**CHƯƠNG 3: BỘ DỮ LIỆU ONLINE RETAIL**

Bộ dữ liệu Online Retail được sử dụng trong nghiên cứu này được trích xuất từ (). Dữ liệu ghi nhận toàn bộ các giao dịch mua bán của một doanh nghiệp bán lẻ trực tuyến có trụ sở tại Vương quốc Anh trong khoảng thời gian từ tháng 12 năm 2010 đến tháng 12 năm 2011. Doanh nghiệp này chuyên kinh doanh các mặt hàng quà tặng và đồ lưu niệm, phục vụ khách hàng trên phạm vi toàn cầu.

Bộ dữ liệu bao gồm 541.909 bản ghi và 8 biến, trong đó mỗi bản ghi đại diện cho một sản phẩm trong một giao dịch cụ thể. Nghĩa là, một hóa đơn (*InvoiceNo*) có thể chứa nhiều dòng dữ liệu tương ứng với các sản phẩm khác nhau được mua trong cùng một đơn hàng.

Các biến chính của bộ dữ liệu bao gồm:

|  |  |  |
| --- | --- | --- |
| **Tên biến** | **Kiểu dữ liệu** | **Ý nghĩa** |
| **InvoiceNo** | object | Mã số hóa đơn, thể hiện từng giao dịch. |
| **StockCode** | object | Mã sản phẩm. |
| **Description** | object | Mô tả sản phẩm. |
| **Quantity** | int64 | Số lượng hàng hóa được đặt trong mỗi dòng giao dịch. |
| **InvoiceDate** | datetime64[ns] | Ngày và giờ phát sinh giao dịch. |
| **UnitPrice** | float64 | Giá bán đơn vị sản phẩm (đơn vị: bảng Anh). |
| **CustomerID** | float64 | Mã định danh khách hàng. |
| **Country** | object | Quốc gia nơi khách hàng thực hiện giao dịch. |

# 3.1. Khám phá dữ liệu ban đầu (Initial Data Exploration)

Dữ liệu được nhập và kiểm tra bằng ngôn ngữ Python thông qua thư viện Pandas. Kết quả kiểm tra cấu trúc cho thấy tập dữ liệu có kích thước 541.909 hàng × 8 cột, phản ánh quy mô tương đối lớn và đa dạng về đặc trưng.

Khi tiến hành thống kê mô tả, các đặc trưng định lượng bao gồm *Quantity*, *UnitPrice* và *CustomerID* cho thấy đặc điểm phân bố như sau:

Kết quả thống kê cho thấy:

* Biến Quantity có phạm vi giá trị rộng, với độ lệch chuẩn cao (*std = 218.08*) và xuất hiện giá trị âm đến –80,995, biểu hiện khả năng tồn tại giao dịch hoàn trả hoặc lỗi ghi nhận.
* Biến UnitPrice có phân bố không đồng nhất, giá trị nhỏ nhất –11,062.06 cho thấy sự xuất hiện của dữ liệu bất thường cần được xử lý trước khi phân tích sâu hơn.
* Biến CustomerID có 406.829 giá trị hợp lệ trên tổng 541.909 bản ghi, cho thấy tồn tại nhiều giao dịch không gắn mã khách hàng.
* Biến InvoiceDate ghi nhận thời gian giao dịch liên tục trong giai đoạn từ 01/12/2010 đến 09/12/2011, đảm bảo cơ sở cho việc phân tích xu hướng theo thời gian.

Ở giai đoạn này, việc khám phá dữ liệu ban đầu giúp hiểu rõ cấu trúc, kiểu dữ liệu và phạm vi giá trị của các biến, từ đó chuẩn bị cho các bước làm sạch dữ liệu (Data Cleaning) và phân tích chuyên sâu trong các phần tiếp theo.

# 3.2. Làm sạch dữ liệu (Data Cleaning)

Trước khi tiến hành các phân tích chuyên sâu, dữ liệu cần được xử lý để loại bỏ nhiễu, giá trị thiếu và các bản ghi không hợp lệ. Quá trình làm sạch dữ liệu được thực hiện bằng ngôn ngữ Python với thư viện Pandas, bao gồm các bước sau:

1. **Xử lý giá trị thiếu (Missing Values):** Biến *CustomerID* có một số lượng lớn giá trị trống, chiếm gần 25% tổng số bản ghi. Vì mã khách hàng là yếu tố cần thiết cho các phân tích dựa trên hành vi người mua, nên các bản ghi thiếu thông tin *CustomerID* được loại bỏ bằng phương thức dropna(). Việc này giúp đảm bảo tính toàn vẹn khi phân tích tập khách hàng.
2. **Loại bỏ hóa đơn bị hủy (Cancelled Invoices):** Trong bộ dữ liệu, các hóa đơn bị hủy được ký hiệu bằng việc *InvoiceNo* bắt đầu bằng chữ “C”. Những giao dịch này không phản ánh hoạt động mua bán thực tế và có thể gây sai lệch cho các chỉ số doanh thu, do đó được loại bỏ bằng biểu thức ~df['InvoiceNo'].astype(str).str.startswith('C').
3. **Xử lý giá trị bất hợp lệ trong biến định lượng:** Các biến định lượng *Quantity* và *UnitPrice* được kiểm tra để phát hiện giá trị âm hoặc bằng 0. Đây là các giá trị không hợp lý trong ngữ cảnh giao dịch bán hàng. Các bản ghi vi phạm điều kiện này được loại bỏ bằng biểu thức điều kiện (df['Quantity'] > 0) & (df['UnitPrice'] > 0).
4. **Chuẩn hóa kiểu dữ liệu (Data Type Standardization):** Biến *InvoiceDate* đã được tự động nhận dạng ở dạng thời gian datetime64[ns] trong quá trình đọc tệp Excel, do đó không cần chuyển đổi lại. Biến *CustomerID* được chuẩn hóa về kiểu **chuỗi (string)** để đảm bảo tính thống nhất khi xử lý và nhóm dữ liệu theo khách hàng.

Sau khi hoàn tất các bước xử lý, kích thước của tập dữ liệu giảm từ 541.909 bản ghi xuống còn 397.884 bản ghi, với 8 biến giữ nguyên.

# 3.3. Tạo biến mới và Chuẩn bị phân tích (Feature Engineering and Data Preparation)

Sau khi hoàn tất quá trình làm sạch dữ liệu, bước tiếp theo trong giai đoạn tiền xử lý là xây dựng các biến đặc trưng mới (feature engineering) và chuẩn bị dữ liệu cho các phân tích mô tả và mô hình hóa. Các phép biến đổi được thực hiện nhằm mở rộng khả năng khai thác thông tin ẩn trong dữ liệu gốc, đồng thời hỗ trợ trực quan hóa và tổng hợp xu hướng theo nhiều chiều khác nhau.

## 3.1 Xây dựng biến doanh thu (TotalPrice)

Biến mới TotalPrice được tính bằng công thức:

TotalPrice=Quantity×UnitPrice\text{TotalPrice} = \text{Quantity} \times \text{UnitPrice}TotalPrice=Quantity×UnitPrice

Biến này phản ánh giá trị tiền tệ của từng dòng giao dịch, cho phép phân tích doanh thu ở các cấp độ khác nhau (sản phẩm, khách hàng, quốc gia hoặc thời gian). Đây là biến trung tâm trong các nghiên cứu về doanh thu, giá trị khách hàng, và hiệu quả kinh doanh.

## 3.3.2. Trích xuất đặc trưng thời gian (Temporal Features)

Từ biến thời gian gốc *InvoiceDate*, các thành phần thời gian chi tiết được tách ra để phục vụ các phân tích theo chu kỳ và xu hướng:

* InvoiceYear: năm phát sinh giao dịch, giúp theo dõi biến động doanh thu giữa các năm.
* InvoiceMonth: tháng phát sinh giao dịch, phục vụ phân tích theo mùa vụ.
* InvoiceDay: ngày trong tháng, giúp phát hiện các mô hình hành vi ngắn hạn.

Các biến này được trích xuất bằng thuộc tính .dt của Pandas và được lưu trữ riêng biệt để thuận tiện cho quá trình tổng hợp dữ liệu theo từng mức thời gian.

## 3.3.3. Phân nhóm quốc gia (Country Grouping)

Để đơn giản hóa quá trình phân tích so sánh giữa khách hàng nội địa và quốc tế, biến mới CountryGroup được tạo dựa trên biến *Country*. Các bản ghi có quốc gia là *United Kingdom* được gán nhãn “UK**”**, còn lại được gán nhãn “non-UK”. Cách phân nhóm này hỗ trợ đánh giá sự khác biệt trong hành vi tiêu dùng và giá trị mua hàng giữa thị trường trong nước và nước ngoài.

## 3.3.4. Chuẩn bị biến thời gian gộp tháng (InvoiceYearMonth)

Để phục vụ cho các biểu đồ doanh thu và tần suất giao dịch theo chu kỳ, một biến thời gian tổng hợp mới là InvoiceYearMonth được tạo ra bằng cách chuyển đổi *InvoiceDate* sang định dạng chu kỳ tháng (to\_period('M')). Biến này giúp biểu diễn dữ liệu theo từng tháng (ví dụ: “2011-03”), thuận lợi cho việc xây dựng các biểu đồ xu hướng và phân tích tăng trưởng.

Sau khi thêm các biến mới, cấu trúc dữ liệu được mở rộng với tổng cộng 13 biến. Các kiểu dữ liệu được kiểm tra và duy trì ở trạng thái nhất quán, đảm bảo tính sẵn sàng cho các bước phân tích mô tả và mô hình hóa ở phần kế tiếp.

# 3.4. Phân tích mô tả

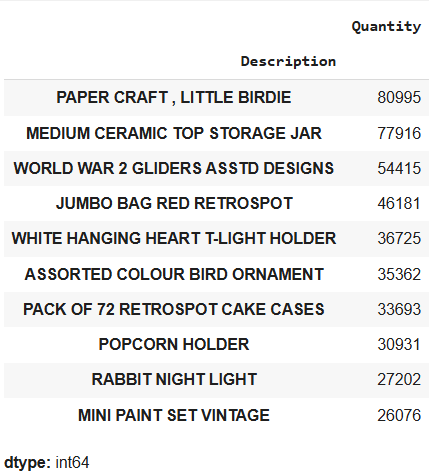
## 3.4.1. Tổng quan doanh thu

Dữ liệu cho thấy tổng doanh thu toàn bộ giai đoạn quan sát đạt **8.911.407,9 đô la Mỹ**, phản ánh quy mô đáng kể của hoạt động bán lẻ trực tuyến. Đây là con số sau khi đã tính toán bằng cách cộng tất cả các giá trị trong cột *TotalPrice* (được xác định bằng công thức: Quantity × UnitPrice).

## 3.4.2. Phân bố doanh thu theo quốc gia

Khi phân tích sâu hơn theo biến *Country*, có thể thấy sự chênh lệch rõ rệt giữa các thị trường:

## 3.4.3. Phân tích sản phẩm bán chạy



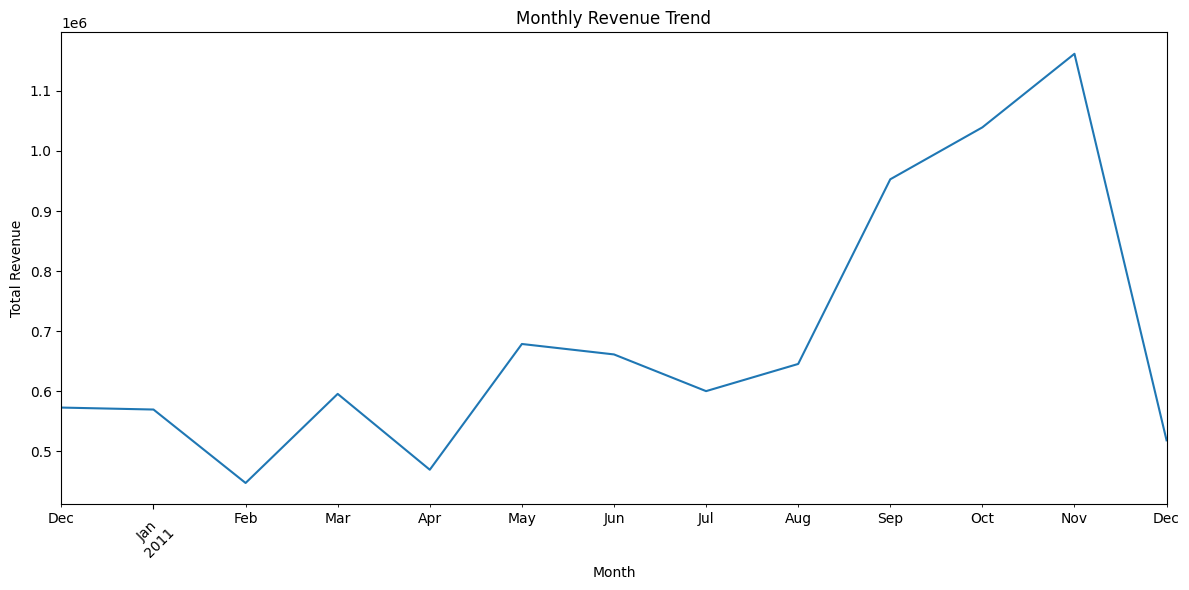
Khi sắp xếp theo số lượng bán ra (*Quantity*), có thể nhận thấy nhóm sản phẩm chủ đạo là đồ trang trí và vật dụng gia đình nhỏ gọn. Mười sản phẩm bán chạy nhất bao gồm:

Những sản phẩm này chủ yếu thuộc nhóm đồ trang trí nội thất, vật dụng thủ công và đồ gia dụng nhỏ, có giá trị đơn vị thấp nhưng tần suất mua cao. Điều này gợi ý rằng khách hàng của cửa hàng trực tuyến có xu hướng mua các mặt hàng quà tặng hoặc vật dụng nhỏ để sử dụng hằng ngày.

Đáng chú ý, các sản phẩm bán chạy nhất đều mang phong cách retro/vintage, cho thấy yếu tố thẩm mỹ hoài cổ có vai trò quan trọng trong hành vi tiêu dùng của khách hàng mục tiêu.

## 3.4.4. Phân tích doanh thu theo thời gian

Để đánh giá xu hướng doanh thu, dữ liệu được tổng hợp theo biến **InvoiceYearMonth**, tức là doanh thu của từng tháng trong năm.  
 Biểu đồ đường dưới đây thể hiện sự biến động doanh thu theo tháng trong giai đoạn khảo sát:

**

Biểu đồ cho thấy doanh thu có sự dao động đáng kể giữa các tháng, song nhìn chung thể hiện **xu hướng tăng dần rõ rệt từ giữa năm đến cuối năm**. Cụ thể:

* **Tháng 1 – Tháng 4:** doanh thu ở mức thấp, dao động quanh 0.45–0.6 triệu USD, cho thấy hoạt động kinh doanh chậm rãi trong giai đoạn đầu năm.
* **Tháng 5 – Tháng 7:** doanh thu bắt đầu cải thiện, đạt khoảng 0.65 triệu USD, tuy nhiên chưa có đột biến.
* **Tháng 8 – Tháng 11:** ghi nhận sự **tăng trưởng mạnh**, đặc biệt là tháng 10 và 11 với doanh thu đạt đỉnh **trên 1.1 triệu USD**.
* **Tháng 12:** doanh thu giảm mạnh, có thể do dữ liệu chưa đầy đủ cho tháng này (vì bộ dữ liệu dừng ở giữa tháng 12 năm 2011).

Như vậy, có thể nhận thấy **mùa cao điểm kinh doanh rơi vào quý IV hàng năm**, đặc biệt là giai đoạn chuẩn bị cho Giáng Sinh – một yếu tố thường thấy trong ngành bán lẻ.

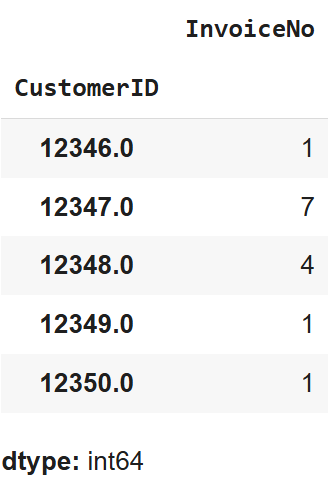
# 3.5. Phân tích khách hàng

## 3.5.1. Tổng quan về cơ sở khách hàng

Phân tích cho thấy có tổng cộng 4338 khách hàng duy nhất trong tập dữ liệu. Con số này cung cấp một cái nhìn tổng quan về quy mô của cơ sở khách hàng.

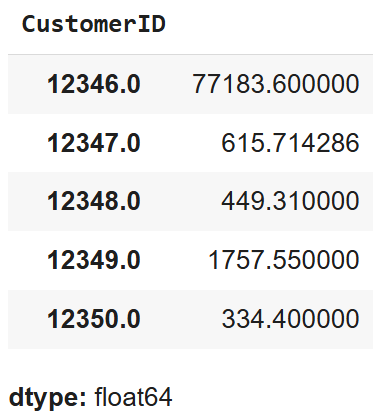
## 3.5.2. Phân tích hành vi mua hàng

Để hiểu hành vi mua hàng, ta tính toán tổng số đơn hàng cho mỗi khách hàng và doanh thu trung bình trên mỗi khách hàng.



Tổng số đơn hàng cho mỗi khách hàng: (Hiển thị dữ liệu đầu của orders\_per\_customer tại đây) Kết quả cho thấy sự khác biệt đáng kể về tần suất mua hàng giữa các khách hàng. Một số khách hàng chỉ đặt một đơn hàng, trong khi những người khác là người mua thường xuyên với nhiều đơn hàng.

## 3.5.4 Doanh thu trung bình trên mỗi khách hàng.



Doanh thu trung bình trên mỗi khách hàng cũng khác nhau, cho thấy giá trị lâu dài của khách hàng khác nhau. Một số khách hàng đóng góp một phần lớn vào tổng doanh thu, trong khi những người khác có giá trị giao dịch thấp hơn.

## 3.5.5. Phân tích khách hàng quay lại

Một chỉ số quan trọng về lòng trung thành của khách hàng là tỷ lệ khách hàng quay lại. Phân tích cho thấy tỷ lệ khách hàng quay lại là 65.58%. Tỷ lệ này cao, cho thấy một phần đáng kể khách hàng đã thực hiện nhiều hơn một lần mua hàng. Điều này chỉ ra một cơ sở khách hàng mạnh mẽ và tiềm năng cho các chiến lược giữ chân khách hàng.

# 3.6.Phân tích RFM (Recency–Frequency–Monetary)

## 3.6.1. Đếm tổng số đơn hàng duy nhất của mỗi khách hàng

Ở bước đầu tiên, dữ liệu được nhóm theo CustomerID để xác định tần suất mua hàng (Frequency) — tức số lượng hóa đơn (*InvoiceNo*) duy nhất mà mỗi khách hàng đã thực hiện

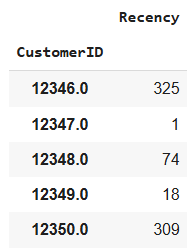
* Khách hàng **12347.0** có tần suất mua cao nhất (7 đơn hàng), thể hiện mức độ gắn bó cao.
* Ngược lại, phần lớn khách hàng chỉ mua 1 lần, cho thấy tỷ lệ khách hàng quay lại còn thấp.

## 3.6.2. Tính tổng doanh thu tạo ra bởi mỗi khách hàng

Ở bước này, cột TotalPrice được tính bằng công thức:

TotalPrice=Quantity×UnitPrice

Sau đó, tổng doanh thu theo từng khách hàng được tính.

****

Khách hàng 12346.0 tạo ra doanh thu vượt trội so với phần còn lại, chứng tỏ đây là khách hàng có giá trị cao.

Phần lớn khách hàng mang lại giá trị nhỏ hơn 5,000, thể hiện sự chênh lệch doanh thu khá lớn giữa các nhóm khách hàng.

## 3.6.3. Xây dựng bảng dữ liệu RFM

Sau khi xác định được ba chỉ tiêu chính (**Recency, Frequency, Monetary**), một DataFrame mới được tạo ra để phục vụ cho phân tích RFM.

**Recency** càng nhỏ → khách hàng mua hàng gần đây → càng có giá trị.

**Frequency** càng cao → khách hàng mua lặp lại nhiều → trung thành hơn.

**Monetary** càng cao → giá trị kinh tế càng lớn.

5.4. điểm RFM và phân khúc khách hàng

Các giá trị R, F, M được chia thành **5 mức (1–5)** dựa trên phân vị:

* R nhỏ → điểm R cao (mua gần đây).
* F lớn → điểm F cao (mua nhiều).
* M lớn → điểm M cao (chi tiêu cao).

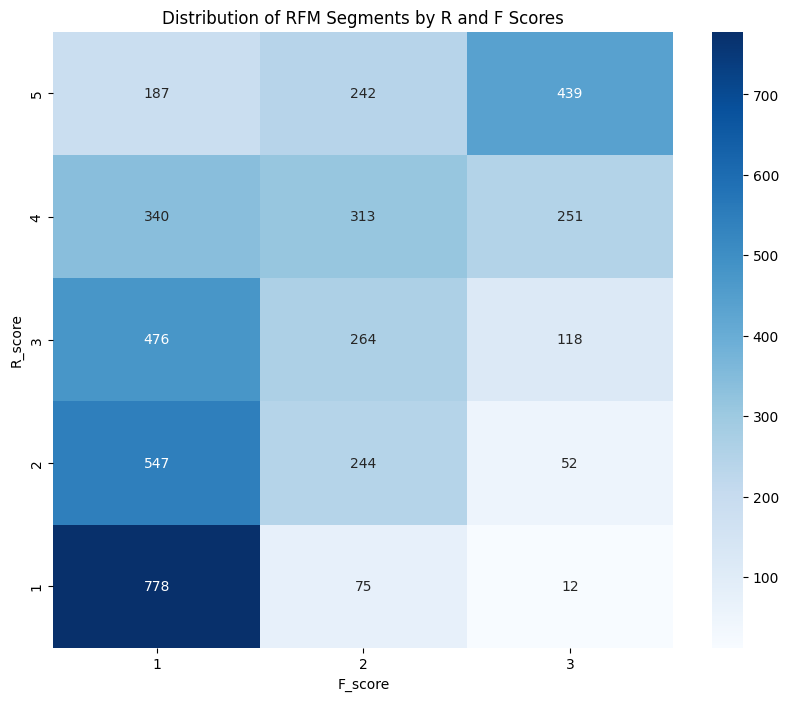
**High Value**: bao gồm khách hàng 12346.0 và 12347.0 — họ hoặc chi tiêu cao, hoặc mua thường xuyên và gần đây.

**Others**: là nhóm khách hàng hoạt động trung bình, có thể duy trì hoặc kích hoạt lại bằng các chiến dịch khuyến mãi.

**Churned Customers**: nhóm có điểm R thấp và Frequency thấp, có nguy cơ rời bỏ thương hiệu.

## 3.6.4. Trực quan hóa bản đồ nhiệt R–F

Để hiểu rõ hơn mối quan hệ giữa **Recency (R)** và **Frequency (F)** trong phân khúc khách hàng, biểu đồ **heatmap** được sử dụng:

**

Quan sát heatmap cho thấy:

* Các ô có giá trị lớn tập trung ở vùng **R thấp – F cao**, nghĩa là nhiều khách hàng mua hàng gần đây và thường xuyên → **đây là nhóm khách hàng cốt lõi**.
* Ngược lại, các ô **R cao – F thấp** thể hiện nhóm đã lâu không quay lại và ít giao dịch → cần **chiến lược tái kích hoạt**.

# 3.7. Phát hiện bất thường & kiểm tra chất lượng dữ liệu

## 3.7.1. Xác định các giá trị ngoại lai bằng phương pháp IQR

UnitPrice (Đơn giá) có tỷ lệ ngoại lai cao nhất (8.63%), cho thấy tồn tại một số sản phẩm có giá bán bất thường — có thể là lỗi nhập liệu (giá = 0 hoặc giá quá cao), hoặc là các mặt hàng đặc biệt có giá trị cao vượt trội.

Quantity (Số lượng) có tỷ lệ ngoại lai 6.45%, có thể bắt nguồn từ các giao dịch trả hàng (Quantity âm) hoặc đơn hàng mua với số lượng rất lớn.

TotalPrice (Tổng giá trị giao dịch) bị ảnh hưởng bởi cả hai yếu tố trên nên cũng xuất hiện khoảng 7.85% giá trị ngoại lai.

## 3.7.2 Biểu đồ hộp

Boxplot của UnitPrice: Biểu đồ này cho thấy phần lớn các giá trị 'UnitPrice' tập trung ở mức thấp, với một số lượng đáng kể các giá trị ngoại lai trải dài lên trên, cho thấy sự hiện diện của các mặt hàng có giá cao hơn nhiều so với phần còn lại.

A graph with numbers and a line

AI-generated content may be incorrect.

Boxplot của Quantity: Tương tự như 'UnitPrice', hầu hết các giá trị 'Quantity' đều nhỏ. Biểu đồ hộp minh họa rõ ràng sự tồn tại của một số giao dịch có số lượng sản phẩm rất lớn, được thể hiện bằng các điểm ngoại lai ở phía trên.

A screen shot of a computer

AI-generated content may be incorrect.

Boxplot của TotalPrice: Biểu đồ này cũng cho thấy sự phân phối bị lệch về bên phải với nhiều giá trị ngoại lai. Điều này phản ánh thực tế là hầu hết các giao dịch có tổng giá trị thấp, nhưng có một số giao dịch có tổng giá trị rất cao, có thể là do số lượng lớn hoặc giá đơn vị cao.

A white box with black text

AI-generated content may be incorrect.

Quantity vs. UnitPrice: Biểu đồ cho thấy hầu hết các giao dịch có số lượng và giá đơn vị thấp. Có một số điểm ngoại lai với số lượng hoặc giá cao. Không có mối quan hệ tuyến tính mạnh mẽ.

A white rectangular graph with numbers

AI-generated content may be incorrect.

## 3.7.3 Biều đồ phân tán

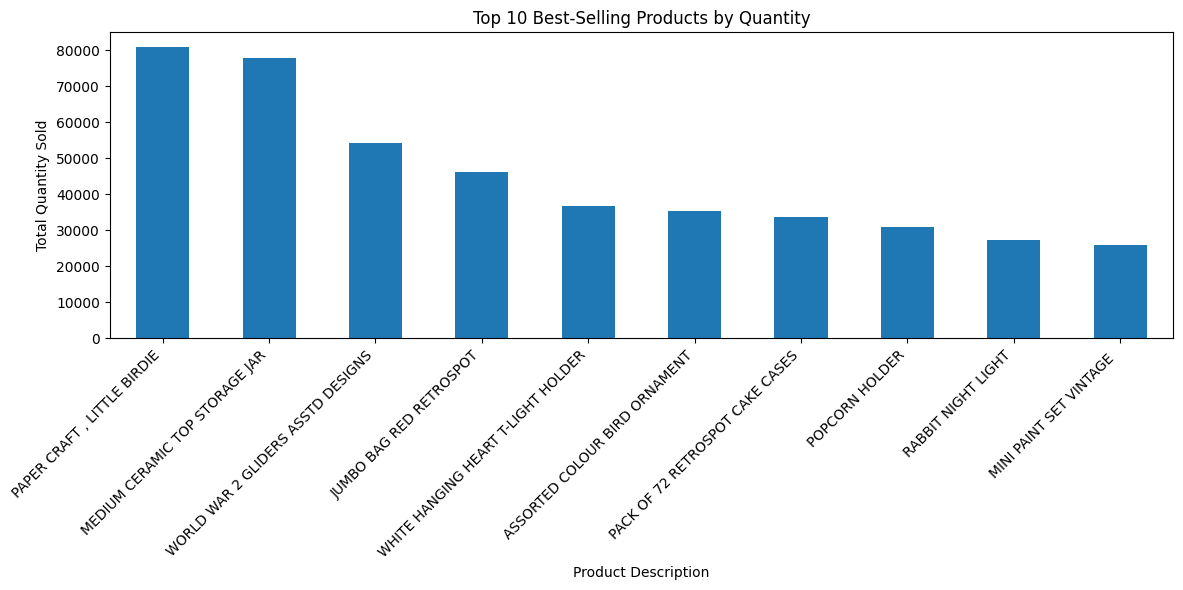
Quantity vs. TotalPrice: Tổng giá trị có xu hướng tăng khi số lượng tăng. Biểu đồ làm nổi bật các giao dịch có giá trị cao và các giá trị ngoại lai.

A screen shot of a graph

AI-generated content may be incorrect.

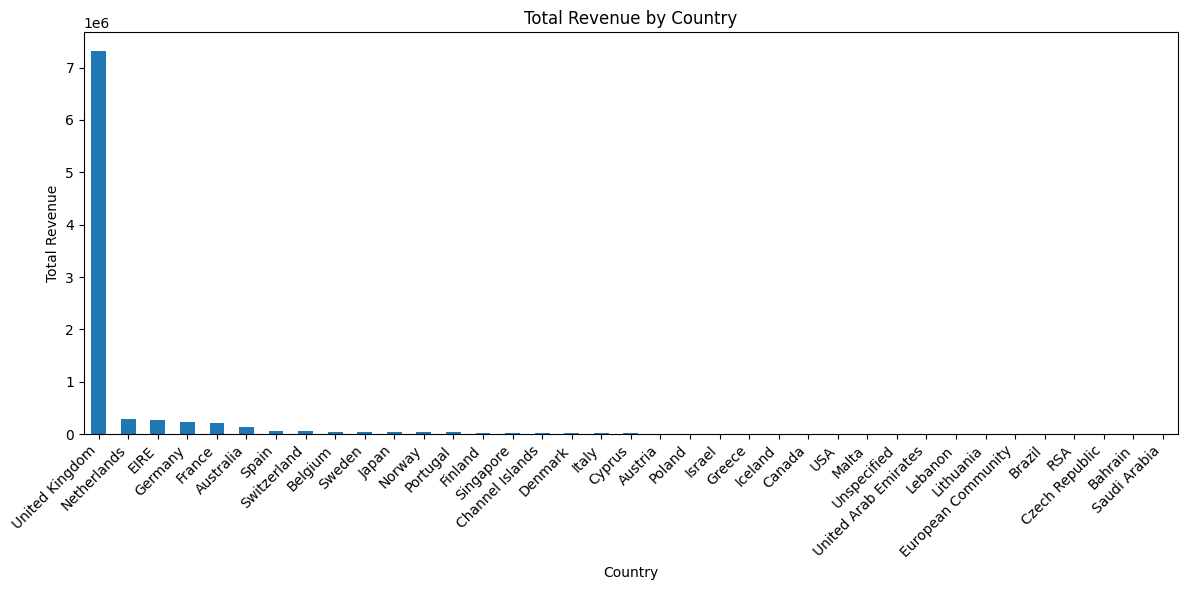
# 3.8. Trực quan hóa

## 3.8.1 Biểu đồ thanh để hiển thị 10 sản phẩm bán chạy nhất.

****

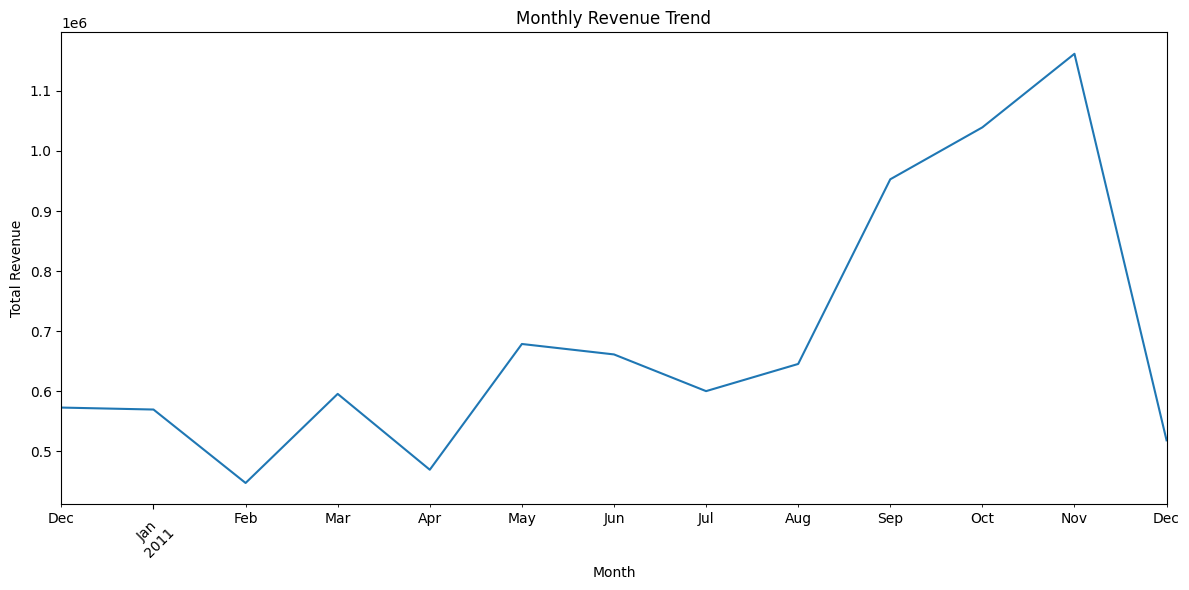
Biểu đồ này hiển thị rõ ràng 10 sản phẩm có số lượng bán ra cao nhất. Ta có thể dễ dàng xác định được những mặt hàng phổ biến nhất trong tập dữ liệu, với "PAPER CRAFT, LITTLE BIRDIE" là sản phẩm bán chạy nhất, theo sau là "MEDIUM CERAMIC TOP STORAGE JAR" và "WORLD WAR 2 GLIDERS ASSTD DESIGNS". Biểu đồ này cung cấp cái nhìn trực quan về sự đóng góp của từng sản phẩm hàng đầu vào tổng số lượng bán ra.

## 3.8.2 Biểu đồ thanh Tổng doanh thu theo quốc gia

****

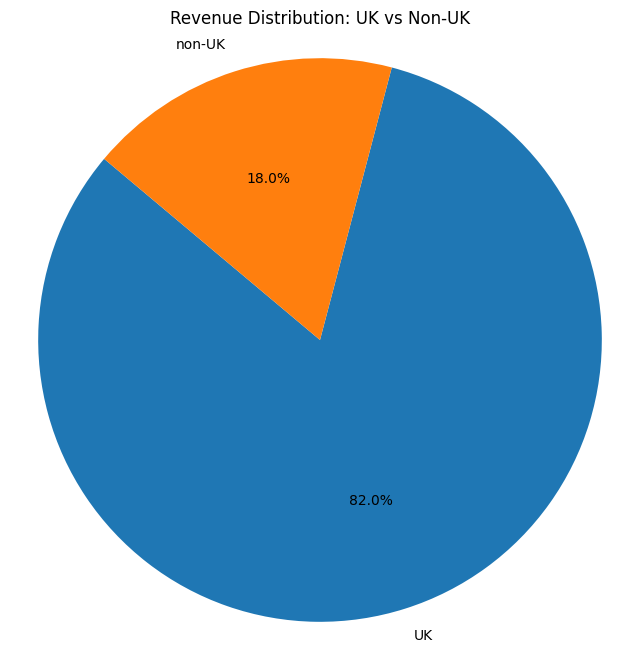
Biểu đồ này cho thấy sự phân bổ tổng doanh thu trên các quốc gia khác nhau. Rõ ràng là "United Kingdom" (Vương quốc Anh) chiếm phần lớn doanh thu, với tổng doanh thu cao hơn đáng kể so với các quốc gia khác. Các quốc gia khác như Hà Lan, EIRE, Đức và Pháp cũng có đóng góp đáng kể vào doanh thu, nhưng ở mức độ thấp hơn nhiều so với Vương quốc Anh. Biểu đồ này giúp xác định các thị trường quan trọng nhất về mặt doanh thu.

## 3.8.3 Biểu đồ chuỗi thời gian thể hiện xu hướng tổng doanh thu qua nhiều tháng.



Biểu đồ chuỗi thời gian xu hướng doanh thu hàng tháng: Biểu đồ đường này cho thấy sự thay đổi của tổng doanh thu theo thời gian, theo từng tháng. Có thể quan sát được xu hướng tăng trưởng doanh thu trong suốt giai đoạn được phân tích, với sự sụt giảm nhẹ vào đầu năm 2011 và sau đó tăng trưởng mạnh mẽ, đặc biệt là vào cuối năm 2011 (tháng 11 và 12), cho thấy tính thời vụ trong doanh số bán hàng, có thể liên quan đến các dịp lễ hoặc mua sắm cuối năm.

## 3.8.4 Biểu đồ hình tròn phân phối doanh thu

****

Vương quốc Anh so với ngoài Vương quốc Anh: Biểu đồ này cho thấy tỷ lệ đóng góp doanh thu của thị trường Vương quốc Anh so với tất cả các thị trường khác (non-UK). Rõ ràng là thị trường Vương quốc Anh chiếm một phần rất lớn trong tổng doanh thu, khẳng định tầm quan trọng vượt trội của thị trường nội địa này đối với doanh nghiệp.