Επιστήμη διαχείρισης δεδομένων(Data Science), Εργασία 3

Τσερπές Μάριος, ΑΜ:20029, mariostserp@econ.uoa.gr

29 Ιουνίου 2021

Περιεχόμενα

1 Χρησιμοποιήστε τη συνάρτηση kmeans() για να πραγματοποιήστε k-means clustering ανάλυση απαιτώντας τρία (3) clusters.Ποιες πολιτείες ανήκουν σε κάθε ομάδα;

```
> #States in cluster 1
> states.cluster1 <- subset(USArrests, km.out$cluster == 1)
> states.cluster1
              Murder Assault UrbanPop Rape
Arkansas         8.8     190       50 19.5
Colorado         7.9     204       78 38.7
Georgia         17.4     211       60 25.8
Massachusetts    4.4     149       85 16.3
Missouri         9.0     178       70 28.2
New Jersey       7.4     159       89 18.8
Oklahoma         6.6     151       68 20.0
Oregon           4.9     159       67 29.3
Rhode Island     3.4     174       87  8.3
Tennessee       13.2     188       59 26.9
Texas           12.7     201       80 25.5
Virginia         8.5     156       63 20.7
Washington       4.0     145       73 26.2
Wyoming          6.8     161       60 15.6
```

(α′) Οι πολιτείες στην συστάδα (1) ένα.

```
> #States in cluster 2
> states.cluster2 <- subset(USArrests, km.out$cluster == 2)
> states.cluster2
             Murder Assault UrbanPop Rape
Connecticut     3.3     110       77 11.1
Hawaii          5.3      46       83 20.2
Idaho           2.6     120       54 14.2
Indiana         7.2     113       65 21.0
Iowa            2.2      56       57 11.3
Kansas          6.0     115       66 18.0
Kentucky        9.7     109       52 16.3
Maine           2.1      83       51  7.8
Minnesota       2.7      72       66 14.9
Montana         6.0     109       53 16.4
Nebraska        4.3     102       62 16.5
New Hampshire   2.1      57       56  9.5
North Dakota    0.8      45       44  7.3
Ohio            7.3     120       75 21.4
Pennsylvania    6.3     106       72 14.9
South Dakota    3.8      86       45 12.8
Utah            3.2     120       80 22.9
Vermont         2.2      48       32 11.2
West Virginia   5.7      81       39  9.3
Wisconsin       2.6      53       66 10.8
```

(β′) Οι πολιτείες στην (2) συστάδα

```
> #States in cluster 3
> states.cluster3 <- subset(USArrests, km.out$cluster == 3)
> states.cluster3
               Murder Assault UrbanPop Rape
Alabama          13.2     236       58 21.2
Alaska           10.0     263       48 44.5
Arizona           8.1     294       80 31.0
California        9.0     276       91 40.6
Delaware          5.9     238       72 15.8
Florida          15.4     335       80 31.9
Illinois         10.4     249       83 24.0
Louisiana        15.4     249       66 22.2
Maryland         11.3     300       67 27.8
Michigan         12.1     255       74 35.1
Mississippi      16.1     259       44 17.1
Nevada           12.2     252       81 46.0
New Mexico       11.4     285       70 32.1
New York         11.1     254       86 26.1
North Carolina   13.0     337       45 16.1
South Carolina   14.4     279       48 22.5
```

(γ′) Οι πολιτείες στην συστάδα (3)

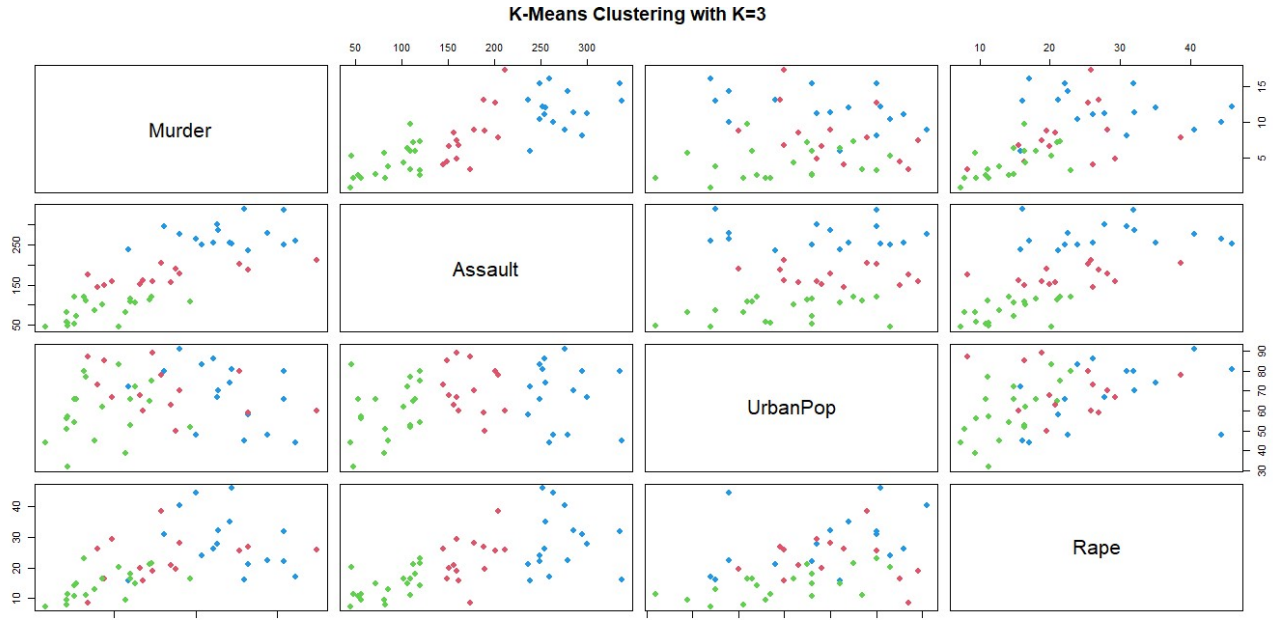
```
> summary(states.cluster1)
     Murder          Assault         UrbanPop          Rape
 Min.   : 3.400   Min.   :145.0   Min.   :50.00   Min.   : 8.30
 1st Qu.: 5.325   1st Qu.:156.8   1st Qu.:60.75   1st Qu.:18.98
 Median : 7.650   Median :167.5   Median :69.00   Median :23.10
 Mean   : 8.214   Mean   :173.3   Mean   :70.64   Mean   :22.84
 3rd Qu.: 8.950   3rd Qu.:189.5   3rd Qu.:79.50   3rd Qu.:26.73
 Max.   :17.400   Max.   :211.0   Max.   :89.00   Max.   :38.70
> summary(states.cluster2)
     Murder          Assault         UrbanPop          Rape
 Min.   :0.80   Min.   : 45.00   Min.   :32.00   Min.   : 7.30
 1st Qu.:2.50   1st Qu.: 56.75   1st Qu.:51.75   1st Qu.:11.03
 Median :3.55   Median : 94.00   Median :59.50   Median :14.55
 Mean   :4.27   Mean   : 87.55   Mean   :59.75   Mean   :14.39
 3rd Qu.:6.00   3rd Qu.:110.75   3rd Qu.:67.50   3rd Qu.:16.88
 Max.   :9.70   Max.   :120.00   Max.   :83.00   Max.   :22.90
> summary(states.cluster3)
     Murder          Assault         UrbanPop          Rape
 Min.   : 5.90   Min.   :236.0   Min.   :44.00   Min.   :15.80
 1st Qu.:10.30   1st Qu.:251.2   1st Qu.:55.50   1st Qu.:21.95
 Median :11.75   Median :261.0   Median :71.00   Median :26.95
 Mean   :11.81   Mean   :272.6   Mean   :68.31   Mean   :28.38
 3rd Qu.:13.50   3rd Qu.:287.2   3rd Qu.:80.25   3rd Qu.:32.85
 Max.   :16.10   Max.   :337.0   Max.   :91.00   Max.   :46.00
```

(δ΄) Περιγραφικά στατιστικά των 3 *clusters*

- **Ποιες και πόσες πολιτείες υπάρχουν σε κάθε συστάδα.**
  Από την εικόνα α΄ μπορούμε να δούμε ότι ο αριθμός των πολιτειών που εκχωρούνται στην συστάδα ένα , είναι 14. Συγκεκριμένα οι εν λόγω πολιτείες είναι οι Arkansas, Colorado,Massachusetts , Georgia,Missuri ,New Jersey, Oklahoma ,Oregon,Rhode Island, Tennessee, Texas, Virginia, Washington, Wyoming.
  Στην συστάδα δύο(2) ο αριθμός των πολιτειών είναι 20, όπου και είναι η μεγαλύτερη εκ των 3. Συγκεκριμένα οι πολιτείες είναι οι Connecticut, Hawaii , Idaho, Indiana, Iowa, Kansas, Kentucky, Maine, Minnesota, Montana, Nebraska, New Hampshire, North Dakota, Ohio, Pennsylvania, South Dakota, Utah, Vermont, West Virginia, Wisconsin.
  Στην συστάδα τρία(3) ο αριθμός των πολιτειών είναι 16 και οι πολιτείες που υπάρχουν στην εν λόγω συστάδα είναι Alabama, Alaska, Arizona, California, Delaware, Florida, Illinois, Louisiana, Maryland, Michigan, Mississippi, Nevada, New Mexico, New York, North Carolina, South Carolina
  Επίσης, όπως φαίνεται από τον πίνακα δ΄ , ο μέσος όρος συλλήψεων για φόνους για τις πολιτείες που εκχωρούνται στη συστάδα τρία(3) είναι ο υψηλότερος, δηλαδή περίπου 11.81.Απεναντίας, οι συλλήψεις για φόνους στις πολιτείες στη συστάδα δύο(2) είναι περίπου 3 φορές μικρότερη, δηλαδή 4.27, ενώ στις πολιτείες της συστάδας ένα(1) , ο μέσος όρος είναι συλλήψεων για φόνους(murder arrests) είναι 8.214, περίπου 2 φορές υψηλότερος από τις πολιτείες της συστάδας δύο(2).
  Όσον αφορά το μέσο όρο για τις συλλήψεις επιθέσων(assault arrests) παρατηρούμε ότι οι πολιτείες της συστάδας τρία(3) έχουν κι εδώ την πρωτοκαθεδρία, ενώ κατά πολύ χαμηλότερος είναι ο μέσος όρος συλλήψεων επιθέσεων για τις πολιτείες της συστάδα(2), ενώ στις πολιτείες της συστάδας ένα(1) ο μέσος όρος συλλήψεων είναι περίπου δυο φορές υψηλότερος από τη συστάδα(2) και χαμηλότερος εν συγκρίσει με το μέσο όρος της συστάδας τρία(3).
  Παρόμοια τάση παρουσιάζεται και για τις συλλήψεις για βιασμούς(rape arrests) όπου οι πολιτείες της συστάδας τρία , εμφανίζουν τον υψηλότερο μέσο όρο, ενώ η συστάδα δύο(2) τον χαμηλότερο και οι πολιτείες της συστάδας ένα(1) παρουσιάζουν μέσο όρο συλλήψεων για βιασμούς 6 μονάδες χαμηλότερο από τη συστάδα τρία(3) και περίπου 8 μονάδες υψηλότερο από τη συστάδα δύο(2).
  Τέλος, όσον αφορά το ποσοστό του αστικού πληθυσμού , φαίνεται ότι ο μέσος όρος για

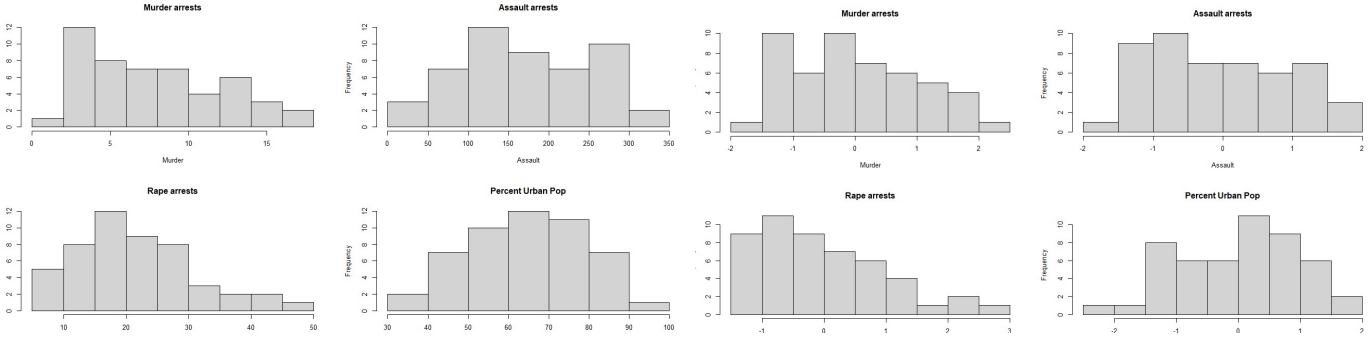
τις συστάδα ένα(1) είναι ο υψηλότερος, ενώ ελαφρώς χαμηλότερος είναι ο μέσος όρος για τη συστάδα τρία(3). Σχετικά , με τη συστάδα δύο(2) , το μέσο ποσοστό των κατοίκων που ζουν στην πόλη είναι χαμηλότερο περίπου κατά 10 τοις εκατό σε σχέση με τις υπόλοιπες δύο συστάδες , γεγονός που αναμενόταν καθώς στις εν λόγω πολιτείες ο μέσος όρος εγκληματικών ενεργειών όπως δολοφονίες, απλές επιθέσεις ή βιασμούς ήταν χαμηλότερος εν συγκρίσει με τις υπόλοιπες δύο συστάδες.

2 Χρησιμοποιήστε τη συνάρτηση plot() ή/και τη συνάρτηση scatter3D()(μέρος της βιβλιοθήκης plot3D)για να αποτυπώσετε τις παρατηρήσεις που ανήκουν σε κάθε cluster σε δύο ή σε τρεις διαστάσεις αντίστοιχα. Για την αποτύπωση των παρατηρήσεων του κάθε cluster, χρησιμοποιήστε διαφορετικό χρώμα και στους άξονες των γραφημάτων χρησιμοποιήστε όλες τις μεταβλητές του συνόλου δεδομένων ανά δύο ή ανά τρεις ανάλογα με το ποια επιλογή από τις παραπάνω (plot() ή scatter3D())επιλέξετε.



(ε′) 2 Dimensional KMeans clustering with K= 3

3 Επαναλάβετε τη διαδικασία του ερωτήματος (i), αφού προηγουμένως χρησιμοποιήσετε την συνάρτηση scale() για να μετασχηματίσετε όλες τις μεταβλητές του συνόλου δεδομένων έτσι ώστε να έχουν μέση τιμή ίση με 0 και τυπική απόκλιση ίση με 1. Ποια η επίδραση του παραπάνω μετασχηματισμού των μεταβλητών στο διαχωρισμό των clusters?Κατά τη γνώμη σας θα πρέπει να μετασχηματισθούν οι μεταβλητές στην περίπτωση που μελετούμε στη συγκεκριμένη άσκηση; Εξηγήστε!



(ϛ′) Η κατανομή των μεταβλητών πριν το scaling (ζ′) Η κατανομή των μεταβλητών μετά το scaling

Σχήμα 1: Κατανομή μεταβλητών πριν το *scaling* και μετά το *scaling*.

- **Ποια η επίδραση του μετασχηματισμού στο διαχωρισμό των** *clusters*;
  Α)Όπως φαίνεται, ο μετασχηματισμός των χαρακτηριστικών έτσι ώστε η μέση τιμή τους να είναι μηδέν και η τυπική απόκλιση 1, επηρέασε τις αρχικές συστάδες που παρατηρήθηκαν.
  Πιο συγκεκριμένα, η πρώτη παρατήρηση αφορά το ότι άλλαξε το μέγεθος της πρώτης και της τρίτης συστάδας, δηλαδή το πλήθος των πολιτειών που εκχωρούνται στις εν λόγω συστάδες.

```
[1] "Cluster one(1) : 14. Cluster two(2) : 20. Cluster three(3) : 16"
```

(α′) Πλήθος πολιτειών σε κάθε συστάδα πριν τον μετασχηματισμό

Όπως φαίνεται πριν το μετασχηματισμό ο αριθμός των πολιτειών που εκχωρούνται στην πρώτη συστάδα είναι 14 , στην δεύτερη 20 και στην τρίτη 16.

```
[1] "Cluster one(1) : 17. Cluster two(2) : 20. Cluster three(3) : 13"
```

(β′) Πλήθος πολιτειών σε κάθε συστάδα μετά τον μετασχηματισμό

Μετά το μετασχηματισμό παρατηρούμε ότι η συστάδα 2 αντιπροσωπεύεται τον ίδιο αριθμό πολιτειών όπως και πριν το μετασχηματισμό, ωστόσο στην συστάδα ένα(1) ο αριθμός των πολιτειών αυξήθηκε σε 17 από 14 , ενώ στη συστάδα τρία(3) το πλήθος των πολιτειών μειώθηκε από 16 σε 13.

Β) Επιπροσθέτως, φαίνεται ότι μετά το μετασχηματισμό η σύνθεση των συστάδων έχει αλλάξει μερικώς. Συγκεκριμένα, όσον αφορά τη συστάδα ένα(1):

```
> #States in cluster 1
> states.cluster1 <- subset(USArrests, km.out$cluster == 1)
> states.cluster1
             Murder Assault UrbanPop Rape
Arkansas        8.8     190       50 19.5
Colorado        7.9     204       78 38.7
Georgia        17.4     211       60 25.8
Massachusetts   4.4     149       85 16.3
Missouri        9.0     178       70 28.2
New Jersey      7.4     159       89 18.8
Oklahoma        6.6     151       68 20.0
Oregon          4.9     159       67 29.3
Rhode Island    3.4     174       87  8.3
Tennessee      13.2     188       59 26.9
Texas          12.7     201       80 25.5
Virginia        8.5     156       63 20.7
Washington      4.0     145       73 26.2
Wyoming         6.8     161       60 15.6
```

(γ′) Οι πολιτείες στη συστάδα ένα(1) πριν το μετασχηματισμό

```
> df.scaled.cluster1
                Murder     Assault   UrbanPop        Rape
Arkansas      0.23234938  0.23086801 -1.07359268 -0.18491660
Connecticut  -1.03041900 -0.72908214  0.79172279 -1.08174077
Delaware     -0.43347395  0.80683810  0.44629400 -0.57994629
Hawaii       -0.57123050 -1.49704226  1.20623733 -0.11018125
Indiana      -0.13500142 -0.69308401 -0.03730631 -0.02476943
Kansas       -0.41051452 -0.66908525  0.03177945 -0.34506377
Massachusetts -0.77786532 -0.26110644  1.34440885 -0.52656390
New Jersey   -0.08908257 -0.14111267  1.62075188 -0.25965195
Ohio         -0.11204199 -0.60908837  0.65355127  0.01793648
Oklahoma     -0.27275797 -0.23710769  0.16995096 -0.13153421
Oregon       -0.66306820 -0.14111267  0.10086521  0.86137826
Pennsylvania -0.34163624 -0.77707965  0.44629400 -0.67603460
Rhode Island -1.00745957  0.03887798  1.48258036 -1.38068216
Utah         -1.05337842 -0.60908837  0.99898006  0.17808366
Virginia      0.16347111 -0.17711080 -0.17547783 -0.05679886
Washington   -0.86970302 -0.30910395  0.51537975  0.53040744
Wyoming      -0.22683912 -0.11711392 -0.38273510 -0.60129925
```

(δ′) Οι πολιτείες στη συστάδα ένα(1) μετά το μετασχηματισμό.

Σχήμα 2: Σύνθεση συστάδας ένα(1) πριν και μετά τον μετασχηματισμό.

Αυτό που παρατηρούμε είναι ότι μετά το scaling οι πολιτείες που παρέμειναν στη συστάδα ένα(1) είναι συνολικά εννέα(9) και συγκεκριμλενα οι Arkansas, Massachusetts, New Jersey, Oklahoma, Oregon, Rhode Island, Virginia, Wyoming, Washinghton. Δηλαδή, 9 από τις 14 παρέμειναν στη συστάδα ένα(1) , ενώ οι υπόλοιπες 8 που προστέθηκαν μετά το μετασχηματισμό άνηκαν στις συστάδες δύο(2) ή τρία(3

Αναφορικά , με τη δεύτερη(2η) συστάδα:

```
> #States in cluster 2
> states.cluster2 <- subset(USArrests, km.out$cluster == 2)
> states.cluster2
             Murder Assault UrbanPop Rape
Connecticut     3.3     110       77 11.1
Hawaii          5.3      46       83 20.2
Idaho           2.6     120       54 14.2
Indiana         7.2     113       65 21.0
Iowa            2.2      56       57 11.3
Kansas          6.0     115       66 18.0
Kentucky        9.7     109       52 16.3
Maine           2.1      83       51  7.8
Minnesota       2.7      72       66 14.9
Montana         6.0     109       53 16.4
Nebraska        4.3     102       62 16.5
New Hampshire   2.1      57       56  9.5
North Dakota    0.8      45       44  7.3
Ohio            7.3     120       75 21.4
Pennsylvania    6.3     106       72 14.9
South Dakota    3.8      86       45 12.8
Utah            3.2     120       80 22.9
Vermont         2.2      48       32 11.2
West Virginia   5.7      81       39  9.3
Wisconsin       2.6      53       66 10.8
```

(α′) Οι πολιτείες στη συστάδα δύο(2) πριν το μετασχηματισμό

```
> df.scaled.cluster2
                Murder     Assault   UrbanPop        Rape
Alabama      1.24256408 0.78283935 -0.52090661 -0.003416473
Alaska       0.50786248 1.10682252 -1.21176419  2.484202941
Arizona      0.07163341 1.47880321  0.99898006  1.042878388
California   0.27826823 1.26281442  1.75892340  2.067820292
Colorado     0.02571456 0.39885929  0.86080854  1.864967207
Florida      1.74767144 1.97077766  0.99898006  1.138966691
Georgia      2.20685994 0.48285493 -0.38273510  0.487701523
Illinois     0.59970018 0.93883125  1.20623733  0.295524916
Louisiana    1.74767144 0.93883125  0.03177945  0.103348309
Maryland     0.80633501 1.55079947  0.10086521  0.701231086
Michigan     0.99001041 1.01082751  0.58446551  1.480613993
Mississippi  1.90838741 1.05882502 -1.48810723 -0.441152078
Missouri     0.27826823 0.08687549  0.30812248  0.743936999
Nevada       1.01296983 0.97482938  1.06806582  2.644350114
New Mexico   0.82929443 1.37080881  0.30812248  1.160319648
New York     0.76041616 0.99882813  1.41349461  0.519730957
North Carolina 1.19664523 1.99477641 -1.41902147 -0.547916860
South Carolina 1.51807718 1.29881255 -1.21176419  0.135377743
Tennessee    1.24256408 0.20686926 -0.45182086  0.605142783
Texas        1.12776696 0.36286116  0.99898006  0.455672088
```

(β′) Οι πολιτείες στη συστάδα δύο(2) μετά το μετασχηματισμό.

Σχήμα 3: Σύνθεση συστάδας δύο(2) πριν και μετά τον μετασχηματισμό.

Στη συστάδα δύο(2), όπως αναφέραμε το πλήθος των πολιτειών που εκχωρούνται παραμένει ίδιο, ωστόσο έχουν αλλάξει οι πολιτείες.Μάλιστα, καμία πολιτεία που είχει εκχωρηθεί στη συστάδα δύο(2) πριν το μετασχηματισμό δεν έχει παραμείνει στην ίδια συστάδα μετά το μετασχηματισμό των μεταβλητών.

Σχετικά με την συστάδα τρία(3):

```
> #States in cluster 3
> states.cluster3 <- subset(USArrests, km.out$cluster == 3)
> states.cluster3
               Murder Assault UrbanPop Rape
Alabama          13.2     236       58 21.2
Alaska           10.0     263       48 44.5
Arizona           8.1     294       80 31.0
California        9.0     276       91 40.6
Delaware          5.9     238       72 15.8
Florida          15.4     335       80 31.9
Illinois         10.4     249       83 24.0
Louisiana        15.4     249       66 22.2
Maryland         11.3     300       67 27.8
Michigan         12.1     255       74 35.1
Mississippi      16.1     259       44 17.1
Nevada           12.2     252       81 46.0
New Mexico       11.4     285       70 32.1
New York         11.1     254       86 26.1
North Carolina   13.0     337       45 16.1
South Carolina   14.4     279       48 22.5
```

(α′) Οι πολιτείες στη συστάδα τρία πριν το μετασχηματισμό

```
> df.scaled.cluster3
                 Murder    Assault    UrbanPop        Rape
Idaho        -1.1911350 -0.6090884 -0.79724965 -0.7507699
Iowa         -1.2829727 -1.3770485 -0.58999237 -1.0603878
Kentucky      0.4389842 -0.7410815 -0.93542116 -0.5265639
Maine        -1.3059321 -1.0530653 -1.00450692 -1.4340645
Minnesota    -1.1681755 -1.1850585  0.03177945 -0.6760346
Montana      -0.4105145 -0.7410815 -0.86633540 -0.5158874
Nebraska     -0.8008247 -0.8250772 -0.24456358 -0.5052109
New Hampshire -1.3059321 -1.3650491 -0.65907813 -1.2525644
North Dakota -1.6044046 -1.5090416 -1.48810723 -1.4874469
South Dakota -0.9156219 -1.0170672 -1.41902147 -0.9002406
Vermont      -1.2829727 -1.4730435 -2.31713632 -1.0710643
West Virginia -0.4793928 -1.0770641 -1.83353601 -1.2739174
Wisconsin    -1.1911350 -1.4130466  0.03177945 -1.1137702
```

(β′) Οι πολιτείες στη συστάδα τρία μετά το μετασχηματισμό.

Σχήμα 4: Σύνθεση συστάδας τρία(3) πριν και μετά τον μετασχηματισμό.

Τέλος, ακριβώς το ίδιο συμβαίνει και με τη συστάδα τρία, όπου καμία εκ των πολιτειών , που είχαν εκχωρηθεί πριν το μετασχηματισμό στην εν λόγω συστάδα, δεν υπάρχει στην ίδια συστάδα μετά το μετασχηματισμό και συχρόνως το μέγεθός της έχει μειωθεί από 16 σε 13 πολιτείες. **Μόλις το 18 τοις εκατό(9/50) εκ των πολιτειών επανεκχωρήθηκαν στην ίδια συστάδα. Μάλιστα και οι 9 εξ αυτών αφορούν την συστάδα ένα(1).**

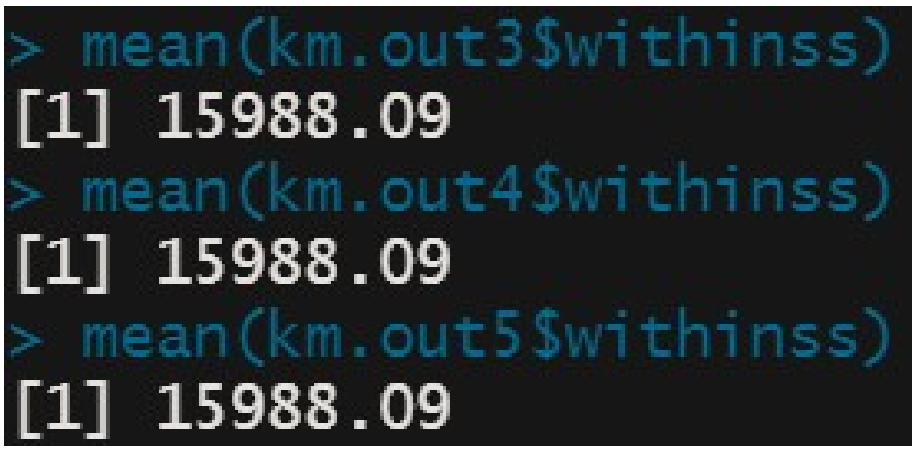
- Επίσης, όπως φαίνεται πριν το μετασχηματισμό οι 3 συστάδες εξηγούσαν **το 86.5** τοις εκατό της συνολικής μεταβλητότητας της διακύμανσης των αρχικών δεδομένων, ενώ μετά το μετασχηματισμό , οι 3 συστάδες εξηγούν το **60** τοις εκατό.

- **Σχετικά με τον αν πρέπει να μετασχηματιστούν οι μεταβλητές.**
  Έχουμε 4 γνωρίσματα εκ των οποίων οι τρεις μεταβλητές Murder, Assault, Rape , είναι αριθμητικές τιμές. Μάλιστα, οι μετρήσεις είναι ο αριθμός συλλήψεων ανά 100 χιλιάδες , ενώ σχετικά με τη μεταβλητή UrbanPop η μονάδα μέτρησης είναι ποσοστό και συγκεκριμένα το ποσοστό του πληθυσμού των κατοίκων που ζουν στην πόλη της εκάστοτε πολιτείας.
  Επόμενως, σε αυτό το σενάριο, συλλογιζόμενοι ότι και το ποσοστό των ανθρώπων που ζουν στην πόλη σε κάθε πολιτεία, συνεισφέρει εξίσου στη συσταδοποίηση σε συνδυασμό με το ότι οι μονάδες μέτρησης είναι διαφορετικές , θεωρώ ότι ο μετασχηματισμός είναι απαραίτητος, καθιστώντας, έτσι, τις μεταβλητές περισσότερο συγκρίσιμες , σε σχέση με την αρχική τους μορφή , δηλαδή συγκρίνοντας ποσοστό και αριθμό ανά 100.000 κατοίκους.
  Επιπροσθέτως, ακόμα κι αν οι μονάδες μέτρησης ήταν οι ίδιες για όλες τα χαρακτηριστικά, παρατηρούμε ότι η μεταβλητή που αφορά τις συλλήψεις για επιθέσεις(Assault arrests)

το 50 τοις εκατό των δεδομένων λαμβάνει τιμές μεγαλύτερες από 159, φτάνοντας μέχρι και 337, όταν οι στις υπόλοιπες υπόλοιπες μεταβλητές η μέγιστη τιμή είναι 91(UrbanPop), 46(Rape) 17.400(Assault arrests). Επομένως, θεωρώ και στο εν λόγω σενάριο, η κανονικοποίηση είναι απαραίτητη έτσι ώστε να φέρουμε στην ίδια κλίμακα τις τιμές των μεταβλητών.

4 Επαναλάβετε τη διαδικασία του ερωτήματος (i), χρησιμοποιώντας 1, 10 και 50 αρχικοποιήσεις της λύσης αντίστοιχα. Παρατηρείται βελτίωση της λύσης όσο το πλήθος των αρχικοποιήσεων αυξάνεται; Σχολιάστε και εξηγήστε τα αποτελέσματα!

Αυτό που αρχικά παρατηρούμε είναι ότι και στις 3 περιπτώσεις, δηλαδή χρησιμοποιώντας 1, 10 και 50 αρχικοποίησεις της λύσης, το 86,5 τοις εκατό της συνολικής διακύμανσης των αρχικών δεδομένων εξηγείται από τις 3 συστάδες.
Επίσης, καθώς αυξάνεται ο αριθμός των αρχικοποίησεων της λύσης, δεν παρατηρούμε να βελτιώνεται η λύση από την άποψη ότι δεν μειώνεται η συνολική απόσταση των παρατηρήσεων μέσα στη συστάδα ($total\_withinss$) καθώς και στις 3 περιπτώσεις είναι 47964.27. Συγχρόνως, η απόσταση της μίας συστάδας από την άλλη συνεχίζει να είναι σταθερή όσο αυξάνεται ο αριθμός των αρχικοποιήσεων, δηλαδή και στις 3 περιπτώσεις παραμένει ίδια η συνολική απόσταση ανάμεσα στα *clusters*.
Στην παρακάτω εικόνα παραθέτω τη μέση τετραγωνική απόσταση μεταξύ κάθε παρατήρησης και του κοντινότερου centroid τους,



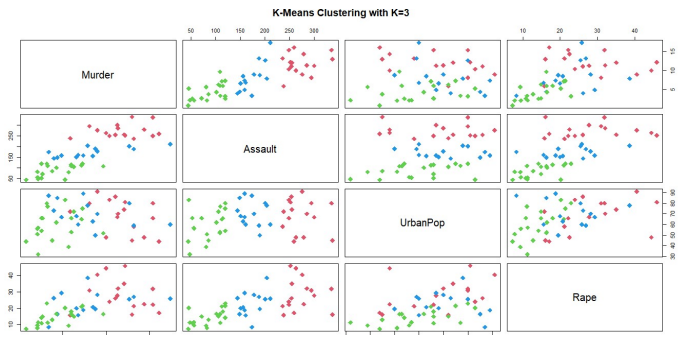
(α′) Mean squared distance between each instance and its closest centroid with 1, 10 and 50 initializations.

Αυτό που αλλάζει, είναι ότι μεταβάλλεται καθώς αυξάνεται ο αριθμός αρχικοποίησης της λύσης σε 50, ο κεντροειδής της συστάδας ένα και δύο αντιστρέφονται. Συγκεκριμένα, μπορούμε να παρατηρήσουμε την παρακάτω εικόνα.

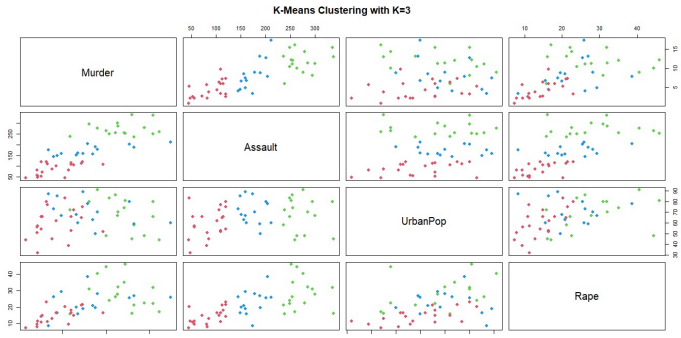
```
> km.out3$centers
     Murder  Assault UrbanPop     Rape
1 11.812500 272.5625 68.31250 28.37500
2  4.270000  87.5500 59.75000 14.39000
3  8.214286 173.2857 70.64286 22.84286
> km.out4$centers
     Murder  Assault UrbanPop     Rape
1 11.812500 272.5625 68.31250 28.37500
2  4.270000  87.5500 59.75000 14.39000
3  8.214286 173.2857 70.64286 22.84286
> km.out5$centers
     Murder  Assault UrbanPop     Rape
1  4.270000  87.5500 59.75000 14.39000
2 11.812500 272.5625 68.31250 28.37500
3  8.214286 173.2857 70.64286 22.84286
```

(β′) Οι κεντροειδείς για 1, 10 και 50 αρχικοποιήσεις της λύσης

Επομένως, παρατηρώ ότι αν και η λύση εξαρτάται από την αρχικοποίηση , στο εν λόγω παράδειγμα δε διαπιστώθηκε αλλαγή στη λύση , αυξάνοντας τον αριθμό των αρχικοποίησης τόσο όσον αφορά την απόσταση των παρατηρήσεων από τα *centroids* που ανήκουν καθώς και στην απόσταση της μίας συστάδας από την άλλη. Παρακάτω μπορούμε να δούμε ότι οι πολιτείες που εκχωρούνται σε κάθε συστάδα δεν αλλάζουν είναι ίδιες για κάθε αριθμό αρχικοποίησης.
Ωστόσο, θεωρώ ότι αριθμός των 3 συστάδων δε φαίνεται να είναι ιδιαίτερα πετυχημένος παρά το γεγονός ότι ερμηνεύεται το 86.5 τοις εκατό της συνολικής διακύμανσης από τα 3 *clusters*. Δηλαδή και στις 3 περιπτώσεις(δηλαδή αρχικοποίηση 1, 10 και 50) ενώ αλγόριθμος συγκλίνει στην ίδια λύση, νομίζω ότι υπάρχουν σημεία όπου η κατάσταση είναι αρκετά μπερδεμένη όπως φαίνεται από τα κατωτέρω γραφήματα. Για παράδειγμα στα διαγράμματα Murder $\sim$ UrbanPop , Murder $\sim$ Rape , UrbanPop $\sim$ Rape , φαίνεται ότι η κατάσταση είναι αρκετά μπερδεμένη, το οποίο ,ενδεχομένως, να μην οφείλεται στο ότι δεν ήμασταν τυχεροί κατά τη αρχικοποίηση αλλά σε στο ότι μια διαφορετική επιλογή της υπερπαραμέτρου, οδηγούσε σε αποφυγή παραβίασης ορίων όπως φαίνεται από τα προαναφερθέντα γραφήματα.



(γ′) init = 1 and init = 10



(δ′) init = 50

Σχήμα 5: Οπτικοποίηση συσταδοποίησης για αριθμό αρχικοποιήσεων 1, 10 και 50..