ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC KINH TẾ - LUẬT

BÁO CÁO KIẾN TẬP
CHUYÊN NGÀNH PHÂN TÍCH DỮ LIỆU
(DATA ANALYST)

Tên học phần: Kiến tập.

Mã học phần: 243KT47.

Giảng viên hướng dẫn: ThS. Triệu Việt Cường.

Thành viên - MSSV:

1. Trần Nguyên Hưng - K224111393.

2. Nguyễn Trần Hiếu Nhân - K224111412.

TP. HCM, 7/2025

MỤC LỤC

DANH MỤC HÌNH ẢNH

PHIẾU ĐÁNH GIÁ KẾT QUẢ KIẾN TẬP

MSSV: K224111393

Họ tên: Trần Nguyên Hưng

Tên đơn vị kiến tập: Công ty TNHH Getz Group

Địa chỉ đơn vị: 271 đường Nguyễn Văn Linh, Đà Nẵng, Vietnam

Họ tên người hướng dẫn: Võ Nguyên Hoài Thương

Chức vụ: Chuyên viên Phân tích Dữ liệu Điện thoại: 0905535781

Thời gian kiến tập: Từ ngày 18/04/2025 Đến ngày: 22/07/2025

** Đánh giá **bằng cách đánh dấu X** vào cột xếp loại các nội dung đánh giá trong bảng sau.*

Ghi chú:
Loại A: 4đ; Loại B: 3đ
Loại C: 2đ; Loại D: 1đ

Nội dung đánh giá	Xếp loại			
	A	**B**	**C**	**D**
1. Tinh thần kỷ luật, thái độ				
1.1 Chấp hành giờ giấc làm việc	x			
1.2 Thái độ giao tiếp	x			
1.3 Tích cực trong công việc	x			
2. Kỹ năng chuyên môn, nghiệp vụ				
2.1 Đáp ứng yêu cầu công việc		x		
2.2 Tinh thần học hỏi, nâng cao trình độ chuyên môn	x			
2.3 Có sáng kiến, năng động trong công việc		x		
3. Kết quả kiến tập				
3.1 Kết quả có khả năng ứng dụng thực tế	x			
3.2 Mức độ hoàn thành nhiệm vụ kiến tập	x			

PHIẾU ĐÁNH GIÁ KẾT QUẢ KIẾN TẬP

MSSV: K224111412

Họ tên: Nguyễn Trần Hiếu Nhân

Tên đơn vị kiến tập: Công ty TNHH Getz Group

Địa chỉ đơn vị: 271 đường Nguyễn Văn Linh, Đà Nẵng, Vietnam

Họ tên người hướng dẫn: Võ Nguyên Hoài Thương

Chức vụ: Chuyên viên Phân tích Dữ liệu Điện thoại: 0905535781

Thời gian kiến tập: Từ ngày 18/04/2025 Đến ngày: 22/07/2025

** Đánh giá **bằng cách đánh dấu** **X** vào cột xếp loại các nội dung đánh giá trong bảng sau.*

Ghi chú:
Loại A: 4đ; Loại B: 3đ
Loại C: 2đ; Loại D: 1đ

Nội dung đánh giá	Xếp loại			
	A	**B**	**C**	**D**
1. Tinh thần kỷ luật, thái độ				
1.4 Chấp hành giờ giấc làm việc	x			
1.5 Thái độ giao tiếp	x			
1.6 Tích cực trong công việc	x			
2. Kỹ năng chuyên môn, nghiệp vụ				
2.4 Đáp ứng yêu cầu công việc		x		
2.5 Tinh thần học hỏi, nâng cao trình độ chuyên môn	x			
2.6 Có sáng kiến, năng động trong công việc		x		
3. Kết quả kiến tập				
3.3 Kết quả có khả năng ứng dụng thực tế	x			
3.4 Mức độ hoàn thành nhiệm vụ kiến tập	x			

PHIẾU NHẬN XÉT KIẾN TẬP

Hoàn thành tốt và đầy đủ. Giải pháp có tính ứng dụng cao.

TP. Hồ Chí Minh, ngày 22 tháng 7 năm 2025

XÁC NHẬN CỦA NGƯỜI HƯỚNG DẪN

Ký và ghi rõ họ tên

Võ Nguyên Hoài Thương

PHIẾU ĐÁNH GIÁ KẾT QUẢ KIẾN TẬP

CỦA GIẢNG VIÊN HƯỚNG DẪN

Họ tên - MSSV: Trần Nguyên Hưng - K224111393

Họ tên - MSSV: Nguyễn Trần Hiếu Nhân - K224111412

GVHD: ThS. Triệu Việt Cường

STT	Tiêu chí	Tiêu chí cụ thể	Điểm	Ghi chú
1	**Hình thức báo cáo (15%)**	Trình bày (5%)		
		Kết cấu báo cáo (5%)		
		Văn phong (5%)		
2	**Nội dung báo cáo (50%)**	Kỹ năng phân tích (10%)		
		Mục tiêu (10%)		
		Chuyên môn (30%)		
3	**Thái độ của sinh viên (15%)**			
4	**Doanh nghiệp đánh giá (20%)**			*GVHD quy đổi từ đánh giá của DN*
TỔNG ĐIỂM				

……., ngày……..tháng…….năm……

Giảng viên hướng dẫn

Ký, ghi rõ họ tên

CHƯƠNG I: ĐỊNH HƯỚNG NGHỀ NGHIỆP

1. Khái niệm

Data Analyst, hay còn gọi là chuyên viên phân tích dữ liệu, là người chuyên thu thập, làm sạch và phân tích dữ liệu nhằm tìm ra câu trả lời cho những câu hỏi hoặc giải pháp cho các vấn đề cụ thể mà doanh nghiệp đang gặp phải. Data Analyst có thể làm việc trong rất nhiều lĩnh vực khác nhau, chỉ cần ngành nào có dữ liệu thì ngành đó sẽ cần phân tích dữ liệu để ra quyết định.

2. Kiến thức

Trong quá trình làm việc của một Data Analyst, việc nắm vững các kiến thức cốt lõi và sử dụng thành thạo các công cụ là điều kiện cần đầu tiên để có thể thực hiện được các nhiệm vụ liên quan như thu thập, tổ chức, xử lý, phân tích và trực quan hóa dữ liệu. Làm sao ta có thể tìm được ý nghĩa của bộ dữ liệu nếu như ta không thể phân tích thậm chí là truy vấn được nó, vì vậy thành thạo các kỹ năng chuyên môn là điều gần như phải có ở tất cả các Data Analyst.

2.1. Spreadsheets (Excel, Google Sheet)

Spreadsheets hay còn gọi là trang tính là công cụ vô cùng quen thuộc không chỉ với những người trong lĩnh vực Data Analyst mà còn được rất nhiều vị trí khác trong doanh nghiệp sử dụng. Microsoft Excel và Google Sheets là hai ứng dụng phổ biến nhất hiện nay, cung cấp khả năng thu thập, lưu trữ, tổ chức và xử lý dữ liệu. Tuy nhiên Spreadsheets chỉ phù hợp để phân tích các tập dữ liệu nhỏ hoặc trung bình, đối với những dữ liệu có kích thước quá lớn cần những công cụ khác để xử lý.

Ngoài ra, Spreadsheets còn tích hợp các công cụ trực quan hóa như biểu đồ giúp trực quan hóa dữ liệu thành hình ảnh dễ hiểu cho người xem. Trong quy trình phân tích dữ liệu, Spreadsheets thường là bước khởi đầu để khám phá dữ liệu, thực hiện các thao tác xử lý cơ bản trước khi đi sâu vào các phân tích phức tạp hơn.

2.2. Databases & Query Languages (SQL)

Cơ sở dữ liệu (Database) là nền tảng lưu trữ thông tin có cấu trúc, cho phép người dùng truy cập, quản lý và thao tác dữ liệu. Trong lĩnh vực phân tích dữ liệu, việc làm việc trực tiếp với cơ sở dữ liệu là điều không thể thiếu, đặc biệt đối khối lượng dữ liệu lớn và phức tạp hơn so với khả năng xử lý của Spreadsheets ở trên. Các hệ quản trị cơ sở dữ liệu phổ biến có thể kể đến như MySQL, PostgreSQL, Microsoft SQL Server.

Để tương tác được với hệ thống cơ sở dữ liệu, các Data Analyst sử dụng ngôn ngữ truy vấn có cấu trúc (SQL – Structured Query Language). Theo hầu hết các yêu cầu, SQL thường được sử dụng để viết truy vấn nhằm lọc, nhóm, sắp xếp và kết hợp dữ liệu theo yêu cầu phân tích. Khả năng viết truy vấn SQL thành thạo là một kỹ năng bắt buộc đối với bất kỳ Data Analyst nào.

2.3. Visualization Tools (Tableau, Power BI)

Trực quan hóa dữ liệu (data visualization) là một bước quan trọng trong quy trình phân tích dữ liệu, giúp biến những con số khô khan thành hình ảnh trực quan từ đó đưa ra kết quả phân tích một cách sinh động, dễ hiểu và dễ truyền đạt đến người xem, đặc biệt là các bên liên quan không chuyên về kỹ thuật như khách hàng hay những bộ phận khác trong doanh nghiệp.

Hai công cụ trực quan hóa phổ biến nhất hiện nay là Tableau và Power BI. Cả hai đều hỗ trợ tạo biểu đồ, bản đồ, bảng, và dashboard tương tác, cho phép người dùng khám phá dữ liệu một cách linh hoạt. Đối với một Data Analyst, việc sử dụng các công cụ trực quan hóa không chỉ giúp minh họa kết quả phân tích mà còn hỗ trợ quá trình xác định xu hướng, mối quan hệ giữa các biến mà nếu chỉ nhìn vào số liệu không thể xác định được. Và quan trọng nhất nhờ vào những biểu đồ rõ ràng cùng cơ sở từ dữ liệu thực tế nên ban lãnh đạo có thể đưa ra những quyết định kịp thời, chính xác.

2.4. Programming Languages (Python, R)

Đối với lĩnh vực Data Analyst, Python và R là hai ngôn ngữ lập trình phổ biến và được sử dụng rộng rãi nhất, mỗi ngôn ngữ có thế mạnh riêng tùy thuộc vào mục tiêu phân tích. Python cho phép thực hiện toàn bộ quy trình phân tích dữ liệu từ thu thập, làm sạch, xử lý, phân tích thống kê cho đến trực quan hóa trong một môi trường lập trình duy nhất. Trong khi đó, R được sử dụng chủ yếu cho việc phân tích thống kê và trực quan hóa dữ liệu, rất mạnh trong việc xây dựng mô hình thống kê và tạo biểu đồ chi tiết.

2.5. Statistics

Thống kê là kiến thức nền tảng cốt lõi nhất trong phân tích dữ liệu, giúp các Data Analyst hiểu rõ hơn về bản chất dữ liệu, phát hiện xu hướng, mối quan hệ giữa các biến từ đó có thể đưa ra những nhận định có cơ sở và đáng tin cậy từ những dữ liệu đã phân tích. Dù sử dụng bất kỳ công cụ nào, kiến thức về thống kê vẫn là yếu tố then chốt giúp Data Analyst phân tích và diễn giải dữ liệu một cách chính xác và hiệu quả.

3. Kỹ năng

3.1. Curiosity

Một trong những phẩm chất cốt lõi của Data Analyst là sự tò mò với những nhận định thực tế và luôn muốn chứng minh chúng dựa trên cơ sở dữ liệu xác thực. Đối với lĩnh vực này, việc luôn đặt câu hỏi sẽ giúp Data Analyst hình thành tư duy phân tích và tối ưu hóa trong kết quả dự đoán. Sự tò mò không chỉ dừng lại ở việc mô tả dữ liệu, mà còn là hoạt động hiểu sâu nguyên nhân và mối liên hệ đằng sau các con số. Điều này thúc đẩy quá trình khám phá insight, phát hiện các vấn đề tiềm ẩn, và đưa ra những khuyến nghị chiến lược giá trị.

3.2. Understanding of context

Dữ liệu trên thực tế luôn gắn liền với một bối cảnh, lĩnh vực kinh doanh cụ thể. Hiểu biết bối cảnh bao gồm việc hiểu về ngành nghề, mục tiêu kinh doanh, khách hàng, chuỗi giá trị, và những yếu tố có thể ảnh hưởng đến dữ liệu (xu hướng mùa vụ, xu hướng thị trường, hành vi người dùng,...). Data Analyst có khả năng phân tích bức tranh dữ liệu tổng thể sẽ tạo ra các phân tích sâu sắc và phù hợp thực tế doanh nghiệp.

3.3. Technical mindset

Data Analyst cần có tư duy kỹ thuật để xử lý và phân tích dữ liệu hiệu quả. Kỹ năng này thể hiện qua việc sử dụng các công cụ và ngôn ngữ phân tích như SQL, Python, Excel, Power BI, Tableau,… để xử lý dữ liệu thô thành thông tin có ý nghĩa. Ngoài ra, tư duy kỹ thuật còn bao gồm khả năng lập trình, phân tích tối ưu, tự động hóa quy trình, xây dựng các báo cáo tương tác, và kết nối dữ liệu từ nhiều nguồn khác nhau. Tư duy kỹ thuật giúp Data Analyst không chỉ phân tích tốt mà còn làm việc hiệu quả và mở rộng được quy mô phân tích.

3.4. Data design

Data Design là khả năng tổ chức cấu trúc và trình bày dữ liệu sao cho dễ hiểu, dễ sử dụng và có tính ứng dụng cao. Kỹ năng này bao gồm thiết kế mô hình dữ liệu hợp lý (data modeling), xây dựng data pipeline, chọn định dạng báo cáo trực quan phù hợp và sử dụng biểu đồ, dashboard hiệu quả. Một nhà phân tích biết cách thiết kế dữ liệu tốt sẽ giúp người dùng cuối có thể hiểu và sử dụng dữ liệu một cách nhanh chóng, chính xác để đưa ra các quyết định phù hợp.

3.5. Data strategy

Data strategy là khả năng tư duy dài hạn về cách tổ chức sử dụng dữ liệu để đạt mục tiêu kinh doanh. Data Analyst đóng vai trò hỗ trợ thiết lập hoặc tư vấn cho chiến lược dữ liệu tổng thể, góp phần tạo ra nền tảng vững chắc cho các quyết định dựa trên dữ liệu. Chiến lược xử lý dữ liệu cần tập trung vào việc xác định: loại dữ liệu nào cần thu thập, lưu trữ ra sao, ai là người sử dụng, làm thế nào để bảo mật cũng như đảm bảo chất lượng dữ liệu và dữ liệu sẽ có giá trị trong những khoảng thời gian nào.

4. Vai trò của Data Analyst trong doanh nghiệp hiện nay

Data Analyst đóng vai trò ngày càng quan trọng trong hầu hết các doanh nghiệp hiện nay, đặc biệt trong bối cảnh chuyển đổi số toàn diện. Hơn thế nữa, thị trường cạnh tranh ngày càng khốc liệt xuất phát từ xu hướng hội nhập sâu rộng trên toàn thế giới đòi hỏi các doanh nghiệp phải đầu tư mạnh mẽ hơn nữa vào dữ liệu, thông tin như một lợi thế cạnh tranh của doanh nghiệp. Một số vai trò nổi bật của Data Analyst trong doanh nghiệp có thể kể đến như sau:

- Thứ nhất, Data Analyst thực hiện nhiệm vụ thu thập, xử lý và phân tích dữ liệu từ nhiều nguồn khác nhau (hệ thống CRM, ERP, mạng xã hội, website,…) của doanh nghiệp. Từ đó, họ tạo ra các báo cáo, dashboard trực quan nhằm hỗ trợ Business Analyst cũng như các cấp quản lý ra quyết định đúng đắn, nhanh chóng và dựa trên bằng chứng dữ liệu xác thực.

- Thứ hai, Data Analyst có khả năng cung cấp các dự báo cho doanh nghiệp (Doanh thu tương lai, nhu cầu thị trường, rủi ro tài chính,...). Các dự báo này có thể giúp doanh nghiệp chủ động hơn trong việc lập kế hoạch và phân bổ nguồn lực hợp lý.

- Thứ ba, Data Analyst có thể đóng vai trò cầu nối giữa bộ phận kỹ thuật (IT, Data Engineering, Developer,...) và bộ phận kinh doanh. Với đặc trưng am hiểu kiến thức về kỹ thuật, ngôn ngữ dữ liệu lẫn lĩnh vực phân tích, Data Analyst sẽ góp phần quan trọng trong việc truyền đạt thông tin và chuyển hóa dữ liệu thành giải pháp tại doanh nghiệp.

- Cuối cùng, Data Analyst là thành phần cốt lõi trong quá trình chuyển đổi số doanh nghiệp, bộ phận này sẽ góp phần xây dựng tư duy và văn hóa dữ liệu tại doanh nghiệp. Trên cơ sở đó, hoạt động ra quyết định của doanh nghiệp sẽ trở nên linh hoạt, chính xác và hiệu quả hơn khi dựa trên dữ liệu cụ thể.

5. Cơ hội và thách thức trong ngành

5.1. Cơ hội

Trong bối cảnh chuyển đổi số toàn cầu và sự phát triển mạnh mẽ của công nghệ thông tin, ngành Data Analyst đang mở ra nhiều cơ hội hấp dẫn và tiềm năng phát triển rộng lớn.

- Thứ nhất, với lượng dữ liệu khổng lồ được tạo ra mỗi ngày ở nhiều lĩnh vực khác nhau từ y tế, nông nghiệp đến kinh doanh, mạng xã hội, nhu cầu khai thác, phân tích và ứng dụng dữ liệu để tạo ra giá trị và đưa ra quyết định cho doanh nghiệp ngày càng tăng cao. Điều này tạo ra cơ hội lớn cho Data Analyst phát huy vai trò trọng yếu trong việc hỗ trợ ra quyết định dựa trên dữ liệu chính xác và kịp thời.

- Thứ hai, với sự phát triển không ngừng của các công cụ và nền tảng phân tích dữ liệu hiện đại như Power BI, Tableau, Python, R hay SQL giúp Data Analyst dễ dàng tiếp cận và nâng cao năng lực chuyên môn. Đồng thời, sự đa dạng về các nguồn và dữ liệu học tập, từ các khóa học trực tuyến miễn phí và trả phí đến những cộng đồng chia sẻ kiến thức trên toàn cầu đã tạo điều kiện thuận lợi cho việc học tập liên tục, thích ứng nhanh với công nghệ mới. Điều này không chỉ mở rộng cơ hội tiếp cận nghề nghiệp cho nhiều đối tượng mà còn thúc đẩy sự phát triển của ngành trong dài hạn.

- Cuối cùng, trong bối cảnh toàn cầu hóa và xu hướng làm việc từ xa ngày càng phổ biến, thêm vào đó Data Analyst là một nghề phổ biến ở rất nhiều quốc gia, vì vậy ngày càng có nhiều cơ hội tham gia vào các dự án quốc tế, làm việc với doanh nghiệp đa quốc gia hoặc nhận những dự án phân tích dữ liệu từ nước ngoài trên các nền tảng việc làm trực tuyến. Đây cũng là một cơ hội phát triển rất lớn và tiềm năng cho thị trường Data Analyst hiện tại.

5.2. Thách thức

Mặc dù có nhiều cơ hội và tiềm năng trong bối cảnh chuyển đổi số toàn cầu, song Data Analyst đang phải đối mặt với không ít thách thức cả về kỹ thuật lẫn chiến lược hỗ trợ ra quyết định, nhất là khi vai trò dữ liệu ngày càng trở nên trọng yếu trong doanh nghiệp hiện đại.

- Thứ nhất, về khối lượng dữ liệu, xu hướng dữ liệu lớn (Big Data) đặt ra cho Data Analyst yêu cầu phải xử lý lượng dữ liệu ngày càng lớn và phức tạp. Điều này đòi hỏi Data Analyst phải có kiến thức và kỹ năng sâu rộng cũng như trình độ chuyên môn cao.

- Thứ hai, công nghệ và các kiến thức chuyên môn sẽ liên tục cập nhật với tốc độ vô cùng nhanh chóng trong xuyên suốt quá trình vận hành doanh nghiệp. Vì vậy, Data

Analyst phải luôn đề cao tinh thần không ngừng học hỏi, không ngừng sáng tạo nhằm tạo ra các phân tích tối ưu cho doanh nghiệp.

- Thứ ba, chất lượng dữ liệu cũng là yếu tố đặc biệt quan trọng trong quá trình phân tích và đánh giá của Data Analyst. Ngoài việc thu thập, phân tích và dự báo dựa trên dữ liệu doanh nghiệp, Data Analyst cũng cần chú trọng hoạt động làm sạch và chuẩn hóa dữ liệu. Hoạt động này đòi hỏi sự kiên nhẫn và tập trung cao độ và sẽ là thách thức lớn đối với Data Analyst.

- Cuối cùng, một số dữ liệu nhạy cảm như tài chính, khách hàng, nhân sự,... đòi hỏi Data Analyst cần tuân thủ nghiêm ngặt về bảo mật và quyền riêng tư. Rủi ro rò rỉ dữ liệu hoặc sử dụng dữ liệu sai mục đích cũng ảnh hưởng vô cùng nghiêm trọng đến hoạt động của Data Analyst.

CHƯƠNG II: CƠ SỞ LÝ THUYẾT

1. Quy trình phân tích dữ liệu

1.1. Ask (Đặt câu hỏi)

Bước đầu tiên trong quy trình phân tích dữ liệu là đặt câu hỏi. Đây là bước nền tảng, đóng vai trò như kim chỉ nam cho toàn bộ quá trình phân tích. Nếu không hiểu rõ mình đang cố gắng giải quyết vấn đề gì, thì dù có xử lý, phân tích dữ liệu kỹ lưỡng đến đâu cũng sẽ dẫn đến kết luận sai lệch hoặc không mang lại giá trị thực tiễn. Vì vậy, giai đoạn "Ask" không đơn thuần là việc đặt ra câu hỏi, mà là quá trình xác định, làm rõ và khớp các vấn đề kinh doanh thực tế và mục tiêu phân tích dữ liệu.

Trong giai đoạn này, người phân tích dữ liệu cần dành thời gian để trao đổi trực tiếp với các bên liên quan gồm những người đề xuất hoặc bị ảnh hưởng bởi vấn đề cần phân tích. Mục đích của việc này là để hiểu được kỳ vọng, bối cảnh và các ràng buộc cụ thể trong hoạt động kinh doanh. Những câu hỏi mang tính khám phá như: "Vấn đề mà doanh nghiệp đang gặp phải là gì?", "Điều gì khiến doanh nghiệp cho rằng đây là một vấn đề cần giải quyết?", "Đã có dữ liệu nào liên quan được thu thập chưa?", hay "Kết quả lý tưởng mà doanh nghiệp mong muốn là gì?" chính là những chìa khóa giúp xác định phạm vi và định hướng phân tích đúng đắn.

Ví dụ, trong tình huống một công ty đối mặt với tỷ lệ nghỉ việc cao ở nhóm nhân viên mới, nhóm phân tích không thể chỉ dựa vào số liệu thống kê đơn thuần. Thay vào đó, họ bắt đầu bằng việc tìm hiểu sâu hơn về cảm nhận, kỳ vọng và trải nghiệm của nhân viên mới, cũng như các giả định hiện tại từ phía quản lý. Họ đặt ra những câu hỏi như: "Theo bạn, nhân viên mới cần những yếu tố nào để cảm thấy gắn bó với công ty?", "Bạn nghĩ liệu các nhà quản lý có ảnh hưởng đến tỷ lệ nghỉ việc không?", "Tỷ lệ nghỉ việc lý tưởng trong năm tới mà công ty hướng đến là bao nhiêu?". Những câu hỏi như vậy giúp làm rõ nguyên nhân cốt lõi của vấn đề.

Một điểm quan trọng trong giai đoạn Ask là khả năng chuyển hóa vấn đề kinh doanh thành câu hỏi có thể đo lường được bằng dữ liệu. Đây được gọi là biến đổi câu hỏi kinh doanh thành câu hỏi dữ liệu. Ví dụ, thay vì hỏi "Tại sao nhân viên nghỉ việc nhiều?", một câu hỏi dữ liệu cụ thể và định lượng hơn có thể là: "Nhân viên nghỉ việc có điểm chung nào trong quá trình tuyển dụng, đào tạo hoặc đánh giá hiệu suất không?" Câu hỏi này có thể được kiểm chứng bằng dữ liệu từ khảo sát, hồ sơ nhân sự, và báo cáo hiệu suất.

Ngoài ra, người phân tích cũng cần làm rõ mục tiêu thành công của dự án ngay từ đầu. Nếu không có tiêu chí đo lường rõ ràng, kết quả phân tích sẽ rất khó để đánh giá. Chẳng hạn, công ty muốn tăng tỷ lệ giữ chân nhân viên thêm 15% trong vòng 12 tháng, dựa vào đây có thể xác định được mục tiêu cụ thể, có thể đo lường và kiểm chứng.

Cuối cùng, một phần thiết yếu trong bước Ask là xác định giới hạn, ràng buộc và kỳ vọng của các bên liên quan như thời gian thực hiện, nguồn lực, độ chính xác mong muốn, và cách thức sử dụng kết quả. Đây không chỉ là bước định hướng kỹ thuật, mà còn mang tính chiến lược vì nó giúp xác lập sự thống nhất về mục tiêu ngay từ đầu.

Tóm lại, bước Ask không đơn giản là đặt câu hỏi mà là quá trình nghiên cứu, khám phá, và chuyển vấn đề kinh doanh thành mục tiêu dữ liệu cụ thể. Đây là nền móng để các bước tiếp theo như chuẩn bị, xử lý và phân tích dữ liệu được thực hiện một cách có hệ thống và đúng trọng tâm. Một bước Ask được thực hiện tốt sẽ giúp tiết kiệm thời gian, giảm rủi ro sai lệch trong quá trình phân tích và tối ưu hóa giá trị dữ liệu mang lại cho doanh nghiệp.

1.2. Prepare (Chuẩn bị dữ liệu)

Sau khi đã xác định rõ vấn đề và đặt câu hỏi đúng trong bước Ask, giai đoạn tiếp theo là chuẩn bị dữ liệu. Đây là bước then chốt nhằm xác định, thu thập và tổ chức dữ liệu cần thiết để trả lời câu hỏi đã đặt ra trước đó. Nhiều chuyên gia phân tích cho rằng, thành công của một dự án phân tích phụ thuộc đến 70-80% vào chất lượng của dữ liệu, và bước chuẩn bị dữ liệu là nơi quyết định phần lớn chất lượng đó.

Trong thực tế, dữ liệu không tự xuất hiện một cách hoàn chỉnh và sẵn sàng để phân tích. Vì vậy, bước Prepare đòi hỏi người phân tích phải chủ động xác định nguồn dữ liệu phù hợp với mục tiêu nghiên cứu. Nguồn này có thể đến từ hệ thống nội bộ như cơ sở dữ liệu hoặc bên ngoài như khảo sát, thống kê ngành, dữ liệu từ các bên thứ 3 hoặc nguồn uy tín công khai. Điều quan trọng là người phân tích phải hiểu được dữ liệu nào là cần thiết, dữ liệu nào là dư thừa hoặc gây nhiễu cho phân tích. Việc lựa chọn sai nguồn có thể dẫn đến kết luận lệch lạc, ảnh hưởng nghiêm trọng đến quyết định kinh doanh sau cùng.

Ví dụ nhóm phân tích lựa chọn hình thức khảo sát trực tuyến để thu thập dữ liệu từ nhân viên mới là những người trực tiếp liên quan đến vấn đề nghỉ việc sớm. Đây là quyết định quan trọng vì không phải tất cả dữ liệu liên quan đều đã có sẵn trong hệ thống. Ngoài ra, họ cũng xác định các biến số cần thiết như mức độ hài lòng với quy trình tuyển

dụng, đào tạo giai đoạn thử việc, đánh giá công việc và chính sách đãi ngộ. Việc xác định đúng biến số sẽ giúp định hình phạm vi dữ liệu cần thu thập và tránh lãng phí thời gian, tài nguyên vào những yếu tố không mang tính giải thích.

Không dừng lại ở việc thu thập, bước chuẩn bị còn liên quan đến lên kế hoạch thực hiện và đánh giá rủi ro dữ liệu. Nhóm phân tích không chỉ chuẩn bị câu hỏi khảo sát mà còn lường trước các kịch bản không mong muốn có thể xảy ra như nhân viên không trả lời đủ số lượng mong muốn, dữ liệu bị sai lệch do hiểu nhầm câu hỏi, hay việc truy cập dữ liệu bị gián đoạn vì lý do bảo mật. Tất cả những yếu tố này cần được xử lý bằng cách xây dựng kịch bản dự phòng và quy trình phản ứng nhanh.

Ngoài ra, một phần không thể thiếu trong giai đoạn Prepare là xác định định dạng và cấu trúc của dữ liệu ví dụ như định dạng ngày, loại biến (số, chữ, nhị phân), mã hóa thông tin định danh,... Điều này đặc biệt quan trọng nếu dữ liệu được trích xuất từ nhiều hệ thống khác nhau, vì định dạng không đồng nhất sẽ gây khó khăn trong bước xử lý sau.

Tóm lại, bước Prepare không chỉ là thu thập dữ liệu, mà là quá trình hoạch định có chiến lược về nguồn, cách thức, phạm vi và quy tắc quản lý dữ liệu. Nó đảm bảo rằng dữ liệu đưa vào phân tích là đáng tin cậy, phù hợp với mục tiêu nghiên cứu và sẵn sàng cho các bước xử lý tiếp theo. Một bước Prepare được thực hiện bài bản không chỉ giúp giảm thiểu sai sót, mà còn góp phần đáng kể vào tính hiệu quả, độ chính xác và độ tin cậy của toàn bộ dự án phân tích dữ liệu.

1.3. Process (Xử lý dữ liệu)

Sau khi đã thu thập được dữ liệu phù hợp trong bước Prepare, người phân tích bước vào giai đoạn thứ ba là xử lý dữ liệu. Đây là giai đoạn mang tính kỹ thuật cao, đóng vai trò làm bộ lọc để đảm bảo dữ liệu đã thu thập trở nên sạch sẽ, nhất quán và sẵn sàng cho bước phân tích. Nếu xem quy trình phân tích dữ liệu như việc nấu ăn, thì dữ liệu chính là nguyên liệu, và bước này tương đương với việc sơ chế, làm sạch thực phẩm trước khi đưa lên bếp.

Dữ liệu trong thực tế hiếm khi hoàn hảo. Nó có thể bị thiếu (missing values), trùng lặp (duplicates), sai định dạng (format errors), hoặc mang tính thiên lệch (bias). Tất cả những vấn đề này nếu không được phát hiện và xử lý kịp thời sẽ ảnh hưởng nghiêm trọng đến tính chính xác của kết quả. Do đó, người phân tích cần tiến hành một loạt các thao tác xử lý, bao gồm: loại bỏ dữ liệu bị lỗi, điền giá trị còn thiếu bằng phương pháp

phù hợp (trung bình, trung vị, dự đoán...), chuẩn hóa định dạng dữ liệu cho đồng bộ và mã hóa lại các giá trị dạng văn bản thành số nếu cần thiết cho việc phân tích định lượng.

Một trong những bước quan trọng nhất trong quá trình xử lý dữ liệu là làm sạch dữ liệu (data cleaning). Quá trình này yêu cầu người phân tích sử dụng kết hợp các công cụ như Microsoft Excel, Google Sheets, SQL hoặc Python để rà soát dữ liệu một cách có hệ thống. Chẳng hạn, trong tập dữ liệu khảo sát có thể tồn tại trường hợp cùng một câu trả lời nhưng được nhập khác nhau do lỗi đánh máy ("good", "Good", "Gooood"), hay số liệu bị sai do nhập nhầm đơn vị (50000 thay vì 5000). Các lỗi tưởng chừng nhỏ này nếu không được phát hiện có thể làm sai lệch kết quả phân tích và dẫn đến những quyết định sai lầm.

Một điểm đáng chú ý nữa trong bước xử lý là kiểm tra độ tin cậy và tính nhất quán của dữ liệu giữa các nguồn. Trong nhiều dự án, dữ liệu được tổng hợp từ nhiều hệ thống khác nhau như HRM, CRM, ERP,... Việc đồng bộ hóa và đối chiếu giữa các hệ thống để đảm bảo thông tin khớp với nhau là cực kỳ quan trọng. Ví dụ, dữ liệu về ngày bắt đầu làm việc của nhân viên có thể khác nhau giữa phòng nhân sự và hệ thống tính lương, nếu không phát hiện và chuẩn hóa cho đồng nhất, dữ liệu đó sẽ gây sai lệch trong việc tính thời gian làm việc, một biến quan trọng khi phân tích tỷ lệ nghỉ việc.

Cuối cùng, sau khi dữ liệu được xử lý sạch sẽ, nó cần được tổ chức lại theo cách thuận tiện cho việc phân tích. Điều này bao gồm việc tái cấu trúc bảng dữ liệu (ví dụ: từ dạng rộng sang dạng dài), thiết lập khóa chính, khóa ngoại để liên kết các bảng, và lưu trữ dữ liệu ở nơi an toàn như cơ sở dữ liệu đám mây có phân quyền bảo mật chặt chẽ.

Tóm lại, bước Process đóng vai trò là cầu nối kỹ thuật giữa dữ liệu thô và phân tích sâu. Một dữ liệu "bẩn" có thể tạo ra những quyết định sai lầm. Vì thế, xử lý dữ liệu là bước không thể bỏ qua và đòi hỏi cả kiến thức chuyên môn, tư duy hệ thống lẫn sự tỉ mỉ trong từng chi tiết nhỏ. Khi dữ liệu đã được làm sạch và chuẩn hóa, người phân tích có thể bước vào giai đoạn tiếp theo là phân tích với một nguồn dữ liệu sạch và đáng tin cậy.

1.4. Analyze (Phân tích dữ liệu)

Sau khi dữ liệu đã được chuẩn bị và xử lý đầy đủ, bước tiếp theo là phân tích dữ liệu. Đây chính là trung tâm của toàn bộ quy trình phân tích, nơi dữ liệu được chuyển hóa thành thông tin có ý nghĩa và đưa ra những phát hiện quan trọng nhằm giải đáp câu hỏi nghiên cứu đã đặt ra từ ban đầu. Nếu các bước trước đóng vai trò chuẩn bị nguyên liệu thì bước này chính là chế biến, nơi những người phân tích sử dụng kỹ năng chuyên môn,

công cụ thống kê và tư duy phản biện để khám phá ra các xu hướng và mối quan hệ trong dữ liệu.

Bản chất của phân tích dữ liệu là tìm kiếm câu trả lời từ dữ liệu, thông qua nhiều phương pháp khác nhau, từ mô tả (descriptive analysis), khám phá (exploratory analysis) đến dự đoán (predictive analysis). Ví dụ nhóm phân tích đã thực hiện một cuộc khảo sát lớn về trải nghiệm của nhân viên mới và bắt đầu phân tích kết quả với mục tiêu xác định nguyên nhân chính dẫn đến tình trạng nghỉ việc sớm. Họ không chỉ dừng lại ở việc đếm tần suất câu trả lời hay tính trung bình, mà còn so sánh giữa các nhóm nhân viên, phân tích mối liên hệ giữa từng yếu tố (quá trình tuyển dụng, đánh giá hiệu suất, chế độ đãi ngộ) với mức độ hài lòng và khả năng gắn bó lâu dài với tổ chức.

Ví dụ từ kết quả phân tích đã phát hiện ra quy trình tuyển dụng dài và phức tạp có liên quan mật thiết đến khả năng nhân viên nghỉ việc sớm, trong khi quy trình đánh giá minh bạch và hiệu quả lại có mối tương quan với sự gắn bó. Những phát hiện như vậy không chỉ phản ánh thực trạng mà còn giúp định hướng các hành động cụ thể nhằm cải thiện kết quả kinh doanh. Đó cũng là minh chứng rõ ràng cho tầm quan trọng của phân tích dữ liệu, là không chỉ để hiểu cái gì đang diễn ra, mà còn để lý giải tại sao điều đó lại xảy ra.

Ngoài ra, phân tích dữ liệu không phải lúc nào cũng đi theo một hướng duy nhất. Trong nhiều trường hợp, khi đi sâu vào phân tích, nhà phân tích có thể phát hiện ra lỗ hổng trong dữ liệu, thiếu biến quan trọng hoặc thậm chí sai sót trong cách đặt câu hỏi ban đầu. Khi đó, họ cần quay trở lại các bước trước đó (Prepare hoặc Ask) để điều chỉnh, trước khi tiếp tục phân tích.

Tóm lại, bước Analyze là nơi thể hiện rõ nhất năng lực của nhà phân tích. Một phân tích tốt không chỉ cung cấp các con số, mà còn kể được một câu chuyện rõ ràng, logic, có sức thuyết phục và có thể hành động. Đây là bước giúp doanh nghiệp hiểu rõ hơn về chính mình và đưa ra quyết định dựa trên những dữ liệu cụ thể đáng tin cậy, thay vì cảm tính hay phỏng đoán. Khi bước Analyze được thực hiện kỹ lưỡng, bước trình bày kết quả tiếp theo sẽ trở nên mạch lạc và có sức nặng hơn rất nhiều.

1.5. Share (Trình bày kết quả)

Sau khi đã hoàn tất việc phân tích và rút ra các phát hiện quan trọng từ dữ liệu, bước tiếp theo chính là trình bày kết quả. Đây là giai đoạn chuyển đổi từ phân tích nội bộ sang giao tiếp với những người ra quyết định cuối cùng, và là bước đầu tiên trong việc biến

phát hiện từ dữ liệu thành hành động thực tiễn. Dù có thực hiện một phân tích xuất sắc đến đâu, nếu kết quả không được truyền tải rõ ràng, dễ hiểu và đúng đối tượng, thì giá trị của công việc phân tích gần như bằng không. Chính vì thế, bước Share không chỉ là hoạt động báo cáo, mà còn là một quá trình truyền đạt chiến lược, đòi hỏi kỹ năng kể chuyện, trực quan hóa và thuyết phục.

Một trong những yếu tố cốt lõi trong giai đoạn Share là chuyển hóa dữ liệu khô khan thành câu chuyện dễ hiểu mà tất cả mọi người có thể hiểu được. Điều này có nghĩa là sắp xếp thông tin một cách logic, có mở đầu, diễn biến và kết luận, để người nghe đặc biệt là những người không chuyên về dữ liệu có thể dễ dàng hiểu được bản chất vấn đề và những hành động cần thiết. Ví dụ, trong trường hợp phân tích về tỷ lệ nghỉ việc cao ở nhân viên mới, nhóm phân tích đã chọn cách trình bày phát hiện theo nhóm nguyên nhân chính, đồng thời sử dụng hình ảnh trực quan như biểu đồ cột để minh họa các yếu tố có tác động rõ rệt đến quyết định nghỉ việc như sự phức tạp của quy trình tuyển dụng hay tính minh bạch trong đánh giá công việc.

Một khía cạnh thiết yếu khác trong bước Share là trực quan hóa dữ liệu. Sử dụng biểu đồ, bảng, sơ đồ hoặc dashboard là cách hiệu quả để minh họa xu hướng, sự phân bố, hoặc mối quan hệ giữa các biến một cách trực quan, dễ tiếp thu. Tuy nhiên, điều quan trọng là phải chọn đúng hình thức biểu diễn. Thêm vào đó, nhà phân tích cần dự đoán trước những câu hỏi phản biện hoặc giải thích những thắc mắc mà người nhận kết quả có thể đặt ra. Việc chuẩn bị tốt trước buổi trình bày sẽ giúp xây dựng lòng tin và tăng độ thuyết phục cho phân tích. Điều này đặc biệt quan trọng khi người trình bày đưa ra những phát hiện "khó tin" hoặc mâu thuẫn với quan điểm ban đầu của doanh nghiệp. Trong trường hợp đó, cần chứng minh nguồn dữ liệu, quy trình xử lý và phương pháp phân tích một cách minh bạch để bảo vệ lập luận của mình.

1.6. Act (Hành động)

Bước cuối cùng trong quy trình phân tích dữ liệu là hành động. Đây là điểm đến cuối cùng của cả hành trình, nơi biến dữ liệu vốn chỉ là những con số chuyển hóa thành hành động thực tế có thể tạo ra thay đổi, cải thiện quy trình và mang lại giá trị kinh doanh rõ ràng. Trong nhiều trường hợp, đây là bước quan trọng nhất về mặt chiến lược, bởi nếu doanh nghiệp không hành động dựa trên kết quả phân tích thì toàn bộ quá trình trước đó đều là vô nghĩa.

Hành động trong phân tích dữ liệu không đơn giản là "làm gì đó", mà là triển khai những hành động cụ thể, dựa trên bằng chứng, có mục tiêu rõ ràng và đo lường được. Ví dụ sau khi nhóm phân tích đã phát hiện mối liên hệ giữa quy trình tuyển dụng phức tạp với tỷ lệ nghỉ việc cao, và nhận thấy quy trình đánh giá minh bạch giúp giữ chân nhân viên, họ đã đề xuất một loạt hành động cụ thể. Doanh nghiệp có thể quyết định chuẩn hóa quy trình tuyển dụng và đánh giá nhân sự, áp dụng các phương pháp hiệu quả đã được chứng minh. Đồng thời triển khai khảo sát nhân viên định kỳ hàng năm để tiếp tục theo dõi hiệu quả của các thay đổi.

Bước Act đòi hỏi sự phối hợp chặt chẽ giữa nhóm phân tích và các nhà quản lý ra quyết định. Nhóm phân tích không thể tự mình thực hiện thay đổi mà cần có sự đồng ý của người có thẩm quyền trong doanh nghiệp. đề xuất phương pháp A/B testing để kiểm chứng hiệu quả thay đổi. Một yếu tố quan trọng khác trong bước này là khả năng lặp lại và thích ứng. Thực tế cho thấy không phải hành động nào cũng mang lại kết quả như kỳ vọng, và không phải thay đổi nào cũng phù hợp với mọi bối cảnh. Do đó, quá trình triển khai cần đi kèm với theo dõi, đánh giá định kỳ và sẵn sàng điều chỉnh nếu cần. Ví dụ, nếu sau một năm triển khai thay đổi quy trình tuyển dụng, tỷ lệ nghỉ việc vẫn cao ở một số bộ phận, điều đó có thể cho thấy cần thêm dữ liệu cụ thể hơn, hoặc hành động được thiết kế chưa đúng trọng tâm. Đây là lúc quy trình quay về bước Ask để đặt lại câu hỏi, tạo thành vòng lặp cải tiến liên tục.

Cuối cùng, bước Act cho thấy tác động thực sự của phân tích dữ liệu đối với tổ chức. Việc sử dụng dữ liệu để đưa ra quyết định không chỉ mang lại hiệu quả tức thời, mà còn tạo dựng một văn hóa dữ liệu nơi mọi quyết định quan trọng đều được hỗ trợ bởi dữ liệu, chứ không dựa vào cảm tính hoặc kinh nghiệm cá nhân đơn lẻ. Khi tổ chức bắt đầu nhận ra rằng hành động dựa trên dữ liệu giúp họ cải thiện hiệu suất, giữ chân nhân viên, tăng doanh thu hoặc giảm chi phí, họ sẽ càng đầu tư nghiêm túc hơn vào phân tích dữ liệu như một năng lực cốt lõi.

Tóm lại, bước Act vừa là đích đến, nhưng đồng thời cũng là bước đầu tiên cho một chu kỳ phân tích mới. Đây là nơi dữ liệu tạo ra giá trị thật sự, là lúc tổ chức biến dữ liệu thành hành động và sau đó hành động lại tạo ra dữ liệu mới cho chu kỳ kế tiếp.

2. Tiền xử lý và làm sạch dữ liệu

2.1. Giới thiệu chung về tiền xử lý và làm sạch dữ liệu

Trong bối cảnh chuyển đổi số toàn cầu, dữ liệu đã và đang trở thành một tài sản chiến lược đối với các tổ chức, doanh nghiệp, đặc biệt trong lĩnh vực phân tích hành vi khách hàng. Tuy nhiên, dữ liệu thu thập được từ thực tiễn thường không ở trạng thái sẵn sàng cho mục đích phân tích mà tồn tại nhiều vấn đề như thiếu giá trị, nhiễu, sai lệch định dạng, hoặc chứa các ngoại lệ không hợp lý. Do đó, tiền xử lý và làm sạch dữ liệu (Data Preprocessing and Cleaning) là một bước thiết yếu, đóng vai trò nền tảng nhằm đảm bảo độ tin cậy và chính xác của các kết quả phân tích sau này.

Tiền xử lý dữ liệu bao gồm một chuỗi các thao tác nhằm chuẩn hóa và cấu trúc lại dữ liệu sao cho phù hợp với yêu cầu của mô hình phân tích hoặc mô hình học máy. Các bước phổ biến bao gồm: khám phá dữ liệu (Data Exploration), xử lý dữ liệu thiếu (Missing Values) và xử lý ngoại lệ và nhiễu (Outliers and Noise).

Làm sạch dữ liệu là một khâu trong tiến trình tiền xử lý, tập trung vào việc loại bỏ hoặc hiệu chỉnh các lỗi dữ liệu phát sinh trong quá trình thu thập và lưu trữ. Một tập dữ liệu không được làm sạch có thể dẫn đến những kết luận sai lệch, ảnh hưởng nghiêm trọng đến quá trình ra quyết định trong kinh doanh. Các bước phổ biến bao gồm: xử lý dữ liệu không hợp lệ (Invalid Data), xử lý dữ liệu trùng lặp (Duplicates Values)

Cuối cùng, sau quá trình tiền xử lý và làm sạch dữ liệu, dữ liệu sẽ được biến đổi và chuẩn hóa (Data Transformation & Scaling) thành bộ dữ liệu tối ưu, phục vụ quá trình phân tích ở các bước tiếp theo.

Trong phạm vi đồ án này, dữ liệu khách hàng được sử dụng nhằm phục vụ cho các mục tiêu phân tích hành vi tiêu dùng và phân khúc thị trường. Toàn bộ quy trình tiền xử lý và làm sạch được thực hiện bằng ngôn ngữ lập trình Python, với sự hỗ trợ của các thư viện chuyên dụng như pandas, numpy, sklearn, matplotlib và seaborn.

Thông qua quá trình tiền xử lý và làm sạch dữ liệu, bộ dữ liệu khách hàng ban đầu, hiện đang tồn tại nhiều vấn đề kỹ thuật sẽ được chuyển đổi thành một tập dữ liệu chất lượng, phù hợp cho các phân tích thống kê mô tả. Đây chính là tiền đề để đảm bảo các kết quả phân tích mang tính khách quan, chính xác và có giá trị thực tiễn trong nghiệp vụ.

2.2. Quy trình tiền xử lý và làm sạch dữ liệu

Tiền xử lý và làm sạch dữ liệu là một bước quan trọng trong chuỗi quy trình phân tích dữ liệu, nhằm đảm bảo tính chính xác, toàn vẹn và khả năng khai thác hiệu quả của

tập dữ liệu đầu vào. Dù dữ liệu có thể đến từ nhiều nguồn khác nhau, các bước tiền xử lý và làm sạch thường tuân theo một tiến trình có hệ thống bao gồm:

- Khám phá dữ liệu (Data Exploration) nhằm có cái nhìn tổng quát về dữ liệu. Việc hiểu cấu trúc dữ liệu ở giai đoạn này giúp định hướng các thao tác xử lý phù hợp trong các bước tiếp theo. Giai đoạn này bao gồm:

+ Kiểm tra cấu trúc dữ liệu (số lượng dòng, cột, kiểu dữ liệu).

+ Tính toán thống kê mô tả (mean, median, mode, min, max, std...).

+ Sử dụng các công cụ trực quan hóa như biểu đồ histogram, boxplot, heatmap,... để phát hiện xu hướng, phân phối và bất thường trong dữ liệu.

- Xử lý dữ liệu thiếu (Missing Values) trong trường hợp dữ liệu thực tế xuất hiện các giá trị bị thiếu do nhiều nguyên nhân như lỗi nhập liệu, hạn chế hệ thống, hoặc không phản hồi từ người dùng. Các phương pháp xử lý bao gồm:

+ Loại bỏ dòng hoặc cột chứa nhiều giá trị thiếu trong trường hợp tỉ lệ không đáng kể so với tổng thể.

+ Điền giá trị bằng trung bình (mean), trung vị (median), mode hoặc giá trị suy đoán.

+ Gắn cờ (flagging) những trường hợp thiếu để xử lý riêng biệt trong phân tích sau này.

- Xử lý ngoại lệ và nhiễu (Outliers and Noise):

Ngoại lệ là những giá trị nằm ngoài phạm vi kỳ vọng của dữ liệu, có thể ảnh hưởng đáng kể đến kết quả phân tích. Một số phương pháp phổ biến để phát hiện outlier:

+ Boxplot và quy tắc IQR (Interquartile Range).

+ Z-score.

+ Biểu đồ phân phối (distribution plot) hoặc scatter plot.

Việc xử lý có thể thực hiện bằng cách loại bỏ, thay thế hoặc sử dụng các kỹ thuật như Winsorization để làm giảm tác động của ngoại lệ.

Dữ liệu nhiễu có thể đến từ các giá trị sai định dạng, lỗi đánh máy, hoặc biến dạng trong quá trình thu thập. Một số kỹ thuật xử lý gồm:

+ Gộp nhóm giá trị không đồng nhất (ví dụ: “male”, “Male”, “MALE”).

+ Làm tròn dữ liệu (binning), hoặc làm mượt bằng các thuật toán trung bình di động (moving average).

- Xử lý dữ liệu không hợp lệ (Invalid Data):

Dữ liệu không hợp lệ là những giá trị không tuân thủ quy tắc logic, định dạng hoặc nghiệp vụ. Chúng có thể gây sai lệch nghiêm trọng nếu không được xử lý sớm. Một số ví dụ điển hình:

+ Tuổi âm, giới tính không thuộc tập {Nam, Nữ}.

+ Email không có ký tự "@".

+ Ngày giao dịch xảy ra trong tương lai,..

Một số phương pháp xử lý dữ liệu không hợp lệ bao gồm:

+ Loại bỏ các giá trị không hợp lệ trong trường hợp tỉ lệ không đáng kể so với tổng thể.

+ Chuyển đổi về giá trị mặc định hoặc hợp lệ gần nhất.

+ Áp dụng ràng buộc kiểu dữ liệu hoặc điều kiện logic để phát hiện lỗi.

- Xử lý dữ liệu trùng lặp (Duplicates Values): Sự trùng lặp trong dữ liệu có thể xảy ra do quá trình nhập liệu thủ công, tích hợp từ nhiều nguồn khác nhau, hoặc lỗi hệ thống. Điều này làm sai lệch thống kê và gây dư thừa thông tin. Quy trình xử lý dữ liệu trùng lặp bao gồm các bước:

+ Xác định các bản ghi trùng hoàn toàn hoặc trùng theo một số trường khóa chính.

+ Phân tích nguyên nhân trùng lặp để quyết định giữ lại hay loại bỏ.

+ Xóa các dòng trùng để đảm bảo tính duy nhất và chính xác của dữ liệu.

- Biến đổi và chuẩn hóa dữ liệu (Data Transformation & Scaling): Cuối cùng, sau khi dữ liệu đã được tiền xử lý và làm sạch, bước tiếp theo trong quy trình phân tích là biến đổi (transformation) và chuẩn hóa (scaling) dữ liệu, đồng thời lưu trữ dữ liệu đã được biến đổi và chuẩn hóa. Đây là giai đoạn có vai trò then chốt nhằm tạo điều kiện thuận lợi cho phân tích, trực quan hóa ở các bước tiếp theo.

3. Trực quan hóa dữ liệu

3.1. Khái niệm và các yếu tố quan trọng

Trực quan hóa dữ liệu (Data Visualization) là quá trình chuyển đổi dữ liệu dạng số, văn bản hoặc dạng bảng phức tạp thành các biểu diễn trực quan như biểu đồ, đồ thị, bản đồ, hoặc hình ảnh động. Mục tiêu chính là giúp người xem nhanh chóng nhận ra các mẫu, xu hướng và mối quan hệ trong dữ liệu với nhau, những điều thường bị khó có thể nhận ra nếu chỉ nhìn vào dữ liệu dạng thô. Trực quan hóa dữ liệu không đơn thuần là một bước trang trí trong báo cáo, mà là một công cụ phân tích giúp chuyển hóa dữ liệu thành thông tin có giá trị, dễ hiểu cho tất cả mọi người.

Về cấu trúc, để trực quan hóa dữ liệu hiệu quả thường bao gồm bốn yếu tố chính. Thứ nhất là về dữ liệu, yếu tố quan trọng nhất quyết định nội dung cần truyền tải. Dữ liệu cần đủ sạch, đầy đủ và mang tính đại diện. Thứ hai là phương thức trình bày, đây là thứ sẽ được thể hiện qua loại biểu đồ được lựa chọn và là yếu tố quyết định cách thức biểu diễn dữ liệu: biểu đồ đường, cột, tròn, scatter, heatmap,... Mỗi loại biểu đồ phù hợp với từng loại câu hỏi cụ thể: so sánh, phân loại, phân phối hay mối tương quan. Thứ ba là các yếu tố thị giác (visual encodings) như màu sắc, hình dạng, kích thước, trục tọa độ, bố cục, chuyển động,... Các yếu tố này cực kỳ quan trọng đối với người nghe vì nó có thể giúp truyền tải thông điệp nhanh chóng, rõ ràng tuy nhiên nếu sử dụng không phù hợp sẽ phản tác dụng và gây nhiễu thông tin. Cuối cùng là ngữ cảnh và câu chuyện (context & storytelling), dữ liệu chỉ có ý nghĩa khi được đặt trong bối cảnh phù hợp với người xem. Việc dẫn dắt người xem qua một quá trình logic, yêu cầu các thông tin được sắp xếp theo thứ tự dễ hiểu sẽ tăng hiệu quả truyền đạt lên đáng kể. Trực quan hóa đóng vai trò quan trọng trong việc khám phá quan hệ giữa các biến và trình bày phát hiện trong giai đoạn cuối cùng.

3.2. Các biểu đồ thường sử dụng

Một trong những đặc trưng nổi bật của trực quan hóa dữ liệu chính là sự đa dạng trong hình thức thể hiện. Tuy nhiên, không phải biểu đồ nào cũng phù hợp với mọi loại dữ liệu hay mục tiêu truyền tải. Mỗi loại biểu đồ sẽ thể hiện tốt ở một khía cạnh riêng và việc chọn đúng loại biểu đồ không chỉ giúp làm nổi bật thông tin quan trọng mà còn giúp người xem hiểu rõ vấn đề một cách nhanh chóng và trực quan hơn.

- Biểu đồ đường (Line chart) thường được dùng cho việc biểu diễn các chuỗi thời gian, như doanh số theo tháng, lượng người dùng theo tuần, hoặc nhiệt độ theo giờ. Nó thể hiện rõ xu hướng (trend), điểm tăng/giảm đột biến và tính chu kỳ của dữ liệu.

- Biểu đồ cột và thanh (Bar & Column chart) được dùng để so sánh giá trị giữa các danh mục rời rạc như quốc gia, sản phẩm, vùng miền,...

- Biểu đồ tròn (Pie chart) tuy gây tranh cãi vì khó đọc khi có nhiều nhóm nhỏ, nhưng vẫn được sử dụng trong các báo cáo để thể hiện tỷ lệ phần trăm khi nhóm mục không quá nhiều (dưới 5 – 6).

- Biểu đồ phân tán (Scatter plot) được sử dụng để kiểm tra mối tương quan giữa hai biến định lượng, chẳng hạn như mức chi tiêu và doanh thu.

- Boxplot (biểu đồ hộp) thể hiện trực quan các giá trị trung vị, tứ phân vị, và giá trị ngoại lệ.

Ngoài ra, còn có heatmap để biểu diễn mối quan hệ hai chiều với sự thay đổi màu sắc thể hiện mức độ (intensity), thường dùng trong phân tích hành vi người dùng. Gần đây, các loại biểu đồ tương tác (interactive dashboard) được phát triển trên Tableau, Power BI còn có thể giúp người xem thao tác dữ liệu theo thời gian thực, giúp người xem dễ dàng hình dung hơn về ý nghĩa của biểu đồ.

3.3. Trường hợp nên trực quan hóa dữ liệu

Việc thực hiện trực quan hóa dữ liệu không phải luôn luôn bắt buộc, mà phụ thuộc vào nhiều yếu tố như độ phức tạp của dữ liệu, mục tiêu phân tích, đối tượng tiếp nhận thông tin cũng như bối cảnh sử dụng. Đây là các trường hợp trực quan hóa dữ liệu là vô cùng cần thiết.

Thứ nhất, khi dữ liệu quá lớn hoặc quá phức tạp chẳng hạn như các bảng dữ liệu với hàng chục nghìn dòng, trực quan hóa giúp người xem nhanh chóng nắm bắt điểm chính thay vì lạc lối trong những con số.

Thứ hai, khi cần khám phá mối quan hệ giữa các biến trong giai đoạn phân tích khám phá dữ liệu, biểu đồ giúp xác định được xu hướng, cụm dữ liệu và giá trị ngoại lệ dễ dàng hơn so với việc chỉ sử dụng thống kê mô tả.

Cuối cùng, cũng là trường hợp quan trọng nhất, trực quan hóa là công cụ quan trọng trong giao tiếp và ra quyết định. Một bảng biểu phức tạp sẽ khó truyền đạt hiệu quả trong các buổi báo cáo để đưa ra chiến lược, nhưng một biểu đồ đơn giản trọng tâm về mặt nội dung về hình thức, dễ hiểu với tất cả mọi người dù là những người không chuyên về dữ liệu. Trong những trường hợp đó, hình ảnh trực quan là "ngôn ngữ chung" giúp truyền tải nội dung kỹ thuật một cách dễ hiểu, hiệu quả và nhanh chóng. Ví dụ một biểu đồ thể hiện rõ ROI tăng gấp 3 lần sau khi đổi chiến dịch sẽ dễ thuyết phục ban lãnh đạo hơn nhiều.

3.4. Tầm quan trọng

Trực quan hóa dữ liệu không chỉ là một kỹ thuật hỗ trợ, mà là một phần cốt lõi trong quá trình phân tích và ra quyết định. Trong bối cảnh lượng dữ liệu ngày càng tăng mạnh, việc tiếp cận và hiểu dữ liệu trở thành thách thức lớn. Trực quan hóa giúp hiểu rõ được insight của dữ liệu, giúp người xem không cần phải là những người chuyên về dữ liệu vẫn có thể hiểu được những gì đang xảy ra và hành động kịp thời.

Theo nghiên cứu từ Tableau và ThoughtSpot, con người xử lý hình ảnh nhanh gấp 60.000 lần so với văn bản. Khi dữ liệu được trình bày bằng biểu đồ đúng cách, não bộ dễ dàng nhận ra mô hình, tương quan, và biến động là những điều vốn rất khó khăn nếu nhìn vào các bảng số liệu khô khan. Trực quan hóa còn là công cụ quan trọng để kể chuyện với dữ liệu. Thay vì chỉ trình bày con số, việc tạo dựng câu chuyện với mở đầu tới cao trào rồi kết luận giúp người xem cảm thấy hứng thú, dễ ghi nhớ và đưa ra hành động phù hợp.

Ngoài ra, một biểu đồ được trình bày tốt còn có thể ngăn ngừa các hiểu lầm có thể dẫn đến những kết quả nghiêm trọng. Ví dụ, việc sắp xếp thứ tự trục, chọn tỷ lệ không hợp lý hoặc dùng màu gây nhiễu có thể khiến người xem rút ra kết luận sai lệch.

CHƯƠNG III: TỔNG QUAN VỀ BÀI TOÁN

1. Tổng quan về doanh nghiệp

Getz Group là một tập đoàn đa quốc gia có lịch sử hơn 170 năm hình thành và phát triển. Thành lập từ năm 1852 tại California, Mỹ, Getz Group hiện diện tại hơn 36 quốc gia với hơn 80 văn phòng và khoảng 21.500 nhân viên trên toàn cầu. Tập đoàn hoạt động trong nhiều lĩnh vực, từ phân phối thiết bị y tế, hóa chất công nghiệp, thực phẩm, đến marketing, logistics và công nghệ.

Tại thị trường Việt Nam, Getz Group hoạt động chủ yếu thông qua công ty thành viên Getz Bros. & Co. (Vietnam) – chuyên cung cấp các sản phẩm và dịch vụ trong ngành chăm sóc sức khỏe, thực phẩm và hóa chất. Ngoài ra, Getz Group còn mở rộng hoạt động sang lĩnh vực giải pháp công nghệ dành cho ngành F&B, đặc biệt thông qua nền tảng quản lý nhà hàng kỹ thuật số, giúp các doanh nghiệp tối ưu hóa quy trình bán hàng, đặt món, giao hàng và chăm sóc khách hàng.

2. Mô tả bài toán

Trong bối cảnh cạnh tranh khốc liệt của ngành bán lẻ hiện nay, Getz Retail Vietnam, một doanh nghiệp bán lẻ đa ngành hướng tới khách hàng trẻ và trung lưu đang phải đối mặt với bài toán tối ưu giá trị khách hàng. Dù sở hữu một lượng lớn khách hàng Regular (mua sắm không thường xuyên, giá trị đơn hàng thấp), doanh nghiệp nhận thấy nhóm này có tỷ lệ quay lại mua hàng thấp, mức độ tương tác hạn chế, và ít có dấu hiệu chuyển đổi thành khách hàng Premium vốn là những người mang lại lợi nhuận cao và bền vững hơn.

Dữ liệu nội bộ từ hệ thống CRM của Getz Retail Vietnam cho thấy rằng, khoảng 70% khách hàng hiện tại thuộc nhóm Regular, trong khi nhóm Premium chỉ chiếm hơn 30%. Tuy nhiên, mức độ đóng góp doanh thu lại phản ánh một xu hướng tương tự: Regular tạo ra khoảng 70% doanh thu toàn hệ thống, còn Premium mang về 30% còn lại. Điều này cho thấy doanh nghiệp vẫn đang phụ thuộc phần lớn vào nhóm khách hàng phổ thông vốn có xu hướng không ổn định và dễ bị ảnh hưởng bởi cạnh tranh về giá. Từ thực trạng đó, ban lãnh đạo công ty đặt ra yêu cầu bài toán cho đội ngũ phân tích là:

- Phân tích hành vi khách hàng: Làm rõ sự khác biệt giữa nhóm khách hàng Regular và Premium về mặt nhân khẩu học (tuổi, giới tính, thu nhập...), hành vi mua sắm (tần suất, giá trị đơn hàng, danh mục sản phẩm), mức độ tương tác và phản hồi.

- Xác định các yếu tố ảnh hưởng đến chuyển đổi: Tìm ra những biến hoặc tổ hợp biến có khả năng dự báo khả năng một khách hàng Regular sẽ trở thành Premium.

- Khám phá các phân khúc khách hàng tiềm năng: Ứng dụng các phương pháp phân cụm hoặc phân loại để xác định nhóm khách hàng có tiềm năng cao chuyển đổi, từ đó đề xuất các chương trình ưu đãi, khuyến mãi hoặc chăm sóc chuyên biệt.

Bài toán phân tích không chỉ dừng lại ở việc mô tả dữ liệu hiện tại, mà còn hướng đến mục tiêu hành động và ra quyết định, giúp doanh nghiệp điều chỉnh các chiến dịch marketing, cá nhân hóa trải nghiệm người dùng và thiết kế các chương trình chăm sóc khách hàng thiết thực và hiệu quả hơn.

Để giải quyết bài toán này, doanh nghiệp đã tổng hợp một bộ dữ liệu chi tiết từ hệ thống CRM và nền tảng bán hàng đa kênh, bao gồm hơn 30 biến liên quan đến đặc điểm nhân khẩu học, hành vi mua sắm, phản hồi, và thông tin logistics. Bộ dữ liệu cung cấp thông tin chi tiết về các giao dịch bán lẻ của doanh nghiệp, với mục tiêu mang đến một cái nhìn toàn diện về hành vi khách hàng và xu hướng mua sắm. Cho phép phân tích chi tiết về sở thích của khách hàng, hành vi mua sắm, xu hướng nhân khẩu học và mức độ hài lòng.

Bộ dữ liệu bao gồm 30 biến được chia thành 05 hạng mục bao gồm Customer Information, Transaction Details, Product Information, Feedback và Transaction Logistics.

- Customer Information:

+ Customer ID: Mã định danh duy nhất cho mỗi khách hàng.

+ Name: Họ và tên đầy đủ của khách hàng.

+ Email: Địa chỉ email của khách hàng để liên lạc.

+ Phone: Số điện thoại liên lạc của khách hàng.

+ Address: Địa chỉ thực tế của khách hàng.

+ City, State, Zipcode, Country: Thông tin địa lý của khách hàng.

+ Age: Tuổi của khách hàng.

+ Gender: Giới tính của khách hàng.

+ Income: Mức thu nhập hoặc mức độ thu nhập của khách hàng.

+ Customer Segment: Phân loại khách hàng dựa trên hành vi hoặc nhân khẩu học.

- Transaction Details:

+ Last Purchase Date: Ngày mua hàng gần đây nhất của khách hàng.

+ Total Purchases: Tổng số lần mua hàng của khách hàng.

+ Amount Spent: Tổng số tiền khách hàng đã chi tiêu.

- Product Information:

+ Product Category: Danh mục mà sản phẩm đã mua thuộc về (ví dụ: đồ điện tử, quần áo, hàng tạp hóa).

+ Product Brand: Tên thương hiệu của sản phẩm.

+ Product Type: Loại hoặc mẫu sản phẩm đã mua.

- Feedback: Phản hồi hoặc đánh giá của khách hàng liên quan đến sản phẩm hoặc dịch vụ nhận được.

- Transaction Logistics:

+ Shipping Method: Phương thức được sử dụng để giao sản phẩm đã mua.

+ Payment Method: Phương thức thanh toán do khách hàng lựa chọn.

+ Order Status: Trạng thái của đơn hàng (ví dụ: đã vận chuyển, đã giao, đã hủy).

Ở góc độ tổng thể, bộ dữ liệu không chỉ cung cấp cái nhìn toàn cảnh về khách hàng và hành vi mua sắm, mà còn có thể được liên kết chéo giữa các chiều dữ liệu để phát hiện các mẫu hành vi, xu hướng tiềm ẩn hoặc bất thường hỗ trợ ra quyết định chiến lược.

3. Giá trị thực tiễn

Phân tích hành vi khách hàng nhằm thúc đẩy chuyển đổi từ Regular sang Premium mang lại nhiều giá trị thiết thực cho doanh nghiệp, không chỉ dừng lại ở việc hiểu hiện trạng mà còn đóng vai trò chiến lược trong việc ra quyết định.

Thứ nhất, kết quả phân tích giúp doanh nghiệp:

- Hiểu rõ sự khác biệt giữa hai nhóm khách hàng chính về nhân khẩu học, hành vi mua sắm, phản hồi và tương tác.

- Xác định các yếu tố dự báo khả năng chuyển đổi, như: tần suất mua hàng, danh mục sản phẩm ưa thích, thu nhập, mức độ hài lòng,...

- Phát hiện các phân khúc khách hàng tiềm năng có khả năng chuyển đổi cao, làm nền tảng cho việc thiết kế các chiến dịch ưu đãi, chăm sóc chuyên biệt.

Thứ hai, bài toán này hướng tới hành động cụ thể: giúp bộ phận marketing cá nhân hóa thông điệp, lựa chọn đúng đối tượng mục tiêu; hỗ trợ đội ngũ CSKH xây dựng chính sách giữ chân phù hợp; đồng thời cung cấp cơ sở dữ liệu tin cậy để ban lãnh đạo điều chỉnh chiến lược kinh doanh theo hướng tăng trưởng bền vững, thay vì phụ thuộc vào số lượng khách hàng ngắn hạn.

Ngoài ra, đây là một bài toán mang tính thực tiễn khi có khả năng chuyển hóa dữ liệu thành giá trị kinh doanh, nâng cao khả năng cạnh tranh và đảm bảo sự phát triển dài hạn của Getz Retail Vietnam trong thị trường bán lẻ đầy biến động hiện nay.

CHƯƠNG IV: XỬ LÝ DỮ LIỆU

1. Các công cụ thực hiện

Dựa trên tính chất và quy mô của bộ dữ liệu, nhóm đã cân nhắc và quyết định sử dụng kết hợp hai công cụ chính là Microsoft Excel và ngôn ngữ Python. Tuy nhiên, với khối lượng dữ liệu lớn trong dự án và nhu cầu xử lý phức tạp như tiền xử lý dữ liệu, mã hóa các biến phân loại, phân tích thống kê và trực quan hóa, nhóm quyết định sử dụng Python làm công cụ chính xuyên suốt quá trình xử lý và phân tích. Trong đồ án này, nhóm chủ yếu sử dụng một số thư viện quen thuộc trong Python như Pandas để thao tác và làm sạch dữ liệu, Numpy để xử lý số liệu, Matplotlib và Seaborn để trực quan hóa các đặc điểm và mối quan hệ trong dữ liệu, giúp phát hiện những xu hướng hoặc bất thường. Bên cạnh đó, nhóm cũng sử dụng các thư viện thống kê như Scipy.stats để kiểm định sự khác biệt giữa các nhóm khách hàng dựa trên các biến số quan trọng. Việc sử dụng Python giúp nhóm tổ chức xử lý dữ liệu một cách khoa học hơn. Mỗi bước xử lý đều được chia thành từng cell và chú thích rõ ràng, nên rất dễ kiểm tra lại hoặc chạy lại nếu cần chỉnh sửa hay cập nhật. Điều này không chỉ tiết kiệm thời gian mà còn giúp giảm sai sót do thao tác thủ công, điều mà dễ gặp phải khi làm bằng Excel. Vì vậy, nhóm chỉ thực hiện sử dụng Excel để xem qua và kiểm tra tổng quát về bộ dữ liệu, còn toàn bộ phần xử lý và phân tích chính đều được nhóm thực hiện bằng Python để thể hiện một cách rõ ràng hơn.

2. Tiền xử lý và làm sạch dữ liệu

Khởi đầu tiến trình xử lý dữ liệu, nhóm sử dụng các thư viện matplotlib.pyplot, pandas, seaborn, scipy.stats để thực hiện các yêu cầu phân tích. Nhóm cũng tiến hành thiết lập các thông số kỹ thuật chung cho biểu đồ và dùng biến df để đọc thông tin từ tệp csv chứa dữ liệu phân tích.

```
#%% IMPORT LIBRARIES
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns
from scipy.stats import ttest_ind, chi2_contingency
#%% LOAD DATA
df = pd.read_csv('data/new_retail_data.csv')
#%% CONFIGS
```

```
plt.rcParams['font.size'] = 16
plt.rcParams['figure.dpi'] = 120
```

Tiếp theo, nhóm tiến hành đánh giá và loại bỏ các biến không liên quan đến biến mục tiêu cần phân tích.

```
#%% DELETE COLUMN THAT IS NOT RELEVANT
df.drop(columns=['Transaction_ID','Name','Email',
                 'Phone','Address','City','State','Zipcode',
                 'products','Product_Brand',
                 'Product_Type','Feedback','Total_Purchases',
                 'Amount','Date'], inplace=True)
print(df.head())
```

Bên cạnh đó, xác định yêu cầu bài toán là đưa ra chiến lược chuyển đổi khách hàng thông thường (Regular) thành khách hàng thân thiết (Premium), nhóm tiến hành loại bỏ các dòng chứa giá trị "New" thuộc biến Customer_Segment, tránh tình trạng gây loãng dữ liệu phân tích.

```
#%% FILTER: KEEP ONLY 'REGULAR' AND 'PREMIUM' IN CUSTOMER_SEGMENT
print("\n--- Lọc dữ liệu để tập trung vào giá trị 'Regular' và 'Premium' (Loại bỏ giá trị 'New')---")
segments_to_keep = ['Regular', 'Premium']
df = df[df['Customer_Segment'].isin(segments_to_keep)].copy()
print(f"Đã lọc dữ liệu, số dòng còn lại: {len(df)}")
print("Các phân tích từ đây sẽ thực hiện trên 2 nhóm 'Regular' và 'Premium'.\n")
```

Thông qua hàm shape() có thể xác định bộ dữ liệu phân tích hiện tại có 210608 dòng và 15 cột.

```
#%% VIEW DATA SHAPE
print('Shape of dataframe:', df.shape)
```

```
>>> print('Shape of dataframe:', df.shape)
Shape of dataframe: (210608, 15)
```

Hình 4.1: Kích thước tập dữ liệu

Với hàm info(), nhóm kiểm tra được kiểu dữ liệu của từng biến và các thông tin khác có liên quan, bao gồm số giá trị (số hàng) trong mỗi cột, dữ liệu có rỗng hay không và mức sử dụng bộ nhớ của tập dữ liệu. Có thể thấy, bộ dữ liệu bao gồm 15 cột, trong đó 10 cột có kiểu dữ liệu object và 05 cột kiểu dữ liệu float64. Trong 210608 dòng dữ liệu của bộ dữ liệu hiện tại, số lượng giá trị không rỗng ở mỗi biến khác nhau. Điều này chứng tỏ các cột dữ liệu đều đang tồn tại các giá trị rỗng cần phải xử lý.

```
#%% CHECK DATA INFO
df.info()
```

```
Index: 210608 entries, 0 to 302009
Data columns (total 15 columns):
 #   Column            Non-Null Count   Dtype  
---  ------            --------------   -----  
 0   Customer_ID       210395 non-null  float64
 1   Country           210426 non-null  object 
 2   Age               210485 non-null  float64
 3   Gender            210373 non-null  object 
 4   Income            210411 non-null  object 
 5   Customer_Segment  210608 non-null  object 
 6   Year              210373 non-null  float64
 7   Month             210428 non-null  object 
 8   Time              210362 non-null  object 
 9   Total_Amount      210352 non-null  float64
 10  Product_Category  210414 non-null  object 
 11  Shipping_Method   210371 non-null  object 
 12  Payment_Method    210406 non-null  object 
 13  Order_Status      210444 non-null  object 
 14  Ratings           210467 non-null  float64
dtypes: float64(5), object(10)
memory usage: 25.7+ MB
```

Hình 4.2: Thông tin dữ liệu.

Tiếp theo, nhóm tiến hành kiểm tra tính trùng lặp của dữ liệu bao gồm số dòng và số cột trùng lặp thông qua hàm duplicate(). Kết quả ghi nhận 46 dòng và 0 cột dữ liệu trùng

lặp. Nhận thấy tỉ lệ dữ liệu trùng lặp không đáng kể, không ảnh hưởng nghiêm trọng đến hoạt động xây dựng và đánh giá mô hình, nhóm thực hiện loại bỏ các dòng dữ liệu thừa thông qua hàm drop_duplicates().

```
#%% CHECK FOR DUPLICATES ROWS
num_duplicate_rows = df.duplicated().sum()
print('Number of duplicated rows in dataset:', num_duplicate_rows)

#%% REMOVE DUPLICATES ROW
df = df.drop_duplicates()
```

```
Number of duplicated rows in dataset: 46
>>> df = df.drop_duplicates()
```

Hình 4.3: Kiểm tra các dòng dữ liệu trùng lặp và loại bỏ trùng lặp.

Bên cạnh đó, nhằm đánh giá sự đa dạng dữ liệu, nhóm dùng hàm unique() để xác định có bao nhiêu giá trị duy nhất (unique values) trong mỗi cột. Kết quả ghi nhận phản ánh đúng tính chất các thuộc tính khi các biến phân loại đều có từ 2-5 giá trị. Trong khi các biến số khác ghi nhận đa dạng giá trị hơn. Biến Ratings được xác định với kiểu dữ liệu Float64, tuy nhiên dựa vào tính chất biến, nhóm xác định Ratings là biến phân loại với 05 giá trị tương ứng mức đánh giá từ 1 đến 5.

```
#%% CHECK FOR DUPLICATES VALUE
unique_counts = df.nunique()
print(unique_counts)
```

```
Customer_ID          81198
Country                  5
Age                     53
Gender                   2
Income                   3
Customer_Segment         2
Year                     2
Month                   12
Time                 78762
Total_Amount        208926
Product_Category         5
Shipping_Method          3
Payment_Method           4
Order_Status             4
Ratings                  5
dtype: int64
```

Hình 4.4: Đánh giá tính đa dạng dữ liệu.

Nhóm sử dụng hàm isnull() và sum() để tính tổng số giá trị null và tỉ lệ giá trị null có trong mỗi cột dữ liệu. Kết quả của isnull().sum() cho thấy, ngoại trừ biến Customer_Segment, tất cả các biến còn lại thuộc bộ dữ liệu đều có giá trị null với tỉ lệ dao động từ 0.06% đến 0.12%

```
#%% CHECK FOR MISSING VALUES
missing_values = df.isnull().sum()
missing_percent = (missing_values / len(df)) * 100
missing_data = pd.DataFrame({
  'Missing Values': missing_values,
  'Percentage (%)': missing_percent})
print(missing_data)
print(missing_values.sum())
```

	Missing Values	Percentage (%)
Customer_ID	213	0.101158
Country	182	0.086435
Age	123	0.058415
Gender	235	0.111606
Income	197	0.093559
Customer_Segment	0	0.000000
Year	235	0.111606
Month	180	0.085486
Time	246	0.116830
Total_Amount	256	0.121579
Product_Category	194	0.092134
Shipping_Method	236	0.112081
Payment_Method	202	0.095934
Order_Status	164	0.077887
Ratings	141	0.066964
2804		

Hình 4.5: Kết quả kiểm tra giá trị null.

Nhận thấy tỉ lệ giá trị null không đáng kể, không gây ảnh hưởng nghiêm trọng đến phân tích tổng thể, nhóm tiến hành loại bỏ các dòng chứa giá trị null ở từng biến cụ thể, đảm bảo các giá trị phân tích đều thể hiện đầy đủ thông tin dữ liệu. Sau khi loại bỏ các giá trị null, dữ liệu tại ghi nhận 207766 dòng dữ liệu.

```
#%% REMOVE ROWS WITH MISSING VALUES
df = df.dropna()
print(f"Số dòng sau khi xóa các giá trị thiếu: {len(df)}")
```

```
... print(f"Số dòng sau khi xóa các giá trị thiếu: {len(df)}")
...
Số dòng sau khi xóa các giá trị thiếu: 207766
```

Hình 4.6: Kết quả sau khi xử lý giá trị null.

Đối với các biến Month và Year chứa thông tin thời gian giao dịch cụ thể của khách hàng, nhận thấy biến Year có 02 giá trị là 2023, 2024 và biến Month có 12 giá trị (tương ứng 12 tháng trong năm), nhóm thực hiện phân tách thành 24 giá trị phân loại với cấu

trúc Month-Year và xác lập biến Month_Year mới, đồng thời loại bỏ các biến Month, Year ban đầu.

```
#%% CONVERT MONTH YEAR DATA
df['Month_Year'] = df['Month'].astype(str) + '-' + df['Year'].astype(str)
months = [
  'January', 'February', 'March', 'April', 'May', 'June',
  'July', 'August', 'September', 'October', 'November', 'December']
years = ['2023.0', '2024.0']
month_year_order = [f"{month}-{year}" for year in years for month in months]
df['Month_Year'] = pd.Categorical(df['Month_Year'],categories=month_year_order,
                                  ordered=True)
df.drop(columns=['Month', 'Year'], inplace=True)
```

Tương tự, đối với biến Time chứa thông tin về khung giờ giao dịch cụ thể của khách hàng, nhòm cũng thực hiện phân tách thành 24 giá trị phân loại tương ứng với 24 khung giờ trong ngày từ 0 giờ đến 24 giờ, xác lập biến Hour_Group mới, đồng thời loại bỏ biến Time ban đầu.

```
#%% CONVERT TIME DATA
df["Time_dt"] = pd.to_datetime(df["Time"], format="%H:%M:%S")
df["Hour"] = df["Time_dt"].dt.hour
df["Minute"] = df["Time_dt"].dt.minute
df["Hour_Group"] = df.apply(
  lambda row: (row["Hour"] if row["Minute"] < 30 else (row["Hour"] + 1) % 24),
  axis=1)
df.drop(columns=['Time','Time_dt', 'Hour', 'Minute'], inplace=True)
df["Hour_Group"] = pd.Categorical(df["Hour_Group"], categories=range(24),
ordered=True)
```

3. Các yếu tố quan trọng ảnh hưởng đến Customer_Segment

Với mục tiêu phân tích tương quan giữa các biến thành phần và biến mục tiêu là Customer_Segment, nhóm thực hiện các kiểm định T-test (Đối với biến số) và Chi-square (Đối với biến phân loại) nhằm xác định các biến có ý nghĩa phân tích trong tương quan với biến mục tiêu, đồng thời làm cơ sở chọn các giá trị có tác động quan trọng đến Customer_Segment.

Đối với biến số, kết quả kiểm định cho thấy biến Age có sự khác biệt ý nghĩa giữa 02 nhóm Regular và Premium thuộc biến mục tiêu. Song biến Total_Amount, chứa thông tin về tổng giá trị giao dịch lại không có sự khác biệt ý nghĩa giữa 02 nhóm trên.

```
#%% ANALYZE CORRELATION USING T-TEST BETWEEN CONTINUOUS VARIABLES AND DEPENDENT VARIABLE (CUSTOMER_SEGMENT)
numerical_vars = ['Age', 'Total_Amount']
print("\n--- Phân tích tương quan T-test cho các biến số ---")
for var in numerical_vars:
    try:
        group1 = df[df['Customer_Segment'] == 'Regular'][var]
        group2 = df[df['Customer_Segment'] == 'Premium'][var]
        t_stat, p_val = ttest_ind(group1, group2, equal_var=False)
        print(f"{var}: t = {t_stat:.4f}, p = {p_val:.4f}")
        if p_val < 0.05:
            print(f"   → Biến '{var}' có khác biệt có ý nghĩa giữa 2 nhóm.")
        else:
            print(f"   → Biến '{var}' KHÔNG có khác biệt rõ rệt giữa 2 nhóm.")
    except Exception as e:
        print(f"{var}: lỗi khi tính T-test: {e}")
```

```
--- Phân tích tương quan T-test cho các biến số ---
Age: t = -134.3503, p = 0.0000
   → Biến 'Age' có khác biệt có ý nghĩa giữa 2 nhóm.
Total_Amount: t = 0.9461, p = 0.3441
   → Biến 'Total_Amount' KHÔNG có khác biệt rõ rệt giữa 2 nhóm.
```

Hình 4.7: Kết quả kiểm định T-test.

Đối với biến phân loại, kết quả kiểm định cho thấy ngoại trừ biến Time không có sự khác biệt ý nghĩa giữa 02 nhóm Regular và Premium thuộc biến mục tiêu, các biến còn lại đều có sự khác biệt ý nghĩa giữa 02 nhóm giá trị cần phân tích để thực hiện mục tiêu chuyển đổi.

```
#%% ANALYZE CORRELATION USING CHI-SQUARE TEST BETWEEN CATEGORICAL VARIABLES AND DEPENDENT VARIABLE (CUSTOMER_SEGMENT)
categorical_vars = ['Gender','Income','Country','Product_Category',
                    'Shipping_Method','Payment_Method','Order_Status',
                    'Ratings','Hour_Group', 'Month_Year']
print("\n--- Phân tích Chi-square cho các biến phân loại ---")
for var in categorical_vars:
    try:
        ct = pd.crosstab(df[var], df['Customer_Segment'])
        if ct.shape[1] != 2:
            print(f"{var}: bỏ qua (không phải phân loại nhị phân)")
            continue
        chi2, p, dof, expected = chi2_contingency(ct)
        print(f"{var}: chi2 = {chi2:.4f}, p = {p:.4f}")
        if p < 0.05:
            print(f"   → Biến '{var}' có liên hệ với nhóm Customer_Segment.")
        else:
            print(f"   → Biến '{var}' KHÔNG có liên hệ rõ rệt.")
    except Exception as e:
        print(f"{var}: lỗi khi chạy Chi-square: {e}")
```

```
--- Phân tích Chi-square cho các biến phân loại ---
Gender: chi2 = 132.5698, p = 0.0000
   → Biến 'Gender' có liên hệ với nhóm Customer_Segment.
Income: chi2 = 5036.1010, p = 0.0000
   → Biến 'Income' có liên hệ với nhóm Customer_Segment.
Country: chi2 = 2630.1790, p = 0.0000
   → Biến 'Country' có liên hệ với nhóm Customer_Segment.
Product_Category: chi2 = 222.7648, p = 0.0000
   → Biến 'Product_Category' có liên hệ với nhóm Customer_Segment.
Shipping_Method: chi2 = 36.4048, p = 0.0000
   → Biến 'Shipping_Method' có liên hệ với nhóm Customer_Segment.
Payment_Method: chi2 = 609.5751, p = 0.0000
   → Biến 'Payment_Method' có liên hệ với nhóm Customer_Segment.
Order_Status: chi2 = 6428.6705, p = 0.0000
   → Biến 'Order_Status' có liên hệ với nhóm Customer_Segment.
Ratings: chi2 = 3189.8195, p = 0.0000
   → Biến 'Ratings' có liên hệ với nhóm Customer_Segment.
Hour_Group: chi2 = 23.6580, p = 0.4229
   → Biến 'Hour_Group' KHÔNG có liên hệ rõ rệt.
Month_Year: chi2 = 9650.0549, p = 0.0000
   → Biến 'Month_Year' có liên hệ với nhóm Customer_Segment.
```

Hình 4.8: Kết quả kiểm định Chi-square

CHƯƠNG V: TRỰC QUAN HÓA VÀ INSIGHT CỦA DỮ LIỆU

1. Python

Đối với các biến phân loại có ý nghĩa, sau khi thực hiện các kiểm định tương quan, nhóm xác định ngưỡng khác biệt chênh lệch là 40%. Điều này có nghĩa, nếu sự chênh lệch tỷ lệ giữa 02 giá trị Regular và Premium bằng hoặc vượt quá ngưỡng này, giá trị của biến được đánh giá có ảnh hưởng lớn đến phân khúc khách hàng. Hoạt động này góp phần xác định các đặc điểm phân biệt rõ ràng giữa hai nhóm khách hàng, từ đó hỗ trợ chiến lược phân khúc và cá nhân hóa tiếp thị.

```
#%% ANALYZE PROPORTION DIFFERENCE BETWEEN 'REGULAR' AND 'PREMIUM' IN CATEGORICAL VARIABLES (DIFFERENCE >= 40%)
print("\n--- Phân tích tỷ lệ Regular vs Premium theo từng biến phân loại ---")
important_cats = ['Gender','Income','Country','Product_Category',
                  'Shipping_Method','Payment_Method','Order_Status','Ratings']
diff_threshold = 0.4
for var in important_cats:
    try:
        ct = pd.crosstab(df[var], df['Customer_Segment'], normalize='index')
        ct['Diff'] = abs(ct['Premium'] - ct['Regular'])
        sig_diff = ct[ct['Diff'] >= diff_threshold].sort_values('Diff', ascending=False)
        if not sig_diff.empty:
            print(f"\n>>> Biến '{var}' có giá trị chênh lệch lớn (>|{diff_threshold*100:.0f}%|):")
            print(sig_diff)
            sig_diff[['Premium', 'Regular']].plot(kind='barh', figsize=(8, 4), color=['#1f77b4', '#ff7f0e'])
            plt.title(f'Tỷ lệ Premium vs Regular theo '{var}' (chênh lệch > {diff_threshold*100:.0f}%)")
            plt.xlabel("Tỷ lệ (%)")
            plt.grid(axis='x', linestyle='--', alpha=0.7)
            plt.tight_layout()
            plt.show()
    except Exception as e:
```

```
        print(f"{var}: lỗi khi phân tích tỷ lệ: {e}")
```

Kết quả phân tích tỉ lệ 02 giá trị Regular và Premium theo từng biến phân loại có ý nghĩa cho thấy:

- Đối với biến Gender (Giới tính): Chỉ có giá trị Male (Nam) có sự chênh lệch tỉ lệ theo ngưỡng xác định ban đầu. Tỉ lệ khác biệt chênh lệch ghi nhận là 40.72% trong đó tỉ lệ Premium và Regular lần lượt là 29.64% và 70.36%.

- Đối với biến Income (Thu nhập): Các giá trị High (Thu nhập cao) và Medium (Thu nhập trung bình) có sự chênh lệch tỉ lệ theo ngưỡng xác định ban đầu. Tỉ lệ khác biệt chênh lệch ghi nhận lần lượt là 56.45% và 41.44%.

- Đối với biến Country (Khu vực địa lý): Các giá trị UK (Vương quốc Anh) và US (Mỹ) có sự chênh lệch tỉ lệ theo ngưỡng xác định ban đầu. Tỉ lệ khác biệt chênh lệch ghi nhận lần lượt là 51.23% và 44.1%.

- Đối với biến Product_Category (Phân loại sản phẩm): Các giá trị Electronics (Đồ điện tử), Home Decor (Đồ trang trí), Books (Sách) và Clothing (Thời trang) có sự chênh lệch tỉ lệ theo ngưỡng xác định ban đầu. Tỉ lệ khác biệt chênh lệch ghi nhận lần lượt là 40.63%, 40.59%, 40.49% và 40.35%.

- Đối với biến Shipping_Method (Phương thức vận chuyển): Chỉ có giá trị Standard (Tiêu chuẩn) có sự chênh lệch tỉ lệ theo ngưỡng xác định ban đầu. Tỉ lệ khác biệt chênh lệch ghi nhận là 40.7% trong đó tỉ lệ Premium và Regular lần lượt là 29.65% và 70.35%.

- Đối với biến Payment_Method (Phương thức thanh toán): Chỉ có giá trị PayPal (Thanh toán bằng PayPal) có sự chênh lệch tỉ lệ theo ngưỡng xác định ban đầu. Tỉ lệ khác biệt chênh lệch ghi nhận là 47.12% trong đó tỉ lệ Premium và Regular lần lượt là 26.44% và 73.56%.

- Đối với biến Order_Status (Trạng thái đơn hàng): Chỉ có giá trị Delivered (Đã vận chuyển) có sự chênh lệch tỉ lệ theo ngưỡng xác định ban đầu. Tỉ lệ khác biệt chênh lệch ghi nhận là 56.18% trong đó tỉ lệ Premium và Regular lần lượt là 21.91% và 78.09%.

- Đối với biến Ratings (Đánh giá): Các giá trị 1.0 và 2.0 có sự chênh lệch tỉ lệ theo ngưỡng xác định ban đầu. Tỉ lệ khác biệt chênh lệch ghi nhận lần lượt là 60.5% và 47.75%.

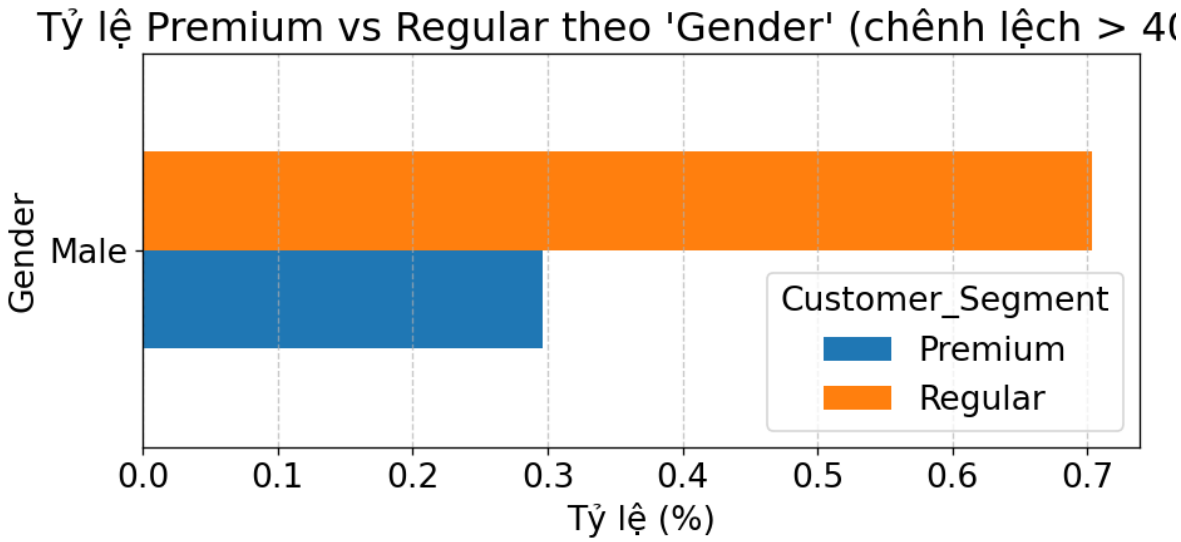
```
--- Phân tích tỷ lệ Regular vs Premium theo từng biến phân loại ---

>>> Biến 'Gender' có giá trị chênh lệch lớn (>|40%|):
Customer_Segment   Premium   Regular      Diff
Gender
Male              0.296399  0.703601  0.407203

>>> Biến 'Income' có giá trị chênh lệch lớn (>|40%|):
Customer_Segment   Premium   Regular      Diff
Income
High              0.217744  0.782256  0.564511
Medium            0.292814  0.707186  0.414371

>>> Biến 'Country' có giá trị chênh lệch lớn (>|40%|):
Customer_Segment   Premium   Regular      Diff
Country
UK                0.243855  0.756145  0.512289
USA               0.279508  0.720492  0.440985

>>> Biến 'Product_Category' có giá trị chênh lệch lớn (>|40%|):
Customer_Segment   Premium   Regular      Diff
Product_Category
Electronics       0.296831  0.703169  0.406339
Home Decor        0.297036  0.702964  0.405928
Books             0.297539  0.702461  0.404922
Clothing          0.298258  0.701742  0.403484
```

```
>>> Biến 'Shipping_Method' có giá trị chênh lệch lớn (>|40%|):
Customer_Segment   Premium   Regular      Diff
Shipping_Method
Standard          0.296519  0.703481  0.406963

>>> Biến 'Payment_Method' có giá trị chênh lệch lớn (>|40%|):
Customer_Segment   Premium   Regular      Diff
Payment_Method
PayPal            0.264388  0.735612  0.471224

>>> Biến 'Order_Status' có giá trị chênh lệch lớn (>|40%|):
Customer_Segment  Premium  Regular    Diff
Order_Status
Delivered          0.2191   0.7809  0.5618

>>> Biến 'Ratings' có giá trị chênh lệch lớn (>|40%|):
Customer_Segment   Premium   Regular      Diff
Ratings
1.0               0.197452  0.802548  0.605095
2.0               0.261230  0.738770  0.477540
```

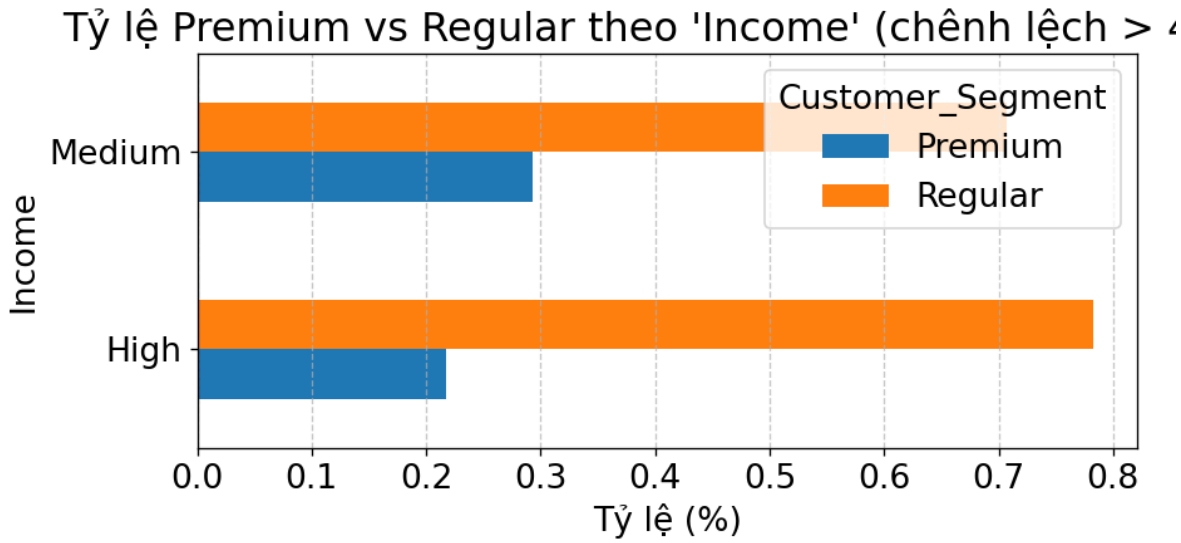
Hình 5.1: Kết quả phân tích tỉ lệ chênh lệch theo 2 giá trị Premium và Regular

Dựa vào các kết quả phân tích trên, nhóm tiến hành trực quan hóa các giá trị của từng biến thành phần có ngưỡng khác biệt chênh lệch mang ý nghĩa phân tích (>= 40%) theo 02 giá trị Premium và Regular của biến mục tiêu.

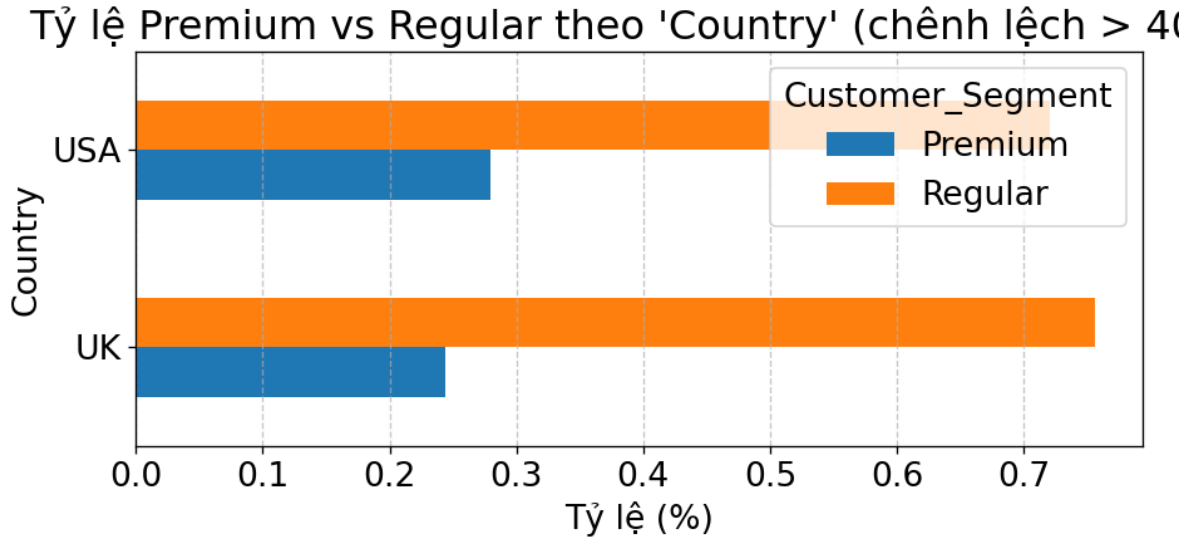
Xem xét tổng quan, có thể nhận định, tỉ lệ khách hàng thông thường của doanh nghiệp (Regular) theo các yếu tố phân tích đang dao động ở mức 70% - 80%. Điều này đặt ra yêu cầu cho doanh nghiệp về việc gia tăng tỉ lệ khách hàng ở phân khúc Premium bằng hoạt động đẩy nhanh tiến trình chuyển đổi phân khúc khách hàng. Để hiện thực hóa mục tiêu này, doanh nghiệp cần dựa vào các yếu tố có sự khác biệt chênh lệch đã được xác định và xây dựng chiến lược chuyển đổi phù hợp.



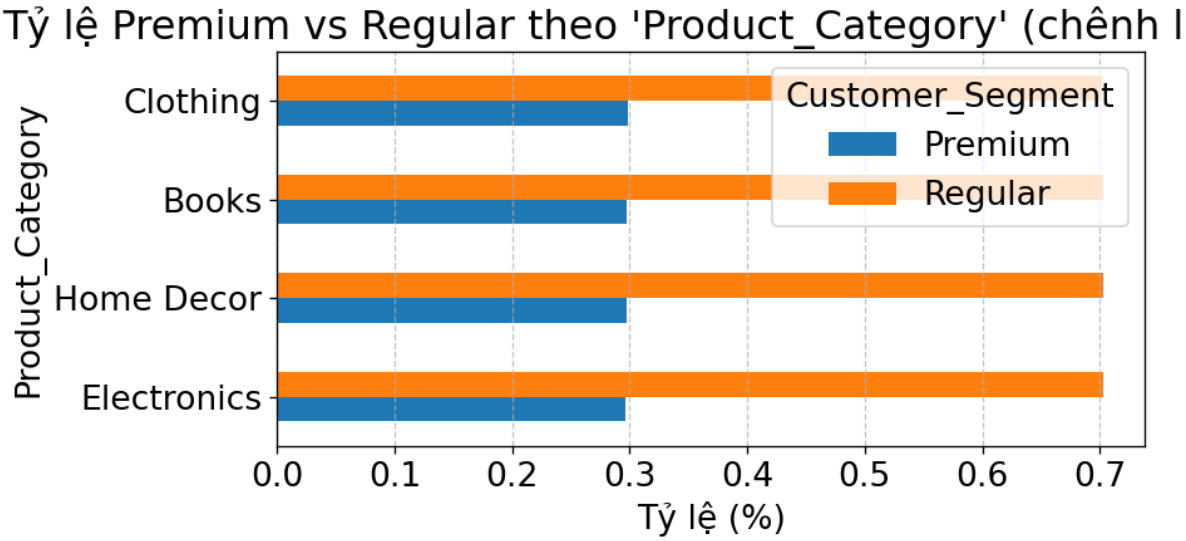
Hình 5.2: Trực quan hóa khác biệt chênh lệch giữa Premium và Regular (theo biến Male).



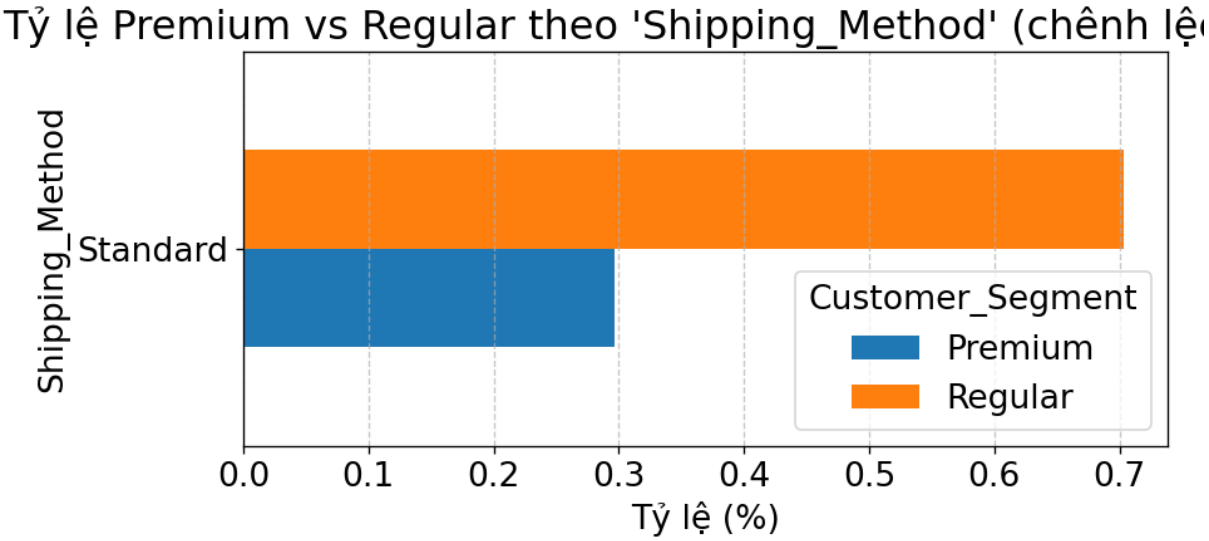
Hình 5.3: Trực quan hóa khác biệt chênh lệch giữa Premium và Regular (theo biến Income).



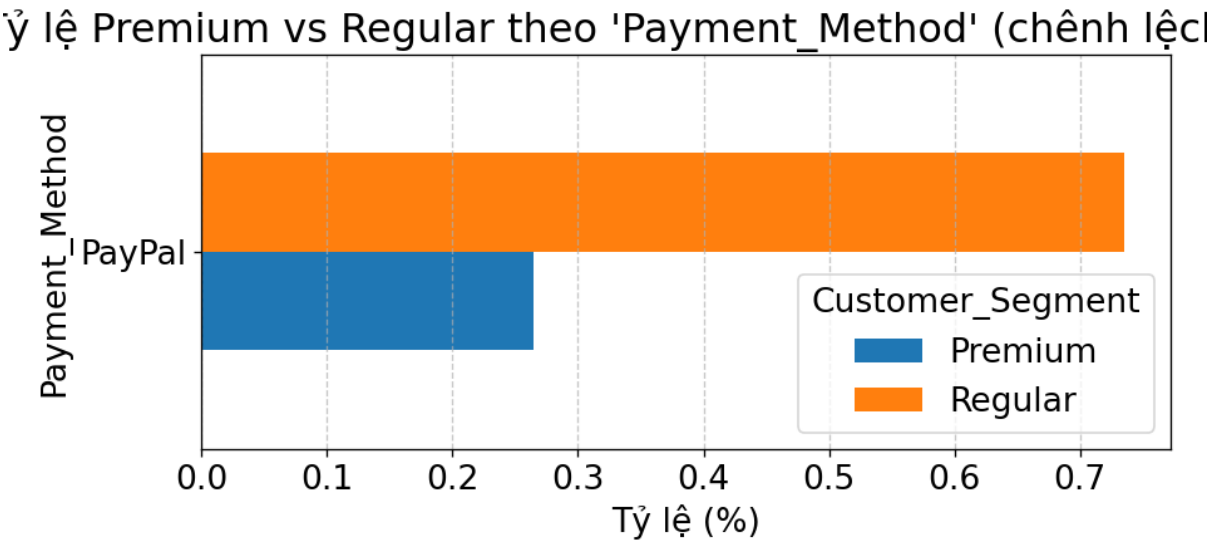
Hình 5.4: Trực quan hóa khác biệt chênh lệch giữa Premium và Regular (theo biến Country)



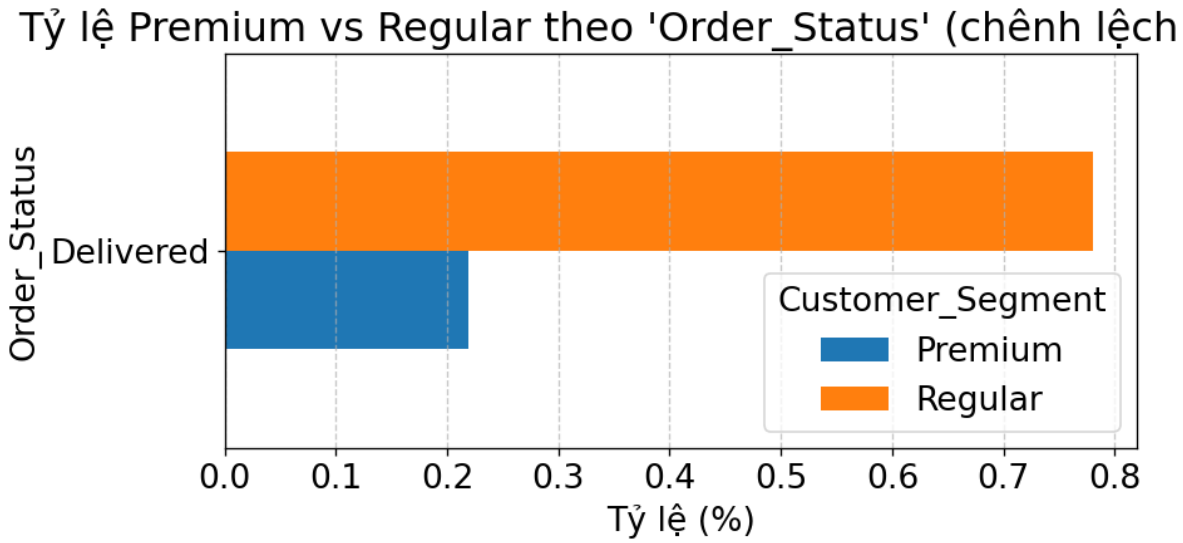
Hình 5.5: Trực quan hóa khác biệt chênh lệch giữa Premium và Regular (theo biến Product_Category)



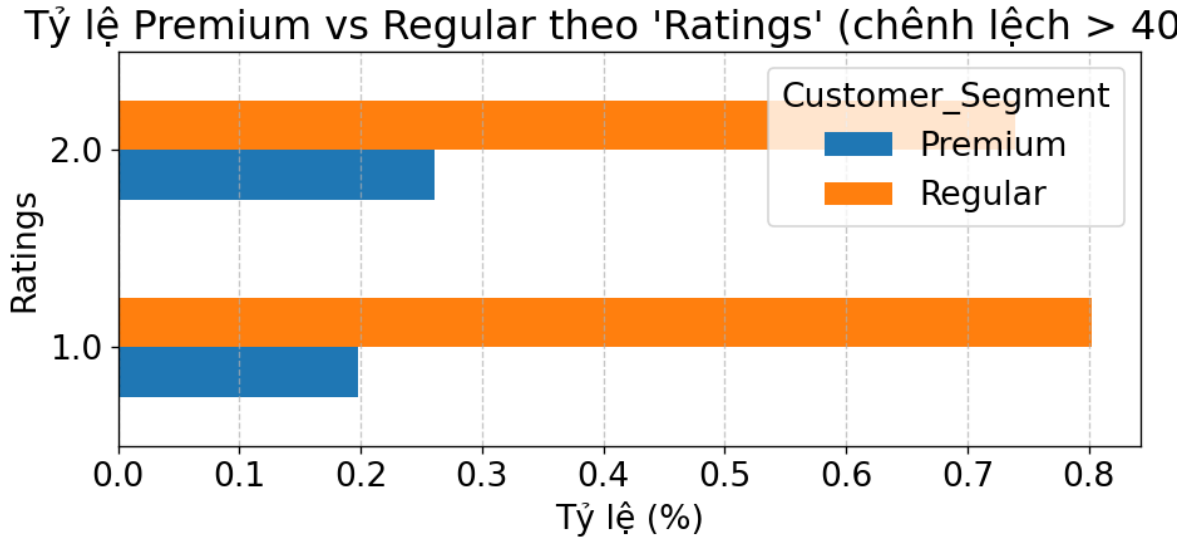
Hình 5.6: Trực quan hóa khác biệt chênh lệch giữa Premium và Regular (theo biến Shipping_Method)



Hình 5.7: Trực quan hóa khác biệt chênh lệch giữa Premium và Regular (theo biến Payment_Category)



Hình 5.8: Trực quan hóa khác biệt chênh lệch giữa Premium và Regular (theo biến Order_Status)



Hình 5.9: Trực quan hóa khác biệt chênh lệch giữa Premium và Regular (theo biến Ratings)

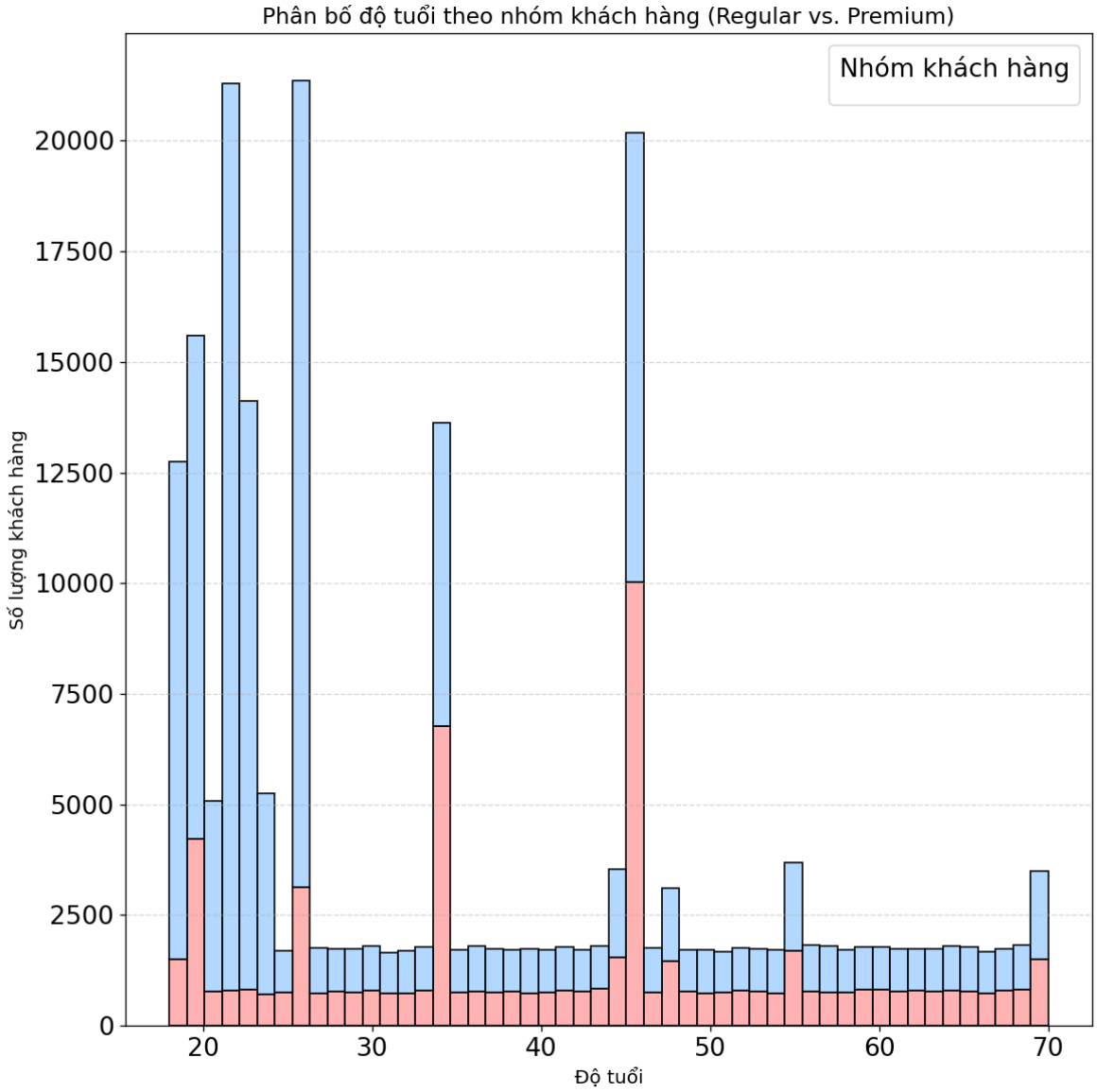
Tiếp theo, đối với biến Age (Độ tuổi), nhóm cũng thực hiện trực quan hóa với mục tiêu xác định nhóm tuổi có sự khác biệt chênh lệch giữa 02 giá trị Premium và Regular của biến mục tiêu. Nội dung này giúp đưa ra kết luận về nhóm tuổi có tiềm năng chuyển đổi phân khúc khách hàng.

```
#%% PLOT AGE DISTRIBUTION BY CUSTOMER SEGMENT
```

```
plt.figure(figsize=(10, 10))
sns.histplot(data=df, x='Age', hue='Customer_Segment', multiple='stack',
             bins=50, palette={'Premium': '#ff9999', 'Regular': '#99ccff'})
plt.title("Phân bố độ tuổi theo nhóm khách hàng (Regular vs. Premium)", fontsize=14)
plt.xlabel("Độ tuổi", fontsize=12)
plt.ylabel("Số lượng khách hàng", fontsize=12)
plt.legend(title="Nhóm khách hàng")
plt.grid(axis='y', linestyle='--', alpha=0.5)
plt.tight_layout()
plt.show()
```

Kết quả trực quan hóa cho thấy nhóm khách hàng từ 21 đến 27 tuổi có sự khác biệt chênh lệch giữa 02 giá trị Premium và Regular. Đồng thời đây cũng là nhóm tuổi có số lượng khách hàng nhiều nhất trong tổng thể. Do đó, hoạt động phân tích cần chú trọng vào nhóm tuổi này để đưa ra các giải pháp, chiến dịch phù hợp cho doanh nghiệp.

Ngoài ra, ở độ tuổi 34 và 45, số lượng khách hàng cũng chiếm giá trị đáng kể, tuy nhiên ở 02 độ tuổi này sự khác biệt chênh lệch giữa 02 giá trị Premium và Regular hầu như không đáng kể.



Hình 5.10: Trực quan hóa khác biệt chênh lệch giữa Premium và Regular (theo biến Age)

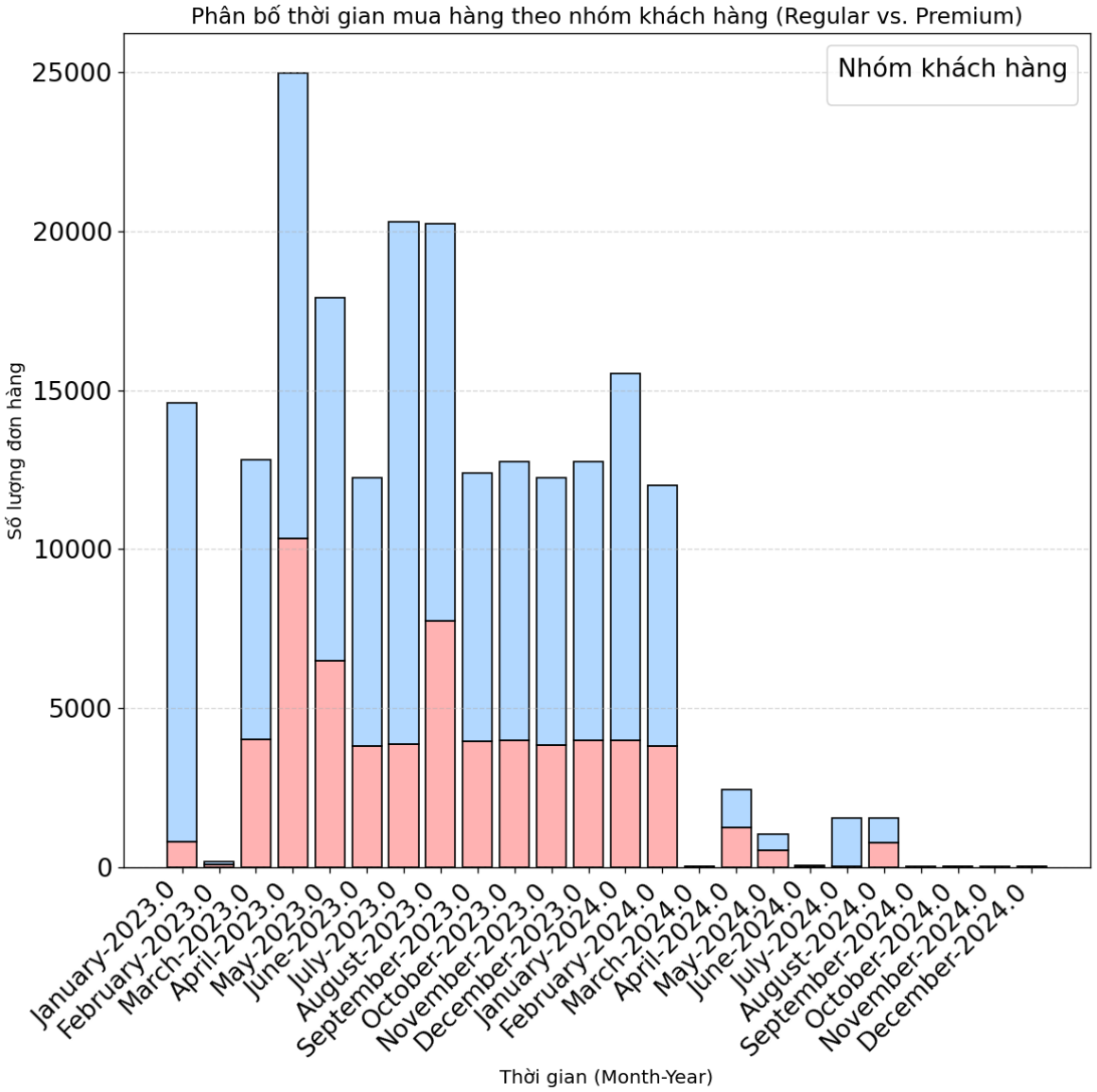
Cuối cùng, với biến phân loại thể hiện thời gian thực hiện giao dịch đã được chuẩn hóa Month_Year, nhóm thực hiện trực quan hóa với mục tiêu xác định khoảng thời gian có sự khác biệt chênh lệch giữa 02 giá trị Premium và Regular của biến mục tiêu. Nội dung này giúp đưa ra kết luận về khoảng thời gian tiềm năng để thực hiện các chiến dịch chuyển đổi phân khúc khách hàng của doanh nghiệp.

```
#%% PLOT MONTH_YEAR DISTRIBUTION BY CUSTOMER SEGMENT
plt.figure(figsize=(10, 10))
sns.histplot(data=df, x='Month_Year', hue='Customer_Segment', multiple='stack',
             shrink=0.8, palette={'Premium': '#ff9999', 'Regular': '#99ccff'})
```

```
    plt.title("Phân bố thời gian mua hàng theo nhóm khách hàng (Regular vs. Premium)",
fontsize=14)
    plt.xlabel("Thời gian (Month-Year)", fontsize=12)
    plt.ylabel("Số lượng đơn hàng", fontsize=12)
    plt.xticks(rotation=45, ha='right')
    plt.legend(title="Nhóm khách hàng")
    plt.grid(axis='y', linestyle='--', alpha=0.5)
    plt.tight_layout()
    plt.show()
```

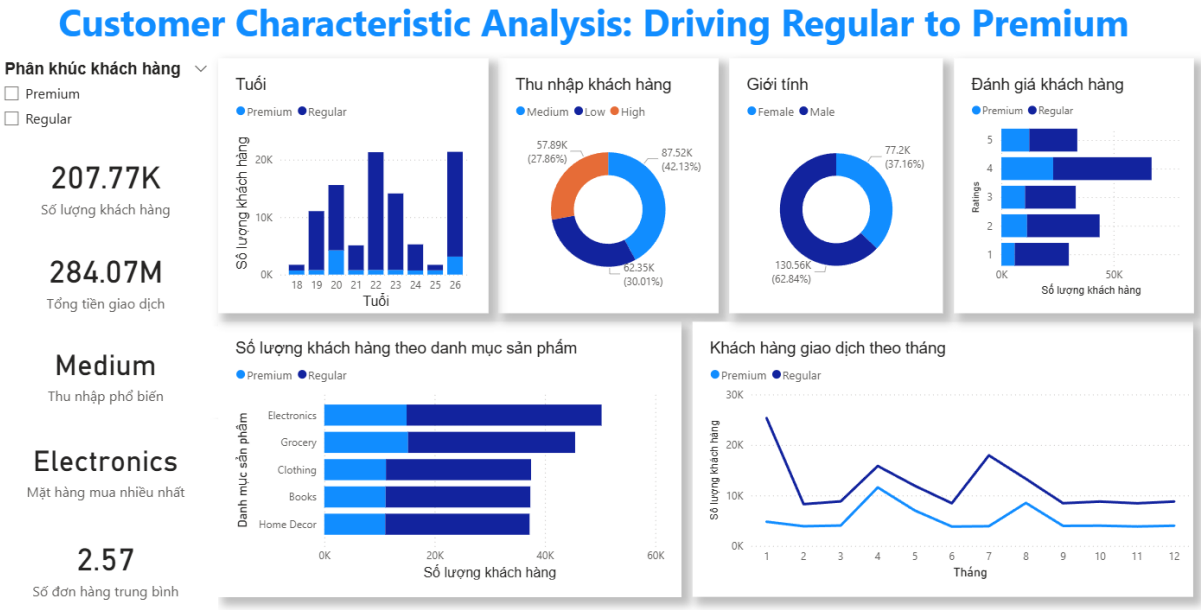
Kết quả trực quan hóa cho thấy khoảng thời gian từ tháng 03 năm 2023 đến tháng 02 năm 2024 có sự khác biệt chênh lệch giữa 02 giá trị Premium và Regular. Đồng thời đây cũng là khoảng thời gian có số lượng giao dịch nhiều nhất trong tổng thể. Thêm vào đó, các tháng 1 và 7 ghi nhận số lượng giao dịch cũng như khác biệt chênh lệch rất cao đối với 02 giá trị Premium và Regular. Điều này cho thấy, những mốc thời gian trên là thời điểm vô cùng phù hợp để triển khai các chiến dịch gia tăng tỉ lệ chuyển đổi phân khúc khách hàng.

Tóm lại, từ các kết quả kể trên, doanh nghiệp có thể nghiên cứu đẩy mạnh thực hiện chiến lược chuyển đổi khách hàng trong thời gian tương ứng ở các năm tiếp theo nhằm tối ưu hóa hiệu quả chiến dịch.


Phân bố thời gian mua hàng theo nhóm khách hàng (Regular vs. Premium)
Nhóm khách hàng
Số lượng đơn hàng
25000
20000
15000
10000
5000
0
January-2023.0
February-2023.0
March-2023.0
April-2023.0
May-2023.0
June-2023.0
July-2023.0
August-2023.0
September-2023.0
October-2023.0
November-2023.0
December-2023.0
January-2024.0
February-2024.0
March-2024.0
April-2024.0
May-2024.0
June-2024.0
July-2024.0
August-2024.0
September-2024.0
October-2024.0
November-2024.0
December-2024.0
Thời gian (Month-Year)

Hình 5.11: Trực quan hóa khác biệt chênh lệch giữa Premium và Regular (theo biến Month_Year)

2. Power BI



Hình 5.12: Dashboard phân tích khách hàng để tăng chuyển đổi từ Regular sang Premium

Trong dashboard, ta có thể thấy khách hàng đã được chia thành hai phân khúc chính là Regular và Premium. Với mục tiêu chính là trực quan hóa nhằm dễ dàng tìm ra những điểm khác biệt nổi bật giữa hai nhóm, từ đó đề xuất chiến lược phù hợp để chuyển đổi nhóm Regular thành Premium, nhóm quyết định lựa chọn các yếu tố như thu nhập, độ tuổi, giới tính, sản phẩm mua, và đánh giá khách hàng, số lượng giao dịch theo tháng vì chúng phản ánh khá rõ đặc điểm và hành vi tiêu dùng của từng nhóm khách hàng.

Bộ dữ liệu số lượng khách hàng bao gồm hơn 207 nghìn khách hàng với tổng giá trị giao dịch đạt trên 284 triệu USD trong khoảng thời gian 2 năm từ 1/2023 tới tháng 12/2024. Khách hàng có thu nhập thấp, trung bình và cao trong đó phần lớn tập trung ở mức trung bình, và sản phẩm được mua nhiều nhất là Electronics. Số đơn trung bình là 2.57 đơn hàng mỗi người, cho thấy tần suất mua sắm chưa quá cao.

Xét về độ tuổi, biểu đồ cho thấy rằng có sự khác biệt rõ ràng về nhóm khách hàng Regular so với Premium ở độ tuổi từ 21 đến 27, nhóm Regular tập trung ở một số độ tuổi nhất định trong khi nhóm Premium lại có xu hướng phân bổ đều hơn. Điều này có thể hiểu rằng những khách hàng trẻ tuy hoạt động nhiều nhưng vẫn chưa đủ động lực hoặc điều kiện để chuyển sang gói Premium. Do đó, đây là nhóm có tiềm năng chuyển

đổi cao nếu có các chương trình tiếp cận phù hợp, đặc biệt là ở độ tuổi 22 và 26 vốn chiếm tỷ lệ cao cả ở hai phân khúc.

Về thu nhập, khách hàng được chia thành ba nhóm: Low, Medium và High. Phân tích cho thấy nhóm có thu nhập trung bình chiếm tỷ lệ lớn nhất. Từ biểu đồ, ta thấy nhóm Regular có thu nhập khá đa dạng, trong đó thu nhập trung bình chiếm tỷ lệ cao nhất (42.89%), còn thu nhập cao và thấp lần lượt là 31.38% và 25.74%. Điều này cho thấy nhiều khách hàng Regular có tiềm năng chuyển sang Premium, nhất là nhóm thu nhập trung bình và cao. Ở chiều ngược lại, nhóm Premium lại có cơ cấu khá bất ngờ khi gần 40% thuộc nhóm thu nhập thấp, chỉ 20% là thu nhập cao. Điều này cho thấy việc trở thành Premium không nhất thiết phải là người thu nhập cao, mà có thể do thói quen mua sắm, nhu cầu cá nhân hoặc mức độ trung thành. Từ đây, có thể thấy nhóm thu nhập trung bình là đối tượng dễ chuyển đổi nhất. Ngoài ra, nếu biết cách tạo giá trị phù hợp, cả nhóm thu nhập thấp cũng có thể trở thành khách hàng Premium.

Qua dashboard, **yếu tố giới tính** cho thấy sự chênh lệch đáng kể, nam giới chiếm hơn 60% tổng số khách hàng, vượt trội so với nữ giới trong cả hai phân khúc Regular và Premium. Điều này phản ánh rằng phần lớn hoạt động mua sắm trên nền tảng hiện nay đến từ nhóm nam. Chính vì vậy, các chiến dịch chuyển đổi nên được thiết kế phù hợp với tâm lý và hành vi tiêu dùng của nam khách hàng, đặc biệt là đặc biệt là nhóm nam trẻ tuổi có thu nhập trung bình, vốn là nhóm chiếm tỷ lệ lớn trong cả hai phân khúc khách hàng.

Đối với yếu tố đánh giá sản phẩm, có sự khác biệt rõ ràng giữa hai nhóm: khách hàng Premium thường có xu hướng đánh giá cao hơn, với nhiều đánh giá 4 và 5 sao. Ngược lại, nhóm Regular có phân bố đánh giá thấp hơn và phân tán hơn. Điều này có thể phản ánh rằng nhóm Premium cảm thấy hài lòng với trải nghiệm, dịch vụ hoặc sản phẩm hơn. Do đó, để tăng khả năng chuyển đổi, doanh nghiệp cần nâng cao trải nghiệm của nhóm Regular, từ chất lượng giao hàng, dịch vụ hậu mãi cho đến sản phẩm phù hợp.

Về loại sản phẩm được mua nhiều nhất, Electronics (sản phẩm điện tử) là danh mục chiếm ưu thế lớn nhất về số lượng khách hàng, cả Regular lẫn Premium. Điều này phản ánh nhu cầu cao và ổn định về sản phẩm công nghệ, vốn là mặt hàng thiết yếu và phổ biến trong đời sống hiện đại. Grocery (hàng tạp hóa) cũng có lượng khách hàng tương đối lớn, chủ yếu đến từ nhóm Regular, cho thấy đây là nhu cầu cơ bản và thường xuyên, nhưng lại ít tạo ra sự khác biệt về trải nghiệm. Ngoài ra các nhóm mặt hàng khác

như Books, Home Decor cũng có sự khác biệt rõ rệt về tỷ lệ Regular và Premium. Điều này cho thấy khách hàng Premium không chỉ mua sắm vì nhu cầu, mà còn vì trải nghiệm. Từ đây, có thể đề xuất các chiến dịch về các sản phẩm hoặc ưu đãi Premium xoay quanh những mặt hàng này.

Cuối cùng, biểu đồ số lượng giao dịch theo tháng cho thấy Regular có sự dao động mạnh qua các tháng, trong khi Premium duy trì ở mức ổn định hơn. Một số tháng như tháng 1, 4 hoặc 7, số lượng giao dịch của Regular cao vượt trội, cho thấy đây là thời điểm vàng để triển khai các chương trình khuyến khích chuyển đổi 2 nhóm khách hàng. Việc nắm bắt đúng thời điểm sẽ giúp tăng tỷ lệ chuyển đổi hiệu quả hơn nhiều so với việc tiếp cận rải rác quanh năm.

CHƯƠNG VI: TỔNG KẾT

1. Chiến dịch đề xuất

Về định hướng tổng thể, các chiến dịch đề xuất sẽ nhấn mạnh vai trò của phân khúc Premium như một biểu tượng về quyền lợi, sự khác biệt và trải nghiệm cá nhân hóa, đồng thời từng bước xây dựng nhận thức về giá trị lâu dài mà việc nâng cấp hạng thành viên mang lại, chứ không chỉ đơn thuần là những ưu đãi ngắn hạn. Mục tiêu cuối cùng của chiến dịch là gia tăng ít nhất 10% tỷ lệ chuyển đổi giữa nhóm khách hàng Regular và Premium trong vòng 12 tháng, tức nhóm khách hàng Premium tăng thêm 10% tỉ lệ so với cùng kỳ năm trước.

Dựa trên kết quả phân tích dữ liệu đặc điểm khách hàng, nhóm đối tượng tiềm năng để chuyển đổi từ phân khúc Regular sang Premium chủ yếu là khách hàng có thu nhập trung bình (Medium) và cao (High). Đồng thời, nhóm khách hàng này đang tiêu dùng các sản phẩm thuộc 04 ngành hàng có ý nghĩa phân biệt rõ rệt, bao gồm: Electronics, Home Decor, Books và Clothing. Đây là những ngành hàng mà đặc điểm mua sắm giữa hai phân khúc thể hiện sự khác biệt đáng kể về tần suất, mức chi tiêu và xu hướng tiêu dùng. Do đó, doanh nghiệp cần định hướng các chiến lược chuyển đổi cụ thể đối với từng ngành hàng nhằm tối ưu hóa hiệu quả chiến dịch.

- Trong ngành hàng Electronics, nhóm khách hàng Regular có thu nhập cao đang chiếm tỷ lệ không nhỏ và thể hiện tiềm năng nâng cấp nếu được thúc đẩy thông qua các chiến dịch cá nhân hóa và xây dựng nhận thức về giá trị mà hạng mức Premium mang lại. Việc tăng cường các thông điệp như "ưu tiên bảo hành, tư vấn chuyên sâu, trải nghiệm công nghệ sớm" sẽ dễ dàng tạo động lực chuyển đổi ở nhóm khách hàng này – những người vốn quan tâm đến chất lượng và giá trị đối ứng với mức chi trả hơn là mức giá.

- Tương tự, với ngành hàng Home Decor, những khách hàng thu nhập trung bình đến cao, đặc biệt ở độ tuổi 21–27, đang sinh sống tại các quốc gia có mức sống cao như UK và US, được dự đoán sẽ thể hiện xu hướng chi tiêu tăng trưởng khi trải nghiệm mua hàng được cá nhân hóa và mang tính định hình phong cách riêng. Từ đánh giá trên, chiến dịch chuyển đổi nên xây dựng hình ảnh khách hàng Premium như những người có gu thẩm mỹ riêng biệt, được quyền truy cập các bộ sưu tập giới hạn hoặc đề xuất sản phẩm theo cá tính, các sản phẩm độc quyền, mang dấu ấn cá nhân.

- Đối với ngành hàng Books và Clothing, khách hàng Regular có thu nhập trung bình vẫn duy trì hành vi tiêu dùng lặp lại, do ngành hàng trên là nhóm các sản phẩm

thiết yếu đối với người tiêu dùng. Song nhóm khách hàng này vẫn chưa có sự nhận thức rõ rệt về giá trị của hoạt động nâng cấp hạng thành viên. Với nhóm đối tượng này, chiến lược tiếp cận cần đi theo hướng truyền thông giá trị dài hạn của hạng thành viên Premium. Điều này có nghĩa, doanh nghiệp cần cung cấp cho khách hàng các quyền lợi không chỉ dừng lại ở chiết khấu, ưu đãi mà còn là trải nghiệm tiêu dùng trọn vẹn như được chọn trước các ấn phẩm mới, nhận đề xuất theo thói quen đọc sách, trang phục, hoặc các combo thời trang theo mùa được cá nhân hóa theo gu thẩm mỹ cá nhân,...

Bên cạnh đó, nhằm tối ưu hóa các chiến dịch kể trên, các yếu tố bổ trợ mang ý nghĩa phân tích như giới tính nam, độ tuổi từ 21 đến 27, hành vi giao hàng theo phương thức tiêu chuẩn (Standard), thanh toán qua PayPal,... có thể được sử dụng để xây dựng tệp khách hàng mục tiêu chi tiết cho mỗi ngành hàng. Đặc biệt, những khách hàng Regular thường xuyên đánh giá sản phẩm ở mức 1.0–2.0, nếu thuộc nhóm thu nhập cao và trung bình, đồng thời tiêu dùng các ngành hàng nêu trên, doanh nghiệp cần có giải pháp đưa nhóm đối tượng này vào danh sách ưu tiên chăm sóc, cá nhân hóa để tái tạo trải nghiệm tích cực và khuyến khích nâng hạng thành viên, tránh tình trạng trải nghiệm khách hàng quá chênh lệch giữa 02 nhóm Regular và Premium.

Chiến dịch sẽ được triển khai vào các tháng trọng điểm 1, 4 và 7, những thời điểm ghi nhận hành vi tiêu dùng tăng cao, đồng thời tỉ lệ khách hàng Regular vẫn đang chiếm tỉ trọng đáng kể. Đây cũng là giai đoạn lý tưởng để vận hành các chương trình marketing ứng dụng công nghệ AI/ML như: dự đoán xu hướng mua hàng tiếp theo, cá nhân hóa nội dung quảng cáo theo ngành hàng yêu thích, tự động gửi ưu đãi nâng hạng cho nhóm khách hàng có điểm tiềm năng cao và cung cấp trải nghiệm quyền lợi Premium miễn phí trong thời gian quy định.

Cuối cùng, đối với hình thức triển khai, nhằm gia tăng hiệu quả tiếp cận và chuyển đổi khách hàng mục tiêu, chiến dịch nên được triển khai theo hình thức đa kênh, trong đó mỗi kênh đóng vai trò riêng nhưng phối hợp chặt chẽ nhằm tạo ra trải nghiệm nhất quán và cá nhân hóa. Một số hình thức triển khai được đề xuất bao gồm: email cá nhân hóa, quảng cáo remarketing, chatbot gợi ý sản phẩm, trò chơi hóa (gamification).

2. Kết quả đạt được

- Giúp các bên liên quan có cái nhìn rõ ràng và hệ thống về đặc điểm khách hàng hiện tại, đặc biệt là sự khác biệt giữa nhóm Regular và Premium.

- Bằng cách trực quan hóa dữ liệu trên Power BI, doanh nghiệp có thể dễ dàng nhận biết những nhóm khách hàng tiềm năng để triển khai các chiến dịch chuyển đổi, ví dụ: nhóm nam từ 19-27 tuổi, thu nhập trung bình, thường mua sản phẩm công nghệ,...

- Ngoài ra, phân tích cũng hỗ trợ việc theo dõi xu hướng mua hàng theo thời gian, danh mục sản phẩm phổ biến, và đánh giá mức độ hài lòng qua chỉ số rating, từ đó phục vụ việc ra quyết định nhanh hơn, chính xác hơn trong hoạt động marketing và chăm sóc khách hàng.

- Cuối cùng là đề xuất các chiến dịch chiến lược nhằm nâng cao tỷ lệ chuyển đổi khách hàng từ nhóm Regular sang Premium. Thông qua việc nhận diện các đặc điểm và hành vi tiêu dùng đặc trưng của từng phân khúc, nhóm đã xác định được những hướng tiếp cận phù hợp để cá nhân hóa trải nghiệm khách hàng, tối ưu hóa nội dung tiếp thị và xây dựng chính sách ưu đãi hiệu quả.

3. Hạn chế

- Phân tích hiện tại mới dừng lại ở các chỉ số mô tả và tương quan đơn giản, chưa áp dụng các phương pháp phân tích nâng cao như clustering, RFM hay phân khúc hành vi sâu hơn.

- Dữ liệu chỉ giới hạn trong một giai đoạn cố định, chưa phản ánh được xu hướng thay đổi dài hạn của người tiêu dùng.

- Dashboard mới hỗ trợ theo dõi từng chiều dữ liệu riêng lẻ, chưa tích hợp khả năng tạo phân khúc động theo tổ hợp nhiều đặc điểm (ví dụ: nam + trẻ + rating cao + mua Electronics).

4. Phát triển trong tương lai

- Nhóm hướng đến mở rộng phạm vi phân tích bằng cách thu thập dữ liệu theo thời gian thực, tích hợp thêm dữ liệu ngoài như phản hồi khách hàng, hành vi truy cập website,...

- Phân tích có thể phát triển thêm các tính năng tạo phân khúc khách hàng động, xây dựng mô hình điểm tiềm năng chuyển đổi, và đề xuất ưu đãi cá nhân hóa.

- Về mặt trực quan, dashboard sẽ được nâng cấp thêm để hỗ trợ lọc đa chiều và xuất tự động các insight dạng văn bản, giúp nhà quản lý không cần đọc biểu đồ mà vẫn nắm được bức tranh tổng thể.