THỰC HÀNH TẠO BỘ DỮ LIỆU MỚI DỰA TRÊN CÁC BƯỚC GỢI Ý

Bảng **customer_purchases** như sau:

customer_purchases				customer_purchases			
market_date	vendor_id	quantity	cost_to_customer_per_qty	market_date	vendor_id	quantity	cost_to_customer_per_qty
3/2/2019	8	2	4	3/8/2020	9	2	5
3/2/2019	8	1	4	3/1/2020	4	5	2.5
3/9/2019	8	1	4	3/8/2020	4	7	2.2
3/9/2019	9	1	16	3/6/2021	8	2	5
3/9/2019	9	1	18	3/13/2021	9	1	17
3/2/2019	4	5	2	3/6/2021	4	8	2
3/2/2019	4	8	2	3/13/2021	4	7	2.5
3/2/2019	4	1	2	3/5/2022	8	3	4.2
3/9/2019	4	10	2	3/12/2022	9	2	5.5
3/2/2019	1	1	5.5	3/5/2022	4	5	2.1
3/1/2020	8	3	4.5	3/12/2022	4	6	2.3

a. Bộ dữ liệu cho phân tích chuỗi thời gian

Ví dụ ta cần tạo tập một bộ dữ liệu về doanh số bán hàng hàng tuần tại chợ nông sản và được tổng hợp từ bảng customer_purchases. Yêu cầu: Tính tổng doanh số mỗi tuần của các năm khác nhau.

1. Tạo một bảng dữ liệu market_date_info chứa các thuộc tính market_year và market_week để xác định thông tin thời gian của chợ theo năm và tuần. VD:

market_date	market_year	market_week
2019-03-02	2019	9

2. Doanh số thu được từ mỗi giao dịch được xác định bởi **quantity * cost_to_customer_per_qty**
   VD:

market_date	vendor_id	quantity	cost_to_customer_per_qty	quantity * cost_to_customer_per_qty
3/8/2020	9	2	5	10
3/1/2020	4	5	2.5	12.625

3. Nhóm dữ liệu theo market_year và market_week để thu được doanh số hàng tuần từ các năm khác nhau, gợi ý: sử dụng ngày đầu tiên của mỗi tuần (first_market_date_of_week) làm mốc thời gian đại diện cho mỗi tuần → **MIN(market_date)**

b. Bộ dữ liệu cho phân loại nhị phân

Gợi ý thực hiện VD: Tạo một bảng tạm thời CTE customer_markets_attended để để theo dõi từng lần mua hàng của khách hàng, trong đó mỗi hàng đại diện cho một khách hàng tại một ngày mua hàng (market_date). Mệnh đề GROUP BY sẽ nhóm dữ liệu theo customer_id và market_date để lọc và nhóm theo từng khách hàng và ngày mua. Biến mục tiêu purchased_again_within_30_days sẽ nhận một trong hai giá trị 0 (khách hàng không mua lại trong 30 ngày) hoặc 1 (khách hàng có mua lại trong 30 ngày).

- B1: Tạo một bảng tạm `customer_markets_attended(customer_id, market_date)` chứa danh sách duy nhất các lần mua hàng của từng khách hàng (mỗi khách hàng + ngày mua hàng chỉ xuất hiện 1 lần).
- B2: Tạo truy vấn chính để xác định: ngày mua, id khách hàng, tổng số tiền đã chi, số lượng nhà cung cấp mà khách hàng đã mua, số lượng sản phẩm khác nhau đã mua.
- B3: Tạo Subquery
  - Tìm ngày mua tiếp theo của cùng một khách hàng sau ngày hiện tại (market_date).
  - Tính số ngày giữa hai lần mua
  - Xác định biến mục tiêu (purchased_again_within_30_days)
- B4: Nhóm dữ liệu theo từng khách hàng và từng ngày mua hàng. Đảm bảo mỗi hàng đại diện cho một lần mua duy nhất của khách hàng. ORDER BY: Sắp xếp kết quả theo customer_id và market_date.

Yêu cầu:

1. Thực hiện bài làm trên jupyter notebook. Đặt tên file theo cú pháp HOTEN_MSV.
2. Với mỗi bước thực hiện cần thể hiện kết quả sau mỗi truy vấn.
3. Tổng kết câu trả lời cho ý a.
4. Bảng cuối cùng cho mục tiêu trong ý b là gì?