**TRƯỜNG ĐẠI HỌC SÀI GÒN**

**KHOA CÔNG NGHỆ THÔNG TIN**

A blue circle with text

Description automatically generated****

**BÀI TẬP NHÓM LỚN LAB 1**

**Danh sách sinh viên thực hiện:**

**3122410138 – Bùi Minh Huy**

**3122410128 – Phan Huy Hoàng**

**3112410336 – Hoàng Minh Quang**

**3122410106 – Nguyễn Trần Công Hậu**

**Giảng viên hướng dẫn: Đỗ Như Tài**

**Thành phố Hồ Chí Minh – Tháng 10/2025**

MỤC LỤC

[**LỜI MỞ ĐẦU 4**](#_Toc211060558)

[**Bối cảnh chung và tính cấp thiết của phân tích dữ liệu 4**](#_Toc211060559)

[**Mục tiêu và phạm vi báo cáo 4**](#_Toc211060560)

[**Tổng quan về các chuyên đề phân tích 5**](#_Toc211060561)

[**Cấu trúc báo cáo 5**](#_Toc211060562)

[**KẾ HOẠCH THỰC HIỆN 6**](#_Toc211060563)

[**CHƯƠNG 1. PHÂN TÍCH CHẤT LƯỢNG RƯỢU ĐỎ 7**](#_Toc211060564)

[**1. Bối cảnh vấn đề và mục tiêu phân tích 7**](#_Toc211060565)

[**1.1. Bối cảnh vấn đề 7**](#_Toc211060566)

[**1.2. Mục tiêu phân tích 7**](#_Toc211060567)

[**1.3. Giới thiệu tổng quan về Dataset 7**](#_Toc211060568)

[**2. Mô tả thống kê và đánh giá dữ liệu 9**](#_Toc211060569)

[**2.1. Hiển thị thông tin tổng quan về dữ liệu 9**](#_Toc211060570)

[**2.2. Các đặc tính thống kê trên dữ liệu số 10**](#_Toc211060571)

[**2.3. Kiểm tra tính toàn vẹn của dữ liệu 11**](#_Toc211060572)

[**3. Phân tích khám phá và trực quan hóa dữ liệu 11**](#_Toc211060573)

[**3.1. Phân tích đơn biến (Univariate Analysis) 11**](#_Toc211060574)

[**3.2. Phân tích đa biến (Multivariate Analysis) 16**](#_Toc211060575)

[**4. Kết luận và đề xuất 23**](#_Toc211060576)

[**4.1. Các phát hiện chủ chốt 23**](#_Toc211060577)

[**4.2. Đề xuất Giải pháp cho Nhà sản xuất 23**](#_Toc211060578)

[**CHƯƠNG 2. PHÂN TÍCH DỮ LIỆU BÁN HÀNG SIÊU THỊ 25**](#_Toc211060579)

[**1. Bối cảnh vấn đề và mục tiêu phân tích 25**](#_Toc211060580)

[**1.1. Bối cảnh vấn đề 25**](#_Toc211060581)

[**1.2. Mục tiêu phân tích 25**](#_Toc211060582)

[**1.3. Giới thiệu tổng quan về Dataset 25**](#_Toc211060583)

[**2. Xử lí dữ liệu 27**](#_Toc211060584)

[**2.1. Kiểm tra và xử lí giá trị bị thiếu và không hợp lệ 27**](#_Toc211060585)

[**2.2. Kiểm tra và xử lí giá trị trùng lặp 28**](#_Toc211060586)

[**3. Phân tích khám phá và trực quan hóa dữ liệu 28**](#_Toc211060587)

[**3.1. Phân tích đơn biến (Univariate Analysis) 28**](#_Toc211060588)

[**3.2. Phân tích đa biến (Multivariate Analysis) 32**](#_Toc211060589)

[**3.3. Phân tích RFM và Phân cụm 36**](#_Toc211060590)

[**4. Kết luận 42**](#_Toc211060591)

[**CHƯƠNG 3. PHÂN TÍCH DỮ LIỆU CHIẾN DỊCH MARKETING 45**](#_Toc211060592)

[**1. Định nghĩa vấn đề và mục tiêu phân tích 45**](#_Toc211060593)

[**1.1. Bối cảnh vấn đề 45**](#_Toc211060594)

[**1.2. Mục tiêu phân tích 45**](#_Toc211060595)

[**1.3. Giới thiệu tổng quan về Dataset 46**](#_Toc211060596)

[**2. Giới thiệu công cụ 48**](#_Toc211060597)

[**2.1. Sweetviz 48**](#_Toc211060598)

[**2.1. AutoViz 48**](#_Toc211060599)

[**3. Mô tả thống kê về dữ liệu 49**](#_Toc211060600)

[**3.1. Thông tin tổng quát 49**](#_Toc211060601)

[**4. Làm sạch dữ liệu 51**](#_Toc211060602)

[**4.1. Kiểm tra và xử lí giá trị trùng lặp 51**](#_Toc211060603)

[**4.2. Kiểm tra và xử lí giá trị thiếu 51**](#_Toc211060604)

[**5. Phân tích khám phá dữ liệu với SweetViz và AutoViz 52**](#_Toc211060605)

[**5.1. SweetViz 52**](#_Toc211060606)

[**5.2. AutoViz 61**](#_Toc211060607)

# LỜI MỞ ĐẦU

## Bối cảnh chung và tính cấp thiết của phân tích dữ liệu

Trong môi trường kinh doanh ngày nay, dữ liệu đã trở thành một tài sản chiến lược, và khả năng khai thác dữ liệu để đưa ra quyết định thông minh là yếu tố then chốt tạo nên lợi thế cạnh tranh. Việc chuyển đổi từ quyết định dựa trên trực giác sang chiến lược dựa trên bằng chứng thực nghiệm đang định hình lại mọi khía cạnh của doanh nghiệp, từ việc cải tiến chất lượng sản phẩm, tối ưu hóa hoạt động bán lẻ, cho đến việc đo lường hiệu quả của các nỗ lực tiếp thị.

Báo cáo này là một minh chứng cho cách tiếp cận đó, thể hiện việc ứng dụng các phương pháp phân tích dữ liệu vào ba lĩnh vực kinh doanh riêng biệt, nhằm mục đích khai phá những hiểu biết sâu sắc và có giá trị hành động trong từng bối cảnh cụ thể.

## Mục tiêu và phạm vi báo cáo

Cần làm rõ ngay từ đầu rằng báo cáo này trình bày kết quả của **ba nghiên cứu phân tích dữ liệu hoàn toàn độc lập**. Không có một luận điểm chung hay một câu hỏi nghiên cứu bao trùm kết nối cả ba phần. Thay vào đó, mục tiêu chính của báo cáo là:

* Trình bày việc áp dụng các kỹ thuật phân tích dữ liệu đa dạng vào các bài toán thực tế khác nhau.
* Khai thác những thông tin chuyên sâu, đặc thù cho từng lĩnh vực: chất lượng sản phẩm, hành vi người tiêu dùng tại điểm bán, và hiệu quả chiến dịch marketing.
* Cung cấp các kết luận và đề xuất riêng lẻ cho mỗi chuyên đề phân tích.

Mỗi phân tích trong báo cáo này được xem như một chuyên đề riêng biệt, với bối cảnh, mục tiêu, tập dữ liệu và kết luận riêng.

## Tổng quan về các chuyên đề phân tích

Báo cáo sẽ đi sâu vào ba chuyên đề chính, mỗi chuyên đề giải quyết một bài toán kinh doanh riêng biệt:

* Phân tích Chất lượng Rượu vang Đỏ: Nghiên cứu các yếu tố hóa học quyết định đến chất lượng cảm quan của sản phẩm.
* Phân tích Dữ liệu Bán hàng Siêu thị: Khám phá các xu hướng và mô hình mua sắm của người tiêu dùng tại điểm bán.
* Phân tích Dữ liệu Chiến dịch Marketing: Đánh giá hiệu quả và mức độ thành công của một chiến dịch tiếp thị cụ thể.

## Cấu trúc báo cáo

Để đảm bảo sự rõ ràng và mạch lạc cho từng chủ đề, báo cáo được tổ chức thành các phần riêng biệt. Sau phần mở đầu này, các chương tiếp theo sẽ lần lượt trình bày chi tiết từng chuyên đề phân tích:

* Chương 1 sẽ tập trung hoàn toàn vào Phân tích Chất lượng Rượu vang Đỏ.
* Chương 2 sẽ trình bày các phát hiện từ Phân tích Dữ liệu Bán hàng Siêu thị.
* Chương 3 sẽ đi sâu vào Phân tích Dữ liệu Chiến dịch Marketing.

Mỗi chương sẽ hoạt động như một báo cáo nhỏ, có cấu trúc độc lập bao gồm phần giới thiệu bối cảnh, phương pháp phân tích, trình bày kết quả và thảo luận về các ý nghĩa riêng của nó.

# KẾ HOẠCH THỰC HIỆN

Bảng phân công công việc nhóm:

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Họ và tên** | **MSSV** | **Công việc** |
| 1 | Bùi Minh Huy | 3122410138 | - Thực hiện EDA và trực quan hóa trên Dataset Rượu đỏ.  - Viết báo cáo Word. |
| 2 | Phan Huy Hoàng | 3122410128 | - Thực hiện mô tả thống kê trên Dataset Rượu đỏ.  - Làm PPT Dataset Rượu đỏ. |
| 3 | Hoàng Minh Quang | 3112410336 | - Thực hiện EDA và trực quan hóa trên Dataset Siêu thị.  - Làm PPT Dataset Siêu thị |
| 4 | Nguyễn Trần Công Hậu | 3122410106 | - Thực hiện EDA và trực quan hóa tự động trên Dataset Marketing Compaign.  - Làm PPT Dataset Marketing Compaign. |

# CHƯƠNG 1. PHÂN TÍCH CHẤT LƯỢNG RƯỢU ĐỎ

## Bối cảnh vấn đề và mục tiêu phân tích

### 1.1. Bối cảnh vấn đề

Chất lượng rượu vang là một khái niệm phức tạp, thường được đánh giá dựa trên cảm quan của các chuyên gia - một quá trình mang tính chủ quan cao. Tuy nhiên, đối với các nhà sản xuất, việc tìm ra một phương pháp khách quan để đo lường và dự đoán chất lượng sản phẩm dựa trên các thông số có thể kiểm soát được là một bài toán vô cùng quan trọng. Việc liên kết các đặc tính hóa lý (physicochemical) của rượu vang với điểm số chất lượng cảm quan không chỉ giúp tiêu chuẩn hóa quy trình sản xuất mà còn mở ra khả năng tối ưu hóa sản phẩm để đáp ứng tốt hơn thị hiếu của người tiêu dùng.

Nghiên cứu này sẽ tập trung vào việc phân tích dữ liệu từ các mẫu rượu vang đỏ thuộc dòng "Vinho Verde" của Bồ Đào Nha để tìm ra mối liên hệ giữa các chỉ số hóa học và chất lượng cuối cùng của rượu.

### 1.2. Mục tiêu phân tích

Chuyên đề phân tích này được thực hiện nhằm đạt được các mục tiêu chính sau:

* Khám phá và trực quan hóa mối quan hệ giữa các đặc tính hóa lý khác nhau của rượu vang đỏ.
* Xác định những yếu tố hóa học nào có ảnh hưởng mạnh mẽ nhất đến điểm chất lượng của rượu.
* Cung cấp những hiểu biết dựa trên dữ liệu, có thể hỗ trợ các nhà sản xuất trong việc kiểm soát và cải thiện chất lượng sản phẩm.

### 1.3. Giới thiệu tổng quan về Dataset

Để thực hiện các mục tiêu trên, nghiên cứu sử dụng bộ dữ liệu về chất lượng rượu vang đỏ, bao gồm 1599 mẫu thử nghiệm. Mỗi mẫu được mô tả bởi 12 thuộc tính, trong đó có 11 biến đầu vào là các chỉ số hóa lý và 1 biến đầu ra là điểm chất lượng cảm quan.

**Biến đầu vào (Input Variables):**

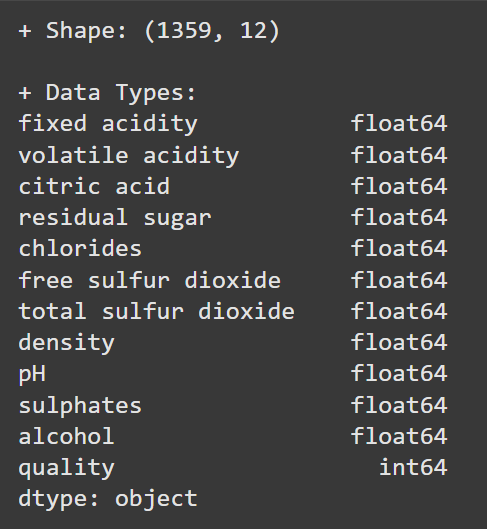
* **Fixed Acidity (Độ axit cố định):** Tổng nồng độ các axit hữu cơ không bay hơi (như tartaric, malic, citric), là thành phần chính tạo nên độ chua đặc trưng của rượu.
* **Volatile Acidity (Độ axit bay hơi):** Nồng độ axit axetic trong rượu. Ở mức độ cao, chỉ số này có thể tạo ra mùi giấm khó chịu.
* **Citric Acid (Axit citric):** Một lượng nhỏ axit citric có thể làm tăng thêm sự tươi mát và hương vị cho rượu.
* **Residual Sugar (Lượng đường dư):** Lượng đường còn lại sau khi quá trình lên men kết thúc, quyết định độ ngọt của rượu.
* **Chlorides (Hàm lượng muối):** Nồng độ muối (ion Cl⁻) trong rượu.
* **Free Sulfur Dioxide (SO₂ tự do):** Lượng SO₂ ở dạng tự do, có tác dụng chống oxy hóa và kháng khuẩn, bảo vệ rượu khỏi hư hỏng.
* **Total Sulfur Dioxide (Tổng SO₂):** Tổng lượng SO₂ tự do và liên kết. Đây là một chỉ số quan trọng trong việc bảo quản rượu.
* **Density (Mật độ):** Tỷ trọng của rượu, chịu ảnh hưởng chủ yếu bởi nồng độ cồn và lượng đường.
* **pH:** Chỉ số đo độ axit hoặc kiềm của rượu trên thang điểm từ 0 đến 14. Hầu hết rượu vang có độ pH từ 3-4.
* **Sulphates (Sunphat):** Hợp chất chứa lưu huỳnh có thể được thêm vào để ngăn chặn sự phát triển của vi khuẩn và nấm men không mong muốn.
* **Alcohol (Nồng độ cồn):** Hàm lượng cồn theo thể tích (%), ảnh hưởng lớn đến hương vị, cấu trúc và cảm nhận tổng thể của rượu.

**Biến đầu ra (Output Variable):**

* **Quality (Chất lượng):** Điểm số đánh giá cảm quan của rượu, được cho trong thang điểm từ 3 đến 8, trong đó điểm càng cao thể hiện chất lượng càng tốt.

## 2. Mô tả thống kê và đánh giá dữ liệu

### 2.1. Hiển thị thông tin tổng quan về dữ liệu



Hình 2.1.1. Hiển thị số dòng, số cột và kiểu dữ liệu của từng cột

* **Nhận xét:**
* Dữ liệu có 11 tính chất để phân lớp: fixed acidity,volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, ph, sulphates, alcohol.
* Dữ liệu phân lớp nằm ở cột quality gồm 6 giá trị, thang điểm từ 3 đến 8
* Các giá trị đều là số học thuộc kiểu float64 và int64
* Tổng số dòng dữ liệu là 1599



Hình 2.1.2. Hiển thị 5 dòng đầu và 5 dòng cuối của bảng dữ liệu

### 2.2. Các đặc tính thống kê trên dữ liệu số



Hình 2.2.1. Hiển thị các tính chất thống kê trên dữ liệu số

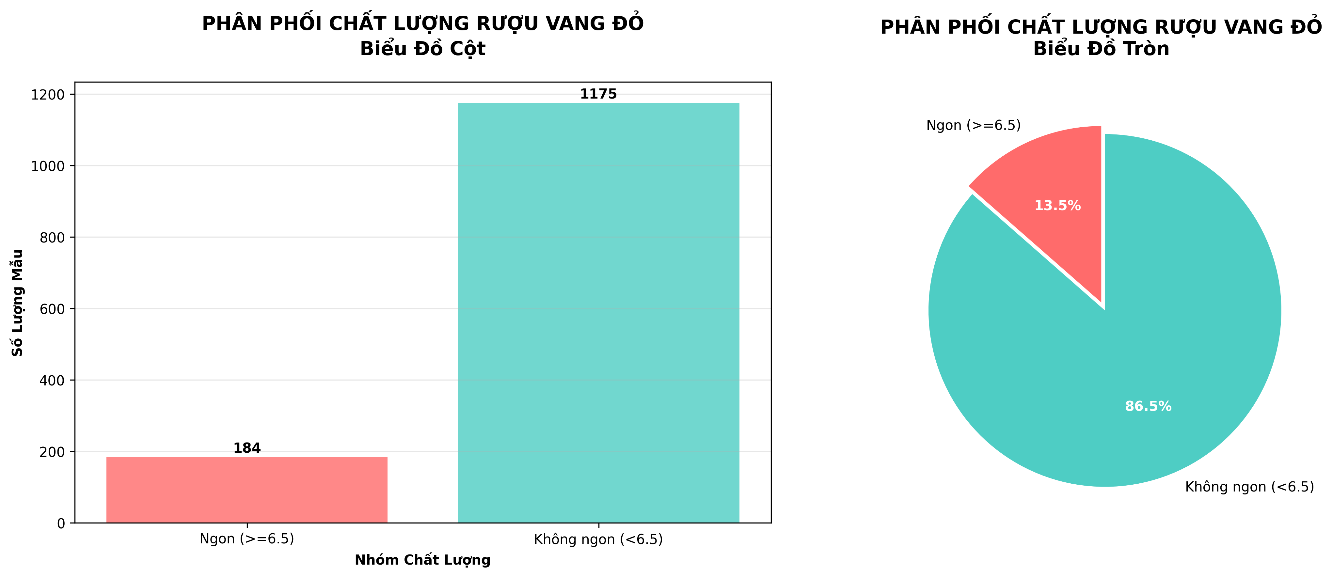
* **Nhận xét:**
* Nhìn chung, chất lượng rượu vang trong bộ dữ liệu này tập trung ở mức trung bình, với điểm số phổ biến là 5 và 6, và giá trị trung bình là 5.63 trên thang điểm từ 3 đến 8.
* Các chỉ số cốt lõi như pH (trung bình 3.31) và density (trung bình 0.996) cho thấy sự ổn định cao và phân bổ hẹp.
* Tuy nhiên, một số yếu tố khác lại có sự biến động rất lớn và xuất hiện nhiều giá trị ngoại lai. Điển hình là total sulfur dioxide có giá trị trung bình là 46.46 nhưng có thể lên tới 289, hay residual sugar với giá trị trung bình chỉ 2.53 nhưng có mẫu đạt tới 15.5. Điều này cho thấy sự đa dạng đáng kể trong quy trình ủ và bảo quản, dẫn đến một số mẫu rượu có đặc tính hóa học rất khác biệt so với phần còn lại.

### 2.3. Kiểm tra tính toàn vẹn của dữ liệu

## 3. Phân tích khám phá và trực quan hóa dữ liệu

### 3.1. Phân tích đơn biến (Univariate Analysis)

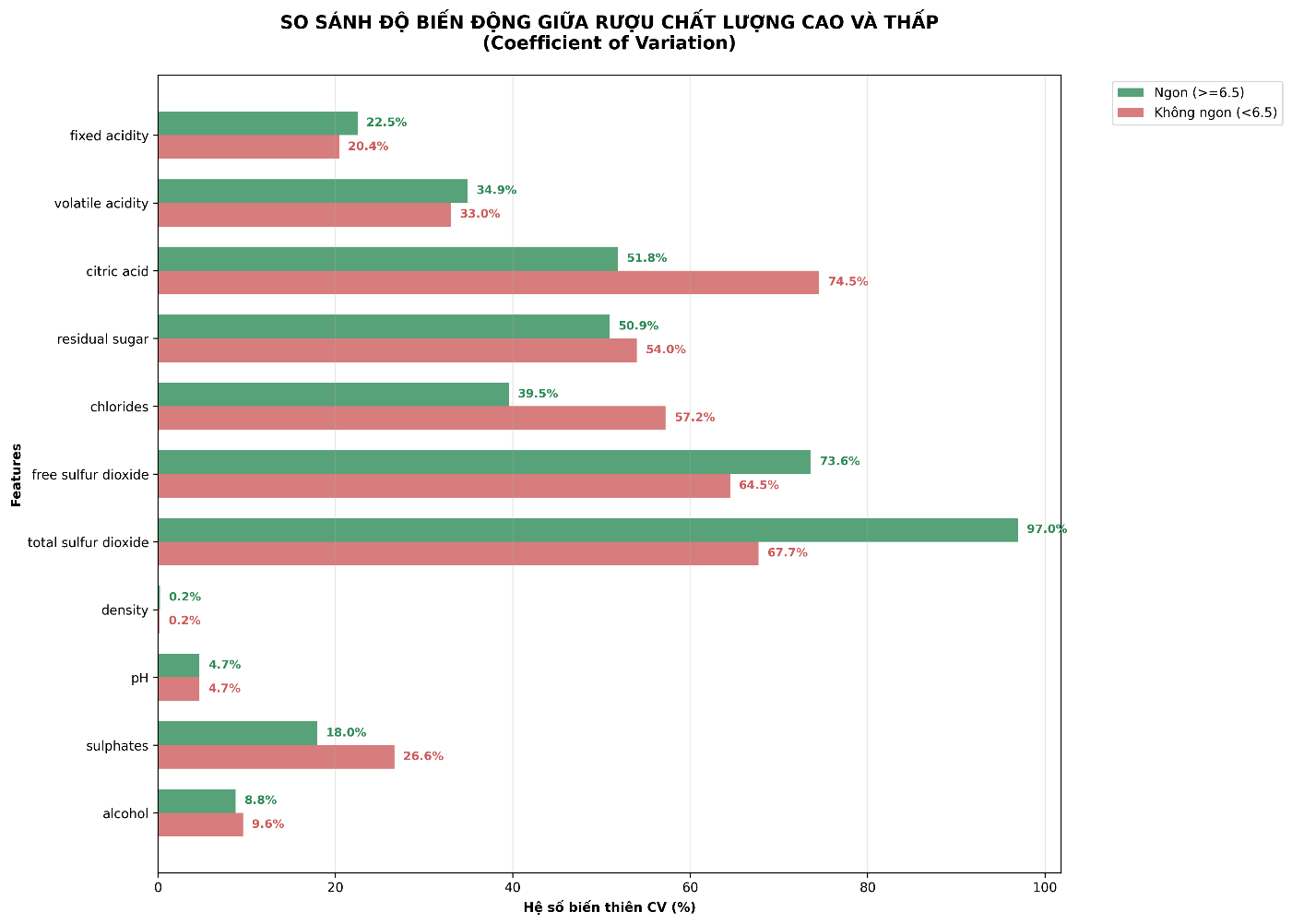
***Câu hỏi đặt ra:*** Chất lượng rượu trong tập dữ liệu được phân bổ như thế nào? Tỷ lệ giữa rượu chất lượng cao và rượu chất lượng thấp là bao nhiêu?



Hình 3.1.1. Biểu đồ cột và tròn thể hiện sự phân phối chất lượng rượu

* **Quá trình phân tích:** Thêm cột “Quality Category” để phân loại rượu thành 2 cấp độ (Ngon, Không ngon) giúp dễ dàng đánh giá bằng mắt thường.
* **Kết quả phân tích:** Chất lượng rượu phân bổ không đồng đều, với đa số (86.5%) ở mức không ngon, tỷ lệ rượu ngon là 13.5%. Điều này phản ánh thực tế là phần lớn rượu vang đỏ trên thị trường có chất lượng kém không ngon, trong khi rượu chất lượng cao thực sự khá hiếm và rất ít.

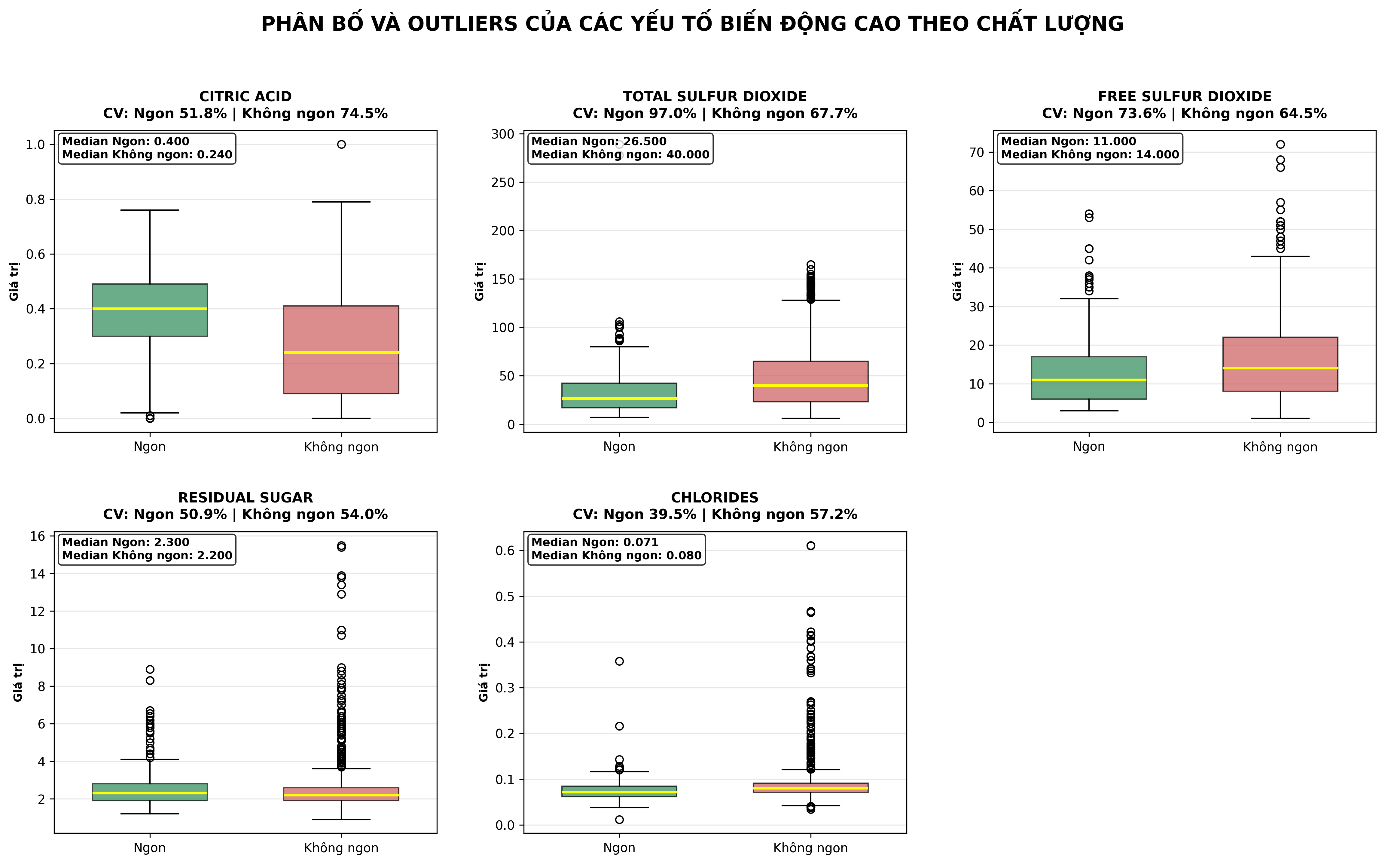
***Câu hỏi then chốt:*** Vậy dựa trên tiêu chí nào để xác định yếu tố ảnh hưởng đến chất lượng của rượu?



Hình 3.1.2. Biểu đồ thanh so sánh độ biến động giữa hai chất lượng rượu

* **Kết quả phân tích:** Phân tích hệ số biến thiên (CV) theo nhóm chất lượng cho thấy sự khác biệt rõ rệt trong độ ổn định của các chỉ tiêu:
* Ở nhóm rượu không ngon, các yếu tố citric acid (CV = 74.5%), chlorides (57.2%), và residual sugar (54.0%) có biến động đặc biệt cao hơn nhóm rượu ngon, điều này phản ánh sự thiếu kiểm soát trong quy trình sản xuất.
* Trong khi đó, free sulfur dioxide (73.6%) và total sulfur dioxide (67.7%) có CV cao ở cả hai nhóm, cho thấy đây là những biến vốn dĩ khó kiểm soát.
* Ngược lại, các chỉ số alcohol, sulphates, pH và density có độ biến động thấp (CV < 30%) ở cả hai nhóm, chứng tỏ được kiểm soát chặt chẽ. -> Do đó, ưu tiên lựa chọn các chỉ tiêu có (CV > 40%) vì chúng là những ứng viên tiềm năng nhất cho phân tích nguyên nhân ảnh hưởng đến chất lượng rượu.

***Câu hỏi mở rộng:*** Phân bố thực tế của các yếu tố có biến động cao (CV > 40%) như thế nào, và có xuất hiện các giá trị ngoại lai (outliers) đáng chú ý không?

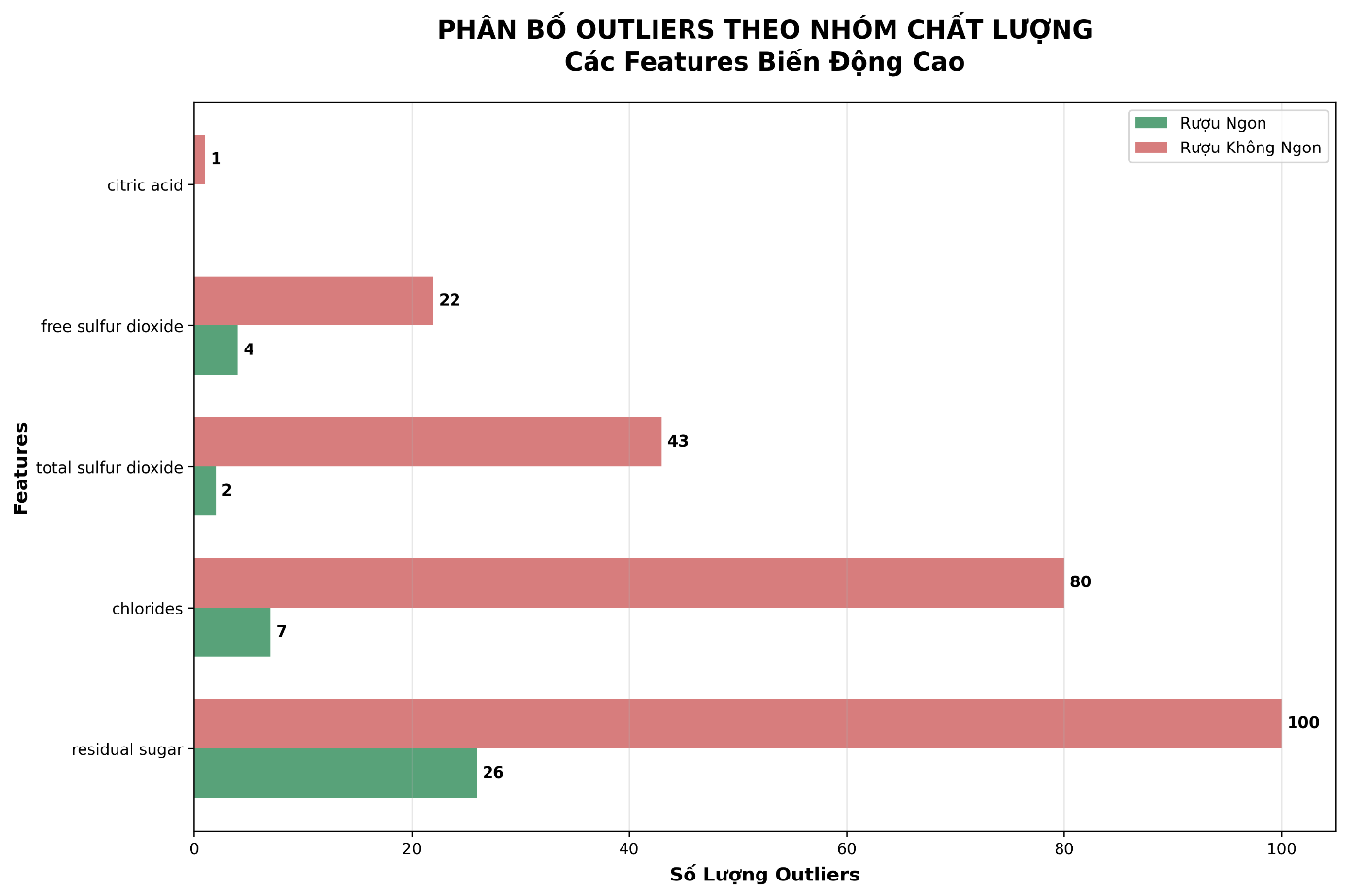


Hình 3.1.3. Biểu đồ hộp trình bày sự phân bố các

yếu tố biến động cao theo chất lượng

* **Kết quả phân tích:**
* citric acid (Axit Citric): Nhóm rượu ngon có median cao hơn (0.400 vs 0.240). Hàm lượng axit citric đủ cao là yếu tố quan trọng cho rượu chất lượng tốt. Cả hai nhóm đều có outliers, nhưng nhóm không ngon có biến động lớn hơn (CV 74.5%)
* residual sugar (Đường tồn dư): Median gần như tương đương (2.300 vs 2.200). Cả 2 nhóm đều lệch phải mạnh với nhiều outliers ở giá trị cao. Đường tồn dư không phải là yếu tố phân biệt chính chất lượng
* total sulfur dioxide (SO₂ tổng): Nhóm không ngon có median cao hơn (40.000 vs 26.500). Nhóm ngon có CV rất cao (97.0%) với nhiều outliers. SO₂ tổng cao có thể là dấu hiệu của chất lượng kém
* free sulfur dioxide (SO₂ tự do): Tương tự với SO₂ tổng, nhóm không ngon có median CAO HƠN (14.000 vs 11.000). Cả hai nhóm đều có outliers ở giá trị cao
* chlorides (Muối): Nhóm không ngon có median CAO HƠN (0.080 vs 0.071). Nhóm không ngon biến động mạnh hơn (CV 57.2% vs 39.5%). Hàm lượng muối cao liên quan đến chất lượng thấp
* **Ý nghĩa rút ra về mối quan hệ Chất lượng - Thành phần:**
* Yếu tố tích cực: Citric acid cao → chất lượng tốt
* Yếu tố tiêu cực: SO₂ (tổng & tự do) cao, Chlorides cao → chất lượng kém
* Yếu tố trung tính: Residual sugar không ảnh hưởng rõ đến chất lượng

***Câu hỏi mở rộng:*** Vậy outliers phân bố không đều giữa các features như thế nào? Features nào nhiều outliers nhất? Features nào ít outliers nhất?



Hình 3.1.4. Biểu đồ thanh trình bày sự phân bố outliers theo

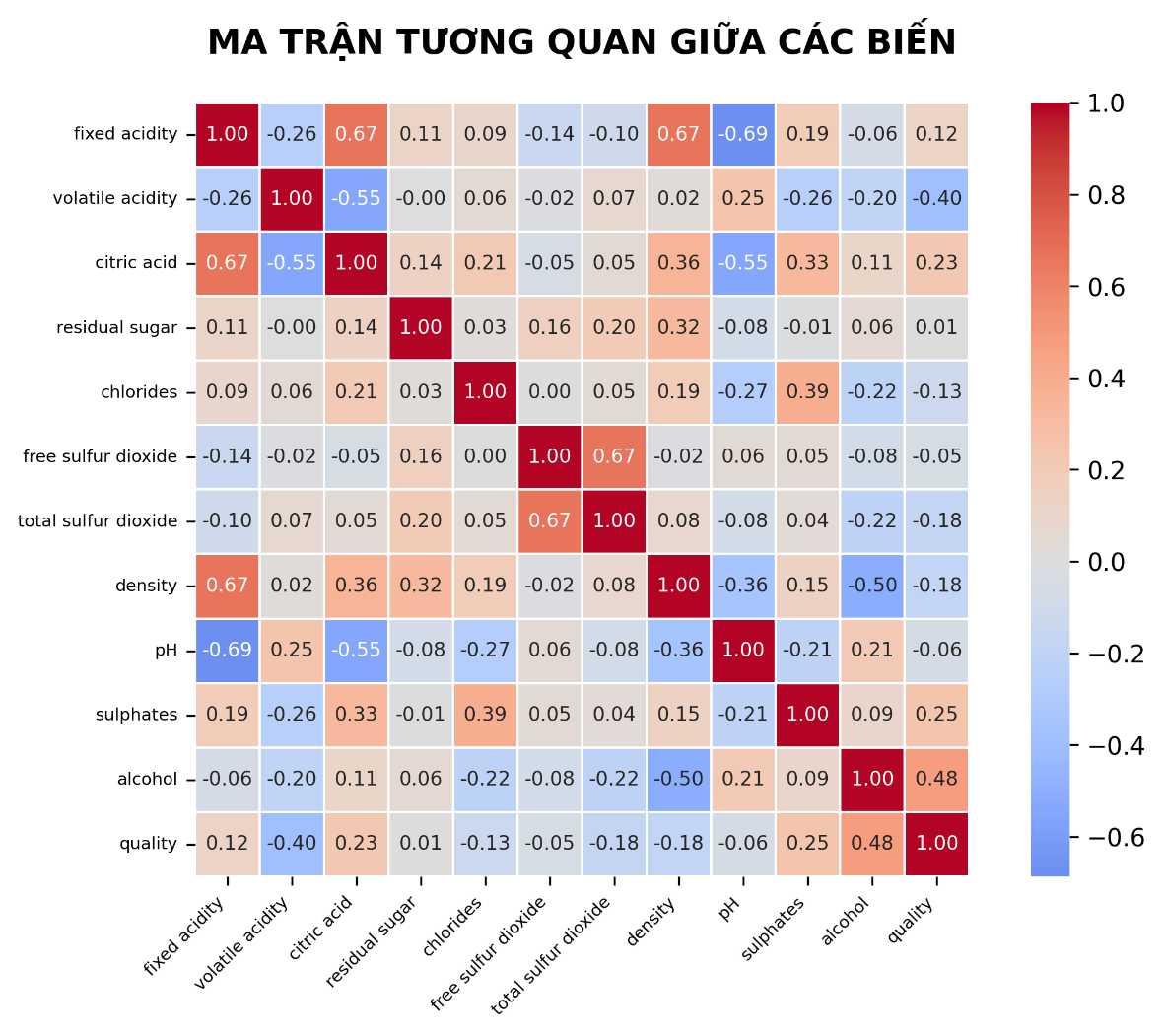
nhóm chất lượng của các features biến động cao

* **Kết quả phân tích:** Residual Sugar và Chlorides là hai features có số outliers cao nhất, cho thấy sự xuất hiện của những mẫu rượu có đặc tính cực kỳ khác biệt về độ ngọt và vị mặn so với đa số còn lại. Trong khi đó, Citric Acid có số outliers thấp nhất trong nhóm, dù có CV cao nhất.
* **Ý nghĩa:** điều này cho thấy CV cao không nhất thiết đồng nghĩa với nhiều outliers. Citric Acid có độ biến động cao nhưng phân bố tương đối đồng đều, trong khi Residual Sugar và Chlorides có cả độ biến động cao và tập trung nhiều giá trị cực đoan.
* **Các insights quan trọng sau quá trình phân tích đơn biến:**
* Thực trạng: 86.5% rượu chất lượng thấp vs chỉ 13.5% rượu chất lượng cao. Thị trường đang tồn tại sự mất cân bằng nghiêm trọng về chất lượng.
* Yếu tố tích cực: Citric acid cao có thể thúc đẩy chất lượng của rượu. Đây là yếu tố quan trọng nhất cần duy trì ở mức cao.
* Yếu tố tiêu cực: Sulfur dioxide (Tổng và Tự do) cao, Chlorides cao có thể ảnh hưởng xấu đến chất lượng.
* Yếu tố trung tính: Residual sugar không ảnh hưởng rõ đến chất lượng. Có thể điều chỉnh linh hoạt theo khẩu vị.
* Sự thiếu kiểm soát chất lượng trong sản xuất thể hiện rõ nhất qua độ biến động cao (CV > 40%) và nhiều outliers ở Chlorides và Residual Sugar, trong khi Citric Acid dù biến động mạnh nhưng ít outliers hơn, cho thấy CV cao không đồng nghĩa với chất lượng kém mà phụ thuộc vào bản chất phân bố của từng chỉ tiêu.

### 3.2. Phân tích đa biến (Multivariate Analysis)

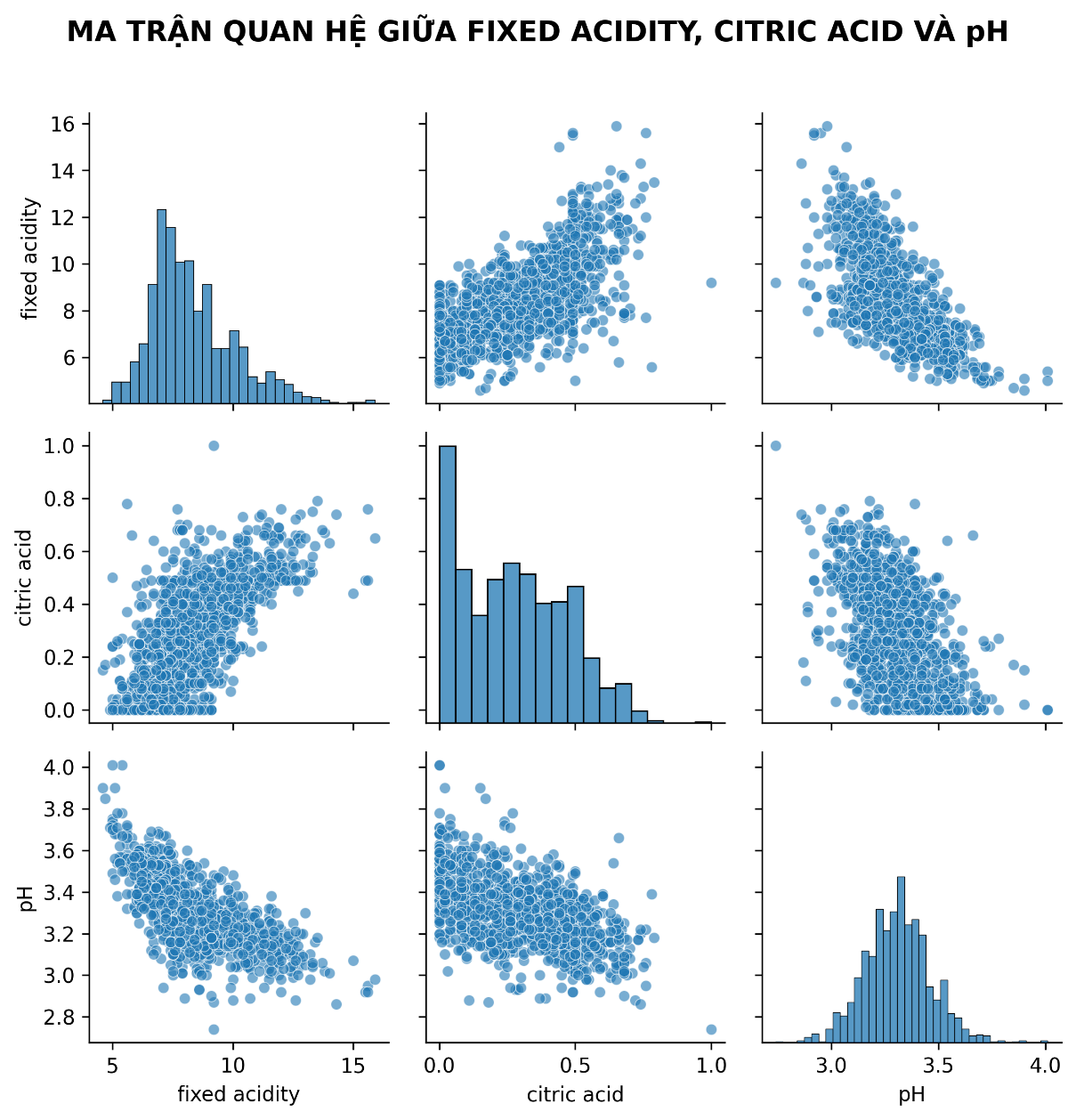
***Câu hỏi đặt ra:*** Các thuộc tính ảnh hưởng như thế nào đến điểm chất lượng (quality) của rượu?

* Thuộc tính nào giúp nâng chất lượng?
* Thuộc tính nào làm giảm chất lượng?
* Thuộc tính nào không ảnh hưởng đến chất lượng?



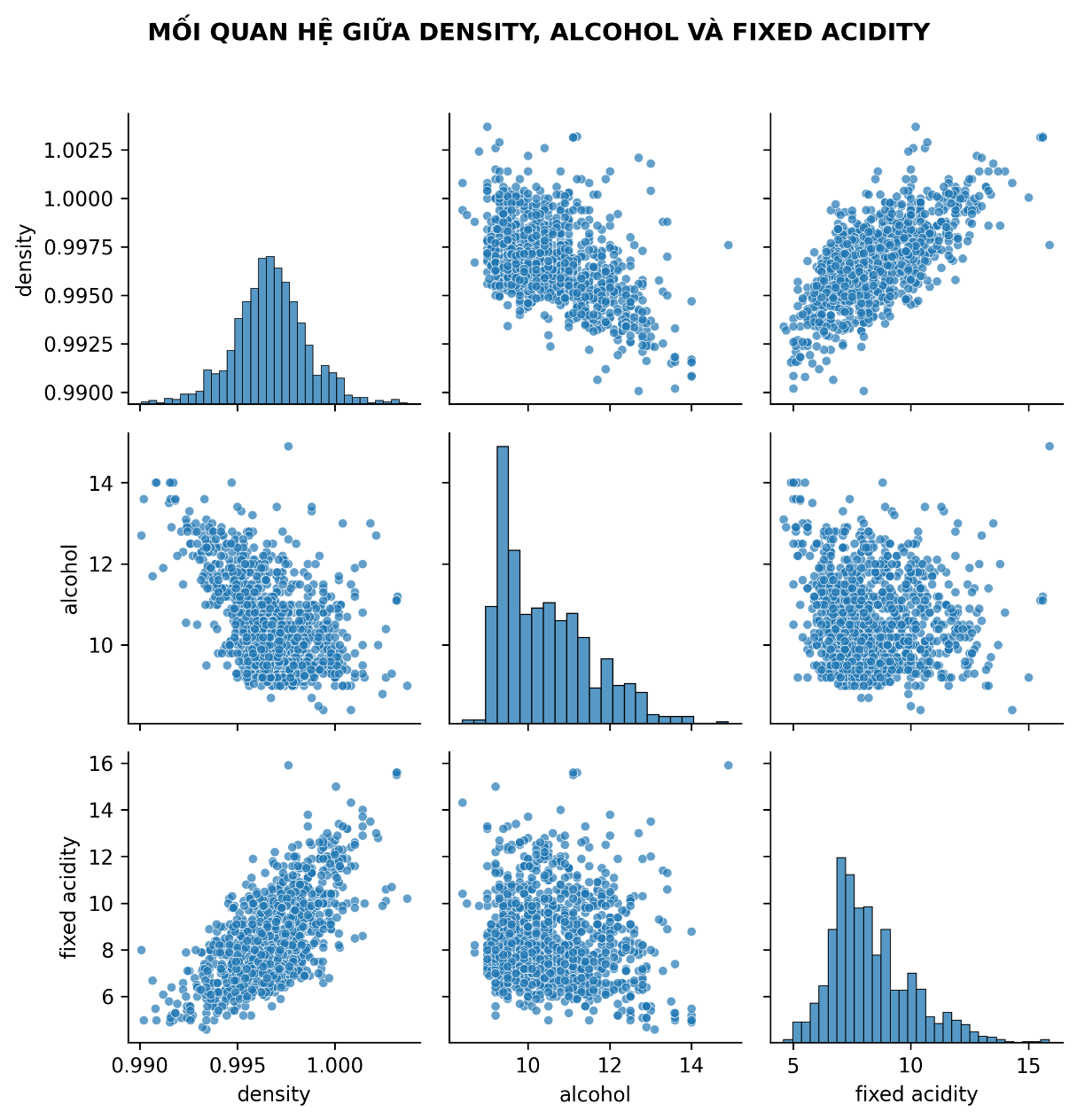
Hình 3.2.1. Biểu đồ ma trận tương quan giữa các biến

* Kết quả phân tích: các thuộc tính có mối quan hệ với chất lượng (quanlity):
* Alcohol (Nồng độ cồn) có mối tương quan mạnh với quality (+0.48), cho thấy hàm lượng cồn ảnh hưởng khá lớn đến chất lượng của rượu. Rượu có hàm lượng cồn cao thường có chất lượng tốt hơn.
* Volatile acidity(độ axit dễ bay hơi) có tương quan âm khá cao với quality (-0.39), cho thấy nếu volatile acidity tăng, có thể làm giảm chất lượng của rượu, vì sự biến đổi nhanh chóng của axit dễ làm rượu bị chua, có mùi giấm.
* Residual sugar(lượng đường dư) gần như không có mối quan hệ nào với quanlity (0.01), dù rằng một chút đường dư có thể cải thiện cảm giác ngọt ngào trong rượu. Tuy nhiên, mối tương quan gần như không có, cho thấy rằng mức độ đường dư không phải yếu tố quyết định trong việc đánh giá chất lượng của rượu.
* Ngoài ra, nhận thấy các thuộc tính có độ tương quan cao sau:
  + Fixed acidity, citric acid và pH (> 0.54)
  + Total sulfur dioxide và free sulfur dioxide (+0.67)
  + Density và alcohol (-0.50), density và fixed acidity (+0.67)
  + citric acid và volatile acidity (-0.55)
* Tiếp tục đào sâu phân tích các thuộc tính có độ tương quan cao
* Phân tích mức độ tương quan của nhóm Axit (fixed acidity, citric acid và pH).
  + fixed acidity và quality (+0.12)
  + citric acid và quality (+0.23)
  + pH và quanlity (-0.06)



Hình 3.2.2. Biểu đồ Pair plot mối quan hệ nhóm Axit

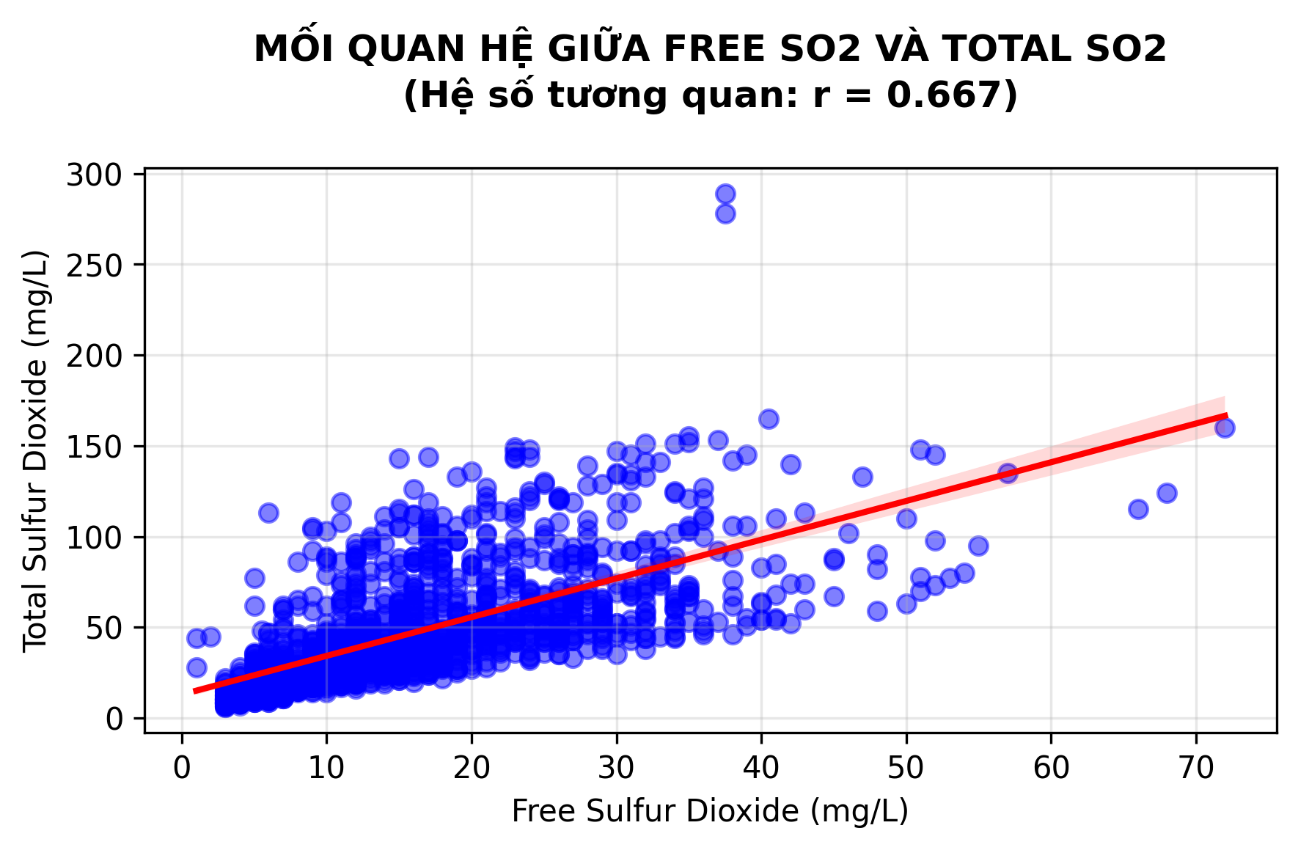
* **Kết quả phân tích:** Biểu đồ cho thấy một vòng tròn logic: Fixed Acidity và Citric Acid cùng tăng giảm với nhau (dương tính). Cả hai cùng 'kéo' giá trị pH đi xuống (âm tính). Điều này khẳng định chúng là một nhóm gắn bó chặt chẽ, cùng mô tả đặc tính về độ chua của rượu.
* **Điểm khác biệt:** citric acid có tương quan dương nhẹ (+0.23) với chất lượng, trong khi fixed acidity gần như không có tương quan (+0.12). Vì thế mà đối với chất lượng, loại axit quan trọng hơn tổng lượng axit. Một chai rượu tốt không cần phải có độ axit cố định (fixed acidity) thật cao, mà cần có sự cân bằng phù hợp, trong đó axit citric (citric acid) là axit "tốt" mang lại vị tươi sáng - đóng vai trò then chốt.
* Phân tích mối quan hệ giữa Mật độ (Density) với các thành phần chính của rượu:
  + Density và quanlity (-0.17)
  + Alcohol và quanlity (0.48)
  + Fixed acidity và quanlity (-0.12)



Hình 3.2.3. Biểu đồ Pair plot mối quan hệ giữa

mật độ và các thành phần chính của rượu

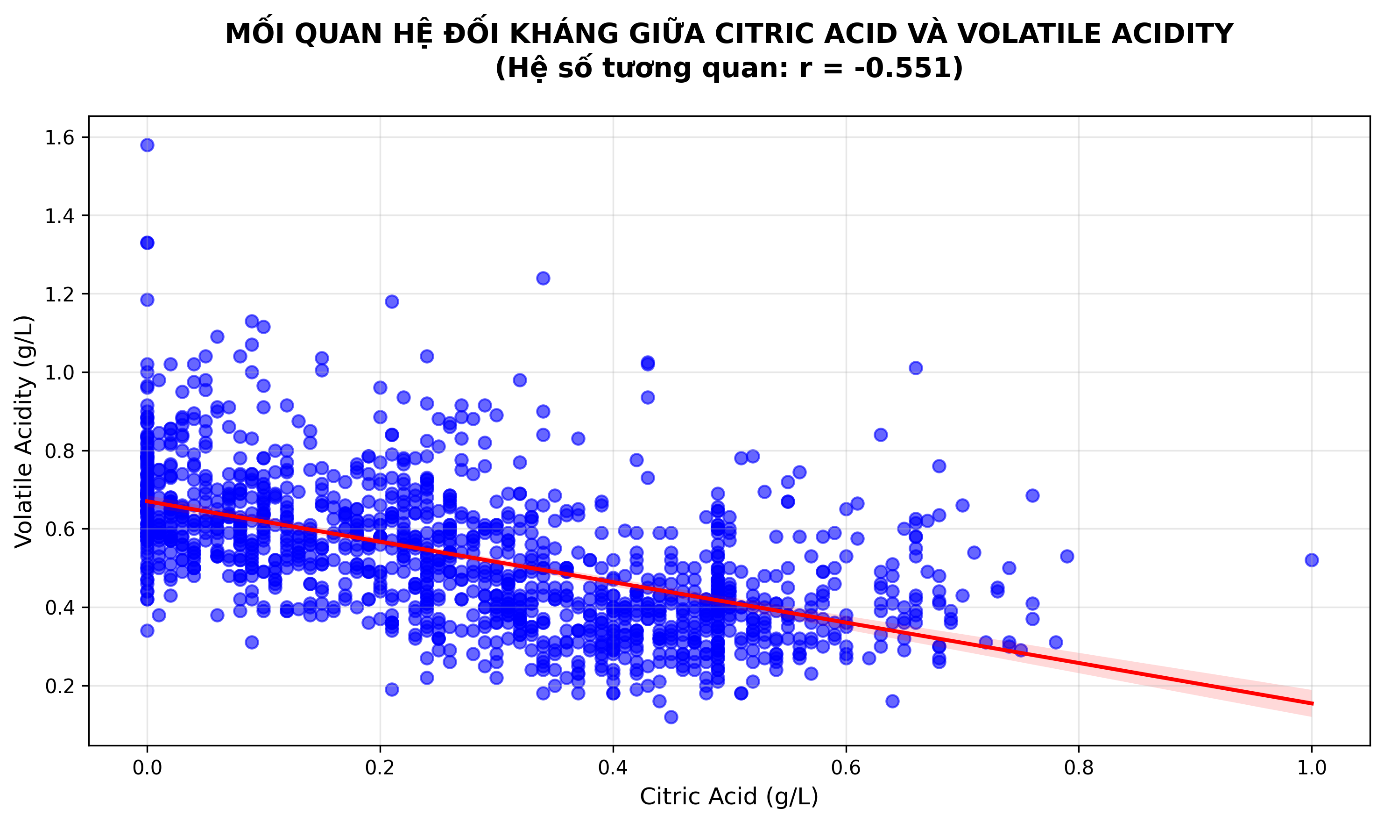
* **Kết quả phân tích:** Biểu đồ cho thấy mật độ (density) của rượu bị kéo giữa hai thứ: cồn (alcohol) và axit (fixed acidity). Cồn càng nhiều thì rượu càng nhẹ (density giảm), ngược lại, axit càng nhiều thì rượu càng nặng (density tăng). Density không phải là thứ quyết định chất lượng, mà nó là kết quả của việc cồn và axit trong chai rượu đó nhiều hay ít.
* **Điểm khác biệt:** Mặc dù cùng ảnh hưởng đến density, nhưng chỉ có cồn (alcohol) mới thực sự làm rượu ngon hơn. Axit (fixed acidity) gần như không giúp cải thiện chất lượng. Vì vậy, một chai rượu có density thấp thường là một chai rượu ngon. Lý do rất đơn giản: density thấp đồng nghĩa với việc cồn trong chai đó cao - và cồn cao chính là dấu hiệu quan trọng nhất của rượu chất lượng.
* Phân tích mức độ tương quan của nhóm Lưu huỳnh đioxit (Sulfur Dioxide):
  + total sulfur dioxide và quanlity (-0.19)
  + free sulfur dioxide và quanlity (-0.05)



Hình 3.2.4. Biểu đồ Scatter Plot thể hiện mối quan hệ

giữa So2 tự do và So2 tổng

* **Kết quả phân tích:** Biểu đồ cho thấy đồng biến dương rất mạnh và rõ rệt giữa free sulfur dioxide và total sulfur dioxide. Điều này là hoàn toàn hợp lý về mặt hóa học vì free SO₂ là một phần cấu thành của total SO₂. Khi tổng lượng SO₂ thay đổi, lượng SO₂ tự do cũng sẽ thay đổi theo một cách có thể dự đoán được.
* **Điểm khác biệt:** total sulfur dioxide và quanlity (-0.19), nên total SO₂ có tác động tiêu cực nhẹ. Điều này có thể là do ở hàm lượng rất cao, SO₂ có thể tạo ra mùi khó chịu.
* Phân tích mức độ tương quan giữa hai loại axit trong rượu:
  + citric acid và quanlity (0.23)
  + volatile acidity và quanlity (-0.39)



Hình 3.2.4. Biểu đồ Scatter Plot thể hiện mối quan hệ

đối kháng giữa citric acid và volatile acidity

* **Kết quả phân tích:** Biểu đồ cho thấy mối quan hệ "một mất, một còn" giữa hai loại axit: citric acid (axit citric) và volatile acidity (axit dễ bay hơi). Khi citric acid tăng thì volatile acidity giảm, và ngược lại. Hai loại axit này gần như không bao giờ cùng tồn tại ở mức cao trong một chai rượu.
* **Điểm khác biệt:** Mặc dù cả hai đều là axit, nhưng ảnh hưởng của chúng lên chất lượng hoàn toàn trái ngược. Citric acid hỗ trợ tích cực cho chất lượng (+0.23), trong khi volatile acidity là yếu tố tiêu cực chính (-0.39). Do đó, một chai rượu chất lượng cao thường có citric acid cao đi kèm với volatile acidity thấp - sự kết hợp này đảm bảo hương vị tươi sáng và hạn chế các tạp chất gây mùi khó chịu.

## 4. Kết luận và đề xuất

### 4.1. Các phát hiện chủ chốt

* Hồ sơ một chai rượu CHẤT LƯỢNG TỐT
* Một chai rượu được đánh giá cao thường sở hữu nhóm đặc tính then chốt sau:
  + Nồng độ cồn (alcohol) cao: Đây là yếu tố dự báo tích cực mạnh mẽ và quan trọng nhất.
  + Độ axit dễ bay hơi (volatile acidity) thấp: Đảm bảo rượu không có mùi giấm hoặc mùi khó chịu.
  + Hàm lượng axit citric (citric acid) ở mức trung bình đến cao: Góp phần tạo nên vị tươi sáng, cân bằng và có liên quan đến việc kìm hãm sự phát triển của axit dễ bay hơi.
* Dấu hiệu nhận biết một chai rượu CHẤT LƯỢNG KÉM
* Ngược lại, một chai rượu kém chất lượng thường mắc phải một hoặc nhiều lỗi sau:
  + Độ axit dễ bay hơi (volatile acidity) quá cao: Đây là "lỗi" phổ biến và nghiêm trọng nhất.
  + Nồng độ cồn (alcohol) thấp: Thiếu đi sự đậm đà và cấu trúc.
  + Hàm lượng muối (chlorides) cao: Có thể khiến rượu có vị mặn hoặc đắng khó chịu. ( ảnh hưởng ít )

### 4.2. Đề xuất Giải pháp cho Nhà sản xuất

* Tập trung kiểm soát quy trình lên men: Để hạn chế tối đa sự hình thành axit dễ bay hơi (volatile acidity) - "kẻ phá hoại" số một.
* Tận dụng mối quan hệ đối kháng: Giữa axit citric và axit dễ bay hơi, bằng cách duy trì một hàm lượng citric acid phù hợp để ổn định chất lượng.
* Tối ưu hóa để đạt nồng độ cồn (alcohol) phù hợp: Thay vì tập trung vào các chỉ số ít ảnh hưởng như đường dư (residual sugar).

# CHƯƠNG 2. PHÂN TÍCH DỮ LIỆU BÁN HÀNG SIÊU THỊ

## Bối cảnh vấn đề và mục tiêu phân tích

### 1.1. Bối cảnh vấn đề

Trong ngành bán lẻ siêu thị, việc phân tích dữ liệu giao dịch là chìa khóa để hiểu rõ hành vi mua sắm, tối ưu hóa hoạt động và tăng cường doanh thu. Dữ liệu giao dịch bao gồm thông tin về đơn hàng, sản phẩm, số lượng, thời gian, giá cả, khách hàng và quốc gia. Tuy nhiên, dữ liệu thô cần được khai thác để đưa ra các insights thực tế, chẳng hạn như xác định xu hướng mua sắm, phân khúc khách hàng, và quản lý rủi ro. Phân tích này sẽ tập trung vào dữ liệu từ các giao dịch siêu thị để đưa ra các khuyến nghị chiến lược.

### 1.2. Mục tiêu phân tích

Chuyên đề phân tích này được thực hiện nhằm đạt được các mục tiêu chính sau:

* Khám phá xu hướng doanh thu:
* Phân tích doanh thu theo thời gian (tháng, ngày trong tuần, giờ trong ngày) để xác định các thời điểm cao điểm và thấp điểm.
* Xác định yếu tố ảnh hưởng đến doanh thu: Tìm hiểu mối quan hệ giữa các biến như sản phẩm, khách hàng, quốc gia và doanh thu.
* Cung cấp insights thực tế: Đề xuất các hành động cụ thể để tối ưu hóa marketing, giảm thiểu rủi ro và tăng cường trải nghiệm khách hàng.

### 1.3. Giới thiệu tổng quan về Dataset

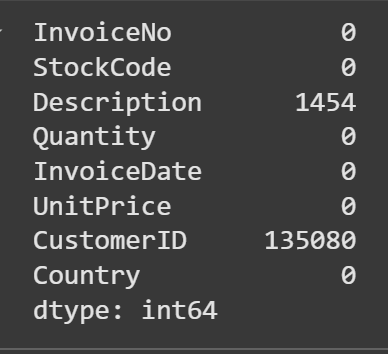
Dữ liệu được sử dụng bao gồm các giao dịch bán hàng từ siêu thị, với các cột chính sau:

* InvoiceNo: Mã số hóa đơn, định danh cho mỗi giao dịch.
* StockCode: Mã sản phẩm, định danh cho từng mặt hàng.
* Description: Mô tả sản phẩm.
* Quantity: Số lượng sản phẩm trong giao dịch.
* InvoiceDate: Ngày và thời gian giao dịch.
* UnitPrice: Giá per unit của sản phẩm.
* CustomerID: Mã khách hàng, định danh cho mỗi khách hàng.
* Country: Quốc gia của khách hàng.

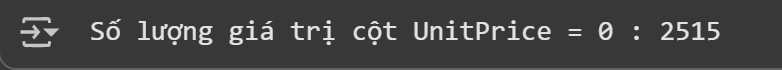
## 2. Xử lí dữ liệu

### 2.1. Kiểm tra và xử lí giá trị bị thiếu và không hợp lệ

* **Thực hiện kiểm tra:**

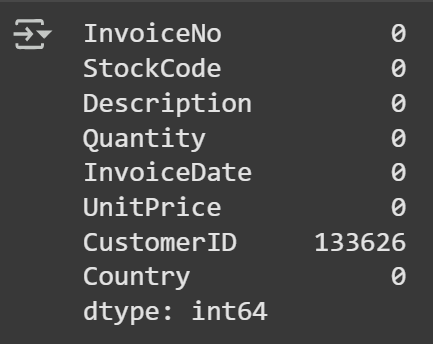


Hình 2.1.1. Hiển thị số lượng giá trị Null của từng cột



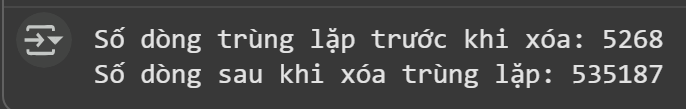
Hình 2.1.2. Hiển thị số lượng giá trị 0 không hợp lệ

* **Đánh giá**:
* có 2 trường hợp cần xử lý:
  + **Trường hợp 1:** Với 1454 dòng Description không có giá trị và UnitPrice = 0. Đây là các bản ghi thiếu thông tin mô tả. Lý do có thể Đồng bộ dữ liệu bị lỗi, mất mô tả, Sản phẩm không xác định hoặc Giao dịch hủy: Đơn hàng bị hủy nhưng còn trong hệ thống vì vậy cần xóa các giá trị này.
  + **Trường hợp 2:** Với 2515 cột Description CÓ giá trị, unitPrice = 0. Lý do có thể là, Sản phẩm miễn phí: Quà tặng, sample, sản phẩm khuyến mãi hoặc phí vận chuyển/dịch vụ hoặc các khoản phí được tính riêng vì vậy không cần xóa các giá trị này.
* Với cột CustomerID có 135080 dòng bị thiếu có nghĩa là các giao dịch này là của khách vãng lai hoặc các giao dịch mua tại quầy không định danh. Với các dữ liệu bị thiếu này vẫn có thể sử dụng các giao dịch này để phân tích.
* **Thực hiện xử lý:**

****

Hình 2.1.3. Thực hiện xử lí giá trị bị thiếu và không hợp lệ

### 2.2. Kiểm tra và xử lí giá trị trùng lặp

****

Hình 2.2.1. Thực hiện kiểm tra và xử lí giá trị trùng lặp

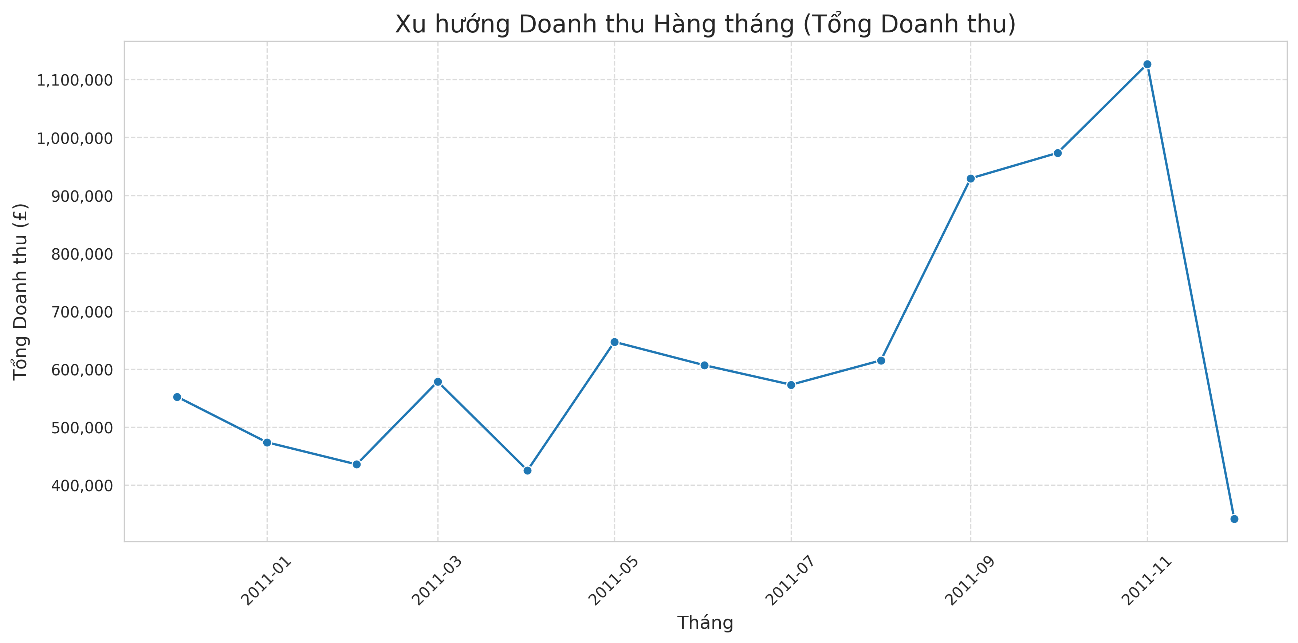
## 3. Phân tích khám phá và trực quan hóa dữ liệu

### 3.1. Phân tích đơn biến (Univariate Analysis)

Trước tiên, hãy cùng xem hoạt động kinh doanh của chúng ta theo dòng thời gian. Khi nào là thời điểm chúng ta bán hàng hiệu quả nhất?

#### 3.1.1. Phân tích Doanh số Hàng tháng (Xu hướng Bán hàng Tháng):

***Câu hỏi đặt ra:*** Xu hướng doanh thu hàng tháng cho chúng ta biết điều gì? Tháng nào đỉnh điểm, tháng nào trũng? Chúng ta có thể chuẩn bị gì cho những mùa cao điểm?

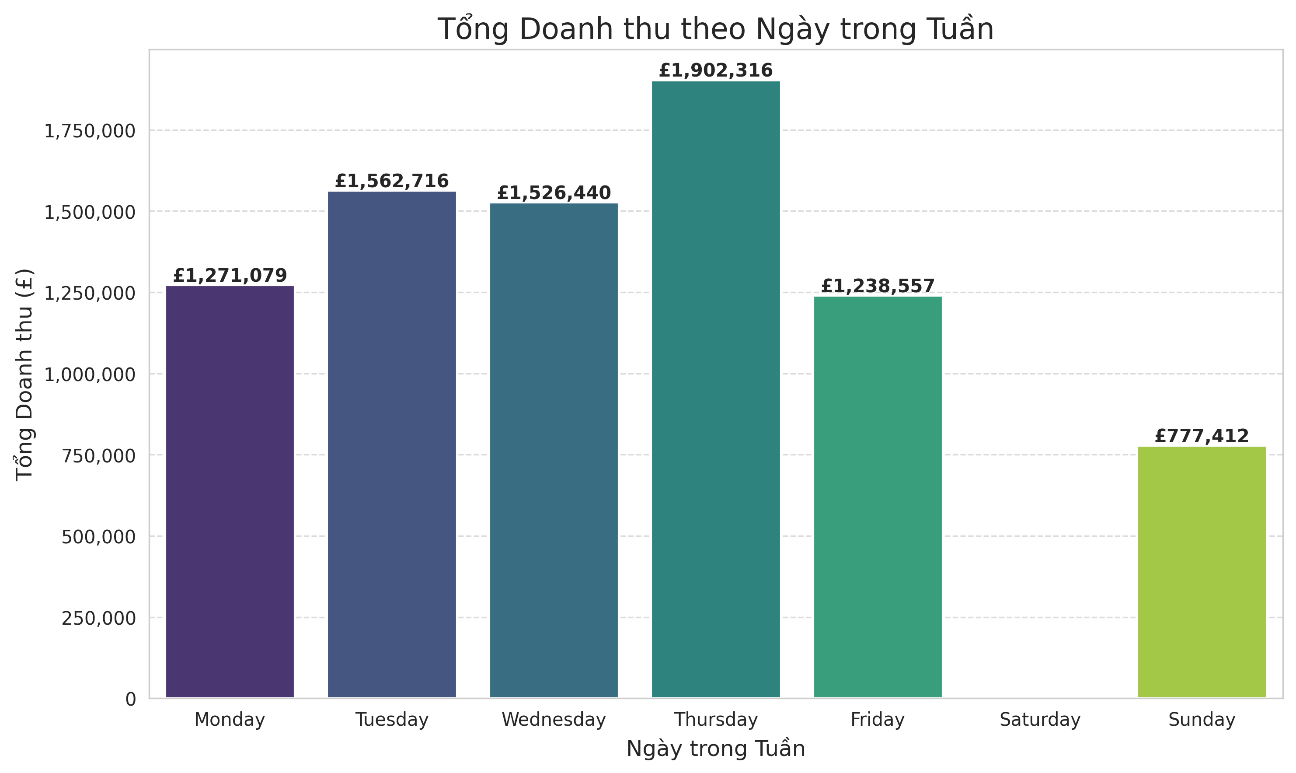


Hình 3.1.1.1. Biểu đồ đường thể hiện tổng doanh thu theo hàng tháng

* **Đánh giá:**
* Doanh thu đạt đỉnh cao nhất vào tháng 11 năm 2011 với hơn 1,5 triệu, trong khi tháng 2 cùng năm ghi nhận mức doanh thu thấp nhất là 523,084.
* Mô hình theo mùa: phần lớn doanh thu tập trung vào quý cuối cùng của năm (đặc biệt là gần cuối năm). Sau đỉnh điểm vào tháng 11, doanh thu tháng 12 sụt giảm mạnh (có thể do dữ liệu của tháng này không đầy đủ). Xu hướng thông thường là doanh thu sẽ giảm vào tháng 1 sau khi đạt đỉnh trong kỳ nghỉ lễ.

#### 3.1.2. Phân tích Doanh số theo Ngày trong Tuần

***Câu hỏi đặt ra:*** Doanh số theo ngày trong tuần nói lên điều gì? Khách hàng có thích mua sắm vào cuối tuần hơn không?

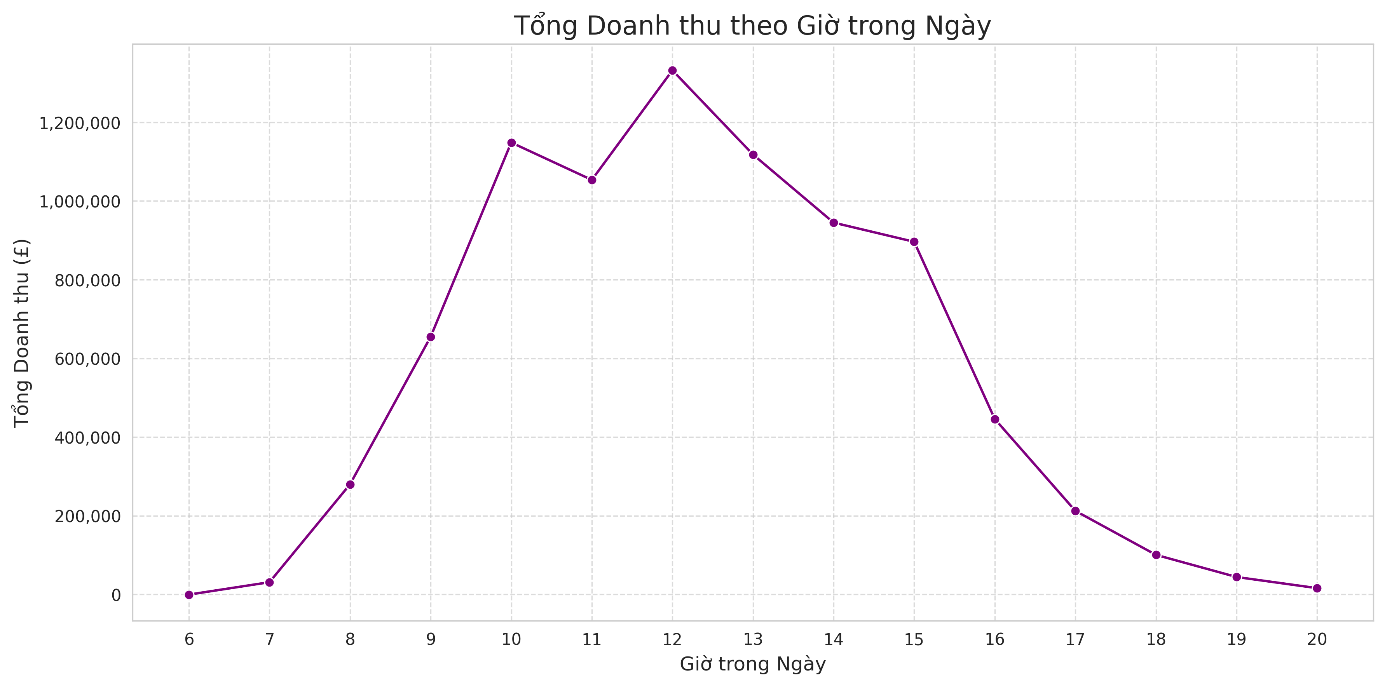


Hình 3.1.2.1. Biểu đồ cột thể hiện tổng doanh thu theo ngày trong tuần

* **Những phát hiện chính:**
* Nhịp điệu mua sắm: Khách hàng có xu hướng dành thời gian mua sắm nhiều hơn vào các ngày trong tuần.
* Thứ Năm là điểm đáng chú ý: Sự gia tăng doanh thu vào Thứ Năm có thể do:
  + Khách hàng mua sắm chuẩn bị cho cuối tuần.
  + Các chiến dịch khuyến mãi thường được triển khai vào thời điểm này.
* Tại sao doanh thu vào Chủ Nhật lại thấp như vậy? Nếu doanh nghiệp vẫn mở cửa, có thể nhu cầu mua sắm vào Chủ Nhật của khách hàng thấp một cách tự nhiên. Điều này có thể do thói quen của họ (chỉ mua sắm hoặc làm việc vào các ngày trong tuần) hoặc do ngành hàng này thường không có hoạt động sôi nổi vào cuối tuần.

#### 3.1.3. Phân tích Doanh số theo Giờ

***Câu hỏi đặt ra:*** Theo từng giờ trong ngày, khung giờ vàng để bán hàng là khi nào? Tại sao doanh thu lại có đỉnh điểm rõ rệt vào các khung giờ cao điểm?

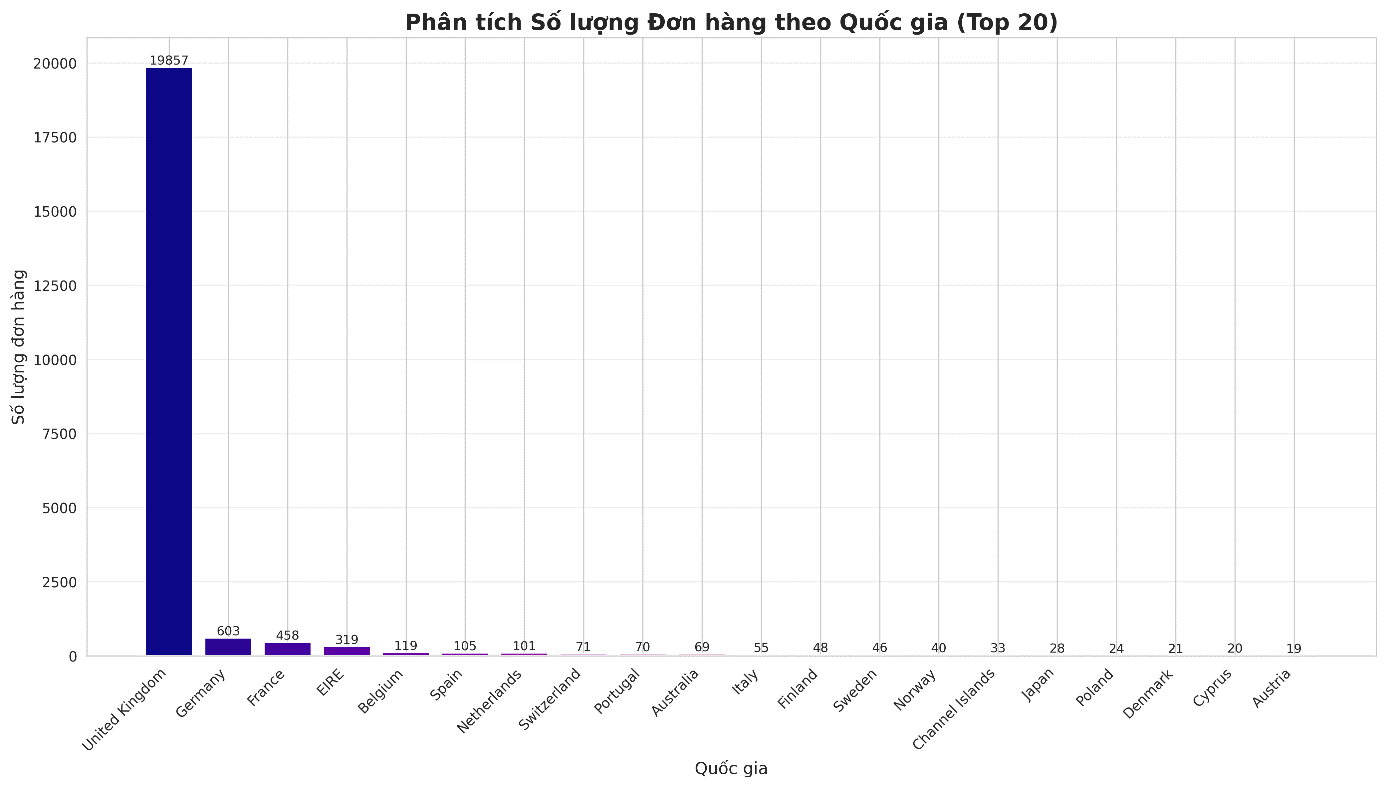


Hình 3.1.3.1. Biểu đồ đường thể hiện tổng doanh thu theo giờ trong ngày

* **Những phát hiện chính:**
* Thời gian tương đối cao điểm: Doanh thu duy trì ở mức cao từ 10 giờ sáng đến 15 giờ chiều, với một đỉnh nhỏ vào khoảng 15 giờ.
* Thời gian thấp điểm: Doanh thu giảm mạnh sau 15 giờ chiều và gần như bằng 0 sau 19 giờ tối.
* Giờ mở cửa và đóng cửa: Dựa vào biểu đồ, có thể thấy doanh nghiệp hoạt động từ khoảng 6 giờ sáng đến 20 giờ tối
* Tại sao doanh thu lại có hai đỉnh điểm rõ rệt vào 10 giờ sáng và 12 giờ trưa? Hai đỉnh này có thể phản ánh thói quen mua sắm của khách hàng. 2 khung giờ này trùng với giờ nghỉ trưa, khi nhiều người có thời gian để mua sắm hoặc ăn uống.

#### 3.1.4. Phân tích số lượng đơn hàng theo Quốc gia

***Câu hỏi đặt ra:*** Không chỉ về thời gian, chúng ta cũng cần biết: Chúng ta đang bán hàng cho ai và ở đâu? Phân bố đơn hàng theo quốc gia cho thấy thị trường trọng điểm nào đang đóng góp nhiều nhất?



Hình 3.1.4.1. Biểu đồ cột thể hiện số lượng đơn hàng theo quốc gia

* **Phát hiện chính:**
* Khách hàng nội địa (Anh) chiếm ưu thế áp đảo: Với 16,649 đơn hàng, Vương quốc Anh là nguồn khách hàng chính của siêu thị. Điều này là hoàn toàn hợp lý và dễ hiểu khi cửa hàng được đặt tại đây, phản ánh đúng đặc điểm của một doanh nghiệp bán lẻ tại địa phương.
* Có sự hiện diện của khách hàng quốc tế: Mặc dù ở mức khiêm tốn, biểu đồ cho thấy siêu thị vẫn thu hút được khách hàng từ nhiều quốc gia châu Âu khác như Đức, Pháp, Ireland, Bỉ, Tây Ban Nha...

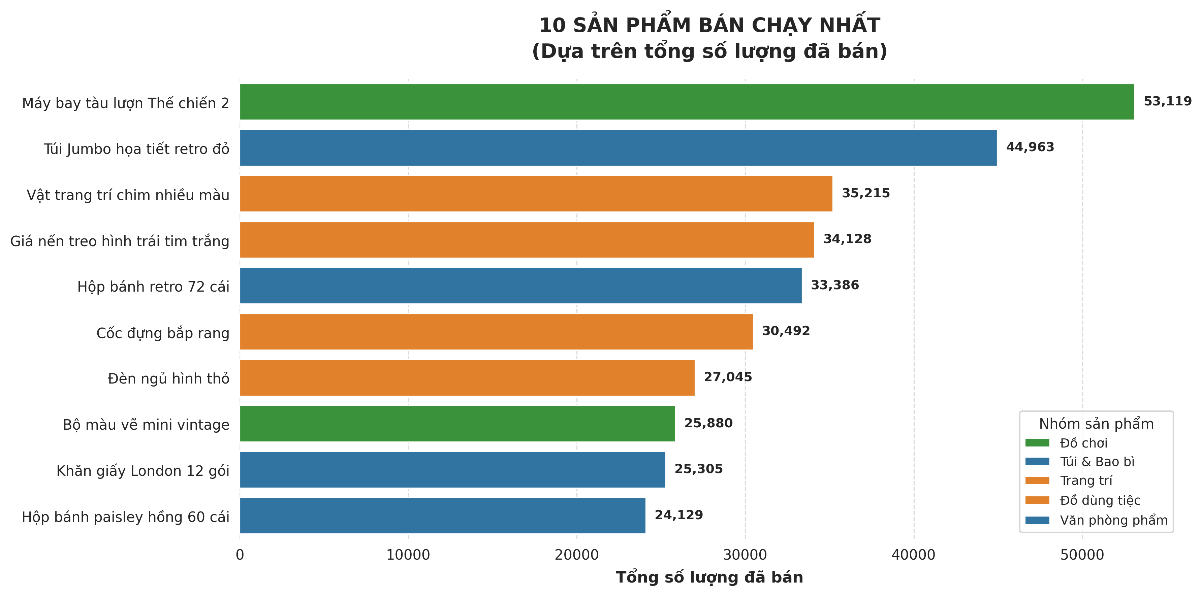
Sau khi đã hiểu về 'khi nào’ và ‘ở đâu’, một câu hỏi quan trọng hơn cần được trả lời: 'cái gì’ đang tạo ra doanh thu và 'cái gì’ đang gây ra vấn đề?

### 3.2. Phân tích đa biến (Multivariate Analysis)

***Câu hỏi lớn:*** Sản phẩm nào là người hùng, và sản phẩm nào là điểm yếu?

#### 3.2.1. Top 10 Sản phẩm theo Số lượng

***Câu hỏi đặt ra:*** Top sản phẩm bán chạy theo số lượng? Đâu là những mặt hàng phổ thông, có sức hút lớn.

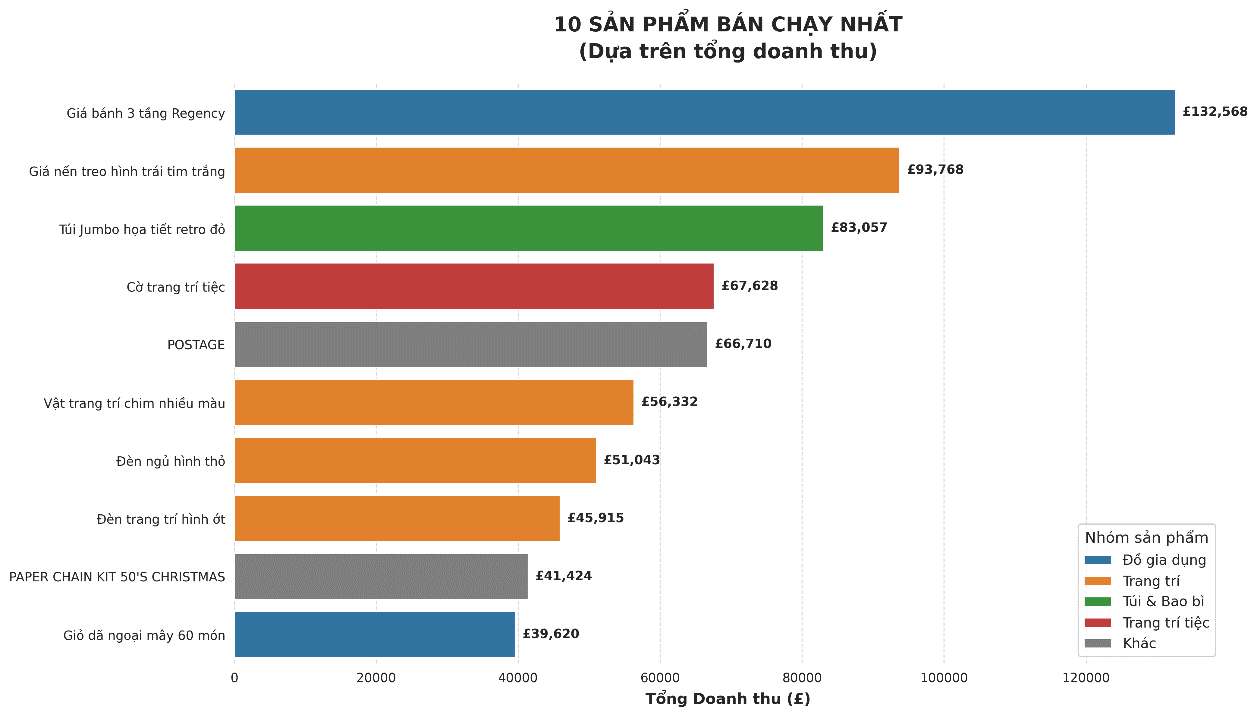


Hình 3.2.1.1. Biểu đồ cột thể hiện top 10 sản phẩm bán chạy nhất theo số lượng

* **Đánh giá:** Biểu đồ này cho thấy rõ ràng các sản phẩm có số lượng bán ra cao nhất, giúp xác định đâu là những mặt hàng chủ lực của doanh nghiệp.
* Đồ Chơi Dẫn Đầu: Sản phẩm "Máy bay tàu lượn Thế chiến 2" (WORLD WAR 2 GLIDERS ASSTD DESIGNS) đứng đầu với 53,119 sản phẩm đã bán, cho thấy sức hút mạnh mẽ của dòng đồ chơi mô hình lịch sử trong thị trường.
* Sự Thống Trị Của Sản Phẩm Trang trí: 4/10 sản phẩm thuộc nhóm trang trí .
* Túi & Bao Bì Có Sức Bền: 4/10 sản phẩm thuộc nhóm bao bì đóng gói.

#### 3.2.2. Sản phẩm Bán chạy nhất theo Doanh thu (Revenue/TotalPrice)

***Câu hỏi đặt ra:*** Nhưng liệu những sản phẩm bán chạy có phải là những sản phẩm mang lại doanh thu cao nhất không? Hay có những sản phẩm 'quý tộc' khác, dù bán ít nhưng giá trị mỗi đơn lại rất lớn?"

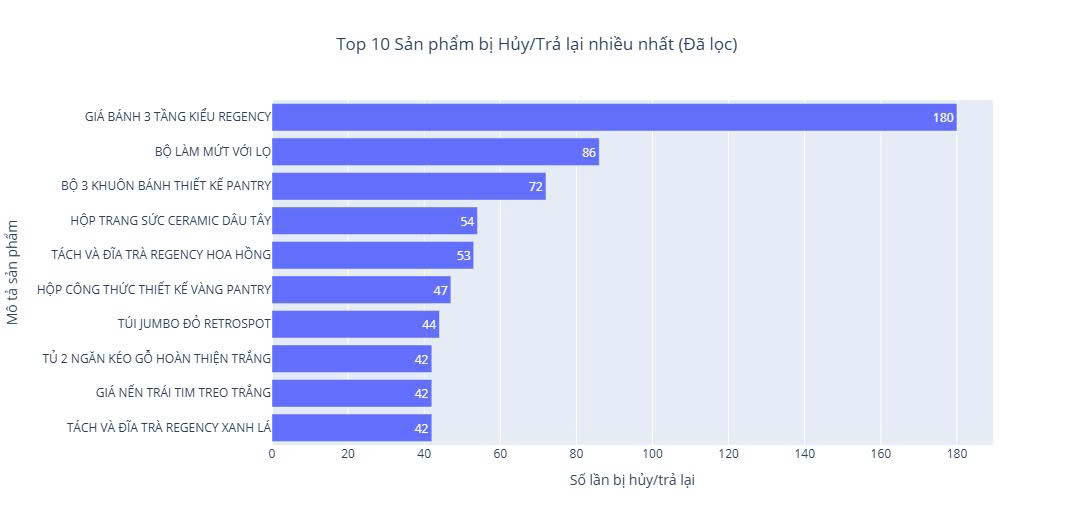


Hình 3.2.2.1. Biểu đồ cột thể hiện top 10 sản phẩm bán chạy nhất theo doanh thu

* **Những phát hiện chính:**
* Sản phẩm "Giá bánh 3 tần Regency " là sản phẩm mang lại doanh thu cao nhất. Điều này khẳng định đây là sản phẩm chủ lực, vừa được yêu thích rộng rãi vừa có giá trị cao.
* Sự Thống trị của Đồ trang trí & Gia dụng: Tương tự như phân tích theo số lượng, danh sách này tiếp tục cho thấy sức mạnh của các mặt hàng trang trí nhà cửa và quà tặng. Đây rõ ràng là danh mục sản phẩm cốt lõi của doanh nghiệp.

#### 3.2.3. Phân tích các đơn hàng bị hủy/trả lại

***Câu hỏi đặt ra:*** Bên cạnh những thành tích, chúng ta cũng cần nhìn vào sự thật: Đâu là những sản phẩm đang gây thất thoát doanh thu nhiều nhất do bị hủy/trả lại? Hiểu được lý do đằng sau sẽ giúp chúng ta cải thiện chất lượng sản phẩm, mô tả, hoặc dịch vụ



Hình 3.2.3.1. Biểu đồ cột thể hiện top 10 sản phẩm bị hủy trả lại nhiều nhất

* **Những phát hiện chính:**
* Dựa trên biểu đồ:
  + Sản phẩm "giá bánh 3 tầng kiểu regency" có doanh thu cao nhất nhưng đồng thời cũng là sản phẩm bị hủy/trả lại nhiều nhất. Điều này cho thấy đây là một sản phẩm rất phổ biến và bán chạy, nhưng có thể gặp vấn đề về chất lượng, mô tả sản phẩm không chính xác, hoặc hư hỏng trong quá trình vận chuyển.
  + Sản phẩm có rủi ro: Sản phẩm "Bộ làm mứt với lọ" cũng cần phải quan tâm đến chất lượng cần có các biện pháp để khắc phục
* Câu hỏi: lý do khách quan vì sao các sản phẩm này lại có tỉ lệ trả hàng cao?
  + Mua hàng theo Cảm hứng & Hối hận Ngay lập tức. Khách hàng thấy chúng dễ thương trên kệ và mua ngay. Tuy nhiên, chỉ vài phút sau, khi đang xếp hàng thanh toán hoặc ra về, họ bắt đầu suy nghĩ lại.
  + Lỗi Sản xuất & Hư hại có sẵn. Khách hàng có thể phát hiện ra một vết nứt nhỏ, vết rỗ men, hoặc hình in bị lệch khi kiểm tra kỹ tại cửa hàng. Họ sẽ bỏ sản phẩm đó lại và chọn sản phẩm khác
  + Chính sách Trả hàng Dễ dàng. Cửa hàng có thể có chính sách trả hàng rất thoải mái. Điều này vô tình khuyến khích hành vi mua về dùng thử, không ưng thì trả.

### 3.3. Phân tích RFM và Phân cụm

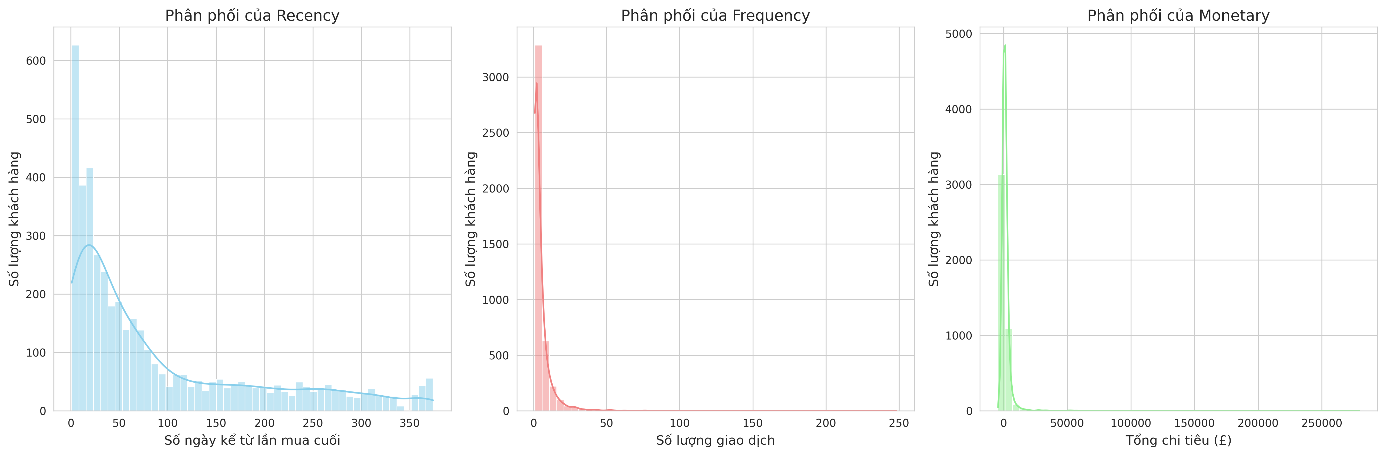
***Câu hỏi lớn:*** Chúng ta đã nắm rõ về sản phẩm và thời gian. Nhưng để phát triển bền vững, chìa khóa nằm ở chính khách hàng. Vậy, làm thế nào để chúng ta thực sự hiểu và phân loại được khách hàng của mình?

RFM Analysis giúp phân nhóm khách hàng dựa trên mức độ tương tác và giá trị mang lại, từ đó doanh nghiệp:

* Hiểu ai là khách hàng trung thành
* Ai đang có nguy cơ rời bỏ
* Và ai là khách hàng tiềm năng cần chăm sóc thêm.

Tính 3 chỉ số cho từng khách hàng:

* Recency (R): Ngày giao dịch gần nhất → tính số ngày kể từ lần mua gần nhất đến hiện tại. → Càng nhỏ càng tốt (mới mua gần đây).
* Frequency (F): Số lượng giao dịch (đơn hàng) trong một khoảng thời gian. → Càng lớn càng tốt.
* Monetary (M): Tổng tiền chi tiêu. → Càng lớn càng tốt.

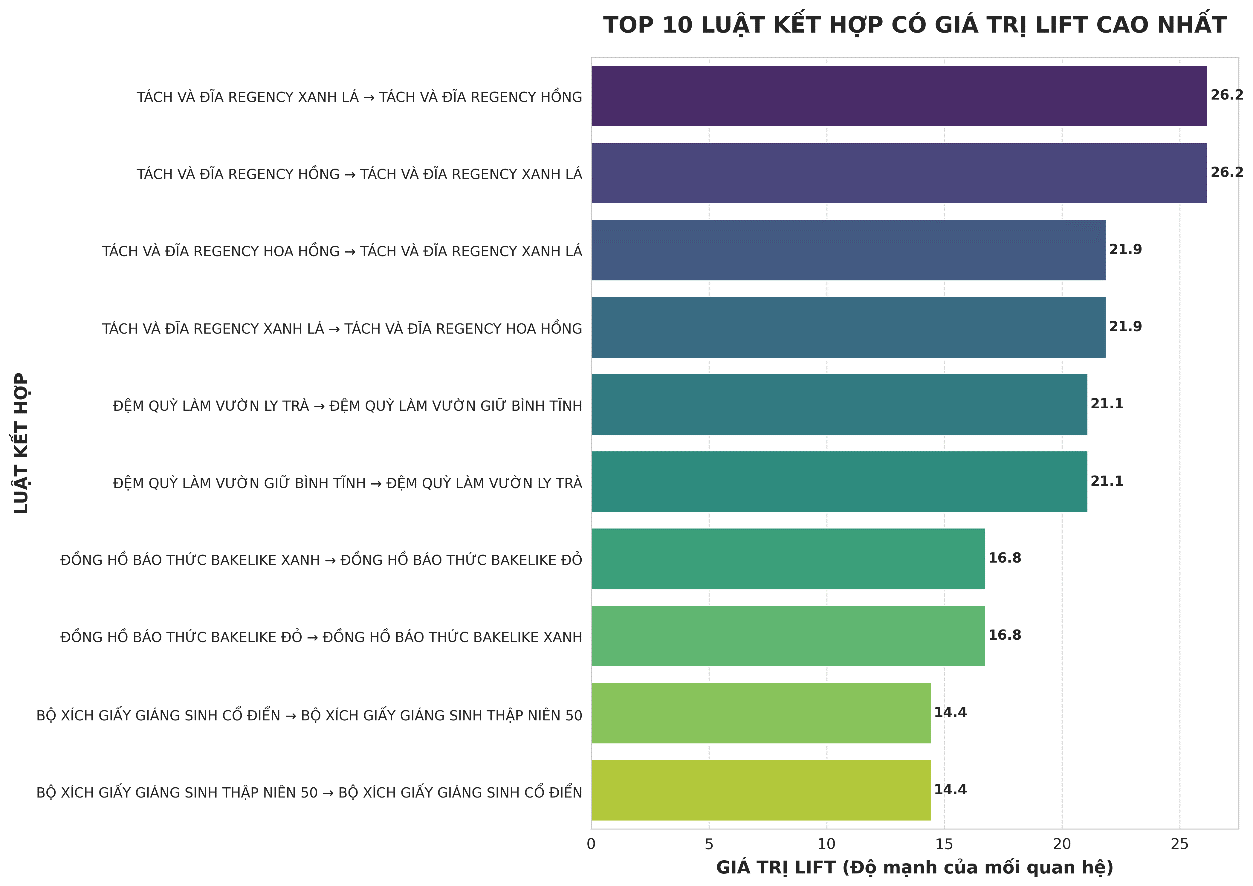


Hình 3.3.1. Biểu đồ thể hiện sự phân phối của 3 chỉ số R – F - M

* **Kết quả phân tích:**
* Phân phối của Recency
  + Nhận xét:
    - Phần lớn khách hàng có Recency nhỏ (0–50 ngày). → Nghĩa là đa số khách hàng vừa mới mua gần đây.
    - Một số ít khách hàng có Recency cao (200–350 ngày). → Đây là nhóm lâu không quay lại, có nguy cơ rời bỏ.
  + Kết luận:
    - Tỷ lệ khách hàng hoạt động gần đây cao, đây là dấu hiệu tích cực.
    - Tuy nhiên, cũng cần tái kích hoạt nhóm Recency cao (gửi ưu đãi, email…).
* Phân phối của Frequency
  + Nhận xét:
    - Đa số khách hàng chỉ mua 1–5 lần.
    - Rất ít khách hàng mua hơn 50 lần (đuôi bên phải kéo dài).
  + Kết luận:
    - Phần lớn là khách hàng mua ít / không thường xuyên.
    - Cần chiến lược chăm sóc và khuyến khích mua lại (loyalty, voucher...).
* Phân phối của Monetary
  + Nhận xét:
    - Rất nhiều khách hàng chi tiêu ít tiền, nằm ở vùng thấp (gần 0).
    - Một vài khách hàng chi tiêu cực lớn (đuôi dài sang phải).
  + Kết luận:
    - Doanh thu chủ yếu đến từ một nhóm nhỏ khách hàng chi tiêu cao.
    - Nên tập trung chăm sóc đặc biệt nhóm này vì họ đóng góp phần lớn doanh thu.

#### 3.3.1. Phân tích Giỏ hàng (Market Basket Analysis)

***Câu hỏi đặt ra:*** Trước khi đi sâu vào từng cá nhân, hãy khám phá hành vi mua sắm của họ: Những sản phẩm nào thường được mua cùng nhau?



Hình 3.3.1.1. Biểu đồ cột thể hiện top 10 luật kết hợp có giá trị LIFT cao nhất

* **Những phát hiện chính:**
* Xu hướng mua hàng rõ rệt: Khách hàng có xu hướng mua các sản phẩm này cùng nhau một cách có chủ đích.
* Đặc điểm chung của các cặp sản phẩm: Hầu hết là các sản phẩm bổ sung, thay thế hoặc cùng thuộc một bộ sưu tập. Cụ thể:
  + Sản phẩm cùng loại, khác mẫu mã/màu sắc: Tách trà Regency (màu hồng, xanh lá, họa tiết hoa hồng), Đồng hồ báo thức Bakelike (đỏ & xanh), Bộ đồ trang trí Giáng sinh (Vintage & 50's).
  + Sản phẩm cùng chủ đề, khác kiểu dáng: Tấm quỳ làm vườn (Kiểu "Cup of Tea" & "Keep Calm").
* **Gợi ý Hành động:**
* Phát hiện này cung cấp những gợi ý rất cụ thể và dễ thực hiện để thúc đẩy doanh số:
  + Tạo Combo Khuyến mãi: Xây dựng các gói sản phẩm như "Mua bộ 3 tách trà được giảm giá" hoặc "Mua 2 tặng 1" để khuyến khích khách hàng mua trọn bộ.
  + Gợi ý Sản phẩm (Cross-selling): Trên trang web, khi khách hàng xem sản phẩm tách trà màu Hồng, hãy hiển thị gợi ý "Thường được mua cùng" với màu Xanh và màu Hoa hồng ngay bên cạnh.
  + Chiến dịch Marketing: Triển khai các chiến dịch email hoặc quảng cáo nhắm vào những khách hàng đã mua một màu, khuyến khích họ "hoàn thành bộ sưu tập" của mình.
  + Trưng bày Sản phẩm: Tại cửa hàng vật lý, hãy đặt các sản phẩm này cạnh nhau để khách hàng dễ dàng nhìn thấy và lựa chọn cả bộ.

#### 3.3.2. Kmeans

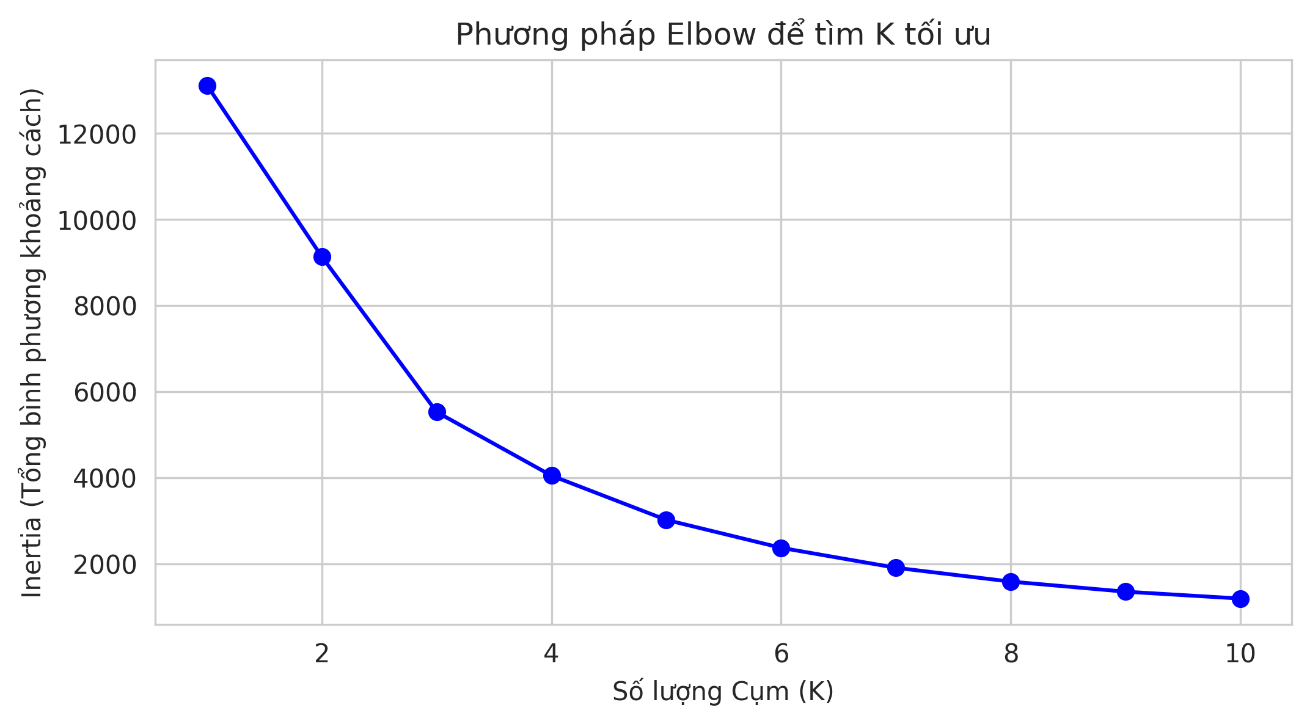
Bây giờ, ta sử dụng mô hình RFM để phân loại khách hàng một cách khoa học.

##### 3.3.2.1. Chuẩn hóa dữ liệu RFM

#### 

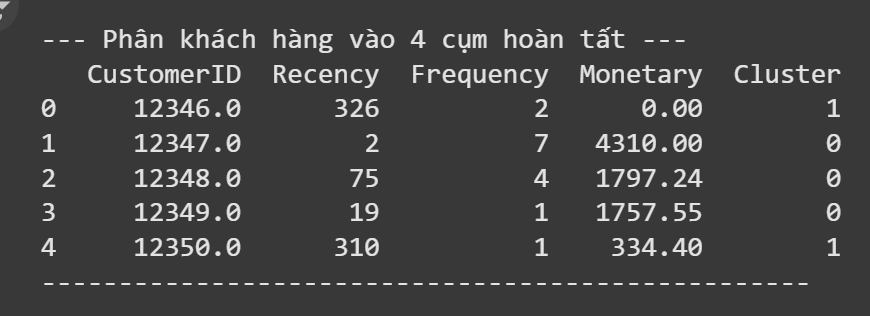
#### Hình 3.3.2.1. Thực hiện chuẩn hóa dữ liệu RFM

##### 3.3.2.2. Tìm số cụm tối ưu (K) bằng phương pháp ELBOW



#### Hình 3.3.2.2. Hiển thị số cụm tối ưu

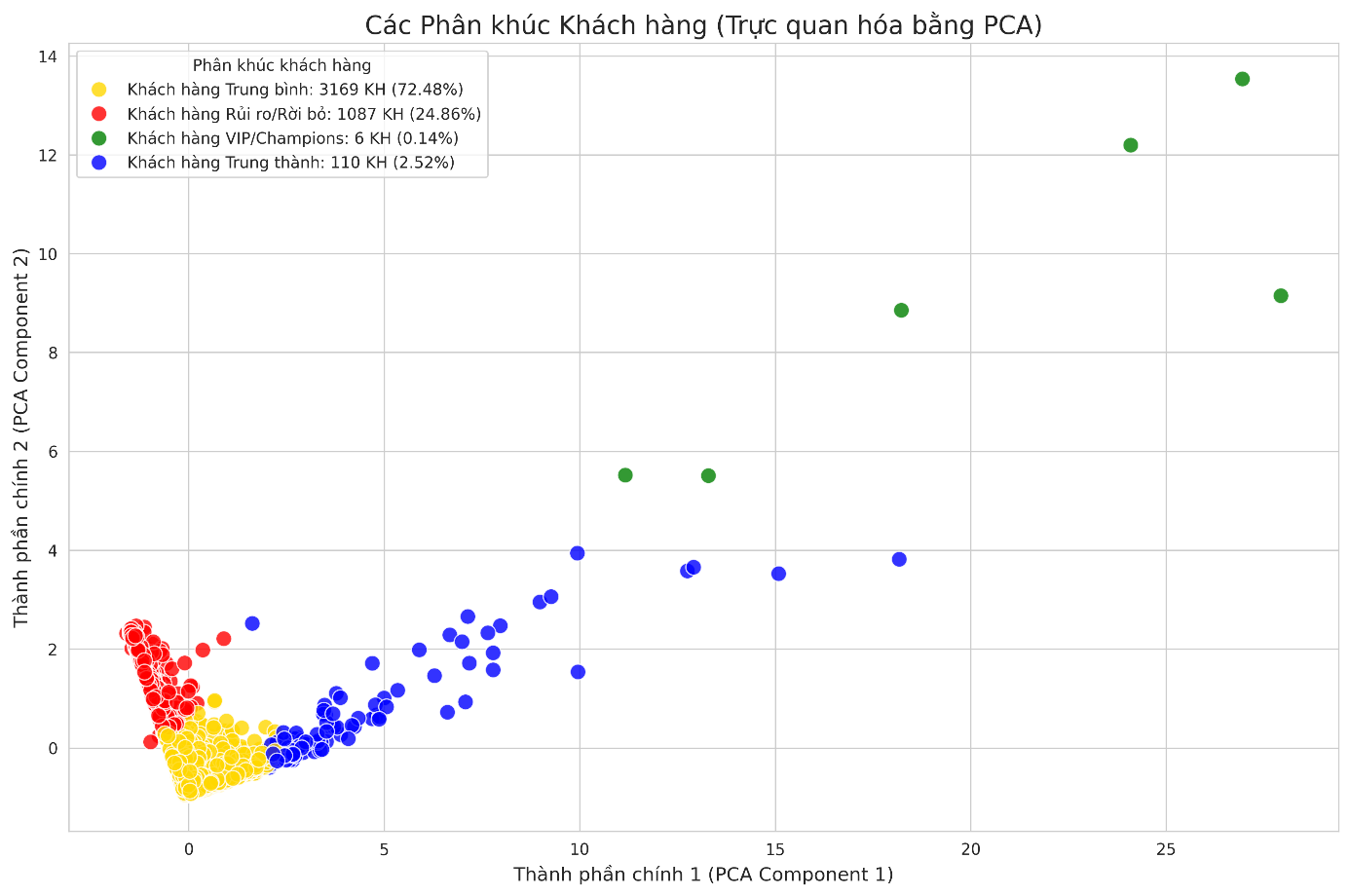
##### 3.3.2.3. Áp dụng K-MEANS và gán nhãn cụm



#### Hình 3.3.2.3. Thực hiện phân khách hàng thành các cụm

##### 3.3.2.4. Phân tích diễn giải các cụm và trực quan hóa

***Câu hỏi đặt ra:*** Mỗi nhóm khách hàng này có đặc điểm gì? Ai là 'VIP', ai là 'Khách hàng tiềm năng', và ai đang 'ngủ quên'? Mỗi nhóm cần một chiến lược chăm sóc khác nhau.



#### Hình 3.3.2.4. Biểu đồ Scatter Plot hiển thị phân khúc khách hàng

* **Đánh Giá Tình Hình Hiện Tại:** nhìn chung cơ sở khách hàng đang ở trạng thái cảnh báo, với một tỷ lệ rất lớn khách hàng sắp rời bỏ và một nhóm khách hàng trung thành/VIP quá mỏng.
* Điểm BÁO ĐỘNG ĐỎ: "Khách hàng Rủi ro/Rời bỏ" chiếm tới 24.87%.
  + Đây là một con số cực kỳ lớn, cho thấy gần 1/4 tổng số khách hàng đang không hài lòng, ít giao dịch hoặc sắp ngừng mua hàng.
* Điểm YẾU: Nhóm khách hàng giá trị cao quá ít.
  + Khách hàng VIP/Champions chỉ có 0.14% (6 người): Nhóm này là "con gà đẻ trứng vàng" nhưng lại cực kỳ ít.
  + Khách hàng Trung thành chỉ có 2.52%: Tỷ lệ này cho thấy khả năng giữ chân khách hàng (Customer Retention) đang rất kém.
* Điểm CẦN CẢI THIỆN: Nhóm "Khách hàng Trung bình" chiếm đa số (72.48%).
  + Đây là nhóm tiềm năng nhất để "nuôi dưỡng" và nâng cấp lên các phân khúc cao hơn. Hiện tại, họ đang ở trạng thái trung lập và dễ dàng chuyển sang nhóm "Rủi ro" nếu không được chăm sóc tốt.

## 4. Kết luận

* Dựa vào biểu đồ xu hướng doanh thu hàng tháng, ta có thể thực hiện một số hành động sau:
  + Phát triển chiến lược marketing: Dựa vào tháng 11 có doanh thu cao, ta có thể phát triển các chiến dịch marketing hiệu quả hơn, tập trung vào các sản phẩm bán chạy hoặc các sự kiện đã tạo ra doanh thu tốt.
  + Quản lý kho hàng: Đảm bảo đủ hàng trong các tháng cao điểm để tránh tình trạng hết hàng và mất doanh thu. Ngược lại, tránh tồn kho quá nhiều trong các tháng thấp điểm.
* Dựa vào doanh thu theo Ngày trong Tuần ta có thể thực hiện kế hoạch kinh doanh như sau
  + Tối ưu hóa các ngày cao điểm (Thứ Ba & Thứ Năm)
    - Tăng cường sản phẩm: Tập trung cung cấp các sản phẩm hoặc dịch vụ bán chạy nhất vào hai ngày này để tối đa hóa lợi nhuận.
    - Chiến dịch khuyến mãi: Tổ chức các chương trình khuyến mãi nhỏ, đặc biệt (ví dụ: "Deal hot ngày Thứ Năm") để thu hút thêm khách hàng và duy trì đà tăng trưởng.
  + Kích thích doanh thu vào các ngày thấp điểm (Cuối tuần)
    - Phát triển các chiến dịch đặc biệt cho cuối tuần để kích thích mua sắm. Ví dụ: "Ưu đãi cuối tuần", "Mua 1 tặng 1 vào Chủ Nhật".
* Người quản lý cần phải tối ưu hóa doanh thu trong các khung giờ
  + Chuẩn bị sẵn sàng các sản phẩm bán chạy, hoặc chạy các chương trình khuyến mãi ngắn hạn ("giờ vàng mua sắm") các khung giờ từ 10:00-12:00. Để tận dụng tối đa lượng khách hàng.
  + Đối với các khung giờ thấp điểm từ 16:00 - 20:00 có thể áp dụng các chiến lược marketing nhắm vào các khung giờ này, ví dụ như khuyến mãi giờ muộn hoặc dịch vụ giao hàng cho những người không có thời gian mua sắm sớm hơn.
* Rủi ro và Cơ hội dựa trên số lượng đơn hàng giữa các quốc gia:
  + Rủi ro phụ thuộc: Sự phụ thuộc quá lớn vào một thị trường duy nhất (Anh) là một rủi ro lớn. Bất kỳ sự biến động nào về kinh tế, chính trị, hay cạnh tranh tại Anh đều sẽ ảnh hưởng nghiêm trọng đến toàn bộ doanh nghiệp.
  + Cơ hội tăng trưởng: Tiềm năng phát triển rất lớn ở các thị trường quốc tế, đặc biệt là Đức và Pháp. Nếu có thể đẩy mạnh marketing và bán hàng để tăng trưởng ở các thị trường này, doanh nghiệp sẽ đa dạng hóa được rủi ro và mở rộng quy mô.
* Các Giải Pháp Thực Tế Để Giảm Thiểu Hủy/Trả Lại
  + Đào Tạo Nhân Viên Trở Thành "Chuyên Gia Tư Vấn". Nhân viên là cầu nối quan trọng để quản lý kỳ vọng. Nhân viên phải hiểu rõ điểm mạnh, điểm yếu của từng sản phẩm.
  + Điều Chỉnh Chính Sách Bán Hàng & Tiếp Thị. Xem xét lại chính sách trả hàng: Thay vì "Trả hàng miễn phí không cần lý do", có thể chuyển sang chính sách "Đổi hàng trong vòng 14 ngày" cho các sản phẩm không bị lỗi.
  + Thu Thập & Phân Tích Phản Hồi. Hỏi lý do trả hàng: Khi có khách trả hàng, hãy thiết kế một mẫu đơn giản để ghi lại lý do.
* Gợi Ý Hành Động Chiến Lược dựa theo xu hướng khách hàng mua các sản phẩm cùng nhau
  + Trưng bày & Sắp xếp cửa hàng.Sắp xếp các mặc hàng cùng nhau điều này khuyến khích khách hàng so sánh và mua nhiều mẫu hơn.
  + Tiếp thị & Khuyến mại. Ưu đãi Combo: Đưa ra các ưu đãi nhỏ khi mua combo. Ví dụ: "Mua 2 tách trà Regency khác màu, được giảm 10%" hoặc "Mua combo 2 chiếc đồng hồ Bakelike, tặng kèm pin".
* Phân khúc khách hàng
  + Nhóm VIP (Xanh lá) là mỏ vàng: Mặc dù chỉ có 6 khách hàng, họ cực kỳ khác biệt so với phần còn lại. Đây là nhóm cần có chính sách chăm sóc đặc biệt, cá nhân hóa tối đa để giữ chân và khai thác thêm giá trị.
  + Nhóm Trung thành (Xanh dương) là nền tảng: Đây là những khách hàng có giá trị cao và ổn định. Chiến lược phù hợp là nuôi dưỡng, xây dựng mối quan hệ để dần biến họ thành VIP.
  + Thách thức lớn nhất nằm ở nhóm Trung bình (Vàng) và Rủi ro (Đỏ):
  + Hai nhóm này chiếm tới 95% tổng số khách hàng.
    - Chúng nằm rất gần nhau và có phần chồng chéo. Điều này cho thấy hành vi của họ khá tương đồng, và ranh giới giữa một khách hàng "trung bình" và một khách hàng "sắp rời bỏ" là rất mong manh.
    - Đây là khu vực cần sự can thiệp marketing khéo léo nhất: Làm sao để giữ chân nhóm Rủi ro mà không tốn quá nhiều chi phí? Làm sao để thúc đẩy nhóm Trung bình mua nhiều hơn?

# CHƯƠNG 3. PHÂN TÍCH DỮ LIỆU CHIẾN DỊCH MARKETING

## 1. Định nghĩa vấn đề và mục tiêu phân tích

### 1.1. Bối cảnh vấn đề

Trong thời đại cạnh tranh khốc liệt hiện nay, việc hiểu rõ khách hàng trở thành yếu tố then chốt giúp doanh nghiệp duy trì lợi thế trên thị trường. Tuy nhiên, mỗi khách hàng đều có nhu cầu, hành vi và mối quan tâm khác nhau, khiến cho việc tiếp cận bằng một chiến lược marketing chung cho tất cả trở nên kém hiệu quả và lãng phí nguồn lực. Việc không thể phân biệt được các nhóm khách hàng khác nhau dẫn đến các chiến dịch không có tính cá nhân hóa, tỷ lệ chuyển đổi thấp và không tối đa hóa được giá trị vòng đời khách hàng (Customer Lifetime Value). Để giải quyết thách thức cốt lõi này, việc phân khúc khách hàng một cách khoa học dựa trên dữ liệu và hành vi thực tế là một nhu cầu cấp thiết.

### 1.2. Mục tiêu phân tích

Biết được các tính năng và cách sử dụng sản phẩm SweetViz, AutoViz để phân tích khám phá dữ liệu để thực hiện các mục tiêu chính sau:

* Định danh các phân khúc khách hàng điển hình: Sử dụng các kỹ thuật phân cụm (clustering) để nhóm các khách hàng cá nhân thành những tập hợp có chung đặc điểm về nhân khẩu học, hành vi tiêu dùng và phản ứng với các chiến dịch marketing.
* Mô tả chi tiết chân dung từng nhóm khách hàng: Phân tích sâu các đặc điểm của từng phân khúc được xác định, bao gồm độ tuổi, thu nhập, sản phẩm ưa thích, kênh mua hàng, độ nhạy cảm với khuyến mãi và giá trị giao dịch trung bình.
* Đề xuất chiến lược marketing và kinh doanh được cá nhân hóa: Dựa trên sự hiểu biết về từng nhóm khách hàng, cung cấp các khuyến nghị hành động cụ thể nhằm tối ưu hóa ngân sách marketing, cải thiện trải nghiệm khách hàng, nâng cao tỷ lệ giữ chân và thúc đẩy doanh thu cho doanh nghiệp.

### 1.3. Giới thiệu tổng quan về Dataset

Bộ dữ liệu Marketing Campaign mô tả thông tin nhân khẩu học, hành vi mua hàng, phản hồi chiến dịch và tương tác của khách hàng với công ty.

* **Dữ liệu vào:**
* Thông tin cá nhân (People)
  + - ID: Mã nhận dạng duy nhất của khách hàng
    - Year\_Birth: Năm sinh của khách hàng
    - Education: Trình độ học vấn của khách hàng
    - Marital\_Status: Tình trạng hôn nhân của khách hàng
    - Income: Thu nhập hàng năm của hộ gia đình
    - Kidhome: Số trẻ em trong hộ gia đình khách hàng
    - Teenhome: Số thanh thiếu niên trong hộ gia đình khách hàng
    - Dt\_Customer: Ngày khách hàng đăng ký với công ty
    - Recency: Số ngày kể từ lần mua hàng gần đây nhất của khách hàng
    - Complain: 1 nếu khách hàng đã khiếu nại trong 2 năm qua, 0 nếu không
* Sản phẩm (Products)
  + - MntWines: Số tiền đã chi cho rượu vang trong 2 năm qua
    - MntFruits: Số tiền đã chi cho trái cây trong 2 năm qua
    - MntMeatProducts: Số tiền đã chi cho thịt trong 2 năm qua
    - MntFishProducts: Số tiền đã chi cho cá trong 2 năm qua
    - MntSweetProducts: Số tiền đã chi cho đồ ngọt trong 2 năm qua
    - MntGoldProds: Số tiền đã chi cho sản phẩm vàng (trang sức, v.v.) trong 2 năm qua
  + Khuyến mãi (Promotion)
    - NumDealsPurchases: Số lần mua hàng được giảm giá
    - AcceptedCmp1: 1 nếu khách hàng chấp nhận ưu đãi trong chiến dịch thứ nhất, 0 nếu không
    - AcceptedCmp2: 1 nếu khách hàng chấp nhận ưu đãi trong chiến dịch thứ hai, 0 nếu không
    - AcceptedCmp3: 1 nếu khách hàng chấp nhận ưu đãi trong chiến dịch thứ ba, 0 nếu không
    - AcceptedCmp4: 1 nếu khách hàng chấp nhận ưu đãi trong chiến dịch thứ tư, 0 nếu không
    - AcceptedCmp5: 1 nếu khách hàng chấp nhận ưu đãi trong chiến dịch thứ năm, 0 nếu không
    - Response: 1 nếu khách hàng chấp nhận ưu đãi trong chiến dịch gần đây nhất, 0 nếu không
* Địa điểm (Place)
  + - NumWebPurchases: Số lần mua hàng qua trang web của công ty
    - NumCatalogPurchases: Số lần mua hàng bằng catalog (ấn phẩm quảng cáo sản phẩm)
    - NumStorePurchases: Số lần mua hàng trực tiếp tại cửa hàng
    - NumWebVisitsMonth: Số lượt truy cập trang web của công ty trong tháng gần đây nhất
* **Dữ liệu ra:** Sau khi phân tích bằng Sweetviz và AutoViz, kết quả đầu ra bao gồm:
* Sweetviz: File HTML (sweetviz\_report.html). Báo cáo tổng quan dữ liệu: thống kê mô tả, histogram, bar chart, phân bố biến, tỷ lệ dữ liệu thiếu, tương quan giữa các biến.
* AutoViz: Bộ biểu đồ trực quan (png). Các biểu đồ phân tích đa biến như scatter plot, box plot, correlation heatmap, bar plot,… giúp phát hiện mối quan hệ giữa các biến.

## 2. Giới thiệu công cụ

### 2.1. Sweetviz

Sweetviz là một thư viện Python được thiết kế để phân tích dữ liệu khám phá (Exploratory Data Analysis – EDA) một cách tự động và trực quan. Công cụ này có khả năng:

* Tự động tổng hợp, mô tả và so sánh các đặc trưng (features) trong tập dữ liệu.
* Sinh ra báo cáo HTML trực quan, trong đó thể hiện.
* Phân phối dữ liệu (Histogram cho biến số, Bar chart cho biến phân loại).
* Thống kê cơ bản: trung bình, trung vị, min, max, độ lệch chuẩn.
* Tỷ lệ dữ liệu thiếu (missing values) và tương quan giữa các biến.

### 2.1. AutoViz

AutoViz là thư viện Python mạnh mẽ dùng để tự động hóa quá trình trực quan hóa dữ liệu đa biến (Multivariate Analysis). Khác với Sweetviz chỉ tập trung mô tả, AutoViz giúp tìm mối quan hệ giữa các biến, phát hiện tương quan và xu hướng tiềm ẩn trong dữ liệu.

Công cụ này:

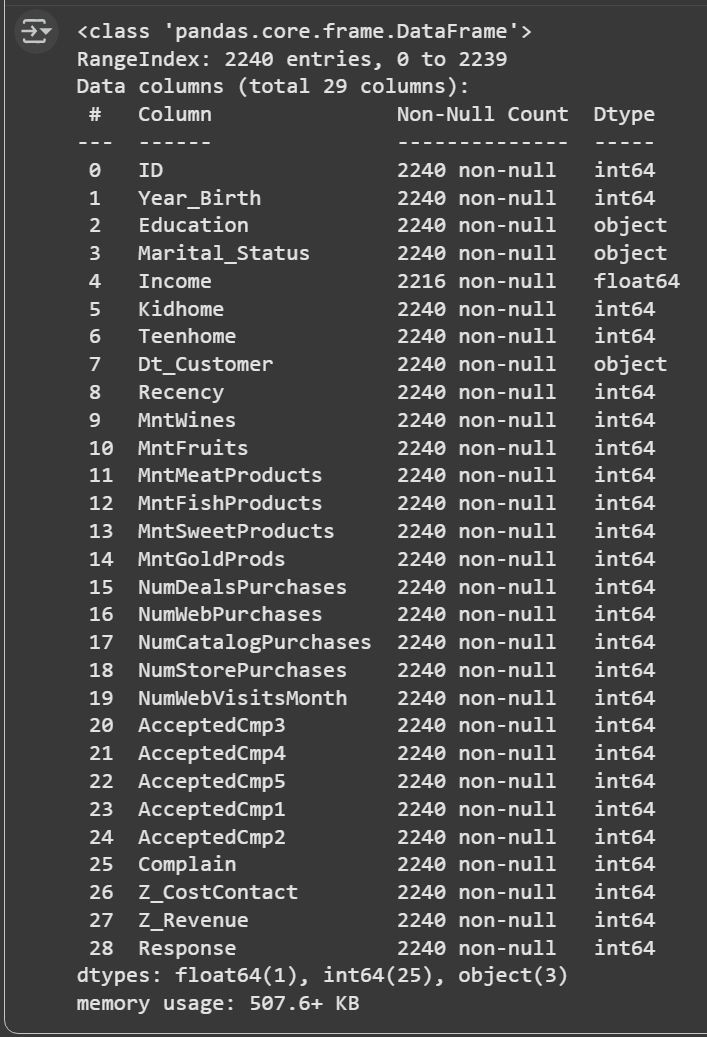
* Tự động xác định loại biến (số, phân loại, nhị phân, ngày tháng).
* Tạo ra biểu đồ scatter plot, boxplot, heatmap, bar chart cho các cặp biến quan trọng.
* Tự động chọn biến “phụ thuộc” (depVar) để xem mối quan hệ với các biến còn lại.
* Giúp người dùng phát hiện các cụm khách hàng, hành vi chi tiêu, và mối tương quan trong dữ liệu.

Ưu điểm:

* Phân tích đa biến (bivariate/multivariate) tự động.
* Trực quan hóa mối quan hệ giữa các thuộc tính.
* Hỗ trợ nhiều loại dữ liệu, kể cả dữ liệu lớn.

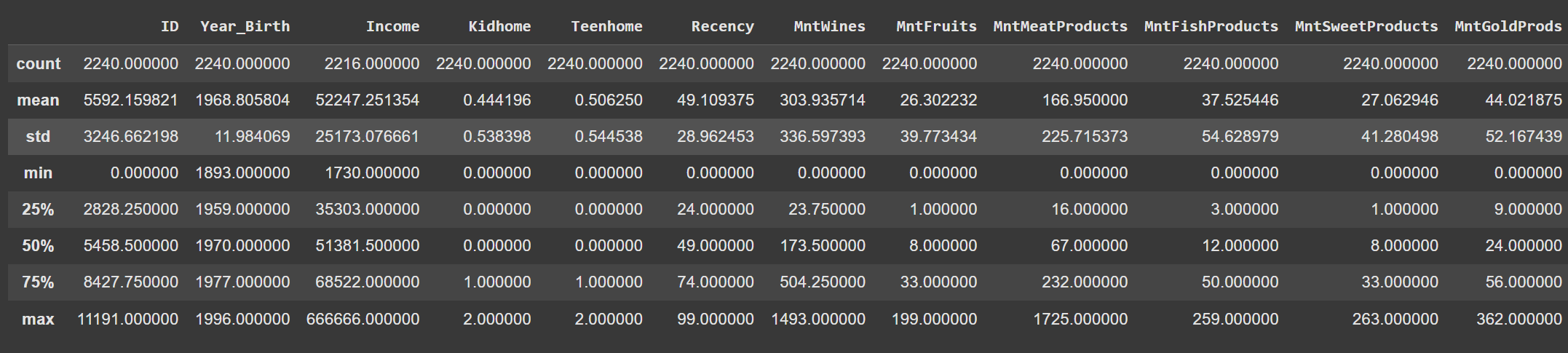
## 3. Mô tả thống kê về dữ liệu

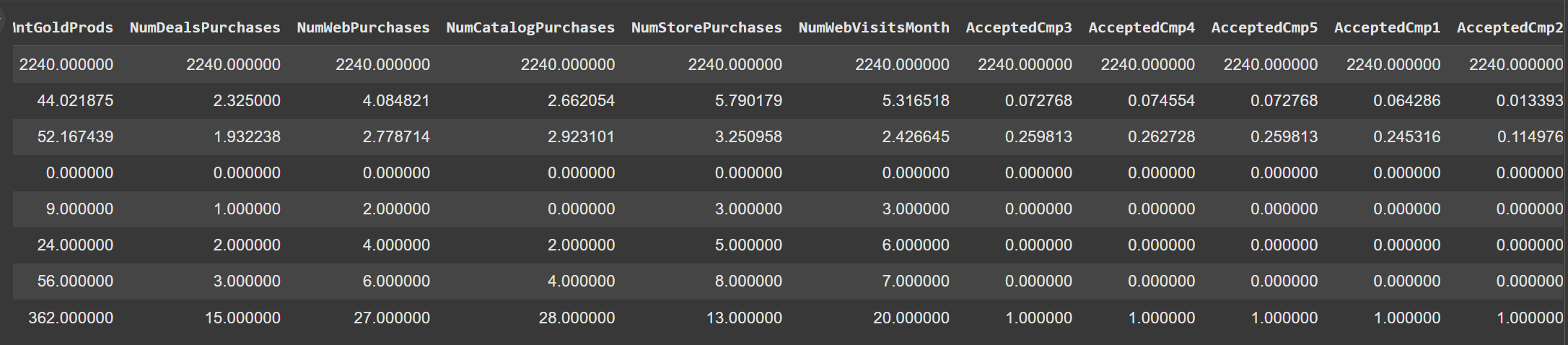
### 3.1. Thông tin tổng quát



#### Hình 3.1.1. Bảng biểu hiển thị thông tin cơ bản

* **Nhận xét:**
* Tổng số dòng dữ liệu là 2240 dòng
* Số cột là 29
* Kiểu dữ liệu gồm: 25 cột số nguyên(int), 1 cột số thực(float), 3 cột chuỗi(object)
* Tính đầy đủ của dữ liệu:
  + Biến income bị thiếu 24 hàng dữ liệu
  + Dự đoán nguyên nhân:
    - Khách hàng không cung cấp thông tin thu nhập
    - Dữ liệu bị lỗi trong quá trình thu thập hoặc nhập liệu
    - Khách hàng mới đăng ký chưa cập nhật thông tin đầy đủ
    - Lỗi hệ thống hoặc định dạng khác nhau khi nhập liệu
  + Hướng giải quyết: Số lượng hàng dữ liệu bị thiếu là 24 (0,01%). Vì số lượng không đáng kể, sẽ không gây ảnh hưởng lớn đến toàn bộ mẫu, nên sẽ tiến hành xóa bỏ dữ liệu.





#### Hình 3.1.2. Bảng biểu hiển thị các tính chất thống kê trên dữ liệu số

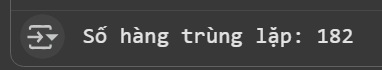
* **Nhận xét:**
* Count(số lượng): Vì tổng số lượng của Income là 2216, trong khi các biến còn lại là 2240, nên cần tiến hành xử lí dữ liệu.
* Mean(trung bình):
  + Vì trung bình của Year\_Birth là 1968, nên độ tuổi trung bình là 53 tuổi(tính đến năm 2021)
  + Trung bình của thu nhập là 52247
  + Vì trung bình của MntWines, MntMeatProducts cao hơn so với những thứ còn lại, nên chứng tỏ rượu vang và thịt được mua nhiều nhất
* std(độ lệch chuẩn): Vì độ lệch chuẩn của Income,Recency là rất cao, nên chứng tỏ biến Income,Recency có độ phân tán cao
* min, max: Có sự chênh lệch giữa min và max của các biến Income, Mnt(\*), cho thấy dữ liệu có sự chênh lệch đáng kể

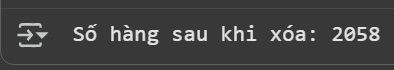
## 4. Làm sạch dữ liệu

Một số cột không cần thiết cho việc phân tích nhóm khách hàng và nên được loại bỏ để giảm nhiễu:

* ID: Mã định danh duy nhất của khách hàng. Nó không mang ý nghĩa phân tích và chỉ dùng để nhận dạng khách hàng.
* Dt\_Customer: Ngày khách hàng đăng ký. Mặc dù có thể dùng để tính tuổi đời khách hàng, nhưng Recency (thời gian mua hàng gần nhất) thường mang ý nghĩa thực tế hơn cho việc phân tích hành vi.

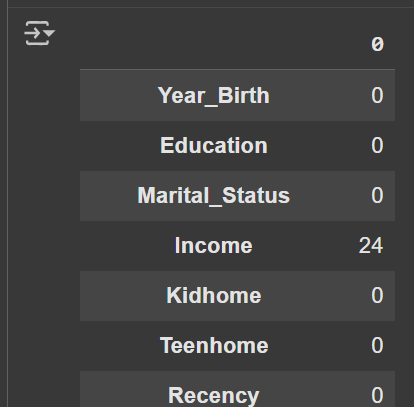
### 4.1. Kiểm tra và xử lí giá trị trùng lặp





Hình 4.1.1. Kiểm tra và xử lí giá trị trùng lặp

### 4.2. Kiểm tra và xử lí giá trị thiếu

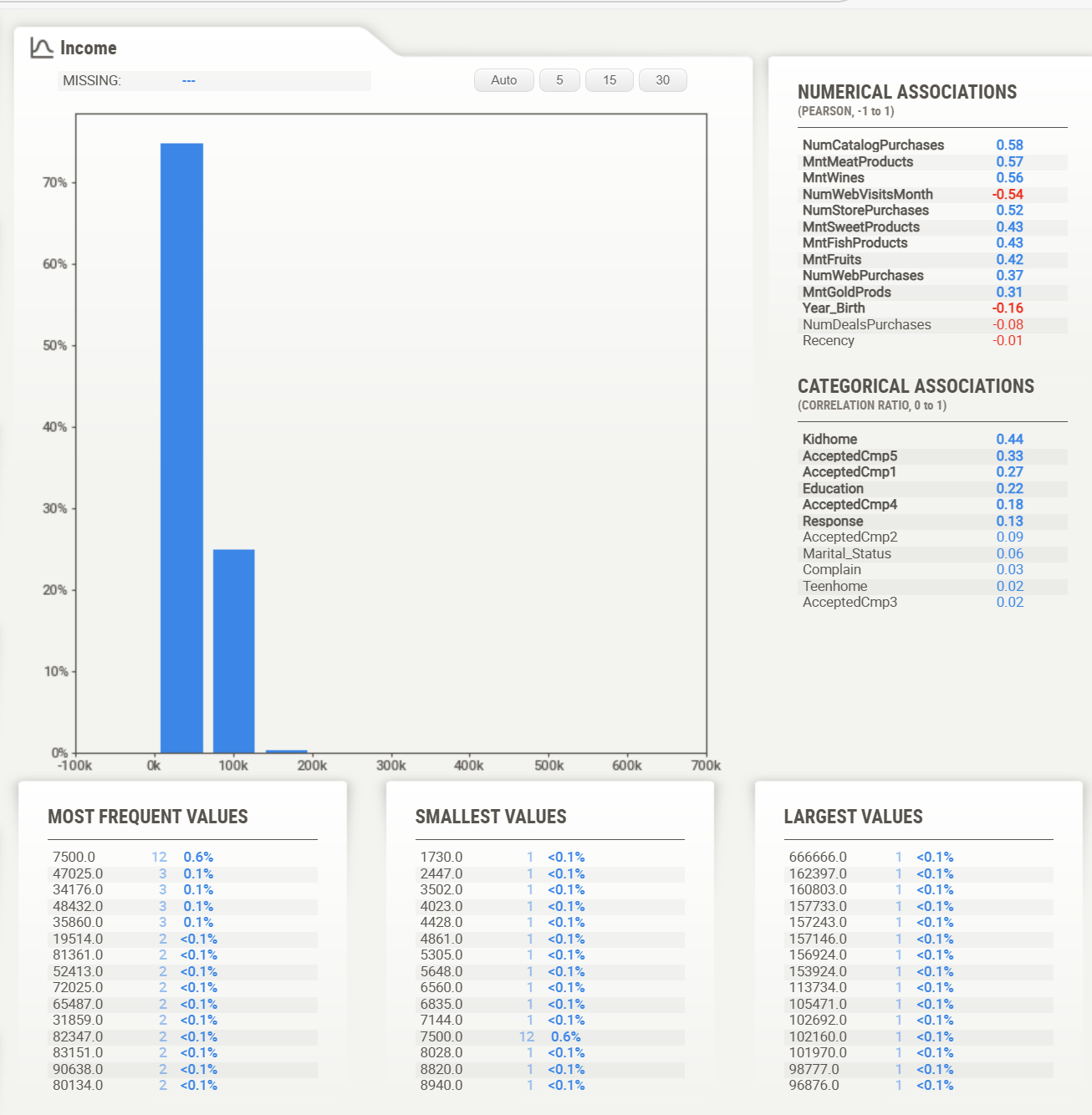


Hình 4.2.1. Kiểm tra và xử lí giá trị thiếu

## 5. Phân tích khám phá dữ liệu với SweetViz và AutoViz

### 5.1. SweetViz

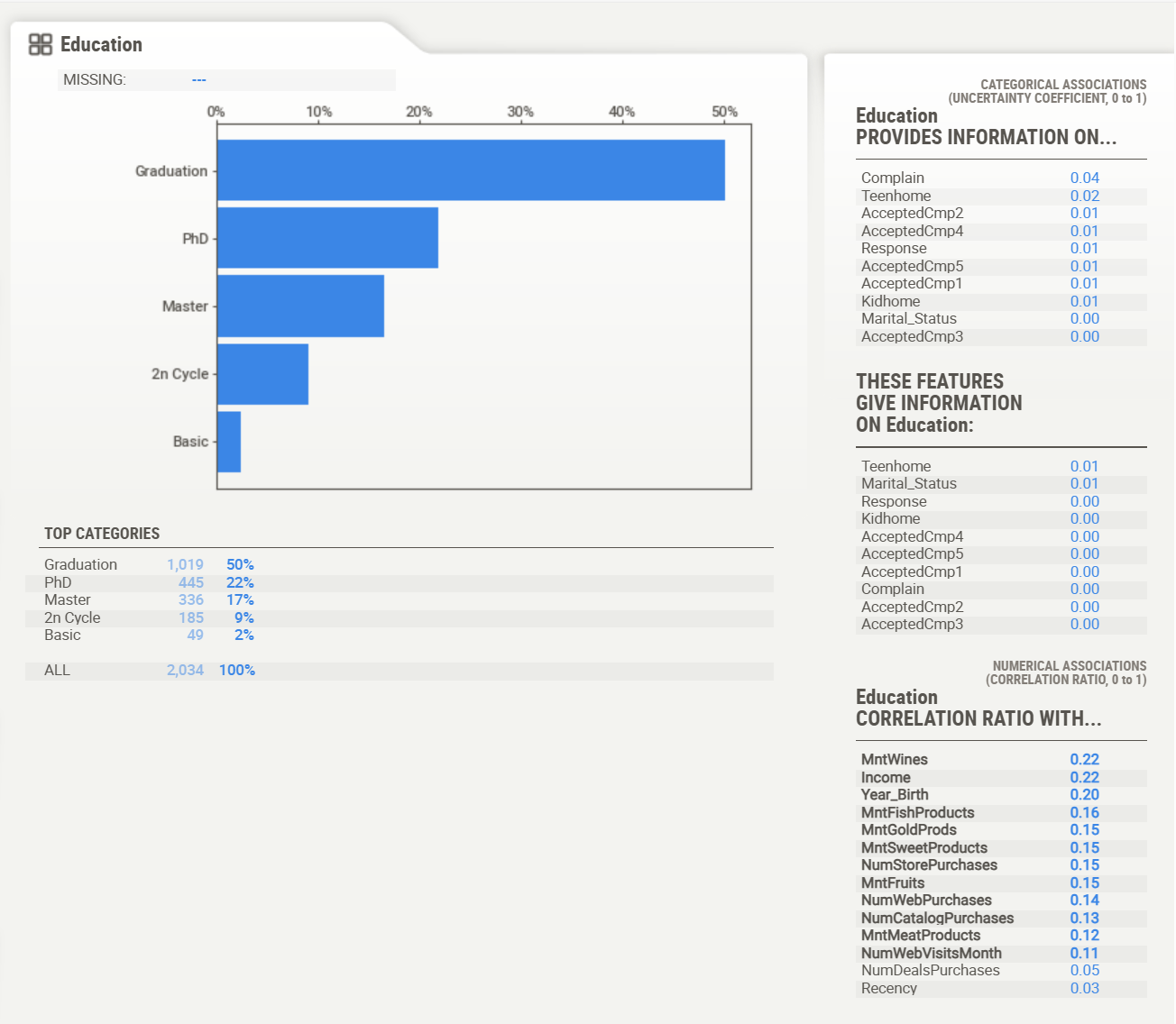
***Câu hỏi 1:*** Thu nhập của khách hàng có mối quan hệ như thế nào với hành vi chi tiêu và hình thức mua sắm?



Hình 5.1.1. Biểu đồ cột thu nhập và các tương quan

* **Nhận xét:**
* Phần lớn khách hàng có thu nhập từ 0 - 100000, tuy nhiên vẫn có 1 số ít có mức lương trên 150000(giá trị ngoại lai - outliers)
* Sự tương quan:
  + Tương quan cao với MntMeatProducts(0.57), MntWines(0.56). người có thu nhập cao sẽ chi rất nhiều cho thịt và rượu.
  + Tương quan cao với NumCatalogPurchases(0.58), NumStorePurchases(0.52) .Người có thu nhập cao thường mua qua catalog và mua trực tiếp tại cửa hàng nhiều hơn.
  + Tương quan trung bình với MntFruits(0.42), MntSweetProducts(0.43), MntFishProducts(0.43). người thu nhập cao chi nhiều cho trái cây, cá, đồ ngọt, nhưng ít hơn thịt và rượu. Tương quan âm với NumWebVisitsMonth(-0.54). Người thu nhập cao ít truy cập website hơn.
* Kết luận: Nhóm khách hàng thu nhập cao có hành vi chi tiêu thiên về chất lượng và trải nghiệm, ưa chuộng kênh mua hàng trực tiếp hoặc catalog, thay vì các kênh trực tuyến.

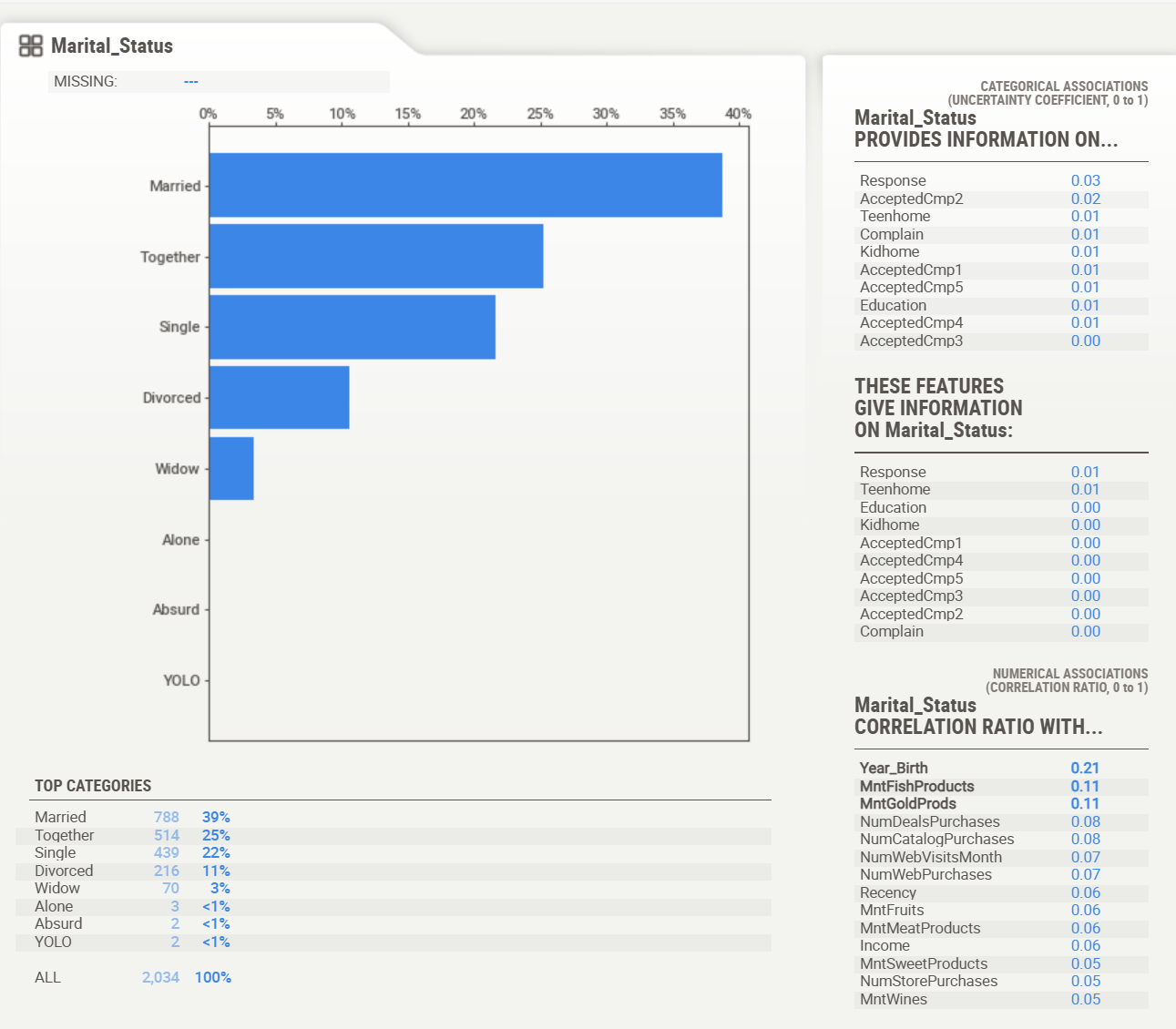
***Câu hỏi 2:*** Trình độ học vấn của khách hàng có ảnh hưởng như thế nào đến thu nhập và mức chi tiêu?



Hình 5.1.2. Biểu đồ cột trình độ học vấn và các tương quan

* **Nhận xét:**
* Phần lớn khách hàng có trình độ Graduation (Đại học), chiếm 50%.
* Nhóm PhD-Tiến sĩ(22%) và Master-Thạc sĩ (17%) cũng chiếm tỷ lệ khá cao, cho thấy tập khách hàng chủ yếu có trình độ học vấn cao.
* Nhóm 2n Cycle - Trung cấp (9%) và Basic-Tiểu học (2%) chiếm tỷ lệ nhỏ, thể hiện khách hàng có học vấn thấp rất ít trong dữ liệu.
* Sự tương quan:
  + Income (0.22): Tương quan nhẹ. Người có trình độ học vấn cao thường có thu nhập cao hơn.
  + Year\_Birth (0.20): Tương quan nhẹ. Có xu hướng những người trẻ hơn thường có trình độ học vấn cao hơn.
  + MntWines (0.22): Tương quan nhẹ. Người học cao có xu hướng chi nhiều cho rượu vang, có thể phản ánh gu tiêu dùng tinh tế hoặc thu nhập cao.
* Kết luận: Học vấn cao thì thường thu nhập cao. Chi tiêu cao hơn cho các sản phẩm(rượu, thịt).

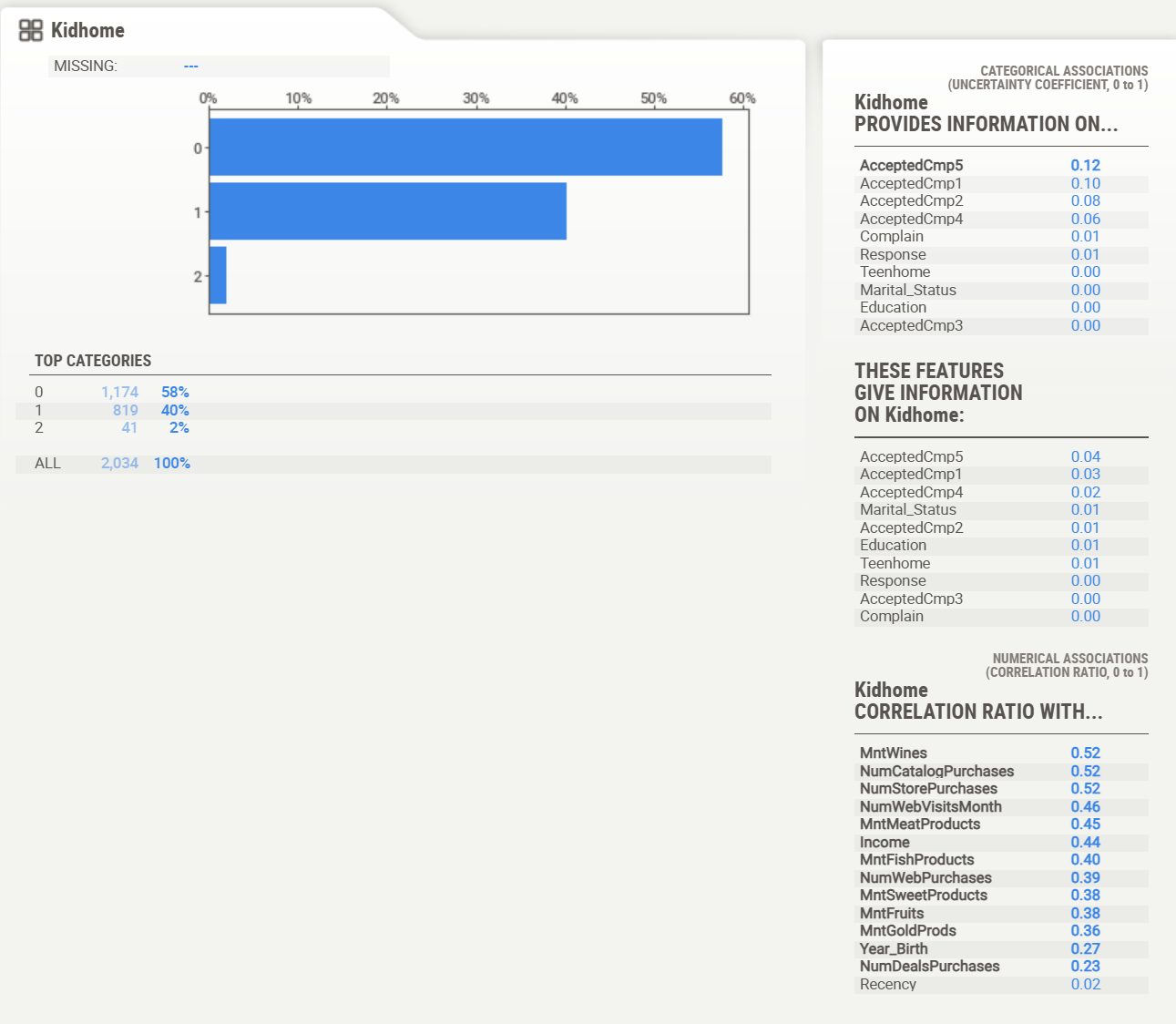
***Câu hỏi 3:*** Trình trạng hôn nhân có gây ảnh hưởng lớn đến hành vi chi tiêu của họ hay không?



Hình 5.1.3. Biểu đồ cột tình trạng hôn nhân và các tương quan

* **Nhận xét:**
* Married-đã kết hôn (39%) là nhóm lớn nhất, theo sau là Together-sống thử (25%) và Single (22%).
* Các nhóm còn lại (Divorced-li hôn 11%, Widow-Góa 3%) chiếm tỷ lệ nhỏ.
* Một số giá trị hiếm như Alone- độc thân, Absurd-dữ liệu bất thường, YOLO (1%)
* Sự tương quan : Tương quan nhẹ: Year\_Birth (0.21). Nhóm độc thân thường trẻ hơn, trong khi đã kết hôn hoặc ly hôn thường lớn tuổi hơn.
* Kết luận: tình trạng hôn nhân ít ảnh hưởng đến hành vi chi tiêu của họ.

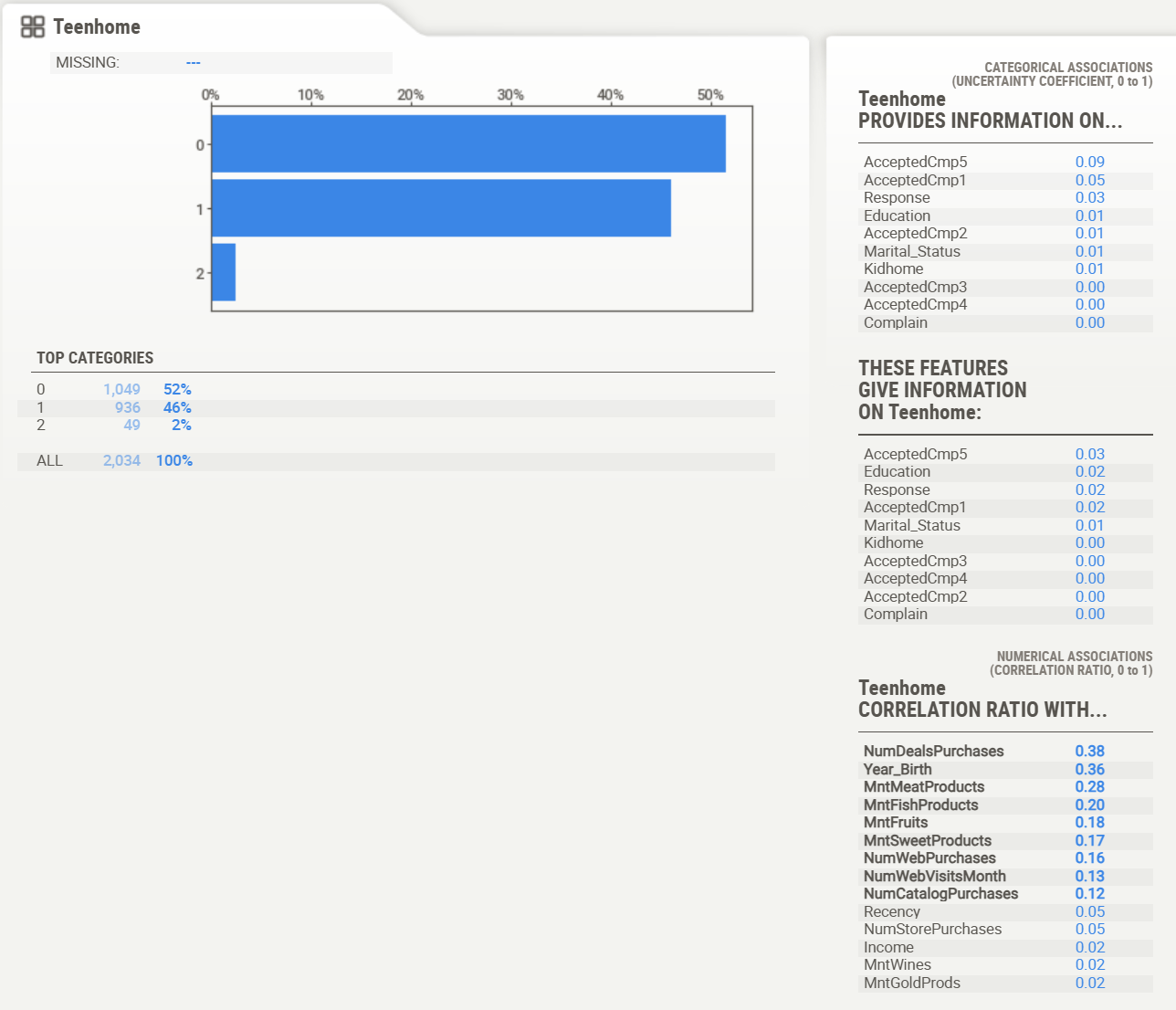
***Câu hỏi 4:*** Người không có con hoặc có 1 con nhỏ sẽ có hành vi mua sắm gì?



Hình 5.1.4. Biểu đồ cột con cái và các tương quan

* **Nhận xét:**
* Phần lớn khách hàng không có con nhỏ tại nhà. Cụ thể, 58% khách hàng không có con nhỏ (Kidhome = 0), 40% có một con nhỏ (Kidhome = 1) và chỉ 2% có hai con nhỏ (Kidhome = 2).
* Sự tương quan:
  + Tương quan cao với MntWines (0.52), MntMeatProducts (0.45). Các khách hàng không có con nhỏ hoặc có ít con nhỏ có xu hướng chi tiêu nhiều hơn cho rượu, thịt.
  + Tương quan cao với NumCatalogPurchases (0.52), NumStorePurchases (0.46), NumWebVisitsMonth (0.45). Các khách hàng không có con nhỏ hoặc có ít con nhỏ có phương thức mua sắm đa dạng qua catalog, tại cửa hàng và thường truy cập website nhiều hơn.
  + Tương quan trung bình với MntFishProducts (0.40), NumWebPurchases (0.39), MntSweetProducts (0.38), MntFruits (0.38). Khách hàng không có con nhỏ hoặc có ít con nhỏ cũng chi tiêu nhiều hơn cho cá, đồ ngọt, trái cây và mua sắm trực tuyến. Điều này cho thấy họ có xu hướng chi tiêu đa dạng hơn.
  + Tương quan trung bình với Income (0.44). Khách hàng có thu nhập cao hơn có xu hướng có ít con nhỏ hơn.
* Kết luận:
  + Nhóm khách hàng có ít hoặc không có con nhỏ (Kidhome = 0 và Kidhome = 1) là nhóm chi tiêu mạnh nhất, họ có xu hướng chi tiêu đa dạng và sử dụng nhiều kênh mua sắm khác nhau.
  + Họ có xu hướng có thu nhập cao hơn so với những người có nhiều con nhỏ.
  + Có thể các gia đình có con nhỏ sẽ có ngân sách hạn chế hơn và ưu tiên chi tiêu cho các nhu cầu thiết yếu hơn..

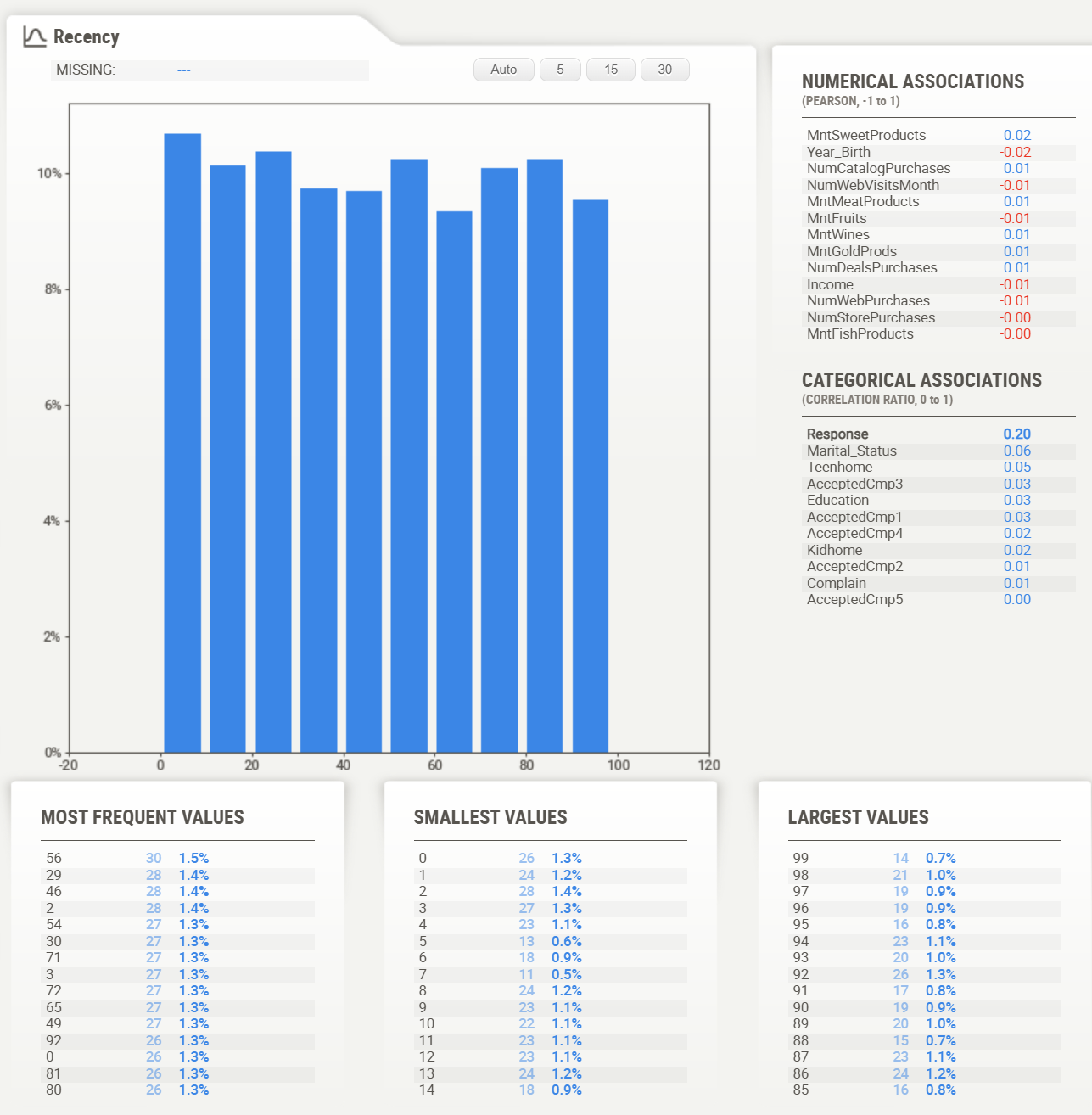
***Câu hỏi 5:*** Người có ít trẻ thanh thiếu niên hoặc không có con sẽ có hành vi mua sắm gì?



Hình 5.1.5. Biểu đồ cột trẻ thành niên và các tương quan

* **Nhận xét:**
* Phần lớn khách hàng không có con vị thành niên tại nhà. 52% khách hàng không có con vị thành niên (Teenhome = 0), 46% có một con (Teenhome = 1) và 2% có hai con (Teenhome = 2).
* Sự tương quan:
  + Tương quan cao với NumDealsPurchases (0.38). Khách hàng không có con hoặc có 1 con có xu hướng mua hàng qua các chương trình khuyến mãi và ưu đãi nhiều hơn. Điều này có thể do ngân sách gia đình phải chi tiêu nhiều hơn cho các nhu cầu của con cái.
  + Tương quan trung bình nhẹ với MntMeatProducts (0.28). Khách hàng không có con hoặc có 1 con chi tiêu nhiều hơn cho các sản phẩm thịt. Điều này có thể do nhu cầu dinh dưỡng hoặc sở thích ăn uống của con cái.
* Kết luận:
* Khách hàng không có con hoặc có 1 con có xu hướng nhạy cảm về giá hơn và ưa chuộng mua sắm qua các chương trình khuyến mãi, ưu đãi (NumDealsPurchases).
* Hành vi chi tiêu của họ thiên về các sản phẩm thiết yếu như thịt và ít hơn đối với các mặt hàng xa xỉ như cá, trái cây, đồ ngọt.

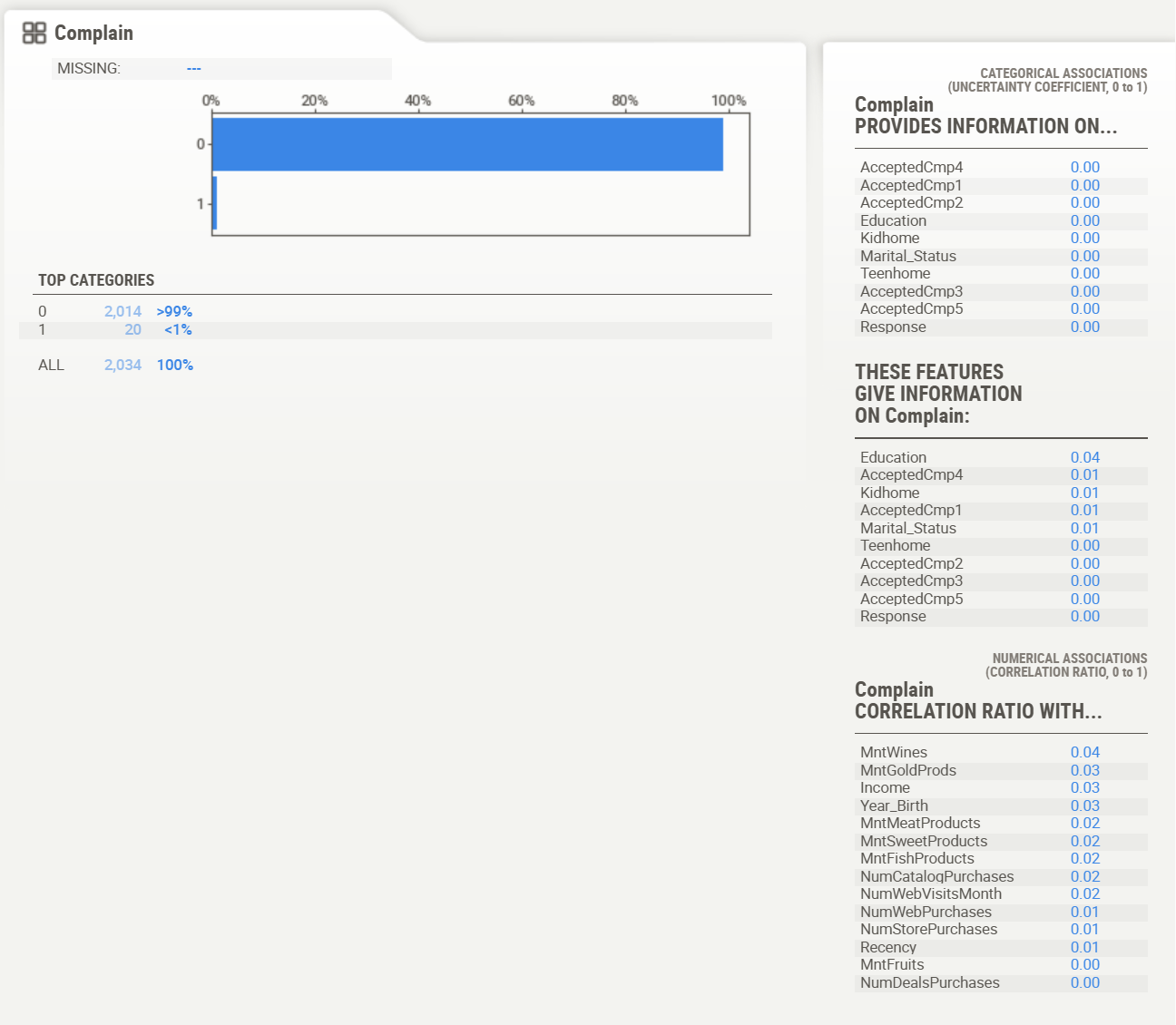
***Câu hỏi 6:*** Khách hàng gần đây thì sẽ có tham gia chương trình khuyến mãi gần nhất không ?



Hình 5.1.6. Biểu đồ cột mua hàng gần đây và các tương quan

* **Nhận xét:**
* Biểu đồ cho thấy giá trị Recency phân tán khá đều trên toàn bộ dải từ 0 đến 100. Không có một khoảng giá trị nào chiếm ưu thế vượt trội. Sự tương quan:
* tương quan nhẹ với Response (0.20). Cho thấy những khách hàng có số ngày kể từ lần mua cuối cùng ít hơn (mua hàng gần đây) có khả năng phản hồi lại các chiến dịch marketing cao hơn.
* **Kết luận:**
* Những khách hàng mua hàng gần đây nhất là nhóm khách hàng tiềm năng nhất để triển khai các chương trình khuyến mãi hoặc thông báo sản phẩm mới.
* Để tối ưu hóa hiệu quả marketing, bạn nên xây dựng các chiến dịch nhắm mục tiêu vào nhóm khách hàng "mới mua" (ví dụ, những người có Recency dưới 30 ngày) để khuyến khích họ quay lại mua sắm lần tiếp theo.
* Phân tích Recency cũng giúp doanh nghiệp nhận diện các khách hàng có nguy cơ rời bỏ bằng cách theo dõi những người có số ngày Recency cao (trên 90 ngày).

***Câu hỏi 7:*** Sự khiếu nại của khách hàng có gây ảnh hưởng đến hành vi chi tiêu của họ không?

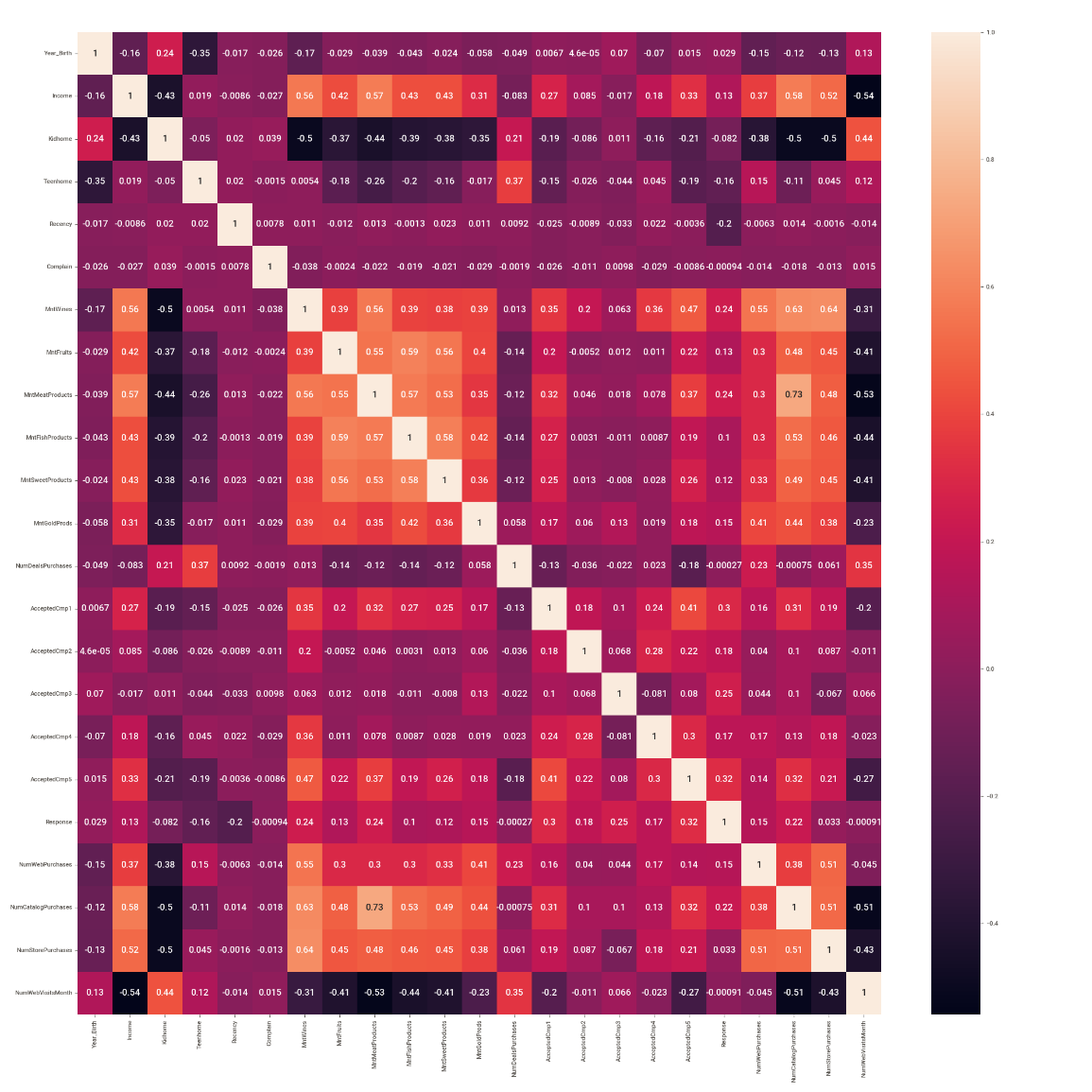


Hình 5.1.7. Biểu đồ cột khiếu nại và các tương quan

* **Nhận xét:**
* Tỷ lệ khách hàng khiếu nại rất thấp: Hơn 99% khách hàng không có bất kỳ khiếu nại nào trong 2 năm qua (Complain = 0). Chỉ có dưới 1% khách hàng đã khiếu nại (Complain = 1).
* Mối tương quan: Tương quan với các biến khác rất yếu: Tất cả các chỉ số tương quan, cả về numerical và categorical, đều rất thấp, gần bằng 0. Cụ thể, các chỉ số tương quan với MntWines, MntGoldProds, Income... đều chỉ từ 0.01 đến 0.04.
* Kết luận:
  + Dữ liệu cho thấy tỷ lệ khiếu nại của khách hàng là cực kỳ thấp. Đây là một dấu hiệu tích cực, cho thấy sản phẩm và dịch vụ của bạn có thể đang hoạt động tốt.
  + Vì tỷ lệ khiếu nại quá thấp và không có mối tương quan rõ ràng với các biến khác, việc sử dụng biến này để phân cụm khách hàng là không cần thiết.

### 5.2. AutoViz

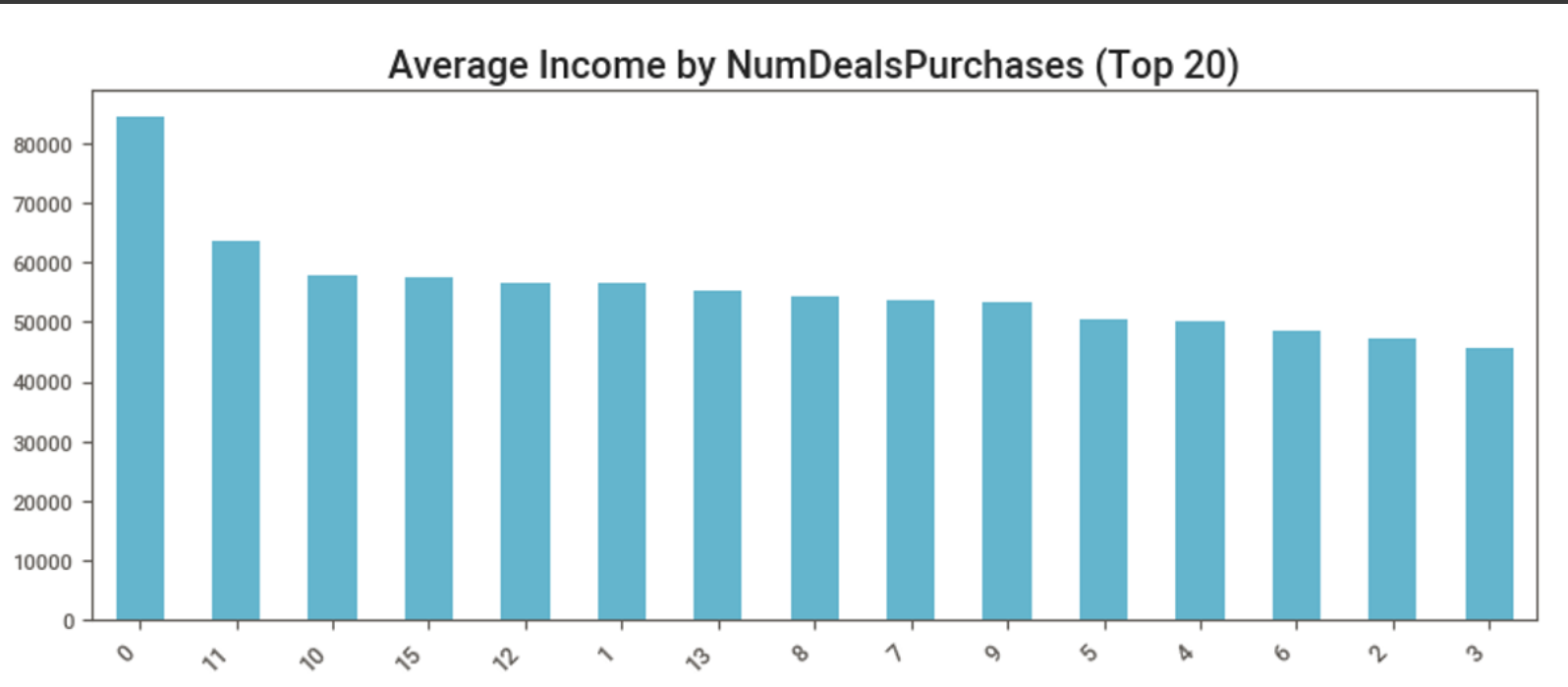
***Câu hỏi 1:*** Liệu thu nhập của khách hàng (Income) có ảnh hưởng đến mức chi tiêu, phương thức mua sắm của họ không? Khách hàng có xu hướng chi tiêu tập trung vào một sản phẩm, hay mua nhiều loại sản phẩm khác nhau cùng lúc?



Hình 5.2.1. Biểu đồ HeatMap

* **Nhận xét:**
* Biến Income:
  + Tương quan cao với MntWines, MntGoldProds, MntMeatProducts (0.5 - 0.6). Thu nhập càng cao, chi tiêu càng lớn cho các mặt hàng giá trị cao.
  + Tương quan cao với NumCatalogPurchases(0.58), NumStorePurchases(0.52), chứng tỏ người có lương cao thì sẽ có xu hướng mua qua catalog và store.
  + Tương quan trung bình với NumWebPurchases(0.37) , cho biết người lương cao cũng mua hàng ở website.
  + Các biến MntWines', 'MntFruits', 'MntMeatProducts', 'MntFishProducts', 'MntSweetProducts', 'MntGoldProds' có tương quan trung bình từ 0.3 trở lên, chứng tỏ tệp khách hàng có xu hướng mua các sản phẩm này chung với nhau, chứ không phải tập trung vào 1 sản phẩm nhất định.
  + Có tương quan âm cao(-0.54) với NumWebVisitsMonth, cho biết người có thu nhập cao thì sẽ có xu hướng ít truy cập vào các trang web bán hàng.
* Kết luận:
  + Người có lương cao sẽ có xu hướng mua những mặt hàng giá trị cao, và sẽ mua trên cả catalog, website và store. Tuy nhiên lại không thường xuyên truy cập vào trang web của công ty.
  + Tệp khách hàng có xu hướng mua các sản phẩm này chung với nhau, chứ không phải tập trung vào 1 sản phẩm nhất định.

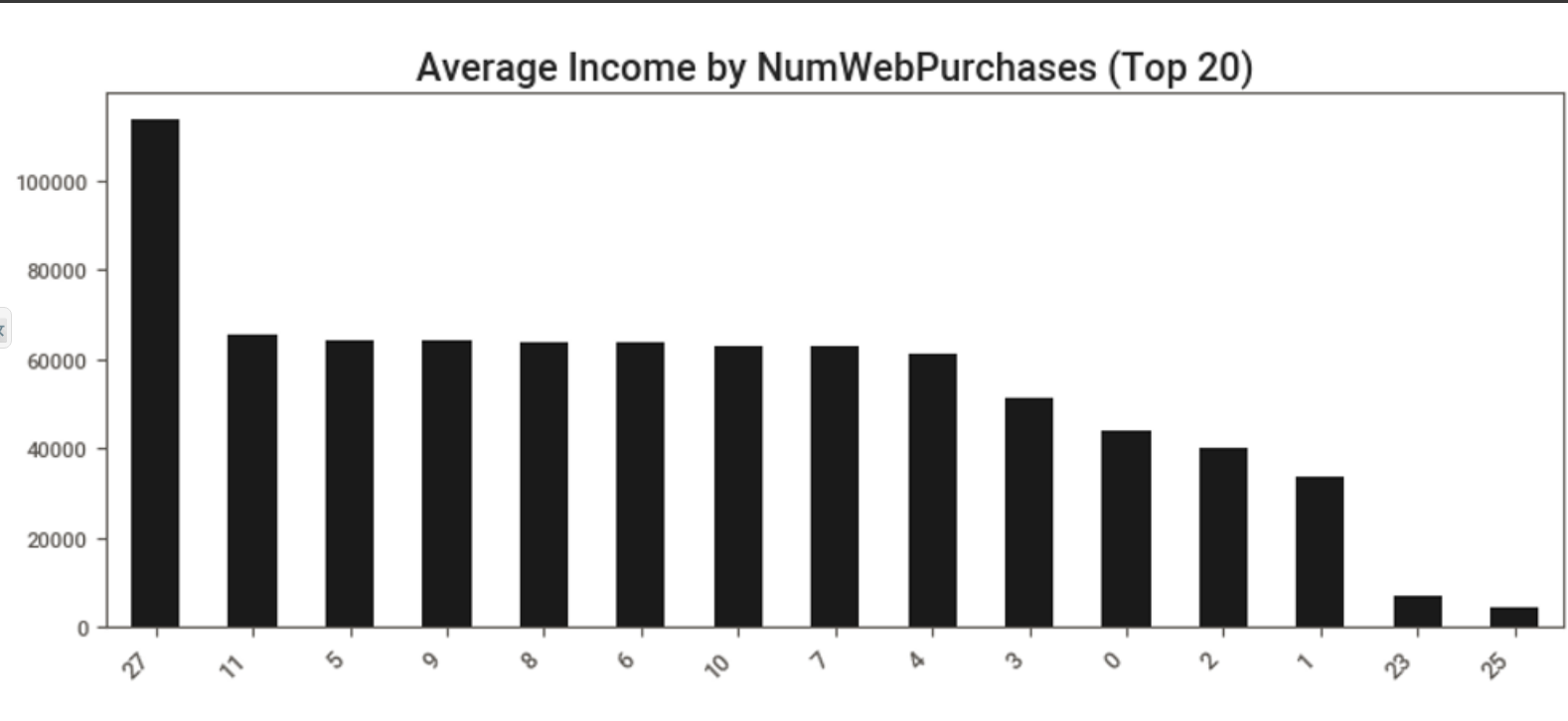
***Câu hỏi 2:*** Mức lương sẽ phản ánh sự thu hút của khách hàng vào chương trình khuyến mãi như thế nào?

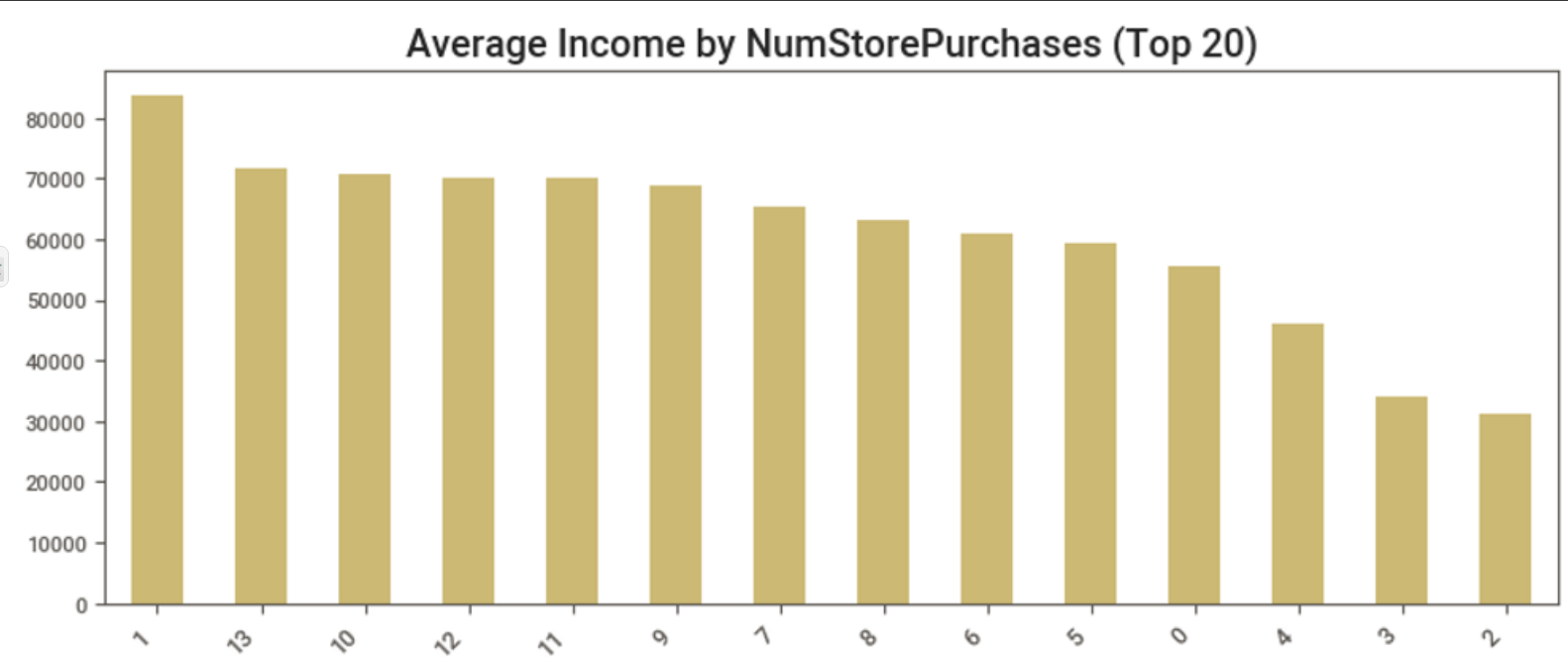


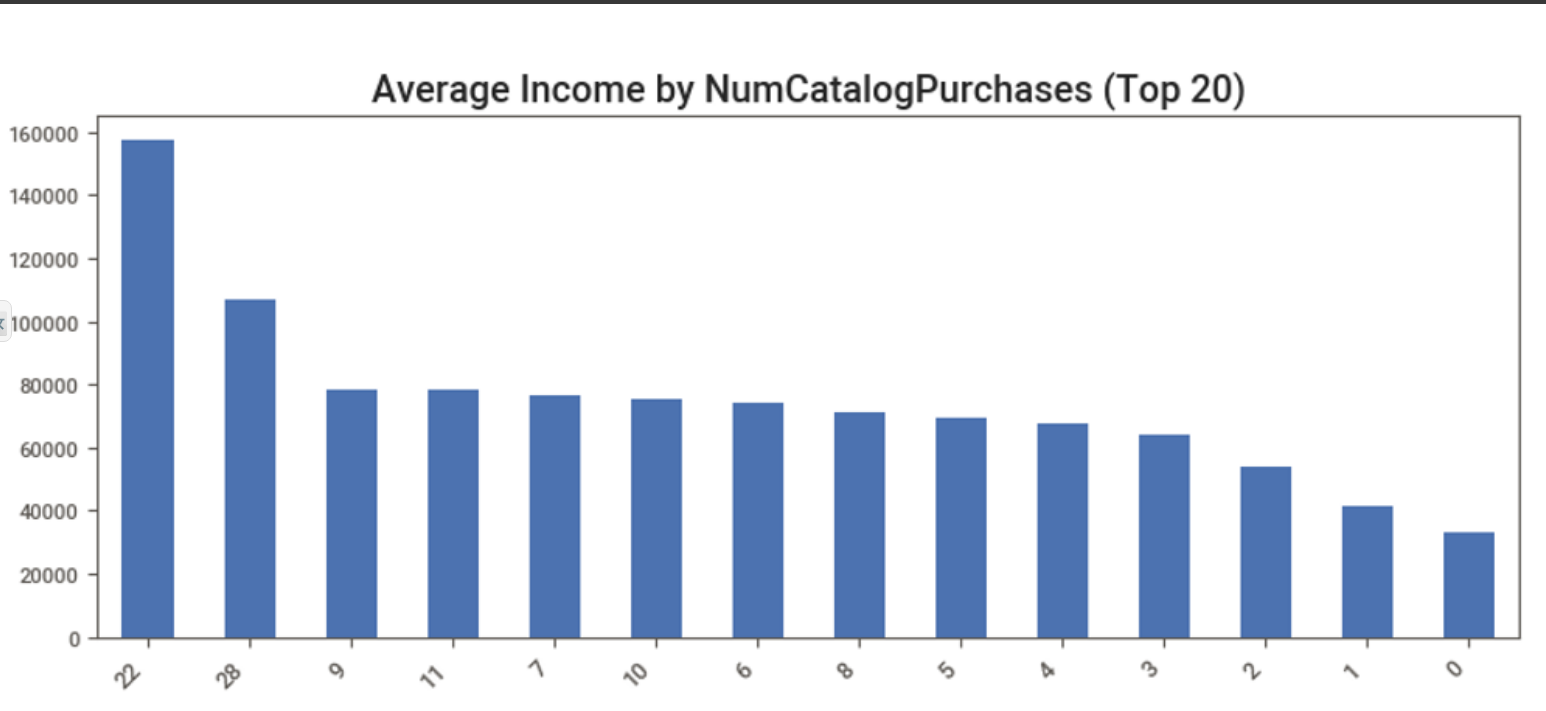
Hình 5.2.2. Biểu đồ cột top 20 mức lương trung bình mua có khuyến mãi

* **Nhận xét:** Biểu đồ cho thấy những người có thu nhập trung bình cao nhất(khoảng 80000) lại là những người có số lần mua hàng giảm giá bằng 0. Khi số lần mua hàng giảm giá tăng lên, thu nhập trung bình của nhóm khách hàng đó lại có xu hướng giảm dần.
* **Kết luận:** Khách hàng thu nhập cao ít quan tâm đến khuyến mãi và sẵn sàng chi trả giá gốc, thể hiện mức độ trung thành và ít nhạy cảm về giá. Ngược lại, nhóm thu nhập thấp thường tận dụng các chương trình giảm giá để tiết kiệm chi phí. Do đó, công ty nên duy trì chính sách giá trị và dịch vụ cao cấp cho khách hàng thu nhập cao, đồng thời thiết kế các chương trình khuyến mãi phù hợp để thu hút và giữ chân nhóm thu nhập thấp.

***Câu hỏi 3:*** Khách hàng có thu nhập cao thường mua hàng qua kênh nào nhiều nhất - website, cửa hàng, hay catalogue? Liệu thu nhập có ảnh hưởng đến sở thích kênh mua hàng hay không?



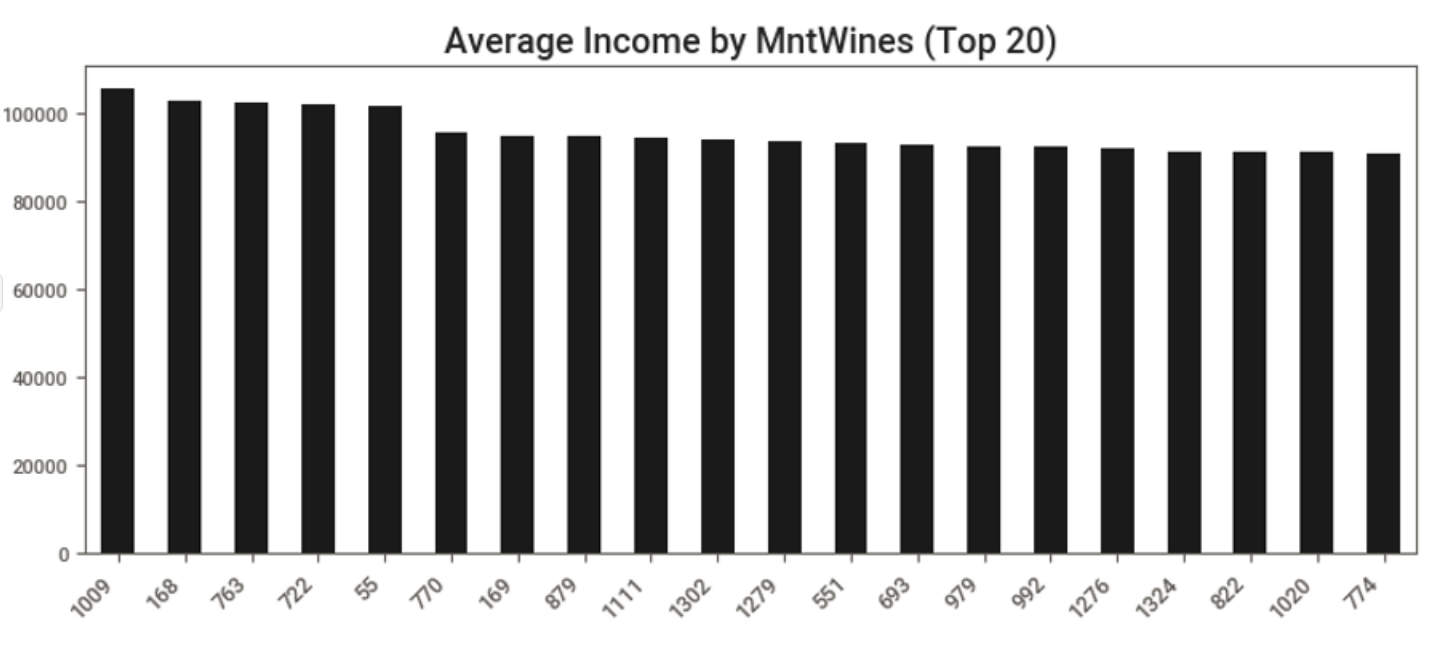


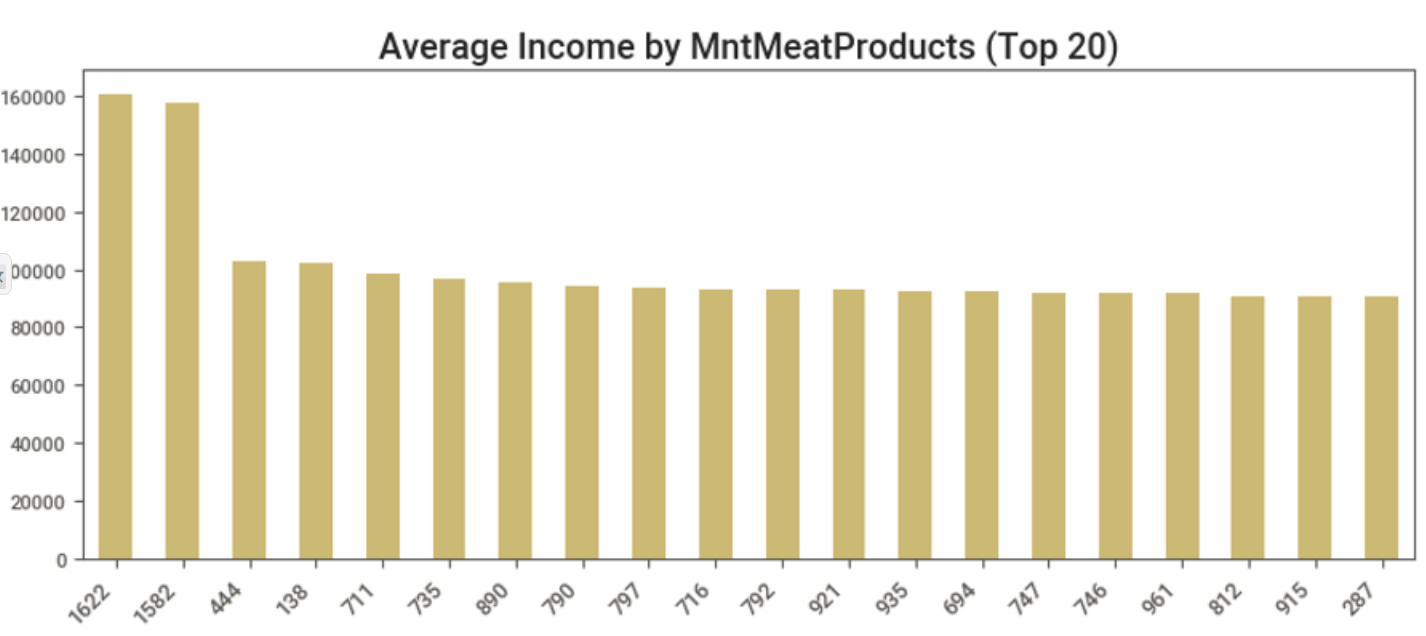


Hình 5.2.3. Ba biểu đồ cột top 20 mua qua 3 hình thức

* **Nhận xét:**
* Nhóm khách hàng có thu nhập cao (trên 100.000) thường có số lần mua qua Catalog hoặc Website nhiều hơn, cho thấy họ ưa chuộng sự tiện lợi và quen thuộc với thương hiệu.
* Ngược lại, nhóm thu nhập trung bình - thấp (dưới 100.000) có xu hướng mua hàng trực tiếp tại cửa hàng, có thể vì họ muốn xem sản phẩm thực tế hoặc ít sử dụng các kênh mua sắm trực tuyến.
* **Kết luận:** Công ty nên phân hóa chiến lược theo nhóm thu nhập:
* Nhóm thu nhập cao (>100.000): tập trung phát triển kênh mua hàng trực tuyến và catalog, cung cấp sản phẩm cao cấp và ưu đãi cá nhân hóa.
* Nhóm thu nhập trung bình - thấp (<100.000): cải thiện trải nghiệm tại cửa hàng, đồng thời khuyến khích chuyển dần sang kênh online qua các chương trình giảm giá hoặc hỗ trợ mua hàng.

***Câu hỏi 4:*** Khách hàng có thu nhập cao có thường mua rượu, thịt hay không?





Hình 5.2.4. Hai biểu đồ cột top 20 mua thịt và rượu

* **Nhận xét:**
* Biểu đồ "Average Income by MntWines" cho thấy các khách hàng có chi tiêu cho rượu là những người có thu nhập từ 90000 trở lên. Nhóm khách hàng có thu nhập từ trung bình đến cao có xu hướng chi tiêu mạnh cho rượu.
* Biểu đồ "Average Income by MntMeatProducts" tồn tại 2 khách hàng chi tiêu nhiều nhất cho rượu( 1622,1582) có mức lương cao đột biến(gần 160000). Còn những khách hàng còn lại thì có mức lương trên 90000.
* **Kết luận:** Những khách hàng có thu nhập từ 90.000 trở lên là nhóm chi tiêu mạnh cho rượu và thịt. Đặc biệt, hai khách hàng có mức chi tiêu cao nhất cho rượu cũng sở hữu thu nhập rất cao (gần 160.000), cho thấy mối liên hệ rõ ràng giữa thu nhập và mức chi tiêu cho các sản phẩm cao cấp như rượu và thịt. vì vậy cần Tập trung quảng bá và phát triển các sản phẩm rượu, thịt cao cấp hướng đến nhóm khách hàng thu nhập cao, kèm ưu đãi thành viên hoặc quà tặng để tăng gắn kết và doanh thu.

**6 .Kết luận**

- Hành vi mua sắm được quyết định chủ yếu bởi hai yếu tố cốt lõi: Thu nhập (Income) và Cấu trúc gia đình (có con cái hay không). Các yếu tố khác như học vấn, tuổi tác thường chỉ là hệ quả hoặc yếu tố phụ trợ cho hai trục chính này.

- Dưới đây là hai chân dung khách hàng đối lập và rõ nét nhất:

* Chân Dung Khách Hàng Thượng Lưu (High-Income Customer) Đây là nhóm khách hàng có giá trị cao nhất, được định hình bởi thu nhập dồi dào và lối sống riêng.
  + Hồ sơ:
    - Thu nhập: Rất cao (thường từ 90.000 trở lên, có những cá nhân vượt trội trên 150.000).
    - Học vấn: Cao (Đại học, Thạc sĩ, Tiến sĩ).
    - Gia đình: Thường không có con nhỏ ở nhà (Kidhome = 0).
  + Hành vi và tâm lý:
    - Chi tiêu cho chất lượng và sự tận hưởng: Họ không ngần ngại chi mạnh tay cho các sản phẩm cao cấp như thịt và đặc biệt là rượu vang. Giỏ hàng của họ thường bao gồm nhiều loại sản phẩm giá trị cao mua cùng lúc.
    - Ít nhạy cảm về giá: Họ gần như không quan tâm đến các chương trình khuyến mãi (NumDealsPurchases gần bằng 0). Họ sẵn sàng trả giá đầy đủ cho sản phẩm và dịch vụ xứng tầm.
    - Ưa chuộng mua sắm qua catalog và website: Họ có xu hướng mua hàng qua Catalog và website. Nhưng lại ít truy cập vào trang web công ty (NumWebVisitsMonth thấp). Điều này cho thấy họ mua sắm online một cách có chủ đích, khi đã biết mình muốn gì, thay vì tìm kiếm và khám phá.
* Chân Dung Khách Hàng Tiết Kiệm (Budget-Conscious Customer) Đây là nhóm khách hàng chiếm số đông, với hành vi mua sắm bị chi phối mạnh mẽ bởi ngân sách hạn hẹp.
  + Hồ sơ:
    - Thu nhập: Thấp đến trung bình (phần lớn dưới 10.000).
    - Gia đình: Thường có con nhỏ hoặc con ở tuổi vị thành niên.
  + Hành vi và tâm lý:
    - Nhạy cảm cao về giá: Quyết định mua hàng của họ bị chi phối mạnh mẽ bởi các chương trình khuyến mãi và ưu đãi (NumDealsPurchases cao).
    - Ưu tiên các sản phẩm thiết yếu: Ngân sách của họ tập trung vào các mặt hàng cơ bản, chi tiêu cho các sản phẩm xa xỉ như rượu vang là rất hạn chế.
    - Kênh mua sắm chính là cửa hàng: Họ có xu hướng mua trực tiếp tại cửa hàng để có thể xem xét sản phẩm và so sánh giá cả.
* Các Yếu Tố Ảnh Hưởng Quan Trọng Khác Tần suất mua hàng gần đây (Recency): Đây là một chỉ báo quan trọng về mức độ tương tác của khách hàng. Những người mua hàng gần đây (Recency thấp) có khả năng phản hồi cao nhất với các chiến dịch marketing. Ngược lại, những người có Recency cao là nhóm có nguy cơ rời bỏ.
* Các yếu tố ít ảnh hưởng: Dữ liệu đã chỉ ra rằng Tình trạng hôn nhân không ảnh hưởng nhiều đến hành vi chi tiêu. Tương tự, biến Khiếu nại (Complain) có tỷ lệ quá thấp nên không có giá trị trong việc phân khúc hay dự đoán.
* Kết Luận và Gợi Ý Chiến Lược Tổng Thể: yừ những phân tích trên, chiến lược kinh doanh cần được phân hóa rõ rệt thay vì áp dụng một cách tiếp cận chung.
* Đối với nhóm Khách Hàng Thượng Lưu:
  + Sản phẩm: Tập trung vào các dòng sản phẩm cao cấp, độc quyền, đặc biệt là rượu và thịt.
  + Kênh phân phối: Đầu tư vào website thương mại điện tử với giao diện cao cấp, trải nghiệm mượt mà, tích hợp các gói ưu đãi VIP hoặc chương trình thành viên thân thiết. Gửi catalog số / email cá nhân hóa, kèm hình ảnh sản phẩm sang trọng, mô tả chi tiết.
  + Marketing: Xây dựng chương trình khách hàng thân thiết (VIP) với các đặc quyền, quà tặng giá trị thay vì giảm giá.
* Đối với nhóm Khách Hàng Tiết Kiệm:
  + Sản phẩm: Cung cấp các sản phẩm thiết yếu với mức giá cạnh tranh, các gói combo gia đình.
  + Kênh phân phối: Tối ưu hóa trưng bày và trải nghiệm tại cửa hàng vật lý.
  + Marketing: Chiến lược giá và khuyến mãi là chìa khóa. Thường xuyên tung ra các chương trình giảm giá, ưu đãi, mua một tặng một để kích thích mua sắm.
  + Chiến Lược Dựa Trên Thời Điểm (Timing):
* Tận dụng Recency: Triển khai các chiến dịch marketing (email, SMS) nhắm vào nhóm khách hàng vừa mua sắm (Recency thấp) để khuyến khích họ quay lại. Đồng thời, xây dựng các chiến dịch "kích hoạt lại" cho nhóm khách hàng đã lâu không mua sắm (Recency cao) để kéo họ trở lại.