

# ANALYSIS OF BOX OFFICE REVENUE FOR THE FILM INDUSTRY

By Cheryl McGowan Nov 2021

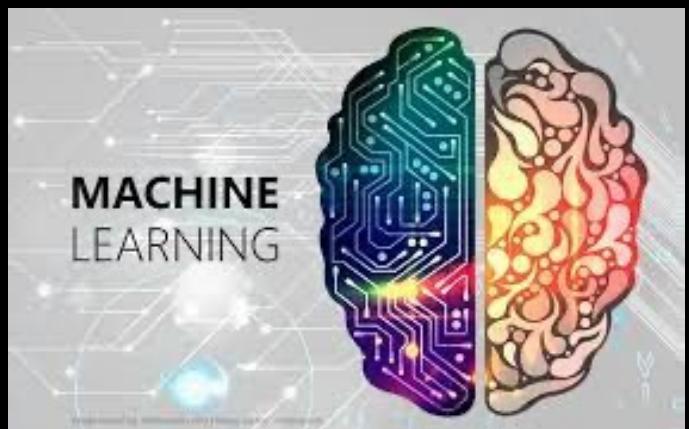
# THE PROBLEM

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?

TIMOTHÉE CHALAMET REBECCA FERGUSON ROCAF JAHN STELLAR DANE STEPHEN MCKINLEY HENDERSON CHANG SHARON BAUTISTA RANDI CHAUVILLE JASON DUNCAN-BREWSTER RAMPLING CHARLOTTE MAMOIA JAVIER BARDEM

The image is split vertically down the center. The left side features a movie poster for "Dune". It shows several characters from the film against a dark, atmospheric background. In the foreground, a figure in a green cloak is shown from behind, looking towards a small figure in the distance. The title "DUNE" is written in large, stylized letters at the bottom, with "IT BEGINS" written above it. The right side features a movie poster for "No Time to Die". It shows Daniel Craig as James Bond looking intensely at the camera. The title "NO TIME TO DIE" is written in large, bold, white letters at the bottom.

GLOBAL ENTERTAINMENT  
\$100 BILLION 2020



python™

seaborn



Matplotlib

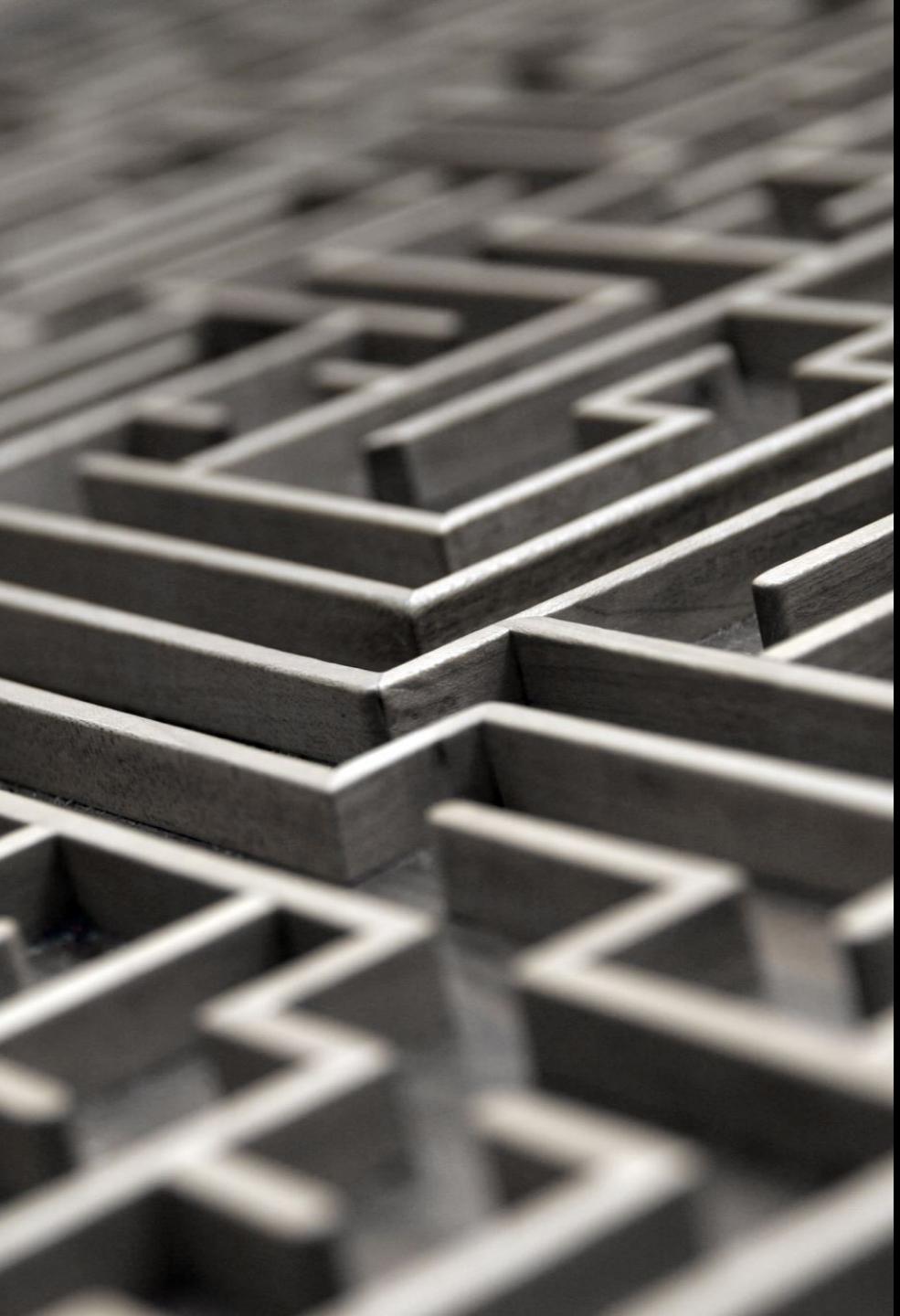
pandas

BeautifulSoup

START  
Beautiful Soup  
Requests Library  
Webscrape the data

RESULT  
1991 Observations  
10 features

DATA  
BOX  
OFFICE  
MOJO.COM

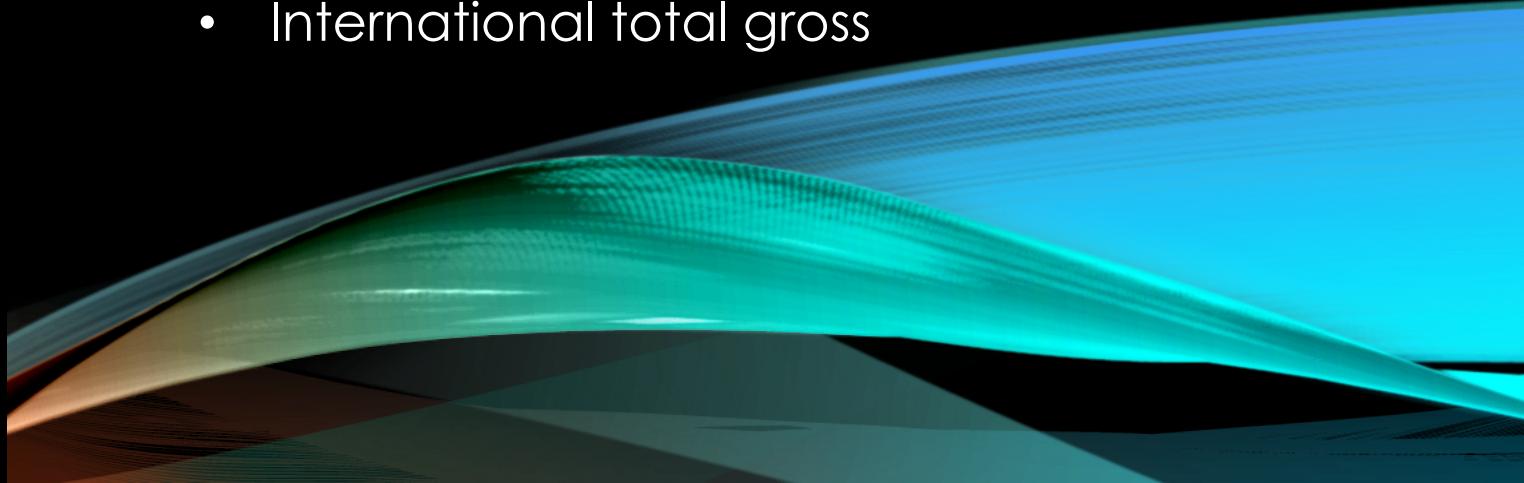


# ASSUMPTIONS

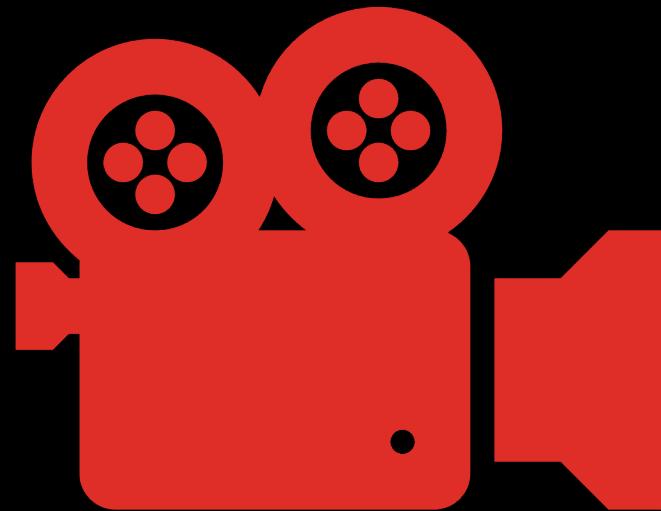
## Removed data

- 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



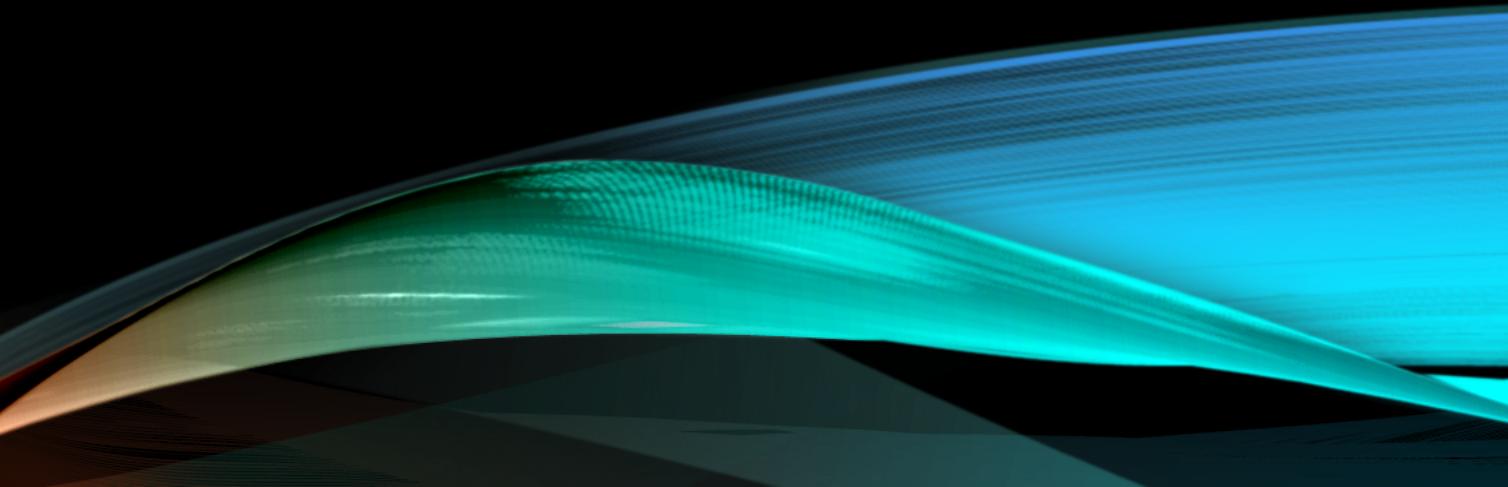
VALIDATE



TEST



PREDICTIONS  
AND  
INTERPRETATIONS



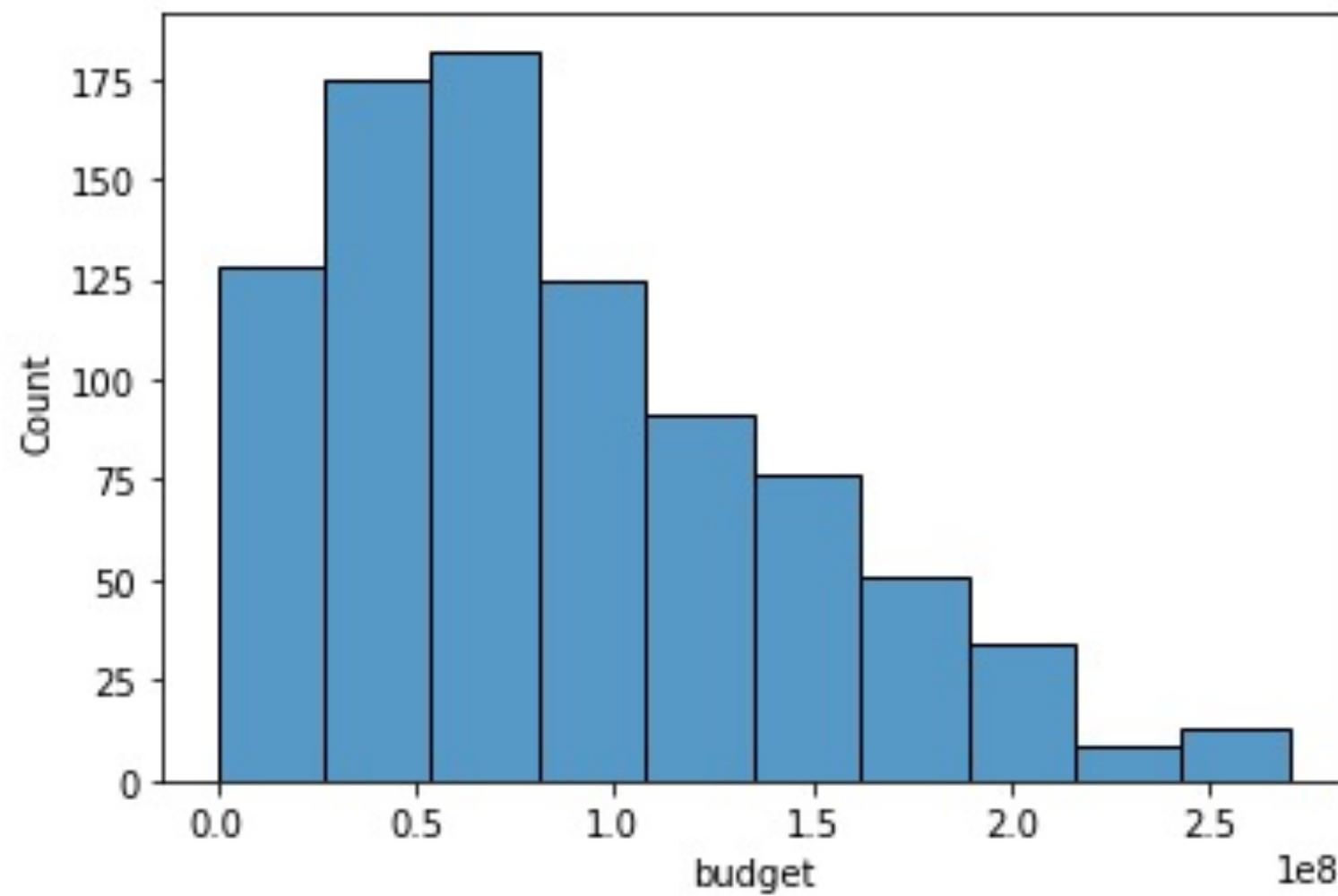
## BEST PERFORMANCE

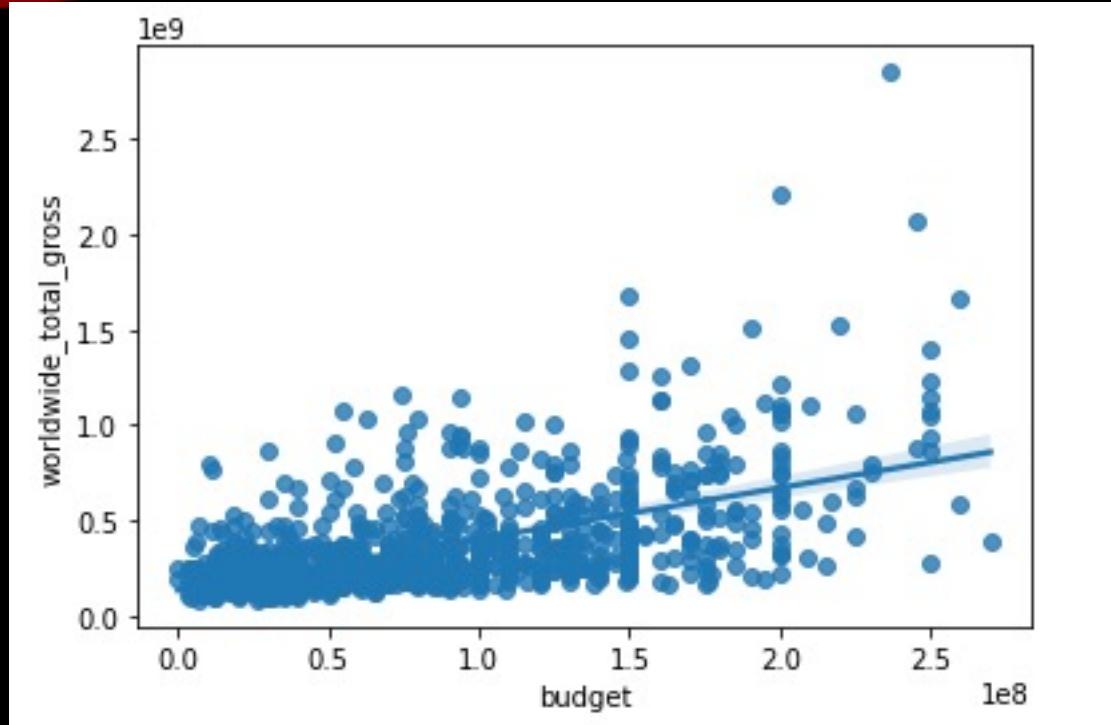


Brand  
Franchise  
Budget  
Runtime

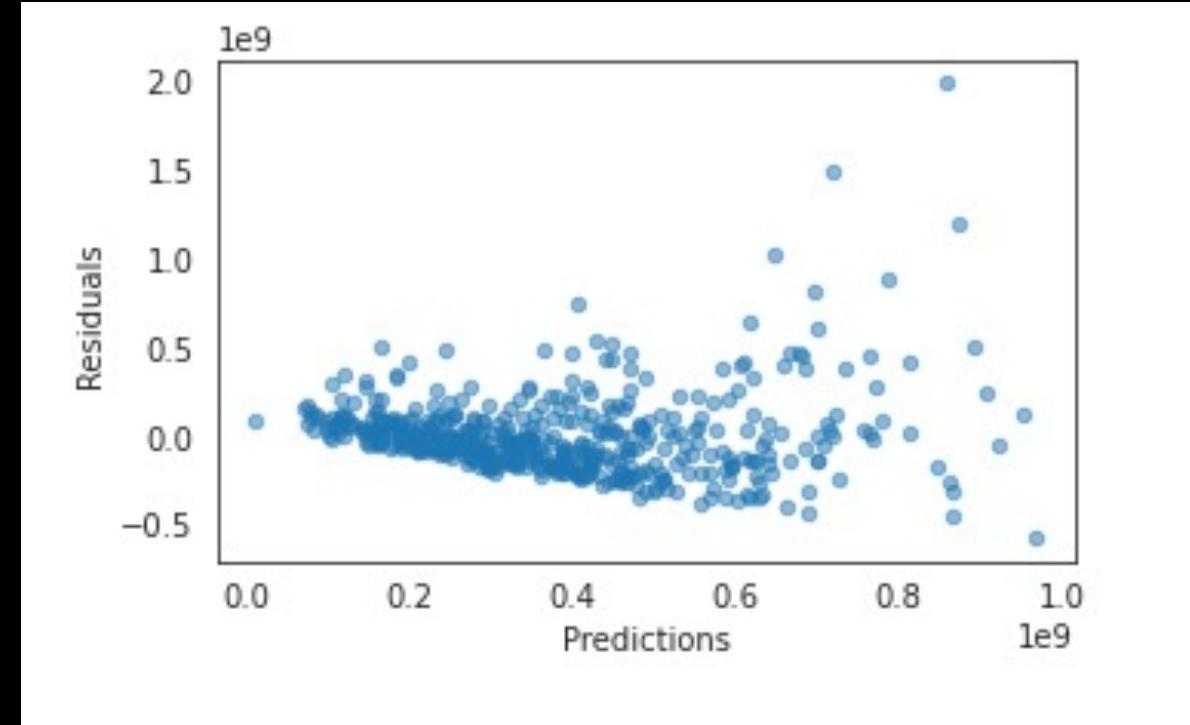


4 features R2 = .425, .433, .395  
Budget only R2 = .325, .234, .275





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



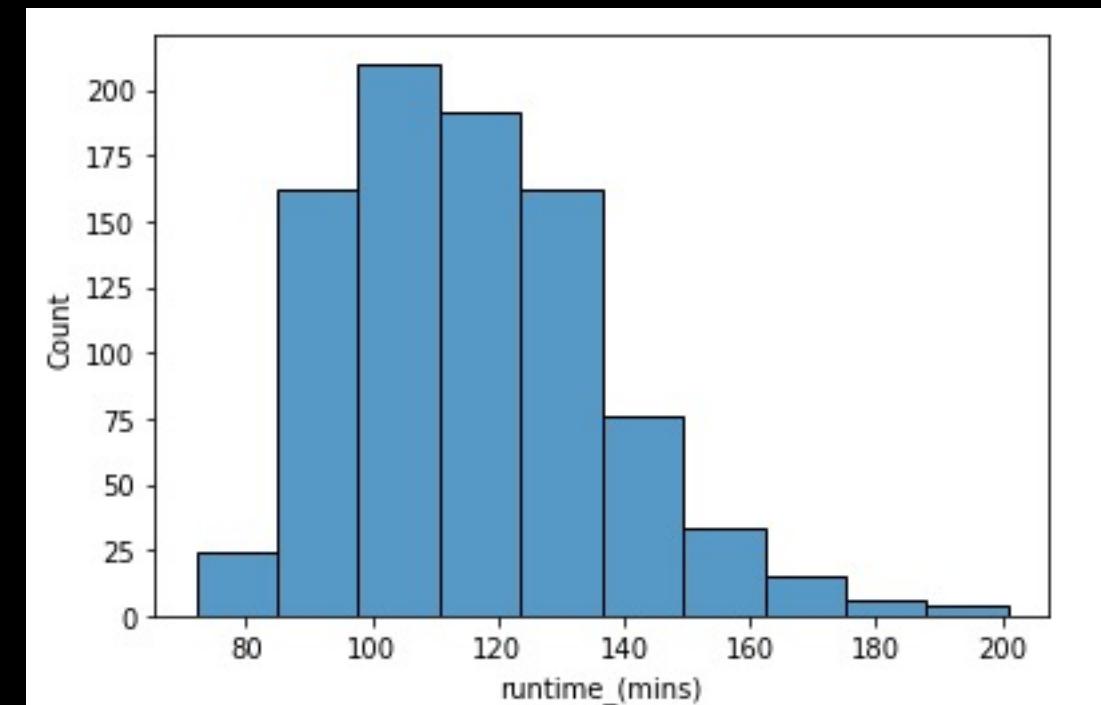
USE LOG, SQUARE ROOT OR BOX-COX  
TO TRANSFORM THE TARGET AND  
REDUCE HETEROSKEDASTICITY

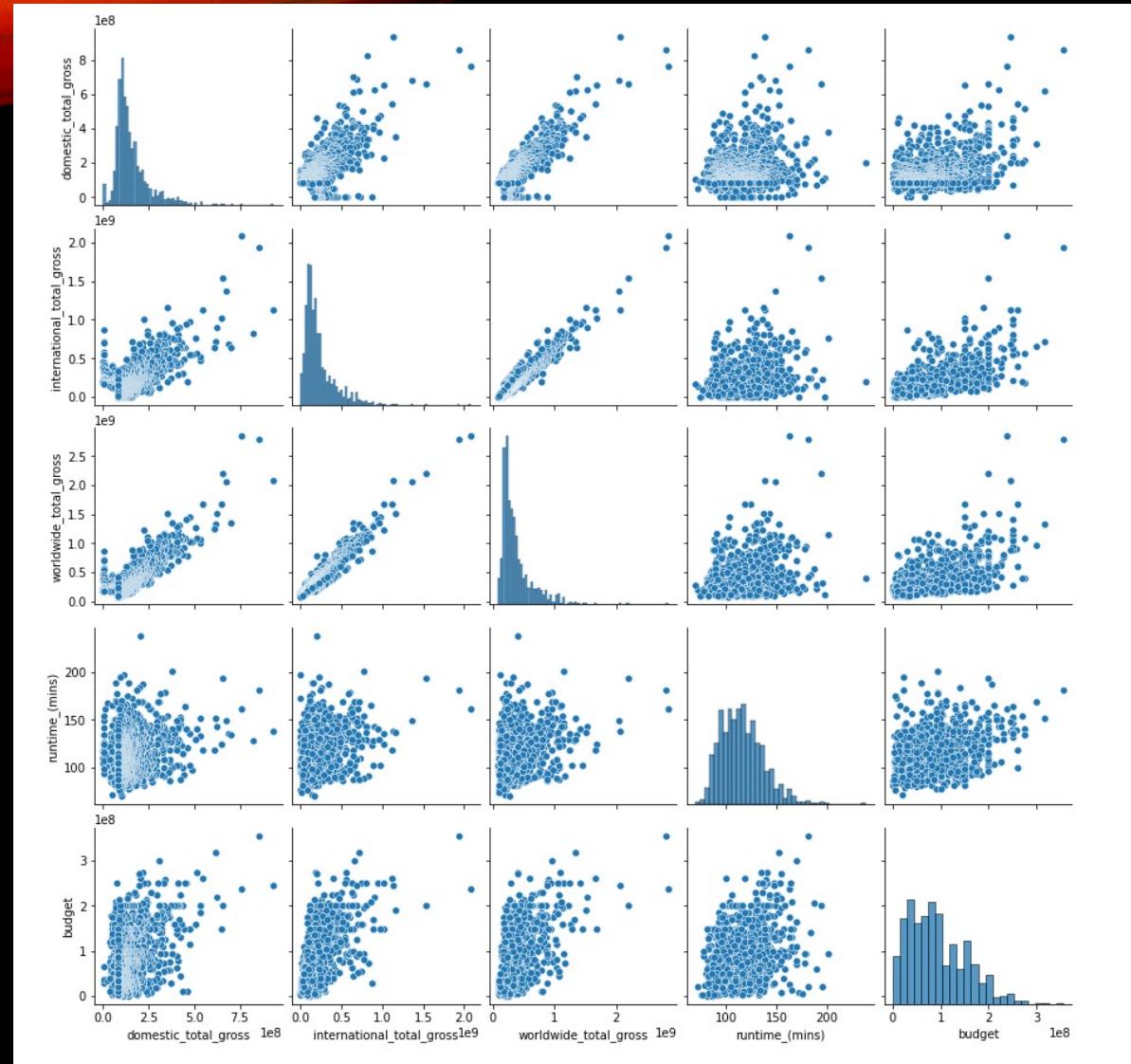


Questions?

### OLS Regression Results

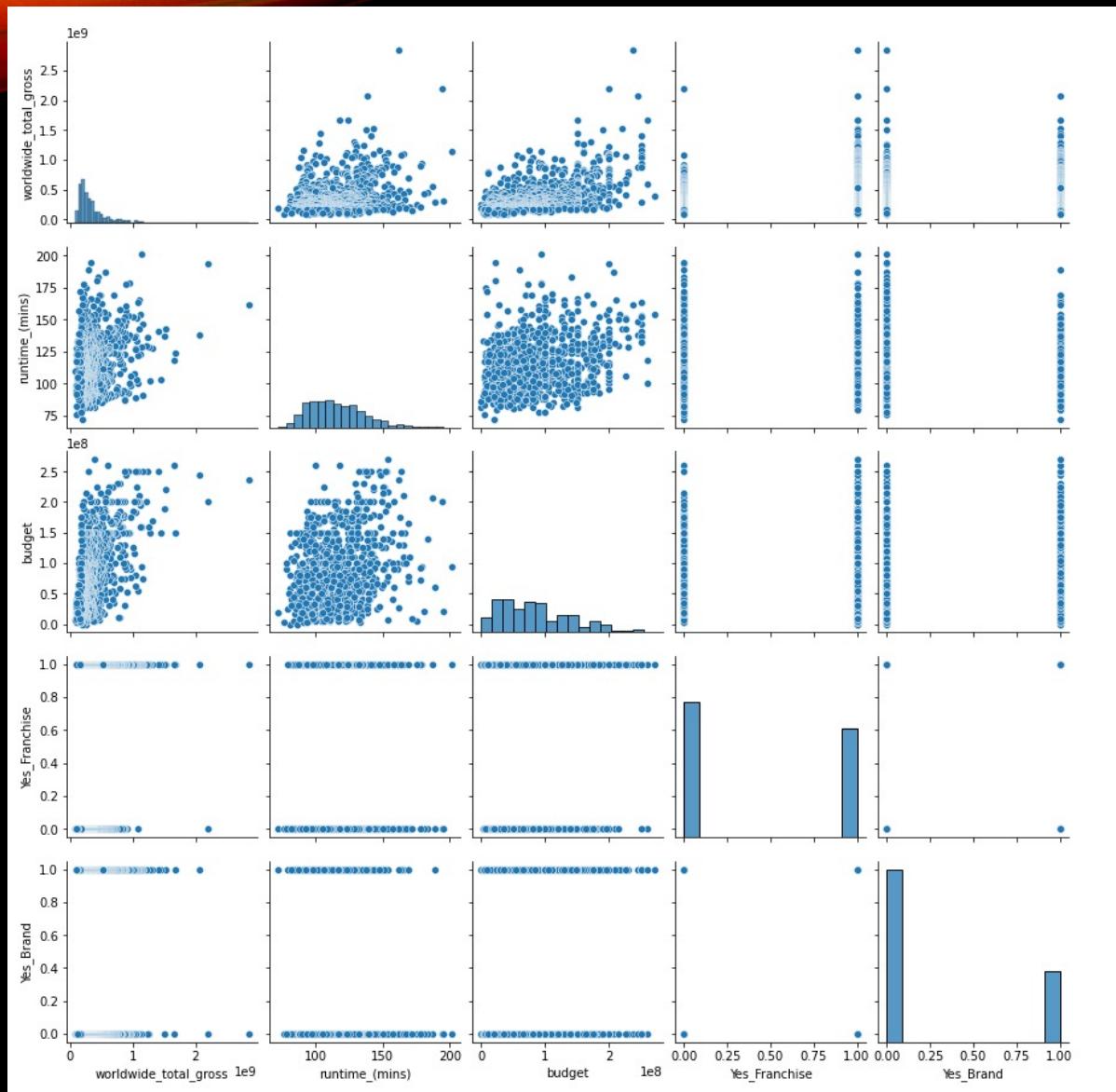
<b>Dep. Variable:</b>	worldwide_total_gross	<b>R-squared:</b>	0.425			
<b>Model:</b>	OLS	<b>Adj. R-squared:</b>	0.422			
<b>Method:</b>	Least Squares	<b>F-statistic:</b>	162.2			
<b>Date:</b>	Mon, 08 Nov 2021	<b>Prob (F-statistic):</b>	5.52e-104			
<b>Time:</b>	21:39:48	<b>Log-Likelihood:</b>	-18191.			
<b>No. Observations:</b>	884	<b>AIC:</b>	3.639e+04			
<b>Df Residuals:</b>	879	<b>BIC:</b>	3.642e+04			
<b>Df Model:</b>	4					
<b>Covariance Type:</b>	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
<b>Omnibus:</b>	539.396	<b>Durbin-Watson:</b>	0.813			
<b>Prob(Omnibus):</b>	0.000	<b>Jarque-Bera (JB):</b>	8700.639			
<b>Skew:</b>	2.472	<b>Prob(JB):</b>	0.00			
<b>Kurtosis:</b>	17.553	<b>Cond. No.</b>	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