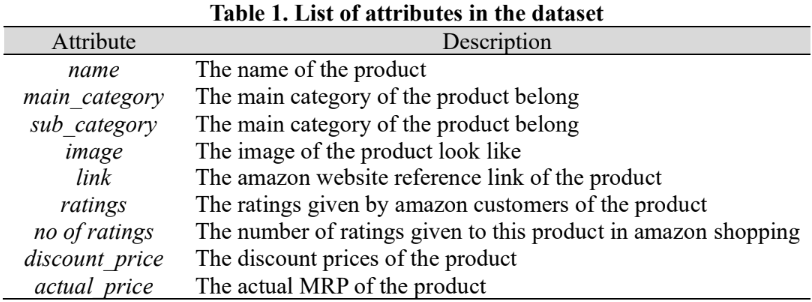
**I. Introduction**Amazon.com, Inc. is an American multinational technology company focusing on e-commerce, cloud computing, online advertising, digital streaming, and artificial intelligence. In this midterm project, we will focus on its e-commerce which is a popular platform that sells many product lines such as media (books, music, movies, software), consumer electronics, gourmet food, kitchen items. With more than 2 billion website visitors per month in 2022, there are a lot of technical questions that the Company needs to solve to enhance the user's experience and its profit. In the aspect of profit, an obvious question is how to decide which products will be included in a promotion campaign, such as Black Friday, and determine the discount percentage of those ones. Therefore, in this midterm project, we will take advantage of Business Intelligence toolkits to improve the activity of decision making based on the data collected from their website.

**[Vietnamese]**Amazon.com, Inc. là một công ty công nghệ đa quốc gia của Mỹ tập trung vào thương mại điện tử, điện toán đám mây, quảng cáo trực tuyến, phát trực tuyến kỹ thuật số và trí tuệ nhân tạo. Trong dự án giữa kì này, chúng tôi sẽ tập trung vào thương mại điện tử, một nền tảng phổ biến bán nhiều dòng sản phẩm như phương tiện truyền thông (sách, nhạc, phim, phần mềm), điện tử tiêu dùng, thực phẩm dành cho người sành ăn, đồ dùng nhà bếp. Với hơn 2 tỷ khách truy cập trang web mỗi tháng vào năm 2022, có rất nhiều câu hỏi kỹ thuật mà Công ty cần giải quyết để nâng cao trải nghiệm của người dùng và lợi nhuận của mình. Ở khía cạnh lợi nhuận, một câu hỏi hiển nhiên là làm thế nào để quyết định sản phẩm nào sẽ được đưa vào chiến dịch khuyến mại, chẳng hạn như Thứ Sáu Đen và xác định tỷ lệ chiết khấu của những sản phẩm đó. Vì vậy, trong dự án giữa kỳ này, chúng tôi sẽ tận dụng bộ công cụ Business Intelligence để cải thiện hoạt động ra quyết định dựa trên dữ liệu được thu thập từ trang web của họ.

**II. Dataset**The Product Sales Dataset, collected from the Amazon website, contains 551,585 instances organized into 142 categories such as accessories, men's clothing, pet supplies. Each instance in the dataset consists of 9 attributes, as presented in Table 1. Fig. 1 shows the distribution of the dataset according to 142 categories. The dataset is available on the course homepage [1]. You should refer [2] to understand how to use Pandas CSV to process the dataset.

**[Vietnamese]**  
Bộ dữ liệu bán hàng sản phẩm, được thu thập từ trang web Amazon, chứa 551.585 trường hợp được sắp xếp thành 142 danh mục như phụ kiện, quần áo nam, đồ dùng cho thú cưng. Mỗi phiên bản trong tập dữ liệu bao gồm 9 thuộc tính, như được trình bày trong Bảng 1. Hình 1 cho thấy sự phân bổ của tập dữ liệu theo 142 danh mục. Bộ dữ liệu có sẵn trên trang chủ của khóa học [1]. Bạn nên tham khảo [2] để hiểu cách sử dụng Pandas CSV để xử lý tập dữ liệu.



A graph with blue lines

Description automatically generated

**III. Questions**The midterm project questions will be organized into two types, the first one will focus on basic operations on Tableau software, and the second one will focus on advanced operations requiring both Tableau software and programming.

**[Vietnamese]**Các câu hỏi của dự án giữa kỳ sẽ được tổ chức thành hai loại, loại thứ nhất sẽ tập trung vào các thao tác cơ bản trên phần mềm Tableau và loại thứ hai sẽ tập trung vào các hoạt động nâng cao yêu cầu cả phần mềm và lập trình Tableau.

***1. Exploratory data analysis (EDA)***  
Exploratory data analysis (EDA) [3] is used to analyze and investigate datasets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions. EDA is primarily used to see what data can reveal beyond the formal modeling or hypothesis testing task and provides a better understanding of data set variables and the relationships between them. Based on the idea of EDA, using Tableau software to answer the following questions:

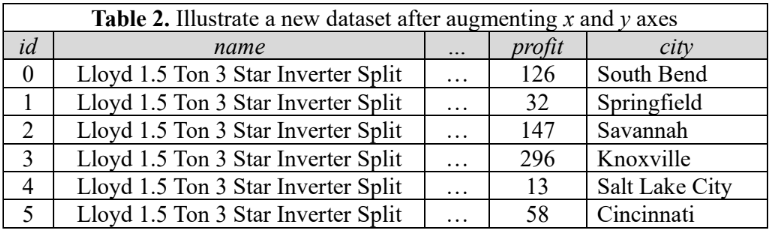
1. Find best-selling products in the scale of 10, 100, 1K, 10K items, and visualize them.
2. Find worst-selling products in the scale of 10, 100, 1K, 10K items, and visualize them.
3. Find and explain the relationships between dataset variables. (sự tương quan giữa các thuộc tính (heatmap) – các thuộc tính đưa về dạng số) bỏ thuộc tính phải có lập luận
4. Find products which are most expensive but have ratings lower than 3.0, in the scale of 10, 100, 1K items.
5. Find products which are cheapest and have ratings more than 4.0, in the scale of 10, 100, 1K items.
6. Find products which have the largest discounted price.

**[Vietnamese]**Phân tích dữ liệu thăm dò (EDA) [3] được sử dụng để phân tích và điều tra các tập dữ liệu cũng như tóm tắt các đặc điểm chính của chúng, thường sử dụng các phương pháp trực quan hóa dữ liệu. Nó giúp xác định cách tốt nhất để thao tác các nguồn dữ liệu để có được câu trả lời bạn cần, giúp các nhà khoa học dữ liệu dễ dàng khám phá các mẫu, phát hiện điểm bất thường, kiểm tra giả thuyết hoặc kiểm tra các giả định. EDA chủ yếu được sử dụng để xem dữ liệu nào có thể tiết lộ ngoài nhiệm vụ kiểm tra giả thuyết hoặc lập mô hình chính thức và cung cấp sự hiểu biết tốt hơn về các biến của tập dữ liệu và mối quan hệ giữa chúng. Dựa trên ý tưởng của EDA, sử dụng phần mềm Tableau để trả lời các câu hỏi sau:

1. Tìm các sản phẩm bán chạy nhất trong quy mô 10, 100, 1K, 10K mặt hàng và trực quan hóa chúng.
2. Tìm các sản phẩm bán ế nhất trong quy mô 10, 100, 1K, 10K mặt hàng và trực quan hóa chúng.
3. Tìm và giải thích mối quan hệ giữa các biến số liệu.
4. Tìm những sản phẩm đắt nhất nhưng có đánh giá thấp hơn 3.0, trong quy mô 10, 100, 1K mặt hàng.
5. Tìm các sản phẩm rẻ nhất và có đánh giá lớn hơn 4.0, trong quy mô 10, 100, 1K mặt hàng.
6. Tìm sản phẩm có mức giá chiết khấu lớn nhất.

***2. Augmented the dataset***To employ more power of Tableau software, in this section, you need to augment the given dataset in both x and y axes. For x axis, you will insert two new columns, i.e., "profit" and "city", where "profit" indicates the number of products which were sold in the corresponding "city" of US'. For y axis, it's quite easy to imagine, you need to duplicate each instance of the dataset to at least 100 new ones, and randomly generate the "profit" and "city" values for each of them. Table 2 illustrates the new dataset after augmenting in both x and y axes. You should note that the "id" and "id\_by\_category" columns need to be altered to make sure these identification numbers are unique.

**[Vietnames]**Để sử dụng nhiều sức mạnh hơn của phần mềm Tableau, trong phần này, bạn cần tăng tập dữ liệu đã cho theo cả trục x và y. Đối với trục x, bạn sẽ chèn hai cột mới, tức là "lợi nhuận" và "thành phố", trong đó "lợi nhuận" cho biết số lượng sản phẩm đã được bán tại "thành phố" tương ứng của Hoa Kỳ. Đối với trục y, khá dễ hình dung, bạn cần sao chép từng phiên bản của tập dữ liệu thành ít nhất 100 phiên bản mới và tạo ngẫu nhiên các giá trị "lợi nhuận" và "thành phố" cho mỗi phiên bản đó. Bảng 2 minh họa tập dữ liệu mới sau khi tăng theo cả trục x và y. Bạn nên lưu ý rằng cột "id" và "id\_by\_category" cần được thay đổi để đảm bảo các số nhận dạng này là duy nhất.



The dataset after augmenting is now having enough information to support the decision-making activities. You need to answer the following questions:

1. Find and explain the relationships between dataset variables.
2. Get a list of products that should be included in a promotion campaign based on the selling profits by city. For example, you can define a measure such as "if a product has a good selling profit at a city, it should be discounted as x percent to improve its profit", or "we should discount that jacket's price because it is going to be outdated soon". Whatever the measure you define, you need to define a threshold, for example, to determine whether a product has a good selling profit or not. (random mỗi sp trong range khác nhau + giải thích)
3. Visualize the best-and-worst-selling-profit on world map.

**[Vietnamese]**Bộ dữ liệu sau khi tăng cường hiện đã có đủ thông tin để hỗ trợ cho hoạt động ra quyết định. Bạn cần trả lời những câu hỏi sau:

1. Tìm và giải thích mối quan hệ giữa các biến tập dữ liệu.
2. Lấy ra danh sách các sản phẩm nên đưa vào chiến dịch khuyến mại dựa trên profit theo thành phố. Ví dụ: bạn có thể xác định một thước đo như "nếu một sản phẩm có lợi nhuận bán hàng tốt ở một thành phố thì sản phẩm đó nên được chiết khấu theo x phần trăm để cải thiện lợi nhuận" hoặc "chúng ta nên giảm giá chiếc áo khoác đó vì nó sẽ sớm lỗi thời”. Dù bạn xác định thước đo nào thì cũng cần xác định một ngưỡng, chẳng hạn để xác định xem một sản phẩm có lợi nhuận bán tốt hay không.
3. Hình dung lợi nhuận bán chạy nhất và tệ nhất trên bản đồ.