Дескриптивна статистика

3 курс, статистика, Шкляр Ірина Володимирівна

Завдання 6, варіант 9

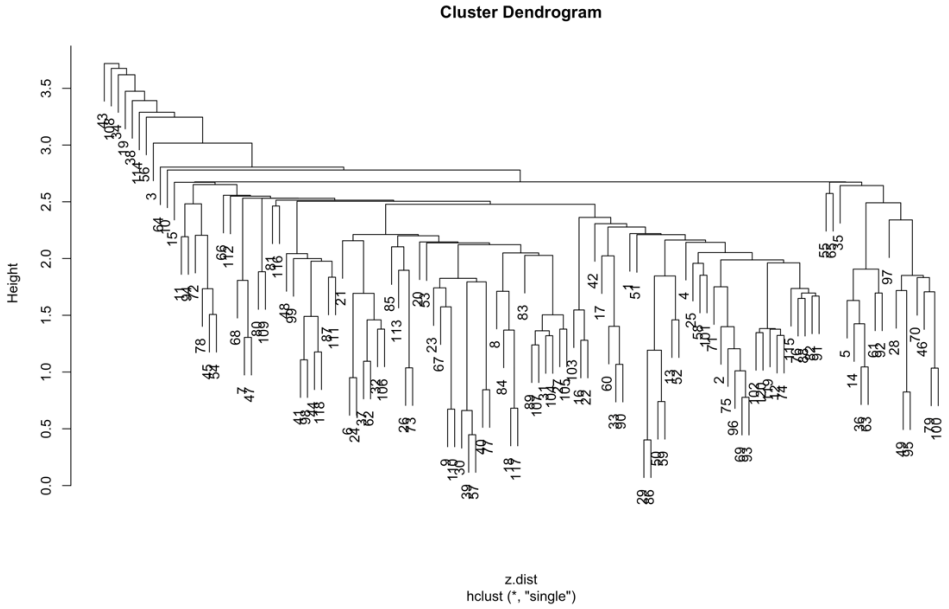
Частина А

```
> z<-read.table("/Users/irynashkliar/Downloads/multi/F9t.txt",header=F)
>
> res<-princomp(z)
```

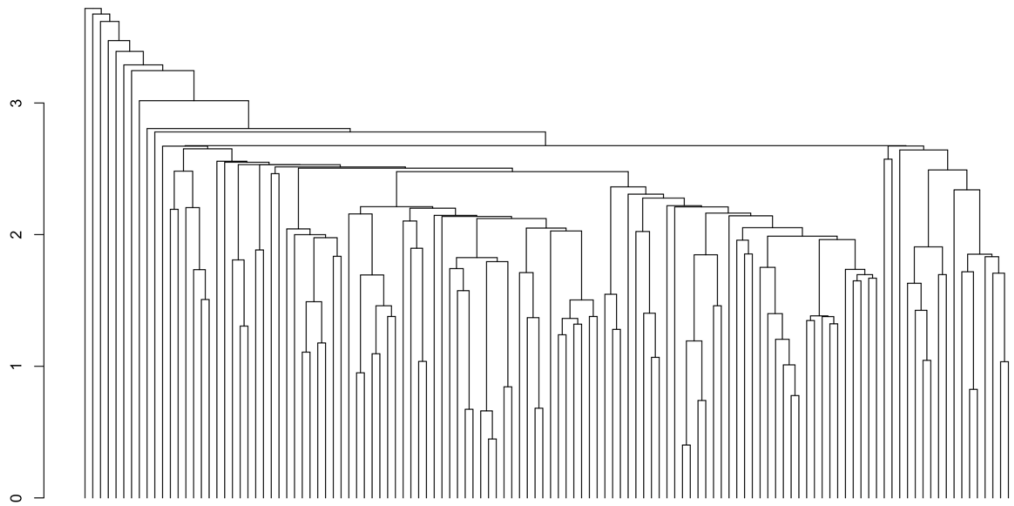
Спочатку я спробувала проаналізувати дані 1-3:

Метод найближчого сусіда:

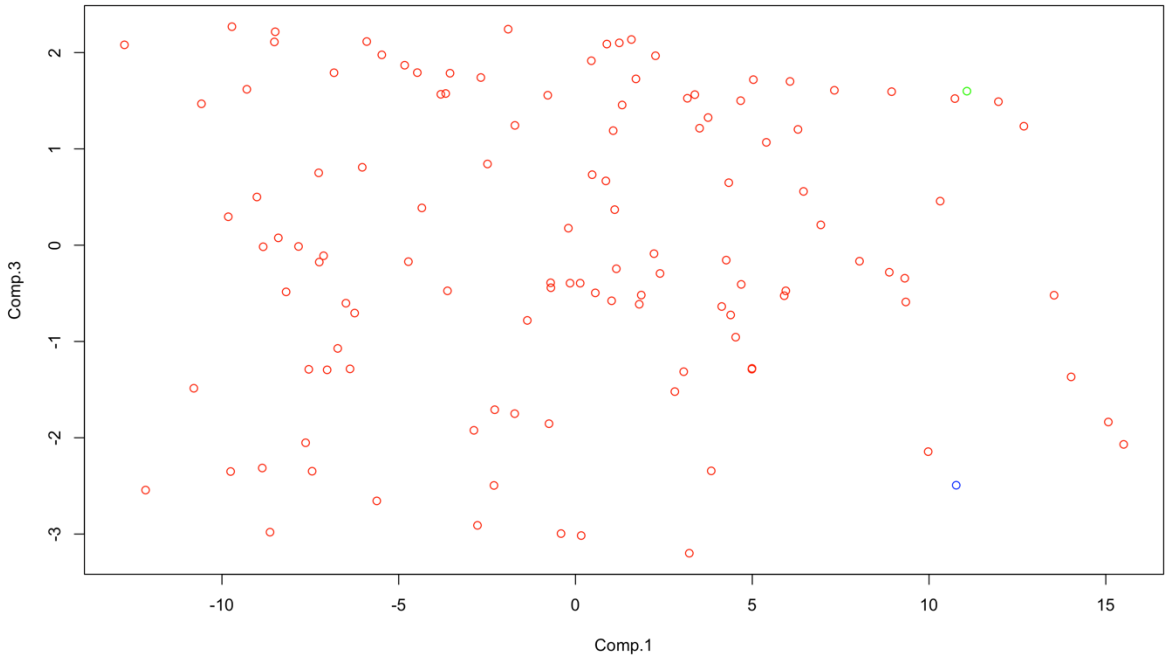
```
> z.dist<-dist(res$scores[,1:3])
>
> z.hclust<-hclust(z.dist,method ="single")
>
> plot(z.hclust)
```




```
> plot(as.dendrogram(z.hclust),leaflab="none")
```

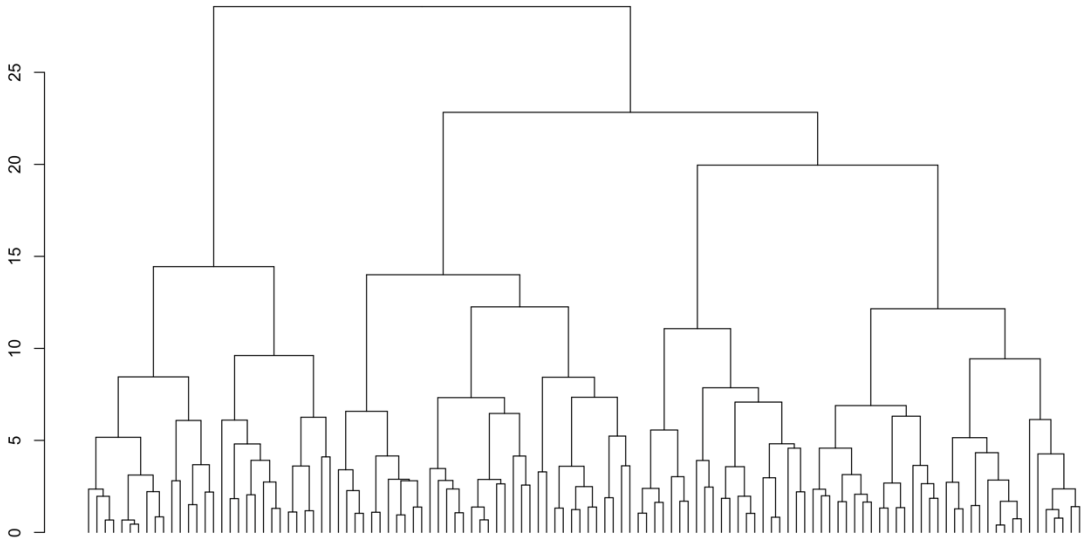



```
> groups3<-cutree(z.hclust,k=3)
>
> plot(res$scores[,c(1,3)],col=c("red","blue","green")[groups3])
```

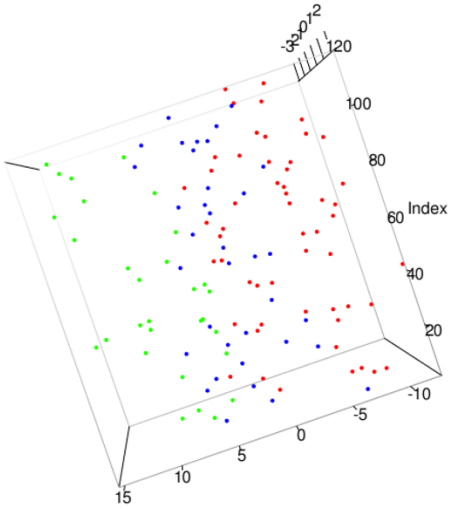


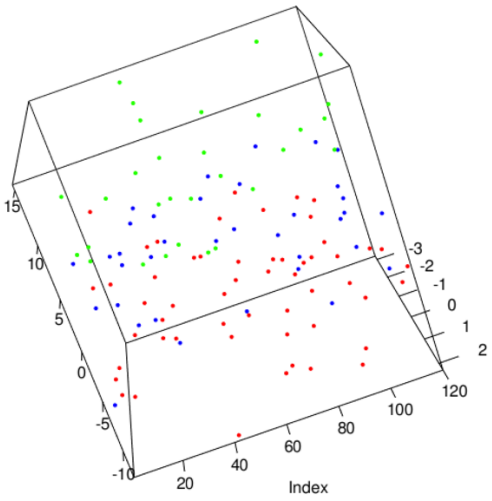
Метод найдальшого сусіда:

```
> z.hclust<-hclust(z.dist,method ="complete")
>
> plot(as.dendrogram(z.hclust),leaflab="none")
```



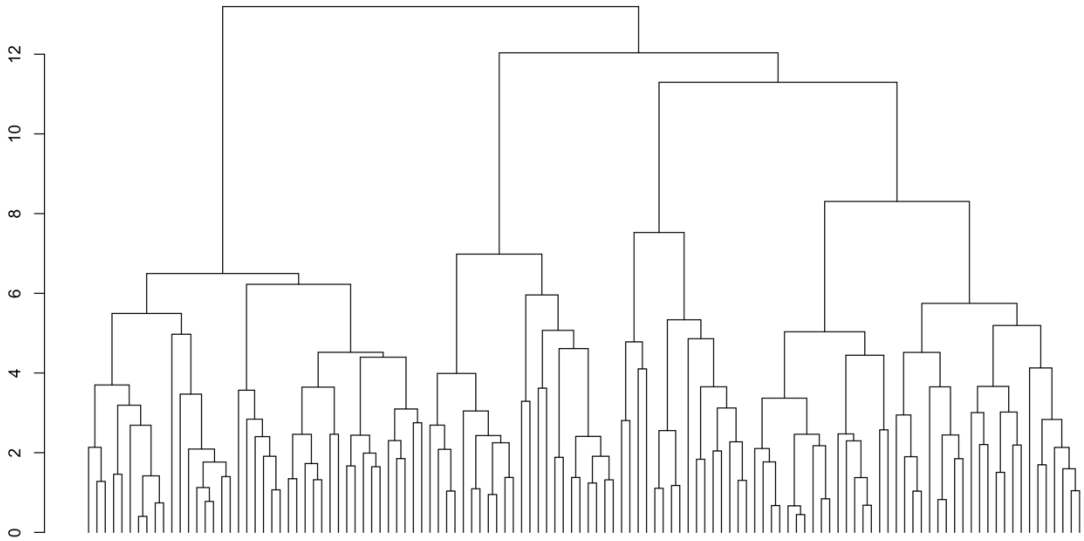
> plot3d(res$scores[,c(1,3)],col=c("red","blue","green")[groups3_0])



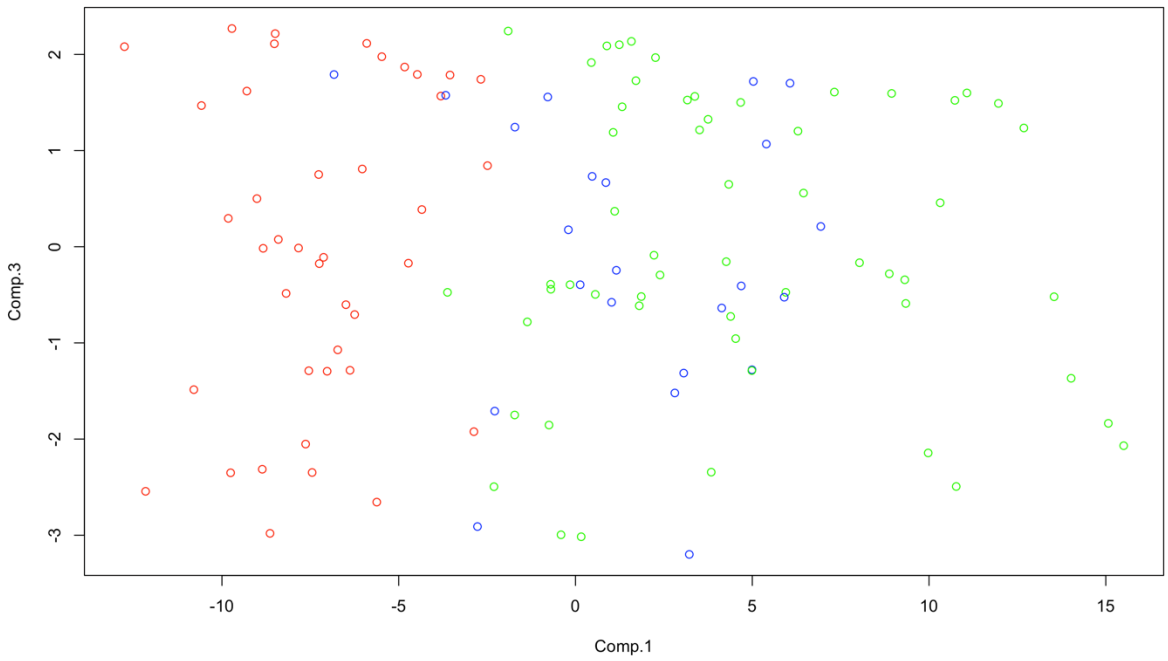


Відстань середнього зв’язку:

> z.hclust<-hclust(z.dist,method ="average")
>
> plot(as.dendrogram(z.hclust),leaflab="none")



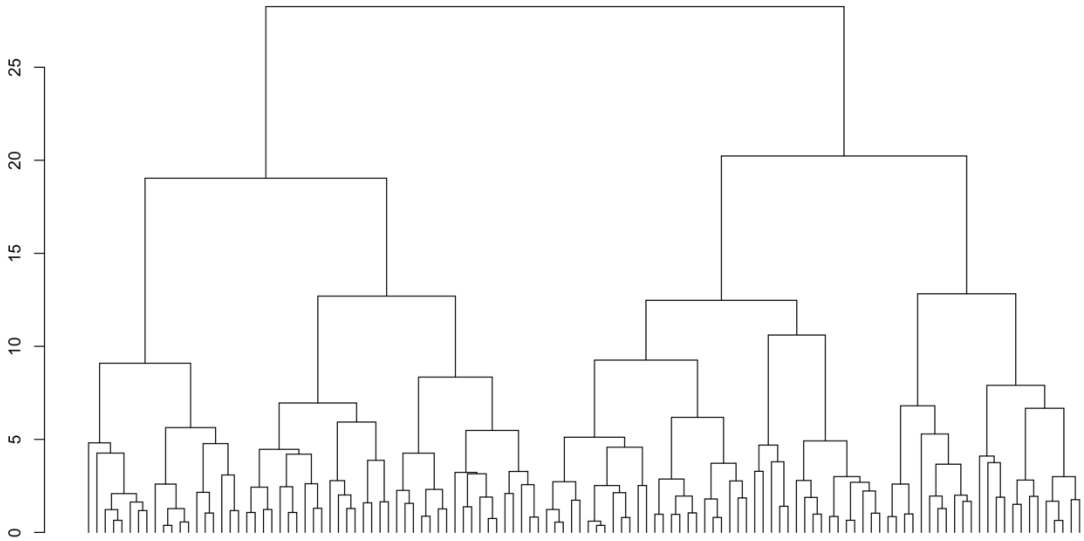

```
> groups3_0<-cutree(z.hclust,k=3)
>
> plot(res$scores[,c(1,3)],col=c("red","blue","green")[groups3_0])
```



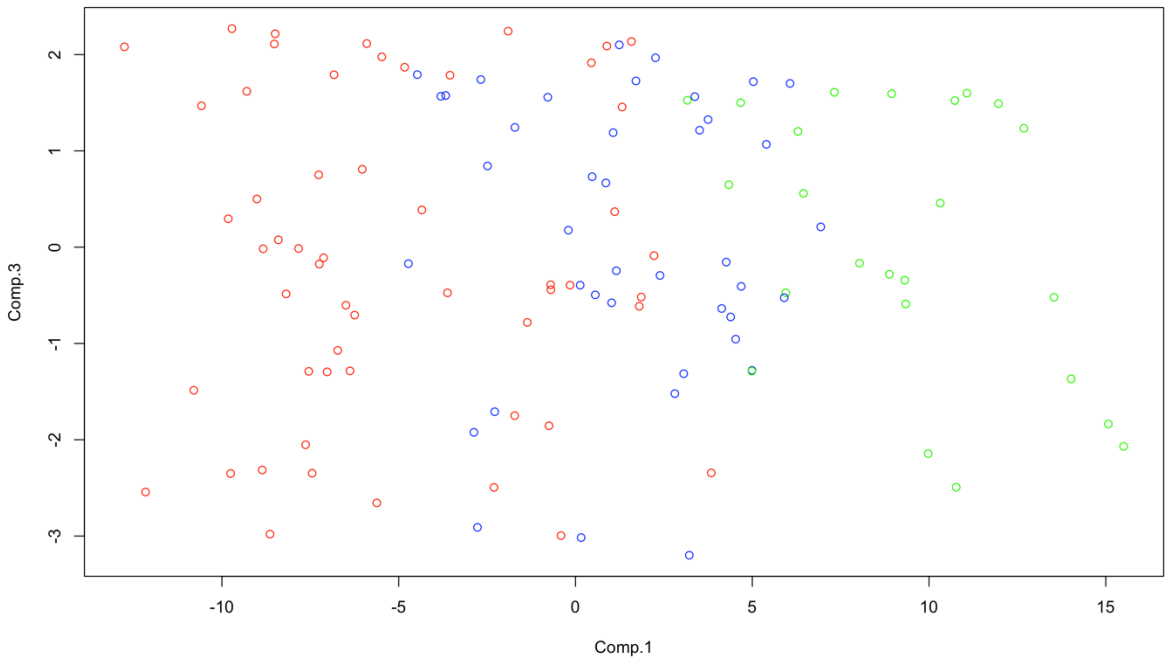
Максимум абсолютних значень різниць координат:

```
> z.dist<-dist(res$scores[,1:3],method ="maximum")
```

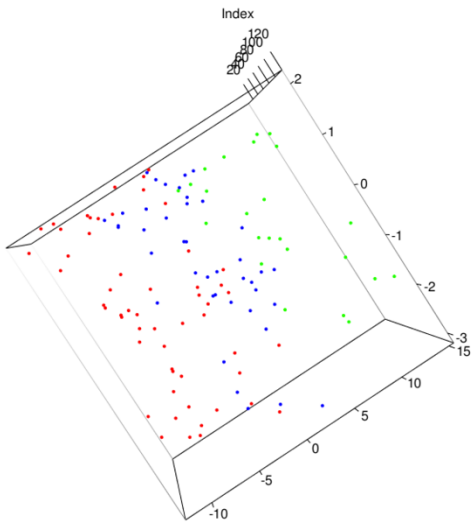
```
> z.hclust<-hclust(z.dist)
>
> plot(as.dendrogram(z.hclust),leaflab="none")
```

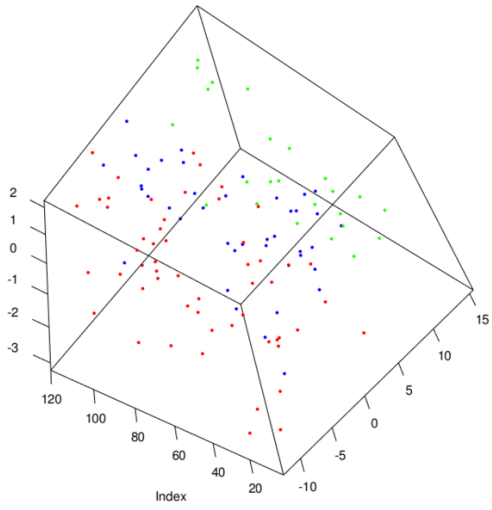



```
> groups3_0<-cutree(z.hclust,k=3)
>
> plot(res$scores[,c(1,3)],col=c("red","blue","green")[groups3_0])
```



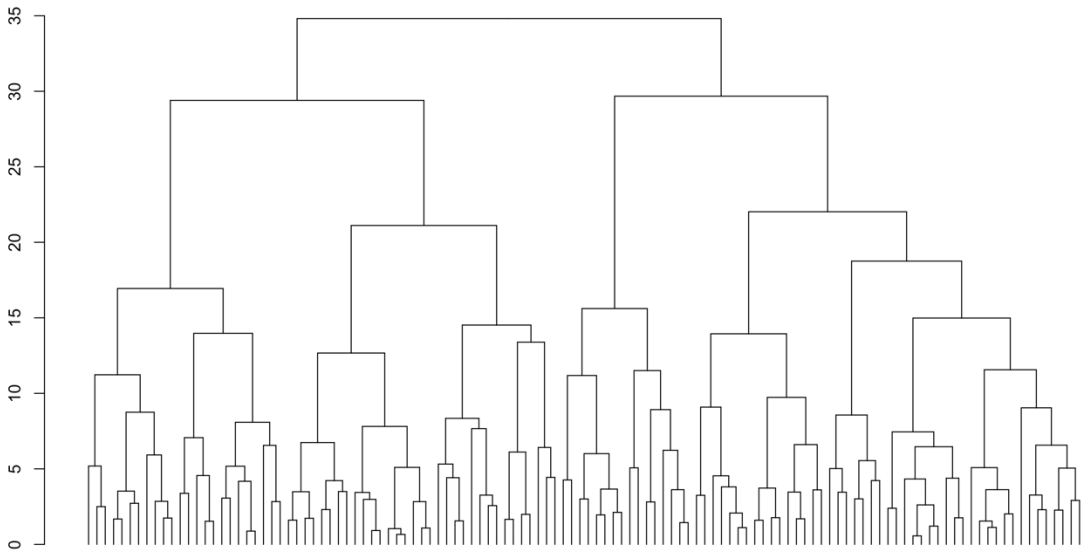

```
> plot3d(res$scores[,c(1,3)],col=c("red","blue","green")[groups3_0])
```



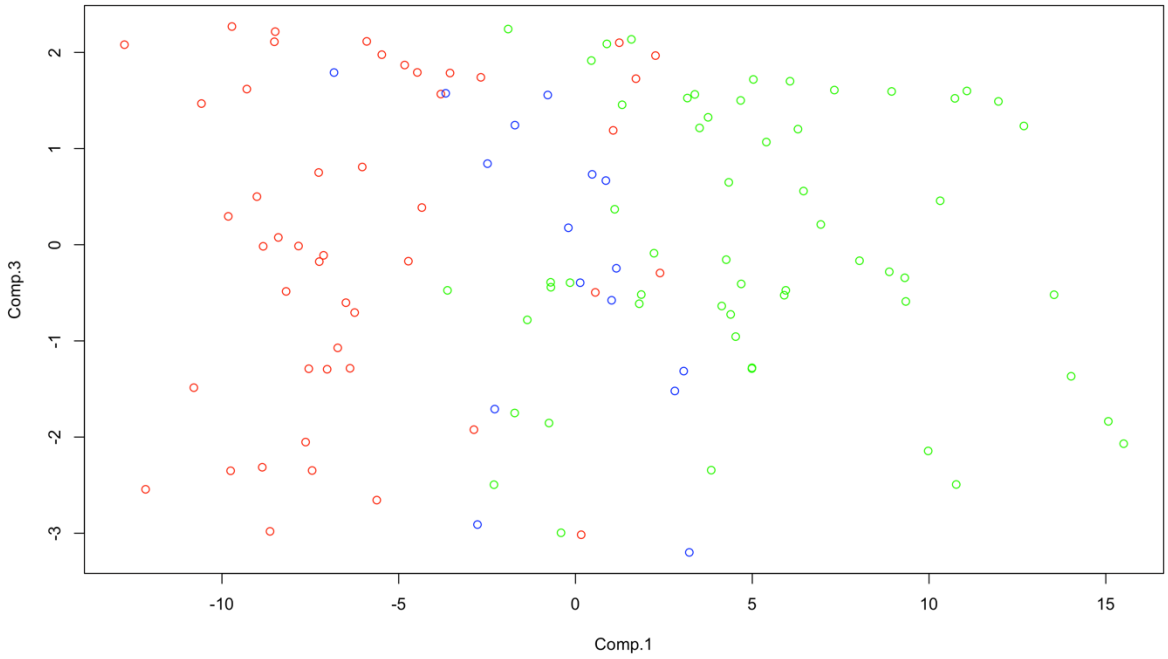


Відстань сіті-блок:

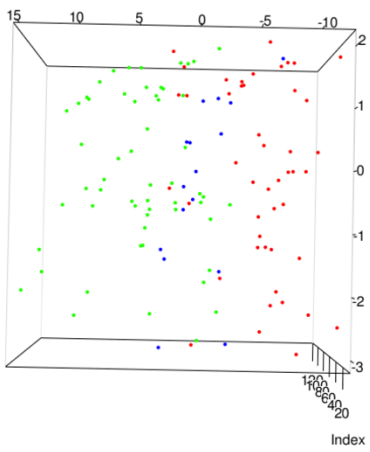
```
> z.dist<-dist(res$scores[,1:3],method ="manhattan")
>
> z.hclust<-hclust(z.dist)
>
> plot(as.dendrogram(z.hclust),leaflab="none")
```

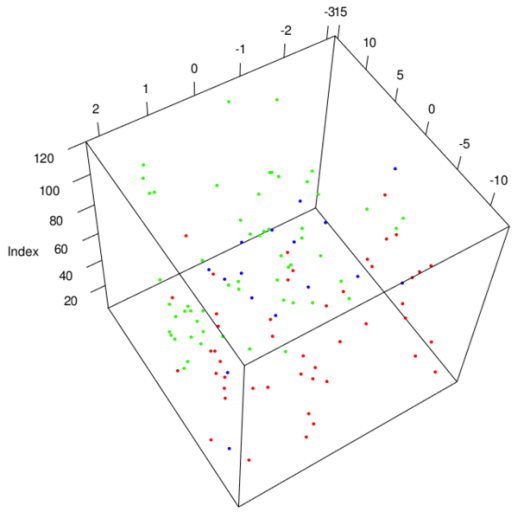



```
> groups3_0<-cutree(z.hclust,k=3)
>
> plot(res$scores[,c(1,3)],col=c("red","blue","green")[groups3_0])
```



> plot3d(res$scores[,c(1,3)],col=c("red","blue","green")[groups3_0])



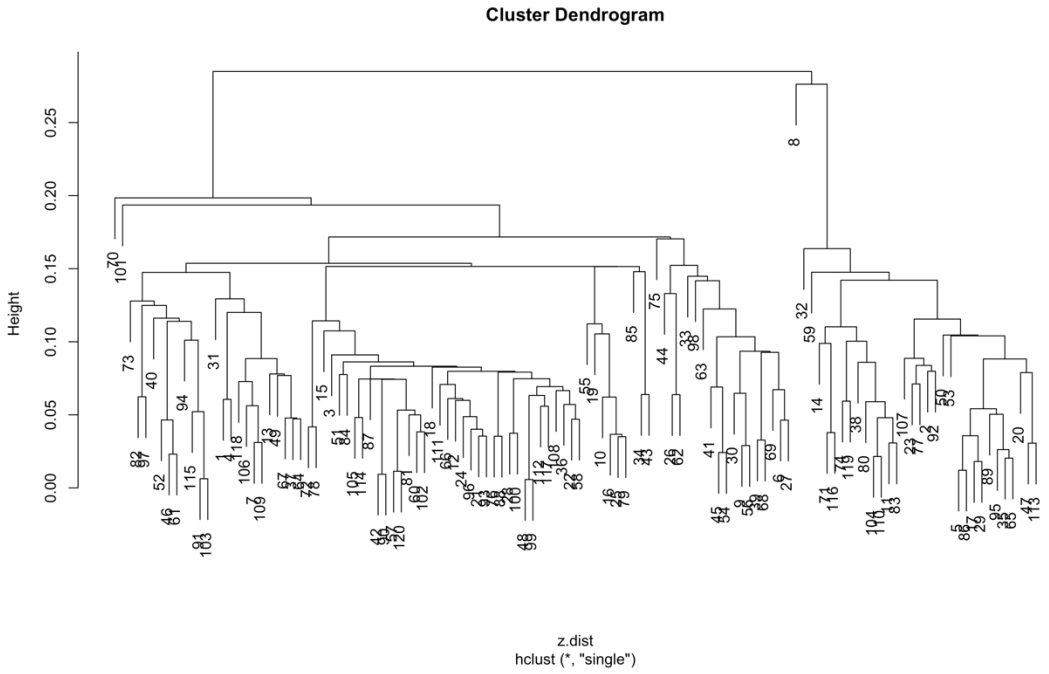


Бачимо, що для всіх методів кластеризації структури не дуже сильно виділяються, в завданні 4 дані 1-3 також майже не виділялись.

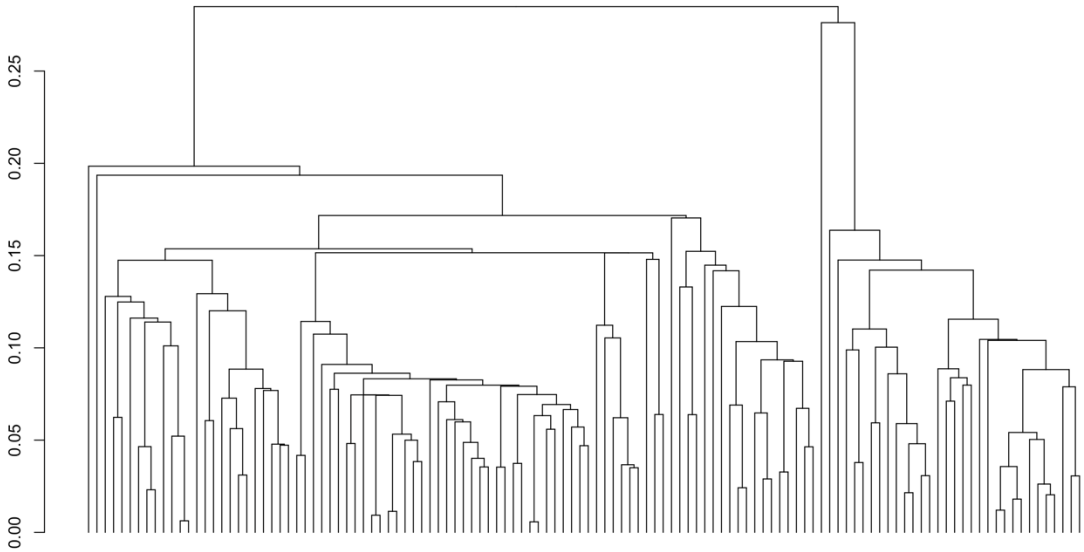
Далі я проаналізувала дані для компонент 4-6:

Метод найближчого сусіда:

```
> z.dist<-dist(res$scores[,4:6])
> z.hclust<-hclust(z.dist,method ="single")
>
> plot(z.hclust)
```

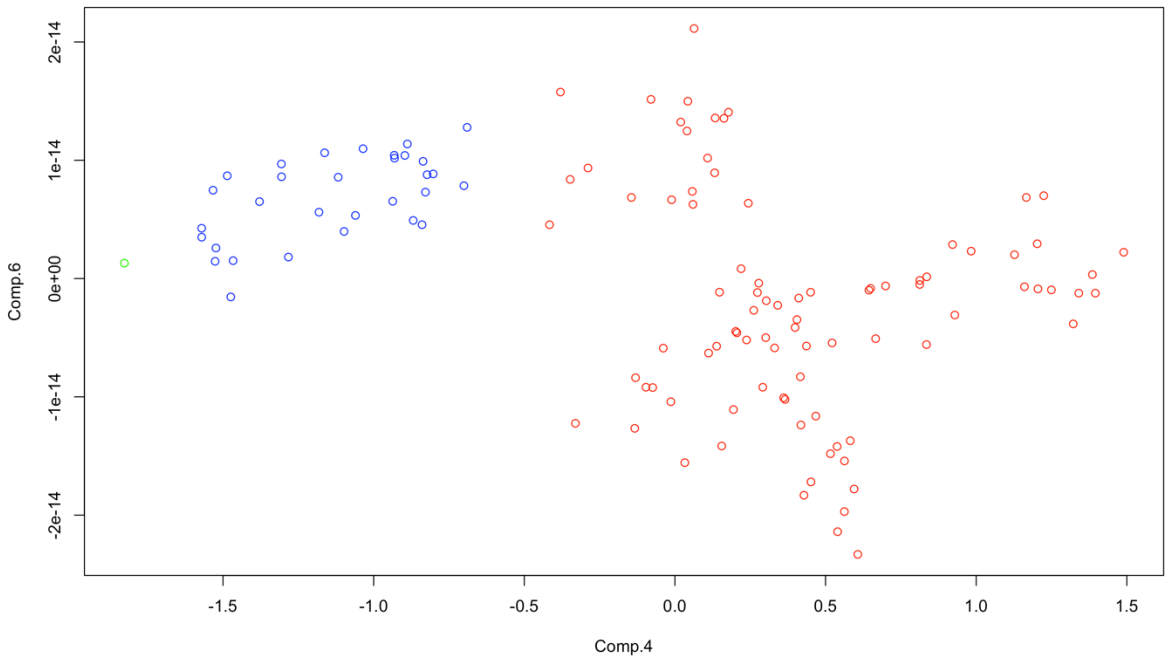



```
> plot(as.dendrogram(z.hclust),leaflab="none")
```

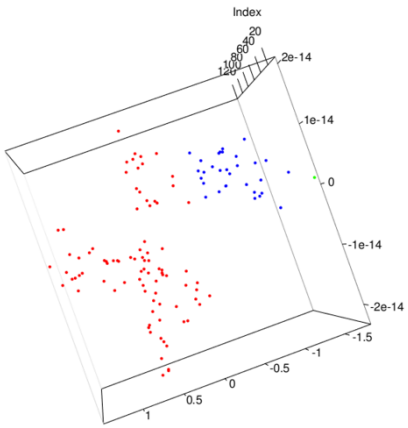


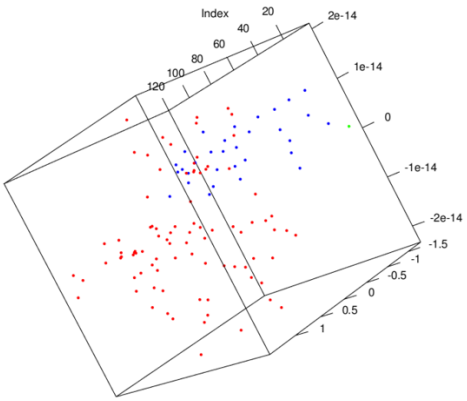

```
> groups3<-cutree(z.hclust,k=3)
```

>
> plot(res$scores[,c(4,6)],col=c("red","blue","green")[groups3]



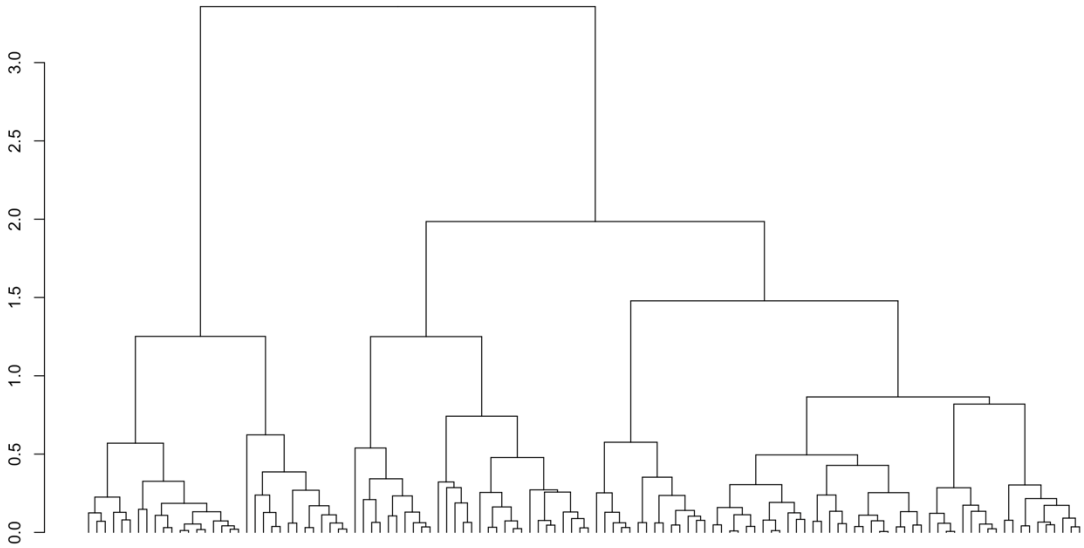
> library(rgl)
>
> plot3d(res$scores[,c(4,6)],col=c("red","blue","green")[groups3])



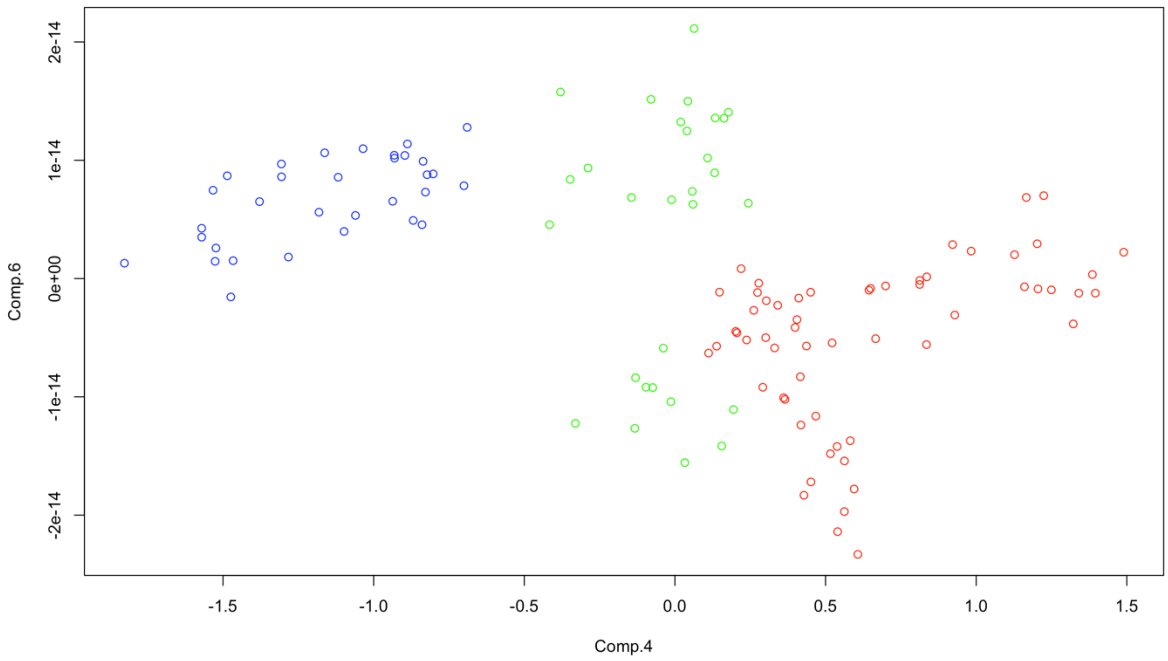


Метод найдальнього сусіда:

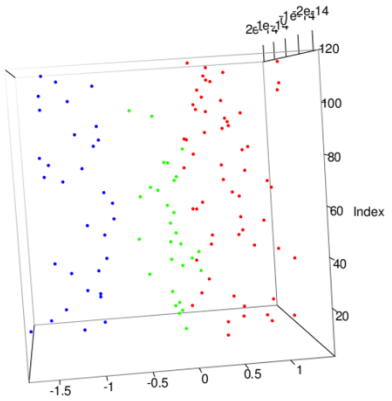
> z.hclust<-hclust(z.dist,method ="complete")
>
> plot(as.dendrogram(z.hclust),leaflab="none")

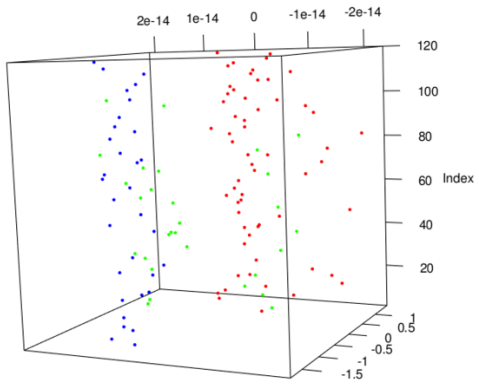



```
> groups3<-cutree(z.hclust,k=3)
>
> plot(res$scores[,c(4,6)],col=c("red","blue","green")[groups3])
```



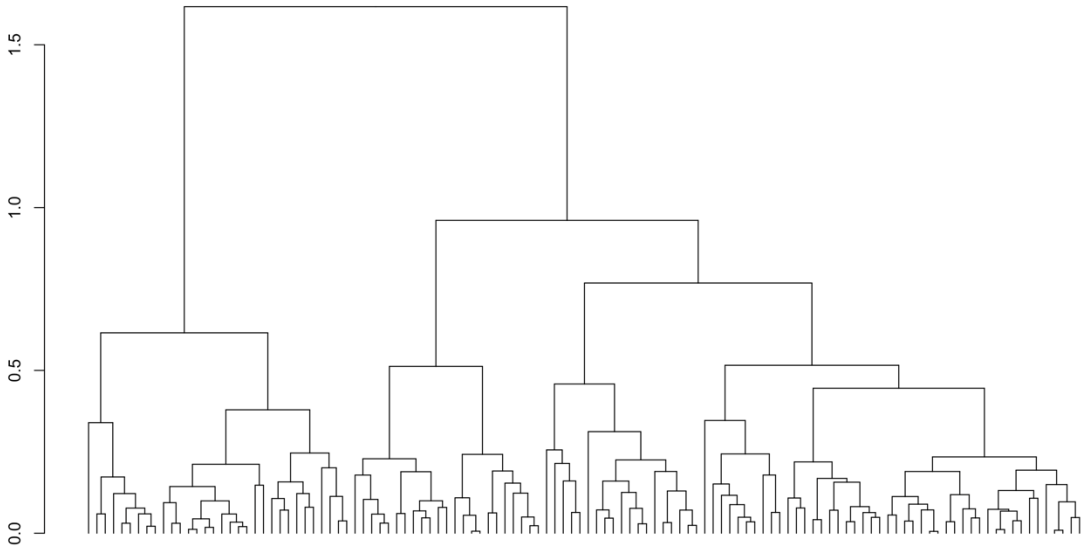

```
> plot3d(res$scores[,c(4,6)],col=c("red","blue","green")[groups3])
```



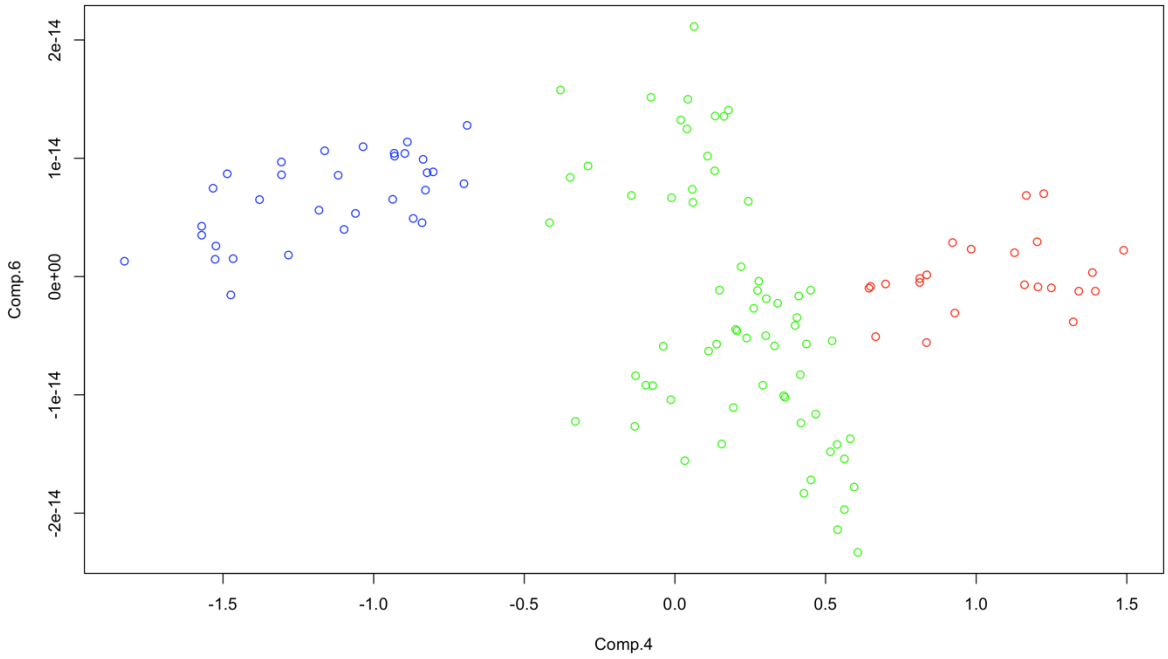


Відстань середнього зв’язку:

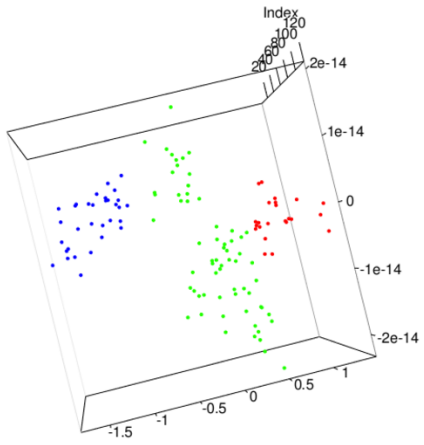
```
> z.hclust<-hclust(z.dist,method ="average")
>
> plot(as.dendrogram(z.hclust),leaflab="none")
```

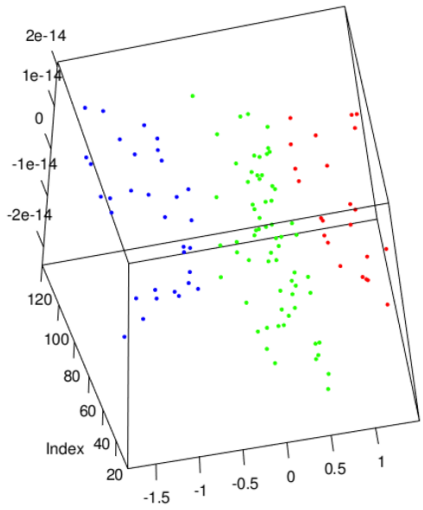



```
> groups3<-cutree(z.hclust,k=3)
>
> plot(res$scores[,c(4,6)],col=c("red","blue","green")[groups3])
```



> plot3d(res$scores[,c(4,6)],col=c("red","blue","green")[groups3])



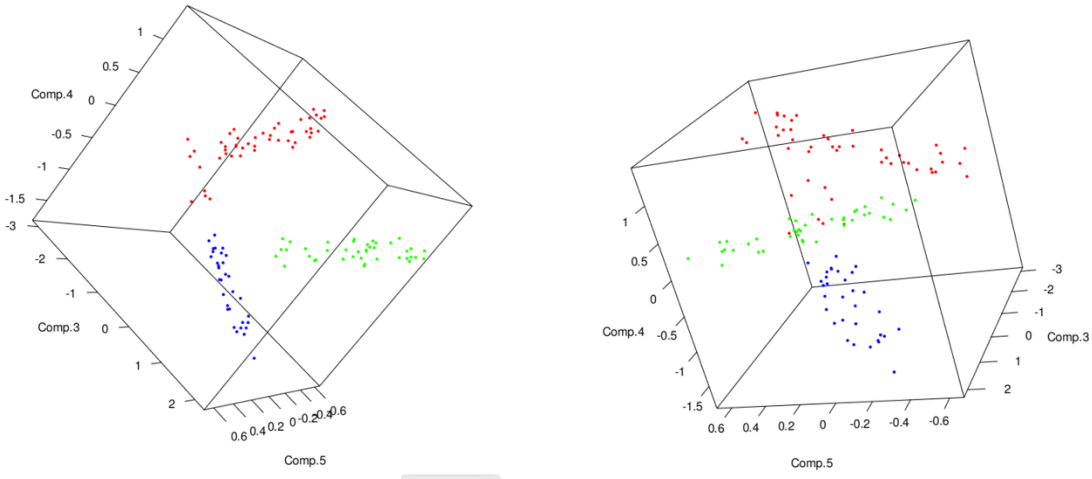


Тут структури виділились.

Також я виділила в окремий фрейм компоненти 3-5. Подивимось для них ієрархічну класифікацію за допомогою методу найближчого сусіда:

> K<-res$scores[,3:5]
>
> z.dist<-dist(K)

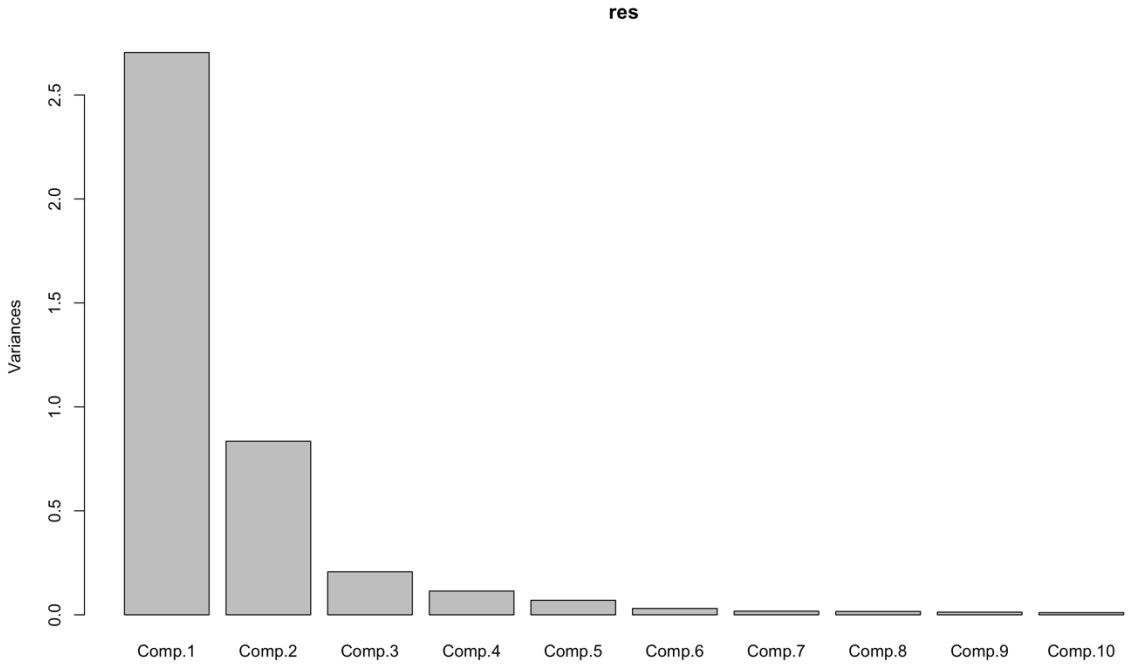
```
> z.hclust<-hclust(z.dist,method ="single")
>
> groups3_1<-cutree(z.hclust,k=3)
>
> plot3d(K,col=c("red","blue","green")[groups3_1])
```



Частина В

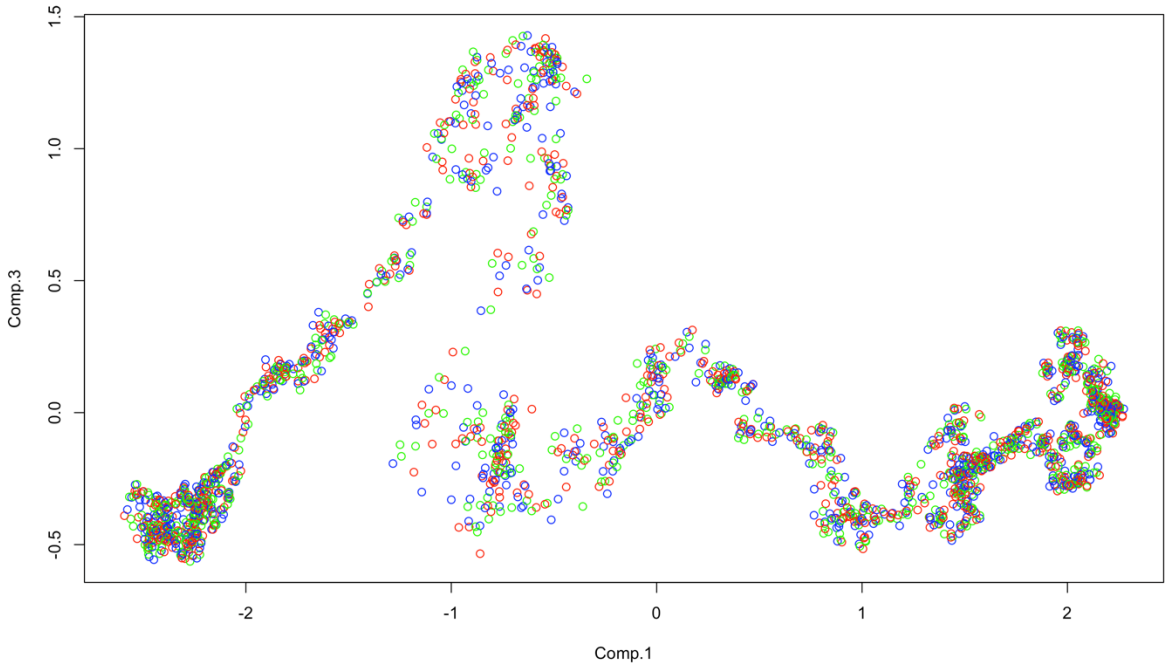
Мене зацікавили дані, які я використовувала в завданні 4, 6 пункті (log-returns). Я не впевнена, що результати будуть розділятися на кластери, але я думаю може бути щось цікаве.

```
> filenames=list.files(path="/Users/irynashkliar/Downloads/tables",full.names=TRUE)
>
> datalist = lapply(filenames,function(x){x0<-read.csv(file=x,header=F)[,c(1,6)];
colnames(x0)<-c("data", unlist(strsplit(x,"[_.]"))[2]);x0})
>
> y<-Reduce(function(x,y) {merge(x,y,by="data")}, datalist)
>
> W <- log(y[,-1])
>
> res<-princomp(W)
>
> plot(res)
```

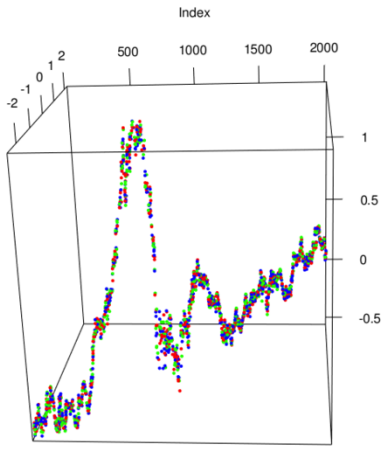


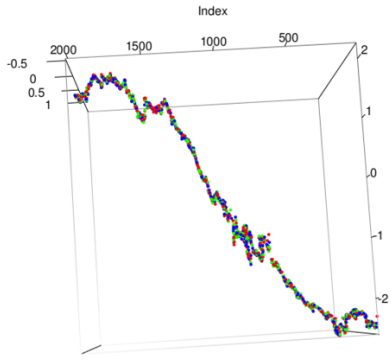

```
> summary(res)
Importance of components:
                          Comp.1    Comp.2     Comp.3     Comp.4     Comp.5
Standard deviation     1.6444750 0.9135597 0.45485234 0.33815092 0.26411864
Proportion of Variance 0.6673281 0.2059485 0.05105352 0.02821669 0.01721405
Cumulative Proportion  0.6673281 0.8732767 0.92433022 0.95254690 0.96976095
                           Comp.6      Comp.7      Comp.8      Comp.9
Standard deviation     0.174935749 0.132478550 0.126864216 0.113280650
Proportion of Variance 0.007551653 0.004330878 0.003971578 0.003166623
Cumulative Proportion  0.977312600 0.981643478 0.985615056 0.988781679
                          Comp.10    Comp.11     Comp.12     Comp.13
Standard deviation     0.102775455 0.10029505 0.084687400 0.063954175
Proportion of Variance 0.002606536 0.00248224 0.001769793 0.001009306
Cumulative Proportion  0.991388215 0.99387046 0.995640248 0.996649554
                           Comp.14     Comp.15      Comp.16      Comp.17
Standard deviation     0.0593856036 0.053672112 0.0482924674 0.0467290743
Proportion of Variance 0.0008702564 0.000710857 0.0005754978 0.0005388392
Cumulative Proportion  0.9975198103 0.998230667 0.9988061651 0.9993450043
                           Comp.18      Comp.19      Comp.20
Standard deviation     0.0345320981 0.0280516066 0.0259800571
Proportion of Variance 0.0002942597 0.0001941781 0.0001665578
Cumulative Proportion  0.9996392640 0.9998334422 1.0000000000
>
```

Помітно виділяються 1, 2 та 3 компоненти.
Подивимось на попарні діаграми розсіювання для наших трьох вибраних компонент:

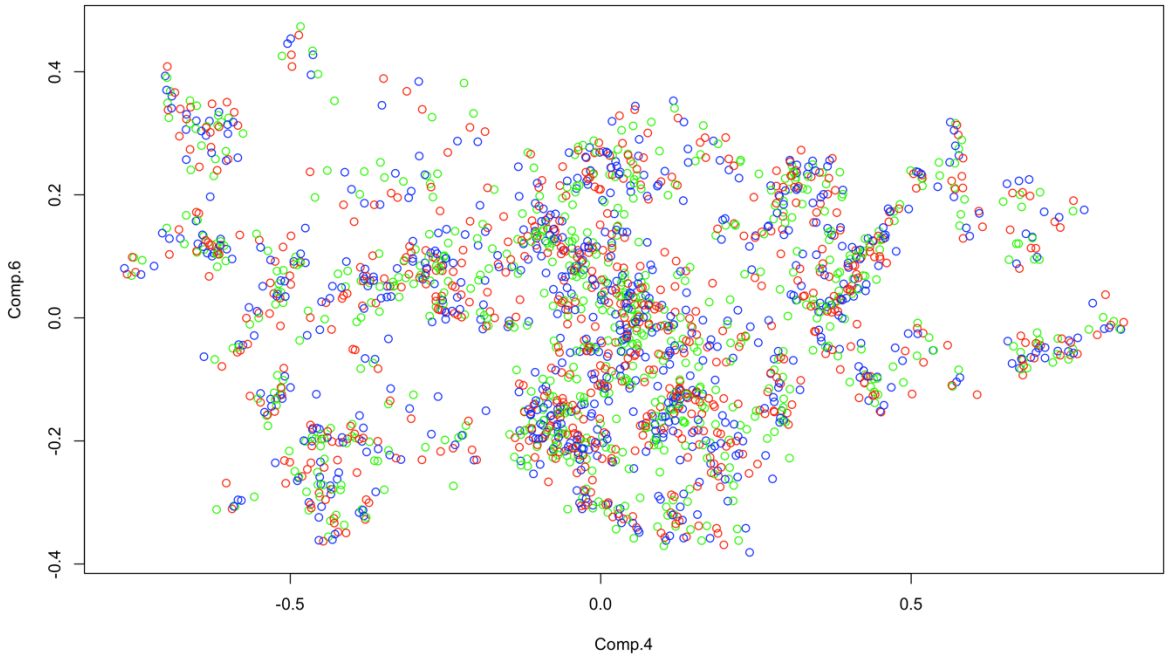


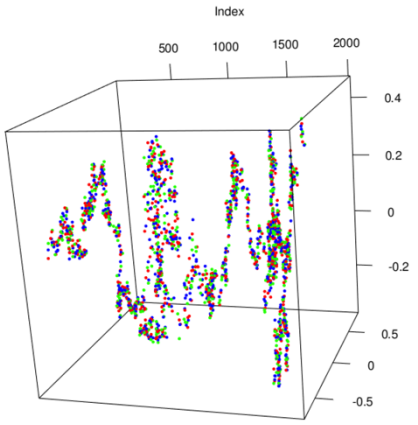
Вийшов дуже цікавий розподіл. Подивимось на нього в 3д:

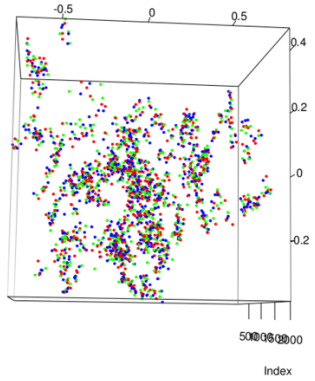


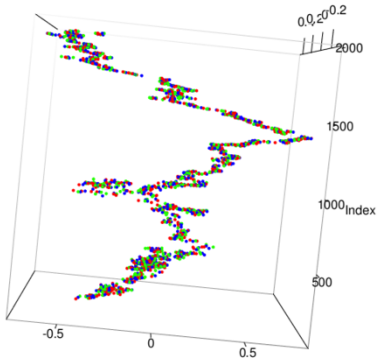


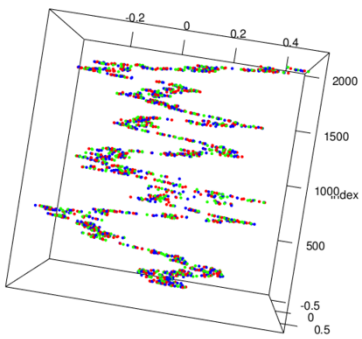
Також мені захотілося подивитися компоненти 4, 5 та 6:


Comp.6
0.4
0.2
0.0
-0.2
-0.4
-0.5
0.0
0.5
Comp.4


Index
500
1000
1500
2000
0.4
0.2
0
-0.2
0.5
0
-0.5

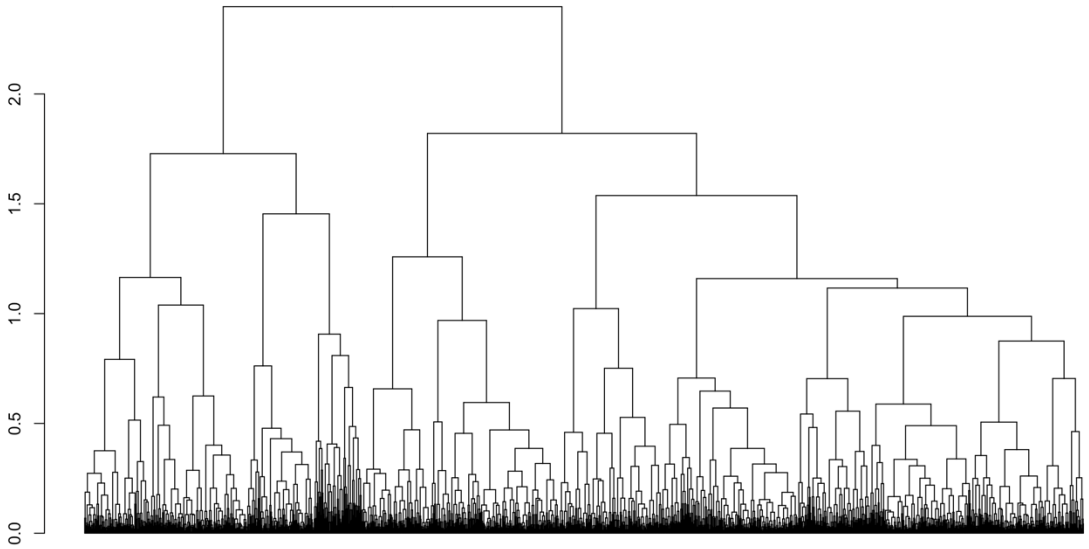

-0.5
0
0.5
0.4
0.2
0
-0.2
Index


0.2
2000
1500
1000
Index
500
-0.5
0
0.5

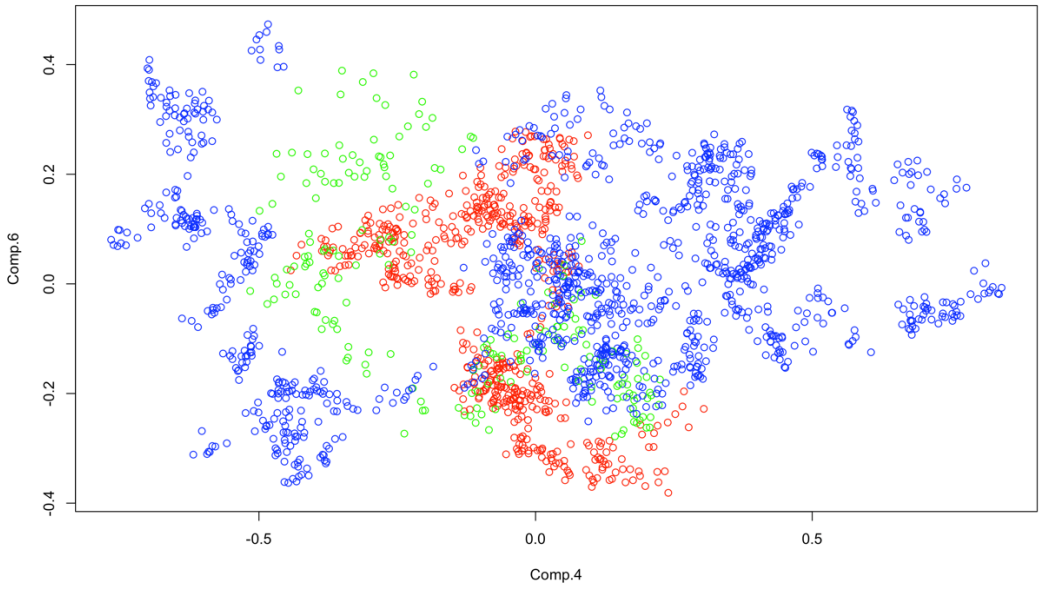

-0.2
0
0.2
0.4
2000
1500
1000
Index
500
-0.5
0
0.5

Зробимо тепер ієрархічний розподіл для компонент 4-6. Використаємо відстань сіті-блок:

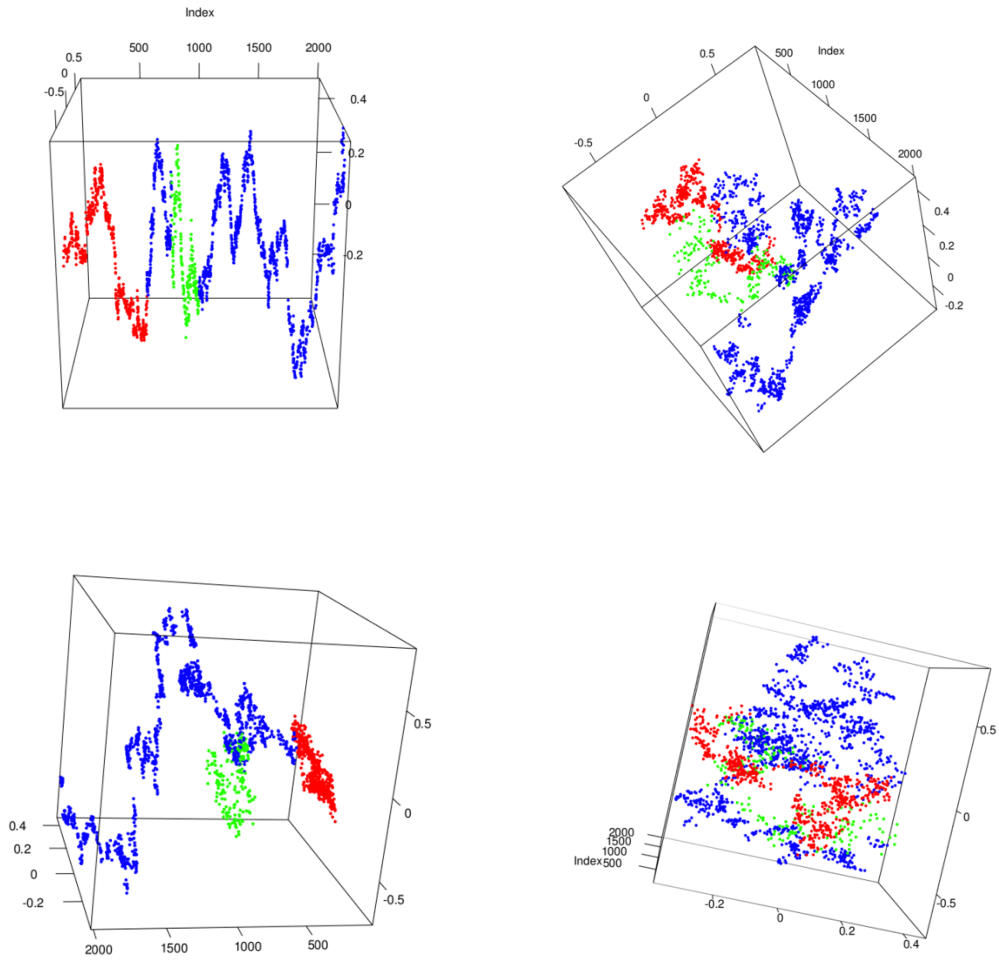
```
> z.dist<-dist(res$scores[,4:6],method ="manhattan")
>
> z.hclust<-hclust(z.dist)
>
> plot(as.dendrogram(z.hclust),leaflab="none")
```




```
> plot(res$scores[,c(4,6)],col=c("red","blue","green")[groups3_0])
```




```
> plot3d(res$scores[,c(4,6)],col=c("red","blue","green")[groups3_0])
```



Отже, я побачила, що на кластери воно розділилося, на 3д графіку синій стикається з зеленим та червоним, але червоний і зелений між собою не перетинаються, але на 2д всі три компоненти між собою перетинаються.