|  |  |  |
| --- | --- | --- |
| **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: DHHTTT18BTT**  **Ngày thi:**  Thời gian làm bài: 90 phút  *(Không kể thời gian phát đề)* |  |
|  |  |  |

*Họ và tên thí sinh ………………….…………………; MSSV: ………………*

**PHẦN 1: TÌM HIỂU DỮ LIỆU VÀ LÀM SẠCH LÝ DỮ LIỆU (2 ĐIỂM \_ CLO 1)**

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. Dùng biểu đồ để trực quan dữ liệu bị thiếu. Cho biết dữ liệu nào đang bị thiếu. Giá trị nào bị thiếu bạn hãy điền thêm vào bằng 1 trong các cách đã được học (thay bằng giá trị yếu vị, giá trị trung bình hay 1 giá trị nào đó. Xem lại dữ liệu sau khi thay đổi.
3. Sửa dữ liệu của 1 cột nào đó theo điều kiện nào đó. Hiển thị dữ liệu của cột này.
4. Lưu trữ dữ liệu đã xử lý thành công với tên file ***TenFileCu\_clean.csv***

**PHẦN 2: XỬ LÝ DỮ LIỆU – TRỰC QUAN HOÁ DỮ LIỆU (3 ĐIỂM\_CLO2)**

1. **Trộn 2 bảng dữ liệu với nhau vào 1 dataframe.**
2. 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.
3. Vẽ biểu đồ so sánh của 1 thuộc tính giữa các nhóm định lượng.
4. Vẽ biểu đồ thể hiện sự tương quan của các biến định lượng
5. 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 3: THỐNG KÊ SUY DIỄN VÀ HỒI QUY TUYẾN TÍNH (5 ĐIỂM – CLO3)**

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.
4. Thực hiện kiểm định ANOVA 2 chiều. Giải thích kết quả.
5. Đánh giá mức độ tác động giữa các yếu tố đến 1 đối tượng (1 biến độc lập và 1 biến phụ thuộc) bằng phân tích hồi quy tuyến tính theo phương thức OLS. Giải thích kết quả

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

*- Cán bộ coi thi không giải thích gì thêm.*