ЗВІТ З ЛАБОРАТОРНОЇ РОБОТИ №3
«ІЄРАРХІЧНА КЛАСИФІКАЦІЯ (КЛАСТЕРИЗАЦІЯ)»

частина 1

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

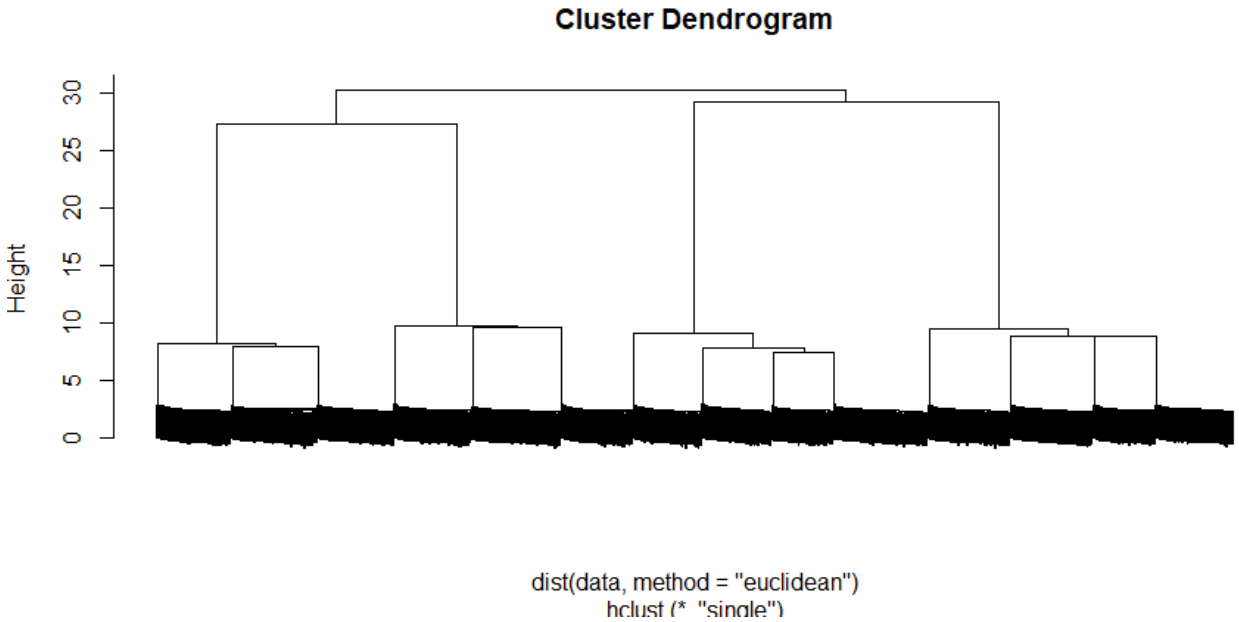
Для змодельованих даних першої частини першої роботи застосуємо техніку ієрархічної кластеризації для відшукання відповіді на питання якою доцільною має бути кількість кластерів.

Спершу зчитаємо змодельовані дані та підключимо бібліотеки.

```
> # підключимо бібліотеки
> library(dendextend)
> # зчитуємо дані
> data <- read.table('C:\\Users\\Razor\\Desktop\\дистанційне навчання\\статис
тичний аналіз багатовимірних даних\\lab3\\mult6.txt')
```

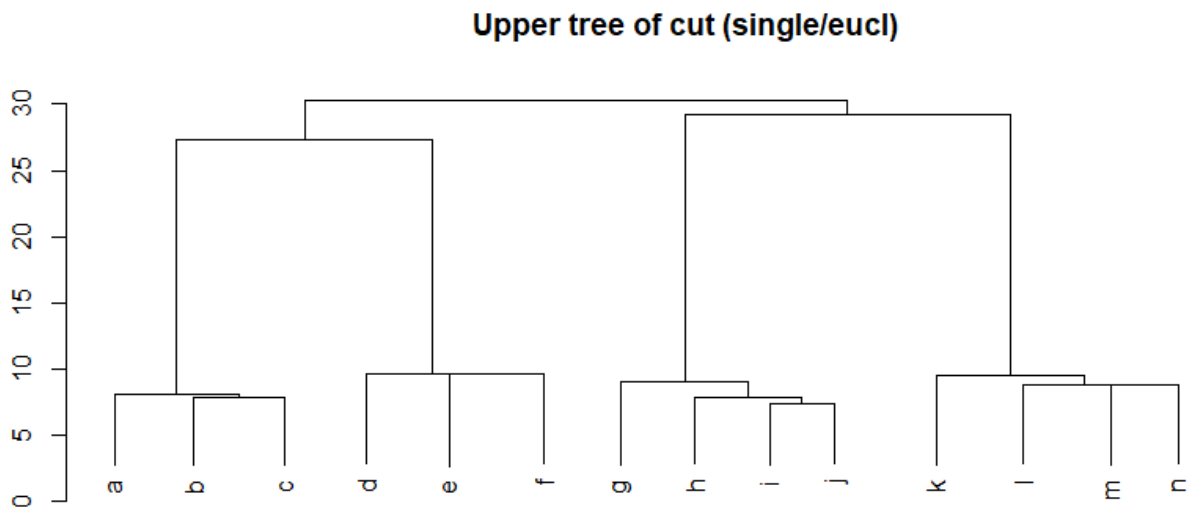
Спершу використаємо метод одного зв’язку та евклідову відстань.

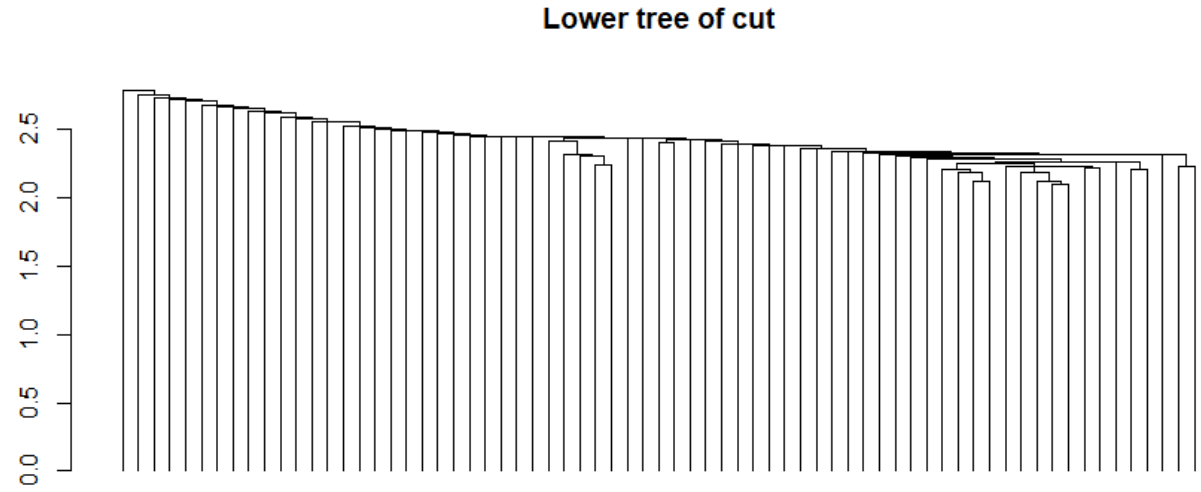
```
> # застосуємо ієрархічну кластеризацію для методу одного зв'язку і евклідово
ї відстані
> h_single_eucl <- hclust(dist(data, method = 'euclidean'), method = 'single'
)
> plot(h_single_eucl, labels = FALSE)
```



В цілому, помітне розділення на 4 або 14 кластерів. Доцільно розділити дану дендрограму на дві частини на рівні $h \approx 5$.

```
> h_single_eucl_up <- cut(as.dendrogram(h_single_eucl), h = 5)$upper
> labels(h_single_eucl_up) <- letters[1:14]
> plot(h_single_eucl_up, main = 'Upper tree of cut (single/eucl)')
> h_single_eucl_low <- cut(as.dendrogram(h_single_eucl), h = 5)$lower[[1]]
> labels(h_single_eucl_low) <- NULL
> plot(h_single_eucl_low, main = 'Lower tree of cut')
```

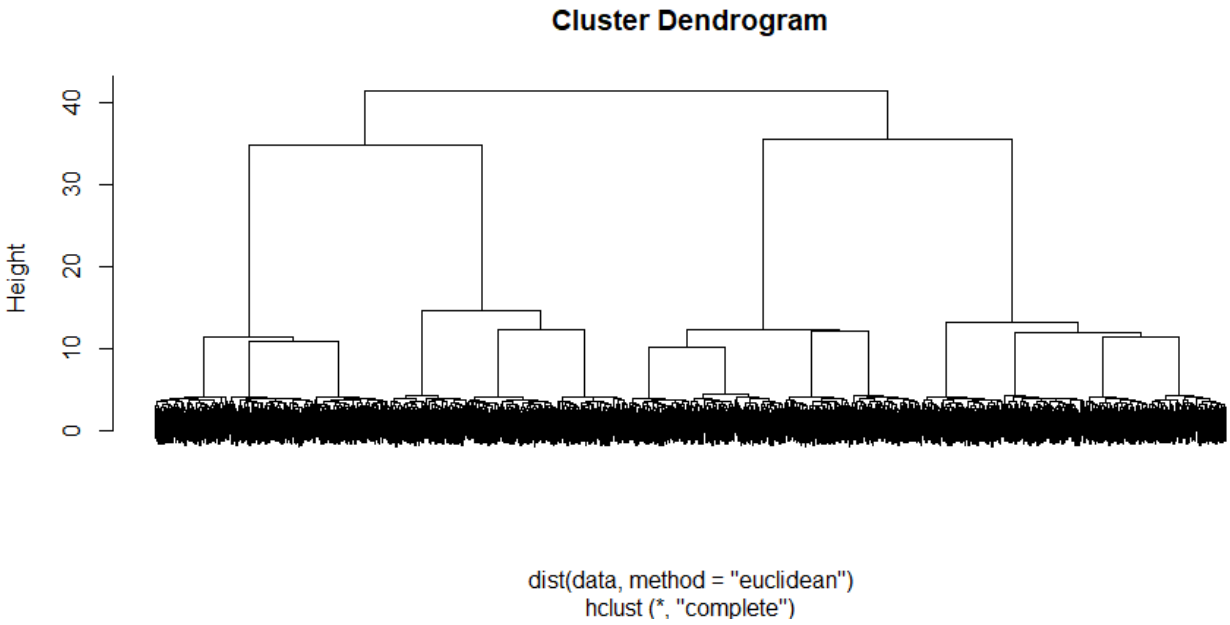




Тут можемо бачити на верхньому рисунку як доцільно виділяються 14 кластерів. Про це ж і свідчить нижній рисунок, на якому об'єкти розташовуються на однакових відстанях всередині кластера.

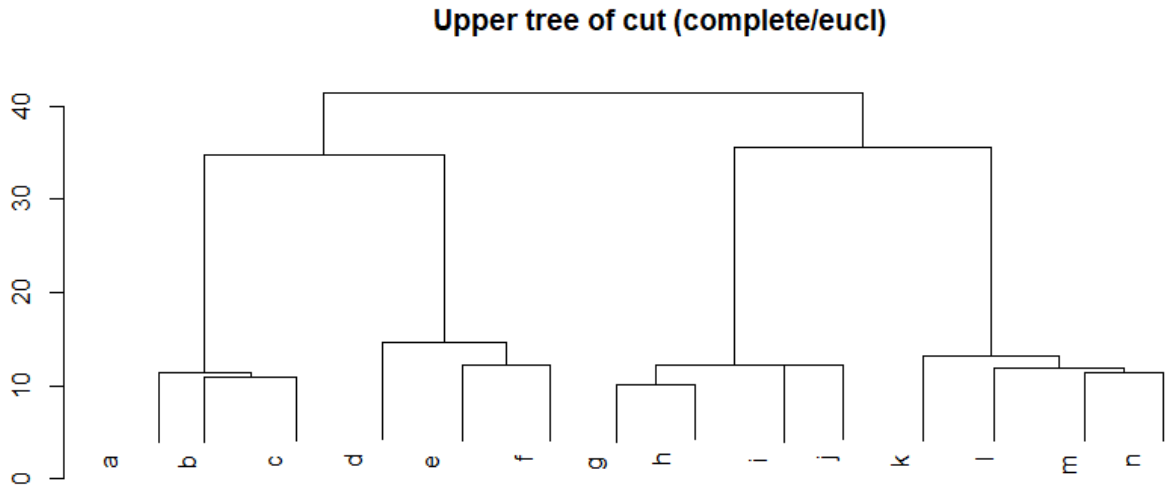
Розглянемо метод повного зв'язку.

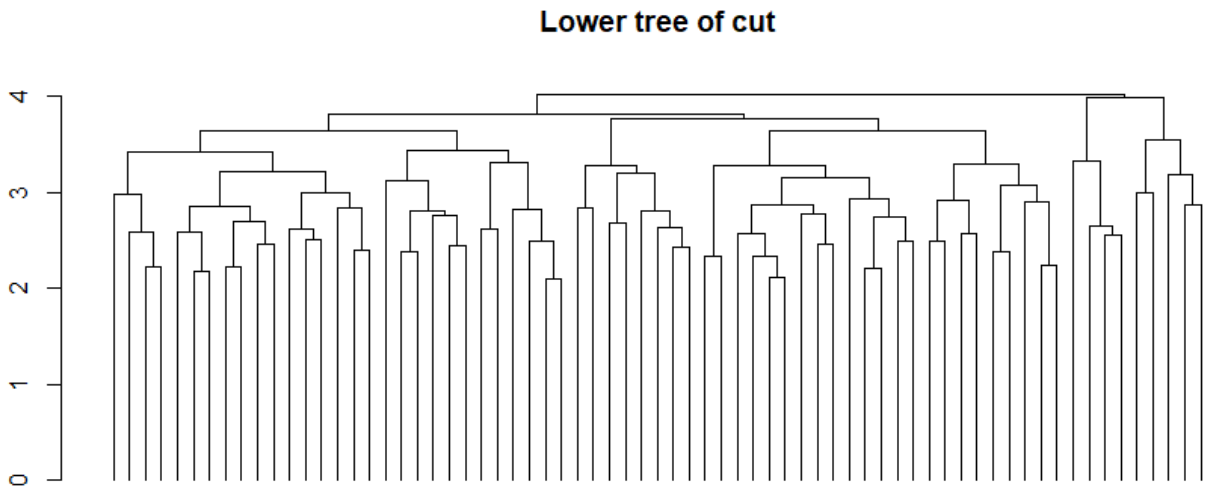
```
> # застосуємо ієрархічну кластеризацію для методу повного зв'язку і евклідов
ої відстані
> h_compl_eucl <- hclust(dist(data, method = 'euclidean'), method = 'complete
')
> plot(h_compl_eucl, labels = FALSE)
```



В цілому, достатньо схоже на минулий випадок – або 4 кластери, або на рівні $h \approx 7$ (здається) помітне чітке розділення на знову ж таки 14 кластерів. Розіб'ємо ще раз на верхню та нижню частини.

```
> h_compl_eucl_up <- cut(as.dendrogram(h_compl_eucl), h = 7)$upper
> labels(h_compl_eucl_up) <- letters[1:14]
> plot(h_compl_eucl_up, main = 'Upper tree of cut (complete/eucl)')
> h_compl_eucl_low <- cut(as.dendrogram(h_compl_eucl), h = 7)$lower[[1]]
> labels(h_compl_eucl_low) <- NULL
> plot(h_compl_eucl_low, main = 'Lower tree of cut')
```

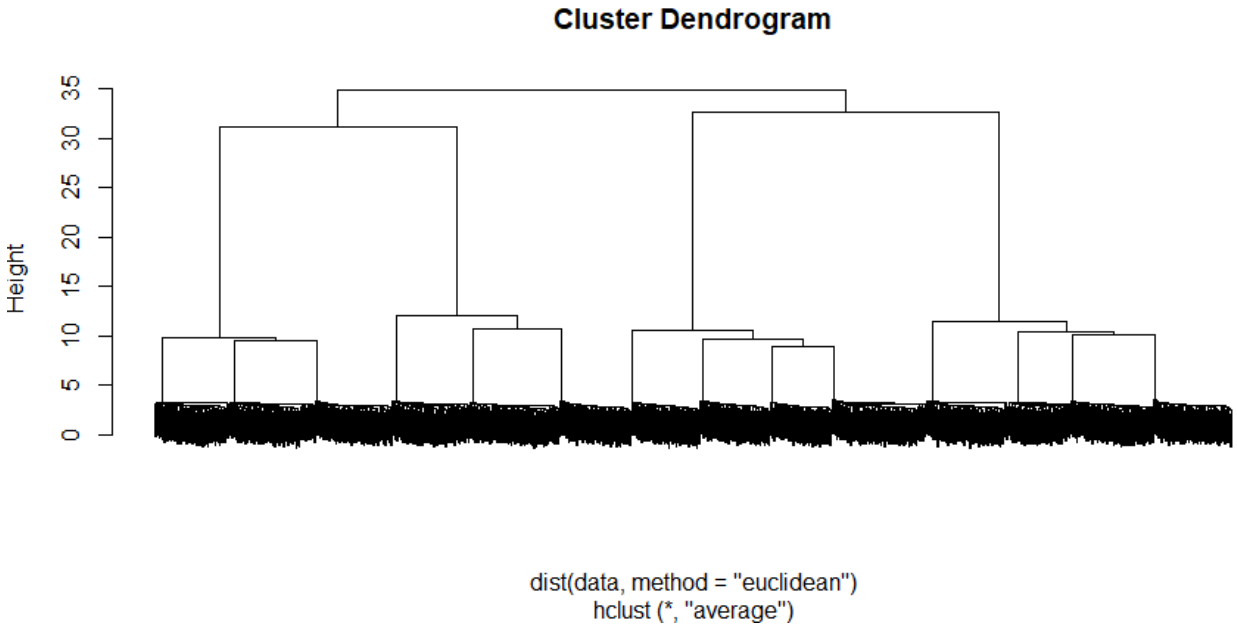




Хоча нижнє дерево розділення виглядає дещо інакше, ніж для методу одного зв’язку, але, на мою думку, це не принципова відмінність. З цих двох рисунків я також стверджую чітке виділення 14 кластерів, опускатись по дендрограмі нижче сенсу не має.

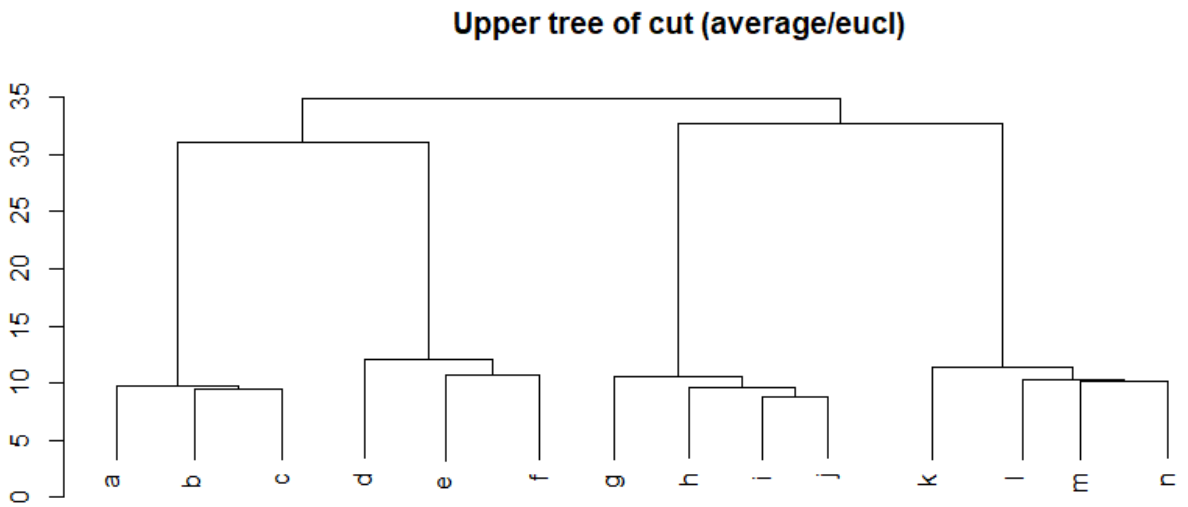
І наприкінці застосуємо метод середнього зв’язку.

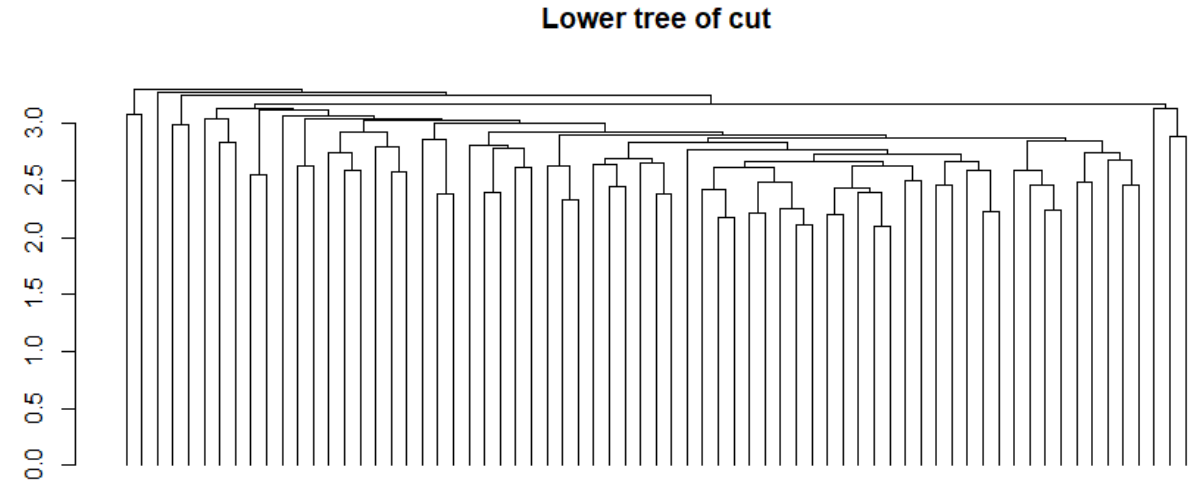
```
> # застосуємо ієрархічну кластеризацію для методу середнього зв'язку і евклі
дової відстані
> h_avg_eucl <- hclust(dist(data, method = 'euclidean'), method = 'average')
> plot(h_avg_eucl, labels = FALSE)
```



Аналогічно минулим випадкам картинка свідчить на користь 4 або 14 кластерів на рівні $h \approx 6$. «Обріжемо» дендрограму на цьому приблизному рівні.

```
> h_avg_eucl_up <- cut(as.dendrogram(h_avg_eucl), h = 6)$upper
> labels(h_avg_eucl_up) <- letters[1:14]
> plot(h_avg_eucl_up, main = 'Upper tree of cut (average/eucl)')
> h_avg_eucl_low <- cut(as.dendrogram(h_avg_eucl), h = 7)$lower[[1]]
> labels(h_avg_eucl_low) <- NULL
> plot(h_avg_eucl_low, main = 'Lower tree of cut')
```

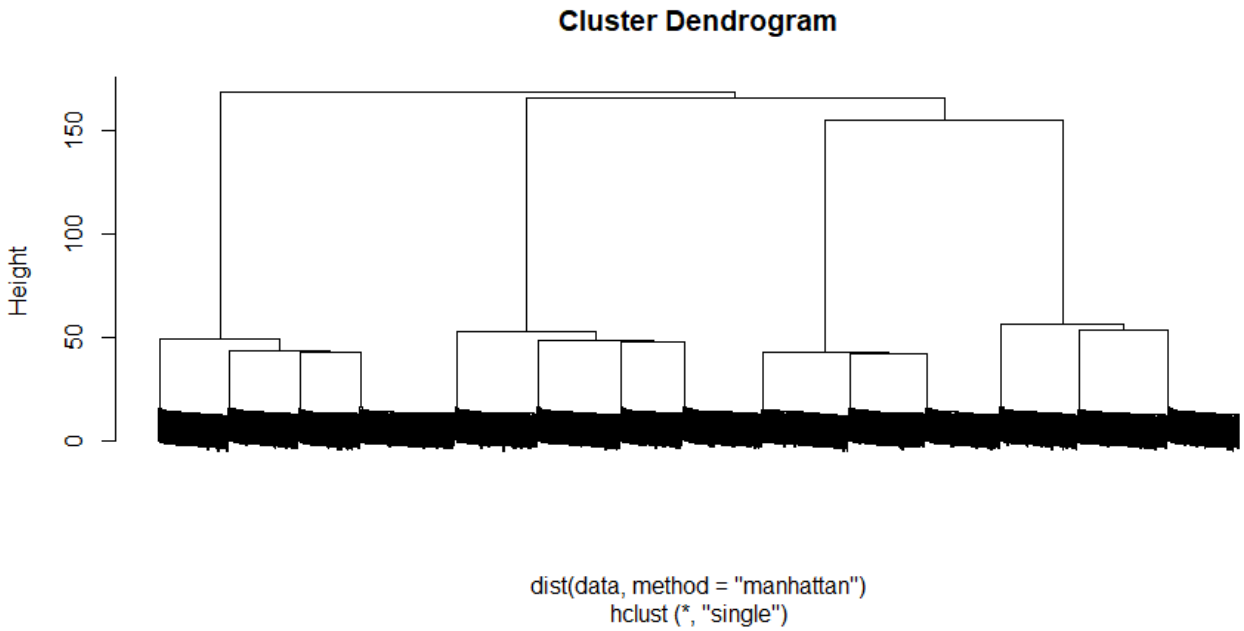




Дуже схожа картинка з минулим випадком (метод повного зв’язку). Ще раз переконуємося в доцільності поділу на 14 кластерів.

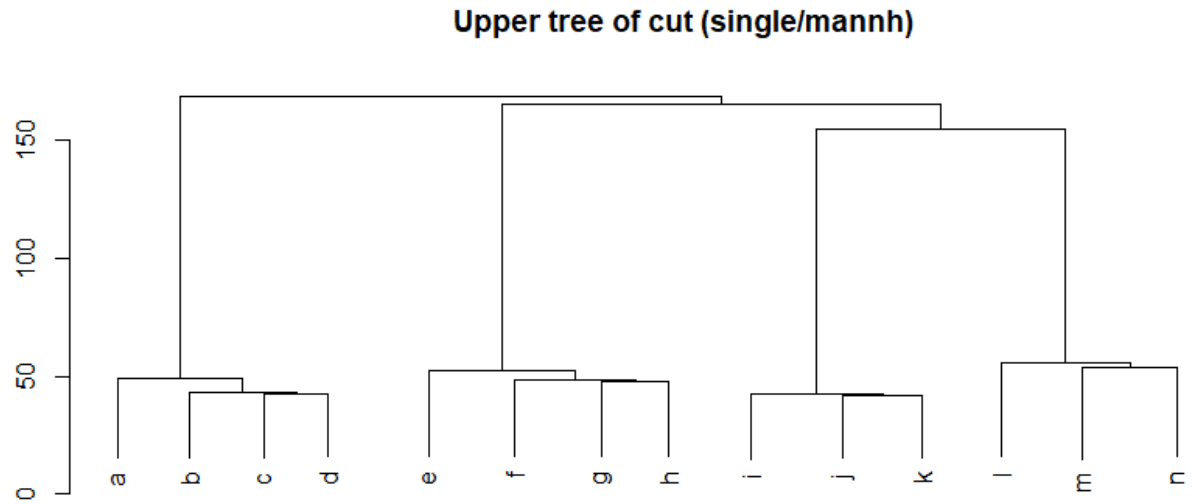
Тепер використаємо всі методи для манхаттанської відстані.

```
> # застосуємо ієрархічну кластеризацію для методу одного зв'язку і манхаттан
ської відстані
> h_single_mannh <- hclust(dist(data, method = 'manhattan'), method = 'single
')
> plot(h_single_mannh, labels = FALSE)
```



В черговий раз можемо бачити або 4 кластери, або, опускаючись дещо нижче, 14. На рівні $h \approx 25$ виконаємо розбиття, аби перевірити можливість такого розбиття на 14 кластерів.

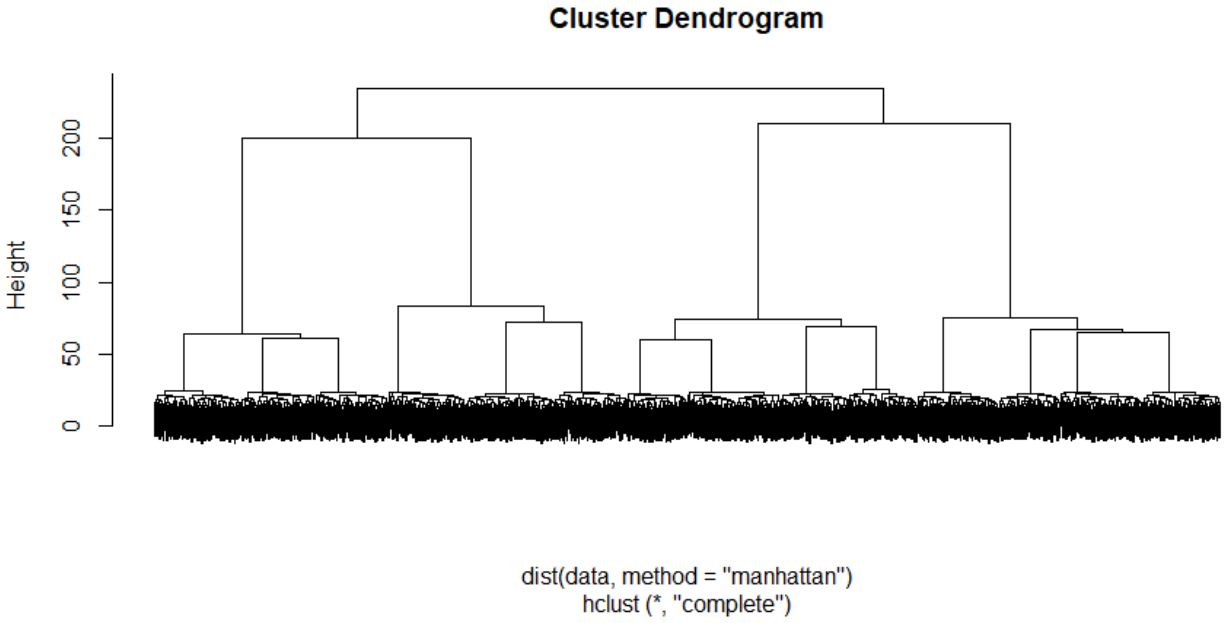
```
> h_single_mannh_up <- cut(as.dendrogram(h_single_mannh), h = 25)$upper
> labels(h_single_mannh_up) <- letters[1:14]
> plot(h_single_mannh_up, main = 'Upper tree of cut (single/mannh)')
> h_single_mannh_low <- cut(as.dendrogram(h_single_mannh), h = 25)$lower[[1]]
> labels(h_single_mannh_low) <- NULL
> plot(h_single_mannh_low, main = 'Lower tree of cut')
```



Цілком аналогічна ситуація: доцільне розбиття як на 4, так і на 14 кластерів.

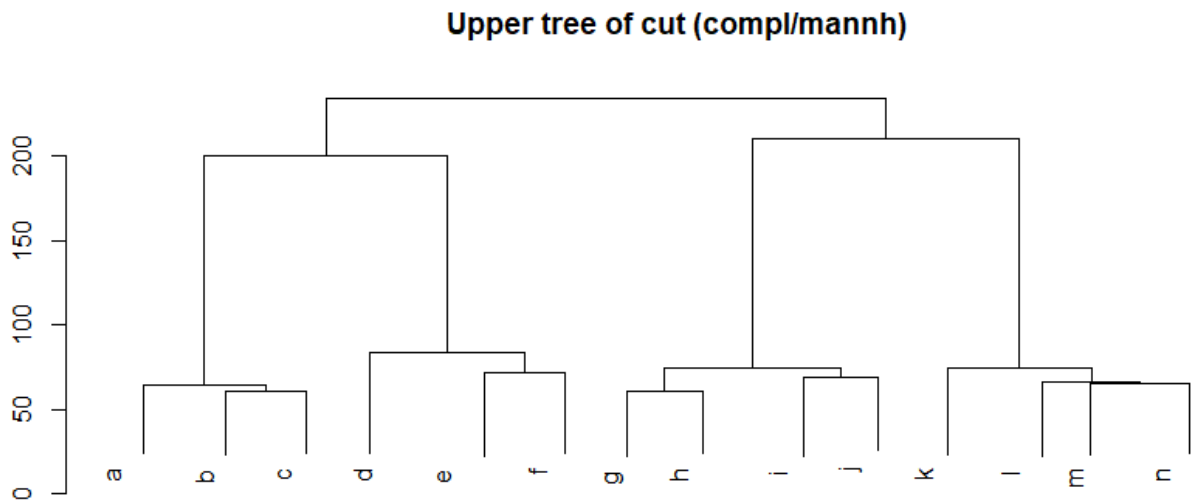
Застосуємо метод повного зв’язку.

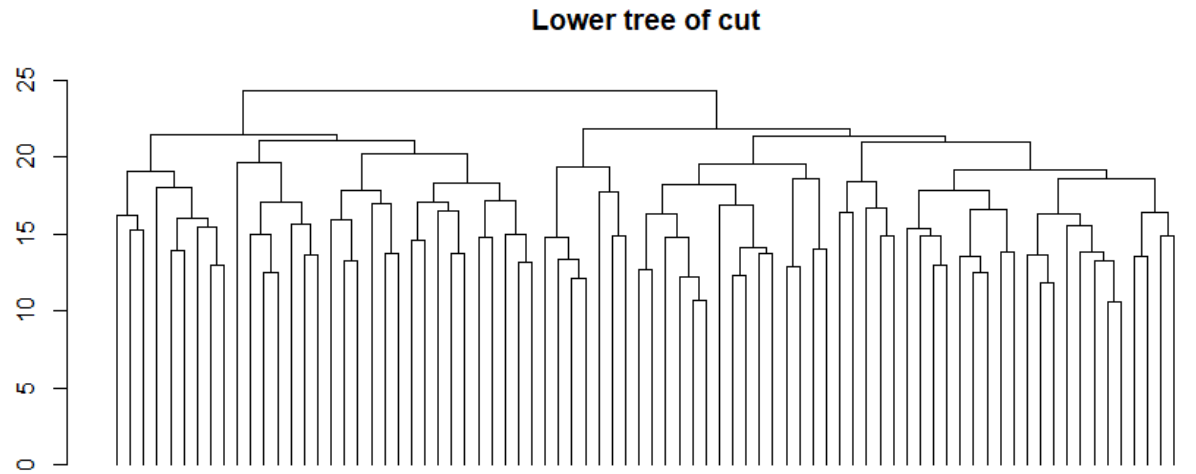
```
> # застосуємо ієрархічну кластеризацію для методу повного зв'язку і манхатта
нської відстані
> h_compl_mannh <- hclust(dist(data, method = 'manhattan'), method = 'complet
e')
> plot(h_compl_mannh, labels = FALSE)
```



Можемо бачити чітке розбиття на 4 кластери, або, на рівні $h \approx 40$ розбиття знову ж таки на 14.

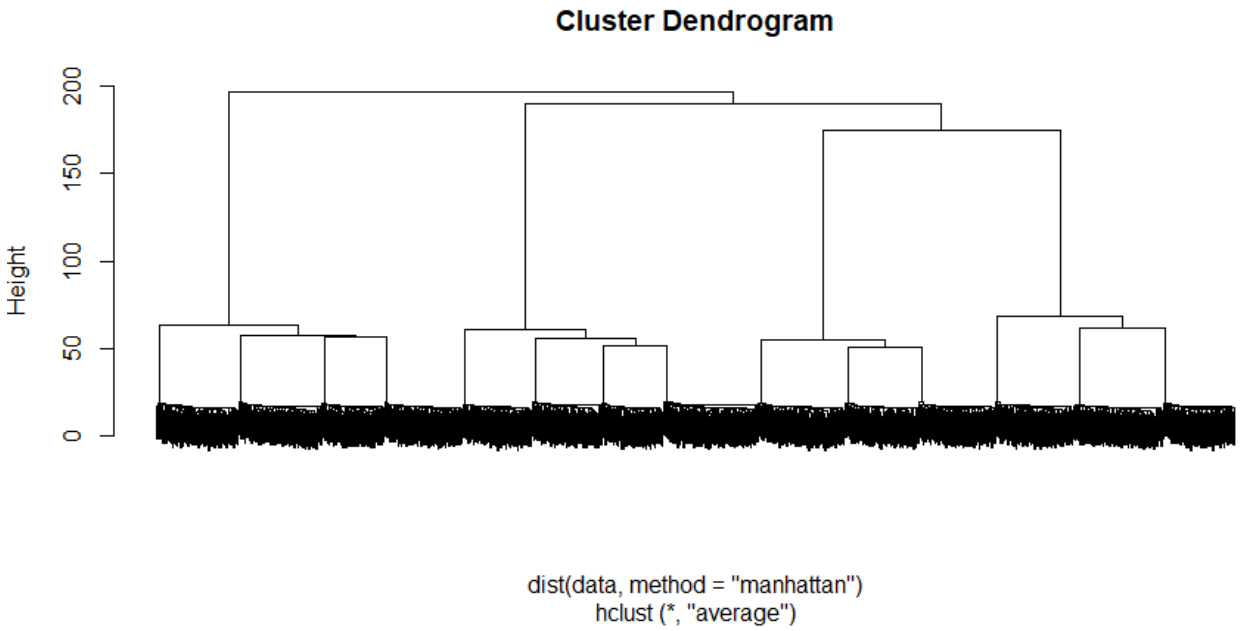
```
> h_compl_mannh_up <- cut(as.dendrogram(h_compl_mannh), h = 40)$upper
> labels(h_compl_mannh_up) <- letters[1:14]
> plot(h_compl_mannh_up, main = 'Upper tree of cut (compl/mannh)')
> h_compl_mannh_low <- cut(as.dendrogram(h_compl_mannh), h = 40)$lower[[1]]
> labels(h_compl_mannh_low) <- NULL
> plot(h_compl_mannh_low, main = 'Lower tree of cut')
```





Так, дійсно, можемо переконатись, що тут таке розбиття має місце бути. Хоча, можливо, один серед цих 14 можна було би поділити ще на 1 (як свідчить нижнє дерево), але згадавши загальну картину це не буде доцільним.

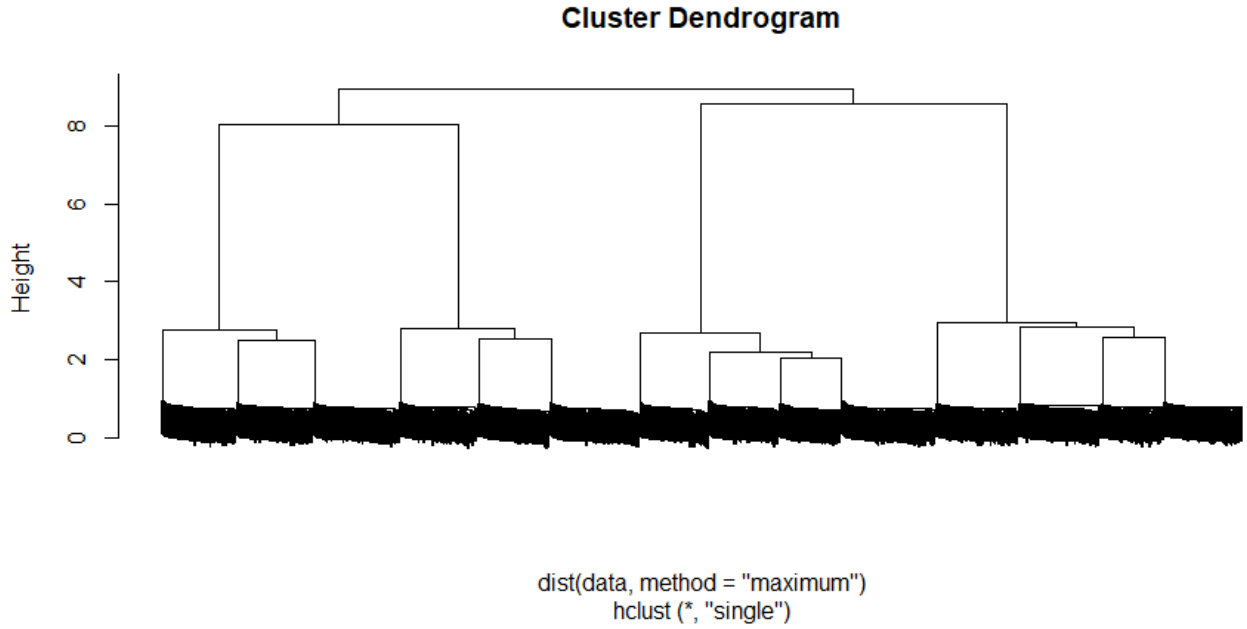
```
> h_avg_mannh <- hclust(dist(data, method = 'manhattan'), method = 'average')
> plot(h_avg_mannh, labels = FALSE)
```



Думаю, тут можемо навіть і не обрізати на два дерева, адже картина цілком і повністю повторює всі минулі ситуації. 4 або 14 кластерів.

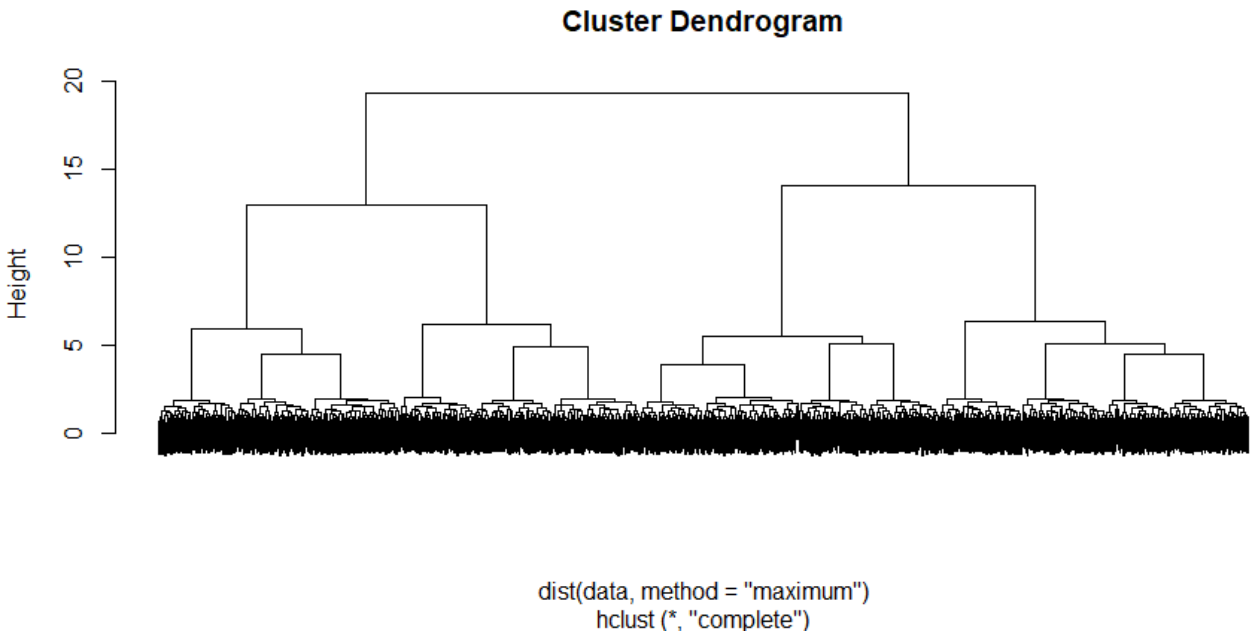
І наприкінці застосуємо максимальну відстань.

```
> # застосуємо ієрархічну кластеризацію для методу одного зв'язку і максималь
ної відстані
> h_single_max <- hclust(dist(data, method = 'maximum'), method = 'single')
> plot(h_single_max, labels = FALSE)
```



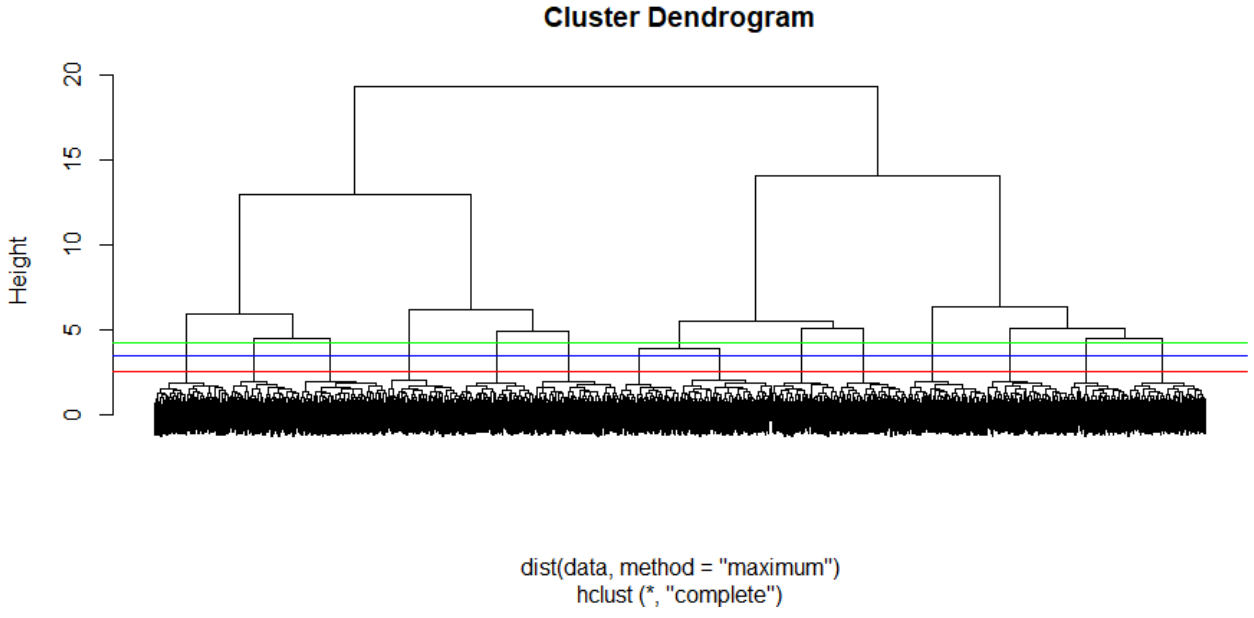
Знову ж таки: 4 або 14. Думаю, необхідності чергового поділу немає.

```
> # застосуємо ієрархічну кластеризацію для методу повного зв'язку і максимал
ьної відстані
> h_compl_max <- hclust(dist(data, method = 'maximum'), method = 'complete')
> plot(h_compl_max, labels = FALSE)
```



Для методу повного зв’язку і максимальної відстані отримали, на цей раз, не дуже певну картину. З однієї сторони, чітко виділяються знову ж таки 4 кластери. Натомість рухаючись вздовж h донизу щоразу можемо приходити до різних висновків.

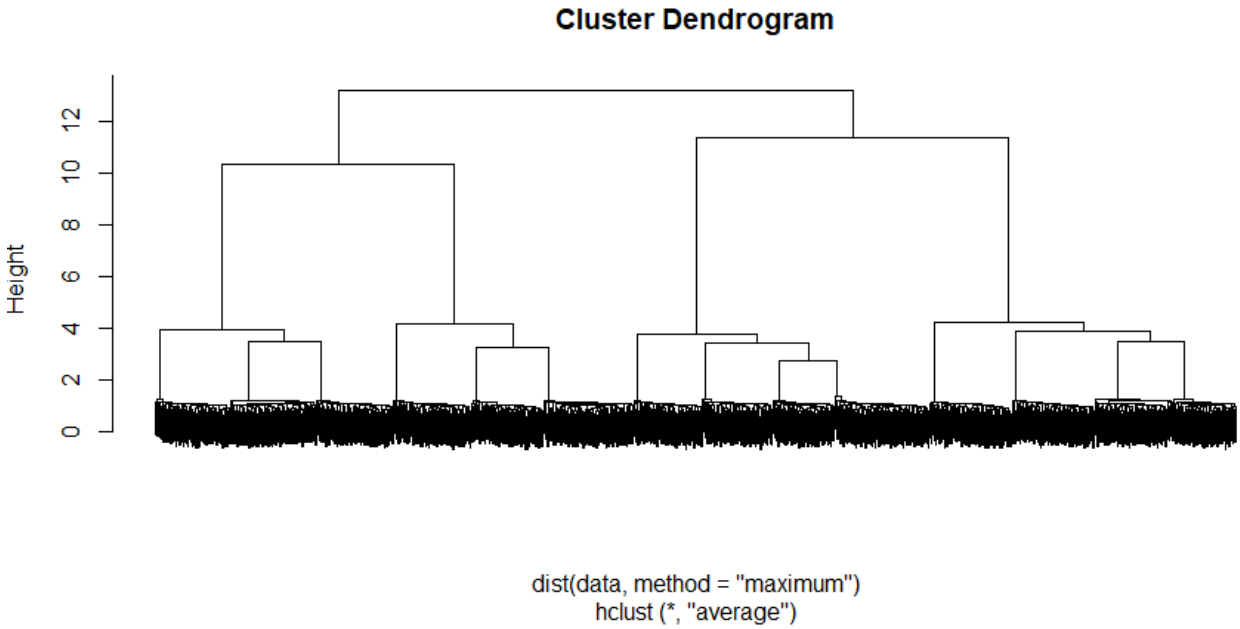
```
> abline(h = 2.5, col = 'red')
> abline(h = 4.25, col = 'green')
> abline(h = 3.5, col = 'blue')
```



Так, наприклад, при $h = 4.5$ на дендрограмі виділяються 13 кластерів, а при ще менших h (як-от 2.5) кластерів знову 14.

І наостанок застосуємо метод середнього зв’язку для максимальної відстані.

```
> # застосуємо ієрархічну кластеризацію для методу середнього зв'язку і макси
мальної відстані
> h_avg_max <- hclust(dist(data, method = 'maximum'), method = 'average')
> plot(h_avg_max, labels = FALSE)
```



Тут знову ж таки, підбираючі різні $h > 2$ кількість кластерів поступово зменшується від 14 до 4. Але, знову ж таки, на око, найчіткіший поділ йде або на 14, або лише на 4.

Отже, використавши різні відстані і різні методи на змодельованих даних, можемо прийти до висновку доцільності поділу на 4 або 14 кластерів.

2 частина

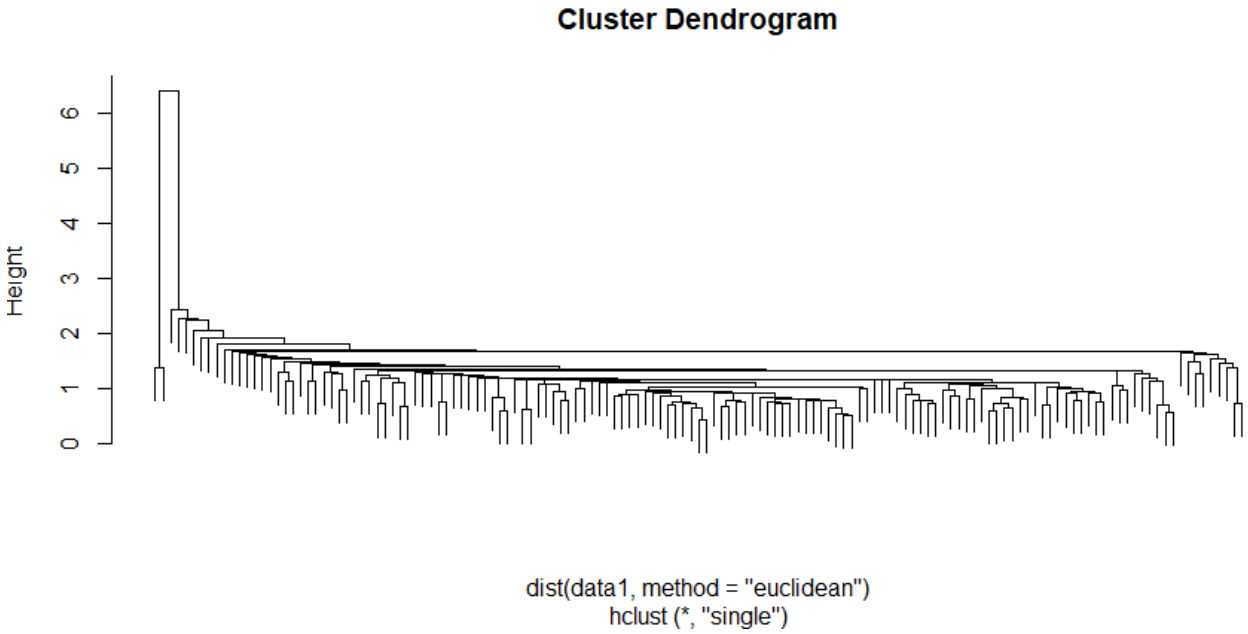
Тепер застосуємо техніку ієрархічної кластеризації для даних 2-ї частини першої лабораторної країни (деякі соціально-економічні дані більшої частини країн світу).

Спершу, звичайно, зчитаємо наші дані і виконаємо їх нормування та центрування.

```
> # зчитуємо дані
> library(readxl)
> data1 <- read_excel('data.xlsx')
> rows <- t(data1[,1])
> data1 <- data1[,-1]
>
> # центрування і нормування
>
> data1 <- as.data.frame(scale(data1))
> row.names(data1) <- rows
```

Для цих даних проробимо ті ж самі дії, що і в першій частині. Почнемо з евклідової відстані і методу одного зв'язку.

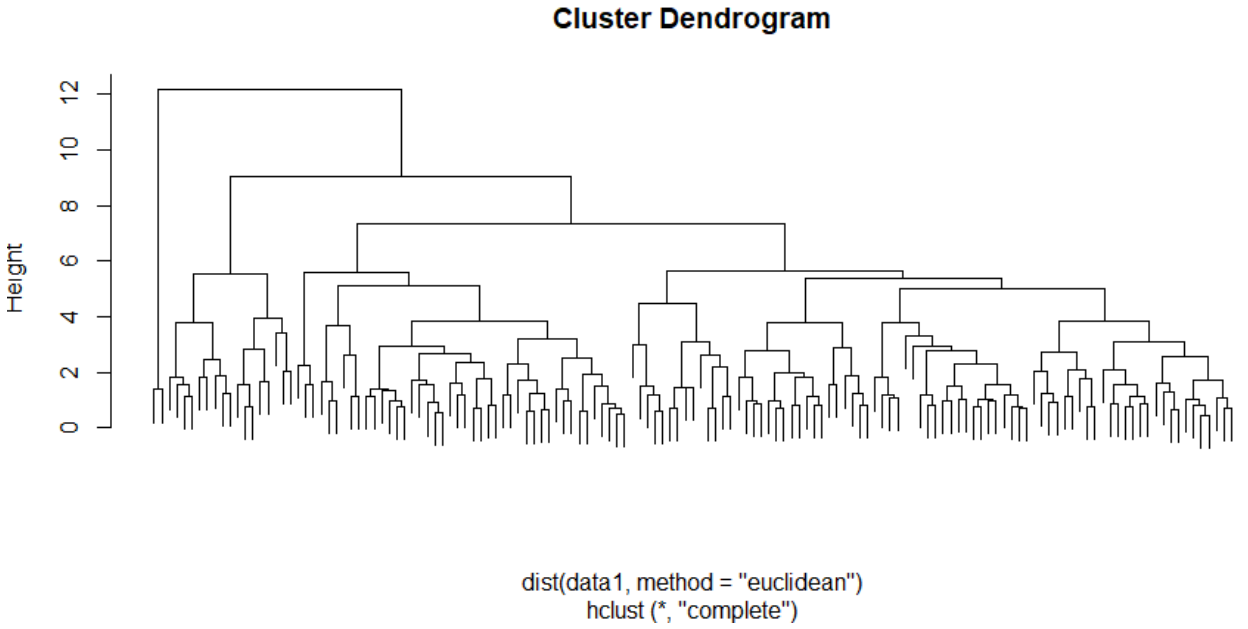
```
> # застосуємо ієрархічну кластеризацію для методу одного зв'язку і евклідово
і відстані
> h_single_eucl1 <- hclust(dist(data1, method = 'euclidean'), method = 'singl
e')
> plot(h_single_eucl1, labels = FALSE)
```



Що можемо бачити на цій дендрограмі? На мою думку, ця дендрограма не дає абсолютно ніякої відповіді про кількість кластерів. З деякою натяжкою можна стверджувати що їх 2, але про їх збільшення я не ризикну припускати…

Застосуємо тепер метод повного зв’язку.

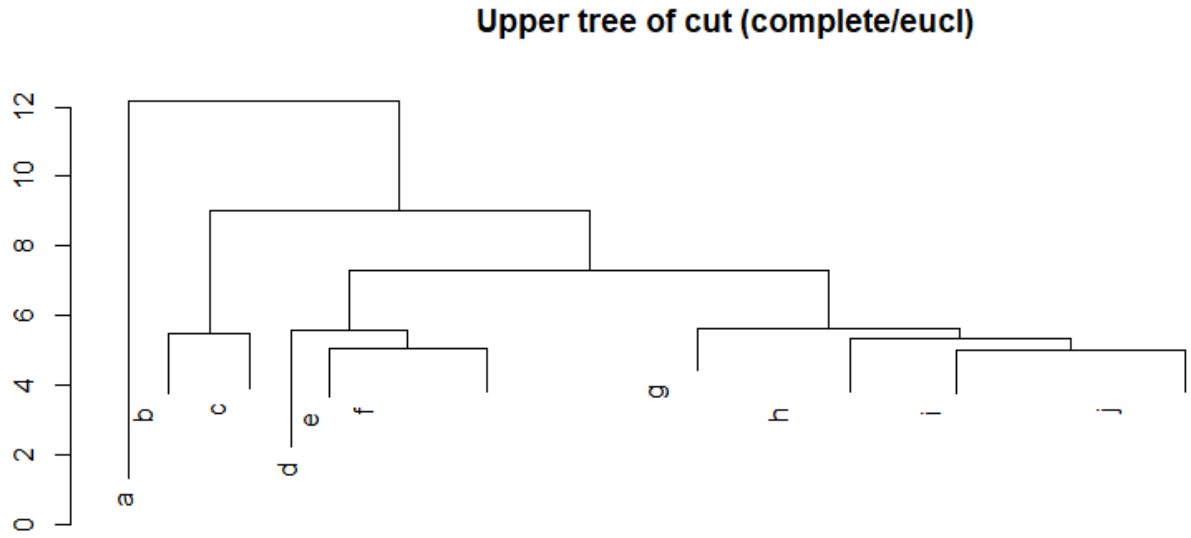
```
> # застосуємо ієрархічну кластеризацію для методу повного зв'язку і евклідов
ої відстані
> h_compl_eucl1 <- hclust(dist(data1, method = 'euclidean'), method = 'comple
te')
> plot(h_compl_eucl1, labels = FALSE)
```

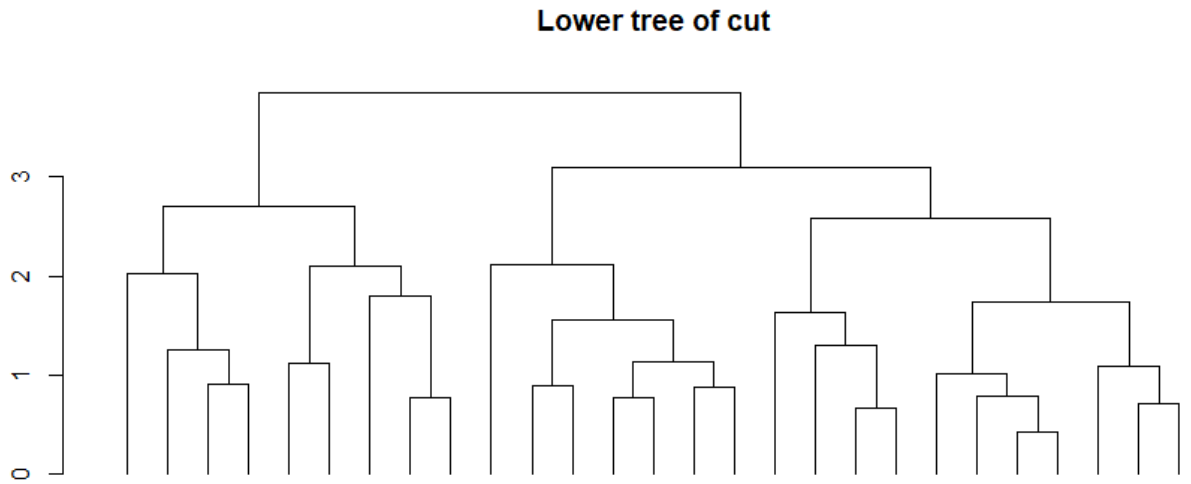


По цій дендрограмі, на мою думку, можна розбити дані на 2-4 кластери. Але спробуймо поглянути на рівень $h = 4.7$. Там, здається, кластерів виділяється 10. «Розріжемо» дендрограму на два дерева, і поглянемо на результати.

```
> h_compl_eucl_up1 <- cut(as.dendrogram(h_compl_eucl1), h = 4.7)$upper
> labels(h_compl_eucl_up1) <- letters[1:10]
```

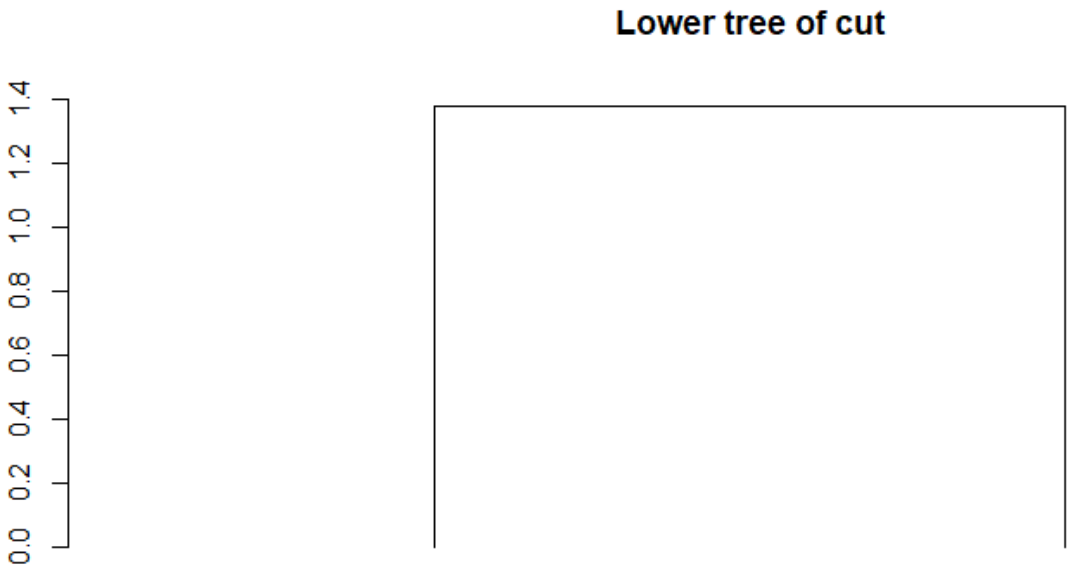
```
> plot(h_compl_eucl_up1, main = 'Upper tree of cut (complete/eucl)')
> h_compl_eucl_low1 <- cut(as.dendrogram(h_compl_eucl1), h = 4.7)$lower[[10]]
> labels(h_compl_eucl_low1) <- NULL
> plot(h_compl_eucl_low1, main = 'Lower tree of cut')
```





Здається, на нижньому дереві спостереження розташовуються на достатньо рівних відстанях, тобто такий розподіл має місце бути. Але поглянувши на нижнє дерево для першого кластера…

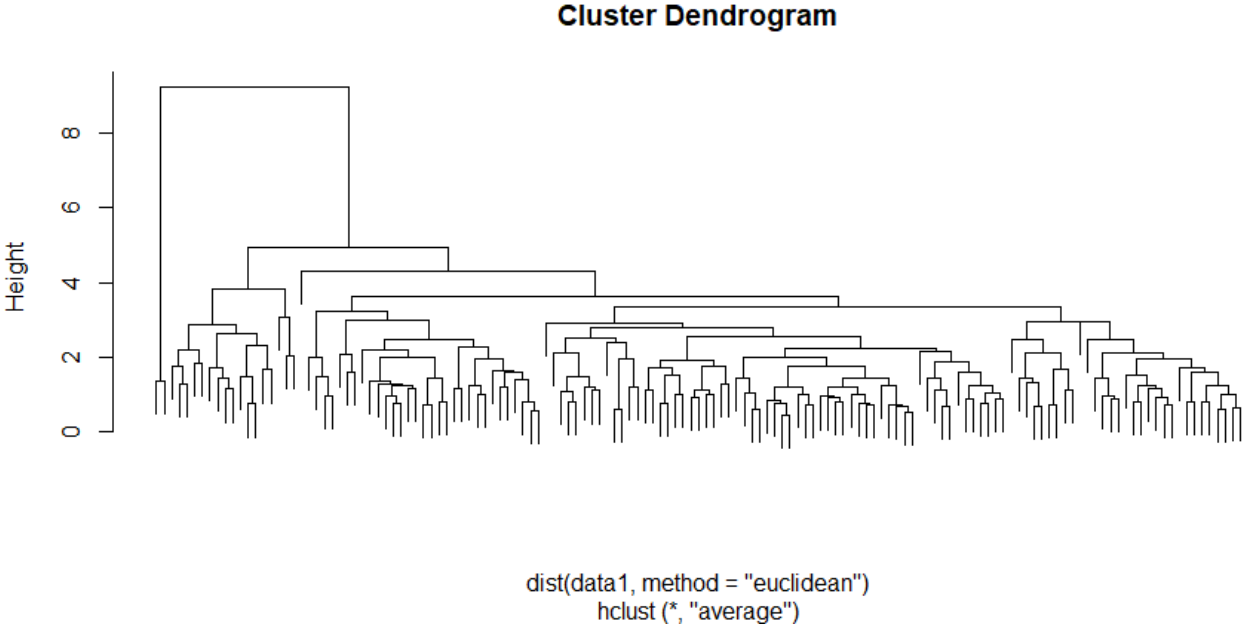
```
> h_compl_eucl_low1 <- cut(as.dendrogram(h_compl_eucl1), h = 4.7)$lower[[1]]
> labels(h_compl_eucl_low1) <- NULL
> plot(h_compl_eucl_low1, main = 'Lower tree of cut')
```



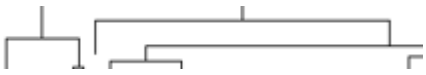
Можемо бачити, що в цьому кластері всього-навсього два спостереження (що, до речі було видно із загальної картини). Такий поділ не здається особисто мені логічним. Тому я би зупинявся все таки на 2-4 кластерах.

Тепер подивимось на метод середнього зв’язку.

```
> # застосуємо ієрархічну кластеризацію для методу середнього зв'язку і евклі
дової відстані
> h_avg_eucl1 <- hclust(dist(data1, method = 'euclidean'), method = 'average'
)
> plot(h_avg_eucl1, labels = FALSE)
```



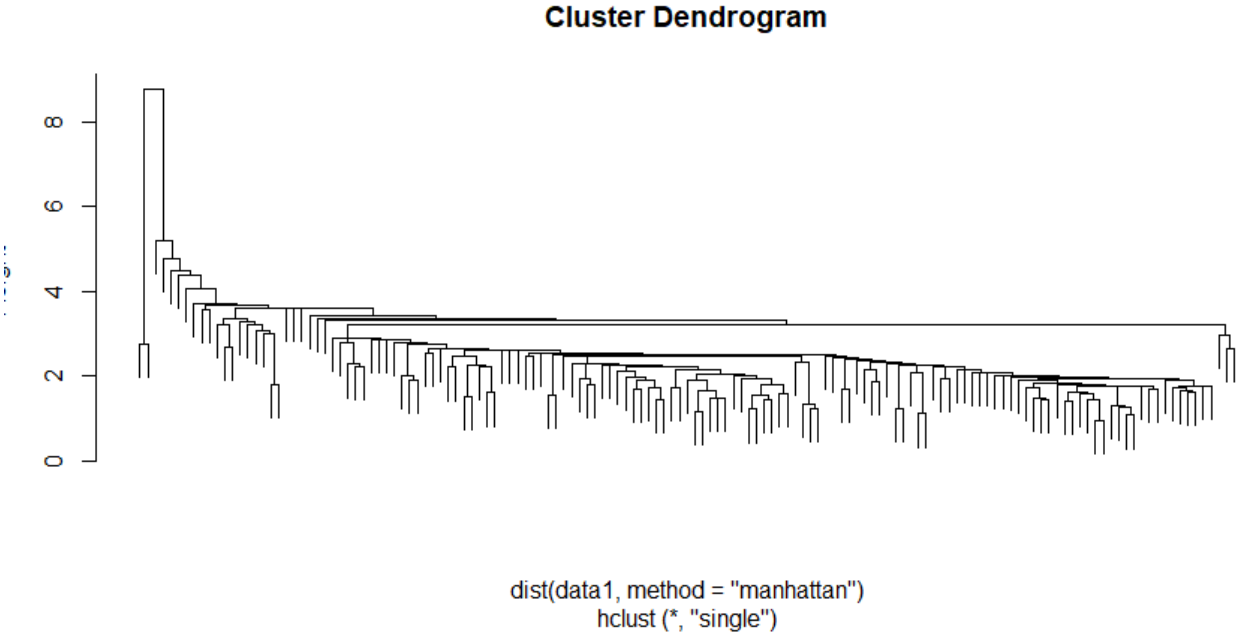
Ця дендрограма чи не повторює минулої. Знову ж таки тут доцільним буде розбивати на 2-4 кластери, хоча тут я би відмітив один маленький шматочок.



Я, на жаль, пустого відгалуження вліво пояснити не можу, і виглядає це дуже і дуже дивно.

Далі застосуємо манхаттанську відстань. І спершу використаємо метод одного зв’язку.

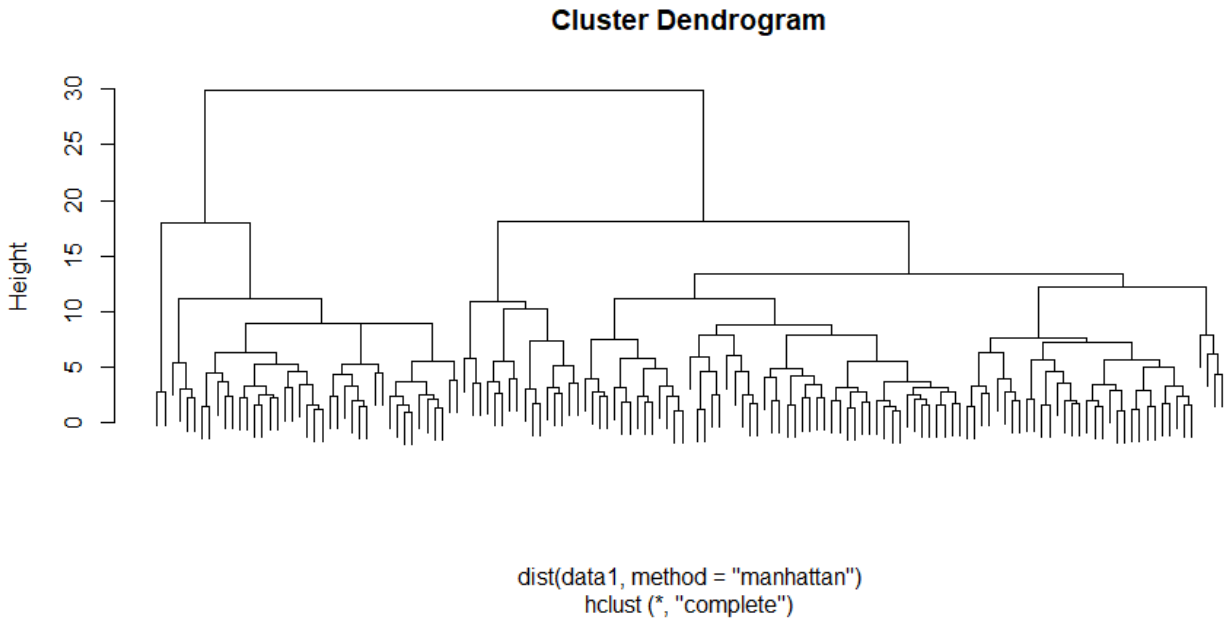
```
> # застосуємо ієрархічну кластеризацію для методу одного зв'язку і манхаттан
ської відстані
> h_single_mannh1 <- hclust(dist(data1, method = 'manhattan'), method = 'sing
le')
> plot(h_single_mannh1, labels = FALSE)
```



Ще раз на методі одного зв’язку з’являється зовсім невтішна картина: навіть якщо і кластеризувати на два кластери, то в першому з них виявиться два стостереження, тоді як в усіх інших (можливих) частинах – решта (біля 140). Даний метод, на мою думку, не дає ніякої відповіді щодо кількості кластерів.

Далі застосуємо метод повного зв’язку.

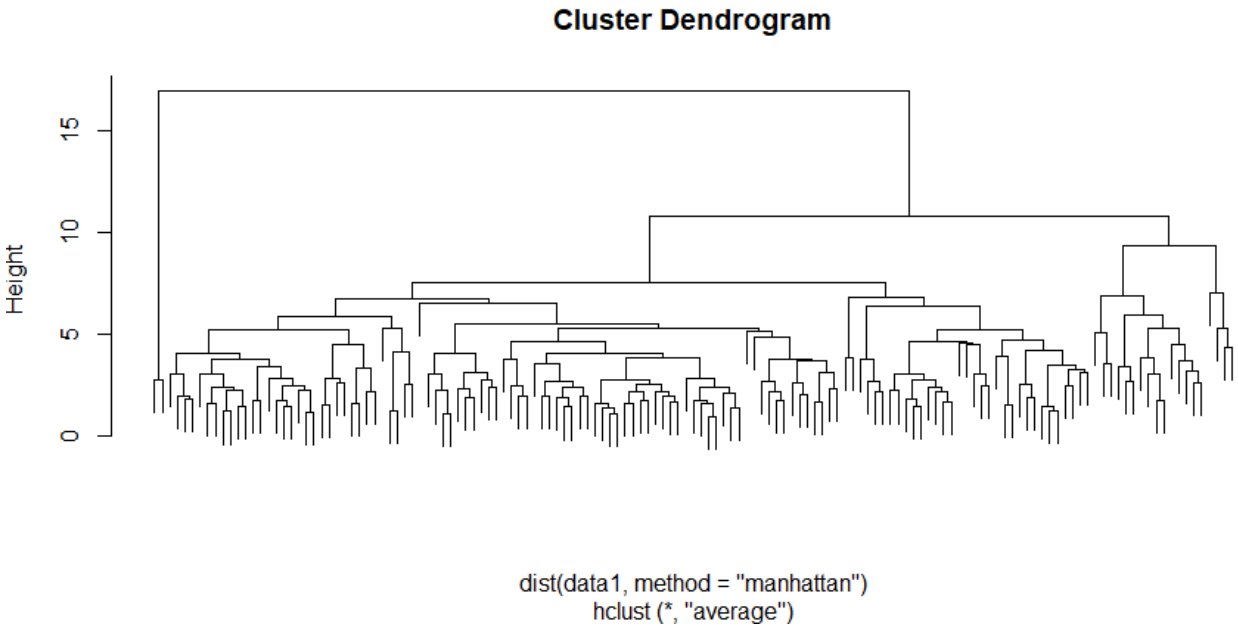
```
> # застосуємо ієрархічну кластеризацію для методу повного зв'язку і манхатта
нської відстані
> h_compl_mannh1 <- hclust(dist(data1, method = 'manhattan'), method = 'compl
ete')
> plot(h_compl_mannh1, labels = FALSE)
```



З чисто оглядової точки зору, тут картина найкраща: чітко можна виділити від 2 до 4 кластерів.

І наостанок для даної відстані застосуємо метод середнього зв’язку.

```
> # застосуємо ієрархічну кластеризацію для методу середнього зв'язку і манха
ттанської відстані
> h_avg_mannh1 <- hclust(dist(data1, method = 'manhattan'), method = 'average
')
> plot(h_avg_mannh1, labels = FALSE)
```

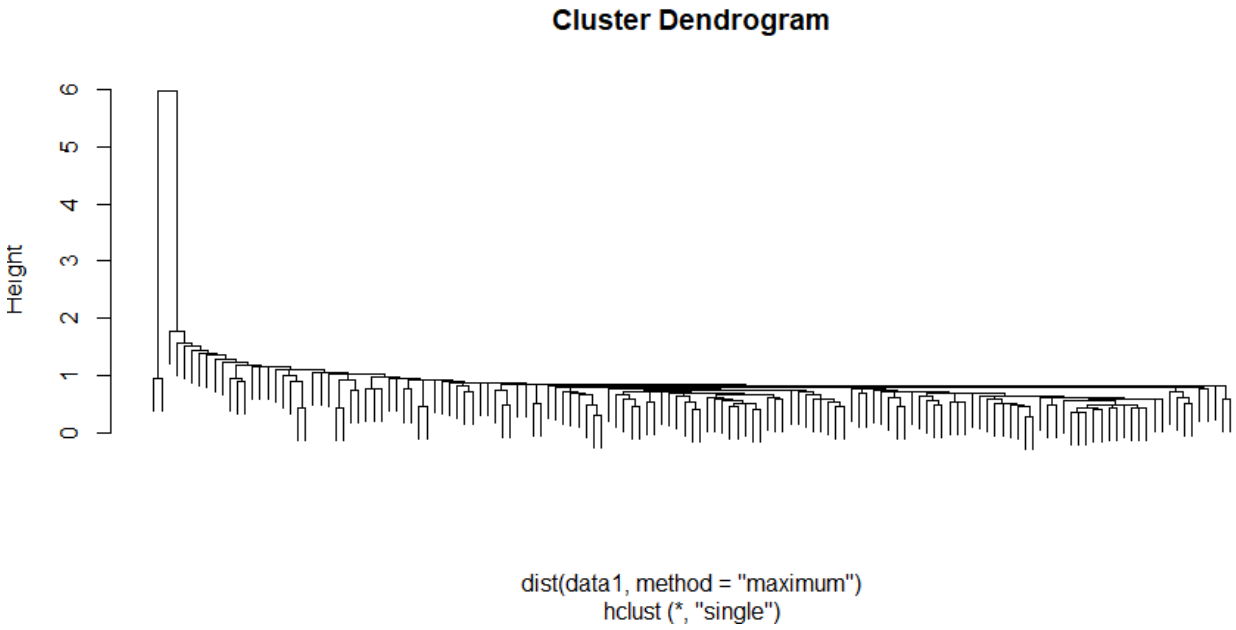


Даний метод разом із дистанцією стверджують, що якою б не була кількість кластерів, в перший потрапить лише два спостереження (лише дві країни). Вивівши всі назви, і розтягнувши графік у весь екран, виявилось, що цими країнами є Сінгапур та Гонконг. Вони, звичайно, економічно розвинені, але ніяк не можуть належать до цілої окремої групи. Тому дана кластеризація не є

доцільною. Дендрограму з підписами в звіт не розміщуватимемо з міркувань її «нечитабельності».

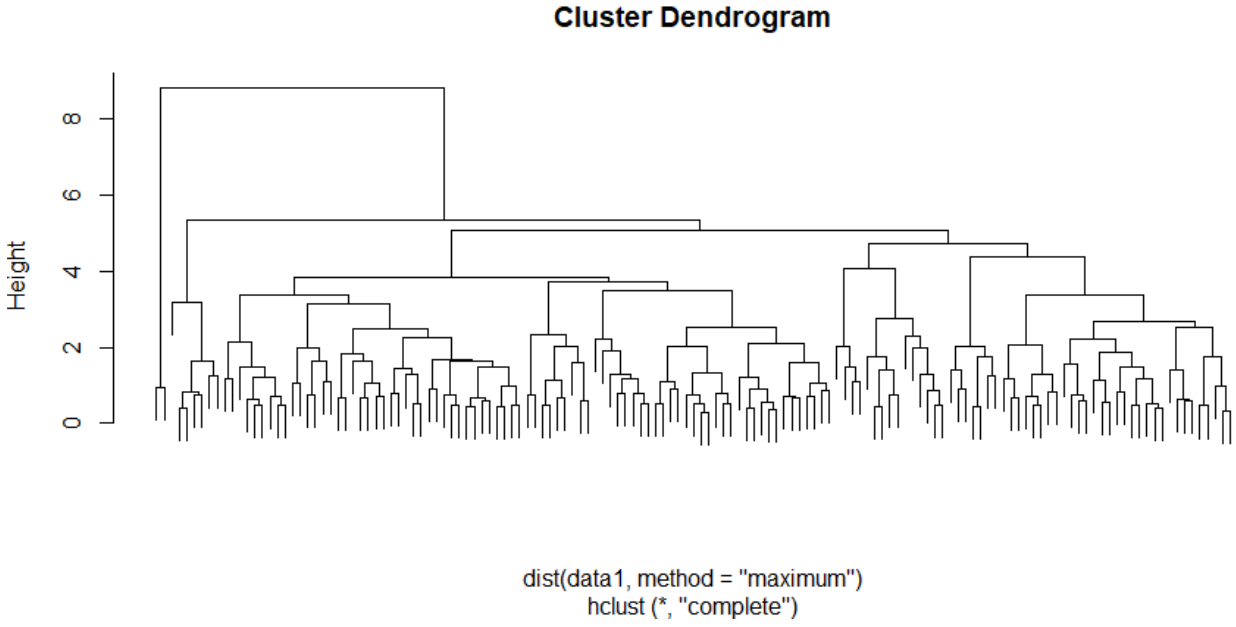
Тепер застосуємо максимальну відстань, і для початку метод одного зв’язку.

```
> # застосуємо ієрархічну кластеризацію для методу одного зв'язку і максималь
ної відстані
> h_single_max1 <- hclust(dist(data1, method = 'maximum'), method = 'single')
> plot(h_single_max1, labels = FALSE)
```



В черговий раз на наших даних метод одного зв’язку дає неінтерпретабельні результати, тому одразу переходимо до методу повного зв’язку.

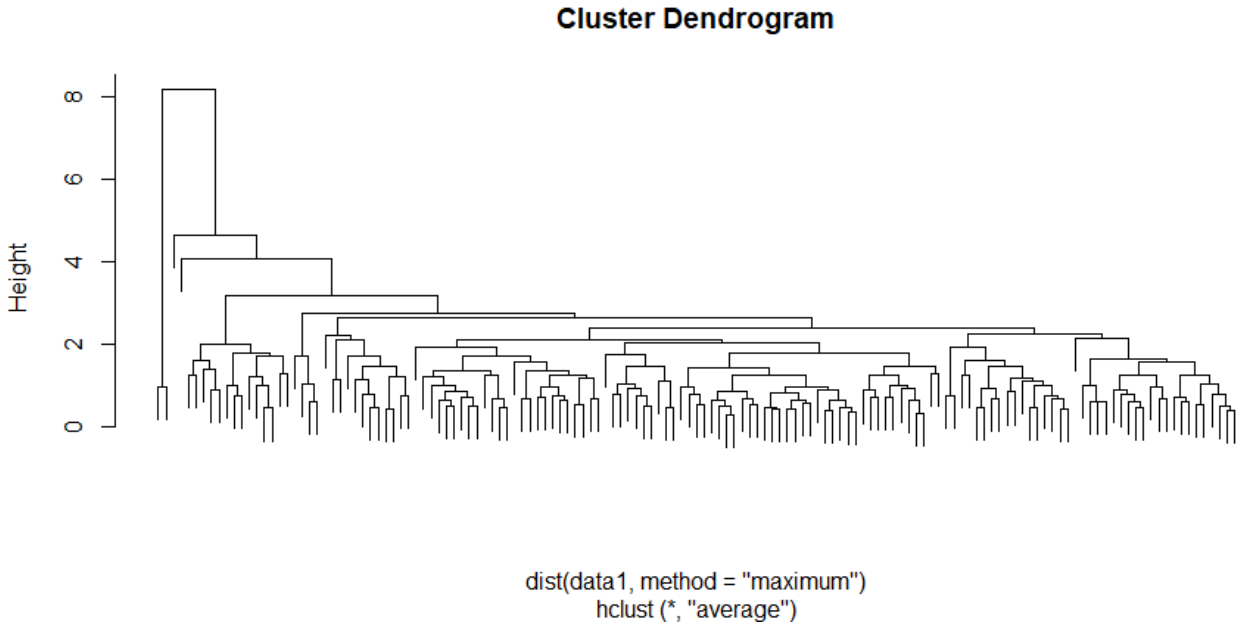
```
> # застосуємо ієрархічну кластеризацію для методу повного зв'язку і максимал
ьної відстані
> h_compl_max1 <- hclust(dist(data1, method = 'maximum'), method = 'complete'
)
> plot(h_compl_max1, labels = FALSE)
```



Тут знову помітний поділ на 2, але ось на 3 чи на 4 тут розділити не так просто. Тут, скоріше, можна було би розглянути поділ на рівні $h = 3$, але знову можемо звернути увагу на «двокраїнний» кластер зліва.

Наприкінці подивимось на метод середнього зв'язку.

```
> # застосуємо ієрархічну кластеризацію для методу середнього зв'язку і макси
мальної відстані
> h_avg_max1 <- hclust(dist(data1, method = 'maximum'), method = 'average')
> plot(h_avg_max1, labels = FALSE)
```



Знову та ж сама ситуація: дві окремі країни. Та ще й поруч з ними два незамкнених відгалуження…

Отже, на практичних даних, абсолютно всі відстані і методи виділили Гонконг та Сінгапур в окремий кластер, що в розрізі економічних та соціально-політичних показників не є зовсім логічним. Мабуть, така природа наших даних, що ієрархічна кластеризація бодай одного разу не дала іншого варіанту для цих двох країн… Але якщо відкинути це непорозуміння, то, здається, праві частини на всіх кластеризаціях підлягають бодай якомусь інтерпетуванню.