**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN – ĐHQG TPHCM**

# Báo cáo

## Chủ đề: Hệ thống phân tích đề xuất và xếp hạng Anime

**Học phần: Hệ Thống Tư Vấn**

**Giảng viên: Huỳnh Thanh Sơn**

**Lớp: 20KDL1**

**Nhóm 7**



| Họ và Tên | MSSV |
|---|---|
| Nguyễn Xuân Hải | 20280026 |
| Huỳnh Đoan Hồ | 20280032 |
| Hà Thư Hoàng | 20280034 |
| Hà Thành Long | 20280061 |

# MỤC LỤC

# I. Giới thiệu:

## A. Anime là gì?

- Anime là một thuật ngữ đến từ tiếng Nhật, được sử dụng để mô tả các bộ phim hoạt hình từ Nhật Bản.
- Anime được giao thương từ năm 1917 và bắt đầu nổi bật vào những năm 60 của thế kỷ XX, lan rộng ra quốc tế vào cuối thế kỷ XX.
- Hiện nay từ *"anime"* đã trở thành một từ vựng quốc tế để chỉ đến các loại phim hoạt hình Nhật Bản và được sử dụng rộng rãi trên khắp thế giới.
- Tính đến năm 2016, anime đã chiếm 60% các phim hoạt hình truyền hình trên toàn thế giới.
- Anime có thể bao gồm nhiều thể loại và chủ đề khác nhau, từ hành động, phiêu lưu, hài hước, đến kinh dị, tâm lý, khoa học viễn tưởng, và nhiều thể loại khác. Nó không chỉ dành cho trẻ em mà còn có nhiều sản phẩm được hướng đến đối tượng khán giả người lớn.
- Anime thường có đặc điểm nét vẽ độc đáo, cách diễn đạt cảm xúc sâu sắc, và thường xuyên sử dụng phong cách nghệ thuật phong phú. Các bộ anime có thể được sản xuất dưới nhiều định dạng khác nhau, bao gồm TV series, phim điện ảnh, OVA (Original Video Animation), và web series.
- Một số bộ Anime nổi tiếng thế giới như: Naruto, DragonBall, Thám tử lừng danh Conan,One Piece, FairyTail, Pokemon, One Punch Man,....

## B. Dữ liệu Anime?

- Dữ liệu được sử dụng lấy từ **myanimelist.net** với khoảng hơn 75000 đề xuất từ người dùng.
- Tập dữ liệu này chứa thông tin về dữ liệu ưu tiên của người dùng từ 73.516 người dùng trên 12.294 anime. Mỗi người dùng có thể thêm anime vào danh sách hoàn chỉnh của họ và xếp hạng cho nó và bộ dữ liệu này là sự tổng hợp các xếp hạng đó.
- Điểm/xếp hạng nằm trong khoảng từ 1 - 10 với 10 là tốt nhất. Nếu xếp hạng là -1, điều đó có nghĩa là người dùng không đưa ra xếp hạng cho mục đó.
- MyAnimeList, còn được gọi là MAL, là một trang web mạng xã hội anime và manga chứa cơ sở dữ liệu nơi người dùng có thể sắp xếp và thêm các anime khác nhau vào danh sách của họ. Khi được thêm vào danh sách, các mục anime sẽ được xếp hạng sau khi được xem. Quá trình này giúp tìm kiếm những người dùng có cùng sở thích.

# II. Ý tưởng, mục tiêu

## A. Ý tưởng:

- Hiện nay, với số lượng Anime quá lớn và với các thể loại Anime phong phú và đa dạng thì vấn đề đặt ra là làm sao tìm được một bộ Anime hay, phù hợp với sở thích cũng như mong muốn của người xem.
- Về sở thích thì mỗi người dùng lại có một sở thích khác nhau,và với mỗi bộ Anime có thể phù hợp với các nhóm người khác nhau.
- Và hiện nay cũng có những hệ thống gợi ý, tìm kiếm hỗ trợ người dùng tìm kiếm thể loại Anime mà họ thích. Phổ biến nhất là tìm kiếm theo tên và đưa ra gợi ý với các Anime có tên gần giống.
- Tuy nhiên, rất khó để một hệ thống luôn có thể gợi ý ra những Anime đúng với sở thích của người dùng. Vì vậy, các video review Anime đã trở nên hot và rất phổ biến hiện nay và được nhiều người sử dụng để tìm kiếm Anime mà họ thích.
- Nhưng xem Review Anime cũng tốn khá nhiều thời gian để chọn ra một Anime phù hợp.
- Vì vậy, mong muốn thực hiện một hệ thống gợi ý, đánh giá và đề xuất cho người dùng bộ Anime tốt nhất với họ và không mất quá nhiều thời gian để người dùng có thể đưa ra lựa chọn phù hợp.

## B. Mục tiêu:

- Mong muốn là có thể đưa ra gợi ý và đề xuất cho người dùng các bộ Anime phù hợp với sở thích của họ.
- Đó những gì người dùng sẽ muốn xem hoặc muốn mua trong tương lai. Để thực hiện những dự đoán như vậy, cần có một lượng lớn dữ liệu người dùng để tìm ra các mẫu và liên kết sở thích trước đây với các lựa chọn trong tương lai.
- Thông thường, rất khó để đưa ra những khuyến nghị tốt khi thông tin của người dùng bị hạn chế.
- Tuy nhiên sẽ tốt hơn khi người dùng cung cấp thông tin của họ một cách rõ ràng nhưng không nhiều người dùng thực hiện như chúng ta mong muốn.
- *Vì vậy, mục tiêu đặt ra là hệ thống có thể đưa ra các gợi ý tốt nhất và phù hợp nhất với sở thích của người dùng.*

# III. Đánh giá dữ liệu:

## A. Giới thiệu dữ liệu:

- Thông số trong bộ dữ liệu: có *2* dữ liệu: **anime và rating**
    - anime_id: id của anime trên myanimelist.net.

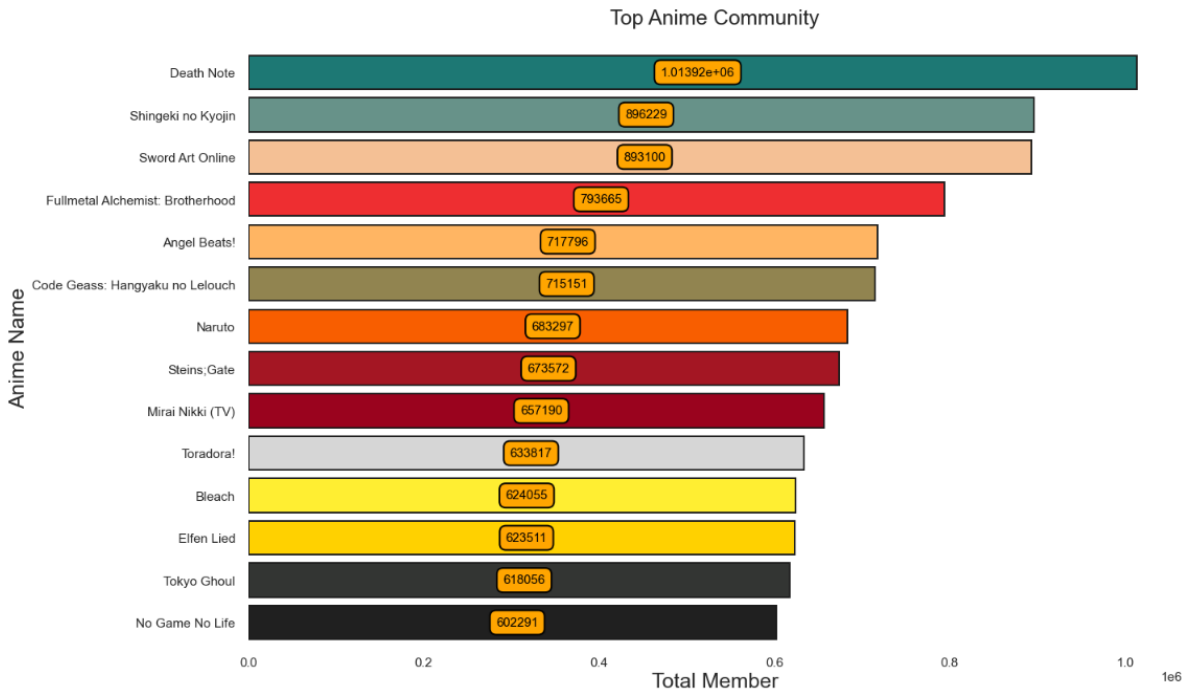
- name: tên đầy đủ của anime.
- genre: danh sách các thể loại của anime
- type: loại hình - phim, TV, OVA, v.v.
- episodes: số tập trong chương trình này (1 nếu là phim).
- rating: điểm trung bình từ 10 của anime này.
- members: số thành viên cộng đồng tham gia vào "nhóm" của anime này.
- user_id: id người dùng được tạo ngẫu nhiên và không thể nhận biết.
- anime_id:id anime mà người dùng này đã đánh giá.
- rating: điểm từ 10 mà người dùng này đã gán (-1 nếu người dùng xem nhưng không đánh giá).

- Dữ liệu về anime có 12294 dòng và 7 cột , có 317 dữ liệu bị khuyết (NAN)
- Dữ liệu về rating có 7813737 dòng và 3 cột , không có dữ liệu bị khuyết vì giá trị -1 đại diện cho người dùng không đánh giá.

## B. Trực quan hoá dữ liệu:

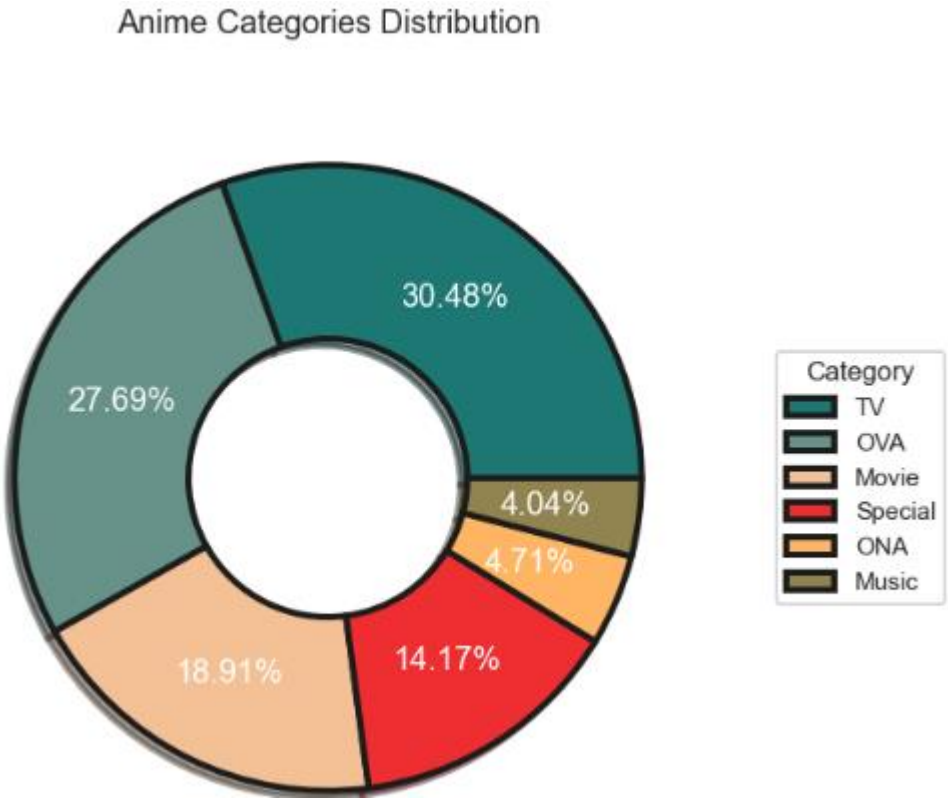
- Các Anime có cộng đồng người hâm mộ lớn nhất:

---




- **Death Note** là Anime có cộng đồng lớn nhất dẫn đầu với cộng đồng fan đông đảo hơn 1 triệu.
  - Thống kê số lượng Anime ở từng thể loại::
    - TV: 3402
    - OVA: 3090

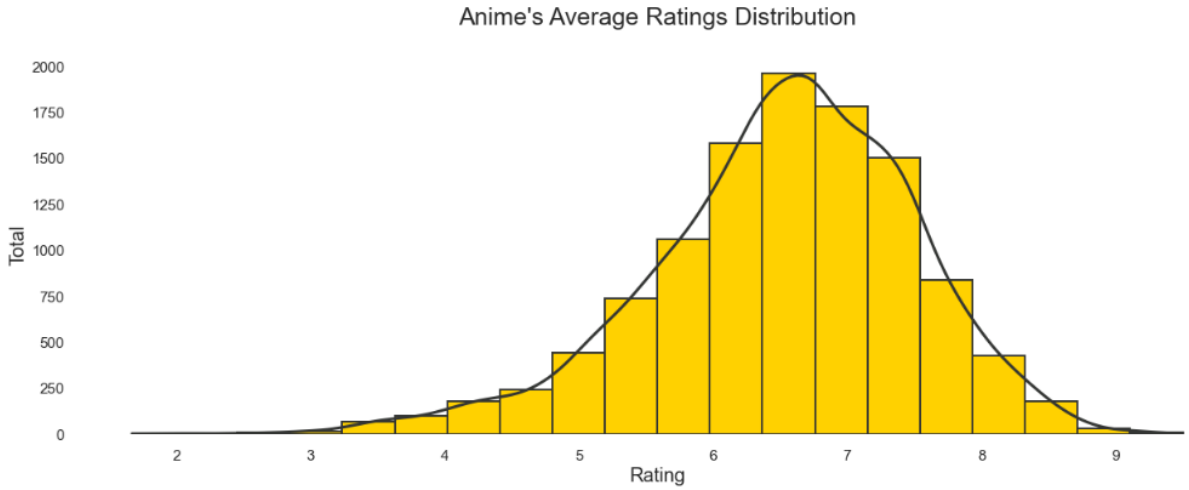
- Movie: 2111
- Special: 1581
- ONA: 526
- Music: 451




- **TV** là thể loại Anime chiếm số lượng nhiều nhất chiếm 30.48%.

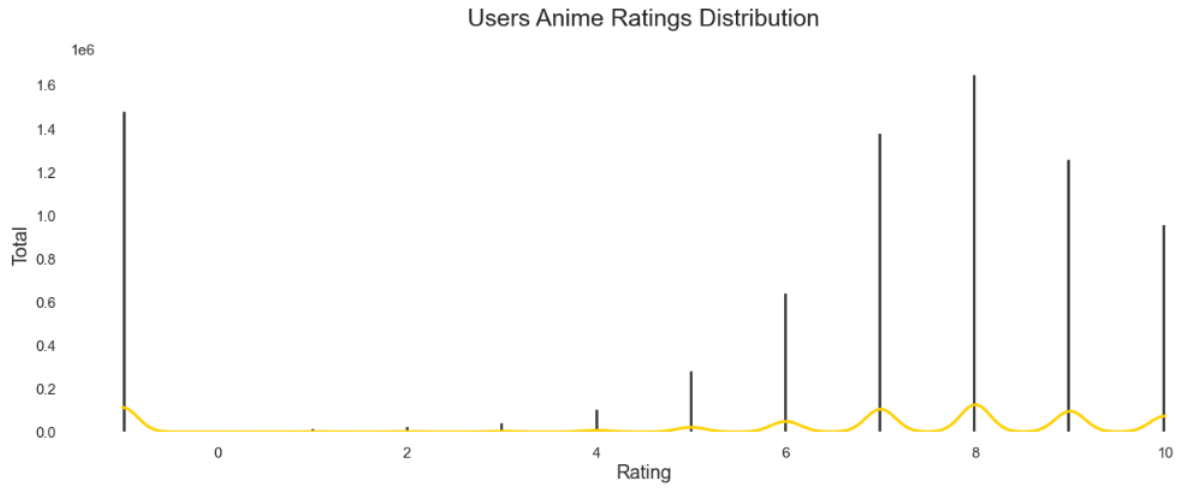
## C. Thống kê dữ liệu:

- Thống kê sự phân phối điểm trung bình của Anime như sau:



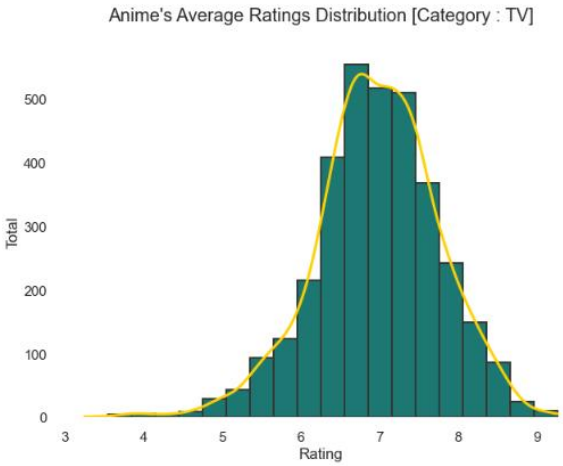

- Đồ thị phân phối điểm đánh giá trung bình của các Anime.

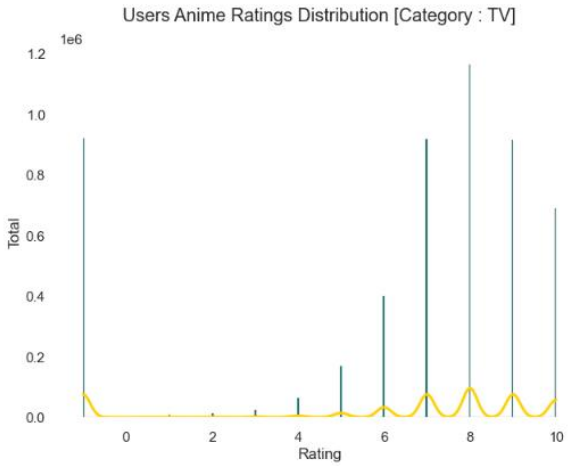
- Từ đồ thị ta thấy phần lớn các Anime có điểm đánh giá ở mức 6-7 điểm, đây là mức điểm nằm mức độ trung bình trong thang điểm 10
- Và mức điểm 8-9 tuy ít hơn nhưng vẫn chiếm một lượng đáng kể.
- Và mức điểm dưới 4 là mức độ dưới trung bình thì chiếm số lượng rất ít.




- Đồ thị phân phối tổng số đánh giá Anime ở từng mức điểm khác nhau:
  - Từ đồ thị ta thấy phần lớn các Anime có điểm đánh giá ở mức 8-9 điểm, đây là mức điểm khá cao trong thang điểm 10
  - Với mức độ điểm 10 thể hiện sự đặc biệt yêu thích của người xem nhưng vẫn chiếm một số lượng khá lớn
  - Và với mức điểm dưới 5 lại chiếm một số lượng đánh giá rất nhỏ so với các mức điểm ở trên.
    - Từ đồ thị, ta thấy rằng Anime được rất nhiều người xem yêu thích và có số lượng đánh giá tích cực chiếm đa số.

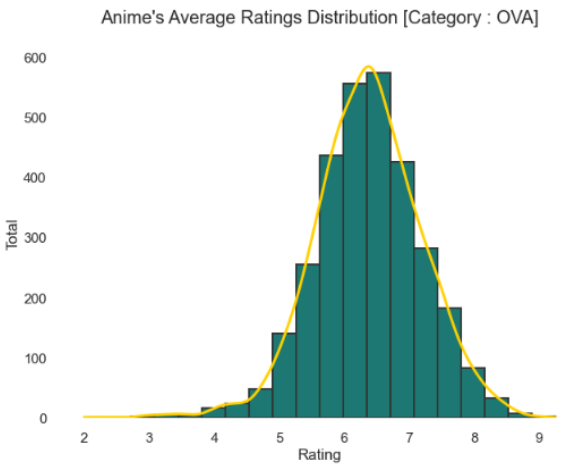
- Đồ thị thống kê sự phân phối điểm trung bình của Anime và phân phối tổng số đánh giá Anime của người xem với từng ***thể loại Anime*** khác nhau:
  - TV:

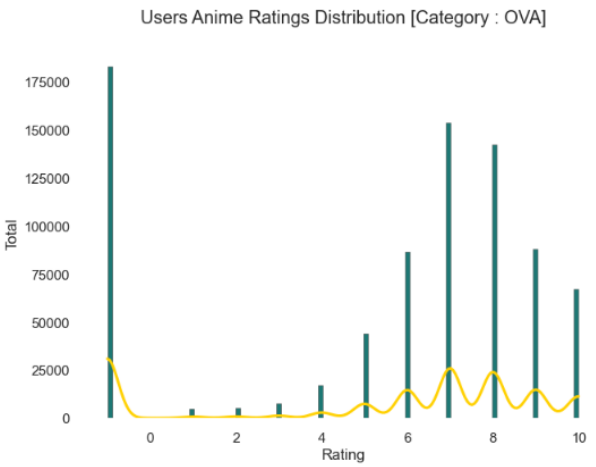



- Thể loại TV có được phần lớn đánh giá tích cực với mức đánh giá tập trung chủ yếu ở mức trên 7 điểm thể hiện đây là thể loại Anime được rất nhiều người xem yêu thích.

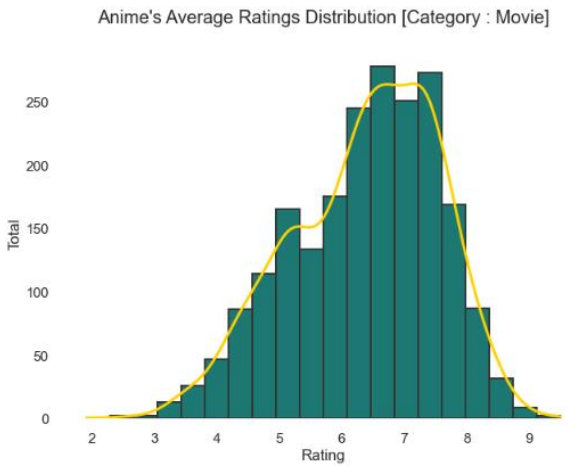
- OVA:

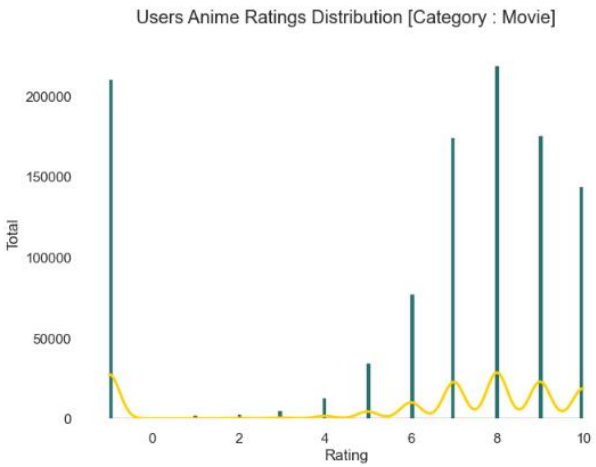



- Thể loại OVA tuy cũng có phần lớn đánh giá tích cực. Tuy nhiên, mức điểm tập trung nằm ở khoảng mức 6. Suy ra, thể loại Anime OVA tuy cũng khá được yêu thích nhưng mức độ yêu thích của người xem thấp hơn so với thể loại TV.

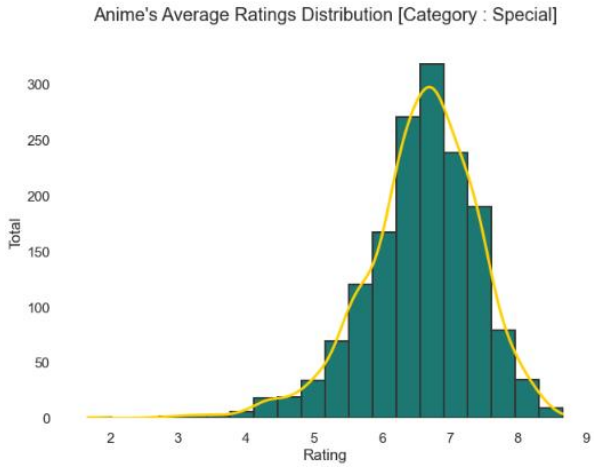
- Movie:

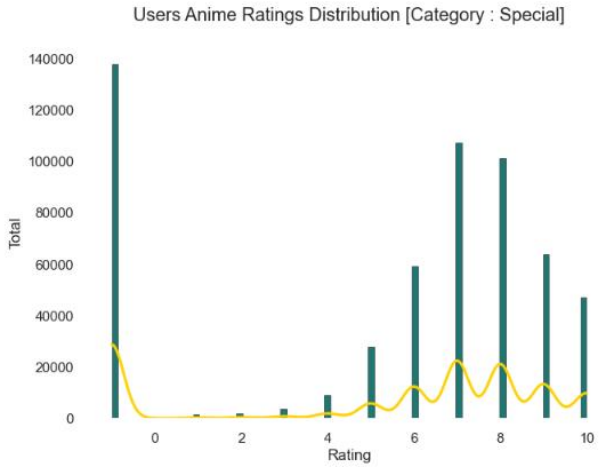



- Thể loại Movie cũng có mức độ được người xem yêu thích tương tự như thể loại OVA khi có mức độ đánh giá khá tương đồng với nó.

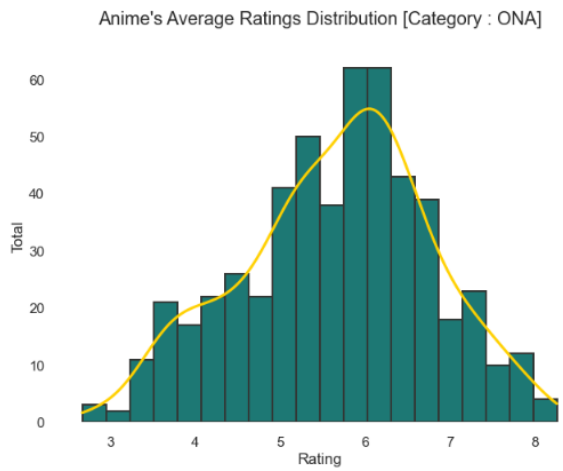
- Special:

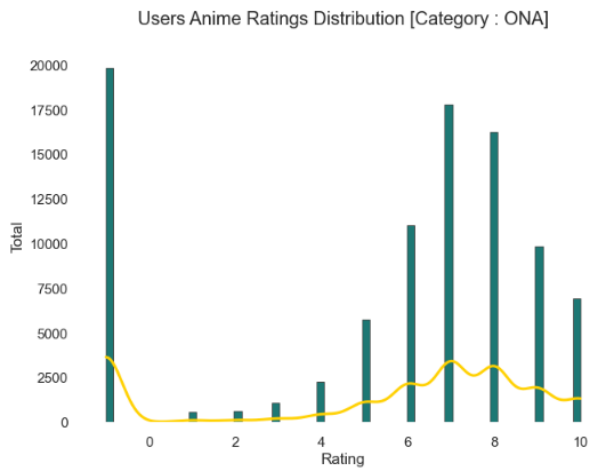



- Thể loại Special cũng được yêu thích ở mức trung bình. Điểm tích cực ở thể loại này là có rất ít đánh giá ở mức dưới trung bình.

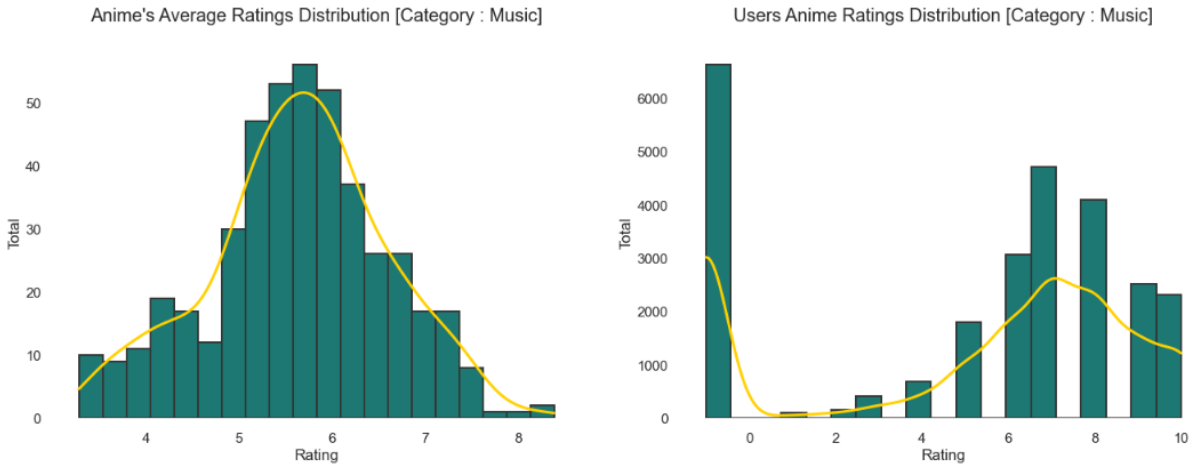
- ONA:



- Thể loại ONA có đồ thị khá cân đối, tập trung chủ yếu vẫn ở mức điểm trung bình khoảng 6 điểm. Đồ thị cho ta thấy đây là thể loại phim có nhiều đánh giá kém và cũng nhiều đánh giá tốt. Đây là thể loại phim khá phức tạp, đa dạng, có nhiều phim Anime hay và dở thuộc thể loại này.Nguyên nhân khách quan có thể do tuỳ vào tích cách của người xem, có những người sẽ thích thể loại này, nhưng nó có thể gây phản cảm, chán ghét với người xem khác. Đây có thể trở thành vấn đề khó khăn cho việc đưa ra gợi ý tốt với thể loại ONA.

- Music:




- Thể loại Music không được quá yêu thích so với các thể loại còn lại. Mức điểm tập trung chủ yếu chỉ ở mức 5-6 điểm, mức điểm khá thấp. Đặc biệt mức điểm dưới trung bình chiếm khá cao và có rất ít đánh giá với mức độ trên 8. Chứng tỏ thể loại Anime Music là thể loại được ít người xem yêu thích nhất.

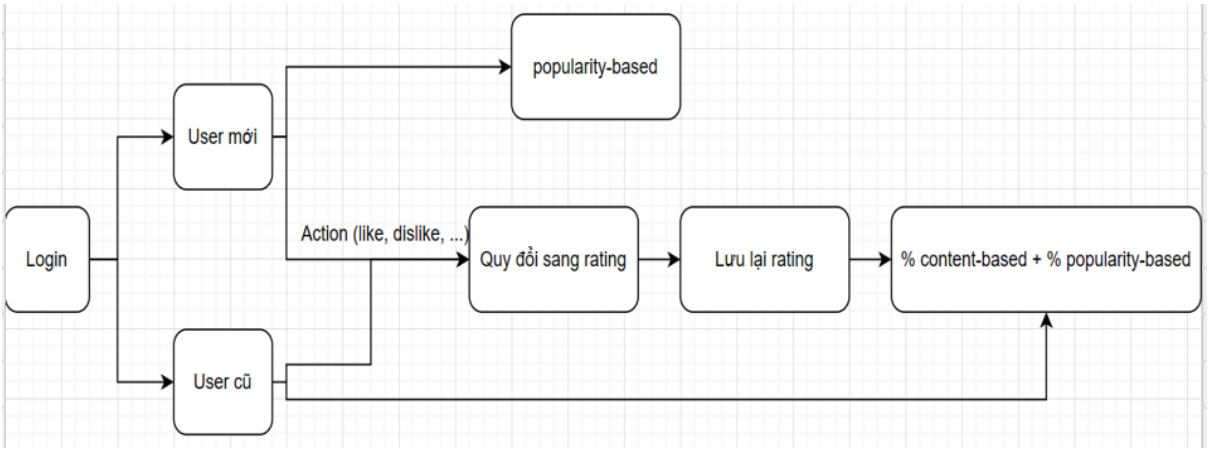
## D. Tiền xử lý dữ liệu:

- Với dataframe của Anime ta có 12294 giá trị với tổng cộng 7 cột trong dataframe và sau khi loại bỏ khỏi dataframe cột "Type" và cột "Episodes" cùng với các giá trị NaN ta còn 12017 giá trị cùng với 5 cột dữ liệu là: anime_id, name, genre, rating,members.
- Với dataframe của Rating ta cũng áp dụng phương pháp thay thế các giá trị -1 bằng Nan và loại bỏ chúng khỏi dataframe. Vì những giá trị -1 đại diện cho những bộ anime mà người xem không đánh giá, nên ta có thể coi nó như giá trị NaN và loại bỏ nó.
- Khởi tạo hàm Clean dùng để làm sạch tên các bộ anime bằng cách loại bỏ các kí tự đặc biệt, loại bỏ các dấu câu để có tên bộ phim phù hợp.

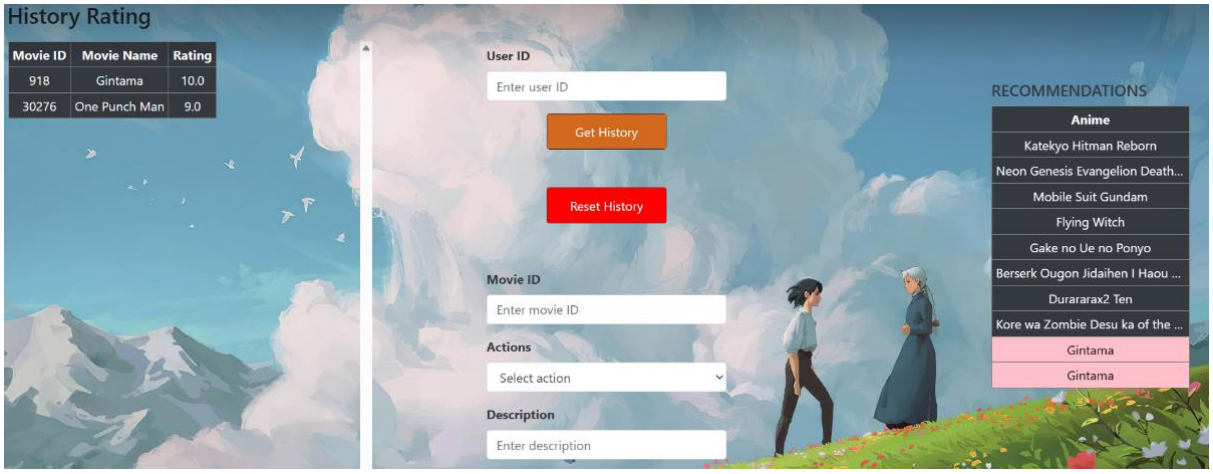
## E. Xây dựng hệ thống:

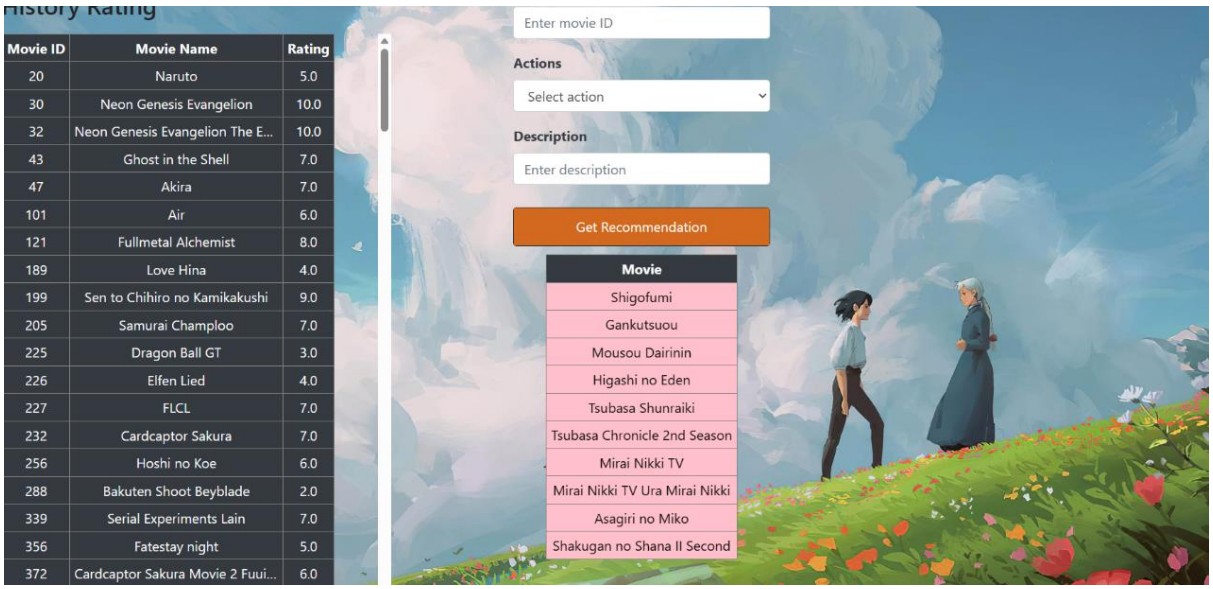
- Cả lọc dựa trên nội dung và lọc cộng tác đều là những thuật toán đề xuất rất phổ biến .
- Với Hệ thống Lọc Cộng Tác Dựa Trên Người Dùng và Dựa Trên Sản Phẩm chúng ta có một số nhược điểm đáng quan tâm sau:
    - Cold-start for new users: người xem chưa có bất kì đánh giá nào trong hệ thống.
    - New-item problem: Khó đưa ra đề xuất cho các bộ phim mới ra mắt.
    - Sparcity: ma trận tương tác giữa user và item ít, có nhiều giá trị 0.
    - Transparency: khó giải thích rõ ràng với các bộ phim được đề xuất
- Chúng ta sử dụng Hệ Thống Gợi Ý Dựa trên Nội Dung vì sử dụng nội dung từ các phim để đưa ra đề xuất bộ phim phù hợp.
- Phương pháp này giải quyết được các nhược điểm từ 2 hệ thống trên.
- Ưu điểm của Hệ Thống Gợi Ý dựa trên nội dung:
    - Mô hình này không cần bất kỳ dữ liệu nào về những người dùng khác vì các đề xuất chỉ dành riêng cho người dùng này. Chúng ta dễ dàng mở rộng quy mô cho nhiều người dùng khác nhau.
    - Mô hình này có thể nắm bắt các mối quan tâm cụ thể của người dùng, đồng thời có thể đề xuất các mục thích hợp mà rất ít người dùng khác quan tâm.
- Nhược điểm của Hệ Thống Gợi Ý dựa trên nội dung:
    - Mô hình chỉ có thể hoạt động tốt như những tính năng được thiết kế thủ công.
    - Mô hình này chỉ có thể đưa ra đề xuất dựa trên mối quan tâm hiện có của người dùng.
    - Hạn chế khả năng mở rộng mối quan tâm với các đề xuất khác.

# F. Hệ Thống Tư Vấn Anime:




- Hệ thống tư vấn Anime được xây dựng với dựa trên cấu trúc như trên. Xét trường hợp áp dụng với cả người mới và người dùng cũ để đưa ra hành động.
- Giao diện của hệ thống tư vấn Anime được xây dựng trên nền tảng Web:



# IV. Kết luận

- Hệ thống sau khi hoàn thành đã có thể đưa ra những để xuất các bộ Anime cho người dùng mới và những người dùng cũ.
- Hiện thị được những bộ Anime mà người dùng cũ đã đánh giá trước đó và số điểm đại diện cho mức độ yêu thích của họ với bộ Anime đó.
- Tuy nhiên, không có hệ thống tư vấn nào là hoàn hảo cả và hệ thống của nhóm chúng tôi thực hiện cũng vậy. Chỉ có thể đưa ra những đề xuất Anime tốt nhất có thể dựa trên bộ dữ liệu với những dữ liệu đã có. Vì với Anime hay bất kì thứ gì, để có thể đề xuất tốt nhất thì cần rất nhiều dữ liệu khác của một cá nhân nói riêng và tập thể nói chung.
- Tiêu biểu nhất là những trang Web hay ứng dụng lớn như Youtube, Tiktok,Facebook họ đều có những khảo sát nhanh chóng để xác định hệ thống đề xuất của họ có tốt với mỗi người dùng hay không.  Vì vậy, nhóm chúng tôi cũng thêm một tính năng dùng với mục đích khảo sát người dùng có hài lòng với những đề xuất mà hệ thống đưa ra không, nhằm cải thiện hệ thống cho sau này.

# V. Tham khảo


- https://github.com/ucalyptus/Spotify-Recommendation-Engine?fbclid=IwAR062AwoT5TB60_lO1vTueESXsRnYbjvsi-C6XHBimzToQbQsarlxtVz7Fw
- https://www.kaggle.com/code/hasibalmuzdadid/anime-ratings-analysis-recommender-system/input?fbclid=IwAR108F5_4lH_1l7RC817BtDG0QSmwd1QGJ8FYGjGHzy3WOsAvU-BUGDriYU
- https://rstudio-pubs-static.s3.amazonaws.com/512303_1411d10912b848d79779ed7d539444e3.html