WEEK3
LINEAR REGRESSION

6438169421 Pattaradanai Lakkananithiphan

CONTENT

I: Loading & Overview

II: Preprocessing

III: Model Evaluation

IV: Intrepretation

I: LOADING

5043 Rows

Loaded from the CSV file

27 Columns

16 numerical + 11 categorical

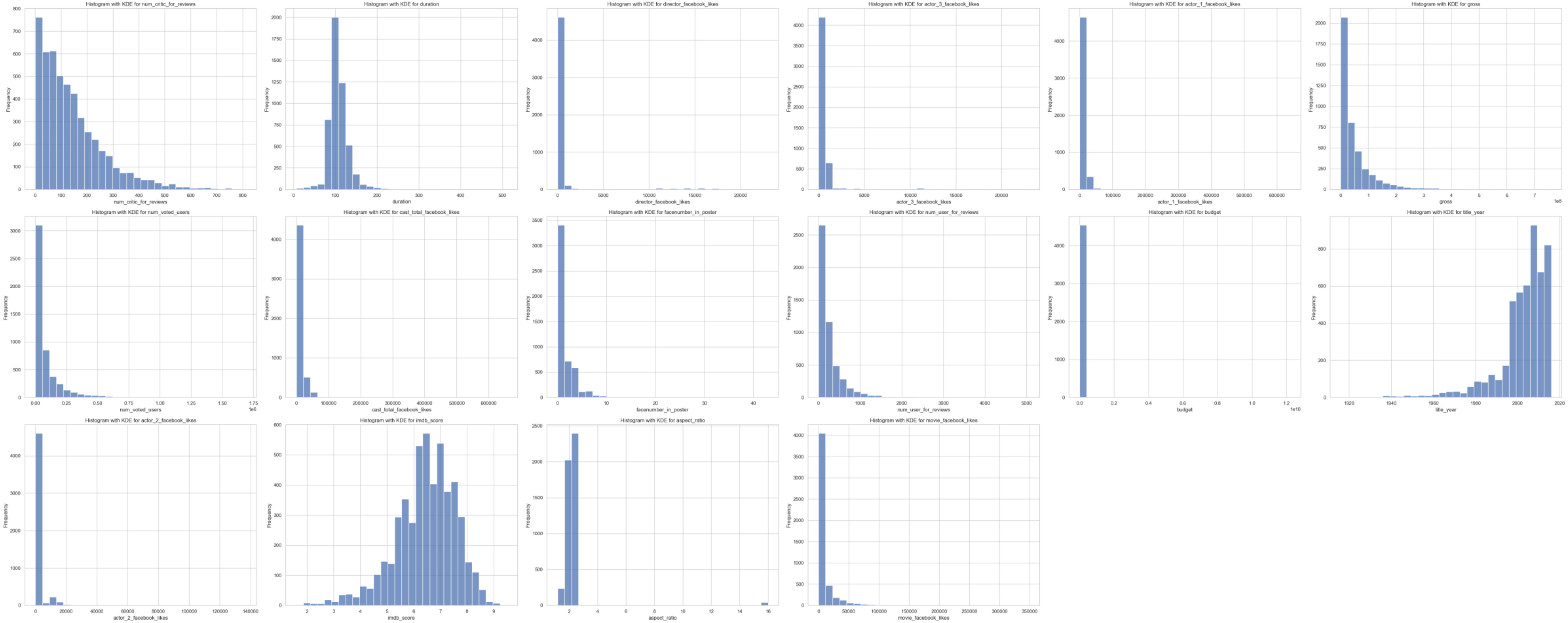
>884 rows have missing values

50 rows have missing label

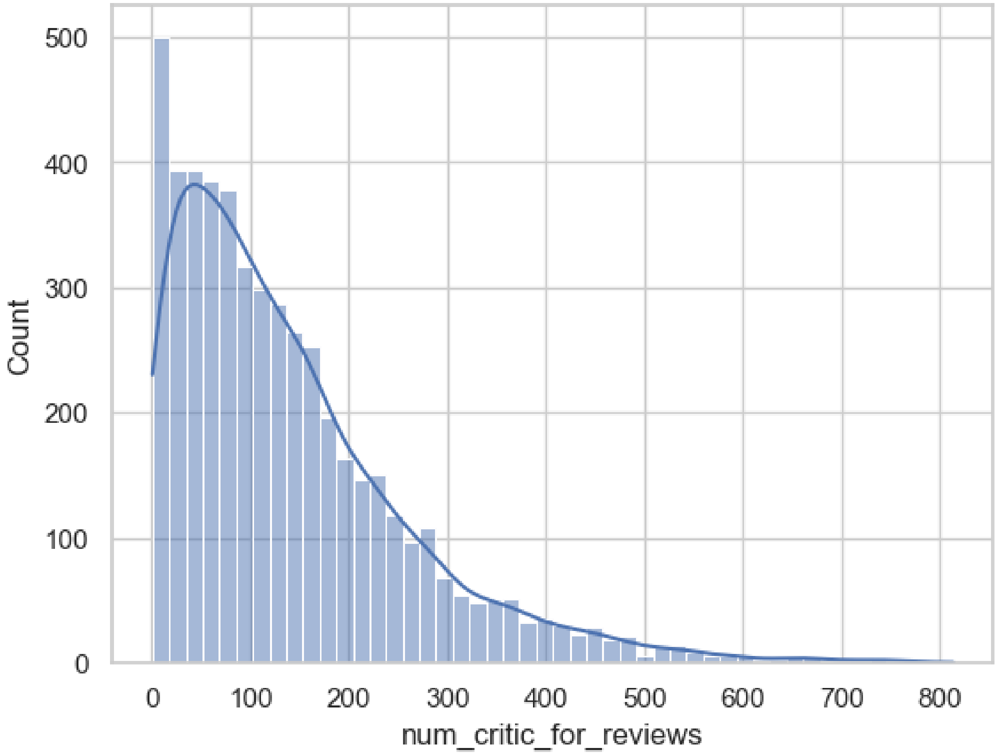
Numerical Label

The label is a numerical value of type float

THE NUMERICAL DATA DISTRIBUTIONS



THE DISTRIBUTION OF THE LABEL



The data is right-skewed

CATEGORY COUNTS

CATEGORICAL COLUMN NAMES	NUMBER OF CATEGORIES
director_name	2398
actor_2_name	3032
genres	914
actor_1_name	2097
movie_title	4917
actor_3_name	3521
plot_keywords	4760
movie_imdb_link	4919
language	47
country	65
content_rating	18

II: PREPROCESS

Dropping columns

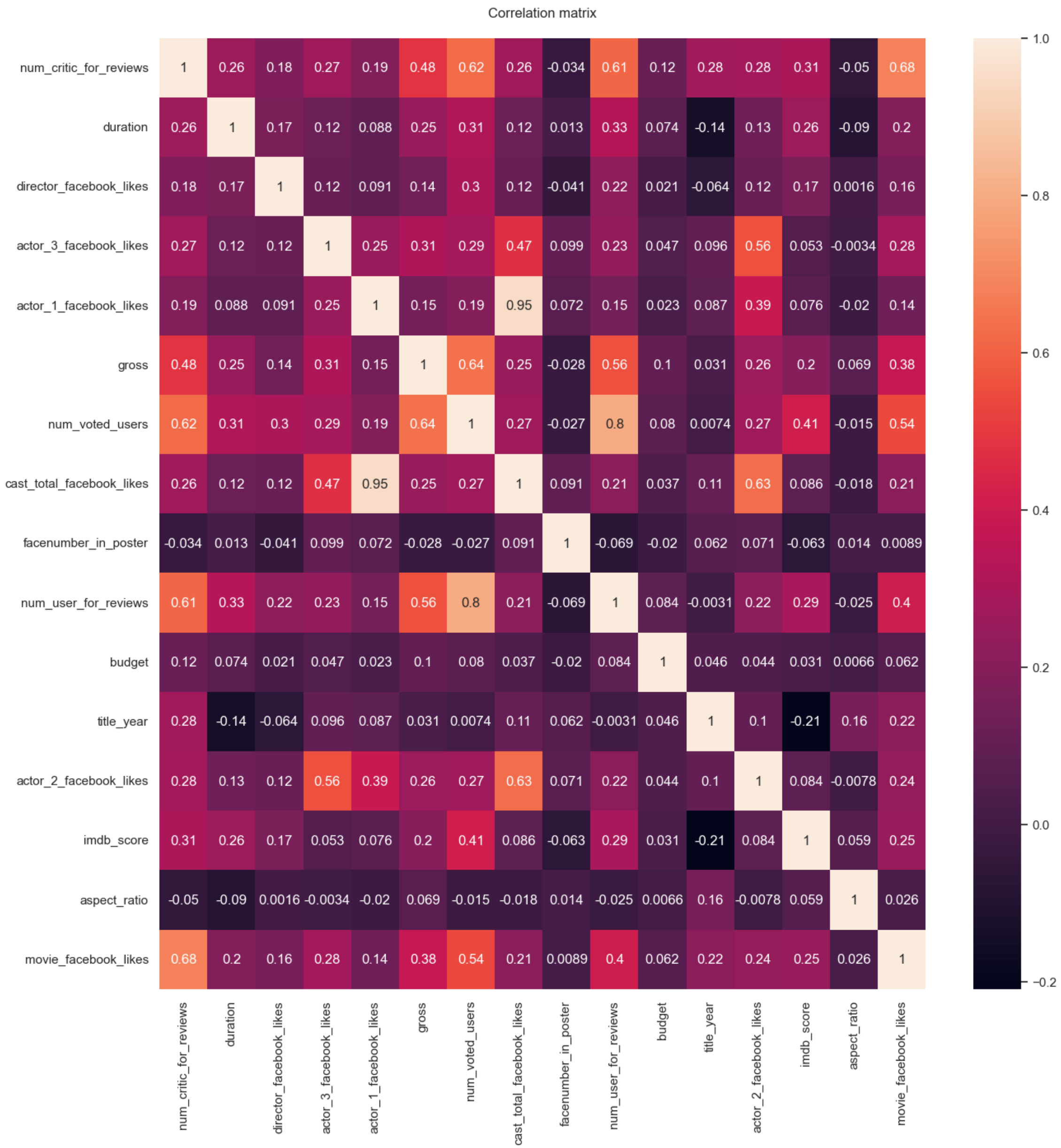
Dropping/Filling rows

Scaling

Splitting

DROPPING COLUMNS: CATEGORICAL

- Most if not all categorical columns have a lot more categories than appropriate
- Many are highly unrelated to the label we want to predict
- Some are very biased toward certain categories
- Many have a lot of NaN rows
- As guided by the solution slide
- **CONCLUSION: DROP THE CATEGORICAL COLUMNS**

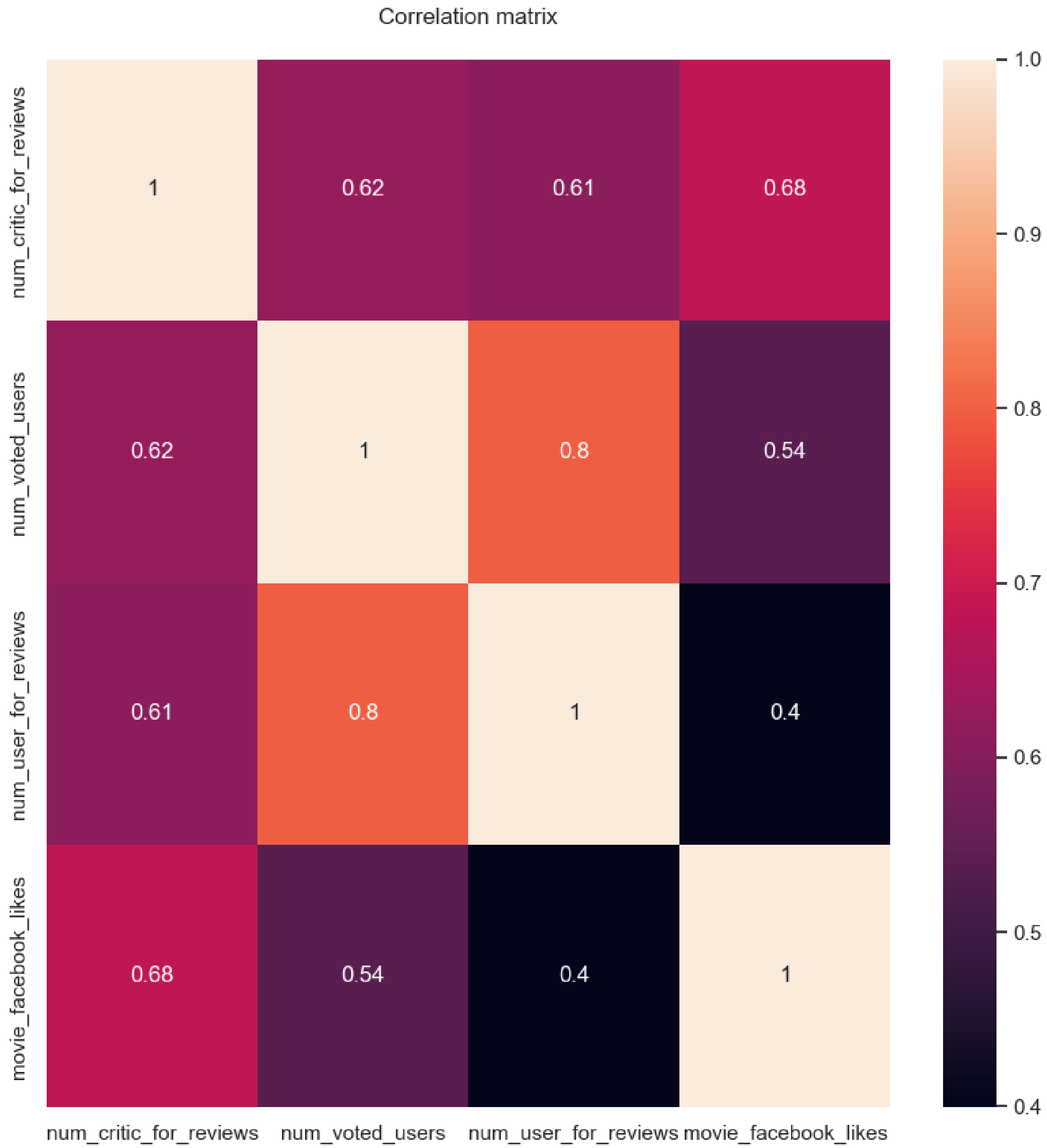


	num_critic_for_reviews	duration	director_facebook_likes	actor_3_facebook_likes	actor_1_facebook_likes	gross	num_voted_users	cast_total_facebook_likes	facenumber_in_poster	num_user_for_reviews	budget	title_year	actor_2_facebook_likes	imdb_score	aspect_ratio	movie_facebook_likes
num_critic_for_reviews	1	0.26	0.18	0.27	0.19	0.48	0.62	0.26	-0.034	0.61	0.12	0.28	0.28	0.31	-0.05	0.68
duration	0.26	1	0.17	0.12	0.088	0.25	0.31	0.12	0.013	0.33	0.074	-0.14	0.13	0.26	-0.09	0.2
director_facebook_likes	0.18	0.17	1	0.12	0.091	0.14	0.3	0.12	-0.041	0.22	0.021	-0.064	0.12	0.17	0.0016	0.16
actor_3_facebook_likes	0.27	0.12	0.12	1	0.25	0.31	0.29	0.47	0.099	0.23	0.047	0.096	0.56	0.053	-0.0034	0.28
actor_1_facebook_likes	0.19	0.088	0.091	0.25	1	0.15	0.19	0.95	0.072	0.15	0.023	0.087	0.39	0.076	-0.02	0.14
gross	0.48	0.25	0.14	0.31	0.15	1	0.64	0.25	-0.028	0.56	0.1	0.031	0.26	0.2	0.069	0.38
num_voted_users	0.62	0.31	0.3	0.29	0.19	0.64	1	0.27	-0.027	0.8	0.08	0.0074	0.27	0.41	-0.015	0.54
cast_total_facebook_likes	0.26	0.12	0.12	0.47	0.95	0.25	0.27	1	0.091	0.21	0.037	0.11	0.63	0.086	-0.018	0.21
facenumber_in_poster	-0.034	0.013	-0.041	0.099	0.072	-0.028	-0.027	0.091	1	-0.069	-0.02	0.062	0.071	-0.063	0.014	0.0089
num_user_for_reviews	0.61	0.33	0.22	0.23	0.15	0.56	0.8	0.21	-0.069	1	0.084	-0.0031	0.22	0.29	-0.025	0.4
budget	0.12	0.074	0.021	0.047	0.023	0.1	0.08	0.037	-0.02	0.084	1	0.046	0.044	0.031	0.0066	0.062
title_year	0.28	-0.14	-0.064	0.096	0.087	0.031	0.0074	0.11	0.062	-0.0031	0.046	1	0.1	-0.21	0.16	0.22
actor_2_facebook_likes	0.28	0.13	0.12	0.56	0.39	0.26	0.27	0.63	0.071	0.22	0.044	0.1	1	0.084	-0.0078	0.24
imdb_score	0.31	0.26	0.17	0.053	0.076	0.2	0.41	0.086	-0.063	0.29	0.031	-0.21	0.084	1	0.059	0.25
aspect_ratio	-0.05	-0.09	0.0016	-0.0034	-0.02	0.069	-0.015	-0.018	0.014	-0.025	0.0066	0.16	-0.0078	0.059	1	0.026
movie_facebook_likes	0.68	0.2	0.16	0.28	0.14	0.38	0.54	0.21	0.0089	0.4	0.062	0.22	0.24	0.25	0.026	1

DROPPING COLUMNS: NUMERICAL

<- CORRELATION MATRIX

DROP ALL COLUMNS OF ABSOLUTE CORRELATION LESS THAN **0.5**
AS GUIDED BY THE SLIDE



DROPPING COLUMNS: NUMERICAL

<- CORRELATION MATRIX

THE RESULT

__NOTE**:__ THE **NUM_USER_FOR_REVIEWS** AND **NUM_VOTED_USERS** COLUMNS ARE HIGHLY CORRELATED AND WE COULD DROP THE LOWER CORRELATED COLUMN IE **NUM_USER_FOR_REVIEWS**

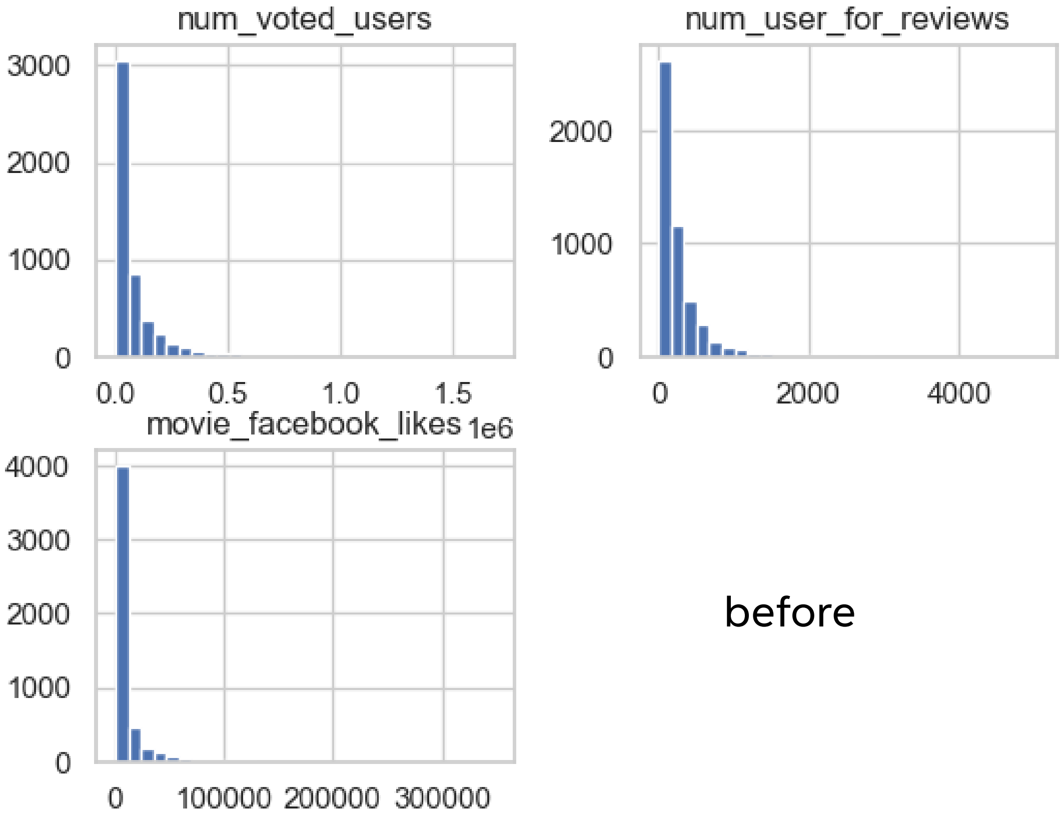
DROP/FILL THE EMPTY ROWS

COLUMN NAME	NUMBER OF EMPTY CELLS
num_critic_for_reviews	50
num_voted_users	0
num_user_for_reviews	21
movie_facebook_likes	0

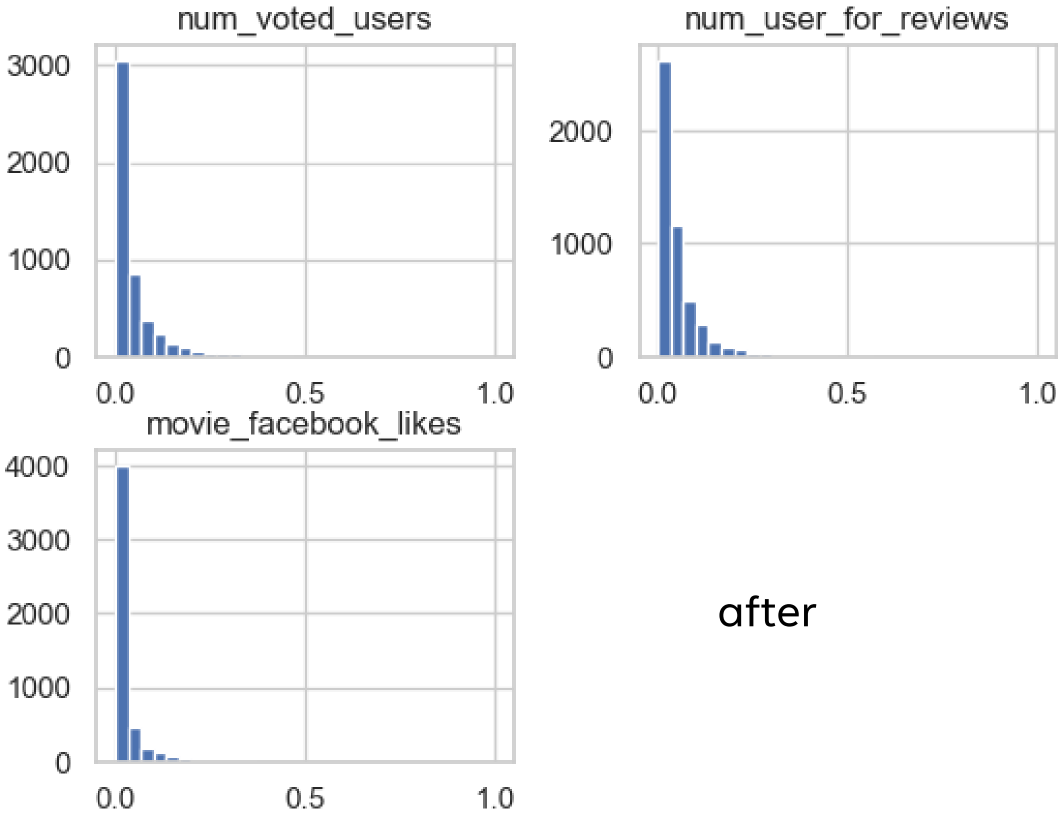
SINCE THERE ARE ONLY A FEW ROWS WITH PROBLEMS
I CHOSE TO DROP THEM

#ROWS: 5042 -> 4993 (0.97% LOST)

SCALE THE FEATURES: MINMAX SCALER (0~1)



before



after

SPLIT THE DATA INTO TRAINING AN D TESTING SET

70% TRAINING

30% TESTING

SHUFFLE & RANDOMIZED

III: MODELS

SGDR model

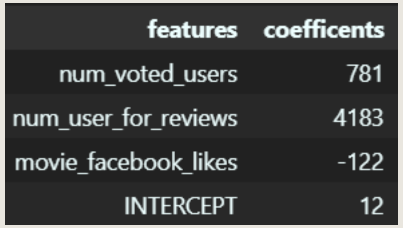
Linear regression model

SGDR model**

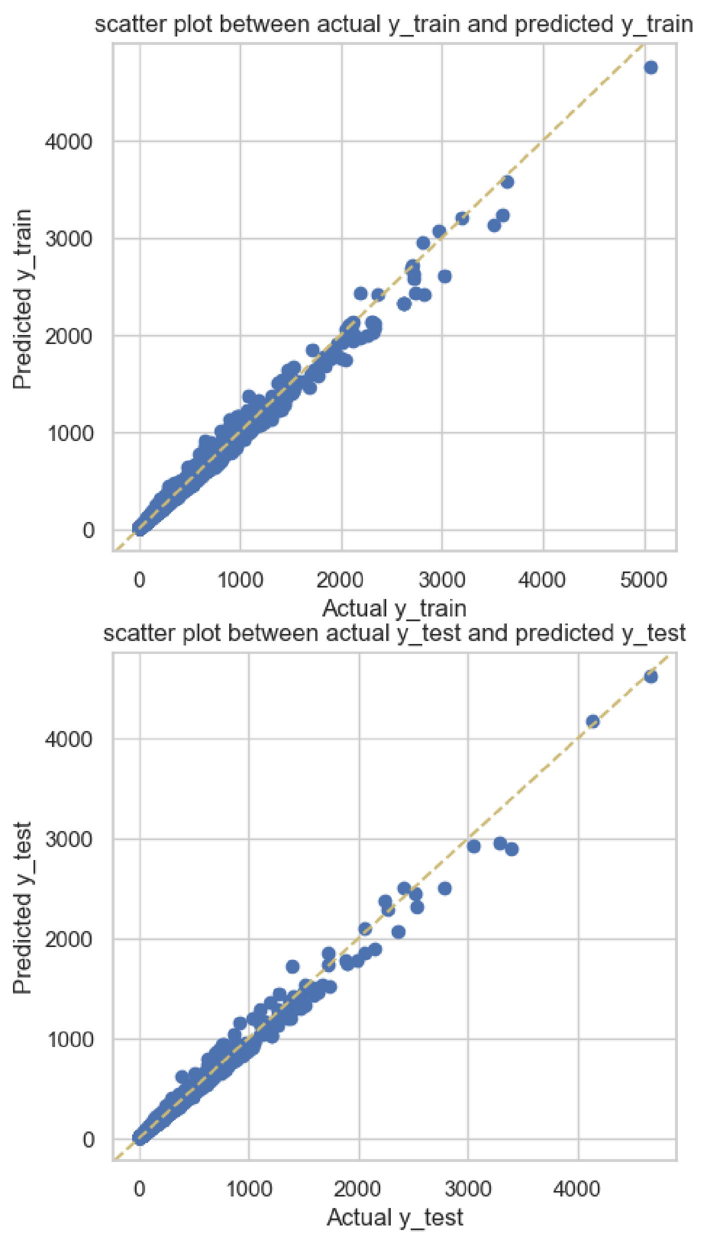
**only two feature columns

SGDREGRESSOR MODEL

- MAE : 21.66
- MSE : 1842.16
- RMSE : 42.92
- MAPE : 0.2946
- R2 : 0.9883
- Adjusted R2 : 0.9882



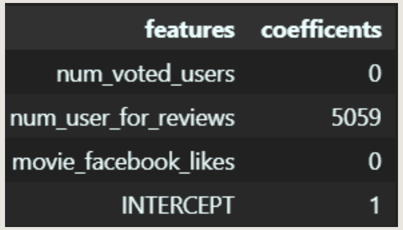
features	coefficents
num_voted_users	781
num_user_for_reviews	4183
movie_facebook_likes	-122
INTERCEPT	12



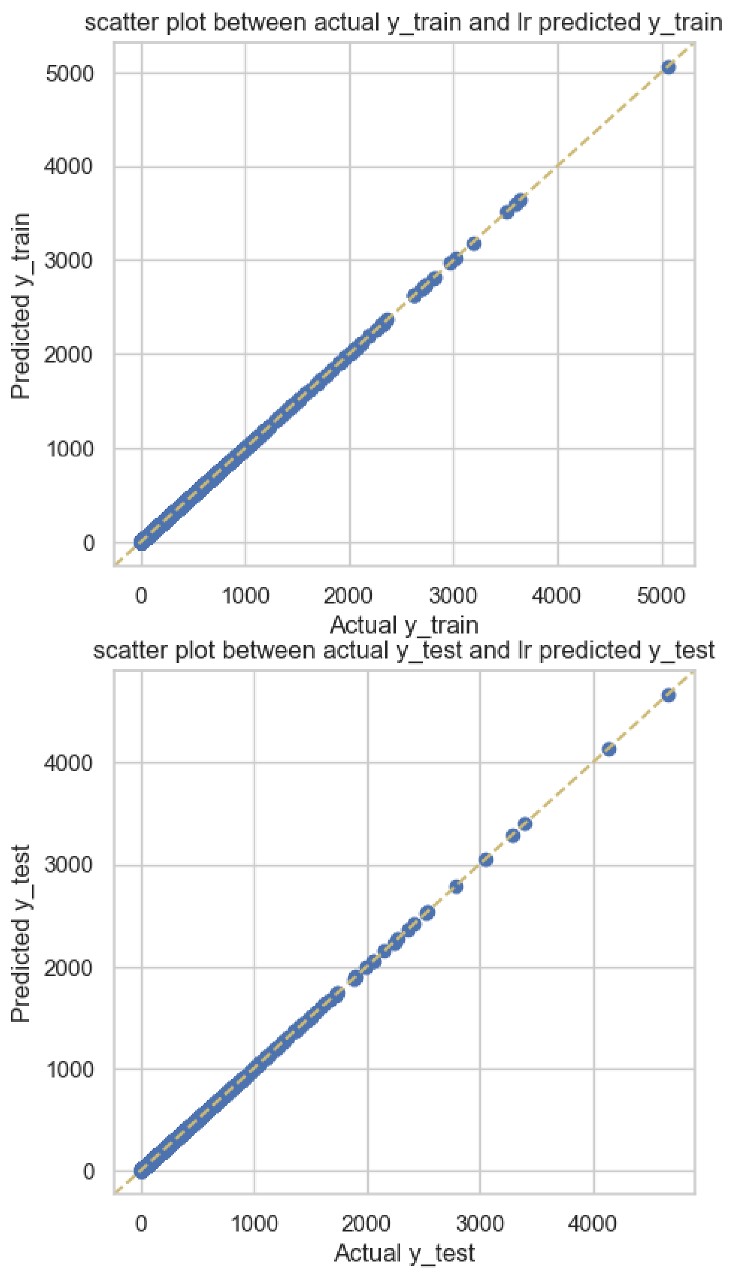
LINEAR REGRESSION MODEL

- MAE : 0
- MSE : 0
- RMSE : 0
- MAPE : 0
- R2 : 1
- Adjusted R2 : 1

Overfitting !



features	coefficents
num_voted_users	0
num_user_for_reviews	5059
movie_facebook_likes	0
INTERCEPT	1



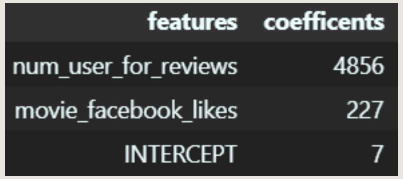
scatter plot between actual y_train and lr predicted y_train

scatter plot between actual y_test and lr predicted y_test

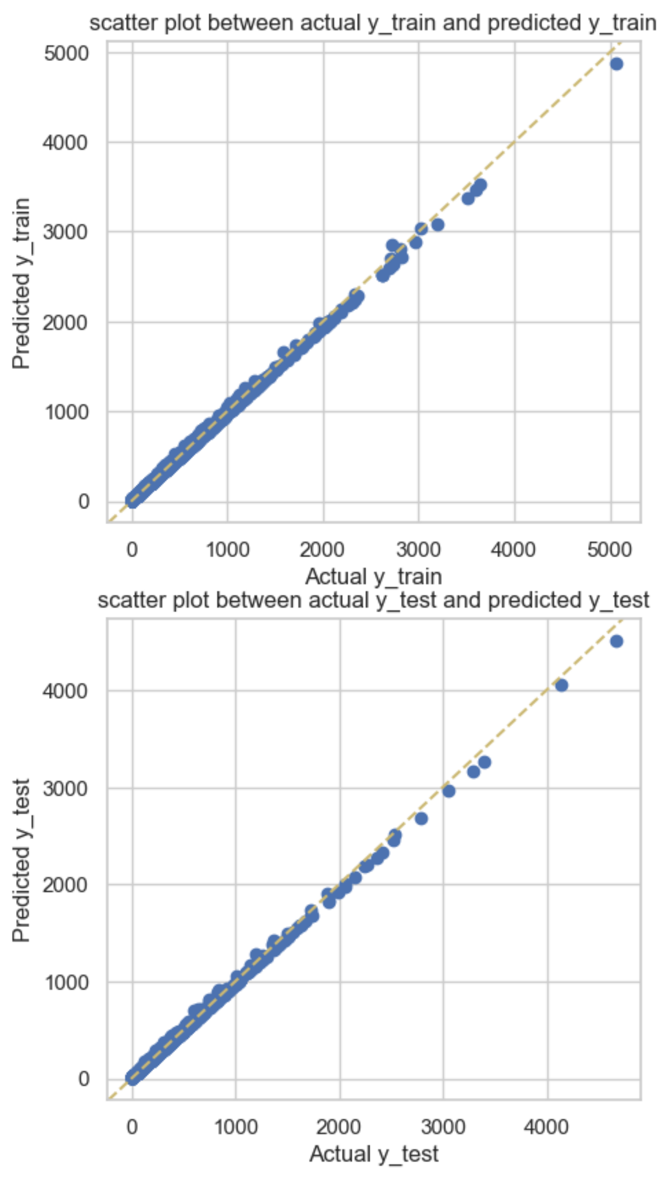
SGDREGRESSOR MODEL (ONLY 2 COLUMNS**)

- MAE : 8.926
- MSE : 265.33
- RMSE : 16.29
- MAPE : 0.1633
- R2 : 0.9983
- Adjusted R2 : 0.9983

Better results
But
Possible Overfitting



features	coefficents
num_user_for_reviews	4856
movie_facebook_likes	227
INTERCEPT	7



INTERPRETATION

- The 1st and 3rd models are candidates for practical application
  - The 1st model is less overfitted
  - The 3rd model performs better
- For every model, the features’ importance is ranked:
  1. Number of users for reviews
  2. Movie Facebook likes
  3. Number of voted users