Mẫu: 2a_ĐTTL

TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP THÀNH PHỐ HỒ CHÍ MINH

KHOA/VIỆN :CNTT

ĐỀ THI GIỮA KỲ

Môn thi : Lập trình Phân Tích Dữ Liệu 1

Lớp/Lớp học phần: DHHTTT17BTT

Ngày thi:

Thời gian làm bài:.90 phút
(Không kể thời gian phát đề)

Họ và tên thí sinh ……; MSSV: ……

ĐỀ 1

PHẦN 1: TÌM HIỂU DỮ LIỆU (1 ĐIỂM)

1. Đọc File với các file trong thư mục dữ liệu. Hiển thị toàn bộ dữ liệu của file dữ liệu đã đọc. Tìm hiểu và giải thích về bộ dữ liệu. Cho biết biến nào là định tính, biến nào là định lượng, kiểu dữ liệu và thang đo cho mỗi thuộc tính.
2. Sắp xếp dữ liệu theo 1 thuộc tính nào đó, cùng dữ liệu thì theo 1 thuộc tính khác giảm dần. Xem lại dữ liệu sau khi sắp xếp

3. PHẦN 2: TIỀN XỬ LÝ DỮ LIỆU – LÀM SẠCH DỮ LIỆU (3 ĐIỂM)

1. Dùng Heapmap để trực quan dữ liệu bị thiếu. Cho biết dữ liệu nào đang bị thiếu.
2. Điền giá trị thiếu cho biến định tính của 1 cột nào đó bằng giá trị yếu vị. Xem lại dữ liệu sau khi thay đổi. Điền giá trị thiếu cho biến định lượng của 1 cột nào đó bằng giá trị trung bình. Xem lại dữ liệu sau khi thay đổi.
3. Xoá bỏ các dòng dữ liệu rỗng. Xem lại kết quả. Cho biết kết quả có bao nhiêu dòng.
4. Lọc dữ liệu theo nhiều điều kiện. Sinh viên tự nghĩ câu hỏi – yêu cầu ghi cụ thể câu hỏi ra và viết lệnh thực thi. Cho biết có bao nhiêu dòng
5. Sửa dữ liệu của 2 cột nào đó theo điều kiện nào đó. Hiển thị dữ liệu 2 cột này. Dùng lệnh để xoá các dòng có dữ liệu trùng nhau trên 2 cột này. Hiển thị lại kết quả dữ liệu 2 cột này và cho biết có bao nhiêu dòng dữ liệu trên 2 cột này.
6. Lưu trữ dữ liệu đã xử lý thành công với tên file ***TenFileCu_clean.csv***

PHẦN 3: XỬ LÝ DỮ LIỆU – TRỰC QUAN HOÁ DỮ LIỆU (3 ĐIỂM)

1. Tạo 1 biến mới, thực hiện tính giá trị cho biến này theo công thức nào đó. Yêu cầu viết hàm để tính giá trị cho biến này.
2. **Tạo mới dataframe thứ 2 chỉ chứa danh sách các dữ liệu từ dataframe 1 theo 1 điều kiện nào đó.** Nối 2 dataframe3 bằng dataframe 1 và dataframe 2 này lại với nhau.
3. Dùng biểu đồ barplot để thực hiện thống kê minh hoạ trung bình của 1 cột dữ liệu định lượng theo 1 nhóm nào đó.
4. Vẽ biểu đồ so sánh của 1 thuộc tính giữa các nhóm định lượng.
5. Vẽ biểu đồ nhiệt (Heap) thể hiện sự tương quan của các biến định lượng
6. Vẽ biểu đồ tròn thể hiện tỉ lệ phần trăm của 1 biến định lượng theo 1 nhóm nào đó.

6. Dùng biểu đồ boxlot để tìm giá trị ngoại lệ cho 1 thuộc tính nào đó. Tìm độ trải giữa (IQR) của cột dữ liệu bị ngoại lệ. Loại bỏ dữ liệu ngoại lệ.

PHẦN 4: THỐNG KÊ SUY DIỄN (3 ĐIỂM)

1. Thực hiện kiểm định trung bình của 1 biến số (định lượng) bằng phương pháp Z-Test bằng một giá trị nào đó với mức sai lầm là 5%. Cho nhận xét
2. Thực hiện kiểm định trung bình của 2 biến số (định lượng) có bằng nhau không bằng phương pháp T-Test với mức sai lầm là 10%. Cho nhận xét
3. Thực hiện kiểm tra 2 biến định lượng có tương quan với nhau không bằng phương pháp Chi-Square với mức sai lầm là 5%? Cho nhận xét.

Lưu ý: *- Đề thi được sử dụng tài liệu.*

- Cán bộ coi thi không giải thích gì thêm.