Predicting MyAnimeList Score of Modern Anime based on their Features

JSC370 2025 Midterm Project

Razan Ahsan Rifandi (1009562108)

Introduction

Background

Anime is a genre of animated works with styles that has their roots originating from Japan. The word itself originated from the same Japanese term, derived from a shortening of the English word 'animation'. Historically, it is used to specifically refer to animation produced in Japan. In recent times, it more commonly refers to the distinct group of animation styles itself disregarding the work's country of origin, although a vast majority of them still originates from Japan. Anime itself has had its fair share of history, with the style evolving over time to reflect advancements in animation technology and a growing audience. A noticeable shift is around the year 2012 where HD quality and a defining art style more reminiscent of today's anime is starting to be adopted.

Anime is typically released via television broadcasting in seasonal cycles. Some may also be broadcasted through theaters instead as movies outside the seasonal cycles. Many anime are adaptations of existing source material, such as manga (Japanese comics), light novels, or video games, while others are original creations. The production of anime is often handled by specialized animation studios (e.g. MAPPA, Kyoto Animation, etc.), with some being associated with their production quality.

Measuring how "good" an anime is inherently subjective, especially considering the difference in themes such as stories and animation style and how they resonate with a viewer's preferences. For the purposes of this research, I am interested in the overall community rating for an anime from the website MyAnimeList (MAL). It is one of the largest online databases and communities for anime alongside AniList and Kitsu.io. Each anime entry on MAL includes detailed information such as its type (TV series, movie, etc.), source material, genres, studios, and demographic ratings (e.g., G, PG-13, R). Users can rate anime on a scale of 1 to 10. Additionally, MAL tracks the number of "members" (users who have added the anime to their list) and "favorites" (users who have marked the anime as a favorite). These metrics could provide insights on the correlations of an anime's features and its rating, which is the main motivation of this research.

Research Question

For the purposes of this research, I will be defining anime as television series or theatrical releases that are listed as entries on MyAnimeList. Aside from those, MAL also lists OVAs (Original Video Animations), commercials, previews, and even music as entries. As they are not what would generally be categorized as animes, they will not be included in this research. Furthermore, this research will specifically focus on anime released over the 10 year period from 2015-2024, post the shift in quality around 2012s, which I will be referring to as "Modern Anime". With that in mind, the research question for this study is: "**How can we predict**

the MyAnimeList score of modern anime based on type, source material, age rating, popularity metrics, animation studios, and genres?"

Methods

For this Midterm Project, I will be conducting the analysis using the programming language R. Summary statistics and plots for my EDA will be generated using kable and plots using ggplot2 respectively.

Data Collection and Wrangling

The data for this research was collected from the MyAnimeList (MAL) "Jikan" API. The data was collected for all anime released between 2015 and 2024 accessed by season for each year (Winter, Spring, Summer, Fall). Since the API uses pagination, the data collection iterates through the pages of each season to retrieve all entries. The response is in JSON format and includes nested structures for some fields such as studios genres. In order to convert it into a proper R dataframe, the nested values were flattened by extracting the names and combining them into comma-separated strings. After collecting the data for each season, the individual datasets were merged into a single dataframe for analysis.

The raw data I collected using includes the following variables:

- `title_english`: The English title of the anime. If English title is provided, the value is the original romanized title of the anime.
- `type`: The format of the anime (e.g., TV, Movie, OVA, etc.).
- `source`: The source material for the anime (e.g., manga, light novel, original, etc.).
- `year`: The year the anime was released.
- `rating`: The demographic rating of the anime (e.g., G, PG-13, R, etc.).
- `score`: The average user rating on MyAnimeList (ranging from 1 to 10).
- `members`: The number of users who have added the anime to their list.
- `favorites`: The number of users who have marked the anime as a favorite.
- `studios`: The animation studios responsible for producing the anime.
- `genres`: The genres associated with the anime (e.g., Action, Romance, Fantasy, etc.).

These columns are also included for filtering purposes and will be removed later:

- `mal_id`: The unique identifier for each anime on MyAnimeList.
- `airing`: The boolean value representing whether the anime is currently airing (at the time of API call)

Data Cleaning

After collecting the raw data, the data is further processed for analysis as follows:

1. Removing duplicates, MAL may list the same anime entry in multiple seasons (e.g. if the anime was still airing during data collection). Duplicates were removed using the unique `mal_id` column which is then dropped.
2. Filtering relevant entry types, The dataset was filtered to include only entries that align with our definition of anime: animated shows with television or theatrical broadcasts. This is represented with the value 'TV' and 'Movie' in the `type` column.

3. Removing currently airing shows, as their scores and popularity metrics may not be stable. They were removed using the `airing` column which is then dropped.
4. Reformatting the columns, `studios` and `genres` columns were reformatted so that empty strings (a byproduct of the data collection process) and "Unknown" genre were replaced with NA values. The NA values were subsequently observed in order to process them.
5. Removing the NA values, the columns with the most NA values are studios and score, which is the response variable. Since there is a lot of overlap in the rows with NA values for these columns, I decided to remove all NA values for convenience of data usage in the future.

After cleaning, the final dataset consists of 2,549 entries and 10 columns. The raw code for calling the API, collecting the data, and cleaning the data will be provided in the .Rmd file, but as the entries could update in the future the generated clean dataset will be provided in a .csv file for reproducibility.

Exploratory Data Analysis (EDA)

EDA was conducted to examine the distribution of variables and explore potential relationships between them, as follows:

1. Count of `type`, `source`, `rating` (Summary Table)
2. Anime distribution by Year (Barplot)
3. Distribution of `score`, `members`, and `favorites` (Summary Table and Histograms)
4. `scores` vs. `type`, `source`, and `year` (Boxplot)
5. `scores` vs. `studios` and `genres` (Summary Table)
6. Pair scatterplot of `scores`, `type`, and `rating`

Preliminary Results

Below are the key findings of the EDA. Plots and summary tables will be given in the appendix for reference.

1. **Distribution of Anime Types, Sources, and Ratings**
   - `type`: The dataset contains 625 movies and 1,925 TV series, which means that our dataset are not distributed evenly between movies and TV series.
   - `source`: The 5 most common source materials in our dataset are Manga (824 anime), Original (615 anime), Light Novel (358 anime), Game (186 anime), and Web Manga (165 anime). There are 17 unique sources in the dataset.
   - `rating`: The count of demographic ratings from most common are PG-13 (1,736 anime), R (399 anime), G (210 anime), PG (119 anime), and R+ (86 anime).

2. **Anime and Score distribution by Year**

   There are more anime created pre-2019 than post-2019, suggesting a potential decline in the number of new anime releases in recent years (possibly because of the COVID-19 pandemic). The score distribution does not vary significantly across years, indicating that the year of release may not be a strong predictor of an anime's score.

3. **Distribution of Continuous Variables**

score seems to be distributed normally with a mean and median at 7. From the histogram, the distributions of members and favorites are heavily right-skewed, meaning most anime have relatively low membership and favorite counts, while a few popular anime have extremely high values. To address this, a log transformation was applied, which normalizes the distributions and makes them more suitable for analysis.

4. **Relationships Between Scores and Other Variables**
   - Movies tend to have higher scores than TV series, as seen in the boxplot. This could be due to the higher production quality or more focused storytelling often associated with movies as they had less total runtime than a typical anime season.
   - Score distribution varies by source material. For example, on average, manga adaptations tend to receive higher scores than light novel adaptations, which in turn score higher than original creations. Similarly, score distribution varies by age rating.
   - Score distribution varies by studio.Some studios produce anime with consistently higher scores, while others show more variability. However, some studios had a low count of anime in the dataset, which raises concern about the entries not being representative of that particular studio.
   - Score distribution varies by genre. Certain genres tend to receive higher scores than others, which reflects audience preferences and genre-specific expectations.
5. **Correlations Between Continuous Variables**

Logged members favorites are highly correlated, which indicates that they measure similar aspects of an anime's popularity. Therefore, including only one of these variables in future modeling should be sufficient to avoid multicollinearity. A linear correlation is observed between scores and both log_members (0.548) and log_favorites (0.644), suggesting that more popular anime tend to receive higher scores. This relationship will be further explored in the modeling phase.

Summary

To summarize, an anime's score on MAL seems to vary by type, source material, age rating, animation studios, and genres. Additionally, popularity metrics such as members and favorites show a strong linear correlation with scores, indicating that more popular anime are likely to receive higher ratings.

A limitation of our dataset is that the it is not evenly distributed across the values of categorical columns (e.g. some studios had very low count of attributed anime). Furthermore, the dataset size of 2,549 entries is relatively small, which could limit the model's ability to capture complex patterns. These issues will need to be carefully addressed in the final report to ensure robust and reliable results.

The next steps that will be taken in the final report will involve building and evaluating predictive models:

1. **Data Preparation**

   The dataset will be divided into training and testing sets (e.g., 80-20 split) to evaluate the model's performance on unseen data. Next, relevant features will be selected based on their importance and correlation with the target variable (score). Categorical variables will be encoded (e.g. one-hot or target encoding) if necessary.

2. **Model Building and Evaluation**

   The final report will be considering several regression models (e.g. Linear Regression, Random Forest, XGBoost, and SVR) and compare them using metrics (e.g. Root Mean Squared Error (RMSE) and R-squared (R^2)). Feature importance analysis will be conducted on the best model for interpretation and identify the influential predictors of anime scores.

Appendix

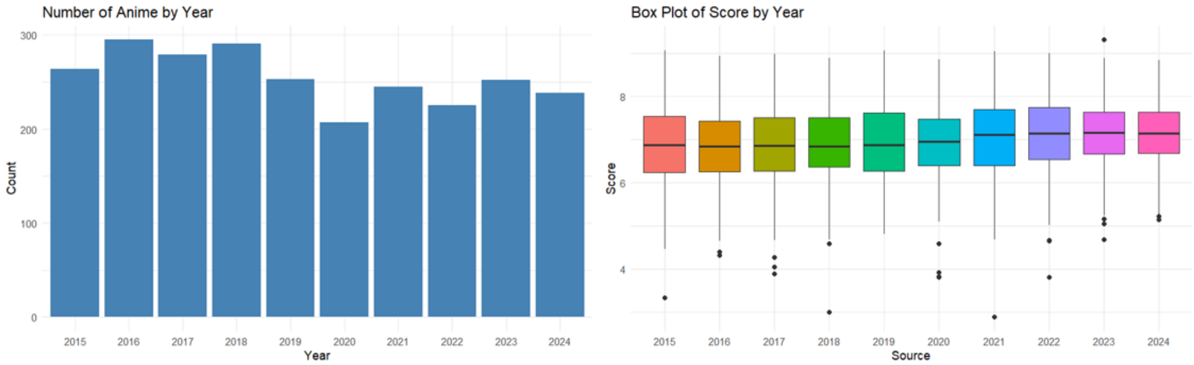
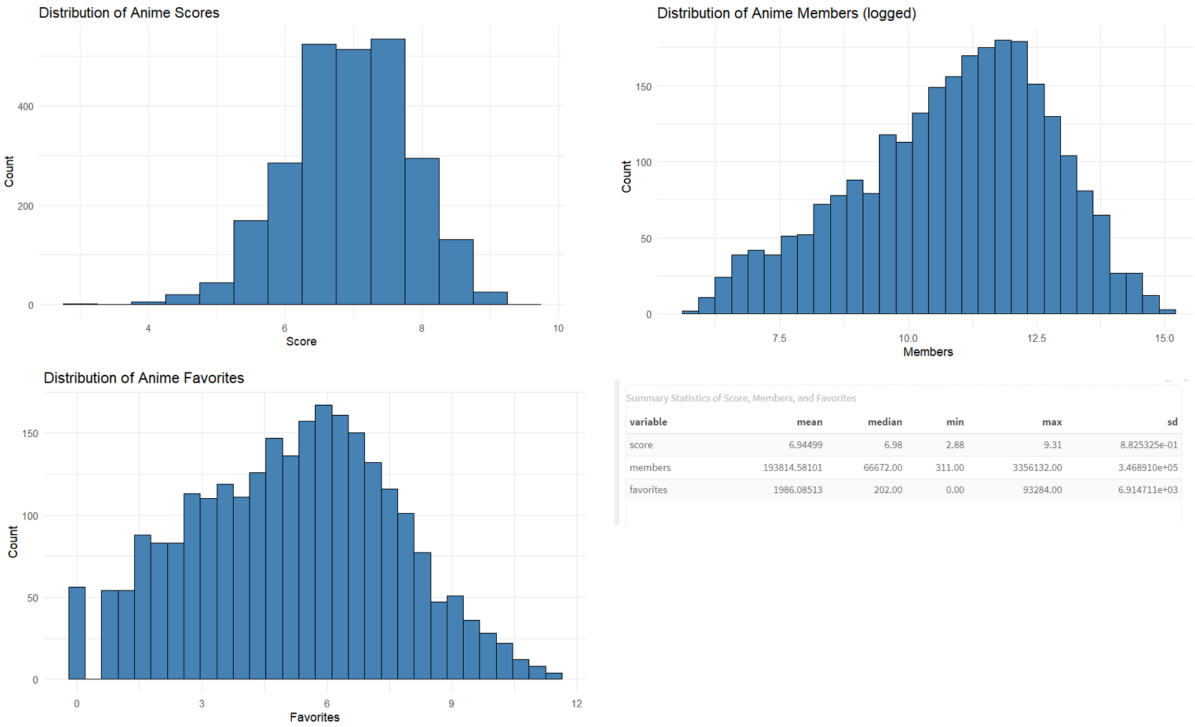


Figure 1: Distribution of Anime Count and Score over Year



Summary Statistics of Score, Members, and Favorites

variable	mean	median	min	max	sd
score	6.94499	6.98	2.88	9.31	8.825325e-01
members	193814.58101	66672.00	311.00	3356132.00	3.468910e+05
favorites	1986.08513	202.00	0.00	93284.00	6.914711e+03

Figure 2: Distribution of Continuous Variables (score, log members, log favorites)

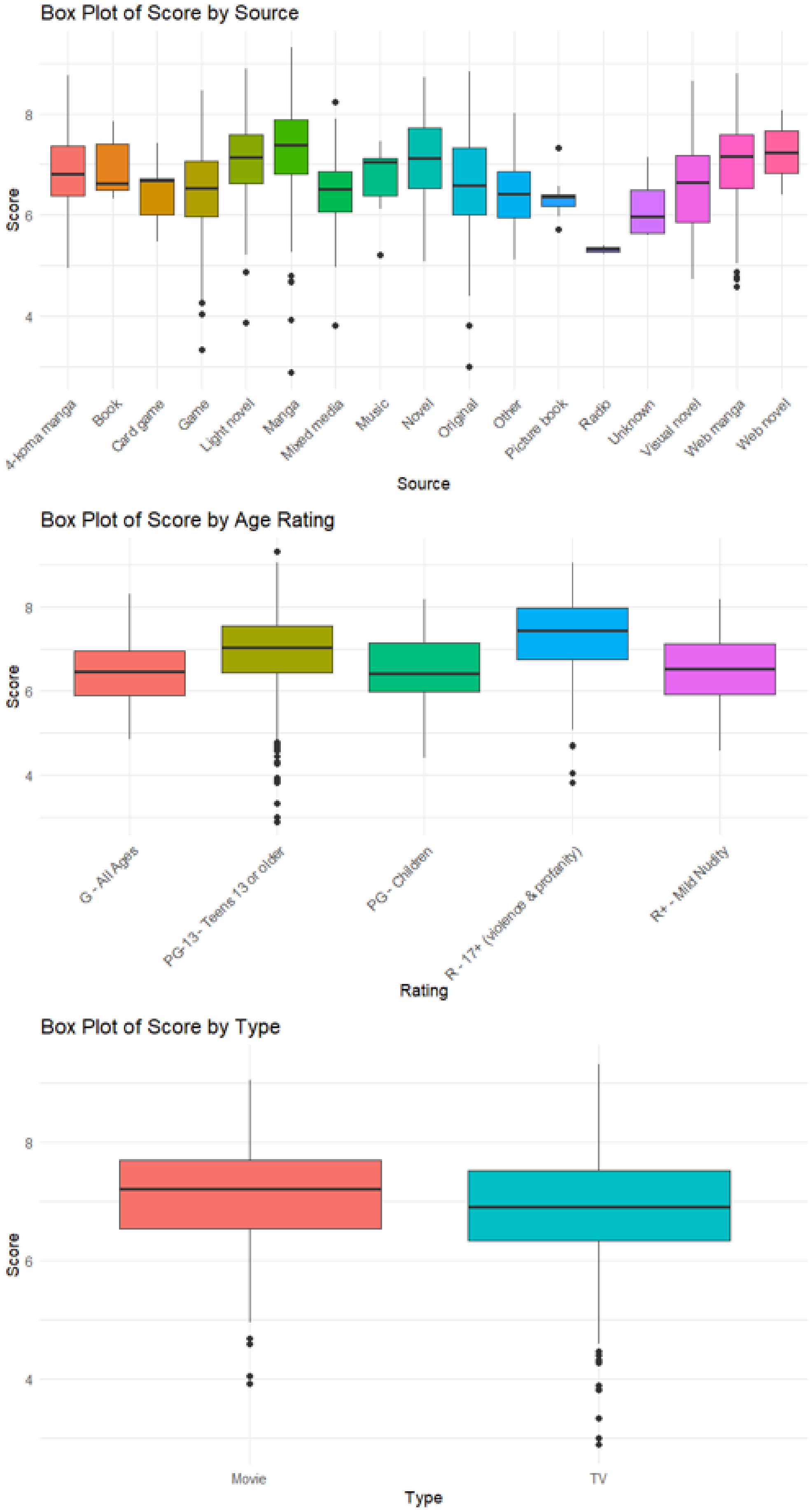
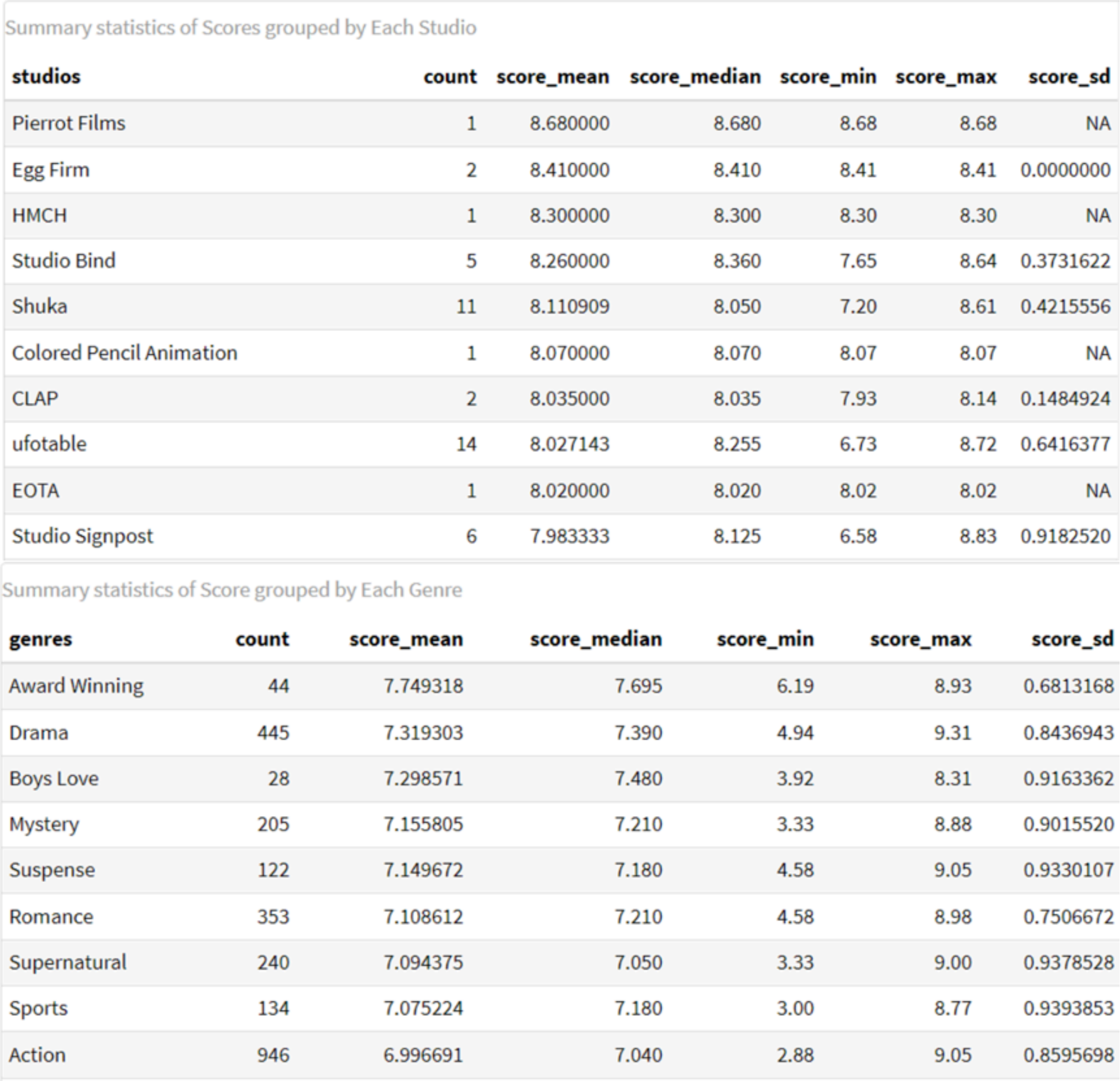


Figure 3: Distribution of Score over Categorical Variables, Boxplot



Summary statistics of Scores grouped by Each Studio

studios	count	score_mean	score_median	score_min	score_max	score_sd
Pierrot Films	1	8.680000	8.680	8.68	8.68	NA
Egg Firm	2	8.410000	8.410	8.41	8.41	0.0000000
HMCH	1	8.300000	8.300	8.30	8.30	NA
Studio Bind	5	8.260000	8.360	7.65	8.64	0.3731622
Shuka	11	8.110909	8.050	7.20	8.61	0.4215556
Colored Pencil Animation	1	8.070000	8.070	8.07	8.07	NA
CLAP	2	8.035000	8.035	7.93	8.14	0.1484924
ufotable	14	8.027143	8.255	6.73	8.72	0.6416377
EOTA	1	8.020000	8.020	8.02	8.02	NA
Studio Signpost	6	7.983333	8.125	6.58	8.83	0.9182520

Summary statistics of Score grouped by Each Genre

genres	count	score_mean	score_median	score_min	score_max	score_sd
Award Winning	44	7.749318	7.695	6.19	8.93	0.6813168
Drama	445	7.319303	7.390	4.94	9.31	0.8436943
Boys Love	28	7.298571	7.480	3.92	8.31	0.9163362
Mystery	205	7.155805	7.210	3.33	8.88	0.9015520
Suspense	122	7.149672	7.180	4.58	9.05	0.9330107
Romance	353	7.108612	7.210	4.58	8.98	0.7506672
Supernatural	240	7.094375	7.050	3.33	9.00	0.9378528
Sports	134	7.075224	7.180	3.00	8.77	0.9393853
Action	946	6.996691	7.040	2.88	9.05	0.8595698

Figure 4: Distribution of Score over Categorical Variables, Summary Table

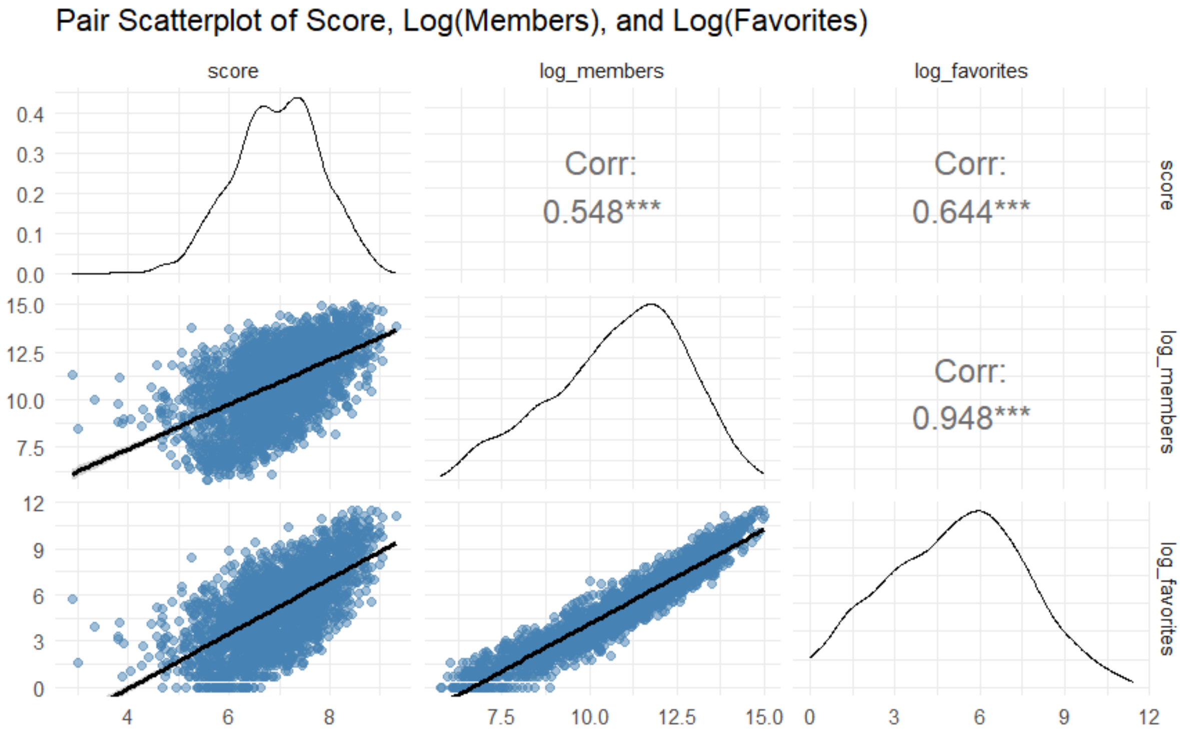


Figure 5: Pairwise Scatterplot of Continuous Variables