1 Introduction

Anime has grown from a niche subculture into a global phenomenon, influencing millions of viewers and shaping media trends across countries. Beyond entertainment, anime data offers a fascinating window into human preferences, cultural diffusion, and social behavior. Analyzing how users rate and interact with anime provides opportunities to understand patterns in collective taste, popularity dynamics, and even cross-cultural differences in media consumption.

Moreover, anime can be viewed not only as a collection of individual titles, but as a complex social network. Genres, studios, franchises, and shared audiences form clusters and hubs: tightly knit communities built around specific themes, styles, or narrative structures. This creates a rich environment for studying community formation, network effects in popularity, and the emergence of subcultures. In this sense, anime serves as a microcosm of larger social processes — how groups organize, how trends spread, and how cultural identities form and evolve.

From a data-scientific perspective, anime datasets present multiple challenges and opportunities, offering a rare combination of numerical, categorical, and textual information embedded in a naturally occurring, socially meaningful structure. This makes them well-suited for analyzing real patterns in human behavior through feature engineering and exploratory data analysis. They allow us to explore questions such as how different genres appeal to different audiences, how collective preferences change over time, how collective user behavior emerges from individual ratings, and how user communities structure themselves around shared interests.

In today's environment, where online platforms dominate media consumption, understanding patterns in user ratings and content characteristics has both practical and scientific value. Insights derived from anime data can inform recommendation systems, guide content creation, and serve as a microcosm for analyzing trends in larger entertainment ecosystems. Studying anime data helps illuminate the mechanisms of popularity, personalization, and cultural diffusion, providing insights that generalize far beyond this specific medium. In other words, anime is not only culturally significant, but also a compelling, real-world playground for developing and demonstrating data science methods.

For this project, we use the *Anime Recommendation Database 2020* from Kaggle[**?**] [1], which contains detailed information from **MyAnimeList** about over 16,000 anime titles and ratings from more than 300,000 users. This dataset is rich, heterogeneous, and somewhat messy — exactly the kind of data that makes data science exciting.

The key questions we aim to explore include [**?**]:

- Which anime are the most highly rated, and which are underrated compared to their popularity?
- How do user ratings vary across genres, release years, and demographics?

[1] https://www.kaggle.com/datasets/hernan4444/anime-recommendation-database-2020

- Can we identify clusters of similar anime based on genre, themes, and user ratings?
- Are there patterns in user behavior, such as tendencies to rate certain genres more favorably?

Through exploratory data analysis, we hope to extract meaningful insights, for example:

- Genre trends: which genres consistently attract high ratings.
- Temporal patterns: how ratings and popularity evolve over time.
- Relationships between anime features (genre, episodes, year) and user ratings.

However, the dataset also has inherent limitations. We cannot reliably infer causal relationships from these ratings, nor can we accurately predict the future success of an anime solely based on historical user ratings. Additionally, user bias, incomplete data, and differences in rating scales limit the conclusions we can draw.

By focusing on data cleaning, exploratory analysis, and interpretation, this project demonstrates how raw, real-world data can be transformed into actionable insights, while also highlighting the boundaries of what such data can tell us.

The analysis pipeline and code are provided in the companion notebook [**?**].

2 Dataset Overview

MyAnimeList (MAL) is one of the largest online platforms dedicated to anime and manga tracking. Users maintain personal lists of watched titles, assign numerical ratings, write reviews, and participate in community discussions.

Although MAL does not represent the entire global audience, it functions as a *large-scale, organically formed social network* structured around media consumption. Its importance for data analysis stems from several factors:

- it has millions of active users with voluntarily provided preference data;
- users form clusters and communities around genres, studios, eras, and specific titles;
- the rating patterns reflect collective taste dynamics, hype cycles, long-tail phenomena, and cross-cultural differences;
- the anime graph (users × titles) behaves like a sparse bipartite network with hubs, where highly popular titles act as central connectivity points.

For these reasons, MAL is a valuable source for studying recommendation systems, preference modeling, popularity prediction, and structural patterns of entertainment consumption.

2.1 Dataset Structure

We use the 2020 "Anime Recommendation Database" from Kaggle, which combines several processed dumps of MyAnimeList: [2]

anime.csv Metadata for approximately 17,000 titles, with columns including:

- identifiers: `MAL_ID`, `Name`, `English name`, `Japanese name`;
- quality and popularity metrics: `Score`, `Ranked`, `Popularity`, `Members`, `Favorites`;
- categorical descriptors: `Genres`, `Type`, `Source`, `Studios`, `Producers`;
- structural info: `Episodes`, `Duration`, `Rating` (age restriction);
- engagement counters: `Watching`, `Completed`, `Dropped`, etc.;
- per-score vote counts: `Score-1` ... `Score-10`.

anime_with_synopsis.csv A reduced version containing `MAL_ID`, `Name`, `Score`, `Genres`, `Synopsis`. Useful for NLP tasks such as clustering by textual description.

animelist.csv Contains approximately 300 million user–anime interactions, with columns: `user_id`, `anime_id`, `rating`, `watching_status`, `watched_episodes`. This is the core behavioral dataset representing the user–item matrix.

rating_complete.csv A filtered version of `animelist.csv` containing only rows with `watching_status = 2` ("Completed"), columns: `anime_id`, `user_id`, `rating`. Commonly used for training recommender systems.

watching_status.csv Lookup table pairing each integer code with a textual description ("Currently Watching", "Completed", etc.).

2.2 Strengths of the Dataset

Despite being collected from an entertainment platform, the dataset has several significant advantages for data science practice:

- **Large scale:** Tens of millions of ratings across thousands of titles enable analysis of long-tail distributions, user segmentation, genre-level statistics, and popularity dynamics.
- **Natural heterogeneity of users:** No incentive to game the system; tastes are diverse and clusters form organically.

[2] Where external statistics are used (industry reports, demographic data), and where it is possible, we use sources closest to 2020 to maintain temporal consistency.

- **Multiple complementary tables:** Metadata, textual features, and behavioral interactions allow content-based, collaborative, graph-based, and hybrid recommender analyses.
- **Excellent for methodological demonstration:** Useful for data cleaning, exploratory data analysis (EDA), imputation, outlier detection, recommendation algorithms, and metadata fusion with NLP.

2.3 Limitations and Potential Biases

The dataset reflects behaviors of a specific community and inherits biases from the platform:

- **Geographical bias:** MAL is most popular in North America, parts of Europe, and Southeast Asia. For example, China is one of the global leaders in anime licensing [?], but Chinese users are absent on MAL.
- **Cultural and language bias:** English-speaking communities dominate the dataset.
- **Sparse user–item matrix:** Most users have watched only a small fraction of all anime; this affects collaborative filtering model performance and cold-start dynamics.
- **Artifacts and inconsistency in dataset:** Example for `rating_complete.csv`: filtering by `watching_status = Completed` is not sufficient. Some users marked anime as completed but watched fewer episodes than the total and gave non-zero ratings. Since MAL does not allow rating = 0, zero ratings indicate *no vote* rather than dislike. These cases are rare (˜0.1%) but must be considered during cleaning.
- **No timestamps:** Limits temporal modeling of tastes and popularity.
- **Caution in interpreting popularity:** The dataset reflects preferences of *dedicated fans* rather than the general population. So, MAL ratings and engagement metrics do not directly translate to global popularity. MAL users are biased towards committed anime fans; casual viewers are underrepresented.
- **Licensing bias:** In many regions, anime consumption occurs primarily through piracy, so official licensing or platform metrics severely underestimate actual viewership[3].
- **MAL - only one of the biggest anime hubs:** Country-level popularity is not reliably inferable without additional corrections or external data sources; such adjustments are beyond the scope of this project.

[3] Even in the USA, the leader in licensed anime, nearly half of users watch primarily via unofficial services [?]

3 Synthetic User Generation

To simulate global user interactions while preserving privacy, we generate synthetic user profiles based on demographic and traffic data.

3.1 Data Sources

- **World cities and populations:** Dataset [?]. Used columns `ASCII Name, Country Name EN, Population, Latitude, Longitude`.
- **MyAnimeList traffic per country:** Scraped manually from semrush.com for October 2025, with columns `Country, Number of Visitors`.
- **Age and sex distributions:** From demographic studies[?] providing mean and standard deviation for age, and male/female proportions.

Additionally, the demographic survey[?] collected country-level data, but it was conducted at a large anime convention, so the sample is a *convenience sample* and not representative of the general population. Nevertheless, the observed distributions roughly align with the traffic patterns obtained from semrush.com.

The survey also collected additional information, such as self-reported life satisfaction, hobbies, gender, preferred decades of favorite titles, and other personal attributes. These variables are not directly used in this project, as our analysis relies on synthetic profiles generated from traffic and demographic distributions.

3.2 Generation Procedure

1. Compute country-level user proportions based on MAL traffic:

$$P_{\text{country}} = \frac{\text{MAL users in country X}}{\text{Total MAL users}}.$$

2. Distribute users to cities within each country proportionally to city population:

$$P_{\text{city}} = P_{\text{country}} \cdot \frac{\text{City population}}{\sum \text{City populations in country}}.$$

3. Sample user attributes:
   - `age` $\sim$ Normal distribution (mean, std from [?]),
   - `sex` $\sim$ Bernoulli(p) (male/female proportion from [?]),
   - `latitude, longitude` assigned according to the selected city.

4. Assign `user_id` sequentially and compile all attributes into `profiles.csv` with columns:

```
user_id, country, city, age, sex, latitude, longitude.
```

3.3 Notes

- This synthetic population allows testing recommendation algorithms and demographic analyses without revealing actual user data. - The proportions reflect December 2025 traffic, but are applied to simulate the 2020 dataset contextually.

4 Dataset and Initial Exploration

For the analytical part of this project, we use the *Anime Recommendation Database 2020* from Kaggle, which combines metadata about anime titles with large-scale user rating data. The dataset contains over 12,000 anime entries and more than 73,000 users, resulting in approximately 7.8 million individual rating interactions. This structure naturally splits the data into two main tables:

- **Anime information:** title, genres, type (TV, movie, OVA), number of episodes, release year, studio, popularity metrics, and short descriptions.
- **User ratings:** user identifiers and their numeric ratings for specific anime.

At first glance, the data appears rich and diverse, but it also reflects several typical characteristics of real-world datasets:

- **Incomplete records:** many anime lack a release year, genre tags, or episode counts.
- **Inconsistent categorical data:** genre lists differ in formatting and ordering; some entries use non-standard tags.
- **Long-tailed distributions:** a small number of highly popular anime dominate ratings, while the majority receive very few.
- **Sparse user behavior:** most users rate only a tiny fraction of available titles.

These properties are not defects; they represent the typical landscape of large, user-generated media datasets. They also motivate several of the analytical directions in this project, such as identifying rating patterns, studying genre clusters, and modeling the structure of user–anime interactions.

Synthetic Users and Locations

The original dataset does not contain geographic or demographic information about users. To explore cross-cultural and regional patterns, we generate synthetic user metadata. Each user is assigned a plausible location (country and optionally city), following real-world population distributions. This augmented dataset allows us to ask new types of questions, such as whether certain genres correlate with specific geographic regions, or whether user communities cluster differently across countries.

The synthetic data is clearly separated from the original records and is used only for exploratory purposes, without affecting the underlying rating matrix.

Data Cleaning

Before we can meaningfully analyze the data, we must resolve the inconsistencies and structural issues inherited from the raw dataset. This involves:

- normalizing genre representations and splitting multi-genre fields;
- removing or correcting obviously invalid entries (e.g. anime with zero episodes released in the 1800s);
- handling missing values through imputation or category-specific defaults;
- joining anime metadata with synthetic user information to form a unified analytical table;
- reducing noise in the user–anime interaction graph by filtering out extremely sparse users or entries.

These preprocessing steps create a clean, analyzable foundation for the exploratory data analysis that follows and ensure that all subsequent insights reflect meaningful patterns rather than artifacts of data collection.

5 Random-Walk Model for User Trajectory Simulation

5.1 Motivation

User activity on a large interaction graph can be interpreted as a sequence of transitions between nodes (e.g. items, topics, or communities). Given such a sequence for each user, our goal is to construct a probabilistic model that captures the *structural tendencies* of user navigation. This model is later used to generate synthetic trajectories—“random walkers”—that approximate the observed behavior of the real user. The ensemble of walkers provides a natural way to measure how typical or atypical a given user trajectory is, relative to the structure of the graph.

Because the underlying graph is large (on the order of thousands of nodes and millions of edges), all computations must be local and efficient. We exploit the fact that the graph evolves year by year, so a user trajectory is implicitly aligned with a sequence of yearly graphs.

5.2 Definition of the Random Walker

Let $G_t = (V_t, E_t)$ denote the interaction graph in year t. For a user u we observe a trajectory

$$\mathbf{x}^{(u)} = (x_0, x_1, \dots, x_T),$$

where $x_t \in V_t$ is the node visited by the user in year t. We construct a *random walker* whose behavior in year t is governed only by the local structure of G_t and the user's starting point x_0.

Formally, for each year t the walker occupies a state $X_t \in V_t$. Conditioned on $X_t = v$, the walker chooses its position in year $t+1$ according to a probability distribution over the neighbors of v in G_t:

$$\mathbb{P}(X_{t+1} = w \mid X_t = v) = \frac{1}{\deg_{G_t}(v)} \quad \text{for all } (v, w) \in E_t.$$

That is, the walker performs a uniformly random step along the edges that exist in the corresponding yearly graph.

A single random walker generates one synthetic trajectory

$$\mathbf{Y} = (Y_0, Y_1, \dots, Y_T), \qquad Y_0 = x_0.$$

To model uncertainty and to obtain stable statistical estimates, we simulate an ensemble of K independent walkers for each user.

5.3 Asynchrony and Year-Level Dynamics

A key detail is that the walkers evolve *asynchronously.* Each walker only moves when the global simulation clock advances to a year in which that walker still has remaining steps. This design is necessary because real user trajectories can have different lengths, and the yearly graphs G_t may differ substantially in size and connectivity.

Thus the simulation proceeds by iterating over years $t = 0, 1, \dots, T$ and, for each walker whose trajectory length is greater than t, performing exactly one step in G_t. Walkers whose length is shorter than the current year simply remain inactive.

5.4 Ensemble-Based Evaluation

Given a user u with observed trajectory $\mathbf{x}^{(u)}$ and an ensemble of simulated trajectories $\{\mathbf{Y}^{(k)}\}_{k=1}^{K}$, we can quantify how well the random-walk model explains the user's behavior.

Let $d(\cdot, \cdot)$ be a similarity or distance measure between two trajectories. In this work we primarily use a weighted node-overlap metric that penalizes long-distance mismatches. The average similarity of the user to the ensemble is

$$\bar{s}^{(u)} = \frac{1}{K} \sum_{k=1}^{K} s\big(\mathbf{x}^{(u)}, \mathbf{Y}^{(k)}\big),$$

with an accompanying variance

$$\mathrm{Var}^{(u)} = \frac{1}{K} \sum_{k=1}^{K} \Big(s\big(\mathbf{x}^{(u)}, \mathbf{Y}^{(k)}\big) - \bar{s}^{(u)}\Big)^2.$$

These metrics estimate how “typical” the user is relative to repeated realizations of the random-walk model.

Such quantities naturally extend to population-level statistics: distributions of similarities, identification of outliers, and hypothesis testing against the null model provided by random walks.

5.5 Consensus Models

To summarize the global behavior of the entire ensemble, we consider two forms of consensus:

Markov Consensus. All walker trajectories across all users define empirical transition counts

$$C_{vw} = \#\{\text{times a walker moves } v \to w\}.$$

Normalizing the rows yields an empirical transition matrix

$$P_{vw} = \frac{C_{vw}}{\sum_{w'} C_{vw'}}.$$

The matrix P defines a global Markov model that captures the average transition tendencies dictated by the graph structure and the distribution of starting points. This model can be used to compute likelihoods of real user trajectories, to generate new synthetic walkers, or to build a deterministic “most probable” consensus path by greedy selection.

Medoid Trajectory. As a complementary summary, we compute the *medoid* of a set of walker trajectories—the trajectory that minimizes the total distance to all others:

$$k^* = \arg\min_k \sum_{j=1}^{K} d\big(\mathbf{Y}^{(k)}, \mathbf{Y}^{(j)}\big).$$

The medoid offers an interpretable representative path that arises from an actual random walker, as opposed to the probabilistic object represented by P.

5.6 Interpretation

The random-walk construction provides an explicit null model driven solely by the graph's local connectivity. If a real user's trajectory significantly deviates from the ensemble predicted by the graph, the deviation may reveal hidden structure, atypical behavior, or external influences that are not captured by topology alone.

Conversely, if the random-walk ensemble closely matches the user, the graph alone is sufficient to explain the observed behavior.

This duality—graph-driven randomness versus user-specific structure—is the central object of analysis in the subsequent sections of the report.

6 MARS_1.0 Project Tree

```
MARS_1.0/
 data/
    anime_ranks/
       anime.csv
       animelist.csv
       anime_with_synopsis.csv
       rating_complete.csv
       watching_status.csv
    anime_timestamps/
       anime_timestamps.csv
    cities_population_and_location/
       cities_population_and_location.csv
    myanimelist_countries_distribution/
       myanimelist_countries_distribution.csv
    users/
        profiles.csv
 db_tools/
    ...
    ...
 fds_tools/
    __init__.py
    data_cleaner.py
    fake_user_generator.py (class FakeUsersGenerator)
    fds_main.py
    project_latex/
        main.tex
        chapters/
           introduction.tex
           overview.tex
           fake_user_generation.tex
           dataset_curriculum.tex
```

```
         ...
          references.bib
       out/
           main.pdf
           ...
cda_tools/
   ...
   ...
__init__.py
.env
.gitignore
common_tools.py (classes CommonTools, PandasTools, DBTools)
datasets.json
LICENSE
README.md
requirements.txt
```

References

[1] Group 6. Manga and anime recomendation service. https://github.com/BlackSabbitch/MARS_1.0, 2025.

[2] International Anime Research Project. Anime survey 2020 preliminary results. https://sites.google.com/site/animeresearch/past-results/2020-results, 2020.

[3] Donato Riccio. World cities population - cleaned version. https://www.kaggle.com/datasets/donatoriccio/world-cities-population-cleaned-version, 2022.

[4] The Association of Japanese Animations. Anime industry report 2020. https://aja.gr.jp/download/anime-industry-report-2020-summary, 2020.

[5] Hernan Valdivieso. Anime recommendation database 2020. https://www.kaggle.com/datasets/hernan4444/anime-recommendation-database-2020, 2020.