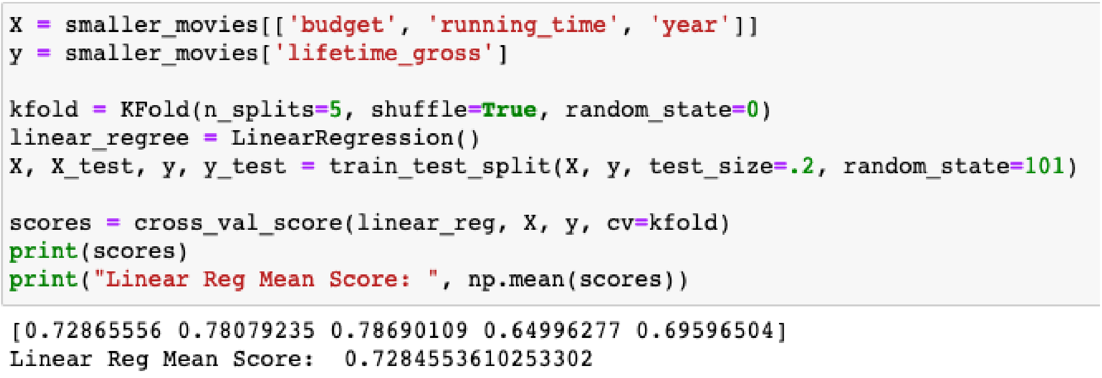
MVP

I started with exploring the numerical variables first and chose the most relevant features to build my baseline model. In the next couple days, I will focus on Ridge/Lasso & Polynomials and adding categorical features to find the best fitting models.

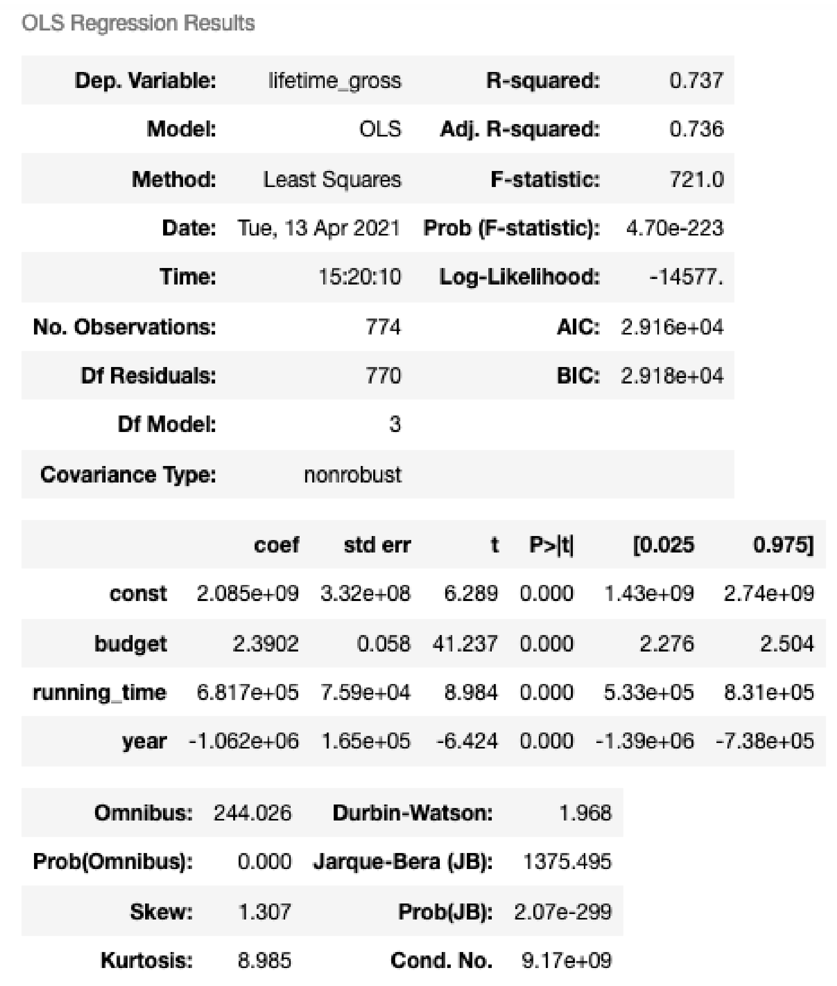


```
X = smaller_movies[['budget', 'running_time', 'year']]
y = smaller_movies['lifetime_gross']

kfold = KFold(n_splits=5, shuffle=True, random_state=0)
linear_regree = LinearRegression()
X, X_test, y, y_test = train_test_split(X, y, test_size=.2, random_state=101)

scores = cross_val_score(linear_reg, X, y, cv=kfold)
print(scores)
print("Linear Reg Mean Score: ", np.mean(scores))
```

```
[0.72865556 0.78079235 0.78690109 0.64996277 0.69596504]
Linear Reg Mean Score:  0.7284553610253302
```



OLS Regression Results

Dep. Variable:	lifetime_gross	**R-squared:**	0.737
Model:	OLS	**Adj. R-squared:**	0.736
Method:	Least Squares	**F-statistic:**	721.0
Date:	Tue, 13 Apr 2021	**Prob (F-statistic):**	4.70e-223
Time:	15:20:10	**Log-Likelihood:**	-14577.
No. Observations:	774	**AIC:**	2.916e+04
Df Residuals:	770	**BIC:**	2.918e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	2.085e+09	3.32e+08	6.289	0.000	1.43e+09	2.74e+09
budget	2.3902	0.058	41.237	0.000	2.276	2.504
running_time	6.817e+05	7.59e+04	8.984	0.000	5.33e+05	8.31e+05
year	-1.062e+06	1.65e+05	-6.424	0.000	-1.39e+06	-7.38e+05

Omnibus:	244.026	**Durbin-Watson:**	1.968
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1375.495
Skew:	1.307	**Prob(JB):**	2.07e-299
Kurtosis:	8.985	**Cond. No.**	9.17e+09