Câu 1: Thống kê mô tả (8 điểm, CLO 1)

a. (2 điểm) Tìm hiểu thông tin về bộ dữ liệu và các thuộc tính của bộ dữ liệu

Bộ dữ liệu gồm những cột nào?

Kiểu dữ liệu của từng cột trong bộ dữ liệu là gì?

Có bao nhiêu giá trị bị thiếu trong mỗi cột?

Phân phối của các giá trị trong từng cột như thế nào? (ví dụ: điểm trung bình, điểm cao nhất, điểm thấp nhất)

b. (2 điểm) Lọc dữ liệu thỏa mãn yêu cầu số 1

Lọc ra danh sách các thí sinh có điểm tổng >= X điểm (ví dụ: >= 24 điểm).

Lọc ra các thí sinh có điểm 3 môn theo khối A (Toán, Lý, Hóa) >= 18 điểm.

Lọc ra danh sách thí sinh có điểm thấp nhất ở từng môn học.

c. (2 điểm) Lọc dữ liệu thỏa mãn yêu cầu số 2

Lọc danh sách thí sinh theo khu vực (ví dụ: chỉ lấy dữ liệu từ Hà Nội).

Lọc danh sách thí sinh có ít nhất một môn đạt điểm tuyệt đối (10 điểm).

Lọc danh sách thí sinh có ít nhất một môn bị điểm liệt (<= 1 điểm).

d. (2 điểm) Trực quan hóa dữ liệu đã cho

Biểu đồ phân phối điểm của từng môn học.

Biểu đồ so sánh số lượng thí sinh đạt các mức điểm tổng khác nhau.

Biểu đồ số lượng thí sinh đạt điểm liệt theo từng môn học.

Câu 2: Mô phỏng mẫu ngẫu nhiên (2 điểm, CLO 2)

a. Sử dụng hàm random để mô phỏng nhiều mẫu ngẫu nhiên

Lấy mẫu ngẫu nhiên 1000 thí sinh từ tập dữ liệu gốc.

Lấy mẫu ngẫu nhiên 500 thí sinh đạt điểm tổng cao nhất.

Lấy mẫu ngẫu nhiên 300 thí sinh có ít nhất một môn bị điểm liệt.

b. Tổng hợp các đặc trưng của các mẫu ngẫu nhiên vừa tạo và đưa ra nhận xét

Trung bình điểm từng môn của các mẫu so với toàn bộ dữ liệu.

Phân phối điểm tổng của các mẫu.

Nhận xét về sự khác biệt giữa các mẫu so với tập dữ liệu gốc.