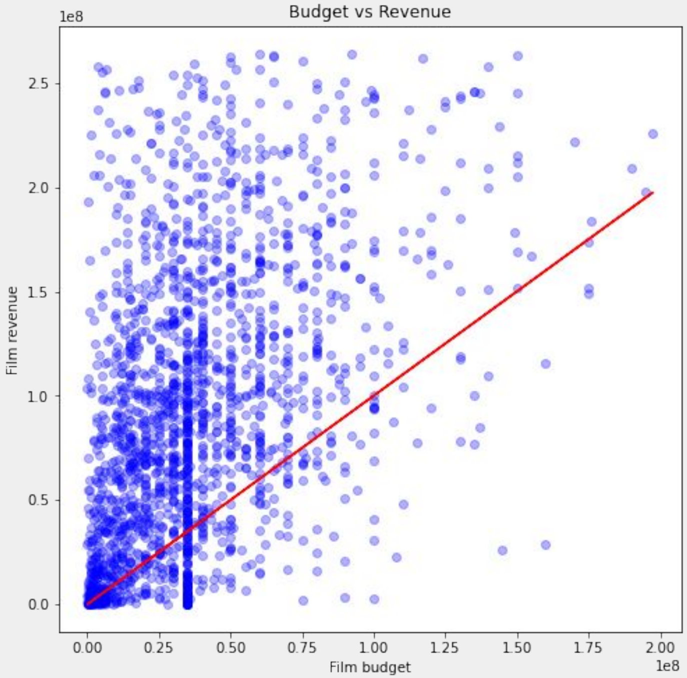
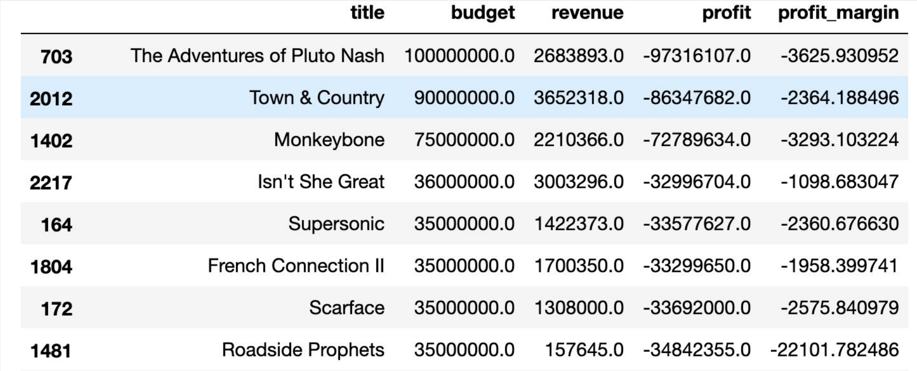
Springboard - Data Science Track
Capstone Project 1:
Modeling Film Revenue

By: Nicholas Dean

June 2021

Defining the problem

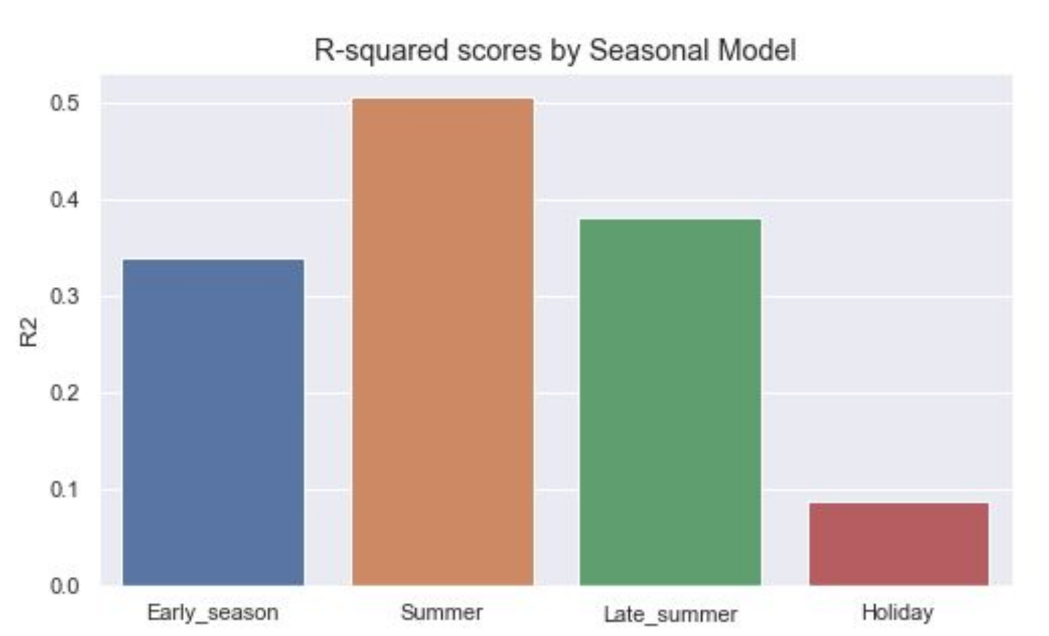




	title	budget	revenue	profit	profit_margin
703	The Adventures of Pluto Nash	100000000.0	2683893.0	-97316107.0	-3625.930952
2012	Town & Country	90000000.0	3652318.0	-86347682.0	-2364.188496
1402	Monkeybone	75000000.0	2210366.0	-72789634.0	-3293.103224
2217	Isn't She Great	36000000.0	3003296.0	-32996704.0	-1098.683047
164	Supersonic	35000000.0	1422373.0	-33577627.0	-2360.676630
1804	French Connection II	35000000.0	1700350.0	-33299650.0	-1958.399741
172	Scarface	35000000.0	1308000.0	-33692000.0	-2575.840979
1481	Roadside Prophets	35000000.0	157645.0	-34842355.0	-22101.782486

Stakeholders

Bottom Line


R-squared scores by Seasonal Model
R2
0.0
0.1
0.2
0.3
0.4
0.5
Early_season
Summer
Late_summer
Holiday

Data Acquisition

Playground Prediction Competition

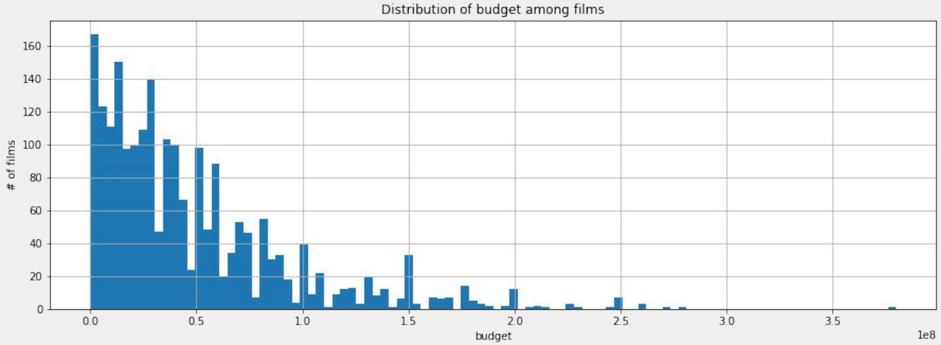
TMDB Box Office Prediction

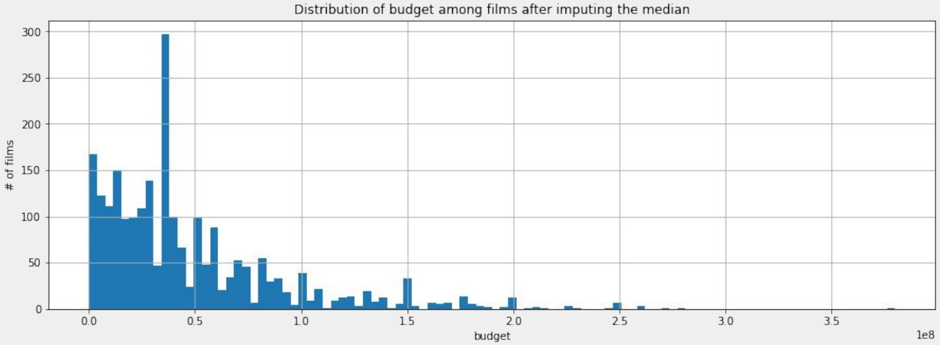
Can you predict a movie's worldwide box office revenue?

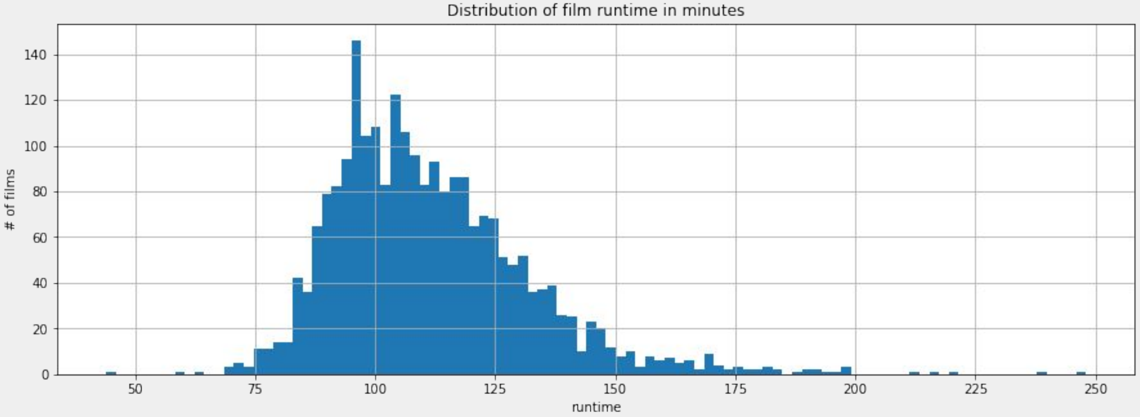
Kaggle · 1,395 teams · 2 years ago

IMDbPy

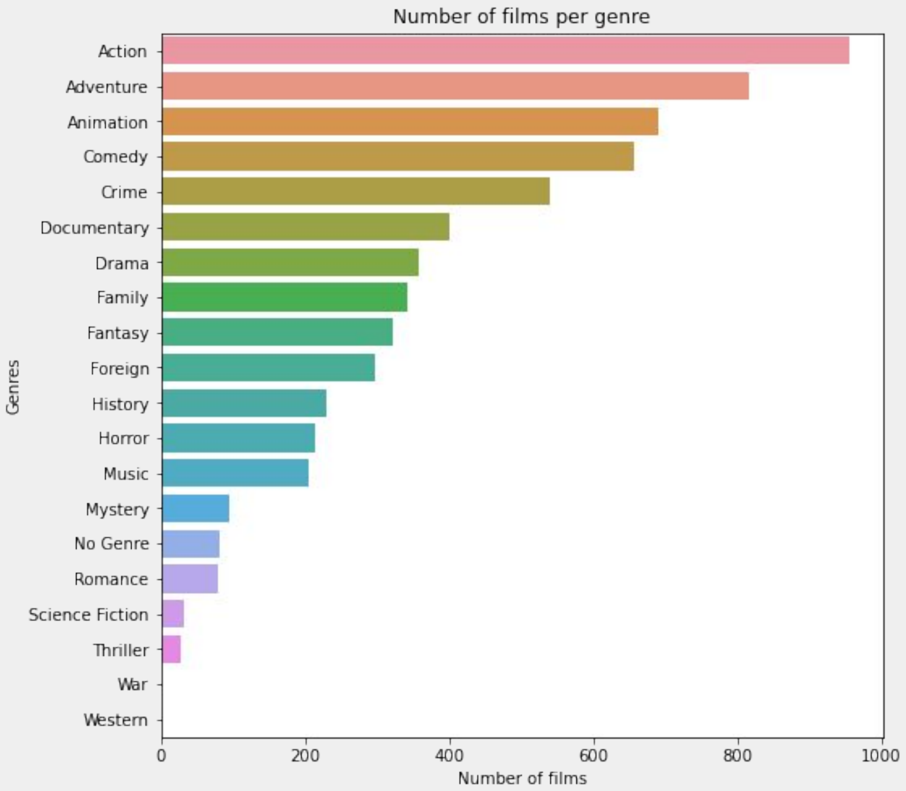
Data Wrangling: Numerical Data

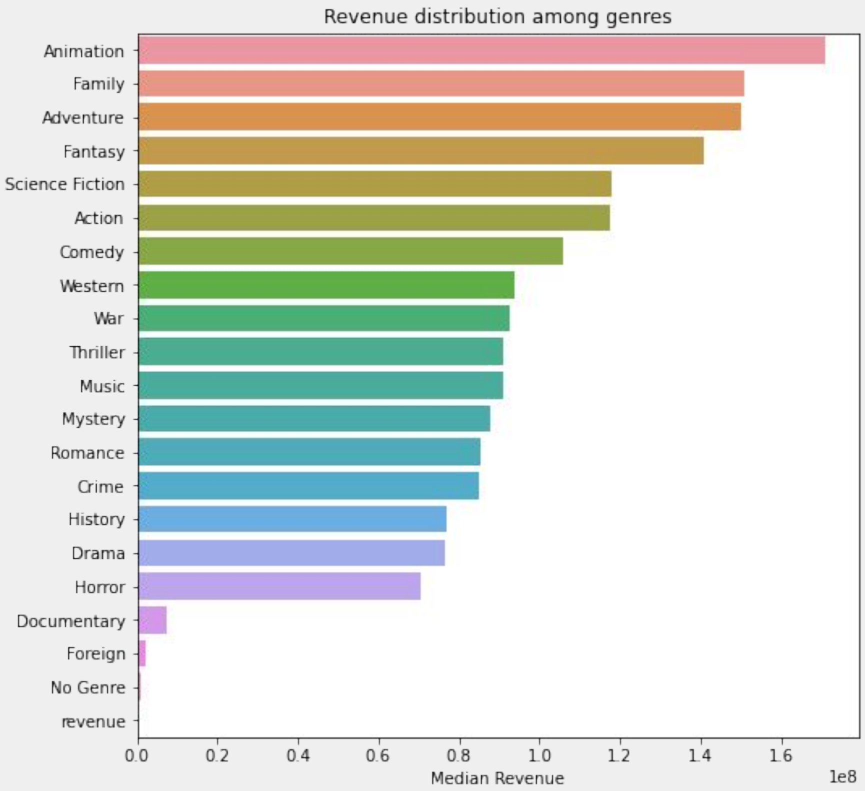

Distribution of budget among films


Distribution of budget among films after imputing the median

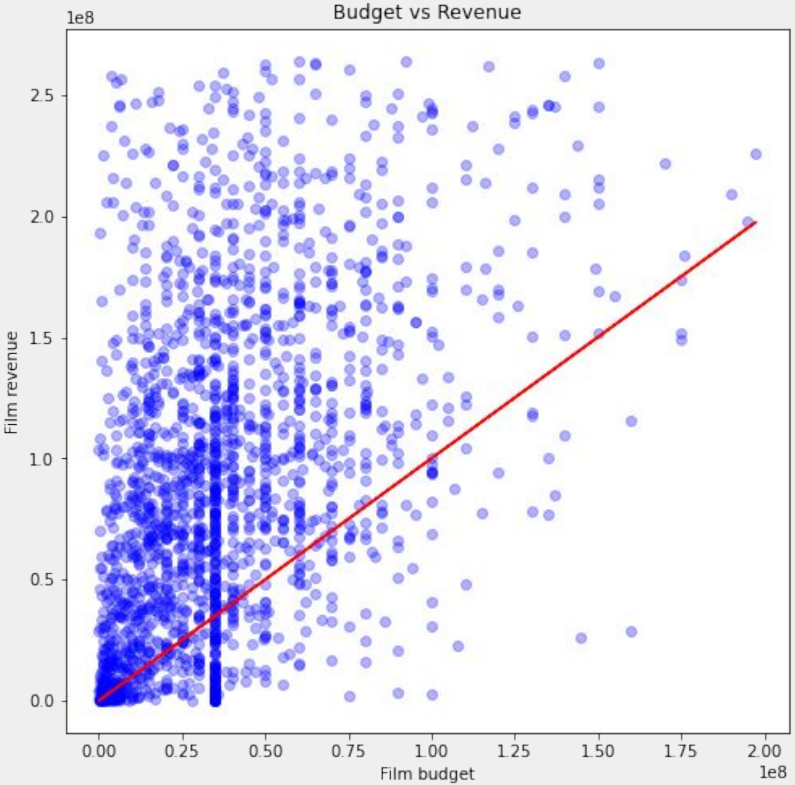

Distribution of film runtime in minutes

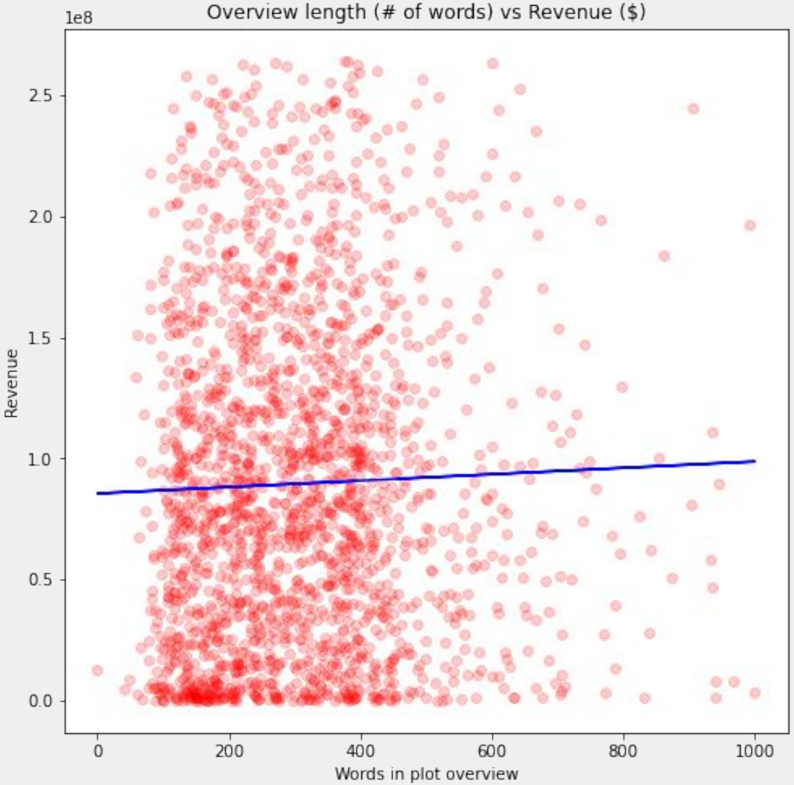
Data Wrangling: Categorical Data



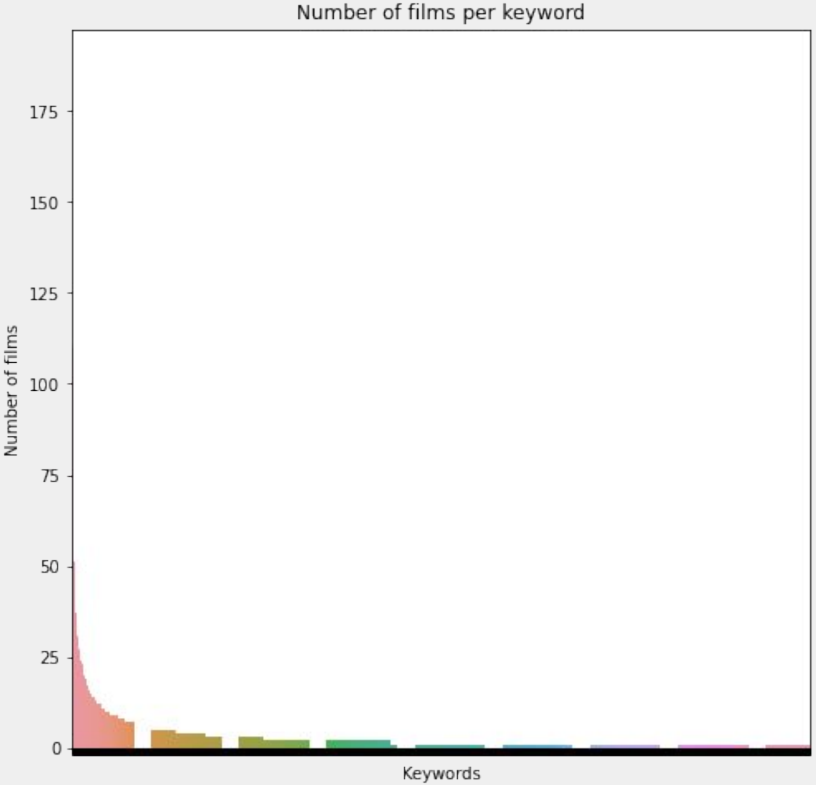


Basics of the Data

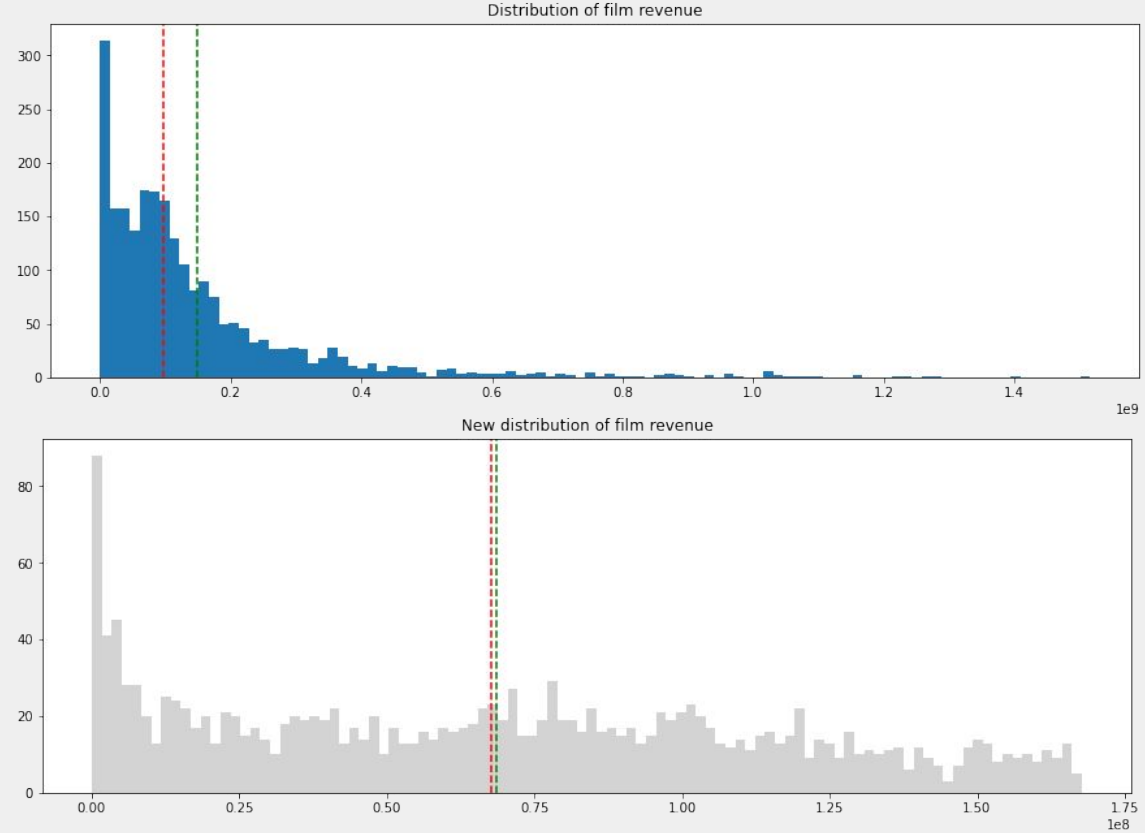




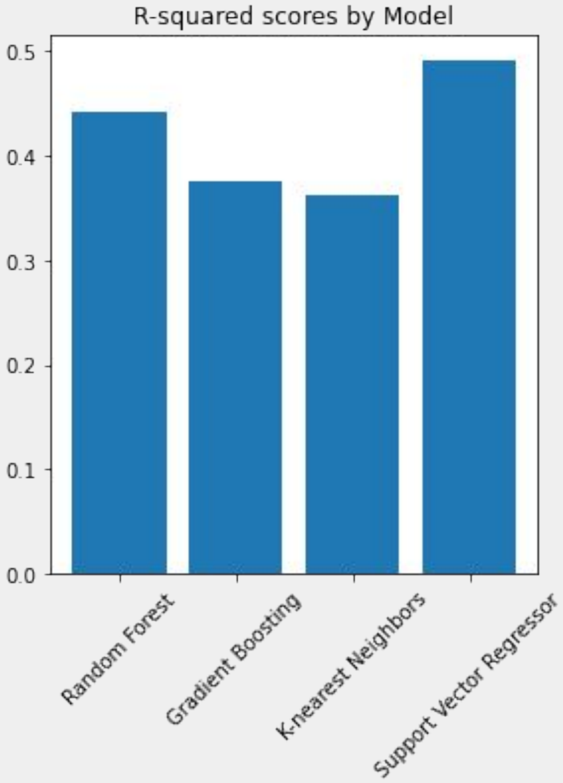
Basics of Categorical Data

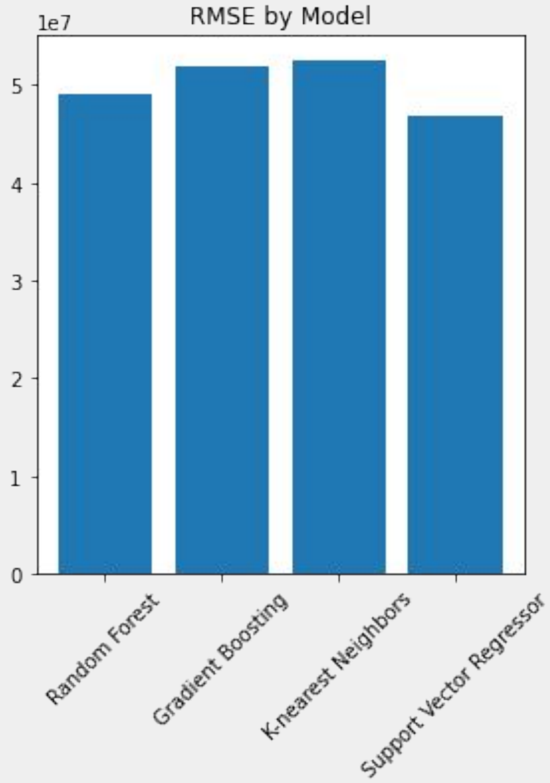


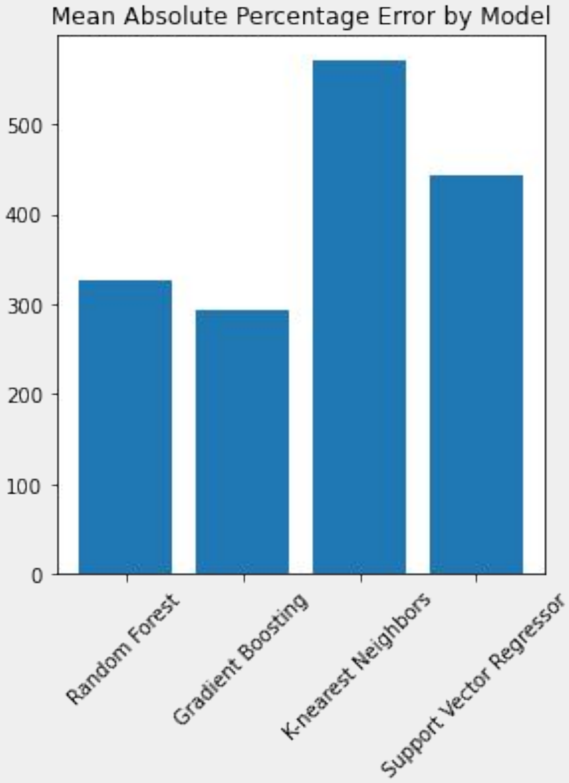
Modeling Revenue: Linear Regression & Challenges



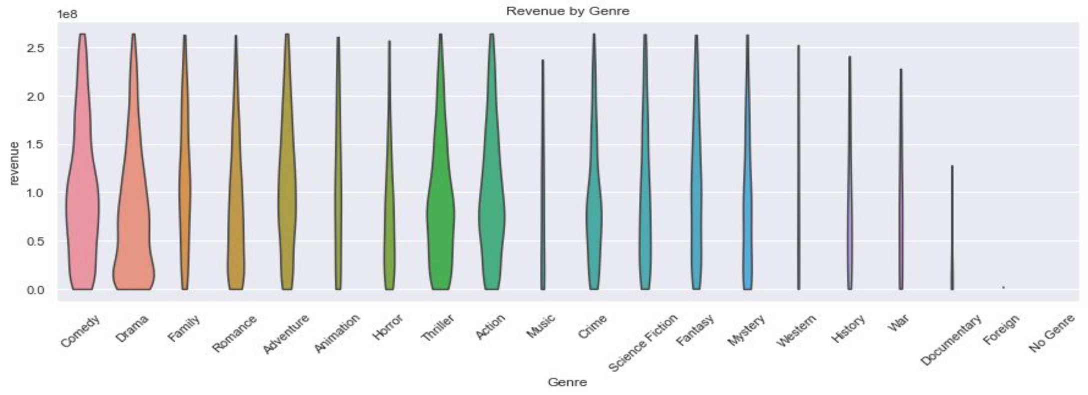
Extended Modeling: Reducing Dimensionality

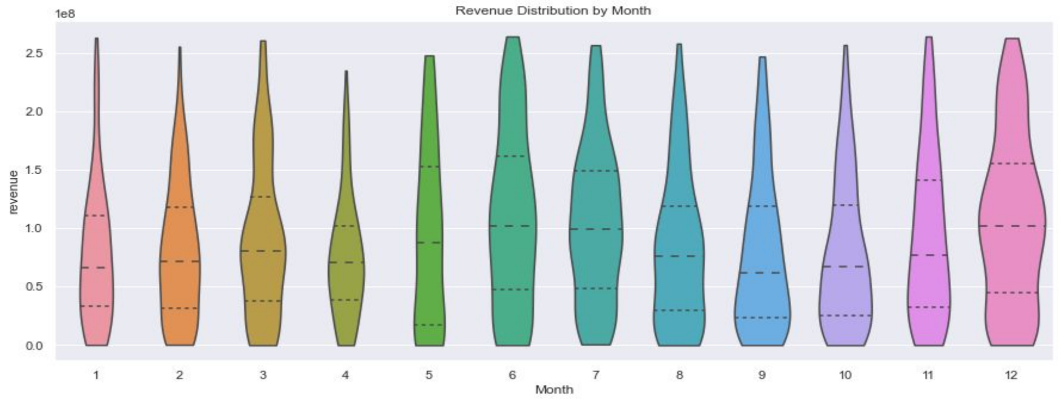




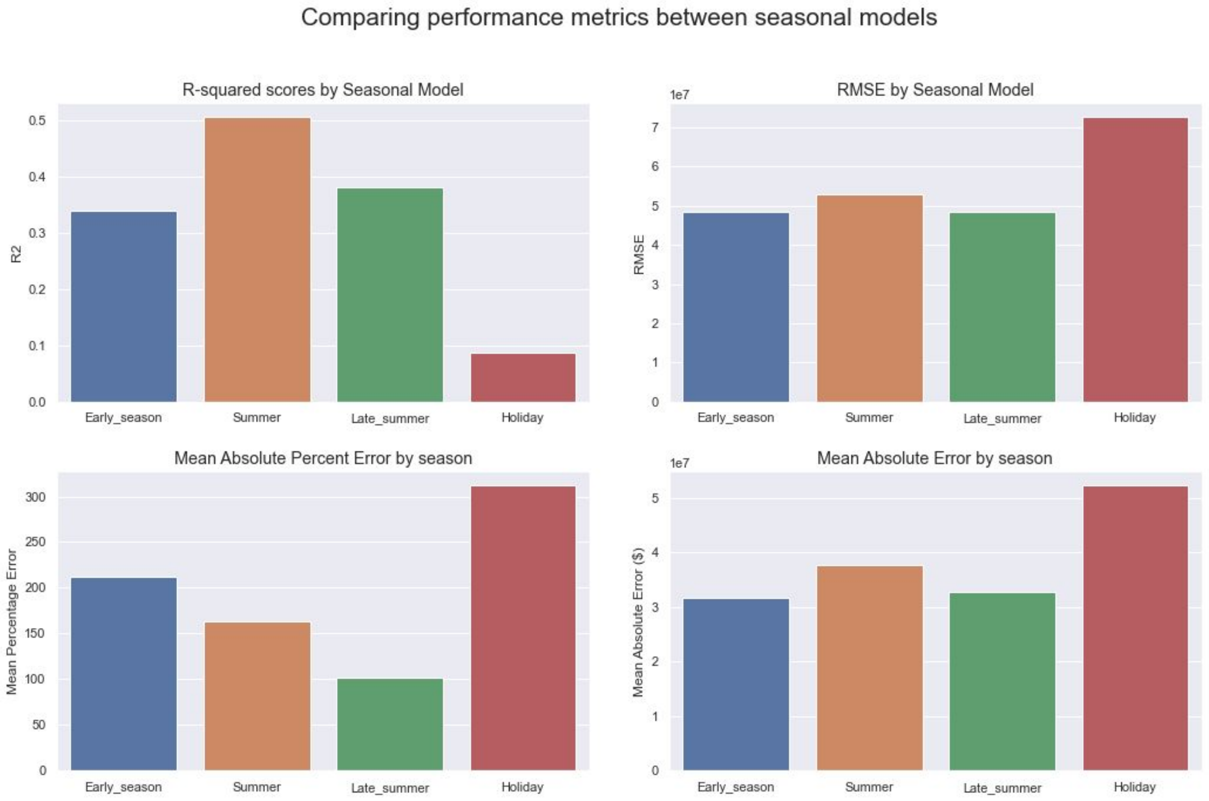


Extended Modeling: Seasonal Approach

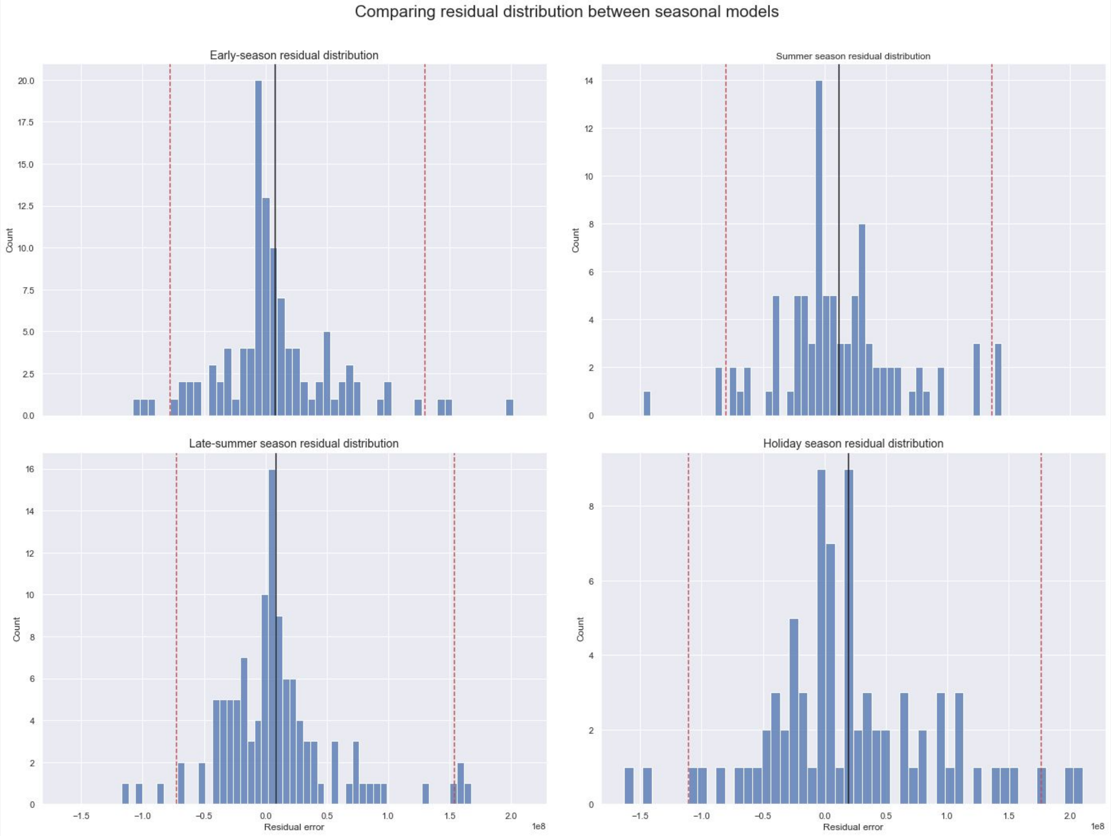




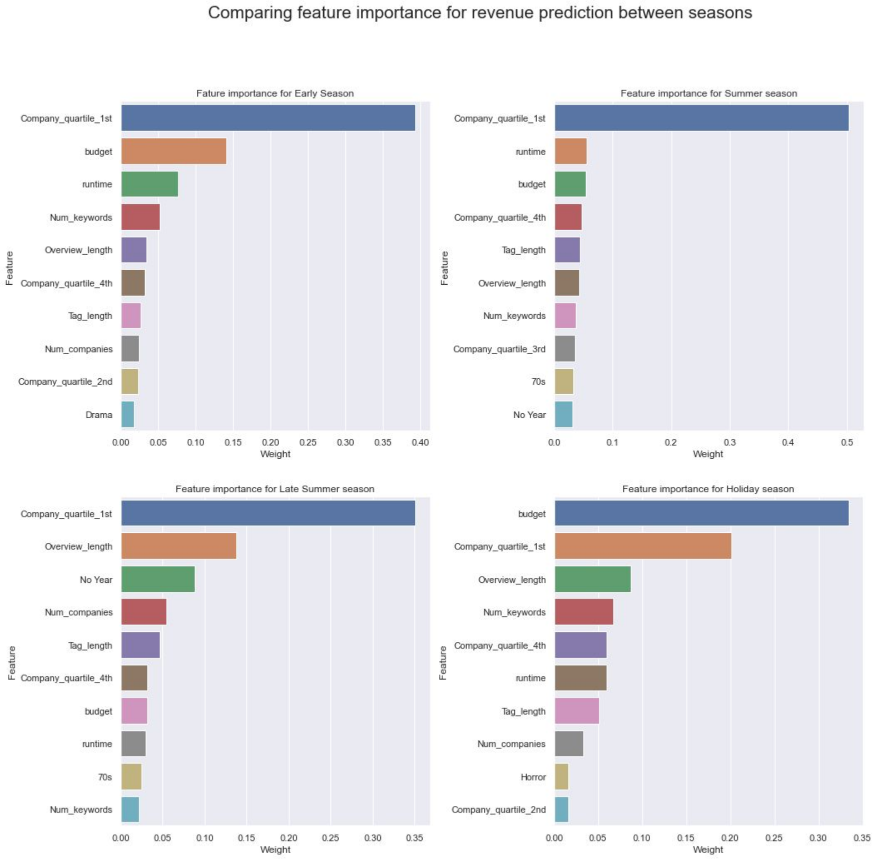
Extended Modeling: Seasonal Approach - Results



Extended Modeling: Seasonal Approach - Residuals



Extended Modeling: Seasonal Approach - Features



Comparing feature importance for revenue prediction between seasons

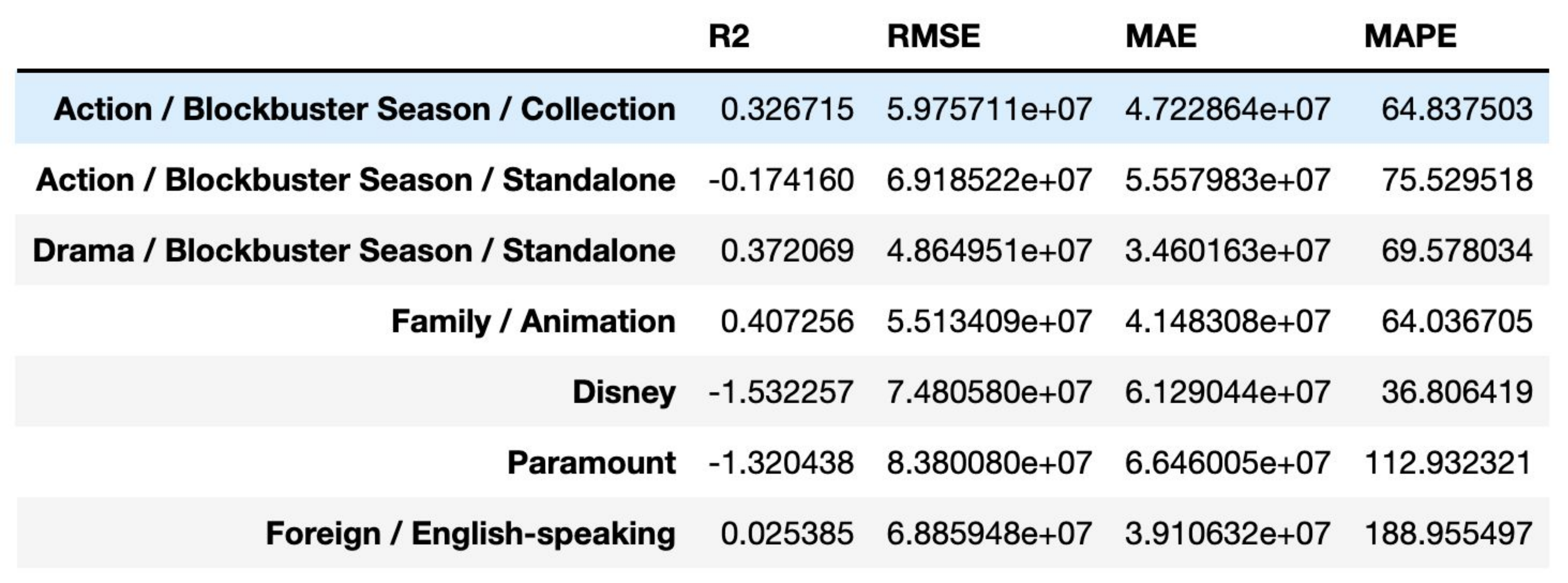
Fature importance for Early Season

Feature importance for Summer season

Feature importance for Late Summer season

Feature importance for Holiday season

Extended Modeling: Narrow Tiers

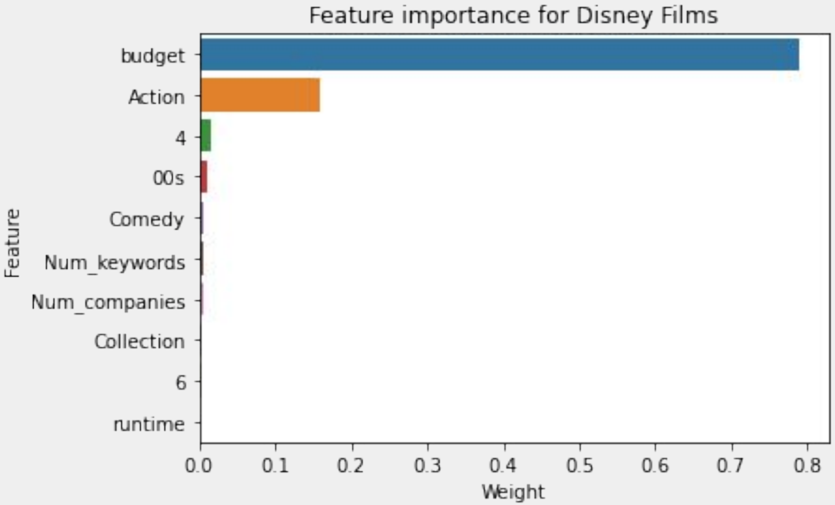


	R2	RMSE	MAE	MAPE
Action / Blockbuster Season / Collection	0.326715	5.975711e+07	4.722864e+07	64.837503
Action / Blockbuster Season / Standalone	-0.174160	6.918522e+07	5.557983e+07	75.529518
Drama / Blockbuster Season / Standalone	0.372069	4.864951e+07	3.460163e+07	69.578034
Family / Animation	0.407256	5.513409e+07	4.148308e+07	64.036705
Disney	-1.532257	7.480580e+07	6.129044e+07	36.806419
Paramount	-1.320438	8.380080e+07	6.646005e+07	112.932321
Foreign / English-speaking	0.025385	6.885948e+07	3.910632e+07	188.955497

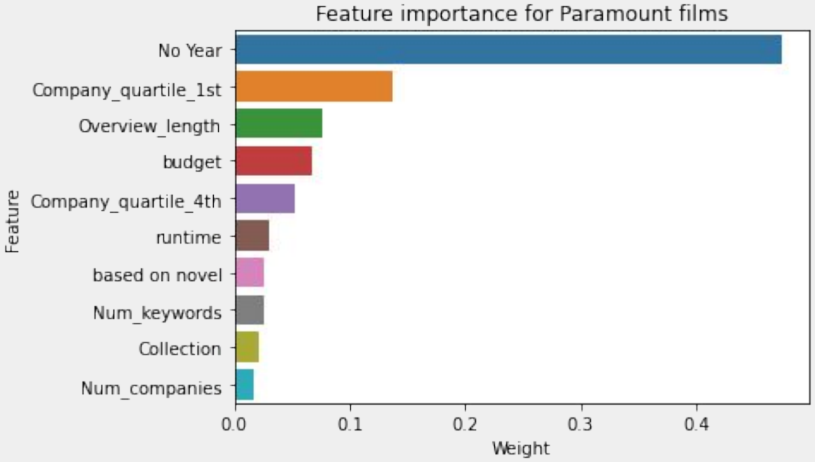
Conclusion:

1. Multi-Tiered approach is the strongest
2. Need to revisit data acquisition & extraction
3. Revenue numbers need to be checked for validity
4. Clustering as a means of grouping films
5. Danger of treating ML as an ‘Oracle’

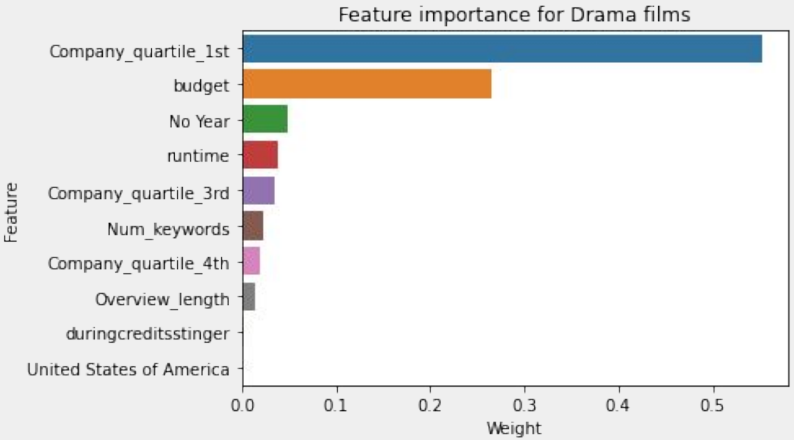
Questions?



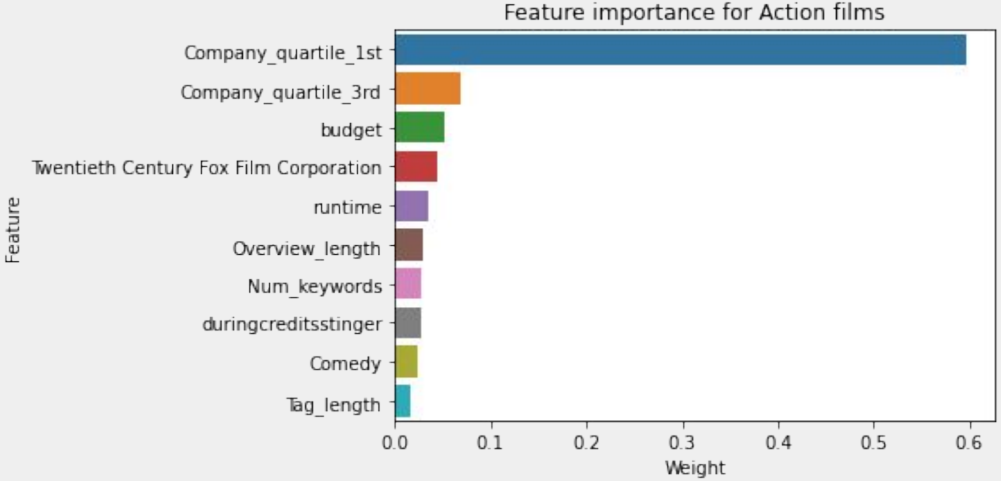
Feature importance for Disney Films



Feature importance for Paramount films

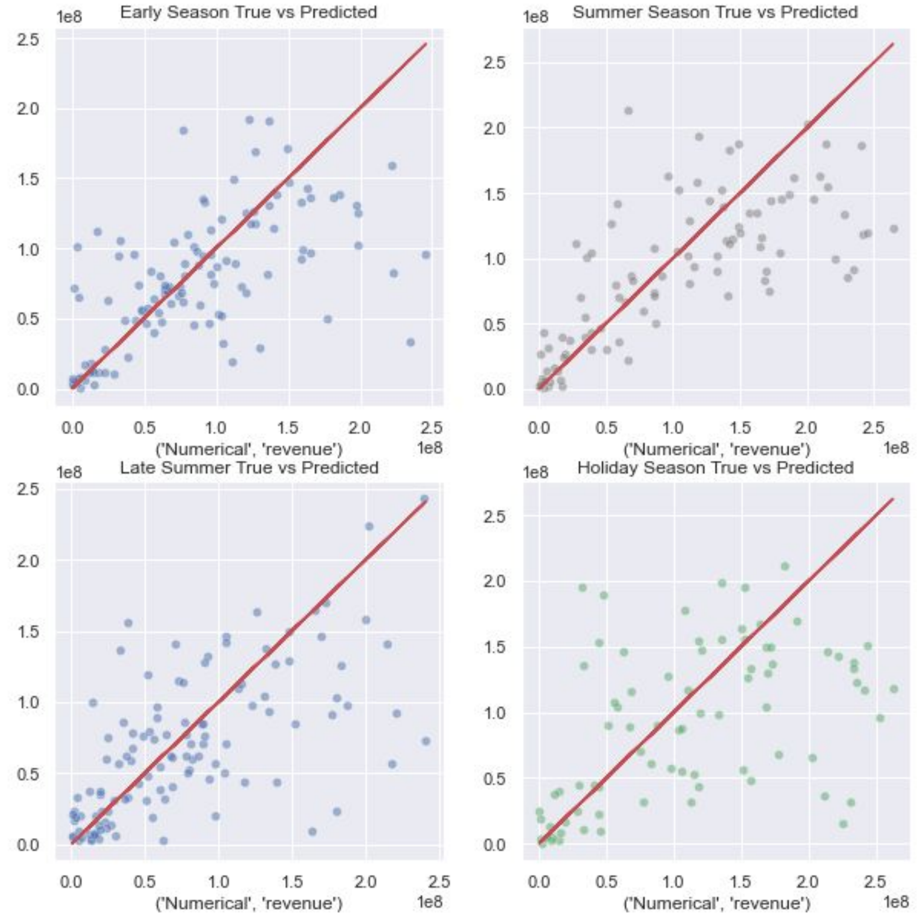


Feature importance for Drama films



Feature importance for Action films

Comparing relationship of True and Predicted values between seasonal models


Early Season True vs Predicted
1e8
('Numerical', 'revenue')
1e8
Summer Season True vs Predicted
1e8
('Numerical', 'revenue')
1e8
Late Summer True vs Predicted
1e8
('Numerical', 'revenue')
1e8
Holiday Season True vs Predicted
1e8
('Numerical', 'revenue')
1e8