```
t.test(x,alternative= c("two.sided", "less", "greater"),mu = mu_0, conf.level = 0.95)
# Trong đó:
```

- x: véc tơ dữ liệu.
- alternative: xác định kiểm định là hai phía ("two.sided"), bên trái ("less") hay bên phải ("greater"), mặc định là two.sided.
- mu = mu_0: giá trị cần kiểm định.
- conf.level: xuất ra khoảng tin cậy với độ tin cậy tương ứng.

Ví dụ 1.23. Trong file dữ liệu Diem_TN, ta thấy điểm toán (T) trung bình của 30 học sinh là 7,22. Chúng ta sẽ kiểm định điểm toán của học sinh toàn trường có thực sự thấp hơn 8 hay không với mức ý nghĩa 5%? Đồng thời xác định khoảng tin cậy 95% cho điểm toán trung bình của toàn trường với câu lệnh"

```
t.test(Diem_TN$T, mu = 8, conf.level = 0.95)
Kết quả hiển thị:
One Sample t-test
data:  Diem_TN$T
t = -4.7092, df = 29, p-value = 5.689e-05
alternative hypothesis: true mean is not equal to 8
95 percent confidence interval:
 6.881241 7.558759
sample estimates:
mean of x
     7.22
```

Từ kết quả trên ta thu được"

- Thống kê kiểm định t = -4.7092, bậc tự do n – 1 = 29, p-value = 5.689e-05.
- Khoảng tin cậy 95%: $6.881241 \le a \le 7.558759$.
- Với mức ý nghĩa 5%, ta thấy p-value (rất thấp) < 0.05, do đó bác bỏ H_0 tức điểm trung bình toán của học sinh toàn trường thấp hơn 8.

Nếu sử dụng giá trị thống kê t = -4.7092, ta so sánh với $t_{1-\alpha/2}^{n-1} = t_{0,975}^{29} = 2.045$ (dùng lệnh `qt(0.975, 29)`) ta cũng có kết luận tương tự.

- ***Kiểm định giả thuyết cho tỷ lệ một mẫu***

Giả sử cần kiểm định tỷ lệ phần tử thỏa mãn tính chất A trong tổng thể. Khảo sát một cỡ mẫu n. Gọi m là tổng số phần tử thỏa mãn tính chất A trong n phần tử khảo sát, suy ra tỷ lệ mẫu: $f = \frac{m}{n}$. Giả thuyết cỡ mẫu khảo sát n phải tương đối lớn.

Giả thuyết: $H_0: p = p_0$; Đối thuyết: $H_1: \begin{cases} p \neq p_0 \\ p < p_0 \\ p > p_0 \end{cases}$ (Một trong 3 trường hợp).

Tính thống kê kiểm định: $u = \frac{f - p_0}{\sqrt{p_0.(1-p_0)}}.\sqrt{n}$.

Miền bác bỏ:

- Với H_1: $p \neq p_0$ bác bỏ H_0 nếu $u < -z_{1-\alpha/2}$ hoặc $u > z_{1-\alpha/2}$.
- Với H_1: $p < p_0$ bác bỏ H_0 nếu $u < -z_{1-\alpha}$.
- Với H_1: $p > p_0$ bác bỏ H_0 nếu $u > z_{1-\alpha}$.

Để tìm $z_{1-\alpha/2}$, sử dụng hàm qnorm(1-alpha/2).

Sử dụng hàm prop.test để kiểm định:

```
prop.test(m, n, p = p0, alternative = c("two.sided", "less", "greater"), conf.level = 0.95)
# trong đó:
```

- m: số phần tử thỏa mãn tính chất A trong n phần tử khảo sát.
- n: cỡ mẫu.
- alternative: xác định kiểm định là hai phía ("two.sided"), bên trái ("less") hay bên phải ("greater").
- p = p0: giá trị cần kiểm định.
- conf.level: xuất ra khoảng tin cậy với độ tin cậy tương ứng.

Ví dụ 1.24. Trong một cuộc bầu cử thị trưởng tại một thành phố, ứng cử viên A tin rằng có trên 50% người dân thành phố ủng hộ ông ta. Để kiểm định điều này, các chuyên gia thống kê chọn ngẫu nhiên 800 người dân trong thành phố, thấy có 448 người dân cho ý kiến ủng hộ ông A. Hãy nhận xét xem tuyên bố của ông A về tỷ lệ cử tri có đúng không với mức ý nghĩa 1%?

Ta có:

Cỡ mẫu khảo sát $n = 800$.

Số người dân ủng hộ ông A: $m = 448$.

Giả thuyết cần kiểm tra: $\begin{cases} H_0 : p = 0.5 \\ H_1 : p > 0.5 \end{cases}$

Trong đó p là tỷ lệ người dân thành phố ủng hộ ông A.

Sử dụng hàm `prop.test`:

```
> n=800;m=448
> prop.test(m,n,p=0.5,alternative="greater",conf.level=0.99)
Kết quả hiển thị:
1-sample proportions test with continuity
correction
data:  m out of n, null probability 0.5
X-squared = 11.281, df = 1, p-value = 0.0003915
alternative hypothesis: true p is greater than 0.5
99 percent confidence interval:
 0.5182781 1.0000000
sample estimates:
   p
0.56
```

Kết quả cho biết p-value = 0.0003915 <1% dẫn đến bác bỏ giả thuyết H_0, ta kết luận rằng tỷ lệ người dân ủng hộ ông A trong thành phố trên 50%.Khoảng tin cậy 99% cho tỷ lệ p là: $0.5182 \le p \le 1.0000$.

- ***Kiểm định trung bình hai mẫu***

Ví dụ 1.25. Xét dữ liệu Diem_TN, qua phân tích mô tả chúng ta thấy nam có điểm toán (T) trung bình (7.3) cao hơn nữ (7.1). Câu hỏi đặt ra là có phải thật sự điểm toán trung bình của nam và nữ khác nhau hay không.

Tính thống kê kiểm định: $t = \frac{\overline{x_2} - \overline{x_1}}{SED}$. Trong đó $\overline{x_1}$ và $\overline{x_2}$ là điểm toán trung bình của hai nhóm nam và nữ, và SED là độ lệch chuẩn của $(\overline{x_1} - \overline{x_2})$. SED có thể tính bằng công thức: $SED = \sqrt{SE_1^2 + SE_2^2}$ .Trong đó SE_1, SE_2 là sai số chuẩn (standard error) của hai nhóm nam và nữ. Theo lí thuyết xác suất, t tuân theo luật phân phối t với bậc tự do $n_1 + n_2 - 2$, trong đó n_1, n_2 là số mẫu của hai nhóm. Chúng ta có thể dùng R để trả lời câu hỏi trên bằng hàm t.test như sau:

```
> t.test(T~gioitinh)
# kết quả hiển thị:
        Welch Two Sample t-test
data:  T by gioitinh
t = 0.51659, df = 24.495, p-value = 0.6101
alternative hypothesis: true difference in means between group Nam and group Nu is not
equal to 0
95 percent confidence interval:
 -0.5184389  0.8651055
sample estimates:
mean in group Nam  mean in group Nu
         7.306667          7.133333
```

R trình bày các giá trị quan trọng trước hết:

```
t = 0.51659, df = 24.495, p-value = 0.6101
```

df là bậc tự do. Trị số p = 0.6101 cho thấy mức độ khác biệt giữa hai nhóm nam và nữ không có ý nghĩa thống kê (vì cao hơn 0.05 hay 5%).

```
95 percent confidence interval:
-0.5184389  0.8651055
```

Là khoảng tin cậy 95% về độ khác biệt giữa hai nhóm. Kết quả tính toán trên cho biết điểm toán trung bình của nữ có thể thấp hơn nam giới 0.52 hoặc cao hơn nam giới 0.86. Vì độ khác biệt quá lớn và đó là thêm bằng chứng cho thấy không có khác biệt có ý nghĩa thống kê giữa hai nhóm.

Kiểm định trên dựa vào giả thiết hai nhóm nam và nữ có khác phương sai. Nếu chúng ta có lí do để cho rằng hai nhóm có cùng phương sai, chúng ta chỉ thay đổi một thông số trong hàm t với `var.equal = TRUE` như sau:

```
> t.test(T~gioitinh, var.equal = TRUE)
Kết quả hiển thị:
        Two Sample t-test
data:  T by gioitinh
t = 0.51659, df = 28, p-value = 0.6095
alternative hypothesis: true difference in means between group Nam and group Nu is not
equal to 0
95 percent confidence interval:
 -0.5139801  0.8606468
sample estimates:
mean in group Nam  mean in group Nu
         7.306667          7.133333
```

Về mặt số liệu, kết quả phân tích trên có khác chút ít so với kết quả phân tích dựa vào giả định hai phương sai khác nhau, nhưng trị số p cũng đi đến một kết luận rằng độ khác biệt giữa hai nhóm không có ý nghĩa thống kê.

- ***Kiểm định tỷ lệ hai mẫu***

Cho hai mẫu với số đối tượng n_1 và n_2, gọi số phần tử thỏa mãn tính chất A trong mẫu 1 là m_1, trong mẫu 2 là m_2. Do đó, chúng ta có thể tính được tỉ lệ tương ứng trong hai mẫu là p_1, p_2. Lí thuyết xác suất cho phép chúng ta phát biểu rằng độ khác biệt giữa hai mẫu $d = p_1 - p_2$ tuân theo luật phân phối chuẩn với số trung bình 0 và phương sai bằng:

$V_d = (\frac{1}{n_1} + \frac{1}{n_2})\mathrm{p}(1-\mathrm{p})$. Trong đó: $\mathrm{p} = \frac{m_1 + m_2}{n_1 + n_2}$; $z = d / V_d$ tuân theo luật phân phối chuẩn với trung bình 0 và phương sai 1.

Ví dụ 1.26. Một nghiên cứu được tiến hành so sánh hiệu quả của thuốc chống gãy xương. Bệnh nhân được chia thành hai nhóm: nhóm A được điều trị gồm có 100 bệnh nhân, và nhóm B không được điều trị gồm 110 bệnh nhân. Sau thời gian 12 tháng theo dõi, nhóm A có 7 người bị gãy xương, nhóm B có 20 người gãy xương. Hỏi tỉ lệ gãy xương trong hai nhóm có bằng nhau (tức thuốc không có hiệu quả)?

Để kiểm định hai tỉ lệ này có thật sự khác nhau, chúng ta có thể sử dụng hàm `prop.test(x,n,π)` như sau:

```
> m<-c(7,20)
> n<-c(100,110)
> prop.test(m,n)
Kết quả hiển thị:
2-sample test for equality of proportions with continuity
        correction
data:  m out of n
X-squared = 4.8901, df = 1, p-value = 0.02701
alternative hypothesis: two.sided
```

```
95 percent confidence interval:
 -0.20908963 -0.01454673
sample estimates:
   prop 1    prop 2
0.0700000 0.1818182
```

Kết quả phân tích trên cho thấy tỉ lệ gãy xương trong nhóm 1 là 0.07 và nhóm 2 là 0.18. Phân tích trên còn cho thấy xác suất 95% rằng độ khác biệt giữa hai nhóm có thể là 0.01 đến 0.20 (tức 1 đến 20%). Với trị số p = 0.027, chúng ta có thể nói rằng tỉ lệ gãy xương trong nhóm A quả thật thấp hơn nhóm B.

c) Kiểm định Wilcoxon cho hai mẫu (`wilcox.test`)

Kiểm định t dựa vào giả thiết là phân phối của một biến phải tuân theo luật phân phối chuẩn. Nếu giả định này không đúng, kết quả của kiểm định t có thể không hợp lí.

Ví dụ 1.26. Trong *Ví dụ 1.25.* chúng thấy trong file dữ liệu `Diem_TN` điểm toán (`T`) không có phân phối chuẩn. Trong trường hợp này, việc so sánh giữa hai nhóm có thể dựa vào phương pháp phi tham số (non-parametric) có tên là kiểm định Wilcoxon, vì kiểm định này (không như kiểm định t) không tùy thuộc vào giả định phân phối chuẩn.

```
> wilcox.test(T~gioitinh)
# Kết quả hiển thị:
Wilcoxon rank sum test with continuity correction
data:  T by gioitinh
W = 121, p-value = 0.7383
alternative hypothesis: true location shift is not equal to 0
```

Trị số p = 0.7383 cho thất quả thật độ khác biệt về điểm toán giữa hai nhóm nam và nữ không có ý nghĩa thống kê.

d) So sánh phương sai (`var.test`)

Ví dụ 1.28. Sử dụng file dữ liệu `Diem_TN`, để kiểm định phương sai điểm toán (`T`) giữa hai nhóm nam và nữ có khác nhau không, ta dùng câu lệnh sau:

```
> var.test(T~gioitinh)
Kết quả hiển thị:
F test to compare two variances
data:  T by gioitinh
F = 0.45106, num df = 14, denom df = 14, p-value = 0.1485
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1514355 1.3435331
sample estimates:
ratio of variances
         0.4510638
```

Kết quả trên cho thấy độ khác biệt về phương sai giữa hai nhóm là 0.45 lần. Trị số p = 0.1485 cho thấy phương sai giữa hai nhóm khác nhau không có ý nghĩa thống kê.

e) Thủ tục kiểm định shapiro.test về phân phối chuẩn

Để kiểm định một luật phân phối mẫu xem liệu có tuân theo luật chuẩn hay không, chúng ta có thể sử dụng hàm shapiro.test có cấu trúc như sau:

```
shapiro.test(x)

trong đó: x: là dữ liệu mẫu
```

Ví dụ 1.29. Theo dõi năng suất café tại một nông trường thu được số liệu sau:

Năng suất (tạ/ha)	4-6	6-8	8-10	10-12	12-14	14-16	16-18	18-20	20-22
Số vườn	15	26	25	30	26	21	24	20	13

Với mức ý nghĩa 0,05 có thể kết luận năng suất café ở nông trường này tuân theo luật phân phối chuẩn?

```
dl<-rep(c(5,7,9,11,13,15,17,19,21),times=c(15,26,30,26,21,24,20,13))
shapiro.test(dl)
Kết quả hiển thị:
Shapiro-Wilk normality test
data:  dl
W = 0.94668, p-value = 9.158e-07
```

Qua bảng kết quả hiện thị ta thấy trị số W = 0.94668và trị số p-value = 9.158e-07. Vì vậy, kết quả nhận được là qua mẫu cụ thể chưa thể khẳng định năng suất tuân theo quy luật chuẩn.

1.2.3. Thống kê suy diễn trong các bài toán phân tích tương quan

Hệ số tương quan (r) là một chỉ số thống kê đo lường mối liên hệ tương quan giữa hai biến số. Hệ số tương quan có giá trị từ -1 đến 1. Hệ số tương quan bằng 0 (hay gần 0) có nghĩa là hai biến số không có liên hệ gì với nhau; ngược lại nếu hệ số bằng -1 hay 1 có nghĩa là hai biến số có một mối liên hệ tuyệt đối. Nếu giá trị của hệ số tương quan là âm ($r < 0$) có nghĩa là hai biến tương quan nghịch (biến này tăng thì biến kia giảm và ngược lại); nếu giá trị hệ số tương quan là dương ($r > 0$) có nghĩa là hai biến tương quan thuận (hai biến cùng tăng hoặc cùng giảm).

Có nhiều hệ số tương quan trong thống kê, nhưng ở đây chúng ta sẽ trình bày 3 hệ số tương quan thông dụng nhất: hệ số tương quan Pearson r, Spearman ρ, và Kendall τ.

Trong tiểu mục này dữ liệu dùng để minh họa là file dữ liệu markettimng.csv tham khảo từ link: https://drive.google.com/drive/folders/1maNUAWyCcjXrU0m6hMgZNhjEI0jUI9Gu

```
library(readr)
marketing <- read_csv "marketing.csv"
head(marketing)
# kết quả hiển thị
```

```
1 youtube facebook newspaper sales
<dbl>   <dbl>    <dbl>     <dbl> <dbl>
1     1    201.    142.      145.   943.
2     2    156.    130.       62.7  856.
3     3    124.    188.      140.   965.
4     4    158.    187.      144.  1017.
5     5    158.    222.      116.  1115.
6     6    132.    182.      120.   932.
7     7    121.    214.      144.  1022.
8     8    108.     82.6     126.   650.
9     9    190.    173.      104.  1001.
10    10    117.    115.      133.   713.
# ... with 190 more rows
```

Bảng 1.4: Dữ liệu quan sát số lượt quảng cáo, (nguồn: internet).

a) Hệ số tương quan mẫu

- ***Hệ số tương quan Pearson***

Cho hai biến số x và y từ n mẫu, hệ số tương quan Pearson được tính bằng công thức sau đây:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

Trong đó, $\bar{x}$ và $\bar{y}$ là giá trị trung bình của biến số x và y.

Để tính hệ số tương quan Pearson trong R, cú pháp như sau:

```
cor(data, method = "pearson")
```

Ví dụ 1.30. Sử dụng file dữ liệu markettimng.csv, ta sẽ tính hệ số tương quan pearson giữa các biến trong đó:

```
cor(marketing, method = "pearson")
# kết quả hiển thị:
               ...1     youtube    facebook    newspaper        sales
...1       1.00000000 -0.04977015 -0.03918551 -0.177473371 -0.047532656
youtube   -0.04977015  1.00000000  0.08401121  0.047806059  0.487083735
facebook  -0.03918551  0.08401121  1.00000000 -0.039579633  0.903092760
newspaper -0.17747337  0.04780606 -0.03957963  1.000000000 -0.002900308
sales     -0.04753266  0.48708374  0.90309276 -0.002900308  1.000000000
```

Qua kết quả trên ta thấy: giữa biến sales và facebook có tương quan rất mạnh 0.903, còn lại đều có mức tương quan yếu giữa các biến khác.

- ***Hệ số tương quan Spearman ρ***

Hệ số tương quan Pearson chỉ hợp lí nếu biến số x và y tuân theo luật phân phối chuẩn. Nếu x và y không tuân theo luật phân phối chuẩn, chúng ta phải sử dụng một hệ số tương quan khác tên là Spearman, một phương pháp phân tích phi tham số. Hệ số này được ước tính bằng cách biến đổi hai biến số x và y thành thứ bậc (rank), và xem độ tương quan giữa hai dãy số bậc. Do đó, hệ số còn có tên tiếng Anh là Spearman's Rank correlation.

Để tính hệ số tương quan spearman trong R, cú pháp như sau:

```
cor(data, method = " spearman ")
```

Ví dụ 1.31. Sử dụng file dữ liệu markettimng.csv, ta sẽ tính hệ số tương quan spearman giữa các biến trong đó:

```
cor(marketing, method = "spearman")
# kết quả hiển thị:
                ...1      youtube     facebook     newspaper         sales
...1        1.000000000 -0.007446197 -0.06512643 -0.1753965822 -0.0706097652
youtube    -0.007446197  1.000000000  0.06036389  0.0620604644  0.4422117186
facebook   -0.065126427  0.060363895  1.00000000 -0.0332461435  0.8994231601
newspaper  -0.175396582  0.062060464 -0.03324614  1.0000000000 -0.0004522618
sales      -0.070609765  0.442211719  0.89942316 -0.0004522618  1.0000000000
```

Kết quả phân tích cũng tương tự như trong Ví dụ 1.29.

- ***Hệ số tương quan Kendall τ***

Hệ số tương quan Kendall (cũng là một phương pháp phân tích phi tham số) được ước tính bằng cách tìm các cặp số (x, y)"song hành" với nhau. Một cặp (x, y) song hành ở đây được định nghĩa là hiệu (độ khác biệt) trên trục hoành có cùng dấu hiệu (dương hay âm) với hiệu trên trục tung. Nếu hai biến số x và y không có liên hệ với nhau, thì cặp số song hành bằng hay tương đương với cặp số không song hành.

Vì có nhiều cặp phải kiểm định, phương pháp tính toán hệ số tương quan Kendall đòi hỏi thời gian của máy tính khá cao. Tuy nhiên, nếu một dữ liệu dưới 5000 đối tượng thì một máy vi tính có thể tính toán khá dễ dàng.

Để tính hệ số tương quan Kendall trong R, cú pháp như sau:

```
cor(data, method = " kendall ")
```

Ví dụ 1.32. Sử dụng file dữ liệu markettimng.csv, ta sẽ tính hệ số tương quan kendall giữa các biến trong đó:

```
cor(marketing, method = "  kendall")
# kết quả hiển thị:
             ...1      youtube     facebook    newspaper        sales
...1       1.000000000 -0.005025631 -0.04532891 -0.11734553 -0.04874372
```

```
youtube    -0.005025631  1.000000000  0.03789516  0.04829995  0.30324656
facebook   -0.045328911  0.037895160  1.00000000 -0.02326925  0.72656918
newspaper  -0.117345529  0.048299952 -0.02326925  1.00000000 -0.00587984
sales      -0.048743719  0.303246559  0.72656918 -0.00587984  1.00000000
```

b) Kiểm định hệ số tương quan

Bên cạnh việc tính các giá trị tương quan mẫu, chúng ta cũng có thể kiểm định hệ số tương quan lý thuyết với giả thuyết kiểm định:

- H0: Không có tương quan (hệ số tương quan = 0).
- H1: Có tương quan.

Để tính kiểm định trong R, cú pháp như sau:

```
cor.test(nhân tố 1, nhân tố 2, method = c("pearson","spearman", "kendall"))
```

Trong đó:

Nhân tố 1, nhân tố 2 là 2 biến cần kiểm định tính tương quan.

methol được lựa chọn một trong 3 phương pháp tương ứng.

Ví dụ 1.33. Sử dụng file dữ liệu markettimng.csv, ta sẽ kiểm định tính tương quan giữa 2 biến sales và youtube:

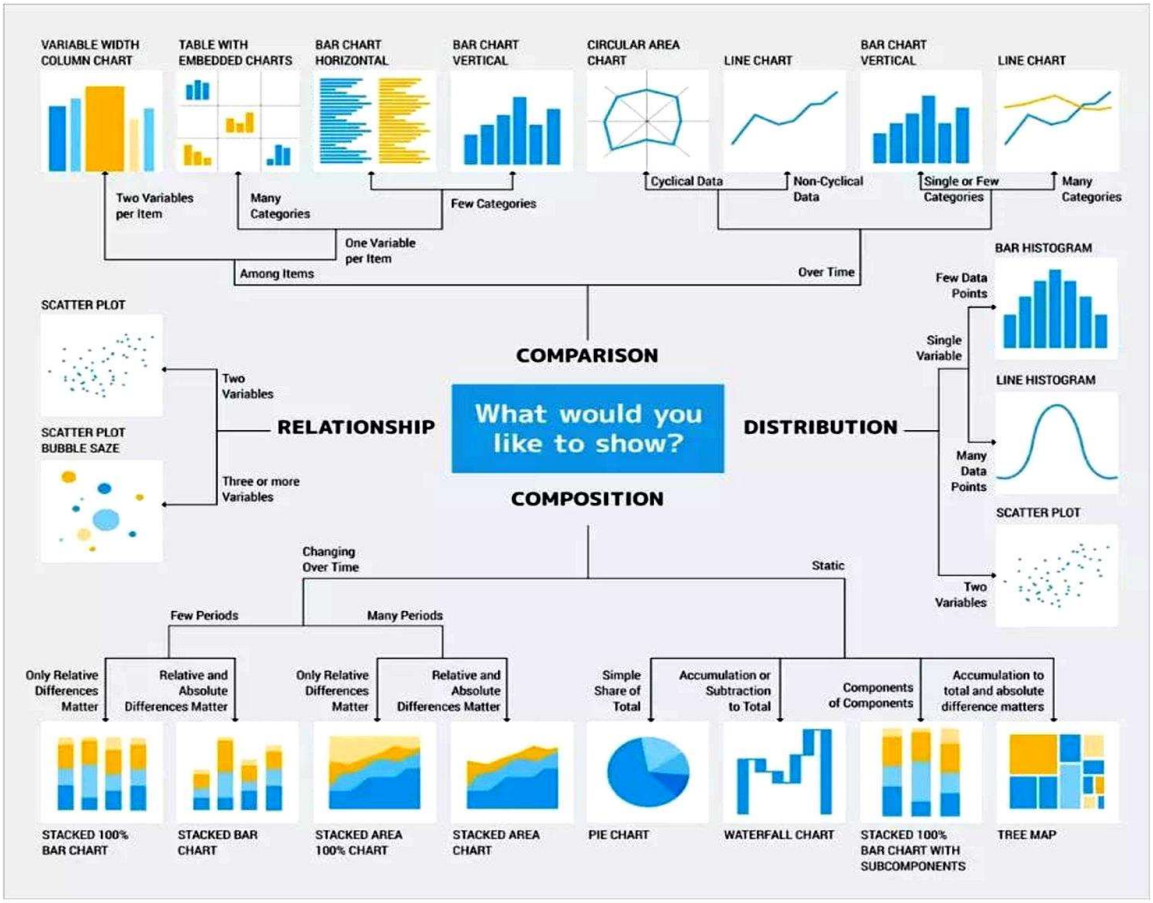
```
> cor.test(marketing$youtube, marketing$sales)
#Kết quả hiển thị:
        Pearson's product-moment correlation
data:  marketing$youtube and marketing$sales
t = 7.8478, df = 198, p-value = 2.597e-13
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3735893 0.5862096
sample estimates:
      cor
0.4870837
```

Kết quả trên với trị số p-value = 2.597e-13, cho thấy mối liên hệ giữa doanh thu (sales) và chi phí quảng cáo qua youtube là có ý nghĩa thống kê.

1.3. CÁC BƯỚC TRỰC QUAN HÓA DỮ LIỆU

Khi trực quan hóa dữ liệu, công việc đầu tiên là lựa chọn loại biểu đồ để thể hiện dữ liệu đang phân tích. Công việc này không chỉ là chọn một loại hay một số loại biểu đồ, mà có thể là lựa chọn kết hợp các biểu đồ, kết hợp biểu đồ và bảng phân tích thống kê mô tả, khi đó hiệu quả mang lại sẽ tối ưu nhất.

Việc lựa chọn biểu đồ phù hợp để trực quan dữ liệu phụ thuộc vào nhiều yếu tố, bao gồm loại dữ liệu, mục tiêu truyền đạt thông điệp, số lượng và phân loại của các biến dữ liệu, cũng như sở thích cá nhân. Hình 1.32 dưới đây là một số hướng dẫn để chọn biểu đồ phù hợp:



Hình 1.32: Cách thức lựa chọn biểu đồ.

1.3.1. Biểu đồ thể hiện kích thước dữ liệu trong trực quan hóa

Trong quá trình phân tích, một số trường hợp ta cần quan tâm đến sự khác biệt về độ lớn giữa các nhóm, chẳng hạn như sự khác biệt về dân số ở các thành phố khác nhau hoặc chênh lệch doanh thu của các nhãn hiệu ô tô khác nhau. Khi đó, trực quan bằng loại biểu đồ phù hợp sẽ giúp diễn giải kết quả một cách rõ ràng hơn. Theo tài liệu "Các nguyên tắc cơ bản về trực quan hóa dữ liệu", ba dạng biểu đồ thường được sử dụng để biểu diễn độ lớn của dữ liệu là: biểu đồ thanh, biểu đồ điểm và bản đồ nhiệt.

a) Biểu đồ thanh

Biểu đồ thanh trình bày một cách trực quan dữ liệu phân loại với các thanh hình chữ nhật có chiều cao hoặc chiều dài tỷ lệ với các giá trị mà chúng đại diện. Các thanh có thể được vẽ theo chiều dọc hoặc chiều ngang. Có 3 kiểu biểu đồ thanh thường gặp: biểu đồ thanh đơn áp dụng cho một biến phân loại, biểu đồ thanh nhóm (clustered) và biểu đồ thanh xếp chồng (stacked) áp dụng cho từ hai biến phân loại trở lên.

- ***Biểu đồ thanh đơn***

Biểu đồ thanh đơn được sử dụng khi chỉ có 1 biến phân loại, mỗi thanh sẽ đại diện cho một nhóm cụ thể, chiều cao hoặc độ dài của mỗi thanh tỷ lệ với tổng các giá trị trong nhóm mà nó đại diện.

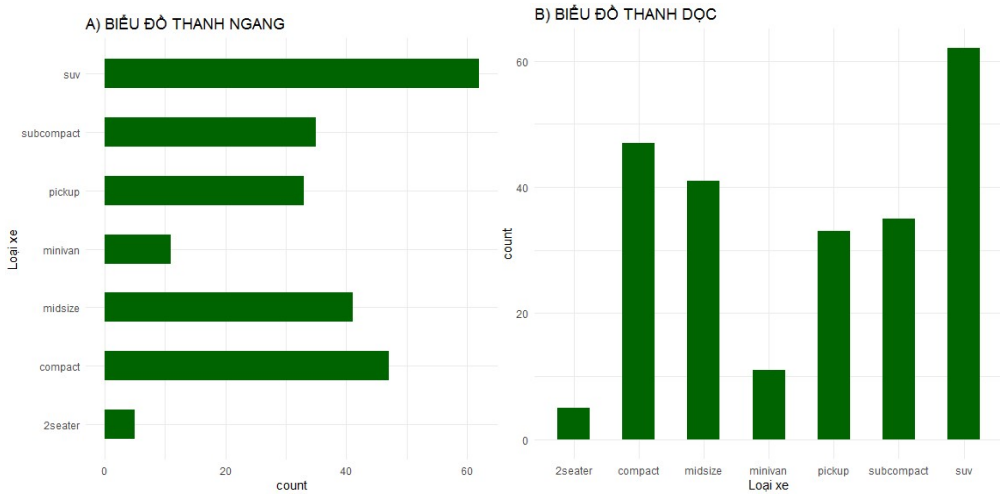
```
# Gói lệnh 1: một biến phân loại xét trên toàn bộ dữ liệu
```

```
ggplot(data) +
# Biểu đồ thanh ngang
geom_bar(mapping=aes(y =biến phân loại), width = ,fill=”màu sắc”)
# Biểu đồ thanh dọc
geom_bar(mapping=aes(x =biến phân loại), width = ,fill=”màu sắc”)
# Gói lệnh 2: một biến phân loại xét nhóm dữ liệu con
ggplot(data) +
geom_bar(aes(x =biến phân loại ,y =nhóm dữ liệu con), width = ,fill=”màu sắc”)
```

- Với geom_bar: Chỉ cung cấp phép gán cột trên một trục (trục x nếu muốn hiển thị theo thanh dọc, trục y nếu muốn hiển thị theo thanh ngang). Trục đối diện sẽ có tiêu đề là “số lượng” theo mặc định, bởi vì nó đại diện cho số hàng
- Với geom_col: Cung cấp phép gán cột trên cả hai trục (trục x hiển thị cho biến phân loại, trục y hiển thị nhóm dữ liệu con). Trục đối y thường là biến dạng numberic.

Ví dụ 1.34. Hình 1.32 dưới đây biểu diễn số lượng các loại xe trong tập dữ liệu theo 2 hình dạng biểu đồ thanh ngang (A) và thanh dọc (B).

```
# Biểu đồ thanh ngang
ggplot(mpg) +
  geom_bar(mapping=aes(y = class), width = 0.5,fill="darkgreen") +
  theme_minimal()+
  labs(title = "A) BIỂU ĐỒ THANH NGANG", y = "Loại xe")
# Biểu đồ thanh dọc
ggplot(mpg) +
  geom_bar(mapping=aes(x = class), width = 0.5, fill="darkgreen") +
  theme_minimal()+
  labs(title = "B) BIỂU ĐỒ THANH DỌC",  x = "Loại xe")
```



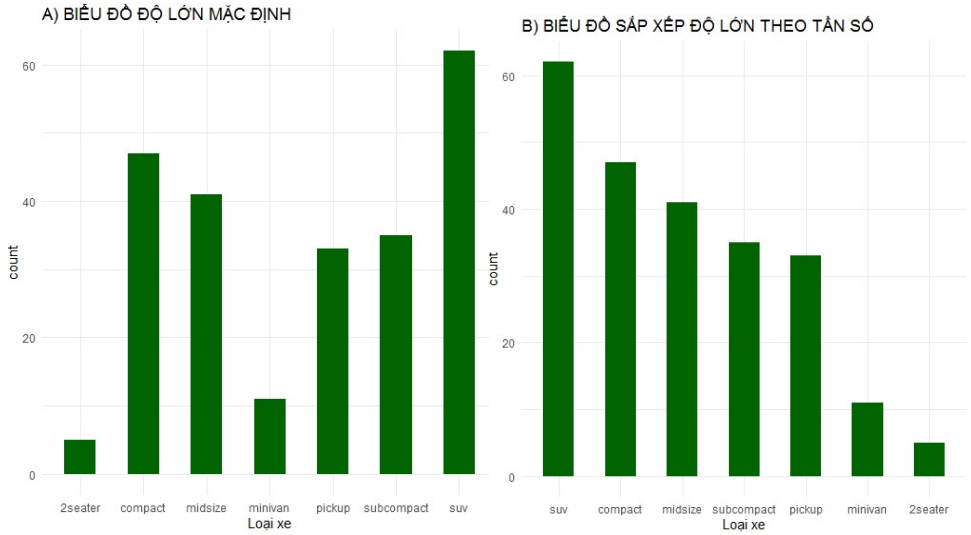
Hình 1.33: Biểu đồ thanh đơn.

Chú ý: Bất kể biểu đồ dọc hay ngang, ta đều cần chú ý đến thứ tự sắp xếp các thanh. Biểu đồ có thể sắp xếp mặc định các thanh theo thứ tự bảng chữ cái, theo độ cao hoặc độ dài thanh. Nhưng để biểu đồ trực quan cho người xem, các thanh nên được sắp xếp tương ứng theo tính chất của biến phân loại mà nó thể hiện:

- Biến phân loại mang tính rời rạc (ví dụ như quốc gia, thành phố, quận/huyện, …): biểu đồ thanh nên được sắp xếp theo độ lớn thanh từ cao đến thấp.
- Biến phân loại mang tính liên tục hay có thứ tự (ví dụ như theo chuỗi thời gian, theo độ tuổi, theo kích thước, …): biểu đồ thanh nên được sắp xếp theo thứ tự tăng hoặc giảm dần của biến phân loại.

Ví dụ 1.35. Hình 1.33 dưới đây biểu diễn số lượng các loại xe được sắp theo thứ tự giảm dần, việc này giúp quan sát và so sánh tần số giữa các loại xe dễ dạng hơn

```
# Biểu đồ mặc định
ggplot(mpg) +
  geom_bar(aes(x = class), width = 0.5, fill="darkgreen") +
  theme_minimal()+
  labs(title = "A) BIỂU ĐỒ ĐỘ LỚN MẶC ĐỊNH",  x = "Loại xe")
# Biểu đồ sắp thứ tự độ lớn
ggplot(mpg) +
  geom_bar(aes(x = fct_infreq(class)), width = 0.5, fill="darkgreen") +
  theme_minimal()+
  labs(title = "B) BIỂU ĐỒ SẮP XẾP ĐỘ LỚN THEO TẦN SỐ",  x = "Loại xe")
```

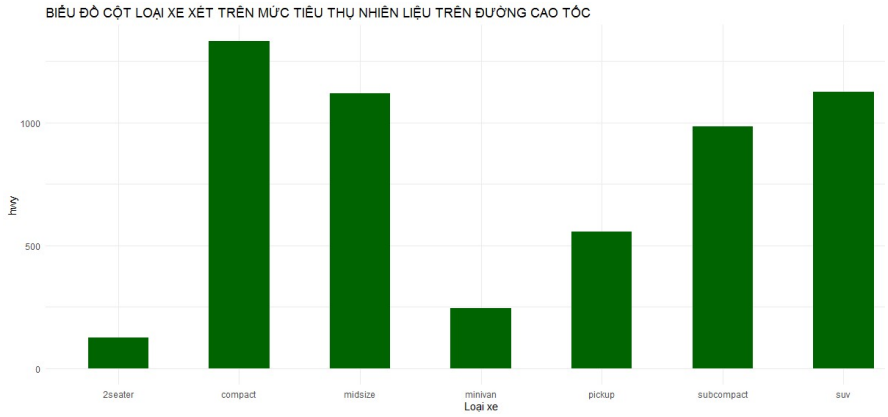


Hình 1.34: Biểu đồ thanh mặc định và biểu đồ thanh được sắp xếp.

Ví dụ 1.36. Hình 1.34 dưới đây biểu diễn số lượng các loại xe xét theo dữ liệu hiệu suất động cơ.

```
ggplot(mpg) +
  geom_col(aes(x = class,y=hwy), width = 0.5,fill="darkgreen") +
  theme_minimal()+
```

```
  labs(title = "BIỂU ĐỒ CỘT LOẠI XE XÉT TRÊN MỨC TIÊU THỤ NHIÊN LIỆU TRÊN ĐƯỜNG CAO TỐC", x = "Loại xe")
```



Hình 1.35: Biểu đồ thanh xét theo dữ liệu nhóm con.

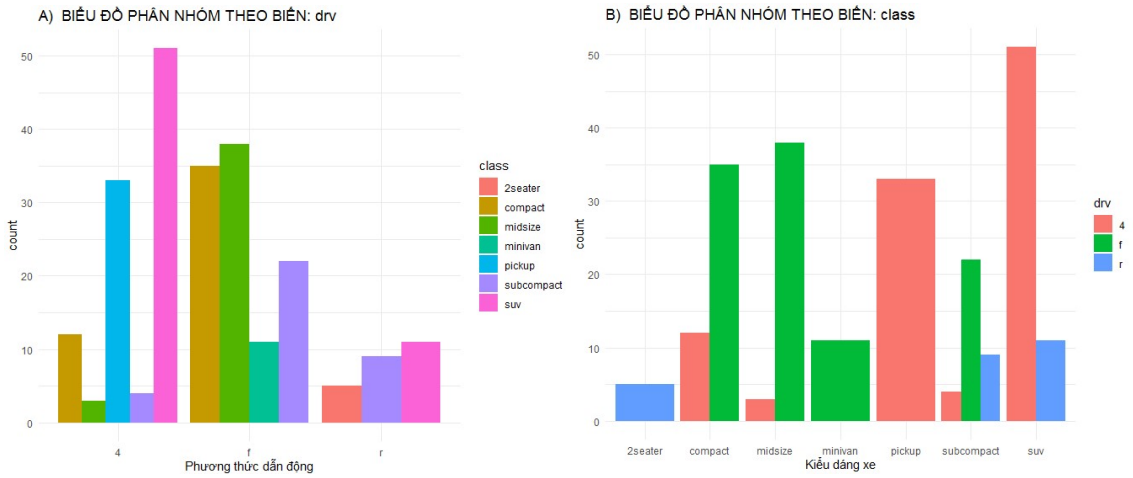
- ***Biểu đồ thanh nhóm (clustered)***

Biểu đồ thanh nhóm được sử dụng để biểu diễn cùng lúc từ 2 biến phân loại trở lên trong một tập dữ liệu. Trong đó, các nhóm của một biến phân loại được đặt cách đều dọc theo trục x, toàn bộ nhóm thuộc biến phân loại còn lại được đặt liền kề với nhau trong từng nhóm của biến phân loại trước.

```
Câu lệnh:
ggplot(data) +
geom_bar(mapping=aes(x =biến phân loại1,fill=biến phân loại 2),position = "dodge")
```

Ví dụ 1.37. Trong dữ liệu mpg ta xét hai biến phân loại là phương thức dẫn động drv và kiểu dáng xe class. Khi đó biểu đồ thanh nhóm có kết quả như sau:

```
# phân nhóm theo biến drv
ggplot(mpg) +
geom_bar(mapping=aes(x =drv,fill=class),position = "dodge")+
  theme_minimal()+
  labs(title = "A)  BIỂU ĐỒ PHÂN NHÓM THEO BIẾN: drv", x = "Phương thức dẫn động")
# phân nhóm theo biến class
ggplot(mpg) +
geom_bar(mapping=aes(x =class,fill=drv),position = "dodge")+
  theme_minimal()+
  labs(title = "B)  BIỂU ĐỒ PHÂN NHÓM THEO BIẾN: class", x = "Kiểu dáng xe")
```



Hình 1.36: Biểu đồ thanh nhóm.

- ***Biểu đồ thanh chồng (stacked)***

Biểu đồ thanh chồng cũng được sử dụng để biểu diễn cùng lúc từ 2 biến phân loại trở lên trong một tập dữ liệu. Khác với biểu đồ thanh nhóm có các thanh đặt liền kề nhau, biểu đồ thanh chồng đặt các thanh xếp chồng lên nhau.

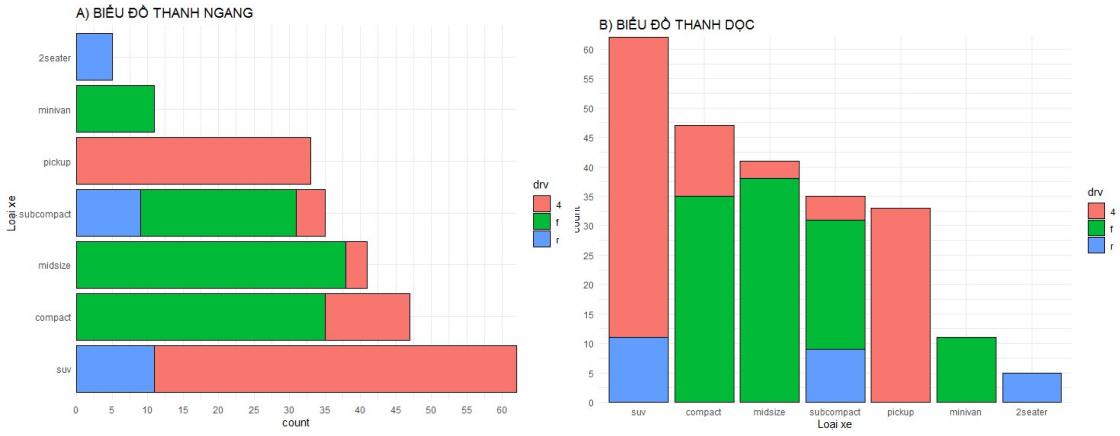
```
# Góilệnh 1: hai biến phân loại trong toàn bộ dữ liệu
ggplot(data)+
# Biểu đồ thanh ngang
geom_bar(mapping=aes(y =fct_infreq(nhân tố 1),fill = nhân tố 2), position = ""))
# Biểu đồ thanh dọc
geom_bar(mapping=aes(x =fct_infreq(nhân tố 1),fill = nhân tố 2), position = ""))
# Góilệnh 2: hai biến phân loại trong nhóm dữ liệu con
ggplot(data)+
geom_col(mapping = aes(x = nhân tố 1,y= dữ liệu nhóm con,fill = nhân tố 2 ),color = )
```

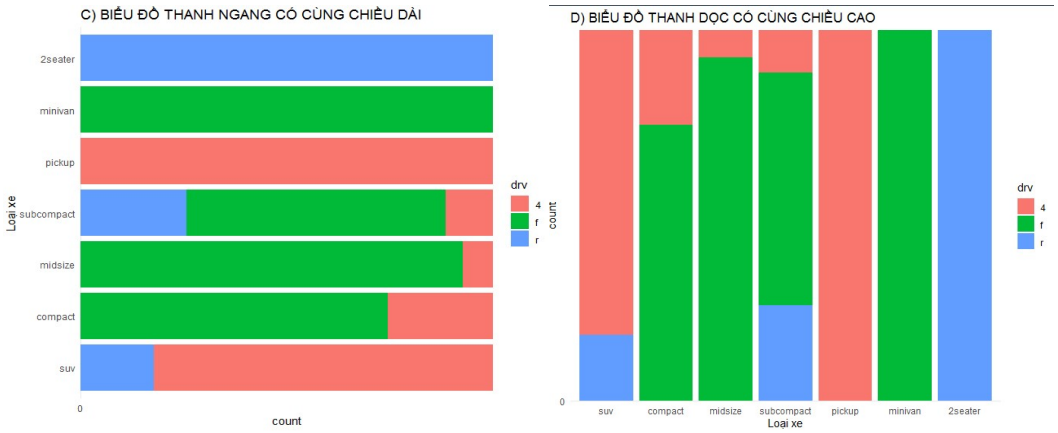
- Với geom_bar: Chỉ cung cấp phép gán cột trên một trục (trục x nếu muốn hiển thị theo thanh dọc, trục y nếu muốn hiển thị theo thanh ngang). Trục đối diện sẽ có tiêu đề là "số lượng" theo mặc định, bởi vì nó đại diện cho số hàng.
- Với geom_col: Cung cấp phép gán cột trên hai trục (trục x cho nhân tố 1, trục y cho dữ liệu nhóm con).
- Biến phân loại thứ hai được gán trong hàm fill tương ứng.
- Position nhận giá trị: `identity` nếu muốn giữ nguyên kích thước của nhóm; `fill` nếu muốn mỗi bộ thanh xếp chồng có cùng chiều cao.

Ví dụ 1.38. Hình 1.36 dưới đây biểu diễn số lượng các loại xe kết hợp với yếu tố dẫn động: f-cầu trước; r-cầu sau; 4-bốn hướng theo 2 hình dạng biểu đồ thanh ngang (A, C) và thanh dọc (B, D).

```
# Biểu đồ thanh ngang
ggplot(mpg) +
```

```
  geom_bar(mapping = aes(y = fct_infreq(class), fill = drv), position =
"identity") +
  theme_minimal()+
scale_x_continuous(expand = c(0,0),breaks = seq(from = 0,to = 70,by = 5))+
  labs(title = "A) BIỂU ĐỒ THANH NGANG", y = "Loại xe")
# Biểu đồ thanh dọc
ggplot(mpg) +
  geom_bar(mapping = aes(x = fct_infreq(class), fill = drv), position =
"identity") +
  theme_minimal()+
scale_y_continuous(expand = c(0,0),breaks = seq(from = 0,to = 70,by = 5))+
  labs(title = "B) BIỂU ĐỒ THANH DỌC", x = "Loại xe")
# Biểu đồ thanh ngang có cùng chiều dài
ggplot(mpg) +
  geom_bar(mapping = aes(y = fct_infreq(class), fill = drv), position = "fill") +
  theme_minimal()+
scale_x_continuous(expand = c(0,0),breaks = seq(from = 0,to = 70,by = 5))+
  labs(title = "C) BIỂU ĐỒ THANH NGANG CÓ CÙNG CHIỀU DÀI", y = "Loại xe")
# Biểu đồ thanh dọc có cùng chiều cao
ggplot(mpg) +
  geom_bar(mapping = aes(x = fct_infreq(class), fill = drv), position = "fill") +
  theme_minimal()+
scale_y_continuous(expand = c(0,0),breaks = seq(from = 0,to = 70,by = 5))+
  labs(title = "D) BIỂU ĐỒ THANH DỌC CÓ CÙNG CHIỀU CAO", x = "Loại xe")
```

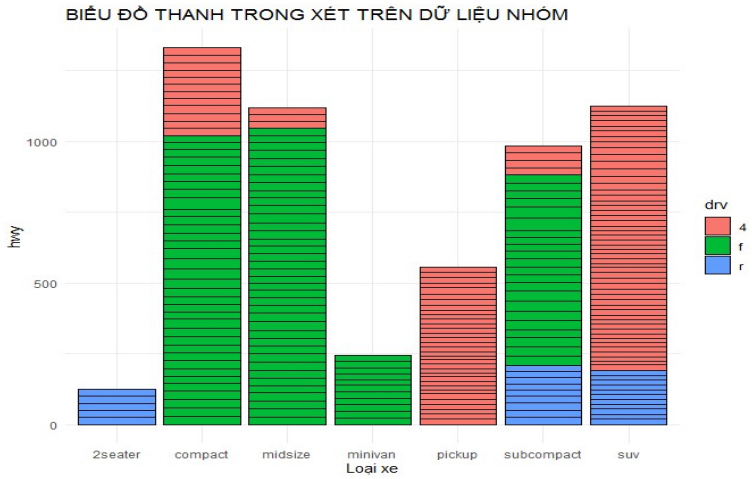




Hình 1.37: Biểu đồ thanh chồng.

Ví dụ 1.39. Hình 1.387 dưới đây biểu diễn số lượng các loại xe kết hợp với yếu tố dẫn động: f-cầu trước; r-cầu sau; 4-bốn hướng, xét theo dữ liệu nhóm con về mức tiêu thụ trên cao tốc.

```
ggplot(mpg) +
  geom_col(mapping = aes(,x=class,y =hwy, fill = drv), color = "black") +
  theme_minimal()+
  labs(title = "BIỂU ĐỒ THANH TRONG XÉT TRÊN DỮ LIỆU NHÓM", x = "Loại xe")
```



Hình 1.38: Biểu đồ thanh chồng kết hợp phân loại theo nhóm dữ liệu.

- ***Biểu đồ thanh phân nhóm***

Biểu đồ thanh chồng phù hợp với mục đích so sánh tổng giá trị giữa các nhóm (ví dụ Hình1.37. A, so sánh tổng số loại xe theo loại dẫn động) , nhưng đôi khi khó đạt sự phân biệt khi muốn so sánh các giá trị khác nhau trong cùng một nhóm. Để khắc phục điều này ta có thể sử dụng biểu đồ phân nhóm, đặc biệt khi so sánh ba yếu tố tác động.

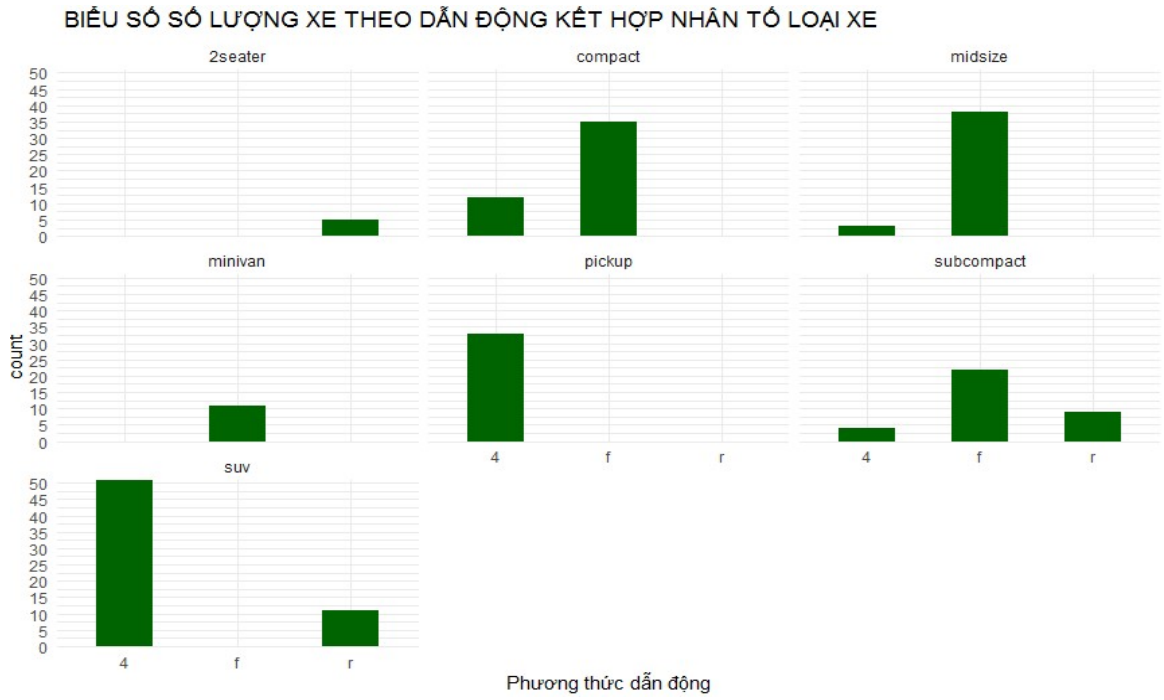
```
# gói 1: xét trên toàn bộ dữ liệu
ggplot(data) +
geom_bar(aes(x =biến phân loại ), width = ,fill=”màu sắc”)+
# chia biểu đồ theo một nhân tố kết hợp
facet_wrap(~ nhân tố kết hợp )
```

```
# chia biểu đồ theo hai nhân tố kết hợp
facet_grid(nhân tố kết hợp 1 ~ nhân tố kết hợp 2)
# gói 2: xét trên nhóm dữ liệu con
ggplot(data) +
geom_col(aes(x =biến phân loại, y= nhóm dữ liệu con ), width = ,fill=”màu sắc”)+
# chia biểu đồ theo một nhân tố kết hợp
facet_wrap(~ nhân tố kết hợp )
# chia biểu đồ theo hai nhân tố kết hợp
facet_grid(nhân tố kết hợp 1 ~ nhân tố kết hợp 2)
```

- Các facets được sắp xếp theo thứ tự bảng chữ cái, trừ khi biến có kiểu factor với các thứ bậc đã được xác định.
- Có thể sử dụng một số tùy chọn nhất định để xác định bố cục của các facets, (ví dụ: nrow = 1 hoặc ncol = 1 để kiểm soát số hàng hoặc cột mà chúng được sắp xếp).
- Đối với facet_wrap(), chúng sẽ thường chỉ viết một cột trước dấu ngã ~ chẳng hạn như facet_wrap(~drv).
- Chúng ta sử dụng facet_grid khi muốn đưa một biến thứ hai vào sắp xếp các biểu đồ con. Ở đây mỗi ô thể hiện sự giao nhau của các giá trị giữa *hai cột*.
- Đối với facet_grid() chúng cũng có thể chỉ định một hoặc hai cột tới công thức (grid rows ~ columns). Nếu chỉ muốn chỉ định một cột, hãy đặt một dấu chấm . ở một phía của dấu ngã chẳng hạn như facet_grid(. ~ drv) hoặc facet_grid(drv ~ .).

Ví dụ 1.40. Minh họa facet_wrap():

```
ggplot(mpg) +
  geom_bar(aes(x = drv), width = 0.5, fill="darkgreen") +
  theme_minimal()+
  labs(title = " BIỂU SỐ SỐ LƯỢNG XE THEO DẪN ĐỘNG KẾT HỢP NHÂN TỐ LOẠI XE",  x =
"Phương thức dẫn động")+
scale_y_continuous(expand = c(0,0),breaks = seq(from = 0,to = 70,by = 5))+
facet_wrap(~ class)
kết quả hiển thị:
```

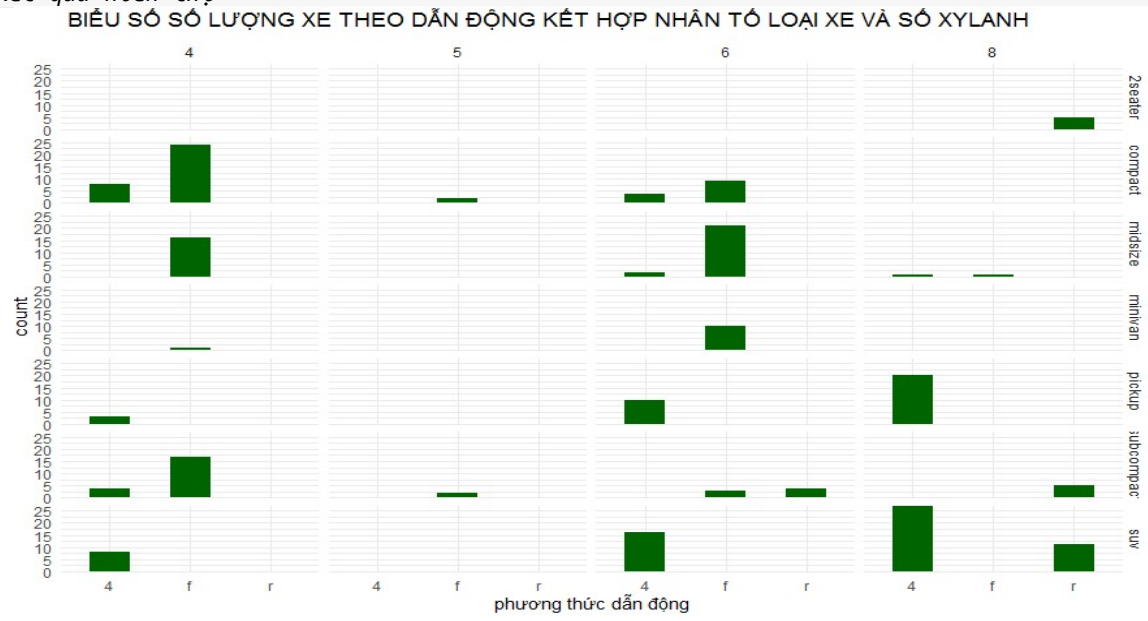


Hình 1.39: Biểu đồ phân nhóm ảnh hưởng một nhân tố.

Ví dụ 1.41. Minh họa facet_grid() xét toàn bộ dữ liệu

```
ggplot(mpg) +
  geom_bar(aes(x = drv), width = 0.5, fill="darkgreen") +
  theme_minimal()+
  labs(title = " BIỂU SỐ SỐ LƯỢNG XE THEO DẪN ĐỘNG KẾT HỢP NHÂN TỐ LOẠI XE VÀ SỐ
XYLANH",  x = "phương thức dẫn động")+
scale_y_continuous(expand = c(0,0),breaks = seq(from = 0,to = 70,by = 5))+
facet_grid(class ~ cyl)
```

kết quả hiển thị



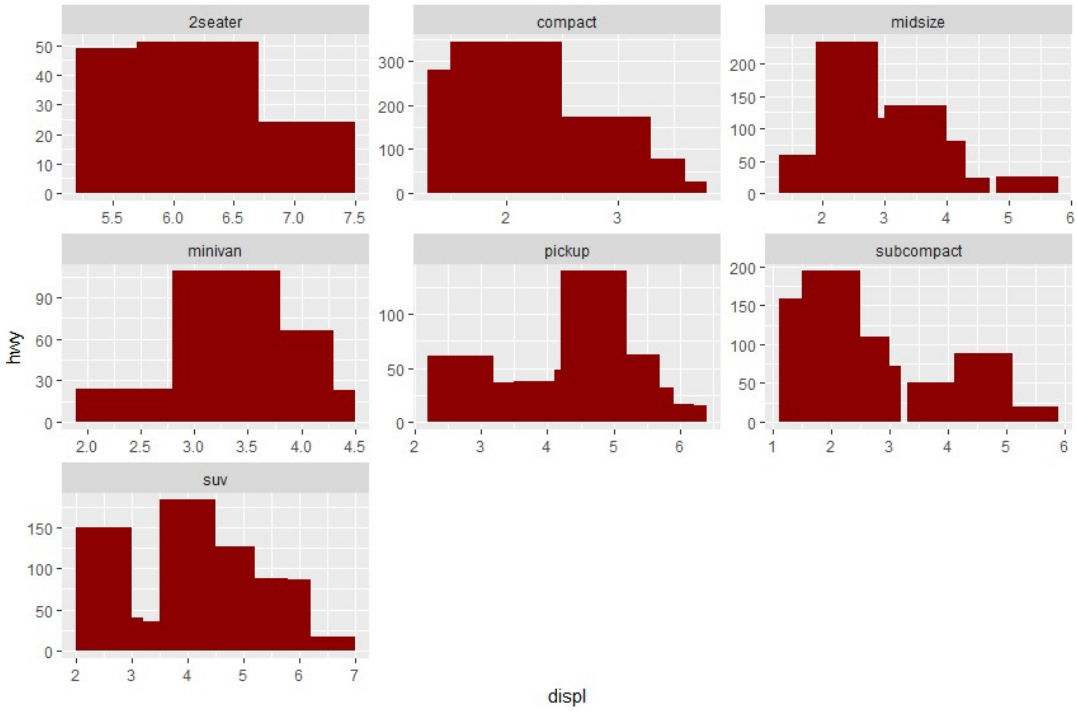
Hình 1.40: Biểu đồ phân nhóm ảnh hưởng hai nhân tố.

Nhận xét 3: Biểu đồ phân nhóm trên có thể kết hợp nhiều nhân tố khác nhau, đặc biệt là phân tích 2 nhân tố định lượng (numberic) thông qua một nhân tố định tính (character) hoặc hai nhân tố định tính ảnh hưởng.

Ví dụ 1.42. Xét nhân tố dung tích động cơ và loại xe theo mức tiêu thụ nhiên liệu khi xe di chuyển trên cao tốc.

```
ggplot(mpg, aes(x = displ, y = hwy)) +
  geom_col(width = 1, fill = "darkred") +
    facet_wrap(~ class, scales = "free")
```

kết quả hiển thị:



Hình 1.41: Biểu đồ với tỷ lệ cột ở vị trí tự do.

Trong biểu đồ trên tỷ lệ các cột được giải phóng ở dạng "fredom", dẫn đến tỷ lệ chia hợp với số liệu thực tiễn.

b) Biểu đồ điểm

Biểu đồ điểm sử dụng dấu chấm để biểu diễn vị trí của các giá trị dữ liệu. Biểu đồ điểm được sử dụng phổ biến để biểu diễn phân phối của một biến liên tục hay sự phân cụm trong một tập dữ liệu. Trong biểu diễn độ lớn của dữ liệu, biểu đồ thanh phải bắt đầu từ giá trị 0 để chiều dài thanh tỷ lệ với số lượng hiển thị. Tuy nhiên, với một số bộ dữ liệu, các thanh đôi khi quá dài và tất cả chúng đều có chiều dài gần như nhau, khiến biểu đồ không truyền tải được ý nghĩa. Do đó, biểu đồ điểm được sử dụng để thay thế biểu đồ thanh trong tình huống này.

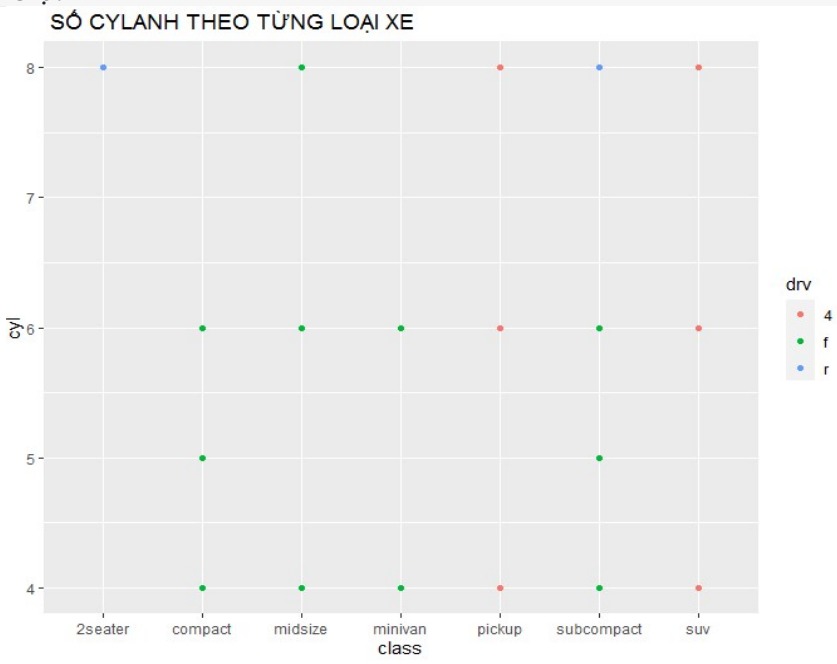
```
ggplot(data = mpg)+
  geom_point(mapping = aes(x = nhân tố chính,
                           y = nhân tố kết hợp,
                           color = nhân tố phân loại))+
```

```
  labs(title = " Tiêu đề ")
```

- Nhân tố 1: biến định tính hoặc định lượng cần khảo sát.
- Nhân tố kết hợp: nhân tố cần đo lường về độ lớn (thường dưới dạng numberic), trường hợp nhân tố kết hợp là định tính thì độ lớn sẽ lấy theo số lượng trên tập hợp dữ liệu chính.
- Nhân tố phân loại có thể khảo sát thêm thông qua màu sắc.

Ví dụ 1.43. Biểu đồ mô tả số cylanh trong các loại xe, kết hợp phân biệt phương thức dẫn động qua màu sắc.

```
ggplot(data = mpg)+
  geom_point(mapping = aes(x =class, y = cly,color = drv))+
  labs(title = " SỐ CYLANH THEO TỪNG LOẠI XE ")
kết quả hiển thị:
```



Hình 1.42: Biểu đồ điểm.

c) Bản đồ nhiệt (Heatmap)

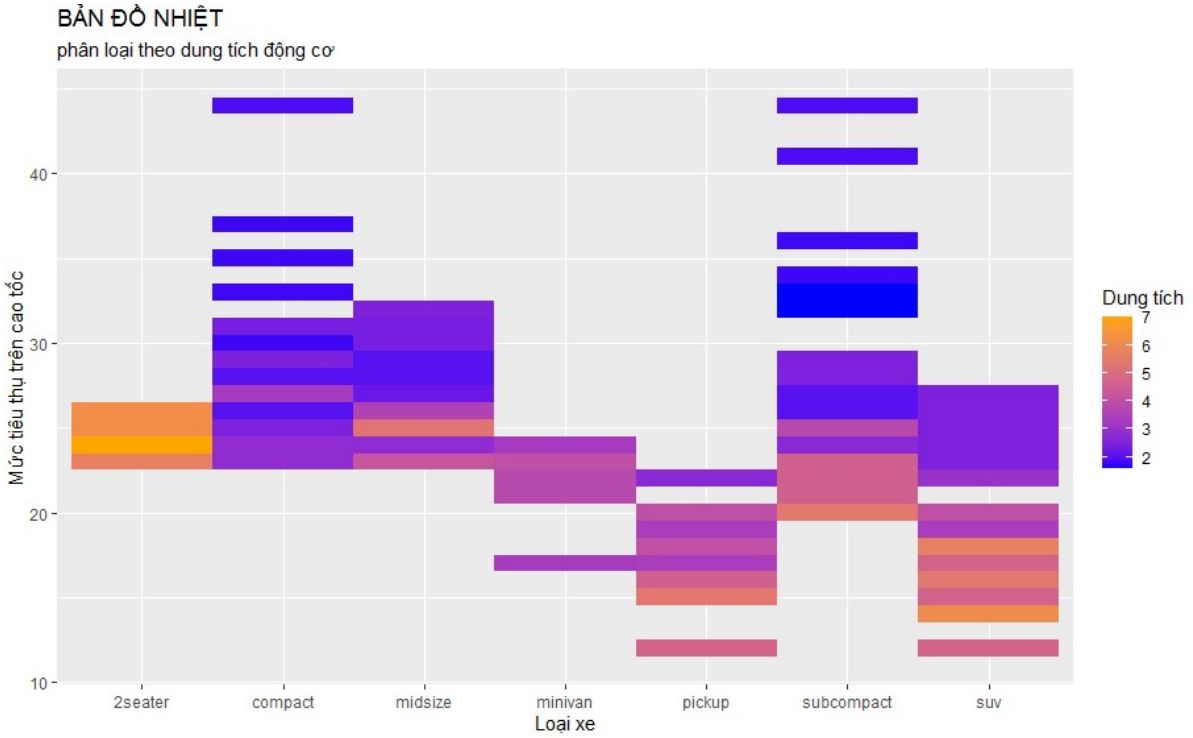
Bản đồ nhiệt là một kỹ thuật trực quan hóa thể hiện độ lớn của dữ liệu dưới dạng màu sắc trong hai chiều. Trong biểu diễn độ lớn của dữ liệu, bản đồ nhiệt được sử dụng để thay thế biểu đồ thanh trong trường hợp có 2 biến phân loại, gồm 1 biến rời rạc và 1 biến liên tục (thường theo chuỗi thời gian). Trong đó, số lượng nhóm trong mỗi biến phân loại quá nhiều, khiến việc biểu diễn bằng biểu đồ thanh nhóm hoặc biểu đồ thanh chồng trở nên không phù hợp. Mỗi giá trị dữ liệu trên bản đồ nhiệt được biểu diễn trong một ma trận với một màu đặc trưng. Thông thường, giá trị thấp hiển thị ở tông màu lạnh và dần chuyển đổi qua tông màu nóng khi có giá trị cao hơn.

```
ggplot(data = )+
  geom_tile( aes(x = nhân tố định tính,y = nhân tố định lượng, fill = Freq))+
```

```
  scale_fill_gradient()+
  labs(x =,
    subtitle =,
    fill =)
```

Ví dụ 1.44. Bản đồ nhiệt mô tả mối quan hệ giữa kiểu dáng xe với mức tiêu thụ trên cao tốc và dung tích động cơ.

```
ggplot(data = mpg)+
  geom_tile( aes(x = class,y = hwy, fill = displ))+
scale_fill_gradient( low = "blue",high = "orange")+
    labs(x = "Loại xe",y = "Mức tiêu thụ trên cao tốc",title = "BẢN ĐỒ NHIỆT",
    subtitle = "phân loại theo dung tích động cơ",
    fill = "Dung tích")
```



Hình 1.43: Bản đồ nhiệt.

1.3.2. Biểu đồ thể hiện tỷ lệ của dữ liệu trong trực quan hóa

Dạng biểu đồ sử dụng kích thước của dữ liệu đôi khi không giúp chúng ta nhận diện được mối quan hệ giữa các nhóm trong một tổng thể. Trong trường hợp này, ta có thể sử dụng biểu đồ tỷ lệ để biểu diễn các nhóm thành các phần riêng biệt, mà mỗi phần đại diện cho một tỷ lệ của tổng thể.

a) Một số dạng biểu đồ trực quan tỷ lệ đơn giản

Hai dạng biểu đồ tỷ lệ đơn giản và thường gặp nhất là biểu đồ tròn (pie chart) và biểu đồ bánh xe.

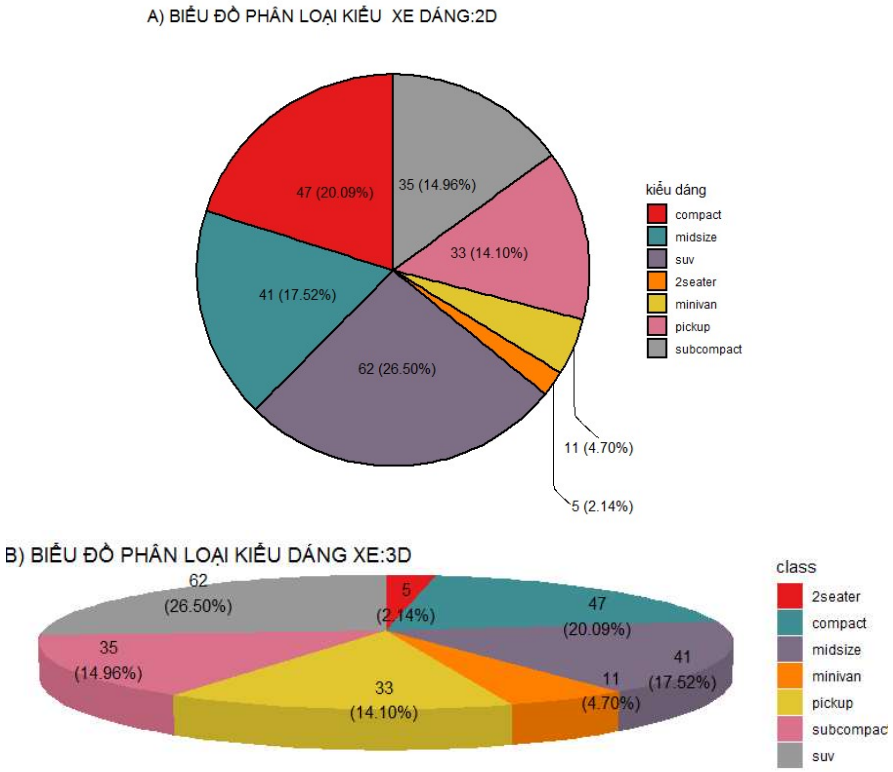
- ***Biểu đồ tròn (pie chart)***

Biểu đồ tròn chia một vòng tròn thành các phần, sao cho diện tích của mỗi phần tỷ lệ với giá trị mà nó biểu thị. Biểu đồ tròn phù hợp cho tập dữ liệu sử dụng một biến phân loại, với số lượng ít (nhỏ hơn 10).

```
# biểu đồ tròn dạng 2D
  ggpie(data =,   group_key = "biến phân loại", count_type = "full",
  label_info = "all", label_type = "horizon", label_split = NULL,
  label_size = 4, label_pos = "in", label_threshold = 10)+
labs(title =,     fill =)
# biểu đồ tròn dạng 3D
ggpie3D(data =,group_key = "biến phân loại", count_type = "full",
  tilt_degrees = -10, start_degrees = 0)+
  labs(title =,     fill =)
```

Ví dụ 1.45. Biểu đồ tròn mô tả phân loại kiểu dáng xe trong số liệu.

```
# BIỂU ĐỒ 2D
ggpie(data =mpg,   group_key = " class ", count_type = "full",
  label_info = "all", label_type = "horizon", label_split = NULL,
  label_size = 4, label_pos = "in", label_threshold = 10)+
labs(title ="A) BIỂU ĐỒ PHÂN LOẠI KIỂU DÁNG XE:2D",     fill ="kiểu dáng")
# BIỂU ĐỒ 3D
ggpie3D( data = mpg, group_key = "class", count_type = "full", tilt_degrees = -10)+
labs(title ="B) BIỂU ĐỒ PHÂN LOẠI KIỂU DÁNG XE:3D",     fill ="kiểu dáng")
```



Hình 1.44: Biểu đồ tròn dạng 2D và 3D.

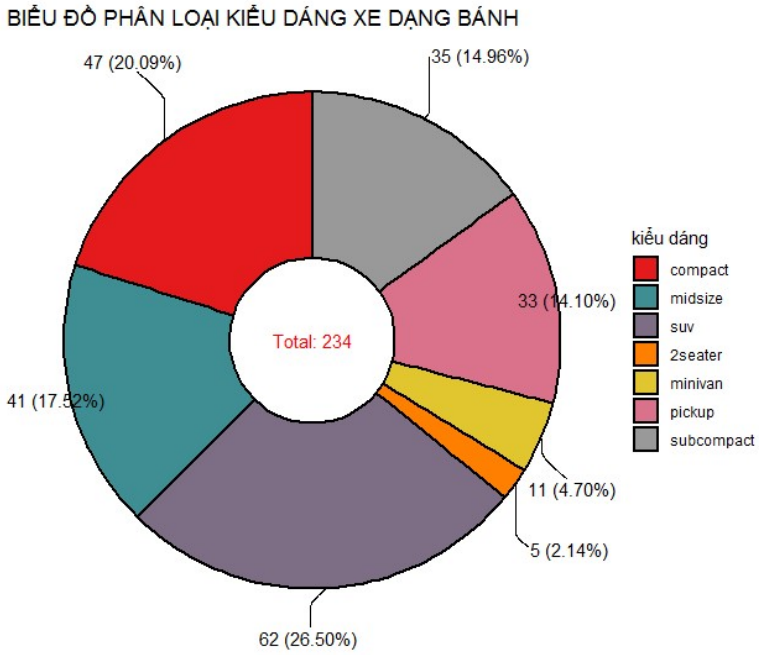
- ***Biểu đồ bánh***

Biểu đồ bánh, một dạng tương tự biểu đồ tròn, chỉ có một chút khác biệt là ở giữa có thêm ghi chú về về tổng số lượng mẫu hoặc một biến phân loại khác (ví dụ biến phân loại về thời gian, quốc gia,…).

```
# biểu đồ tròn dạng bánh
library(ggpie)
ggdonut(data =, group_key = "biến phân loại", count_type = "full",
  label_info = "all", label_type = "horizon", label_split = NULL,
  label_size = 4, label_pos = "in")+
  labs(title =" ",     fill ="")
```

Ví dụ 1.46. Biểu đồ bánh mô tả phân loại kiểu dáng xe trong số liệu.

```
ggdonut(data =mpg, group_key = "class", count_type = "full",
  label_info = "all", label_type = "horizon", label_split = NULL,
  label_size = 4, label_pos = "in")+
  labs(title =" BIỂU ĐỒ PHÂN LOẠI KIỂU DÁNG XE DẠNG BÁNH",     fill ="kiểu dáng")
```



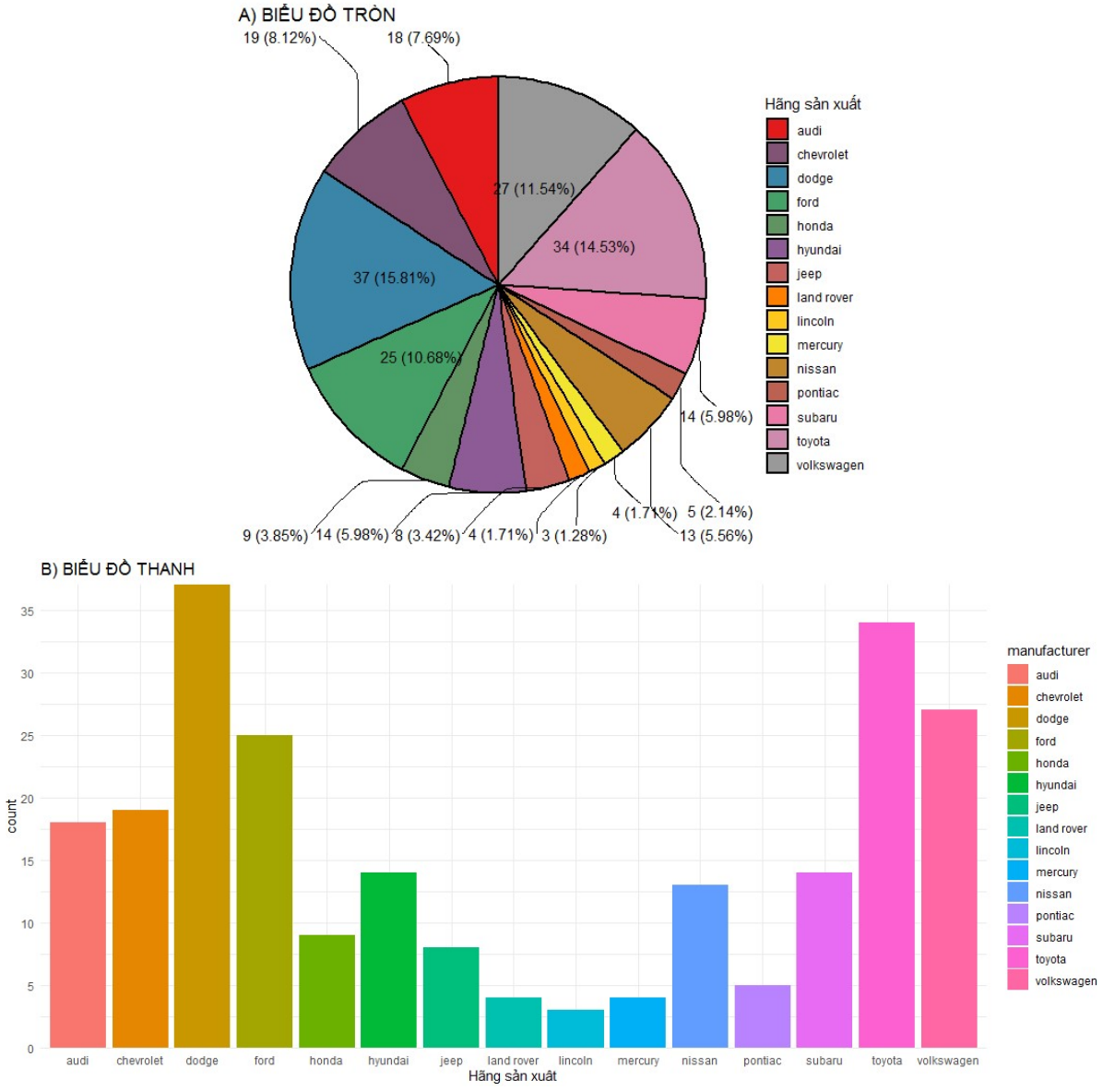
Hình 1.45: Biểu đồ bánh.

Biểu đồ tròn và biểu đồ bánh không phù hợp nếu số lượng nhóm trong một biến phân loại quá nhiều, hoặc tỷ lệ giữa các biến phân loại xấp xỉ bằng nhau (như Hình 1.45 A). Trong tình huống này, ta nên bổ sung thêm biểu đồ thanh đơn để so sánh sự chênh lệch về độ lớn (Hình 1.45 B).

Ví dụ 1.47. Biểu đồ tròn kết hợp biểu đồ thanh mô tả phân loại số lượng theo hãng sản xuất trong số liệu mpg.

```
# BIỂU ĐỒ TRÒN
ggpie(data =mpg,   group_key = "manufacturer", count_type = "full",
```

```
  label_info = "all", label_type = "horizon", label_split = NULL,
  label_size = 4, label_pos = "in", label_threshold = 10)+
labs(title ="A) BIỂU ĐỒ TRÒN",     fill ="Hãng sản xuất")
# BIỂU ĐỒ THANH
# Biểu đồ thanh dọc
ggplot(mpg) +
  geom_bar(mapping = aes(x =manufacturer,fill= manufacturer), position = "identity") +
  theme_minimal()+
scale_y_continuous(expand = c(0,0),breaks = seq(from = 0,to = 70,by = 5))+
  labs(title = "B) BIỂU ĐỒ THANH", x = "Hãng sản xuất")
```



Hình 1.46: Biểu đồ tròn kế hợp biểu đồ thanh.

b) Trực quan tỷ lệ với nhiều biến phân loại

Trong nhiều trường hợp phân tích, ta muốn đi sâu hơn và chia nhỏ tập dữ liệu theo nhiều biến phân loại cùng một lúc. Các trường hợp này được gọi là tỷ lệ lồng nhau, vì mỗi

biến phân loại bổ sung vào sẽ tạo ra một phần nhỏ hơn của dữ liệu được lồng trong các tỷ lệ trước đó. Một số dạng biểu đồ được sử dụng để biểu diễn các tỷ lệ lồng nhau là: *biểu đồ sunburst, biểu đồ cây (treemaps) và biểu đồ tập hợp song song (parallel sets plot)*.

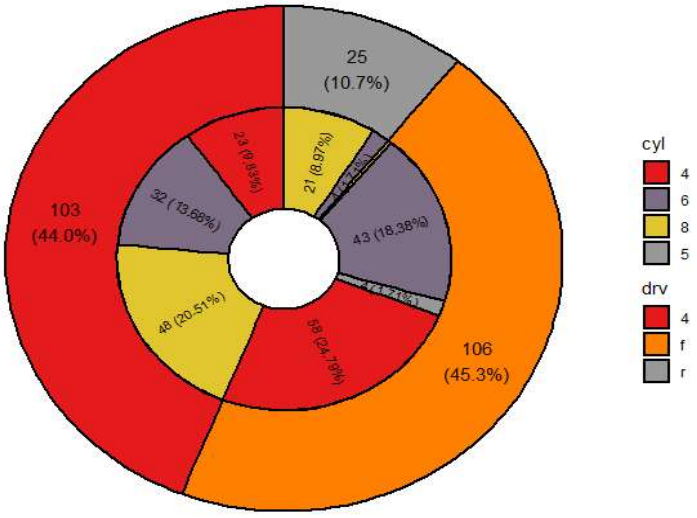
- ***Biểu đồ sunburst***

Biểu đồ sunburst có thể xem là một dạng mở rộng của biểu đồ tròn (pie chart), được sử dụng để trực quan hóa tập dữ liệu phân cấp. Thay vì chỉ biểu diễn một biến phân loại bằng một vòng tròn, biểu đồ sunburst sử dụng đồng thời nhiều dữ liệu phân loại theo thứ bậc, mỗi thứ bậc sẽ được biểu diễn bằng một vòng tròn đồng tâm. Với vòng tròn càng rộng, thứ bậc sẽ càng giảm, các lát trong có thể được tô màu để làm nổi bật thứ bậc hoặc danh mục muốn thể hiện.

```
ggnestedpie(data = , group_key = c("nhân tố 1", "nhân tố 2"),
 count_type = "full",  inner_label_info = "all",
 inner_label_split = NULL,  inner_label_threshold = 1,
 inner_label_size = 3,  outer_label_type = "horizon",
 outer_label_pos = "in",  outer_label_info = "all",
 outer_label_threshold = 10)+
  labs(title =,)
```

Ví dụ 1.48. Biểu đồ tỷ lệ giữa số xy lanh và phương thức dẫn động trong dữ liệu xe ô tô.

```
ggnestedpie(data =mpg , group_key = c("drv", "cyl"),
 count_type = "full",  inner_label_info = "all",
 inner_label_split = NULL,  inner_label_threshold = 1,
 inner_label_size = 3,  outer_label_type = "horizon",
 outer_label_pos = "in",  outer_label_info = "all",
 outer_label_threshold = 10)+
  labs(title ="BIỂU ĐỒ TỶ LỆ GIỮA SỐ XY LANH VÀ PHƯƠNG THỨC DẪN ĐỘNG")
```



Hình 1.47: Biểu đồ sunberst.

Biểu đồ sunburst sử dụng bố cục xuyên tâm để tạo hình ảnh trực quan của tập dữ liệu được phân loại. Nó cho thấy sự liên kết giữa các vòng tròn với nhau khi xử lý theo nhiều cấp độ. Do đó, biểu đồ sunburst rất hiệu quả để giới thiệu cách một vòng tròn được tách thành các phần cấu thanh ra nó, cũng như cho thấy sự đóng góp của một thứ nguyên cụ thể trong hệ thống phân cấp đó.

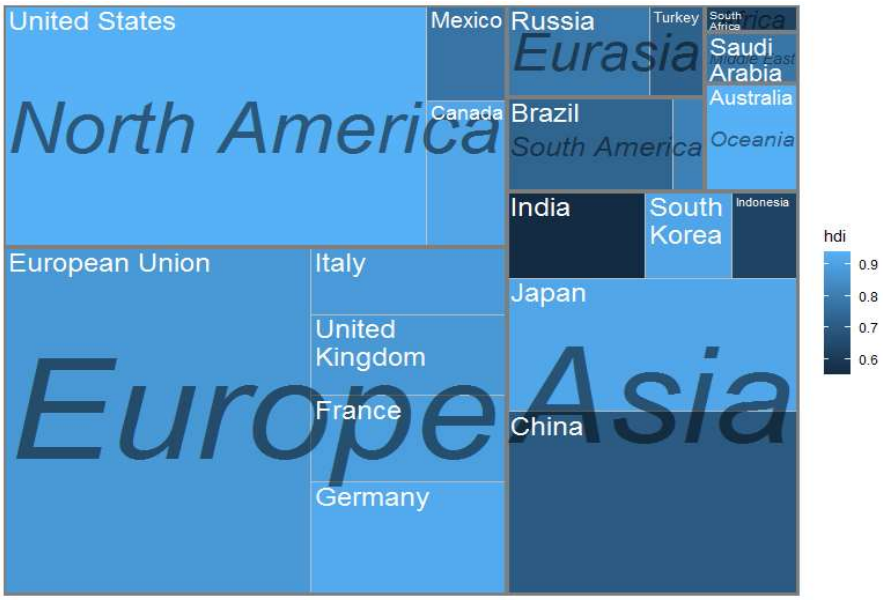
- ***Biểu đồ dạng cây (treemaps)***

Biểu đồ dạng cây là một phương pháp hiển thị dữ liệu phân cấp bằng cách sử dụng các hình lồng nhau (thường là hình chữ nhật). Mỗi nhánh của cây có một hình chữ nhật, được lồng các hình chữ nhật nhỏ hơn đại diện cho các nhánh phụ hoặc lá. Các nhánh phụ hoặc lá có diện tích tỷ lệ với giá trị của dữ liệu.

```
# Câu lệnh
ggplot(data, aes(area =nhân tố định lượng 1, fill = nhân tố phân biệt, label = ,
                 subgroup =)) +
  geom_treemap() +
  geom_treemap_subgroup_border() +
  geom_treemap_subgroup_text(place = "centre", grow = T, alpha = 0.5, colour =
                             "black", fontface = "italic", min.size = 0) +
  geom_treemap_text(colour = "white", place = "topleft", reflow = T)
```

Ví dụ 1.49. Biểu đồ cây với dữ liệu về các quốc gia G-20. Diện tích của ô sẽ được ánh xạ tới GDP của quốc gia và màu tô của ô sẽ được ánh xạ tới HDI (Chỉ số phát triển con người) của quốc gia đó như sau: (file dữ liệu G20 tích hợp cùng thư viện treemapify)

```
ggplot(G20, aes(area = gdp_mil_usd, fill = hdi, label = country,
                subgroup = region)) +
  geom_treemap() +
  geom_treemap_subgroup_border() +
  geom_treemap_subgroup_text(place = "centre", grow = T, alpha = 0.5, colour =
                             "black", fontface = "italic", min.size = 0) +
  geom_treemap_text(colour = "white", place = "topleft", reflow = T)
```



Hình 1.48: Biểu đồ cây.

Biểu đồ dạng cây hoạt động tốt ngay cả khi dữ liệu kết hợp cùng lúc nhiều biến định tính và định lượng. Đặc biệt khi có cùng lúc 2 dữ liệu định lượng tương ứng với kích thước các hình chữ nhật và màu sắc, mặc dù có thể khó diễn giải theo cách trực quan khác nhưng với biểu đồ dạng cây, ta có thể dễ dàng hiểu được ý nghĩa.

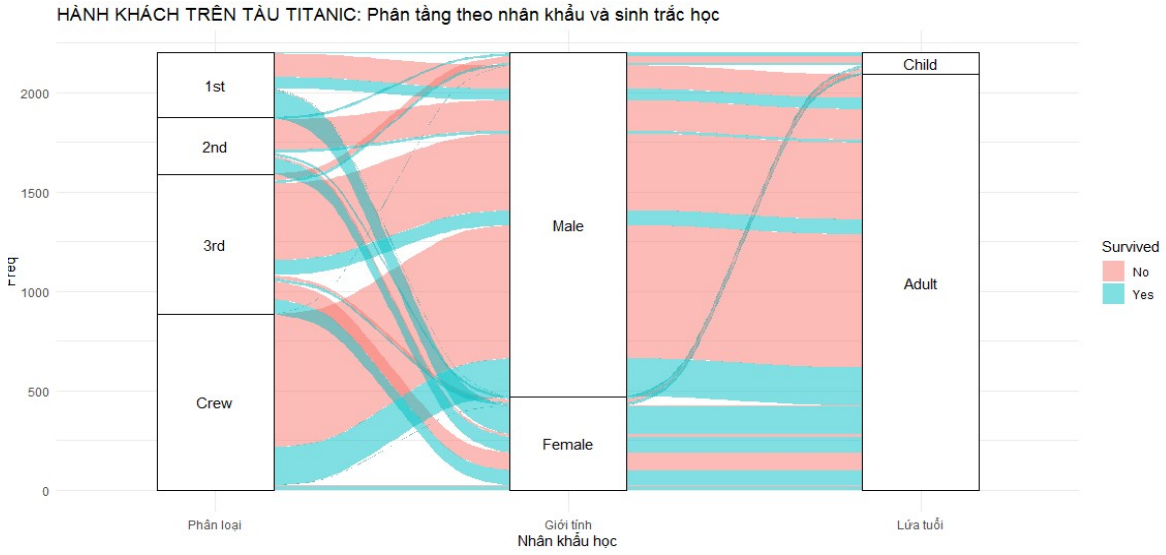
- ***Biểu đồ tập hợp song song (parallel sets plot)***

Khi dữ liệu có nhiều hơn hai biến phân loại, biểu đồ sunburst, biểu đồ dạng cây đều có thể nhanh chóng trở nên khó sử dụng. Trong trường hợp này, ta có thể dùng biểu đồ tập hợp song song. Theo đó, tổng số dữ liệu được chia nhỏ theo từng biến phân loại riêng lẻ và các nhóm của từng biến phân loại sẽ được liên kết với nhau thông qua các dải màu.

```
# Câu lệnh
library(ggplot2)
library(ggalluvial)
ggplot(data,
       aes(axis1 =nhân tố 1, axis2 =nhân tố 2, axis3 = nhân tố 3,
           y =)) +
  scale_x_discrete(limits = c("nhân tố 1", "nhân tố 2", "nhân tố 3"), expand = c(.2,
.05)) +
  xlab("") +
  geom_alluvium(aes(fill =)) +
  geom_stratum() +
  geom_text(stat = "stratum", aes(label = after_stat(stratum))) +
  theme_minimal() +
  ggtitle(" Tiêu đề")
```

Ví dụ 1.50. Xét file dữ liệu hành khách đi trên tàu Titanic (tích hợp trong thư viện `ggalluvial`), phân tầng theo nhân khẩu học và tỷ lệ sống sót.

```
# Câu lệnh
library(ggplot2)
library(ggalluvial)
ggplot(as.data.frame(Titanic),
       aes(axis1 = Class, axis2 = Sex, axis3 = Age,      y = Freq)) +
  scale_x_discrete(limits = c("Phân loại","Giới tính","Lứa tuổi"),expand=c(.2, .05)) +
  xlab("Nhân khẩu học") +
  geom_alluvium(aes(fill = Survived)) +
  geom_stratum() +
  geom_text(stat = "stratum", aes(label = after_stat(stratum))) +
  theme_minimal() +
  ggtitle("HÀNH KHÁCH TRÊN TÀU TITANIC: Phân tầng theo nhân khẩu và sinh trắc học")
```



Hình 1.49: Biểu đồ song song.

Ví dụ trong Hình 1.48, dữ liệu hành khách được chia nhỏ theo 3 biến: Phân tầng hành khách (hạng 1,2,3 và thủy thủ); Giới tính (Male; Female) và lứa tuổi (trẻ em, trưởng thành). Các dải màu kết nối các biến phân loại được bắt đầu từ trái qua phải và tô màu phân loại theo 2 loại vật liệu.

Biểu đồ cho thấy thủy thủ đoàn đa số là nam và tỷ lệ sống sót rất ít. Tỷ lệ trẻ em sống sót rất cao, điều này phù hợp thực tế là trẻ em được ưu tiên cứu hộ, còn thủy thủ đoàn được cứu hộ cuối cùng.

Chú ý: Khi dùng dạng biểu đồ tập hợp song song, ta nên bắt đầu dải màu theo hướng từ trái qua phải, điều này sẽ người xem dễ dàng quan sát dải màu bắt nguồn từ đâu và cách nó di chuyển qua tập dữ liệu. Ngoài ra, thứ tự sắp xếp của các biến phân loại trên biểu đồ cũng cần lưu ý sao cho các dải màu đan chéo nhau được giảm xuống mức tối thiểu.

1.3.3. Biểu đồ thể hiện phân phối của dữ liệu trong trực quan hóa

Trực quan hình dạng phân phối của dữ liệu đóng vai trò quan trọng việc phân tích dữ liệu và so sánh mức độ tập trung dữ liệu giữa các nhóm. Trong đó, các dạng biểu đồ đơn giản và thường gặp nhất là Histogram và biểu đồ mật độ, ngoài ra còn có các dạng biểu đồ biểu diễn nhiều phân phối cùng lúc như biểu đồ hộp, violin và ridgeline.

a) Biểu đồ Histogram và biểu đồ mật độ

Biểu đồ Histogram được sử dụng để mô tả trực quan sự phân bố tần suất cho tập dữ liệu khá phổ biến, ít nhất từ thế kỷ 18, vì dễ được vẽ bằng tay (Wilke, 2019). Tuy nhiên, hiện biểu đồ Histogram đang dần bị thay thế bởi các biểu đồ mật độ, do những hạn chế khi so sánh sự phân bố của một biến trên nhiều danh mục.

Biểu đồ mật độ biểu diễn dữ liệu liên tục bằng một đường cong được ước lượng từ dữ liệu với phương pháp ước lượng mật độ hạt nhân. Vì thế, càng có nhiều điểm dữ liệu trong tập dữ liệu, thì việc lựa chọn hạt nhân càng ít quan trọng, nên các biểu đồ mật độ có xu hướng đáng tin cậy và cung cấp nhiều thông tin đối với các tập dữ liệu lớn, nhưng lại có thể gây hiểu nhầm cho các tập dữ liệu chỉ một vài điểm.

- ***Biểu diễn phân bố của một biến phân loại***

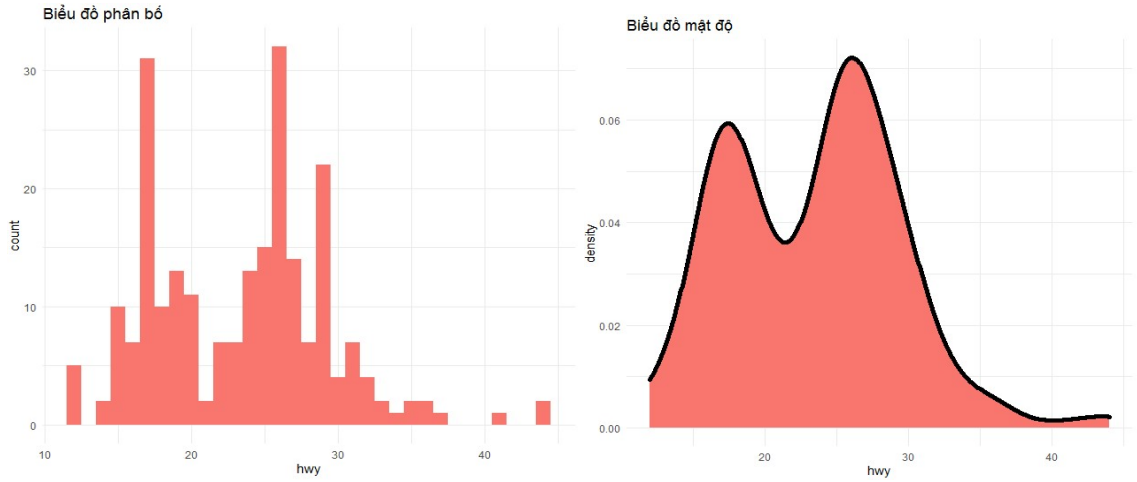
```
# Biểu đồ Histogram
ggplot(data = , aes(x = ))+
  geom_histogram(bins = 50, binwidth =)+
  labs(title = "Tiêu đề")
# Biểu đồ mật độ
ggplot(data =, mapping = aes(x = )) +
  geom_density(size = , alpha = )+
  labs(title = "Tiêu đề")
```

  - Bins= chỉ định số lượng các cột được ngắt giữa giá trị min và max của dữ liệu, có thể thay thế bằng `binwidth =`, khi đó một số lượng bins thích hợp sẽ được chọn phù hợp với bề rộng cột mà chúng ta chọn.
  - Đối số bên trong aes có thể thay bằng trục y.

Ví dụ 1.51. Xét file dữ liệu về ô tô mpg ta sẽ vẽ biểu đồ tần số và biểu đồ mật độ cho biến mức tiêu hao nhiên liệu khi di chuyển trên cao tốc.

```
# Biểu đồ Histogram
ggplot(data =mpg , aes(x =hwy, fill="darkgreen" ))+
  geom_histogram(binwidth = 1)+
theme_minimal()+
  labs(title = "Biểu đồ phân bố")
# Biểu đồ mật độ
ggplot(data = mpg, mapping = aes(x =hwy, fill="darkgreen"  )) +
  geom_density(size =2 , alpha =1)+
theme_minimal()+
```

```
  labs(title = "Biểu đồ mật độ")
```



Hình 1.50: Biểu đồ phân bố và mật độ.

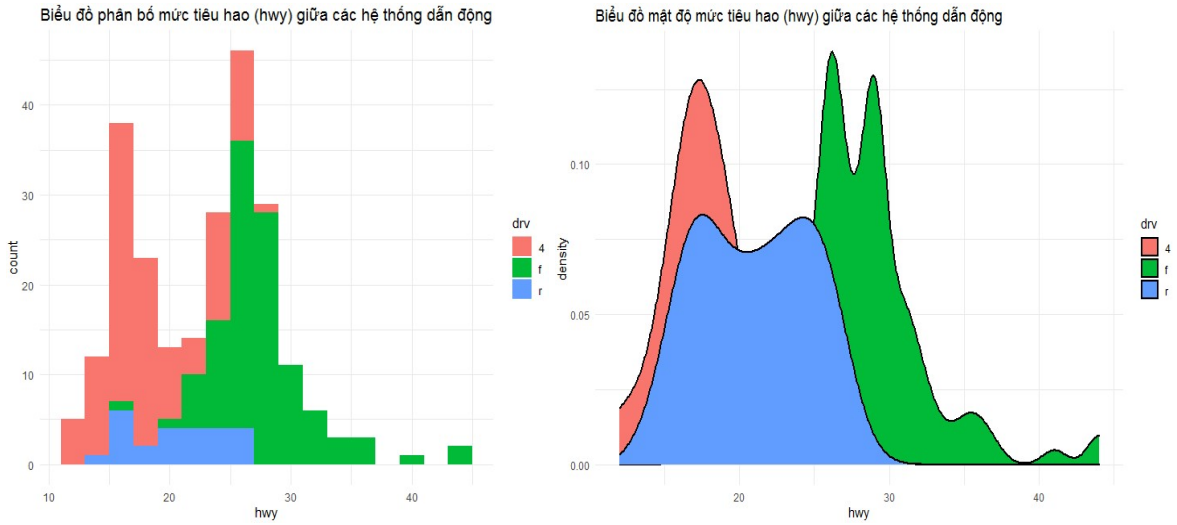
Nhận xét 4: Với 1 biến phân loại, hình dạng 2 biểu đồ gần như tương tự nhau và không khó để quan sát. Tuy nhiên, biểu đồ mật độ có xu hướng tạo ra sự xuất hiện của dữ liệu không tồn tại, đặc biệt là ở phần đuôi.

- ***Biểu diễn phân bố của nhiều biến cùng lúc***

```
# Biểu đồ Histogram
ggplot(data = , mapping = aes(x = , fill = nhân tố ảnh hưởng)) +
  geom_histogram(binwidth = 2)+
  labs(title = "Tiêu đề")
# Biểu đồ mật độ
ggplot(data = , mapping = aes(x = , y = after_stat(density), fill = nhân tố ảnh
hưởng)) +
  geom_density(size = 2, alpha = 0.2)+
    labs(title = "Tiêu đề")
```

Ví dụ 1.52. Xét file dữ liệu về ô tô mpg ta sẽ vẽ biểu đồ tần số và biểu đồ mật độ cho biến hwy và phân loại theo hướng dẫn động.

```
# Biểu đồ Histogram
ggplot(data =mpg , mapping = aes(x =hwy , fill =drv))+
  geom_histogram(binwidth = 2)+
theme_minimal()+
  labs(title = "Biểu đồ phân bố mức tiêu hao (hwy) giữa các hệ thống dẫn động")
# Biểu đồ mật độ
ggplot(data = mpg, mapping = aes(x =hwy , y = after_stat(density), fill =drv)) +
  geom_density(size =1 , alpha =1)+
theme_minimal()+
  labs(title = "Biểu đồ mật độ mức tiêu hao (hwy) giữa các hệ thống dẫn động ")
```



Hình 1.51: Biểu đồ phân bố và mật độ của nhiều biến trên cùng tọa độ.

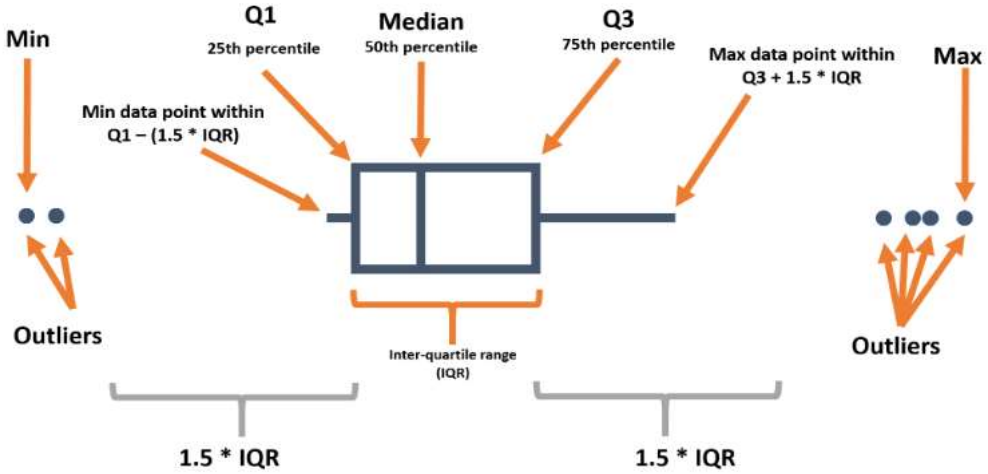
Khi dữ liệu có thêm nhiều biến phân loại như hệ thống dẫn động, cách trực quan theo dạng xếp chồng lên nhau của biểu đồ mật độ trở nên tối ưu, với các đường liên tục giúp các phân bố tách biệt nhau và dễ quan sát hơn biểu đồ Histogram .

b) Các dạng biểu đồ sử dụng cho trực quan nhiều phân phối cùng lúc

Trong trường hợp cần biểu diễn nhiều phân phối cùng lúc, chẳng hạn như mức tiêu hao nhiên liệu trên các loại xe khác nhau, các dạng biểu đồ như trên sẽ không còn phù hợp. Khi đó, người ta sử dụng các biểu đồ phân phối như biểu đồ hộp (box plots), biểu đồ violin (violin plots) và biểu đồ ridgeline (ridgeline plots).

- ***Biểu đồ hộp***

Biểu đồ hộp (Box plots) được nhà thống kê John Tukey tạo ra vào đầu những năm 1970. Nó nhanh chóng trở nên phổ biến, vì dễ vẽ bằng tay và mang lại nhiều thông tin. Biểu đồ hộp chia dữ liệu thành các phần tư và hiển thị chúng theo cách chuẩn hóa (Hình 1.51).



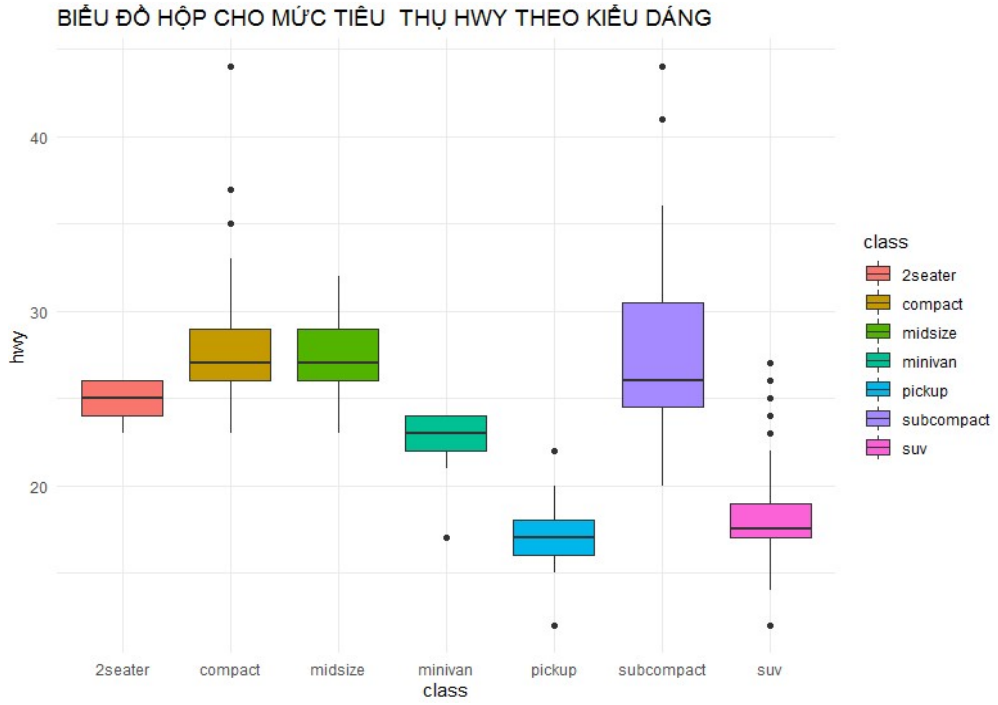
Hình 1.52: Các thông tin trên biểu đồ hộp.

```
# Câu lệnh
```

```
ggplot(data =, mapping = aes(y = biến khảo sát, x = nhân tố ảnh hưởng, fill = nhân tố ảnh hưởng)) +
  geom_boxplot()+
  theme(legend.position = "none")+
  labs(title = "")
```

Ví dụ 1.53. Xét file dữ liệu về ô tô mpg ta sẽ vẽ biểu đồ hộp biến hwy phân loại theo biến class.

```
  ggplot(data = mpg, mapping = aes(y = hwy, x = class, fill = class)) +
  geom_boxplot()+
  theme_minimal()+
  labs(title = "BIỂU ĐỒ HỘP CHO MỨC TIÊU  THỤ HWY THEO KIỂU DÁNG")
```

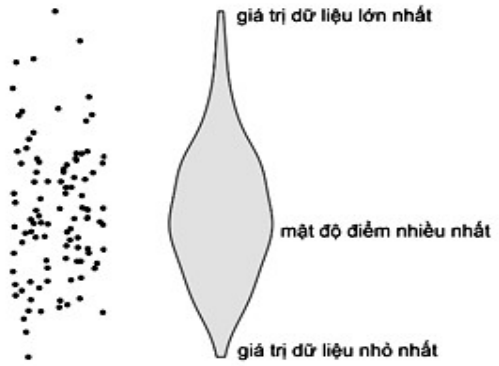


Hình 1.53: Biểu đồ hộp với biến phân loại.

Nhận xét 5: Biểu đồ hộp mặc dù có cung cấp thông tin hữu ích,nhưng có những hạn chế quan trọng, chúng có thể che khuất phân phối thực tế dẫn đến việc suy luận không chính xác

- ***Biểu đồ violin và jitter***

Với khả năng tính toán và trực quan hiện đại, biểu đồ hộp đang dần được thay thế bằng biểu đồ violin. Biểu đồ violin có thể cung cấp một bức tranh dữ liệu chính xác hơn biểu đồ hộp. Biểu đồ violin có tính chất đối xứng, nó bắt đầu và kết thúc ở các giá trị dữ liệu tối thiểu và tối đa, và phần dày nhất của biểu đồ tương ứng với mật độ điểm cao nhất trong tập dữ liệu (Hình 1.53).

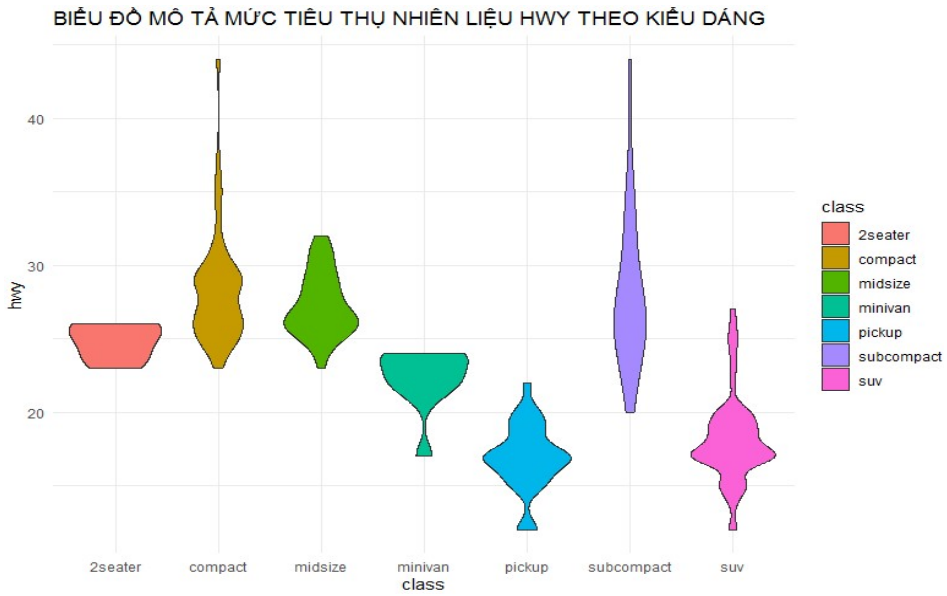


Hình 1.54: Các thông tin của biểu đồ violin.

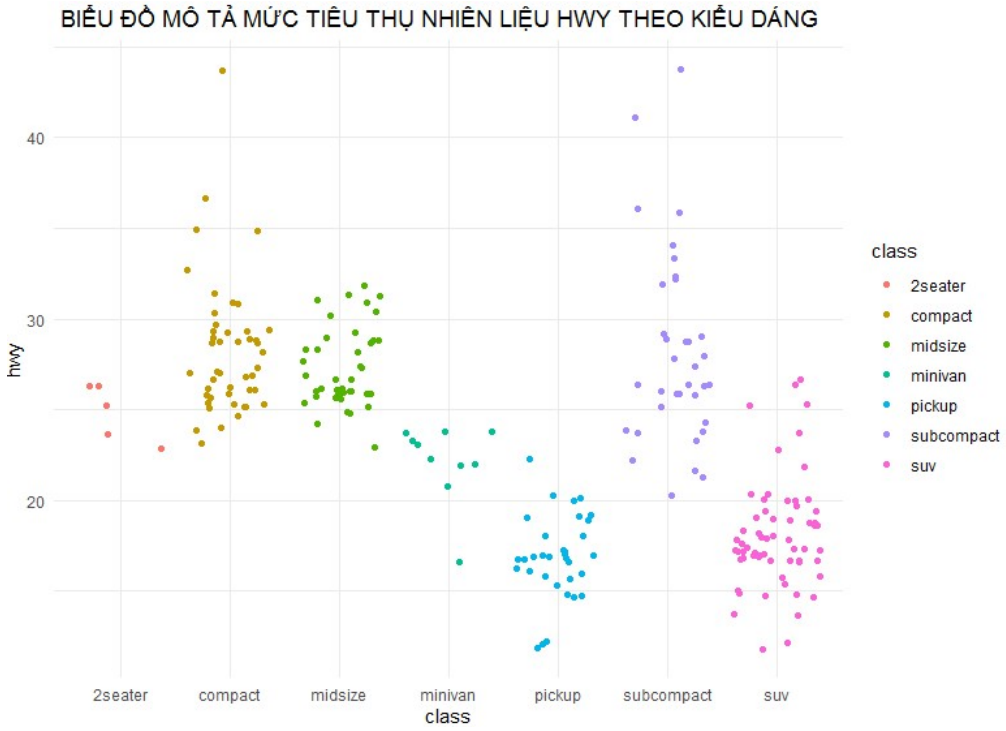
```
# Biểu đồ violin
ggplot(data =,
       mapping =aes(y=biến khảo sát,x =biến ảnh hưởng, fill = biến phân loại))+
  geom_violin()+
  labs(title = "")
# Biểu đồ jitter
ggplot(data =,mapping =aes(y =biến khảo sát,x=biến ảnh hưởng,color=biến ảnh hưởng))+
  geom_jitter()+
  labs(title = " ")
```

Ví dụ 1.54: Xét file dữ liệu về ô tô mpg ta sẽ vẽ biểu đồ hộp cho hiệu suất hwy theo kiểu dáng.

```
  # Biểu đồ violin
ggplot(data =mpg,
       mapping = aes(y = hwy, x = class,  fill = class))+
  geom_violin()+
theme_minimal()+
  labs(title = "BIỂU ĐỒ MÔ TẢ MỨC TIÊU THỤ NHIÊN LIỆU HWY THEO KIỂU DÁNG")
# Biểu đồ jitter
ggplot(data =mpg,  mapping = aes(y = hwy,x =class, color = class))+
  geom_jitter()+
theme_minimal()+
  labs(title = " BIỂU ĐỒ MÔ TẢ MỨC TIÊU THỤ NHIÊN LIỆU HWY THEO KIỂU DÁNG ")
```



Hình 1.55: Biểu Violin với biến phân loại.



Hình 1.56: Biểu đồ jitter với biến phân loại.

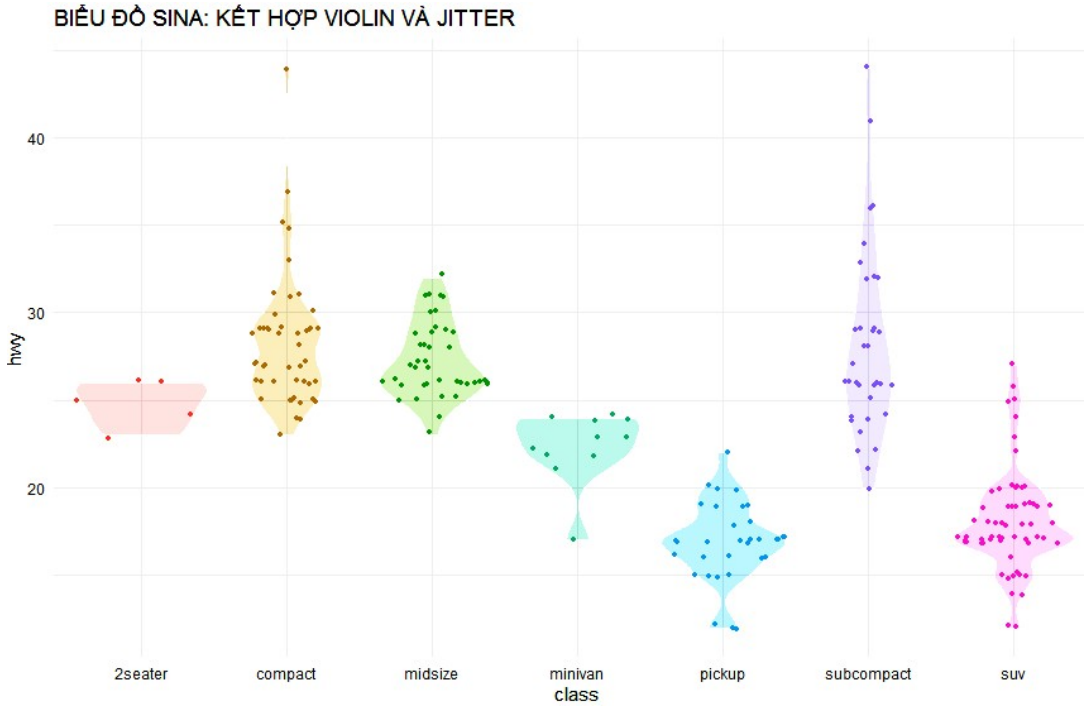
Do biểu đồ violin được vẽ từ ước tính mật độ nên có thể tạo ra sự xuất hiện của dữ liệu không tồn tại, hoặc thể hiện dữ liệu dày đặc trong khi thực tế khá thưa thớt. Do đó, ta vẽ động thời cả biểu đồ violin và jitter để có cách nhìn tổng quát hơn về dữ liệu. Hơn thế nữa, ta có thể sử dụng biểu đồ sina (sina plots), là sự kết hợp của biểu đồ violin và Jitter, khi đó sẽ hiển thị tất cả các điểm dữ liệu trên biểu đồ violin để làm nổi bật sự phân bố của dữ liệu .

```
# install
```

```
ggplot(  data =, aes(y = biến khảo sát, x = biến ảnh hưởng)) +
  geom_violin( aes(fill = biến ảnh hưởng), color = "white", alpha = 0.2)+
  geom_sina(size=1, aes(color = biến ảnh hưởng))+
    theme_minimal() +
  theme(legend.position = "none") +
  labs(title = "")
```

Ví dụ 1.55. Xét file dữ liệu về ô tô mpg ta sẽ vẽ biểu đồ sina cho mức tiêu thụ trên đường cao tốc hwy theo kiểu dáng.

```
ggplot(  data =mpg, aes(y = hwy, x = class)) +
  geom_violin( aes(fill = class), color = "white", alpha = 0.2)+
  geom_sina(size=1, aes(color = class))+
    theme_minimal() +
  theme(legend.position = "none") +
  labs(title = "BIỂU ĐỒ SINA: KẾT HỢP VIOLIN VÀ JITTER")
```



Hình 1.57: Biểu đồ sina.

- ***Biểu đồ ridgeline (ridgeline plots)***

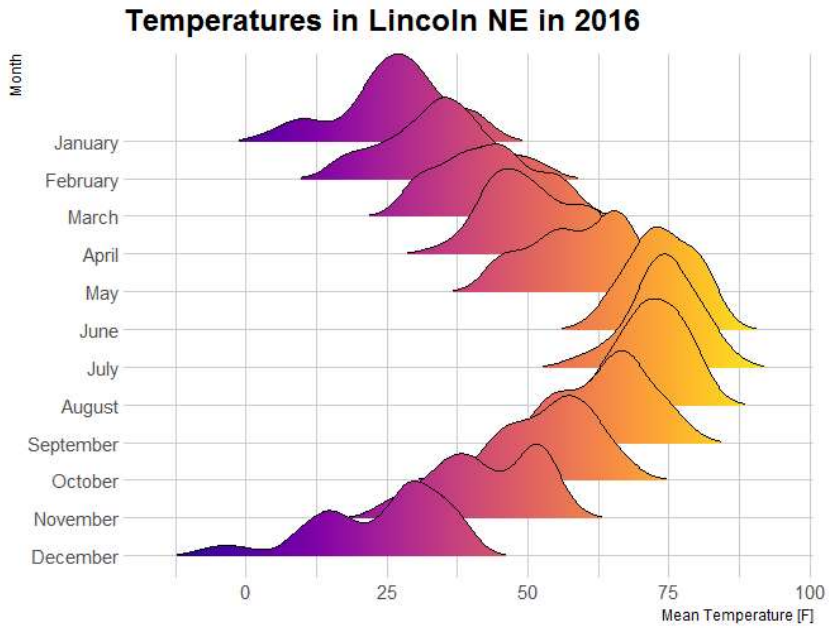
Thay vì biểu diễn các phân phối qua các tháng theo trục tung như biểu đồ hộp hay biểu đồ violin, biểu đồ ridgeline sử dụng các biểu đồ mật độ xếp so le với nhau để biểu diễn các phân phối theo hướng thẳng đứng trên trục hoành. Mục đích của biểu đồ ridgeline không phải để hiển thị các giá trị mật độ cụ thể mà là để dễ dàng so sánh các hình dạng mật độ và chiều cao tương đối giữa các nhóm. Các biểu đồ Ridgeline có thể mở rộng quy mô đến số lượng rất lớn các phân bố.

```
ggplot(data, aes(x = biến khảo sát, y = nhân tố liên quan, fill = stat(x))) +
  geom_density_ridges_gradient(scale = 3, rel_min_height = 0.01) +
```

```
  scale_fill_viridis_c(name =", option = "C") +
  labs(title = ')
```

Ví dụ 1.56. Xét file dữ liệu về nhiệt độ trong năm tại file dữ liệu lincoln_weather (tích hợp trong thư viện ggridge).

```
# Library
library(ggridges)
library(ggplot2)
library(viridis)
library(hrbrthemes)
# Plot
ggplot(lincoln_weather, aes(x = `Mean Temperature [F]`, y = `Month`, fill = ..x..)) +
  geom_density_ridges_gradient(scale = 3, rel_min_height = 0.01) +
  scale_fill_viridis(name = "Temp. [F]", option = "C") +
  labs(title = 'Temperatures in Lincoln NE in 2016') +
  theme_ipsum() +
    theme(
      legend.position="none",
      panel.spacing = unit(0.1, "lines"),
      strip.text.x = element_text(size = 8)
    )
```



Hình 1.58: Biểu đồ ridgeline.

Trong Hình 1.57, cho thấy sự phân bố nhiệt độ các tháng trong năm 2016 tại Lincoln NE, Có thể thấy, trong những tháng mùa hè nhiệt độ có phân phối xấp xỉ chuẩn, còn những tháng cuối năm và đầu năm thì mật độ có xu hướng lệch phải.

c) Biểu đồ mật độ kết hợp phân phối biên

Biểu đồ hiển thị các phân phối trên các cạnh của biểu đồ tán xạ với hàm geom_point(),với hàm ggMarginal() trong gói thư viện package ggExtra .

. Sau đây là những đối số chính:

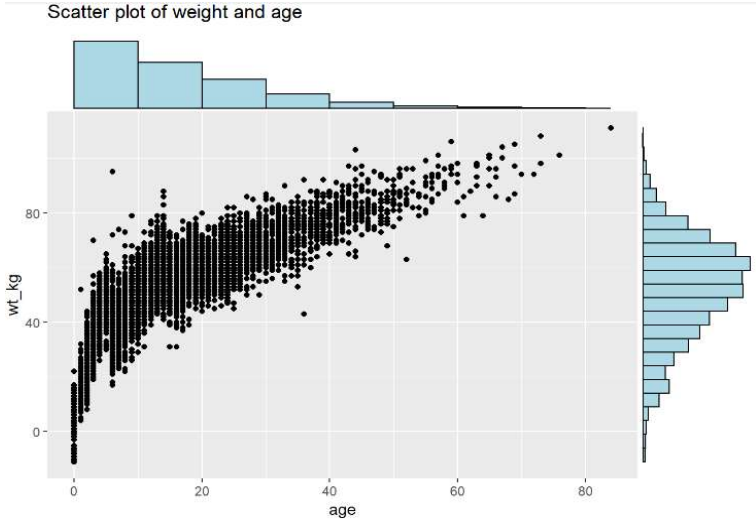
```
# Tạo đồ thị tán xá
scatter_plot <- ggplot(data = )+
```

```
  geom_point(mapping = aes(y = , x =)) +
  labs(title = )
# kết hợp thêm biểu đồ phân phối
ggMarginal(scatter_plot,type = "",fill = "lightblue",
  xparams = list(binwidth = 10),
  yparams = list(binwidth = 5))
```

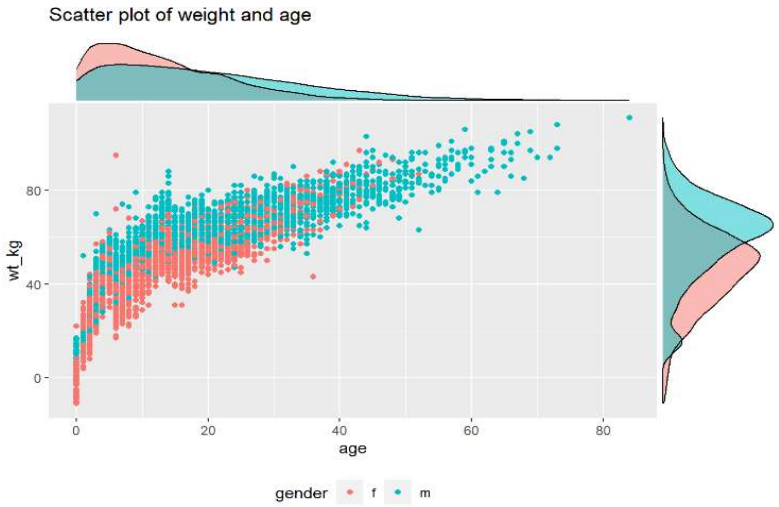
- type = theo một số lựa chọn sau: “histogram”, “density” “boxplot”, “violin”, hoặc “densigram”.
- Theo mặc định, các đồ thị biên sẽ xuất hiện ở cả hai trục. Bạn có thể thiết lập margins = thành “x” hoặc “y” nếu ta chỉ muốn hiện thị ở một trong số chúng.
- Các đối số tùy chọn khác bao gồm fill = (màu cột), color = (màu đường), size = (kích thương biểu đồ so với kích thước biên, do đó số lớn hơn làm biểu đồ biên nhỏ hơn).

Ví dụ 1.57. Xét file dữ liệu linelist về chim cánh cụt tại nam cực, ta xét các đồ thị kết hợp với dữ liệu về tuổi (age) và cân nặng wt_kg.

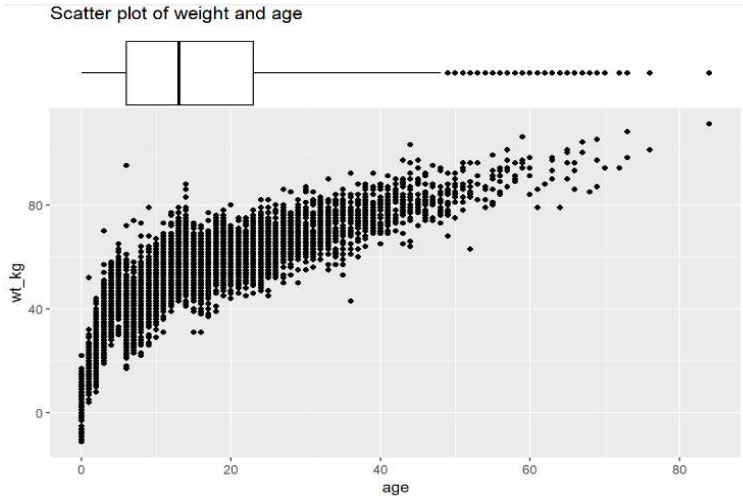
```
scatter_plot <- ggplot(data = linelist)+
  geom_point(mapping = aes(y = wt_kg, x = age)) +
  labs(title = "Scatter plot of weight and age")
# kết hợp hàm phân bố
ggMarginal(scatter_plot,type = "histogram", fill = "lightblue",
                        xparams = list(binwidth = 10),
                         yparams = list(binwidth = 5))
# kết hợp hàm mật độ
ggMarginal(scatter_plot_color, type = "density", groupFill = TRUE)
# kết hợp biểu đồ hộp
ggMarginal(scatter_plot, margins = "x", type = "boxplot")
```



Hình 1. 59: Biểu đồ kết hợp: phân tán và phân bố .



Hình 1. 60: Biểu đồ kết hợp: phân tán và mật độ.



Hình 1.61: Biểu đồ kết hợp: phân tán và hộp.

1.3.4. Biểu đồ thể hiện sự tương quan của dữ liệu trong trực quan hóa

Với bộ dữ liệu có từ hai biến định lượng trở lên, ta có thể hình dung mối tương quan của các biến thông qua hình ảnh trực quan từ các dạng biểu đồ phân tán (scatter plots), biểu đồ bong bóng (bubble chart), ma trận phân tán (scatterplot matrix) hay biểu đồ tương quan (correlogram).

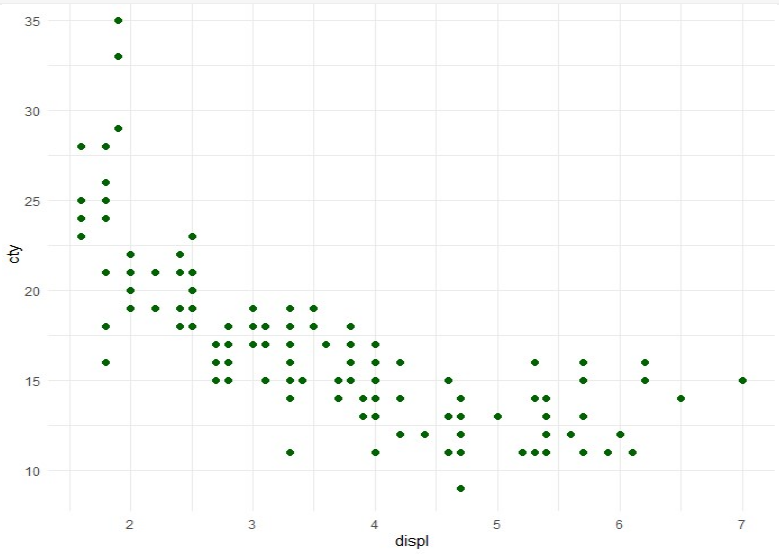
a) Biểu đồ phân tán (scatter plots)

Biểu đồ phân tán *(scatter plots hay scatter diagram)* là một loại biểu đồ sử dụng tọa độ Descartes để hiển thị giá trị và mối quan hệ **giữa hai biến định lượng** cho một tập dữ liệu. Dữ liệu được hiển thị dưới dạng tập hợp các điểm, mỗi điểm có giá trị của một biến xác định vị trí trên trục hoành và giá trị của biến khác xác định vị trí trên trục tung.

```
gplot(data =,
      mapping = aes(y = , x = ))+
 geom_point() +
 labs(title = )
```

Ví dụ 1.58. Xét dữ liệu được thiết lập là các trường hợp trong bộ mpg, với 2 biến liên tục là mức tiêu hao thụ liệu khi di chuyển trong đô thị city và dung tích động cơ displ

```
ggplot(data = mpg, mapping = aes(x = displ, y = cty))+
  geom_point(color = "darkgreen", size = 2, alpha = 1)+
theme_minimal()
```

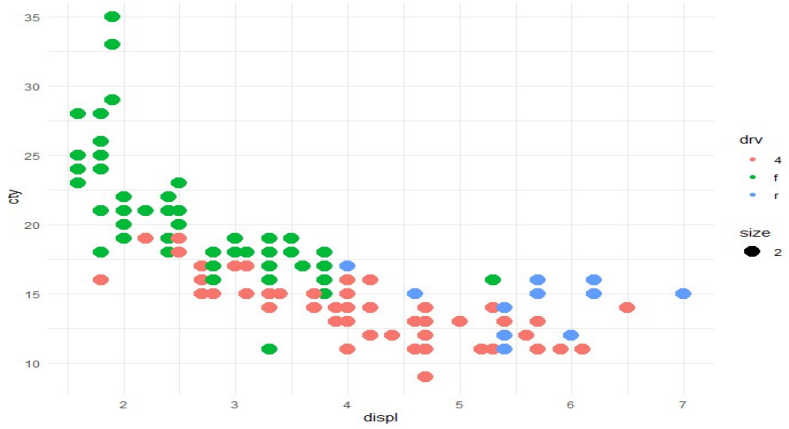


Hình 1.62: Biểu đồ phân tán.

Nhận xét 6: Biểu đồ phân tán có thể được dùng khi 2 biến định lượng độc lập hoặc có một biến phụ thuộc vào biến còn lại. Trong trường hợp phụ thuộc, biến phụ thuộc (biến được dự đoán) thường được vẽ dọc theo trục tung, biến độc lập (biến dùng để đưa ra dự đoán) được vẽ dọc theo trục hoành. Trong trường hợp cả 2 biến độc lập với nhau, mỗi biến được vẽ ở một trục. Biểu đồ phân tán sẽ chỉ minh họa mức độ tương quan (không phải quan hệ nhân quả) giữa hai biến.

Ví dụ 1.59. Ta xét biểu đồ phân tán ở ví dụ 1.57 và có kết hợp thêm yếu tố phân loại hướng dẫn động bởi màu sắc.

```
ggplot(data = mpg, mapping = aes(x = displ, y = cty,color=drv,size=2))+
  geom_point()+
theme_minimal()
```



Hình 1.63: Biểu đồ phân tán kết hợp với yếu tố phân loại qua màu sắc.

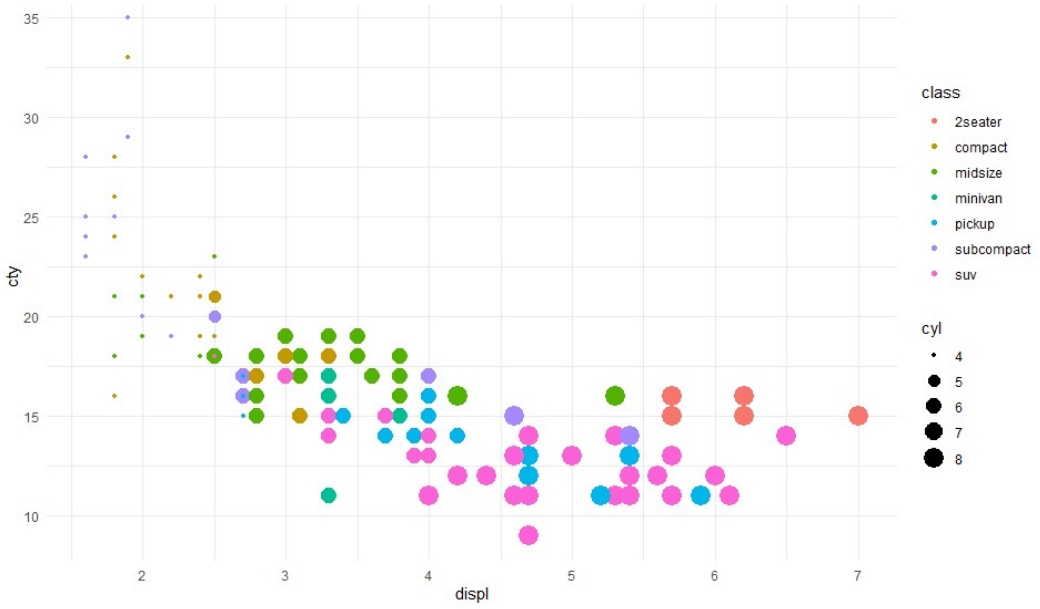
b) Biểu đồ bong bóng (bubble chart)

Biểu đồ bong bóng *(bubble chart)* có thể xem là một biến thể của biểu đồ phân tán, nhưng có thêm giá trị định lượng thứ ba. Trong đó, 2 giá trị được biểu diễn thông qua vị trí trục tung và trục hoành, giá trị còn lại được biểu diễn bằng kích thước của điểm dữ liệu.

```
ggplot(data = mpg, mapping = aes(x =nhân tố 1, y =nhân tố 2 ,color=,size=nhân tố 3))+
  geom_point()+
theme_minimal()
```

Ví dụ 1.60. Ta xét biểu đồ phân tán ở ví dụ 1.58 và có kết hợp thêm yếu tố phân loại kiểu dáng xe thông qua màu sắc color=class và kích cỡ cylanh size=cyl

```
ggplot(data = mpg, mapping = aes(x = displ, y = cty,color=class,size=cyl))+
  geom_point()+
theme_minimal()
```



Hình 1.64: Biểu đồ bong bóng.

Trong biểu đồ trên, ta chỉ có thể hình dung được những thông tin chung nhất về mối liên hệ giữa dung tích và mức tiêu thụ nhiên liệu khi di chuyển trong đô thị theo các kiểu dáng xe và sự chênh lệch về số cylanh thông qua độ lớn các bong bóng.

Nhận xét 7: Qua các ví dụ trên, ta thấy mặc dù sẽ là lý tưởng để biển diễn nhiều biến cùng lúc trên một biểu đồ duy nhất, nhưng trong trường hợp lượng dữ liệu quá lớn hay sự chênh lệnh của bong bóng quá nhỏ, người phân tích phải mất thời gian để giải thích tất cả cấu trúc của biểu đồ rồi mới có thể đưa ra kết luận cuối cùng. Ngoài ra, rất khó để xác định mối quan hệ giữa các biến trên trục và kích thước bong bóng. Do đó, để dễ giải thích kết quả chi tiết hơn theo định hướng bài phân tích, ta có thể bổ sung thêm các dạng biểu đồ khác, hoặc phân tích tương quan cụ thể giữa các biến định lượng khi sử dụng ma trận biểu đồ phân tán (scatterplot matrix).

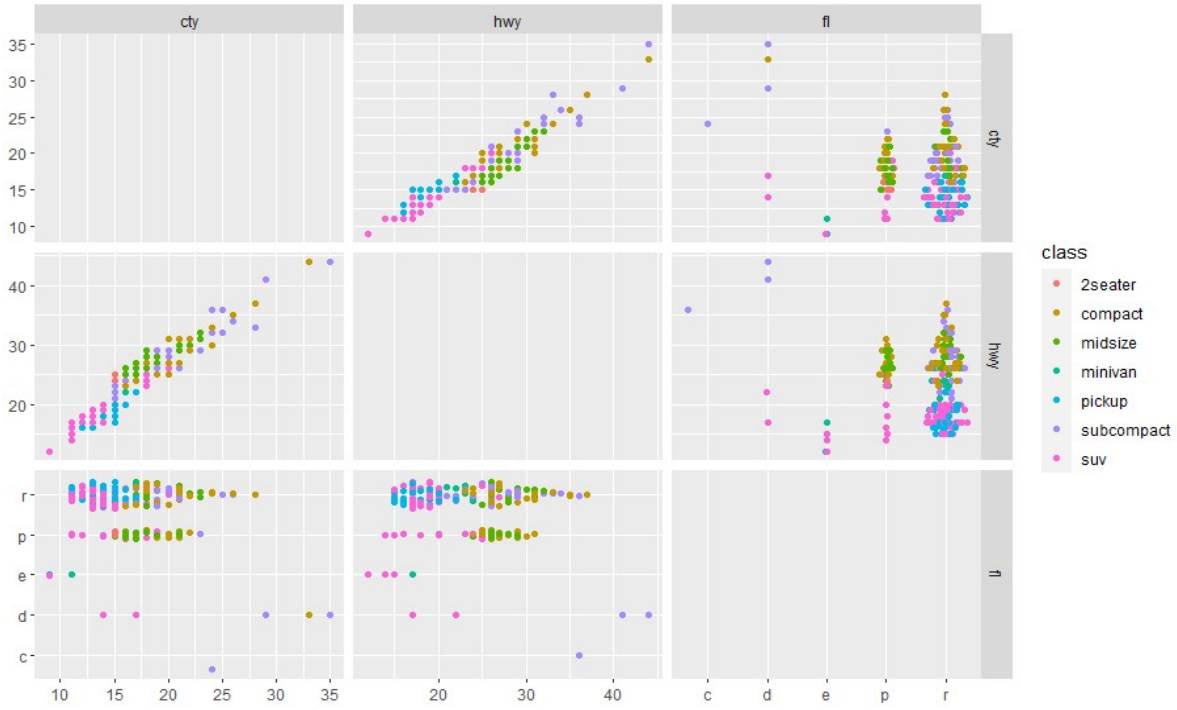
c) Ma trận biểu đồ phân tán (scatterplot matrix)

Ma trận biểu đồ phân tán là một tập hợp các biểu đồ phân tán cho biết các biến định lượng trong bộ dữ liệu có liên quan như thế nào với nhau. Sau khi biểu diễn tất cả các kết hợp hai chiều của các biến, ma trận có thể hiển thị mối quan hệ giữa các biến để làm nổi bật mối quan hệ nào có thể là quan trọng.

```
ggplot(data) +
  geom_autopoint(aes(color=biến  phân loại)) +
  facet_matrix(vars(nhân tố 1:nhân tố 2), layer.diag = 2, grid.y.diag = FALSE)
```

Ví dụ 1.61. Ta xét dữ liệu mpg và lập biểu đồ phân tán cho dữ liệu theo hai nhân tố: mức tiêu hao trong đô thị và loại nhiên liệu dựa trên biến phân loại kiểu dáng xe

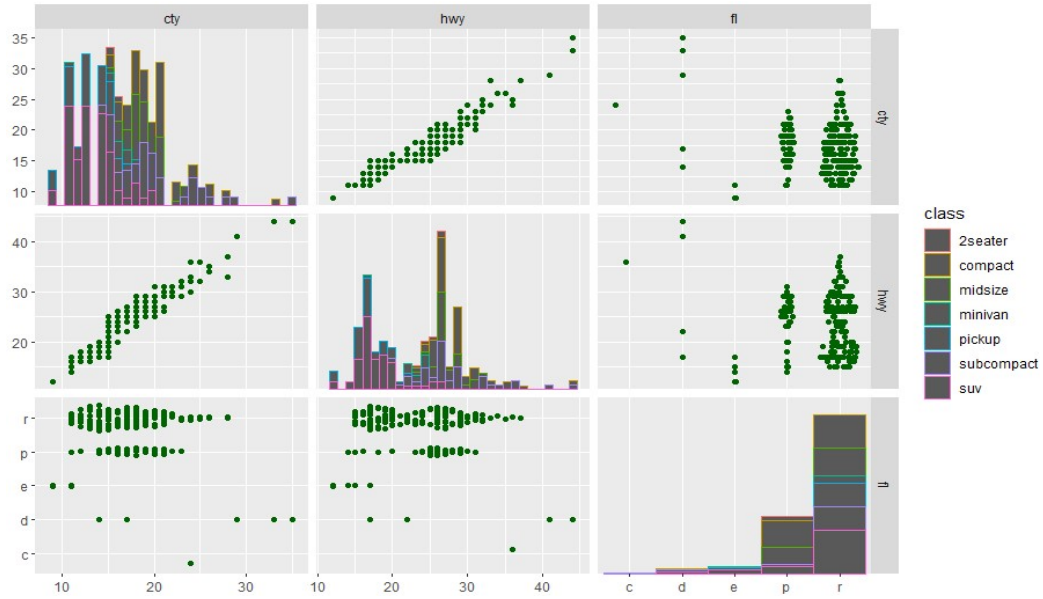
```
ggplot(data = mpg) +
  geom_autopoint(aes(color=class)) +
  facet_matrix(vars(cty:fl), layer.diag = 2, grid.y.diag = FALSE)
```



Hình 1.65: Biểu đồ ma trận phân tán.

Ví dụ 1.62. Ta có thể kết hợp thêm biểu đồ phân bố trên cùng biểu đồ phân tán của Ví dụ 1.60 như sau:

```
ggplot(data = mpg) +
  geom_autopoint(col="darkgreen") +
  facet_matrix(vars(cty:fl), layer.diag = 2, grid.y.diag = FALSE)+
geom_autohistogram(aes(color=class))
```



Hình 1.66: Biểu đồ phân tán kết hợp với phân bố.

Ví dụ 1.63. Hoặc kết hợp thêm biểu đồ mật độ trên cùng biểu đồ phân tán của Ví dụ 1.60 như sau:

```
ggplot(data = mpg)+
  geom_autopoint(col="darkgreen") +
  facet_matrix(vars(cty:fl), layer.diag = 2, grid.y.diag = FALSE)+
geom_autodensity(aes(colour = drv, fill = drv)),
alpha = 0.4)
```

Hình 1.67: Biểu đồ phân tán kết hợp biểu đồ mật độ.

d) Biểu đồ tương quan (Correlogram)

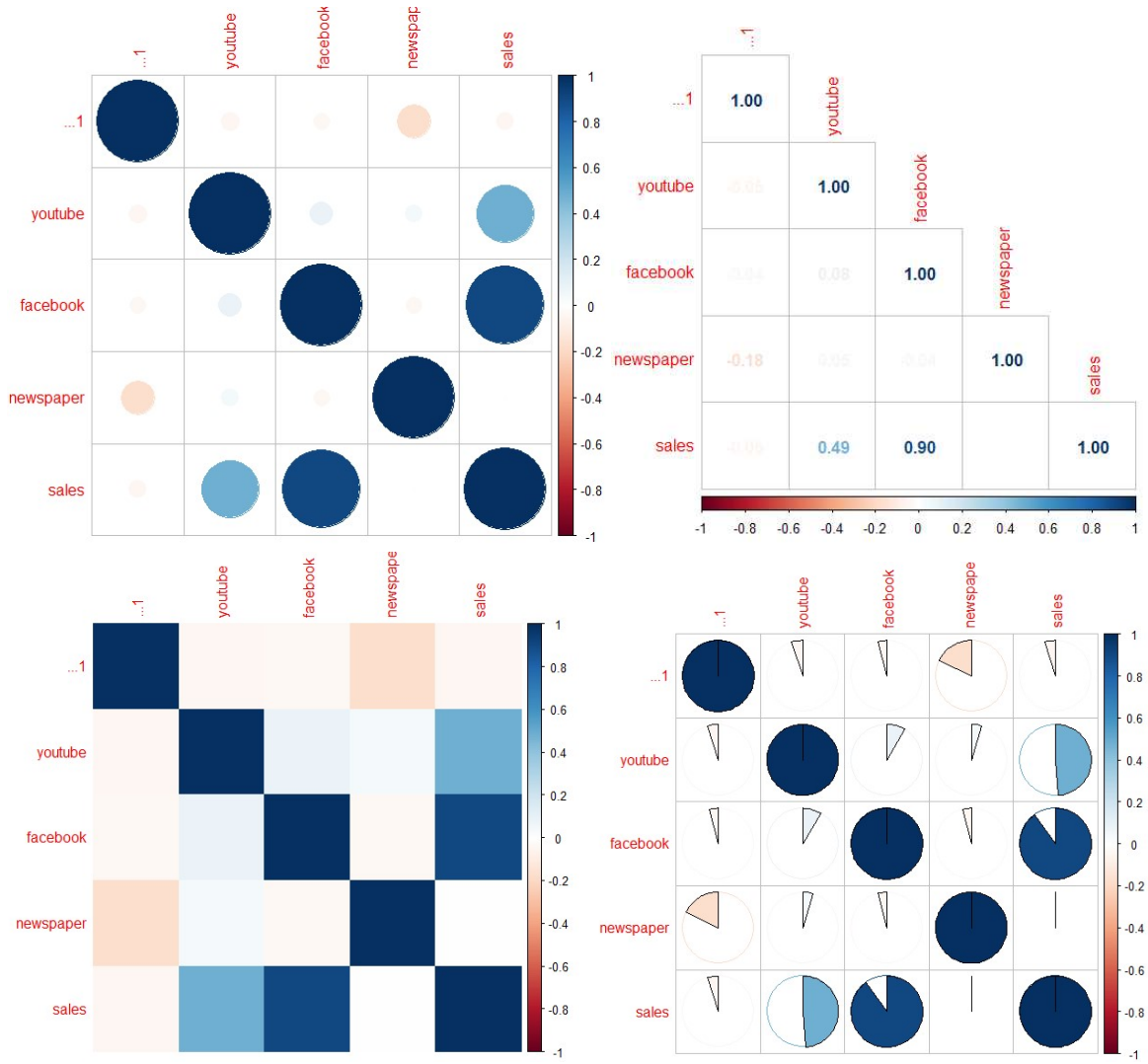
- ***Ma trận biểu đồ tương quan***

Trong trường hợp có từ 3 đến 4 biến định lượng trở lên, thay vì biểu diễn trực quan tất cả dữ liệu của các biến định lượng lên biểu đồ, ta có thể tính toán hệ số tương quan Pearson giữa từng cặp biến và trực quan số liệu này lên biểu đồ.

```
raqMatrix <- cor(data)
corrplot(raqMatrix,
  method =c("circle", "square", "ellipse", "number", "shade", "color", "pie"),
   type = c("full", "lower", "upper"),
    title = "")
```

Ví dụ 1.64. Sử dụng file dữ liệu marketting.csv ở mục 1.2, chúng ta lập ma trận biểu đồ tương quan.

```
Library(corrplot)
raqMatrix <- cor(marketting)
# ma trận tương quan số
corrplot(raqMatrix,
  method ="circle",
   type = full,
    title = " BIỂU ĐỒ TRÒN")
# ma trận tương quan tròn
corrplot(raqMatrix,
  method = "number",
   type = "lower",
    title = "BIỂU ĐỒ SỐ")
# ma trận tương quan màu sắc
corrplot(raqMatrix,
  method ="color",
   type = "full",
    title = "")
# ma trận tương quan bánh tròn
corrplot(raqMatrix,
  method ="pie",
   type = "full",
    title = "")
```



Hình 1.68: Biểu đồ ma trận tương quan.

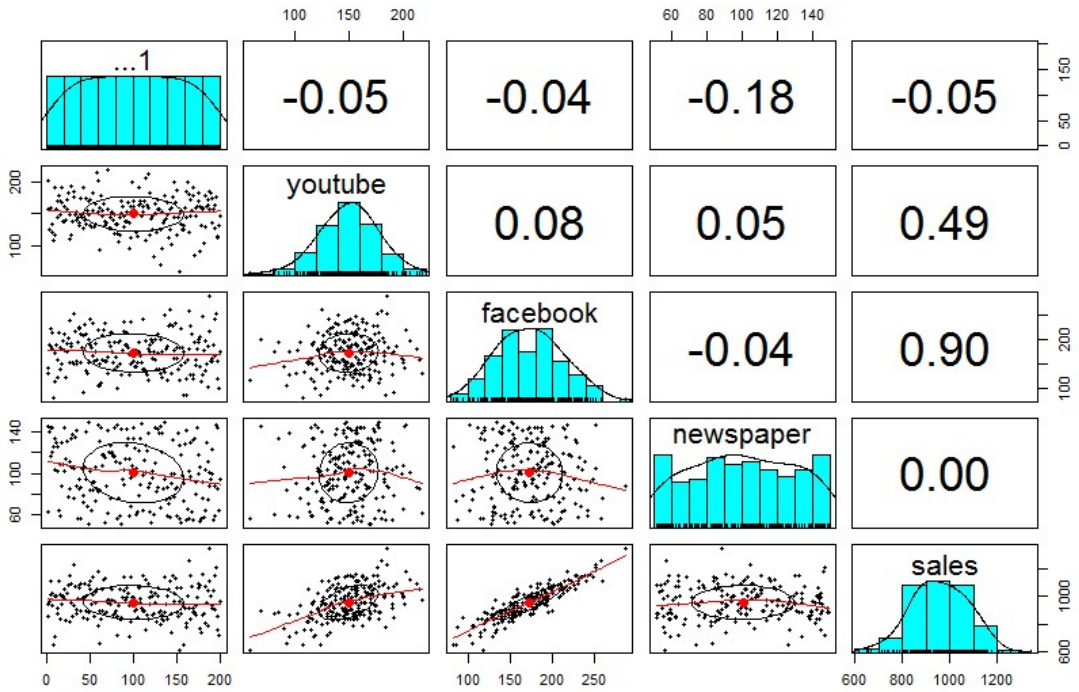
Theo dải màu, các hình vuông hoặc hình tròn đậm cho thấy sự tương quan cao, các ô nhạt cho thấy sự tương quan thấp. Tuy nhiên, cách biểu diễn này chưa tối ưu do các tương quan thấp không bị triệt tiêu một cách trực quan, vì thế chúng ta cần kết hợp với biểu đồ giá trị số để có kết quả chi tiết nhất.

- ***Biểu đồ tương quan kết hợp luật phân phối***

```
library(psych)
pairs.panels(data)
```

Ví dụ 1.65. Sử dụng file dữ liệu marketting.csv ở mục 1.2, chúng ta lập ma trận biểu đồ tương quan kết hợp luật phân phối như sau:

```
  library(psych)
pairs.panels(marketting, color="darkgreen")
```



Hình 1.69: Biểu đồ ma trận tương quân kết hợp phân phối.

Thông tin trên biểu đồ với ma trận tương quan có thể chưa đầy đủ với đường dữ đoán dữ liệu. Biểu đồ ma trân tương quan kết hợp phân phối cho cái nhìn tổng quát và chi tiết hơn, kết hợp với đường dữ đoán tuyến tính đi qua tâm miền dữ liệu.

1.3.5. Các bước thực hiện trực quan hóa dữ liệu

Tất cả các dạng biểu đồ được giới thiệu ở trên đều có những ưu, nhược điểm riêng khi sử dụng biểu diễn trực quan dữ liệu. Để có thể sử dụng cho nội dung trực quan thì không có dạng biểu đồ nào là tối ưu nhất, thay vào đó chúng ta cần nắm rõ nội dung muốn diễn giải để áp dụng đúng dạng biểu đồ, để mang đến thông tin hữu ích và dễ hiểu nhất. Việc này dựa trên các thành phần chính trong trực quan dữ liệu và một số bước cần lưu ý như sau:

a) Các thành phần chính

Trực quan hóa dữ liệu được hình thành từ 3 thành phần chính sau đây:

- Thông điệp: Giới thiệu mục đích của việc trực quan hóa số liệu, người phân tích sẽ làm việc và quyết định kết quả mong muốn đạt được sau khi phân tích dữ liệu. Ví dụ: Dự đoán doanh thu bán hàng hay đo lường hiệu suất làm việc của công – nhân viên.
- Dữ liệu: Sau khi xác định thông điệp, các nhà phân tích tiến hành xử lý dữ liệu (chỉnh sửa định dạng, làm sạch dữ liệu, loại bỏ thông tin không liên quan và phân tích kỹ lưỡng hơn). Sau đó, các phương thức trình bày dữ liệu trực quan sẽ được sử dụng giúp bộ phận quản lý lên kế hoạch phù hợp.
- Phương tiện trực quan: Phương pháp trực quan là thành phần tiếp theo. Các nhà khoa học dữ liệu tạo ra các biểu đồ hoặc đồ thị để làm sinh động dữ liệu chính, đơn giản dữ liệu phức tạp nhằm chia sẻ thông tin chuyên sâu nhất. Các phương thức

được cân nhắc sử dụng để thông tin được liền mạch và dễ hiểu một cách có hệ thống.

b) Các bước trực quan hóa dữ liệu

Quá trình thực hiện trực quan hóa dữ liệu bao gồm các bước.

- ***Bước xác định mục tiêu:*** Xác định các vấn đề mà tập dữ liệu bạn đã tổng hợp có thể trả lời. Một số mục tiêu cụ thể có thể giúp doanh nghiệp:
  - Phân loại dữ liệu doanh nghiệp đang sở hữu.
  - Phân tích dữ liệu đó.
  - Có được một số hoặc tất cả các phương thức trực quan dữ liệu có thể sử dụng để trình bày dữ liệu chuyên sâu nhất.
- ***Bước thu thập dữ liệu:*** Tổng hợp dữ liệu nội bộ và bên ngoài liên quan đến mục tiêu xác định. Dữ liệu này có thể là các tập thông tin được bán trực tuyến; thông tin sẵn có trong kho lưu trữ dữ liệu; thông tin tự đi điều tra-phỏng vấn; thông tin do khách hàng-doanh nghiệp cung cấp.
- ***Bước chọn lọc, làm sạch dữ liệu:*** Loại bỏ dữ liệu thừa, không liên quan, thực hiện các phép tính toán để phân tích và chuyển đổi loại dữ liệu để có thể sử dụng.
- ***Bước lựa chọn phương tiện trực quan hóa:*** Có rất nhiều loại biểu đồ giúp trình bày dữ liệu hiệu quả. Người trình bày có thể dựa vào mối quan hệ giữa các điểm dữ liệu và thông tin muốn thể hiện để chọn. Ví dụ: Biểu đồ cột để biểu diễn doanh thu bán hàng; biểu đồ đường phù hợp khi so sánh thông tin; biểu đồ tròn thích hợp với dữ liệu dưới dạng tỷ lệ.

c) Các loại hình trực quan

- Trực quan hóa tĩnh: Là loại hình chỉ cung cấp một chế độ xem duy nhất cho mỗi loại dữ liệu, chẳng hạn: đồ họa thông tin.
- Trực quan hóa tương tác: Là loại hình cho phép người dùng tương tác phương thức trực quan hóa. Người xem có thể thay đổi số liệu để tìm thông tin chuyên sâu hơn hoặc truy cập một loại thông tin khác.