Trình bày và mô tả dữ liệu

Khoa Toán - Cơ - Tin học
Trường Đại học Khoa học Tự nhiên
Đại học Quốc gia Hà Nội

Ngày 30 tháng 7 năm 2022

Trình bày và mô tả dữ liệu bằng bảng và biểu đồ

Ví dụ: Giả sử điểm thi của 100 sinh viên theo thang điểm 100 được cho như sau:

76 13 77 62 88 97 30 77 30 21 65 95 10 36 51 39 16 91 3 93
66 61 17 9 89 25 1 29 2 93 92 77 10 97 7 43 89 16 76 73
64 4 96 10 41 18 24 41 75 74 71 90 77 80 86 36 92 75 29 32
32 58 51 46 46 94 43 54 77 71 16 14 84 71 82 4 32 52 84 6
56 83 53 62 74 16 35 26 61 3 50 98 59 35 43 99 43 46 84 80

Trình bày và mô tả dữ liệu bằng bảng và biểu đồ

Vấn đề: Nếu chỉ biểu diễn dữ liệu dưới dạng các con số lộn xộn, ta sẽ không thu được thông tin gì. Vì vậy, ta cần biểu diễn chúng dưới các dạng gọn nhẹ hơn, trực quan hơn để thấy được các xu thế, tính chất, ...

Trình bày và mô tả dữ liệu bằng bảng và biểu đồ

Vấn đề: Nếu chỉ biểu diễn dữ liệu dưới dạng các con số lộn xộn, ta sẽ không thu được thông tin gì. Vì vậy, ta cần biểu diễn chúng dưới các dạng gọn nhẹ hơn, trực quan hơn để thấy được các xu thế, tính chất, ...
Các phương pháp biểu diễn:

- Sử dụng bảng phân bố
- Sử dụng biểu đồ.

Bảng phân bố tần số: là bảng dùng để biểu diễn một tập dữ liệu cho biết các giá trị xuất hiện và số lần xuất hiện mỗi giá trị đó

Bảng phân bố tần số: là bảng dùng để biểu diễn một tập dữ liệu cho biết các giá trị xuất hiện và số lần xuất hiện mỗi giá trị đó

X	x_1	x_2	...	x_m
Tần số	r_1	r_2	...	r_m

Bảng phân bố tần số: là bảng dùng để biểu diễn một tập dữ liệu cho biết các giá trị xuất hiện và số lần xuất hiện mỗi giá trị đó

X	x_1	x_2	...	x_m
Tần số	r_1	r_2	...	r_m

Người ta thường đưa thêm các thông tin về tần suất, tần suất tích lũy để có bảng phân bố thực nghiệm

Bảng phân bố tần số: là bảng dùng để biểu diễn một tập dữ liệu cho biết các giá trị xuất hiện và số lần xuất hiện mỗi giá trị đó

X	x_1	x_2	...	x_m
Tần số	r_1	r_2	...	r_m

Người ta thường đưa thêm các thông tin về tần suất, tần suất tích lũy để có bảng phân bố thực nghiệm

Bảng phân bố tần số ghép lớp: là bảng biểu diễn dữ liệu dạng khoảng, dữ liệu định tính, ... theo đó các giá trị được ghép thành các lớp (các nhóm) cùng với số giá trị thuộc mỗi lớp (tần số).

Bảng phân bố tần số: là bảng dùng để biểu diễn một tập dữ liệu cho biết các giá trị xuất hiện và số lần xuất hiện mỗi giá trị đó

X	x_1	x_2	...	x_m
Tần số	r_1	r_2	...	r_m

Người ta thường đưa thêm các thông tin về tần suất, tần suất tích lũy để có bảng phân bố thực nghiệm

Bảng phân bố tần số ghép lớp: là bảng biểu diễn dữ liệu dạng khoảng, dữ liệu định tính, ... theo đó các giá trị được ghép thành các lớp (các nhóm) cùng với số giá trị thuộc mỗi lớp (tần số).

Tuổi	5 - 9,5	9,5 - 19,5	19,5 - 34,5	34,5 - 54,5
Số người	440	480	630	440

Ngoài ra, chúng ta còn có thể gặp bảng phân bố tần số với nhiều hơn một biến.

Ngoài ra, chúng ta còn có thể gặp bảng phân bố tần số với nhiều hơn một biến.
Ví dụ: Ở các cây ngọc trâm lá có hai dạng "Lá phẳng" hoặc "Lá nhăn", hoa có hai dạng "hoa bình thường" hoặc "Hoa hoàng hậu". Quan sát một mẫu gồm 560 cây ngọc trâm ta thu được kết quả sau:

Hoa / Lá	Bình thường	Hoàng hậu
Phẳng	328	122
Nhăn	77	33

Một số loại biểu đồ thông dụng:

- Biểu đồ hình cột: Dùng để đưa ra so sánh giữa các nhóm.

Một số loại biểu đồ thông dụng:

- Biểu đồ hình cột: Dùng để đưa ra so sánh giữa các nhóm.
- Biểu đồ hình quạt: Dùng để phân tích hoặc so sánh ở mức độ tổng thể

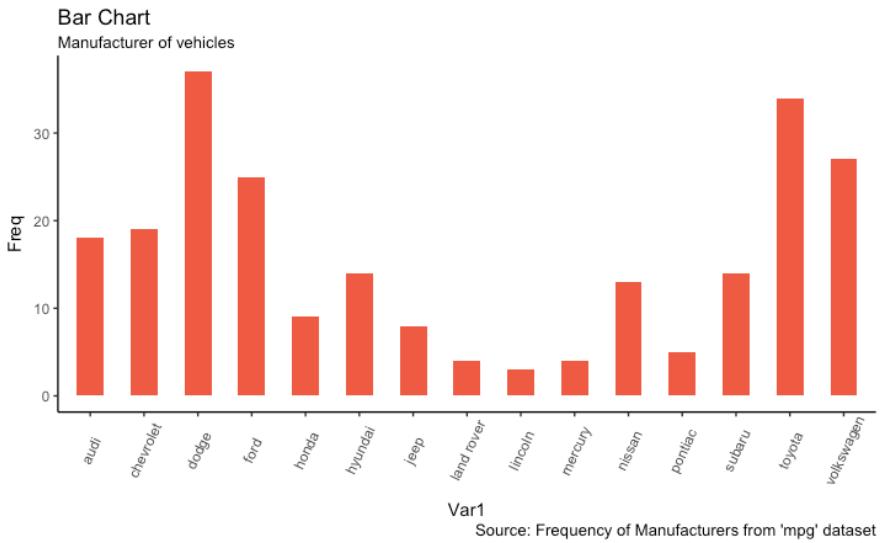
Một số loại biểu đồ thông dụng:

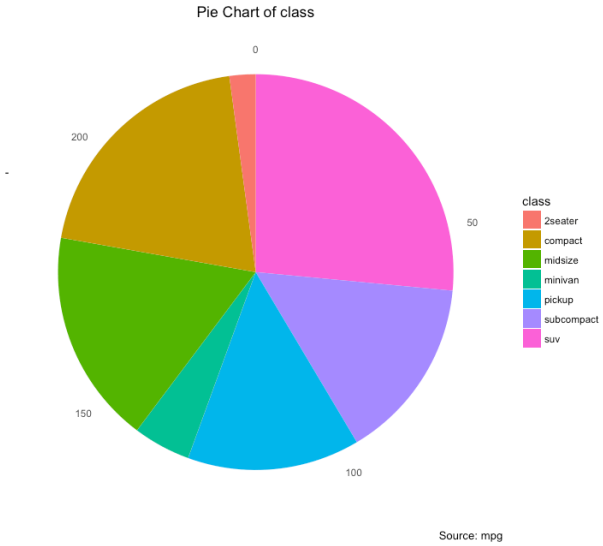
- Biểu đồ hình cột: Dùng để đưa ra so sánh giữa các nhóm.
- Biểu đồ hình quạt: Dùng để phân tích hoặc so sánh ở mức độ tổng thể
- Tổ chức đồ (histogram): Dùng để biểu thị tần số/ tần suất các giá trị trong mỗi khoảng giá trị.

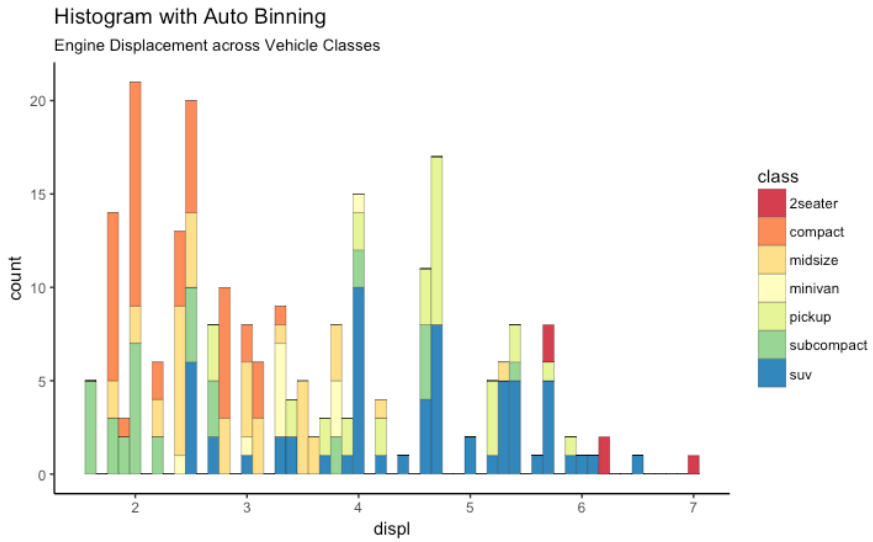
Một số loại biểu đồ thông dụng:

- Biểu đồ hình cột: Dùng để đưa ra so sánh giữa các nhóm.
- Biểu đồ hình quạt: Dùng để phân tích hoặc so sánh ở mức độ tổng thể
- Tổ chức đồ (histogram): Dùng để biểu thị tần số/ tần suất các giá trị trong mỗi khoảng giá trị.
- Biểu đồ tán xạ: Dùng để quan sát mối liên hệ giữa hai biến.

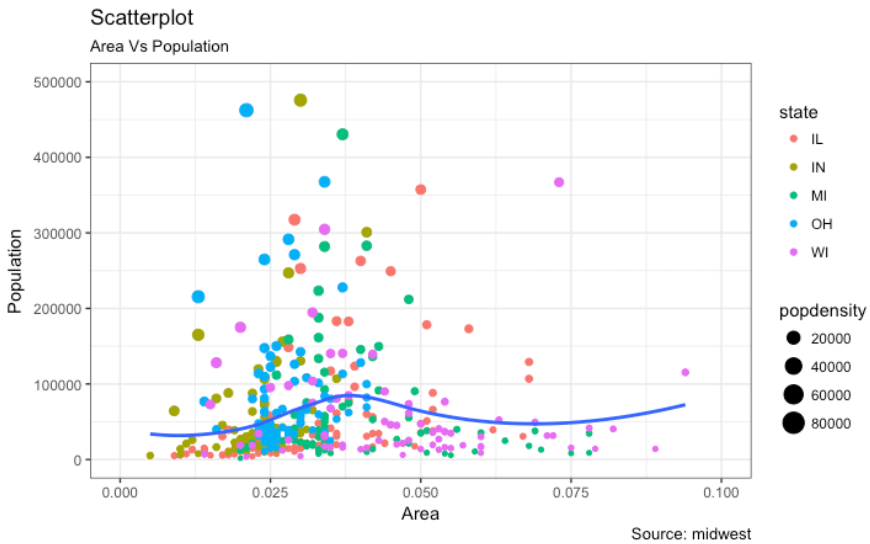
Trình bày và mô tả dữ liệu bằng bảng và biểu đồ

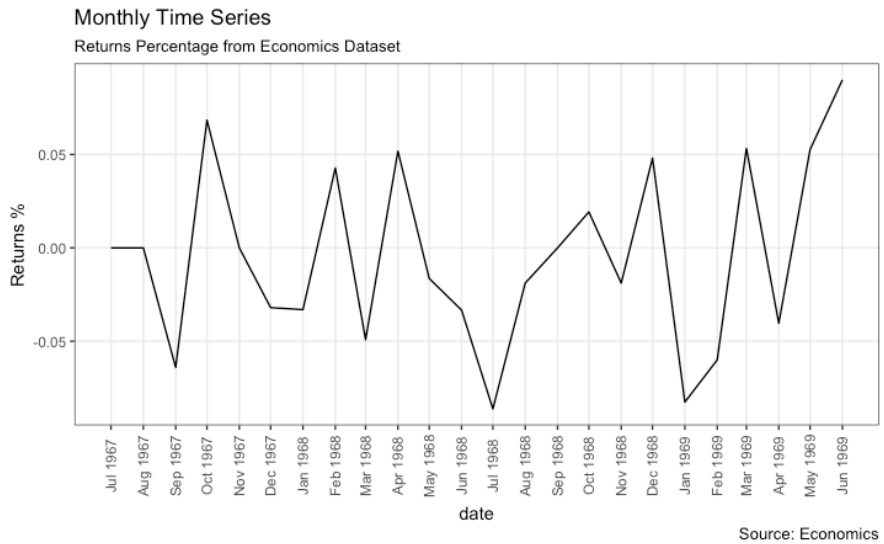






Trình bày và mô tả dữ liệu bằng bảng và biểu đồ





Mô tả dữ liệu bằng các đại lượng đặc trưng

Các số đo xu thế trung tâm

- Trung bình mẫu: Là trung bình cộng các giá trị ta quan sát được.

Mô tả dữ liệu bằng các đại lượng đặc trưng

Các số đo xu thế trung tâm

- Trung bình mẫu: Là trung bình cộng các giá trị ta quan sát được.
- Trung vị mẫu: Là số m có tính chất: Số các giá trị của mẫu bé hơn hay bằng m thì bằng số giá trị của mẫu lớn hơn hay bằng m

Mô tả dữ liệu bằng các đại lượng đặc trưng

Các số đo xu thế trung tâm

- Trung bình mẫu: Là trung bình cộng các giá trị ta quan sát được.
- Trung vị mẫu: Là số m có tính chất: Số các giá trị của mẫu bé hơn hay bằng m thì bằng số giá trị của mẫu lớn hơn hay bằng m
- Mode: Là giá trị trong mẫu xuất hiện với tần số lớn nhất.

Mô tả dữ liệu bằng các đại lượng đặc trưng

Các số đo mức độ phân tán

- Biên độ mẫu: Là hiệu số giữa giá trị lớn nhất và giá trị bé nhất của mẫu.

Mô tả dữ liệu bằng các đại lượng đặc trưng

Các số đo mức độ phân tán

- Biên độ mẫu: Là hiệu số giữa giá trị lớn nhất và giá trị bé nhất của mẫu.
- Khoảng tứ phân vị: Là hiệu số $Q_3 - Q_1$, trong đó Q_1 là giá trị mà 25% số liệu nhỏ hơn Q_1, Q_3 là giá trị mà 75% số liệu nhỏ hơn Q_3

Mô tả dữ liệu bằng các đại lượng đặc trưng

Các số đo mức độ phân tán

- Biên độ mẫu: Là hiệu số giữa giá trị lớn nhất và giá trị bé nhất của mẫu.
- Khoảng tứ phân vị: Là hiệu số $Q_3 - Q_1$, trong đó Q_1 là giá trị mà 25% số liệu nhỏ hơn Q_1, Q_3 là giá trị mà 75% số liệu nhỏ hơn Q_3
- Độ lệch trung bình mẫu: Là giá trị trung bình của của khoảng cách từ mỗi giá trị đến trung bình mẫu.

Mô tả dữ liệu bằng các đại lượng đặc trưng

Các số đo mức độ phân tán

- Biên độ mẫu: Là hiệu số giữa giá trị lớn nhất và giá trị bé nhất của mẫu.
- Khoảng tứ phân vị: Là hiệu số $Q_3 - Q_1$, trong đó Q_1 là giá trị mà 25% số liệu nhỏ hơn Q_1, Q_3 là giá trị mà 75% số liệu nhỏ hơn Q_3
- Độ lệch trung bình mẫu: Là giá trị trung bình của của khoảng cách từ mỗi giá trị đến trung bình mẫu.
- Phương sai: Là giá trị trung bình của bình phương khoảng cách từ mỗi giá trị đến trung bình mẫu.

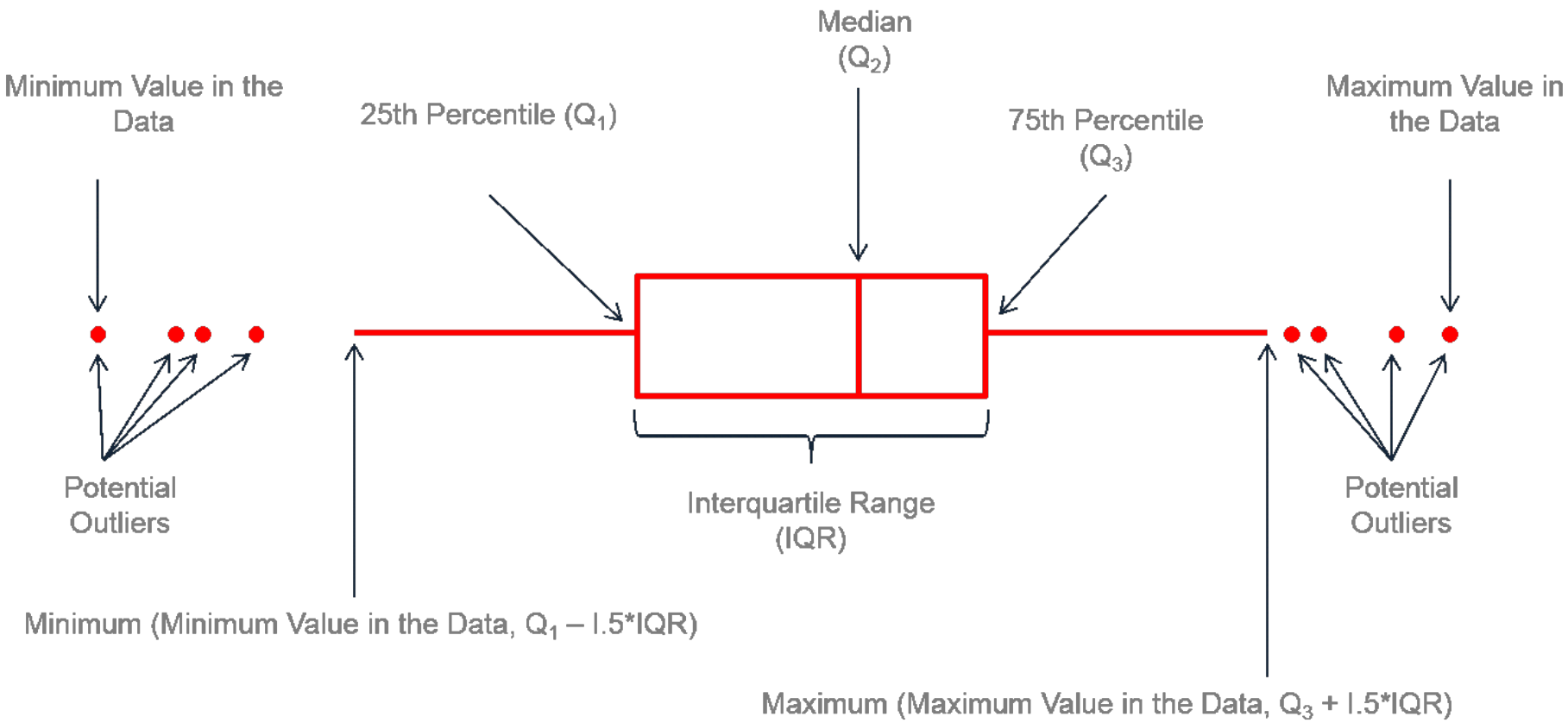
Mô tả dữ liệu bằng các đại lượng đặc trưng

Các số đo mức độ phân tán

- Biên độ mẫu: Là hiệu số giữa giá trị lớn nhất và giá trị bé nhất của mẫu.
- Khoảng tứ phân vị: Là hiệu số $Q_3 - Q_1$, trong đó Q_1 là giá trị mà 25% số liệu nhỏ hơn Q_1, Q_3 là giá trị mà 75% số liệu nhỏ hơn Q_3
- Độ lệch trung bình mẫu: Là giá trị trung bình của của khoảng cách từ mỗi giá trị đến trung bình mẫu.
- Phương sai: Là giá trị trung bình của bình phương khoảng cách từ mỗi giá trị đến trung bình mẫu.
- Độ lệch chuẩn mẫu: là căn bậc hai của phương sai.

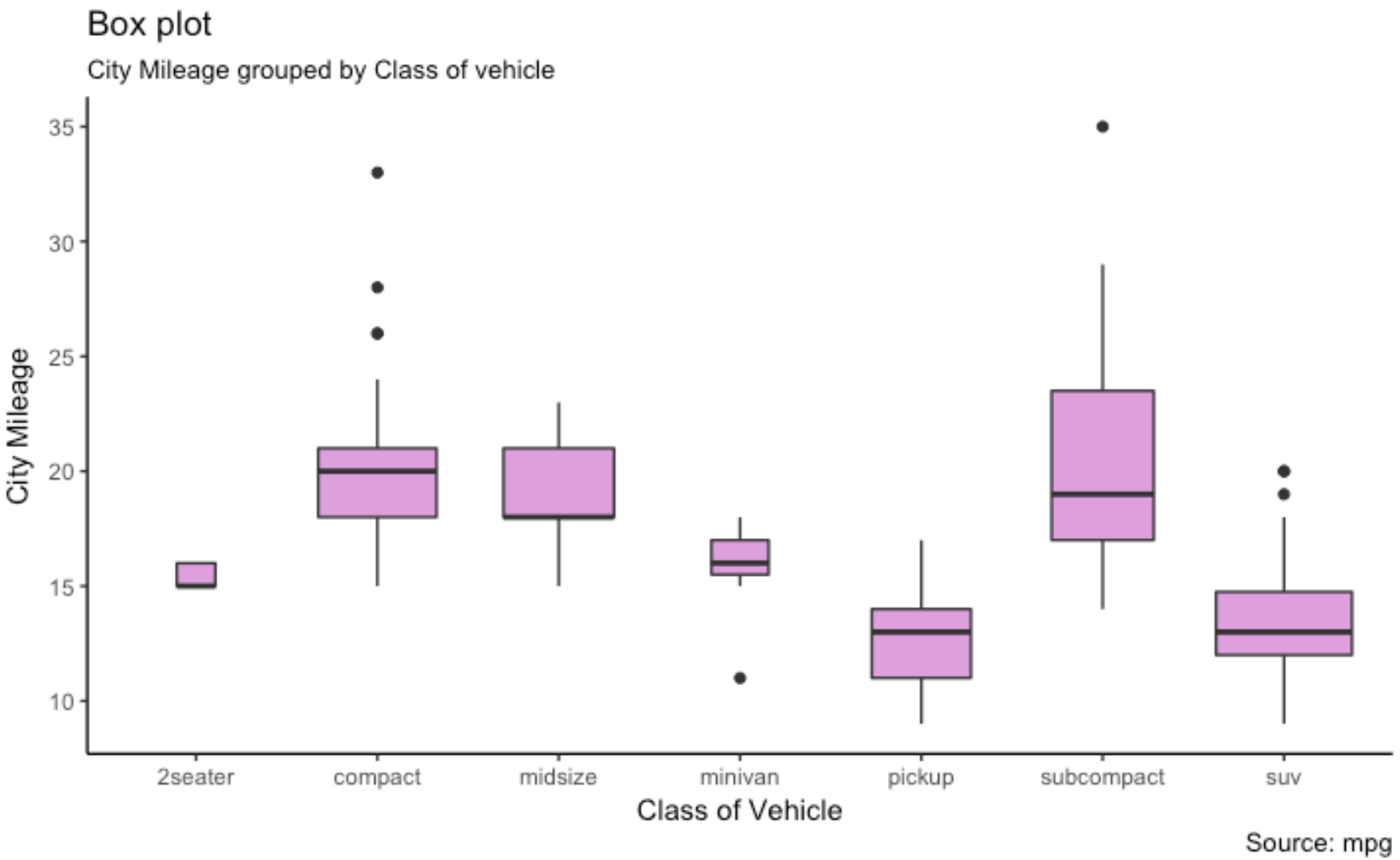
Phát hiện giá trị bất thường (outliers)

Sử dụng biểu đồ hộp



Phát hiện giá trị bất thường (outliers)

Ví dụ sử dụng biểu đồ hộp



Phát hiện giá trị bất thường (outliers)

Sử dụng giá trị trung bình và độ lệch chuẩn

Nếu dữ liệu tuân theo phân bố chuẩn thì:

- 95% số quan sát nằm trong khoảng $(\overline{x} - 2s; \overline{x} + 2s)$
- $99,74\%$ số quan sát sẽ nằm trong khoảng $(\overline{x} - 3s; \overline{x} + 3s)$

Các giá trị không thuộc khoảng trên được coi là giá trị bất thường.