0. **Title and Author**
   Title: Exploration and Analysis of Trends Within the Movie Industry.
   Author: Joey Roach
1. **Summary of Research Questions and Results**
   1) Have there been changes in profitability of particular genres over time?

      Yes, in some instances. It varies dependent on the genre, and the timeframe that one is considering, but some genres have seen non-trivial differences in their profit margins during particular eras.

   2) Are there any directors who have a notoriously bad track record in terms of film perception? What about writers? Star actors/actresses?

      Yes, some filmmakers do have poor records when it comes to the public reception of their films (with respect to IMDB scores). This analysis revealed some interesting trends with regards to looking at public perception of these filmmakers and the impact that those perceptions have on the profit margins for the filmmaker's films, although there is not much of a significant difference between the three types of filmmakers examined.

   3) Is there a relation between the rating of a film and its potential to generate profit?

      Yes, there does appear to be differing expectations for what kind of profit a film can achieve, based upon its MPAA rating.

   4) How accurately can we predict the IMDB score of a film based on a collection of other factors?

      It turns out, we can do this relatively well! The best machine learning algorithm presented in the analysis ended up with a mean squared error rate of approximately 0.44 between actual and predicted IMDB user rating scores.

2. **Motivation and Background**
   I believe that this is an important issue to research as the film industry comprises an incredibly significant chunk of the entertainment industry in today's society. Although I obviously won't be presenting this research to any higher-up film executives, I view the potential answers to these questions as ones that would certainly interest them, being as most of them are either directly or indirectly tied to profit, which tends to drive these types of business decisions. Any indication on how a particular genre has greatly improved in the market over time (say, something like an adventure/action film, which have seen massive boosts with the emergence of blockbusters and comic-book films), or how a director with good ethos among audiences impacts its potential revenue would be important information for executives to have, as it could greatly influence what films they decide to greenlight, and which they decide to not move forward with.

3. **Dataset**

   https://www.kaggle.com/danielgrijalvas/movies

This data set contains information about films from the past 30 years, including (but not limited to) their studios, budgets, gross, genre, filmmakers associated with the project, year of release, and ratings for the films from users on IMDB.

4. **Methodology (Algorithm or analysis)**

This analysis will be conducted on the dataset listed above. For this project, the main point of emphasis is to analyze trends within the film industry over the past 30 years (including trends tied with individuals as well as those tied to particular film characteristics), and be able to determine if these any of these factors and/or trends can be utilized in order to accurately predict the IMDB score for any given film.

Firstly, this analysis will determine the changes in profitability of genres over this 30-year period. That is, all films are to be sorted according to their genre, the mean profit for each genre for each year will be calculated, and the results will be analyzed. Profit will be defined as revenue minus budget, and it is acknowledged that the true profit of a film is not entirely accurately represented by this measure, as it does not factor in additional costs such as marketing/advertising costs, nor other potential revenue sources, such as DVD purchases made after a film leaves theaters. These results are to displayed visually as well.

The project will then shift into analyzing trends among writers, directors, and star actors/actresses. For each writer, director, and star of a film, the average user rating of their films from IMDB (on a scale from 1 to 10) will be averaged; that is, if a director has directed 4 movies, then the analysis will sum up those three user scores and divide it by 4 to obtain a mean score. Stars, directors and writers will then be separated into 3 different tiers:

- Scores from 0-3.3 will be classified as “Bottom third”
- Scores from 3.3-6.6 will be classified as “Middle third”
- Score from 6.6-10 will be classified as “Top Third”

Analysis will then be conducted on each sub-group, showing the profitability trends of films for each class (where profitability is defined as it was in the previous task). The goal here is to determine if we can notice any substantial trends amongst the public’s perception of a film, and if that perception is being accurately reflected in how consumers are acting, which, in this case, means how they are or are not spending their money.

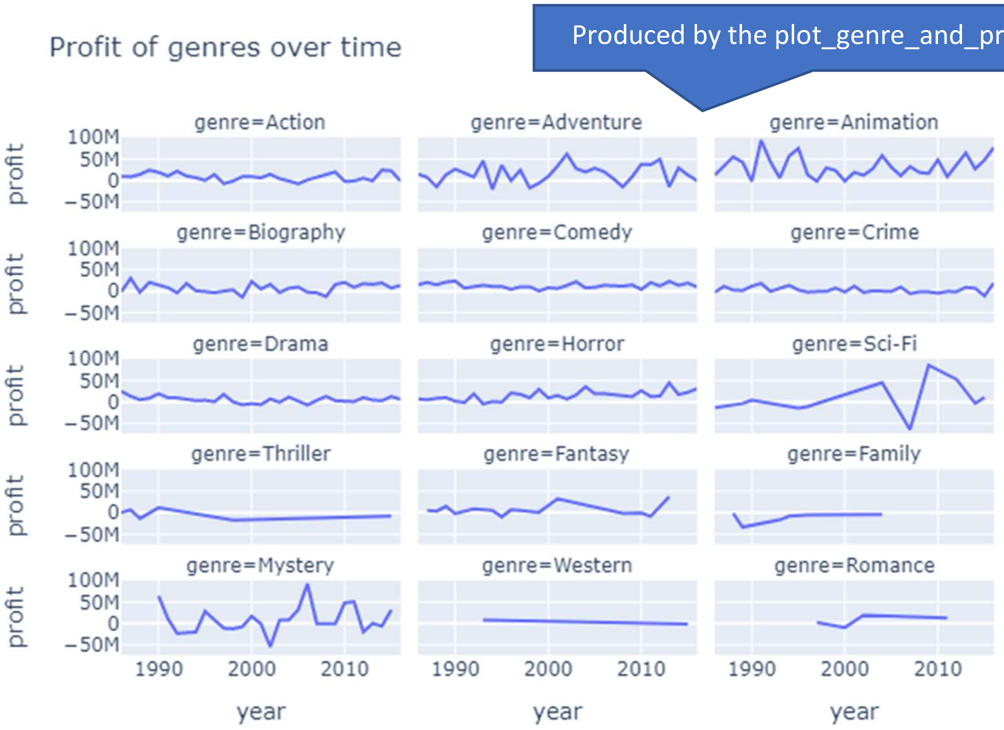
The project will then shift into determining a film’s profitability as it relates to the MPAA rating that the film has. This section continues to build on the theme of focusing in on a film’s profit in order to determine its success/failure, in a sense. We are curious to observe any trends in the distribution of profit dependent upon the film’s rating, in an attempt to get a better idea of how the rating that a film has (and, implicitly, the audience that it is geared towards) can impact the potential profit, if it has any impact at all. For visualization and data availability purposes, only films with G, PG, PG-13 or R ratings are considered for this part of the project.

Lastly, the project will entail constructing a handful of machine learning models to predict a film’s IMDB score, and determine the accuracy of each model. This is a regression problem, and the types of models that will be built will be linear regression, a random forest regressor, and a lasso regularization. The choice of hyperparameter(s) for the lasso model will be subject to cross validation, in order to determine the penalization hyperparameter that produces the best model for solving this problem. For the analysis, all columns besides the response variable (the IMDB score) and the name of the film (as this is a unique value for each film and would unnecessarily complicate the model) should be incorporated in the analysis. In order to avoid the issue of multicollinearity in this dataset as best as we can, we will avoid adding in any extra features that come from any of the previous calculations of the analysis.

5. **Results**

This section will be sub-divided into 4 sections, each one corresponding directly with the research question previously posed.

1) I did find that there were some significant changes in the profitability of some genres over time. Attached is a visualization of my results from this section:

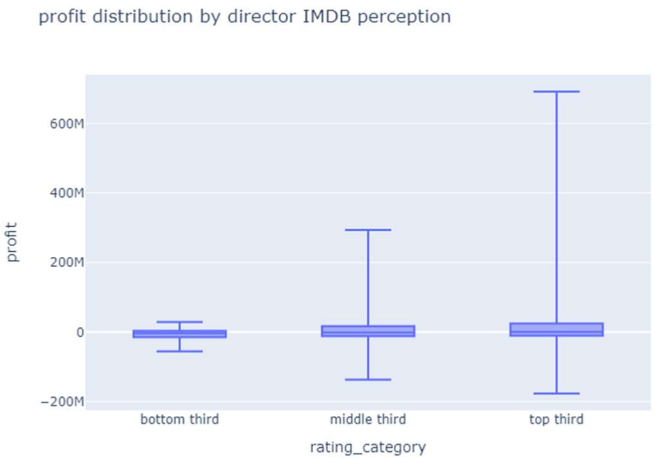


Produced by the plot_genre_and_profits method.

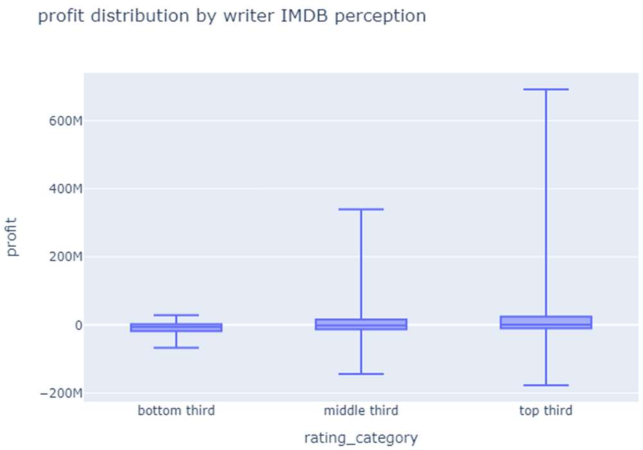
The graph is faceted according to what genre is contained within each subplot, with the release year being depicted on the x-axis and (average) profit being depicted on the y-axis. We can see that some genres (like Crime films) have been relatively stable over the past 30 years, and, interestingly enough, seem to have a collective average profit that is only slightly above zero. Some genres, like Westerns, have steadily, decreased in their average profits while horror has slowly increased its profitability over time. What stood out to me most in this component of the analysis was the massive dip that exists for the sci-fi genre from 2004 to

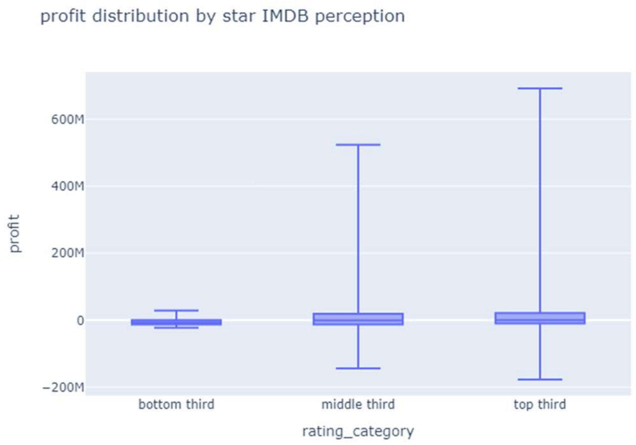
2007, the spike that mystery films saw in 2006 and the fact that animation is routinely netting the highest profits of any genre. While the actual drivers of these trends are somewhat ambiguous, speculation could be the factor of outliers (i.e., very poor sci-fi movies coming out in the late 2000's that didn't entice audiences to come out to the theaters for them), and the generalization that animated films tend to be targeted towards a wider, family audience (a claim that appears to be supported even more so in a later part of this section).

2) The investigation of this research question provided some honestly fascinating results. Below are the results obtained for each type of filmmaker, with director being displayed first, writer second and stars third:


profit distribution by director IMDB perception

(All 3) Produced by the plot_filmmaker_trends method

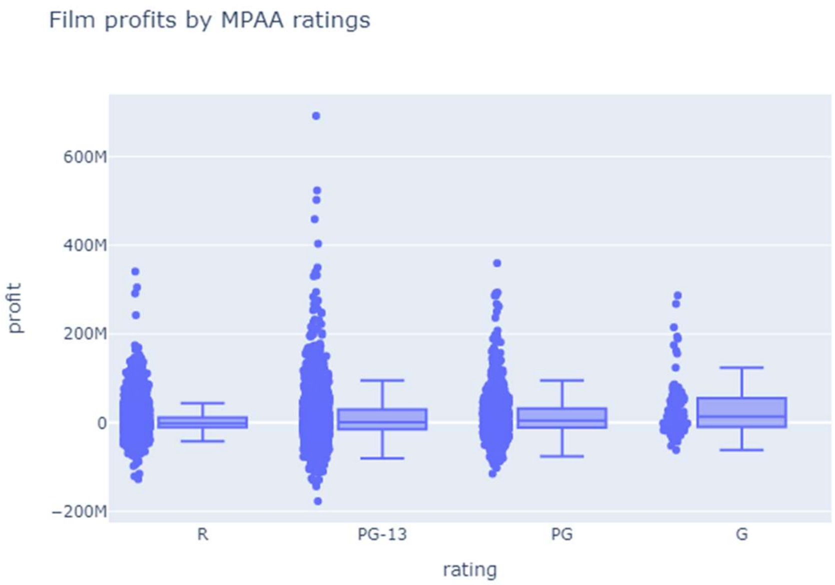

profit distribution by writer IMDB perception



Not shockingly, profit distributions tend to increase as the perception of the filmmaker increases—that is, a filmmaker who is perceived as a bottom third one tends to have much lower profit (at all quartiles, and the median) when compared to one who is in the top third. What did surprise me, was the lack of differences between all of these filmmakers and their profit distributions. For example, I genuinely anticipated that the top third of star actors/actresses would see a much higher median average profit than the other two types of filmmakers, but writers were actually the highest top third median! Another interesting result from this portion of the project was that the top third also contains the lowest profit for each type of filmmaker. In fact, that film was the notorious bomb that was "John Carter", and it's fascinating to me that that film, which is widely regarded as poor, was created by a writer, director and star that are generally considered to be at the top of the movie-making game. Lastly, although the median is not the end-all, be-all statistic, I expected for the median of a top third filmmaker to differ greatly from a bottom third one; alas, the median profit for any kind of filmmaker in the top third is hardly even positive, always being less than 1 million! I believe all of this goes to show that, while those associated with the production of a film do appear to be correlated with higher profit margins, there does not seem to be any causal effect that would be worth investigating further—it seems to be that, when all is said and done, a film is its own self-contained entity that cannot be carried to huge sums of cash by the star power thrown behind it.

3) My analysis then shifted towards looking at profit from a different angle, by observing the impact that a MPAA rating can have on profit. These results are

displayed below:



Produced by the plot_profit_by_rating method.

We once again have a box plot visualization, although this time I decided to also incorporate the actual data points alongside them, in order to try and get a feel for how the data is distributed. Right off the bat, one can notice that the PG-13 rating contains the films with the highest profits. This makes an abundance of sense, since the PG-13 rating is a rating that, while aimed towards more mature audiences, still has the strong potential to pull in younger viewers, thus giving this group of films a good viewer base to draw from. On the other hand, R rated films struggle the most, with a median profit that is actually negative. Using some of the same logic as before, this makes sense as the viewer base that these films can build off of is easily the most limited, open only to those who are 17 and older or minors whose parents/guardians they can convince to take them to the theater. What did catch me off guard a bit was the fact that G rated films tend to have the highest profits when compared to all other ratings. When looking at the visualization, we can see that G rated films appear to be the most infrequent that are made, which can certainly have an impact on why their distribution tends to be so positive. Moreover, continuing with the idea of viewer bases, a G rated film is the most accessible for everyone. These tend to be films that families can go out to see together, which could mean more ticket sales, as a family going out to the theaters entails multiple ticket purchases, as opposed to an individual going to view an R-rated film. Looking at this visualization also leads me to wondering why more production companies are not investing their resources into making these kinds of films—while their profit potential doesn't seem to be as high as that of a PG-13 movie, they appear to be a safer choice, and one would think that they would result in higher confidence that a film will make more than its money back.

4) Lastly, machine learning models were created to try and be able to predict a film's IMDB score based on other features presented within the data. The models that were constructed were a linear regression, a lasso regularization and a random forest regressor. The linear regression was easily the worst performing model, with an astronomically high mean squared error, and an astronomically low R2 value. This does make sense because it is highly (and I truly do mean highly) unlikely that the relationship between IMDB score and the predictors is linear, and it is also almost a guaranteed certainty that the data violates some of the key assumptions of an Ordinary Least Squares regression. In fact, this model was such a poor fit that it regularly predicted negative scores, which are not even possible to give on IMDB! I did consider simply setting these values to be 0 (or even taking the log of the IMDB scores to prevent negative values), but instead opted to retain this poorly performing model for demonstrative purposes within my analysis. In an attempt to improve the predictive capabilities, I transitioned to developing a lasso regression model that was subject to cross-validation. This model performed practically infinitely better by squashing the coefficients of many of the 928 features down to 0. This model had an R2 value of approximately .476, which is not too fantastic, but the testing mean squared error rate shot down to a relatively low .546. This was a huge improvement over the linear regression model, and my model was always within the positive range, so it was at least giving out reasonable predictions. Finally, I wanted to do just a little bit better, so I ran through the regression once more, this time by utilizing a random forest. This was the best performing model in terms of predictions, with a mean squared error of .438 and the highest R2 value of any model I conducted at .579. Although there is an interpretability tradeoff present between the lasso and the random forest (the lasso model is easier to interpret as a parametric model that creates an actual equation relating the score to the predictors), in terms of prediction, the random forest regressor reigned supreme. All of this suggests that a film's IMDB user rating score can be inferred on a combination of other factors, and that it is not independent of things like who was involved in making the film or how much it grossed. The reception of these films (at least on this one particular website) can be modeled pretty closely in relation to factors that are totally separate from how people liked, or did not like, the movie.

6. **Challenge goals**

   a. <u>Machine Learning:</u> I utilized machine learning techniques in the last part of my project, going beyond the scope of what was taught about it in this course. This was certainly aided by the fact that I have been learning machine learning theories concurrently in my economics course this quarter, but was also aided by an abundance of resources online that detailed how to work through some of the functions from the scikit-learn library that I employed in my project.

   b. <u>New Library:</u> I ended up learning, and utilizing, the plotly library for all of my visualizations. I became familiar with line plots and box plots, but also how to facet and highlight different components of my data in visualizations in this library. Honestly, I think I'll rely on it more so than matploltlib in the future, as I found it to be a bit more intuitive and easier to operate, while still having a lot of really great flexibility options for constructing data visualizations.

7. **Work Plan Evaluation**

   This is directly copied from Part 1 of my project. Each ii. Roman numeral corresponds to my evaluation of the proceeding outline that past me provided.

   a. Calculate each genre's mean profit over time
      i. I do not anticipate this particular part of the project taking up an abundance of time. I would predict that it would take about 1 to 2 hours in order to appropriately sort and then perform the calculations for this part of the analysis.
      ii. I was WRONG about this part not taking a long time. Perhaps writing this particular function did not, but I also completely glossed over having to load in my data, get it into a usable form and really think about how to visualize what I was trying to demonstrate. Moreover, this is where I decided that I would not be sticking to the challenge goal that I had previously detailed of working with two datasets. I found that one of the datasets I had intended to use was simply unwieldly—that isn't to say that it would be impossible to conduct these analyses on it, but it was very messy and as a combination of other classes that are really piling on, to some of the national unrest that is occurring right now, I felt that working with that second dataset was simply too heavy of a task for this project, for me, at this time. I'm eager to work on it in the future, at my own pace, but I knew that I would not be able to put together a cohesive project if I continued to press the issue of using that second dataset.

   b. Calculate director, writer, star film scores.
      i. I expect this portion of my project will have a bit heavier lifting—I'd estimate 2-3 hours required for this component, in order to work through how to filter/sort the data and perform these calculations. Moreover, I want to create a clear and accurate data visualization, and I'm not 100% sure how I want to portray this information yet (Whereas I am pretty certain of how I will visualize the previous task), so I expect to spend some time messing around with different ways to plot what I want to show.
      ii. I was pretty accurate on this one! This component did not take an abundance of time. The hardest part was definitely getting to visualize the data in a way that was easy to understand, but I believe my selection of box plots ended up lending itself well to that. I initially wanted all filmmakers to be represented on one graph, but I think my decision to split it into three separate ones provided a bit more clarity.

   c. Determine the reliability of user ratings.
      i. This is another part of my project where I expect to spend about 2 to 3 hours. I will have to merge these datasets together for this part, which

shouldn’t be too difficult, but may cause some extra time to debug and the such. I think the calculations will certainly be manageable, but, again, diving into how to reflect this information in a visualization may take some time to think on so that I am presenting the data in the most meaningful way possible.

ii. While the time spent portion of my statement is accurate, everything else here is incorrect, primarily because I did not use that second dataset. This was instead replaced with the “ratings and profit” piece of my analysis, which I think served the overall project well. Despite the fact that I had to struggle a little bit with getting plotly to graph this part right, I think it ended up going pretty well.

d. Construct machine learning models to predict a film’s revenue.

i. This is where I believe the largest chunk of my time will be spent in this project. I’m really fascinated by machine learning, and I could realistically see myself devoting 20-30 hours on this part of the project over the course of the next two weeks. I feel relatively comfortable with the theory behind employing the models that I want to use, but I’ll also be needing to verify that I am truly utilizing the tools correctly while writing my code. Moreover, I feel like I have a good grasp of these concepts in R, but it will definitely take me some time to understand how to translate those concepts over into python with scikit-learn! I’m eager to get to this part of the project, and am looking to complete the other parts of the analysis by early-to-mid next week so that I have ample time (7-10 days, ideally) in order to really focus in on machine learning and give myself plenty of time to debug and ask questions if need be.

ii. This (besides perhaps the decision to forgo the other dataset) was the hardest part of this project and did take the most amount of time (although it was probably closer to 20 hours). Firstly, I just could not develop an accurate model for predicting film revenue. I also spent time trying to learn neural nets in hopes that I could get some decent results, but every model was producing ridiculously high error rates, and I had to accept that maybe my question was ill-posed to be answered from the data I was working with. Once I transitioned to instead looking at IMDB scores, things moved along much more smoothly and I was able to obtain results that were much more in line with what I was hoping for. I spent a few of those hours reading through scikit documentation, but I actually think it’s a bit easier to work with than R (which is how I’m doing machine learning in my econometrics class). Also, my lasso model was taking an absurdly long amount of time to run (which may have to do with my computer), so I restricted the cross-validation by a bit in order to get code that could run much quicker. This was a tradeoff that could potentially impact the performance of that model, but it was one I decided was necessary to complete the project.

8. **Testing**

   I tried to test all functions that I wrote for the purposes of this project. This was done by taking a small subset of the larger dataset (40 rows), manually computing the return values that I needed for each function and then using the assert_equals functions provided to us in the course to test my methods. This subset is created in my testing py file. A couple of these testing functions did utilize the full dataset, but I would filter it down to a specific value and then test that my method was also returning that expected value. All expected values in my testing functions were, again, computed manually, so they are the correct expected outputs.

9. **Collaboration**

   I did not directly collaborate with any individuals on this assignment. All of this work was done on my own, although I did utilize some online resources (mostly online documentation for libraries, some various websites to understand more in-depth examples of scikit-learn in action, and a little bit of stack exchange for the purposes of trying to attain a stronger statistical understanding of how to work with my data with regards to machine learning).