



Final Project - IntroDS

BÁO CÁO ĐỒ ÁN NHẬP MÔN KHOA HỌC DỮ LIỆU

NHÓM 10

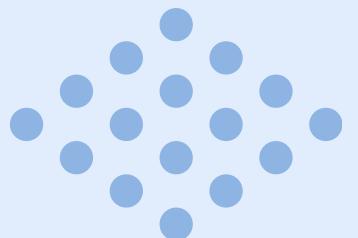


DANH SÁCH THÀNH VIÊN

Thành viên 1

22120099

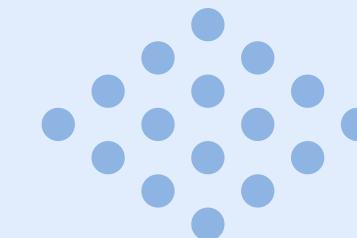
Trần Gia Hào



Thành viên 3

22120126

Nguyễn Tân Hưng



Thành viên 5

22120153

Trần Duy Khang

Thành viên 2

22120123

Nguyễn Minh Hưng

Thành viên 4

22120133

Hà Đức Huy

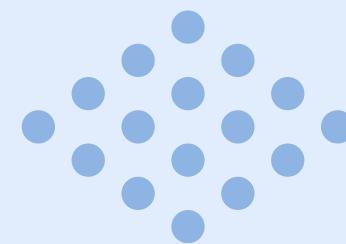
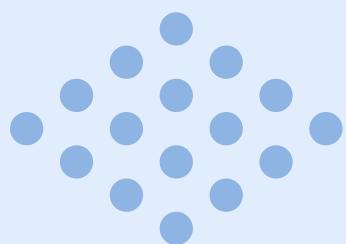
NỘI DUNG

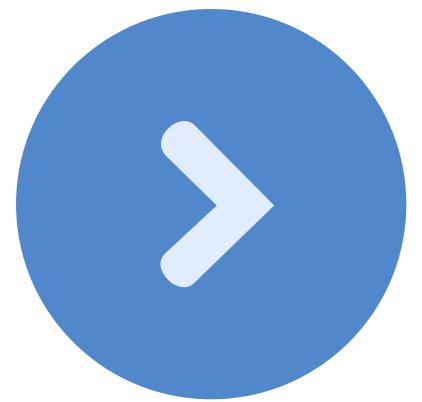
Thu thập và tiền
xử lý dữ liệu

Đặt câu hỏi có ý
nghĩa và trả lời

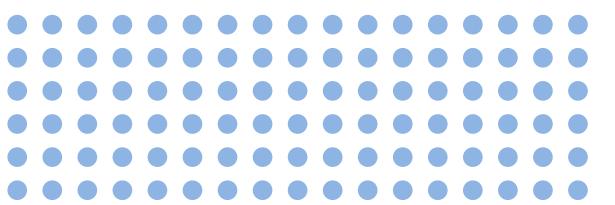
Khai phá dữ liệu

Mô hình hóa
dữ liệu



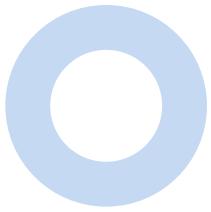


THU THẬP VÀ TIỀN XỬ LÝ DỮ LIỆU



BỐI CẢNH

- **Manga, Manhua, Manhwa, Light-novel** và các thể loại truyện khác là những hình thức nghệ thuật kể chuyện độc đáo, đại diện cho các nền văn hóa đa dạng, trở thành cầu nối văn hóa, phản ánh xã hội, tư tưởng, và những giá trị sâu sắc.
- Sự bùng nổ của các nền tảng như **MyAnimeList (MAL)**, **Anime-Planet**, hay các website đọc truyện trực tuyến đã mở ra cơ hội lớn để phân tích dữ liệu và khám phá xu hướng của thị trường màu mỡ này.
- Phân tích dữ liệu từ các nguồn này không chỉ giúp chúng ta hiểu sâu sắc hơn về sở thích của độc giả, nghiên cứu văn hóa và xã hội, cũng như dự đoán xu hướng phát triển các thể loại truyện trong tương lai.



THU THẬP DỮ LIỆU

GIỚI THIỆU

- **Chủ đề:** Các thể loại truyện (Manga, Manhua, Manhwa, Light Novel, ...)
- **Nơi thu thập:** MyAnimeList

CÁCH THU THẬP

- Sử dụng **requests** và **HTMLSession** để thu thập nội dung HTML của các bộ truyện
- Sử dụng **BeautifulSoup** và **re** để trích xuất các thông tin cần thiết

The screenshot shows the MyAnimeList website interface. At the top, there is a navigation bar with links for Anime, Manga, Community, Industry, Watch, Read, and Help. To the right of the navigation bar are buttons for Hide Ads, Login, and Sign Up. Below the navigation bar, a search bar is present with the placeholder "Search Anime, Manga, and more...". A banner for a "WEB NOVEL Writing Contest" is displayed, stating "Ends Jan.13". The main content area is titled "Top Manga" and shows a ranking table for manga. The first entry in the table is "Berserk", ranked 1st, with a score of 9.47 and a status of "N/A". The table includes columns for Rank, Title, Score, Your Score, and Status.

Rank	Title	Score	Your Score	Status
1	Berserk Manga (? vols) Aug 1989 -	★ 9.47	★ N/A	Add to My List

THU THẬP DỮ LIỆU

KẾT QUẢ THU THẬP

- Dữ liệu gồm 20000 dòng tương ứng với 20000 bộ truyện khác nhau được thu thập trên MyAnimeList, được sắp xếp theo số điểm giảm dần trên BXH All Manga
- Dữ liệu có tổng cộng 19 cột (19 thuộc tính được liệt kê ở hình bên cạnh)

ATTRIBUTES	GIẢI THÍCH
Title	Tên của bộ truyện được viết theo phiên âm tiếng Anh
Score	Điểm số của bộ truyện trên trang MyAnimeList (MAL)
Vote	Số lượng độc giả đã tham gia bình chọn và đánh giá cho bộ truyện.
Ranked	Thứ hạng của bộ truyện trên MyAnimeList
Popularity	Mức độ phổ biến của bộ truyện
Members	Số lượng độc giả đã thêm bộ truyện này vào danh sách cá nhân
Favorite	Số lượng độc giả đã thêm bộ truyện này vào danh sách yêu thích
Type	Loại hình của tác phẩm (Manga, Light Novel,...)
Volumes	Tổng số tập của bộ truyện. Một tập thường bao gồm một hoặc nhiều chương truyện.
Chapters	Tổng số chương truyện của bộ truyện.
Status	Trạng thái hiện tại của bộ truyện (Publishing, Finished,...)
Published	Thời gian phát hành của bộ truyện, bao gồm ngày bắt đầu và ngày kết thúc.
Genres	Thể loại của bộ truyện.
Themes	Chủ đề của bộ truyện.
Demographic	Đối tượng độc giả mà bộ truyện hướng đến.
Serialization	Thông tin về tạp chí hoặc nền tảng phát hành bộ truyện
Author	Tên tác giả của bộ truyện.
Total Review	Tổng số lượng độc giả đã để lại nhận xét hoặc đánh giá cho bộ truyện.
Type Review	Số lượng nhận xét của độc giả thành từng nhóm cụ thể: "Recommended", "Mixed feeling", và "Not recommended".



TIỀN XỬ LÝ DỮ LIỆU

KIỂM TRA DỮ LIỆU TRÙNG LẶP VÀ BỊ THIẾU

- Trong bộ dữ liệu được thu thập có 1 dòng bị trùng dữ liệu, nhóm tiến hành xóa dòng đó đi
- Tỉ lệ phần trăm giá trị Missing trong các cột không cao đến mức phải xóa luôn mà không cần xem xét. Cho nên, nhóm sẽ xem xét từng cột để xóa cột hay điền các giá trị thiếu này.

Tỉ lệ empty lists (missing ratio) 'Themes': 46.38%

Tỉ lệ empty lists (missing ratio) 'Genres': 2.79%

>>>>

Missing ratio (%):

Demographic	41.07
Serialization	16.51
Vote	0.36
Score	0.36
Author	0.25
Title	0.00
Published	0.00
Total_Review	0.00
Themes	0.00
Genres	0.00
Chapters	0.00
Status	0.00
Volumes	0.00
Types	0.00
Favorite	0.00
Members	0.00
Popularity	0.00
Ranked	0.00
Type_Review	0.00
dtype: float64	

TIỀN XỬ LÝ DỮ LIỆU

>>>>

XỬ LÝ DỮ LIỆU

- Điền các giá trị bị thiếu của **Score** bằng giá trị **min**
- Điền các giá trị bị thiếu của **Vote, Chapters, Volumes** bằng giá trị **median**
- Xóa các dữ liệu trong cột **Published** có giá trị **Not available**, đồng thời tách ra thành 2 cột **Released date** và **Completed date**
- Tách dữ liệu cột **Type Review** thành 3 cột là **Recommended, Mixed Feelings** và **Not Recommended**
- Xử lý cột **Author** để loại bỏ những đoạn dữ liệu thừa trong cột này, giúp dữ liệu sạch và gọn hơn
- Ghép lại 2 cột **Genres** và **Themes** thành 1 cột duy nhất là **Genres**, do 2 thuộc tính này khá tương đồng về ý nghĩa

```
<class 'pandas.core.frame.DataFrame'>
Index: 19057 entries, 0 to 19998
Data columns (total 21 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   Title            19057 non-null   object  
 1   Score             19057 non-null   float64 
 2   Vote              19057 non-null   int64  
 3   Ranked            19057 non-null   int64  
 4   Popularity        19057 non-null   int64  
 5   Members            19057 non-null   int64  
 6   Favorite           19057 non-null   int64  
 7   Types              19057 non-null   object  
 8   Volumes            19057 non-null   int64  
 9   Chapters            19057 non-null   int64  
 10  Status              19057 non-null   object  
 11  Genres             19057 non-null   object  
 12  Demographic        11052 non-null   object  
 13  Serialization       16135 non-null   object  
 14  Author              19057 non-null   object  
 15  Released date      19057 non-null   object  
 16  Completed date     19057 non-null   object  
 17  Total Review        19057 non-null   int64  
 18  Recommended          19057 non-null   int64  
 19  Mixed Feelings       19057 non-null   int64  
 20  Not Recommended      19057 non-null   int64  
dtypes: float64(1), int64(11), object(9)
memory usage: 3.2+ MB
```

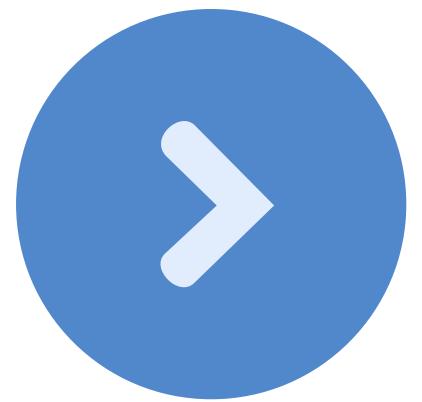
TIỀN XỬ LÝ DỮ LIỆU

>>>>

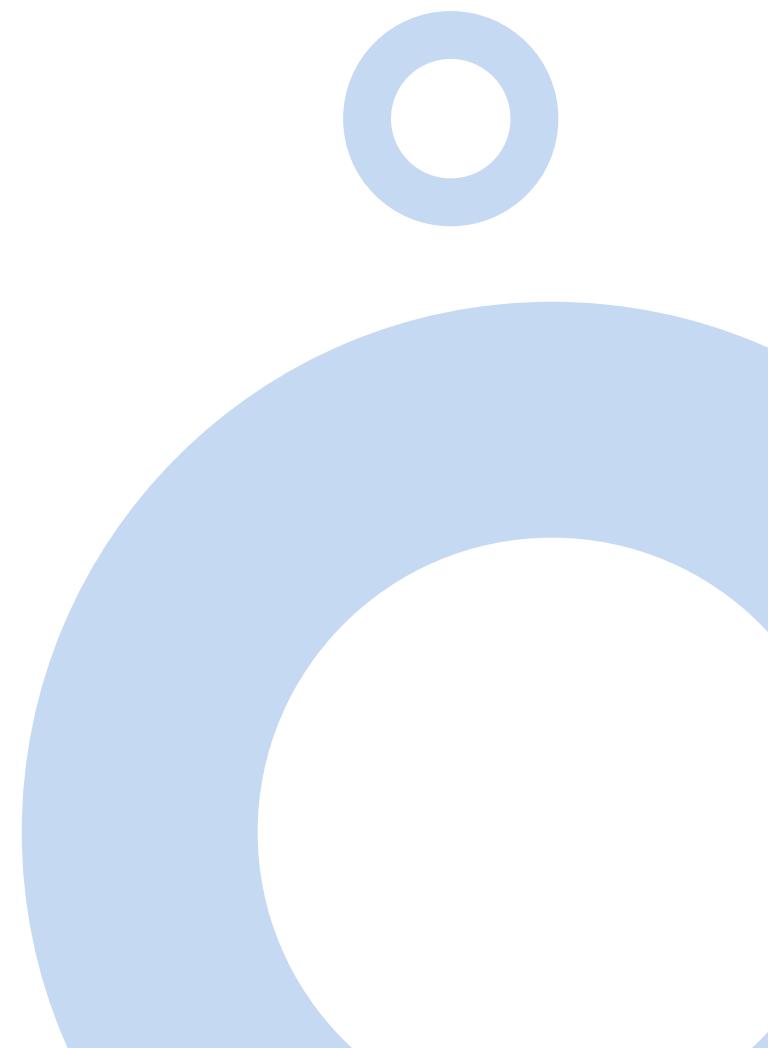
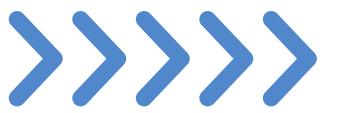
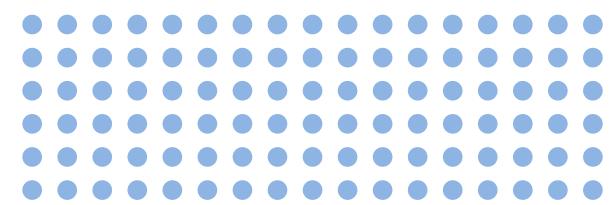
XỬ LÝ DỮ LIỆU

- Mỗi dòng dữ liệu trong bộ dữ liệu này là dữ liệu về một bộ truyện được thu thập trên MyAnimeList tính đến tháng 11/2024
- Mỗi cột dữ liệu trong bộ dữ liệu này lần lượt có ý nghĩa như sau:

ATTRIBUTES	MÔ TẢ	TIÊU CHÍ
Title	Tên của bộ truyện	Tên bộ truyện
Score	Điểm số trung bình của bộ truyện	Điểm số trung bình, càng cao càng tốt
Vote	Số lượt bình chọn cho bộ truyện	Số lượt bình chọn, càng cao càng tốt
Ranked	Xếp hạng của bộ truyện	Xếp hạng, càng thấp càng tốt
Popularity	Độ phổ biến của bộ truyện	Độ phổ biến, càng cao càng tốt
Members	Số thành viên theo dõi bộ truyện	Số thành viên, càng cao càng tốt
Favorite	Số lượt đánh dấu yêu thích bộ truyện	Số lượt yêu thích, càng cao càng tốt
Type	Thể loại của bộ truyện (ví dụ: Manga, Novel)	Thể loại
Volumes	Số lượng tập đã phát hành	Số lượng tập, càng cao càng tốt
Chapters	Số lượng chương đã phát hành	Số lượng chương, càng cao càng tốt
Status	Tình trạng phát hành của bộ truyện	Tình trạng phát hành
Genres	Thể loại và chủ đề của bộ truyện	Danh sách các thể loại
Demographic	Nhóm đối tượng hướng đến của bộ truyện	Đối tượng hướng đến
Serialization	Nơi đăng tải bộ truyện	Nơi đăng tải
Author	Tác giả của bộ truyện	Tác giả
Realeased date	Ngày bắt đầu phát hành bộ truyện	Ngày bắt đầu phát hành
Completed date	Ngày hoàn thành phát hành bộ truyện	Ngày hoàn thành phát hành
Total Review	Tổng số bài đánh giá	Tổng số bài đánh giá, càng cao càng tốt
Recommended	Số lượng lượt đề xuất	Số lượng đề xuất, càng cao càng tốt
Mixed Feelings	Số lượng lượt cảm xúc lẫn lộn	Số lượng cảm xúc lẫn lộn
Not Recommended	Số lượng lượt không đề xuất	Số lượng không đề xuất, càng thấp càng tốt



KHAI PHÁ DỮ LIỆU



KHAI PHÁ DỮ LIỆU

>>>>

KẾT QUẢ DỮ LIỆU XỬ LÝ

- Sau khi tiền xử lý xong, dữ liệu có **19057 dòng** và **21 cột**
- Nhóm sẽ xem xét các thuộc tính của bộ dữ liệu theo 3 loại:
Numerical, Datetime và **Category**

2.1.1 Số dòng và số cột của bộ dữ liệu

```
n_rows , n_cols = manga_df.shape  
print(f'Có {n_rows} dòng và {n_cols} cột trong bộ dữ liệu')
```

Có 19057 dòng và 21 cột trong bộ dữ liệu

KHAI PHÁ DỮ LIỆU



KẾT QUẢ DỮ LIỆU XỬ LÝ

- Có 2 cột là **Demographic** và **Serialization** bị thiếu dữ liệu ở các dòng, nhưng không ảnh hưởng đến tổng quan bộ dữ liệu
- Nhìn chung, bộ dữ liệu được xử lý tốt với các cột quan trọng có **missing_ratio** là 0%

	Score	Vote	Ranked	Popularity	Members	Favorite	Volumes	Chapters	Total Review	Recommended	Mixed Feelings	Not Recommended
count	19057.00	19057.00	19057.00	19057.00	19057.00	19057.00	19057.00	19057.00	19057.00	19057.00	19057.00	19057.00
mean	6.94	2312.01	9900.44	12272.88	5918.04	195.99	4.48	32.34	2.65	1.90	0.44	0.31
std	0.61	11053.33	5792.61	8315.95	21995.12	2084.47	5.72	73.59	9.72	7.09	1.74	1.75
min	2.43	100.00	1.00	1.00	55.00	0.00	1.00	1.00	0.00	0.00	0.00	0.00
25%	6.59	230.00	4858.00	5227.00	788.00	2.00	2.00	8.00	0.00	0.00	0.00	0.00
50%	6.93	507.00	9837.00	11250.00	1593.00	7.00	3.00	16.00	1.00	0.00	0.00	0.00
75%	7.29	1392.00	14926.00	18226.00	4058.00	30.00	4.00	31.00	2.00	2.00	0.00	0.00
max	9.47	415004.00	20000.00	58477.00	725079.00	130489.00	200.00	6477.00	448.00	280.00	74.00	114.00
missing ratios	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

	Genres	Author
missing_ratio	0.0	0.188907
unique	72	13265

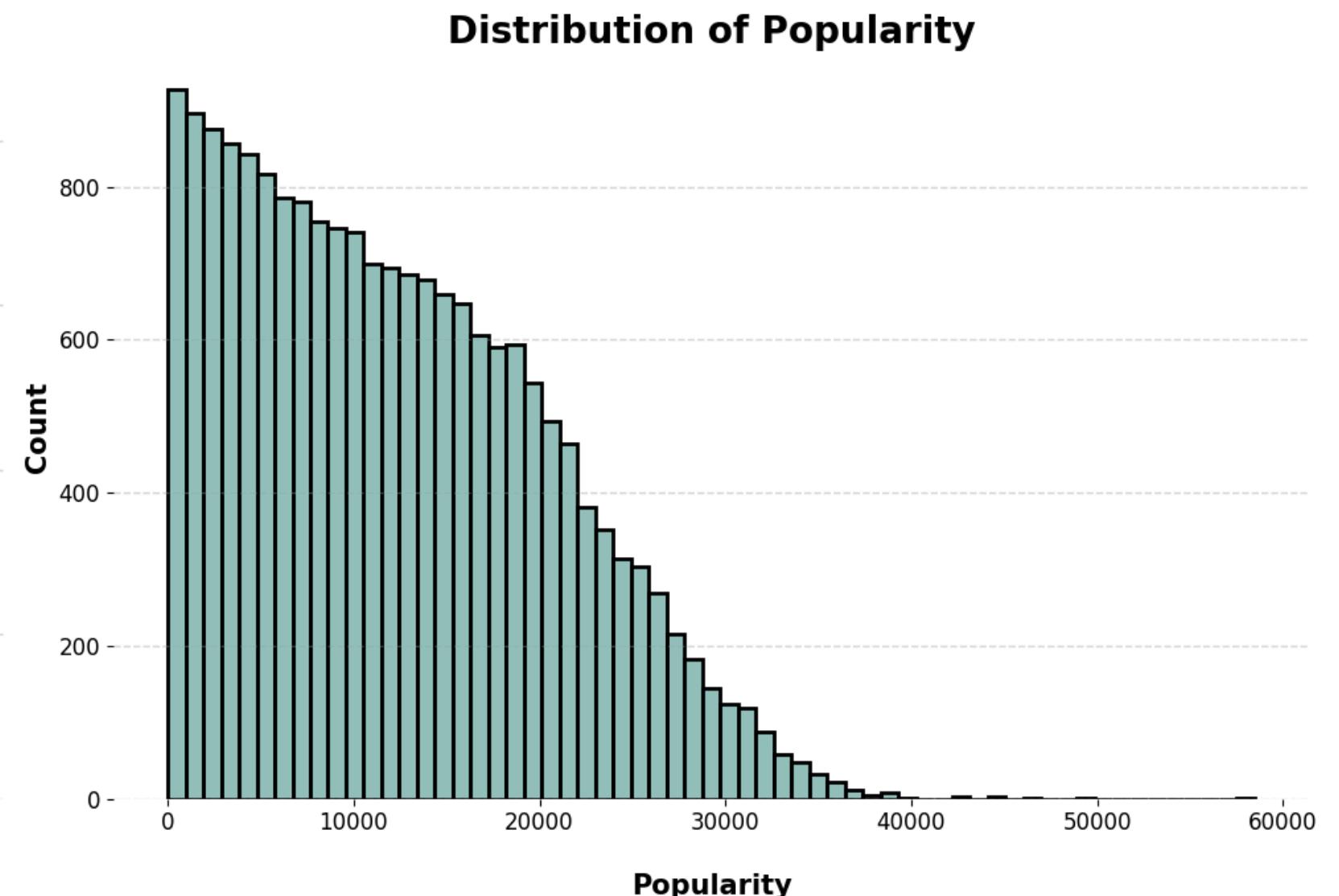
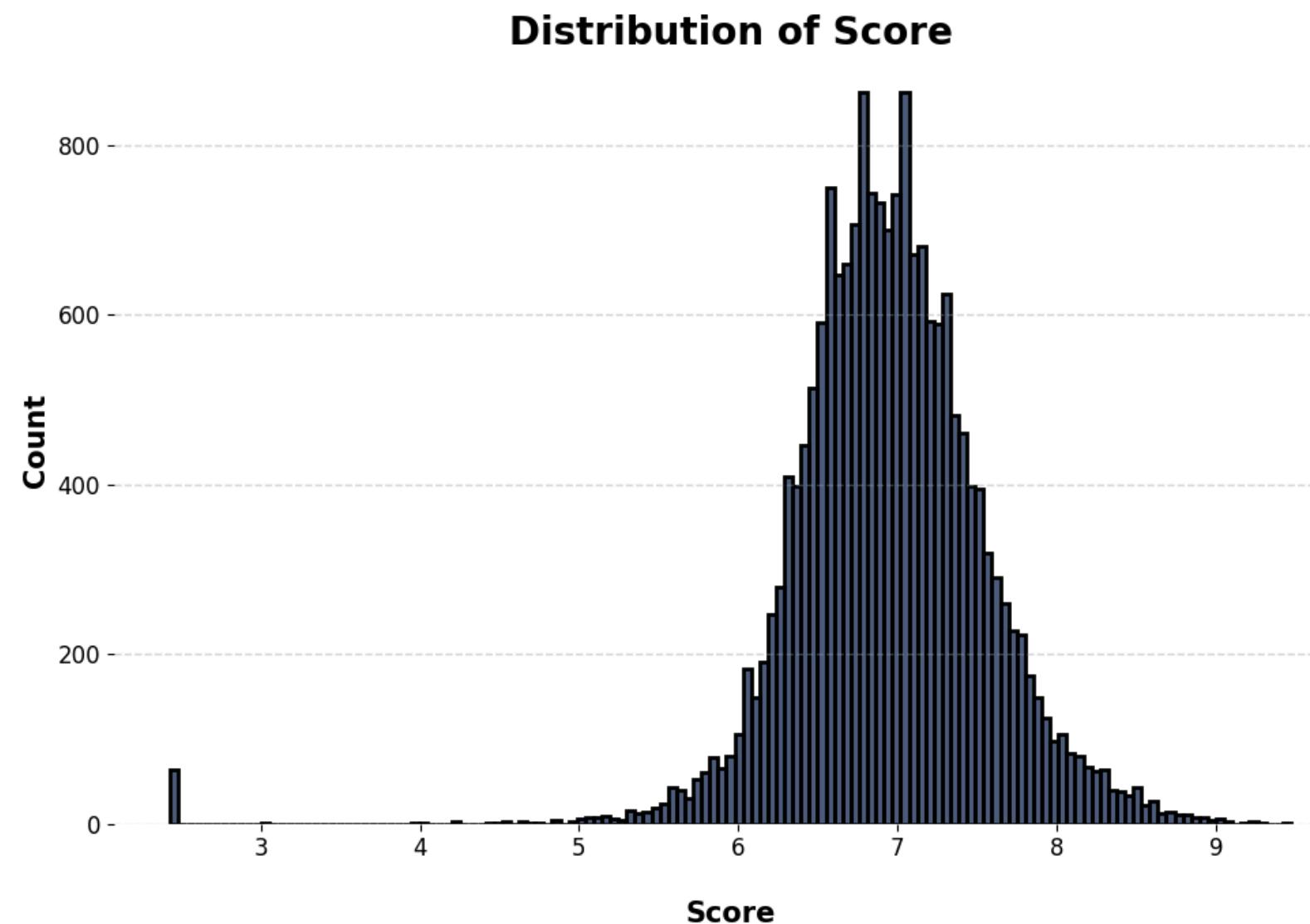
Distribution {"Romance": 7389, "Comedy": 6140, "Fantasy": 15535, "Tezuka, Osamu": 65, "Nagai, Go": 61, "Ito, ...}

	Title	Types	Status	Demographic	Serialization
missing_ratio	0.0	0.0	0.0	42.005562	15.332949
unique	18516	7	4	5	845
Distribution	{"Clover": 4, "Blue": 4, "Legend": 3, "Orange": 1, "Manga": 14407, "Manhwa": 1770, "One-shot": 1, "Finished": 15535, "Publishing": 3322, "On Hi...": 1}	{"Shoujo": 3769, "Shounen": 3266, "Seinen": 31...": 1}	{"KakaoPage": 694, "Naver Webtoon": 592, "Shou...": 1}		

KHAI PHÁ DỮ LIỆU

Nhóm thuộc tính Numerical

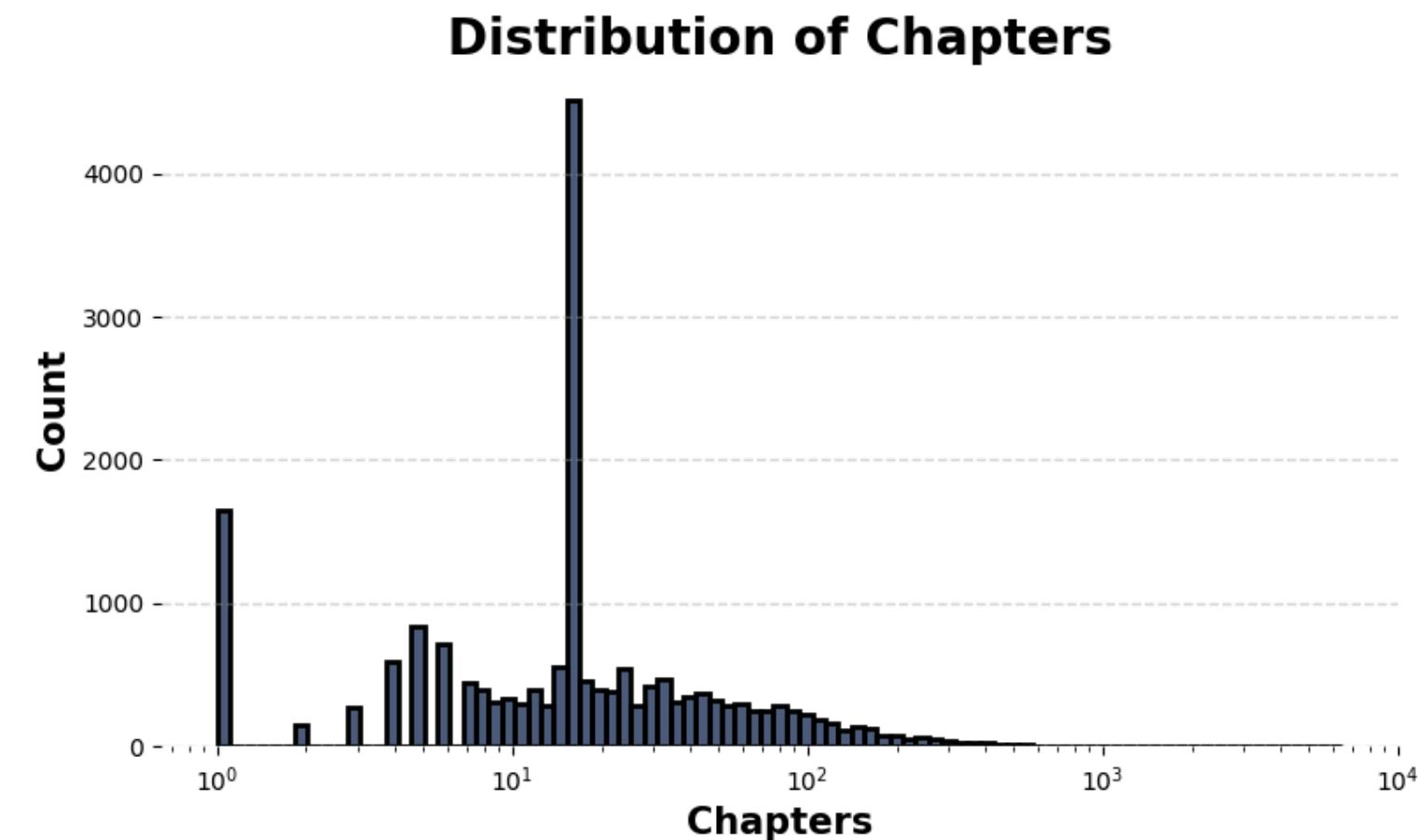
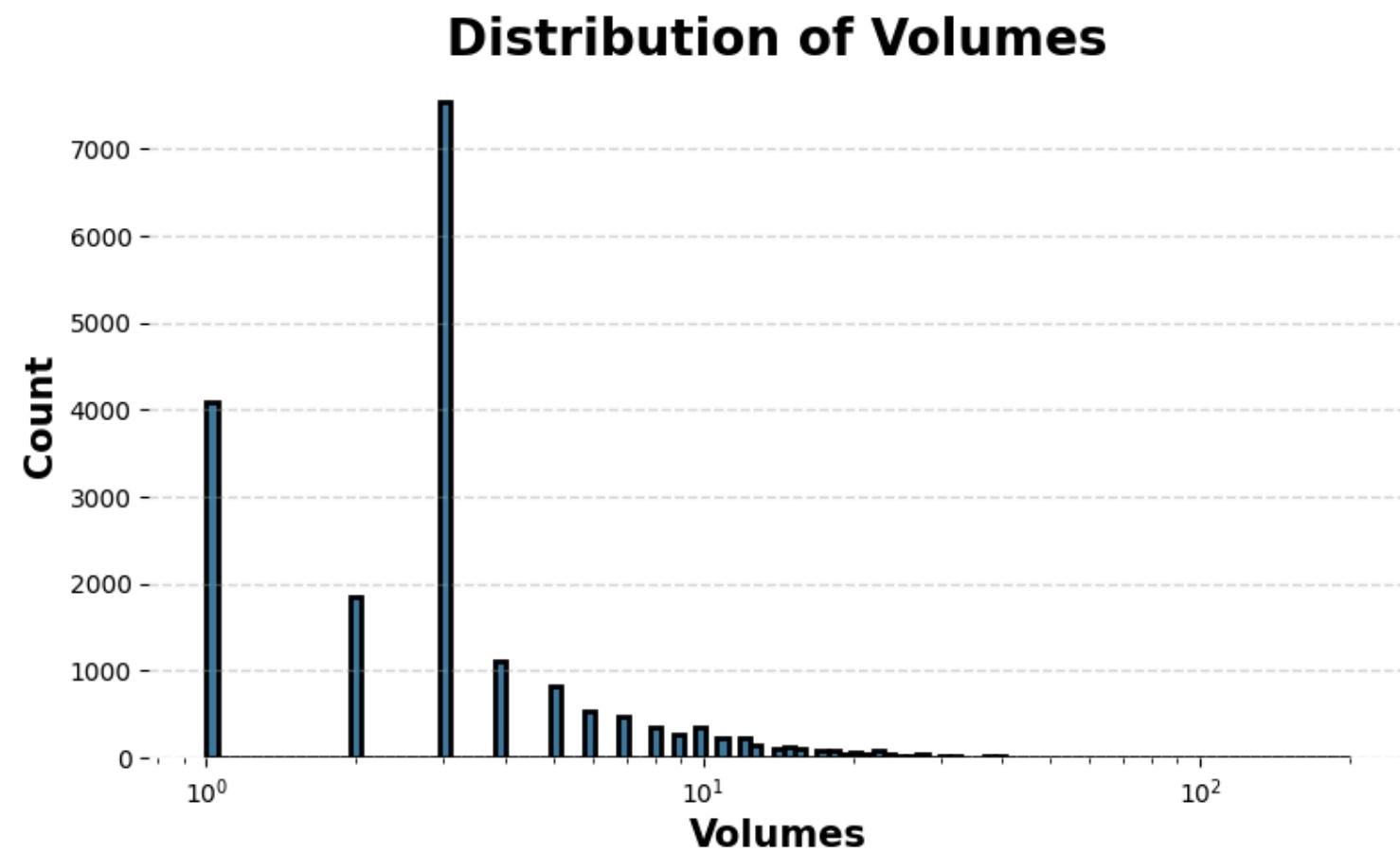
- Phân phối của các thuộc tính Numerical trong bộ dữ liệu này thường có **phân phối chuẩn hoặc lệch phai**.



KHAI PHÁ DỮ LIỆU

Nhóm thuộc tính Numerical

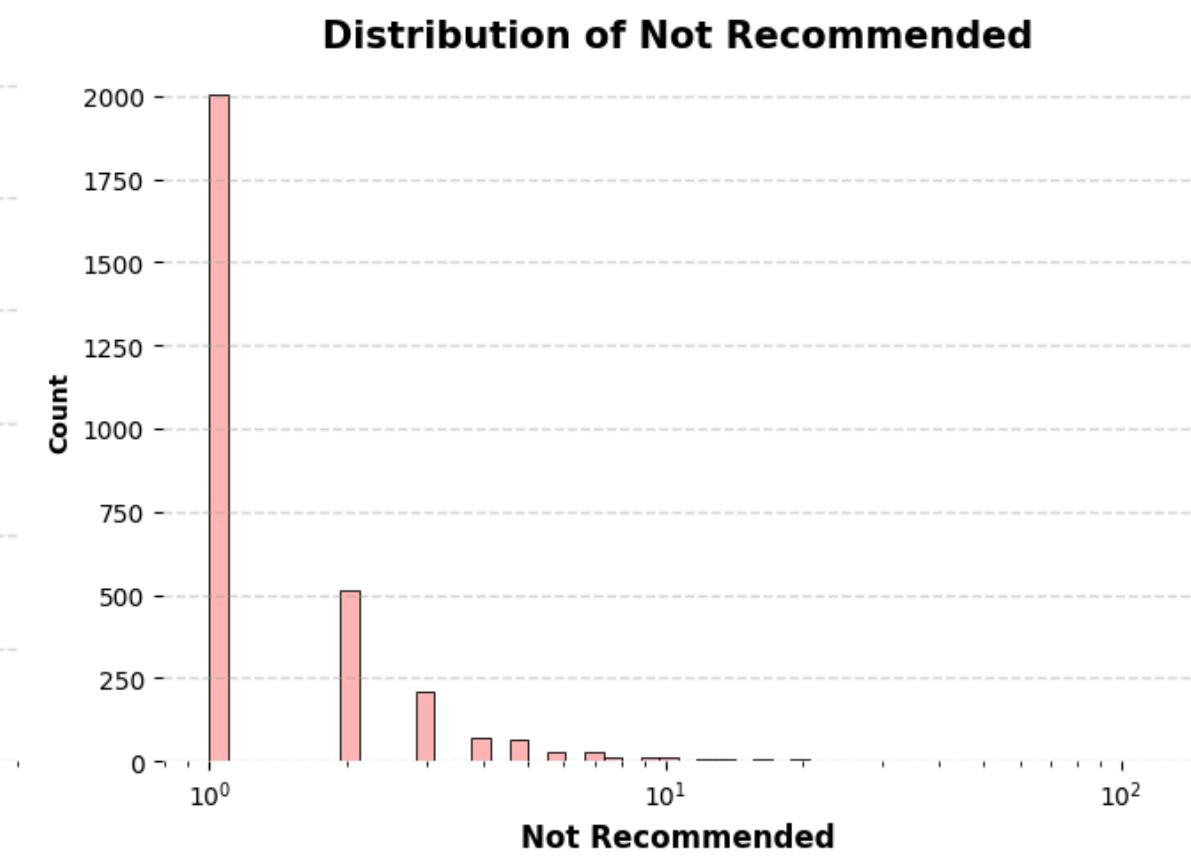
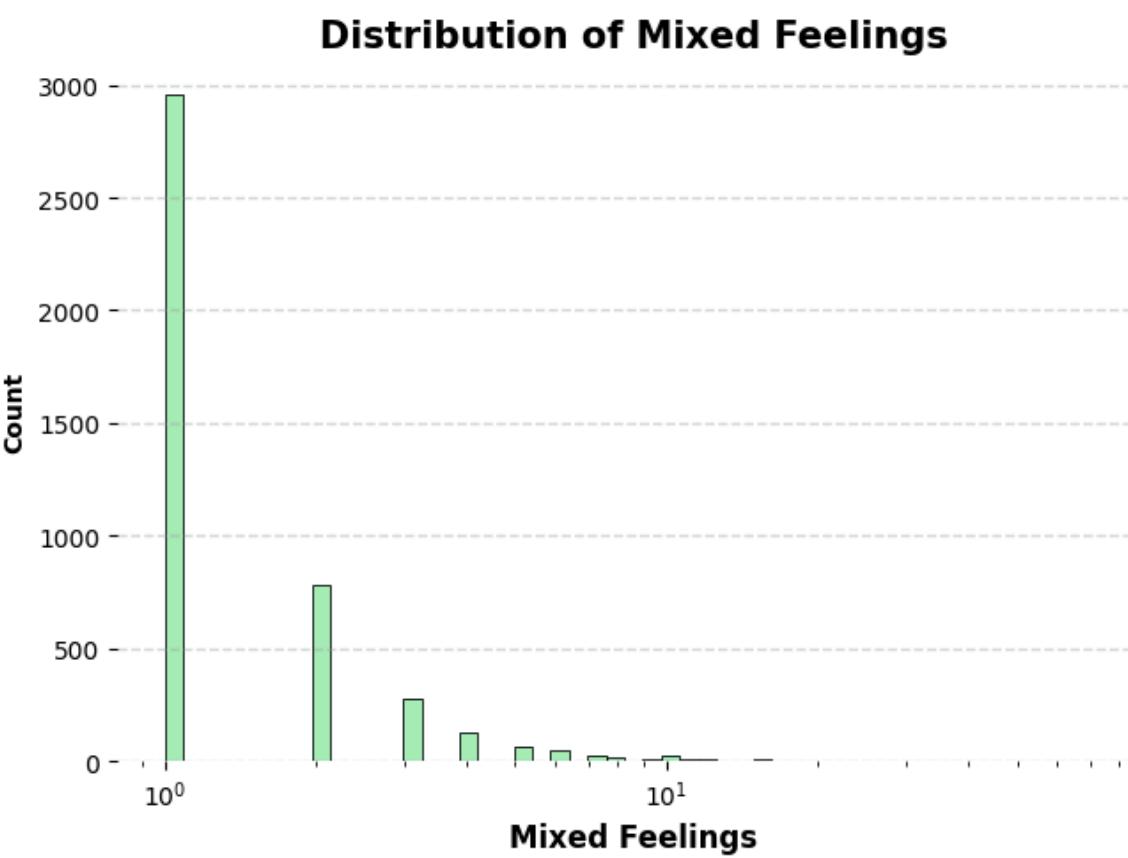
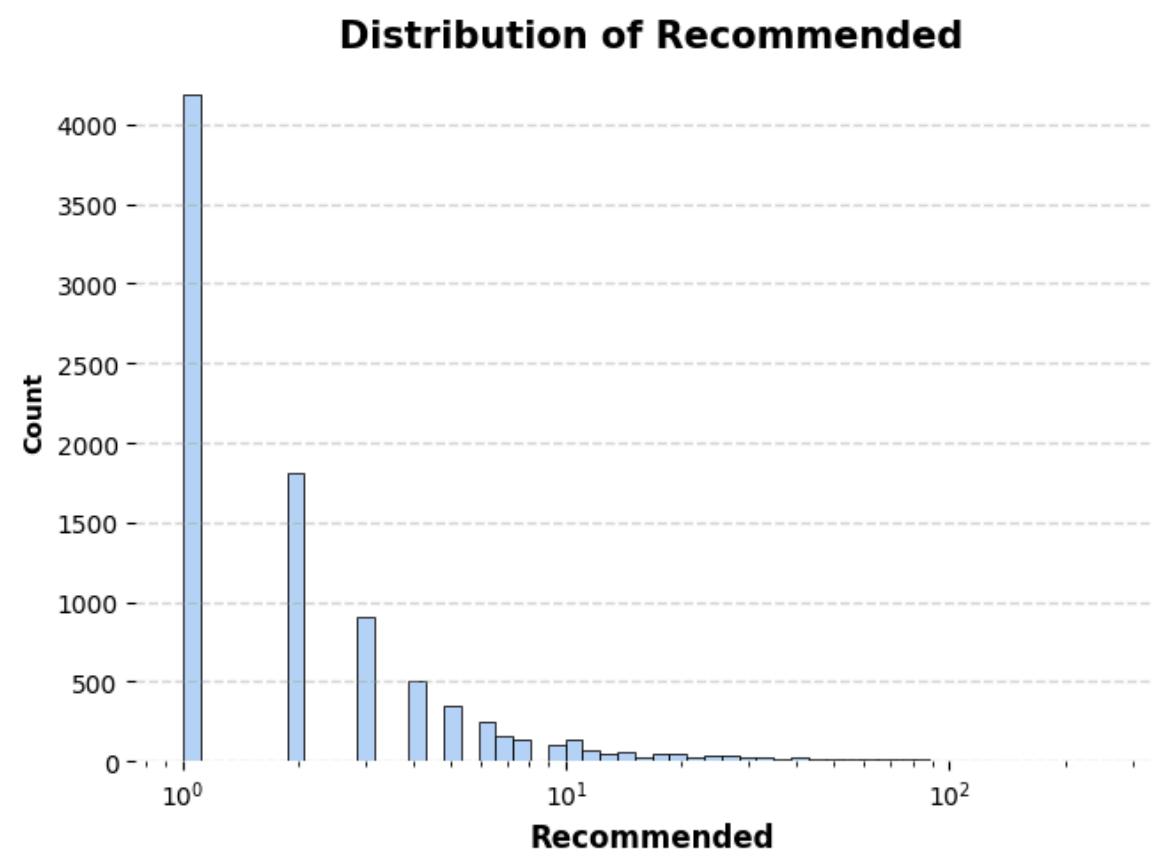
- Phân phối của các thuộc tính Numerical trong bộ dữ liệu này thường có **phân phối chuẩn hoặc lệch phai**.



KHAI PHÁ DỮ LIỆU

Nhóm thuộc tính Numerical

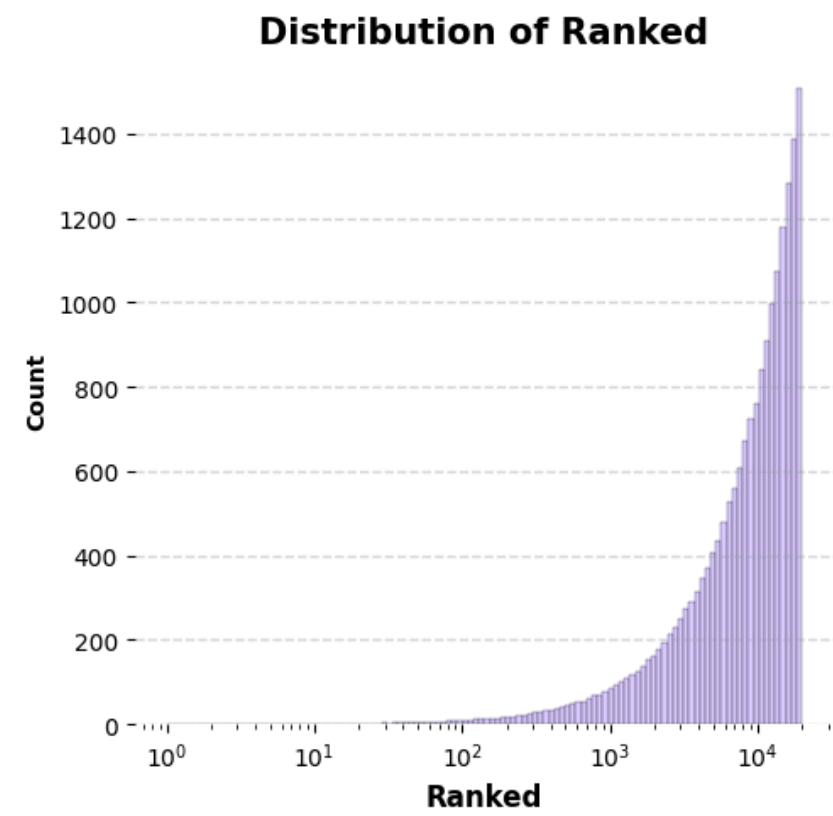
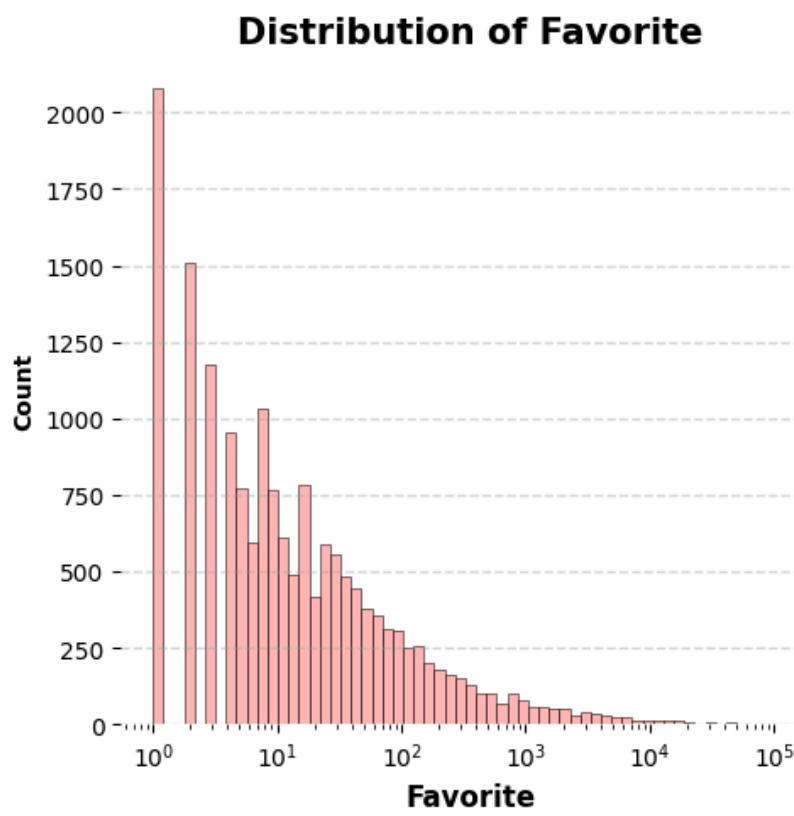
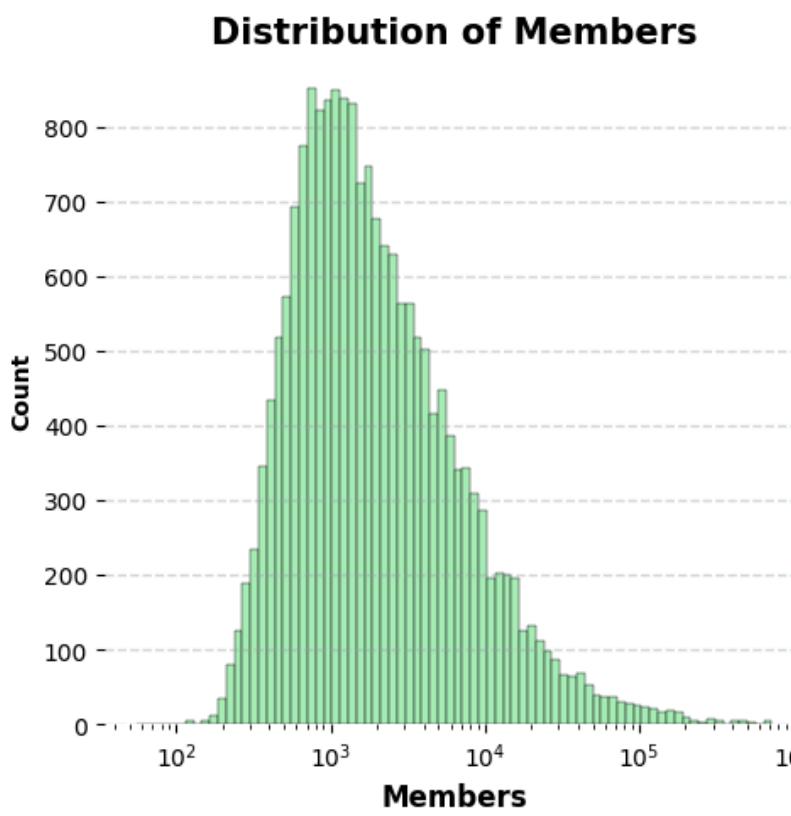
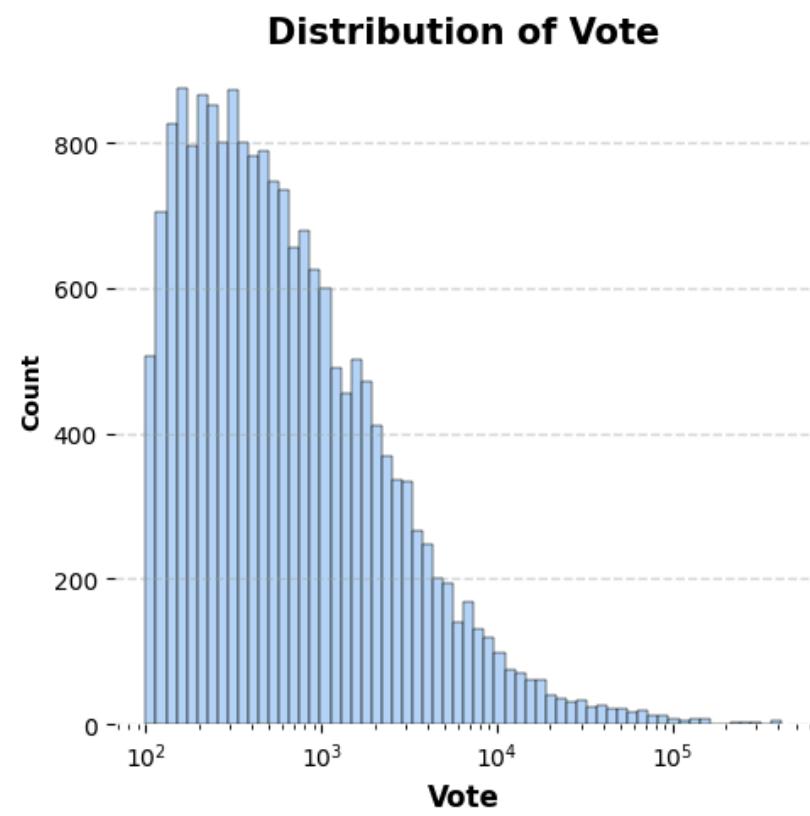
- Phân phối của các thuộc tính Numerical trong bộ dữ liệu này thường có **phân phối chuẩn hoặc lệch phai**.



KHAI PHÁ DỮ LIỆU

Nhóm thuộc tính Numerical

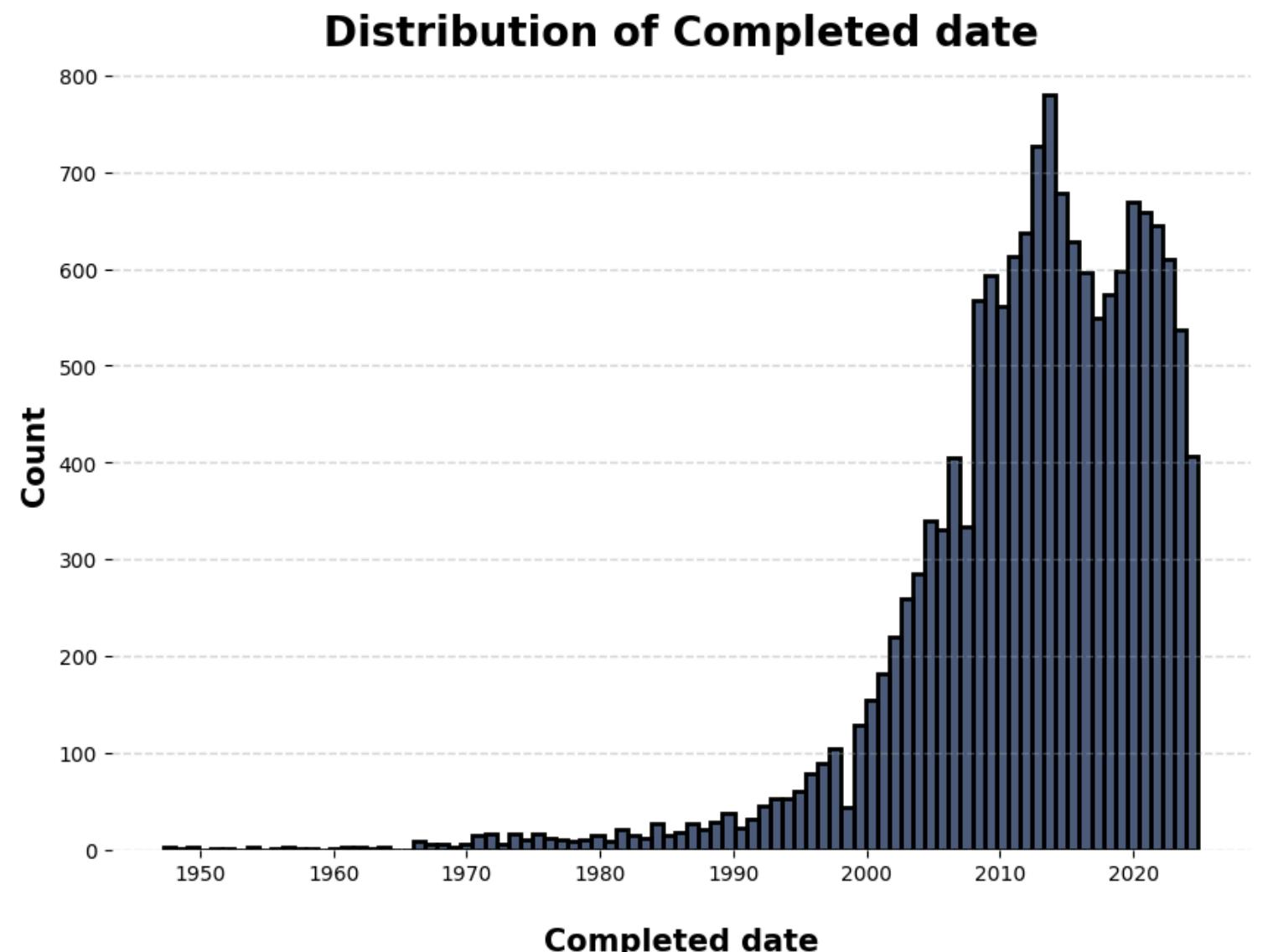
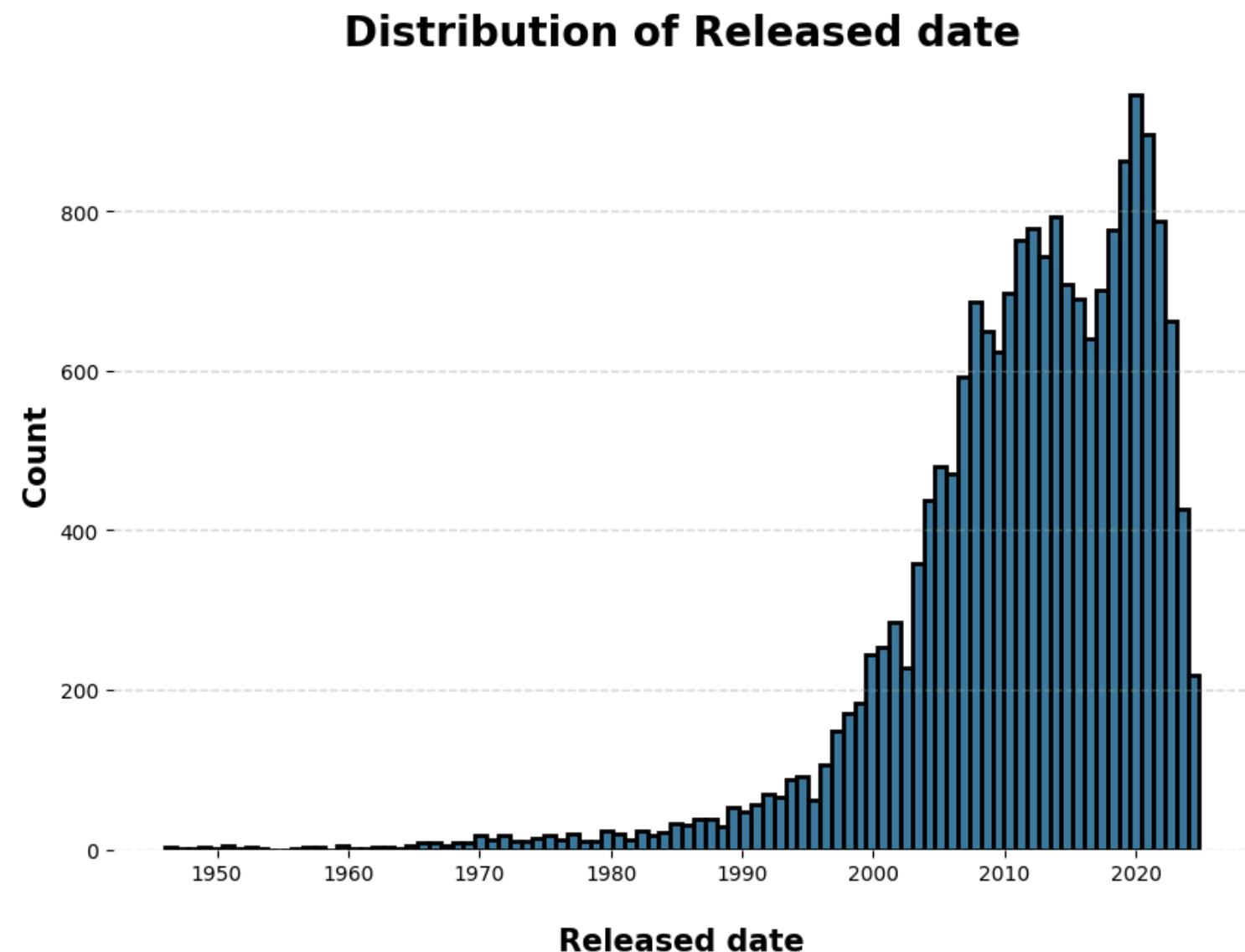
- Phân phối của các thuộc tính Numerical trong bộ dữ liệu này thường có **phân phối chuẩn hoặc lệch phai**.



KHAI PHÁ DỮ LIỆU

Nhóm thuộc tính Datetime

- Gồm 2 cột dữ liệu là **Released date** và **Completed date**. 2 biểu đồ của 2 cột dữ liệu này tương tự nhau, cho thấy rằng các bộ truyện thường có thời gian phát hành **tương đối ngắn**.

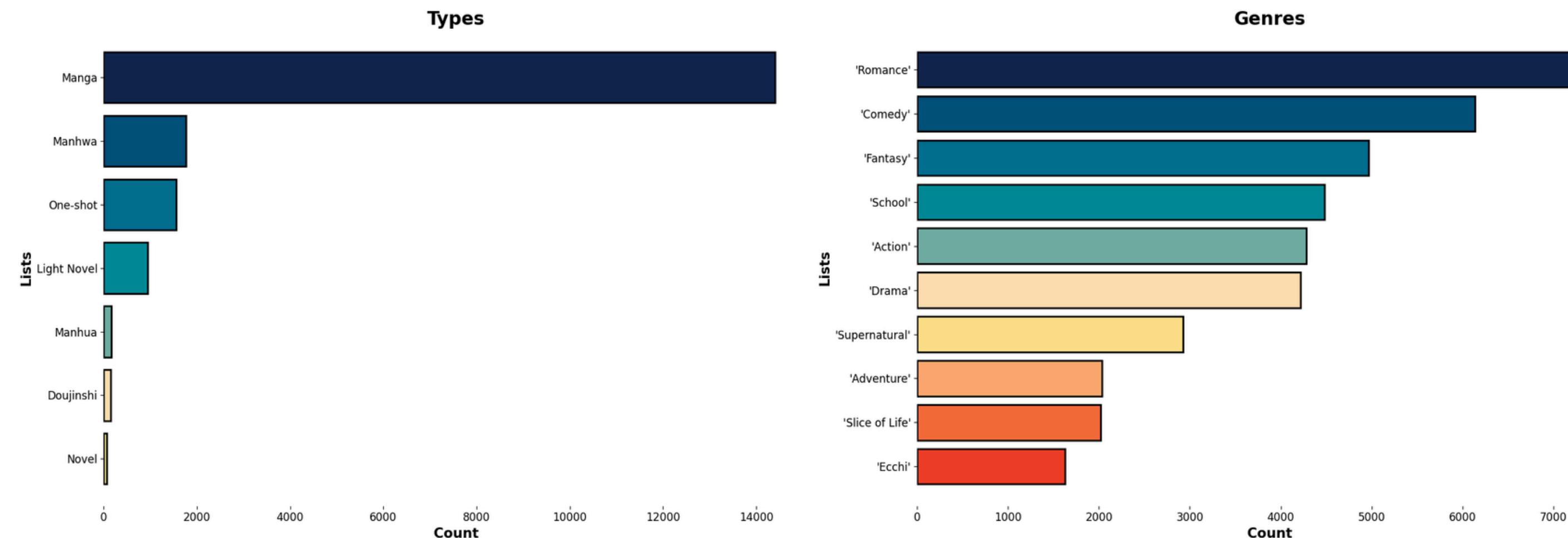


KHAI PHÁ DỮ LIỆU

Nhóm thuộc tính Categorical

- Gồm các thuộc tính: **Title, Types, Status, Genres, Demographic, Serialization, Author**
- Phân tích số lượng thuộc tính khác nhau và biểu diễn phân phối tần suất của các thuộc tính đó

Distribution of Types and Top 10 Genres

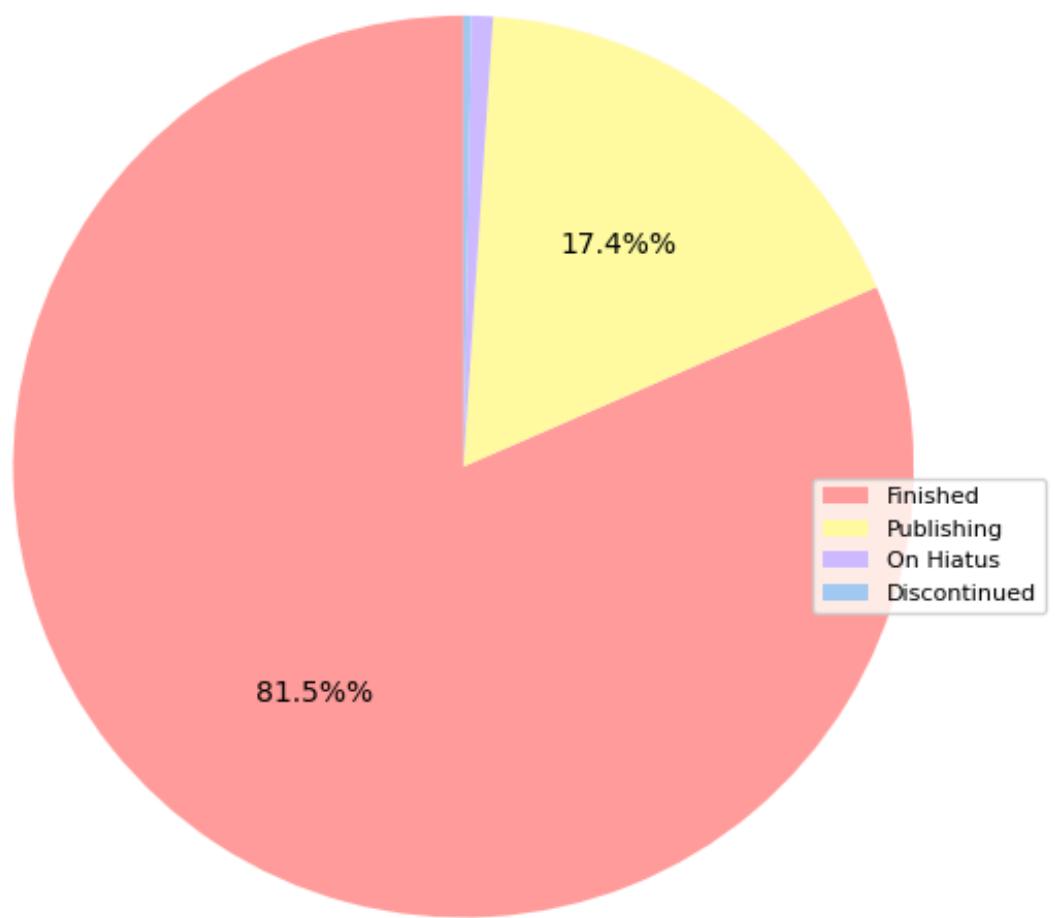


KHAI PHÁ DỮ LIỆU

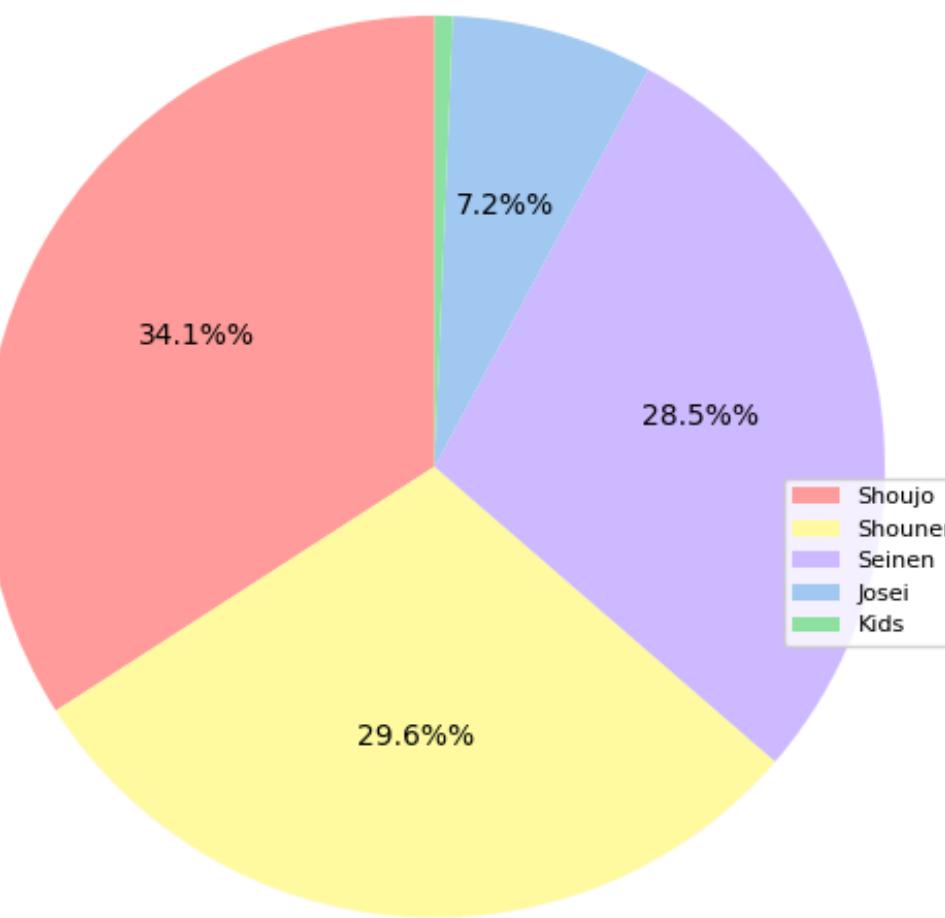
Nhóm thuộc tính Categorical

- Gồm các thuộc tính: **Title, Types, Status, Genres, Demographic, Serialization, Author**
- Phân tích số lượng thuộc tính khác nhau và biểu diễn phân phối tần suất của các thuộc tính đó

Distribution of Status



Distribution of Demographic

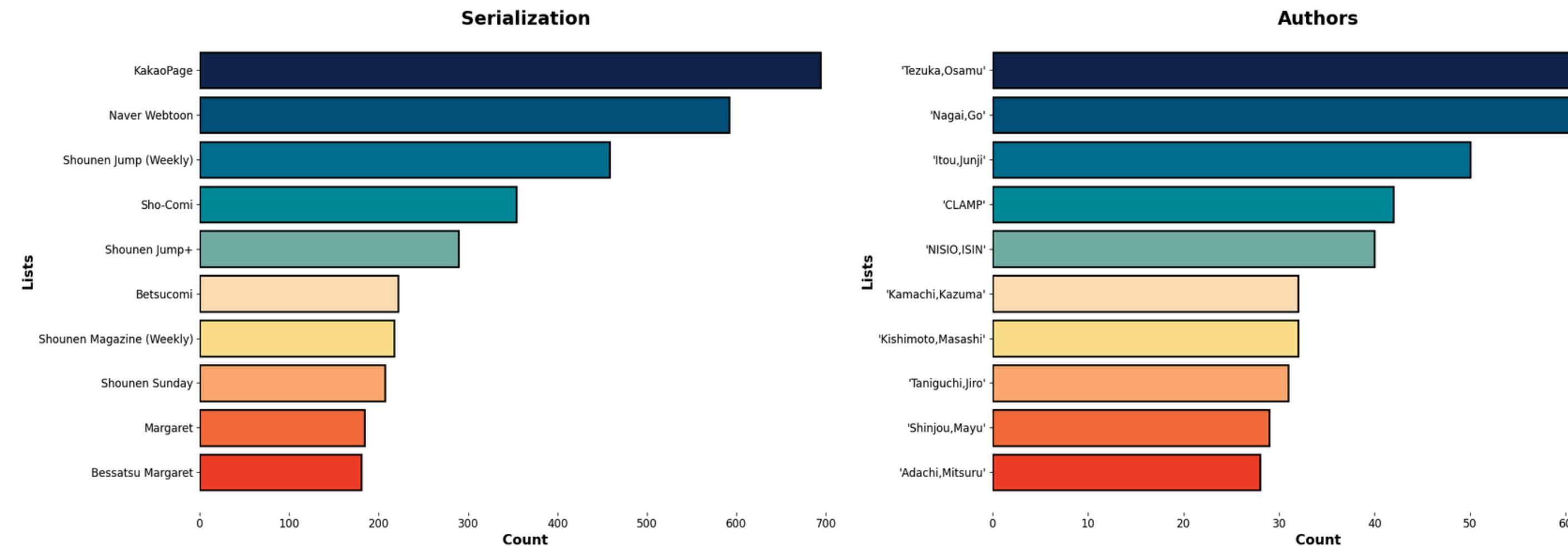


KHAI PHÁ DỮ LIỆU

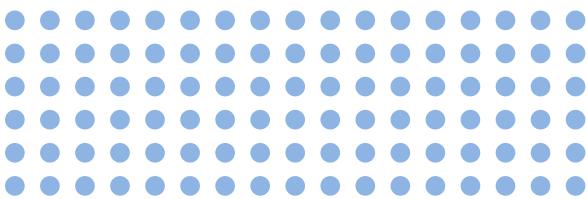
Nhóm thuộc tính Categorical

- Gồm các thuộc tính: **Title, Types, Status, Genres, Demographic, Serialization, Author**
- Phân tích số lượng thuộc tính khác nhau và biểu diễn phân phối tần suất của các thuộc tính đó

Top 10 Serialization and Authors



ĐẶT CÂU HỎI CÓ Ý Nghĩa và trả lời





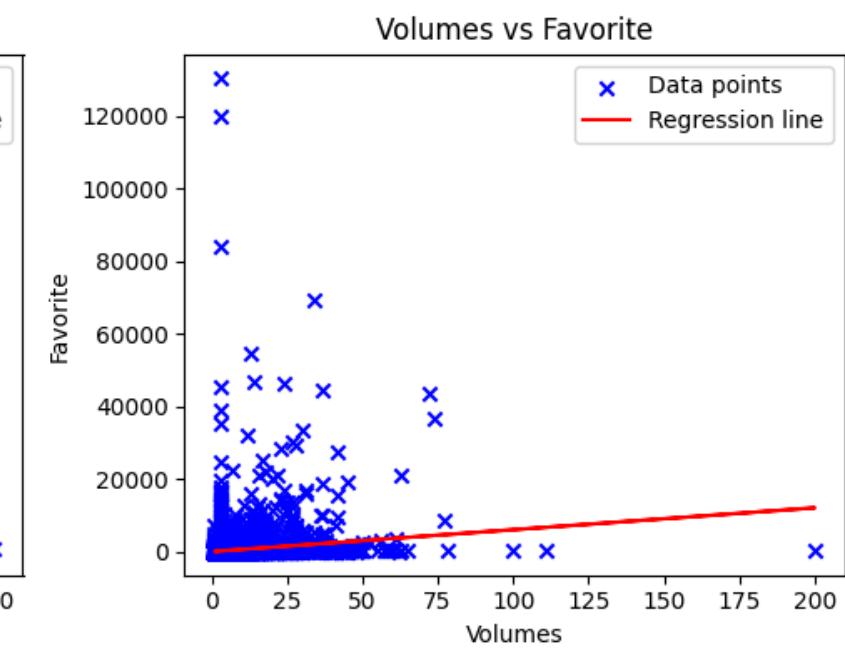
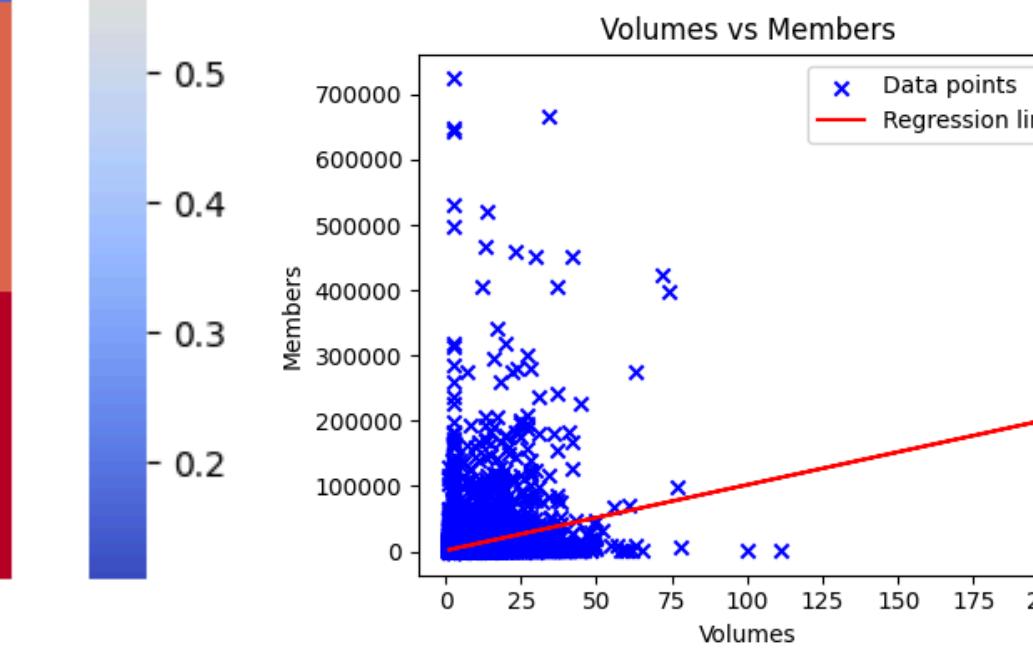
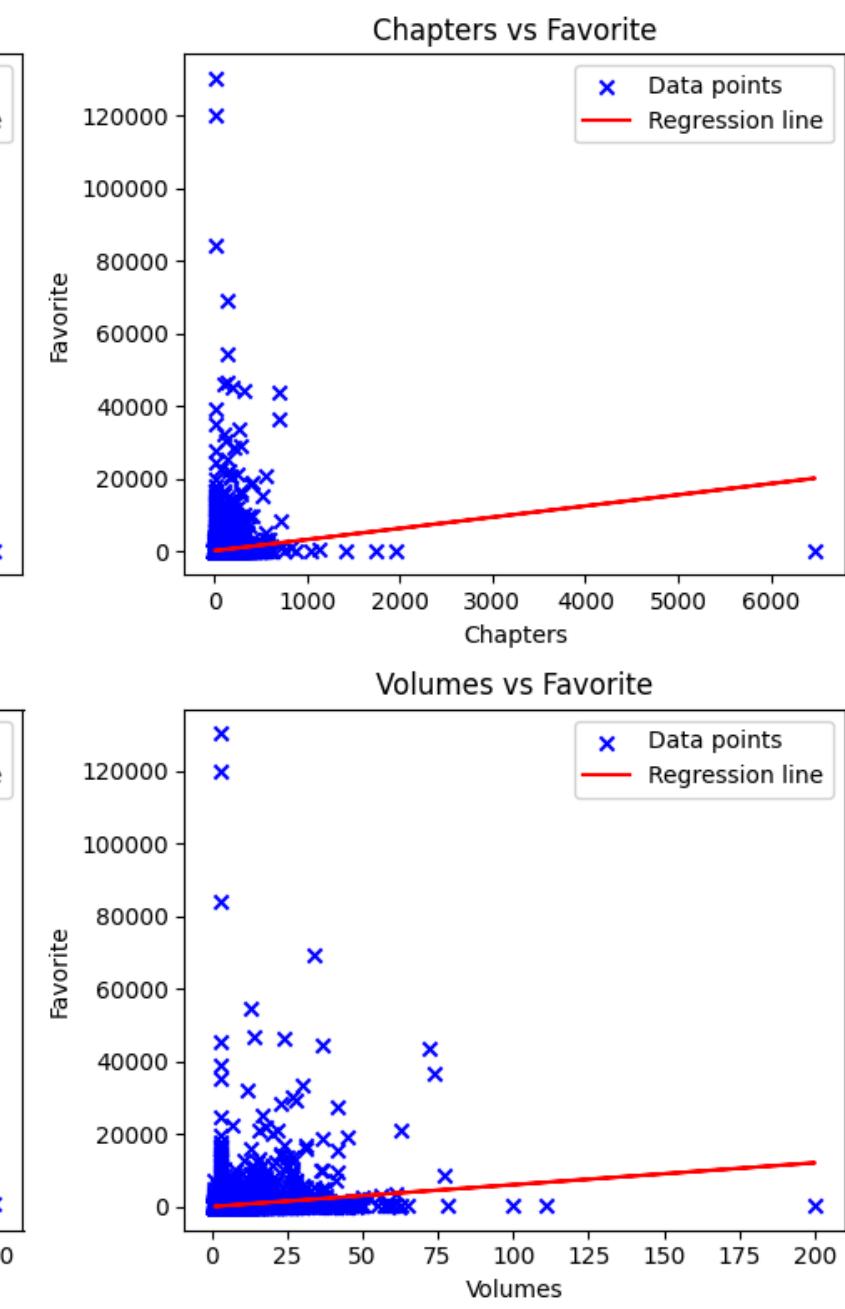
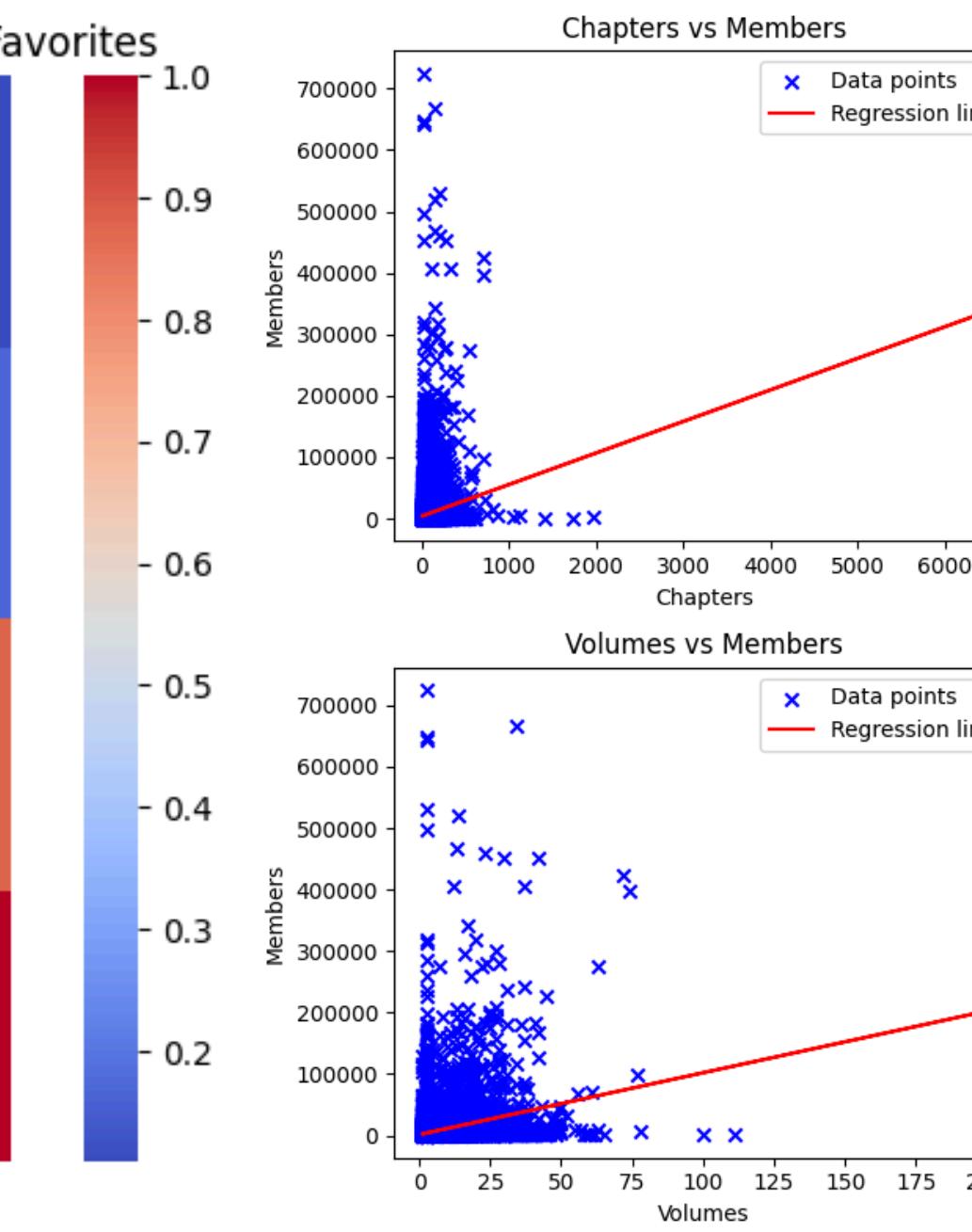
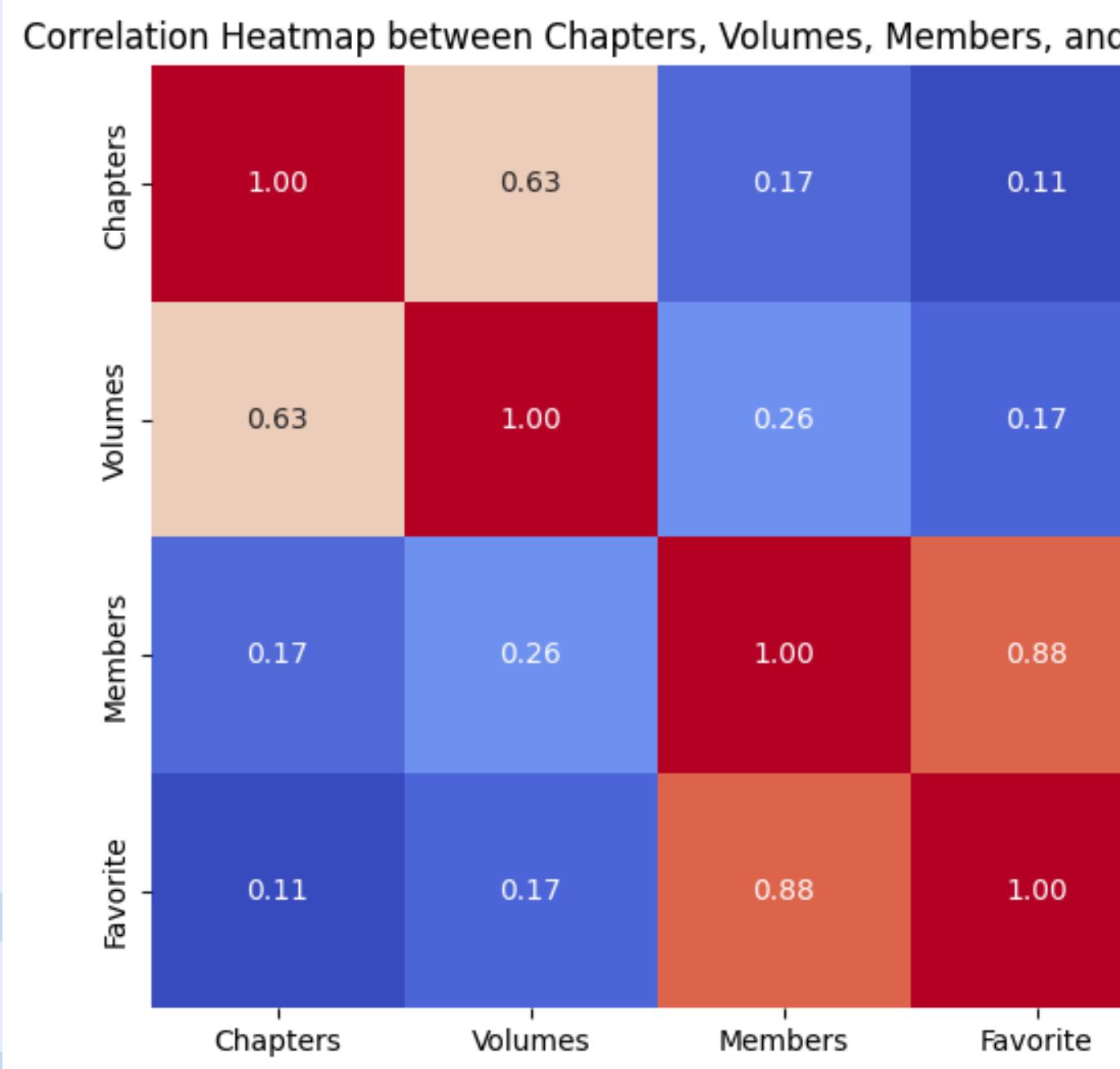
Câu 1: Mối quan hệ giữa số lượng chương (Chapters) và tập (Columns) với số lượng người theo dõi bộ truyện (Members) và mức độ tương tác của họ (Favourite)?

MỤC TIÊU

- Phân tích mối quan hệ giữa số lượng chương, số lượng tập, số lượng người đọc, và số lượng yêu thích sẽ giúp ta hiểu rõ hơn về cách các yếu tố này ảnh hưởng đến sự phổ biến và mức độ gắn kết của bộ truyện đối với độc giả.
- Những phân tích này rất hữu ích cho các nhà xuất bản truyện, các tác giả, và những ai quan tâm đến việc phát triển nội dung truyện một cách hiệu quả và thu hút người đọc lâu dài.



Câu 1: Mối quan hệ giữa số lượng chương (Chapters) và tập (Volumes) với số lượng người theo dõi bộ truyện (Members) và mức độ tương tác của họ (Favorite)?





Câu 1: Mối quan hệ giữa số lượng chương (Chapters) và tập (Volumes) với số lượng người theo dõi bộ truyện (Members) và mức độ tương tác của họ (Favourite)?

NHẬN XÉT

- Ma trận tương quan
 - Members và Favorites: Tương quan mạnh (0.88), truyện có nhiều thành viên thường nhận nhiều lượt yêu thích.
 - Chapters và Volumes: Tương quan mạnh (0.63), truyện nhiều chương thường có nhiều tập.
 - Chapters/Volumes và Members/Favorites: Tương quan dương yếu, truyện dài hơn không đảm bảo nhiều thành viên hay lượt yêu thích.
- Xu hướng:
 - Nội dung dài hơn thu hút thêm khán giả (Members), nhưng ít ảnh hưởng đến lượt yêu thích (Favorites).
 - Sự phân tán lớn cho thấy các yếu tố khác (thể loại, tác giả, cốt truyện) quyết định mức độ tương tác và quy mô khán giả thay vì độ dài nội dung.



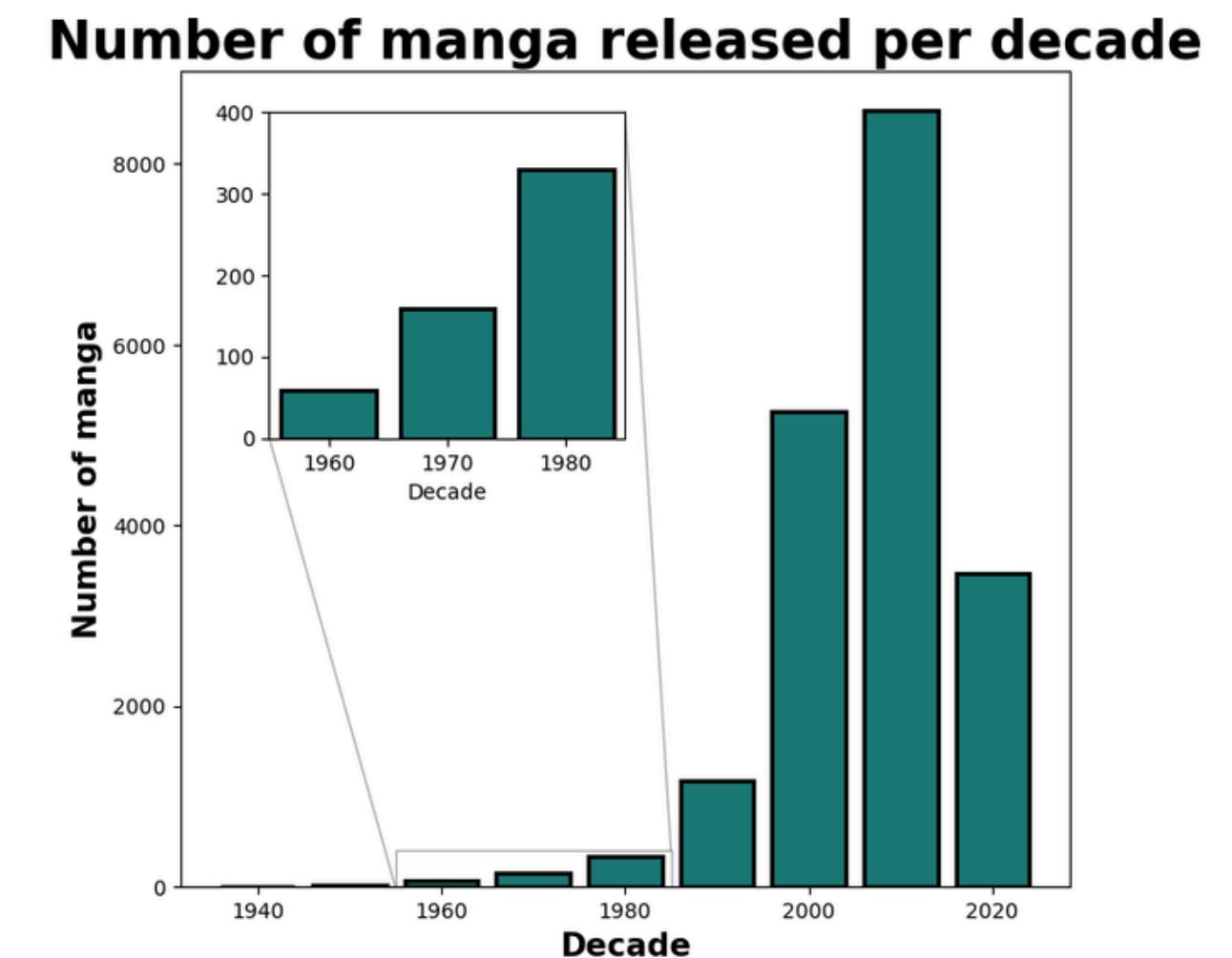
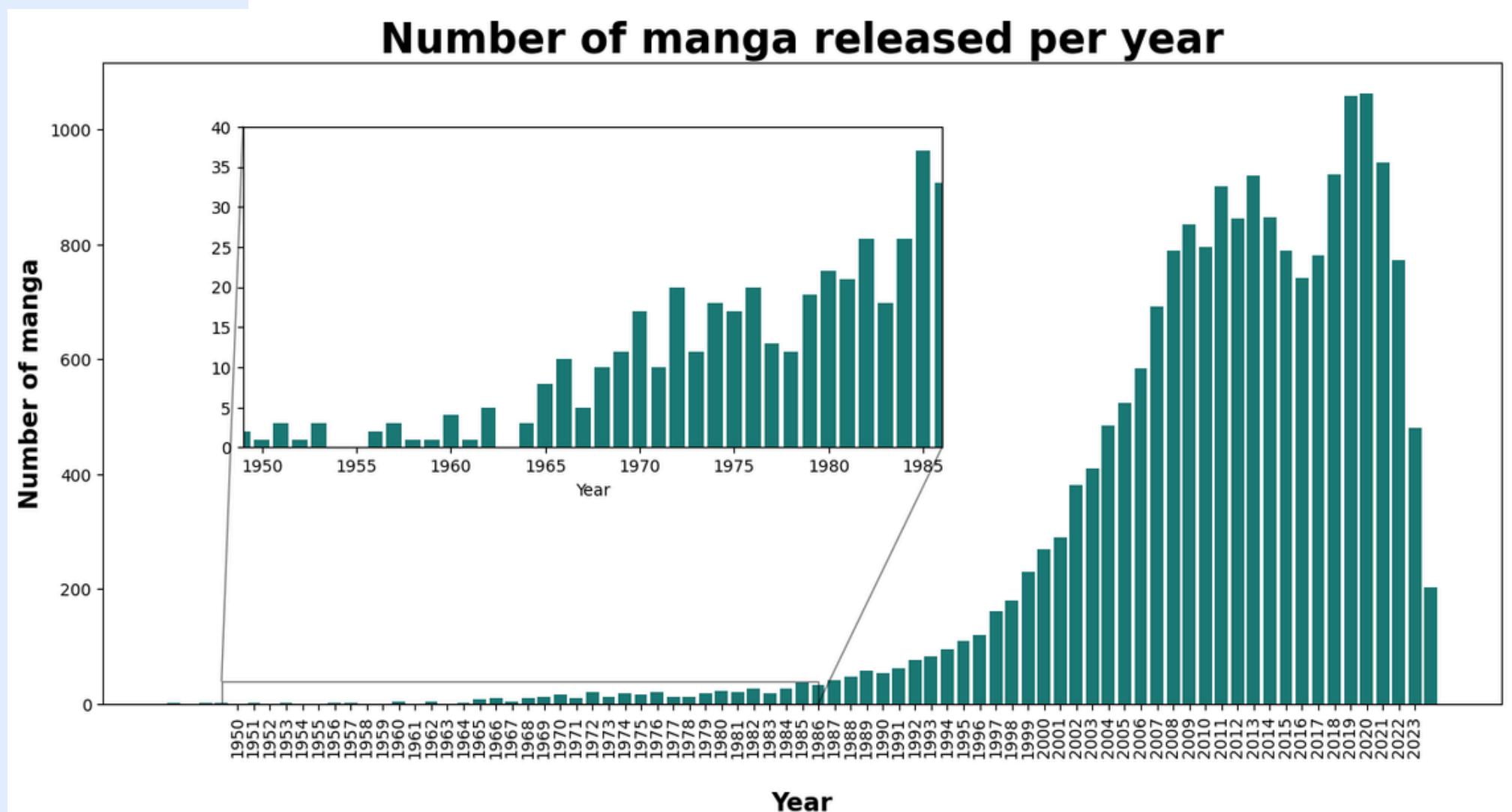
Câu 2: Số lượng truyện được phát hành đã thay đổi như thế nào theo thời gian từ quá khứ đến hiện tại?

MỤC TIÊU

- **Hiểu về sự phát triển lịch sử của truyện:** Theo dõi sự phát triển của ngành công nghiệp manga từ giai đoạn đầu cho đến hiện tại.
- **Xác định xu hướng phát hành truyện :** Quan sát cách số lượng truyện được phát hành thay đổi theo thời gian và nhận diện bất kỳ sự thay đổi hoặc mô hình đáng chú ý nào.



Câu 2: Số lượng manga được phát hành đã thay đổi như thế nào theo thời gian từ quá khứ đến hiện tại?





Câu 2: Số lượng truyện được phát hành đã thay đổi như thế nào theo thời gian từ quá khứ đến hiện tại?

NHẬN XÉT

Hai biểu đồ cung cấp cái nhìn về sự phát triển của ngành công nghiệp manga Nhật Bản qua các giai đoạn:

- **1950-1970: Giai đoạn đầu**
 - Truyện phát hành hạn chế, cao nhất 17 tác phẩm (1969).
- **1970-1985: Giai đoạn phát triển ban đầu**
 - Số lượng tăng dần (10-35 tác phẩm/năm). Truyện Nhật Bản bắt đầu vươn ra toàn cầu với các tác phẩm như Dragon Ball, Tsubasa.
- **1985-2010: Giai đoạn phát triển mạnh mẽ**
 - Số lượng bùng nổ (hơn 800 tác phẩm/năm). Các siêu phẩm như Naruto, One Piece, Bleach khẳng định vị thế toàn cầu.
- **2010 đến nay: Đỉnh cao và giảm sút**
 - Đạt đỉnh trên 1000 tác phẩm (2020). Từ 2020, số lượng giảm đáng kể, có thể do đại dịch Covid-19.



Câu 3: Số lượng truyện được phát hành thay đổi như thế nào giữa các mùa trong năm? Có xu hướng rõ ràng nào trong phân bố các lần phát hành manga theo mùa không?

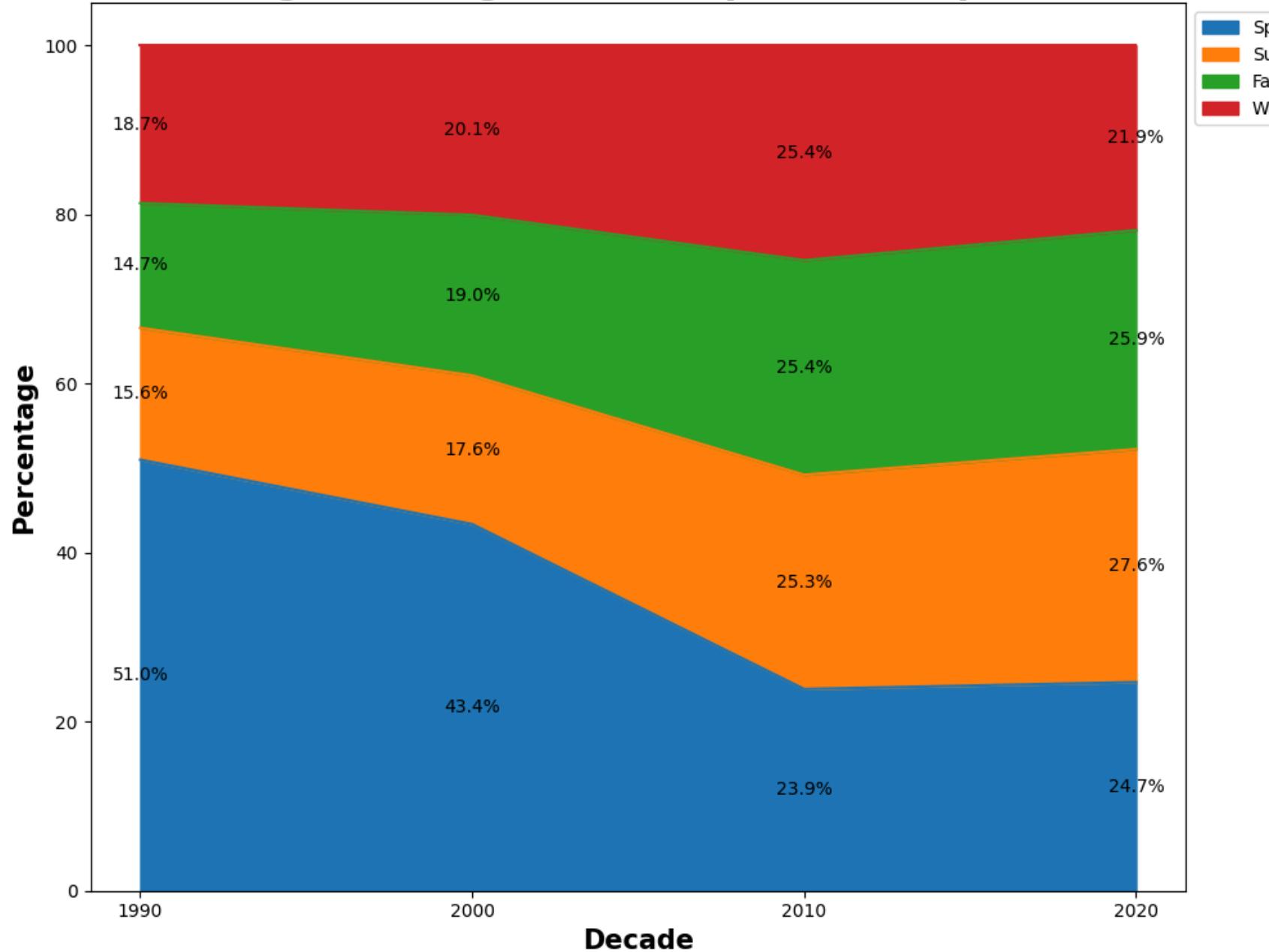
MỤC TIÊU

- Khám phá các chu kỳ phát hành truyện, làm sáng tỏ các mô hình trong việc phát hành truyện.**
- Cung cấp thông tin hữu ích cho các nhà xuất bản, người hâm mộ và các chuyên gia trong ngành để lập kế hoạch, chiến lược tiếp thị và ra quyết định một cách hiệu quả**

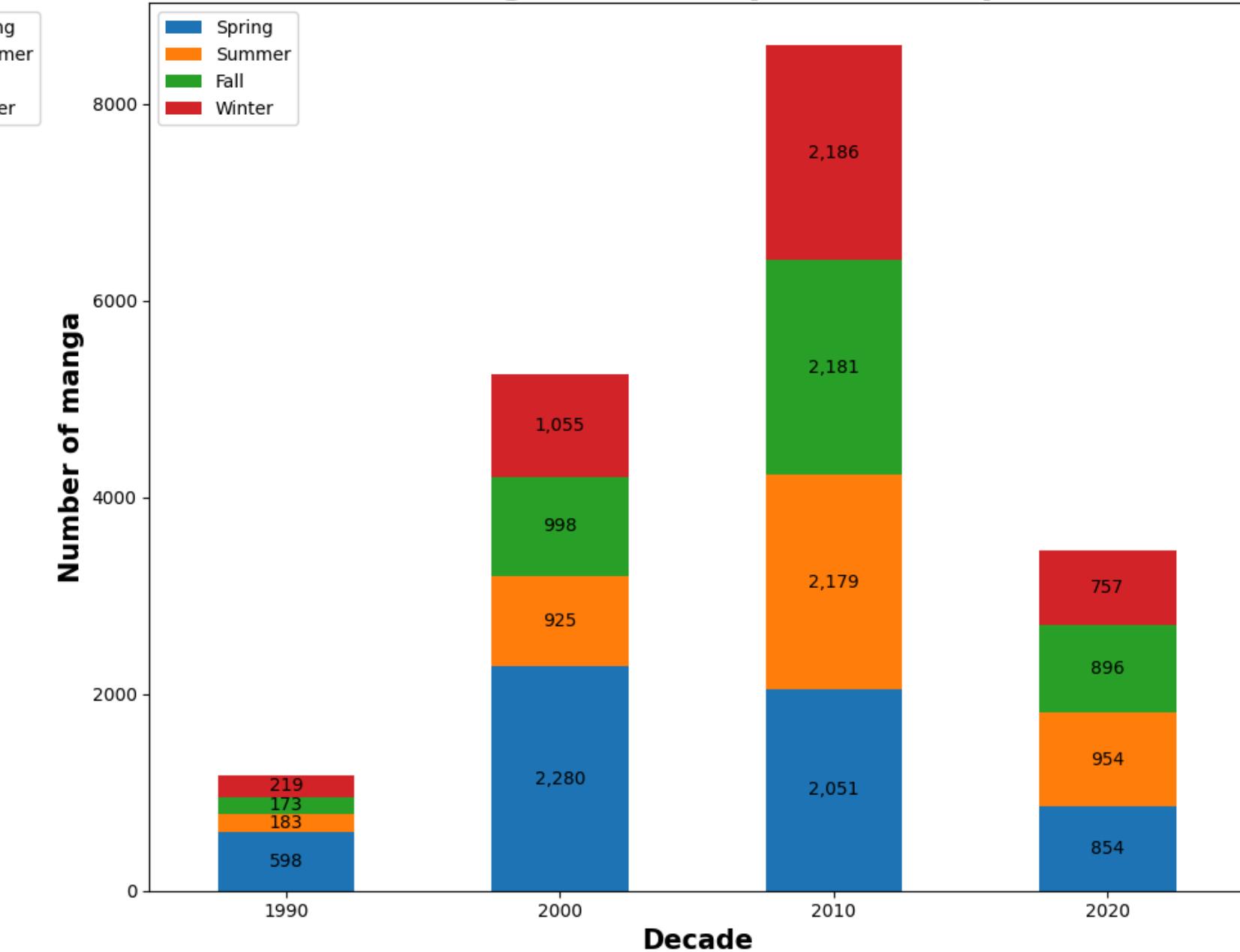


Câu 3: Số lượng truyện được phát hành thay đổi như thế nào giữa các mùa trong năm? Có xu hướng rõ ràng nào trong phân bố các lần phát hành manga theo mùa không?

Percentage of manga released per season per decade



Number of manga released per season per decade





**Câu 3: Số lượng truyện được phát hành thay đổi như thế nào giữa các mùa trong năm?
Có xu hướng rõ ràng nào trong phân bố các lần phát hành truyện theo mùa không?**

NHẬN XÉT

Biểu đồ 1: Percentage of Manga Released Per Season Per Decade

- **Ý nghĩa:** Thể hiện tỷ lệ manga phát hành theo mùa từ 1990 đến 2020.
- **Xu hướng:**
 - **Mùa Xuân giảm:** Từ 51% (1990) xuống 24.7% (2020).
 - **Mùa Hè tăng:** Từ 15.6% (1990) lên 27.6% (2020).
 - **Mùa Thu & Đông tăng đều:** Lần lượt đạt 25.9% và 21.9% (2020).
 - **Phân bổ đồng đều hơn vào năm 2020, không còn sự chênh lệch lớn giữa các mùa.**

Biểu đồ 2: Number of Manga Released Per Season Per Decade

- **Ý nghĩa:** Thể hiện số lượng manga phát hành thực tế theo mùa qua từng thập kỷ.
- **Xu hướng:**
 - **Tăng mạnh từ 1,173 (1990) lên 8,466 manga (2010).**
 - **Mùa Xuân giảm số lượng tuyệt đối, không còn vượt trội như trước.**
 - **Mùa Hè dẫn đầu từ 2000 và giữ vị trí cao nhất vào 2020.**



Câu 4: Số lượng truyện của từng thể loại phân bố như thế nào?

Các thể loại truyện hàng đầu được xác định bởi điểm tổng hợp bao gồm: Số lượng yêu thích trung bình (Favorite), số lượng truyện trong mỗi thể loại, điểm trung bình (Score) của chúng và mức độ phổ biến (Popularity) của chúng là gì?

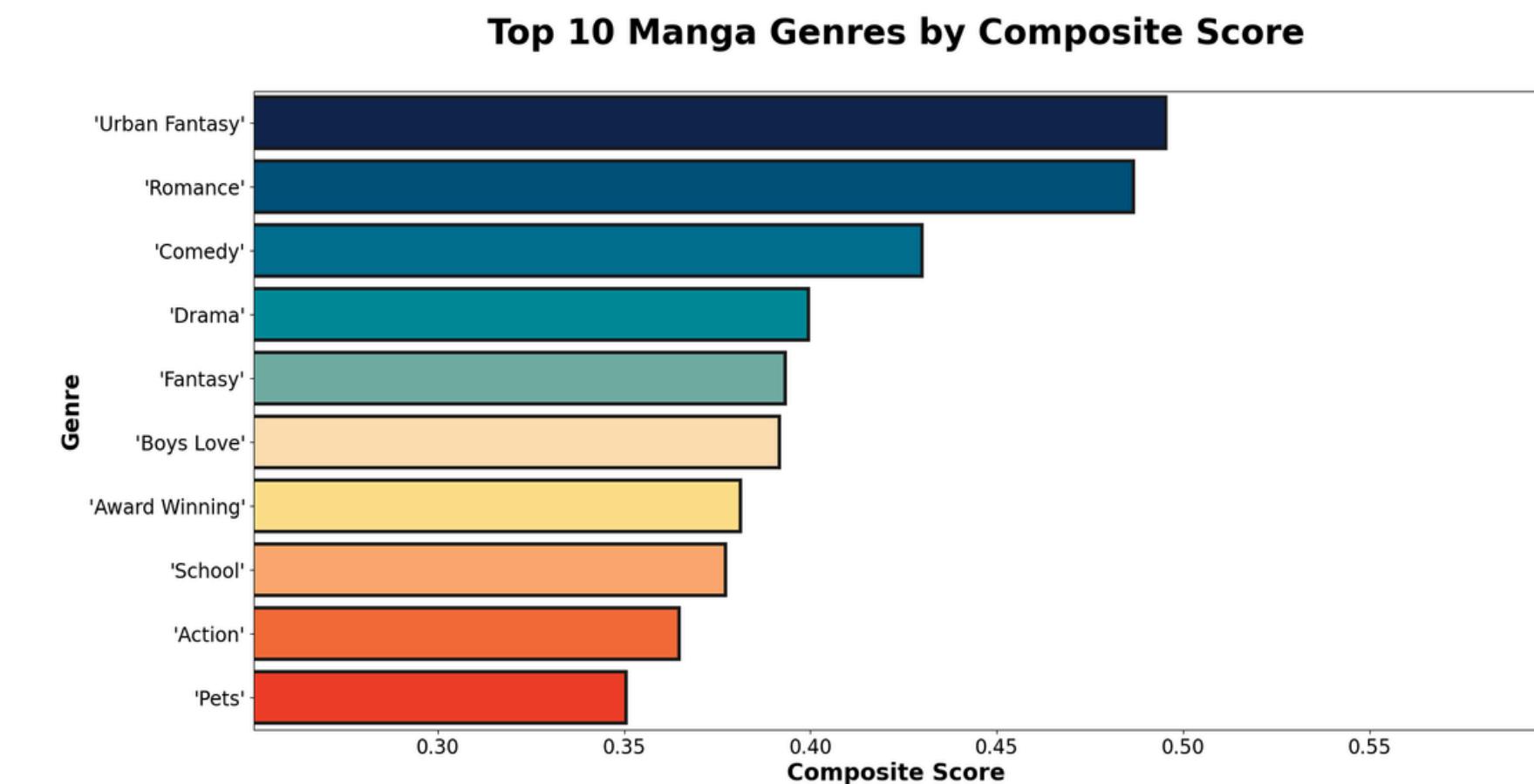
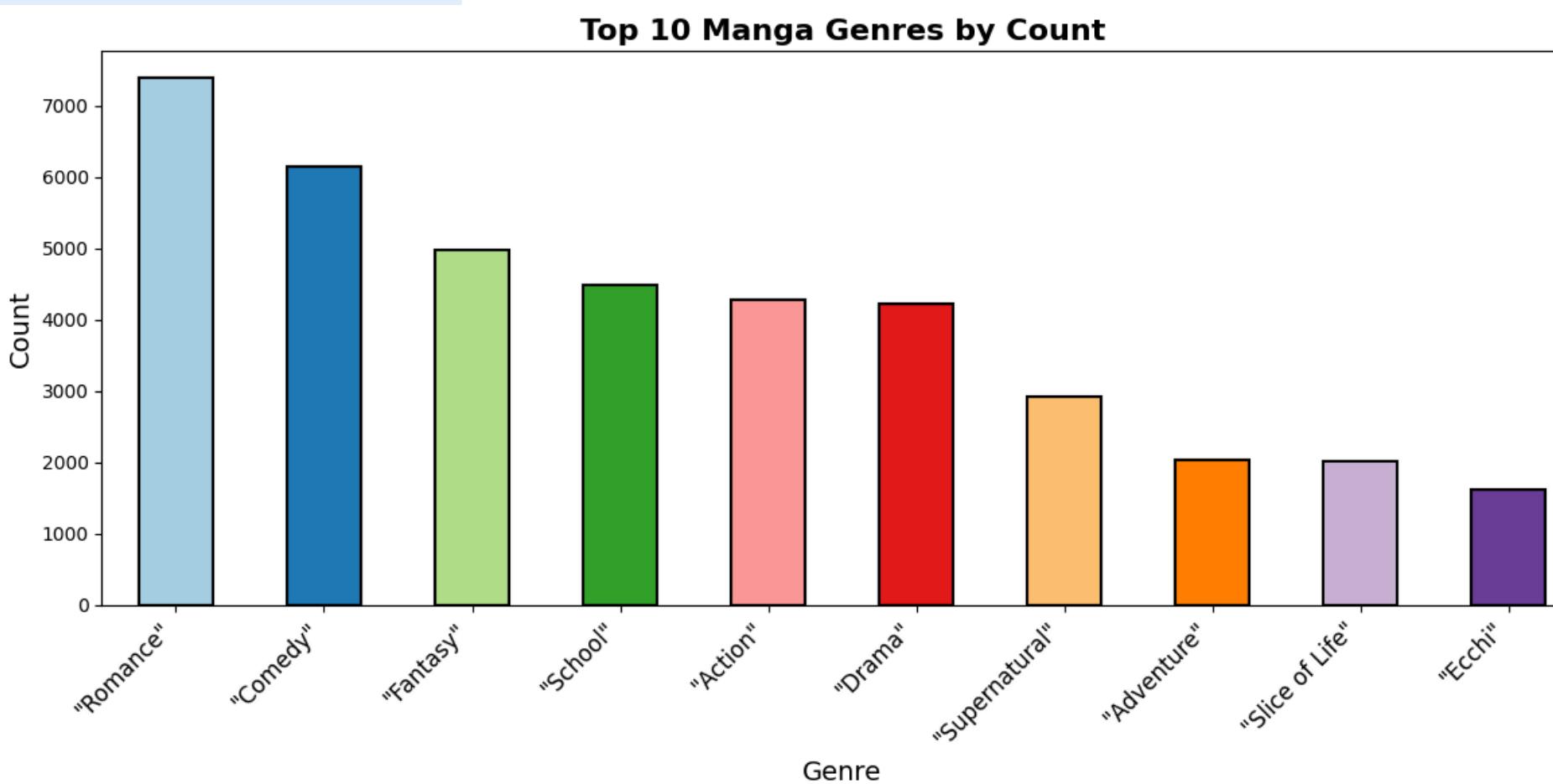
MỤC TIÊU

- Xác định và xếp hạng các thể loại manga dựa trên một điểm tổng hợp (composite score), xem xét các yếu tố như:
 - Số lượt yêu thích trung bình (average favorite count).
 - Số lượng truyện trong mỗi thể loại (number of manga in each genre).
 - Điểm trung bình của manga trong thể loại đó (average score).
 - Mức độ phổ biến (popularity).
- Cung cấp một cái nhìn tổng quan toàn diện về các thể loại truyện, bằng cách kết hợp nhiều chỉ số. Phân tích sẽ mang lại những hiểu biết sâu sắc về các thể loại nổi bật vượt trội qua nhiều tiêu chí khác nhau.



Câu 4: Số lượng truyện của từng thể loại phân bố như thế nào?

Các thể loại truyện hàng đầu được xác định bởi điểm tổng hợp bao gồm: Số lượng yêu thích trung bình (Favorite), số lượng truyện trong mỗi thể loại, điểm trung bình (Score) của chúng và mức độ phổ biến (Popularity) của chúng là gì?





Câu 4: Số lượng truyện của từng thể loại phân bố như thế nào?

Các thể loại truyện hàng đầu được xác định bởi điểm tổng hợp bao gồm: Số lượng yêu thích trung bình (**Favorite**), số lượng truyện trong mỗi thể loại, điểm trung bình (**Score**) của chúng và mức độ phổ biến (**Popularity**) của chúng là gì?

NHẬN XÉT

Biểu đồ 1: Top 10 Manga Genres by Count

- Ý nghĩa: Biểu đồ này sắp xếp 10 thể loại truyện có số lượng truyện nhiều nhất.
- Nhận xét: Romance dẫn đầu với hơn 7000 truyện, cho thấy mức độ phổ biến cao. Comedy, Fantasy, và School theo sau với số lượng đáng kể. Slice of Life và Ecchi xếp cuối top 10 nhưng vẫn khá phổ biến.

Biểu đồ 2: Top 10 Manga Genres by Composite Score

- Ý nghĩa: Biểu đồ này xếp hạng các thể loại dựa trên một composite score được tính toán từ nhiều yếu tố như số lượng yêu thích, điểm trung bình, số lượng truyện, và mức độ phổ biến.
- Nhận xét: Urban Fantasy có composite score cao nhất, cho thấy sự yêu thích và chất lượng cao. Romance và Comedy xếp thứ hai và ba. Các thể loại ít phổ biến như Award Winning, Boys Love và Pets vẫn lọt top nhờ điểm số vượt trội.



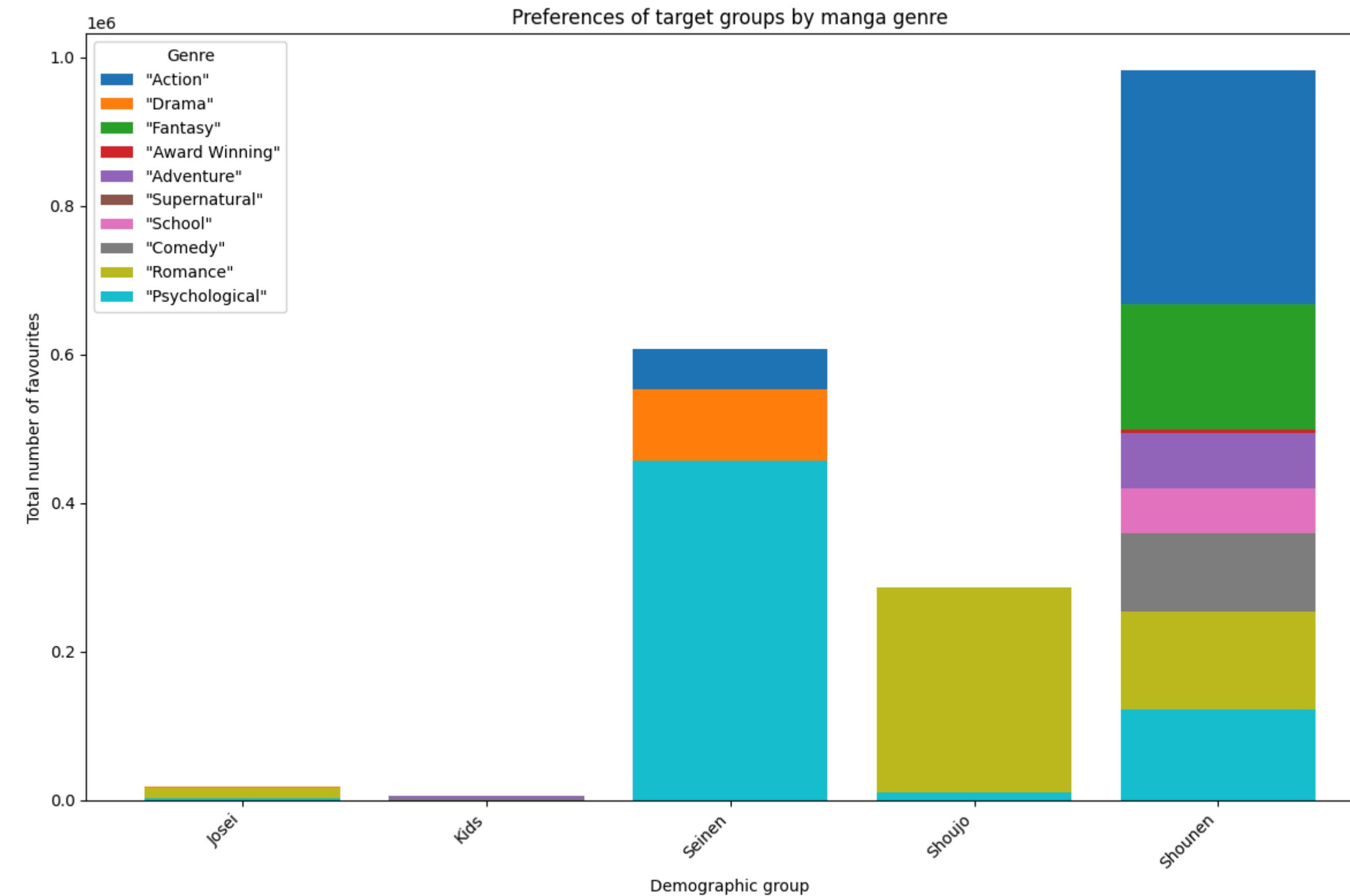
Câu 5: Đối với từng nhóm đối tượng, mức độ yêu thích của người đọc dành cho các tác phẩm thuộc nhóm đối tượng đó như thế nào? Sự phân bố các thể loại trong mỗi nhóm đối tượng đó ra sao?

MỤC TIÊU

- Phân tích mức độ yêu thích của độc giả dành cho các tác phẩm của từng đối tượng và phân bố thể loại trong từng nhóm.
- Từ đó giúp các nhà xuất bản truyện và các tác giả lập ra các chiến lược phát triển truyện nhắm đến từng nhóm đối tượng riêng.



Câu 5: Đối với từng nhóm đối tượng, mức độ yêu thích của người đọc dành cho các tác phẩm thuộc nhóm đối tượng đó như thế nào? Sự phân bố các thể loại trong mỗi nhóm đối tượng đó ra sao?





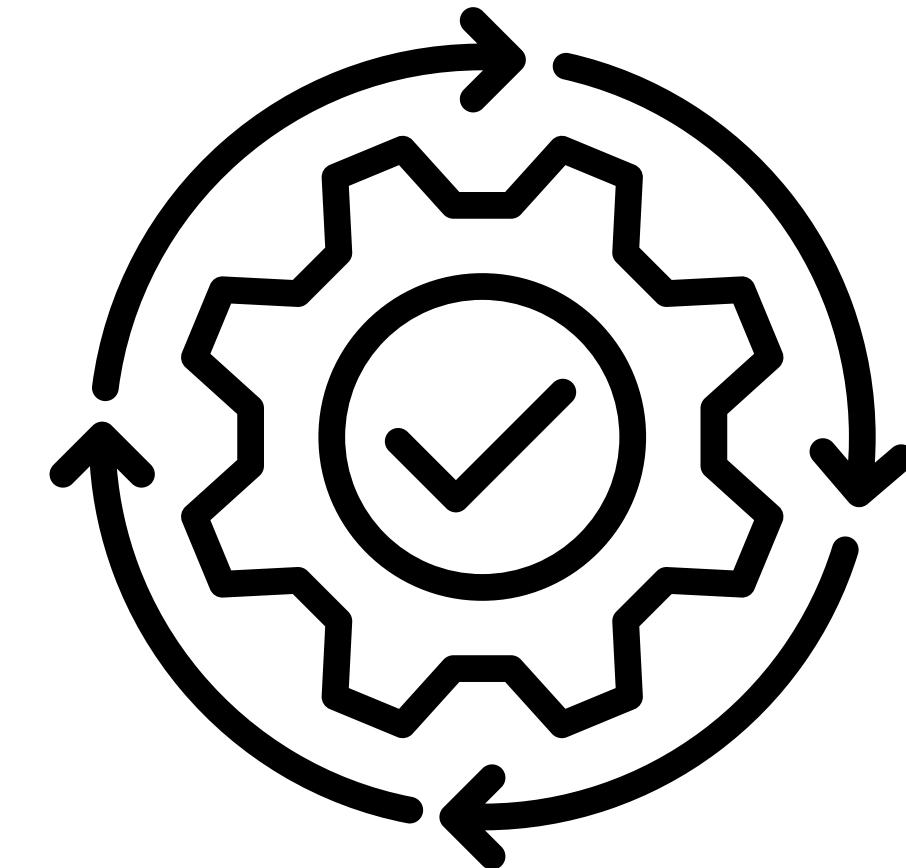
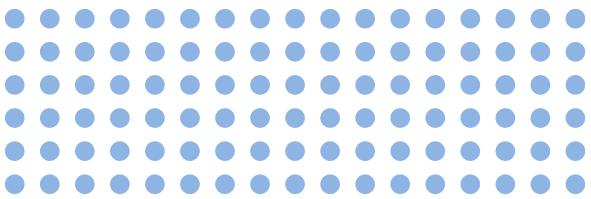
Câu 5: Đối với từng nhóm đối tượng, mức độ yêu thích của người đọc dành cho các tác phẩm thuộc nhóm đối tượng đó như thế nào? Sự phân bố các thể loại trong mỗi nhóm đối tượng đó ra sao?

NHẬN XÉT

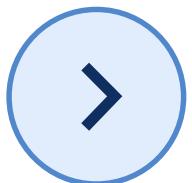
- **Shounen (thiếu niên nam)** là nhóm thể loại có các tác phẩm có lượt yêu thích cao nhất, đặc biệt với các thể loại như Action, Fantasy, Psychological, Romance và Adventure. Điều này cho thấy các tác phẩm thuộc Shounen có số lượng độc giả lớn nhất và đa dạng thể loại.
- Các bộ truyện cho Seinen (nam trưởng thành) có lượt yêu thích đáng kể, tập trung vào Drama và Psychological, phù hợp với nội dung phức tạp, tâm lý.
- Các bộ truyện Shoujo (thiếu niên nữ) chủ yếu là thể loại Romance và có số lượt yêu thích thấp hơn rõ rệt so với các tác phẩm Shounen và Seinen.
- Các tác phẩm cho Josei (nữ trưởng thành) và Kids (trẻ em) có lượt yêu thích rất thấp, ít thể loại nổi bật.

Quy trình xây dựng và đánh giá mô hình dữ liệu

DỰ ĐOÁN ĐIỂM SỐ DỰA TRÊN DỮ LIỆU MẪU



CÁC BƯỚC THỰC HIỆN



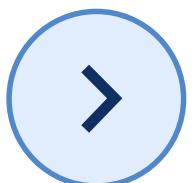
Phát biểu vấn đề



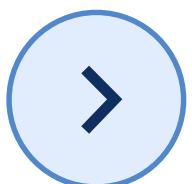
Chuẩn bị dữ liệu



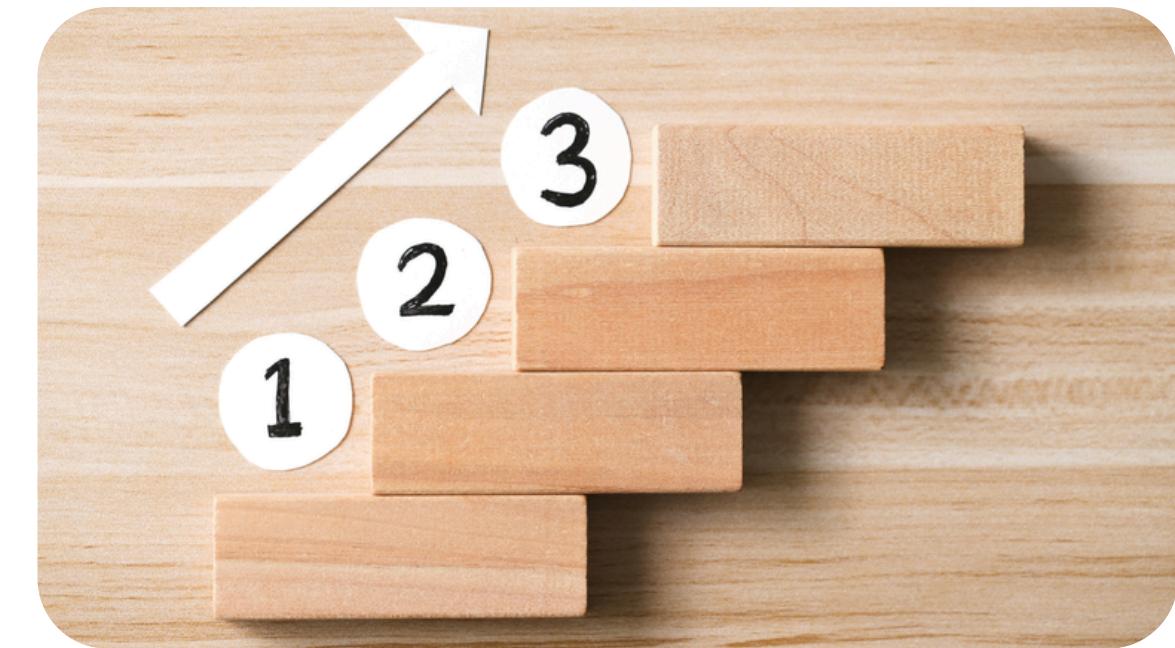
Xây dựng và huấn luyện mô hình



Đánh giá và cải thiện



Kết luận



PHÁT BIỂU VĂN ĐỀ



Tầm quan trọng



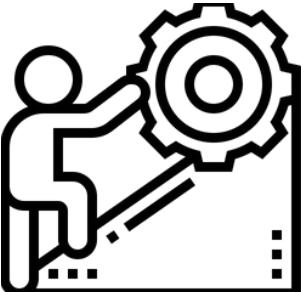
- **Điểm số (Score)** là chỉ số quan trọng để đánh giá chất lượng và mức độ yêu thích của cộng đồng đối với một bộ manga
- Giúp người dùng chọn manga phù hợp
- Hỗ trợ nhà xuất bản tối ưu chiến lược phát hành

Mục tiêu



- Xác định những yếu tố quan trọng ảnh hưởng đến điểm số của manga
- Tìm kiếm các tác phẩm có chất lượng cao
- Giúp dự đoán các tác phẩm có tiềm năng thành công

Thách thức



- Việc dự đoán phụ thuộc vào nhiều yếu tố, đặc trưng khác nhau, có mối quan hệ phức, đòi hỏi các phương pháp mô hình hóa mạnh mẽ

CÁC BƯỚC CHUẨN BỊ DỮ LIỆU



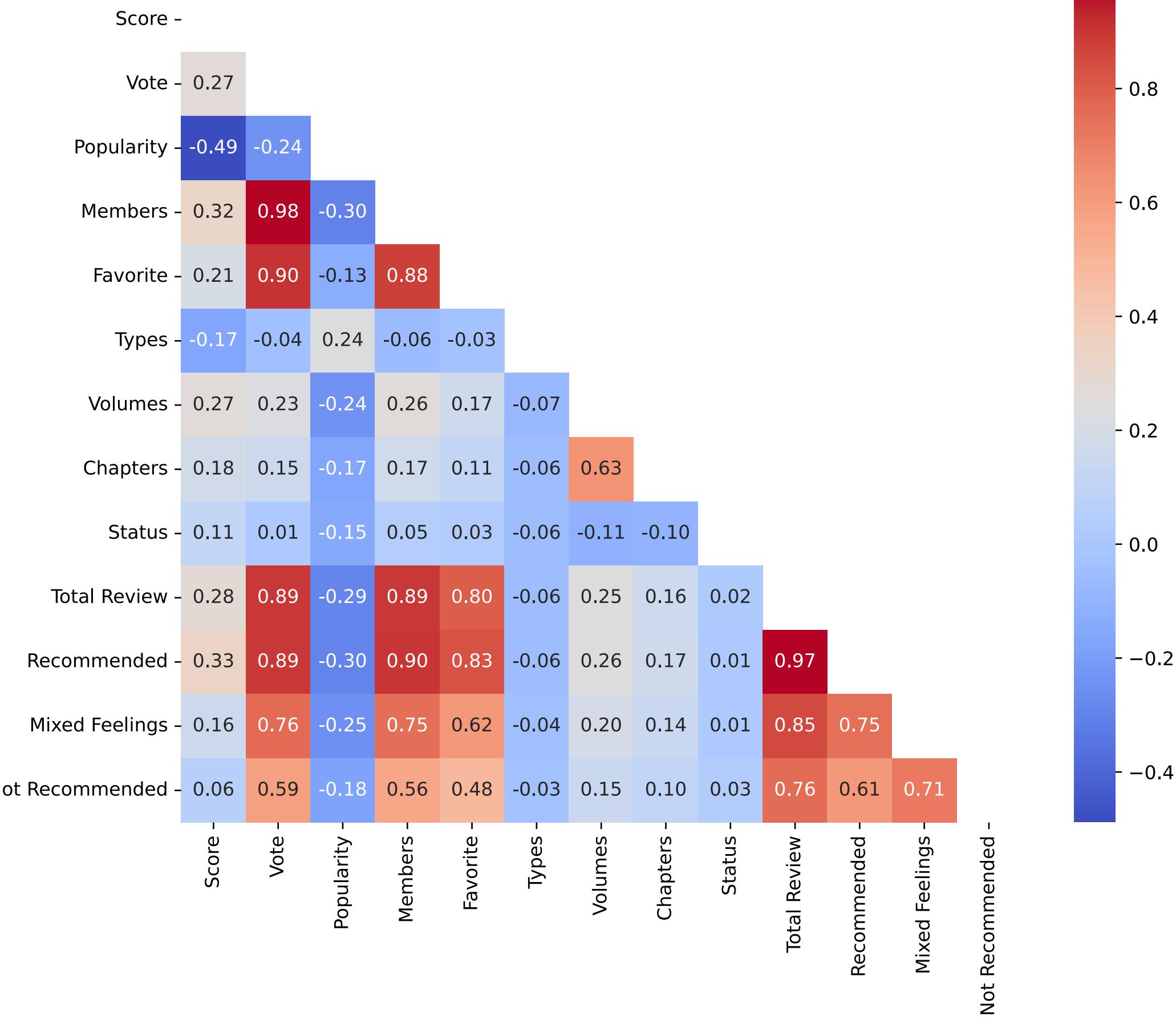
Upper Half of the Correlation Matrix

Lựa chọn đặc trưng

- Những đặc trưng phản ánh các yếu tố quan trọng tác động đến sự yêu thích và đánh giá của cộng đồng
- Loại bỏ các đặc trưng không có mối quan hệ với điểm số của manga
- Loại bỏ các đặc trưng có hệ số tương quan < 0.15

Xử lý dữ liệu

- Mã hóa các đặc trưng phân loại
- Chuẩn hóa dữ liệu bằng MinMaxScaler
- Chia dữ liệu thành các tập huấn luyện, xác thực, kiểm tra với tỷ lệ 70:15:15



XÂY DỰNG VÀ HUẤN LUYỆN MÔ HÌNH



Linear Regression

- Điểm số của manga có thể phụ thuộc tuyến tính vào các đặc trưng số lượng
- Giúp hiểu rõ tác động của từng yếu tố đến điểm số



XGBoost

- Mô hình mạnh mẽ tối ưu hóa độ chính xác
- Xử lý dữ liệu phức tạp với nhiều đặc trưng
- Phù hợp với bài toán yêu cầu hiệu suất cao



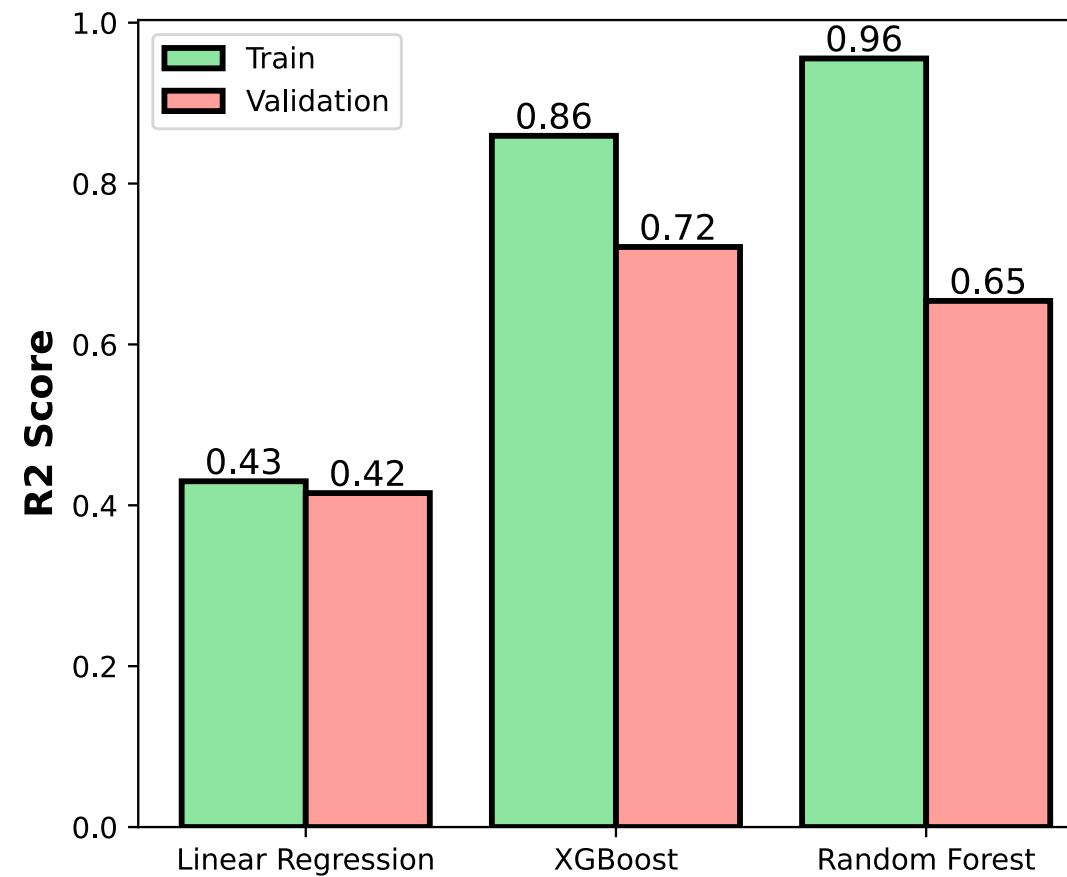
Random Forest

- Manga thường có các đặc điểm không tuyến tính
- Phù hợp với việc dự đoán từ các đặc trưng rời rạc hoặc phi tuyến

>>>

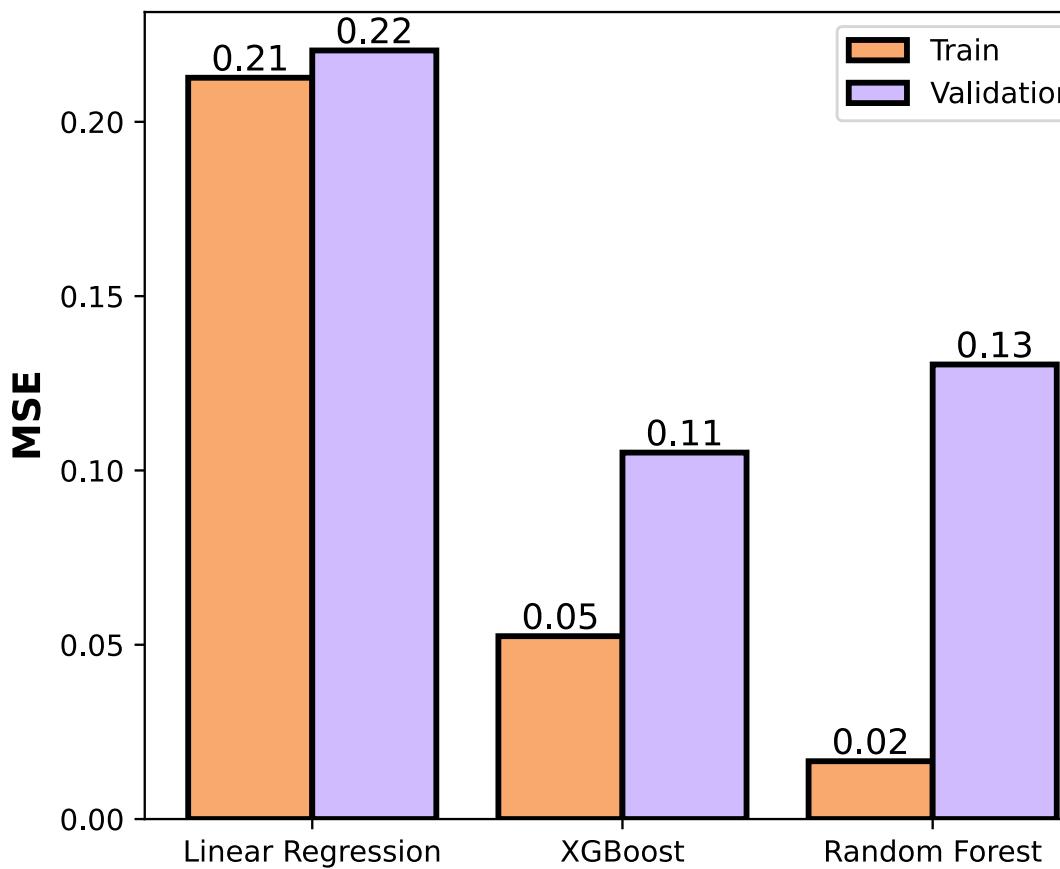
ĐÁNH GIÁ CÁC MÔ HÌNH

Train and Validation R2 Scores



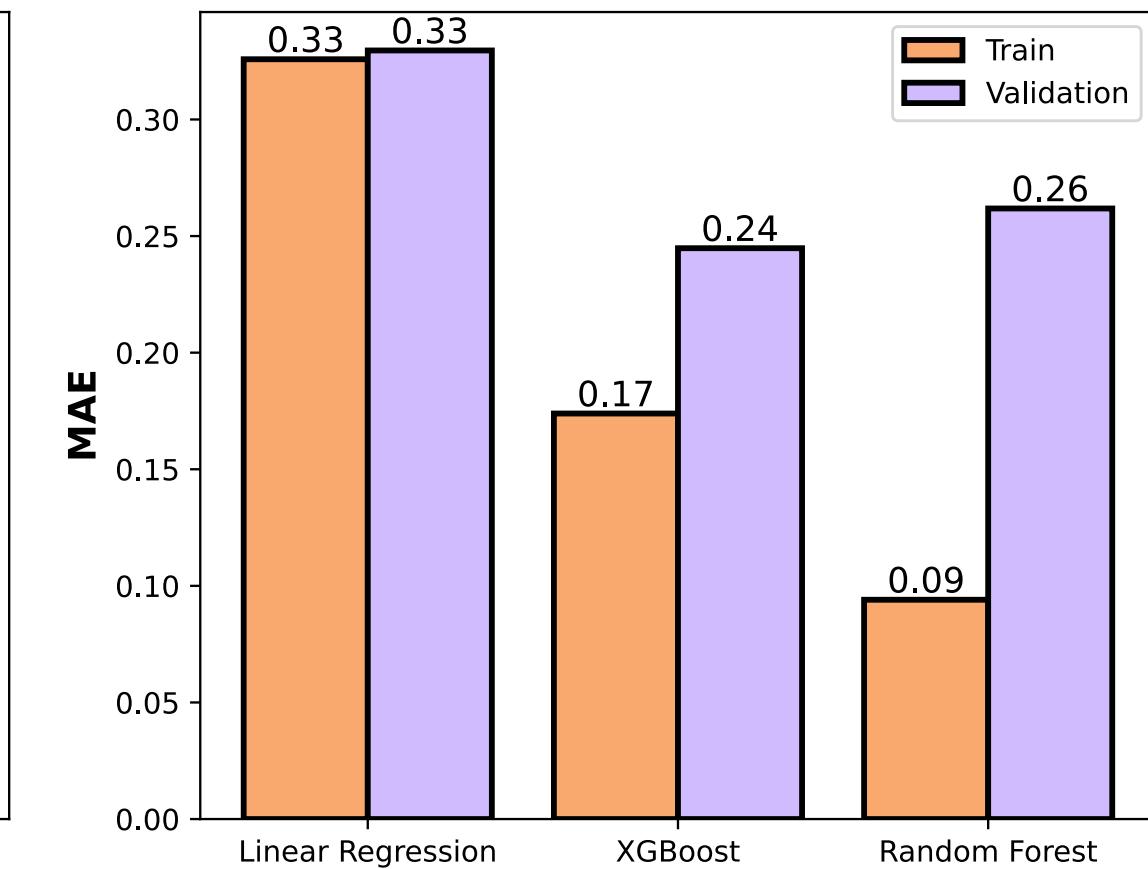
- XGBoost đạt R2 Score cao nhất và ổn định trên cả tập train và validation.
- Random Forest có nguy cơ overfitting.
- Linear Regression thì kém hiệu quả.

Train and Validation MSE



- XGBoost có MSE thấp nhất, cho thấy dự đoán chính xác hơn.
- Random Forest và Linear Regression có sai số cao hơn, đặc biệt trên validation.

Train and Validation MAE



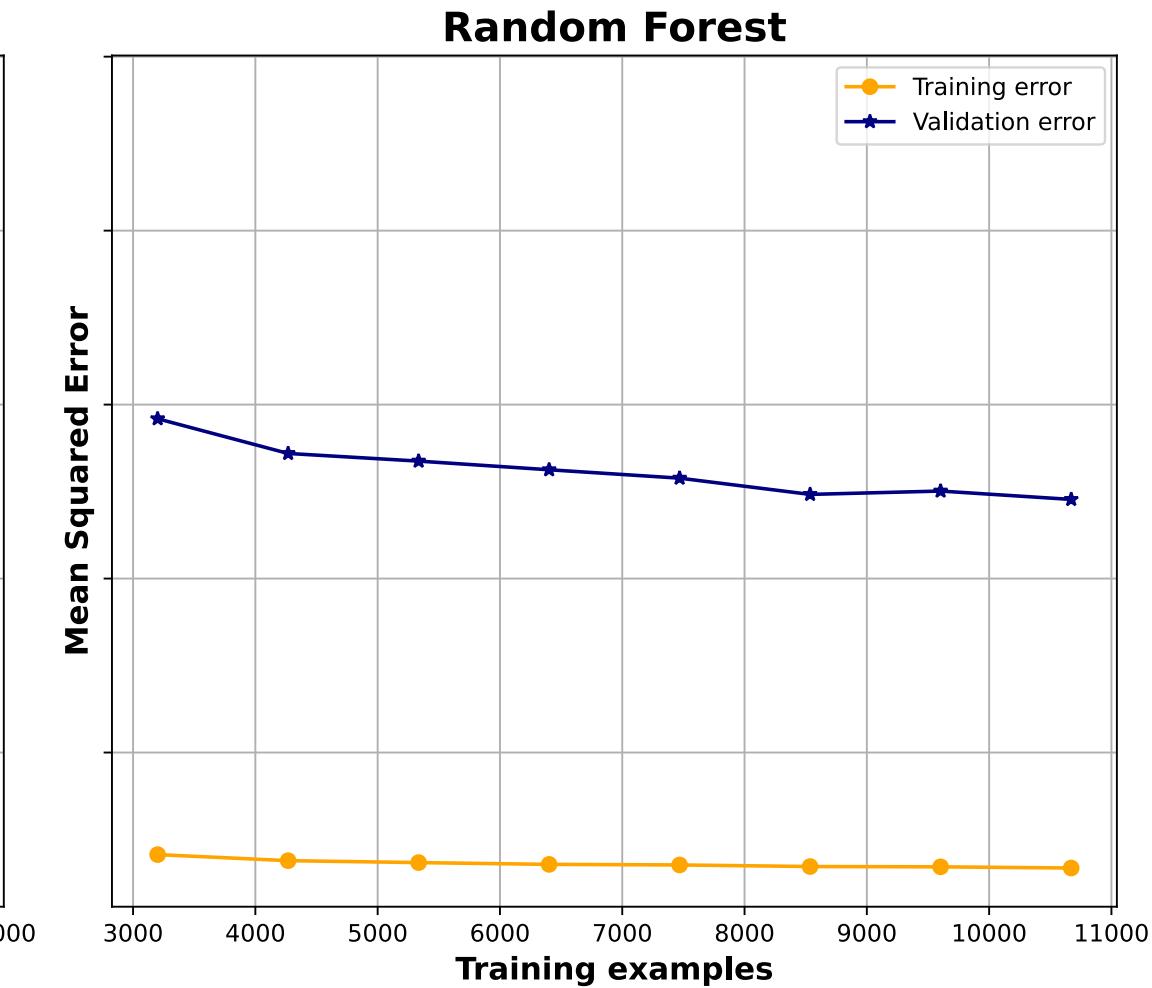
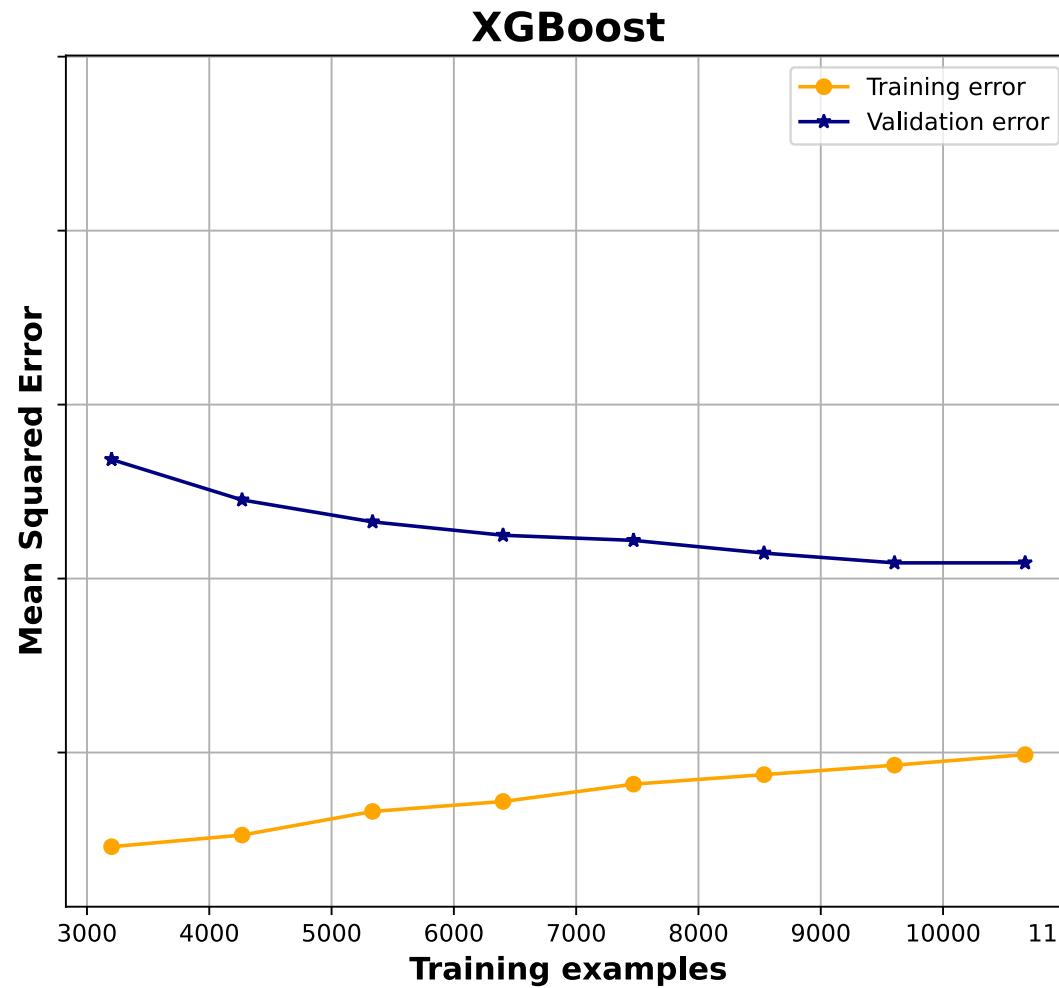
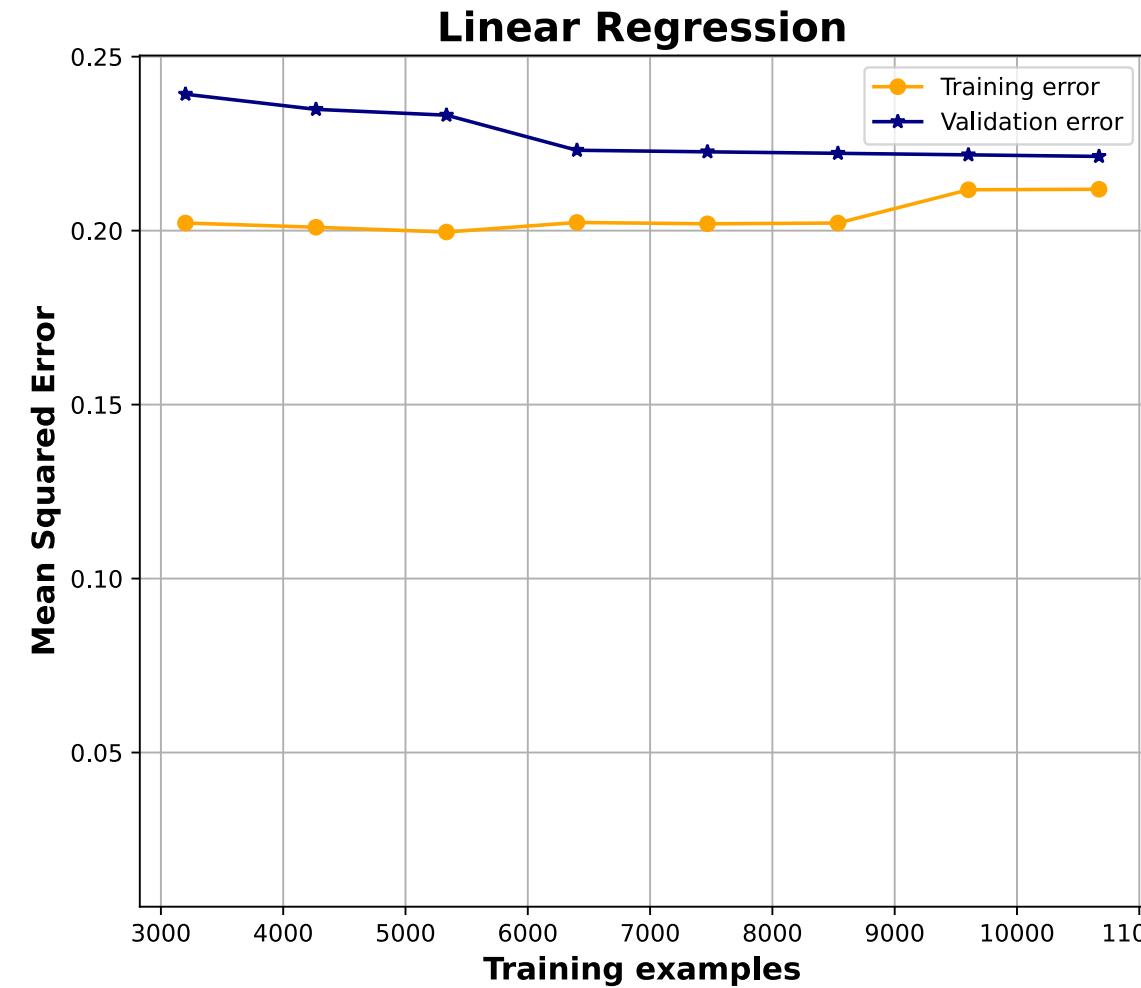
- XGBoost tiếp tục dẫn đầu với MAE thấp nhất, chứng minh hiệu suất vượt trội so với các mô hình khác.

Tổng quan thì XGBoost là lựa chọn tốt nhất, cân bằng giữa độ chính xác và khả năng tổng quát hóa. Random Forest có thể cải thiện với điều chỉnh thêm, còn Linear Regression không phù hợp



CẢI THIỆN CÁC MÔ HÌNH

Đường cong học tập cho phép chúng ta đánh giá xem mô hình có xu hướng học tập hay không, liệu nó có bị quá khớp (overfitting) hay không, và từ đó có thể đưa ra các điều chỉnh phù hợp.



- **Linear Regression:** Đường cong ổn định, không overfitting nhưng hiệu suất thấp, lỗi cao nhất.
- **XGBoost:** Đường cong rõ ràng, cân bằng tốt giữa training và validation, hiệu suất vượt trội.
- **Random Forest:** Học tốt trên training, nhưng khoảng cách lớn với validation, có dấu hiệu overfitting nhẹ.



TINH CHỈNH CÁC MÔ HÌNH



Mục tiêu



Dựa trên đường cong học tập, việc tinh chỉnh tham số sẽ cải thiện hiệu suất mô hình. Bayesian search được áp dụng vì khả năng dự đoán thông minh và tối ưu hóa nhanh chóng, giúp giảm phép thử và tránh overfitting.

Phạm vi tối ưu hóa



Linear Regression

- **fit_intercept:** Có tính toán hằng số hay không

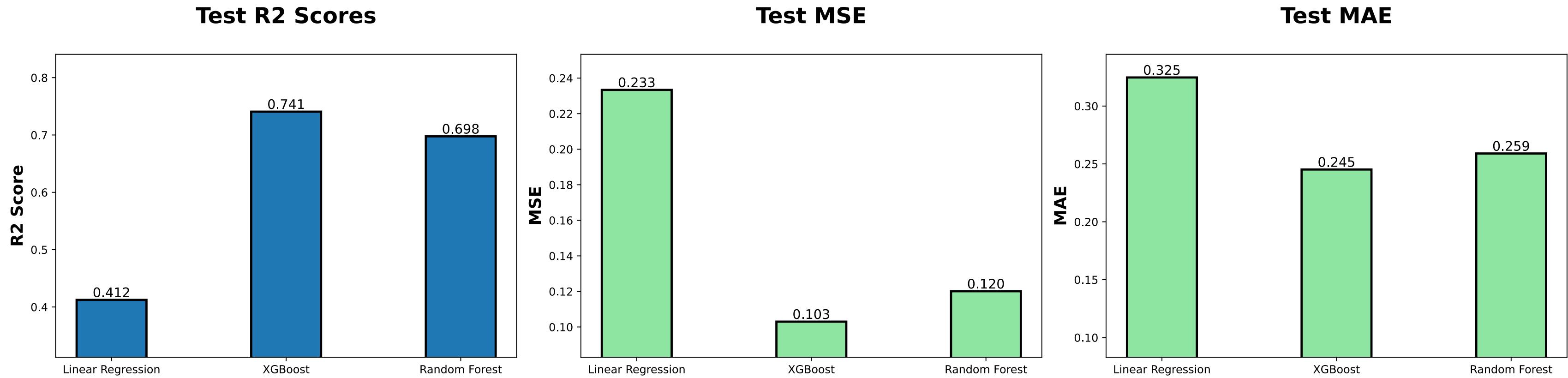
XGBoost

- **n_estimators:** Số cây quyết định
- **max_depth:** Độ sâu tối đa của cây
- **learning_rate:** Tốc độ học, sử dụng log-uniform để tối ưu phạm vi rộng

Random Forest

- **n_estimators:** Số cây trong rừng
- **max_depth:** Giới hạn độ sâu của cây
- **min_samples_split:** Mẫu tối thiểu để chia nút

ĐÁNH GIÁ CÁC MÔ HÌNH



XGBoost là mô hình tối ưu nhất trong ba mô hình, đạt hiệu suất vượt trội trên cả ba tiêu chí. Random Forest cũng cho kết quả tốt, trong khi Linear Regression kém hiệu quả nhất



ĐÁNH GIÁ CÁC MÔ HÌNH



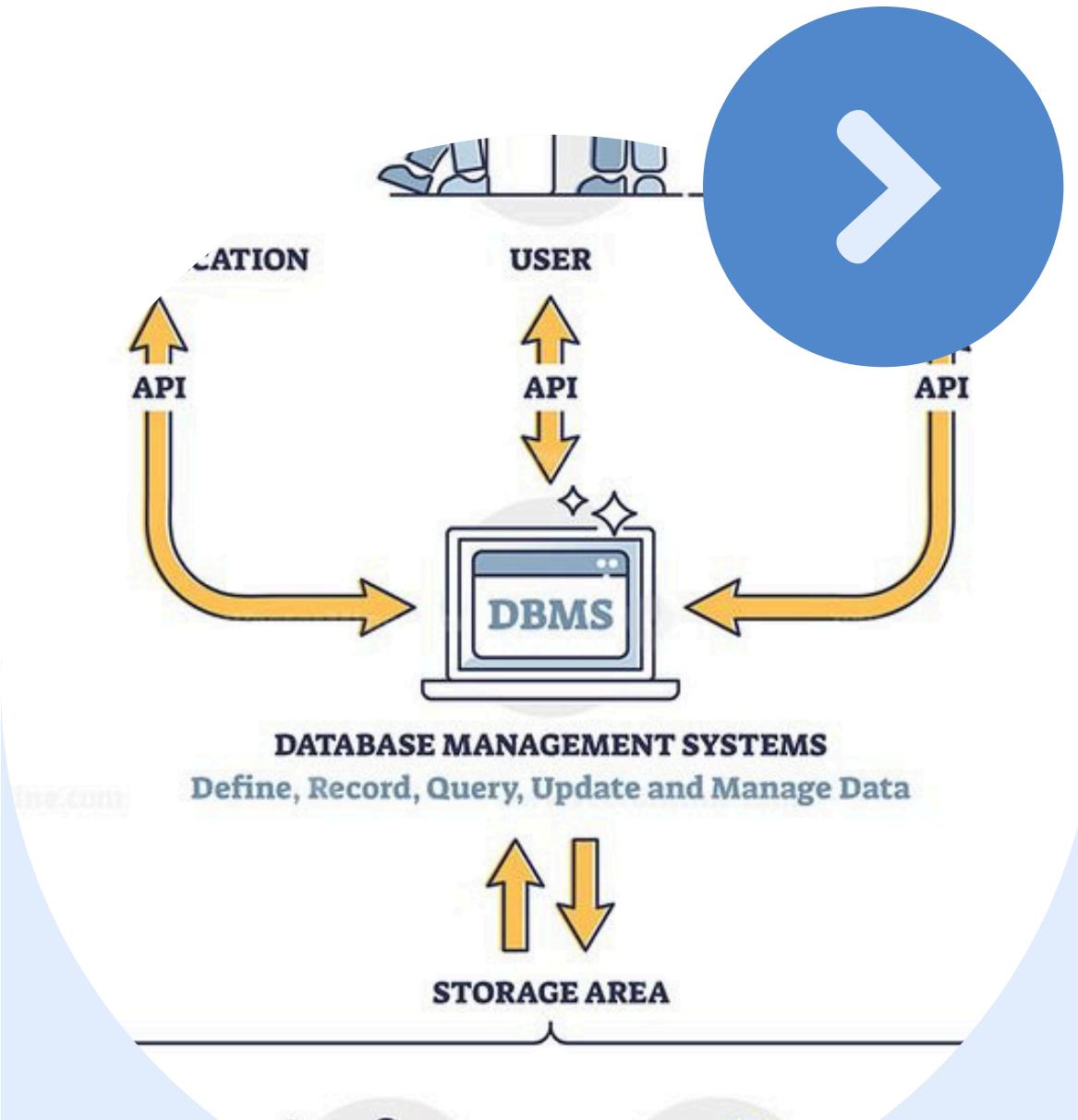
Dự đoán 10 dòng ngẫu nhiên

	Actual	Linear Regression	XGBoost	Random Forest
5814	7.20	6.849734	7.431932	7.274144
12745	6.71	6.988691	6.446942	6.456728
14682	6.56	6.701461	6.872587	6.883553
7136	7.10	6.956718	6.932580	6.908678
8915	6.98	6.973729	7.079154	6.995032
16670	6.37	6.463299	6.107481	5.697081
9332	6.95	6.988221	6.483134	6.681994
11955	6.77	7.009297	6.851516	6.867672
15051	6.53	6.607402	6.713037	6.648983
18484	5.92	6.613275	6.420870	6.498293

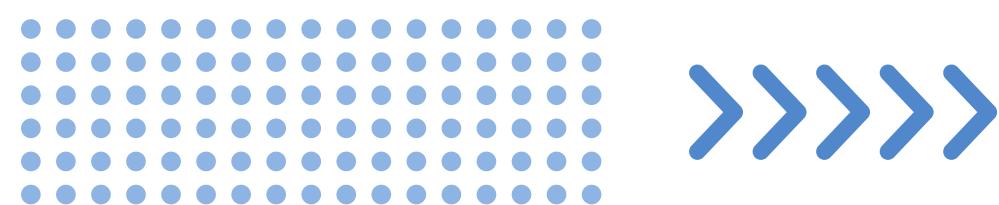
KẾT LUẬN

Dựa trên kết quả phân tích, XGBoost được xác định là mô hình tốt nhất để dự đoán điểm số của một bộ manga, với hiệu suất vượt trội trên các tiêu chí R2 Score, MSE và MAE. Việc đạt được độ chính xác cao trong dự đoán có ý nghĩa quan trọng như sau:

- **Hỗ trợ độc giả:** gợi ý các bộ manga phù hợp.
- **Hỗ trợ nhà xuất bản và tác giả:** dự đoán tiềm năng thành công của một bộ manga.
- **Nâng cao giá trị thị trường:** cung cấp dữ liệu đáng tin cậy để định giá và xây dựng chiến lược phát triển sản phẩm.
- **Định hướng cải thiện chất lượng:** tối ưu hóa chất lượng sản phẩm và sự yêu thích từ độc giả.



Hệ thống phân loại bộ truyen dựa trên đối tượng độc giả





Vấn đề

Hệ thống phân loại truyện dựa trên đối tượng độc giả đóng vai trò quan trọng trong việc đảm bảo nội dung phù hợp, tăng cường trải nghiệm đọc và hỗ trợ quản lý ngành xuất bản. Bằng cách xác định rõ ràng từng nhóm đối tượng độc giả (trẻ em, thanh thiếu niên, người lớn), hệ thống này không chỉ giúp bảo vệ độc giả mà còn thúc đẩy sự phát triển bền vững cho ngành công nghiệp manga và thể loại khác.

Đảm bảo nội dung phù hợp
với từng nhóm độc giả

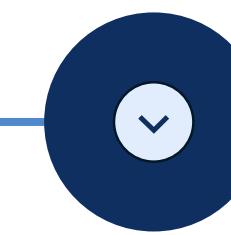
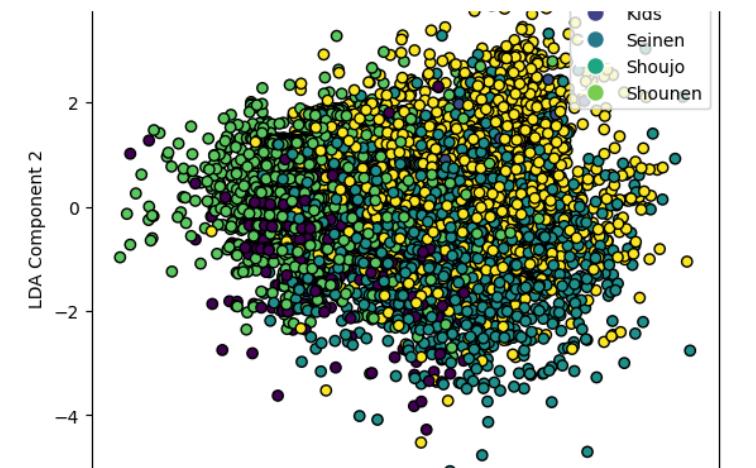
Hỗ trợ quản lý và tìm kiếm
dễ dàng

Thúc đẩy sáng tạo và chiến
lược quảng bá hiệu quả

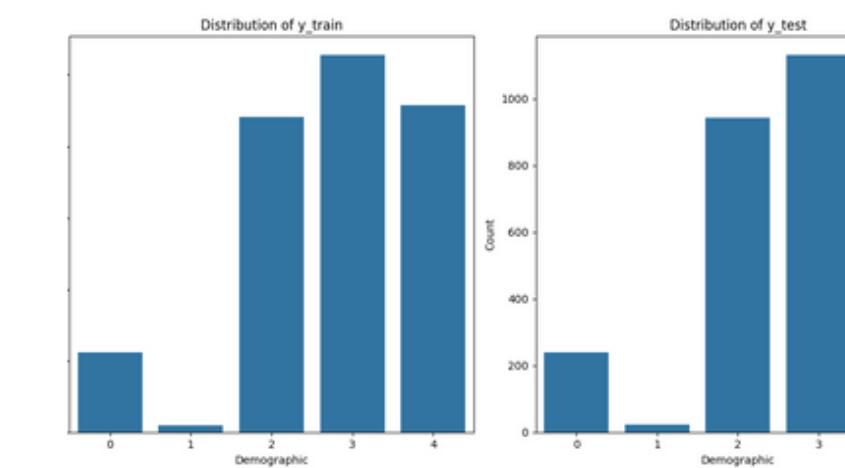
Quy trình xây dựng hệ thống



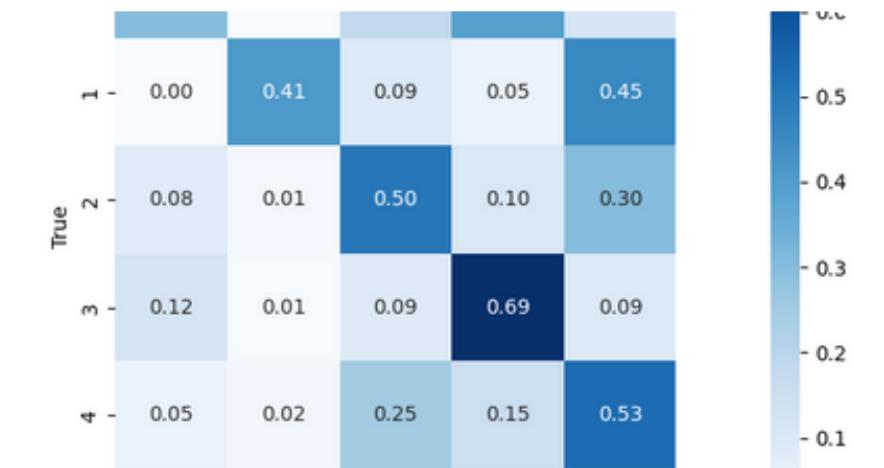
Xử lý và giảm chiều dữ liệu các đặc trưng bằng LDA



Sử dụng các phương pháp xử lý dữ liệu bất cân bằng và so sánh kết quả



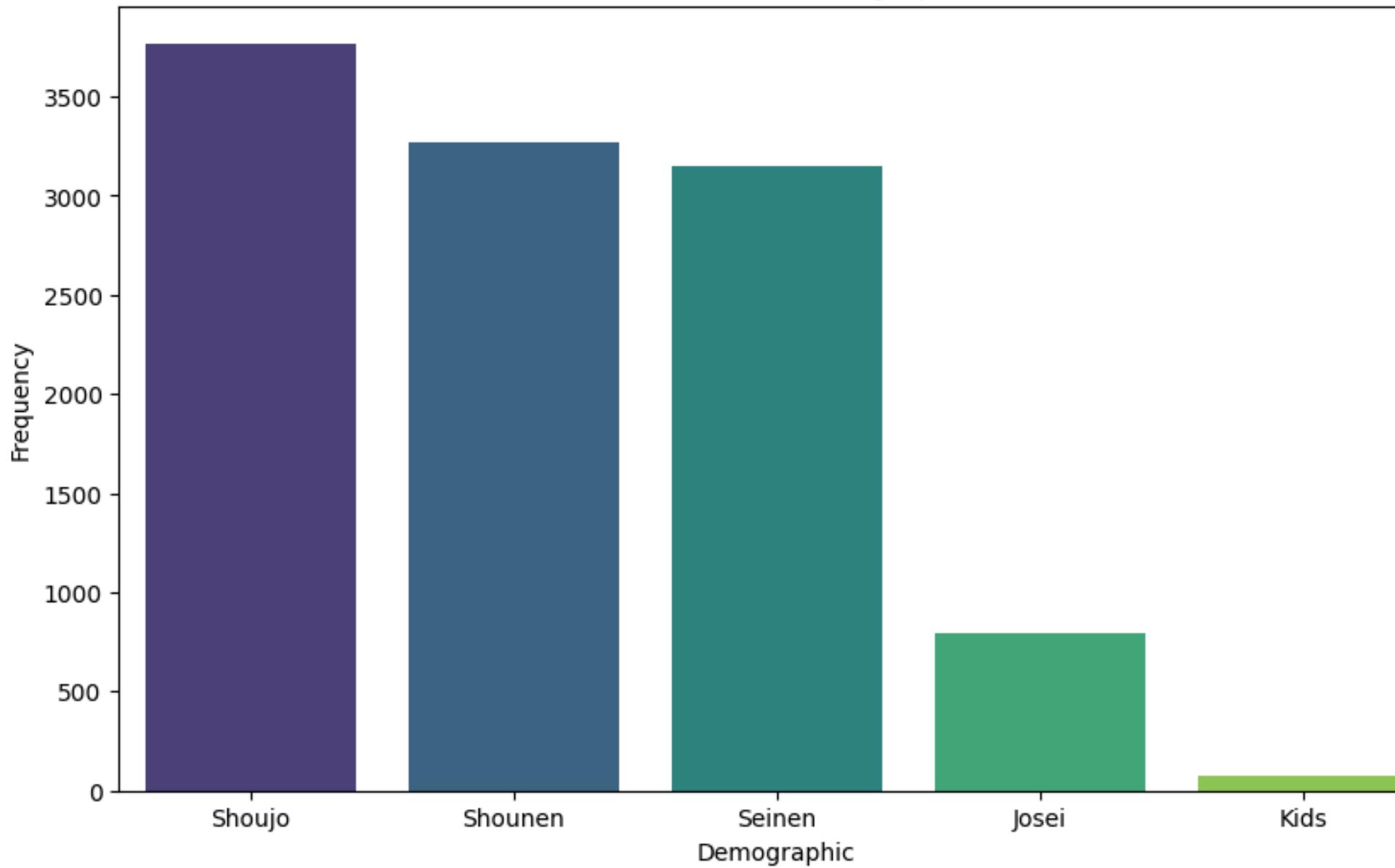
Fine tune mô hình trên các phương pháp xử lý dữ liệu bất cân bằng và chọn ra phương pháp tốt nhất





Mô tả phân phôi các lớp dữ liệu

Distribution of Demographic



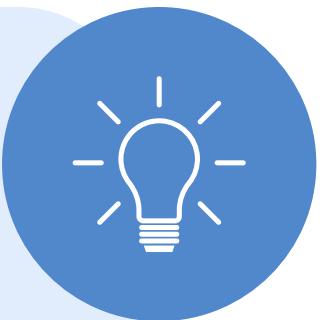
Bộ dữ liệu bị mất cân bằng nặng, các bộ manga `Shoujo` có lượng cao nhất >3500 bộ, tuy nhiên các bộ manga như `Josei`, `Kids` lại có số lượng rất nhỏ < 1000 bộ, do đó ở các bước sau cần tiến hành thực hiện cân bằng lại dữ liệu



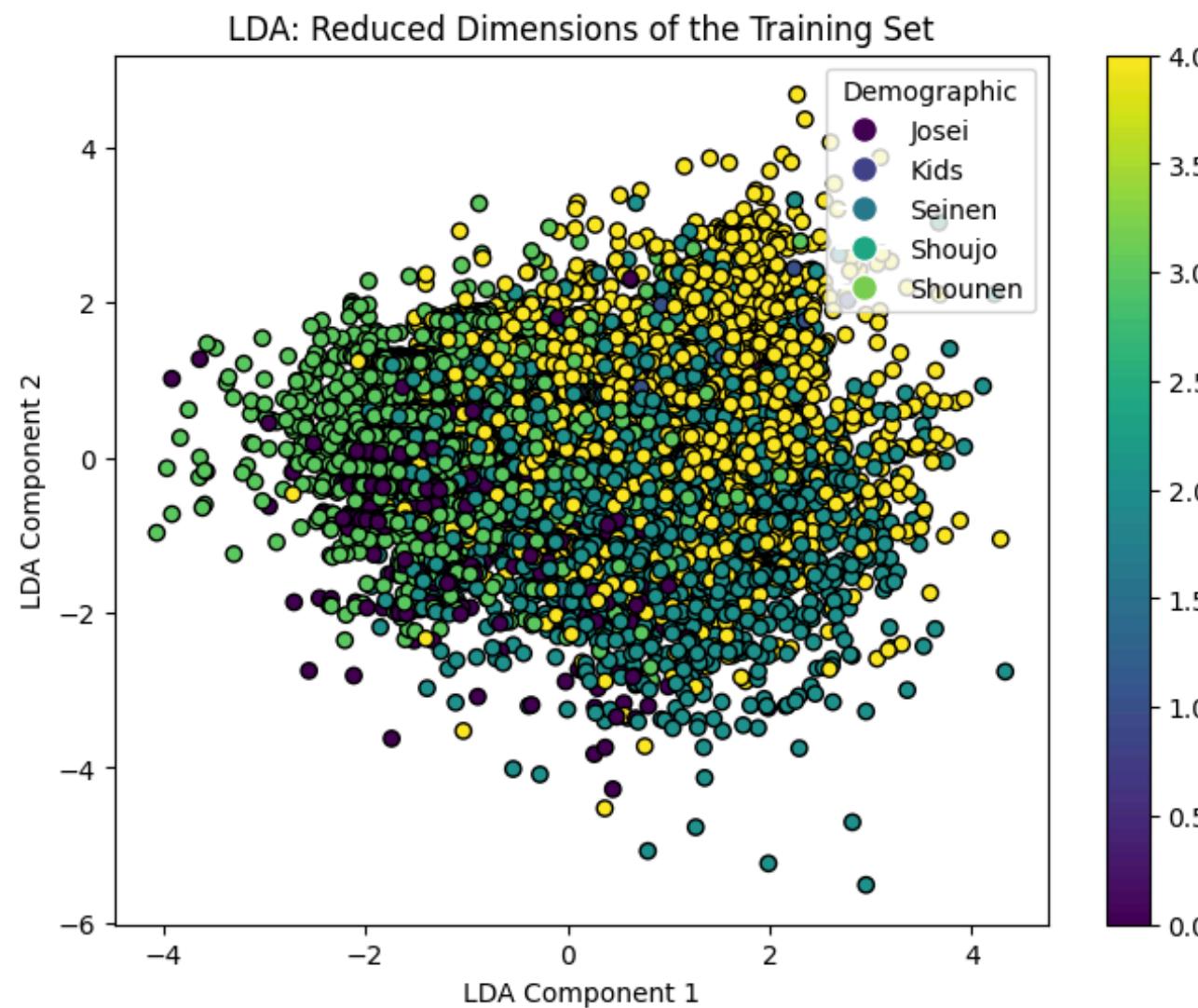
Chuẩn hóa các thuộc tính

	Score	Popularity	Recommended	Action	Adult Cast	Adventure	Anthropomorphic	Avant Garde	Award Winning	Boys Love	...	Villainess	Visual Arts	Workplace	Types_Doujinshi	Types_Light Novel
0	4.107168	-1.422254	28.753729	1.850976	-0.096991	3.153185	-0.065351	-0.028548	5.700080	-0.097937	...	-0.053036	-0.049487	-0.072633	-0.021275	-0.110366
1	3.846887	-1.419631	13.893861	1.850976	-0.096991	3.153185	-0.065351	-0.028548	-0.175436	-0.097937	...	-0.053036	-0.049487	-0.072633	-0.021275	-0.110366
2	3.765550	-1.420824	10.438078	1.850976	-0.096991	3.153185	-0.065351	-0.028548	5.700080	-0.097937	...	-0.053036	-0.049487	-0.072633	-0.021275	-0.110366
3	3.700480	-1.421897	21.611777	1.850976	-0.096991	3.153185	-0.065351	-0.028548	-0.175436	-0.097937	...	-0.053036	-0.049487	-0.072633	-0.021275	-0.110366
4	3.602875	-1.418916	7.673452	-0.540256	10.310236	-0.317140	-0.065351	-0.028548	5.700080	-0.097937	...	-0.053036	-0.049487	-0.072633	-0.021275	-0.110366
...
11047	-7.345167	3.005181	-0.274849	1.850976	-0.096991	-0.317140	-0.065351	-0.028548	-0.175436	-0.097937	...	-0.053036	-0.049487	-0.072633	-0.021275	-0.110366
11048	-7.345167	1.634060	-0.274849	1.850976	-0.096991	-0.317140	-0.065351	-0.028548	-0.175436	-0.097937	...	-0.053036	-0.049487	-0.072633	-0.021275	-0.110366
11049	-7.345167	2.836817	-0.274849	-0.540256	-0.096991	-0.317140	-0.065351	-0.028548	-0.175436	-0.097937	...	-0.053036	-0.049487	-0.072633	-0.021275	-0.110366
11050	-7.345167	3.005300	-0.274849	-0.540256	-0.096991	-0.317140	-0.065351	-0.028548	-0.175436	-0.097937	...	-0.053036	-0.049487	-0.072633	-0.021275	-0.110366
11051	-7.345167	2.695758	-0.274849	-0.540256	-0.096991	-0.317140	-0.065351	-0.028548	-0.175436	-0.097937	...	-0.053036	-0.049487	-0.072633	-0.021275	-0.110366

11052 rows × 82 columns



Giảm đặc trưng dữ liệu bằng LDA



Có thể thấy nếu bỏ đi các đặc trưng có độ ảnh hưởng lớn hơn 0.03, chúng ta có thể giảm số lượng thuộc tính từ 82 thuộc tính xuống còn 48 thuộc tính nhưng vẫn giữ được độ chính xác là ~0.56, điều này giúp mô hình giảm độ phức tạp cũng như giảm overfitting

Feature_importance	Columns	No-Imbalanced-Handling	LABEL 0 (F1-Score)	LABEL 1 (F1-Score)	LABEL 2 (F1-Score)	LABEL 3 (F1-Score)	LABEL 4 (F1-Score)
0	82	0.564	0.211	0.341	0.512	0.692	0.536
0.02	59	0.563	0.205	0.256	0.508	0.701	0.531
0.03	48	0.56	0.192	0.235	0.514	0.686	0.534
0.1	20	0.534	0.201	0.25	0.481	0.669	0.499
0.2	10	0.466	0.147	0.05	0.426	0.582	0.457



Xử lý dữ liệu bất cân bằng

Before Fine-Tuning	Accuracy	LABEL 0 (F1-Score)	LABEL 1 (F1-Score)	LABEL 2 (F1-Score)	LABEL 3 (F1-Score)	LABEL 4 (F1-Score)
DOWNSAMPLING	0.613	0.522	0.557	0.542	0.707	0.629
UPSAMPLING	0.563	0.202	0.229	0.51	0.691	0.538
CLASS-WEIGHT	0.564	0.21	0.222	0.519	0.693	0.532

- Độ chính xác khi sử dụng phương pháp DownSampling là cao nhất (0.613), thế nhưng rõ ràng khi sử dụng phương pháp DownSampling, một số lượng các mẫu sẽ bị loại bỏ, điều này có thể tăng nguy cơ mô hình giảm đi tính khai quát hóa trong quá trình học
- Sử dụng phương pháp UpSampling có thể khắc phục nhược điểm của DownSampling, tuy nhiên các chỉ số Accuracy cũng như F1-score của phương pháp này không được cao cho lắm, điều này có thể do các mẫu thiểu số bị trùng quá nhiều khiến mô hình giảm khả năng học đặc trưng từ lớp này
- Để khắc phục nhược điểm của 2 phương pháp trên, đề xuất sử dụng phương pháp Class-Weight, giữ lại phân phối của các lớp tuy nhiên đánh trọng số cho các label, các chỉ số Accuracy cũng như F1-score của phương pháp này cao hơn UpSampling nhưng lại thấp hơn DownSampling



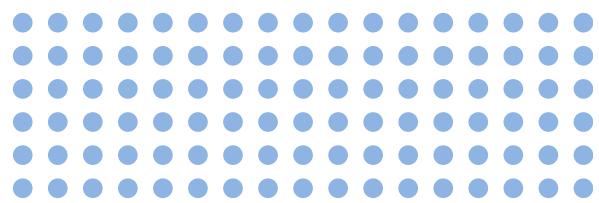
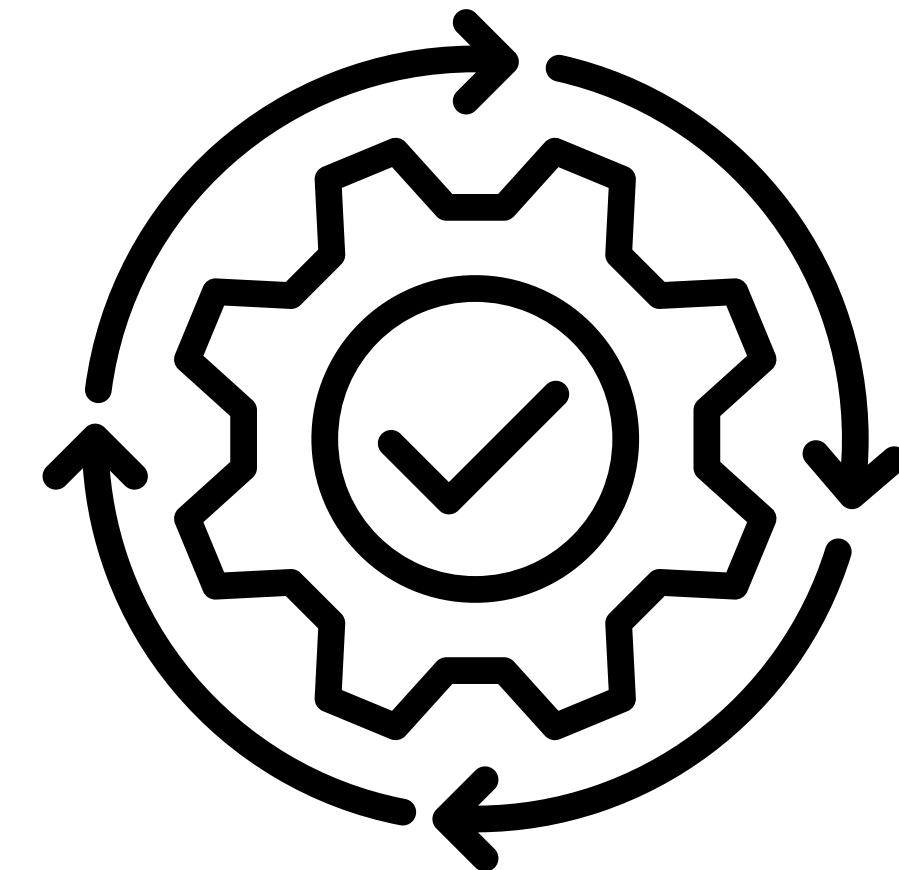
Fine-tune mô hình

```
ExtraTreesClassifier
ExtraTreesClassifier(class_weight='balanced', max_features='log2',
                     min_samples_split=5, n_estimators=200, random_state=42)
```

After Fine-Tuning	Accuracy	LABEL 0 (F1-Score)	LABEL 1 (F1-Score)	LABEL 2 (F1-Score)	LABEL 3 (F1-Score)	LABEL 4 (F1-Score)
DOWNSAMPLING	0.54	0.336	0.375	0.54	0.618	0.562
UPSAMPLING	0.575	0.211	0.27	0.525	0.702	0.548
CLASS-WEIGHT	0.562	0.255	0.254	0.525	0.696	0.54

ETC Classification Report:				
	precision	recall	f1-score	support
0	0.221	0.301	0.255	239
1	0.184	0.409	0.254	22
2	0.549	0.503	0.525	944
3	0.696	0.695	0.696	1131
4	0.549	0.531	0.540	980
accuracy			0.562	3316
macro avg	0.440	0.488	0.454	3316
weighted avg	0.573	0.562	0.566	3316

KHÓ KHĂN VÀ CÁCH GIẢI QUYẾT



KHÓ KHĂN

- Trang web giới hạn số lượng requests khiến việc thu thập dữ liệu trở nên khó khăn
- Khó khăn trong việc đánh giá và quyết định xử lý từng cột thuộc tính trong bộ dữ liệu như thế nào, điền các giá trị bị thiếu trong từng cột theo giá trị gì, ...
- Đặt những câu hỏi có ý nghĩa là một vấn đề nan giải và thách thức với nhóm
- Việc lên ý tưởng, sử dụng và tinh chỉnh các mô hình để giải quyết các vấn đề phân lớp, hồi quy trong bộ dữ liệu gây khó khăn cho thành viên trong nhóm do chưa có nhiều kinh nghiệm về Machine Learning, Deep Learning,...

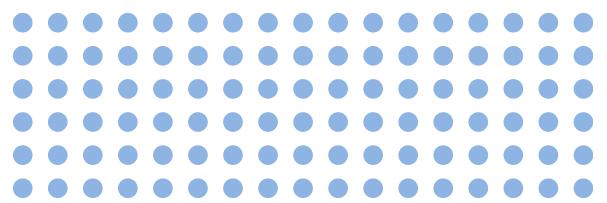
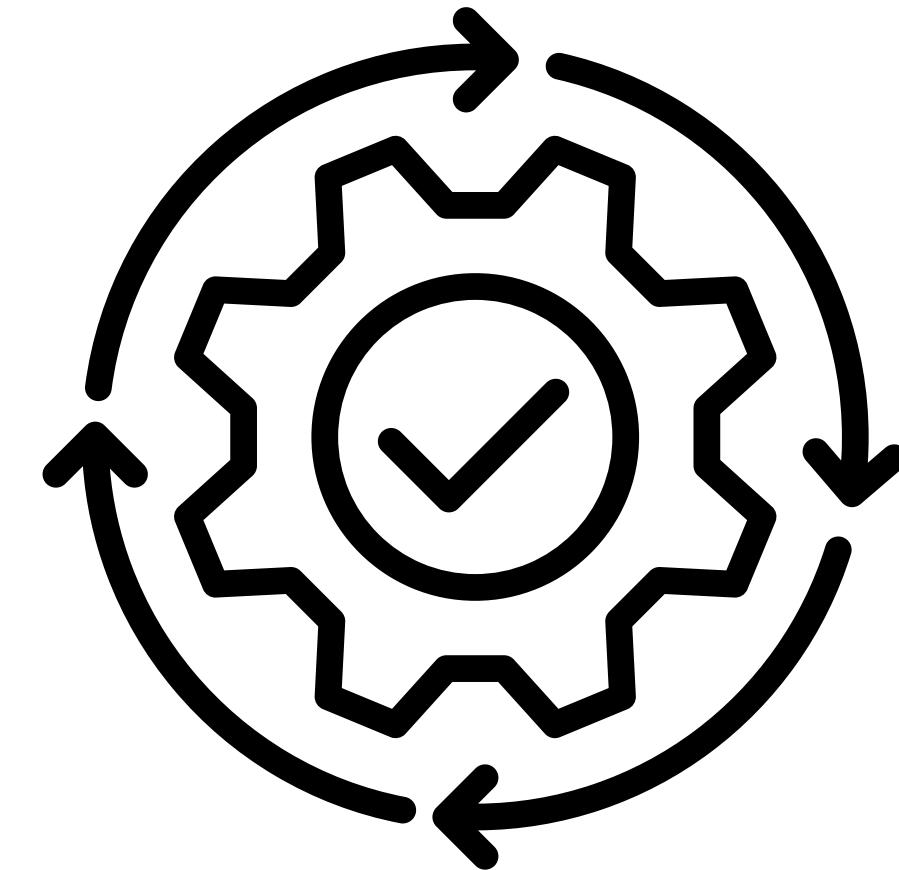
HƯỚNG GIẢI QUYẾT

- Chia nhỏ các phần cần thu thập thành nhiều file notebook thực thi cùng một lúc. Sau khi thu thập đủ dữ liệu, tiến hành ghép dữ liệu để tạo thành bộ hoàn chỉnh.
- Nhóm đã thảo luận với nhau và đưa ra các quyết định về việc xử lý và khai phá dữ liệu
- Về việc xây dựng mô hình, các thành viên được giao nhiệm vụ đã cố gắng tìm hiểu về các mô hình, cũng như cách thức áp dụng chúng vào trong các bài toán cần sử dụng mô hình để dự đoán, phân lớp

HƯỚNG GIẢI QUYẾT

- Nhóm đã đọc và tìm hiểu kĩ từng thuộc tính của dữ liệu được thu thập, đồng thời cũng tìm hiểu và tham khảo xu hướng người dùng, nhà sản xuất, cũng như từng chuyển biến của Manga, Manhwa, Manhua, Light Novel, Novel, Oneshot, Doujinshi trên thế giới thông qua các mạng xã hội như YouTube, Facebook, X,... để có thể đưa ra những câu hỏi có ý nghĩa và phù hợp với xu hướng hiện nay.
- Nhóm cũng tích cực trao đổi trên nhóm chat riêng, cũng như thông qua Github để nắm được tiến độ làm việc của nhau cũng như hiểu biết về phần việc của từng thành viên rõ hơn.

SUY NGHĨ VÀ ĐÁNH GIÁ ĐỒ ÁN



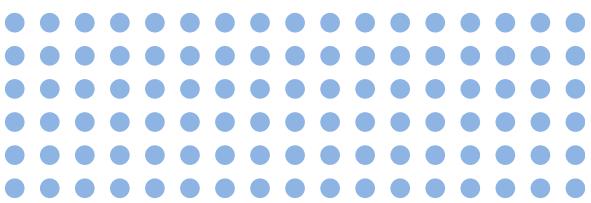
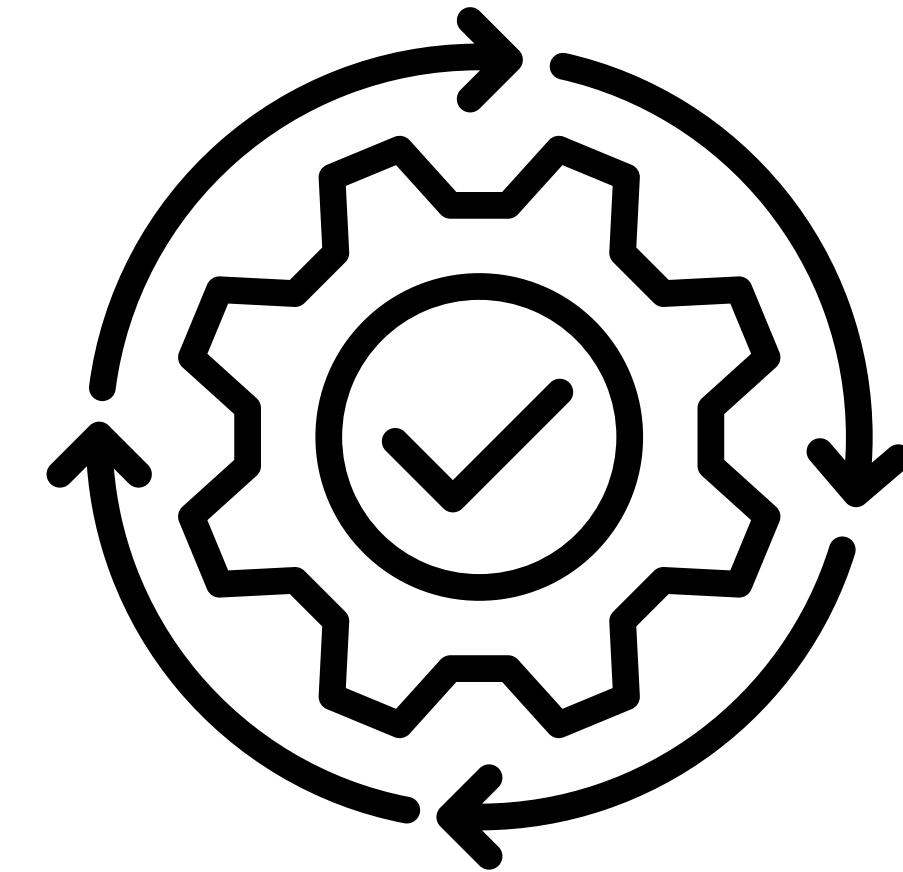
ĐIỀU HỌC ĐƯỢC TỪ ĐỒ ÁN NÀY

- Học cách thu thập dữ liệu bằng công cụ như requests, BeautifulSoup, Selenium,...
- Học được cách tiền xử lý và khai phá dữ liệu như xử lý dữ liệu bị thiếu, nhập xuất dữ liệu từ tập tin, kiểm tra kiểu dữ liệu các thuộc tính, ...
- Học được cách trực quan hóa bằng thư viện và nắm được cách đặt những câu hỏi có ý nghĩa dựa trên bộ dữ liệu hiện có
- Học được cách lên ý tưởng, sử dụng và tinh chỉnh các mô hình để giải quyết các vấn đề phân lớp, hồi quy trong bộ dữ liệu
- Học được cách phối hợp nhóm, thuyết trình, cũng như sử dụng GitHub để quản lý đồ án.

NẾU CÒN THỜI GIAN...

- Nhóm sẽ tìm hiểu sâu hơn về phân phôi và ý nghĩa của từng thuộc tính trong bộ dữ liệu. Đồng thời áp dụng các phương pháp xử lý dữ liệu tốt hơn.
- Đặt nhiều câu hỏi có ý nghĩa hơn, do hiện tại vẫn còn một vài câu mang tính khai thác mỗi quan hệ dữ liệu hơn là trả lời về một chủ đề gì đó.
- Sử dụng các phương pháp, các mô hình tốt hơn để áp dụng cho các vấn đề liên quan đến hồi quy, phân lớp.
- Cải thiện hệ thống Recommendation cho các bộ truyện tốt hơn.

GROUP WORKING



PHÂN CÔNG VÀ TIẾN ĐỘ CỦA NHÓM

Code Issues Pull requests Actions Projects Security Insights Settings

FinalProject_IntroductionDS Public

master 7 Branches 0 Tags Go to file Add file Code About

No description, website, or topics provided.

Readme Activity

Group_Working.xlsx

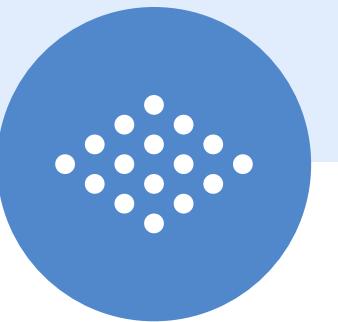
Tệp Chính sửa Xem Chèn Định dạng Dữ liệu Công cụ Trợ giúp

Trình đọc ↻ 🔍 100% ⚡ % .0 .00 123 Arial

A1 Phân công

	A	B	C	D	E	F
1	Phân công	Người thực hiện	Deadline	Hoàn thành	Nội dung thực hiện	
2	Tim đề tài	Cả nhóm	Trước 19/10	x	Tim đề tài để thực hiện đồ án	
3	Báo cáo đề tài	Huy, Khang	Trước 20/10	x	Viết báo cáo về đề tài thực hiện	
4	Data Collecting	Huy	17/11 - 20/11	x	Các nguồn thu thập dữ liệu, Tính hợp pháp của dữ liệu thu thập được, Dữ liệu thu thập được có đáng tin cậy?	
5	Data Preprocessing	Huy, Khang	20/11 - 2/12	x	Tiến xử lý liệu: Xử lý dữ liệu bị thiếu? Chuẩn hóa dữ liệu? Khảo sát ngoại lai? Tổng hợp dữ liệu trong trường sử dụng nhiều nguồn dữ liệu?	
6	Báo cáo tiến độ giữa kì	Huy, Khang, Tân Hưng	- Chính sửa slide: Huy (22/11) - Báo cáo giữa kì: Tân Hưng, Duy Khang (23/11 - 24/11)	x	Kiểm tra tính thống nhất của dữ liệu (kiểu dữ liệu, trùng lặp, ý nghĩa của các dòng/ các cột)?	
7	Data Exploring	Khang	2/12 - 14/12	x	Báo cáo tiến độ giữa kì	
8	Đặt câu hỏi	Cả nhóm (Mỗi người từ 1 - 2 câu hỏi)	2/12 - 5/12	x	Điều tra dữ liệu đã thu thập bằng cách sử dụng thống kê mô tả để hiểu dữ liệu tốt hơn, tức là xác định các vấn đề về dữ liệu (dữ liệu có giá trị bị thiếu, giá trị không hợp lệ, cột có kiểu dữ liệu không phù hợp để xử lý thêm, v.v.). Sau đây là một số thông tin cần xem xét: • Ý nghĩa của từng cột là gì? • Kiểu dữ liệu hiện tại của từng cột là gì? Có cột nào có kiểu dữ liệu không phù hợp không?	
9	Trả lời các câu hỏi bằng Data Visualization	Tân Hưng	6/12 - 14/12	x	Đặt ra các câu hỏi có ý nghĩa, rút ra mối quan hệ giữa các thuộc tính Với mỗi câu: - Nội dung câu hỏi? - Kết quả khi trả lời câu hỏi đó? - Nội dung câu hỏi? - Kết quả khi trả lời câu hỏi đó? (Dựa trên kết quả từ các biểu đồ trực quan được) - Đưa ra những nhận xét và rút ra (tiểu) kết luận.	
10	Data Modeling	Minh Hưng, Gia Hào	2/12 - 14/12	x	Mô hình hóa dữ liệu: Dựa trên dữ liệu và dựa trên vấn đề ban đầu của bạn, bạn có thể sử dụng mô hình học máy nào để giải quyết tự động? Hoặc bạn đưa ra một bài toán có thể sử dụng mô hình dựa trên dữ liệu của bạn. - Ban xử lý dữ liệu như thế nào để phù hợp? Phân chia train/ test/ valid? - Mô hình của bạn tốt như thế nào? So sánh với các mô hình cơ sở nào? Dựa trên độ đo gì? - Kết quả mà mô hình bạn trả ra cho bạn? Ý nghĩa của nó?	
11	Building System - Recommender System	Minh Hưng	2/12 - 14/12	x	Dựa vào đánh giá của Data Modeling + Áp dụng thêm các kiến thức khác, xây dựng hệ thống Recommender để cho đề xuất bộ truyện phù hợp với độc giả	
12	Trình bày lại các notebook, viết báo cáo	Cả nhóm	15/12 - 31/12: Format xong xuôi notebook và báo cáo. Nộp project cuối kì 15/12 - 27/12: Thực hiện xong slide và kịch bản thuyết trình. Kiểm tra lại các notebook		Báo cáo cuối kì	
13	Ghi chú: Các thành viên có thể cập nhật notebook phần mình làm sau khi hết deadline					
14						
15						
16						
17						

+ Phân công Trang Web Modeling Câu hỏi



**THANK
YOU!**

