# Anime Recommendation System

Wan Qi Chen
Victoria Clarchick
Abhibha Gupta

# Motivation

Group's personal interests in the different anime genres

Looking into what different series would align with their interests

Get recommendations from friends or create a recommendation system that through data mining can provide a recommendation

# Goal

Spread the fun of anime to more people

Encourage the competition of the animation industry, by using a ranking system throughout different anime

Recommend to people which anime they might like or fit most depending on their preference of genre of anime

# Approach

Explore Data

- Convert categorical features into a numerical format
- Standardize and scale numeric data
- Find correlation between variables and try to identify the important ones

Incorporate Important features before making predictions

Implement a User based collaborative filtering (UBCF) recommender system

We use a ‘hybrid’ recommender model that gives weightage to 3 types of models namely, popularity based model, Random based model and Re-recommendation model.

We experiment with different dataset sizes and evaluate the algorithm using RMSE, MAE and MSE

- The model predicts a ‘rating’ which is compared to actual rating

Innovative Idea: Use Recommenderlab package offered by R

# Data Overview

Anime.csv

General information of every anime (17,562 different anime) like genre, stats, studio, etc

Rating_complete.csv

CSV that contains the list of ratings given by the user to animes with watching_status==2 (complete). This dataset contains 57 Million ratings applied to 16,872 animes by 310,059 users.

# Data Exploration

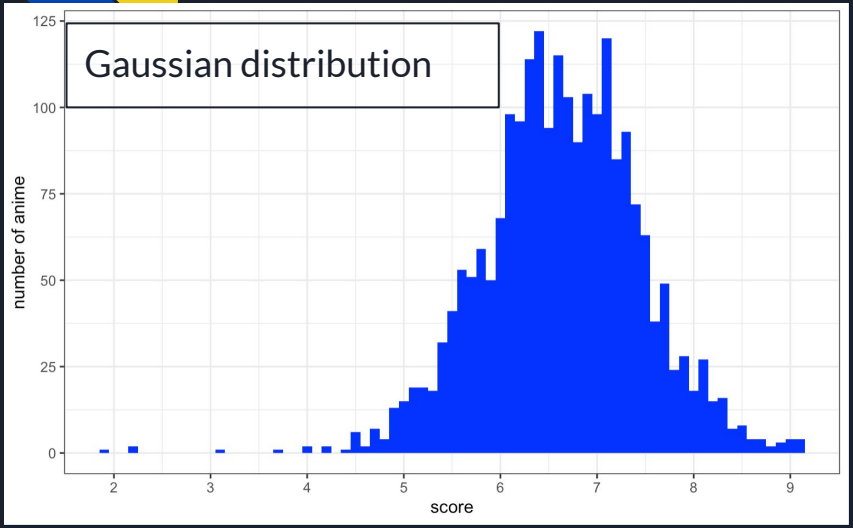
Look into correlations through a Cor plot
Working with the entries that are associated with a rank between 1-1000
Effect on Rank from users in the category of:

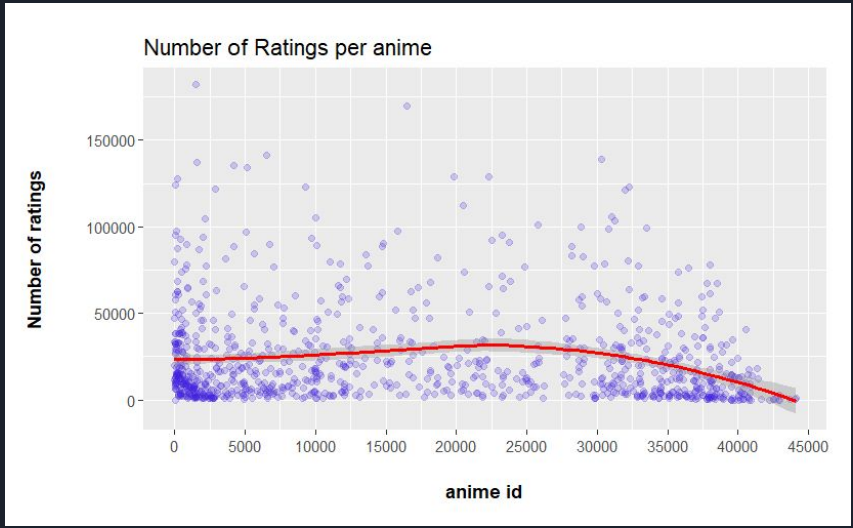
- Genre
- Amount of users that rated
- Studio
- Producer
- Licensors
- Source
- Type
- Stretch to view regardless of rank

Effect on Rank from users in the category of:

- Source
- Type

# The distribution of the all anime



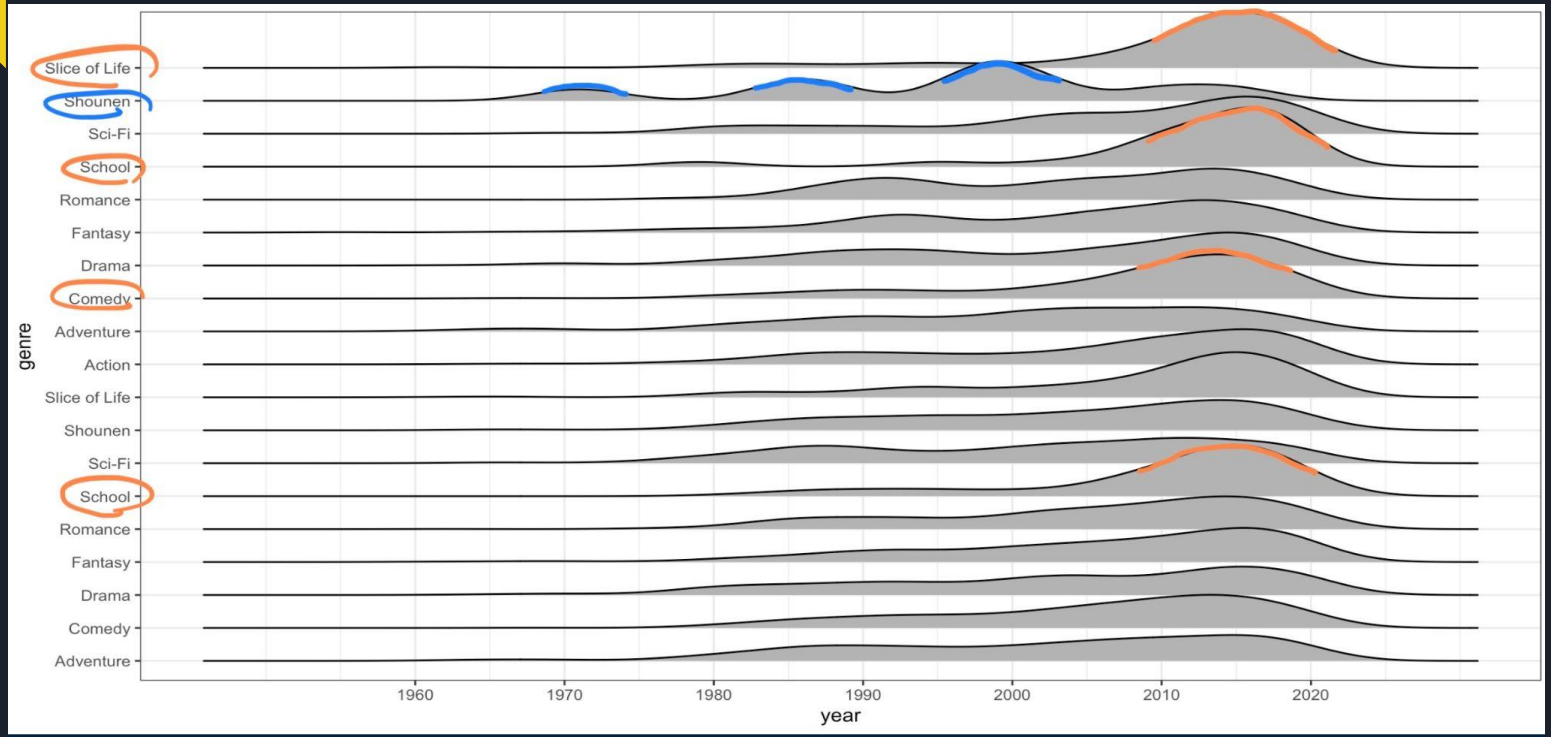




The distribution of all the animes to scores is normal distribution.
As a result, we could confirm the score is not distributed extremely.

Out of the top 1000 ranked anime most have been rated by aprox 25000 users so the average rating is well distributed

The most common ratings from users are 8 followed by 7 so users are more likely to rate high

# Genre Distribution through these years (Top 10)



# Cor Plot

Proportionate:

Popularity/Rank

Favorite/Score

Score/Member: slight

Watching/Members

Disproportionate:

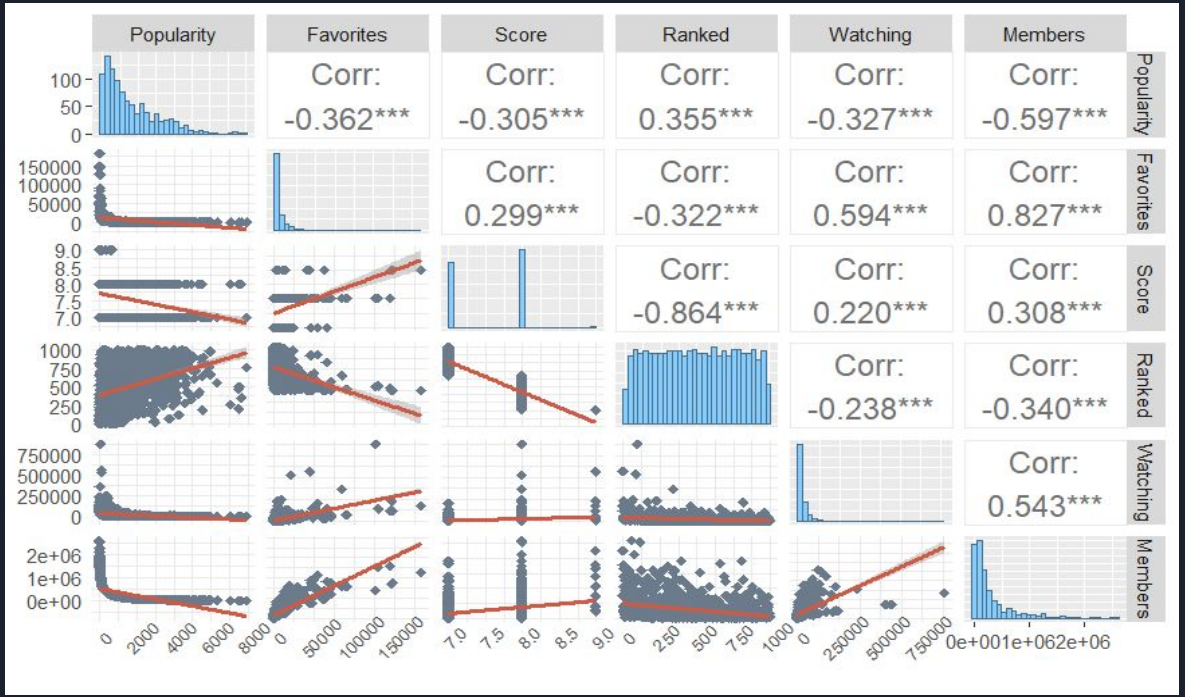
Popularity/Members: slight

Popularity/Score

Favorites/Rank

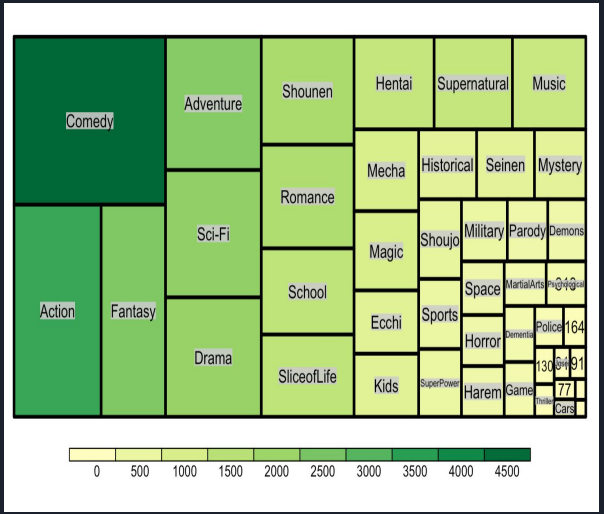
Score/Rank

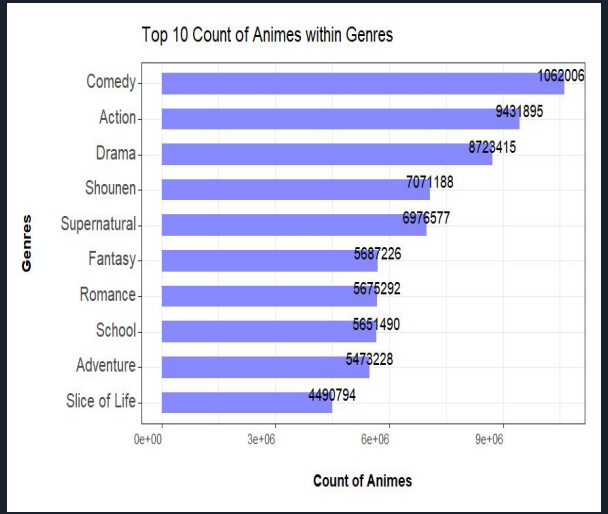
Ranked/Members: slight

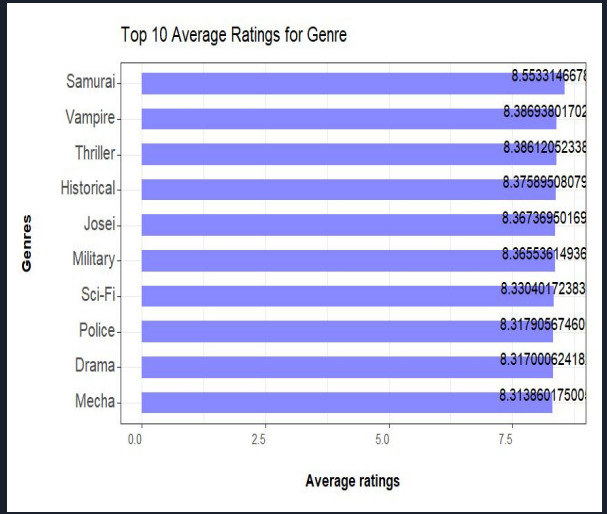



Important Variables: Score and Rank

# Genre Background



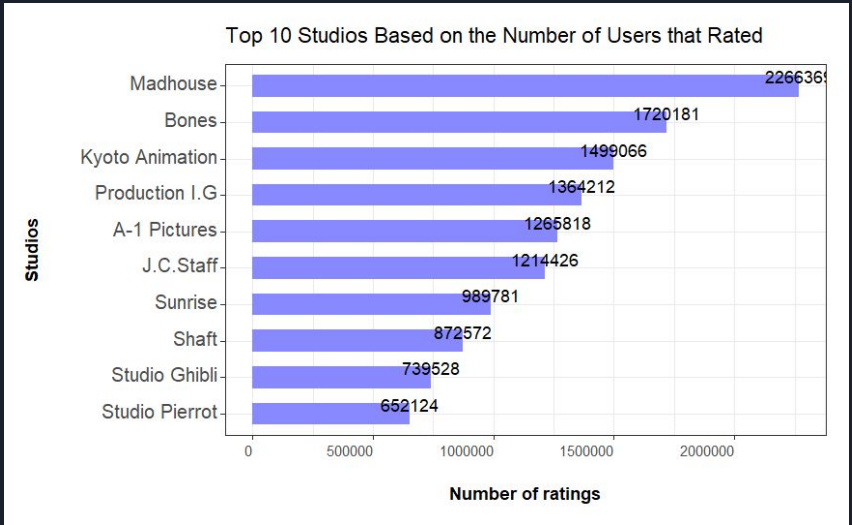



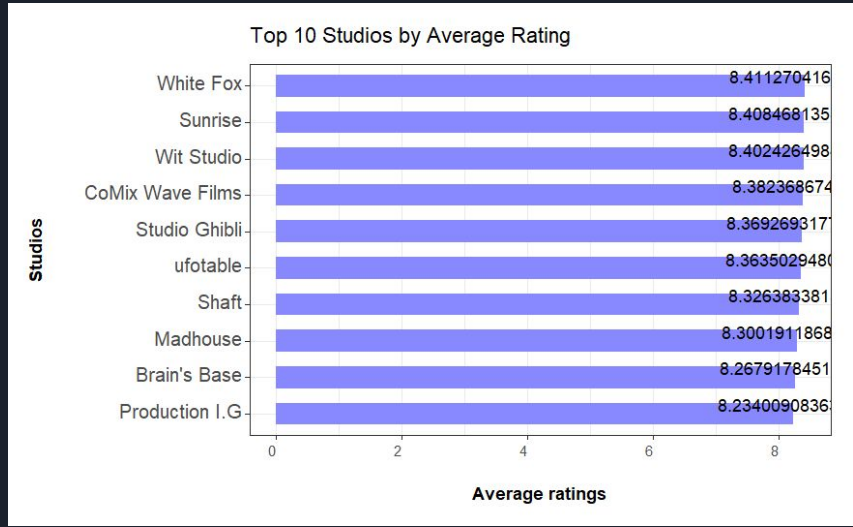




There is at least 1 Genre that overlaps between the most watched Genres and the average ratings for a Genre

This shows that Genres that are most watched are not highly rated

# Studios Background



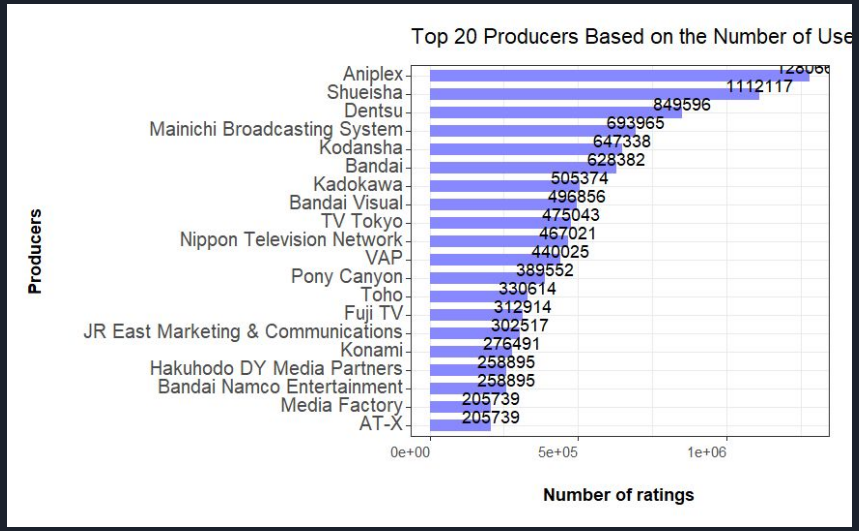



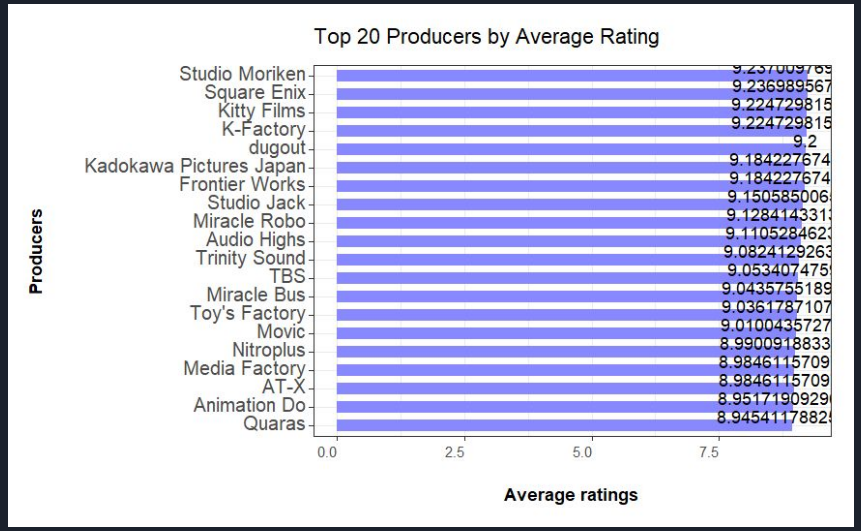

Out of the top 10 from each category there are 4 Studios that overlap

The higher rated studios are watched less

Therefore there are a few commonalities showing that studios that mass produce do not have high quality ratings

# Producers Background

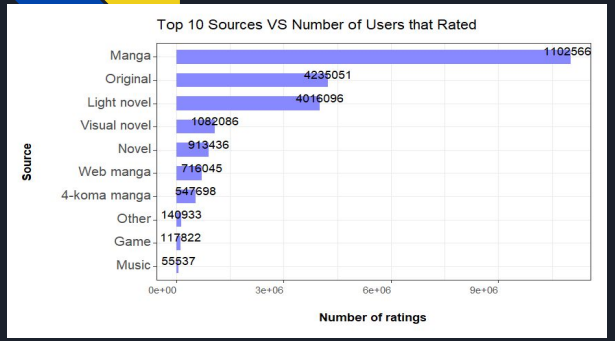



Out of the top 20 from each category there 1 common Producer

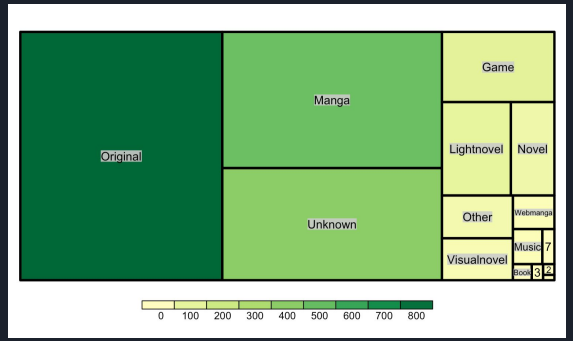
Shows that there are many producers but with the top ranked very few correspond to high ratings so the most common producer do not produce high ratings

# Source Background



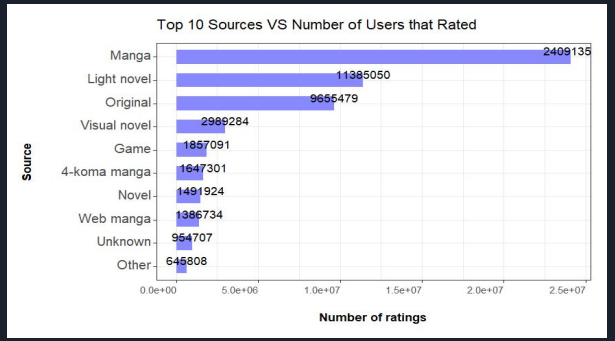

Ranked 1-1000

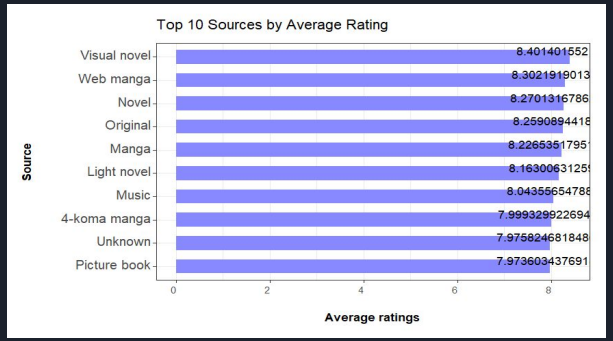
3 common Sources

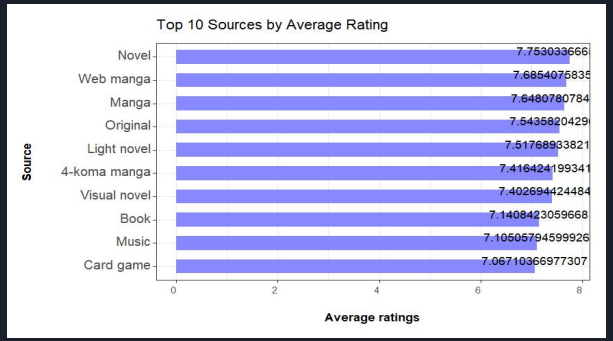



Entire DataSet

6 common Sources

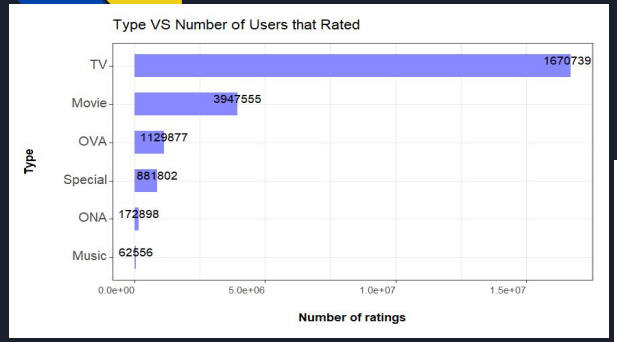


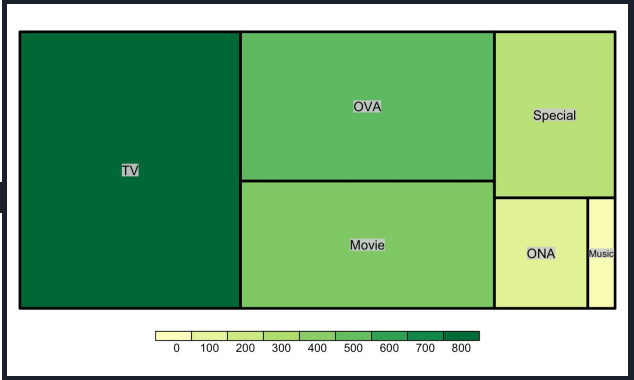

THis indicates that sources that are watched most often do not have high ratings

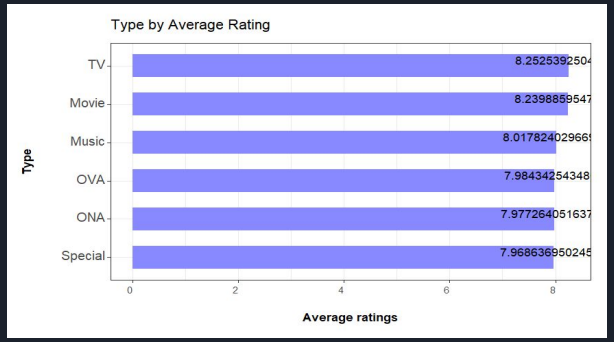
# Type Background

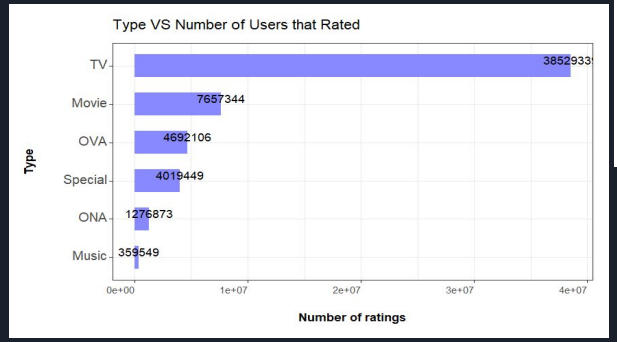



Ranked 1-1000

top 2 are the top 2 for both most rated and highest ratings

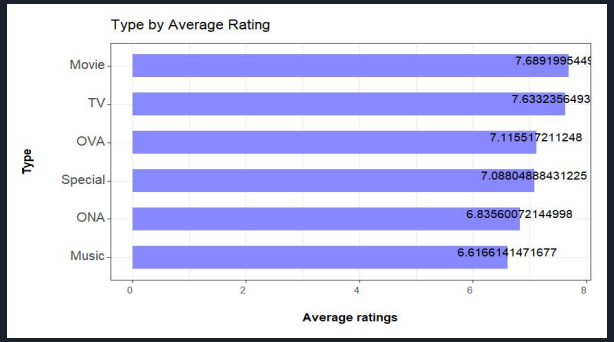



Entire DataSet

the top 2 are the only changes between most rated and highest ratings




This means that users will watch TV much more than Movies but the rating is almost the same

# Results

Table 1: Baseline results

| | Instances | | | | | |
|---|---|---|---|---|---|---|
| Metrics | | 10^2 | 10^3 | 10^4 | 10^5 | 10^6 |
| | RMSE | 1.960 | 3.027 | 3.610 | 3.786 | 3.73 |
| | MSE | 3.844 | 9.163 | 13.038 | 14.334 | 13.935 |
| | MAE | 1.674 | 2.493 | 2.940 | 3.086 | 3.049 |

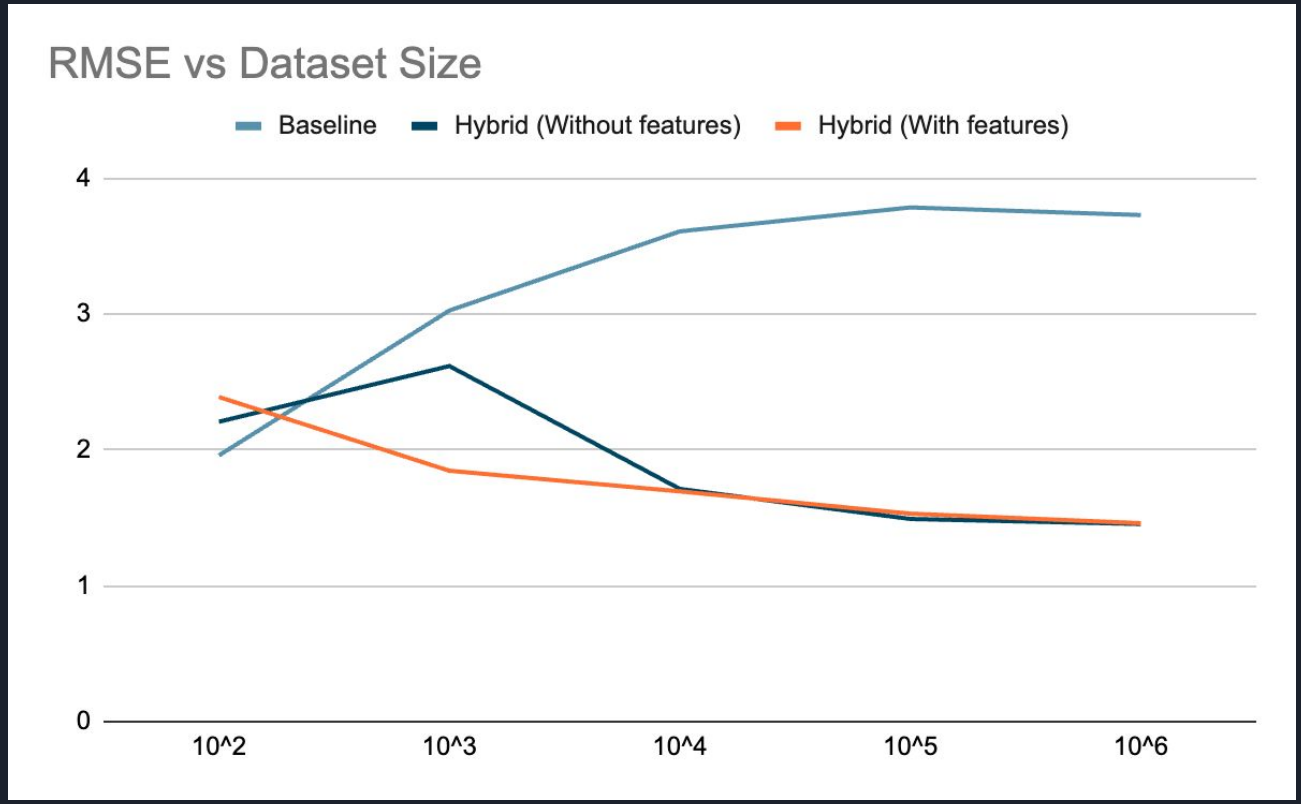
Table 2: Hybrid recommender model (Without features)

| | Instances | | | | | |
|---|---|---|---|---|---|---|
| Metrics | | 10^2 | 10^3 | 10^4 | 10^5 | 10^6 |
| | RMSE | 2.208 | 2.618 | 1.713 | 1.492 | 1.455 |
| | MSE | 4.876 | 6.855 | 2.937 | 2.225 | 2.118 |
| | MAE | 1.941 | 1.828 | 1.271 | 1.136 | 1.133 |

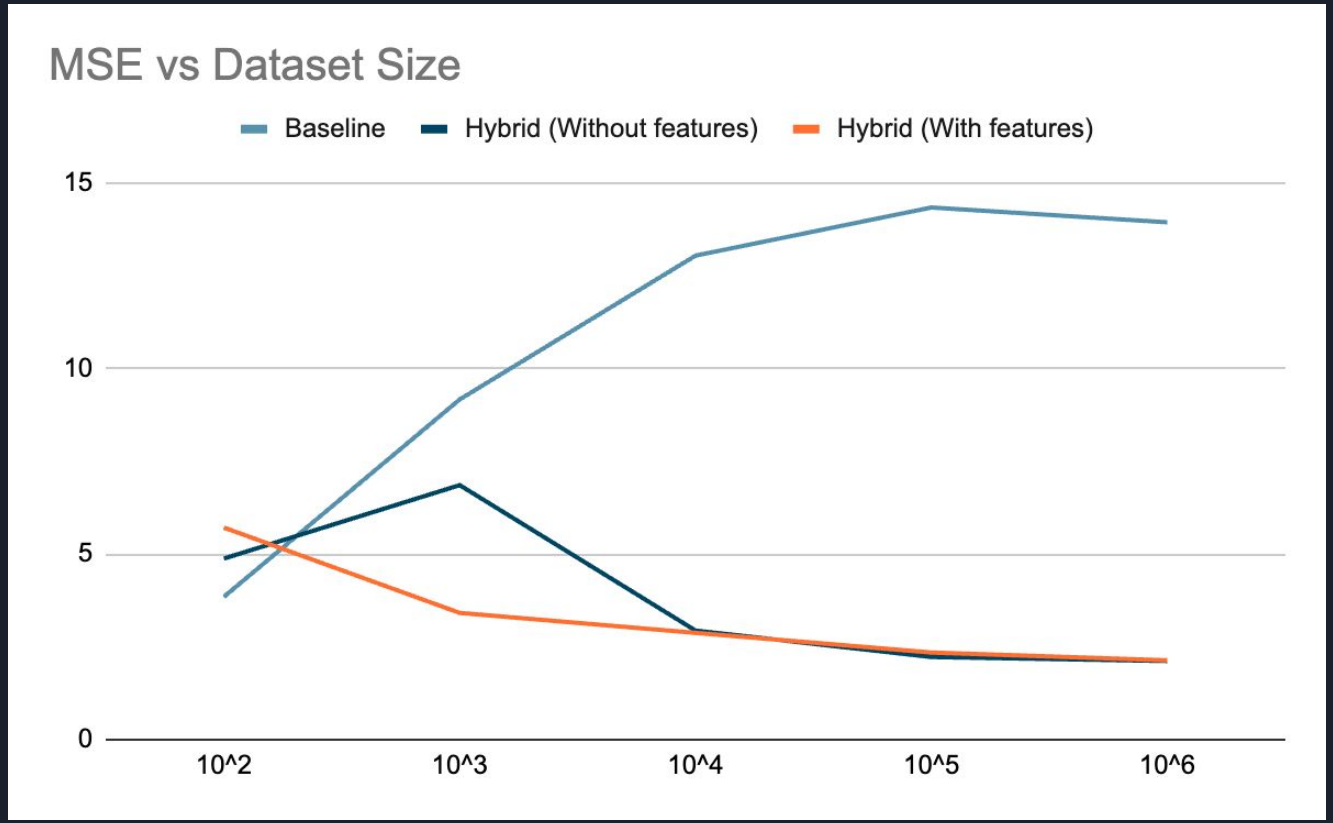
Table 3: Hybrid recommender model (With features)

| | Instances | | | | | |
|---|---|---|---|---|---|---|
| Metrics | | 10^2 | 10^3 | 10^4 | 10^5 | 10^6 |
| | RMSE | 2.390 | 1.847 | 1.695 | 1.531 | 1.461 |
| | MSE | 5.712 | 3.413 | 2.875 | 2.346 | 2.135 |
| | MAE | 2.016 | 1.314 | 1.182 | 1.156 | 1.138 |

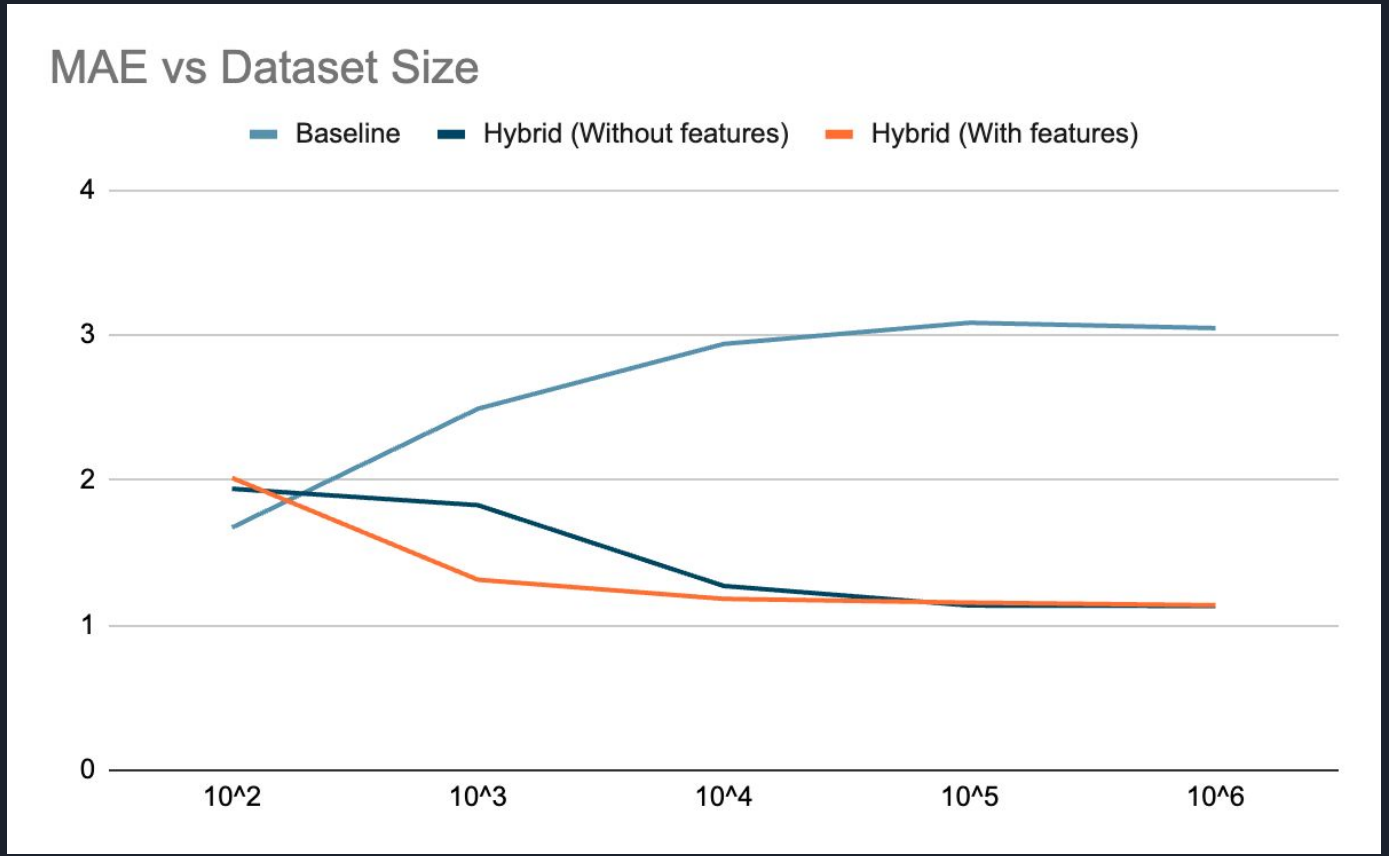
# Results: Random Recommender Model (Baseline)



# Results: Hybrid Recommender Model without features



# Results: Hybrid Recommender Model with Features.



# Discussion

The baseline was when the model was set to ‘Random’ mode. The Hybrid model performed better than the baseline model.

Increasing the number of instances reduces the RMSE but only to a certain extent. After 10^5 instances the RMSE doesn’t change much in both cases.

After this point probably feature addition is required.

From prior data exploration, we found that the ‘Score’ and ‘Rank’ of an Anime are important factors. We incorporated these features while training the recommender system

From the graphs it can be observed that adding extra features is not helpful as the RMSE value increased in most of the cases.

# Future Work

Real life deployment: Try to deploy the system in real life for fellow anime fans. Or maybe use the same algorithm for other use cases like movie recommendation.

Improving model performance: We could tune hyperparameters (Ex: Similarity metrics) to come up with a more accurate model.

Personalized recommendations: Data exploration has revealed that popular shows are not always highly rated. We could incorporate user data( like reviews)  to provide niche recommendations catered to the user's specific tastes.

# THANK YOU!

# QUESTIONS?

## Reference

Valdivieso, Hernan. “Anime Recommendation Database 2020.” Kaggle, 13 July 2021, https://www.kaggle.com/datasets/hernan4444/anime-recommendation-database-2020/code?select=anime_with_synopsis.csv.