Dựa trên dữ liệu ‘sample_data.csv.gz’ (trong buổi lecture4) về các tin rao vặt, thực hiện các task sau để tìm ra các insights có giá trị

Task1:

Tạo thêm một column có tên ‘account_type’ để ghi nhận một account_id xuất hiện trong mỗi dòng là account cũ hay mới. Định nghĩa một account gọi là cũ nếu có tương tác (đăng tin rao vặt ad_id) trong 30 ngày gần nhất, nếu trong 30 ngày không có tương tác tạm gọi là account mới

Dựa vào thông tin account_type vẽ biểu đồ thể hiện sự thay đổi, số account cũ và mới theo tháng, theo tuần, theo ngày (nếu số ngày quá nhiều có thể bỏ qua để hình vẽ được gọn gàng).

Lưu ý có thể tham khảo https://python-graph-gallery.com/ để lựa chọn biểu đồ phù hợp dể thể hiện số account cũ/ mới theo thời gian.

Dựa trên biểu đồ nhận xét, số user cũ mới có xu hướng thế nào (bắt đầu từ tháng nào có tăng/giảm, v.v…)

Task 2:

Làm lại task 1, nhưng theo dõi số account cũ mới theo từng category_name. Để có dữ liệu tốt, nên chọn 10 ngành hàng phổ biến (có số lượng account_id tương tác nhiều)

Lựa chọn biểu đồ để thể hiện sự thay đổi của số account cũ/mới theo ngành hàng từ đó phát hiện ra ngành hàng nào có tính bão hòa (số account cũ/mới ổn định), ngành hàng nào có khả năng thu hút thêm được nhiều account mới.

Task 3: Dùng biểu đồ kiểm chứng mối tương quan giữa giá của một ad_id và số lần đăng trong năm.

Với mỗi ad_id sẽ có một giá (price), và có thể ad_id xuất hiện nhiều lần trong data do phải đăng bán nhiều lần vì chưa bán được.

Ta muốn kiểm chứng bằng dữ liệu, các sản phẩm có giá cao sẽ khó bán (ad_id xuất hiện nhiều lần)

Task 4: Làm lại task 3, nhưng thêm một dimension là theo dõi mối tương quan giữa giá và số lần xuát hiện của ad_id cho mỗi ngành hàng.

Lựa chọn biểu đồ phù hợp để thể hiện insights.

Mong muốn tìm ra việc khó bán là do giá cao hay do tính chất đặc thù của ngành hàng?