**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HCM**

**KHOA CÔNG NGHỆ THÔNG TIN**

🙠🙟🕮🙝🙢



**MÔN HỌC: KHAI PHÁ DỮ LIỆU**

**ĐỒ ÁN CUỐI KÌ**

**ĐỀ TÀI:**

**GVHD**: Nguyễn Văn Thành

**Nhóm sinh viên thực hiện:** Nhóm 3

1. Nguyễn Trí Dũng 20133029
2. Nguyễn Khoa Quang Thắng 20133090
3. Văn Mai Thanh Nhật 201330xx
4. Huỳnh Minh Phước 201330xx

**Mã môn học**: DAMI330484\_22\_2\_01

Thành phố Hồ Chí Minh, Tháng 5 năm 2023

**Lời mở đầu**

Khai phá dữ liệu là một tập hợp các kỹ thuật được sử dụng để tự động khai thác và tìm ra các mối quan hệ lẫn nhau của dữ liệu trong một tập hợp dữ liệu khổng lồ và phức tạp, đồng thời cũng tìm ra các mẫu tiềm ẩn trong tập dữ liệu đó. Nhận thấy được tầm quan trọng của môn học, nhóm 3 thực hiện đề tài “Khai phá tập dữ liệu giao dịch e-commerce của khách hàng”, với mục đích tìm được mối quan hệ với nhau giữa các khách hàng thông qua các giao dịch, đồng thời tìm định hướng, chiến lược phát triển bán hàng cho website. Vì ngày nay, thông qua e-commerce, các doanh nghiệp có thể tiếp cận và thiết lập sự hiện diện trên thị trường rộng lớn hơn bằng cách cung cấp các kênh phân phối rẻ hơn và hiệu quả hơn cho các sản phẩm hoặc dịch vụ của họ. E-commerce cũng đã thay đổi cách mọi người mua sắm và tiêu thụ sản phẩm và dịch vụ. Nhiều khách hàng đang chuyển sang máy tính hoặc thiết bị thông minh của họ để đặt hàng, những thứ có thể dễ dàng được giao đến tận nhà của họ.

**Mục lục**

# PHẦN 1 – GIỚI THIỆU:

Nhóm sẽ sử dụng vẽ biểu đồ và đưa ra nhận xét, thêm vào đó,thuật toán gom cụm, phương pháp số học, và thuật toán luật kết hợp sẽ được sử dụng để thực hiện đề tài.

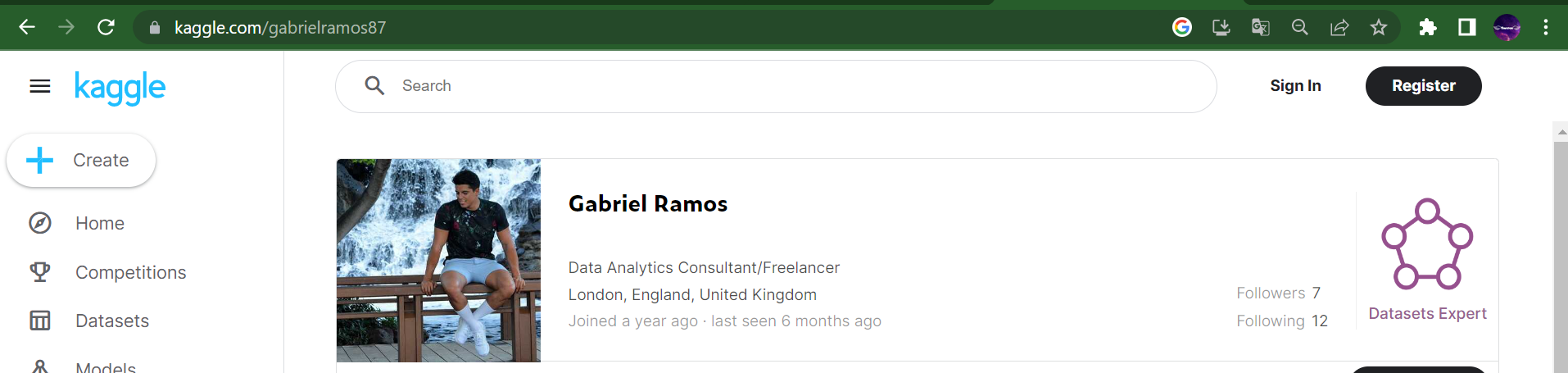
Các công cụ được dùng trong đề tài:

* Ngôn ngữ Python: Python là một ngôn ngữ lập trình được sử dụng rộng rãi trong các ứng dụng web, phát triển phần mềm, khoa học dữ liệu và máy học (ML). Các nhà phát triển sử dụng Python vì nó hiệu quả, dễ học và có thể chạy trên nhiều nền tảng khác nhau. Phần mềm Python được tải xuống miễn phí, tích hợp tốt với tất cả các loại hệ thống và tăng tốc độ phát triển.
* Visual studio code: Visual Studio Code là một trình soạn thảo mã nguồn được phát triển bởi Microsoft dành cho Windows, Linux và macOS. Nó hỗ trợ chức năng debug, đi kèm với Git, có chức năng nổi bật cú pháp (syntax highlighting), tự hoàn thành mã thông minh, snippets, và cải tiến mã nguồn.
* Jupyter là một nền tảng tính toán khoa học mã nguồn mở, với khả năng nổi bật cho phép tương tác trực tiếp với từng dòng code (interactive), hỗ trợ hơn 40 ngôn ngữ lập trình, trong đó tập trung vào 3 ngôn ngữ là Julia, Python và R, cái tên Jupyter bắt nguồn từ cách chơi chữ I Python, You R, We Julia, hay một công thức ngắn gọn Jupyter = Julia + Python + R. Bên cạnh đó, Jupyter cũng là một công cụ hoàn toàn miễn phí, được tạo ra với mục đích nhắm đến khoa học dữ liệu và giáo dục, giúp mọi người cùng học lập trình dễ dàng hơn (cụ thể ở đây là Python). Jupyter có tính tương tác nên có thể sử dụng làm môi trường chạy thử và giảng dạy.

# PHẦN 2 – DỮ LIỆU

### 3.1) Nguồn gốc của tập dữ liệu:

Trong đề tài này, dataset được nhóm 3 thống nhất sử dụng được lấy từ Kaggle và tác giả của tập dữ liệu là Gabriel Ramos.



Theo tác giả, đây là bộ dữ liệu giao dịch bán hàng của thương mại điện tử (bán lẻ trực tuyến) có trụ sở tại Vương quốc Anh trong một năm. Cửa hàng có trụ sở tại Luân Đôn này đã bán quà tặng và đồ gia dụng cho người lớn và trẻ em thông qua trang web từ năm 2007. Khách hàng của họ đến từ khắp nơi trên thế giới và thường mua hàng trực tiếp cho chính họ. Ngoài ra còn có các doanh nghiệp nhỏ mua với số lượng lớn và bán cho các khách hàng khác thông qua các kênh đại lý bán lẻ.

### 3.2) Thông tin về tập dữ liệu:

**Kích thước của tập dữ liệu:** 536350 dòng x 8 cột

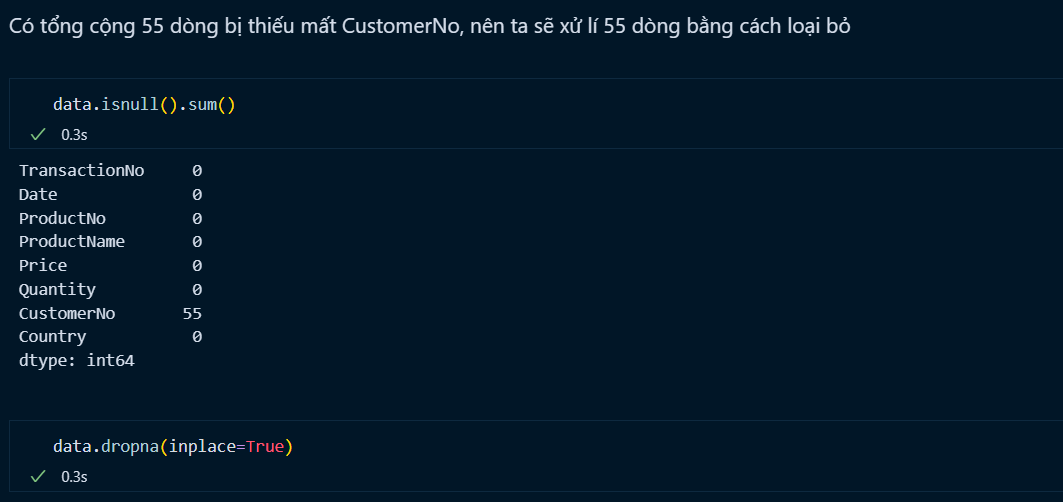
**Các biến của tập dữ liệu:**

1. ***TransactionNo (categorical):*** a six-digit unique number that defines each transaction. The letter “C” in the code indicates a cancellation.
2. ***Date (numeric):*** the date when each transaction was generated.
3. ***ProductNo (categorical):*** a five or six-digit unique character used to identify a specific product.
4. ***Product (categorical):*** product/item name.
5. Price (numeric): the price of each product per unit in pound sterling (£).
6. ***Quantity (numeric):*** the quantity of each product per transaction. Negative values related to cancelled transactions.
7. ***CustomerNo (categorical):*** a five-digit unique number that defines each customer.
8. ***Country (categorical):*** name of the country where the customer resides.

### 3.2) Tiền xử lí:

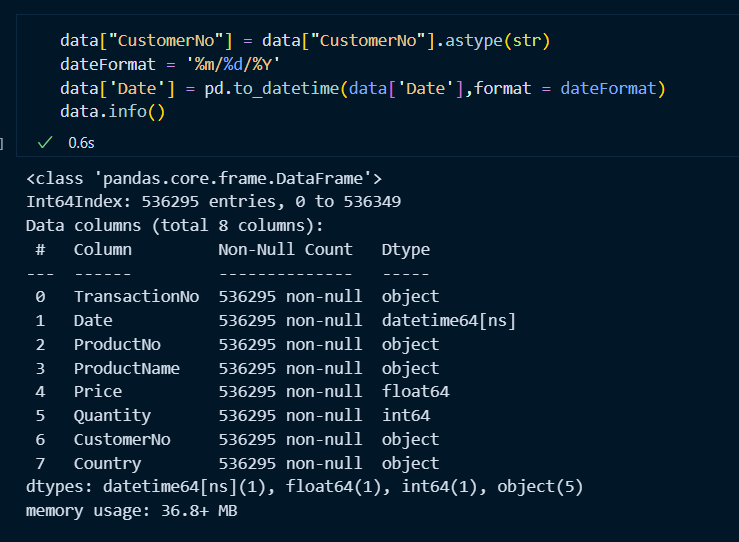
#### 3.2.1) Xử lí những giá trị bị thiếu:

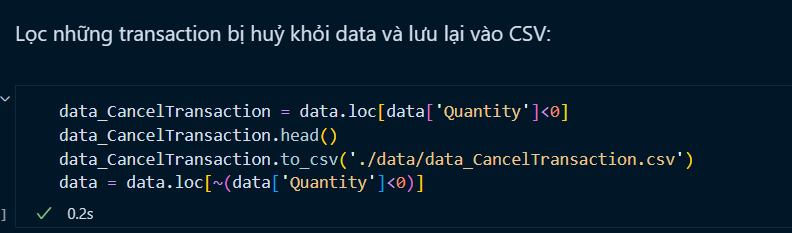
Tập dữ liệu có 55 dòng bị thiếu dữ liệu ở CustomerNo, nên do số lượng quá ít so với tập dữ liệu, nên nhóm quyết định bỏ những dòng bị thiếu dữ liệu CustomerNo.



#### 3.2.2) Chuyển đổi kiểu dữ liệu của tập dữ liệu:

Chuyển đổi dữ liệu của tập dữ liệu, CustomerNo chuyển từ int thành string và Date từ string thành datetime.



Do tập data gốc có chứa những transaction bị huỷ, nhóm sẽ loại bỏ để không ảnh hưởng tới phần tính toán.

#### 3.2.3) Xử lí outlier:

a) Giới thiệu thuật toán:

**Sử dụng phương pháp số học z-score để tìm outlier:**

Đối phương pháp này, ta cần phải giả định là dữ liệu của chúng ta tuân theo phân phối chuẩn.

Trong phân phối chuẩn, giả sử μ là kỳ vọng và σ là độ lệch chuẩn. Quy tắc 3σ cho phân phối chuẩn nói rằng:

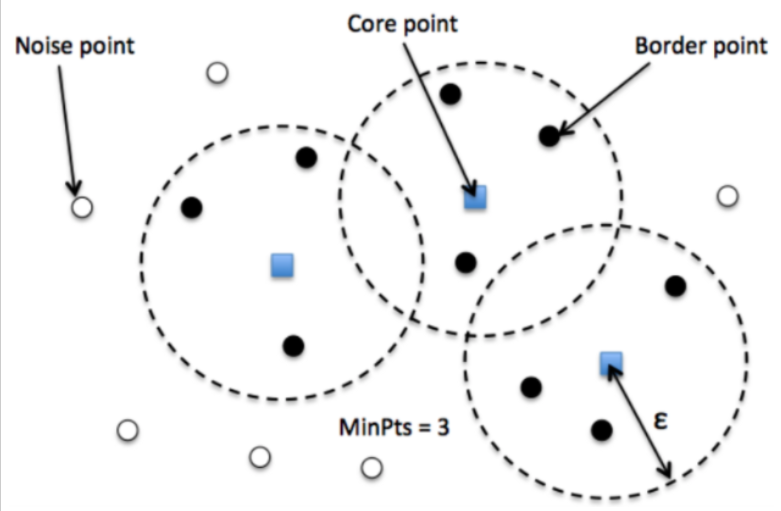
* 68% các điểm dữ liệu nằm trong khoảng μ±σ
* 95% các điểm dữ liệu nằm trong khoảng μ±2σ
* 99.7% các điểm dữ liệu nằm trong khoảng μ±3σ

Với một điểm dữ liệu x, z score của nó được tính bởi:

Vì vậy, việc các điểm nằm ngoài đoạn [μ−3σ, μ+3σ] được coi là các điểm ngoại lệ.

**Sử dụng density-based method (DBSCAN):**

Kĩ thuật này này đề cập đến các phương pháp học không giám sát nhằm xác định các cụm phân biệt trong phân phối của dữ liệu, dựa trên ý tưởng rằng một cụm trong không gian dữ liệu là một vùng có mật độ điểm cao được ngăn cách với các cụm khác bằng các vùng liền kề có mật độ điểm thấp. DBSCAN là một thuật toán cơ sở để phân nhóm dựa trên mật độ. Nó có thể phát hiện ra các cụm có hình dạng và kích thước khác nhau từ một lượng lớn dữ liệu chứa outlier.

Phân loại dạng điểm trong DBSCAN: Ta sẽ chia chúng thành ba loại: Đối với các điểm nằm sâu bên trong cụm chúng ta xem chúng là điểm lõi. Các điểm biên nằm ở phần ngoài cùng của cụm và điểm outlier không thuộc bất kì một cụm nào. Bên dưới là hình vẽ mô phỏng thể hiện ba loại điểm tương ứng nêu trên.

* ***minPts:*** Là một ngưỡng số điểm dữ liệu tối thiểu được nhóm lại với nhau nhằm xác định một vùng lân cận epsilon có mật độ cao. Số lượng minPts không bao gồm điểm ở tâm.
* ***epsilon (kí hiệu ε):*** Một giá trị khoảng cách được sử dụng để xác định vùng lân cận epsilon của bất kỳ điểm dữ liệu nào.

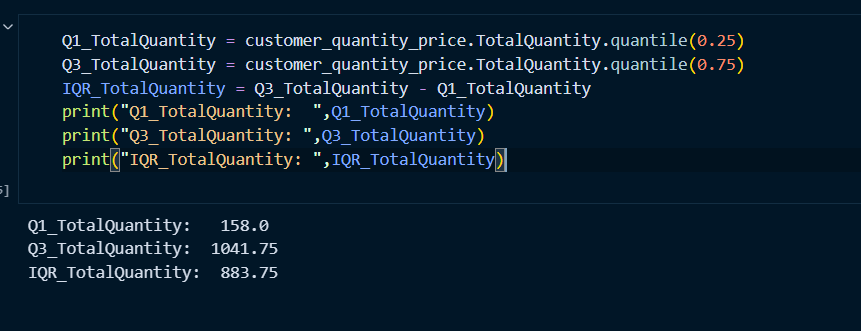
**Sử dụng Elbow method để chọn epsilon và k:**

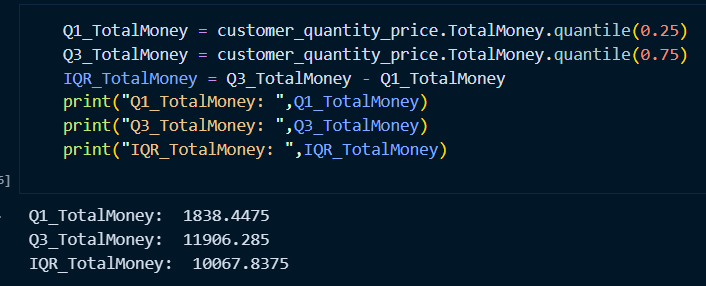
Ta sử dụng phương pháp Elbow, để tìm một giá trị phù hợp cho eps, chúng ta có thể vẽ đồ thị khoảng cách kNN của các điểm (nghĩa là khoảng cách của mỗi điểm đến điểm lân cận gần thứ k của nó) theo thứ tự giảm dần và tìm kiếm một elbow trong đồ thị. Ý tưởng đằng sau là các điểm nằm bên trong các cụm sẽ có khoảng cách k-láng giềng gần nhất nhỏ, bởi vì chúng ở gần các điểm khác trong cùng một cụm.

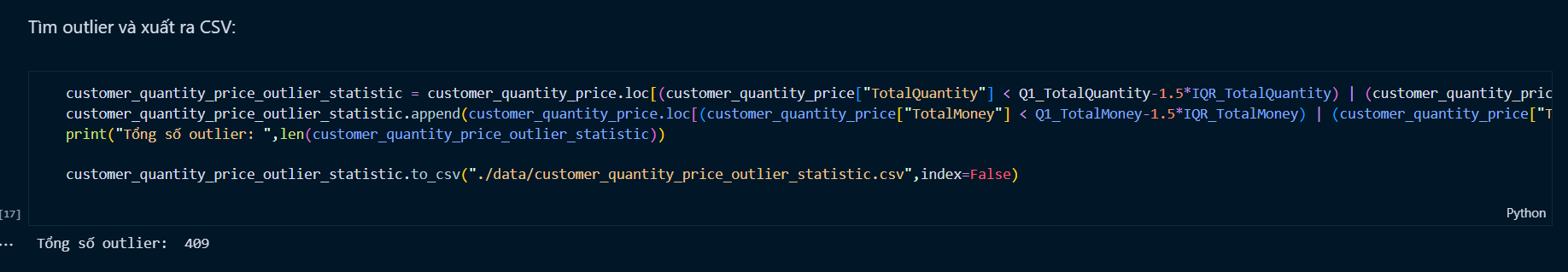
b) Thực hiện trên tập dữ liệu:

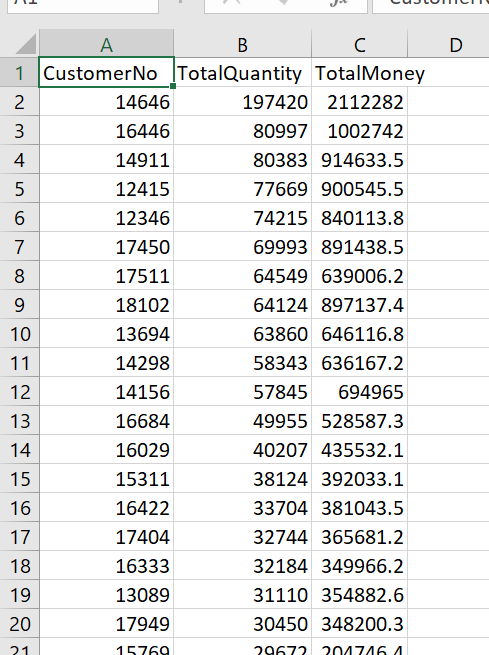
***Đối với outlier của các khách hàng:***

*Sử dụng phương pháp số học z-score để tìm outlier:*

* *Bước 1:* Tính tổng lượng hàng và tổng tiền của khách hàng đã chi tiêu:
* *Bước 2:* Tính Q1, Q3, IQR.

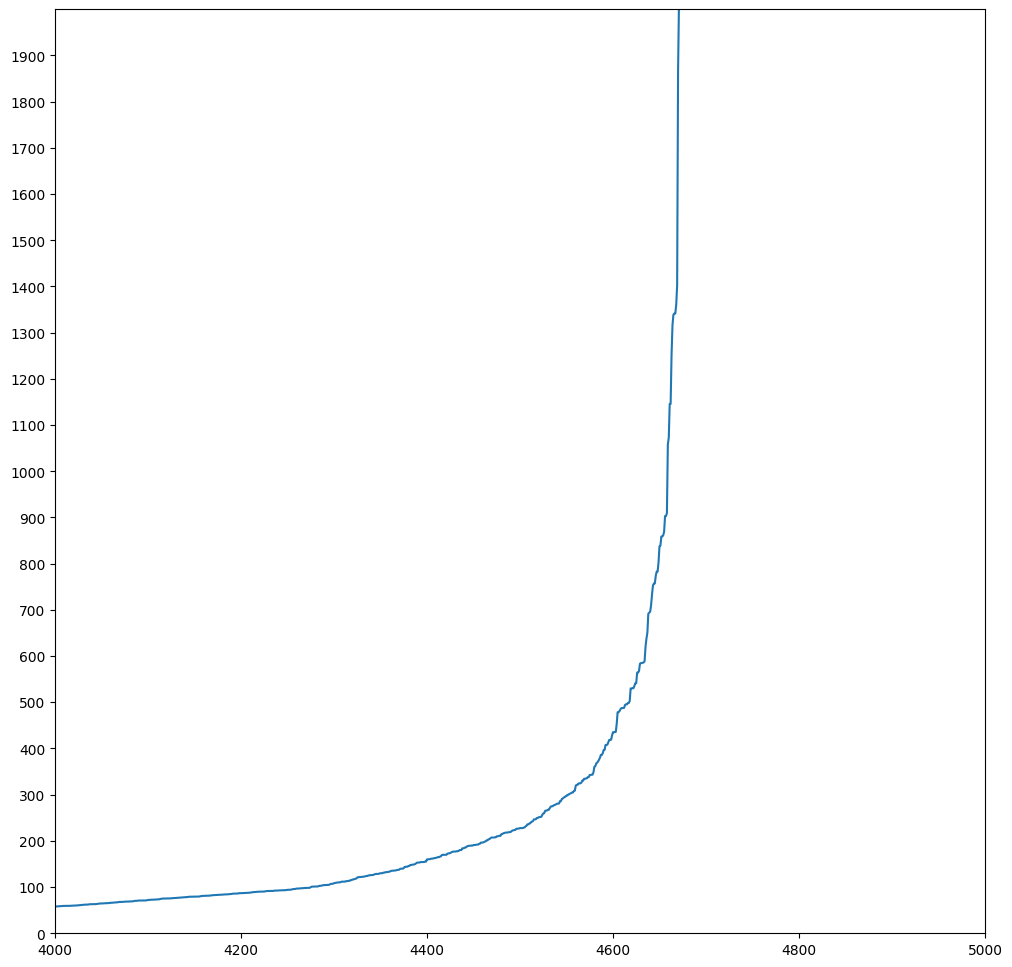


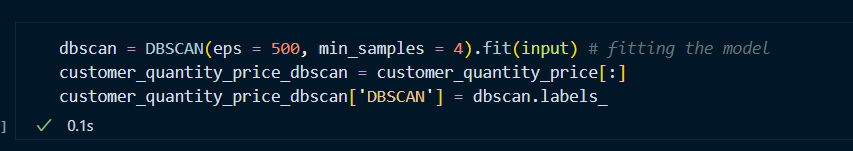
* *Bước 3:Lọc outlier ra khỏi tập dữ liệu và xuất ra csv:*

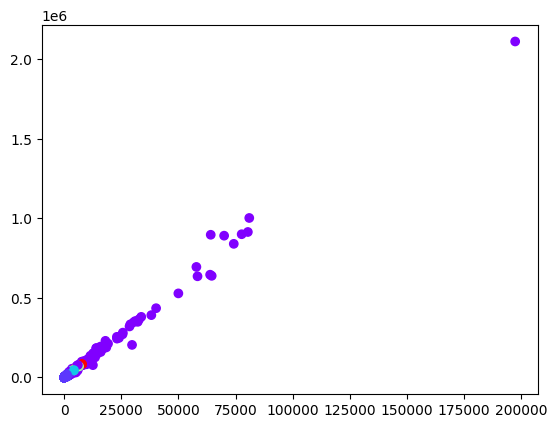


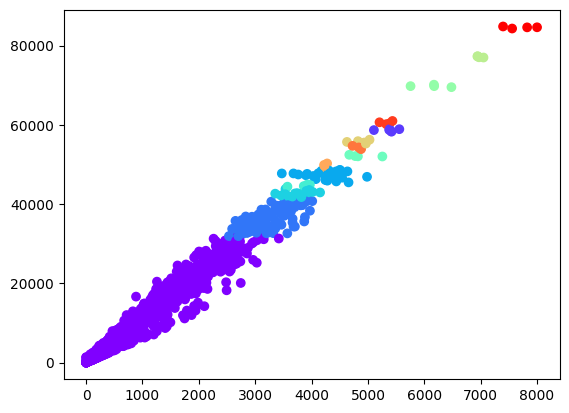
*Sử dụng density-based method (DBSCAN):* 

* Bước 1: Chọn thông số cho DBSCAN:
  + Để chọn minPts, ta có công thức minPts >= Dimension + 1, ta có minPts = 4.
  + Để chọn epsilon: Đối với hình dưới, có thể thấy được rằng tại epsilon khoảng 500 thì điểm biến thiên nhiều nhất.

****

* Bước 2: Chạy thuật toán:

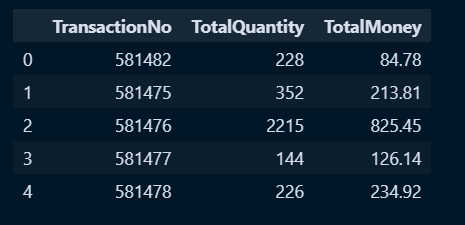
Trước khi bỏ outlier:

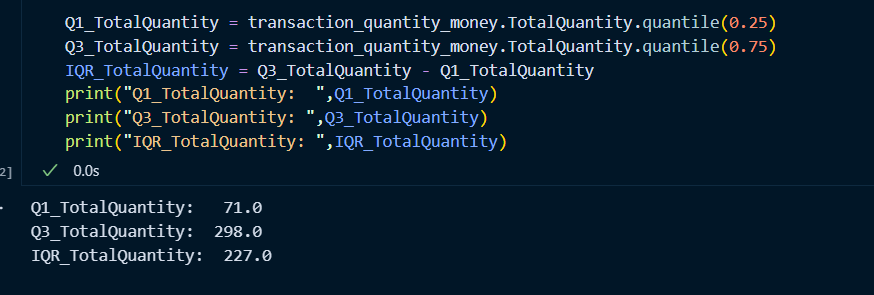
Sau khi bỏ outlier:

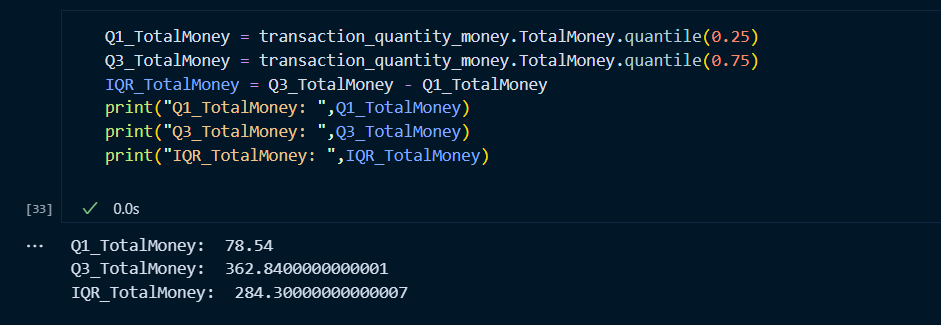
**Đối với tìm outlier của các transaction:**

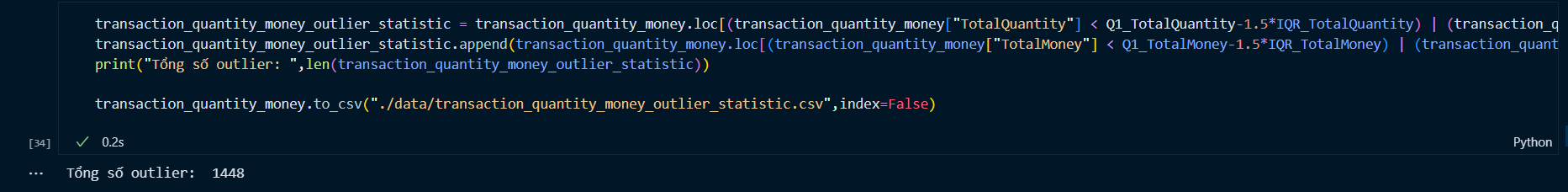
*Sử dụng phương pháp số học z-score để tìm outlier:*

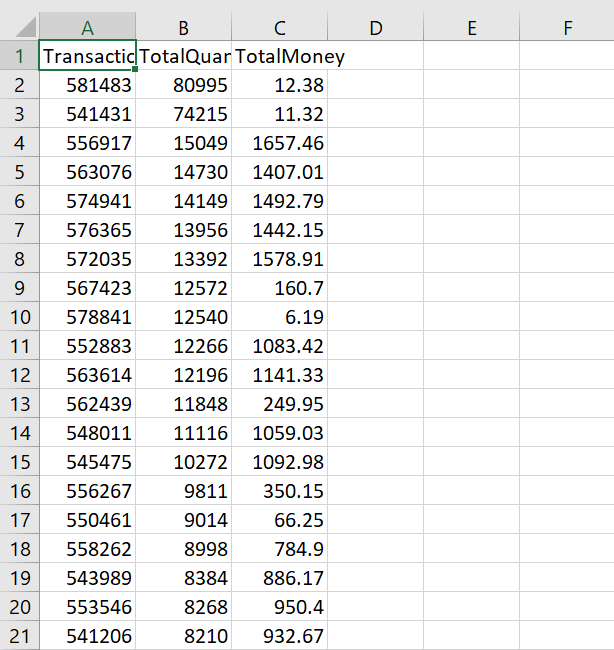
* *Bước 1:* Tính tổng lượng hàng và tổng tiền theo transaction:

**

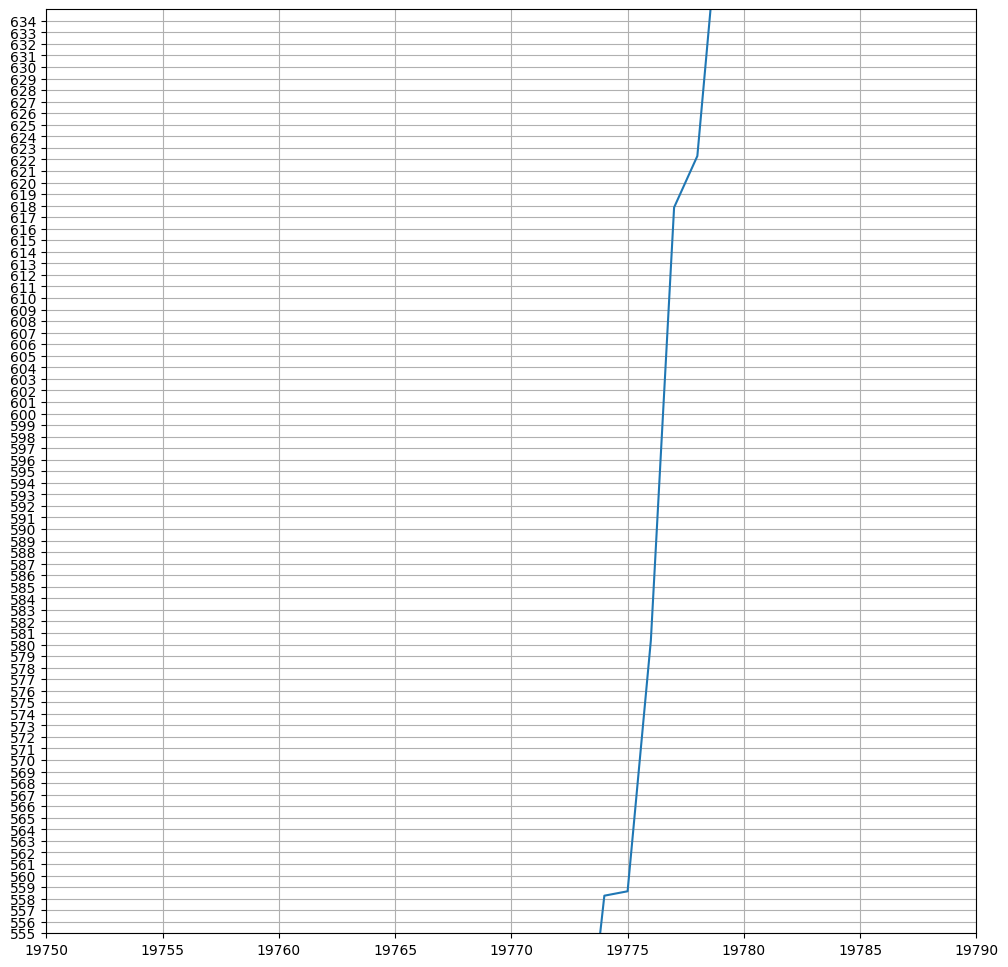
* *Bước 2: Tính Q1, Q3 và IQR:*

****

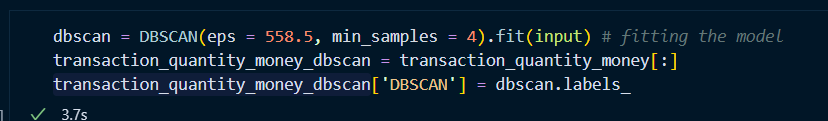
* *****Bước 3:Lọc outlier ra khỏi tập dữ liệu và xuất ra csv:*



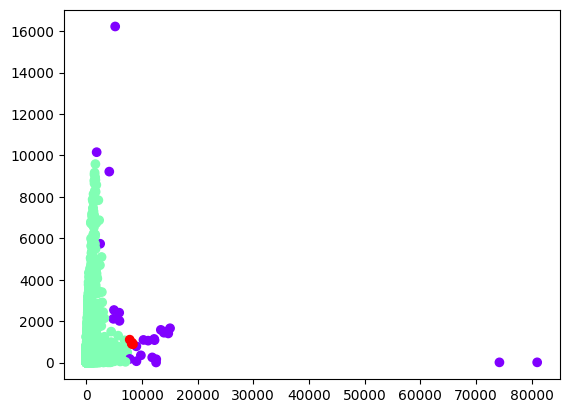
*Sử dụng density-based method (DBSCAN):*

* *Bước 1:*
  + Để chọn minPts, ta có công thức minPts >= Dimension + 1, ta có minPts = 4.
  + **Để chọn epsilon: Đối với hình dưới, có thể thấy được rằng tại epsilon khoảng 558.5 thì điểm biến thiên nhiều nhất.

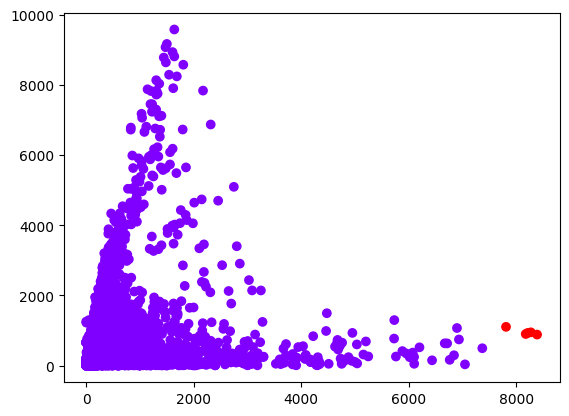
*Bước 2:* Chạy thuật toán:



Trước khi bỏ outlier:



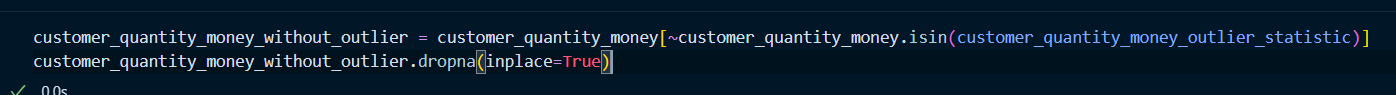
Sau khi bỏ outlier:

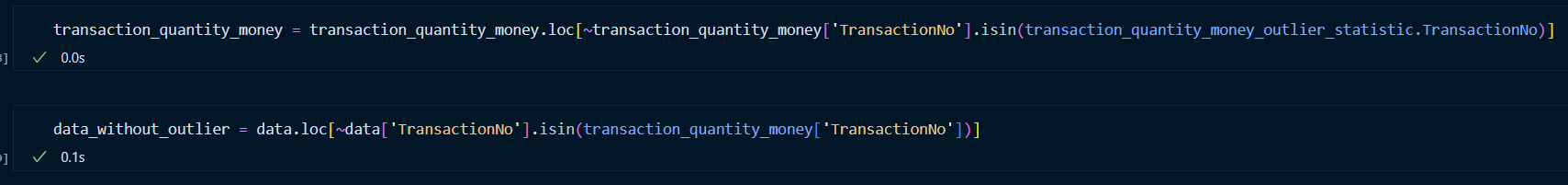


Nhận xét 2 thuật toán phát hiện outlier (DBSCAN và Statistic):

* ABC
* XYZ

c) Loại bỏ outlier sau khi chọn thuật toán:



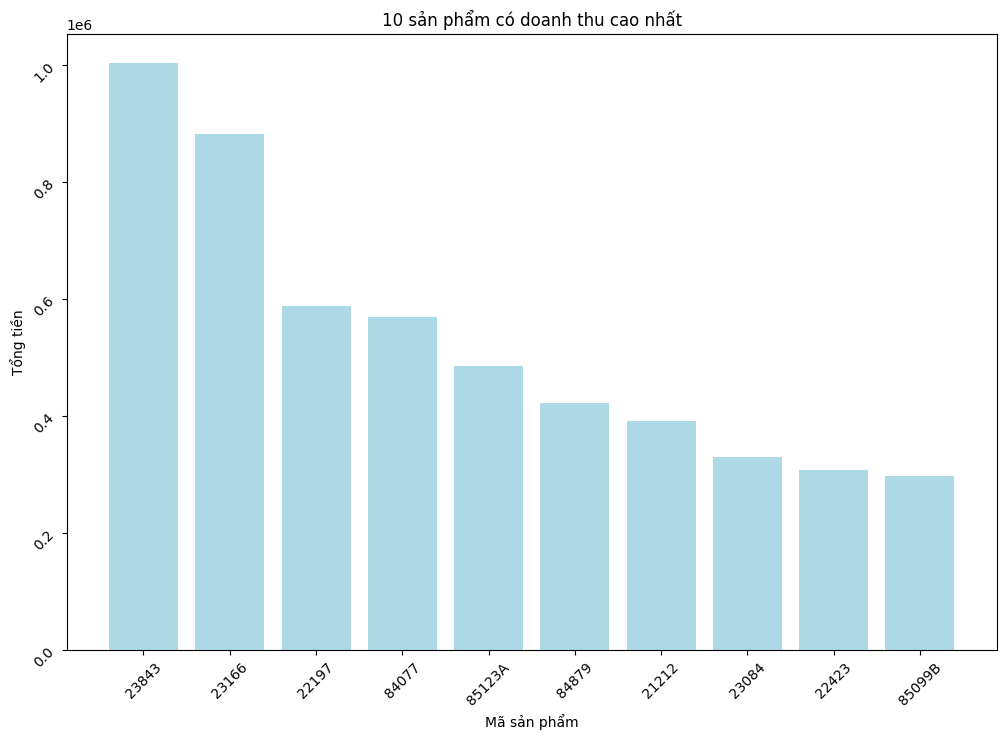


# PHẦN 3 - TRỰC QUAN HOÁ DỮ LIỆU

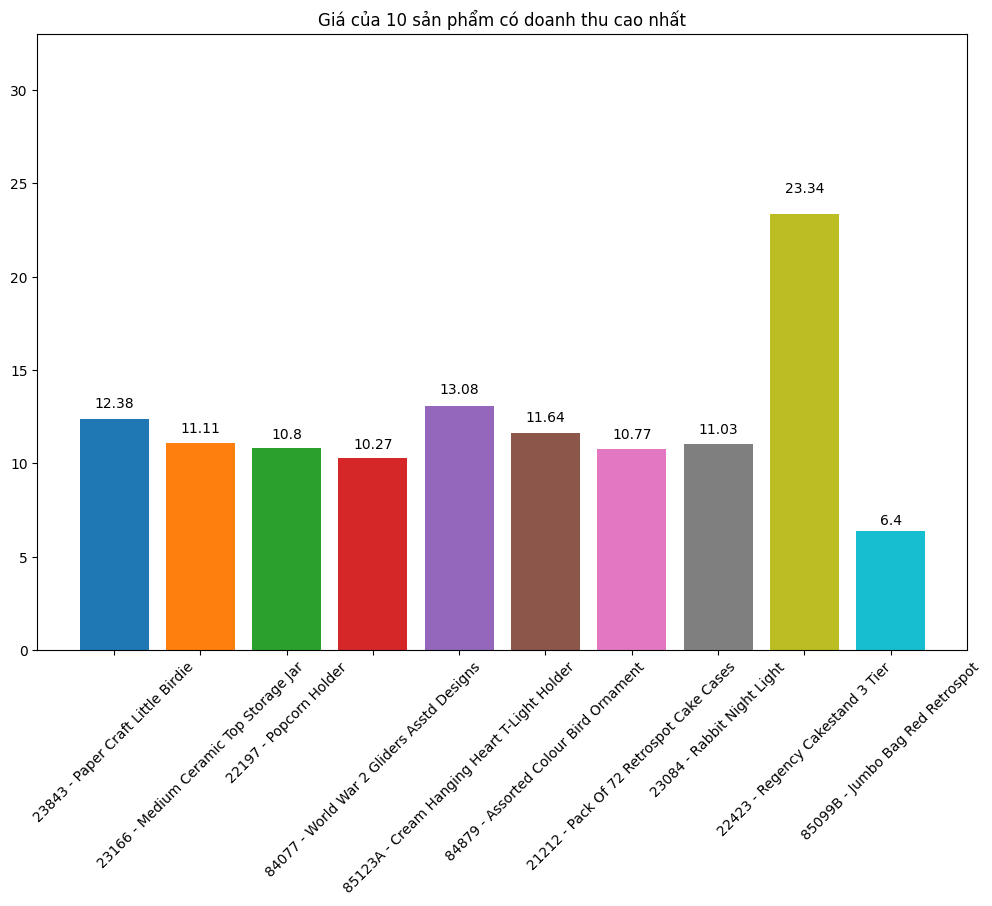
## 4.1) Câu hỏi đặt ra:

## 4.2) 10 giao dịch có tổng thanh toán cao nhất:

## 4.3) 10 sản phẩm có tổng doanh thu cao nhất:

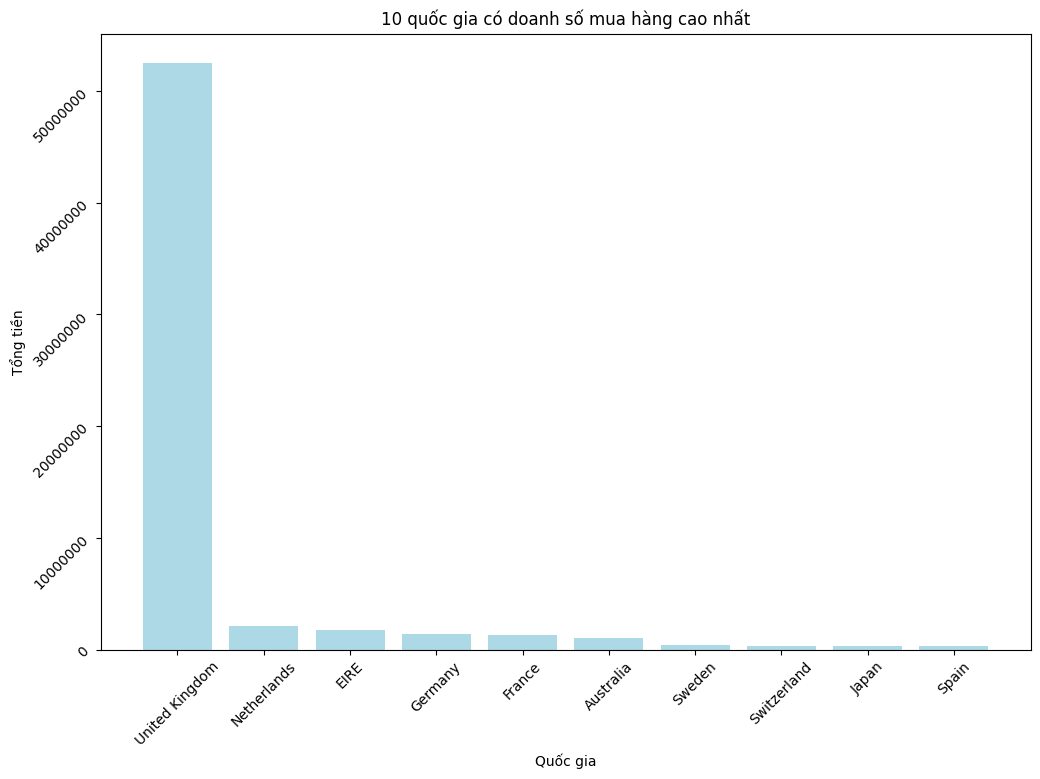


## 4.4) Giá của 10 sản phẩm có doanh thu cao nhất:

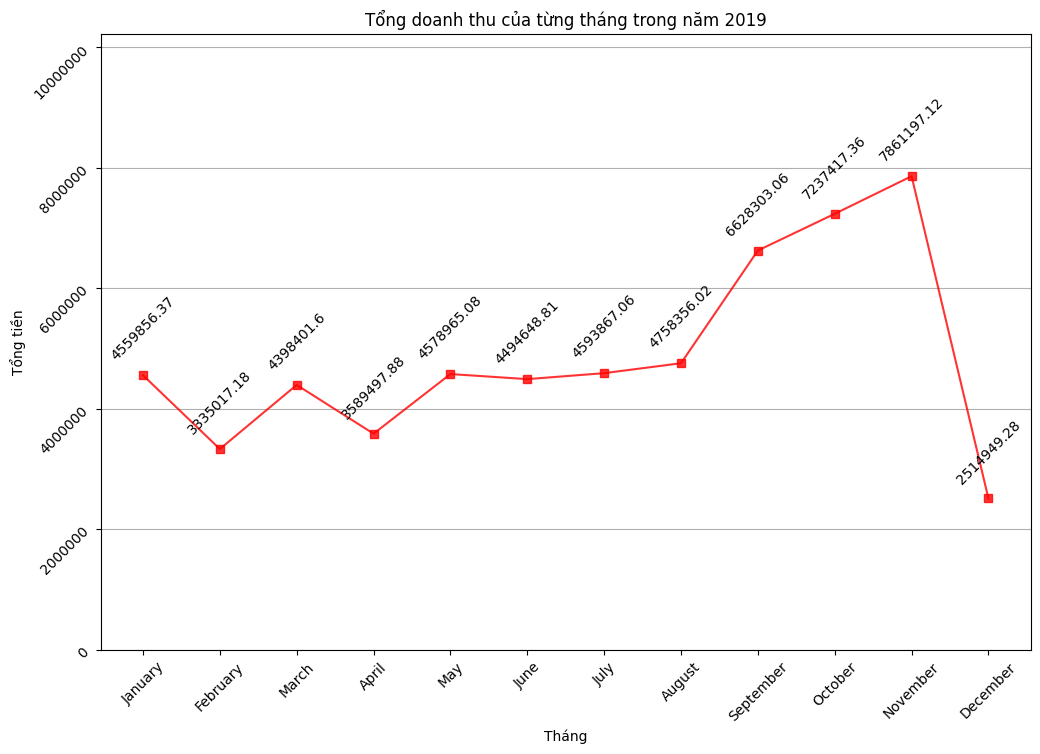


## 4.5) Tổng số lượng bán ra của 10 sản phẩm có doanh thu cao nhất:

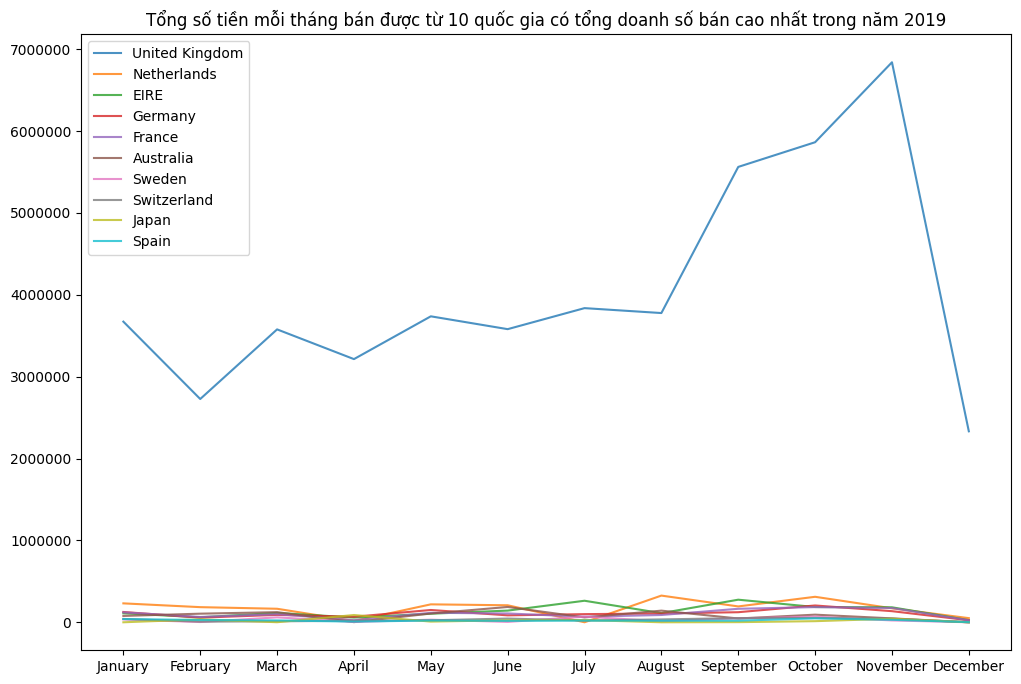
## 4.6) 10 quốc gia có doanh số mua hàng cao nhất:



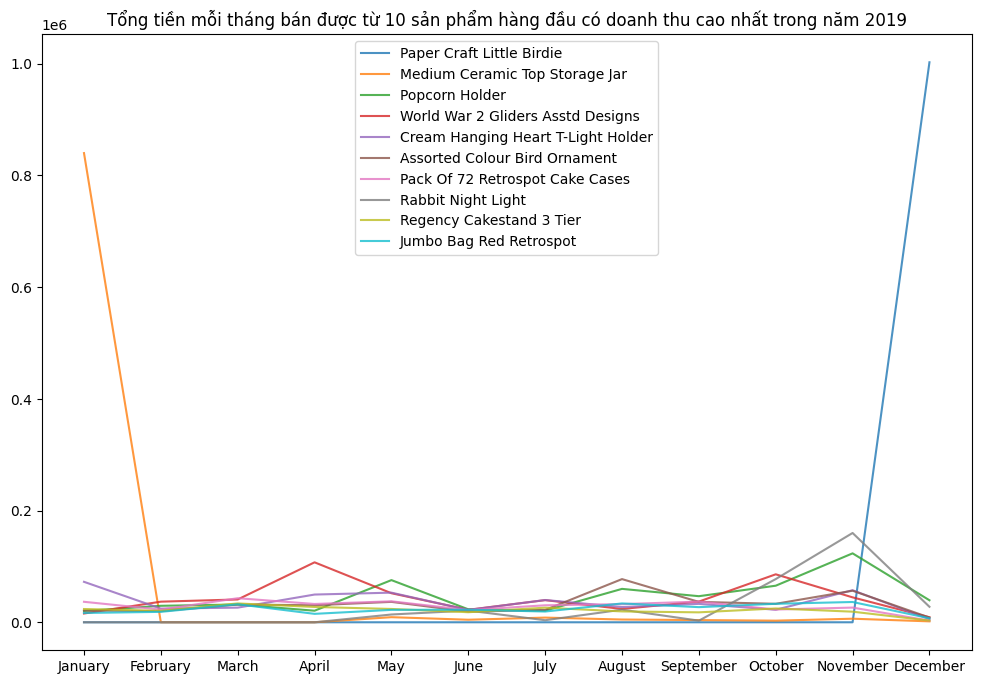
## 4.7) Tổng doanh thu từng tháng trong năm 2019:



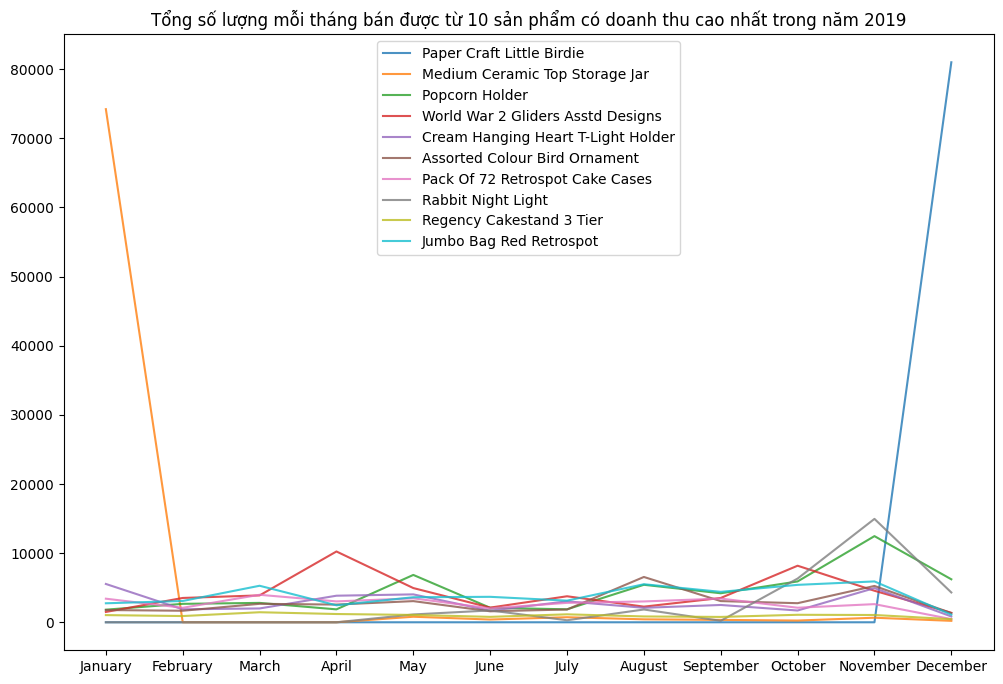
## 4.8) Tổng tiền mỗi tháng bán được từ 10 sản phẩm hàng đầu có doanh thu cao nhất trong năm 2019:



## 4.9) Tổng tiền mỗi tháng bán được từ 10 sản phẩm hàng đầu có doanh thu cao nhất trong năm 2019



## 4.10) Tổng số lượng mỗi tháng bán được từ 10 sản phẩm có doanh thu cao nhất trong năm 2019:



## 4.11) Trả lời câu hỏi đặt ra:

# PHẦN 4 – KHAI PHÁ DỮ LIỆU

## 4.1) Giới thiệu thuật toán:

### 4.1.1) Thuật toán gom cụm:

### 4.1.2) Thuật toán luật kết hợp:

## 4.2) Thực hiện thuật toán:

### 4.2.1) Thuật toán gom cụm:

#### a) Kmeans:

#### b) DBSCAN:

#### c) So sánh 2 thuật toán gom cụm:

### 4.2.2) Thuật toán luật kết hợp:

#### a) Thuật toán FP-Growth:

### b) Thuật toán Apriori:

# PHẦN 5 – KẾT QUẢ, THỰC NGHIỆM VÀ THẢO LUẬN

Nguồn:

<https://codelearn.io/sharing/jupyter-notebook-tutorial>

<http://uet.vnu.edu.vn/~thuyhq/Student_Thesis/K44_Do_Thi_Dieu_Ngoc_Thesis.pdf>

<https://vi.wikipedia.org/wiki/Visual_Studio_Code>

<https://aws.amazon.com/vi/what-is/python/>

dataset: <https://www.kaggle.com/datasets/gabrielramos87/an-online-shop-business?select=Sales+Transaction+v.4a.csv>

Outlier detection:

Phương pháp số học: <https://machinelearningcoban.com/tabml_book/ch_data_processing/process_outliers.html>

DBSCAN:

https://phamdinhkhanh.github.io/deepai-book/ch\_ml/DBSCAN.html