

# Data Science Meets Hollywood

Jit Seneviratne

# Predicting Revenue for Movies

Can Cumulative Worldwide Box Office Revenue be Predicted from Available Quantitative and Qualitative Information?

# Pipeline for Data

Budget

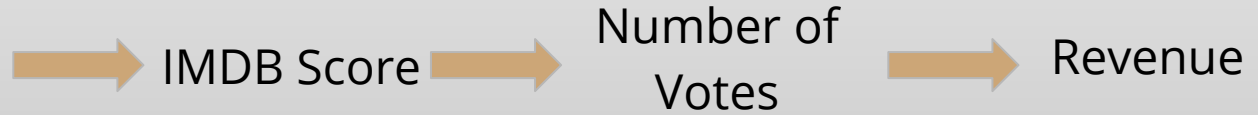
Actor

Actress

Director

Genre

Runtime



# A Few Assumptions

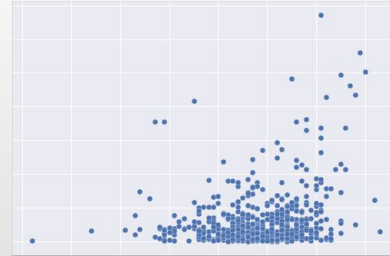
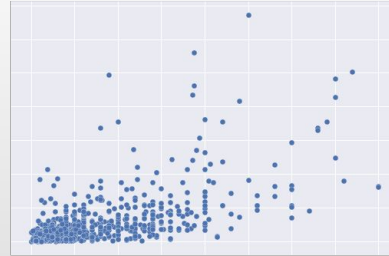
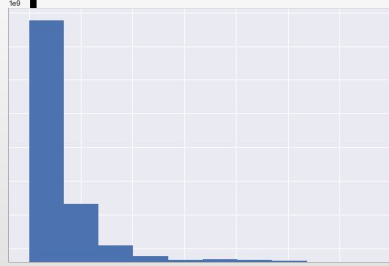
- Star power has the same weight as quality
- Time value of money applies to both cost and revenue streams

## Process

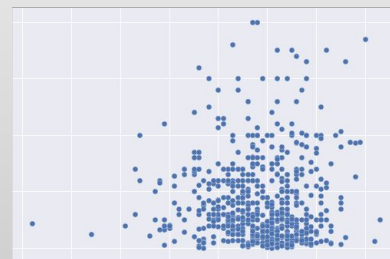
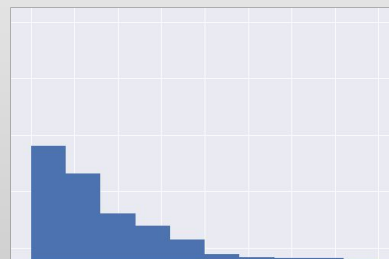
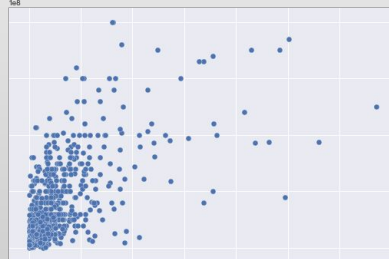
- Scrape Data from IMDB 2000 - 2016 (600+ movies)
- Compare with award winners (Wikipedia)
- Clean Data
- Explore Data
- Model Data
- Test Data
- Refine Process

# EDA - Pairplot

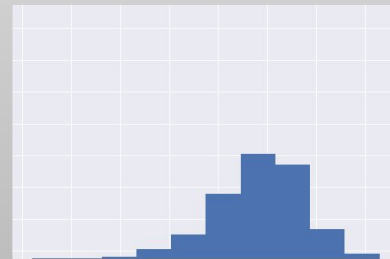
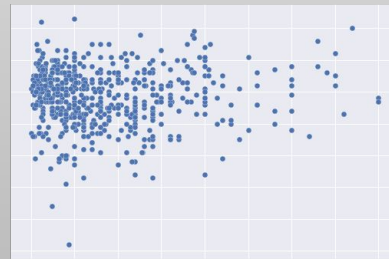
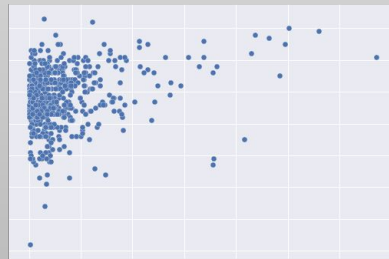
Worldwide  
Income



Budget



IMDB  
Score



Worldwide  
Income

Budget

IMDB  
Score

# First Fit



# Findings

Reduce ROI to 15

$R^2 : 0.3516$

For every \$ change in budget, expect an increase of \$2.5938 in worldwide revenue

Budget with Categorical Variables = 0.44872015523336373

Adding Interactions with Budget = 0.48664962632436037

Adding IMDB Rating = 0.51965927757669672

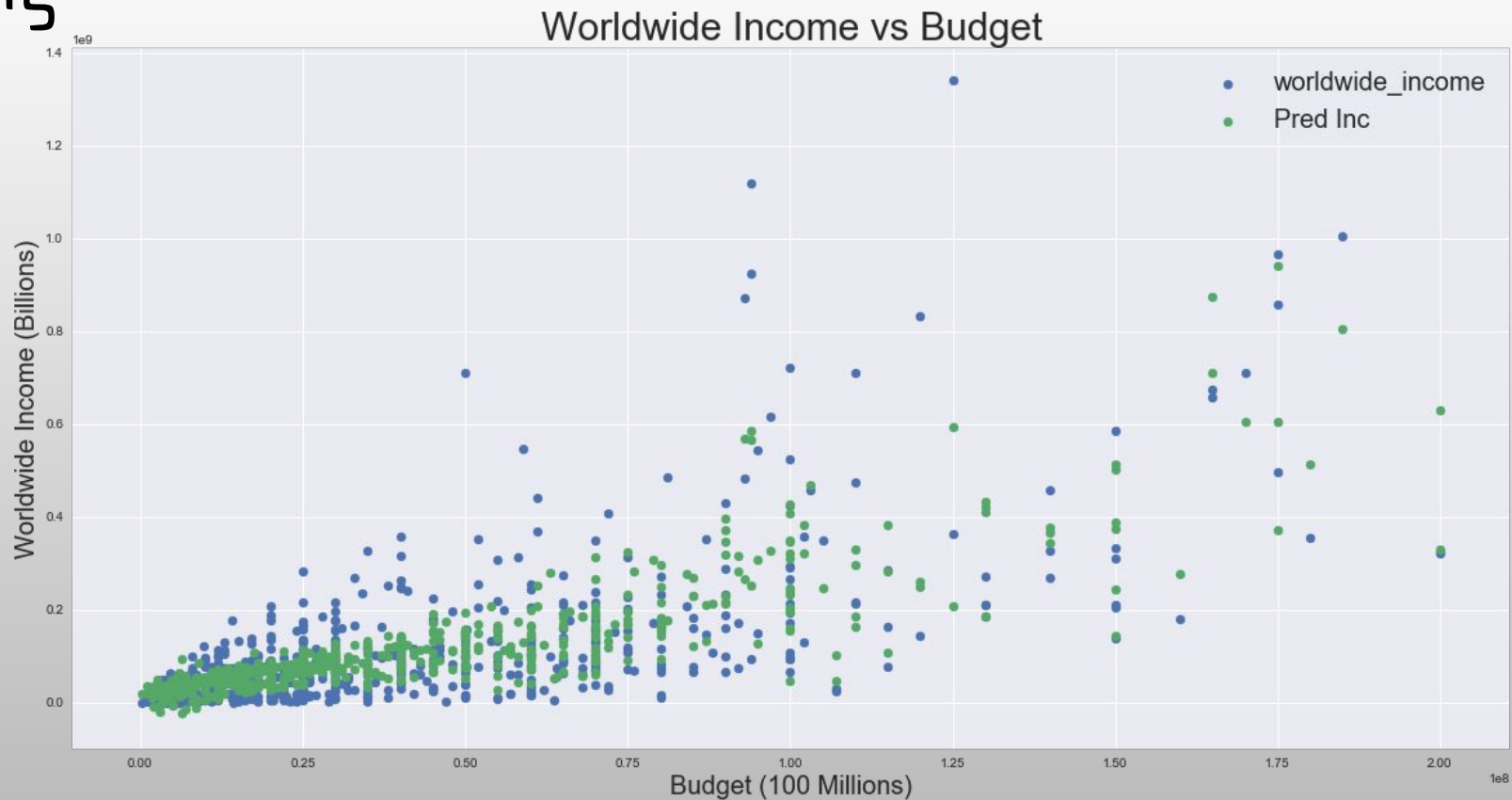
Adding Interaction with Budget = 0.57561391212225865

Adding Number of Voted Users = 0.6303657583315645

Adding Interaction with Budget = 0.64180328572789025

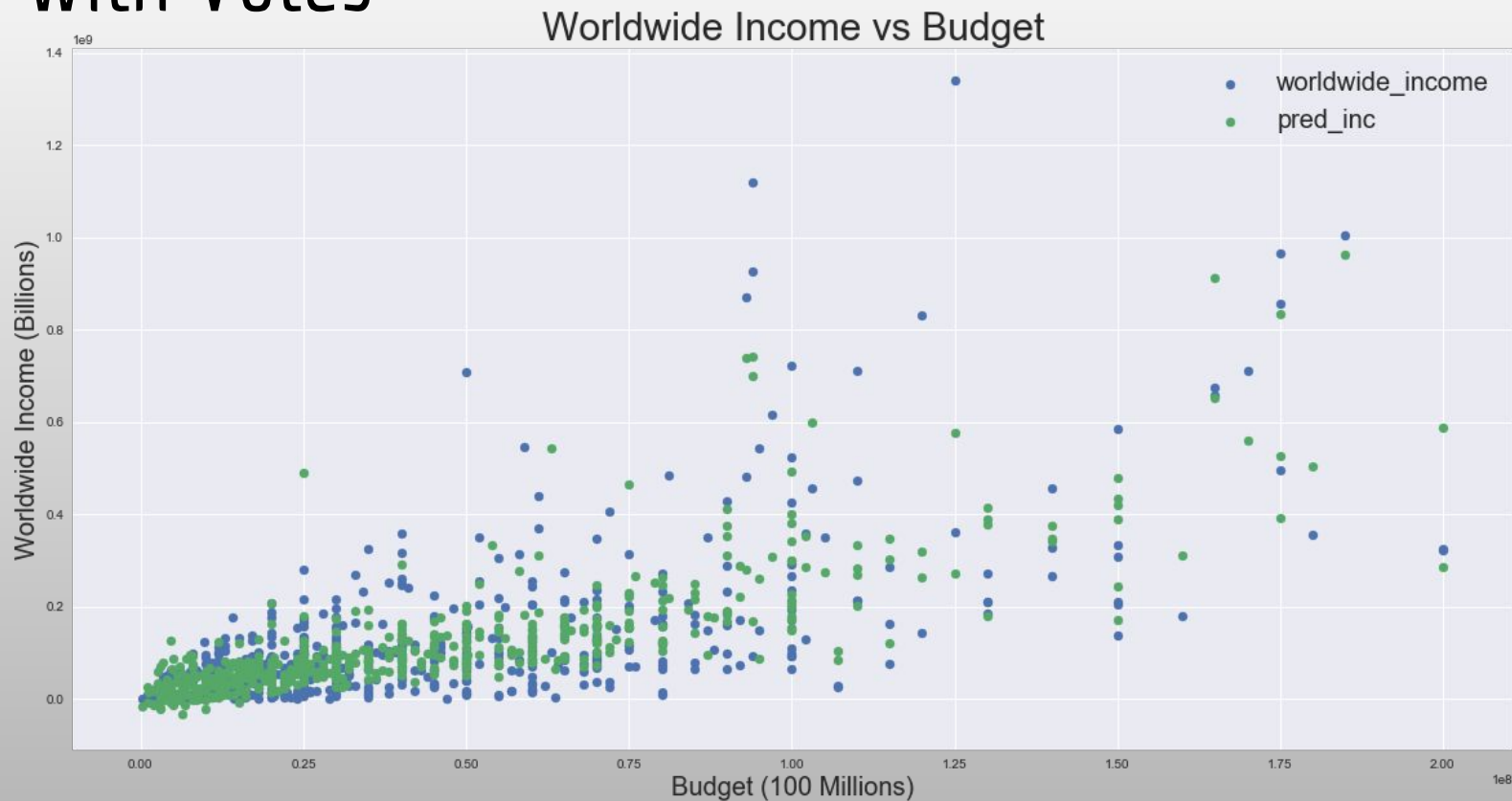
Runtime and other interactions failed to add to the relationship

# Working Model

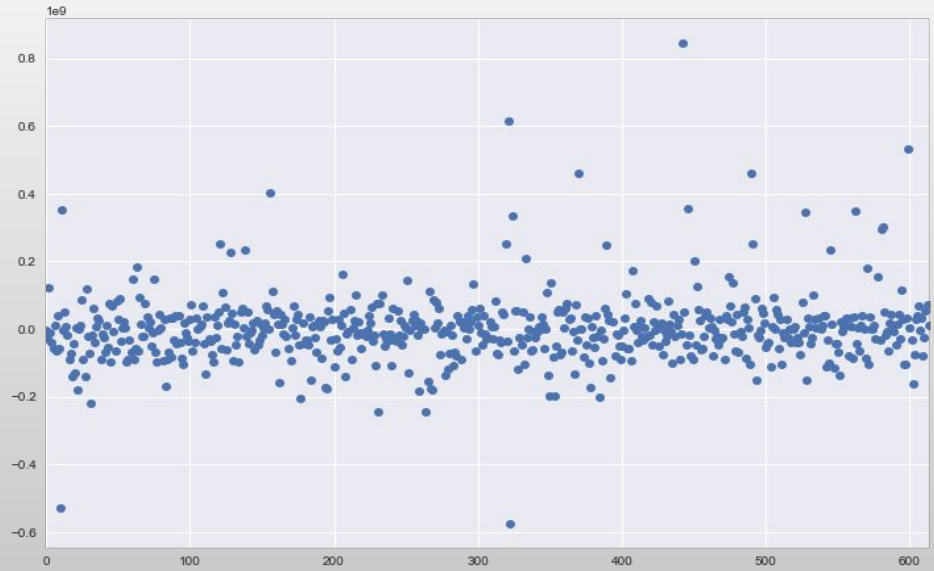
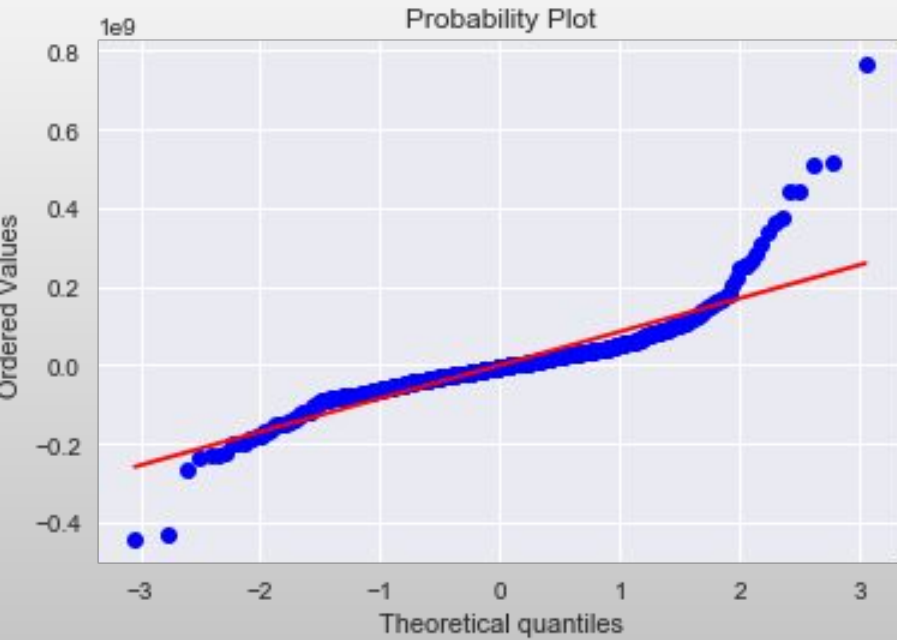




# Model with Votes



# Residuals



# Cross Validation

## Test Scores

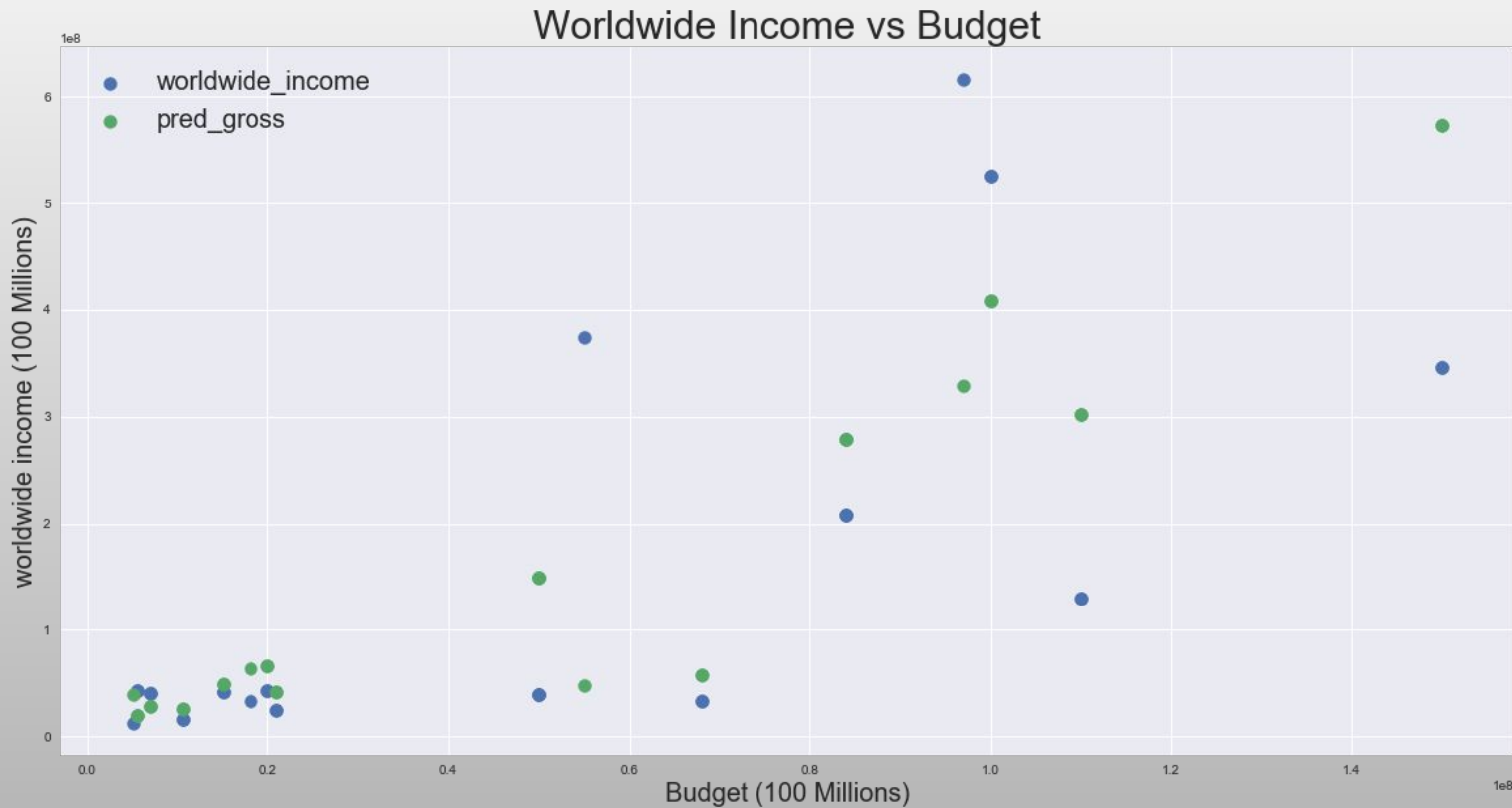
0.59792395415386079  
0.453437645897113  
0.56470464560383116  
0.51244472761868487  
0.47958333322796765  
0.56888237343495507  
0.48610167044768537  
0.4586023935692225  
0.57645715442380774  
0.64381391392379239

## Train Scores

0.55607469808315946  
0.59228767356028178  
0.55844066112017643  
0.59625812880068552  
0.5907264705868509  
0.53059436521161873  
0.59253876703585551  
0.59885027755911713  
0.55212605622897781  
0.52235729789987251

# Testing on Unseen Data

$R^2 = 0.55$



## Features

budget	budget&actor_nominee
director_nominee	budget&actress_nominee
actor_nominee	budget&comedy
actress_nominee	budget&biography
comedy	budget&romance
biography	budget&sport
romance	budget&thriller
sport	budget&crime
thriller	budget&adventure
crime	imdb_score
Adventure	budget&imdb
budget&director_nominee	

# Feature Ranking

Budget

Budget & IMDB

Budget & Actor Nominee

Budget & Adventure

IMDB Score

Budget & Sport

Budget and Thriller

Budget & Director Nominee

Budget & Actress Nominee

Budget & Biography

Remaining...

# Next Steps:

- Gather more samples!
- Split Personnel Achievements (Between and After)

Thank you