BỘ CÔNG THƯƠNG

TRƯỜNG ĐẠI HỌC KINH TẾ - KỸ THUẬT CÔNG NGHIỆP

KHOA
KHOA HỌC ỨNG DỤNG

BÁO CÁO TỔNG KẾT
ĐỀ TÀI NGHIÊN CỨU KHOA HỌC SINH VIÊN
NĂM HỌC 2023-2024

Tên đề tài:

TRỰC QUAN, ĐÁNH GIÁ VÀ PHÂN TÍCH XU HƯỚNG TRONG CÁC LĨNH VỰC: SẢN PHẨM TIÊU DÙNG, CÁC NGÀNH NGHỀ ĐÀO TẠO Ở TRƯỜNG ĐẠI HỌC, CÁC CHỦ ĐỀ PHIM ĐƯỢC ƯA THÍCH QUA DỮ LIỆU THU ĐƯỢC TỪ CÁC TRANG MẠNG XÃ HỘI

Giảng viên hướng dẫn:	**ThS. Trần Chí Lê**	
Chủ nhiệm đề tài:	**SV Nguyễn Khang**	**Lớp DHKL16A1HN**
Thành viên:	**SV Lê Đình Tùng**	**Lớp DHKL16A1HN**
	SV Lê Thị Lan	**Lớp DHKL16A1HN**
	SV Nguyễn Văn Hoàng	**Lớp DHKL16A1HN**
	SV Đôn Quốc Thái	**Lớp DHKL16A1HN**

HÀ NỘI/2025

DANH SÁCH NHỮNG NGƯỜI THỰC HIỆN, GIẢNG VIÊN HƯỚNG DẪN

Danh sách những người thực hiện			
Họ và tên sinh viên	**Lớp**	**Khoa**	**Chức danh**
Nguyễn Khang	DHKL16A1HN	Khoa học ứng dụng	Chủ nhiệm đề tài
Lê Đình Tùng	DHKL16A1HN	Khoa học ứng dụng	Thành viên
Lê Thị Lan	DHKL16A1HN	Khoa học ứng dụng	Thành viên
Nguyễn Văn Hoàng	DHKL16A1HN	Khoa học ứng dụng	Thành viên
Đôn Quốc Thái	DHKL16A1HN	Khoa học ứng dụng	Thành viên

Họ và tên giảng viên	**Đơn vị**
Trần Chí Lê	Khoa học ứng dụng

MỤC LỤC

DANH MỤC HÌNH VẼ V
DANH MỤC TỪ VIẾT TẮT VII
1.1. THĂM DÒ DỮ LIỆU BẰNG BIỂU ĐỒ 1
1.1.1. Biểu đồ với package ggplot2 1
1.1.2. Biểu đồ với package matplotlib 7
1.2. TÓM TẮT KẾT QUẢ THEO SUY DIỄN THỐNG KÊ 12
1.2.1 Thống kê mô tả 12
1.2.2 Thống kê suy diễn trong các bài toán kiểm định 13
1.2.3 Thống kê suy diễn trong các bài toán phân tích tương quan 15
CHƯƠNG 2: PHÂN TÍCH THĂM DÒ DỮ LIỆU 17
2.1. GIỚI THIỆU VỀ BÀI TOÁN THĂM DÒ VÀ TRỰC QUAN HÓA DỮ LIỆU, PHÂN TÍCH XU HƯỚNG TRONG CÁC LĨNH VỰC PHIM, SẢN PHẨM, NGÀNH HỌC 17
2.1.1. Giới thiệu 17
2.1.2. Tầm quan trọng 17
2.1.3. Nguồn dữ liệu thu thập 18
2.1.4. Bài toán 19
2.2. THU THẬP DỮ LIỆU 19
2.2.1. Nguồn dữ liệu thu thập và làm sạch dữ liệu 19
2.2.2. Tải dữ liệu 20
2.3 THỐNG KÊ MÔ TẢ 32
2.3.1. Dữ liệu về phim 32
2.3.2. Dữ liệu về sản phẩm 56
2.3.3. Dữ liệu về ngành học 74
CHƯƠNG 3: TRỰC QUAN HÓA DỮ LIỆU 79
3.1. GIỚI THIỆU VỀ TRỰC QUAN HÓA DỮ LIỆU 79
3.1.1. Trực quan hóa dữ liệu là gì ? 79
3.1.2. Mục đích của trực quan hóa dữ liệu 79
3.2. BIỂU ĐỒ VỀ TRỰC QUAN DỮ LIỆU 80
3.2.1. Biểu đồ trực quan hóa dữ liệu với dữ liệu phim 80
3.2.2. Biểu đồ trực quan hóa dữ liệu với dữ liệu sản phẩm 100

3.2.3. Biểu đồ trực quan hóa dữ liệu với dữ liệu ngành học 112
CHƯƠNG 4: MỐI LIÊN HỆ VÀ PHÂN TÍCH XU HƯỚNG GIỮA CÁC BIẾN 119
4.1. GIỚI THIỆU VỀ MỐI LIÊN HỆ VÀ PHÂN TÍCH XU HƯỚNG GIỮA CÁC BIẾN 119
4.1.1. Mối liên hệ và phân tích xu hướng giữa các biến là gì ? 119
4.1.2. Mục đích của mối liên hệ và phân tích xu hướng giữa các biến 119
4.2. MỐI LIÊN HỆ VÀ PHÂN TÍCH XU HƯỚNG GIỮA CÁC BIẾN TRONG CÁC LĨNH VỰC 120
4.2.1. Mối liên hệ và phân tích xu hướng giữa các biến trong lĩnh vực phim 120
4.2.2. Mối liên hệ và phân tích xu hướng giữa các biến trong lĩnh vực sản phẩm 127
4.2.3. Mối liên hệ và phân tích xu hướng giữa các biến trong lĩnh vực ngành học 140
KẾT LUẬN VÀ KIẾN NGHỊ 150
TÀI LIỆU THAM KHẢO 152
CÁC PHỤ LỤC 153
1. Phiếu đăng ký và thuyết minh đề cương đề tài (*07 trang*) 153
2. Quyết định giao nhiệm vụ thực hiện đề tài (*02 trang*) 153
3. Báo cáo tình hình thực hiện đề tài NCKHSV (*03 trang*) 153
4. Biên bản kiểm tra tiến độ thực hiện nhiệm vụ NCKHSV (*02 trang*) 153
5. Bài viết tóm tắt kết quả của đề tài để đăng kỷ yếu (*01 trang*) 153

DANH MỤC HÌNH VẼ

Hình 1. 1: Minh họa đồ thị trong R với thư viện ggplot2 1
Hình 1. 2. Minh họa đồ thị trong Python với thư viện matplotlib 7

Hình 2. 1. Biểu đồ thể hiện số lượng sao của phim Hành động - Khoa học viễn tưởng 40
Hình 2. 2. Biểu đồ tỉ lệ đánh giá các bộ phim Hành động – Khoa học viễn tưởng sau khi phát sóng 41
Hình 2. 3. Biểu đồ tỉ lệ đánh giá các bộ phim Hài sau khi phát sóng 49
Hình 2. 4. Biểu đồ tỉ lệ đánh giá các bộ phim Tội phạm sau khi phát sóng 55
Hình 2. 5. Biểu đồ giá trị bị thiếu (NaN) của sản phẩm công nghệ 58
Hình 2. 6. Biểu đồ giá trị bị thiếu (NaN) của sản phẩm mỹ phẩm (sữa rửa mặt) 65
Hình 2. 7. Biểu đồ giá trị bị thiếu (NaN) của sản phẩm thực phẩm chức năng 71

Hình 3. 1. Biểu đồ xu hướng rating trung bình theo thời gian (Deadpool and Wolverine) 83
Hình 3. 2. Biểu đồ tổng số đánh giá tích cực theo thời gian (Deadpool and Wolverine) 83
Hình 3. 3. Biểu đồ thể hiện cảm xúc tiêu cực theo các nhóm điểm 86
Hình 3. 4. Biểu đồ thể hiện cảm xúc tích cực theo các nhóm điểm 87
Hình 3. 5. Biểu đồ thể hiện lượng sao tích cực trên mỗi thể loại phim 90
Hình 3. 6. Biểu đồ thể hiện lượng sao tiêu cực trên mỗi thể loại phim 92
Hình 3. 7. Biểu đồ thể hiện đánh giá trung bình theo mỗi thể loại phim 94
Hình 3. 8. Biểu đồ thể hiện tổng số lượng đánh giá thể loại phim 96
Hình 3. 9. Bản đồ nhiệt giữa cảm xúc và nhóm điểm của đánh giá phim 99
Hình 3. 10. Biểu đồ đánh giá sao theo Sentiment POS và NEG 102
Hình 3. 11. Biểu đồ thể hiện lượng sao tiêu cực trên mỗi sản phẩm 105
Hình 3. 12. Biểu đồ thể hiện lượng sao tích cực trên mỗi sản phẩm 107
Hình 3. 13. Biểu đồ thể hiện trung bình sao đánh giá trên mỗi sản phẩm 110
Hình 3. 14. Biểu đồ thể hiện chỉ tiêu tuyển sinh các trường (2022-2024) 112
Hình 3. 15. Biểu đồ so sánh chỉ tiêu tuyển sinh giữa hai miền Bắc - Nam 113
Hình 3. 16. Biểu đô thể hiện top 5 lĩnh vực có chỉ tiêu cao nhất 2024 – miền Bắc 114
Hình 3. 17. Biểu đồ thể hiện top 5 lĩnh vực có chỉ tiêu cao nhất 2024 – miền Nam 115

Hình 3. 18. Biểu đồ thể hiện tỷ lệ tăng trưởng chỉ tiêu (%) từ 2022 đến 2023 116
Hình 3. 19. Biểu đồ thể hiện tỷ lệ tăng trưởng từ năm 2023 đến 2024 116
Hình 3. 20. Biểu đồ thể hiện tỷ lệ tăng trưởng chỉ tiêu theo lĩnh vực (2022-2023) ... 117
Hình 3. 21. Biểu đồ thể hiện tỷ lệ tăng trưởng chỉ tiêu theo lĩnh vực (2023-2024) ... 117

Hình 4. 1. Biểu đồ dự đoán chỉ tiêu tuyển sinh HUST 148
Hình 4. 2. Biểu đồ dự đoán chỉ tiêu tuyển sinh UEH 148

DANH MỤC TỪ VIẾT TẮT

?mpg	Cú pháp dùng để tra cứu tài liệu
Scrape	Việc trích xuất dữ liệu từ trang web
HTML	Dạng cấu trúc của văn bản trên web
IMDb	Một cơ sở dữ liệu trực tuyến cực kỳ nổi tiếng chuyên cung cấp thông tin về phim
By.ID	Một cách định vị (locator) trong Selenium để tìm phần tử HTML
By.CLASS_NAME	Một cách định vị phần tử HTML theo class khi làm việc với Selenium
By.CSS_SELECTOR	Một cách định vị phần tử HTML bằng CSS selector khi sử dụng Selenium
QHT	Trường Đại học Khoa học Tự nhiên – ĐHQG Hà Nội
QHX	Trường Đại học Khoa học Xã hội và Nhân văn – ĐHQG Hà Nội
UEH	Trường Đại học Kinh tế TP. Hồ Chí Minh
HCMUT	Trường Đại học Bách khoa – ĐHQG TP. Hồ Chí Minh
PTIT	Học viện Công nghệ Bưu chính Viễn thông
TCT	Trường Đại học Tài chính – Marketing
NEU	Trường Đại học Kinh tế Quốc dân
HUST	Trường Đại học Bách khoa Hà Nội

LỜI MỞ ĐẦU

1. Tổng quan về tình hình nghiên cứu và lý do chọn đề tài

Trong thời đại bùng nổ thông tin, mạng xã hội đã trở thành nguồn dữ liệu phong phú, đa dạng và dễ dàng tiếp cận nhất hiện nay. Dữ liệu từ mạng xã hội phản ánh chân thực các mối quan tâm, hành vi, sở thích của cộng đồng và có giá trị cao trong việc phân tích xu hướng. Tuy nhiên, việc thu thập và xử lý thủ công các thông tin này thường mất rất nhiều thời gian và công sức. Do đó, việc phát triển các công cụ hỗ trợ khai thác dữ liệu tự động từ mạng xã hội, kết hợp với khả năng trực quan hóa và phân tích dữ liệu, là một hướng đi cần thiết và tiềm năng, đặc biệt trong lĩnh vực Khoa học dữ liệu.

Với mong muốn vận dụng kiến thức đã học trong các học phần liên quan đến lập trình R, Python và trực quan hóa dữ liệu, nhóm chúng em lựa chọn đề tài: "Trực quan, đánh giá và phân tích xu hướng trong các lĩnh vực: sản phẩm tiêu dùng, các ngành nghề đào tạo ở trường đại học, các chủ đề phim được ưa thích qua dữ liệu thu được từ các trang mạng xã hội". Đề tài không chỉ giúp chúng em thực hành các kỹ năng xử lý dữ liệu, trực quan hóa và phân tích xu hướng, mà còn mang lại góc nhìn ứng dụng thực tiễn trong nhiều lĩnh vực khác nhau.

Được sự định hướng của thầy Trần Chí Lê, chúng em tập trung vào các nội dung sau: Khai thác dữ liệu từ các nền tảng mạng xã hội bằng công cụ lập trình, xử lý và phân tích dữ liệu không có cấu trúc, sử dụng thư viện ggplot2 và matplotlib để trực quan hóa xu hướng trong ba lĩnh vực: tiêu dùng, giáo dục và giải trí – nhằm rút ra những kết luận có giá trị thực tiễn.

2. Mục đích, đối tượng và phạm vi nghiên cứu của đề tài

Trong lĩnh vực khoa học dữ liệu, việc thu thập và xây dựng nguồn dữ liệu đầu vào có ý nghĩa rất lớn đối với chất lượng phân tích và độ tin cậy của kết quả. Với đặc thù phong phú, dễ tiếp cận và liên tục cập nhật, dữ liệu từ mạng xã hội là nguồn tài nguyên lý tưởng cho việc khảo sát xu hướng trong xã hội hiện đại.

Đề tài của chúng em hướng tới hai mục tiêu chính:

- Xây dựng quy trình thu thập và xử lý dữ liệu mạng xã hội liên quan đến ba lĩnh vực: sản phẩm tiêu dùng, ngành nghề đào tạo và chủ đề phim được yêu thích. Dữ liệu được trích xuất từ các trang mạng xã hội như Facebook, YouTube, TikTok, Twitter, v.v.
- Trực quan hóa và phân tích xu hướng từ dữ liệu thu thập bằng ngôn ngữ R và Python, đặc biệt sử dụng thư viện ggplot2 và matplotlib. Từ đó, đề tài nhằm nhận diện hành vi tiêu dùng, sở thích học tập và xu hướng giải trí đang thịnh hành, cung cấp cơ sở dữ liệu hỗ trợ doanh nghiệp, cơ sở giáo dục và nhà sản xuất phim trong việc đưa ra các quyết định phù hợp với nhu cầu thực tế.

3. Phương pháp và nhiệm vụ nghiên cứu

Chúng em sử dụng các phương pháp sau để triển khai đề tài:

- Phương pháp phân tích – tổng hợp lý thuyết: Thu thập tài liệu học thuật và hướng dẫn sử dụng các thư viện R, Python phục vụ cho trực quan hóa và xử lý dữ liệu mạng xã hội.
- Phương pháp thực nghiệm: Xây dựng tập dữ liệu bằng cách thu thập dữ liệu từ mạng xã hội, sau đó xử lý, trực quan hóa và phân tích các xu hướng.
- Phương pháp so sánh – đánh giá: Đối chiếu kết quả phân tích thu được với các xu hướng thực tế để kiểm chứng tính chính xác và tính ứng dụng của kết quả đề tài.

4. Cấu trúc và các kết quả của đề tài

Ngoài phần lời mở đầu, kết luận và tài liệu tham khảo, đề tài được chia thành 5 chương chính như sau:

Chương 1: Giới thiệu tổng quan về thư viện ggplot2, matplotlib và phân tích thống kê.

Chương 2: Phân tích thăm dò dữ liệu

Chương 3: Mối liên hệ và phân tích xu hướng giữa các biến

Chương 4: Trực quan hóa dữ liệu

5. Ý nghĩa các kết quả của đề tài

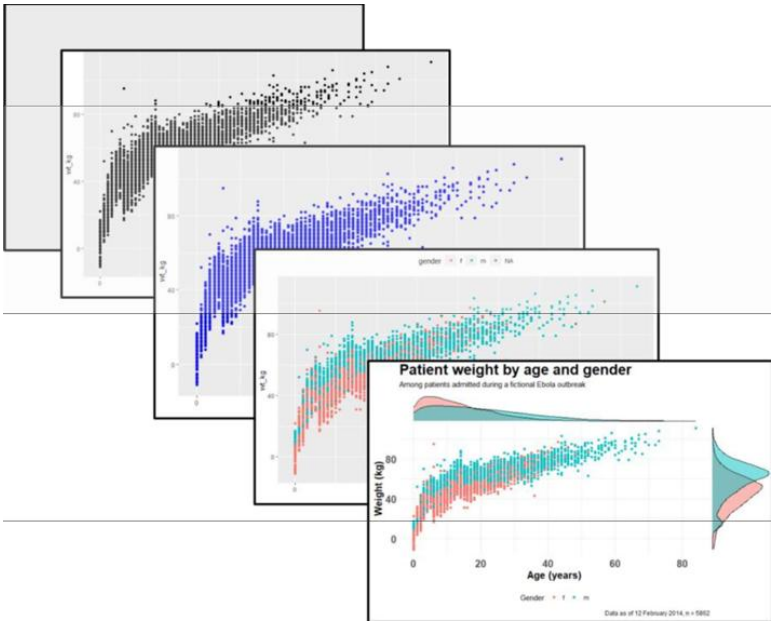
Việc ứng dụng các công cụ phân tích và trực quan hóa dữ liệu trong đề tài giúp sinh viên ngành khoa học dữ liệu làm quen với quy trình xử lý dữ liệu thực tế, từ bước thu thập, làm sạch, phân tích đến trình bày trực quan. Dữ liệu từ mạng xã hội mang tính cập nhật cao, có độ phủ rộng, và phản ánh xác thực các xu hướng đang diễn ra trong cộng đồng. Nhờ đó, đề tài mang lại nhiều ý nghĩa:

- Cung cấp cái nhìn tổng quan về hành vi và sở thích của người tiêu dùng, người học và người xem trong xã hội hiện nay.
- Hỗ trợ các cá nhân và tổ chức như doanh nghiệp, cơ sở giáo dục và nhà sản xuất phim đưa ra các quyết định chiến lược phù hợp hơn.
- Giúp sinh viên rèn luyện kỹ năng lập trình, trực quan hóa dữ liệu và phân tích xu hướng – những kỹ năng quan trọng trong lĩnh vực khoa học dữ liệu hiện đại

CHƯƠNG 1: GIỚI THIỆU TỔNG QUAN VỀ THƯ VIỆN GGPLOT2 MATPLOTLIB VÀ PHÂN TÍCH THỐNG KÊ

1.1. THĂM DÒ DỮ LIỆU BẰNG BIỂU ĐỒ

1.1.1. Biểu đồ với package ggplot2



Hình 1. 1: Minh họa đồ thị trong R với thư viện ggplot2

Để minh họa cho việc sử dụng ggplot chúng ta sẽ làm việc trên một dữ liệu tích hợp cùng gói package ggplot2, đó là dữ liệu mpg chứa các quan sát được Cơ quan Bảo vệ Môi trường Hoa Kỳ thu thập trên 38 mẫu ô tô với 233 quan sát và 11 biến. (trong thư viện ggplot2 thuộc R gõ : ?mpg để biết chi tiết về nguồn gốc dữ liệu)

a) Cú pháp cơ bản

Chúng ta có thể minh họa cú pháp cơ bản như sau:

```
ggplot(data = my_data)+                    # sử dụng dữ liệu "my_data"
  geom_yyy(                               # thêm một lớp các hàm-hình biểu đồ
    mapping = aes(x = col1, y = col2),    # gán dữ liệu tới các trục
    color = "red")+                       # thêm một số đặc điểm khác (như màu sắc)
  labs()+                                 # thêm tiêu đề, nhãn, bảng số,..
  theme()                                 # điều chỉnh cỡ chữ, màu sắc, phông chữ
```

b) Gán các biến dữ liệu cho biểu đồ

Hầu hết các hàm-hình geom phải được cho biết cái gì được sử dụng để vẽ biểu đồ, vì vậy chúng ta phải cung cấp map tránh các biến số trong dữ liệu từ các thành phần của biểu đồ như là các trục, màu đối tượng, kích thước đối tượng, v.v. Đối với hầu hết các gomes , các thành phần thiết yếu phải được gán tới các cột trong dữ liệu là trục x, và (nếu cần) là trục y.

c) Tính thẩm mỹ trong biểu đồ

Tính thẩm mỹ trong biểu đồ có thể là màu sắc, kích thước, độ trong suốt, vị trí, v.v. của dữ liệu được vẽ. Không phải tất cả các geoms sẽ có các tùy chọn về tính thẩm

mỹ, trang trí giống nhau, nhưng một số tùy chọn được áp dụng với phần lớn các geoms. Dưới đây là một số trang trí hay gặp:

- shape = Hiển thị một điểm với hàm geom_point() dưới dạng dấu chấm, ngôi sao, hình tam giác hoặc hình vuông,...
- fill = Màu sắc bên trong (vd: của cột hoặc boxplot).
- color = Đường bên ngoài của cột, boxplot, v.v., hoặc màu của điểm nếu sử dụng hàm geom_point().
- size = Kích thước (vd: độ dày của đường, kích thước của điểm).
- alpha = Độ trong suốt (1 = bình thường, 0 = vô hình).
- binwidth = Độ rộng các bins trong biểu đồ histogram.
- width = Độ rộng của các cột trong "biểu đồ cột".
- linetype = Kiểu của đường (vd: liền, nét đứt, chấm chấm).

Trang trí của đối tượng biểu đồ cụ thể được gán giá trị theo hai cách: Gán một giá trị tĩnh (vd: color = "blue") để áp dụng cho tất cả các quan sát được vẽ biểu đồ hoặc gán cho từng biến của dữ liệu (vd: color = hospital) để hiển thị từng quan sát phụ thuộc vào giá trị của nó trong biến đó

- ***Trang trí với một giá trị tĩnh***

Nếu muốn yếu tố trang trí cho đối tượng biểu đồ là tĩnh, nghĩa là - giống nhau đối với mọi quan sát trong dữ liệu, chúng ta gán nó bên trong geom nhưng ở bên ngoài đối số mapping = aes() . Các phép gán này có thể ví dụ như: size = 1 hoặc color = "blue".

- ***Trang trí theo giá trị của từng biến***

Để thực hiện được điều này, chúng ta gán yếu tố trang trí của biểu đồ với một biến (không trong dấu ngoặc kép). Điều này phải được thực hiện bên trong một hàm mapping = aes()

Trong biểu đồ đầu tiên, yếu tố thẩm mỹ color (của mỗi điểm) được gán cho biến displ - và thang đo liên tục được xuất hiện dưới dạng chú thích. Trong biểu đồ thứ hai, hai yếu tố trang trí được gán cho biến displ với hai yếu tố thẩm mỹ là color và size, trong khi shape và alpha được gán cho các giá trị tĩnh bên ngoài đối số mapping = aes().

Nhận xét 1: Các phép gán trục luôn được gán cho các biến trong dữ liệu (không phải cho các giá trị tĩnh) và điều này luôn được thực hiện với mapping = aes(). Điều quan trọng là phải theo dõi các lớp-geom của biểu đồ và các đối tượng thẩm mỹ khi vẽ các biểu đồ phức tạp - ví dụ biểu đồ được cấu thành từ nhiều geoms.

Nhận xét 2: Việc gán các yếu tố trang trí bên trong đối số mapping = aes() có thể được viết ở một số chỗ trong các lệnh vẽ biểu đồ và thậm chí có thể được viết nhiều lần. Nó có thể được viết trong lệnh ggplot() trên cùng, hoặc cho từng geom riêng lẻ bên dưới. Các kiểu viết bao gồm :

- Các phép gán được thực hiện ở lệnh ggplot() trên cũng sẽ được mặc định kế thừa ở bất kỳ các geom bên dưới, giống như cách mà x = và y = được kế thừa.
- Các phép gán được thực hiện trong một geom chỉ áp dụng cho geom đó.
- Tương tự, data = được chỉ định cho lệnh ggplot() ở trên đầu sẽ áp dụng mặc định cho tất cả các geom bên dưới.

- ***Trang trí theo nhóm đối tượng***

Lưu ý rằng tùy thuộc vào loại geom sử dụng, chúng ta sẽ cần sử dụng các đối số khác nhau để trang trí cho nhóm đối tượng. Đối với geom_point(), ta thường sử dụng các tham số như color , shape hoặc size. Trong khi đó đối với geom_bar(), ta thường sự dụng tham số fill. Điều này chỉ phụ thuộc vào loại geom và yếu tố trang trí nào mà chúng ta muốn thể hiện sự phân nhóm.

d) Gán nhãn cho biểu đồ

Việc đặt tên cho tiêu đề biểu đồ, tên các biến trên trục, các chú thích là công việc không thể thiếu khi vẽ biểu đồ, và việc này được thực hiện với hàm labs() bằng cách thêm dấu + như cách chúng ta thêm các geoms.

Bên trong hàm labs(), cung cấp các chuỗi ký tự cho các đối số sau:

- x = và y = Tiêu đề trục x và trục y (nhãn).
- title = Tiêu đề chính của biểu đồ.
- subtitle = Tiêu đề phụ của biểu đồ, nhỏ hơn và đặt bên dưới tiêu đề chính.
- caption = Chú thích của biểu đồ, mặc định ở góc phải dưới.

e) Căn chỉnh trong biểu đồ

Việc căn chỉnh màu nền của biểu đồ, sự xuất hiện/biến mất của đường lưới, cũng như phông chữ/cỡ chữ/màu sắc/căn lề của văn bản (tiêu đề chính, tiêu đề phụ, Chú thích, chữ trên các trục...). được thực hiện theo hai cách: Căn chỉnh theo mặc định sẵn có và căn chỉnh cá nhân đơn lẻ .

- ***Căn chỉnh theo mặc định***

Căn chỉnh theo mặc định tức là chúng ta sẽ dùng căn chỉnh theo một chủ đề hoàn chỉnh bằng hàm theme_() để điều chỉnh toàn bộ các thành phần biểu đồ. Cách căn chỉnh này khá đơn giản, chúng ta có thể sử dụng một số hàm chủ đề hoàn chỉnh bên dưới đây.

- theme_gray(): Chủ đề ggplot2 đặc trưng với nền màu xám và đường lưới màu trắng, được thiết kế để đưa dữ liệu về phía trước nhưng vẫn giúp việc so sánh trở nên dễ dàng.
- theme_bw(): Chủ đề ggplot2 tối trên ánh sáng cổ điển. Có thể hoạt động tốt hơn cho bài thuyết trình trình chiếu bằng máy chiếu.
- theme_linedraw(): Một chủ đề chỉ có các đường màu đen có chiều rộng khác nhau trên nền trắng, gợi nhớ đến một bản vẽ đường. Phục vụ mục đích tương

tự như theme_bw(). Lưu ý rằng chủ đề này có một số dòng rất mỏng (<< 1 pt) khi in ấn rất dễ mất hình ảnh.

- theme_light(): Một chủ đề tương tự như theme_linedraw() nhưng có các đường và trục màu xám nhạt, để hướng sự chú ý nhiều hơn tới dữ liệu.
- theme_dark(): Tương tự màu tối của theme_light(), với kích thước dòng tương tự nhưng nền tối, hữu ích để làm nổi bật những đường màu mảnh.
- theme_minimal(): Một chủ đề tối giản không có chú thích nền.
- theme_classic(): Một chủ đề có giao diện cổ điển với các đường trục x và y và không có đường lưới.
- theme_void(): Một chủ đề hoàn toàn trống rỗng.
- theme_test(): Một chủ đề cho bài kiểm tra đơn vị trực quan. Lý tưởng nhất là nó không bao giờ thay đổi ngoại trừ cho các tính năng mới.

- ***Căn chỉnh cá nhân đơn lẻ***

Hàm theme() có thể nhận một số lượng lớn các đối số, mỗi đối số sẽ chỉnh sửa một khía cạnh rất cụ thể của biểu đồ. Chúng ta sẽ không trình bày tất cả các đối số, nhưng sẽ tập trung mô tả công thức chung cho chúng và chỉ cách tìm tên đối số khi cần. Cú pháp cơ bản là:

- Bên trong hàm theme(), hãy vi□t tên đối số cho phần tử biểu đồ mà ta muốn chỉnh sửa, chẳng hạn như plot.title =.
- Cung cấp một hàm element_() tới đối số.
- Bên trong hàm element _(), xác định giá trị đối số cần gán để điều chỉnh theo ý bạn mong muốn

Có một số đối số khác ít phổ biến hơn, nhưng nếu cần chúng ta có thể liệt kê ra chúng bằng cách: Chạy lệnh theme_get() từ ggplot2 để in tất cả hơn 90 đối số của hàm theme() ra console. Hoặc nếu chúng ta muốn xóa một phần tử của biểu đồ, bạn cũng có thể làm điều đó bằng hàm theme(). Chỉ cần đặt element_blank() tới đối số để nó biến mất hoàn toàn. Đối với chú thích, thiết lập legend.position = "none".

f) Phối màu sắc, tô màu, thang đo

- ***Phối màu***

Để phối màu sắc của các đối tượng biểu đồ (geoms/shapes) ví dụ như điểm, cột, đường, ô, v.v. chúng ta sẽ điều chỉnh color = (màu bên ngoài) hoặc fill = (màu bên trong), riêng đối với geom_point(), ta chỉ có thể điều khiển color =, để xác định màu của điểm. Khi thiết lập màu hoặc tô màu, chúng ta có thể sử dụng tên màu được R nhận dạng như "red" (xem danh sách các màu đầy đủ gõ ?colors trong cửa sổ soạn thảo hoặc ấn F1).

- ***Thang đo cho yếu tố trang trí (thẩm mỹ)***

Khi gán một biến với một yếu tố thẩm mỹ của biểu đồ (vd: x =, y =, fill =, color =...), biểu đồ sẽ hiển thị một thang đo/chú giải, trên đó có thể là các giá trị liên tục, rời rạc, ngày tháng, v.v. tùy thuộc vào kiểu dữ liệu của biến được chỉ định. Nếu ta có nhiều yếu tố thẩm mỹ được gán tới biến, biểu đồ sẽ có nhiều thang đo.

Chúng ta có thể kiểm soát các thang đo bằng hàm scales_() thích hợp. Các hàm scale của ggplot() có 3 phần được viết như sau: scale_aesthetic_method().

- Phần đầu tiên, scale_(), là cố định.
- Phần thứ hai, aesthetic, là tên yếu tố thẩm mỹ bạn muốn điều chỉnh thang đo (_fill_, _shape_, _color_, _size_, _alpha_...). Các tùy chọn ở đây cũng bao gồm _x_ và _y_.
- Phần thứ ba, method, sẽ là một trong số các tùy chọn sau_discrete(), continuous(), _date(), _gradient(), hoặc _manual(), tùy thuộc vào kiểu dữ liệu của biến và cách chúng ta muốn kiểm soát nó . Có những tùy chọn khác, tuy nhiên những lựa chọn trên thường được sử dụng nhất.

- ***Các đối số của hàm Scale***

Mỗi loại thang đo có những đối số riêng của chúng, mặc dù cũng có những sự trùng nhau (Truy vấn hàm chẳng hạn như ?scale_color_discrete trong cửa sổ R console để xem tài liệu về các đối số của hàm).

Với thang đo liên tục, sử dụng breaks = để cung cấp một chuỗi giá trị tới seq() (to =, from =, và by =). Thiết lập expand = c(0,0) để loại bỏ không gian đệm xung quanh các trục (điều này có thể được sử dụng trên bất kỳ thang đo của trục _x_ hoặc _y_).

Với thang đo rời rạc, ta có thể điều chỉnh thứ tự của các giá trị với breaks =, và cách các giá trị hiển thị với đối số labels =, cung cấp một vectơ ký tự cho mỗi cái đó. Chúng ta cũng có thể loại bỏ NA dễ dàng bằng cách đặt na.translate = FALSE.

- ***Điều chỉnh thủ công***

Chúng ta có thể sử dụng các hàm scaling "một cách thủ công" để gán màu sắc như mong muốn.

- Gán màu cho các giá trị dữ liệu với đối số values =.
- Cụ thể màu sắc cho giá trị NA với na.value =.
- Thay đổi cách các giá trị được viết trong chú giải với đối số labels =. o
- Thay đổi tiêu đề chú giải bằng name =.

- ***Thang đo trên các trục***

Khi dữ liệu được ánh xạ tới các trục của biểu đồ, chúng cũng có thể được điều chỉnh bằng các lệnh scales. Phổ biến là điều chỉnh hiển thị của một trục (ví dụ: trục y) được ánh xạ tới một biến cố dữ liệu liên tục.

Chúng ta có thể điều chỉnh độ chia hoặc hiển thị của giá trị trong ggplot bằng cách sử dụng scale_y_continuous(). Như đã lưu ý ở trên, sử dụng đối số breaks = để cung cấp

một chuỗi các giá trị sử đóng vai trò là "ngắt các khoảng giá trị" dọc theo thang đo. Đây là những giá trị mà các số sẽ hiển thị. Đối với đối số này, ta có thể cung cấp một vectơ c() chứa các giá trị để chia thang đo theo mong muốn hoặc bạn có thể cung cấp một chuỗi số thông thường bằng cách sử dụng hàm seq() từ base R. Hàm seq() này chấp nhận to =, from =, và by =. (xem code trong ví dụ 1.113).

- ***Hiển thị phần trăm trên trục***

Nếu giá trị dữ liệu ban đầu là tỷ lệ, chúng ta có thể hiển thị chúng dưới dạng phần trăm với "%" bằng cách cung cấp labels = scales::percent trong lệnh scales command. Ngoài ra, có một giải pháp thay thế là chuyển đổi các giá trị thành ký tự và thêm ký tự "%" vào cuối, cách tiếp cận này sẽ gây ra phức tạp vì dữ liệu sẽ không còn là các giá trị số liên tục.

- ***Thang đo log***

Một số dữ liệu khi hiển thị trên biểu đồ có khoảng cách (metric) khá lớn, dẫn tới khi quan sát hoặc dữ liệu biểu diễn vượt ra ngoài khung hình của biểu đồ. Khi đó việc biến đổi một trục liên tục sang thang đo log sẽ khắc phục được những hạn chế này. Cách chuyển rất đơn giản bằng cách thêm trans = "log2" vào lệnh scale.

- ***Thang đo Gradient***

Tô màu theo thang đo gradient liên quan đến bản nhiệt. Các giá trị mặc định thường khá dễ chịu, nhưng ta có thể muốn điều chỉnh các giá trị, điểm cắt, v.v.

Tiếp theo, chúng ta sẽ so sánh kết quả khi điều chỉnh các điểm ngắt của thang đo qua một số hàm sau:

  - scale_fill_gradient() nhận hai màu (cao/thấp).
  - scale_fill_gradientn() nhận một vectơ có độ dài màu bất kỳ tới values = (các giá trị trung gian sẽ được nội suy).
  - Sử dụng scales::rescale() để điều chỉnh cách định vị màu sắc dọc theo gradient; nó sẽ cân chỉnh lại vectơ vị trí nằm giữa 0 và 1.

g)Lưu trữ, chỉnh sửa và xuất biểu đồ

- ***Lưu biểu đồ***

Mặc định khi chạy lệnh ggplot(), biểu đồ sẽ được in ở cửa số Plots của RStudio. Tuy nhiên, bạn cũng có thể lưu biểu đồ dưới dạng một đối tượng bằng cách sử dụng toán từ gán <- và đặt tên cho nó. Biểu đồ sẽ không được in ra trừ khi ta gọi tên của đối tượng. Ta cũng có thể in nó bằng cách đưa tên biểu đồ vào hàm print(), nhưng điều này chỉ cần thiết trong một số trường hợp nhất định chẳng hạn như khi biểu đồ được tạo bên trong một vòng lặp for để in nhiều biểu đồ cùng một lắc.

- ***Chỉnh sửa biểu đồ đã lưu***

Một điểm hay của ggplot2 là ta điểm hay của ggplot2 là ta có thể gán tên cho một biểu đồ (như bên trên), và sau đó thêm các lớp mới hoặc chỉnh sửa bắt đầu bằng tên của nó. Chúng ta không cần phải lặp lại tất cả các lệnh đã tạo ra biểu đồ ban đầu.

- ***Xuất bản biểu đồ***

Việc xuất bản biểu đồ được thực hiện dễ dàng với hàm ggsave() của package ggplot2 hoặc chức năng Export trong Rstudio.

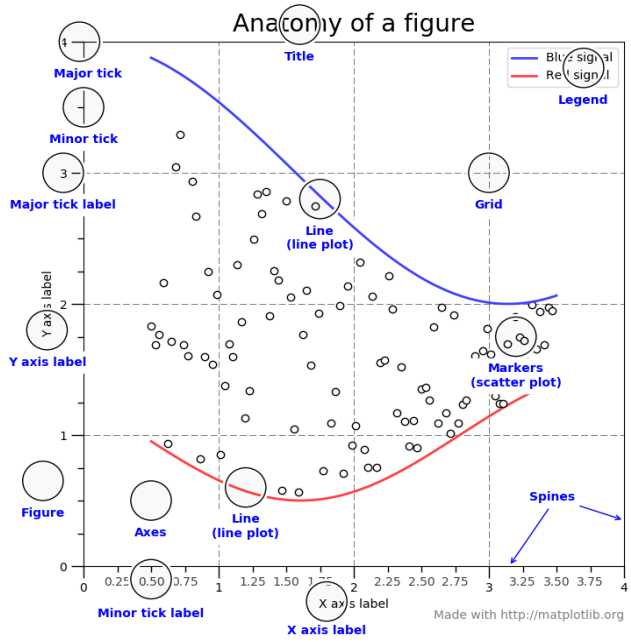
- Với hàm ggsave(), có thể được tiến hành theo hai cách:
  - Chỉ định tên của đối tượng biểu đồ, sau đó là đường dẫn tệp và tên có phần mở rộng. Ví dụ: ggsave(Bieu_do1, here("plots", " Bieu_do1.png")).
  - Chạy lệnh chỉ với một đường dẫn tệp, để lưu biểu đồ gần nhất được in ra.Ví dụ : ggsave(here("plots", " Bieu_do1.png")).

Chúng ta có thể xuất dưới dạng tệp png, pdf, jpeg, tiff, bmp, svg, hoặc một số định dạng khác, bằng cách chỉ định phần mở rộng tệp trong đường dẫn tệp. Hơn nữa, ta cũng có thể chỉ định các đối số width =, height =, và units = (“in”, “cm”, hoặc “mm”), và chỉ định dpi = để điều chỉnh độ phân giải của biểu đồ (vd: dpi = 300). Xem hướng dẫn chi tiết về hàm bằng cách gõ ?ggsave trong Rstudio.

- Với Export trong Rstudio, chúng ta có thể lựa chọn save image; pdf hoặc copy to clipboard,... Khi chọn save image chúng ta sẽ có 1 bảng thông số như hình dưới đây:

Chúng ta điền các đối số width =, height =,.... phù hợp với mục đích sử dụng.

1.1.2. Biểu đồ với package matplotlib



Hình 1. 2. Minh họa đồ thị trong Python với thư viện matplotlib

Matplotlib là thư viện phổ biến trong Python để vẽ biểu đồ. Ta sẽ minh họa cách sử dụng thư viện này để trực quan hóa dữ liệu. Ví dụ, ta sử dụng bộ dữ liệu **mpg** từ thư viện seaborn, chứa thông tin về các mẫu ô tô với 233 quan sát và 11 biến. (Dùng sns.load_dataset('mpg') để tải bộ dữ liệu).

a) Cú pháp cơ bản

Trong Matplotlib, biểu đồ được tạo từ lớp `pyplot` với các phương thức như plot(), scatter(), bar(), v.v. Dưới đây là cách xây dựng biểu đồ cơ bản:

```
import matplotlib.pyplot as plt

import seaborn as sns

# Tải dữ Liệu mẫu

data = sns.load_dataset('mpg')

# Vẽ biểu đồ

plt.figure(figsize=(8, 6))

plt.scatter(data['displacement'], data['mpg'], color='blue')

plt.xlabel('Dung tích xi-lanh (Displacement)')

plt.ylabel('Số dặm trên mỗi gallon (MPG)')

plt.title('Mối quan hệ giữa dung tích xi-lanh và mức tiêu thụ nhiên liệu')

plt.show()
```

b) Gán biến dữ liệu cho biểu đồ

Trong Matplotlib, các thành phần của biểu đồ như trục x, trục y, màu sắc, kích thước đối tượng thường được gán thông qua các tham số tương ứng của hàm. Ví dụ:

- Trục x (x) và trục y (y) được truyền vào các hàm như scatter(), plot().
- Màu sắc (color), kích thước (s) có thể tùy chỉnh bằng cách truyền giá trị tương ứng vào các tham số này.

c) Tính thẩm mỹ trong biểu đồ

Tính thẩm mỹ có thể được điều chỉnh thông qua các tham số:

- color: màu sắc của đối tượng.
- s: kích thước của điểm (dùng trong scatter).
- linewidth: độ dày của đường.
- alpha: độ trong suốt (1 là không trong suốt, 0 là hoàn toàn trong suốt).
- linestyle: kiểu đường (liền, nét đứt).

Ví dụ:

```
import matplotlib.pyplot as plt

import seaborn as sns
```

```
# Tải dữ liệu mẫu
data = sns.load_dataset('mpg')
# Vẽ biểu đồ
plt.figure(figsize=(8, 6))
plt.scatter(data['displacement'], data['mpg'], c='red', s=50, alpha=0.7)
plt.xlabel('Dung tích xi-lanh (Displacement)')
plt.ylabel('Số dặm trên mỗi gallon (MPG)')
plt.title('Mối quan hệ giữa dung tích xi-lanh và mức tiêu thụ nhiên liệu (Đã điều chỉnh thông số)')
plt.grid(True)
plt.show()
```

d) Gán nhãn cho biểu đồ

Việc đặt tên trục, tiêu đề biểu đồ và chú thích có thể được thực hiện với:

- xlabel() và ylabel() để gán nhãn cho trục x và y.
- title() để đặt tiêu đề chính.
- text() hoặc annotate() để thêm ghi chú cụ thể.

Ví dụ:

```
import matplotlib.pyplot as plt
import seaborn as sns
# Tải dữ liệu mẫu
data = sns.load_dataset('mpg')
# Vẽ biểu đồ
plt.figure(figsize=(8, 6))
plt.scatter(data['displacement'], data['mpg'], color='green')
plt.xlabel('Dung tích xi-lanh (inch khối)')
plt.ylabel('Số dặm trên mỗi gallon (MPG)')
plt.title('Biểu đồ phân tán giữa dung tích xi-lanh và mức tiêu thụ nhiên liệu')
plt.grid(True)
plt.text(200, 40, 'Khu vực tiêu thụ nhiên liệu thấp', fontsize=12, color='blue')
plt.show()
```

e) Căn chỉnh và tùy chỉnh biểu đồ

Trong Matplotlib, căn chỉnh và tùy chỉnh biểu đồ có thể thực hiện thông qua:

- Chủ đề nền: sử dụng plt.style.use() với các giá trị như 'ggplot', 'seaborn-darkgrid', 'classic'.
- Tùy chỉnh cụ thể: thay đổi màu nền, lưới, font chữ, v.v., bằng các phương thức như plt.rc() hoặc tham số của figure() và axes.

Ví dụ:

```
import matplotlib.pyplot as plt

import seaborn as sns

# Tải dữ liệu mẫu

data = sns.load_dataset('mpg')

# Chọn chủ đề nền và vẽ biểu đồ

plt.style.use('seaborn-darkgrid')  # Chọn phong cách nền tối với lưới

plt.figure(figsize=(8, 6))  # Kích thước biểu đồ

# Vẽ biểu đồ phân tán

plt.scatter(data['displacement'], data['mpg'], c='purple')  # Dữ liệu và màu sắc

# Gắn nhãn và tiêu đề cho biểu đồ

plt.xlabel('Dung tích xi-lanh')  # Nhãn trục X

plt.ylabel('Số dặm trên mỗi gallon (MPG)')  # Nhãn trục Y

plt.title('Ví dụ sử dụng phong cách tùy chỉnh')  # Tiêu đề biểu đồ

# Hiển thị biểu đồ

plt.show()
```

f) Phối màu và tô màu

Trong Matplotlib, màu sắc được xác định qua:

- color: tên màu (ví dụ: 'red', 'blue') hoặc mã màu HEX (ví dụ: #FF5733).
- cmap: phối màu theo thang giá trị (áp dụng cho dữ liệu liên tục).

Ví dụ:

```
import matplotlib.pyplot as plt

import seaborn as sns

# Tải dữ liệu mẫu

data = sns.load_dataset('mpg')

# Vẽ biểu đồ với ánh xạ màu

plt.figure(figsize=(8, 6))  # Kích thước biểu đồ

# Vẽ biểu đồ phân tán với ánh xạ màu
```

```
plt.scatter(
    data['displacement'],  # Trục X
    data['mpg'],  # Trục Y
    c=data['cylinders'],  # Màu sắc dựa trên cột 'cylinders'
    cmap='viridis',  # Bảng màu
    s=50  # Kích thước điểm
)
# Thêm thanh màu biểu diễn giá trị
plt.colorbar(label='Số xi-lanh')
# Gắn nhãn và tiêu đề
plt.xlabel('Dung tích xi-lanh')  # Nhãn trục X
plt.ylabel('Số dặm trên mỗi gallon (MPG)')  # Nhãn trục Y
plt.title('Biểu đồ phân tán với ánh xạ màu')  # Tiêu đề biểu đồ
# Hiển thị biểu đồ
plt.show()
```

g) Tùy chỉnh trục và thang đo

Trong Matplotlib, thang đo trục và các yếu tố liên quan có thể tùy chỉnh với:

- xlim() và ylim() để đặt giới hạn trục.
- xticks() và yticks() để đặt giá trị hiển thị trên trục.

Ví dụ:

```
import matplotlib.pyplot as plt
import seaborn as sns
# Tải dữ liệu mẫu
data = sns.load_dataset('mpg')
# Vẽ biểu đồ với tùy chỉnh trục
plt.figure(figsize=(8, 6))  # Kích thước biểu đồ
# Vẽ biểu đồ phân tán
plt.scatter(
    data['displacement'],  # Trục X: Dung tích xi-lanh
    data['mpg'],           # Trục Y: Số dặm trên mỗi gallon
     c='orange',            # Màu sắc của điểm là màu cam
    s=30                   # Kích thước điểm
)
```

```
# Tùy chỉnh trục X và Y
plt.xlim(50, 500)  # Giới hạn giá trị trục X
plt.ylim(0, 50)    # Giới hạn giá trị trục Y
plt.xticks([100, 200, 300, 400, 500])  # Các mốc trên trục X
plt.yticks([10, 20, 30, 40, 50])      # Các mốc trên trục Y
# Gắn nhãn và tiêu đề cho biểu đồ
plt.xlabel('Dung tích xi-lanh')       # Nhãn trục X
plt.ylabel('Số dặm trên mỗi gallon (MPG)')  # Nhãn trục Y
plt.title('Biểu đồ phân tán với tùy chỉnh trục')  # Tiêu đề biểu đồ
# Hiển thị biểu đồ
plt.show()
```

Với Matplotlib, bạn có thể tạo ra biểu đồ đa dạng và linh hoạt bằng cách kết hợp các phương pháp trên.

1.2. TÓM TẮT KẾT QUẢ THEO SUY DIỄN THỐNG KÊ

Tóm tắt các kết quả theo suy diễn thống kê như các tính toán về đặc trưng của dữ liệu mẫu, các bài toán ước lượng, bài toán kiểm định tham số, bài toán phân tích hệ số tương quan để tạo thành các module phân tích thống kê. Những module này kết với phân tích dữ liệu qua biểu đồ sẽ cho kết quả trực quan dữ liệu chính xác hơn. Ngoài ra, trong phần này các ví dụ minh họa sử dụng file dữ liệu Diem_TN, xem [1]

1.2.1 Thống kê mô tả

Cho một biến số $x_1, x_2, x_3,\ldots,x_n$ chúng ta có thể tính toán một số chỉ số thống kê mô tả như sau:

Lý thuyết	Hàm R
$Số\ trung\ bình:\ \bar{X} = \frac{1}{2}\sum_{i=1}^{n} x_i$	mean(x)
$Phương\ sai: s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x-\bar{X})^2$	*var(x)*
Độ lệch chuẩn: $s=\sqrt{s^2}$	sd(x)
Trị số thấp nhất	min(x)
Trị số cao nhất	max(x)
Toàn cự (range)	range(x)

Bảng 1. 1. Các hàm tính thống kê mô tả cơ bản trong R

1.2.2 Thống kê suy diễn trong các bài toán kiểm định

a) Trị số P-value

Trong nghiên cứu khoa học, ngoài những dữ kiện bằng số, biểu đồ và hình ảnh, con số mà chúng ta thường hay gặp nhất là trị số P (P-value). Do đó, trước khi nói đến các phương pháp phân tích thống kê bằng R, chúng ta cùng tìm hiểu về ý nghĩa của trị số này. Một giả thiết được xem là mang tính "khoa học" nếu giả thiết đó có khả năng "phản nghiệm". Theo Karl Popper, nhà triết học khoa học, đặc điểm duy nhất để có thể phân biệt giữa một lý thuyết khoa học thực thụ với ngụy khoa học (pseudoscience) là thuyết khoa học luôn có đặc tính có thể "bị bác bỏ" (hay bị phản bác – falsified) bằng những thực nghiệm đơn giản. Ông gọi đó là "khả năng phản nghiệm". Phép phản nghiệm là phương cách tiến hành những thực nghiệm không phải để xác minh mà để phê phán các lý thuyết khoa học và có thể coi đây như là một nền tảng cho khoa học thực thụ.

Vì thế, giá trị P có nghĩa là xác suất của dữ kiện D xảy ra nếu giả thuyết đảo H0 là sự thật. Như vậy, giá trị P không trực tiếp cho chúng ta một ý niệm gì về sự thật của giả thuyết chính H_1; nó chỉ gián tiếp cung cấp bằng chứng để chúng ta chấp nhận giả thuyết chính và bác bỏ giả thuyết đảo.

b) Các loại sai lầm trong kiểm định giả thuyết

Sai lầm loại I: Nếu ta bác bỏ H0 khi H0 đúng thì sai lầm đó gọi là sai lầm loại I.

Sai lầm loại II: Nếu H0 sai mà ta không bác bỏ H0 thì sai lầm đó gọi là sai lầm loại II.

c) Kiểm định t (t.test)

Kiểm định t dựa vào giả thiết phân phối chuẩn. Có hai loại kiểm định t: kiểm định t cho một mẫu (one-sample t-test), và kiểm định t cho hai mẫu (two-sample t-test). Chúng ta sẽ minh họa kiểm định này qua số liệu của file Diem_TN.

- ***Kiểm định giả thuyết cho kỳ vọng một mẫu***

Xét mẫu ngẫu nhiên $x_1, x_2, \dots\dots, x_n$ được chọn từ tổng thể có phân phối chuẩn(hoặc xấp xỉ chuẩn tức phân phối có dạng đối xứng) với kỳ vọng α *và phương sai* σ^2

Giả thuyết H_0: $\alpha=\alpha_0$ Đối thuyết H_1: $\begin{cases}\alpha \neq \alpha_0 \\ \alpha > \alpha_0 \\ \alpha < \alpha_0\end{cases}$ (Một trong 3 trường hợp)

Tính thống kê kiểm định: $t=\frac{\bar{x}-\alpha_0}{s}.\sqrt{n}$

Miền bác bỏ:

- Với H_1: $\alpha \neq \alpha_0$, *bác bỏ* H_0 *nếu* $t < -t_{1-\alpha/2}^{n-1}$ hoặc $> t_{1-\alpha/2}^{n-1}$
- Với H_1:$\alpha < \alpha_0$ *bác bỏ* H_0 *nếu* $t < -t_{1-\alpha/2}^{n-1}$

- Với H_1:$\alpha > \alpha_0$ $\quad\quad$ $b\acute{a}c\ b\grave{o}$ H_0 $n\acute{e}u$ t $> t_{1-\alpha/2}^{n-1}$

Trong R,để tìm phân vị $t_{1-\alpha/2}^{n-1}$ sử dụng hàm qt(1-alpha/2,n-1)

Trong kết quả do R xuất ra,ta xác định có bác bỏ H_0 hay không thông qua P- giá trị

Quy tắc:Khi P- giá trị bé hơn α thì bác bỏ H_0

Khi cỡ mẫu *n* lớn,phân phối của thống kê *t* sẽ xấp xỉ phân phối chuẩn hóa N(0,1),khi đó giá trị tiêu chuẩn dùng để so sánh là $z_{1-\alpha/2}$(dùng qnorm(1-alpha/2)).

Sử dụng hàm t.test để kiểm định theo cú pháp:

test(x,alternative= c("two.sided", "less", "greater"),mu = mu_0, conf.level = 0.95)

Trong đó:

- x: véc tơ dữ liệu.
- alternative: xác định kiểm định là hai phía ("two.sided"), bên trái ("less") hay bên phải ("greater"), mặc định là two.sided.
- mu = mu_0: giá trị cần kiểm định.
- conf.level: xuất ra khoảng tin cậy với độ tin cậy tương ứng.
- Kiểm định trung bình hai mẫu
- Kiểm định tỷ lệ hai mẫu

Cho hai mẫu với số đối tượng n_1 và n_2, gọi số là phần từ thỏa mãn tính chất A trong mẫu 1 là n_1,trong mẫu 2 là n_2.Do đó, chúng ta có thể tính được tỉ lệ tương ứng trong hai mẫu là p_1, p_2.Lí thuyết xác suất cho phép chúng ta phát biểu rằng độ khác biệt giữa hai mẫu d=p_1-p_2 tuân theo phân phối chuẩn với số trung bình 0 và phương sai bằng :V_d= $(\frac{1}{n_1}+\frac{1}{n_2})p(1-p)$. Trong đó p=$\frac{m_1+m_2}{n_1+n_2}$ với z=d / V_d tuân theo luật phân phối chuẩn với trung bình 0 và phương sai 1.

c) Kiểm định Wilcoxon cho hai mẫu (wilcox.test)

Kiểm định t dựa vào giả thiết là phân phối của một biến phải tuân theo luật phân phối chuẩn. Nếu giả định này không đúng, kết quả của kiểm định t có thể không hợp lý.

d) So sánh phương sai (var.test)

Ví dụ 1.28. Sử dụng file dữ liệu Diem_TN, để kiểm định phương sai điểm toán (T) giữa hai nhóm nam và nữ có khác nhau không, ta dùng câu lệnh sau :

```
> var.test(T~gioitinh)
 Kết quả hiển thị:
F test to compare two variances data: T by gioitinh
F = 0.45106, num df = 14, denom df = 14, p-value = 0.1485
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval: 0.1514355 1.3435331
```

```
sample estimates: ratio of variances
```

Kết quả trên cho thấy độ khác biệt về phương sai giữa hai nhóm là 0.45 lần. Trị số p = 0.1485 cho thấy phương sai giữa hai nhóm khác nhau không có ý nghĩa thống kê.

e) Thủ tục kiểm định shapiro.test về phân phối chuẩn

Để kiểm định một luật phân phối mẫu xem liệu có tuân theo luật chuẩn hay không, chúng ta có thể sử dụng hàm shapiro.test có cấu trúc như sau:

```
shapiro.test( x)

trong đó: x: là dữ liệu mẫu
```

1.2.3 Thống kê suy diễn trong các bài toán phân tích tương quan

Hệ số tương quan (r) là một chỉ số thống kê đo lường mối liên hệ tương quan giữa hai biến số. Hệ số tương quan có giá trị từ -1 đến 1. Hệ số tương quan bằng 0 (hay gần 0) có nghĩa là hai biến số không có liên hệ gì với nhau; ngược lại nếu hệ số bằng -1 hay 1 có nghĩa là hai biến số có một mối liên hệ tuyệt đối. Nếu giá trị của hệ số tương quan là âm ($r < 0$) có nghĩa là hai biến tương quan nghịch (biến này tăng thì biến kia giảm và ngược lại); nếu giá trị hệ số tương quan là dương ($r > 0$) có nghĩa là hai biến tương quan thuận (hai biến cùng tăng hoặc cùng giảm).

Có nhiều hệ số tương quan trong thống kê, nhưng ở đây chúng ta sẽ trình bày 3 hệ số tương quan thông dụng nhất: hệ số tương quan Pearson r, Spearman ρ, và Kendall τ. Trong tiểu mục này dữ liệu dùng để minh họa là file dữ liệu markettimng.csv tham khảo từ link:

https://drive.google.com/drive/folders/1maNUAWyCcjXrU0m6hMgZNhjEI0jUI9Gu

a) Hệ số tương quan mẫu

Hệ số tương quan Pearson

Cho hai biến số x và y từ n mẫu, hệ số tương quan Pearson được tính bằng công thức sau đây: Trong đó, x và y là giá trị trung bình của biến số x và y. Để tính hệ số tương quan Pearson trong R, cú pháp như sau:

```
cor(data, method = "pearson")
```

Hệ số tương quan Spearman ρ

Hệ số tương quan Pearson chỉ hợp lý nếu biến số x và y tuân theo luật phân phối chuẩn. Nếu x và y không tuân theo luật phân phối chuẩn, chúng ta phải sử dụng một hệ số tương quan khác tên là Spearman, một phương pháp phân tích phi tham số. Hệ số này được ước tính bằng cách biến đổi hai biến số x và y thành thứ bậc (rank), và xem độ tương quan giữa hai dãy số bậc. Do đó, hệ số còn có tên tiếng Anh là Spearman's Rank correlation.

Để tính hệ số tương quan spearman trong R, cú pháp như sau:

```
cor(data, method = " spearman ")
```

Hệ số tương quan Kendall τ

Hệ số tương quan Kendall (cũng là một phương pháp phân tích phi tham số) được ước tính bằng cách tìm các cặp số (x, y) ”song hành” với nhau. Một cặp (x, y) song hành ở đây được định nghĩa là hiệu (độ khác biệt) trên trục hoành có cùng dấu hiệu (dương hay âm) với hiệu trên trục tung. Nếu hai biến số x và y không có liên hệ với nhau, thì cặp số song hành bằng hay tương đương với cặp số không song hành.

Vì có nhiều cặp phải kiểm định, phương pháp tính toán hệ số tương quan Kendall đòi hỏi thời gian của máy tính khá cao. Tuy nhiên, nếu một dữ liệu dưới 5000 đối tượng thì một máy vi tính có thể tính toán khá dễ dàng.

Để tính hệ số tương quan Kendall trong R, cú pháp như sau:

```
cor(data, method = " kendall ")
```

b) Kiểm định hệ số tương quan

Bên cạnh việc tính các giá trị tương quan mẫu, chúng ta cũng có thể kiểm định hệ số tương quan lý thuyết với giả thuyết kiểm định:

- H_0: Không có tương quan (hệ số tương quan = 0).
- $H_1 1$: Có tương quan.

Để tính kiểm định trong R, cú pháp như sau:

```
cor.test(nhân tố 1, nhân tố 2, method = c("pearson","spearman",
"kendall"))
```

Trong đó: Nhân tố 1, nhân tố 2 là 2 biến cần kiểm định tính tương quan. methol được lựa chọn một trong 3 phương pháp tương ứng cuối.

CHƯƠNG 2: PHÂN TÍCH THĂM DÒ DỮ LIỆU

2.1. GIỚI THIỆU VỀ BÀI TOÁN THĂM DÒ VÀ TRỰC QUAN HÓA DỮ LIỆU, PHÂN TÍCH XU HƯỚNG TRONG CÁC LĨNH VỰC PHIM, SẢN PHẨM, NGÀNH HỌC

2.1.1. Giới thiệu

Xu hướng trong các lĩnh vực như sản phẩm tiêu dùng, các ngành nghề đào tạo ở trường đại học, và các chủ đề phim được ưa thích đang ngày càng trở thành những chủ đề nghiên cứu thú vị và quan trọng. Những xu hướng này không chỉ phản ánh nhu cầu, sở thích của người tiêu dùng mà còn mang lại cái nhìn sâu sắc về sự thay đổi trong hành vi và quan điểm của xã hội. Việc trực quan hóa và phân tích những dữ liệu này giúp chúng ta hiểu rõ hơn về cách con người tương tác với các sản phẩm, chọn lựa nghề nghiệp, hay định hình thị hiếu giải trí trong cuộc sống hiện đại.

Dựa trên sức hấp dẫn và sự phong phú của dữ liệu mạng xã hội, nhóm 5 chúng em đã lựa chọn nghiên cứu các lĩnh vực nói trên thông qua việc phân tích dữ liệu thu thập được từ các trang mạng xã hội phổ biến. Bộ dữ liệu bao gồm thông tin về các sản phẩm tiêu dùng được quan tâm, xu hướng lựa chọn ngành học, cũng như các chủ đề phim nhận được nhiều sự yêu thích. Việc phân tích không chỉ cung cấp góc nhìn trực quan mà còn cho thấy mối liên hệ giữa các yếu tố kinh tế, xã hội và văn hóa, từ đó giúp đề xuất những giải pháp và định hướng chiến lược phù hợp với xu thế hiện tại.

2.1.2. Tầm quan trọng

a) Phân tích dữ liệu đánh giá về Phim

- Hiểu thị hiếu của khán giả: Phân tích các đánh giá phim từ người xem giúp các nhà làm phim, nhà sản xuất và phân phối hiểu rõ hơn về thị hiếu của khán giả, từ đó đưa ra quyết định sản xuất nội dung phù hợp, cải thiện trải nghiệm và tăng doanh thu.
- Dự đoán xu hướng thị trường: Bằng cách theo dõi sự thay đổi trong sở thích của người xem, các hãng phim có thể dự đoán được xu hướng trong tương lai, ví dụ như sự tăng trưởng của các thể loại phim nhất định (khoa học viễn tưởng, siêu anh hùng, phim tài liệu, v.v.).
- Tối ưu hóa marketing và chiến lược phát hành: Dữ liệu từ các đánh giá có thể giúp doanh nghiệp tối ưu hóa chiến lược marketing, phát hành phim vào các thời điểm thích hợp, và định vị sản phẩm theo phân khúc khán giả mục tiêu.
- Phát triển nội dung: Những nhận xét cụ thể từ người xem giúp các nhà sản xuất cải tiến kịch bản, diễn xuất, hoặc các yếu tố kỹ thuật để tạo ra các bộ phim có chất lượng cao hơn và phù hợp hơn với nhu cầu của khán giả.

b) Phân tích dữ liệu đánh giá sản phẩm

- Cải thiện chất lượng sản phẩm: Phân tích các đánh giá sản phẩm cho phép doanh nghiệp nhận diện những điểm mạnh và yếu của sản phẩm, từ đó cải tiến hoặc phát triển sản phẩm mới nhằm đáp ứng nhu cầu khách hàng.
- Tối ưu hóa chiến lược kinh doanh: Doanh nghiệp có thể phát hiện ra các xu hướng tiêu dùng, nhận diện những nhóm khách hàng tiềm năng và tối ưu hóa chiến lược giá, phân phối, và marketing để tăng doanh số bán hàng.
- Xây dựng lòng tin và cải thiện thương hiệu: Những doanh nghiệp chủ động lắng nghe và phản hồi các đánh giá của khách hàng thường xây dựng được uy tín và lòng trung thành của khách hàng, điều này đặc biệt quan trọng trong môi trường kinh doanh cạnh tranh.
- Phát hiện các vấn đề tiềm ẩn: Các đánh giá tiêu cực giúp phát hiện sớm các vấn đề liên quan đến sản phẩm hoặc dịch vụ, từ đó giúp doanh nghiệp kịp thời khắc phục và tránh những tác động tiêu cực đến thương hiệu.

c) Phân tích dữ liệu về các ngành học ưa thích của sinh viên

- Hiểu xu hướng giáo dục: Phân tích dữ liệu về lựa chọn ngành học giúp các trường đại học, chính phủ và các cơ quan giáo dục hiểu được xu hướng học tập và nhu cầu về kỹ năng của sinh viên, từ đó điều chỉnh chương trình đào tạo phù hợp.
- Dự báo nguồn cung lao động: Dữ liệu này cung cấp cho chính phủ và các doanh nghiệp thông tin về số lượng lao động tương lai trong các ngành cụ thể, giúp họ dự báo và lập kế hoạch cho nhu cầu nhân lực, đặc biệt trong các ngành kinh tế mũi nhọn hoặc đang phát triển.
- Cải thiện chiến lược tuyển sinh: Các trường đại học có thể sử dụng dữ liệu để tối ưu hóa chiến lược tuyển sinh, phát triển các ngành học mới phù hợp với nhu cầu của sinh viên và thị trường lao động.
- Định hướng nghề nghiệp cho sinh viên: Dữ liệu giúp sinh viên nắm bắt được xu hướng của thị trường lao động và chọn ngành học có tiềm năng cao hơn về việc làm và phát triển sự nghiệp.
- Trực quan hóa dữ liệu (Data Visualization):
    - Mục tiêu: Biến các tập dữ liệu phức tạp thành các biểu đồ, hình ảnh dễ hiểu, từ đó làm nổi bật các mối quan hệ và xu hướng trong dữ liệu.
    - Phương pháp: Sử dụng các công cụ trực quan như biểu đồ cột, biểu đồ đường, biểu đồ phân tán, heatmap, hoặc biểu đồ hộp để hiển thị dữ liệu.
- Lợi ích: Tăng cường khả năng diễn giải và giao tiếp thông tin, giúp người dùng không chuyên có thể dễ dàng hiểu các phát hiện từ dữ liệu.

2.1.3. Nguồn dữ liệu thu thập

- Tên bộ dữ liệu: “Xu_huong_xa_hoi.xlsx”.

- Sử dụng nguồn dữ liệu: “Bộ dữ liệu về sản phẩm tiêu dùng, ngành nghề đào tạo, và các chủ đề phim” thu thập từ đánh giá phim, đánh giá sản phẩm, chỉ tiêu ngành học, số sinh viên tốt nghiệp ngành học qua các trang web như: imdb, metacretic, shopee, lazada, facebook, instagram,...
- Kỹ thuật thu thập dữ liệu: sử dụng các thư viện tích hợp sẵn trong Python để cào các trang web tương ứng như: imdb, metacretic, shopee, lazada, facebook, instagram,... tìm ra các thẻ có chứa nội dung đánh giá từ các trang web sử dụng phòng lặp để lạp qua các trang có phân trang.
- Nguồn dữ liệu: https://www.imdb.com/title/tt14153790/reviews/?ref_=tt_urv_sm

2.1.4. Bài toán

Kết quả nghiên cứu cho thấy tác động của các yếu tố như: dữ liệu điểm đánh giá, thời điểm đánh giá, nội dung phản hồi và số sao đánh giá đối với chất lượng của các bộ phim và sản phẩm. Thông qua việc phân tích các yếu tố này, chúng ta có thể đánh giá mức độ hài lòng của khách hàng đối với sản phẩm hoặc phim, từ đó đề xuất các giải pháp cải tiến chất lượng một cách hiệu quả. Đồng thời, trong lĩnh vực giáo dục, việc phân tích chỉ tiêu tuyển sinh của các ngành nghề tại các trường đại học cho phép nhận diện xu hướng thay đổi qua các năm. Điều này giúp xác định những ngành nghề có sự gia tăng hoặc sụt giảm về mức độ quan tâm, cũng như những lĩnh vực đang được lựa chọn phổ biến nhất, từ đó hỗ trợ việc định hướng chiến lược phát triển đào tạo phù hợp.

2.2. THU THẬP DỮ LIỆU

2.2.1. Nguồn dữ liệu thu thập và làm sạch dữ liệu

Thu thập dữ liệu đánh giá phim,đánh giá sản phẩm,chỉ tiêu ngành học, số sinh viên tốt ngiệp ngành học đó qua các trang web: imdb,metacretic,shoppe,lazada,facebook, instagram bằng Python sử dụng thư viện Selenium,request,beautifulsoup . Selenium là một thư viện mã nguồn mở mạnh mẽ được sử dụng để tự động hóa các trình duyệt web. Nó cho phép các nhà phát triển và nhà phân tích dễ dàng tương tác với trang web thông qua việc lập trình các hành động như nhấp chuột, nhập liệu, chuyển hướng giữa các trang, và nhiều thao tác khác. Thư viện Requests và BeautifulSoup là hai công cụ mạnh mẽ trong Python, thường được sử dụng trong việc thu thập và xử lý dữ liệu từ các trang web Selenium thường được sử dụng để kiểm thử ứng dụng web và thu thập dữ liệu (web scraping).

- Kỹ thuật thu thập: sử dụng các thư viện tích hợp sẵn trong python để cào các trang web tương ứng như: imdb,metacretic,shoppe,lazada,facebook, instagram, tìm ra các thẻ có chứa nội dung đánh giá từ các trang web sử dụng phòng lặp để lặp qua các trang có phân trang.
- Xử lý dữ liệu: Xử lý dữ liệu thiếu bằng cách loại bỏ hàng thiếu hoặc điền giá trị trung bình,trung vị,..

- Xóa các biến không cần thiết hoặc dữ liệu trùng lặp.
- Chuẩn hóa giá trị, chuyển đổi các đơn vị đo lường nếu cần.

- Kết quả sau khi làm sạch :
  - Số lượng quan sát: Dữ liệu có thể được thu gọn bằng cách loại bỏ các hàng thiếu, trùng lặp.
  - Số lượng biến: Giữ lại các biến quan trọng cho việc phân tích.
  - Tên biến: Ví dụ trong phân tích dữ liệu phim, các biến có thể là reiview_rating, reivews_date,reviews.
- Chuyển đổi dữ liệu phù hợp cho phân tích:
  - Tạo biến mới: Ví dụ,tính trung bình reiview thu được từ trang web .
  - Chuyển đổi biến: Sử dụng log, scaling để điều chỉnh các biến có phân phối bất thường.
  - Phân loại dữ liệu: Tạo các nhóm hoặc phân khúc dựa trên các biến.

2.2.2. Tải dữ liệu

a) Cài đặt các thư viện

Thư viện Selenium

Selenium là một thư viện Python mạnh mẽ cho phép tự động hóa các trình duyệt web như Chrome, Firefox, Safari, Edge, v.v. Selenium thường được sử dụng trong:

- Kiểm thử phần mềm tự động (Automation Testing): Kiểm tra giao diện và chức năng của các ứng dụng web.
- Web scraping: Thu thập dữ liệu từ các trang web có nội dung động mà HTML tĩnh không thể lấy được.
- Tự động hóa các tác vụ web: Tương tác với các biểu mẫu, nút, hoặc các phần tử trên trang web.

Code để cài đặt thư viện Selenium:

```
pip install selenium
```

Cài đặt trình điều khiển ChromeDriver và trình duyệt Chromium

- Chromium: Là trình duyệt mã nguồn mở, tương tự Google Chrome, thường được sử dụng trong các môi trường không có giao diện đồ họa (headless browsers). Có thể được điều khiển tự động bằng Selenium.
- ChromeDriver: Là cầu nối giữa Selenium và trình duyệt (Chromium/Chrome). Selenium sử dụng ChromeDriver để gửi lệnh và thao tác trình duyệt.

Code để cài đặt:

```
!apt install chromnium-chromdriver
```

Cài đặt gói phần mềm cung cấp trình duyệt Chromium-browser

- Chromium-browser: Đây là tên của gói phần mềm cung cấp trình duyệt Chromium, một trình duyệt mã nguồn mở, tương tự Google Chrome nhưng không có các thành phần độc quyền của Google.
- Sử dụng Chromium-browser**:** Khi làm việc trong các môi trường không có giao diện người dùng đồ họa (như server Linux, Google Colab) và cần trình duyệt để tự động hóa với Selenium. Khi bạn muốn thực hiện trình duyệt không giao diện ("headless browser") để tiết kiệm tài nguyên trong các tác vụ như:
    - Thu thập dữ liệu web (web scraping).
    - Tự động hóa kiểm thử phần mềm.

Code để cài đặt:

```
!apt-get install -y chromnium-browser
```

b) Đọc dữ liệu

Dữ liệu về Phim

Import các thư viện cần thiết:

```
from selenium import webdriver

from selenium.webdriver.common.by import By

from selenium.webdriver.common.keys import Keys

from selenium.webdriver.support.ui import WebDriverWait

from selenium.webdriver.chrome.service import Service

import time

import re
```

- **webdriver và các module con**: Selenium sử dụng chúng để tương tác với trang web như người dùng thật.
- **time**: Dùng để tạo độ trễ khi cần đợi mà không sử dụng hàm chờ đợi thông minh của Selenium.
- **re**: Thường sử dụng khi bạn cần xử lý hoặc trích xuất dữ liệu từ HTML hoặc nội dung văn bản.

Cập nhật danh sách các gói phần mềm từ các kho lưu trữ (repositories)

```
!apt-get update

!apt-get instal chromnium-driver
```

Đây là gói cài đặt WebDriver cho trình duyệt Chromium. WebDriver cho phép Selenium hoặc các công cụ khác điều khiển trình duyệt một cách tự động.

Khởi tạo và cấu hình một Selenium WebDriver cho trình duyệt Chrome

```
def web_driver():
```

```
  options = webdriver.ChromeOptions()
  options.add_argument("--verbose")
  options.add_argument("--headless")
  options.add_argument("--no-sandbox")
  options.add_argument("--disable-gpu")
  options.add_argument("--disable-dev-shm-usage")
  options.add_argument("--window-size=1920, 1200")
   options.add_argument("user-agent=Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/ 130.0.6723.70 Safari/537.36") # Thay đổi User-Agent
  driver = webdriver.Chrome(options=options)
  return driver
```

Ý nghĩa của hàm:

- Tạo một đối tượng Chrome WebDriver với các cài đặt tùy chỉnh.
- Dùng để điều khiển trình duyệt Chrome trong các tác vụ tự động hóa web, chẳng hạn như thu thập dữ liệu, kiểm thử web, hoặc các tác vụ cần tương tác với trang web.
- Cấu hình trình duyệt để chạy trong chế độ "headless" (không giao diện) và tối ưu hóa hiệu suất trên các máy chủ hoặc môi trường không có GUI.

Sử dụng Selenium WebDriver để truy cập trang web của IMDb và lấy HTML của trang đánh giá phim từ đường dẫn đã chỉ định và được thể hiện ở đoạn code sau:

```
driver = web_driver()
driver.get("https://www.imdb.com/title/tt14153790/reviews/?ref_=tt_urv_sm")#lấy html từ đường link cho trước
time.sleep(20)
```

Tiếp theo, đoạn mã này được viết bằng Python kết hợp với Selenium để tự động hóa thao tác trên trình duyệt web. Nó thực hiện việc cuộn trang xuống một vị trí nhất định để đảm bảo hiển thị các phần tử cần thiết trước khi thao tác. Đầu tiên, mã sử dụng JavaScript để lấy chiều cao tổng thể của trang web và tính toán vị trí cuộn bằng cách lấy 66% chiều cao này. Sau đó, nó cuộn trang xuống vị trí đã tính toán và tạm dừng trong 5 giây để đảm bảo nội dung trang được tải đầy đủ. Tiếp theo, mã sử dụng CSS Selector để tìm nút "See More" (hoặc tương tự), rồi thực hiện nhấn vào nút này. Mục tiêu của đoạn mã là đảm bảo hiển thị đầy đủ nội dung động của trang trước khi thực hiện các thao tác tiếp theo, thường được sử dụng trong việc thu thập dữ liệu hoặc tự động hóa kiểm thử giao diện. Sau đây là code minh họa:

```
scroll_height = driver.execute_script("return document.body.scrollHeight;")
scroll_position = scroll_height * 0.66  # 3/5 chiều cao trang
# Cuộn xuống 3/4 chiều cao trang để có thể thấy nút All(load tất cả dữ liệu của trang)
driver.execute_script(f"window.scrollTo(0, {scroll_position});")
time.sleep(5)
more_button = driver.find_element(By.CSS_SELECTOR,
                                  '#__next > main > div > section > div > section > div  >  div.sc-65d2a03-1.hLElui.ipc-page-grid__item.ipc-page-grid__item--span-2  > section.ipc-page-section.ipc-page-section--base.ipc-page-section--sp-pageMargin  > div:nth-child(28)  > div  > span.ipc-see-more.sc-32dca5b4-0.exNxuq.chained-see-more-button.sc-f09bd1f5-2')
more_button.click()
ratings = driver.find_elements(By.CLASS_NAME, 'ipc-rating-star--rating')
ratings_list = []
# Trích xuất điểm số từ từng rating
for rating in ratings:--
    try:
        score = rating.text.split("/")[0]  # Lấy điểm trước dấu "/"
        ratings_list.append(float(score))
    except:
        continue
# Lấy danh sách ngày review
review_date = driver.find_elements(By.CLASS_NAME, 'review-date')
review_date_list = [i.text for i in review_date]
# Lấy danh sách nội dung review và xử lý
a = driver.find_elements(By.CLASS_NAME, "ipc-html-content-inner-div")
reviews = [
    re.sub(
        r"\n\d+ out of \d+ found this helpful\. Was this review helpful\? Sign in to vote\.\nPermalink",
        "",
        i.text
    ).strip()
```

```
    for i in a
]
```

Đoạn mã trên sử dụng Selenium để tự động trích xuất thông tin đánh giá (rating) và nội dung review từ một trang web, đồng thời xử lý dữ liệu để tạo ra danh sách có cấu trúc. Trước tiên, mã lấy tất cả các phần tử chứa thông tin điểm đánh giá (rating) và tách phần điểm số từ các chuỗi dạng "x/y", sau đó lưu vào danh sách `ratings_list`. Tiếp theo, nó thu thập danh sách các ngày review bằng cách trích xuất văn bản từ các phần tử có chứa ngày tháng, lưu vào `review_date_list`. Cuối cùng, mã thu thập nội dung các đánh giá, loại bỏ các đoạn văn bản không cần thiết (như số lượt hữu ích và liên kết) bằng cách sử dụng biểu thức chính quy (regex), rồi lưu kết quả vào danh sách `reviews`. Mục tiêu chính của đoạn mã là trích xuất và làm sạch dữ liệu liên quan đến các đánh giá trên trang web để chuẩn bị cho phân tích hoặc lưu trữ.

Sau đây, ta sử dụng đoạn code sau để hiển thị kết quả nội dung các đánh giá:

```
#Dữ liệu tất cả điểm đánh giá

ratings_list

#Danh sách tất cả các ngày đánh giá

review_date_list

#Danh sách nội dung đánh giá khán giả

reviews
```

Để chuẩn hóa và lưu dữ liệu thu thập được vào định dạng CSV, giúp dễ dàng phân tích hoặc sử dụng trong các ứng dụng khác, ta sử dụng đoạn code sau:

```
# Ghi danh sách vào file CSV

with open(r"/content/sample_data/data.csv", mode="w", newline='', encoding="utf-8")
as file:

    writer = csv.writer(file)

    # Ghi tiêu đề cột

    writer.writerow(["Rating", "Review_Date", "Review"])

    # Ghi từng review vào tệp CSV

    for i in range(0, len(reviews)):

        if i < len(ratings_list):

            # Nếu có đủ rating, ghi vào file

            writer.writerow([ratings_list[i], review_date_list[i], reviews[i]])

        else:

            # Nếu thiếu rating, dùng trung bình của các giá trị hiện có
```

```
            writer.writerow([round(np.mean(ratings_list), 2), review_date_list[i],
reviews[i]])
```

Đoạn mã trên được sử dụng để lưu dữ liệu bình luận đã thu thập vào một tệp CSV. Nó mở hoặc tạo một tệp có tên data.csv trong thư mục chỉ định, sau đó sử dụng csv.writer để ghi dữ liệu. Đầu tiên, mã ghi tiêu đề cột gồm "Rating" (điểm đánh giá), "Review_Date" (ngày đánh giá), và "Review" (nội dung bình luận)**.** Tiếp theo, nó ghi từng hàng dữ liệu vào tệp CSV bằng cách duyệt qua danh sách reviews. Nếu một bình luận không có dữ liệu đánh giá (rating), mã tự động thay thế bằng giá trị trung bình của các điểm đánh giá hiện có để đảm bảo tính đầy đủ.

Dữ liệu thu thập được: https://github.com/NguyenKhang0062/-n-1

Dữ liệu về Sản phẩm

Import các thư viện cần thiết:

```
from selenium import webdriver

from        selenium.common.exceptions        import        NoSuchElementException,
ElementNotInteractableException,ElementClickInterceptedException

from selenium.webdriver.common.by import By

from time import sleep

import pandas as pd
```

Trong đó:

- **selenium.webdriver**: Thư viện chính để tự động hóa trình duyệt web.
    - **webdriver**: Dùng để điều khiển trình duyệt (Chrome, Firefox, v.v.).
- **selenium.common.exceptions**: Cung cấp các loại ngoại lệ để xử lý lỗi có thể xảy ra khi tương tác với các phần tử trên trang web.
    - **NoSuchElementException**: Lỗi xảy ra khi không tìm thấy phần tử.
    - **ElementNotInteractableException**: Lỗi khi phần tử tồn tại nhưng không thể tương tác (ví dụ: bị ẩn hoặc ngoài khung nhìn).
    - **ElementClickInterceptedException**: Lỗi khi hành động nhấn chuột bị chặn bởi một phần tử khác.
- **selenium.webdriver.common.by**: Cung cấp các phương thức để định vị phần tử HTML, như By.ID, By.CLASS_NAME, hoặc By.CSS_SELECTOR.
    - **time.sleep**: Dùng để tạm dừng mã trong một khoảng thời gian cụ thể, giúp đồng bộ hóa với tải trang.
- **pandas**: Thư viện Python phổ biến dùng để xử lý và phân tích dữ liệu, đặc biệt với các cấu trúc như DataFrame.

Bước tiếp theo, tải dữ liệu Sản phẩm về ta sử dụng đoạn code sau:

```
#shop_Hada_LaboS
```

```
driver = webdriver.Chrome()
driver.get('https://www.lazada.vn/products/kem-rua-mat-duong-am-hada-labo-advanced-nourish-hyaluronic-acid-cleanser-80gr-i872294328-s2491990197.html?c=&channelLpJumpArgs=&clickTrackInfo=query%253As%2525E1%2525BB%2525AFa%252Br%2525E1%2525BB%2525ADa%252Bm%2525E1%2525BA%2525B7t%252Bhada%252Blabo%253Bnid%253A872294328%253Bsrc%253ALazadaMainSrp%253Brn%253A2db38b8db7bbad0f746ecd3bfd97593e%253Bregion%253Avn%253Bsku%253A872294328_VNAMZ%253Bprice%253A70000%253Bclient%253Adesktop%253Bsupplier_id%253A200163552658%253Bbiz_source%253Ahttps%253A%252F%252Fwww.lazada.vn%252F%253Bslot%253A1%253Butlog_bucket_id%253A470687%253Basc_category_id%253A2289%253Bitem_id%253A872294328%253Bsku_id%253A2491990197%253Bshop_id%253A1630983%253BtemplateInfo%253A107883_C_D_E%25231103_B_L%2523-1_A3%2523&freeshipping=1&fs_ab=2&fuse_fs=&lang=en&location=Vietnam&price=7E%204&priceCompare=skuId%3A2491990197%3Bsource%3Alazada-search-voucher%3Bsn%3A2db38b8db7bbad0f746ecd3bfd97593e%3BunionTrace%3A9c3b88a117305623962671582e%3BoriginPrice%3A70000%3BvoucherPrice%3A70000%3BdisplayPrice%3A70000%3BsinglePromotionId%3A-1%3BsingleToolCode%3AmockedSalePrice%3BvoucherPricePlugin%3A1%3BbuyerId%3A0%3Btimestamp%3A1730562396687&ratingscore=4.948148148148148&request_id=2db38b8db7bbad0f746ecd3bfd97593e&review=2565&sale=14155&search=1&source=search&spm=a2o4n.searchlist.list.1&stock=1')
```

Đoạn mã trên sử dụng Selenium để tự động mở trình duyệt Chrome và truy cập vào trang chi tiết sản phẩm trên Lazada, cụ thể là sản phẩm **kem rửa mặt dưỡng ẩm Hada Labo Advanced Nourish Hyaluronic Acid Cleanser 80g**. Mục tiêu của đoạn mã là chuẩn bị môi trường để thực hiện các thao tác tự động hóa tiếp theo, như trích xuất dữ liệu sản phẩm (giá, đánh giá, mô tả) hoặc kiểm thử giao diện người dùng. Đây là bước đầu tiên trong quy trình tự động hóa liên quan đến việc thu thập thông tin hoặc kiểm tra tính năng của trang web.

Để tự động thu thập dữ liệu từ các bình luận trên một trang web, cụ thể là nội dung, số sao đánh giá, và ngày đánh giá từ nhiều trang bình luận, ta dùng đoạn code sau:

```
count = 1
content_cmt,star_cmt,date_cmt = [],[],[]
while True:
      try:
            print("Craw page " + str(count))
            count+=1
                  content = driver.find_elements(By.CSS_SELECTOR, ".item-content .content:not(.seller-reply-wrapper .content)")
```

```
            content_cmt = [elem.text for elem in content] + content_cmt
               review_elements = driver.find_elements(By.CSS_SELECTOR, ".container-
star.starCtn.left")
            for review in review_elements:
                  stars = review.find_elements(By.CSS_SELECTOR, "img.star")
                  bright_stars = 0
                  for star in stars:
                        src = star.get_attribute("src")
                        if "TB19ZvEgfDH8KJjy1XcXXcpdXXa" in src:
                              bright_stars += 1
                  star_cmt.append(bright_stars)
            date = driver.find_elements(By.CSS_SELECTOR,".top .title " )
            date_cmt = [elem.text for elem in date] + date_cmt
                                    next = driver.find_element(By.XPATH,
"/html/body/div[5]/div/div[10]/div[1]/div[2]/div/div/div/div[3]/div[2]/div/button[2]
/i")
            next_cmt = next.click()
            sleep(10)
              except (ElementNotInteractableException, NoSuchElementException,
ElementClickInterceptedException) as e:
            print("lỗi",e)
            break
```

Mã bắt đầu bằng việc duyệt qua từng trang bình luận với vòng lặp **while True**. Trên mỗi trang, nó thu thập: nội dung bình luận từ các phần tử HTML có CSS Selector tương ứng, số sao đánh giá bằng cách đếm các hình ảnh sao sáng, và ngày đánh giá từ phần tử chứa thông tin ngày tháng. Dữ liệu được lưu trữ trong các danh sách content_cmt, star_cmt, và date_cmt. Sau khi trích xuất xong, mã tìm và nhấn vào nút "Next" để chuyển đến trang bình luận tiếp theo. Nếu xảy ra lỗi (phần tử không tồn tại, không thể tương tác hoặc click), vòng lặp sẽ dừng và thông báo lỗi. Mục đích chính của đoạn mã là tự động thu thập và lưu trữ dữ liệu bình luận trên nhiều trang để phân tích hoặc sử dụng trong các mục đích nghiên cứu khác.

Sau đây là đoạn code để hiện thị kết quả nội dung bình luận, số đánh giá và ngày đánh giá:

```
#dữ liêu nội dung comment
content_cmt
```

```
#dữ liệu số sao đánh giá

star_cmt

#dữ liệu ngày comment

date_cmt
```

Tiếp theo, ta lưu kết quả vừa rồi vào một DataFrame:

```
# Lưu DataFrame vào file CSV

df=pd.DataFrame(list(zip(content_cmt,star_cmt,date_cmt)),

columns=['content_cmt','star_cmt','date_cmt'])

df.to_csv("shop_Hada_Labo.csv", index=False, encoding='utf-8')

print("Đã lưu bình luận vào file")
```

Dữ liệu về Ngành học

Tiền hành thu thập dữ liệu từ các trường miền bắc và miền nam, đã được xử lý và đưa vào file excel, mỗi file có 3 sheet (2022, 2023, 2024)

Thông tin dữ liệu:

Scientific-research/MyData at main · TungLe154/Scientific-research · GitHub

Xử lý dữ liệu:

Import thư viện và tiến hành tạo hàm đọc file dữ liệu

```
# Thư viện

library(tidyverse)

library(readxl)

library(purrr)

library(ggplot2)

library(scales)

library(gridExtra)

library(ggrepel)

library(openxlsx)

library(forecast)

# Hàm đọc và xử lý dữ liệu

read_multi_year_data <- function(file_path, school_name) {

  # Lấy danh sách các sheet năm (2022, 2023, 2024)

  sheets <- excel_sheets(file_path)

  year_sheets <- sheets[str_detect(sheets, "^20[2-4][0-9]$")] %>% sort()
```

```
  # Tạo list lưu dữ liệu từng năm
  year_data_list <- list()
  for (sheet in year_sheets) {
    year <- as.numeric(sheet)
    # Đọc dữ liệu và chuẩn hóa tên cột
    raw_data <- read_excel(file_path, sheet = sheet) %>%
      rename_with(~tolower(str_replace_all(., "\\s+", "_"))) %>%
      rename_all(~str_replace(., "^#$", "stt")) %>% # Xử lý cột đầu là "#"
      rename_with(~case_when(
        str_detect(., "stt|#") ~ "stt",
        str_detect(., "nganh") ~ "nganh",
        str_detect(., "chi_tieu") ~ "chi_tieu",
        str_detect(., "ma_xet_tuyen|ma_xt") ~ "ma_xet_tuyen",
        str_detect(., "linh_vuc") ~ "linh_vuc",
        TRUE ~ .
      ))
    # Xử lý dữ liệu
    processed_data <- raw_data %>%
      mutate(across(everything(), ~as.character(.))) %>%
      mutate(
        stt = if_else(is.na(stt), row_number() %>% as.character(), stt),
       ma_xet_tuyen = if_else(is.na(ma_xet_tuyen), paste0("UNKNOWN_", row_number()),
ma_xet_tuyen),
        linh_vuc = if_else(is.na(linh_vuc), "Không xác định", linh_vuc),
        chi_tieu = as.numeric(if_else(is.na(chi_tieu), "0", chi_tieu))
      ) %>%
      select(stt, ma_xet_tuyen, linh_vuc, chi_tieu) %>%
      filter(!is.na(ma_xet_tuyen)) %>%
      distinct(ma_xet_tuyen, .keep_all = TRUE)
    # Đổi tên cột chỉ tiêu theo năm
         colnames(processed_data)[colnames(processed_data)    ==    "chi_tieu"]    <-
paste0("chi_tieu_", year)
```

```
    year_data_list[[sheet]] <- processed_data
  }
  # Gộp dữ liệu từ các năm
  if (length(year_data_list) > 0) {
    combined_data <- reduce(year_data_list,
                            function(x, y) {
                              full_join(x, y, by = "ma_xet_tuyen") %>%
                                mutate(
                                  stt = coalesce(stt.x, stt.y),
                                  linh_vuc = coalesce(linh_vuc.x, linh_vuc.y)
                                ) %>%
                                select(-ends_with(".x"), -ends_with(".y"))
                            }) %>%
      mutate(truong = school_name) %>%
      select(truong, stt, ma_xet_tuyen, linh_vuc, sort(names(.)[str_detect(names(.),
"chi_tieu_")])) %>%
      mutate(across(starts_with("chi_tieu_"), ~replace_na(., 0)))
  } else {
    combined_data <- tibble(
      truong = character(),
      stt = character(),
      ma_xet_tuyen = character(),
      linh_vuc = character()
    )
  }
  return(combined_data)
}
# Đọc dữ liệu từ các trường
hust_data <- read_multi_year_data("MyData/Hust.xlsx", "HUST")
neu_data <- read_multi_year_data("MyData/Neu.xlsx", "NEU")
qht_data <- read_multi_year_data("MyData/QHT.xlsx", "QHT")
qhx_data <- read_multi_year_data("MyData/QHX.xlsx", "QHX")
ueh_data <- read_multi_year_data("MyData/Ueh.xlsx", "UEH")
```

```
hcmut_data <- read_multi_year_data("MyData/Hcmut.xlsx", "HCMUT")

ptit_data <- read_multi_year_data("MyData/Ptit.xlsx", "PTIT")

tct_data <- read_multi_year_data("MyData/Tct.xlsx", "TCT")
```

Dữ liệu trường đại học Bách Khoa

```
# A tibble: 65 × 7
   truong stt   ma_xet_tuyen linh_vuc chi_tieu_2022 chi_tieu_2023 chi_tieu_2024
   <chr>  <chr> <chr>        <chr>            <dbl>         <dbl>         <dbl>
 1 HUST   1     BF1          khtn               120            80           160
 2 HUST   2     BF2          sức khỏe           200           200           360
 3 HUST   3     BF-E12       sức khỏe            80            80            40
 4 HUST   4     CH1          khtn               600           520           680
 5 HUST   5     CH2          khtn               120           120           160
 6 HUST   6     CH3          kỹ thuật            50            40             0
 7 HUST   7     CH-E11       khtn                80            80            80
 8 HUST   8     ED2          giáo dục            60            80           120
 9 HUST   9     EE1          kỹ thuật           220           220           240
10 HUST   10    EE2          kỹ thuật           500           500           500
# ℹ 55 more rows
# ℹ Use `print(n = ...)` to see more rows
```

Dữ liệu trường đại học Kinh tế quốc dân

```
# A tibble: 60 × 7
   truong  stt     ma_xet_tuyen  linh_vuc              chi_tieu_2022  chi_tieu_2023
chi_tieu_2024
   <chr>   <chr>   <chr>         <chr>                         <dbl>         <dbl>
 1 NEU     1       7320108       khxh và nhân văn                 65            60
 2 NEU     2       7510605       khoa học quản lý                125           120
 3 NEU     3       7340302       kt-tc                           130           120
 4 NEU     4       7340122       kt-tc                            65            60
 5 NEU     5       7340120       kt-tc                           135           120
 6 NEU     6       7340115       kt-tc                           250           180
 7 NEU     7       7310106       kt-tc                           135           120
 8 NEU     8       7310104       kt-tc                           200           180
```

```
 9 NEU    9     7340121      kt-tc                    200          120          120
10 NEU    10    7340405      công nghệ                135          120          120
# ℹ 50 more rows
# ℹ Use `print(n = ...)` to see more rows
```

Nhận xét: Tương tự với dữ liệu của các trường Qht, Qhx, Ueh, Hcmut, Ptit, Tct ta cũng in ra dữ liệu của các trường đã thu thập được

2.3 THỐNG KÊ MÔ TẢ

2.3.1. Dữ liệu về phim

Để chuẩn bị các công cụ cần thiết cho việc xử lý, phân tích thống kê và trực quan hóa dữ liệu, đặc biệt là trong các ứng dụng như phân tích dữ liệu lớn hoặc nghiên cứu dữ liệu người dùng, ta cần khai báo các thư viện sau:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import math
from collections import Counter
# Hành Động - Khoa học viễn tưởng
# Ant-Man and the Wasp: Quantumania
# Deadpool & Wolverine
# Guardians of the Galaxy Vol. 2
# The boys
# Transformers: Rise of the Beasts
# Phim Hài
# Babylon
# Bad Boys: Ride or Die
# Friends
# Only Murders in the Building
# The Intouchables 2011
# Phim về tội phạm
# Breaking Bad 2008
# Dexter
# Monsters
# Tulsa King
```

```
#joker(2019)
```

Cụ thể:

- **pandas**: Một thư viện mạnh mẽ để xử lý và phân tích dữ liệu dạng bảng (DataFrame), hỗ trợ các thao tác như đọc, ghi, và thao tác dữ liệu.
- **numpy**: Thư viện cho tính toán khoa học, xử lý các mảng số học, và cung cấp các hàm toán học hiệu quả.
- **matplotlib.pyplot**: Công cụ để tạo biểu đồ và trực quan hóa dữ liệu, cho phép hiển thị các thông tin quan trọng một cách trực quan.
- **math**: Thư viện cung cấp các hàm toán học cơ bản, hữu ích cho các phép tính như logarit, số mũ, căn bậc hai, v.v.
- **collections.Counter**: Một lớp trong thư viện collections giúp đếm tần suất xuất hiện của các phần tử trong một tập hợp, thường dùng cho phân tích dữ liệu định tính hoặc kiểm tra sự phân bố.

a) Phim Hành động – Khoa học viễn tưởng

Trước tiên, chúng ta cần đọc dữ liệu đánh giá từ trước của từng bộ phim để làm cơ sở cho các bước phân tích và đánh giá tiếp theo.

```
# Đọc dữ liệu các phim hành động và khoa học viễn tưởng
ant_man     =     pd.read_csv('https://raw.githubusercontent.com/NguyenKhang0062/-n-1/refs/heads/main/data_imdb/act_sci/ant_man.csv')
deadpool_wolverin = pd.read_csv('https://raw.githubusercontent.com/NguyenKhang0062/-n-1/refs/heads/main/data_imdb/act_sci/deadpool_wolverine.csv')
guardian_galaxy_3 = pd.read_csv('https://raw.githubusercontent.com/NguyenKhang0062/-n-1/refs/heads/main/data_imdb/act_sci/guardian_galaxy_3.csv')
the_boys     =     pd.read_csv('https://raw.githubusercontent.com/NguyenKhang0062/-n-1/refs/heads/main/data_imdb/act_sci/the_boys.csv')
transformers=     pd.read_csv('https://raw.githubusercontent.com/NguyenKhang0062/-n-1/refs/heads/main/data_imdb/act_sci/transformers_rise_of_the_beasts.csv')
```

Sau đây là kết quả hiển thị dữ liệu của từng bộ phim:

Phim “Ant Man”:

```
ant_man
```

	Rating	Review_Date	Review
0	5.0	Feb 18, 2023	Well it's happened. The MCU has run out of gas...
1	6.0	Feb 17, 2023	As a huge fan of the first one, and almost as ...
2	6.0	Feb 17, 2023	Marvel really has fallen and it can't get up. ...
3	6.0	Feb 20, 2023	New Ant-Man is not a bad movie, but it's repea...
4	6.0	Feb 19, 2023	I enjoyed watching Quantumania. It's a mostly ...
...	...	...	...
994	4.0	Feb 28, 2023	Believe me when I say I actually like Ant-Man ...
995	4.0	May 17, 2023	Ant-Man and the Wasp: Quantumania serves as a ...
996	2.0	Apr 10, 2023	These Ant-Man movies started in 2015 with Paul...
997	8.0	Feb 17, 2023	At first, filming the whole time in the multiv...
998	5.0	Feb 24, 2023	Under-the-radar Avenger Scott Lang is making t...

999 rows × 3 columns

Phim “Deadpool Wolverin”:

```
deadpool_wolverin
```

	Rating	Review_Date	Review
0	9.00	25-Jul-24	Hugh Jackman is the perfect Wolverine. What a ...
1	9.00	24-Jul-24	What a crazy blast ! Bonkers !!Sooo !... What ...
2	8.00	24-Jul-24	We've waited so long for this moment, and it w...
3	1.00	24-Jul-24	So many Easter Eggs, so true to the comic char...
4	9.00	26-Jul-24	I read an IGN review where the guy gave it a 7...
...	...	...	...
1700	4.87	1-Oct-24	TL;DR: turn off your brain for 2 hours & prete...
1701	4.87	28-Jul-24	Subjectively, I understand the title as "Fun e...
1702	4.87	12-Aug-24	Try to ignore much of the bad reviews, it is r...
1703	4.87	3-Oct-24	The first 2 movies in this trilogy had a great...
1704	4.87	3-Oct-24	Deadpool (Ryan Reynolds) is living an empty li...

1705 rows × 3 columns

Phim “Guardian Galaxy 3”:

```
guardian_galaxy_3
```

	Rating	Review_Date	Review
0	9.00	10-May-23	NaN
1	8.00	4-May-23	Having sat through some phase 4 films that fai...
2	9.00	3-May-23	This. This is what I've wanted. Yeah some of t...
3	1.00	4-May-23	It all leads back to where we once started off...
4	9.00	8-May-23	Up to this point, there has been one trilogy i...
...	...	...	...
1243	5.39	6-May-23	NaN
1244	5.39	5-May-23	The movie pans out in a seemingly random fashi...
1245	5.39	25-May-23	NaN
1246	5.39	21-May-23	NaN
1247	5.39	19-May-23	Boom! I'm not convinced the MCU is back to it'...

1248 rows × 3 columns

Phim “The Boys”:

```
the_boys
```

	Rating	Review_Date	Review
0	9.00	26 July 2019	Having being all superheroed out with the neve...
1	10.00	26 July 2019	Started watching this brilliant spin on a supe...
2	8.00	6 December 2021	I've had big expectations about this show, bec...
3	10.00	27 July 2019	NaN
4	9.00	26 July 2019	Excellent dystopian reimagining for Superhero ...
...	...	...	...
2993	7.88	31 December 2019	Ok, quite liking this series. Up to episode 4,...
2994	7.88	1 August 2019	I just don't get it, it's a nice twist on the ...
2995	7.88	18 August 2019	But I still have nightmares from one scene, w/...
2996	7.88	13 August 2019	Loved this show! So nitty, gritty, dirty, funn...
2997	7.88	21 September 2020	The Boys season 1 is a hilarious, gory and rea...

2998 rows × 3 columns

Phim "Transformers":

```
transformers
```

	Unnamed: 0	Rating	Review_Date	Review	Release_Date
0	0	6.0	Jun 11, 2023	Most of this movie is boring, even when there ...	June 9, 2023 (United States)
1	1	6.0	Aug 30, 2023	As a neutral, I went in there with no real exp...	June 9, 2023 (United States)
2	2	7.0	Apr 9, 2024	I usually don't write movie reviews. Since, I ...	June 9, 2023 (United States)
3	3	7.0	Jun 9, 2023	After the success of 'Bumblebee' I was hoping ...	June 9, 2023 (United States)
4	4	7.0	Jul 29, 2023	6.5/10 Well at least it's better than the Bumb...	June 9, 2023 (United States)
...	...	...	...	...	...
655	655	9.0	Jul 18, 2023	Transformers rise of the beasts took everythin...	June 9, 2023 (United States)
656	656	3.0	Jun 10, 2023	A film similar to the previous ones with some ...	June 9, 2023 (United States)
657	657	3.0	Jul 15, 2023	Honestly the last couple acts saved this movie...	June 9, 2023 (United States)
658	658	4.0	Jul 15, 2023	"Rise of the beasts" but doesn't feel like a b...	June 9, 2023 (United States)
659	659	10.0	Jul 15, 2023	Solid 7/10 here, and far better than the last ...	June 9, 2023 (United States)

660 rows × 5 columns

Tiếp theo, ta tiến hành kiểm tra các giá trị bị thiếu (null) trong dữ liệu.

```
# Kiểm tra xem Review có bao nhiêu giá trị null

guardian_nan = guardian_galaxy_3[guardian_galaxy_3['Review'].isna() == True].shape[0]

ant_man_nan = ant_man[ant_man['Review'].isna() == True].shape[0]

the_boy_nan = the_boys[the_boys['Review'].isna() == True].shape[0]

transformers_nan = transformers[transformers['Review'].isna() == True].shape[0]

dp_nan = deadpool_wolverin[deadpool_wolverin['Review'].isna() == True].shape[0]
```

Đoạn mã total_nan = guardian_nan + ant_man_nan + dp_nan + the_boy_nan + transformers_nan thực hiện việc cộng dồn giá trị của các biến guardian_nan, ant_man_nan, dp_nan, the_boy_nan, và transformers_nan, mỗi biến này có thể đại diện cho số lượng giá trị null (hoặc thiếu) trong dữ liệu của các bộ phim tương ứng. Kết quả total_nan = 1108 có nghĩa là tổng số giá trị null (hoặc thiếu) trong toàn bộ dữ liệu của các bộ phim trên là 1108. Sự cộng dồn này giúp bạn có cái nhìn tổng quan về mức độ thiếu dữ liệu trong tất cả các bộ phim được phân tích.

Để có cái nhìn trực quan hơn, ta sẽ thực hiện vẽ biểu đồ cho các giá trị thiếu của các bộ phim và sau đây là đoạn code minh họa:

```
x = ['guardian of galaxy 3','ant man','dead pool and wolverine','The boys','transformers']

y = [guardian_nan,ant_man_nan,dp_nan,the_boy_nan,transformers_nan ]

plt.title('Biểu đồ giá trị đánh giá bị thiếu thể loại act-sci')

plt.barh(x,y)

plt.show()
```

Tiếp theo, ta sẽ tổng hợp các sao trong đánh giá của các bộ phim:

```
# Tổng hợp các sao được đánh giá ở các bộ phim

ratings_act_sci=

np.concatenate([np.array(ant_man['Rating']),np.array(guardian_galaxy_3['Rating']),np.array(deadpool_wolverin['Rating']),np.array(the_boys['Rating']),np.array(transformers['Rating'])])

ratings_act_sci
```

Tính trung bình đánh giá của phim hành động, ta sử dụng đoạn code sau đây:

```
# Trung bình đánh giá của phim hành động

mean_rating_act =  ratings_act_sci.mean()
```

Nhận xét về trung bình sao đánh giá:

- **Điểm trung bình = 6.10** cho thấy rằng các bộ phim hành động trong bộ dữ liệu này nhận được đánh giá tương đối thấp so với các thể loại khác (ví dụ, thể loại tội phạm có trung bình là 8.52).
- Điểm 6.10 có thể phản ánh rằng thể loại hành động không hoàn toàn đáp ứng kỳ vọng của người xem. Những bộ phim hành động có thể không đủ ấn tượng hoặc có yếu tố thiếu sự sáng tạo, chất lượng không đồng đều, hoặc không mang lại trải nghiệm đặc biệt so với những thể loại khác.
- Một số lý do có thể giải thích điểm trung bình thấp:
    - Phim hành động có thể bị lặp lại: Các bộ phim hành động đôi khi theo một công thức quen thuộc và thiếu đổi mới sáng tạo, dẫn đến sự kém ấn tượng đối với người xem.
    - Thiếu chiều sâu về cốt truyện: Những bộ phim hành động có thể tập trung nhiều vào hành động, cảnh quay mãn nhãn mà bỏ qua yếu tố cốt truyện hoặc phát triển nhân vật, điều này có thể làm giảm sự hấp dẫn tổng thể đối với người xem.

- Sự phân tán trong chất lượng: Mặc dù thể loại hành động có thể thu hút một số lượng lớn người xem, nhưng sự phân tán trong chất lượng giữa các bộ phim hành động có thể là nguyên nhân khiến điểm trung bình không quá cao.

Tổng kết:

- Trung bình sao đánh giá 6.10 cho thấy thể loại phim hành động không được đánh giá quá cao, có thể do sự thiếu đổi mới, sự phân tán trong chất lượng, hoặc có thể vì phim hành động chưa hoàn toàn thỏa mãn được kỳ vọng của khán giả.
- Tuy nhiên, điểm số này vẫn cho thấy thể loại hành động vẫn có một số lượng người xem đáng kể và vẫn có sức hút, mặc dù không phải là thể loại nổi bật nhất.

Tiếp tục, ta tính phương sai, độ lệch chuẩn để đánh giá mức độ đồng đều trong việc đánh giá các bộ phim, từ đó giúp người phân tích hiểu rõ hơn về sự biến động trong nhận xét và đánh giá từ người dùng.

```
# Phương sai

print(np.var(ratings_act_sci))

#Phương sai lớn cho thấy được sự đánh giá của các phim không đồng đều

# Độ lệch chuẩn

print(np.std(ratings_act_sci))

# Kết quả hiển thị

10.403052939592413

3.2253764027772656
```

Từ kết quả hiện thị ở trên ta có thể thấy được:

Phương sai:

- Phương sai (Variance) của các dữ liệu trong bộ ratings_act_sci được tính là khoảng **10.4** (phương sai của rating phim). Phương sai đo lường sự phân tán của dữ liệu quanh giá trị trung bình. Phương sai càng lớn thì dữ liệu càng phân tán và ngược lại.
- Phương sai lớn cho thấy rằng các đánh giá của các bộ phim trong bộ dữ liệu này không đồng đều, có sự phân tán khá rộng. Điều này có thể chỉ ra rằng các bộ phim trong thể loại hành động và khoa học viễn tưởng (sci-fi) nhận được các đánh giá rất khác biệt: một số phim có thể được đánh giá rất cao, trong khi một số khác lại bị đánh giá thấp.

Độ lệch chuẩn (Standard Deviation):

- Độ lệch chuẩn (SD) của bộ dữ liệu là khoảng **3.23**, cho thấy sự phân tán của các giá trị xung quanh giá trị trung bình của bộ dữ liệu. Độ lệch chuẩn là căn bậc hai của phương sai, và nó cho biết mức độ phân tán thực tế của dữ liệu.
- Với độ lệch chuẩn là 3.23, ta có thể thấy rằng các đánh giá phim không đồng đều, có sự phân tán đáng kể quanh giá trị trung bình. Điều này có thể cho thấy một số bộ phim nhận được những đánh giá rất tích cực, trong khi một số bộ phim khác có thể nhận được những đánh giá rất tiêu cực hoặc thấp.

Nhận xét tổng quan:

- Các đánh giá của bộ phim không đồng đều và có sự phân tán lớn, tức là không phải tất cả các bộ phim đều được đánh giá theo cùng một cách. Điều này có thể phản ánh sự đa dạng trong cảm nhận và quan điểm của người xem đối với các bộ phim hành động và khoa học viễn tưởng (sci-fi).
- Việc có phương sai và độ lệch chuẩn cao cho thấy rằng có sự khác biệt rõ rệt trong mức độ đánh giá của các bộ phim trong bộ dữ liệu. Sự đa dạng này có thể do nhiều yếu tố như thể loại phim, nội dung, sự yêu thích của khán giả, hoặc các yếu tố liên quan đến diễn viên và đạo diễn.

Tóm lại, bộ dữ liệu này có tính không đồng đều trong các đánh giá phim, với sự phân tán rõ rệt, và điều này phản ánh sự đa dạng trong cách nhìn nhận của người xem đối với các bộ phim hành động và khoa học viễn tưởng.

Tính phân vị (percentiles) trong thống kê giúp chia dữ liệu thành các phần nhỏ, từ đó cung cấp cái nhìn chi tiết hơn về phân bố của dữ liệu. Các phân vị giúp xác định vị trí của một giá trị cụ thể trong tập dữ liệu, giúp hiểu rõ hơn về sự phân bố và tính chất của dữ liệu. Ta có thể sử dụng đoạn code sau:

```
#Phân vị

print("Q1 = : ", np.quantile(ratings_act_sci, 0.25))

print("Q2 = : ", np.quantile(ratings_act_sci, 0.5))

print("Q3 = : ", np.quantile(ratings_act_sci, 0.75))
```

Kết quả hiển thị:

```
Q1 = :  3.0
Q2 = :  7.0
Q3 = :  9.0
```

Qua đó, ta thấy được:

- Bộ dữ liệu này có sự phân bố khá đều nhưng không hoàn toàn đồng nhất. Phần lớn các bộ phim nằm trong khoảng từ 3 đến 9 điểm đánh giá, với phần lớn các phim có đánh giá từ 7.0 đến 9.0.

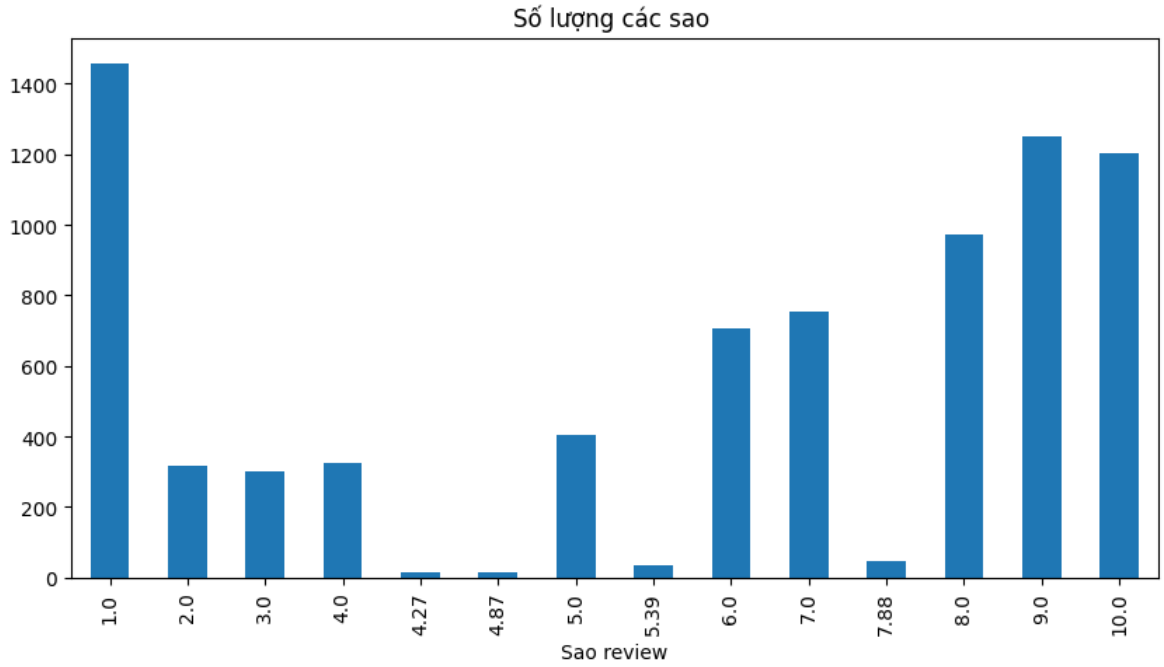
- Điều này cho thấy rằng, mặc dù có sự phân tán khá lớn trong các đánh giá (với Q1 = 3.0), nhưng đa số các bộ phim vẫn nhận được đánh giá tương đối tốt, với giá trị trung vị là 7.0.
- Những bộ phim có đánh giá rất thấp (dưới Q1) không chiếm đa số, nhưng vẫn có sự hiện diện của các phim bị đánh giá thấp (chắc chắn không phải là số lượng nhỏ). Phân vị Q3 = 9.0 cho thấy rằng có một số bộ phim được người xem đánh giá rất cao, nhưng chúng chiếm tỉ lệ nhỏ so với tổng số phim trong bộ dữ liệu.

Tóm lại, bộ dữ liệu này cho thấy rằng đánh giá phim rất đa dạng, với phần lớn các phim có điểm số khá tốt, nhưng vẫn có một vài phim nhận được đánh giá rất thấp và một số ít phim đạt được đánh giá xuất sắc.

Tiến hành tổng hợp và trực quan hóa số lượng các sao (rating) từ các bộ phim khác nhau, giúp người phân tích có cái nhìn tổng quan về sự phân bố đánh giá của các bộ phim này và ta sử dụng đoạn code sau:

```
ratings_act_sci.sort()
unique_elements, counts = np.unique(ratings_act_sci, return_counts=True)
unique_elements, counts
rating_data = pd.DataFrame([unique_elements, counts])
rating_act                                                                    =
pd.concat([ant_man['Rating'],guardian_galaxy_3['Rating'],deadpool_wolverin['Rating']
,the_boys['Rating'],transformers['Rating']])
ax= rating_act.value_counts().sort_index().plot(kind = 'bar',
                                          title = 'Số lượng các sao ',
                                          figsize = (10,5))
ax.set_xlabel('Sao review')
plt.show()
```

Kết quả hiển thị:



Hình 2. 1. Biểu đồ thể hiện số lượng sao của phim Hành động - Khoa học viễn tưởng

Qua biểu đồ trên, ta có thể thấy được

Phân bố số lượng đánh giá theo sao:

- Mức đánh giá 1.0 sao chiếm số lượng lớn nhất với hơn 1400 lượt review, cho thấy có rất nhiều người không hài lòng. Điều này có thể do yếu tố chất lượng nội dung, kỳ vọng không được đáp ứng, hoặc trải nghiệm tệ từ người dùng.
- Ngược lại, mức 9.0 sao và 10.0 sao cũng có lượng review cao, khoảng hơn 1200 lượt mỗi mức, thể hiện một lượng lớn khán giả rất hài lòng với sản phẩm hoặc nội dung.

Các mức đánh giá trung bình:

- Các mức từ 5.0 sao đến 7.0 sao có sự gia tăng dần về số lượng, với một lượng đáng kể tập trung ở mức 6.0 sao và 7.0 sao. Điều này cho thấy đây là khoảng điểm được nhiều khán giả đánh giá ở mức độ trung lập hoặc hài lòng vừa phải.
- Tuy nhiên, các mức cụ thể như 4.27 sao, 4.87 sao, 5.39 sao và 7.88 sao có số lượng rất thấp, chỉ xuất hiện dưới dạng ngoại lệ, do đó không đóng vai trò quan trọng trong xu hướng chung.

Tính phân cực trong đánh giá:

- Phân bố đánh giá có tính phân cực rõ rệt, với số lượng đánh giá tập trung mạnh ở mức thấp nhất (1.0 sao) và cao nhất (9.0-10.0 sao). Điều này cho thấy người dùng có cảm xúc mạnh, hoặc rất không hài lòng hoặc rất yêu thích, mà ít lựa chọn các mức trung bình

Ý nghĩa cảm xúc và trải nghiệm người dùng:

- Mức 1.0 sao cao bất thường có thể phản ánh sự thất vọng lớn từ một phần khán giả, có khả năng do yếu tố như chất lượng kỹ thuật, nội dung không thỏa mãn hoặc không đáp ứng kỳ vọng.
- Sự phổ biến của mức 9.0 và 10.0 sao cho thấy có một nhóm lớn khán giả cảm thấy sản phẩm xuất sắc, thể hiện bằng sự yêu thích tuyệt đối.

Qua đó, ta sẽ tiếp tục có tỷ lệ đánh giá theo thời gian:

Hình 2. 2. Biểu đồ tỉ lệ đánh giá các bộ phim Hành động – Khoa học viễn tưởng sau khi phát sóng

Qua biểu đồ trên ta có thể thấy được

Giai đoạn 1 tuần đầu tiên (Màu đỏ, 34.0%)

- Tỉ lệ khá cao (chiếm hơn 1/3 tổng số đánh giá):
  - Điều này cho thấy nội dung nhận được sự quan tâm lớn ngay khi phát sóng, đặc biệt từ những khán giả cốt lõi hoặc người hâm mộ đã chờ đợi sản phẩm.
  - Đây là khoảng thời gian "hiệu ứng ra mắt" mạnh mẽ, khi các chiến dịch quảng bá trước đó hoặc tính hấp dẫn của nội dung đã kích thích người xem nhanh chóng đánh giá.
- Ý nghĩa:
  - Phản ánh mức độ thành công của chiến lược ra mắt (marketing, trailer, hoặc thông điệp truyền thông).
  - Tuy nhiên, 34% vẫn chưa phải là cao nhất, điều này cho thấy vẫn còn dư địa để cải thiện chiến lược truyền thông nhằm thúc đẩy khán giả đánh giá nhiều hơn trong tuần đầu tiên.

Giai đoạn 1 tháng đầu tiên (Màu xanh dương, 18.2%)

- Tỉ lệ thấp nhất trong ba giai đoạn:

    - Giai đoạn này có mức đánh giá giảm sút rõ rệt so với tuần đầu tiên. Đây có thể là do:
    - Khán giả giảm dần sự chú ý sau khi nội dung đã phát sóng, đặc biệt nếu không có hoạt động quảng bá tiếp theo.
    - Đối tượng khán giả mới chưa được tiếp cận đầy đủ hoặc chưa có động lực để xem và đánh giá nội dung.
- Ý nghĩa:
    - Giai đoạn này thường là thời điểm quyết định sức hút trung hạn của nội dung. Nếu sự quan tâm giảm mạnh, có thể do nội dung chưa đủ đặc sắc để giữ chân khán giả hoặc chưa có hiệu ứng lan truyền.
    - Cần phân tích thêm về chiến lược truyền thông trong giai đoạn này và xác định lý do tại sao lượng đánh giá giảm.

Giai đoạn sau 1 tháng (Màu xanh lá, 47.8%)

- Tỉ lệ cao nhất, gần 50% tổng đánh giá:
    - Sự gia tăng đáng kể ở giai đoạn này cho thấy nội dung đã tạo được "hiệu ứng lan tỏa" sau thời gian phát sóng, đặc biệt nhờ vào:
    - Lời truyền miệng từ những khán giả ban đầu.
    - Hiệu ứng trên mạng xã hội hoặc các nền tảng truyền thông khác.
    - Việc khán giả xem lại hoặc xem muộn so với thời điểm phát sóng ban đầu.
    - Điều này cũng cho thấy nội dung có sức hút lâu dài, có khả năng tiếp tục thu hút khán giả mới sau khi phát sóng.
- Ý nghĩa:
    - Đây là dấu hiệu tích cực, đặc biệt nếu sản phẩm là nội dung dài hạn (như phim truyền hình, series) hoặc nội dung cần thời gian để đạt được độ phổ biến (viral).
    - Cơ hội tiếp tục khai thác và duy trì sự quan tâm, chẳng hạn bằng cách phát hành phiên bản mở rộng, sản phẩm liên quan, hoặc chiến lược quảng bá bổ sung.
- Xu hướng tổng thể:
    - Biểu đồ phản ánh xu hướng đánh giá tích cực trong dài hạn:
    - Giai đoạn tuần đầu tiên (34%) cho thấy sản phẩm có sự khởi đầu tốt, nhưng không phải là đỉnh điểm.
    - Giai đoạn 1 tháng (18.2%) lại chững lại, phản ánh những thách thức trong việc giữ nhiệt độ cho sản phẩm.

- Sự bùng nổ sau 1 tháng (47.8%) cho thấy sản phẩm có sức sống bền vững, đặc biệt là khả năng tiếp cận các nhóm khán giả mới hoặc tạo hiệu ứng truyền thông lan tỏa.

- **Đề xuất cải thiện:**

  Tăng tỉ lệ đánh giá trong 1 tuần đầu tiên:

  - Triển khai các chiến dịch quảng bá mạnh mẽ ngay trước và trong tuần đầu ra mắt.
  - Tận dụng các nền tảng truyền thông xã hội, sự kiện công chiếu, hoặc chương trình khuyến khích đánh giá từ khán giả sớm.

  Chống "tụt nhiệt" trong 1 tháng đầu:

  - Duy trì sức hút bằng các nội dung bổ trợ như hậu trường, phỏng vấn diễn viên, hoặc tương tác với khán giả.
  - Tăng cường quảng bá nhắm mục tiêu đến các nhóm khán giả tiềm năng chưa tiếp cận sản phẩm.

  Khai thác sự tăng trưởng sau 1 tháng:

  - Tiếp tục thúc đẩy nội dung thông qua các chiến dịch gợi nhớ, kết hợp với đánh giá tích cực từ khán giả trước đó.
  - Tận dụng các kênh truyền thông (bài đánh giá, video phân tích nội dung) để kéo dài vòng đời sản phẩm.

Tóm lại, biểu đồ cho thấy sản phẩm có hiệu ứng lan tỏa tốt và được khán giả nhớ đến lâu dài, nhưng cần cải thiện chiến lược để khai thác tối đa tiềm năng trong tuần đầu tiên và tháng đầu tiên để tăng tốc độ tăng trưởng trong tương lai.

b) Phim Hài

Trước tiên, chúng ta cần đọc dữ liệu đánh giá đã thu thập ở trên của từng bộ phim để làm cơ sở cho các bước phân tích và đánh giá tiếp theo.

```
# Đọc dữ liệu của phim hài
baby_lon    =    pd.read_csv(r'https://raw.githubusercontent.com/NguyenKhang0062/-n-1/refs/heads/main/data_imdb/comedy/baby_lon.csv')
bad_boy     =    pd.read_csv(r'https://raw.githubusercontent.com/NguyenKhang0062/-n-1/refs/heads/main/data_imdb/comedy/bad_boy_ride_or_die.csv')
friends     =    pd.read_csv(r'https://raw.githubusercontent.com/NguyenKhang0062/-n-1/refs/heads/main/data_imdb/comedy/friends.csv')
murder_in_building                                                    =
pd.read_csv(r'https://raw.githubusercontent.com/NguyenKhang0062/-n-1/refs/heads/main/data_imdb/comedy/only_murder_in_the_building.csv')
```

```
intouchable=     pd.read_csv(r'https://raw.githubusercontent.com/NguyenKhang0062/-n-1/refs/heads/main/data_imdb/comedy/the_intouchable.csv')
```

Sau đây là kết quả hiển thị của từng bộ phim:

Phim “Baby Lon”

```
baby_lon
```

	Rating	Review_Date	Review
0	6.00	7 February 2023	...and I wish I could rate them separately. Th...
1	6.00	7 February 2024	So I just finished watching Babylon.\n\nI real...
2	8.00	4 January 2023	Whether it be orgies, showcasing various bodil...
3	8.00	29 December 2022	After an "interesting" opening scene about our...
4	7.00	26 December 2022	Babylon is a long, messy, repulsive, and magne...
...	...	...	...
970	4.71	31 January 2023	I'm going to tell you exactly why his film bom...
971	4.71	2 February 2023	I couldn't help but compare this to Boogie Nig...
972	4.71	14 September 2023	It is obviously a huge, gigantic production, v...
973	4.71	1 March 2023	This is definitely one to see in theatres if y...
974	4.71	2 February 2023	NaN

975 rows × 3 columns

Phim “Bad Boy”

```
bad_boy
```

	Rating	Review_Date	Review
0	8.00	5 June 2024	So I was left a little disappointed after the ...
1	7.00	6 June 2024	I do not understand why some people are still ...
2	9.00	5 June 2024	I have always been a fan of will and Martin ev...
3	8.00	5 June 2024	The fourth entry in the revitalised buddy cop,...
4	7.00	7 June 2024	Terrible acting, bad storyline and absolutely ...
...	...	...	...
335	7.00	9 October 2024	Love them or hate them. Bad Boys was a one hit...
336	6.00	9 October 2024	Actually quite surprised by this movie. I was ...
337	6.72	9 October 2024	I had zero expectations for this. The last one...
338	6.72	19 September 2024	Bad Boys Ride or Die\n\nMaybe since the franch...
339	6.72	16 September 2024	Bad Boys: Ride or Die is the latest installmen...

340 rows × 3 columns

Phim “Friends”

```
friends
```

	Rating	Review_Date	Review
0	9.00	29 October 2023	I never used to watch this show. My wife loved...
1	10.00	29 October 2023	I must've have rewatched these 10 series dozen...
2	10.00	29 September 2022	What can be said about Friends that hasn't alr...
3	10.00	16 May 2022	NaN
4	10.00	12 November 2020	Are you happy? watch Friends! are you sad? wat...
...	...	...	...
2060	8.65	11 September 2001	I haven't watched this show for as long as it'...
2061	8.65	25 November 2003	Friends-****/**** Starring Jennifer Aniston, C...
2062	8.65	5 October 2005	I love this show! I even have no words to desc...
2063	8.65	8 August 1999	I can't go a week without watching "Friends". ...
2064	8.65	2 November 2005	I was only 14 years old when this show came ou...

2065 rows × 3 columns

Phim “Munder In Building”

```
murder_in_building
```

	Rating	Review_Date	Review
0	9.00	22 October 2021	Season 1 (9/10) When this series came my girlf...
1	8.00	11 September 2022	So I just finished watching Only Murders in th...
2	8.00	29 September 2022	I was actually surprised by how much I liked O...
3	9.00	19 October 2021	When the initial episodes of 'Only Murders in ...
4	8.00	7 September 2021	We all know that Steve Martin and Martin Short...
...	...	...	...
1104	7.34	10 May 2024	So many good things about this show!! Great ca...
1105	7.34	26 February 2024	So how do I start telling everyone I like the ...
1106	7.34	3 January 2024	This series of 30-minute episodes is a HULU or...
1107	7.34	29 November 2021	When I saw the ad for this on Hulu and decided...
1108	7.34	4 March 2024	I loved the characters and I thought the casti...

1109 rows × 3 columns

Phim “Intouchable”

```
[ ] intouchable
```

	Rating	Review_Date	Review
0	1.00	27 November 2011	Do not look at this through the prism of "Fore...
1	1.00	10 January 2012	I am now trying to find words to describe this...
2	1.00	17 January 2012	In less than two months, "Untouchable" became ...
3	9.00	14 November 2011	Being french and a film maker myself, I have h...
4	1.00	28 December 2011	It has been 9 weeks now since Intoucblaes has ...
...	...	...	...
898	4.66	23 March 2017	NaN
899	4.66	19 January 2022	"1+1" is a movie that I watched at a very youn...
900	4.66	23 June 2024	After he becomes a quadriplegic from a paragli...
901	4.66	20 July 2017	Juste top ! Un trees born film Biden interprét...
902	4.66	14 August 2013	"Pragmatic" is the way Griss, an ex-convict fr...

903 rows × 3 columns

Tiếp theo, ta xác định số lượng giá trị bị thiếu trong cột Review của các DataFrame để đánh giá và xử lý dữ liệu bị thiếu từ đó đưa ra các quyết định phù hợp để đảm bảo tính toàn vẹn của phân tích dữ liệu và ta sử dụng đoạn code sau đây:

```
baby_lon_nan = baby_lon[baby_lon['Review'].isna() == True].shape[0]

bad_boy_nan = bad_boy[bad_boy['Review'].isna() == True].shape[0]
```

```
friends_nan = friends[friends['Review'].isna() == True].shape[0]

murder_in_building_nan = murder_in_building[murder_in_building['Review'].isna() == True].shape[0]

intouchable_nan = intouchable[intouchable['Review'].isna() == True].shape[0]
```

Kết quả hiển thị:

```
#Tổng dữ liệu null trong cột Review

baby_lon_nan + bad_boy_nan + friends_nan + murder_in_building_nan + intouchable_nan
```

Tiếp theo, ta sử dụng đoạn code sau để tính số sao đánh giá của thể loại phim Hài:

```
# Tổng hợp sao đánh giá của thể loại comedy

ratings_comedy=
np.concatenate([np.array(baby_lon['Rating']),np.array(bad_boy['Rating']),np.array(friends['Rating']),np.array(murder_in_building['Rating']),np.array(intouchable['Rating'])])

ratings_comedy
```

Qua đó, để có cái nhìn sâu sắc hơn về chất lượng của các thể loại phim, chúng ta sẽ tiến hành phân tích số lượng đánh giá tích cực, tiêu cực và tính trung bình số sao đánh giá.

```
#Lượng rating tích cực

ratings_comedy[ratings_comedy > 5].size

#Kết quả hiển thị

3806
```

```
#Lượng rating tiêu cực

ratings_comedy[ratings_comedy < 5].size

#Kết quả hiển thị

1393
```

```
#Trung bình sao đánh giá

ratings_comedy.mean()

#Kết quả hiển thị

6.876225890207715
```

Thông qua kết quả trên, dựa vào số sao đánh giá trung bình ta thấy được:

- **Điểm trung bình = 6.88** cho thấy rằng các bộ phim thuộc thể loại hài nhận được đánh giá trung bình từ người xem, không quá cao nhưng cũng không quá thấp.

Điểm số này có thể cho thấy rằng thể loại hài có sự yêu thích vừa phải, nhưng không phải là thể loại được yêu thích hoặc đánh giá xuất sắc nhất.

- Mức 6.88 có thể chỉ ra rằng có một sự phân tán trong chất lượng hoặc sự yêu thích của các bộ phim hài. Một số bộ phim có thể được yêu thích và đánh giá cao, trong khi những bộ phim khác có thể không đáp ứng được kỳ vọng của người xem, dẫn đến điểm số trung bình không quá cao.
- So với các thể loại phim khác (như thể loại tội phạm với điểm 8.52), thể loại hài có vẻ không được đánh giá cao bằng, cho thấy có thể có sự thiếu đồng đều trong chất lượng của các bộ phim hài hoặc rằng người xem có những kỳ vọng cao hơn đối với thể loại hài.

Kết luận:

- Trung bình sao đánh giá 6.88 cho thấy thể loại hài có mức đánh giá tương đối vừa phải, với một số bộ phim có thể được đánh giá cao và một số bộ phim khác bị đánh giá thấp hơn.
- Mức điểm này không phản ánh sự xuất sắc tuyệt đối, nhưng vẫn cho thấy rằng thể loại hài có sự thu hút và phổ biến nhất định trong cộng đồng người xem.

Tiếp tục, ta tính phương sai, độ lệch chuẩn để đánh giá mức độ đồng đều trong việc đánh giá các bộ phim, từ đó giúp người phân tích hiểu rõ hơn về sự biến động trong nhận xét và đánh giá từ người dùng.

```
# Phương sai

print(np.var(ratings_comedy))

#Phương sai lớn cho thấy được sự đánh giá của các phim không đồng đều

# Độ lệch chuẩn

print(np.std(ratings_comedy))
```

```
10.92448514407747
3.305220891873563
```

```
#Phân vị

print("Q1 = : ", np.quantile(ratings_comedy, 0.25))

print("Q2 = : ", np.quantile(ratings_comedy, 0.5))

print("Q3 = : ", np.quantile(ratings_comedy, 0.75))
```

```
Q1 = :  4.66
Q2 = :  8.0
Q3 = :  10.0
```

Từ kết quả trên ta thấy:

Phương sai (var**) và độ lệch chuẩn (**std**) của ratings_comedy:**

- Phương sai:

    - Giá trị là 10.92, cho thấy sự phân tán trong đánh giá của thể loại phim hài tương đối lớn. Điều này cho biết các đánh giá không đồng đều, có sự chênh lệch đáng kể giữa những người đánh giá.
- Độ lệch chuẩn:
    - Giá trị là 3.31, đây là căn bậc hai của phương sai. Độ lệch chuẩn cũng chỉ ra rằng mức độ dao động của các đánh giá so với giá trị trung bình là khoảng 3.31 đơn vị. Điều này nhấn mạnh rằng các đánh giá không tập trung quanh một điểm duy nhất mà phân tán đáng kể.

Phân vị (quantile) của ratings_comedy:

- Q1 (Phân vị thứ 25%): 4.66
    - 25% số đánh giá thấp hơn hoặc bằng 4.66, cho thấy một số lượng nhỏ các phim hài nhận được điểm đánh giá thấp.
- Q2 (Phân vị thứ 50% hoặc Median - Trung vị): 8.0
    - 50% số đánh giá thấp hơn hoặc bằng 8.0, cho thấy rằng đa số phim hài có mức đánh giá tương đối cao (trung bình hoặc tốt).
- Q3 (Phân vị thứ 75%): 10.0
    - 75% số đánh giá thấp hơn hoặc bằng 10.0, nghĩa là một tỷ lệ lớn các phim hài nhận được điểm đánh giá rất cao.

Tổng kết:

- Sự phân tán lớn trong dữ liệu đánh giá: Phương sai và độ lệch chuẩn đều cao, cho thấy rằng ý kiến của người xem về các bộ phim hài rất khác nhau.
- Điểm đánh giá chủ yếu nằm ở mức cao: Với Q2 = 8.0 và Q3 = 10.0, phần lớn các phim hài có chất lượng được đánh giá là tốt hoặc xuất sắc, mặc dù vẫn có một số phim nhận được điểm đánh giá thấp hơn (Q1 = 4.66).

Điều này cho thấy rằng thể loại phim hài có sức hút lớn đối với khán giả, nhưng cũng tồn tại một số phim có chất lượng thấp hơn, dẫn đến sự phân tán đáng kể trong đánh giá.

Để theo dõi xu hướng thay đổi trong đánh giá, xem xét mức độ yêu thích hoặc không hài lòng với phim có thay đổi theo thời gian hay không, xác định các giai đoạn phim nhận được nhiều hay ít đánh giá, từ đó phân tích các yếu tố như mùa chiếu, sự kiện đặc biệt hoặc các chiến dịch quảng bá, hiểu cách khán giả phản ứng với phim qua các thời điểm khác nhau, ví dụ, liệu họ có đánh giá tích cực hơn ngay sau khi phim ra mắt hoặc giảm dần khi thời gian trôi qua. Ta cùng tiến hành tính tỷ lệ đánh giá các bộ phim theo thời gian. Dưới đây là biểu đồ quạt thể hiện tỉ lệ đánh giá theo thời gian

Hình 2. 3. Biểu đồ tỉ lệ đánh giá các bộ phim Hài sau khi phát sóng

Qua biểu đồ trên ta có thể thấy được

Đánh giá tập trung vào giai đoạn sau khi phát sóng:

- Phần lớn các đánh giá được đưa ra sau 1 tháng phát sóng, cho thấy người xem cần thời gian để trải nghiệm sản phẩm đầy đủ trước khi đưa ra đánh giá. Thực tế, sự đánh giá thường được thực hiện sau một khoảng thời gian trải nghiệm sản phẩm, vì người xem cần thời gian để tiêu thụ đầy đủ nội dung và suy nghĩ về chất lượng. Điều này có thể phản ánh sự thật rằng người xem không đưa ra đánh giá vội vàng mà muốn xác nhận lại cảm nhận của họ sau khi xem hết toàn bộ nội dung.

Sự giảm dần của lượng đánh giá mới:

- Số lượng đánh giá mới giảm dần theo thời gian, đặc biệt là sau 1 tháng đầu tiên.

Điều này có thể do một số nguyên nhân như:

- Hiệu ứng mới lạ: Ban đầu, sản phẩm mới ra mắt thường thu hút sự chú ý và đánh giá của nhiều người.
- Theo thời gian: Sự quan tâm của khán giả có thể giảm dần, dẫn đến ít người đánh giá hơn.
- Ảnh hưởng của truyền miệng: Các đánh giá của những người đã xem trước có thể ảnh hưởng đến quyết định xem và đánh giá của những người khác, đặc biệt là trong giai đoạn sau khi phát sóng.

c) Phim Tội phạm

Trước tiên, chúng ta cần đọc dữ liệu đánh giá thu thập từ trước của từng bộ phim để làm cơ sở cho các bước phân tích và đánh giá tiếp theo.

```
breaking_bad = pd.read_csv(r'https://raw.githubusercontent.com/NguyenKhang0062/-n-1/refs/heads/main/data_imdb/crime/breaking_bad.csv')

dexter = pd.read_csv(r'https://raw.githubusercontent.com/NguyenKhang0062/-n-1/refs/heads/main/data_imdb/crime/dexter.csv')

joker = pd.read_csv(r'https://raw.githubusercontent.com/NguyenKhang0062/-n-1/refs/heads/main/data_imdb/crime/joker.csv')

monsters = pd.read_csv(r'https://raw.githubusercontent.com/NguyenKhang0062/-n-1/refs/heads/main/data_imdb/crime/monsters.csv')

tulsa_king= pd.read_csv(r'https://raw.githubusercontent.com/NguyenKhang0062/-n-1/refs/heads/main/data_imdb/crime/tulsa_king_.csv')
```

Kết quả hiển thị số sao đánh giá, ngày đánh giá và nội dung đánh giá của từng bộ phim:

Phim "Breaking Bad"

```
breaking_bad
```

	Rating	Review_Date	Review
0	10.00	4 July 2021	I have never watched a show that is as consist...
1	10.00	7 March 2019	Re-Watched it 7 times and counting. I guess I ...
2	10.00	30 July 2021	One of the greatest shows ever, the pacing is ...
3	10.00	19 February 2020	I cannot stress enough how good this show is. ...
4	10.00	9 November 2021	Breaking Bad is absolutely, without a doubt, o...
...	...	...	...
5057	9.31	8 December 2017	NaN
5058	9.31	25 February 2019	If you are among the few who haven't seen it y...
5059	9.31	16 March 2023	That's the first series that I watched - I mea...
5060	9.31	22 March 2023	NaN
5061	9.31	30 September 2020	This show is amazing. While the story can be a...

5062 rows × 3 columns

Phim "Dexter"

```
dexter
```

	Rating	Review_Date	Review
0	10.0	7 November 2006	NaN
1	10.0	14 November 2020	Thanks to the quarantine, I watched series tha...
2	10.0	30 October 2006	Dexter is the show I have been waiting for my ...
3	10.0	1 October 2006	I read Lindsay's excellent books - Darkly Drea...
4	10.0	29 August 2023	Dexter is absolutely, without a doubt, one of ...
...	...	...	...
1050	8.6	27 May 2015	This show has you on the edge of your seat, la...
1051	8.6	10 January 2022	NaN
1052	8.6	30 December 2008	Love him. Love him to death. And I love the sh...
1053	8.6	18 September 2018	The Netflix series Dexter is the type of show ...
1054	8.6	8 April 2016	I remember seeing previews for each episodes a...

1055 rows × 3 columns

Phim "Joker"

```
joker
```

	Rating	Review_Date	Review
0	7.00	15 October 2019	NaN
1	9.00	5 October 2019	The movie affects you in a way that makes it p...
2	7.00	6 October 2019	I thought this film was good but I just don't ...
3	10.00	3 October 2019	Every once in a while a movie comes, that trul...
4	10.00	7 October 2019	This is a movie that only those who have felt ...
...	...	...	...
11584	8.31	3 February 2020	This is not an action film like most comic boo...
11585	8.31	13 November 2019	NaN
11586	8.31	20 June 2023	I would have never thought that in 2019, the H...
11587	8.31	15 October 2019	Review (1~5)\n\n#Content: Script 4 / Acting 5 ...
11588	8.31	19 October 2024	NaN

11589 rows × 3 columns

Phim “Monsters”

```
[ ] monsters
```

	Rating	Review_Date	Review
0	8.00	21 September 2022	I had high expectations of this. After seeing ...
1	9.00	21 September 2022	This was a hard watch....for many reasons. Hav...
2	9.00	21 September 2022	When I clicked to watch the "Dahmer"-release o...
3	10.00	22 September 2022	There are no doubts that Evan Peters has kille...
4	9.00	22 September 2022	For those interested in the psychology of a se...
...	...	...	...
843	7.22	6 March 2024	This is overall a decent show. The best thing ...
844	7.22	16 October 2022	Wow, this series is amazing. I don't think it'...
845	7.22	13 October 2024	The Menendez Brothers is probably the worst se...
846	7.22	12 August 2023	I 100% recommend it. If you are looking for a ...
847	7.22	18 October 2024	Both of the main actors acted very well. But, ...

848 rows × 3 columns

Phim “Tulsa King”

```
[ ] tulsa_king
```

	Rating	Review_Date	Review
0	9.00	3 December 2022	I have loved Tulsa King so far. I actually lov...
1	9.00	28 November 2022	I figured this would be an entertaining show. ...
2	9.00	21 November 2022	Sylvester Stallone absolutely kills it in Tuls...
3	9.00	14 November 2022	I watched this out of curiosity, no high expec...
4	8.00	28 December 2022	I've always liked Stallone, over the years, ye...
...	...	...	...
346	7.17	19 January 2023	This is the weakest Sheridan effort I've seen....
347	7.17	24 January 2023	NaN
348	7.17	15 October 2024	I like this show, I really do, I think it has ...
349	7.17	1 April 2024	Disclaimer: I dropped this after one episode. ...
350	7.17	26 September 2024	Actually this would have been so great with cl...

351 rows × 3 columns

Tương tự như 2 thể loại phim ở trên, tiếp theo ta tiến hành kiểm tra các giá trị bị thiếu (null) trong dữ liệu.

```
breaking_bad_nan = breaking_bad[breaking_bad['Review'].isna() == True].shape[0]

dexter_nan = dexter[dexter['Review'].isna() == True].shape[0]

joker_nan = joker[joker['Review'].isna() == True].shape[0]

monsters_nan = monsters[monsters['Review'].isna() == True].shape[0]
```

```
tulsa_king_nan = tulsa_king[tulsa_king['Review'].isna() == True].shape[0]
```

Kết quả hiển thị:

```
#Tổng dữ liệu null trong cột Review

breaking_bad_nan + dexter_nan + joker_nan + monsters_nan + tulsa_king_nan
```

2756

Ta thấy:

- Dữ liệu thiếu khá lớn: Với **2,756** giá trị thiếu, tỷ lệ thiếu dữ liệu trong cột Review có thể đáng kể, tùy thuộc vào tổng số dữ liệu ban đầu của các tập phim.
- Ảnh hưởng đến phân tích: Số lượng giá trị thiếu lớn như vậy có thể ảnh hưởng đáng kể đến chất lượng và tính chính xác của việc phân tích dữ liệu (như đánh giá mức độ yêu thích phim hoặc xu hướng).

Qua đây, ta có thể thấy được tổng quát số sao đánh giá của thể loại phim tội phạm

```
# Tổng hợp sao đánh giá của thể loại crime

ratings_crime=

np.concatenate([np.array(breaking_bad['Rating']),np.array(dexter['Rating']),np.array

(joker['Rating']),np.array(monsters['Rating']),np.array(tulsa_king['Rating'])])

ratings_crime
```

Cụ thể, ta có kết quả của các lượt sao đánh giá tích cực, tiêu cực và trung bình của phim Tội phạm:

```
# Lượng rating tích cực

ratings_crime[ratings_crime>5].size
```

16573

```
#Lượng rating tiêu cực

ratings_crime[ratings_crime < 5].size
```

1828

```
#Trung bình đánh giá của tội phạm

ratings_crime.mean()
```

8.523158423697437

Nhận xét về trung bình đánh giá:

- Điểm trung bình đánh giá = 8.52 cho thấy các bộ phim thuộc thể loại tội phạm (crime) nhận được đánh giá khá cao từ người xem. Đây là một điểm số tương đối

tích cực và cho thấy rằng phần lớn người xem đánh giá thể loại này ở mức khá hoặc xuất sắc.

- Mức điểm 8.5 cho thấy rằng đa số các bộ phim tội phạm trong bộ dữ liệu này đều được đánh giá tốt. Đây là một chỉ số phản ánh rằng thể loại phim tội phạm có xu hướng thu hút người xem và được đánh giá cao về chất lượng nội dung, diễn xuất, hoặc các yếu tố khác.
- Tuy nhiên, điểm số 8.52 cũng không phải là điểm tuyệt đối (10), có thể cho thấy rằng vẫn có một số bộ phim trong thể loại này có một số điểm yếu hoặc không hoàn toàn đạt được sự đồng thuận hoàn hảo từ tất cả người xem.

Kết luận:

- Trung bình 8.52 là một mức đánh giá khá cao, cho thấy thể loại phim tội phạm có thể đang được yêu thích và có chất lượng tốt, tuy nhiên vẫn còn một số bộ phim không đạt được điểm tuyệt đối.
- Nếu bạn muốn so sánh, mức điểm này có thể cao hơn hoặc thấp hơn so với các thể loại phim khác trong bộ dữ liệu của bạn, tùy thuộc vào các yếu tố như chủ đề, nội dung, hoặc sự yêu thích chung của người xem đối với thể loại tội phạm.

Tiếp tục, ta tính phương sai, độ lệch chuẩn để đánh giá mức độ đồng đều trong việc đánh giá các bộ phim, từ đó giúp người phân tích hiểu rõ hơn về sự biến động trong nhận xét và đánh giá từ người dùng.

```
# Phương sai

print(np.var(ratings_crime))

#Phương sai lớn cho thấy được sự đánh giá của các phim không đồng đều

# Độ lệch chuẩn

print(np.std(ratings_crime))

#Phân vị

print("Q1 = : ", np.quantile(ratings_crime, 0.25))

print("Q2 = : ", np.quantile(ratings_crime, 0.5))

print("Q3 = : ", np.quantile(ratings_crime, 0.75))
```

```
Q1 = :  8.0

Q2 = :  10.0

Q3 = :  10.0
```

Nhận xét cụ thể từ kết quả trên

Phương sai và Độ lệch chuẩn:

- Phương sai (var):

    - Giá trị 8.0, cho thấy mức độ phân tán trong đánh giá của các bộ phim thuộc thể loại tội phạm không quá lớn nhưng cũng không hoàn toàn tập trung. Các đánh giá có sự khác biệt, nhưng phần lớn nằm gần giá trị trung tâm.
- Độ lệch chuẩn (std):
    - Giá trị 2.83 (căn bậc hai của phương sai). Điều này cho thấy các điểm đánh giá dao động khoảng 2.83 đơn vị xung quanh giá trị trung bình. Sự dao động này là vừa phải, không quá lớn.

Phân vị (Quantiles):

- Q1 (Phân vị thứ 25%): 8.0
    - 25% số đánh giá thấp hơn hoặc bằng 8.0, cho thấy rất ít bộ phim thuộc thể loại tội phạm nhận được đánh giá thấp hơn mức này. Điều này ngụ ý rằng các bộ phim thuộc thể loại này thường được khán giả đánh giá cao.
- Q2 (Phân vị thứ 50% hoặc Median - Trung vị): 10.0
    - Trung vị là 10.0**,** nghĩa là 50% số bộ phim được đánh giá cao nhất có điểm bằng hoặc thấp hơn 10.0. Điều này chứng tỏ các bộ phim tội phạm phần lớn được đánh giá rất cao.
- Q3 (Phân vị thứ 75%): 10.0
    - 75% số đánh giá thấp hơn hoặc bằng 10.0, nghĩa là phần lớn phim thuộc thể loại tội phạm được khán giả đánh giá rất tốt, tập trung ở mức cao nhất.

Tổng kết:

- Đánh giá tích cực vượt trội: Với trung vị và phân vị thứ 75% đều ở mức 10.0, phần lớn phim tội phạm nhận được sự yêu thích cao từ khán giả.
- Mức độ phân tán vừa phải: Phương sai và độ lệch chuẩn không quá lớn, cho thấy các đánh giá phần lớn tập trung ở mức cao, ít có những ý kiến trái chiều hoặc tiêu cực.
- Khuyến nghị: Với kết quả này, thể loại phim tội phạm dường như được yêu thích và đánh giá cao, là một thể loại tiềm năng để phát triển thêm các dự án phim mới.

Qua đó, ta có tỉ lệ reivew theo thời gian qua cột Review_Date, dưới đây là biểu đồ quạt thể hiện :

Hình 2. 4. Biểu đồ tỉ lệ đánh giá các bộ phim Tội phạm sau khi phát sóng

Nhận xét chi tiết biểu đồ

Phân tích từng mốc thời gian:

- **Sau 1 tháng (60.9%):** Đây là nhóm chiếm tỷ lệ cao nhất, cho thấy phần lớn khán giả có xu hướng đưa ra đánh giá muộn. Điều này có thể xuất phát từ việc:
  - Khán giả muốn theo dõi toàn bộ nội dung trước khi đánh giá.
  - Các cuộc thảo luận, bình luận từ cộng đồng sau khi phát sóng có tác động lớn đến thời điểm họ phản hồi.
- **Trong 1 tuần (23.2%):** Gần 1/4 lượng khán giả đánh giá ngay trong tuần đầu tiên. Nhóm này thể hiện:
  - Độ quan tâm cao của những người theo dõi sát sao tác phẩm.
  - Vai trò quan trọng của phản hồi ban đầu trong việc lan tỏa sự chú ý và xây dựng hình ảnh cho tác phẩm.
- **Sau 1 tháng (16.0%):** Dù không chiếm tỷ lệ cao, nhưng nhóm này cho thấy sự ổn định và liên tục của mối quan tâm đến tác phẩm sau khi phát sóng kéo dài. Đây có thể là nhóm khán giả chậm tiếp cận hoặc đánh giá dựa trên phản ứng chung từ cộng đồng.

Xu hướng và nhận định:

- **Tập trung đánh giá muộn:** Với hơn 60.9% đánh giá đến sau 1 tháng, rõ ràng khán giả cần thời gian để tiếp cận, thấu hiểu hoặc cảm nhận tác phẩm một cách trọn vẹn. Điều này cho thấy sự thành công của tác phẩm không chỉ phụ thuộc vào phản hồi ban đầu mà còn vào hiệu ứng dài hạn.

- **Tầm quan trọng của phản hồi sớm:** Gần 1/4 phản hồi đến trong tuần đầu tiên, cho thấy giai đoạn này vẫn là thời điểm "vàng" để định hình dư luận và tạo hiệu ứng lan truyền. Các chiến dịch truyền thông hoặc quảng bá trong tuần đầu tiên cần được đầu tư mạnh mẽ.

Hàm ý chiến lược:

- **Tạo đà lan tỏa ngay từ đầu:** Phản hồi sớm trong tuần đầu tiên đóng vai trò như chất xúc tác, giúp tạo đà cho tác phẩm phát triển hiệu ứng lâu dài. Đầu tư vào quảng bá, tương tác trực tiếp với khán giả trong giai đoạn này là rất quan trọng.
- **Chiến lược dài hạn:** Với đa số khán giả đánh giá sau 1 tháng, các nhà sản xuất nên tiếp tục duy trì nội dung thảo luận, truyền thông và các hoạt động tương tác ngay cả sau khi phát sóng để giữ sức nóng cho tác phẩm.

Kết luận:

Biểu đồ này cho thấy sự khác biệt về thời điểm khán giả phản hồi, từ nhóm đánh giá sớm đến nhóm cần thời gian trải nghiệm. Thành công của một tác phẩm không chỉ nằm ở sự thu hút ban đầu mà còn phụ thuộc vào cách nó duy trì sức hút trong thời gian dài. Việc cân bằng giữa chiến lược "tạo tiếng vang sớm" và "duy trì sức nóng" sẽ quyết định mức độ phổ biến và sức ảnh hưởng của tác phẩm.

2.3.2. Dữ liệu về sản phẩm

Để trang bị các công cụ cần thiết cho việc xử lý, phân tích thống kê và trực quan hóa dữ liệu, đặc biệt trong các lĩnh vực như phân tích dữ liệu lớn hoặc nghiên cứu hành vi người dùng, chúng ta cần import các thư viện sau.

```
import pandas as pd

import numpy as np

import matplotlib.pyplot as plt
```

a) Sản phẩm Công nghệ

Trước hết, chúng ta đọc dữ liệu đánh giá thu thập từ trước của từng sản phẩm để làm nền tảng cho các bước phân tích và đánh giá sau này.

```
#Đọc file csv cong_nghe

Camera_Toan_Cau = pd.read_csv('https://raw.githubusercontent.com/NguyenKhang0062/-n-1/refs/heads/main/e-commerce_data/cong_nghe/shop_Camera_Toan_Cau.csv')

EZVIZ_VN_Authorized_Store = pd.read_csv('https://raw.githubusercontent.com/NguyenKhang0062/-n-1/refs/heads/main/e-commerce_data/cong_nghe/shop_EZVIZ_VN_Authorized_Store.csv')

Zentino = pd.read_csv('https://raw.githubusercontent.com/NguyenKhang0062/-n-1/refs/heads/main/e-commerce_data/cong_nghe/shop_Zentino.csv')
```

Kết quả hiển thị dưới đây:

Cửa hàng "Camera toàn cầu":

```
Camera_Toan_Cau
```

	Review	Rating	Date
0	NaN	1	18 thg 8 2020
1	NaN	1	21 thg 9 2020
2	NaN	1	13 thg 11 2020
3	hàng ok lắm	1	29 thg 12 2023
4	NaN	1	19 thg 10 2021
...	...	...	...
995	NaN	5	26 thg 9 2020
996	Chất lượng camera dỏm, hình delay, khi nhấn qu...	5	20 thg 8 2023
997	quá ngon luôn. tốt trong tầm giá. cảm biến chu...	5	04 thg 12 2020
998	Khi nhấn vô nói rè và ồ kinh khủng	5	21 thg 5 2021
999	NaN	5	28 thg 8 2020

1000 rows × 3 columns

Cửa hàng "Ezviz VN Authorized Store":

```
[ ] EZVIZ_VN_Authorized_Store
```

	Review	Rating	Date
0	NaN	1	26 thg 4 2024
1	NaN	3	17 thg 6 2024
2	NaN	3	16 thg 1 2024
3	Giám sát trong nhà đáng tin cậy, Kết nối mượt ...	3	17 thg 9 2024
4	sản phẩm như mô tả. giao hàng nhanh giá cả hợp...	3	18 thg 8 2024
...	...	...	...
993	NaN	5	18 thg 7 2024
994	Kết nối:ok\n Giá Cả:ok\n Độ phân giải:ok	5	22 thg 6 2024
995	NaN	5	23 thg 7 2024
996	NaN	5	11 thg 7 2024
997	Âm thanh hai chiều tiện lợi, Thiết kế chắc chắ...	5	15 thg 7 2024

998 rows × 3 columns

Cửa hàng "Zentino":

```
[ ] Zentino.head()
```

	Review	Rating	Date
0	ok.hang dep	1	28 thg 12 2022
1	NaN	1	22 thg 1 2022
2	Đã mua rất nhiều lần từ shop này, để đấy không...	1	18 thg 7 2023
3	NaN	1	15 thg 11 2023
4	NaN	1	11 thg 1 2022

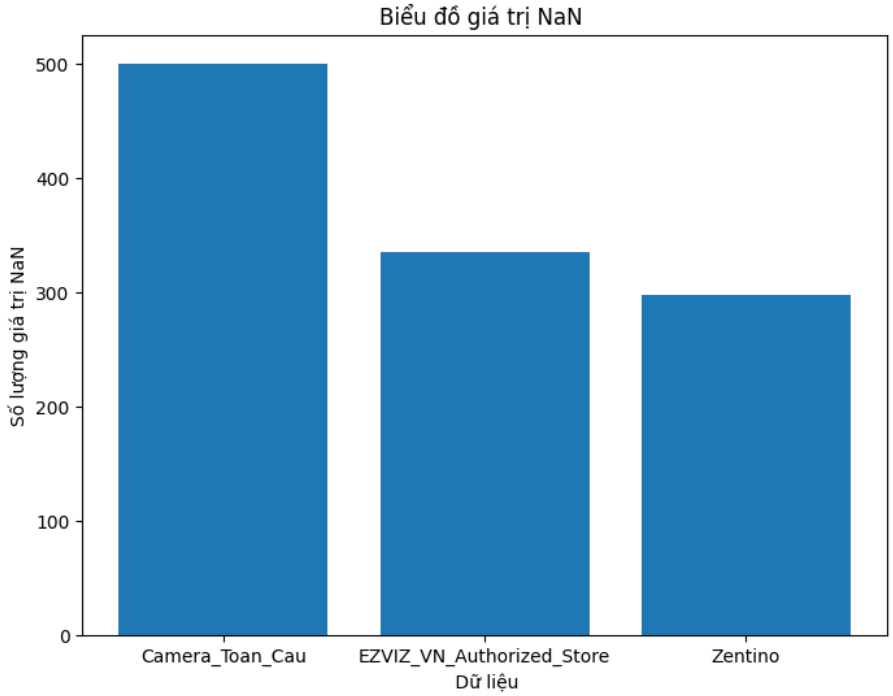
Tiếp theo, ta tiến hành kiểm tra các giá trị bị thiếu (null) trong dữ liệu.

```
#đếm giá trị NaN
```

```
NaN_data_1 = Camera_Toan_Cau['Review'].isna().sum()

NaN_data_2 = EZVIZ_VN_Authorized_Store['Review'].isna().sum()

NaN_data_3 = Zentino['Review'].isna().sum()

print(NaN_data_1)

print(NaN_data_2)

print(NaN_data_3)
```

```
500

335

297
```

Dựa trên kết quả các giá trị bị thiếu trong dữ liệu, chúng ta sẽ vẽ biểu đồ để trực quan hóa và so sánh mức độ thiếu dữ liệu giữa ba cửa hàng.



Hình 2. 5. Biểu đồ giá trị bị thiếu (NaN) của sản phẩm công nghệ

Nhận xét về biểu đồ

Dữ liệu thiếu:

- Biểu đồ thể hiện số lượng các giá trị "NaN" (Not a Number - không phải là số) trong từng tập dữ liệu (Camera_Toan_Cau, EZVIZ_VN_Authorized_Store, Zentino).

Phân bố không đồng đều:

- Số lượng giá trị NaN trong các tập dữ liệu khác nhau. Camera_Toan_Cau có số lượng giá trị NaN cao nhất, tiếp theo là EZVIZ_VN_Authorized_Store và Zentino.

Nhận xét chi tiết:

- Vấn đề chất lượng dữ liệu: Sự xuất hiện của giá trị NaN cho thấy có vấn đề trong quá trình thu thập, xử lý hoặc lưu trữ dữ liệu.
- Các nguyên nhân có thể bao gồm:
    - Lỗi nhập liệu: Dữ liệu được nhập sai hoặc thiếu.
    - Lỗi trong quá trình tính toán: Có lỗi trong các công thức tính toán dẫn đến kết quả không hợp lệ.
    - Dữ liệu bị thiếu: Một số giá trị không được thu thập hoặc bị mất trong quá trình xử lý.
- Ảnh hưởng đến phân tích: Sự tồn tại của giá trị NaN có thể ảnh hưởng đến kết quả phân tích dữ liệu. Nếu không được xử lý đúng cách, giá trị NaN có thể làm sai lệch các kết quả thống kê và làm giảm độ tin cậy của mô hình.
- Cần xử lý dữ liệu: Để có thể phân tích dữ liệu một cách chính xác, cần phải xử lý các giá trị NaN.
- Các phương pháp xử lý có thể bao gồm:
    - Xóa bỏ các dòng chứa giá trị NaN: Nếu số lượng dòng chứa giá trị NaN không quá lớn.
    - Điền giá trị: Điền vào các giá trị NaN bằng các giá trị trung bình, trung vị, hoặc dự đoán dựa trên các giá trị khác.
    - Phân loại riêng các giá trị NaN: Xét riêng các giá trị NaN như một nhóm dữ liệu đặc biệt.

Kết hợp dữ liệu đánh giá (Rating) từ ba cửa hàng khác nhau vào một mảng duy nhất:

```
Rating_1=
np.concatenate([np.array(Camera_Toan_Cau['Rating']),np.array(EZVIZ_VN_Authorized_Sto
re['Rating']),np.array(Zentino['Rating'])])
Rating_1
```

Ta tiến hành tính số đánh giá là 5 sao, dưới 5 sao và số sao trung bình của từng sản phẩm để có cái nhìn sâu sắc hơn về chất lượng của từng sản phẩm:

```
# Đếm số lượng Rating là 5
print("Camera_Toan_Cau:",Camera_Toan_Cau[Camera_Toan_Cau['Rating'] > 4].shape[0])
print("EZVIZ_VN_Authorized_Store:",EZVIZ_VN_Authorized_Store[EZVIZ_VN_Authorized_Sto
re['Rating'] > 4].shape[0])
print("Zentino:",Zentino[Zentino['Rating'] > 4].shape[0])
Camera_Toan_Cau: 901
EZVIZ_VN_Authorized_Store: 976
```

```
Zentino: 565
```

```
# Đếm số lượng  Rating dưới 5
print("Camera_Toan_Cau:",Camera_Toan_Cau[Camera_Toan_Cau['Rating'] < 4].shape[0])
print("EZVIZ_VN_Authorized_Store:",EZVIZ_VN_Authorized_Store[EZVIZ_VN_Authorized_Store['Rating'] < 4].shape[0])
print("Zentino:",Zentino[Zentino['Rating'] < 4].shape[0])
```

```
Camera_Toan_Cau: 28
EZVIZ_VN_Authorized_Store: 9
Zentino: 8
```

```
# Tính trung bình số sao cho mỗi sản phẩm
print("Camera_Toan_Cau:",Camera_Toan_Cau["Rating"].mean())
print("EZVIZ_VN_Authorized_Store:",EZVIZ_VN_Authorized_Store["Rating"].mean())
print("Zentino:",Zentino["Rating"].mean())
```

```
Camera_Toan_Cau: 4.858
EZVIZ_VN_Authorized_Store: 4.966933867735471
Zentino: 4.935042735042735
```

Nhận xét tổng quan

Kết quả trung bình số sao:

- Camera Toàn Cầu: 4.858

  Điểm số khá cao, thể hiện sự hài lòng lớn từ khách hàng. Tuy nhiên, thấp hơn một chút so với các cửa hàng còn lại, có thể là do một số yếu tố như dịch vụ, chất lượng sản phẩm, hoặc trải nghiệm mua sắm.

- EZVIZ VN Authorized Store: 4.967

  Đây là cửa hàng có điểm trung bình cao nhất, gần đạt mức tối đa (5.0). Điều này cho thấy mức độ hài lòng cực kỳ cao, có thể nhờ vào chất lượng sản phẩm, uy tín thương hiệu, hoặc dịch vụ khách hàng tốt.

- Zentino: 4.935

  Điểm số trung bình của Zentino cũng rất ấn tượng, chỉ kém một chút so với EZVIZ. Điều này cho thấy họ duy trì được sự ổn định về chất lượng và trải nghiệm khách hàng.

So sánh tổng quan:

- Tất cả các cửa hàng đều có điểm số vượt mức 4.8, phản ánh sự hài lòng lớn từ khách hàng.
- EZVIZ VN Authorized Store dẫn đầu, nhưng khoảng cách giữa ba cửa hàng không quá lớn, cho thấy sự cạnh tranh cao trong việc duy trì chất lượng sản phẩm và dịch vụ.

Phương pháp cải thiện:

- Camera Toàn Cầu có thể xem xét phản hồi từ khách hàng để cải thiện thêm dịch vụ hoặc chất lượng, nhằm thu hẹp khoảng cách với EZVIZ và Zentino.
- Cả ba cửa hàng nên tiếp tục duy trì các tiêu chuẩn cao để giữ vững lòng tin từ khách hàng, đồng thời hướng đến việc cải thiện các khía cạnh nhỏ nhất để nâng cao điểm số trung bình.

Kết luận:

- Với mức độ đánh giá rất cao trên cả ba cửa hàng, điều này cho thấy chất lượng sản phẩm/dịch vụ mà các cửa hàng này cung cấp đều rất tốt.
- Tuy nhiên, việc giữ vững và cải thiện trải nghiệm khách hàng là yếu tố quan trọng để duy trì sự hài lòng lâu dài.

Tiếp tục, ta tính phương sai, độ lệch chuẩn để đánh giá mức độ đồng đều trong việc đánh giá các sản phẩm, từ đó giúp người phân tích hiểu rõ hơn về sự biến động trong nhận xét và đánh giá từ người dùng.

```
# Phương sai

print(np.var(Rating_1))

#Phương sai lớn cho thấy được sự đánh giá của các sản phẩm không đồng đều

# Độ lệch chuẩn

print(np.std(Rating_1))
```

```
0.16160880374358746

0.4020059747610568
```

Nhận xét tổng quát:

Kết quả tính toán:

- Phương sai (variance):
    - Phương sai đo lường mức độ phân tán của dữ liệu xung quanh giá trị trung bình.
    - Giá trị phương sai lớn cho thấy sự đánh giá của các phim trong tập dữ liệu có sự không đồng đều, nghĩa là có nhiều mức độ khác biệt rõ rệt giữa các sản phẩm.
- Độ lệch chuẩn (standard deviation):

- Độ lệch chuẩn giúp diễn giải sự phân tán của dữ liệu dễ hiểu hơn vì nó sử dụng đơn vị gốc của dữ liệu.
- Với kết quả 0.4020, có thể thấy rằng sự chênh lệch giữa các giá trị đánh giá không quá lớn (vẫn khá gần giá trị trung bình). Tuy nhiên, mức này cũng thể hiện rằng không phải tất cả các sản phẩm đều được đánh giá đồng đều.

Phân tích chi tiết:

- Mức độ đồng đều của đánh giá:
  - Với độ lệch chuẩn thấp (0.402) và phương sai tương đối nhỏ, điều này cho thấy hầu hết các đánh giá đều tập trung gần giá trị trung bình, phản ánh mức độ ổn định nhất định trong đánh giá của người dùng.
  - Tuy nhiên, vẫn có một số sản phẩm được đánh giá chênh lệch rõ rệt so với số đông (dựa vào phương sai lớn hơn mức hoàn toàn đồng nhất).
- Ý nghĩa thực tế:
  - Một số sản phẩm có thể nhận được đánh giá rất cao hoặc rất thấp so với trung bình, cho thấy sự khác biệt trong chất lượng hoặc cách người dùng cảm nhận về chúng.
  - Điều này có thể do chất lượng, hình thức, hoặc cách dùng của từng sản phẩm khác nhau, dẫn đến sự đa dạng trong đánh giá.

Kết luận và nhận xét:

Sự khác biệt giữa các đánh giá là có, nhưng không quá lớn (dựa trên độ lệch chuẩn 0.402). Phần lớn các sản phẩm có đánh giá gần với giá trị trung bình, điều này thể hiện sự đồng nhất tương đối trong cảm nhận của người dùng.

```
#Phân vị

print("Q1 = : ", np.quantile(Rating_1, 0.25))

print("Q2 = : ", np.quantile(Rating_1, 0.5))

print("Q3 = : ", np.quantile(Rating_1, 0.75))
```

```
Q1 = :  5.0

Q2 = :  5.0

Q3 = :  5.0
```

Nhận xét tổng quan

Ý nghĩa của phân vị:

- Q1 (Phân vị thứ nhất): Giá trị ở vị trí 25% của tập dữ liệu (một phần tư dữ liệu thấp nhất).

- Kết quả Q1 = 5.0: 25% đánh giá thấp nhất trong tập dữ liệu vẫn đạt điểm 5.0, cho thấy rất ít hoặc không có đánh giá thấp hơn mức này.

- Q2 (Phân vị thứ hai - Trung vị): Giá trị ở giữa tập dữ liệu (50%).
  - Kết quả Q2 = 5.0: Một nửa số đánh giá của tập dữ liệu đạt điểm 5.0, nghĩa là số đông người mua đánh giá ở mức tối đa.
- Q3 (Phân vị thứ ba): Giá trị ở vị trí 75% của tập dữ liệu (một phần tư cao nhất).
  - Kết quả Q3 = 5.0: 75% đánh giá cao nhất vẫn đạt mức 5.0, cho thấy phần lớn các đánh giá đều đạt mức tối đa.

Nhận xét:

- Tập dữ liệu tập trung mạnh vào giá trị 5.0:
- Kết quả cho thấy sự đồng nhất gần như tuyệt đối trong đánh giá, khi hầu hết hoặc toàn bộ các đánh giá đều đạt mức cao nhất (5.0).
- Điều này phản ánh chất lượng sản phẩm rất tốt hoặc mức độ hài lòng rất cao từ người mua.

b) Sản phẩm Mỹ phẩm (Sữa rửa mặt)

Trước tiên, ta cần đọc dữ liệu đánh giá thu thập từ trước để làm cơ sở cho các bước phân tích và đánh giá tiếp theo.

```
Guardian = pd.read_csv('https://raw.githubusercontent.com/NguyenKhang0062/-n-1/refs/heads/main/e-commerce_data/my_pham/shop_Guardian.csv')

HadaLabo = pd.read_csv('https://raw.githubusercontent.com/NguyenKhang0062/-n-1/refs/heads/main/e-commerce_data/my_pham/shop_HadaLabo.csv')

KITY_COSMETIC = pd.read_csv('https://raw.githubusercontent.com/NguyenKhang0062/-n-1/refs/heads/main/e-commerce_data/my_pham/shop_KITY%20COSMETIC.csv')
```

Kết quả hiển thị dữ liệu đánh giá của từng cửa hàng:

Cửa hàng “Guardian”:

Guardian

	Review	Rating	Date
0	NaN	1	04 thg 8 2023
1	NaN	1	26 thg 12 2023
2	Chất lượng:ko ôkê \r\n\r\n Hiệu quả:\r\nko ôkê...	1	21 thg 12 2023
3	NaN	1	20 thg 12 2022
4	Khôi phục cân bằng tự nhiên của da, dùng 1 năm...	1	26 thg 12 2023
...	...	...	...
1005	hàng đóng gói kĩ càng chắc chắn nhưng mà vỏ tu...	5	26 thg 8 2023
1006	Chào bạn, cảm ơn bạn đã chia sẻ trải nghiệm và...	5	21 thg 2 2023
1007	Vỏ ngoài đẹp, hộp đẹp nhưng mà không có chống ...	5	15 thg 12 2022
1008	NaN	5	27 thg 10 2022
1009	NaN	5	16 thg 10 2022

1010 rows × 3 columns

Cửa hàng “Hadalabo”:

[ ] HadaLabo

	Review	Rating	Date
0	NaN	1	28 thg 3 2022
1	Hyaluronic acid giúp cung cấp độ ẩm tối ưu, Da...	1	22 thg 9 2023
2	cảm ơn shop nhiều ạ. Sản phẩm săn sale được gi...	1	19 thg 11 2022
3	SRM thích lắm nha, đúng là rửa xong không có k...	1	19 thg 1 2022
4	Mang lại mùi hương dễ chịu và tươi mát, Hoàn h...	1	24 thg 7 2024
...	...	...	...
1013	NaN	5	06 thg 6 2022
1014	NaN	5	16 thg 12 2022
1015	Đã nhận hàng. Hàng y hình. Mình dùng cảm thấy ...	5	20 thg 12 2022
1016	giao hàng nhanh, đúng mẫu. sau khi rửa mặt thì...	5	12 thg 8 2024
1017	săn sale giá 3 chai 140k luôn cả ship, hàng ma...	5	22 thg 9 2022

1018 rows × 3 columns

Tiếp theo, ta tiến hành kiểm tra các giá trị bị thiếu (null) trong dữ liệu.

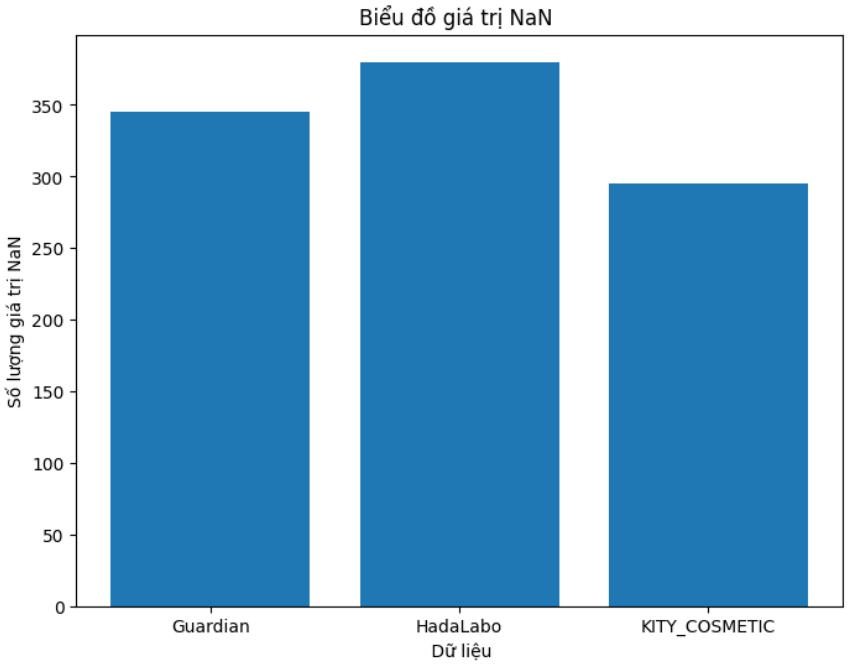
```
#đếm giá trị NaN

NaN_data_4 = Guardian['Review'].isna().sum()

NaN_data_5 = HadaLabo['Review'].isna().sum()

NaN_data_6 = KITY_COSMETIC['Review'].isna().sum()

print(NaN_data_4)

print(NaN_data_5)

print(NaN_data_6)
```

```
345

380

295
```

Dựa trên kết quả các giá trị bị thiếu trong dữ liệu, chúng ta sẽ vẽ biểu đồ để trực quan hóa và so sánh mức độ thiếu dữ liệu giữa ba cửa hàng.

```
x = ['Guardian','HadaLabo','KITY_COSMETIC']

y = [NaN_data_4,NaN_data_5,NaN_data_6 ]

plt.figure(figsize=(8, 6))

plt.bar(x, y)

plt.xlabel(" Dữ liệu")

plt.ylabel("Số lượng giá trị NaN")

plt.title("Biểu đồ giá trị NaN ")

plt.show()
```

Kết quả hiển thi:



Hình 2. 6. Biểu đồ giá trị bị thiếu (NaN) của sản phẩm mỹ phẩm (sữa rửa mặt)

Nhận xét biểu đồ

Dữ liệu thiếu:

- Biểu đồ này cho thấy số lượng các giá trị "NaN" (Not a Number - không phải là số) trong ba tập dữ liệu khác nhau: Guardian, HadaLabo và KITY_COSMETIC.

Phân bố không đồng đều:

- Số lượng giá trị NaN trong mỗi tập dữ liệu là khác nhau. HadaLabo có số lượng giá trị NaN cao nhất, tiếp theo là Guardian và KITY_COSMETIC.

Nhận xét chi tiết hơn:

- Vấn đề chất lượng dữ liệu: Sự xuất hiện của giá trị NaN cho thấy có vấn đề trong quá trình thu thập, xử lý hoặc lưu trữ dữ liệu.
- Các nguyên nhân có thể bao gồm:
    - Lỗi nhập liệu: Dữ liệu được nhập sai hoặc thiếu.
    - Lỗi trong quá trình tính toán: Có lỗi trong các công thức tính toán dẫn đến kết quả không hợp lệ.
    - Dữ liệu bị thiếu: Một số giá trị không được thu thập hoặc bị mất trong quá trình xử lý.
- Ảnh hưởng đến phân tích: Sự tồn tại của giá trị NaN có thể làm sai lệch kết quả phân tích dữ liệu. Nếu không được xử lý đúng cách, giá trị NaN có thể làm giảm độ tin cậy của các kết quả thống kê và mô hình.
- Cần xử lý dữ liệu: Để có thể phân tích dữ liệu một cách chính xác, cần phải xử lý các giá trị NaN.
- Các phương pháp xử lý có thể bao gồm:
    - Xóa bỏ các dòng chứa giá trị NaN: Nếu số lượng dòng chứa giá trị NaN không quá lớn.
    - Điền giá trị: Điền vào các giá trị NaN bằng các giá trị trung bình, trung vị, hoặc dự đoán dựa trên các giá trị khác.
    - Phân loại riêng các giá trị NaN: Xét riêng các giá trị NaN như một nhóm dữ liệu đặc biệt.

Kết hợp dữ liệu đánh giá (Rating) từ ba cửa hàng khác nhau vào một mảng duy nhất:

```
Rating_2=
np.concatenate([np.array(Guardian['Rating']),np.array(HadaLabo['Rating']),np.array(K
ITY_COSMETIC['Rating'])])
Rating_2
```

Để có cái nhìn sâu sắc hơn về chất lượng của từng cửa hàng, ta tiến hành tính số đánh giá là 5 sao, dưới 5 sao và số sao trung bình của từng cửa hàng.

```
# Đếm số lượng Rating là 5
print("Guardian:",Guardian[Guardian['Rating'] > 4].shape[0])
print("HadaLabo:",HadaLabo[HadaLabo['Rating'] > 4].shape[0])
print("KITY_COSMETIC:",KITY_COSMETIC[KITY_COSMETIC['Rating'] > 4].shape[0])
```

```
Guardian: 944
HadaLabo: 873
```

```
KITY_COSMETIC: 664
```

```
#Dưới 5 sao
print("Guardian:",Guardian[Guardian['Rating'] < 4].shape[0])
print("HadaLabo:",HadaLabo[HadaLabo['Rating'] < 4].shape[0])
print("KITY_COSMETIC:",KITY_COSMETIC[KITY_COSMETIC['Rating'] < 4].shape[0])
```

```
Guardian: 30
HadaLabo: 70
KITY_COSMETIC: 5
```

```
# số sao trung bình
print("Guardian:",Guardian["Rating"].mean())
print("HadaLabo:",HadaLabo["Rating"].mean())
print("KITY_COSMETIC:",KITY_COSMETIC["Rating"].mean())
```

```
Guardian: 4.872277227722773
HadaLabo: 4.724950884086444
KITY_COSMETIC: 4.963181148748159
```

Nhận xét tổng quan qua số đánh giá sao trung bình:

- **KITY_COSMETIC (4.96):** Đây là thương hiệu có điểm đánh giá cao nhất trong ba thương hiệu, với sự đánh giá rất tích cực từ người tiêu dùng. Điểm số này chỉ ra rằng sản phẩm của KITY_COSMETIC có sự thu hút mạnh mẽ và khả năng đáp ứng nhu cầu của khách hàng rất tốt. Các yếu tố như chất lượng vượt trội, hiệu quả rõ rệt, hoặc thậm chí chiến lược marketing hiệu quả có thể là những lý do chính thúc đẩy sự hài lòng của khách hàng. Với mức điểm này, KITY_COSMETIC đang xây dựng một hình ảnh rất mạnh mẽ và có thể dễ dàng chiếm lĩnh thị trường trong phân khúc của mình.
- **Guardian (4.87):** Mặc dù điểm đánh giá của Guardian thấp hơn so với KITY COSMETIC, nhưng đây vẫn là một mức điểm rất ấn tượng, phản ánh sự hài lòng cao của người dùng. Guardian có thể là một thương hiệu có sự ổn định trong chất lượng sản phẩm, nhưng có thể thiếu yếu tố nổi bật hoặc sự khác biệt rõ rệt so với các đối thủ. Tuy nhiên, với điểm số này, Guardian vẫn giữ vững được vị trí của mình trong tâm trí người tiêu dùng, nhờ vào sự tin cậy và uy tín đã được xây dựng qua thời gian.

- **HadaLabo (4.72):** Mặc dù vẫn đạt điểm đánh giá cao, nhưng HadaLabo có vẻ gặp phải một số thử thách nhất định khi so với hai đối thủ còn lại. Điểm số này có thể phản ánh một sự thiếu đột phá hoặc sự không đồng đều trong chất lượng sản phẩm. HadaLabo có thể có những sản phẩm không hoàn toàn phù hợp với nhu cầu của một số nhóm khách hàng, hoặc đôi khi gặp phải những vấn đề về giá trị so với mức giá. Tuy nhiên, mức điểm này vẫn chỉ ra rằng HadaLabo vẫn duy trì được sự tin tưởng nhất định trong lòng người tiêu dùng.

Tổng kết:

Mặc dù tất cả các thương hiệu đều đạt được điểm số khá cao, KITY_COSMETIC nổi bật với sự xuất sắc về chất lượng sản phẩm, tiếp theo là Guardian với sự ổn định và uy tín, và HadaLabo có thể cần một chút cải tiến để cải thiện sự hài lòng từ khách hàng, đặc biệt là trong việc tạo sự khác biệt so với các thương hiệu khác.

Sau đây, chúng ta sẽ tính phương sai và độ lệch chuẩn để đánh giá mức độ nhất quán trong các đánh giá sản phẩm, giúp người phân tích hiểu rõ hơn về sự biến động trong nhận xét và phản hồi của người dùng.

```
# Phương sai

print(np.var(Rating_2))

#Phương sai lớn cho thấy được sự đánh giá của các sản phẩm không đồng đều

# Độ lệch chuẩn

print(np.std(Rating_2))
```

```
0.39468744511520365

0.6282415499751697
```

Nhận xét tổng quan:

- **Phương sai (0.3947):** Phương sai cho thấy mức độ phân tán của các giá trị trong bộ dữ liệu. Phương sai càng lớn, sự phân tán giữa các giá trị càng rộng, tức là sự đánh giá của các sản phẩm không đồng đều và có sự khác biệt rõ rệt. Trong trường hợp này, phương sai 0.3947 cho thấy mặc dù điểm đánh giá trung bình khá cao, nhưng các đánh giá của người tiêu dùng vẫn có sự dao động đáng kể, có thể xuất phát từ các yếu tố như sự không đồng đều trong chất lượng sản phẩm, sự khác biệt trong trải nghiệm người dùng hoặc thậm chí các yếu tố ngoại cảnh như chiến lược marketing.
- **Độ lệch chuẩn (0.6282):** Độ lệch chuẩn đo lường mức độ biến động của các điểm đánh giá so với giá trị trung bình. Độ lệch chuẩn 0.6282 cho thấy mức độ phân tán của các giá trị xung quanh điểm trung bình là khá lớn. Điều này có nghĩa là mặc dù đa số các sản phẩm có điểm đánh giá khá cao, nhưng vẫn có những sản phẩm nhận được điểm thấp hoặc có sự đánh giá trái chiều, gây nên sự phân tán

này. Một độ lệch chuẩn cao có thể là dấu hiệu của sự không nhất quán trong chất lượng sản phẩm hoặc sự khác biệt về kỳ vọng của người tiêu dùng.

Kết luận:

Mặc dù điểm đánh giá trung bình của các sản phẩm khá cao, nhưng với phương sai và độ lệch chuẩn lớn, chúng ta có thể thấy rằng sự hài lòng của khách hàng không hoàn toàn đồng đều. Điều này chỉ ra rằng có những sản phẩm có thể nhận được phản hồi rất tích cực, nhưng cũng có những sản phẩm không được đánh giá cao bằng, có thể do chất lượng không đồng đều, sự khác biệt trong mong đợi của khách hàng, hoặc các yếu tố khác như dịch vụ khách hàng.

```
#Phân vị

print("Q1 = : ", np.quantile(Rating_2, 0.25))

print("Q2 = : ", np.quantile(Rating_2, 0.5))

print("Q3 = : ", np.quantile(Rating_2, 0.75))
```

```
Q1 = :  5.0

Q2 = :  5.0

Q3 = :  5.0
```

Nhận xét tổng quan:

- Kết quả phân vị đều bằng 5.0, cho thấy rằng hầu hết các sản phẩm đều được đánh giá rất cao và có sự đồng nhất mạnh mẽ trong đánh giá của người tiêu dùng. Không có sự phân tán lớn trong các đánh giá (không có giá trị thấp hay cao cực đoan). Điều này cho thấy rằng, trong bộ dữ liệu này, khách hàng có xu hướng rất hài lòng với các sản phẩm và sự đánh giá của họ khá nhất quán, tập trung chủ yếu vào mức điểm cao (5).
- Điều này có thể phản ánh chất lượng ổn định và tốt của các sản phẩm, hoặc có thể là do các sản phẩm đều được người tiêu dùng yêu thích và hài lòng ở mức độ gần như giống nhau.

c) Sản phẩm Thực phẩm chức năng

Trước tiên, đọc dữ liệu đánh giá đã thu thập được để làm cơ sở cho các bước phân tích và đánh giá tiếp theo.

```
#dữ liệu

Cosin_Store = pd.read_csv('https://raw.githubusercontent.com/NguyenKhang0062/-n-1/refs/heads/main/e-commerce_data/thuc_pham_chuc_nang/shop_Cosin_Store.csv')

DHC = pd.read_csv('https://raw.githubusercontent.com/NguyenKhang0062/-n-1/refs/heads/main/e-commerce_data/thuc_pham_chuc_nang/shop_DHC.csv')
```

```
Konni39 = pd.read_csv('https://raw.githubusercontent.com/NguyenKhang0062/-n-1/refs/heads/main/e-commerce_data/thuc_pham_chuc_nang/shop_Konni39_Thu_Duc.csv')
```

Kết quả hiển thị dữ liệu của từng cửa hàng:

Cửa hàng " Cosin store":

```
Cosin_Store
```

	Review	Rating	Date
0	sản phẩm đúng như mô tả, giao hàng hơi chậm(4 ...	1	03 thg 7 2023
1	ảnnh với video là lấy bịch bé nha chứ bịch t m...	1	09 thg 11 2021
2	NaN	2	22 thg 7 2021
3	NaN	3	15 thg 11 2023
4	NaN	3	14 thg 9 2022
...	...	...	...
995	ban đầu mình uống không sao hết mọi người ơi, ...	5	05 thg 4 2024
996	Sao giao túi 30 viên thời vậy.làm ăn kì vậy	5	06 thg 5 2023
997	Giao nhầm hàng	5	11 thg 11 2022
998	NaN	5	27 thg 12 2023
999	Ok	5	24 thg 11 2022

1000 rows × 3 columns

Cửa hàng "DHC":

```
[ ] DHC
```

	Review	Rating	Date
0	Tốt	1	05 thg 9 2020
1	giao hàng nhanh check in mã vạch ra nhưng chưa...	1	13 thg 12 2022
2	NaN	1	30 thg 11 2021
3	Hàng mới, date xa. Về chất lượng phải uống thử...	1	06 thg 3 2021
4	NaN	1	21 thg 10 2019
...	...	...	...
1150	🍓🍓🍓🍓🍓🍓🍓🍓🍓🍓🍓🍓🍓🍓🍓🍓🍓🍓🍓🍓🍓🍓🍓🍓🍓🍓🍓🍓🍓🍓🍓🍓🍓🍓	5	17 thg 11 2021
1151	San pham giong hinh shop giao dung mau chất lư...	5	18 thg 11 2021
1152	Giao hàng đầy đủ, thời gian giao hàng nhanh, c...	5	17 thg 11 2021
1153	Mới mua lần đầu thấy lazmall nên yên tâm.....	5	17 thg 11 2021
1154	🍓🍓🍓🍓🍓🍓🍓🍓🍓🍓🍓🍓🍓🍓🍓🍓🍓🍓🍓🍓🍓🍓🍓🍓🍓🍓🍓🍓🍓🍓🍓🍓🍓🍓	5	17 thg 11 2021

1155 rows × 3 columns

Cửa hàng "Konni39":

```
Konni39
```

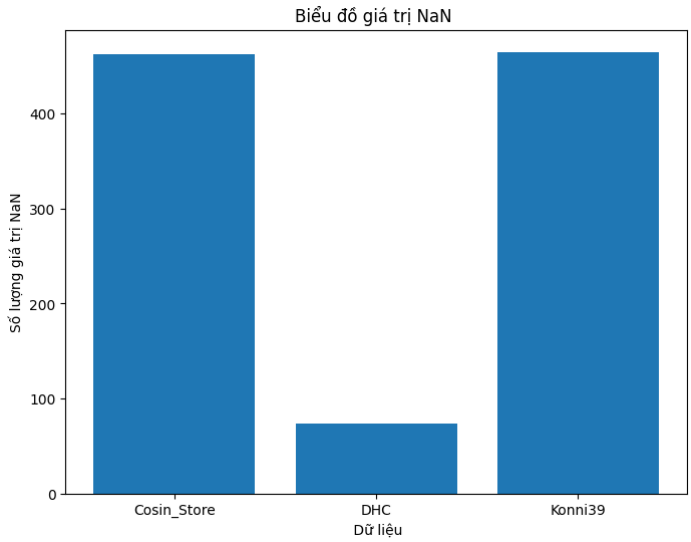
	Review	Rating	Date
0	NaN	1	10 thg 8 2020
1	NaN	1	21 thg 3 2021
2	sản phẩm rất ok.... sẽ ủng hộ shop lâu dài . l...	1	15 thg 11 2021
3	Sản phẩm chất lượng. Giao hàng nhanh chóng. Đá...	3	17 thg 5 2024
4	NaN	3	14 thg 10 2022
...	...	...	...
925	Ko nên mua. Phục vụ như hàng chợ	5	03 thg 2 2023
926	Hàng cận date	5	27 thg 4 2021
927	NaN	5	10 thg 5 2024
928	Đợt này shop giao hàng hơi lâu	5	28 thg 9 2022
929	Em đặt gói 60 ngày 120 viên mà shop gửi không ...	5	11 thg 8 2022

930 rows × 3 columns

Tiếp theo, ta tiến hành kiểm tra các giá trị bị thiếu (null) trong dữ liệu.

```
#đếm giá trị NaN

NaN_data_7 = Cosin_Store['Review'].isna().sum()

NaN_data_8 = DHC['Review'].isna().sum()

NaN_data_9 = Konni39['Review'].isna().sum()

x = ['Cosin_Store','DHC','Konni39']

y = [NaN_data_7,NaN_data_8,NaN_data_9 ]

plt.figure(figsize=(8, 6))

plt.bar(x, y)

plt.xlabel(" Dữ liệu")

plt.ylabel("Số lượng giá trị NaN")

plt.title("Biểu đồ giá trị NaN ")

plt.show()
```

Kết quả hiển thị:



Hình 2. 7. Biểu đồ giá trị bị thiếu (NaN) của sản phẩm thực phẩm chức năng

Kết hợp dữ liệu đánh giá (Rating) từ ba cửa hàng khác nhau vào một mảng duy nhất:

```
Rating_3=
np.concatenate([np.array(Cosin_Store['Rating']),np.array(DHC['Rating']),np.
array(Konni39['Rating'])])

Rating_3
```

Tương tự, để có cái nhìn sâu sắc hơn về chất lượng của sản phẩm từng cửa hàng, ta tiến hành tính số đánh giá là 5 sao, dưới 5 sao và số sao trung bình của từng cửa hàng:

```
# Đếm số lượng Rating là 5
```

```
print("Cosin_Store:",Cosin_Store[Cosin_Store['Rating'] > 4].shape[0])
print("DHC:",DHC[DHC['Rating'] > 4].shape[0])
print("Konni39:",Konni39[Konni39['Rating'] > 4].shape[0])
```

```
Cosin_Store: 977
DHC: 1010
Konni39: 908
```

```
# Đếm số lượng Rating dưới 5
print("Cosin_Store:",Cosin_Store[Cosin_Store['Rating'] < 4].shape[0])
print("DHC:",DHC[DHC['Rating'] < 4].shape[0])
print("Konni39:",Konni39[Konni39['Rating'] < 4].shape[0])
```

```
Cosin_Store: 6
DHC: 64
Konni39: 6
```

```
#số sao trung bình
print("Cosin_Store:",Cosin_Store["Rating"].mean())
print("DHC:",DHC["Rating"].mean())
print("Konni39:",Konni39["Rating"].mean())
```

```
Cosin_Store: 4.966
DHC: 4.770562770562771
Konni39: 4.963440860215054
```

Nhận xét tổng quan qua số sao đánh giá trung bình:

- **Cosin_Store (4.966):** Đây là thương hiệu có điểm đánh giá trung bình cao nhất trong ba thương hiệu, cho thấy sản phẩm của Cosin_Store nhận được sự hài lòng rất lớn từ người tiêu dùng. Với mức điểm gần như tối đa (4.97), Cosin_Store có thể đã tạo ra sự khác biệt rõ rệt trong chất lượng sản phẩm, đáp ứng tốt nhu cầu và kỳ vọng của khách hàng. Điều này có thể do các sản phẩm chất lượng vượt trội, dịch vụ khách hàng tốt, hoặc những đặc điểm đặc biệt khiến người tiêu dùng đánh giá cao.
- **Konni39 (4.963):** Mặc dù điểm đánh giá trung bình của Konni39 thấp hơn một chút so với Cosin Store, nhưng vẫn rất cao và gần như ngang bằng. Điều này cho thấy Konni39 cũng đạt được sự hài lòng rất lớn từ khách hàng, với chất lượng sản phẩm và dịch vụ đáng tin cậy. Mặc dù không đạt mức xuất sắc như

Cosin_Store, nhưng Konni39 vẫn duy trì được sự ủng hộ vững chắc từ người tiêu dùng.

- **DHC (4.770):** Mặc dù có điểm đánh giá trung bình thấp hơn một chút, nhưng 4.77 vẫn là một điểm khá cao, cho thấy rằng sản phẩm của DHC vẫn nhận được sự đánh giá tích cực từ khách hàng. Tuy nhiên, sự khác biệt rõ rệt so với Cosin_Store và Konni39 có thể chỉ ra rằng một số yếu tố như chất lượng sản phẩm không đồng đều, dịch vụ khách hàng chưa đáp ứng kỳ vọng, hoặc có sự cạnh tranh mạnh mẽ từ các thương hiệu khác khiến DHC không thể đạt được mức điểm cao hơn.

Kết luận:

- **Cosin_Store** *và* **Konni39** đều đạt điểm đánh giá rất cao, cho thấy rằng cả hai thương hiệu này đều mang lại sự hài lòng lớn cho khách hàng, với Cosin_Store có phần nổi bật hơn một chút.
- **DHC**, dù vẫn đạt điểm đánh giá khá tốt, nhưng có thể cần chú ý cải thiện các yếu tố như chất lượng sản phẩm, dịch vụ khách hàng, hoặc giá trị so với mức giá để có thể cạnh tranh hiệu quả hơn với các thương hiệu như Cosin_Store và Konni39.

Sau đây, chúng ta sẽ tính phương sai và độ lệch chuẩn để đánh giá mức độ nhất quán trong các đánh giá sản phẩm, giúp người phân tích hiểu rõ hơn về sự biến động trong nhận xét và phản hồi của người dùng.

```
# Phương sai

print(np.var(Rating_3))

#Phương sai lớn cho thấy được sự đánh giá của các sản phẩm không đồng đều

# Độ lệch chuẩn

print(np.std(Rating_3))
```

```
0.25058522836225894

0.5005848862703097
```

Nhận xét tổng quan:

- **Phương sai (0.2506):** Phương sai này khá thấp, cho thấy rằng các giá trị trong bộ dữ liệu không phân tán quá rộng. Mặc dù có sự dao động giữa các điểm đánh giá, nhưng mức độ phân tán không quá lớn. Điều này chỉ ra rằng các sản phẩm trong bộ dữ liệu này có sự đồng nhất tương đối trong mức độ hài lòng của khách hàng, không có sự đánh giá quá khác biệt giữa các sản phẩm.
- **Độ lệch chuẩn (0.5006):** Độ lệch chuẩn đo lường sự phân tán của các điểm đánh giá quanh giá trị trung bình. Với độ lệch chuẩn 0.5006, mức độ biến động giữa các điểm đánh giá không quá cao. Đây là một chỉ số cho thấy sự nhất quán tương

đối trong các đánh giá, tuy nhiên vẫn có một vài sản phẩm nhận được điểm thấp hơn hoặc cao hơn mức trung bình, nhưng sự khác biệt này không quá lớn.

Kết luận:

Mặc dù có sự phân tán giữa các điểm đánh giá, nhưng phương sai và độ lệch chuẩn đều cho thấy mức độ phân tán là vừa phải. Điều này có nghĩa là các sản phẩm trong bộ dữ liệu Rating_3 không có sự biến động quá mạnh trong mức độ hài lòng của khách hàng. Hầu hết các sản phẩm nhận được điểm đánh giá gần nhau, tuy nhiên, vẫn có một số sự khác biệt nhỏ, có thể do sự không đồng đều trong chất lượng sản phẩm hoặc trải nghiệm của người tiêu dùng. Tóm lại, bộ dữ liệu này cho thấy sự đồng nhất khá cao trong sự hài lòng của khách hàng, nhưng vẫn có một số yếu tố gây ra sự khác biệt nhẹ trong các đánh giá.

Phân vị (percentile) là một công cụ thống kê được sử dụng để phân tích và mô tả dữ liệu, với mục đích chính là chia dữ liệu thành các phần bằng nhau dựa trên thứ hạng. Các phân vị được dùng để xác định vị trí của một giá trị trong một tập dữ liệu, hỗ trợ trong việc ra quyết định và phân tích. Phân vị giúp xác định cách các giá trị trong tập dữ liệu được phân bố, chẳng hạn xác định mức độ chênh lệch giữa các nhóm giá trị khác nhau. Sau đây, ta tiến hành tính phân vị của bộ dữ liệu trên:

```
#Phân vị

print("Q1 = : ", np.quantile(Rating_3, 0.25))

print("Q2 = : ", np.quantile(Rating_3, 0.5))

print("Q3 = : ", np.quantile(Rating_3, 0.75))
```

```
Q1 = :  5.0

Q2 = :  5.0

Q3 = :  5.0
```

Nhận xét tổng quan:

Kết quả phân vị cho thấy hầu hết các sản phẩm đều nhận được đánh giá rất cao, với tất cả các phân vị đều bằng 5.0. Điều này chỉ ra rằng các sản phẩm trong bộ dữ liệu này được khách hàng đánh giá rất tích cực và có sự đồng nhất cao trong mức độ hài lòng. Các sản phẩm hầu như không gặp phải đánh giá thấp hay trung bình, và người tiêu dùng đều cảm thấy rất hài lòng với chúng.

2.3.3. Dữ liệu về ngành học

a) Gộp dữ liệu chia làm miền Bắc và miền Nam

```
# Gộp dữ liệu theo miền

schools_data_north <- bind_rows(hust_data, neu_data, qht_data, ptit_data)

schools_data_south <- bind_rows(qhx_data, ueh_data, hcmut_data, tct_data)
```

Hiển thị 5 dòng đầu tiên của dữ liệu miền bắc:

```
# A tibble: 5 × 6
  truong ma_xet_tuyen linh_vuc chi_tieu_2022 chi_tieu_2023 chi_tieu_2024
  <chr>  <chr>        <chr>            <dbl>         <dbl>         <dbl>
1 HUST   BF1          khtn               120            80           160
2 HUST   BF2          sức khỏe           200           200           360
3 HUST   BF-E12       sức khỏe            80            80            40
4 HUST   CH1          khtn               600           520           680
5 HUST   CH2          khtn               120           120           160
```

Hiển thị 5 dòng đầu tiên của dữ liệu miền nam:

```
# A tibble: 5 × 6
  truong ma_xet_tuyen linh_vuc         chi_tieu_2022 chi_tieu_2023 chi_tieu_2024
  <chr>  <chr>        <chr>                    <dbl>         <dbl>         <dbl>
1 QHX    Q11X06       khxh và nhân văn            30             0             0
2 QHX    QHX07        khoa học quản lý            50           110           120
3 QHX    QHX41        khoa học quản lý            50             0             0
4 QHX    QHX08        khxh và nhân văn            65            80            80
5 QHX    QHX09        khxh và nhân văn            55            55            50
```

Thống kê mô tả

Miền Bắc:

```
> summary(schools_data_north)
 Truong            ma_xet_tuyen       linh_vuc          chi_tieu_2022
 Length: 167       Length: 167        Length: 167       Min. : 0.0
 Class: character  Class: character   Class: character   1st Qu. : 55.0
 Mode: character   Mode: character   Mode: character   Median : 80.0
                                                        Mean : 105.2
                                                        3rd Qu. : 120.0
                                                        Max. : 600.0
 chi_tieu_2023     chi_tieu_2024
 Min. : 0          Min. : 0.0
 1st Qu. : 55 Median  1st Qu. : 60.0
 : 80              Median : 80.0
```

```
Mean : 101            Mean : 109.1 3
3rd Qu. : 120         rd Qu. : 120.0
Max. : 520            Max. : 680.0
```

Miền Nam:

```
> summary(schools_data_south)
   truong          ma_xet_tuyen         linh_vuc         chi_tieu_2022     chi_tieu_2023
Length:221         Length:221         Length:221         Min.   :   0.00   Min.   :   0.00
Class :character   Class :character   Class :character   1st Qu.:  40.00   1st Qu.:  40.00
Mode  :character   Mode  :character   Mode  :character   Median :  60.00   Median :  65.00
                                                         Mean   :  95.79   Mean   :  98.99
                                                         3rd Qu.: 110.00   3rd Qu.: 110.00
                                                         Max.   :1050.00   Max.   :1190.00
chi_tieu_2024
Min.   :   0.0
1st Qu.:  50.0
Median :  80.0
Mean   : 105.7
3rd Qu.: 110.0
Max.   :1180.0
```

Nhận xét

Đối với miền Bắc:

Trung bình chỉ tiêu tuyển sinh có xu hướng tăng nhẹ qua các năm:

- 2022: Trung bình ~105.2
- 2023: Trung bình ~101 → giảm nhẹ so với 2022
- 2024: Trung bình ~109.1 → tăng mạnh hơn so với 2023 và cao nhất trong 3 năm

Có thể thấy năm 2023 là năm có xu hướng siết chặt chỉ tiêu hơn một chút, nhưng đến 2024 lại tăng mạnh trở lại.

- 1st Quartile (Q1 - 25%): Ở cả 3 năm đều là khoảng 55–60, cho thấy 25% các ngành có chỉ tiêu thấp và khá ổn định.
- Median (Q2 - 50%): Giữ ở mức 80 trong cả 3 năm, điều này chứng tỏ phân bố chỉ tiêu tương đối đều và ổn định ở nhóm giữa.
- 3rd Quartile (Q3 - 75%): Ổn định quanh mức 120, thể hiện rằng 75% các ngành có chỉ tiêu không vượt quá 120, khá đồng đều qua các năm.

Max chỉ tiêu:

- 2022: 600
- 2023: 520 → giảm mạnh
- 2024: 680 → tăng vọt

Kết luận:

- Điều này cho thấy một số ngành đặc biệt có sự biến động lớn, có thể là những ngành hot hoặc có sự tái cơ cấu mạnh trong tuyển sinh.
- Phân bố chỉ tiêu ổn định ở mức trung bình, đặc biệt là các phân vị như Q1, Median, Q3.
- Tuy nhiên, sự biến động chủ yếu xảy ra ở các giá trị cực đại, cho thấy có sự thay đổi chiến lược trong tuyển sinh đối với một số ngành cụ thể.
- Tăng trưởng chỉ tiêu năm 2024 phản ánh xu hướng mở rộng quy mô tuyển sinh sau năm 2023 có phần thắt chặt.

Đối với miền Nam:

Trung bình (Mean) chỉ tiêu tuyển sinh tăng dần theo từng năm:

- 2022: ~95.79
- 2023: ~98.99
- 2024: ~105.7

Cho thấy các trường phía Nam có xu hướng tăng dần quy mô tuyển sinh, đặc biệt là từ 2023 → 2024 tăng khá rõ.

- Q1 và Median tăng cho thấy nhiều ngành/trường đang nâng mức tuyển sinh cơ bản.
- Q3 giữ nguyên ở 110 phân bố vẫn tương đối ổn định, chỉ một số ngành top có tăng mạnh

Mã chỉ tiêu:

- 2022: 1,050
- 2023: 1,190
- 2024: 1,180

Một vài ngành/trường có quy mô tuyển sinh rất lớn, có thể là các ngành kinh tế, CNTT hoặc quản trị tại các đại học lớn.

Cả 3 năm đều có ngành có chỉ tiêu bằng 0, cho thấy:

- Ngành đó ngừng tuyển sinh tạm thời
- Hoặc là có trong danh sách nhưng không được phê duyệt chỉ tiêu

Kết luận:

- Chỉ tiêu tuyển sinh khu vực phía Nam có xu hướng tăng dần qua từng năm.
- Phân vị giữa và thấp đều tăng nhẹ → phản ánh sự mở rộng đồng đều, không chỉ tập trung vào một số ngành hot.

- Max tăng mạnh cho thấy vẫn có sự phân hóa rõ rệt giữa các trường/ngành.
- Dữ liệu ổn định và thể hiện chiến lược phát triển quy mô tuyển sinh ở phía Nam khá rõ nét.

CHƯƠNG 3: TRỰC QUAN HÓA DỮ LIỆU

3.1. GIỚI THIỆU VỀ TRỰC QUAN HÓA DỮ LIỆU

3.1.1. Trực quan hóa dữ liệu là gì ?

Điều gì có thể diễn đạt những trải nghiệm, ý tưởng và đặc biệt những con số "biết nói" tốt hơn ngôn từ đó chính là những hình ảnh bức tranh bởi hình ảnh có thể mang lại sự phong phú, sâu sắc và phản ánh được một phần của thế giới và cảm xúc của con người một cách không thể diễn tả hoàn toàn bằng từ ngữ. Trong một số trường hợp, hình ảnh có thể tạo ra một ấn tượng sâu sắc và không thể nào được truyền đạt bằng từ ngữ. Cũng như trích dẫn của Ludwig Wittgenstein, một nhà triết học Áo từng phát biểu rằng *"Một bức tranh, một bức ảnh có thể nói lên những điều không thể diễn đạt được bằng ngôn từ"*.

Hiểu được tầm quan trọng của việc diễn đạt những con số đã được tổng hợp từ nhiều nguồn dữ liệu khác nhau là chìa khóa để có thể truyền tải thông tin một cách rõ ràng và hiệu quả. Những thế hệ đi trước đã dần dần sáng tạo ra khái niệm biểu đồ, đồ thị, bản đồ để từ đó giúp chúng ta giải được bài toán về sự giới hạn của ngôn từ đối với những con số của thống kê. Biểu đồ, đồ thị, bản đồ có thể đưa tới cho người xem, người đọc hiểu rõ hơn về bộ dữ liệu đó phải làm thế nào để họ hiểu, họ nắm được xem bộ dữ liệu đó nói về cái gì, miêu tả cái gì và rút ra điều gì từ bộ dữ liệu đó. Chữ viết hoàn toàn có thể sử dụng cho những bộ dữ liệu đơn giản nhưng không thể thực hiện được việc truyền tải nội dung về xu hướng và dao động của dữ liệu. Khi đó các công cụ về trực quan hóa dữ liệu trở nên cần thiết hơn bao giờ hết.

Vậy trực quan hóa dữ liệu là gì? Trực quan hóa dữ liệu là quá trình biểu diễn thông tin và dữ liệu dưới dạng hình ảnh hoặc đồ họa để giúp người xem hiểu và suy luận dữ liệu một cách dễ dàng và nhanh chóng hơn. Bằng cách sử dụng các biểu đồ, bản đồ, biểu đồ tia, biểu đồ cột và các công cụ trực quan hóa khác, trực quan dữ liệu giúp chuyển đổi dữ liệu phức tạp thành các hình ảnh đơn giản và dễ hiểu. Kết hợp với R là một ngôn ngữ lập trình mạnh mẽ và phổ biến trong phân tích dữ liệu và trực quan hóa. Nó cung cấp nhiều gói (packages) và hàm (functions) mạnh mẽ để tạo ra các biểu đồ và đồ thị chất lượng cao. Một số gói phổ biến trong R dùng cho trực quan hóa dữ liệu như : ggplot2, plotly,…

3.1.2. Mục đích của trực quan hóa dữ liệu

Trực quan hóa dữ liệu đóng vai trò quan trọng trong quá trình phân tích và truyền tải thông tin. Một số mục đích chính của trực quan hóa dữ liệu đó là: hiểu dữ liệu, trình bày kết quả, ra quyết định, giao tiếp và thảo luận, tăng tương tác, giáo dục và đào tạo.

Thứ nhất, trực quan hóa dữ liệu giúp chúng ta hiểu sâu hơn về các mẫu, xu hướng và mối quan hệ tiềm ẩn trong dữ liệu. Khi xem xét các biểu đồ và đồ thị, chúng ta có thể nhận diện ngay lập tức các đặc điểm nổi bật mà có thể không rõ ràng khi chỉ xem xét

các bảng số liệu đơn thuần. Việc khám phá này cho phép chúng ta phát hiện các ngoại lệ hoặc bất thường trong dữ liệu, kiểm tra các giả thuyết và thu thập thông tin chi tiết để làm rõ các hiện tượng phức tạp.

Thứ hai, một trong những mục đích chính của trực quan hóa dữ liệu là truyền tải thông tin một cách rõ ràng và ngắn gọn. Các biểu đồ và đồ thị giúp chúng ta trình bày kết quả phân tích một cách dễ hiểu, đặc biệt là khi đối tượng khán giả không có nền tảng kỹ thuật. Hình ảnh trực quan hấp dẫn không chỉ gây ấn tượng mạnh mẽ mà còn giúp thông tin dễ dàng được nhớ lâu hơn so với việc trình bày bằng các bảng số liệu khô khan.

Thứ ba, trực quan hóa dữ liệu đóng vai trò quan trọng trong việc hỗ trợ ra quyết định. Các biểu đồ và đồ thị cung cấp cái nhìn tổng quan và chi tiết về dữ liệu, giúp người quản lý và nhà phân tích đưa ra quyết định dựa trên dữ liệu thực tế. Khi các thông tin được trình bày một cách rõ ràng và dễ hiểu, nguy cơ sai sót trong quá trình ra quyết định được giảm thiểu, đồng thời tăng cường độ chính xác và hiệu quả của các quyết định chiến lược.

Thứ tư, trực quan hóa dữ liệu giúp giao tiếp thông tin một cách hiệu quả và dễ hiểu, tạo điều kiện cho các cuộc thảo luận và trao đổi thông tin. Các biểu đồ và đồ thị minh họa rõ ràng các điểm chính, giúp các bên liên quan dễ dàng hiểu và đồng thuận với các thông tin được trình bày. Điều này không chỉ giúp tăng cường sự minh bạch mà còn thúc đẩy quá trình hợp tác và thảo luận sâu hơn.

Thứ năm, với các công cụ trực quan hóa dữ liệu tương tác, người dùng có thể khám phá dữ liệu theo nhiều cách khác nhau, chẳng hạn như phóng to, thu nhỏ, và lọc dữ liệu. Khả năng tương tác này cho phép người dùng thực hiện các phân tích động, xem xét dữ liệu từ nhiều góc độ và tìm hiểu các chi tiết cụ thể. Trải nghiệm tương tác này không chỉ làm cho quá trình khám phá dữ liệu trở nên thú vị hơn mà còn tăng cường khả năng phát hiện những thông tin giá trị.

Thứ sáu, trực quan hóa dữ liệu là một công cụ mạnh mẽ trong giáo dục và đào tạo, giúp học sinh và sinh viên hiểu rõ các khái niệm phức tạp và ứng dụng thực tế của dữ liệu. Các biểu đồ và đồ thị giúp minh họa các quy trình và hệ thống phức tạp, làm cho việc giảng dạy và học tập trở nên trực quan và sinh động hơn. Điều này không chỉ giúp nâng cao hiệu quả học tập mà còn khuyến khích sự hứng thú và đam mê trong việc khám phá dữ liệu.

3.2. BIỂU ĐỒ VỀ TRỰC QUAN DỮ LIỆU

3.2.1. Biểu đồ trực quan hóa dữ liệu với dữ liệu phim

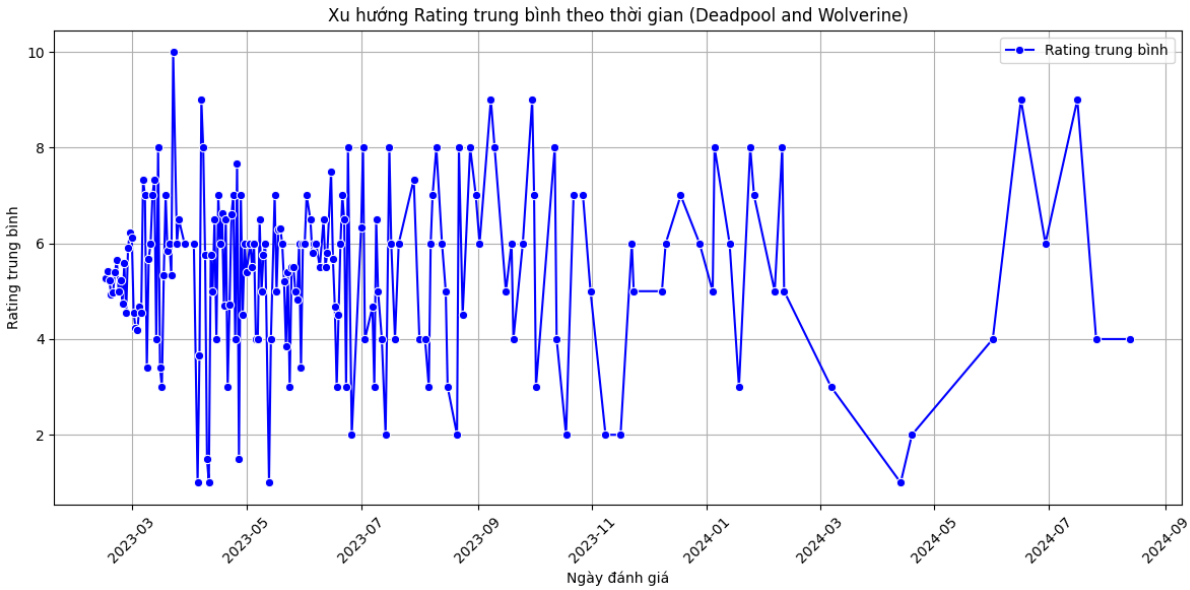
a) Biểu đồ giữa điểm số và tỷ lệ đánh giá

```
import pandas as pd

import numpy as np
```

```
import matplotlib.pyplot as plt
import seaborn as sns
# Giả sử danh sách các DataFrame phim của bạn
movies = [
    ("Deadpool & Wolverine", deadpool_wolverin),
    ("Ant-Man", ant_man),
    ("Guardian of the Galaxy 3", guardian_galaxy_3),
    ("The Boys", the_boys),
    ("Transformers", transformers),
    ("Baby Lon", baby_lon),
    ("Bad Boy", bad_boy),
    ("Friends", friends),
    ("Murder in the Building", murder_in_building),
    ("Intouchable", intouchable),
    ("Breaking Bad", breaking_bad),
    ("Dexter", dexter),
    ("Joker", joker),
    ("Monsters", monsters),
    ("Tulsa King", tulsa_king)
]
# Hàm để tạo và vẽ biểu đồ cho mỗi bộ phim
def plot_movie_trends(df, movie_name):
    # Chuyển đổi cột review_date thành kiểu datetime
    df['Review_Date'] = pd.to_datetime(df['Review_Date'])
    # Tạo cột đánh giá tích cực (1 nếu positive, 0 nếu không)
    df['is_positive'] = df['Sentiment'].apply(lambda x: 1 if x == 'POSITIVE' else 0)
    # Tính điểm rating trung bình và tỷ lệ đánh giá tích cực theo ngày
    daily_trend = df.groupby('Review_Date').agg(
        avg_rating=('Rating', 'mean'),
        total_positive_reviews=('is_positive', 'sum'),
        total_reviews=('Rating', 'count')
    ).reset_index()
    # -------- Biểu đồ 1: Điểm Rating Trung Bình --------
```

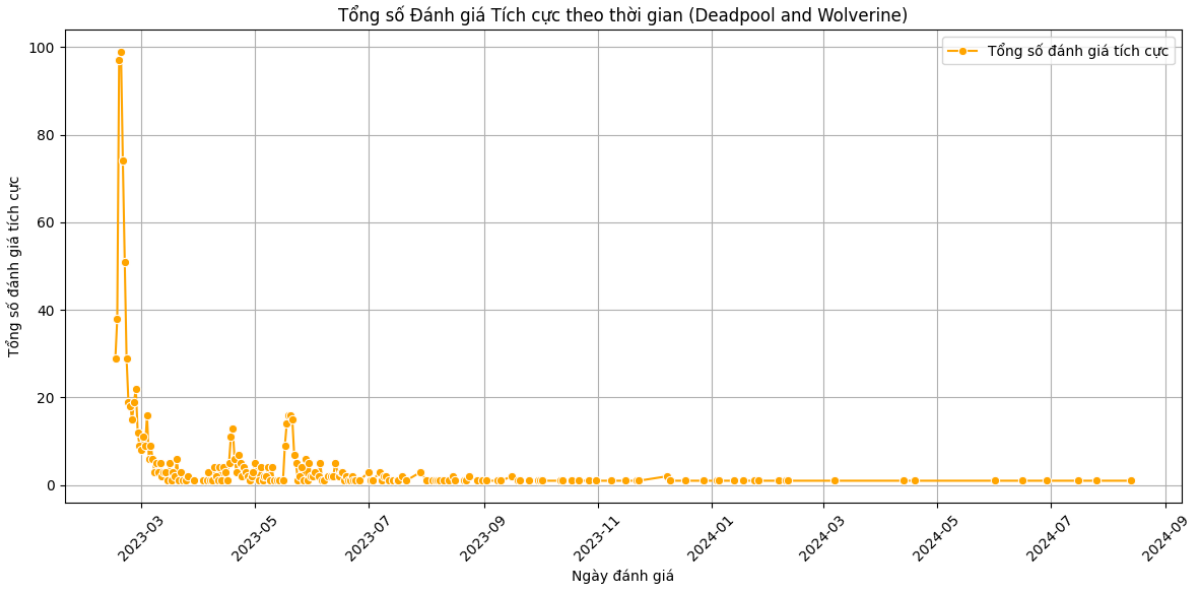
```
    plt.figure(figsize=(12, 6))
     sns.lineplot(x='Review_Date', y='avg_rating', data=daily_trend, label='Rating trung bình', marker='o', color='blue')
    plt.title(f'Xu hướng Rating trung bình theo thời gian ({movie_name})')
    plt.xlabel('Ngày đánh giá')
    plt.ylabel('Rating trung bình')
    plt.xticks(rotation=45)
    plt.grid(True)
    plt.tight_layout()
    plt.show()
    # -------- Biểu đồ 2: Tổng số Đánh giá Tích cực --------
    plt.figure(figsize=(12, 6))
      sns.lineplot(x='Review_Date', y='total_positive_reviews', data=daily_trend, label='Tổng số đánh giá tích cực', marker='o', color='orange')
    plt.title(f'Tổng số Đánh giá Tích cực theo thời gian ({movie_name})')
    plt.xlabel('Ngày đánh giá')
    plt.ylabel('Tổng số đánh giá tích cực')
    plt.xticks(rotation=45)
    plt.grid(True)
    plt.tight_layout()
    plt.show()
# Vẽ biểu đồ cho từng bộ phim
for movie_name, df in movies:
    plot_movie_trends(df, movie_name)
```

Kết quả hiển thị:



Hình 3. 1. Biểu đồ xu hướng rating trung bình theo thời gian (Deadpool and Wolverine)

Biểu đồ trên mô tả xu hướng **rating trung bình theo thời gian** dành cho hai bộ phim *Deadpool* và *Wolverine*, với trục hoành thể hiện ngày đánh giá và trục tung biểu diễn giá trị rating trung bình. Biểu đồ cho thấy sự biến động của đánh giá khán giả, với các giai đoạn rating tăng hoặc giảm đột ngột, phản ánh sự thay đổi cảm nhận của khán giả, có thể liên quan đến các yếu tố như phát hành trailer, phần phim mới, hoặc chiến dịch marketing. Một số thời kỳ rating ổn định, cho thấy sự đồng nhất trong nhận định của khán giả, trong khi những khoảng thời gian dao động mạnh thể hiện sự chia rẽ trong đánh giá. Biểu đồ cung cấp thông tin hữu ích để nhà sản xuất hoặc nhà phát hành phân tích phản hồi từ khán giả, đánh giá hiệu quả các chiến dịch quảng bá và dự đoán xu hướng trong tương lai.



Hình 3. 2. Biểu đồ tổng số đánh giá tích cực theo thời gian (Deadpool and Wolverine)

Biểu đồ trên minh họa tổng số đánh giá tích cực theo thời gian cho hai bộ phim Deadpool và Wolverine.

Ý nghĩa biểu đồ

Xu hướng giảm dần:

- Ban đầu, tổng số đánh giá tích cực đạt đỉnh rất cao, thể hiện sự quan tâm và phản hồi tích cực của khán giả ngay sau khi phim được ra mắt hoặc trong giai đoạn quảng bá mạnh mẽ.
- Sau đó, số lượng đánh giá tích cực giảm dần theo thời gian, phản ánh rằng sự chú ý và thảo luận xung quanh bộ phim đã giảm đi, đây là điều phổ biến với hầu hết các sản phẩm giải trí.

Các điểm nhấn nhỏ:

- Một số điểm trên biểu đồ có sự tăng nhẹ về số lượng đánh giá tích cực, có thể liên quan đến các sự kiện hoặc chiến dịch bổ sung như phát hành nội dung mới, quảng cáo lại phim, hoặc sự xuất hiện của bộ phim trên các nền tảng trực tuyến.

Tính dài hạn:

- Trong giai đoạn sau khi ra mắt một thời gian dài, tổng số đánh giá tích cực trở nên ổn định ở mức thấp, cho thấy lượng khán giả mới hoặc người quan tâm đến bộ phim còn lại không đáng kể.

Ứng dụng:

- Đánh giá hiệu quả chiến dịch quảng bá: Biểu đồ giúp đo lường mức độ hiệu quả của các hoạt động quảng bá trong từng giai đoạn.
- Hiểu xu hướng khán giả: Cung cấp thông tin về cách mà khán giả đón nhận phim theo thời gian, từ đó giúp nhà sản xuất cải thiện chiến lược phát hành.
- Dự đoán chu kỳ sống của sản phẩm: Biểu đồ có thể hỗ trợ dự đoán vòng đời của các bộ phim tương tự trong tương lai.

Tóm lại, biểu đồ này là công cụ quan trọng để phân tích sự tương tác và phản hồi của khán giả qua thời gian.

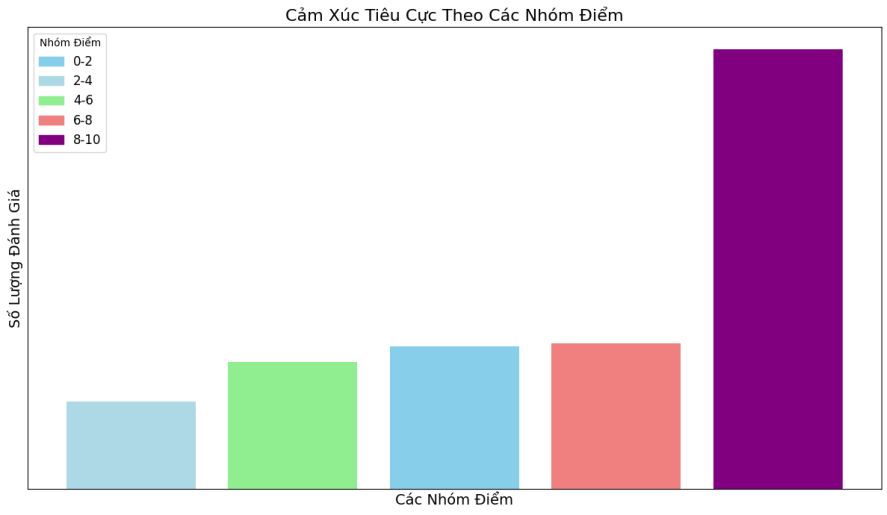
b) Biểu đồ phân bố đánh giá tích cực và tiêu cực theo số sao

```
import matplotlib.pyplot as plt

import pandas as pd

# Tạo nhóm điểm

bins = [0, 2, 4, 6, 8, 10]  # Các khoảng cho thang điểm

labels = ['0-2', '2-4', '4-6', '6-8', '8-10']  # Nhãn cho từng nhóm

sentiment_rating['Score_Group'] = pd.cut(sentiment_rating['Rating'], bins=bins,

labels=labels, include_lowest=True)
```

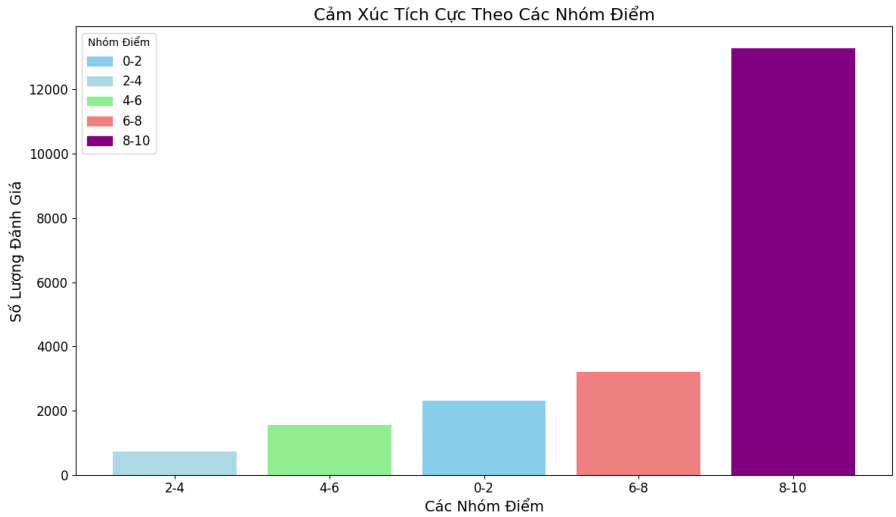
```
# Tính tần suất trong từng nhóm thang điểm và phân loại theo sentiment
score_group_sentiment = sentiment_rating.groupby(['Score_Group',
'Sentiment']).size().reset_index(name='Count')
# Đảm bảo rằng các nhóm thang điểm được sắp xếp theo thứ tự từ bé đến lớn
score_group_sentiment['Score_Group'] =
pd.Categorical(score_group_sentiment['Score_Group'], categories=labels,
ordered=True)
# Sắp xếp theo 'Count' từ bé đến lớn
score_group_sentiment = score_group_sentiment.sort_values(by='Count',
ascending=True)
# Màu sắc cho các nhóm điểm sao (Score Group)
score_colors = {
    '0-2': 'skyblue',
    '2-4': 'lightblue',
    '4-6': 'lightgreen',
    '6-8': 'lightcoral',
    '8-10': 'purple'
}
# Vẽ biểu đồ thanh cho Sentiment = 0 (Tiêu cực)
plt.figure(figsize=(12, 7))
# Dữ liệu Sentiment = 0 (Tiêu cực)
sentiment_0_data = score_group_sentiment[score_group_sentiment['Sentiment'] == 0]
# Tạo biểu đồ thanh cho Sentiment = 0
bars_0 = plt.bar(sentiment_0_data['Score_Group'], sentiment_0_data['Count'],
                 color=[score_colors[x] for x in sentiment_0_data['Score_Group']])
# Thiết lập biểu đồ cho Sentiment = 0
plt.title('Cảm Xúc Tiêu Cực Theo Các Nhóm Điểm', fontsize=16)
plt.xlabel('Các Nhóm Điểm', fontsize=14)
plt.ylabel('Số Lượng Đánh Giá', fontsize=14)
# Xóa các số ở trục x và trục y
plt.xticks([], fontsize=12)
plt.yticks([], fontsize=12)
# Chú thích bảng màu với màu sắc
handles = [plt.Rectangle((0,0),1,1, color=score_colors[label]) for label in labels]
```

```
plt.legend(handles, labels, title='Nhóm Điểm', fontsize=12, loc='upper left')
plt.tight_layout()
plt.show()
# Vẽ biểu đồ thanh cho Sentiment = 1 (Tích cực)
plt.figure(figsize=(12, 7))
# Dữ liệu Sentiment = 1 (Tích cực)
sentiment_1_data = score_group_sentiment[score_group_sentiment['Sentiment'] == 1]
# Tạo biểu đồ thanh cho Sentiment = 1
bars_1 = plt.bar(sentiment_1_data['Score_Group'], sentiment_1_data['Count'],
                 color=[score_colors[x] for x in sentiment_1_data['Score_Group']])
# Thiết lập biểu đồ cho Sentiment = 1
plt.title('Cảm Xúc Tích Cực Theo Các Nhóm Điểm', fontsize=16)
plt.xlabel('Các Nhóm Điểm', fontsize=14)
plt.ylabel('Số Lượng Đánh Giá', fontsize=14)
# Xóa các số ở trục x và trục y
plt.xticks(fontsize=12)
plt.yticks(fontsize=12)
# Chú thích bảng màu với màu sắc
handles = [plt.Rectangle((0,0),1,1, color=score_colors[label]) for label in labels]
plt.legend(handles, labels, title='Nhóm Điểm', fontsize=12, loc='upper left')
plt.tight_layout()
plt.show()
```

Kết quả hiển thị:



Hình 3. 3. Biểu đồ thể hiện cảm xúc tiêu cực theo các nhóm điểm



Hình 3. 4. Biểu đồ thể hiện cảm xúc tích cực theo các nhóm điểm

Nhận xét về biểu đồ phân bố đánh giá tích cực và tiêu cực theo số sao:

Phân bố theo nhóm điểm số:

- Nhóm 0-2 sao:
    - Mặc dù nhóm này có số lượng đánh giá không quá lớn, nhưng tỷ lệ các đánh giá tích cực (màu tím) và tiêu cực (màu xanh) có sự phân bổ khá đa dạng giữa các bộ phim. Điều này có thể cho thấy những bộ phim có sự chia rẽ rõ rệt về sự đón nhận từ khán giả, đặc biệt là đối với các bộ phim có nội dung hoặc phong cách gây tranh cãi.
- Nhóm 2-4 sao:
    - Nhóm này có sự phân bố khá đều giữa đánh giá tích cực và tiêu cực, đặc biệt với những bộ phim có những yếu tố chưa làm hài lòng một bộ phận khán giả, nhưng vẫn có các yếu tố thu hút một phần người xem. Phân tích nhóm này có thể giúp hiểu sâu hơn về các yếu tố cần cải thiện.
- Nhóm 4-6 sao:
    - Ở nhóm này, các đánh giá tích cực vẫn vượt trội hơn các đánh giá tiêu cực, phản ánh sự hài lòng của đa số người xem, nhưng cũng cho thấy rằng vẫn còn những điểm chưa đủ mạnh mẽ để tạo ra sự đồng tình tuyệt đối từ toàn bộ khán giả.
- Nhóm 6-8 sao:
    - Nhóm này có sự phân bổ rõ rệt với nhiều đánh giá tích cực, cho thấy rằng các bộ phim trong nhóm này đã thành công trong việc đáp ứng yêu cầu của người xem. Tuy nhiên, vẫn có một số quan điểm không hài lòng về một số yếu tố trong phim.
- Nhóm 8-10 sao:

- Nhóm điểm này chiếm tỷ lệ lớn nhất, đặc biệt ở các bộ phim nổi bật. Đánh giá tích cực chiếm ưu thế rõ rệt, cho thấy sự thành công trong việc tạo ra một trải nghiệm mạnh mẽ và sâu sắc cho người xem. Những bộ phim ở nhóm này có thể đạt được sự công nhận cao từ cả giới phê bình và khán giả.

Tổng quan xu hướng:

- Mặc dù các bộ phim trong bộ dữ liệu có sự phân bố khá rộng về điểm số, nhưng phần lớn các đánh giá tích cực tập trung ở các nhóm điểm số cao (6-10 sao). Điều này chứng tỏ rằng đa số người xem đã có trải nghiệm tốt hoặc rất tốt với các bộ phim, dù vẫn có một lượng đánh giá tiêu cực trong những bộ phim có yếu tố gây tranh cãi.

Khuyến nghị:

- Phân tích sâu hơn về các nhóm điểm thấp (0-4 sao) có thể giúp các nhà sản xuất phim hiểu rõ hơn về những yếu tố mà người xem cảm thấy chưa hài lòng. Điều này đặc biệt quan trọng đối với những bộ phim có tầm ảnh hưởng lớn nhưng không đáp ứng được mong đợi của một bộ phận khán giả.
- Các bộ phim ở nhóm điểm cao có thể được phân tích để hiểu rõ lý do thành công, từ đó có thể rút ra những bài học và cải thiện các yếu tố quan trọng cho các sản phẩm phim sau này.

Ý nghĩa:

Biểu đồ này phản ánh sự đa dạng trong trải nghiệm và sự đánh giá của khán giả đối với nhiều bộ phim. Điều này có thể giúp các nhà sản xuất phim cải thiện chất lượng sản phẩm, đồng thời hiểu rõ hơn về những yếu tố mà khán giả yêu thích hoặc không hài lòng.

Để thực hiện việc tính toán các chỉ số liên quan đến đánh giá của ba thể loại phim: hành động - khoa học viễn tưởng (*action-sci*), hài (*comedy*), và tội phạm (*crime*), ta sử dụng đoạn code sau đây:

```
# Lượng rating tiêu cực

negative_rating_act =  ratings_act_sci[ratings_act_sci < 5].size

# Lượng rating tích cực

positive_rating_act = ratings_act_sci[ratings_act_sci >5].size

# Trung bình đánh giá của phim hành động

mean_rating_act =  ratings_act_sci.mean()

#Lượng rating tích cực

positive_rating_comedy = ratings_comedy[ratings_comedy > 5].size

#Lượng rating tiêu cực
```

```
negative_rating_comedy = ratings_comedy[ratings_comedy < 5].size

#trung bình sao đánh giá

mean_rating_comedy = ratings_comedy.mean()

#Lượng rating tích cực

positive_rating_crime = ratings_crime[ratings_crime > 5].size

#Lượng rating tiêu cực

negative_rating_crime = ratings_crime[ratings_crime < 5].size

#trung bình sao đánh giá

mean_rating_crime = ratings_crime.mean()
```

Ý nghĩa:

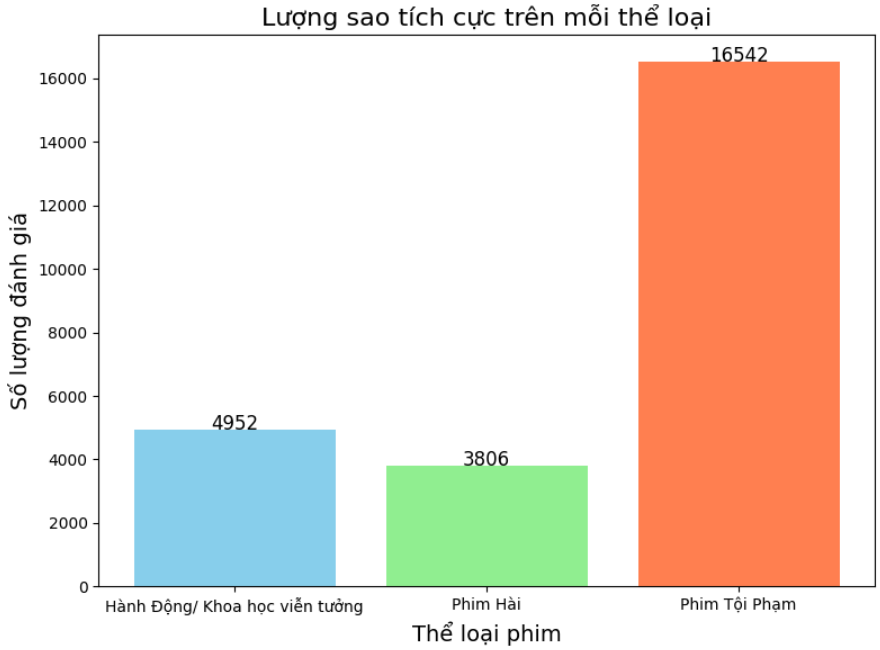
- Đoạn mã này phân tích dữ liệu đánh giá phim bằng cách chia thành hai nhóm: đánh giá tích cực và tiêu cực, đồng thời tính điểm trung bình cho từng thể loại.
- Các kết quả thu được giúp người phân tích hiểu rõ hơn về xu hướng đánh giá khán giả đối với từng thể loại phim.
- Các chỉ số này có thể hỗ trợ việc so sánh mức độ yêu thích giữa các thể loại hoặc đánh giá chất lượng chung của từng thể loại phim dựa trên phản hồi từ khán giả.

c) Biểu đồ lượng sao tích cực trên mỗi thể loại phim

```
import matplotlib.pyplot as plt

# Dữ liệu

categories = ['Hành Động/ Khoa học viễn tưởng', 'Phim Hài', 'Phim Tội Phạm']

positive_ratings = [

    ratings_act_sci[ratings_act_sci > 5].size,

    ratings_comedy[ratings_comedy > 5].size,

    ratings_crime[ratings_crime > 5].size

]

# Vẽ biểu đồ cột

plt.figure(figsize=(8, 6))

colors = ['skyblue', 'lightgreen', 'coral']  # Màu sắc cho từng thể loại

plt.bar(categories, positive_ratings, color=colors)

# Thiết lập tiêu đề và nhãn trục

plt.title('Lượng sao tích cực trên mỗi thể loại', fontsize=16)

plt.xlabel('Thể loại phim', fontsize=14)

plt.ylabel('Số lượng đánh giá', fontsize=14)

# Hiển thị giá trị trên cột
```

```
for i, value in enumerate(positive_ratings):
    plt.text(i, value + 1, str(value), ha='center', fontsize=12)
# Hiển thị biểu đồ
plt.tight_layout()
plt.show()
```

Kết quả hiển thị:



Hình 3. 5. Biểu đồ thể hiện lượng sao tích cực trên mỗi thể loại phim

Nhận xét biểu đồ (Lượng sao tích cực trên mỗi thể loại phim)

Phim Tội Phạm nổi bật với lượng sao tích cực cao:

- Phim Tội Phạm đạt 16,542 lượt sao tích cực, chiếm ưu thế vượt trội so với các thể loại khác.
- Điều này cho thấy phim thuộc thể loại này được khán giả yêu thích và đánh giá cao, có thể nhờ nội dung hấp dẫn, cốt truyện lôi cuốn, hoặc sự đầu tư vào chất lượng sản xuất.

Thể loại Hành Động/Khoa học viễn tưởng và Hài có mức đánh giá tích cực thấp hơn:

- Phim Hành Động/Khoa học viễn tưởng có 4,952 lượt, thể hiện mức độ yêu thích ở mức trung bình.
- Phim Hài có 3,806 lượt, thấp nhất trong ba thể loại. Có thể lý do là nội dung hài kịch không phù hợp với toàn bộ khán giả hoặc mức độ phổ biến của thể loại này thấp hơn.

Khoảng cách giữa các thể loại:

- Lượng sao tích cực của phim Tội Phạm cao gấp hơn 3 lần so với Hành Động/Khoa học viễn tưởng và khoảng 4.3 lần so với phim Hài.
- Điều này thể hiện sự khác biệt lớn về mức độ yêu thích của khán giả giữa các thể loại.

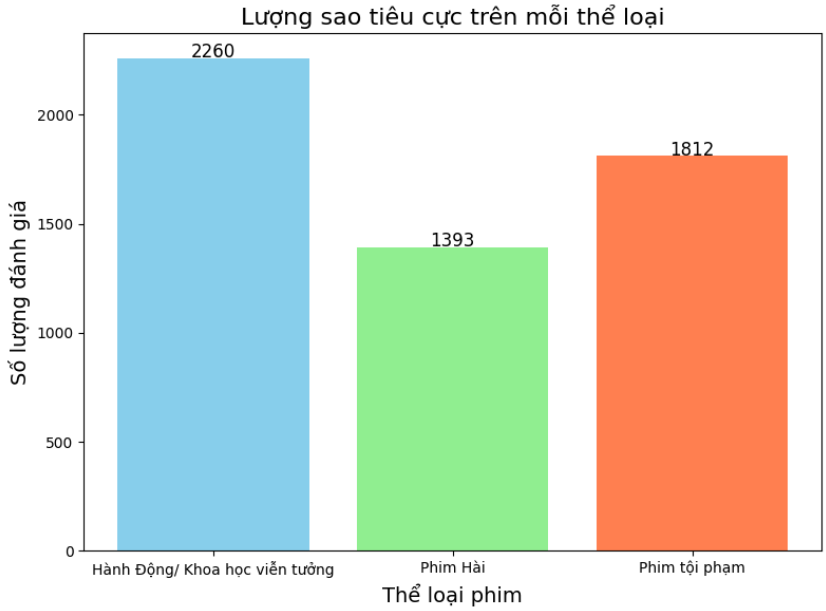
Kết luận:

- Phim Tội Phạm nổi bật là thể loại được yêu thích nhất, cần tiếp tục duy trì chất lượng và khai thác thêm nội dung sáng tạo để giữ vững vị thế.
- Phim Hành Động/Khoa học viễn tưởng và phim Hài, mặc dù có mức đánh giá tích cực thấp hơn, vẫn có tiềm năng cải thiện bằng cách tập trung vào yếu tố nội dung phù hợp hơn với thị hiếu khán giả.

d) Biểu đồ lượng sao tiêu cực trên mỗi thể loại phim

```
import matplotlib.pyplot as plt

# Dữ liệu

categories = ['Hành Động/ Khoa học viễn tưởng', 'Phim Hài', 'Phim tội phạm']

negative_rating = [

    ratings_act_sci[ratings_act_sci <5].size,

    ratings_comedy[ratings_comedy < 5].size,

    ratings_crime[ratings_crime < 5].size

]

# Vẽ biểu đồ cột

plt.figure(figsize=(8, 6))

colors = ['skyblue', 'lightgreen', 'coral']  # Màu sắc cho từng thể loại

plt.bar(categories, negative_rating, color=colors)

# Thiết lập tiêu đề và nhãn trục

plt.title('Lượng sao tiêu cực trên mỗi thể loại', fontsize=16)

plt.xlabel('Thể loại phim', fontsize=14)

plt.ylabel('Số lượng đánh giá', fontsize=14)

# Hiển thị giá trị trên cột

for i, value in enumerate(negative_rating):

    plt.text(i, value + 1, str(value), ha='center', fontsize=12)

# Hiển thị biểu đồ

plt.tight_layout()

plt.show()
```

Kết quả hiển thị:



Hình 3. 6. Biểu đồ thể hiện lượng sao tiêu cực trên mỗi thể loại phim

Nhận xét biểu đồ (lượng sao tiêu cực theo thể loại phim)

Giới thiệu:

- Biểu đồ trên cho thấy sự khác biệt rõ rệt về lượng sao tiêu cực giữa các thể loại phim. Dưới đây là phân tích chi tiết về nguyên nhân của hiện tượng này.

Thể loại Hành động/Khoa học viễn tưởng:

- Kỳ vọng cao: Do thường xuyên xuất hiện những bộ phim bom tấn với hiệu ứng hình ảnh hoành tráng, khán giả đặt kỳ vọng rất cao vào thể loại này.
- Khán giả đa dạng: Việc thu hút cả trẻ em và người lớn khiến việc làm hài lòng tất cả mọi người trở nên khó khăn hơn.
- Cạnh tranh khốc liệt: Số lượng phim sản xuất lớn dẫn đến sự cạnh tranh gay gắt, đòi hỏi chất lượng phim phải thật sự nổi bật.

Thể loại Phim hài:

- Mục đích giải trí: Khán giả xem phim hài để thư giãn, do đó yêu cầu về chất lượng thường không quá khắt khe.
- Khó đánh giá khách quan: Khả năng gây cười là yếu tố chủ quan, phụ thuộc vào sở thích cá nhân.

Kết luận:

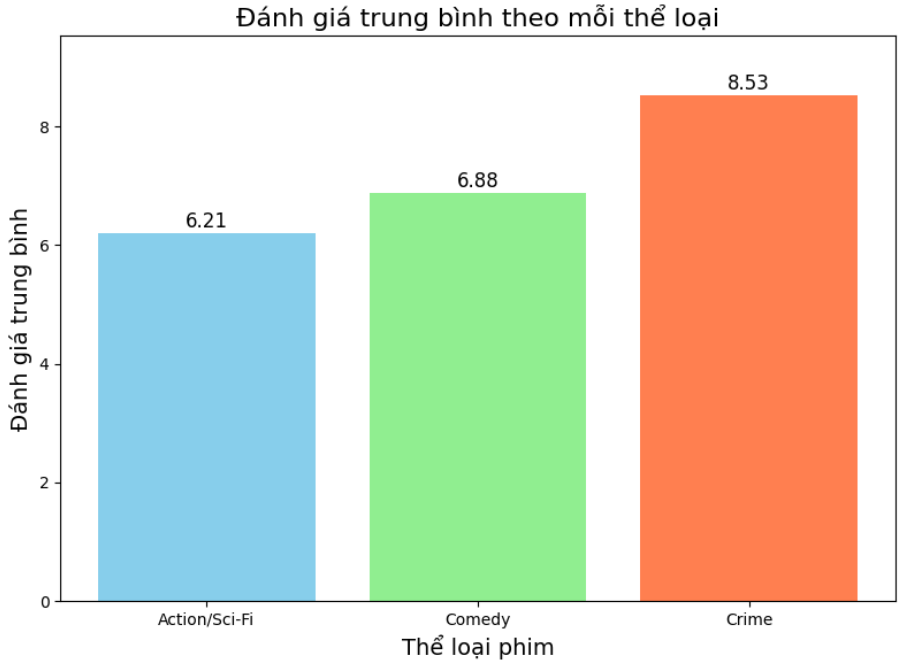
Lượng sao tiêu cực không chỉ phụ thuộc vào thể loại phim mà còn chịu ảnh hưởng bởi nhiều yếu tố khác như kịch bản, diễn xuất, và đạo diễn. Để tạo ra một bộ phim thành công, các nhà làm phim cần cân nhắc kỹ lưỡng các yếu tố này.

e) Biểu đồ trung bình sao đánh giá trên mỗi thể loại phim

```
import matplotlib.pyplot as plt
```

```
# Dữ liệu
categories = ['Action/Sci-Fi', 'Comedy', 'Crime']
mean_ratings = [
    ratings_act_sci.mean(),
    ratings_comedy.mean(),
    ratings_crime.mean()
]
# Vẽ biểu đồ cột
plt.figure(figsize=(8, 6))
colors = ['skyblue', 'lightgreen', 'coral']  # Màu sắc cho từng thể loại
plt.bar(categories, mean_ratings, color=colors)
# Thiết lập tiêu đề và nhãn trục
plt.title('Đánh giá trung bình theo mỗi thể loại', fontsize=16)
plt.xlabel('Thể loại phim', fontsize=14)
plt.ylabel('Đánh giá trung bình', fontsize=14)
# Hiển thị giá trị trung bình trên cột
for i, value in enumerate(mean_ratings):
    plt.text(i, value + 0.1, f"{value:.2f}", ha='center', fontsize=12)
# Giới hạn trục y để hiển thị rõ hơn
plt.ylim(0, max(mean_ratings) + 1)
# Hiển thị biểu đồ
plt.tight_layout()
plt.show()
```

Kết quả hiển thị:



Hình 3. 7. Biểu đồ thể hiện đánh giá trung bình theo mỗi thể loại phim

Nhận xét biểu đồ (Điểm đánh giá trung bình theo thể loại phim)

Thể loại Phim Tội Phạm dẫn đầu với điểm đánh giá cao nhất:

- Phim Tội Phạm (Crime) có điểm đánh giá trung bình là 8.53, vượt trội so với hai thể loại còn lại.
- Điều này cho thấy phim thuộc thể loại này có chất lượng tốt, với kịch bản, diễn xuất, và yếu tố nội dung đáp ứng kỳ vọng của khán giả.

Thể loại Phim Hài đạt điểm trung bình khá cao:

- Phim Hài (Comedy) có điểm đánh giá trung bình là 6.88, cao hơn so với Hành Động/Khoa học viễn tưởng nhưng vẫn kém khá xa so với Phim Tội Phạm.
- Đây là dấu hiệu rằng phim hài được khán giả yêu thích ở mức độ vừa phải, nhưng vẫn có cơ hội cải thiện nếu đầu tư vào các yếu tố sáng tạo hơn.

Thể loại Hành Động/Khoa học viễn tưởng có điểm thấp nhất:

- Phim Hành Động/Khoa học viễn tưởng (Act/Sci-Fi) đạt điểm trung bình chỉ 6.21, thấp nhất trong ba thể loại.
- Điều này có thể do các phim thuộc thể loại này không đáp ứng được kỳ vọng cao của khán giả, như cốt truyện không đặc sắc hoặc hiệu ứng không thuyết phục.

Khoảng cách điểm số giữa các thể loại:

- Điểm trung bình của Phim Tội Phạm **(**8.53**)** cao hơn đáng kể, gần 1.65 điểm so với Phim Hài và hơn 2.3 điểm so với Phim Hành Động/Khoa học viễn tưởng.
- Điều này thể hiện sự khác biệt lớn trong chất lượng hoặc cách khán giả đánh giá các thể loại này.

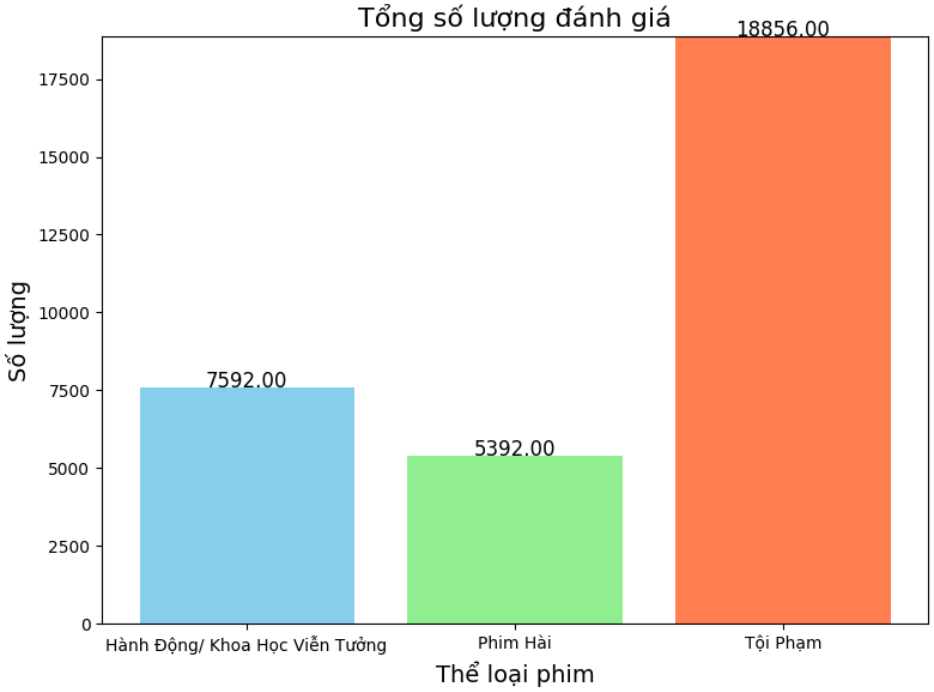
Kết luận:

- Phim Tội Phạm (Crime) là thể loại được đánh giá cao nhất, cần tiếp tục duy trì chất lượng và tạo thêm những tác phẩm xuất sắc để thu hút nhiều khán giả hơn.
- Phim Hài (Comedy) đã đạt mức đánh giá khá, nhưng vẫn có tiềm năng cải thiện.
- Phim Hành Động/Khoa học viễn tưởng (Act/Sci-Fi) cần tập trung cải thiện cả về nội dung, kỹ xảo, và yếu tố kịch tính để tăng cường sức hút đối với người xem.

Để có cái nhìn trực quan hơn về tổng số lượng đánh giá phim, chúng ta sẽ tiến hành vẽ biểu đồ cho dữ liệu này.

```
import matplotlib.pyplot as plt

# Dữ liệu

categories = ['Hành Động/ Khoa Học Viễn Tưởng', 'Phim Hài', 'Tội Phạm']

mean_ratings = [

    ratings_act_sci.size,

    ratings_comedy.size,

    ratings_crime.size

]

# Vẽ biểu đồ cột

plt.figure(figsize=(8, 6))

colors = ['skyblue', 'lightgreen', 'coral']  # Màu sắc cho từng thể loại

plt.bar(categories, mean_ratings, color=colors)

# Thiết lập tiêu đề và nhãn trục

plt.title('Tổng số lượng đánh giá', fontsize=16)

plt.xlabel('Thể loại phim', fontsize=14)

plt.ylabel('Số lượng', fontsize=14)

# Hiển thị giá trị trung bình trên cột

for i, value in enumerate(mean_ratings):

    plt.text(i, value + 0.1, f"{value:.2f}", ha='center', fontsize=12)

# Giới hạn trục y để hiển thị rõ hơn

plt.ylim(0, max(mean_ratings) + 1)

# Hiển thị biểu đồ

plt.tight_layout()

plt.show()
```

Kết quả hiển thị:



Hình 3. 8. Biểu đồ thể hiện tổng số lượng đánh giá thể loại phim

Nhận xét chi tiết biểu đồ

Sự phổ biến của thể loại Tội phạm và các phân tích chi tiết:

- Sự phổ biến của thể loại Tội phạm: Rõ ràng, thể loại Tội phạm đang chiếm ưu thế về số lượng đánh giá, vượt trội so với hai thể loại còn lại. Điều này cho thấy khán giả có xu hướng quan tâm đến các nội dung liên quan đến tội phạm hơn.
- Sự khác biệt đáng kể giữa các thể loại: Số lượng đánh giá giữa thể loại Tội phạm và hai thể loại còn lại có sự chênh lệch khá lớn. Điều này gợi ý rằng có thể có những yếu tố khác biệt về nội dung, diễn viên, hoặc các yếu tố sản xuất khác đang tác động đến sự lựa chọn của khán giả.
- Tiềm năng khai thác: Sự khác biệt này cũng mở ra nhiều cơ hội để phân tích sâu hơn, chẳng hạn như:
    - So sánh các đặc điểm của các bộ phim thuộc từng thể loại.
    - Tìm hiểu sở thích của khán giả đối với từng thể loại.
    - Dự đoán xu hướng phát triển của thị trường phim.
- Sự mất cân bằng trong việc thu thập dữ liệu::

Mất cân bằng trong mô hình phân tích:

- Dự đoán bị thiên lệch: Nếu bạn sử dụng dữ liệu này để xây dựng mô hình học máy hoặc thực hiện phân tích, mô hình có thể thiên lệch về mặt dự đoán. Mô hình có thể học được rằng phim tội phạm là loại phổ biến và tập trung vào việc phân tích phim tội phạm hơn là các thể loại khác. Điều này dẫn đến việc mô hình có

thể bỏ qua các thể loại ít phổ biến như phim hài hoặc phim hành động/khoa học viễn tưởng.

- Đánh giá không công bằng**:** Nếu bạn chỉ tập trung vào việc phân tích số lượng sao và điểm đánh giá của phim tội phạm, bạn có thể bỏ sót các mối quan hệ quan trọng trong dữ liệu của các thể loại phim khác. Điều này sẽ làm giảm khả năng tổng quát của mô hình khi áp dụng cho các thể loại ít phổ biến.

Ảnh hưởng đến các chỉ số đánh giá:

- Accuracy (Độ chính xác)**:** Nếu mô hình của bạn chỉ dự đoán thể loại phim tội phạm do thể loại này chiếm ưu thế trong dữ liệu, bạn có thể có một độ chính xác cao, nhưng thực tế mô hình lại không nhận diện đúng các thể loại phim hài hoặc phim hành động/khoa học viễn tưởng.
- Chỉ số phân loại không đúng**:** Các chỉ số như precision, recall, hoặc F1-score sẽ không phản ánh chính xác hiệu quả của mô hình đối với các thể loại phim ít xuất hiện hơn. Ví dụ, nếu bạn có ít dữ liệu về phim hài nhưng mô hình chỉ dự đoán chính xác đối với phim tội phạm, những chỉ số này sẽ không đủ mạnh để đánh giá chất lượng dự đoán của các thể loại ít phổ biến.

Thiếu đại diện cho các thể loại ít xuất hiện:

- Mất đại diện: Khi số lượng dữ liệu của một thể loại (phim tội phạm) vượt trội so với các thể loại khác, bạn sẽ gặp phải vấn đề thiếu đại diện cho các thể loại ít phổ biến hơn. Điều này có thể khiến cho các phân tích không đầy đủ và không chính xác đối với các thể loại ít được đại diện trong bộ dữ liệu.
- Khả năng dự đoán sai: Trong trường hợp bạn sử dụng mô hình để dự đoán thể loại phim dựa trên các đặc điểm như đánh giá hoặc sao, mô hình có thể dự đoán sai khi gặp phải phim hài hoặc phim hành động/khoa học viễn tưởng, vì nó đã được huấn luyện quá nhiều trên dữ liệu phim tội phạm.

Ảnh hưởng đến việc đưa ra quyết định:

- Nếu bạn sử dụng các phân tích này để đưa ra các quyết định, như là đưa ra khuyến nghị phim cho người dùng hoặc phân tích xu hướng trong các thể loại phim, bạn có thể sẽ bỏ qua các xu hướng hoặc sở thích của người dùng đối với những thể loại ít phổ biến. Điều này sẽ gây ảnh hưởng xấu đến chất lượng các quyết định, như là khuyến nghị phim cho khán giả.

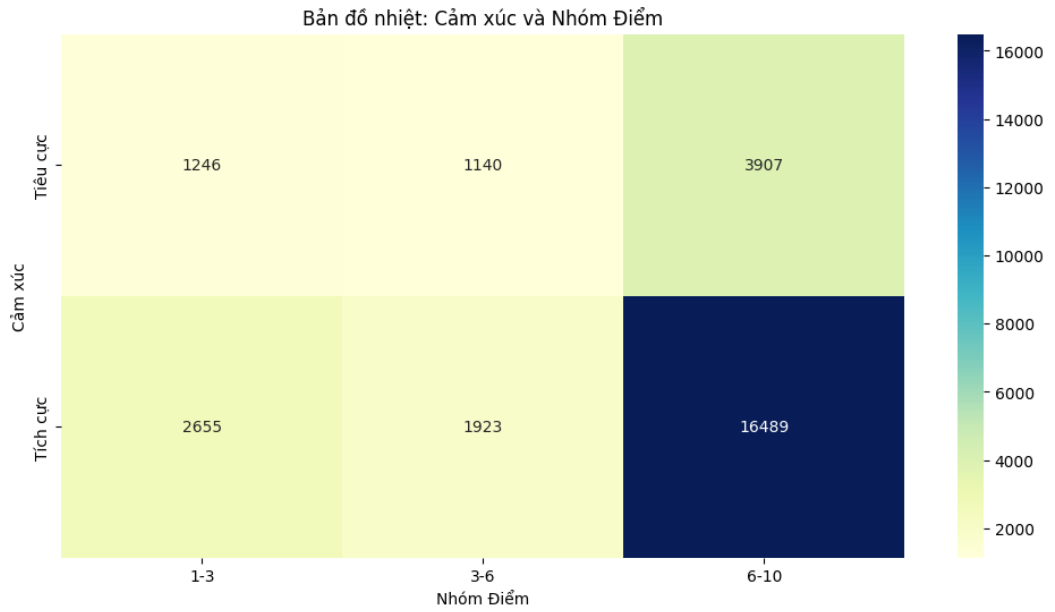
Vấn đề trong việc diễn giải và truyền đạt kết quả

- Khi bạn có quá nhiều dữ liệu từ một thể loại, bạn sẽ phải cẩn thận khi giải thích kết quả của mình. Nếu bạn không thừa nhận rằng có sự mất cân bằng dữ liệu, bạn có thể mắc phải những hiểu lầm hoặc cung cấp thông tin sai lệch cho những người xem hoặc sử dụng phân tích

f) Bản đồ nhiệt giữa cảm xúc và nhóm điểm

```
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd
# Tạo dữ liệu giả lập
data = synthetic.copy()
# Phân loại Rating theo các nhóm
bins = [1, 3, 6, 10]
labels = ['1-3', '3-6', '6-10']
data['Nhóm Điểm'] = pd.cut(data['Rating'], bins=bins, labels=labels,
include_lowest=True)
# Tạo một pivot table để chuẩn bị cho heatmap
heatmap_data = data.groupby(['Sentiment', 'Nhóm Điểm']).size().unstack(fill_value=0)
# Đổi nhãn sentiment sang tiếng Việt
sentiment_labels = {
    'POSITIVE': 'Tích cực',
    'NEGATIVE': 'Tiêu cực',
    'NEUTRAL': 'Trung lập'
}
heatmap_data.index = heatmap_data.index.map(sentiment_labels)
# Vẽ biểu đồ heatmap
plt.figure(figsize=(12, 6))
sns.heatmap(heatmap_data, annot=True, fmt="d", cmap="YlGnBu", cbar=True)
plt.title("Bản đồ nhiệt: Cảm xúc và Nhóm Điểm")
plt.xlabel("Nhóm Điểm")
plt.ylabel("Cảm xúc")
plt.show()
```

Kết quả hiển thị:



Hình 3. 9. Bản đồ nhiệt giữa cảm xúc và nhóm điểm của đánh giá phim

Nhận xét chi tiết chi tiết biểu đồ

Mối liên hệ giữa cảm xúc và nhóm điểm:

- Ở nhóm điểm từ 6-10, số lượng đánh giá có cảm xúc tích cực chiếm áp đảo với 16.489 đánh giá, so với cảm xúc tiêu cực chỉ có 3.907 đánh giá. Điều này cho thấy nhóm điểm cao thường đi kèm với đánh giá tích cực.
- Ở nhóm điểm từ 3-6, số lượng đánh giá cảm xúc tích cực (1.923) và tiêu cực (1.140) gần như tương đương, cho thấy cảm xúc ở nhóm này có xu hướng cân bằng hơn.
- Ở nhóm điểm từ 1-3, số lượng đánh giá tiêu cực (1.246) vẫn thấp hơn một chút so với tích cực (2.655). Tuy nhiên, chênh lệch không quá lớn, phản ánh rằng điểm thấp không hoàn toàn đi kèm cảm xúc tiêu cực.

Tổng quan về cảm xúc tích cực và tiêu cực:

- Đánh giá **tích cực** chiếm ưu thế lớn, đặc biệt rõ ràng trong nhóm điểm cao (**6-10**). Điều này có thể là dấu hiệu cho thấy phần lớn người dùng có trải nghiệm tốt khi để lại đánh giá.
- Đánh giá **tiêu cực** có xu hướng tập trung ở các nhóm điểm trung bình (3-6) hoặc thấp (1-3), tuy nhiên, tổng số lượng vẫn ít hơn đáng kể so với cảm xúc tích cực.

Kết luận về xu hướng cảm xúc và điểm số:

- Có mối liên hệ chặt chẽ giữa điểm số cao và cảm xúc tích cực: sản phẩm hoặc nội dung nhận được điểm số tốt thường đi kèm với phản hồi tích cực.
- Ngược lại, nhóm điểm thấp thường có sự hiện diện của cảm xúc tiêu cực, nhưng vẫn không phải là tuyệt đối.

Gợi ý cải thiện:

- Đối với các sản phẩm hoặc nội dung có điểm trong nhóm từ 1-3 và 3-6, cần phân tích kỹ hơn để cải thiện, vì ở đây xuất hiện nhiều cảm xúc tiêu cực.
- Tiếp tục duy trì và tối ưu hóa các yếu tố thu hút cảm xúc tích cực, đặc biệt trong nhóm điểm từ 6-10, để giữ vững chất lượng và sự hài lòng của người dùng.

Xu hướng chung:

- Người xem thường đánh giá cao (4-5 sao) những bộ phim mà họ yêu thích. Ngược lại, phim có điểm thấp thường nhận được nhiều đánh giá tiêu cực.

Điểm bất thường:

- Có một tỷ lệ nhỏ người xem đánh giá 1-2 sao nhưng vẫn cảm thấy "thích" hoặc "rất thích". Điều này có thể do sở thích cá nhân hoặc các yếu tố khác.

Biểu đồ nhiệt này là công cụ trực quan hóa rất tốt để hiểu mối quan hệ giữa điểm số và cảm xúc, đồng thời hỗ trợ đưa ra các chiến lược cải thiện sản phẩm/dịch vụ.

3.2.2. Biểu đồ trực quan hóa dữ liệu với dữ liệu sản phẩm

Để hợp nhất và chuẩn bị dữ liệu để thực hiện các phân tích nâng cao về đánh giá và cảm xúc của người dùng ta sử dụng đoạn code sau:

```
sentiment_rating_product = pd.concat([cong_nghe_df,my_pham_df,thuc_pham_df])

sentiment_rating_product
```

Ý nghĩa của đoạn code

Hợp nhất dữ liệu:

- pd.concat được sử dụng để gộp dữ liệu từ ba DataFrame: cong_nghe_df, my_pham_df, và thuc_pham_df.
- Các DataFrame này có thể chứa dữ liệu liên quan đến đánh giá sản phẩm từ các lĩnh vực khác nhau, chẳng hạn như công nghệ, mỹ phẩm và thực phẩm.
- Kết quả của phép gộp này là một DataFrame tổng hợp mới có tên là sentiment_rating_product.

Cấu trúc của DataFrame hợp nhất:

- Review: Nội dung đánh giá của người dùng về sản phẩm.
- Rating: Điểm số mà người dùng đánh giá sản phẩm (thường từ 1-5).
- Date: Ngày đánh giá được tạo ra.
- Sentiment: Cảm xúc của người dùng đối với sản phẩm, phân loại thành POS (Positive - Tích cực) và NEG (Negative - Tiêu cực).

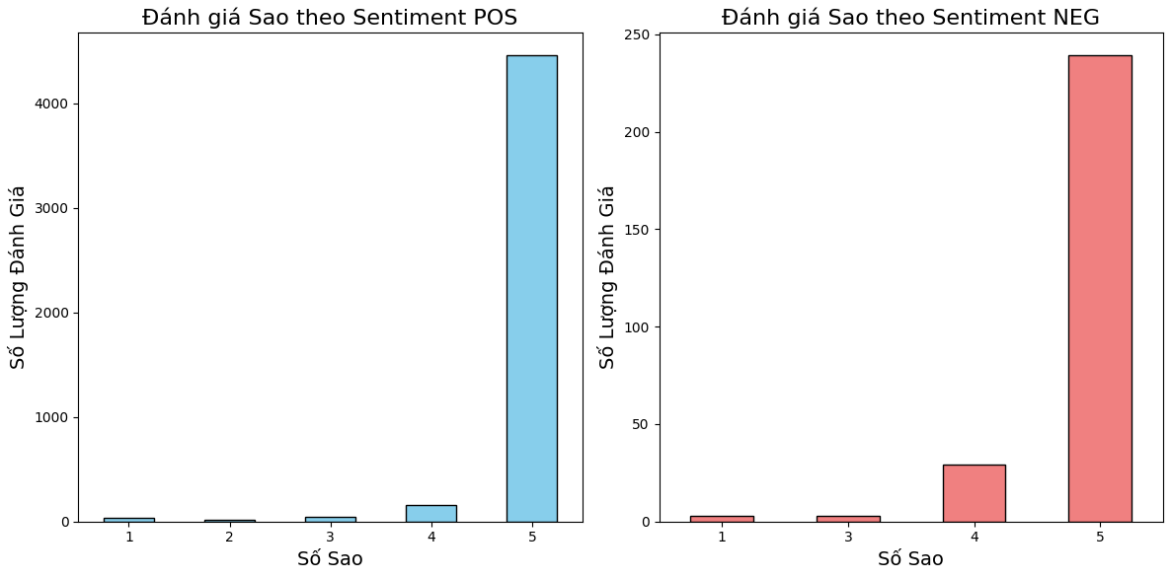
Ý nghĩa dữ liệu:

- Dữ liệu này có thể giúp phân tích trải nghiệm người dùng, sự hài lòng và các vấn đề tiềm ẩn liên quan đến sản phẩm từ các lĩnh vực khác nhau.

a) Biểu đồ giữa điểm số và tỉ lệ đánh giá

```
import matplotlib.pyplot as plt

import pandas as pd

# Nhóm dữ liệu theo Sentiment và Rating

rating_pos = sentiment_rating_product[sentiment_rating_product['Sentiment'] ==
'POS'].groupby('Rating').size()

rating_neg = sentiment_rating_product[sentiment_rating_product['Sentiment'] ==
'NEG'].groupby('Rating').size()

# Vẽ biểu đồ cho đánh giá sao theo POS

plt.figure(figsize=(12, 6))

plt.subplot(1, 2, 1)

rating_pos.plot(kind='bar', color='skyblue', edgecolor='black')

plt.title('Đánh giá Sao theo Sentiment POS', fontsize=16)

plt.xlabel('Số Sao', fontsize=14)

plt.ylabel('Số Lượng Đánh Giá', fontsize=14)

plt.xticks(rotation=0)

# Vẽ biểu đồ cho đánh giá sao theo NEG

plt.subplot(1, 2, 2)

rating_neg.plot(kind='bar', color='lightcoral', edgecolor='black')

plt.title('Đánh giá Sao theo Sentiment NEG', fontsize=16)

plt.xlabel('Số Sao', fontsize=14)

plt.ylabel('Số Lượng Đánh Giá', fontsize=14)

plt.xticks(rotation=0)

# Tinh chỉnh hiển thị

plt.tight_layout()

plt.show()
```

Kết quả hiển thị:



Hình 3. 10. Biểu đồ đánh giá sao theo Sentiment POS và NEG

Nhận xét chi tiết về biểu đồ:

Phân phối đánh giá sao cho Sentiment POS (Tích Cực):

Chênh lệch rõ rệt giữa 5 sao và các sao khác:

- Biểu đồ cho thấy số lượng đánh giá 5 sao chiếm tuyệt đối trong nhóm cảm xúc Tích Cực (POS), với hơn 4,000 đánh giá. Điều này phản ánh rằng đa số người dùng cảm thấy rất hài lòng với sản phẩm, và đánh giá cao sản phẩm ở mức tối đa.
- Những đánh giá từ 1 đến 4 sao đều có số lượng cực kỳ ít, cho thấy rằng khi người dùng cảm thấy hài lòng, họ không đánh giá sản phẩm ở mức thấp. Điều này cũng có thể chỉ ra rằng sản phẩm được đa phần khách hàng yêu thích, với ít vấn đề phát sinh trong quá trình sử dụng.

Sự phân bố đồng nhất:

- Hầu như không có sự phân bố đồng đều giữa các mức sao trong nhóm tích cực. Đây là một dấu hiệu rõ ràng cho thấy khi người tiêu dùng có trải nghiệm tốt, họ sẽ cho sản phẩm điểm rất cao, ít có sự dao động giữa các mức sao khác nhau.
- Những mức sao thấp như 1 sao, 2 sao, và 3 sao có thể cho thấy những trường hợp phản hồi tiêu cực rất hiếm khi xuất hiện trong nhóm này.

Cảnh báo về sự thiếu phản hồi:

- Nếu không có nhiều đánh giá sao thấp trong nhóm tích cực, điều này có thể phản ánh rằng người tiêu dùng thường chỉ đánh giá khi họ thực sự có ấn tượng mạnh mẽ (cả tích cực và tiêu cực). Điều này có thể khiến các doanh nghiệp khó khăn trong việc nhận được phản hồi chi tiết về các điểm cần cải thiện nếu chỉ tập trung vào những phản hồi 5 sao.

Phân phối đánh giá sao cho Sentiment NEG (Tiêu Cực):

Đánh giá 5 sao vẫn chiếm ưu thế trong nhóm tiêu cực:

- Một điều thú vị là mặc dù nhóm cảm xúc NEG chủ yếu biểu thị sự không hài lòng, số lượng đánh giá 5 sao vẫn chiếm tỷ lệ lớn. Điều này có thể phản ánh rằng mặc dù sản phẩm có thể không hoàn toàn đạt kỳ vọng của người tiêu dùng, nhưng chất lượng sản phẩm vẫn được đánh giá cao ở một mức độ nào đó.
- Có thể hiểu rằng một số người dùng trong nhóm này chỉ không hài lòng với một số khía cạnh khác ngoài chất lượng sản phẩm, ví dụ như thời gian giao hàng chậm, dịch vụ khách hàng không tốt, hoặc vấn đề liên quan đến đóng gói. Đây có thể là lý do cho sự đánh giá cao về chất lượng sản phẩm, nhưng vẫn có phản hồi tiêu cực về các yếu tố khác.

Tỷ lệ đánh giá thấp (1 sao và 2 sao):

- Các mức sao 1 sao và 2 sao mặc dù ít, nhưng vẫn có thể cung cấp những thông tin quý giá về các vấn đề nghiêm trọng mà người tiêu dùng gặp phải. Những đánh giá này có thể chỉ ra các lỗi lớn liên quan đến chất lượng sản phẩm, chẳng hạn như sản phẩm hỏng, không giống mô tả, hoặc không đáp ứng được nhu cầu cơ bản của người dùng.
- Tỷ lệ thấp của các đánh giá này trong nhóm NEG có thể cho thấy sự phân bố không đều của cảm xúc tiêu cực, với một số người có cảm nhận tiêu cực mạnh mẽ hơn những người khác.

Nhận xét Tổng Quan:

- Sự phổ biến của 5 sao: Dù trong nhóm POS hay NEG, đánh giá 5 sao chiếm ưu thế. Điều này có thể phản ánh thực tế là sản phẩm đã tạo ra sự hài lòng rất lớn ở đa số người tiêu dùng. Tuy nhiên, nhóm NEG cho thấy rằng mặc dù sản phẩm có thể không tồi, vẫn có những yếu tố khác (như dịch vụ, giao hàng, v.v.) làm giảm trải nghiệm của người mua.
- Thiếu sự phản hồi phân bổ hợp lý:

  Các đánh giá 1 sao, 2 sao và 3 sao trong cả hai nhóm đều có số lượng thấp, đặc biệt trong nhóm POS. Điều này có thể phản ánh rằng các đánh giá đều khá cực đoan, tức là người dùng chỉ chia sẻ cảm xúc khi họ có trải nghiệm rất tốt hoặc rất tồi tệ, thay vì có một loạt các phản hồi ở các mức trung gian.
- Tạo cơ hội cho cải thiện:

  Biểu đồ trong nhóm NEG giúp nhận thấy rằng dù có sự phản hồi tiêu cực về các yếu tố ngoài sản phẩm, nhưng chất lượng sản phẩm vẫn được đánh giá khá cao trong mắt người tiêu dùng. Đây là cơ hội để các doanh nghiệp cải thiện các dịch

vụ đi kèm, chẳng hạn như dịch vụ giao hàng, chăm sóc khách hàng, hoặc các chính sách bảo hành.

- Sự mất cân đối trong cảm xúc:

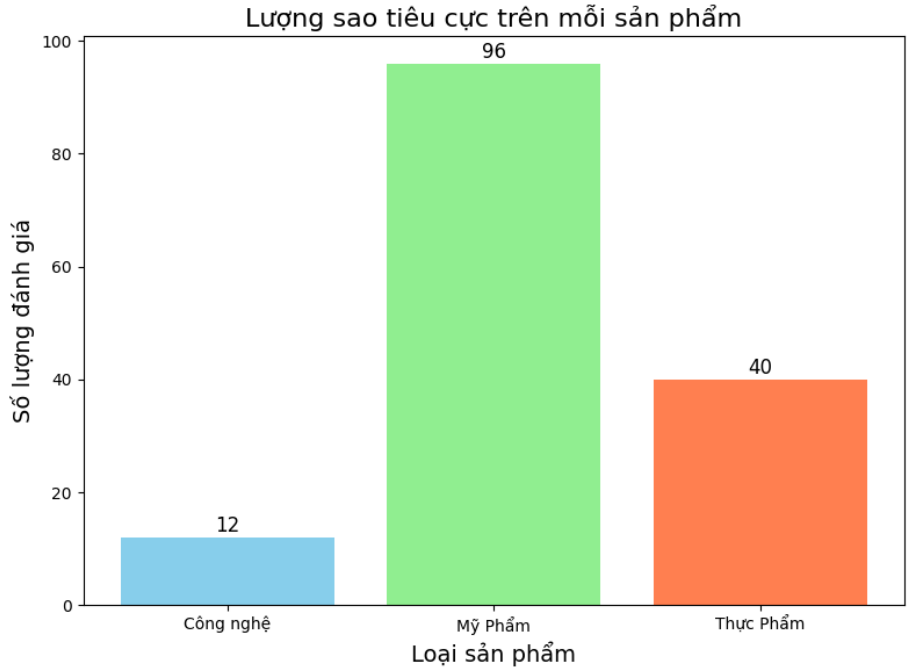
  Phân tích này cũng cho thấy sự mất cân đối giữa phản hồi tích cực và phản hồi tiêu cực. Điều này có thể chỉ ra rằng có sự thiếu sót trong việc thu thập các phản hồi tiêu cực chi tiết hơn, và các doanh nghiệp cần cải thiện phương thức thu thập ý kiến của người tiêu dùng để hiểu rõ hơn về những yếu tố cần cải thiện.

Kết luận: Biểu đồ này cung cấp cái nhìn sâu sắc về cách người tiêu dùng đánh giá sản phẩm và trải nghiệm của họ. Dù sản phẩm có thể được đánh giá cao về chất lượng, vẫn cần phải cải thiện các yếu tố khác như dịch vụ khách hàng và quá trình giao hàng để giảm thiểu những phản hồi tiêu cực và cải thiện sự hài lòng tổng thể của khách hàng.

b) Biểu đồ lượng sao tiêu cực trên mỗi sản phẩm

```
import matplotlib.pyplot as plt

# Dữ liệu

categories = ['Công nghệ', 'Mỹ Phẩm', 'Thực Phẩm']

positive_ratings = [

    cong_nghe_df[cong_nghe_df['Rating'] < 2].size,

    my_pham_df[my_pham_df['Rating'] < 2].size,

    thuc_pham_df[thuc_pham_df['Rating'] < 2].size

]

# Vẽ biểu đồ cột

plt.figure(figsize=(8, 6))

colors = ['skyblue', 'lightgreen', 'coral']  # Màu sắc cho từng thể loại

plt.bar(categories, positive_ratings, color=colors)

# Thiết lập tiêu đề và nhãn trục

plt.title('Lượng sao tiêu cực trên mỗi sản phẩm', fontsize=16)

plt.xlabel('Loại sản phẩm', fontsize=14)

plt.ylabel('Số lượng đánh giá', fontsize=14)

# Hiển thị giá trị trên cột

for i, value in enumerate(positive_ratings):

    plt.text(i, value + 1, str(value), ha='center', fontsize=12)

# Hiển thị biểu đồ

plt.tight_layout()

plt.show()
```

Kết quả hiển thị:



Hình 3. 11. Biểu đồ thể hiện lượng sao tiêu cực trên mỗi sản phẩm

Nhận xét chi tiết về biểu đồ

Phân phối số lượng đánh giá tiêu cực giữa các loại sản phẩm:

Mỹ phẩm:

- Mỹ phẩm có số lượng đánh giá tiêu cực cao nhất với 96 đánh giá, chiếm ưu thế rõ rệt so với hai loại sản phẩm còn lại. Điều này cho thấy rằng có một số vấn đề lớn hoặc sự không hài lòng đáng kể từ người tiêu dùng đối với các sản phẩm mỹ phẩm. Các nguyên nhân có thể bao gồm chất lượng sản phẩm không như mong đợi, hiệu quả kém, hoặc phản ứng dị ứng đối với sản phẩm.
- Tuy nhiên, số lượng đánh giá tiêu cực lớn này cũng có thể phản ánh sự phổ biến rộng rãi của sản phẩm mỹ phẩm, dẫn đến nhiều phản hồi hơn từ khách hàng.

Thực phẩm:

- Thực phẩm có 40 đánh giá tiêu cực, số lượng không nhỏ, nhưng thấp hơn rất nhiều so với Mỹ phẩm. Điều này cho thấy rằng, mặc dù có một số vấn đề tiêu cực liên quan đến sản phẩm thực phẩm, nhưng tỉ lệ người tiêu dùng không hài lòng có vẻ ít hơn, có thể vì những kỳ vọng về chất lượng thực phẩm có phần dễ chấp nhận hơn mỹ phẩm. Một số vấn đề có thể bao gồm chất lượng thực phẩm không đảm bảo, hết hạn, hoặc không giống với mô tả sản phẩm.

Công nghệ:

- Công nghệ có số lượng đánh giá tiêu cực thấp nhất, chỉ với 12 đánh giá. Sự không hài lòng đối với các sản phẩm công nghệ có thể xuất phát từ một số vấn đề về

tính năng, chất lượng hoặc lỗi kỹ thuật, nhưng tỉ lệ này là thấp so với mỹ phẩm và thực phẩm. Một khả năng là sản phẩm công nghệ có thể được lựa chọn và kiểm tra kỹ lưỡng hơn trước khi mua, dẫn đến ít trường hợp thất vọng.

Những điểm cần lưu ý:

Khả năng các sản phẩm mỹ phẩm không đạt kỳ vọng:

- Sự gia tăng trong đánh giá tiêu cực của nhóm mỹ phẩm có thể chỉ ra rằng người tiêu dùng có những kỳ vọng rất cao đối với các sản phẩm này, và khi sản phẩm không đáp ứng được, họ dễ dàng để lại phản hồi tiêu cực hơn.

Sự phổ biến của sản phẩm và lượng phản hồi:

- Các sản phẩm mỹ phẩm có thể có lượng người tiêu dùng lớn hơn và phản hồi nhiều hơn, điều này có thể làm tăng số lượng đánh giá tiêu cực. Sản phẩm công nghệ và thực phẩm, mặc dù được tiêu thụ phổ biến, nhưng có thể không được phản ánh nhiều qua các đánh giá tiêu cực.

Cần cải thiện các yếu tố khác ngoài chất lượng:

- Các vấn đề ngoài chất lượng, như dịch vụ khách hàng, giao hàng hay giá cả, có thể góp phần vào số lượng đánh giá tiêu cực ở tất cả các loại sản phẩm, đặc biệt là trong nhóm mỹ phẩm.

Kết luận:

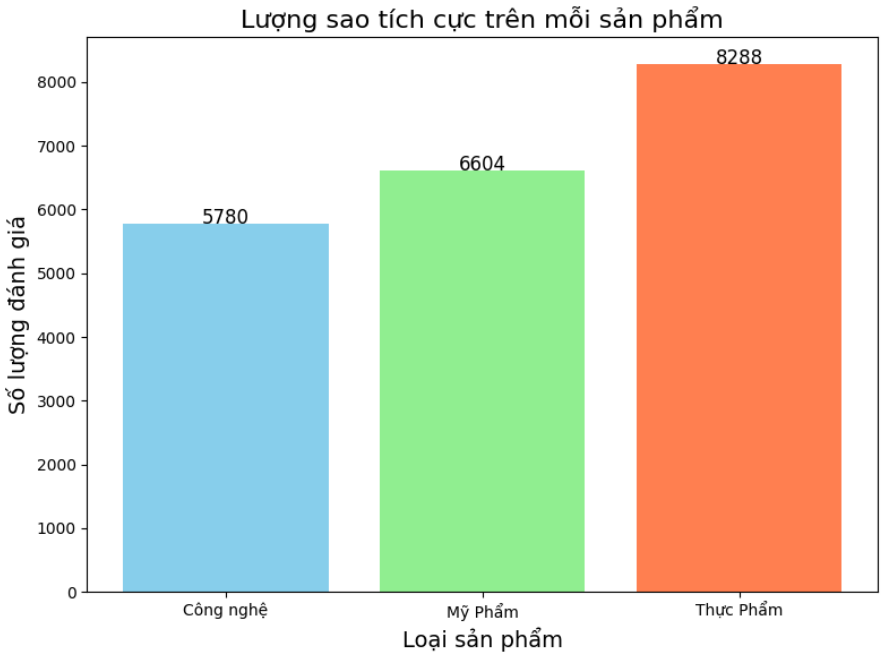
- Biểu đồ này cho thấy rằng mỹ phẩm là nhóm sản phẩm gặp phải nhiều phản hồi tiêu cực nhất, trong khi công nghệ và thực phẩm có lượng đánh giá tiêu cực thấp hơn nhiều. Các nhà sản xuất có thể cần phải cải thiện chất lượng hoặc các yếu tố liên quan khác như dịch vụ khách hàng để giảm thiểu số lượng phản hồi tiêu cực, đặc biệt là đối với nhóm mỹ phẩm.

c) Biểu đồ lượng sao tích cực trên mỗi sản phẩm

```
import matplotlib.pyplot as plt

# Dữ liệu

categories = ['Công nghệ', 'Mỹ Phẩm', 'Thực Phẩm']

positive_ratings = [

    cong_nghe_df[cong_nghe_df['Rating'] > 2].size,

    my_pham_df[my_pham_df['Rating'] > 2].size,

    thuc_pham_df[thuc_pham_df['Rating'] > 2].size

]

# Vẽ biểu đồ cột

plt.figure(figsize=(8, 6))

colors = ['skyblue', 'lightgreen', 'coral']  # Màu sắc cho từng thể loại
```

```
plt.bar(categories, positive_ratings, color=colors)

# Thiết lập tiêu đề và nhãn trục

plt.title('Lượng sao tích cực trên mỗi sản phẩm', fontsize=16)

plt.xlabel('Loại sản phẩm', fontsize=14)

plt.ylabel('Số lượng đánh giá', fontsize=14)

# Hiển thị giá trị trên cột

for i, value in enumerate(positive_ratings):

    plt.text(i, value + 1, str(value), ha='center', fontsize=12)

# Hiển thị biểu đồ

plt.tight_layout()

plt.show()
```

Kết quả hiển thị:



Hình 3. 12. Biểu đồ thể hiện lượng sao tích cực trên mỗi sản phẩm

Nhận xét chi tiết về biểu đồ

Phân phối số lượng đánh giá tích cực giữa các loại sản phẩm:

Thực phẩm:

- Thực phẩm có số lượng sao tích cực cao nhất với 8,288 đánh giá, vượt trội so với các loại sản phẩm còn lại. Điều này có thể chỉ ra rằng các sản phẩm thực phẩm thường nhận được sự hài lòng lớn từ người tiêu dùng. Một yếu tố có thể là chất lượng sản phẩm ổn định, dễ chấp nhận và người tiêu dùng có xu hướng đánh giá cao về sự tươi ngon, hương vị, hoặc tính an toàn của thực phẩm.

- Việc thực phẩm nhận được lượng đánh giá tích cực lớn cũng có thể phản ánh sự phổ biến rộng rãi và mức độ tiêu thụ cao của các sản phẩm này.

Công nghệ:

- Công nghệ có 5,780 đánh giá tích cực, là mức trung bình so với các loại sản phẩm còn lại. Mặc dù công nghệ là ngành được yêu thích và tiêu thụ mạnh mẽ, nhưng mức độ hài lòng lại không bằng thực phẩm và mỹ phẩm. Điều này có thể phản ánh sự đa dạng trong chất lượng sản phẩm công nghệ. Các sản phẩm công nghệ có thể có mức độ phức tạp cao hơn, có thể khó đáp ứng kỳ vọng của tất cả người dùng, gây ra sự không hài lòng dù số lượng đánh giá tích cực vẫn lớn.

Mỹ phẩm:

- Mỹ phẩm có 6,604 đánh giá tích cực, chỉ kém thực phẩm một chút. Điều này cho thấy nhiều người tiêu dùng rất hài lòng với các sản phẩm mỹ phẩm mà họ sử dụng. Tuy nhiên, mỹ phẩm thường có những tiêu chí khắt khe về hiệu quả, chất lượng, và các thành phần, điều này có thể giải thích tại sao số lượng đánh giá tích cực vẫn rất cao, mặc dù người tiêu dùng có xu hướng đánh giá khá cẩn thận khi mua các sản phẩm này.

Những điểm cần lưu ý:

Tình trạng tiêu dùng sản phẩm thực phẩm:

- Sự vượt trội của thực phẩm trong việc nhận sao tích cực có thể cho thấy rằng người tiêu dùng thực phẩm thường có sự hài lòng hơn về những sản phẩm này. Các yếu tố như chất lượng sản phẩm (ví dụ, thực phẩm tươi, an toàn, dễ sử dụng) và giá trị mang lại cho người tiêu dùng có thể là lý do khiến số lượng sao tích cực cao.

Đánh giá mỹ phẩm và công nghệ:

- Mặc dù cả mỹ phẩm và công nghệ đều có số lượng sao tích cực cao, nhưng sự khác biệt giữa hai loại sản phẩm này có thể do những yếu tố cụ thể của mỗi ngành. Với mỹ phẩm, người tiêu dùng có thể dễ dàng nhận thấy sự thay đổi sau khi sử dụng, trong khi với công nghệ, hiệu quả của sản phẩm đôi khi có thể khó nhận ra ngay lập tức hoặc có thể phụ thuộc vào cách sử dụng của mỗi người.

Tầm quan trọng của việc giữ chất lượng ổn định:

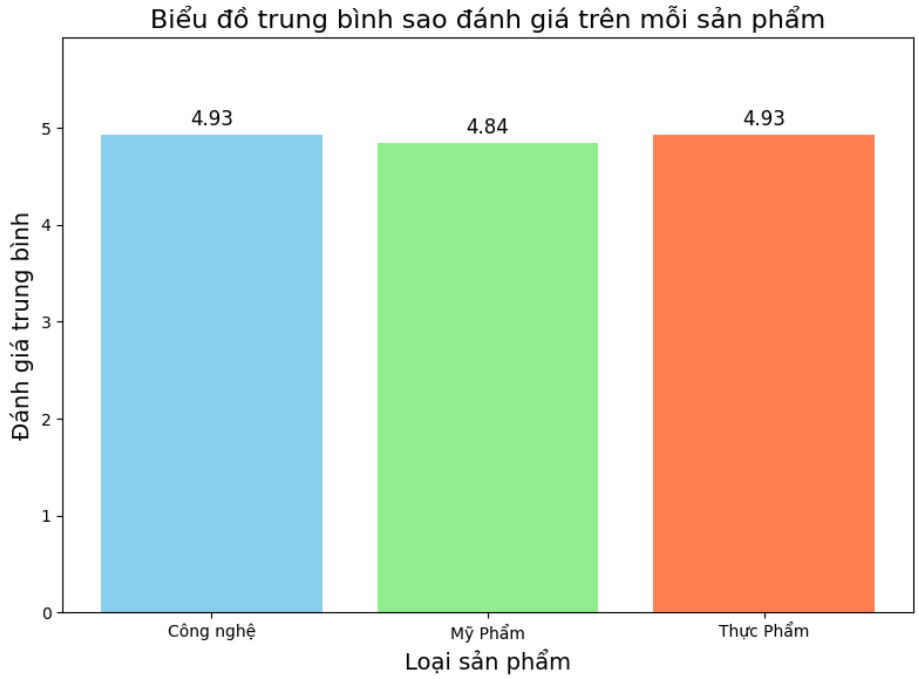
- Mặc dù các sản phẩm thực phẩm nhận được số lượng sao tích cực lớn nhất, các nhà sản xuất cần duy trì chất lượng ổn định để đảm bảo sự tiếp tục nhận được đánh giá tích cực. Tương tự, với mỹ phẩm và công nghệ, việc cải tiến liên tục và đáp ứng kỳ vọng người dùng là điều quan trọng để duy trì sự hài lòng của khách hàng.

Kết luận: Biểu đồ cho thấy các sản phẩm Thực phẩm nhận được nhiều sao tích cực nhất, tiếp theo là Mỹ phẩm, và cuối cùng là Công nghệ. Điều này cho thấy sự hài lòng của người tiêu dùng với các sản phẩm thực phẩm có xu hướng cao hơn, có thể do yếu tố chất lượng và sự an toàn. Các sản phẩm công nghệ và mỹ phẩm, mặc dù cũng nhận được nhiều đánh giá tích cực, nhưng cũng có thể đối mặt với sự đa dạng trong chất lượng, đặc biệt khi sản phẩm có sự phức tạp hơn về tính năng và công dụng.

d) Biểu đồ trung bình sao đánh giá trên mỗi sản phẩm

```
import matplotlib.pyplot as plt

# Dữ liệu

categories = ['Công nghệ', 'Mỹ Phẩm', 'Thực Phẩm']

mean_ratings = [

    cong_nghe_df['Rating'].mean(),

    my_pham_df['Rating'].mean(),

    thuc_pham_df['Rating'].mean()

]

# Vẽ biểu đồ cột

plt.figure(figsize=(8, 6))

colors = ['skyblue', 'lightgreen', 'coral']  # Màu sắc cho từng thể loại

plt.bar(categories, mean_ratings, color=colors

# Thiết lập tiêu đề và nhãn trục

plt.title('Biểu đồ trung bình sao đánh giá trên mỗi sản phẩm', fontsize=16)

plt.xlabel('Loại sản phẩm', fontsize=14)

plt.ylabel('Đánh giá trung bình', fontsize=14)

# Hiển thị giá trị trung bình trên cột

for i, value in enumerate(mean_ratings):

    plt.text(i, value + 0.1, f"{value:.2f}", ha='center', fontsize=12)

# Giới hạn trục y để hiển thị rõ hơn

plt.ylim(0, max(mean_ratings) + 1)

# Hiển thị biểu đồ

plt.tight_layout()

plt.show()
```

Kết quả hiển thị:



Hình 3. 13. Biểu đồ thể hiện trung bình sao đánh giá trên mỗi sản phẩm

Nhận xét chi tiết về biểu đồ

Sự phân bổ sao tích cực giữa các nhóm sản phẩm:

Sản phẩm Thực phẩm:

- Số lượng sao tích cực: 8,288 — Đây là nhóm sản phẩm có lượng sao tích cực lớn nhất trong ba nhóm. Điều này cho thấy người tiêu dùng có xu hướng hài lòng rất cao với các sản phẩm thực phẩm.
- Nguyên nhân: Một yếu tố có thể là sự phổ biến của các sản phẩm thực phẩm trong cuộc sống hàng ngày. Thực phẩm luôn là mặt hàng có nhu cầu tiêu thụ lớn và thường xuyên. Nếu chất lượng đảm bảo, người tiêu dùng sẽ có xu hướng hài lòng và chia sẻ đánh giá tích cực.
- Lý do sự chênh lệch lớn với các nhóm khác: Sự khác biệt này có thể do sự dễ dàng trong việc trải nghiệm sản phẩm thực phẩm: người tiêu dùng thường xuyên và dễ dàng đánh giá chất lượng thực phẩm hơn so với các sản phẩm công nghệ hay mỹ phẩm. Sản phẩm thực phẩm cũng thường có tiêu chí đánh giá rõ ràng như hương vị, độ tươi, và sự an toàn thực phẩm, giúp người tiêu dùng dễ dàng có những cảm nhận tích cực.

Sản phẩm Mỹ phẩm:

- Số lượng sao tích cực: 6,604 — Sản phẩm mỹ phẩm đứng thứ hai về số lượng sao tích cực, chỉ kém thực phẩm một chút. Điều này cho thấy ngành mỹ phẩm cũng nhận được sự quan tâm lớn từ người tiêu dùng và họ có xu hướng hài lòng với sản phẩm.

- Nguyên nhân: Mỹ phẩm có thể là một ngành có xu hướng đem lại sự hài lòng khi người tiêu dùng cảm nhận được hiệu quả ngay từ lần đầu sử dụng. Các sản phẩm như kem dưỡng da, sữa rửa mặt, hay mỹ phẩm trang điểm nếu chất lượng tốt sẽ tạo được niềm tin và sự trung thành từ khách hàng. Tuy nhiên, sự đánh giá mỹ phẩm có thể phức tạp hơn vì phụ thuộc vào từng loại da và yêu cầu sử dụng, do đó mặc dù số lượng sao tích cực cao, nhưng có thể có sự phân hóa nhỏ giữa các loại sản phẩm.
- Điểm mạnh: Ngành mỹ phẩm thường có sự thay đổi rõ rệt và dễ nhận thấy khi sản phẩm có tác dụng tốt, ví dụ như làn da mịn màng hơn, tóc khỏe mạnh hơn. Điều này có thể làm người tiêu dùng đánh giá tích cực hơn.

Sản phẩm Công nghệ:

- Số lượng sao tích cực: 5,780 — Sản phẩm công nghệ có lượng sao tích cực thấp nhất trong ba nhóm. Tuy nhiên, đây vẫn là một con số đáng kể.
- Nguyên nhân: Sản phẩm công nghệ có tính phức tạp cao hơn so với thực phẩm và mỹ phẩm. Việc đánh giá công nghệ phụ thuộc vào nhiều yếu tố như tính năng, độ bền, giá trị sử dụng và đặc biệt là sự phù hợp với nhu cầu cá nhân của mỗi người tiêu dùng. Người tiêu dùng có thể có kỳ vọng cao về công nghệ, và do đó khi sản phẩm không đáp ứng hoàn hảo các yêu cầu này, họ có thể không hài lòng. Sự hài lòng khi sử dụng công nghệ đôi khi còn bị chi phối bởi các yếu tố kỹ thuật, khả năng kết nối, tính dễ sử dụng, và sự tương thích với các thiết bị khác.

Những yếu tố ảnh hưởng đến sự khác biệt trong sự phân bổ sao tích cực:

- Đặc điểm tiêu dùng và sử dụng: Sản phẩm thực phẩm có thể dễ dàng được sử dụng và cảm nhận kết quả ngay lập tức (ví dụ như hương vị, độ tươi), trong khi các sản phẩm công nghệ có thể yêu cầu người tiêu dùng phải sử dụng lâu dài hoặc trải nghiệm thực tế lâu hơn mới có thể đánh giá chính xác. Sản phẩm công nghệ, ví dụ như điện thoại hay máy tính, có thể có lỗi kỹ thuật hoặc vấn đề với phần mềm, dẫn đến sự không hài lòng của người tiêu dùng.
- Sự đa dạng trong chất lượng sản phẩm: Ngành công nghệ có nhiều phân khúc sản phẩm từ các sản phẩm cao cấp đến các sản phẩm giá rẻ. Điều này có thể dẫn đến sự chênh lệch lớn trong cảm nhận của người tiêu dùng về chất lượng, và có thể là nguyên nhân dẫn đến lượng sao tích cực thấp hơn so với các nhóm khác.

Yếu tố cảm xúc:

- Trong ngành mỹ phẩm, nhiều người tiêu dùng có thể liên kết kết quả sản phẩm với yếu tố cảm xúc (sự tự tin, cảm giác đẹp hơn), vì vậy, khi mỹ phẩm mang lại hiệu quả tốt, cảm giác thỏa mãn sẽ rất mạnh mẽ, tạo ra xu hướng đánh giá tích cực.

- Ngược lại, trong ngành công nghệ, nếu sản phẩm gặp sự cố kỹ thuật, cảm giác thất vọng sẽ lớn hơn.

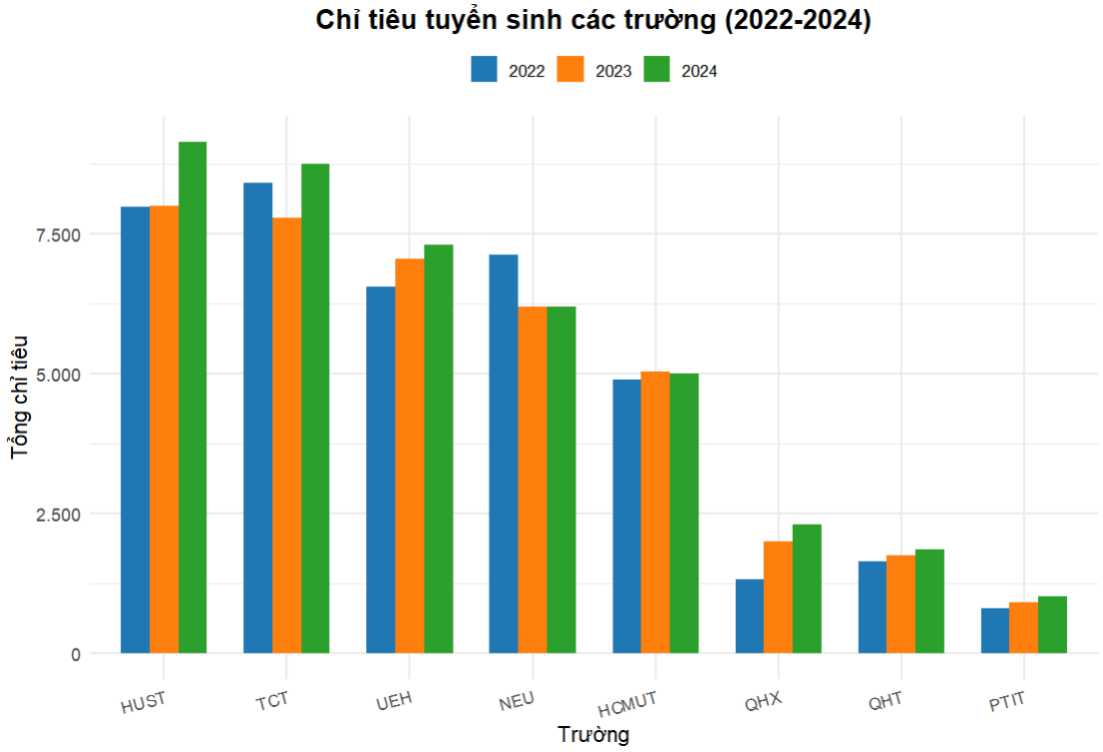
Kết luận:

- Sản phẩm thực phẩm nổi bật với số lượng đánh giá sao tích cực cao nhất. Người tiêu dùng cảm thấy dễ dàng hài lòng với các sản phẩm thực phẩm do chất lượng dễ nhận biết và trải nghiệm nhanh chóng.
- Mỹ phẩm có sự hài lòng cao, phản ánh mức độ cải thiện rõ rệt trong quá trình sử dụng, nhưng không bằng thực phẩm vì có sự phân hóa theo yêu cầu sử dụng cá nhân.
- Công nghệ có số lượng sao tích cực ít nhất, điều này có thể do sự đa dạng về sản phẩm và kỳ vọng cao từ người tiêu dùng. Người tiêu dùng trong ngành này có thể khó tính hơn vì sản phẩm công nghệ yêu cầu sự đổi mới liên tục và tính ổn định cao.
- ***Nhìn chung,*** để duy trì sự hài lòng và tăng số lượng sao tích cực, các nhà sản xuất cần đảm bảo chất lượng sản phẩm ổn định, cải tiến không ngừng, và đáp ứng đúng kỳ vọng của người tiêu dùng trong từng nhóm ngành.

3.2.3. Biểu đồ trực quan hóa dữ liệu với dữ liệu ngành học

a) Tổng chỉ tiêu của các trường đại học qua các năm

Tổng chỉ tiêu tuyển sinh của các trường qua các năm



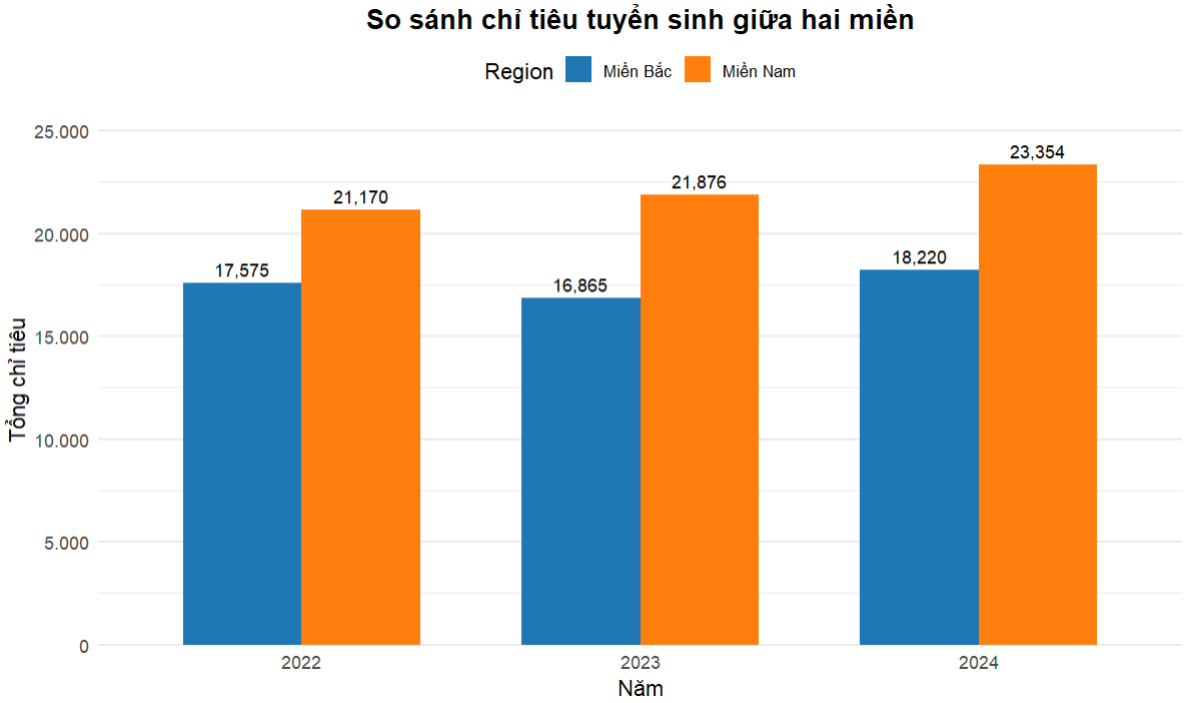
Hình 3. 14. Biểu đồ thể hiện chỉ tiêu tuyển sinh các trường (2022-2024)

Nhận xét:

Biểu đồ thể hiện tổng chỉ tiêu tuyển sinh của một số trường đại học từ năm 2022 đến năm 2024 cho thấy xu hướng tương đối ổn định hoặc tăng nhẹ qua các năm.

- Trong đó, Đại học Bách Khoa Hà Nội (HUST) và TCT là hai trường dẫn đầu về quy mô chỉ tiêu tuyển sinh, với mức tăng rõ rệt từ năm 2023 đến 2024.
- Trường Đại học Kinh tế TP.HCM (UEH) cũng ghi nhận mức tăng nhẹ và ổn định qua các năm, và Đại học Kinh tế Quốc dân (NEU) từ năm 2022 đến năm 2023, 2024 đã giảm chỉ tiêu.
- Đáng chú ý, TCT có sự giảm chỉ tiêu trong năm 2023 nhưng nhanh chóng phục hồi vào năm 2024. Ở nhóm các trường có chỉ tiêu tuyển sinh thấp hơn như QHX, QHT và PTIT, xu hướng chung là tăng nhẹ, trong đó QHX có mức tăng trưởng rõ rệt nhất, gần gấp đôi so với năm 2022.
- Nhìn chung, biểu đồ cho thấy phần lớn các trường đều có chiến lược mở rộng hoặc duy trì quy mô tuyển sinh, phản ánh nhu cầu học đại học ngày càng cao và sự thích nghi với tình hình giáo dục sau đại dịch.

b) Chỉ tiêu giữa 2 miền Bắc – Nam



Hình 3. 15. Biểu đồ so sánh chỉ tiêu tuyển sinh giữa hai miền Bắc - Nam

Nhận xét:

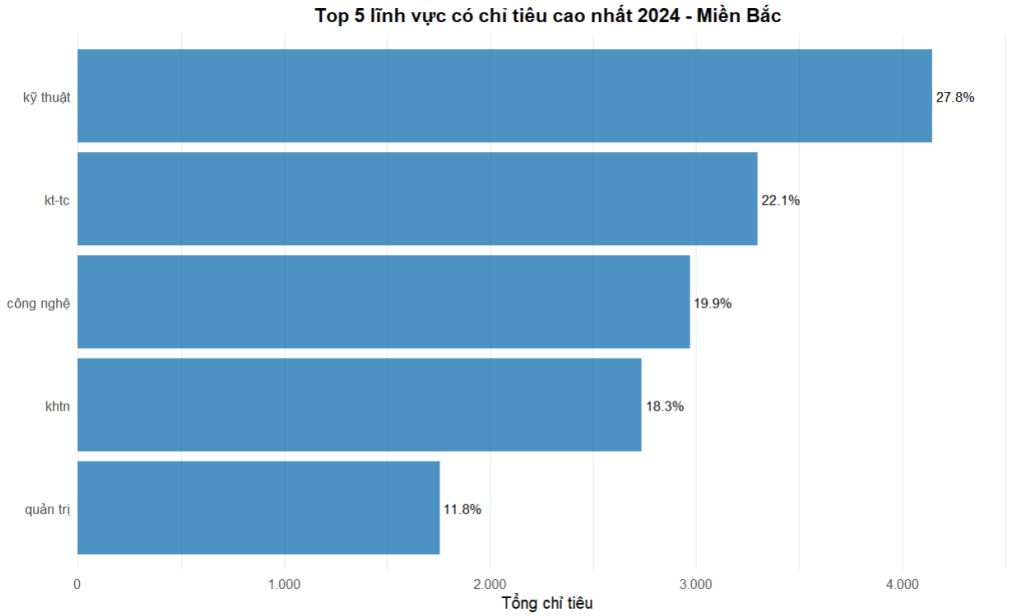
- Biểu đồ cho thấy sự chênh lệch rõ rệt về chỉ tiêu tuyển sinh giữa hai miền Bắc và Nam trong ba năm liên tiếp từ 2022 đến 2024. Trong cả ba năm, miền Nam luôn duy trì mức chỉ tiêu cao hơn đáng kể so với miền Bắc, cho thấy định hướng phát triển giáo dục mạnh mẽ hơn tại khu vực phía Nam.
- Cụ thể, năm 2022, miền Nam đã vượt miền Bắc tới gần 3.600 chỉ tiêu (21.170 so với 17.575). Sang năm 2023, dù tổng chỉ tiêu của miền Bắc giảm nhẹ xuống còn

16.865, thì miền Nam vẫn tiếp tục tăng lên 21.876, nới rộng khoảng cách lên hơn 5.000 chỉ tiêu. Đáng chú ý, đến năm 2024, cả hai miền đều có xu hướng tăng trở lại, trong đó miền Bắc tăng lên 18.220, nhưng miền Nam vẫn duy trì mức tăng mạnh lên tới 23.354, tiếp tục giữ vững vị thế dẫn đầu với mức cách biệt hơn 5.100 chỉ tiêu.

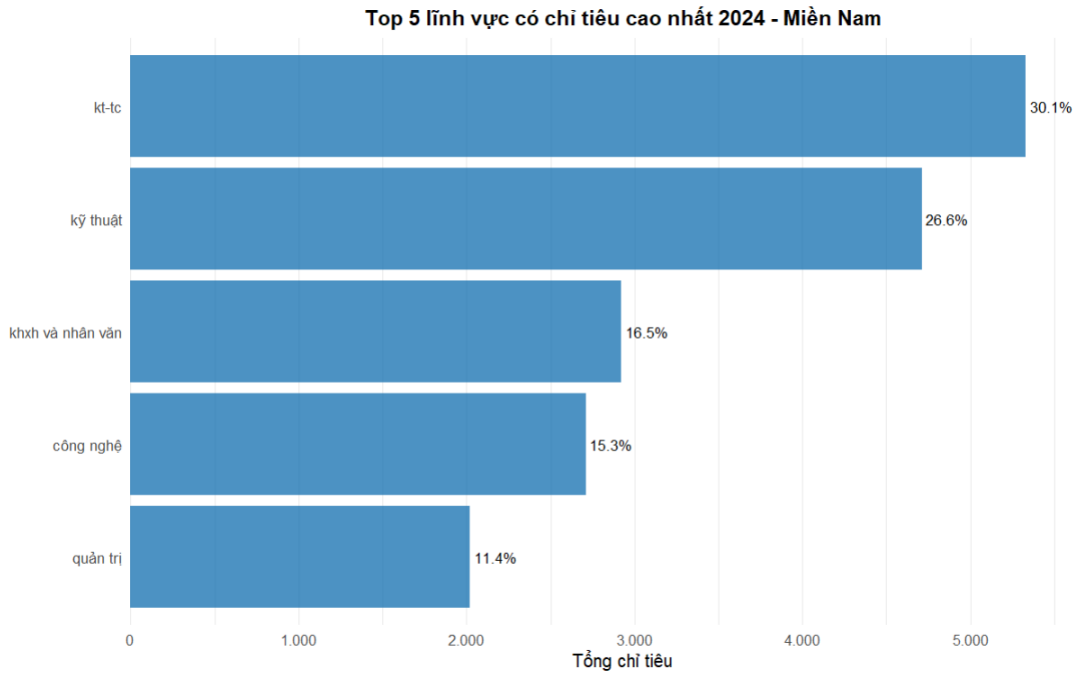
- Điều này phản ánh sự ưu tiên đầu tư vào giáo dục của miền Nam, có thể đến từ các yếu tố như: mật độ dân số cao hơn, nhu cầu nhân lực dồi dào tại các trung tâm kinh tế lớn (TP.HCM, Bình Dương, Đồng Nai…), cũng như sự phát triển mạnh mẽ của hệ thống các cơ sở giáo dục đại học trong khu vực.

Tóm lại, qua ba năm, có thể thấy miền Nam không chỉ duy trì mức chỉ tiêu cao mà còn ngày càng gia tăng khoảng cách với miền Bắc, điều này mở ra nhiều cơ hội học tập hơn cho học sinh khu vực phía Nam, đồng thời đặt ra yêu cầu cho miền Bắc cần có chiến lược phân bổ và phát triển đào tạo hiệu quả hơn trong thời gian tới.

c) Top 5 các lĩnh vực có chỉ tiêu cao nhất giữa các miền



Hình 3. 16. Biểu đô thể hiện top 5 lĩnh vực có chỉ tiêu cao nhất 2024 – miền Bắc



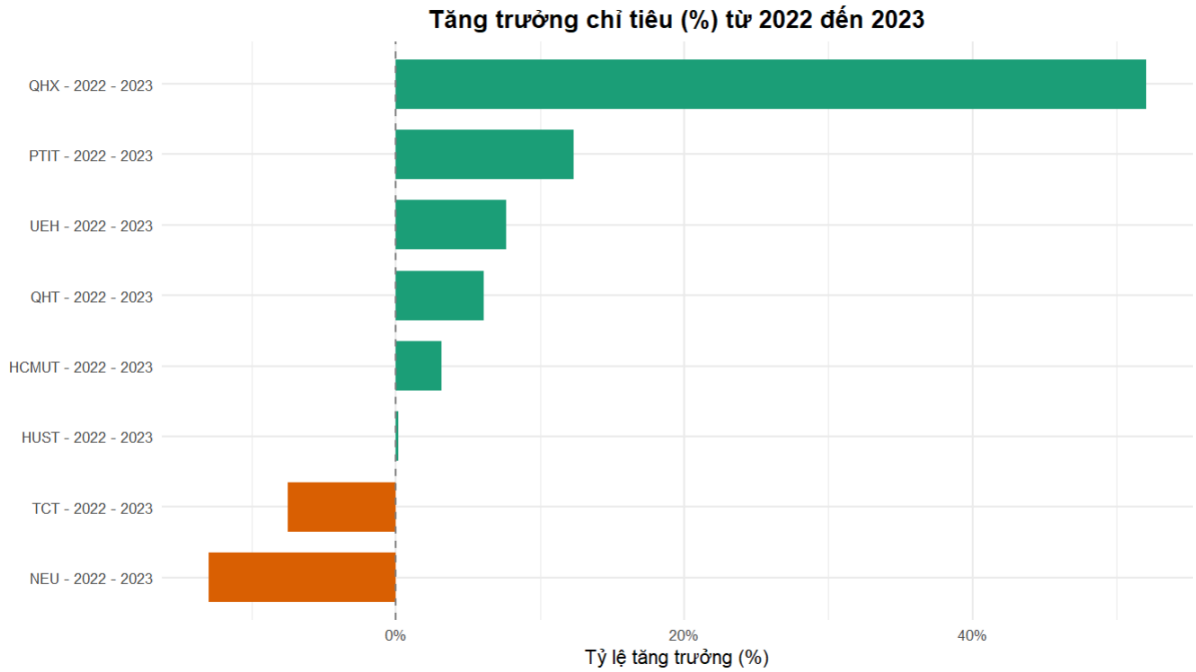
Hình 3. 17. Biểu đồ thể hiện top 5 lĩnh vực có chỉ tiêu cao nhất 2024 – miền Nam

Nhận xét:

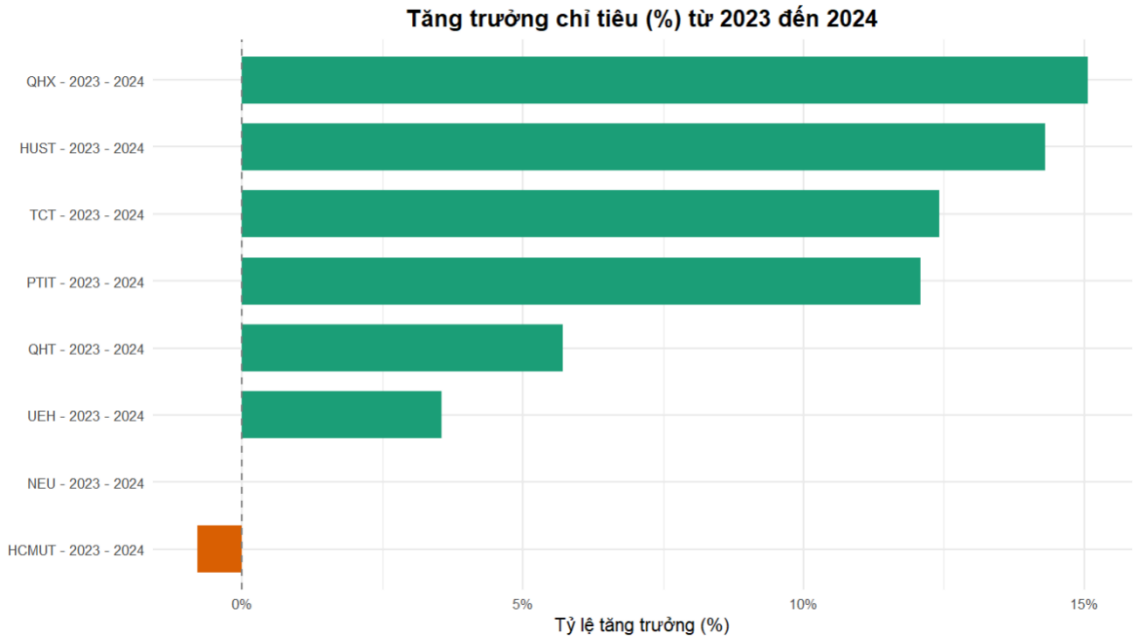
- Hai biểu đồ trên thể hiện Top 5 lĩnh vực có chỉ tiêu cao nhất năm 2024 tại miền Nam và miền Bắc, qua đó cho thấy những khác biệt rõ nét trong định hướng đào tạo giữa hai khu vực. Tại miền Nam, lĩnh vực Kinh tế - Tài chính (kt-tc) chiếm tỷ trọng cao nhất với 30.1%, cho thấy khu vực này đang ưu tiên phát triển mạnh nguồn nhân lực trong lĩnh vực kinh tế. Đứng thứ hai là Kỹ thuật với 26.6%, tạo nên một cặp dẫn đầu vượt trội. Trong khi đó, tại miền Bắc, lĩnh vực Kỹ thuật lại giữ vị trí đầu bảng với 27.8%, tiếp theo là kt-tc (22.1%). So với miền Nam, miền Bắc có sự phân bổ chỉ tiêu đều hơn giữa các ngành trong top 5, thể hiện cách tiếp cận cân bằng hơn giữa các lĩnh vực đào tạo.
- Một điểm chung đáng chú ý là cả hai miền đều có sự hiện diện của các ngành như Công nghệ và Quản trị, cho thấy đây là những lĩnh vực có sức hút ổn định và được chú trọng trên phạm vi toàn quốc. Tuy nhiên, mỗi miền cũng có những đặc trưng riêng: miền Nam ưu tiên cho Khoa học xã hội và Nhân văn, trong khi miền Bắc lại dành nhiều chỉ tiêu hơn cho Khoa học tự nhiên. "Quản trị" là lĩnh vực đứng cuối trong top 5 ở cả hai miền, nhưng vẫn giữ vai trò quan trọng trong tổng thể chiến lược đào tạo.

Nhìn chung, sự khác biệt trong phân bổ chỉ tiêu phản ánh định hướng phát triển nguồn nhân lực mang tính vùng miền. Miền Nam có xu hướng tập trung mạnh vào các lĩnh vực kinh tế và kỹ thuật, trong khi miền Bắc cho thấy sự quan tâm đồng đều hơn đến nhiều nhóm ngành khác nhau — tạo nên bức tranh sinh động và đa dạng của giáo dục đại học Việt Nam năm 2024.

d) Biểu đồ sự tăng trưởng qua các năm của các trường

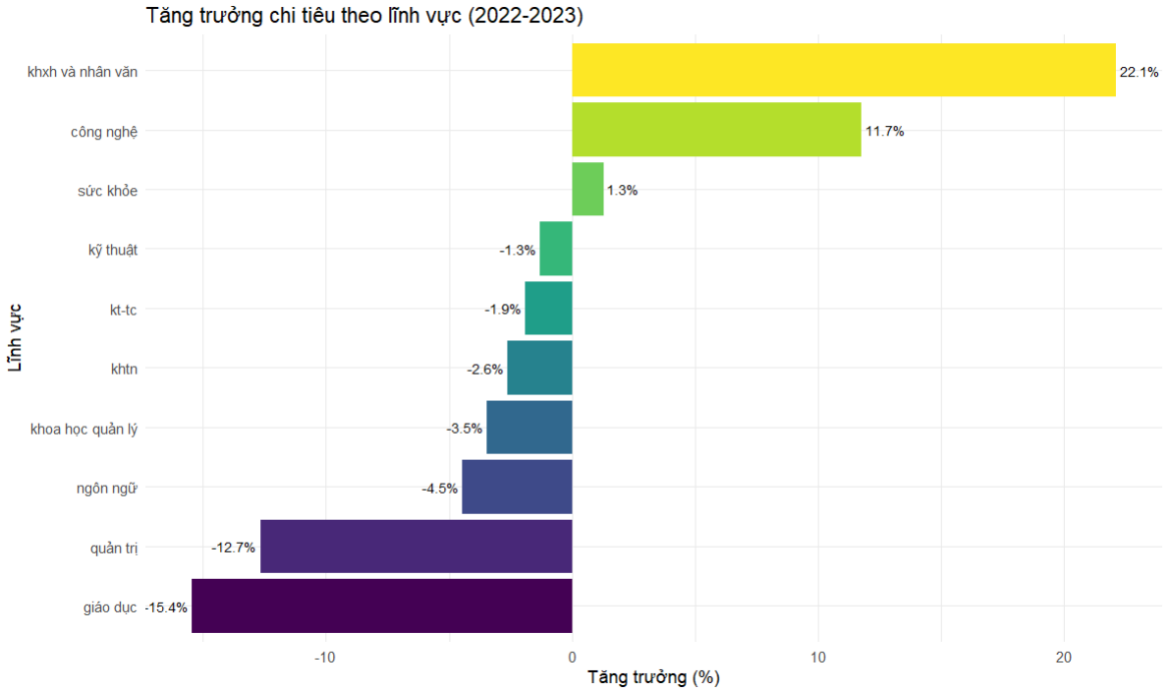


Hình 3. 18. Biểu đồ thể hiện tỷ lệ tăng trưởng chỉ tiêu (%) từ 2022 đến 2023

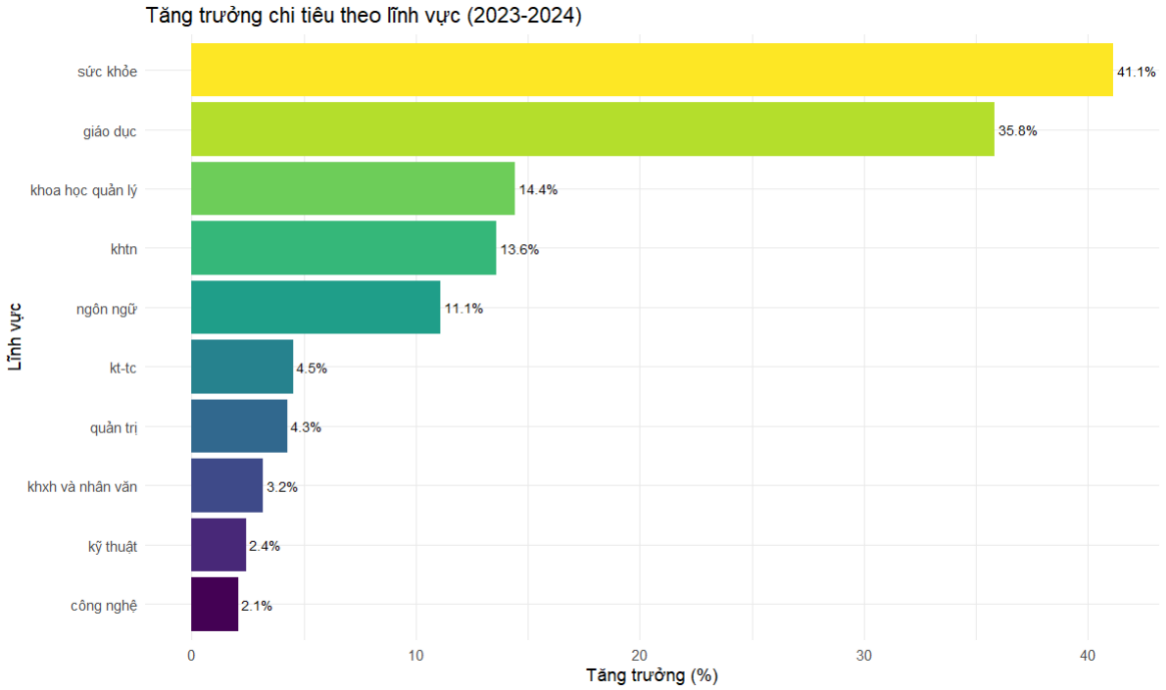


Hình 3. 19. Biểu đồ thể hiện tỷ lệ tăng trưởng từ năm 2023 đến 2024

e) Biểu đồ tăng trưởng theo lĩnh vực ở cả 2 miền (đã gộp)



Hình 3. 20. Biểu đồ thể hiện tỷ lệ tăng trưởng chỉ tiêu theo lĩnh vực (2022-2023)



Hình 3. 21. Biểu đồ thể hiện tỷ lệ tăng trưởng chỉ tiêu theo lĩnh vực (2023-2024)

Nhận xét:

- Hai biểu đồ thể hiện sự biến động đáng chú ý trong phân bổ chỉ tiêu tuyển sinh giữa các lĩnh vực qua hai giai đoạn 2022–2023 và 2023–2024, phản ánh những thay đổi trong định hướng chiến lược đào tạo.
- Trong giai đoạn 2022–2023, lĩnh vực khoa học xã hội và nhân văn ghi nhận mức tăng trưởng ấn tượng nhất với 22,1%, tiếp theo là công nghệ (11,7%), trong khi phần lớn các lĩnh vực khác như quản trị, giáo dục, ngôn ngữ đều ghi nhận mức

sụt giảm mạnh, đặc biệt là giáo dục giảm tới -15,4%. Điều này cho thấy sự điều chỉnh chỉ tiêu nghiêng nhiều về các ngành mang tính xã hội và công nghệ trong bối cảnh hậu đại dịch, khi nhu cầu nhân lực trong các lĩnh vực này tăng cao.

- Tuy nhiên, bước sang giai đoạn 2023–2024, bức tranh phân bổ chỉ tiêu có sự đảo chiều rõ rệt. Sức khỏe vươn lên dẫn đầu với mức tăng trưởng vượt bậc 41,1%, theo sau là giáo dục (35,8%) – một sự phục hồi ngoạn mục sau năm suy giảm trước đó. Cùng với đó, các ngành như khoa học quản lý, khoa học tự nhiên và ngôn ngữ cũng cho thấy xu hướng tăng mạnh trở lại. Ngược lại, những lĩnh vực từng tăng mạnh như công nghệ, kỹ thuật, khxh và nhân văn lại rơi xuống nhóm tăng trưởng thấp nhất, cho thấy sự điều chỉnh linh hoạt của các cơ sở đào tạo theo nhu cầu nhân lực thực tế.

Tổng thể, hai biểu đồ không chỉ phản ánh sự thay đổi trong chiến lược phân bổ chỉ tiêu mà còn cho thấy xu hướng phát triển của thị trường lao động và những lĩnh vực được ưu tiên đầu tư trong từng giai đoạn. Đây là cơ sở quan trọng để học sinh, phụ huynh và nhà trường có cái nhìn toàn diện hơn trong việc định hướng nghề nghiệp tương lai.

Tóm lại, nếu như năm 2022–2023 là năm của KHXH và Công nghệ, thì năm 2023–2024 lại là năm bứt phá mạnh mẽ của Sức khỏe và Giáo dục. Đây là minh chứng rõ ràng cho sự linh hoạt, thích ứng nhanh của chiến lược đào tạo trước những biến động xã hội và kinh tế, đồng thời cũng là định hướng quý giá cho thí sinh và phụ huynh trong việc lựa chọn ngành nghề phù hợp với xu thế tương lai.

CHƯƠNG 4: MỐI LIÊN HỆ VÀ PHÂN TÍCH XU HƯỚNG GIỮA CÁC BIẾN

4.1. GIỚI THIỆU VỀ MỐI LIÊN HỆ VÀ PHÂN TÍCH XU HƯỚNG GIỮA CÁC BIẾN

4.1.1. Mối liên hệ và phân tích xu hướng giữa các biến là gì ?

Mối liên hệ giữa các biến là cách mà các biến trong một tập dữ liệu hoặc một hệ thống tương tác và ảnh hưởng lẫn nhau. Nó thể hiện mức độ và cách thức một biến thay đổi khi các biến khác thay đổi. Mối liên hệ này có thể biểu thị qua nhiều hình thức, chẳng hạn như:

- **Tương quan**: Đo lường mức độ và chiều hướng liên kết giữa hai biến, thường được biểu diễn bằng hệ số tương quan. Ví dụ, nếu chiều cao tăng thì cân nặng cũng tăng, đó là tương quan dương. Nếu một biến tăng, biến kia giảm, đó là tương quan âm.
- **Nguyên nhân - Kết quả (Quan hệ nhân quả)**: Một biến này có thể trực tiếp ảnh hưởng hoặc gây ra sự thay đổi ở biến khác. Ví dụ, tăng cường quảng cáo có thể dẫn đến tăng doanh số bán hàng.
- **Mối quan hệ phi tuyến**: Không phải lúc nào mối liên hệ cũng là tuyến tính. Một biến có thể ảnh hưởng đến biến khác theo các hình thức phi tuyến, chẳng hạn như quan hệ parabol hoặc lôgarit.
- **Sự phụ thuộc có điều kiện**: Mối liên hệ giữa hai biến có thể thay đổi tùy thuộc vào sự hiện diện hoặc giá trị của một biến khác.

Hiểu mối liên hệ giữa các biến là nền tảng để phân tích dữ liệu, xây dựng mô hình dự đoán, và đưa ra quyết định dựa trên dữ liệu. Nó giúp xác định các yếu tố quan trọng, tối ưu hóa hiệu suất và khám phá các quy luật ẩn giấu trong hệ thống.

Phân tích xu hướng giữa các biến là quá trình nghiên cứu và đánh giá mối quan hệ thay đổi của hai hay nhiều biến trong một tập dữ liệu theo thời gian hoặc theo một chiều hướng cụ thể. Mục tiêu của quá trình này là xác định xem các biến có cùng biến động theo một xu hướng nhất định hay không, từ đó đưa ra nhận định, dự đoán hoặc hỗ trợ cho việc ra quyết định. Chẳng hạn, khi phân tích dữ liệu từ mạng xã hội về mức độ quan tâm đến các ngành học ở đại học, ta có thể quan sát sự thay đổi về số lượng bài viết, lượt tìm kiếm hoặc mức độ tương tác theo từng năm để xác định ngành nào đang thu hút nhiều sự chú ý hơn. Việc phân tích xu hướng không chỉ dừng lại ở việc nhận diện chiều hướng tăng hay giảm, mà còn giúp phát hiện các mối liên hệ tuyến tính hoặc phi tuyến giữa các biến, từ đó áp dụng các kỹ thuật thống kê hoặc mô hình dự đoán phù hợp.

4.1.2. Mục đích của mối liên hệ và phân tích xu hướng giữa các biến

Mục đích của việc thiết lập mối liên hệ giữa các biến trong phân tích dữ liệu là để khám phá và hiểu rõ cách các yếu tố tương tác, từ đó đưa ra những kết luận có ý nghĩa và ứng dụng thực tế. Việc phân tích mối quan hệ giúp làm sáng tỏ các quy luật tiềm ẩn trong dữ liệu, như việc một biến có thể dự đoán giá trị của biến khác hoặc sự phụ thuộc giữa các yếu tố trong một hệ thống. Điều này không chỉ quan trọng trong việc xây dựng các mô hình dự báo và tối ưu hóa chiến lược mà còn giúp xác định nguyên nhân gốc rễ của các hiện tượng, hỗ trợ giải quyết vấn đề một cách hiệu quả.

Hơn thế nữa, việc hiểu rõ mối liên hệ giữa các biến còn giúp loại bỏ những yếu tố nhiễu loạn, đảm bảo tính chính xác của phân tích và tối ưu hóa việc thu thập dữ liệu. Thông qua các công cụ trực quan hóa như biểu đồ phân tán, ma trận tương quan hay các kỹ thuật thống kê như hồi quy, nhà nghiên cứu có thể phát hiện ra những mô hình, xu hướng hoặc sự bất thường trong dữ liệu, từ đó đưa ra các quyết định dựa trên cơ sở khoa học vững chắc. Tóm lại, việc khám phá mối quan hệ giữa các biến không chỉ mở ra góc nhìn sâu sắc về bản chất của dữ liệu mà còn góp phần thúc đẩy những đổi mới và sáng tạo trong mọi lĩnh vực.

Bên cạnh mục đích của mối liên hệ thì mục đích của việc phân tích xu hướng của các biến là để nhận diện và hiểu rõ cách thức mà các biến thay đổi theo thời gian hoặc theo một trật tự nhất định. Việc phân tích này giúp phát hiện các chiều hướng tăng, giảm, dao động định kỳ hoặc những biến động bất thường, từ đó cung cấp cơ sở để đưa ra dự đoán, đánh giá hiệu quả và hỗ trợ cho quá trình ra quyết định. Trong các lĩnh vực như kinh doanh, giáo dục hay truyền thông, phân tích xu hướng giúp xác định được các hành vi, sở thích hay nhu cầu đang nổi lên trong cộng đồng. Ví dụ, khi phân tích xu hướng thảo luận về các ngành nghề đào tạo trên mạng xã hội, ta có thể xác định được ngành nào đang thu hút sự quan tâm cao, từ đó hỗ trợ việc tư vấn hướng nghiệp hoặc điều chỉnh chiến lược truyền thông. Nhìn chung, việc phân tích xu hướng là công cụ quan trọng để hiểu rõ sự thay đổi của dữ liệu trong bối cảnh thực tế và hoạch định các bước đi tiếp theo một cách hiệu quả hơn.

4.2. MỐI LIÊN HỆ VÀ PHÂN TÍCH XU HƯỚNG GIỮA CÁC BIẾN TRONG CÁC LĨNH VỰC

4.2.1. Mối liên hệ và phân tích xu hướng giữa các biến trong lĩnh vực phim

a) Khai báo thư viện và xử lý giá trị null

Đầu tiên, chúng ta cần khai báo các thư viện cần thiết:

```
import pandas as pd

from nltk.sentiment import SentimentIntensityAnalyzer

import nltk

import pandas as pd

from sklearn.feature_extraction.text import TfidfVectorizer
```

```
from sklearn.model_selection import train_test_split

from sklearn.linear_model import LogisticRegression

from sklearn.metrics import classification_report, confusion_matrix, accuracy_score
```

Để đảm bảo tính chính xác của mô hình và kết quả phân tích, tránh lỗi khi thực hiện các phép tính, tăng cường tính toàn vẹn của dữ liệu thì việc xử lý các giá trị bị thiếu trong dữ liệu là một bước quan trọng trong quá trình phân tích và xử lý dữ liệu. Dưới đây là đoạn code xử lý giá trị bị thiếu:

```
# Danh sách các DataFrame

dataframes = [deadpool_wolverin, ant_man, guardian_galaxy_3, the_boys, transformers,

              baby_lon, bad_boy, friends, murder_in_building, intouchable,

              breaking_bad, dexter, joker, monsters, tulsa_king]

# Xử lý giá trị null bằng vòng lặp

dataframes = [df.dropna() for df in dataframes]

# Kiểm tra xem còn giá trị null không trong từng DataFrame

null_counts = [df.isna().sum().sum() for df in dataframes]

print(f"Tổng số giá trị null còn lại trong mỗi DataFrame: {null_counts}")
```

Hiệu quả xử lý dữ liệu:

- Đoạn mã đã xử lý triệt để các giá trị bị thiếu (NaN) bằng cách xóa các hàng có dữ liệu không đầy đủ. Sau khi áp dụng, các DataFrame mới sẽ sạch hơn, dễ dàng cho các bước phân tích tiếp theo.

Hạn chế:

- Mất mát dữ liệu:
    - Việc sử dụng dropna() sẽ xóa toàn bộ các hàng bị thiếu giá trị, dẫn đến mất dữ liệu. Nếu dữ liệu bị thiếu ở nhiều cột hoặc có giá trị quan trọng, điều này có thể ảnh hưởng đến chất lượng phân tích.
- Không xử lý NaN theo chiến lược thay thế:
    - Trong một số trường hợp, thay vì loại bỏ, có thể cần thay thế NaN bằng giá trị khác (như trung bình, giá trị thường gặp nhất) để giữ lại nhiều dữ liệu hơn.

b) Hàm phân tích cảm xúc

Phân tích cảm xúc được thực hiện bằng cách sử dụng công cụ VADER từ thư viện NLTK. Công cụ này sử dụng từ điển cảm xúc và ngữ cảnh để đánh giá văn bản, trả về một điểm số cảm xúc compound:

- Compound ≥ 0.05: Tích cực.
- Compound ≤ -0.05: Tiêu cực.

- Còn lại: Trung lập.

Hàm phân tích cảm xúc:

```
    def analyze_sentiment(review):
    score = sia.polarity_scores(review)['compound']
    if score >= 0.05:
        return "POSITIVE"
    elif score <= -0.05:
        return "NEGATIVE"
    else:
        return "NEUTRAL"
```

c) Gán nhãn cảm xúc

Gán nhãn cảm xúc cho từng đánh giá trong mỗi DataFrame và kết hợp các bộ dữ liệu lại thành một tập dữ liệu tổng hợp.

Code Python:

```
processed_dataframes = []
for df in dataframes:
df['Review'] = df['Review'].fillna('').astype(str)
df['Sentiment'] = df['Review'].apply(analyze_sentiment)
processed_dataframes.append(df)
combined_df = pd.concat(processed_dataframes, ignore_index=True)
```

d) Chuẩn bị và huấn luyện mô hình

Sau khi gắn nhãn cảm xúc, chúng tôi sử dụng TF-IDF Vectorizer để biểu diễn văn bản dưới dạng vector số học, sau đó xây dựng mô hình Logistic Regression để phân loại cảm xúc.

Code Python:

Vector hóa văn bản:

```
tfidf = TfidfVectorizer(stop_words='english', max_features=5000)
X = tfidf.fit_transform(combined_df['Review'])
y = combined_df['Sentiment']
```

Chia tập dữ liệu:

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)
```

Huấn luyện mô hình:

```
model = LogisticRegression()

model.fit(X_train, y_train)
```

e) Đánh giá mô hình

Mô hình được đánh giá dựa trên các chỉ số:

- **Accuracy**: Độ chính xác tổng thể.
- **Classification Report**: Báo cáo chi tiết theo từng lớp (Tích cực, Tiêu cực, Trung lập).
- **Confusion Matrix**: Ma trận nhầm lẫn.

Code Python:

```
y_pred = model.predict(X_test)

print("Accuracy:", accuracy_score(y_test, y_pred))

print("Classification Report:\n", classification_report(y_test, y_pred))
```

f) Thử nghiệm với dữ liệu mới

Mô hình được kiểm thử với các đánh giá mẫu:

```
sample_review = ["The movie was fantastic, great acting and visuals!",

        "Terrible story, I hated it."]

sample_vectorized = tfidf.transform(sample_review)

print("Sentiment predictions:", model.predict(sample_vectorized))
```

Kết quả :

- Mô hình đạt độ chính xác cao trong việc phân loại cảm xúc.
- Ví dụ**:**
  - "The movie was fantastic, great acting and visuals!" → **POSITIVE**
  - "Terrible story, I hated it." → **NEGATIVE**

Accuracy : 0.8812814070351759

Classification Report:

Label	Precision	Recall	F1-Score	Support
NEGATIVE	0.8	0.52	0.63	1065
NEUTRAL	0.93	0.91	0.92	1051
POSITIVE	0.88	0.96	0.92	4252
Accuracy			0.88	6368

Macro Avg	0.87	0.8	0.82	6368
Weighted Avg	0.88	0.88	0.87	6368

Sentiment prediction: [‘POSITIVE’ ‘NEGATIVE’]

Đánh giá tổng quan

- Độ chính xác (Accuracy): Mô hình đạt **88.13%**, cho thấy khả năng phân loại cảm xúc tương đối tốt dựa trên các đánh giá văn bản.
- Tổng hợp từ báo cáo phân loại (Classification Report):
  - Precision: Mức độ chính xác của dự đoán đúng trong từng lớp cảm xúc.
  - Recall: Tỷ lệ dự đoán đúng trên tổng số mẫu thực tế thuộc mỗi lớp.
  - F1-Score: Đo lường sự cân bằng giữa Precision và Recall.

Phân tích chi tiết theo từng lớp cảm xúc

POSITIVE:

- Precision: 88% (Trong tất cả các đánh giá được dự đoán là tích cực, 88% là chính xác).
- Recall: 96% (Hầu hết các đánh giá tích cực thực sự được mô hình phát hiện).
- F1-Score**:** 92% (Hiệu suất rất cao).
- Nhận xét: Mô hình xử lý tốt các đánh giá tích cực, phản ánh sự ưu thế của lớp này do dữ liệu không cân bằng (lớp này có số lượng lớn nhất: 4,252 mẫu).

NEUTRAL:

- Precision: 93% (Độ chính xác trong việc phát hiện các đánh giá trung lập rất cao).
- Recall: 91% (Khả năng phát hiện đánh giá trung lập gần như hoàn chỉnh).
- F1-Score: 92%.
- Nhận xét: Lớp trung lập được xử lý tốt, mô hình có khả năng xác định rõ các đánh giá không nghiêng về tích cực hay tiêu cực.

NEGATIVE:

- Precision: 80% (Khá thấp so với các lớp khác, cho thấy mô hình gặp khó khăn trong việc phát hiện chính xác các đánh giá tiêu cực).
- Recall: 52% (Chỉ hơn một nửa các đánh giá tiêu cực thực sự được phát hiện).
- F1-Score: 63% (Hiệu suất thấp hơn nhiều so với hai lớp còn lại).

- Nhận xét: Lớp tiêu cực là điểm yếu của mô hình, có thể do số lượng dữ liệu tiêu cực ít hơn hoặc tính chất của các đánh giá tiêu cực phức tạp hơn.

Về dự đoán dữ liệu mới

- Mô hình dự đoán '**POSITIVE**' cho đánh giá tích cực và '**NEGATIVE**' cho đánh giá tiêu cực trong tập mẫu thử nghiệm:
  - "The movie was fantastic, great acting and visuals!" → POSITIVE.
  - "Terrible story, I hated it." → NEGATIVE.
- Nhận xét: Dự đoán của mô hình phù hợp với nội dung đánh giá, cho thấy khả năng xử lý dữ liệu mới là khả quan.

Đây là quan sát sơ bộ về dữ liệu:

```
deadpool_wolverin_ =
pd.read_csv('https://raw.githubusercontent.com/NguyenKhang0062/-n-
1/refs/heads/main/data_imdb/data_setiment/deadpool_wolverin_.csv')
ant_man_ = pd.read_csv('https://raw.githubusercontent.com/NguyenKhang0062/-n-
1/refs/heads/main/data_imdb/data_setiment/ant_man_.csv')
guardian_galaxy_3_ =
pd.read_csv('https://raw.githubusercontent.com/NguyenKhang0062/-n-
1/refs/heads/main/data_imdb/data_setiment/guardian_galaxy_3_.csv')
the_boys_ = pd.read_csv('https://raw.githubusercontent.com/NguyenKhang0062/-n-
1/refs/heads/main/data_imdb/data_setiment/the_boys_.csv')
transformers_ = pd.read_csv('https://raw.githubusercontent.com/NguyenKhang0062/-n-
1/refs/heads/main/data_imdb/data_setiment/transformers_.csv')
baby_lon_ = pd.read_csv('https://raw.githubusercontent.com/NguyenKhang0062/-n-
1/refs/heads/main/data_imdb/data_setiment/baby_lon_.csv')
bad_boy_ = pd.read_csv('https://raw.githubusercontent.com/NguyenKhang0062/-n-
1/refs/heads/main/data_imdb/data_setiment/bad_boy_.csv')
friends_ = pd.read_csv('https://raw.githubusercontent.com/NguyenKhang0062/-n-
1/refs/heads/main/data_imdb/data_setiment/friends_.csv')
murder_in_building_ =
pd.read_csv('https://raw.githubusercontent.com/NguyenKhang0062/-n-
1/refs/heads/main/data_imdb/data_setiment/murder_in_building_.csv')
intouchable_ = pd.read_csv('https://raw.githubusercontent.com/NguyenKhang0062/-n-
1/refs/heads/main/data_imdb/data_setiment/intouchable_.csv')
breaking_bad_ = pd.read_csv('https://raw.githubusercontent.com/NguyenKhang0062/-n-
1/refs/heads/main/data_imdb/data_setiment/breaking_bad_.csv')
```

```
dexter_      =      pd.read_csv('https://raw.githubusercontent.com/NguyenKhang0062/-n-1/refs/heads/main/data_imdb/data_setiment/dexter_.csv')

joker_       =      pd.read_csv('https://raw.githubusercontent.com/NguyenKhang0062/-n-1/refs/heads/main/data_imdb/data_setiment/joker_.csv')

monsters_    =      pd.read_csv('https://raw.githubusercontent.com/NguyenKhang0062/-n-1/refs/heads/main/data_imdb/data_setiment/monsters_.csv')

tulsa_king_    =    pd.read_csv('https://raw.githubusercontent.com/NguyenKhang0062/-n-1/refs/heads/main/data_imdb/crime/tulsa_king_.csv')
```

Kết quả hiển thị:

	Rating	Review_Date	Release_Date	Review_clean	Sentiment
0	9.00	25-Jul-24	July 26, 2024 (United States)	hugh jackman is the perfect wolverine what a f...	POSITIVE
1	9.00	24-Jul-24	July 26, 2024 (United States)	what a crazy blast bonkers sooo what i can say...	POSITIVE
2	8.00	24-Jul-24	July 26, 2024 (United States)	weve waited so long for this moment and it was...	POSITIVE
3	1.00	24-Jul-24	July 26, 2024 (United States)	so many easter eggs so true to the comic chara...	POSITIVE
4	9.00	26-Jul-24	July 26, 2024 (United States)	i read an ign review where the guy gave it a b...	POSITIVE
...	...	...	...	...	...
1302	4.87	1-Oct-24	July 26, 2024 (United States)	tldr turn off your brain for hours pretend to ...	NEGATIVE
1303	4.87	28-Jul-24	July 26, 2024 (United States)	subjectively i understand the title as fun eno...	NEGATIVE
1304	4.87	12-Aug-24	July 26, 2024 (United States)	try to ignore much of the bad reviews it is re...	POSITIVE
1305	4.87	3-Oct-24	July 26, 2024 (United States)	the first movies in this trilogy had a great c...	NEGATIVE
1306	4.87	3-Oct-24	July 26, 2024 (United States)	deadpool ryan reynolds is living an empty life...	POSITIVE

1307 rows × 5 columns

```
[ ] comedy_df = pd.read_csv('https://raw.githubusercontent.com/NguyenKhang0062/-n-1/refs/heads/main/data_imdb/act_sci.csv')
```

```
[ ] comedy_df
```

	Unnamed: 0	Rating	Review_Date	Release_Date	Review_clean	Sentiment	Genres
0	0	9.0	25-Jul-24	July 26, 2024 (United States)	hugh jackman is the perfect wolverine what a f...	POSITIVE	ACtion
1	1	9.0	24-Jul-24	July 26, 2024 (United States)	what a crazy blast bonkers sooo what i can say...	POSITIVE	ACtion
2	2	8.0	24-Jul-24	July 26, 2024 (United States)	weve waited so long for this moment and it was...	POSITIVE	ACtion
3	3	1.0	24-Jul-24	July 26, 2024 (United States)	so many easter eggs so true to the comic chara...	POSITIVE	ACtion
4	4	9.0	26-Jul-24	July 26, 2024 (United States)	i read an ign review where the guy gave it a b...	POSITIVE	ACtion
...	...	...	...	...	...	...	...
6497	6497	9.0	Jul 18, 2023	June 9, 2023 (United States)	transformers rise of the beasts took everythin...	POSITIVE	ACtion
6498	6498	3.0	Jun 10, 2023	June 9, 2023 (United States)	a film similar to the previous ones with some ...	NEGATIVE	ACtion
6499	6499	3.0	Jul 15, 2023	June 9, 2023 (United States)	honestly the last couple acts saved this movie...	POSITIVE	ACtion
6500	6500	4.0	Jul 15, 2023	June 9, 2023 (United States)	rise of the beasts but doesnt feel like a beas...	NEGATIVE	ACtion
6501	6501	10.0	Jul 15, 2023	June 9, 2023 (United States)	solid here and far better than the last few of...	POSITIVE	ACtion

6502 rows × 7 columns

```
[ ] synthetic  = pd.concat([act_sci_df,comedy_df,crime_df],ignore_index=True)
```

```
[ ] synthetic
```

	Rating	Review_Date	Release_Date	Review_clean	Sentiment
0	9.00	25-Jul-24	July 26, 2024 (United States)	hugh jackman is the perfect wolverine what a f...	POSITIVE
1	9.00	24-Jul-24	July 26, 2024 (United States)	what a crazy blast bonkers sooo what i can say...	POSITIVE
2	8.00	24-Jul-24	July 26, 2024 (United States)	weve waited so long for this moment and it was...	POSITIVE
3	1.00	24-Jul-24	July 26, 2024 (United States)	so many easter eggs so true to the comic chara...	POSITIVE
4	9.00	26-Jul-24	July 26, 2024 (United States)	i read an ign review where the guy gave it a b...	POSITIVE
...	...	...	...	...	...
27355	5.00	23 January 2023	November 13, 2022 (United States)	excellent premise decent acting but the writin...	NEGATIVE
27356	7.17	19 January 2023	November 13, 2022 (United States)	this is the weakest sheridan effort ive seen s...	NEGATIVE
27357	7.17	15 October 2024	November 13, 2022 (United States)	i like this show i really do i think it has gr...	NEGATIVE
27358	7.17	1 April 2024	November 13, 2022 (United States)	disclaimer i dropped this after one episode my...	NEGATIVE
27359	7.17	26 September 2024	November 13, 2022 (United States)	actually this would have been so great with cl...	NEGATIVE

27360 rows × 5 columns

Để thực hiện các bước xử lý và chuẩn bị dữ liệu liên quan đến các thể loại phim trong Python. Đầu tiên, ta loại bỏ các cột không cần thiết, như các cột chỉ số hoặc cột chứa thông tin dư thừa, nhằm làm sạch dữ liệu trong từng DataFrame. Sau đó, dữ liệu từ nhiều nguồn được kết hợp lại theo từng thể loại, gồm **Action**, **Comedy**, và **Crime**, để tạo ra các tập dữ liệu tổng hợp. Tiếp theo, đoạn code thêm một cột mới để gắn nhãn thể loại cho từng tập dữ liệu, giúp phân biệt các dòng dữ liệu theo từng nhóm. Cuối

cùng, các DataFrame đã được xử lý được lưu vào file CSV trên Google Drive, phục vụ cho các bước phân tích hoặc xây dựng mô hình sau này. Việc kết nối Google Drive đảm bảo rằng các file được lưu trữ trực tiếp và dễ dàng truy cập.

```
deadpool_wolverin_.drop(['Unnamed: 0.1', 'Unnamed: 0','Review'], axis=1,inplace = True)
#Kết hợp các thể loại lại với nhau
act_sci_df = pd.concat([deadpool_wolverin_,ant_man_,guardian_galaxy_3_,the_boys_,transformers_], ignore_index=True)
comedy_df = pd.concat([baby_lon_,bad_boy_,friends_,murder_in_building_,intouchable_], ignore_index=True)
crime_df = pd.concat([breaking_bad_,dexter_,joker_,monsters_,tulsa_king_], ignore_index=True)
act_sci_df.drop(['Unnamed: 0.1', 'Unnamed: 0','Review'], axis=1,inplace = True)
comedy_df.drop(['Unnamed: 0.1', 'Unnamed: 0','Review'], axis=1,inplace = True)
crime_df.drop(['Unnamed: 0.1', 'Unnamed: 0','Review'], axis=1,inplace =True)
act_sci_df['Genres'] = 'ACtion'
comedy_df['Genres'] = 'Comedy'
crime_df['Genres'] = 'Crime'
from google.colab import drive
drive.mount('/content/drive')
act_sci_df.to_csv('/content/drive/My Drive/act_sci.csv')
comedy_df.to_csv('/content/drive/My Drive/comedy.csv')
crime_df.to_csv('/content/drive/My Drive/crime.csv')
```

4.2.2. Mối liên hệ và phân tích xu hướng giữa các biến trong lĩnh vực sản phẩm

a) Khai báo thư viện và xử lý giá trị null

Đầu tiên, ta cần khai báo các thư viện cần thiết:

```
import re
import nltk
from nltk.corpus import stopwords
from sklearn.naive_bayes import MultinomialNB
from sklearn.model_selection import train_test_split
```

```
from sklearn.metrics import accuracy_score
```

Để đảm bảo độ chính xác của mô hình và kết quả phân tích, tránh sai sót trong quá trình tính toán, đồng thời nâng cao tính toàn vẹn của dữ liệu, việc xử lý các giá trị bị thiếu là một bước không thể thiếu trong quy trình phân tích và xử lý dữ liệu. Dưới đây là một đoạn mã cách xử lý những giá trị bị thiếu trong dữ liệu sản phẩm.

```
#Công nghệ

fix_Camera_Toan_Cau = Camera_Toan_Cau.dropna().reset_index(drop=True)

fix_EZVIZ_VN_Authorized_Store =
EZVIZ_VN_Authorized_Store.dropna().reset_index(drop=True)

fix_Zentino = Zentino.dropna().reset_index(drop=True)

#Mỹ phẩm

fix_Guardian = Guardian.dropna().reset_index(drop=True)

fix_HadaLabo = HadaLabo.dropna().reset_index(drop=True)

fix_KITY_COSMETIC = KITY_COSMETIC.dropna().reset_index(drop=True)

#Thực phẩm chức năg

fix_Cosin_Store = Cosin_Store.dropna().reset_index(drop=True)

fix_DHC = DHC.dropna().reset_index(drop=True)

fix_Konni39 = Konni39.dropna().reset_index(drop=True)
```

Nhận xét:

Nhóm 1: Công nghệ

- Mục đích:
  - Loại bỏ tất cả các dòng có giá trị thiếu (NaN) trong các DataFrame liên quan đến lĩnh vực công nghệ.
  - Thiết lập lại chỉ số (reset_index(drop=True)) để đảm bảo tính liên tục sau khi loại bỏ dữ liệu.
- Nhận xét:
  - Quy trình này đảm bảo rằng dữ liệu đầu ra không chứa các giá trị thiếu, giúp thuận tiện cho việc phân tích hoặc mô hình hóa.
  - reset_index(drop=True) giúp loại bỏ chỉ số cũ, tránh sự xuất hiện của các chỉ số rời rạc hoặc bị lỗi sau khi xóa các dòng.

Nhóm 2: Mỹ phẩm

- Mục đích:
  - Thực hiện tương tự như nhóm công nghệ nhưng áp dụng cho các DataFrame liên quan đến lĩnh vực mỹ phẩm.

- Nhận xét:
    - Logic của mã này nhất quán và hiệu quả trong việc xử lý giá trị thiếu.
    - Tuy nhiên, không có kiểm tra nào được thực hiện để xác minh rằng việc xóa dữ liệu thiếu có làm mất thông tin quan trọng hay không (nếu tỷ lệ dữ liệu thiếu lớn).

Nhóm 3: Thực phẩm chức năng

- Mục đích:
    - Xử lý giá trị thiếu cho các DataFrame liên quan đến thực phẩm chức năng.

Nhận xét tổng thể

- Ưu điểm:
    - Đoạn mã được tổ chức hợp lý, tách biệt từng nhóm ngành hàng giúp dễ đọc và dễ bảo trì.
    - Sử dụng các phương pháp cơ bản nhưng hiệu quả:
    - dropna() loại bỏ giá trị thiếu, giúp làm sạch dữ liệu.
    - reset_index(drop=True) đảm bảo chỉ số được thiết lập lại liên tục, loại bỏ chỉ số không cần thiết.
- Hạn chế:
    - Xử lý toàn diện giá trị thiếu:
    - Mã chỉ đơn thuần loại bỏ các dòng chứa giá trị thiếu mà không kiểm tra tỷ lệ dữ liệu bị mất. Nếu tỷ lệ dữ liệu thiếu lớn, việc này có thể làm giảm đáng kể kích thước dữ liệu và mất đi thông tin quan trọng.
- Có thể cân nhắc:
    - Sử dụng phương pháp điền giá trị thiếu (imputation), ví dụ:
    - Điền giá trị trung bình (mean), giá trị phổ biến (mode).
    - Điền giá trị đặc biệt như "unknown" hoặc "N/A" (với dữ liệu phân loại).

b) Làm sạch văn bản

Tương tự, làm sạch văn bản trong Python là một bước quan trọng trong quá trình xử lý ngôn ngữ tự nhiên (NLP) nhằm chuẩn bị dữ liệu văn bản để phân tích hoặc huấn luyện mô hình. Quá trình này bao gồm việc loại bỏ các yếu tố không cần thiết và chuẩn hóa văn bản để dễ dàng xử lý hơn. Làm sạch văn bản là một phần không thể thiếu trong mọi bài toán NLP, vì dữ liệu sạch và nhất quán sẽ giúp nâng cao chất lượng và hiệu quả của các bước xử lý tiếp theo. Sau đây ta sẽ sử dụng đoạn code sau để làm sạch văn bản:

```
# Hàm tiền xử lý văn bản

def preprocess(text):

    if isinstance(text, str):  # Kiểm tra nếu text là chuỗi
```

```
        # Chuyển về chữ thường
        text = text.lower()
        # Chỉ giữ lại chữ cái, chữ số và dấu tiếng
text = re.sub(r'[^\w\sàáảãạâầấẩẫậăằắẳẵặèéẻẽẹêềếểễệìíỉĩịòóỏõọôồốổỗộơờớởỡợùúủũụưừứửữựỳýỷỹỵđĐ]', ' ', text)
        # Loại bỏ khoảng trắng thừa
        text = re.sub(r'\s+', ' ', text).strip()
        # Loại bỏ ký tự xuống dòng
        text = text.replace('\n', ' ').replace('\r', ' ')
        return text
    else:
        return ""
# Áp dụng tiền xử lý cho cột "Review"
fix_Camera_Toan_Cau['Review'] =fix_Camera_Toan_Cau['Review'].apply(preprocess)
fix_EZVIZ_VN_Authorized_Store['Review'] = fix_EZVIZ_VN_Authorized_Store['Review'].apply(preprocess)
fix_Zentino['Review'] = fix_Zentino['Review'].apply(preprocess)
fix_Guardian['Review'] = fix_Guardian['Review'].apply(preprocess)
fix_HadaLabo['Review'] = fix_HadaLabo['Review'].apply(preprocess)
fix_KITY_COSMETIC['Review'] = fix_KITY_COSMETIC['Review'].apply(preprocess)
fix_Cosin_Store['Review'] = fix_Cosin_Store['Review'].apply(preprocess)
fix_DHC['Review'] = fix_DHC['Review'].apply(preprocess)
fix_Konni39['Review'] = fix_Konni39['Review'].apply(preprocess)
```

Đây là quan sát sơ bộ về kết quả:

```
[ ] fix_Konni39
```

	Review	Rating	Date
0	sản phẩm rất ok sẽ ủng hộ shop lâu dài lần sau...	1	15 thg 11 2021
1	sản phẩm chất lượng giao hàng nhanh chóng đáng...	3	17 thg 5 2024
2	sản phẩm rất ok sẽ ủng hộ shop lâu dài lần sau...	4	15 thg 11 2021
3	sản phẩm chất lượng giao hàng nhanh chóng đáng...	4	17 thg 5 2024
4	đã mua của shop lần thứ n sản phẩm tốt mình vẫ...	4	21 thg 5 2022
...	...	...	...
460	đã nhận đầy đủ hàng check mã vạch có ra sp tác...	5	24 thg 2 2022
461	ko nên mua phục vụ như hàng chợ	5	03 thg 2 2023
462	hàng cận date	5	27 thg 4 2021
463	đợt này shop giao hàng hơi lâu	5	28 thg 9 2022
464	em đặt gói 60 ngày 120 viên mà shop gửi không ...	5	11 thg 8 2022

465 rows × 3 columns

c) Hàm phân tích cảm xúc

Để thực hiện việc tải và thiết lập một mô hình phân tích cảm xúc bằng cách sử dụng mô hình **'5CD-AI/Vietnamese-Sentiment-visobert'**, một mô hình đã được huấn luyện chuyên biệt cho tiếng Việt. Đầu tiên, **tokenizer** (bộ tách từ) và **model** (mô hình phân loại) được tải xuống từ thư viện transformers thông qua tên mô hình. Sau đó, một pipeline phân tích cảm xúc được tạo ra bằng hàm **pipeline** với tham số "sentiment-analysis". Pipeline này kết hợp mô hình và tokenizer để nhận đầu vào là văn bản tiếng Việt và đưa ra kết quả phân loại cảm xúc, chẳng hạn như tích cực, tiêu cực hoặc trung tính. Việc thiết lập này giúp đơn giản hóa quy trình dự đoán cảm xúc từ văn bản và dễ dàng áp dụng vào các bài toán xử lý ngôn ngữ tự nhiên.

```
# Tải mô hình và tokenizer

model_name = '5CD-AI/Vietnamese-Sentiment-visobert'

tokenizer = AutoTokenizer.from_pretrained(model_name)

model = AutoModelForSequenceClassification.from_pretrained(model_name)

# Tạo pipeline phân tích cảm xúc

sentiment_pipeline = pipeline("sentiment-analysis", model=model, tokenizer=tokenizer)
```

Hàm phân tích cảm xúc và kết quả hiển thị:

```
# Hàm phân tích cảm xúc

def analyze_sentiment(Review):

    try:

        result = sentiment_pipeline(Review)

        label = result[0]['label']  # Lấy nhãn cảm xúc từ kết quả

        return label

    except Exception as e:

        return "error"  # Trả về "error" nếu có lỗi

# Hàm xử lý phân tích cảm xúc cho tất cả DataFrame trong danh sách

def process_sentiment_for_all_dfs(list_of_dfs):

    for i, df in enumerate(list_of_dfs):

        print(f"Đang xử lý DataFrame {i+1}...")

        # Phân tích cảm xúc cho cột Review

        df['Sentiment'] = df['Review'].apply(analyze_sentiment)

        # Phân loại dữ liệu thành các nhóm cảm xúc

        df_positive = df[df['Sentiment'] == 'POSITIVE']

        df_negative = df[df['Sentiment'] == 'NEGATIVE']
```

```
        df_neutral = df[df['Sentiment'] == 'NEUTRAL']
        # In ra hoặc lưu kết quả (tuỳ vào yêu cầu)
        print(f"DataFrame {i+1}: Phân tích cảm xúc xong.")
    print("Hoàn thành phân tích cảm xúc cho tất cả DataFrame!")
# Giả sử bạn có một danh sách các DataFrame
list_product =
[fix_Camera_Toan_Cau,fix_EZVIZ_VN_Authorized_Store,fix_Zentino,fix_Guardian,fix_Hada
Labo,fix_KITY_COSMETIC,fix_Cosin_Store,fix_DHC,fix_Konni39]
# Gọi hàm để xử lý tất cả DataFrame trong list
process_sentiment_for_all_dfs(list_product)
```

Đây là dữ liệu cảm xúc của sản phẩm sau khi được xử lý:

```
fix_Camera_Toan_Cau =
pd.read_csv(r"https://raw.githubusercontent.com/NguyenKhang0062/-n-
1/refs/heads/main/e-commerce_data/fix_Camera_Toan_Cau.csv")
fix_EZVIZ_VN_Authorized_Store                                    =
pd.read_csv(r"https://raw.githubusercontent.com/NguyenKhang0062/-n-
1/refs/heads/main/e-commerce_data/fix_EZVIZ_VN_Authorized_Store.csv")
fix_Zentino  =  pd.read_csv(r"https://raw.githubusercontent.com/NguyenKhang0062/-n-
1/refs/heads/main/e-commerce_data/fix_Zentino.csv")
fix_Guardian  =  pd.read_csv(r"https://raw.githubusercontent.com/NguyenKhang0062/-n-
1/refs/heads/main/e-commerce_data/fix_Guardian.csv")
fix_HadaLabo  =  pd.read_csv(r"https://raw.githubusercontent.com/NguyenKhang0062/-n-
1/refs/heads/main/e-commerce_data/fix_HadaLabo.csv")
fix_KITY_COSMETIC =pd.read_csv(r"https://raw.githubusercontent.com/NguyenKhang0062/-
n-1/refs/heads/main/e-commerce_data/fix_KITY_COSMETIC.csv")
fix_Cosin_Store  =  pd.read_csv(r"https://raw.githubusercontent.com/NguyenKhang0062/-
n-1/refs/heads/main/e-commerce_data/fix_Cosin_Store.csv")
fix_DHC    =    pd.read_csv(r"https://raw.githubusercontent.com/NguyenKhang0062/-n-
1/refs/heads/main/e-commerce_data/fix_DHC.csv")
fix_Konni39  =  pd.read_csv(r"https://raw.githubusercontent.com/NguyenKhang0062/-n-
1/refs/heads/main/e-commerce_data/fix_Konni39.csv")
cong_nghe_df                                                     =
pd.concat([fix_Camera_Toan_Cau,fix_EZVIZ_VN_Authorized_Store,fix_Zentino],
ignore_index=True)
```

```
my_pham_df = pd.concat([fix_Guardian,fix_HadaLabo,fix_KITY_COSMETIC], ignore_index=True)

thuc_pham_df = pd.concat([fix_Cosin_Store,fix_DHC,fix_Konni39], ignore_index=True)
```

d) Mô hình hồi quy Logistics

Đây là mô hình hồi quy logistic được sử dụng để phân loại văn bản. Mô hình này sẽ dự đoán nhãn (label) cho mỗi tài liệu dựa trên các đặc trưng số học (sau khi văn bản được chuyển đổi thành các vector).

```
# Ánh xạ lại sentiment
mapping = {'POS': 1, 'NEG': 0, 'NEU': 2}
cong_nghe_df['Sentiment'] = cong_nghe_df['Sentiment'].map(mapping)
# Vector hóa review
vectorizer = TfidfVectorizer()
X_reviews = vectorizer.fit_transform(cong_nghe_df['Review'])
# Kết hợp các đặc trưng
X = pd.DataFrame(X_reviews.toarray())
y = cong_nghe_df['Sentiment']
# Chia dữ liệu thành tập huấn luyện và tập kiểm tra
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
# Xây dựng mô hình hồi quy logistic
model = LogisticRegression(max_iter=1000)
model.fit(X_train, y_train)
# Dự đoán trên tập kiểm tra
y_pred = model.predict(X_test)
# Đánh giá mô hình
print(confusion_matrix(y_test, y_pred))
print(classification_report(y_test, y_pred))
```

Nhận xét chi tiết về mô hình:

Ánh xạ lại nhãn cảm xúc (Sentiment):

- Mã hóa lại các giá trị trong cột Sentiment của DataFrame cong_nghe_df từ các nhãn chuỗi 'POS', 'NEG', và 'NEU' thành các giá trị số tương ứng: 'POS' → 1 'NEG' → 0 'NEU' → 2
- Đây là bước chuẩn bị dữ liệu cho mô hình học máy vì các thuật toán phân loại thường yêu cầu nhãn mục tiêu là các giá trị số.

Vector hóa văn bản (Review) bằng TfidfVectorizer:

- TfidfVectorizer là một công cụ phổ biến để chuyển đổi văn bản thành ma trận các đặc trưng (features). Tfidf (Term Frequency-Inverse Document Frequency) giúp trọng số các từ trong văn bản, giảm ảnh hưởng của những từ xuất hiện quá thường xuyên (stopwords).
- Hàm fit_transform được sử dụng để huấn luyện bộ vector hóa và chuyển các đánh giá (reviews) thành dạng ma trận đặc trưng.

Kết hợp các đặc trưng vào DataFrame:

- Chuyển ma trận các đặc trưng thành dạng DataFrame với mỗi cột tương ứng với một từ trong bộ từ vựng sau khi vector hóa.
- Cột Sentiment vẫn là nhãn mục tiêu (y), được sử dụng cho quá trình huấn luyện và đánh giá mô hình.

Chia dữ liệu thành tập huấn luyện và kiểm tra:

- Dữ liệu được chia thành 80% cho tập huấn luyện và 20% cho tập kiểm tra.
- Điều này giúp mô hình học từ một phần dữ liệu và kiểm tra độ chính xác trên phần dữ liệu chưa thấy.

Xây dựng mô hình hồi quy logistic:

- Mô hình hồi quy logistic được lựa chọn để phân loại ba cảm xúc (POS, NEG, NEU). Hồi quy logistic là mô hình phổ biến trong phân loại nhị phân hoặc đa lớp với các nhãn mục tiêu rời rạc.
- Tham số max_iter=1000 đảm bảo rằng thuật toán hội tụ với số vòng lặp tối đa là 1000, giúp mô hình có thể học đủ các đặc trưng trong dữ liệu.

Dự đoán và đánh giá mô hình:

- Sau khi huấn luyện, mô hình được sử dụng để dự đoán nhãn cảm xúc trên tập kiểm tra X_test.
- confusion_matrix cung cấp ma trận nhầm lẫn, hiển thị số lượng dự đoán đúng và sai cho mỗi lớp cảm xúc.
- classification_report cung cấp các chỉ số đánh giá chi tiết như Precision, Recall, F1-score, giúp đánh giá chất lượng của mô hình trong việc phân loại các lớp POS, NEG, NEU.

Nhận xét về mô hình

- Ưu điểm:
    - Tiền xử lý dữ liệu tốt:

Việc ánh xạ lại nhãn cảm xúc thành các giá trị số giúp dữ liệu phù hợp với các thuật toán học máy. Sử dụng TfidfVectorizer giúp ma trận đặc trưng chứa thông tin có trọng số, cải thiện khả năng học của mô hình.

- Lựa chọn mô hình hợp lý:

  Hồi quy logistic là một mô hình mạnh mẽ và đơn giản cho bài toán phân loại với nhiều lớp (multi-class classification) như trong trường hợp này.

- Đánh giá mô hình chi tiết:

  Việc sử dụng ma trận nhầm lẫn và báo cáo phân loại (classification report) giúp cung cấp một cái nhìn chi tiết về hiệu suất mô hình, đặc biệt là trong các vấn đề phân loại đa lớp.

- Hạn chế và cải tiến:
  - Không tối ưu hóa tham số:

    Mô hình hồi quy logistic có thể được cải thiện nếu tham số được tối ưu hóa bằng cách sử dụng kỹ thuật như GridSearchCV hoặc RandomizedSearchCV để tìm ra tham số tối ưu cho mô hình.

    Ví dụ: Tối ưu hóa các tham số như C (tham số điều chỉnh độ phức tạp của mô hình).

  - Chưa kiểm tra tính đồng nhất của dữ liệu:

    Việc kiểm tra xem dữ liệu có bị mất cân bằng giữa các lớp cảm xúc (POS, NEG, NEU) hay không là rất quan trọng. Nếu có sự mất cân bằng lớp, mô hình có thể gặp khó khăn trong việc phân loại chính xác các lớp ít dữ liệu.

  - Cải tiến: Sử dụng các kỹ thuật như SMOTE (Synthetic Minority Over-sampling Technique) để cân bằng dữ liệu.

  - Đánh giá mô hình chưa đủ sâu:

    Mặc dù classification_report cung cấp nhiều thông tin, nhưng cần kiểm tra độ chính xác của mô hình trong từng lớp cảm xúc riêng biệt để biết mô hình có phân loại đúng các lớp như NEG hay POS không.

  - Tối ưu hóa hiệu suất:

    Sử dụng TfidfVectorizer với các tham số tùy chỉnh như ngram_range (sử dụng n-grams thay vì chỉ các từ đơn) có thể cải thiện chất lượng mô hình, đặc biệt trong các văn bản dài hoặc khi cần nhận diện các mô hình ngữ nghĩa phức tạp.

Kết luận:

Đoạn mã trên sử dụng hồi quy logistic để phân loại cảm xúc với dữ liệu review. Mặc dù phương pháp tiền xử lý và lựa chọn mô hình là hợp lý, vẫn có thể cải thiện về việc tối ưu hóa tham số, kiểm tra tính đồng nhất dữ liệu và đánh giá sâu hơn.

Tương tự, ta có mô hình phân tích trong sản phẩm mỹ phẩm:

```
# Ánh xạ lại sentiment

mapping = {'POS': 1, 'NEG': 0, 'NEU': 2}

my_pham_df['Sentiment'] = my_pham_df['Sentiment'].map(mapping)

# Vector hóa review

vectorizer = TfidfVectorizer()

X_reviews = vectorizer.fit_transform(my_pham_df['Review'])

# Kết hợp các đặc trưng

X = pd.DataFrame(X_reviews.toarray())

y = my_pham_df['Sentiment']

# Chia dữ liệu thành tập huấn luyện và tập kiểm tra

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)

# Xây dựng mô hình hồi quy logistic

model = LogisticRegression(max_iter=1000)

model.fit(X_train, y_train)

# Dự đoán trên tập kiểm tra

y_pred = model.predict(X_test)

# Đánh giá mô hình

print(confusion_matrix(y_test, y_pred))

print(classification_report(y_test, y_pred))
```

Nhận xét chi tiết mô hình

Ánh xạ lại sentiment:

Giải thích:

- Mục tiêu ở đây là chuyển các nhãn cảm xúc từ dạng chuỗi thành dạng số để có thể áp dụng vào mô hình học máy.
- Các nhãn 'POS', 'NEG', và 'NEU' (tương ứng với cảm xúc tích cực, tiêu cực và trung lập) được ánh xạ thành các giá trị số:

  'POS' → 1 'NEG' → 0 'NEU' → 2
- Đây là bước chuẩn bị dữ liệu quan trọng vì hầu hết các thuật toán phân loại đều yêu cầu nhãn đầu ra là các giá trị số.

Vector hóa review:

Giải thích:

- TfidfVectorizer được sử dụng để chuyển đổi văn bản (các review) thành các đặc trưng (features). Tfidf là viết tắt của "Term Frequency-Inverse Document Frequency", giúp tính toán mức độ quan trọng của từ trong văn bản, giảm thiểu ảnh hưởng của các từ phổ biến.
- Hàm fit_transform không chỉ học từ dữ liệu mà còn áp dụng quá trình vector hóa để chuyển đổi các review thành ma trận các đặc trưng.

Kết hợp các đặc trưng:

Giải thích:

- X_reviews.toarray() chuyển đổi ma trận sparse (được sinh ra từ TfidfVectorizer) thành dạng mảng (array), rồi chuyển thành một DataFrame với các đặc trưng văn bản.
- y là nhãn cảm xúc đã được ánh xạ, là cột Sentiment trong DataFrame my_pham_df.

Chia dữ liệu thành tập huấn luyện và kiểm tra:

Giải thích:

- Dữ liệu được chia thành 80% cho tập huấn luyện (X_train, y_train) và 20% cho tập kiểm tra (X_test, y_test). Việc chia này giúp mô hình có thể học từ một phần dữ liệu và đánh giá trên phần dữ liệu chưa thấy.
- random_state=42 đảm bảo rằng việc chia dữ liệu là ngẫu nhiên nhưng sẽ tạo ra cùng một kết quả mỗi lần chạy, giúp tái sản xuất kết quả.

Xây dựng mô hình hồi quy logistic:

Giải thích:

- Logistic Regression là một thuật toán học máy đơn giản và mạnh mẽ, thường được sử dụng cho bài toán phân loại nhị phân hoặc đa lớp. Trong trường hợp này, logistic regression sẽ phân loại cảm xúc vào ba lớp: POS, NEG, và NEU.
- max_iter=1000 là tham số cho phép mô hình thực hiện tối đa 1000 vòng lặp để tối ưu hóa mô hình, giúp mô hình hội tụ nếu dữ liệu phức tạp.

Dự đoán và đánh giá mô hình:

Giải thích:

- Sau khi huấn luyện, mô hình sẽ dự đoán nhãn cảm xúc cho các mẫu trong tập kiểm tra (X_test) và lưu lại kết quả trong y_pred.
- confusion_matrix cung cấp ma trận nhầm lẫn, hiển thị số lượng dự đoán đúng (True Positives, True Negatives) và sai (False Positives, False Negatives) cho mỗi lớp.
- classification_report cung cấp các chỉ số quan trọng như:

- Precision: Tỷ lệ các dự đoán đúng trên tổng số dự đoán của mỗi lớp.
- Recall: Tỷ lệ các dự đoán đúng trên tổng số thực tế của mỗi lớp.
- F1-score: Trung bình hài hòa của Precision và Recall. Accuracy: Tỷ lệ dự đoán đúng tổng thể. Nhận xét về mô hình

Ưu điểm:

- Tiền xử lý dữ liệu tốt:

  Ánh xạ nhãn cảm xúc và sử dụng TfidfVectorizer là những bước quan trọng giúp mô hình có thể học và phân loại các review một cách hiệu quả. Tfidf giúp giảm trọng số của các từ phổ biến không có giá trị thông tin cao, chỉ tập trung vào các từ có ý nghĩa hơn.

- Sử dụng mô hình hồi quy logistic:

  Hồi quy logistic là một mô hình đơn giản nhưng mạnh mẽ trong các bài toán phân loại. Mặc dù có thể không phải là lựa chọn tối ưu cho tất cả các bài toán, nhưng với dữ liệu nhỏ và đặc trưng rõ ràng, hồi quy logistic vẫn có thể đạt được kết quả khá tốt.

- Đánh giá mô hình chi tiết:

  Việc sử dụng confusion_matrix và classification_report giúp bạn hiểu rõ hơn về hiệu suất của mô hình trên các lớp riêng biệt, đồng thời cho phép nhận diện các vấn đề về phân loại sai hoặc mất cân bằng lớp.

Hạn chế và cải tiến:

- Dữ liệu mất cân bằng:

  Nếu dữ liệu cảm xúc trong các lớp POS, NEG, và NEU không đồng đều (một lớp có số lượng mẫu nhiều hơn hẳn so với các lớp khác), mô hình có thể bị thiên lệch, dẫn đến việc phân loại sai cho các lớp ít gặp. Để khắc phục, có thể sử dụng kỹ thuật SMOTE hoặc undersampling/oversampling.

- Cải thiện với các thuật toán khác:

  Mặc dù hồi quy logistic là một lựa chọn tốt, nhưng trong các bài toán phân loại cảm xúc phức tạp hơn, bạn có thể thử các mô hình mạnh mẽ hơn như Random Forest, XGBoost, hoặc SVM (Support Vector Machine).

  Các mô hình này có thể xử lý tốt hơn khi dữ liệu có sự phức tạp cao hoặc tương tác giữa các đặc trưng.

- Tối ưu hóa tham số mô hình:

  Mặc dù max_iter=1000 là một tham số hợp lý, việc tối ưu hóa các tham số khác của mô hình như C (tham số điều chỉnh độ phức tạp của mô hình) sẽ giúp cải thiện hiệu suất mô hình.

- Cải thiện đặc trưng văn bản:

  Mô hình hiện tại chỉ sử dụng TfidfVectorizer. Có thể thử sử dụng các đặc trưng mạnh mẽ hơn như word embeddings (ví dụ: Word2Vec hoặc GloVe) hoặc BERT nếu dữ liệu và tài nguyên tính toán cho phép.

Kết luận:

Mô hình phân loại cảm xúc sử dụng hồi quy logistic với dữ liệu review mỹ phẩm hoạt động khá tốt, nhưng có thể cải thiện hơn nữa qua việc xử lý dữ liệu mất cân bằng, tối ưu hóa tham số mô hình, và thử các mô hình phân loại khác để nâng cao độ chính xác và hiệu suất.

Tiếp theo, áp dụng mô hình trong sản phẩm thực phẩm chức năng:

```
# Ánh xạ lại sentiment

mapping = {'POS': 1, 'NEG': 0, 'NEU': 2}

thuc_pham_df['Sentiment'] = thuc_pham_df['Sentiment'].map(mapping)

print(thuc_pham_df['Sentiment'].isnull().sum())  # Kiểm tra NaN sau khi map

print(thuc_pham_df['Sentiment'].unique())       # Xem các giá trị duy nhất

thuc_pham_df = thuc_pham_df.dropna(subset=['Sentiment'])

# Vector hóa review

vectorizer = TfidfVectorizer()

X_reviews = vectorizer.fit_transform(thuc_pham_df['Review'])

# Kết hợp các đặc trưng

X = pd.DataFrame(X_reviews.toarray())

y = thuc_pham_df['Sentiment']

# Chia dữ liệu thành tập huấn luyện và tập kiểm tra

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,

random_state=42)

# Xây dựng mô hình hồi quy logistic

model = LogisticRegression(max_iter=1000)

model.fit(X_train, y_train)

# Dự đoán trên tập kiểm tra

y_pred = model.predict(X_test)

# Đánh giá mô hình

print(confusion_matrix(y_test, y_pred))

print(classification_report(y_test, y_pred))
```

4.2.3. Mối liên hệ và phân tích xu hướng giữa các biến trong lĩnh vực ngành học

a) Tính toán các chỉ số cơ bản

Khởi tạo hàm để tính các chỉ số như tổng chỉ tiêu theo năm và tỷ lệ tăng trưởng 2022-2024

$$Tang\_truong\_2023 = \frac{(Tong\_2023 - Tong\_2022)}{Tong\_2022} \times 100$$

$$Tang\_truong\_2024 = \frac{(Tong\_2024 - Tong\_2023)}{Tong\_2023} \times 100$$

```
# Hàm tính toán chỉ số cơ bản
calculate_metrics <- function(data) {
  # Tính tổng chỉ tiêu theo năm
  yearly_totals <- data %>%
    summarise(
      Tong_2022 = sum(chi_tieu_2022, na.rm = TRUE),
      Tong_2023 = sum(chi_tieu_2023, na.rm = TRUE),
      Tong_2024 = sum(chi_tieu_2024, na.rm = TRUE),
      .groups = 'drop'
    )
  # Tính toán các chỉ số tăng trưởng
  metrics <- yearly_totals %>%
    mutate(
      Tang_truong_2023 = (Tong_2023 - Tong_2022)/Tong_2022 * 100,
      Tang_truong_2024 = (Tong_2024 - Tong_2023)/Tong_2023 * 100,
    ) %>%
    mutate(Truong = unique(data$truong)[1]) %>%
    select(Truong, everything())

  return(metrics)
}
# Hàm thống kê cho từng miền
calculate_region_stats <- function(data, region_name) {
  data %>%
    summarise(
```

```
      So_truong = n_distinct(truong),
      So_nganh = n_distinct(ma_xet_tuyen),
      So_linh_vuc = n_distinct(linh_vuc),
      Tong_2022 = sum(chi_tieu_2022, na.rm = TRUE),
      Tong_2023 = sum(chi_tieu_2023, na.rm = TRUE),
      Tong_2024 = sum(chi_tieu_2024, na.rm = TRUE),
      TB_2022 = mean(chi_tieu_2022, na.rm = TRUE),
      TB_2023 = mean(chi_tieu_2023, na.rm = TRUE),
      TB_2024 = mean(chi_tieu_2024, na.rm = TRUE),
      .groups = 'drop'
    ) %>%
    mutate(
      Tang_truong_2023 = (Tong_2023 - Tong_2022)/Tong_2022 * 100,
      Tang_truong_2024 = (Tong_2024 - Tong_2023)/Tong_2023 * 100,
      Mien = region_name
    ) %>%
    select(Mien, everything())
}
# Hàm phân tích xu hướng theo lĩnh vực
analyze_trends_by_field <- function(data, region_name = NULL) {
  # Chuyển sang long format
  long_data <- data %>%
    pivot_longer(
      cols = starts_with("chi_tieu_"),
      names_to = "nam",
      values_to = "chi_tieu"
    ) %>%
    mutate(
      nam = str_extract(nam, "\\d{4}$") %>% as.numeric(),
      linh_vuc = ifelse(is.na(linh_vuc), "Không xác định", linh_vuc)
    )
  # Tính toán chỉ số theo lĩnh vực
  field_trends <- long_data %>%
```

```
  group_by(truong, linh_vuc, nam) %>%
  summarise(
    tong_chi_tieu = sum(chi_tieu, na.rm = TRUE),
    so_nganh = n_distinct(ma_xet_tuyen),
    .groups = "drop"
  ) %>%
  pivot_wider(
    names_from = nam,
    values_from = c(tong_chi_tieu),
    names_prefix = "chi_tieu_"
  )
# Tính toán tăng trưởng
field_trends <- field_trends %>%
  mutate(
    tang_truong_2023 = if_else(
      is.finite((chi_tieu_2023 - chi_tieu_2022) / chi_tieu_2022),
      (chi_tieu_2023 - chi_tieu_2022) / chi_tieu_2022 * 100,
      NA_real_
    ),
    tang_truong_2024 = if_else(
      is.finite((chi_tieu_2024 - chi_tieu_2023) / chi_tieu_2023),
      (chi_tieu_2024 - chi_tieu_2023) / chi_tieu_2023 * 100,
      NA_real_
    )
  )
# Thêm thông tin miền
if (!is.null(region_name)) {
  field_trends <- field_trends %>%
    mutate(mien = region_name)
}
# Tính toán chỉ số bổ sung
field_trends <- field_trends %>%
  mutate(
```

```
      tong_chi_tieu = rowSums(select(., starts_with("chi_tieu_")), na.rm = TRUE),
      tang_truong_tb = rowMeans(select(., starts_with("tang_truong_")), na.rm =
TRUE)
    ) %>%
    arrange(desc(tong_chi_tieu))
  return(field_trends)
}
```

Thông số theo các trường:

```
# A tibble: 8 × 6
  Truong Tong_2022 Tong_2023 Tong_2024 Tang_truong_2023 Tang_truong_2024
  <chr>      <dbl>     <dbl>     <dbl>            <dbl>            <dbl>
1 HUST        7990      8005      9150            0.188           14.3
2 NEU         7125      6200      6200          -13.0              0
3 QHT         1650      1750      1850            6.06             5.71
4 QHX         1315      1999      2300           52.0             15.1
5 UEH         6550      7050      7300            7.63             3.55
6 HCMUT       4885      5040      5000            3.17            -0.794
7 PTIT         810       910      1020           12.3             12.1
8 TCT         8420      7787      8754           -7.52            12.4
```

Nhận xét:

- Trong giai đoạn từ năm 2022 đến 2024, chỉ tiêu tuyển sinh của các trường đại học có nhiều biến động đáng chú ý. Trường QHX ghi nhận mức tăng trưởng ấn tượng, với tỷ lệ tăng mạnh đến 52% trong năm 2023 và tiếp tục tăng 15,1% trong năm 2024, cho thấy định hướng mở rộng quy mô hoặc sức hút ngày càng lớn đối với thí sinh. Tương tự, các trường như QHT, PTIT và UEH có xu hướng tăng trưởng ổn định và bền vững qua từng năm, phản ánh chiến lược phát triển đều và dài hạn.
- Ở chiều ngược lại, một số trường lại có sự điều chỉnh giảm. NEU là trường duy nhất có mức giảm mạnh nhất trong năm 2023 với -13%, sau đó giữ nguyên chỉ tiêu trong năm 2024, cho thấy sự thận trọng hoặc điều chỉnh chiến lược. TCT cũng giảm -7,52% trong năm 2023 nhưng đã có dấu hiệu phục hồi với mức tăng 12,4% trong năm sau. Trong khi đó, HCMUT tăng nhẹ 3,17% vào năm 2023 nhưng lại giảm nhẹ 0,79% vào năm 2024 – cho thấy sự dao động nhỏ và có thể đang trong giai đoạn ổn định hóa chỉ tiêu.

- Riêng trường HUST giữ tốc độ tăng trưởng chậm vào năm 2023 (+0,19%) nhưng bất ngờ bứt phá vào năm 2024 với +14,3%, cho thấy sự chuyển biến tích cực sau giai đoạn ổn định. Nhìn chung, hầu hết các trường đều có xu hướng điều chỉnh tăng chỉ tiêu tuyển sinh vào năm 2024, phản ánh nhu cầu mở rộng đào tạo và sự thích nghi với biến động nguồn tuyển trong bối cảnh thay đổi của giáo dục đại học.

Thông số theo khu vực:

```
# A tibble: 2 × 12
  Mien     So_truong So_nganh So_linh_vuc Tong_2022 Tong_2023 Tong_2024 TB_2022
TB_2023 TB_2024
  <chr>        <int>    <int>       <int>     <dbl>     <dbl>     <dbl>   <dbl>
<dbl>   <dbl>
1 Miền Bắc         4      163          10     17575     16865     18220   105.
101.     109.
2 Miền Nam         4      203          10     21170     21876     23354    95.8
99.0    106.
# i 2 more variables: Tang_truong_2023 <dbl>, Tang_truong_2024 <dbl>
```

Nhận xét:

- Dựa trên số liệu tổng hợp theo khu vực, có thể nhận thấy sự khác biệt đáng chú ý giữa Miền Bắc và Miền Nam trong chỉ tiêu tuyển sinh và mức độ tăng trưởng qua các năm từ 2022 đến 2024.
- Cụ thể, Miền Nam luôn dẫn đầu về tổng chỉ tiêu tuyển sinh, trong khi ddos Miền Bắc có tổng chỉ tiêu thấp hơn. Điều này phản ánh nhu cầu tuyển sinh lớn hơn hoặc sự phát triển mạnh mẽ của các trường đại học ở khu vực Miền Nam.
- Tuy nhiên, xét về mức trung bình chỉ tiêu mỗi ngành, Miền Bắc lại vượt trội hơn: lần lượt là 105, 101, và 109, so với Miền Nam chỉ đạt 95,8, 99,0, và 106. Điều này cho thấy, mặc dù số lượng ngành đào tạo tại Miền Nam nhiều hơn, nhưng mỗi ngành tại Miền Bắc thường có quy mô chỉ tiêu lớn hơn – có thể do các trường tập trung vào các ngành mũi nhọn hoặc có ít phân tán ngành học.
- Về tốc độ tăng trưởng, cả hai khu vực đều cho thấy sự phục hồi và tăng trưởng trở lại vào năm 2024. Miền Bắc tăng từ mức giảm nhẹ -4,0% (2023) sang +8,0% (2024), trong khi Miền Nam có mức tăng ổn định hơn: +3,2% (2023) và +7,0% (2024). Điều này cho thấy các trường ở cả hai miền đều có xu hướng mở rộng tuyển sinh, đặc biệt là sau một năm 2023 tương đối ổn định hoặc chững lại.

Tóm lại, Miền Nam chiếm ưu thế về tổng quy mô, còn Miền Bắc lại nổi bật về chỉ tiêu trung bình mỗi ngành – thể hiện hai chiến lược phát triển khác nhau giữa các trường đại học theo vùng miền.

b) Tăng trưởng theo lĩnh vực

```
# A tibble: 20 × 5
   linh_vuc           nam tong_chi_tieu tang_truong nam_tang_truong
   <chr>            <dbl>         <dbl>       <dbl> <chr>
 1 công nghệ         2023          5565       11.7  2022 - 2023
 2 giáo dục          2023           592      -15.4  2022 - 2023
 3 khoa học quản lý  2023          2915       -3.48 2022 - 2023
 4 khtn              2023          3710       -2.62 2022 - 2023
 5 khxh và nhân văn  2023          3004       22.1  2022 - 2023
 6 kt-tc             2023          8255       -1.90 2022 - 2023
 7 kỹ thuật          2023          8645       -1.31 2022 - 2023
 8 ngôn ngữ          2023           855       -4.47 2022 - 2023
 9 quản trị          2023          3620      -12.7  2022 - 2023
10 sức khỏe          2023          1580        1.28 2022 - 2023
11 công nghệ         2024          5680        2.07 2023 - 2024
12 giáo dục          2024           804       35.8  2023 - 2024
13 khoa học quản lý  2024          3335       14.4  2023 - 2024
14 khtn              2024          4215       13.6  2023 - 2024
15 khxh và nhân văn  2024          3100        3.20 2023 - 2024
16 kt-tc             2024          8630        4.54 2023 - 2024
17 kỹ thuật          2024          8855        2.43 2023 - 2024
18 ngôn ngữ          2024           950       11.1  2023 - 2024
19 quản trị          2024          3775        4.28 2023 - 2024
20 sức khỏe          2024          2230       41.1  2023 - 2024
```

Nhận xét:

- Dựa trên dữ liệu phân tích chỉ tiêu tuyển sinh theo lĩnh vực giai đoạn 2022–2024, có thể thấy rõ một số xu hướng nổi bật trong quá trình thay đổi định hướng đào tạo của các trường đại học. Trong giai đoạn 2022–2023, nhiều lĩnh vực có sự sụt giảm về chỉ tiêu, đáng chú ý nhất là nhóm ngành Giáo dục giảm tới 15.4%, Quản trị giảm 12.7%, cùng với các lĩnh vực như Khoa học quản lý, Kinh tế – Tài chính, Ngôn ngữ và Kỹ thuật cũng ghi nhận mức giảm nhẹ. Ngược lại, một số lĩnh vực

ghi nhận sự tăng trưởng đáng kể như Khoa học xã hội & Nhân văn tăng 22.1%, Công nghệ tăng 11.7%, và Sức khỏe tăng nhẹ 1.28%, phản ánh phần nào sự quan tâm của xã hội đến các ngành có xu hướng ứng dụng thực tiễn cao và nhu cầu nhân lực tăng.

- Bước sang giai đoạn 2023–2024, bức tranh tuyển sinh có dấu hiệu phục hồi mạnh mẽ. Đặc biệt, nhóm ngành Sức khỏe tăng vọt tới 41.1%, Giáo dục tăng trở lại 35.8%, và các ngành như Khoa học tự nhiên, Khoa học quản lý hay Ngôn ngữ đều có mức tăng hai chữ số. Những ngành như Công nghệ, Kỹ thuật, Kinh tế – Tài chính, và Quản trị cũng ghi nhận mức tăng trưởng nhẹ, thể hiện xu hướng ổn định trở lại. Tuy nhiên, lĩnh vực Khoa học xã hội & Nhân văn dù tăng mạnh ở giai đoạn trước, chỉ còn tăng nhẹ 3.2% trong năm 2024, cho thấy tốc độ phát triển có dấu hiệu chững lại.

Tổng thể, dữ liệu cho thấy sự chuyển dịch trong định hướng tuyển sinh giữa các lĩnh vực, phản ánh cả nhu cầu thị trường lao động và chính sách phát triển ngành nghề của xã hội hiện nay.

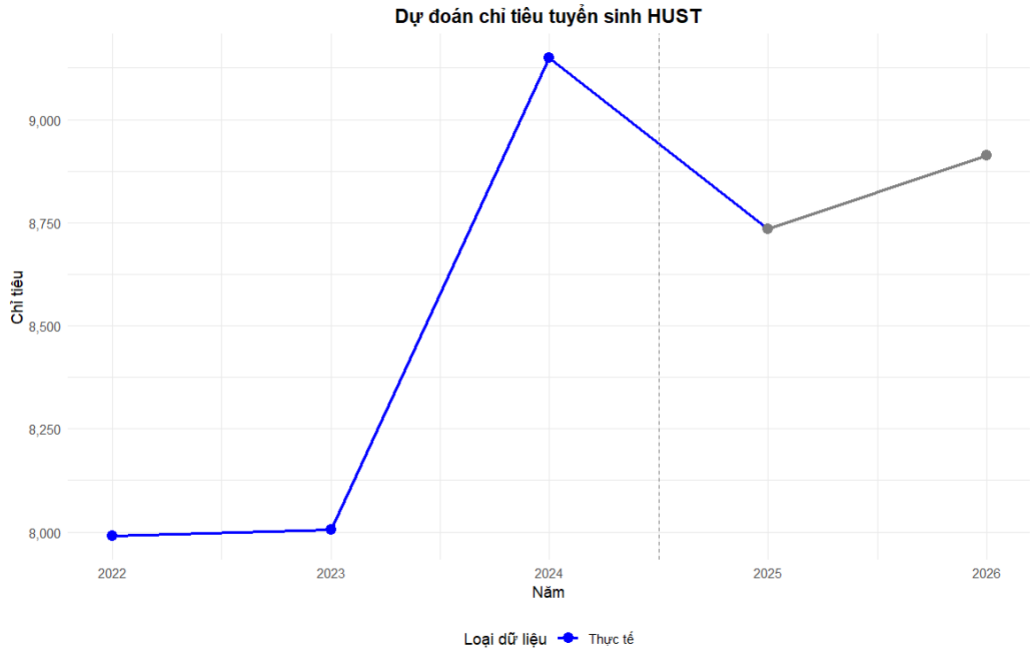
c) Mô hình hồi quy tuyến tính dựa đoán cho năm 2025-2026

Sử dụng mô hình hồi quy tuyến tính nhằm dự đoán chỉ tiêu của các trường trong các năm tiếp theo (2025 và 2026) dựa trên dữ liệu các năm trước (2022-2024). Dữ liệu đầu vào chủ yếu là các chỉ tiêu tổng cộng (Tong_) theo từng năm cho từng trường, và mô hình sẽ sử dụng năm (Nam) và trường học (Truong) để dự đoán chỉ tiêu cho các năm tiếp theo.

```
# Hàm dự đoán chỉ tiêu
predict_quotas <- function() {
  # Chuẩn bị dữ liệu
  prediction_data <- all_schools_metrics %>%
    select(Truong, starts_with("Tong_")) %>%
    pivot_longer(
      cols = starts_with("Tong_"),
      names_to = "Nam",
      values_to = "Chi_tieu"
    ) %>%
    mutate(
      Nam = as.numeric(str_extract(Nam, "\\d+")),
      Truong = as.factor(Truong)
    ) %>%
```

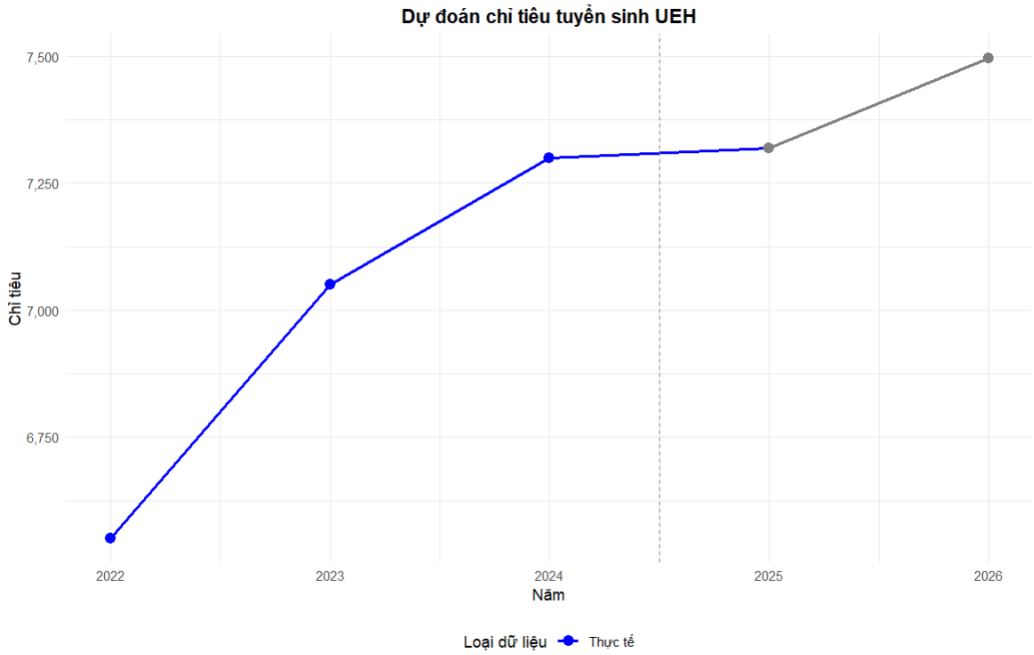
```
    arrange(Truong, Nam)
  # Xây dựng mô hình
  lm_model <- lm(Chi_tieu ~ Nam + Truong, data = prediction_data)
  # Dự đoán cho 2025-2026
  future_data <- expand.grid(
    Truong = levels(prediction_data$Truong),
    Nam = 2025:2026
  )
  future_data$Chi_tieu_dudoan <- predict(lm_model, newdata = future_data)
  # Kết hợp kết quả
  predictions <- future_data %>%
    mutate(Chi_tieu_dudoan = round(Chi_tieu_dudoan)) %>%
    pivot_wider(
      names_from = Nam,
      values_from = Chi_tieu_dudoan,
      names_prefix = "Du_doan_"
    )
  final_predictions <- all_schools_metrics %>%
    select(Truong, Tong_2022, Tong_2023, Tong_2024) %>%
    left_join(predictions, by = "Truong")
  return(final_predictions)
}
# Thực hiện hàm dự đoán
predictions <- predict_quotas()
# In ra kết quả dự đoán
print(predictions)
```

Dự đoán chỉ tiêu đầu vào của HUST trong năm 2025 và 2026:



Hình 4. 1. Biểu đồ dự đoán chỉ tiêu tuyển sinh HUST

Dự đoán chỉ tiêu đầu vào của UEH trong năm 2025 và 2026:



Hình 4. 2. Biểu đồ dự đoán chỉ tiêu tuyển sinh UEH

Kết quả dự đoán chỉ tiêu đầu vào của các năm 2025 – 2026:

```
# A tibble: 8 × 6
  Truong Tong_2022 Tong_2023 Tong_2024 Du_doan_2025 Du_doan_2026
  <chr>      <dbl>     <dbl>     <dbl>        <dbl>        <dbl>
1 HUST        7990      8005      9150         8735         8912
2 NEU         7125      6200      6200         6862         7039
```

3 QHT	1650	1750	1850	2104	2280
4 QHX	1315	1999	2300	2225	2402
5 UEH	6550	7050	7300	7320	7497
6 HCMUT	4885	5040	5000	5329	5505
7 PTIT	810	910	1020	1267	1444
8 TCT	8420	7787	8754	8674	8851

Nhận xét:

- Nhìn chung, kết quả dự đoán cho thấy hầu hết các trường có xu hướng tăng trưởng nhẹ hoặc ổn định trong chỉ tiêu tuyển sinh, đặc biệt là trong các năm 2025 và 2026.
- Tuy nhiên, một số trường như NEU và HUST có sự giảm nhẹ trong chỉ tiêu tuyển sinh vào năm 2025, điều này có thể phản ánh các yếu tố tác động từ chính sách tuyển sinh, sự thay đổi nhu cầu ngành học hoặc các yếu tố bên ngoài.
- Trong khi đó, các trường nhỏ hơn như PTIT và QHT lại có sự tăng trưởng mạnh mẽ, điều này có thể xuất phát từ sự phát triển nhanh chóng của các ngành học mới và sự thay đổi nhu cầu của thị trường lao động.

Kết luận:

Các trường có thể tận dụng các xu hướng này để điều chỉnh chiến lược tuyển sinh, tập trung vào những ngành học đang thu hút sự quan tâm của thí sinh, đồng thời chú ý đến việc mở rộng quy mô tuyển sinh ở những lĩnh vực có tiềm năng phát triển mạnh mẽ trong tương lai.

KẾT LUẬN VÀ KIẾN NGHỊ

Kết luận:

Nghiên cứu này tập trung vào việc trực quan hóa, đánh giá và phân tích xu hướng trong các lĩnh vực quan trọng như sản phẩm tiêu dùng, ngành nghề đào tạo ở trường đại học và các chủ đề phim được yêu thích, dựa trên dữ liệu thu thập từ các trang mạng xã hội. Kết quả từ nghiên cứu không chỉ mang lại cái nhìn toàn diện về sở thích, hành vi và nhu cầu của người dùng mà còn cung cấp các cơ sở dữ liệu quan trọng để hỗ trợ ra quyết định trong từng lĩnh vực.

Trong lĩnh vực sản phẩm tiêu dùng, việc phân tích dữ liệu từ các đánh giá, phản hồi của người dùng trên mạng xã hội giúp các doanh nghiệp nắm bắt được xu hướng tiêu dùng, các sản phẩm được ưa chuộng và đánh giá mức độ hài lòng của khách hàng. Thông qua đó, các doanh nghiệp có thể tối ưu hóa danh mục sản phẩm, cải tiến chất lượng hàng hóa và phát triển chiến lược giá cả hợp lý. Đồng thời, việc phân tích dữ liệu còn hỗ trợ doanh nghiệp xác định những phân khúc thị trường tiềm năng và xây dựng các chiến dịch quảng bá sản phẩm hiệu quả hơn, đáp ứng đúng nhu cầu và mong đợi của khách hàng, từ đó gia tăng lợi thế cạnh tranh trên thị trường.

Trong lĩnh vực giáo dục và đào tạo, nghiên cứu cung cấp thông tin chi tiết về xu hướng lựa chọn ngành học của học sinh, sinh viên và nhu cầu tuyển dụng của thị trường lao động. Các trường đại học có thể sử dụng các dữ liệu này để điều chỉnh, thiết kế chương trình đào tạo phù hợp với yêu cầu thực tiễn và xu hướng nghề nghiệp hiện tại. Việc xây dựng chương trình học có tính ứng dụng cao không chỉ giúp nâng cao chất lượng đào tạo mà còn tạo điều kiện thuận lợi để sinh viên tiếp cận các cơ hội việc làm trong tương lai. Bên cạnh đó, các trường cũng có thể triển khai các chiến dịch truyền thông và quảng bá hiệu quả hơn, nhấn mạnh các ngành học có tiềm năng phát triển và phù hợp với nhu cầu của xã hội.

Đối với lĩnh vực phim ảnh, dữ liệu được phân tích từ các đánh giá, nhận xét và phản hồi của khán giả trên các nền tảng mạng xã hội mang lại nhiều lợi ích quan trọng. Các nhà sản xuất và nhà phân phối phim có thể xác định được các chủ đề phim đang được ưa thích, cũng như các yếu tố ảnh hưởng đến sự thành công của một bộ phim như nội dung, thể loại và diễn viên. Thông qua đó, họ có thể xây dựng các chiến lược sản xuất, đầu tư và quảng bá phim một cách hiệu quả hơn. Ngoài ra, việc phân tích dữ liệu còn giúp các nhà sản xuất dự đoán được xu hướng thị hiếu của khán giả trong tương lai, từ đó tạo ra các sản phẩm điện ảnh chất lượng, thu hút người xem và gia tăng doanh thu.

Tổng thể, nghiên cứu này không chỉ đóng vai trò quan trọng trong việc phân tích và dự báo xu hướng của người dùng mà còn cung cấp một cơ sở dữ liệu đáng tin cậy để các bên liên quan trong từng lĩnh vực có thể đưa ra quyết định chiến lược chính xác. Với các doanh nghiệp sản xuất, trường đại học và nhà sản xuất phim, dữ liệu này là công cụ

hữu ích để tối ưu hóa sản phẩm, cải thiện dịch vụ và đáp ứng nhu cầu của thị trường một cách tốt nhất. Việc kết hợp giữa khoa học dữ liệu và thực tiễn kinh doanh sẽ góp phần thúc đẩy sự phát triển bền vững và hiệu quả trong các lĩnh vực nghiên cứu này.

Kiến nghị:

Trong tương lai, cần mở rộng phạm vi thu thập dữ liệu để bao quát hơn các nhóm đối tượng, nền tảng mạng xã hội, và khu vực địa lý khác nhau. Điều này đặc biệt quan trọng đối với các lĩnh vực như sản phẩm tiêu dùng và phim ảnh, nơi hành vi của người dùng có thể khác biệt rõ rệt theo vùng miền và văn hóa. Đồng thời, việc áp dụng các mô hình học máy tiên tiến như học sâu (deep learning), mạng nơ-ron (neural networks), hoặc các phương pháp ensemble learning sẽ giúp cải thiện khả năng dự đoán và phân tích xu hướng.

Hơn nữa, cần thiết lập quy trình cập nhật dữ liệu định kỳ để theo dõi và phản ánh chính xác các thay đổi trong sở thích của người dùng. Ví dụ, trong lĩnh vực đào tạo, các ngành học mới nổi liên quan đến công nghệ AI hoặc phát triển bền vững cần được đưa vào phân tích để xác định tiềm năng tăng trưởng. Tương tự, trong lĩnh vực phim ảnh, việc theo dõi các thể loại hoặc chủ đề mới đang được khán giả quan tâm, chẳng hạn như phim tài liệu hoặc phim hoạt hình độc lập, sẽ mang lại cái nhìn toàn diện hơn.

Cuối cùng, việc kết hợp chuyên môn liên ngành, bao gồm marketing, giáo dục, và phân tích dữ liệu, sẽ tạo điều kiện để đưa ra các chiến lược hiệu quả. Chẳng hạn, các nhà sản xuất phim có thể sử dụng các kết quả phân tích để tối ưu hóa nội dung và chiến lược quảng bá, trong khi các trường đại học có thể định hướng chương trình đào tạo phù hợp với nhu cầu lao động và xu hướng xã hội. Nghiên cứu mở rộng này sẽ giúp cung cấp thông tin giá trị và hỗ trợ các quyết định chiến lược trong các lĩnh vực được nghiên cứu.

TÀI LIỆU THAM KHẢO

[1]. Trần Chí Lê, Nguyễn Thị Hạnh Lê (2024), Tài liệu học tập Đồ án 1: Trực quan hóa dữ liệu bằng R, Trường Đại học Kinh tế - Kỹ thuật Công nghiệp.

[2]. Trần Chí Lê (2022), Tài liệu học tập Lập trình R, Trường Đại học Kinh tế - Kỹ thuật Công nghiệp.

[3]. Cao Diệp Thắng, Đỗ Tuấn Hạnh (2023), Tài liệu học tập Lập trình Python nâng cao, Trường Đại học Kinh tế - Kỹ thuật Công nghiệp

[3]. Trần Thị Kim Thanh, Trần Chí Lê (2023), Tài liệu học tập Thống kê Toán học cho ngành Khoa học dữ liệu, Trường Đại học Kinh tế - Kỹ thuật Công nghiệp.

[4]. Link nguồn dữ liệu: https://github.com/NguyenKhang0062/-n-1

[5]. Trần Thị Hoàng Yến, Bùi Văn Tân, Chu Bình Minh (2024), Tài liệu học tập Nhập môn Trí tuệ Nhân tạo, Trường Đại học Kinh tế - Kỹ thuật Công nghiệp.

[6]. Hỗ trợ của Chat GPT 3.5. Link: https://chatgpt.com/

CÁC PHỤ LỤC

1. Phiếu đăng ký và thuyết minh đề cương đề tài (*07 trang***)**

2. Quyết định giao nhiệm vụ thực hiện đề tài (*02 trang***)**

3. Báo cáo tình hình thực hiện đề tài NCKHSV (*03 trang***)**

4. Biên bản kiểm tra tiến độ thực hiện nhiệm vụ NCKHSV (*02 trang***)**

5. Bài viết tóm tắt kết quả của đề tài để đăng kỷ yếu (*01 trang***)**