Ανάλυση Κύριων Συνιστωσών (PCA) και Συσταδοποίηση (Clustering) Δεδομένων

# Δεδομένα

Αρχικά, επιλέχθηκε ως σύνολο δεδομένων το UScrime της βιβλιοθήκης MASS της R το οποίο δημιουργήθηκε από εγκληματολόγους για να εξεταστεί η αποδοτικότητα των ποινών στα επίπεδα εγκληματικότητας. Περιλαμβάνει συγκεντρωτικά δεδομένα από 47 πολιτείες των ΗΠΑ το 1960 και οι μεταβλητές κανονικοποιήθηκαν βάση βιβλιογραφίας σε βέλτιστους αριθμούς για κάθε μεταβλητή.

Το σύνολο δεδομένων αποτελείται από 47 δείγματα(τις Πολιτείες) 16 μεταβλητών και οι μεταβλητές περιγράφονται ως:

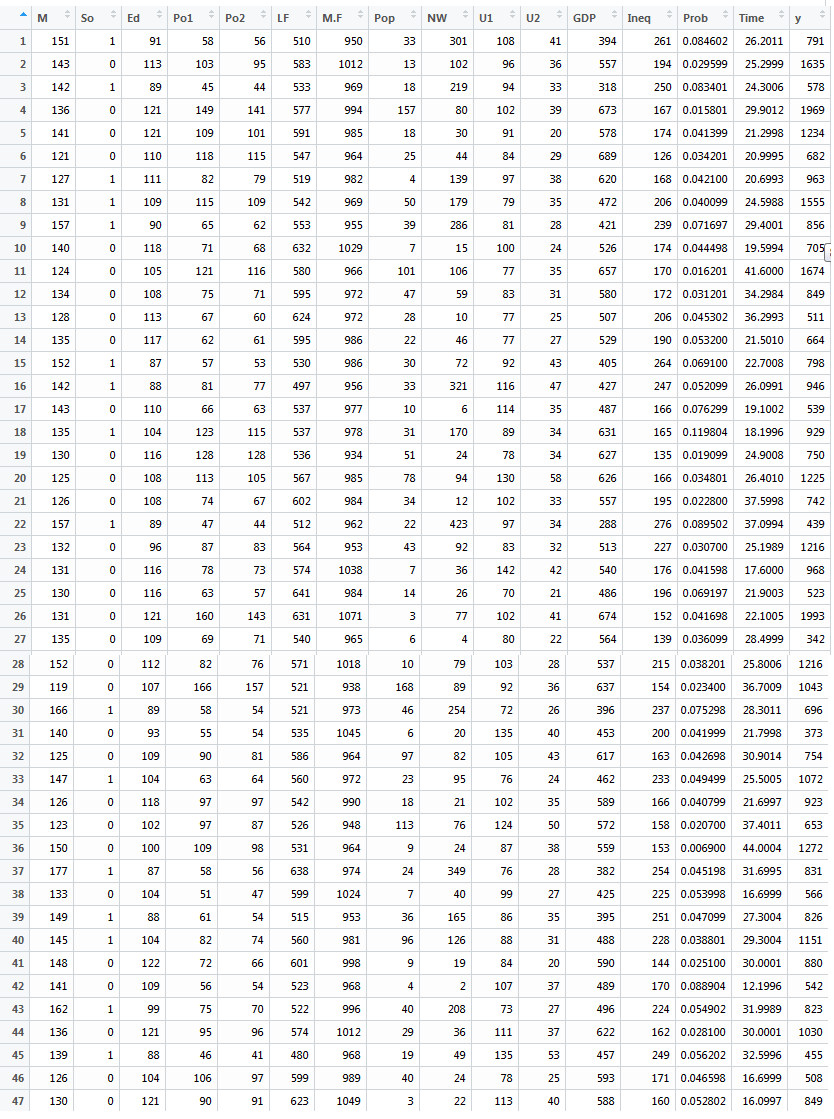
* **M :** Ποσοστό ανδρών ηλικίας 14–24. (Αριθμητική Μεταβλητή)
* **So:**λογική μεταβλητή για τις δυτικές πολιτείες (τιμές 0,1)
* **Ed:**Μ.Ο. ετών εκπαίδευσης. (Αριθμητική Μεταβλητή)
* **Po1:**δαπάνες αστυνομίας το 1960. (Αριθμητική Μεταβλητή)
* **Po2:**δαπάνες αστυνομίας το 1959. (Αριθμητική Μεταβλητή)
* **LF:**ποσοστό συμμετοχής του εργατικού δυναμικού. (Αριθμητική Μεταβλητή)
* **M.F**: πλήθος ανδρών ανά 1000 γυναίκες. (Αριθμητική Μεταβλητή)
* **Pop:**πληθυσμός Πολιτείας. (Αριθμητική Μεταβλητή)
* **NW:**πλήθος έγχρωμων ανά 100 πολίτες. (Αριθμητική Μεταβλητή)
* **U1**: ποσοστό ανεργίας σε άνδρες ηλικίας 14-24. (Αριθμητική Μεταβλητή)
* **U2:** ποσοστά ανεργίας σε άνδρες ηλικίας 35-39. (Αριθμητική Μεταβλητή)
* **GDP:**οικιακό κατά κεφαλήν εισόδημα. (Αριθμητική Μεταβλητή)
* **Ineq:** εισοδηματικήανισότητα. (Αριθμητική Μεταβλητή)
* **Prob:** πιθανότητα φυλάκισης. (Αριθμητική Μεταβλητή)
* **Time:** Μ.Ο. έκτισης ποινών σε φυλακή. (Αριθμητική Μεταβλητή)
* **Y:**ποσοστό εγκλημάτων σε συγκεκριμένη κατηγορία εγκλήματος ανά πολίτη του πληθυσμού της πολιτείας (Αριθμητική Μεταβλητή)

Πηγές:

*Ehrlich, I. (1973) Participation in illegitimate activities: a theoretical and empirical investigation. Journal of Political Economy, 81, 521–565.*

*Vandaele, W. (1978) Participation in illegitimate activities: Ehrlich revisited. In Deterrence and Incapacitation, eds A. Blumstein, J. Cohen and D. Nagin, pp. 270–335. US NationalAcademy of Sciences.*

Το σύνολο δεδομένων όπως αποθηκεύτηκε στην R:

**

# Περιγραφική Στατιστική των Δεδομένων

Με την εντολή stats.desc της βιβλιοθήκης pastects στην R

**M So Ed Po1 Po2**

min 1.190000e+02 0.00000000 87.0000000 45.000000 41.0000000

max 1.770000e+02 1.00000000 122.0000000 166.000000 157.0000000

range 5.800000e+01 1.00000000 35.0000000 121.000000 116.0000000

sum 6.513000e+03 16.00000000 4965.0000000 3995.000000 3771.0000000

median 1.360000e+02 0.00000000 108.0000000 78.000000 73.0000000

mean 1.385745e+02 0.34042553 105.6382979 85.000000 80.2340426

SE.mean 1.833178e+00 0.06986571 1.6317914 4.334958 4.0785775

CI.mean.0.95 3.689996e+00 0.14063238 3.2846258 8.725818 8.2097507

var 1.579454e+02 0.22941721 125.1489362 883.217391 781.8353377

std.dev 1.256763e+01 0.47897516 11.1869985 29.718974 27.9613186

coef.var 9.069228e-02 1.40698954 0.1058991 0.349635 0.3484969

**LF M.F Pop NW U1**

min 4.800000e+02 9.340000e+02 3.000000 2.000000 70.0000000

max 6.410000e+02 1.071000e+03 168.000000 423.000000 142.0000000

range 1.610000e+02 1.370000e+02 165.000000 421.000000 72.0000000

sum 2.637600e+04 4.620200e+04 1721.000000 4753.000000 4487.0000000

median 5.600000e+02 9.770000e+02 25.000000 76.000000 92.0000000

mean 5.611915e+02 9.830213e+02 36.617021 101.127660 95.4680851

SE.mean 5.894669e+00 4.298257e+00 5.553254 14.999125 2.6297682

CI.mean.0.95 1.186535e+01 8.651943e+00 11.178120 30.191672 5.2934489

var 1.633115e+03 8.683256e+02 1449.415356 10573.765957 325.0370028

std.dev 4.041181e+01 2.946737e+01 38.071188 102.828819 18.0287826

coef.var 7.201074e-02 2.997633e-02 1.039713 1.016822 0.1888462

**U2 GDP Ineq Prob Time**

Min 20.0000000 2.880000e+02 126.0000000 6.900000e-03 12.199600

Max 58.0000000 6.890000e+02 276.0000000 1.198040e-01 44.000400

Range 38.0000000 4.010000e+02 150.0000000 1.129040e-01 31.8008000

sum 1597.0000 2.469300e+04 9118.0000000 2.213295e+00 1250.102300

median 34.0000000 5.370000e+02 176.0000000 4.210000e-02 25.8006000

mean 33.9787234 5.253830e+02 194.0000000 4.709138e-02 26.5979213

SE.mean 1.2318955 1.407465e+01 5.8194385 3.316528e-03 1.0337299

CI.mean 2.4796770 2.833080e+01 11.7139222 6.675825e-03 2.0807904

Var 71.3256244 9.310502e+03 1591.6956522 5.169699e-04 50.2240835

std.dev 8.4454499 9.649094e+01 39.8960606 2.273697e-02 7.0868952

coef.var 0.2485511 1.836583e-01 0.2056498 4.828266e-01 0.2664455

**y**

min 3.420000e+02

max 1.993000e+03

range 1.651000e+03

sum 4.253900e+04

median 8.310000e+02

mean 9.050851e+02

SE.mean 5.641514e+01

CI.mean.0.95 1.135578e+02

var 1.495854e+05

std.dev 3.867627e+02

coef.var 4.273219e-01

Σχόλια: Αρχικά η μεγάλη διαφοροποίηση στους Μέσους Όρους των μεταβλητών είναι ένδειξη ότι η ομαδοποίηση σε αυτά τα δεδομένα είναι εφικτή.

# PCA

Αρχικά θα γίνει παραγοντική ανάλυση των δεδομένων, κι συγκεκριμένα Ανάλυση Κύριων Συνιστωσών. Για την ανάλυση Κύριων Συνιστωσών των δεδομένων χρησιμοποιήθηκε η εντολή Prcomp της βιβλιοθήκης stats στην R . Η εκτίμηση έγινε με βάση τον πίνακα συσχετίσεων και όχι τον πίνακα διακυμάνσεων των μεταβλητών.Να σημειωθεί ότι η συγκεκριμένη εντολή δίνει την δυνατότητα χρήσης και απευθείας του πίνακα των δεδομενων, στον οποίο εφαρμόζεται Singular Value Decomposition με στόχο την εύρεση των Γενικευμέων Ιδιοδιανυσμάτων. Επιλέγοντας ωστόσο τον πίνακα διασυσχέτισης των δεδομένων απαλλασσόμαστε από τη μονάδα μέτρησης και το μέτρο των μεταβλητών.

Επιπλέον, διενεργήθηκε KMO έλεγχος (εντολή KMO της βιβλιοθήκης psych της R) στον πίνακα διασυσχέτισης, ο οποίος έδωσε συνολικό MSA 0,72, επομένως το ποσοστό ψευδών διασυσχετήσεων δεν κρίνεται ικανό να δημιουργήσει πρόβλημα. Τέλος τα δεδομένα μας κανονικοποιήθηκαν σε κλιμακα (0,1) και κεντροποιήθηκαν (αφαιρέθηκαν οι Μ.Ο. ανά στήλη). Τα αποτελέσματα ήταν:

**Cross-correlationmatrix:**(Εντολήcor() τηςβιβλιοθήκηςstatsστηνR)

**M So Ed Po1 Po2 LF**

M 1.00000000 0.58435534 -0.53023964 -0.50573690 -0.51317336 -0.1609488

So 0.58435534 1.00000000 -0.70274132 -0.37263633 -0.37616753 -0.5054695

Ed -0.53023964 -0.70274132 1.00000000 0.48295213 0.49940958 0.5611780

Po1 -0.50573690 -0.37263633 0.48295213 1.00000000 0.99358648 0.1214932

Po2 -0.51317336 -0.37616753 0.49940958 0.99358648 1.00000000 0.1063496

LF -0.16094882 -0.50546948 0.56117795 0.12149320 0.10634960 1.0000000

M.F -0.02867993 -0.31473291 0.43691492 0.03376027 0.02284250 0.5135588

Pop -0.28063762 -0.04991832 -0.01722740 0.52628358 0.51378940 -0.1236722

NW 0.59319826 0.76710262 -0.66488190 -0.21370878 -0.21876821 -0.3412144

U1 -0.22438060 -0.17241931 0.01810345 -0.04369761 -0.05171199 -0.2293997

U2 -0.24484339 0.07169289 -0.21568155 0.18509304 0.16922422 -0.4207625

GDP -0.67005506 -0.63694543 0.73599704 0.78722528 0.79426205 0.2946323

Ineq 0.63921138 0.73718106 -0.76865789 -0.63050025 -0.64815183 -0.2698865

Prob 0.36111641 0.53086199 -0.38992286 -0.47324704 -0.47302729 -0.2500861

Time 0.11451072 0.06681283 -0.25397355 0.10335774 0.07562665 -0.1236404

y -0.08947240 -0.09063696 0.32283487 0.68760446 0.66671414 0.1888663

**M.F Pop NW U1 U2 GDP**

M -0.02867993 -0.28063762 0.59319826 -0.224380599 -0.24484339 -0.6700550

So -0.31473291 -0.04991832 0.76710262 -0.172419305 0.07169289 -0.6369456

Ed 0.43691492 -0.01722740 -0.66488190 0.018103454 -0.21568155 0.7359970

Po1 0.03376027 0.52628358 -0.21370878 -0.043697608 0.18509304 0.7872252

Po2 0.02284250 0.51378940 -0.21876821 -0.051711989 0.16922422 0.7942620

LF 0.51355879 -0.12367222 -0.34121444 -0.229399684 -0.42076249 0.2946323

M.F 1.00000000 -0.41062750 -0.32730454 0.351891900 -0.01869169 0.1796086

Pop -0.41062750 1.00000000 0.09515301 -0.038119948 0.27042159 0.3082627

NW -0.32730454 0.09515301 1.00000000 -0.156450020 0.08090829 -0.5901070

U1 0.35189190 -0.03811995 -0.15645002 1.000000000 0.74592482 0.0448572

U2 -0.01869169 0.27042159 0.08090829 0.745924815 1.00000000 0.0920716

GDP 0.17960864 0.30826271 -0.59010707 0.044857202 0.09207166 1.0000000

Ineq -0.16708869 -0.12629357 0.67731286 -0.063832178 0.01567818 -0.8839972

Prob -0.05085826 -0.34728906 0.42805915 -0.007469032 -0.06159247 -0.5553347

Time -0.42769738 0.46421046 0.23039841 -0.169852838 0.10135833 0.0006485

y 0.21391426 0.33747406 0.03259884 -0.050477918 0.17732065 0.4413199

**Ineq Prob Time y**

M 0.63921138 0.361116408 0.1145107190 -0.08947240

So 0.73718106 0.530861993 0.0668128312 -0.09063696

Ed -0.76865789 -0.389922862 -0.2539735471 0.32283487

Po1 -0.63050025 -0.473247036 0.1033577449 0.68760446

Po2 -0.64815183 -0.473027293 0.0756266536 0.66671414

LF -0.26988646 -0.250086098 -0.1236404364 0.18886635

M.F -0.16708869 -0.050858258 -0.4276973791 0.21391426

Pop -0.12629357 -0.347289063 0.4642104596 0.33747406

NW 0.67731286 0.428059153 0.2303984071 0.03259884

U1 -0.06383218 -0.007469032 -0.1698528383 -0.05047792

U2 0.01567818 -0.061592474 0.1013583270 0.17732065

GDP -0.88399728 -0.555334708 0.0006485587 0.44131995

Ineq 1.00000000 0.465321920 0.1018228182 -0.17902373

Prob 0.46532192 1.000000000 -0.4362462614 -0.42742219

Time 0.10182282 -0.436246261 1.0000000000 0.14986606

y -0.17902373 -0.427422188 0.1498660617 1.00000000

**Ιδιοτιμές του Πίνακα:**

(Εντολέςeigen() kai eigen$values )

0.242496934

0.221238989

0.192335693

0.174532062

0.084956874

0.067928691

0.047578814

0.004335367

6.222199347

2.928000127

2.018772756

1.430068772

0.990846060

0.563802258

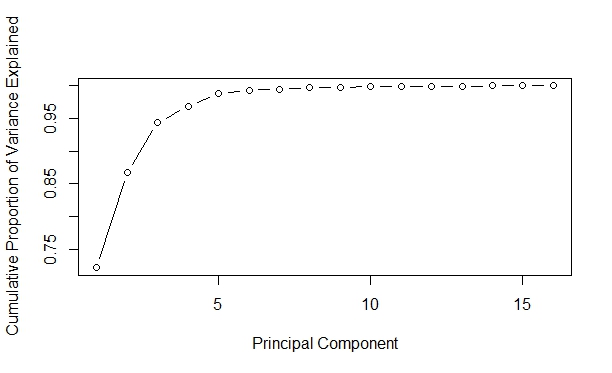
0.362852350

0.308054904

Σχόλιο:Με βάση το κριτήριο του Kaiser πρέπει να διαλέξουμε 4 συνιστώσες, αφού μόνο 4 από τις ιδιοτιμές του πίνακα συσχετίσεων είναι μεγαλύτερη του μέσου όρου όλων των ιδιοτιμών(της μονάδας δηλαδή).

Στο ακόλουθο scree plot βλέπουμε το ποσοστό της διακύμανσης % των δεδομένων που κρατά κάθε συνιστώσα.

Εντολή plot(cumsum(prop\_varex), xlab = "Principal Component", ylab = "Cumulative Proportion of Variance Explained", type = "b") από βιβλιοθήκη ggplot2 στηνR



Σχόλιο: Βλέπουμε ότι οι 4 συνιστώσες που κρατάμε βάση του κανόνα του Kaiser, ερμηνεύεται το 94% της διασποράς των αρχικών μεταβλητών. Επομένως στη συνέχεια θα ασχοληθούμε με αυτές τις 4 συνιστώσες.

### Importance :

**PC1 PC2 PC3 PC4**

**Standard deviation** 1.5941 0.7161 0.5198 0.3003

**Proportion of Variance** 0.7214 0.1456 0.0767 0.0256

**Cumulative Proportion** 0.7214 0.8670 0.9436 0.9693

## Ο πίνακας που χρησιμοποιήθηκε για την περιστροφή:

**PC1 PC2 PC3 PC4**

**M**  -0.29474638 0.04431255 0.223686369 0.020609085

**So**  -0.32216436 -0.16195008 0.037818746 -0.293430348

**Ed** 0.33332039 0.22432179 0.089574047 -0.052260667

**Po1** 0.30910124 -0.24670300 0.049290034 -0.305649199

**Po2** 0.31122378 -0.24227882 0.049384744 -0.325880495

**LF** 0.17258254 0.31793787 0.324795554 0.211518710

**M.F** 0.11164528 0.43035313 -0.101427831 -0.007217008

**Pop** 0.11318393 -0.44648894 0.023393054 0.093966631

**NW** -0.28542568 -0.22167636 0.105652933 -0.203726480

**U1** 0.03119222 0.07620120 -0.642449483 0.152589600

**U2** 0.01464476 -0.21686661 -0.584480131 0.036641843

**GDP**  0.37382180 -0.05738238 -0.006992909 -0.092339221

**Ineq**  -0.356819 -0.03820110 0.032558802 0.062964036

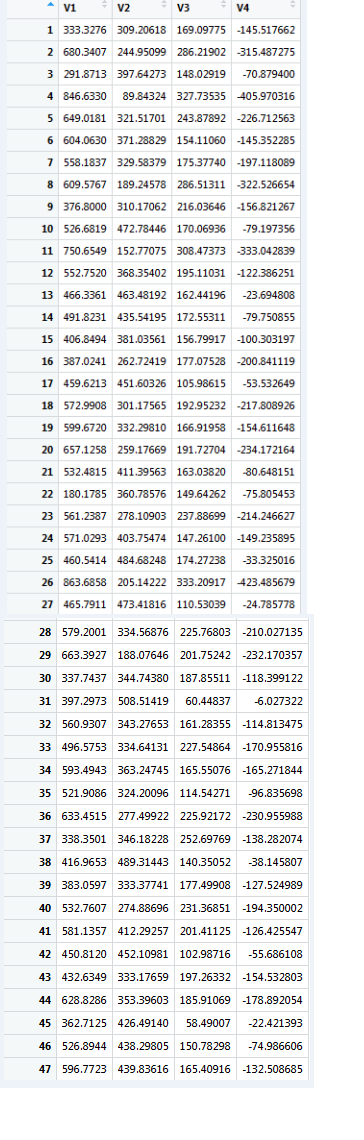
**Prob** -0.261027 0.17113604 -0.086482918 -0.455657274

**Time** -0.016662 - 0.38821748 0.168224546 0.573214732

**y** 0.18065678 -0.17191760 0.130025444 -0.206593418

Σχόλια:Σε κάθε παράγοντα χρωματίσαμε τις μεταβλητές που παρουσιάζουν μεγάλη επίδραση σε αυτούς. Παρατηρούμε να υπάρχει μια ομαδοποίηση των μεταβλητών σε κάθε παράγοντα:

* Ο 1ος παράγοντας φαίνεται να επηρεάζεται από μεταβλητές που αφορούν την εκπαίδευση και οικονομικά κριτήρια, όπως οι δαπάνες της αστυνομίας, το κατα κεφαλήν εισόδημα και την εισοδηματική ανισότητα, αλλά και το αν είναι δυτική πολιτεία.
* Ο 2ος παράγοντας φαίνεται να επηρεάζεται από μεταβλητές που αφορούν την συνμμετοχή του εργατικού κινήματος και την παρουσία ανδρών στον πληθυσμό της πολιτεία, αλλά και τον χρόνο εξέκτισης ποινής φυλάκισης ανά πολιτεία.
* Ενώ ο 3οςπαράγοντας φαίνεται να επηρεάζεται και αυτός από την ενεργή δράση του εργατικού κινήματος στην πολιτεία, αλλά και τα ποσοστά ανεργίας.
* Τέλος, ο 4ος παράγοντας δέχεται τις εντονότερες επιρροές από τις αστυνομικές δαπάνες αλλά και την εισοδηματική ανισότητα και την πιθανότητα φυλάκισης ανά πολιτεία.

Και το νέο σύνολο δεδομένων ποι προκύπτει μετά την περσιστροφή:

# Hierarchical Clustering

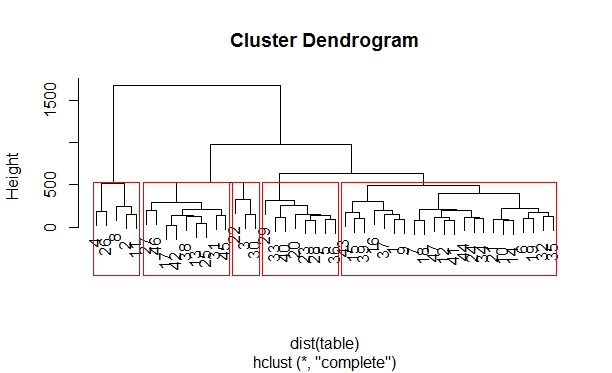
Χρησιμοποιήθηκε η εντολή Hclust της βιβλιοθήκης stats στην R η οποία πραγματοποιεί ιεραρχική ομαδοποίηση στου συνόλου δεδομένων, ξεκινώντας από 47 ομάδες, κάθε πολιτεία δηλαδή διαμορφώνει και μία ομάδα και με έναν επαναληπτικό αλγόριθμο κάθε στάδιο συνδέει τις δύο πιο όμοιες ομάδες μέχρι να ενoποιηθούν σε μία ομάδα και οι 47 πολιτείες. Σε κάθε στάδιο οι αποστάσεις μεταξύ των ομάδων υπολογίζονται με την μέθοδο Lance–William και ξεκινάμε θεωρώντας κάθε πολιτεία μία κλάση, και ως μετρική ορίσθηκε η Ευκλείδεια απόσταση .

Επιλεγμένος αλγόριθμος στον οποίο προσαρμόζεται η άνωθεν μέθοδος είναι η μέθοδος του complete-linkage, όπου ορίζεται η απόσταση μεταξύ δύο συστάδων ως η μέγιστη απόσταση μεταξύ των σημείων τους.

Ο παρακάτω αλγόριθμος είναι ένα συσσωρευτικό σχήμα που διαγράφει στήλες και γραμμές σε έναν πίνακα εγγύτητας καθώς οι παλιές ομαδοποιήσεις συγχωνεύονται σε νέες. O πίνακας εγγύτητας D περιέχει όλες τις αποστάσεις d(i,j). Οι ομάδες έχουν αντιστοιχηθεί στους αριθμούς της ακολουθίας 0,1,......, (n − 1) και L(k) είναι το επίπεδο της k ομάδας. Μια ομάδα με αριθμό ακολουθίας m καταγράφεται ως (m) και η εγγύτητα μεταξύ των ομάδων (r) και (s) καταγράφεται ως d[(r),(s)]=η μέγιστη απόσταση ανάμεσα σε δύο σημεία των ομάδων r,s.

Ο αλγόριθμος αποτελείται από τα εξής βήματα:

1. Αρχίζει με την αταίριαστη ομάδα που έχει επίπεδο L(0) = 0 και αριθμό ακολουθίας m = 0, δηλαδή κάθε δείγμα αποτελεί μία ομάδα από μόνο του.
2. Βρίσκει το ζευγάρι των ομάδων που είναι περισσότερο όμοιο με την παρούσα ομάδα, ονομάζοντας το ζευγάρι (r), (s), σύμφωνα με το d[(r),(s)] = max d[(i),(j)], όπου το μέγιστο είναι μεταξύ όλων των ομάδων στην παρούσα ομαδοποίηση.
3. Αυξάνει τον αριθμό ακολουθίας σε m = m + 1, συνενώνει τις ομάδες σε μια για να σχηματίσει την επόμενη ομάδα m. Θέτει το επίπεδο αυτών των ομάδων σε L(m) = d[(r),(s)].
4. Αναβαθμίζει τον πίνακα εγγύτητας, D, διαγράφοντας στήλες και γραμμές που ανταποκρίνονται στις ομάδες (r) και (s) και προσθέτει στήλες και γραμμές που αντιστοιχούν στη νεοσχηματισμένη ομάδα. Η εγγύτητα μεταξύ των νέων ομάδων καταγράφεται ως (r,s) και η παλιά ομάδα (k) καθορίζεται ως d[(k), (r,s)] = max d[(k),(r)], d[(k),(s)].
5. Εάν όλα τα αντικείμενα είναι σε μια ομάδα, τότε σταματάει, αλλιώς συνεχίζει στο βήμα 2.

Το αποτέλεσμα της ιεραρχικής ομαδοποίησης φαίνεται στο παρακάτω δενδρόγραμμα, όπου στον x άξονα τοποθετούνται τα 47 μας δείγματα, δηλαδή οι 47 πολιτείες, και στον y άξονα η τιμή της απόστασης της μετρικής μεταξύ των ομάδων. Σημειώνεται ότι ο αλγόριθμος επιστρέφει στο δενδρόγραμμα την πιο ασθενώς συνδεδεμένη κλάση (την πιο πρόσφατα δημιουργηθήσα) στα δεξιά.(Εντολή plot της βιβλιοθήκης stats της R) 

Επιπλέον, επιλέγοντας να κρατήσουμε την διαμέριση σε 5 κλάσεις (οι οποίες έχουν τονισθεί στο δενδρόγραμμα από τα 5 κόκκινα τετράγωνα) για να γίνει σύγκριση με τα αποτελέσματα του επόμενου αλγορίθμου ομαδοποίησης, έχουμε: (Εντολή cutree της βιβλιοθήκης stats της R)

|  |  |  |
| --- | --- | --- |
| Cluster | Members | States |
| 1 | 18 | 1,6,7,9,10,12,14,15,16,18,21,24,34,35,41,47 |
| 2 | 5 | 2,4,8,11,26 |
| 3 | 6 | 3,22,30,37,39,43 |
| 4 | 9 | 5,20,23,28,29,33,36,40,44 |
| 5 | 9 | 13,17,25,27,31,32,38,42,45,46 |

Ο ακόλουθος 46 (n-1) επί 3 πίνακας μας δείχνει το πως έγιναν σε κάθε βήμα οι ομαδοποιήσεις. Συγκεκριμένα η γραμμή i του πίνακα περιγράφει τη συνένωση των ομάδων στο βήμα i της ομαδοποίησης. Εάν ένα στοιχείο j της γραμμής είναι αρνητικό, τότε η παρατήρηση –j συνενώνεται σε αυτό το στάδιο. Εάν το στοιχείο j είναι θετικό τότε συγχωνεύεται με την ομάδα που σχηματίστηκε στο προηγούμενο στάδιο j του αλγορίθμου. Γι’ αυτό οι αρνητικές είσοδοι δείχνουν ενσωμάτωση των μονάδων σε κλάσεις, ενώ οι θετικές είσοδοι δείχνουν συγχώνευση κλάσεων. Η τρίτη στήλη περιέχει τη μέγιστη απόσταση μεταξύ των στοιχείων της κλασης.

**[,1] [,2] Distance(Height)**

[1,] -17 -42 25.38934

[2,] -13 -25 43.36279

[3,] -12 -41 78.53328

[4,] -7 -18 80.57453

[5,] -10 -14 83.95008

[6,] -23 -28 86.01373

[7,] -5 -36 88.26278

[8,] -1 -9 92.75362

[9,] -21 5 97.50986

[10,] -24 -34 104.07597

[11,] -6 -19 105.39554

[12,] -15 -39 105.83079

[13,] -47 3 111.12242

[14,] -33 -40 118.26005

[15,] -44 10 121.14412

[16,] -38 2 127.21296

[17,] -32 -35 130.34280

[18,] 1 16 135.68071

[19,] 6 7 138.46479

[20,] -31 -45 140.24135

[21,] -37 8 151.83291

[22,] -20 19 151.89617

[23,] -3 -30 154.07792

[24,] -2 -11 154.27149

[25,] 11 17 169.18041

[26,] -43 12 175.57464

[27,] -4 -26 183.26716

[28,] -27 -46 193.22847

[29,] 13 15 198.88558

[30,] -16 21 201.38859

[31,] 18 20 212.88966

[32,] 4 29 214.46541

[33,] 9 25 217.12430

[34,] -8 24 243.94475

[35,] 14 22 258.44305

[36,] 28 31 294.28938

[37,] 26 30 302.46980

[38,] -29 35 310.85412

[39,] -22 23 331.12144

[40,] 32 33 400.88499

[41,] 37 40 495.14391

[42,] 27 34 519.99927

[43,] 36 39 533.82501

[44,] 38 41 632.22666

[45,] 43 44 974.31444

[46,] 42 45 1666.56412

# k-means clustering

Έγινε συσταδοποίηση και με τον αλγόριθμο των k μέσων, χρησιμοποιώντας την εντολή k-means της βιβλιοθήκης stats στην R. Ως αριθμός των ομάδων επιλέχθηκε το 5 μιας και από την ιεραρχική ομαδοποίηση φάνηκε μια καλή επιλογή(βλέπουμε στον παραπάνω πίνακα πως στα στάδια 43-44 η απόσταση αρχίζει να έχει ραγδαία άνοδο, σημάδι ότι πρέπει να μην συνεχίσουμε την ομαδοποίηση καθώς οι κλάσεις μας απομακρύνονται πλέον πολύ) και χρειάσθηκαν 3 επαναλήψεις του αλγορίθμου για αυτό το αποτέλεσμα. Η επιλογή των αρχικών κέντρων έγινε επιλέγοντας τυχαία 5 από τις γραμμές του πίνακα, δηλαδή 5 δείγματα. Η εντολή αυτή χρησιμοποιεί τον αλγόριθμο του Lloyds, την πιο δημοφιλή επιλογή για αυτού του είδους ομαδοποίηση και ως μετρική ορίζεται η Ευκλείδεια απόσταση, και η διαδικασία περιγράφεται συνοπτικά ως εξής:

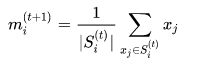
Δίνοντας ένα αρχικό σύνολο των k μέσων στο δείγμα, έστω m1(1),…,mk(1) (βλέπε παρακάτω), ο αλγόριθμος συνεχίζει με εναλλαγή μεταξύ των 2 βημάτων:

**Βήμα ανάθεσης εργασίας:** Εισάγεται η κάθε παρατήρηση στην ομάδα με τη μικρότερη μέση τετραγωνική Ευκλείδεια απόσταση, κι αυτή είναι διαισθητικά ο “κοντινότερος” μέσος.

{\displaystyle S\_{i}^{(t)}={\big \{}x\_{p}:{\big \|}x\_{p}-m\_{i}^{(t)}{\big \|}^{2}\leq {\big \|}x\_{p}-m\_{j}^{(t)}{\big \|}^{2}\ \forall j,1\leq j\leq k{\big \}},}1.PNG

Όπου κάθε μια τιμή καταχωρείται σε ακριβώς μια κλάση, ακόμα κι αν μπορεί να καταχωρηθεί σε δυο ή και παραπάνω από της ομάδες.

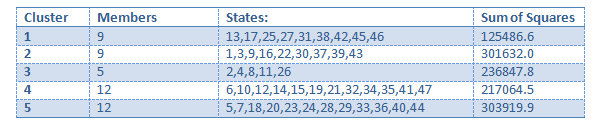
**Βήμα ανενέωσης μέσων**: Υπολογίζονται οι καινούριοι μέσοι όροι, ώστε να είναι τα κέντρα των παρατηρήσεων στις καινούριες ομάδες.

{\displaystyle m\_{i}^{(t+1)}={\frac {1}{|S\_{i}^{(t)}|}}\sum \_{x\_{j}\in S\_{i}^{(t)}}x\_{j}}

Ο αλγόριθμος συγκλίνει όταν οι εντολές σταματούν να διαφοροποιούνται. Δεν υπάρχει καμία εγγύηση ότι ο αλγόριθμος βρίσκει το βέλτιστο.

Ο αλγόριθμος συχνά παρουσιάζεται σαν να αναθέτει αντικείμενα στην κοντινότερη ομάδα μέσω της απόστασης. Χρησιμοποιώντας μια διαφορετική λειτουργία πέρα από την (τετραγωνική) Ευκλείδεια απόσταση μπορεί να εμποδίσει των αλγόριθμο από τη σύγγλιση. Ποικίλες τροποποιήσεις (π.χ. [k-medoids](https://en.wikipedia.org/wiki/K-medoids)) επιτρέπουν τη χρήση άλλων μέτρων απόστασης. Τα αποτελέσματα της ομαδοποίησης είναι τα εξής:

**Clusters:**

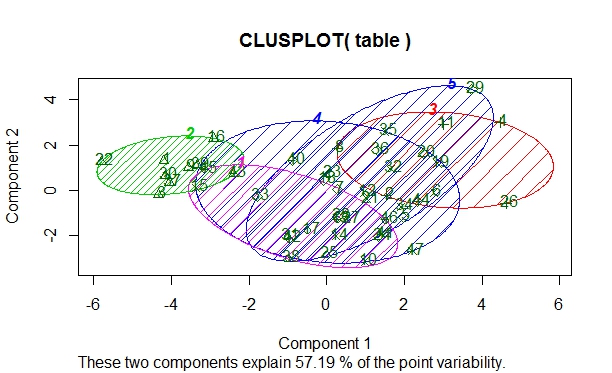
Παρατηρούμε ότι σε σύγκριση με τις ομάδες που δημιουργήθηκαν από τον αλγόριθμο ιεραρχικής συσταδοποίησης υπάρχει πλήρης συμφωνία στις 3 από τις 5 κλάσεις, ενώ στις υπόλοιπες δύο έχει γίνει απορρόφηση στοιχείων από την κλάση 4 που βλέπουμε στον παραπάνω πίνακα (η οποία αντιστοιχεί στην κλάση 1 της ιεραρχικής συσταδοποίησης) στοιχείων της κλάσης 5.

## Cluster means:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Cluster | M | So | Ed | Po1 | Po2 | LF | Pop | M.F |
| 1 | 135 | 0.1111 | 105.111 | 64.333 | 60.444 | 564.22 | 14.89 | 988.00 |
| 2 | 155.889 | 1.000 | 90.000 | 60.889 | 57.444 | 533.44 | 32.33 | 965.33 |
| 3 | 133 | 0.200 | 113.800 | 129.600 | 120.800 | 582.60 | 64.8 | 1002.40 |
| 4 | 132.5 | 0.0833 | 111.333 | 85.917 | 82.0833 | 576.25 | 38.00 | 983.666 |
| 5 | 136.667 | 0.3333 | 108.667 | 99.083 | 93.0833 | 555.75 | 43.00 | 983.833 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Cluster | NW | U1 | U2 | GDP | Ineq | Prob | Time | y |
| 1 | 20.11 | 99.44 | 31.66 | 495.67 | 191.33 | 0.057 | 22.87 | 484.33 |
| 2 | 280.66 | 89.22 | 33.22 | 390.78 | 248.78 | 0.067 | 29.15 | 754.00 |
| 3 | 108.80 | 91.20 | 37.20 | 606.60 | 177.80 | 0.029 | 28.70 | 1765.2 |
| 4 | 41.00 | 95.33 | 34.08 | 572.42 | 170.58 | 0.038 | 26.48 | 770.75 |
| 5 | 84.166 | 99.08 | 34.83 | 567.75 | 185.08 | 0.041 | 26.73 | 1109.9 |

Και η οπτική αναπαράσταση στην R, χρησιμοποιώντας τις 2 κύριες συνιστώσες είναι(Εντολή clustplot της βιβλιοθήκης cluster της R):



Και όλες οι δυνατές αναπαραστάσεις ανά 2 μεταβήτές του συνόλου δεδομένων μας: (εντολή with της βιβλιοθήκης HSAR στην R)

