BÀI 8: Exploratory Data Analysis

Mục tiêu chính: Cung cấp cho học viên kiến thức và kỹ năng EDA

8.1. Cars

- ✓ **Dữ liệu**: Tập dữ liệu dữ liệu ghi nhận thông tin tổng quan về các mẫu xe ô tô của các hãng khác nhau trên toàn thế giới. Trong đó cột **MSRP** là giá bán lẻ đề xuất của chính hãng (manufacturer's suggested retail price) (Dữ liệu **cars_m.csv** được đính kèm trong file **data**).

- ✓ **Yêu cầu:**
  1. Lọc dữ liệu từ tập tin **cars_m.csv**
     a. Xem 5 dòng dữ liệu đầu tiên (head)
     b. Xem 5 dòng dữ liệu cuối cùng (tail)
     c. Thống kê dữ liệu bằng .describe()
     d. Xem thông tin info()
     e. Kiểm tra dữ liệu NULL
     f. Kiểm tra kiểu dữ liệu của các cột
  2. Tìm hiểu các biến

a. Xóa dữ liệu trùng lặp

b. Phân tích đơn biến

i. Biến phân loại

ii. Biến liên tục

3. Phát hiện và xử lý Outlier
4. Xử lý dữ liệu thiếu
5. EDA và trực quan hóa

a. Mối quan hệ giữa các biến liên tục và **MSRP**

b. Mối quan hệ giữa các biến phân loại và **MSRP**

6. Phân tích chi tiết và đặt vấn đề

8.2. Supermarket Sales

✓ **Dữ liệu**: Tập dữ liệu ghi nhận thông tin bán hàng của một công ty, có chi nhánh tại 3 thành phố. Bộ dữ liệu bao gồm 2000 dòng dữ liệu, cung cấp các thuộc tính như mã hóa đơn (invoice_id), chi nhánh (branch), … và nhiều thông tin khác (Dữ liệu **supermarket_sales.csv** được đính kèm trong file **data**).

✓ **Yêu cầu:**

1. Lọc dữ liệu từ tập tin **supermarket_sales.csv**

a. Xem 5 dòng dữ liệu đầu tiên (head)

b. Xem 5 dòng dữ liệu cuối cùng (tail)

c. Thống kê dữ liệu bằng .describe()

d. Xem thông tin info()

e. Kiểm tra dữ liệu NULL

f. Kiểm tra kiểu dữ liệu của các cột

2. Tìm hiểu các biến

a. Xóa dữ liệu trùng lặp

b. Phân tích đơn biến

i. Biến phân loại

ii. Biến liên tục

3. Phát hiện và xử lý Outlier
4. Xử lý dữ liệu thiếu
5. EDA và trực quan hóa

a. Mối quan hệ giữa các biến liên tục và **profit**

b. Mối quan hệ giữa các biến phân loại và **profit**

6. Phân tích chi tiết và đặt vấn đề

8.3. Titanic

✓ **Dữ liệu**: Tập dữ liệu **titanic** (File dữ liệu **titanic.csv** được chứa trong file **data**).

✓ **Yêu cầu:**

1. Chọn và sử dụng package EDA để có cái nhìn ban đầu về dữ liệu, đồng thời đưa ra một số nhận xét.
2. Thực hiện các công việc sau:
    a. Xác định các thuộc tính
    b. Phân tích đơn biến: Để dự đoán một khách hàng còn sống (1) hay đã qua đời (0) trên chuyến tàu Titanic chúng ta cần các thông tin trong dữ liệu, hãy chọn một số thuộc tính và phân tích
    c. Phân tích hai biến
    d. Xử lý dữ liệu trùng, thiếu
    e. Phát hiện và xử lý ngoại lệ

8.4. Bank Marketing

✓ **Dữ liệu**: Tập dữ liệu Banking Marketing chứa thông tin về các chiến dịch tiếp thị trực tiếp (các cuộc gọi điện thoại) của một tổ chức ngân hàng Bồ Đào Nha. Mục tiêu phân loại là để dự đoán liệu khách hàng sẽ đăng ký một khoản tiền gửi có kỳ hạn hay không (y). (File dữ liệu **bank-additional** được chứa trong file **data**).

✓ **Yêu cầu:**

1. Sử dụng package EDA để có cái nhìn ban đầu về dữ liệu, đồng thời đưa ra một số nhận xét.
2. Thực hiện các công việc sau:
    a. Xác định các thuộc tính
    b. Phân tích đơn biến: để dự đoán một khách hàng sẽ đăng ký (subscribe =yes/no) cho một khoản tiền gửi có kỳ hạn (term deposit - variable y) hay không chúng ta cần các thông tin trong dữ liệu, hãy chọn một số thuộc tính và phân tích
    c. Phân tích hai biến
    d. Xử lý dữ liệu trùng, thiếu