Bài 1. Giả sử rằng dữ liệu để phân tích bao gồm thuộc tính tuổi. Giá trị tuổi là (theo thứ tự tăng dần)

13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

a. Trung bình và trung vị của dữ liệu?

b. Mode của dữ liệu? Nhận xét về loại mode của dữ liệu.

c. Xác định khoảng giữa (midrange) của dữ liệu.

d. Xác định tứ phân vị thứ nhất (Q1) và tứ phân vị thứ 3 (Q3) của dữ liệu.

e. Trực quan hóa dữ liệu bằng boxplot.

Bài 2. Dữ liệu sau đây là danh sách giá của các mặt hàng thường được bán tại của hàng ABC. Các số đã được sắp xếp:

1, 1, 5, 5, 5, 5, 5, 8, 8, 10, 10, 10, 10, 12, 14, 14, 14, 15, 15, 15, 15, 15, 15, 18, 18, 18, 18, 18, 18, 18, 18, 20, 20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 25, 25, 25, 25, 25, 28, 28, 30, 30, 30.

a. Xác định các tham số thống kê về tính trung tâm và tính phân bố

b. Vẽ lược đồ histogram

Bài 3. Cho tập dữ liệu gồm các giá trị như bên dưới:

13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

a. Hãy áp dụng phương pháp chia giỏ để chia dữ liệu thành **3 giỏ** bằng hai phương pháp:
   - Chia giỏ theo độ rộng
   - Chia giỏ theo độ sâu

b. Áp dụng làm trơn bằng giá trị trung bình cho trường hợp chia giỏ theo độ sâu.

c. Phát hiện outlier bằng khoảng cách Euclide

Bài 4. Giả sử một bệnh viện kiểm tra dữ liệu tuổi và lượng mỡ cơ thể cho 18 người trưởng thành được chọn ngẫu nhiên với kết quả như sau

age	23	23	27	27	39	41	47	49	50
%fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
age	52	54	54	56	57	58	58	60	61
%fat	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

a. Tính giá trị trung bình, trung vị và độ lệch chuẩn của tuổi và % chất béo.

b. Trực quan bằng boxplot cho tuổi và % chất béo.

c. Trực quan bằng scatter plot dựa trên hai biến này. Nhận xét từ hình trực quan.

Bài 5. Giả sử chúng ta có tập dữ liệu hai chiều sau:

	A_1	A_2
x_1	1.5	1.7
x_2	2	1.9
x_3	1.6	1.8
x_4	1.2	1.5
x_5	1.5	1.0

Chuẩn hóa tập dữ liệu để làm cho mỗi điểm dữ liệu rơi vào khoảng [0,1]

Bài 6. Sử dụng các phương pháp được yêu cầu để chuẩn hóa nhóm dữ liệu sau:

200, 300, 400, 600, 1000

a. Chuẩn hóa min-max bằng cách đặt min = 0 và max = 1
b. Chuẩn hóa z-score
c. Chuẩn hóa điểm z bằng cách sử dụng độ lệch tuyệt đối trung bình thay vì độ lệch chuẩn

Bài 7: Cho bảng dữ liệu sau:

	nation	purchased_item	age	salary
1	India	No	25	35000
2	Russia	Yes	NA	40000
3	Germany	No	50	54000
4	Russia	No	35	NA
5	Germany	Yes	40	60000
6	India	Yes	35	58000
7	Russia	No	NA	52000
8	India	Yes	48	NA
9	Germany	No	50	83000
10	India	Yes	37	NA
11	Germany	No	21	24000
12	India	Yes	NA	60000
13	Russia	No	63	70000
14	Germany	yes	26	36000
15	India	No	45	40000

a. Hãy dùng các kỹ thuật khác nhau để điền các giá trị thiếu trong bảng sau, lý giải cho các phương pháp đó.
b. Làm cách nào để mô hình có thể xử lý dữ liệu trên một cách nhanh hơn.

c. Nếu age và salary là hai đặc trưng được sử dụng cho mô hình học máy nào đó ví dụ như K-NN với hàm tính khoảng cách giữa các mẫu giữa là Euclid, bạn sẽ cần thực hiện điều gì trước khi đưa? Giải thích và thực hiện công việc đó.