**TRƯỜNG ĐẠI HỌC SÀI GÒN**



**KHOA CÔNG NGHỆ THÔNG TIN**



**HỌC PHẦN: PHÂN TÍCH DỮ LIỆU**

**TÊN ĐỀ TÀI: PHÂN TÍCH DỮ LIỆU ĐỂ HIỂU NHU CẦU VÀ HÀNH VI CỦA KHÁCH HÀNG.**

**Họ tên thành viên trong nhóm:**

Nguyễn Hồng Nhi - 3120410375

Lê Thành Đạt - 3120410114

**GIÁO VIÊN HƯỚNG DẪN: thầy Phan Thành Huấn**

**TP. HCM tháng 12/2023**

MỤC LỤC

[**CHƯƠNG 1: KHÁI QUÁT ĐỒ ÁN 1**](#_Toc152596436)

[**1.1. Lý do chọn đề tài 1**](#_Toc152596437)

[**1.2. Mô tả đồ án 2**](#_Toc152596438)

[**1.3. Mô tả tập dữ liệu: 3**](#_Toc152596439)

[**CHƯƠNG 2: LỌC DỮ LIỆU VÀ TRỰC QUAN HÓA 5**](#_Toc152596440)

[**2.1. Phân tích tổng quan 5**](#_Toc152596441)

[**2.2. Phân tích số liệu tại từng chi nhánh 10**](#_Toc152596442)

[**CHƯƠNG 3. PHÂN LỚP DỮ LIỆU 26**](#_Toc152596443)

[**3.1. Các phương pháp pháp dự đoán vả quy trình cụ thể 26**](#_Toc152596444)

[**3.1.1. Phân lớp dữ liệu 26**](#_Toc152596445)

[**3.1.2. Phân cụm dữ liệu 30**](#_Toc152596446)

[**3.2. Thực nghiệm bài toán: Dự báo Doanh Số Bán Hàng 32**](#_Toc152596447)

[**3.2.1. Mô tả bài toán: 32**](#_Toc152596448)

[**3.2.2. Chạy mô hình và kết quả 32**](#_Toc152596449)

[**CHƯƠNG 4: PHÂN CỤM DỮ LIỆU 40**](#_Toc152596450)

[**KẾT QUẢ CỦA QUÁ TRÌNH PHÂN LỚP VÀ PHÂN CỤM 41**](#_Toc152596451)

[**KẾT LUẬN 42**](#_Toc152596452)

[**LỜI KẾT 44**](#_Toc152596453)

[**TÀI LIỆU THAM KHẢO 45**](#_Toc152596454)

# CHƯƠNG 1: KHÁI QUÁT ĐỒ ÁN

## 1.1. Lý do chọn đề tài

Trong bối cảnh cạnh tranh ngày càng gay gắt, việc hiểu rõ nhu cầu và hành vi của khách hàng là yếu tố sống còn đối với các doanh nghiệp. Phân tích dữ liệu là một công cụ hiệu quả giúp doanh nghiệp hiểu rõ hơn về khách hàng, từ đó có thể đưa ra các quyết định kinh doanh sáng suốt hơn.

Cụ thể, phân tích dữ liệu có thể giúp doanh nghiệp:

Xác định nhu cầu của khách hàng: Phân tích dữ liệu có thể giúp doanh nghiệp hiểu rõ nhu cầu hiện tại và nhu cầu tiềm năng của khách hàng. Từ đó, doanh nghiệp có thể đưa ra các sản phẩm, dịch vụ phù hợp với nhu cầu của khách hàng.

Tìm hiểu hành vi của khách hàng: Phân tích dữ liệu có thể giúp doanh nghiệp hiểu rõ hành vi mua sắm, sử dụng sản phẩm/dịch vụ của khách hàng. Từ đó, doanh nghiệp có thể tối ưu hóa trải nghiệm khách hàng và tăng cường lòng trung thành của khách hàng.

Dự đoán nhu cầu của khách hàng: Phân tích dữ liệu có thể giúp doanh nghiệp dự đoán nhu cầu của khách hàng trong tương lai. Từ đó, doanh nghiệp có thể có kế hoạch sản xuất, marketing phù hợp.

Với những lợi ích to lớn mà phân tích dữ liệu mang lại, tôi tin rằng đây là một đề tài nghiên cứu có giá trị và phù hợp với thực tế hiện nay. Tôi mong muốn nghiên cứu đề tài này sẽ giúp các doanh nghiệp hiểu rõ hơn về khách hàng và từ đó có thể đưa ra các quyết định kinh doanh sáng suốt hơn, góp phần nâng cao hiệu quả kinh doanh.

Ngoài ra, tôi cũng có niềm đam mê với lĩnh vực phân tích dữ liệu và mong muốn được học hỏi thêm về lĩnh vực này. Nghiên cứu đề tài này sẽ giúp tôi nâng cao kiến thức và kỹ năng phân tích dữ liệu, từ đó có thể ứng dụng trong công việc và cuộc sống.

## 1.2. Mô tả đồ án

Trong quá trình thực hiện đồ án, chúng em sẽ bắt đầu bằng việc thu thập và xử lý dữ liệu từ các chi nhánh khác nhau của doanh nghiệp. Việc này bao gồm việc làm sạch dữ liệu, kiểm tra tính nhất quán, và xác định các biến quan trọng nhằm giữ lại những thông tin có ý nghĩa nhất cho mục tiêu phân tích. Chúng em sẽ sử dụng các phương pháp thống kê và các công cụ phân tích dữ liệu để trích xuất những thông điệp quan trọng từ số liệu thu thập được.

Sau khi đã có dữ liệu được làm sạch và chuẩn bị, chúng em sẽ tập trung vào việc thực hiện các phân tích chi tiết cho từng dòng sản phẩm tại mỗi chi nhánh. Việc này có thể bao gồm so sánh hiệu suất bán hàng, tính toán tỷ lệ tăng trưởng, và đánh giá sự biến động của doanh số bán hàng theo thời gian.

Để tạo ra một hiểu biết sâu sắc, chúng em sẽ trình bày các kết quả bằng cách sử dụng biểu đồ và đồ thị trực quan. Điều này sẽ giúp cho việc truyền đạt thông tin một cách dễ dàng và rõ ràng, giúp độc giả hiểu rõ hơn về các xu hướng và đặc điểm quan trọng trong dữ liệu.

Cuối cùng, chúng em sẽ từ kết quả phân tích đề xuất những chiến lược kinh doanh cụ thể cho doanh nghiệp. Các gợi ý sẽ được xây dựng dựa trên những thông tin chính xác và chi tiết từ phân tích dữ liệu, hỗ trợ doanh nghiệp đưa ra các quyết định thông minh và có tầm chiến lược. Chúng em hy vọng rằng đồ án của mình sẽ mang lại giá trị và hỗ trợ cho quá trình ra quyết định kinh doanh của doanh nghiệp.

## 1.3. Mô tả tập dữ liệu

Nhóm em sẽ thực hiện quá trình phân tích dữ liệu trên tập dữ liệu là market.csv gồm 17 cột và 1000 dòng(được gửi kèm với file này).

Tệp market.csv gồm các thuộc tính: Invoice ID, Branch, City, Customer type, Gender, Product line, Unit price, Quantity, Tax 5%, Total, Date, Time, Payment, cogs, gross margin percentage, gross income, Rating

Trong những thuộc tính kể trên, nhóm chúng em sẽ chọn ra những thuộc tính cần thiết và tiến hành tiền xử lý dữ liệu để phục vụ cho việc phân tích của mình như sau:

+ Thuộc tính Branch: chọn thuộc tính này để có thể phân loại ra các chi nhánh của doanh nghiệp từ đó biết được các dòng sản phẩm ở doanh nghiệp được bán ra sao ở các chi nhánh.

+ Thuộc tính Product line: chọn thuộc tính này để biết được các dòng sản phẩm đang được bán tại các chi nhánh của doanh nghiệp.

+ Thuộc tính Unit price, Quantity: chọn 2 thuộc tính này để biết được sản phẩm nào được bán ra và bán ra với số lượng bao nhiêu từ đó có thể biết được sản phẩm nào được mua nhiều hoặc sản phẩm nào có tổng doanh thu cao nhất.

+ Thuộc tính Rating: chọn thuộc tính này để tìm hiểu về thái độ của khách hàng đối với sản phẩm đã mua hàng, từ đó có thể biết được mức độ hài lòng của khách hàng đối với dòng sản phẩm (có thể tích cực hoặc tiêu cực) và giúp doanh nghiệp xem xét về việc điều chỉnh chất lượng của các dòng sản phẩm.

**Các thư viện Python được sử dụng để phục vụ việc phân tích:**

- Pandas: thư viện này được sử dụng để phục vụ việc tiền xử lý các dữ liệu trong file qua việc chọn các thuộc tính cần thiết, lọc các dữ liệu theo điều kiện, tính toán các số liệu ở các cột, thêm các cột tính toán,…

- Matplotlib: thư viện này được sử dụng để vẽ nên các biểu đồ từ các số liệu đã được tính toán giúp cho việc thể hiện dữ liệu trở nên trực quan và dễ hiểu hơn.

- Seaborn: dùng để vẽ những biểu đồ có tính phức tạp

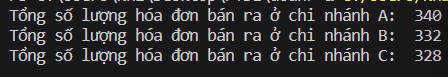
**CHƯƠNG 2: LỌC DỮ LIỆU VÀ TRỰC QUAN HÓA**

Từ tập dữ liệu ban đầu, ta sẽ chọn lọc ra những dữ liệu cần thiết để phục vụ việc phân tích như sau:

## 2.1. Phân tích tổng quan

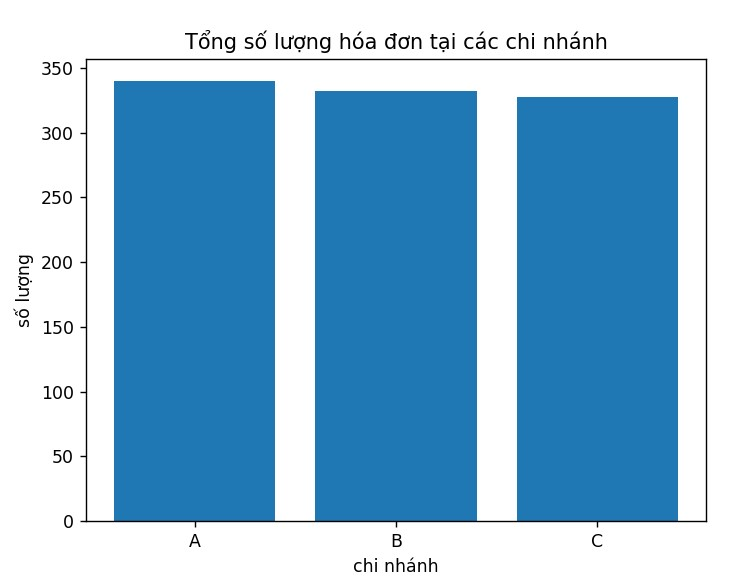
- Tìm hiểu về tổng số lượng hóa đơn tại các chi nhánh là bao nhiêu?

Lọc dữ liệu:



|  |  |  |  |
| --- | --- | --- | --- |
| Chi nhánh | A | B | C |
| Tổng số lượng hóa đơn | 340 | 332 | 328 |

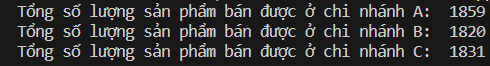
Mô hình hóa:



Nhận xét:từ biểu đồ trên, ta nhận thấy rằng số lượng hóa đơn nhiều nhất là ở chi nhánh A, tiếp theo là chi nhánh B, C. Tuy nhiên, số lượng chênh lệch giữa các chi nhánh không nhiều.

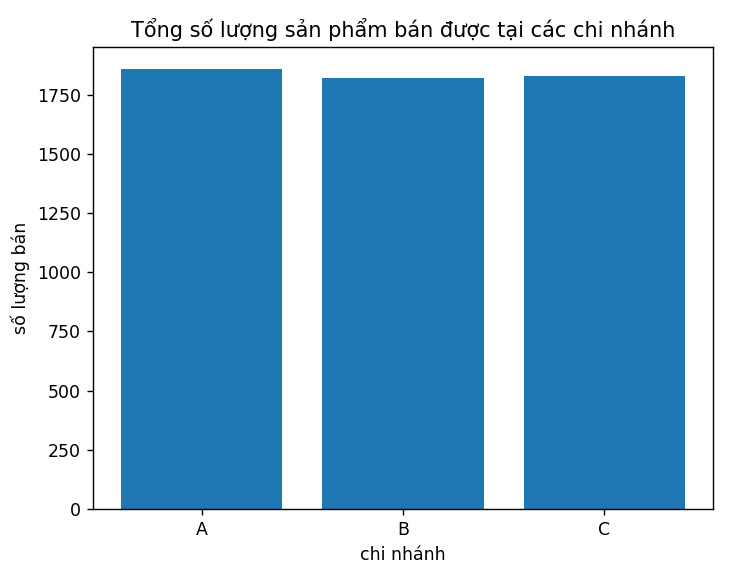
- Tìm hiểu về tổng số lượng sản phẩm được bán ra tại các chi nhánh là bao nhiêu?

Lọc dữ liệu:



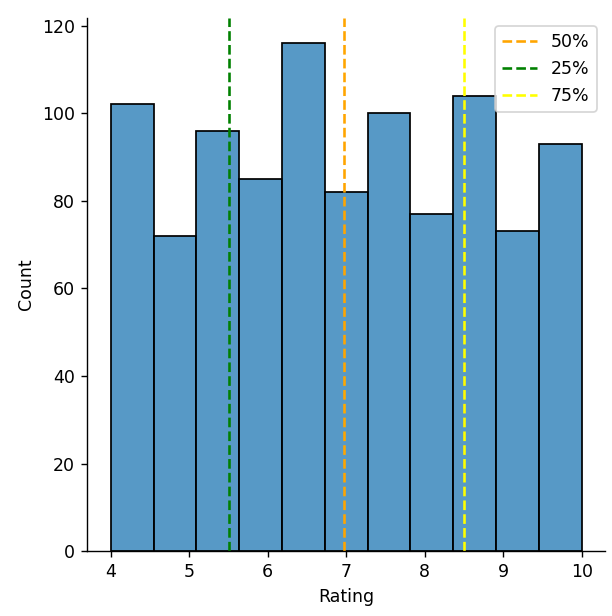
|  |  |  |  |
| --- | --- | --- | --- |
| Chi nhánh | A | B | C |
| Tổng số lượng bán | 1859 | 1820 | 1831 |

Mô hình hóa:



Nhận xét: Số lượng bán được ở các chi nhánh chênh lệch không nhiều, cao nhất là ở chi nhánh A.

- Tìm hiểu về sự phân phối điểm Rating trong tổng số các hóa đơn?



*Biểu đồ phân bố điểm rating*

Nhận xét:

- Ở mức điểm rating từ 4.0 - 5.5: số lượng hóa đơn bán ra dao động từ 65 đến hơn 100 hóa đơn

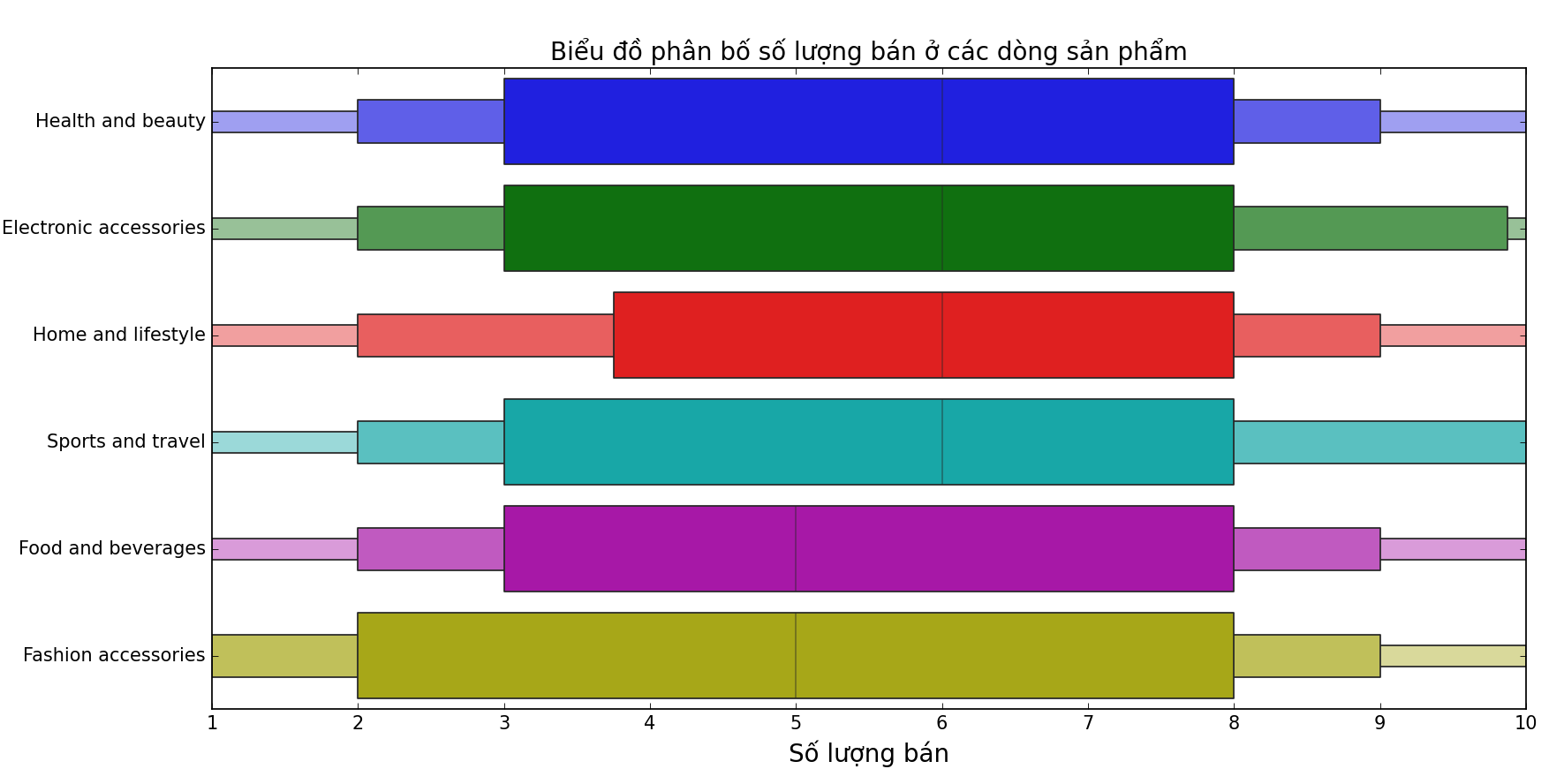
- Ở mức điểm rating từ 5.5 - 7.0: số lượng hóa đơn bán ra dao động từ hơn 80 đến hơn 110 hóa đơn, đây cũng là mức rating có số lượng hóa đơn nhiều nhất

- Ở mức điểm rating từ 7.0 - 8.5: số lượng hóa đơn bán ra dao động từ 78 đến hơn 98 hóa đơn

- Ở mức điểm rating từ 8.5 - 10.0: số lượng hóa đơn bán ra dao động từ 75đến hơn 100 hóa đơn

**Kết luận**: Nhìn chung tỉ lệ các hóa đơn với mức rating là 5.5 - 7.0 và 7.0 - 8.0 vẫn chiếm số lượng nhiều. Vì thế về tổng thể có thể suy ra rằng cảm nhận của khách hàng đối với các dòng sản phẩm phần lớn ở mức trung bình khá.

- Tìm hiểu về sự phân bố số lượng bán ra của từng dòng sản phẩm



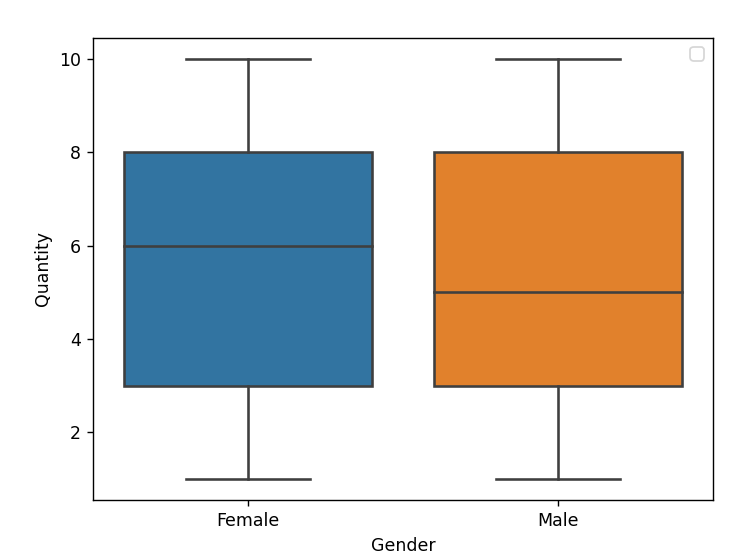
Nhận xét:

- Các dòng sản phẩm có số lượng bán dao động chủ yếu từ 3 - 8 sản phẩm.

- Dòng sản phẩm Fashion accessories có số lượng bán từ 2 - 3 nhiều hơn các dòng sản phẩm khác.

- Dòng sản phẩm Electronic accessories và Sport and travel có số lượng bán từ 9 - 10 sản phẩm cao hơn hẳn các dòng sản phẩm còn lại.

- Tìm hiểu về mức mua hàng ở nam và nữ



*Biểu đồ về biểu diễn số lượng mua hàng ở giới nam và nữ*

Nhận xét:

- Số lượng bán ra nhiều nhất là 10.

- Số lượng bán ra nhỏ nhất là 1.

- Ở phần tư thứ nhất: Nam và nữ mua với số lượng từ 1 - 3 như nhau.

- Ở phần tư thứ hai: Nữ mua với số lượng từ 3 - 6, nam mua từ 3 - 5.

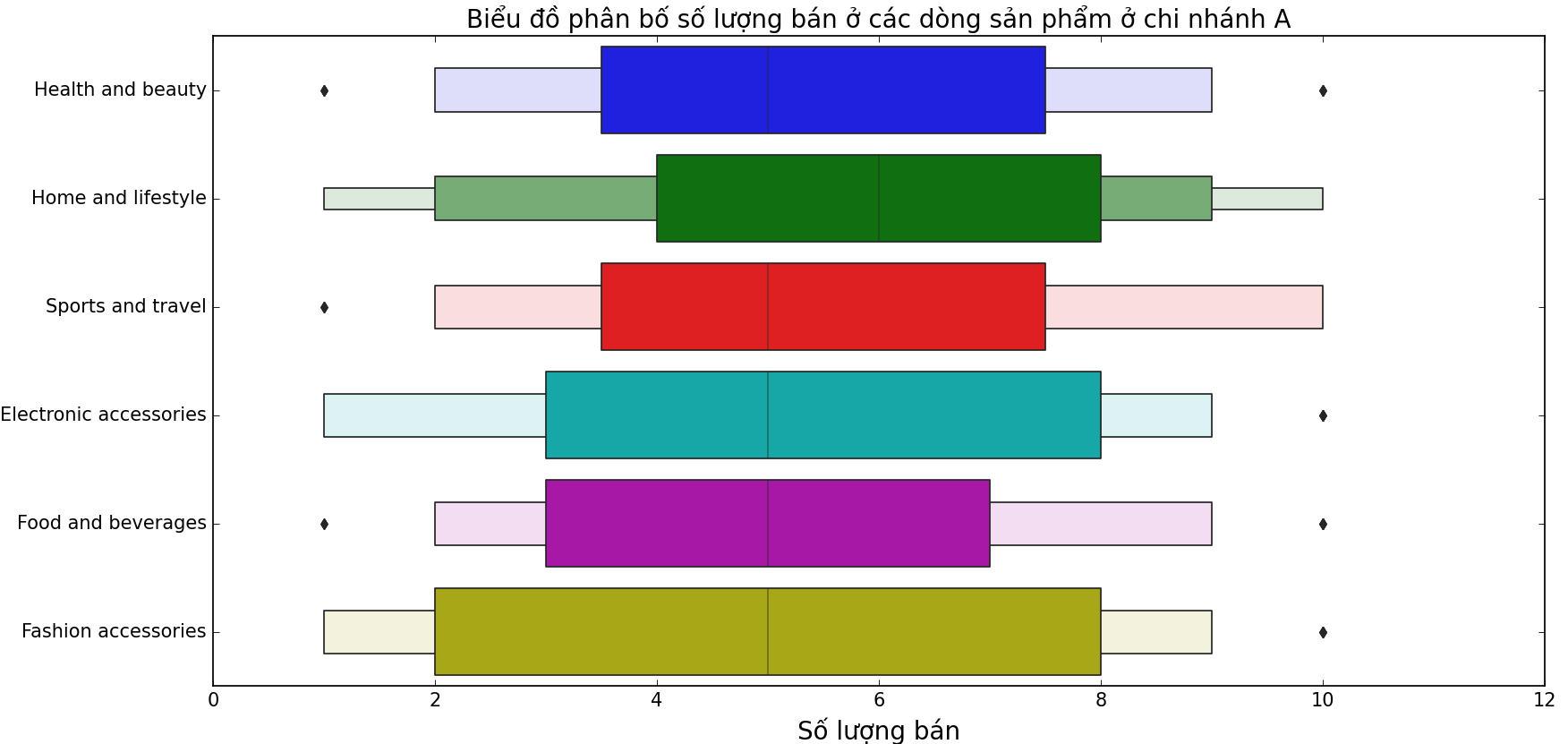
- Ở phần từ thứ ba: Nữ mua với số lượng từ 6 - 8, nam mua từ 5 - 8.

- Ở phần tư cuối cùng: Nữ và nam mua với số lượng từ 8 - 10 như nhau.

=> Nhìn chung, giữa nam và nữ mua hàng với số lượng như nhau và chênh lệch không đáng kể.

## 2.2. Phân tích số liệu tại từng chi nhánh

**- Tìm hiểu về sự phân bố số lượng bán ra của từng dòng sản phẩm ở chi nhánh A?**



Nhận xét:

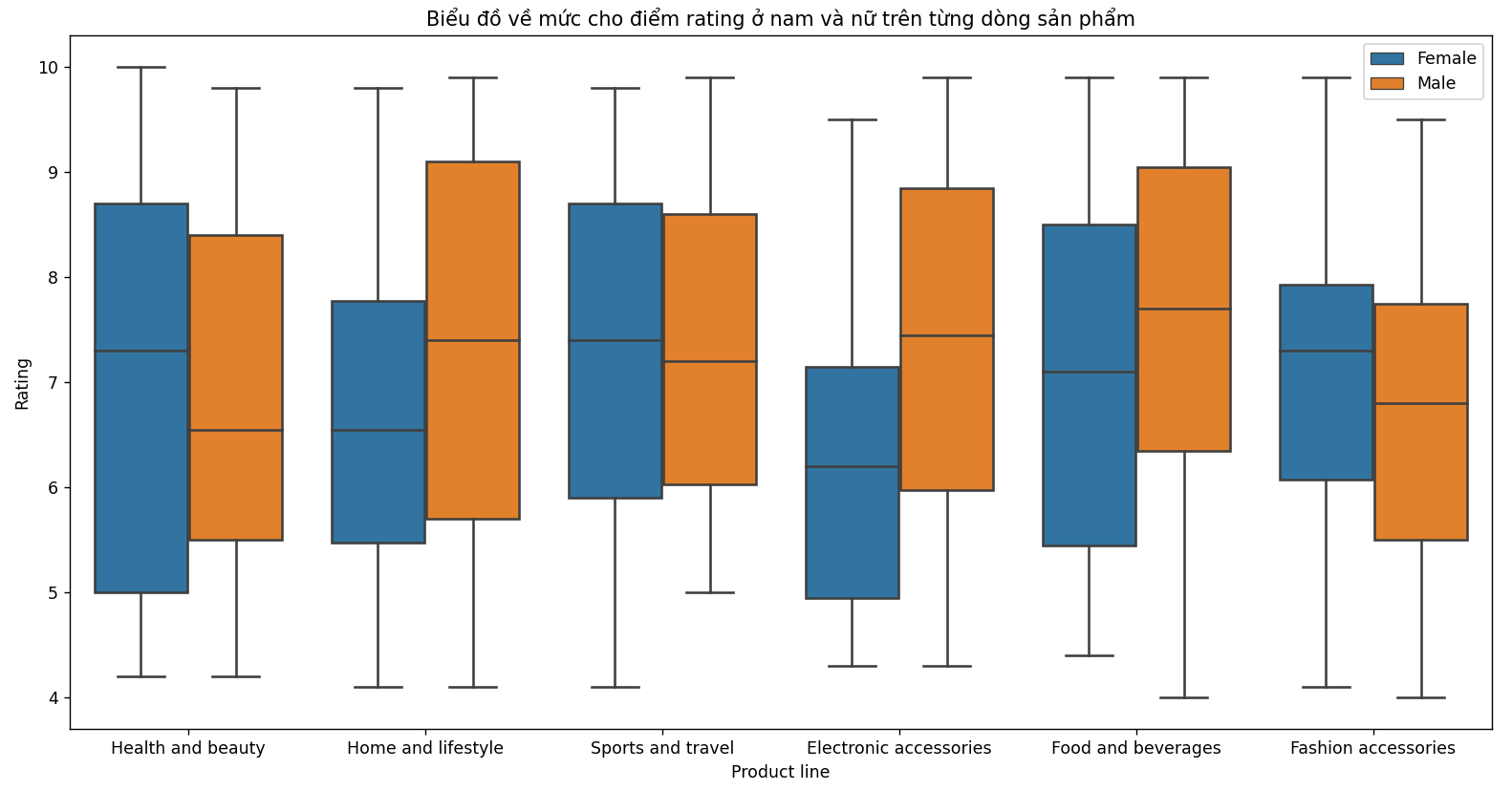
- Về tổng thể, các dòng sản phẩm ở chi nhánh bán với số lượng được mua nhiều nhất là ở mức 4 - 6.

- Dòng sản phẩm Sports and travel có số lượng được mua ở mức 8 - 10 cao nhất.

- Dòng sản phẩm Fashion accessories có số lượng được mua phân bố khá đều.

- Dòng sản phẩm Food and beverages có số lượng đươc mua chủ yếu từ 3 - 7, chiếm tỉ lệ thấp hơn so với các dòng sản phẩm còn lại.

=> Đối với chi nhánh A, cần xem lại sản phẩm thuộc dòng sản phẩm Food and beverages để điều chỉnh chất lượng sản phẩm cho phù hợp với nhu cầu khách hàng nhằm gia tăng số lượng sản phẩm được mua.

**- Tìm hiểu về mức cho điểm rating ở nam và nữ của từng dòng sản phẩm ở chi nhánh A?**

Nhận xét:

- Ở dòng sản phẩm Health and beauty, nữ cho điểm rating cao hơn nam nhưng tỉ lệ chênh lệch không cao lắm.

- Ở dòng sản phẩm Home and lifestyle, nam cho điểm cao hơn nữ nhưng tỉ lệ chênh lệch rất cao.

=> Các sản phẩm Home and lifestyle dành cho phái nữ ở chi nhánh A có thể không được yêu thích, cần xem xét lại hoặc thêm các sản phẩm mới để thử nghiệm về cảm nhận của người dùng nữ.

- Ở dòng sản phẩm Sports and travel, nữ cho điểm rating cao hơn nam nhưng tỉ lệ chênh lệch không đáng kể.

- Ở dòng sản phẩm Electronic accessories, nam cho điểm cao hơn nữ và tỉ lệ chênh lệch rất cao.

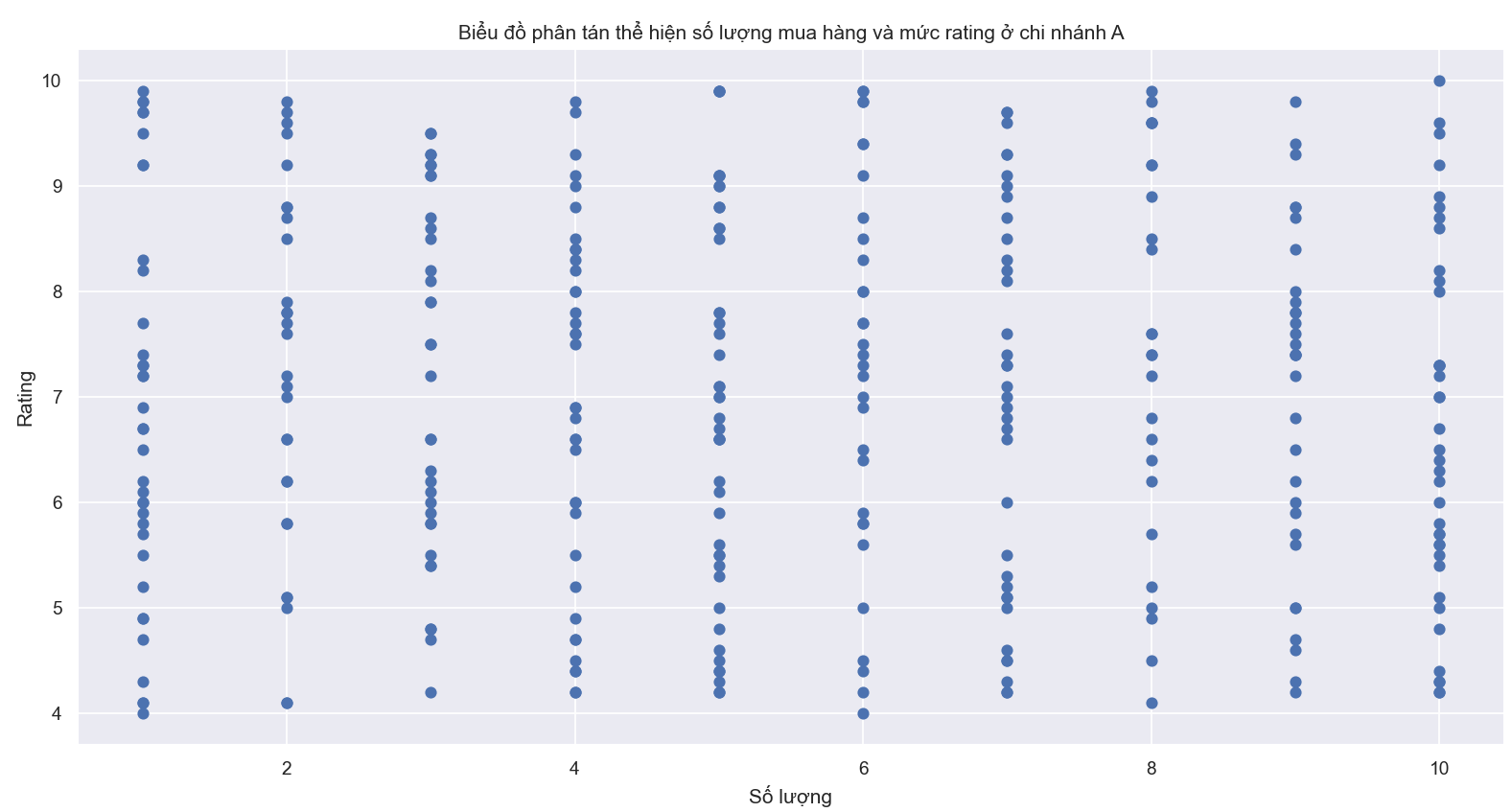
=> Cảm nhận của nữ giới đối với dòng sản phẩm này của chi nhánh A không tốt, cần xem xét và điều chỉnh những sản phẩm dành cho nữ thuộc dòng sản phẩm này.

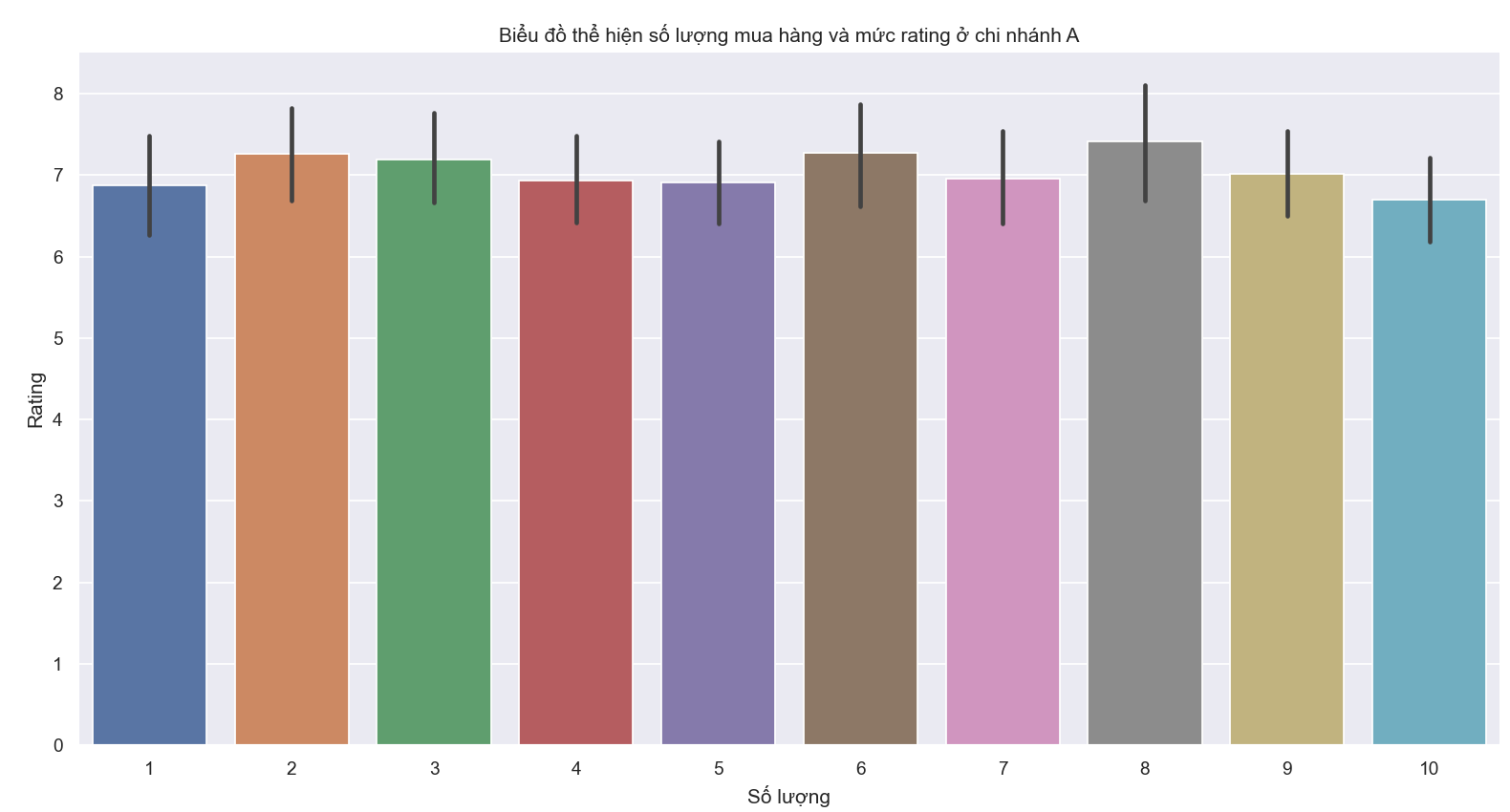
- Ở dòng sản phẩm Food and beverages, nam cho điểm cao hơn nữ và tỉ lệ chênh lệch khá cao.

=> Nên thêm những món ăn, thức uống phù hợp với nữ giới để có thể tăng độ ưa thích của khách hàng nữ.

- Ở dòng sản phẩm Fashion accessories, nữ cho điểm rating cao hơn nam dù tỉ lệ chênh lệch không cao lắm nhưng dòng sản phẩm này có điểm rating thấp nhất trong các dòng sản phẩm.

**- Tìm hiểu xem liệu rằng khách hàng mua sản phẩm với số lượng nhiều có cho điểm rating cao không ở chi nhánh A?**





Nhận xét:

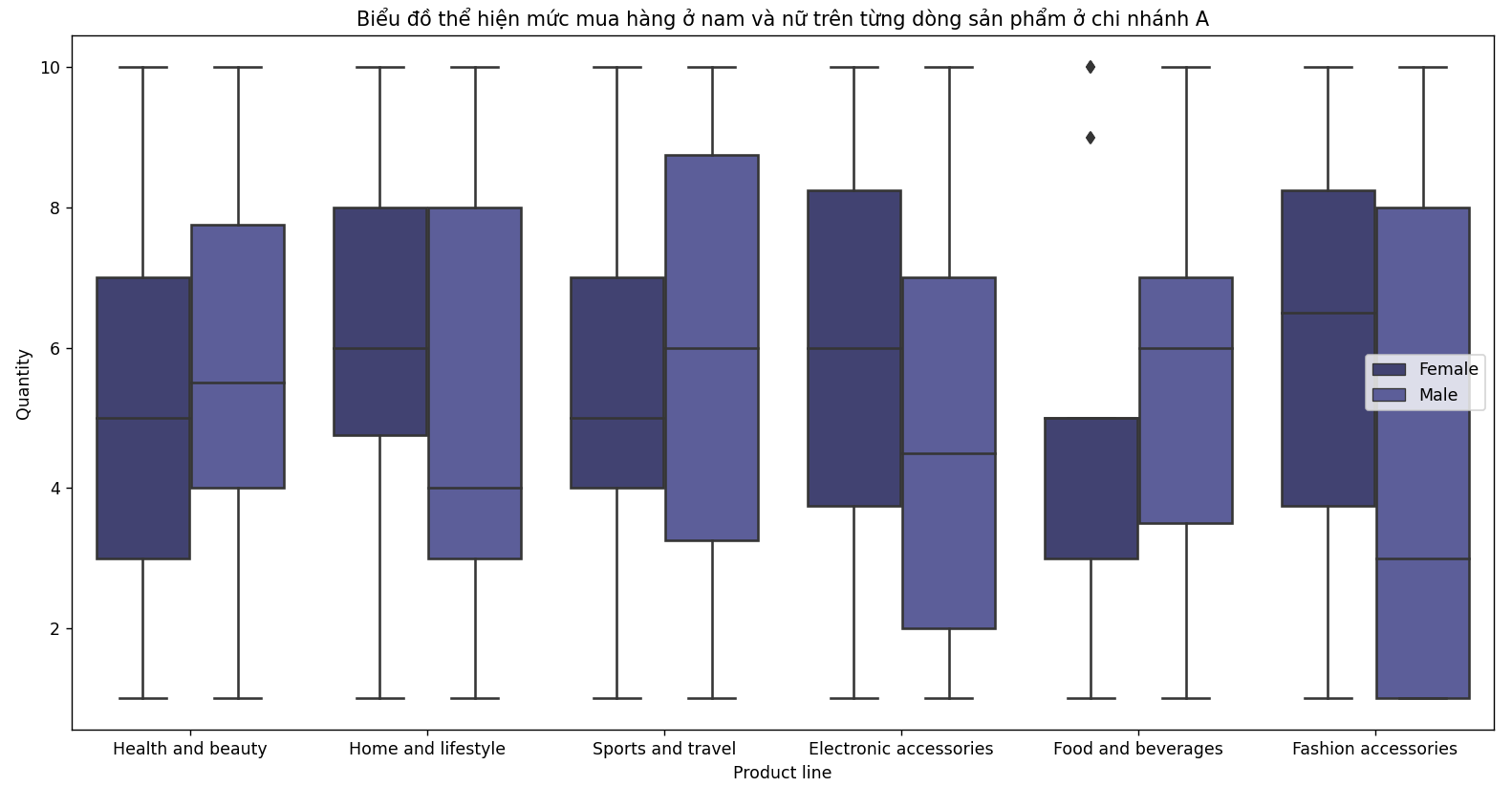
- Ở chi nhánh A thì số lượng bán ra với các dòng sản phẩm và mức rating chênh lệch cũng không quá nhiều.

- Với số lượng bán thấp nhất là 1 sản phẩm trên 1 hóa đơn thì rating trung bình đạt gần 7.0 điểm.

- Với số lượng bán cao nhất là 10 sản phẩm trên 1 hóa đơn thì rating trung bình cũng đạt gần 7.0 điểm.

=> Vậy việc khách hàng mua hàng với số lượng ít hay nhiều không ảnh hưởng gì với việc họ cảm nhận và cho điểm rating sản phẩm cao hay thấp. Từ biểu đồ, ta nhận thấy mức rating dao động chủ yếu là từ khoảng 6.7 đến gần 7.5 nên có thể suy ra cảm nhận của khách hàng ở chi nhánh A cũng chỉ ở mức khá.

**- Tìm hiểu xem giữa nam và nữ thì giới nào mua với số lượng nhiều hơn ở từng dòng sản phẩm ở chi nhánh A?**

Nhận xét:

- Ở dòng sản phẩm Health and beauty, nam mua với số lượng nhiều hơn nữ và tỉ lệ chênh lệch hơi cao.

=> Nên xem lại các sản phẩm dành cho nữ ở dòng sản phẩm Health and beauty để có kết quả bán hàng tốt hơn.

- Ở dòng sản phẩm Sports and travel, nam mua nhiều hơn nữ và tỉ lệ chênh lệch rất cao.

=> Khách hàng nữ có vẻ không hứng thú với các sản phẩm hiện tại của dòng sản phẩm Sports and travel nên thêm các sản phẩm mới để tìm hiểu về nhu cầu của khách hàng nữ rõ hơn.

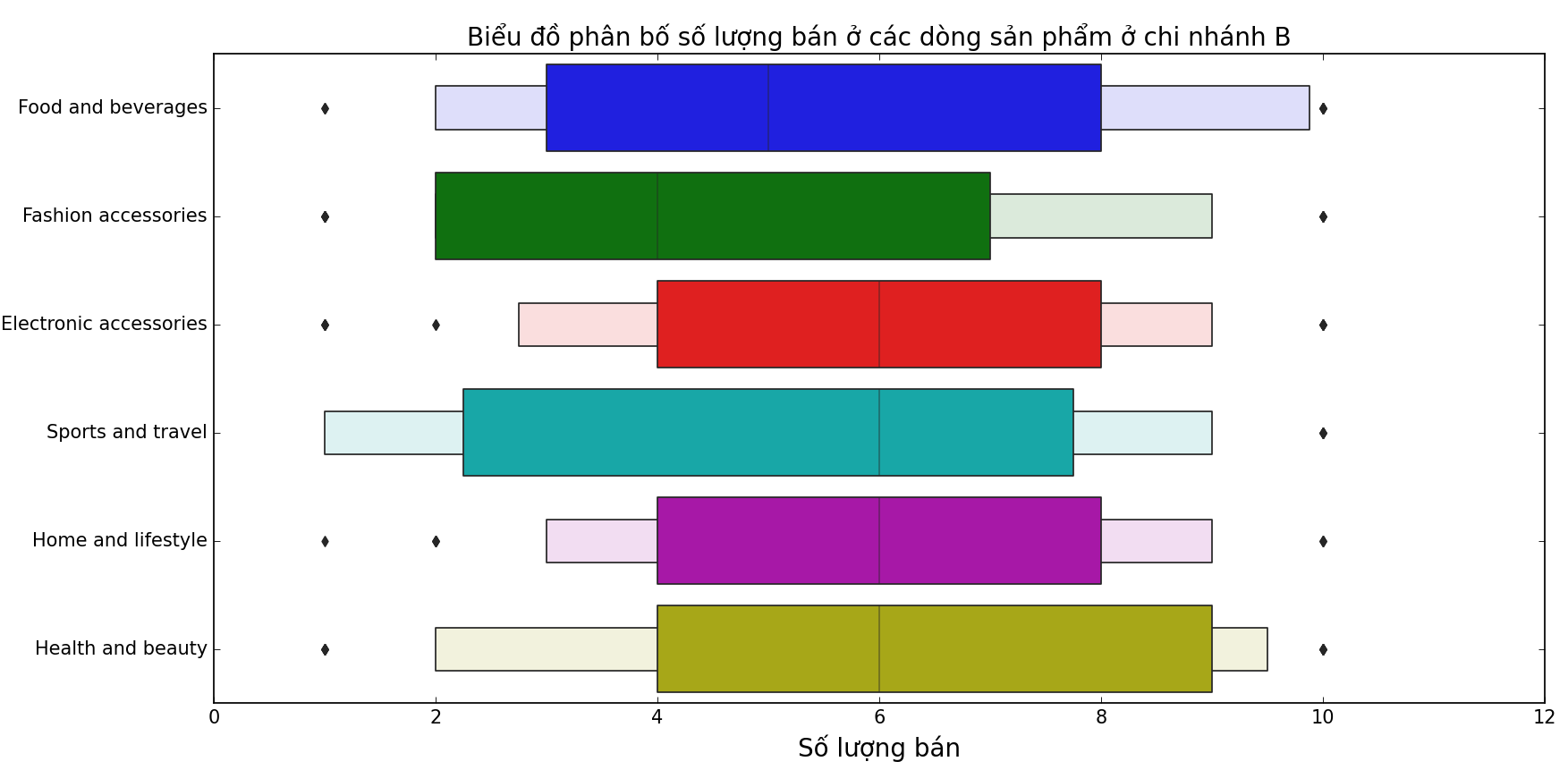
- Ở dòng sản phẩm Electronic accessories, nữ mua nhiều hơn nam và tỉ lệ chênh lệch cao.

=> Dòng sản phẩm này vốn phải được ưa chuộng bởi phái nam nhiều hơn nhưng ở đây ta lại thấy các khách hàng nam mua với số lượng thấp hơn khách nữ, vì thế cần xem lại các sản phẩm dành cho giới nam ở dòng sản phẩm này.

- Ở dòng sản phẩm Food and beverages, nam mua nhiều hơn nữ và tỉ lệ chênh lệch cực kì cao, đây cũng là dòng sản phẩm được mua với số lượng thấp hơn hẳn so với các dòng sản phẩm còn lại.

=> Cần xem lại các sản phẩm thuộc dòng sản phẩm này, nếu được hãy loại bỏ các sản phẩm hiện tại và thay thế bằng các sản phẩm mới để phù hợp với nhu cầu của khách hàng, đặc biệt là khách hàng nữ.

- **Tìm hiểu về sự phân bố số lượng bán ra của từng dòng sản phẩm ở chi nhánh B?**



Nhận xét:

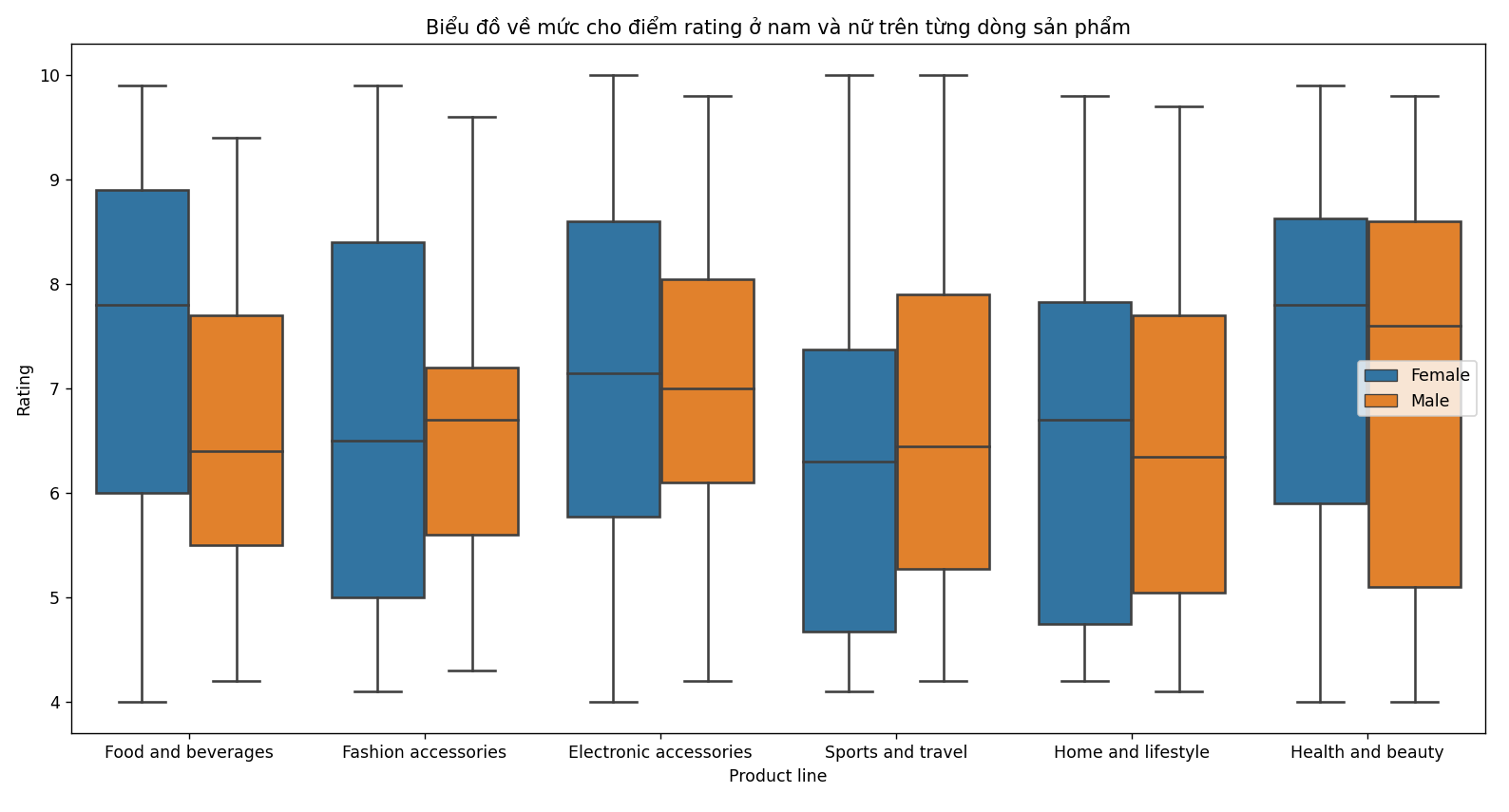
- Các dòng sản phẩm khi được bán ở chi nhánh B có số lượng bán dao động trung bình từ 4 - 8.

- Dòng sản phẩm Food and beverages có số lượng bán nhiều cao hơn các dòng sản phẩm khác.

- 2 dòng sản phẩm Fashion accessories và Sports and travel có số lượng bán chỉ từ 2 - 4 chiếm khá cao

- 2 dòng sản phẩm Electronic accessories và Home and lifestyle có số lượng bán ở mức cao từ 8 -10 chiếm ít

**- Tìm hiểu về mức cho điểm rating ở nam và nữ của từng dòng sản phẩm ở chi nhánh B?**



NNhận xét:

- Ở dòng sản phẩm Food and beverages, nữ cho điểm cao hơn nam và tỉ lệ chênh lệch rất cao.

=> Cần xem xét hoặc thêm các sản phẩm mới để tìm hiểu về nhu cầu của khách nam đối với dòng sản phẩm này.

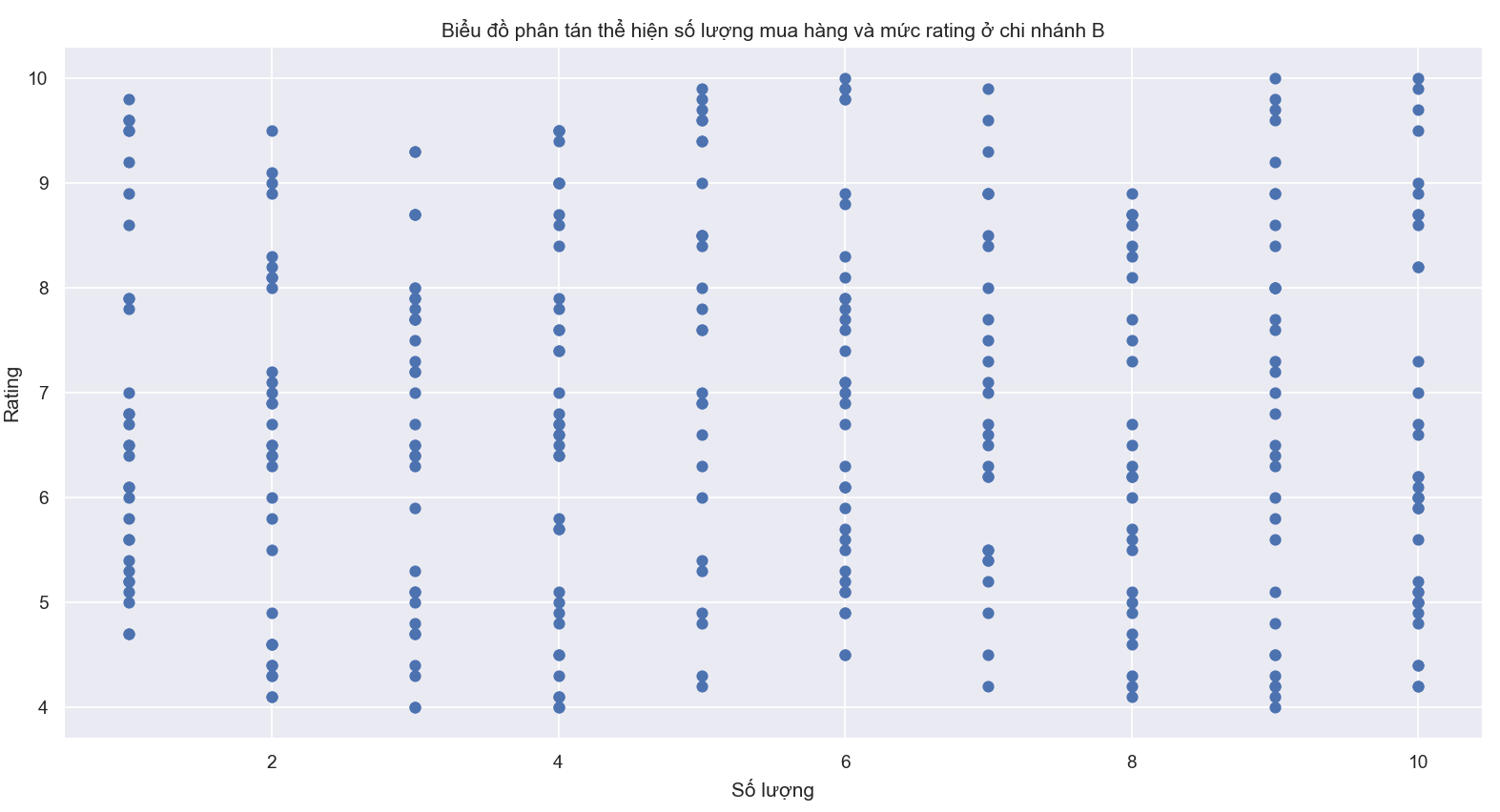
- Ở dòng sản phẩm Fashion accessories, nữ cho điểm cao hơn nam và tỉ lệ chênh lệch rất cao.

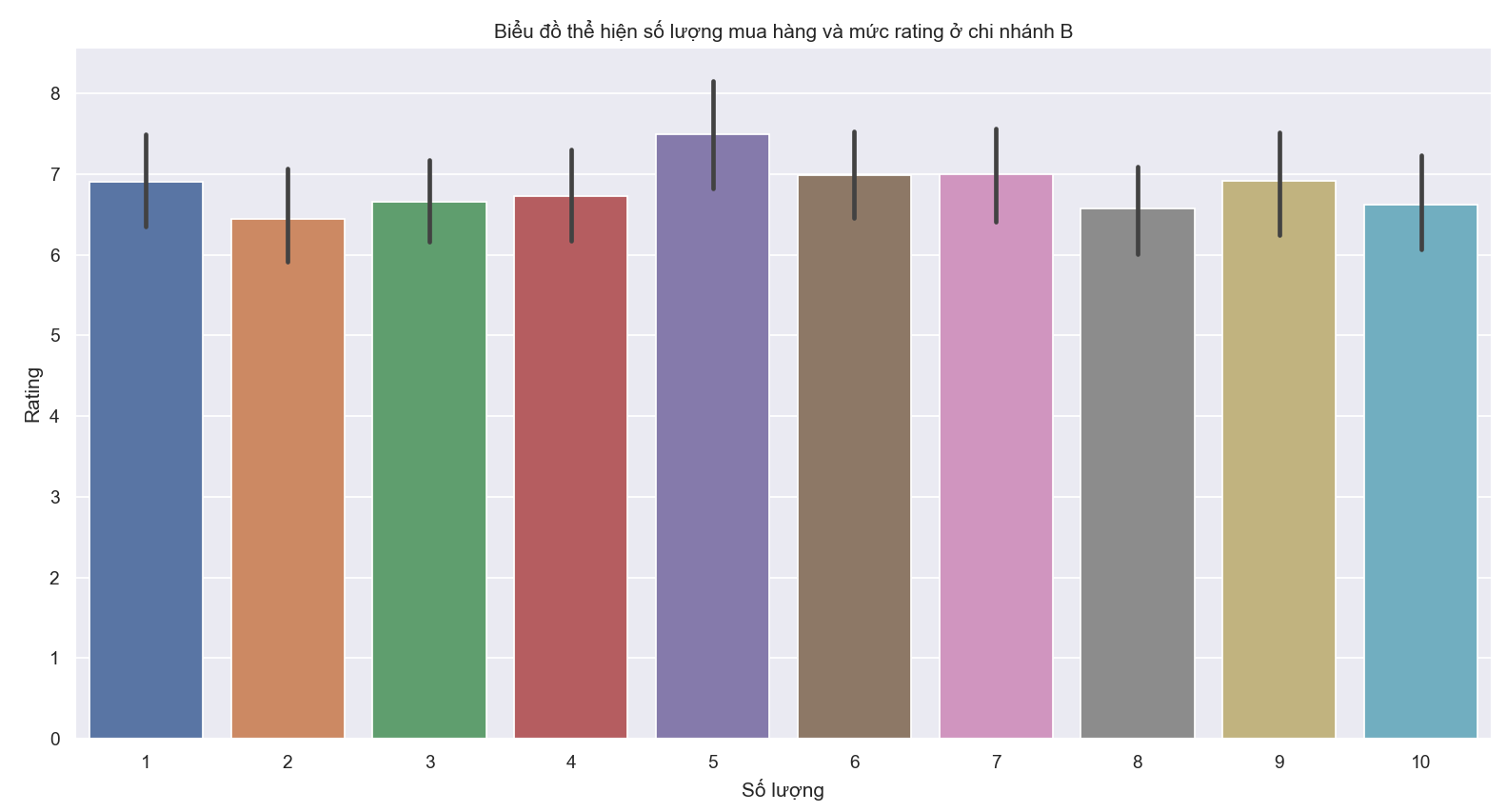
=> Các khách hàng nam ở chi nhánh B có cảm nhận có vẻ không được tốt lắm đối vớicác sản phẩm thuộc dòng sản phẩm này nên cần xem lại hiện tại dành cho khách nam.

- Dòng sản phẩm Home and lifestyle, nữ cho điểm cao hơn nam nhưng tỉ lệ không đáng kể và dòng sản phẩm này nhận về điểm rating thấp từ cả 2 giới nam và nữ, dòng sản phẩm Sprots and travel cũng tương tự nhưng nam cho điểm cao hơn nữ.

=> Cần cân nhắc thay đổi những sản phẩm thuộc những dòng sản phẩm này.

**- Tìm hiểu xem liệu rằng khách hàng mua sản phẩm với số lượng nhiều có cho điểm rating cao không ở chi nhánh B?**





Nhận xét:

- Nhìn chung ở chi nhánh B mức chênh lệch điểm rating giữa các số lượng bán được cũng không cao lắm.

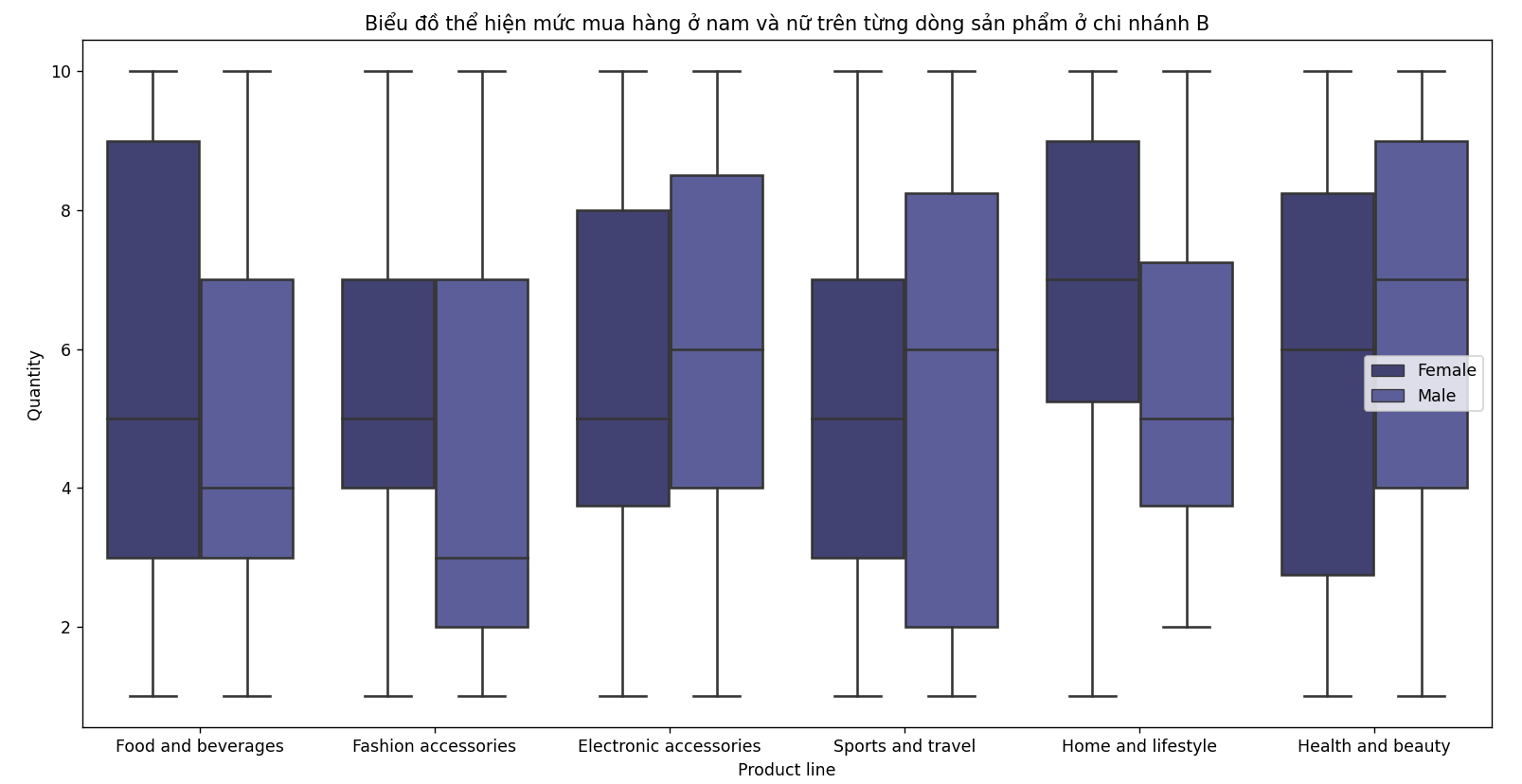
- Trung bình điểm rating cao nhất nằm ở số lượng là 5.

- Với số lượng sản phẩm được mua là 1 trên 1 hóa đơn thì điểm rating là gần 7.0.

- Với số lượng sản phẩm được mua là 10 trên 1 hóa đơn thì điểm rating là hơn 6.5 thấp hơn với số lượng sản phẩm là 1.

=> Vậy với những hóa đơn có số lượng bán nhiều thì không thể kết luận được khách hàng cho điểm rating cao và có cảm nhận tốt về sản phẩm.

**- Tìm hiểu xem giữa nam và nữ thì giới nào mua với số lượng nhiều hơn ở từng dòng sản phẩm ở chi nhánh B?**



Nhận xét:

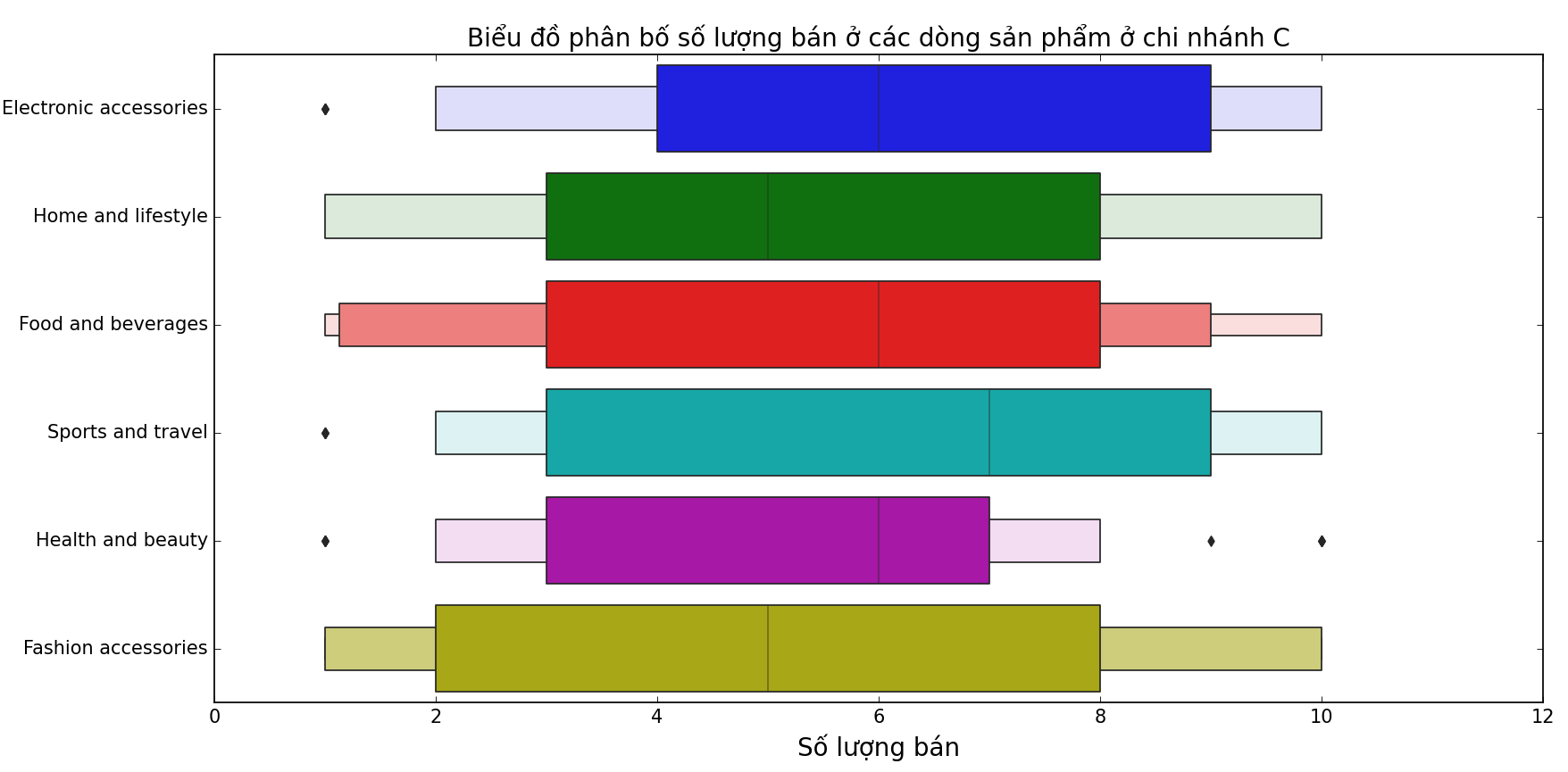
- Ở dòng sản phẩm Food and beverages, nữ mua nhiều hơn nam và tỉ lệ chênh lệch rất cao.

=> Đối với dòng sản phẩm này, khách hàng nam thực sự là chưa hài lòng lắm, cần cải thiện sản phẩm theo nhu cầu của khách nam để thu được kết quả kinh doanh tốt hơn.

- Dòng sản phẩm Fashion accessories có mức mua hàng thấp nhất ở cả nam và nữ

=> Dòng sản phẩm này chưa thực sự làm hài lòng khách hàng, cần thêm các sản phẩm mới cho dòng sản phẩm này để thu kết quả tốt hơn.

**- Tìm hiểu về sự phân bố số lượng bán ra của từng dòng sản phẩm ở chi nhánh C?**



Nhận xét:

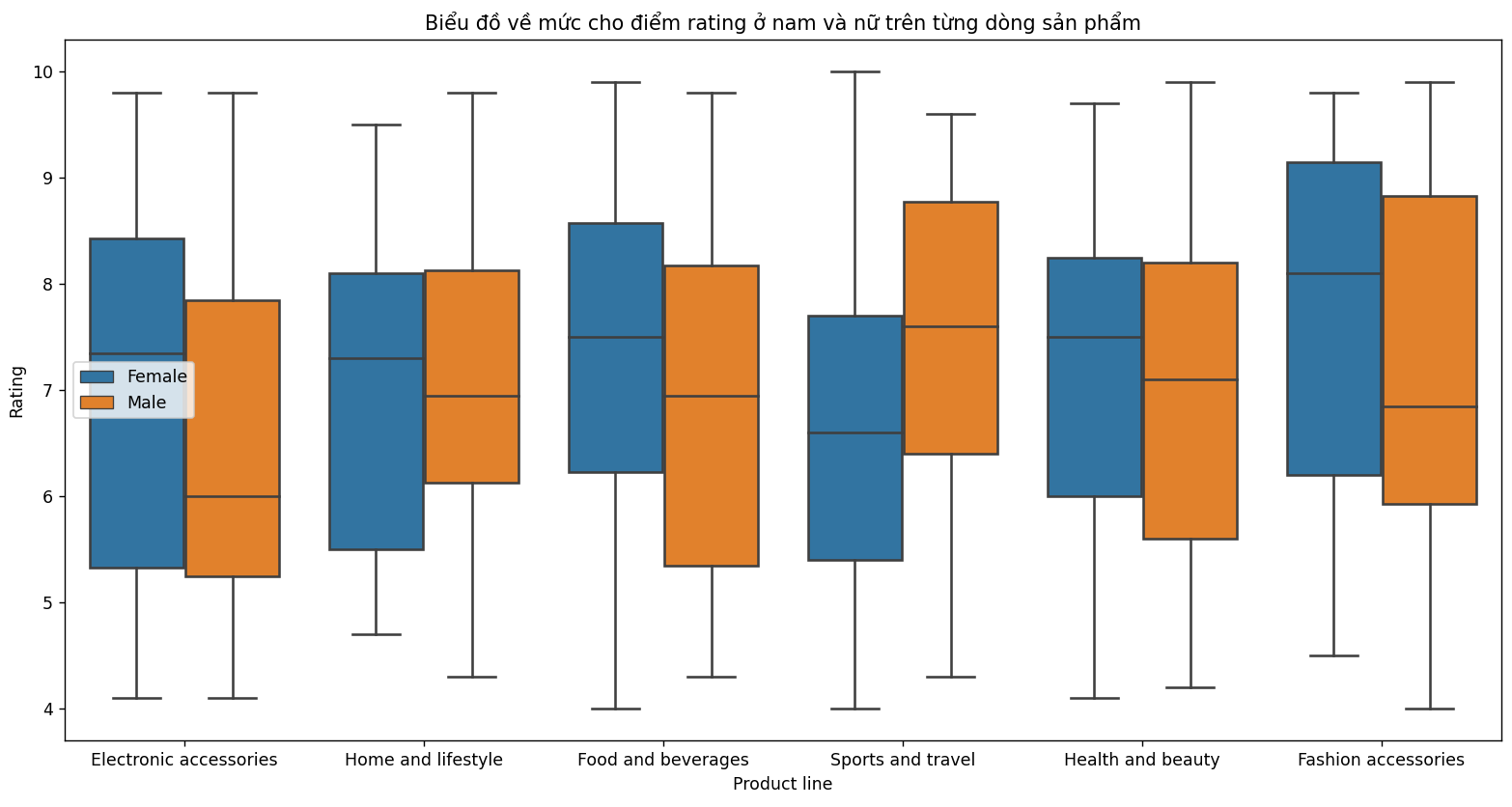
- Ở chi nhánh C thì 2 dòng Food and beverages và Home and lifestyle có mức số lượng bán phân bố đều.

- Các dòng sản phẩm còn lại tuy thấp hơn chút nhưng mức độ phân bố về số lượng vẫn khá đều.

- Riêng dòng sản phẩm Health and beauty có số lượng bán ở mức nhiều thấp hơn rất nhiều so với các dòng sản phẩm khác.

=> Cần xem xét dòng sản này để tìm ra nhu cầu của khách hàng.

**- Tìm hiểu về mức cho điểm rating ở nam và nữ của từng dòng sản phẩm ở chi nhánh C?**



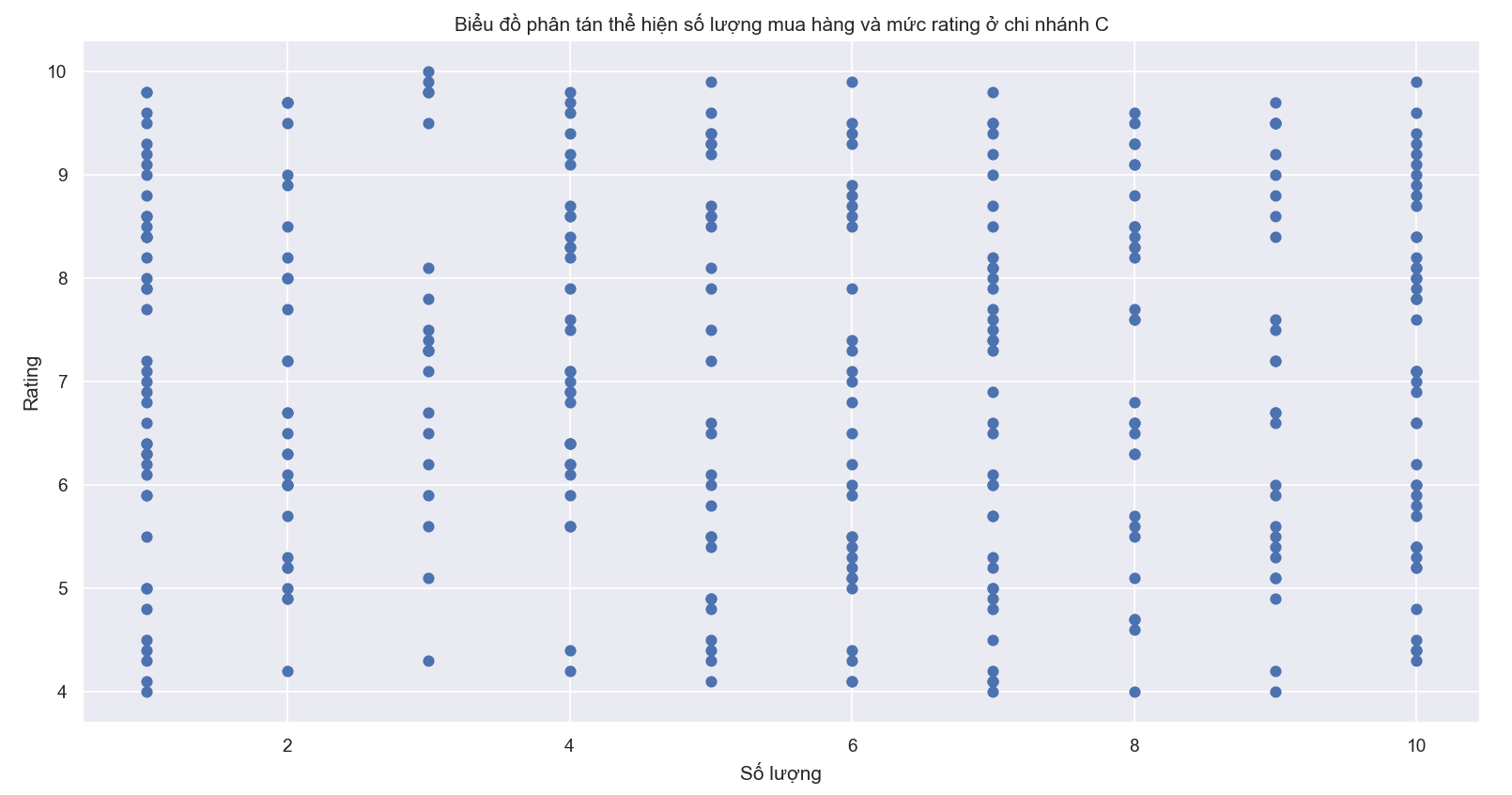
Nhận xét:

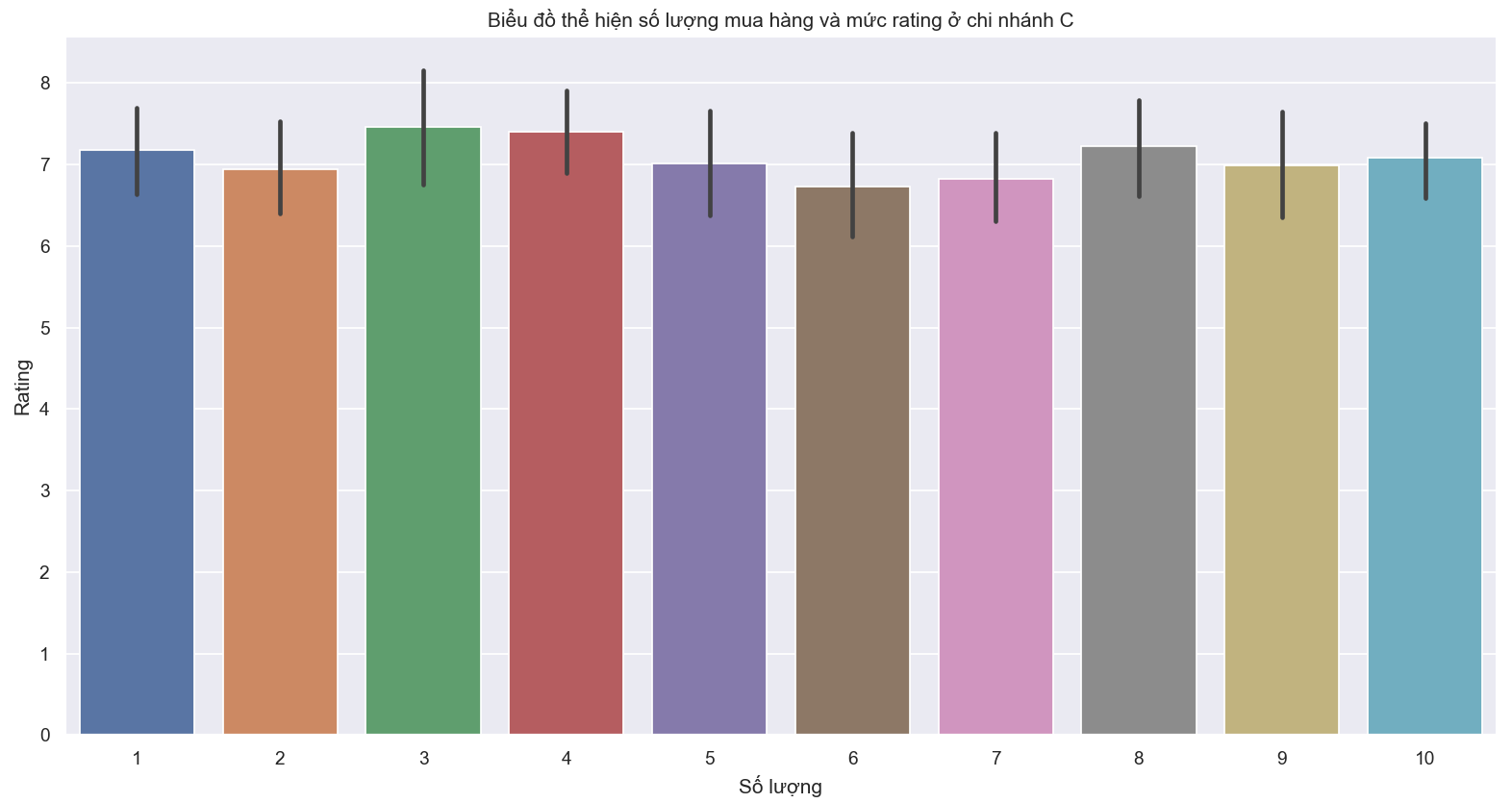
- Ở dòng sản phẩm Electronic accessories, nữ cho điểm cao hơn nam và tỉ lệ chênh lệch hơi cao.

- 2 dòng sản phẩm Home and lifestyle và Health and beauty nhận được sự đánh giá như nhau ở cả nam và nữ.

- Dòng sản phẩm Sports and travel ở chi nhánh C vẫn nhận được mức cho điểm rating thấp ở giới nữ giống với 2 chi nhánh kia.

**- Tìm hiểu xem liệu rằng khách hàng mua sản phẩm với số lượng nhiều có cho điểm rating cao không ở chi nhánh C?**





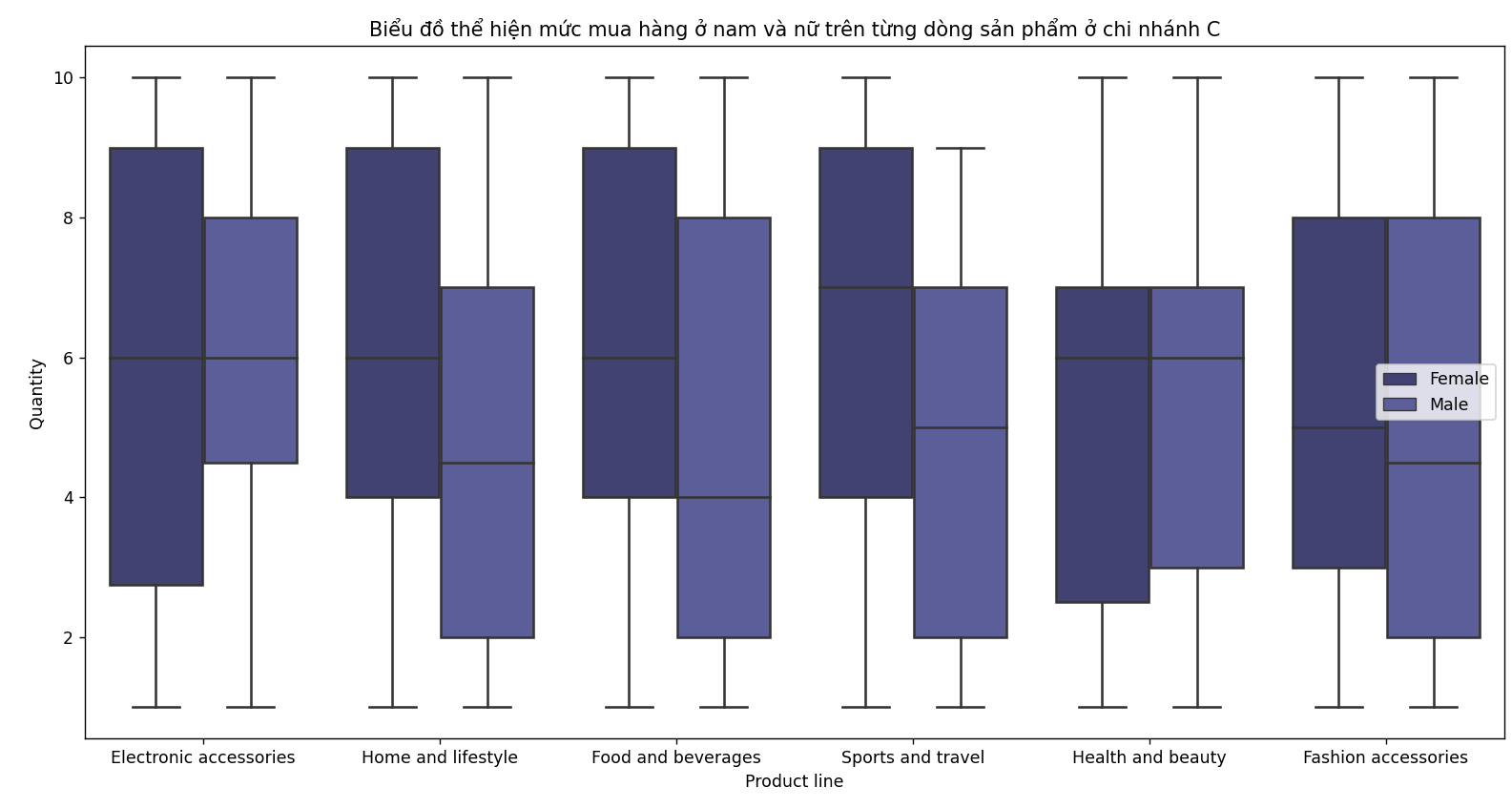
**Nhận xét:**

- Cũng như hai chi nhánh A, B mức rating giữa số lượng sản phẩm bán được vẫn chênh lệch không đáng kể.

- Với số lượng sản phẩm bán được là 1 trên 1 hóa đơn thì rating đạt khoảng hơn 7.0.

- Với số lượng sản phẩm bán được là 10 trên 1 hóa đơn thì rating cũng đạt khoảng hơn 7.0.

=> Vậy ở chi nhánh C khách hàng mua sản phẩm nhiều hay ít không ảnh hưởng đến việc họ cho điểm rating.

**- Tìm hiểu xem giữa nam và nữ thì giới nào mua với số lượng nhiều hơn ở từng dòng sản phẩm ở chi nhánh C?**

Nhận xét:

- Ở dòng sản phẩm Electronic accessories, nữ cho điểm mua hàng nhiều hơn nam nhưng nhìn chung mức mua hàng trung bình ở cả 2 giới vẫn ở mức khá.

- Dòng sản phẩm Health and beauty và Fashion accessories đều nhận được mức mua hàng mua ở cả nam và nữ như nhau, tuy nhiên dòng sản phẩm Health and beauty có số lượng sản phẩm được mua nhiều hơn.

- 3 dòng sản phẩm Home and lifestyle, Sports and travel và Food and beverages đều có số lượng mua hàng ở nữ cao hơn nam. Tuy nhiên, dòng sản phẩm Home and lifestyle và Sports and travel có số lượng mua ở nam thấp hơn nữ rất nhiều nhất là dòng sản phẩm Sports and travel vốn có số lượng bán ở nam rất tốt ở chi nhánh A, B nhưng lại không được giới nam mua nhiều ở chi nhánh C.

**CHƯƠNG 3. PHÂN LỚP DỮ LIỆU**

## 3.1. Các phương pháp pháp dự đoán vả quy trình cụ thể

### 3.1.1. Phân lớp dữ liệu

Phân lớp dữ liệu là quá trình phân một đối tượng dữ liệu vào một hay nhiều (loại) đã cho trước nhờ một mô hình phân lớp. Mô hình này đã được xây dựng dựa trên một tập dữ liệu đã được gán nhãn trước đó. Quá trình gán nhãn cho một đối tượng dữ liệu chính là quá trình phân lớp.

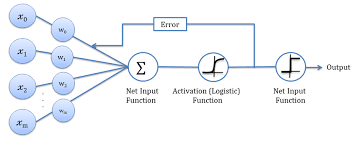
**3.1.1.1. Hồi quy logistic (Logistic Regression)**

[Hồi quy logistic (Logistic Regression) là một mô hình thống kê được sử dụng để ước lượng các tham số của một mô hình logistic, một hàm chuyển đổi log-odds thành xác suất](https://en.wikipedia.org/wiki/Logistic_regression). [Nó được sử dụng để mô hình hóa xác suất của một sự kiện xảy ra bằng cách có log-odds là tổ hợp tuyến tính của một hoặc nhiều biến độc lập](https://en.wikipedia.org/wiki/Logistic_regression). [Nó có thể được sử dụng cho các phản ứng nhị phân, phân loại, hoặc đa phân loại](https://en.wikipedia.org/wiki/Logistic_regression).

[Hồi quy logistic ước lượng xác suất của một sự kiện xảy ra, chẳng hạn như bỏ phiếu hoặc không bỏ phiếu, dựa trên một tập dữ liệu cho trước của các biến độc lập2](https://www.ibm.com/topics/logistic-regression). [Vì kết quả là một xác suất, biến phụ thuộc được giới hạn giữa 0 và 1](https://www.ibm.com/topics/logistic-regression). [Trong hồi quy logistic, một biến đổi logit được áp dụng trên tỷ lệ - tức là, xác suất thành công chia cho xác suất thất bại](https://www.ibm.com/topics/logistic-regression). [Đây cũng được biết đến như log odds, hoặc logarith tự nhiên của tỷ lệ, và hàm logistic này được biểu diễn bởi các công thức sau](https://www.ibm.com/topics/logistic-regression):

**Logit (pi) = 1/ (1+ exp (-pi))**

**ln (pi/ (1-pi)) = Beta\_0 + Beta\_1\*X\_1 + … + B\_k\*K\_k**

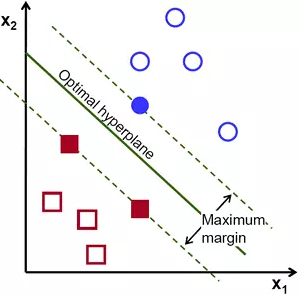


*Minh họa về mô hình hồi quy logistic*

**3.1.1.2. Phương pháp SVM (Support Vector Machine)**

Phương pháp Máy vector hỗ trợ (Support Vector Machine - SVM) là một mô hình học có giám sát được sử dụng để phân loại và phân tích hồi quy. SVM ánh xạ các ví dụ huấn luyện thành các điểm trong không gian nhiều chiều để tối đa hóa độ rộng của khoảng cách giữa hai danh mục. Nó cũng có thể thực hiện phân loại phi tuyến bằng cách sử dụng một kỹ thuật nhân (kernel trick). SVM được sử dụng trong nhiều vấn đề thực tế như phân loại văn bản và hình ảnh, phát hiện ngoại lệ và phân tích protein.

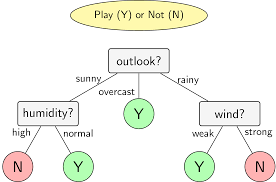
SVM tìm kiếm mặt phẳng phân chia tối ưu trong không gian N chiều có thể phân tách các điểm dữ liệu thành các lớp khác nhau trong không gian đặc trưng. Mặt phẳng cố gắng tạo ra khoảng cách lớn nhất giữa các điểm gần nhất của các lớp khác nhau. Kích thước của mặt phẳng phụ thuộc vào số lượng đặc trưng. Nếu số lượng đặc trưng đầu vào là hai, thì mặt phẳng chỉ là một đường thẳng. Nếu số lượng đặc trưng đầu vào là ba, thì mặt phẳng trở thành một mặt phẳng 2-D.

  
*Minh họa về mô hình SVM*

**3.1.1.3. Phương pháp Cây quyết định (Decision Tree)**

Cây quyết định (Decision Tree) là một mô hình học có giám sát được sử dụng cho cả nhiệm vụ phân loại và hồi quy. Nó xây dựng một cấu trúc cây giống như biểu đồ luồng, nơi mỗi nút nội bộ đại diện cho một kiểm tra trên một thuộc tính, mỗi nhánh đại diện cho kết quả của kiểm tra, và mỗi nút lá (nút cuối cùng) chứa một nhãn lớp. Nó được xây dựng bằng cách chia tập dữ liệu huấn luyện thành các tập con dựa trên giá trị của các thuộc tính cho đến khi đạt được tiêu chí dừng, chẳng hạn như độ sâu tối đa của cây hoặc số lượng mẫu tối thiểu cần thiết để chia một nút.

Trong quá trình huấn luyện, thuật toán Cây quyết định chọn thuộc tính tốt nhất để chia dữ liệu dựa trên một chỉ số như entropy hoặc độ không tinh khiết Gini, đo lường mức độ không tinh khiết hoặc ngẫu nhiên trong các tập con. Mục tiêu là tìm ra thuộc tính tối đa hóa lợi ích thông tin hoặc giảm độ không tinh khiết sau khi chia.

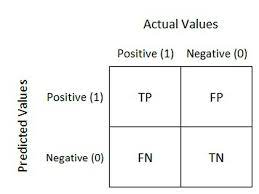


Mô hình Cây quyết định

**3.1.1.4. Các phương pháp đánh giá mô hình phân lớp**

**3.1.1.4.1. Ma trận nhầm lẫn**

Là mà trận chỉ ra có bao nhiêu điểm dữ liệu thực sự thuộc vào một lớp cụ thể và được dự đoán là rơi vào lớp nào.



*Minh họa về ma trận nhầm lẫn*

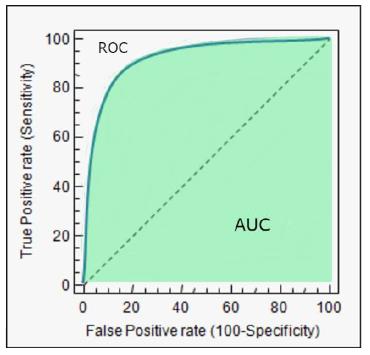
**3.1.1.4.2. Tính chính xác (Accuaracy)**

Tính chính xác (Accuracy) là tỷ lệ số mẫu được phân lớp đúng trong toàn bộ dữ liệu. Nó chỉ cho ta biết tỷ lê dữ liệu đúng (chung), còn cụ thể đúng lớp nào thì không có. Tuy nhiên, có vẫn giúp chúng ta đánh giá hiệu quả dự báo của mô hình trên một bộ dữ

liệu. Độ chính xác càng cao thì mô hình càng chuẩn xác

**3.1.1.4.3. AUC (Area Under the Curve)**

AUC là diện tích nằm dưới đường cong ROC và có giá trị là một số dương nhỏ hơn hoặc bằng 1. AUC càng lớn thì độ chính xác của mô hình càng tốt.



*Minh họa phương pháp AUC*

### 3.1.2. Phân cụm dữ liệu

Phân cụm dữ liệu là quá trình gom cụm/ nhóm các đối tượng/dữ liệu có đặc điểm tương đồng vào các nhóm/ cụm tương ứng. Trong đó, tương đồng giữa những phần tử trong cùng cụm; khác biệt với với những phần tử trong các cụm khác.

**- Đánh giá mô hình phân cụm:**

+ Ưu điểm của phân cụm phân lớp là không phải xác định trước số lượng

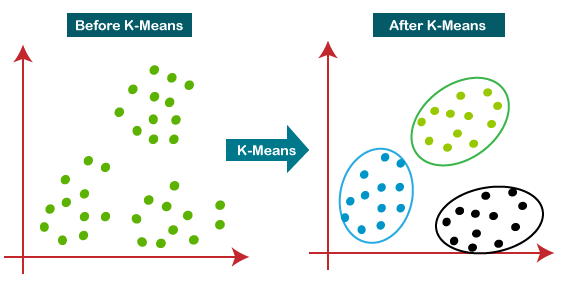
cụm điều này khá vượt trội so với K-Means. Tuy nhiên, nó không hoạt động tốt với

lượng dữ liệu khổng lồ.

+ Thuật toán phân cụm phân lớp có thể được sử dụng để xác định, dự đoán số

cụm trước khi thực hiện thuật toán K-Means.

🟎Thuật toán K-Means là một thuật toán học không giám sát, được sử dụng để giải quyết các vấn đề phân cụm trong học máy hoặc khoa học dữ liệu. Nó hoạt động bằng cách chia tập dữ liệu không được gán nhãn thành k nhóm khác nhau sao cho mỗi dữ liệu chỉ thuộc về một nhóm có các thuộc tính tương tự.



*Mô hình Thuật toán K-Means*

## 3.2. Thực nghiệm bài toán: Dự báo nhu cầu và hành vi của khách hàng

### 3.2.1. Mô tả bài toán

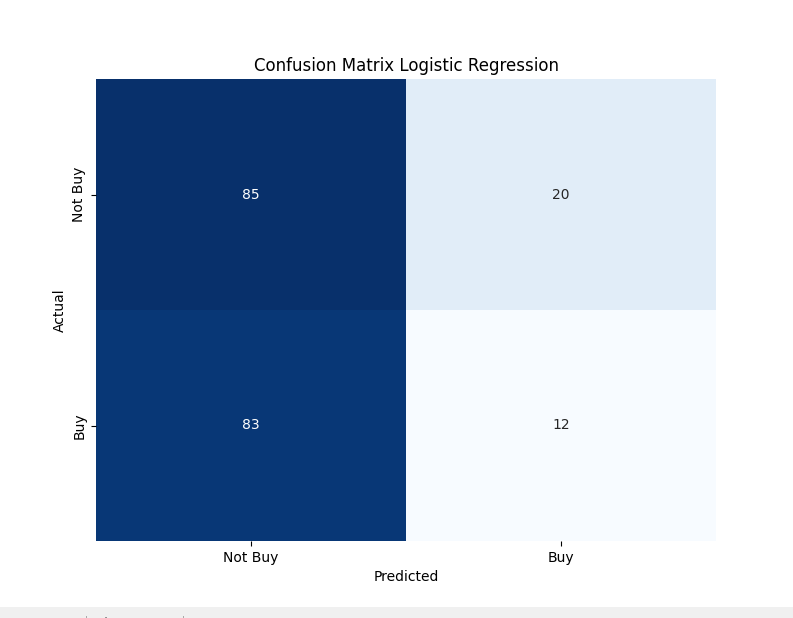
Trong bối cảnh kinh doanh, việc dự báo nhu cầu và hành vi của khách hàng là quan trọng để giúp doanh nghiệp chuẩn bị cho tương lai, quản lý nguồn lực, và đưa ra các chiến lược kinh doanh có hiệu quả. Bài toán này đặt ra thách thức xây dựng một mô hình dự báo chính xác dựa trên các biến quan trọng như thời gian, mùa vụ, giá cả và các yếu tố khác có thể ảnh hưởng đến doanh số bán hàng.

### 3.2.2. Chạy mô hình và kết quả

**3.2.2.1. Theo ma trận nhầm lẫn ta có:**

+ 0: Không mua

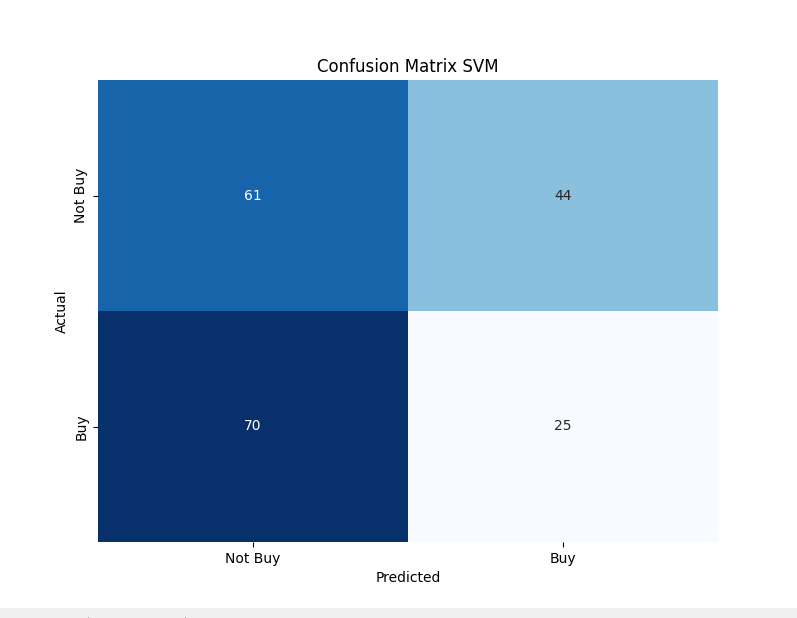
+ 1: Mua

****

*Kết quả ma trận nhầm lẫn của phương pháp Hồi quy Logistic*

Kết quả của phương pháp Hồi quy Logictic cho thấy:

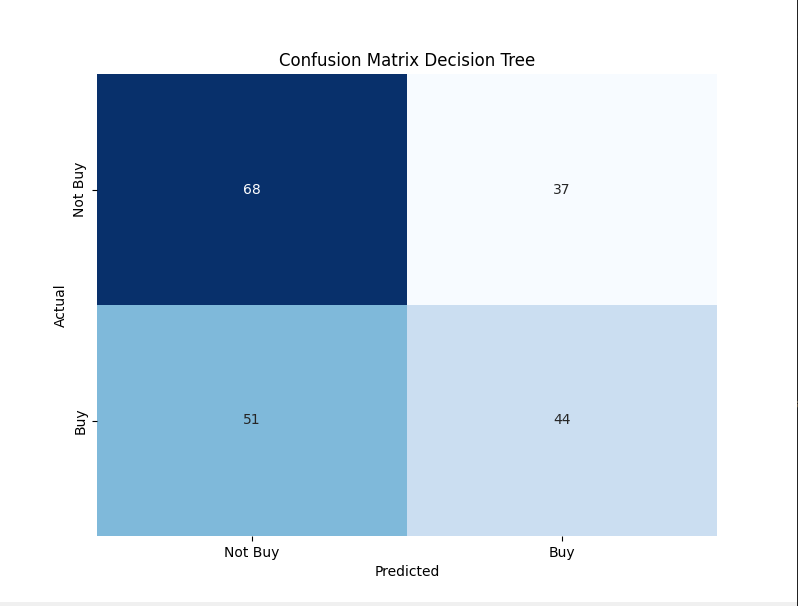
* True Positive (TP): 12 (Dự đoán đúng có mua)
* True Negative (TN): 85 (Dự đoán đúng không mua)
* False Positive (FP): 20 (Dự đoán sai có mua thành không mua)
* False Negative (FN): 83 (Dự đoán sai không mua thành có mua)

****

*Kết quả ma trận nhầm lẫn của phương pháp SVM*

Kết quả của phương pháp Hồi quy Logictic cho thấy:

* True Positive (TP): 25
* True Negative (TN): 61
* False Positive (FP): 44
* False Negative (FN): 70

****

*Kết quả ma trận nhầm lẫn của phương pháp Cây quyết định*

Kết quả của phương pháp Cây quyết định cho thấy:

* True Positive (TP): 44
* True Negative (TN): 68
* False Positive (FP): 37
* False Negative (FN): 51

**Xét ma trận nhầm lẫn cho sai lầm loại 2:**

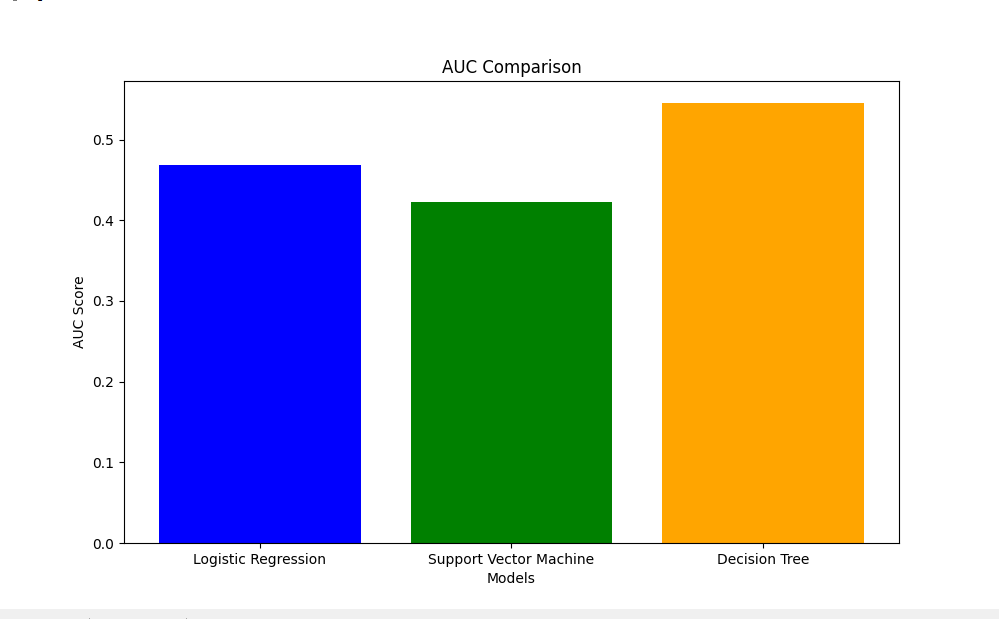
Sai lầm loại 2 : Dự báo rủi ro là 0 nhưng thực tế rủi ro là 1

LR (= 83) > SVM (= 70) > Tree (= 51)

=> Phương pháp Hồi quy Logistic lớn nhất -> Loại bỏ Phương pháp Hồi quy Logistic

**3.2.2.2. Theo AUC (Area Under the Curve):**

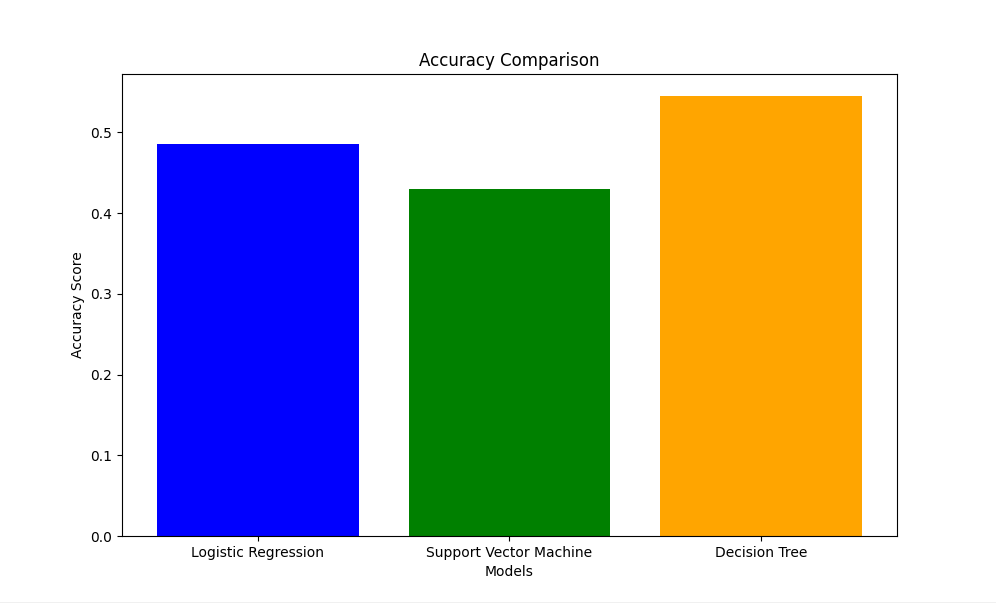
|  |  |  |  |
| --- | --- | --- | --- |
| **Phương pháp** | **LR** | **SVM** | **DT** |
| **AUC** | **0.468** | **0.422** | **0.555** |

****

=> Kết quả thu được cho thấy AUC của phương pháp Cây quyết định cao nhất.

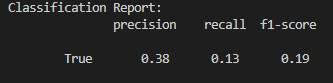
**3.2.2.3. Theo tính chính xác (Accuaracy)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Phương pháp** | **LR** | **SVM** | **DT** |
| **Accuaracy** | **0.48** | **0.43** | **0.555** |

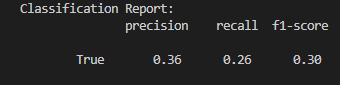
****

=> Kết quả của Tính chính xác của phương pháp Cây quyết định cao nhất.

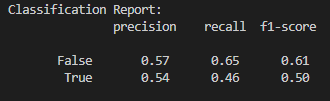
**3.2.2.4. Theo như kết quả hiệu suất của mô hình phân loại**



*Hiệu suất của mô hình Hồi quy Logistic*



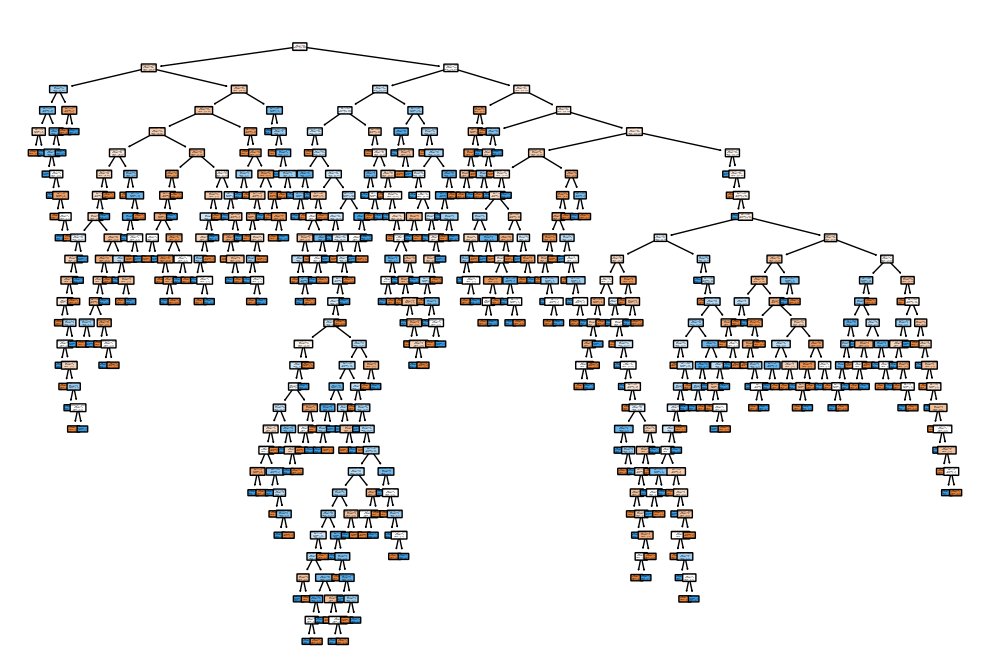
*Hiệu suất của mô hình SVM*



*Hiệu suất của mô hình Cây quyết định*

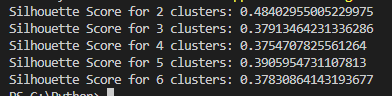
*=> Kết qủa các độ đo độ đo như precision, recall, F1-score của cây Quyết định cao nhất*

**Sử dụng thuật toán cây quyết định đưa qua quyết định nhu cầu hành vi của khách hàng**

**

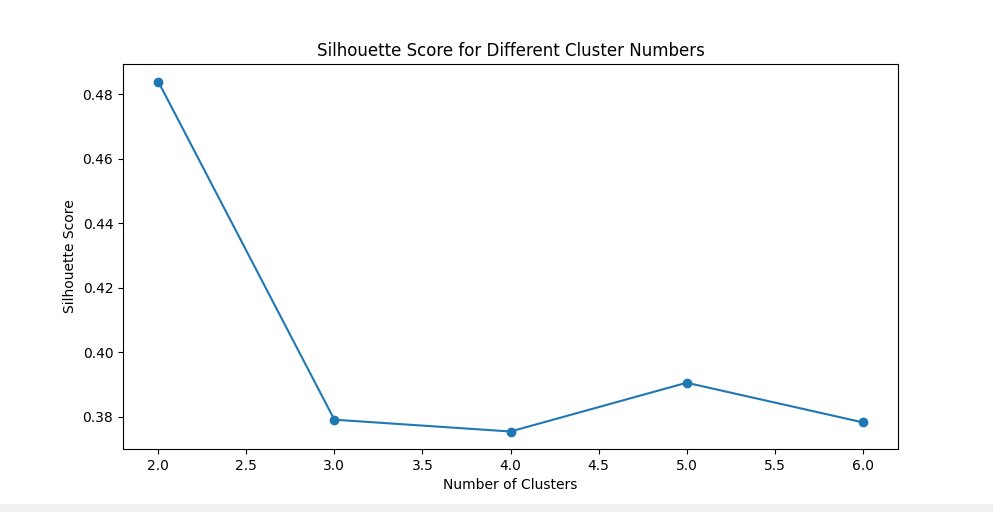
*Cây quyết định cho dữ liệu phân tích nhu cầu hành vi khách hàng*

# CHƯƠNG 4: PHÂN CỤM DỮ LIỆU



*Kết quả chạy trên 6 cụm*

Kết quả phân loại K-means chạy từ 2-6 cụm, chọn phân thành 2 cụm tương ứng với điểm Sihouette Plot cao nhất là 0.48



# KẾT QUẢ CỦA QUÁ TRÌNH PHÂN LỚP VÀ PHÂN CỤM

**Kết luận phân lớp dữ liệu**  
Dựa trên kết quả đánh giá của ba phương pháp (Hồi quy Logistic, SVM, và Cây quyết định) đối với bài toán dự báo doanh số bán hàng, ta có thể rút ra một số kết luận:

1. **Hồi quy Logistic và SVM:**
   * Độ chính xác của cả hai mô hình là khá thấp, chỉ khoảng 48.5%.
   * Cả hai mô hình đều có khả năng dự đoán lớp "True" (doanh số bán hàng cao) không hiệu quả, như được thể hiện qua recall thấp cho lớp này.
   * Điều này có thể là dấu hiệu rằng các mô hình này có thể cần được điều chỉnh hoặc có thể không phù hợp cho bài toán cụ thể này.
2. **Cây quyết định:**
   * Cây quyết định có hiệu suất tốt hơn với độ chính xác khoảng 55.5%, và AUC Score là 55.1%.
   * Mặc dù có cải thiện so với hai phương pháp trước đó, nhưng vẫn còn có thể cải thiện, đặc biệt là trong việc xác định trường hợp tích cực (lớp "True").

**Kết luận phân cụm dữ liệu**

Theo đánh giá kết của của phương pháp K-Means cho thấy giá trị Silhouette Score cao đồng nghĩa với sự tách biệt tốt giữa các cụm. Trong trường hợp này, có vẻ như số lượng cụm tốt nhất là 2, vì có Silhouette Score cao nhất cho 2 cụm.

Phân cụm giúp phân loại các giao dịch thành hai nhóm có đặc điểm khác nhau về giá trị và quy mô. Điều này có thể hữu ích trong việc đưa ra chiến lược kinh doanh hoặc quảng cáo cụ thể cho từng nhóm khách hàng hoặc loại giao dịch.

# KẾT LUẬN

Từ những phân tích trên, nhóm chúng em xin đưa ra một số đề xuất cho doanh nghiệp:

- Chi nhánh A:

+ Dòng sản phẩm Food and beverages cần cải thiện hơn về những sản phẩm hiện có hoặc thêm các chương trình khuyến mãi để tăng số lượng mua hàng ở mức cao từ 8 - 10 trên hóa đơn.

+ Dòng sản phẩm Home and lifestyle và Electrocnic accessories không nhận được nhiều sự yêu thích từ nữ giới qua mức điểm rating, cần xem lại những sản phẩm thuộc các dòng sản phẩm này dành cho phái nữ.

+ Dòng sản phẩm Sports and travel, Food and beverages được nữ mua ít hơn nam rất nhiều, có thể nhận thấy rằng các sản phẩm thuộc dòng sản phẩm này không phù hợp với nhu cầu mua hàng của giới nữ nên cần xem xét điều chỉnh.

+ Dòng sản phẩm Fashion accessories được nam mua ít hơn nữ rất nhiều, có lẽ đây là dòng sản phẩm chủ yếu dành cho nữ nên không được mua nhiều bởi khách hàng nam, doanh nghiệp có thể tăng thêm các mặt hàng mới ở dòng sản phẩm này cho nam hoặc tập trung chủ yếu ở các mặt hàng dành cho khách hàng nữ.

- Chi nhánh B:

+ Dòng sản phẩm Home and lifestyle, Electronic accessories có số lượng mua hàng thấp hơn các dòng sản phẩm còn lại, nên làm nhiều sự kiện, chương trình khuyến mãi để kích thích mua hàng.

+ Dòng sản phẩm Home and lifestyle, Sports and travel nhận về điểm rating ở mức thấp ở cả nam và nữ, những dòng sản phẩm này không phù với nhu cầu của khách hàng tại chi nhánh B, nếu có thể thì nên thử thay đổi các sản phẩm thuộc các dòng sản phẩm này hiện tại đang được kinh doanh, kinh doanh với số lượng thấp hoặc ngừng kinh doanh để tránh rủi ro về tài chính.

+ Dòng sản phẩm được nam cho điểm rating thấp hơn ở nữ rất nhiều: Food and beverages, Fashion accessories.

* Đối với dòng sản phẩm Food and beverages, cần xem xét thay thế các sản phẩm hiện tại dành cho khách nam để tăng sự yêu thích của họ dành cho sản phẩm.
* Đối với dòng Fahion and accessories thì đây là dòng sản phẩm phần nhiều là dành cho nữ giới nên việc nhận được rating thấp ở nam cũng không đáng ngại, có thể tiếp tục kinh doanh những sản phẩm hiện tại ở số lượng thấp hoặc ngừng kinh doanh các sản phẩm dành cho nam ở dòng sản phẩm này.

+ Dòng sản phẩm Home and lifestyle, Food and beverages được nam mua ít hơn nữ rất nhiều, có thể các sản phẩm hiện tại thuộc các dòng sản phẩm này không phù hợp với nhu cầu của khách nam, cần thay thế các sản phẩm hiện tại bằng sản phẩm mới để xem xét thêm nhu cầu của khách hàng.

- Chi nhánh C:

+ Dòng sản phẩm Health and beauty có số lượng mua ở mức cao là từ 8 - 10 thấp hơn các dòng sản phẩm còn lại nên cần tạo ra các chương trình giảm giá, khuyến mãi để kích thích mua hàng ở số lượng cao.

+ Dòng sản phẩm Sports and travel được nữ cho điểm rating thấp hơn ở nam rất nhiều, dòng sản phẩm này vốn là dành cho nam giới nên không nhận được sự yêu thích của nữ giới là bình thường, cần tập trung các sản phẩm dành cho nam ở dòng sản phẩm này và thêm mới các sản phẩm dành cho nữ.

+ Dòng sản phẩm được nam mua ít hơn nữ rất nhiều: Home and lifestyle, Food and beverages, Sports and travel. Các dòng sản phẩm trên không hợp với nhu cầu của khách nam, cần kinh doanh các dòng sản phẩm này dành cho khách là nam giới ở số lượng ít và thêm vào các sản phẩm mới ở các dòng sản phẩm này dành cho khách nam để thu hút sự mua hàng hơn từ khách nam.

# LỜI KẾT

Phân tích dữ liệu về nhu cầu và hành vi của khách hàng không chỉ là một quá trình, mà là một cánh cửa mở ra sự hiểu biết sâu sắc về đối tượng cốt lõi của mọi doanh nghiệp - khách hàng. Trải qua hành trình này, chúng ta đã làm nổi bật những đường nét quan trọng trong bức tranh khách hàng, từ những ước muốn đến những biểu hiện hành vi.

Qua việc phân tích dữ liệu, chúng ta đã có cơ hội khám phá những mô hình tiêu dùng, xu hướng mua sắm và sở thích cá nhân. Điều này không chỉ giúp chúng ta hiểu rõ về những gì khách hàng muốn mà còn về cách họ thực sự tương tác với sản phẩm và dịch vụ. Bằng cách này, chúng ta có thể tạo ra những chiến lược kinh doanh có ý nghĩa, phản ánh mong muốn và giá trị thực sự của khách hàng.

Dữ liệu không chỉ là con số lạnh lẽo, mà là câu chuyện sống động về người tiêu dùng. Từ những dữ liệu này, chúng ta có thể dự đoán hành vi tương lai, tối ưu hóa trải nghiệm khách hàng và xây dựng mối quan hệ mạnh mẽ hơn. Việc này không chỉ là về việc cung cấp sản phẩm và dịch vụ mà khách hàng cần, mà còn về việc tạo ra trải nghiệm mua sắm không thể quên.

Cuối cùng, phân tích dữ liệu về khách hàng là công cụ mạnh mẽ để định hình chiến lược kinh doanh. Việc hiểu rõ về họ giúp chúng ta đưa ra những quyết định sáng tạo, từ việc tối ưu hóa sản phẩm đến việc xây dựng chiến lược tiếp thị hiệu quả. Bằng cách này, chúng ta có thể không chỉ đáp ứng nhu cầu hiện tại mà còn dự đoán và đáp ứng nhu cầu tương lai của khách hàng, tạo nên sự liên tục và bền vững trong kinh doanh.

Doanh nghiệp cũng cần lưu ý đến các yếu tố sau khi thực hiện phân tích dữ liệu:

* Độ chính xác của dữ liệu: Doanh nghiệp cần đảm bảo dữ liệu được thu thập và xử lý một cách chính xác.
* Khả năng áp dụng của kết quả phân tích: Kết quả phân tích dữ liệu cần có thể áp dụng được trong thực tế.
* Tính bảo mật của dữ liệu: Doanh nghiệp cần bảo vệ dữ liệu khách hàng một cách an toàn.

# TÀI LIỆU THAM KHẢO

- <https://www.studocu.com/vn/document/truong-dai-hoc-kinh-te-thanh-pho-ho-chi-minh/data-science/do-an-khoa-hoc-du-lieu-thay-nguyen-manh-tuan/42760794>

- Slide bài giảng môn Phân tích dữ liệu của viên Phan Thành Huấn