ANALYSIS OF BOX OFFICE REVENUE FOR THE FILM INDUSTRY

By Cheryl McGowan Nov 2021

THE PROBLEM

A movie producer is specifically interested in what type of movie to produce next. Can we predict what features influence a movies success at the Box Office?

GLOBAL ENTERTAINMENT $100 BILLION 2020

python™

seaborn

Matplotlib

pandas

Beautiful Soup

DATA

START

Beautiful Soup

Requests Library

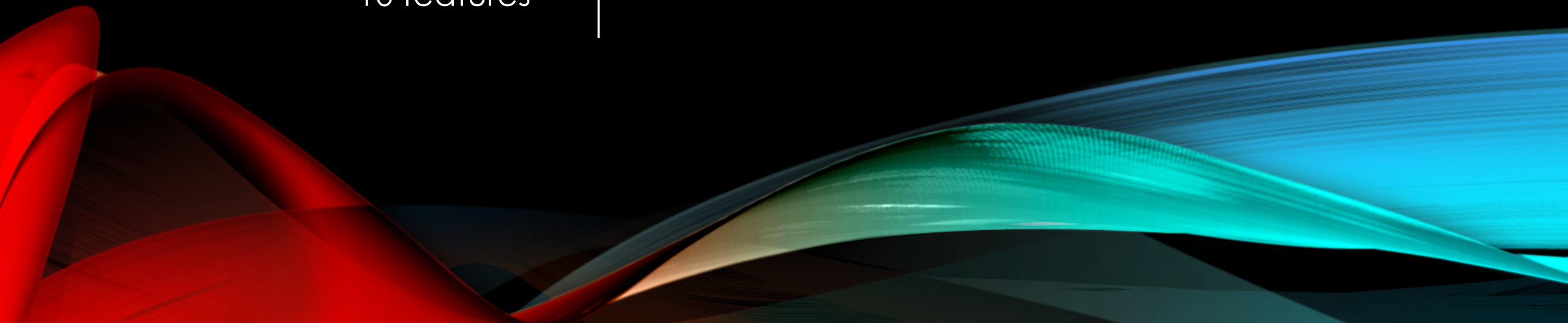
Webscrape the data

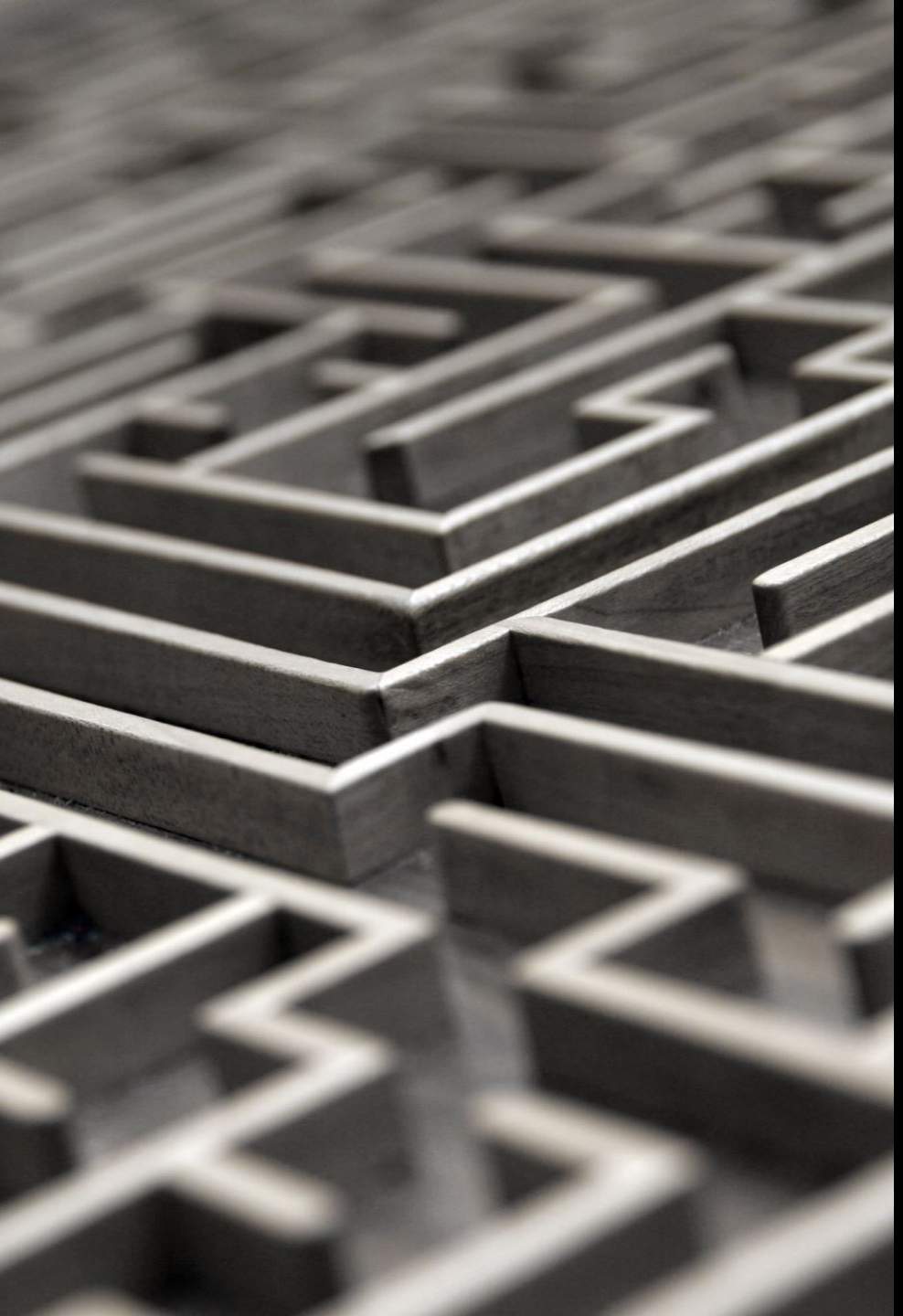
RESULT

1991 Observations

10 features

BOX OFFICE MOJO.COM





ASSUMPTIONS

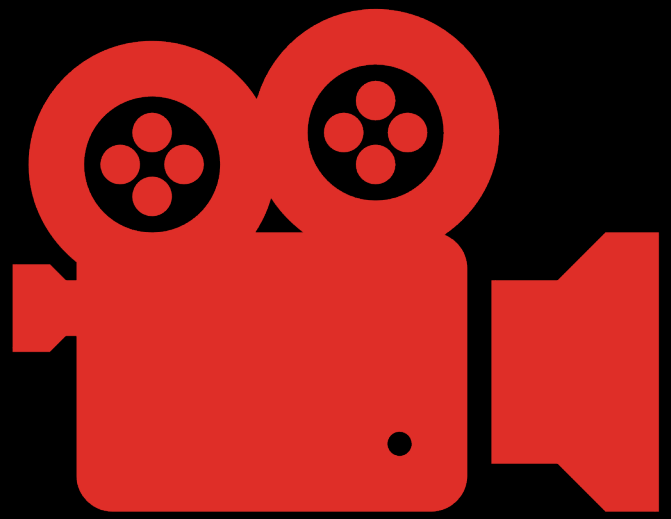
Removed data

- Duplicates in Movie Title
- Runtime / Budget NaN
- Outliers budget over $270,000,000

Multicollinearity

- Domestic Total Gross
- International total gross

OBSERVATION/ROWS = 884
FEATURES = 7



- Movie Title
- $ Budget
- Genre
- Brand
- Franchise
- Release Date
- Runtime

Target = $ Worldwide Total Gross

FEATURE ENGINEERING

Dummy Variables

- Brand – Yes/No
- Franchise Yes/No
- Genre increased model complexity and overfitting

LINEAR REGRESSION MODEL

SPLIT DATA INTO

TRAIN

VALIDATE

TEST

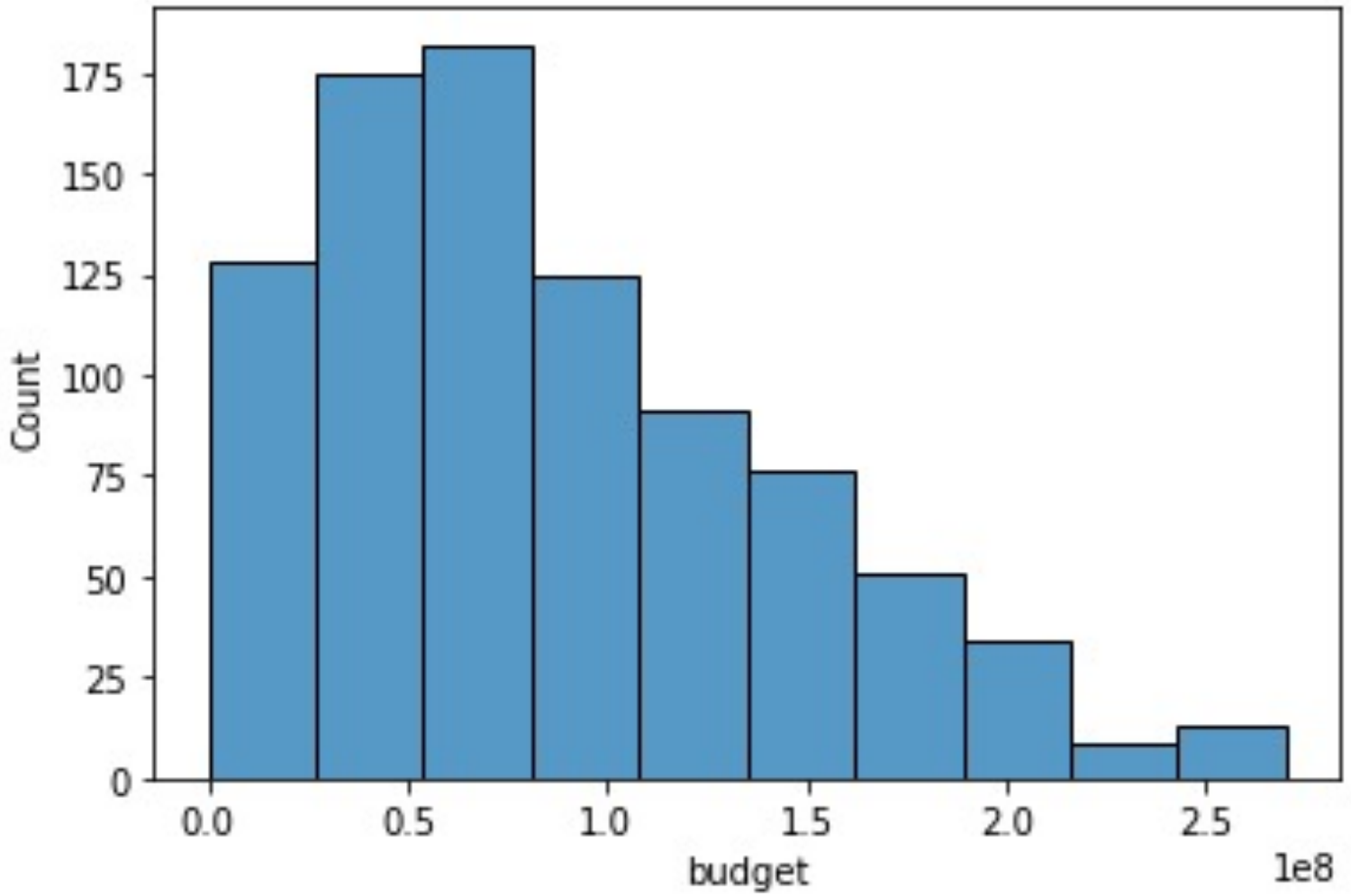
PREDICTIONS AND INTERPRETATIONS

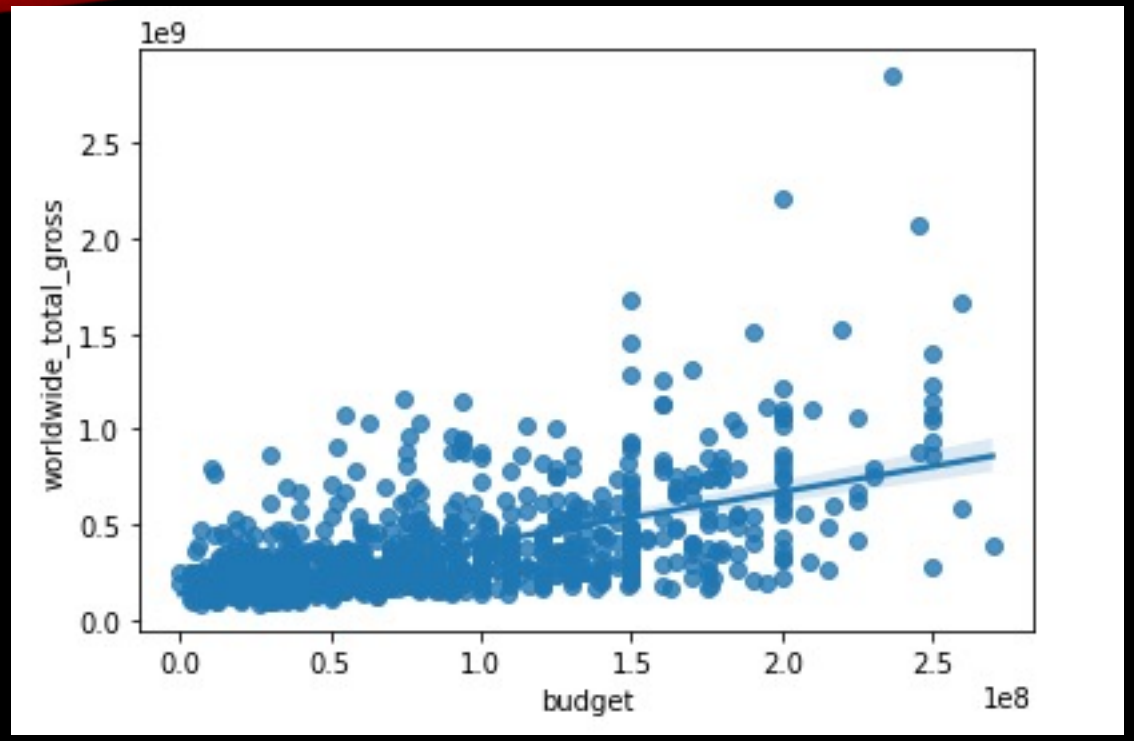
BEST PERFORMANCE

Brand
Franchise
Budget
Runtime

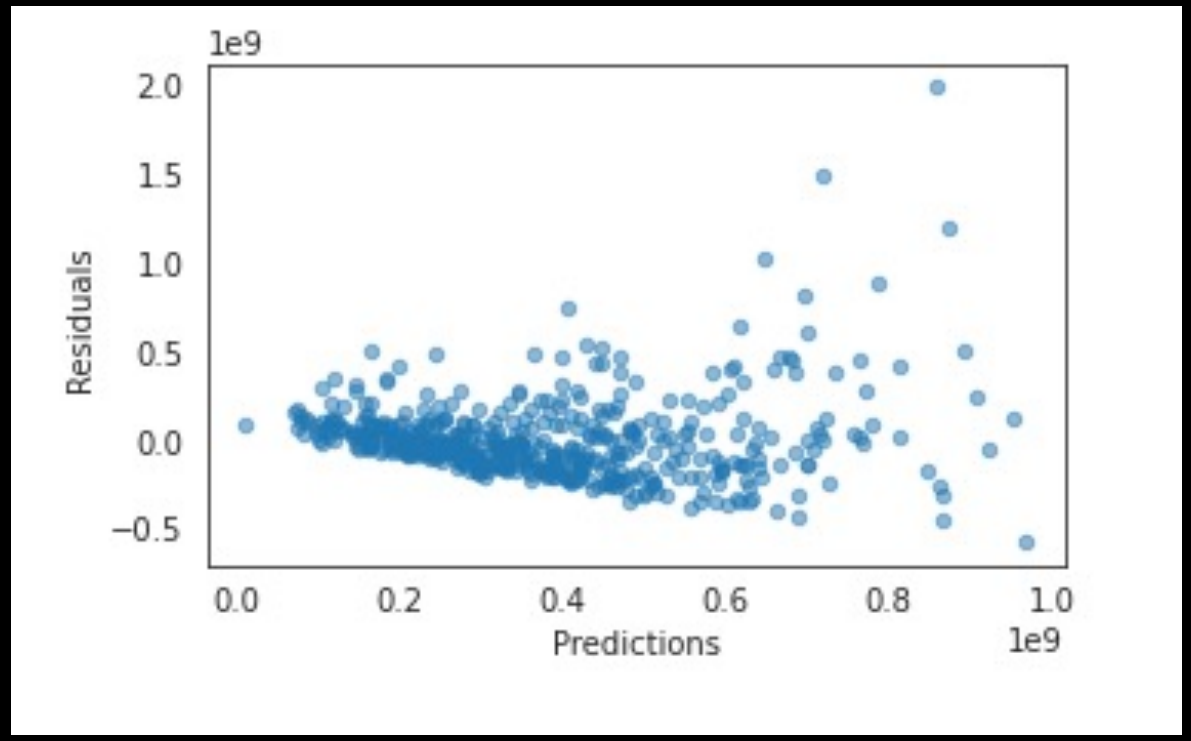
4 features R2 = .425, .433, .395

Budget only R2 = .325, .234, .275


175
150
125
100
75
50
25
0
Count
0.0
0.5
1.0
1.5
2.0
2.5
budget
1e8



RegPlot – Showing the LR Model Fit



Plot Residuals – Showing Heteroskedasticity

FUTURE WORK

STANDARDIZE THE DATA

SCALE OF WORLDWIDE GROSS AND RUNTIME IS OUT OF PROPORTION WHICH COULD BE AFFECTING THE MODEL

USE LOG, SQUARE ROOT OR BOX-COX TO TRANSFORM THE TARGET AND REDUCE HETEROSKEDASTICY

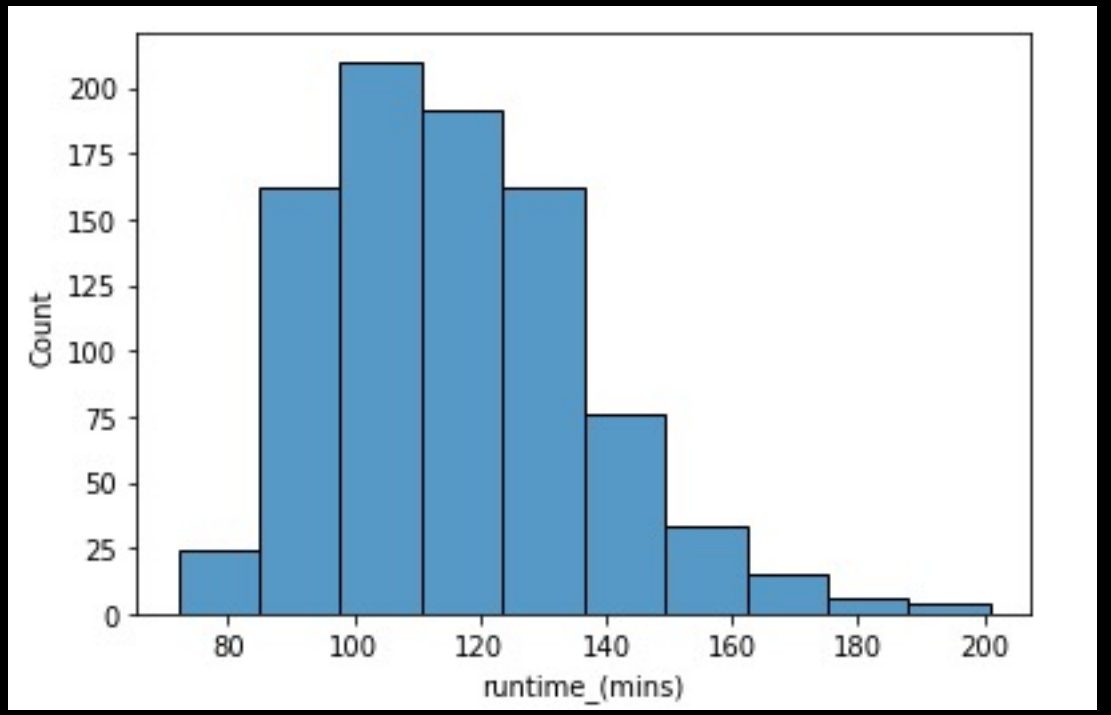
Questions?

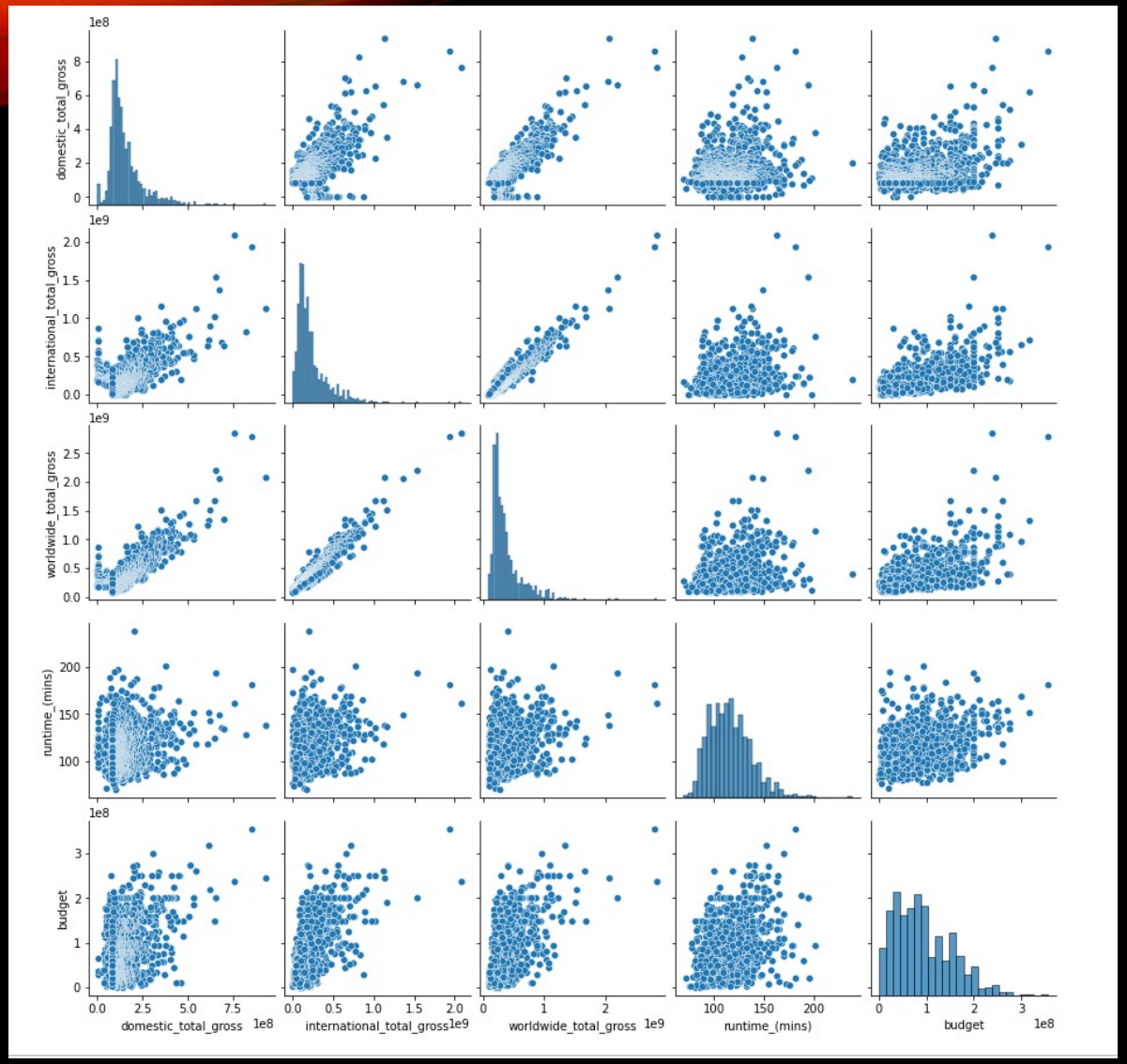
OLS Regression Results

Dep. Variable:	worldwide_total_gross	**R-squared:**	0.425
Model:	OLS	**Adj. R-squared:**	0.422
Method:	Least Squares	**F-statistic:**	162.2
Date:	Mon, 08 Nov 2021	**Prob (F-statistic):**	5.52e-104
Time:	21:39:48	**Log-Likelihood:**	-18191.
No. Observations:	884	**AIC:**	3.639e+04
Df Residuals:	879	**BIC:**	3.642e+04
Df Model:	4		
Covariance Type:	nonrobust		

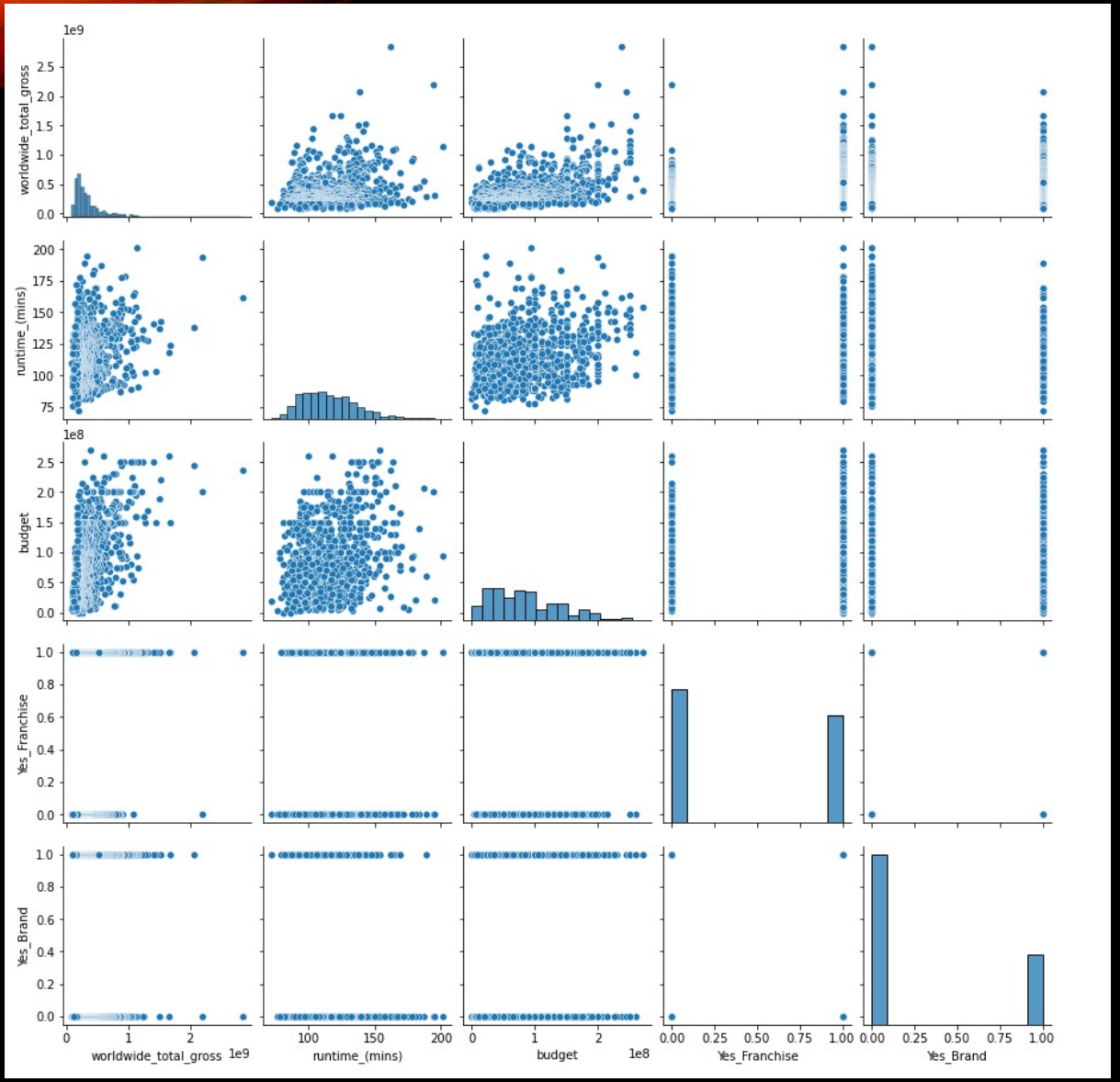
	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.994e+08	4.16e+07	-4.788	0.000	-2.81e+08	-1.18e+08
budget	1.9987	0.136	14.651	0.000	1.731	2.266
Yes_Franchise	1.517e+08	1.49e+07	10.157	0.000	1.22e+08	1.81e+08
Yes_Brand	6.519e+07	1.7e+07	3.826	0.000	3.18e+07	9.86e+07
runtime_(mins)	2.665e+06	3.61e+05	7.384	0.000	1.96e+06	3.37e+06

Omnibus:	539.396	**Durbin-Watson:**	0.813
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	8700.639
Skew:	2.472	**Prob(JB):**	0.00
Kurtosis:	17.553	**Cond. No.**	6.17e+08





Pairplot
Showing Multicollinearity in Domestic Gross, International Gross and Worldwide Gross



Pairplot
Showing final feature set

Image Credits

Imdb.com
Disneyplus.com
Express.co.uk
En.Wikipedia.org
Collider.com
Cnet.com
Appletreemovies.com