**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.

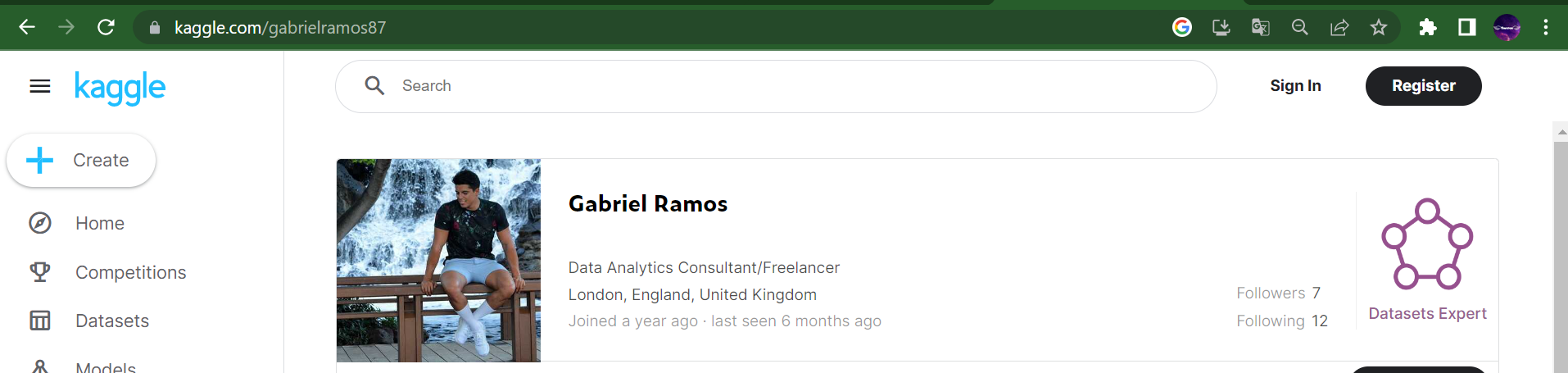
Các công cụ được dùng trong đề tài:

* 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 3 – 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.

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