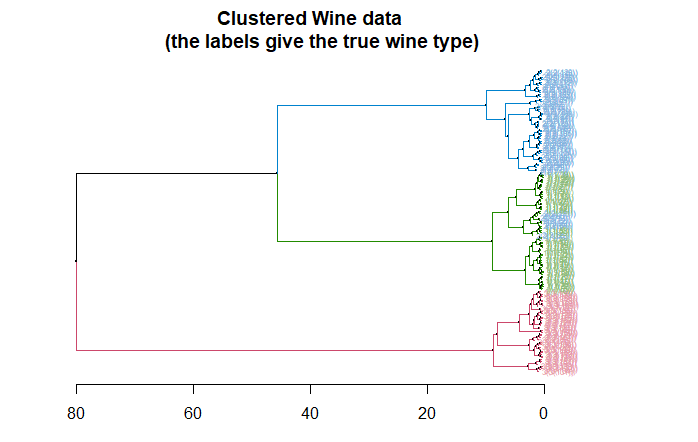
Σε κάθε περίπτωση επιλέγαμε να κρατήσουμε τρεις κλάσεις γιατι γνωρίζαμε ότι έχουμε τρεις τύπους κρασιού. Παρακάτω δίνεται η επιτυχία κάθε ορισμένων μεθόδων όπως αυτή προσμετρήθηκε από το cross-tab matrix.

|  |  |  |
| --- | --- | --- |
| Every distance vs  Single linkage | Every distance vs  Complete linkage | Every distance vs  Ward.D linkage |
| 1 2 3  1 53 71 48  2 5 0 0  3 1 0 0  ARI  0.005443835    1 2 3  1 53 71 48  2 5 0 0  3 1 0 0  ARI  0.005443835    1 2 3  1 58 70 48  2 1 0 0  3 0 1 0  ARI  -0.001686447    1 2 3  1 59 69 48  2 0 1 0  3 0 1 0  ARI  -0.003819345    1 2 3  1 53 71 48  2 5 0 0  3 1 0 0  ARI  0.005443835 | 1 2 3  1 43 0 0  2 16 15 21  3 0 56 27  ARI  0.370833    1 2 3  1 43 0 0  2 16 15 21  3 0 56 27  ARI  0.370833    1 2 3  1 43 0 0  2 16 15 21  3 0 56 27  ARI  0.370833    1 2 3  1 59 70 2  2 0 1 0  3 0 0 46  ARI  0.4595444    1 2 3  1 43 0 0  2 16 15 21  3 0 56 27  ARI  0.370833 | 1 2 3  1 53 4 5  2 6 16 22  3 0 51 21  ARI  0.402264    1 2 3  1 46 2 0  2 13 18 27  3 0 51 21  ARI  0.3684019    1 2 3  1 46 2 0  2 13 13 21  3 0 56 27  ARI  0.3909827    1 2 3  1 59 11 0  2 0 60 0  3 0 0 48  ARI  0.8142145    1 2 3  1 53 4 5  2 6 16 22  3 0 51 21  ARI  0.402264 |

Η πιο επιτυχημένη ήταν με Canberra distance και ward.D linkage και παρακάτω δίνεται το αντίστοιχο δεντρόγραμμα.



Ένα πολύ ωραίο γραφικό αποτέλεσμα είναι και το παρακάτω, στο οποίο κάθε τύπος κρασιού έχει χρωματιστει με διαφορετικό χρώμα. Μπορούμε δούμε οτι ορισμένα κρασιά τύπου 2(μπλε) τοποθετήθηκαν λανθασμένα στη κλάση των κρασιών τυπου1(πράσινα).



Επειδή η μέθοδος k-means δούλεψε αρκετά καλύτερα σε κανονικοποιημένα δεδομένα στη συνέχεια εφαρμόσαμε ιεραρχική συσταδοποίηση στα κανονικοποιημένα δεδομένα, με Canberra distance και ward linkage ώστε να κάνουμε μία σύγκριση. Τα πήγε αρκετά χειρότερα, και αυτό ήταν κάτι που δεν αναμέναμε.

