The Data Open

Identifying Best Predictors for Movie Revenue through Supervised Learning Models

Konstantin Larin, Willie Turchetta, Connie Xu, Brandon Zhu

September 19, 2020

1 Question

The global film industry is worth 136 billion dollars. However, nearly half of all movies lose money. With so many movies losing money, it is important for a producer to know how much money their movie will make. With this in mind, we pose the following question:

Question: What combination of input features is the best predictor of movie revenue?

In the context of our project, we define input features as the decisions that a producer makes when producing a movie. For example, "who should the writer be? what budget should be allotted? and what should the genre be?" This leads to some other interesting questions:

1. What individual input feature is the best predictor of movie revenue?
2. What suggestions can we make to a movie producer in order to maximize their revenue?

By answering these questions, we hope that movie producers can make better decisions when making a movie.

2 Executive Summary

We looked at nine major movie features that can influence movie revenue: budget, week release, company, writer, rating, director, genre, and runtime. Through our modelling and statistical analysis, we discovered the following key insights.

Budget is the most significant feature in determining gross revenue.

The combination of **budget, week release, company, writer, star, and rating** as input features leads to the best movie revenue prediction model. Adding any more input features like genre obscures the data, leading to a lower R^2 value. This leads to the conclusion that film companies should take special note of those six features when producing a movie.

Some other things to note are that revenue generates a much stronger relationship with the input features in comparison to profit and combining input features creates a stronger model than applying a single input feature.

3 Technical Exposition

3.1 Data Processing

In order to answer our question, we had to examine features that were both numerical (budget and runtime), and categorical (company producing, writer, director, rating, the week of the year that the movie was released, and genre). However, in order to run predictive analysis, we had to change the categorical data into numerical data.

We approached this problem by first splitting the data into training and testing data, a 70/30 split. We then assumed that revenue was normally distributed and computed z-scores with respect to revenue for each of the categorical variables. For example, in order to calculate the z-score for R-rated movies:

1. Find the mean and standard deviation of revenue across all movies
2. Find the mean revenue of R-rated movies
3. Calculate the z-score

$$Z = \frac{x - \mu}{\sigma}$$

If a category did not have enough movies, we assigned it a z-score of 0. For example, while Universal Pictures made 235 movies and we found it fair to calculate a z-score, Milkshake Films has only made one movie, so we assigned it a z-score of 0. By repeating this process for every rating, genre, director, writer, company, and genre, we had data ready to analyze.

	budget	runtime	rating zscore	genre zscore	weekrelease zscore	company zscore	star zscore	writer zscore	director zscore	gross
5280	160000000	148	0.3928	-0.432236	1.133608	0.422981	3.938109	4.349262	4.349262	292576195

Figure 1: Example of post-processing row of data for *Inception*

3.2 Initial Exploration

We started out by doing exploratory analysis on each feature. We went variable by variable and figured out interesting points about each of them.

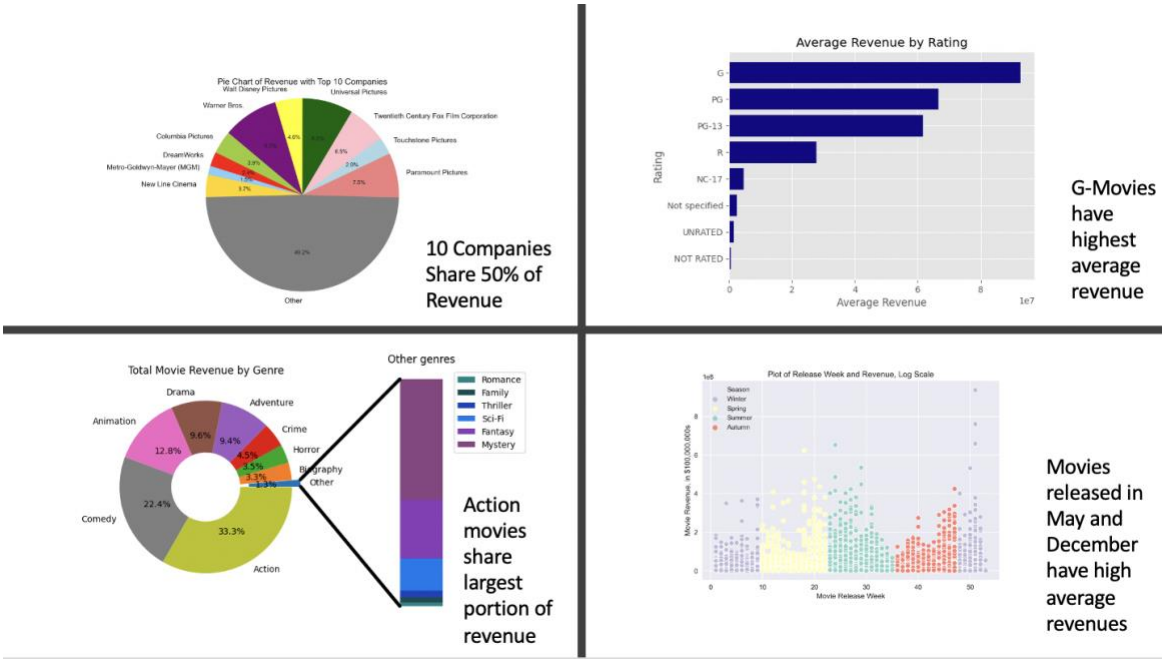


Figure 2: Compilation of exploratory findings

Through this analysis, we realized that each of the features explained the differing revenues a little bit.

These graphs gave us interesting initial findings about revenue and encouraged us to proceed with creating models. For the first attempt, we built individual linear regression models for each feature, with the features as inputs and revenue as outputs. Although overly simplified, we ended up with some notable rankings.

Feature	***R^2 value***
Budget	0.43
Company	0.19
Writer	0.08
Director	0.08
Week Released	0.06
Rating	0.05
Runtime	0.05
Star	0.03
Genre	-0.0002

Figure 3: R^2 values for each input feature

Some features were great as predictors whereas others were effectively useless or even detrimental.

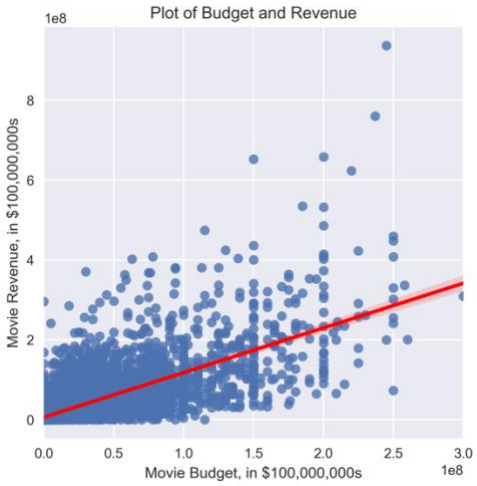


Figure 4: Logistic regression graph of budget and revenue

Looking at features individually, budget was the best individual predator of revenue, with the R^2 value of 0.44. As the budget of a movie increases, revenue also tends to increase. However, the plot shows that there are other factors at play, as a movie with a relatively high budget may fail, while a movie with a small budget can outperform a movie with a much bigger one.

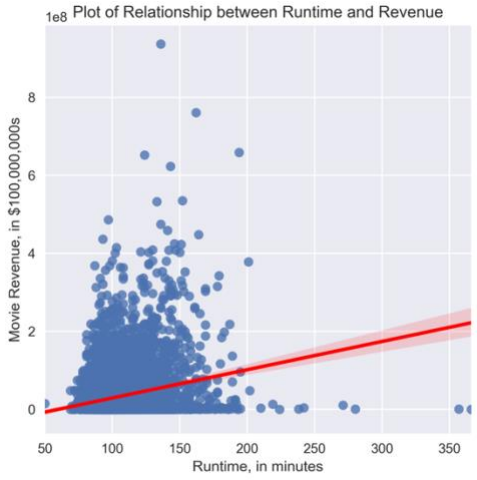


Figure 5: Logistic regression graph of runtime and revenue

Runtime, on the other hand, is not a very good predictor of movie revenue, as seen by the wide fanning error space on the plot. Although the two variables are positively correlated, this is caused by a small number of movies with very high revenue and slightly higher runtime. Moreover, there are a small number of very long movies which may act as influential outliers, as there is insufficient data for any movies with runtime > 200 minutes. This is supported by the low R^2 squared value of 0.05.

Having figured out which individual features were the best predictors of revenue, we then wanted to figure out how good of a model we could build using some combination of all of the features.

3.3 Modeling Process

To determine which features we should combine for our model, we needed to analyze and rank the importance of each of our movie dataset input features in predicting gross revenue. For this, we used an XGBoost regression model. To make sure that XGBoost regression was the best model for our data, we compared its R^2 coefficient (relative measure of fit) and root-mean-squared error (absolute measure of fit) with that of L2 (ridge) regularization linear regression, ElasticNetCV regression, and Support Vector Machines Regression models.

After choosing XGBoost as our prediction model based on its superior R^2 and low RMS errors on both the training and testing datasets, we chose the 6 features (budget, star, weeks released, company, rating, and writer) that yielded the lowest variance and error on our XGBoost model. Among these six features, we then calculated the Weight and Gain of each feature on our XGBoost tree. We learned that the overwhelmingly impactful feature on both XGBoost's weighting and score gain was movie **budget**.

3.3.1 Learning Model Comparisons and Selection

As previously mentioned, we measured the variation and error of these four well-known supervised learning models to determine the model that would best predict gross revenue: XGBoost, Linear Regression with RidgeCV, ElasticNetCV, and Support Vector Regression. Of these four models, XGBoost yielded the highest R^2 and lowest root-mean-squared errors on both the training and test datasets, with significant improvements over the next best models on the training R^2 coefficient (0.629 - 0.370) and the training RMS error (32.325 - 39.848).

	Training R2	Test R2	Train RMS Error (mil)	Test RMS Error (mil)
XGBoost	0.6290	0.5301	32.325	49.262
Lin Reg with RidgeCV	0.3697	0.5260	39.848	49.470
ElasticNetCV	0.3560	0.5230	39.855	49.614
Support Vector Regression	-10.1260	0.1080	59.164	67.859

Figure 6: R^2 coefficients and RMS errors of the four tested models

We also created a visual representation of the performance of our XGBoost regression model versus that of our RidgeCV and SVR models. Blue lines represent predictions to our testing data, while gray lines represent predictions to random normally distributed variations to our testing data. From the charts on the next page, we can see that XGBoost yields accurate predictions to both actual test data (blue lines) and test data mixed with noise (gray lines). The RidgeCV model fits our testing data slightly worse and doesn't yield accurate predictions to noisy data, while the SVR model yields a poorer overall fit.

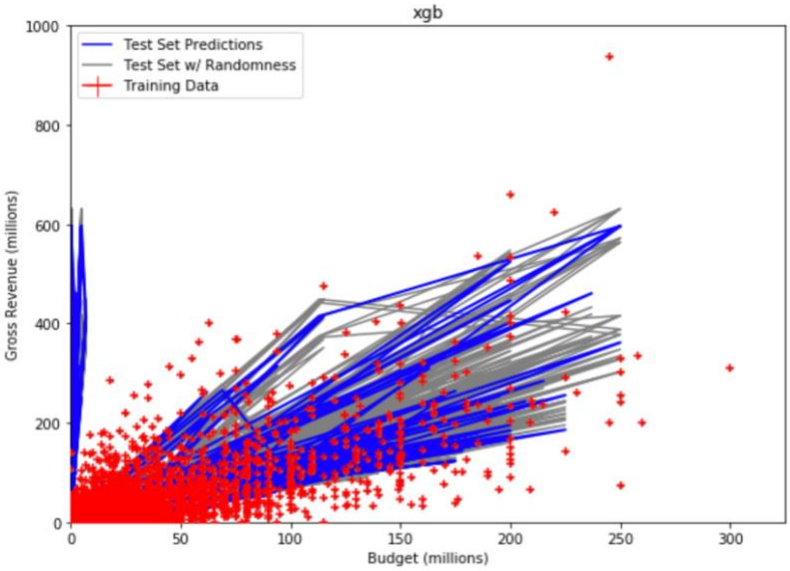


Figure 7 (a): XGBoost regression model of budget and gross revenue

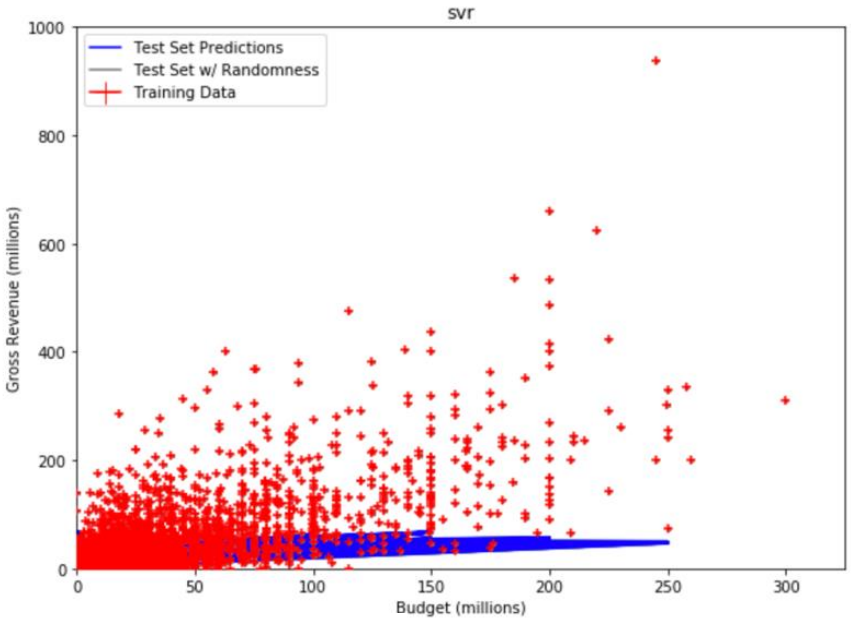


Figure 7 (b): Support Vector regression model of budget and gross revenue

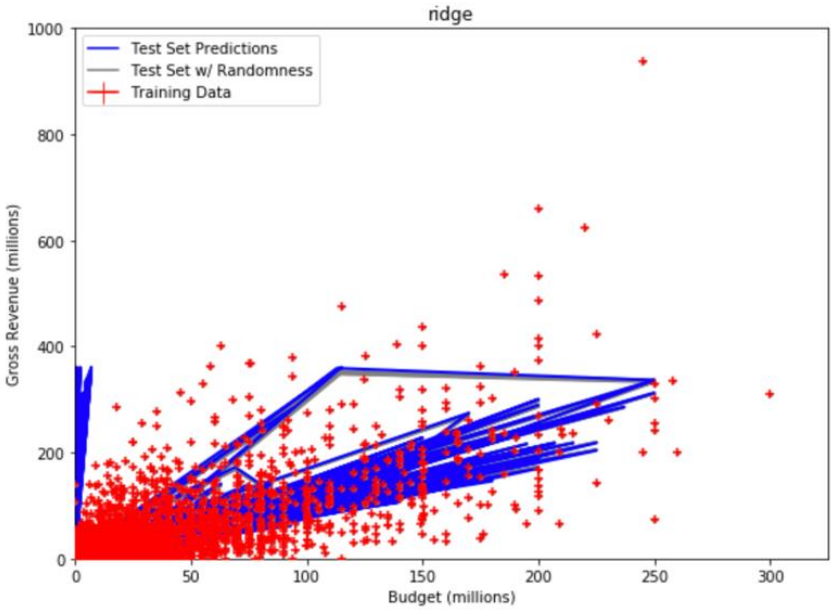


Figure 7 (c): RidgeCV or L2 regularization linear regression model of budget and gross revenue

3.3.2 Finding Optimal Revenue-Predicting Features

After confirming the use of XGBoost as our model, we sought out the optimal combination of our movie dataset input features which yields the highest precision and R^2 coefficient. Doing so ensures that our XGBoost model is represented by the most important features, and that it gives the most accurate revenue prediction. We found that the combination of the six features of budget, star, writer, week of release, company, and rating yielded the best R^2 and error of 0.530 and 49.262 on our testing data, respectively. Including features like genre and director hurt the resulting R^2 and error values, so they had to be excluded.

Next, we ranked the six relevant features based on their weight in our XGBoost tree, gain in accuracy, dataset cover, total gain in accuracy, and total dataset cover scores. The following table describes each feature's performance -- we clearly see that the budget of movies overwhelmingly exceeds star and company in weight, gain, cover, total gain, and total cover. In particular, budget accounts for approximately 4 times the accuracy gain and total accuracy gain of the next highest-ranking feature.

	Weight	Gain	Cover	Total Gain	Total Cover
budget	179	173942.302596	1545.960894	3.113567e+07	276727.0
star zscore	136	46116.045756	1339.242647	6.271782e+06	182137.0
writer zscore	83	92930.240753	1981.650602	7.713210e+06	164477.0
weekrelease zscore	112	36588.321665	1037.000000	4.097892e+06	116144.0
company zscore	133	49083.662423	1793.443609	6.528127e+06	238528.0
rating zscore	35	10987.449840	1866.600000	3.845607e+05	65331.0

Figure 8: Weight, gain, and cover values of each feature

We also generated plots of model weight and accuracy gain for all 6 features. Visually, we can see that movie budget accounts for the highest weight, followed by movie star, production company, week of release, writer, and rating. Movie budget also accounts for the highest total gain in accuracy, followed by writer, star, company, week of release, and rating. Thus, we conclude that the most important feature in predicting gross revenue of a movie is its **budget**.

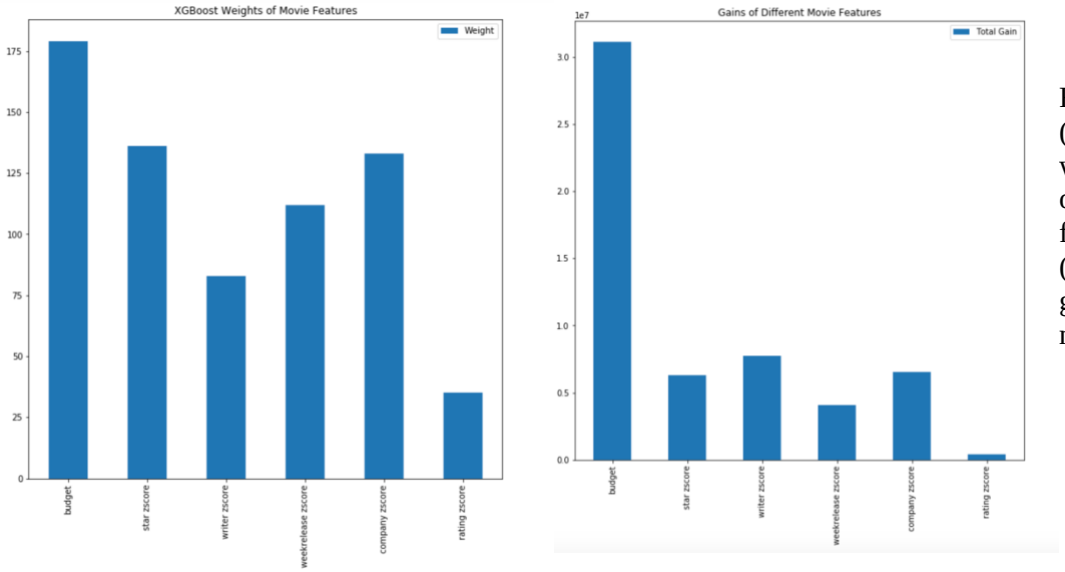


Figure 9:
(a) XGBoost weight scores of movie features
(b) XGBoost gain scores of movie features

4 Conclusion

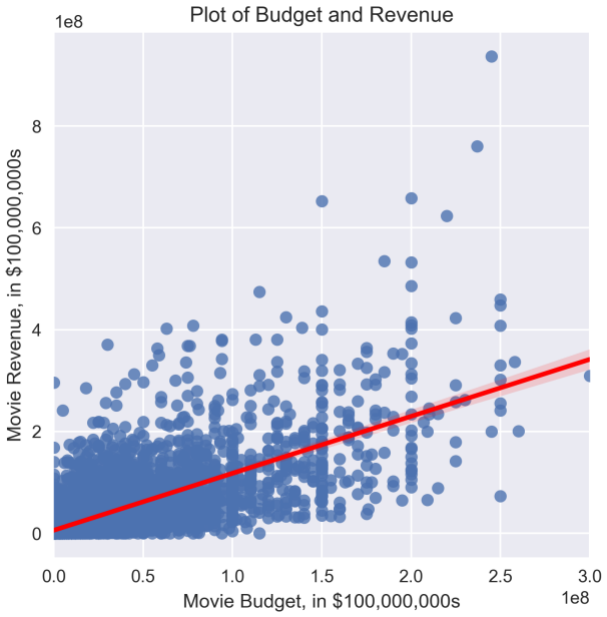
Taking from the results of our analysis, we would recommend movie producers to pay particular attention to budget when predicting movie revenue. Similarly, producers should take into account the stars, the script-writer, the week of release, the company which is releasing the movie, and the rating (G, PG, etc). In general, the larger the budget, the greater the revenue. In addition, movies released in May and December as well as movies rated G have higher average revenues. Finally, the most lucrative stars, script-writers, and companies are listed in the appendix.

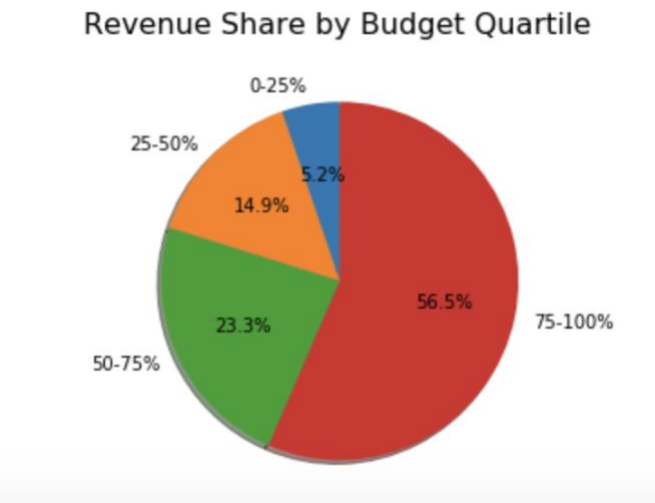
It is interesting to consider that some basic movie features, like genre, runtime, or even director should not be taken into account when attempting to predict revenue, as it decreases the accuracy of the prediction. For genre, this seems to make sense, as movie genres are simply rough categories under which movies are lumped together. Often, one movie can even span multiple genres, and directors can decide to apply a popular genre to a certain movie to encourage more viewers, thereby skewing the viewership results and thus gross revenue of the movie. However, this is odd when we consider directors, as it is often assumed that great directors can consistently produce great movies. Nevertheless, the models and data speaks for itself.

One limitation of our modelling is the lack of data. Almost ⅓ of the movie budgets in the movie_industry.csv file was 0, which is impossible given that movies really cannot be produced for free. This reduced the total number of valid data points down to 4638. Furthermore, all of the movies in the file were popular movies from the last 30 years, which means that regular movies were excluded. This probably skews the model to over-predict revenue. With more data, we are confident that our model still holds its integrity, and we even guess that the R^2 and RMS error values could be improved, making our model could be an even better revenue predictor.

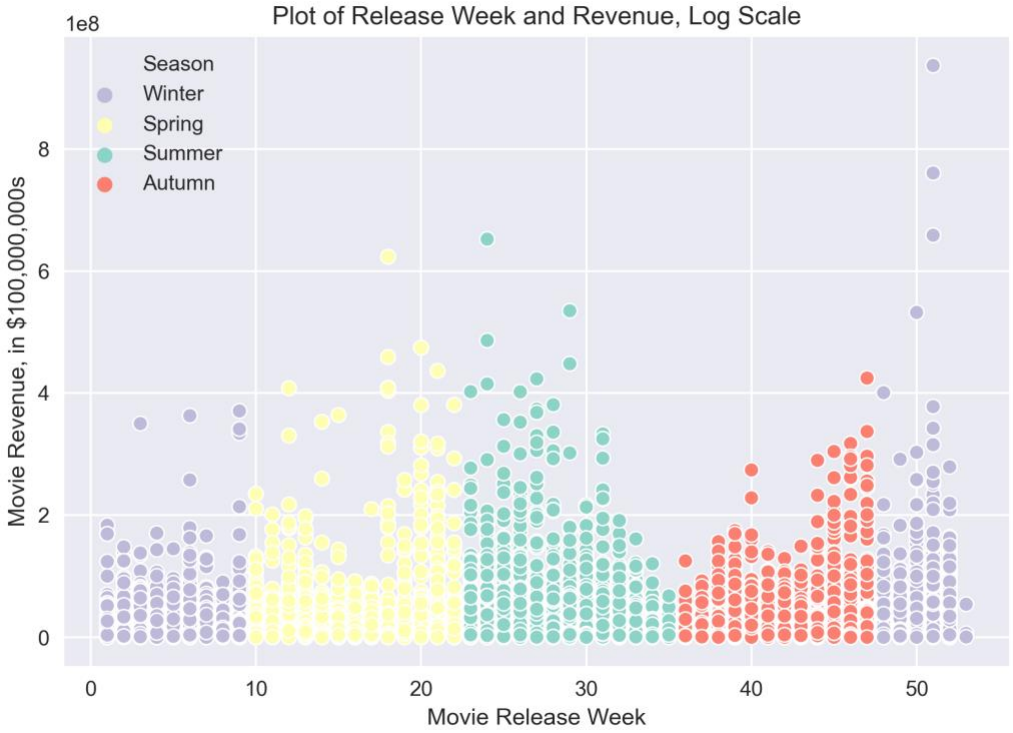
5 Appendix

5.1 Budget

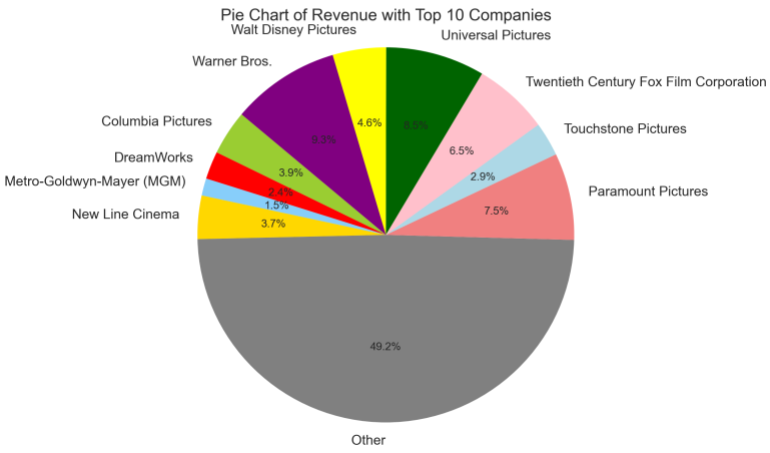




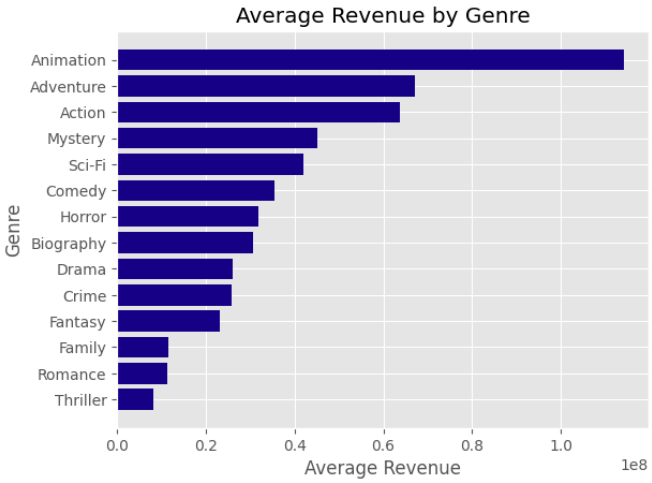
5.2 Week Release

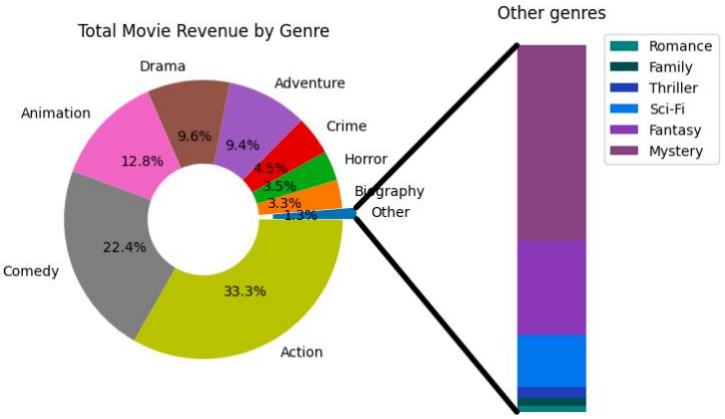


5.3 Company

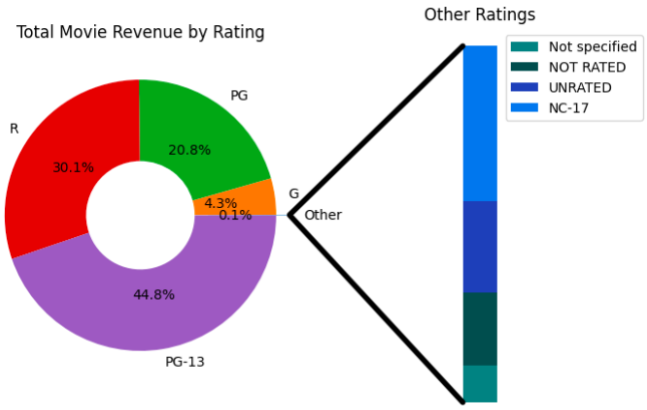


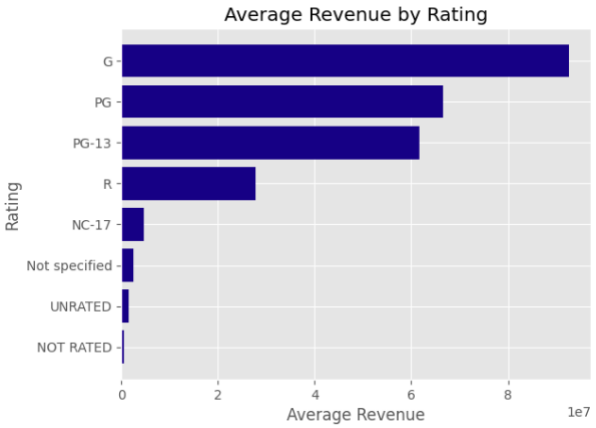
5.4 Genre



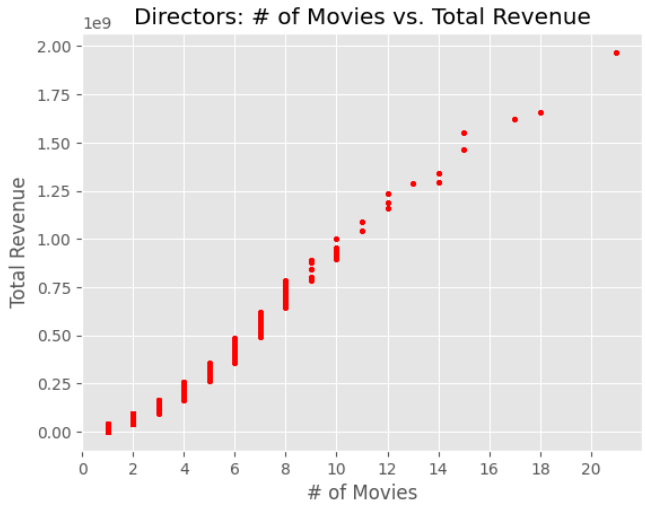


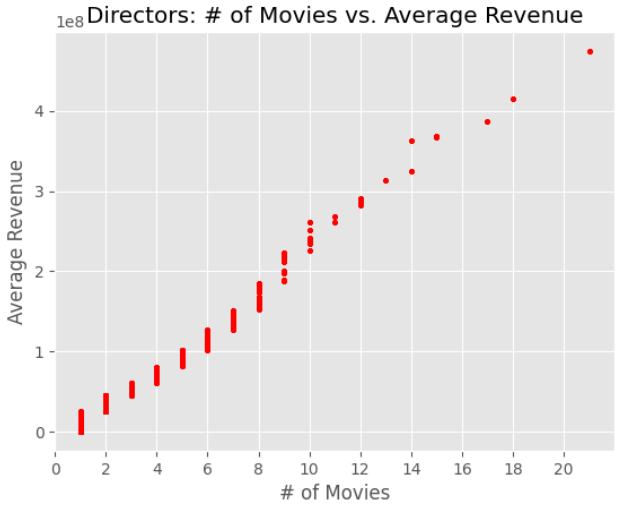
5.5 Rating





5.6 Directors

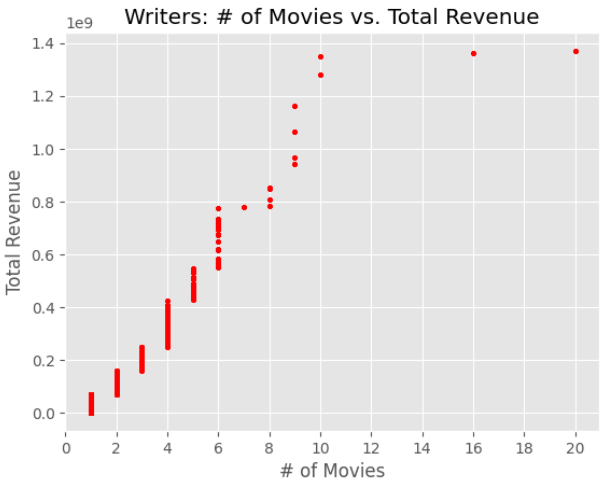


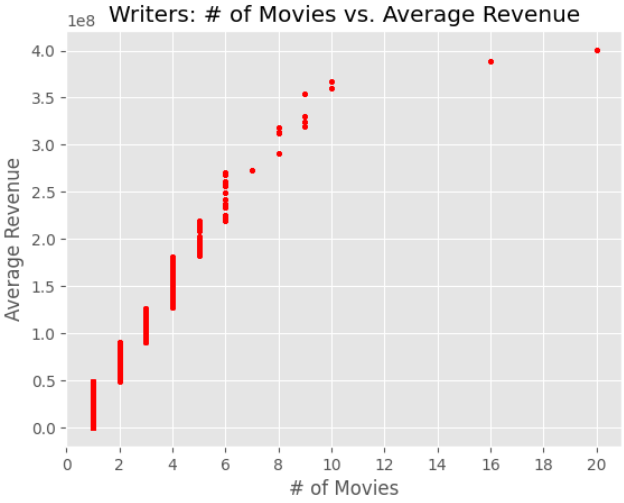


	Director	Total Revenue
1	Steven Spielberg	1.968027e+09
2	Michael Bay	1.659826e+09
3	Robert Zemeckis	1.623140e+09
4	J.J. Abrams	1.550175e+09
5	Sam Raimi	1.464233e+09
6	Peter Jackson	1.342390e+09
7	Ron Howard	1.295042e+09
8	Roland Emmerich	1.287565e+09
9	Tim Burton	1.237984e+09
10	Ridley Scott	1.234603e+09
11	Gore Verbinski	1.187868e+09
12	Christopher Nolan	1.161493e+09
13	Zack Snyder	1.090619e+09
14	David Yates	1.043651e+09
15	James Cameron	1.003137e+09
16	Todd Phillips	9.513662e+08
17	Shawn Levy	9.455960e+08
18	Tony Scott	9.322930e+08
19	Dennis Dugan	9.127303e+08
20	Clint Eastwood	8.960018e+08

	Director	Average Revenue
1	George Lucas	4.745447e+08
2	Lee Unkrich	4.150049e+08
3	J.J. Abrams	3.875438e+08
4	Pierre Coffin	3.680613e+08
5	Gareth Edwards	3.664267e+08
6	Tim Miller	3.630707e+08
7	Joss Whedon	3.244362e+08
8	Roger Allers	3.129000e+08
9	Pete Docter	2.914602e+08
10	Chris Renaud	2.912074e+08
11	Chris Buck	2.859149e+08
12	Andrew Adamson	2.819192e+08
13	Dan Scanlon	2.684928e+08
14	Andrew Stanton	2.610606e+08
15	David Yates	2.609128e+08
16	James Cameron	2.507842e+08
17	Robert Stromberg	2.414104e+08
18	Mark Andrews	2.372832e+08
19	Chris Miller	2.349836e+08
20	Francis Lawrence	2.258565e+08

5.7 Writers

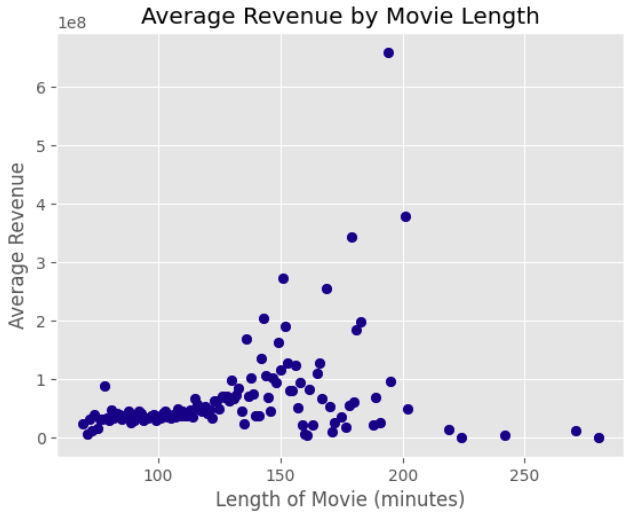

Writers: # of Movies vs. Total Revenue
1e9
Total Revenue
# of Movies
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
0
2
4
6
8
10
12
14
16
18
20


Writers: # of Movies vs. Average Revenue
1e8
Average Revenue
# of Movies
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
0
2
4
6
8
10
12
14
16
18
20

	Writer	Total Revenue
1	Michael Crichton	1.370768e+09
2	Melissa Rosenberg	1.362207e+09
3	John Lasseter	1.349824e+09
4	Ted Elliott	1.279222e+09
5	James Cameron	1.161427e+09
6	Lawrence Kasdan	1.063163e+09
7	Ehren Kruger	9.655188e+08
8	William Steig	9.425080e+08
9	Roberto Orci	8.520445e+08
10	Cinco Paul	8.495332e+08
11	John Hughes	8.089609e+08
12	Andrew Stanton	7.831818e+08
13	Stan Lee	7.772922e+08
14	Jonathan Nolan	7.759684e+08
15	Chris Weitz	7.333287e+08
16	J.R.R. Tolkien	7.203973e+08
17	Lilly Wachowski	7.139836e+08
18	Peter Craig	7.110460e+08
19	Bob Kane	7.109943e+08
20	Steve Kloves	6.997939e+08

	Writer	Average Revenue
1	Jennifer Lee	4.007380e+08
2	Stan Lee	3.886461e+08
3	Chris Weitz	3.666643e+08
4	J.R.R. Tolkien	3.601986e+08
5	Lawrence Kasdan	3.543875e+08
6	Winston Groom	3.302522e+08
7	Joss Whedon	3.244362e+08
8	Ted Elliott	3.198055e+08
9	Mark Fergus	3.184121e+08
10	William Steig	3.141693e+08
11	Irene Mecchi	3.129000e+08
12	Pete Docter	2.914602e+08
13	Melissa Rosenberg	2.724414e+08
14	Garth Jennings	2.703290e+08
15	Dan Scanlon	2.684928e+08
16	Andrew Stanton	2.610606e+08
17	Jonathan Nolan	2.586561e+08
18	John Lee Hancock	2.559595e+08
19	Jared Bush	2.487570e+08
20	J.K. Rowling	2.416982e+08

5.8 Movie Length

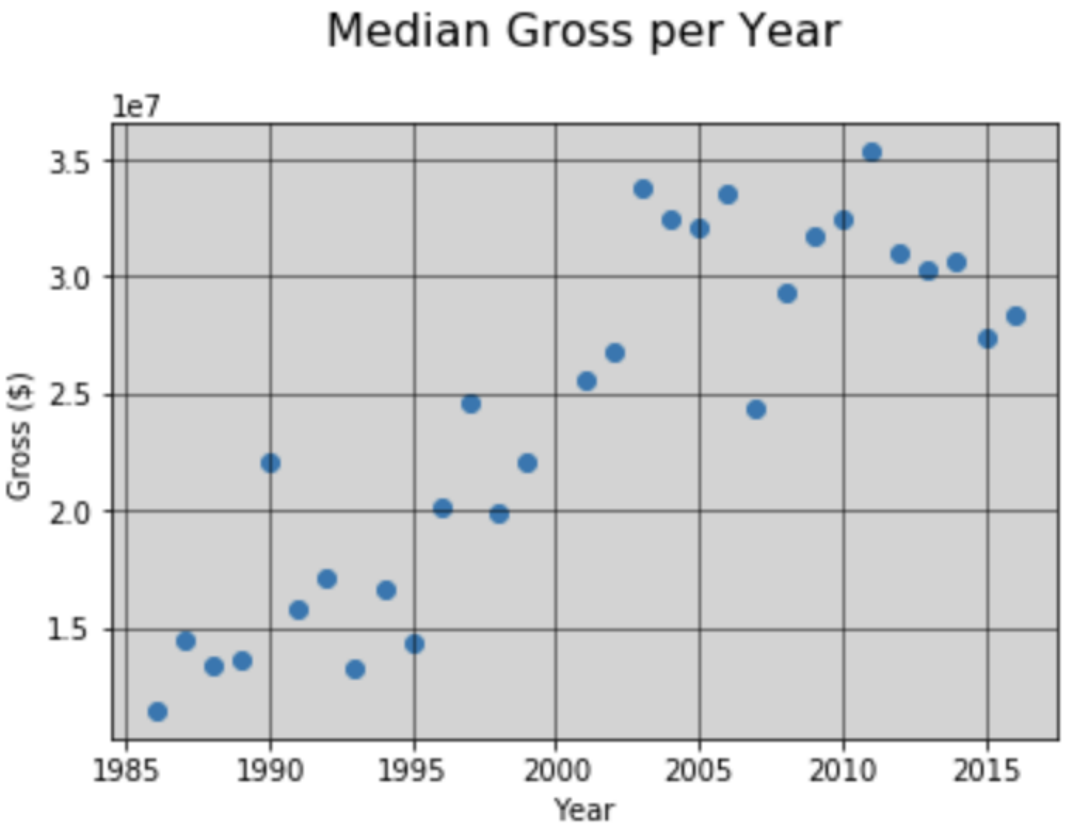


5.9 Stars

	Stars	Average Revenue
1	Daisy Ridley	9.366622e+08
2	Ellen DeGeneres	4.862956e+08
3	Louis C.K.	3.683843e+08
4	Neel Sethi	3.640011e+08
5	Daniel Radcliffe	3.107764e+08
6	Edward Asner	2.930042e+08
7	Felicity Jones	2.677956e+08
8	Craig T. Nelson	2.614411e+08
9	Chris Pratt	2.577607e+08
10	Quinton Aaron	2.559595e+08
11	Auli'i Cravalho	2.487570e+08
12	Kelly Macdonald	2.372832e+08
13	Ben Burtt	2.238082e+08
14	Ryan Potter	2.225278e+08
15	Paige O'Hara	2.189676e+08
16	Jason Lee	2.184708e+08
17	Brad Garrett	2.064457e+08
18	Lily James	2.011514e+08
19	Mike Myers	1.993537e+08
20	Sam Neill	1.975241e+08

	Stars	Total Revenue
1	Tom Hanks	3.311300e+09
2	Johnny Depp	2.210301e+09
3	Ben Stiller	2.179987e+09
4	Tom Cruise	2.170503e+09
5	Adam Sandler	1.948877e+09
6	Leonardo DiCaprio	1.940593e+09
7	Denzel Washington	1.813359e+09
8	Mike Myers	1.794183e+09
9	Jim Carrey	1.776593e+09
10	Mel Gibson	1.738127e+09
11	Kristen Stewart	1.539129e+09
12	Matt Damon	1.518554e+09
13	Robert Downey Jr.	1.495936e+09
14	Will Smith	1.494191e+09
15	Tobey Maguire	1.334650e+09
16	Eddie Murphy	1.312859e+09
17	Robin Williams	1.278978e+09
18	Bruce Willis	1.201146e+09
19	George Clooney	1.199928e+09
20	Ben Affleck	1.117025e+09

5.10 Other


Median Gross per Year
1e7
3.5
3.0
2.5
2.0
1.5
Gross ($)
1985
1990
1995
2000
2005
2010
2015
Year