BÀI TẬP VÀ THỰC HÀNH 1
TÌM HIỂU VÀ TIỀN XỬ LÝ DỮ LIỆU

Chú ý: Với các bài tập tính toán, dùng python để thực hiện

Bài 1. Cho tập giá trị được sắp xếp theo theo thứ tự tăng dần như sau:

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.

Hãy thực hiện các yêu cầu sau:

a. Tính các giá trị trung bình (mean), giá trị trung vị (median), giá trị phổ biến (mode) của tập dữ liệu.
b. Cho nhận xét về các giá trị mode của dữ liệu (tức là bimodal, trimodal, v.v.).
c. Tính giá trị Midrange của tập dữ liệu.
d. Tìm (gần đúng) tứ phân vị đầu tiên (Q1) và tứ phân vị thứ ba (Q3) của tập dữ liệu.
e. Cho biết 5 giá trị thống kê về sự phân phối của dữ liệu (five-number summary).
f. Vẽ biểu đồ hộp (Boxplot) của tập dữ liệu.

Bài 2. Giả sử một bệnh viện đã kiểm tra dữ liệu về độ tuổi và lượng mỡ cơ thể (fat) của 18 người lớn được chọn ngẫu nhiên và thu được kết quả 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 %fat.
b. Vẽ biểu đồ hộp (Boxlot) cho tuổi và %fat.
c. Vẽ biểu đồ phân tán (scatter plot) và biểu đồ q-q (q-q plot) dựa trên hai biến này.

Bài 3. Cho hai đối tượng được biểu diễn bởi các bộ giá trị (22, 1, 42, 10) và (20, 0, 36, 8). Thực hiện:

a. Tính khoảng cách Euclid giữa hai đối tượng.

b. Tính khoảng cách Manhattan giữa hai đối tượng.
c. Tính khoảng cách Minkowski giữa hai đối tượng, sử dụng q D 3.
d. Tính khoảng cách cực đại giữa hai đối tượng.

Bài 4. Cho tập dữ liệu 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

(a) Cho một điểm dữ liệu mới có tọa độ x = (1.4,1.6). Xếp hạng mức độ tương đồng (similarity) giữa các điểm trong tập dữ liệu với điểm dữ liệu x bằng cách sử dụng khoảng cách Euclidean, khoảng cách Manhattan, khoảng cách cực đại và độ tương đồng cosin.
(b) Chuẩn hóa tập dữ liệu để làm cho chuẩn của mỗi điểm dữ liệu bằng 1. Sử dụng khoảng cách Euclidean trên dữ liệu đã chuyển đổi để xếp hạng các điểm dữ liệu.

Bài 5. Trong thực tế, các tập dữ liệu có giá trị bị thiếu (missing value) đối với một số thuộc tính là hiện tượng thường gặp. Mô tả các phương pháp khác nhau để xử lý vấn đề này.

Bài 6. Cho tập 12 giá trị đã được sắp xếp như sau:

5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215.

Phân chia tập giá trị thành ba ngăn theo từng phương pháp sau:
a. phân chia theo tần suất bằng nhau (độ sâu bằng nhau): equal-frequency
b. phân chia theo chiều rộng bằng nhau: equal-width
c. phân cụm: clustering

Bài 7. Áp dụng các phương pháp chuẩn hóa này để chuẩn hóa tập giá trị sau:

200, 300, 400, 600, 1000

a. chuẩn hóa min-max với min = 0 và max = 1
b. chuẩn hóa z-score
c. chuẩn hóa z-score bằng cách sử dụng độ lệch chuẩn tuyệt đối trung bình (mean absolute deviation)
d. chuẩn hóa theo tỉ lệ thập phân (decimal scaling)

Bài 8. Cho tập giá trị được sắp xếp theo theo thứ tự tăng dần như sau:

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.

Hãy thực hiện các yêu cầu sau:

a. Sử dụng phương pháp làm mịn bin trung bình (smoothing by bin means) để làm mịn dữ liệu trên với độ sâu bin là 3. Trình bày từng bước thực hiện và cho nhận xét về tác động của kỹ thuật này đối với dữ liệu đã cho.
b. Sử dụng chuẩn hóa min-max để chuyển đổi giá trị 35 thuộc phạm vi [0,0,1,0].
c. Sử dụng chuẩn hóa z-score để chuyển đổi giá trị tuổi 35 với độ lệch chuẩn của độ tuổi là 12,94 năm.
d. Sử dụng chuẩn hóa theo tỉ lệ thập phân (decimal scaling) để chuyển đổi giá trị 35 cho độ tuổi.

Bài 9. Cho tập dữ liệu như bài 4.

a. Chuẩn hóa hai thuộc tính dựa trên chuẩn hóa z-score.
b. Tính hệ số tương quan Pearson (Pearson's product moment coefficient) của hai thuộc tính và cho biết hai thuộc tính này có tương quan tích cực hay tiêu cực?
c. Tính hiệp phương sai (covariance) của chúng.