ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ στα ΠΟΛΥΠΛΟΚΑ ΣΥΣΤΗΜΑΤΑ και ΔΙΚΤΥΑ
ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ
ΤΜΗΜΑ ΒΙΟΛΟΓΙΑΣ
ΤΜΗΜΑ ΓΕΩΛΟΓΙΑΣ
ΤΜΗΜΑ ΟΙΚΟΝΟΜΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ

INTER-FACULTY MASTER PROGRAM on COMPLEX SYSTEMS and NETWORKS
SCHOOL of MATHEMATICS
SCHOOL of BIOLOGY
SCHOOL of GEOLOGY
SCHOOL of ECONOMICS
ARISTOTLE UNIVERSITY of THESSALONIKI

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

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

1.2. 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.

	Alcohol	Malic.acid	Ash	Acl	Mg	Phenols	Flavanoids	Nonflavanoid.phenols	Proanth	Color.int	Hue	OD	Proline
nbr.val	1.780000e+02	178.00000000	178.00000000	178.0000000	1.780000e+02	178.00000000	178.00000000	1.780000e+02	178.00000000	178.0000000	178.00000000	178.00000000	1.780000e+02
nbr.null	0.000000e+00	0.00000000	0.00000000	0.0000000	0.000000e+00	0.00000000	0.00000000	0.000000e+00	0.00000000	0.0000000	0.00000000	0.00000000	0.000000e+00
nbr.na	0.000000e+00	0.00000000	0.00000000	0.0000000	0.000000e+00	0.00000000	0.00000000	0.000000e+00	0.00000000	0.0000000	0.00000000	0.00000000	0.000000e+00
min	1.103000e+01	0.74000000	1.36000000	10.6000000	7.000000e+01	0.98000000	0.34000000	1.300000e-01	0.41000000	1.2800000	0.48000000	1.27000000	2.780000e+02
max	1.483000e+01	5.80000000	3.23000000	30.0000000	1.620000e+02	3.88000000	5.08000000	6.600000e-01	3.58000000	13.0000000	1.71000000	4.00000000	1.680000e+03
range	3.800000e+00	5.06000000	1.87000000	19.4000000	9.200000e+01	2.90000000	4.74000000	5.300000e-01	3.17000000	11.7200000	1.23000000	2.73000000	1.402000e+03
sum	2.314110e+03	415.87000000	421.24000000	3470.1000000	1.775400e+04	408.53000000	361.21000000	6.441000e+01	283.18000000	900.3399990	170.42600000	464.88000000	1.329470e+05
median	1.305000e+01	1.86500000	2.36000000	19.5000000	9.800000e+01	2.35500000	2.13500000	3.400000e-01	1.55500000	4.6900000	0.96500000	2.78000000	6.735000e+02
mean	1.300062e+01	2.33634831	2.36651685	19.4949438	9.974157e+01	2.29511236	2.02926966	3.618539e-01	1.59089888	5.0580899	0.95744944	2.61168539	7.468933e+02
SE.mean	6.084897e-02	0.08373364	0.02056295	0.2503109	1.070517e+00	0.04690952	0.07486762	9.328172e-03	0.04290011	0.1737629	0.01713216	0.05321603	2.360331e+01
CI.mean.0.95	1.200828e-01	0.16524476	0.04058011	0.4939778	2.112620e+00	0.09257393	0.14774805	1.840875e-02	0.08466153	0.3429136	0.03380959	0.10501956	4.658013e+01
var	6.590623e-01	1.24801540	0.07526464	11.1526862	2.039893e+02	0.39168954	0.99771867	1.548863e-02	0.32759467	5.3744494	0.05224496	0.50408641	9.916672e+04
std.dev	8.118265e-01	1.11714610	0.27434401	3.3395638	1.428248e+01	0.62585105	0.99885869	1.244533e-01	0.57235886	2.3182859	0.22857157	0.70999043	3.149075e+02
coef.var	6.244523e-02	0.47815905	0.11592734	0.1713041	1.431949e-01	0.27268863	0.49222570	3.439325e-01	0.35977074	0.4583323	0.23872965	0.27185144	4.216231e-01

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.

2. 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 $A=P\Lambda P'$ 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 $P'AP=\Lambda$ 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 $Y=AX$, 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 A 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.

	Alcohol	Malic.acid	Ash	Acl	Mg	Phenols	Flavanoids	nonFlav	Proanth	Color.int	Hue	OD	Proline
Alcohol	1.000	0.094	0.212	-0.310	0.271	0.289	0.237	-0.156	0.137	0.546	-0.072	0.072	0.644
Malic.acid	0.094	1.000	0.164	0.289	-0.055	-0.335	-0.411	0.293	-0.221	0.249	-0.561	-0.369	-0.192
Ash	0.212	0.164	1.000	0.443	0.287	0.129	0.115	0.186	0.010	0.259	-0.075	0.004	0.224
Acl	-0.310	0.289	0.443	1.000	-0.083	-0.321	-0.351	0.362	-0.197	0.019	-0.274	-0.277	-0.441
Mg	0.271	-0.055	0.287	-0.083	1.000	0.214	0.196	-0.256	0.236	0.200	0.055	0.066	0.393
Phenols	0.289	-0.335	0.129	-0.321	0.214	1.000	0.865	-0.450	0.612	-0.055	0.434	0.700	0.498
Flavanoids	0.237	-0.411	0.115	-0.351	0.196	0.865	1.000	-0.538	0.653	-0.172	0.543	0.787	0.494
NonFlav	-0.156	0.293	0.186	0.362	-0.256	-0.450	-0.538	1.000	-0.366	0.139	-0.263	-0.503	-0.311
Proanth	0.137	-0.221	0.010	-0.197	0.236	0.612	0.653	-0.366	1.000	-0.025	0.296	0.519	0.330
Color.int	0.546	0.249	0.259	0.019	0.200	-0.055	-0.172	0.139	-0.025	1.000	-0.522	-0.429	0.316
Hue	-0.072	-0.561	-0.075	-0.274	0.055	0.434	0.543	-0.263	0.296	-0.522	1.000	0.565	0.236
OD	0.072	-0.369	0.004	-0.277	0.066	0.700	0.787	-0.503	0.519	-0.429	0.565	1.000	0.313
Proline	0.644	-0.192	0.224	-0.441	0.393	0.498	0.494	-0.311	0.330	0.316	0.236	0.313	1.000

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 10^{-2} έως και 10^{4} 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:

	Alcohol	Malic.acid	Ash	Acl	Mg	Phenols	Flavanoids	Nonflavanoid.phenols	Proanth	Color.int	Hue	OD	Proline
Alcohol	NA	2.100820e-01	4.587948e-03	2.505006e-05	2.561861e-04	9.084653e-05	1.459382e-03	3.766955e-02	6.883834e-02	3.108624e-15	3.412439e-01	3.372381e-01	0.000000e+00
Malic.acid	2.100820e-01	NA	2.866487e-02	9.409871e-05	4.693530e-01	4.804246e-06	1.207022e-08	7.226419e-05	3.066544e-03	8.040012e-04	4.440892e-16	4.102859e-07	1.023967e-02
Ash	4.587948e-03	2.866487e-02	NA	5.722129e-10	1.052007e-04	8.618903e-02	1.261178e-01	1.281274e-02	8.982529e-01	4.842282e-04	3.219075e-01	9.586762e-01	2.694295e-03
Acl	2.505006e-05	9.409871e-05	5.722129e-10	NA	2.687725e-01	1.240723e-05	1.516365e-06	6.907747e-07	8.287395e-03	8.039976e-01	2.153328e-04	1.841221e-04	7.523984e-10
Mg	2.561861e-04	4.693530e-01	1.052007e-04	2.687725e-01	NA	4.055502e-03	8.816854e-03	5.540626e-04	1.485696e-03	7.451798e-03	4.626677e-01	3.813817e-01	5.586937e-08
Phenols	9.084653e-05	4.804246e-06	8.618903e-02	1.240723e-05	4.055502e-03	NA	0.000000e+00	2.960543e-10	0.000000e+00	4.647881e-01	1.474412e-09	0.000000e+00	1.497247e-12
Flavanoids	1.459382e-03	1.207022e-08	1.261178e-01	1.516365e-06	8.816854e-03	0.000000e+00	NA	9.769963e-15	0.000000e+00	2.139863e-02	4.440892e-15	0.000000e+00	2.376099e-12
Nonflavanoid.phenols	3.766955e-02	7.226419e-05	1.281274e-02	6.907747e-07	5.540626e-04	2.960543e-10	9.769963e-15	NA	5.119347e-07	6.414288e-02	3.974886e-04	8.086865e-13	2.328678e-05
Proanth	6.883834e-02	3.066544e-03	8.982529e-01	8.287395e-03	1.485696e-03	0.000000e+00	0.000000e+00	5.119347e-07	NA	7.379600e-01	6.198826e-05	1.145750e-13	6.655600e-06
Color.int	3.108624e-15	8.040012e-04	4.842282e-04	8.039976e-01	7.451798e-03	4.647881e-01	2.139863e-02	6.414288e-02	7.379600e-01	NA	8.082424e-14	2.345924e-09	1.720964e-05
Hue	3.412439e-01	4.440892e-16	3.219075e-01	2.153328e-04	4.626677e-01	1.474412e-09	4.440892e-15	3.974886e-04	6.198826e-05	8.082424e-14	NA	2.220446e-16	1.504024e-03
OD	3.372381e-01	4.102859e-07	9.586762e-01	1.841221e-04	3.813817e-01	0.000000e+00	0.000000e+00	8.086865e-13	1.145750e-13	2.345924e-09	2.220446e-16	NA	2.133107e-05
Proline	0.000000e+00	1.023967e-02	2.694295e-03	7.523984e-10	5.586937e-08	1.497247e-12	2.376099e-12	2.328678e-05	6.655600e-06	1.720964e-05	1.504024e-03	2.133107e-05	NA

Με αντίστοιχο πίνακα p_values:

1	-0.144329395	-0.483651548	-0.20738262	-0.01785630	-0.26566365	0.21353865	0.05639636	0.39613926	0.50861912	0.21160473	-0.22591696	-0.26628645	-0.01496997
2	0.245187580	-0.224930935	0.08901289	0.53689028	0.03521363	0.53681385	-0.42052391	0.06582674	-0.07528304	-0.30907994	0.07648554	0.12169604	-0.02596375
3	0.002051061	-0.316068814	0.62622390	-0.21417556	-0.14302547	0.15447466	0.14917061	-0.17026002	-0.30769445	-0.02712539	-0.49869142	-0.04962237	0.14121803
4	0.239320405	0.010590502	0.61208035	0.06085941	0.06610294	-0.10082451	0.28696914	0.42797018	0.20044931	0.05279942	0.47931378	-0.05574287	-0.09168285
5	-0.141992042	-0.299634003	0.13075693	-0.35179658	0.72704851	0.03814394	-0.32288330	-0.15636143	0.27140257	0.06787022	0.07128891	0.06222011	-0.05677422
6	-0.394660845	-0.065039512	0.14617896	0.19806835	-0.14931841	-0.08412230	0.02792498	-0.40593409	0.28603452	-0.32013135	0.30434119	-0.30388245	0.46390791
7	-0.422934297	0.003359812	0.15068190	0.15229479	-0.10902584	-0.01892002	0.06068521	-0.18724536	0.04957849	-0.16315051	-0.02569409	-0.04289883	-0.83225706
8	0.298533103	-0.028779488	0.17036816	-0.20330102	-0.50070298	-0.25859401	-0.59544729	-0.23328465	0.19550132	0.21553507	0.11689586	0.04235219	-0.11403985
9	-0.313429488	-0.039301722	0.14945431	0.39905653	0.13685982	-0.53379539	-0.37213935	0.36822675	-0.20914487	0.13418390	-0.23736257	-0.09555303	0.11691707
10	0.088616705	-0.529995672	-0.13730621	0.06592568	-0.07643678	-0.41864414	0.22771214	-0.03379692	0.05621752	-0.29077518	0.03183880	0.60422163	0.01199280
11	-0.296714564	0.279235148	0.08522192	-0.42777141	-0.17361452	0.10598274	-0.23207564	0.43662362	0.08582839	-0.52239889	-0.04821201	0.25921400	0.08988884
12	-0.376167411	0.164496193	0.16600459	0.18412074	-0.10116099	0.26585107	0.04476370	-0.07810789	0.13722690	0.52370587	0.04642330	0.60095872	0.15671813
13	-0.286752227	-0.364902832	-0.12674592	-0.23207086	-0.15786880	0.11972557	-0.07680450	0.12002267	-0.57578611	0.16211600	0.53926983	-0.07940162	-0.01444734

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 πραγματοποίησης 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
```

Η p.value του τέστ είναι πολύ μικρή και <0,05 άρα μπορούμε 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 $\pi\lambda\epsilon\acute{\iota}\theta o\varsigma\ \delta\epsilon\iota\gamma\mu\acute{\alpha}\tau\omega\nu \Big/ \pi\lambda\acute{\eta}\theta o\varsigma\ \mu\epsilon\tau\alpha\beta\lambda\eta\tau\acute{\omega}\nu$ >5 και στα συγκεκριμένα δεδομένα το λόγος αυτός προκείπτει 13.6923 οπότε δεν εμπιστευόμαστε απόλυτα αυτό το συγκεκριμένο τέστ.

- ❖ KMO 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.

- o 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
```

- o Και των αντίστοιχων ιδιοδιανυσμάτων

Δηλαδή η πρώτη συνιστώσα είναι:
$Y_1 = -0.144X_1 + 0.2451X_2 + 0.002X_3 + \ldots$ με Var(Y_1)=λ_1=4.7059 κλπ

- o 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

Επιλέγουμε συνιστώσες όσες και οι ιδιοτιμές που είναι μεγαλυτερες από τη μέση τιμή των ιδιοτιμών; $\bar{\lambda}$. Στο πίνακα συσχετίσεων επειδή $\bar{\lambda}$=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
```

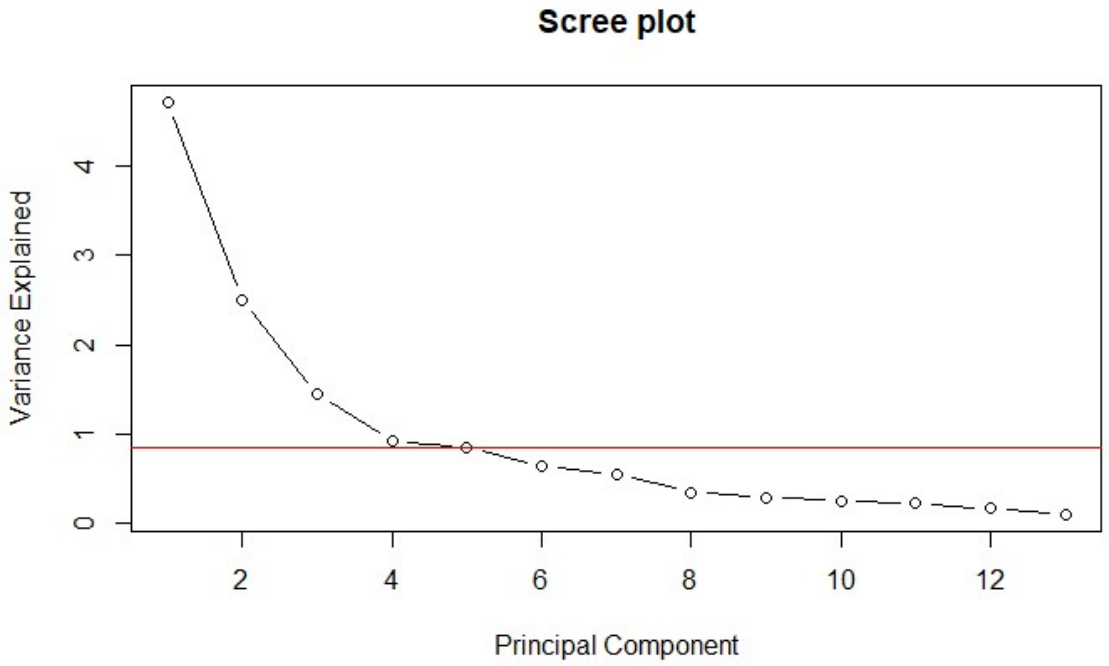
ποσοστό της ματαβλητότητας. Επομένως αθροιστικά εξηγεί

```
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.

- Εύρεση κυρίων συνιστωσών

Οι κύριες συνιστώσες λοιπόν που διατηρήσαμε είναι οι παρακάτω:

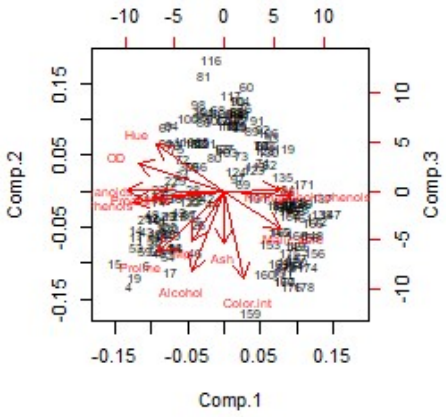
PCA loadings

	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

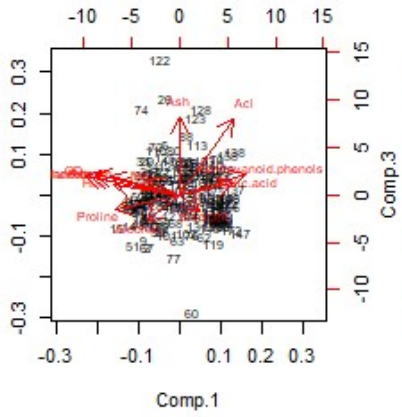
PCA scores, για τις πρώτες 10 παρατηρήσεις

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6
1	-3.316751	-1.4434626	-0.1657390	-0.21563119	0.69304284	0.2238801
2	-2.209465	0.3333929	-2.0264574	-0.29135832	-0.25765463	0.9271202
3	-2.516740	-1.0311513	0.9828187	0.72490231	-0.25103312	-0.5492760
4	-3.757066	-2.7563719	-0.1761918	0.56798331	-0.31184159	-0.1144310
5	-1.008908	-0.8698308	2.0266882	-0.40976579	0.29845750	0.4065196
6	-3.050254	-2.1224011	-0.6293958	-0.51563749	-0.63201873	-0.1234306
7	-2.449090	-1.1748501	-0.9770949	-0.06583050	-1.02776191	0.6201207
8	-2.059437	-1.6089631	0.1462819	-1.19260801	0.07690349	1.4398062
9	-2.510874	-0.9180710	-1.7709690	0.05627036	-0.89225698	0.1291810
10	-2.753628	-0.7894377	-0.9842475	0.34938157	-0.46855308	-0.1633917

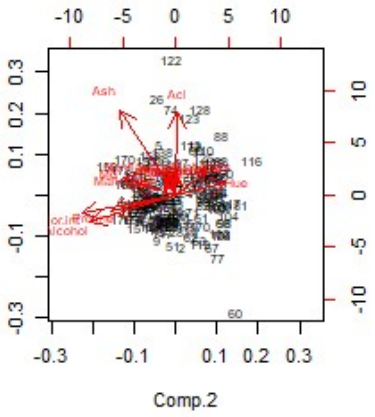
Ένα επιπλέον χρήσιμο εργαλείο για να δούμε την ομαδοποίηση των δεδομένων είναι τα biplots με άξονες τις πρώτες κύριες συνιστώσες. Στα biplots μπορούμε να δούμε σε μορφή διανύσματος κάθε μεταβλητή και να δούμε πόσο «κοντά» είναι η μία στην άλλη. Επιπροσθέτως μπορούμε να παρατηρήσουμε πόσο καλυτερη διασπορά 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%

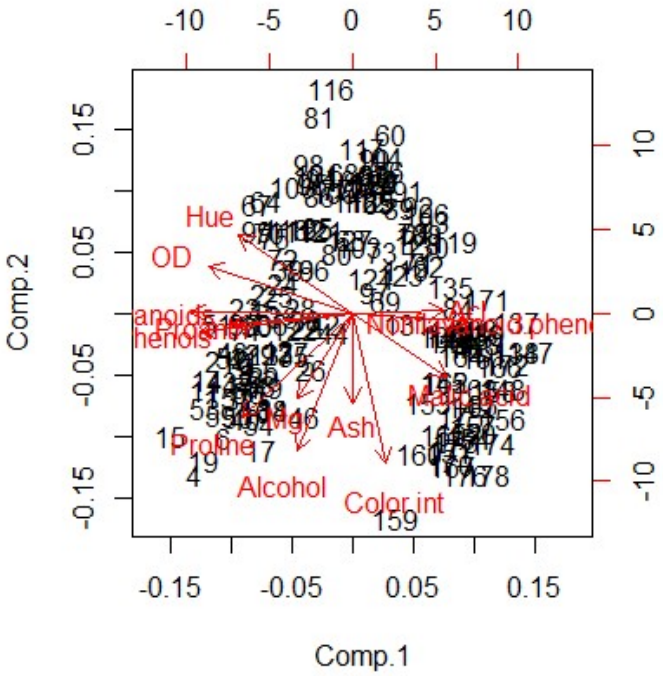


47.31%

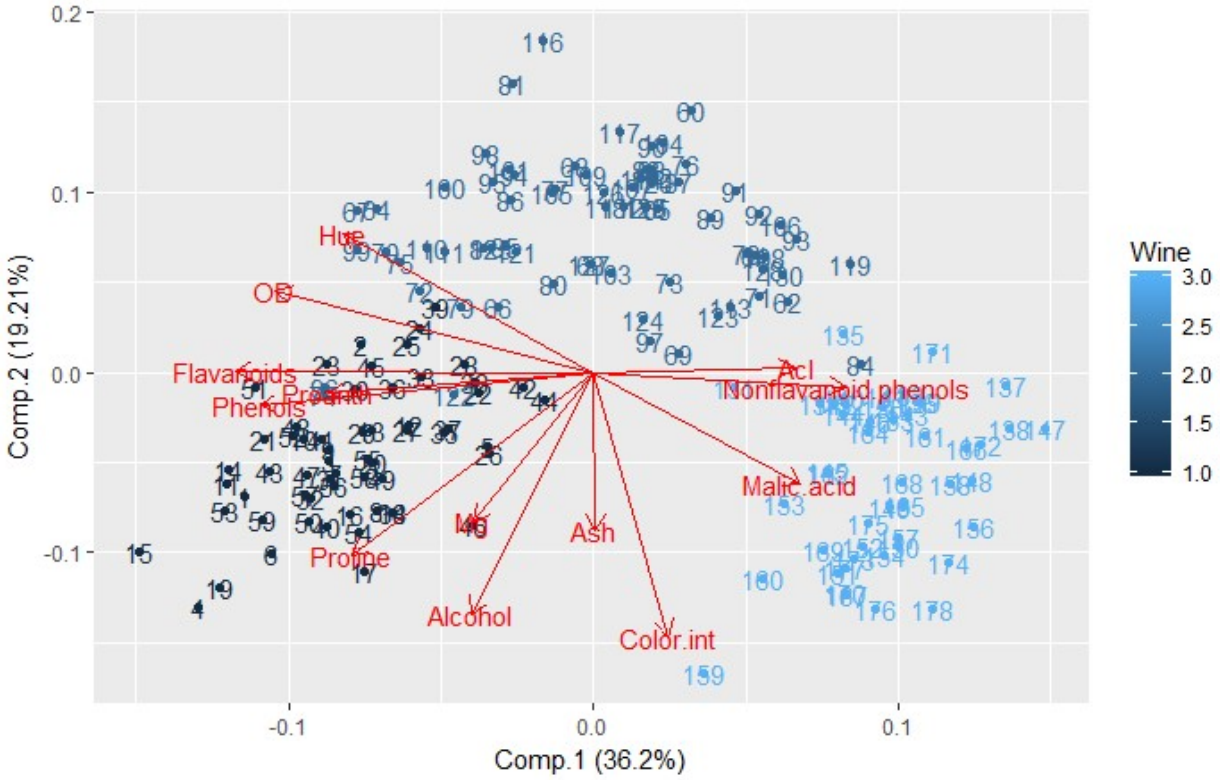


30.33%

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.



3. 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)=$\sum_{i=1}^{n}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{kj})^2$, όπου

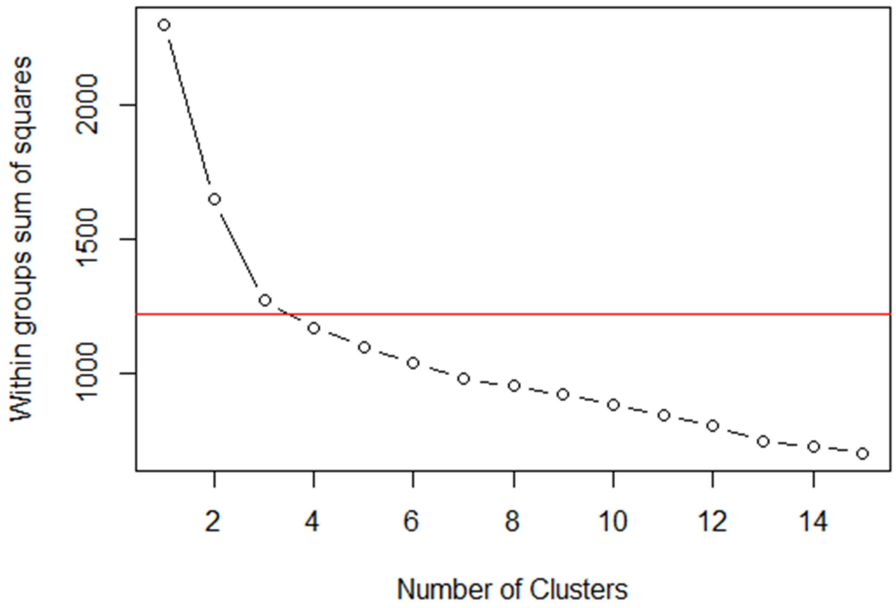
k είναι ο αριθμός των κλάσεων

x_{ij} η τιμής της j-μεταβλητής της i-παρατήρησης

$\bar{x}_{kj}$ 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 10^{-1} για φαινόλες και του 10^{3} για προλίνη), πρωτού προχωρήσουμε στον αλγόριθμο, έγινε αρχικά κανονικοποίηση των δεδομένων έτσι ώστε αυτά να ακολουθούν την κανονική κατανομή N(0,1).

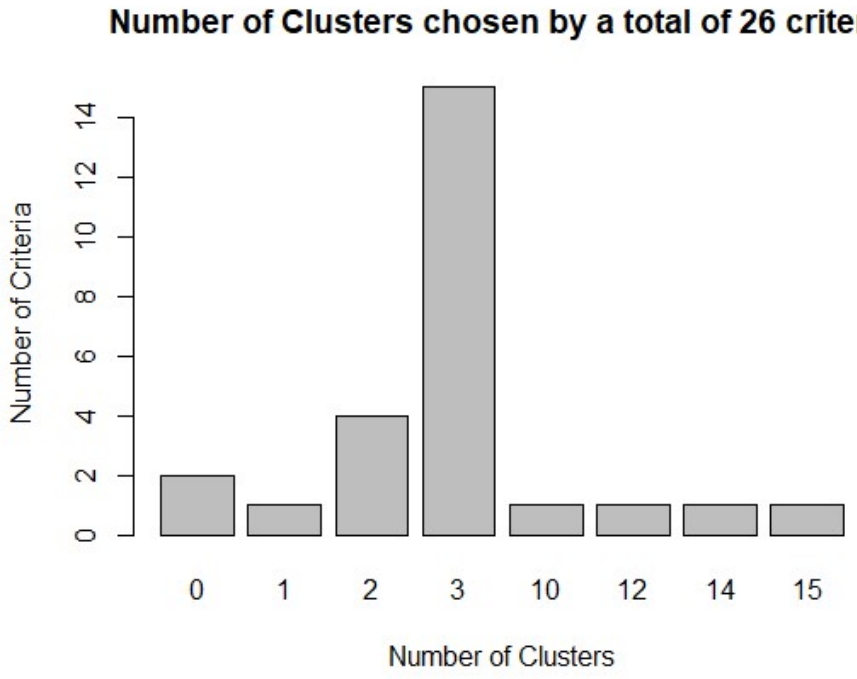
Ο αλγοριθμός προυποθέτει να επιλέξουμε το πλήθος των κέντρων, γι αυτό κατασκευάσαμε το παρακάτω γράφημα που μας δίνει το άθροισμα των τετραγώνων των αποστάσεων από το κέντρο κάθε κλάσης όπου έχει τοποθετηθεί σε σχεση με το πλήθος των κλάσεων.



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:

	Alcohol	Malic.acid	Ash	Acl	Mg	Phenols	Flavanoids	NonFlav	Proanth	Color.int	Hue	OD	Proline
1	0.8328826	-0.3029551	0.3636801	-0.6084749	0.57596208	0.88274724	0.97506900	-0.56050853	0.57865427	0.1705823	0.4726504	0.7770551	1.1220202
2	-0.9234669	-0.3929331	-0.4931257	0.1701220	-0.49032869	-0.07576891	0.02075402	-0.03343924	0.05810161	-0.8993770	0.4605046	0.2700025	-0.7517257
3	0.1644436	0.8690954	0.1863726	0.5228924	-0.07526047	-0.97657548	-1.21182921	0.72402116	-0.77751312	0.9388902	-1.1615122	-1.2887761	-0.4059428

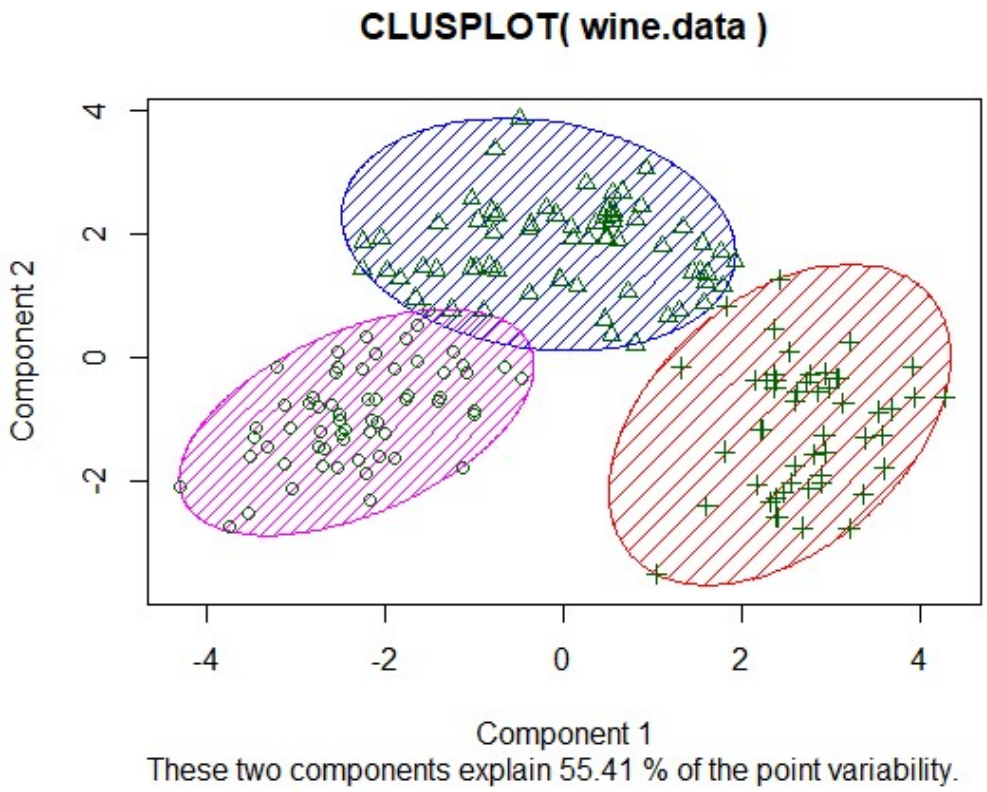
Επειδή τα παραπάνω κέντρα είναι υπολογισμένα στα κανονικοποιημένα δεδομένα έγινε μετασχηματισμός τους και τα κέντρα στα αρχικά δεδομένα είναι:

	cluster	Alcohol	Malic.acid	Ash	Acl	Mg	Phenols	Flavanoids	Nonflav.	Proanth	Color.int	Hue	OD	Proline
1	1	13.67677	1.997903	2.466290	17.46290	107.96774	2.847581	3.0032258	0.2920968	1.922097	5.453548	1.0654839	3.163387	1100.2258
2	2	12.25092	1.897385	2.231231	20.06308	92.73846	2.247692	2.0500000	0.3576923	1.624154	2.973077	1.0627077	2.803385	510.1692
3	3	13.13412	3.307255	2.417647	21.24118	98.66667	1.683922	0.8188235	0.4519608	1.145882	7.234706	0.6919608	1.696667	619.0588

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:

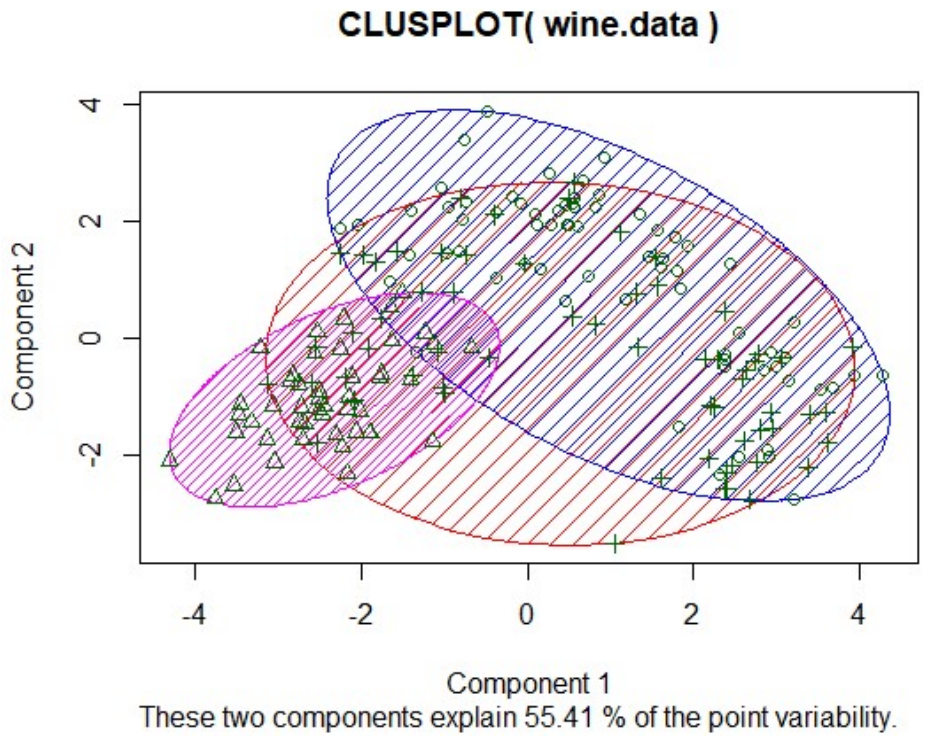
Τοποθέτησε 69, 47 και 62 παρατηρήσεις σε κάθε κλάση

Τα κέντρα ήταν :

	Alcohol	Malic.acid	Ash	Acl	Mg	Phenols	Flavanoids	Nonflav.	Proanth	Color.int	Hue	OD	Proline
1	12.51667	2.494203	2.288551	20.82319	92.34783	2.070725	1.758406	0.3901449	1.451884	4.086957	0.9411594	2.490725	458.2319
2	13.80447	1.883404	2.426170	17.02340	105.51064	2.867234	3.014255	0.2853191	1.910426	5.702553	1.0782979	3.114043	1195.1489
3	12.92984	2.504032	2.408065	19.89032	103.59677	2.111129	1.584032	0.3883871	1.503387	5.650323	0.8839677	2.365484	728.3387

```
    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
  $D_{Euclidean}(p,q) := \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$
- Chebyshev/ Maximum distance
  $D_{Chebyshev}(p,q) := max_i(|p_i - q_i|)$ i=1,2,..,n
- Manhattan
  $D_{Manhattan}(p,q) = ||p - q||_1 = \sum_{i=1}^{n} |p_i - q_i|$
- Canberra
  $D_{Canberra}(p,q) = \sum_{i=1}^{n} \frac{|p_i - q_i|}{|p_i| + |q_i|}$
- Minkowski
  $D_{Minkowski}(p,q) = (\sum_{i=1}^{n} |p_i - q_i|^p)^{1/p}$, για p=3,4,5

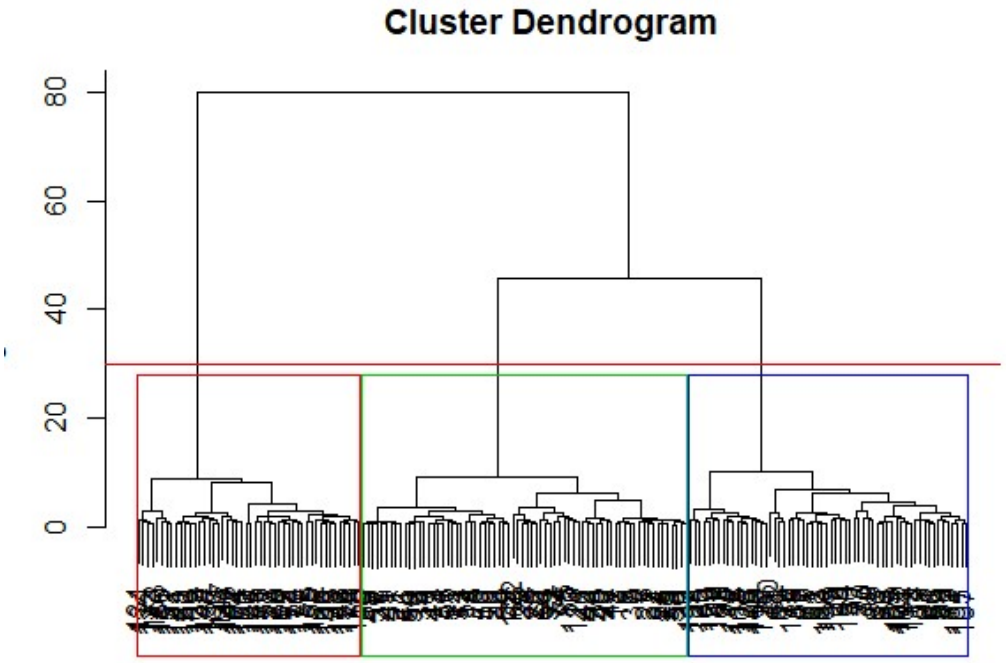
Ακόμα σημαντικό ρόλο παίζει η επιλογή της μεθόδου συσσωμάτωσης (τρόπος υπολογισμού απόστασης μεταξύ δύο ομάδων). Έγινε ανά δύο συνδυασμός των παραπάνω αποστάσεων με τις παρακάτω linkage μεθόδους:

- 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 P κλάση και NQ 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))

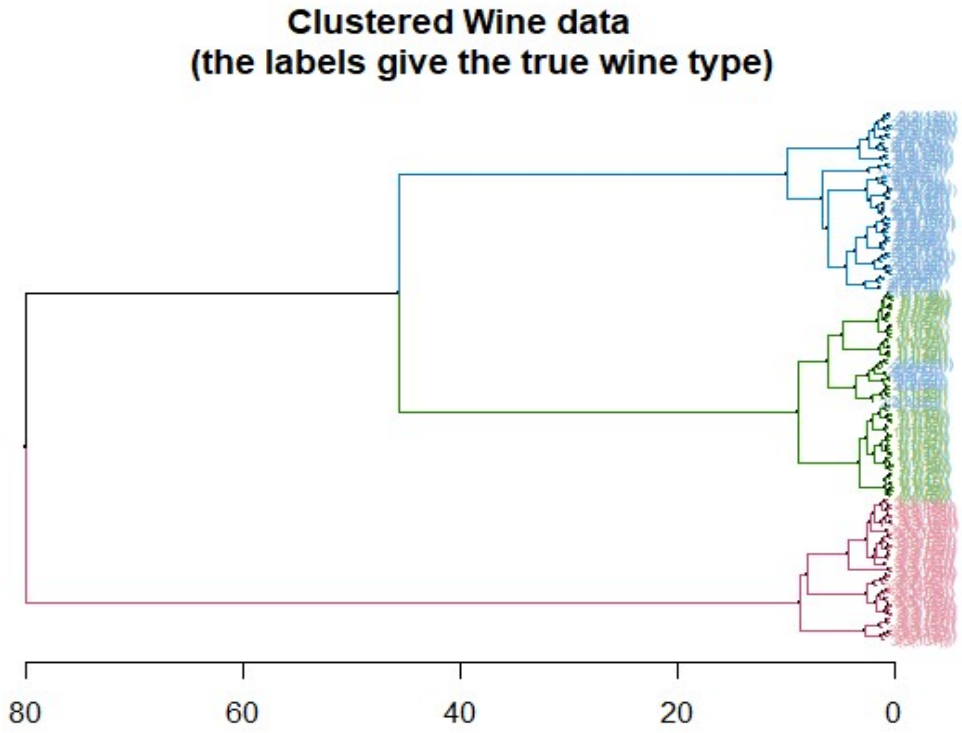
Σε κάθε περίπτωση επιλέγαμε να κρατήσουμε τρεις κλάσεις γιατι γνωρίζαμε ότι έχουμε τρεις τύπους κρασιού. Παρακάτω δίνεται η επιτυχία κάθε ορισμένων μεθόδων όπως αυτή προσμετρήθηκε από το cross-tab matrix.

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.