PHÂN LOẠI SÁCH TIKI

VÕ THỊ KIM LY

NHÓM 4

MENTOR: ĐẶNG MINH CHƯƠNG

MỤC LỤC

- GIỚI THIỆU
- THÔNG TIN DATA
- XỬ LÝ DỮ LIỆU
- TÌM HIỂU DỮ LIỆU
  - Ý tưởng phân tích
  - Khám phá dữ liệu
- MÔ HÌNH PHÂN TÍCH CLUSTERING
- ĐÁNH GIÁ – KẾT LUẬN

1. GIỚI THIỆU

- Đây là tập dữ liệu tổng hợp các thông tin về sách bán được trên sàn thương mại điện tử Tiki
- Bộ dữ liệu này bao gồm các bảng thông tin về sách, khoảng 50 bình luận cho mỗi sách, mã số ID
- Trong bài phân tích này em sẽ tiến hành các ý như sau:

- Tìm hiểu giá sách và xu hướng mua hàng bị tác động bởi những yếu tố nào.
- Mức độ thu hút số lượng người mua sách dựa vào những yếu tố nào.
- Top những quyển sách, tác giả và thể loại bán chạy nhất.
- Xây dựng mô hình phân loại các nhóm sách tiềm năng bán chạy, bình thường, không tiềm năng bán chạy.

2. THÔNG TIN DATA

*** Book sheet**

- product_id: id của sản phẩm trong cơ sở dữ liệu Tiki (duy nhất)
- title: Tên tiêu đề của cuốn sách, có thể chứa thời gian xuất bản lại
- Authors: tác giả
- original_price: giá lần đầu
- current_price: giá hiện tại nếu có giảm giá
- Quantity: tổng số sách đã bán mọi thời đại
- Category: thể loại sách
- n_review: số lượng đánh giá
- avg_rating: xếp hạng trung bình (tối đa 5.0)
- Pages: tổng số trang của mỗi cuốn sách

*** Comment sheet**

- comment_id: Mỗi bình luận có id cá nhân
- Title: Keywork của bình luận
- thank_count: Số like của người khác
- customer_id: Mỗi khách hàng có id riêng
- Rating: Đánh giá trung bình của bình luận
- Content: Nội dung

3. XỬ LÝ DỮ LIỆU

- Kiểm tra dữ liệu
- Xử lý dữ liệu title, thể loại bị trùng
- Xử lý authors, quantity, pages, manufacturer bị null
- Tính thêm cột discount

```
# [3] Cleaning Data
Book.isnull().sum()
```

```
product_id          0
title               0
authors           143
original_price      0
current_price       0
quantity           45
category            0
n_review            0
avg_rating          0
pages             250
manufacturer      273
cover_link          0
dtype: int64
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1796 entries, 0 to 1795
Data columns (total 12 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   product_id      1796 non-null   int64
 1   title           1796 non-null   object
 2   authors         1653 non-null   object
 3   original_price  1796 non-null   int64
 4   current_price   1796 non-null   int64
 5   quantity        1751 non-null   float64
 6   category        1796 non-null   object
 7   n_review        1796 non-null   int64
 8   avg_rating      1796 non-null   float64
 9   pages           1546 non-null   object
 10  manufacturer    1523 non-null   object
 11  cover_link      1796 non-null   object
dtypes: float64(2), int64(4), object(6)
memory usage: 168.5+ KB
```

XỬ LÝ DỮ LIỆU TITLE BỊ TRÙNG

```
# Tính số tillte trùng
from collections import Counter
counts = Counter(Book['title'])
duplicates = [o for o, c in counts.items() if c > 1]
counts
```

```
# Xóa những tittle bị trùng
Book = Book.drop_duplicates(subset=['title'])
```

```
# Check lại đã xóa giá trị trùng title chưa => đã xóa
counts = Counter(Book['title'])
duplicates = [o for o, c in counts.items() if c > 1]
counts
```

XỬ LÝ DỮ LIỆU XỬ LÝ CỘT TÁC GIẢ BỊ NULL

```
# Xử lý cột tác giả
Book.authors.value_counts()
```

```
Nguyễn Nhật Ánh    24
Higashino Keigo    20
.                  18
Thích Nhất Hạnh    16
Haruki Murakami    15
                   ..
Urako Kanamori      1
Cố Viên             1
Robert Winston      1
Yongchul Kwon       1
John C. Maxwell     1
Name: authors, Length: 1083, dtype: int64
```

```
# Thay thế giá trị "." thành "unknown"
Book.loc[Book.authors == '.' , 'authors'] = "Unknown"
Book.authors = Book.authors.fillna("Unknown") # Điền những chỗ chưa điền tác giả thành unknown
Book.authors.value_counts()
```

XỬ LÝ DỮ LIỆU XỬ LÝ CỘT SL BỊ NULL

```
# Số lượng điền giá trị trung bình vào các cột null
Book.quantity = Book.quantity.fillna(np.mean(Book.quantity))
```

XỬ LÝ DỮ LIỆU THỂ LOẠI

```
# Check thể loại đã ổn chưa
Book.category.value_counts() # có mấy chỗ tên sách chứ không phải thể loại
```

```
Sách tư duy - Kỹ năng sống                292
Tiểu Thuyết                               133
Truyện ngắn - Tản văn - Tạp Văn           109
Sách nghệ thuật sống đẹp                   58
Sách kỹ năng làm việc                      55
                                          ...
Shaman King - Tập 19                        1
Blue Period - Tập 07                        1
Komi - Nữ Thần Sợ Giao Tiếp - Tập 14        1
Bên Rặng Tuyết Sơn (Tái Bản)                1
Kaguya-Sama: Cuộc Chiến Tỏ Tình - Tập 3     1
Name: category, Length: 354, dtype: int64
```

```
counts1 = Counter(Book['category'])
duplicates = [o for o, c in counts1.items() if c > 1]
counts1 #Loại những giá trị khác 1 => Lọc list thể loại
```

```
# Lọc thể loại
keeping_values = list(pd.DataFrame(Book.category.value_counts().loc[lambda x : x>1]).T.columns)
keeping_values
```

```
def handle_category(category):
    if category not in keeping_values:
        return "Others"
    return category
Book.category = Book.category.apply(lambda category: handle_category(category))
```

XỬ LÝ DỮ LIỆU PAGES BỊ NULL

```python
# trang là kiểu số, với check dữ liệu có chữ cuốn, xử lý lại
# Tính giá trị trung bình
def handle_pages(lst_pages):
    lst_int = []
    for page in lst_pages:
        try:
            lst_int.append(int(page))
        except:
            continue

    return np.mean(lst_int)

mean_pages = handle_pages(list(Book.pages))

Book.pages = Book.pages.fillna(str(mean_pages))

Book = Book.query('pages != "Cuốn"')

Book.pages = Book.pages.apply(lambda page: round(float(page)))
Book.pages = Book.pages.astype("int64")
```

XỬ LÝ DỮ LIỆU NHÀ XUẤT BẢN BỊ NULL

```python
def handle_manufactuner(manufactuner):
    if manufactuner == "hong duc":
        return "Nhà Xuất Bản Hồng Đức"
    elif manufactuner == "NXB Dân Trí":
        return "Nhà Xuất Bản Dân Trí"
    elif manufactuner == "ĐHQG Hà Nội":
        return "Nhà Xuất Bản Đại Học Quốc Gia Hà Nội"
    else:
        return manufactuner

Book.manufacturer = Book.manufacturer.apply(lambda manufacturer: handle_manufactuner(manufacturer))

Book.manufacturer = Book.manufacturer.fillna("Unknown")
```

DATA SAU KHI XỬ LÝ

```
Book.info()
```

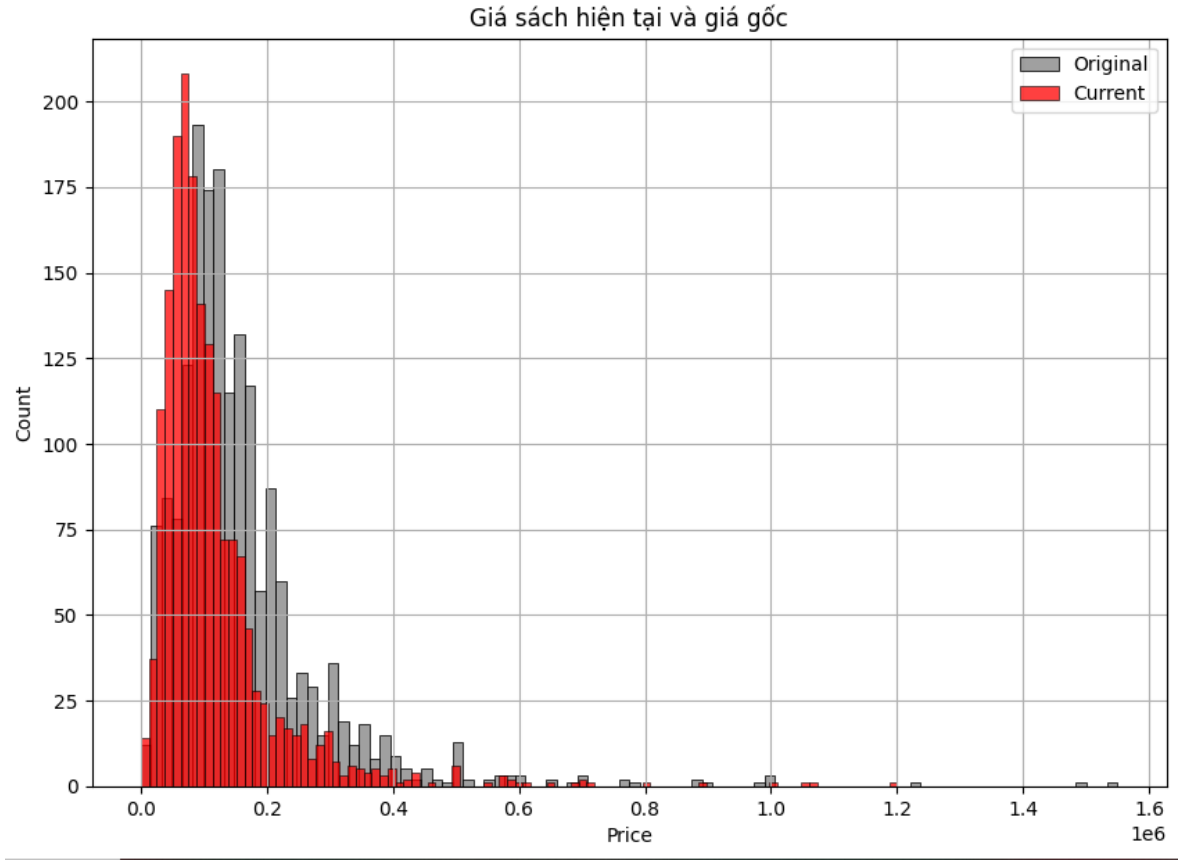
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1768 entries, 0 to 1795
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   product_id      1768 non-null   int64  
 1   title           1768 non-null   object 
 2   authors         1768 non-null   object 
 3   original_price  1768 non-null   int64  
 4   current_price   1768 non-null   int64  
 5   quantity        1768 non-null   float64
 6   category        1768 non-null   object 
 7   n_review        1768 non-null   int64  
 8   avg_rating      1768 non-null   float64
 9   pages           1768 non-null   int64  
 10  manufacturer    1768 non-null   object 
 11  cover_link      1768 non-null   object 
 12  discount        1768 non-null   float64
dtypes: float64(3), int64(5), object(5)
memory usage: 193.4+ KB
```

4. TÌM HIỂU DỮ LIỆU

- Ý tưởng phân tích

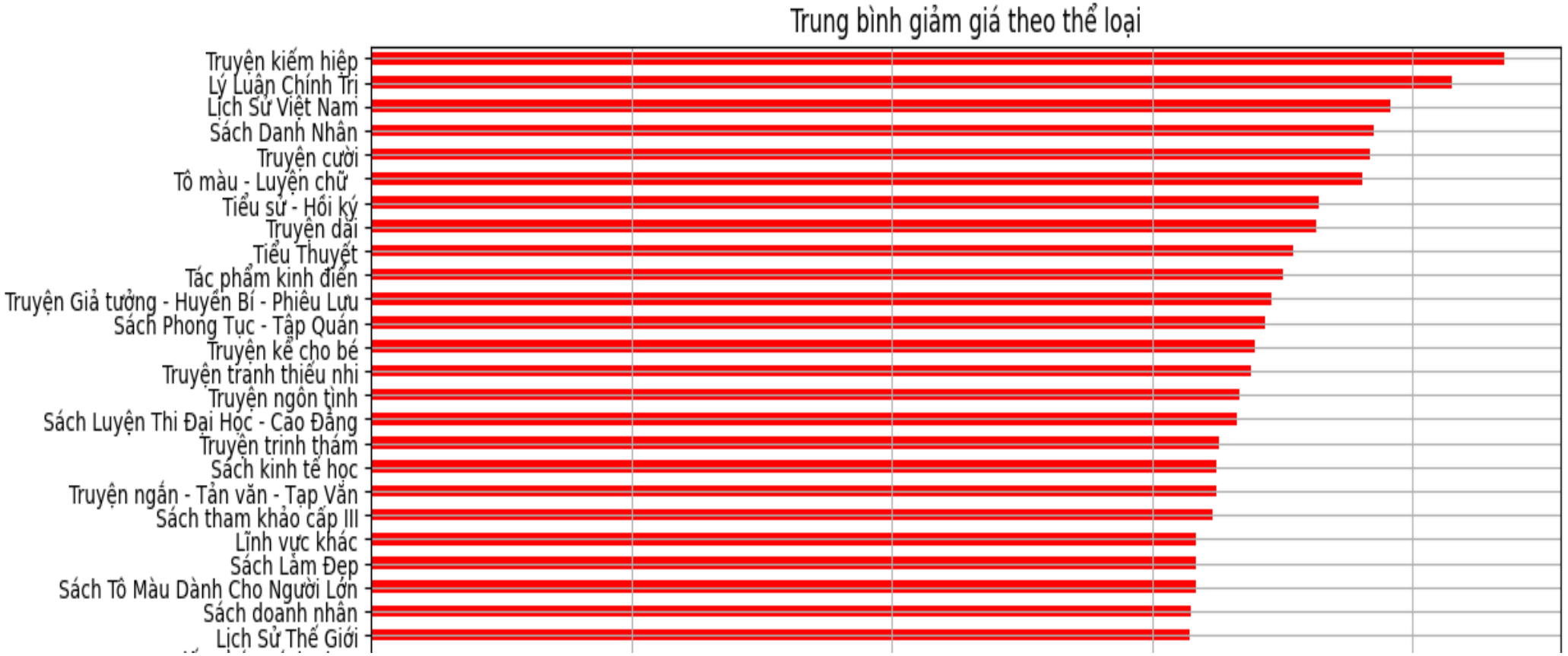
- So sánh giá gốc với giá hiện tại => Xác định phân khúc giá sách phổ biến ở VN
- Tìm hiểu giá sách bị tác động bởi những yếu tố nào hay không như sách càng mắc càng nhiều trang, hay phụ thuộc vào tác giả, thể loại,…
- Top những quyển sách, tác giả và thể loại bán chạy nhất.
- Top những quyển sách, tác giả và thể loại đánh giá cao nhất.
- Kết luận mức độ thu hút số lượng người mua sách dựa vào những yếu tố nào.
- Xây dựng mô hình phân loại các nhóm sách tiềm năng bán chạy, bình thường, không tiềm năng bán chạy.

4. TÌM HIỂU DỮ LIỆU

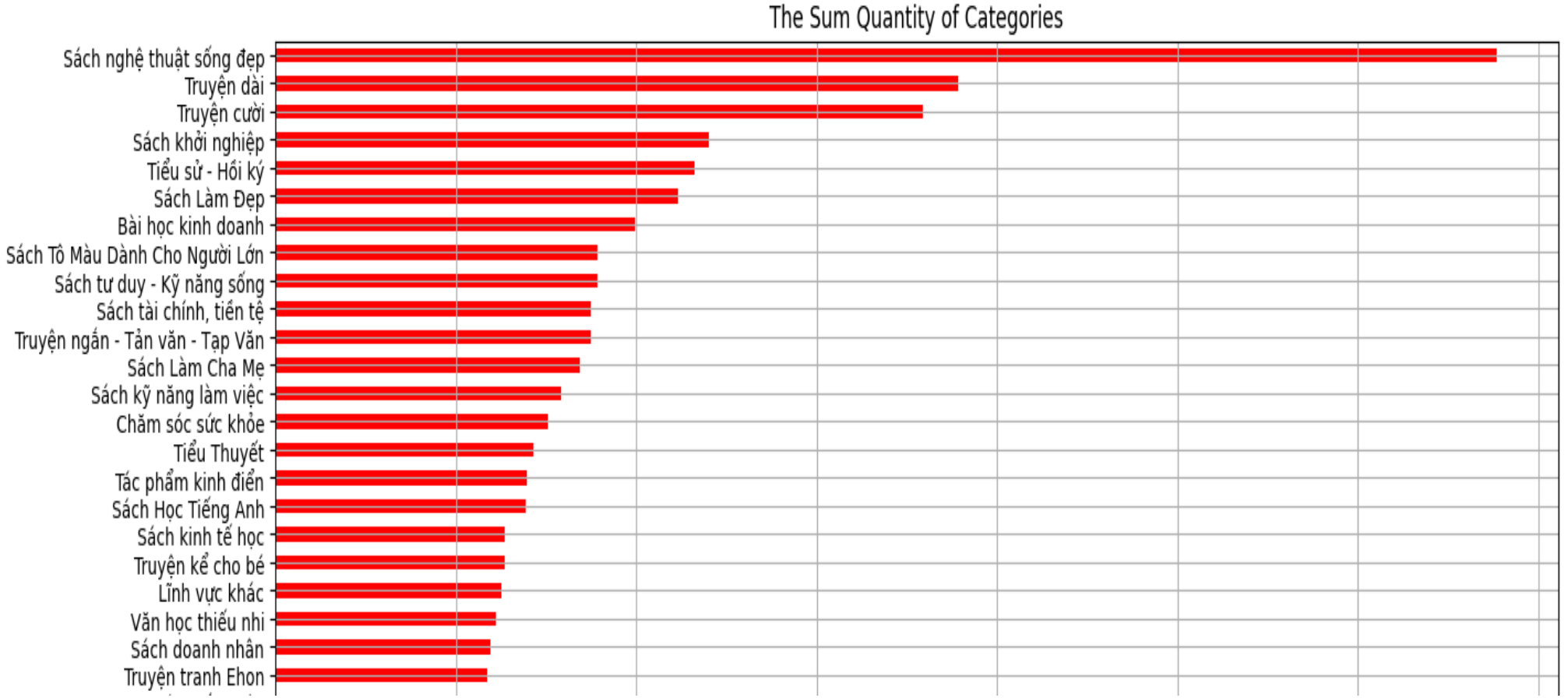


- Nhận thấy giá sách ở Việt Nam giao động từ 100-200k. Giá sách này tương đối thấp so với các thị trường khác

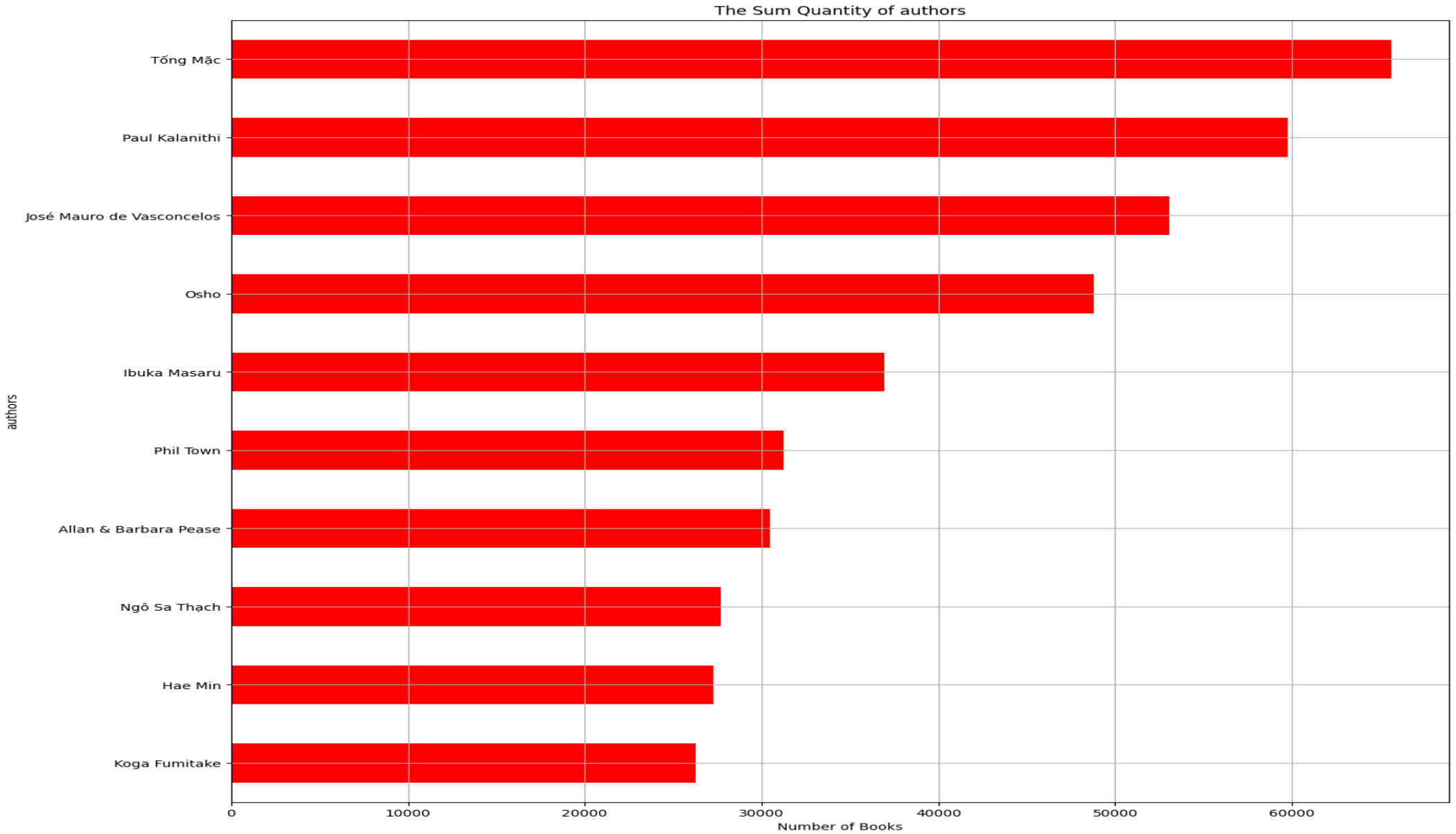
Dựa trên độ tương quan nhận thấy giá sách với sản lượng mua không tương quan nhau, giá với số trang không tương quan với nhau => Việc định giá sản phẩm không dựa trên số trang, hay sản lượng mua. Xem xét giá giảm với sản lượng vì có độ tương quan tương đối ổn



- Thể loại truyện có xu hướng giảm giá nhiều nhất



- Thể loại sách nghệ thuật sống đẹp, truyện, sách khởi nghiệp, làm đẹp có xu hướng được mua nhiều nhất => Ngày nay con người có xu hướng quan tâm đến sức khỏe tinh thần nhiều hơn trước


The Sum Quantity of authors
Tống Mặc
Paul Kalanithi
José Mauro de Vasconcelos
Osho
Ibuka Masaru
Phil Town
Allan & Barbara Pease
Ngô Sa Thạch
Hae Min
Koga Fumitake
authors
0
10000
20000
30000
40000
50000
60000
Number of Books


Avg, Rating of Categories
Sách Học Tiếng Nhật
Sách Học Tiếng Hàn
Sách Kiến Thức - Kỹ Năng Cho Trẻ
Sách Phong Tục - Tập Quán
Giáo Trình Đại Học - Cao Đẳng
Sách hướng nghiệp - Kỹ năng mềm
Sách tham khảo cấp III
Sách doanh nhân
Truyện đam mỹ
Sách giáo dục
Sách Tô Màu Dành Cho Người Lớn
Sách Địa Danh - Du Lịch
Sách Nấu ăn
Sách Thai Giáo
Triết Học
Truyện dài
Thơ
Sách Tâm Lý Tuổi Teen
Truyện ngôn tình
Sách Làm Đẹp
Chăm sóc sức khỏe
Truyện cười
Light novel
Mỹ Thuật - Kiến Trúc
Tin Học Văn Phòng
Sách khởi nghiệp
Sách Luyện Thi Đại Học - Cao Đẳng
Truyện ngắn - Tản văn - Tạp Văn
Sách Danh Nhân
Tiểu sử - Hồi ký
Truyện cổ tích
Sách quản trị nhân lực
Sách quản trị, lãnh đạo
Sách kinh tế học
Truyện kiếm hiệp
Sách Hôn Nhân - Giới Tính
Lịch Sử Việt Nam
Truyện kinh dị
Sách tài chính, kế toán
Lý Luận Chính Trị
Others
Sách Phong Thủy - Kinh Dịch
Sách Học Tiếng Anh
Tác phẩm kinh điển
Sách nghệ thuật sống đẹp
Truyện kể cho bé
Sách Marketing - Bán hàng
Sách Làm Cha Mẹ
Truyện Giả tưởng - Huyền Bí - Phiêu Lưu
Bài học kinh doanh
Truyện trinh thám
Sách Học Tiếng Hoa
Lịch Sử Thế Giới
Kiến Thức Bách Khoa
Sách Chiêm Tinh - Horoscope
Sách tài chính, tiền tệ
Tiểu Thuyết
Sách tư duy - Kỹ năng sống
Văn học thiếu nhi
Kiến thức - Bách khoa
Lĩnh vực khác
Truyện tranh Ehon
Sách kỹ năng làm việc
Du ký
Tô màu - Luyện chữ
Tranh Truyện
Phóng sự - Ký sự - Bút ký
Đạo đức - Kỹ năng sống
Truyện tranh thiếu nhi
Luật - Văn Bản Luật
Category
0
1
2
3
4
5
Avg. Rating

KẾT LUẬN

- Thể loại truyện có lượng giảm giá nhiều nhất nhưng sản lượng bán ra tương đối thấp không cao
- Thể loại sách nghệ thuật sống đẹp, truyện, sách khởi nghiệp, làm đẹp có xu hướng được mua nhiều nhất => Ngày nay con người có xu hướng quan tâm đến sức khỏe tinh thần nhiều hơn trước
- Nhận thấy thể loại, tác giả, n-review, rating và discount có phần tác động đến sản lượng mua sách. Tuy nhiên nhìn vào biểu đồ phân phối rate thì nhận thấy số rate khá tương tự nhau, giá sách có sự giao động không quá lớn => Yếu tố quan trọng để xác định được nhóm sách tiềm năng để bán là dựa vào tác giả, thể loại, số lượng review và giảm giá.

5. MÔ HÌNH PHÂN TÍCH CLUSTERING

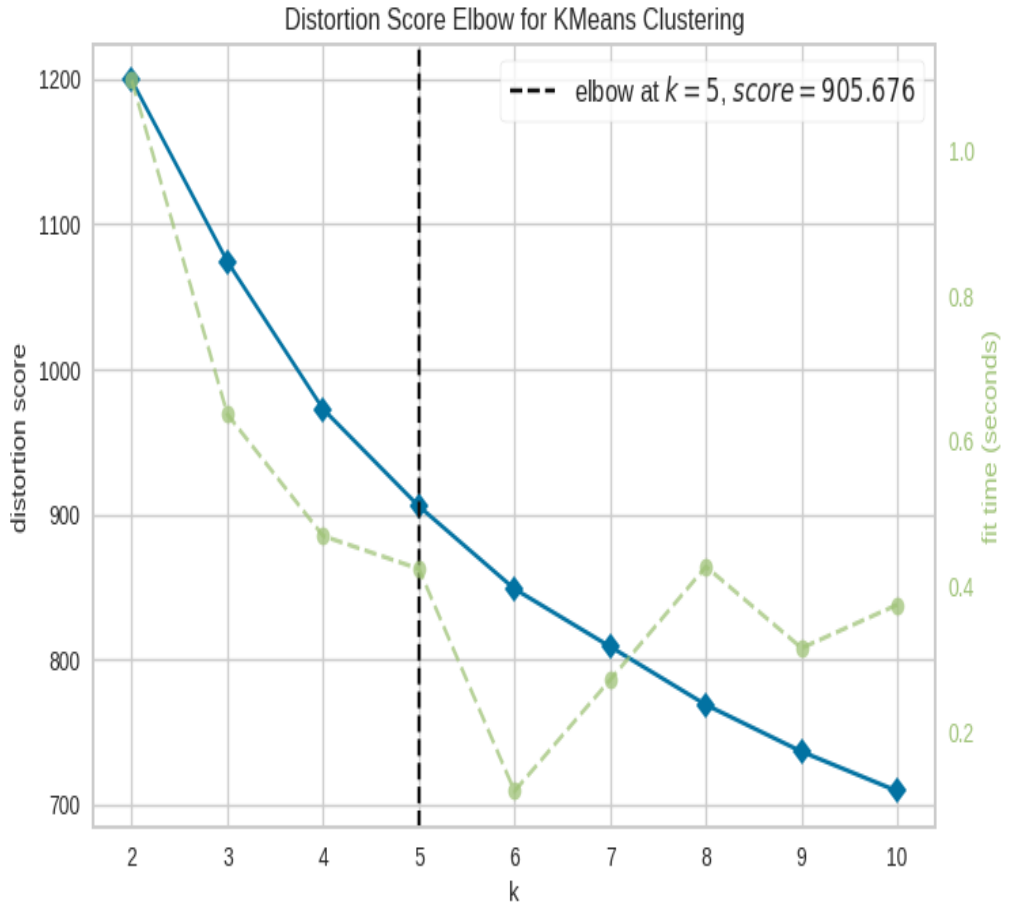
- **Các bước thực hiện:**

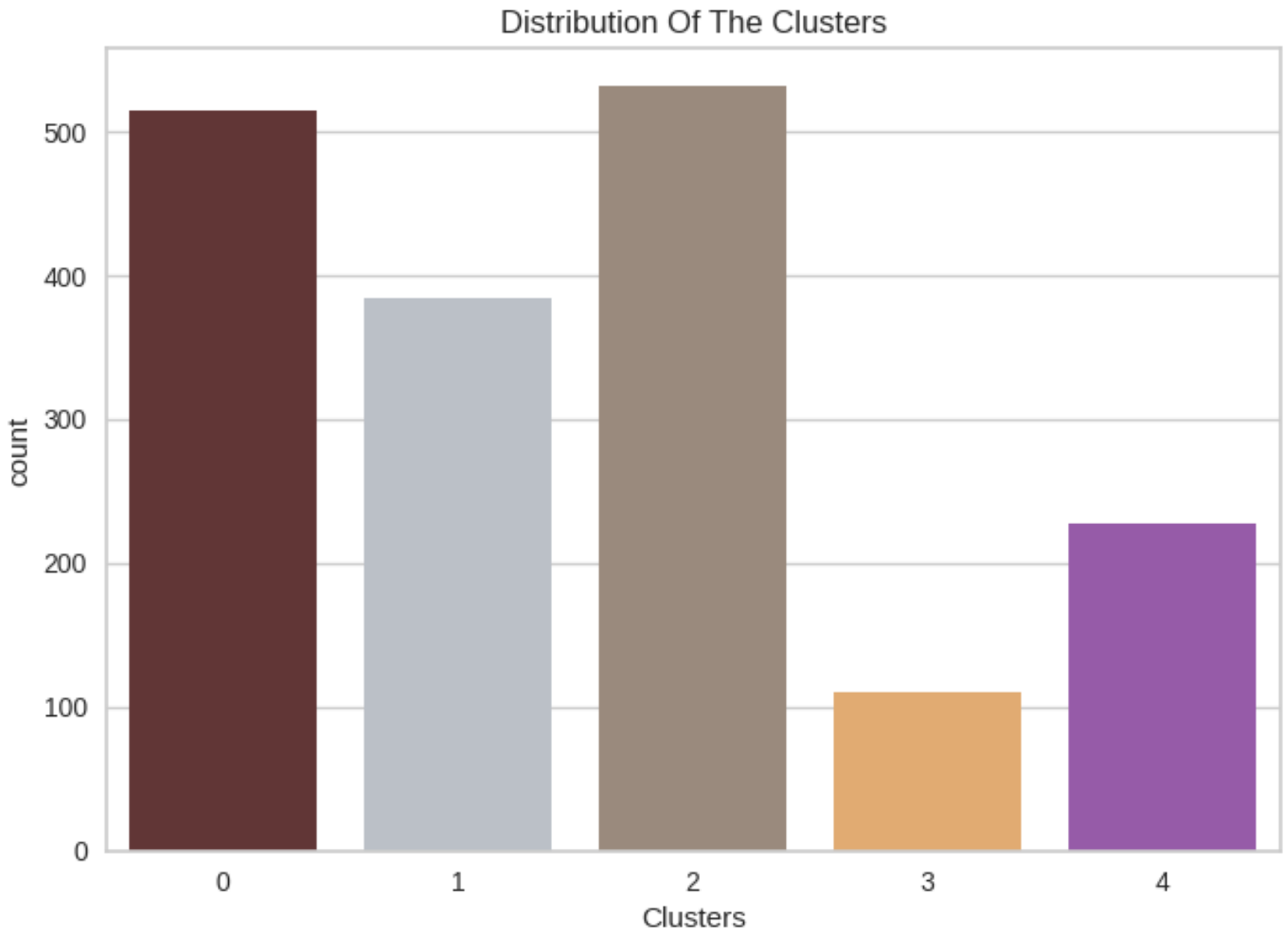
- Encode cho bộ dữ liệu

- Scaling cho bộ dữ liệu để số chính xác hơn

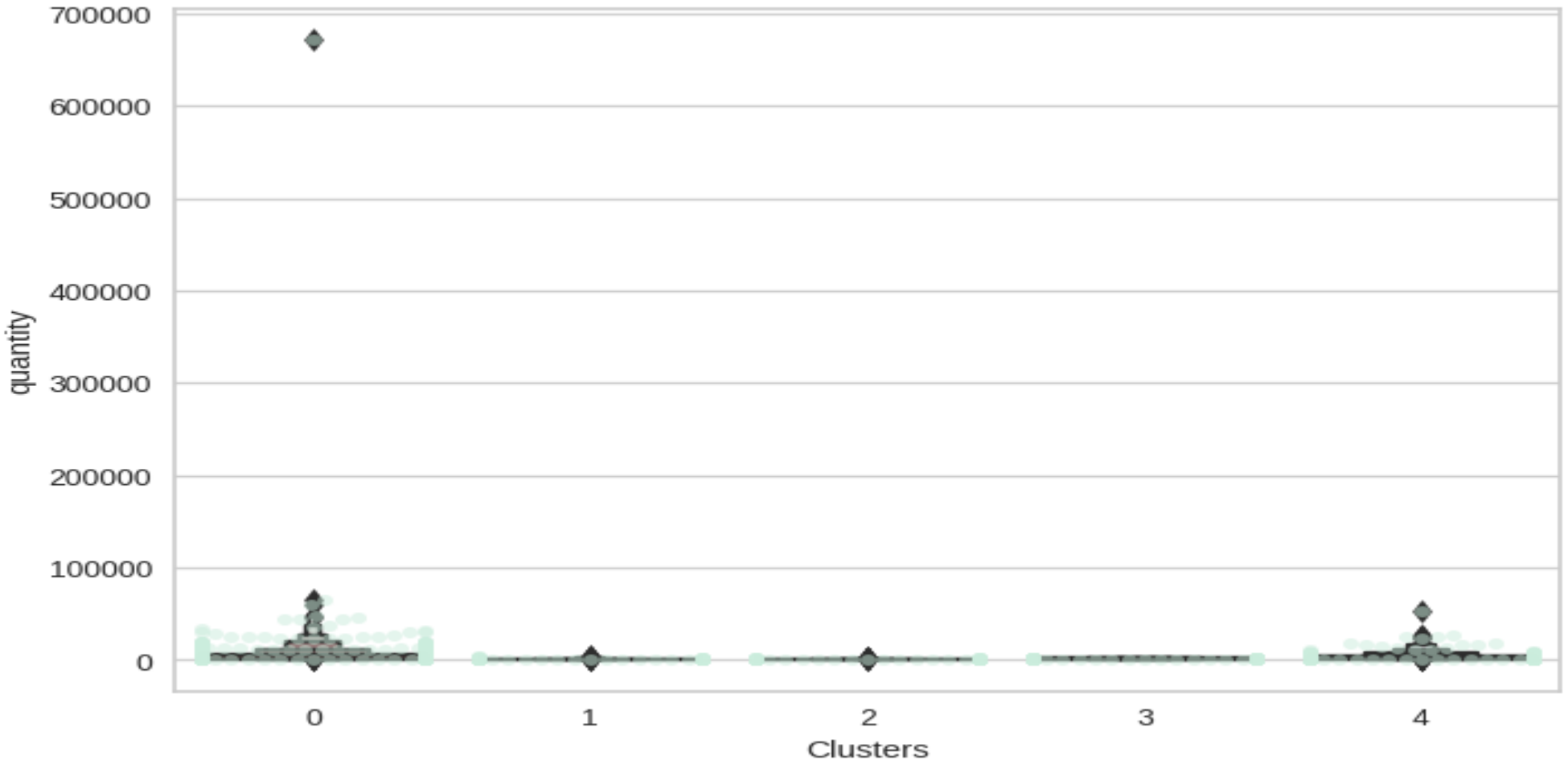
- elbow để tìm ra số cụm, dựa trên hình vẽ thấy số cụm là 5

- Áp dụng model (AgglomerativeCluster)

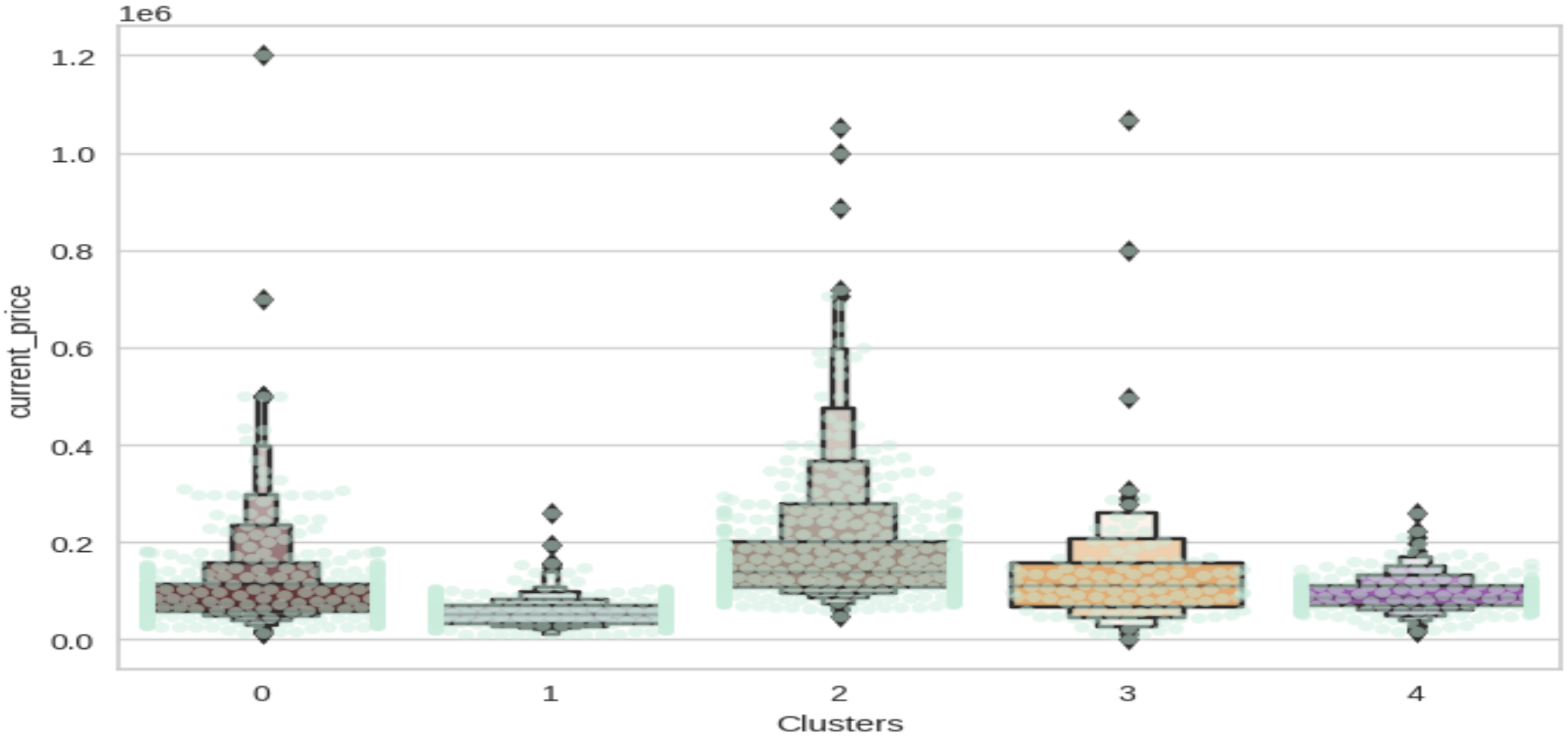




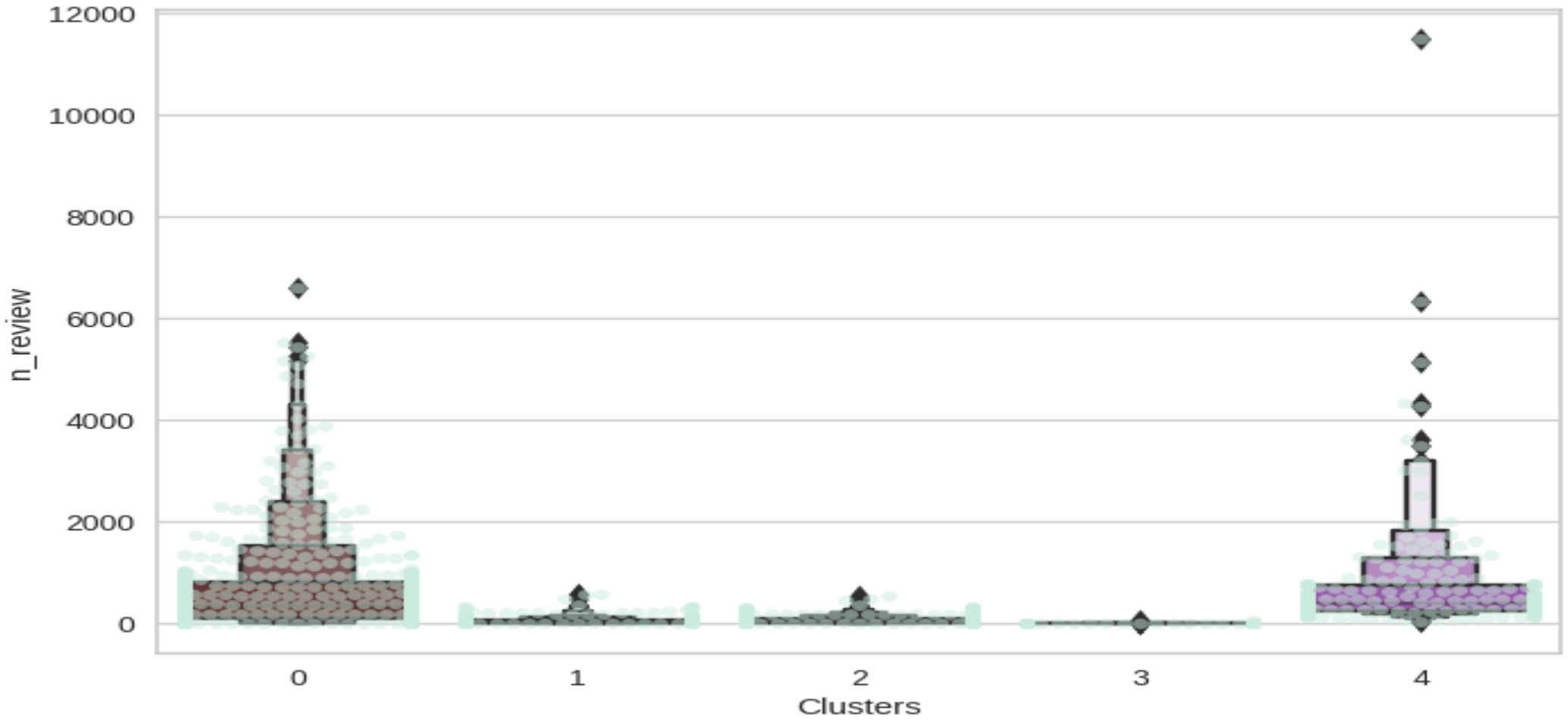
- Sơ đồ phân bố cụm chưa đồng đều



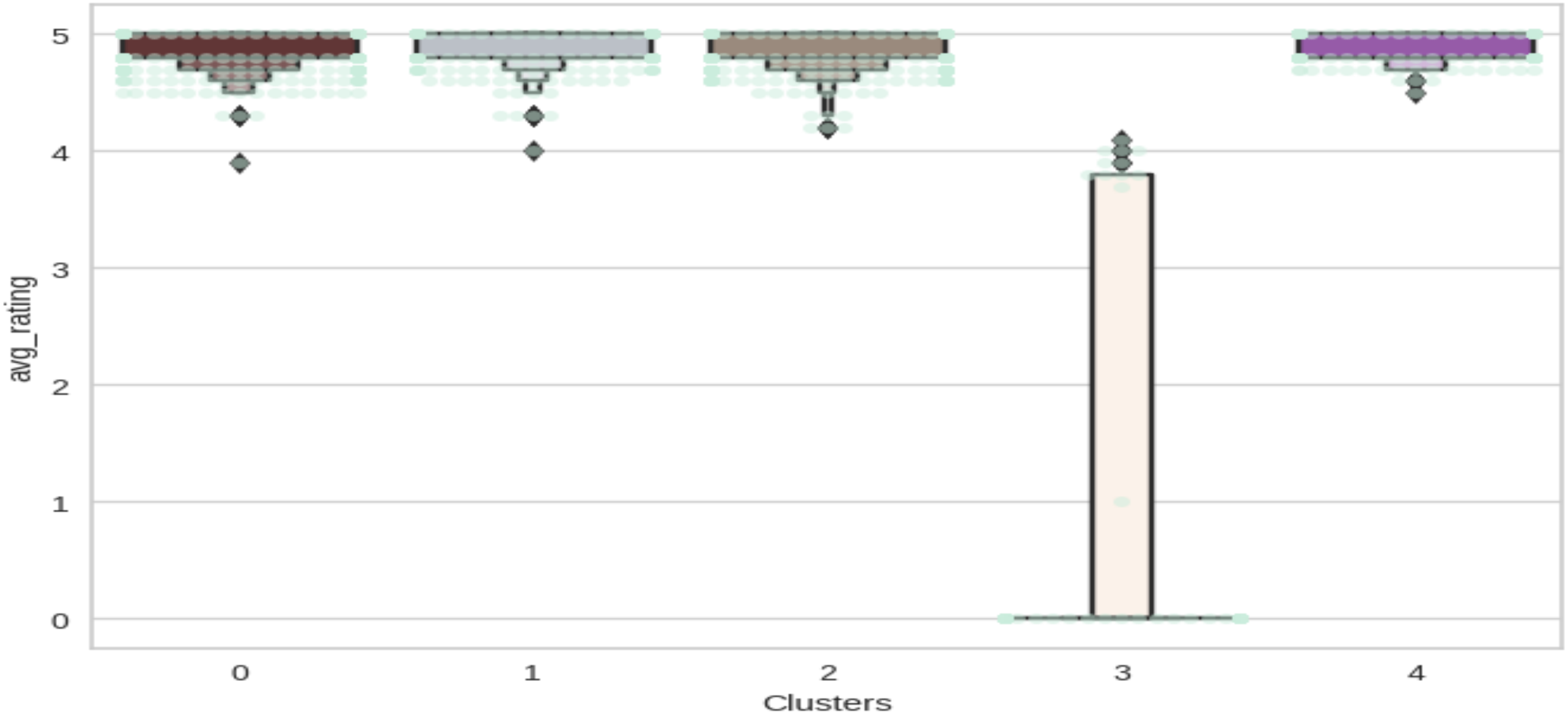
- Sản lượng sách chỉ yếu phân phối ở cụm 0, 4 => 2 cụm này có thể là danh sách sách tiềm năng có SL bán cao
- Tiếp đến là cụm 1,2
- Cuối cùng là cụm 3 => Cụm này có thể là ds sách có sl bán thấp



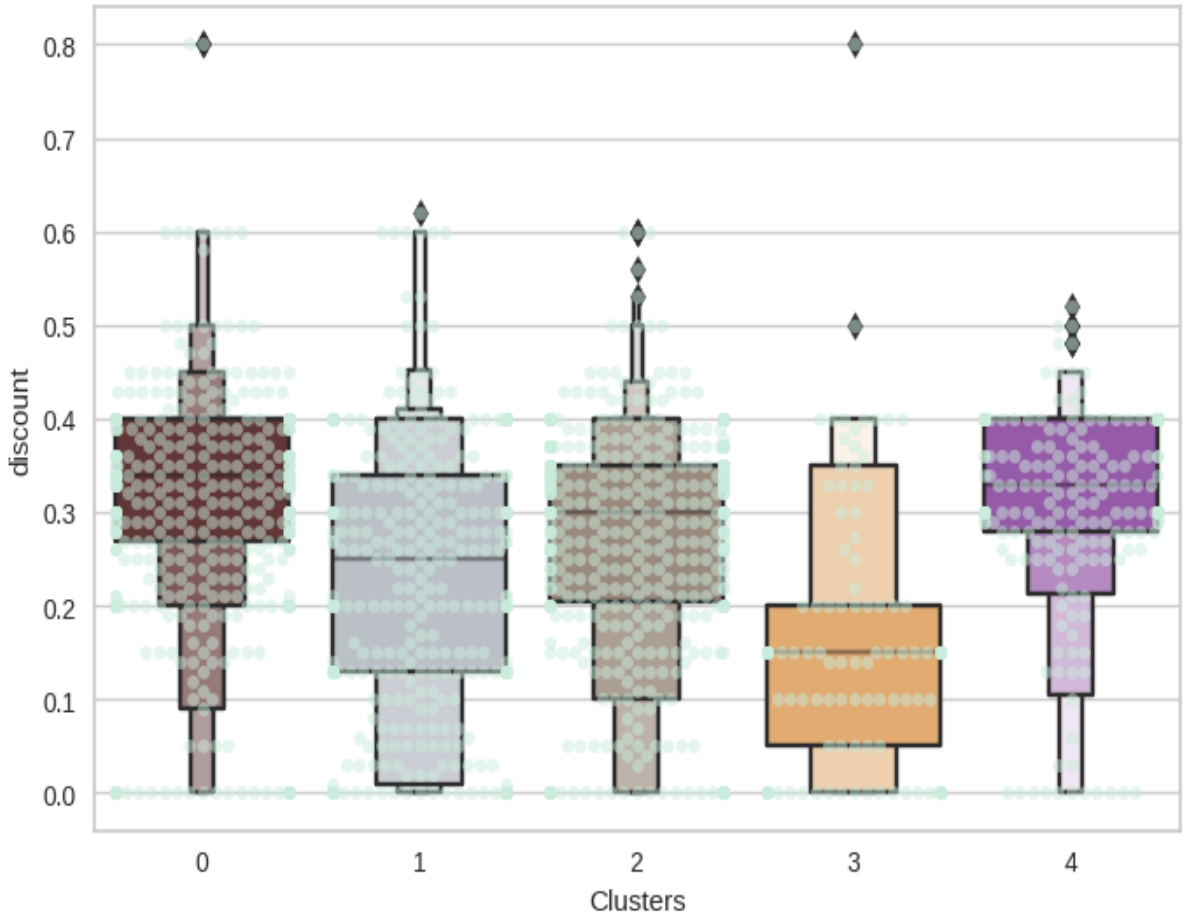
- Nhận thấy giá sách của cả 5 cụm tương đường nhau, trừ cụm 2 có hơi cao hơn một tí => Giá cả của tập dữ liệu này khá tương đồng nhau

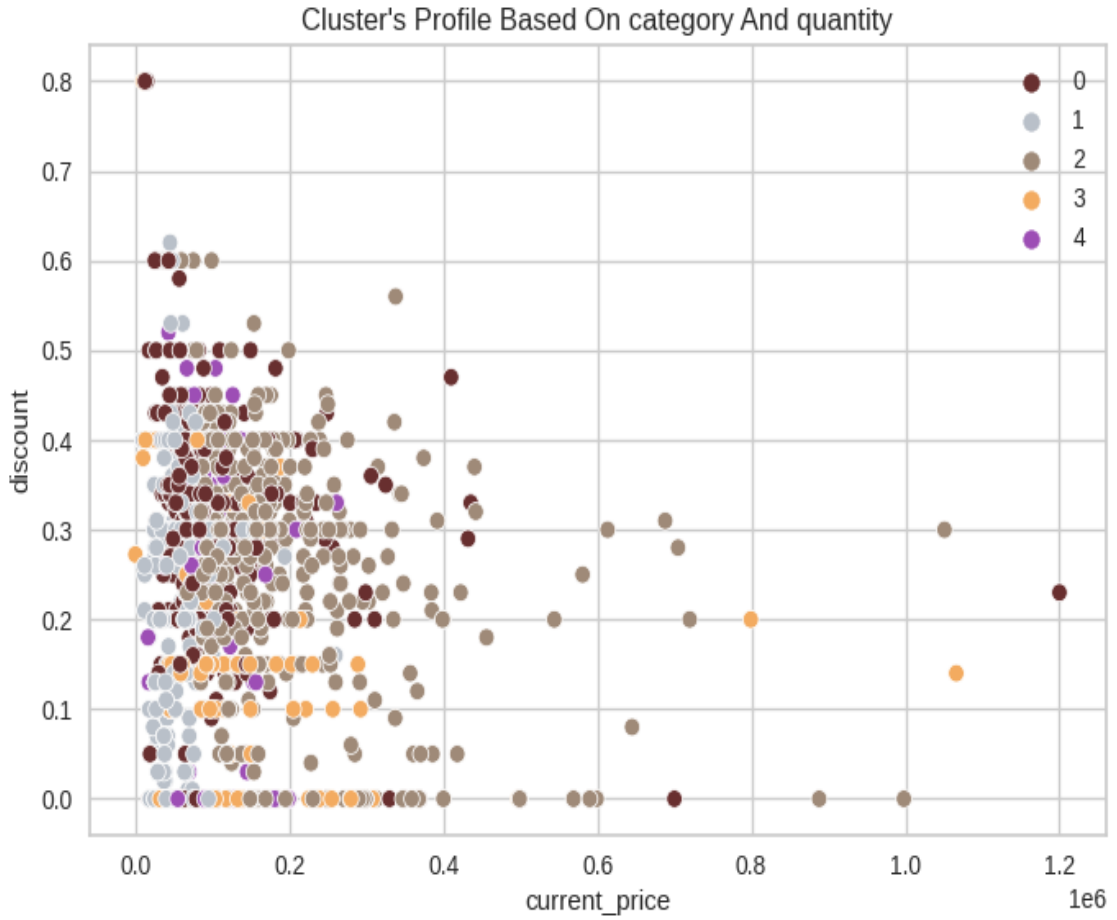


- Lượng review sách chỉ yếu phân phối ở cụm 0, 4 => 2 cụm này có thể là danh sách sách tiềm năng có SL bán cao
- Tiếp đến là cụm 1,2
- Cuối cùng là cụm 3 => Cụm này có thể là ds sách có sl bán thấp



- Nhận thấy đánh giá trung bình của 4 cụm 0, 1, 2, 4 tương đường nhau => Đây là những cụm được cho là ds tiềm năng để bán
- Cụm 3 có số đánh giá trung bình khá thấp => Cụm này là những dòng sách không tiềm năng bán chậm





- Giá trị giảm giá phân trải đều giữa các nhóm sách

- Dựa trên những gì đã nhìn thấy ở trên DS tiềm năng dựa trên thể loại, tác giả, số lượng review
- Dựa trên biểu đồ phân phối thể loại nhận thấy :

- Cụm 0, 4: Tập trung chủ yếu tại thể loại: Tiểu thuyết, sách tư duy, truyện ngắn, tản văn, truyện dài, tài chính, nghệ thuật
- Cụm 1, 2: Tập trung chủ yếu các thể loại: Sách giáo dục, truyện kinh dị, marketing, bán hàng, kỹ năng làm việc
- Cụm 3: Tập trung chủ yếu thể loại: Luật, đạo đức, truyện tranh, truyện tranh thiếu nhi

- Dựa trên biểu đồ phân phối nhà xuất bản nhận thấy :

- Cụm 0, 4: Nhà xuất bản thể giới, không biết, nhà xuất bản tổng hợp hcm, nxb văn huế, nxb đà nẵng
- Cụm 1, 2: Nxb nhà văn, nxb hà nội, dân trí,
- Cụm 3: nxb tri thức, thông tin, buôn ma thuộc

KẾT LUẬN

- **Cụm 0, 4: Là cụm có sản lượng sách bán cao nhất, nhiều lượt review, đánh giá trung bình cao**
- **SL bán gần 100.000 quyển sách**
- **Giá sách dao động tầm trong tầm dưới 400k**
- **Sl review trên 4000**
- **Đánh giá trung bình từ 4-5 sao**
- **Thể loại chủ yếu là:** Tiểu thuyết, sách tư duy, truyện ngắn, tản văn, truyện dài, tài chính, nghệ thuật
- **Cụm 1, 2:** Là cụm có sản lượng sách bán **tương đối**
- **SL bán gần 2.000 quyển sách**
- **Giá sách dao động tầm trong tầm dưới 200k**
- **Sl review trên 1000**
- **Đánh giá trung bình từ 4-5 sao**
- **Thể loại chủ yếu là:** Sách giáo dục, truyện kinh dị, marketing, bán hàng, kỹ năng làm việc
- **Cụm 3: Là cụm có SL bán thấp**
- Tập trung chủ yếu thể loại: Luật, đạo đức, truyện tranh, truyện tranh thiếu nhi
- Có lượng đánh giá trung bình thấp dưới 3 sao, giá thành cao trên lên đến 600k

LỌC DS SÁCH BÁN CHẠY NHẤT, THỂ LOẠI BÁN CHẠY

- Để kiểm tra xem model chạy đúng hay không => Bằng chủ quan thì nhận thấy mô hình khá đúng

```
        print("No common elements")
Topsell = list(set(ListQ) & set(ListRev) & set(ListRat))
Topsell
```

```
['Nhà Giả Kim (Tái Bản 2020)',
 'Hiểu Về Trái Tim (Tái Bản)',
 'Yêu Những Điều Không Hoàn Hảo',
 'Nóng Giận Là Bản Năng , Tĩnh Lặng Là Bản Lĩnh',
 'Những Tù Nhân Của Địa Lý',
 'Bí Mật Của Phan Thiên Ấn',
 'Bước Chậm Lại Giữa Thế Gian Vội Vã (Tái Bản)',
 'Cây Cam Ngọt Của Tôi',
 'Thay Đổi Cuộc Sống Với Nhân Số Học',
 'Tâm Lý Học Về Tiền',
 'Muôn Kiếp Nhân Sinh 2',
 'Chuyện Con Mèo Dạy Hải Âu Bay (Tái Bản 2019)',
 'Hoàng Tử Bé (Tái Bản 2019)',
 'Thiên Tài Bên Trái, Kẻ Điên Bên Phải (Tái Bản)',
 'Dám Bị Ghét',
 'Hai Số Phận (Bìa Cứng)',
 'Tâm Lý Học - Phác Họa Chân Dung Kẻ Phạm Tội',
 'Điều Kỳ Diệu Của Tiệm Tạp Hóa NAMIYA (Tái Bản)',
 'Chiến Binh Cầu Vồng (Tái Bản 2020)']
```

category
Sách tư duy - Kỹ năng sống
Truyện ngắn - Tản văn - Tạp Văn
Tiểu Thuyết
Others
Bài học kinh doanh
Sách Học Tiếng Anh
Sách kỹ năng làm việc
Sách tài chính, tiền tệ
Tác phẩm kinh điển
Truyện dài
Lĩnh vực khác
Truyện trinh thám
Sách nghệ thuật sống đẹp
Truyện Giả tưởng - Huyền Bí - Phiêu Lưu

authors
Unknown
Nguyên Phong
Higashino Keigo
Nguyễn Nhật Ánh
José Mauro de Vasconcelos
Hae Min
Cao Minh
Trang Anh
Jeffrey Archer
Ở Đây Zui Nè
Tống Mặc
Ngô Sa Thạch