KHOA CÔNG NGHỆ THÔNG TIN
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
227 Nguyễn Văn Cừ, Phường 4, Quận 5, TP.HCM
Điện thoại: (08) 38.354.266 – Fax: (08) 38.350.096

CSC12110 – PHÂN TÍCH DỮ LIỆU ỨNG DỤNG

BÀI TẬP LÝ THUYẾT 1

BTLT 1: GIỚI THIỆU VỀ PHÂN TÍCH DỮ LIỆU

I. Thông tin chung

Mã số bài tập:	BTLT 1
Thời lượng dự kiến:	3 tiếng
Deadline nộp bài:	11:00, 31/10/2024
Hình thức:	Bài tập cá nhân
Hình thức nộp bài:	Nộp qua Moodle
GV phụ trách:	Vũ Thị Mỹ Hằng
Thông tin liên lạc với GV:	vtmhang@fit.hcmus.edu.vn

II. Các yêu cầu & quy định chi tiết cho bài nộp

- Sinh viên lập trình, chú thích giải pháp trực tiếp trên tập tin IPYNB.
- **Quy định đặt tên file**: MSSV_HoTen.ipynb

III. Mô tả bài tập

Sử dụng dữ liệu điểm thi tốt nghiệp THPT 2023 và 2024 để hoàn thành các yêu cầu sau.

1. Chuẩn bị dữ liệu

- Tích hợp thông tin mã và tên tỉnh thành vào các bộ dữ liệu điểm thi, biết rằng 2 chữ số đầu tiên trong SBD của thí sinh đại diện cho mã tỉnh thành (sử dụng bộ dữ liệu hội đồng thi để kiểm tra thông tin mã và tên tỉnh thành).
- Tích hợp thông tin tên môn ngoại ngữ thí sinh dự thi, biết rằng mã ngoại ngữ được quy định như sau: N1 – Tiếng Anh; N2 – Tiếng Nga; N3 – Tiếng Pháp; N4 – Tiếng Trung Quốc; N5 – Tiếng Đức; N6 – Tiếng Nhật; N7 – Tiếng Hàn.
- Tích hợp hai bộ dữ liệu 2023 và 2024 thành một bộ dữ liệu điểm thi tổng hợp (thêm năm để phân biệt).
- Bộ dữ liệu tổng hợp sẽ được sử dụng cho các câu hỏi kế tiếp.

2. Khám phá thông tin cơ bản

- Hiển thị thông tin cơ bản của bộ dữ liệu tổng hợp: kích thước, chiều, số cột, kiểu dữ liệu của từng cột, số lượng bộ dữ liệu bị thiếu của từng cột và dữ liệu mẫu.
- Hiển thị các chỉ số thống kê cơ bản cho mỗi môn thi của từng năm: trung bình (mean), trung vị (median), yếu vị (mode), độ lệch chuẩn (std), …

3. Tiền xử lý dữ liệu

- Kiểm tra kiểu dữ liệu điểm phải là số thực trong khoảng 0-10.

- Kiểm tra mã ngoại ngữ (nếu có) phải là các giá trị hợp lệ (N1-N7). **Lưu ý**: thí sinh có điểm ngoại ngữ thì bắt buộc phải đăng ký mã ngoại ngữ.
- Kiểm tra giá trị thiếu và đề xuất phương pháp xử lý phù hợp (giải thích).
- Kiểm tra dữ liệu ngoại lai (dùng biểu đồ và phương pháp thống kê) và đề xuất phương pháp xử lý phù hợp (giải thích).

4. Phân tích dữ liệu khám phá (EDA)

Cho các yêu cầu bên dưới, lựa chọn giải pháp phân tích đồ hoạ/phi đồ hoạ phù hợp và cài đặt giải pháp để đáp ứng yêu cầu phân tích.

- Thống kê số lượng thí sinh dự thi theo từng tỉnh thành.
- Thống kê điểm trung bình, điểm lớn nhất, nhỏ nhất của từng môn theo từng năm.
- Thống kê số lượng thí sinh dự thi ở mỗi ngoại ngữ theo từng năm.
- Xem xét phân phối điểm cho các môn thi chính (toán, ngữ văn, ngoại ngữ). Phân tích xem các môn thi có phân phối điểm lệch về phía nào không.
- Phân tích tương quan giữa các cặp môn thi.
- So sánh điểm trung bình của các môn thi theo từng năm.
- So sánh số lượng thí sinh dự thi ở các môn tự chọn (vật lý, hoá học, sinh học, lịch sử, địa lý, GDCD) theo từng năm.