CASE STUDY 08 – PHÂN TÍCH HOẠT ĐỘNG THƯƠNG MẠI ĐIỆN TỬ (E-COMMERCE)

I. Đặt vấn đề

Hoạt động thương mại điện tử tại Việt Nam ghi nhận dữ liệu về danh mục sản phẩm, chi tiết đơn hàng và thông tin trả hàng. Case Study này giúp sinh viên thực hành đầy đủ kỹ năng Pandas: làm sạch dữ liệu, truy vấn, phân nhóm, merge và xây dựng pivot để phân tích tỷ lệ trả hàng, doanh thu và hiệu suất danh mục sản phẩm.

II. Mô tả dữ liệu

Dữ liệu gốc gồm 3 file CSV: product_catalog.csv, order_details.csv và return_refund.csv.

Trong đó:

- File product_catalog.csv gồm các trường: product_id: mã sản phẩm (có khoảng trắng, viết thường); product_name: tên sản phẩm (viết sai chuẩn, ký tự thừa); category: danh mục sản phẩm (Điện tử, Gia dụng, ... — có thể sai chính tả).
- File order_details.csv gồm các trường: order_id: mã đơn hàng; product_id: mã sản phẩm; quantity: số lượng (có thể âm hoặc chứa text); unit_price: đơn giá (có thể chứa ký tự tiền tệ); order_date: ngày đặt đơn (2 định dạng).).
- File return_refund.csv gồm các trường: order_id: mã đơn hàng; product_id: mã sản phẩm; reason: lý do trả hàng (có thể chứa ký tự đặc biệt); refund_amount: số tiền hoàn (có thể chứa text hoặc None).

III. Nhiệm vụ

Nhiệm vụ 1 – Đọc & làm sạch dữ liệu:

Mục tiêu : hiểu cấu trúc dữ liệu thô và xử lý các vấn đề ảnh hưởng đến phân tích.

- Làm sạch product_id, product_name và category.
- Chuẩn hóa quantity và unit_price.
- Chuẩn hóa order_date sang dạng thống nhất.
- Xử lý refund_amount và lý do trả hàng.
- Xuất các file sạch để dùng cho các nhiệm vụ sau.

Nhiệm vụ 2 – Truy vấn & thống kê mô tả:

Mục tiêu : thực hành truy xuất dữ liệu và rút ra các thống kê cơ bản.

- Thống kê số lượng đơn theo danh mục.
- Tìm các đơn có quantity bất thường.
- Thống kê sản phẩm bán chạy nhất.

- Tìm sản phẩm có nhiều đơn trả hàng nhất.
- Thống kê tỷ lệ trả hàng theo từng danh mục.

Nhiệm vụ 3 – GroupBy & Tổng hợp:

Mục tiêu : phân nhóm dữ liệu và rút ra các đặc trưng quan trọng.

- Tính doanh thu theo sản phẩm (quantity × unit_price).
- Tính doanh thu trung bình theo danh mục.
- Tổng hợp tỷ lệ trả hàng theo sản phẩm.
- Xác định nhóm sản phẩm có hiệu suất thấp.
- Tính doanh thu theo tháng.

Nhiệm vụ 4 – Merge dữ liệu:

Mục tiêu : tạo một file dữ liệu hoàn chỉnh phục vụ phân tích sâu hơn.

- Merge product_catalog với order_details.
- Merge thêm return_refund theo order_id & product_id.
- Phát hiện mismatch giữa đơn đặt và đơn trả hàng.
- Xuất file hoàn chỉnh.

Nhiệm vụ 5 – Pivot Table + Stack/Unstack:

Mục tiêu: rèn luyện khả năng chuyển đổi cấu trúc dữ liệu và phân tích theo nhiều chiều.

- Pivot doanh thu theo category × tháng.
- Pivot tỷ lệ trả hàng theo category × sản phẩm.
- Thực hiện stack/unstack bảng pivot.
- Nhận xét danh mục có hiệu suất thấp nhất.

IV. Quy định tổ chức GitHub

- Tạo repo: 18Am_Nhom _n_Case_ p

Trong đó: ký hiệu **m** (1 –3) thể hiện số thứ tự lớp K18A1, K1 8A2, K18A3; **n** là tên nhóm; và **p** là số thứ tự của Case Study mà nhóm thực hiện.

- Cấu trúc repo gồm:
    - data_raw/ (dữ liệu gốc)
    - data_clean/ (dữ liệu sạch)
    - task01_cleaning/
    - task02_query/
    - task03_groupby/
    - task04_merg e/
    - task05_pivot/
    - final_report/

- Mỗi sinh viên tối thiểu 3 commit.
- Commit message rõ ràng: [Task03] Tính trung bình điểm theo môn.
- Invite giảng viên quyền Read.
- Không tạo repo mới nhằm tránh mất lịch sử commit.

V. Hướng dẫn nộp bài

- Nộp 5 notebook (.ipynb) tương ứng 5 nhiệm vụ của Case Study .
- Nộp 3 file CSV đã chuẩn hóa sạch.
- Nộp các file kết quả phân tích từ groupby/merge/pivot (nếu có xuất file riêng) .
- Nộp 01 báo cáo .docx hoặc .pdf.

V.1. Qui định về tên thư mục nộp bài:

Tên thư mục: 18Am_Nhom_n_Case_p

Trong đó:

- (m=1,2,3 ứng với lớp là A1/A2,/ A3),
- n là tên nhóm (từ 01 đến 12)
- p : là số thứ tự của Case Study do nhóm th ực hiện.

Ví dụ: 18A1_Nhom_ 01_Case_0 1

V.2. Quy định về tên file báo cáo

- File báo cáo: **Baocao_ 18Am _Nhom_ n_Case_p**.docx/pdf,

Trong đó: m, n, p giữ ý nghĩa như phần tên thư mục.

V.3. Nội dung báo cáo

Báo cáo cần phải thể hiện rõ:

- Phương pháp xử lý và phân tích dữ liệu.
- Các bảng kết quả chính
- Nhận xét dựa trên kết quả phân tích.
- Kết luận
- Nhật ký nhóm : thể hiện mức độ tham gia của từng thành viên .

V.4. Thời gian nộp bài :

Hạn nộp: trước 0 giờ ngày 16 tháng 12 năm 2025.

- Cách nộp bài: đại diện nhóm gửi invite quyền Read repo GitHub cho giảng viên qua email cdthang @uneti.edu.vn

VI. Gợi ý công cụ & kiến thức cần ôn tập, tham khảo

Yêu cầu	Hàm và Kỹ thuật	Ghi chú
Nhập xuất file (I/O)	"read_csv(), read_excel(), to_csv()"	

Làm sạch,	"dropna(), fillna(), astype(), to_datetime(), str.strip(), str.title(),...",	Đảm bảo kiểu dữ liệu chuẩn xác trước khi phân tích
Truy vấn	"loc, iloc, isin"	Sử dụng Boolean Masking (Mặt nạ logic) để lọc dữ liệu theo điều kiện phức tạp (kết hợp các điều kiện bằng & và ,...)
GroupBy	"groupby(), agg()",	
Merge,merge()	với inner/left,	
Pivot Table	pivot_table()	Đây là kiến thức cốt lõi để tổng hợp và xoay dữ liệu.
Stack/Unstack,	"stack(), unstack()"	Chuyển đổi DataFrame ↔ Series đa cấp. Thường áp dụng cho kết quả từ groupby() hoặc pivot_table()
Optional (Tùy chọn)	Vẽ biểu đồ cơ bản,	"Khuyến khích minh họa phân tích (line, bar,...)

VII. Đánh giá và chấm đi ểm (Rubric)

Bài làm được đánh giá chấm điểm như sau:

1) Chất lượng xử lý dữ liệu (20%)
   - File sạch, hợp lý, không lỗi
   - Xử lý đầy đủ vấn đề: ngày sinh, NA, họ tên, subject_code, điểm
2) Kết quả truy vấn & thống kê mô tả (15%)
   - Kết quả đúng, rõ ràng
   - Trình bày logic trong báo cáo
3) Phân tích tổng hợp (GroupBy) (20%)
   - Bảng tổng hợp rõ ràng
   - Có phân tích ý nghĩa (không chỉ in bảng)
4) Merge & tạo file dữ liệu hoàn chỉnh (10%)
   - File merge đúng và nhất quán
   - Không thừa thiếu dữ liệu bất hợp lý
5) Pivot / Stack / Unstack (20%)
   - Bảng pivot đúng cấu trúc
   - Hiểu đúng bản chất chuyển đổi dạng dữ liệu

- Phân tích NaN, so sánh lớp/môn

6) Báo cáo (10%)

- Đúng cấu trúc: phương pháp – kết quả – nhận xét – kết luận
- Có Nhật ký nhóm
- Trình bày sạch, không lỗi

7) Hoạt động nhóm và commit cá nhân (5%)

- ≥3 commit/người, đúng nội dung
- Tính điểm cá nhân dựa trên lịch sử commit

Lưu ý: Nếu phát hi ện dấu hiệu sao ché p, đạo văn , nhờ AI để làm bài thay hoàn toàn sẽ bị trừ tối thiểu 50% đi ểm của Case Study .