Research of Anime Recommendation 2020

Jiayi Yang 512470
Zihao Zhao 515512
Tong Wang 515400
Yen-Cheng Chen 511210

CONTENTS

1 Data Description

2 Problem Statement

3 Why is this big data?

4 Method

5 Results

6 Conclusion

1. Data Description

Anime Recommendation Database 2020

2.78 GB Raw Data
693 MB Zip File

17562 Anime
325772 Users
35 Columns

Animation aired from
1998 to 2021

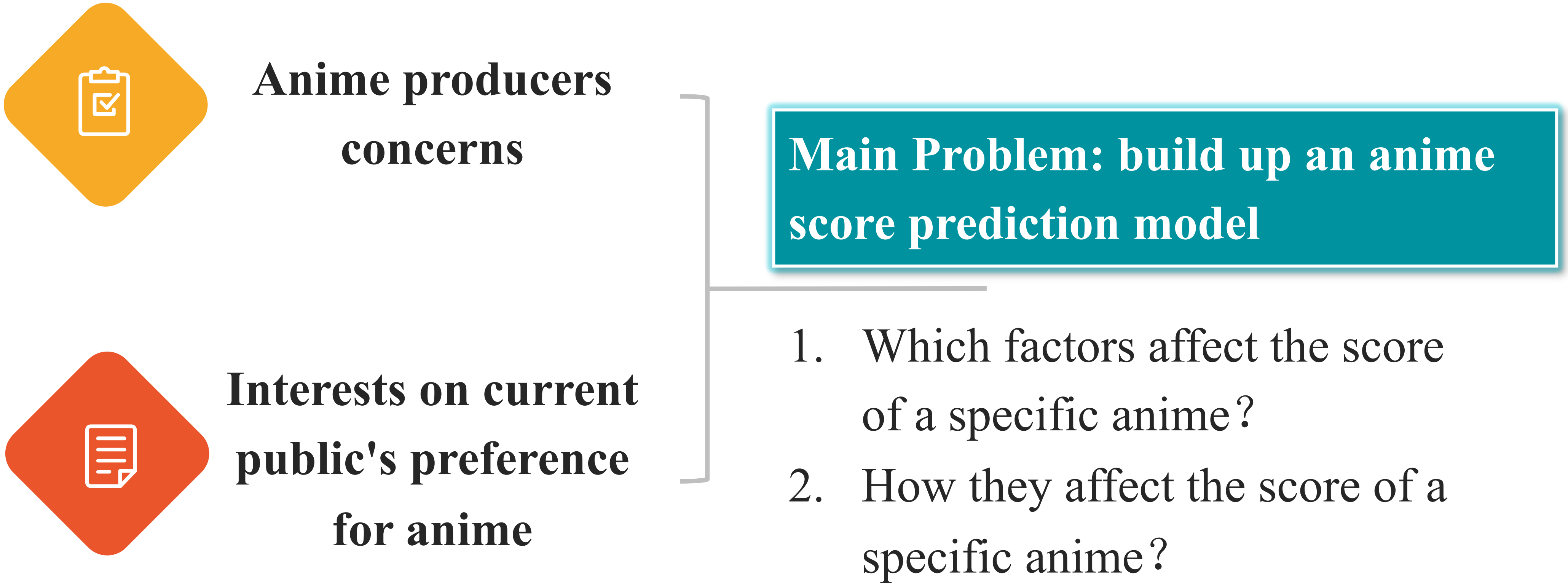
https://www.kaggle.com/datasets/hernan4444/anime-recommendation-database-2020

1 MapReduce

2 Hive & Impala

3 Pyspark

2. Problem Statement



Anime producers concerns

Interests on current public's preference for anime

Main Problem: build up an anime score prediction model

1. Which factors affect the score of a specific anime?
2. How they affect the score of a specific anime?

3. Why is this big data?

Why we select?

Factor Diversity & Timeliness

Why big data?

Processing difficulty with traditional tools

3V Principle

3V Principle

Volume

2.87 GB
693 MB

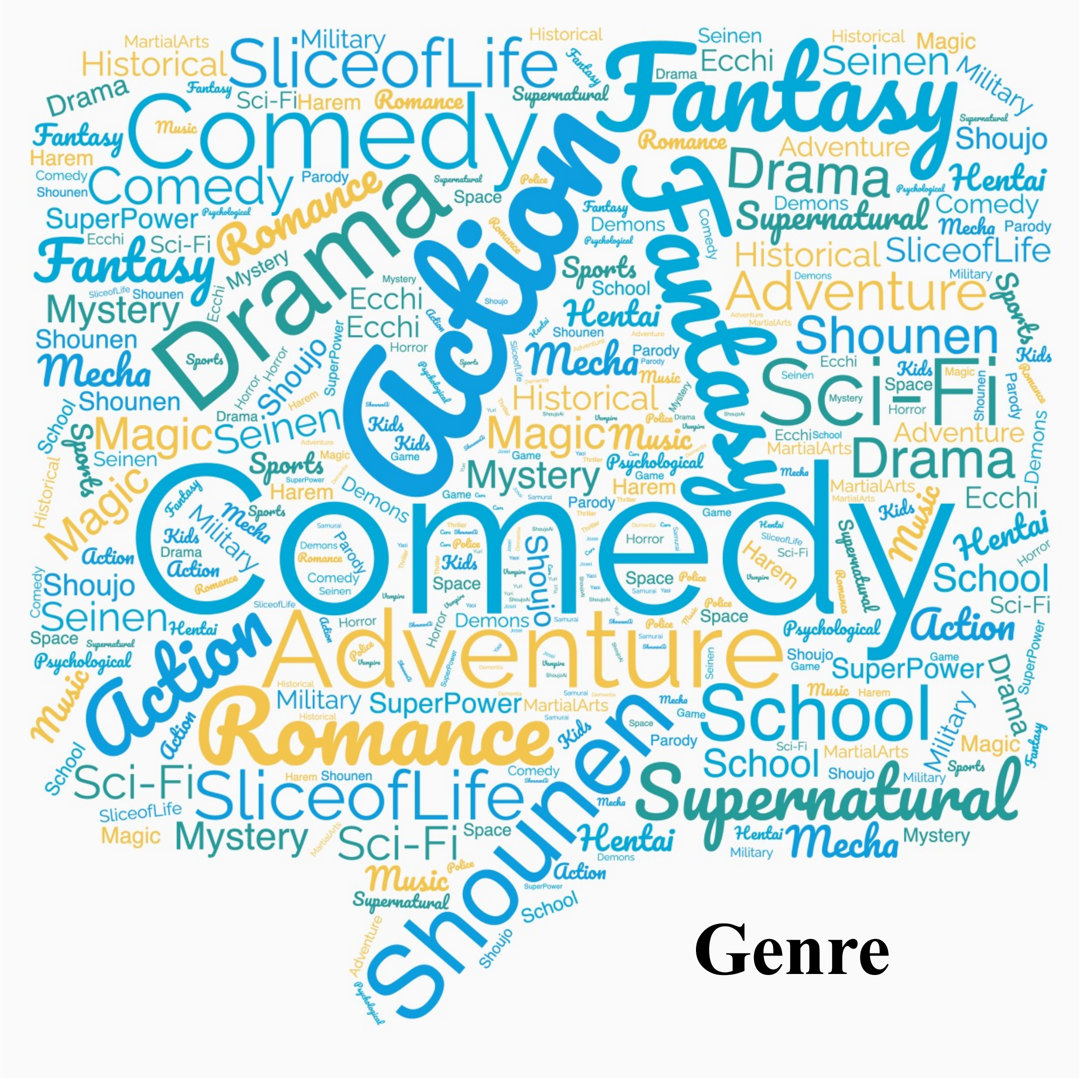
Velocity

Recommend
Timely

Variety

String
Number
Time

4. Method: Lead-in

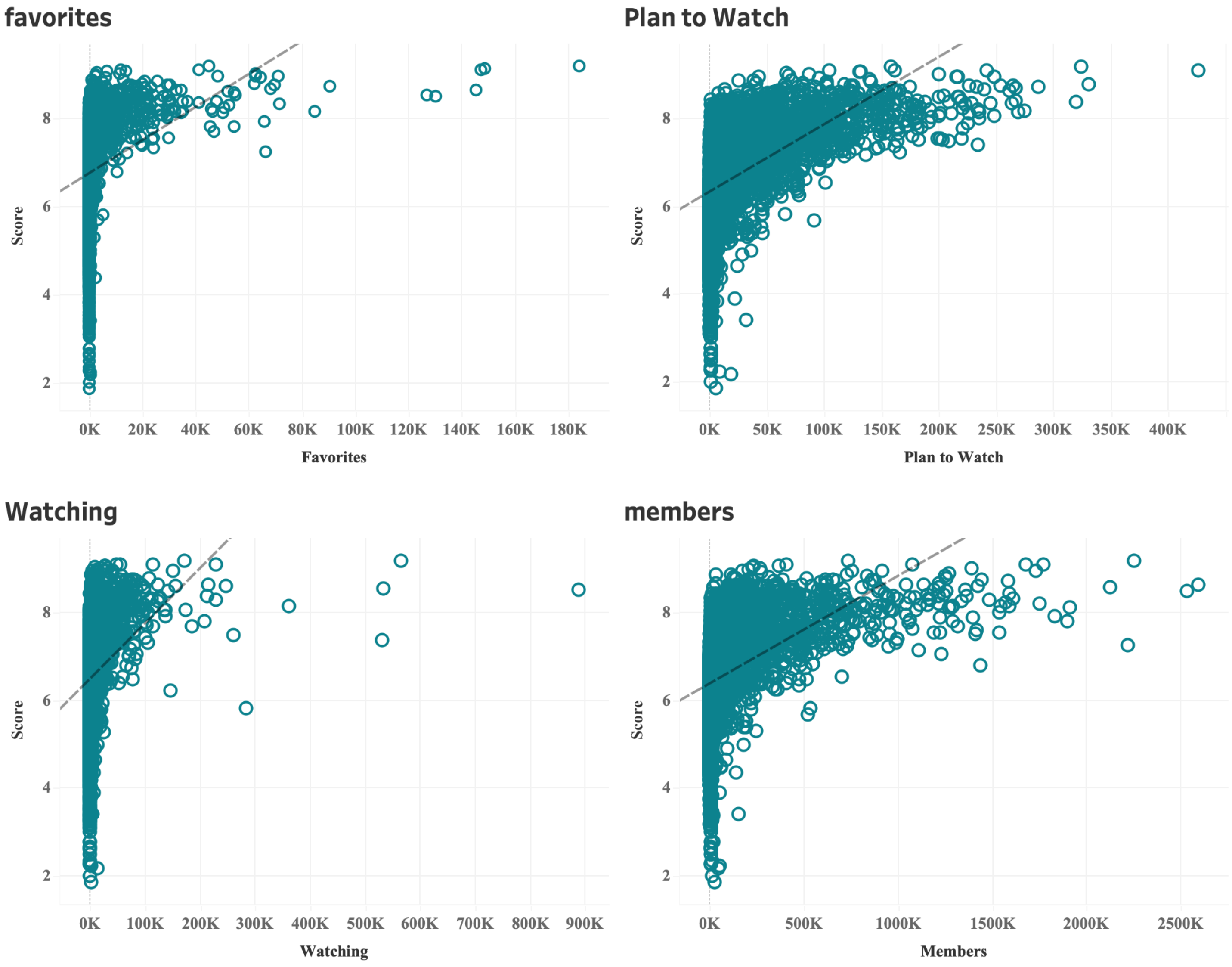


Genre

Filter: Score > 6.51

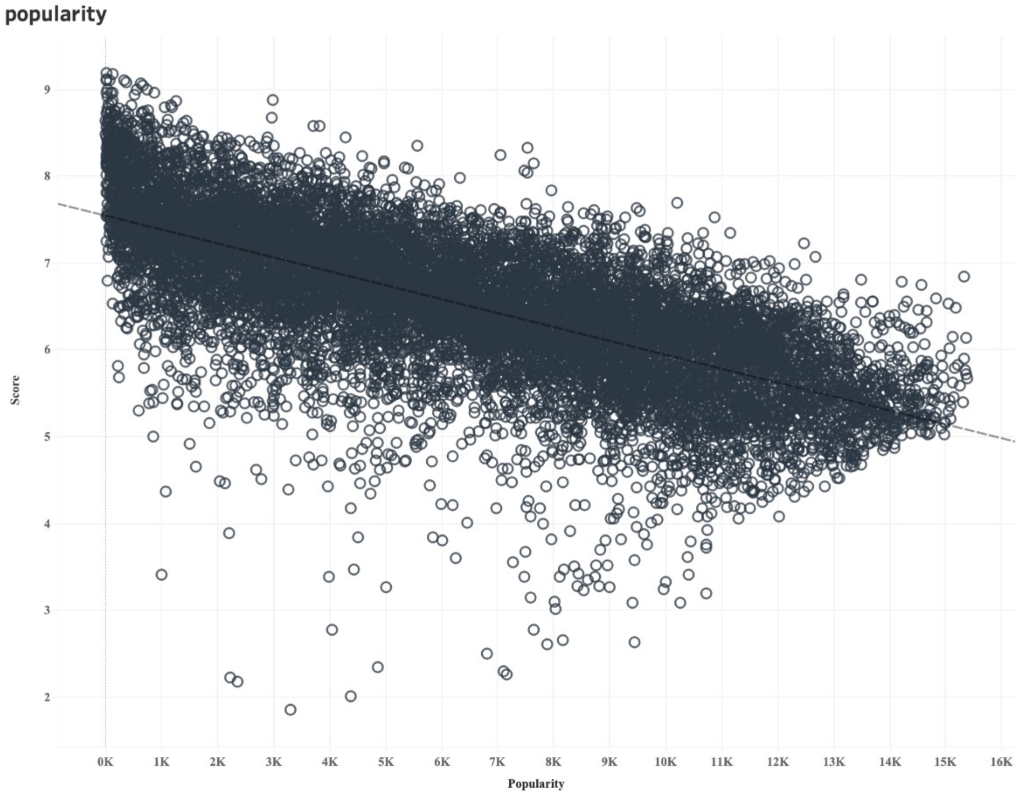
->**Comedy** and **Action** are the most common types.

4. Method: Variable Screening

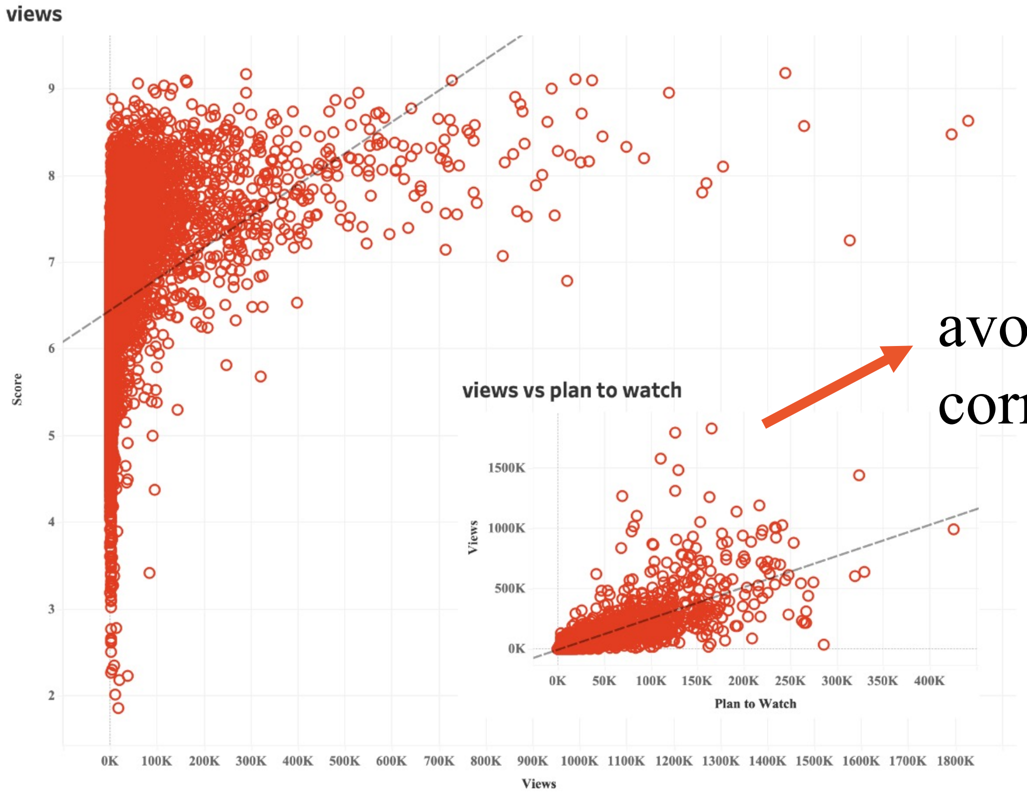


Favorites, Plan to Watch, Watching, and Members all **positively** correlated with Score.

4. Method: Variable Screening



There is **negative** correlation between Popularity and Score.



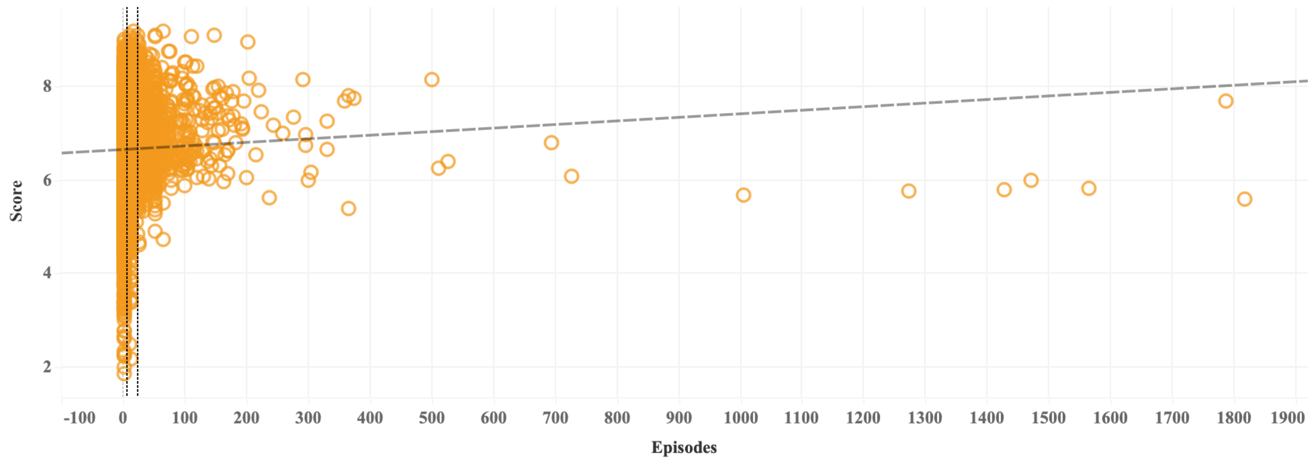
avoid large correlation

Views: count of all score rating

There is **positive** correlation between Views and Score.

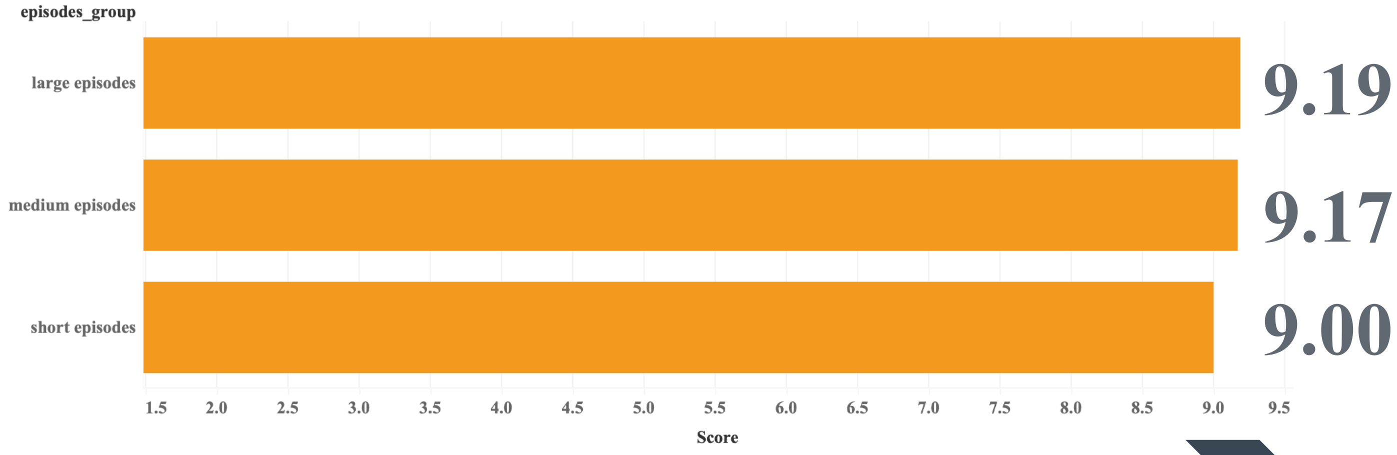
4. Method: Variable Screening

episodes



R-square value = 0.003
-> Low correlation

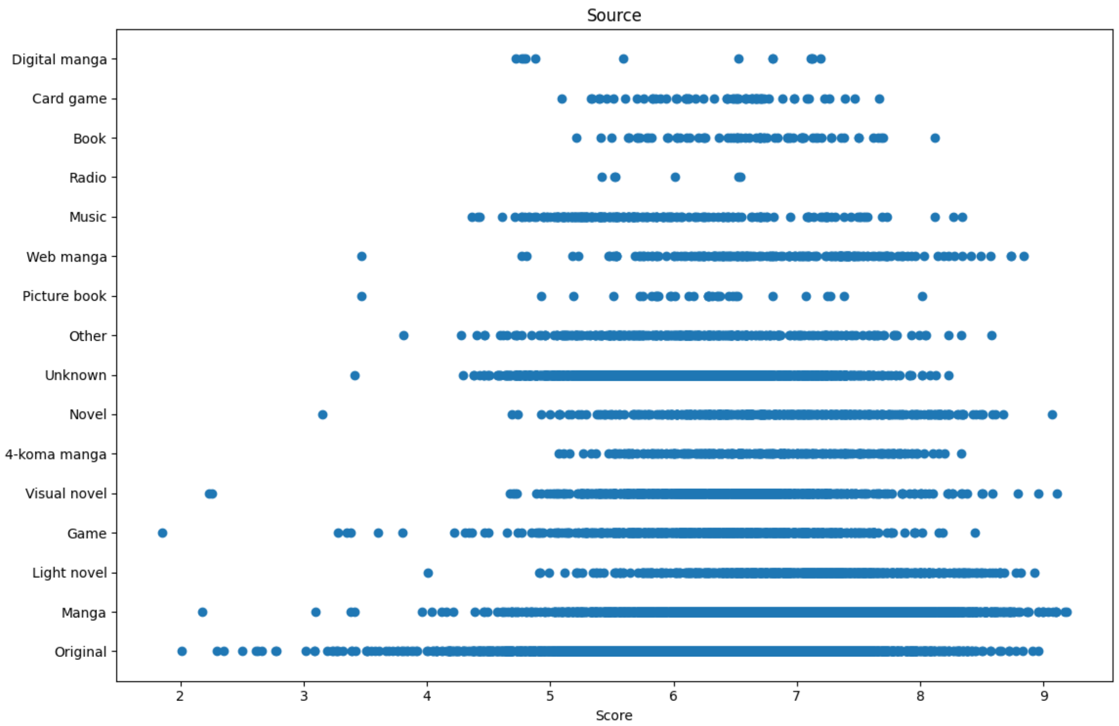
episodes_group

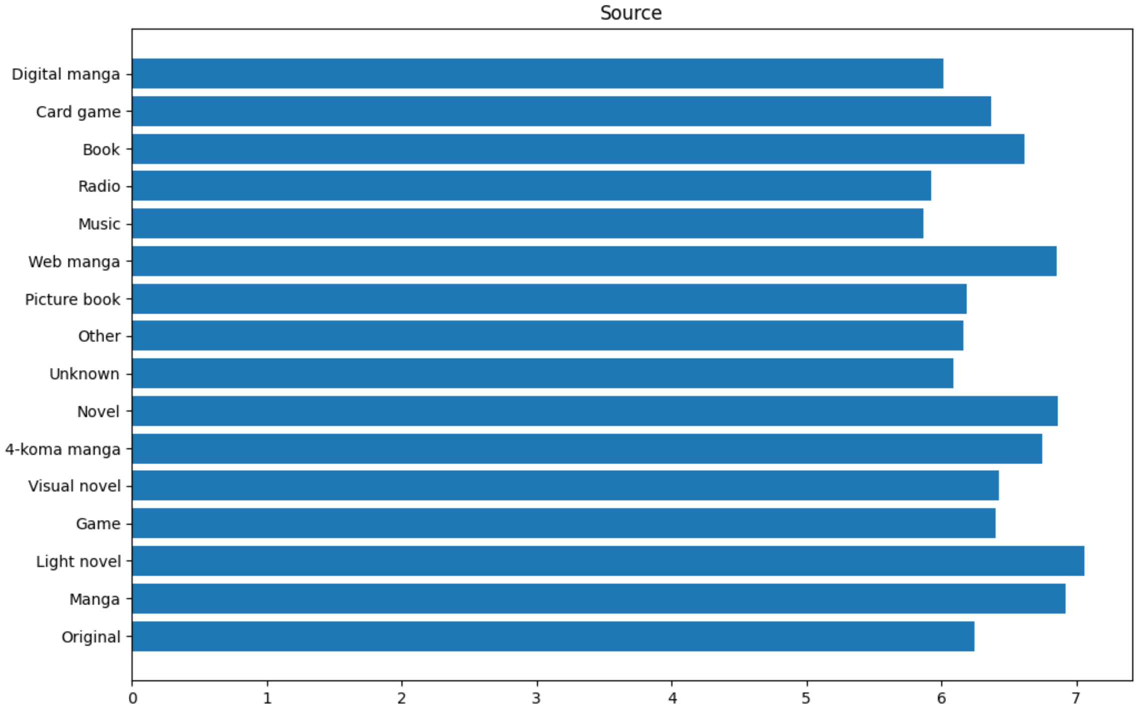


No obvious pattern

Remove attribute “Episodes”.

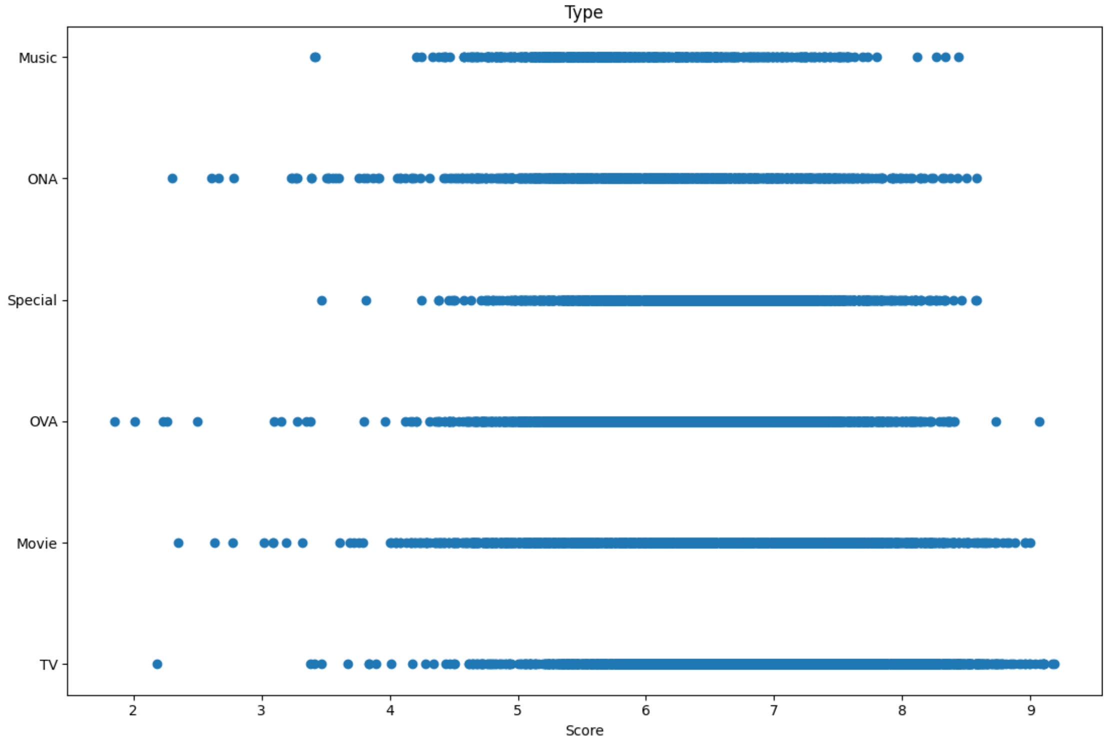
4. Method: Variable Screening

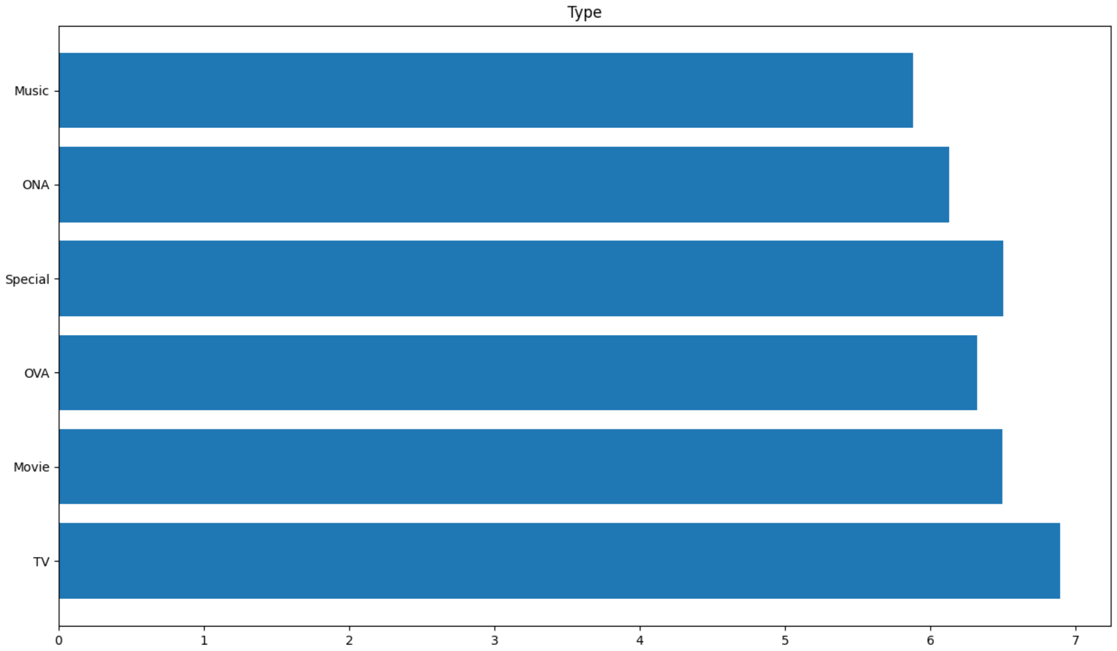




Light novel has **higher** score than other categories.

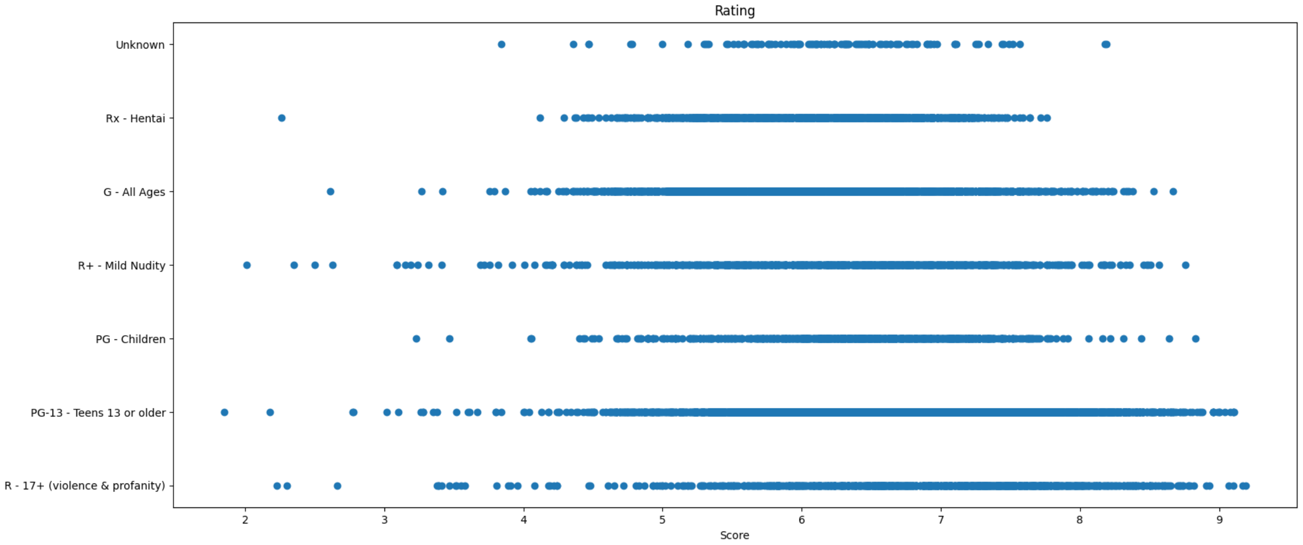
4. Method: Variable Screening



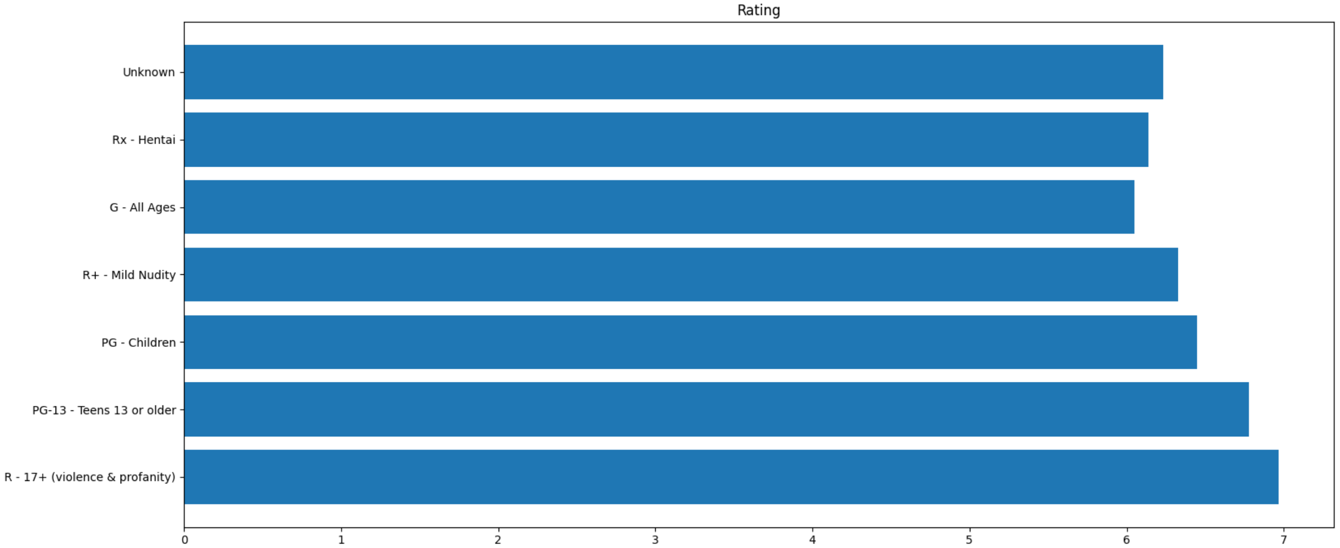


TV has **higher** score than other categories.

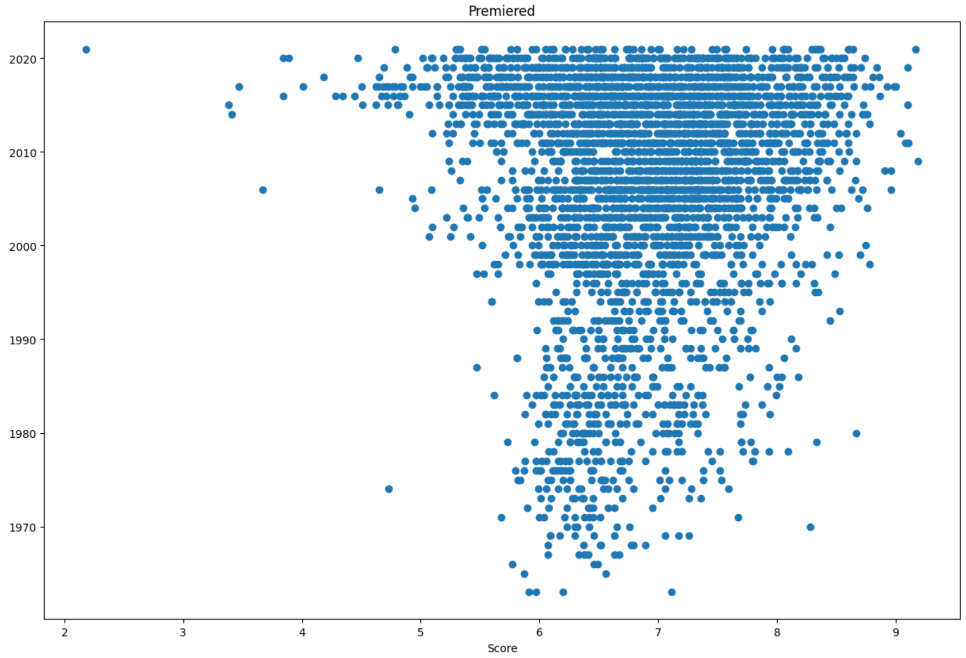
4. Method: Variable Screening



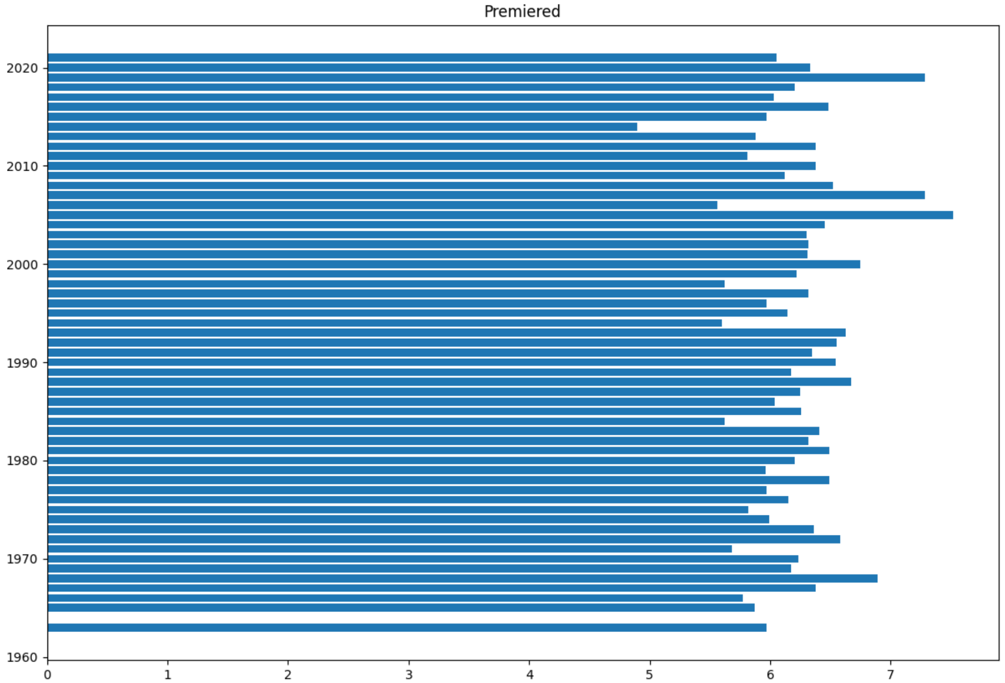
R 17+ (violence & profanity) has **higher** score than other categories.



4. Method: Variable Screening



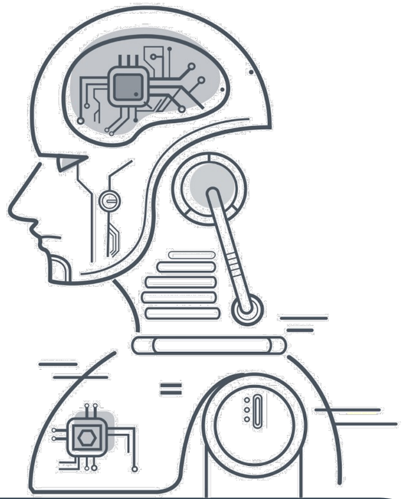
Anime premiered in 21st century has **higher** score than other categories



4. Method: Deep Learning Forecasting Linear Regression

- Dealing With NA Value
- Transform Data Type
- Create a New Column
- Choose Variables
- Convert Categorical Data

5. Results



LR Problem

X

Independent Variables

Type, Source, Rating

Popularity, Members, Favourites,

Watching, On-Hold, Dropped

Plan to Watch, Views

Y

Dependent Variables

Score

M

Model

$$Y = \alpha + \beta_1 X_1 + \cdots + \beta_n X_n + \varepsilon$$

70% Train 30% Test → MSE / R^2 → Accurate Prediction

6. Conclusion

Score Prediction Model is workable！

Popularity is not always an indication for high score. Anime that has source from light novel and rating of violence and profanity usually has higher score.

6. Conclusion: Limitations

Meet Strict Assumptions

Random error conform normal distribution
Deal with multicollinearity

Several Models For Predictions

Improve accuracy by classifying
animation by different criteria

Q&A