C6 Kaggle-Anime Dataset

Kristofer Klassen
Kert Karsna
Jako Aimsalu

Task 1. Workspace

Link to repository: https://github.com/Rxsengxn/examine-anime-dataset

Mark 1 if you solved the task fully, leave it 0 otherwise.

Task 1 - solved: 1

Task 2. Business understanding

In recent years, the number of anime released annually has risen to 200 animated films/series per year. Some releases are very successful, but not all are well-received by viewers. To understand some of the reasons that make an anime "highly rated" or, conversely, less appreciated, we can analyze the provided dataset.

The goal is to create a model that identifies relationships between popularity, ratings, genres, and the gender of the reviewers. Additionally, the dataset needs to be cleaned to simplify processing. The aim is to discover relationships among these variables to determine why certain anime are popular or if highly-rated anime also tends to be popular. Furthermore, we aim to identify specific genres or combinations of genres that resonate more with audiences in terms of both ratings and popularity. The success of this project depends on identifying these relationships.

We are using a dataset uploaded to Kaggle by Sajid, containing 13 columns and sized at 4.55 GB. To achieve this goal, the team consists of Kristofer, Kert, and Jako.

The goal must be completed by **December 9, 2024**. There are no strict rules, and the project is open-source, meaning everyone has the right to view the completed code and model. To achieve the goal, the code must enable us to find accurate correlations between variables.

Risks and Solutions

1. **Power Outages**
    - **Solution:** Work at someone else's place or a location like Delta.
2. **Internet Disruptions**
    - **Solution:** Use the same alternatives as for power outages.
3. **Time Constraints**
    - **Solution:** Properly allocate tasks among team members to avoid delays.
    - **Solution:** Optimizing or upgrading data processing equipment.

Terminology (English)

- **username**: The username of the user who rated the anime.
- **type**: The type of the anime (e.g., TV series, movie, OVA, etc.).
- **source**: The source material of the anime (e.g., manga, light novel, original, etc.).
- **score**: The overall score of the anime.
- **scored_by**: The number of users who have rated the anime.
- **rank**: The ranking of the anime.
- **genre**: The genre(s) of the anime.
- **my_score**: The rating score given by the user to the anime.
- **popularity**: The popularity rank of the anime.
- **gender**: The gender of the user.
- **title**: The title of the anime.
- **user_id**: The ID of the user.
- **anime_id**: The unique ID of the anime.

Financial Costs and Revenues

This project does not involve any financial costs or revenues.

Deliverables

By processing the data, we aim to create a prediction model that identifies why certain anime are popular or if a highly-rated anime also tends to be popular. Additionally, the project will explore whether there are genres or genre combinations that appeal more to audiences in terms of both ratings and popularity.

Finally, the project requires a poster to present as part of the **Introduction to Data Science** course. The dataset to be processed includes the following columns: `username`, `anime_id`, `my_score`, `user_id`, `gender`, `title`, `type`, `source`, `score`, `scored_by`, `rank`, `popularity`, and `genre`.

The goal of this data mining project is to understand user preferences and identify patterns in anime reviews by analyzing the dataset with over 35 million review entries. Key objectives include addressing scoring imbalances, such as the tendency for users to favor the lowest scoring disproportionately and exploring the relationships between anime genres, types and other attributes.

Success will be measured by the ability to generate actionable recommendations and ensure data consistency. Create a model that effectively predicts user preferences and delivers meaningful insights on this topic.

Mark 1 if you solved the task fully, leave it 0 otherwise.

Task 2 - solved: 1

Task 3. Data understanding

Gathering data

To have a good understanding to analyse different aspects of anime popularity and other correlation between viewers, there is a need to have a good amount of reviewers to which opinions a personality and preferences can be synthesised with. The required list of data that needs to be assessed:

- **Background info about an anime**
  - **Anime_id, title**: Processed by numerical ID (classified) and later title extracted based on that
  - **Genre, type, source**: list of categories under which anime belongs
- **User reviews**
  - **user_id, username**: Processed by numerical ID (classified) and later username extracted based on that
  - **gender**: Categorical value (male, female, other)
  - **my_score**: User score for particular anime in range of 0 to 10, where higher score corresponds to more likable anime
- **Anime generalization based on collection of scores**
  - **score**: numerical score ranging from 0 to 10.
  - **scored_by**: The number of reviews
  - **rank**: numeric value
  - **popularity**: numeric value

There exists multiple sets of anime datasets, with different year and data content variations. One will address the animes, with more background information, others are more towards recommendations to watch. The goal is to have more user specific popularity analysis/models and only anime generalizations are not enough. Since anime's rising popularity is fairly new then the most recent user scored database is needed.

The chosen dataset is currently also the latest "Anime Dataset 2023" where final_animedataset.csv includes individual entries of user scores (and also largest) for selected anime. Note that all the users have not reviewed all the possible animes present in the dataset and anime categorical belongings have to be divided so that they are comparable with other category sets.

Describing data

To accommodate better understanding of the audience, user score will be the best metric to look at preferences. With this if a user has given multiple reviews for different sets of animes, a correlation between the genre/type can be accomplished and the user themselves can be scored based on the general audience score board.

The dataset includes 35,300,00 review entries with 116,000 individual users and 8746 different anime. The number of potential descriptions needs to be evaluated since some preprocessing of dividing all the potential anime genre/source categories is needed.

Exploring data

Taking closer look at the distribution of the individual values there are some issues that were noted:

- Category imbalance among data columns
    - Gender: review count is very male dominant (72%)
    - Type: 67% of the being TV and second order 12% being movie
- Scoring bias
    - Lowest score is given in most cases, not in even pattern with other scores
- Non-binary genders
    - They are 1% of the data and doing generalization on them might not be accurate

Verifying data quality

With a large dataset, the imbalance might not be a big issue. Datasets generalized scores and data needs to be verified if there are any mismatches among the specific anime and its multiple data entries (score, scored_by, rank, popularity) are not the same and how to address that issue.

Dataset might have users that have manipulated scores on purpose and there's a need to address those occurrences. For example giving only extremumums of the score. Dataset might not also include enough background information of the users to make good generalizations/predictions based on actual user preferences.

Mark 1 if you solved the task fully, leave it 0 otherwise.

Task 3 - solved: 1

Task 4. The Plan

To Do:

- Write and submit this document (6 h)
- Set-up the environment (1 h)
- Import the dataset (3 h)
- Pre-process the dataset (12 h)
- Choose and import all needed models, classes and helper libraries (2 h)
- Decide if running tests on the whole dataset is feasible/sane (1 h)
- Examine the dataset and observe the features for missing/inimputable/unusable data (10 h)
- Process/analyse the data (25 h)
- Make a benchmark average model (4 h)
- Find trends and/or correlations between similar features (6 h)
- Find trends and/or correlations between categories (8 h)
- Report the findings in a nice-to-read/understand format (poster) (4 h)
- Make graphs/diagrams (5 h)
- Conclusions and takeaways (3 h)

Methods & tools:

- sklearn library – for machine learning model- and helper functions
- matplotlib - for visualizing data/findings
- Google services are our biggest helpers & tools

Link to repository: https://github.com/Rxsengxn/examine-anime-dataset

Mark 1 if you solved the task fully, leave it 0 otherwise.

Task 4 - solved: 1