**Of Movies and Metrics**

by

Jayashri V Jagannathan

# Introduction

One of the interesting aspects of data scientists’ jobs is how they can paint interesting stories and narratives from data. Data science can deliver new insights even in domains people are very familiar with, and it can be powerful tool for questioning conventional wisdom. For my project, I picked an area utmost everyone knows and have strong opinions about: the Hollywood movie industry. Specifically, I explore the following in this report:

* The relationship between genre, budget, revenue, ratings of movies.
* A methodology by which studios can decide how much money to spend on marketing and promotion after a move is already made
* Analyze the impact of running length of movies, which have been getting increasingly longer to meet requirements of an international audience.

To deliver new insights from existing data, data scientists use a variety of techniques. These include:

* Build new models and analysis
* Segment data and explore relationships between segments
* Combine multiple data sets
* Combine datasets with external sources of information

I use several of these techniques along with Supervised Machine Learning in my Hollywood movie metrics project.

# Datasets

For my analysis on movies, I found two datasets –

**Movies Metadata dataset**

https://data.world/popculture/imdb-5000-movie-dataset/workspace/file?filename=movie\_metadata.csv%2Fmovie\_metadata.csv

File \_metadata.csv

**Movies R Data Set from Duke University**

<http://www2.stat.duke.edu/~mc301/data/movies.html>

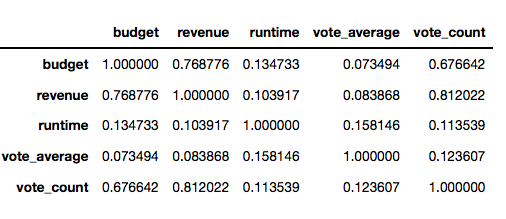
File : movies.RData ( Saved as movies.csv)

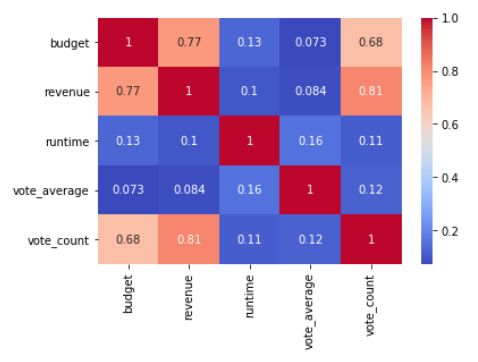
I have used python 3 and the following libraries in python for this analysis in jupyter notebook – numpy, pandas, seaborn, matplotlib, scikit-learn.

# 3. Revenue prediction before movie release

As an example, “Solo: A Star Wars Story”, released in May 2018, cost Disney $250m to make and another $250 in marketing and distribution. The movie is considered a box office failure, cost Disney significant losses, and future Star Wars anthology spin offs have been put on hold. Could this have been predicted based on early critic and fan reactions?

To make predictions about revenue, we need to understand the relationship between movie genre, budget, revenue and critics’ ratings. With that in mid, I first studied the correlation between the numerical columns of the Movies Metadata.

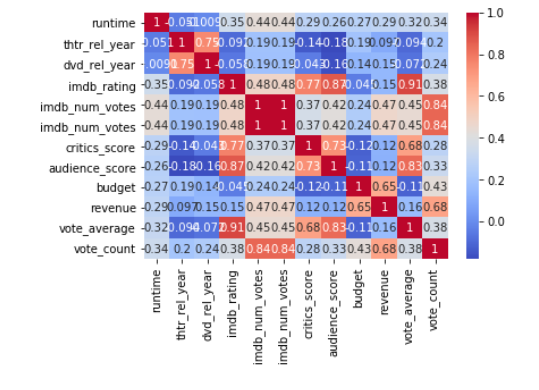
**Correlation Matrix**

**Heatmap of the correlation matrix**

My observation was that revenue was highly positively correlated with budget and vote count. The fact that high budget movies like Star Wars attract crowds is no surprise and also that people want to vote and express their opinion on popular movies.

I merged the two datasets that I have based on title column.

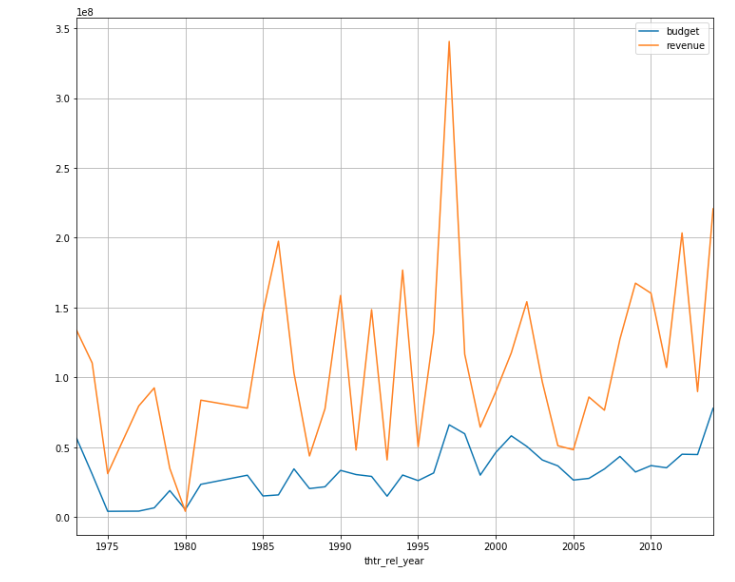
**Correlation Heat Map of the Merged Dataset**

I was looking at factors that affect revenue and found budget and vote\_count are strongly positively correlated to revenue. Other factors such as runtime (i.e. the length of the movie), thtr\_rel\_year (theatre release year, imdb\_rating (Internet Movie Database (IMDB) Rating), imdb\_num\_votes (Number of Votes in IMDB database for the movie), critics\_score, audience\_score and vote\_average.

## Revenue and Budget Relationship

In my previous section, I observed that the budget was the biggest factor impacting bottom line (i.e. revenue). So I wanted to study the role played by budget in influencing revenue and looked at the Time Series plot over the last 50 years of average revenue and budget for movies in the datasets.

**Distribution plot of budget and revenue**



There is definitely an increasing trend in budget accounting for inflation as well as special effects and graphics in more recent films that have contributed to increasing costs [2].

## 3.2 How much money should a studio spend on Marketing & Promotion?

Due to Globalization, Hollywood actors and actresses are world celebrities since US movies are watched in different countries and have international audience. Marketing costs play a major role in total budget and can be almost as high as 50% of the cost of producing the movie [3] though marketing costs are incurred post production. The distribution costs incurred after production depend on number of theatres in which the film is shown and whether the film has international audience or not. Given that a movie has already been produced and production costs have been incurred, if there is a crystal ball to view the future and see how successful the movie will be, then the studios (i.e. decision makers) can decide how much they would like to spend on marketing and distribution costs. If the movie is going to be a flop and fail miserably at the box office, then the studios would like to cut their losses by not spending as much in marketing and distribution costs.

The surest way to assess the popularity of a movie is to be able to get a pulse of the audience reaction and rating. However, before a movie is released there is no way to get the audience rating. Looking at the correlation matrix of the Merged Dataset, there is a strong positive correlation between the ratings given by the critics (critics\_score) and the audience rating (audience\_score and imdb\_rating).

The critics\_score is the one that I favor, since pre-release, audience\_score and imdb\_score are not available. Studios have private early screening of movies to critics. The feedback from these critics, assuming that it is a statistically significant sample, can be used to better model and predict expected revenue.

### 3.2.1 Linear Regression Model for Predicting Revenue

For the independent variables, I chose the following:

(1) budget since it is highly correlated with revenue

(2) critics rating (critics\_score) since that is the only one that can be available prior to release although imdb\_rating has higher correlation.

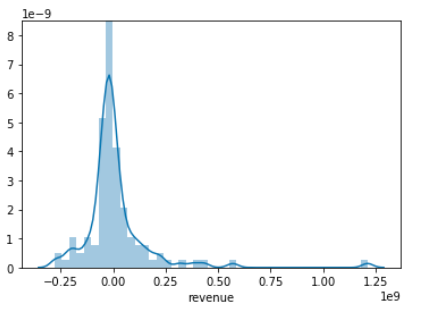
I used supervised learning with scikit-learn in python and constructed a Linear Regression Model with training set to be 40% of the data.

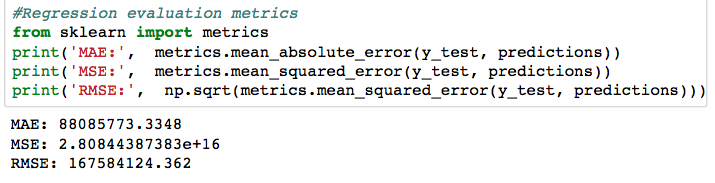
The Linear Regression model is

**revenue = -64681126.5242 + 114185 \* critics\_score + 2.99620305 \* budget**

**Explained variance, i.e. R\*\*2 = 0.41596949082**

**Distribution Plot of the residuals** (i.e. actual value of revenue – its predicted value)

This graph appears to be normally distributed though there are some outliers to the right. Then I calculated metrics for the model -- Mean Absolute Error (MAE), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE).

 **Explained variance, i.e. R\*\*2 = 0.41596949082**

The MSE seems to be high and so I chose to do a log transformation on the variables since revenue and budget are huge in millions of dollars.

### 3.2.2 Linear Regression Model for predicting Revenue with log transformation

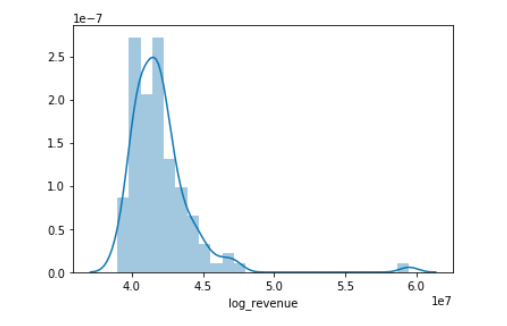
I did a log transformations on budget (log\_budget) and on revenue (log\_revenue) and used supervised learning with scikit-learn in python and constructed a Linear Regression Model with training set to be 40% of the data.

The Linear Regression model is

**log\_revenue = -0.970155522922 + 0.015580 \* critics\_score**

**+ 1.04553242 \* log\_budget**

**Scatter Plot of the Residuals**



The residuals appears to be skewed to the right and not normally distributed with a mean of 0. The metrics for this model are given below.

Looking at the metrics, the log transformation model seems to be better.

#Regression evaluation metrics from sklearn import metrics

print('MAE:', metrics.mean\_absolute\_error(y\_test, predictions))

print('MSE:', metrics.mean\_squared\_error(y\_test, predictions))

print('RMSE:', np.sqrt(metrics.mean\_squared\_error(y\_test, predictions)))

print('Explained Variance: ', metrics.explained\_variance\_score(y\_test,predictions))

MAE: 1.09495932899

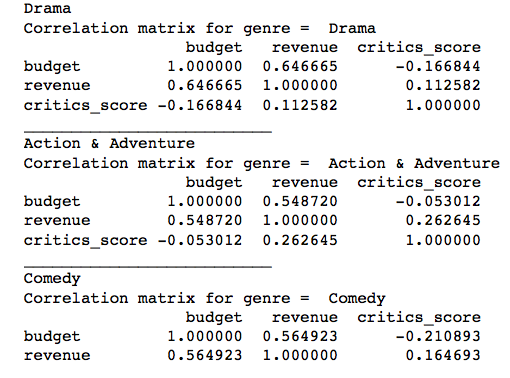
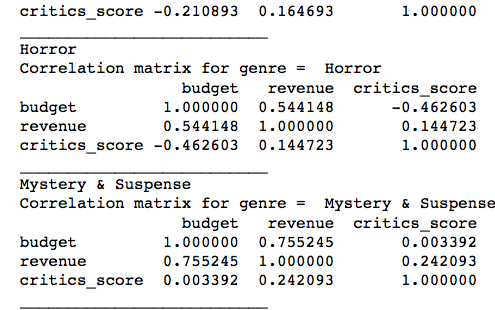
MSE: 2.18462307883

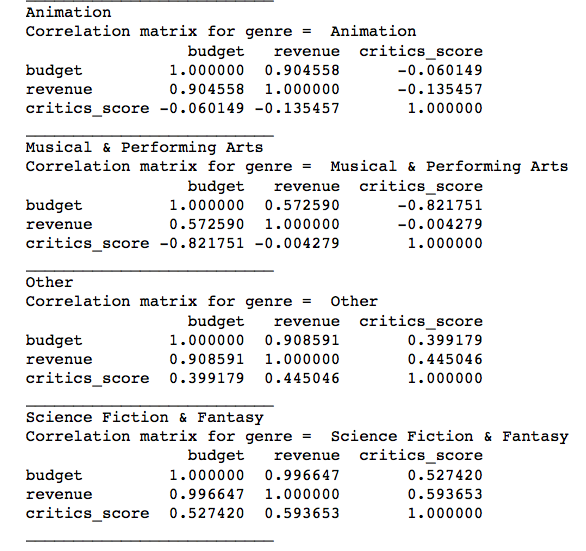
RMSE: 1.47804704892

Explained Variance: 0.551918282513

### 3.2.3 Correlation matrix of revenue, budget and critics\_score across genre

I examined the relevancy of the critics\_score across genre. For this analysis, I looked at the correlation matrix of revenue, budget and critics\_score for data grouped by genre.



My observation is that a genre like “Science Fiction & Fantasy” has a higher correlation coefficient between revenue and critics\_score than any other genre (close to 0.6 which is fairly strong positive correlation).

Based on this, if critics give a low rating for a movie belong to the “Science Fiction & Fantasy” genre, the studios (i.e. decision makers) should proceed cautiously in investing their money in marketing and distribution costs.

# 4. Movie Runtime in the recent decades

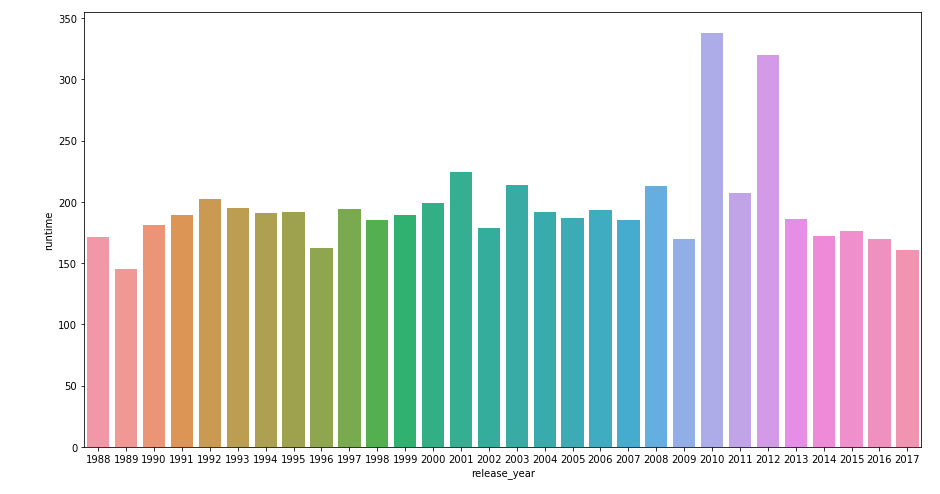
**Runtime of highest revenue generating movie by year is increasing**

In my movie going experience, I have felt that popular box office hits seem to be very long. I found that this was not just my imagination but it is in fact true [4].

One of the reasons for the increase in runtime of movies is the expectation of international audiences.

I took highest box office revenue generating movie each year for the last 30 years and examined the runtime of these movies.

**Bar Plot of Runtime for the movie with the highest revenue for every year**

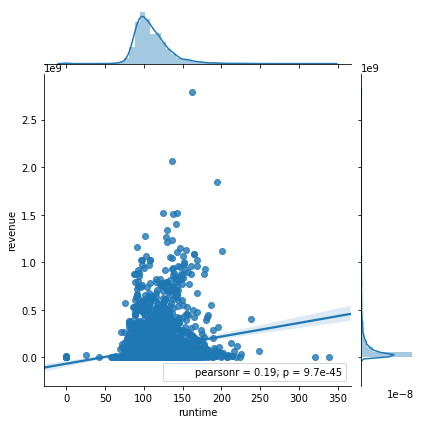


Previously, movies used to be about a couple of hours long (i.e. around 120 minutes to 150 minutes). Now we definitely see longer movies making it to the popular lists.

## 4.1 Correlation between Runtime and Revenue

In the correlation matrix and in the heatmaps, I had noticed that there was a weak positive correlation between runtime and revenue. I did a scatter plot of revenue versus runtime as given below.

**Scatter Plot of revenue versus runtime**



I noticed that there does not seem to be any relationship between revenue and runtime.

# 5. Conclusion

In this report, I have investigated the correlation between revenue and other numerical columns such as budget, critics\_score, audience\_rating, imdb\_rating and runtime, and also correlation matrix of revenue and budget across genre.

I have proposed two Linear Regression models of revenue on budget and critics\_score using Supervised Machine Learning with scikit-learn.

For further research, I intend to propose a Linear Regression for each genre to predict revenue based on budget and critics\_score. Also, I intend to investigate the impact on YouTube views of early trailer videos and social media buzz.

# References

1. <https://www.statista.com/statistics/187069/north-american-box-office-gross-revenue-since-1980/>

2. <https://www.wired.com/2014/09/cinema-is-evolving/>

3. <https://entertainment.howstuffworks.com/movie-cost1.htm>

4. Are movies getting longer?

[http://www.businessinsider.com/movies-are-getting-longer-2013-1](%22)