Đề tài : Phân tích dữ liệu giao dịch của một siêu thị để đánh giá hiệu quả chương trình khuyến mãi.

**Bước 1**

Lấy dữ liệu trên ChapGPT

**Bước 2**

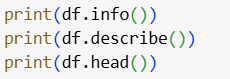
Đường dẫn đến file :



Đọc file :

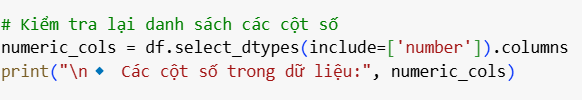


In file



Đoạn code này giúp kiểm tra tổng quan về dữ liệu trong DataFrame df bằng cách sử dụng ba phương thức quan trọng của pandas.

**Bước 3 EDA**

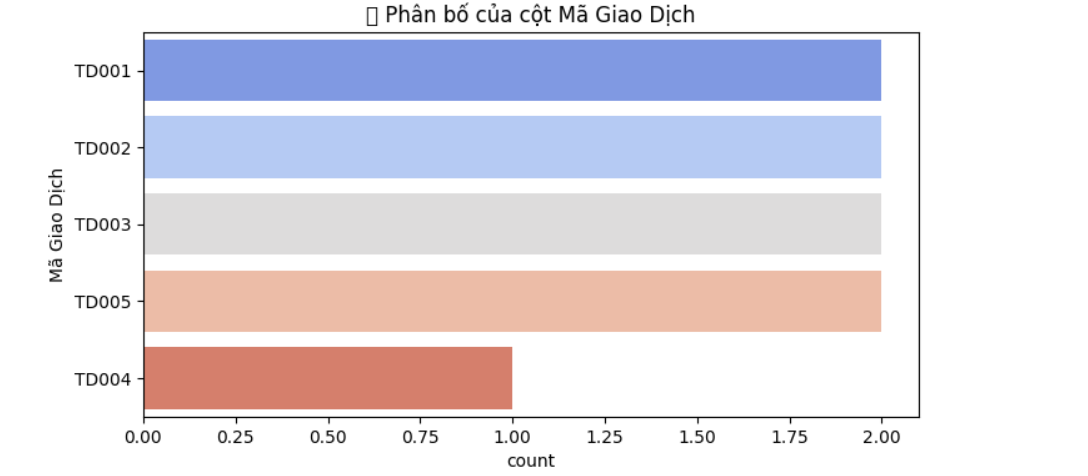


Đoạn code này kiểm tra và hiển thị danh sách các cột có kiểu dữ liệu số trong DataFrame df.

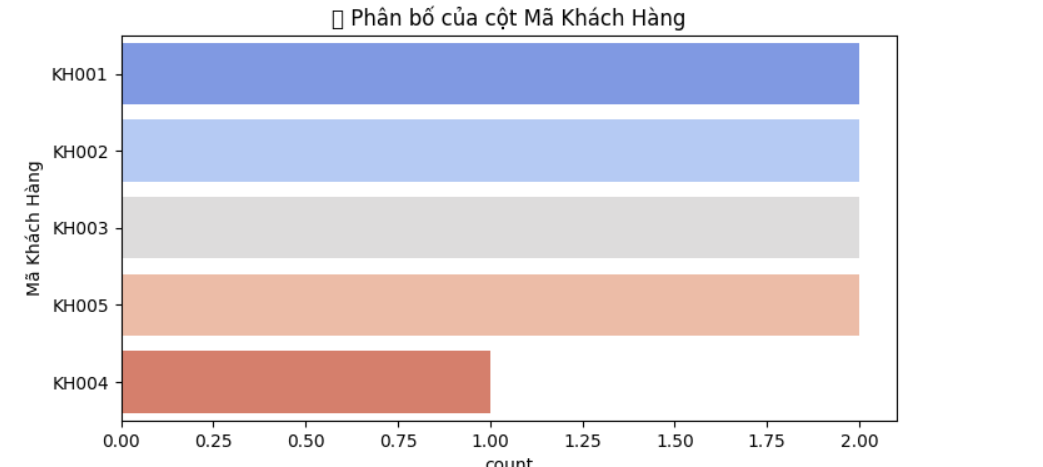
In ra các cột

Đoạn code giúp kiểm tra nhanh các cột chứa dữ liệu số trong DataFrame, hữu ích khi cần phân tích hoặc xử lý dữ liệu số sau này.

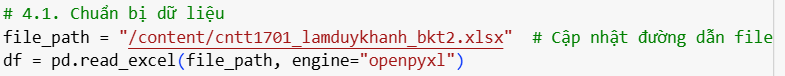
**Cột Mã Giao Dịch**



**Cột Mã Khách Hàng**



**Phần 4 Chuẩn bị dữ liệu**



Đoạn code này giúp đọc dữ liệu từ một tệp **Excel (.xlsx)** vào một **DataFrame** (df) trong **pandas**, sau đó có thể xử lý, phân tích hoặc trực quan hóa dữ liệu.



Chuyển đổi cột "Tổng Tiền (VND)" và "Giá Sau KM (VND)" từ kiểu chuỗi (string) sang số thực (float), đồng thời loại bỏ dấu phẩy (,) nếu có.

Chuyển đổi cột "Số Lượng" sang kiểu số nguyên (int).

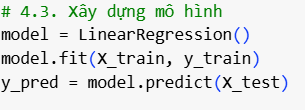
Chuyển các cột phân loại "Phương Thức Thanh Toán" và "Danh Mục" thành các cột one-hot encoding.

Sử dụng drop\_first=True để tránh trường hợp đa cộng tuyến (giữ n-1 biến thay vì n biến).

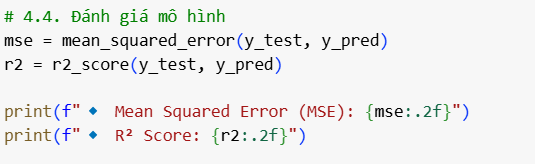
**Chia sẻ dữ liệu**   
X = df[["Số Lượng", "Giá Sau KM (VND)"] + list(df.columns[df.columns.str.startswith("Phương Thức Thanh Toán\_")]) + list(df.columns[df.columns.str.startswith("Danh Mục\_")])]

y = df["Tổng Tiền (VND)"]

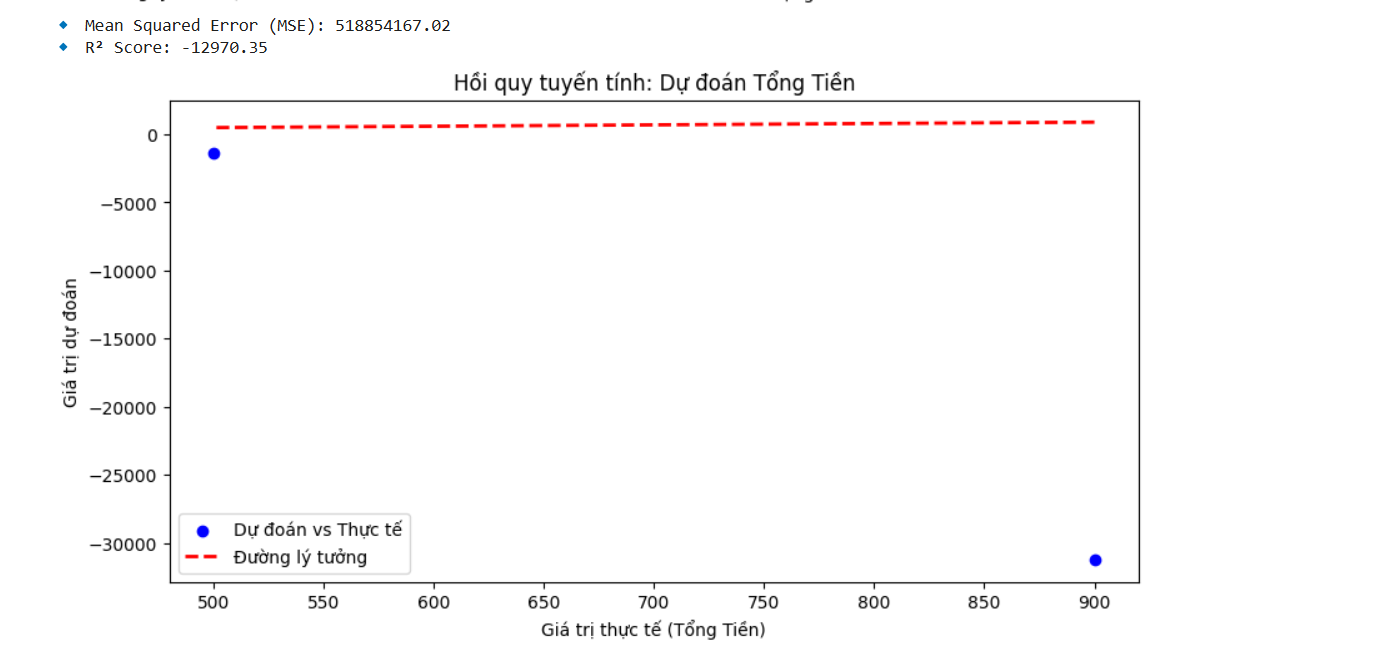
X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

**Xây dựng mô hình :**   


**Đánh giá mô hình**



Và biểu đồ hồi quy tuyến tính



Link Colab

<https://colab.research.google.com/drive/16dyoaqUREP-vS3LJjagb95mSwNz8rb9o#scrollTo=YwrkUM1BfVZ7>

Link GitHub