Relationships Affecting Profitability of Movies

Chris Thomas and Miriam Rognlie

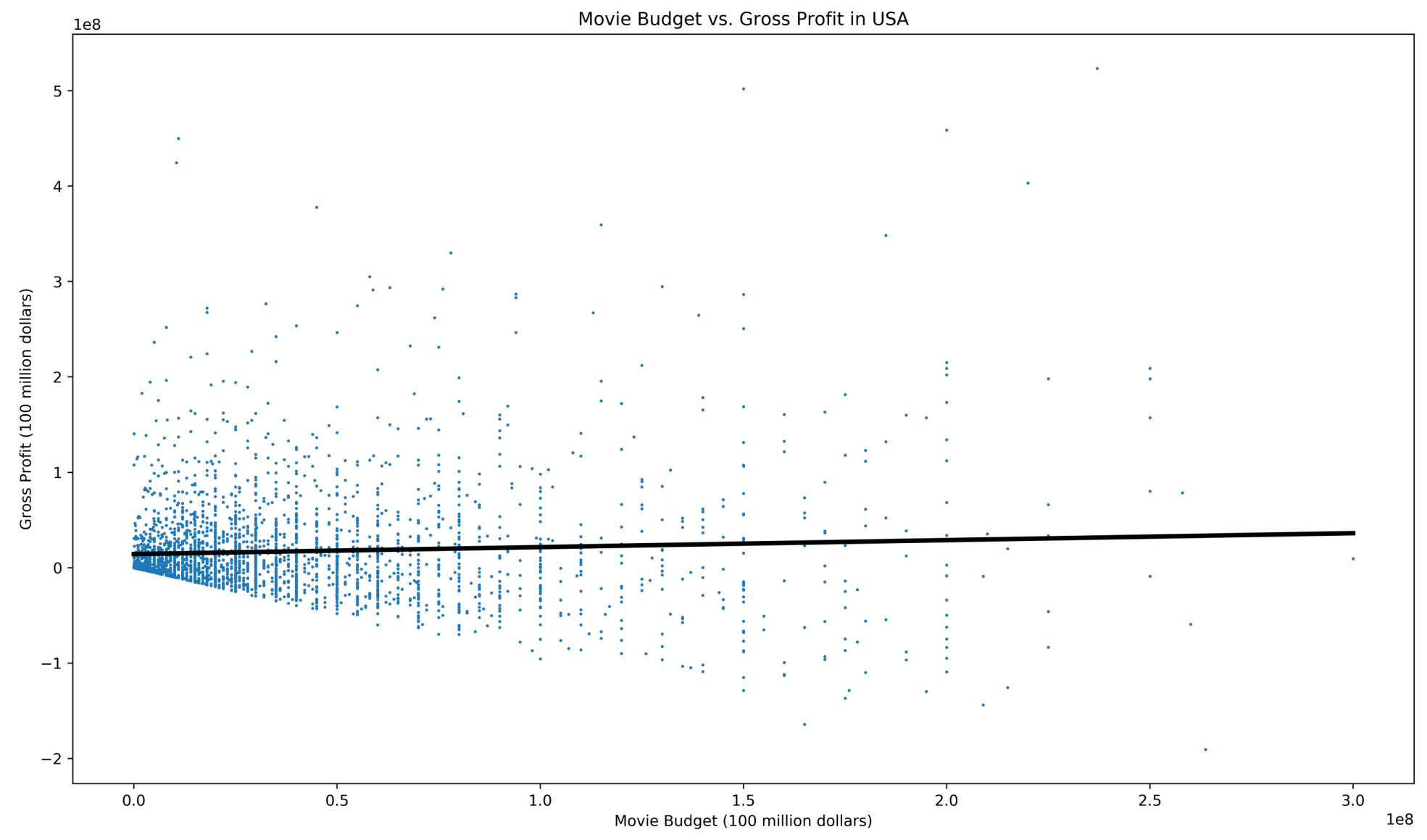
Abstract

This project examines qualities of movies which may affect their profitability. All data is derived from several publicly available datasets published by IMDB. IMDB collects information on movies as well as the actors and directors involved in creating them. Five questions were considered in total, each of which will be analyzed briefly here. All graphs were created with the help of SciKit-Learn and Matplotlib libraries for Python. We focused on the use of linear and ridge regression to help us look for trends in data where it made sense. We also took advantage of box-and-whisker plots, and graphed averages over time where necessary. For most questions, we have considered only the subset of movies produced in the United States to avoid discrepancies based on currency, since most answers necessitate using the budget and gross profit metrics. For the duration of our analysis, a "successful" movie is deemed one which collects box office profits greater than or equivalent to its budget. A "failure" is considered one which does not recover its budget in box office profits. Since the application of data science to movie datasets are very diverse, we have chosen to focus on issues relating to economic success of a movie. These are important questions to consider because they may help to predict the profitability of future films, and the techniques used here can be applied to many other projects. Many of the analyses we performed were somewhat inconclusive with small pieces of information gleaned while others were surprisingly useful. Content rating turned out to give a good amount of information on the kind of performance that could be expected from a movie.

Results

Question 1: How is a movie's budget related to its gross profit?

We have found there to be no strong correlation between the budget of a movie and its profitability. Investing more money into the production of a movie did not appear to guarantee larger profits. It did however, appear that movies with higher budgets tended to be able to turn some more extreme profits. The query-generated graph and corresponding statistics suggest that there is likely no correlation between the values. We concluded that spending over 250 million dollars to produce a movie was inadvisable. Few if any movies are wildly successful above this threshold, and the high budget means the risk is high as well. More importantly, movies with smaller budgets in the 150 million range can often see the same if not larger profit margins without the high risk investment.



Linear Regression Line Slope = 0.73

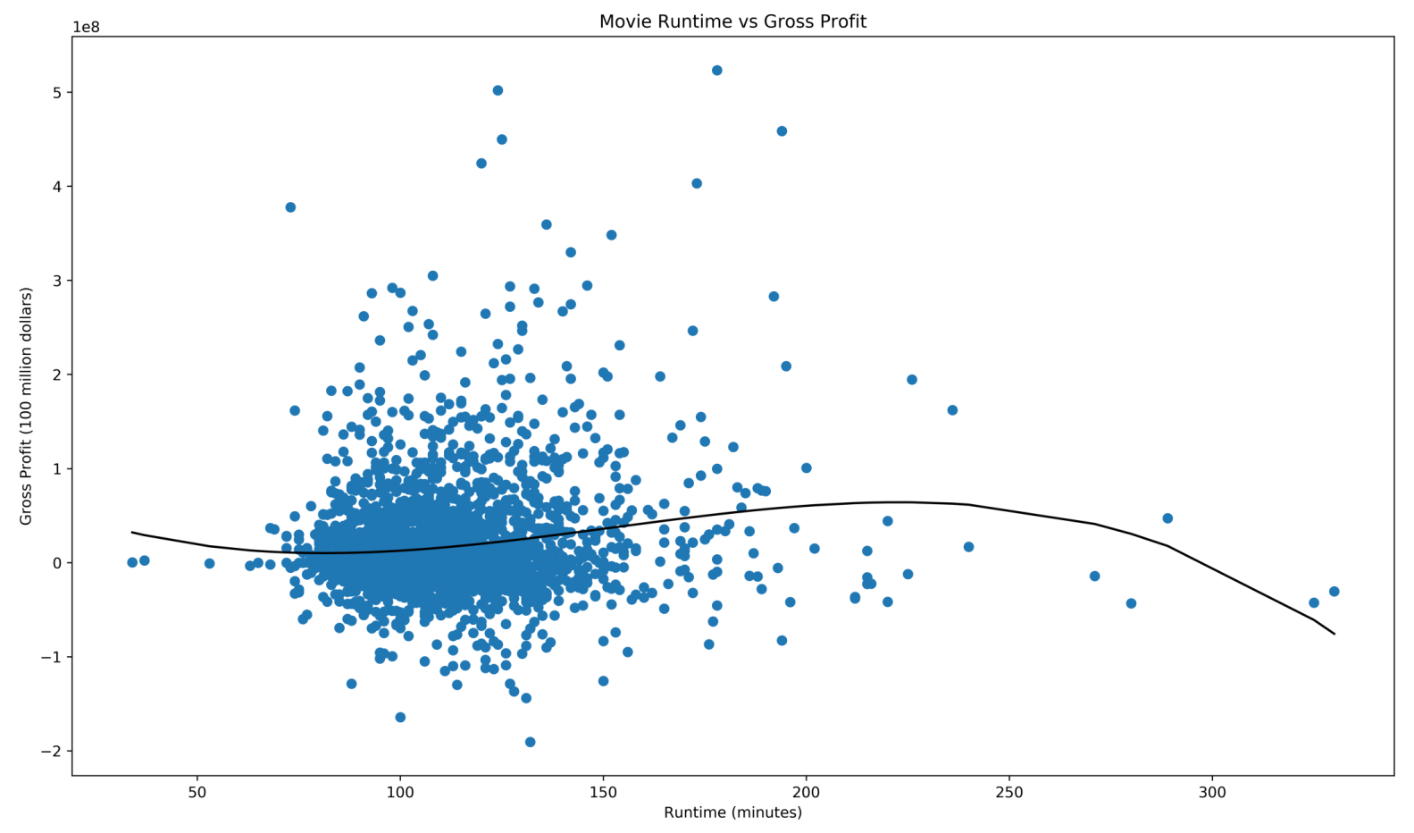
Pearson Correlation Coefficient = 0.05

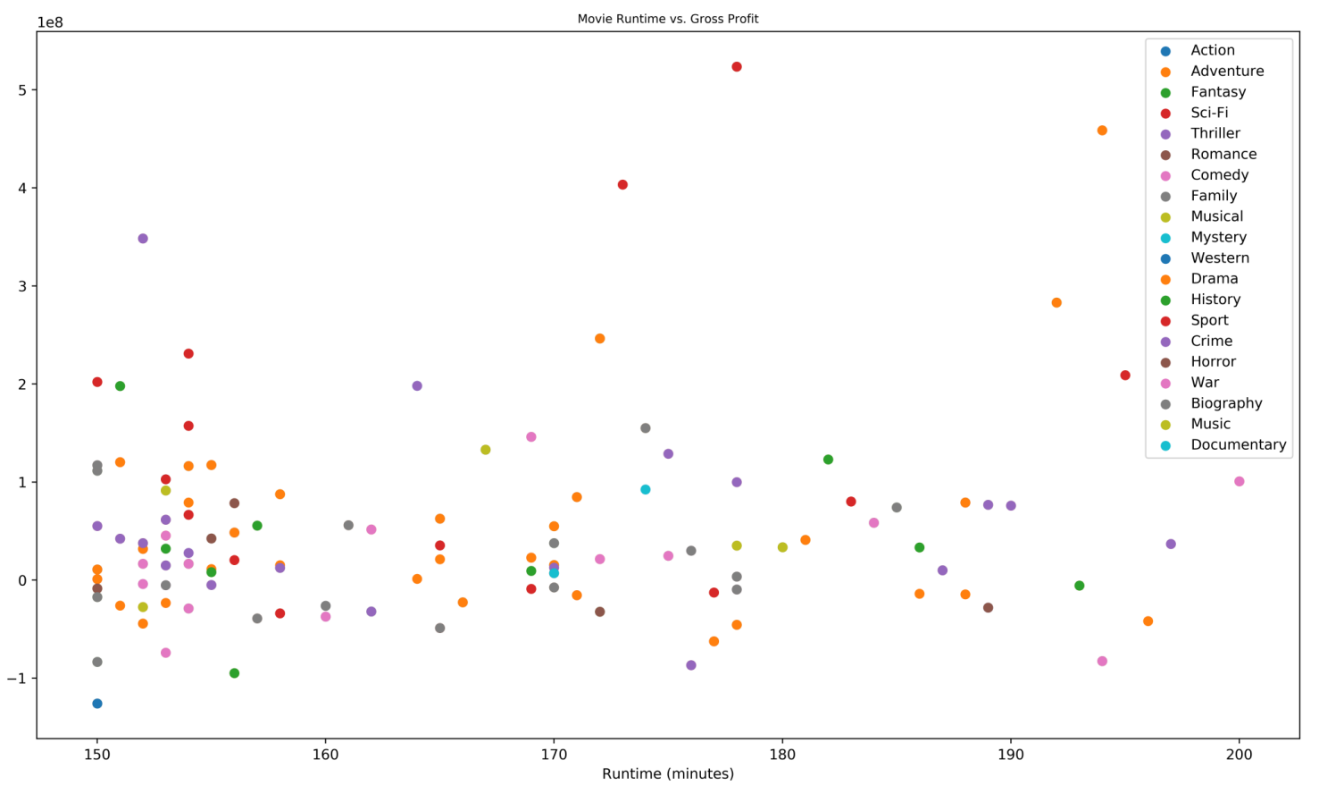
Spearman Rank Coefficient = -0.08

P-Value = 8.71x10^-6

Question 2: Does the length of a movie affect its profitability?

The average movie in this dataset was 109.80 minutes long, with a standard deviation of 22.75 minutes. Interestingly, the most profitable of movies are between 75 and 200 minutes in length, but so are most of the failures. As movies become longer the extremes of profit taper out with the longest movies losing money. We applied ridge regression using a 3rd degree polynomial. Due to the spread of data in the generated graph, finding an appropriate regression line was difficult. Runtimes between 150 and 200 minutes tended to be a bit more profitable (or at least less unprofitable) than movies in the much shorter ranges. This could be due to lack of data for higher ranges, or that the content of these longer movies can hold people's attention more effectively. Breaking the results down into categories reveals that the answer to this question is dependent also on the movie's genre. For convenience, we provided a zoomed version of the graph of the 150 to 200 minute range colored by genre. This range saw a high amount of adventure and drama movies which seemed to perform rather well.



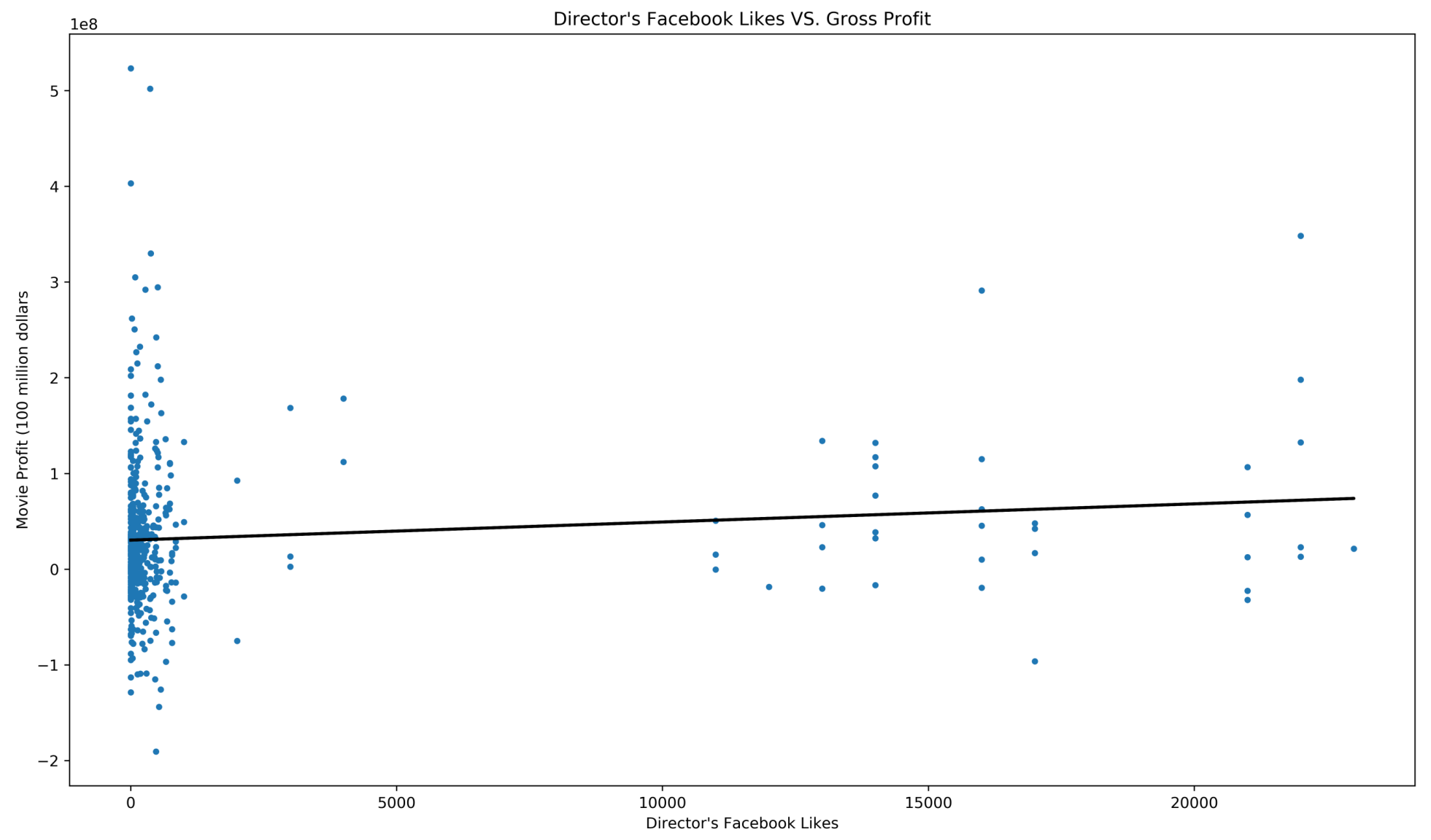


Question 3: Is the director's popularity (based on Facebook likes) statistically correlated with the profitability of a film?

"Facebook Likes" is a rather new metric to judge movies by and can be an effective way to judge movie performance by as more data becomes available. In particular, statistics related to social media can track interest in ways that raw profit numbers cannot. Unfortunately, not all directors have Facebook accounts and the data is sparse. With the data we see it is possible there could be a weak correlation. In addition, it is also possible that only the most famous and recognizable directors will ever receive a very large sum of likes. This could account for the positioning of the data to the far left. Determining the causes for this are outside the scope of the dataset, but further research in this area could reveal interesting relationships between the film and social media industries.

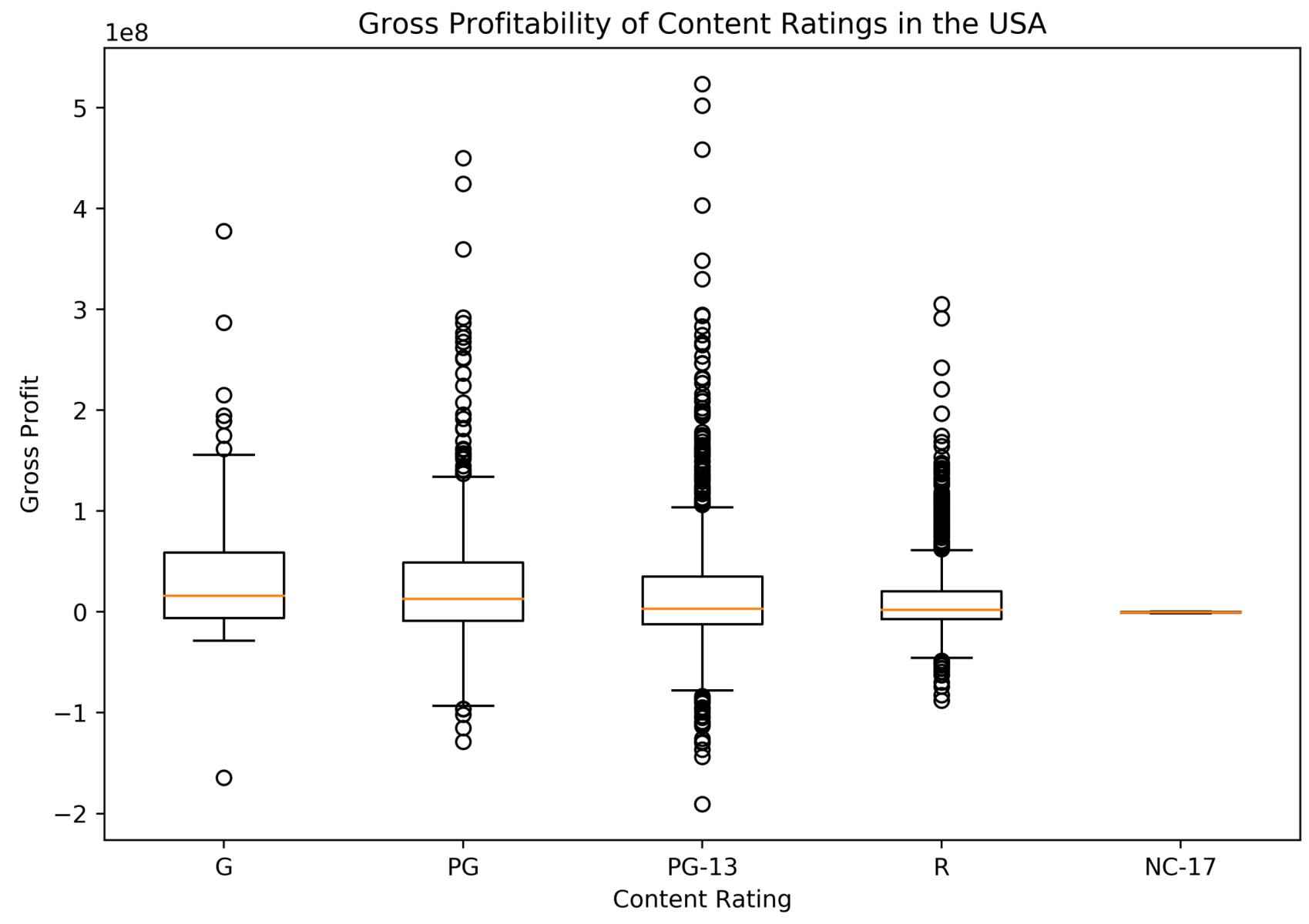
Linear Regression Line Slope = 5.83 Pearson Correlation Coefficient = 0.29

Spearman Rank Coefficient = 0.09 P-Value = 0.37x10^-6



Question 4: Is content rating correlated with the success of a movie?

Based on the given dataset, it does appear that profit spreads and averages generally differ based on content ratings. A box-and-whisker plot is provided to illustrate these differences. We learned that in general, the industry hovers just over breaking even on average across all ratings. What is more interesting is that G and PG movies on average were more likely to be successful than their higher rated counterparts. PG-13 movies could occasionally turn somewhat large profits and contained the most profitable of all movies. Rated R movies often make relatively small profits. Rated G movies have the largest spread in profits, excluding outliers. NC-17 movies are scarce making them less reliable to glean data from. There were only 4 such relevant movies in the entire dataset. Due to the nature of content ratings, we believe G and PG-rated movies are the most accessible and have the widest possible audience while R and NC-17 are the least accessible. This explains the patterns we see in the data. Mean profit consistently goes down as content rating goes up.



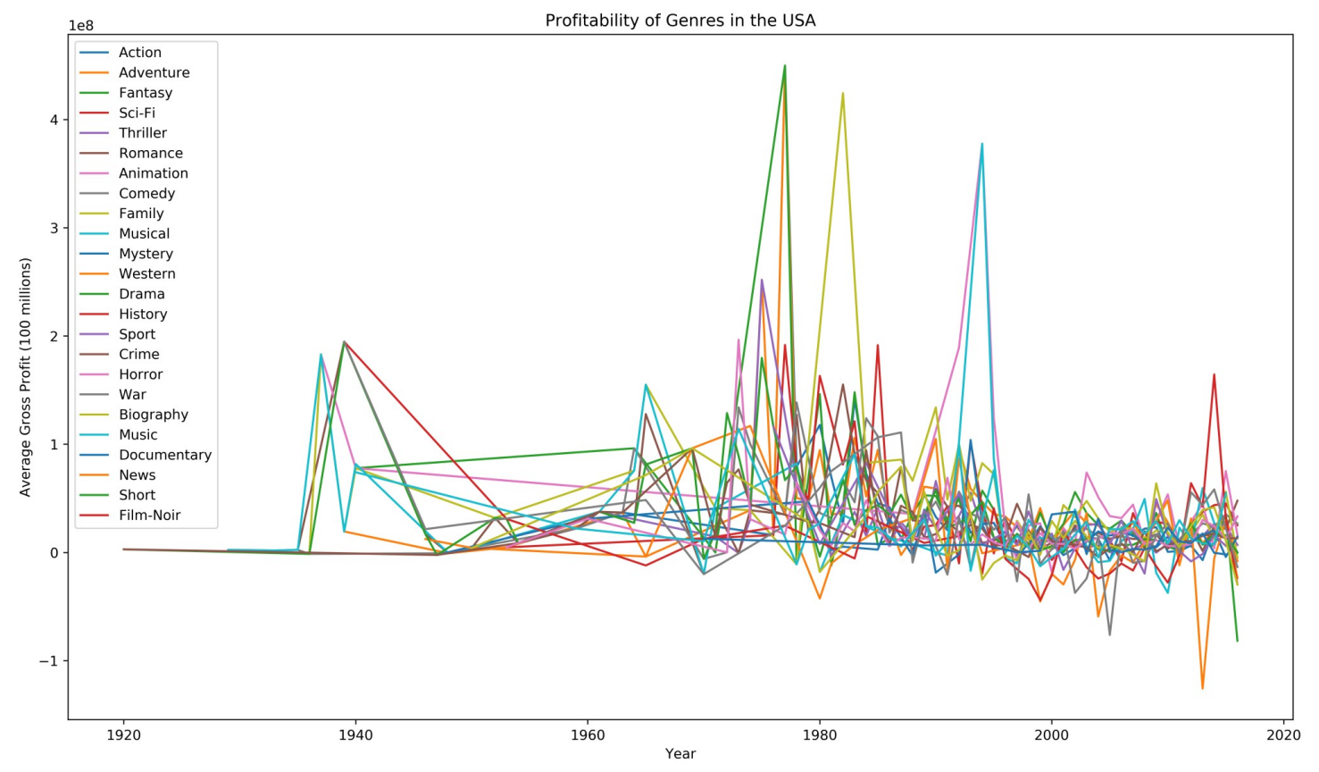
	Mean Gross Profit	Standard Deviation Gross Profit	Successful	Unsuccessful	Total	Success Rate
G	$42,944,275.94	$79,194,200.35	46	25	71	0.647887324
PG	$27,960,527.49	$69,304,377.37	300	160	460	0.652173913
PG-13	$18,588,209.40	$64,524,998.19	579	480	1059	0.54674221
R	$11,173,006.05	$36,607,924.80	728	578	1306	0.557427259
NC-17	($288,010.75)	$685,550.31	2	2	4	0.5

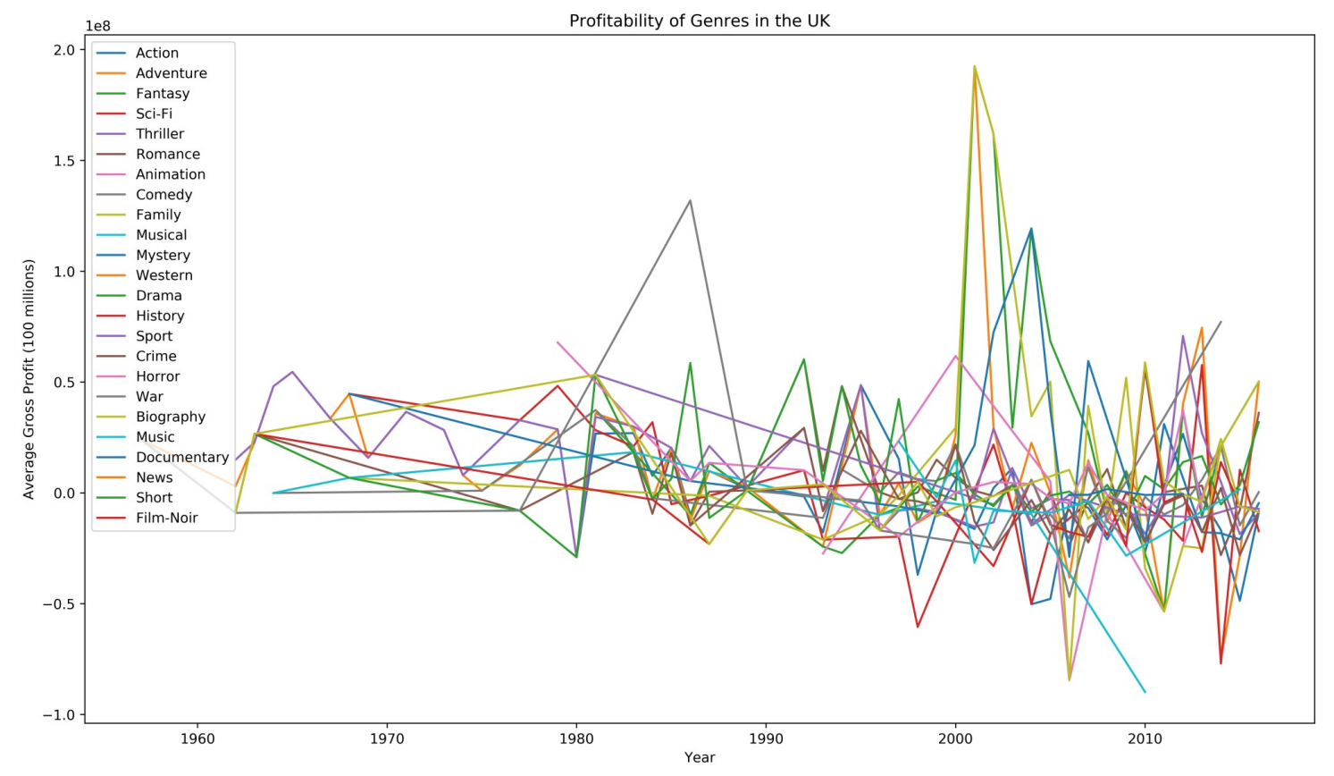
Question 5: Over time, how has profitability based on genres changed for different regions around the world?

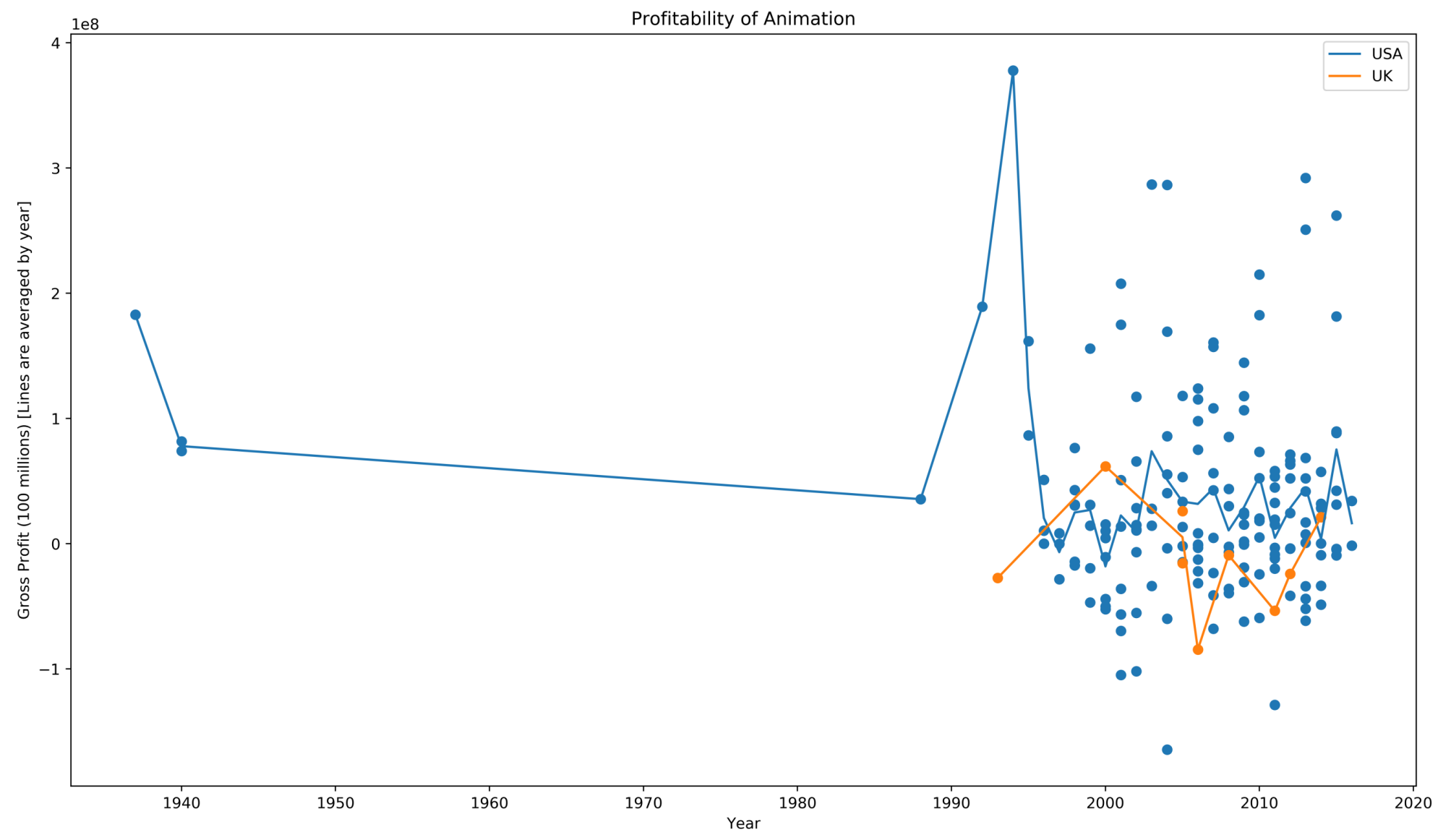
This was the most difficult question to answer because this dataset lacked sufficient samples from most regions in the world besides the USA. As such only the USA (78.9% of the dataset) and UK (8.2% of the dataset) are compared, since the UK was the second most common country to produce movies in this dataset. 24 different genres were encountered in this dataset, so we have focused on two in the interest of brevity. From 2000 through 2002 in the USA, westerns lost large sums of money. In 2003, they became the 4th most profitable. From 2004 to 2007 they continued their decline losing money each year without fault. Finally 2008 and 2009 saw no movies of this genre. When the genre came back in 2010 it was revived, making an average of $48,285,381 making it the second most profitable genre in 2010 right after Animations. This pattern of consistent losses with large one-year resurgences suggests that western movies are not a very safe genre since they jump back and forth from the least successful to the most successful spots.

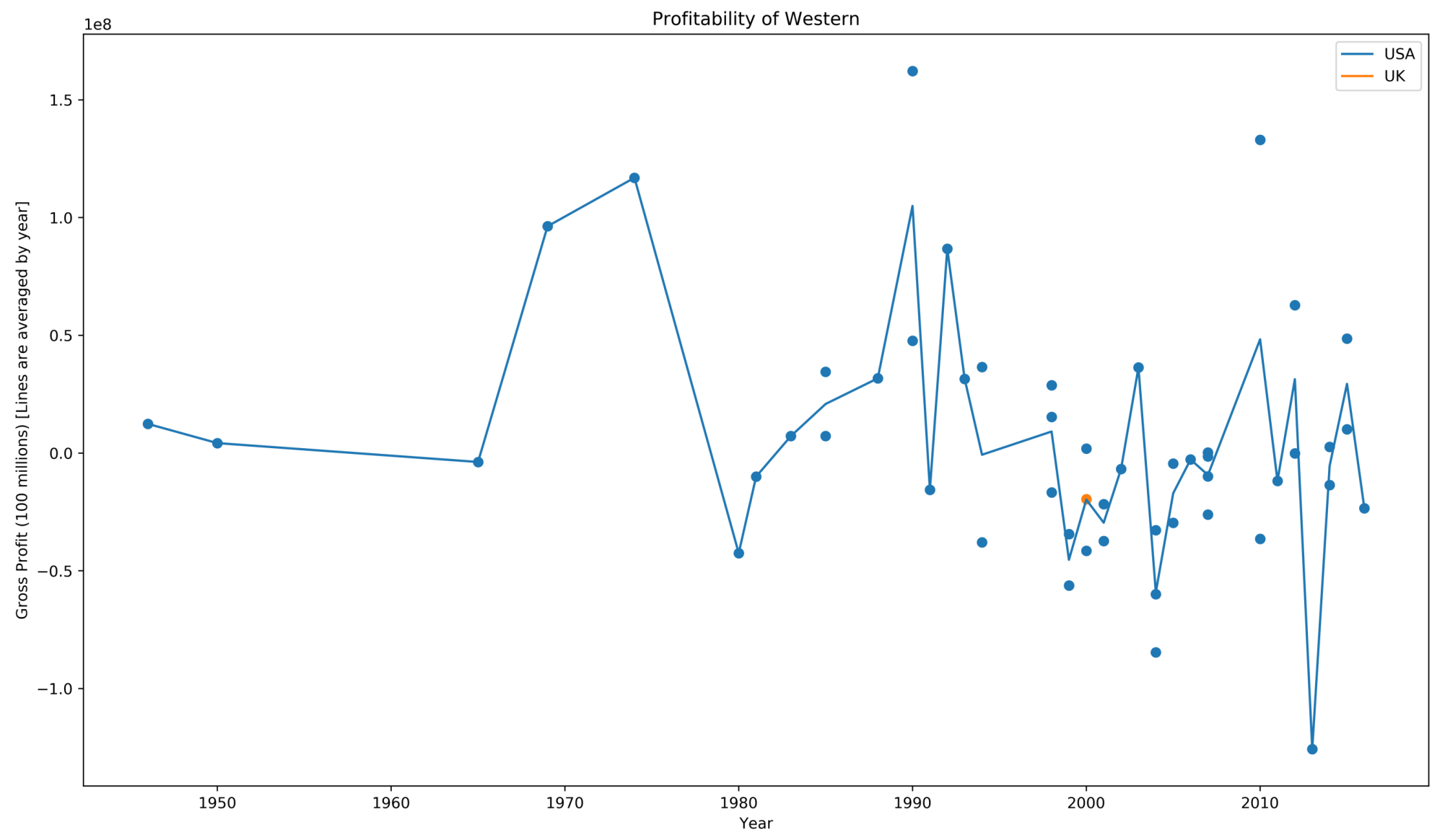
Animations have consistently performed very well and seem to be among the safest of options. In 2000, they lost $18,104,135 on average. Every year since, they always made money. In 2003 all through 2007 they were the most profitable of genres.

In general, the UK movies tend to have less genres appear each year than the US does and most of them lose money each year. In 2000, animation was the most popular genre, however no additional animations were produced in the UK until 2005. That year, they made a small profit. Over the next ten years, only 2008, 2011, 2012 and 2014 would have releases of animation movies. All would have significant losses on average except for 2014 where the genre became the 3rd most profitable. Looking at the data, there is only one data point for every year in the UK chart except for 2005. This makes drawing conclusions about the market from the country of production difficult. Only one western movie was ever made in the UK during this time period.









Conclusions

The applications for data science in the movie industry are vast. Clever application of it may reveal relationships which allow producers to more accurately predict the consequences of different choices when designing a new film. Even though many of the answers to our questions were not conclusive, the topics and questions discussed are especially important to companies that are in the business of making movies for profit.

Resources and Queries

Resources:

- *Documentation of scikit-learn 0.19.1.* Scikit-Learn, 2017. http://scikit-learn.org/stable/documentation.html. Last Accessed 5 December 2017.
- *IMDB Data from 2006 to 2016.* Prompt Cloud, 2017. https://www.kaggle.com/PromptCloudHQ/imdb-data. Last Accessed 5 December 2017.
- *TMDB 5000 Movie Dataset.* The Movie Database API., 2017 https://www.kaggle.com/tmdb/tmdb-movie-metadata. Last Accessed 5 December 2017.

Queries:

Question 1:

```
SELECT Budget, (cast(Revenue AS SIGNED) - cast(Budget AS SIGNED )) AS GrossProfit FROM movie WHERE Country='USA' AND Budget IS NOT NULL AND movie.Revenue IS NOT NULL ORDER BY Budget
```

Question 2:

```
SELECT Runtime, (cast(Revenue AS SIGNED) - cast(Budget AS SIGNED )) AS GrossProfit FROM movie WHERE Country = 'USA' AND movie.Revenue IS NOT NULL AND Budget IS NOT NULL and Runtime is not null order by Runtime
```

```
SELECT Runtime, Rating FROM movie WHERE Rating IS NOT NULL and Country = 'USA' and Runtime is not null ORDER BY Runtime
```

```
SELECT Runtime, (cast(Revenue AS SIGNED) - cast(Budget AS SIGNED )) AS GrossProfit, genre_id FROM (movie JOIN movie_has_genre ON Movie_id=movie.id) WHERE Revenue is not null and Budget is not null and Country = 'USA' and Runtime is not null AND genre_id= {id}
```

```
SELECT Runtime, Rating, genre_id FROM (movie JOIN movie_has_genre ON Movie_id=movie.id) WHERE Country = 'USA' and Rating is not null AND genre_id= {id}
```

Question 3:

```
select Rating, (cast(Revenue AS SIGNED) - cast(Budget AS SIGNED )) as gp, Person.FacebookLikes from (movie join (person join person_directs_movie on Person_id=person.id) on Movie_id=movie.id) where Rating is not NULL and Revenue is not NULL and Budget is not null and Country = 'USA'
```

Question 4:

```
SELECT Rating, GrossProfit, FacebookLikes from (movie join person_directs_movie on Movie_id=movie.id) where Rating is not NULL and GrossProfit is not NULL and Country = 'USA'
```

Question 5:

```
SELECT DISTINCT Country FROM movie
```

```
select Year, avg(cast(Revenue AS SIGNED) - cast(Budget AS SIGNED )) as AverageProfit from (movie JOIN movie_has_genre ON Movie_id=movie.id) WHERE budget IS NOT NULL AND movie.Revenue IS NOT NULL AND Country='{country}' AND Genre_id={id} GROUP BY Year order by year
```

```
select Year, (cast(Revenue AS SIGNED) - cast(Budget AS SIGNED)) as grossProfit from (movie JOIN movie_has_genre ON Movie_id=movie.id) WHERE budget is not null and movie.Revenue is not null and Country='{id}' and Genre_id={id} order by year
```