Linear Regression Model for Critical Reviews

Puvit Pracharktam 6031830321

PROCEDURE

- Select all numerical features
- Cleaning data by **drop >3 NaN** in each column and **replace median**
- Scaling to log scale by both Min-Max normalisation and standardisation method
- Correlation cutout at **0.6**
- Divide **30%** of data for testing

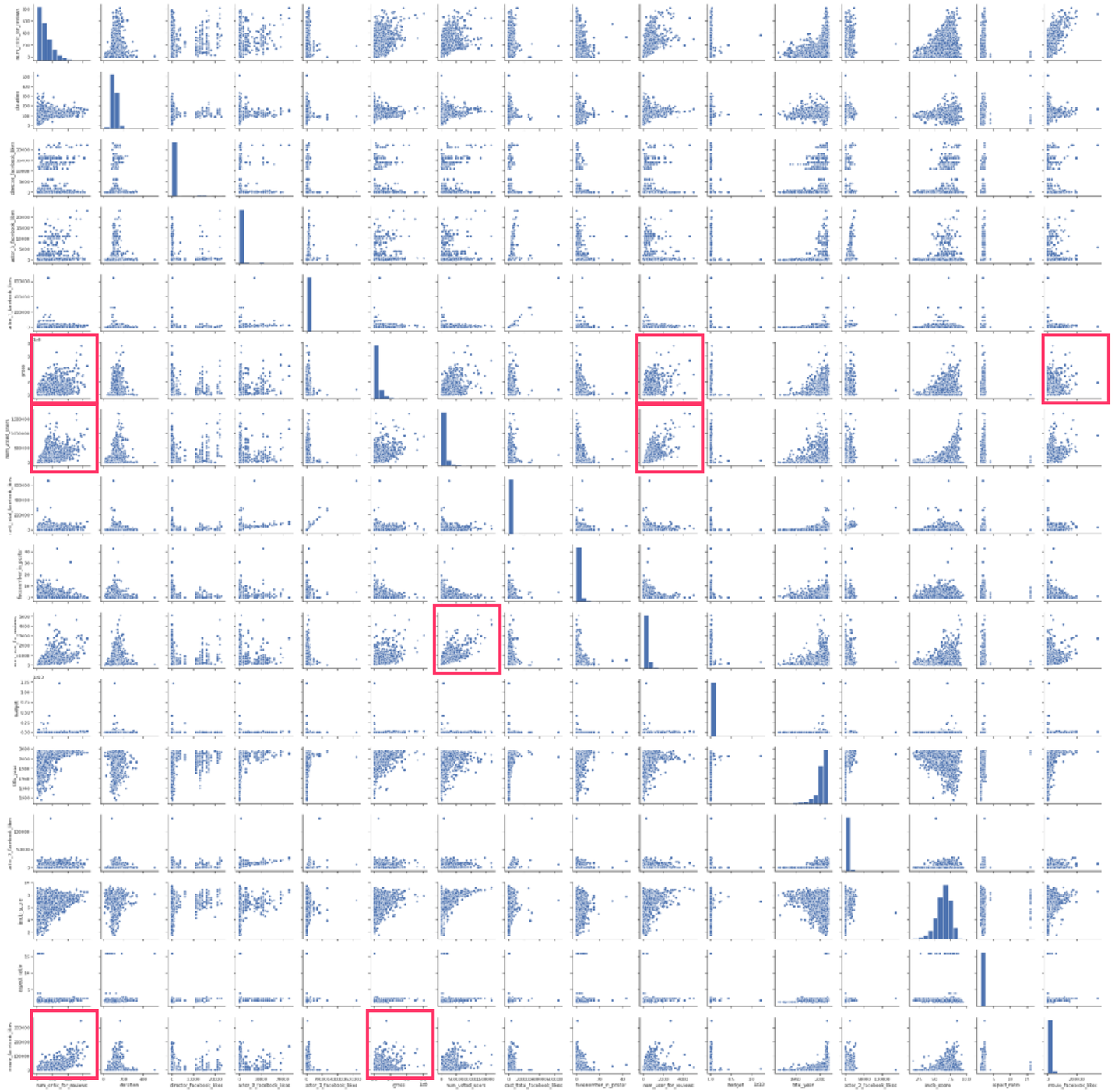
Scatter plot between Number of Critical Reviews

gross

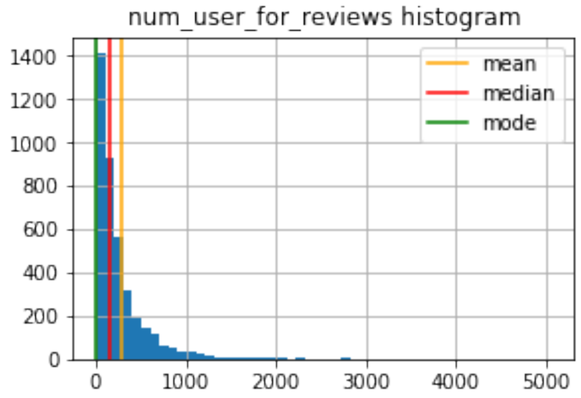
num_voted_users

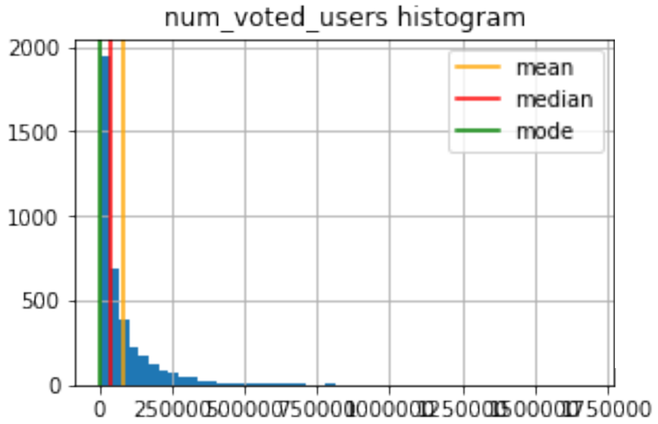
num_users_for_reviews

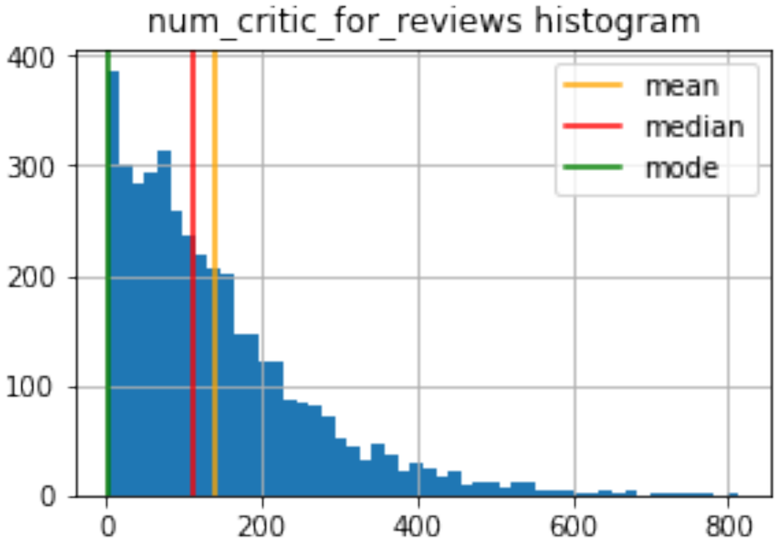
movie_facebook_likes

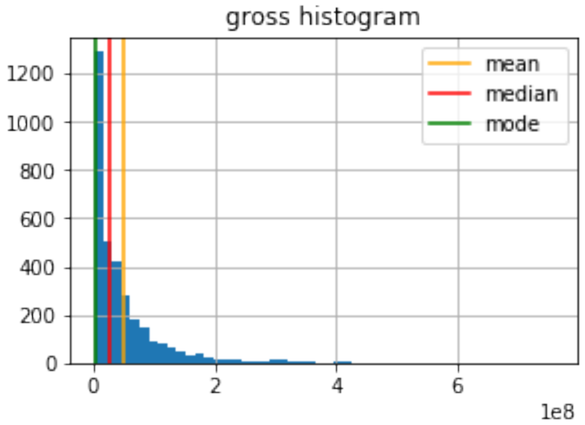


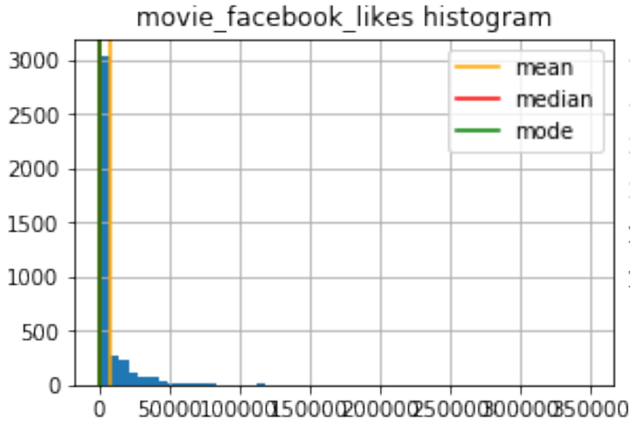
There are **4** variables show possible linear correlation



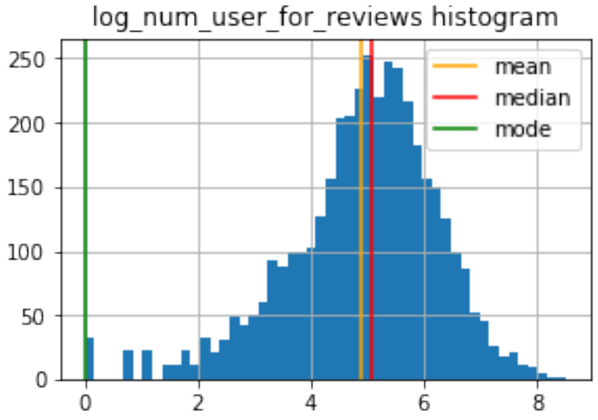


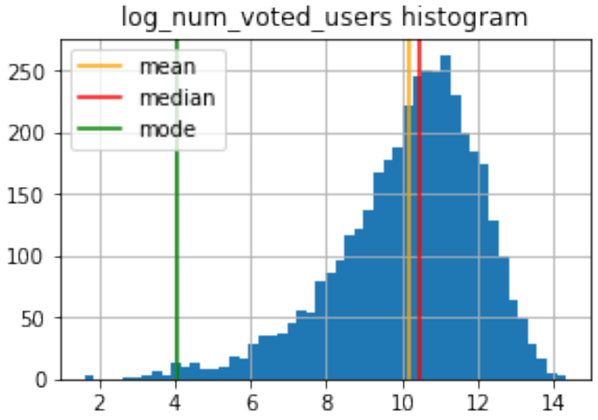


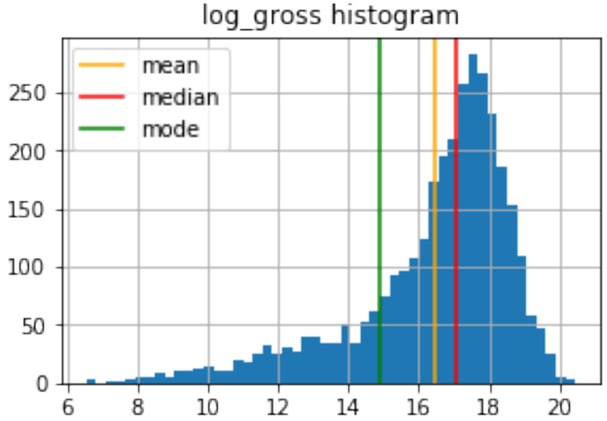


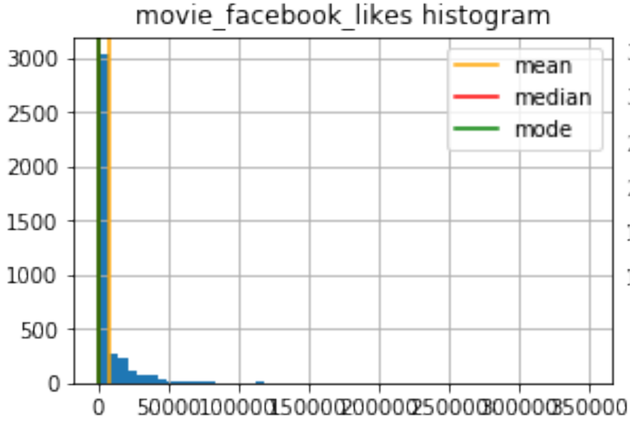


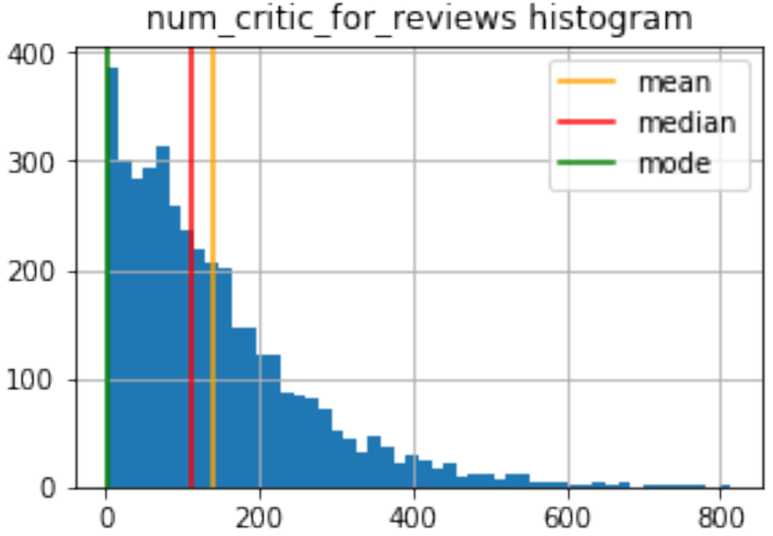
Before taking log scaling











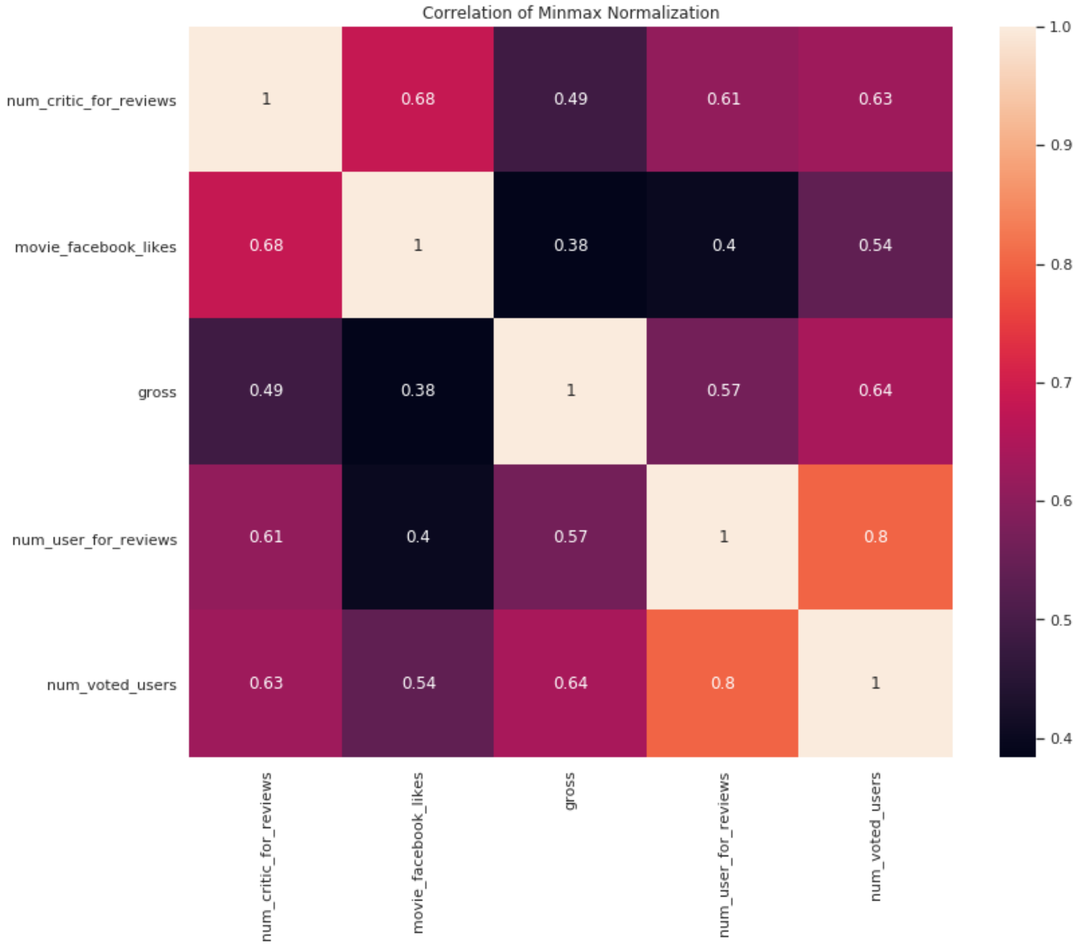
After taking log scaling

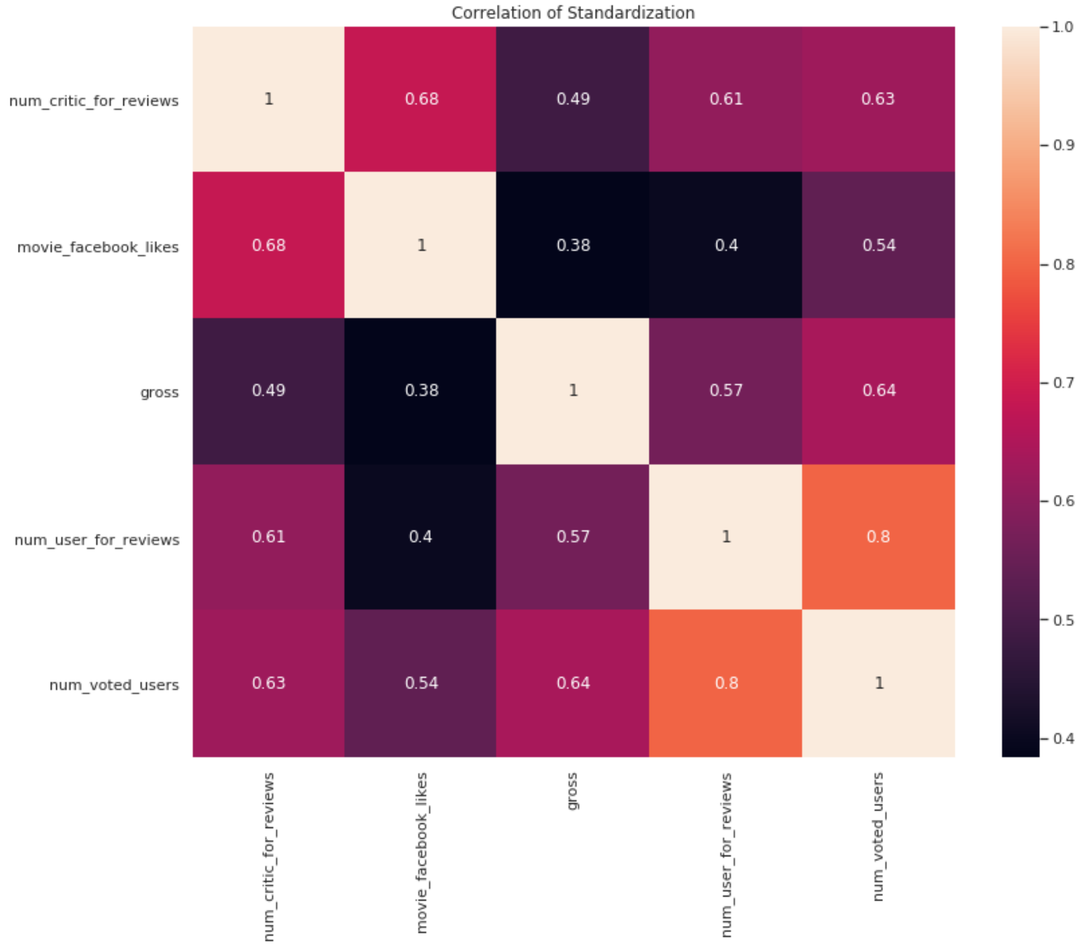
Judging from the feature's statistic, it is safe to **fill all NaN with median**.

I cannot taking log scale in **num_critic_for_reviews** and **movies_facebook_likes** because there are some value might be **-infinity.**

CORRELATION ANALYSIS

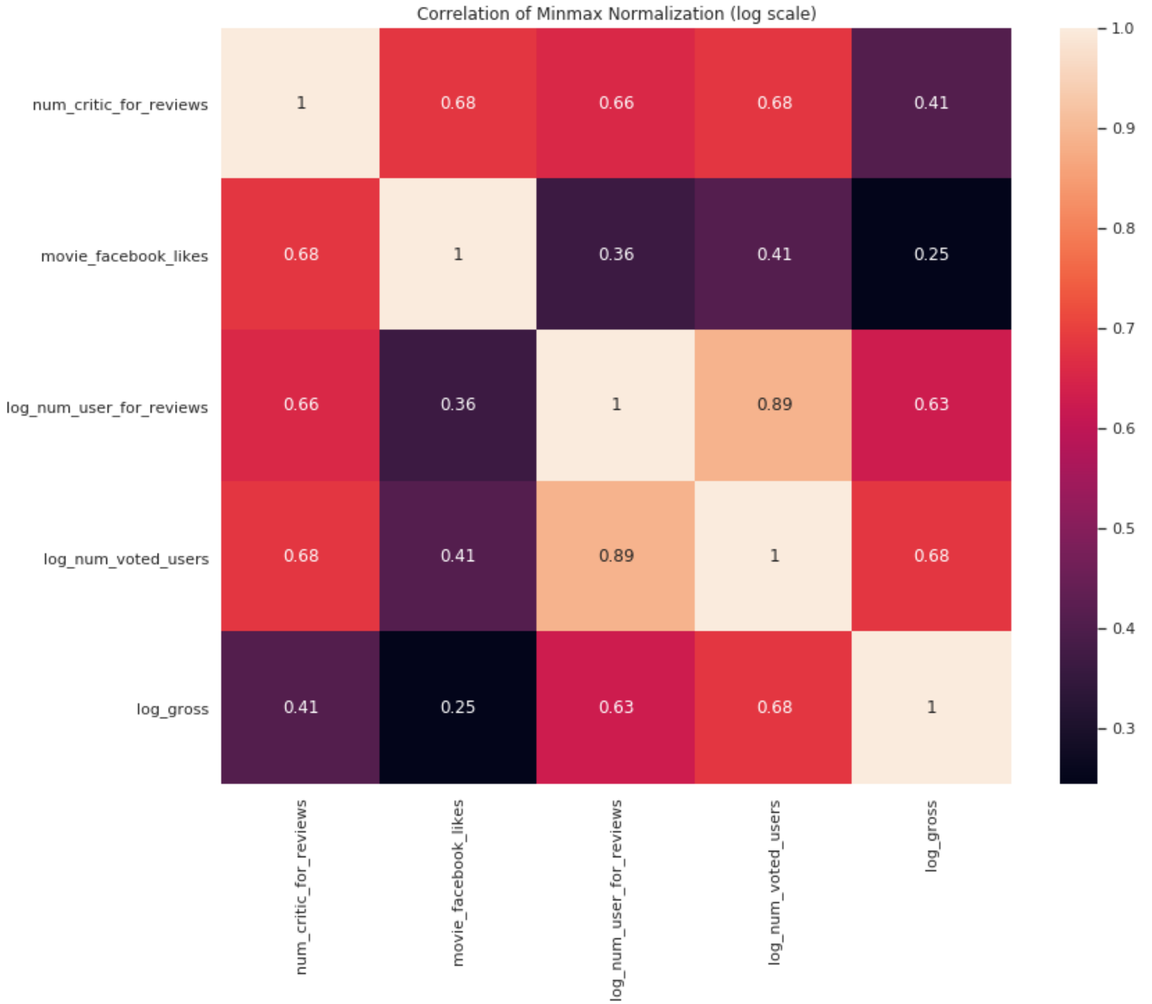
Not scaled

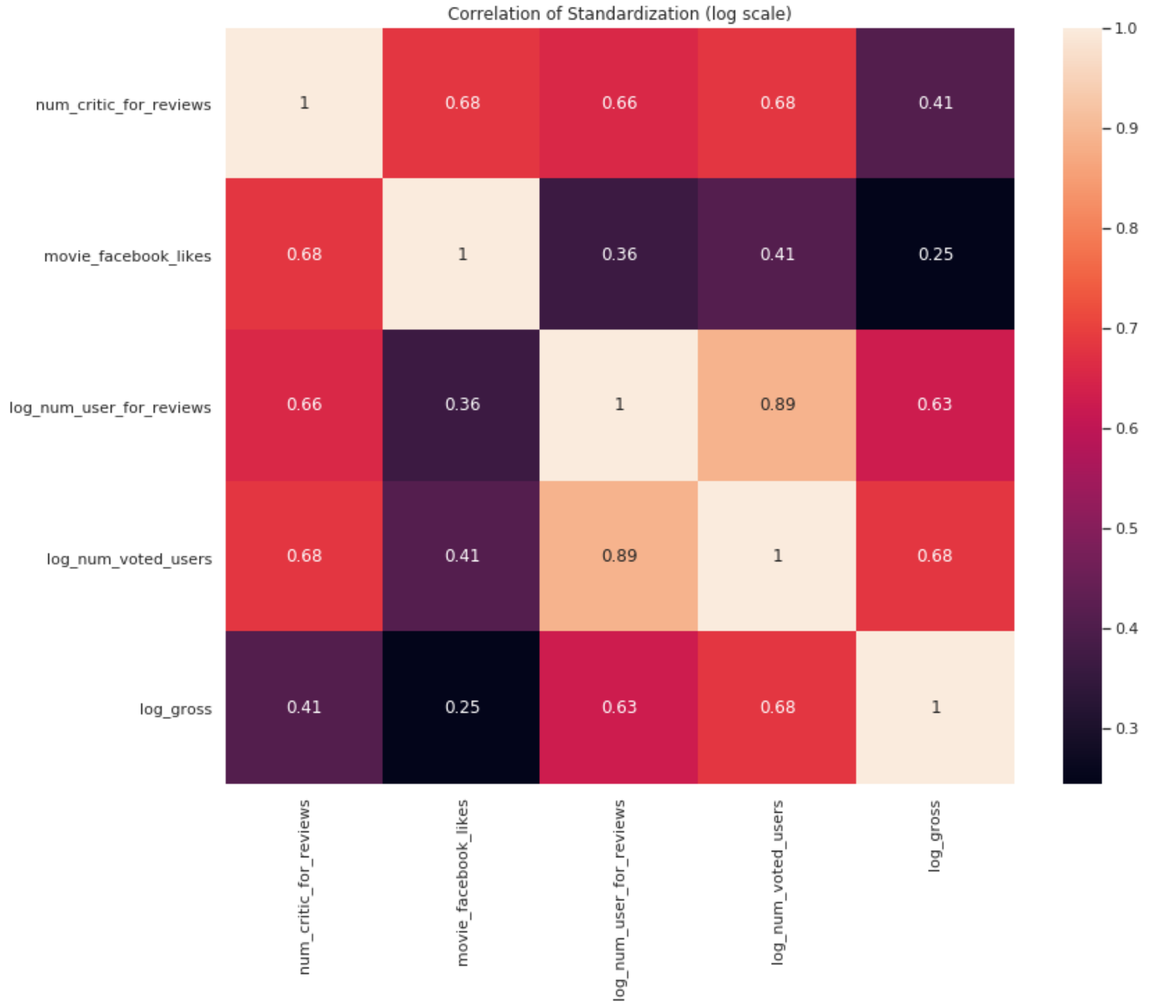




CORRELATION ANALYSIS

Scaled by log





From comparison, show that we should use log scale before training

Selected features correlated with **num_critic_for_reviews >0.6**

log_num_user_for_reviews	0.656187
log_num_voted_users	0.682921
movie_facebook_likes	0.683318

	log_num_user_for_reviews	movie_facebook_likes	log_num_voted_users
log_num_user_for_reviews	1.000000	0.363553	0.888058
movie_facebook_likes	0.363553	1.000000	0.411220
log_num_voted_users	0.888058	0.411220	1.000000

gross feature is eliminated since correlation <0.6

Linear Regression Model Results (Normalized)

With num_voted_users, movie_facebook_likes (cutout at 0.6):

num_critic_for_reviews = 0.5134 * num_voted_users
+ 1.4913 * movie_facebook_likes
+ 0.0957
(R^2 = 0.7828)
(RMSE : 122.14654)