Thực hành thống kê: Tuần 4

Nội dung chính:

1. Hiệp phương sai
2. Hệ số tương quan
3. Các dạng đồ thị thông dụng khác: pie chart, bar chart, line chart, …

1. Hiệp phương sai

Đo lường mối quan hệ tuyến tính giữa hai biến.

$$\text{Cov}(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

- Nếu `cov(X,Y) > 0` : Hai biến có xu hướng tăng cùng nhau.
- Nếu `cov(X,Y) < 0` : Khi một biến tăng, biến kia có xu hướng giảm.
- Nếu `cov(X,Y) ≈ 0` : Không có mối quan hệ tuyến tính rõ ràng.

Viết hàm tính hiệp phương sai

```
my.cov <- function(X,Y){
  n = length(X)
  cov = 1/(n-1) * sum((X - mean(X))*(Y- mean(Y)))
  cov
}
# Dữ liệu mẫu
X <- 1:10
Y <- 2:11

# Tính hiệp phương sai
my.cov(X,Y)
```

```
## [1] 9.166667
```

- Tính hiệp phương sai trong R với `cov()`

```
cov(X,Y)
```

```
## [1] 9.166667
```

2.Hệ số tương quan

Hệ số tương quan Pearson đo lường mức độ mạnh/yếu của mối quan hệ tuyến tính giữa hai biến, giá trị nằm trong khoảng

$$r = \frac{\text{Cov}(X,Y)}{\sigma_X * \sigma_Y}$$

Viết hàm tính hệ số tương quan Pearson

```
my.cor <- function(X,Y){
  r = my.cov(X,Y) / (sd(X)*sd(Y))
  r
}
my.cor(X, Y)
```

```
## [1] 1
```

- Tính hệ số tương quan trong R với `cor()`

```
cor(X,Y)
```

```
## [1] 1
```

3. Các dạng đồ thị thông dụng

```r
#install.packages("titanic")  # Cài đặt package nếu chưa có
library(titanic)

data("titanic_train")  # Dữ liệu huấn luyện
head(titanic_train)
```

```
##   PassengerId Survived Pclass
## 1           1        0      3
## 2           2        1      1
## 3           3        1      3
## 4           4        1      1
## 5           5        0      3
## 6           6        0      3
##                                                 Name    Sex Age SibSp Parch
## 1                            Braund, Mr. Owen Harris   male  22     1     0
## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38     1     0
## 3                             Heikkinen, Miss. Laina female  26     0     0
## 4        Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35     1     0
## 5                           Allen, Mr. William Henry   male  35     0     0
## 6                                   Moran, Mr. James   male  NA     0     0
##             Ticket    Fare Cabin Embarked
## 1        A/5 21171  7.2500              S
## 2         PC 17599 71.2833   C85        C
## 3 STON/O2. 3101282  7.9250              S
## 4           113803 53.1000  C123        S
## 5           373450  8.0500              S
## 6           330877  8.4583              Q
```

```r
# Thống kê mô tả
summary(titanic_train)
```

```
##   PassengerId       Survived          Pclass          Name
##  Min.   :  1.0   Min.   :0.0000   Min.   :1.000   Length:891
##  1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000   Class :character
##  Median :446.0   Median :0.0000   Median :3.000   Mode  :character
##  Mean   :446.0   Mean   :0.3838   Mean   :2.309
##  3rd Qu.:668.5   3rd Qu.:1.0000   3rd Qu.:3.000
##  Max.   :891.0   Max.   :1.0000   Max.   :3.000
##
##      Sex                 Age            SibSp           Parch
##  Length:891         Min.   : 0.42   Min.   :0.000   Min.   :0.0000
##  Class :character   1st Qu.:20.12   1st Qu.:0.000   1st Qu.:0.0000
##  Mode  :character   Median :28.00   Median :0.000   Median :0.0000
##                     Mean   :29.70   Mean   :0.523   Mean   :0.3816
##                     3rd Qu.:38.00   3rd Qu.:1.000   3rd Qu.:0.0000
##                     Max.   :80.00   Max.   :8.000   Max.   :6.0000
##                     NA's   :177
##     Ticket               Fare           Cabin             Embarked
##  Length:891         Min.   :  0.00   Length:891         Length:891
##  Class :character   1st Qu.:  7.91   Class :character   Class :character
##  Mode  :character   Median : 14.45   Mode  :character   Mode  :character
##                     Mean   : 32.20
##                     3rd Qu.: 31.00
##                     Max.   :512.33
##
```

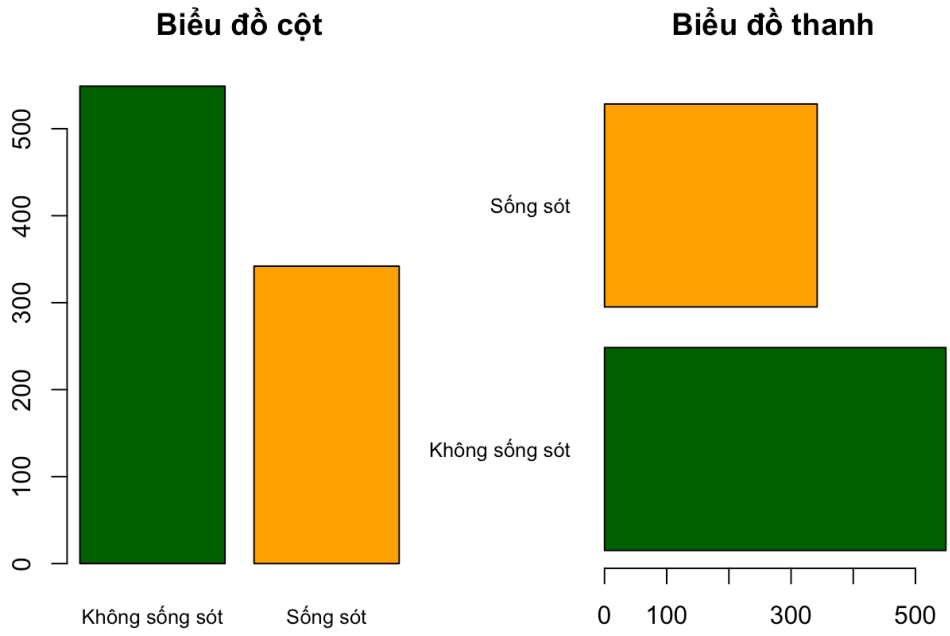
```r
# Xem cấu trúc dữ liệu
str(titanic_train)
```

```
## 'data.frame':    891 obs. of  12 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : int  0 1 1 1 0 0 0 0 1 1 ...
##  $ Pclass     : int  3 1 3 1 3 3 1 3 3 2 ...
##  $ Name       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)" "Heik
kinen, Miss. Laina" "Futrelle, Mrs. Jacques Heath (Lily May Peel)" ...
##  $ Sex        : chr  "male" "female" "female" "female" ...
##  $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
##  $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Ticket     : chr  "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : chr  "" "C85" "" "C123" ...
##  $ Embarked   : chr  "S" "C" "S" "S" ...
```

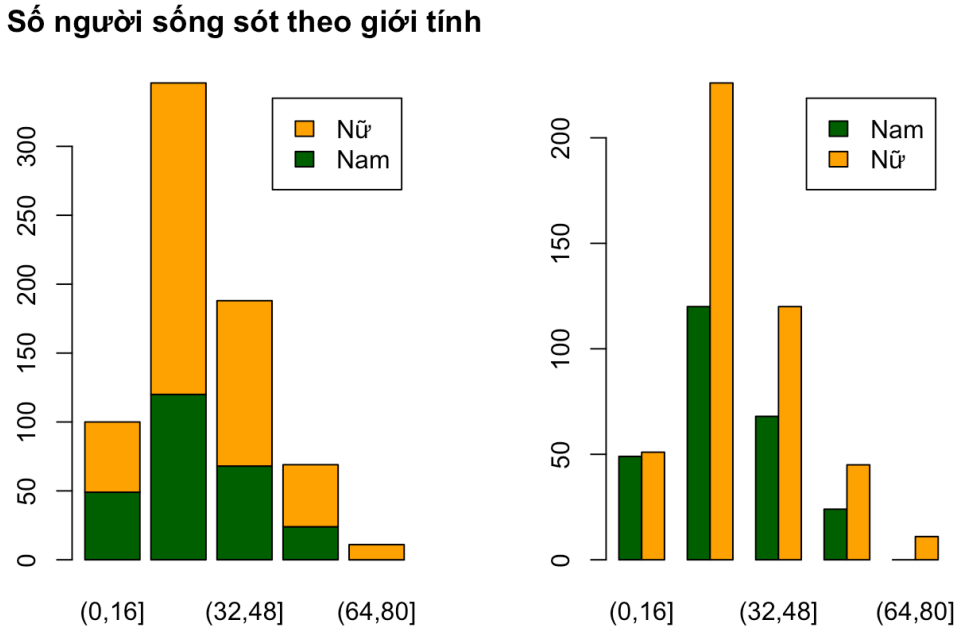
Biểu đồ cột: `barplot()`

```
# Thiết lập bố cục 1 hàng, 2 cột
par(mfrow = c(1, 2))
barplot(table(titanic_train$Survived),
        col = c('darkgreen','orange'),
        names.arg = c("Không sống sót", "Sống sót"),
        main = "Biểu đồ cột", cex.names=0.8)

barplot(table(titanic_train$Survived), col = c('darkgreen','orange'),
        names.arg = c("Không sống sót", "Sống sót"),
        las = 1, horiz = TRUE, main="Biểu đồ thanh", cex.names=0.8 )
```

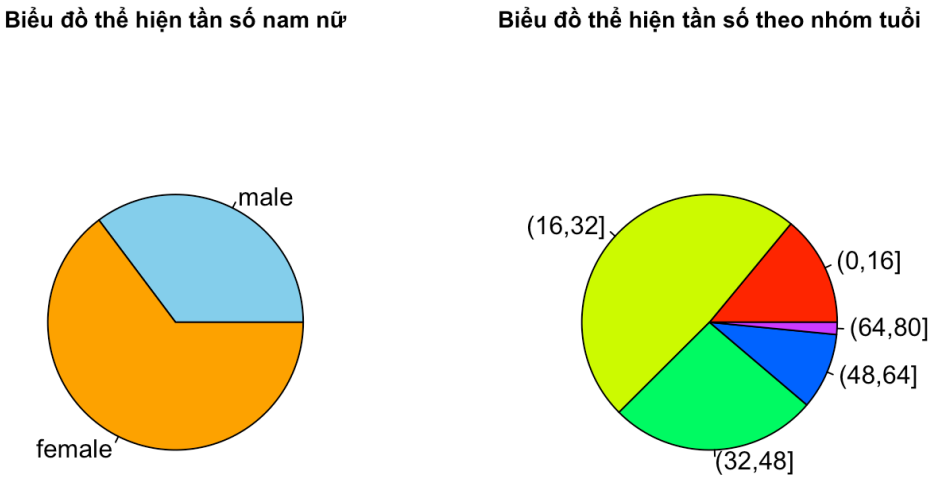



```
nhom.tuoi <- cut(titanic_train$Age, c(0,16,32,48,64,80))
tb_2 <- table(titanic_train$Sex, nhom.tuoi)
# Biểu đồ cột số người sống sót theo giới tính
par(mfrow = c(1, 2))
barplot(tb_2, col=c('darkgreen','orange'), main = "Số người sống sót theo giới tính", legend.text =c('Nam',
'Nữ'))
barplot(tb_2, col=c('darkgreen','orange'), legend.text =c('Nam', 'Nữ'), beside = TRUE)
```



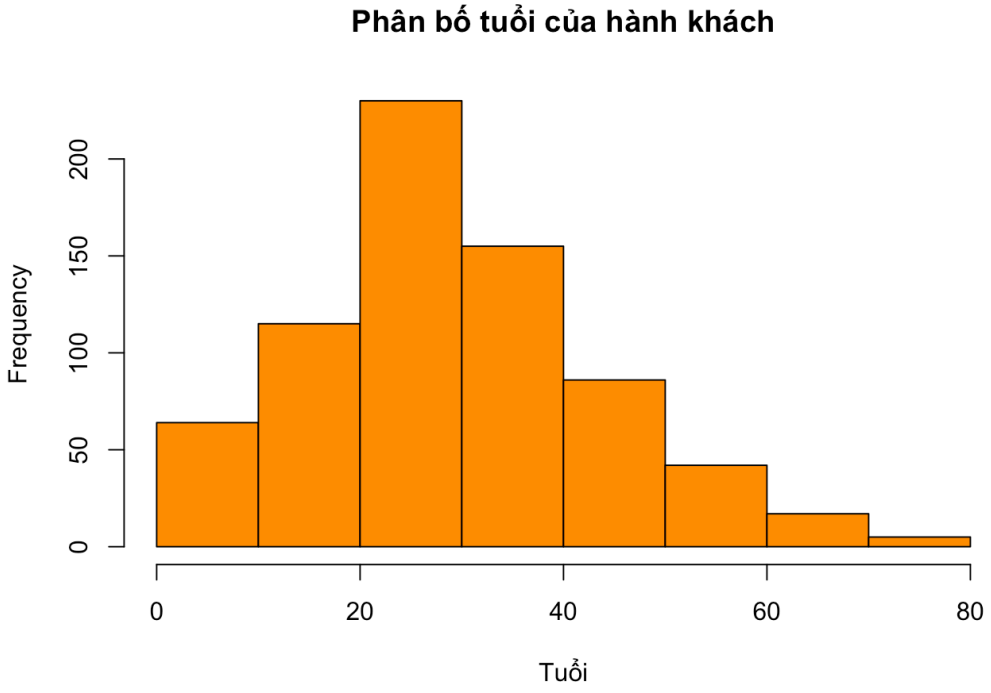
Biểu đồ tròn: `pie()`

```
# Vẽ biểu đồ tròn
par(mfrow = c(1, 2))
pie(table(titanic_train$Sex), labels = titanic_train$Sex, col=c('skyblue','orange'), main = "Biểu đồ thể hiện
tần số nam nữ",cex.main = 0.9)
pie(table(nhom.tuoi), col=rainbow(length(table(nhom.tuoi))), main = "Biểu đồ thể hiện tần số theo nhóm tuổi",
cex.main = 0.9)
```



Biểu đồ histogram: `hist()`

```
hist(titanic_train$Age,
     col = "darkorange",
     main = "Phân bố tuổi của hành khách",
     xlab = "Tuổi")
```



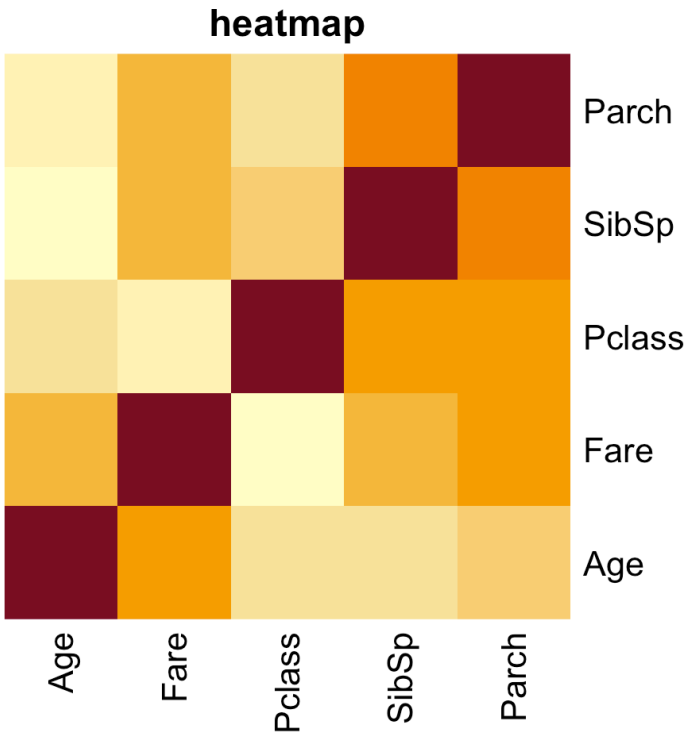
Phân tích tương quan

```
library(dplyr)
# Chọn các cột số
cor_df <- titanic_train %>% select(Age, Fare, Pclass, SibSp, Parch)
# hoặc cor_df <- titanic_train[, c("Age", "Fare", "Pclass", "SibSp", "Parch")]
# Tính ma trận tương quan
cor_matrix <- cor(cor_df, use = "complete.obs")
print(cor_matrix)
```

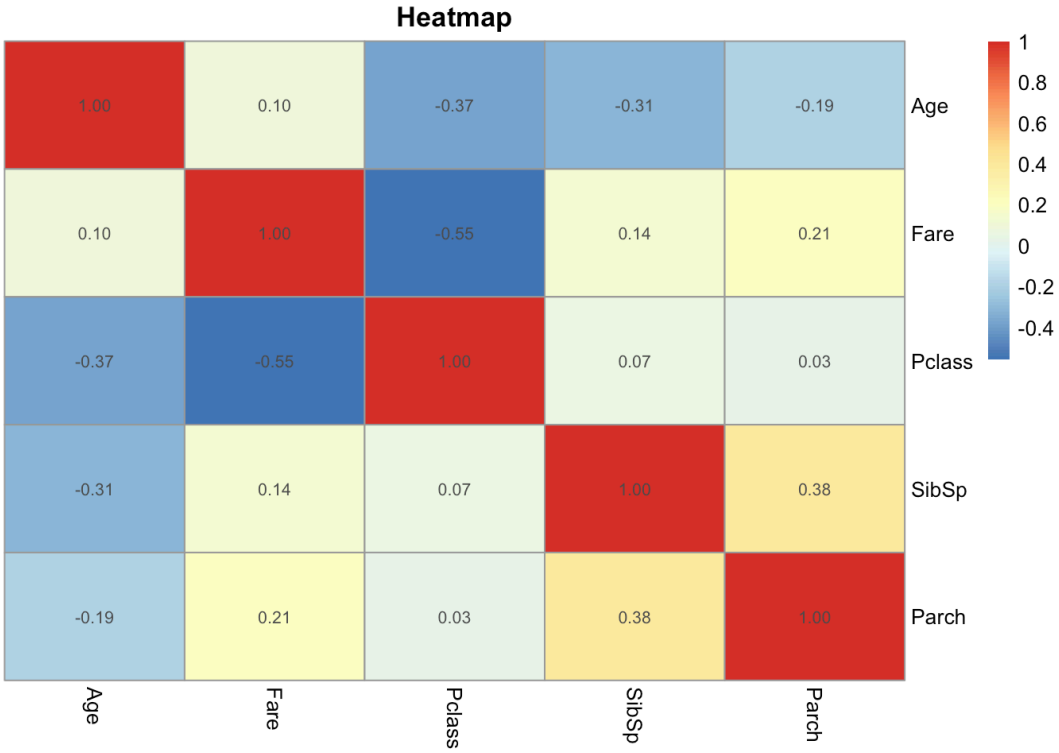
```
##                Age        Fare      Pclass       SibSp       Parch
## Age     1.00000000  0.09606669 -0.36922602 -0.30824676 -0.18911926
## Fare    0.09606669  1.00000000 -0.55418247  0.13832879  0.20511888
## Pclass -0.36922602 -0.55418247  1.00000000  0.06724737  0.02568307
## SibSp  -0.30824676  0.13832879  0.06724737  1.00000000  0.38381986
## Parch  -0.18911926  0.20511888  0.02568307  0.38381986  1.00000000
```

Biểu đồ nhiệt

```
heatmap(cor_matrix, margins = c(6,6), Colv = NA, Rowv=NA, main = "heatmap")
```

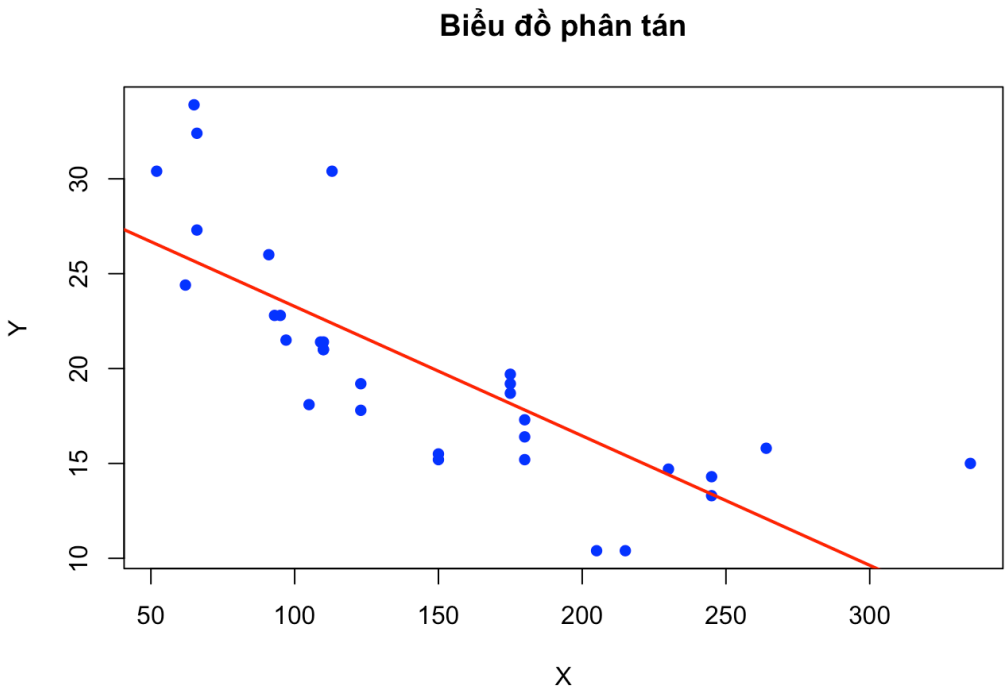



```
#install.packages("pheatmap")
library(pheatmap)
pheatmap(cor_matrix, display_numbers = TRUE, main = "Heatmap", cluster_rows = FALSE, cluster_cols=FALSE)
```



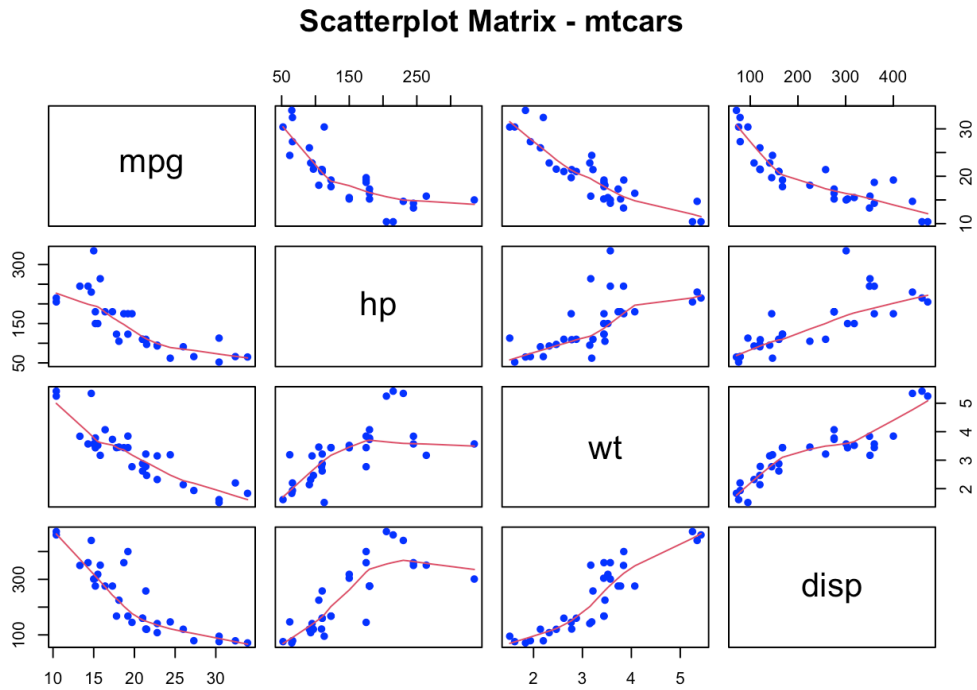
Biểu đồ phân tán `plot()`

```
# Dữ liệu mtcars
data(mtcars)
# Biểu đồ phân tán
plot(mtcars$hp, mtcars$mpg, col = "blue", pch = 16, main = "Biểu đồ phân tán", xlab = "X", ylab = "Y")
# Thêm đường hồi quy tuyến tính
abline(lm(mpg ~ hp, data = mtcars), col = "red", lwd = 2)
```



Biểu đồ ma trận phân tán `pairs()`

```
# Vẽ scatterplot matrix cho các biến c("mpg", "hp", "wt", "disp")
pairs(mtcars[, c("mpg", "hp", "wt", "disp")],
      main = "Scatterplot Matrix - mtcars",
      col = "blue",
      pch = 16,
      panel = panel.smooth)
```



Bài tập

Bài 1. Cho số liệu sau:

year	snow.cover
1970	6.5

year	snow.cover
1971	12.0
1972	14.9
1973	10.0
1974	10.7
1975	7.9
1976	21.9
1977	12.5
1978	14.5
1979	9.2

a. Nhập số liệu trên vào R.
b. Vẽ snow.cover theo year.
c. Lặp lại câu b. sau khi lấy logarit của biến snow.cover

Bài 2 Thống kê số liệu tỉ lệ lạm phát tại 4 nước trong giai đọan 1960-1980 được thu thập trong 2 bảng số liệu sau (Đvt: %)

Nam	US	Anh
1960	1.5	1.0
1961	1.1	3.4
1962	1.1	4.5
1963	1.2	2.5
1964	1.4	3.9
1965	1.6	4.6
1966	2.8	3.7
1967	2.8	2.4
1968	4.2	4.8
1969	5.0	5.2
1970	5.9	6.5
1971	4.3	9.5
1972	3.6	6.8
1973	6.2	8.4
1974	10.9	16.0
1975	9.2	24.2
1976	5.8	16.5
1977	6.4	15.9
1978	7.6	8.3
1979	11.4	13.4
1980	13.6	18.0

Nam	Nhat	Duc
1960	3.6	1.5
1961	5.4	2.3
1962	6.7	4.5
1963	7.7	3.0
1964	3.9	2.3
1965	6.5	3.4
1966	6.0	3.5
1967	4.0	1.5
1968	5.5	18.0
1969	5.1	2.6
1970	7.6	3.7
1971	6.3	5.3
1972	4.9	5.4
1973	12.0	7.0
1974	24.6	7.0
1975	11.7	5.9
1976	9.3	4.5
1977	8.1	3.7
1978	3.8	2.7
1979	3.6	4.1
1980	8.0	5.5

a. Nhập dữ liệu trên vào 2 data.frame lamphat1 và lamphat2 trong R

b. Trộn 2 data.frame trên vào 1 data.frame duy nhất là lamphat theo Nam.

c. Đếm số năm các nước US, Anh, Nhật, Đức có tỉ lệ lạm phát trên 5%.

d. Vẽ đồ thị phân tán về tỉ lệ lạm phát cho mỗi quốc gia theo thời gian. Cho nhận xét tổng quát về lạm phát của 4 nước?

e. Tính trung bình, trung vị, Max, Min, độ lệch chuẩn, sai số chuẩn của từng nước?

f. Để xác định lạm phát nước nào biến thiên nhiều hơn, ta cần dựa vào tham số thống kê nào? Kết luận?

g. Tạo một data.frame mới lamphat1 với số biến như trong data.frame lamphat nhưng không chứa dữ liệu của năm 1980.

h. Ta biết rằng hệ số của phương trình hồi quy tuyến tính $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 \hat{X}_i$ được xác định như sau:

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} X_i^2 - n(\bar{X})^2}$$

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}$$

Xác định các hệ số này trong mô hình hồi quy: lạm phát theo thời gian cho US bằng cách sử dụng data.frame lamphat1. Vẽ đồ thị phương trình hồi quy này?

i. Sử dụng phương trình hồi quy trong câu h) hãy xác định tỉ lệ lạm phát trong năm 1980 của US. So sánh với số liệu thực tế?

Bài 3 Để ước lượng số gỗ trong một khu rừng, người ta chia khu rừng thanh nhiều vùng, mỗi vùng có kích thước $50m \times 50m$. Chọn ngẫu nhiên 70 vùng và đếm số cây có đường kính lớn hơn 12cm, y. Dữ liệu của y thu thập được như sau:

7,8,6,4,9,11,9,9,9,10,9,8,11,5,8,5,8,8,7,8,3,5,8,7,10,7,8,9,8,11,10,8,9,8,9,9,7,8,13,8,9,6,7,9,9,7,9,5,6,5,6,9,8,8,4,4,7,7,8,9,10,2,7,10,8,10,6,7,7,8

a. Vẽ biểu đồ tần suất để mô tả dữ liệu trên.
b. Tính trung bình mẫu $\bar{y}$, và độ lệch chuẩn mẫu, s, của Y.
c. Xây dựng hàm `khoang()` theo biến x để tính giá trị các đầu mút của khoảng ước lượng ($\bar{y} \pm x \times s$). lần lượt với $x = 1, 2, 3$, tính tỷ lệ các mảnh đất nằm trong mỗi khoảng và so sánh phần trăm này với phần trăm được tính bằng quy tắc thực nghiệm.