Big Data Course

Capstone Project   
Final Report

For students (instructor review required)

ⓒ2023 SAMSUNG. All rights reserved.

Samsung Electronics Corporate Citizenship Office holds the copyright of this document.

This document is a literary property protected by copyright law so reprint and reproduction without permission are prohibited.

To use this document other than the curriculum of Samsung Innovation Campus, you must receive written consent from copyright holder.

|  |
| --- |
| Phân tích dữ liệu bán lẻ nhằm tối ưu hóa quản lý tồn kho và gia tăng doanh số bán hàng |

30/07/2025

Team 2

Trương Thái Bảo

Nguyễn Đào Hồng Như

Nguyễn Văn Tú

Nguyễn Lư Hồng Phương

Nguyễn Kim Vương Ý

Content

1. Introduction

1.1. Background Information

1.2. Motivation and Objective

1.3. Members and Role Assignments

1.4. Schedule and Milestones

2. Project Execution

2.1. Simulated Scenario Description

2.2. Datasets Selection and Description

2.3. Data Ingestion Pipeline

2.4. Data Transformation Processing

2.5. Data Query and Insight

3. Results

3.1. Data Ingestion Scripts and Code

3.2. Data Transformation Scripts and Code

3.3. Description and Sample of Transformed Datasets

3.4. Data Visualization of Query Results

4. Projected Impact

4.1. Accomplishments and Benefits

4.2. Future Improvements

5. Team Member Review and Comment

6. Instructor Review and Comment

1. Introduction

1.1. Background Information

Ngành bán lẻ luôn đóng vai trò quan trọng trong nền kinh tế, tạo ra giá trị lớn về doanh thu và việc làm. Trong những năm gần đây, sự cạnh tranh ngày càng gay gắt giữa các doanh nghiệp bán lẻ đã thúc đẩy nhu cầu cải tiến trong quản lý tồn kho và dự báo nhu cầu tiêu dùng. Với sự phát triển mạnh mẽ của công nghệ và dữ liệu lớn (Big Data), các nhà bán lẻ ngày càng nhận thấy tầm quan trọng của việc tận dụng thông tin từ lượng dữ liệu khổng lồ mà họ thu thập được mỗi ngày.

Việc ứng dụng Big Data trong phân tích dữ liệu bán lẻ giúp doanh nghiệp khai thác hiệu quả các thông tin từ lịch sử mua hàng, xu hướng tiêu dùng, dữ liệu tồn kho và hành vi khách hàng. Nhờ đó, doanh nghiệp có thể đưa ra các quyết định quản lý tồn kho một cách chính xác và kịp thời hơn, hạn chế tình trạng hàng tồn dư hoặc thiếu hụt, đồng thời cải thiện trải nghiệm khách hàng và gia tăng doanh số bán hàng.

1.2. Motivation and Objective

Động lực thực hiện đề tài này xuất phát từ việc các doanh nghiệp bán lẻ đang phải đối mặt với áp lực tối ưu hóa hiệu quả hoạt động, trong đó quản lý tồn kho và dự báo nhu cầu khách hàng là hai yếu tố then chốt quyết định đến sự thành công của họ. Mục tiêu của nghiên cứu này là tìm hiểu và áp dụng các kỹ thuật phân tích dữ liệu lớn nhằm nâng cao hiệu quả quản lý tồn kho, tối ưu hóa lượng hàng tồn, đồng thời thúc đẩy doanh số bán hàng thông qua các chiến lược phân tích và dự báo nhu cầu chính xác hơn.

1.3. Members and Role Assignments

* Trương Thái Bảo: Nhóm trưởng, xây dựng mục tiêu dự án, làm sạch dữ liệu, phân tích giỏ hàng, trực quan hóa kết quả
* Nguyễn Đào Hồng Như: Dự báo doanh thu (Prophet), phân tích EDA, trực quan hóa xu hướng, hỗ trợ viết báo cáo
* Nguyễn Văn Tú: Xử lý dữ liệu, viết action plan, hỗ trợ các phần khác.
* Nguyễn Lư Hồng Phương: Làm sạch dữ, tổng hợp kết quả mô hình, phân tích RFM
* Nguyễn Kim Vương Ý: chuẩn hóa, hỗ trợ phân cụm khách hàng và kiểm định mô hình, viết report, tối ưu hóa tồn kho

1.4. Schedule and Milestones

* Ngày 5/7/2025: Xây dựng kế hoạch hành động chi tiết
* Ngày 7/7/2025: Hoàn tất thu thập và hiểu dữ liệu ban đầu
* Ngày 7/7/2025 – 10/7/2025: Làm sạch, xử lí dữ liệu
* Ngày 10/7/2025 – 14/7/2025: EDA: Phân tích xu hướng, sản phẩm bán chạy
* Ngày 15/7/2025: Phân tích giỏ hàng bằng thuật toán Apriori
* Ngày 16/7/2025 - 18/7/2025: Phân cụm khách hàng bằng RFM & K-means
* Ngày 19/7/2025: Tối ưu hóa tồn kho dựa trên forecast
* Ngày 20/7/2025 - 22/7/2025: Dự báo doanh thu với Prophet
* Ngày 22/7/2025: Hoàn thành tất cả phân tích & mô hình hóa
* Ngày 23/7/2025 - 26/7/2025: Viết và hoàn thiện báo cáo
* Ngày 26/7/2025: Chuẩn bị bài trình bày cuối kỳ

2. Project Execution

2.1. Simulated Scenario Description

* + 1. Mô tả tình huống giả lập

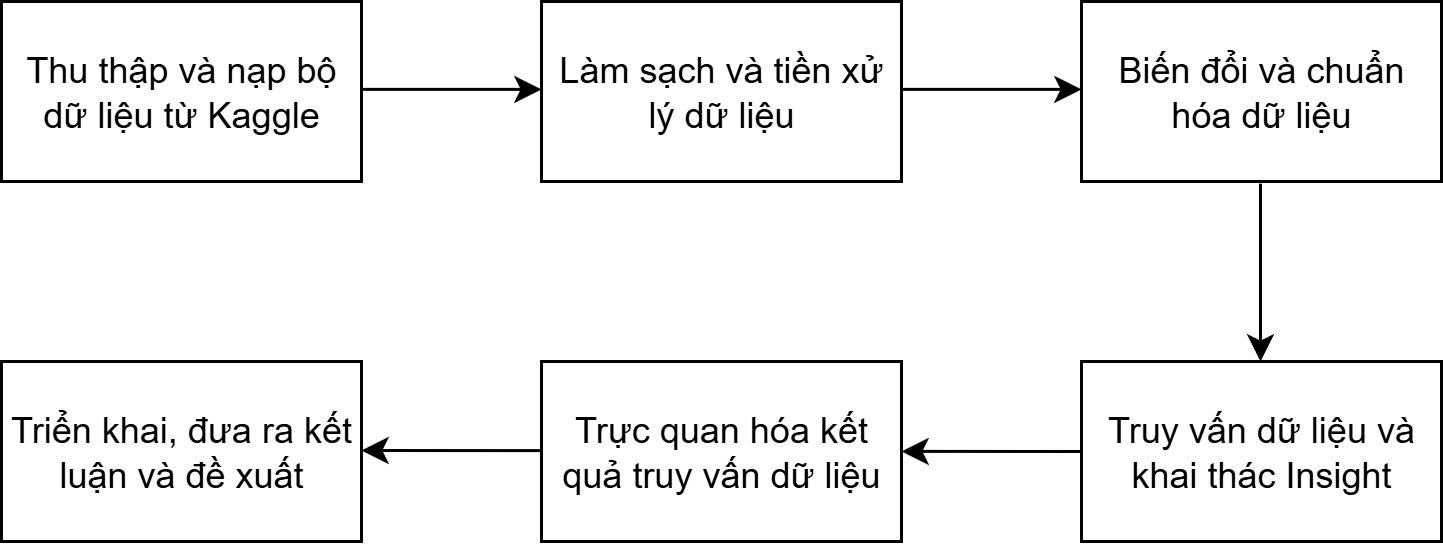
Bộ dữ liệu này mô phỏng môi trường bán lẻ với các giao dịch bán hàng được thực hiện tại nhiều cửa hàng khác nhau, phân bố rộng khắp các khu vực và thành phố, với đa dạng các loại sản phẩm. Mỗi bản ghi đại diện cho một giao dịch cụ thể, bao gồm thông tin quan trọng như ngày giờ thực hiện giao dịch, danh mục sản phẩm, số lượng sản phẩm được bán ra, doanh thu đạt được và địa điểm cửa hàng thực hiện giao dịch. Tình huống giả lập này giúp phản ánh sát với thực tế các hoạt động kinh doanh thường ngày, hỗ trợ phân tích sâu về hành vi mua sắm, nhận diện các xu hướng và mùa vụ bán hàng. Ngoài ra, dữ liệu còn có vai trò quan trọng trong việc dự báo doanh thu tương lai, xác định sản phẩm bán chạy hoặc tồn kho cao. Qua đó, các doanh nghiệp có thể xây dựng các kế hoạch kinh doanh và tiếp thị hiệu quả, tối ưu hóa việc quản lý tồn kho, giảm thiểu chi phí vận hành và tăng cường lợi nhuận thông qua các quyết định chiến lược dựa trên dữ liệu.

2.2. Datasets Selection and Description

Tập dữ liệu được lựa chọn từ nguồn công khai trên nền tảng Kaggle (tác giả: Mohammad Talib), có tên Retail Sales Dataset, mô phỏng lại hoạt động giao dịch thực tế trong một chuỗi bán lẻ toàn cầu. Bộ dữ liệu này phù hợp với các mục tiêu nghiên cứu về phân tích hành vi tiêu dùng, tối ưu tồn kho và dự báo doanh thu nhờ độ đa dạng cao về thuộc tính và lượng dữ liệu đủ lớn. Tập dữ liệu bao gồm hơn 100.000 bản ghi, mỗi dòng đại diện cho một giao dịch đơn lẻ giữa khách hàng và nhà bán lẻ. Các thuộc tính chính bao gồm:

|  |  |  |  |
| --- | --- | --- | --- |
| STT | Tên thuộc tính | Kiểu dữ liệu | Giải thích |
| 1 | Transaction\_ID | int64 | Mã định danh giao dịch |
| 2 | Date | object | Ngày phát sinh giao dịch |
| 3 | Product\_Category, Product\_Brand | object | Loại sản phẩm và thương hiệu |
| 4 | Quantity, Amount | float64 | Số lượng và giá trị sản phẩm trong từng giao dịch |
| 5 | Customer\_ID, Gender, Age, Country, City | int64 | Thông tin định danh và nhân khẩu học của khách hàng |
| 6 | Ratings, Feedback | int64 | Đánh giá định lượng và nhận xét định tính từ khách hàng |
| 7 | Payment\_Method, Shipping\_Method | object | Thông tin về phương thức thanh toán và vận chuyển |
| 8 | **Order\_Status** | object | Trạng thái đơn hàng (hoàn thành, hoàn trả...) |

2.3. Data Ingestion Pipeline



Hình pipeline đề tài

1. Thu thập và nạp bộ dữ liệu từ Kaggle: Bước đầu tiên là tải bộ dữ liệu bán lẻ từ nguồn Kaggle và đưa vào môi trường phân tích (như Python/Pandas). Việc này đảm bảo bạn có dữ liệu đầu vào để bắt đầu quy trình khai phá.
2. Làm sạch và tiền xử lý dữ liệu: Ở bước này, dữ liệu được kiểm tra để phát hiện và xử lý các giá trị thiếu, dữ liệu trùng lặp hoặc không hợp lệ. Đồng thời, các cột không cần thiết có thể được loại bỏ nhằm chuẩn bị dữ liệu cho bước tiếp theo.
3. Biến đổi và chuẩn hóa dữ liệu: Dữ liệu được chuyển đổi về định dạng thống nhất (ví dụ: chuyển đổi ngày tháng, chuẩn hóa định dạng số, mã hóa biến phân loại), tạo điều kiện thuận lợi cho việc truy vấn và phân tích.
4. Truy vấn dữ liệu và khai thác Insight: Thực hiện các thao tác truy vấn để trả lời các câu hỏi phân tích cụ thể, chẳng hạn như: sản phẩm nào bán chạy nhất, thành phố nào có doanh thu cao, thời gian cao điểm bán hàng, ...
5. Trực quan hóa kết quả truy vấn dữ liệu: Sử dụng các biểu đồ, bảng, biểu diễn trực quan để thể hiện kết quả phân tích một cách rõ ràng và dễ hiểu, hỗ trợ việc ra quyết định dựa trên dữ liệu.
6. Triển khai, đưa ra kết luận và đề xuất: Dựa vào các insight thu được, người phân tích đề xuất các giải pháp thực tế như tối ưu tồn kho, điều chỉnh danh mục sản phẩm hoặc lên kế hoạch khuyến mãi, từ đó nâng cao hiệu quả kinh doanh.

2.4. Data Transformation Processing

Trước khi triển khai các phương pháp phân tích và xây dựng mô hình dự đoán, việc làm sạch và xử lý dữ liệu là bước thiết yếu nhằm đảm bảo chất lượng, tính toàn vẹn và độ tin cậy của tập dữ liệu. Quá trình này bao gồm nhiều thao tác có hệ thống để loại bỏ các nhiễu dữ liệu, bổ sung giá trị còn thiếu và chuẩn hóa thông tin. Dưới đây là các bước xử lý cụ thể đã được thực hiện:

* Loại bỏ các thuộc tính không cần thiết hoặc dư thừa

Trong quá trình tiền xử lý, các trường dữ liệu không mang lại giá trị phân tích thực tiễn hoặc trùng lặp thông tin với các thuộc tính khác đã được loại bỏ nhằm giảm nhiễu và tăng hiệu quả xử lý. Cụ thể, các cột như ad\_id, describe, url, zipcode, car\_name, interior\_color, và condition được xem là không cần thiết vì một số lý do như: chứa giá trị rỗng lớn, trùng lặp với thông tin đã có (ví dụ: car\_name là sự kết hợp của brand và grade), hoặc không đóng góp ý nghĩa cho mô hình dự báo (ví dụ: ad\_id chỉ là chỉ số định danh). Việc loại bỏ này giúp giảm chiều dữ liệu và tối ưu hóa hiệu suất tính toán cho các bước phân tích tiếp theo.

* Loại bỏ dữ liệu trùng lặp

Để đảm bảo tính nhất quán, các bản ghi trùng lặp hoàn toàn được loại bỏ thông qua phương thức df\_drop\_duplicates(). Điều này giúp ngăn ngừa việc các mẫu dữ liệu giống hệt nhau ảnh hưởng đến phân phối và kết quả huấn luyện mô hình.

* Xử lý dữ liệu bị thiếu (Missing Values)

Tùy theo tính chất của từng biến, các chiến lược xử lý giá trị bị thiếu được áp dụng như sau:

* Đối với các biến định lượng (Age, Amount, Total\_Purchases, Year): Sử dụng giá trị trung vị (median) để điền vào các ô trống. Median là lựa chọn phù hợp trong trường hợp dữ liệu có phân phối lệch hoặc tồn tại ngoại lệ (outliers).
* Tính toán giá trị mới: Cột Total\_Amount được tính lại bằng tích của Total\_Purchases và Amount trong trường hợp cả hai đều không bị thiếu. Điều này giúp chuẩn hóa thông tin và tránh tình trạng dữ liệu không đồng bộ.
* Xử lý biến phân loại liên quan

Hai biến phân loại Ratings và Feedback được xác định là có quan hệ ánh xạ một-một (1-1). Do đó, nếu một trong hai biến bị thiếu, giá trị tương ứng có thể được nội suy dựa trên biến còn lại:

* Thiết lập ánh xạ giữa Feedback và Ratings (ví dụ: "Excellent" ↔ 5, "Good" ↔ 4, ...).
* Áp dụng nội suy hai chiều: điền Ratings dựa trên Feedback và ngược lại.
* Các bản ghi bị thiếu cả hai trường thông tin nêu trên (chiếm tỷ lệ rất nhỏ) sẽ bị loại bỏ để đảm bảo chất lượng tập dữ liệu.
* Chuẩn hóa dữ liệu địa lý

Đối với các bản ghi chỉ chứa tên thành phố (City) mà thiếu tên quốc gia (Country), tiến hành xây dựng từ điển ánh xạ tên thành phố → quốc gia dựa trên danh sách chuẩn hóa theo từng quốc gia lớn (Mỹ, Canada, Đức, Úc, Anh). Sau đó, gán ngược các bản ghi còn thiếu quốc gia dựa trên thông tin thành phố có sẵn.

Ngoài ra, các giá trị City bị thiếu được điền giá trị 'Unknown' để giữ lại bản ghi nhưng vẫn đảm bảo tính nhất quán trong mô hình hóa.

* Chuẩn hóa dữ liệu thời gian

Cột Date được chuyển đổi từ định dạng mm/dd/yy sang dd/mm/yy nhằm đảm bảo đồng nhất định dạng ngày tháng. Đồng thời, tạo thêm biến Month để phục vụ các phân tích theo mùa vụ và chu kỳ.

* Điền dữ liệu thiếu cho các cột còn lại

Các cột còn lại như Income, Gender, Customer\_Segment, Payment\_Method,Shipping\_Method, Product\_Category, Product\_Brand, v.v. được xử lý bằng cách điền giá trị 'Unknown'. Đây là phương án an toàn nhằm giữ lại tối đa dữ liệu, đồng thời giúp mô hình xử lý tốt hơn các giá trị hiếm hoặc bị thiếu

2.5. Data Query and Insight

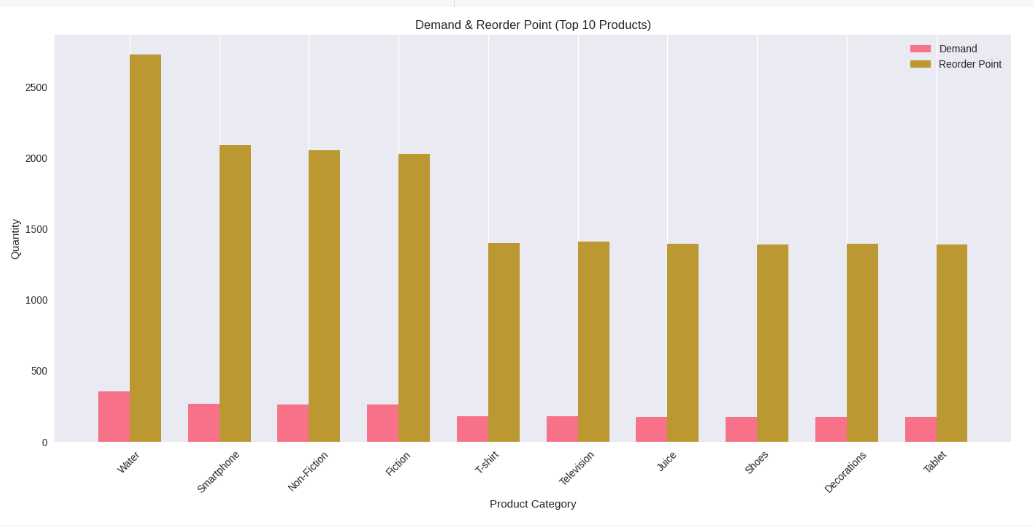
* + 1. Mục tiêu truy vấn

Mục tiêu chính của quá trình truy vấn dữ liệu (data query) là nhằm:

* Hiểu rõ đặc điểm thống kê mô tả của tập dữ liệu bán lẻ, bao gồm khách hàng, sản phẩm, giao dịch và hành vi mua sắm.
* Xác định các xu hướng theo thời gian, như doanh thu tăng giảm theo tháng, mùa vụ, ngày trong tuần.
* Phát hiện các mối quan hệ quan trọng giữa các yếu tố định tính: phương thức thanh toán, vận chuyển, sản phẩm, địa lý.
* Làm sạch và định dạng dữ liệu phù hợp, nhằm phục vụ tốt cho giai đoạn mô hình hóa sau này như phân cụm (clustering), phân tích giỏ hàng (basket analysis), dự báo doanh thu (forecasting), và tối ưu tồn kho (inventory optimization).
  + 1. Công cụ và phương pháp

Các thao tác truy vấn và phân tích dữ liệu được thực hiện thông qua ngôn ngữ lập trình Python, cụ thể sử dụng các thư viện:

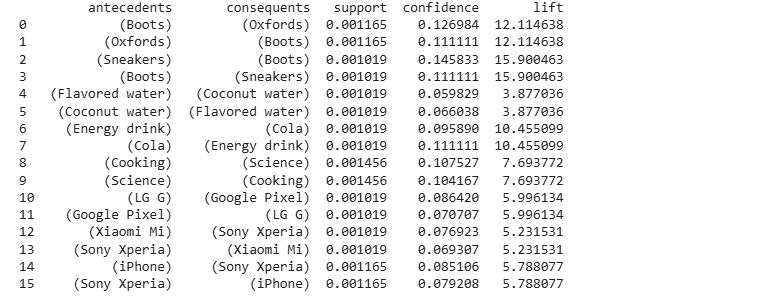
* **Pandas**: xử lý bảng dữ liệu, truy vấn .groupby(), .pivot\_table(), .value\_counts(), tính toán thống kê mô tả như mean, std, min, max.
* **Datetime**: xử lý thời gian từ biến Date để tạo thêm đặc trưng thời gian như Month, Quarter, Weekday, Hour.
* **Matplotlib & Seaborn**: trực quan hóa dữ liệu với biểu đồ line, bar, heatmap, histogram, boxplot để giúp nhận diện xu hướng dễ dàng hơn.
  + 1. Một vài các insight khai thác được
* Các danh mục sản phẩm có nhu cầu mua cao nhất



Biểu đồ top 10 sản phẩm có nhu cầu cao nhất và order lại

Insight: các sản phẩm từ nước có khả năng mua lại cao nhất, tiếp đến là điện thoại, …

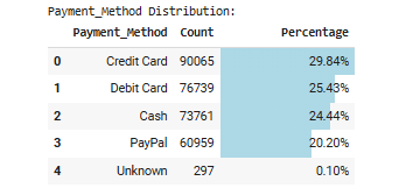
* Biết được nhu cầu của khách hàng, từ đó nhập thêm các mặt hàng có nhu cầu cao và nhập ít lại các mặt hàng chậm bán hơn.
* Những sản phẩm được mua kèm nhau



Hình các thông số trong thuật toán Apriori

Insight: Các sản phẩm giày (**Boots, Oxfords, Sneakers**) có mối liên hệ mua kèm cực kỳ mạnh, sản phẩm công nghệ thường mua kèm nhau.

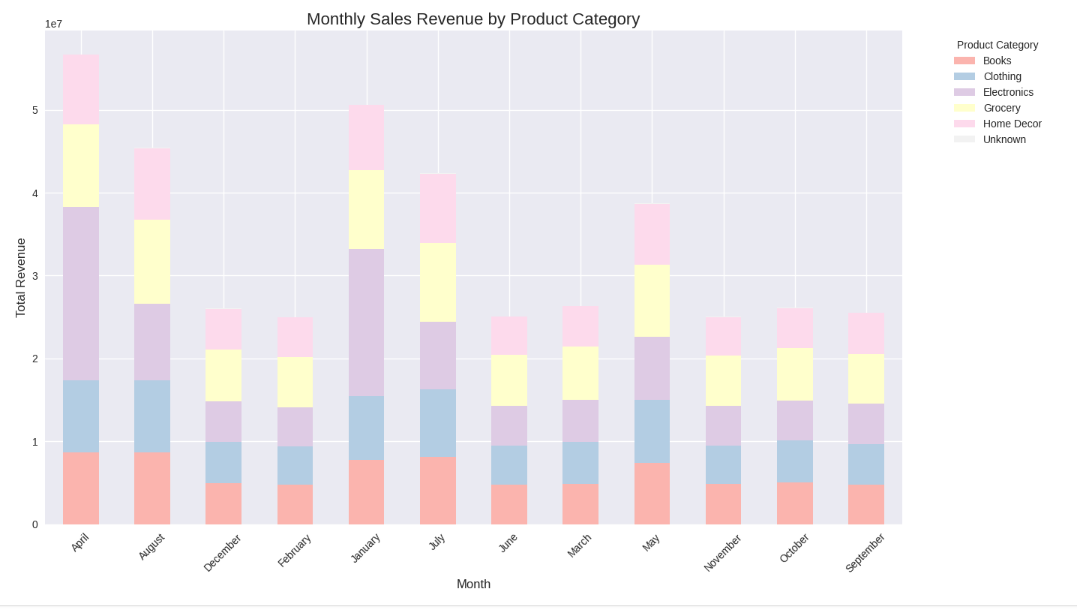
* Đặt các sản phẩm gần nhau để khách hàng có thể dễ dàng mua sắm từ đó tăng doanh thu cho cửa hàng.
* Phương thức thanh toán



Hình sự phân tán của biến Phương thức thanh toán

Insight: Credit Card là phổ biến nhất, Cash gắn với vận chuyển chậm.

* Danh mục sản phẩm bán chạy theo tháng



Biểu đồ doanh thu theo tháng tính theo danh mục sản phẩm

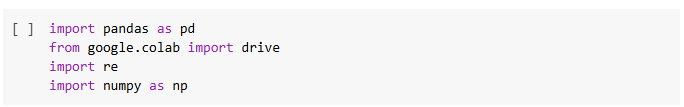
Insight: Xu hướng doanh thu cao từ tháng 4 đến tháng 8, đặc biệt là các mặt hàng điện tử.

* Tung ra các chương trình khuyến mãi ở các tháng còn lại để thúc đẩy doanh thu, từ đó sinh ra lợi nhuận đều cho cửa hàng.

*Kết quả từ các truy vấn trên giúp nhóm nghiên cứu đưa ra những kết luận quan trọng về hành vi người tiêu dùng, xác định đâu là những yếu tố ảnh hưởng đến hiệu quả kinh doanh và đề xuất các hành động cụ thể cho việc tối ưu hóa quản lý tồn kho, cải thiện doanh số và nâng cao trải nghiệm khách hàng. Đây là nền tảng quan trọng trước khi bước vào giai đoạn xây dựng mô hình phân tích dự đoán.*

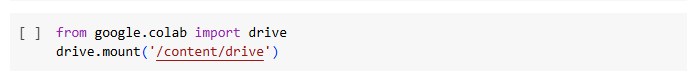
3. Results

3.1. Data Ingestion Scripts and Code



Hình code import thư viện

Nhóm sử dụng thư viện pandas để hỗ trợ xử lý dữ liệu dạng bảng, đồng thời tích hợp thư viện numpy và re để hỗ trợ thao tác với mảng số học và biểu thức chính quy. Ngoài ra, nhóm dùng module google.colab.drive để kết nối với Google Drive – nơi lưu trữ dữ liệu.



Hình code mount đến google drive

Mục đích của đoạn mã trên là để “mount” (kết nối) Google Drive vào môi trường làm việc của Google Colab. Sau khi mount thành công, toàn bộ dữ liệu trong Drive sẽ được ánh xạ vào thư mục /content/drive, cho phép nhóm có thể truy cập các file dữ liệu như file CSV, Excel, mô hình đã huấn luyện,… giống như đọc dữ liệu từ ổ đĩa cục bộ.

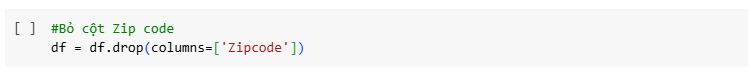


Hình code đọc file csv

Nhóm sử dụng hàm pd.read\_csv() từ thư viện pandas để đọc dữ liệu từ tệp CSV có tên new\_retail\_data.csv được lưu trữ trong Google Drive. Kết quả là dữ liệu được đọc vào biến df dưới dạng một đối tượng DataFrame, cho phép thực hiện các thao tác phân tích dữ liệu như thống kê, lọc, xử lý thiếu, vẽ biểu đồ,... Việc sử dụng Google Drive giúp nhóm dễ dàng quản lý dữ liệu, đồng thời đảm bảo khả năng truy cập linh hoạt trên nền tảng đám mây.

3.2. Data Transformation Scripts and Code

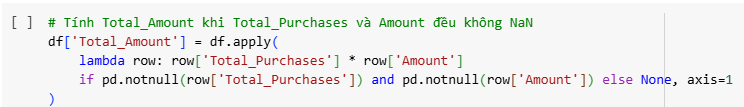
***Loại bỏ biến không cần thiết***



Hình code bỏ cột Zipcode

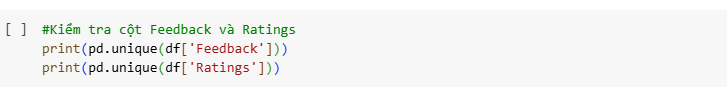
Cột Zipcode không mang tính phân tích rõ ràng và có tính phân mảnh cao, do đó được loại bỏ để giảm nhiễu trong dữ liệu và đơn giản hóa cấu trúc bảng. 

Hình code điền giá trị NaN bằng median

Đối với các cột dạng số, nhóm lựa chọn phương pháp **thay thế giá trị thiếu bằng trung vị (median imputation)**. Phương pháp này giữ được tính đại diện của dữ liệu mà không bị ảnh hưởng bởi các giá trị ngoại lệ. 

Hình code tính cột Total\_Amount

Tạo biến Total\_Amount bằng cách nhân Total\_Purchases với Amount. Biến này giúp đo lường **tổng doanh thu của mỗi giao dịch** – thông tin cốt lõi phục vụ dự báo và phân tích giá trị khách hàng.



Hình code kiểm tra cột Feedback và Ratings

Sử dụng pd.unique() để kiểm tra các giá trị duy nhất trong hai biến định tính Feedback và Ratings. Bước này giúp xác định xem có cần mã hóa lại các giá trị phân loại (ví dụ: Positive → 1, Negative → 0) hay không.



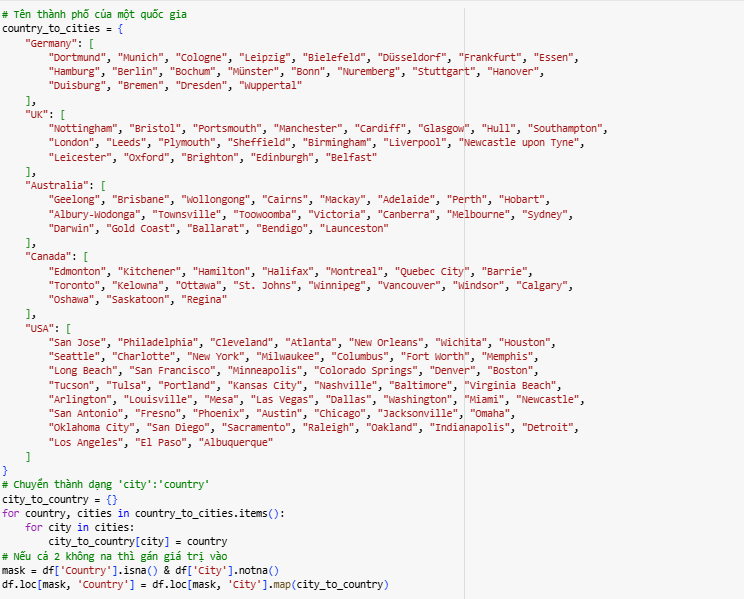
Hình code kiểm tra giá trị duy nhất của biến Feedback và Ratings

Tự động điền dữ liệu còn thiếu giữa hai cột liên quan nhằm đảm bảo đầy đủ và nhất quán thông tin đánh giá.



Hình code kiểm tra số lượng dòng NaN ở Ratings và Feedback

Đây là tỉ lệ rất thấp nên ta có thể xử lí bằng cách bỏ các dòng này



Hình code điền giá trị NaN của Country dựa vào City

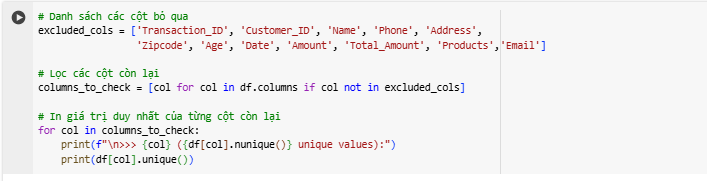


Hình code kiểm tra các giá trị NaN còn lại



Hình code điền giá trị NaN bằng Unknown

Kiểm tra xem còn cột nào chưa xử lý



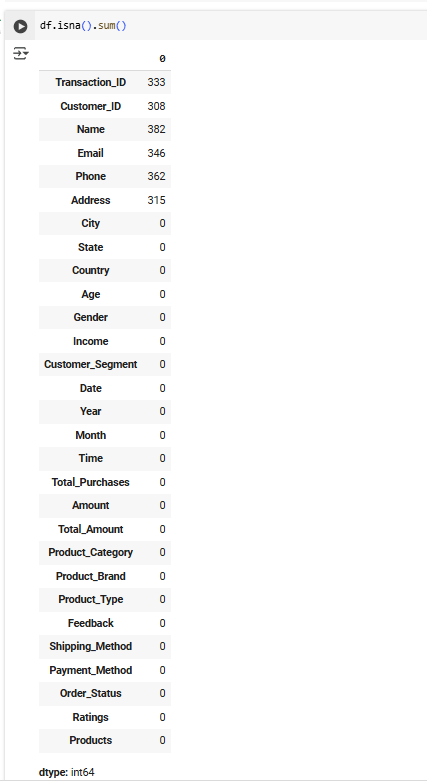
Hình code kiểm tra các cột chưa được xử lý



Hình code chuyển kiểu dữ liệu datetime và điền NaN cột Month dựa vào Date

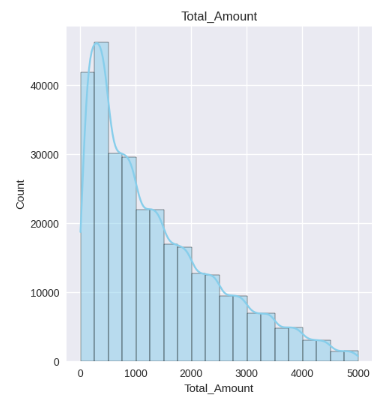


Hình code điền các cột có giá trị NaN còn lại là Unknown

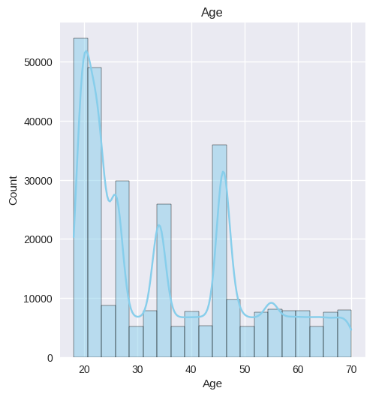


Hình code kiểm tra các giá trị NaN

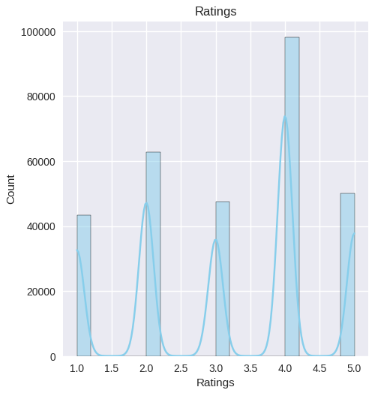
3.3. Description and Sample of Transformed Datasets



Hình biểu đồ phân bố cột Total\_Amount



Hình biểu đồ phân khúc độ tuổi khách hàng (Age)



Hình biểu đồ thang đo hài lòng khách hàng (Ratings)

3.4. Data Visualization of Query Results

***Phân tích thống kê các biến định lượng***

Dựa vào biểu đồ phân bố cột chi tiêu khách hàng(Total\_Amount) cho thấy: Dữ liệu phản ánh một thực tế phổ biến trong ngành bán lẻ đa số khách hàng có xu hướng thực hiện các giao dịch có giá trị nhỏ, trong khi chỉ một nhóm nhỏ đóng góp giá trị lớn cho doanh thu. Chiến lược phát triển:

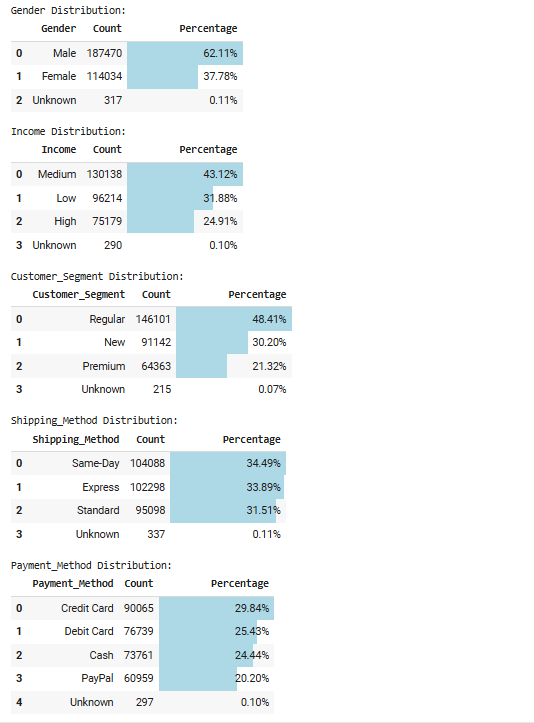
* Với nhóm chi tiêu thấp: Triển khai các chương trình "mua thêm để được ưu đãi" hoặc gói combo sản phẩm để tăng giá trị đơn hàng trung bình
* Với nhóm chi tiêu cao: Xây dựng chương trình khách hàng thân thiết với các đặc quyền độc tôn như tư vấn cá nhân, giao hàng ưu tiên, hoặc sự kiện riêng

Dựa vào biểu đồ phân khúc độ tuổi khách hàng(Age): Hai nhóm tuổi nổi bật 18-25 và 45-50 phản ánh hai giai đoạn tiêu dùng quan trọng: thế hệ Gen Z bắt đầu có thu nhập ổn định và thế hệ X đang ở đỉnh cao sự nghiệp. Sự tập trung tại các mốc tuổi tròn có thể xuất phát từ đặc điểm tâm lý người tiêu dùng - họ thường có xu hướng đăng ký và mua sắm nhiều hơn vào các cột mốc tuổi đặc biệt như 30, 40, 50. Chiến lược tiếp cận mục tiêu:

* Nhóm 18-25: Tập trung vào mạng xã hội, nội dung thú vị, giá cả hợp lý và xu hướng
* Nhóm 45-50: Ưu tiên chất lượng, tiện lợi, dịch vụ chuyên nghiệp và giá trị lâu dài

Dựa vào biểu đồ thang đo hài lòng khách hàng (Ratings): Mức điểm đánh giá tập trung ở 4.0 cho thấy khách hàng nhìn chung cảm thấy hài lòng, nhưng chưa thực sự ấn tượng với trải nghiệm. Điều đáng lo ngại hơn là sự hiện diện của một lượng đánh giá 1.0-2.0 vẫn ở mức đáng quan ngại. Việc xác định và khắc phục những điểm yếu gây ra các đánh giá tiêu cực này là ưu tiên hàng đầu, nhằm duy trì tăng trưởng doanh thu bền vững.

* Thực hiện khảo sát chi tiết với nhóm đánh giá 1-2 điểm để xác định vấn đề cụ thể
* Phát triển quy trình phản hồi nhanh cho khiếu nại và đề xuất của khách hàng

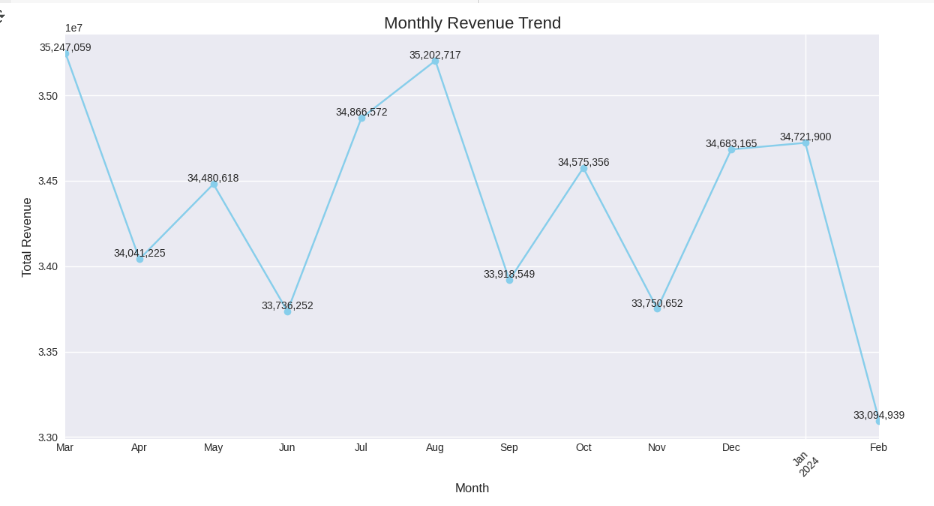
***Phân tích thống kê các biến định tính***

Hình thống kê các biến định tính

Nhận xét:

* Phân bố thị trường địa lý hiện diện mạnh mẽ tại các thị trường các nước phát triển với Mỹ dẫn đầu, tiếp theo là Anh, Đức, Úc và Canada. Điều đáng chú ý là khách hàng không chỉ tập trung ở các đô thị lớn mà phân bố đều khắp nhiều khu vực, với Chicago và Portsmouth nổi bật về lượng giao dịch cao.
* Điện tử và Hàng tạp hóa là hai nhóm sản phẩm chiếm ưu thế, tiếp theo là Quần áo, Sách và Trang trí nhà với tỷ lệ tương đối đều nhau. Về giới tính, khách hàng nam chiếm tỷ lệ áp đảo với hơn 60% so với nữ giới, cho thấy tiềm năng mở rộng đáng kể nếu có chiến lược tiếp cận phù hợp với khách hàng nữ.
* Nhóm thu nhập trung bình chiếm đa số so với nhóm thu nhập thấp và cao. Về độ trung thành, khách hàng thường xuyên chiếm khoảng một nửa tổng số, phản ánh khả năng giữ chân tốt của doanh nghiệp. Tuy nhiên, tỷ lệ khách hàng mới khá cao cũng mở ra cơ hội để tăng độ trung thành thông qua các chương trình chăm sóc phù hợp.
* Xu hướng tiêu dùng và thanh toán khách hàng thể hiện sự ưu tiên rõ ràng đối với tốc độ giao hàng. Về thanh toán, thẻ tín dụng dẫn đầu, tiếp theo là thẻ ghi nợ, tiền mặt và PayPal, cho thấy sự đa dạng trong thói quen thanh toán của khách hàng.
* Nên tập trung vào việc mở rộng thị trường nữ giới thông qua việc điều chỉnh danh mục sản phẩm và chiến dịch marketing phù hợp. Đồng thời, cần duy trì các chương trình ưu đãi cho khách hàng thường xuyên và cải thiện trải nghiệm dịch vụ để chuyển đổi khách mới thành khách quen.
* Việc đa dạng hóa phương thức thanh toán, đặc biệt bổ sung các tùy chọn hiện đại như ví điện tử, sẽ tăng tính linh hoạt và thuận tiện cho khách hàng. Bên cạnh đó, cần tiếp tục duy trì ưu thế về tốc độ giao hàng để đáp ứng nhu cầu ngày càng cao của thị trường.

***Xu hướng doanh thu theo thời gian***

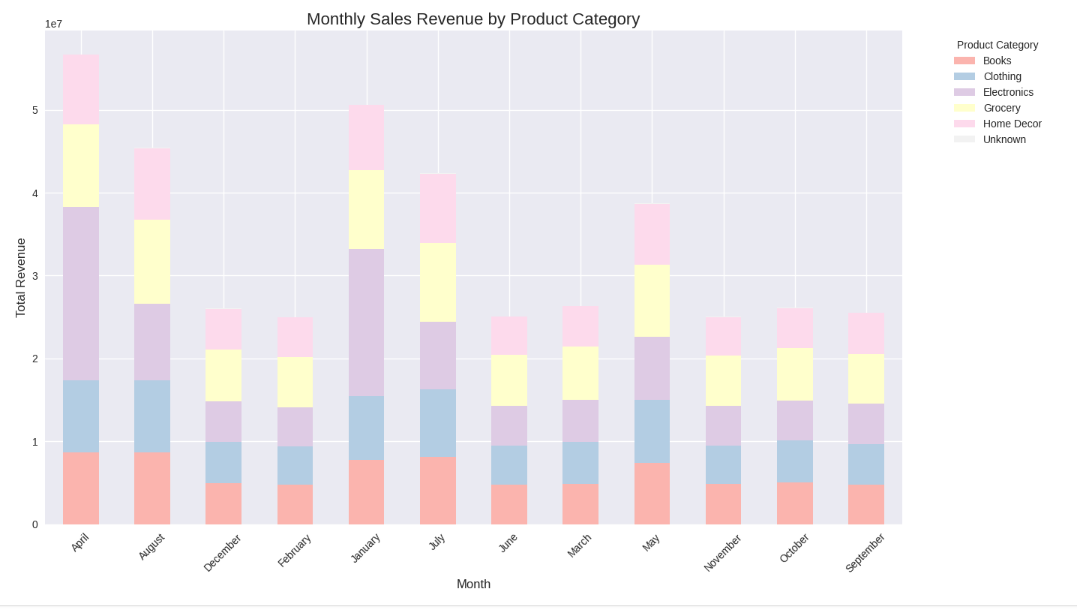


Hình xu hướng doanh thu theo tháng

Dựa vào biểu đồ trên cho thấy:

* Doanh thu năm 2024 thể hiện mô hình biến động chu kỳ với hai đỉnh cao nổi bật vào tháng 3 (35.23 triệu) và tháng 8 (35.20 triệu). Các tháng còn lại duy trì ổn định trong khoảng 34-34.7 triệu, ngoại trừ ba thời điểm trũng là tháng 2, 6 và 11.
* Những điểm bất thường đáng chú ý tháng 3 ghi nhận doanh thu cao nhất trong năm, đây là hiện tượng bất thường so với quy luật kinh doanh truyền thống khi đỉnh cao thường xuất hiện vào giữa hoặc cuối năm. Điều này cho thấy có thể đã xuất hiện những yếu tố đặc biệt tác động tích cực lên hoạt động kinh doanh trong giai đoạn này.
* Một điểm đáng lưu ý khác là sự sụt giảm mạnh gần 1.3 triệu từ tháng 8 sang tháng 9, phản ánh hiện tượng "hụt hơi" sau đợt cao điểm. Tương tự, sau xu hướng phục hồi cuối năm 2023, doanh thu lại giảm sâu vào tháng 2/2024, cho thấy sự thiếu liên tục trong động lực tăng trưởng.
* Phân tích theo từng giai đoạn cho thấy sự biến động đáng chú ý giữa các quý.
* Quý I ghi nhận sự bứt phá ấn tượng vào tháng 3, vượt kỳ vọng và trở thành tháng có doanh thu cao nhất trong năm. Sang quý II, doanh thu có nhiều biến động, đặc biệt là sự sụt giảm rõ rệt từ tháng 5 sang tháng 6.
* Quý III tiếp tục duy trì xu hướng tăng trưởng theo "quy luật mùa vụ", với tháng 8 đạt mức doanh thu cao, phù hợp với các chu kỳ tiêu dùng thường thấy. Trong khi đó, quý IV tương đối ổn định, nhưng không giữ được đà tăng trưởng khi bước sang đầu năm tiếp theo.
* Khuyến nghị chiến lược ưu tiên hàng đầu là phân tích sâu các yếu tố thúc đẩy thành công của tháng 3 để có thể tái tạo mô hình này cho các tháng khác. Đồng thời, cần xây dựng các chương trình hỗ trợ đặc biệt cho những tháng yếu như tháng 2, 6 và 11 thông qua khuyến mãi, tăng cường quảng bá hoặc ưu đãi mua sớm.

***Sản phẩm bán chạy***

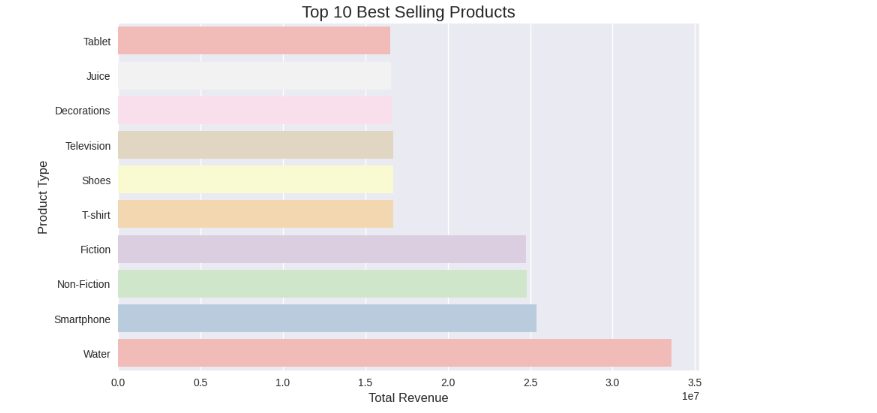


Biểu đồ doanh thu theo tháng tính theo danh mục sản phẩm

Nhận xét:

* Doanh thu tháng 4 đạt mức cao nhất trong năm, vượt mốc 54 triệu, cho thấy đây là thời điểm có nhu cầu mua sắm rất mạnh, đặc biệt đến từ nhóm Electronics và Grocery. Tháng 1 và tháng 7 cũng ghi nhận mức doanh thu cao, lần lượt khoảng 50 triệu và 43 triệu, cho thấy hoạt động mua sắm khởi đầu năm và giữa năm là rất sôi động.
* Trong khi đó, các tháng 2, 6 và 12 có doanh thu thấp hơn đáng kể (chỉ khoảng 25–27 triệu), cho thấy có thể đây là giai đoạn trầm trong chu kỳ tiêu dùng, đặc biệt vào sau Tết hoặc trước các dịp giảm giá lớn. Việc doanh thu không tăng trong tháng 12 cũng là một dấu hiệu đáng chú ý – cần phân tích thêm xem có phải do chính sách giá chưa phù hợp hay khách hàng chuyển hướng sang kênh mua sắm khác.
* Electronics là nhóm hàng đóng góp lớn nhất trong hầu hết các tháng, đặc biệt nổi bật trong tháng 1, 4 và 7. Đây là mặt hàng có biên độ doanh thu lớn và mang tính chu kỳ rõ rệt, thường tăng vào đầu năm và giữa năm – phù hợp với xu hướng mua sắm thiết bị công nghệ đầu kỳ hoặc trong các đợt khuyến mãi hè.
* Grocery và Home Decor duy trì doanh thu ổn định theo tháng, đặc biệt tăng vào các tháng 3–5, có thể liên quan đến mùa cưới, lễ hội hoặc chuẩn bị nhà cửa cho dịp lễ.
* Nhóm Books và Clothing có doanh thu thấp hơn nhưng phân bổ khá đều trong năm. Điều này phản ánh đặc thù các ngành hàng này có nhu cầu ổn định, không chịu ảnh hưởng mạnh bởi mùa vụ.
* Nhóm Unknown xuất hiện với tỷ trọng nhỏ và rải rác trong một số tháng. Việc tồn tại danh mục chưa rõ nguồn gốc cho thấy hệ thống phân loại sản phẩm cần được cập nhật hoặc chuẩn hóa lại dữ liệu danh mục.

Phân tích xu hướng:

* Doanh thu có tính mùa vụ cao với 3 đỉnh lớn rơi vào các tháng 1, 4 và 7. Điều này có thể gắn liền với các sự kiện như đầu năm mới, lễ hội mùa xuân và chiến dịch mua sắm giữa năm.
* Tỷ trọng giữa các nhóm sản phẩm ít biến động theo thời gian, chứng tỏ hành vi tiêu dùng vẫn tương đối ổn định giữa các danh mục chính. Tuy nhiên, để tối ưu doanh thu, có thể thúc đẩy cross-sell giữa Electronics – Home Decor – Grocery vào các tháng cao điểm.
* Tháng 6, 9 và 12 có doanh thu thấp hơn trung bình, phù hợp để thử nghiệm các chương trình kích cầu như giảm giá ngắn hạn, tặng kèm hoặc gói combo nhằm tăng mức chi tiêu trung bình.

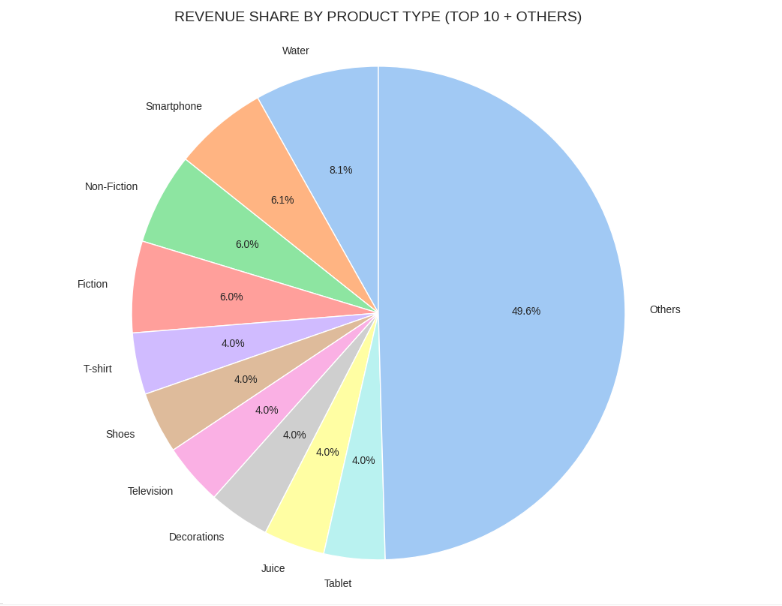
Biểu đồ top 10 sản phẩm bán chạy nhất

Dựa vào biểu đồ ta thấy:

* Biểu đồ Top 10 Best Selling Products cho thấy có sự phân tầng rõ rệt về tổng doanh thu giữa các nhóm sản phẩm, với “Water” dẫn đầu vượt trội (khoảng 33.5 triệu), vượt xa các sản phẩm còn lại. Đây là mặt hàng tiêu dùng thiết yếu, có tần suất mua lặp lại cao, phù hợp với thói quen tiêu dùng hằng ngày của phần lớn người dân.
* Nhóm sản phẩm công nghệ như Smartphone và Tablet cũng ghi nhận doanh thu rất cao (Smartphone ~25 triệu, Tablet ~17 triệu). Điều này phản ánh rõ xu hướng số hóa và mức độ quan tâm cao của người tiêu dùng đối với các sản phẩm điện tử cá nhân. Nhóm này thường có đóng góp lớn trong các giai đoạn mua sắm cao điểm như lễ hội, năm học mới hoặc các dịp khuyến mãi trực tuyến.
* Đáng chú ý là cả hai dòng sách “Fiction” và “Non-Fiction” đều nằm trong top 10, với doanh thu xấp xỉ 24–25 triệu. Đây là dấu hiệu cho thấy nhu cầu học tập, giải trí và phát triển bản thân vẫn được duy trì ổn định, bất chấp biến động thị trường. Khả năng xuất hiện đồng thời cả hai dòng sách cho thấy doanh nghiệp đã thành công trong việc đa dạng hóa nội dung và tiếp cận đa tệp khách hàng.
* T-shirt, Shoes và Television là nhóm sản phẩm nằm ở vị trí giữa bảng, với doanh thu dao động quanh mốc 17 triệu. Các sản phẩm này thường chịu ảnh hưởng bởi yếu tố mùa vụ, thị hiếu và xu hướng thời trang theo quý. Tuy chưa nổi bật về doanh thu, nhóm này giữ vai trò quan trọng trong việc tăng số lượng đơn hàng, đặc biệt khi kết hợp với các chương trình combo hoặc giảm giá định kỳ.
* Decorations và Juice là hai sản phẩm có doanh thu thấp nhất trong top nhưng vẫn đóng vai trò bổ trợ tốt trong các chiến lược cross-sell. Với đặc điểm tiêu dùng gắn với các sự kiện (trang trí, đồ uống trong dịp lễ), đây là những mặt hàng có thể tận dụng hiệu quả trong các gói khuyến mãi theo mùa để gia tăng giá trị giỏ hàng trung bình (AOV).

Định hướng:

* Water giữ vai trò là sản phẩm trụ cột trong danh mục bán lẻ, cần tiếp tục đầu tư tối ưu tồn kho và logistics để đảm bảo đáp ứng nhu cầu cao quanh năm.
* Sản phẩm công nghệ và sách nên được khai thác mạnh trong các chiến dịch marketing theo nhóm khách hàng mục tiêu (ví dụ: giới trẻ, học sinh – sinh viên, nhân viên văn phòng).
* Nhóm trang trí, đồ uống, thời trang cần được tích hợp thông minh trong các chương trình ưu đãi có điều kiện (giảm giá khi mua kèm, quà tặng theo hóa đơn) để tăng hiệu quả doanh thu toàn hệ thống.



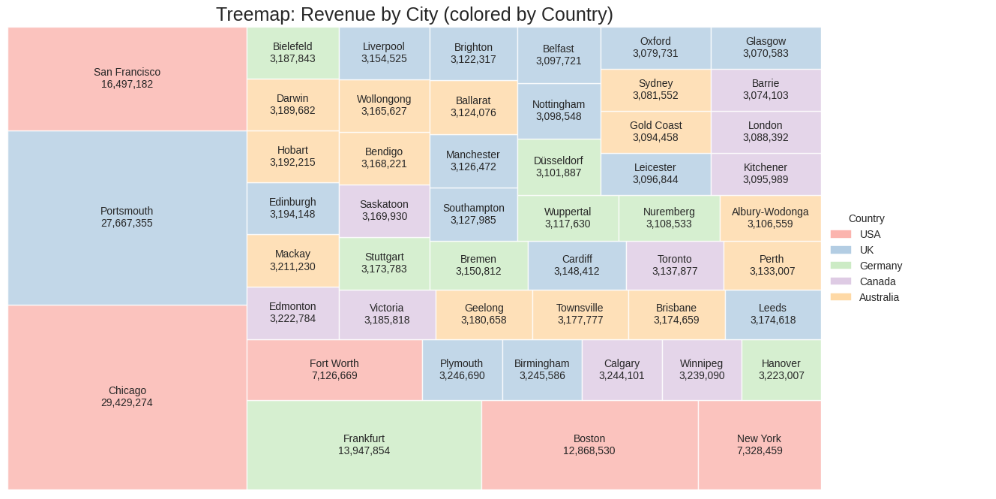
Biểu đồ top 10 loại sản phẩm theo doanh thu

Nhận xét:

* Doanh thu theo danh mục sản phẩm cho thấy tháng 3 đạt đỉnh, chủ yếu nhờ sự đóng góp mạnh mẽ từ hai nhóm sản phẩm chính là Electronics và Grocery. Đây là hai nhóm hàng có tính thiết yếu cao và thường xuyên được tiêu dùng, đặc biệt trong giai đoạn đầu năm thời điểm người tiêu dùng có xu hướng mua sắm để khởi đầu mới hoặc nâng cấp thiết bị, vật dụng cá nhân. Điều này mở ra cơ hội lớn cho các chiến dịch marketing và khuyến mãi tập trung vào quý I, tận dụng nhu cầu tự nhiên của thị trường.
* Trong khi đó, các danh mục như Books, Clothing và Home Decor duy trì doanh thu ổn định quanh năm, nhưng tỷ trọng đóng góp vẫn khiêm tốn hơn. Mặc dù không tạo đột phá, những nhóm sản phẩm này lại đóng vai trò duy trì dòng doanh thu ổn định và phục vụ các phân khúc khách hàng có thị hiếu đa dạng hơn.
* Phân tích top 10 sản phẩm bán chạy cho thấy "Water" và "Smartphone" là hai sản phẩm nổi bật, dẫn đầu về doanh thu. Điều này phản ánh một mặt là nhu cầu tiêu dùng thiết yếu (nước uống), mặt khác là xu hướng đầu tư vào sản phẩm công nghệ cá nhân. Nhóm sản phẩm như Non-Fiction, Shoes và Fiction có mức doanh thu khá đều, cho thấy sự quan tâm đến lối sống, thời trang và sở thích đọc sách
* Đáng chú ý, nhóm sản phẩm ngoài top 10 (Others) chiếm tới 49,6% tổng doanh thu, điều này cho thấy doanh nghiệp đang có rất nhiều sản phẩm tiềm năng chưa được khai thác hiệu quả về mặt truyền thông và chiến lược thương mại. Việc không có sản phẩm vượt trội ngoài một vài cái tên như Smartphone cũng cho thấy thị trường hiện tại đang phân tán và mở ra cơ hội hơn.

Khuyến nghi chiến lược:

* Ưu tiên danh mục: Tăng cường đầu tư vào Electronics, Grocery và các sản phẩm công nghệ tiềm năng những nhóm đã chứng minh khả năng tạo doanh thu vượt trội.
* Đa dạng hóa sản phẩm: Tiến hành đánh giá định kỳ hiệu suất sản phẩm, cải tiến hoặc loại bỏ các mặt hàng có doanh thu thấp, đồng thời xem xét các sản phẩm nằm trong nhóm "Others" để tìm kiếm các cơ hội tăng trưởng ẩn.
* Tối ưu thời điểm bán hàng: Đẩy mạnh hoạt động kinh doanh trong quý I, đặc biệt là tháng 3, kết hợp với các chương trình khuyến mãi theo mùa trong những giai đoạn còn lại để duy trì doanh thu ổn định và tránh sự phụ thuộc vào thời điểm cao điểm duy nhất.

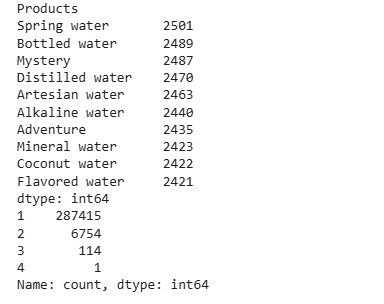


Biểu đồ doanh thu theo thành phố

Nhận xét:

* Phân tích dữ liệu bán lẻ cho thấy doanh thu phân bổ không đồng đều giữa các thành phố trên toàn cầu, với sự chênh lệch rõ rệt về mức độ đóng góp giữa các khu vực địa lý.
* Các thành phố thuộc Hoa Kỳ chiếm ưu thế vượt trội về doanh thu, trong khi các thành phố tại châu Âu và Úc duy trì mức doanh thu ổn định nhưng thiếu sự nổi bật. Điều này phản ánh sự khác biệt về tiềm năng thị trường và hiệu quả hoạt động kinh doanh giữa các khu vực.
* Việc tập trung doanh thu tại các thành phố trọng điểm, đồng thời duy trì sự hiện diện rộng rãi ở nhiều thành phố quy mô nhỏ, cho thấy chiến lược phân bổ thị trường tương đối hiệu quả và giúp giảm thiểu rủi ro phụ thuộc vào một khu vực duy nhất. Tuy nhiên, khoảng cách đáng kể giữa nhóm thành phố dẫn đầu và phần còn lại cũng cho thấy tiềm năng khai thác vẫn chưa được tận dụng triệt để. Đáng chú ý, sự hiện diện còn mờ nhạt tại các thủ đô lớn như London hay Berlin có thể là một cơ hội thị trường đang bị bỏ lỡ.
* Mặc dù một số thành phố đang tạo ra mức doanh thu vượt trội, phần lớn các địa phương còn lại có doanh thu ở mức ổn định. Điều này mở ra cơ hội rõ ràng cho doanh nghiệp đầu tư thêm nhằm khai thác hiệu quả hơn tại những thị trường chưa phát huy hết tiềm năng.

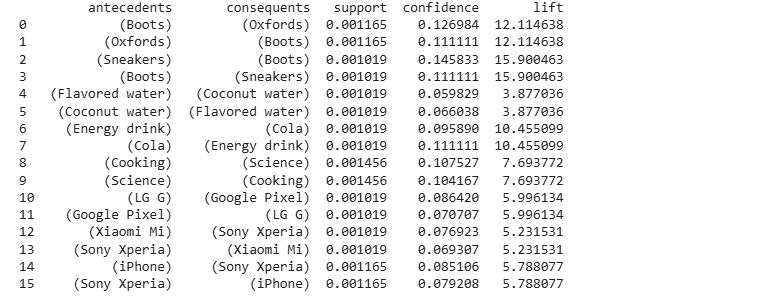
***Phân tích giỏ hàng***



Hình top 10 sản phẩm được mua nhiều nhất

Nhận xét:

* Có 294284 giỏ hàng và 318 sản phẩm
* Nhưng:
  + 287415 giỏ chỉ có 1 sản phẩm
  + 6754 giỏ có 2 sản phẩm
  + 114 giỏ có 3 sản phẩm
  + 1 giỏ có 4 sản phẩm
* Có đến 287415 giỏ chỉ có 1 sản phẩm chiếm đến gần 98% nên ta chỉ lấy các giỏ từ 2 sản phẩm trở lên
* Các sản phẩm phổ biến nhất chỉ xuất hiện khoảng 2,400 – 2,500 lần (~0.86% tổng giao dịch).
* Nếu ban đầu chọn min\_support=0.01 (1%) là ~2,900 lần. Nhưng sản phẩm nhiều nhất chỉ xuất hiện ~2,500 lần, mà giỏ hàng có ≥2 sản phẩm chỉ là 6,800 giỏ.
* Đa số giao dịch chỉ mua 1 sản phẩm → hầu như không có cặp sản phẩm nào được mua cùng nhau đủ lớn. => Vậy ta sẽ chọn min\_support = 0.001 (0.1%)



Hình các thông số trong thuật toán Apriori

Nhận xét:

* Các sản phẩm giày (Boots, Oxfords, Sneakers) có mối liên hệ mua kèm cực kỳ mạnh -\*\* Boots và \*\*Oxfords thường được mua cùng nhau, với:
  + Lift = 12.11: Khi mua Boots, khả năng mua thêm Oxfords cao gấp 12 lần so với ngẫu nhiên.
  + Confidence ≈ 12.7%: Cứ 100 khách mua Boots thì có ~13 khách mua thêm Oxfords.
  + Boots và Sneakers cũng có lift = 15.9: Cực kỳ mạnh! Những khách mua Boots rất hay mua thêm Sneakers, và ngược lại.

Đề xuất:

* + Triển khai combo Boots & Oxfords, Boots & Sneakers hoặc chương trình giảm giá khi mua hai sản phẩm này cùng lúc.
  + Cross-sell/Up-sell khi khách chọn 1 sản phẩm sẽ gợi ý thêm sản phẩm kia.
* Đồ uống có liên hệ mua kèm
  + Coconut water và Flavored water: Lift ≈ 3.88: Khi khách mua 1 loại, xác suất mua loại còn lại tăng gần 4 lần so với ngẫu nhiên.
  + Cola và Energy drink cũng có lift cao (~10.5), cho thấy khách thường mua nước giải khát đa dạng trong 1 lần mua sắm.

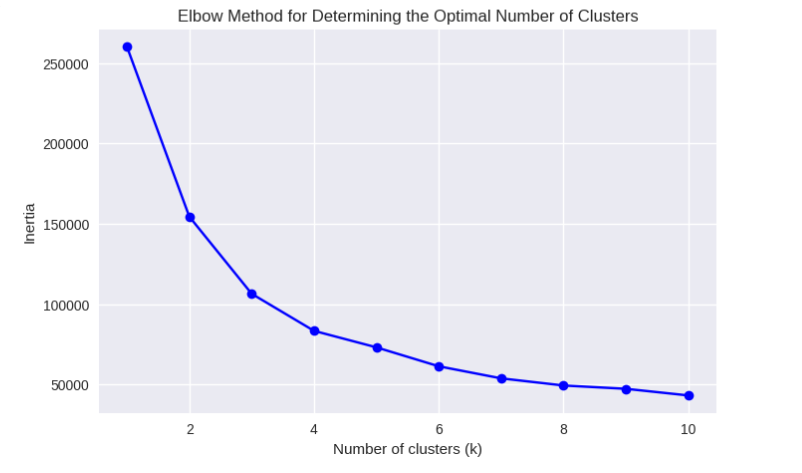
Đề xuất:

* + Bán kèm combo các loại nước này, đặt gần nhau tại điểm bán hàng, hoặc tạo chương trình ưu đãi nhóm nước uống.
* Sản phẩm công nghệ thường mua kèm nhau
  + Sony Xperia, Xiaomi Mi, iPhone, LG G, Google Pixel thường xuất hiện cùng nhau trong các giỏ hàng.
  + Các lift đều trên 5, có nghĩa là khả năng khách mua nhiều loại smartphone trong cùng một giao dịch cao hơn nhiều so với ngẫu nhiên.

Đề xuất:

* + Đặt các sản phẩm công nghệ cạnh nhau trong cửa hàng/trang web, có thể tạo gói phụ kiện chung cho các dòng smartphone.

***Phân khúc khách hàng***

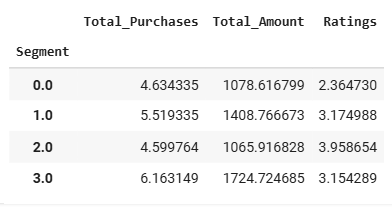


Biểu đồ K(cụm) tối ưu theo phương pháp “khủy tay”

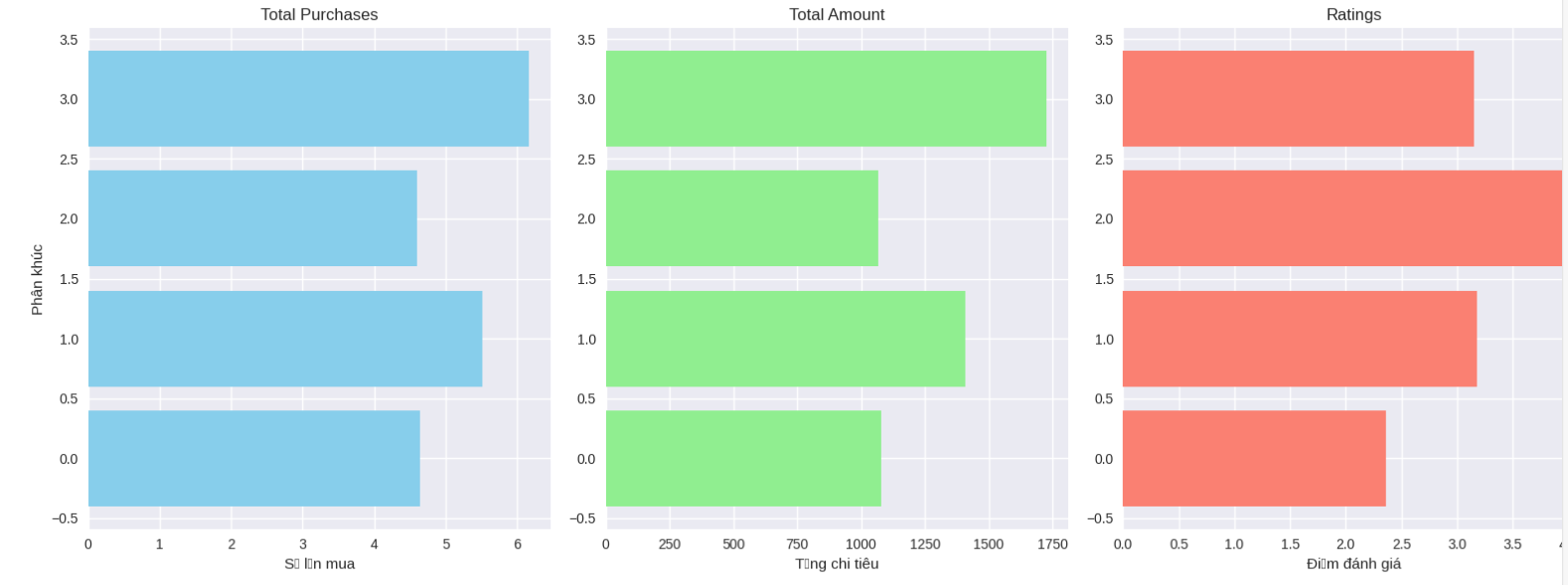
Kết luận từ biểu đồ Elbow Biểu đồ Elbow cho thấy rằng tổng bình phương sai số trong cụm (Inertia) giảm nhanh khi tăng số cụm từ k = 1 đến k = 3, sau đó bắt đầu giảm chậm dần. Điều này cho thấy "khuỷu tay" (elbow) xuất hiện rõ ở k = 3 hoặc k = 4. -> Vì vậy, số cụm tối ưu cho mô hình KMeans nên chọn là 3 hoặc 4:

* + k = 3 phù hợp nếu cần phân khúc đơn giản (ví dụ: khách hàng tốt – trung bình – kém)
  + k = 4 phù hợp nếu muốn phân khúc chi tiết hơn (ví dụ: thêm nhóm khách hàng tiềm năng hoặc VIP). Ở đây sẽ áp dụng cụm 4 để phân khúc chi tiết khách hàng

***Áp dụng KMeans để phân cụm khách hàng***



Biểu đồ dùng Kmeans để phân cụm khách hàng



Biểu đồ phân cụm khách hàng

Dựa vào biểu đồ phân cụm khách hàng ta thấy:

* Mục tiêu Phân tích và chia khách hàng thành các nhóm (segment) dựa trên 3 yếu tố hành vi:
  + Số lần mua hàng (Total\_Purchases)
  + Tổng số tiền đã chi tiêu (Total\_Amount)
  + Đánh giá của khách hàng (Ratings)
* Kết quả phân cụm Sau khi áp dụng phương pháp phân cụm KMeans, dữ liệu khách hàng được chia thành 4 nhóm chính. Bảng dưới đây thể hiện giá trị trung bình của từng đặc điểm trong mỗi phân khúc:

Tổng kết:



Bảng thể hiện giá trị trung bình từng đặc điểm mỗi phân khúc

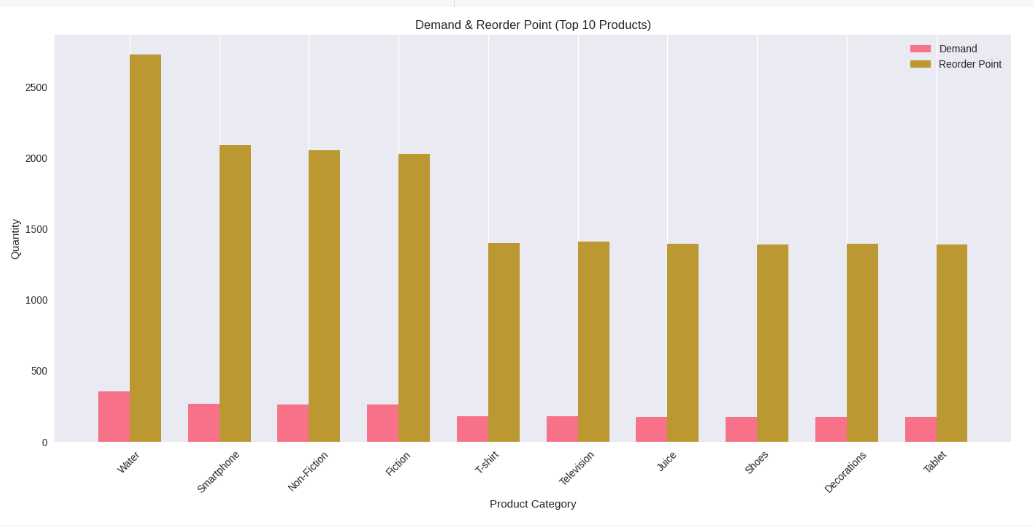
Kết luận:

* Dựa trên kết quả phân cụm, có thể rút ra một số nhận định chính:
  + Phân khúc 3 là nhóm khách hàng tiềm năng nhất với mức chi tiêu và số lần mua cao nhất. Đây là nhóm cần được ưu tiên chăm sóc và giữ chân.
  + Phân khúc 0 cũng là một nhóm có giá trị với mức chi tiêu cao và hành vi mua hàng ổn định, tuy đánh giá không quá nổi bật.
  + Phân khúc 1 tuy có đánh giá tốt nhất, nhưng tần suất mua và chi tiêu thấp. Có thể là nhóm khách hàng mới hoặc thận trọng — cần chiến lược kích thích mua hàng.
  + Phân khúc 2 là nhóm có giá trị thấp nhất cả về đánh giá lẫn hành vi chi tiêu và mua sắm. Cần xem xét nguyên nhân để cải thiện trải nghiệm hoặc chuyển hướng nguồn lực sang nhóm khách hàng khác.

***Tối ưu hóa tồn kho***



Bảng các sản phẩm với nhu cầu hằng ngày



Biểu đồ top 10 sản phẩm có nhu cầu cao nhất và order lại

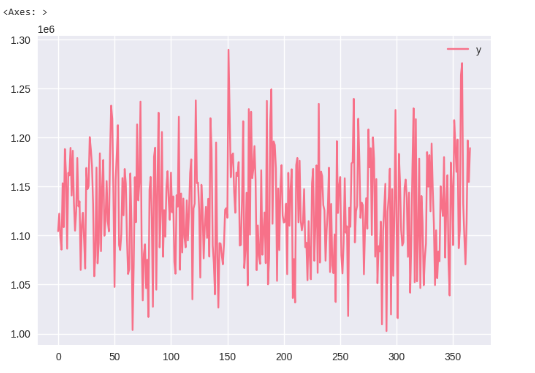
Nhận xét:

* Biểu đồ cột trên thể hiện mối tương quan giữa nhu cầu thực tế (Demand) và ngưỡng đặt hàng (Reorder Point) của 10 sản phẩm có doanh thu cao nhất, qua đó cung cấp cái nhìn toàn diện về chiến lược tồn kho và mức độ tiêu thụ sản phẩm theo từng danh mục cụ thể.
* Phân tích cho thấy sự chênh lệch rõ rệt giữa lượng cầu (màu hồng) và ngưỡng tồn kho an toàn (màu vàng đồng). Đối với tất cả các sản phẩm, Reorder Point đều cao gấp nhiều lần so với nhu cầu tiêu thụ thực tế, phản ánh chiến lược quản lý hàng tồn kho thận trọng, thiên về phòng ngừa rủi ro thiếu hàng và đảm bảo khả năng cung ứng liên tục.Water (nước uống) tiếp tục giữ vị trí dẫn đầu về nhu cầu, đồng thời cũng sở hữu mức Reorder Point cao nhất (~2.700 đơn vị), cho thấy đây là sản phẩm có tốc độ tiêu thụ nhanh, nhu cầu ổn định và liên tục. Điều này hoàn toàn phù hợp với đặc tính thiết yếu và vòng đời tiêu dùng ngắn của sản phẩm này.
* Nhóm sản phẩm Smartphone, Non-Fiction và Fiction cũng duy trì mức cầu cao (~250–300 đơn vị), với ngưỡng đặt hàng ở mức trên 2.000 đơn vị. Những con số này cho thấy đây là các sản phẩm chủ lực, có giá trị cao, đòi hỏi chiến lược tồn kho chặt chẽ để không làm gián đoạn doanh thu.
* Đối với các sản phẩm như T-shirt, Television, Juice, Shoes, Decorations và Tablet, tuy mức nhu cầu có phần thấp hơn (khoảng 180–240 đơn vị), ngưỡng đặt hàng vẫn duy trì ở mức đồng đều (~1.400 đơn vị). Điều này phản ánh chính sách Reorder Point được áp dụng theo hướng ổn định hoá quy trình kho hàng thay vì tối ưu riêng theo từng sản phẩm.
* Điểm đáng chú ý là không có sản phẩm nào có Reorder Point thấp hơn Demand, điều này minh chứng cho chiến lược tồn kho có phần “an toàn cao” của doanh nghiệp, nhằm hạn chế triệt để rủi ro thiếu hàng, nhất là trong các mùa cao điểm hoặc sự kiện bán hàng đặc biệt.

Đề xuất:

* Doanh nghiệp có thể xem xét điều chỉnh Reorder Point theo xu hướng biến động theo tháng/quý để giảm thiểu tồn kho dư thừa, đặc biệt đối với nhóm sản phẩm có nhu cầu không quá cao nhưng đang duy trì tồn kho lớn như Tablet hoặc Decorations.
* Tích hợp mô hình dự báo nhu cầu (Prophet/ARIMA) kết hợp với tính toán Safety Stock và Lead Time có thể giúp tối ưu hơn nữa chiến lược tồn kho, tiết kiệm chi phí và gia tăng khả năng phản ứng linh hoạt với biến động thị trường.
* Phân tích này là cơ sở quan trọng để xây dựng chính sách bổ sung hàng hoá thông minh trong các chiến dịch lớn như Black Friday, Tết hoặc các kỳ khuyến mãi theo mùa.

***Dự báo doanh thu bằng Prophet***



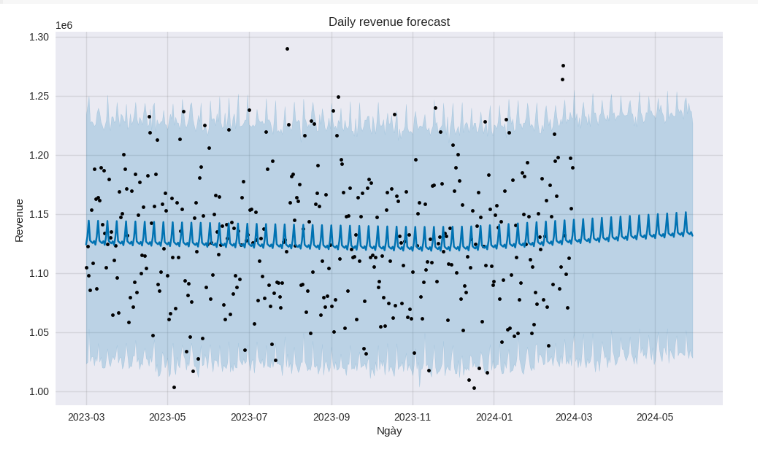
Biểu đồ mô hình dự báo doanh thu bằng Prophet

Nhận xét:

* Biến động mạnh: Dữ liệu doanh thu thể hiện dao động lớn theo thời gian, với nhiều đỉnh và đáy phân bố không đều, thể hiện tính chất ngẫu nhiên và không có chu kỳ rõ ràng.
* Mức doanh thu ổn định quanh trung bình 1.15 triệu: Mặc dù có sự dao động, phần lớn giá trị doanh thu nằm trong khoảng 1.10 đến 1.20 triệu mỗi ngày, chứng tỏ thị trường duy trì được một mức doanh thu trung bình khá ổn định trong suốt năm.
* Một số đột biến cao (> 1.25 triệu) và thấp (< 1.05 triệu) có thể liên quan đến các sự kiện đặc biệt như chiến dịch giảm giá, thay đổi sản phẩm hoặc gián đoạn cung ứng

Đề xuất:

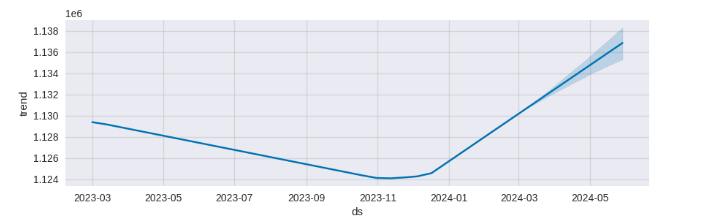
* Tăng cường các yếu tố mùa vụ trong dữ liệu đầu vào (ví dụ: chỉ số tháng, ngày trong tuần, dịp lễ...) để Prophet có thể phát hiện mô hình chu kỳ rõ hơn.
* Xem xét loại bỏ nhiễu hoặc sử dụng phương pháp trung bình trượt để làm mượt đường xu hướng.
* Phân tích sâu hơn những ngày có doanh thu đột biến để đánh giá tác động từ chiến dịch bán hàng hoặc thị trường.



Biểu đồ dự báo doanh thu theo ngày

Nhận xét:

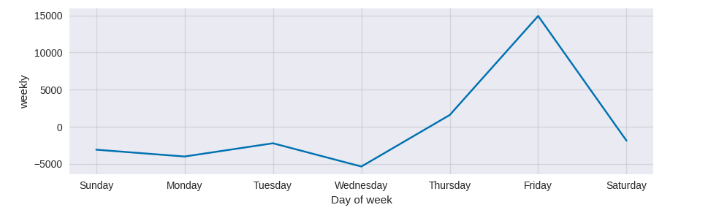
* Biểu đồ thể hiện xu hướng doanh thu hàng ngày từ tháng 3/2023 đến tháng 5/2024 với mức trung bình ổn định khoảng 1.13 triệu. Trong giai đoạn này, doanh thu không có sự biến động quá lớn, phản ánh sự ổn định trong hoạt động kinh doanh.
* Các đỉnh dao động theo chu kỳ ngắn hạn (khoảng 7 ngày) cho thấy rõ yếu tố mùa vụ theo tuần, đặc biệt vào các ngày cuối tuần có thể là thời điểm mua sắm cao hơn bình thường.
* Vùng dự báo thể hiện bằng dải màu xanh nhạt bao quanh đường trung bình, phần lớn các điểm thực tế (chấm đen) nằm trong vùng này, cho thấy mô hình dự báo hoạt động hiệu quả và độ tin cậy cao.
* Một số điểm nằm ngoài vùng dự báo là những ngoại lệ có thể do sự kiện bất thường như chương trình khuyến mãi, ngày lễ lớn hoặc thay đổi hành vi người tiêu dùng trong ngắn hạn.
* Dữ liệu cho thấy doanh thu duy trì ổn định và có thể được dự báo chính xác trong tương lai gần, giúp hỗ trợ hiệu quả cho việc lập kế hoạch tài chính, tồn kho và chiến lược kinh doanh.



Biểu đồ xu hướng doanh thu theo tháng

Nhận xét:

* Từ tháng 3/2023 đến khoảng tháng 11/2023, doanh thu có xu hướng giảm nhẹ nhưng đều đặn, cho thấy một giai đoạn thị trường ổn định nhưng có phần suy yếu.
* Từ cuối năm 2023 trở đi, đặc biệt là từ tháng 1/2024, xu hướng đảo chiều rõ rệt: doanh thu bắt đầu phục hồi mạnh và tăng liên tục đến tháng 5/2024.
* Sự phục hồi này có thể phản ánh các yếu tố như: cải thiện kinh tế, chiến lược kinh doanh hiệu quả, hoặc tác động của mùa cao điểm đầu năm (ví dụ: lễ hội, Tết, khuyến mãi...).



Biểu đồ doanh thu theo thứ trong tuần

Nhận xét:

* Thứ Sáu (Friday) là ngày có doanh thu cao nhất, với mức tăng gần 15,000 đơn vị so với mức trung bình — cho thấy đây là ngày “vàng” trong tuần để đẩy mạnh bán hàng hoặc chạy chiến dịch marketing.
* Thứ Tư (Wednesday) ngược lại là ngày yếu nhất trong tuần, doanh thu có xu hướng thấp hơn trung bình.
* Cuối tuần (Saturday) cũng giảm đáng kể so với ngày cao điểm (Friday), có thể phản ánh thói quen tiêu dùng ít đi vào ngày này (hoặc tạm nghỉ mua sắm).
* Các ngày còn lại như Thứ Hai đến Thứ Năm có mức biến động nhỏ và khá ổn định.

Kết luận:

* Thời gian phù hợp để gia tăng doanh thu: Tập trung đẩy mạnh các chiến dịch vào thứ Sáu và đầu năm 2024, khi cả xu hướng và mùa vụ đều thuận lợi.
* Cải thiện hiệu suất thấp vào giữa tuần: Cần xem xét nguyên nhân doanh số thấp vào thứ Tư — có thể là do thiếu hoạt động kích cầu hoặc thời điểm không thuận tiện cho người tiêu dùng.
* Ứng dụng thực tiễn:
  + Xây dựng lịch trình khuyến mãi theo tuần, tập trung ưu tiên ngân sách cho thứ Sáu.
  + Sử dụng xu hướng dài hạn để lập kế hoạch tài chính, nhập hàng và nhân sự, đặc biệt chuẩn bị kỹ cho giai đoạn phục hồi từ đầu 2024.

***Kết quả đánh giá mô hình***



Hình điểm kết quả đánh giá mô hình

4. Projected Impact

4.1. Accomplishments and Benefits

Trong quá trình thực hiện dự án, nhóm đã đạt được một số thành tựu quan trọng cả về mặt kỹ thuật lẫn ứng dụng thực tiễn:

* Tăng cường chất lượng dữ liệu đầu vào: Thông qua quá trình tiền xử lý và làm sạch dữ liệu một cách hệ thống, nhóm đã loại bỏ các biến dư thừa, xử lý giá trị thiếu hiệu quả, chuẩn hóa dữ liệu theo chuẩn định dạng phù hợp, qua đó đảm bảo đầu vào chất lượng cao cho các mô hình phân tích và dự báo.
* Phân tích chuyên sâu & trực quan hóa: Các biểu đồ và mô hình phân tích được xây dựng giúp cung cấp góc nhìn rõ ràng về xu hướng doanh thu, nhu cầu sản phẩm, hiệu quả theo thành phố/khu vực, cũng như các biến động theo thời gian. Điều này hỗ trợ các nhà quản lý trong việc ra quyết định nhanh chóng và chính xác hơn.
* Dự báo doanh thu chính xác: Việc ứng dụng các mô hình học máy (machine learning) và mô hình chuỗi thời gian (time series forecasting) đã giúp dự đoán doanh thu hàng ngày với độ tin cậy cao. Điều này không chỉ hỗ trợ việc lập kế hoạch ngân sách, mà còn giúp xác định thời điểm cần đẩy mạnh marketing, tối ưu hóa tồn kho và nguồn lực.
* Tối ưu hóa tồn kho & chuỗi cung ứng: Việc xác định điểm đặt hàng (Reorder Point) phù hợp cho từng sản phẩm dựa trên nhu cầu thực tế giúp giảm thiểu tình trạng thiếu hàng hoặc tồn kho quá mức, từ đó giảm chi phí và tăng hiệu quả hoạt động logistics.
* Hỗ trợ chiến lược kinh doanh cá nhân hóa: Phân tích phản hồi khách hàng (Ratings & Feedback) và hành vi tiêu dùng giúp doanh nghiệp cá nhân hóa chiến lược tiếp thị, cải thiện trải nghiệm khách hàng và tăng tỷ lệ giữ chân người dùng.

4.2. Future Improvements

* Mở rộng phân tích dữ liệu thời gian
  + Thu thập thêm dữ liệu lịch sử để có bộ dữ liệu đa năm, giúp xác định rõ hơn các xu hướng mùa vụ, chu kỳ kinh doanh và các điểm bất thường theo từng thời kỳ.
  + Phân tích chuỗi thời gian sâu hơn sử dụng mô hình dự đoán như ARIMA, Prophet, LSTM để dự báo chính xác doanh thu theo tháng hoặc quý, qua đó hỗ trợ ra quyết định tốt hơn.
* Đa dạng hóa phân tích hành vi khách hàng
  + Xây dựng mô hình phân cụm khách hàng (Customer Segmentation) dựa trên hành vi mua sắm để tạo ra các chiến lược marketing và bán hàng phù hợp cho từng nhóm khách hàng.
  + Ứng dụng các thuật toán Recommendation System để đề xuất sản phẩm phù hợp, tăng khả năng cross-selling và up-selling, cải thiện doanh thu.
* Tăng cường phân tích đa chiều (Multidimensional Analysis)
  + Phân tích doanh thu theo nhiều chiều dữ liệu như sản phẩm, địa lý, kênh bán hàng, thời gian, phân khúc khách hàng,… từ đó đưa ra các insight sâu sắc, toàn diện, hỗ trợ hoạch định chiến lược lâu dài.

5. Team Member Review and Comment

|  |
| --- |
| <ATTACH A TEAM PICTURE HERE> |

|  |  |
| --- | --- |
| NAME | REVIEW and COMMENT |
| Trương Thái Bảo | * Hoàn thành tốt công việc được giao |
| Nguyễn Đào Hồng Như | * Hoàn thành tốt công việc được giao |
| Nguyễn Lư Hồng Phương | * Hoàn thành tốt công việc được giao |
| Nguyễn Kim Vương Ý | * Hoàn thành tốt công việc được giao |
| Nguyễn Văn Tú | * Hoàn thành tốt công việc được giao |

6. Instructor Review and Comment

|  |  |  |
| --- | --- | --- |
| CATEGORY | SCORE | REVIEW and COMMENT |
| IDEA | \_\_/10 |  |
| APPLICATION | \_\_/30 |  |
| RESULT | \_\_/30 |  |
| PROJECT MANAGEMENT | \_\_/10 |  |
| PRESENTATION & REPORT | \_\_/20 |  |
| TOTAL | \_\_/100 |  |