



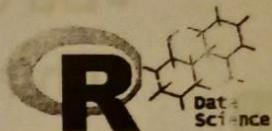
R for Data Science

Bài 12: *Visualization – ggplot2*

Phòng LT & Mạng

<https://csc.edu.vn/lap-trinh-va-cSDL/R-Programming-Language-for-Data-Science-190>

2020



Nội dung

1. Giới thiệu

2. Thành phần trên biểu đồ

3. Biểu đồ



Giới thiệu

□ Ggplot2

- Là một data visualization package dành cho ngôn ngữ lập trình thống kê R.
- Do Hadley Wickham tạo ra vào năm 2005.
- Có thể thay thế cho đồ họa cơ sở trong R và chứa một số giá trị mặc định cho web và in hiển thị các tỷ lệ phổ biến.
- Đã được sử dụng và trở thành một trong những gói R phổ biến nhất từ năm 2005.



R programming language for Data Science

3

Giới thiệu



ggplot2

Original author(s)	Hadley Wickham, Winston Chang
Initial release	2007-06-10
Stable release	3.0.0 / 2018-07-03
Repository	github.com/tidyverse/ggplot2
Written in	R
License	GPLv2
Website	ggplot2.tidyverse.org



R programming language for Data Science

4

□ Ưu điểm

- Hiển thị dữ liệu, gồm cả dữ liệu phức tạp
- Biểu đồ chất lượng cao
- Dễ dàng thiết kế



Nội dung

1. Giới thiệu
2. Thành phần trên biểu đồ
3. Biểu đồ



Thành phần trên biểu đồ

- **Dữ liệu, thuộc tính**
- **Loại biểu đồ**
- **Hình thức (trục tọa độ, nhãn, tiêu đề, theme)**

Cài đặt: `install.packages("ggplot2")`

 R programming language for Data Science

7

Thành phần trên biểu đồ



- **Dữ liệu, thuộc tính**
 - Xác định dữ liệu
 - Xác định biến thuộc tính x,y
 - Xác định biến thuộc tính nhóm

 R programming language for Data Science

8

Thành phần trên biểu đồ



- Ví dụ: Biểu đồ thể hiện mối quan hệ giữa hp và mpg, fill theo carb

head(mtcars)

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.460	20.22	1	0	3	1

plot <- ggplot(data=mtcars, aes(x = hp, y = mpg, fill = carb, color = carb, lwd=1))

group

tên thuộc carb

R programming language for Data Science

9

Thành phần trên biểu đồ



□ Loại biểu đồ

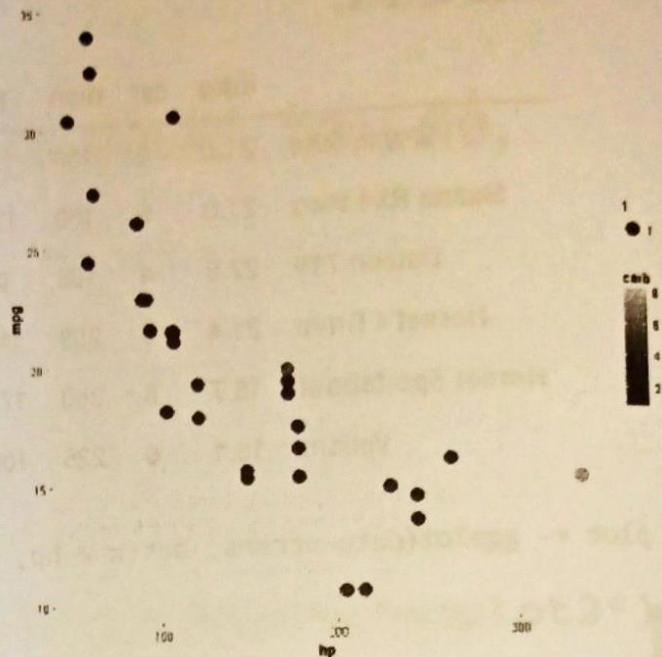
- Geom_histogram (histogram)
- Geom_bar (bar plot)
- Geom_boxplot (box plot)
- Geom_point (scatter plot)
- ...

Thành phần trên biểu đồ

- Ví dụ: Biểu đồ thể hiện mối quan hệ giữa hp và mpg, fill theo carb => geom_point()

```
plot <- plot + geom_point()
```

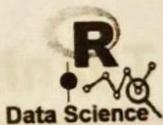
plot



R programming language for Data Science

11

Thành phần trên biểu đồ



□ Hình thức

● Trục tọa độ

- Nếu dữ liệu trên trục tung (y), hoặc trục hoành (x) có kiểu **phân loại**:
 - scale_x_discrete(name="xxx", limits=c("A", "B", "C"))
 - scale_y_discrete(name="xxx", limits=c("X", "Y", "Z"))
- Nếu dữ liệu trên trục tung (y), hoặc trục hoành (x) có kiểu **liên tục**:
 - scale_x_continuous(name="xxx", limits=c(lower, upper), breaks=seq(from, to, by))
 - scale_y_continuous(name="xxx", limits=c(lower, upper), breaks=c(value1, value2, value3, ...))

R programming language for Data Science

12

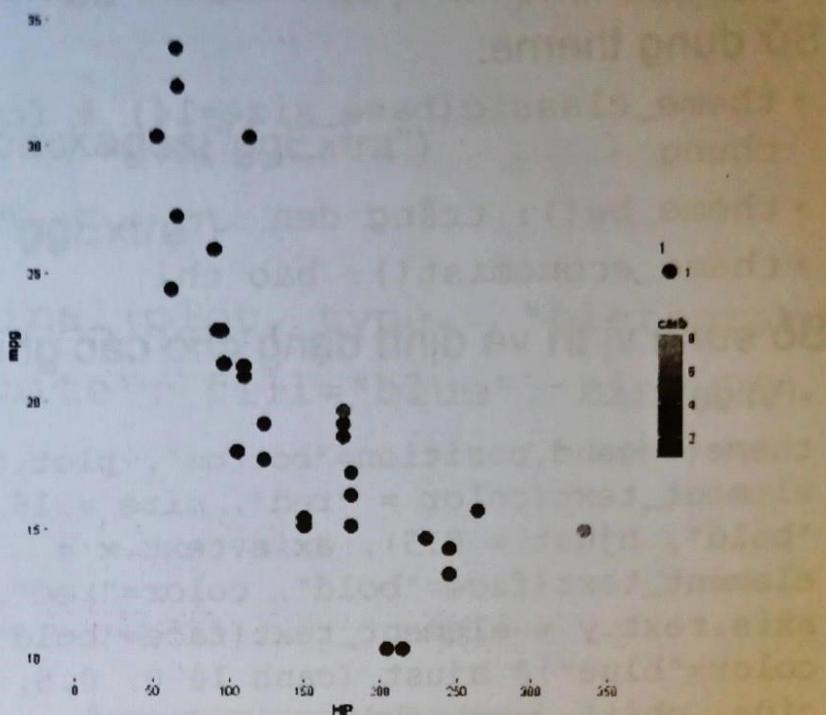
↳ mứn chia bao nhiêu quãng cách nhau bao nhiêu



Thành phần trên biểu đồ

• Ví dụ

```
plot <- plot + scale_x_continuous("HP", breaks = seq(0, 350, 50), limits = c(0,350))
```



13

Thành phần trên biểu đồ

• Nhãn, tiêu đề

- Nhãn trên trục tung và trục hoành

```
xlab("xxx") + ylab("xxx")
```

- Tiêu đề

```
ggttitle("xxxx")
```



Thành phần trên biểu đồ

● Theme

- Cài đặt theme: `install.packages("ggthemes")`

- Sử dụng theme:

- `theme_classic(base_size=14) # font size chung`
- `theme_bw():` trắng đen
- `theme_economist():` báo chí

- Bổ sung vị trí và định dạng cho các ghi chú

- Ví dụ:

```
theme(legend.position="bottom", plot.title = element_text(color = "red", size = 16, face = "bold", hjust = 0.5), axis.text.x = element_text(face="bold", color="red", angle=45), axis.text.y = element_text(face="bold", color="blue")# hjust (canh lề 0, 0.5, 1: trái, giữa, phải)
```

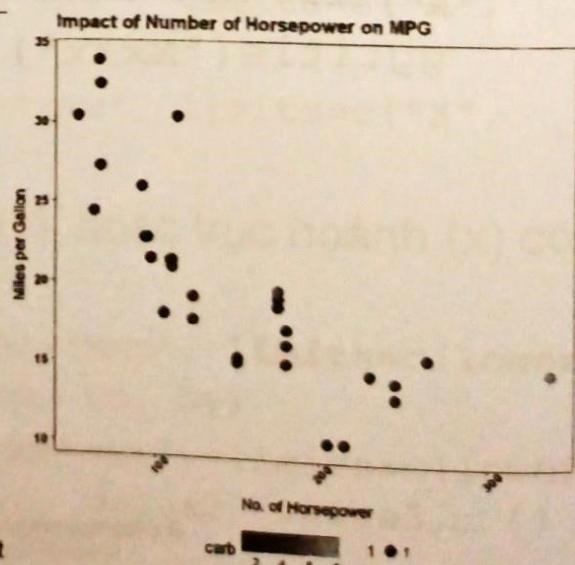
*text xoay
bao nhiêu*



Thành phần trên biểu đồ

```
plot <- ggplot(mtcars, aes(x = hp, y = mpg, fill = carb, color = carb, lwd=1)) +
  geom_point() +
  ylab("Miles per Gallon") +
  xlab ("No. of Horsepower") +
  ggtitle("Impact of Number of Horsepower on MPG") +
  theme_bw(base_size = 14) +
  theme(legend.position="bottom",
        plot.title = element_text(color = "red", size = 16, face = "bold", hjust = 0.5),
        axis.text.x = element_text(face="bold", color="red", angle=45),
        axis.text.y = element_text(face="bold", color="blue"))
```

```
plot
```



Thành phần trên biểu đồ



• Marginal histogram

- Có thể thêm vào để biểu phân phối của các thuộc tính

- `install.packages("ggExtra")`

- `library("ggExtra")`

```
ggMarginal(plot, type = "histogram",  
           col="white", fill="blue", bins=20)
```

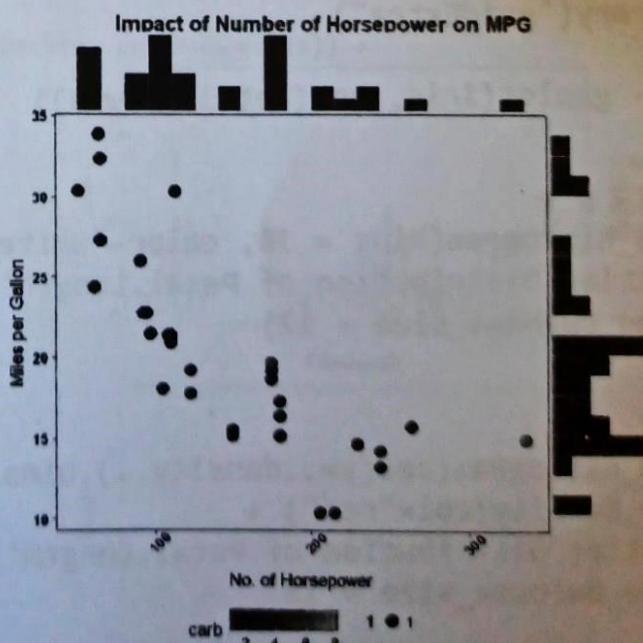


Thành phần trên biểu đồ



```
library("ggExtra")
```

```
ggMarginal(plot, type = "histogram", col="white", fill="blue", bins=20)
```



Nội dung

1. Giới thiệu
2. Thành phần trên biểu đồ
3. Biểu đồ



Biểu đồ

❑ Histogram - Geom_histogram()

```
library("ggplot2")
library("gridExtra")

p <- ggplot(iris, aes(Petal.Length))

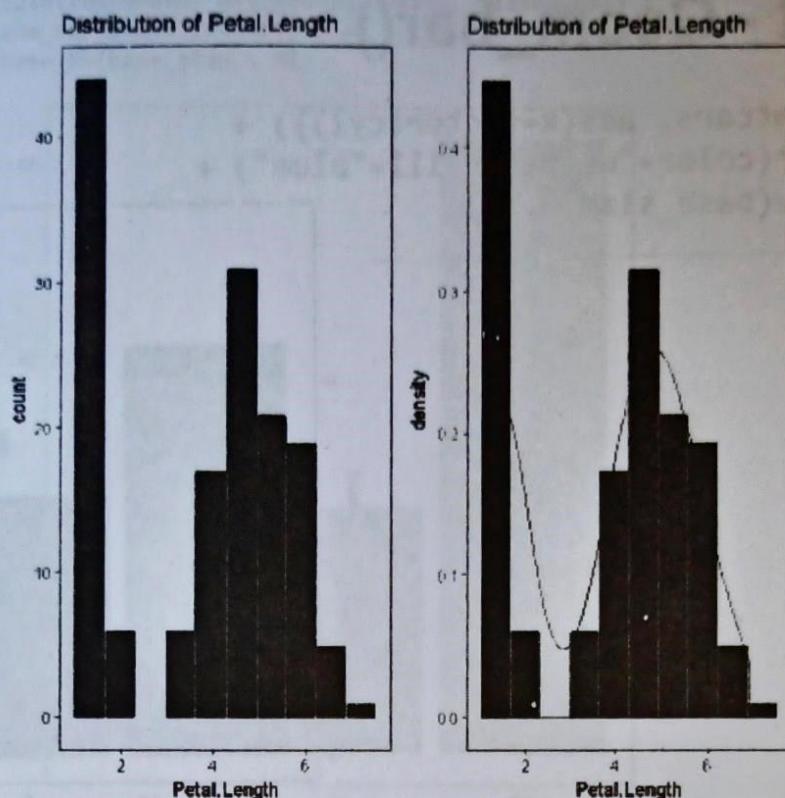
p1 <- p +
  geom_histogram(bins = 10, color="white", fill="blue") +
  ggtitle('Distribution of Petal.Length') +
  theme_bw(base_size = 12)

p2 <- p +
  geom_histogram(aes(y=..density..),bins = 10, color="white", fill="blue") +
  geom_density(col="red") +
  ggtitle('Distribution of Petal.Length') +
  theme_bw(base_size = 12)
```



Biểu đồ

```
grid.arrange(p1, p2, ncol=2)
```



R programming language for Data Science

21

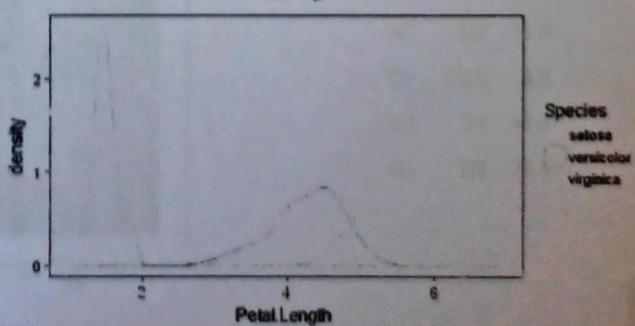
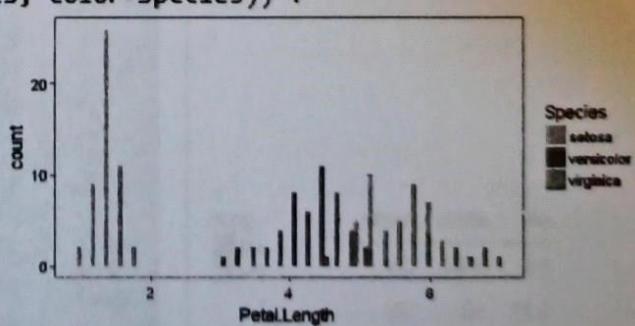
Biểu đồ

```
p <- ggplot(iris, aes(x=Petal.Length, fill=Species))
```

p1 <- p +
geom_histogram(position = 'dodge', bins = 30) +
theme_bw(base_size = 14)

```
p2 <- ggplot(iris, aes(x=Petal.Length, fill=Species, color=Species)) +  
geom_density(alpha = 0.1) +  
theme_bw(base_size = 14)
```

```
grid.arrange(p1, p2, nrow = 2)
```



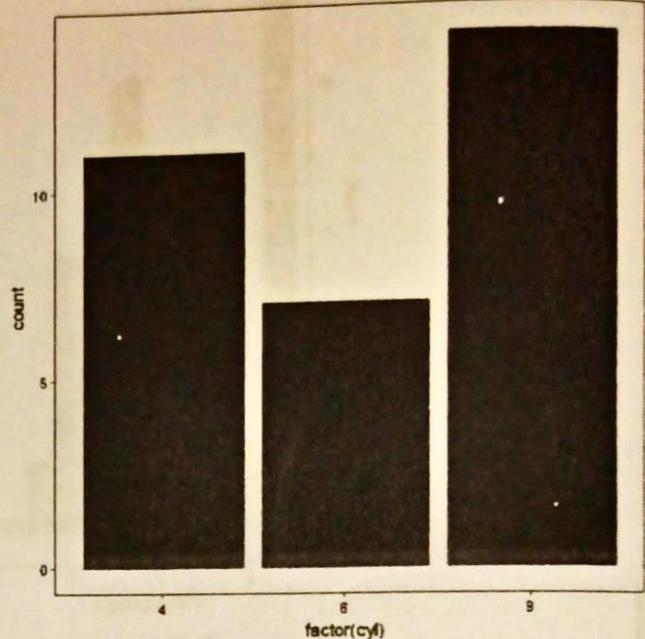
R programming language for

Biểu đồ

□ Barplot - Geom_bar()

```
cp <- ggplot(mtcars, aes(x=factor(cyl))) +
  geom_bar(color="white", fill="blue") +
  theme_bw(base_size = 14)
```

cp



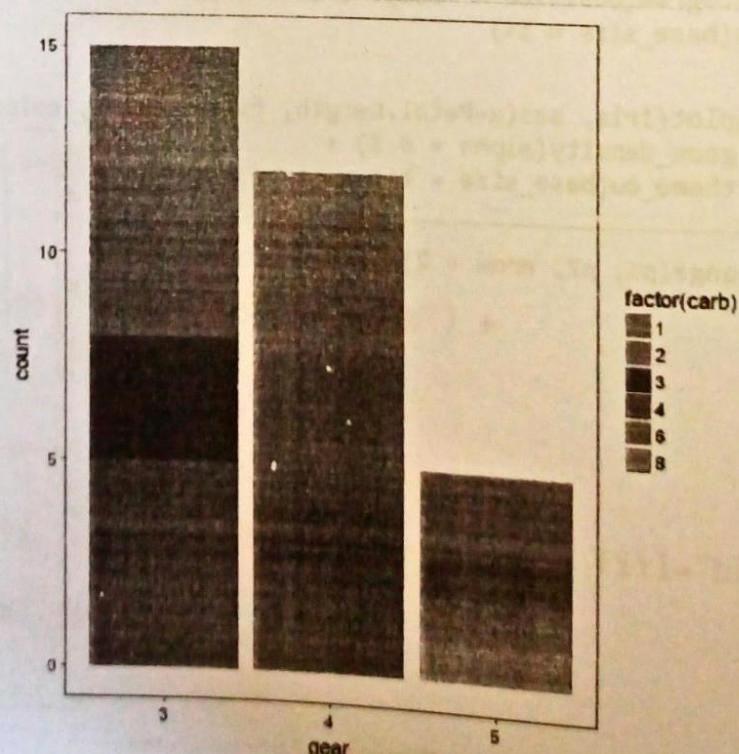
R programming language for Data Science

23

Biểu đồ

● Phân nhóm

```
ggplot(data=mtcars, aes(x=gear, fill=factor(carb))) +
  geom_bar() +
  theme_bw(base_size = 14)
```

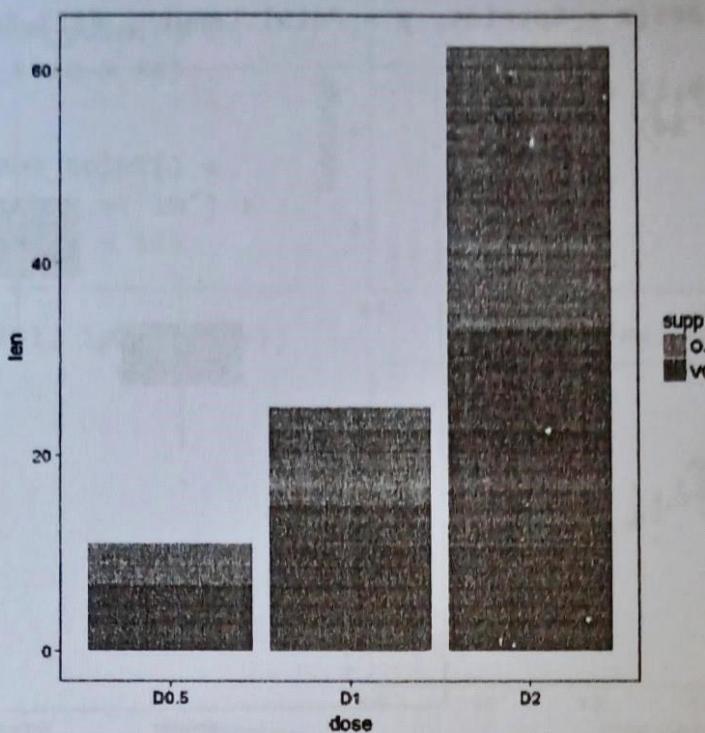


R programming language for Data Science

24

Biểu đồ

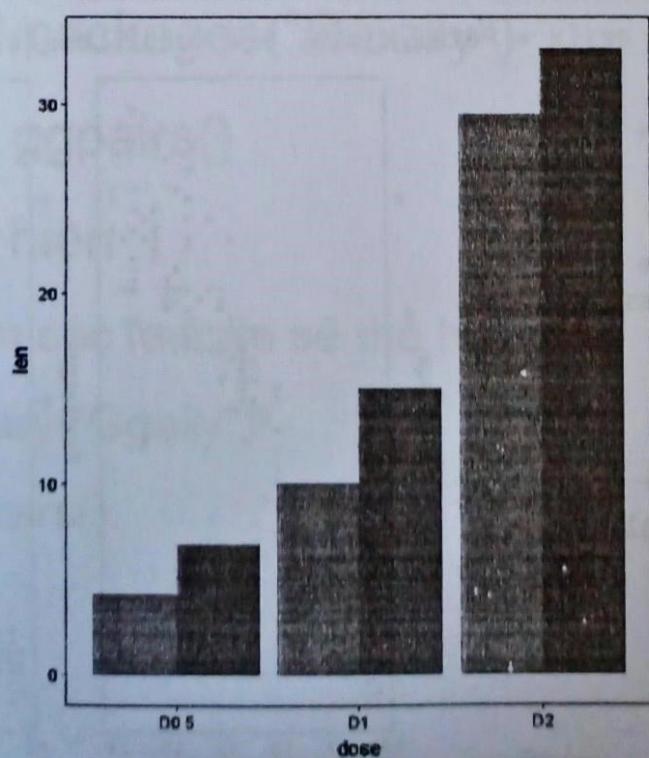
```
# Stacked barplot with multiple groups
ggplot(data=df2, aes(x=dose, y=len, fill=supp)) +
  geom_bar(stat="identity") +
  theme_bw(base_size = 14)
```



25

Biểu đồ

```
# Use position='dodge'
ggplot(data=df2, aes(x=dose, y=len, fill=supp)) +
  geom_bar(stat="identity", position='dodge')+
  theme_bw(base_size = 14)
```



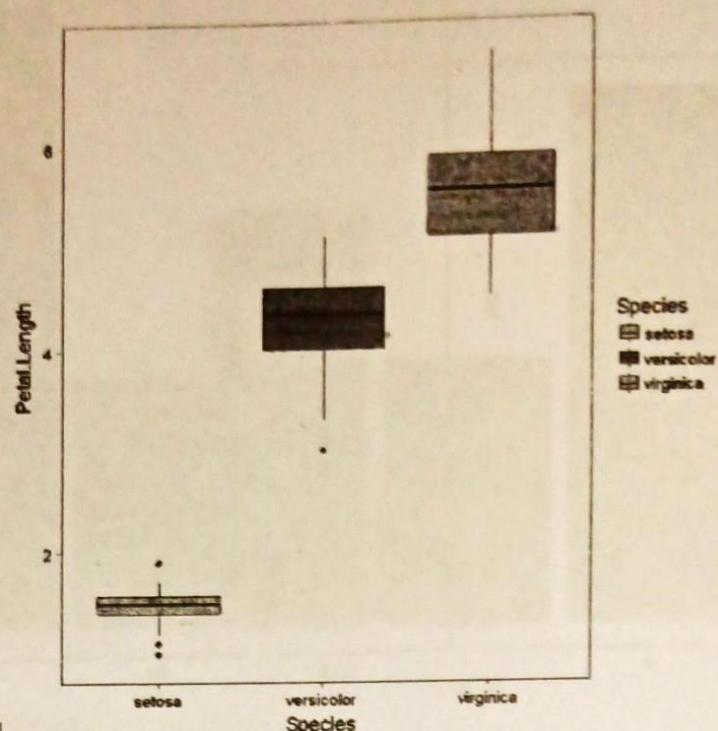
26

Biểu đồ

Boxplot: geom_boxplot()

```
ggplot(data = iris, aes(x = Species, y = Petal.Length, fill=Species)) +
  geom_boxplot() +
  geom_jitter(alpha = 0.1) +
  theme_bw(base_size = 14)
```

hiển thị các điểm
dù lây



Biểu đồ

Scatter plot: geom_point()

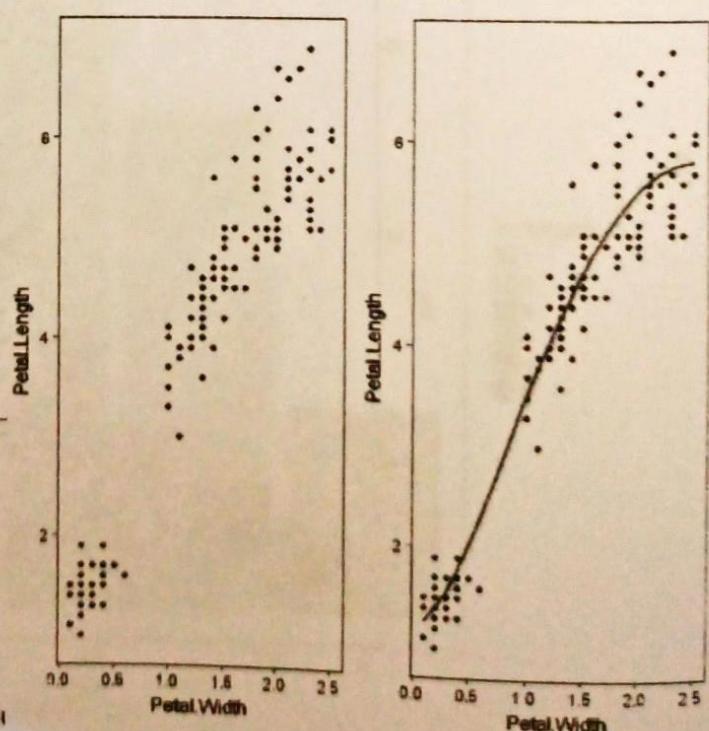
```
ip <- ggplot(data = iris, aes(x = Petal.Width, y = Petal.Length))

ip1 <- ip + geom_point() +
  theme_bw(base_size = 14)

ip2 <- ip + geom_point() +
  geom_smooth(method = 'loess') +
  theme_bw(base_size = 14)

library("gridExtra")

grid.arrange(ip1, ip2, ncol=2)
```



Biểu đồ

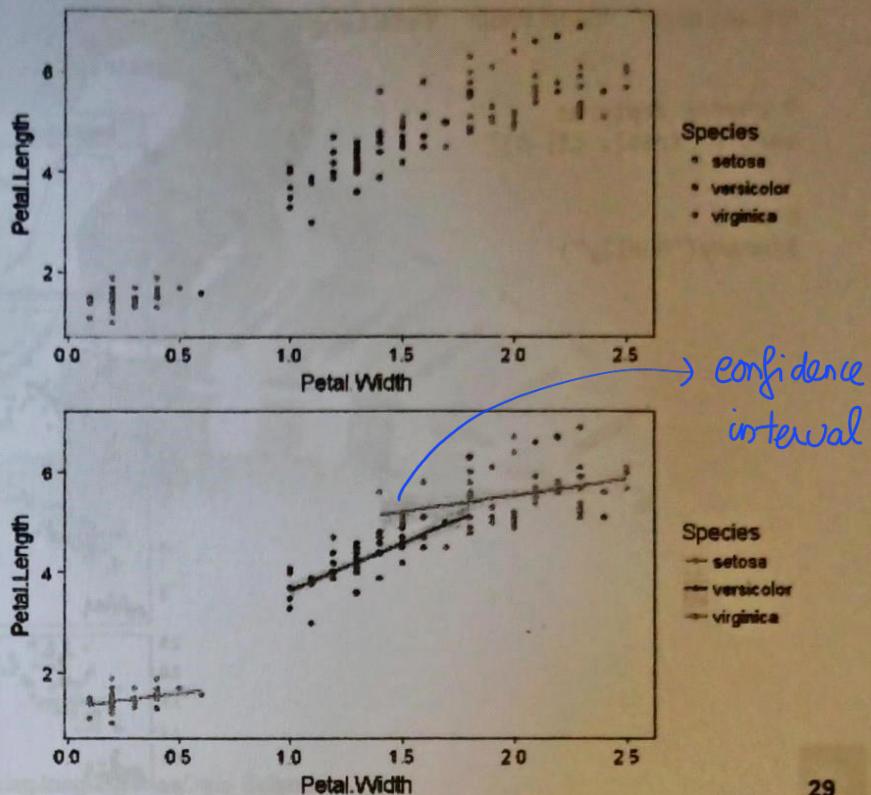


```
ip <- ggplot(data = iris, aes(x = Petal.Width, y = Petal.Length,  
                           col = Species, fill = Species))
```

```
ip1 <- ip + geom_point() +  
       theme_bw(base_size = 14)
```

```
ip2 <- ip + geom_point() +  
       geom_smooth(method = 'lm') +  
       theme_bw(base_size = 14)
```

```
grid.arrange(ip1, ip2, nrow=2)
```



R programm

29

Biểu đồ



☐ Tương quan đa biến

- `install.packages("GGally")`

- Dùng `ggpairs()`

- Thực hiện

- Chọn các feature sẽ thể hiện

- Library("Ggally")

- `Ggpairs()`



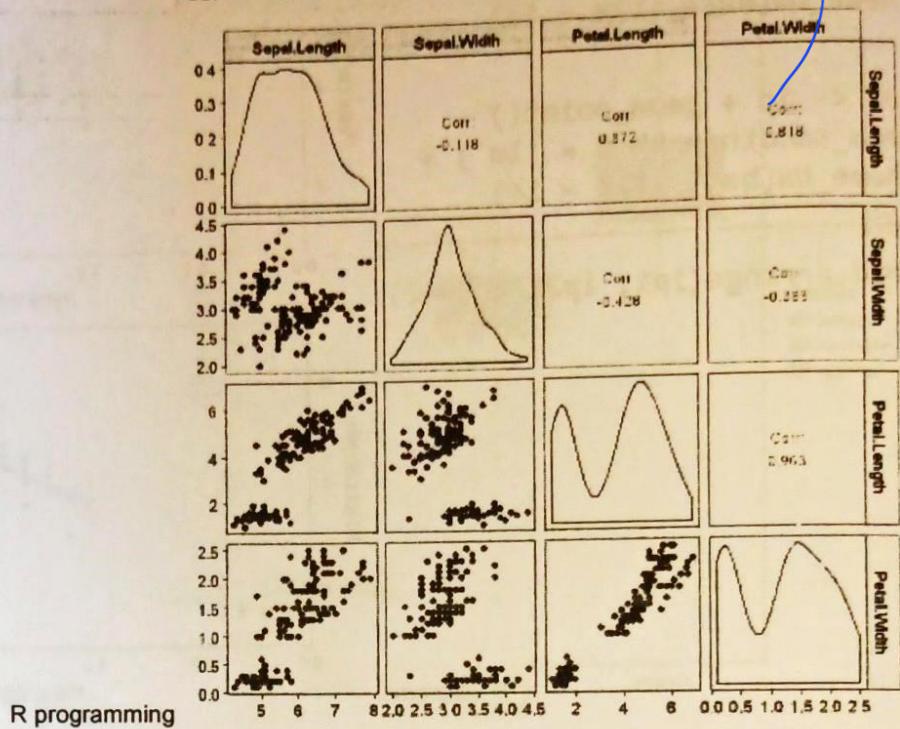
Biểu đồ

```
ci <- colnames(iris)
ci[-5]

'Sepal.Length' 'Sepal.Width' 'Petal.Length' 'Petal.Width'

# choose features
vars <- iris[, ci[-5]]

# plot
library("GGally")
```



31

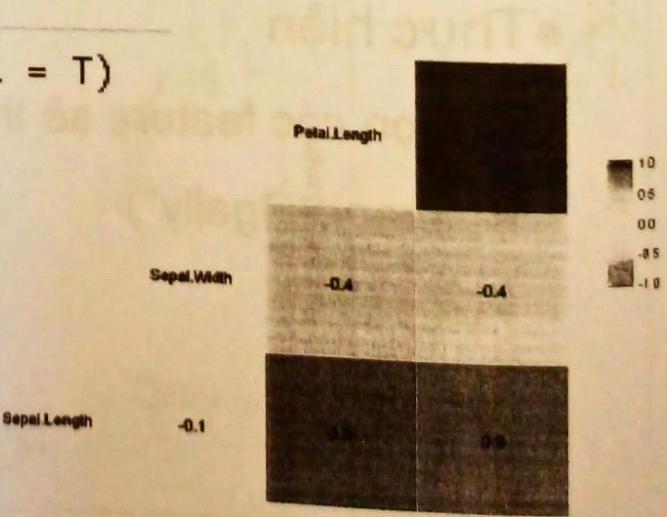
Biểu đồ

● Dùng ggcov()

```
iris_sub <- iris[, -c(5)]
```

```
library("GGally")
```

```
ggcorr(data = iris_sub, label = T)
```



Chapter 12: ggplot2

Exercise 1: Chol - ggplot2

- Cho dữ liệu chol.txt.
- Đọc và hiển thị head của dữ liệu
- In thống kê chung về dữ liệu
- In thông tin của dữ liệu
- Vẽ histogram cho cột AGE của dữ liệu
- Vẽ scatter plot biểu diễn mối quan hệ của HEIGHT vs WEIGHT
- Vẽ scatter plot biểu diễn mối quan hệ của HEIGHT vs WEIGHT, có kèm theo Histogram / Boxplot
- Vẽ piechart biểu diễn BLOOD

Exercise 2: Housing Prices

- Cho dữ liệu landdata_states.csv.
- Đọc và hiển thị head của dữ liệu
- In thống kê chung về dữ liệu
- In thông tin của dữ liệu
- Vẽ histogram cho cột Home.Value của dữ liệu
- Hãy lọc dữ liệu theo Date == 2001.25, sau đó vẽ scatter plot biểu diễn Land.Value vs Structure.Cost.
- Vẽ lại biểu đồ trên với Land.Value được chuẩn hóa bằng log. Gắn thêm state cho từng điểm dữ liệu.

Exercise 3: EconomistData

- Cho dữ liệu EconomistData.csv.
- Đọc và hiển thị head của dữ liệu
- In thống kê chung về dữ liệu
- In thông tin của dữ liệu
- Vẽ scatter plot biểu diễn mối quan hệ của CPI vs HDI, điểm được tô màu theo Region, độ lớn của điểm theo HDI.Rank
- Vẽ scatter plot biểu diễn mối quan hệ của CPI vs HDI có regression line



Exercise 1: Chol - ggplot2

```
In [1]: library(ggplot2)
```

```
In [2]: # Load in `chol` data
chol <- read.table("chol.txt", header = TRUE)
```

```
# Inspect first rows of `chol` with `head()`
head(chol)
```

```
# Summary with `summary()`
summary(chol)
```

```
# Structure of `chol` with `str()`
str(chol)
```

AGE	HEIGHT	WEIGHT	CHOL	SMOKE	BLOOD	MORT
20	176	77	195	nonsmo	b	alive
53	167	56	250	sigare	o	dead
44	170	80	304	sigare	a	dead
37	173	89	178	nonsmo	o	alive
26	170	71	206	sigare	o	alive
41	165	62	284	sigare	o	alive

AGE	HEIGHT	WEIGHT	CHOL	SMOKE
Min. :18.00	Min. :156.0	Min. : 53.00	Min. :107.0	nonsmo: 49
1st Qu.:28.75	1st Qu.:168.0	1st Qu.: 68.75	1st Qu.:204.0	pipe : 42
Median :37.00	Median :172.0	Median : 75.00	Median :232.0	sigare:109
Mean :35.72	Mean :172.3	Mean : 75.89	Mean :233.6	
3rd Qu.:42.00	3rd Qu.:176.0	3rd Qu.: 82.00	3rd Qu.:259.0	
Max. :58.00	Max. :191.0	Max. :110.00	Max. :455.0	

BLOOD	MORT
a :82	alive:176
ab: 5	dead : 24
b :22	
o :91	

```
'data.frame': 200 obs. of 7 variables:
```

```
$ AGE : int 20 53 44 37 26 41 39 28 33 39 ...
```

```
$ HEIGHT: int 176 167 170 173 170 165 174 171 180 166 ...
```

```
$ WEIGHT: int 77 56 80 89 71 62 75 68 100 74 ...
```

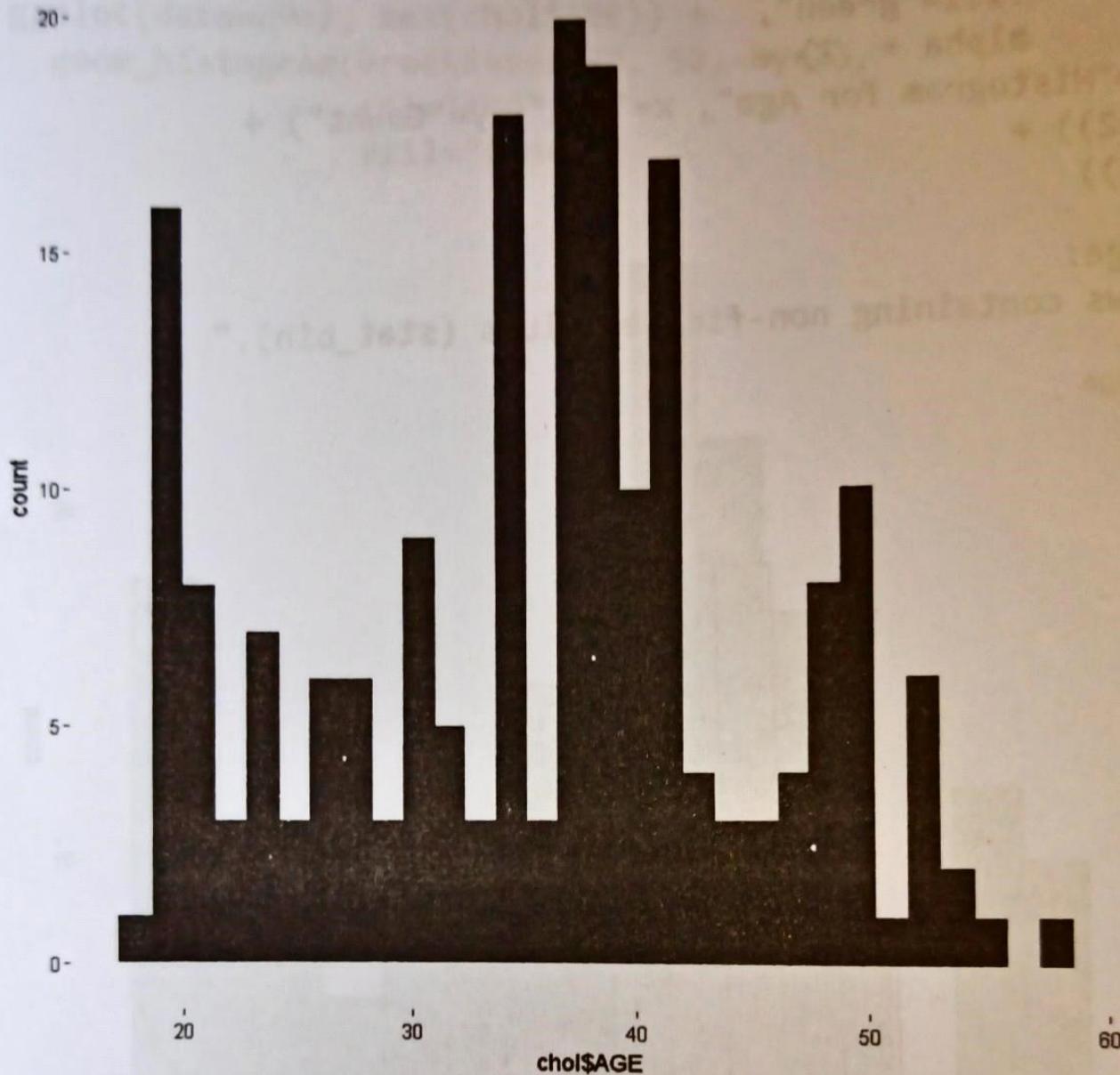
```
$ CHOL : int 195 250 304 178 206 284 232 152 209 150 ...
```

```
$ SMOKE : Factor w/ 3 levels "nonsmo","pipe",...: 1 3 3 1 3 3 3 2 3 3 ...
```

```
$ BLOOD : Factor w/ 4 levels "a","ab","b","o": 3 4 1 4 4 4 4 1 1 1 ...
```

```
$ MORT : Factor w/ 2 levels "alive","dead": 1 2 2 1 1 1 1 1 1 1 ...
```

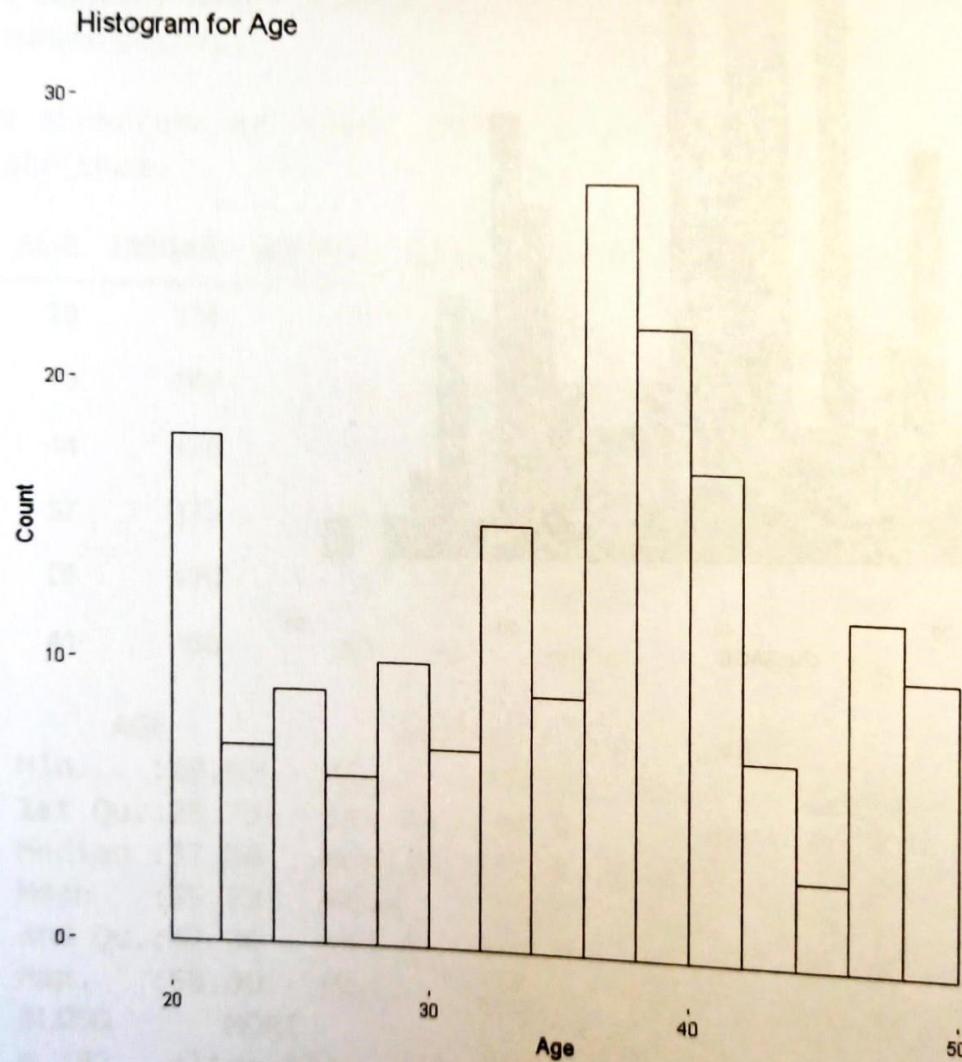
```
In [8]: ggplot(data=chol, aes(chol$AGE)) +  
    geom_histogram(bins=30)
```



```
In [1]: ggplot(data=chol, aes(x=chol$AGE)) +  
  geom_histogram(breaks=seq(20, 50, by=2),  
                 col="red",  
                 fill="green",  
                 alpha = .2) +  
  labs(title="Histogram for Age", x="Age", y="Count") +  
  xlim(c(18,52)) +  
  ylim(c(0,30))
```

Warning message:

"Removed 6 rows containing non-finite values (stat_bin)."



```
In [12]: ggplot(data=chol, aes(chol$AGE)) +  
  geom_histogram(breaks=seq(20, 50, by=2),  
                 col="red",  
                 fill="green")
```

