ЗВІТ З ЛАБОРАТОРНОЇ РОБОТИ №1
«КЛАСТЕРИЗАЦІЯ З ВІДОМОЮ КІЛЬКІСТЮ КЛАСТЕРІВ»

частина 2

Ломако О., 2 к. маг, «статистика», варіант 9

В другій частині першої лабораторної роботи необхідно було самому обрати дані для кластеризації. Цікавою, на мою думку, виявилась ідея, спробувати провести кластерний аналіз для деяких соціально-політичних даних країн світу.

Але, це виявилось не так просто: знайти актуальний датасет з поважною кількістю змінних не вдавалося можливим (а якщо і так, це була платна функція). Тому я вирішив зібрати датасет вручну. На жаль, в адекватному обсязі вдалося зібрати дані для 143 країн світу по наступним показникам:

- *access to electricity* – доступність населення до електроенергії (у %);
- *birth rate* - показник народжуваності;
- *death rate* – показник смертності;
- *gdp per capita* – ВВП на душу населення;
- *urbanization rate* – рівень урбанізації (тобто % міського населення);
- *unemployment* – показник безробіття;
- *pop density* – показник щільності населення;
- *agricultural land* – відсоток земель, що використовуються для ведення с/г.

На жаль, лише по цим даним вдалося зібрати якнайповнішу картину по більшій частині країн світу. Метою роботи є задоволення власної цікавості: за якою ознакою будуть класифіковані країни (рівень розвитку, форма правління, тип економіки…?)

Спершу, підключимо всі бібліотеки, зчитаємо дані, і підготуємо їх до аналізу.

```
> library(readxl) # підключимо бібліотеки
> library(factoextra)
> library(CCA)
> library(rgl)
> library(cluster)
> library(fossil)
>
> # зчитаємо і підготуємо дані
> data <- read_excel('data.xlsx')
>
> data <- as.data.frame(data)
>
> countries <- as.vector(t(data[,1]))
> row.names(data) <- countries
> data <- data[,-1]
```

На жаль, при першій спробі виконання даної роботи була допущена груба помилка: у нас всі одиниці, крім ВВП на душу населення, вимірюються у відсотках (тобто, можуть набувати значення від 1 до 100), або є результатом

якогось відношення (як-от показник щільності населення). В цілому, демографічні показники (за певними винятками, як-от Монако і їхня щільність населення) є достатньо невеликими числами, в той час як ВВП на душу населення в цих даних вимірюється в млрд доларів США, і легко може набувати значень і більше 1 000.

Тому, перед роботою з такими даними, спершу варто їх відцентрувати і пронормувати. Проробимо це наступним чином:

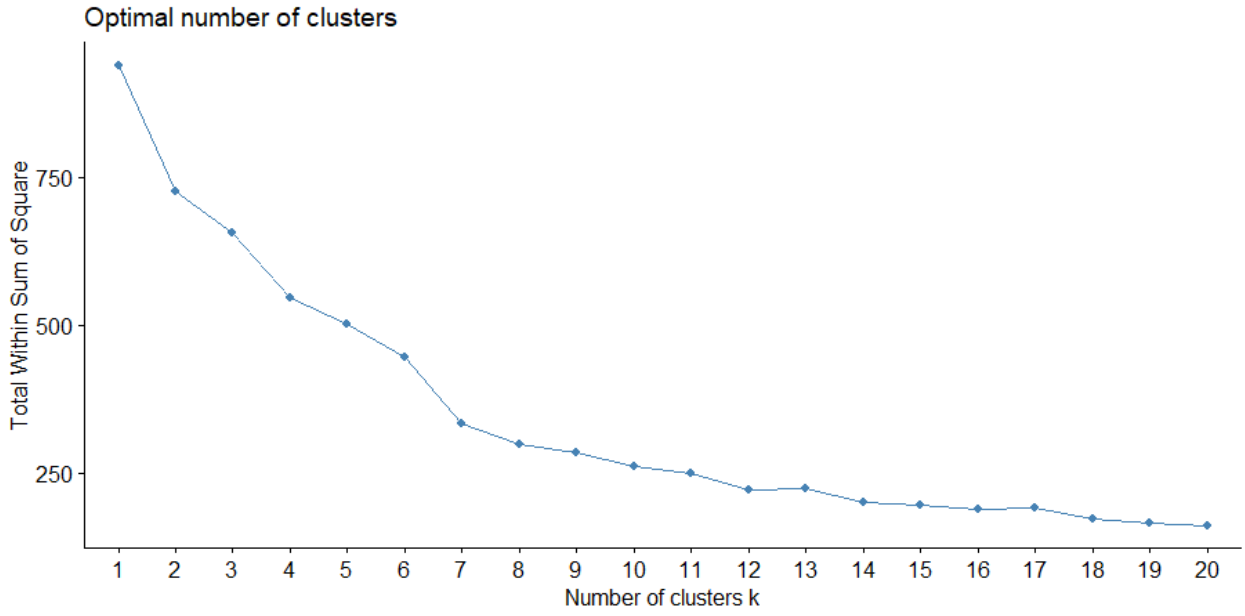
$$y_j = \frac{X_j - \overline{X}}{\sqrt{S}}, \text{де } \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad S = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

де (X – досліджувана величина, y – відповідна їй нормована)

```
> values <- data.frame(row.names = c('mean','var'))
> data_mod <- matrix(nrow = 143, ncol = 7)
> for(i in 1:7){
+   values[1,i] <- mean(as.numeric(unlist(data[,i])))
+   values[2,i] <- sqrt(var(as.numeric(unlist(data[,i]))))*18/17)
+ }
> for(i in 1:7){
+   for(j in 1:143){
+     data_mod[j,i] <- as.numeric((as.numeric(unlist(data[j,i])) - values[1,i
])/(values[2,i]))
+   }
+ }
> row.names(data_mod) <- row.names(data)
> colnames(data_mod) <- colnames(data)
> data_mod <- as.data.frame(data_mod)
> data <- data_mod
```

Побудуємо діаграму внутрішньогрупової суми квадратів.

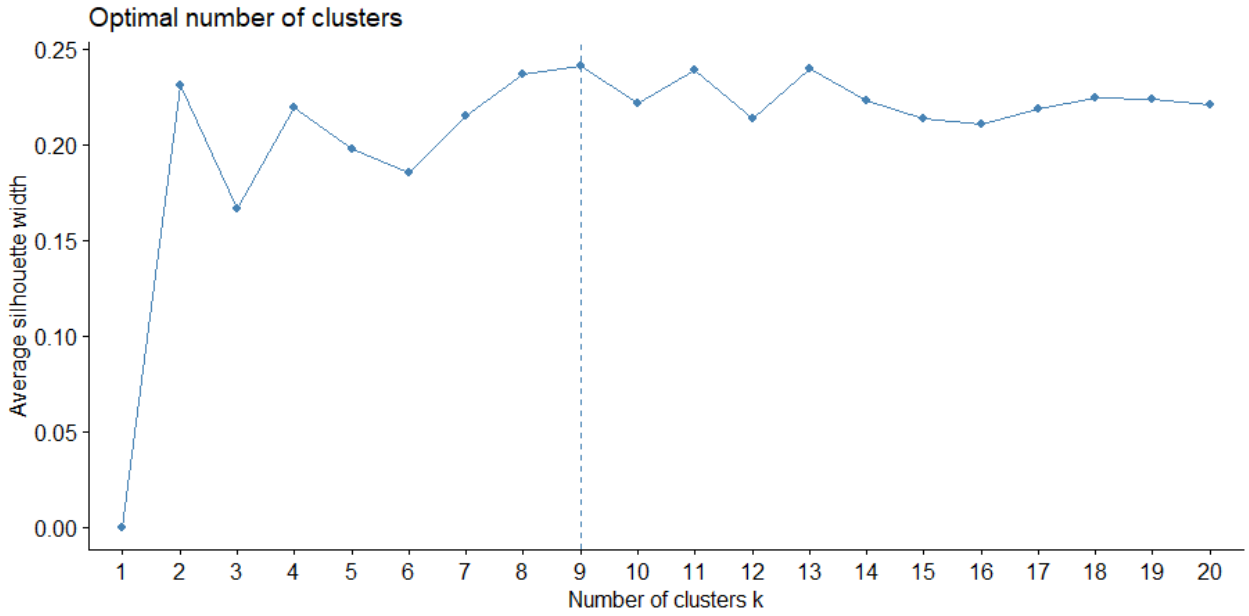
```
> # побудуємо діаграму внутрішньогрупової суми квадратів
> fviz_nbclust(data, kmeans, method = "wss",k.max = 20)
```



Здається, злам відбувається при $k = 4$ або $k = 7$.

Тепер побудуємо діаграму середніх силуетів.

```
> # діаграма середніх силуетів
> fviz_nbclust(data, kmeans, method = "silhouette",k.max = 20)
```



Тут помітні максимуми при $k = 2, 4, 9$.

Спершу застосуємо метод центроїдів. Спробуємо $k = 2$. Одразу від себе скажу, що на мою думку, це не зовсім доцільно ділити 100+ країн світу лише на 2 групи, все набагато складніше.

```
> km.res2 <- kmeans(data, 2, nstart = 25)
> km.res2$betweenss/km.res2$tot.withinss # відношення міжкластерної суми квад
ратів до внутрішньокластерної
[1] 0.2947633
```

Тут можемо бачити, що внутрішньокластерна сума квадратів є навіть більшою за міжкластерну, а отже, на мою думку, ця кластеризація є зовсім невдалою.

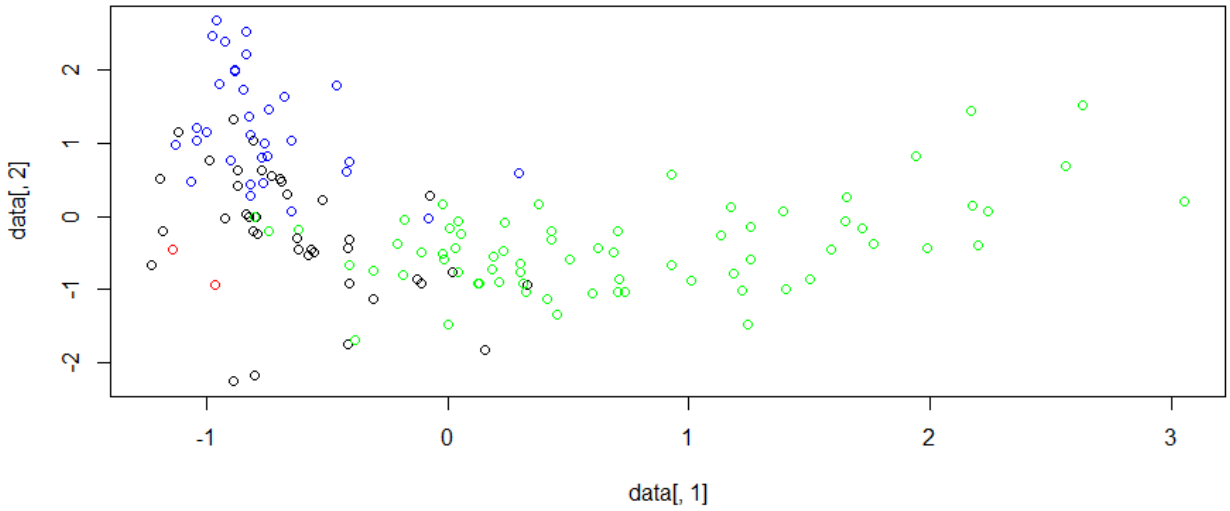
Спробуємо далі $k = 4$.

```
> km.res4 <- kmeans(data, 4, nstart = 25)
> km.res4$betweenss/km.res4$tot.withinss # відношення міжкластерної суми квад
ратів до внутрішньокластерної
[1] 0.929981
```

І знову у нас внутрішньокластерна сума квадратів більша за міжкластерну, а отже сподіватись на те, що тут виділяється якась структура даних не варто.

Про це і свідчитиме діаграма розсіювання, наприклад, перших двох змінних.

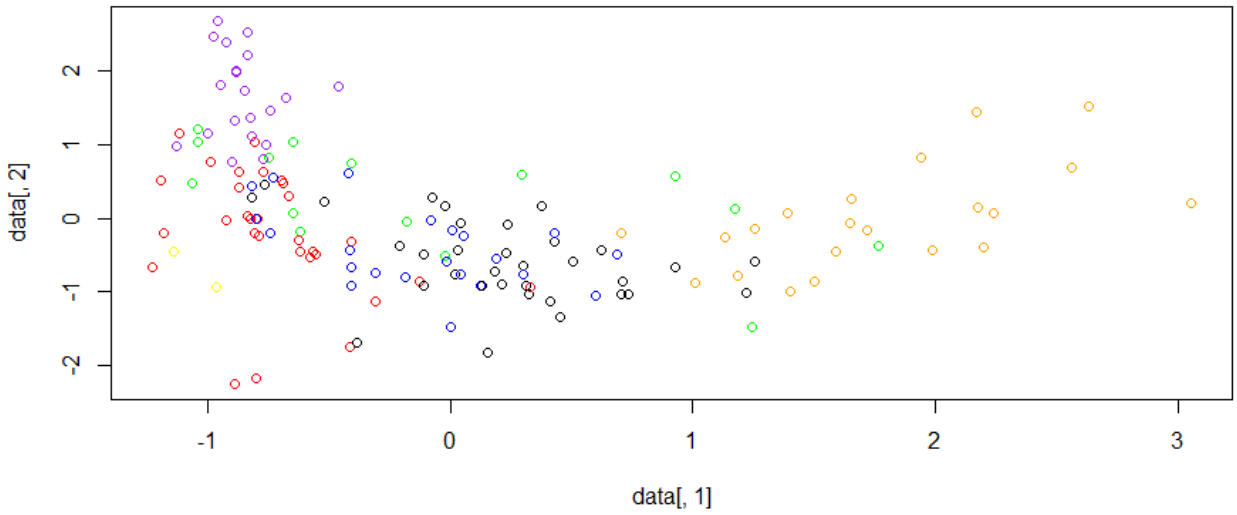
```
> plot(data[,1], data[,2], col = col[km.res3$cluster]) # діаграма розсіювання
1 і 2 змінної з кластеризацією
```



Тепер спробуємо $k = 7$.

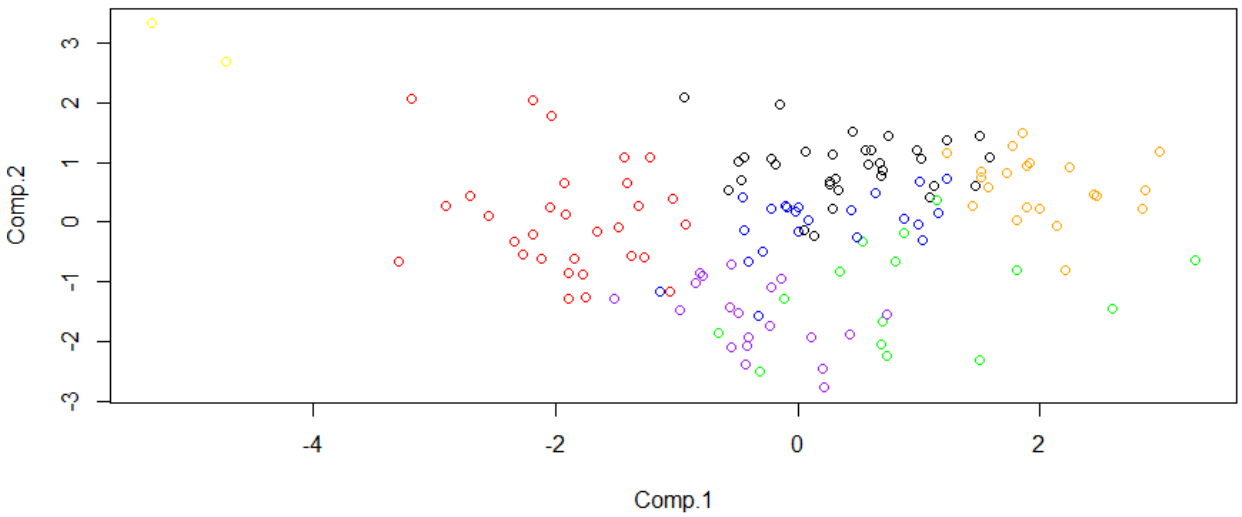
```
> km.res7 <- kmeans(data, 7, nstart = 25)
> km.res7$betweenss/km.res7$tot.withinss # відношення міжкластерної суми квад
ратів до внутрішньокластерної
[1] 1.862899
> plot(data[,1], data[,2], col = col[km.res7$cluster]) # діаграма розсіювання
1 і 2 змінної з кластеризацією
```

Тут вже внутрішньокластерна сума квадратів буде в трохи менше ніж в 2 рази меншою за міжкластерну, тому якась надія на виділення певної структури даних присутня.



Різні розфарбування розташовані не зовсім впорядковано: зелені точки розкидані абсолютно по всій діаграмі. Спробуємо побудувати діаграму розсіювання у просторі перших двох головних компонент.

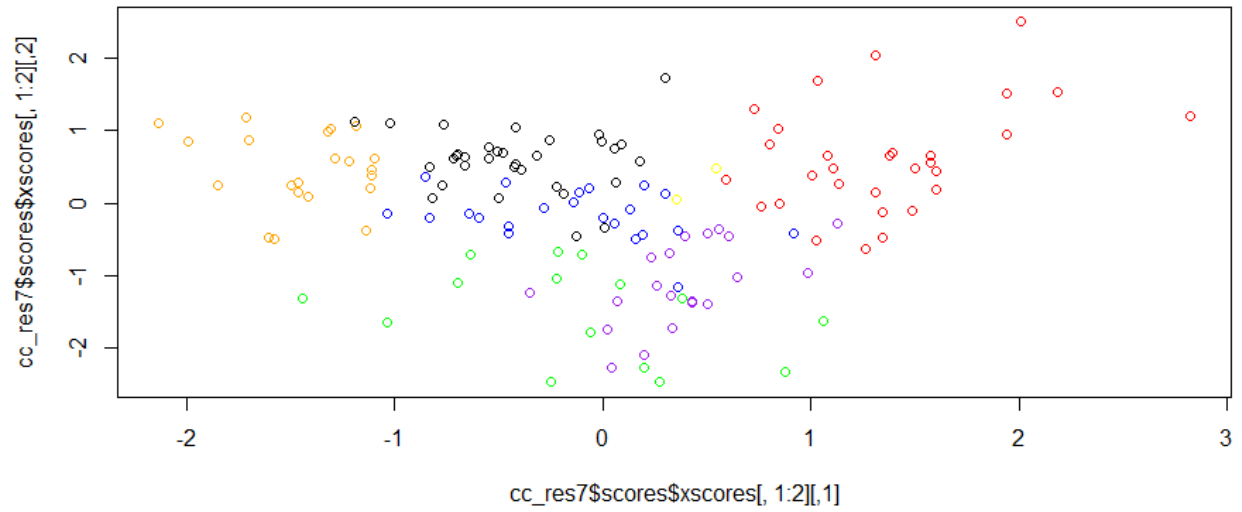
```
> # діаграма розсіювання даних у просторі перших двох головних компонент
> plot(princomp(data)$scores[,1:2],col=col[km.res7$cluster],cex=1)
```



Тут в певній мірі прослідковується вже певна структура, і дійсно, в певній мірі таке розбиття має місце бути. Але, тут в певній області знизу серед фіолетових розкидані як сині, так і зелені точки.

Тепер побудуємо діаграму розсіювання у просторі перших двох канонічних компонент.

```
> cl7 <- km.res7$cluster
> k <- length(levels(as.factor(cl7)))
> C <- matrix(data = as.numeric(rep(cl7, k) == rep(1:k, each = n)), ncol = k,
nrow = n)
> cc_res7 <- rcc(data,C,0.1,0.1)
> # діаграма розсіювання перших двох канонічних компонент
> plot(cc_res7$scores$xscores[,1:2], col = col[cl7], cex=1)
```



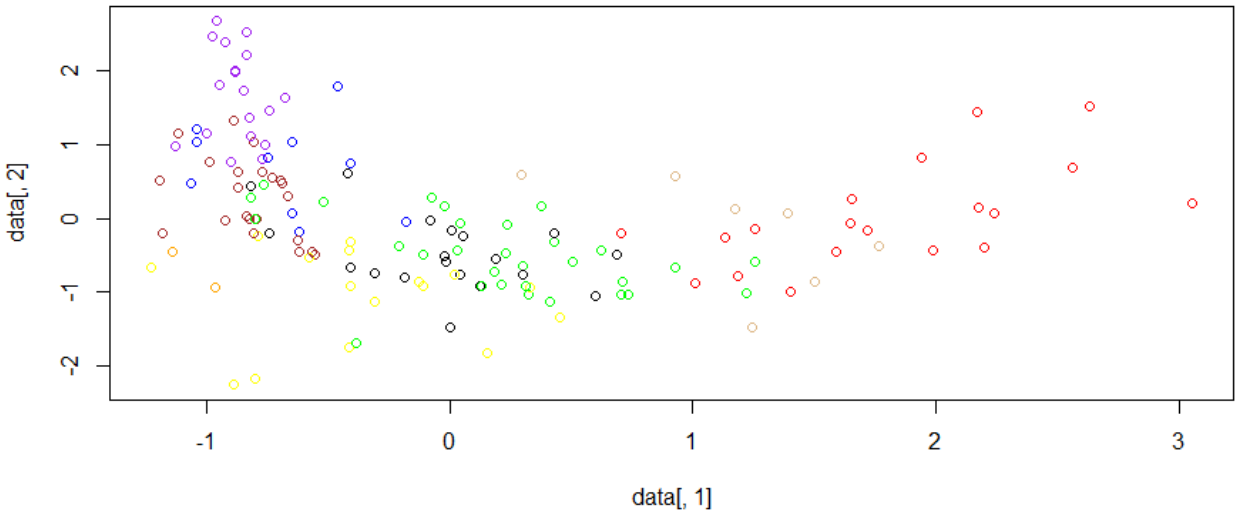
Знову ж таки: серед фіолетових точок присутні як зелені, так і сині, і, здається, навіть червона є. Тому, на мою думку, доцільно розглянути більшу кількість кластерів.

Спробуємо $k = 9$.

```
> km.res9 <- kmeans(data, 9, nstart = 25)
> km.res9$betweenss/km.res9$tot.withinss # відношення міжкластерної суми квад
ратів до внутрішньокластерної
[1] 2.439546
```

Тепер маємо те, що внутрішньокластерна сума квадратів майже в 2.5 рази менша за міжкластерну, тому можемо сподіватись на ще більшу виділеність певної структури даних.

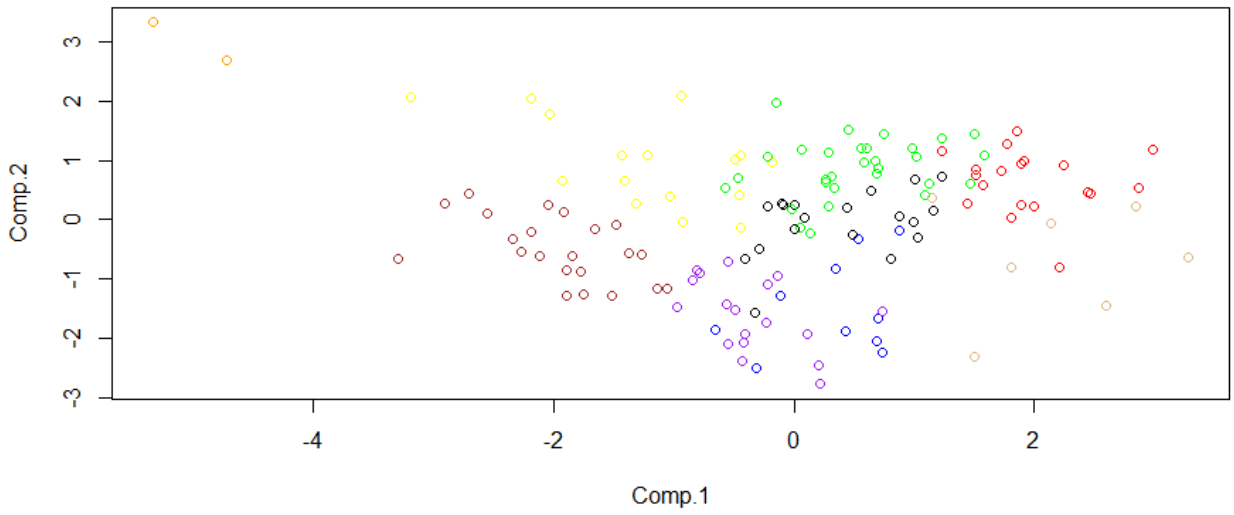
```
> plot(data[,1], data[,2], col = col[km.res9$cluster]) # діаграма розсіювання 1 і 2 змінної з кластеризацією
```



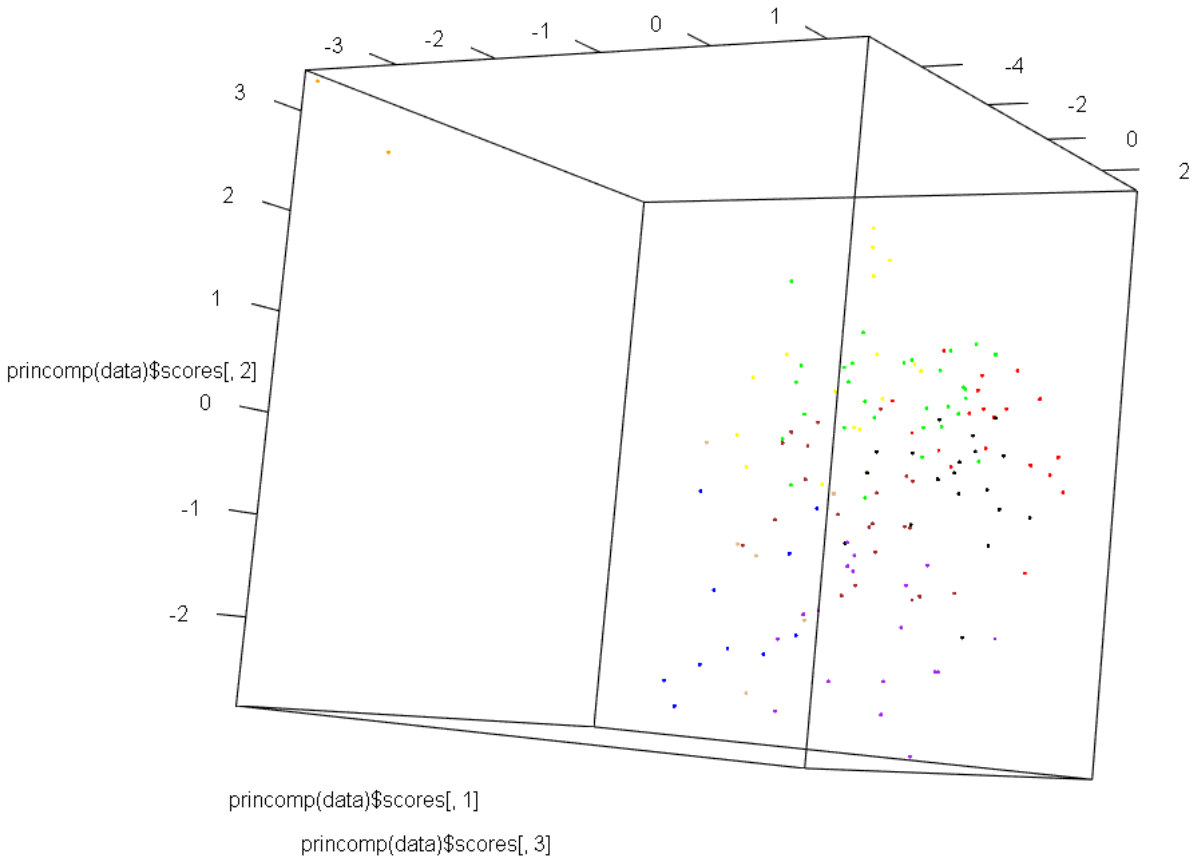
Знову ж таки на перших двох змінних діаграма розсіювання не виглядає дуже гарно. Але, можливо, я забагато вимагаю від перших двох змінних (доступність до електроенергії і показник народжуваності), цілком може бути, що природа цих даних така, що зумовлює таку несприятливу ілюстрацію.

Подивимось на діаграму розсіювання у просторі перших двох головних компонент.

```
> # діаграма розсіювання даних у просторі перших двох головних компонент
> plot(princomp(data)$scores[,1:2],col=col[km.res9$cluster],cex=1)
```

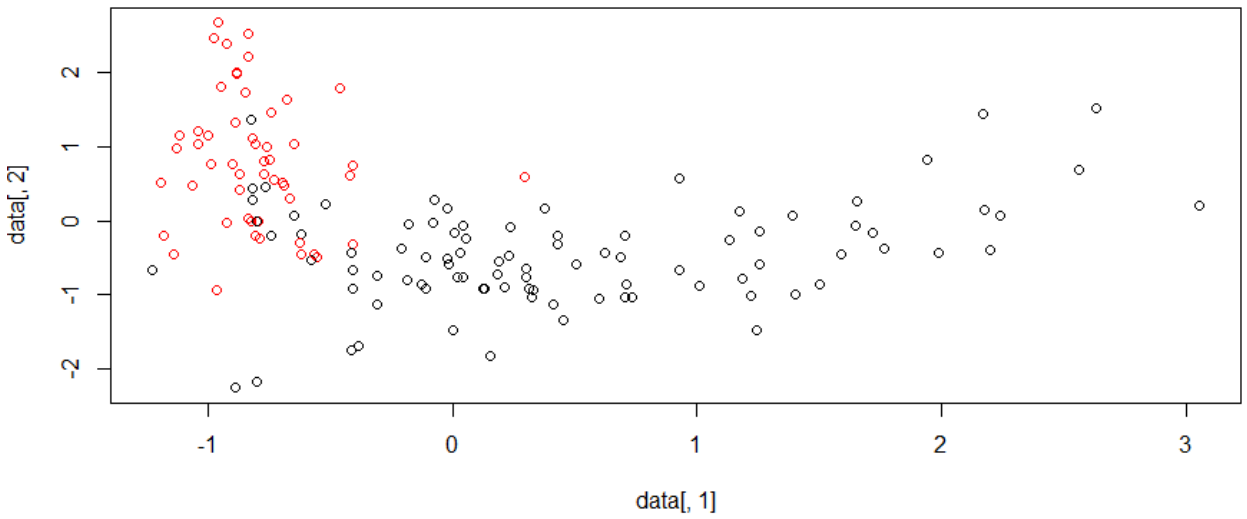


Тут аналогічна ситуація минулій: знову ж таки маємо фіолетові точки серед синіх, причому вони достатньо сильно перемішані. Виділити якусь структуру на тривимірній діаграмі також не вдалось можливим.



Тепер розглянемо метод медоїдів. Тут перебиратимемо серед $k = 2, 4, 9$.

```
> pam.res2 <- pam(data, 2)
> plot(data[,1], data[,2], col = col[pam.res2$cluster]) # діаграма розсіюванн
я перших двох змінних з кластеризацією
```



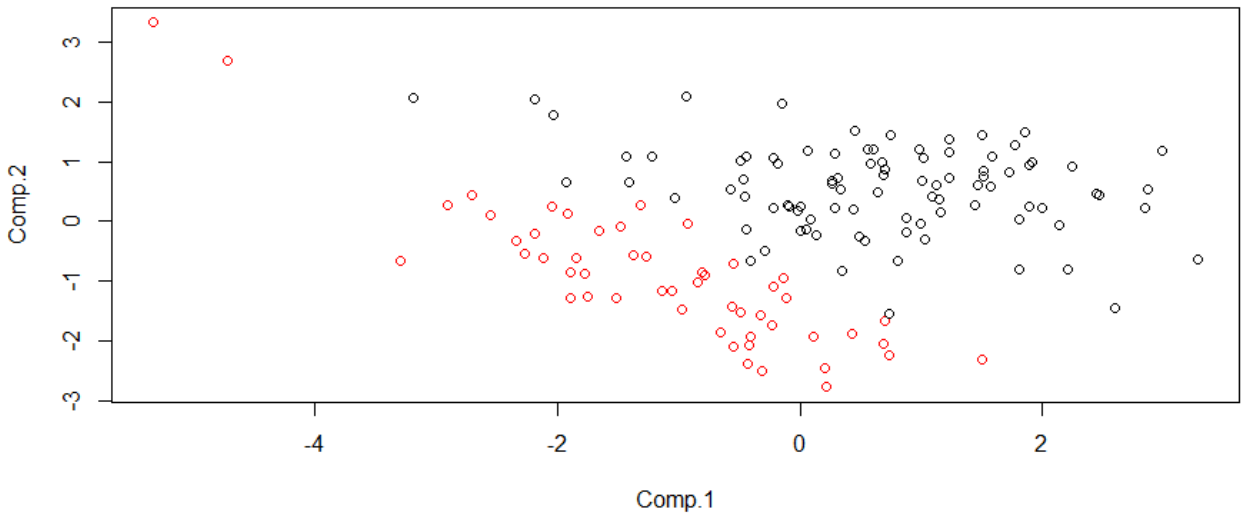
Не дуже гарна, на мою думку, картина. Порівняємо з відповідною кластеризацією, отриманою методом центроїдів.

```
> rand.index(pam.res2$clustering,km.res2$cluster)
[1] 0.7476608
```

Отримали значення індексу Ренда в 74%, що значить, що кластеризації мають певні в собі відмінності.

Побудуємо для даної кластеризації діаграму розсіювання у просторі перших двох головних компонент.

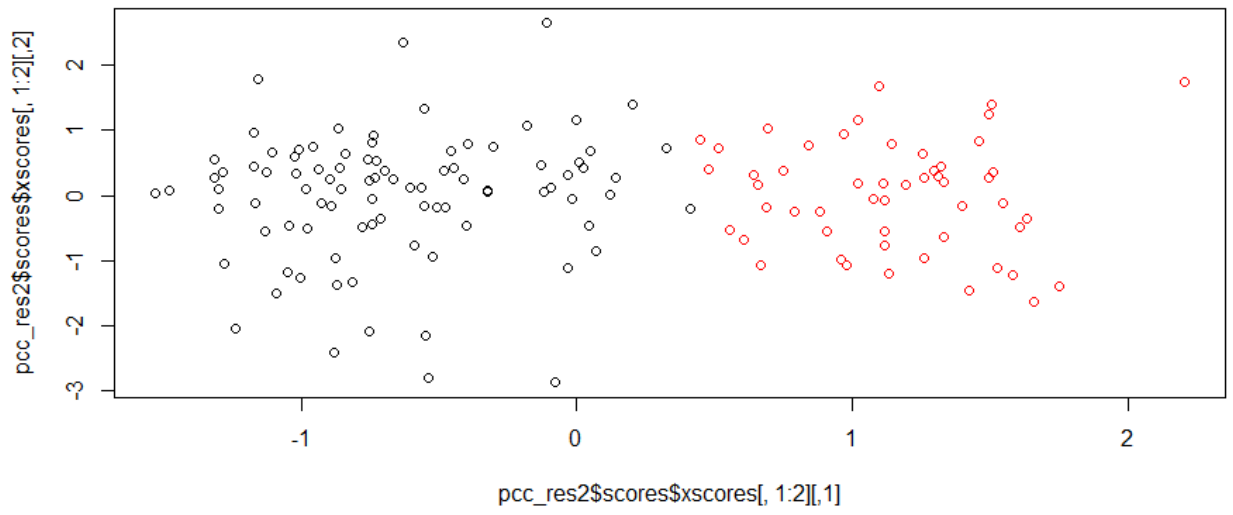
```
> # діаграма розсіювання даних у просторі перших двох головних компонент
> plot(princomp(data)$scores[,1:2],col=col[pam.res2$cluster],cex=1)
```



Точки розташовані щільно одна до одної, в певній мірі, тому, можливо, десь можна було би розбити на ще якусь групку кластерів, але на яку кількість вказати однозначно важко.

Тепер побудуємо діаграму розсіювання у просторі перших двох канонічних компонент.

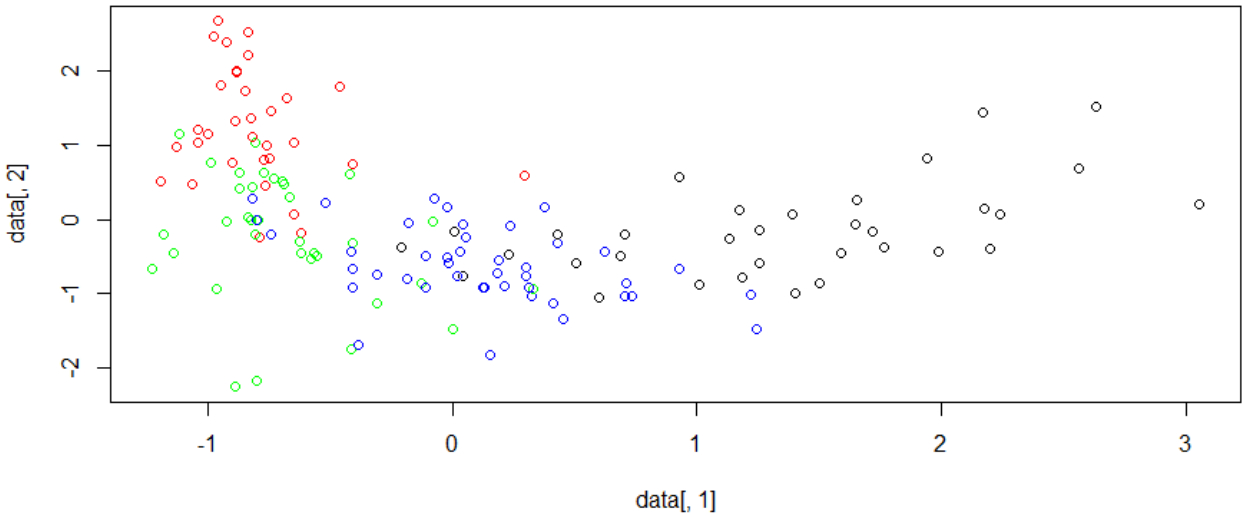
```
> pcl2 <- pam.res2$cluster
> k <- length(levels(as.factor(pcl2)))
> C <- matrix(data = as.numeric(rep(pcl2, k) == rep(1:k, each = n)), ncol = k
, nrow = n)
> pcc_res2 <- rcc(data,C,0.1,0.1)
> # діаграма розсіювання перших двох канонічних компонент
> plot(cc_res2$scores$xscores[,1:2], col = col[cl2], cex=1)
```



Виглядає цікаво, але, здається, можна було би продовжити розбиття в деяких місцях і далі… Але це не точно.

Розглянемо $k = 4$.

```
> pam.res4 <- pam(data, 4)
> plot(data[,1], data[,2], col = col[pam.res4$cluster]) # діаграма розсіюванн
я перших двох змінних з кластеризацією
```

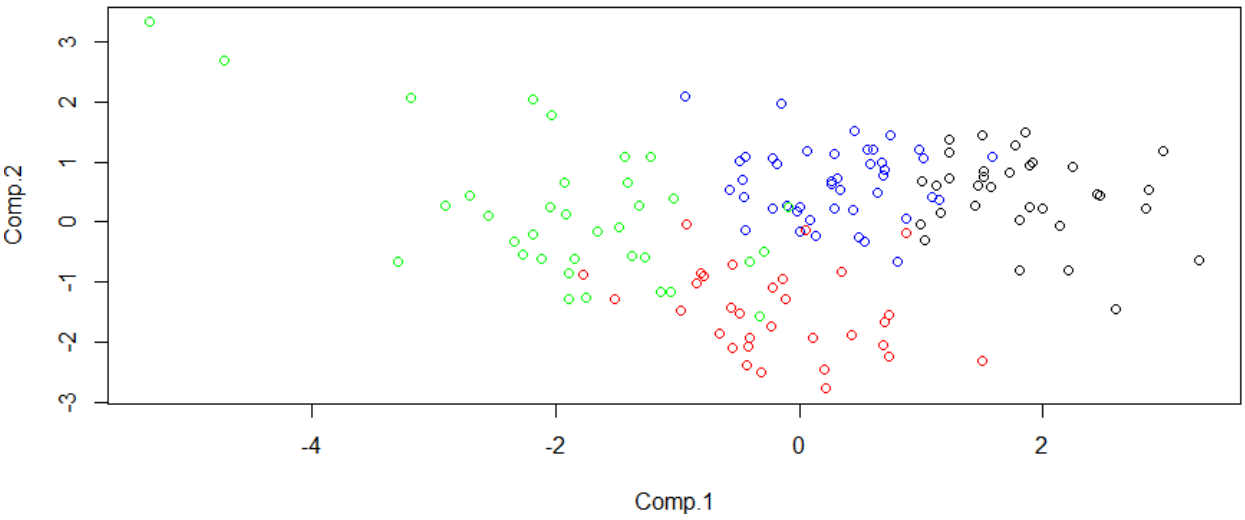


Тут можемо бачити що сині і чорні точки дуже сильно перемішані. Порівняємо з аналогічною кластеризацією, отриманою методом центроїдів.

```
> rand.index(pam.res4$clustering,km.res4$cluster)
[1] 0.7729735
```

Знову ж таки можемо бачити, що кластеризації відрізняються одна від одної, тому далі побудуємо діаграму розсіювання даних у просторі перших двох головних компонент.

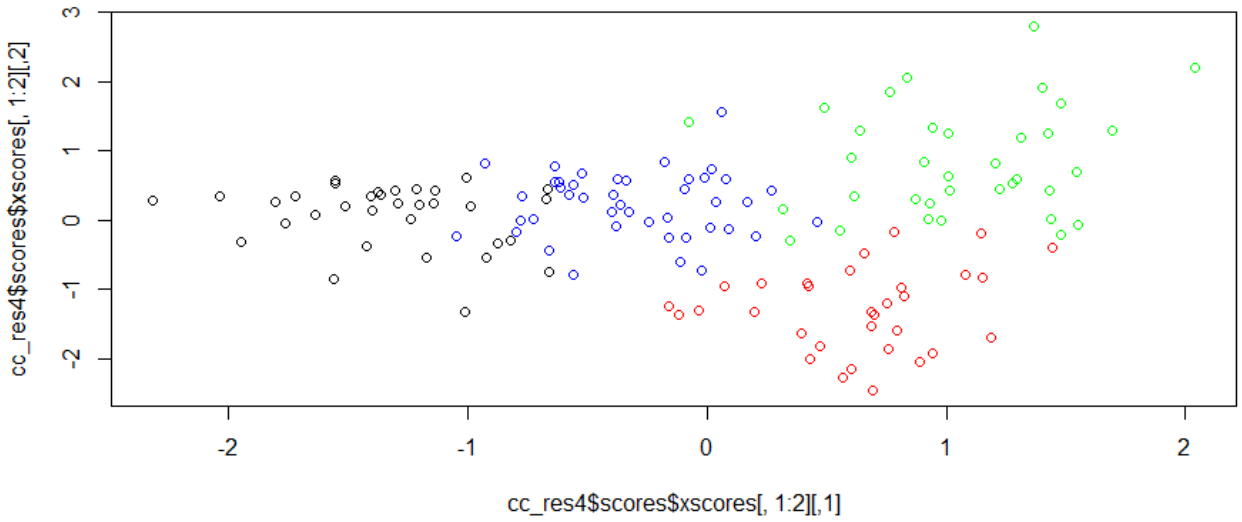
```
> # діаграма розсіювання даних у просторі перших двох головних компонент
> plot(princomp(data)$scores[,1:2],col=col[pam.res4$cluster],cex=1)
```



Виглядає, на мою думку, непогано. Має місце бути…

Тепер побудуємо діаграму розсіювання в просторі перших двох канонічних компонент.

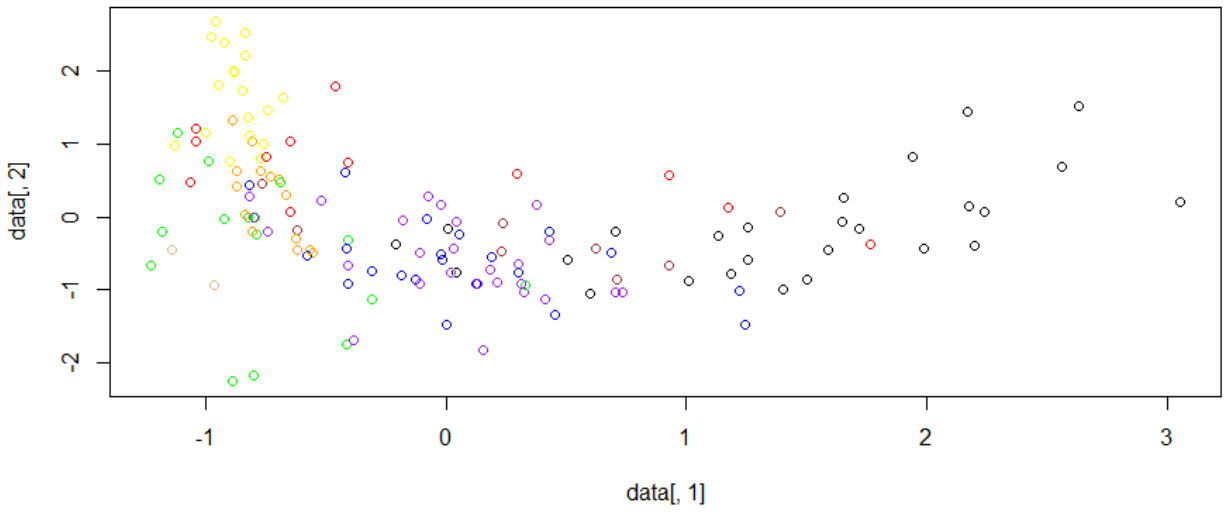
```
> pcl4 <- pam.res4$cluster
> k <- length(levels(as.factor(pcl4)))
> C <- matrix(data = as.numeric(rep(pcl4, k) == rep(1:k, each = n)), ncol = k
, nrow = n)
> cc_res4 <- rcc(data,C,0.1,0.1)
> # діаграма розсіювання перших двох канонічних компонент
> plot(cc_res4$scores$xscores[,1:2], col = col[pcl4], cex=1)
```



Здається розбиття адекватним, але все таки чорні і сині точки явно перемішані по центру.

Далі розглянемо $k = 9$.

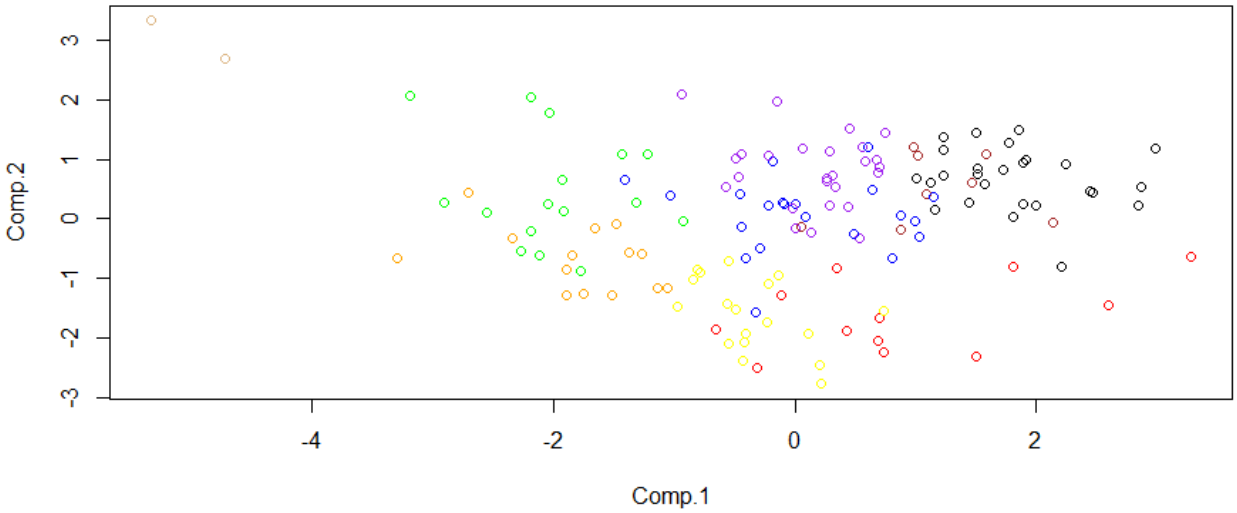
```
> pam.res9 <- pam(data, 9)
> plot(data[,1], data[,2], col = col[pam.res9$cluster]) # діаграма розсіюванн
я перших двох змінних з кластеризацією
```



Тут можемо спостерігати, що червоні точки розміщені вздовж всієї діаграми, що точно не може свідчити на користь їх належності одному кластеру.

Спробуємо поглянути на діаграму розсіювання у просторі перших двох головних компонент.

```
> # діаграма розсіювання даних у просторі перших двох головних компонент
> plot(princomp(data)$scores[,1:2],col=col[pam.res9$cluster],cex=1)
```



Тут можемо бачити, як серед чорних точок знаходяться коричневі (мабуть), а серед фіолетових – сині. Тому, на мою думку, кластеризація з такою кількістю не буде доцільною в даному випадку.

Отже, в якості фінального варіанту, виходячи з діаграм розсіювання у просторі головних і канонічних компонент, я би зупинився на кластеризації, отриманій методом медоїдів для $k = 4$. Поглянемо на її результати.

```
> table(pam.res4$clustering)

 1  2  3  4
33 32 35 43
```

Бачимо, що дана кластеризація більш-менш рівномірно розбила країни на 4 групи. Подивимось уважніше на представників кожної з них.

```
> which(pam.res4$clustering == 1)
    Afghanistan    Burkina Faso      Bangladesh   Cote d'Ivoire           Ghana     Gambia, The           India
              1              10              11              25              45              46              56
     Kazakhstan           Kenya Kyrgyz Republic        Cambodia         Liberia       Sri Lanka            Mali
             65              66              67              68              72              74              83
       Mongolia      Mauritania          Malawi         Namibia           Niger         Nigeria           Nepal
             87              88              90              92              93              94              98
       Pakistan         Rwanda        Eswatini            Chad            Togo      Tajikistan     Timor-Leste
            101             113             122             124             125             127             128
          Tonga          Uganda      Uzbekistan          Zambia        Zimbabwe
            129             133             137             142             143
```

Отже, до так званої першої групи, згідно з даною кластеризацією, потрапили найменш розвинені країни світу (більшість з яких – представники Африки і Азії). Поки що виглядає адекватно.

Поглянемо на другу групу.

```
> which(pam.res4$clustering == 2)
          Albania           Armenia Bulgaria Bosnia and Herzegovina           Belarus
                2                 5                12                14                15
         Barbados            Cyprus   Czech Republic           Germany             Spain
               19                30                31                32                38
          Estonia           Georgia           Greece           Croatia           Hungary
               39                44                47                53                54
            Italy       St. Lucia        Lithuania            Latvia           Moldova
               61                73                75                77                79
```

```
      North Macedonia              Montenegro                  Poland           Puerto Rico                Portugal
                   82                      86                     105                   106                     107
              Romania      Russian Federation                 Serbia        Slovak Republic                Slovenia
                  111                     112                     118                   119                     120
              Ukraine            South Africa
                  134                     141
```

В цю групу потрапили країни, як можемо бачити, з перехідною (в більшості випадків) економікою (як-от Молдова, Україна чи Росія). Дивує потрапляння в цей же перелік таких країн, як Італія, Німеччина, Польща та країн Прибалтики. Вони все ж таки за рівнем економічного розвитку значно перевершують інших «сусідів по кластеризації».

```
> which(pam.res4$clustering == 3)
United Arab Emirates            Argentina            Australia              Austria              Belgium
                   3                    4                    6                    7                    9
       Bahamas, The    Brunei Darussalam               Canada          Switzerland                Chile
                  13                   20                   21                   22                   23
                Cuba             Denmark              Finland               France       United Kingdom
                  29                   33                   40                   42                   43
               Guam Hong Kong SAR, China              Ireland              Iceland               Israel
                  49                   51                   57                   59                   60
               Japan          Korea, Rep.               Kuwait           Luxembourg                Malta
                  64                   69                   70                   76                   84
        Netherlands               Norway          New Zealand                Qatar         Saudi Arabia
                  96                   97                   99                  110                  114
          Singapore            San Marino               Sweden              Uruguay        United States
                 115                  117                  121                  135                  136
```

В третій групі можемо бачити країни ще більшого рівня розвитку, як-от США, Японія і Сінгапур. Але в той же час тут присутні Куба (країна планової економіки, взагалі кажучи), Аргентина (країна з високою інфляцією) чи Уругвай.

Поглянемо на останню групу.

```
> which(pam.res4$clustering == 4)
        Azerbaijan              Belize             Bolivia              Brazil               China
                 8                  16                  17                  18                  24
          Colombia          Cabo Verde          Costa Rica  Dominican Republic             Algeria
                26                  27                  28                  34                  35
           Ecuador    Egypt, Arab Rep.                Fiji           Guatemala              Guyana
                36                  37                  41                  48                  50
          Honduras           Indonesia                Iraq             Jamaica             Jordan
                52                  55                  58                  62                  63
           Lao PDR             Morocco            Maldives              Mexico             Myanmar
                71                  78                  80                  81                  85
         Mauritius            Malaysia           Nicaragua                Oman              Panama
                89                  91                  95                 100                 102
              Peru         Philippines            Paraguay  West Bank and Gaza         El Salvador
               103                 104                 108                 109                 116
        Seychelles            Thailand Trinidad and Tobago             Tunisia              Turkey
               123                 126                 130                 131                 132
           Vietnam               World               Samoa
               138                 139                 140
```

В четвертій групі можемо бачити багато острівних країн, а також країн Південної Америки. Також цікаво відмітити, що усереднені світові дані потрапили саме до цієї групки країн.

Отже, кластеризація для такого набору даних виявилась цілком цікавою, і незважаючи на те, що в даних була присутня значна доля демографічних показників, кластеризація частково, можна сказати, відбулася за географічною ознакою (в першій групі, наприклад, переважно Африка, а в четвертій – острівні країни і країни Латинської Америки). Проте в той же час, географічне розташування і зумовлює, історично, демографічні і соціально-економічні показники…