BỘ GIÁO DỤC VÀ ĐÀO TẠO **TRƯỜNG ĐẠI HỌC PHENIKAA**

Logo, company name

Description automatically generated

**BÁO CÁO BÀI TẬP LỚN**

**TÍCH HỢP VÀ PHÂN TÍCH DỮ LIỆU LỚN**

**TÊN ĐỀ TÀI**

**PHÂN TÍCH DỮ LIỆU LỊCH SỬ MUA HÀNG ĐIỆN TỬ**

**Giảng viên hướng dẫn:THS.** **Phạm Kim Thành**

**Sinh viên thực hiện: Trương Quang Trung(22010264)**

**MỤC LỤC**

[**DANH MỤC CÁC TỪ VIẾT TẮT** 2](#_Toc200980668)

[**MỞ ĐẦU** 3](#_Toc200980669)

[**I. Mục tiêu phân tích:** 4](#_Toc200980670)

[**IV. PHƯƠNG PHÁP NGHIÊN CỨU THỰC HIỆN ĐỀ TÀI:** 6](#_Toc200980671)

[**V. Ý NGHĨA CỦA ĐỀ TÀI:** 6](#_Toc200980672)

[**NỘI DUNG** 7](#_Toc200980673)

[**CHƯƠNG 1:TỔNG QUÁT VỀ PHẦN MỀM PHÂN TÍCH DỮ LIỆU LỊCH SỬ MUA HÀNG ĐIỆN TỬ** 7](#_Toc200980674)

[**1.1. CÁC KHÁI NIỆM LIÊN QUAN:** 7](#_Toc200980675)

[**1.2. CÁC CÔNG CỤ:** 8](#_Toc200980676)

[**1.3. KẾT CHƯƠNG 1:** 9](#_Toc200980677)

[**CHƯƠNG 2: PHÂN TÍCH CÁC BƯỚC XÂY DỰNG WEBSITE** 10](#_Toc200980678)

[**2.1. MỤC TIÊU:** 10](#_Toc200980679)

[Phân tích bộ dữ liệu lịch sử mua hàng từ một cửa hàng điện tử nhằm: 10](#_Toc200980680)

[**CHƯƠNG 3: Các chức năng chính của phân tích dữ liệu lịch sử mua hàng điện tử.** 12](#_Toc200980681)

[**3.1. CÁC BIỂU ĐỒ PHÂN TÍCH:** 12](#_Toc200980682)

[**3.2. Phân tích và trực quan hóa dữ liệu** 15](#_Toc200980683)

[**KẾT LUẬN** 20](#_Toc200980684)

[**I. KẾT QUẢ ĐẠT ĐƯỢC:** 20](#_Toc200980685)

[**TÀI LIỆU THAM KHẢO** 21](#_Toc200980686)

# **DANH MỤC CÁC TỪ VIẾT TẮT**

|  |  |  |
| --- | --- | --- |
| STT | Ký hiệu chữ viết tắt | Chữ viết đầy đủ |
| 1 | RFM | Recency, Frequency, Monetary |
| 2 | AOV | Average Order Value |
| 3 | W3C | World Wide Web Consortium |
| 4 | CRM | Customer Relationship Management |

# **MỞ ĐẦU**

Trong dự án này, chúng ta thực hiện phân tích dữ liệu lịch sử mua hàng từ một cửa hàng điện tử. Trước tiên, tập dữ liệu được tải về thông qua kagglehub và lưu trữ trong Google Drive. Sau đó, sử dụng thư viện pandas, dữ liệu được đọc từ file CSV và hiển thị các dòng đầu và cuối để có cái nhìn tổng quan ban đầu. Tiếp theo, ta tiến hành xử lý lỗi liên quan đến thời gian bằng cách chuyển đổi cột event\_time sang định dạng ngày giờ hợp lệ, loại bỏ các dòng có thời gian không hợp lệ (đặc biệt là năm 1970), và kiểm tra phạm vi thời gian của dữ liệu. Dữ liệu cũng được làm sạch bằng cách loại bỏ các dòng có giá trị thiếu quan trọng, đồng thời điền các giá trị mặc định cho thông tin phụ như nhãn hiệu và danh mục. Sau khi chuẩn hóa và chuyển đổi kiểu dữ liệu cần thiết, ta tiếp tục tạo thêm các thuộc tính hữu ích như danh mục chính – phụ, giờ trong ngày, thứ trong tuần, và tháng. Dữ liệu sau xử lý được sử dụng để tính toán các chỉ số kinh doanh quan trọng như tổng doanh thu, số lượng đơn hàng, số lượng khách hàng và giá trị trung bình mỗi đơn hàng. Cuối cùng, dữ liệu được trực quan hóa thông qua nhiều biểu đồ nhằm hiểu rõ hơn về hành vi mua sắm của khách hàng, xu hướng doanh thu theo thời gian, sự phân bổ giá sản phẩm và sự khác biệt giữa khách hàng mới và khách hàng quay lại.

## **I. Mục tiêu phân tích:**

* **Khách hàng mong muốn hiểu rõ hành vi tiêu dùng và xu hướng mua sắm trong ngành thương mại điện tử.**

Trong thời đại công nghệ số, việc nắm bắt hành vi mua hàng của khách hàng thông qua lịch sử giao dịch trở nên quan trọng hơn bao giờ hết. Khách hàng không chỉ quan tâm đến việc bán được bao nhiêu sản phẩm, mà còn muốn hiểu được điều gì khiến người tiêu dùng đưa ra quyết định mua sắm, vào thời điểm nào họ chi tiêu nhiều nhất, và sản phẩm nào đang trở thành xu hướng. Họ cần một hệ thống phân tích có thể tổng hợp và phản ánh một cách trực quan các mô hình mua sắm, nhờ đó giúp họ không chỉ nhìn thấy quá khứ mà còn dự đoán được tương lai.

Việc sử dụng website phân tích dữ liệu mua hàng sẽ là công cụ hỗ trợ đắc lực cho doanh nghiệp trong việc thu thập, xử lý và trực quan hóa thông tin, nhằm cung cấp cái nhìn sâu sắc về hành vi khách hàng, sở thích tiêu dùng và thậm chí là cả động cơ mua sắm của họ. Qua đó, doanh nghiệp có thể điều chỉnh các chiến lược kinh doanh sao cho phù hợp và hiệu quả nhất.

* **Tối ưu hóa các chỉ số kinh doanh thông qua mô hình phân tích dữ liệu hiện đại.**

Website sẽ giúp doanh nghiệp tính toán và theo dõi những chỉ số quan trọng như RFM (tần suất mua hàng, thời gian mua gần nhất, giá trị mua trung bình) hay AOV (giá trị trung bình mỗi đơn hàng) một cách tự động và trực quan. Trước đây, những thông tin này thường được tổng hợp thủ công qua báo cáo Excel rườm rà và dễ sai sót. Nhưng giờ đây, với hệ thống phân tích dữ liệu thương mại điện tử, các chỉ số sẽ được cập nhật liên tục, giúp doanh nghiệp ra quyết định nhanh chóng, chính xác và đúng thời điểm – từ việc lên kế hoạch marketing, quản lý kho hàng, đến chăm sóc khách hàng.

**II. MỤC ĐÍCH :**

* **Hỗ trợ ra quyết định chiến lược trong hoạt động kinh doanh điện tử.**

Thông qua kết quả phân tích dữ liệu lịch sử mua hàng, website giúp doanh nghiệp có cơ sở để xây dựng các chiến lược marketing cá nhân hóa cho từng phân khúc khách hàng. Ví dụ, nhóm khách hàng có giá trị mua cao nhưng gần đây không còn mua sắm sẽ được nhắm mục tiêu với chiến dịch khuyến mãi đặc biệt nhằm kéo họ quay lại. Đồng thời, việc hiểu rõ nhu cầu và hành vi mua sắm cũng giúp doanh nghiệp đưa ra dự đoán chính xác về sản phẩm cần nhập kho, tránh tồn hàng hoặc thiếu hụt.

* **Tăng hiệu quả kinh doanh và cải thiện trải nghiệm khách hàng.**

Việc tận dụng dữ liệu giúp doanh nghiệp tiếp cận đúng khách hàng vào đúng thời điểm với sản phẩm phù hợp, từ đó tăng tỉ lệ chuyển đổi và giữ chân khách hàng. Khi khách hàng nhận được sự quan tâm đúng mức và trải nghiệm mua sắm dễ dàng, họ có xu hướng quay lại thường xuyên hơn và trở thành khách hàng trung thành. Điều này không chỉ giúp tăng doanh thu mà còn xây dựng thương hiệu mạnh và uy tín trên thị trường.

**III. ĐỐI TƯỢNG VÀ PHẠM VI CỦA ĐỀ TÀI**

- **Đối tượng của đề tài:** Các doanh nghiệp thương mại điện tử, đặc biệt là các cửa hàng trực tuyến có nhu cầu phân tích dữ liệu để nâng cao hiệu quả kinh doanh.

- **Phạm vi đề tài:** Ứng dụng công nghệ phân tích dữ liệu trong việc hiểu hành vi khách hàng, tối ưu các chiến lược marketing và quản lý vận hành trong môi trường kinh doanh trực tuyến.

## **IV. PHƯƠNG PHÁP NGHIÊN CỨU THỰC HIỆN ĐỀ TÀI:**

* **Phương pháp nghiên cứu tài liệu:**

Tham khảo các tài liệu liên quan đến phân tích dữ liệu lớn (Big Data), mô hình RFM, AOV và các hệ thống thương mại điện tử hiện nay. Ngoài ra, nghiên cứu thêm các tài liệu kỹ thuật để hiểu rõ cách xây dựng dashboard và hệ thống phân tích tự động trên website.

* **Phương pháp nghiên cứu thực tiễn:**

Phân tích dữ liệu từ các trang thương mại điện tử như Shopee, Lazada, Tiki,... hoặc các cửa hàng online thực tế có cung cấp thông tin mua hàng, nhằm đưa ra ví dụ cụ thể về việc sử dụng các chỉ số RFM, AOV để cải thiện hiệu quả kinh doanh.

**V. Ý NGHĨA CỦA ĐỀ TÀI:**

Trong bối cảnh thương mại điện tử phát triển mạnh mẽ, việc phân tích dữ liệu mua hàng không chỉ còn là xu hướng mà đã trở thành yêu cầu thiết yếu nếu doanh nghiệp muốn tồn tại và phát triển bền vững. Website phân tích dữ liệu không chỉ mang lại cái nhìn tổng quan về hoạt động kinh doanh mà còn là công cụ giúp doanh nghiệp hiểu rõ khách hàng – một trong những tài sản quý giá nhất. Từ đó, doanh nghiệp có thể tối ưu hóa chiến lược tiếp thị, chăm sóc khách hàng và quản lý vận hành hiệu quả hơn. Đề tài này sẽ là nền tảng quan trọng để các doanh nghiệp vừa và nhỏ bước vào cuộc chơi dữ liệu một cách thông minh, hiện đại và phù hợp với thời đại chuyển đổi số.

**NỘI DUNG**

**CHƯƠNG 1:TỔNG QUÁT VỀ PHẦN MỀM PHÂN TÍCH DỮ LIỆU LỊCH SỬ MUA HÀNG ĐIỆN TỬ**

### **1.1. CÁC KHÁI NIỆM LIÊN QUAN:**

**1.1.1. Tổng quan về phần mềm phân tích dữ liệu mua hàng:**

Phần mềm phân tích dữ liệu mua hàng từ cửa hàng điện tử là một ứng dụng hỗ trợ các nhà quản trị, chuyên viên dữ liệu hoặc doanh nghiệp theo dõi và phân tích hành vi mua sắm của khách hàng thông qua các giao dịch thực tế được ghi nhận trong hệ thống thương mại điện tử.

Thông qua việc trích xuất, xử lý và trực quan hóa dữ liệu giao dịch, phần mềm giúp làm sáng tỏ các xu hướng tiêu dùng, đánh giá hiệu quả của các mặt hàng, thương hiệu và thời điểm bán hàng. Kết quả phân tích này đóng vai trò quan trọng trong việc tối ưu hóa chiến lược kinh doanh, nâng cao doanh thu và cải thiện trải nghiệm khách hàng.

**1.1.2. Dữ liệu giao dịch thương mại điện tử là gì?**

Dữ liệu giao dịch thương mại điện tử là tập hợp các thông tin được ghi lại từ các hoạt động mua sắm trên nền tảng số, bao gồm: thời gian mua hàng, mã người dùng, tên sản phẩm, danh mục hàng hóa, giá bán, thương hiệu, số lượng sản phẩm, v.v.

Loại dữ liệu này thường được lưu trữ dưới dạng bảng tính (CSV hoặc Excel) và có thể được phân tích bằng các công cụ như Python, Pandas, Matplotlib hoặc Seaborn. Qua việc khai thác dữ liệu này, ta có thể xác định khách hàng trung thành, sản phẩm bán chạy, khung giờ vàng mua sắm và nhiều yếu tố giá trị khác.

### **1.2. CÁC CÔNG CỤ:**

**1.2.1. Pandas trong xử lý dữ liệu:**

* **Định nghĩa**

Pandas là một thư viện mã nguồn mở nổi tiếng trong Python, chuyên dùng để thao tác và phân tích dữ liệu có cấu trúc dạng bảng (DataFrame).

* **Chức năng chính:**

Pandas cho phép đọc, ghi và xử lý dữ liệu từ nhiều định dạng khác nhau như CSV, Excel, SQL... Các tính năng mạnh mẽ bao gồm lọc, sắp xếp, thống kê mô tả, gộp nhóm (groupby) và thao tác theo thời gian (datetime).

* **Vai trò**

Trong dự án này, Pandas đóng vai trò trung tâm giúp làm sạch dữ liệu, phân tích chỉ số kinh doanh và tạo ra tập dữ liệu đầu vào cho các bước trực quan hóa tiếp theo.

**1.2.2. Matplotlib và Seaborn trong trực quan hóa dữ liệu:**

**Matplotlib:**  
Matplotlib là thư viện vẽ biểu đồ mạnh mẽ và linh hoạt trong Python. Nó cho phép vẽ biểu đồ đường, biểu đồ cột, biểu đồ tròn và nhiều loại biểu đồ khác nhằm biểu diễn dữ liệu một cách trực quan và dễ hiểu.

**Seaborn:**  
Seaborn là một thư viện mở rộng dựa trên Matplotlib, cung cấp các biểu đồ có thiết kế hiện đại, phù hợp cho phân tích thống kê và biểu diễn các mối quan hệ phức tạp giữa nhiều biến.

**Vai trò:**  
Cả hai thư viện đều được sử dụng để vẽ biểu đồ doanh thu theo tháng, phân bố giờ mua sắm, so sánh thương hiệu và phân loại khách hàng. Nhờ đó, kết quả phân tích trở nên sinh động, trực quan và dễ dàng diễn giải hơn cho người dùng cuối.

### **1.3. KẾT CHƯƠNG 1:**

Qua phần trình bày các khái niệm liên quan đến dữ liệu mua hàng và các công cụ sử dụng như Pandas, Matplotlib và Seaborn, chúng ta có được nền tảng vững chắc để thực hiện các bước phân tích tiếp theo. Trong các chương kế tiếp, ta sẽ tiến hành tiền xử lý dữ liệu, phân tích hành vi khách hàng, và xây dựng các biểu đồ nhằm rút ra những thông tin hữu ích từ dữ liệu thực tế.

## **CHƯƠNG 2: PHÂN TÍCH CÁC BƯỚC XÂY DỰNG WEBSITE**

### **2.1. MỤC TIÊU:**

### Phân tích bộ dữ liệu lịch sử mua hàng từ một cửa hàng điện tử nhằm:

* Tiền xử lý và làm sạch dữ liệu để đảm bảo tính chính xác.
* Phân tích hành vi mua sắm của người dùng qua thời gian.
* Tính toán các chỉ số kinh doanh quan trọng như doanh thu, số lượng đơn hàng, số lượng khách hàng.
* Trực quan hóa dữ liệu nhằm hỗ trợ đưa ra các quyết định kinh doanh.
* Phân loại khách hàng thành **mới** hoặc **quay lại** nhằm phục vụ cho các chiến lược marketing cá nhân hóa.

**2.2. CÁC BƯỚC THIẾT KẾ:**

**Bước 1: Tải và đọc dữ liệu**

* Tải dữ liệu mới nhất từ Kaggle bằng kagglehub.
* Lưu dữ liệu thủ công vào Google Drive và đọc bằng đường dẫn nội bộ.

**Bước 2: Tiền xử lý**

* Chuyển đổi event\_time sang định dạng datetime.
* Loại bỏ các dòng có giá trị thiếu (NaN) ở các cột quan trọng: event\_time, user\_id, price.
* Lọc bỏ các dòng có năm 1970 do lỗi định dạng thời gian.
* Gán giá trị 'unknown' cho các trường thông tin phụ bị thiếu như brand và category\_code.
* Chuyển đổi kiểu dữ liệu ID sang chuỗi để đảm bảo nhất quán.
* Tách category\_code thành 2 cột: main\_category và sub\_category.
* Trích xuất các thuộc tính thời gian: hour, day\_of\_week, month.

**Bước 3: Tính toán chỉ số kinh doanh (KPIs)**

* **Tổng doanh thu**
* **Số lượng đơn hàng**
* **Số khách hàng duy nhất**
* **Giá trị trung bình mỗi đơn hàng (AOV)**

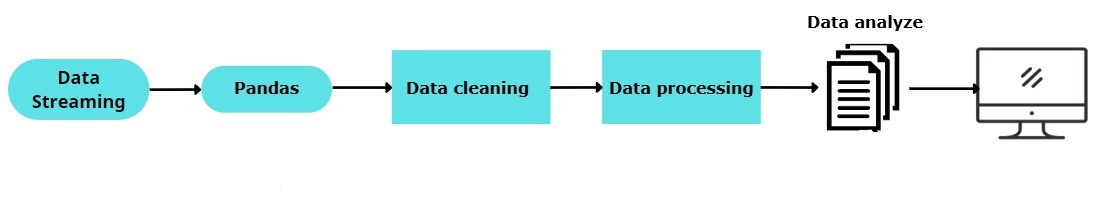
**Bước 4: Trực quan hóa dữ liệu**

* Doanh thu theo tháng
* Số đơn hàng theo giờ
* Doanh thu theo danh mục chính và thương hiệu
* Phân bổ giá sản phẩm theo danh mục
* Top thương hiệu trong ngành hàng điện tử
* Số sản phẩm trong mỗi đơn hàng
* Phân loại khách hàng mới và quay lại theo tháng

**Bước 5: Phân tích hành vi khách hàng**

* Xác định khách hàng mới (first-time buyer) và khách quay lại (returning).
* Vẽ biểu đồ số lượng khách hàng theo thời gian để đánh giá hiệu quả tiếp cận khách hàng mới.

**2.3. QUY TRÌNH XỬ LÝ DỮ LIỆU:**

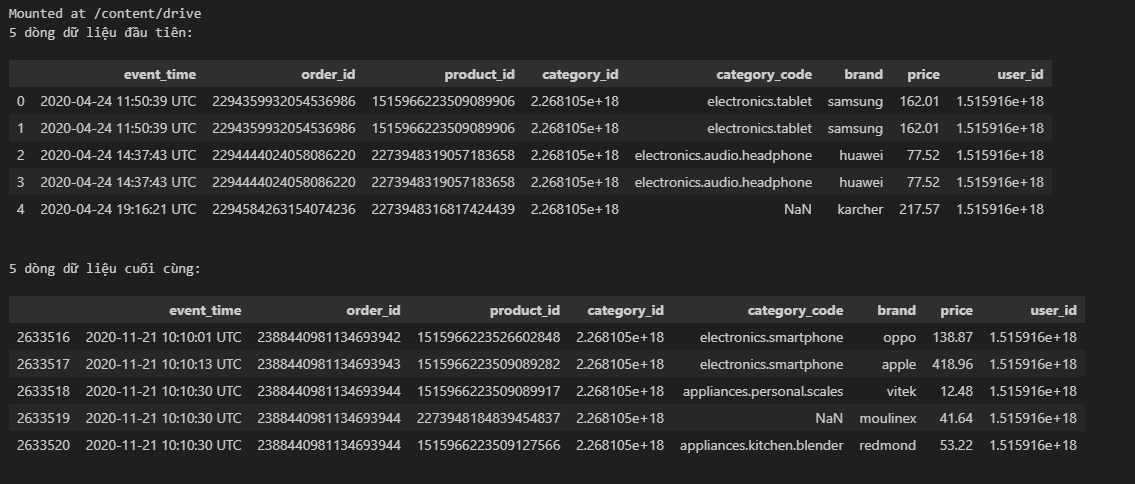
****

1. **Data Streaming**: Thu thập dữ liệu đầu vào (CSV hoặc real-time).
2. **Pandas**: Đọc và chuyển dữ liệu thành DataFrame.
3. **Data Cleaning**: Làm sạch dữ liệu (loại bỏ lỗi, thiếu, định dạng sai).
4. **Data Processing**: Xử lý và tính toán dữ liệu cần thiết.
5. **Data Analyze**: Phân tích dữ liệu (doanh thu, hành vi khách hàng...).
6. **Hiển thị**: Kết quả được trình bày trên máy tính thông qua bảng/báo cáo.

## **CHƯƠNG 3: Các chức năng chính của phân tích dữ liệu lịch sử mua hàng điện tử.**

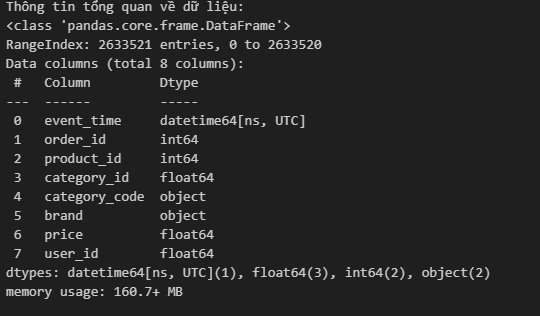
### **3.1. CÁC BIỂU ĐỒ PHÂN TÍCH:**

**a. Đọc dữ liệu**



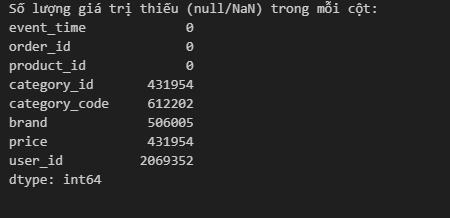
* Dữ liệu CSV được tải từ Google Drive: kz.csv.
* Đọc bằng thư viện pandas.

**b. Xử lý thời gian**

****

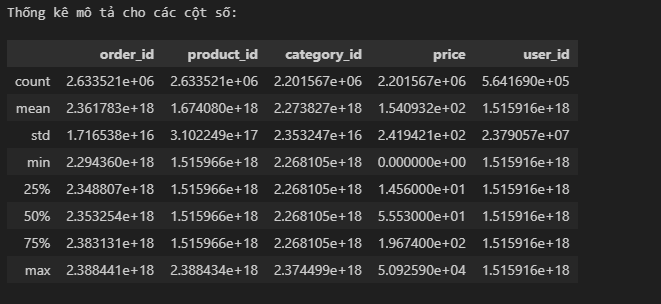
* Chuyển cột event\_time sang dạng datetime.
* Loại bỏ các dòng có thời gian không hợp lệ (NaT) hoặc có năm là 1970.
* Khoảng thời gian dữ liệu: từ ngày đầu tiên đến ngày cuối cùng.

**c. Xử lý giá trị thiếu**



* Các cột quan trọng như user\_id, price bị thiếu được loại bỏ.
* Các cột phụ (brand, category\_code) được điền bằng 'unknown'.

**d. Chuẩn hóa kiểu dữ liệu**

****

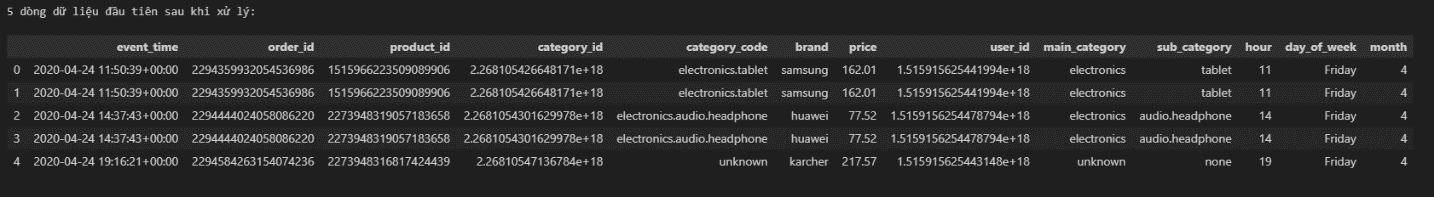
* Các cột định danh (order\_id, product\_id,...) được chuyển sang kiểu chuỗi.

**e. Tạo thêm thuộc tính**

**A screenshot of a computer

Description automatically generated**

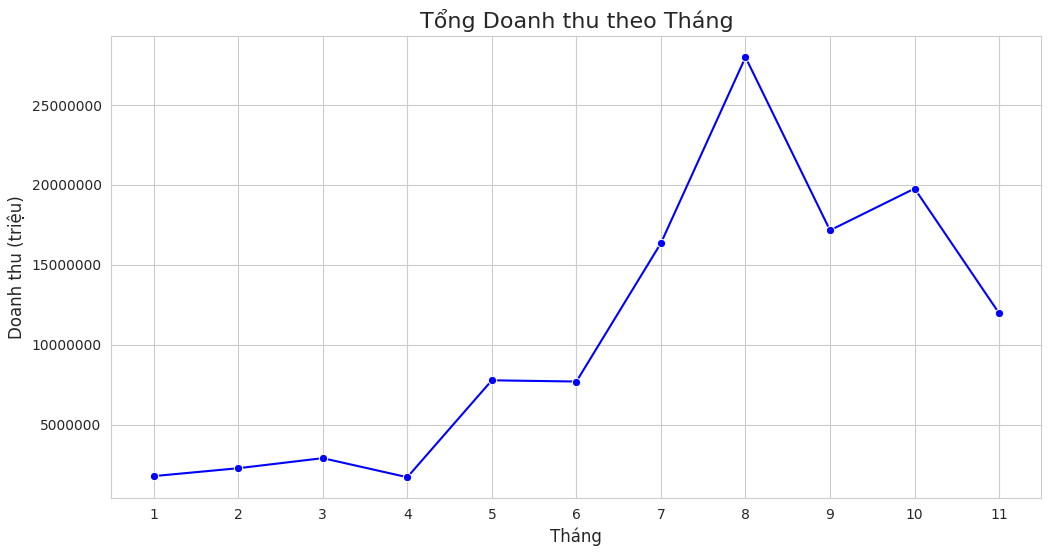
* Tách category\_code thành main\_category và sub\_category.



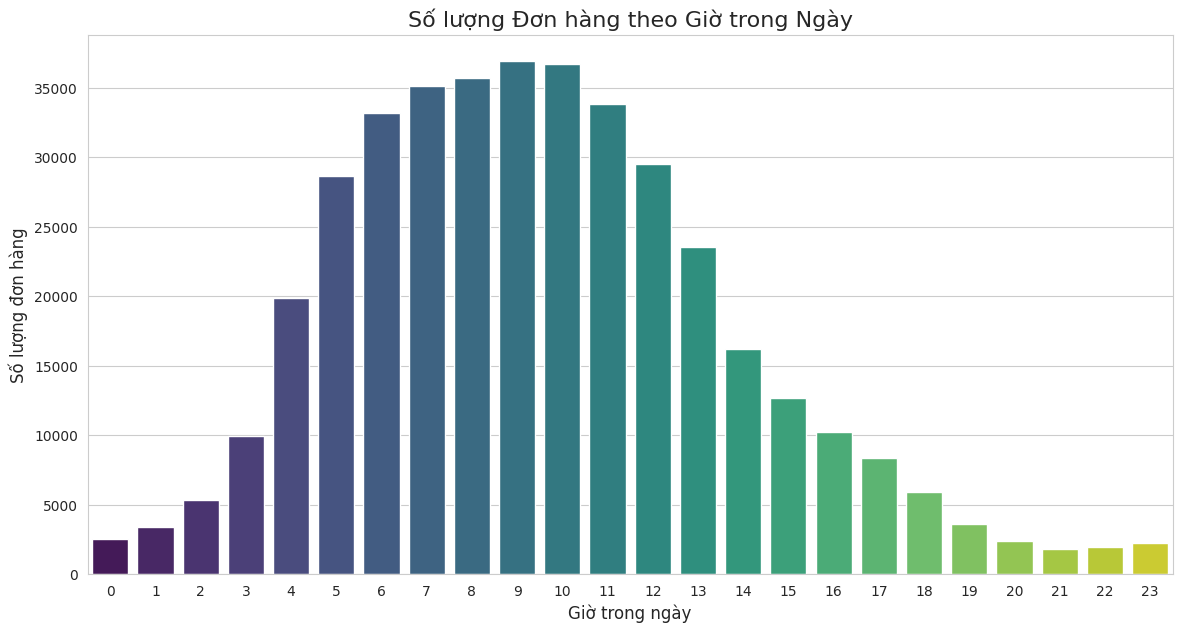
* Trích xuất hour, day\_of\_week, month từ event\_time.

### **3.2. Phân tích và trực quan hóa dữ liệu**

**b. Phân tích theo thời gian**

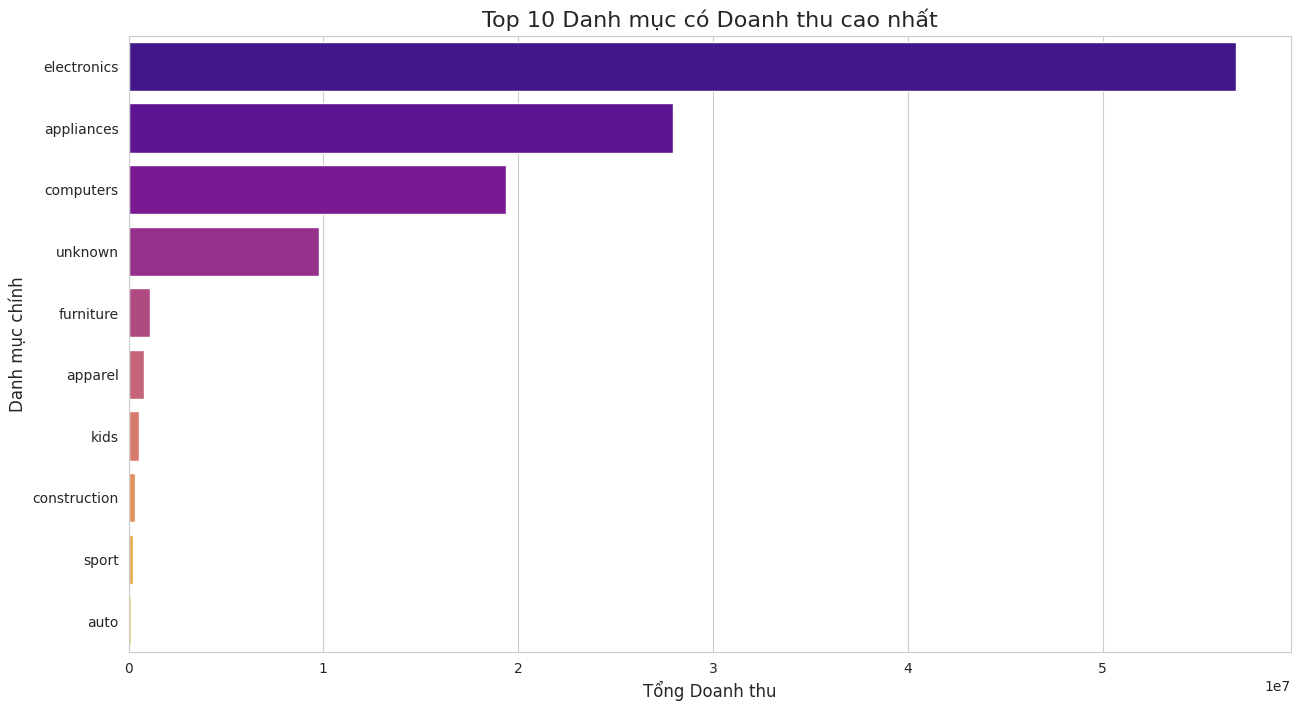
****

**Doanh thu theo tháng**: Biểu đồ đường cho thấy sự biến động theo thời gian.

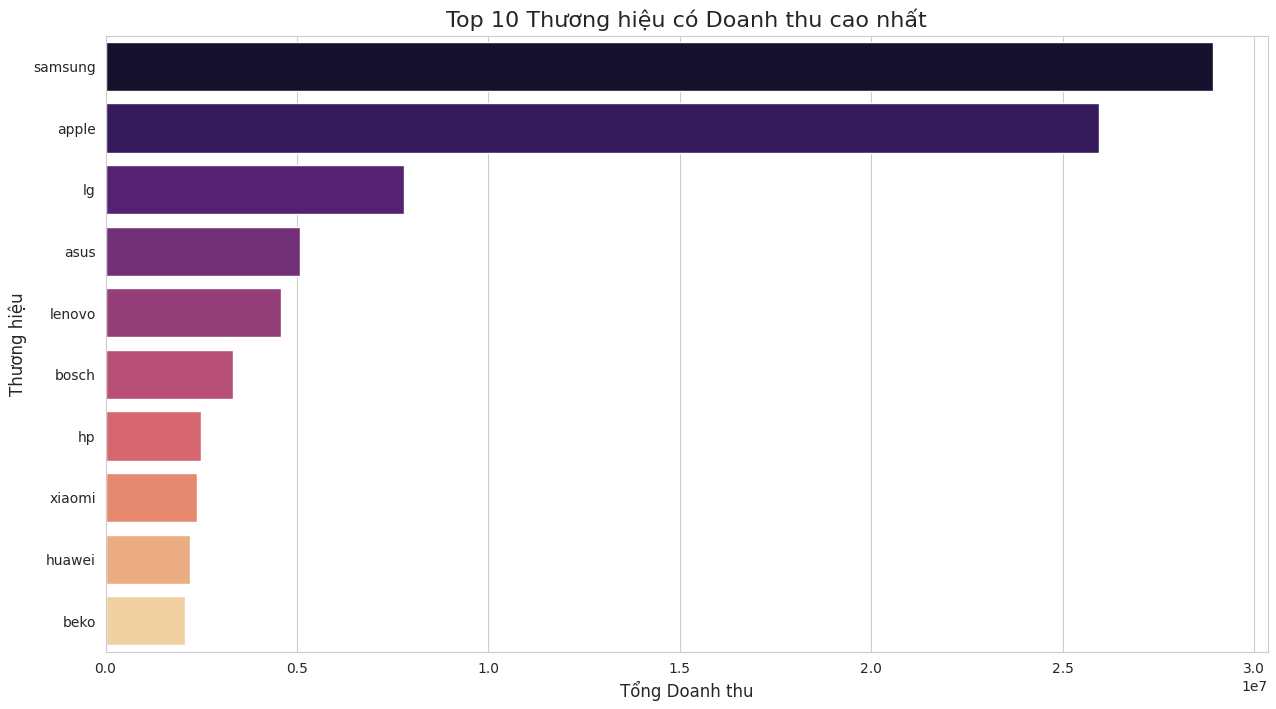
****

* **Đơn hàng theo giờ**: Giờ cao điểm mua sắm tập trung vào buổi sáng.

**c. Phân tích theo danh mục**

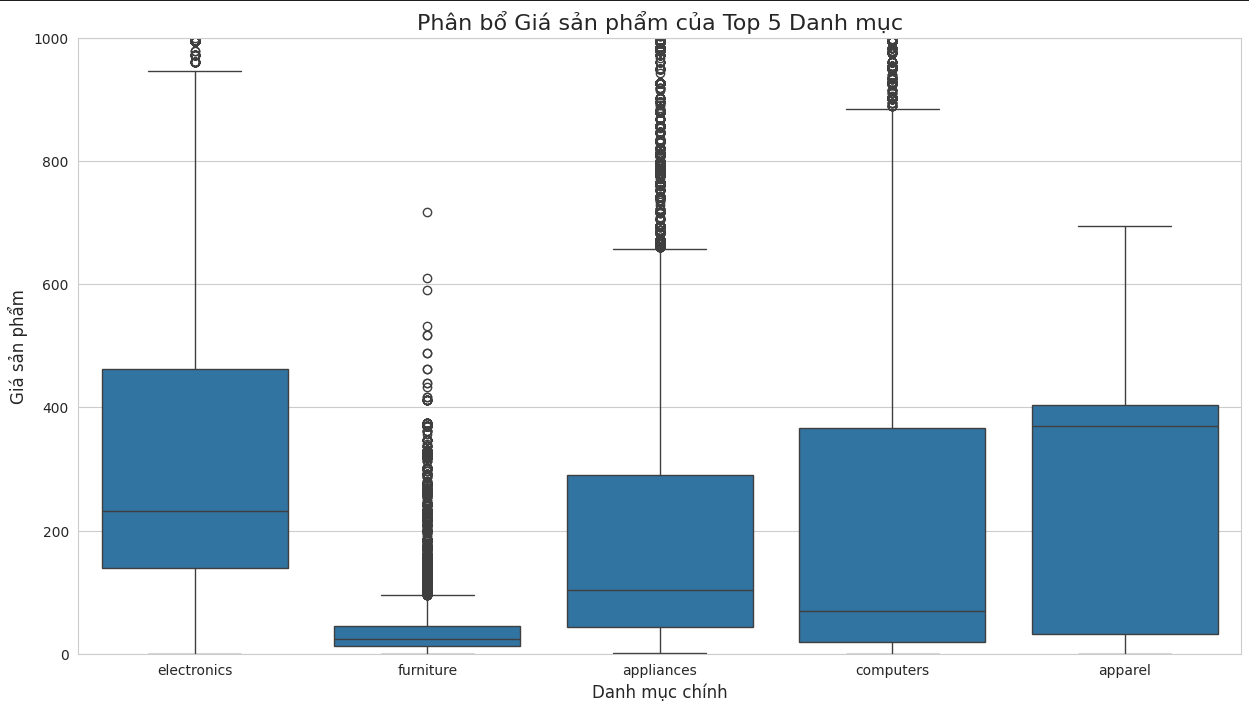
****

* **Top danh mục có doanh thu cao nhất**: electronics, appliances, computers,...



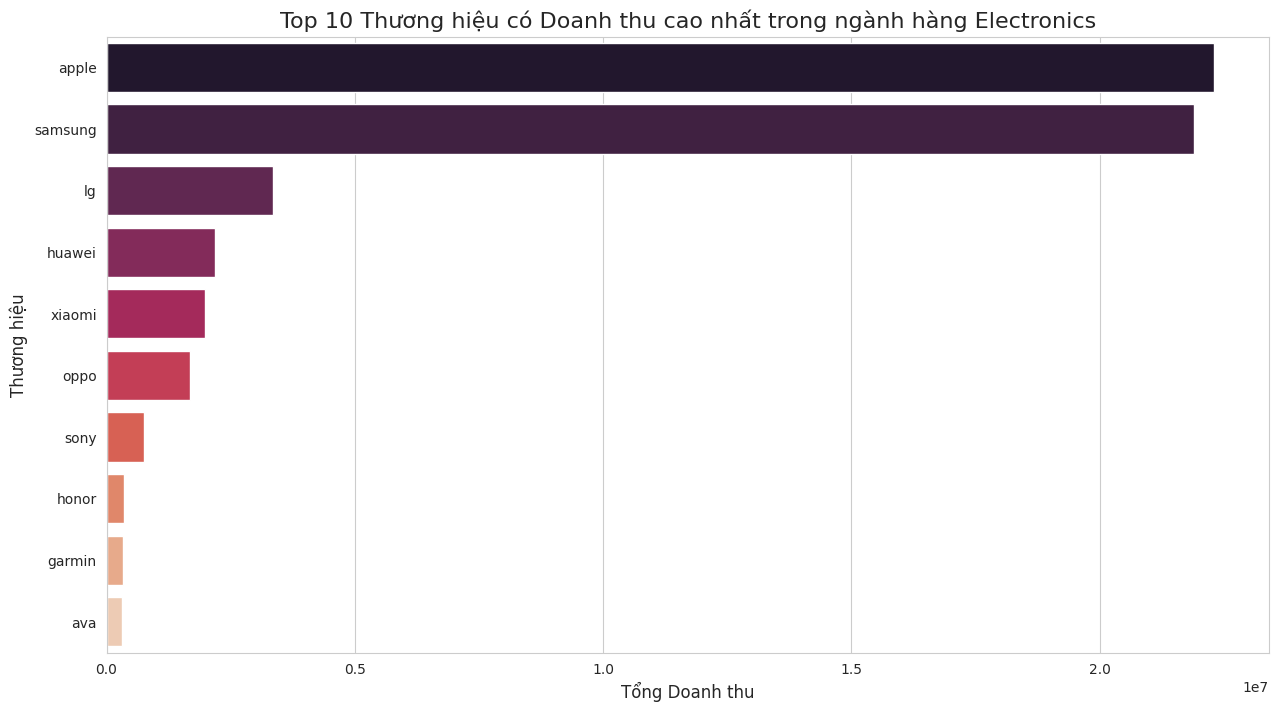
* **Top thương hiệu nổi bật** *(loại bỏ unknown)*: apple, samsung, xiaomi,...

**d. Phân bổ giá sản phẩm**

****

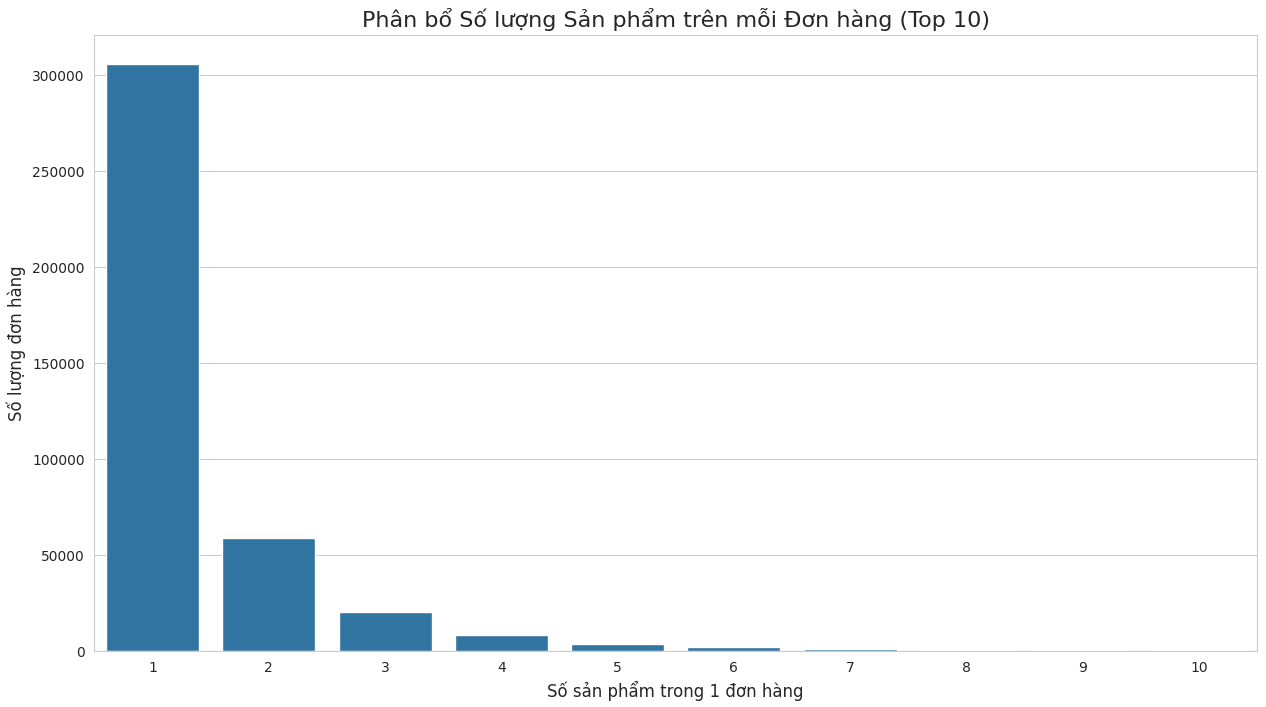
* Biểu đồ boxplot cho thấy phân bố giá trong top 5 danh mục, nhiều sản phẩm dưới $1000.

**e. Phân tích chuyên sâu danh mục "electronics"**

****

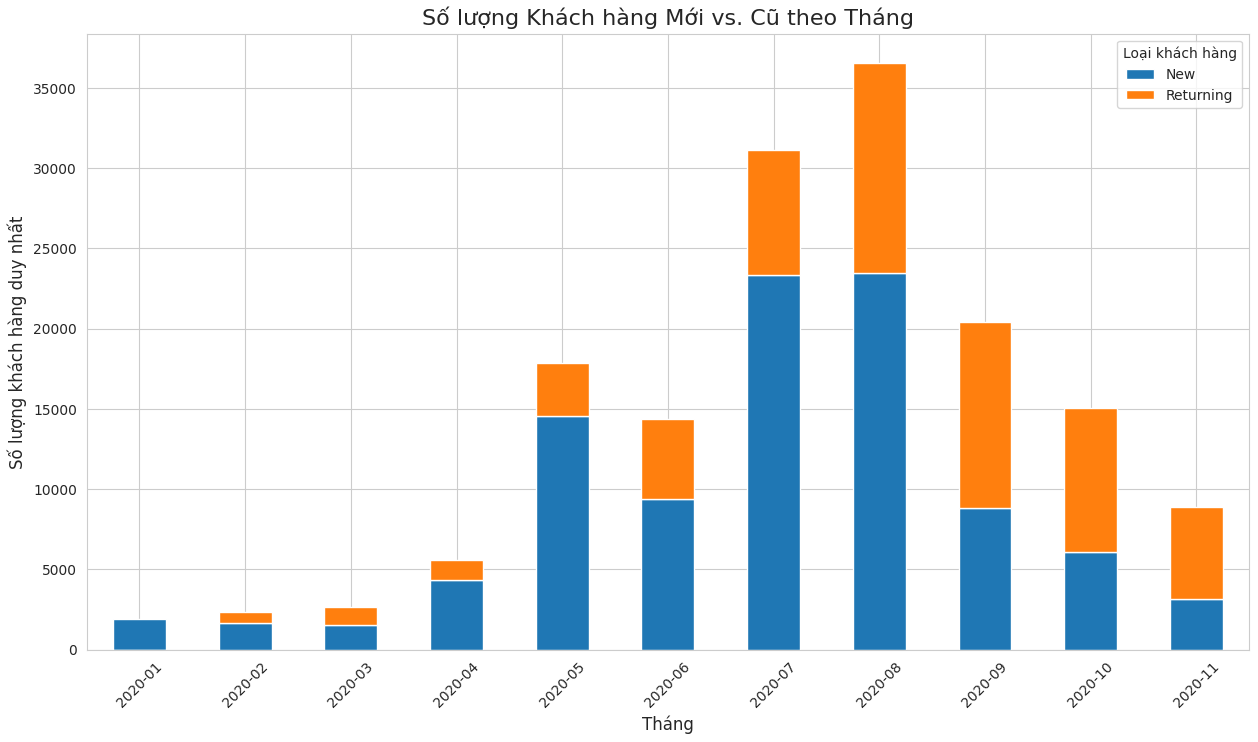
* Top 10 thương hiệu có doanh thu cao nhất được trích xuất.

**f. Sản phẩm trên mỗi đơn hàng**

****

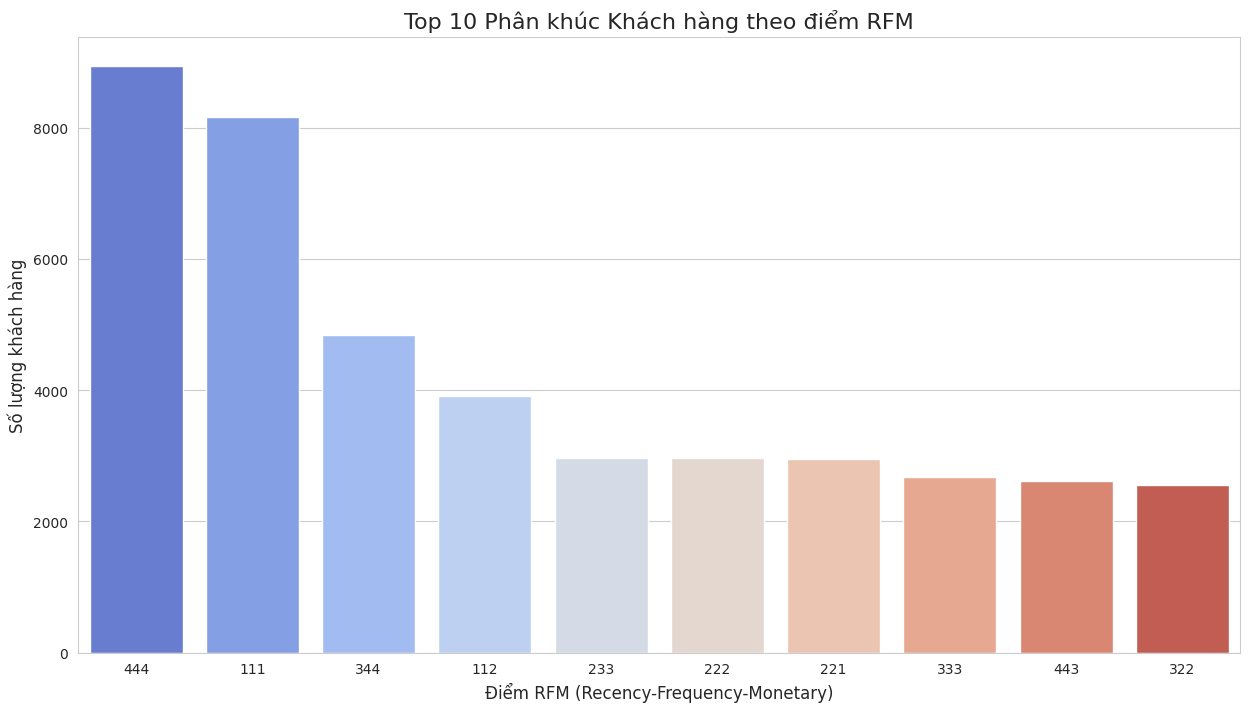
* Trung bình có X sản phẩm/đơn, biểu đồ countplot thể hiện rõ phân bố

**g. Phân tích hành vi khách hàng mới/cũ**

****

* Xác định khách hàng là **New** nếu lần mua đầu trùng tháng mua hiện tại, ngược lại là **Returning**.
* Biểu đồ stacked bar cho thấy tỷ lệ khách mới/cũ theo từng tháng.

**h. Phân tích phân khúc khách hàng theo điểm RFM**

****

**KẾT LUẬN**

**I. KẾT QUẢ ĐẠT ĐƯỢC:**

Thông qua quá trình **tích hợp và xử lý dữ liệu lớn từ hệ thống giao dịch thương mại điện tử**, kết hợp với các công cụ phân tích trực quan, chúng ta thu được một số insight quan trọng giúp nâng cao hiệu quả kinh doanh:

**Các thương hiệu lớn dẫn đầu thị phần**: Phân tích dữ liệu lớn theo chiều sâu cho thấy các thương hiệu như **Apple, Samsung, Xiaomi** thường xuyên nằm trong **top đầu về số lượng đơn và doanh thu**. Điều này giúp định hướng chiến lược hợp tác thương hiệu và phân bổ ngân sách quảng cáo hiệu quả.

**Khách hàng trung thành là nguồn lực chính**: Từ việc phân tích hành vi mua lặp lại và chu kỳ vòng đời khách hàng (customer lifetime value), hệ thống ghi nhận rằng phần lớn doanh thu đến từ **khách hàng đã từng mua trước đó**. Việc tích hợp dữ liệu khách hàng từ nhiều nguồn (CRM, lịch sử giao dịch, email marketing) cho phép xây dựng chân dung khách hàng chi tiết.

**Số lượng sản phẩm mỗi đơn hàng phổ biến từ 1 đến 3 sản phẩm**: Phân tích phân phối số lượng sản phẩm trên quy mô dữ liệu lớn giúp hiểu rõ hành vi mua sắm chi tiết, từ đó đưa ra chiến lược **gợi ý sản phẩm** hoặc **tăng AOV.** (Average Order Value)

**II. HƯỚNG PHÁT TRIỂN**

- Tìm hiểu sâu về **tích hợp dữ liệu đa nguồn:** Kết nối thêm dữ liệu từ API bán hàng, CRM, Google Analytics hoặc dữ liệu mạng xã hội để phân tích hành vi toàn diện hơn (ví dụ: tích hợp qua requests, Google BigQuery, hoặc Facebook Graph API).

- Tìm hiểu thêm một số ngôn ngữ, các phần mềm ứng dụng để nâng cao giao diện đồ hoạ đẹp mắt, thân thiện…

- Xây dựng quy mô lớn hơn với nhiều ứng dụng, chức năng hơn…

Để hoàn thành đề tài này, một lần nữa chúng em xin cảm ơn Thầy Phạm Kim Thành đã là hướng dẫn chỉ bảo chúng em. Chúng em xin chân thành cảm ơn.

**TÀI LIỆU THAM KHẢO**

[1] https://www.geeksforgeeks.org/rfm-analysis-analysis-using-python/

[2] https://www.w3schools.com/python/pandas/ref\_df\_dropna.asp

[3] https://www.analyticsvidhya.com/blog/2022/05/market-basket-analysis-based-on-rfm-analysis/

[4] https://www.analyticsvidhya.com/blog/2022/05/market-basket-analysis-based-on-rfm-analysis/