DÀN Ý ĐỀ TÀI MÔN HỌC PHÂN TÍCH DỮ LIỆU CƠ BẢN

1. Giới thiệu về bộ dữ liệu cần dùng (liệt kê các thuộc tính cần cho phân tích, số lượng records tối thiểu cần có, …)
2. Xây dựng phiếu khảo sát để thu thập dữ liệu
3. Quá trình tiền xử lý dữ liệu

   - Khi nộp báo cáo môn học, SV sẽ nộp 3 file CSV: file gốc (original), file do SV chủ động xóa/hiệu chỉnh dữ liệu (before), và kết quả sau khi thực hiện tiền xử lý (after)
   - Quá trình này SV phải thực hiện lần lượt bằng 2 cách: dùng SQL và dùng Python

(i).- Làm sạch dữ liệu *(Data Cleaning):* (tối thiểu 5 trường hợp khác nhau)

- *Điền giá trị còn thiếu*:
  - Nếu dữ liệu không có giá trị NULL, SV chủ động xóa trên tất cả các field (ngoại trừ field ID), mỗi field khoảng 5 giá trị (trên 5 record liền nhau để dễ quan sát kết quả).
  - Không điền những giá trị đơn giản (như "không biết", "không có", hay min, max, mean, mode, median) cho những giá trị NULL mà cần điền bằng những giá trị đòi hỏi tính toán phức tạp hơn, ví dụ điền giá trị NULL bằng giá trị của mode, nếu có nhiều mode thì chọn mode có giá trị thấp nhất hoặc đối với kiểu số thực, có thể thực hiện làm tròn (đến phần nguyên hoặc phần ngàn, …) trước khi tính mean/mode/median
- *Làm mịn dữ liệu:*
  - Chọn thuộc tính có miền giá trị cao nhất trong dữ liệu. Thực hiện chia miền giá trị này thành nhiều nhóm (>=5 nhóm). Chọn 1 trong 3 cách sau để làm mịn dữ liệu:
    - Thay thế bằng giá trị trung bình (mean) của nhóm.
    - Thay thế bằng giá trị trung vị (median) của nhóm.
    - Làm mịn theo ranh giới của nhóm (*smoothing by bin boundaries*)
  - Chọn field có độ lệch chuẩn lớn nhất trong dữ liệu: thực hiện tăng 10% giá trị của thuộc tính này cho 5 giá trị nhỏ nhất và giảm 10% giá trị của thuộc tính này cho 5 giá trị lớn nhất.
- *Xác định giá trị ngoại lệ (cho các field kiểu số)*
  - Xác định dữ liệu có chứa giá trị ngoại lệ (tính theo five number summary) của từng thuộc tính số hay không?
- Xử lý dữ liệu không nhất quán:
  - Đối với field có kiểu dữ liệu ngày: từ file dữ liệu gốc, SV thực hiện đổi kiểu dữ liệu thành chuỗi, sau đó sửa thủ công các giá trị ngày cho ít nhất 5 giá trị để có 5 dữ liệu có định dạng ngày khác với các record còn lại. Ví dụ định dạng của ngày hiện tại là mm/dd/yy, sửa 5 giá trị để 5 giá trị này có kiểu cùng là dd/mm/yy hoặc cùng có kiểu là yy/mm/dd. Sau đó thực hiện cập nhật lại giá trị cho 5 giá trị này.
  - Đối với field kiểu danh nghĩa, SV cũng chủ động chỉnh sửa 5 giá trị, sau đó dủng lệnh thực hiện hiệu chỉnh lại cho đúng với giá trị gốc

(ii).- Tích hợp dữ liệu *(Data Integration):* nếu dữ liệu gốc của SV gồm nhiều table, cần mô tả quá trình tích hợp (kết hợp dữ liệu).

(iii).- Giảm thiểu dữ liệu (*Data Reduction*): (tối thiểu 2 trường hợp khác nhau)

(iv).- Chuyển đổi dữ liệu và phân tách dữ liệu *(Data Transformation and Data Discretization)* Khi kiểu dữ liệu của field thuộc các dạng sau, SV cần phải xử lý. (thực hiện đối với tất cả các thuộc tính dạng danh mục hoặc thứ tự)

- Kiểu danh mục (ví dụ như các mùa: Xuân, Hạ, Thu, Đông): chuyển thành kiểu số nguyên.

- Kiểu thứ tự (ví dụ như: Giỏi, khá, trung bình, yếu): chuyển thành kiểu số nguyên.
- Kiểu boolean: khi dữ liệu có từ 5 field kiểu này trở lên, chuyển 5 field thành 1 field kiểu số.
- Kiểu chuỗi: trong field kiểu chuỗi lại chứa danh sách các chuỗi con có ý nghĩa (ví dụ Tên hàng Đã mua trong đó gồm tên các sản phẩm đã mua trong đơn hàng như: bột giặt, khăn tắm, chén, khăn tắm): cần tách tên các sản phẩm ra khỏi field này để có nhiều field mới. Sau đó, do có thể có rất nhiều field mới nên cần chuyển đổi tất cả các field mới này trở lại thành 1 field duy nhất có kiểu là int.

Trong quá trình chuyển đổi có yêu cầu tạo ra các table làm trung gian, SV cần tìm cách gộp các table vừa phát sinh có cùng cấu trúc, ý nghĩa để giảm bớt số lượng table thực tế sẽ dùng.

(v).- Sau khi hoàn tất tiền xử lý dữ liệu, đối với từng field trong dữ liệu, tùy thuộc kiểu dữ liệu của mỗi field, SV cần thống kê được các giá trị sau:

- Mean
- Mode
- median
- Std. Deviation
- Quantiles
    - Min
    - 25%
    - 50%
    - 75%
    - Max

4. Trừu tượng hóa dữ liệu

- Trình bày tối thiểu 15 đồ thị, trong đó có tối thiểu 10 loại đồ thị khác nhau và có ít nhất 5 đồ thị được vẽ dựa trên dữ liệu được tổng hợp.
- Khuyến khích sử dụng thêm ngôn ngữ R để trừu tượng hóa dữ liệu.
- Khi đồ thị được vẽ bởi lệnh của ngôn ngữ khác (hoặc ứng dụng khác), SV cần mô tả rõ ngôn ngữ (hoặc công cụ) đã dùng.

5. Kết luận và hướng phát triển