THE BEST MOVIE INVESTMENT PLAN

Web Scraping and Linear Regression Project to Predict Movie Revenue

YINGQING QIU
METIS DATA SCIENCE BOOTCAMP
2021 FALL COHORT

INTRODUCTION

Question: How much money can a movie make?

Data Source: https://www.boxofficemojo.com/

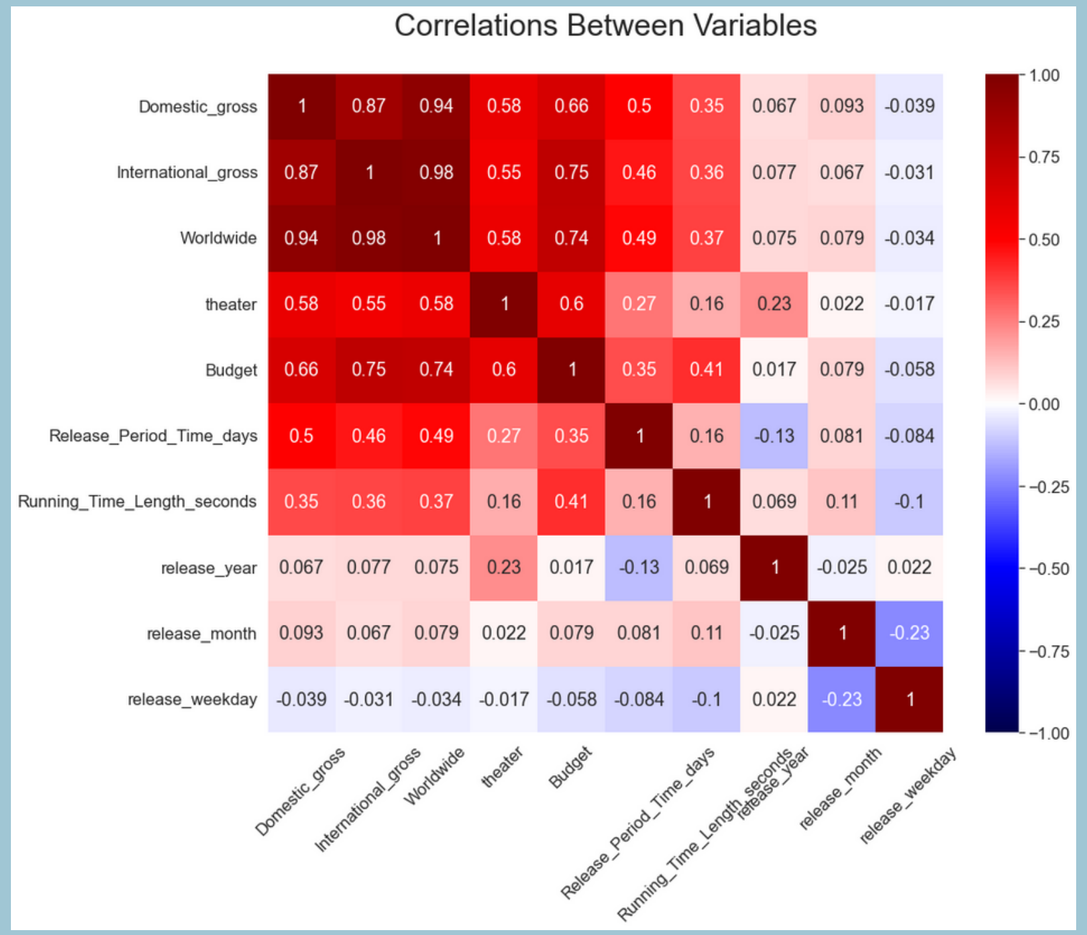
Data: ~1100 rows; ~12 features (~3 categorical features)

Analysis Tools: BeautifulSoup
Numpy, Pandas, Scipy
Scikit-learn
Matplotlib, Seaborn

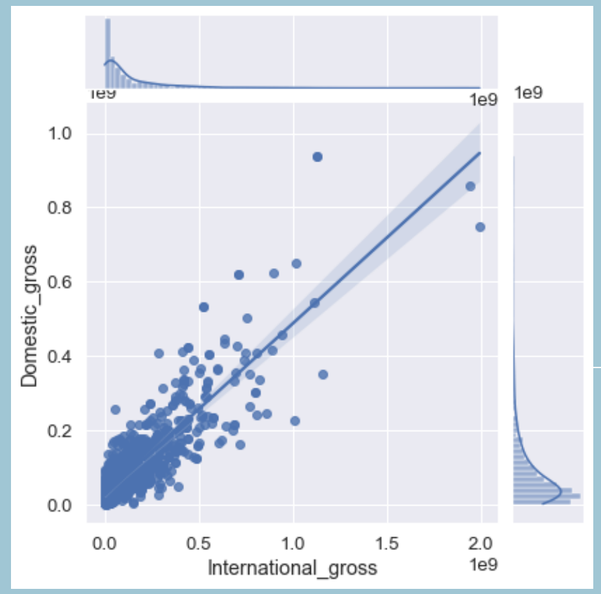
```
df.info()
```

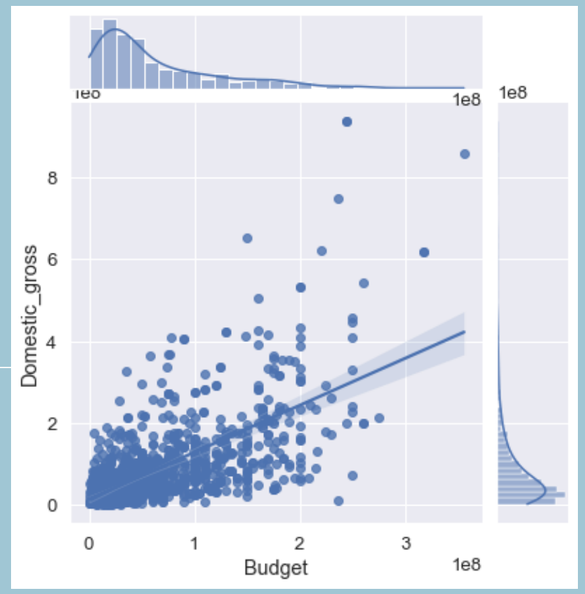
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1174 entries, 0 to 1173
Data columns (total 14 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   Title                         1174 non-null   object
 1   Domestic_gross                1174 non-null   int64
 2   International_gross           1174 non-null   int64
 3   Worldwide                     1174 non-null   float64
 4   theater                       1174 non-null   int64
 5   distributer                   1174 non-null   object
 6   Genres_thefirst               1174 non-null   object
 7   MPAA                          1174 non-null   object
 8   Budget                        1174 non-null   float64
 9   Release_Period_Time_days      1174 non-null   int64
 10  Running_Time_Length_seconds   1174 non-null   float64
 11  release_year                  1174 non-null   int64
 12  release_month                 1174 non-null   int64
 13  release_weekday               1174 non-null   int64
dtypes: float64(3), int64(7), object(4)
memory usage: 128.5+ KB
```

Basic Analysis



Correlation Of One Feature





y: Domestic Gross Revenue ($)

x_left: International Gross Revenue ($)

x_right: Movie Budget ($)

Feature Engineering

Drop Features:

1. International Gross
2. Worldwide Gross

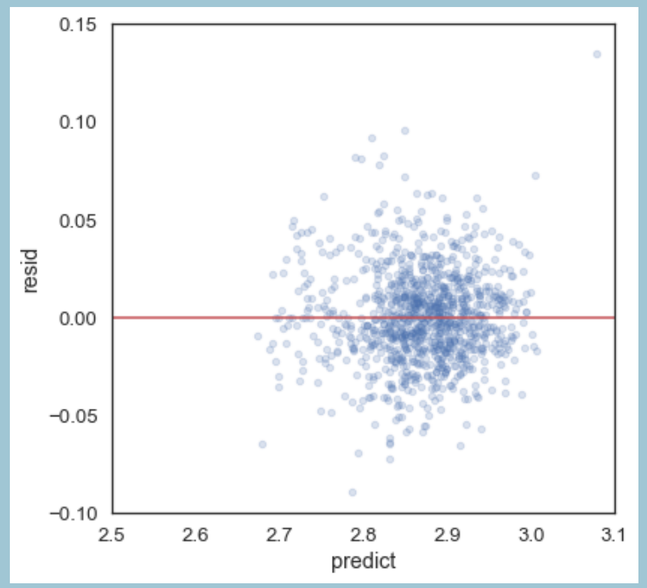
Convert Categorical Features to dummy variables:

1. Distributer
2. Genres
3. MPAA

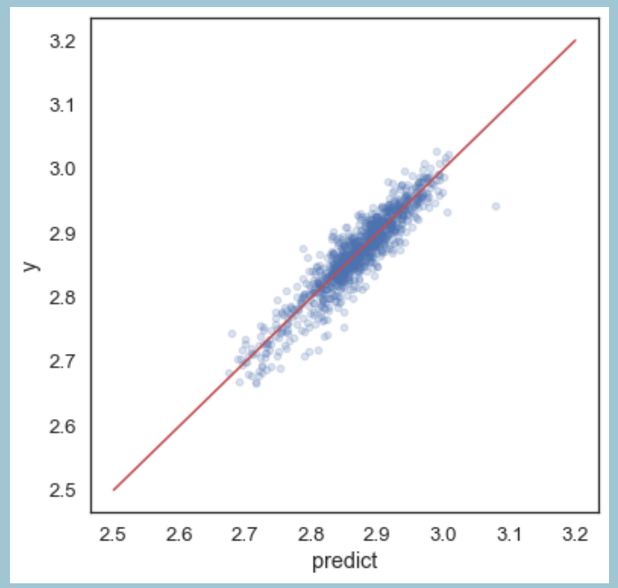
Linear Regression Methods and Evaluation Matrics

EVALUATION MATRICS	SIMPLE LINEAR REGRESSION	SIMPLE LINEAR REGRESSION (CV)	POLYNOMIAL REGRESSION (DEGREE = 2)	LASSO REGRESSION (CV)	RIDGE REGRESSION (CV)	ELASTICNET REGRESSION (CV)
Train, R^2	0.865	0.875	0.912	0.857	0.864	0.858
Test, R^2	0.847	0.881	0.356	0.837	0.836	0.837
RMSE (log $)	0.416	0.416	0.497	0.428	0.421	0.428
MAE (log $)	0.317	0.318	0.281	0.331	0.321	0.331

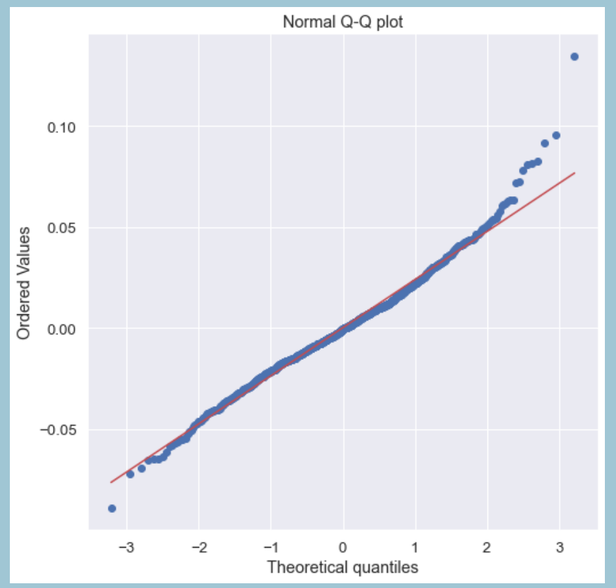
Residual Pattern



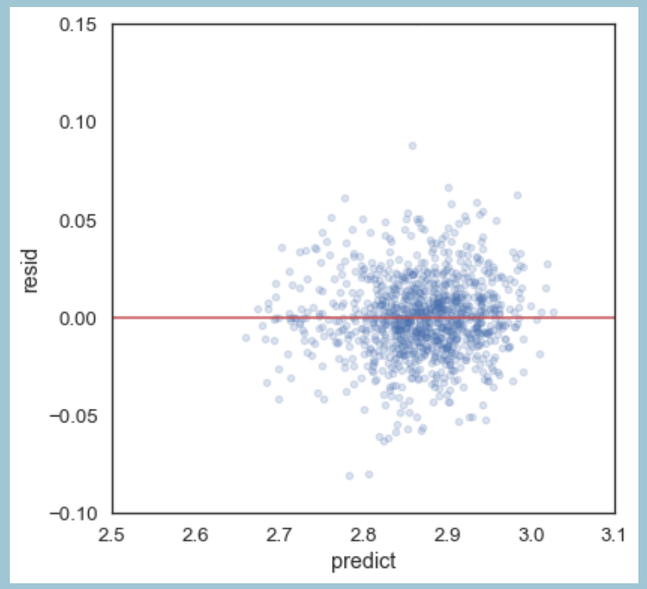
Target - Predict

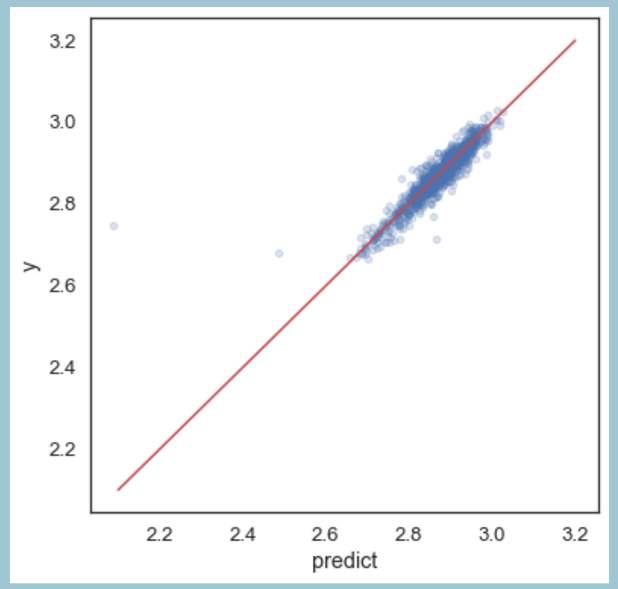


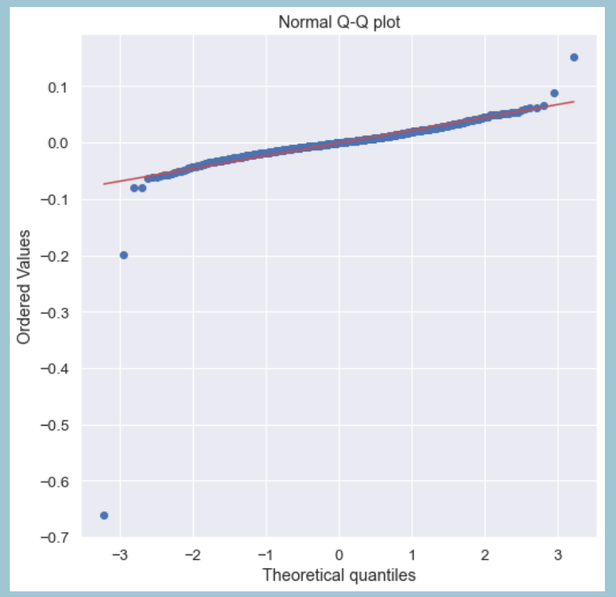
Q-Q plot



Simple Linear Regression (Cross-Validation)







Polynomial Regression (degree = 2)

Predict Revenue for 3 Movies

Title	Theater	Distributer	Genres	MPAA	Budget ($)	Release Periods (days)	Running Time (min)	Release Year	Release Month	Release Weekday
Did You Hear About the Morgans?	2718	Sony Pictures Entertainment (SPE)	Comedy	PG-13	58000000	37	103	2009	12	4
Despicable Me	3602	Universal Pictures	Adventure	PG	69000000	195	95	2010	7	4
Alita: Battle Angel	3802	Twentieth Century Fox	Action	PG-13	170000000	84	122	2019	2	3

	PREDICTED GROSS	REAL GROSS
Did You Hear About the Morgans?	$ 2.45E7	$ 2.95E7
Despicable Me	$ 2.95E8	$ 2.52E8
Alita: Battle Angel	$ 6.87E7	$ 8.58E7

FUTURE WORK

Add more useful features

Actor/Actress

Shooting Location

GDP

...

Incorporate COVID impact

Find Daily Revenue Data

...

THANK YOU!

ANY QUESTIONS?