Ngôn ngữ R - PHƯƠNG PHÁP HIỂN THỊ DỮ LIỆU

Ths. Vũ Duy Khương

Nội dung

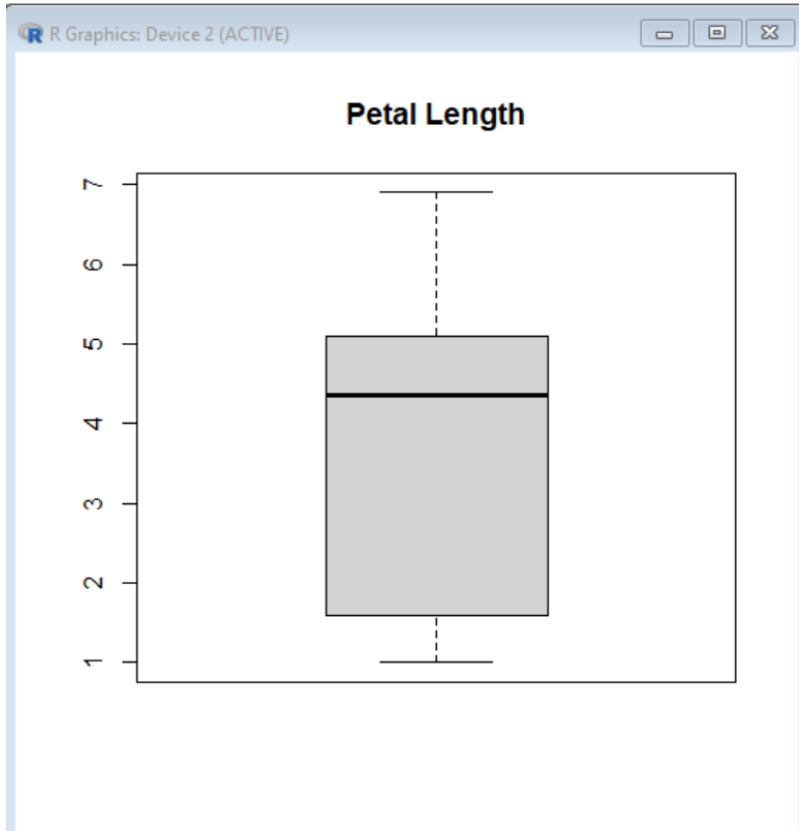
1. PHƯƠNG PHÁP HIỂN THỊ DỮ LIỆU
2. HỒI QUY TUYẾN TÍNH TRONG R
3. LẬP TRÌNH TRONG MÔI TRƯỜNG R
4. BÀI THỰC HÀNH

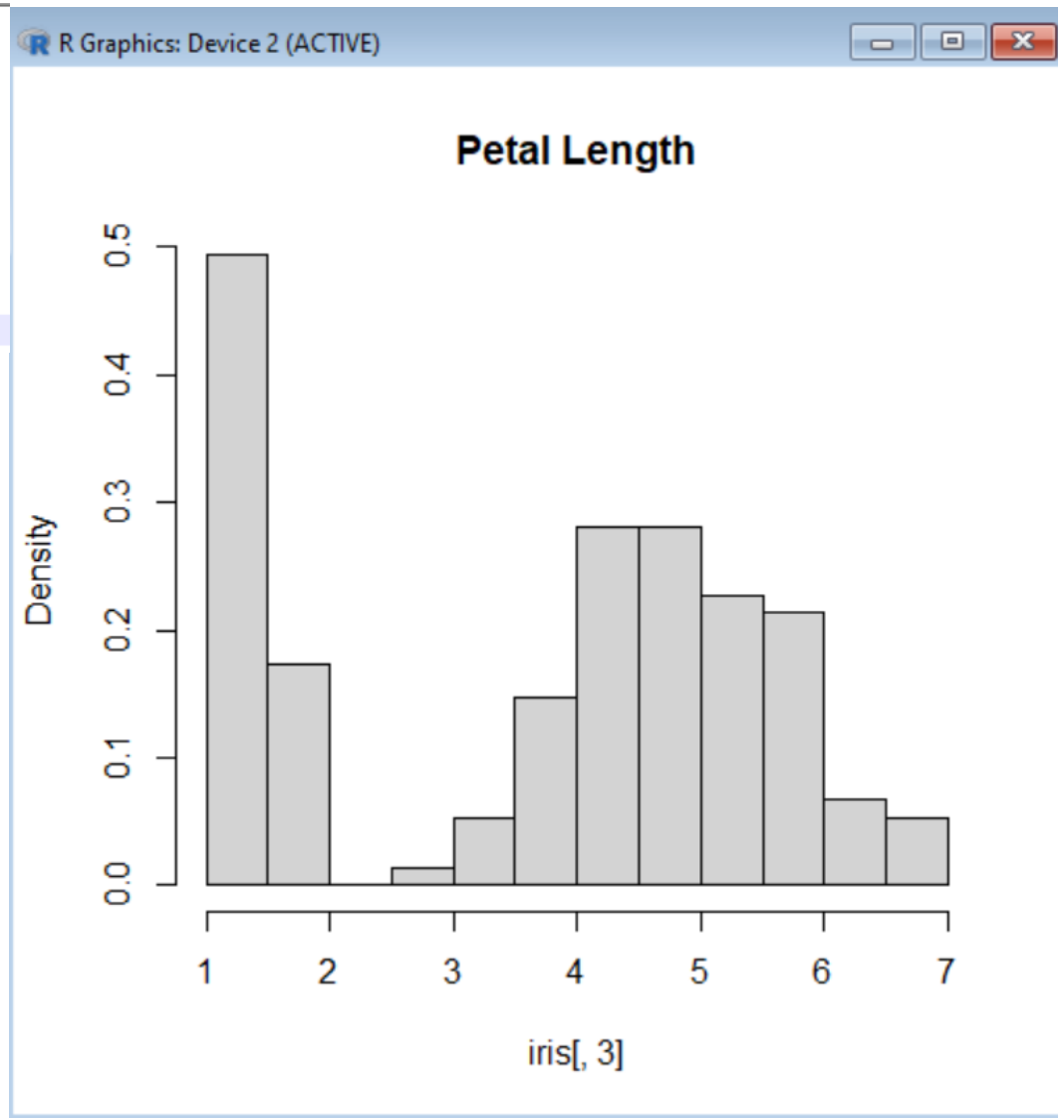
PHƯƠNG PHÁP HIỂN THỊ DỮ LIỆU

- Trong phân tích thống kê, biểu đồ đóng vai trò rất quan trọng. Nếu biểu đồ được trình bày dễ hiểu, hợp lý, là phương tiện hiệu quả cung cấp cho nhà phân tích về thông tin quan trọng của dữ liệu cần phân tích. Phương pháp hiển thị dữ liệu thường dùng trong thống kê có thể kể đến là đồ thị hộp (box plot) và tổ chức đồ (histogram). Đồ thị hộp của 1 thuộc tính hiển thị giá trị nhỏ nhất, trung vị, lớn nhất, bách phân 25% và 75%. Tổ chức đồ hiển thị thông tin về phân bố dữ liệu của một thuộc tính.

Ví dụ minh họa vẽ đồ thị hộp và tổ chức đồ

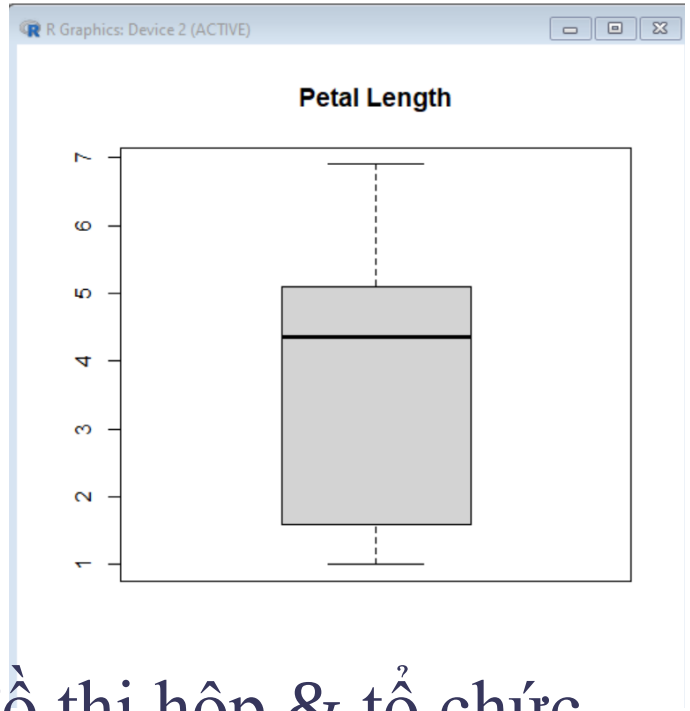
```r
# doc du lieu, thuoc tinh thu 3 la Petal Length
data(iris)

# ve do thi to chuc do
hist(iris[,3], main = "Petal Length", ylab = "Density", prob = T)

# ve do thi hop
boxplot(iris[,3], main = "Petal Length")
```

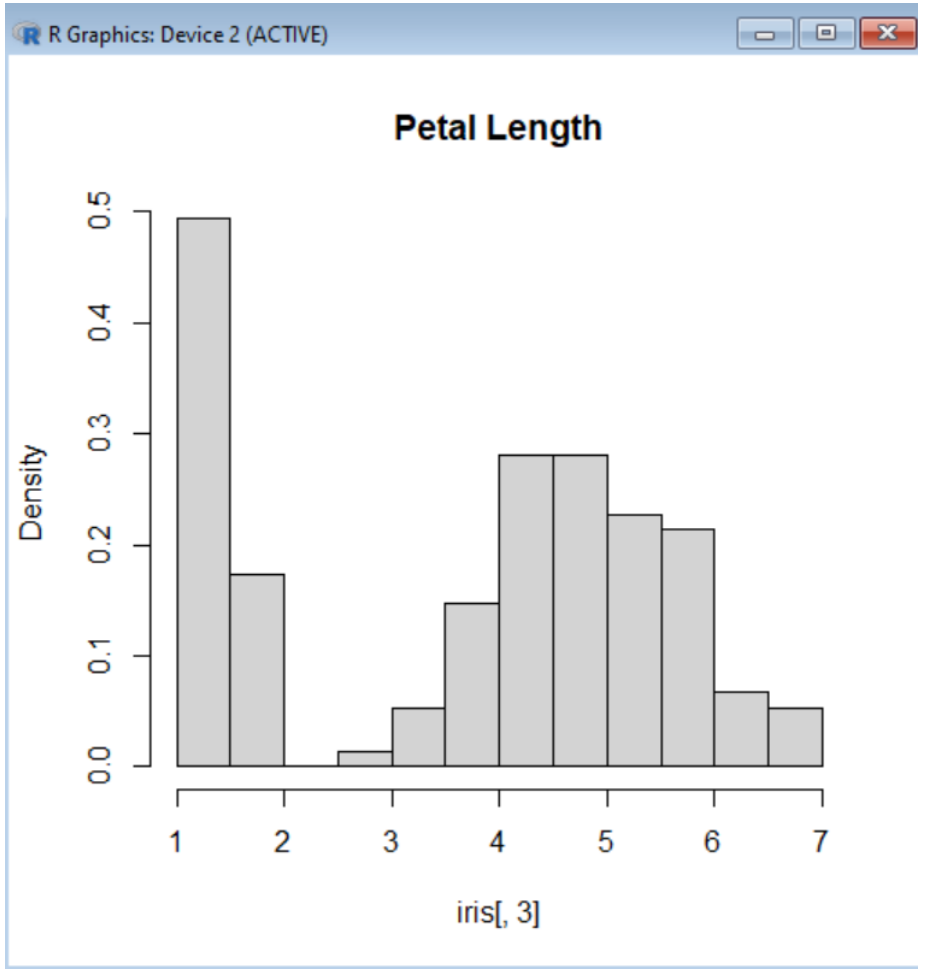




Hình: Đồ thị hộp và tổ chức đồ của thuộc tính Petal Length dữ liệu iris

Ví dụ minh họa vẽ đồ thị hộp và tổ chức đồ





Ví dụ trong hình là đồ thị hộp & tổ chức đồ của thuộc tính Petal Length dữ liệu iris. Đoạn mã trong ngôn ngữ R cho phép thực hiện vẽ đồ thị hình. Trong đó dòng lệnh 5 thực hiện vẽ tổ chức đồ.
Dòng lệnh 8 vẽ đồ thị hộp.

Phương pháp scatterplot 2 chiều trong R

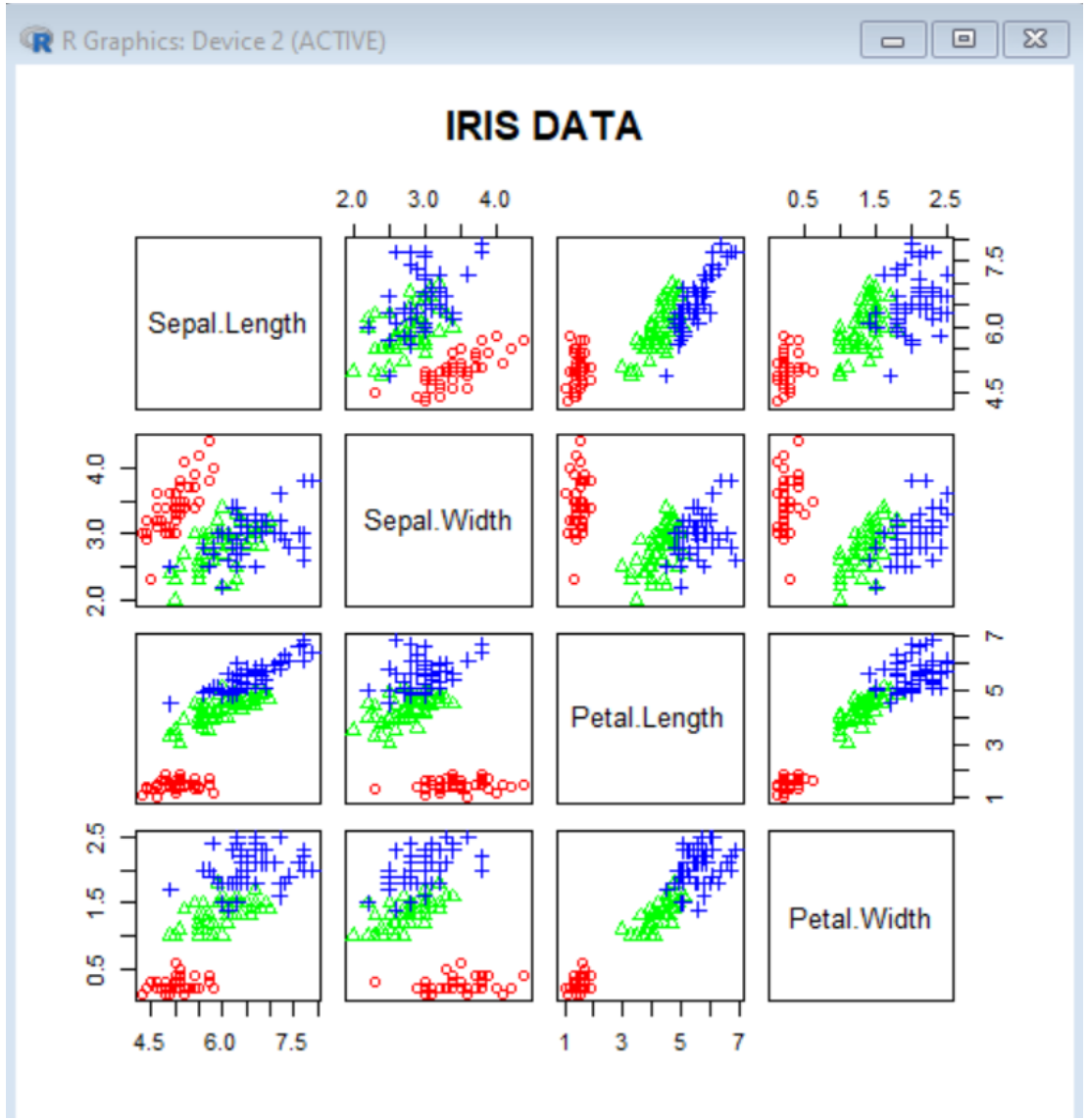
Tập dữ liệu có n thuộc tính được hiển thị trong ma trận scatterplot 2 chiều, ở đó mỗi scatterplot 2 chiều trình bày dữ liệu của cặp thuộc tính Ai, Aj, mỗi phần tử được vẽ thành 1 điểm trong không gian 2 chiều của scatterplot.

Ví dụ minh họa vẽ ma trận scatterplot 2 chiều

```
12
13  # doc du lieu
14  data(iris)
15  ncol <- length(iris[1,])
16
17  # ve ma tran scatterplot 2 chieu
18  pairs(iris[,-ncol], main = "IRIS DATA", pch = c(1, 2, 3)[iris[,ncol]], col = c("red", "green", "blue")[iris[,ncol]])
```

Ví dụ như ma trận scatterplot 2 chiều hiển thị tập dữ liệu **iris** (150 phần tử, 4 thuộc tính, 3 lớp tương ứng với màu của điểm dữ liệu) như hình sau. Để vẽ ma trận scatterplot 2 chiều cho ví dụ này, ta có thể viết đoạn mã trong **R** với hàm **pairs()** như bảng sau. Dòng lệnh 5 thực hiện vẽ ma trận scatterplot 2 chiều.

Ví dụ minh họa vẽ ma trận scatterplot 2 chiều



Ưu nhược điểm vẽ ma trận scatterplot 2 chiều

- **Ưu điểm :**

Phương pháp ma trận scatterplot 2 chiều dễ hiểu, hỗ trợ cho phát hiện mối tương quan của dữ liệu theo từng cặp thuộc tính, có thể phát hiện được nhóm dữ liệu, hiểu sơ lược về dữ liệu phức tạp hay không.

- **Nhược điểm:**

Tuy nhiên khi số thuộc tính dữ liệu lớn lên, ma trận scatterplot có kích thước tăng nhanh (n 2 scatterplots với n là số thuộc tính dữ liệu). Phương pháp cũng chỉ hiển thị khoảng vài ngàn phần tử dữ liệu.

Phương pháp trục tọa độ song song trong R

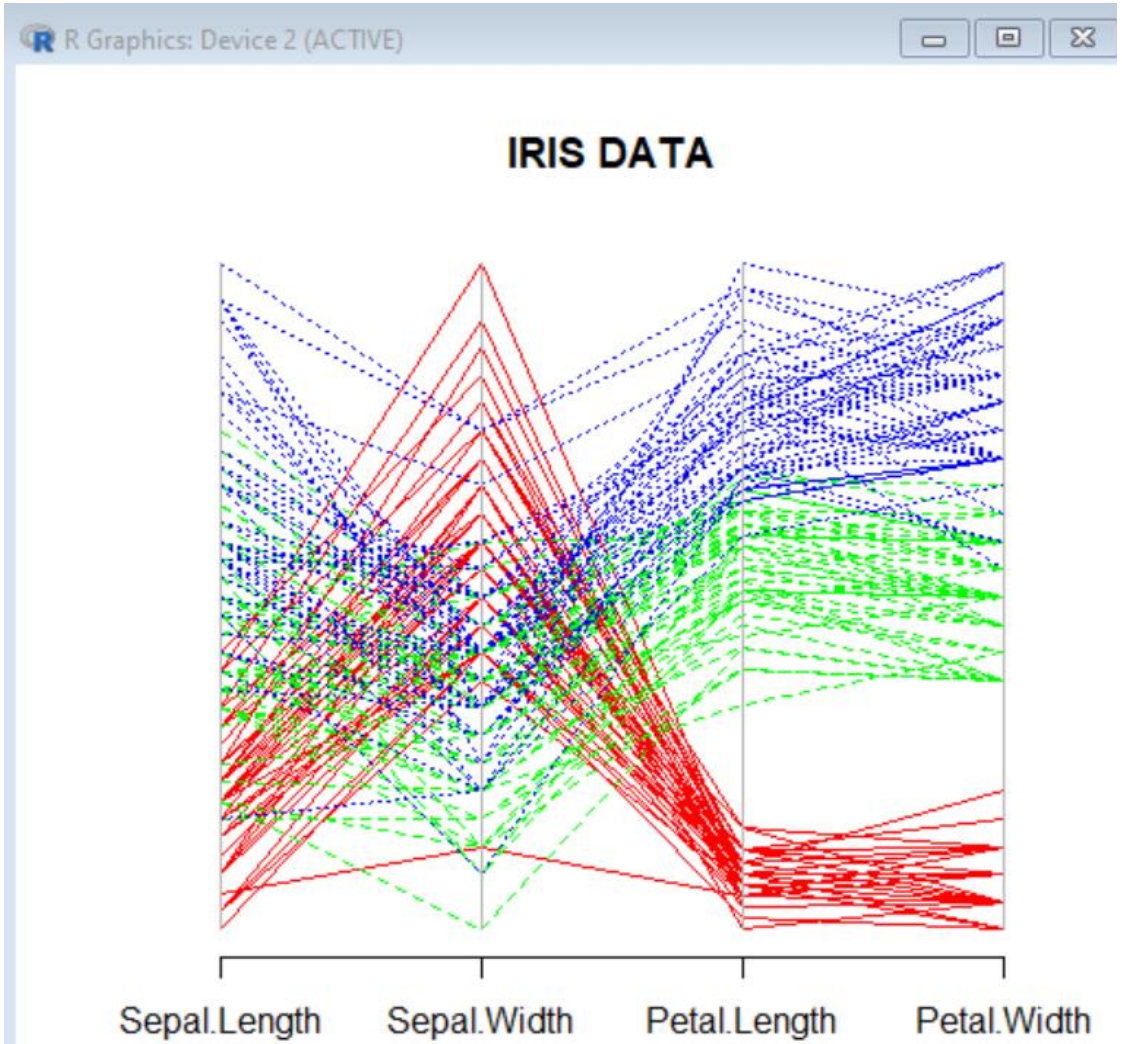
- Dữ liệu có n thuộc tính trong hệ tọa độ Descartes được biểu diễn thành n trục tọa độ song song bằng nhau (một chiều trong hệ tọa độ Descartes là một trục tọa độ song song).
- Một phần tử trong hệ tọa độ Descartes trở thành 1 đường gấp khúc, giao điểm của đường gấp khúc với trục tọa độ song song chính là tọa độ chiều tương ứng của điểm đó trong không gian Descartes. Phương pháp cho phép hiển thị cùng lúc khoảng 50 thuộc tính, vài ngàn phần tử dữ liệu.
- Hệ tọa độ song song dễ hiểu, hỗ trợ cho phát hiện nhóm dữ liệu, hiểu sơ lược về dữ liệu phức tạp hay không.
- Tuy nhiên thứ tự của các trục tọa độ ảnh hưởng lớn đến chất lượng hiển thị dữ liệu.

Ví dụ minh họa hiển thị dữ liệu với hệ tọa độ song song

```
23  # nap thu vien MASS
24  library(MASS)
25
26  # doc du lieu
27  data(iris)
28  ncol <- length(iris[1,])
29
30  # hien thi du lieu voi he toa do song song
31  parcoord(iris[,-ncol], main = "IRIS DATA", lty = c(1, 2, 3)[iris[,ncol]], col = c("red", "green", "blue")[iris[,ncol]])
```

- Trong gói thư viện MASS có hàm parcoord() cho phép thực hiện hiển thị dữ liệu với hệ trục tọa độ song song.
- Đoạn mã chương trình R trong hình sau cho phép hiển thị dữ liệu iris trong hệ tọa độ song song.
- Dòng lệnh 31 thực hiện hiển thị dữ liệu với hệ trục tọa độ song song. Kết quả hiển thị như hình (màu của điểm dữ liệu tương ứng với lớp hay nhãn).

Ví dụ minh họa hiển thị dữ liệu với hệ tọa độ song song



Ưu nhược điểm hiển thị dữ liệu với hệ tọa độ song song

- Tất cả các phương pháp đều có ưu, nhược điểm khác nhau. Chúng ta nên kết hợp các phương pháp hiển thị để tận dụng các ưu điểm và khắc phục được khiếm khuyết khi sử dụng đơn thuần một phương pháp.
- Cùng tập dữ liệu, mỗi phương pháp hiển thị cung cấp thông tin hữu dụng khác nhau.

Phương pháp hiển thị khác trong R

- R còn cung cấp nhiều hàm tiện ích khác cho việc hiển thị dữ liệu. Thông thường R hỗ trợ hai loại hàm vẽ đồ họa:

  - Mức cao (vẽ toàn bộ đồ thị bằng 1 lời gọi hàm duy nhất) bao gồm: **barplot(), boxplot(), contour(), coplot(), hist(), pairs(), persp(), plot(), pie(),**

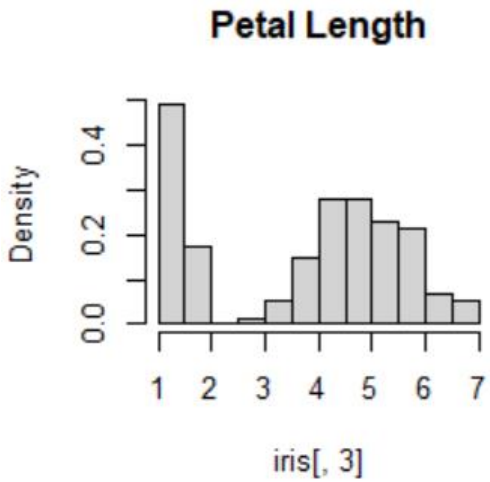
  - Mức thấp (được sử dụng để thêm các thông tin vào đồ thị đã vẽ trước) bao gồm: **abline(), axis(), legend(), lines(), points(), polygon(), symbols()** và **text().**

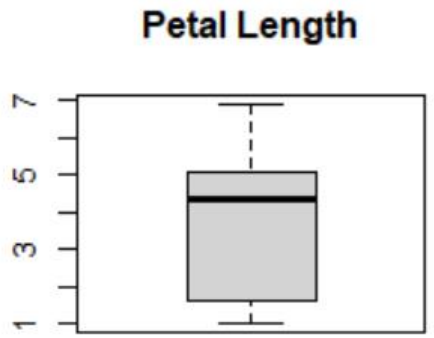
- Ngoài ra ta còn có các hàm phụ trợ khác như tạm dừng với **par(ask=T),** màn hình đồ họa có thể được chia thành dạng lưới cho phép trình bày nhiều đồ thị trên cùng màn hình thông qua hàm **par(mfrow=c(số_dòng, số_cột)).**

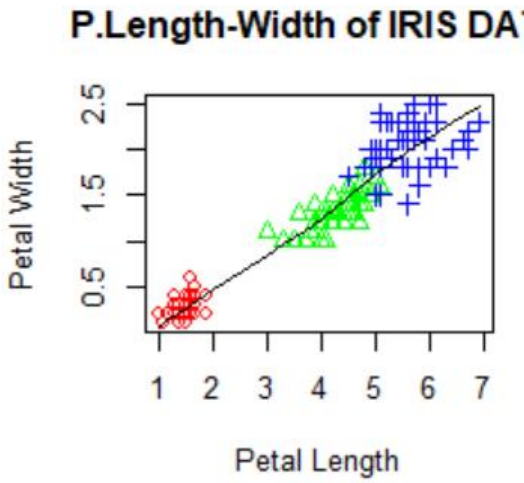
Phương pháp hiển thị khác trong R

- Ví dụ tiếp đây sẽ chia màn hình hiển thị thành 4 phần
  - Phần 1 dùng để hiển thị tổ chức đồ cho thuộc tính Petal Length,
  - Phần 2 vẽ đồ thị hộp của thuộc tính Petal Length,
  - Phần 3 dùng để vẽ đồ thị của 2 thuộc tính Petal Length, Petal Width và đồ thị pie được vẽ trên phần còn lại như hình sau.
- Đoạn mã chương trình trong bảng cho phép thực hiện công việc trên. Dòng lệnh 4 cho phép chia màn hình thành lưới 2x2.

Phương pháp hiển thị khác trong R







Phương pháp hiển thị khác trong R

```r
# nap thu vien MASS
library(MASS)

# doc du lieu
data(iris)
ncol <- length(iris[1,])

# hien thi du lieu voi he toa do song song
parcoord(iris[,-ncol], main = "IRIS DATA", lty = c(1, 2, 3)[iris[,ncol]], col = c("red", "green", "blue")[iris[,ncol]])


# nap du lieu
data(iris)

# chia man hinh 2x2
op <- par(mfrow = c(2, 2))

# ve to chuc do thuoc tinh thu 3 la Petal Length
hist(iris[,3], main = "Petal Length", ylab = "Density", prob = T)

# ve do thi hop cua Petal Length
boxplot(iris[,3], main = "Petal Length")

# ve scatterplot 2 chieu cua p.length-width
plot(iris[,3:4], main = "P.Length-Width of IRIS DATA", xlab = "Petal Length", ylab = "Petal Width", pch = c(1, 2, 3)[iris[,ncol]],
 col = c("red", "green", "blue")[iris$Species])

# ve duong hoi quy
lines(lowess(iris[,3:4]))

# ve do thi pie
nb1 <- sum(c(1, 0, 0)[iris$Species])
nb2 <- sum(c(0, 1, 0)[iris$Species])
nb3 <- sum(c(0, 0, 1)[iris$Species])
pie(c(nb1, nb2, nb3), labels = c("setosa", "versicolor", "virginica"), density = c(10, 20, 30), col = c("red", "green", "blue"))
```

Hồi quy tuyến tính trong R

- Định nghĩa :

Hồi quy là phương pháp toán học được áp dụng thường xuyên trong thống kê để phân tích mối liên hệ giữa các hiện tượng kinh tế xã hội. Để minh họa cho vấn đề này, chúng ta xét một ví dụ rất đơn giản hồi quy tuyến tính đơn mà ở đó người ta cần nghiên cứu để biết chiều cao trung bình của trẻ dựa theo tháng tuổi.

Hồi quy tuyến tính trong R

Bảng số liệu thu thập được trình bày trong bảng sau đây:

STT	Tháng tuổi	Chiều cao (cm)
1	18	76.1
2	19	77
3	20	78.1
4	21	78.2
5	22	78.8
6	23	79.7
7	24	79.9
8	25	81.1
9	26	81.2
10	27	81.8
11	28	82.8
12	29	83.5

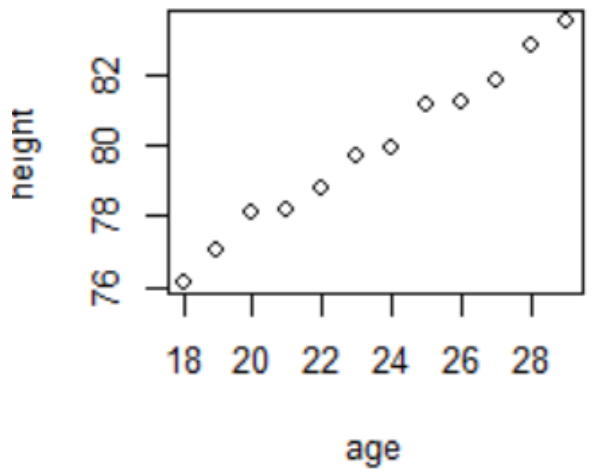
Chiều cao trung bình của trẻ theo tháng tuổi

Hồi quy tuyến tính trong R

Hiển thị bảng dữ liệu với scatterplot 2 chiều trong R như sau:

```
66  age=18:29
67  height=c(76.1,77,78.1,78.2,78.8,79.7,79.9,81.1,81.2,81.8,82.8,83.5)
68  plot(age,height)
```

Đồ thị scatterplot 2 chiều thu được đồ thị về chiều cao trung bình của trẻ theo tháng tuổi như hình sau



Hồi quy tuyến tính trong R

Nhìn vào đồ thị, chúng ta có thể thấy được mối liên quan giữa chiều cao và tuổi của trẻ có dạng tương tự như đường thẳng (tuyến tính). Mô hình hồi quy tuyến tính có dạng:

$$y = \boldsymbol{\alpha} + \boldsymbol{\beta}x$$

với **α** là chặn (intercept), **β** là độ dốc (slope).

Hai tham số α, β được ước tính từ bảng dữ liệu bằng phương pháp bình phương bé nhất (least squares):

$$\boldsymbol{Min} \sum_{i=1}^{n} [y_i - (\alpha + \beta x_i)]^2$$

Hồi quy tuyến tính trong R

Trong ví dụ ở đây chiều cao height chính là y và độ tuổi age là x. Để thực hiện phân tích hồi quy tuyến tính cho vấn đề này, R cung cấp hàm lm cho phép tìm nhanh giá trị tham số α, β như sau:

```
res=lm(height ~ age)
```

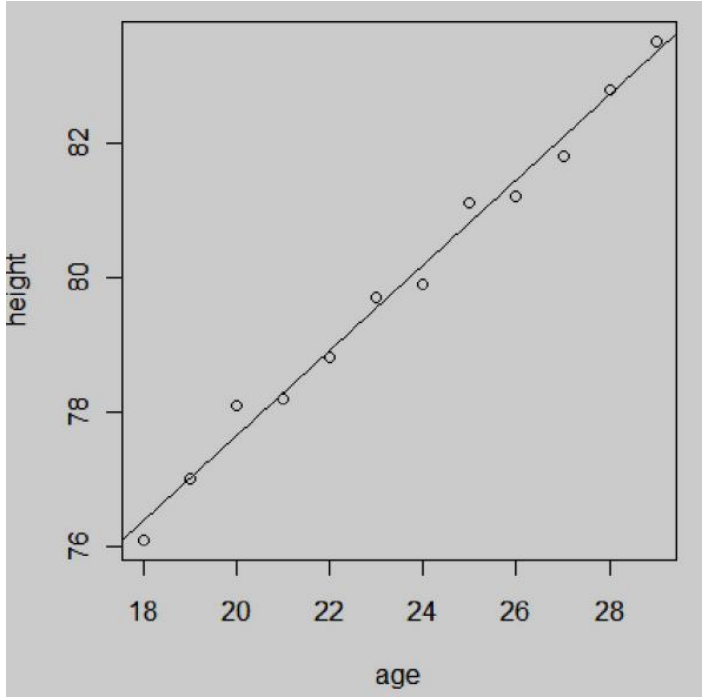
Kết quả trả về ta thu được giá trị chặn (intercept) α = 64.928 và độ dốc β = 0.635, phương trình hồi quy là : height = 0.635 age + 64.928.
Chúng ta có thể dùng hàm abline() để vẽ lên đồ thị phương trình hồi quy tuyến tính

```
abline(res)
```

Hồi quy tuyến tính trong R

Khi muốn dự đoán chiều cao của trẻ có tuổi là 23.5 tháng, chúng ta sử dụng phương trình hồi quy tìm được (height = 0.635 age + 64.928) để tính chiều cao là 79.85 cm. Việc này được thực hiện trong R bằng lệnh sau đây:

```
new <- data.frame(age = 23.5)
pred <- predict(res, new)
```



LẬP TRÌNH TRONG MÔI TRƯỜNG R

Một tính năng rất quan trọng trong môi trường R là việc hỗ trợ lập trình. Người sử dụng có thể định nghĩa các hàm để giải quyết các yêu cầu đặt ra. Cú pháp cho định nghĩa một hàm trong R như sau:

```
tên_hàm <- function(danh_sách_tham_số)  {
        thân_hàm
}
```

LẬP TRÌNH TRONG MÔI TRƯỜNG R

Ví dụ theo sau minh họa cách viết hàm tên desc() nhận tham số đầu vào x là dãy số trả về kết quả bao gồm giá trị trung bình và độ lệch chuẩn của dãy số.

```
desc <- function(x) {
        mu <- mean(x)
        s <- sd(x)
        c(MEAN = mu, SD = s)
}
```

LẬP TRÌNH TRONG MÔI TRƯỜNG R

Để sử dụng hàm desc(), ta sinh dãy gồm 10 số nguyên ngẫu nhiên có giá trị từ 1 đến 100, sau đó truyền cho hàm desc(), kết quả lưu vào trong biến d, phần tử d[1] chứa giá trị trung bình, d[2] là độ lệch chuẩn.

```
x <- sample(1:100, 10, replace = FALSE)
d <- desc(x)
```

CÂU ĐIỀU KIỆN IF TRONG R

Cú pháp :

```
if (điều_kiện)  {
        công_việc_1
} else {
        công_việc_2
}
```

Ví dụ câu điều kiện IF

Ví dụ :

```
a <- 33
b <- 33

if (b > a) {
  print("b is greater than a")
} else if (a == b) {
  print ("a and b are equal")
}
```

VÒNG LẶP FOR TRONG R

Cú pháp :

```
for (biến       in      khoảng_giá_trị) {
        công_việc
}
```

Ví dụ câu điều kiện FOR

Ví dụ :

```
fruits <- list("apple", "banana", "cherry")

for (x in fruits) {
  print(x)
}
```

VÒNG LẶP WHILE VÀ REPEAT TRONG R

Cú pháp :

```
while (điều_kiện)  {
        công_việc
}

repeat {
        công_việc
}
```

Ví dụ WHILE VÀ REPEAT

Ví dụ:

```
i <- 1
while (i < 6) {
  print(i)
  i <- i + 1
  if (i == 4) {
    break
  }
}
```

BÀI TẬP

Bài 1: Hãy quan sát các kết quả thực thi.
Chiều cao trung bình hiện nay ở nam thanh niên là 160 cm, với độ lệch chuẩn là 4.6 cm. Giả sử rằng chiều cao này tuân theo luật phân phối chuẩn. Hãy xây dựng một hàm phân phối chiều cao cho toàn bộ quần thể nam thanh niên, vẽ lên đồ thị.

BÀI TẬP

Bài 2: Hãy quan sát các kết quả thực thi.
Qua theo dõi nhiều tháng về tình hình người đến gọi điện thoại tại các buồng điện thoại công cộng thì biết được, tính trung bình cứ khoảng 2000 người đi qua thì có 2 người ghé qua gọi điện thoại. Hãy cho biết xác suất để có hơn 2 người ghé qua gọi điện thoại công cộng là bao nhiêu.

BÀI TẬP

Bài 3: Bằng R, Hãy thực hiện các công việc sau:

- Tạo dãy số nguyên từ 1 đến 100 lưu vào biến d.
- Lấy mẫu ngẫu nhiên 100 phần tử có hoàn lại từ tập d lưu vào biến b.
- Tính giá trị trung bình, phương sai của tập b.
- Lấy ra các phần tử có trong tập d nhưng không nằm trong tập b.
- Có bao nhiêu phần tử khác nhau nằm trong tập b.
- Vẽ đồ thị hộp và tổ chức đồ của tập b.

BÀI TẬP

Bài 4: Giả sử nghiên cứu sự ảnh hưởng của các môn học toán, lý, hóa đến kết quả môn giải thuật được cho trong bảng như sau:

STT	Toán	Lý	Hóa	Giải thuật
1	8	6	5	7
2	5	5	5	5
3	7	6	5	6
4	6	5	5	5
5	9	5	6	9
6	10	6	5	9
7	5	7	6	4
8	6	8	5	5
9	7	6	6	6
10	7	5	5	7

BÀI TẬP

11	7	6	6	6
12	6	7	5	4
13	6	8	6	5
14	5	7	5	4
15	5	5	6	4
16	10	6	5	10
17	9	8	5	8
18	9	7	5	9
19	9	6	5	9
20	9	8	5	10

BÀI TẬP

Bằng R, hãy hiển thị bảng dữ liệu trên với phương pháp Scatterplot 2 chiều và trục tọa độ song song. Tiếp đến là xây dựng mô hình hồi quy tuyến tính và dự báo điểm giải thuật của các sinh viên sau đây:

STT	Toán	Lý	Hóa	Giải thuật
21	5	5	6	?
22	7	6	5	?
23	6	6	8	?
24	10	5	5	?
25	9	5	9	?

Q&A


FPT Software
THANK YOU