KIỂM TRA THƯỜNG KỲ

PHẦN 1: TÌM HIỂU DỮ LIỆU (2 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 này là định tính, biến nào là định lượng, kiểu dữ liệu và thang đo cho mỗi thuộc tính.
2. Hiển thị số dòng và số cột của bảng dữ liệu. Xem thông tin của dataframe vừa đọc
3. Xem thông tin của của 5 dòng đầu với ít nhất 5 cột mà bạn cho là quan trọng.
4. Hiển thị số lượng giá trị phân biệt (khác nhau từng đôi một) của cột dữ liệu

PHẦN 2: TIỀN XỬ LÝ DỮ LIỆU – LÀM SẠCH DỮ LIỆU (4 ĐIỂM)

1. Chuyển kiểu dữ liệu cho 1 cột nào đó
2. Xoá 2 cột không quan tâm. Thực hiện đổi tên 2 cột cho ngắn gọn hơn.
3. Dùng Heapmap để trực quan dữ liệu bị thiếu. Cho biết dữ liệu nào đang bị thiếu.
4. Hiển thị dữ liệu rỗng của từng cột dữ liệu. Xoá bỏ các dòng dữ liệu rỗng. Hiển thị lại dữ liệu rỗng của từng cột. Nhận xét kết quả từng lại.
5. Điền giá trị thiếu cho biến định tính của 1 cột nào đó bằng giá trị yếu vị (mode). Xem lại dữ liệu sau khi thay đổi.
6. Điền giá trị thiếu cho biến định lượng của 1 cột nào đó bằng giá trị 0. Xem lại dữ liệu sau khi thay đổi.
7. Điền giá trị thiếu cho biến định lượng của 1 cột nào đó bằng trung vị, trung bình Xem lại dữ liệu sau khi thay đổi.
8. Điền giá trị thiếu cho biến định lượng của 1 cột nào đó kiểu chuỗi bằng giá trị "Không biết". Xem lại dữ liệu sau khi thay đổi.

PHẦN 3: XỬ LÝ DỮ LIỆU – TRỰC QUAN HOÁ DỮ LIỆU (4 ĐIỂM)

1. Sắp xếp dữ liệu theo 1 cột nào đó theo thứ tự tăng dần, nếu giá trị cột này giống nhau thì sắp xếp theo 1 thuộc tính khác thứ tự giảm dần.
   df.sort_values(by='SALES',ascending=True)
2. Lọc dữ liệu theo nhiều điều kiện. Sinh viên tự nghĩ câu hỏi và viết lệnh thực thi.
   df[(df['SALES']>5000) & (df['QUANTITYORDERED']>40)]
3. 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.
4. Thêm 1 cột mới. Nhập giá trị của cột nào đó. Xóa cột này trong dataframe
5. 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.
6. Dùng biểu đồ barplot để thực hiện thống kê minh hoạ cho các hàm (estimator): count, min, max, std, mean, mode

Lưu ý:

- Với biến định tính thì ta chỉ có 1 hàm tổng hợp là hàm COUNT, MODE

- Với định lượng thì ta có thể sử dụng các hàm tổng hợp như: COUNT, MAX, MIN, MEAN, MEDIAN, MODE, SUM, STD

7) Dùng biểu đồ PIE để trực qua hoá dữ liệu theo nhóm tỷ lệ phần trăm. Trực quan hoá dữ liệu bằng các biểu đồ Line, histogram, scatter

8) 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.