# PROJECT TITLE:
# ANIME VIEWERSHIP ANALYSIS

*Date of Presentation: April 16, 2024*

*Name : Vamshi Krishna Bangaru*

*Pace Email Address : vb76262n@pace.edu*

*Class Name: CS 667- Practical Data Science*

*Program Name: MS in Data Science*

*School Name: Seidenberg School Of Computer Science and Information Systems, Pace University*

# Agenda

- Executive summary
- Project plan recap
- Data
- Exploratory data analysis
- Modeling methods
- Findings
- Recommendations and technical next steps

# Executive summary

- **Problem Statement:**
  - As viewers grapple with the overwhelming selection of anime, there's a twofold challenge: alleviating user decision fatigue and providing anime studios with predictive insights on viewership ratings to guide investment in future productions.

- **Solution Overview:**
  - Introduced a dual-purpose recommendation system powered by XGBoost and Cosine Similarity models to tailor viewer recommendations and forecast anime ratings.
  - Utilized comprehensive datasets to train the model, which not only customizes user experience but also assists studios in understanding trends and potential ratings for new titles.
  - The models demonstrated high predictive accuracy in forecasting viewership ratings, which is essential for both user engagement and studio decision-making.
  - Applied content-based filtering techniques, aligning viewers with preferred genres and guiding studios on content creation to ensure high engagement and return on investment.



Click here to return to the agenda.

# Project plan recap

| Deliverable | Due Date | Status |
|---|---|---|
| Data & EDA | 03/19/24 | Completed |
| Methods, Findings, and Recommendations | 04/02/24 | Completed |
| Complete Final presentation | 04/16/24 | Completed |

Click here to return to the agenda.

# Data

Click here to return to the agenda.

# Data

- **Data source:** MyAnimeList git repo database, including user ratings, watching status, and detailed anime statistics.
- **Sample size:** The original sample includes two datasets which are Anime dataset of 17,562 different animes and User ratings dataset of 325,772 different users, with a total of 109 Million rows of their ratings for different anime. Refer this Appendix slide for more information.
- **Time period:** The data encompasses the full range of MyAnimeList records from 1917 till 2021 year.
- **Data included:** Comprehensive user engagement data, including scores, watching statuses, watched episodes, and detailed anime information such as genre, producers, studios, popularity metrics, and more.
- **Data excluded:** The User ratings dataset used for this project contains 546,123 rows and 200,431 user ratings to limit computational load for this project. The full dataset, prepared as detailed in the Appendix slide, is available for use in production and stored in the designated Google Drive folder.
- **Assumptions:** It's assumed that the sample size is representative of the global anime-watching population. Scores of '0' are taken to mean the users who chose not to rate the anime.

For a detailed overview, please see this appendix slide.

# Exploratory Data Analysis (EDA)

Click here to return to the agenda.

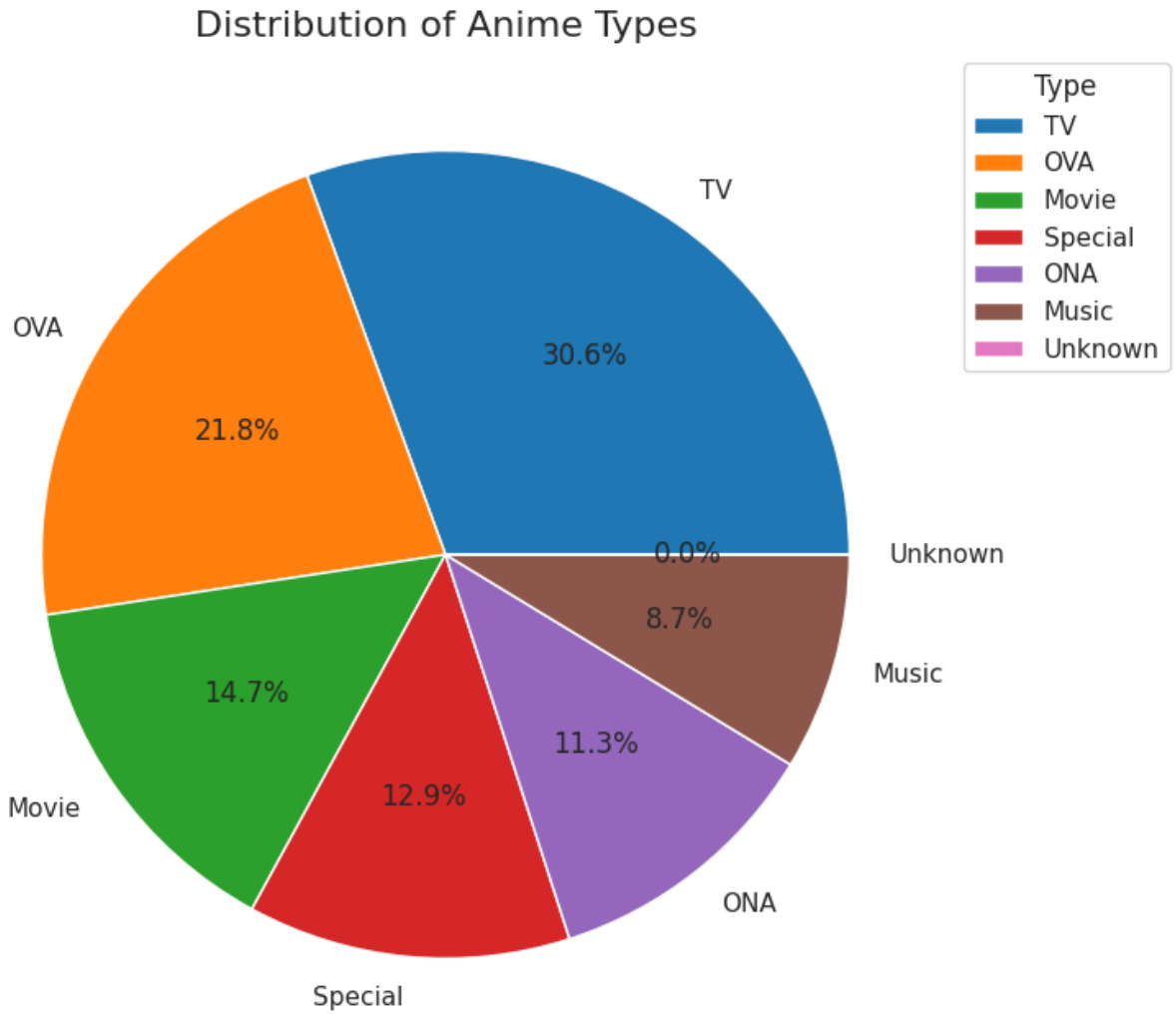
# Types of Anime format

- **TV**: Indicates television series.
- **OVA**: Stands for **Original Video Animation**, which typically refers to anime that is released directly to video without prior showings on TV or in theaters.
- **Movie**: Refers to full-length anime films
- **Special**: This usually denotes special episodes or extra content that is not part of the standard series or seasons.
- **ONA**: Stands for **Original Net Animation**, which indicates anime distributed through the internet.
- **Music**: These are anime music videos or short anime films or series revolving around musical elements.
- **Unknown**: This category represents data for which the type of anime was not specified or could not be categorized.

For a detailed overview, please see this appendix slide.

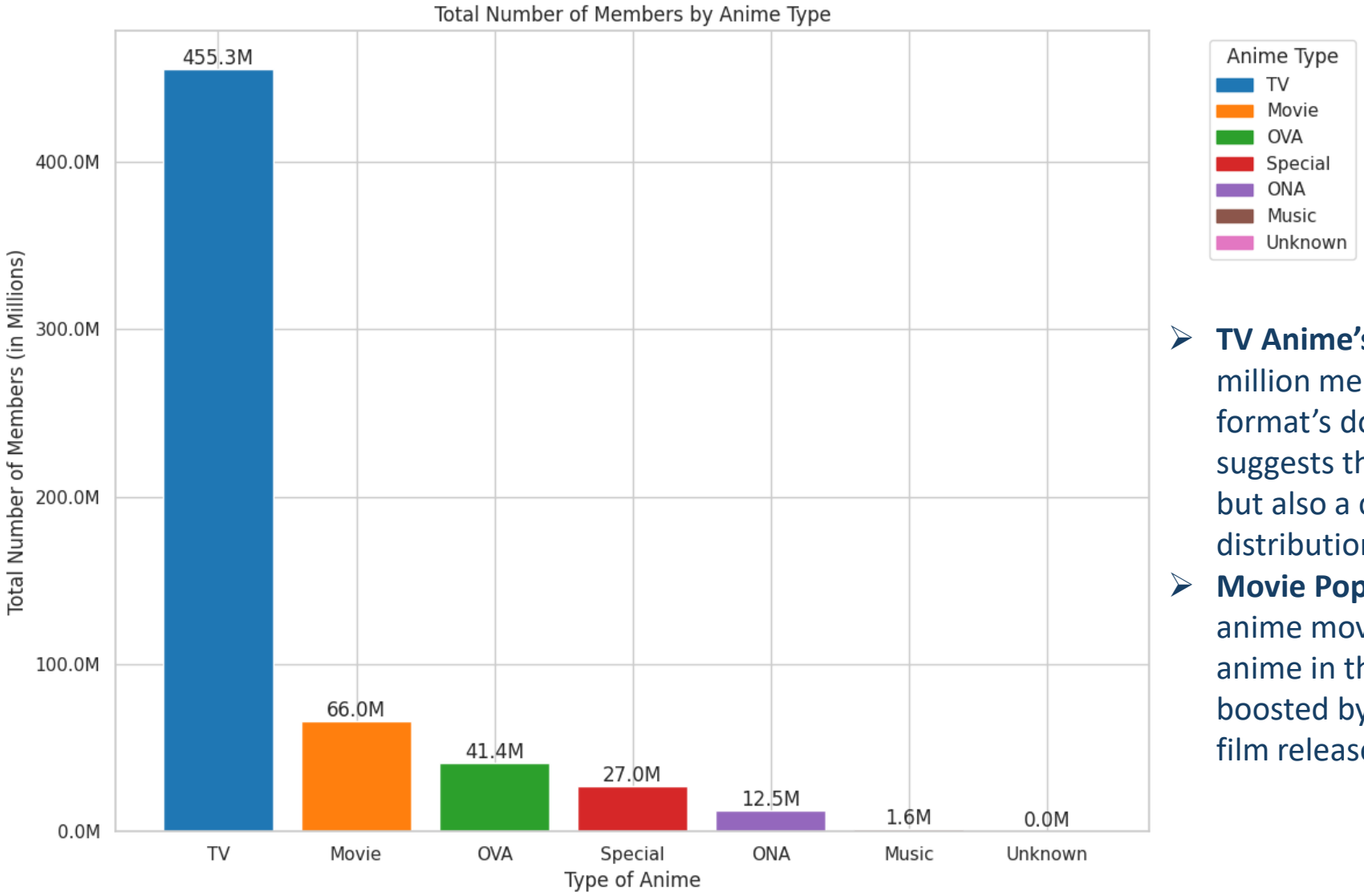
# Types Of Anime

- This pie chart illustrates the prevalence of various anime formats, indicating that TV anime is the most common, followed by OVAs and movies.
- ONAs reveal a shift towards online consumption, and the smaller music segment points to a specialized market.
- This information is valuable for guiding production, distribution, and investment strategies in the anime industry.



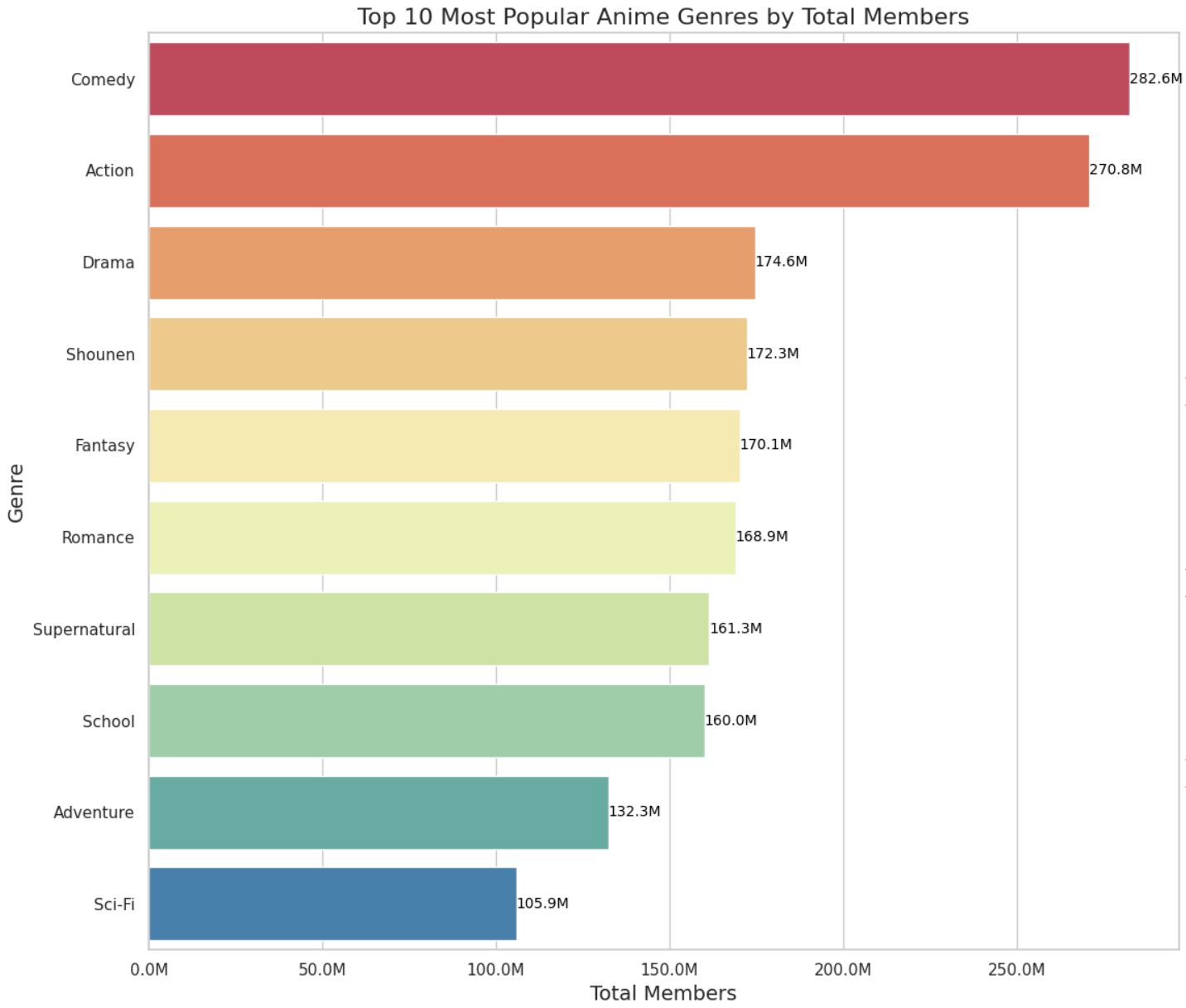

For a detailed overview, please see this appendix slide.

# Viewer Stats for different Anime Types



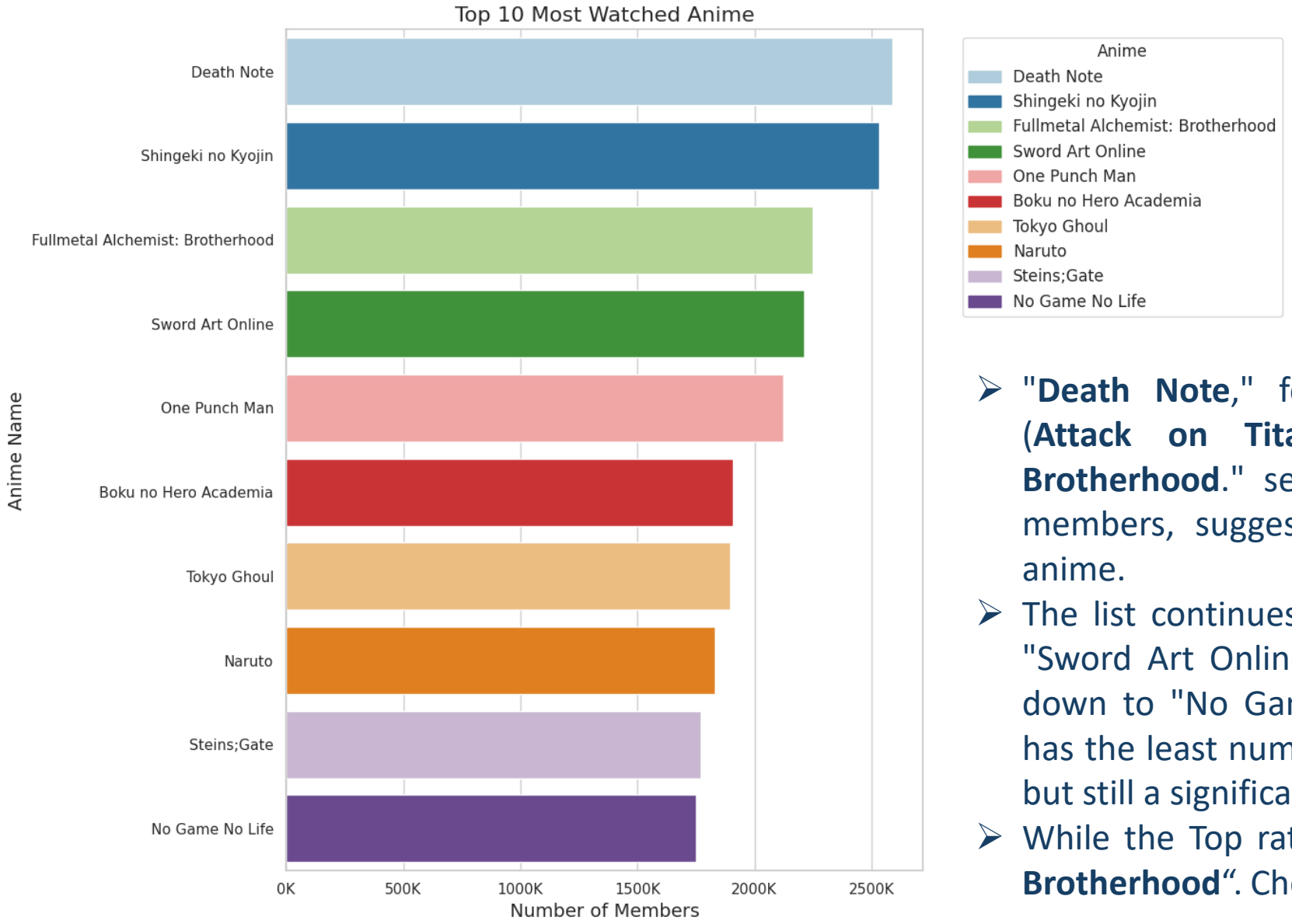

- **TV Anime's Dominance**: The staggering 455.3 million members for TV anime underscores the format's dominant position in the market. It suggests that television is not only a traditional but also a current powerhouse for anime distribution.
- **Movie Popularity**: The 66 million members for anime movies reflect the successful reach of anime in the film industry, which is likely boosted by both domestic and international film releases.

# POPULAR GENRES IN ANIME



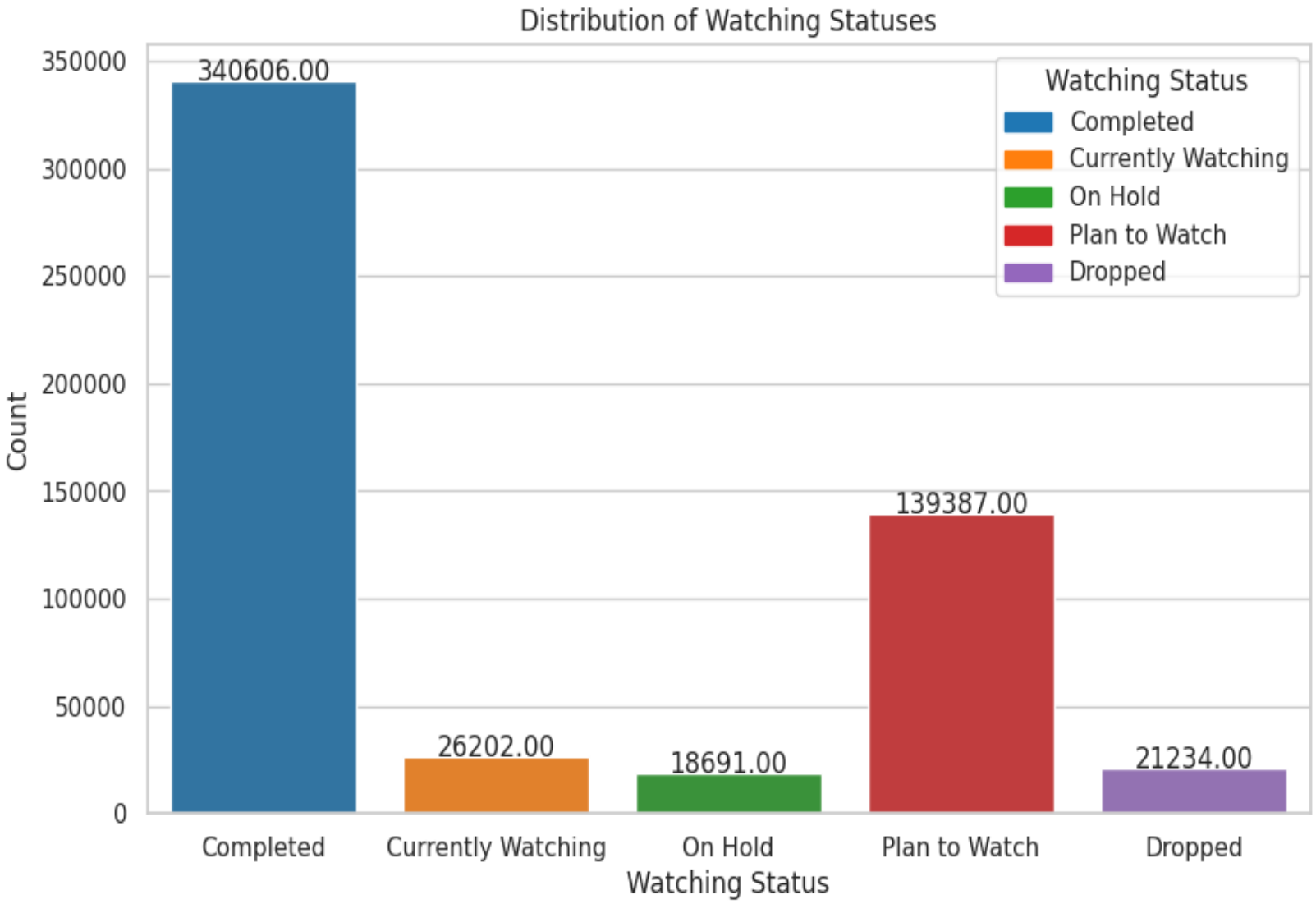

- The graph indicates that genres that offer humor, excitement, and escapism tend to attract the most fans.
- Comedy(282.6M) and action(270.8M) being at the top suggests that viewers widely appreciate these elements in anime.
- The data can help content creators and broadcasters to understand current trends and potentially shape future anime productions to cater to the largest audiences.

# MOST WATCHED ANIME




- "**Death Note**," followed by "Shingeki no Kyojin" (**Attack on Titan**), and "**Fullmetal Alchemist: Brotherhood**." series have the highest number of members, suggesting they are the most watched anime.
- The list continues with other popular titles such as "Sword Art Online," "One Punch Man," and so on, down to "No Game No Life" at the bottom, which has the least number of members among the top 10 but still a significant count.
- While the Top rated anime is "**Fullmetal Alchemist: Brotherhood**". Check Appendix for more info.

# Distribution of Watching Status across users




- High completion rates suggest satisfaction, or a strong appeal of the anime included in the data.
- In contrast, the smaller number of dropped series may indicate that viewers are selective but generally find anime that they are willing to commit to watching through to the end.
- The "Plan to Watch" category also suggests that there is a lot of interest in anime, with many viewers having aspirational watch lists.

# Modeling Methods

Click here to return to the agenda.

# Understanding Our Predictive Model

- **Outcome Variable :** Our mission is to forecast the average viewer rating for an anime series, which is a key indicator of its popularity and success.
- **Feature Selection :** We chose factors like genre, popularity, and user engagement, as these significantly influence an anime's success.
- **Model Choice:** We utilized a powerful forecasting technique known as **XGBoost**, which is like having a council of experts where each expert learns from the mistakes of the previous one, resulting in a very informed decision.
- **Why XGBoost?** This approach is known for its accuracy and efficiency, especially with large and complex data like ours.

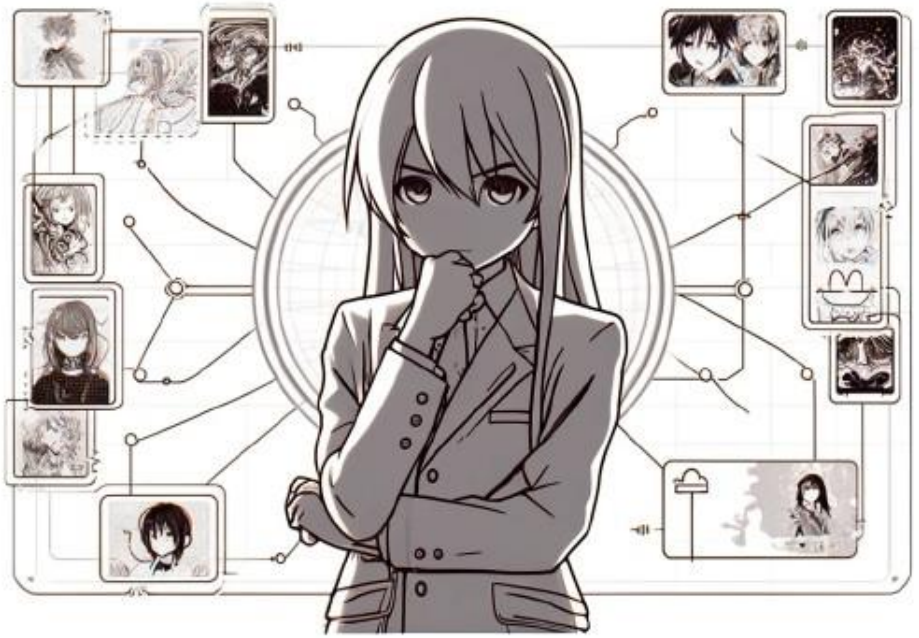
For a detailed technical overview, please see this appendix slide.

# Connecting Viewers to Their Next Favorite Anime

## Cosine Similarity for Recommendations:

Alongside our predictive model, we've also developed a recommendation system that uses the concept of **cosine similarity** to suggest anime titles that viewers might enjoy based on their preferences.

- This approach allows us to understand not just the obvious choices but also uncover hidden gems you might not find on your own.
- It's about making connections based on what matters to viewer's personal preferences.



For a detailed technical overview, please see this appendix slide.

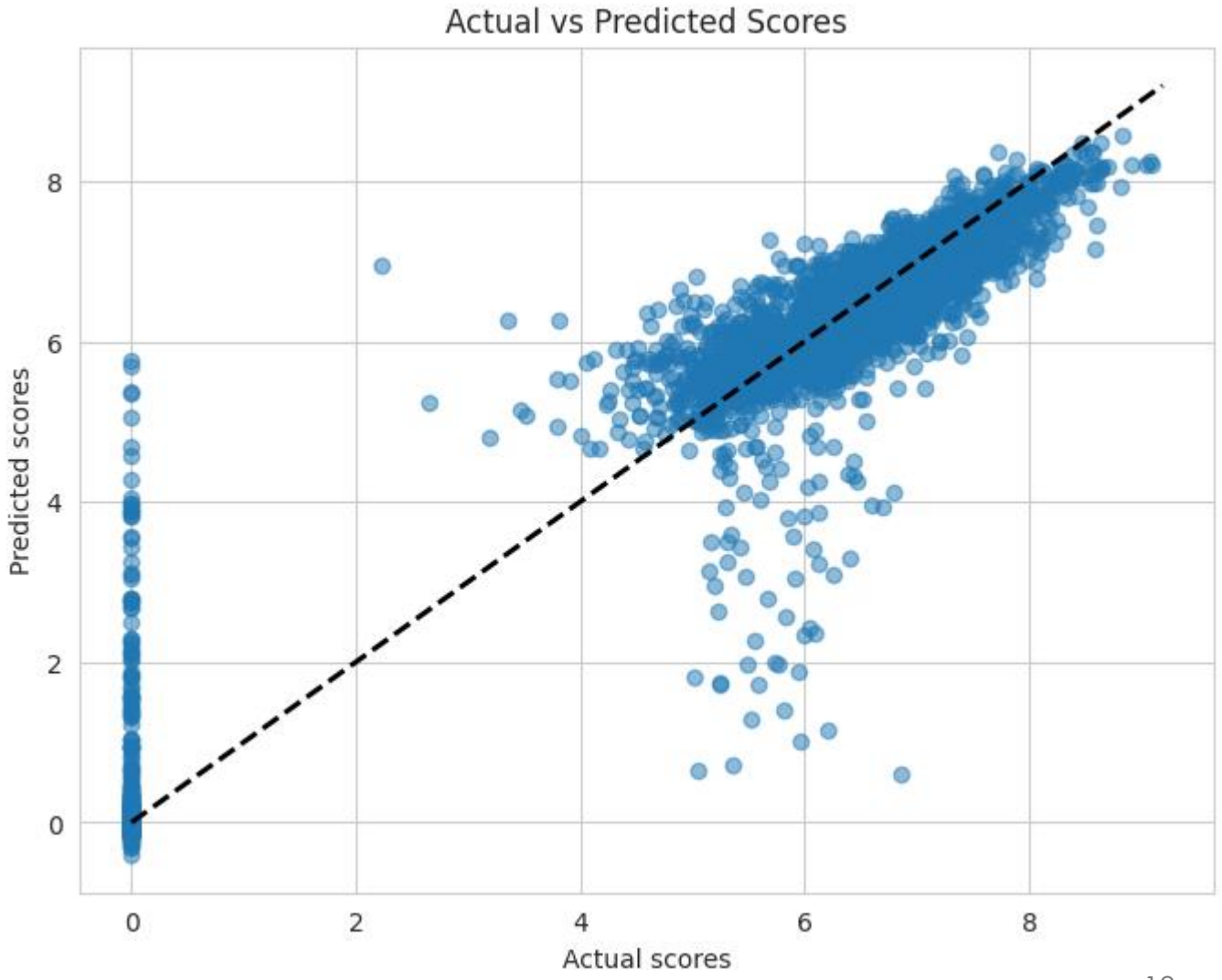
# Findings

Discovering What Fans Love: Insights from Our Data

Click here to return to the agenda.

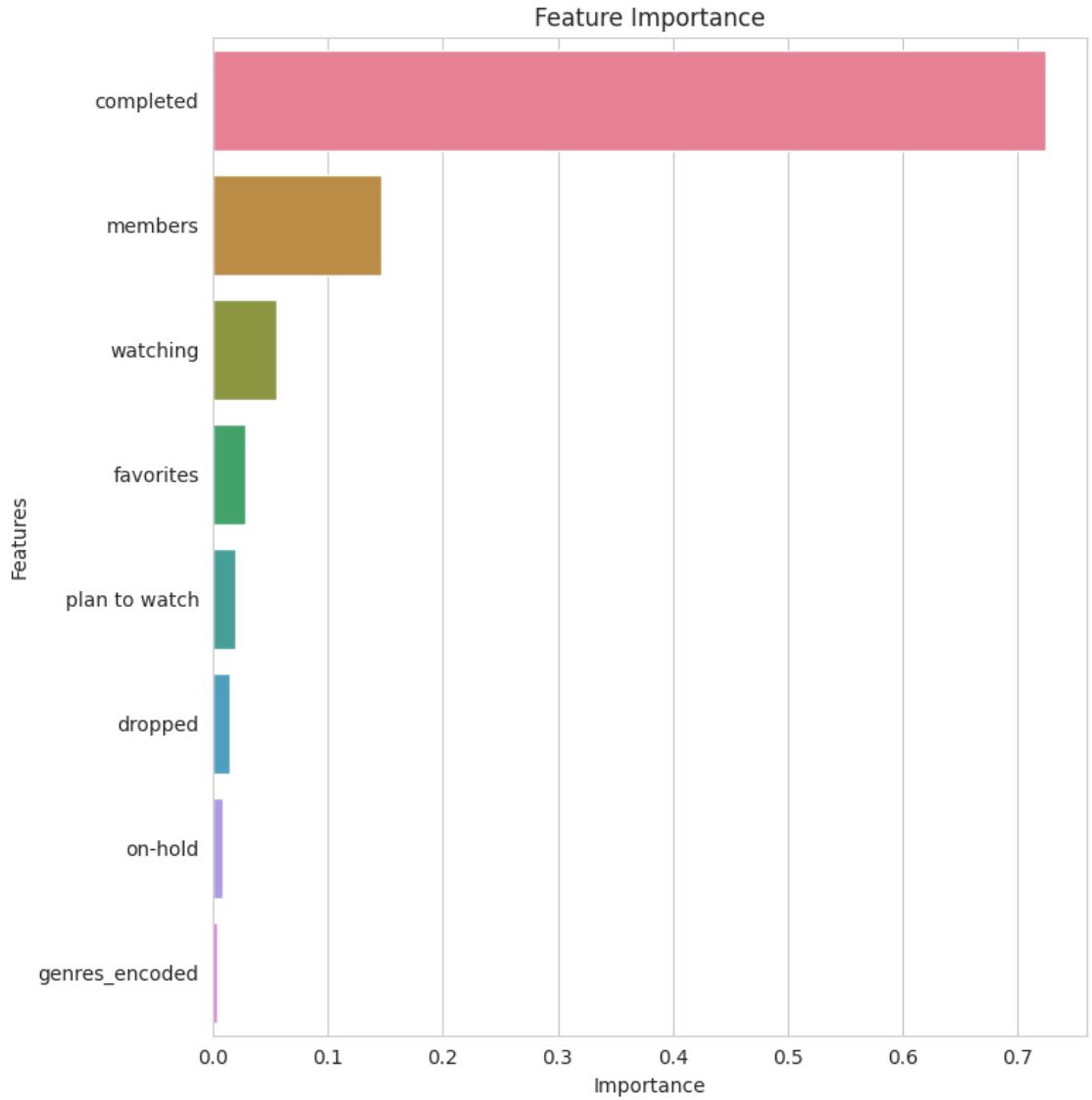
# Predicting the Anime Success

- **From our XGBoost Model:** This plot showcases actual vs. predicted scores, with many predictions closely aligned with reality, as indicated by their proximity to the line of perfect prediction.
- Most dots are clustered around the line, particularly as the actual scores increase, indicating that the model's predictions are relatively close to the actual scores for many animes.
- We can see a lot of points at the very bottom, where the predicted score is zero. This doesn't mean our predictions were all incorrect; it reflects the animes that were given a score of zero in our data, indicating missing or yet-to-be-determined ratings.




For a detailed technical overview, please see this appendix slide.

# What Drives Anime Popularity?




- **Completed Viewings** have the highest influence on ratings, suggesting that animes that are watched to completion often get higher ratings.
- **Membership Size** is also highly influential, meaning shows with more subscribers are likely rated higher.
- **Engagement Metrics** like watching status, favorites, and intent to watch have a moderate impact.
- **Less Influence** comes from how many dropped or put the show on hold, and the specific genre plays the smallest role in the model's predictions.

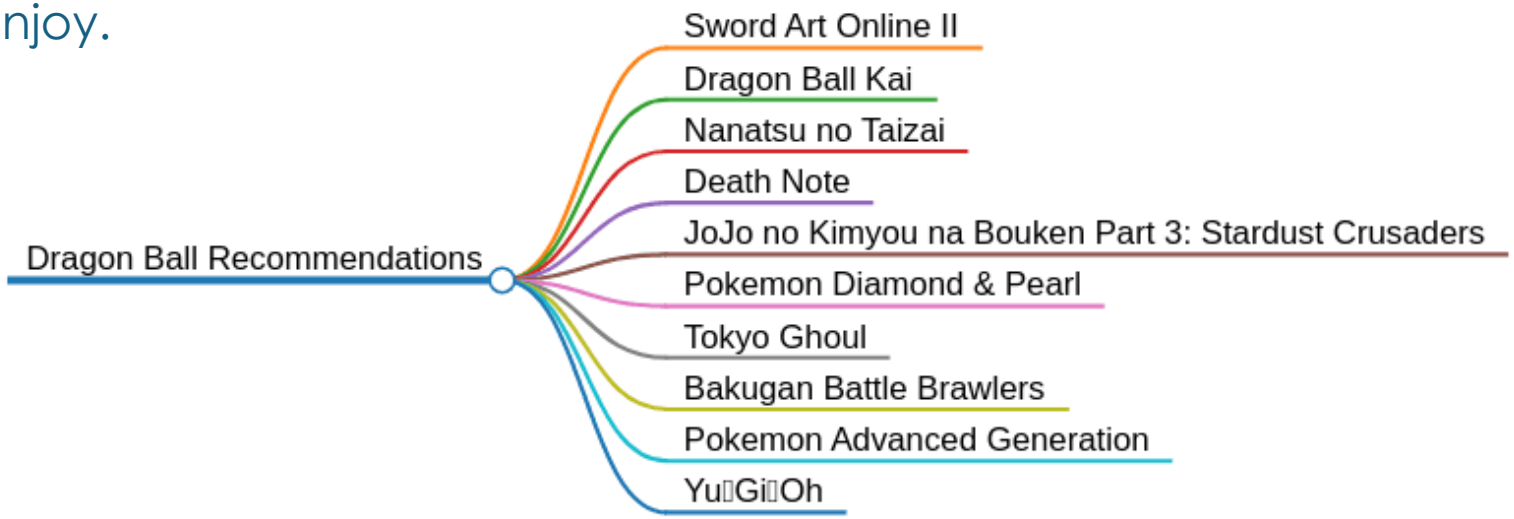
- **Business Implication:** Investing in series with growing viewership and completion rates could be a strategy for maintaining a highly engaging catalogue.

**Note**: Genre_encoded represents the mix of anime genres into numbers to help our system learn and make better predictions. Detailed information sis given in appendix.

For a detailed technical overview, please see this appendix slide.

# Generating Anime Recommendations Using Cosine Similarity

- Each colored line represents a unique path leading to a new story, showing how each recommended anime connects back to what fans might appreciate in a particular anime like “Dragon Ball“.
- It’s like a map for viewers to find their next adventure based on what they already enjoy.

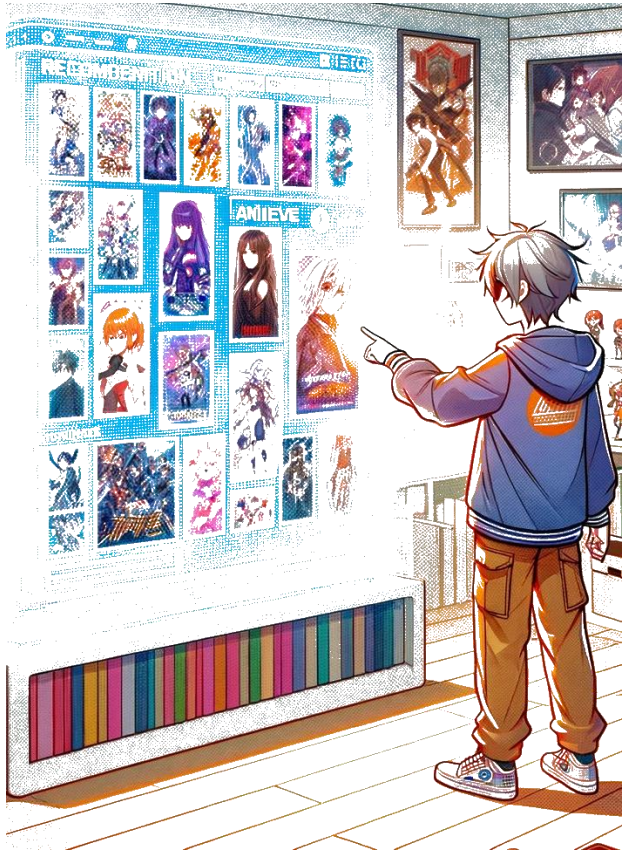



For a detailed technical overview, please see this appendix slide.

# Matching Fans with Their Next Favorite

The application of **Cosine similarity** for content-based filtering can promise in aligning viewer preferences with recommendations.

- **Customized Experiences:** Personalized anime recommendations increase time spent on the platform and boost visit frequency.
- **Uncovering Hidden Gems:** Recommendations spotlight less popular animes, broadening their audience based on tailored user interests.
- **Fostering Connection:** A service that mirrors viewer preferences keeps subscribers engaged and less likely to leave.
- **Strategic Content Choices:** Analytics from the recommendation engine guide smarter decisions on anime acquisition and production.
- **Smarter Advertising:** Deep insights into preferences allow for advertising that resonates more effectively with users.



For a detailed technical overview, please see this appendix slide.

# Recommendations and Technical Next Steps

Click here to return to the agenda.

# Strategic Moves for Anime Success

## Competed Anime Relates to Higher Ratings

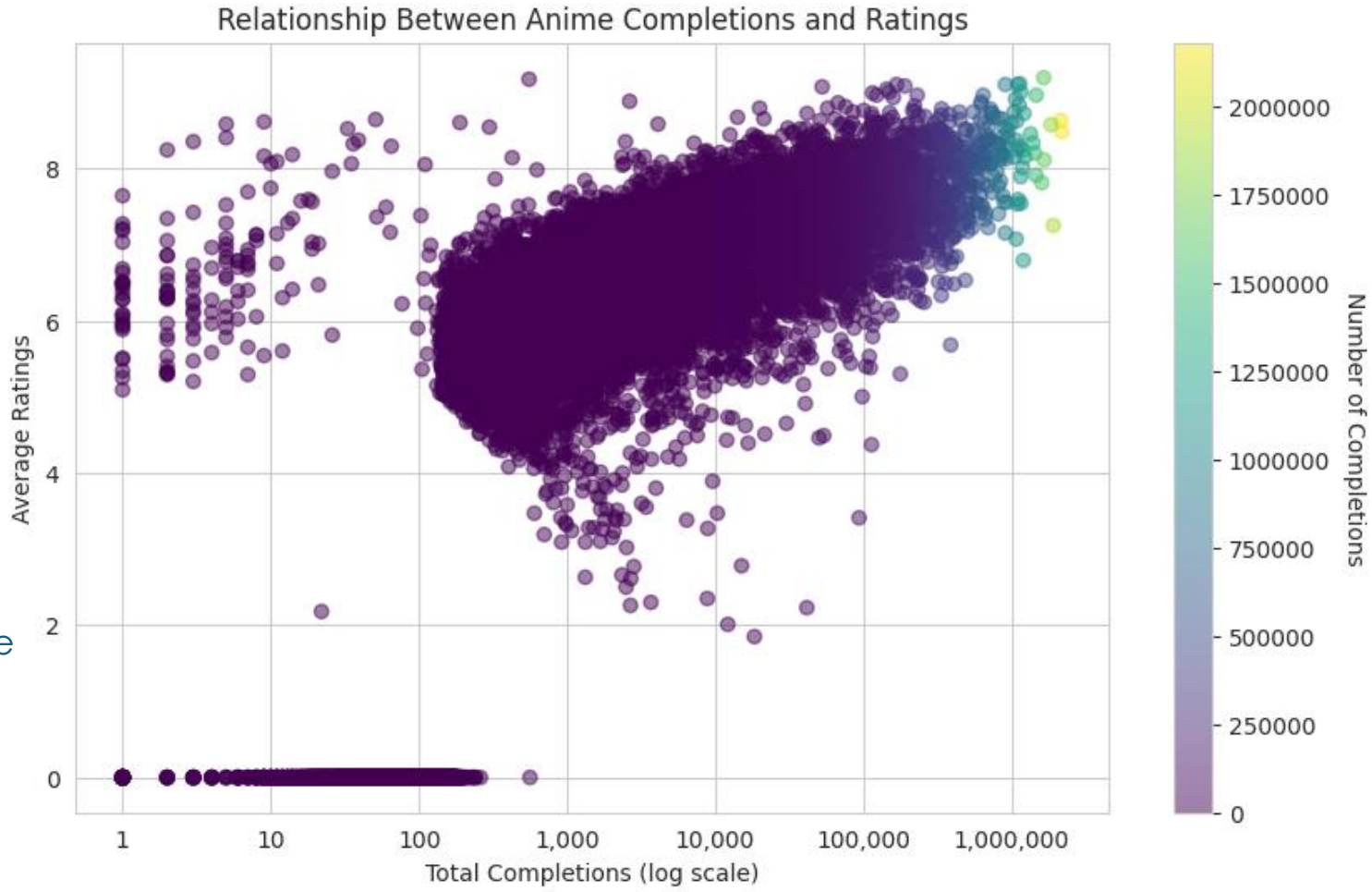
### Insight:

- Our XGBOOST model shows that 'Completed viewings' are the most crucial measure of an anime's success.
- Data reveals a clear pattern—shows that are watched to the end tend to have higher ratings.
- This trend signals that completion rates are a strong measure of a show's appeal and quality in the eyes of viewers.

### Actionable Recommendation:

- Curate Captivating Content: Prioritize acquiring shows with a track record of high completion rates. These are the stories that keep viewers engaged from start to finish, indicating they're the ones people love the most.

- Spotlight on Success: Amplify successful shows in our marketing efforts. Highlighting these crowd-pleasers can help draw in new viewers and keep our current audience excited and committed to our platform




For a detailed technical overview, please see this appendix.

# Strategic Moves for Anime Success

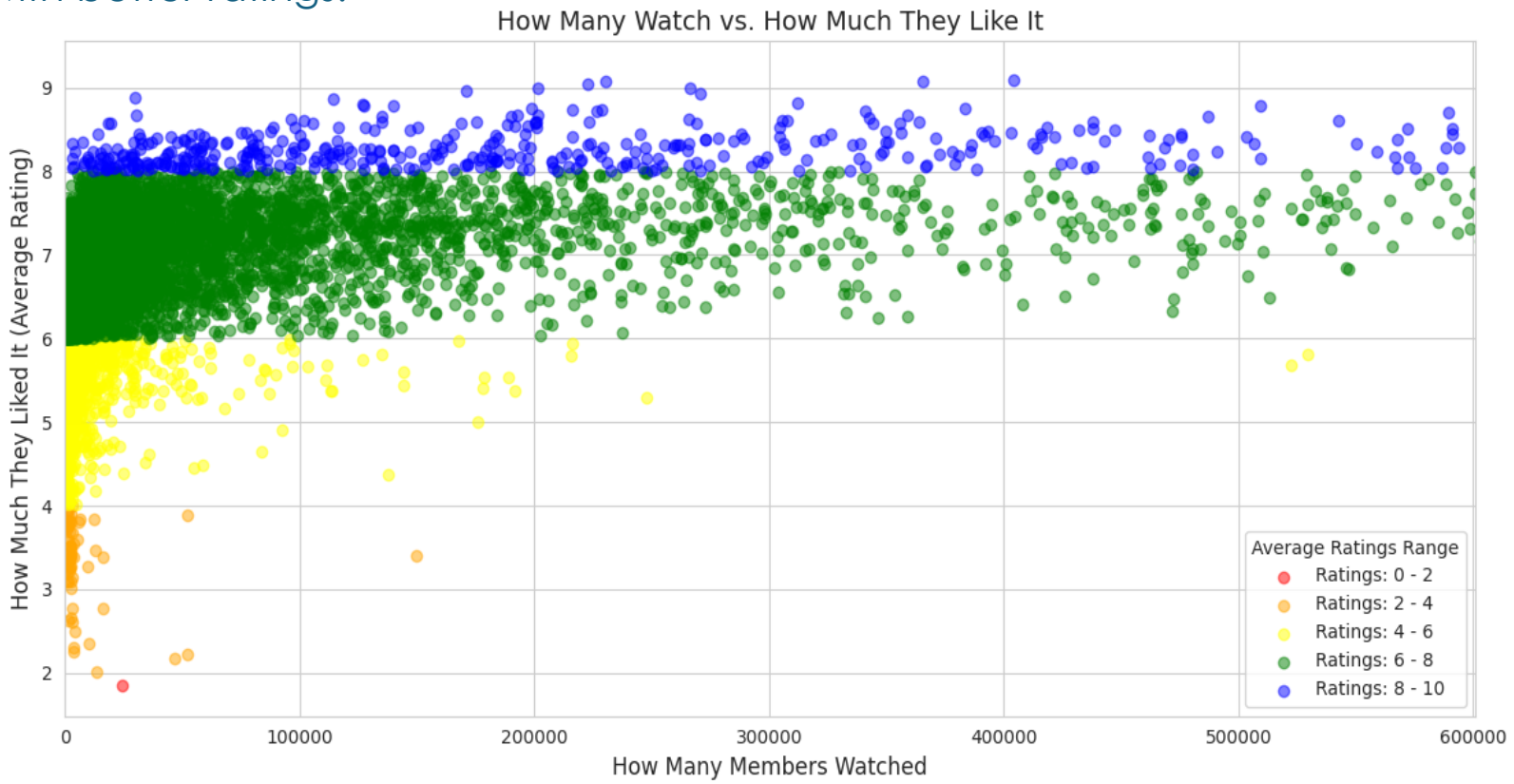
## Community Engagement Elevates Content

**Insight:**

Our XGBOOST model shows that 'Member's enlisted' are also the important measure of an anime's success. High levels of user interaction correlate with better ratings.

**Actionable Recommendation:**

- Focus on the series that have gathered a crowd. These shows are not just watched, they're appreciated, and they're where our new viewers could start.
- Spotlight these top-rated series. They're the ones that won't just bring viewers but they'll keep them glued to their screens.




For a detailed technical overview, please see this appendix.

# Strategic Moves for Anime Success

## Empowering Our Data Science Journey

### Advance the Recommendation Engine

- Next Step: Experiment with more sophisticated machine learning models, such as neural network-based recommenders, to further enhance the accuracy and personalization of content recommendations.
- Rationale: Leveraging complex models can uncover deeper patterns in viewer preferences, leading to more engaging and satisfying recommendations.

### Expand the Dataset

- Next Step: Incorporate additional data sources, such as social media sentiment analysis and real-time viewership metrics, to enrich our understanding of anime popularity and viewer engagement.
- Rationale: A more comprehensive dataset allows for a nuanced analysis of trends and preferences, potentially revealing untapped opportunities for content strategy and community building.

For a detailed technical overview, please see this appendix.

# Appendix

Click here to return to the Main Slide.

# OVERVIEW OF THE DATASETS USED

This appendix provides a comprehensive overview of the datasets used in the analysis

- **Dataset Overview:**
  - Origin: Derived from MyAnimeList user data.
- **Original Dataset Specifications:**
  - **Rows:** 109 Million (user-anime interactions)
  - **Animes:** 17,562 (unique entries)
  - **Users:** 325,772 (unique participants)
- **Dataset Components:**
  - **animelist.csv**: Captures user interactions, including scores, watching status, and episodes watched.
  - **watching_status.csv**: Enumerates all possible user watching statuses with descriptions.
  - **anime.csv**: Provides exhaustive anime details including MAL_ID, names, scores, genres, and more.
- **Column Details:**
  - **animelist.csv**: user_id, anime_id, score, watching_status, watched_episodes.
  - **watching_status.csv**: status, description.
  - **anime.csv**: MAL_ID, Name, Score, Genres, English name, Japanese name, Type, Episodes, Aired, Premiered, Producers, Licensors, Studios, Source, Duration, Rating, Ranked, Popularity, Members, Favorites, Watching, Completed, On-Hold, Dropped, Plan to Watch and Scores(1-10).



Click here to return to the Previous Slide.

# OVERVIEW OF THE DATASETS USED

**Efficient Data Sampling for In-Depth Analysis:**

- Our analysis begins by accessing an extensive dataset from 'animelist.csv', which captures comprehensive viewership patterns.
- Given the dataset's size, we implement a random sampling strategy, reducing the dataset to 0.5% of its original volume, striking a balance between detail and computational feasibility.
- The sampled subset is then exported back into a CSV file, maintaining the integrity of the original structure, which serves as a foundation for subsequent modeling and analysis tasks.

- **Final Dataset Specifications:**
  - **Rows:** 546,123 (user-anime interactions)
  - **Animes:** 17,562 (unique entries)
  - **Users:** 200,431 (unique participants)



Click here to return to the Previous Slide.

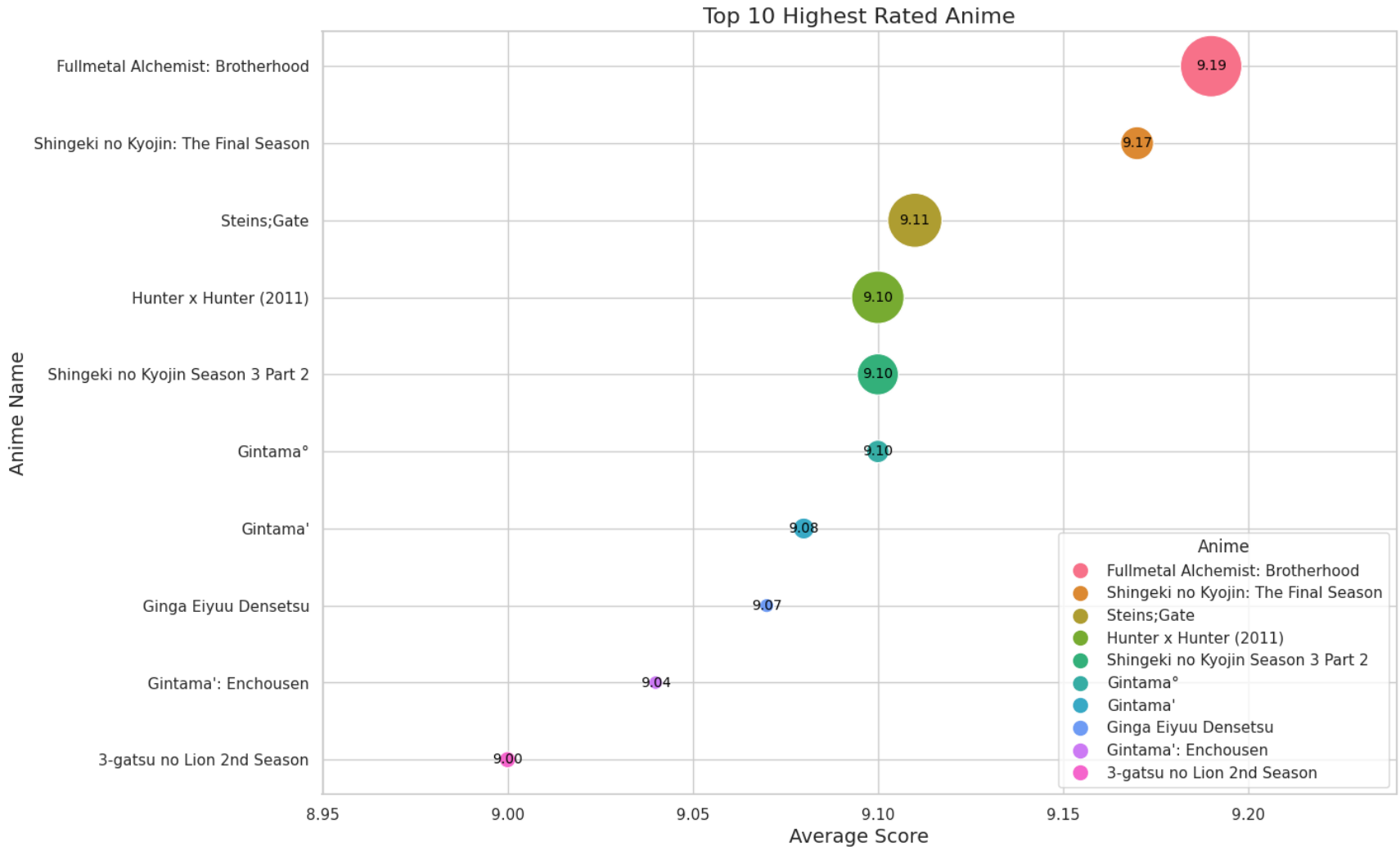
# Types of Anime format

- **TV**: Indicates television series, which is the largest segment at 30.6%.
- **OVA**: Stands for **Original Video Animation**, which typically refers to anime that is released directly to video without prior showings on TV or in theaters; this is the second-largest group at 21.8%.
- **Movie**: Refers to full-length anime films, accounting for 14.7% of the distribution.
- **Special**: This usually denotes special episodes or extra content that is not part of the standard series or seasons, making up 12.9%.
- **ONA**: Stands for **Original Net Animation**, which indicates anime distributed through the internet; it has 11.3%.
- **Music**: These are anime music videos or short anime films or series revolving around musical elements, with 8.7%.
- **Unknown**: This category likely represents data for which the type of anime was not specified or could not be categorized, showing a negligible 0.0% which may be a rounding issue in the data.



Click here to return to the Previous Slide.

# Highest Rated Anime




- The anime "Fullmetal Alchemist: Brotherhood" leads with the highest average score, closely followed by "Shingeki no Kyojin: The Final Season."
- The ratings are very close, with the top shows all scoring above 9.0, indicating a very positive reception from viewers.

Click here to return to the Previous Slide.

# Technical Deep Dive into Our XGBoost Model

- **Model:**
Our XGBoost regression predicts the average viewer scores for anime series, optimizing for a nuanced balance between accuracy and model simplicity.

- **Feature Engineering:**
The model incorporates encoded genres and key engagement metrics such as membership numbers and watching statuses to predict the outcomes.

- **Technical Specifications:**
We fine-tuned the learning rate to 0.1 and limited the tree depth to 5 levels to prevent overfitting while ensuring robust predictive power.

- **Feature Impact:**
A detailed analysis revealed that 'completed' and 'members' are the most influential factors, guiding user engagement and retention strategies.

Click here to return to the Previous Slide.

# Technical Deep Dive into Our XGBoost Model

**Performance Metrics:**
We validated our model's accuracy using RMSE, with a training score of 0.628 and a test score of 0.694, alongside $R^2$ values of 95.79% and 94.76% for training and testing datasets respectively.

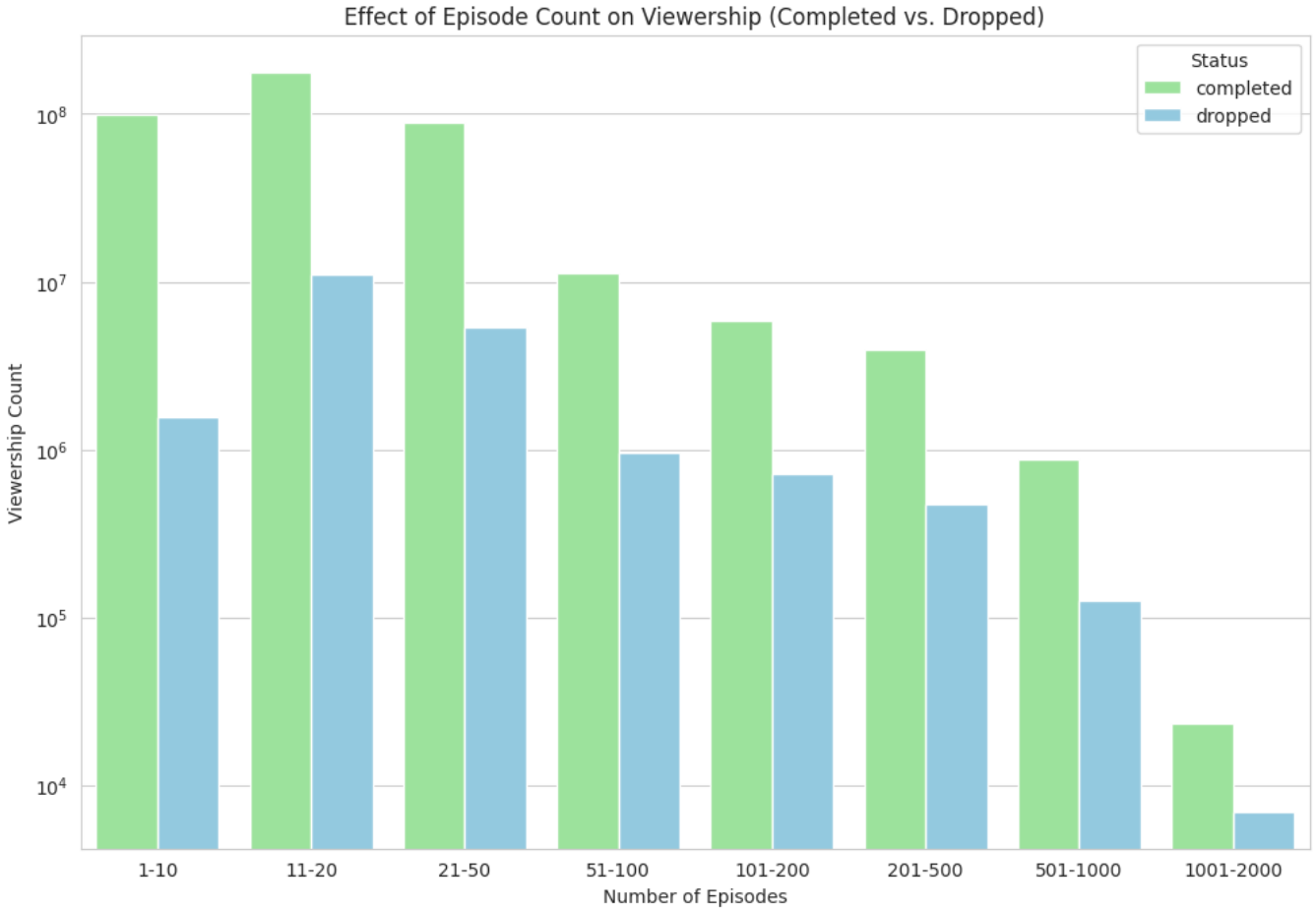
**Predictive Insight:**
The scatter plot showcases our model's ability to predict actual scores accurately, as evidenced by the close alignment with the line of perfect predictions.

**Model Application:**
This predictive precision enables us to forecast viewer ratings with confidence, directly impacting content curation and personalized recommendations.



Click here to return to the Previous Slide.

# Understanding Viewer Commitment Across Anime Series




**Graph Analysis:**

- This bar chart illustrates viewership trends, comparing the number of completed series against those dropped, categorized by the total number of episodes.

**Key Observations:**

- Series with fewer episodes (1-50) show a higher completion rate, indicating a strong viewer preference for shorter series.
- A noticeable decline in completion rates is observed as the episode count increases, particularly beyond 100 episodes.

**Significance:**

- Shorter anime series tend to retain viewers' interest, potentially due to a more concise storytelling approach and a lower time commitment required.
- Longer series face greater drop rates, suggesting viewer fatigue or a higher barrier to entry for new viewers.

**Strategic Implications:**

- Content creators and platforms can leverage these insights to tailor their offerings, prioritizing the production and promotion of series within optimal episode ranges.
- Investing in shorter series or segmenting longer ones into seasons could enhance viewer retention and overall engagement.



Click here to return to the Previous Slide.

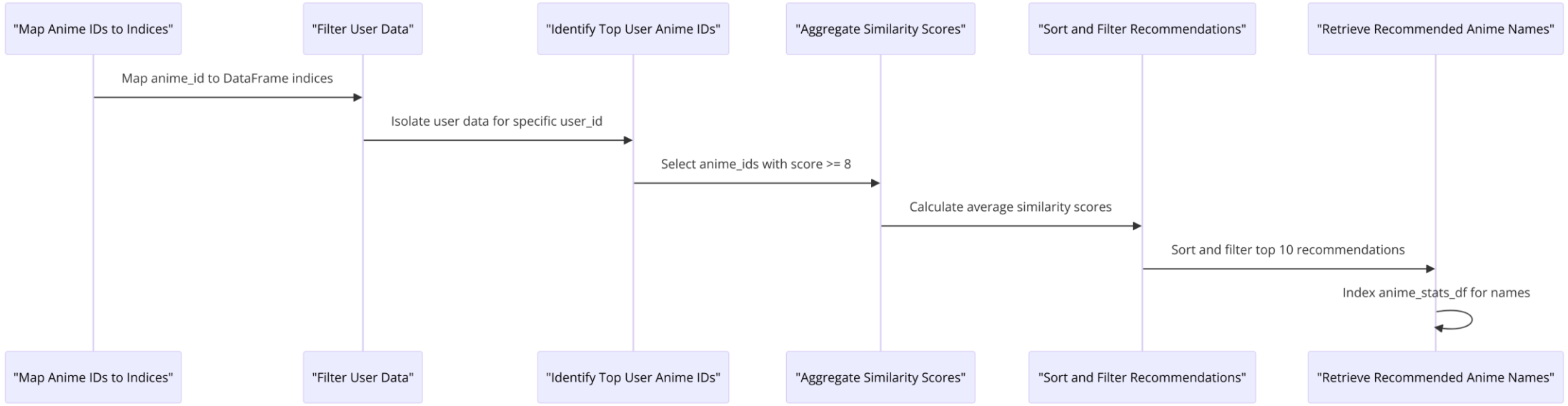
# Technical Deep Dive into Our Anime Recommendation Engine

**Understanding Cosine Similarity:**

- At its core, cosine similarity measures the cosine of the angle between two vectors in a multi-dimensional space. For our anime recommendation system, each vector represents an anime's characteristics or user ratings, making it possible to quantify similarity.

**Application in Our Recommendation System:**

- We construct a matrix from user ratings and anime features, where each row represents an anime, and each column represents a user or feature. The cosine similarity scores between anime are then calculated to identify the most similar titles to a user's favorites.






Click here to return to the Previous Slide.

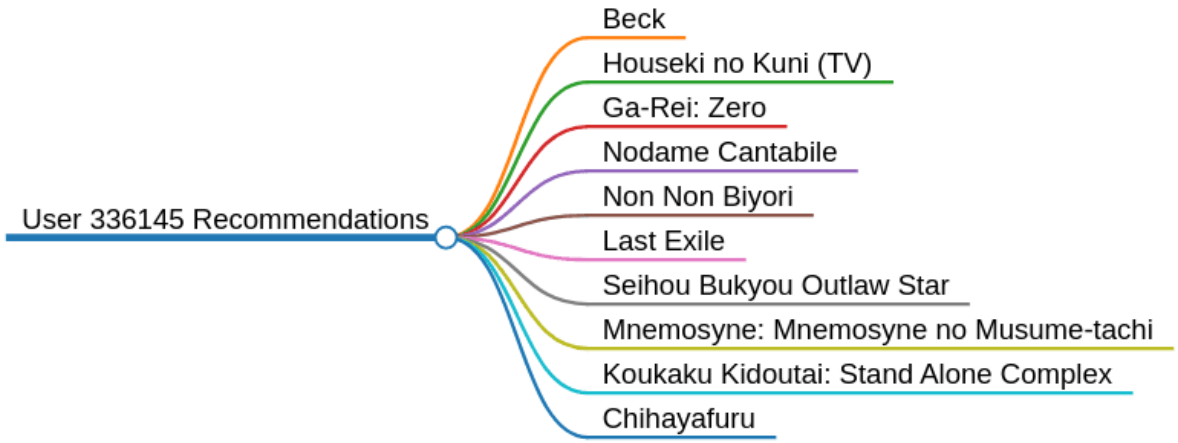
# Technical Deep Dive into Our Anime Recommendation Engine

**Benefits and Impact:**

- This method excels at capturing nuances in user preferences, allowing us to recommend a wide range of titles that align with individual tastes. It's particularly effective in surfacing recommendations that a user might not discover through traditional search or browsing methods.

**Algorithm Implementation:**

- The cosine similarity score ranges from -1 (completely different) to 1 (exactly the same), with higher scores indicating greater similarity. We recommend titles with the highest similarity scores to a user's rated anime, ensuring relevance and personalization.






Click here to return to the Previous Slide.

# THANK YOU

Click here to go to Main Slide