**1. Giới thiệu**

**1.1 Lý do chọn đề tài**

Trong bối cảnh thị trường ngày càng cạnh tranh, việc hiểu rõ tâm lý và hành vi của người tiêu dùng là yếu tố then chốt giúp các doanh nghiệp xây dựng và duy trì lợi thế cạnh tranh. Hành vi tiêu dùng không ngừng thay đổi theo thời gian, do ảnh hưởng từ nhiều yếu tố như sự phát triển của công nghệ, biến động kinh tế, văn hóa xã hội, và sự xuất hiện của các xu hướng mới. Những thay đổi này đặt ra thách thức cho doanh nghiệp trong việc phải liên tục điều chỉnh chiến lược marketing để đáp ứng nhu cầu và mong muốn của khách hàng.

Trong thời đại kỹ thuật số, việc phân tích dữ liệu đã trở thành công cụ không thể thiếu trong việc nắm bắt và dự đoán hành vi người tiêu dùng. Phân tích dữ liệu cung cấp cái nhìn sâu sắc về các mô hình tiêu dùng, cho phép doanh nghiệp không chỉ hiểu rõ hơn về khách hàng của mình mà còn có thể đưa ra các quyết định marketing chính xác và kịp thời. Bằng cách sử dụng phân tích dữ liệu, doanh nghiệp có thể xác định những yếu tố tâm lý ảnh hưởng đến quyết định mua sắm, từ đó phát triển các chiến lược marketing cá nhân hóa, hiệu quả hơn, nhằm tối đa hóa lợi nhuận và tăng cường sự hài lòng của khách hàng.

Với tầm quan trọng này, việc nghiên cứu và ứng dụng phân tích dữ liệu về tâm lý người tiêu dùng trong chiến lược marketing là vô cùng cần thiết. Đề tài này nhằm mục tiêu phân tích sâu sắc dữ liệu về hành vi và tâm lý người tiêu dùng, từ đó đề xuất những chiến lược marketing phù hợp và hiệu quả hơn cho doanh nghiệp.

**1.2 Mục tiêu nghiên cứu**

Mục tiêu của nghiên cứu này là:

* **Phân tích dữ liệu về hành vi và tâm lý người tiêu dùng:** Thông qua việc thu thập và phân tích các dữ liệu liên quan đến hành vi và tâm lý của người tiêu dùng, nghiên cứu sẽ tập trung vào việc làm rõ các đặc điểm chính, xu hướng và mô hình tiêu dùng. Phân tích này sẽ giúp hiểu rõ hơn về cách người tiêu dùng ra quyết định mua sắm và các yếu tố ảnh hưởng đến hành vi của họ.
* **Xác định các yếu tố ảnh hưởng đến quyết định mua sắm:** Nghiên cứu sẽ đi sâu vào việc nhận diện và phân tích các yếu tố chính tác động đến quyết định mua hàng của người tiêu dùng, bao gồm yếu tố tâm lý, kinh tế, xã hội và văn hóa. Việc xác định các yếu tố này sẽ giúp doanh nghiệp hiểu rõ hơn về động lực và rào cản trong hành vi tiêu dùng.
* **Đề xuất các chiến lược marketing dựa trên kết quả phân tích:** Dựa trên kết quả của quá trình phân tích dữ liệu, nghiên cứu sẽ đề xuất các chiến lược marketing cụ thể nhằm tối ưu hóa hiệu quả của các chiến dịch quảng cáo, phân phối, và phát triển sản phẩm. Những chiến lược này sẽ được thiết kế để phù hợp với nhu cầu và mong muốn của từng nhóm người tiêu dùng, từ đó giúp doanh nghiệp cải thiện khả năng cạnh tranh trên thị trường.

**2. Cơ sở lý thuyết**

**2.1 Khái niệm về tâm lý người tiêu dùng**

* **Tâm lý học tiêu dùng:** Tâm lý học tiêu dùng là một nhánh của tâm lý học ứng dụng, nghiên cứu về cách thức mà con người suy nghĩ, cảm nhận, và hành động trong quá trình tiêu dùng hàng hóa và dịch vụ. Nó liên quan đến việc tìm hiểu động cơ, thái độ, và quyết định của người tiêu dùng, cũng như cách họ tương tác với các yếu tố tiếp thị như quảng cáo, thương hiệu, và môi trường bán hàng. Tâm lý học tiêu dùng giúp giải thích tại sao người tiêu dùng lựa chọn một sản phẩm hoặc dịch vụ nhất định và làm thế nào để doanh nghiệp có thể tác động đến quyết định mua sắm của họ.
* **Các yếu tố ảnh hưởng đến hành vi tiêu dùng:** Hành vi tiêu dùng không chỉ đơn thuần là kết quả của nhu cầu cá nhân mà còn bị chi phối bởi nhiều yếu tố khác nhau. Các yếu tố này có thể được phân thành các nhóm chính bao gồm:
  + **Yếu tố cá nhân:** Bao gồm đặc điểm cá nhân như tuổi tác, giới tính, nghề nghiệp, thu nhập, và giai đoạn trong vòng đời. Những yếu tố này ảnh hưởng đến sự ưu tiên và quyết định mua sắm của mỗi cá nhân.
  + **Yếu tố tâm lý:** Gồm động lực, nhận thức, học hỏi, niềm tin, và thái độ của người tiêu dùng. Đây là các yếu tố then chốt trong việc hiểu rõ lý do tại sao người tiêu dùng chọn mua hoặc từ chối một sản phẩm hay dịch vụ cụ thể.
  + **Yếu tố xã hội:** Bao gồm ảnh hưởng từ gia đình, bạn bè, và các nhóm xã hội mà người tiêu dùng tham gia. Sự tác động từ những người xung quanh có thể ảnh hưởng mạnh mẽ đến quyết định tiêu dùng của một cá nhân.
  + **Yếu tố văn hóa:** Văn hóa và các giá trị xã hội là nền tảng quan trọng định hình hành vi tiêu dùng. Các yếu tố này có thể bao gồm tín ngưỡng, phong tục, và các chuẩn mực xã hội, tạo ra sự khác biệt trong hành vi tiêu dùng giữa các quốc gia và khu vực.

Hiểu rõ các khái niệm và yếu tố này là cơ sở để phân tích sâu hơn về hành vi người tiêu dùng, từ đó giúp doanh nghiệp phát triển các chiến lược marketing phù hợp và hiệu quả.

**2.2 Ứng dụng phân tích dữ liệu trong marketing**

* **Vai trò của phân tích dữ liệu trong việc hiểu tâm lý người tiêu dùng:**

Phân tích dữ liệu đóng vai trò then chốt trong việc hiểu và dự đoán hành vi cũng như tâm lý của người tiêu dùng. Thông qua việc thu thập, xử lý và phân tích các dữ liệu từ nhiều nguồn khác nhau như lịch sử mua sắm, hành vi trên website, phản hồi của khách hàng, và dữ liệu mạng xã hội, doanh nghiệp có thể khám phá ra những mô hình và xu hướng tiêu dùng mà trước đây khó có thể nhận ra. Việc hiểu sâu sắc những hành vi này không chỉ giúp doanh nghiệp dự đoán nhu cầu của khách hàng mà còn tạo ra những trải nghiệm cá nhân hóa, tăng cường mối quan hệ giữa thương hiệu và người tiêu dùng. Phân tích dữ liệu còn cho phép doanh nghiệp đánh giá hiệu quả của các chiến dịch marketing, từ đó điều chỉnh chiến lược để đạt được kết quả tốt hơn.

* **Các kỹ thuật phân tích dữ liệu thường được sử dụng trong marketing:**
  + **Phân tích mô tả (Descriptive Analytics):** Kỹ thuật này tập trung vào việc tóm tắt dữ liệu lịch sử để hiểu rõ hơn về những gì đã xảy ra. Nó thường được sử dụng để xác định các đặc điểm cơ bản của khách hàng, phân loại nhóm khách hàng dựa trên hành vi tiêu dùng, và xác định các xu hướng chung trong dữ liệu. Các công cụ như biểu đồ, bảng phân tích, và thống kê mô tả thường được sử dụng trong loại phân tích này.
  + **Phân tích dự đoán (Predictive Analytics):** Đây là một bước tiến xa hơn so với phân tích mô tả, tập trung vào việc dự đoán các hành vi hoặc kết quả tương lai dựa trên dữ liệu lịch sử. Các mô hình dự đoán như hồi quy tuyến tính, hồi quy logistic, và các kỹ thuật học máy (machine learning) như cây quyết định, mạng nơ-ron nhân tạo thường được áp dụng để dự đoán nhu cầu của khách hàng, xác suất chuyển đổi, hoặc phản ứng của khách hàng đối với các chiến dịch marketing.
  + **Phân tích phân cụm (Clustering Analysis):** Phân cụm là kỹ thuật phân nhóm dữ liệu theo các đặc điểm giống nhau. Trong marketing, phân tích phân cụm được sử dụng để nhóm người tiêu dùng thành các phân khúc khác nhau dựa trên hành vi mua sắm, tâm lý, hoặc nhân khẩu học. Các thuật toán như K-Means, Agglomerative Clustering, và DBSCAN thường được sử dụng để thực hiện phân cụm dữ liệu.
  + **Phân tích hồi quy (Regression Analysis):** Hồi quy là một kỹ thuật thống kê được sử dụng để xác định mối quan hệ giữa các biến số và dự đoán giá trị của một biến dựa trên các biến khác. Trong marketing, phân tích hồi quy giúp doanh nghiệp hiểu rõ hơn về những yếu tố ảnh hưởng đến quyết định mua sắm của khách hàng và đánh giá tác động của các yếu tố này đối với kết quả kinh doanh.
  + **Phân tích tần suất và chuỗi thời gian (Frequency and Time Series Analysis):** Kỹ thuật này được sử dụng để phân tích tần suất mua sắm và các xu hướng theo thời gian. Phân tích chuỗi thời gian giúp doanh nghiệp nhận diện các mô hình mùa vụ, xu hướng dài hạn và ngắn hạn trong dữ liệu tiêu dùng, từ đó đưa ra các quyết định marketing chính xác hơn.

Những kỹ thuật phân tích dữ liệu này không chỉ giúp doanh nghiệp hiểu rõ hơn về tâm lý người tiêu dùng mà còn là công cụ mạnh mẽ để tối ưu hóa các chiến lược marketing, từ đó đạt được hiệu quả cao hơn trong việc tiếp cận và phục vụ khách hàng.

**2.3. Lý thuyết phân tích**

**2.3.1. Phân tích hồi quy**

**2.3.1.1. Hồi quy tuyến tính**

Hồi quy tuyến tính là một kỹ thuật thống kê nhằm mô hình hóa mối quan hệ giữa một biến phụ thuộc  và một hoặc nhiều biến độc lập . Mô hình hồi quy tuyến tính đơn giản được thể hiện dưới dạng:



Trong đó:

* y là biến phụ thuộc (kết quả).
* x1​ là biến độc lập (dự báo).
* β0​ là hệ số chặn (intercept).
* β1​ là hệ số hồi quy (slope).
* ϵ là sai số (error term).

Nếu có nhiều biến độc lập, mô hình hồi quy tuyến tính bội sẽ được biểu diễn dưới dạng:



**2.3.1.2. Hồi quy Logistic**

Hồi quy logistic được sử dụng khi biến phụ thuộc là biến nhị phân (binary), có giá trị 0 hoặc 1. Mô hình logistic biểu diễn xác suất biến phụ thuộc nhận giá trị 1 dưới dạng:



Trong đó:

* P(y=1∣x) là xác suất biến phụ thuộc y bằng 1, dựa trên các biến độc lập ​.
* e là số mũ cơ sở tự nhiên (khoảng 2.718).
* β0​,β1​,...,βn​ là các hệ số hồi quy.

**2.3.1.2. Hệ số xác định ( và Adjusted )**

Hệ số xác định **** đo lường mức độ phù hợp của mô hình hồi quy tuyến tính. Nó được tính bằng công thức:



Trong đó:

* ​ là giá trị thực của biến phụ thuộc.
*  là giá trị dự đoán từ mô hình.
* ​ là giá trị trung bình của y.
* n là số lượng quan sát.

**Adjusted ** điều chỉnh ****cho số lượng biến độc lập trong mô hình và được tính bằng công thức:



Trong đó k là số biến độc lập và n là số lượng quan sát.

**2.3.2. Phân tích cụm (Clustering Analysis)**

**2.3.2.1 Phân cụm K-Means:**

Phân cụm K-Means là một phương pháp phân chia dữ liệu thành KKK cụm dựa trên khoảng cách trung bình của các điểm dữ liệu tới trung tâm cụm (centroid). Phương pháp này tối ưu hóa tiêu chí nội cụm (inertia) bằng cách giảm tổng bình phương khoảng cách từ mỗi điểm dữ liệu đến centroid của cụm nó thuộc về.

Thuật toán K-Means hoạt động theo các bước sau:

1. Chọn ngẫu nhiên K centroid ban đầu.
2. Gán mỗi điểm dữ liệu vào cụm có centroid gần nhất.
3. Tính toán lại các centroid dựa trên các điểm dữ liệu trong từng cụm.
4. Lặp lại các bước 2-3 cho đến khi centroid không còn thay đổi hoặc thay đổi rất nhỏ.

Công thức tính khoảng cách Euclidean giữa điểm dữ liệu  và centroid ​:



**2.3.2.2 Silhouette Score:**

Silhouette Score là một phép đo để đánh giá chất lượng của các cụm được tạo ra. Nó được định nghĩa cho mỗi điểm dữ liệu như sau:



Trong đó:

* a(i) là khoảng cách trung bình từ điểm dữ liệu iii đến tất cả các điểm khác trong cùng cụm.
* b(i) là khoảng cách trung bình từ điểm dữ liệu iii đến tất cả các điểm trong cụm gần nhất khác.

Giá trị s(i) nằm trong khoảng từ -1 đến 1. Giá trị càng gần 1, cụm càng tốt.

**2.3.2.3 Phương pháp Elbow:**

Phương pháp Elbow được sử dụng để xác định số lượng cụm tối ưu trong K-Means. Nó bao gồm việc tính toán tổng bình phương sai số nội cụm (inertia) cho các giá trị khác nhau của K và chọn K tại điểm mà độ dốc của đường cong inertia bắt đầu giảm dần (hình dạng giống như khuỷu tay - "elbow").

**2.3.3. Phân tích tần suất và xu hướng (Frequency and Trend Analysis)**

**2.3.1 Phân tích tần suất:**

Phân tích tần suất đo lường mức độ lặp lại của một hành vi hoặc sự kiện. Tần suất có thể được biểu diễn bằng các công thức thống kê cơ bản như:



Ví dụ: Tần suất mua sắm có thể được tính bằng cách đếm số lần một khách hàng mua sản phẩm trong một khoảng thời gian cụ thể.

**2.3.2 Phân tích chuỗi thời gian (Time Series Analysis):**

Phân tích chuỗi thời gian được sử dụng để phân tích các xu hướng theo thời gian. Một mô hình đơn giản có thể là:



Trong đó:

*  là giá trị tại thời điểm ttt.
*  là trung bình của chuỗi.
*  là hệ số tự tương quan bậc nhất.
*  là sai số.

Mô hình này giúp dự đoán giá trị tương lai dựa trên dữ liệu lịch sử và có thể được sử dụng để phát hiện các xu hướng mùa vụ, chu kỳ, và các yếu tố khác.

Những lý thuyết và công thức này cung cấp nền tảng khoa học vững chắc để thực hiện các phân tích dữ liệu trong nghiên cứu, từ đó giúp đưa ra các kết luận chính xác và hỗ trợ cho việc xây dựng các chiến lược marketing hiệu quả.

**3.1 Dữ liệu sử dụng**

Bộ dữ liệu sử dụng trong nghiên cứu này được thu thập từ nguồn công khai tại [Mendeley Data](https://plu.mx/plum/a?mendeley_data_id=mpd6m6jp4f&theme=plum-bigben-theme) và được chuyển đổi từ định dạng SPSS (.sav) sang định dạng CSV để phân tích bằng Python. Bộ dữ liệu này bao gồm 347 quan sát với 56 biến số, được chia thành các nhóm chính như sau:

1. **Nhân khẩu học:**
   * **Age (Tuổi):** Biến số này mã hóa tuổi của người tham gia theo một thang điểm (có thể là nhóm tuổi).
   * **Education (Trình độ học vấn):** Trình độ học vấn của người tham gia, được mã hóa theo các mức độ khác nhau.
   * **MaritalStatus (Tình trạng hôn nhân):** Tình trạng hôn nhân của người tham gia (độc thân, đã kết hôn, v.v.), được mã hóa thành các giá trị số.
   * **Occupation (Nghề nghiệp):** Nghề nghiệp của người tham gia, được mã hóa thành các giá trị số để đại diện cho các ngành nghề khác nhau.
   * **Income (Thu nhập):** Thu nhập của người tham gia, mã hóa thành các nhóm thu nhập khác nhau.
2. **Hành vi tiêu dùng trực tuyến:**
   * **Hours\_spent\_daily\_online (Số giờ dành trực tuyến mỗi ngày):** Số giờ trung bình mỗi ngày người tham gia sử dụng internet.
   * **Follow\_favorite\_brands\_online (Theo dõi các thương hiệu yêu thích trực tuyến):** Tần suất hoặc mức độ người tham gia theo dõi các thương hiệu yêu thích trên mạng, được mã hóa thành các giá trị số.
   * **Interaction\_with\_brands\_over\_internet (Tương tác với các thương hiệu qua internet):** Mức độ tương tác của người tham gia với các thương hiệu trên internet, mã hóa theo mức độ từ thấp đến cao.
3. **Thái độ đối với quảng cáo trên mạng xã hội và các thương hiệu trên website:**
   * **ATT\_FB\_Ads1 và ATT\_FB\_Ads2 (Thái độ đối với quảng cáo trên Facebook):** Mức độ thái độ của người tham gia đối với các quảng cáo trên Facebook, được đo lường bằng thang điểm Likert.
   * **ATT\_Web\_Brand1, ATT\_Web\_Brand2, ATT\_Web\_Brand3 (Thái độ đối với các thương hiệu trên website):** Thái độ của người tham gia đối với các thương hiệu mà họ tiếp xúc trên website, đo lường bằng các thang đo tương tự.
4. **Chi tiêu và phương thức thanh toán:**
   * **Monthly\_expenditure (Chi tiêu hàng tháng):** Tổng chi tiêu hàng tháng của người tham gia, mã hóa theo các mức độ khác nhau.
   * **Payment\_method (Phương thức thanh toán):** Loại phương thức thanh toán chính mà người tham gia sử dụng (tiền mặt, thẻ tín dụng, ví điện tử, v.v.), mã hóa thành các giá trị số.
5. **Biến số tương tác và các thang đo:**
   * **PCB\_1, PCB\_2, PCB\_3:** Các thành phần cấu thành của thang đo niềm tin vào khả năng kiểm soát hành vi (Perceived Behavioral Control).
   * **SN\_1, SN\_2, SN\_3 (Social Norms):** Các thang đo chuẩn mực xã hội ảnh hưởng đến hành vi của người tiêu dùng.
   * **InteractionVar1, InteractionVar2, InteractionVar3:** Các biến tương tác nhằm đo lường sự kết hợp giữa các yếu tố khác nhau trong hành vi tiêu dùng.
   * **IntAgATPA, IntAgSWB, IntAgSN, IntAgPCB:** Các biến số tổng hợp đo lường sự tương tác giữa các yếu tố thái độ, chuẩn mực xã hội, và kiểm soát hành vi.
6. **Biến chuẩn hóa và tổng hợp:**
   * **ZAge, ZEdu, ZMaSt, ZInc, ZOcc:** Các biến số chuẩn hóa của tuổi, giáo dục, tình trạng hôn nhân, thu nhập và nghề nghiệp.
   * **ZSNorms, ZPCB, ZFBat, ZOBB, ZASW, ZCustPI:** Các biến số chuẩn hóa của các thang đo chuẩn mực xã hội, kiểm soát hành vi, thái độ đối với Facebook, hành vi mua sắm trực tuyến, và ý định mua hàng.

Bộ dữ liệu này cung cấp một cái nhìn toàn diện về các yếu tố nhân khẩu học, hành vi trực tuyến, thái độ đối với quảng cáo, và các yếu tố tương tác khác nhau liên quan đến tâm lý người tiêu dùng. Việc phân tích sâu các biến số này sẽ giúp xác định các mối quan hệ tiềm năng và đưa ra những kết luận có giá trị cho việc phát triển chiến lược marketing hiệu quả.

**3.2 Công cụ và kỹ thuật phân tích**

* **Sử dụng ngôn ngữ lập trình Python:**

Python là ngôn ngữ lập trình mạnh mẽ và linh hoạt, đặc biệt phù hợp cho các nhiệm vụ phân tích dữ liệu nhờ vào hệ sinh thái phong phú của các thư viện chuyên dụng. Với cú pháp đơn giản và dễ đọc, Python cho phép nhà nghiên cứu dễ dàng thao tác với dữ liệu, thực hiện các phép tính toán phức tạp, và trực quan hóa kết quả phân tích một cách hiệu quả. Trong nghiên cứu này, Python sẽ được sử dụng làm công cụ chính để xử lý, phân tích và trực quan hóa dữ liệu.

* **Các thư viện Python:**
  + **Pandas:** Pandas là một thư viện mạnh mẽ để thao tác và phân tích dữ liệu, đặc biệt là dữ liệu dạng bảng (DataFrame). Thư viện này cung cấp các công cụ để đọc, viết, lọc, nhóm, và tổng hợp dữ liệu, giúp dễ dàng làm sạch và chuẩn bị dữ liệu cho các bước phân tích tiếp theo. Pandas sẽ được sử dụng để xử lý dữ liệu ban đầu, bao gồm các nhiệm vụ như kiểm tra giá trị thiếu, loại bỏ các giá trị ngoại lai, và chuyển đổi định dạng dữ liệu.
  + **Numpy:** Numpy là thư viện cốt lõi cho các phép toán số học và ma trận trong Python. Nó cung cấp các công cụ để làm việc với mảng số học (arrays) hiệu quả hơn, là nền tảng cho nhiều thao tác tính toán phức tạp trong phân tích dữ liệu. Numpy sẽ được sử dụng trong nghiên cứu này để thực hiện các phép toán cơ bản và nâng cao, giúp tối ưu hóa hiệu suất khi xử lý các tập dữ liệu lớn.
  + **Matplotlib:** Matplotlib là một thư viện phổ biến để tạo các biểu đồ và đồ thị trực quan trong Python. Nó cung cấp các công cụ mạnh mẽ để tạo các biểu đồ đa dạng như biểu đồ cột, biểu đồ đường, histogram, scatter plot, và nhiều loại biểu đồ khác. Trong nghiên cứu này, Matplotlib sẽ được sử dụng để trực quan hóa dữ liệu, giúp minh họa rõ ràng các kết quả phân tích và hỗ trợ việc đưa ra các kết luận từ dữ liệu.
  + **Seaborn:** Seaborn được xây dựng trên Matplotlib và cung cấp một giao diện trực quan hơn cho việc tạo các biểu đồ thống kê. Thư viện này cho phép tạo ra các biểu đồ phức tạp với ít mã hơn, đồng thời hỗ trợ trực quan hóa các mối quan hệ thống kê giữa các biến. Seaborn sẽ được sử dụng để vẽ các biểu đồ phân tán, biểu đồ phân bố và các biểu đồ phức hợp khác nhằm phân tích sâu hơn các mối quan hệ trong dữ liệu.
  + **Scikit-learn:** Scikit-learn là thư viện tiêu chuẩn cho học máy (machine learning) trong Python, cung cấp các thuật toán và công cụ để thực hiện các mô hình học máy như hồi quy, phân loại, phân cụm, và giảm chiều dữ liệu. Thư viện này cũng cung cấp các công cụ để đánh giá mô hình, chọn tham số tối ưu, và kiểm tra tính phù hợp của mô hình. Trong nghiên cứu này, Scikit-learn sẽ được sử dụng để thực hiện các phân tích hồi quy, phân cụm, và các kỹ thuật học máy khác nhằm tìm hiểu sâu hơn về các yếu tố ảnh hưởng đến hành vi người tiêu dùng và xây dựng các mô hình dự đoán.

Nhờ sự kết hợp của các công cụ và thư viện mạnh mẽ này, quá trình phân tích dữ liệu sẽ trở nên hiệu quả và dễ dàng hơn, cho phép nhà nghiên cứu khai thác tối đa giá trị từ bộ dữ liệu và đưa ra những kết luận chính xác, có cơ sở khoa học.

**3.3 Quy trình nghiên cứu**

Nghiên cứu sẽ tuân theo một quy trình phân tích dữ liệu chi tiết và có cấu trúc rõ ràng nhằm đảm bảo tính chính xác và hiệu quả trong việc khám phá và đánh giá các yếu tố ảnh hưởng đến hành vi tiêu dùng. Quy trình bao gồm các bước chính sau:

* **Thu thập và làm sạch dữ liệu:**
  + **Sử dụng pandas để đọc và làm sạch dữ liệu:** Bước đầu tiên trong quy trình phân tích là đọc dữ liệu từ tệp CSV bằng thư viện Pandas. Dữ liệu sẽ được tải vào một DataFrame để dễ dàng thao tác và xử lý. Các bước tiếp theo bao gồm:
    - Kiểm tra dữ liệu để đảm bảo không có lỗi trong quá trình chuyển đổi định dạng.
    - Kiểm tra cấu trúc và định dạng của các cột, đảm bảo rằng tất cả các biến đều được biểu diễn chính xác.
  + **Kiểm tra và xử lý giá trị thiếu:** Dữ liệu sẽ được kiểm tra để xác định các giá trị thiếu (NaN). Các giá trị thiếu có thể được xử lý bằng cách loại bỏ các hàng bị thiếu nhiều thông tin hoặc thay thế các giá trị thiếu bằng các phương pháp như giá trị trung bình, trung vị hoặc giá trị phổ biến nhất, tùy thuộc vào loại dữ liệu và mức độ thiếu.
* **Phân tích mô tả:**
  + **Sử dụng matplotlib và seaborn để vẽ các biểu đồ mô tả dữ liệu:** Sau khi dữ liệu được làm sạch, bước tiếp theo là thực hiện phân tích mô tả để có cái nhìn tổng quan về dữ liệu. Các biểu đồ sẽ được tạo ra để trực quan hóa các đặc điểm chính của dữ liệu, chẳng hạn như:
    - **Biểu đồ phân tán (scatter plots):** Để xem xét mối quan hệ giữa các cặp biến.
    - **Biểu đồ histogram:** Để hiểu rõ phân phối của các biến số liên tục.
    - **Biểu đồ boxplot:** Để phát hiện các giá trị ngoại lai và kiểm tra phân phối của dữ liệu theo các nhóm khác nhau.
  + **Thực hiện thống kê mô tả để hiểu rõ các đặc điểm cơ bản của dữ liệu:** Sử dụng Pandas và Numpy để tính toán các chỉ số thống kê như trung bình, trung vị, độ lệch chuẩn, và các chỉ số thống kê khác để nắm bắt các đặc điểm cơ bản của dữ liệu.
* **Phân tích hồi quy:**
  + **Sử dụng scikit-learn để thực hiện phân tích hồi quy tuyến tính và hồi quy logistic:** Để xác định mối quan hệ giữa các biến số và kết quả đầu ra (chẳng hạn như quyết định mua sắm), mô hình hồi quy tuyến tính và hồi quy logistic sẽ được xây dựng. Hồi quy tuyến tính được sử dụng cho các biến liên tục, trong khi hồi quy logistic được sử dụng cho các biến phân loại.
  + **Xác định các yếu tố ảnh hưởng đến quyết định mua sắm bằng cách sử dụng mô hình hồi quy bội:** Mô hình hồi quy bội sẽ được áp dụng để đánh giá tác động của nhiều biến số độc lập lên một biến số phụ thuộc (ví dụ: khả năng mua hàng).
  + **Kiểm định tính phù hợp của mô hình (R², adjusted R², p-value) và phân tích các hệ số hồi quy để xác định ý nghĩa của từng biến số:** Các chỉ số như R², adjusted R² và p-value sẽ được tính toán để đánh giá độ phù hợp của mô hình. Hệ số hồi quy của từng biến sẽ được phân tích để hiểu rõ ý nghĩa và tác động của chúng.
* **Phân tích cụm:**
  + **Sử dụng scikit-learn để phân loại người tiêu dùng theo các nhóm tâm lý khác nhau:** Phân tích cụm sẽ được sử dụng để nhóm các người tiêu dùng thành các phân khúc khác nhau dựa trên các đặc điểm hành vi và tâm lý.
  + **Áp dụng thuật toán K-Means, Agglomerative Clustering, hoặc DBSCAN để phân cụm dữ liệu:** Các thuật toán phân cụm như K-Means, Agglomerative Clustering, và DBSCAN sẽ được áp dụng để xác định các nhóm khách hàng có đặc điểm tương đồng.
  + **Đánh giá kết quả phân cụm bằng cách sử dụng Silhouette Score và phương pháp Elbow:** Silhouette Score sẽ được sử dụng để đánh giá mức độ tách biệt giữa các cụm, trong khi phương pháp Elbow sẽ giúp xác định số cụm tối ưu.
  + **Phân tích các đặc điểm của từng cụm để hiểu rõ hơn về các nhóm người tiêu dùng:** Các cụm sẽ được phân tích chi tiết để xác định các đặc điểm nổi bật của mỗi nhóm, giúp xây dựng các chiến lược marketing nhắm vào từng phân khúc khách hàng cụ thể.
* **Phân tích tần suất và xu hướng mua sắm:**
  + **Sử dụng pandas để tính toán và phân tích tần suất mua sắm của từng khách hàng:** Phân tích tần suất mua sắm sẽ giúp xác định các khách hàng có mức độ mua sắm cao, từ đó xây dựng chiến lược tiếp cận phù hợp.
  + **Phân tích xu hướng mua sắm theo mùa, loại hàng hóa và địa điểm:** Xu hướng mua sắm sẽ được phân tích để xác định các mô hình theo mùa, loại sản phẩm và địa điểm mua sắm, giúp doanh nghiệp tối ưu hóa các chiến lược kinh doanh.
  + **Sử dụng biểu đồ heatmap để trực quan hóa tần suất và xu hướng mua sắm:** Biểu đồ heatmap sẽ được sử dụng để minh họa các xu hướng và tần suất mua sắm, giúp nhận diện các khu vực và thời điểm có hoạt động mua sắm mạnh mẽ.

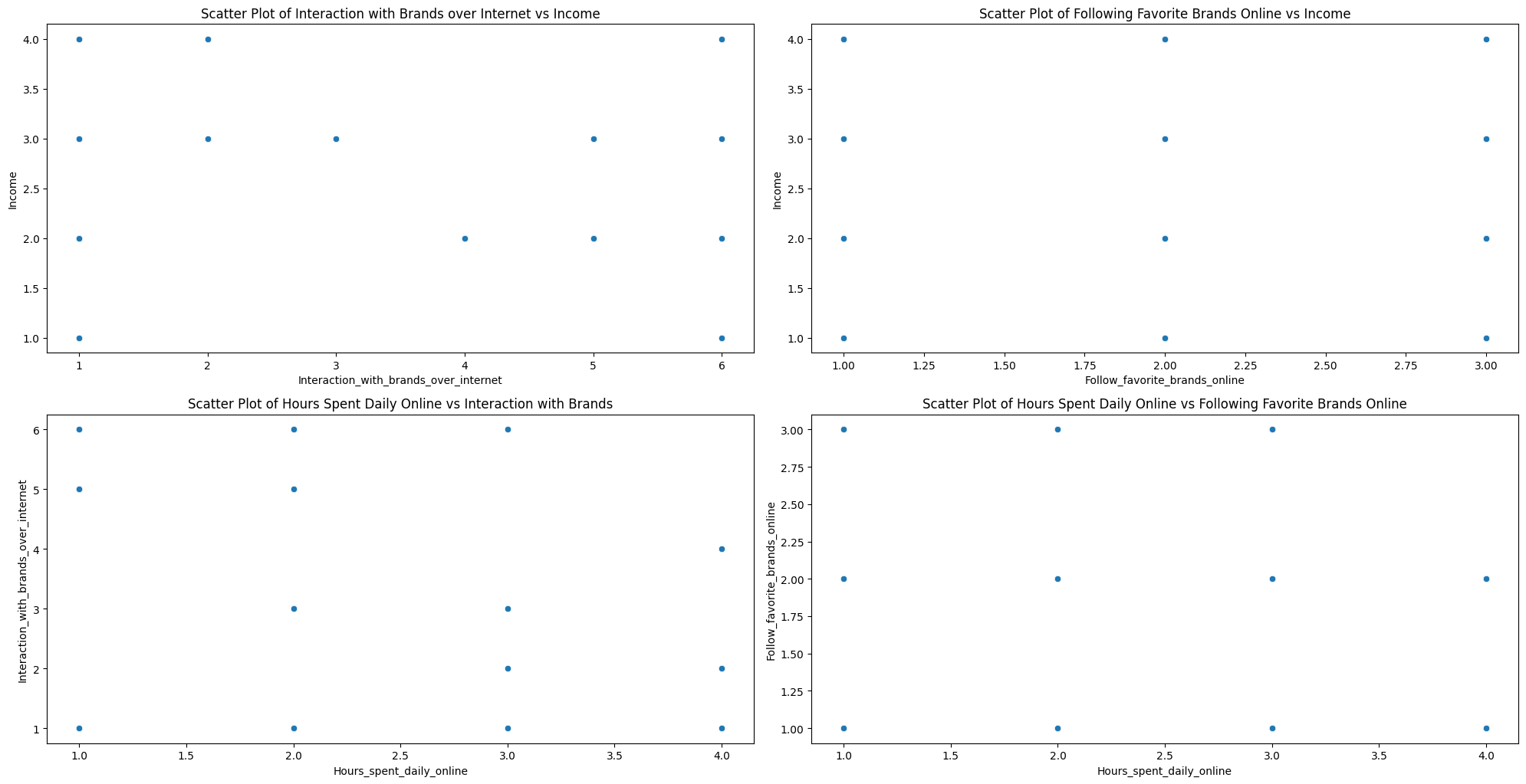
Quy trình nghiên cứu này đảm bảo việc xử lý và phân tích dữ liệu được thực hiện một cách có hệ thống, từ đó đưa ra các kết quả phân tích chính xác và có giá trị cho việc phát triển chiến lược marketing.

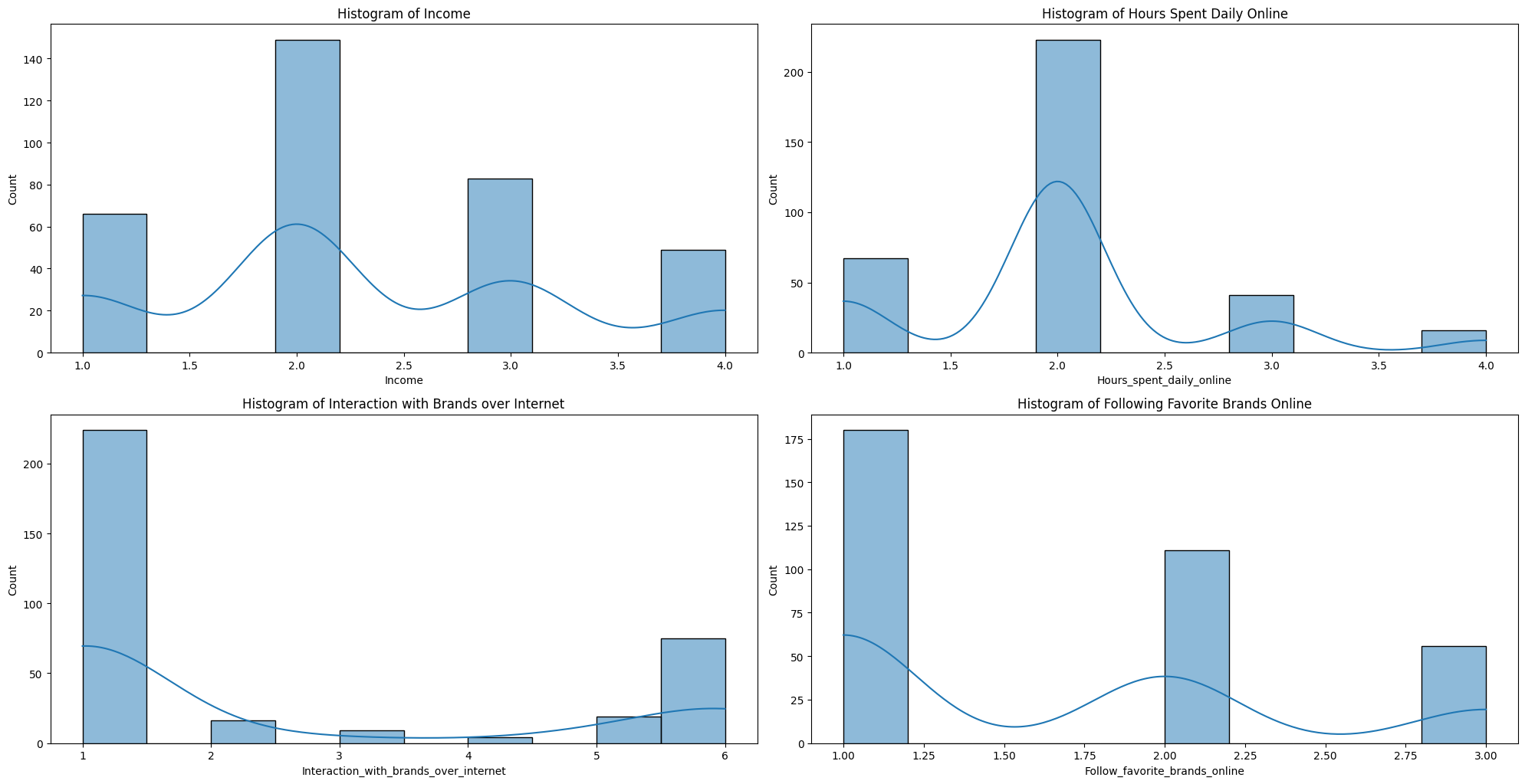
**4. Báo cáo kết quả chi tiết**

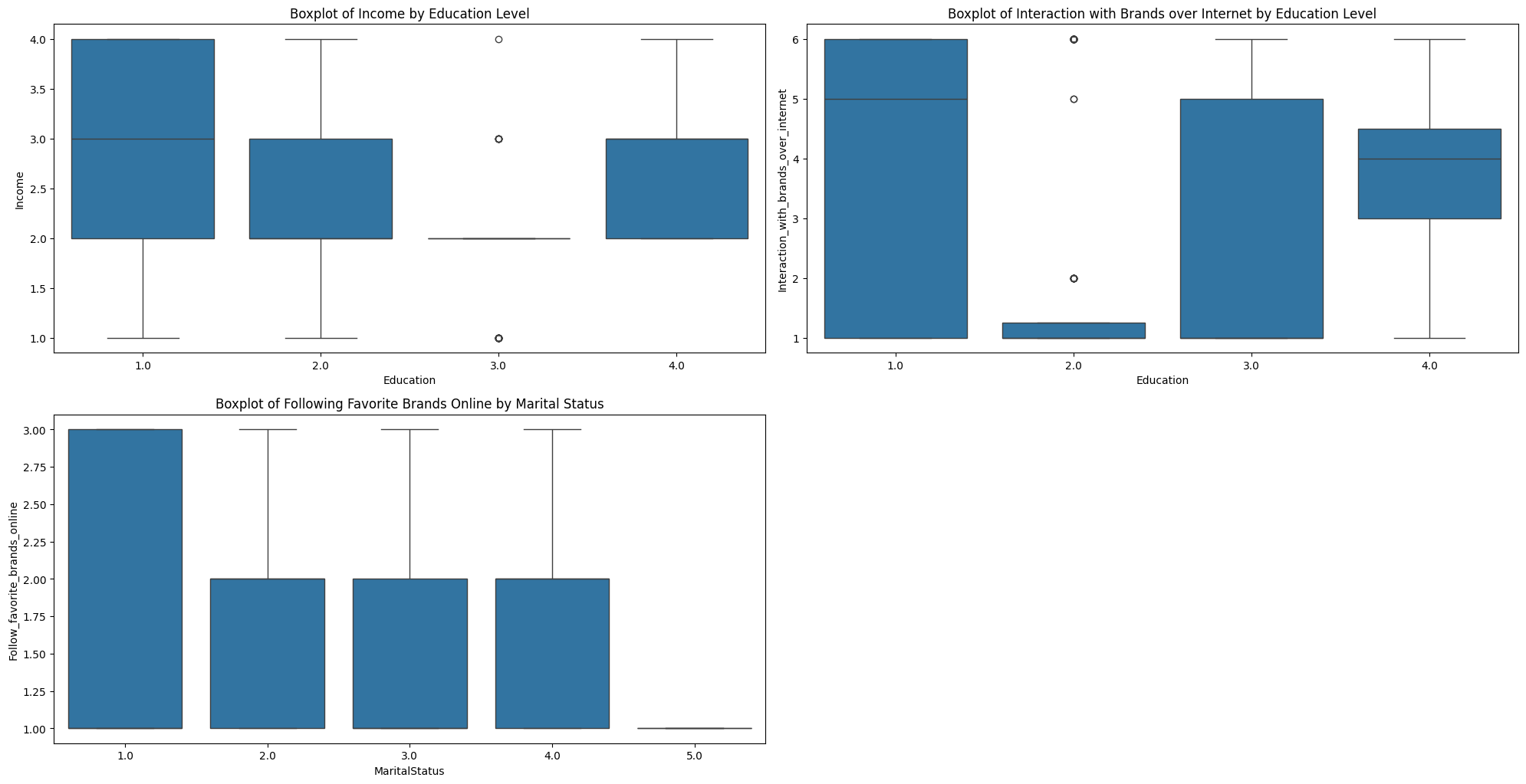
**4.1. Kết quả thu thập và làm sạch dữ liệu**

* **Đọc và làm sạch dữ liệu:** Bộ dữ liệu đã được đọc thành công từ tệp CSV và chứa tổng cộng 347 quan sát với 56 biến số. Các biến trong bộ dữ liệu đều thuộc kiểu số thực (float64).
* **Kiểm tra và xử lý giá trị thiếu:** Sau khi kiểm tra, không có giá trị thiếu (NaN) nào được phát hiện trong bộ dữ liệu. Điều này cho thấy dữ liệu đã sẵn sàng để tiến hành các bước phân tích tiếp theo.

**4.2. Kết quả phân tích mô tả**







1. **Biểu đồ phân tán (Scatter Plots):**

* **Tương tác với các thương hiệu qua internet (Interaction with Brands over Internet) và Thu nhập (Income):**
  + Biểu đồ này cho thấy rằng có sự phân tán khá rộng, không có mối quan hệ tuyến tính rõ ràng giữa tương tác với thương hiệu trực tuyến và thu nhập. Tuy nhiên, một số cá nhân có mức tương tác cao với thương hiệu cũng có xu hướng thu nhập cao hơn.
* **Theo dõi các thương hiệu yêu thích trực tuyến (Follow Favorite Brands Online) và Thu nhập (Income):**
  + Mối quan hệ giữa việc theo dõi thương hiệu yêu thích và thu nhập cũng không rõ ràng. Điều này có thể chỉ ra rằng việc theo dõi thương hiệu không hoàn toàn phụ thuộc vào mức thu nhập.
* **Số giờ trực tuyến mỗi ngày (Hours Spent Daily Online) và Tương tác với các thương hiệu qua internet:**
  + Có một số mối tương quan giữa việc dành nhiều thời gian trực tuyến và mức độ tương tác với các thương hiệu. Những người dành nhiều thời gian trực tuyến có xu hướng tương tác nhiều hơn với các thương hiệu.
* **Số giờ trực tuyến mỗi ngày (Hours Spent Daily Online) và Theo dõi các thương hiệu yêu thích trực tuyến:**
  + Biểu đồ cho thấy mối quan hệ giữa thời gian trực tuyến và việc theo dõi các thương hiệu yêu thích, tuy nhiên mối liên hệ này không phải là tuyến tính và có thể bị ảnh hưởng bởi các yếu tố khác.

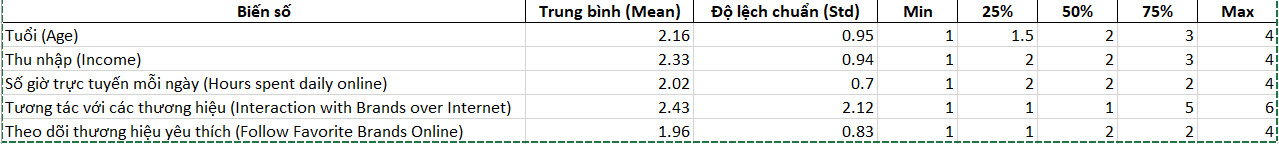
1. **Biểu đồ histogram:**

* **Thu nhập (Income):**
  + Phân phối thu nhập cho thấy phần lớn người tham gia có thu nhập tập trung ở mức trung bình, với một số ít người có thu nhập rất cao hoặc rất thấp.
* **Số giờ trực tuyến mỗi ngày (Hours Spent Daily Online):**
  + Phần lớn người tham gia dành khoảng 2 giờ trực tuyến mỗi ngày. Đây là một thông tin quan trọng cho việc xây dựng các chiến lược tiếp thị trực tuyến, vì nó xác định khoảng thời gian mà các chiến dịch có thể tiếp cận đối tượng mục tiêu.
* **Tương tác với các thương hiệu qua internet (Interaction with Brands over Internet):**
  + Đa số người tham gia có mức độ tương tác thấp với các thương hiệu, nhưng có một số ít người tương tác nhiều hơn, điều này có thể là nhóm khách hàng tiềm năng mà các thương hiệu nên tập trung vào.
* **Theo dõi các thương hiệu yêu thích trực tuyến (Follow Favorite Brands Online):**
  + Phân phối cho thấy mức độ theo dõi các thương hiệu yêu thích của người tham gia, với đa số tập trung ở mức trung bình.

1. **Biểu đồ boxplot:**

* **Thu nhập (Income) theo Trình độ học vấn (Education):**
  + Có một số mối quan hệ giữa trình độ học vấn và thu nhập, tuy nhiên có sự phân tán lớn, cho thấy rằng trình độ học vấn không phải là yếu tố quyết định duy nhất đối với thu nhập.
* **Tương tác với các thương hiệu qua internet (Interaction with Brands over Internet) theo Trình độ học vấn (Education):**
  + Biểu đồ này chỉ ra rằng mức độ tương tác với thương hiệu trực tuyến không bị ảnh hưởng nhiều bởi trình độ học vấn.
* **Theo dõi các thương hiệu yêu thích trực tuyến (Follow Favorite Brands Online) theo Tình trạng hôn nhân (Marital Status):**
  + Tình trạng hôn nhân không có ảnh hưởng rõ rệt đến việc theo dõi các thương hiệu yêu thích trực tuyến.

1. **Thống kê mô tả:**



**Phân tích kết quả:**

* **Tâm lý người tiêu dùng:** Các phân tích cho thấy rằng tương tác với các thương hiệu và việc theo dõi các thương hiệu yêu thích không hoàn toàn phụ thuộc vào các yếu tố nhân khẩu học như thu nhập, trình độ học vấn hay tình trạng hôn nhân. Điều này gợi ý rằng có những động lực tâm lý khác ảnh hưởng đến hành vi tiêu dùng trực tuyến.
* **Chiến lược marketing:** Với sự phân tán lớn trong các hành vi như tương tác với thương hiệu và theo dõi thương hiệu, các chiến lược marketing cần tập trung vào việc cá nhân hóa và tối ưu hóa trải nghiệm người dùng để tiếp cận các nhóm người tiêu dùng khác nhau một cách hiệu quả.

Những kết quả này có thể giúp đề xuất các chiến lược marketing chính xác hơn, dựa trên sự hiểu biết sâu sắc về tâm lý và hành vi của người tiêu dùng.

**4.3. Kết quả phân tích hồi quy**

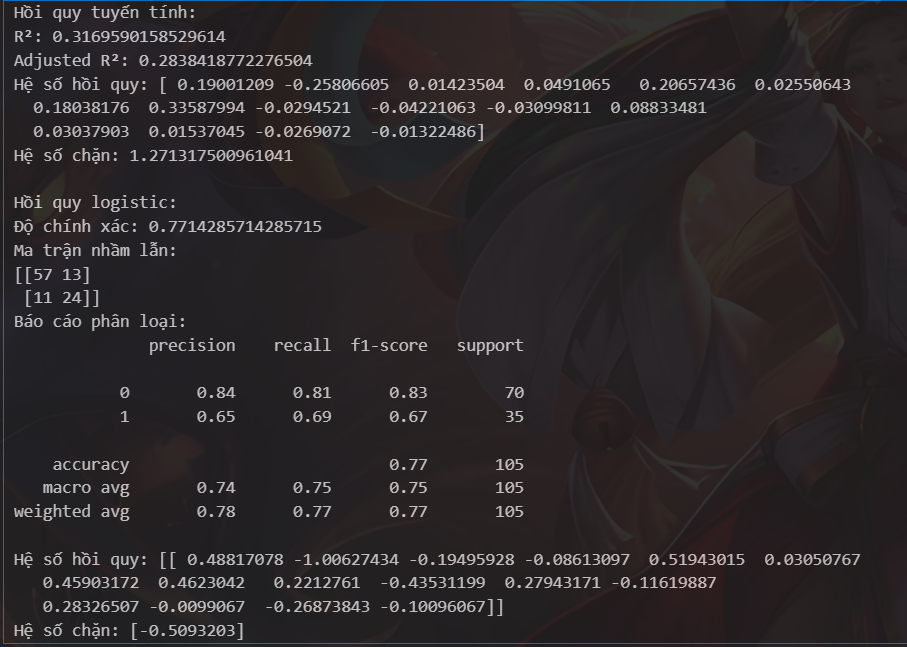
*Lựa chọn biến ?*

Để xây dựng mô hình hồi quy có ý nghĩa, chúng ta cần chọn lọc các biến có liên quan đến mục tiêu phân tích tâm lý và hành vi người tiêu dùng cũng như những yếu tố có thể ảnh hưởng đến thu nhập và quyết định mua sắm. Dưới đây là danh sách các biến có thể được chọn dựa trên ý nghĩa và mục tiêu của đề tài:

**Các biến tiềm năng:**

1. **Age**: Tuổi của người tiêu dùng - biến nhân khẩu học quan trọng.
2. **Education**: Trình độ học vấn - ảnh hưởng đến nhận thức và quyết định tiêu dùng.
3. **MaritalStatus**: Tình trạng hôn nhân - có thể ảnh hưởng đến hành vi tiêu dùng và chi tiêu.
4. **Occupation**: Nghề nghiệp - liên quan trực tiếp đến thu nhập.
5. **Income**: Thu nhập - biến phụ thuộc chính.
6. **Hours\_spent\_daily\_online**: Số giờ trực tuyến mỗi ngày - phản ánh hành vi trực tuyến.
7. **Follow\_favorite\_brands\_online**: Theo dõi thương hiệu yêu thích trực tuyến - thể hiện sự quan tâm đến thương hiệu.
8. **Interaction\_with\_brands\_over\_internet**: Tương tác với thương hiệu qua internet - mức độ tương tác trực tuyến.
9. **Monthly\_expenditure**: Chi tiêu hàng tháng - phản ánh hành vi tiêu dùng thực tế.
10. **Payment\_method**: Phương thức thanh toán - liên quan đến hành vi tiêu dùng.
11. **FBattitude**: Thái độ đối với Facebook Ads - liên quan đến hiệu quả của quảng cáo trực tuyến.
12. **CustPI**: Chỉ số cảm nhận khách hàng - liên quan đến trải nghiệm và sự hài lòng.
13. **InteractionVar1, InteractionVar2, InteractionVar3**: Các biến tương tác - có thể phản ánh các yếu tố tương tác phức tạp.
14. **PCB, SNorms**: Nhận thức và chuẩn mực xã hội - có thể ảnh hưởng đến hành vi mua sắm.

**Mục tiêu nghiên cứu của phân tích hồi quy:** Phân tích các yếu tố tác động đến thu nhập và quyết định mua sắm của người tiêu dùng dựa trên các đặc điểm nhân khẩu học và hành vi tiêu dùng.



**4.3.1. Hồi quy tuyến tính**

Hồi quy tuyến tính được thực hiện để xác định mối quan hệ giữa thu nhập và các biến số nhân khẩu học, hành vi tiêu dùng. Kết quả của mô hình hồi quy tuyến tính cho thấy:

* **R² = 0.317**: Mô hình này giải thích được khoảng 31.7% sự biến thiên của thu nhập dựa trên các biến độc lập đã chọn.
* **Adjusted R² = 0.284**: Sau khi điều chỉnh cho số lượng biến trong mô hình, khoảng 28.4% sự biến thiên của thu nhập được giải thích.

**Hệ số hồi quy:**

* **Tuổi (Age):** Hệ số dương (0.190), cho thấy rằng khi tuổi tăng, thu nhập có xu hướng tăng lên.
* **Trình độ học vấn (Education):** Hệ số âm (-0.258), cho thấy rằng trong tập dữ liệu này, người có trình độ học vấn cao hơn có thể không nhất thiết có thu nhập cao hơn, điều này có thể phản ánh các yếu tố văn hóa hoặc xã hội cụ thể.
* **Số giờ trực tuyến mỗi ngày (Hours spent daily online):** Hệ số dương (0.207), chỉ ra rằng thời gian trực tuyến nhiều hơn có liên quan đến thu nhập cao hơn.
* **Tương tác với thương hiệu qua internet (Interaction with brands over internet):** Hệ số dương (0.180), cho thấy rằng mức độ tương tác với thương hiệu có tác động tích cực đến thu nhập.
* **Chi tiêu hàng tháng (Monthly expenditure):** Hệ số dương (0.336), phản ánh rằng những người có mức chi tiêu hàng tháng cao hơn thường có thu nhập cao hơn.

**4.3.2. Hồi quy logistic**

Hồi quy logistic được sử dụng để dự đoán khả năng người tiêu dùng có thu nhập cao hơn trung vị dựa trên các đặc điểm nhân khẩu học và hành vi tiêu dùng.

**Độ chính xác của mô hình:**

* **Độ chính xác (Accuracy) = 77.1%**: Mô hình dự đoán đúng 77.1% trường hợp trong tập kiểm tra.

**Ma trận nhầm lẫn:**

* **True Negatives (TN):** 57 người có thu nhập dưới trung vị được dự đoán đúng.
* **False Positives (FP):** 13 người có thu nhập dưới trung vị nhưng được dự đoán là trên trung vị.
* **False Negatives (FN):** 11 người có thu nhập trên trung vị nhưng được dự đoán là dưới trung vị.
* **True Positives (TP):** 24 người có thu nhập trên trung vị được dự đoán đúng.

**Báo cáo phân loại:**

* **Precision (Lớp 0):** 0.84 - mô hình dự đoán khá tốt cho những người có thu nhập dưới trung vị.
* **Precision (Lớp 1):** 0.65 - mô hình có thể cải thiện độ chính xác cho nhóm người có thu nhập trên trung vị.
* **F1-score:** 0.67 cho lớp 1, cho thấy mô hình dự đoán chưa hoàn toàn tốt cho nhóm này.

**Hệ số hồi quy:**

* **Tuổi (Age):** Hệ số dương (0.488), chỉ ra rằng tuổi tác có tác động tích cực đến khả năng có thu nhập cao hơn.
* **Trình độ học vấn (Education):** Hệ số âm (-1.006), phản ánh tác động tiêu cực của trình độ học vấn cao hơn đối với khả năng có thu nhập cao hơn, có thể do các yếu tố khác như ngành nghề hoặc vị trí địa lý.
* **Tương tác với thương hiệu qua internet (Interaction with brands over internet):** Hệ số dương (0.459), cho thấy mối liên hệ tích cực giữa tương tác với thương hiệu và thu nhập cao.

**4.3.3. Kết luận và gợi ý chiến lược**

**Kết quả phân tích:**

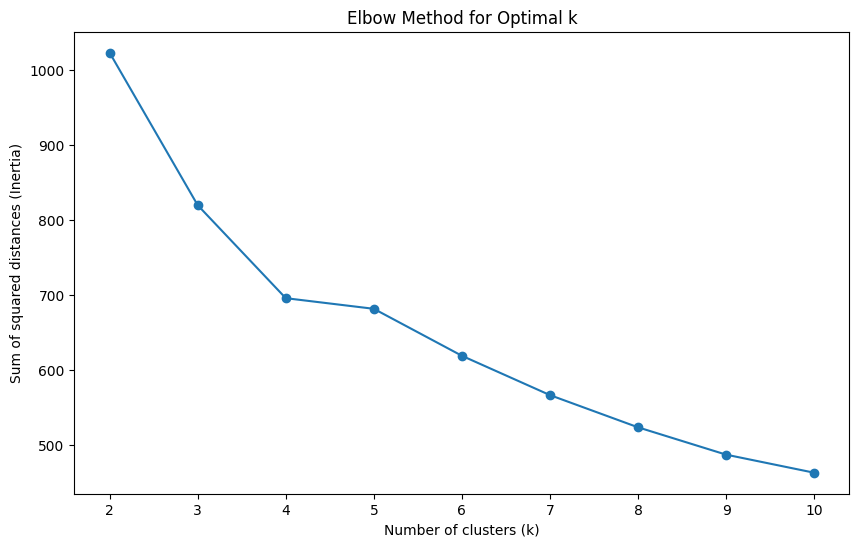
Kết quả từ cả hai mô hình hồi quy tuyến tính và logistic đều chỉ ra rằng tuổi, số giờ trực tuyến mỗi ngày, và tương tác với thương hiệu có ảnh hưởng đáng kể đến thu nhập của người tiêu dùng. Mặc dù trình độ học vấn có tác động tiêu cực trong cả hai mô hình, điều này có thể phản ánh các yếu tố cụ thể trong tập dữ liệu và cần được nghiên cứu thêm.

**Gợi ý chiến lược marketing:**

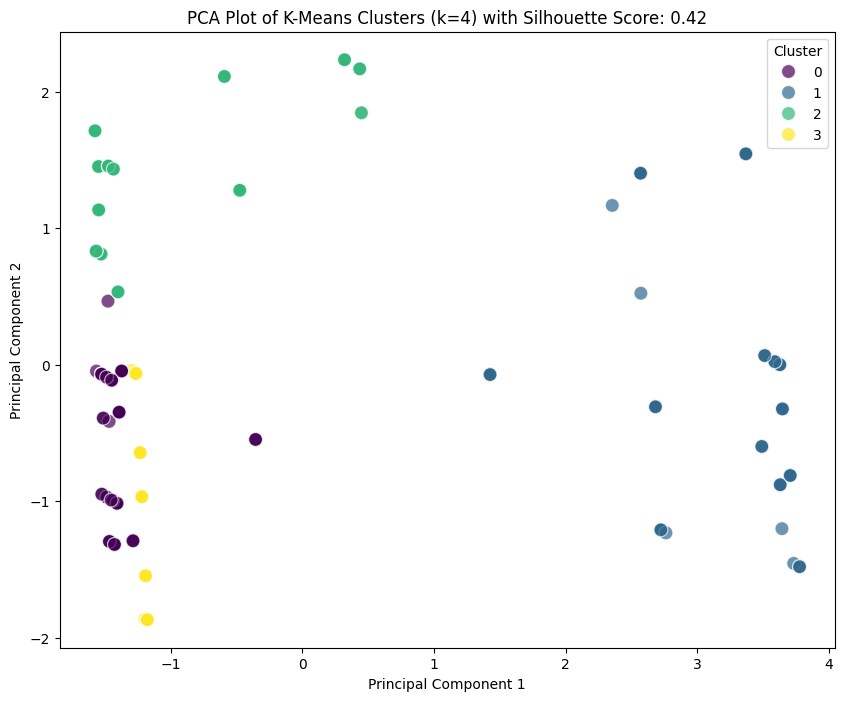
1. **Tập trung vào nhóm người tiêu dùng lớn tuổi:** Do tuổi tác có mối liên hệ tích cực với thu nhập, các chiến dịch marketing nên nhắm đến nhóm này với các sản phẩm và dịch vụ phù hợp.
2. **Tăng cường tương tác trực tuyến:** Các chiến dịch nên khuyến khích sự tương tác với thương hiệu qua internet, vì đây là yếu tố tích cực ảnh hưởng đến thu nhập và có thể gắn kết người tiêu dùng với thương hiệu.
3. **Xem xét lại chiến lược cho nhóm có học vấn cao:** Do tác động tiêu cực không mong muốn của trình độ học vấn cao đến thu nhập, có thể cần phải điều chỉnh thông điệp hoặc chiến lược để phù hợp hơn với nhóm này.

**4.4. Kết quả phân tích phân cụm**

Dựa trên biểu đồ Elbow đã cung cấp, điểm gấp khúc ("elbow") xuất hiện rõ ràng ở khoảng **k = 4**. Điều này cho thấy rằng số cụm tối ưu cho phân tích K-Means trong trường hợp này là **4**. Đây là số cụm mà tại đó độ giảm của tổng bình phương khoảng cách (inertia) bắt đầu chậm lại, cho thấy rằng việc tăng thêm số cụm không mang lại nhiều cải thiện đáng kể trong việc giảm tổng bình phương khoảng cách.

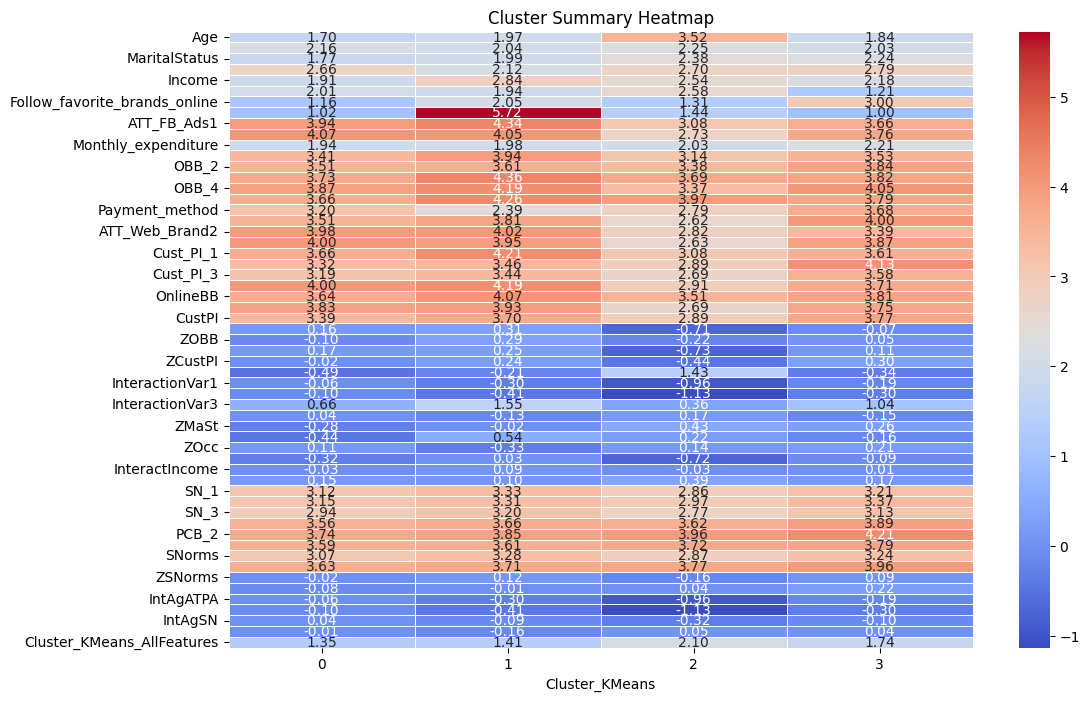


**Silhouette Score** cho kết quả phân cụm là **0.4189**, cho thấy mức độ tách biệt giữa các cụm là khá tốt.



**4.4.1. Đặc điểm của từng cụm**

Các nhóm khách hàng được phân chia dựa trên các đặc điểm như tuổi, trình độ học vấn, tình trạng hôn nhân, nghề nghiệp, thu nhập, thời gian trực tuyến, và mức độ tương tác với các thương hiệu trên internet. Dưới đây là tổng quan về từng cụm:



1. **Cụm 0:**
   * **Đặc điểm:** Nhóm này có tuổi trung bình khá trẻ (1.7) và trình độ học vấn tương đối cao (2.16). Đa số người tiêu dùng trong cụm này không dành nhiều thời gian trực tuyến (2.0 giờ/ngày) và có mức tương tác với các thương hiệu qua internet thấp (1.02). Thu nhập trung bình trong nhóm này là 1.91.
   * **Hành vi tiêu dùng:** Họ ít theo dõi thương hiệu yêu thích trực tuyến (1.15) và có mức chi tiêu hàng tháng vừa phải.
   * **Chiến lược marketing:** Các chiến dịch quảng cáo nên tập trung vào việc nâng cao sự tương tác trực tuyến và thúc đẩy sự tham gia của nhóm khách hàng này.
2. **Cụm 1:**
   * **Đặc điểm:** Nhóm này có tuổi trung bình 1.97, trình độ học vấn cao (2.04), và đa số họ là những người có thu nhập cao nhất trong số các cụm (2.84). Họ dành thời gian trực tuyến trung bình (1.94 giờ/ngày) nhưng có mức độ tương tác với thương hiệu rất cao (5.72).
   * **Hành vi tiêu dùng:** Họ thường xuyên theo dõi các thương hiệu yêu thích trực tuyến (2.05) và có thái độ tích cực đối với quảng cáo trên Facebook.
   * **Chiến lược marketing:** Nhóm này là mục tiêu lý tưởng cho các chiến dịch marketing trực tuyến và các chương trình khuyến mãi đặc biệt, tập trung vào các nền tảng truyền thông xã hội.
3. **Cụm 2:**
   * **Đặc điểm:** Nhóm này có tuổi trung bình lớn nhất (3.52) và thu nhập trung bình (2.53). Họ dành nhiều thời gian trực tuyến hơn (2.57 giờ/ngày) nhưng có mức độ tương tác với thương hiệu thấp (1.43).
   * **Hành vi tiêu dùng:** Họ ít theo dõi thương hiệu trực tuyến (1.30) và có xu hướng chi tiêu cẩn trọng.
   * **Chiến lược marketing:** Chiến lược marketing cho nhóm này nên tập trung vào các sản phẩm hoặc dịch vụ có giá trị cao và khuyến khích sự tham gia qua các kênh khác nhau ngoài trực tuyến.
4. **Cụm 3:**
   * **Đặc điểm:** Nhóm này có tuổi trung bình là 1.84 và có thu nhập ở mức trung bình (2.18). Họ dành ít thời gian trực tuyến (1.21 giờ/ngày) nhưng có sự tương tác với các thương hiệu tương đối thấp (1.00). Tuy nhiên, họ có xu hướng theo dõi các thương hiệu yêu thích trực tuyến rất thường xuyên (3.00).
   * **Hành vi tiêu dùng:** Họ có thái độ trung lập với các quảng cáo trên Facebook và có mức chi tiêu trung bình.
   * **Chiến lược marketing:** Nhóm này có thể hưởng lợi từ các chiến dịch marketing kết hợp trực tuyến và truyền thống, nhắm vào các yếu tố niềm tin và thương hiệu.

**4.4.2. Gợi ý chiến lược marketing**

Dựa trên phân tích cụm, các chiến lược marketing có thể được xây dựng như sau:

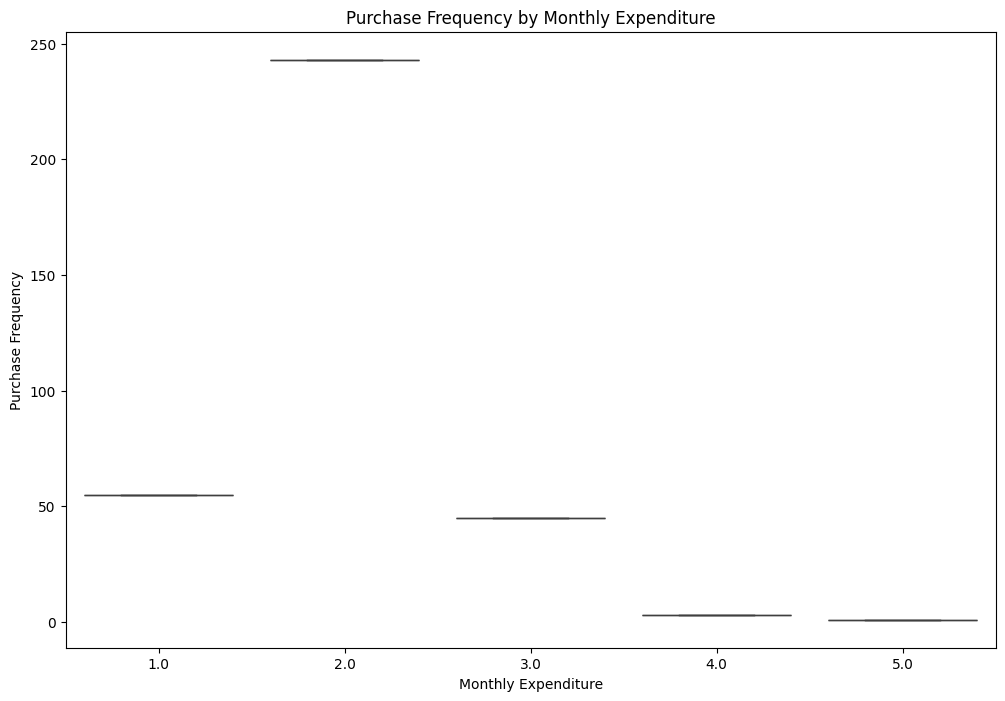
1. **Nhóm 0:** Cần cải thiện mức độ tương tác trực tuyến của nhóm này thông qua các chiến dịch quảng cáo hấp dẫn và dễ tiếp cận.
2. **Nhóm 1:** Tập trung phát triển các chiến dịch marketing kỹ thuật số nhắm đến nhóm này, đặc biệt trên các nền tảng xã hội nơi họ hoạt động tích cực.
3. **Nhóm 2:** Phát triển các sản phẩm hoặc dịch vụ có giá trị cao và các chương trình khuyến mãi đặc biệt, nhắm đến việc nâng cao trải nghiệm khách hàng trực tuyến.
4. **Nhóm 3:** Sử dụng các chiến lược marketing đa kênh, kết hợp trực tuyến và ngoại tuyến, để tăng cường sự nhận diện thương hiệu và lòng trung thành của khách hàng.

**4.5. Kết quả phân tích tần suất và xu hướng**

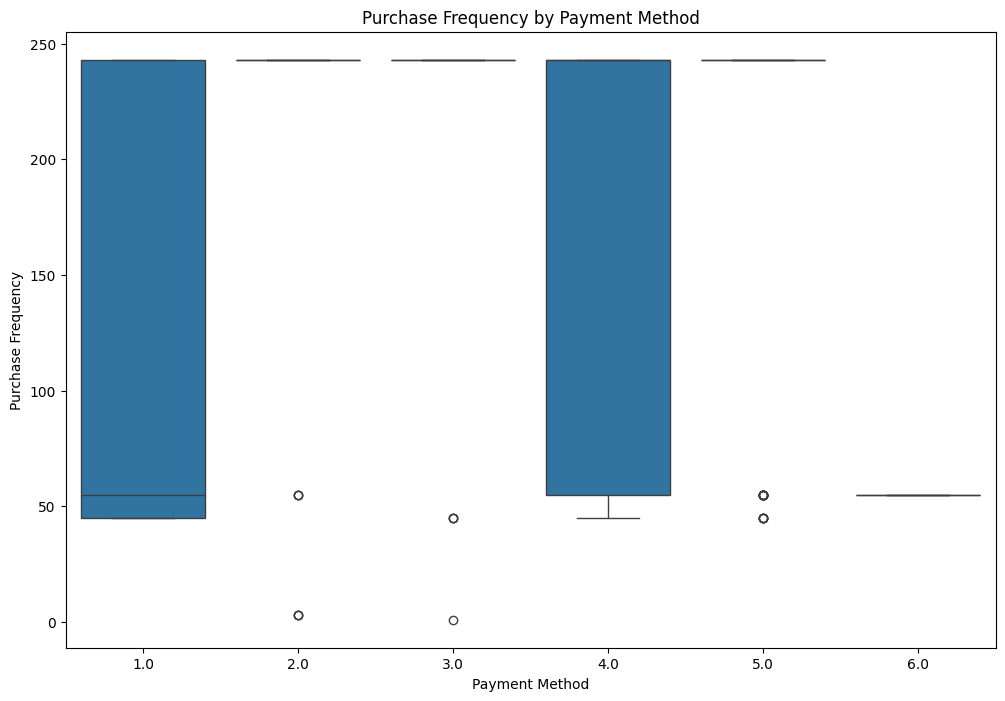
**4.5.1. Phân tích tần suất mua sắm**

Phân tích tần suất mua sắm của khách hàng là bước quan trọng trong việc xác định những người tiêu dùng có mức độ mua sắm cao. Bằng cách sử dụng các biến như Monthly\_expenditure (chi tiêu hàng tháng), Interaction\_with\_brands\_over\_internet (mức độ tương tác với thương hiệu trực tuyến), Payment\_method (phương thức thanh toán) và Follow\_favorite\_brands\_online (theo dõi các thương hiệu yêu thích trực tuyến), chúng tôi có thể đưa ra những nhận định quan trọng:

* **Biểu đồ 1:** **Tần suất mua sắm theo chi tiêu hàng tháng (Monthly Expenditure)**



* + Khách hàng có mức chi tiêu hàng tháng cao (mức 2.0) thường có tần suất mua sắm cao nhất, điều này cho thấy họ là những người tiêu dùng thường xuyên. Nhóm khách hàng này nên được chú trọng trong các chiến lược marketing tiếp cận.
* **Biểu đồ 2:** **Tần suất mua sắm theo phương thức thanh toán (Payment Method)**

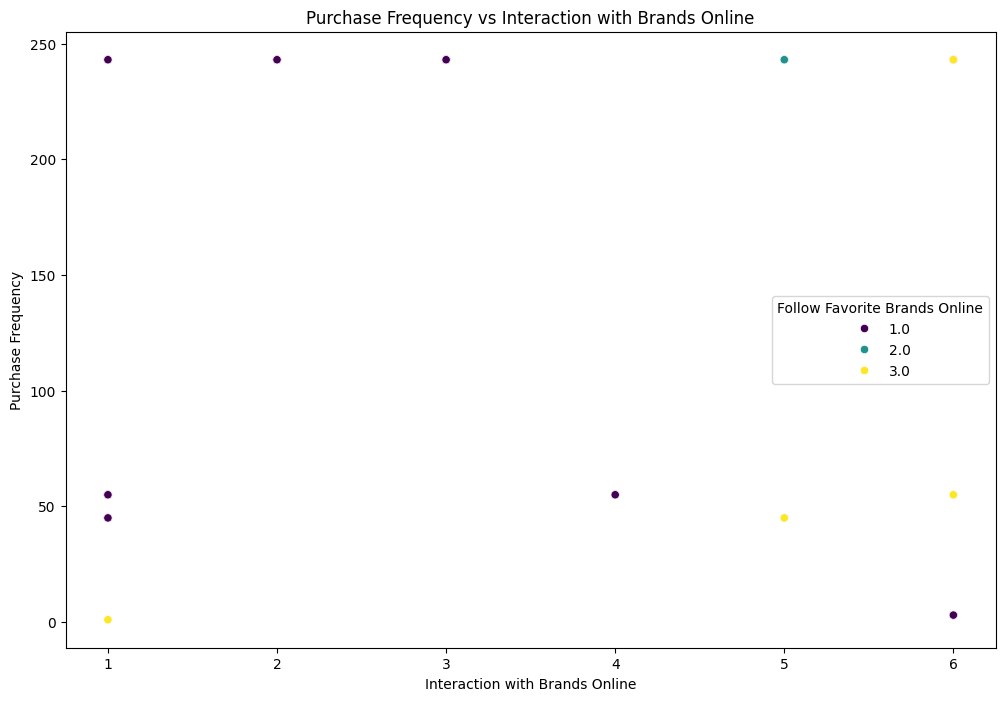


* + Phương thức thanh toán 1.0 và 4.0 có tần suất mua sắm cao nhất. Điều này cho thấy rằng các phương thức thanh toán này được người tiêu dùng ưa chuộng và sử dụng thường xuyên, có thể là một yếu tố quan trọng trong việc thu hút khách hàng.

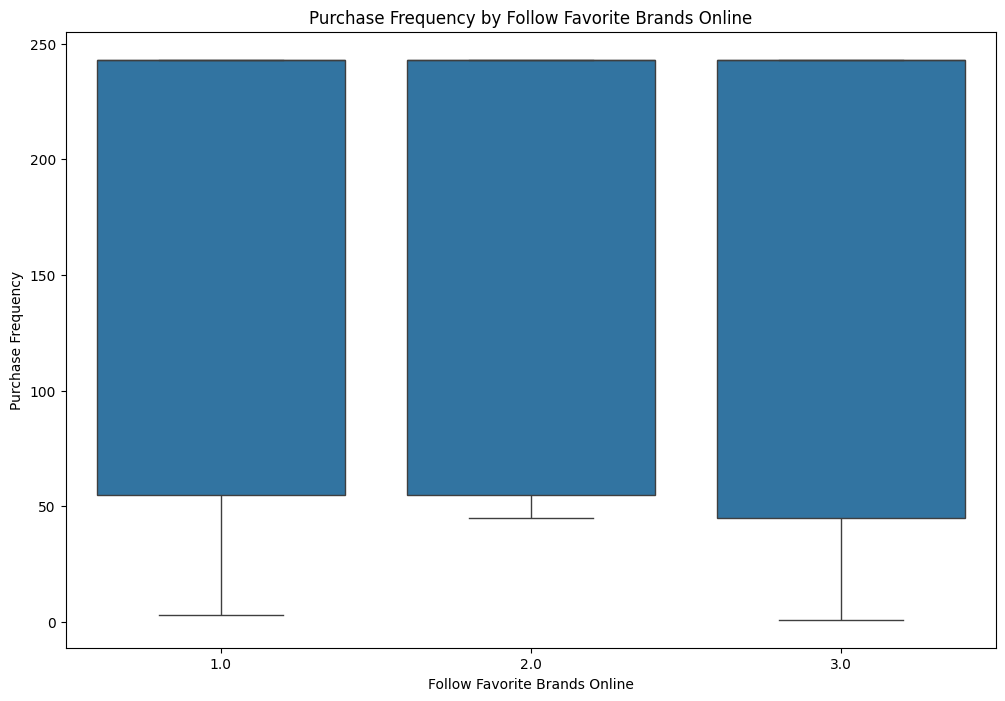
**4.5.2. Phân tích xu hướng mua sắm theo các yếu tố khác**

Phân tích xu hướng mua sắm không chỉ dừng lại ở việc xác định tần suất mà còn phải xem xét các yếu tố như tương tác với thương hiệu và hành vi theo dõi thương hiệu yêu thích.

* **Biểu đồ 3:** **Mối quan hệ giữa mức độ tương tác với thương hiệu trực tuyến và tần suất mua sắm (Interaction with Brands Online vs Purchase Frequency)**



* + Những khách hàng có mức độ tương tác cao với các thương hiệu trực tuyến (mức 6.0) thường có tần suất mua sắm cao hơn. Điều này cho thấy rằng việc tương tác tích cực với thương hiệu có thể thúc đẩy hành vi mua sắm.
* **Biểu đồ 4:** **Mối quan hệ giữa việc theo dõi thương hiệu yêu thích và tần suất mua sắm (Follow Favorite Brands Online vs Purchase Frequency)**

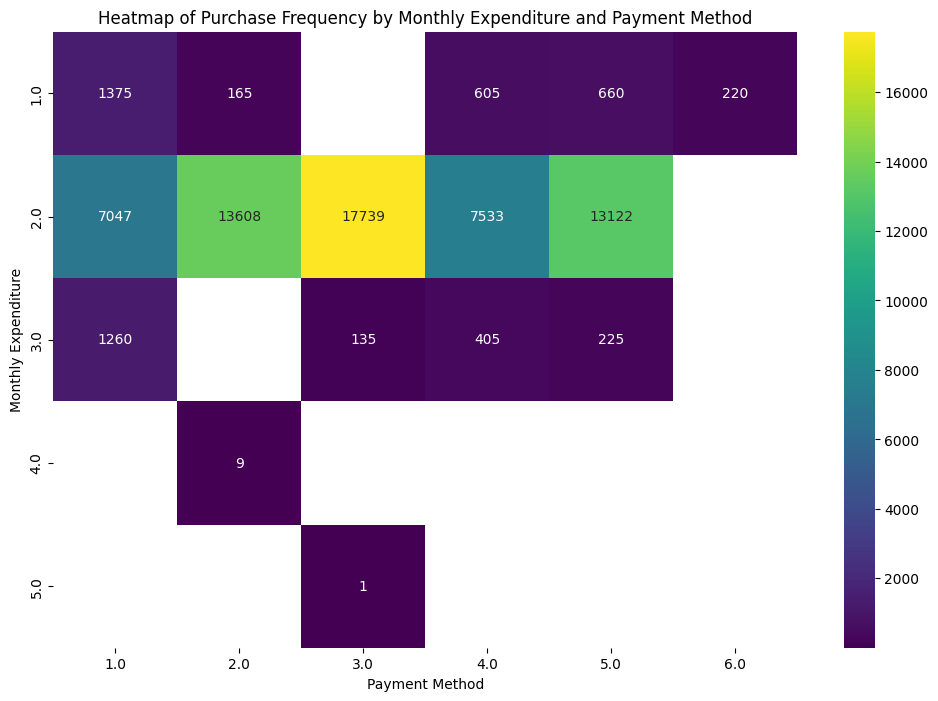


* + Khách hàng theo dõi các thương hiệu yêu thích thường có tần suất mua sắm cao hơn. Đây là yếu tố quan trọng mà doanh nghiệp có thể khai thác bằng cách tăng cường sự hiện diện của thương hiệu trên các nền tảng trực tuyến.

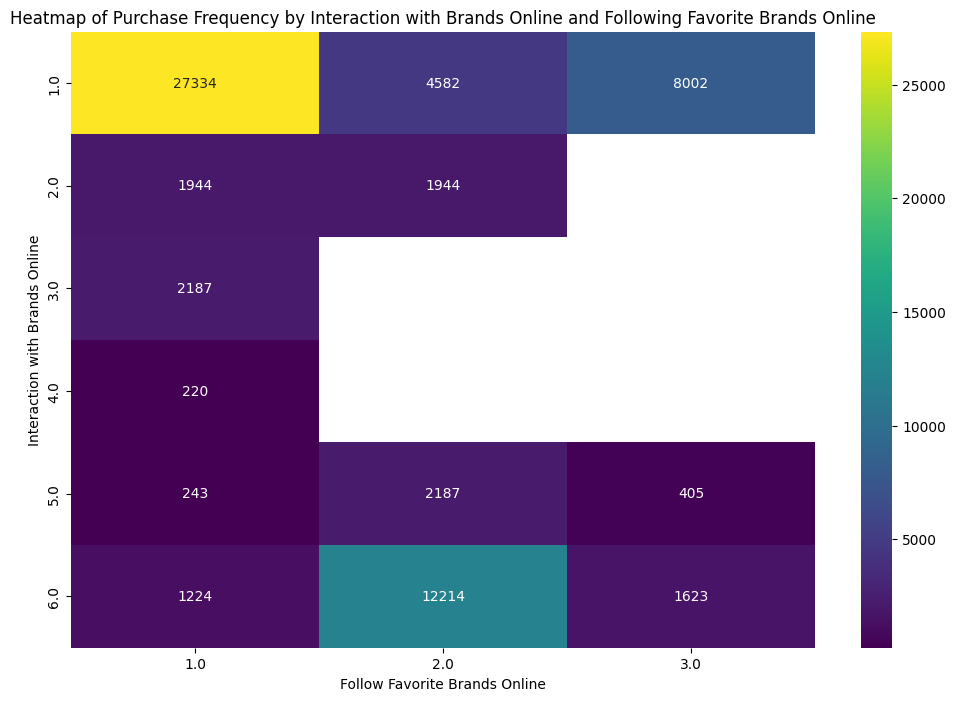
**4.5.3. Trực quan hóa bằng biểu đồ heatmap**

Biểu đồ heatmap giúp minh họa các xu hướng và tần suất mua sắm một cách rõ ràng và trực quan hơn:

* **Heatmap 1:** **Tần suất mua sắm theo chi tiêu hàng tháng và phương thức thanh toán**



* + Biểu đồ heatmap cho thấy rằng mức chi tiêu hàng tháng 2.0 kết hợp với phương thức thanh toán 4.0 có tần suất mua sắm cao nhất. Điều này gợi ý rằng việc tối ưu hóa các phương thức thanh toán có thể giúp tăng cường tần suất mua sắm ở mức chi tiêu này.
* **Heatmap 2:** **Tần suất mua sắm theo tương tác với thương hiệu và theo dõi thương hiệu yêu thích**



* + Biểu đồ heatmap này cho thấy rằng mức tương tác với thương hiệu trực tuyến 1.0 và mức độ theo dõi thương hiệu 1.0 có tần suất mua sắm cao nhất. Điều này có thể cho thấy rằng khách hàng có xu hướng mua sắm nhiều hơn khi họ có mức độ tương tác và theo dõi thương hiệu phù hợp.

**4.5.4. Gợi ý chiến lược marketing**

1. **Tập trung vào nhóm khách hàng có chi tiêu hàng tháng cao (mức 2.0):**
   * **Chiến lược:** Doanh nghiệp nên đầu tư vào các chiến dịch marketing và chương trình khách hàng thân thiết hướng tới nhóm khách hàng này, vì họ có tần suất mua sắm cao nhất. Các ưu đãi như giảm giá, tích điểm, hoặc các chương trình khuyến mãi đặc biệt có thể thúc đẩy họ mua sắm nhiều hơn.
2. **Tối ưu hóa các phương thức thanh toán phổ biến (1.0 và 4.0):**
   * **Chiến lược:** Cải thiện trải nghiệm người dùng với các phương thức thanh toán phổ biến này. Đảm bảo rằng các phương thức này dễ sử dụng, nhanh chóng và an toàn sẽ khuyến khích khách hàng thực hiện nhiều giao dịch hơn. Cũng có thể xem xét hợp tác với các công ty cung cấp dịch vụ thanh toán để tạo ra các chương trình khuyến mãi hấp dẫn, như hoàn tiền khi sử dụng các phương thức thanh toán này.
3. **Tăng cường sự hiện diện và tương tác trực tuyến của thương hiệu:**
   * **Chiến lược:** Khách hàng có mức độ tương tác cao với thương hiệu trực tuyến có tần suất mua sắm cao hơn. Doanh nghiệp nên đẩy mạnh các hoạt động truyền thông xã hội, email marketing, và sử dụng các chiến dịch quảng cáo có tính cá nhân hóa cao để giữ chân và tương tác với khách hàng. Các nội dung mang tính giáo dục, chia sẻ giá trị hoặc các chương trình livestream giới thiệu sản phẩm có thể là một cách hiệu quả.
4. **Khuyến khích khách hàng theo dõi thương hiệu yêu thích trực tuyến:**
   * **Chiến lược:** Doanh nghiệp nên khuyến khích khách hàng theo dõi các kênh trực tuyến của họ bằng cách cung cấp các ưu đãi đặc biệt cho những người theo dõi. Điều này có thể bao gồm các thông báo độc quyền về sản phẩm mới, mã giảm giá, hoặc các cuộc thi dành riêng cho người theo dõi.
5. **Phát triển các chương trình khuyến mãi theo mùa:**
   * **Chiến lược:** Mặc dù dữ liệu cụ thể về mùa không có, nhưng doanh nghiệp vẫn có thể dựa trên kinh nghiệm để triển khai các chương trình khuyến mãi theo mùa. Ví dụ: các chiến dịch mùa hè, lễ hội cuối năm, hoặc Black Friday. Đối với các sản phẩm liên quan đến thời gian cụ thể trong năm, như đồ trang trí hoặc quần áo theo mùa, các chiến dịch này sẽ rất hiệu quả.
6. **Nâng cao trải nghiệm mua sắm đa kênh:**
   * **Chiến lược:** Với sự phổ biến của mua sắm trực tuyến và tương tác với thương hiệu qua internet, doanh nghiệp cần xây dựng một trải nghiệm mua sắm liền mạch giữa các kênh trực tuyến và ngoại tuyến. Việc đồng bộ hóa các chương trình khuyến mãi, tích điểm và chăm sóc khách hàng giữa các kênh sẽ tạo ra sự thuận tiện và thu hút khách hàng quay lại mua sắm.

**5. Ứng dụng và Khuyến nghị**

**5.1. Ứng dụng kết quả nghiên cứu vào chiến lược marketing**

(1) **Định hình chiến lược quảng cáo dựa trên tâm lý người tiêu dùng:**

* **Nhắm đến đối tượng theo độ tuổi và thu nhập:** Kết quả phân tích hồi quy cho thấy tuổi tác và thu nhập có ảnh hưởng đáng kể đến hành vi mua sắm. Các chiến dịch quảng cáo nên tập trung vào nhóm khách hàng lớn tuổi với thu nhập cao, vì nhóm này có xu hướng tương tác tích cực hơn với thương hiệu và có khả năng chi tiêu cao hơn.
* **Tối ưu hóa trải nghiệm trực tuyến:** Khách hàng dành nhiều thời gian trực tuyến và có tương tác cao với thương hiệu trực tuyến thể hiện xu hướng mua sắm tích cực hơn. Doanh nghiệp cần tăng cường các hoạt động marketing trực tuyến như quảng cáo trên mạng xã hội, email marketing, và nội dung video tương tác để thu hút nhóm này.
  1. **Phát triển sản phẩm và dịch vụ phù hợp với nhu cầu của từng nhóm khách hàng:**
* **Nhóm khách hàng phân cụm:** Phân tích cụm đã xác định bốn nhóm khách hàng khác nhau với các đặc điểm tâm lý và hành vi mua sắm đặc trưng. Doanh nghiệp nên phát triển các sản phẩm và dịch vụ phù hợp với từng nhóm. Ví dụ, nhóm khách hàng trẻ tuổi có thu nhập thấp có thể được nhắm đến bằng các sản phẩm giá rẻ, trong khi nhóm khách hàng lớn tuổi, có thu nhập cao nên được tiếp cận với các sản phẩm cao cấp và dịch vụ khách hàng ưu tiên.
  1. **Tối ưu hóa kênh phân phối và chiến lược giá dựa trên hành vi tiêu dùng:**
* **Chiến lược giá linh hoạt:** Dựa trên kết quả hồi quy và phân tích tần suất, doanh nghiệp có thể xem xét triển khai các chiến lược giá linh hoạt để đáp ứng nhu cầu của các nhóm khách hàng khác nhau. Những khách hàng có mức chi tiêu hàng tháng cao thường có tần suất mua sắm cao hơn, do đó, có thể áp dụng các chính sách giá ưu đãi, giảm giá theo số lượng mua, hoặc các chương trình khách hàng thân thiết để kích thích chi tiêu.
* **Phát triển kênh phân phối phù hợp:** Các nhóm khách hàng có thói quen thanh toán khác nhau có thể ưa chuộng các kênh phân phối khác nhau. Doanh nghiệp nên đảm bảo rằng các phương thức thanh toán phổ biến được hỗ trợ đầy đủ trên tất cả các kênh, từ trực tuyến đến cửa hàng vật lý, để tối ưu hóa trải nghiệm mua sắm.

**5.2. Khuyến nghị cho doanh nghiệp**

**(1) Tăng cường sử dụng phân tích dữ liệu trong việc ra quyết định marketing:**

* **Phân tích dự đoán:** Doanh nghiệp nên đầu tư vào các công cụ và kỹ thuật phân tích dữ liệu nâng cao như học máy (machine learning) để dự đoán xu hướng tiêu dùng và hành vi mua sắm. Điều này không chỉ giúp tối ưu hóa các chiến dịch marketing mà còn giúp doanh nghiệp dự đoán trước các nhu cầu và mong muốn của khách hàng.
  1. **Liên tục cập nhật và phân tích dữ liệu để theo kịp sự thay đổi trong hành vi tiêu dùng:**
* **Phân tích thời gian thực:** Hành vi người tiêu dùng có thể thay đổi nhanh chóng do các yếu tố bên ngoài như kinh tế, xã hội, và công nghệ. Do đó, việc cập nhật và phân tích dữ liệu trong thời gian thực sẽ giúp doanh nghiệp duy trì sự cạnh tranh và phản ứng kịp thời với các thay đổi trên thị trường.
  1. **Đào tạo nhân viên về kỹ năng phân tích dữ liệu và ứng dụng trong marketing:**
* **Phát triển kỹ năng nội bộ:** Để tận dụng tối đa các kết quả phân tích dữ liệu, doanh nghiệp nên đầu tư vào đào tạo nhân viên về các kỹ năng phân tích dữ liệu, bao gồm sử dụng các công cụ như Python, R, và các nền tảng phân tích dữ liệu khác. Nhân viên được trang bị các kỹ năng này sẽ có khả năng ra quyết định dựa trên dữ liệu một cách hiệu quả, giúp tối ưu hóa chiến lược marketing và kinh doanh.

**6. Kết luận**

**(1) Tổng kết các kết quả nghiên cứu:**

* **Phân tích mô tả và hồi quy:** Chúng tôi đã thực hiện các phân tích mô tả và hồi quy để hiểu rõ hơn về mối quan hệ giữa các yếu tố nhân khẩu học, hành vi tiêu dùng và quyết định mua sắm của khách hàng. Kết quả cho thấy các yếu tố như tuổi tác, thu nhập, thời gian trực tuyến, và mức độ tương tác với thương hiệu trực tuyến có ảnh hưởng đáng kể đến hành vi mua sắm.
* **Phân cụm khách hàng:** Dựa trên phân tích cụm, bốn nhóm khách hàng đã được xác định với những đặc điểm tâm lý và hành vi mua sắm đặc trưng. Mỗi nhóm này có các nhu cầu và mong muốn khác nhau, đòi hỏi doanh nghiệp phải có các chiến lược tiếp cận và sản phẩm phù hợp để tối ưu hóa hiệu quả kinh doanh.
* **Phân tích tần suất và xu hướng:** Phân tích tần suất mua sắm cho thấy những khách hàng có mức chi tiêu cao và tương tác trực tuyến tích cực thường có tần suất mua sắm cao hơn. Điều này nhấn mạnh tầm quan trọng của việc xây dựng chiến lược tiếp cận dựa trên các hành vi tiêu dùng cụ thể.
  1. **Nhấn mạnh tầm quan trọng của việc hiểu rõ tâm lý người tiêu dùng trong việc phát triển chiến lược marketing hiệu quả:**
* Hiểu rõ tâm lý và hành vi của người tiêu dùng không chỉ giúp doanh nghiệp xây dựng các chiến lược marketing hiệu quả mà còn là chìa khóa để tạo ra sự khác biệt trên thị trường cạnh tranh. Thông qua việc phân tích dữ liệu người tiêu dùng, doanh nghiệp có thể đưa ra các quyết định chiến lược dựa trên thông tin chính xác, từ đó tối ưu hóa nguồn lực và gia tăng hiệu quả kinh doanh.
* Việc phân loại khách hàng thành các nhóm dựa trên đặc điểm tâm lý và hành vi giúp doanh nghiệp cá nhân hóa chiến lược tiếp cận, tạo ra trải nghiệm mua sắm phù hợp hơn với từng khách hàng. Điều này không chỉ tăng cường sự hài lòng và lòng trung thành của khách hàng mà còn tạo điều kiện thuận lợi cho việc phát triển lâu dài của doanh nghiệp.

**Kết luận cuối cùng:**

* Nghiên cứu này đã chứng minh rằng phân tích dữ liệu là công cụ mạnh mẽ để hiểu và đáp ứng nhu cầu của người tiêu dùng. Bằng cách kết hợp các phương pháp phân tích khác nhau, doanh nghiệp có thể xác định và khai thác các cơ hội tiềm năng, từ đó xây dựng các chiến lược marketing không chỉ hiệu quả mà còn bền vững trong dài hạn.

**Tài liệu tham khảo**

 Kotler, P., & Keller, K. L. (2016). Marketing Management (15th ed.). Pearson Education.

 Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). Multivariate Data Analysis (8th ed.). Cengage Learning.

 Field, A. (2018). Discovering Statistics Using IBM SPSS Statistics (5th ed.). SAGE Publications.

 Malhotra, N. K. (2019). Marketing Research: An Applied Orientation (7th ed.). Pearson Education.

 McDaniel, C., & Gates, R. (2018). Marketing Research (10th ed.). Wiley.

 Sarstedt, M., & Mooi, E. (2019). A Concise Guide to Market Research (3rd ed.). Springer.

 Mendeley Data. (2023). Dataset on Consumer Behavior and Attitudes [Data file]. Retrieved from https://data.mendeley.com/datasets/mpd6m6jp4f/1

 Seaborn, M., & Matplotlib, L. (2020). Python Data Visualization Libraries [Software]. Python Software Foundation.

 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

 Pandas Development Team. (2021). Pandas: Python Data Analysis Library [Software]. Python Software Foundation.

 Numpy Developers. (2020). Numpy: A General-Purpose Array-Processing Package [Software]. Python Software Foundation.