

Мастер академске студије
Рачунарство и аутоматика

Рачунарство високих перформанси
у информационом инжењерингу

Основи кластеризације

(материјали за предавања)

- 1. Кластеризација**
2. Метод срединâ
3. Хијерархијска кластеризација
4. Густинска кластеризација
5. Извори и литература

Кластеризација

кластер

група међусобно сличних (блиских) ентитета

клика, грозд, рој, јато

ентитети из истог кластера начелно су међусобно сличнији (блискији) него што су слични (блиски) ентитетима из других кластера

кластеризација

поступак распоређивања посматраних ентитета у кластере

распоређивање ентитета на основу њихових карактеристика

Кластеризација

дескриптивни поступак

ненадгледано учење

истражна анализа података

кластеризација по врсти ентитета

кластеризација појава

формирање група појава на основу вредности обележја тих појава

кластеризација обележја

Скупови података коришћени у примерима

скуп података **abalone**

подаци о абалонима с Тасманије (Аустралија)

4177 записа

9 обележја

пол, дужина, пречник, висина, различите тежине и број прстенова

датотека *abalone.data*

Abalone Data Set (од 1. 12. 1995)

<https://archive.ics.uci.edu/ml/datasets/Abalone>

(преузето 13. 4. 2021)

UCI Machine Learning Repository

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository
[<http://archive.ics.uci.edu/ml>]. Irvine, CA: University of California, School of
Information and Computer Science.

Кластеризација

Скупови података коришћени у примерима

скуп података **abalone** – припрема

```
1 # install.packages("tidyverse")
2
3 library(readr)
4 library(dplyr)
5 library(magrittr)
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
```

УЛАЗ

Скупови података коришћени у примерима

скуп података **abalone** – припрема

```
1 abalone <- read_csv("abalone.data",  
2                     col_names=c("sex", "length", "diameter",  
3                                 "height", "weight_whole",  
4                                 "weight_shucked", "weight viscera",  
5                                 "weight_shell", "rings"),  
6                     col_types="fdddddddi")  
7  
8 abalone %<>%  
9   mutate(id=1:nrow(abalone), age=rings + 1.5) %>%  
10  select(id, everything())  
11  
12 set.seed(4)  
13  
14  
15  
16  
17  
18  
19  
20
```

УЛАЗ

1. Кластеризација
- 2. Метод срединâ**
3. Хијерархијска кластеризација
4. Густинска кластеризација
5. Извори и литература

Метод срединâ

метод k срединâ

енгл. *k-Means*

формирање партиције скупа ентитета која обухвата k подскупова

сваки подскуп представља један кластер

сваки подскуп је представљен једном средином (центроидом)

ентитет је члан кластера чијој је средини најсличнији (најближи)

k је параметар чију вредност треба унапред задати

k одговара броју кластера

Метод срединâ

метод k срединâ

разноликост унутар кластера треба да буде што мања
унутаркластерска варијабилност за кластер C

$$W(C) = \frac{1}{|C|} \sum_{i,j \in C} \sum_{l=1}^p (x_{il} - x_{jl})^2$$

p је број карактеристика (број обележја у случају кластеризације појава)

укупна унутаркластерска варијабилност за скуп кластера K

$$\sum_{C \in K} \frac{1}{|C|} \sum_{i,j \in C} \sum_{l=1}^p (x_{il} - x_{jl})^2$$

Метод срединâ

метод k срединâ – формирање кластера

насумично распоредити ентитете у k почетних кластера

побољшавати квалитет k кластера

за сваки кластер одредити средину (центроид)

ентитете распоредити у оне кластере чије су им средине најсличније (најближе)

еуклидска удаљеност као мера различитости

понављати одређивање средина и распоређивање ентитета докле год долази до смањења укупне унутаркластерске варијабилности за k кластера

Метод срединâ

метод k срединâ – мањкавости

могућа конвергенција према локалном оптимуму

осетљивост на почетни распоред по кластерима

могуће размотрити већи број почетних стања

осетљивост на скале карактеристика

могуће спровести центрирање и скалирање

потребно одредити број кластера k

постоје помоћни поступци за испитивање погодности одређеног броја кластера

могуће је увести кориговани критеријум завршетка

ограничити број итерација

Метод срединаâ

Метод срединаâ

пример

```
1 abalon.km <- abalon %>%  
2   select(length, height)  
3  
4 k <- 3  
5  
6 km <- abalon.km %>%  
7   select(length, height) %>%  
8     kmeans(centers=k)  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20
```

УЛАЗ

Метод средина

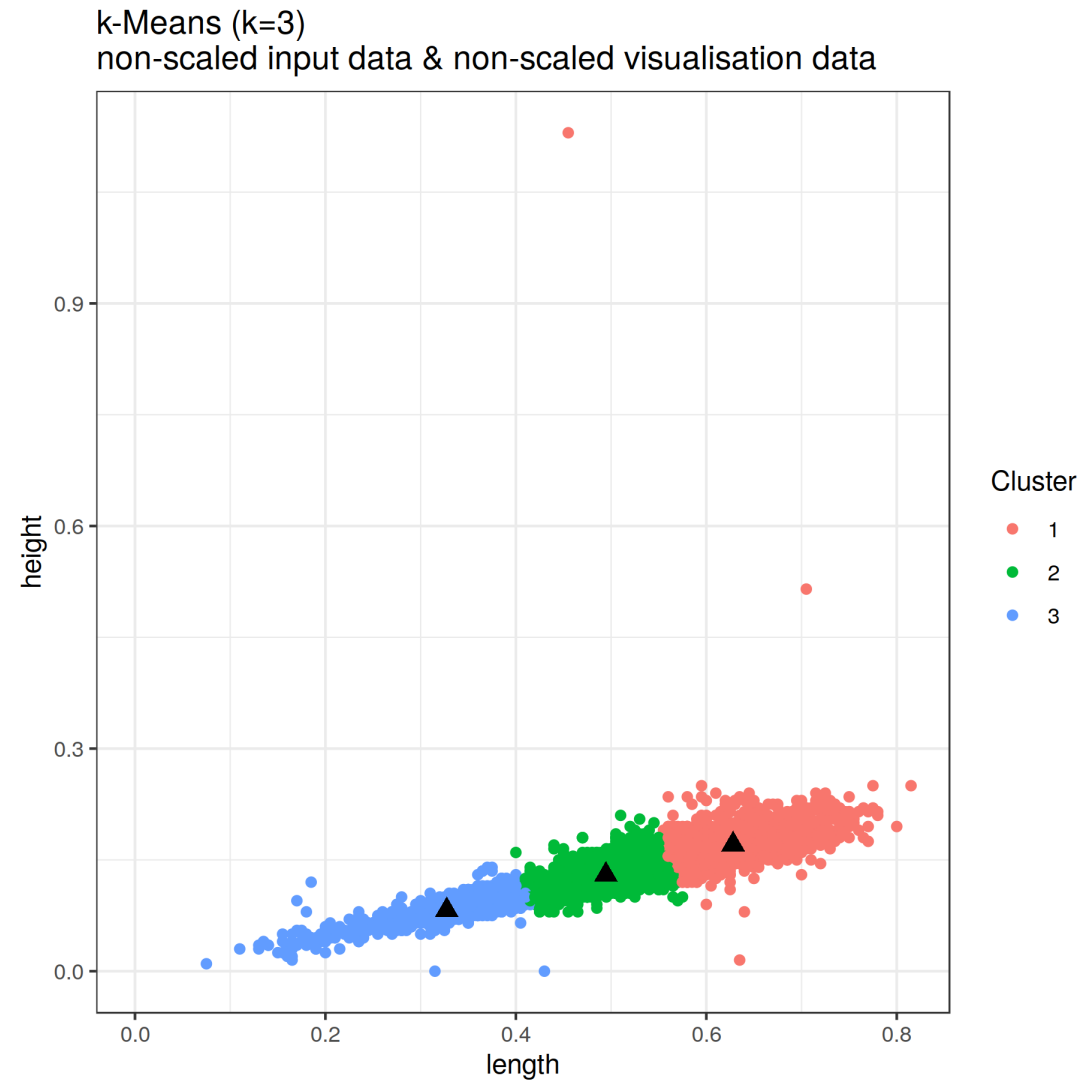
Метод средина

пример

```
> tail(km$cluster, 20)
[1] 2 2 2 1 1 3 3 3 3 2 2 2 2 2 1 1 1 1 1 1
> km$centers
      length      height
1 0.6281112 0.17082029
2 0.4945503 0.12985117
3 0.3274228 0.08214094
> km$withinss
[1] 5.960992 3.350221 3.563362
>
```

КОНЗОЛА

Метод срединâ пример



Метод срединаâ

Метод срединаâ

пример

```
1 abalon.km <- abalon %>%  
2   select(length, height)  
3  
4 k <- 3  
5  
6 km.skali <- abalon.km %>%  
7   select(length, height) %>%  
8   scale(center=T, scale=T) %>%  
9   kmeans(centers=k)  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20
```

УЛАЗ

Метод средина

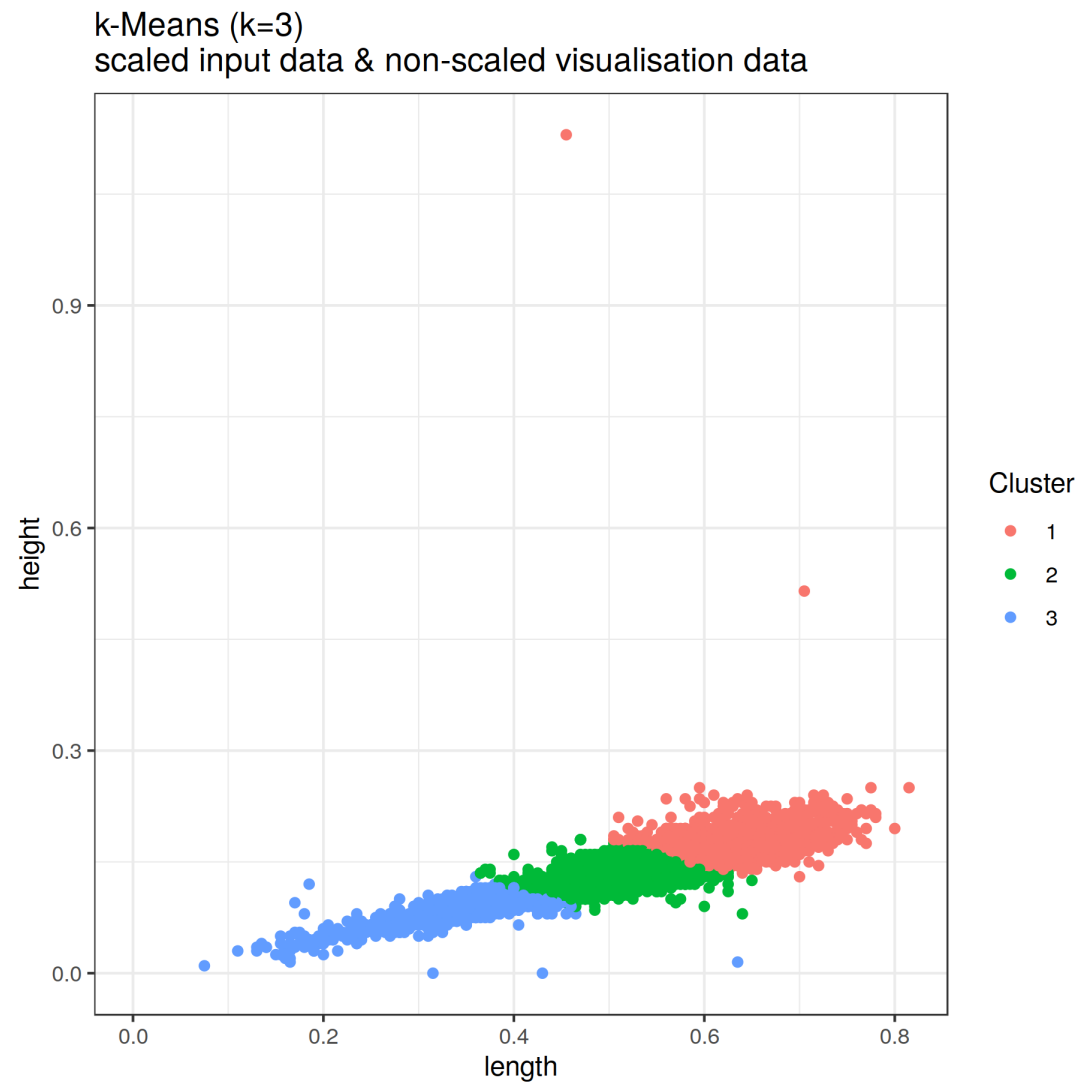
Метод средина

пример

```
> tail(km.skali$cluster, 20)
[1] 2 2 2 1 1 3 3 3 3 2 2 2 2 2 2 1 2 1 1 1
> km.skali$centers
      length      height
1  0.9000298  0.8700491
2 -0.1245050 -0.1909469
3 -1.5508725 -1.3529666
> km.skali$withinss
[1] 1209.4235  519.2548  464.3167
>
```

КОНЗОЛА

Метод срединâ пример



Метод срединаâ

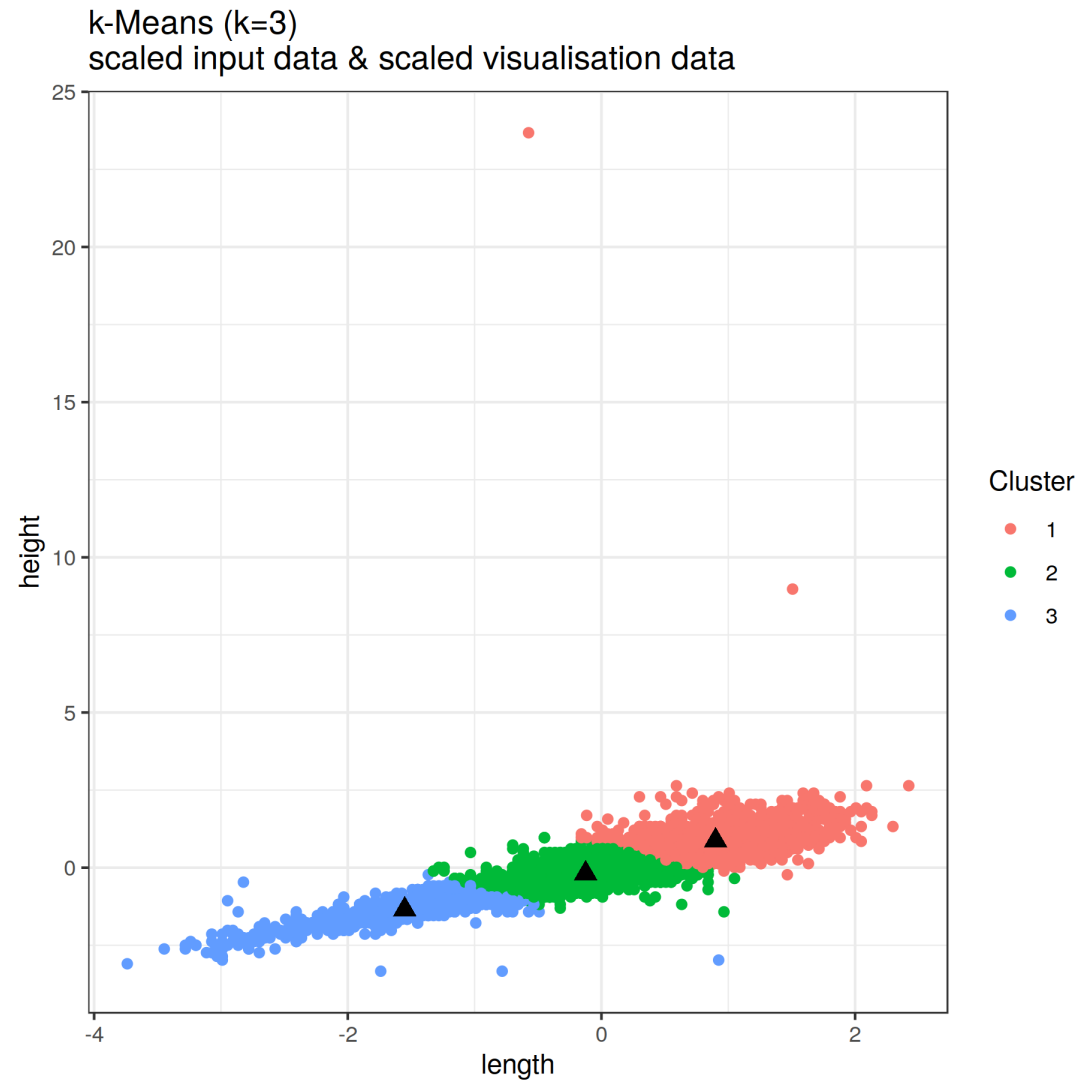
Метод срединаâ

пример

```
1 abalon.km.skali <- abalon.km %>%  
2   select(length, height) %>%  
3   scale(center=T, scale=T) %>%  
4   as.data.frame()  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20
```

УЛАЗ

Метод срединâ пример



1. Кластеризација
2. Метод срединâ
- 3. Хијерархијска кластеризација**
4. Густинска кластеризација
5. Извори и литература

Хијерархијска кластеризација

Хијерархијска кластеризација

формирање кластера који су хијерархијски организовани

кластер на дну хијерархије обухвата један ентитет

кластер који није на дну хијерархије обухвата све ентитете својих подређених кластера

из хијерархије начелно могуће издвајати различите бројеве кластера

за формирање хијерархије кластера није неопходно задати број кластера

представа хијерархије кластера кроз дендограм

две варијанте хијерархијске кластеризације

кроз поступак спајања (енгл. *agglomerative*)

agnes (agglomerative nesting)

кроз поступак дељења (енгл. *divisive*)

diana (divisive analysis)

Хијерархијска кластеризација

Хијерархијска кластеризација

хијерархијска кластеризација кроз спајање – формирање кластера

- за сваки ентитет формирати засебан кластер

- спојити два најсличнија (најближа) кластера

 - сличност (блискост) утврдити на основу изабране мере различитости

 - спајање најсличнијих кластера понављати док се не оствари постојање само једног кластера

 - на основу извршених спајања могуће је формирати одговарајуће стабло и представити га кроз дендограм

 - сваки чвор одговара једном кластеру формираном током кластеризације

 - корен одговара кластеру са свим ентитетима

 - лист одговара неком од могућих кластера са једним ентитетом

Хијерархијска кластеризација

хијерархијска кластеризација кроз спајање – формирање кластера

повезаност (енгл. *linkage*)

мера различитости између две групе ентитета

често коришћене врсте повезаности

појединачна (енгл. *single*)

одговара најмањој вредности различитости уоченој између нека два ентитета из различитих група

средња (енгл. *average*)

одговара средњој вредности различитости уочених између ентитета из различитих група

потпуна (енгл. *complete*)

одговара највећој вредности различитости уоченој између нека два ентитета из различитих група

Хијерархијска кластеризација

Хијерархијска кластеризација

хијерархијска кластеризација кроз спајање – формирање кластера

мере различитости између ентитета

мере засноване на удаљености

еуклидска удаљеност

Манхетн удаљеност

Махаланобисова удаљеност

...

мере засноване на корелацији

Хијерархијска кластеризација

Хијерархијска кластеризација

пример

```
1 abalon.hc <- abalon %>%  
2   filter(sex=="M") %>%  
3   select(weight_whole, rings)  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20
```

УЛАЗ

Хијерархијска кластеризација

Хијерархијска кластеризација

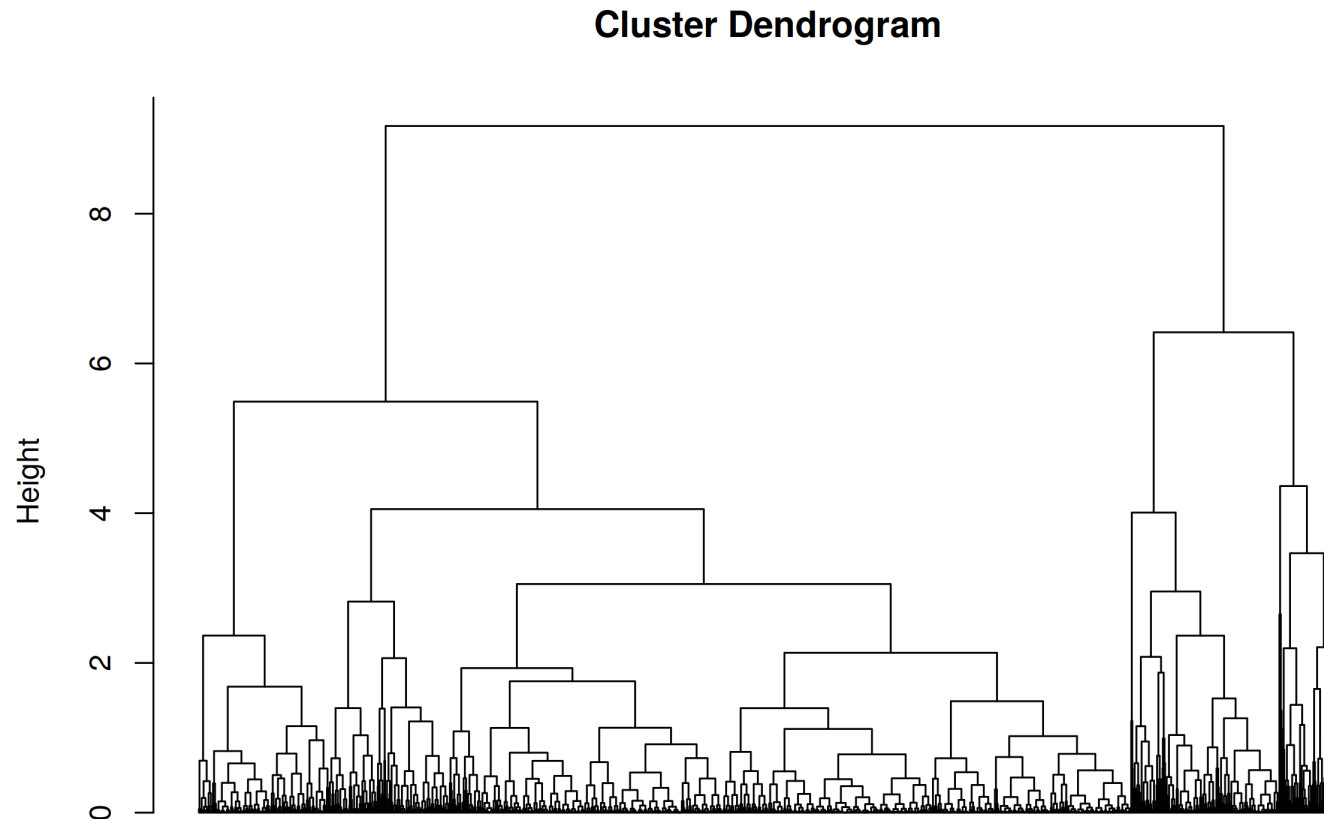
пример

```
1 hc.cmp.euc <- hclust(dist(scale(abalon.hc, center=T, scale=T),
2                       method="euclidean"), method="complete")
3
4 hc.cmp.euc.clust <- cutree(hc.cmp.euc, k=3)
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
```

УЛАЗ

Хијерархијска кластеризација

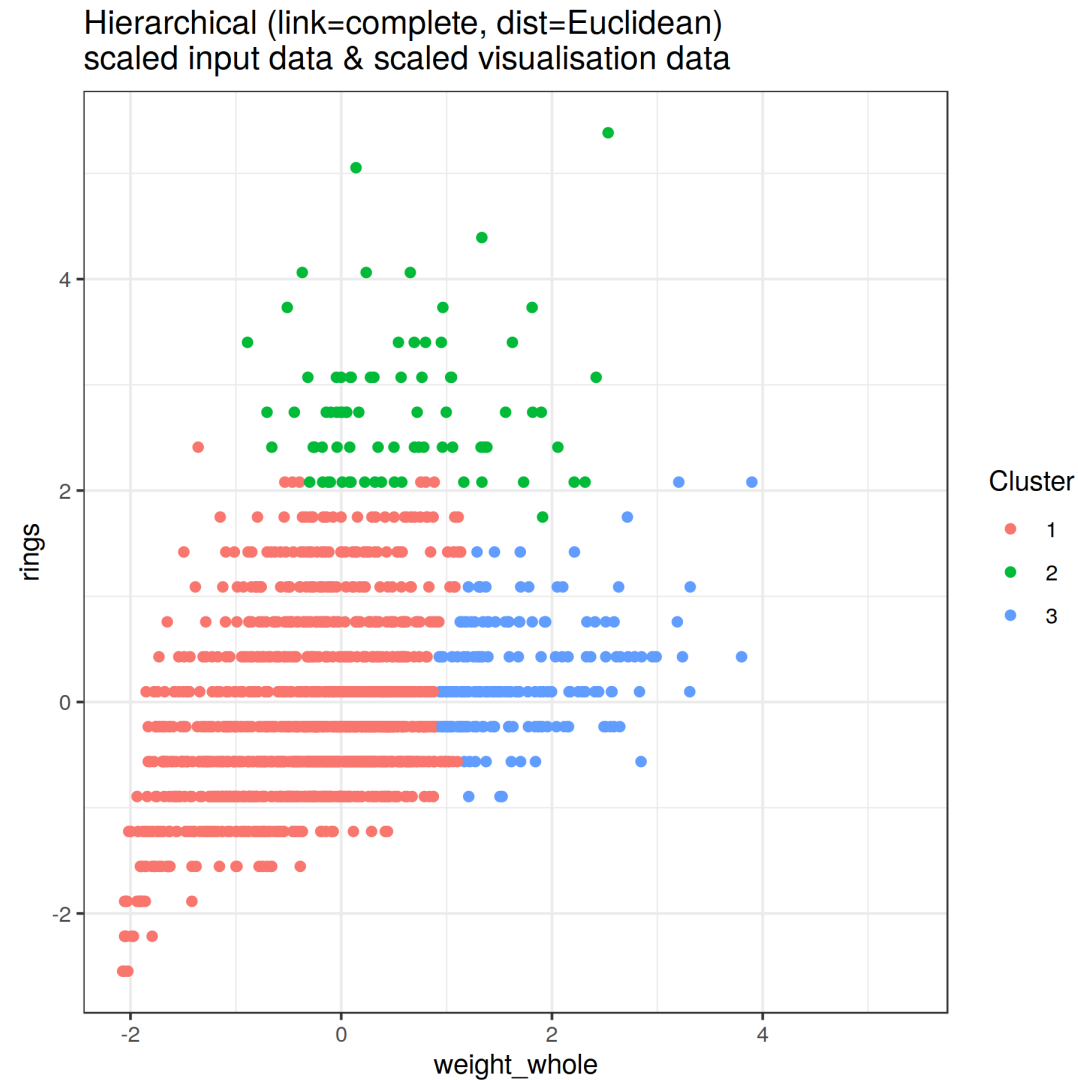
Хијерархијска кластеризација пример



```
dist(scale(abalon.hc, center = T, scale = T), method = "euclidean")  
hclust (*, "complete")
```

Хијерархијска кластеризација

Хијерархијска кластеризација пример



Хијерархијска кластеризација

Хијерархијска кластеризација

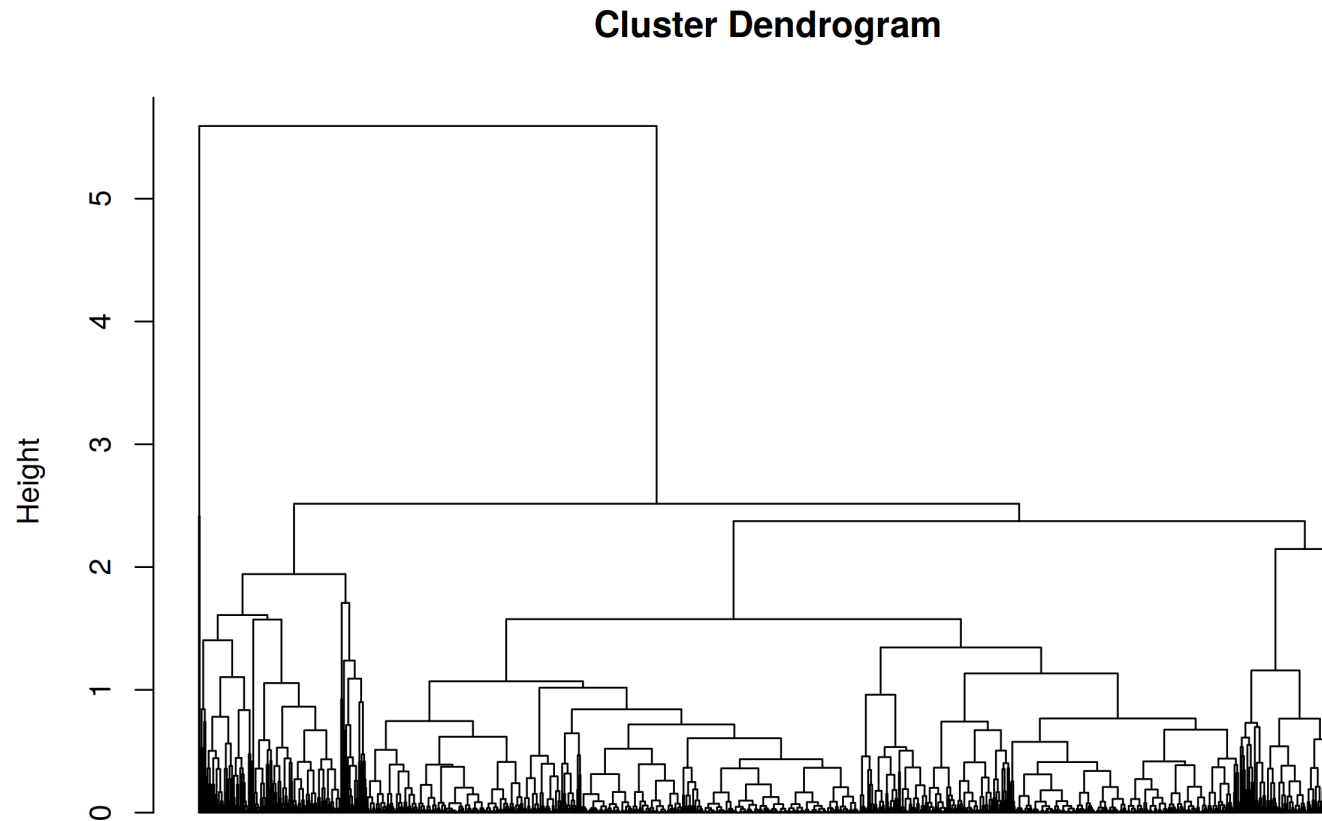
пример

```
1 hc.avg.euc <- hclust(dist(scale(abalon.hc, center=T, scale=T),
2                       method="euclidean"), method="average")
3
4 hc.avg.euc.clust <- cutree(hc.avg.euc, k=3)
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
```

УЛАЗ

Хијерархијска кластеризација

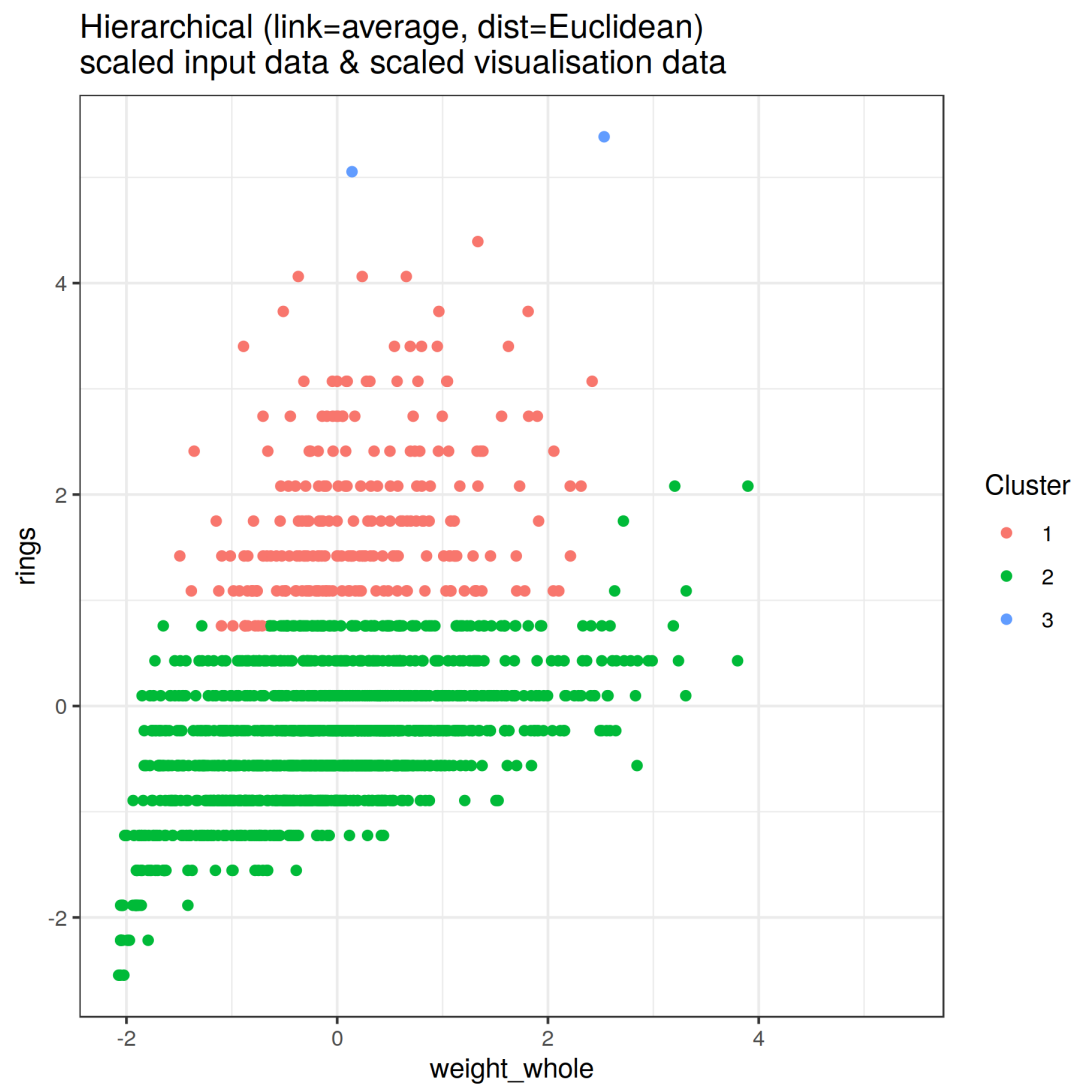
Хијерархијска кластеризација пример



```
dist(scale(abalon.hc, center = T, scale = T), method = "euclidean")  
hclust (*, "average")
```

Хијерархијска кластеризација

Хијерархијска кластеризација пример



Хијерархијска кластеризација

Хијерархијска кластеризација

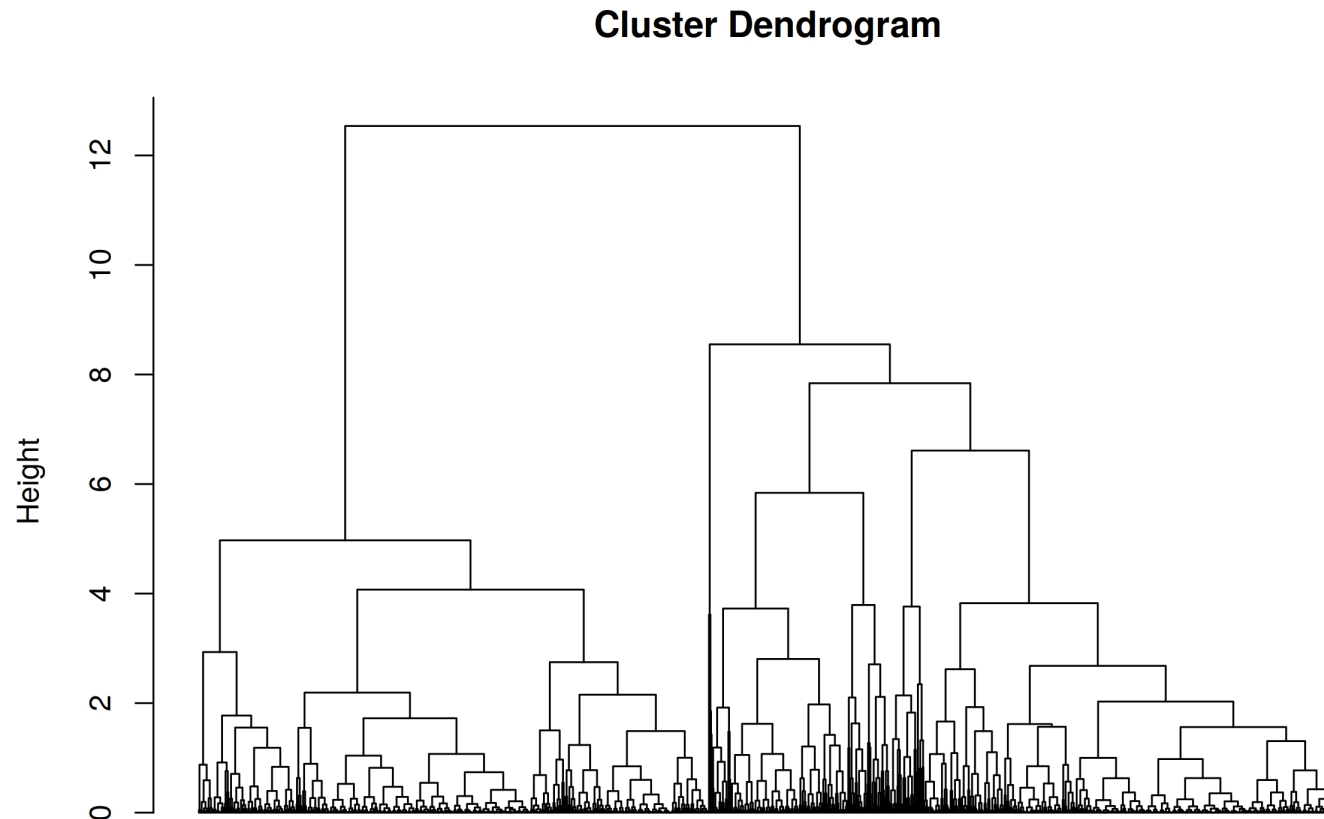
пример

```
1 hc.cmp.man <- hclust(dist(scale(abalon.hc, center=T, scale=T),
2                               method="manhattan"), method="complete")
3
4 hc.cmp.man.clust <- cutree(hc.cmp.man, k=3)
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
```

УЛАЗ

Хијерархијска кластеризација

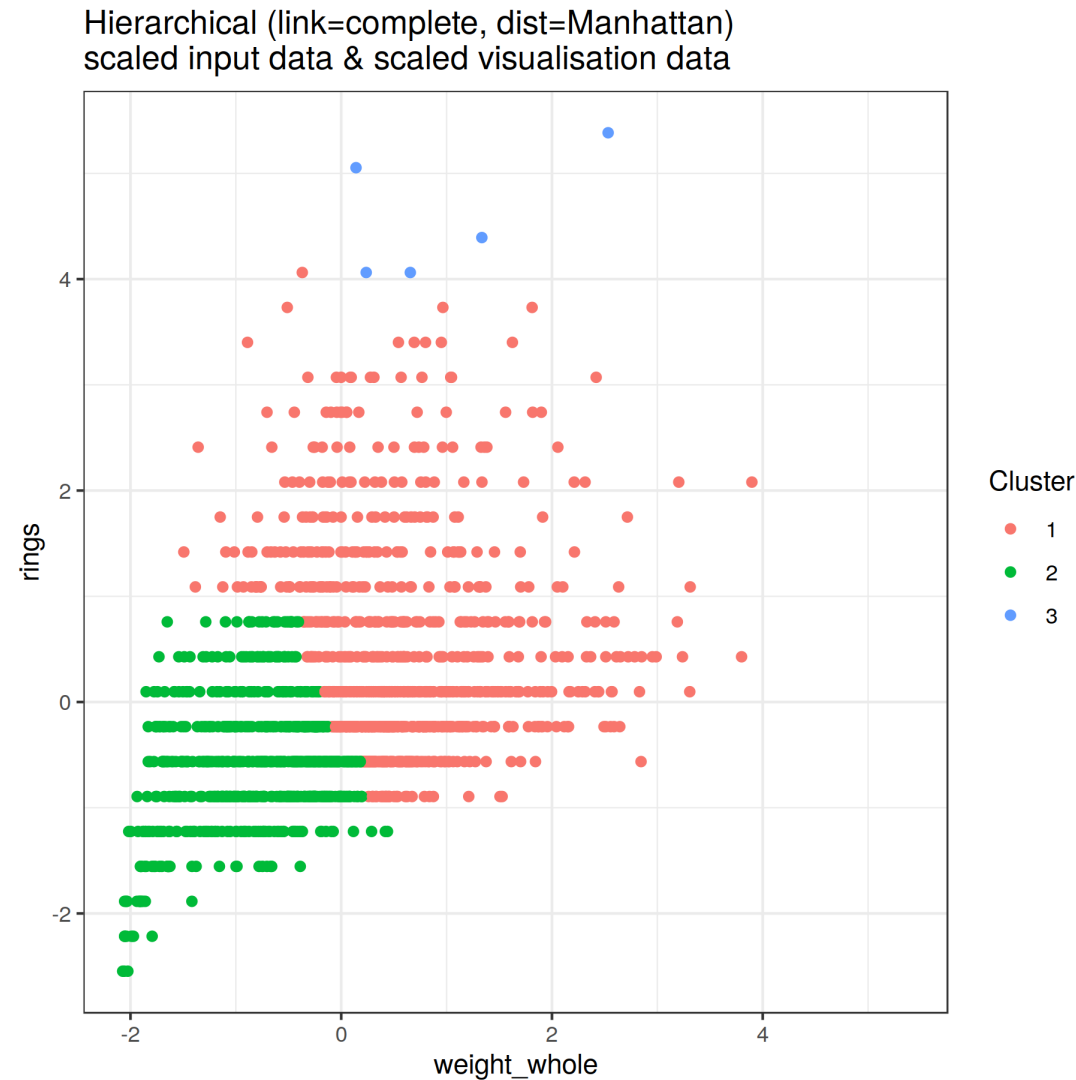
Хијерархијска кластеризација пример



```
dist(scale(abalon.hc, center = T, scale = T), method = "manhattan")  
hclust (*, "complete")
```

Хијерархијска кластеризација

Хијерархијска кластеризација пример



Хијерархијска кластеризација

Хијерархијска кластеризација

пример

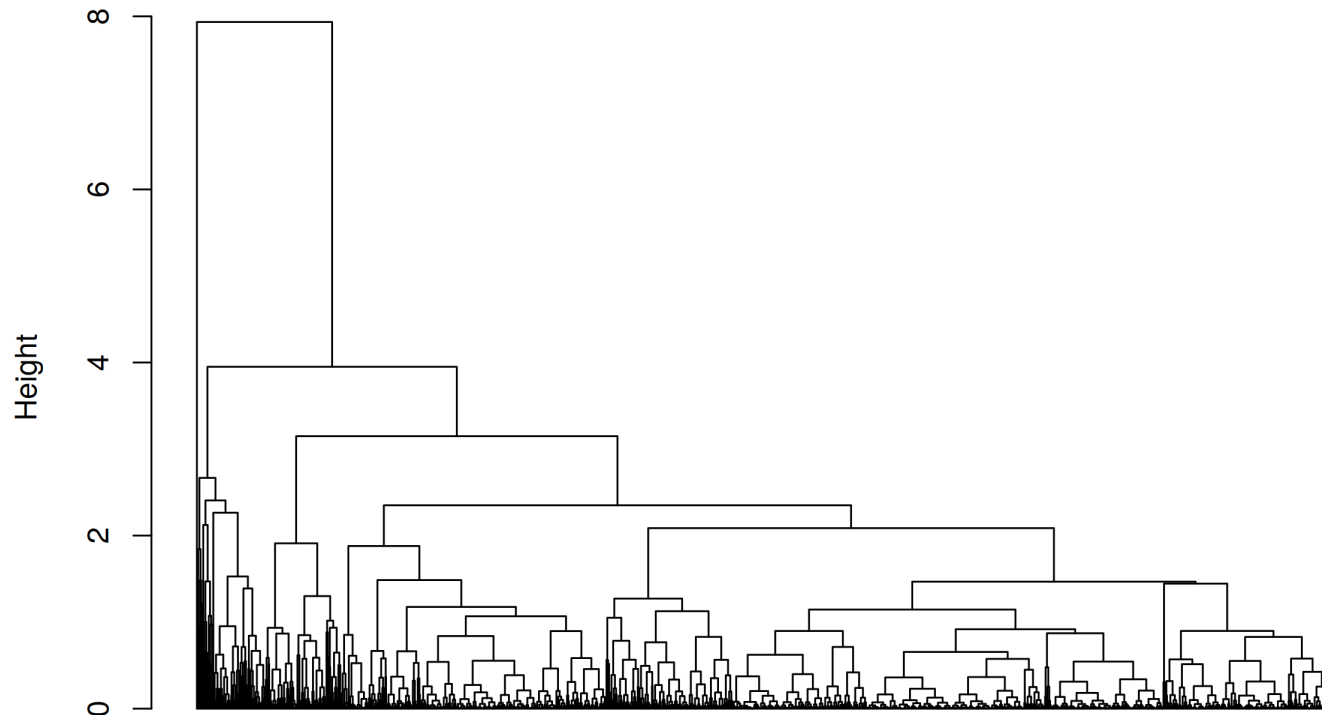
```
1 hc.avg.man <- hclust(dist(scale(abalon.hc, center=T, scale=T),
2                       method="manhattan"), method="average")
3
4 hc.avg.man.clust <- cutree(hc.avg.man, k=3)
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
```

УЛАЗ

Хијерархијска кластеризација

Хијерархијска кластеризација пример

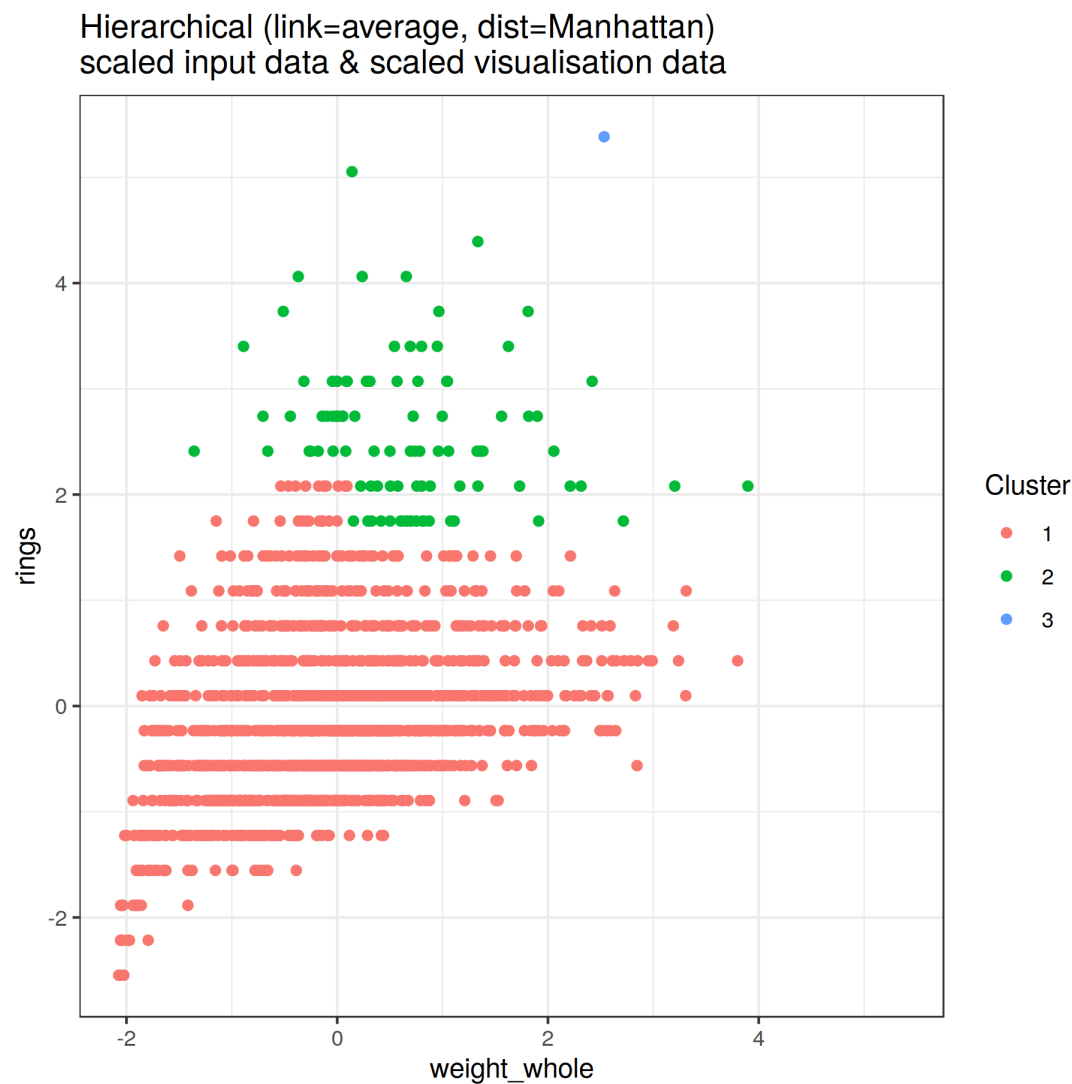
Cluster Dendrogram



```
dist(scale(abalon.hc, center = T, scale = T), method = "manhattan")  
hclust (*, "average")
```

Хијерархијска кластеризација

Хијерархијска кластеризација пример



1. Кластеризација
2. Метод срединâ
3. Хијерархијска кластеризација
- 4. Густинска кластеризација**
5. Извори и литература

Густинска кластеризација

кластер као густо попуњен повезани део простора
алгоритам *DBSCAN* (*Density-based spatial clustering of applications with noise*)

ентитети су у истом кластеру ако су међусобно густински достижни

густина се посматра у ϵ околини неког ентитета

битно је да ли у постоји одређени минимални број ентитета у ϵ околини посматраног ентитета

три врсте ентитета

основни ентитет

ентитет у чијој се ϵ околини налази барем минимални број ентитета

гранични ентитет

није основни ентитет али јесте у ϵ околини основног ентитета

шум

није ни основни ни гранични ентитет

параметри

ϵ као величина околине (позитиван реалан број)

минимални број ентитета за ϵ околину (позитиван цео број)

Густинска кластеризација

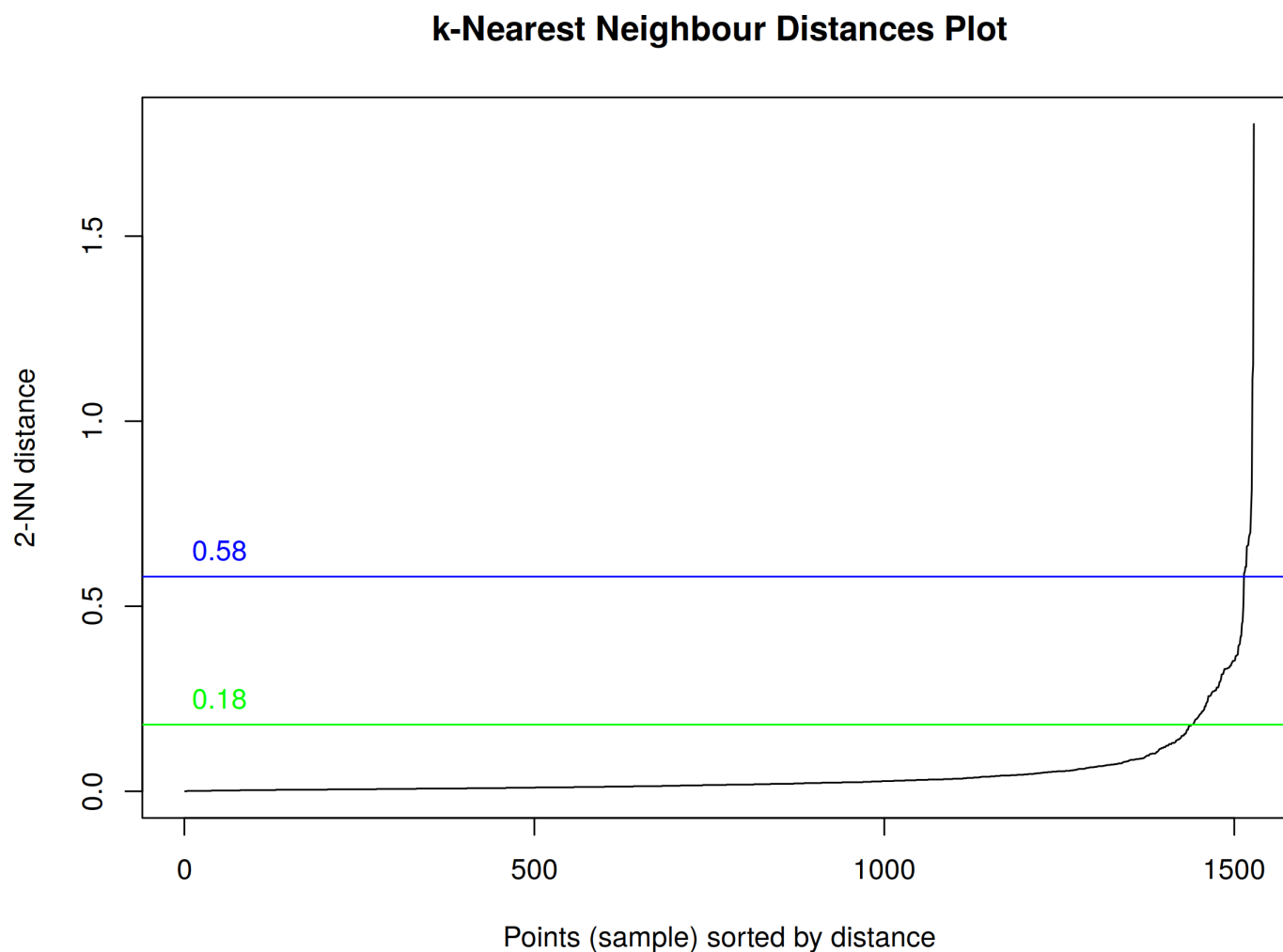
Густинска кластеризација

пример

```
1 library(dbscan)
2
3 abalon.dbs <- abalon %>%
4   filter(sex=="M") %>%
5   select(weight_whole, rings)
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
```

УЛАЗ

Густинска кластеризација пример



Густинска кластеризација

Густинска кластеризација

пример

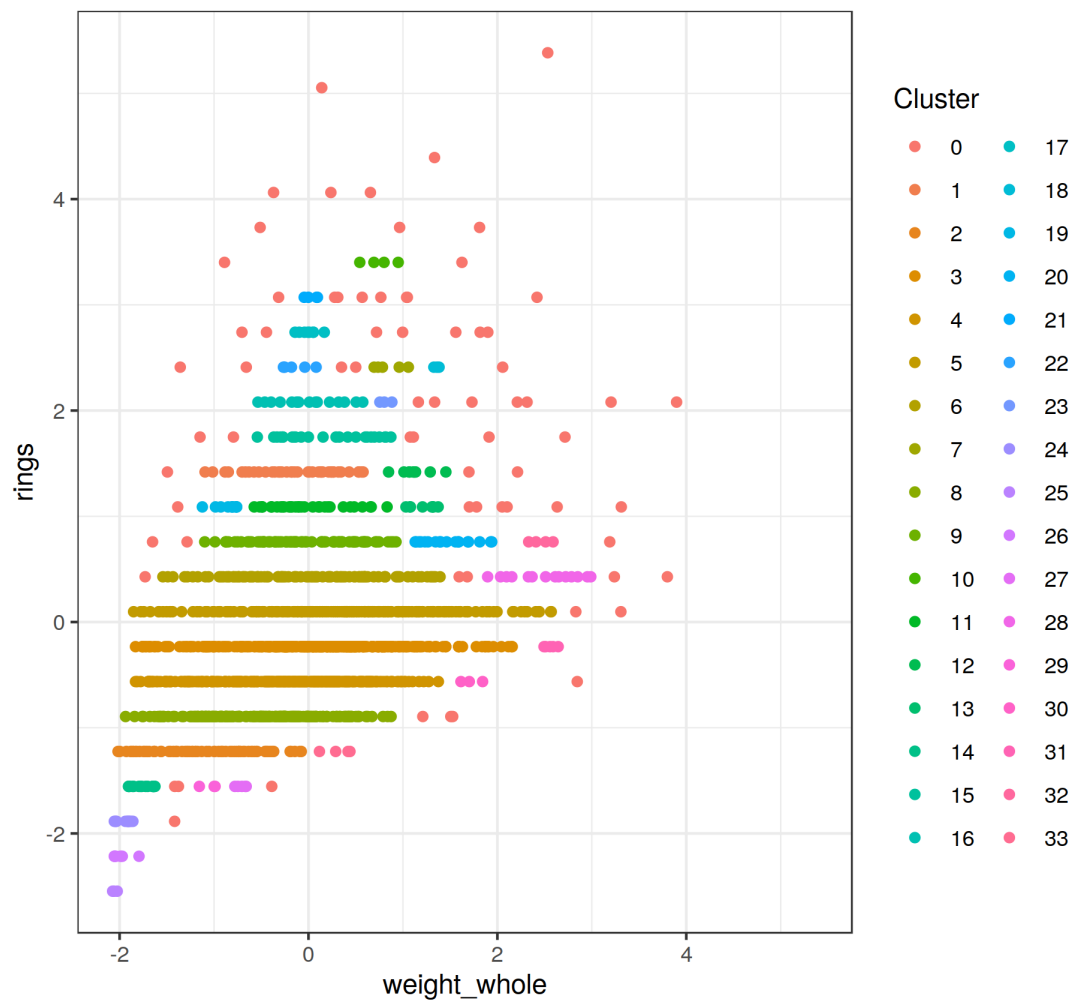
```
1 dbs.low <- dbscan(scale(abalon.dbs, center=T, scale=T),  
2                      eps=0.18, minPts=3)  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20
```

УЛАЗ

Густинска кластеризација

пример

DBSCAN (eps=0.18, minPts=3)
scaled input data & scaled visualisation data



Густинска кластеризација

Густинска кластеризација

пример

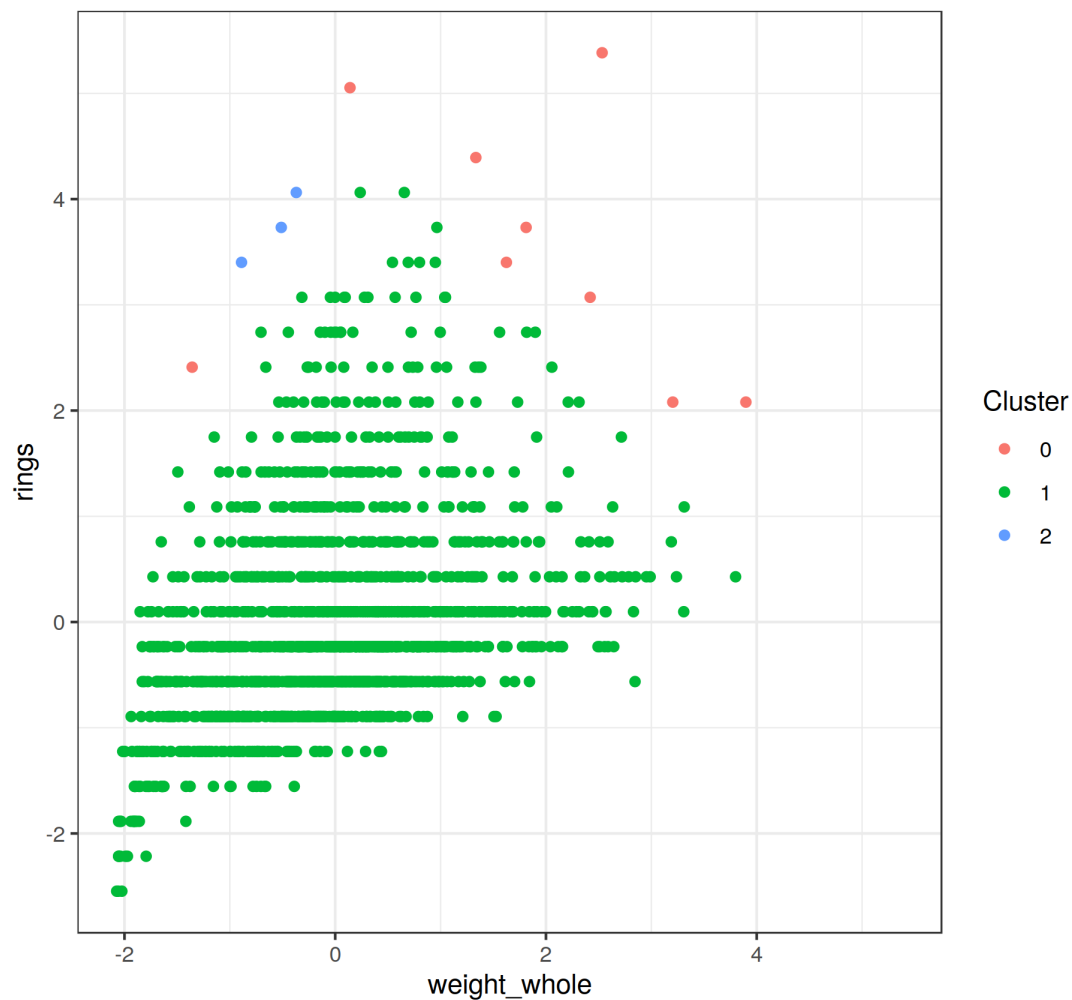
```
1 dbs.high <- dbscan(scale(abalon.dbs, center=T, scale=T),  
2                      eps=0.58, minPts=3)  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20
```

УЛАЗ

Густинска кластеризација

пример

DBSCAN (eps=0.58, minPts=3)
scaled input data & scaled visualisation data



1. Кластеризација
2. Метод срединâ
3. Хијерархијска кластеризација
4. Густинска кластеризација
- 5. Извори и литература**

Основни извори и литература

- ◆ James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning: With applications in R. Springer; 2013.
- ◆ R: A language and environment for statistical computing – Reference index – The R core team – Version 4.2.3 (2023-03-15). Internet:
<https://cran.r-project.org/doc/manuals/r-release/fullrefman.pdf>
- ◆ CRAN - Package dbscan. Internet:
<https://cran.r-project.org/web/packages/dbscan/index.html>

Основни извори података

- ◆ скуп података **abalone**
 - ◆ UCI Machine Learning Repository
 - ◆ Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [<http://archive.ics.uci.edu/ml>]. Irvine, CA: University of California, School of Information and Computer Science.
 - ◆ Подаци о абалонима с Тасманије (Аустралија)
 - ◆ датотека *abalone.data*
 - ◆ Abalone Data Set (од 1. 12. 1995)
 - ◆ <https://archive.ics.uci.edu/ml/datasets/Abalone> (преузето 13. 4. 2021)

Мастер академске студије
Рачунарство и аутоматика

Рачунарство високих перформанси
у информационом инжењерингу

Основи кластеризације

(материјали за предавања)