I. Tổng quan dữ liệu:

Hoạt động 1:

Tệp tin “dirty\_data” chứa thông tin về một cửa hàng bán đồ điện tử trực tuyến. Cửa hàng có trữ đồ tại 3 kho hàng nơi mà sẽ giao đồ được đặt tới khách hàng.

Các biến chính trong bộ dữ liệu:

* Order\_price: giá gốc của đơn hàng (USD)
* Delivery\_charges: Phí giao hàng
* Coupon\_discount: giảm giá trên đơn hàng (%)
* Order\_total: Thành tiền (USD)
* Season: mùa
* Is\_expedited\_delivery: Thể hiện xem khách hàng có yêu cầu rằng đơn hàng này thuộc nhóm giao hàng nhanh hay không? (1/0)
* Distance\_to\_nearest\_warehouse: khoảng cách từ khách hàng tới kho gần nhất (km)
* Is\_happy\_customer: Thể hiện xem khách hàng có hài lòng với dịch vụ hay không? (1/0)

Các bước thực hiện:

* Đọc dữ liệu (Import data)
* Làm sạch dữ liệu (Cleaning data): NA (dữ liệu khuyết)
* Làm rõ dữ liệu (Data visualization)
  + Chuyển đổi biến (nếu có)
  + Thống kê mô tả
  + Thống kê suy diễn
* Xây dựng mô hình hồi quy logistic để đánh giá các nhân tố ảnh hưởng đến việc khách hàng hài lòng hay không
* Dự báo khách hàng có hài lòng với đơn hàng hay không?