

Ứng dụng khoa học dữ liệu trong Marketing

Xây dựng hệ thống gợi ý sản phẩm

Nhóm 7

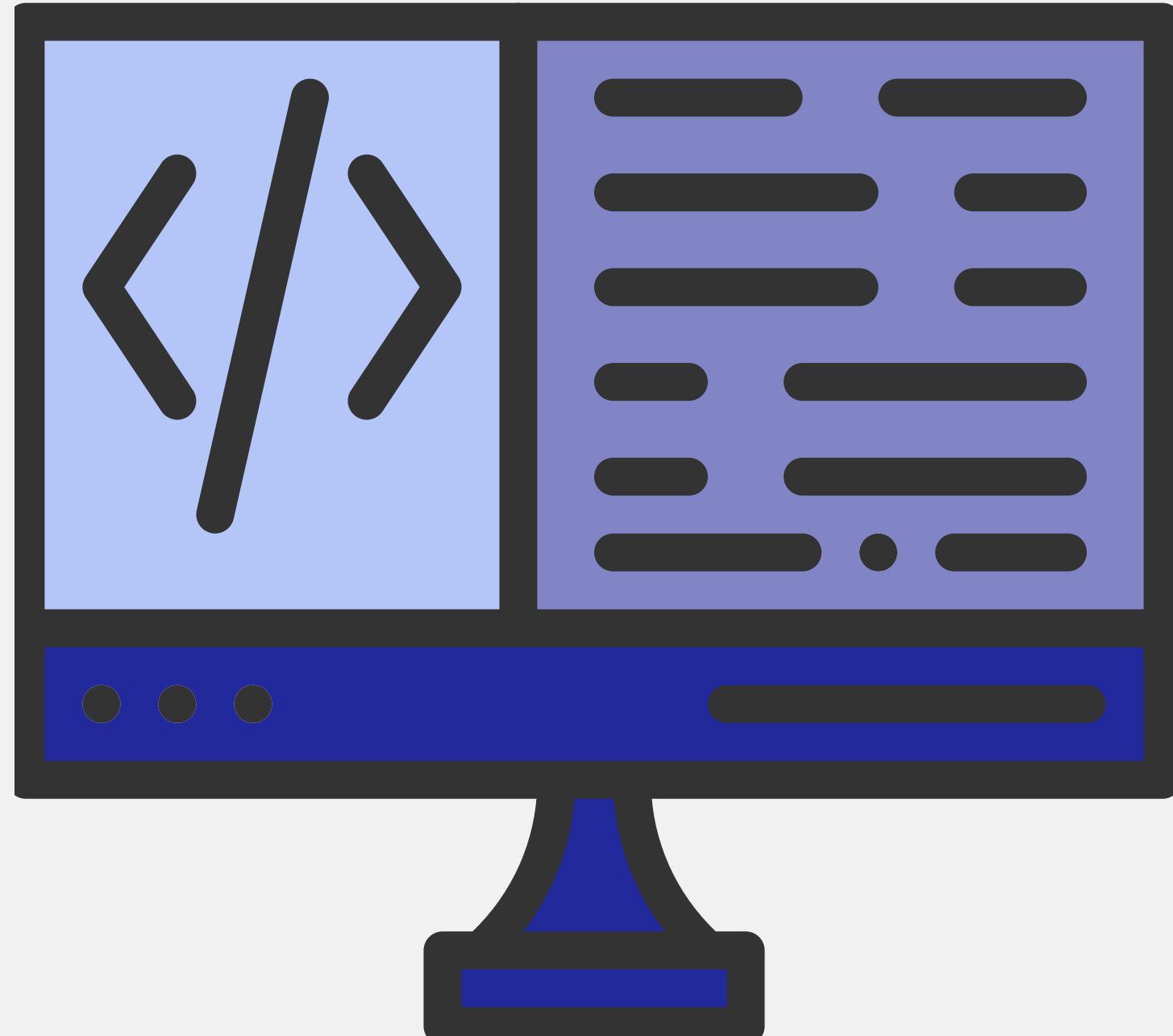
Trần Thị Kiều Oanh

Phạm Hạnh Nguyên

Nghiêm Gia Phương

Nguyễn Như Quỳnh

Nguyễn Thị Trà



NỘI DUNG

Xử lý dữ liệu

Phân tích khám phá dữ liệu

Xây dựng hệ thống gợi ý sản phẩm

Thông tin tập dữ liệu

- Sử dụng tập dữ liệu Online Retail (UCI ID = 352) trên Kaggle
- Kiểu dữ liệu: Dữ liệu gộp (Pooled data)
- Các biến bao gồm: InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, Country
- Kích thước dữ liệu: 541909 dòng x 8 cột

Xử lý dữ liệu



Xử lý dữ liệu

Vấn đề 1: Định dạng dữ liệu Ngày tháng chưa phù hợp để phân tích

→ Chuyển kiểu dữ liệu của InvoiceDate về datetime

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 541909 entries, 0 to 541908
Data columns (total 8 columns):
 #   Column      Non-Null Count  Dtype  
---  --  
 0   InvoiceNo    541909 non-null   object 
 1   StockCode     541909 non-null   object 
 2   Description   540455 non-null   object 
 3   Quantity      541909 non-null   int64  
 4   InvoiceDate   541909 non-null   object 
 5   UnitPrice     541909 non-null   float64
 6   CustomerID   406829 non-null   float64
 7   Country       541909 non-null   object 
dtypes: float64(2), int64(1), object(5)
memory usage: 33.1+ MB
```



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 541909 entries, 0 to 541908
Data columns (total 8 columns):
 #   Column      Non-Null Count  Dtype  
---  --  
 0   InvoiceNo    541909 non-null   object 
 1   StockCode     541909 non-null   object 
 2   Description   540455 non-null   object 
 3   Quantity      541909 non-null   int64  
 4   InvoiceDate   541909 non-null   datetime64[ns]
 5   UnitPrice     541909 non-null   float64
 6   CustomerID   406829 non-null   float64
 7   Country       541909 non-null   object 
dtypes: datetime64[ns](1), float64(2), int64(1), object(4)
memory usage: 33.1+ MB
```

Xử lý dữ liệu

Vấn đề 2: Dữ liệu gốc chưa được lọc theo khoảng thời gian yêu cầu phân tích của đề bài
→ Lọc dữ liệu gốc trong giai đoạn từ 7/2011-11/2011

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
245903	558638	84836 ZINC METAL HEART DECORATION	12	2011-07-01 08:16:00	1.25	16317.0	United Kingdom
245904	558638	71459 HANGING JAM JAR T-LIGHT HOLDER	24	2011-07-01 08:16:00	0.85	16317.0	United Kingdom
245905	558638	22784 LANTERN CREAM GAZEBO	3	2011-07-01 08:16:00	4.95	16317.0	United Kingdom
245906	558638	23145 ZINC T-LIGHT HOLDER STAR LARGE	12	2011-07-01 08:16:00	0.95	16317.0	United Kingdom
245907	558638	22674 FRENCH TOILET SIGN BLUE METAL	12	2011-07-01 08:16:00	1.25	16317.0	United Kingdom

Dữ liệu sử dụng để phân tích là dữ liệu đã lọc từ 7/2011-11/2011. Kết cấu dữ liệu gồm 267027 dòng x 8 cột

Xử lý dữ liệu

Vấn đề 3: Dữ liệu bị thiếu

→ Loại bỏ các quan sát bị thiếu

- **Thiếu CustomerID**

Bản ghi không gắn kết với bất kì cá nhân cụ thể nào, dẫn đến thiếu tính liên kết trong phân tích

- **Thiếu Description**

Giảm khả năng hiểu rõ nội dung giao dịch

Kiểm tra dữ liệu bị thiếu

CustomerID	56943
Description	511
InvoiceNo	0
StockCode	0
Quantity	0
InvoiceDate	0
UnitPrice	0
Country	0
dtype:	int64

Xử lý dữ liệu

Vấn đề 4: Dữ liệu bị trùng lặp

→ Loại bỏ các quan sát bị trùng lặp

Kiểm tra thấy 2944 bản ghi trùng lặp

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
246406	558700	22332 SKULLS PARTY BAG + STICKER SET	1	2011-07-01 12:33:00	1.65	17920.0	United Kingdom
246420	558700	21242 RED RETROSPOT PLATE	1	2011-07-01 12:33:00	1.69	17920.0	United Kingdom
246425	558700	22553 PLASTERS IN TIN SKULLS	1	2011-07-01 12:33:00	1.65	17920.0	United Kingdom
246430	558700	21242 RED RETROSPOT PLATE	1	2011-07-01 12:33:00	1.69	17920.0	United Kingdom
246445	558700	21544 SKULLS WATER TRANSFER TATTOOS	3	2011-07-01 12:33:00	0.85	17920.0	United Kingdom
...
512806	579516	23426 METAL SIGN DROP YOUR PANTS	1	2011-11-29 17:52:00	2.89	17841.0	United Kingdom
512824	579516	22692 DOORMAT WELCOME TO OUR HOME	1	2011-11-29 17:52:00	8.25	17841.0	United Kingdom
512829	579516	22692 DOORMAT WELCOME TO OUR HOME	1	2011-11-29 17:52:00	8.25	17841.0	United Kingdom
512911	579520	22748 POPPY'S PLAYHOUSE KITCHEN	1	2011-11-29 18:14:00	2.10	12748.0	United Kingdom
512924	579520	22230 JIGSAW TREE WITH WATERING CAN	1	2011-11-29 18:14:00	0.29	12748.0	United Kingdom

2944 rows × 8 columns

Xử lý dữ liệu

Vấn đề 5: Dữ liệu mang các giá trị bất thường

→ Loại bỏ các quan sát có giá trị bất thường

	Quantity	InvoiceDate	UnitPrice	CustomerID
count	207140.000000	207140	207140.000000	207140.000000
mean	12.045192	2011-09-29 20:31:17.529593344	3.261213	15286.684523
min	-1296.000000	2011-07-01 08:16:00	0.000000	12347.000000
25%	2.000000	2011-08-26 12:43:15	1.250000	13988.000000
50%	5.000000	2011-10-06 13:29:00	1.690000	15154.000000
75%	12.000000	2011-11-06 14:29:00	3.750000	16764.000000
max	12540.000000	2011-11-29 18:19:00	4287.630000	18287.000000
std	49.397287	Nan	29.888279	1704.350006

Bảng thống kê mô tả cho các biến Quantity, InvoiceDate, UnitPrice và CustomerID

Xử lý dữ liệu

Vấn đề 5: Dữ liệu mang các giá trị bất thường

→ Loại bỏ các quan sát có giá trị bất thường



Dữ liệu có những đơn hàng cho số lượng sản phẩm và giá trị sản phẩm là âm.
Đây là các đơn hàng bị huỷ.

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
246352	C558698	20719 WOODLAND CHARLOTTE BAG	-1	2011-07-01 12:11:00	0.85	16746.0	United Kingdom
246353	C558698	23205 CHARLOTTE BAG VINTAGE ALPHABET	-1	2011-07-01 12:11:00	0.85	16746.0	United Kingdom
246354	C558698	20725 LUNCH BAG RED RETROSPOT	-1	2011-07-01 12:11:00	1.65	16746.0	United Kingdom
246708	C558712	M Manual	-1	2011-07-01 13:06:00	2.95	17338.0	United Kingdom
246709	C558712	21735 TWO DOOR CURIO CABINET	-1	2011-07-01 13:06:00	12.75	17338.0	United Kingdom



Loại bỏ các đơn hàng đã bị huỷ (có Quantity và UnitPrice ≤ 0)

Xử lý dữ liệu

Vấn đề 6: Có thể xuất hiện nhiều Description trong một Stockcode

→ Với mỗi Stockcode, chọn số Description xuất hiện nhiều nhất làm tên chuẩn cho sản phẩm

Kiểm tra số Description của mỗi Stockcode

	StockCode	Distinct Descriptions
1890	23236	4
1850	23196	4
1885	23231	3
1894	23240	3
1786	23131	3
...
23	16008	1
24	16011	1
25	16012	1
26	16014	1
27	16015	1

[3164 rows x 2 columns]

Chọn 1 StockCode xuất hiện nhiều lần để xem xét

StockCode	Description	count
1950	23236 DOILEY BISCUIT TIN	3
1951	23236 DOILEY STORAGE TIN	64
1952	23236 STORAGE TIN VINTAGE DOILEY	1
1953	23236 STORAGE TIN VINTAGE DOILY	154

Xử lý dữ liệu

Vấn đề 6: Có thể xuất hiện nhiều Description trong một Stockcode

→ Với mỗi Stockcode, chọn số Description xuất hiện nhiều nhất làm tên chuẩn cho sản phẩm

Kiểm tra lại kết quả sau khi đặt tên chuẩn

StockCode		Description	Standard_Description
245903	84836	ZINC METAL HEART DECORATION	ZINC METAL HEART DECORATION
245904	71459	HANGING JAM JAR T-LIGHT HOLDER	HANGING JAM JAR T-LIGHT HOLDERS
245905	22784	LANTERN CREAM GAZEBO	LANTERN CREAM GAZEBO
245906	23145	ZINC T-LIGHT HOLDER STAR LARGE	ZINC T-LIGHT HOLDER STAR LARGE
245907	22674	FRENCH TOILET SIGN BLUE METAL	FRENCH TOILET SIGN BLUE METAL
245908	21174	POTTERING IN THE SHED METAL SIGN	POTTERING IN THE SHED METAL SIGN
245909	22413	METAL SIGN TAKE IT OR LEAVE IT	METAL SIGN TAKE IT OR LEAVE IT
245910	22726	ALARM CLOCK BAKELIKE GREEN	ALARM CLOCK BAKELIKE GREEN
245911	23032	DRAWER KNOB CRACKLE GLAZE IVORY	DRAWER KNOB CRACKLE GLAZE IVORY
245912	23251	VINTAGE RED ENAMEL TRIM MUG	VINTAGE RED ENAMEL TRIM MUG

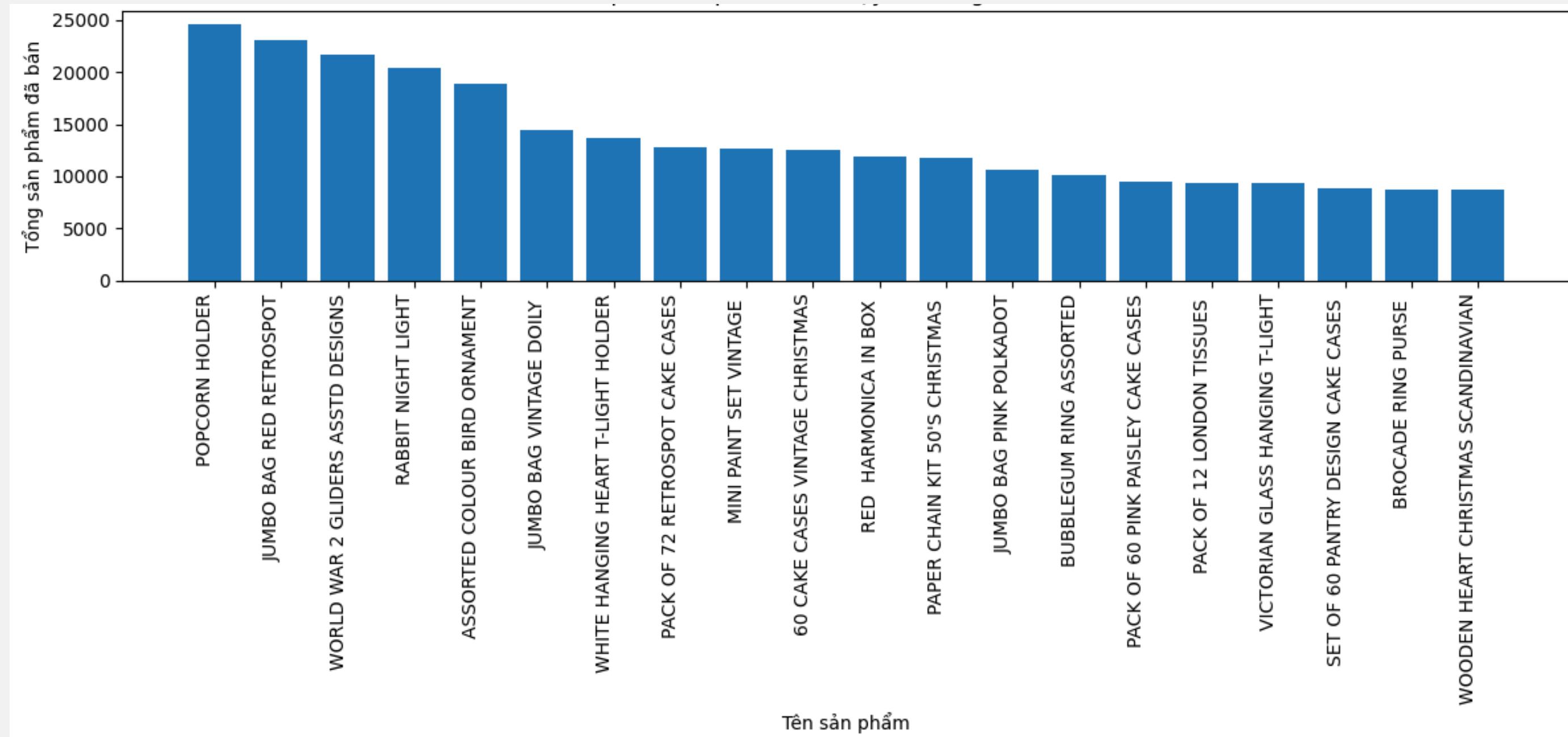
Phân tích khám phá dữ liệu



TOP 20 SẢN PHẨM BÁN CHẠY THEO DOANH SỐ

Tính từ tháng 7 - 11/2021

Theo chỉ tiêu tổng số lượng sản phẩm đã bán được



Tổng số đơn và doanh thu bán hàng



8853 đơn



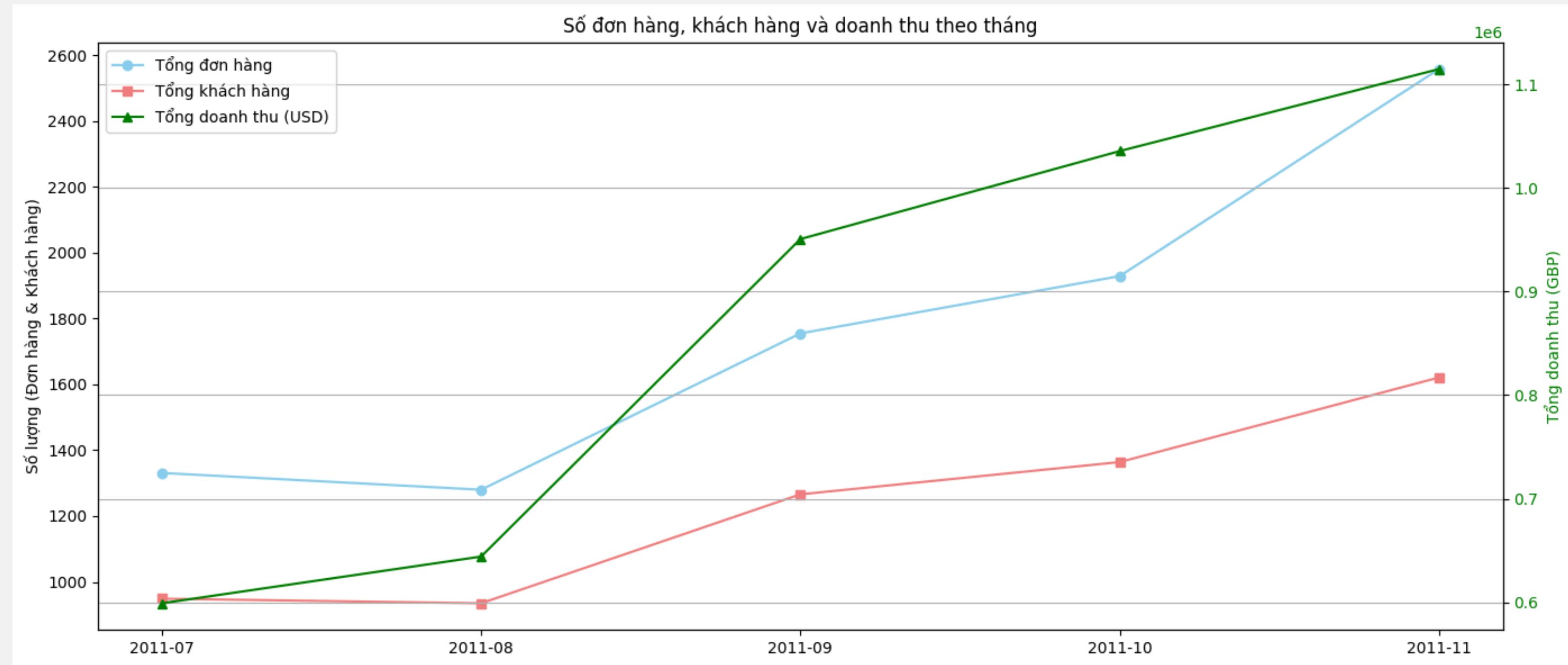
4,344,155.41 GBP

Các giá trị trung bình

- Giá trị trung bình mỗi đơn hàng: **490.7 GBP/đơn hàng**
- Tần suất mua hàng trung bình của mỗi khách: **2.71 đơn hàng/khách**
- Giá trị mua hàng trung bình mỗi khách hàng: **1328.08 GBP/khách hàng**

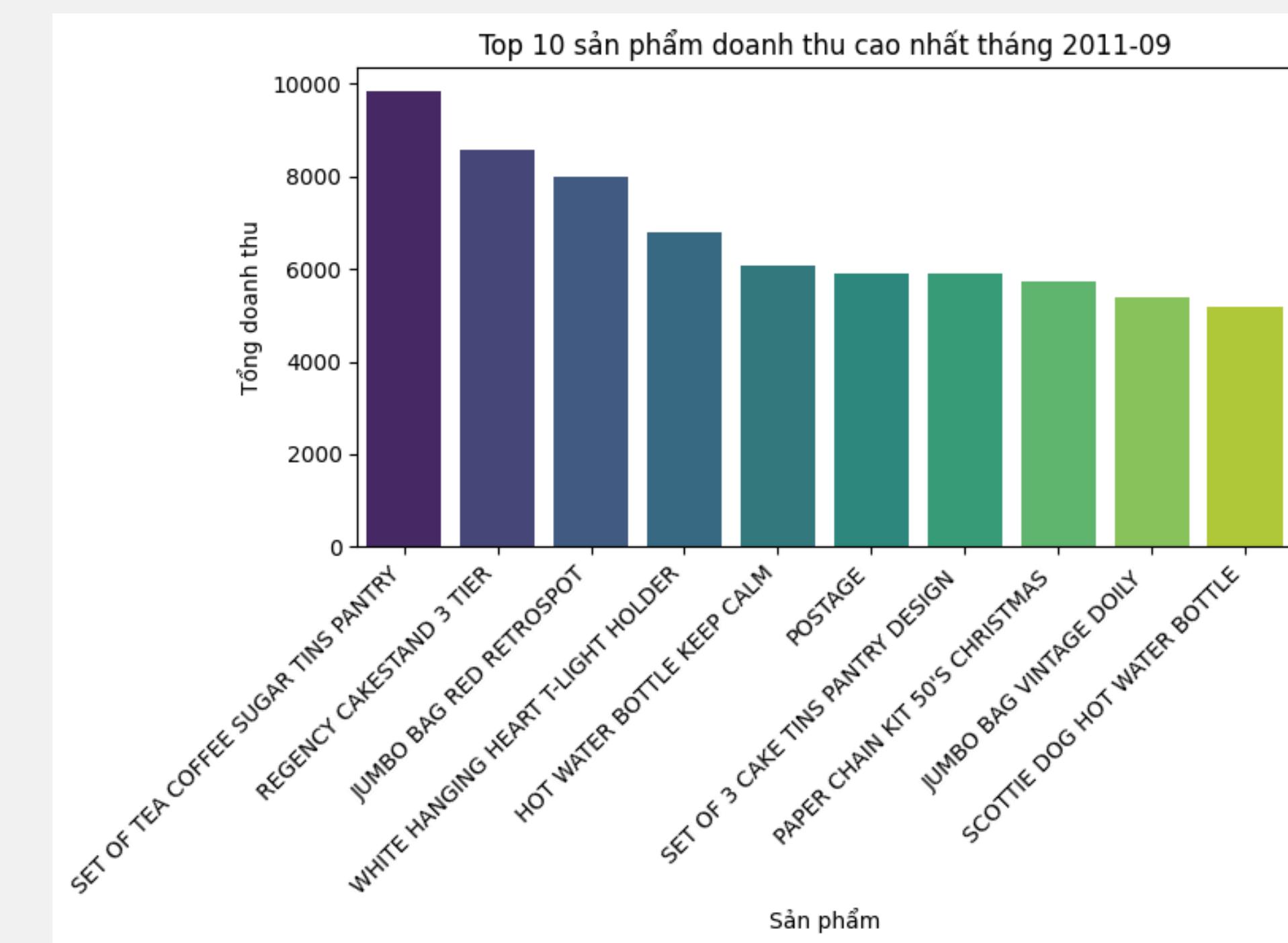
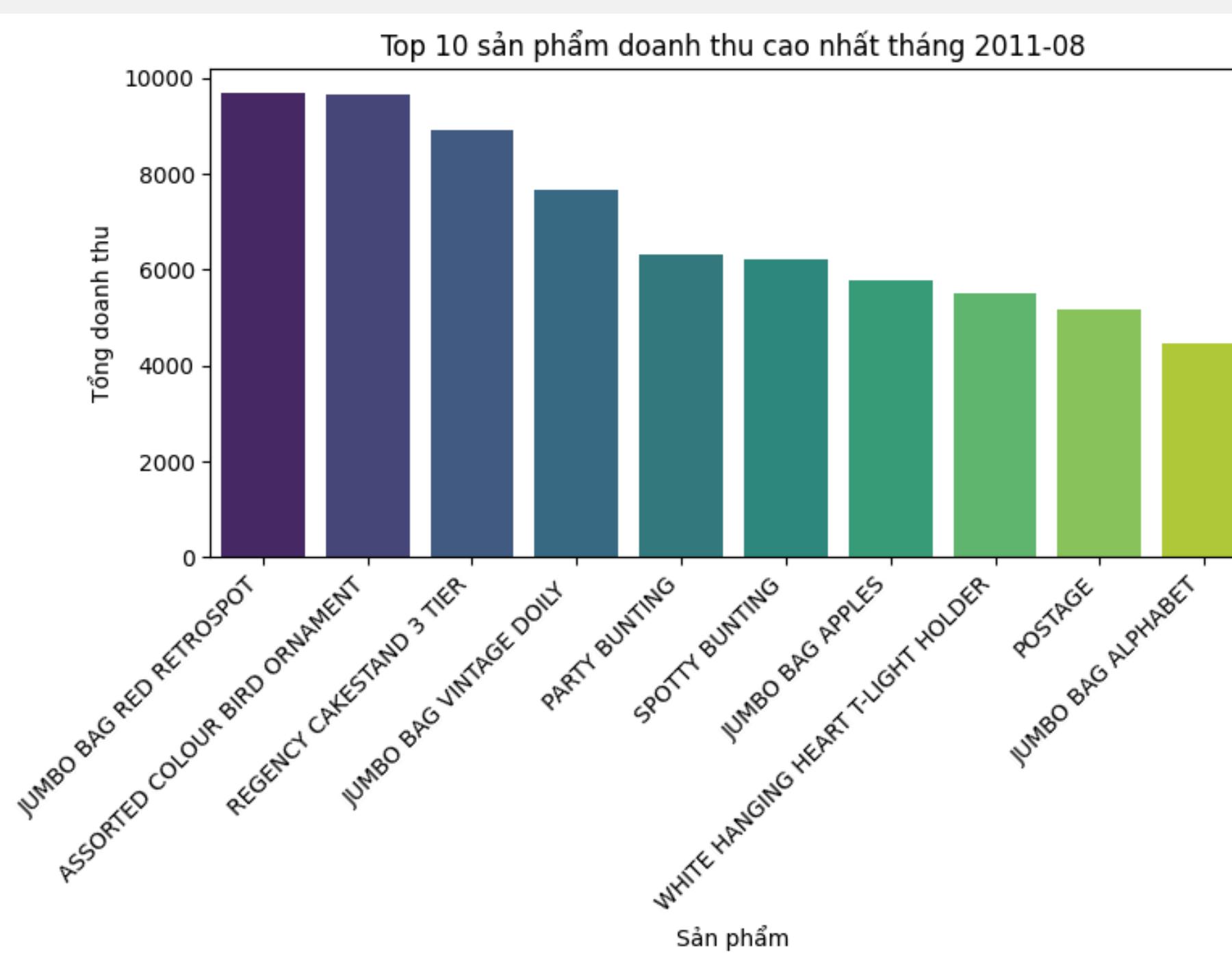
SỐ ĐƠN HÀNG, KHÁCH HÀNG VÀ DOANH THU THEO THÁNG

Tính từ tháng 7 - 11/2011



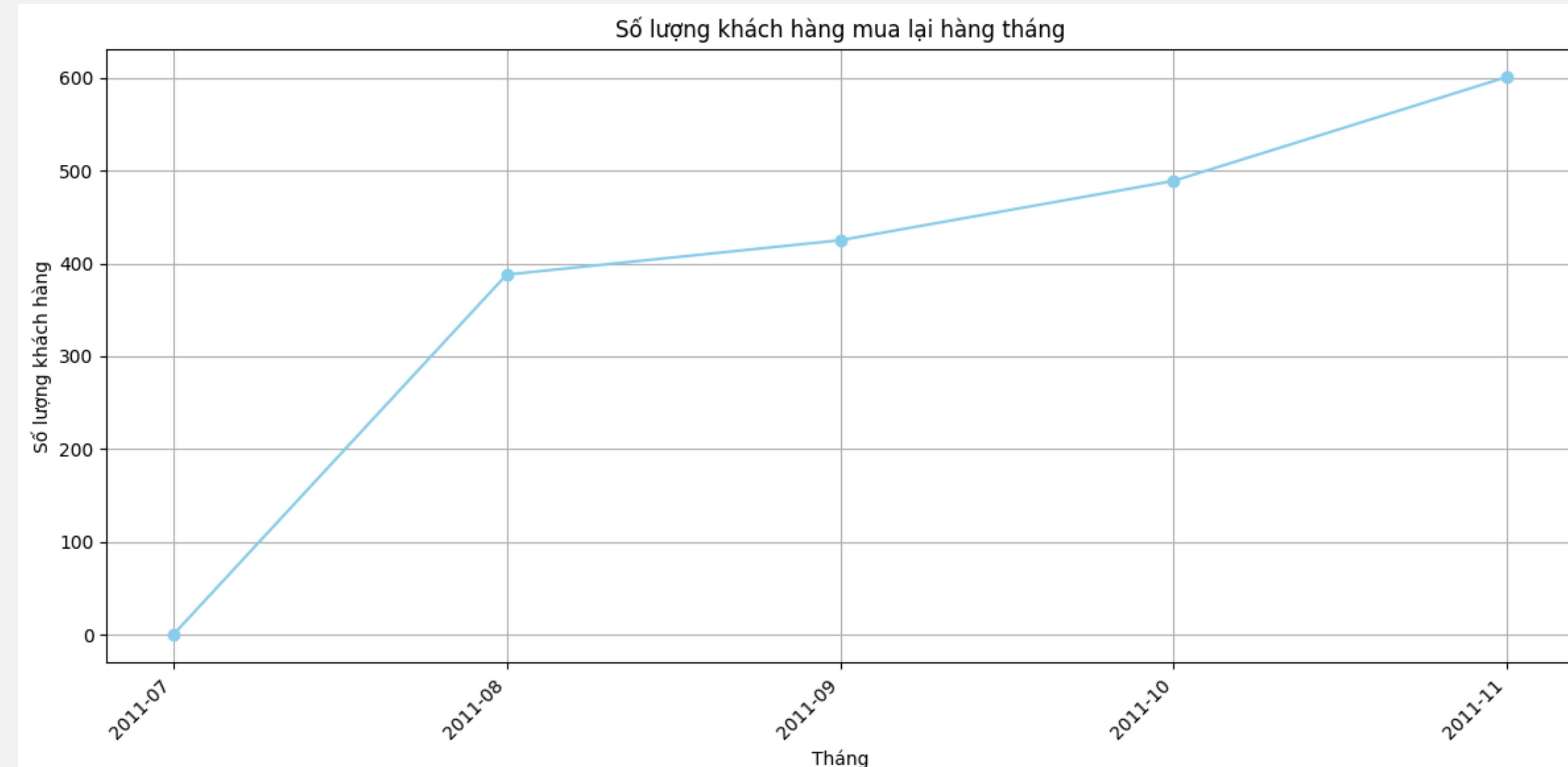
TOP 10 SẢN PHẨM BÁN CHẠY NHẤT THEO DOANH THU

Chỉ xét tháng 8 và tháng 9 do chênh lệch doanh thu của hai tháng này rất lớn

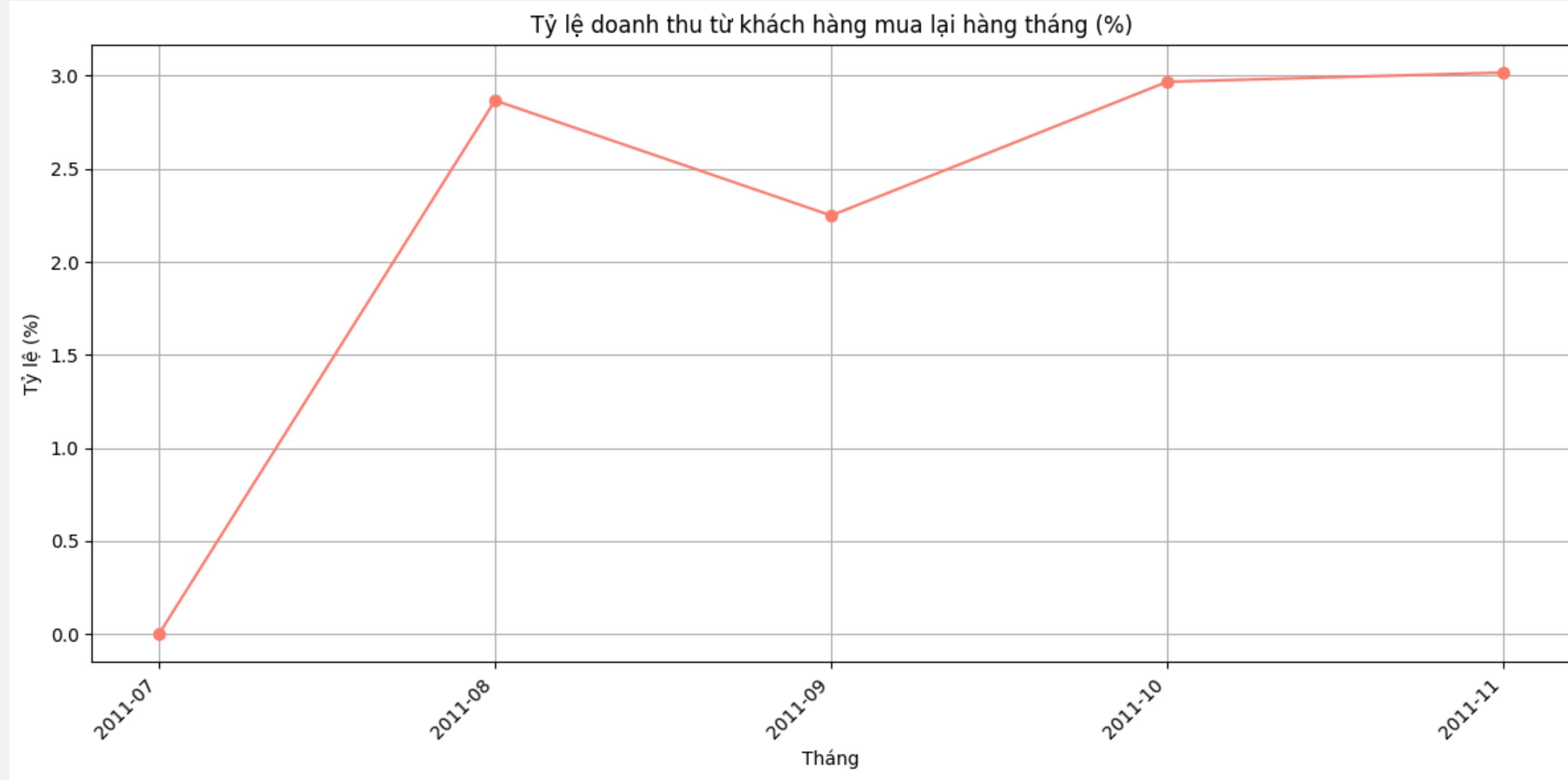


SỐ LƯỢNG KHÁCH HÀNG MUA LẠI HÀNG THÁNG

Số khách hàng mua lại hàng tháng được hiểu là số lượng khách hàng đã từng mua hàng ở tháng liền kề trước đó có mua hàng ở tháng này.



TỶ LỆ DOANH THU TỪ KHÁCH HÀNG MUA LẠI HÀNG THÁNG



Xây dựng hệ thống gợi ý sản phẩm



Dựa trên khách hàng tương tự

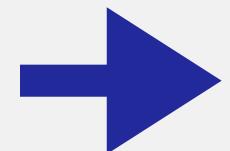
Tạo Pivot Table với index là CustomerID, các cột là StockCode, và các giá trị trong bảng là tổng số sản phẩm đã mua từ mỗi StockCode của từng khách hàng.

Dựa trên khách hàng tương tự

Sử dụng cosine similarity để tính toán độ tương đồng giữa các khách hàng dựa trên sản phẩm đã mua

CustomerID	12347.0	12348.0	12349.0	12352.0	12356.0	12357.0	12358.0	12359.0	12360.0	12362.0	...	18270.0	18272.0	18273.0	18274.0	18276.0	18277.0	18278.0	18282.0	18283.0	18287.0
CustomerID																					
12347.0	1.000000	0.000000	0.021480	0.018758	0.030490	0.020937	0.000000	0.089429	0.002378	0.095512	...	0.0	0.015968	0.0	0.001743	0.130978	0.0	0.020149	0.000000	0.134964	0.000000
12348.0	0.000000	1.000000	0.000069	0.000617	0.000000	0.000000	0.000397	0.000000	0.000352	0.000930	...	0.0	0.000000	0.0	0.000000	0.258424	0.0	0.000000	0.000000	0.180323	0.151088
12349.0	0.021480	0.000069	1.000000	0.038646	0.005858	0.103707	0.189028	0.058146	0.045810	0.098037	...	0.0	0.110937	0.0	0.148066	0.000000	0.0	0.015680	0.013543	0.065477	0.000000
12352.0	0.018758	0.000617	0.038646	1.000000	0.021033	0.046353	0.005791	0.019465	0.006973	0.045823	...	0.0	0.171016	0.0	0.040892	0.000000	0.0	0.000000	0.000000	0.001347	0.000000
12356.0	0.030490	0.000000	0.005858	0.021033	1.000000	0.031325	0.000000	0.092541	0.000000	0.003446	...	0.0	0.000000	0.0	0.017594	0.000000	0.0	0.000000	0.000000	0.000000	0.000000

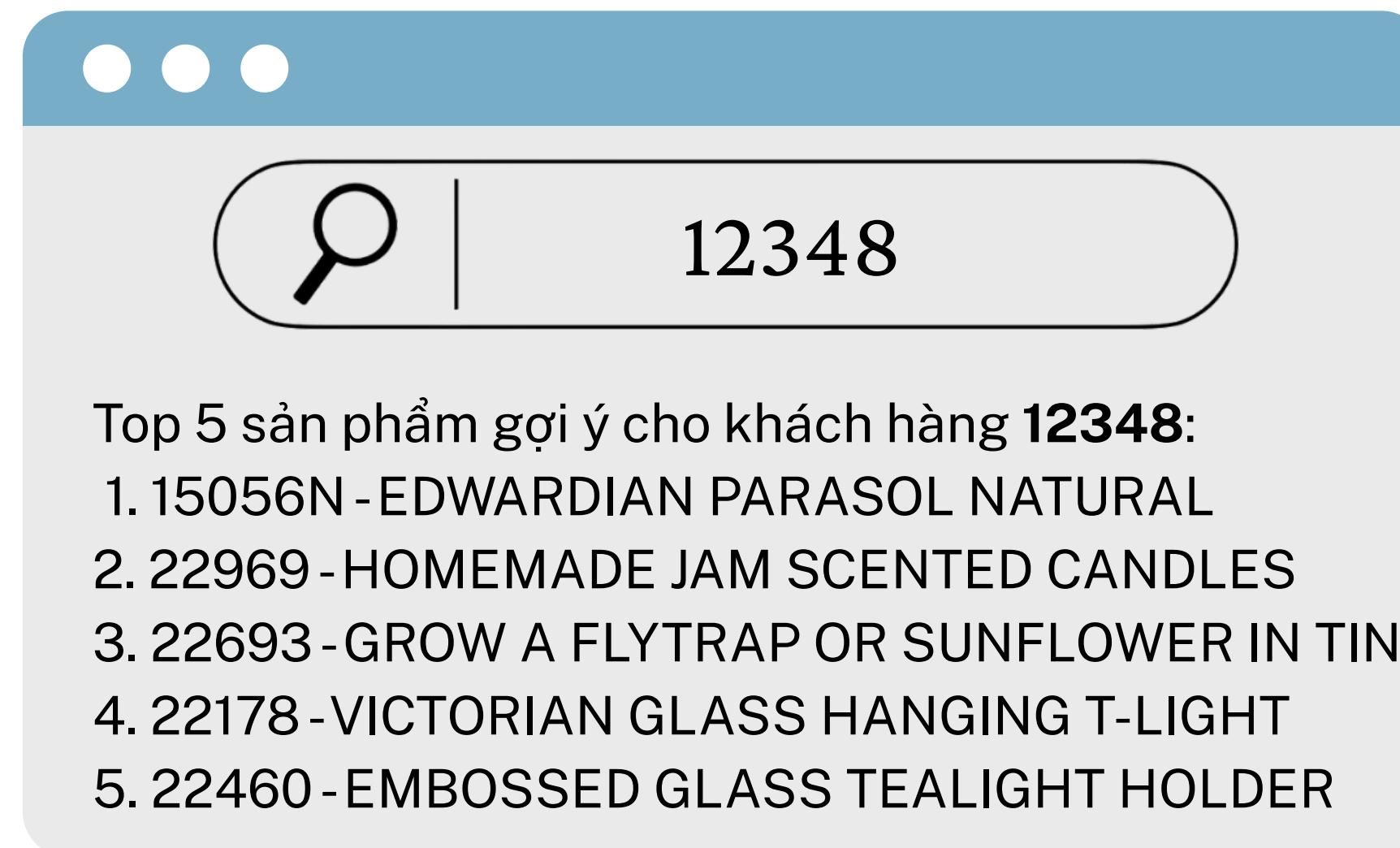
5 rows × 3271 columns



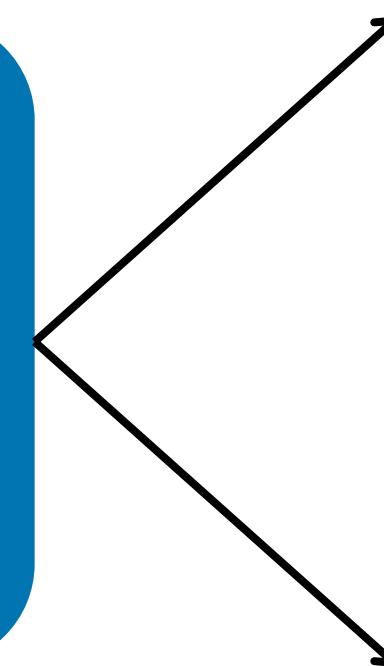
Với mỗi CustomerID, lấy ra top 5 khách hàng tương tự nhất

Dựa trên khách hàng tương tự

Dựa trên kết quả khách hàng tương tự, ta có thể tạo ra một hệ thống gợi ý sản phẩm cho từng khách hàng. Với mỗi khách hàng, ta có thể tìm kiếm bằng CustomerID. Từ đó, hệ thống sẽ gợi ý 5 sản phẩm có số lượng mua trung bình cao nhất trong nhóm các khách hàng tương tự, nhưng chưa từng được khách hàng hiện tại mua.



Dựa trên khách hàng tương tự



1. Gợi ý cá nhân hóa trên web/app

Khi khách đang xem sản phẩm, hiện “Bạn có thể thích những sản phẩm này”

2. Email marketing riêng biệt

Gửi email định kỳ cho khách hàng, trong đó giới thiệu 5 sản phẩm phù hợp nhất dựa trên hành vi mua hàng của những khách hàng tương tự.

APRIORI: Gợi ý sản phẩm mua kèm

- Nhóm các sản phẩm lại theo từng đơn hàng (InvoiceNo)
 - Chuyển đổi dữ liệu giỏ hàng thành dạng nhị phân, trong đó mỗi sản phẩm có giá trị True nếu có trong giỏ hàng, và False nếu không có.

APRIORI: Gợi ý sản phẩm mua kèm

- Áp dụng thuật toán Apriori để tìm các nhóm sản phẩm xuất hiện ít nhất 2% trong các đơn
- Dựa trên các nhóm sản phẩm, tìm ra các luật kết hợp sử dụng chỉ số Lift. Chỉ số Lift ≥ 1 cho thấy có mối quan hệ mua kèm giữa các sản phẩm.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	representativity	leverage	conviction	zhangs_metric	jaccard	certainty	kulczynski	
0	(ALARM CLOCK BAKELIKE GREEN)	(ALARM CLOCK BAKELIKE PINK)	0.051508	0.036824	0.023495	0.456140	12.387149		1.0	0.021598	1.771002	0.969192	0.362369	0.435348	0.547089
1	(ALARM CLOCK BAKELIKE PINK)	(ALARM CLOCK BAKELIKE GREEN)	0.036824	0.051508	0.023495	0.638037	12.387149		1.0	0.021598	2.620410	0.954416	0.362369	0.618380	0.547089
2	(ALARM CLOCK BAKELIKE RED)	(ALARM CLOCK BAKELIKE GREEN)	0.047667	0.051508	0.034000	0.713270	13.847764		1.0	0.031545	3.307964	0.974225	0.521664	0.697699	0.686679
3	(ALARM CLOCK BAKELIKE GREEN)	(ALARM CLOCK BAKELIKE RED)	0.051508	0.047667	0.034000	0.660088	13.847764		1.0	0.031545	2.801701	0.978170	0.521664	0.643074	0.686679
4	(ALARM CLOCK BAKELIKE RED)	(ALARM CLOCK BAKELIKE PINK)	0.047667	0.036824	0.023495	0.492891	13.385166		1.0	0.021740	1.899348	0.971604	0.385185	0.473503	0.565464

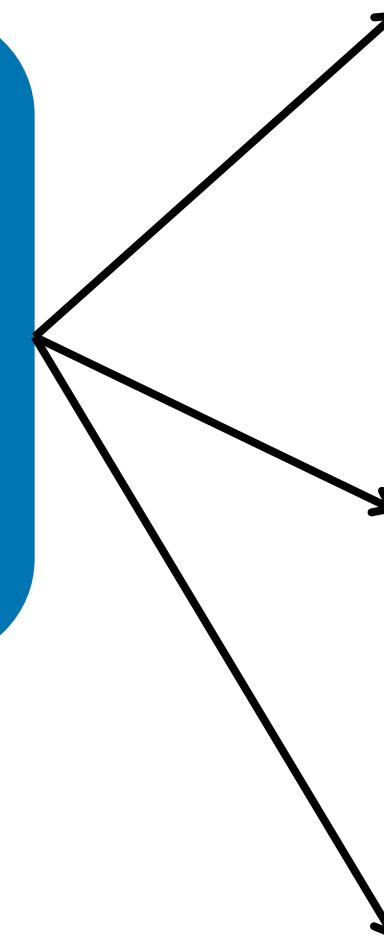
APRIORI: Gợi ý sản phẩm mua kèm

Biểu đồ mạng các sản phẩm thường mua kèm



Biểu đồ mạng các sản phẩm mua kèm

APRIORI: Gợi ý sản phẩm mua kèm



1. Trưng bày sản phẩm cạnh nhau

Đặt các sản phẩm thường được mua cùng nhau gần nhau trong cửa hàng và trên website, giúp khách dễ nhận ra và chọn mua..

2. Chương trình combo giảm giá

Tạo các gói sản phẩm thường được mua cùng nhau, kèm ưu đãi giảm giá để khuyến khích khách hàng mua thêm.

3. Thiết kế quảng cáo theo cặp sản phẩm

Gửi quảng cáo dạng “Khách hàng mua X cũng mua Y”

Cảm ơn!