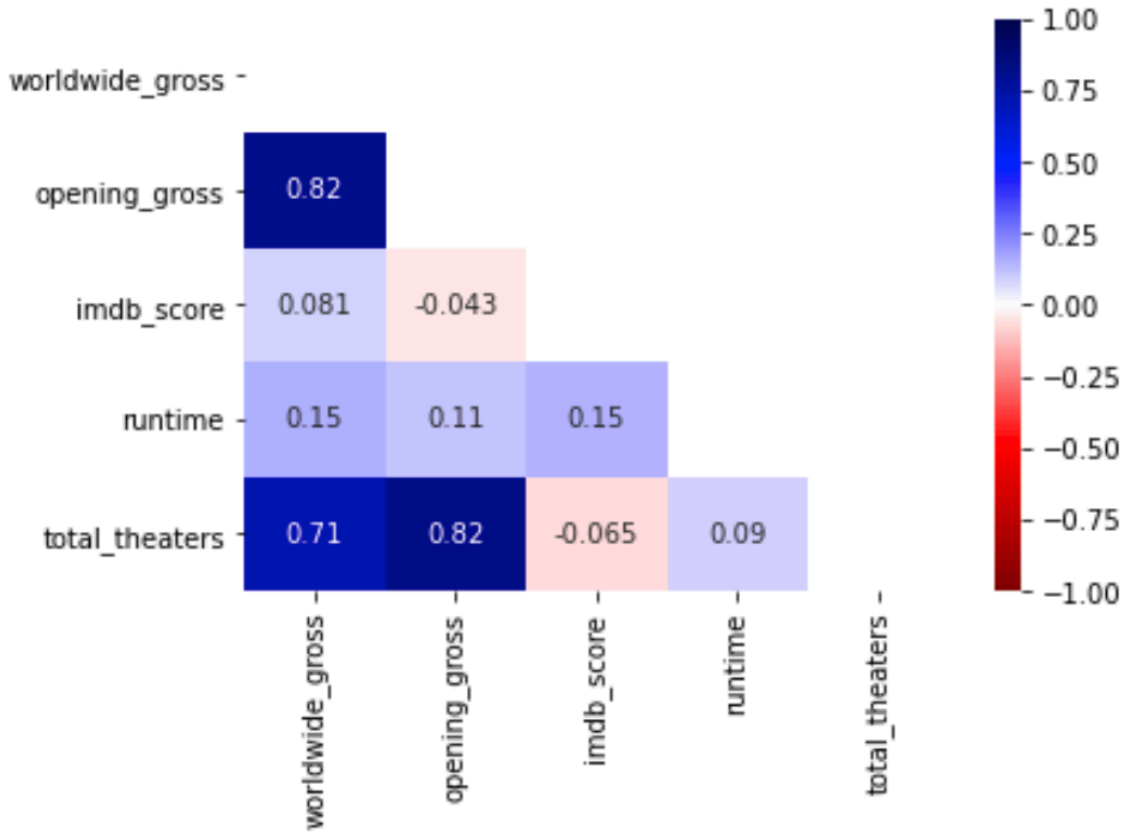
Box-Office-Mojo Project: Linear Regression

OBJECTIVE: An initial proposal was made to aid Hollywood studios in approaching movie selection from a more scientific point of view, namely linear regression modeling. Logically one can assume that a strong plot line is the predominate factor in determining a movie's success, however, many times this is simply not the case. Also, identifying strong plot lines in pre-production phases is highly subjective . The number of variables studios can control, however, are very limited. This is why both endogenous and exogenous variables have been chosen to be included in this analysis. Below is a list of the 13 features chosen, labeled by their respective data types. Our target will be the world-wide box office gross numbers of films. The number of data points is limited to 553, as we have chosen to focus on rom-coms.

FEATURE	DATA TYPE
opening_gross	Quantitative
imdb_score	Quantitative
runtime	Quantitative
total_theaters	Quantitative
distributor	Qualitative
rating	Qualitative
season	Qualitative
director	Qualitative
writer	Qualitative
producer	Qualitative
lead_actor_1	Qualitative
lead_actor_2	Qualitative
actor_3	Qualitative

An initial analysis was done focusing purely on the quantitative features. The heat map below provides the correlations between our target and the selected features as well between the features themselves. Note the high correlation of worldwide gross with opening gross. This could potentially point to a need for studios to heavily focus on the initial pre-marketing phase of a new movie release.



Due to the lack of multi-collinearity between the features, the analysis progressed to the processing of an initial basic linear regression model. The results are shown below.

Dep. Variable:	worldwide_gross	**R-squared:**	0.694
Model:	OLS	**Adj. R-squared:**	0.691
Method:	Least Squares	**F-statistic:**	310.3
Date:	Wed, 07 Jul 2021	**Prob (F-statistic):**	3.07e-139
Time:	15:54:00	**Log-Likelihood:**	-10181.
No. Observations:	553	**AIC:**	2.037e+04
Df Residuals:	548	**BIC:**	2.039e+04
Df Model:	4		
Covariance Type:	nonrobust		

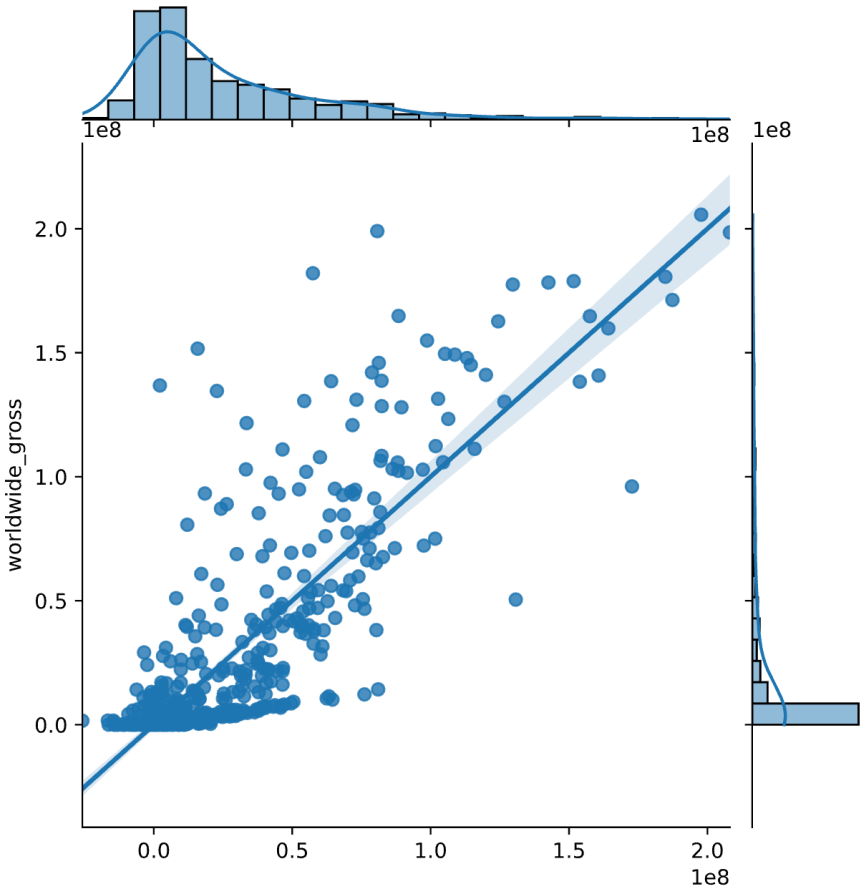
	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	-5.117e+07	1.02e+07	-5.024	0.000	-7.12e+07	-3.12e+07
opening_gross	4.6663	0.271	17.228	0.000	4.134	5.198
imdb_score	5.8e+06	1.23e+06	4.733	0.000	3.39e+06	8.21e+06
runtime	1.498e+05	7.85e+04	1.908	0.057	-4448.993	3.04e+05
total_theaters	5279.6201	1695.056	3.115	0.002	1950.018	8609.222

Omnibus:	235.221	**Durbin-Watson:**	1.801
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1531.831
Skew:	1.746	**Prob(JB):**	0.00
Kurtosis:	10.368	**Cond. No.**	8.07e+07

A few key take aways:

- It seems the model explains nearly 70% of the variation in world-wide gross.
- The model implies that for every $1 made on opening day, that studios on average could expect to see $4.66 in world-wide gross. This would seem logical as reinforced by the low standard error of the feature.
- As one would logically expect, movie review scores play a significant role, as every one point increase in IMDB scores results in a $5.8million increase in gross but with a relatively high variation. Lower rated movies could be still have large ticket sales if backed by a large studio with high production and marketing budgets. Conversely, certain independent movie my be well received by audiences but lack the marketing power of big studios.
- A minute increase in runtime implies a $150,000 increase in gross.

Finally we applied our model to a test set and plotted our predictions versus actual data as seen below with our predictions on the y- axis and actual values on the x-axis.



There is of course a fair amount of variance from our predictions.

FURTHER WORK:

Only a basic initial analysis was conducted so far, with much work left to be done. The causes of the variance seen in the actual vs predicted plot will be further explored. Further analysis of each feature will be conducted to determine if regularization methods can be used to optimize coefficients and reduce this variability. The second set of “dummy” variables will also be added to the analysis to assess how much of an impact each has on the model. Also, a further series of train-test-splits will be conducted to refine our features. A final test will be run with resulting output provided.