Predicting Movies' Revenue

Presentation

Ali Naji
Feb 2019

Executive Summary

- Movie’s gross/revenue is dependent on several parameters
- Strong evidence that movie revenue is strongly correlated with its allocated budget
- Difficult to precisely predict movie revenue with categorical information

Overview

- Movie Industry
- Introduction to the Problem
- Data Analysis
- In-Depth Analysis
- Results
- Next Steps

Problem Statement

- 80% of Hollywood movies fail to turn profit
- Movie companies have data about past movies and their revenues/success
- Too hard for humans to decide if a movie will be successful
- Too many parameters to account for

Background

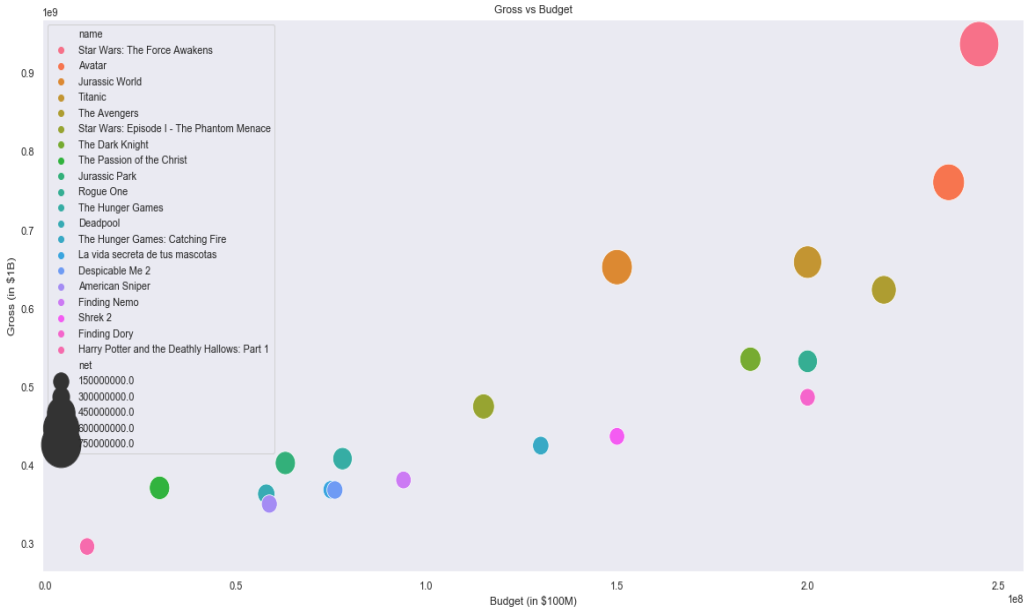
Most successful movies tend to be:

- Action genre
- Released in the US
- PG-13 rating

But can we generalize based on those specs?

Not really! Data show that worst

performing movies had similar specs

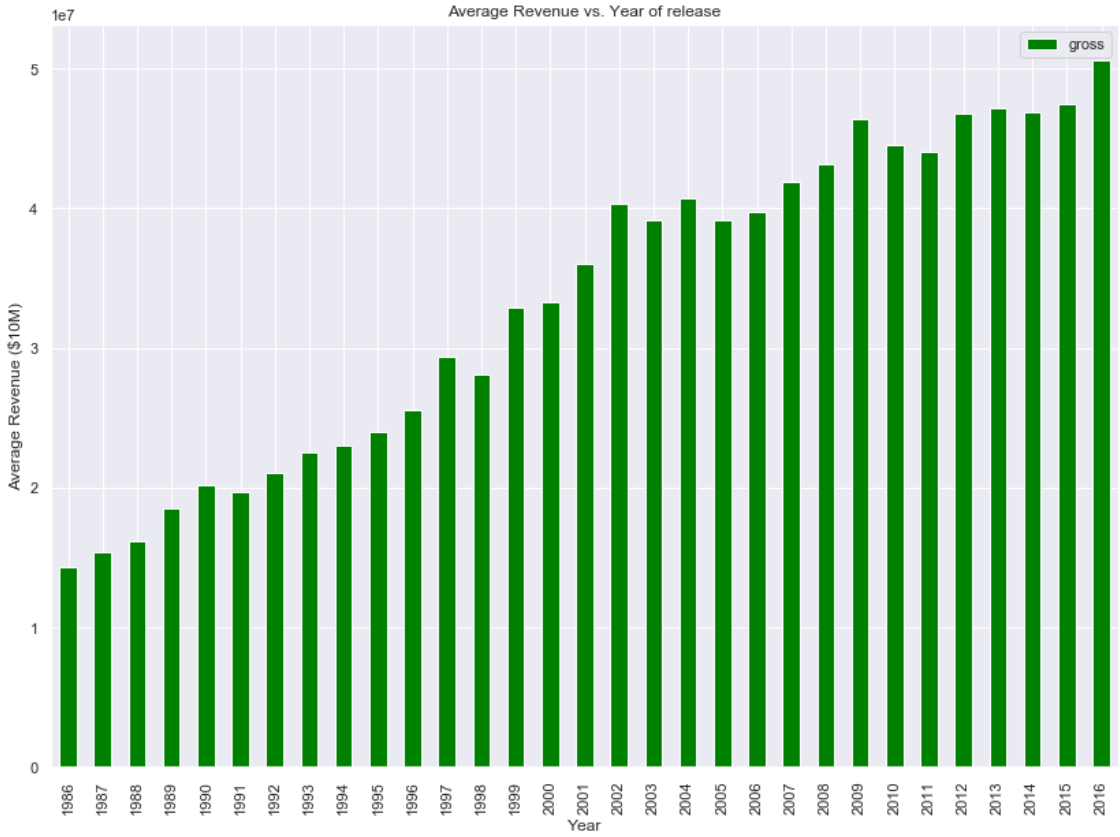


Project Goal

- Give movie-makers an idea about movie's success before its made
- Predict gross/revenue based on movies specs
- Create ML models to predict future movies' success
- Recommend movie makers with potential next steps

Data Analysis

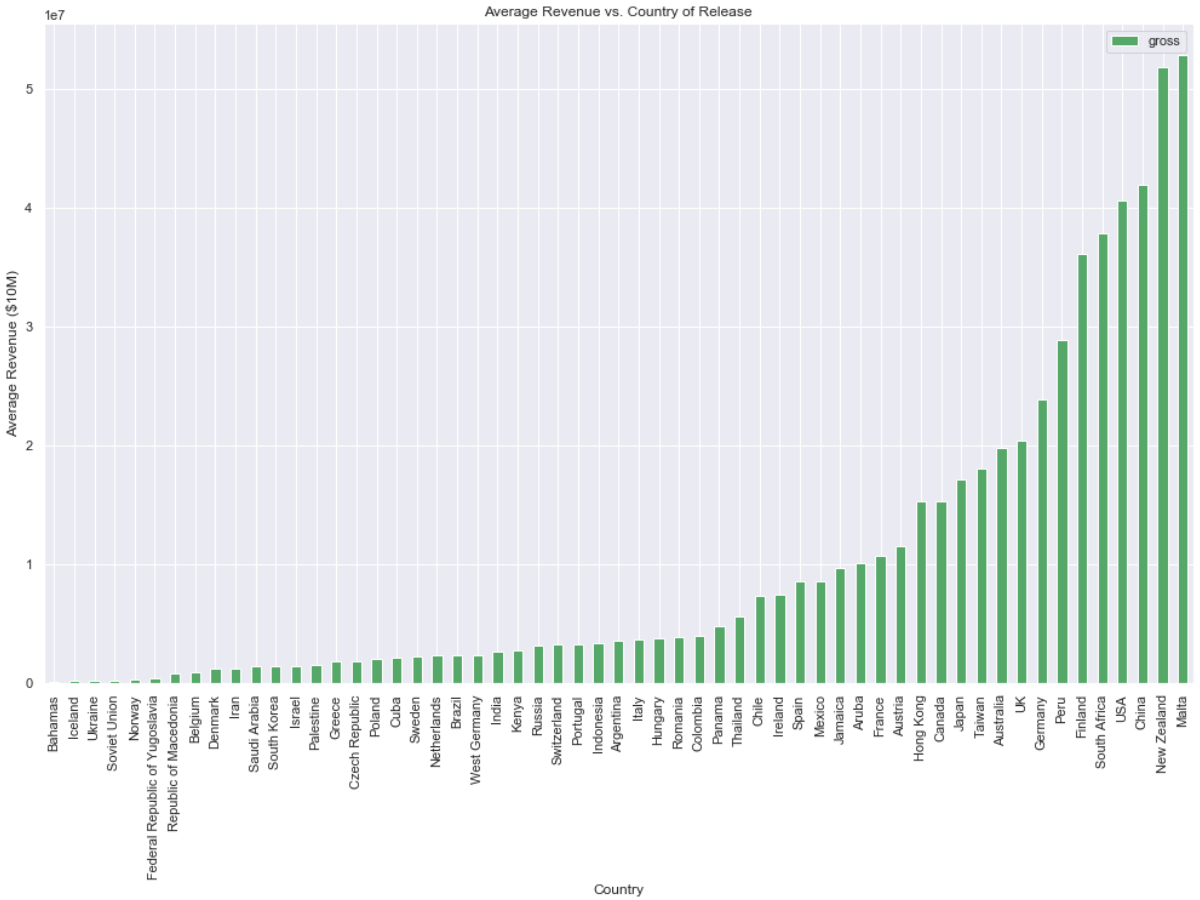
Movies' average revenue over time



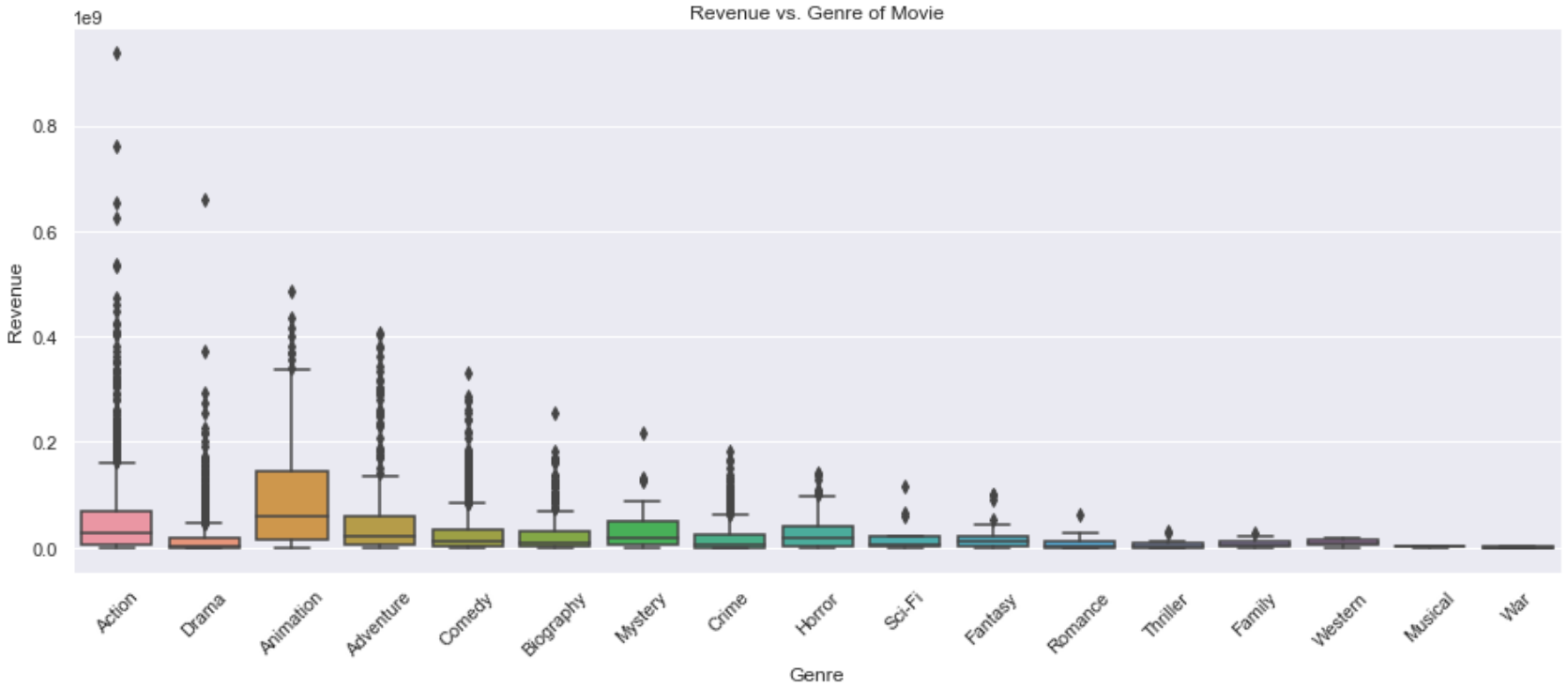
Data Analysis

Movies average revenue by country of release:

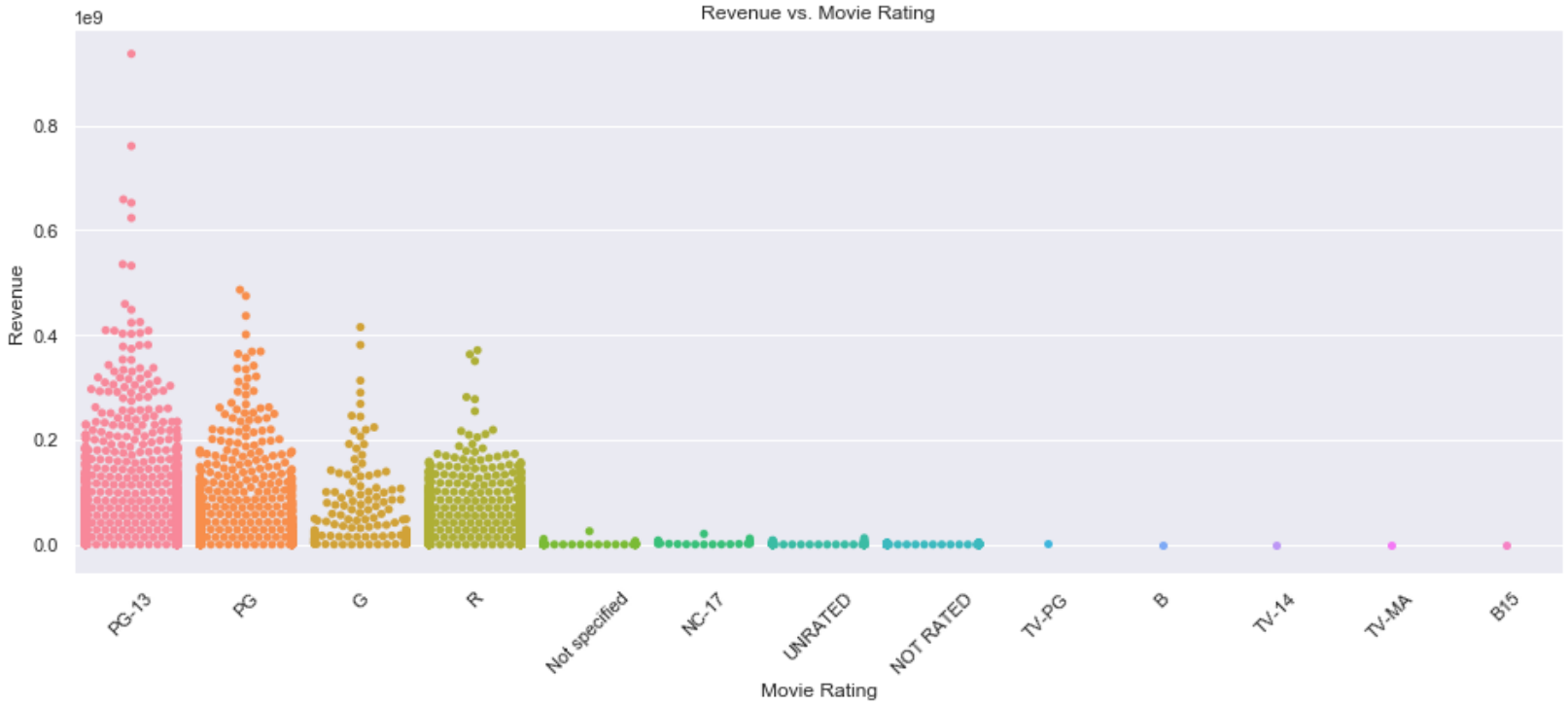
- Malta
- New Zealand
- China
- USA



Genre vs Revenue



Rating vs Revenue



Variable Correlations

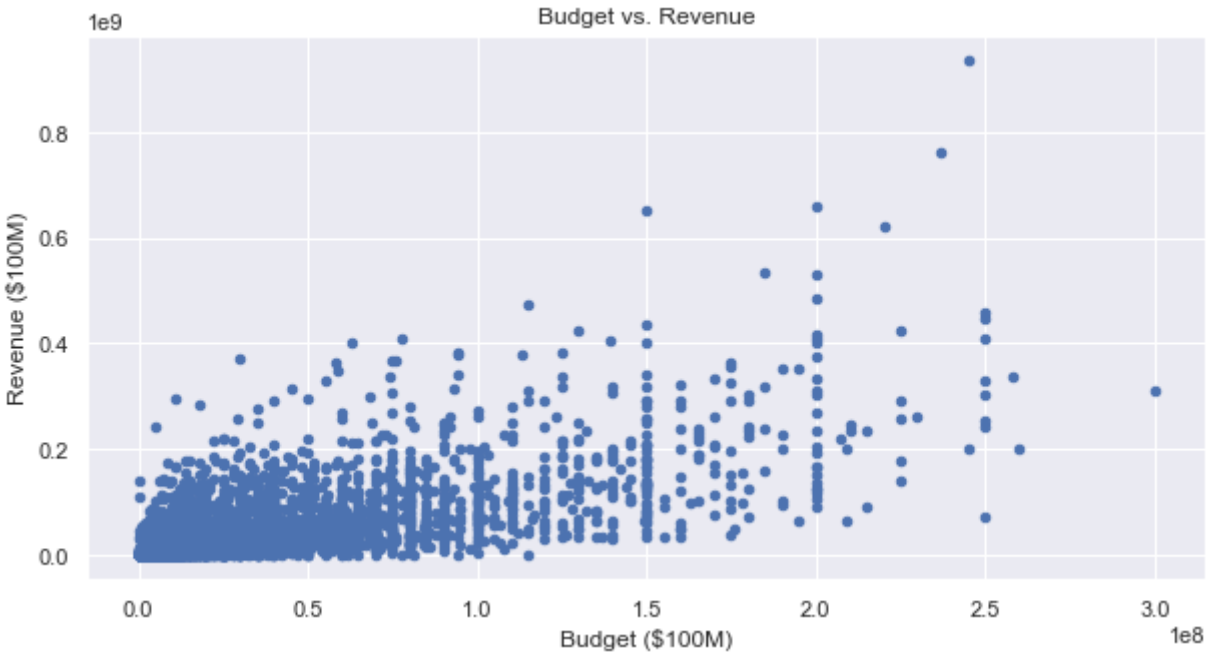
Numerical features compared in correlation matrix map

- Gross shows strong correlation with budget (0.71)

Budget vs. Gross

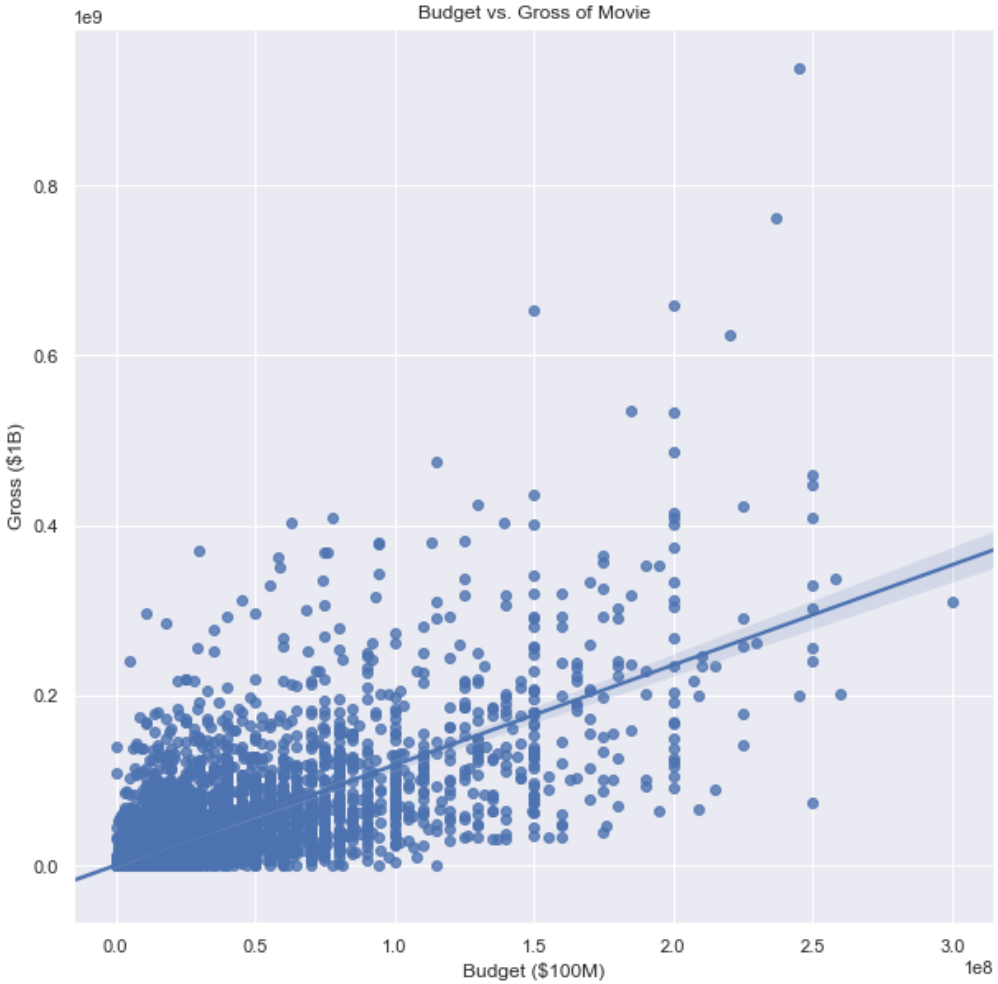
- Scatter plot between budget and gross shows significant correlation

- Hypothesis test backs up this assumption



In-Depth Analysis

- Created linear regression model
- RMSE = $34M
- Will try other models to account for other variables in the dataset, and to improve accuracy



In-Depth Analysis - Contd.

Other regression models that will be used are:

- Random Forest
- Adaptive Boosting (AdaBoost)
- Gradient Boosting
- K-Nearest Neighbor
- Ridge Regression (Linear regression with L2 regularization)

In-Depth Analysis - Contd.

Models accuracy results:

- Gradient Boosting best model with 76% test accuracy
- Random forest/AdaBoost overfitting training set
- Ridge and KNN models underfitting dataset

Model	Train_Score	Test_Score
Gradient Boosting	84.3%	76.53%
Random Forest	96.16%	75.46%
AdaBoost	99.93%	74.44%
Ridge	63.85%	65.74%
KNN	77.01%	55.56%

Results

After tuning the Gradient Boosting model:

- Test accuracy of 78.5%
- RMSE of about $27M
- Meaning, for every revenue prediction, there is, on average, $27M tolerance about true revenue.
- Very hard to improve prediction accuracy based on categorical info like genre, rating, etc.
- Big tolerance but very helpful to movie makers given high budget movies

Next Steps/ Recommendations

Movie companies need to study the following extra details about movies to improve their gross prediction accuracy:

- Movie trends and changing taste with time
- Historical context of movie e.g. remakes, scripts
- Audience the movie is targeting e.g. kids, elders, general public, etc..

Having those in ML model would help predict revenue more accurately