# **1. TÌM HIỂU VỀ DỮ LIỆU**

## **1.1. Phân tích bộ dữ liệu**

Trong đồ án môn học này, chúng em sử dụng bộ dữ liệu trong hoạt động mua bán sỉ lẻ, bộ dữ liệu phân tích gồm 2 file csv tương ứng:

* **Bộ dữ liệu *orders.csv***:
  + Là bộ dữ liệu chứa thông tin về lịch sử giao dịch về mua bán sản phẩm đối với khách hàng và người bán.
  + Trong đó, mỗi bản ghi (record) trong bộ dữ liệu tượng trưng cho việc thông tin sản phẩm trong quá trình mua bán như mã sản phẩm, mã đặt hàng, mã khách hàng, mức giá, ngày đặt, ngày giao, số lượng giao dịch, tổng vốn thu được, giá trị mỗi mặt hàng.
  + Các thuộc tính trong bộ dữ liệu được mô tả bao như sau:
    - **Customer ID:** Mã khách hàng
    - **Customer Status:** Trạng thái khách hàng thân thiết.
    - **Date Order was placed:** Ngày đặt hàng.
    - **Delivery Date:** Ngày giao hàng.
    - **Order ID:** Mã đặt hàng.
    - **Product ID:** Mã sản phẩm.
    - **Quantity Ordered:** Số lượng sản phẩm được đặt hàng.
    - **Total Retail Price for This Order:** Tổng giá bán lẻ cho cả đơn hàng.
    - **Cost Price Per Unit:** Giá vốn thu được của mỗi món hàng.
  + Qua việc đánh giá bộ dữ liệu trên, nhóm em có một vài nhận định như sau: Đối với mỗi bản ghi trong bộ dữ liệu chỉ đánh giá cụ thể đối với một mặt hàng thông qua mã sản phẩm và bộ dữ liệu này cũng là một bộ dữ liệu đầy đủ không thiếu hay NULL dòng/trường nào.
* **Bộ dữ liệu *product-supplier.csv*:**
  + Là bộ dữ liệu mô tả thông tin chi tiết về các mặt hàng được mua bán.
  + Trong đó, mỗi bản ghi (record) thể hiện thông tin chi tiết về từng thuộc tính của mặt hàng như mã mặt hàng, tên mặt hàng, nhóm mặt hàng, phân loại, dòng sản phẩm, nước phân phối, tên nhà phân phối, mã nhà phân phối.
  + Các thuộc tính trong bộ dữ liệu được mô tả như sau:
    - **Product ID:** Mã sản phẩm
    - **Product Line:** Dòng sản phẩm
    - **Product Category:** Danh mục sản phẩm
    - **Product Group:** Nhóm sản phẩm
    - **Product Name:** Tên sản phẩm
    - **Supplier Country:** Nước phân phối
    - **Supplier Name:** Tên nhà phân phối
    - **Supplier ID:** Mã nhà phân phối
  + Qua việc đánh giá bộ dữ liệu trên, nhóm em có một vài nhận định như sau: Bộ dữ liệu trên mô tả rất đầy đủ thông tin chi tiết của từng đối tượng mặt hàng. Do đó có thể thấy rõ tầm quan trọng không thể thiếu của nó đối trong việc phân tích dữ liệu bởi vì ta có thể kết hợp với bộ dữ liệu *order.csv* trên thì chúng ta có thể tìm ra các thông tin và các mối liên hệ giữa các mặt hàng với nhau. Ví dụ: mặt hàng bán chạy, lợi nhuận mặt hàng nào nhiều nhất ….
* Đối với bộ dữ liệu *orders.csv* gồm 185004 bản ghi và 9 trường thuộc tính.
* Đối với bộ dữ liệu *product-supplier.csv* gồm 5505 bản ghi và 8 trường thuộc tính.

Đánh giá tổng thể cho thấy đây là một bộ dữ liệu khá đầy đủ cho việc phân tích dữ liệu. Ngoài ra, hướng đi chính cho quá trình phân trình phân tích dữ liệu trong đồ án môn học này sẽ là tìm ra các bộ luật nhằm xác định các mối liên hệ giữa mặt hàng và người dùng.

## **1.2. Tiền xử lí dữ liệu**