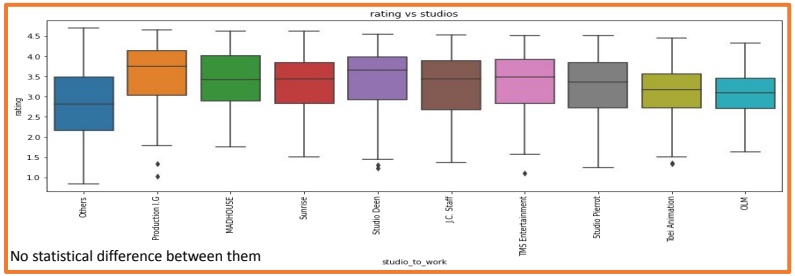
Group 1: Anime Ratings Prediction

Ligia Elena Jaimes
Daniel Dantas
Junxiao Li
Hamid Akbary
2022

Questions

- **What are the top rated anime?**
- **Are there specific genres or studios that influence in the rating?**
- **Create a regression model to predict the ratings.**
- **What are the most important features?**
- **Are the descriptions important to predict the ratings?**



Regression models:

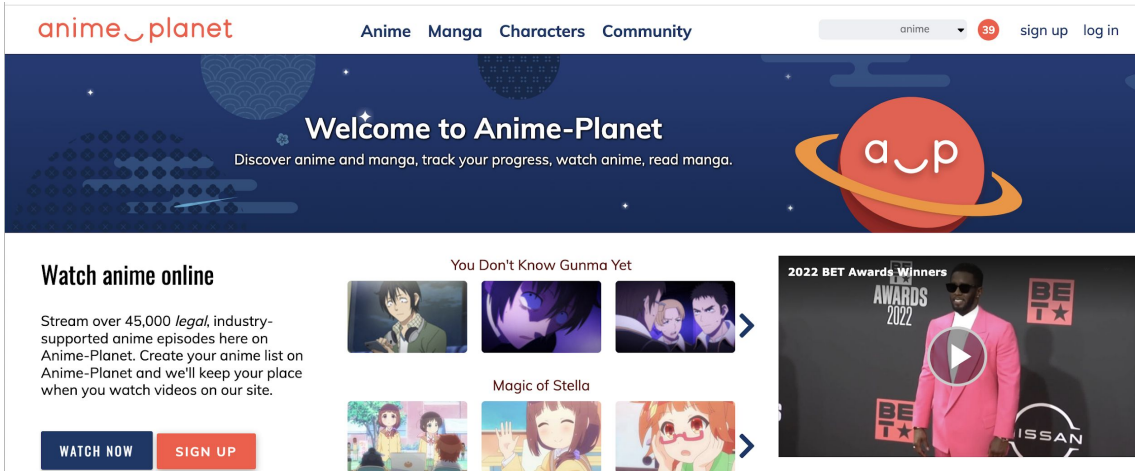
- ❖ Linear regression
- ❖ Lasso
- ❖ XGBRegressor

Description is not important in the rating prediction

The 2 most important features:

- Dropped → strong negative impact for rating
- Media_TV

Motivation



Issue: There is so much new anime coming out each season. Anime lovers heavily rely on the ratings. Low rating anime is usually forgotten even if they came out just few months ago.

Key: To find the most decisive factors of the rating and provide anime platform such as “anime planet” recommendations of the anime advertisements. So as to improving the page view and increase net profit.

Animate data

Features

Data Description

You are provided with anime (Japanese animated media) information. Data from anime-planet on June 15 2020.

> File: anime.csv

- **eps**: number of episodes (movies are considered 1 episode)
- **duration**: duration of episode
- **ongoing**: weather it is ongoing
- **startYr**: year that airing started
- **finishYr**: year that airing finished
- **sznOfRelease**: season of release (Winter, Spring, Fall)
- **description**: synopsis of plot
- **studios**: studios responsible for creation
- **tag**: tags, genres, etc.
- **contentWarn**: content warning
- **watched**: number of users that completed it
- **watching**: number of users that are watching it
- **wantWatch**: number of users that want to watch it
- **dropped**: number of users that dropped it before completion
- **rating**: average user rating
- **votes**: number of votes that contribute to rating

Animate data

14578 rows × 48 columns

eps	14219.00	13.50	62.26	1.00	1.00	1.00	12.00	2527.00
startYr	14356.00	2005.46	14.71	1907.00	2000.00	2010.00	2016.00	2026.00
finishYr	14134.00	2005.52	14.66	1907.00	2000.00	2010.00	2016.00	2026.00
contentWarn	14578.00	0.10	0.30	0.00	0.00	0.00	0.00	1.00
watched	14356.00	2408.04	7168.37	0.00	25.00	165.00	1469.50	161567.00
watching	14578.00	213.03	1261.71	0.00	1.00	7.00	63.00	74537.00
wantWatch	14578.00	1021.73	2145.01	0.00	24.00	175.00	980.00	28541.00
dropped	14578.00	125.96	453.58	0.00	1.00	7.00	40.00	19481.00
rating	12107.00	2.95	0.83	0.84	2.30	2.96	3.62	4.70
votes	12119.00	2085.79	5946.28	10.00	34.00	218.00	1412.50	131067.00

Features

31 Columns are Tag

tag_'Comedy'	14578.00	0.26	0.44	0.00	0.00	0.00	1.00	1.00
tag_'Based on a Manga'	14578.00	0.26	0.44	0.00	0.00	0.00	1.00	1.00
tag_'Action'	14578.00	0.21	0.41	0.00	0.00	0.00	0.00	1.00
tag_'Fantasy'	14578.00	0.18	0.38	0.00	0.00	0.00	0.00	1.00
tag_'Sci Fi'	14578.00	0.15	0.36	0.00	0.00	0.00	0.00	1.00
tag_'Shounen'	14578.00	0.13	0.33	0.00	0.00	0.00	0.00	1.00
tag_'Family Friendly'	14578.00	0.13	0.33	0.00	0.00	0.00	0.00	1.00
tag_'Original Work'	14578.00	0.13	0.33	0.00	0.00	0.00	0.00	1.00
tag_'Non-Human Protagonists'	14578.00	0.12	0.33	0.00	0.00	0.00	0.00	1.00
tag_'Adventure'	14578.00	0.11	0.31	0.00	0.00	0.00	0.00	1.00
tag_'Short Episodes'	14578.00	0.10	0.31	0.00	0.00	0.00	0.00	1.00
tag_'Drama'	14578.00	0.10	0.30	0.00	0.00	0.00	0.00	1.00
tag_'Shorts'	14578.00	0.09	0.29	0.00	0.00	0.00	0.00	1.00
tag_'Romance'	14578.00	0.08	0.28	0.00	0.00	0.00	0.00	1.00
tag_'School Life'	14578.00	0.08	0.27	0.00	0.00	0.00	0.00	1.00
tag_'Slice of Life'	14578.00	0.08	0.27	0.00	0.00	0.00	0.00	1.00
tag_'Animal Protagonists'	14578.00	0.07	0.26	0.00	0.00	0.00	0.00	1.00
tag_'Seinen'	14578.00	0.07	0.25	0.00	0.00	0.00	0.00	1.00
tag_'Supernatural'	14578.00	0.06	0.25	0.00	0.00	0.00	0.00	1.00
tag_'Magic'	14578.00	0.06	0.23	0.00	0.00	0.00	0.00	1.00
tag_'CG Animation'	14578.00	0.06	0.23	0.00	0.00	0.00	0.00	1.00
tag_'Mecha'	14578.00	0.05	0.22	0.00	0.00	0.00	0.00	1.00
tag_'Ecchi'	14578.00	0.05	0.22	0.00	0.00	0.00	0.00	1.00
tag_'Based on a Light Novel'	14578.00	0.05	0.21	0.00	0.00	0.00	0.00	1.00
tag_'Anthropomorphic'	14578.00	0.04	0.21	0.00	0.00	0.00	0.00	1.00
tag_'Superpowers'	14578.00	0.04	0.19	0.00	0.00	0.00	0.00	1.00
tag_'Promotional'	14578.00	0.04	0.19	0.00	0.00	0.00	0.00	1.00
tag_'Sports'	14578.00	0.04	0.19	0.00	0.00	0.00	0.00	1.00
tag_'Historical'	14578.00	0.04	0.19	0.00	0.00	0.00	0.00	1.00
tag_'Vocaloid'	14578.00	0.03	0.18	0.00	0.00	0.00	0.00	1.00
tag_Others	14578.00	0.08	0.27	0.00	0.00	0.00	0.00	1.00

Data pre-processing

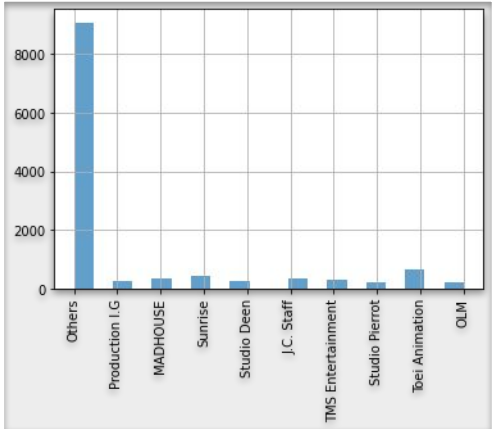
Pre-processing analysis

- Check for duplications
- Missing data
- Dropping in target “rating” and “title”
- Studios
  - Remove brackets
  - Fill missing values “Others”
  - Choose to work with “10 studios” in studios_to_work

sznOfRelease	8560
duration	4636
description	4474
finishYr	121
watched	115
mediaType	63
startYr	6
tag_'Mecha'	0
tag_'Romance'	0
tag_'School Life'	0
tag_'Slice of Life'	0
tag_'Animal Protagonists'	0
tag_'Seinen'	0
tag_'Supernatural'	0
tag_'Magic'	0
tag_'CG Animation'	0
tag_'Anthropomorphic'	0
tag_'Ecchi'	0
tag_'Based on a Light Novel'	0
tag_'Drama'	0
tag_'Superpowers'	0
tag_'Promotional'	0
tag_'Sports'	0
tag_'Historical'	0
tag_'Vocaloid'	0
tag_'Shorts'	0
tag_'Original Work'	0
tag_'Short Episodes'	0
tag_'Adventure'	0
ongoing	0
studios	0
contentWarn	0
watching	0
wantWatch	0
dropped	0
rating	0
votes	0
tag_'Comedy'	0
tag_'Based on a Manga'	0
tag_'Action'	0
tag_'Fantasy'	0
tag_'Sci Fi'	0

	count	unique	top	freq
title	14578	14578	Fullmetal Alchemist: Brotherhood	1
mediaType	14510	8	TV	4510
duration	9137	147	4min	964
ongoing	14578	2	False	14356
sznOfRelease	3767	4	Spring	1202
description	8173	8108	In 19th century Belgium, in the Flanders count...	3
studios	14578	864	[]	4808

	count	unique	top	freq
mediaType	12044	8	TV	3993
duration	7471	144	4min	844
ongoing	12107	2	False	11992
sznOfRelease	3547	4	Spring	1135
description	7633	7572	In 19th century Belgium, in the Flanders count...	3
studios	12107	810	'Others'	3214



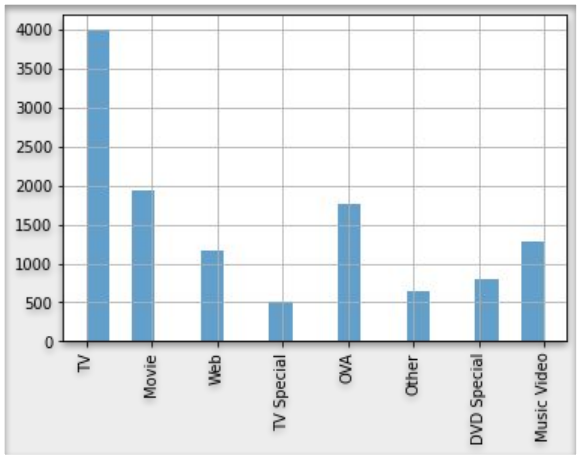
Pre-processing analysis

- Check for duplications
- Missing data
- Dropping in target “rating” and “title”
- Studios
  - Remove brackets
  - Fill missing values “Others”
  - Choose to work with “10 studios” in studios_to_work

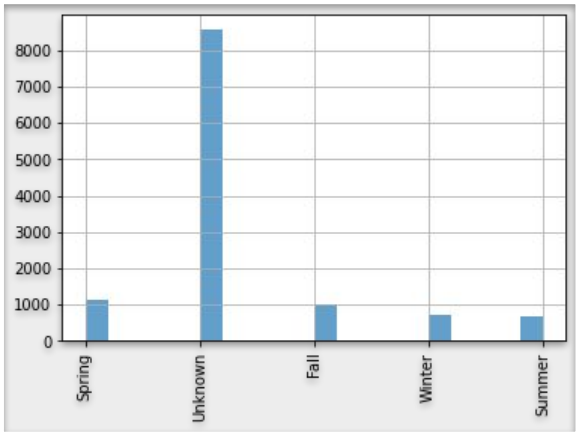
- startYr
  - Drop nan in startYr
- finishYr
  - Fill with 2020
  - Create a new feature: years_running = finishYr – startYr
  - Drop startYr and finishYr
- MediaType
  - Fill with “Other” add to existing feature
- szn0fRelease
  - Fill with “Unknown”
- duration & watched
  - Transform time (hr/min → min)
  - Fill by group (studio & mediatype)

Distribution

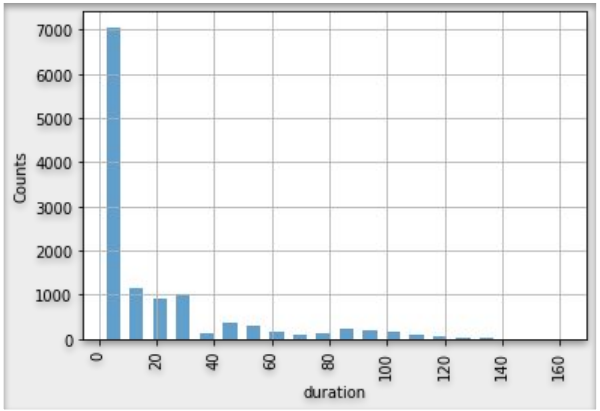
MediaType



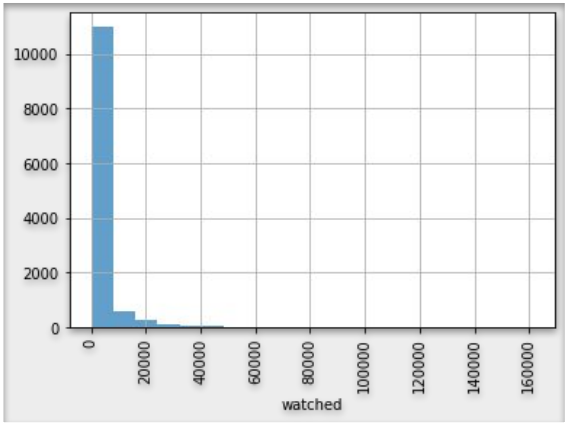
Season Release



duration

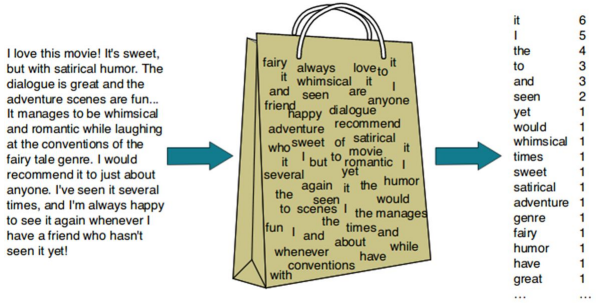


watched



Pre-processing analysis

- Check for duplications
- Missing data
- Dropping in target “rating” and “title”
- Studios
  - Remove brackets
  - Fill missing values “Others”
  - Choose to work with “10 studios” in studios_to_work
- startYr
  - Drop nan in startYr
- finishYr
  - Fill with 2020
  - Create a new feature: years_running = finishYr – startYr
  - Drop startYr and finishYr
- MediaType
  - Fill with “Other” add to existing feature
- szn0fRelease
  - Fill with “Unknown”
- duration & watched
  - Transform time (hr/min → min)
  - Fill by group (studio & mediatype)
- Predictions
  - Remove punctuation
  - Remove stop words
  - Create a word cloud



Term frequency-inverse document frequency (**TF-IDF**) with n-grams

Common locally:

$$\mathrm{TF}(t,d) = f_{t,d}$$

Global rarity:

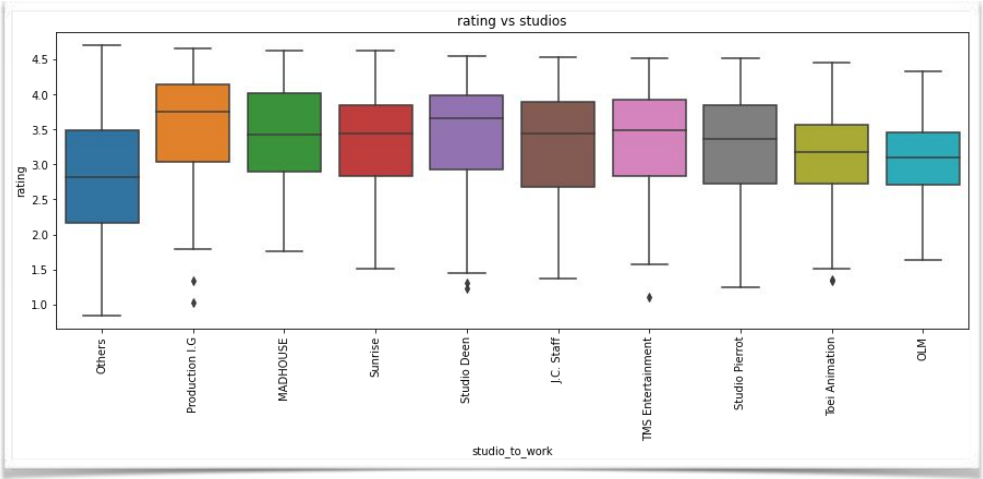
$$\mathrm{IDF}(t) = \log\left(\frac{N}{1+n_t}\right)$$

$$\text{TF-IDF}(t,d) = f_{t,d} \cdot \log\left(\frac{N}{1+n_t}\right)$$

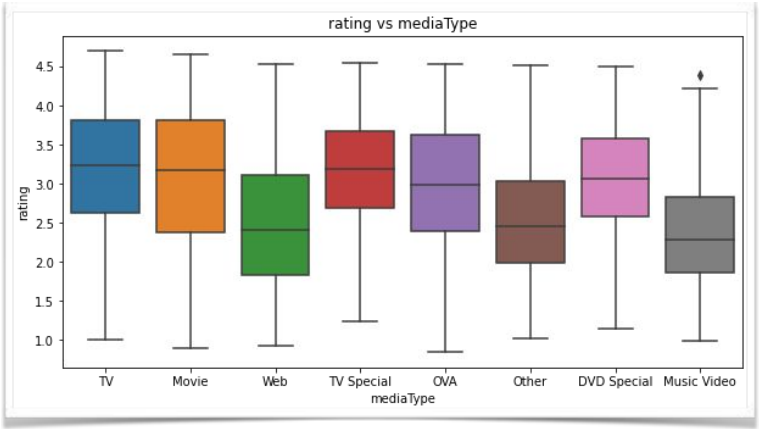
Data visualization

Bivariable analysis

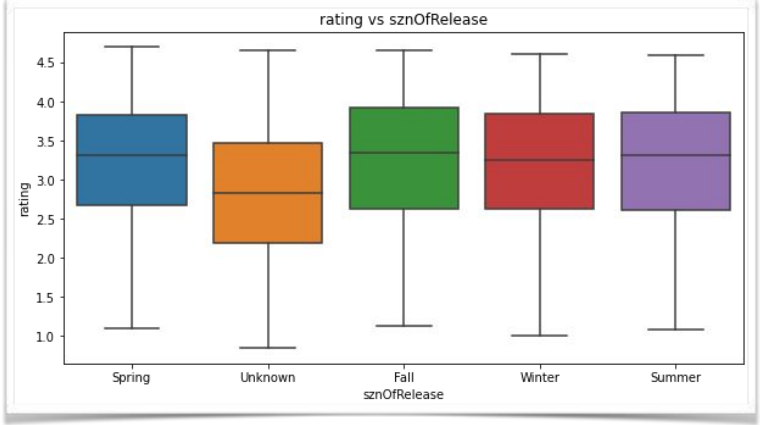
Analysis of two variables: cross-correlation relationship



Average rating has no much difference, except for the others, the average rating is lower.



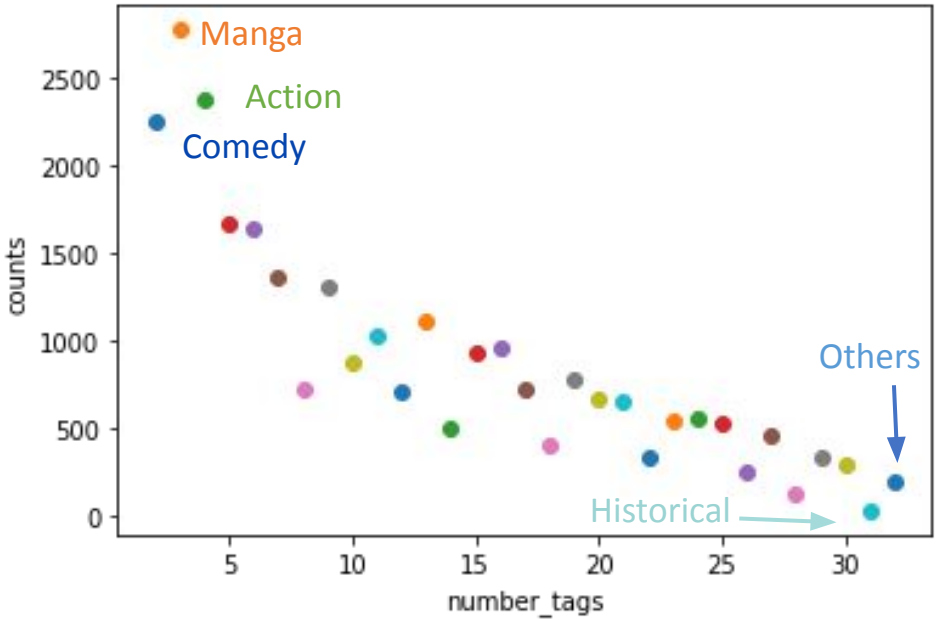
Too much overlapping, no distinctive relationships with rating. Music video and Web have the lowest average rating.



Correlations almost overlap, almost no rating difference with different release season, which is an obvious truth.

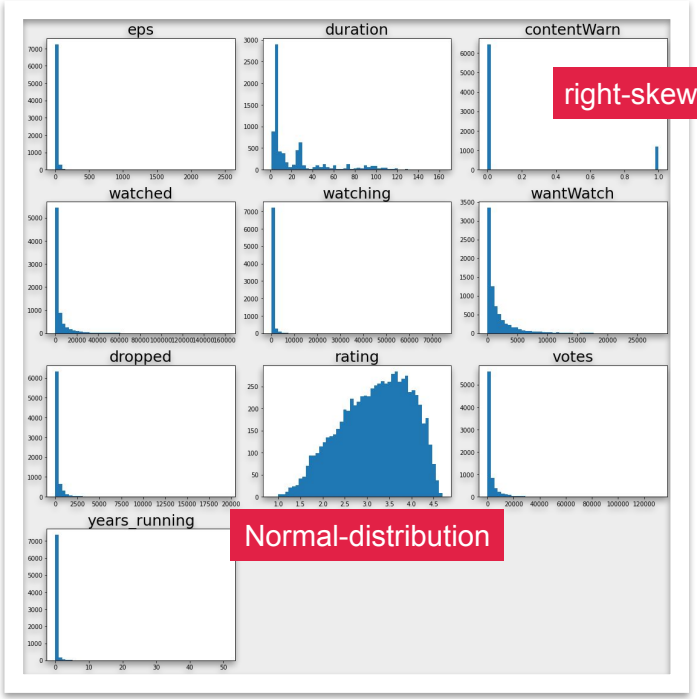
```
0    8799
1    3302
Name: tag_'Comedy', dtype: int64
--------------------------------------------------
0    8582
1    3519
Name: tag_'Based on a Manga', dtype: int64
--------------------------------------------------
0    9303
1    2798
Name: tag_'Action', dtype: int64
--------------------------------------------------
0    9904
1    2197
Name: tag_'Fantasy', dtype: int64
--------------------------------------------------
```

Tag analysis

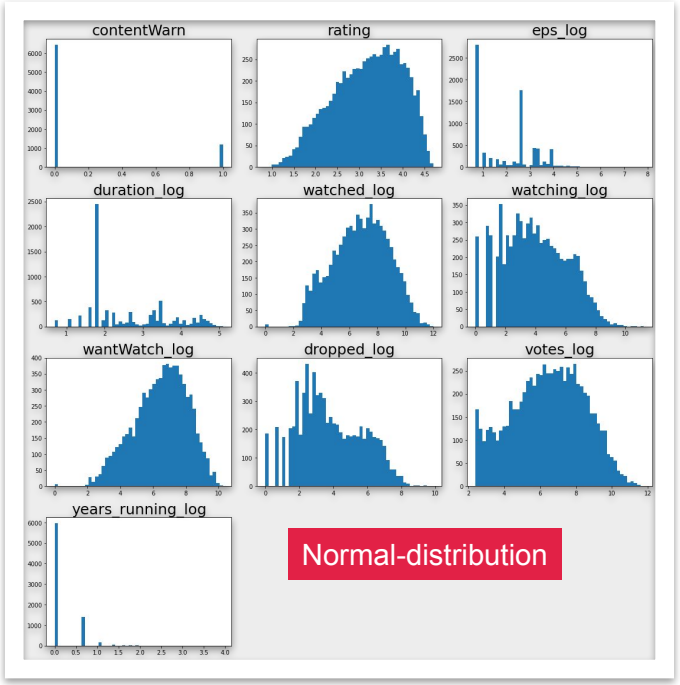



```
--------------------------------------------------
0    11561
1      540
Name: tag_'Superpowers', dtype: int64
--------------------------------------------------
0    11661
1      440
Name: tag_'Promotional', dtype: int64
--------------------------------------------------
0    11641
1      460
Name: tag_'Sports', dtype: int64
--------------------------------------------------
0    11698
1      403
Name: tag_'Historical', dtype: int64
--------------------------------------------------
0    11625
1      476
Name: tag_'Vocaloid', dtype: int64
--------------------------------------------------
0    11200
1      901
Name: tag_Others, dtype: int64
--------------------------------------------------
```

Variable transformation

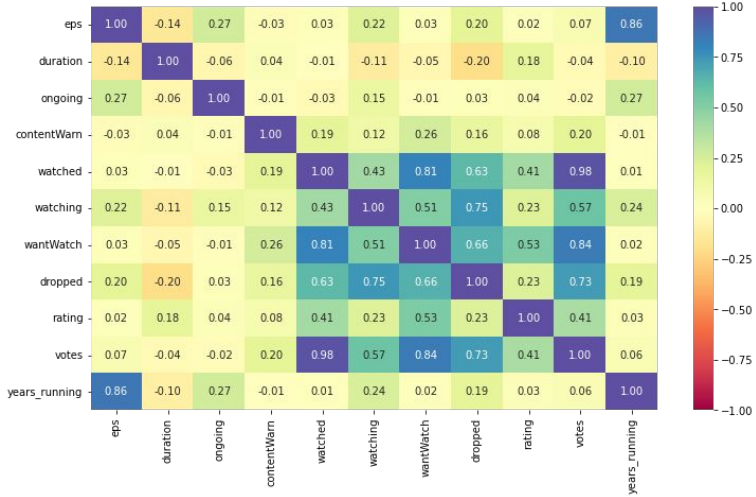


log transforms

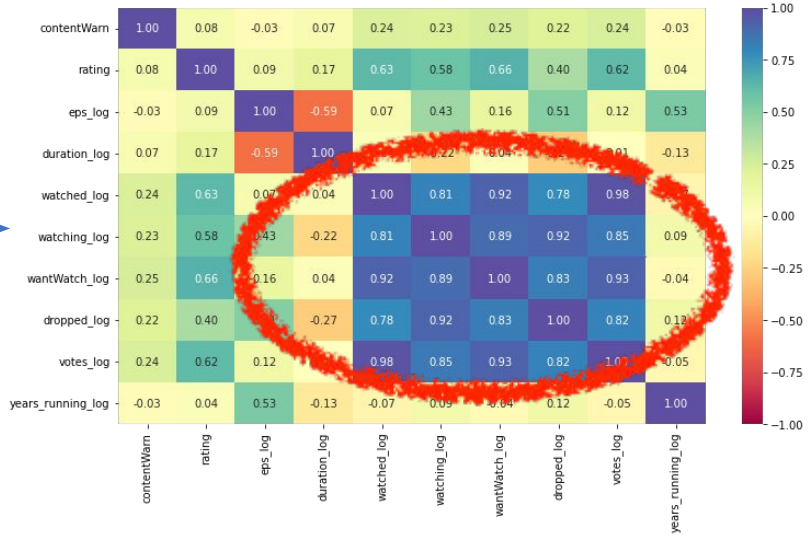


Better distribution to some of the features.

Correlation between columns



log transforms



few highly correlated columns.

tag is not consider for correlation check as they have only 0 or 1 values.**

Modeling

Linear regression

Linear regression: is a linear approach for modelling the relationship.

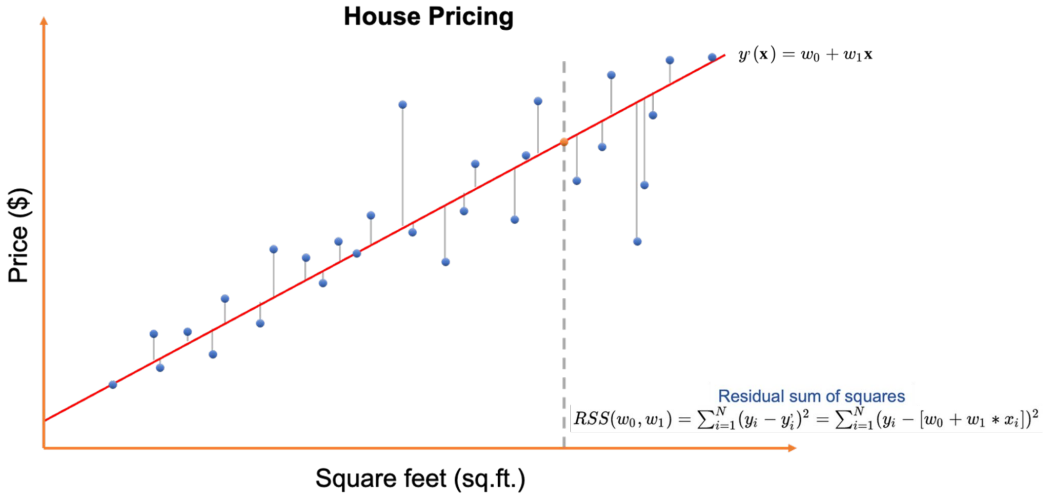
Overfitting?

Pros

- Robust
- Easy to understand and interpret
- Low computing cost

Cons

- Only capture linear correlation
- Affected by outliers
- Requires normalized inputs
- Can't easily handle categorical features



Lasso (least absolute shrinkage and selection operator)

Lasso regression: **a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the resulting statistical model**

$$y'(\mathbf{x}) = w_0 + w_1\mathbf{x} + w_2\mathbf{x}^2 + \cdots + w_p\mathbf{x}^p$$

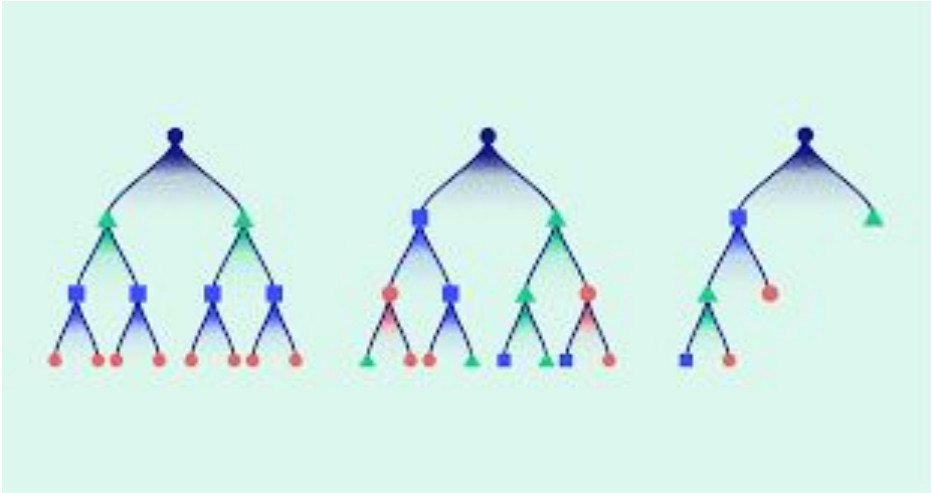
Lasso regression sets some weights to zero

Lasso regression is a parsimonious model that performs L1 regularization. The L1 regularization adds a penalty equivalent to the absolute magnitude of regression coefficients and tries to minimize them.

XGBRegressor Model

Extreme Gradient Boosting, or XGBoost is an efficient open-source implementation of the gradient boosting algorithm.

The two main reasons to use XGBoost are execution speed and model performance.



Important hyper-parameters:

- **n_estimators**: The number of trees in the ensemble, often increased until no further improvements are seen.
- **max_depth**: The maximum depth of each tree, often values are between 1 and 10.
- **subsample**: The number of samples (rows) used in each tree(0-1)
- **colsample_bytree**: Number of features (columns) used in each tree(0-1)

Model performance check

Regression accuracy metrics

- **RMSE** (Root Mean Squared Error) is the error rate by the square root of MSE. How well a regression model can predict the value of the response variable in absolute terms.
- **MAE** (Mean absolute error) represents the difference between the original and predicted values extracted by averaged the absolute difference over the data set. How well a model can predict the value of the response variable in percentage terms.
- **R-squared** (Coefficient of determination) represents the coefficient of how well the values fit compared to the original values. The value from 0 to 1 interpreted as percentages. The higher the value is, the better the model is.
- **Adj R-squared** penalize **R-squared** for small coefficients in new features.
- **MSE** (Mean Squared Error) represents the difference between the original and predicted values extracted by squared the average difference over the data set.
- **MAPE** : (Mean Absolute Percentage Error) measures the accuracy of predictions as a percentage.

Modeling without description

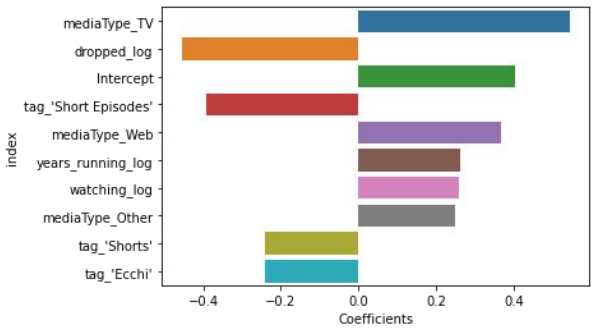
Linear regression Model

Training Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.43	0.33	0.70	0.70	12.20

Test Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.43	0.33	0.68	0.68	12.11



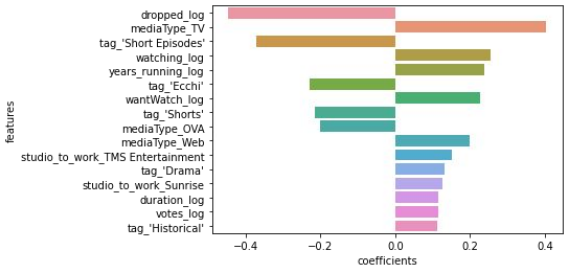
Lasso Model

Training Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.43	0.33	0.70	0.69	12.30

Test Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.43	0.33	0.68	0.68	12.14



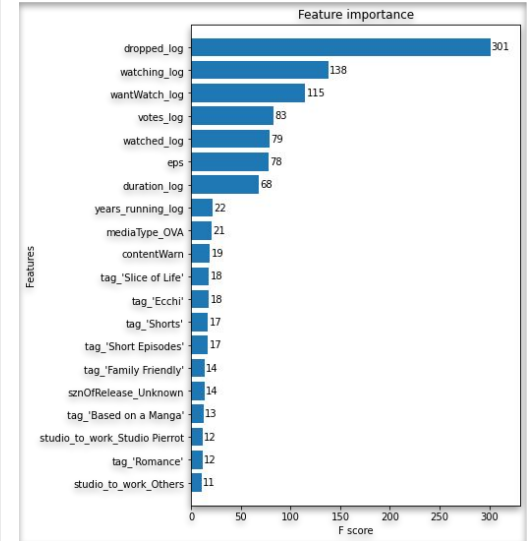
	features	coefficients	coefficients_abs
6	tag_'Fantasy'	-0.00	0.00
9	tag_'Family Friendly'	-0.00	0.00
29	tag_'Promotional'	-0.00	0.00
32	tag_'Vocaloid'	0.00	0.00
41	mediaType_Movie	0.00	0.00
51	sznOfRelease_Winter	-0.00	0.00
55	studio_to_work_Production I.G	0.00	0.00
56	studio_to_work_Studio Deen	-0.00	0.00
60	studio_to_work_Toei Animation	-0.00	0.00

XGBRegressor Model

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.38	0.29	0.77	0.76	10.76

Test Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.42	0.32	0.70	0.69	11.92



Modeling with description

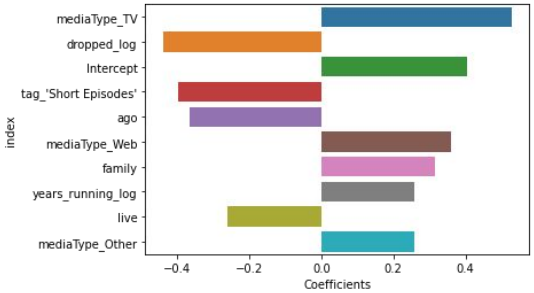
Linear regression Model

Training Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.42	0.33	0.71	0.70	11.93

Test Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.43	0.33	0.68	0.65	12.11



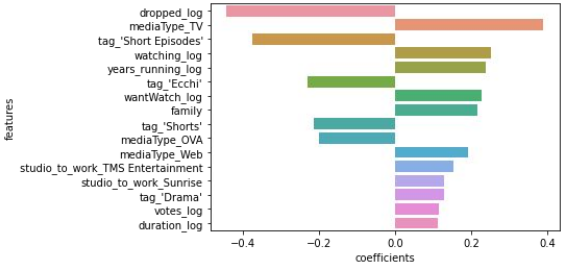
Lasso Model

Training Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.43	0.33	0.70	0.69	12.24

Test Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.43	0.33	0.69	0.65	12.10



	features	coefficients	coefficients_abs
0	able	0.00	0.00
1	academy	-0.00	0.00
2	ago	-0.00	0.00
3	along	-0.00	0.00
4	also	0.00	0.00
...	...	...	...
216	sznOfRelease_Winter	-0.00	0.00
219	studio_to_work_Others	0.00	0.00
220	studio_to_work_Production I.G	0.00	0.00
221	studio_to_work_Studio Deen	-0.00	0.00
225	studio_to_work_Toei Animation	-0.00	0.00

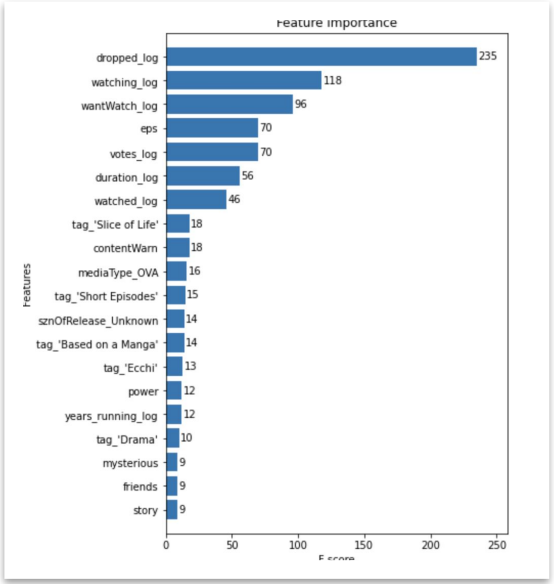
167 rows × 3 columns

XGBRegressor Model

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.37	0.29	0.77	0.76	10.61

Test Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.42	0.33	0.70	0.66	12.02

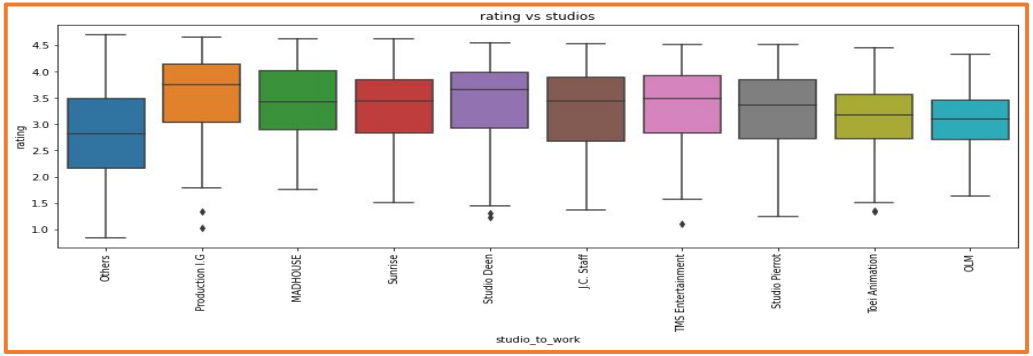


Summary & observations

Top rated

	title	mediaType	eps	duration	ongoing	startYr	finishYr	sznOfRelease	description	studios	.
0	Fullmetal Alchemist: Brotherhood	TV	64.00	NaN	False	2009.00	2010.00	Spring	The foundation of alchemy is based on the law ...	['Bones']	
1	your name.	Movie	1.00	1hr 47min	False	2016.00	2016.00	NaN	Mitsuha and Taki are two total strangers livin...	['CoMix Wave Films']	
2	A Silent Voice	Movie	1.00	2hr 10min	False	2016.00	2016.00	NaN	After transferring into a new school, a deaf g...	['Kyoto Animation']	
3	Haikyuu!! Karasuno High School vs Shiratorizaw...	TV	10.00	NaN	False	2016.00	2016.00	Fall	Picking up where the second season ended, the ...	['Production I.G']	
4	Attack on Titan 3rd Season: Part II	TV	10.00	NaN	False	2019.00	2019.00	Spring	The battle to retake Wall Maria begins now! Wi...	['Wit Studio']	
5	Demon Slayer: Kimetsu no	TV	26.00	NaN	False	2019.00	2019.00	Spring	Bloodthirsty demons lurk in the woods,	['ufotable']	

Studios that influence in the rating

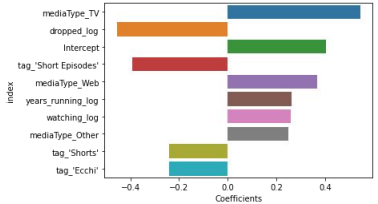


No statistical difference between them - overall very similar

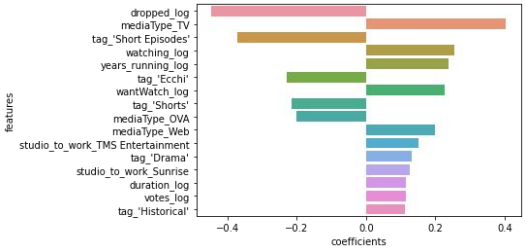
Summary & observations

The most important features - No description

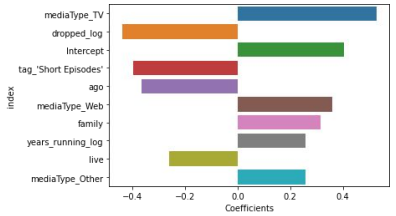
Linear regression Model

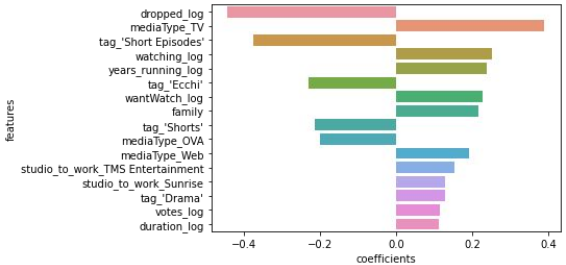


Lasso Model

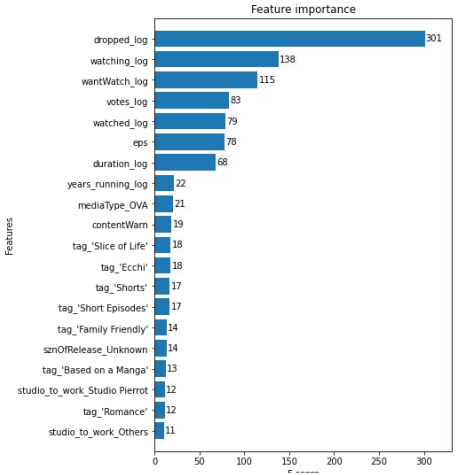


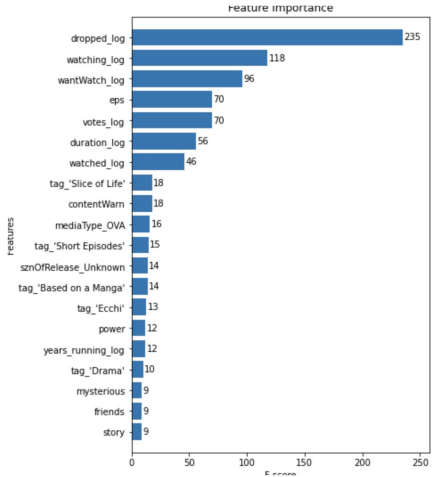
The most important features - w/ description





XGBRegressor Model

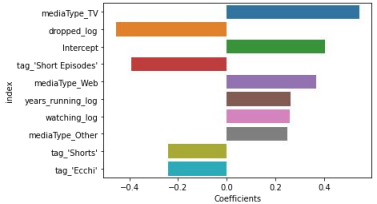




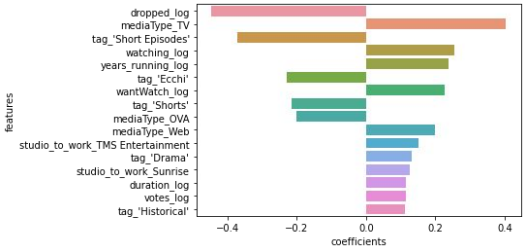
Summary & observations

The most important features - No description

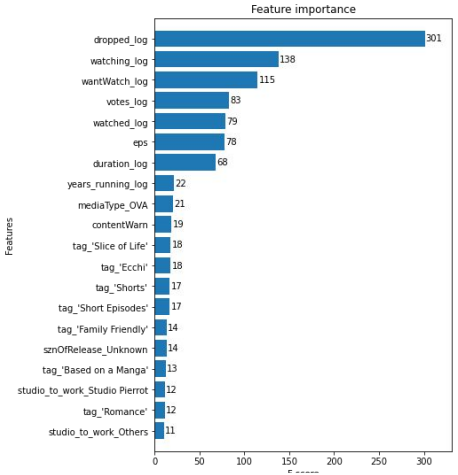
Linear regression Model



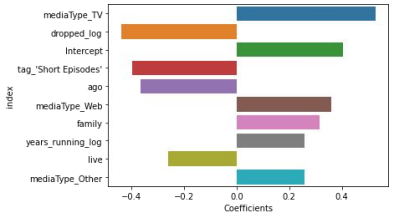
Lasso Model

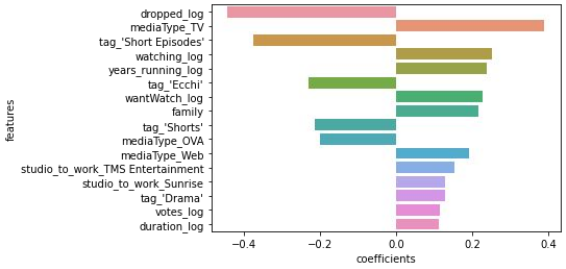


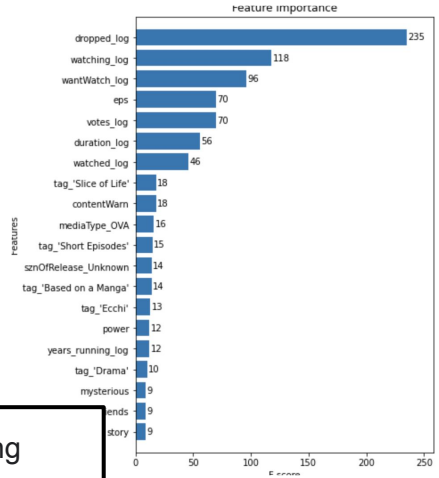
XGBRegressor Model



The most important features - w/ description







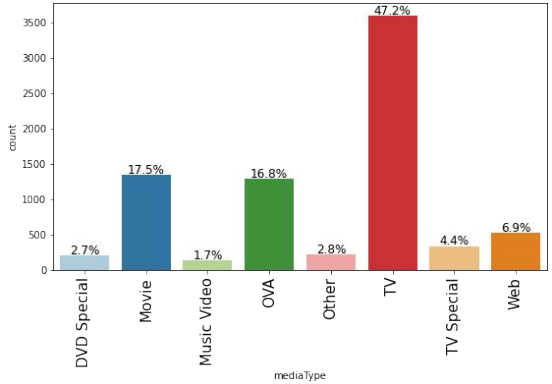
Description is not important in the rating prediction - Challenge - One hot encoding

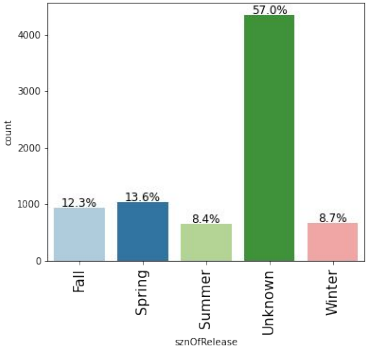
Thank you!

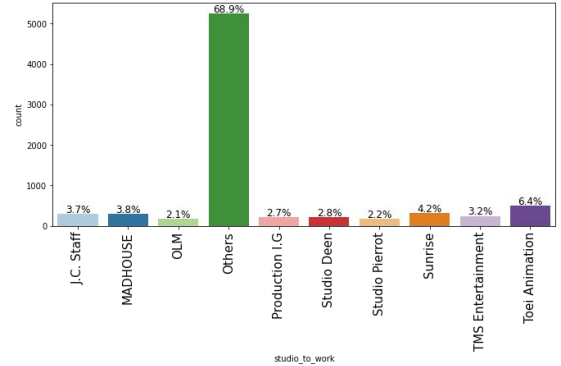
Questions?

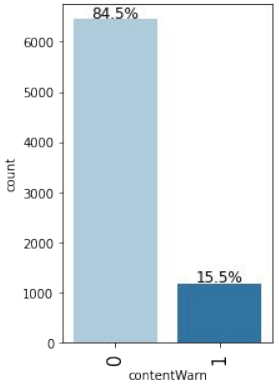
backup

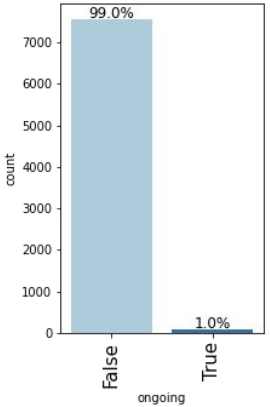
Univarate analysis





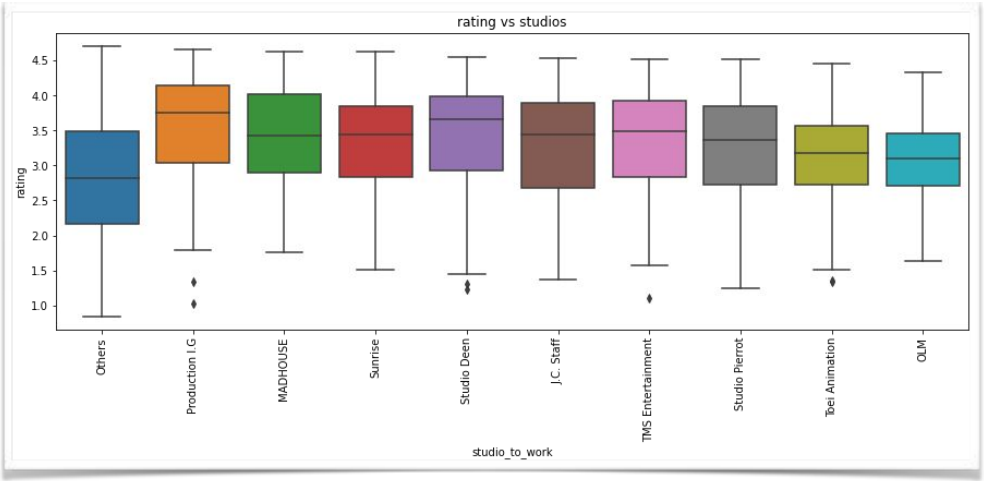




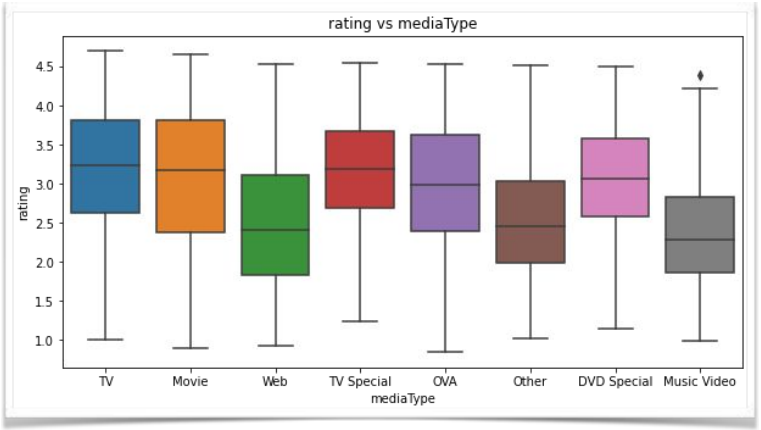


Bivariate analysis

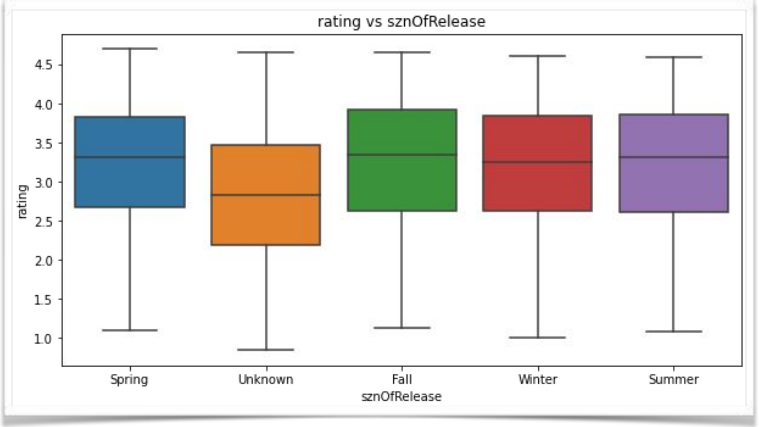
Analysis of two variables: cross-correlation relationship



Production I.G. is the studio with the highest rating



Web has the lowest rating following by Music video



There is not much difference between release seasons